Χειμερινό Εξάμηνο 2017/2018

ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΚΑΙ ΠΡΟΒΛΕΨΗΣ ΤΗΛΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΑΓΟΡΩΝ

Μιχαήλ Κυριακίδης

Ευχαριστούμε τον φοιτητή του μεταπτυχιακού κ. Εμμανουήλ Βάσιλα (Χ.Ε. 2011/12) για την επεξεργασία των "Cases" στο τέλος των κεφαλαίων.

Επίσης ευχαριστούμε την Χριστίνα Καδά και Ελένη Καλοειδά για τις εικόνες περιγραφής του MINITAB από τις σημειώσεις της παράδοσης των Κουτρουμπή - Μιχαλακέλη.

ΠΕΡΙΕΧΟΜΕΝΑ

ΒΙΒΛΙΟΓΡΑΦΙΑ.....6
ΚΕΦΑΛΑΙΟ 1 (ΕΙΣΑΓΩΓΗ ΣΤΗ ΠΡΟΒΛΕΨΗ).....7
1.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΠΡΟΒΛΕΨΗ.....7
1.2 ΟΡΙΑ ΜΕΘΟΔΩΝ ΠΡΟΒΛΕΨΗΣ.....9
1.3 Η ΚΡΙΣΗ ΣΤΗΝ ΠΡΟΒΛΕΨΗ.....10
1.4 ΕΙΝΑΙ Η ΠΡΟΒΛΕΨΗ ΑΠΑΡΑΙΤΗΤΗ;.....12
1.5 ΤΥΠΟΙ ΠΡΟΒΛΕΨΗΣ.....12
1.6 ΜΑΚΡΟΟΙΚΟΝΟΜΙΚΕΣ ΠΡΟΒΛΕΨΕΙΣ.....13
1.7 ΕΠΙΛΟΓΗ ΜΙΑΣ ΜΕΘΟΔΟΥ ΠΡΟΒΛΕΨΗΣ.....14
1.8 Η ΔΙΑΔΙΚΑΣΙΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ.....15
1.9 ΤΑ ΒΗΜΑΤΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ.....15
1.10 ΛΟΓΙΣΜΙΚΟ ΠΡΟΒΛΕΨΗΣ.....17
1.11 ΤΙ ΜΠΟΡΕΙ ΝΑ ΠΡΟΒΛΕΦΘΕΙ.....18
1.12 ΠΑΡΑΔΕΙΓΜΑ ΠΡΟΒΛΕΨΗΣ.....19
1.13 ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ ΠΟΥ ΘΑ ΑΝΑΠΤΥΧΘΟΥΝ.....22
1.14 ΣΥΜΠΕΡΑΣΜΑ.....24
ΚΕΦΑΛΑΙΟ 2 (ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ).....25
2.1 ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΒΑΣΙΚΕΣ ΣΤΑΤΙΣΤΙΚΕΣ ΕΝΝΟΙΕΣ.....25
2.2 ΚΑΤΑΝΟΜΕΣ.....39
2.3 ΔΕΙΓΜΑΤΟΛΗΨΙΑ.....50
2.3.1 ΔΕΙΓΜΑΤΙΚΗ ΜΕΣΗ ΤΙΜΗ.....51
2.3.2 ΚΑΤΑΝΟΜΗ ΤΗΣ ΔΕΙΓΜΑΤΙΚΗΣ ΜΕΣΗΣ ΤΙΜΗΣ.....52
2.3.3 ΔΕΙΓΜΑΤΙΚΗ ΔΙΑΣΠΟΡΑ.....53
2.3.4 ΑΓΝΩΣΤΗ ΔΙΑΣΠΟΡΑ ΠΛΗΘΥΣΜΟΥ.....53
2.4 ΠΑΡΟΥΣΙΑΣΗ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ.....55
2.5 ΕΚΤΙΜΗΤΙΚΗ.....57
2.5.1 ΤΟ ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ.....58
2.5.2 ΣΤΑΤΙΣΤΙΚΕΣ ΥΠΟΘΕΣΕΙΣ ΚΑΙ ΔΙΑΔΙΚΑΣΙΑ ΕΛΕΓΧΟΥ.....61
2.5.3 ΕΛΕΓΧΟΙ.....65
2.5.4 Η P – ΤΙΜΗ ΤΟΥ ΕΛΕΓΧΟΥ.....69
2.6 ΣΥΣΧΕΤΙΣΗ.....71
2.7 ΠΑΛΙΝΔΡΟΜΗΣΗ.....75
2.7.1 Η ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....75
2.7.2 ΠΡΟΣΑΡΜΟΓΗ ΜΙΑΣ ΕΥΘΕΙΑΣ ΣΤΑ ΔΕΔΟΜΕΝΑ.....77
2.8 ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΕΚΤΙΜΗΣΗΣ ΕΥΘΕΙΑΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ...80
2.9 PROBABILITY PLOT.....81
2.10 ΧΡΗΣΗ MINITAB.....82
ΚΕΦΑΛΑΙΟ 3 (ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΜΕΘΟΔΟΥΣ ΠΡΟΒΛΕΨΗΣ).....95
3.1 ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ.....95
3.2 ΠΟΣΟΤΙΚΑ – ΠΟΙΟΤΙΚΑ ΜΟΝΤΕΛΑ.....98
3.2.1 ΠΟΣΟΤΙΚΑ ΜΟΝΤΕΛΑ.....98
3.2.2 ΠΟΙΟΤΙΚΑ ΜΟΝΤΕΛΑ.....101
3.3 ΣΥΝΕΧΕΙΣ – ΔΙΑΚΡΙΤΕΣ ΧΡΟΝΟΛΟΓΙΚΕΣ ΣΕΙΡΕΣ.....101
3.4 ΣΥΝΙΣΤΩΣΕΣ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ.....102

3.4.1 ΤΑΣΗ....102
3.4.2 ΚΥΚΛΙΚΗ ΣΥΝΙΣΤΩΣΑ....103
3.4.3 ΕΠΟΧΙΚΗ ΣΥΝΙΣΤΩΣΑ....104
3.4.4 ΤΥΧΑΙΑ ΣΥΝΙΣΤΩΣΑ....105
3.5 ΣΤΑΣΙΜΟΤΗΤΑ....107
3.5.1 ΟΡΙΣΜΟΙ....107
3.5.2 ΕΛΕΓΧΟΣ ΣΤΑΣΙΜΟΤΗΤΑΣ ΧΡΟΝΟΣΕΙΡΑΣ....108
3.6 ΑΥΤΟΣΥΣΧΕΤΙΣΗ....108
3.7 ΜΕΘΟΔΟΣ ΤΩΝ ΔΙΑΦΟΡΩΝ....113
3.8 Έλεγχος Καταλοιπων....114
3.9 Έλεγχος Σημαντικοτητας Συντελεστων Αυτοσυσχετισης....114
3.10 ΒΑΣΙΚΑ ΣΤΑΔΙΑ ΣΤΗ ΔΙΑΔΙΚΑΣΙΑ ΠΡΟΒΛΕΨΗΣ....125
3.12 Η ΕΠΙΛΟΓΗ ΜΙΑΣ ΤΕΧΝΙΚΗΣ ΠΡΟΒΛΕΨΗΣ....127
3.13 ΑΞΙΟΛΟΓΗΣΗ ΠΡΟΒΛΕΨΕΩΝ....131
3.14 CASE 3-1A MURPHY BROTHERS FURNITURE....134
3.15 ΧΡΗΣΗ MINITAB....140
ΚΕΦΑΛΑΙΟ 4 (ΚΙΝΗΤΟΙ ΜΕΣΟΙ ΚΑΙ ΜΕΘΟΔΟΙ ΕΞΟΜΑΛΥΝΣΗΣ) 144
4.1 ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ....144
4.2 ΜΕΘΟΔΟΣ NAIVE ΚΑΙ ΟΙ ΠΑΡΑΛΛΑΓΕΣ ΤΗΣ....146
4.3 ΜΕΘΟΔΟΙ ΜΕΣΟΥ ΟΡΟΥ....152
4.3.1 ΑΠΛΟΙ ΜΕΣΟΙ....152
4.3.2 ΚΙΝΗΤΟΙ ΜΕΣΟΙ....154
4.3.3 ΔΙΠΛΟΙ ΚΙΝΗΤΟΙ ΜΕΣΟΙ....157
4.4 ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ (ΑΠΛΗ)....161
4.5 ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ ΜΕ ΠΡΟΣΑΡΜΟΓΗ ΣΤΗΝ ΤΑΣΗ (ΔΙΠΛΗ)...169
4.6 ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ ΜΕ ΠΡΟΣΑΡΜΟΓΗ ΣΤΗΝ ΤΑΣΗ ΚΑΙ ΣΤΗΝ ΕΠΟΧΙΚΟΤΗΤΑ (ΤΡΙΠΛΗ)....175
4.7 CASE 4-4 MURPHY BROTHERS FURNITURE....180
4.8 ΧΡΗΣΗ MINITAB....184
ΚΕΦΑΛΑΙΟ 5 (ΧΡΟΝΟΣΕΙΡΕΣ ΚΑΙ ΤΑ ΣΥΣΤΑΤΙΚΑ ΤΟΥΣ).....191
5.1 ΔΙΑΣΠΑΣΗ ΧΡΟΝΟΣΕΙΡΩΝ....191
5.2 ΜΟΝΤΕΛΑ ΔΙΑΣΠΑΣΗΣ ΧΡΟΝΟΣΕΙΡΩΝ....193
5.3 ΠΡΟΣΑΡΜΟΓΗ ΚΑΙ ΑΝΑΛΥΣΗ ΤΗΣ ΤΑΣΗΣ....195
5.4 ΜΟΝΤΕΛΑ ΤΑΣΗΣ....196
5.4.1 ΓΡΑΜΜΙΝΟ ΜΟΝΤΕΛΟ....197
5.4.2 ΠΑΡΑΒΟΛΙΚΟ ΜΟΝΤΕΛΟ....199
5.4.3 ΕΚΘΕΤΙΚΟ ΜΟΝΤΕΛΟ....200
5.4.4 ΜΟΝΤΕΛΟ GOMPERTZ....201
5.4.5 ΜΟΝΤΕΛΟ PEARL REED LOGISTIC....203
5.5 ΑΝΑΛΥΣΗ ΚΑΙ ΔΙΑΧΩΡΙΣΜΟΣ ΤΗΣ ΕΠΟΧΙΚΟΤΗΤΑΣ....204
5.6 ΑΠΟΜΟΝΩΣΗ ΤΗΣ ΚΥΚΛΙΚΟΤΗΤΑΣ....209
5.7 ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΙ ΔΕΙΚΤΕΣ....213
5.8 ΔΕΙΚΤΕΣ ΤΙΜΩΝ....215
5.9 Η ΜΕΘΟΔΟΣ ΑΠΟΣΥΝΘΕΣΗΣ CENSUS II....216
5.10 CASE 5-4 MURPHY BROTHERS FURNITURE....218

5.11 ΧΡΗΣΗ MINITAB.....229
ΚΕΦΑΛΑΙΟ 6 (ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ).....234
6.1 ΕΙΣΑΓΩΓΗ.....234
6.2 ΕΥΘΕΙΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ (ΣΤΑΤΙΣΤΙΚΟ ΜΟΝΤΕΛΟ).....236
6.3 ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΕΚΤΙΜΗΣΗΣ.....241
6.4 ΠΡΟΒΛΕΠΟΝΤΑΣ ΤΟ Y.....242
6.5 ΑΝΑΛΥΣΗ ΤΗΣ ΔΙΑΣΠΟΡΑΣ.....248
6.6 ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ.....253
6.7 ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ.....257
6.8 ΑΝΑΛΥΣΗ ΚΑΤΑΛΟΙΠΩΝ.....262
6.9 ΥΠΟΛΟΓΙΣΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ.....266
6.10 ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΙ ΜΕΤΑΒΛΗΤΩΝ.....267
6.11 ΚΑΜΠΥΛΕΣ ΑΝΑΠΤΥΞΗΣ (GROWTH CURVES).....274
6.12 CASE 6-2 BUTCHER PRODUCTS, INC.....278
ΚΕΦΑΛΑΙΟ 7 (ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΠΑΛΙΝΔΡΟΜΗΣΗ).....286
7.1 ΜΕΡΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ ΠΡΟΒΛΕΨΗΣ.....286
7.2 ΠΙΝΑΚΑΣ ΣΥΣΧΕΤΙΣΗΣ.....287
7.3 ΜΟΝΤΕΛΟ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....289
7.3.1 ΣΤΑΤΙΣΤΙΚΟ ΜΟΝΤΕΛΟ ΓΙΑ ΤΗΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.....291
7.4 ΕΡΜΗΝΕΙΑ ΣΥΝΤΕΛΕΣΤΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....293
7.5 ΣΥΜΠΕΡΑΣΜΑΤΑ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....295
7.5.1 ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΕΚΤΙΜΗΣΗΣ.....297
7.5.2 ΣΗΜΑΝΤΙΚΟΤΗΤΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....299
7.5.3 ΜΕΜΟΝΩΜΕΝΕΣ ΜΕΤΑΒΛΗΤΕΣ ΠΡΟΒΛΕΨΗΣ.....301
7.5.4 ΠΡΟΒΛΕΨΗ ΜΕΛΛΟΝΤΙΚΩΝ ΤΙΜΩΝ.....303
7.6 ΥΠΟΛΟΓΙΣΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ.....303
7.7 ΨΕΥΔΟΜΕΤΑΒΛΗΤΕΣ.....306
7.8 ΠΟΛΥΣΥΓΓΡΑΜΙΚΟΤΗΤΑ.....311
7.9 ΕΠΙΛΕΓΟΝΤΑΣ ΤΗΝ ΚΑΛΥΤΕΡΗ ΕΞΙΣΩΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....317
7.9.1 ΟΛΕΣ ΟΙ ΠΙΘΑΝΕΣ ΠΑΛΙΝΔΡΟΜΗΣΕΙΣ.....321
7.9.2 ΒΗΜΑ ΠΡΟΣ ΒΗΜΑ ΠΑΛΙΝΔΡΟΜΗΣΗ.....324
7.9.3 ΑΝΑΣΤΡΟΦΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.....329
7.9.4 ΕΜΠΡΟΣΘΙΑ ΠΑΛΙΝΔΡΟΜΗΣΗ.....331
7.9.5 ΤΕΛΙΚΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ ΓΙΑ ΤΗ ΒΗΜΑ – ΠΡΟΣ – ΒΗΜΑ ΠΑΛΙΝΔΡΟΜΗΣΗ.....331
7.10 ΔΙΑΓΝΩΣΤΙΚΕΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΑΙ ΑΝΑΛΥΣΗ ΚΑΤΑΛΟΙΠΩΝ.....332
7.11 ΠΡΟΕΙΔΟΠΟΙΗΣΕΙΣ ΠΡΟΒΛΕΨΗΣ.....338
7.11.1 OVERFITTING.....338
7.11.2 ΧΡΗΣΙΜΕΣ ΠΑΛΙΝΔΡΟΜΗΣΕΙΣ – ΜΕΓΑΛΟΙ F-ΛΟΓΟΙ.....338
7.12 CASE 7-1 THE BOND MARKET.....339
ΚΕΦΑΛΑΙΟ 8 (ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕΤΑΞΥ ΧΡΟΝΟΣΕΙΡΩΝ).....348
8.1 ΔΕΔΟΜΕΝΑ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΑΥΤΟΣΥΣΧΕΤΙΣΗ.....348
8.2 ΣΕΙΡΙΑΚΗ ΑΥΤΟΣΥΣΧΕΤΙΣΗ.....350
8.3 ΦΑΙΝΟΜΕΝΗ ΑΥΤΟΣΥΣΧΕΤΙΣΗ – ΛΑΝΘΑΝΟΥΣΑ ΠΑΛΙΝΔΡΟΜΗΣΗ....352

8.4 ΑΥΤΟΣΥΣΧΕΤΙΣΗ ΚΑΙ ΕΛΕΓΧΟΣ DURBIN WATSON.....356
8.5 ΛΥΣΕΙΣ ΓΙΑ ΠΡΟΒΛΗΜΑΤΑ ΑΥΤΟΣΥΣΧΕΤΙΣΗΣ.....363
8.5.1 MODEL SPECIFICATION ERROR.....365
8.5.2 ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΔΙΑΦΟΡΕΣ.....369
8.5.3 ΑΥΤΟΣΥΣΧΕΤΙΖΟΜΕΝΑ ΣΦΑΛΜΑΤΑ-ΓΕΝΙΚΕΥΜΕΝΕΣ ΔΙΑΦΟΡΕΣ 376
8.5.4 ΜΟΝΤΕΛΑ ΑΥΤΟΠΑΛΙΝΔΡΟΜΗΣΗΣ.....379
8.5.5 ΣΥΝΟΨΗ.....381
8.6 ΔΕΔΟΜΕΝΑ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΕΤΕΡΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑΣ.....382
8.7 ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΕΠΟΧΙΚΩΝ ΔΕΔΟΜΕΝΩΝ.....387
8.8 ΟΙΚΟΝΟΜΕΤΡΙΚΕΣ ΠΡΟΒΛΕΨΕΙΣ.....390
8.9 CASE 8-2 BUSINESS ACTIVITY INDEX FOR SPOKANE COUNTY.....392
ΚΕΦΑΛΑΙΟ 9 (ΜΕΘΟΔΟΛΟΓΙΑ BOX – JENKINS (ARIMA)).....405
9.1 ΜΕΘΟΔΟΛΟΓΙΑ BOX – JENKINS.....406
9.1.1 ΜΕΡΙΚΗ ΑΥΤΟΣΥΣΧΕΤΙΣΗ.....408
9.1.2 ΑΥΤΟΠΑΛΙΝΔΡΟΜΙΚΑ ΜΟΝΤΕΛΑ.....414
9.1.3 ΜΟΝΤΕΛΑ ΚΙΝΗΤΟΥ ΜΕΣΟΥ.....416
9.1.4 ΑΥΤΟΠΑΛΙΝΔΡΟΜΙΚΑ ΜΟΝΤΕΛΑ ΚΙΝΗΤΟΥ ΜΕΣΟΥ.....420
9.1.5 ΟΛΟΚΛΗΡΩΜΕΝΗ ΔΙΑΔΙΚΑΣΙΑ.....421
9.1.6 ΣΥΝΟΨΗ.....422
9.2 ΣΤΡΑΤΗΓΙΚΗ ΔΗΜΙΟΥΡΓΙΑΣ ΜΟΝΤΕΛΟΥ.....422
9.2.1 ΒΗΜΑ 1: ΑΝΑΓΝΩΡΙΣΗ ΜΟΝΤΕΛΟΥ.....423
9.2.2 ΒΗΜΑ 2: ΕΚΤΙΜΗΣΗ ΜΟΝΤΕΛΟΥ.....426
9.2.3 ΒΗΜΑ 3: ΕΛΕΓΧΟΣ ΜΟΝΤΕΛΟΥ.....428
9.2.4 ΒΗΜΑ 4: ΠΡΟΒΛΕΠΟΝΤΑΣ ΜΕ ΤΟ ΜΟΝΤΕΛΟ.....430
9.2.5 ΠΑΡΑΔΕΙΓΜΑΤΑ.....431
9.2.6 ΠΕΡΙΟΡΙΣΤΙΚΟΙ ΟΡΟΙ.....461
9.2.7 ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ.....462
9.2.8 ARIMA ΜΟΝΤΕΛΑ ΓΙΑ ΕΠΟΧΙΚΑ ΔΕΔΟΜΕΝΑ.....465
9.2.9 ΣΧΕΣΗ ΕΚΘΕΤΙΚΗΣ ΕΞΟΜΑΛΥΝΣΗΣ ΚΑΙ ARIMA ΜΟΝΤΕΛΩΝ.....488
9.2.10 ΠΛΕΟΝΕΚΤΗΜΑΤΑ – ΜΕΙΟΝΕΚΤΗΜΑΤΑ ARIMA ΜΟΝΤΕΛΩΝ....489
9.3 CASE 9-4 THE LYDIA E. PINKHAM MEDICINE COMPANY.....490
ΚΕΦΑΛΑΙΟ 10 (ΘΕΩΡΙΑ ΥΙΟΘΕΤΗΣΗΣ ΚΑΙ ΔΙΑΧΥΣΗΣ – ADOPTION AND DIFFUSION THEORY).....503
1.1 ΕΙΣΑΓΩΓΗ.....503
1.2 ΘΕΩΡΙΑ ΥΙΟΘΕΤΗΣΗΣ ΚΑΙ ΔΙΑΧΥΣΗΣ.....504
1.3 ΥΠΟΔΕΙΓΜΑ BASS.....506
1.4 ΒΑΣΙΚΕΣ ΥΠΟΘΕΣΕΙΣ ΤΟΥ ΜΟΝΤΕΛΟΥ.....506
1.5 ΙΔΙΟΤΗΤΕΣ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ.....509
1.6 ΤΟ ΔΙΑΚΡΙΤΟ ΑΝΑΛΟΓΟ.....511
1.7 ΠΑΡΑΔΕΙΓΜΑ ΕΥΡΥΖΩΝΙΚΟΤΗΤΑΣ.....512

ΒΙΒΛΙΟΓΡΑΦΙΑ

Οι σημειώσεις ως επί το πλείστον προέρχονται από μετάφραση του παρακάτω βιβλίου:

BUSINESS FORECASTING

(John Hanke, Dean Wichern)

Έχει προστεθεί όμως και πολύ υλικό από την ακόλουθη βιβλιογραφία:

SIMULATION MODELING AND ANALYSIS

(Averill Law, David Kelton)

ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗ

(Murray Spiegel)

SPSS REGRESSION ANALYSIS BY EXAMPLE

(Walter A. Shewhart and Samuel S. Wilks)

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΟΙΚΟΝΟΜΕΤΡΙΑ

(Γεώργιος Χρήστου)

ΤΕΧΝΙΚΕΣ ΠΡΟΒΛΕΨΗΣ ΤΗΛΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΑΓΟΡΩΝ

(Κουτρουμπής, Μιχαλακέλης)

A NEW PRODUCT GROWTH FOR MODEL CONSUMER DURABLES

BASS 1969, **MANAGEMENT SCIENCE,** Vol. **15,** No **5,** p. **215**

ΚΕΦΑΛΑΙΟ 1 (ΕΙΣΑΓΩΓΗ ΣΤΗ ΠΡΟΒΛΕΨΗ)

1.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΠΡΟΒΛΕΨΗ

«*Πρόβλεψη*», μια λέξη που ακούγεται όλο και πιο συχνά στη σημερινή εποχή. Κατά ορισμό, πρόβλεψη είναι η διαδικασία της εκτίμησης ενός μεγέθους σε άγνωστες καταστάσεις. Καθ΄ όλη τη διάρκεια του μαθήματος θα χρησιμοποιήσουμε μεθόδους (μαθηματικές) για την πρόβλεψη τάσεων στην αγορά αβέβαιης φύσης.

Επίσης πρέπει να έχουμε πάντα στο μυαλό μας, ότι η κρίση μας πρέπει να χρησιμοποιείται παράλληλα με τα αριθμητικά αποτελέσματα, αν θέλουμε στο τέλος να έχουμε καλές προβλέψεις, αλλά και να χρησιμοποιήσουμε τις προβλέψεις με τον πιο ωφέλιμο τρόπο.

Όλοι οι άνθρωποι από διαφορετική σκοπιά και για διαφορετικούς λόγους ο καθένας αναζητούν μια πρόβλεψη. Μια εκτίμηση που θα τους βοηθήσει να λάβουν μια απόφαση με μικρότερο ρίσκο. Μικρότερο ρίσκο στην δουλειά τους, σε μία επικείμενη αγορά, στον προγραμματισμό των διακοπών τους, στην επιχείρησή τους, στη ζωή τους.

Κατά τη διάρκεια του 17 αιώνα οι προβλέψεις αποτελούσαν τη καινοτομία της εποχής και τις χρησιμοποιούσαν αρκετοί, οι οποίοι ήθελαν να κάνουν κάποια εκτίμηση του αποτελέσματος του εγχειρήματός τους.

Οι ραγδαίες αλλαγές σε όλους τους τομείς της καθημερινής ζωής και η αβεβαιότητα για την εξέλιξη των καταστάσεων στο μέλλον, έχουν κεντρίσει το ενδιαφέρον τόσο του επιστημονικού – ακαδημαϊκού κόσμου, όσο και του επιχειρηματικού κόσμου. Από την πλευρά των ακαδημαϊκών δημιουργήθηκαν πολλές διαφορετικές μέθοδοι πρόβλεψης, κάποιες εξ΄

ολοκλήρου βασιζόμενες σε θεωρητικό υπόβαθρο και κάποιες άλλες βασιζόμενες περισσότερο στα διαθέσιμα τεχνολογικά μέσα και στις δυνατότητες, που αυτά διαθέτουν. Οι διάφορες μέθοδοι επιλέγονται και εφαρμόζονται ανάλογα με συγκεκριμένες παραμέτρους, μεταβλητές, ή παράγοντες, όπως είναι:

- Το αντικείμενο της πρόβλεψης.
- Ο χρονικός ορίζοντας της πρόβλεψης.
- Τα ήδη υπάρχοντα δεδομένα και το πρότυπο συμπεριφοράς αυτών.
- Το κόστος για την παραγωγή των προβλέψεων.
- Η ευκολία χρήσης της επιλεγμένης μεθόδου πρόβλεψης.
- Η αποδεδειγμένη αξιοπιστία της μεθόδου.

Στην πραγματικότητα αυτό σημαίνει ότι η αληθινή αξία μιας μεθόδου αποτιμάται, όταν εφαρμόζεται σε πραγματικές καταστάσεις. Έτσι, η ολοκληρωμένη εικόνα για μια μέθοδο δημιουργείται μόνο, όταν εφαρμοστεί η θεωρία στην πράξη. Μια συνεργασία λοιπόν ακαδημαϊκού και επιχειρηματικού κόσμου μπορεί να επιφέρει καλά αποτελέσματα. Και αυτό μπορεί να το καταλάβει πολύ απλά ο οποιοσδήποτε, πόσο καλύτερα μοντέλα πρόβλεψης μπορούν να προκύψουν, αν αυτά σχεδιαστούν βάσει των ιδιαίτερων αναγκών της εκάστοτε επιχείρησης.

Από την πλευρά του επιχειρησιακού περιβάλλοντος είναι ολοφάνερη η σημαντικότητα της πρόβλεψης, αφού μια επιχείρηση βρίσκεται καθημερινά στην ανάγκη για λήψη μιας απόφασης αντιμετωπίζοντας, όσο το δυνατόν περισσότερο την αβεβαιότητα του μέλλοντος. Η πρόβλεψη διαδραματίζει πρωτεύοντα ρόλο σε όποια δραστηριότητα λήψης αποφάσεων, καθορισμού, σχεδιασμού και υλοποίησης στρατηγικής, χρονικού προγραμματισμού, καθώς και σχεδιασμού πολιτικής λειτουργίας της επιχείρησης. Είναι σημαντικό να

τονίσουμε, ότι έστω και μία πολύ μικρή βελτίωση στην ακρίβεια των προβλέψεων μπορεί να αποφέρει τεράστια ποσά κέρδους σε μία επιχείρηση. Για αυτό το λόγο οι διάφορες υπηρεσίες πρόβλεψης και τα διάφορα προϊόντα για την παραγωγή προβλέψεων, που γενικά κυκλοφορούν στην αγορά, γνωρίζουν μεγάλη ζήτηση και ταυτόχρονα αναπτύσσονται νέες διαδικασίες πρόβλεψης, λόγω της ανάγκης για ακριβείς προβλέψεις και ιδιαίτερα σε επιχειρήσεις.

Για να μπορέσει όμως κάποιος να χρησιμοποιήσει προς όφελός του τις προβλέψεις, πρέπει να μπορεί να κάνει την πιο κατάλληλη επιλογή. Και αυτό είναι απόρροια της πείρας. Τα διοικητικά στελέχη μεγάλων επιχειρήσεων πρέπει να διαθέτουν την κατάλληλη εμπειρία στις καταστάσεις, που έχουν να αντιμετωπίσουν και οι αποφάσεις τους να μην επηρεάζονται από προσωπικά ενδιαφέροντα, φιλοδοξίες, σκοπιμότητες και μονόπλευρη εξέταση του περιβάλλοντος. Για να επιτευχθεί ο εκάστοτε στόχος πρέπει οι αποφάσεις τους να είναι όσο το δυνατόν πιο αντικειμενικές και να μην βασίζονται μόνο στα "πειραματικά", ή μόνο στα θεωρητικά αποτελέσματα, αλλά να έχουν την ικανότητα να συγκρίνουν και να συνδυάζουν τα αποτελέσματα και των δύο και να λαμβάνουν υπόψη τους τόσο τις οικονομικές καταστάσεις του κλάδου τους όσο και των χωρών που θα ενεργοποιηθούν.

1.2 ΟΡΙΑ ΜΕΘΟΔΩΝ ΠΡΟΒΛΕΨΗΣ

Σε κάποιες περιπτώσεις υπάρχει μεγάλη δυσαρέσκεια σχετικά με τα πολύ μεγάλα σφάλματα των προβλέψεων, όπως και στην *αδυναμία των διαφόρων μεθόδων να προβλέψουν και να προειδοποιήσουν έγκαιρα για επερχόμενες αλλαγές και ανωμαλίες*. Στην πραγματικότητα όμως το περιβάλλον, στο οποίο ζούμε, χαρακτηρίζεται από συνεχείς μεταβολές,

εξελίξεις, μεγάλες διακυμάνσεις και απρόσμενες καταστάσεις και συνθήκες, που αυξάνουν την ανάγκη για προβλέψεις. Ο κύριος λόγος που υπάρχει αυτή η αιωρούμενη δυσανασχέτηση και οι δυσμενείς κριτικές για τις προβλέψεις, είναι ότι πολλοί πιστεύουν, ότι μπορούν να προβλεφθούν τα πάντα. Αυτό όμως καταλήγει να είναι ένας μύθος. Η ύπαρξη σφαλμάτων είναι αναπόφευκτη. Για αυτό το λόγο υπάρχουν αποκλίσεις μεταξύ προβλέψεων και πραγματικών αποτελεσμάτων. Το μόνο που μπορεί να γίνει, για να είναι πιο ωφέλιμα τα αποτελέσματα, είναι να γίνεται, όσο είναι εφικτό φυσικά, ρεαλιστική εκτίμηση των παραμέτρων.

Θα ασχοληθούμε, λοιπόν, με τις μεθόδους που χρησιμοποιούνται για την πρόβλεψη των αβέβαιων επιχειρηματικών τάσεων, έτσι ώστε να βοηθήσουν το management να λάβει καλύτερες αποφάσεις. Οι μέθοδοι αυτές αφορούν συνήθως στη μελέτη ιστορικών στοιχείων και στην επεξεργασία τους και γίνονται προβλέψεις μέσω των κατάλληλων υποδειγμάτων. Κατά τη διάρκεια του μαθήματος θα αναφερθούν λοιπόν διάφοροι τρόποι εξαγωγής προβλέψεων, που βασίζονται σε λογικές μεθόδους διαχείρισης δεδομένων, τα οποία προέρχονται από ιστορικά γεγονότα.

Η ανάλυση παλινδρόμησης και η αυτοπαλινδρόμηση είναι παραδείγματα διαδικασιών πρόβλεψης, που βασίζονται σε δεδομένα και θα συζητηθούν παρακάτω. Αυτές οι διαδικασίες έχουν αποδειχθεί ότι είναι πολύ αποτελεσματικές και συχνά εμφανίζονται σε διάφορα λογισμικά πακέτα.

1.3 Η ΚΡΙΣΗ ΣΤΗΝ ΠΡΟΒΛΕΨΗ

Μαζί με την ανάπτυξη των μεθόδων, που βασίζονται στα

δεδομένα, ο ρόλος της ανθρώπινης κρίσης έχει αυξηθεί σημαντικά κατά τα τελευταία 25 χρόνια. Χωρίς κανένα ιστορικό δεδομένο, η ανθρώπινη κρίση μπορεί να είναι ο μόνος τρόπος, για να γίνουν προβλέψεις για το μέλλον. Σε περιπτώσεις όπου τα δεδομένα είναι διαθέσιμα, η κρίση πρέπει να χρησιμοποιείται για επανεξέταση, αλλά και για να τροποποιεί τις προβλέψεις, που εξάγονται από ποσοτικές διαδικασίες.

Με τον πολλαπλασιασμό των υπολογιστών και τη διαθεσιμότητα των εξελιγμένων λογισμικών πακέτων οι προβλέψεις εξάγονται πολύ εύκολα. Ωστόσο, αυτή η ευκολία των υπολογισμών δεν αντικαθιστά την σκέψη και τη κρίση. Η έλλειψη της εποπτείας και η καταχρηστική χρήση των τεχνικών πρόβλεψης μπορεί να οδηγήσει μια επιχείρηση σε δαπανηρές αποφάσεις.

Από την μια μεριά, βρίσκεται το στέλεχος που λόγω άγνοιας και φόβου των ποσοτικών μεθόδων και των υπολογιστών στηρίζεται αποκλειστικά στη διαίσθηση του. Από την άλλη μεριά, βρίσκεται η πρόγνωση του αναλυτή που βασίζεται στις τελευταίες εξελιγμένες τεχνικές, αλλά είτε δεν έχει την ικανότητα, ή την εμπειρία, ή είναι απρόθυμος να συνδυάσει τη διαδικασία πρόβλεψης με τις ανάγκες της επιχείρησης, ή και του κράτους ανάλογα με το αντικείμενο, ή και να κρίνει αν αυτό, που πρόβλεψε, μπορεί να είναι αληθινό ή όχι.

Συνοψίζοντας λοιπόν ο πιο αποτελεσματικός αναλυτής πρέπει να είναι σε θέση να διαμορφώσει ένα συνδυασμό ποσοτικών και ποιοτικών δεδομένων, προκειμένου να αποφεύγονται μεγάλα σφάλματα.

Θεωρούμε ότι οι ποσοτικές μέθοδοι πρόβλεψης, που θα συζητηθούν στη συνέχεια, είναι μόνο η αρχή για την αποτελεσματική πρόβλεψη των αποτελεσμάτων μιας επιχείρησης. Η ανάλυση, η ανθρώπινη κρίση, η κοινή λογική και η εμπειρία πρέπει να κατευθύνουν

τη διαδικασία της πρόβλεψης.

1.4 ΕΙΝΑΙ Η ΠΡΟΒΛΕΨΗ ΑΠΑΡΑΙΤΗΤΗ;

Αν λάβουμε υπόψη μας ότι μία τράπεζα για να κάνει ρεαλιστικά πλάνα πρέπει να προβλέψει μελλοντικές ισορροπίες μεταξύ δανείων και καταθέσεων, ότι οι κεντρικές τράπεζες, για να ορίσουν τα επιτόκια, πρέπει να προβλέψουν μελλοντικές αυξήσεις ΑΕΠ, ή και υφέσεις, όπως και τις πληθωριστικές πιέσεις, ένας Operational Manager για να σχεδιάσει τη παραγωγή πρέπει να διαθέτει προβλέψεις πωλήσεων, ή μία εταιρεία για να ορίσει το προσωπικό της πρέπει να προβλέψει μελλοντικές πωλήσεις των υπηρεσιών της γίνεται εμφανές ότι οι προβλέψεις είναι αναγκαίες.

1.5 ΤΥΠΟΙ ΠΡΟΒΛΕΨΗΣ

Οι διαδικασίες πρόβλεψης μπορεί αρχικά να διαχωριστούν σε *μακροπρόθεσμες,* ή *βραχυπρόθεσμες*. Οι μακροπρόθεσμες προβλέψεις, είναι αναγκαίες για να χαράξουν την γενική πορεία και στρατηγική της επιχείρησης ή του κράτους. Συνεπώς, σ΄ αυτές δίνουν ιδιαίτερη έμφαση τα ανώτατα διοικητικά στελέχη. Οι βραχυπρόθεσμες προβλέψεις χρειάζονται για τη χάραξη άμεσων στρατηγικών και χρησιμοποιούνται από τα μεσαία στελέχη και τους μάνατζερ, έτσι ώστε να ανταποκριθούν στις ανάγκες του άμεσου μέλλοντος.

Οι προβλέψεις μπορούν επίσης να ταξινομηθούν σε *μίκρο* και *μάκρο*. Για παράδειγμα, ένας διευθυντής μπορεί να ενδιαφέρεται να προβλέψει τον αριθμό των εργαζομένων, που απαιτούνται για τους επόμενους μήνες στην επιχείρηση του (μίκρο πρόβλεψη), ενώ η

κυβέρνηση θέλει να προβλέψει το συνολικό αριθμό των ατόμων, που θα απασχοληθούν σε ολόκληρη τη χώρα (μάκρο πρόβλεψη).

Οι διαδικασίες πρόβλεψης μπορούν επίσης να ταξινομηθούν ανάλογα με το αν είναι περισσότερο *ποσοτικές,* ή *ποιοτικές*. Από την μια μεριά, μία καθαρά ποιοτική τεχνική είναι μία, που δεν απαιτεί καμία εμφανή διαχείριση των δεδομένων. Μόνο η κρίση του αναλυτή απαιτείται. Από την άλλη μεριά, οι καθαρά ποσοτικές τεχνικές δεν απαιτούν ανθρώπινη κρίση. Είναι μηχανικές μαθηματικές διαδικασίες που καταλήγουν σε ποσοτικά αποτελέσματα. Ορισμένες ποσοτικές διαδικασίες, φυσικά, απαιτούν μια πολύ πιο εκλεπτυσμένη διαχείριση των στοιχείων από ότι άλλες.

1.6 ΜΑΚΡΟΟΙΚΟΝΟΜΙΚΕΣ ΠΡΟΒΛΕΨΕΙΣ

Συνήθως όταν μιλάμε για πρόβλεψη σκεφτόμαστε σημαντικές μεταβλητές σε επίπεδο επιχείρησης, που αφορούν πωλήσεις, χρηματοροές, αποθήκες, εργατοώρες προσωπικού και όλα αυτά σε μηνιαία, τριμηνιαία βάση κλπ.

Σε αντίθεση υπάρχει ένα πολύ μεγάλο ενδιαφέρον για πρόβλεψη σημαντικών μεταβλητών ολόκληρης της οικονομίας μιας χώρας. Παραδείγματα είναι το ΑΕΠ, ή ολόκληρων κλάδων μέσω διαφόρων δεικτών, όπως το λιανικό εμπόριο, βιομηχανικός δείκτης, ποσοστό ανεργίας κλπ.

Γιαυτό ακριβώς το λόγο υπάρχει επιτακτική ανάγκη για την βελτιστοποίηση των μεθόδων πρόβλεψης. Μία από τις μεγάλες δυσκολίες επίτευξης του σκοπού αυτού είναι η επίδραση, που μπορεί να έχουν η τιμή του πετρελαίου, ή αλλαγές σε άλλες οικονομίες, που μπορεί να επηρεαστεί η υπό μελέτη οικονομία λόγω της παγκοσμιοποίησης.

Οπότε τίθεται το ερώτημα αν πρέπει να λάβουμε υπόψη μας την αλλαγή των προβλέψεων, που προέρχονται από μαθηματικό τρόπο με τη κρίση του προβλέποντος.

Θεωρητική και πρακτική δουλειά γίνεται ακόμη θεωρώντας τη σημασία της πρόβλεψης στη οικονομική πολιτική και τη προσομοίωση στην προς μελέτη χώρα, ή και σε άλλες.

1.7 ΕΠΙΛΟΓΗ ΜΙΑΣ ΜΕΘΟΔΟΥ ΠΡΟΒΛΕΨΗΣ

Το βασικότερο κριτήριο για την επιλογή μιας μεθόδου πρόβλεψης είναι ότι τα αποτελέσματα πρέπει να διευκολύνουν τη διαδικασία λήψης αποφάσεων των διοικήσεων των εταιρειών. Σπάνια είναι χρήσιμη μια μέθοδος για όλες τις περιπτώσεις. Διαφορετικά προϊόντα, στόχοι και περιορισμοί πρέπει να λαμβάνονται υπόψη κατά την επιλογή μιας μεθόδου.

Με τη διαθεσιμότητα του λογισμικού πρόβλεψης το καλύτερο είναι να σκεφτούμε μεθόδους πρόβλεψης, που μπορούν να εφαρμοστούν ταυτόχρονα. Διάφορες μέθοδοι μπορούν να χρησιμοποιηθούν σε μια δεδομένη κατάσταση. Η μέθοδος που καταλήγει στις πιο ακριβείς προβλέψεις σε μία περίπτωση μπορεί να μην είναι η καλύτερη μέθοδος σε μια άλλη περίπτωση. Ωστόσο, η μέθοδος που θα επιλεγεί θα πρέπει να δώσει μια πρόβλεψη που να είναι ακριβής, έγκυρη και κατανοητή από τη διοίκηση της επιχείρησης, έτσι ώστε η πρόβλεψη που θα γίνει να βοηθήσει να ληφθούν οι καλύτερες αποφάσεις για τον μέλλον της.

Στο κεφάλαιο 3 θα δούμε την τεχνική επιλογής μεθόδου πρόβλεψης.

1.8 Η ΔΙΑΔΙΚΑΣΙΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ

Μία σωστή διαδικασία πρόβλεψης γίνεται εμφανής με την απάντηση μερικών βασικών ερωτημάτων:

- Γιατί χρειάζεται η πρόβλεψη;
- Ποιος θα χρησιμοποιήσει την πρόβλεψη και ποιες είναι οι απαιτήσεις του;
- Τι επίπεδο λεπτομέρειας απαιτείται και ποιο προβλέπεται ότι είναι το κατάλληλο χρονικό διάστημα για να ολοκληρωθεί η διαδικασία;
- Τι δεδομένα είναι διαθέσιμα και τα δεδομένα αυτά είναι επαρκή για να γίνουν οι απαραίτητες προβλέψεις;
- Ποιο είναι το κόστος της πρόβλεψης;
- Πόσο ακριβής πιστεύεται, ότι θα είναι η πρόβλεψη;
- Η πρόβλεψη θα γίνει εγκαίρως για να βοηθήσει τη διαδικασία λήψης αποφάσεων;
- Ο αναλυτής έχει καταλάβει ακριβώς, πώς η πρόβλεψη θα χρησιμοποιηθεί στην επιχείρηση;
- Υπάρχει μια διαδικασία διαθέσιμη για την αξιολόγηση της πρόβλεψης και για να προσαρμοστεί η διαδικασία πρόβλεψης αναλόγως;

1.9 ΤΑ ΒΗΜΑΤΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ

Το γεγονός ότι οι τεχνικές πρόβλεψης βασίζονται σε δεδομένα που προέρχονται από ιστορικά γεγονότα οδηγεί στον προσδιορισμό των παρακάτω πέντε βημάτων:

1. Διατύπωση προβλήματος και συλλογή δεδομένων.

2. Διαχείριση δεδομένων και τελική επιλογή.
3. Κατασκευή υποδείγματος και αξιολόγηση του.
4. Εφαρμογή υποδείγματος (η πραγματική πρόβλεψη).
5. Αξιολόγηση πρόβλεψης.

Στο **πρώτο βήμα**, η διατύπωση του προβλήματος και η συλλογή των δεδομένων αντιμετωπίζονται ως ένα ενιαίο βήμα, διότι είναι αλληλένδετα. Το πρόβλημα καθορίζει τα κατάλληλα δεδομένα. Αν εξετάζεται μια ποσοτική μεθοδολογία πρόβλεψης, θα πρέπει να είναι διαθέσιμα και ακριβή τα κατάλληλα δεδομένα. Εάν τα κατάλληλα δεδομένα δεν είναι διαθέσιμα, τότε το πρόβλημα μπορεί να χρειαστεί να επαναπροσδιοριστεί. Επίσης μπορεί να εμφανιστούν προβλήματα συλλογής και έλεγχου ποιότητας, κάθε φορά που καθίσταται αναγκαία η λήψη κατάλληλων δεδομένων για να γίνει μια πρόβλεψη.

Στο **δεύτερο βήμα**, η διαχείριση των στοιχείων και η τελική επιλογή τους είναι απαραίτητη. Είναι δυνατόν να έχουμε, είτε πάρα πολλά δεδομένα, είτε πολύ λίγα στη διαδικασία πρόβλεψης. Ορισμένα δεδομένα μπορεί να μην είναι σχετικά με το πρόβλημα. Άλλα δεδομένα μπορεί να είναι κατάλληλα, αλλά μόνο σε συγκεκριμένες ιστορικές περιόδους. Συνήθως, απαιτείται κάποια προσπάθεια για να μετατραπούν τα δεδομένα στην κατάλληλη μορφή για τη χρήση συγκεκριμένων διαδικασιών πρόβλεψης.

Το **τρίτο βήμα**, η κατασκευή του μοντέλου και η αξιολόγηση του, περιλαμβάνει την τοποθέτηση του συλλεχθέντων στοιχείων σε ένα μοντέλο – υπόδειγμα πρόβλεψης, που είναι κατάλληλο για την ελαχιστοποίηση των σφαλμάτων προσαρμογής με τελικό στόχο την πρόβλεψη. Όσο πιο απλό είναι το μοντέλο, τόσο το καλύτερο για την αποδοχή της διαδικασίας πρόβλεψης από τους μάνατζερ. Προφανώς, η

ανθρώπινη κρίση περιλαμβάνεται σε αυτό το βήμα.

Το **τέταρτο βήμα**, η εφαρμογή της μεθόδου, είναι η δημιουργία ενός πραγματικού μοντέλου με προβλέψεις, παράγουμε προβλέψεις, μόλις τα κατάλληλα δεδομένα έχουν συλλεχθεί και έχει γίνει η τελική επιλογή. Τα δεδομένα από τις πρόσφατες ιστορικές περιόδους χρησιμοποιούνται αργότερα για να γίνει ο έλεγχος για την ορθότητα της διαδικασίας.

Στο **πέμπτο βήμα**, η αξιολόγηση των προβλέψεων, περιλαμβάνεται η σύγκριση των τιμών με πραγματικές ιστορικές τιμές. Οι προβλέψεις αυτές συγκρίνονται στη συνέχεια με τις γνωστές ιστορικές τιμές και εξετάζονται τα τυχόν σφάλματα που προκύπτουν. Η εξέταση των σφαλμάτων συχνά οδηγεί τον αναλυτή στην τροποποίηση του μοντέλου πρόβλεψης. Ειδικές μέθοδοι μέτρησης των σφαλμάτων αναλύονται στο τέλος του κεφαλαίου 3.

1.10 ΛΟΓΙΣΜΙΚΟ ΠΡΟΒΛΕΨΗΣ

Σήμερα υπάρχει ένας μεγάλος αριθμός λογισμικών πακέτων, που έχουν ειδικά σχεδιαστεί για να παρέχουν στον χρήστη διάφορες μεθόδους πρόβλεψης. Δύο τύποι πακέτων που έχουν πρωταρχική σημασία για τους αναλυτές είναι:

- τα γενικά στατιστικά πακέτα που περιλαμβάνουν την ανάλυση παλινδρόμησης, την ανάλυση χρονολογικών σειρών και άλλων τεχνικών που χρησιμοποιούνται συχνά από τους αναλυτές και
- τα πακέτα πρόβλεψης που έχουν ειδικά σχεδιαστεί για εφαρμογές πρόβλεψης μέσα από ERP συστήματα (SAP).

Παραδείγματα αυτόνομων λογισμικών πακέτων με εργαλεία

πρόβλεψης είναι το **R**, το **Minitab**, το **SAS**, το **Eviews** και το **SPSS**. Υπάρχουν πολλά συμπληρωματικά προγράμματα, τα οποία παρέχουν εργαλεία πρόβλεψης σε περιβάλλον υπολογιστικού φύλλου. Για παράδειγμα, η ανάλυση **ToolPak**, που είναι συμπληρωματική για το Microsoft Excel παρέχει ανάλυση παλινδρόμησης. Σήμερα, υπάρχουν πολλά ολοκληρωμένα συμπληρωματικά πακέτα που έχουν πολλές δυνατότητες πρόβλεψης.

Πολλές από τις τεχνικές που θα δούμε απεικονίζονται με το **Minitab 15 – 16 – 17**. Το Minitab επιλέχθηκε για ευκολία χρήσης.

1.11 ΤΙ ΜΠΟΡΕΙ ΝΑ ΠΡΟΒΛΕΦΘΕΙ

Ένα ερώτημα αναμφισβήτητα ενδιαφέρον και με μεγάλες προεκτάσεις. Τελικά υπάρχουν τομείς και δεδομένα, όπου παράγονται πάντα αξιόπιστες προβλέψεις; Με λίγη σκέψη καταλήγουμε ότι αυτό είναι κάτι πολύ δύσκολο να επιτευχθεί. Ας αναφέρουμε περιληπτικά κάποιους τομείς της καθημερινότητας μας, οι οποίοι συνήθως αναλύονται με προβλέψεις:

- Πρόγνωση μετεωρολογικών συνθηκών.
- Πρόβλεψη κινδύνου πυρκαγιάς.
- Προγνωστικά αγώνων ποδοσφαίρου, μπάσκετ, διεθνών διοργανώσεων, ιππόδρομο.
- Ολυμπιακών αγώνων.
- Πρόβλεψη σεισμών.
- “Αστρολογικές προβλέψεις”.
- Πρόβλεψη πορείας οικονομικών μεγεθών.
- Πρόβλεψη πτώσης / ανόδου τιμής πετρελαίου.

- Πρόβλεψη συνολικής παραγωγής ενέργειας σε μία εγκατάσταση.
- Πρόβλεψη της κυκλοφορίας στην εθνική οδό, ή σε άλλες τοπικές οδούς.
- Πρόβλεψη απόδοσης εργαζομένων σε μια εταιρεία ή μαθητών σε ένα σχολείο.
- Πρόβλεψη ζήτησης ενός συγκεκριμένου είδους κατανάλωσης και εισαγωγής ενός καινοτόμου προϊόντος στην αγορά.

Ο κωμικός Josh Billing σε ένα θέμα σχετικό με τις προβλέψεις είχε πει: «Μην προβλέπεις. Αν η πρόβλεψή σου αποδειχθεί λάθος, κανείς δεν θα την ξεχάσει. Αν αποδειχθεί σωστή, κανείς δεν θα την θυμάται».

Μπορούμε όμως να ακολουθήσουμε αυτή τη λογική; Η αλήθεια είναι ότι οι προβλέψεις είναι αναγκαίες στη ζωή μας. Μας βοηθούν να είμαστε κατάλληλα προετοιμασμένοι για κάποια επικείμενα γεγονότα και να παίρνουμε σωστότερες αποφάσεις, έστω και αν πάντα υπήρχαν και θα υπάρχουν οι μεγάλες αποκλίσεις, που θα ανατρέπουν όλη την προετοιμασία μας.

1.12 ΠΑΡΑΔΕΙΓΜΑ ΠΡΟΒΛΕΨΗΣ

Αυτό το κεφάλαιο δίνει έμφαση στο γεγονός, ότι η πρόβλεψη απαιτεί τον συνδυασμό της ανθρώπινης κρίσης και της μαθηματικής διαχείρισης των δεδομένων, που συλλέγονται. Το παρακάτω παράδειγμα περιγράφει τον τρόπο σκέψης, που προηγείται συχνά μιας διαδικασίας πρόβλεψης σε μια πραγματική επιχείρηση.

Μία επιχείρηση Χ διαφημίζεται με διάφορους τρόπους και μέχρι πρόσφατα δεν είχε μελετήσει την επίδραση, που έχει η διαφήμιση στις πωλήσεις της, αν και ορισμένα στοιχεία είχαν συλλεχθεί και αποθηκευτεί

για 3 χρόνια. Τα στελέχη της αποφάσισαν να ξεκινήσουν να παρακολουθούν τα αποτελέσματα, που είχαν οι διαφημίσεις στον όγκο των πωλήσεων κάθε μήνα. Ήλπιζαν ότι μετά από αρκετούς μήνες τα συλλεγόμενα δεδομένα θα μπορούσαν να εξεταστούν, για να βοηθήσουν στις μελλοντικές διαφημιστικές δαπάνες. Το λογιστήριο άρχισε να επεκτείνει τα ιστορικά αρχεία της επιχείρησης με το να καταγράφει τον όγκο των πωλήσεων κάθε μήνα, μαζί με τα χρήματα που δαπανούσε για την διαφήμιση. Συγκεκριμένα αποθήκευσαν τα ακόλουθα στοιχεία:

- Πωλήσεις σε κύκλο εργασιών.
- Διαφήμιση σε περιοδικά.
- Διαφήμιση στη τηλεόραση.
- Διαφήμιση σε περιοδικά με καθυστέρηση ενός μηνός.
- Διαφήμιση σε περιοδικά με καθυστέρηση δύο μηνών.
- Διαφήμιση στη τηλεόραση με καθυστέρηση ενός μηνός.
- Διαφήμιση στη τηλεόραση με καθυστέρηση δύο μηνών.

Τα στελέχη της πίστευαν, ότι οι πωλήσεις επηρεαζόταν από τον μήνα (υπήρχε δηλαδή εποχικότητα) και αναρωτιόντουσαν για οποιαδήποτε αλλαγή στον όγκο των πωλήσεων. Τέλος, τα στελέχη της πίστευαν, ότι το χρηματικό ποσό που δαπανούσαν για την διαφήμιση της επιχείρησης, μπορεί να έχει επιπτώσεις στη διαφήμιση των ανταγωνιστών της τον επόμενο μήνα. Μετά από μερικούς μήνες συλλογής στοιχείων και ανάλυσης των προηγούμενων αρχείων, το λογιστήριο ολοκλήρωσε μια σειρά δεδομένων για 48 μήνες.

Οι μάνατζερ της επιχείρησης και ειδικά ο πρόεδρος της εταιρείας, θέλουν τώρα να μάθουν ότι μπορούν, από τα στοιχεία που έχουν συλλέξει. Εκτός από το να μάθουν για τις επιπτώσεις της διαφήμισης στον όγκο των πωλήσεων και των ανταγωνιστών τους, ο πρόεδρος

αναρωτιέται για οποιαδήποτε επίπτωση εποχικών δεδομένων στις πωλήσεις. Στο τέλος της συνάντησης ο διευθυντής της παραγωγής έκανε την ακόλουθη δήλωση: «Νομίζω ότι χάνουμε πολύ χρόνο με την συλλογή των δεδομένων και χαζεύοντας μπροστά στους υπολογιστές. Το μόνο που έχετε να κάνετε είναι να μιλήσετε με τους εργαζόμενους, με τους μάνατζερ των καταστημάτων και τους πελάτες για να καταλάβετε τι συμβαίνει. Κάποιοι από εσάς πρέπει να κλείσετε τους υπολογιστές σας, να σηκωθείτε από τα γραφεία σας και να συνομιλήσετε με μερικούς πραγματικούς ανθρώπους».

1.13 ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ ΠΟΥ ΘΑ ΑΝΑΠΤΥΧΘΟΥΝ

TABLE 11-1 Forecasting Models

Method	*Description*	*Applications*	*Computer*		*Chapter*
			Cost	*Necessary?*	
Causal Forecasting Models					
Regression analysis	Explanatory forecasting; assumes a cause-and-effect relationship between the input to a system and its output	Short- and medium-range forecasting of existing products and services; marketing strategies, production, personnel hiring, and facility planning	Low to medium	Usually	6
Multiple regression	Explanatory forecasting; assumes a cause-and-effect relationship between more than one input to a system and its output	Same as above	Low to medium	Yes	7
Time Series Forecasting Models					
Decomposition method	Explanatory forecasting; assumes a cause-and-effect relationship between time and the output of a system; the system is decomposed into its components	Medium-range forecasting for new plant and equipment planning, financing, new-product development, and new methods of assembly; short-range forecasting for personnel, advertising, inventory, financing, and production planning	Low to medium	Yes	5
Moving averages	Used to eliminate randomness in a time series; forecast is based on projection from time series data smoothed by a moving average	Short-range forecasting for operations such as inventory, scheduling, control, pricing, and timing of special promotions; used to compute both the seasonal and the cyclical components for the short-term decomposition method	Low	No	4 and 5

TABLE 11-1 *(Continued)*

Method	*Description*	*Applications*	*Computer*		*Chapter*
			Cost	*Necessary?*	
Exponential smoothing	Similar to moving averages, but values are weighted exponentially, giving more weight to the most recent data	Short-range forecasting for operations such as inventory, scheduling, control, pricing, and timing of special promotions	Low	Yes	4
Autoregressive models	Employed with economic variables to account for relationships between adjacent observations in a time series	Short- and medium-range forecasting for economic data—including price, inventory, production, stocks, and sales—ordered in a time series	Medium	Yes	8 and 9
Box-Jenkins techniques	Does not assume any particular pattern in the historical data of the series to be forecast; uses an iterative approach of identifying and fitting a possibly useful model from a general class of models	Same as above	High	Yes	9
Neural networks	Uses a sophisticated computer program to assimilate relevant data and recognize patterns by "learning" as humans do	Increasing use in a variety of forecasting applications; is in the development phase	High	Yes	10

forecasting process have been described in this book, with an emphasis on learning the techniques commonly used to generate the actual forecasts. The key operational steps in the forecasting process are summarized in Figure 11-1.

The collection and examination of appropriate historical data were described earlier in this book (Chapter 3), as were considerations in selecting a forecasting technique or model. As suggested by Figure 11-1, the next step is usually to forecast several historical periods where the actual values are known. The resulting errors can be summarized in several ways, as discussed in Chapter 3, and this process is continued until a technique with a sufficient cost-benefit ratio is found. The model is then used to forecast future periods, and the results are incorporated into the firm's decision-making process.

1.14 ΣΥΜΠΕΡΑΣΜΑ

Σκοπός μιας πρόβλεψης είναι να μειώσει την αβεβαιότητα των αποφάσεων, που πρέπει να ληφθούν. Η διαδικασία πρόβλεψης πρέπει να προσαρμόζεται σε δύο βασικούς κανόνες:

- Οι προβλέψεις πρέπει να είναι τεχνικώς ορθές και να ανταποκρίνονται στις ανάγκες της επιχείρισης.
- Η διαδικασία πρόβλεψης και τα αποτελέσματά της πρέπει να παρουσιάζονται αποτελεσματικά στην διοίκηση, έτσι ώστε οι προβλέψεις να είναι χρήσιμες στην διαδικασία λήψης αποφάσεων.

Οι αναλυτές δίνουν μεγαλύτερη σημασία στον πρώτο κανόνα και λιγότερη στο δεύτερο. Αν υπάρχουν καλά προετοιμασμένες και οικονομικά αποδοτικές προβλέψεις που θα ωφελήσουν την επιχείρηση, τότε αυτοί που έχουν την εξουσία λήψης αποφάσεων πρέπει να τις χρησιμοποιήσουν. Η ανάγκη να είναι σε θέση κάποιος να παρουσιάσει αποτελεσματικά τις προβλέψεις στην διοίκηση είναι εξίσου σημαντική, με την ανάγκη να είναι σε θέση αυτός να αναπτύξει τις προβλέψεις.

ΚΕΦΑΛΑΙΟ 2 (ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ)

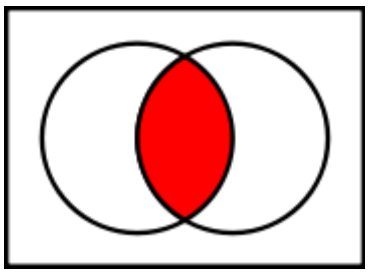
Σ' αυτό το κεφάλαιο θα κάνουμε μια ανασκόπηση στις βασικές στατιστικές έννοιες, που θα χρησιμεύσουν στη συνέχεια. Στην πρόβλεψη ασχολούμαστε συνήθως με ιστορικά δεδομένα, έτσι ώστε να προβλέψουμε το αβέβαιο μέλλον. Για το λόγο αυτό θα εξεταστεί ένα δείγμα δεδομένων, θα γίνει η διαχείριση του με κάποιο τρόπο και θα χρησιμοποιηθούν τα αποτελέσματα του για να γίνουν προβλέψεις. Η αναγκαιότητα συλλογής και επεξεργασίας πληροφοριών προκύπτει από την ανάγκη, να γνωρίσουμε μεγέθη, των οποίων τα χαρακτηριστικά παρουσιάζουν μεταβλητότητα.

2.1 ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΒΑΣΙΚΕΣ ΣΤΑΤΙΣΤΙΚΕΣ ΕΝΝΟΙΕΣ

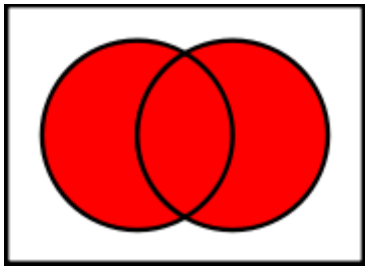
Έστω δύο σύνολα Α και Β. Στα παρακάτω διαγράμματα μπορούμε να παραστήσουμε γραφικά κάποιες πράξεις επί των συνόλων με το Α να είναι αριστερά και το Β δεξιά.

Τα σύνολα αυτά μπορεί να έχουν ως στοιχεία οτιδήποτε μπορούμε να μετρήσουμε και μας ενδιαφέρει, για παράδειγμα μπορεί να είναι εσφαλμένα προϊόντα παραγωγής ανά κατηγορία σφάλματος.

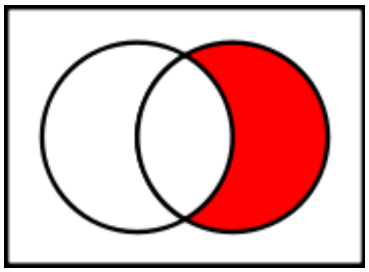
Διαγράμματα Venn



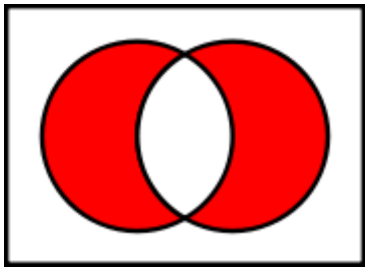
Τομή Συνόλων $A \cap B$ (AND)



Ένωση συνόλων $A \cup B$ *(OR)*



Διαφορά Συνόλων $= A^C \cap B = B \setminus A \ \ (B \ \chi\omega\rho\acute{\iota}\varsigma \ A)$



Συμμετρική Διαφορά Συνόλων $= (A \Delta B) = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B)$

(OR, OR)

Δειγματοχώρος ονομάζεται το σύνολο όλων των δυνατών αποτελεσμάτων κατά την εκτέλεση ενός πειράματος τύχης.

Π.χ. κατά τη ρίψη ενός ζαριού ο δειγματοχώρος είναι το σύνολο {1,2,3,4,5,6}, κατά τη ρίψη ενός νομίσματος ο δειγματοχώρος είναι το σύνολο {Κ,Γ}.

Πεπερασμένος δειγματοχώρος: αν έχει πεπερασμένο πλήθος στοιχείων.

Άπειρος αριθμήσιμος δειγματοχώρος (διακριτός ή διακεκριμένος): Αντιστοίχηση των στοιχείων του με το σύνολο των

φυσικών αριθμών.

Άπειρος μη αριθμήσιμος δειγματοχώρος (συνεχής): Αντιστοίχηση των στοιχείων του ένα προς ένα με τα σημεία ενός διαστήματος του άξονα x της ευθείας των πραγματικών αριθμών.

Γεγονός ή ενδεχόμενο: Ένα υποσύνολο δυνατών αποτελεσμάτων του δειγματοχώρου. Εάν τα σύνολα που αντιστοιχούν στα γεγονότα Α και Β είναι **ξένα** δηλαδή $A \cap B = \varnothing$, τότε τα γεγονότα αυτά λέγονται **ασυμβίβαστα, ή αμοιβαίως αποκλειόμενα**. Τα ενδεχόμενα αυτά δεν μπορούν να συμβούν ταυτόχρονα.

Ο ίδιος ο δειγματοχώρος καλείται **βέβαιο γεγονός,** ενώ το κενό υποσύνολο **αδύνατο γεγονός**.

Η έννοια της πιθανότητας

Σε ένα τυχαίο πείραμα υπάρχει πάντα η αβεβαιότητα, για το αν θα συμβεί ένα ορισμένο γεγονός ή όχι. Σαν μέτρο της πιθανότητας να συμβεί το γεγονός λαμβάνουμε έναν αριθμό μεταξύ 0 και 1. Αν είμαστε βέβαιοι, ότι το γεγονός θα συμβεί, τότε λέμε ότι η πιθανότητα είναι 1 ή 100%. Εάν είμαστε βέβαιοι ότι το γεγονός δεν θα συμβεί, τότε λέμε ότι η πιθανότητα είναι 0.

Υπάρχουν δύο μέθοδοι, με τις οποίες εκτιμάμε (ορίζουμε αυθαίρετα) τη πιθανότητα ενός γεγονότος:

1. **Κλασσική μέθοδος (ή εκ των προτέρων, ή μαθηματική (P)):** Εάν το πείραμα μπορεί να δώσει **n** διαφορετικά αποτελέσματα με τα **h** να είναι ευνοϊκά (δηλαδή αν συμβεί ένα από αυτά να έχουμε γεγονός), τότε η πιθανότητα του γεγονότος είναι **h/n**. Π.χ. Ας εκτιμήσουμε τη πιθανότητα να έλθει κεφάλι ρίχνοντας τίμια ένα κανονικό νόμισμα. Επειδή υπάρχουν δύο εξίσου πιθανά αποτελέσματα δηλαδή κεφάλι ή γράμματα (αγνοώντας την

κατάσταση του νομίσματος να κάτσει όρθιο), τότε η πιθανότητα να έλθει το ευνοϊκό αποτέλεσμα, δηλαδή κεφάλι είναι ½ ή 50%.

2. **Μέθοδος της σχετικής συχνότητας (ή εκ των υστέρων, ή στατιστική (f)):** Εάν μετά από **n** επαναλήψεις του πειράματος, με **n** πολύ μεγάλο, ένα γεγονός παρατηρείται ότι συνέβη **h** φορές, τότε η πιθανότητα του γεγονότος αυτού είναι **h/n**. Π.χ. Σε 1000 ρίψεις ενός νομίσματος ήλθε 532 φορές κεφάλι. Άρα η πιθανότητα να έλθει κεφάλι είναι 532/1000 = 0.532 ή 53.2%. Στην ουσία $P(\Gamma\varepsilon\gamma ονός) \to \lim_{n\to\infty} f(\Gamma\varepsilon\gamma ονός)$.

Αξιώματα Πιθανότητας (Kolmogoroff)

Αξ. 1: Για κάθε γεγονός Α ενός συνόλου είναι $P(A)\geq 0$.

Αξ. 2: Για το βέβαιο γεγονός δ είναι $P(\delta)=1$. Αν δηλαδή το δ είναι όλος ο δειγματοχώρος, τότε η πιθανότητα είναι 1.

Αξ. 3: Προσθετική Ιδιότητα: Αν τα γεγονότα Α και Β είναι ασυμβίβαστα, τότε $P(A\cup B)=P(A)+P(B)$. Αυτό γενικεύεται και για την ένωση περισσότερων γεγονότων.

Θεωρήματα

Αν $A\subset B$ τότε: $P(A)\leq P(B)$ και

$$P(A-B)=P(A)-P(B).$$

Αν A^C είναι το συμπλήρωμα του Α, τότε: $P(A^C)=1-P(A)$.

Αν Α και Β είναι δύο οποιαδήποτε γεγονότα, τότε:

$$P(A\cup B)=P(A)+P(B)-P(A\cap B).$$

Πιθανότητα υπό Συνθήκη

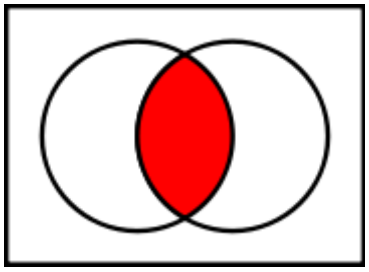
Θεωρούμε δύο γεγονότα Α και Β με P(A)>0. Έστω P(B|A) είναι η πιθανότητα να συμβεί το Β αφού έχει συμβεί το Α. Επειδή το Α έχει

πραγματοποιηθεί, αυτό γίνεται ο νέος δειγματοχώρος και όχι ο δ. Οπότε η πιθανότητα P(B|A) ορίζεται ως:

$$P(B|A)=\frac{P(A\cap B)}{P(A)}$$

Δηλαδή ισούται με τη σχέση της πιθανότητας να συμβεί ένα μέρος του Α με την πιθανότητα να συμβεί όλο το Α.

Μερικοί συγγραφείς όπως ο De Finetti, προτιμούν να εισάγουν την πιθανότητα υπό συνθήκη ως ένα αξίωμα πιθανότητας με το $A\cap B$ να είναι μέρος του Α: $P(A\cap B)=P(A)P(B|A)$



Αυτό σημαίνει ότι η πιθανότητα να συμβούν και τα δύο γεγονότα ισούται με το γινόμενο της πιθανότητας να συμβεί το Α με την πιθανότητα να συμβεί το Β ενώ προηγουμένως συνέβη το Α.

Στην περίπτωση που P(B|A) = P(B), δηλαδή το γεγονός Β είναι ανεξάρτητο του γεγονότος Α, τότε $P(A\cap B)=P(A)P(B)$. Προφανώς για αμοιβαία αποκλειόμενα γεγονότα $P(A\cap B)=0$.

Παραδείγματα

Ποια είναι η πιθανότητα στο ρίξιμο ενός ζαριού να έχουμε 1 ή μονό αριθμό;

Βάσει του ορισμού P(x=1 ή x=μονός)=3/6=0,5.

Βάσει της προσθετικής ιδιότητας P(x=1 ή x=μονός) = P(x=1) + P(x=μονός) = 1/6 + 3/6 = 4/6 = 2/3!!!

Αυτό όμως είναι λάθος γιατί θα πρέπει να αφαιρέσουμε την τομή

τους, διότι τα γεγονότα δεν είναι ξένα μεταξύ τους. Άρα:

$$P(x=1\,\acute{\eta}\,x=\mu ονός)=P(x=1)+P(x=\mu ονός)-P(x=1\cap x=\mu ονός)$$
$$\frac{1}{6}+\frac{3}{6}-\frac{1}{6}=\frac{3}{6}=0{,}5.$$

Μία μηχανή, η οποία παράγει μαζικά μηχανικά μέρη, παρουσιάζει δύο ήδη ελαττωμάτων, Α και Β, τα οποία είναι ανεξάρτητα αλλά όχι αμοιβαία αποκλειόμενα. Το ποσοστό εμφάνισης του Α είναι 7% και του Β είναι 9%. Να βρεθούν:

α) Το ποσοστό εμφάνισης οποιουδήποτε ελαττώματος κατά την παραγωγή.

β) Το ποσοστό των μηχανικών μερών χωρίς ελάττωμα.

γ) Το ποσοστό εμφάνισης μόνο του ελαττώματος Α.

δ) Το ποσοστό εμφάνισης μόνο του ελαττώματος Β.

ε) Το ποσοστό εμφάνισης και των δύο ελαττωμάτων Α και Β ταυτόχρονα.

α) $P(A\cup B)=P(A)+P(B)-P(A\cap B)=7\%+9\%$
$-7\%*9\%=15{,}37\%.$

β) $P((A\cup B)^C)=1-P(A\cup B)=1-15{,}37\%=84{,}63\%.$

γ) Επειδή στο αρχικό ποσοστό των ελαττωματικών μερών ανήκει και η τομή τους, για να βρούμε το καθαρό ποσοστό των ελαττωματικών μερών Α πρέπει να αφαιρέσουμε από αυτό την τομή των Α και Β, όπως φαίνεται παρακάτω. Με άλλα λόγια είναι η πιθανότητα της διαφοράς των δύο. Οπότε το καθαρό ποσοστό θα είναι 7%-7%*9%=6,37%.

$$P(A\setminus B)=P(A\cap B^C)=P(B^C)P(A)=(1-P(B))P(A)$$
$$P(A)-P(A)P(B)=P(A)-P(A\cap B)$$

δ) Ίδιο όπως στο προηγούμενο για το Β: 9%-7%*9%=8,37%.

ε) Το ποσοστό θα είναι η τομή τους: 7%*9%=0,63%.

Τυχαία μεταβλητή και κατανομή πιθανότητας

Ας υποθέσουμε ότι σε κάθε σημείο του δειγματοχώρου αντιστοιχούμε έναν αριθμό. Έχουμε έτσι μία συνάρτηση ορισμένη στο δειγματοχώρο. Η συνάρτηση αυτή καλείται **τυχαία μεταβλητή ή στοχαστική συνάρτηση** και συμβολίζεται με κεφαλαία γράμματα π.χ. **X**.

Διακριτή τυχαία μεταβλητή: Άπειρο αριθμήσιμο πλήθος τιμών.

Συνεχής τυχαία μεταβλητή: Άπειρο μη αριθμήσιμο πλήθος τιμών.

Για δύο νομίσματα έχουμε:

Σημείο Δειγματοχώρου	ΚΚ	ΚΓ	ΓΚ	ΓΓ
X	2	1	1	0

Διακριτές κατανομές πιθανότητας

Έστω X μία τυχαία διακριτή μεταβλητή με τιμές $x_{1,}x_{2,}x_{3,....}$ Ας υποθέσουμε, ότι οι πιθανότητες να πάρει η μεταβλητή τις τιμές αυτές είναι:

$$P(X=x_k)=f(x_k)$$

Γενικότερα ορίζουμε τη **συνάρτηση πιθανότητας, ή κατανομή πιθανότητας** $P(X=x)=f(x),$ έτσι ώστε:

$$1.\quad f(x)\geq 0 \quad 2.\quad \sum_x f(x)=1$$

Συναρτήσεις κατανομής για διακριτές τυχαίες μεταβλητές

Η αθροιστική συνάρτηση κατανομής ή απλά συνάρτηση κατανομής για μια τυχαία διακριτή μεταβλητή X ορίζεται με τη σχέση:

$$P(X\leqslant x)=F(x),\qquad -\infty< x< \infty .$$

Η συνάρτηση κατανομής μπορεί να ληφθεί από τη συνάρτηση πιθανότητας ως:

$$F(x)=P(X\leqslant x)=\sum_{u\leq x} f(u)$$

Συνεχείς κατανομές πιθανότητας

Εάν η Χ είναι μία συνεχής τυχαία μεταβλητή, η πιθανότητα να πάρει η Χ μία ορισμένη τιμή είναι μηδέν. Συνεπώς δεν μπορούμε να ορίσουμε μία συνάρτηση πιθανότητας με τον ίδιο τρόπο, όπως στις διακριτές μεταβλητές. Στην προκείμενη περίπτωση έχει νόημα να βρούμε την πιθανότητα, να βρίσκεται η μεταβλητή ανάμεσα σε δύο διαφορετικές τιμές. Π.χ. Η πιθανότητα ένας ενήλικας να έχει ύψος ακριβώς 1.78 είναι μηδέν. Η πιθανότητα όμως να βρίσκεται ανάμεσα σε 1.65 και 1.84 είναι διάφορη του μηδενός και έχει νόημα να το χρησιμοποιήσουμε.

Από τις ιδιότητες για τις διακριτές μεταβλητές μπορούμε να ορίσουμε αξιωματικά μία συνάρτηση πιθανότητας, όπως παρακάτω:

$$1.\quad f(x)\geq 0 \quad 2.\quad \int_{-\infty}^{\infty} f(x)\,dx = 1$$

Η πιθανότητα να πάρει η Χ τιμές ανάμεσα στα α και β είναι:

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)\,dx$$

Η συνάρτηση f(x) καλείται συνάρτηση πυκνότητας πιθανότητας.

Συναρτήσεις κατανομής για συνεχείς τυχαίες μεταβλητές

Κατ' αναλογία με το αντίστοιχο για διακριτές τιμές ορίζουμε τη συνάρτηση κατανομής F(x) για μία συνεχή τυχαία μεταβλητή με τη σχέση:

$$F(x) = P(X \leqslant x) = P(-\infty < X \leqslant x) = \int_{-\infty}^{x} f(u)\,du$$

Μέση ή Αναμενόμενη Τιμή

Η μέση τιμή, ή μαθηματική ελπίδα, ή αναμενόμενη τιμή ορίζεται:

1) για μια τυχαία διακριτή μεταβλητή με τιμές $x_{1,} x_{2,} x_{3,....}$:

$$E(X)=x_1 P(X=x_1)+ x_2 P(X=x_2)+ ...+ x_n P(X=x_n)=\sum_{i=1}^{n} x_i P(X=x_i)$$

ή $E(X)=x_1 f(x_1)+ x_2 f(x_2)+ ...+ x_n f(x_n)=\sum_{i=1}^{n} x_i f(x_i)$

Στην ειδική περίπτωση που όλες οι πιθανότητες είναι ίδιες τότε:

$$E(X)=\frac{x_1+ x_2+ ...+ x_n}{n}=\frac{\sum_{i=1}^{n} x_i}{n}$$

καλείται και αριθμητικός μέσος όρος, ή μέσος όρος.

2) Για μία συνεχή τυχαία μεταβλητή η αναμενόμενη τιμή ορίζεται ως:

$$E(x)=\int_{-\infty}^{\infty} x f(x)dx$$

Η αναμενόμενη τιμή συμβολίζεται με μ.

Διασπορά

Η διασπορά, ή διακύμανση, ή μεταβλητότητα ορίζεται από τη σχέση:

$$Var(x)=\sigma^2=E[(X-\mu)^2]$$

και μας δίνει το ποσό της μεταβλητότητας των τιμών της μεταβλητής γύρο από το μέσο της.

Τυπική Απόκλιση

Η θετική τετραγωνική ρίζα της διασποράς ονομάζεται τυπική απόκλιση:

$$\sigma=\sqrt{Var(x)}=\sqrt{E[(X-\mu)^2]}$$

1) Για μία διακριτή τυχαία μεταβλητή:

$$\sigma^2 = E[(X-\mu)^2] = \sum_{i=1}^{n} (x_i-\mu)^2 f(x_i)$$

στην ειδική περίπτωση που όλες οι πιθανότητες είναι ίσες:

$$\sigma^2 = E[(X-\mu)^2] = \frac{\sum_{i=1}^{n} (x_i-\mu)^2}{n}$$

2) Για μία συνεχή τυχαία μεταβλητή ισχύει:

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

Ισχύουν επίσης οι σχέσεις:

$$\sigma^2 = E[(X-\mu)^2] = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = ¿$$
$$E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$$

$$Var(cX) = c^2 Var(X)$$

Πολλές φορές χρησιμοποιούμε την τυποποιημένη μεταβλητή $Z = \frac{X-\mu}{\sigma}$, η οποία έχει ως ιδιότητα, ότι $E(Z) = 0 \,και\, Var(Z) = 1$ και είναι επίσης και αδιάστατη.

Πολυδιάστατες διακριτές κατανομές πιθανότητας

Κατά αντιστοιχία αν έχουμε δύο ανεξάρτητες τυχαίες μεταβλητές ισχύουν για διακριτές και συνεχείς τιμές τα ακόλουθα:

Συνάρτηση πιθανότητας, ή κατανομή πιθανότητας

$$P(X = x, Y = y) = f(x, y)$$

$$1.\quad f(x, y) \geq 0 \quad 2.\quad \sum_{x}\sum_{y} f(x, y) = 1$$

Συναρτήσεις κατανομής για διακριτές τυχαίες μεταβλητές

Η αθροιστική συνάρτηση κατανομής, ή απλά συνάρτηση κατανομής για μια τυχαία μεταβλητή X ορίζεται με τη σχέση:

$$P(X \leqslant x, Y \leqslant y) = F(x, y) = \sum_{u \leq x} \sum_{v \leq y} f(u, v)$$

Πολυδιάστατες συνεχείς κατανομές πιθανότητας

Συναρτήσεις πιθανότητας για συνεχείς τυχαίες μεταβλητές

Από τις ιδιότητες για τις διακριτές μεταβλητές μπορούμε να ορίσουμε αξιωματικά μία συνάρτηση, όπως παρακάτω:

$$1. \quad f(x, y) \geq 0 \quad 2. \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1$$

Η πιθανότητα να πάρει η X τιμές ανάμεσα στα α και β και η Y μεταξύ γ και δ είναι:

$$P(\alpha < X < \beta, \gamma < Y < \delta) = \int_{x=\alpha}^{\beta} \int_{y=\gamma}^{\delta} f(x, y)\, dxdy$$

Η συνάρτηση f(x,y) καλείται συνάρτηση πυκνότητας πιθανότητας.

Συναρτήσεις κατανομής για συνεχείς τυχαίες μεταβλητές

Κατ' αναλογία με το αντίστοιχο για διακριτές τιμές ορίζουμε τη συνάρτηση κατανομής F(x,y) για δύο συνεχείς τυχαίες μεταβλητές X<= x και Y<= y, είναι:

$$P(X \leq x, Y \leq y) = F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)\, dxdy$$

Μέση τιμή

$$\mu_X = E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f(x, y)\, dxdy$$

$$\mu_Y = E(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f(x, y)\, dxdy$$

Στην περίπτωση διακριτών μεταβλητών αντικαθιστούμε τα ολοκληρώματα με αθροίσματα.

Διασπορά

$$\sigma_X^2 = E[(X-\mu_X)^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)^2 f(x,y)\,dxdy$$

$$\sigma_Y^2 = E[(Y-\mu_Y)^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-\mu_Y)^2 f(x,y)\,dxdy$$

Στην περίπτωση διακριτών μεταβλητών αντικαθιστούμε τα ολοκληρώματα με αθροίσματα.

Συνδιακύμανση

Για δύο τυχαίες μεταβλητές X και Y μία άλλη αξιοσημείωτη ποσότητα είναι η συνδιακύμανση, ή συνδιασπορά (Covariance), που ορίζεται για δύο διακριτές μεταβλητές από τη σχέση:

$$\sigma_{XY}^2 = E[(X-\mu_X)(Y-\mu_Y)] = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y)$$

και για δύο συνεχείς μεταβλητές από τη σχέση:

$$\sigma_{XY}^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y) f(x,y)\,dx\,dy$$

Η συνδιασπορά μας δίνει το μέσο ποσό της ταυτόχρονης μεταβλητότητας των δύο μεταβλητών από τον μέσο τους. Επίσης η συνδιασπορά είναι το μέτρο σχέσης των δύο μεταβλητών. Δηλαδή αν οι δύο μεταβλητές συσχετίζονται τότε όπως μεταβάλλεται η μία γύρο από τον μέσο της, με τον παρόμοιο τρόπο ή τον αντίθετο θα μεταβάλλεται και η άλλη γύρο από το μέσο της.

Διάμεσος

Η *διάμεσος* (median) ενός συνόλου n τιμών, συμβολίζεται με m στις παρατηρήσεις δείγματος και με M στις παρατηρήσεις πληθυσμού και είναι η τιμή της μεσαίας παρατήρησης, όταν όλες οι παρατηρήσεις είναι ταξινομημένες σε αύξουσα ή φθίνουσα σειρά. Έτσι, μισές παρατηρήσεις

έχουν τιμή μεγαλύτερη από τη διάμεσο και μισές έχουν τιμή μικρότερη από αυτήν. Αν n είναι περιττός αριθμός η διάμεσος είναι η τιμή της (n+1)/2 παρατήρησης. Εάν είναι ζυγός η διάμεσος τιμή είναι ο μέσος όρος των δύο μεσαίων τιμών.

Επικρατούσα Τιμή

Η *επικρατούσα,* ή *τυπική τιμή* (mode) ενός συνόλου παρατηρήσεων ορίζεται ως η τιμή με τη μεγαλύτερη συχνότητα.

Εύρος

Το πιο απλό μέτρο διασποράς ενός συνόλου τιμών είναι η διαφορά της μικρότερης από τη μεγαλύτερη τιμή τους. Η διαφορά αυτή ονομάζεται *εύρος* (range) των τιμών και συμβολίζεται με R.

Εκατοστιαία Σημεία

Συχνά είναι βολικό να διαιρούμε το εμβαδόν μεταξύ μιας καμπύλης πυκνότητας και του οριζόντιου άξονα με τετμημένες, έτσι ώστε το αριστερά μιας τετμημένης εμβαδόν να είναι μέρος του όλου εμβαδού. Το μέρος α του εμβαδού καλείται ποσοστό και η αντίστοιχη τετμημένη α-ποσοστιαίο σημείο.

Ακραίες Τιμές (Outliers)

Ακραίες (outliers) λέγονται οι τιμές, οι οποίες είναι, είτε πολύ μεγαλύτερες, είτε πολύ μικρότερες από τον κύριο όγκο των δεδομένων. Μπορεί να οφείλονται, είτε σε λανθασμένη καταχώρηση, είτε σε ιδιαίτερες συνθήκες, οι οποίες επικρατούσαν κατά την παρατήρηση τους, οπότε θα πρέπει να επανεξεταστούν. Σε κάθε περίπτωση ο αναλυτής θα πρέπει κατ' αρχήν να τις εντοπίσει και στη συνέχεια να τις εξετάσει με προσοχή.

Η Έννοια των βαθμών ελευθερίας

Οι *βαθμοί ελευθερίας* είναι ο αριθμός των ελεύθερων επιλογών που έχουμε στον προσδιορισμό της τιμής των μεταβλητών. Έστω π.χ. ότι θέλουμε να βρούμε τρεις αριθμούς α, β, γ που το άθροισμα τους ισούται με 10, δηλαδή έχουμε α + β + γ = 10.

Είναι προφανές ότι μπορούμε να επιλέξουμε ελεύθερα τους δύο έστω τους α, β και ο τρίτος υποχρεωτικά θα πάρει την τιμή γ = 10 – α – β. Δηλαδή η σχέση α + β + γ = 10 είναι ένας περιορισμός στην επιλογή των α, β, γ που έχει ως αποτέλεσμα να έχουμε 3 – 1 = 2 ελεύθερες επιλογές δηλαδή 2 βαθμούς ελευθερίας. Γενικά ισχύει:

Αριθμός των βαθμών ελευθερίας = αριθμός μεταβλητών – αριθμός περιορισμών

Ασυμμετρία

Εν γένει δεν υπάρχει τιμή της τυχαίας μεταβλητής, ως προς την οποία η συνάρτηση πιθανότητας να είναι συμμετρική.

Αν η τιμή της αδιάστατης ποσότητας:

$$\alpha_3 = \frac{E[(X-\mu)^3]}{\sigma^3}$$

είναι θετική, ή αρνητική, τότε λέμε ότι η καμπύλη είναι ασύμμετρη προς τα δεξιά ή αριστερά.

Κύρτωση

Μία κατανομή που είναι συγκεντρωμένη κοντά στη μέση τιμή παρουσιάζει έντονη κύρτωση. Σαν μέτρο της κύρτωσης παίρνουμε την τιμή της αδιάστατης ποσότητας:

$$\alpha_4 = \frac{E[(X-\mu)^4]}{\sigma^4}$$

Όσο πιο μεγάλη είναι αυτή η ποσότητα, τόσο μεγαλύτερη κύρτωση έχει η κατανομή, δηλαδή τόσο πιο συγκεντρωμένη είναι κοντά στη μέση

τιμή.

2.2 ΚΑΤΑΝΟΜΕΣ

Η ΔΥΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Ας υποθέσουμε ότι εκτελούμε ένα πείραμα τύχης πολλές φορές, (κλήρο, νόμισμα, ...). Κάθε φορά που εκτελούμε το πείραμα, λέμε ότι κάνουμε μία δοκιμή και σε κάθε δοκιμή έχουμε μία πιθανότητα να έχουμε επιτυχία. Σε πολλές περιπτώσεις η πιθανότητα παραμένει σταθερή δεν αλλάζει από δοκιμή σε δοκιμή και οι δοκιμές είναι ανεξάρτητες.

Έστω p η πιθανότητα να πραγματοποιηθεί το γεγονός σε μία μόνο δοκιμή και 1-p η πιθανότητα να μη πραγματοποιηθεί. Η πιθανότητα να πραγματοποιηθεί το γεγονός x φορές σε n δοκιμές είναι:

$$f(x)=P(X=x)=\frac{n!}{x!(n-x)!}\,p^{x}(1-p)^{n-x}$$

όπου,

- p = η πιθανότητα επιτυχίας σε κάθε δοκιμή
- x = ο αριθμός των επιτυχιών ή εμφάνισης του γεγονότος
- n = ο αριθμός των δοκιμών

Μέση τιμή	μ=np
Διασπορά	$\sigma^2=np(1-p)$
Τυπική απόκλιση	$\sigma=\sqrt{np(1-p)}$

Παράδειγμα

Ας υποθέσουμε ότι 8 αντικείμενα επιλέγονται τυχαία από μια γραμμή παραγωγής μιας επιχείρησης, η οποία είναι γνωστό ότι παράγει 5% ελαττωματικά προϊόντα κατά τη διάρκεια του χρόνου. Ποια είναι η

πιθανότητα να ληφθούν μηδέν ελαττωματικά;

Έχουμε: n = 8, p = 0.05, x = 0.

Η πιθανότητα είναι 0.6634 (πίνακας Β – 1, σελ. 523).

Η ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Μία από τις σημαντικότερες κατανομές είναι η κανονική κατανομή ή κατανομή του Gauss. Η συνάρτηση πυκνότητας είναι:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

Η συνάρτηση κατανομής είναι η:

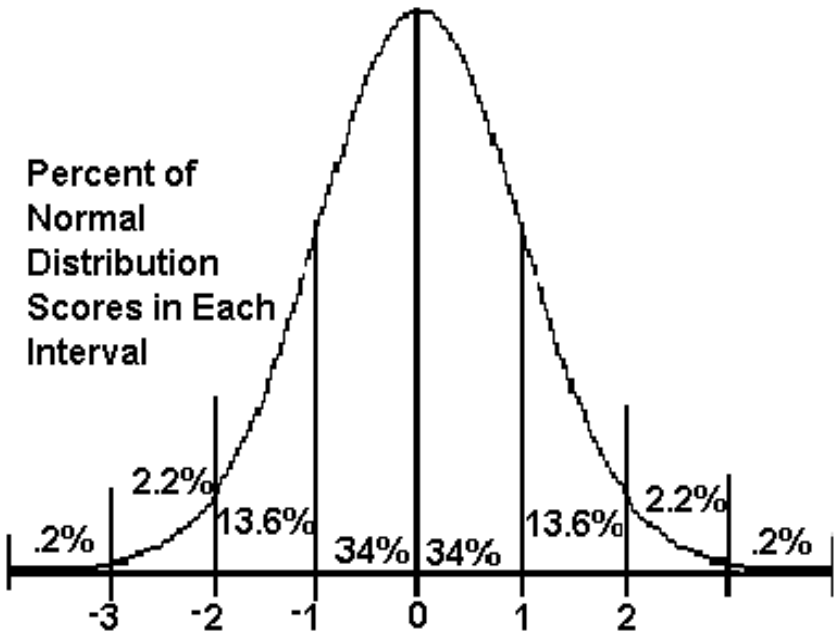
$$F(x)=P(X\le x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-(u-\mu)^2/2\sigma^2}du$$

Ορίζοντας $Z=\frac{X-\mu}{\sigma}$ ως την τυποποιημένη μεταβλητή, έχουμε:

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-z^2/2} \quad \text{και}$$

$$F(z)=P(Z\le z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-u^2/2}du=\frac{1}{2}+\frac{1}{\sqrt{2\pi}}\int_{0}^{z}e^{-u^2/2}du$$

όπου το δεύτερο ολοκλήρωμα είναι η error function.



Μέση τιμή	μ
Διασπορά	σ^2
Τυπική απόκλιση	σ

Παράδειγμα

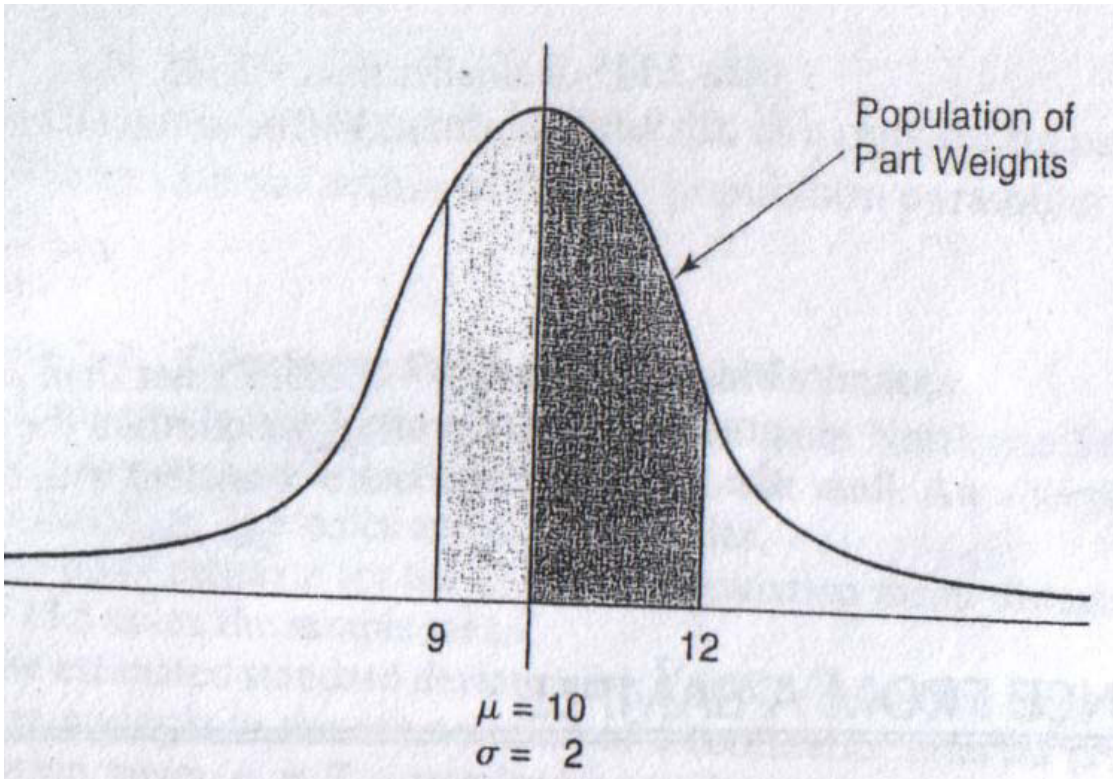
Τα βάρη ενός πληθυσμού προϊόντων κατασκευασμένα από την ίδια μηχανή είναι κατανεμημένα κανονικά με μέσον όρο 10 και τυπική απόκλιση 2, να υπολογιστεί η πιθανότητα P ($9 \leq x \leq 12$).

$$Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{9-10}{2} = -0.50, \quad Z_2 = \frac{X_2 - \mu}{\sigma} = \frac{12-10}{2} = 1.00$$

$Z_1 = -0.50$ αντιστοιχεί στο 0.1915 (πίνακας Β – 2, σελ. 525)

$Z_2 = 1.00$ αντιστοιχεί στο 0.3413 (πίνακας Β – 2, σελ. 525)

Οπότε η συνολική πιθανότητα θα ισούται με 0.1915 + 0.3413 = 0.5328 = 53,28%.



Παράδειγμα (Επιλογή αμοιβαίως αποκλειόμενων επενδύσεων – Ανάλυση κινδύνου με τη χρησιμοποίηση σεναρίων)

Έστω μία τεχνική εταιρεία, η οποία θέλει να κατασκευάσει συγκρότημα κατοικιών με τη διαδικασία της αντιπαροχής. Οι διακυμάνσεις στη ζήτηση διαμερισμάτων, η οποία είναι απόρροια κυρίως των οικονομικών συνθηκών (οικονομική ανάπτυξη, στασιμότητα, ύφεση) και της γενικότερης στεγαστικής πολιτικής της κυβέρνησης θα επηρεάσουν τις χρηματοροές της τεχνικής εταιρείας. Για τη συγκεκριμένη επένδυση αναπτύχθηκαν τρία πιθανά σενάρια και συνδέθηκαν με εκτιμήσεις καθαρών χρηματοροών και πιθανοτήτων. Τα σενάρια φαίνονται στον παρακάτω πίνακα:

Σενάρια συνθηκών	**Χαρακτηρισμός εκτιμήσεων**	**Πιθανότητα πραγματοποίησης f_i**	**Καθαρά παρούσα αξία $(ΚΠΑ)_i$**

Πολύ καλές	Αισιόδοξες	0,3	16000
Κανονικές	Πιο πιθανές	0,5	7000
Κακές	Συντηρητικές	0,2	-2000

Η μέση αναμενόμενη καθαρά παρούσα αξία είναι:

$$E(ΚΠΑ)=\sum_{i=1}^{3}(ΚΠΑ)_i f((ΚΠΑ)_i)=7900.$$

Η διακύμανση είναι:

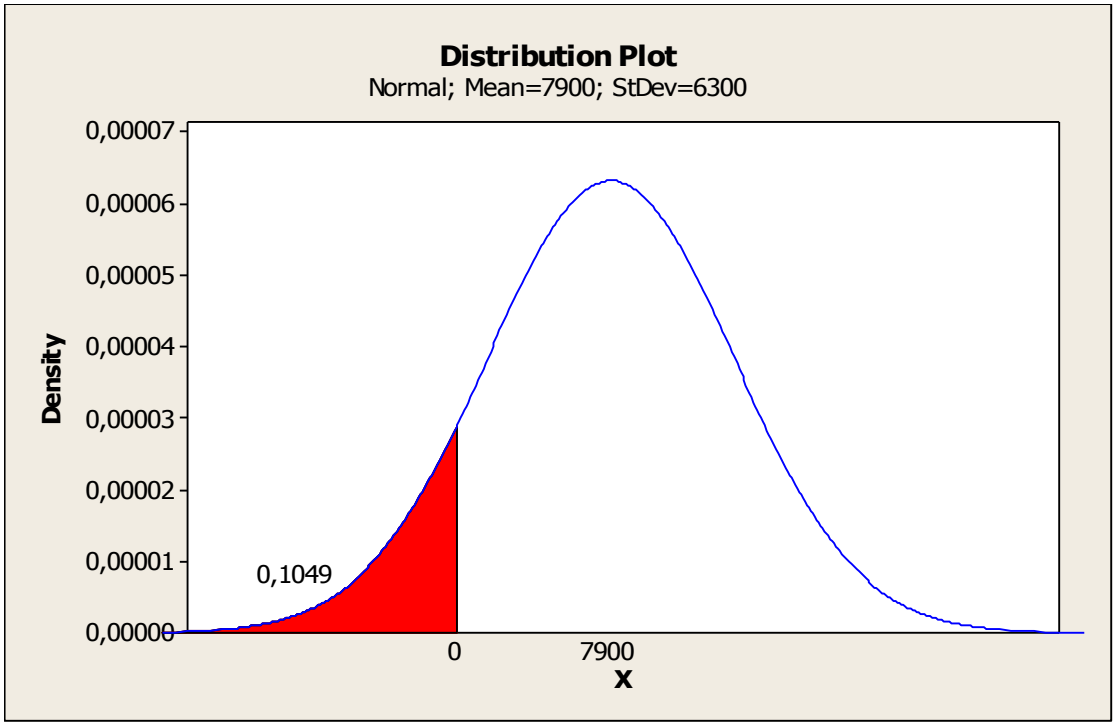
$$\sigma^2_{ΚΠΑ}=\sum_{i=1}^{3}[(ΚΠΑ)_i-E(ΚΠΑ)]^2 f((ΚΠΑ)_i)=39690000.$$

Οπότε η τυπική απόκλιση είναι $\sigma_{ΚΠΑ}=6300.$

Θεωρούμε τώρα ότι η κατανομή, που προσαρμόζεται καλά στα δεδομένα (αν και δεν είναι πολλά), είναι κανονική. Οπότε μπορούμε να υπολογίσουμε πιθανότητες για τρεις περιπτώσεις τιμών της ΚΠΑ.

1. Να είναι: ΚΠΑ > 0. Αυτό αντιστοιχεί σε πιθανότητα 1-0,1049 ή 89,51%:

$$Z=\frac{ΚΠΑ-E(ΚΠΑ)}{\sigma_{ΚΠΑ}}=\frac{0-7900}{6300}=-1.25$$



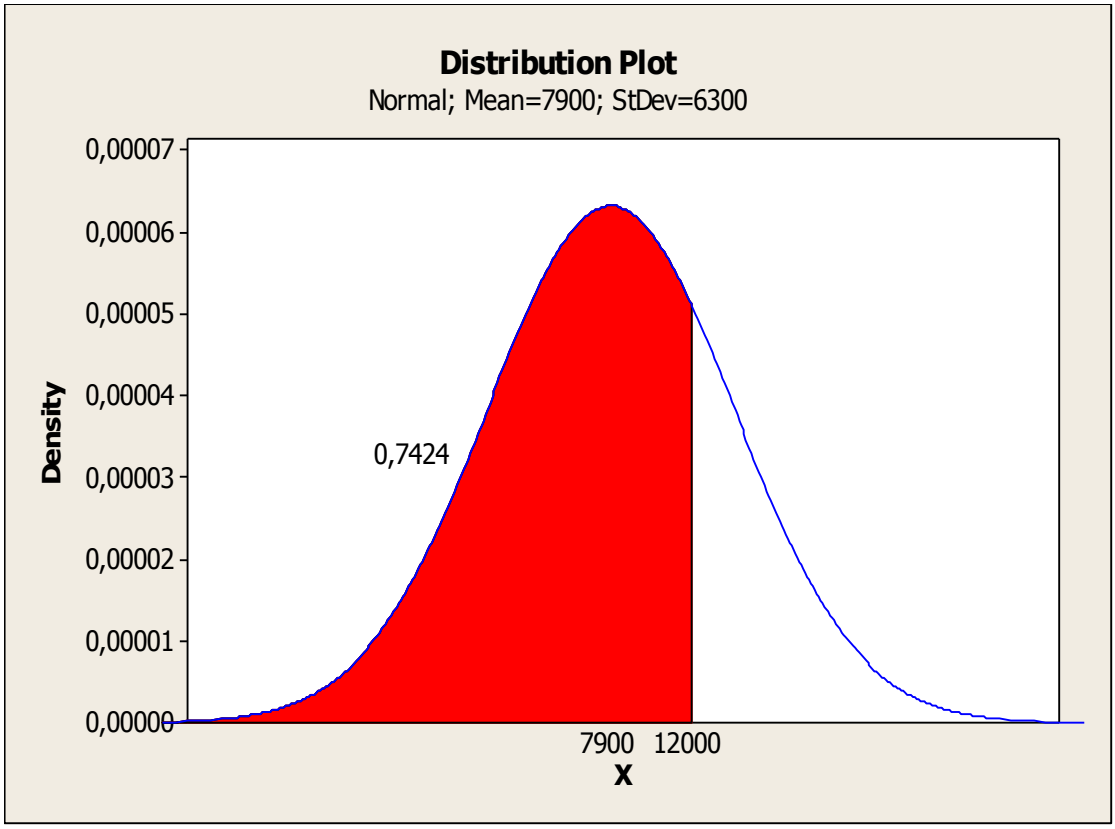
2. Να είναι: 4000 < ΚΠΑ < 12000

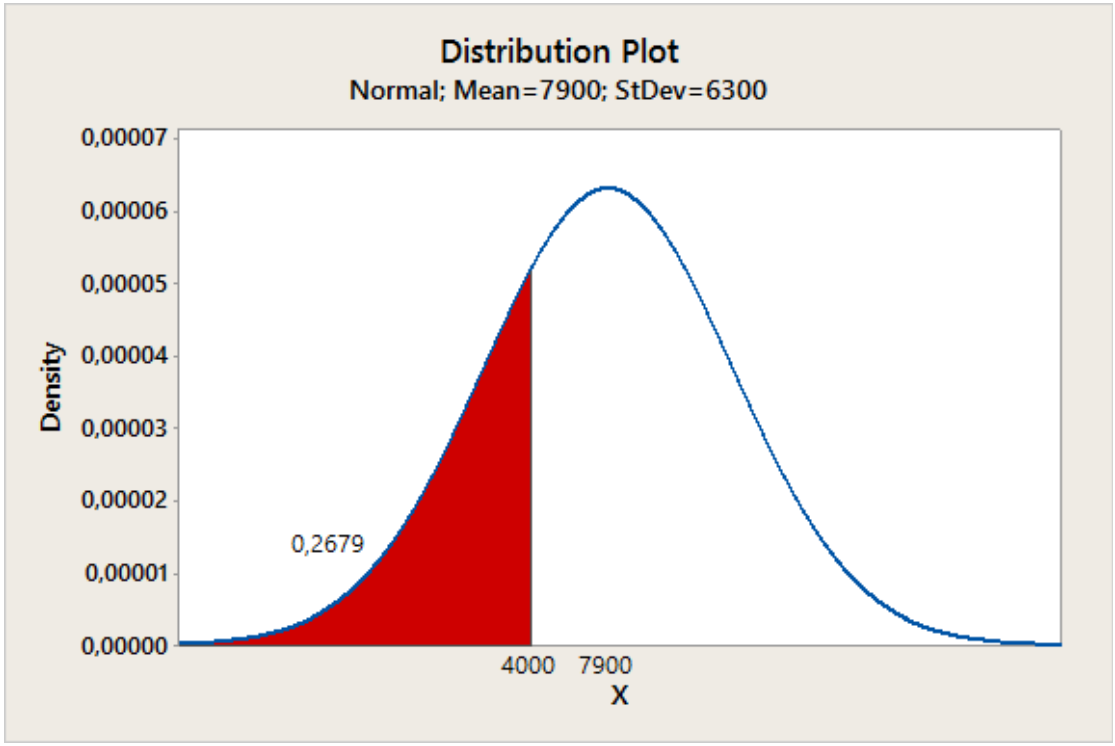
$$Z=\frac{ΚΠΑ-E(ΚΠΑ)}{\sigma_{ΚΠΑ}}=\frac{4000-7900}{6300}=-0.619$$

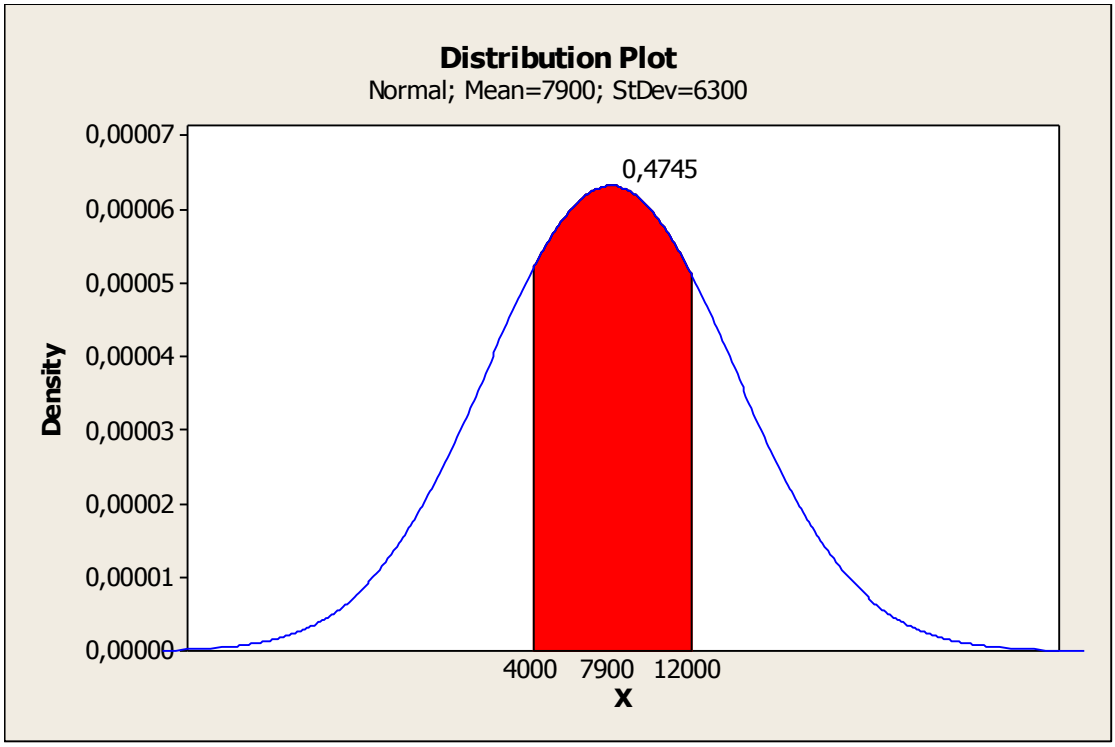
$$Z=\frac{ΚΠΑ-E(ΚΠΑ)}{\sigma_{ΚΠΑ}}=\frac{12000-7900}{6300}=0.65$$

Αυτό αντιστοιχεί σε πιθανότητα 0,7424 – 0,2679 = 0,4745 ή 47,45%.

3. Να είναι: ΚΠΑ > 12000. Η πιθανότητα είναι 1 – 0,7424 = 0,2576 ή 25,76%.


Distribution Plot
Normal; Mean=7900; StDev=6300
Density
0,00007
0,00006
0,00005
0,00004
0,00003
0,00002
0,00001
0,00000
0,7424
7900
12000
X


Distribution Plot
Normal; Mean=7900; StDev=6300
Density
0,00007
0,00006
0,00005
0,00004
0,00003
0,00002
0,00001
0,00000
0,2679
4000
7900
X



Προκειμένου να αποφασίσει η εταιρεία, αν η επένδυση αυτή ενέχει τους λιγότερους κινδύνους μελέτησε και ένα εναλλακτικό σενάριο επένδυσης με τα ακόλουθα χαρακτηριστικά.

Σενάρια συνθηκών	**Χαρακτηρισμός εκτιμήσεων**	**Πιθανότητα πραγματοποίησης f_i**	**Καθαρά παρούσα αξία $(ΚΠΑ)_i$**
Πολύ καλές	Αισιόδοξες	0,3	12000
Κανονικές	Πιο πιθανές	0,5	6000
Κακές	Συντηρητικές	0,2	4000

Η αναμενόμενη τιμή είναι Ε(ΚΠΑ) = 7400 και η τυπική απόκλιση 3104. Οι αντίστοιχες πιθανότητες είναι:

1. Για ΚΠΑ>0, P=100%.
2. Για 4000 < ΚΠΑ < 12000, P =79,57%.
3. Για ΚΠΑ > 12000, P=6,64%.

Αμοιβαία αποκλειόμενες επενδύσεις	Ε(ΚΠΑ)	Πιθανότητα για ΚΠΑ > 0	Πιθανότητα για 4000 < ΚΠΑ < 12000
Επένδυση 1	7900	89,51%	47,42%
Επένδυση 2	7400	100,00%	79,57%

Το ποια επένδυση θα κάνει τελικά η εταιρεία εξαρτάται από το στυλ της. Αν για παράδειγμα η εταιρεία στρατηγικά ρισκάρει θα πραγματοποιήσει την πρώτη επένδυση. Αν όχι τη δεύτερη.

Η ΚΑΤΑΝΟΜΗ χ^2

Έστω ότι η κάθε μία από τις ν ανεξάρτητες τυχαίες μεταβλητές X_1, X_2, X_3,...,X_v ακολουθεί κανονική κατανομή με μέση τιμή 0 και διασπορά 1. Θεωρούμε την τυχαία μεταβλητή:

$$\chi^2 = X_1^2 + X_2^2 + X_3^2 + ... + X_v^2$$

Μπορεί να δειχθεί ότι για $x \geq 0$ η χ^2 ακολουθεί τη κατανομή:

$$F(x) = P(\chi^2 \leq x) = \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^x u^{v/2-1} e^{-u/2} du \,, \forall x > 0$$

και $F(x) = P(\chi^2 \leq x) = 0$ για x<0.

Η κατανομή αυτή καλείται κατανομή χ^2, "χι τετράγωνο", με ν βαθμούς ελευθερίας.

Η συνάρτηση πυκνότητας ορίζεται ως:

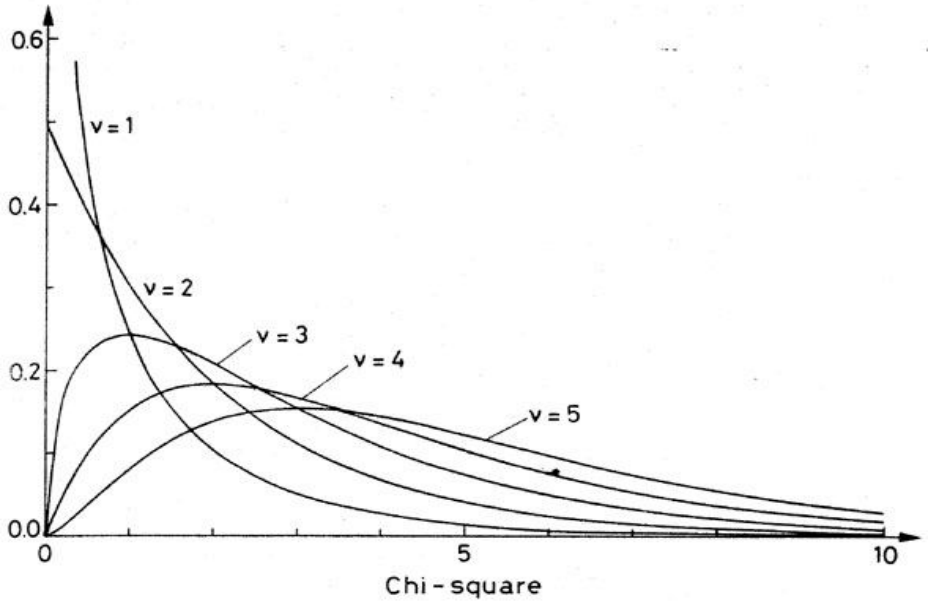
$$f(x) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2} \,, \forall x > 0$$

και $f(x) = 0, \forall x \leq 0$

Μέση τιμή	μ=ν
Διασπορά	$\sigma^2=2\nu$
Τυπική απόκλιση	$\sigma = \sqrt{2\nu}$

Για ν αρκετά μεγάλο $\nu \geq 30$ μπορεί να δειχθεί, ότι η μεταβλητή

$\sqrt{2\chi^2}-\sqrt{2\nu-1}$ ακολουθεί την κανονική κατανομή N(0,1).



Η ΚΑΤΑΝΟΜΗ t ΤΟΥ STUDENT

Μία μεταβλητή με συνάρτηση πυκνότητας:

$$f(t)=\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty<t<\infty$$

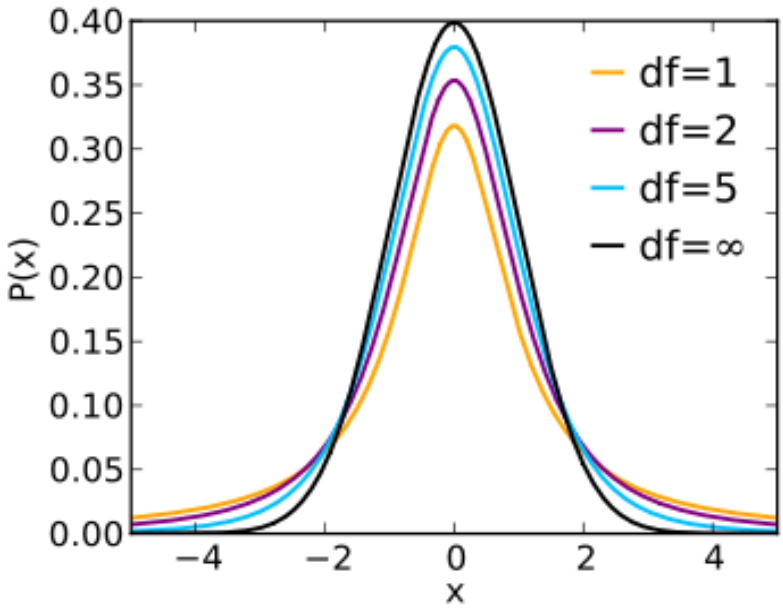
ακολουθεί την κατανομή t του Student με ν βαθμούς ελευθερίας.

Η συνάρτηση κατανομής δίδεται από τη σχέση:

$$F(t)=P(X\le t)=\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\int_{-\infty}^{t}\left(1+\frac{u^2}{v}\right)^{-(v+1)/2}du$$

Μέση τιμή	μ=0
Διασπορά	$\sigma^2=\frac{v}{v-2}, v>2$

Τυπική απόκλιση	$\sigma=\sqrt{\frac{v}{v-2}}, v>2$



Για ν αρκετά μεγάλο $v \geq 30$ η f(t) προσεγγίζει την κανονική κατανομή.

Η ΚΑΤΑΝΟΜΗ F

Μία τυχαία μεταβλητή ακολουθεί την κατανομή F με ν_1 και ν_2 βαθμούς ελευθερίας, εάν έχει συνάρτηση πυκνότητας:

$$f(t)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}t^{\nu_1/2-1}\left(\nu_2+\nu_1 t\right)^{-(\nu_1+\nu_2)/2}, \quad t>0$$

$$\text{και} \quad f(t)=0, \quad \gamma\iota\alpha \quad t\leq 0$$

Μέση τιμή	$\mu=\frac{\nu_2}{\nu_2-2}, \quad \nu_2>2$
Διασπορά	$\sigma^2=\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2}, \quad \nu_2>4$

Τυπική απόκλιση	$\sigma=\sqrt{\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2}}, \quad \nu_2>4$
Πιθανότερη τιμή	$u=\left(\frac{\nu_1-2}{\nu_1}\right)\left(\frac{\nu_2}{\nu_2+2}\right), \quad \nu_1>2$

2.3 ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Στην πράξη θέλουμε να βγάλουμε συμπεράσματα για μια μεγάλη ομάδα ατόμων ή αντικειμένων. Αντί να μελετήσουμε ολόκληρη την ομάδα, τον πληθυσμό όπως λέμε, εξετάζουμε ένα μικρό εύρος αυτού του πληθυσμού, δηλαδή ένα ***δείγμα***. Η διαδικασία λήψεως ενός δείγματος λέγεται ***δειγματοληψία***.

Εάν βγάλουμε ένα αντικείμενο από ένα κουτί, μπορούμε να το ξαναβάλουμε, ή να μη το ξαναβάλουμε πριν βγάλουμε ένα άλλο. Στην πρώτη περίπτωση έχουμε ***δειγματοληψία με επανατοποθέτηση*** και στη δεύτερη ***δειγματοληψία χωρίς επανατοποθέτηση***.

Ένας πεπερασμένος πληθυσμός μπορεί να θεωρηθεί άπειρος σε δειγματοληψία με επανατοποθέτηση, επειδή μπορούμε να πάρουμε δείγματα οποιουδήποτε μεγέθους.

Ως ***στατιστική δειγματική συνάρτηση*** (που είναι συνάρτηση των τυχαίων μεταβλητών του δείγματος) ορίζουμε κάθε ποσότητα, που προκύπτει από ένα δείγμα στη διάρκεια εκτίμησης των παραμέτρων του πληθυσμού, έστω π.χ. το δειγματικό μέσο. Για κάθε παράμετρο θα έχουμε διαφορετική συνάρτηση. Φυσικά η τιμή μιας στατιστικής συνάρτησης μεταβάλλεται από δείγμα σε δείγμα, θα είναι και η ίδια τυχαία μεταβλητή και για κάθε δείγμα θα έχουμε μία τιμή αυτής της στατιστικής συνάρτησης. Άρα μπορούμε να μιλάμε για κατανομή

πιθανότητας αυτής της στατιστικής συνάρτησης και ονομάζεται ***κατανομή δειγματοληψίας*** (sampling distribution) π.χ. του $\overline{X}$.

Στη δειγματοληψία είναι επιθυμητό να πάρουμε ένα δείγμα, το οποίο να αποτελεί μια μικρογραφία του πληθυσμού, να είναι δηλαδή ***αντιπροσωπευτικό***. Αγνοώντας όμως τα χαρακτηριστικά του πληθυσμού είναι αδύνατο να προσδιορίσουμε, ποιο είναι το αντιπροσωπευτικό δείγμα και άρα να το επιλέξουμε. Αποδεικνύεται ότι όταν η δειγματοληψία είναι τυχαία (δηλαδή η πιθανότητα, κάθε στοιχείο να βρεθεί στο δείγμα, να είναι ίδια), τότε η πιθανότητα να προκύψει ένα δείγμα χονδρικά αντιπροσωπευτικό είναι μεγαλύτερη. Θα πρέπει όμως να τονιστεί ότι το τέλεια αντιπροσωπευτικό δείγμα δεν υπάρχει. Έτσι, συμπεραίνοντας για τον πληθυσμό με βάση τις πληροφορίες του δείγματος θα κάνουμε αναπόφευκτα ένα σφάλμα, το οποίο ονομάζεται ***σφάλμα δειγματοληψίας*** (sampling error). Το σφάλμα δειγματοληψίας είναι αναπόφευκτο. Τα συμπεράσματά μας όμως για τον πληθυσμό μπορεί να περιέχουν και άλλου είδους σφάλμα, το οποίο μπορεί να οφείλεται στην κακή μεθοδολογία μέτρησης, ή στη μη τυχαιότητα του δείγματος. Ένα τέτοιο σφάλμα ονομάζεται ***σφάλμα μεροληψίας*** (bias error) και δεν μπορεί να μετρηθεί σε καμιά περίπτωση.

2.3.1 ΔΕΙΓΜΑΤΙΚΗ ΜΕΣΗ ΤΙΜΗ

Εάν οι X_1, X_2, X_3, ..., X_n ανεξάρτητες και ισόνομες τυχαίες μεταβλητές για ένα δείγμα (αντιπροσωπευτικό) μεγέθους n, ως δειγματική μέση τιμή ορίζεται η τυχαία μεταβλητή:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

2.3.2 ΚΑΤΑΝΟΜΗ ΤΗΣ ΔΕΙΓΜΑΤΙΚΗΣ ΜΕΣΗΣ ΤΙΜΗΣ

Έστω f(x) η κατανομή πιθανότητας ενός πληθυσμού. Εάν πάρουμε ένα δείγμα μεγέθους n από τον πληθυσμό είναι φυσικό να ζητήσουμε την κατανομή πιθανότητας της στατιστικής συνάρτησης $\overline{X}$. Αυτή είναι η δειγματοληπτική κατανομή της μέσης τιμής του δείγματος. Σχετικά ισχύουν τα ακόλουθα:

- Η μέση τιμή της δειγματικής μέσης τιμής $\overline{X}$ είναι $E(\overline{X}) = \mu_{\overline{X}} = \mu$.
- Εάν ο πληθυσμός είναι άπειρος ή με επανατοποθέτηση, τότε η διασπορά $\sigma^2_{\bar{X}}$ της δειγματικής μέσης τιμής, ή *τυπικό σφάλμα του δειγματικού μέσου* είναι $E[(\bar{X}-\mu)^2]=\sigma_{\bar{X}}{}^2=\frac{\sigma^2}{n}$, όπου σ^2 είναι η διασπορά του πληθυσμού με n το μέγεθος του δείγματος.
- Εάν ο πληθυσμός είναι κανονικός με μέση τιμή μ και διασπορά σ^2, τότε η δειγματική μέση τιμή ακολουθεί κανονική δειγματοληπτική κατανομή με μέση τιμή μ και διασπορά σ^2/n.
- ***Κεντρικό οριακό θεώρημα*** (central limit theorem): Έστω ο πληθυσμός, από όπου παίρνουμε τα δείγματα, ακολουθεί κάποια κατανομή, όχι αναγκαστικά κανονική, με μέση τιμή μ και διακύμανση σ^2. Καθώς αυξάνει το n (το μέγεθος του δείγματος δηλαδή) η τυποποιημένη μεταβλητή $Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ προσεγγίζει ασυμπτωτικά την τυπική κανονική κατανομή με μέση τιμή 0 και διασπορά 1.

Η $\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}$, που ονομάζεται *τυπικό σφάλμα του δειγματικού μέσου* (standard error of the sample mean), μειώνεται στο μισό με τον τετραπλασιασμό του μεγέθους του δείγματος. Εξάλλου για δεδομένο

μέγεθος του δείγματος το τυπικό σφάλμα είναι ανάλογο προς την τυπική απόκλιση σ του πληθυσμού.

2.3.3 ΔΕΙΓΜΑΤΙΚΗ ΔΙΑΣΠΟΡΑ

Εάν X_1, X_2, X_3, ..., X_n ανεξάρτητες και τυχαίες μεταβλητές για ένα δείγμα μεγέθους n, τότε η τυχαία μεταβλητή, που δίνει τη δειγματική διασπορά ή διακύμανση είναι:

$$S^2=\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+...+(X_n-\bar{X})^2}{n}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$$

Επειδή $E(\bar{X})=\mu$ θα ήταν πολύ ωραίο να είχαμε και $E(S^2)=\sigma^2$. Είναι όμως $E(S^2)=\mu_{S^2}=\frac{n-1}{n}\sigma^2$ που για μεγάλα n>30 δεν διαφέρει και πολύ από το σ^2. Εάν ορίσουμε την μεταβλητή:

$$\hat{S}^2=\frac{n}{n-1}S^2=\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+...+(X_n-\bar{X})^2}{n-1}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$$

τότε θα έχουμε $E(\hat{S}^2)=\sigma^2$.

Επίσης πολλές φορές είναι βολικότερο να χρησιμοποιήσουμε την ακόλουθη μεταβλητή:

$$\frac{nS^2}{\sigma^2}=\frac{(n-1)\hat{S}^2}{\sigma^2}=\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+...+(X_n-\bar{X})^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}$$

Η μεταβλητή αυτή ακολουθεί τη κατανομή χ^2 με n-1 βαθμούς ελευθερίας.

2.3.4 ΑΓΝΩΣΤΗ ΔΙΑΣΠΟΡΑ ΠΛΗΘΥΣΜΟΥ

Σύμφωνα με τα προηγούμενα η τυποποιημένη μεταβλητή

$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ είναι κανονική, εάν ο πληθυσμός είναι κανονικός ή ασυμπτωτικά κανονική, εάν ο πληθυσμός δεν είναι κανονικός, αλλά το n είναι πολύ μεγάλο.

Συνήθως όμως η διασπορά του πληθυσμού δεν είναι γνωστή. Μία λύση είναι να αντικαταστήσουμε την σ με την τυχαία μεταβλητή $\hat{S}$ (την εκτιμήτρια του σ δηλαδή) και να μελετήσουμε τη δειγματοληπτική κατανομή της νέας στατιστικής συνάρτησης $T = \frac{\bar{X} - \mu}{\hat{S} / \sqrt{n}} = \frac{\bar{X} - \mu}{S / \sqrt{n-1}}$.

Αν το n είναι μικρότερο του 30, τότε η τυποποιημένη μεταβλητή T ακολουθεί τη κατανομή t του Student με n-1 βαθμούς ελευθερίας, ενώ εάν το n-1 είναι μεγαλύτερο ή ίσο του 30, τότε η τυποποιημένη μεταβλητή T τείνει ασυμπτωτικά στην κανονική κατανομή.

Παράδειγμα

Υπολογισμός του μέσου όρου, της διακύμανσης και της τυπικής απόκλισης.

Η ακόλουθη συλλογή είναι ηλικίες ατόμων: 23, 38, 42, 25, 60, 55, 42, 32, 35. Για αυτό το δείγμα το n = 10 και

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \ldots + X_{10}}{10} = \frac{402}{10} = 40.2$$

$$\hat{S}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{(23-40.2)^2 + \ldots + (35-40.2)^2}{10-1} = \frac{1,339.6}{9} = 148.84$$

$$\hat{S} = \sqrt{\hat{S}^2} = 12.2$$

2.4 ΠΑΡΟΥΣΙΑΣΗ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Τα στατιστικά δεδομένα παρουσιάζονται οργανωμένα με πίνακες ή (και) με κατάλληλες γραφικές παραστάσεις. Ο τρόπος οργάνωσης και γραφικής τους παράστασης εξαρτάται από το είδος των δεδομένων.

Ειδικότερα διακρίνουμε τις ακόλουθες περιπτώσεις:

ΔΙΑΓΡΑΜΜΑ ΔΙΑΣΠΟΡΑΣ

Τα *διαγράμματα διασποράς* (scatter diagrams) χρησιμοποιούνται για να απεικονίσουν τη σχέση μεταξύ δύο μεταβλητών.

ΡΑΒΔΟΓΡΑΜΜΑ

Ένα *ραβδόγραμμα* (bar chart) κατασκευάζεται ως εξής: στον οριζόντιο άξονα υψώνονται σε ισαπέχοντα σημεία ορθογώνια παραλληλόγραμμα με ύψος ίσο με την αντίστοιχη συχνότητα. Το κοινό πάχος των παραλληλόγραμμων επιλέγεται αυθαίρετα, είναι όμως τέτοιο, ώστε να εξασφαλίζονται κενά μεταξύ διαδοχικών παραλληλογράμμων.

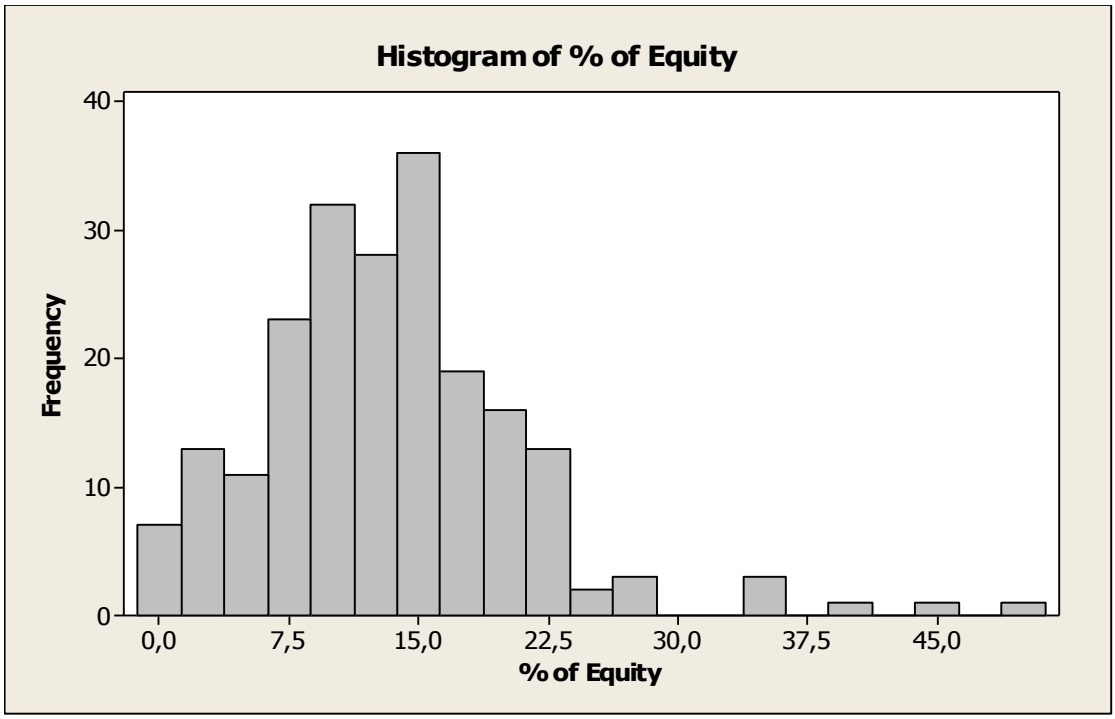
ΧΡΟΝΟΛΟΓΙΚΟ ΔΙΑΓΡΑΜΜΑ

Για χρονολογικά δεδομένα, η πιο συχνά χρησιμοποιούμενη γραφική μορφή είναι ένα *χρονολογικό διάγραμμα* (time plot, run chart), κατά το οποίο τα δεδομένα απεικονίζονται στην πάροδο του χρόνου. Στον οριζόντιο άξονα μετρούμε τον χρόνο και στον κάθετο την τιμή του μεγέθους, οπότε για κάθε παρατήρηση παίρνουμε ένα σημείο στο γράφημα. Στη συνέχεια ενώνουμε τα γειτονικά σημεία με ευθύγραμμα τμήματα. Αυτό συνίσταται, ιδιαίτερα, όταν ο αριθμός των παρατηρήσεων είναι μεγάλος.

ΙΣΤΟΓΡΑΜΜΑ

Το *ιστόγραμμα* (histogram) είναι ίσως η πιο σημαντική γραφική παράσταση των δεδομένων. Αποτελεί μια ακολουθία ορθογωνίων παραλληλογράμμων, όσα και οι τάξεις της κατανομής. Κάθε

παραλληλόγραμμο έχει πλάτος και ύψος ίσα με το εύρος και τη συχνότητα της αντίστοιχης κλάσης. Το σχήμα του αναπαριστά την κατανομή πιθανότητας ή πυκνότητα πιθανότητας.



ΚΥΚΛΙΚΟ ΔΙΑΓΡΑΜΜΑ

Όταν μας ενδιαφέρει η παρουσίαση της ποσοστιαίας αναλογίας των διαφόρων κατηγοριών ενός ολικού μεγέθους χρησιμοποιούμε το *κυκλικό διάγραμμα* (pie chart).

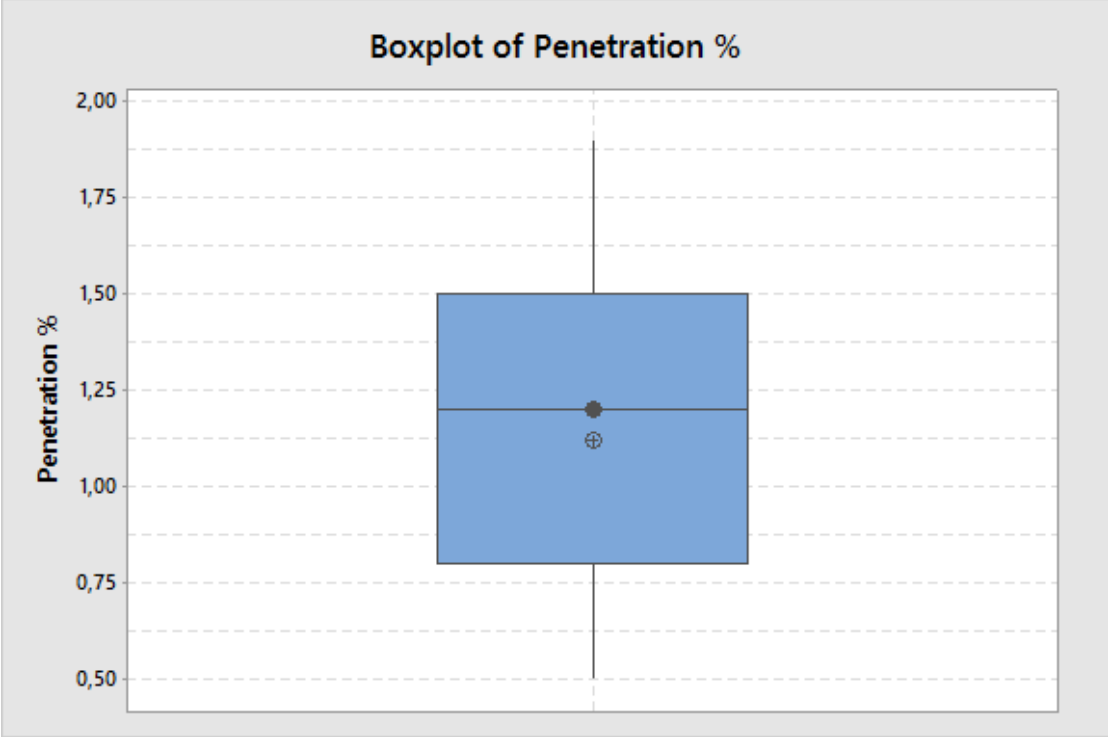
BOX-PLOT

Από ένα Box-Plot μπορούμε να δούμε συμμετρία ή ασυμμετρία, τη διασπορά και ακραίες τιμές. Στο παρακάτω παράδειγμα δίδεται το Box-Plot της διείσδυσης της ευρυζωνικότητας, σε ένα σύνολο τιμών για 23 διαδοχικά τρίμηνα: 0,6 0,6 0,5 0,5 0,9 0,8 0,7 0,8 1,1 1,2 1,0 1,2 1,5 1,5 1,2 1,4 1,9 1,5 1,1 1,2 1,6 1,6 1,3.

Η οριζόντια γραμμή παριστά τον διάμεσο και το σύμβολο από κάτω ακριβώς την μέση τιμή. Το κάτω όριο του κουτιού Q_1 αντιστοιχεί στο 25%-εκατοστιαίο σημείο και το άνω όριο του κουτιού Q_3 αντιστοιχεί

στο 75%-εκατοστιαίο σημείο.

Το άνω άκρο της κάθετης γραμμής αντιστοιχεί στο $Q_3+1,5(Q_3-Q_1)$ και το κάτω άκρο της κάθετης γραμμής αντιστοιχεί στο $Q_3-1,5(Q_3-Q_1)$. Τα ακραία σημεία ταυτοποιούνται για τιμές μεγαλύτερες του άνω άκρου ή μικρότερες του κάτω άκρου και συμβολίζονται στο διάγραμμα με αστερίσκους.



2.5 ΕΚΤΙΜΗΤΙΚΗ

Η μέση τιμή και η διακύμανση, όταν αναφέρονται στον πληθυσμό, ονομάζονται *παράμετροι* (parameters). Οι παράμετροι είναι σταθερές και σχεδόν πάντα άγνωστες ποσότητες. Μπορούμε να εκτιμήσουμε μία παράμετρο, για παράδειγμα τη μέση τιμή ενός δείγματος, έχοντας με αυτό τον τρόπο σημειακή εκτίμηση (point estimate). Μπορούμε όμως να δώσουμε ένα διάστημα, μέσα στο οποίο θα βρίσκεται με μεγάλη πιθανότητα η μέση τιμή (interval estimate).

Αν για παράδειγμα υπολογίζουμε την μέση τιμή μιας μεταβλητής, η οποία παρουσιάζει κάποια διασπορά (τυπική απόκλιση), τότε είναι εύλογο να χρησιμοποιήσουμε τη σημειακή εκτίμηση $\overline{X}$ και να πούμε, ότι θα έχουμε μεγάλη πιθανότητα η εκτίμηση μας για το μ να βρίσκεται σε κάποιο διάστημα, π.χ. $\overline{X}$-σ έως $\overline{X}$+σ, ή $\overline{X}$-2σ έως $\overline{X}$+2σ, κ.ο.κ. Ο αριθμός μπροστά από το σ καθορίζεται κάθε φορά από συγκεκριμένη κατανομή με συγκεκριμένη πιθανότητα.

2.5.1 ΤΟ ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Για πολλές πρακτικές εφαρμογές η εκτίμηση της τιμής μιας παραμέτρου Θ ενός πληθυσμού με την τιμή μιας στατιστικής σε ορισμένο τυχαίο δείγμα θεωρείται αρκετή. Υπάρχουν όμως εφαρμογές, όπου θέλουμε να γνωρίζουμε, πόση εμπιστοσύνη μπορούμε να έχουμε, όταν χρησιμοποιούμε π.χ. τη μέση τιμή ενός τυχαίου δείγματος, για να εκτιμήσουμε την μέση τιμή του πληθυσμού.

Στη στατιστική, το *διάστημα εμπιστοσύνης* (Confidence Interval – CI) είναι ένα διάστημα εκτίμησης μιας παραμέτρου. Αντί να εκτιμούμε την παράμετρο με μία μόνο τιμή, δίνουμε μαζί και το διάστημα πιθανότητας για την παράμετρο αυτή. Συνεπώς, τα διαστήματα εμπιστοσύνης χρησιμοποιούνται για να υποδείξουν την εγκυρότητα της παραμέτρου, που θέλουμε να προβλέψουμε. Η πιθανότητα της τιμής της παραμέτρου να συμπεριλαμβάνεται στα διαστήματα εμπιστοσύνης καθορίζεται από το επίπεδο εμπιστοσύνης (παράμετρος εμπιστοσύνης).

Στις περιπτώσεις αυτές εκτιμούμε ένα διάστημα τιμών [u,v], στο οποίο θα βρίσκεται η τιμή της παραμέτρου Θ με ορισμένη πιθανότητα 1-α, $0 \leq \alpha \leq 1$. Δηλαδή έχουμε:

$$P(u \leq \Theta \leq v) = 1-\alpha, \text{ με } 0 \leq \alpha \leq 1$$

Το [u,v] ονομάζεται *διάστημα εμπιστοσύνης* (confidence interval)

με πιθανότητα ή *επίπεδο εμπιστοσύνης* (level of confidence) 1-α. Οι τιμές u, v ονομάζονται *όρια εμπιστοσύνης* (confidence limits) του διαστήματος και η πιθανότητα α ονομάζεται *επίπεδο σημαντικότητας* (level of significance).

Παράδειγμα

Ένα τυχαίο δείγμα από 500 καταναλωτές συλλέχθηκε από ένα εμπορικό κέντρο για να καθοριστεί η μέση απόσταση, που διανύουν, για να πάνε στο εμπορικό κέντρο. Η ανάλυση κατέληξε στα εξής: $\overline{X}$ = 23.5 χλμ και $\hat{S}=10.4$ χλμ.

Για τη μέση τιμή μ του πληθυσμού με επίπεδο εμπιστοσύνης 1-α, και χρησιμοποιώντας τον ορισμό της τυποποιημένης μεταβλητής Z εκτιμούμε το ακόλουθο διάστημα για τη μ:

$$Z=\frac{\bar{X}-\mu}{\hat{S}/\sqrt{n}} \Rightarrow \quad \mu=\bar{X}\pm Z\frac{\hat{S}}{\sqrt{n}}$$

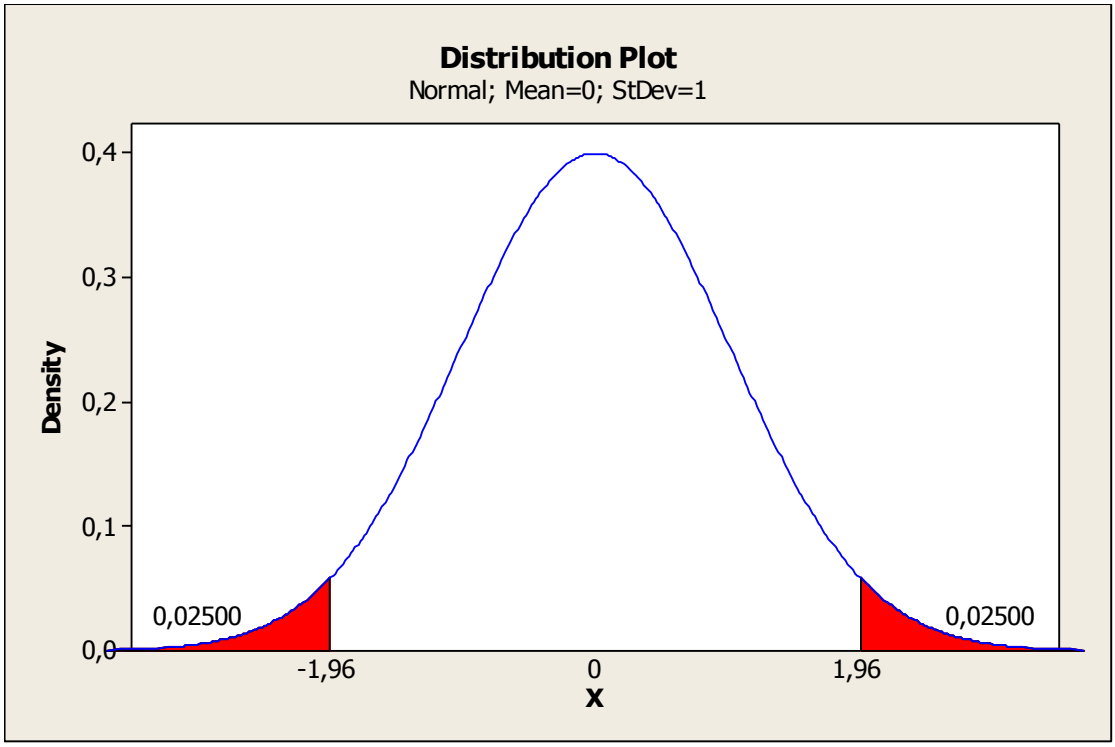
Η Z ακολουθεί κάποια κατανομή, η οποία προέρχεται από τον πληθυσμό, αν είναι γνωστό (για παράδειγμα ξέρουμε ότι ο πληθυσμός είναι κανονικός), ή από μαθηματική ανάλυση (ξέρουμε, ότι η τυποποιημένη μεταβλητή για την μέση τιμή ακολουθεί την κατανομή Student t). Επειδή εδώ το δείγμα μας είναι μεγάλο n>30 χρησιμοποιούμε την κανονική κατανομή.

Ας υποθέσουμε, ότι θέλουμε 95% εμπιστοσύνη, άρα θα έχουμε διάστημα εμπιστοσύνης 95%, οπότε το Z = 1.96. Χρησιμοποιώντας την παραπάνω εξίσωση μαζί με τα αποτελέσματα της έρευνας έχουμε:

$$\mu=23.5\pm1.96\frac{10.4}{\sqrt{500}}$$

$$=23.5 \pm 1.96 \text{ x } (.465)$$

$$=23.5 \pm 0.91 \rightarrow (22.6, 24.4)$$



Συνεπώς, με διάστημα εμπιστοσύνης 95% ο μέσος αριθμός χιλιομέτρων, που διανύουν οι καταναλωτές για να φτάσουν στο εμπορικό κέντρο, είναι ανάμεσα σε 22.6 χλμ και 24.4 χλμ.

Αν το n = 20 και το διάστημα εμπιστοσύνης είναι 95% τότε θα έχουμε:

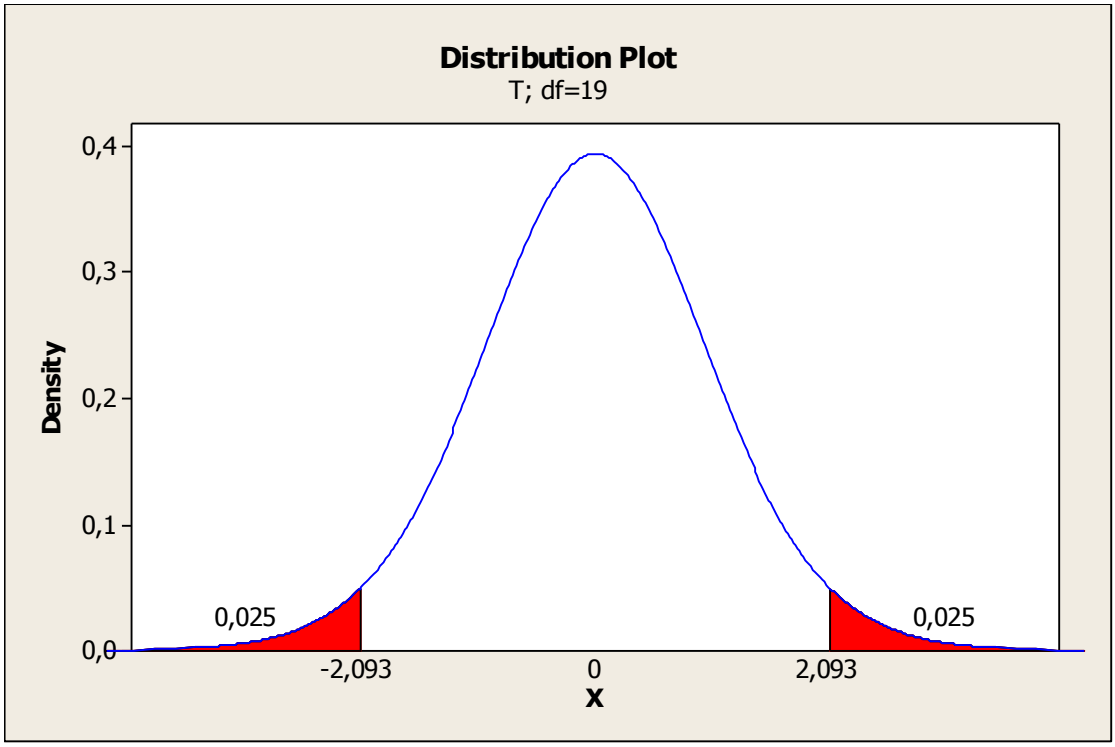
$$\mu = \bar{X} \pm t \frac{\hat{S}}{\sqrt{n}} \quad .$$

Για 19 βαθμούς ελευθερίας το $t_{0.975}$= 2.093 και

$$\mu = 23.5 \pm 2.093 \frac{10.4}{\sqrt{20}}$$

$$= 23.5 \pm 2.093 \text{ x } (2.326)$$

$$= 23.5 \pm 4.868 \rightarrow (18.63, 28.37)$$



2.5.2 ΣΤΑΤΙΣΤΙΚΕΣ ΥΠΟΘΕΣΕΙΣ ΚΑΙ ΔΙΑΔΙΚΑΣΙΑ ΕΛΕΓΧΟΥ

Στη διαδικασία λήψης αποφάσεων κάνουμε συχνά υποθέσεις για τους πληθυσμούς, για τις παραμέτρους των πληθυσμών. Μία τέτοια υπόθεση καλείται *στατιστική υπόθεση* (statistical hypothesis). Στην ουσία η στατιστική υπόθεση είναι μία δήλωση για μια ή περισσότερες παραμέτρους του πληθυσμού, η οποία ελέγχεται με βάση τις παρατηρήσεις ενός ή περισσότερων, αντίστοιχα, τυχαίων δειγμάτων.

Πολλές φορές κάνουμε μία υπόθεση για να την απορρίψουμε. Παρακάτω θα παρουσιάσουμε τη συλλογιστική, με την οποία ελέγχεται μια στατιστική υπόθεση. Για το σκοπό αυτό εξειδικεύουμε δύο αντικρουόμενες υποθέσεις:

- Τη μηδενική υπόθεση (null hypothesis), η οποία είναι αυτή που περιγράφει την υπάρχουσα κατάσταση (status quo) και συμβολίζεται με H_0. Η μηδενική υπόθεση είναι αυτό που περιμένουμε. (Πχ εάν θέλουμε να αποφασίσουμε, αν ένα νόμισμα

είναι κανονικό, τότε δεχόμαστε ότι η πιθανότητα p να έλθει κεφάλι κατά τη ρίψη του είναι 0,5.)

- Την εναλλακτική υπόθεση (alternative hypothesis), η οποία είναι αντίθετη, ασυμβίβαστη της μηδενικής και η οποία υπαγορεύει αλλαγή ενεργειών. Η εναλλακτική υπόθεση συμβολίζεται με H_1. (Πχ η πιθανότητα κατά τη ρίψη ενός νομίσματος να έλθει κεφάλι είναι p=0,7, ή p#0,5, ή p>0,5.)
- Οι δύο υποθέσεις είναι ξένες δηλαδή αμοιβαία αποκλειόμενες.

<u>Διαδικασία Ελέγχου μιας Στατιστικής Υπόθεσης</u>

1. Προσδιορίζουμε τις υποθέσεις, οι οποίες γίνονται δεκτές.
2. Εξειδικεύουμε την μηδενική και την εναλλακτική υπόθεση, έτσι ώστε οι δύο υποθέσεις να διαμερίζουν το σύνολο των δυνατών περιπτώσεων σε δύο υποσύνολα ξένα μεταξύ τους. Έτσι, αν δεν μπορούμε να δεχτούμε την ισχύ της μιας, θα πρέπει να δεχτούμε την άλλη. Εναλλακτικά η απόρριψη της μιας συνεπάγεται την αποδοχή της άλλης.
3. Ορίζουμε το κριτήριο αποφάσεως για την αποδοχή, ή την απόρριψη της μηδενικής υπόθεσης, δηλαδή ορίζουμε κανόνες απόφασης. Για τον σκοπό επιλέγουμε:

   i. *Τη στατιστική του ελέγχου* (test statistic), της οποίας η κατανομή πιθανοτήτων προσδιορίζεται από τις υποθέσεις, που γίνονται δεκτές και τη μηδενική υπόθεση και

   ii. Το επίπεδο σημαντικότητας α του ελέγχου.
4. Παίρνουμε την απόφαση ως εξής: Αν η τιμή της στατιστικής του ελέγχου βρίσκεται στην περιοχή αποδοχής H_0, την δεχόμαστε, διαφορετικά, αν τα αποτελέσματα του δείγματος είναι ***σημαντικά*** διαφορετικά, τότε θα πρέπει να απορρίψουμε την μηδενική

υπόθεση προς χάριν της εναλλακτικής.

Ο έλεγχος μιας στατιστικής υπόθεσης είναι μια διαδικασία λήψης αποφάσεως, στην οποία υπάρχουν αρχικά δύο δυνατές επιλογές: η αποδοχή, ή η απόρριψη της H_0. Για κάθε μία από τις επιλογές αυτές υπάρχουν δύο δυνατές καταστάσεις της φύσης (πραγματικότητες): η H_0 είναι σωστή ή λανθασμένη. Υπάρχουν δηλαδή οι εξής 4 ξένες μεταξύ τους δυνατές περιπτώσεις:

I. Δεχόμαστε την H_0 και η H_0 είναι ορθή

II. Δεχόμαστε την H_0 και η H_0 είναι λανθασμένη

III. Απορρίπτουμε την H_0 και η H_0 να είναι ορθή

IV. Απορρίπτουμε την H_0 και η H_0 να είναι λανθασμένη

Επειδή η απόφαση που παίρνουμε, για το αν θα δεχθούμε ή θα απορρίψουμε τη μηδενική υπόθεση, βασίζεται αποκλειστικά σε μια τιμή μιας στατιστικής (η οποία είναι τυχαία μεταβλητή), κινδυνεύουμε να πάρουμε εσφαλμένη απόφαση. Και αυτό μπορεί να συμβεί στις περιπτώσεις II. και III. Ειδικότερα:

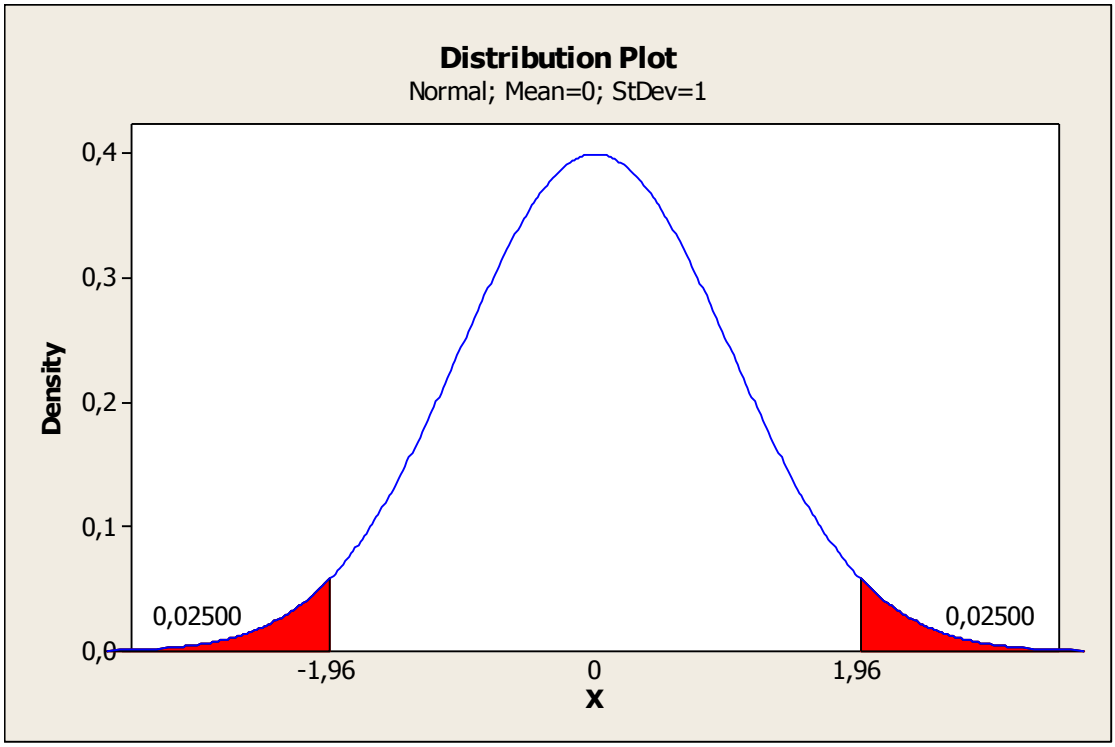
- Όταν απορρίπτουμε την H_0, ενώ είναι σωστή, τότε έχουμε σφάλμα τύπου I. Η πιθανότητα του σφάλματος αυτού ισούται με το επίπεδο σημαντικότητας α του ελέγχου.
- Όταν δεχόμαστε την H_0, ενώ είναι λανθασμένη, τότε έχουμε σφάλμα τύπου II. Η πιθανότητα του σφάλματος αυτού συμβολίζεται με β.

Οι δυνατές αυτές αποφάσεις μπορούν επίσης να παρασταθούν συνοπτικά με τον παρακάτω πίνακα:

Απόφαση	**Κατάσταση της φύσης (πραγματικότητα)**	
	H_0 είναι ορθή	H_0 είναι λανθασμένη

Δεχόμαστε την H_0	Σωστή απόφαση	Σφάλμα τύπου ΙΙ: πιθανότητα β
Απορρίπτουμε την H_0	Σφάλμα τύπου Ι: πιθανότητα α	Σωστή απόφαση

Αν για παράδειγμα έχουμε ένα κανονικό νόμισμα, τότε η πιθανότητα να έλθει κεφάλι είναι 0,5. Αυτό θα αντιπροσωπεύει τη μηδενική υπόθεση. Αν τώρα ρίξουμε το νόμισμα 20 φορές και έλθει κεφάλι τις 16 φορές, τότε θα πρέπει να απορριφθεί η μηδενική υπόθεση.



Στο παραπάνω σχήμα τα όρια ισχύος της μηδενική υπόθεσης καθορίζονται από το άσπρο εμβαδόν, που αντιπροσωπεύει το 95% των περιπτώσεων, ή το επίπεδο σημαντικότητας (κόκκινο εμβαδόν) που είναι α = 1-95% = 0,05. Αν τώρα πάρουμε μία τιμή μεγαλύτερη από 1,96 ή μικρότερη του -1,96 για την τυποποιημένη μεταβλητή, δηλαδή στις κόκκινες περιοχές, τότε λέμε ότι η τιμή της Z είναι σημαντικά διαφορετική, από αυτή που περιμέναμε, άρα θα πρέπει να απορρίψουμε

τη μηδενική υπόθεση. Η κόκκινη περιοχή μας δίνει την πιθανότητα να κάνουμε λάθος απορρίπτοντας την μηδενική υπόθεση. Στην περίπτωσή μας 5%. Με απλά λόγια η συλλογιστική αυτή μας λέει τα εξής: *Επειδή η πιθανότητα να συμβεί ένα γεγονός (στις κόκκινες περιοχές) είναι πάρα πολύ μικρή σε σχέση με αυτό που περιμένουμε και παρόλα αυτά συμβαίνει, τότε είναι σημαντικά διαφορετικό από αυτό που περιμένουμε.* Οπότε απορρίπτουμε τη μηδενική υπόθεση προς χάριν της εναλλακτικής.

Αν λοιπόν για έναν έλεγχο πάρουμε επίπεδο σημαντικότητας 5%, τότε αν απορρίψουμε τη μηδενική υπόθεση, θα είμαστε 95% βέβαιοι, ότι έχουμε πάρει τη σωστή απόφαση και μόνο 5%, ότι έχουμε πάρει λάθος απόφαση. Η περιοχή έξω από το διάστημα εμπιστοσύνης ισχύος της μηδενικής υπόθεσης λέγεται και περιοχή σημαντικότητας.

2.5.3 ΕΛΕΓΧΟΙ

Ένα πρόβλημα ελέγχου ονομάζεται *δίπλευρο* (two – sided), αν οι αποκλίσεις προς οποιαδήποτε κατεύθυνση από τη μηδενική υπόθεση επιβάλλουν διαφορετική ενέργεια, από αυτήν που υπαγορεύεται από την μηδενική υπόθεση.

Ένα πρόβλημα ελέγχου ονομάζεται *μονόπλευρο* (one – sided), αν μόνον οι αποκλίσεις προς μια κατεύθυνση από τη μηδενική υπόθεση επιβάλλουν διαφορετική ενέργεια από ότι στην μηδενική.

Ως παράδειγμα αναφέρουμε το πρόβλημα του ελέγχου υπόθεσης για την μέση τιμή: παίρνουμε πρώτα μία απόφαση με βάση την εμπειρία μας, ότι η μηδενική υπόθεση είναι αληθής, ή στη θέση της μηδενικής υπόθεσης κάποια εναλλακτική υπόθεση ως αληθή. Δηλαδή, ελέγχουμε:

- H_0: $E(Y) = \mu_{Y,0}$ έναντι H_1: $E(Y) > \mu_{Y,0}$ (μονόπλευρη, >)

- H_0: $E(Y) = \mu_{Y,0}$ έναντι H_1: $E(Y) < \mu_{Y,0}$ (μονόπλευρη, <)

- H_0: $E(Y) = \mu_{Y,0}$ έναντι H_1: $E(Y) \# \mu_{Y,0}$ (δίπλευρη)

Οι κατανομές που παίρνουμε για τους ελέγχους είναι:

- Για μεγάλα δείγματα παίρνουμε την κανονική κατανομή, ή αν η δειγματοληπτική κατανομή αναφέρεται, ότι είναι η κανονική κατανομή με γνωστή διασπορά.
- Για μικρά δείγματα ή κανονικό πληθυσμό και άγνωστη διασπορά εν γένει την κατανομή t (συντελεστές παλινδρόμησης, αυτοσυσχέτισης).
- Για μικρά δείγματα και για ελέγχους λόγων διασπορών την κατανομή F (σημαντικότητα παλινδρόμησης).
- Για τον λόγο της εκτιμήτριας του σ, S και του σ, όπως επίσης και για την εκτίμηση του συντελεστή αυτοσυσχέτισης (ως ομάδα) παίρνουμε την χι τετράγωνο κατανομή.

Το επίπεδο σημαντικότητας μοιράζεται και στα δύο άκρα της κατανομής αν έχουμε δίπλευρο έλεγχο. Για μονόπλευρο έλεγχο πηγαίνει όλο στη μία πλευρά.

Παράδειγμα

Ο μάνατζερ μιας εταιρίας πιστωτικών καρτών ισχυρίζεται, ότι ο μέσος χρόνος για να λύσει κάποια προβλήματα είναι 30 μέρες. Για να δούμε αν ισχύει αυτό, συλλέγουμε ένα δείγμα από 15 προβλήματα. Ο μέσος χρόνος για να λυθούν τα 15 αυτά προβλήματα είναι 35.9 μέρες και η τυπική απόκλιση είναι 10.2 μέρες. Αν τα δεδομένα αυτά θεωρούνται ως ένα τυχαίο δείγμα από ένα κανονικό πληθυσμό, μήπως υπάρχουν αμφιβολίες σχετικά με τον ισχυρισμό του μάνατζερ;

Α. Υποθέσεις που γίνονται δεκτές: Κατανομή πληθυσμού κανονική με άγνωστη διασπορά. Η διασπορά του δείγματος είναι $(10.2)^2$

ημέρες και $\hat{S}=10.2$. Επειδή η διασπορά του πληθυσμού δεν είναι γνωστή θα χρησιμοποιήσουμε την στατιστική t του Student: με n-1 = 15-1 = 14 βαθμούς ελευθερίας:

$$t=\frac{\bar{X}-\mu}{\hat{S}/\sqrt{n}}$$

Β. Η μηδενική και η εναλλακτική υπόθεση:

H_0: μ = 30

H_1: μ # 30

Γ. Το κριτήριο αποφάσεως: Για α = 0.05 και δίπλευρο έλεγχο έχουμε:

$$t_{1-\frac{\alpha}{2}}=t_{0.975}=2.145 \quad \text{και} \quad t_{\frac{\alpha}{2}}=t_{0.025}=-2.145$$

Επομένως αν

$$t=\frac{\bar{X}-30}{10.2/\sqrt{15}}<-2.145 \quad \acute{\eta} \quad >2.145,$$

θα απορρίψουμε την H_0.

Δ. Η απόφαση: Επειδή όμως

$$t=\frac{35.9-30}{10.2/\sqrt{15}}=2.240> 2.145$$

είναι μεγαλύτερο του άνω ορίου ισχύος της μηδενικής υπόθεσης απορρίπτουμε την H_0.

Παράδειγμα

Ο αριθμός Χ των αμειβομένων υπερωριών, που κάνουν ετησίως οι εργαζόμενοι σε μεγάλη δημόσια επιχείρηση, προσεγγίζει πολύ καλά την N(43,100). Η διοίκηση της επιχείρησης έδωσε εντολή για μείωση υπερωριών. Αν σε τυχαίο δείγμα n = 20 εργαζομένων ο μέσος αριθμός των υπερωριών, που έκαναν τον επόμενο χρόνο, ισούται με $\bar{x}$ = 32 ώρες

να ελεγχθεί η υπόθεση, ότι δεν μειώθηκε ο μέσος αριθμός των υπερωριών στο σύνολο των εργαζομένων. Η διοίκηση δέχεται πιθανότητα 0.10 να σφάλλει ισχυριζόμενη, ότι μειώθηκαν οι υπερωρίες, ενώ στην πραγματικότητα έμειναν οι ίδιες. Επομένως, το επίπεδο σημαντικότητας του ελέγχου είναι α = 0.10.

Α. Υποθέσεις που γίνονται δεκτές: Κατανομή πληθυσμού κανονική με γνωστή διασπορά $\sigma=\sqrt{100}=10,$ οπότε για το κριτήριο θα χρησιμοποιήσουμε την κανονική κατανομή.

Β. Η μηδενική και η εναλλακτική υπόθεση:

H_0: μ = 43

H_1: μ < 43

Γ. Το κριτήριο αποφάσεως: Για α = 0.10 έχουμε:

$$z_{1-\alpha}=z_{0.90}=1.28 \quad \text{και} \quad z_{\alpha}=z_{0.10}=-1.28.$$

Επομένως αν

$$Z=\frac{\bar{X}-43}{10/\sqrt{20}}<-1.28,$$

θα απορρίψουμε την H_0.

Δ. Η απόφαση: Επειδή όμως

$$Z=\frac{32-43}{10/\sqrt{20}}=-4.9<-1.28,$$

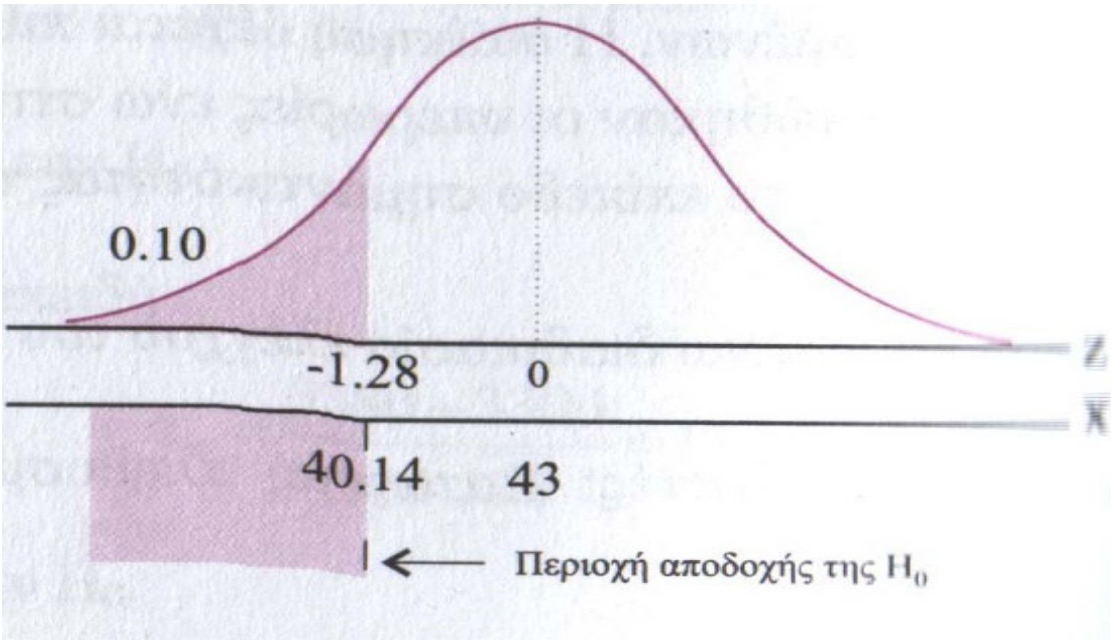
απορρίπτουμε την H_0.

Η τιμή πιθανότητας (P-Value) του ελέγχου ισούται με:

$$P-\text{τιμή}=P(Z<-4.9)=P(Z>4.9)=0.00002$$

Επειδή λοιπόν η πιθανότητα εμφάνισης αυτής της ακραίας τιμής είναι 0.00002 και παρόλα αυτά εμφανίστηκε, απορρίπτουμε τη μηδενική υπόθεση. Οι περιοχές απόρριψης και αποδοχής δίνονται στο σχήμα που

ακολουθεί:



2.5.4 Η P – ΤΙΜΗ ΤΟΥ ΕΛΕΓΧΟΥ

Μετά την παρατήρηση των στοιχείων του δείγματος μπορούμε να προσδιορίσουμε το ελάχιστο επίπεδο σημαντικότητας, στο οποίο μπορεί να απορριφθεί η μηδενική υπόθεση και το οποίο ονομάζεται *p-τιμή* (p-value, prob value) (Minitab → p-value) ή *πιθανότητα σημαντικότητας* (significance probability) (SPSS → sig) του ελέγχου. Όταν απορρίπτουμε την μηδενική υπόθεση, λέμε ότι η διαφορά ανάμεσα στην παρατηρούμενη μέση τιμή του δείγματος και την μ_0 είναι στατιστικά σημαντική και με πιθανότητα σφάλματος α και δεν μπορεί να αποδοθεί στις διακυμάνσεις της δειγματοληψίας. Όταν η p-τιμή του ελέγχου είναι μικρότερη από α, έχουμε μεγαλύτερη εμπιστοσύνη για την στατιστική σημαντικότητα της παρατηρούμενης διαφοράς και συνεπώς για την ισχύ της H_1.

Παράδειγμα

Αυτόματο μηχάνημα κόβει μεταλλικά ελάσματα, των οποίων το μήκος X ακολουθεί την κανονική κατανομή με μ = 5 cm. Ένα περίπου

24ωρο πριν την εμφάνιση της βλάβης, το μήκος των ελασμάτων μειώνεται οπότε, εκτός από την τακτική συντήρηση του μηχανήματος, καθημερινά μετριέται το μήκος σε τυχαίο δείγμα n = 20 ελασμάτων. Αν στο δείγμα μιας μέρας υπολογίσαμε μέσο μήκος $\bar{x}$ = 4.5 cm και s = 1 cm, να ελεγχθεί η υπόθεση, ότι το μέσο μήκος στο σύνολο της παραγωγής δεν μειώθηκε σε επίπεδο σημαντικότητας α = 0.01.

Α. Υποθέσεις που γίνονται δεκτές: Η κατανομή του πληθυσμού είναι κανονική με άγνωστη διασπορά. Οπότε θα χρησιμοποιήσουμε την στατιστική t του Student: με n-1 = 20-1 = 19 βαθμούς ελευθερίας:

Β. Η μηδενική και η εναλλακτική υπόθεση:

$$H_0: \mu = 5$$

$$H_1: \mu < 5$$

Γ. Το κριτήριο αποφάσεως: Για ν = n-1 = 19 βαθμούς ελευθερίας και α = 0.01 βρίσκουμε από την πίνακα Β – 3, $t_{19\,(0.01)} = -2.529$.

Επομένως αν

$$t = \frac{\bar{X} - 5}{\hat{s}/\sqrt{20}} < -2.529,$$

θα απορρίψουμε την H_0, υπέρ της H_1.

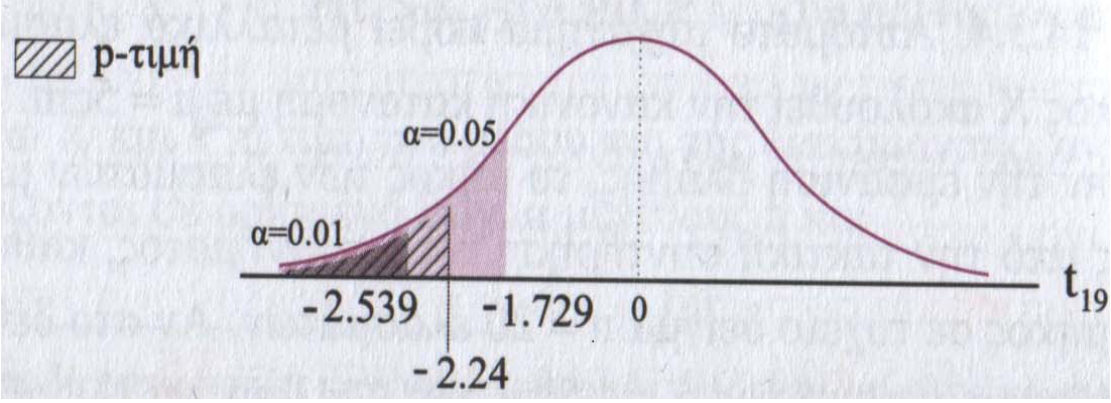
Δ. Η απόφαση: Επειδή όμως

$$t = \frac{4.5 - 5}{1/\sqrt{20}} = -2.24 > -2.529,$$

δεν μπορούμε να απορρίψουμε την H_0.

Η μηδενική υπόθεση όμως απορρίπτεται για μεγαλύτερα επίπεδα σημαντικότητας π.χ. για α = 0.05. Πράγματι, $t_{19\,(0.05)} = -1.729$ και ισχύει t < -1.729.

Στο σχήμα που ακολουθεί βλέπουμε, ότι η τιμή πιθανότητας του ελέγχου είναι μεταξύ του 1% και 5 %.



2.6 ΣΥΣΧΕΤΙΣΗ

Ο *συντελεστής συσχέτισης* (correlation coefficient) μετρά το βαθμό, κατά τον οποίο δύο μεταβλητές σχετίζονται **γραμμικά** μεταξύ τους. Δεν είναι τίποτε άλλο παρά μόνο ένα μέτρο γραμμικής συμμεταβολής δύο τυχαίων μεταβλητών. Στην πράξη χρησιμοποιούμε τον δειγματικό συντελεστή συσχέτισης:

$$r = \frac{s_{XY}}{s_X s_Y}, \quad \text{με}$$

$$s_X{}^2 = \frac{\sum (X - \bar{X})^2}{n} \qquad s_Y{}^2 = \frac{\sum (Y - \bar{Y})^2}{n} \qquad s_{XY}^2 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}$$

$$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{n \sum X^2 - (\sum X)^2}\sqrt{n \sum Y^2 - (\sum Y)^2}}$$

και ο τελεστής Σ δείχνει το άθροισμα για τιμές του i από 1 ως n. Ο συντελεστής r παίρνει τιμές στο κλειστό διάστημα [-1,1], δηλαδή ισχύει: $-1 \leq r \leq 1$.

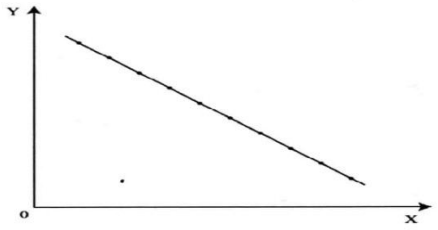
Οι ακραίες τιμές -1 και 1 αντιστοιχούν στην περίπτωση, όπου όλα τα σημεία (X_i,Y_i), i = 1,...,n βρίσκονται επάνω σε μια ευθεία γραμμή με αρνητική ή θετική κλίση, αντίστοιχα. Όταν οι X και Y είναι παρατηρήσεις τυχαίων μεταβλητών, η πιθανότητα να υπολογίσουμε

συντελεστή συσχέτισης ίσο με 1 ή -1 είναι μηδενική. Οι ακραίες τιμές όμως χρησιμεύουν για να ερμηνεύσουμε τις ενδιάμεσες. Όσο πιο κοντά σε μια από αυτές βρίσκεται ο r, τόσο πιο έντονη η γραμμική συσχέτιση των παρατηρήσεων X και Y. Μια τιμή για τον r ίση ή πολύ κοντά στο μηδέν, δηλώνει απουσία γραμμικής σχέσης, αλλά όχι απουσία οποιαδήποτε σχέσης. Το διάγραμμα διασποράς των παρατηρήσεων (X_i,Y_i) μπορεί να είναι διαφωτιστικό στην περίπτωση αυτή.

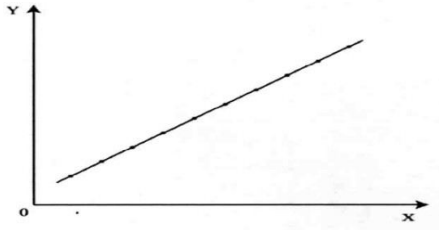
Η γραμμική συσχέτιση μεταξύ δύο μεταβλητών μπορεί να είναι θετική, αρνητική ή οι μεταβλητές να μη συσχετίζονται καθόλου. Θετική συσχέτιση υπάρχει όταν δύο μεταβλητές τείνουν να μεταβάλλονται προς την ίδια κατεύθυνση. Στην περίπτωση αυτή οι τιμές και των δύο μεταβλητών τείνουν να αυξάνονται ή να μειώνονται. Αρνητική συσχέτιση υφίσταται, όταν δύο μεταβλητές τείνουν να μεταβάλλονται προς αντίθετη κατεύθυνση. Στην περίπτωση αυτή όταν οι τιμές της μιας μεταβλητής αυξάνονται οι τιμές της άλλης μειώνονται και αντίστροφα.

Μηδενική συσχέτιση παρατηρείται, όταν οι μεταβολές των τιμών της μιας μεταβλητής δεν συνοδεύονται καθόλου με τις μεταβολές των τιμών της άλλης, ή με άλλα λόγια όταν οι μεταβλητές είναι ανεξάρτητες μεταξύ τους $\sigma_{XY}=0$. Ισχύει:

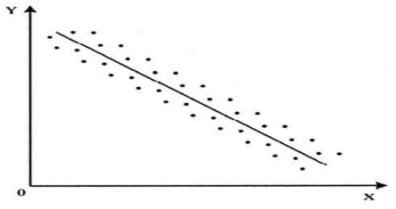
- Εάν r=−1 ή r = 1 έχουμε την μέγιστη δυνατή συσχέτιση.
- Εάν r>0 υπάρχει θετική συσχέτιση, η όποια είναι τόσο πιο ισχυρή, όσο πιο κοντά ο συντελεστής συσχέτισης είναι στο 1.
- Εάν r<0 υπάρχει αρνητική συσχέτιση, η όποια είναι τόσο πιο ισχυρή, όσο πιο κοντά ο συντελεστής συσχέτισης είναι στο –1.
- Για r=0 δεν υπάρχει καμία συσχέτιση μεταξύ των δυο μεταβλητών.
- Όλα αυτά ισχύουν για γραμμική συσχέτιση. Μπορεί όμως να υπάρχει μη γραμμική συσχέτιση και το r να είναι ίσο με μηδέν.



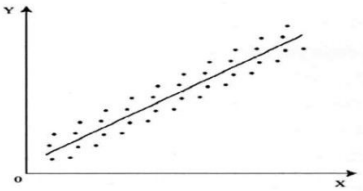
α. Πλήρης γραμμική αρνητική συσχέτιση, r = - 1



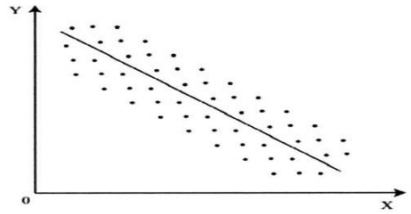
β. Πλήρης γραμμική θετική συσχέτιση, r = + 1



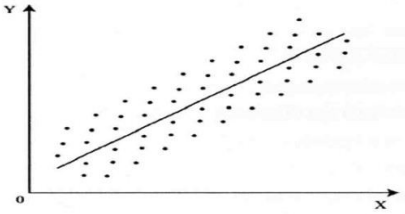
γ. Έντονη γραμμική αρνητική συσχέτιση, r = -0,8



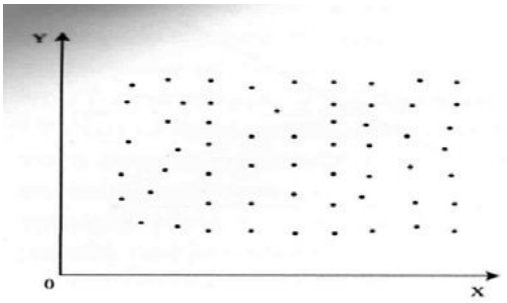
δ. Έντονη γραμμική θετική συσχέτιση, r = +0,8



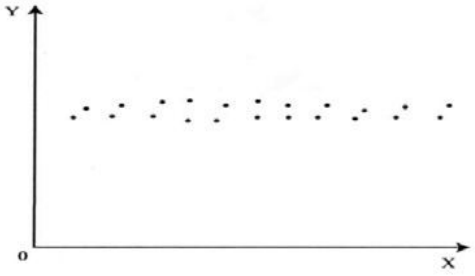
ε. Ασθενής γραμμική αρνητική συσχέτιση, r = - 0,6



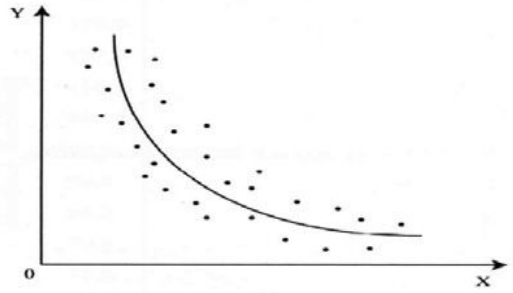
στ. Ασθενής γραμμική θετική συσχέτιση, r = + 0,6



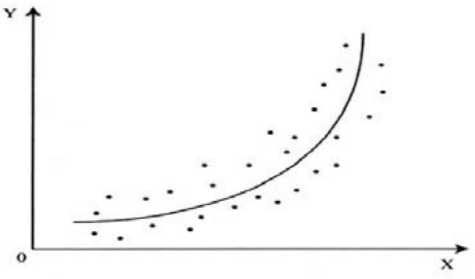
α. Μηδενική συσχέτιση, r = o



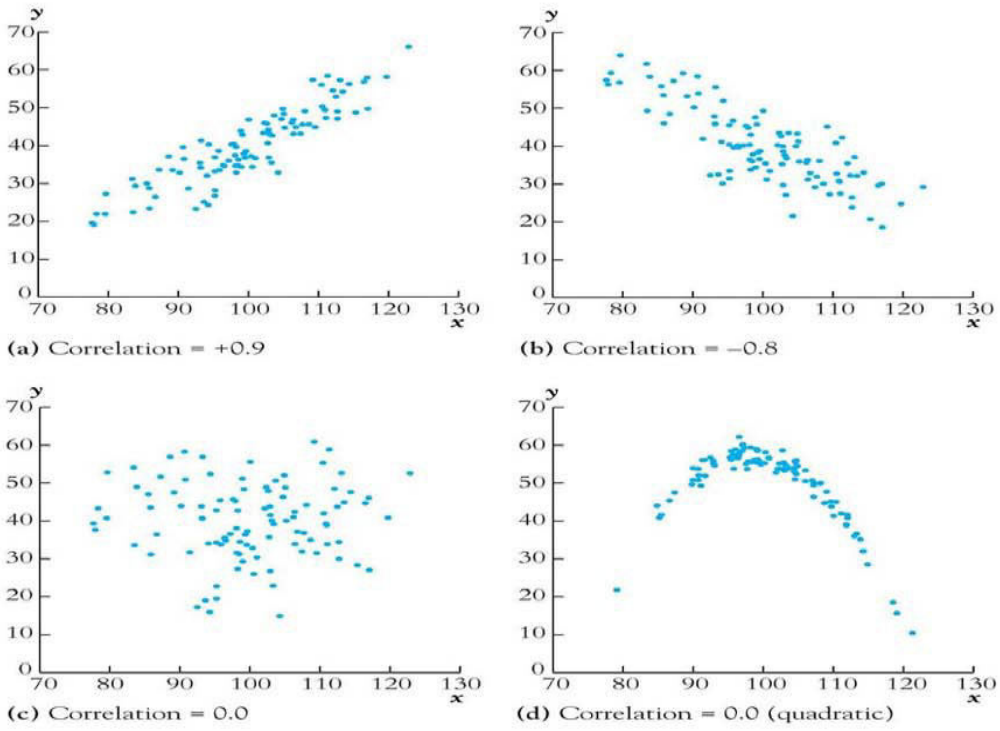
β. Μηδενική συσχέτιση, r = 0



γ. Αρνητική μη γραμμική συσχέτιση



δ. Θετική μη γραμμική συσχέτιση



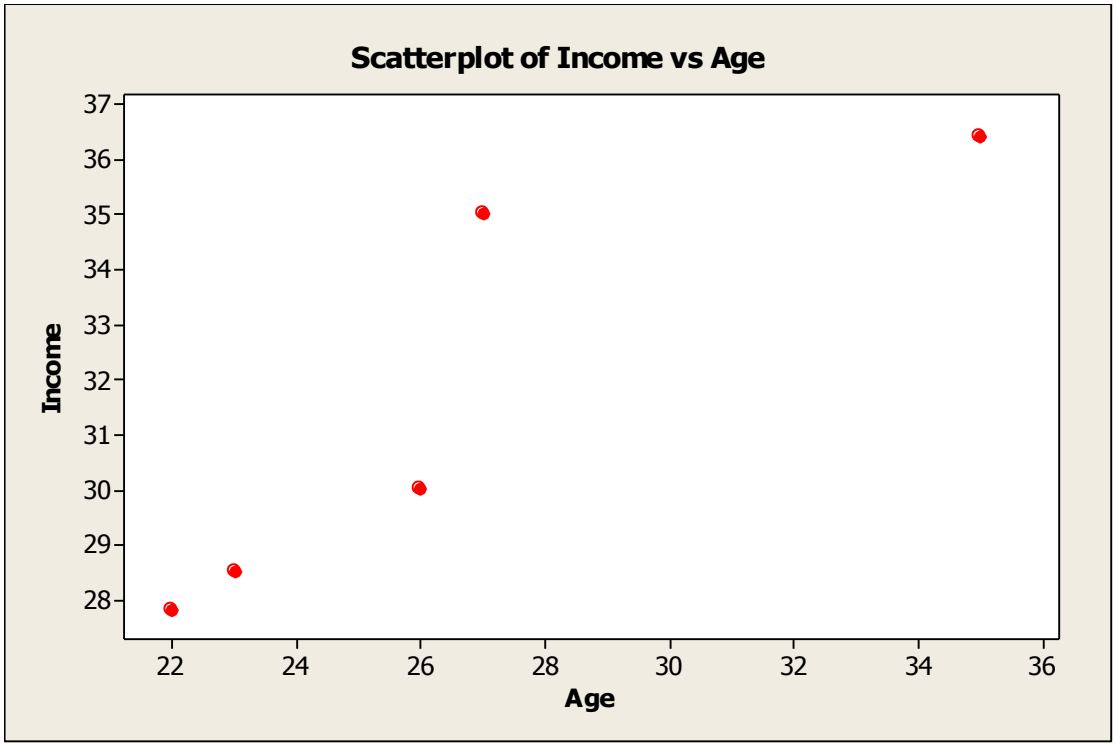
Παράδειγμα Τ2.7

Θα υπολογίσουμε την σχέση της ηλικίας και του εισοδήματος.

ΑΤΟΜΑ	Y	X	Y^2	X^2	XY
1	27.80	22	772.84	484	611.6
2	28.50	23	812.25	529	655.5
3	30.00	26	900.00	676	780.0
4	35.00	27	1,225.00	729	945.0
5	36.40	35	1,324.96	1,225	1,274.0
ΣΥΝΟΛΟ	157.7	133	5,035.05	3,643	4,266.1

$$r=\frac{n\sum XY-\sum X\sum Y}{\sqrt{n\sum X^2-\left(\sum X\right)^2}\sqrt{n\sum Y^2-\left(\sum Y\right)^2}}=¿$$

$$\frac{5*4{,}266.1-133*157.7}{\sqrt{5*3{,}643-133^2}\sqrt{5*5{,}035.05-157.7^2}}=0.89$$



2.7 ΠΑΛΙΝΔΡΟΜΗΣΗ

2.7.1 Η ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Με την ανάλυση της παλινδρόμησης (Regression Analysis) επιδιώκουμε να βρούμε ένα μοντέλο ή υπόδειγμα, δηλαδή μια μαθηματική σχέση, η οποία εκφράζει την εξάρτηση μιας μεταβλητής Y από μία άλλη X. (Για παράδειγμα με βάση το μοντέλο αυτό θα μπορούμε να προσδιορίσουμε το μηνιαίο αριθμό Y των πωλήσεων μιας εταιρείας, όταν δαπανά X νομισματικές μονάδες για διαφήμιση).

Τα ζεύγη (X_i,Y_i) ονομάζονται *σημεία των δεδομένων* (data points). Η γραφική παράσταση των σημείων (X,Y) στο ορθογώνιο σύστημα καρτεσιανών συντεταγμένων ονομάζεται *διάγραμμα διασποράς* (scatter diagram) και μας δίνει πολύτιμες πληροφορίες για το είδος της προσεγγιστικής σχέσης, που συνδέει τα X και Y των δεδομένων μας. Τα σημεία τείνουν να συγκεντρώνονται γύρω από μια ευθεία γραμμή ή μία

καμπύλη. Στην πρώτη περίπτωση μπορούμε επομένως να χρησιμοποιήσουμε μια εξίσωση ευθείας για να περιγράψουμε περιληπτικά την παρατηρούμενη σχέση. Η διαδικασία αυτή ονομάζεται *προσαρμογή μιας ευθείας στα δεδομένα* (fitting a line to data).

Με τη φράση ανάλυση παλινδρόμησης εννοούμε διάφορες γραφικές και αναλυτικές μεθόδους, που σκοπό έχουν την αναζήτηση σχέσεων μεταξύ μιας μεταβλητής, την οποία ονομάζουμε *εξαρτημένη μεταβλητή* (dependent variable) και μιας άλλης ή άλλων μεταβλητών, τις οποίες ονομάζουμε *ανεξάρτητες μεταβλητές* (independent variables). Όταν μια τέτοια σχέση (μοντέλο) βρεθεί, τότε μπορούμε να χρησιμοποιήσουμε το μοντέλο αυτό για να κάνουμε προβλέψεις, για να βρούμε, ποιες από τις ανεξάρτητες μεταβλητές επηρεάζουν περισσότερο την εξαρτημένη μεταβλητή, ή να ελέγξουμε διάφορες υποθέσεις.

Στο σημείο αυτό θα πρέπει να τονίσουμε την διαφορά μεταξύ της ανάλυσης παλινδρόμησης (regression analysis) και της ανάλυσης συσχετίσεως (correlation analysis). Στην ανάλυση παλινδρόμησης η σχέση είναι προς μία κατεύθυνση μόνο. Δηλαδή αγνοεί την πιθανή επίδραση της εξαρτημένης μεταβλητής στις ανεξάρτητες μεταβλητές. Σε μερικές περιπτώσεις, κυρίως σε εργαστηριακά πειράματα, ο ερευνητής μπορεί να προκαθορίσει τις τιμές των ανεξάρτητων μεταβλητών και στη συνέχεια να παρατηρήσει το μέγεθος της εξαρτημένης μεταβλητής. Στην περίπτωση αυτή είναι φανερό, ότι η σχέση είναι προς μία μόνο κατεύθυνση. Σε πολλές περιπτώσεις όμως ο ερευνητής παρατηρεί συγχρόνως όλες τις μεταβλητές. Η ανάλυση παλινδρόμησης, για τέτοια δεδομένα, μπορεί και πάλι να χρησιμοποιηθεί, όταν ο σκοπός της μελέτης είναι η εύρεση της μεταβολής μιας εξ αυτών σε σχέση με τις υπόλοιπες.

Η ανάλυση συσχετίσεως χρησιμοποιείται, όταν κάποιος θέλει να

μελετήσει την ταυτόχρονη σχέση ενός συνόλου μεταβλητών. Στην ανάλυση συσχετίσεως οι σχέσεις μεταξύ των διάφορων μεταβλητών δεν είναι, γενικώς, μονοκατευθυντικές.

Ένα απλό παράδειγμα όπου η χρήση της ανάλυσης συσχετίσεως είναι αναγκαία είναι το εξής: Ας υποθέσουμε ότι κάποιος ερευνητής ενδιαφέρεται να μελετήσει την ταυτόχρονη μεταβολή (σαν συνάρτηση της ηλικίας) του ύψους και του βάρους κάποιου πληθυσμού. Ο σκοπός, στην περίπτωση αυτή, δεν είναι να περιγράψουμε, πως μεταβάλλεται το ύψος συναρτήσει του βάρους ή αντίστροφα, αλλά να μελετήσουμε την από κοινού μεταβολή του ύψους και του βάρους, στον εν λόγω πληθυσμό, από τα νήπια μέχρι τους ηλικιωμένους. Από την άλλη πλευρά, εάν θέλουμε να προβλέψουμε το ύψος από το βάρος, θα πρέπει να χρησιμοποιήσουμε μεθόδους ανάλυσης παλινδρόμησης.

2.7.2 ΠΡΟΣΑΡΜΟΓΗ ΜΙΑΣ ΕΥΘΕΙΑΣ ΣΤΑ ΔΕΔΟΜΕΝΑ

Σε ένα σμήνος σημείων μπορούν να προσαρμοστούν πολλές καμπύλες. Από όλες τις προσεγγιστικές καμπύλες για ένα δεδομένο σμήνος σημείων η καμπύλη με την ιδιότητα $d_1^2 + d_2^2 + ... + d_n^2 = ελάχιστο$ είναι η καμπύλη με την καλύτερη προσαρμογή. Τα d_i είναι οι αποστάσεις της προσαρμοσμένης καμπύλης από τα πραγματικά σημεία.

Η *μέθοδος των ελαχίστων τετραγώνων* (method of least squares) ορίζει ως κριτήριο καλής προσαρμογής την ελαχιστοποίηση του αθροίσματος των τετραγωνικών σφαλμάτων και είναι μια γενική μέθοδος εκτίμησης. Όταν εφαρμόζεται για την προσαρμογή μιας ευθείας σε n ζεύγη (X_i, Y_i) ορίζει την *ευθεία ελαχίστων τετραγώνων* (least squares line). Υποτίθεται ότι τα σφάλματα έχουν μέσο μηδέν και σταθερή διασπορά.

Η μέθοδος των ελαχίστων τετραγώνων υπολογίζει τις τιμές για την

κλίση και το σημείο τομής της ευθείας με τον άξονα Υ που ελαχιστοποιούν το άθροισμα των τετραγώνων των σφαλμάτων μεταξύ των τιμών της Υ και της ευθείας γραμμής. Το *άθροισμα των τετραγώνων των σφαλμάτων* (Sum of Squared Errors) έχει καθιερωθεί να συμβολίζεται διεθνώς με SSE από τα αρχικά των αντίστοιχων αγγλικών όρων και ορίζεται ως:

$$SSE = \sum (y - \hat{y})^2 = \sum (y - \alpha - \beta x)^2$$

$$\text{με} \quad \hat{y} = \alpha + \beta x$$

Για να είναι το SSE ελάχιστο, θα πρέπει $\frac{\partial SSE}{\partial \alpha} = \frac{\partial SSE}{\partial \beta} = 0.$ Μετά την παραγώγιση καταλήγουμε σε ένα σύστημα δύο εξισώσεων με δύο αγνώστους με την ακόλουθη λύση:

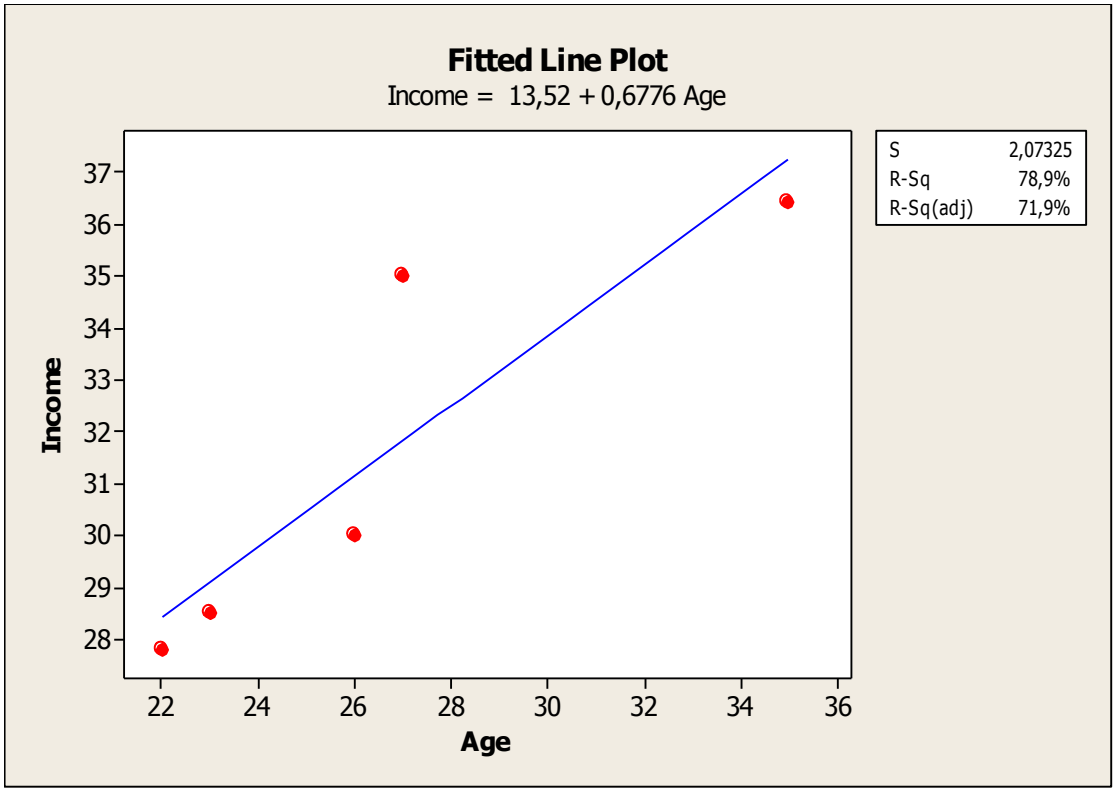
$$\alpha = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}$$

$$\beta = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \quad \text{ή}$$

$$\beta = \frac{\sqrt{\sum (y - \bar{y})^2}}{\sqrt{\sum (x - \bar{x})^2}} r$$

Το β είναι ανάλογο του συντελεστή συσχέτισης και το πρόσημο αυτού εξαρτάται από το πρόσημο του συντελεστή συσχέτισης, επειδή ο παράγοντας μπροστά είναι θετικός.

Εφαρμόζοντας στο προηγούμενο παράδειγμα βρίσκουμε α=13.52, β=0.678 και r=0.89.

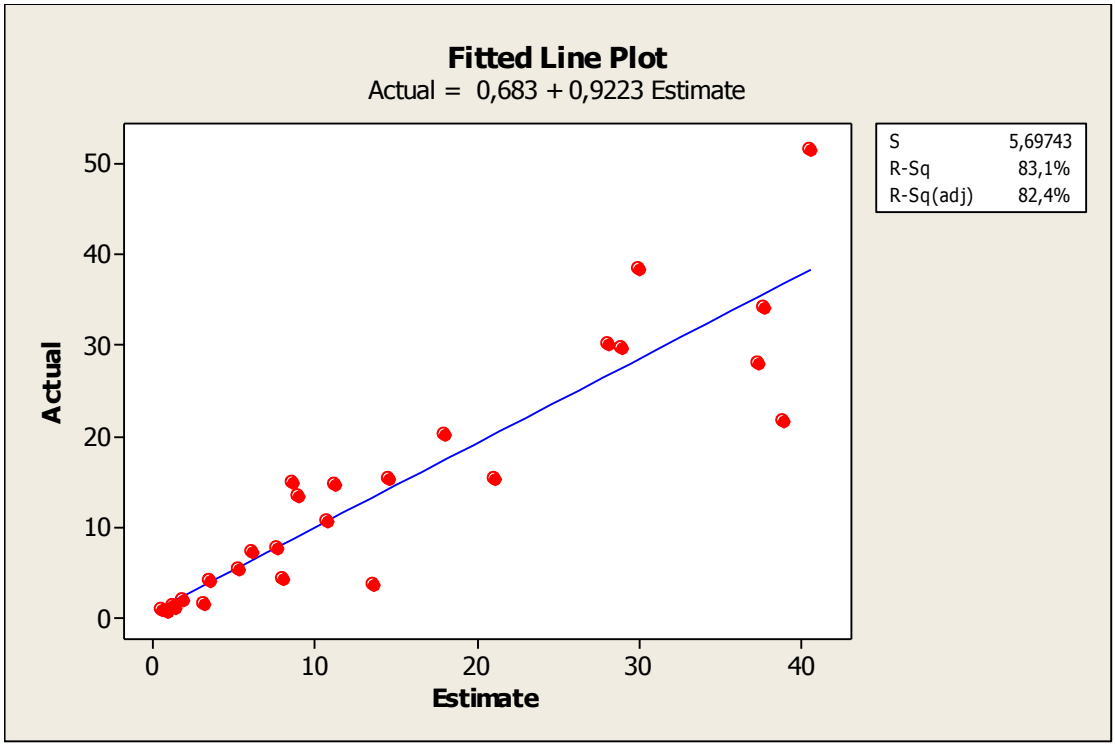


Παράδειγμα Τ2.8

Ο μάνατζερ μιας μεγάλης κατασκευαστικής εταιρείας υποψιάζεται, ότι οι εκτιμώμενες δαπάνες των έργων κατασκευής της εταιρίας δεν είναι πολύ κοντά στις πραγματικές δαπάνες. Τα δεδομένα που χρησιμοποιούνται έχουν συλλεχθεί από προηγούμενα έργα, που έχουν γίνει, έτσι ώστε να αναλυθεί η σχέση μεταξύ των εκτιμώμενων και των πραγματικών δαπανών.

Επειδή ο συντελεστής συσχέτισης είναι πολύ μεγάλος 0,912 και η κλίση της ευθείας 0,9223 μπορούμε να χρησιμοποιήσουμε τη σχέση της ευθείας των ελαχίστων τετραγώνων προκειμένου να κάνουμε προβλέψεις.

Στο παρακάτω διάγραμμα φαίνεται η προσαρμογή της ευθείας στα δεδομένα και η εξίσωσή της.



2.8 ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΕΚΤΙΜΗΣΗΣ ΕΥΘΕΙΑΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Τυπικό σφάλμα εκτίμησης (standard error of estimation) είναι ένα μέτρο του βαθμού διασποράς των τιμών γύρω από τη γραμμή παλινδρόμησης της $Y = a_0 + a_1 x$, ή αλλιώς της διασποράς των σφαλμάτων από την προσαρμοσμένη ευθεία και δίδεται από τη σχέση:

$$S_{y,x} = \sqrt{\frac{\sum (y_i - y_{es})^2}{n-2}}$$

Το σφάλμα αυτό έχει ανάλογες ιδιότητες με εκείνες της τυπικής απόκλισης. Συνεπώς όσο μεγαλύτερο είναι το μέγεθος του δείγματος, τόσο μικρότερο θα είναι το τυπικό σφάλμα εκτίμησης του μέσου και έτσι τόσο περισσότερο αξιόπιστη θα είναι η εκτίμηση του.

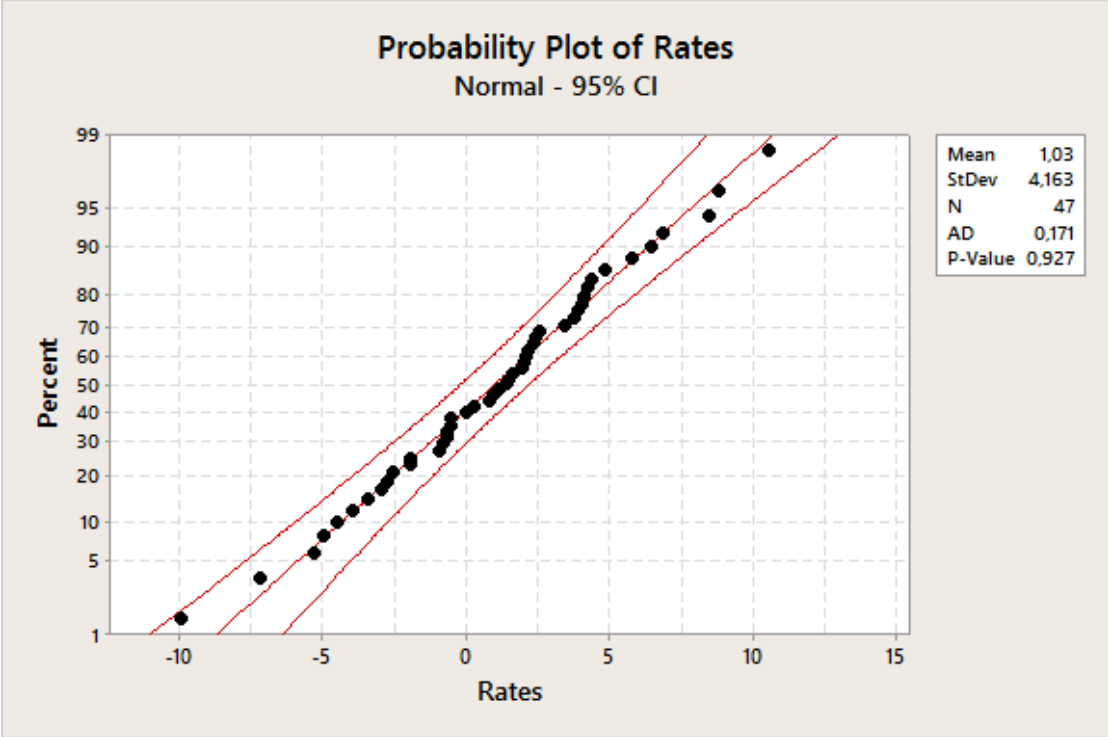
Επίσης επειδή το άθροισμα των τετραγώνων στον αριθμητή είναι το άθροισμα των αποστάσεων μεταξύ των προσαρμοσμένων και

πραγματικών τιμών, μόνο η καμπύλες ελαχίστων τετραγώνων ελαχιστοποιούν την τυπική απόκλιση για δεδομένο μέγεθος δείγματος.

2.9 PROBABILITY PLOT

Ένα άλλο διάγραμμα που χρησιμοποιείται, είναι το Probability Plot και ειδικότερα της κανονικής κατανομής, επειδή τα σφάλματα μέσω προσαρμογής από παλινδρόμηση πρέπει να είναι κανονικά κατανεμημένα.

Ο κάθετος άξονας αντιστοιχεί στα εκατοστιαία σημεία ενώ ο οριζόντιος άξονας οι τιμές του x, οι οποίες έχουν προηγουμένως "σορταριστεί". Πάνω στο διάγραμμα αποτυπώνονται η θεωρητική ευθεία της κανονικής κατανομής και τα σημεία του δείγματος μαζί με το διάστημα εμπιστοσύνης που πρέπει να τοποθετούνται τα σημεία για της ισχύ της κανονικότητας. Επίσης αναγράφεται η τιμή του Anderson-Darling Test και η p-τιμή της.

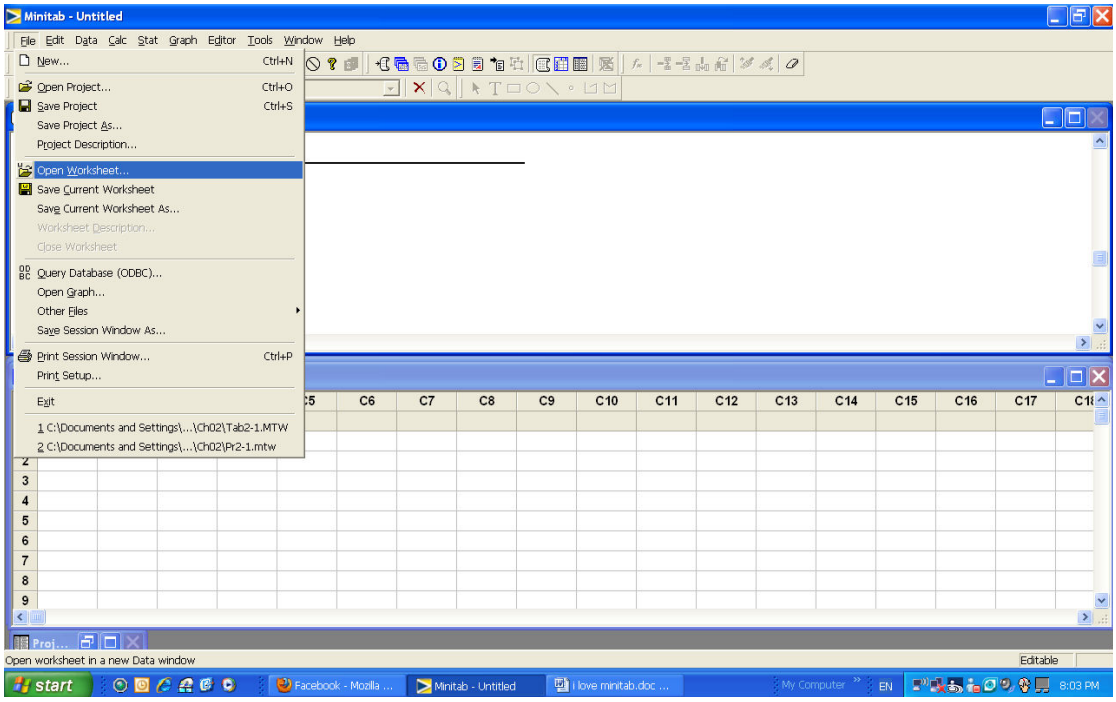


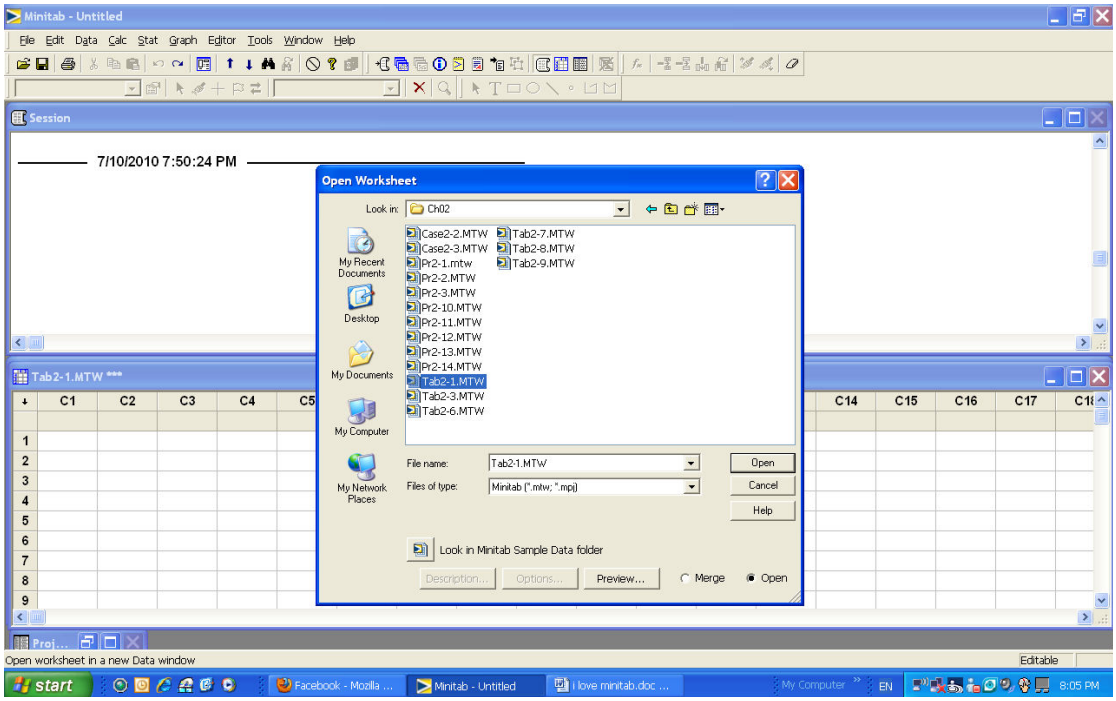
Στο παραπάνω παράδειγμα εξετάζεται η κανονικότητα των Monthly Rates of Return (%) του Standard & Poor' s 500 Stock Market Index.

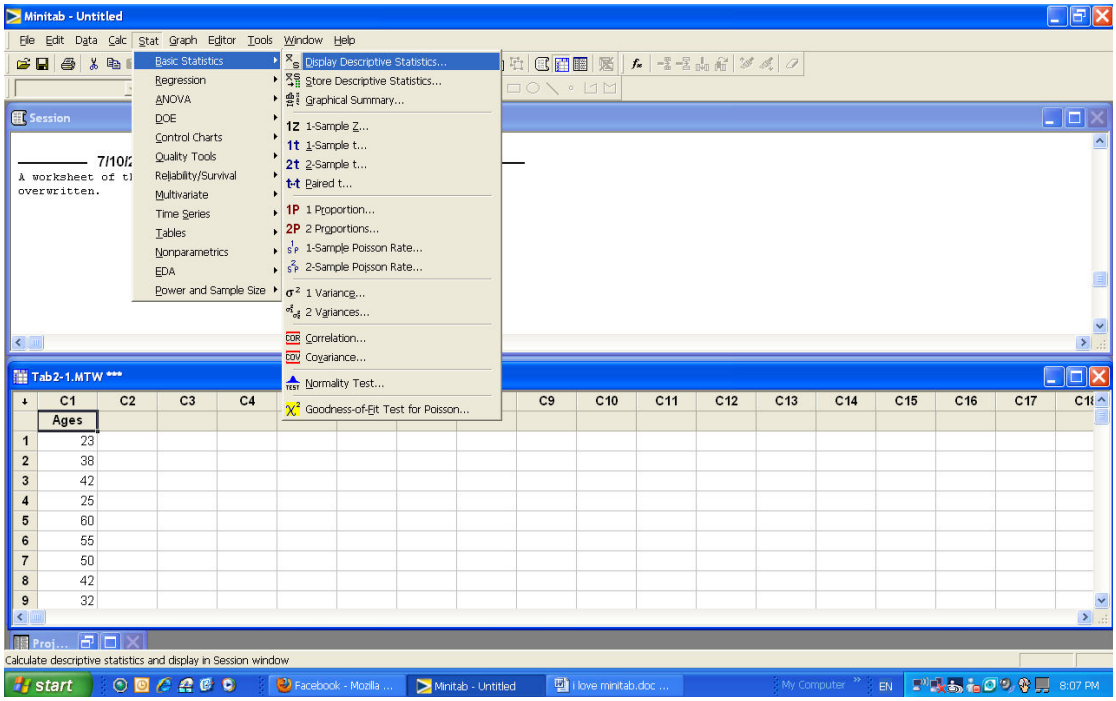
Για κάθε κατανομή υπάρχει ένα αντίστοιχο θεωρητικό Probability Plot.

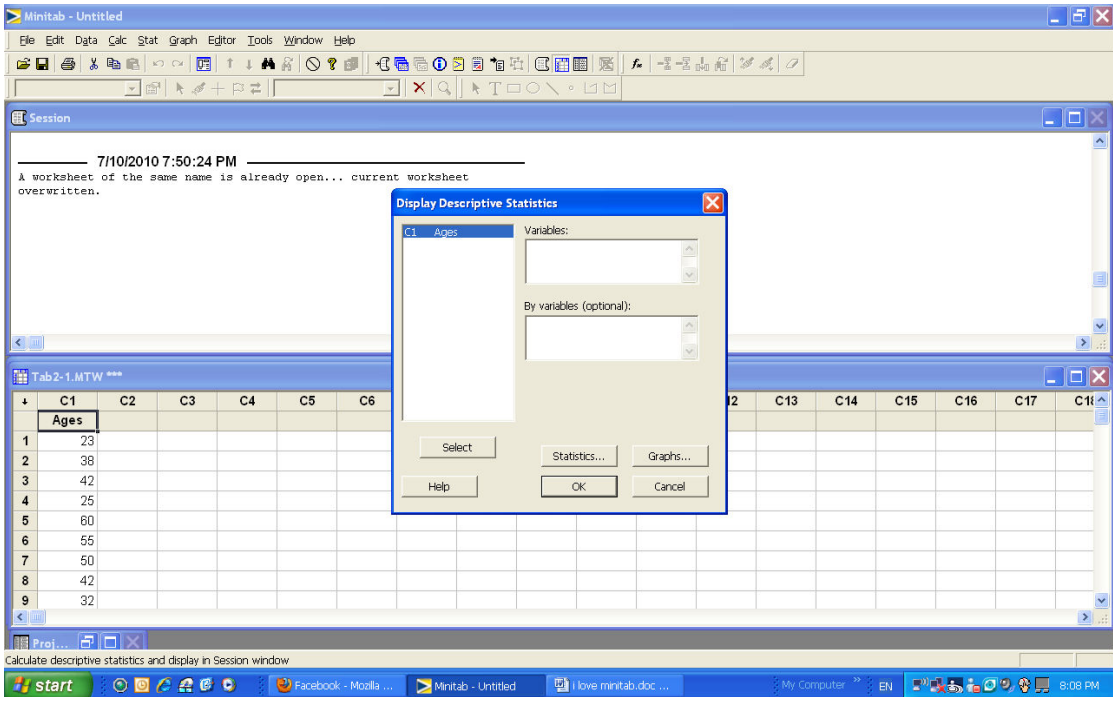
2.10 ΧΡΗΣΗ MINITAB

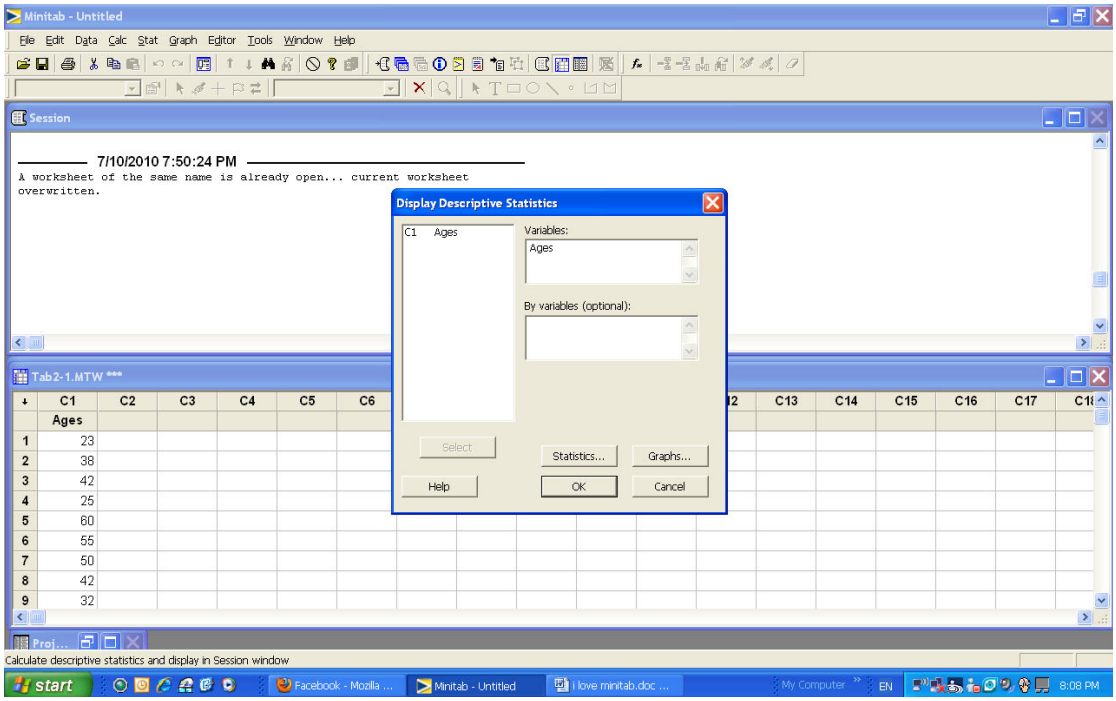
Υπολογισμός Περιγραφικών Στατιστικών Δεδομένων

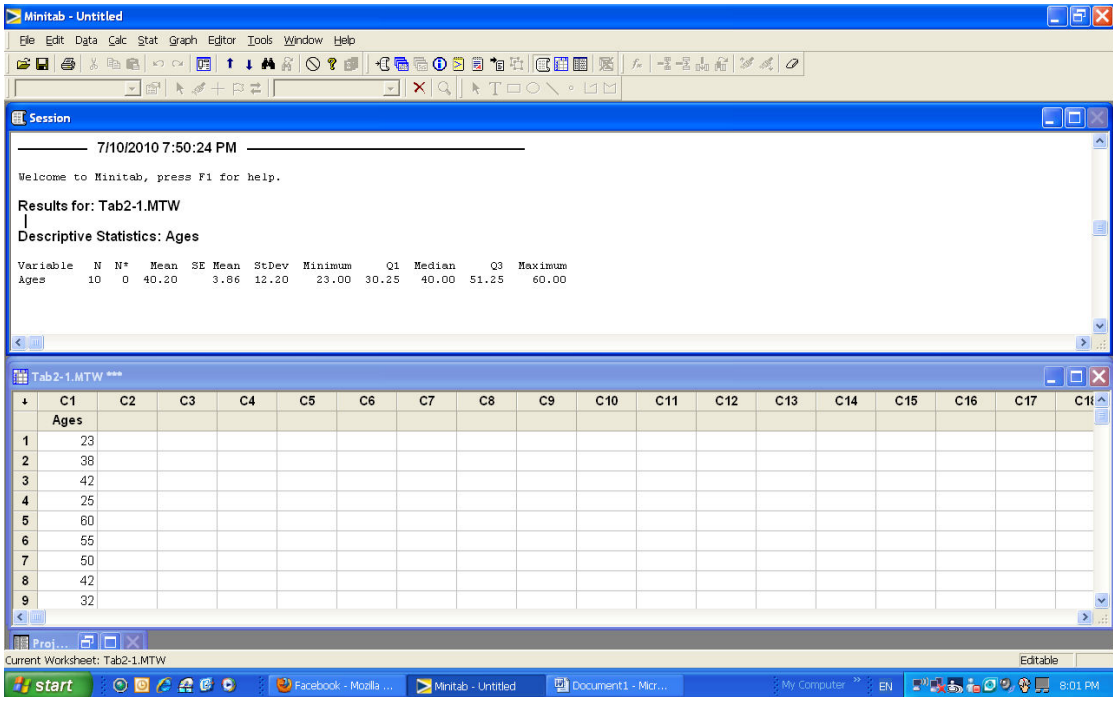



Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/10/2010 7:50:24 PM
Open Worksheet
Look in: Ch02
Case2-2.MTW
Case2-3.MTW
Pr2-1.mtw
Pr2-2.MTW
Pr2-3.MTW
Pr2-10.MTW
Pr2-11.MTW
Pr2-12.MTW
Pr2-13.MTW
Pr2-14.MTW
Tab2-1.MTW
Tab2-3.MTW
Tab2-6.MTW
Tab2-7.MTW
Tab2-8.MTW
Tab2-9.MTW
My Recent Documents
Desktop
My Documents
My Computer
My Network Places
File name: Tab2-1.MTW
Files of type: Minitab (*.mtw; *.mpj)
Open
Cancel
Help
Look in Minitab Sample Data folder
Description...
Options...
Preview...
Merge
Open
Tab2-1.MTW ***
Open worksheet in a new Data window
Editable
start
Facebook - Mozilla ...
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
8:05 PM

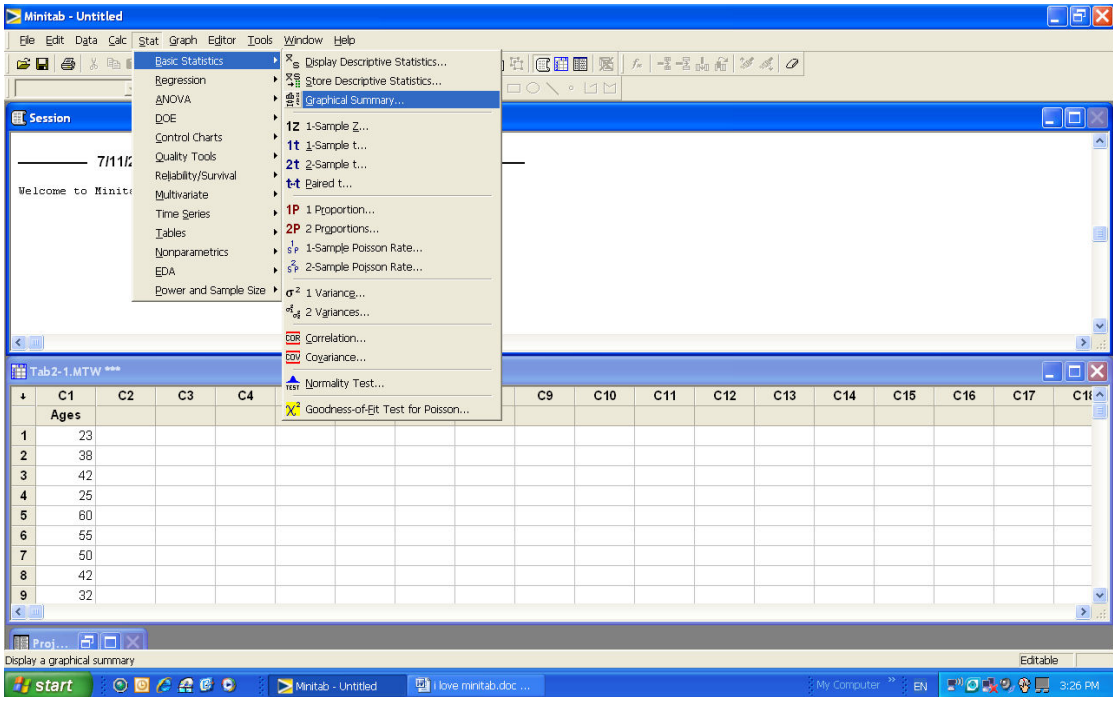

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Basic Statistics
Regression
ANOVA
DOE
Control Charts
Quality Tools
Reliability/Survival
Multivariate
Time Series
Tables
Nonparametrics
EDA
Power and Sample Size
Display Descriptive Statistics...
Store Descriptive Statistics...
Graphical Summary...
1-Sample Z...
1-Sample t...
2-Sample t...
Paired t...
1 Proportion...
2 Proportions...
1-Sample Poisson Rate...
2-Sample Poisson Rate...
1 Variance...
2 Variances...
Correlation...
Covariance...
Normality Test...
Goodness-of-Fit Test for Poisson...
Session
A worksheet of t
overwritten.
Tab2-1.MTW ***
C1 Ages
23
38
42
25
60
55
50
42
32
Calculate descriptive statistics and display in Session window
start
Facebook - Mozilla ...
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
8:07 PM

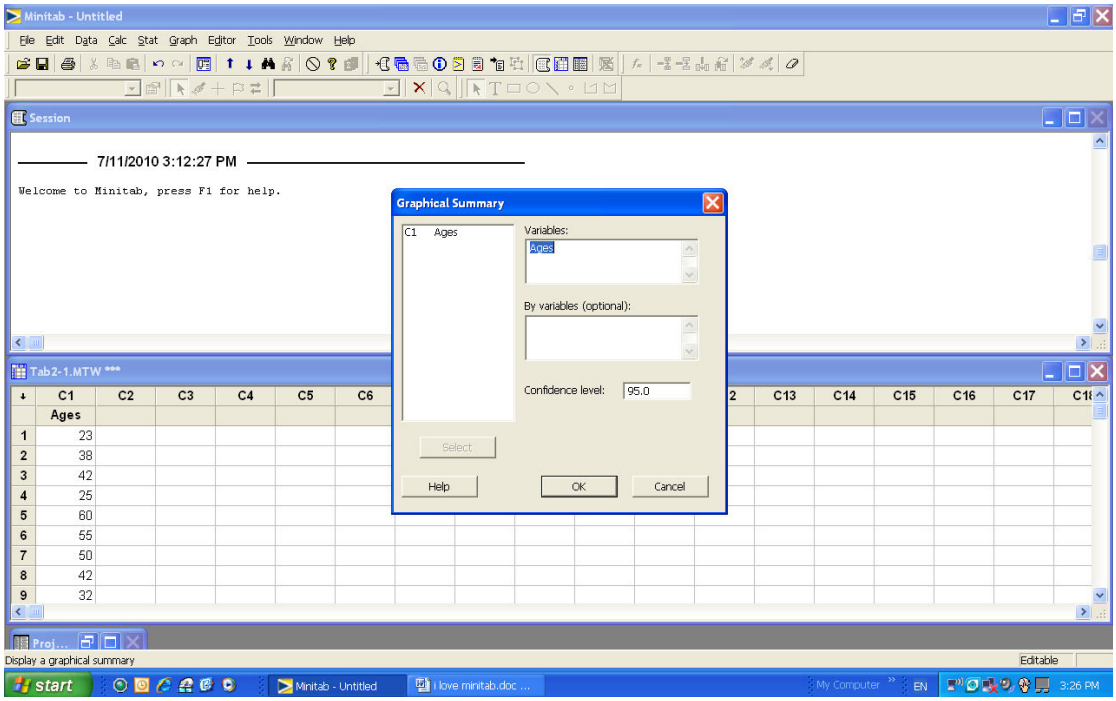

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/10/2010 7:50:24 PM
A worksheet of the same name is already open... current worksheet
overwritten.
Display Descriptive Statistics
C1 Ages
Variables:
By variables (optional):
Select
Statistics...
Graphs...
Help
OK
Cancel
Tab2-1.MTW ***
C1 C2 C3 C4 C5 C6 C13 C14 C15 C16 C17
Ages
1 23
2 38
3 42
4 25
5 60
6 55
7 50
8 42
9 32
Calculate descriptive statistics and display in Session window
start
Facebook - Mozilla ...
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
8:08 PM

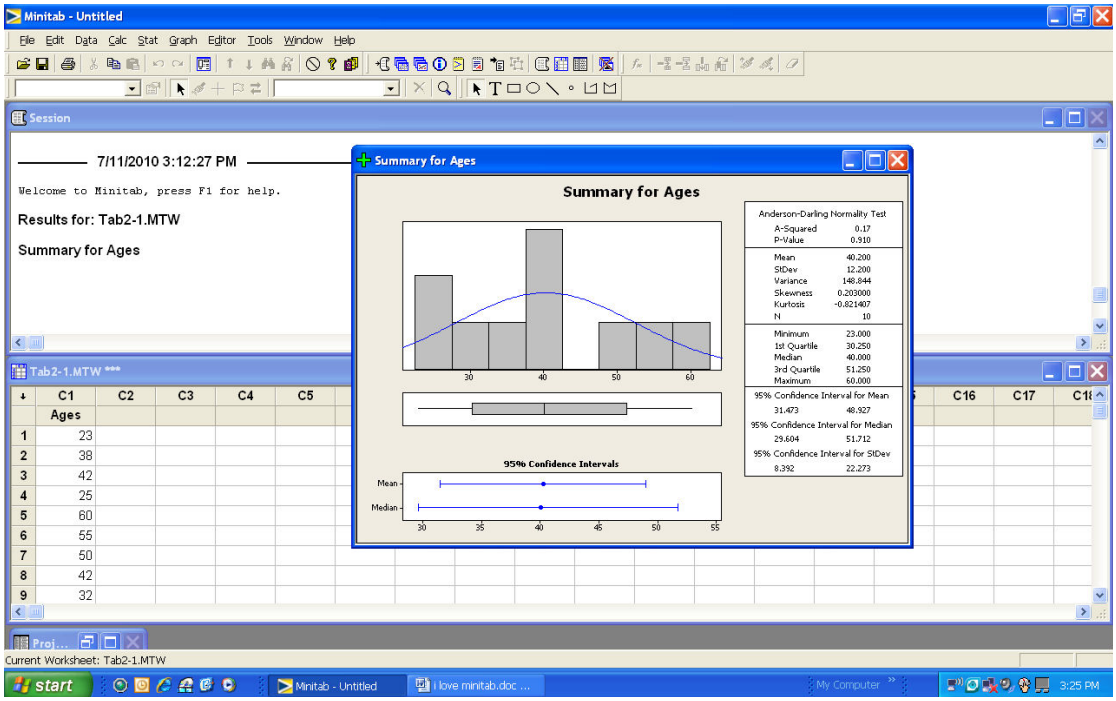

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/10/2010 7:50:24 PM
A worksheet of the same name is already open... current worksheet
overwritten.
Display Descriptive Statistics
C1 Ages
Variables:
Ages
By variables (optional):
Select
Statistics...
Graphs...
Help
OK
Cancel
Tab2-1.MTW ***
C1 C2 C3 C4 C5 C6 C13 C14 C15 C16 C17
Ages
1 23
2 38
3 42
4 25
5 60
6 55
7 50
8 42
9 32
Calculate descriptive statistics and display in Session window
start
Facebook - Mozilla ...
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
8:08 PM



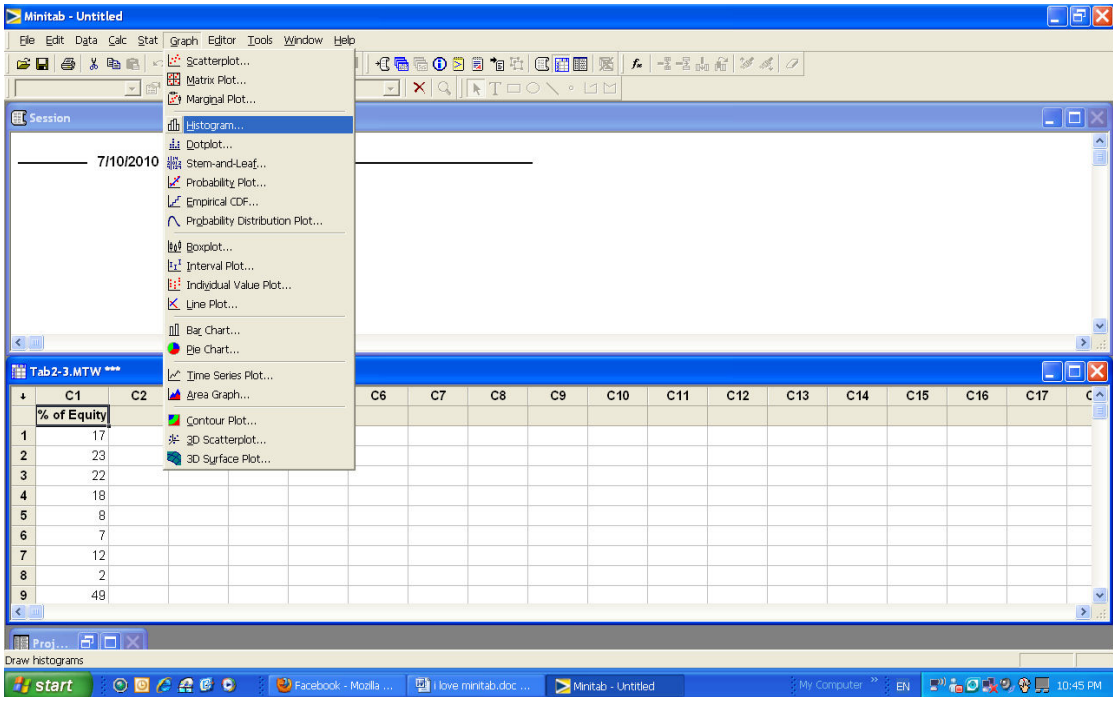
Δημιουργία Ραβδογράμματος

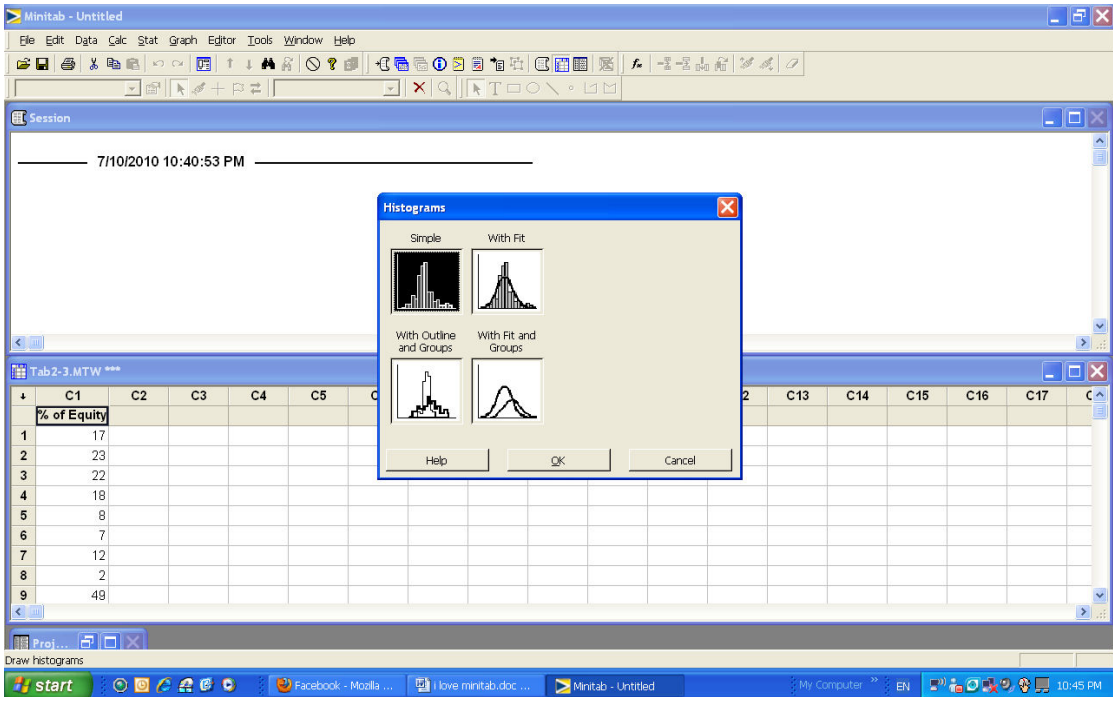


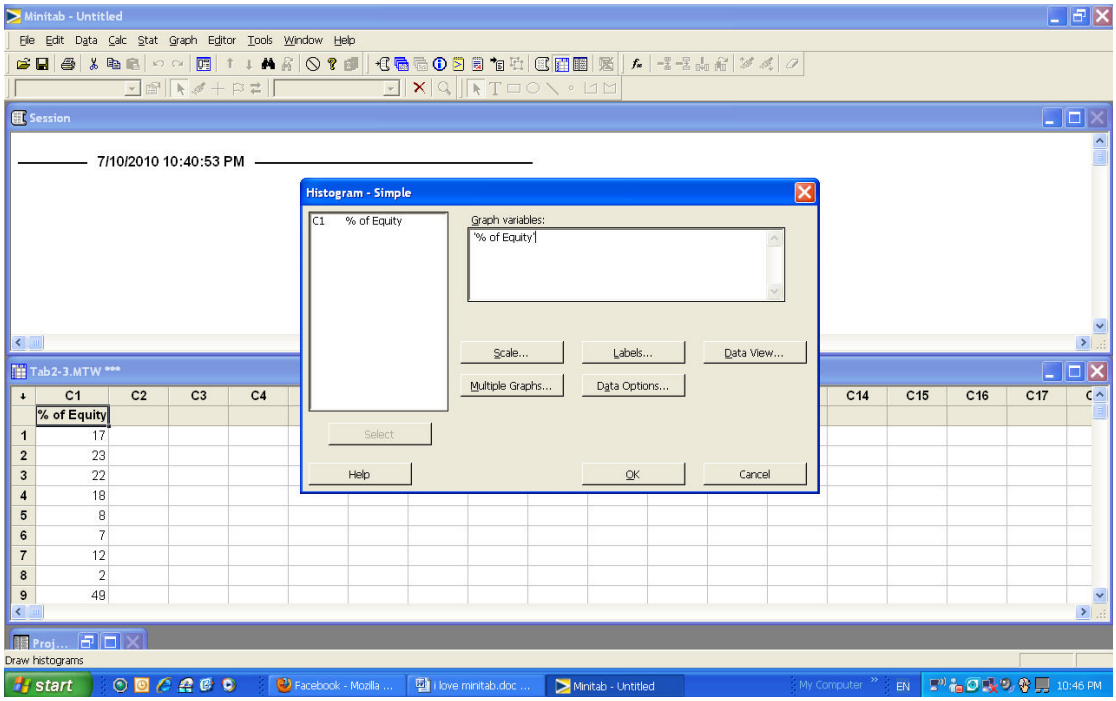

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/11/2010 3:12:27 PM
Welcome to Minitab, press F1 for help.
Graphical Summary
C1 Ages
Variables:
Ages
By variables (optional):
Confidence level: 95.0
Select
Help
OK
Cancel
Tab2-1.MTW ***
C1 C2 C3 C4 C5 C6 C13 C14 C15 C16 C17
Ages
1 23
2 38
3 42
4 25
5 60
6 55
7 50
8 42
9 32
Display a graphical summary
Editable
start
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
3:26 PM

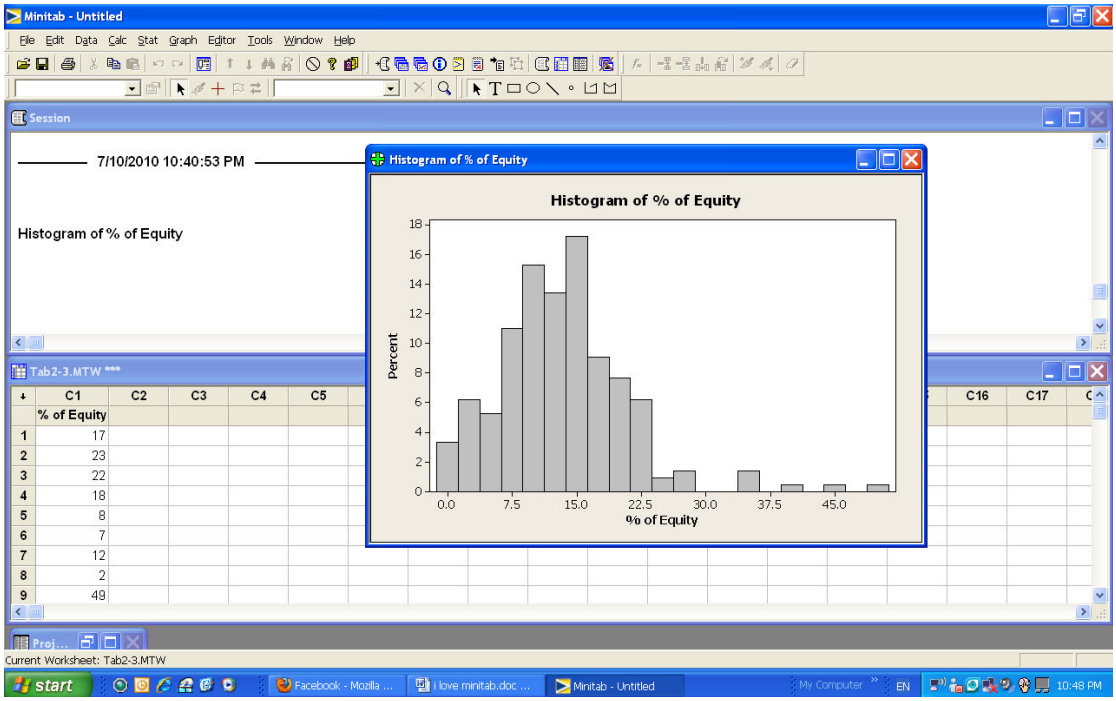

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/11/2010 3:12:27 PM
Welcome to Minitab, press F1 for help.
Results for: Tab2-1.MTW
Summary for Ages
Summary for Ages
Summary for Ages
Anderson-Darling Normality Test
A-Squared 0.17
P-Value 0.910
Mean 40.200
StDev 12.200
Variance 148.844
Skewness 0.203000
Kurtosis -0.821407
N 10
Minimum 23.000
1st Quartile 30.250
Median 40.000
3rd Quartile 51.250
Maximum 60.000
95% Confidence Interval for Mean
31.473 48.927
95% Confidence Interval for Median
29.604 51.712
95% Confidence Interval for StDev
8.392 22.273
95% Confidence Intervals
Mean
Median
Tab2-1.MTW ***
C1 C2 C3 C4 C5 C16 C17
Ages
1 23
2 38
3 42
4 25
5 60
6 55
7 50
8 42
9 32
Current Worksheet: Tab2-1.MTW
start
Minitab - Untitled
I love minitab.doc ...
My Computer
3:25 PM

Δημιουργία Ιστογράμματος

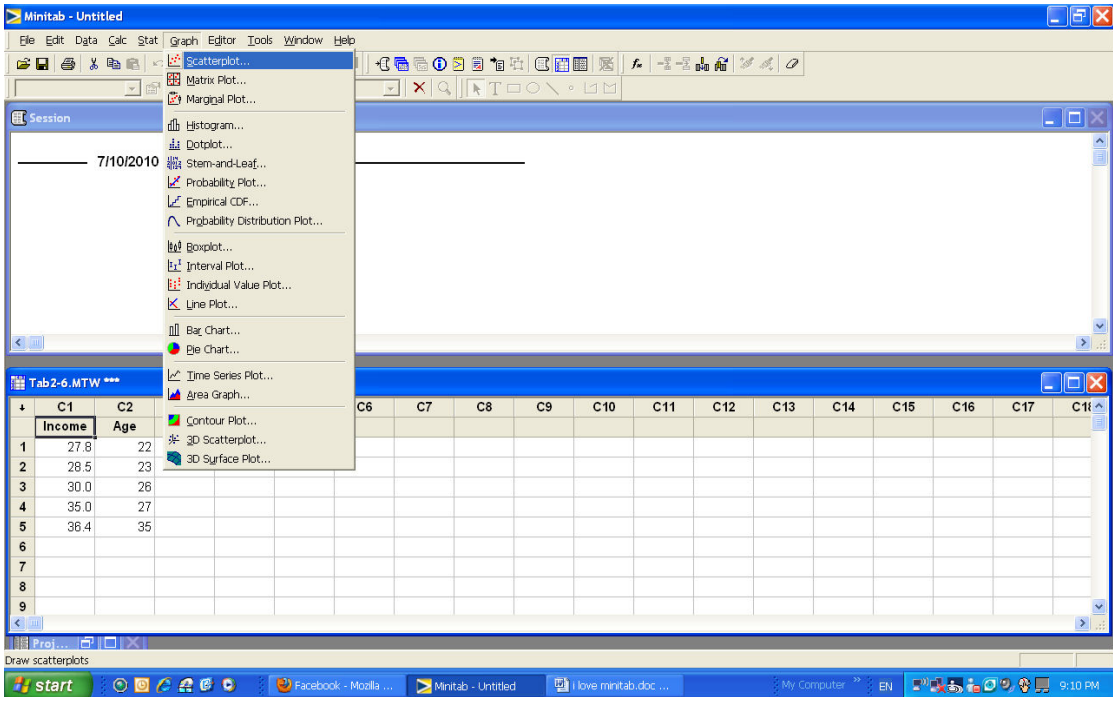


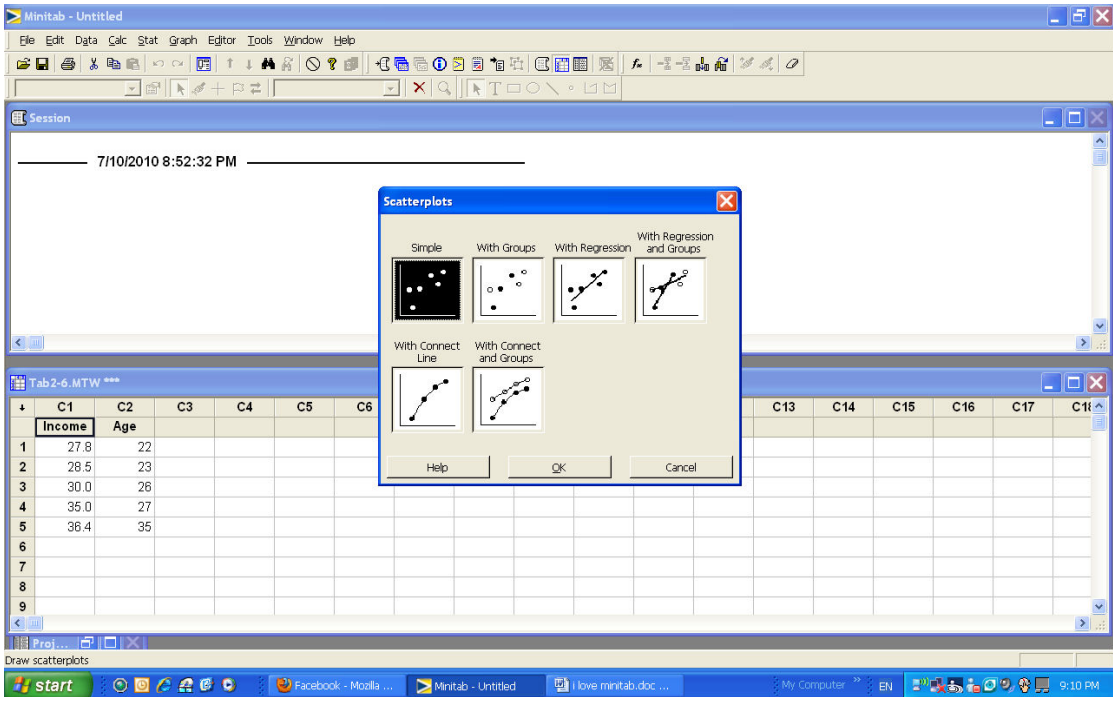


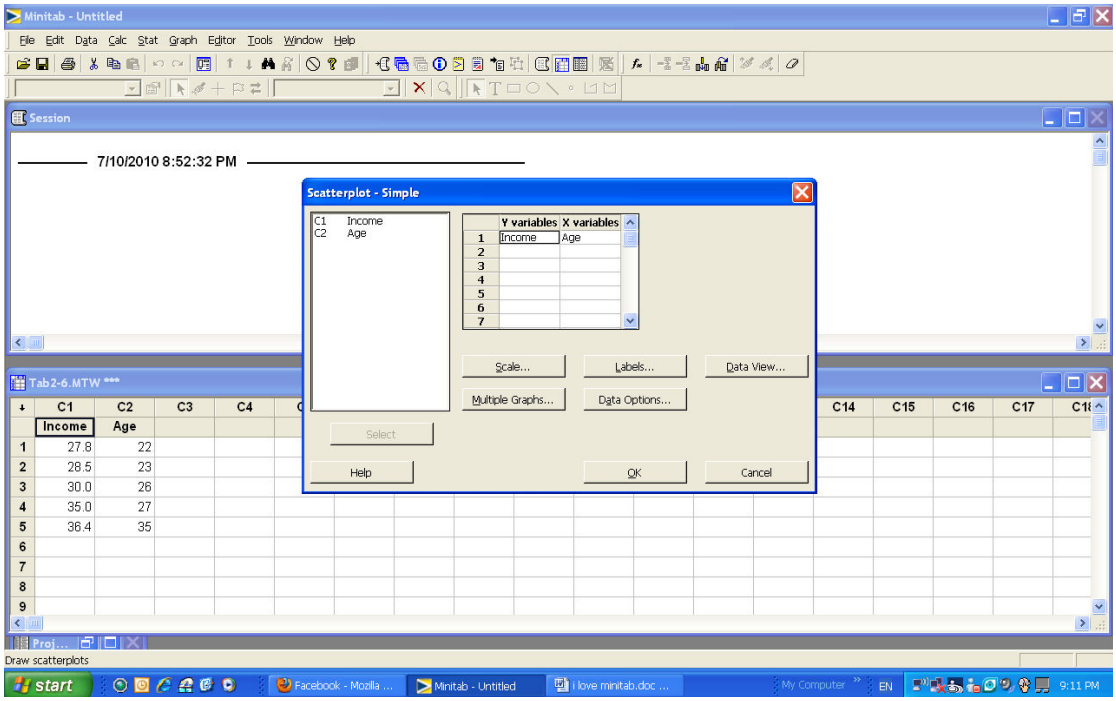

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/10/2010 10:40:53 PM
Histogram - Simple
C1 % of Equity
Graph variables:
'% of Equity'
Scale...
Labels...
Data View...
Multiple Graphs...
Data Options...
Select
Help
OK
Cancel
Tab2-3.MTW ***
% of Equity
17
23
22
18
8
7
12
2
49
Draw histograms
start
Minitab - Untitled
10:46 PM

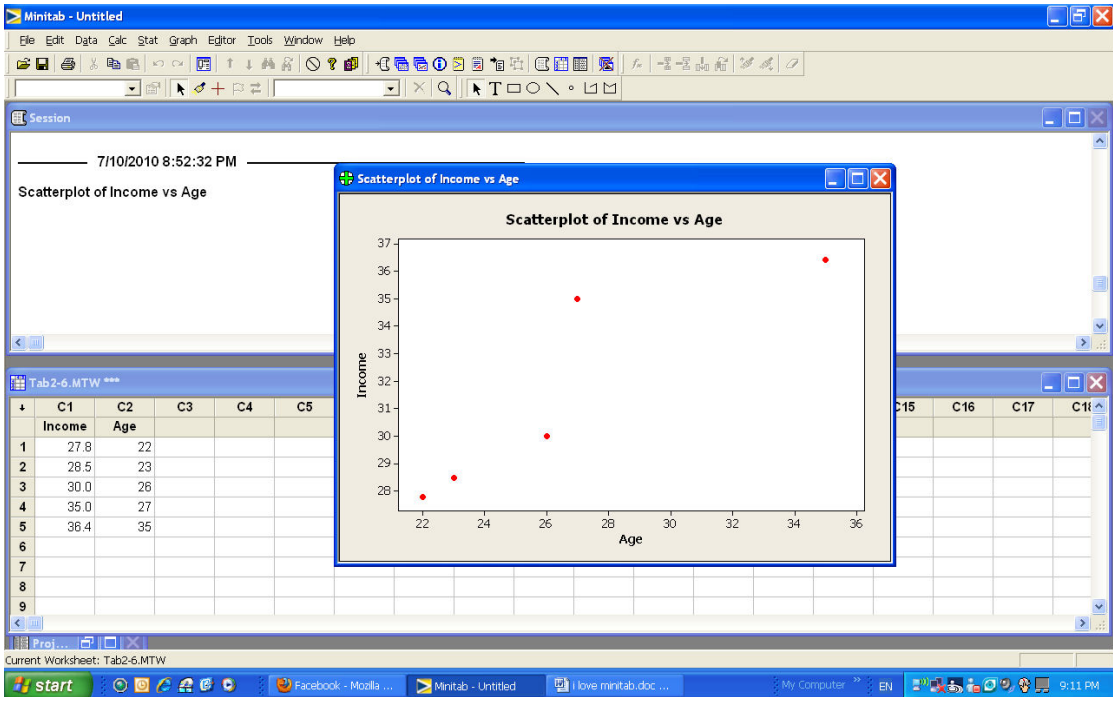

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/10/2010 10:40:53 PM
Histogram of % of Equity
Histogram of % of Equity
Percent
% of Equity
Tab2-3.MTW ***
% of Equity
17
23
22
18
8
7
12
2
49
Current Worksheet: Tab2-3.MTW
start
Minitab - Untitled
10:48 PM

Δημιουργία Διαγράμματος Διασποράς

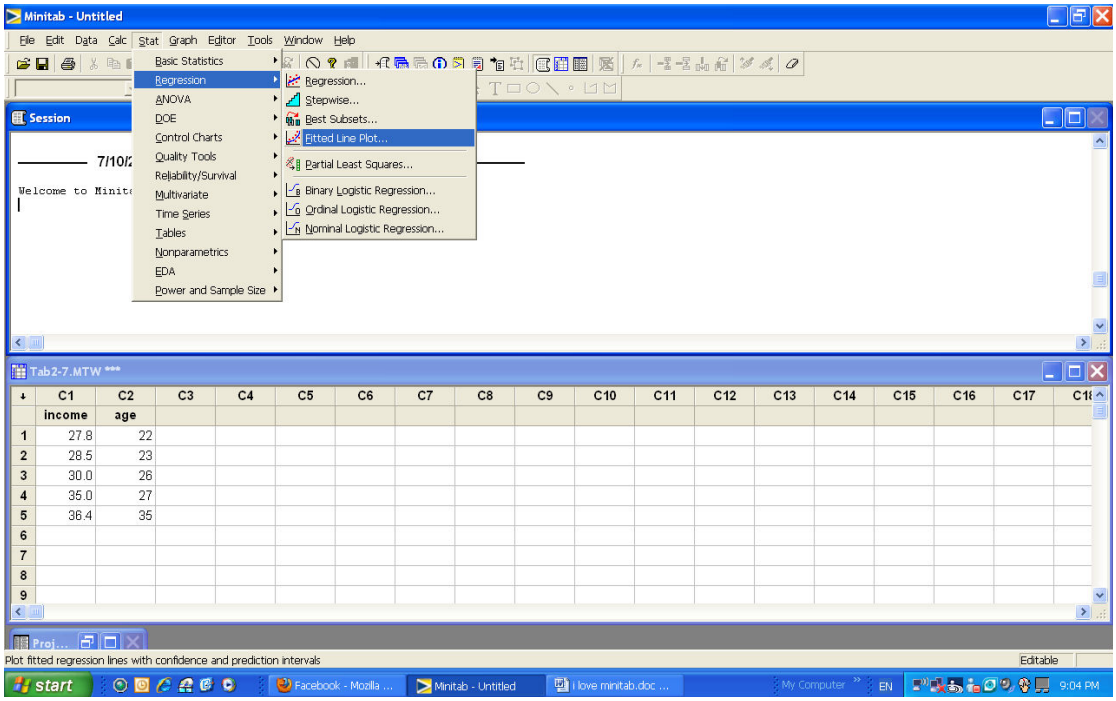


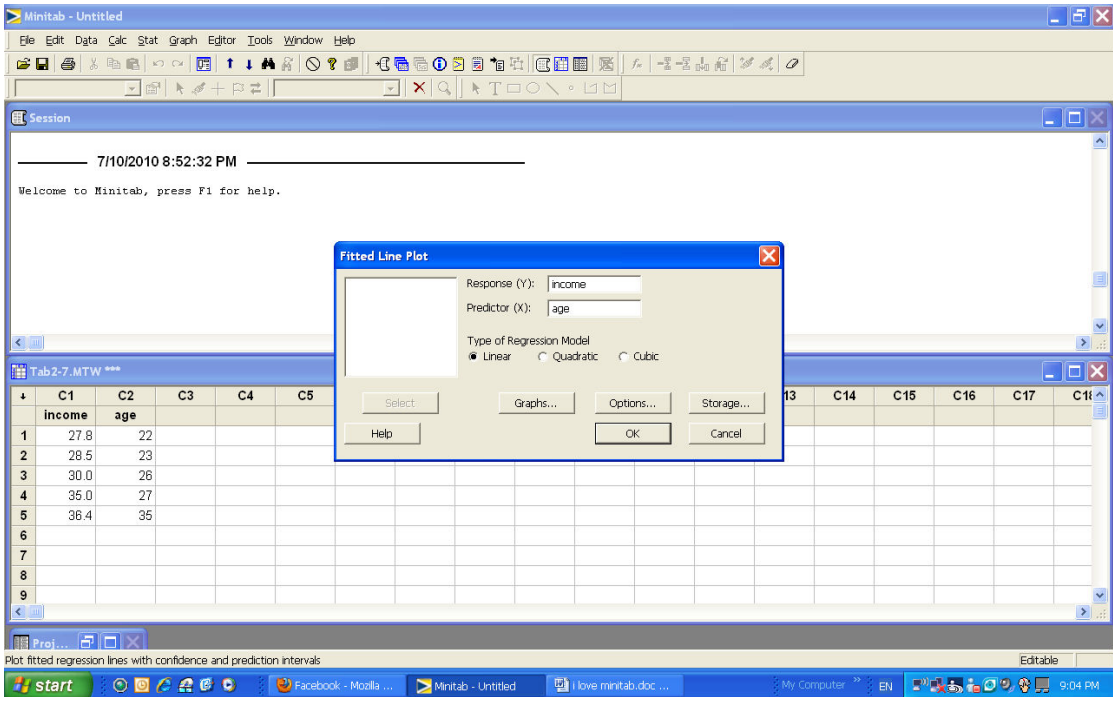


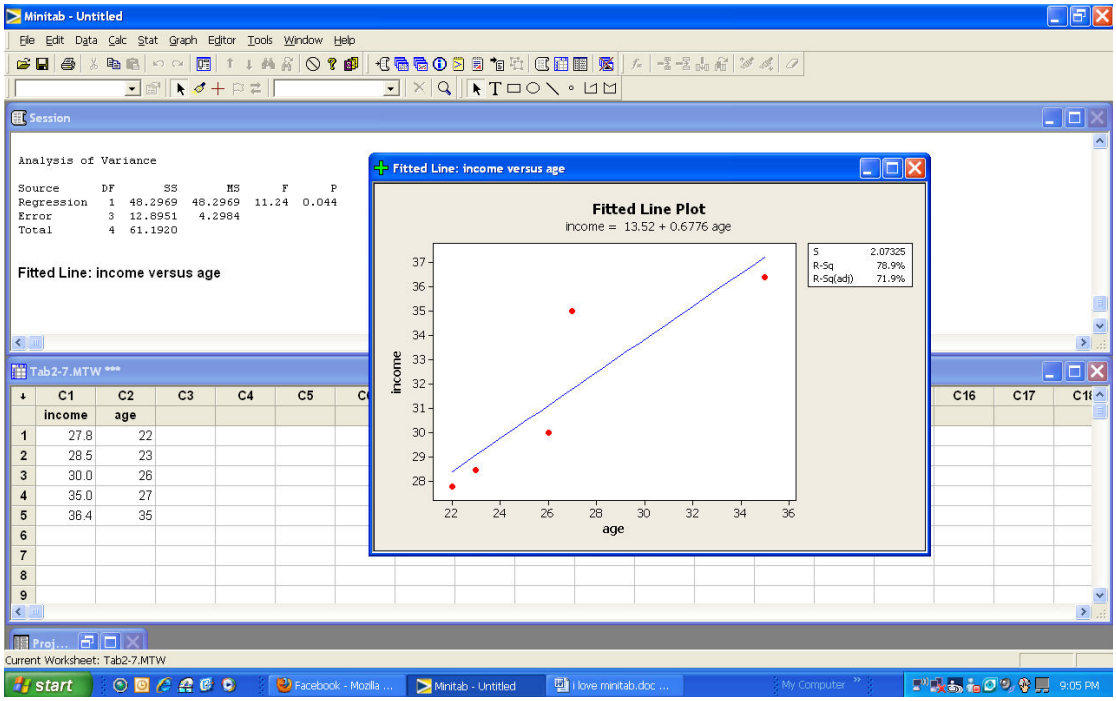

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/10/2010 8:52:32 PM
Scatterplot - Simple
C1 Income
C2 Age
Y variables
X variables
Income
Age
Scale...
Labels...
Data View...
Multiple Graphs...
Data Options...
Select
Help
OK
Cancel
Tab2-6.MTW ***
C1 C2 C3 C4 C14 C15 C16 C17
Income Age
1 27.8 22
2 28.5 23
3 30.0 26
4 35.0 27
5 36.4 35
Draw scatterplots
start
Facebook - Mozilla ...
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
9:11 PM


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/10/2010 8:52:32 PM
Scatterplot of Income vs Age
Scatterplot of Income vs Age
Scatterplot of Income vs Age
Income
Age
37
36
35
34
33
32
31
30
29
28
22
24
26
28
30
32
34
36
Tab2-6.MTW ***
C1 C2 C3 C4 C5 C15 C16 C17
Income Age
1 27.8 22
2 28.5 23
3 30.0 26
4 35.0 27
5 36.4 35
Current Worksheet: Tab2-6.MTW
start
Facebook - Mozilla ...
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
9:11 PM

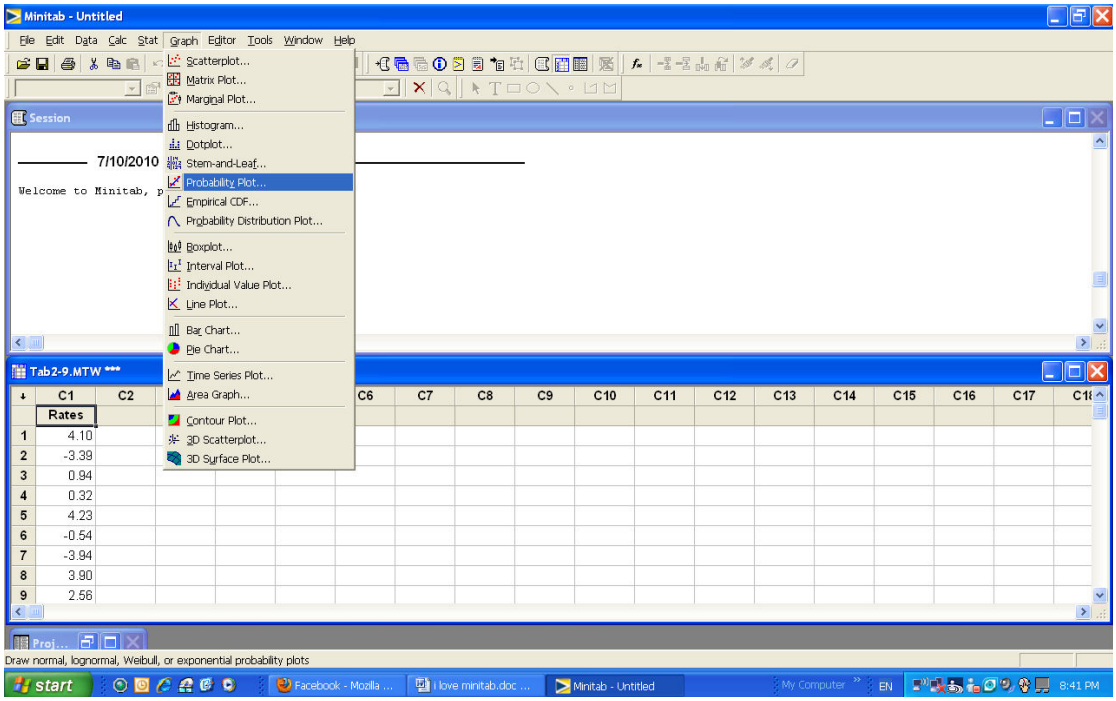
Δημιουργία Διαγράμματος Προσαρμοσμένης Ευθείας

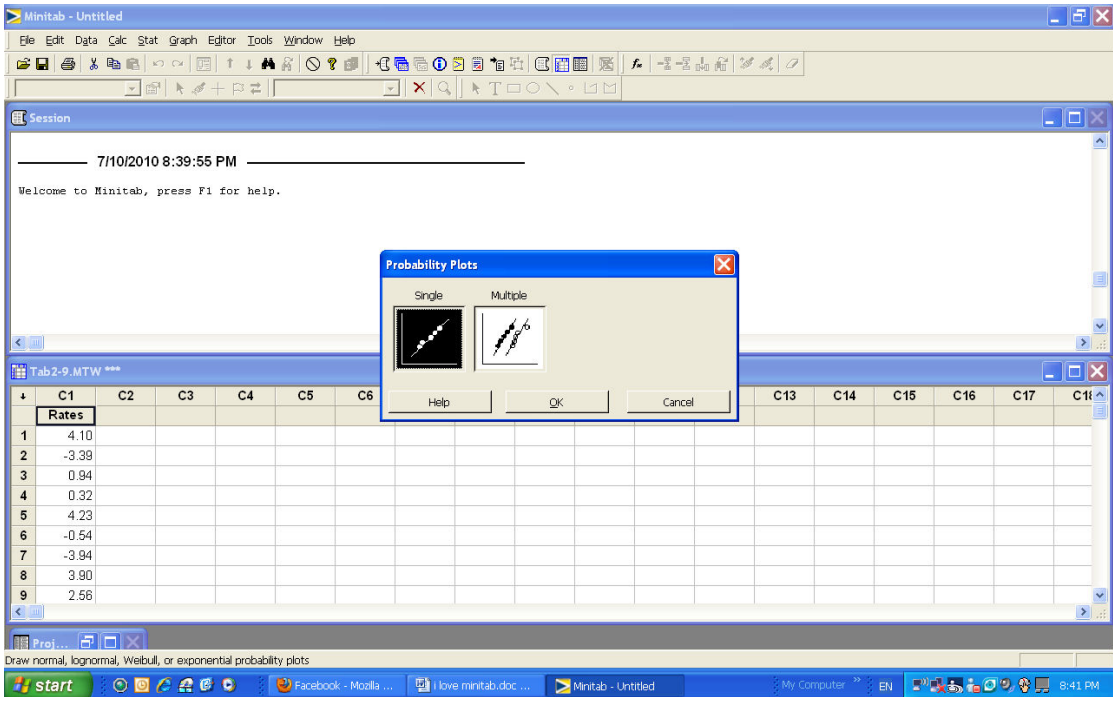


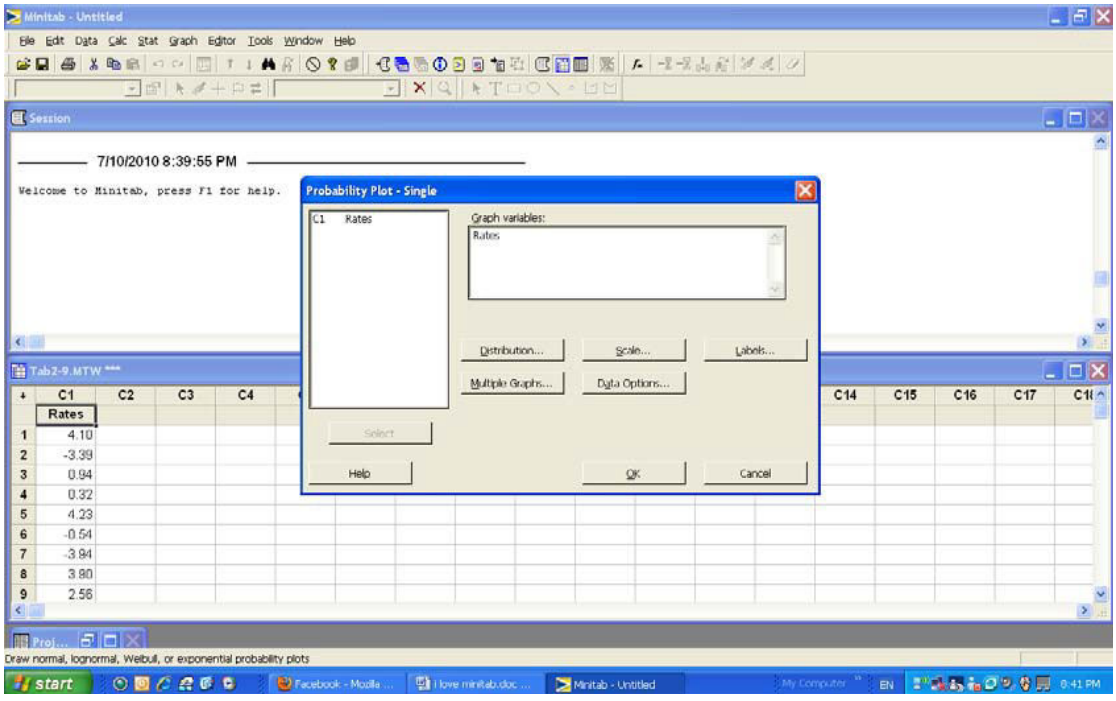


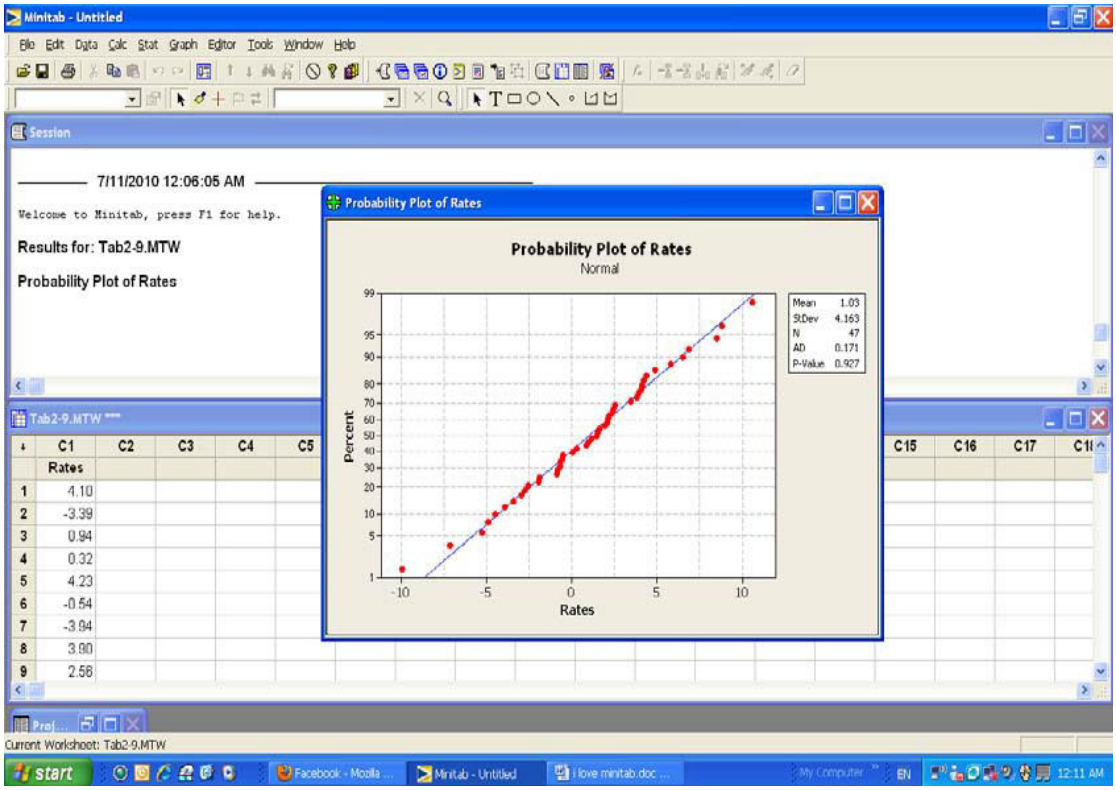


Δημιουργία Διαγράμματος Κανονικής Πιθανότητας




Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/10/2010 8:39:55 PM
Welcome to Minitab, press F1 for help.
Probability Plots
Single
Multiple
Help
OK
Cancel
Tab2-9.MTW ***
C1 Rates
1 4.10
2 -3.39
3 0.94
4 0.32
5 4.23
6 -0.54
7 -3.94
8 3.90
9 2.56
Draw normal, lognormal, Weibull, or exponential probability plots
start
Minitab - Untitled
8:41 PM


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/10/2010 8:39:55 PM
Welcome to Minitab, press F1 for help.
Probability Plot - Single
C1 Rates
Graph variables:
Rates
Distribution...
Scale...
Labels...
Multiple Graphs...
Data Options...
Select
Help
OK
Cancel
Tab2-9.MTW ***
C1 Rates
1 4.10
2 -3.39
3 0.94
4 0.32
5 4.23
6 -0.54
7 -3.94
8 3.90
9 2.56
Draw normal, lognormal, Weibull, or exponential probability plots
start
Minitab - Untitled
8:41 PM


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/11/2010 12:06:05 AM
Welcome to Minitab, press F1 for help.
Results for: Tab2-9.MTW
Probability Plot of Rates
Probability Plot of Rates
Probability Plot of Rates
Normal
Mean 1.03
StDev 4.163
N 47
AD 0.171
P-Value 0.927
Percent
Rates
Tab2-9.MTW ***
C1 C2 C3 C4 C5 C15 C16 C17
Rates
1 4.10
2 -3.39
3 0.94
4 0.32
5 4.23
6 -0.54
7 -3.94
8 3.90
9 2.58
Current Worksheet: Tab2-9.MTW
start
Facebook - Mozilla ...
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
12:11 AM

ΚΕΦΑΛΑΙΟ 3 (ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΜΕΘΟΔΟΥΣ ΠΡΟΒΛΕΨΗΣ)

3.1 ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

Σήμερα υπάρχουν πολλά προγράμματα Στατιστικής, τα οποία είναι εύκολο να εγκατασταθούν σε έναν υπολογιστή και είναι ιδιαίτερα φιλικά στον χρήστη. Έτσι, με μερικές απλές εντολές, που υπαγορεύονται από το ίδιο το πρόγραμμα, μπορούν να γίνουν και οι πιο πολύπλοκες στατιστικές αναλύσεις. Ο αναλυτής όμως, που χρησιμοποιεί τα προγράμματα αυτά, πρέπει να γνωρίζει καλά τη σχετική θεωρία, τη μέθοδο υπολογισμού καθώς και τους περιορισμούς της ανάλυσης, έτσι ώστε να είναι σε θέση:

- Να ελέγξει την ποιότητα των πληροφοριών, οι οποίες θα αποτελέσουν τη βάση της στατιστικής του ανάλυσης. Έτσι π.χ. και οι πιο προχωρημένες τεχνικές θα οδηγήσουν σε λανθασμένα συμπεράσματα, όταν εφαρμόζονται σε πληροφορίες, οι οποίες έχουν συγκεντρωθεί με λανθασμένη μεθοδολογία, ή από άτομα ανειδίκευτα και ανυποψίαστα για το σκοπό της έρευνας. Η κατάσταση αυτή συνοψίζεται διεθνώς με τη ρήση «Garbage In Garbage Out» ή GIGO. Ένα από τα πιο χρονοβόρα και δύσκολα κομμάτια της πρόβλεψης είναι η συλλογή έγκυρων και αξιόπιστων δεδομένων.
- Να επιλέξει τη μέθοδο ανάλυσης, που είναι κατάλληλη για τα δεδομένα και τους σκοπούς της έρευνας.
- Να ερμηνεύσει, να αξιολογήσει τα αποτελέσματα και να σκιαγραφήσει τα συμπεράσματα σχετικά με το υπό μελέτη πρόβλημα.

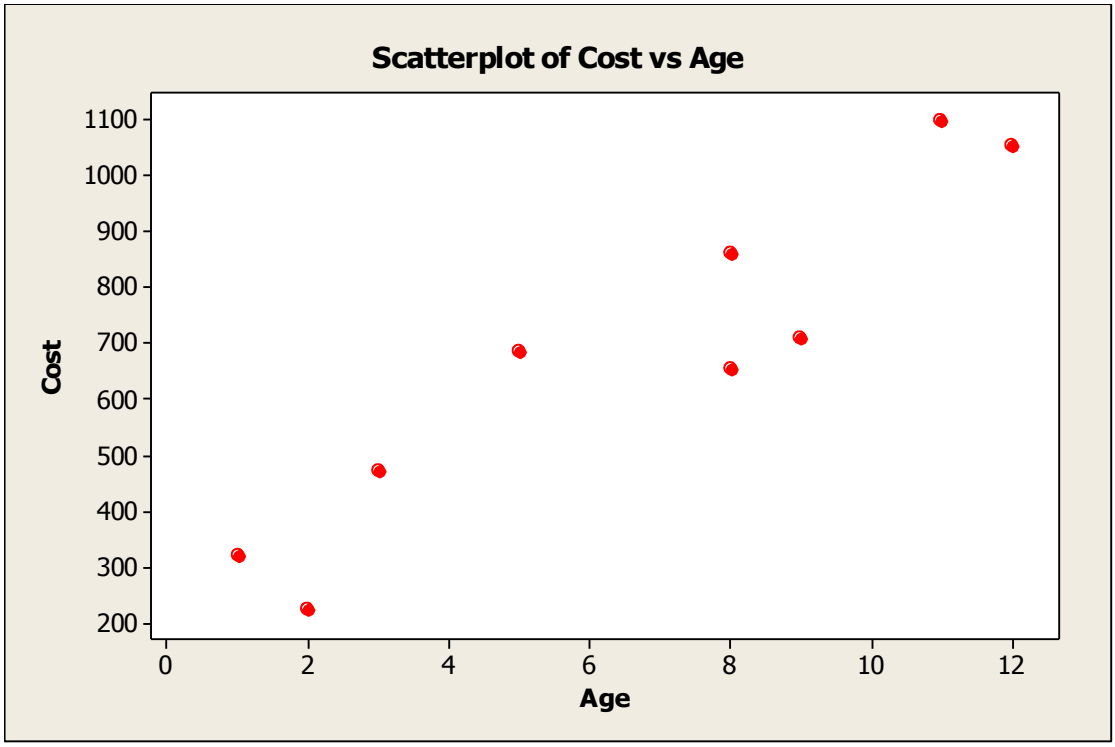
Οι σύγχρονοι υπολογιστές οδήγησαν σε μια συσσώρευση ενός απίστευτα μεγάλου ποσού δεδομένων σχετικά με όλα σχεδόν τα θέματα. Το δύσκολο έργο, που αντιμετωπίζουν οι περισσότεροι αναλυτές, είναι πώς να βρουν σχετικά δεδομένα, που θα τους βοηθήσουν στην επίλυση των προβλημάτων τους. Τέσσερα κριτήρια μπορούν να εφαρμοστούν, για να καθοριστεί αν τα δεδομένα είναι χρήσιμα:

- Τα δεδομένα θα πρέπει να είναι αξιόπιστα και ακριβή. Θα πρέπει να συλλέγονται από αξιόπιστη πηγή και να δίνουμε σημασία στην ακρίβεια.
- Τα δεδομένα θα πρέπει να είναι σχετικά και αντιπροσωπευτικά των περιστάσεων, για τις οποίες θα χρησιμοποιηθούν.
- Τα δεδομένα θα πρέπει να είναι συνεπή. Να αναπροσαρμόζονται ανάλογα με τις συνθήκες.
- Τα δεδομένα θα πρέπει να είναι έγκαιρα. Θα πρέπει να συλλέγονται σε τακτά χρονικά διαστήματα.

Γενικά, δύο τύποι δεδομένων ενδιαφέρουν τους αναλυτές. Ο πρώτος τύπος είναι τα δεδομένα, που συλλέγονται σε ένα και μόνο σημείο στο χρόνο, π.χ. μία ώρα, μία μέρα, μία εβδομάδα, ένας μήνας. Ο δεύτερος τύπος είναι παρατηρήσεις των δεδομένων, που γίνονται με την πάροδο του χρόνου.

Όταν όλες οι παρατηρήσεις προέρχονται από την ίδια χρονική περίοδο τα δεδομένα ονομάζονται *ομαδοποιημένα διαστρωματικά δεδομένα* (cross – sectional data) **(F3-1).**

Στην παρακάτω εικόνα δίδεται για μία συγκεκριμένη χρονική στιγμή το κόστος συντήρησης λεωφορείων σε σχέση με την ηλικία τους.



Οι παρατηρήσεις μιας τυχαίας μεταβλητής, οι οποίες συλλέγονται σε διαδοχικά ισαπέχοντα σημεία του χρόνου, αποτελούν μια *χρονολογική σειρά ή χρονοσειρά* (time series).

Παραδείγματα χρονοσειρών είναι μεταξύ άλλων οι τιμές των μετοχών, η προσφορά χρήματος, ο δείκτης τιμών καταναλωτή, ο δείκτης βιομηχανικής παραγωγής, το Ακαθάριστο Εγχώριο Προϊόν κλπ. Επειδή τα παρελθόντα γεγονότα μπορούν να επηρεάσουν τα μελλοντικά γεγονότα, ο χρόνος είναι μια σημαντική διάσταση – παράμετρος στην ανάλυση χρονοσειρών. Κατά την ανάλυση χρονοσειρών δεν επιχειρείται αναζήτηση σχέσεων αιτίου και αιτιατού, αλλά επιχειρείται η αναζήτηση επαναλαμβανόμενων μοτίβων (patterns) στην ίδια τη χρονοσειρά και υποτίθεται, πως τα μοντέλα αυτά θα συνεχίσουν να ισχύουν και στο μέλλον, προκειμένου να γίνουν κάποιες προβλέψεις.

Με την γραφική παράσταση μιας χρονοσειράς εννοούμε την καμπύλη, που παράγεται σε ένα σύστημα ορθογωνίων αξόνων, όπου ο οριζόντιος άξονας είναι ο χρόνος και ο κατακόρυφος οι μετρούμενες

τιμές του μεγέθους στο αντίστοιχο χρονικό διάστημα, που μετράμε. Με την βοήθεια της γραφικής παράστασης μιας χρονολογικής σειράς μπορούν να γίνουν ορισμένες γρήγορες διαπιστώσεις, όπου ανακαλύπτονται περιοδικότητες, τάσεις και άλλα. Επίσης, υπάρχουν περισσότερο σύνθετες και προχωρημένες περιγραφές, που γίνονται με την βοήθεια των στοχαστικών μοντέλων.

Μία από τις κυριότερες μεθόδους ανάλυσης χρονοσειρών είναι η μέθοδος της αυτοσυσχέτισης (autocorrelation), όπου η χρονοσειρά αναπαρίσταται με ένα δυναμικό μοντέλο (μοντέλο ARIMA – διαδικασία Box & Jenkins) και οι παρατηρήσεις σχετίζονται με το παρελθόν τους.

3.2 ΠΟΣΟΤΙΚΑ – ΠΟΙΟΤΙΚΑ ΜΟΝΤΕΛΑ

Ένα μοντέλο χρησιμοποιείται για την αναπαράσταση μιας διαδικασίας με τρόπο αφαιρετικό, στοχεύοντας σε μια απλοποιητική προσέγγιση του μηχανισμού της. Ειδικότερα, ένα μοντέλο πρόβλεψης αναπαριστά τη διαδικασία, που ακολουθείται, προκειμένου να παραχθούν οι προβλέψεις για το υπό μελέτη μέγεθος.

Τα μοντέλα πρόβλεψης κατηγοριοποιούνται σε δύο μεγάλες ομάδες:

- Τα ποσοτικά μοντέλα (quantitative models)
- Τα ποιοτικά μοντέλα (qualitative models)

3.2.1 ΠΟΣΟΤΙΚΑ ΜΟΝΤΕΛΑ

Τα ποσοτικά μοντέλα εφαρμόζονται, όταν πληρούνται τρεις βασικές προϋποθέσεις:

- Υπάρχει διαθέσιμη πληροφορία από το παρελθόν.
- Η πληροφορία αυτή μπορεί να ποσοτικοποιηθεί σε αριθμητικά

δεδομένα.

- Μπορεί να θεωρηθεί, ότι οι βασικές συνθήκες του παρελθόντος προτύπου συνεχίζονται και στο μέλλον.

Η τρίτη αυτή προϋπόθεση είναι γνωστή ως η "υπόθεση της συνέχειας" (assumption of continuity).

Τα ποσοτικά μοντέλα λοιπόν απαιτούν ικανοποιητικά ποσοτικά ιστορικά δεδομένα του υπό μελέτη μεγέθους. Ένα θέμα, που έχει απασχολήσει αρκετά τους ερευνητές, είναι ποιος είναι ο ελάχιστος αριθμός ιστορικών δεδομένων, που θα μπορούσε να χαρακτηριστεί ικανοποιητικός, ώστε να μπορεί να γίνει ακριβής πρόβλεψη. Αν και η απάντηση εξαρτάται από αρκετές παραμέτρους, ο γενικός κανόνας, που είναι αποδεκτός, είναι ότι ***"ο ιστορικός ορίζοντας πρέπει να είναι τουλάχιστον τετραπλάσιος του ορίζοντα πρόβλεψης"***.

Τα ποσοτικά μοντέλα διακρίνονται σε δύο βασικούς τύπους:

- Το μοντέλο χρονοσειρών (time series model)
- Το επεξηγηματικό μοντέλο (explanatory model)

Το μοντέλο χρονοσειρών (time series model) αποτελεί το πιο διαδεδομένο είδος ποσοτικού μοντέλου πρόβλεψης. Βασίζεται στην υπόθεση, ότι η μεταβολή της τιμής του μεγέθους ακολουθεί ένα συγκεκριμένο πρότυπο ("λανθάνον" πρότυπο), που επαναλαμβάνεται στον χρόνο και παραμένει σταθερό.

Έτσι ο βασικός στόχος των μοντέλων είναι η αναγνώριση του ακολουθούμενου προτύπου των ιστορικών δεδομένων και η προέκτασή του στο μέλλον υπό το καθεστώς των σημερινών συνθηκών. Έτσι, το συγκεκριμένο μοντέλο πρόβλεψης είναι κατάλληλο για να υπολογιστεί το μελλοντικό ύψος των πωλήσεων ενός προϊόντος, του οποίου η ποιότητα, η τιμή και το διαφημιστικό κόστος πρόκειται να παραμείνουν σταθερά.

Δεν θα μπορούσαμε όμως να το χρησιμοποιήσουμε στην περίπτωση, που άλλαζε κάποιο από τα παραπάνω βασικά χαρακτηριστικά, όπως για παράδειγμα η διαφημιστική πολιτική της εταιρείας για το συγκεκριμένο προϊόν. Γίνεται λοιπόν φανερή η αδυναμία του μοντέλου των χρονοσειρών να συσχετίσει το υπό πρόβλεψη μέγεθος, πωλήσεις, με τους παράγοντες, που επηρεάζουν τις πωλήσεις, όπως η τιμή του.

Με μία έννοια, το μοντέλο χρονοσειρών θεωρείται σαν "μαύρο κουτί" (black box) και δεν κάνει καμία προσπάθεια να ανακαλύψει τους συντελεστές που επηρεάζουν τη συμπεριφορά του.

Υπάρχουν τρία βασικά πλεονεκτήματα της θεώρησης του μαύρου κουτιού, που αποτελούν και τους βασικούς λόγους της συχνότατης επιλογής τους:

- Δεν υπάρχει πάντα η δυνατότητα να συσχετίσουμε ένα μεταβαλλόμενο μέγεθος με κάποιους παράγοντες και πολύ περισσότερο να προσδιορίσουμε τον τρόπο αλληλεπίδρασής τους.
- Σε πολλές περιπτώσεις ενδιαφερόμαστε να προσδιορίσουμε μόνο, το τι θα συμβεί και όχι το γιατί.
- Το κόστος που απαιτείται στην περίπτωση αυτή είναι πολύ μικρότερο σε σχέση με άλλες κατηγορίες μοντέλων (όπως το επεξηγηματικό).

Οι μέθοδοι, που περιγράφονται με το συγκεκριμένο μοντέλο, αποτελούν μια ιδιαίτερη κατηγορία των ποσοτικών μεθόδων και ονομάζονται μέθοδοι των χρονοσειρών. Σε αυτές συγκαταλέγονται η αποσύνθεση (decomposition), η εξομάλυνση (smoothing) και οι αυτοπαλινδρομικές μέθοδοι κινητού μέσου όρου (AutoRegressive Moving Average), γνωστότερες με την συντομογραφία ARMA. Οι

διαφορές μεταξύ των παραπάνω μεθόδων εντοπίζονται στην μορφή του συναρτησιακού τύπου, που περιγράφει το σύστημα.

Ένα επεξηγηματικό μοντέλο είναι χρήσιμο στη περιγραφή και επεξήγηση του γιατί και του πως δουλεύουν τα πράγματα, ή γιατί ένα φαινόμενο είναι έτσι.

Τα επεξηγηματικά μοντέλα ούτε είναι τελείως ακριβή, ούτε δίνουν πλήρη επεξήγηση της απόλυτης πραγματικότητας. Μας δίνουν όμως τη δυνατότητα να διασυνδέσουμε τη γνώση, τη παρατήρηση με ένα θεωρητικό υπόβαθρο για τα διάφορα φαινόμενα και γιαυτό το λόγο είναι χρήσιμα. Παράδειγμα επεξηγηματικού μοντέλου είναι η παλινδρόμηση, όπου δημιουργούνται σχέσεις αιτίου αιτιατού και μας δίνουν τη δυνατότητα να επεξηγήσουμε κάποια φαινόμενα.

3.2.2 ΠΟΙΟΤΙΚΑ ΜΟΝΤΕΛΑ

Τα ποιοτικά μοντέλα δεν απαιτούν δεδομένα με τον ίδιο τρόπο, όπως τα ποσοτικά. Οι είσοδοι, που απαιτούνται, εξαρτώνται από τη συγκεκριμένη μέθοδο, που εφαρμόζεται. Συνήθως απαιτείται σημαντική γνώση ποιοτικών χαρακτηριστικών, που μπορούν να προσφερθούν από ιδιαίτερα έμπειρα στελέχη του αντίστοιχου χώρου. Παράδειγμα ποιοτικού μοντέλου, είναι το μοντέλο που χρησιμοποιείται για να προβλέψει, πώς θα επηρεαστεί η κατανάλωση του πετρελαίου από μία πολύ μεγάλη αύξηση της τιμής του πετρελαίου. Στην περίπτωση αυτή χρησιμοποιείται ποιοτικό μοντέλο, αφού τα ιστορικά δεδομένα για αντίστοιχες συνθήκες είναι εξαιρετικά περιορισμένα.

3.3 ΣΥΝΕΧΕΙΣ – ΔΙΑΚΡΙΤΕΣ ΧΡΟΝΟΛΟΓΙΚΕΣ ΣΕΙΡΕΣ

Οι χρονολογικές σειρές διακρίνονται σε συνεχείς χρονολογικές

σειρές και σε διακριτές.

Συνεχείς (continuous) *χρονολογικές σειρές* είναι αυτές, όπου η τιμή του φαινόμενου παρατηρείται συνεχώς. Παράδειγμα συνεχών χρονολογικών σειρών είναι η συνεχόμενη καταγραφή της θερμοκρασίας του αέρα, ή η συνεχής παρακολούθηση των σεισμών.

Διακριτές (discrete) *χρονολογικές σειρές* είναι αυτές, όπου η τιμή του φαινομένου καταγράφεται σε ορισμένα χρονικά διαστήματα. Παράδειγμα διακριτών χρονολογικών σειρών είναι η τιμή μιας μετοχής ανά ημέρα, ή ο αριθμός των ηλιακών κηλίδων ανά έτος, όπου υπάρχουν τιμές σε συγκεκριμένα χρονικά διαστήματα.

3.4 ΣΥΝΙΣΤΩΣΕΣ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

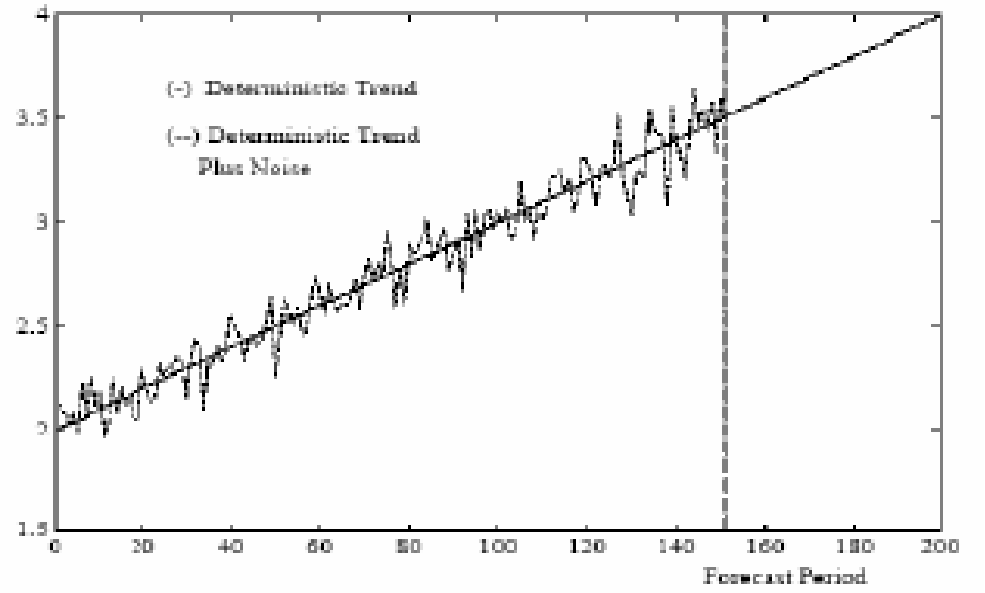
Οι τιμές των χρονολογικών σειρών που παρατηρούμε είναι το αποτέλεσμα της ταυτόχρονης επίδρασης τεσσάρων διαφορετικών συνιστωσών: της τάσης, της κυκλικότητας, της εποχικότητας και των τυχαίων κυμάνσεων.

3.4.1 ΤΑΣΗ

Τάση είναι η μακροχρόνια γενική κίνηση, που ακολουθεί η χρονοσειρά, που αναπαριστά την αύξηση, ή την πτώση των τιμών της σειράς σε μία εκτεταμένη περίοδο του χρόνου. Είναι, δηλαδή, η κατά κατά μέσο όρο απαλλαγμένη από βραχυχρόνιες αυξομειώσεις εξέλιξη της σειράς για μεγάλες χρονικές περιόδους, συνήθως πάνω από 10 έτη. Γι' αυτό το λόγο καλείται και μακροχρόνια τάση και μπορεί να είναι ανοδική ή καθοδική. Η τάση θεωρείται ανύπαρκτη, όταν η κεντρική κίνηση της χρονοσειράς είναι παράλληλη προς τον άξονα του χρόνου, χωρίς να παρουσιάζει τάση προς αύξηση ή μείωση. Η τάση ενσωματώνει τις

μακροχρόνιες εξελίξεις του προς μελέτη μεγέθους, που εκφράζει η μεταβλητή και είναι αποτέλεσμα της εξέλιξης της οικονομίας, των τεχνολογικών μεταβολών, των μακροχρόνιων αλλαγών των διαφόρων βιομηχανικών κλάδων, κ.λ.π.

Ορισμένες μέθοδοι προσδιορισμού της μακροχρόνιας τάσης είναι η μέθοδος των κινητών μέσων, η μέθοδος της ευθείας ελαχίστων τετραγώνων, η μέθοδος της καμπύλης ελαχίστων τετραγώνων και άλλα.

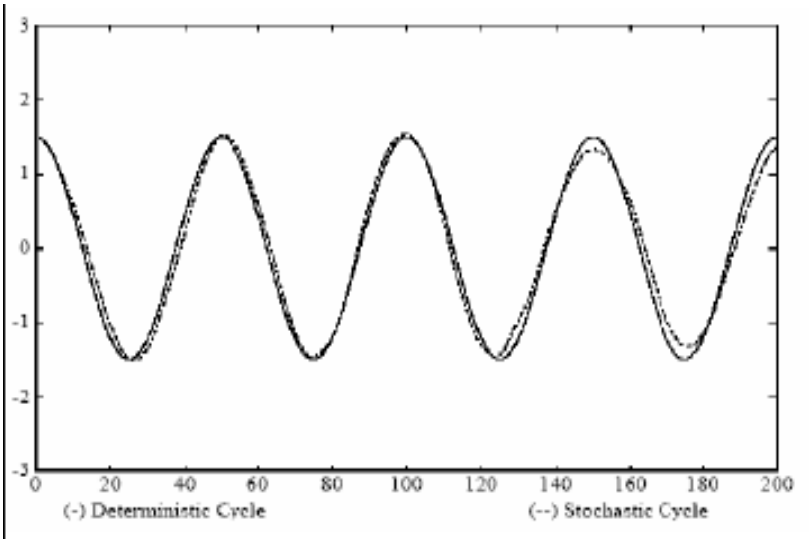


Σχήμα: Χρονοσειρά με μακροχρόνια τάση

3.4.2 ΚΥΚΛΙΚΗ ΣΥΝΙΣΤΩΣΑ

Η κυκλική συνιστώσα αντιπροσωπεύει εκείνες τις επαναλαμβανόμενες κυμάνσεις γύρω από την τάση, που η διάρκειά τους είναι μεγαλύτερη του έτους. Οι κυμάνσεις αυτές έχουν ανοδικές και καθοδικές φάσεις οι οποίες, συνήθως, διαρκούν μερικά έτη. Μία πλήρης κυκλική κύμανση αποτελείται από δύο κάτω σημεία καμπής και ένα άνω σημείο καμπής, που χρονικά παρεμβάλλεται μεταξύ των δύο πρώτων. Ο χρόνος μεταξύ δύο διαδοχικών κάτω ή άνω σημείων καμπής αποτελεί την περίοδο της κυκλικής κύμανσης. Οι οικονομικοί κύκλοι εμφανίζονται

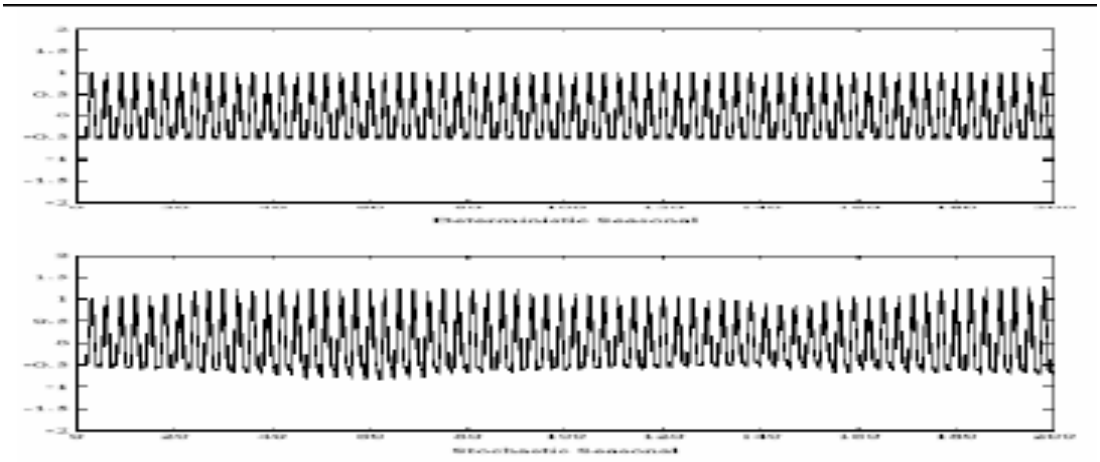
κυρίως στις οικονομικά ανεπτυγμένες οικονομίες. Έτσι, στις δυτικές χώρες τα περισσότερα οικονομικά μεγέθη, όπως τιμές, επενδύσεις, κατανάλωση, εισόδημα, κ.λ.π. παρουσιάζουν κυκλικές κυμάνσεις λίγο, ή πολύ έντονες. Μάλιστα, λόγω του ότι οι οικονομίες αυτές είναι ανοιχτές, οι κυμάνσεις μιας οικονομίας προκαλούν αντίστοιχες κυμάνσεις και στις οικονομίες των άλλων χωρών. Το κυριότερο πρόβλημα που αντιμετωπίζουμε στην εξέταση των οικονομικών κύκλων είναι ότι η διάρκεια τους δεν είναι σταθερή.



3.4.3 ΕΠΟΧΙΚΗ ΣΥΝΙΣΤΩΣΑ

Η εποχική συνιστώσα είναι μια κυκλική κύμανση με περίοδο όμως το έτος, διότι μέσα σ' αυτό εξαντλεί όλες τις ανοδικές και καθοδικές κινήσεις. Επίσης, είναι περιοδική, διότι επαναλαμβάνεται ρυθμικά κάθε έτος. Είναι προφανές ότι η εποχική κύμανση εμφανίζεται μόνο στις χρονοσειρές με εποχικές παρατηρήσεις. Η εποχική κύμανση, που το όνομα της προέρχεται από το γεγονός, ότι συνδέεται με τις εποχές, δεν οφείλεται μόνο στις κλιματολογικές διαφορές μεταξύ των εποχών. Ο μεταβαλλόμενος αριθμός των εργάσιμων ημερών μεταξύ των μηνών του έτους, το διαφορετικό ωράριο των καταστημάτων κ.λ.π. είναι μερικές

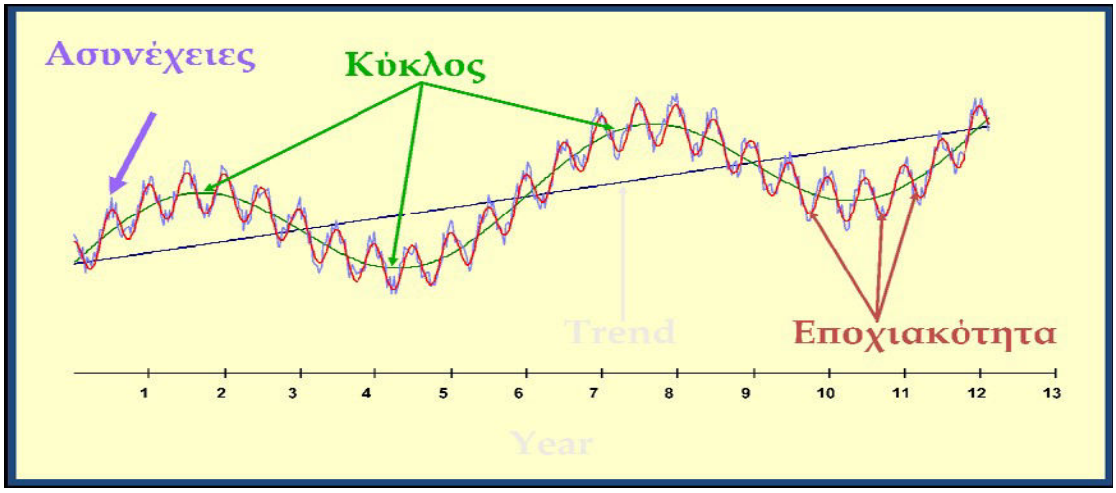
από τις αιτίες των περιοδικών κυμάνσεων, που εμφανίζουν οι χρονοσειρές με εποχικά δεδομένα.



3.4.4 ΤΥΧΑΙΑ ΣΥΝΙΣΤΩΣΑ

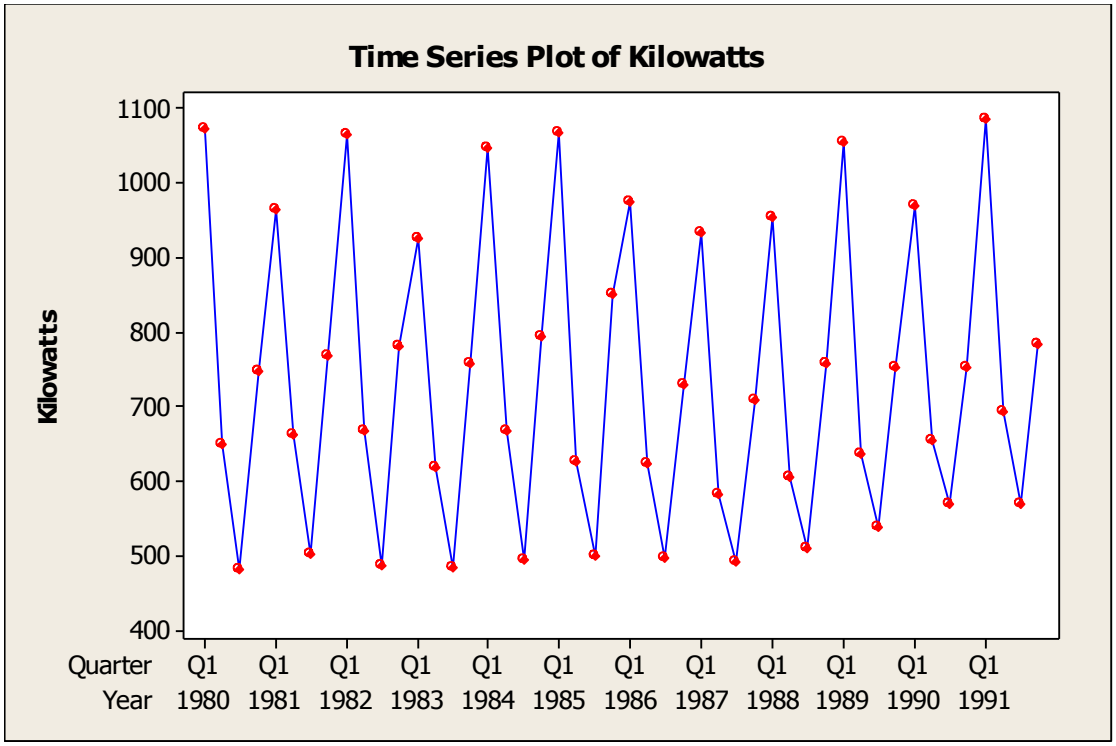
Οποιαδήποτε επίδραση στη διαμόρφωση της τιμής της μεταβλητής, που δεν οφείλεται σε κάποια από τις παραπάνω συνιστώσες, θεωρείται τυχαία ή άρρυθμος κύμανση. Θα πρέπει όμως να σημειωθεί, ότι η τυχαία συνιστώσα εμφανίζεται ακανόνιστα με επιδράσεις, που άλλοτε είναι θετικές και άλλοτε αρνητικές. Οι τυχαίες κυμάνσεις οφείλονται σε όλες εκείνες τις επιδράσεις, που δεν είναι συστηματικές και επομένως δεν μπορούν να προβλεφθούν.

Παραδείγματα τέτοιων επιδράσεων είναι ξαφνικές εξαγγελίες κυβερνητικών μέτρων, απρόβλεπτες αλλαγές τιμών στη διεθνή αγορά, ασυνήθιστες κλιματολογικές συνθήκες, πολιτικές κρίσεις, φυσικές καταστροφές, κ.λ.π.



ΠΑΡΑΔΕΙΓΜΑ (F3-3)

Το παρακάτω διάγραμμα δείχνει την ηλεκτρική ενέργεια, που δαπανάται για μεταφορά νερού για βιομηχανική και οικιακή χρήση στην Washington.



3.5 ΣΤΑΣΙΜΟΤΗΤΑ

3.5.1 ΟΡΙΣΜΟΙ

Υπάρχουν δύο ορισμοί της στασιμότητας. Ο πρώτος ορισμός είναι ο ισχυρός ορισμός (Strict Stationarity – ισχυρή στασιμότητα) και ο δεύτερος ο ασθενής ορισμός (Weak Stationarity – ασθενής στασιμότητα).

Κατά τον ισχυρό ορισμό μία χρονοσειρά είναι στάσιμη, όταν όλες οι ροπές είναι ανεξάρτητες του χρόνου, ή όταν οι ιδιότητες της δεν επηρεάζονται από μια αλλαγή στην αρχή μετρήσεως του χρόνου. Κατά τον δεύτερο ορισμό μία χρονοσειρά είναι στάσιμη, όταν οι δύο πρώτες ροπές (μέση τιμή – διασπορά) είναι ανεξάρτητες του χρόνου. (Ανεξάρτητες του χρόνου δηλαδή σταθερές). Η εποχικότητα και η κυκλικότητα μπορούν να δώσουν ασθενώς στάσιμες χρονοσειρές στα όρια της στατιστικής, που χρησιμοποιούμε (όρια εμπιστοσύνης). Ως παράδειγμα μιας ασθενώς στάσιμης χρονοσειράς είναι η ημιτονοειδής χρονοσειρά, η οποία είναι αυστηρώς περιοδική. Η χρονοσειρά αυτή είναι στάσιμη, διότι η μέση τιμή και η διασπορά είναι σταθερές και ανεξάρτητες του χρόνου.

Με άλλα λόγια, εάν μια χρονοσειρά παρουσιάζει τάση, τότε αυτή δεν θα είναι στάσιμη.

Η στασιμότητα αποτελεί ένα βασικό χαρακτηριστικό των χρονολογικών σειρών, καθώς απλοποιεί την μελέτη τους. Έτσι, βασικός σκοπός της μεθόδου της ανάλυσης είναι να διαχωριστούν και να απομονωθούν τα διάφορα χαρακτηριστικά μιας χρονολογικής σειράς και κυρίως τα στάσιμα από τα μη – στάσιμα χαρακτηριστικά. Συνήθως, σε μια χρονολογική σειρά συνυπάρχουν τρία χαρακτηριστικά. Αυτά είναι η τάση, η περιοδικότητα – εποχικότητα και ο θόρυβος.

3.5.2 ΕΛΕΓΧΟΣ ΣΤΑΣΙΜΟΤΗΤΑΣ ΧΡΟΝΟΣΕΙΡΑΣ

Εάν η χρονοσειρά είναι στάσιμη, τότε τα δεδομένα κυμαίνονται γύρω από ένα σταθερό μέσο, είναι δηλαδή ανεξάρτητα του χρόνου και η διακύμανση να παραμένει σταθερή ή όχι. Συνήθως είναι δυνατόν να ελεγχθεί η στασιμότητα με τη χρήση της γραφικής παράστασης της χρονοσειράς.

- Εάν δεν παρατηρείται αλλαγή της μέσης τιμής με την πάροδο του χρόνου, τότε η χρονοσειρά είναι στάσιμη ως προς τη μέση τιμή.
- Εάν δεν παρατηρείται αλλαγή της διακύμανσης με την πάροδο του χρόνου, τότε η χρονοσειρά είναι στάσιμη ως προς τη διακύμανση.

Το διάγραμμα των αυτοσυσχετίσεων μπορεί επίσης να χρησιμοποιηθεί για τον έλεγχο της στασιμότητας ως προς τη μέση τιμή. ***Οι αυτοσυσχετίσεις στάσιμων χρονοσειρών φθίνουν στο μηδέν με γρήγορο ρυθμό, ενώ για μη στάσιμες χρονοσειρές φθίνουν με αργό ρυθμό, καθώς αυξάνει ο αριθμός των καθυστερήσεων (Lags).***

3.6 ΑΥΤΟΣΥΣΧΕΤΙΣΗ

Ως αυτοσυσχέτιση εννοούμε τη συσχέτιση μιας μεταβλητής με καθυστέρηση μίας ή περισσοτέρων χρονικών περιόδων της ίδιας της μεταβλητής.

Η αυτοσυσχέτιση είναι ένα φαινόμενο, το οποίο κατά κανόνα παρατηρείται, όταν οι παρατηρήσεις των μεταβλητών προέρχονται από χρονοσειρές. Αν μάλιστα αναλογιστεί κάποιος, ότι οι περισσότερες περιπτώσεις μελέτης οικονομικών φαινομένων στηρίζονται σε παρατηρήσεις μεταβλητών σε χρονοσειρές, τότε είναι προφανές, ότι η εμφάνιση της αυτοσυσχέτισης θεωρείται δεδομένη. Αυτό συμβαίνει διότι οι παρατηρούμενες τιμές αντιπροσωπεύουν συνδυασμό διαφόρων

παραγόντων, οι οποίοι επηρεάζουν τη συμπεριφορά των τιμών της εξαρτημένης μεταβλητής. Έτσι, οι διάφοροι παράγοντες θα πρέπει να υφίστανται χρονικά κάποιο βαθμό συσχέτισης. Με άλλα λόγια, η συμπεριφορά αυτών των παραγόντων κατά την τρέχουσα περίοδο είναι πολύ πιθανόν, να είναι όμοια με εκείνη της προηγούμενης περιόδου, δηλώνοντας με αυτόν τον τρόπο την ύπαρξη πιθανής και ιδιαίτερα θετικής συσχέτισης μεταξύ των τιμών, οι οποίες βρίσκονται χρονικά πλησίον μεταξύ τους.

Στην ανάλυση χρονοσειρών μας ενδιαφέρει να ελέγξουμε, αν υπάρχει αλληλεξάρτηση μεταξύ διαδοχικών παρατηρήσεων, οπότε παραβιάζεται μία από τις υποθέσεις των τυχαίων δεδομένων, που είναι η ανεξαρτησία των παρατηρήσεων. Για τον σκοπό αυτό υπολογίζονται οι συντελεστές αυτοσυσχέτισης.

Ο *συντελεστής αυτοσυσχέτισης* (autocorrelation coefficient) συμβολίζεται με r_k και υπολογίζεται ως εξής:

$$r_k = \frac{\sum_{t=k+1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2}$$

Ο r_k υποδηλώνει, πώς οι διαδοχικές παρατηρήσεις (καθυστέρηση ή χρονοκαθυστέρηση k, ή lag k) της χρονοσειράς σχετίζονται μεταξύ τους. Οι συντελεστές αυτοσυσχέτισης για καθυστερήσεις 1, 2, ..., n δημιουργούν τη συνάρτηση αυτοσυσχέτισης (autocorrelation function).

Η σχέση, λοιπόν, που υπάρχει ανάμεσα στο συντελεστή αυτοσυσχέτισης r_k και στο k ονομάζεται *συνάρτηση αυτοσυσχέτισης* (autocorrelation function) και η γραφική απεικόνισή της *διάγραμμα αυτοσυσχέτισης* (correlogram).

Η σημασία της συνάρτησης αυτοσυσχέτισης είναι πολύ μεγάλη, πρώτον γιατί δίνει το μέτρο της συσχέτισης των παρατηρήσεων – μετρήσεων, οι οποίες απέχουν χρονικό διάστημα s και δεύτερον γιατί δείχνει τόσο το βαθμό (ένταση), όσο και το μήκος, ή την χρονική διάρκεια της μνήμης της στοχαστικής διαδικασίας. Δηλαδή, εκφράζει κατά πόσο οι μετρήσεις με χρονική απόσταση s έχουν σχέση μεταξύ τους.

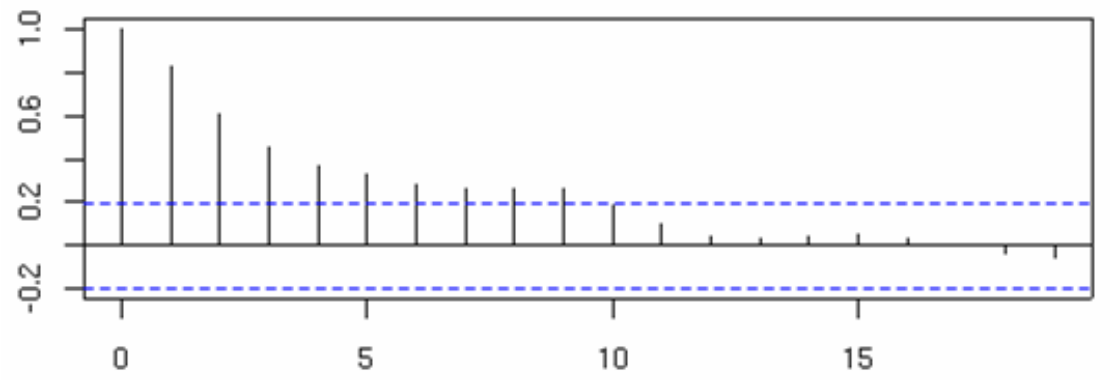
Με τον συντελεστή αυτοσυσχέτισης απαντάμε λοιπόν στα ερωτήματα:

- Είναι τα δεδομένα τυχαία;
- Έχουν τα δεδομένα τάση;
- Είναι τα δεδομένα στάσιμα;
- Είναι τα δεδομένα εποχικά;

Στην περίπτωση, που τα δεδομένα είναι τυχαία, ο συντελεστής αυτοσυσχέτισης θα είναι κοντά στο μηδέν.

Εάν υπάρχει τάση στα δεδομένα, ο συντελεστής αυτοσυσχέτισης της πρώτης χρονοκαθυστέρησης θα είναι συνήθως κοντά στο 1.

Στην περίπτωση, που υπάρχει κάποια εποχικότητα στα δεδομένα, ο συντελεστής αυτοσυσχέτισης θα παρουσιάζει κάποιο peak στη χρονοκαθυστέρηση 4 για δεδομένα τριμήνου, ή 12 για δεδομένα μηνός.



Πως κάποιος αναλυτής μπορεί να ορίσει, αν ο συντελεστής αυτοσυσχέτισης είναι κοντά στο μηδέν ή όχι; Ο Quenouille (1949) έδειξε, ότι η δειγματική κατανομή του συντελεστή αυτοσυσχέτισης τυχαίων δεδομένων ακολουθεί προσεγγιστικά την κανονική κατανομή με μέση τιμή μ=0 και τυπική απόκλιση $S=\frac{1}{\sqrt{n}}$. Επίσης σε κάποια λογισμικά πακέτα χρησιμοποιείται η σχέση:

$$SE(r_k)=\sqrt{\frac{1+2\sum_{i=1}^{k-1} r_i^2}{n}}$$

όπου,

- $SE(r_k)$ = η εκτιμώμενη τυπική απόκλιση του συντελεστή αυτοσυσχέτισης,
- r_i = ο συντελεστής αυτοσυσχέτισης στη χρονοκαθυστέρηση i,
- k = η χρονοκαθυστέρηση,
- n = ο αριθμός των παρατηρήσεων.

Η σχέση αυτή υποθέτει, ότι κάθε αυτοσυσχέτιση πριν τη χρονοκαθυστέρηση k είναι διαφορετική του μηδενός και κάθε αυτοσυσχέτιση για χρονοκαθυστέρηση μεγαλύτερη ή ίση του k είναι μηδέν. Για τον πρώτο συντελεστή αυτοσυσχέτισης η τυπική απόκλιση είναι $S=\frac{1}{\sqrt{n}}$.

Για κάποιο συγκεκριμένο επίπεδο εμπιστοσύνης μία χρονοσειρά μπορεί να θεωρηθεί τυχαία, όταν κάθε συντελεστής αυτοσυσχέτισης βρίσκεται σε ένα διάστημα γύρο από το 0. Το διάστημα ορίζεται ως $0\pm t*SE(r_k)$, όπου t είναι εκατοστιαίο σημείο της κατανομής Student,

ή της κανονικής κατανομής ανάλογα με το n.

Αν και είναι χρήσιμο να εξετάζουμε, αν ο κάθε ένας συντελεστής αυτοσυσχέτισης r_k ξεχωριστά είναι σημαντικά διαφορετικός από το μηδέν, μπορούμε να εξετάσουμε τους συντελεστές όλους μαζί σαν ομάδα. Γιαυτό το λόγο χρησιμοποιούμε τη στατιστική Q των Ljung – Box, η οποία ορίζεται ως:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k},$$

όπου,

- r_k = είναι οι συντελεστές αυτοσυσχέτισης.
- n = ο αριθμός των παρατηρήσεων της χρονοσειράς.
- m = ο αριθμός των χρονοκαθυστερήσεων, που πρέπει να ελεγχθούν.
- k = η εκάστοτε χρονοκαθυστέρηση (lag).

Η στατιστική Q εξετάζει τους συντελεστές αυτοσυσχέτισης σαν ομάδα και όχι ανεξάρτητα τον κάθε ένα. Δηλαδή ο κάθε συντελεστής συνεισφέρει κάτι στη συνολική συσχέτιση μιας χρονοσειράς με τον εαυτό της ή οι παράγοντες επιδρούν αθροιστικά σε κάθε χρονική περίοδο.

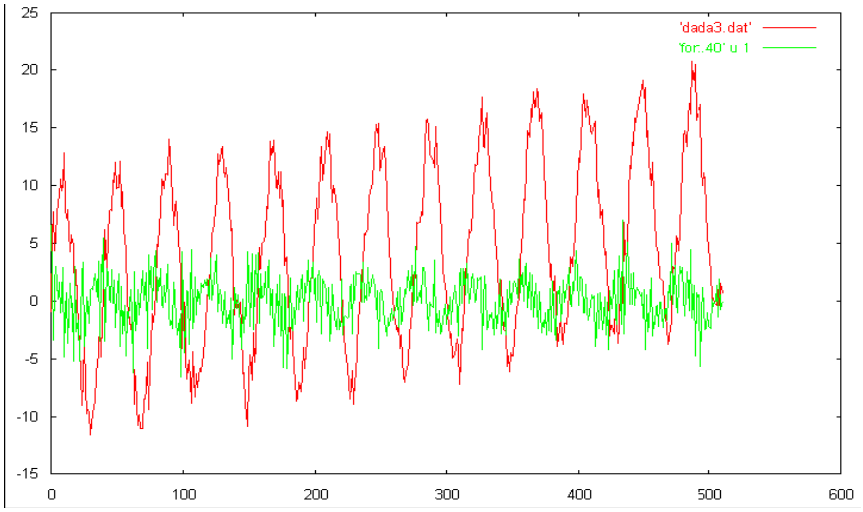
Η στατιστική αυτή ακολουθεί την χ^2 κατανομή με m βαθμούς ελευθερίας, τον αριθμό δηλαδή των χρονοκαθυστερήσεων. Η στατιστική αυτή εφαρμόζεται συνήθως στα κατάλοιπα (residuals).

Ο αριθμός των Lag, που παίρνει το MINITAB "By Default" είναι n/4 αν το n είναι μικρότερο του 240 και $\sqrt{n}+45$, *αν το n είναι μεγαλύτερο από 240, όπου n ο αριθμός των παρατηρήσεων της χρονοσειράς.*

3.7 ΜΕΘΟΔΟΣ ΤΩΝ ΔΙΑΦΟΡΩΝ

Μία μέθοδος ονομαζόμενη μέθοδος των διαφορών χρησιμοποιείται για την απομάκρυνση της τάσης από μη στάσιμα δεδομένα. Η μέθοδος των διαφορών είναι καλή για τον προσδιορισμό του θορύβου και την μετατροπή των μη στάσιμων χρονοσειρών σε στάσιμες. Δηλαδή, οι διαφορές $Y_t - Y_{t-1}$ αφαιρούν την τάση και αφήνουν τα κατάλοιπα.

Η χρονοσειρά των διαφορών είναι πιο μικρή καθώς χάνουμε ένα σημείο (το τελευταίο). Απόρροια του παραπάνω είναι, ότι μειώθηκε το πλάτος, αλλά αυξήθηκε ο θόρυβος.



Το πόσες διαφορές είναι αναγκαίες για την μετατροπή μιας μη στάσιμης χρονοσειράς σε στάσιμη εξαρτάται από τη συνάρτηση της τάσης. Επειδή το Δt είναι στις χρονοσειρές πάντα 1 το $Y_t - Y_{t-1}$ είναι κάτι σαν την αριθμητική παράγωγο. Αν για παράδειγμα έχουμε γραμμική τάση αρκεί η πρώτη παράγωγος (πρώτη διαφορά) για να γίνει η συνάρτηση σταθερά. Αν έχουμε δευτεροβάθμια τάση πρέπει να πάρουμε τη δεύτερη παράγωγο (δύο διαφορές) για να γίνει στάσιμη. Αν έχουμε εκθετική συνάρτηση, πρέπει να κάνουμε μετασχηματισμό, γιατί η

παράγωγος της εκθετικής συνάρτησης είναι η ίδια η συνάρτηση. Δηλαδή η εκθετική συνάρτηση δεν γίνεται ποτέ στάσιμη παίρνοντας διαφορές.

3.8 ΈΛΕΓΧΟΣ ΚΑΤΑΛΟΙΠΩΝ

Θεωρώντας τα κατάλοιπα ως τυχαίους αριθμούς δεν θα πρέπει να αυτοσυσχετίζονται. Ο έλεγχος των καταλοίπων γίνεται με τη στατιστική Q των Ljung & Box, με την οποία ελέγχεται η σημαντικότητα από κοινού ενός αριθμού συντελεστών αυτοσυσχέτισης.

3.9 ΈΛΕΓΧΟΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΣΥΝΤΕΛΕΣΤΩΝ ΑΥΤΟΣΥΣΧΕΤΙΣΗΣ

Α. Υποθέσεις που γίνονται δεκτές: Για κάθε συντελεστή ο έλεγχος σημαντικότητας βασίζεται στη στατιστική συνάρτηση:

$$t=\frac{r_1-\rho_1}{SE(r_1)} \quad \text{και} \quad t=\frac{r_2-\rho_2}{SE(r_2)} \quad \text{κ.λ.π,}$$

η οποία είναι η Student κατανομή με n-1 βαθμούς ελευθερίας, ή η κανονική κατανομή, επειδή ο πληθυσμός είναι κανονικός με γνωστή διασπορά ή αν το n-1 είναι μεγαλύτερο ή ίσο του 30.

Β. Η μηδενική και η εναλλακτική υπόθεση:

Με τον έλεγχο t ελέγχουμε τις υποθέσεις:

$H_0 : \rho_1 = 0$ με εναλλακτική $H_1 : \rho_1 \# 0$ και

$H_0 : \rho_2 = 0$ με εναλλακτική $H_1 : \rho_2 \# 0$

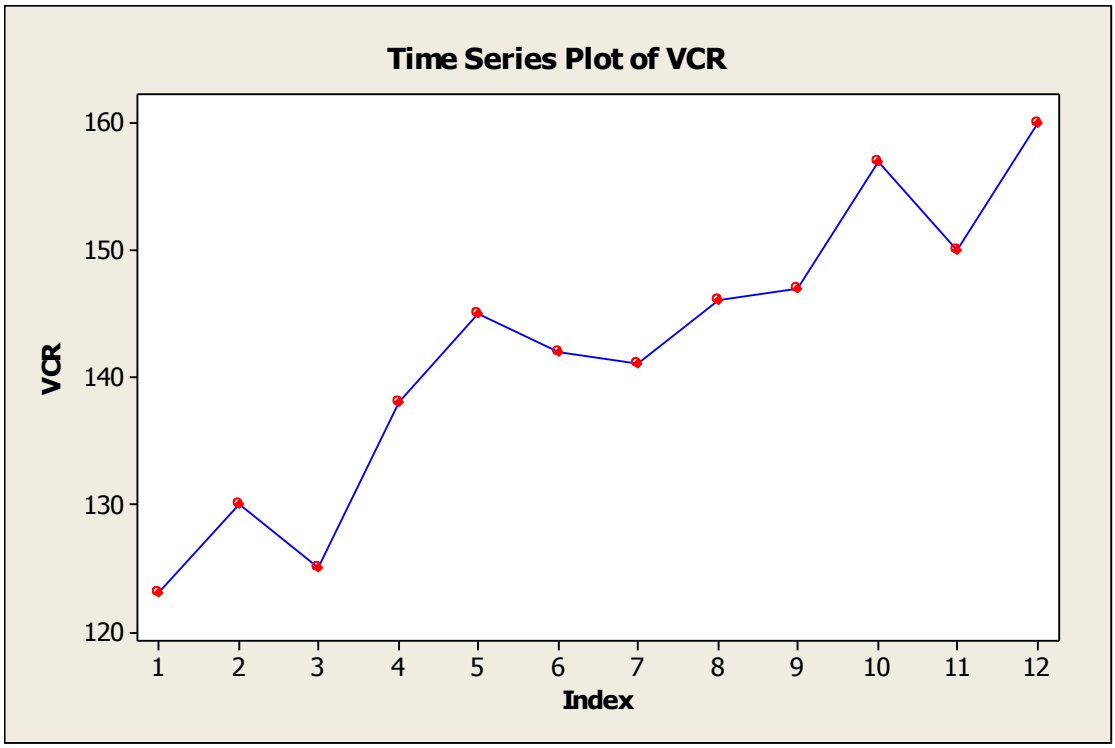
Γ. Το κριτήριο αποφάσεως: Για α = 0.05 και δίπλευρο έλεγχο έχουμε $t_{\frac{\alpha}{2}}=t_{0.025}$ και $t_{1-\frac{\alpha}{2}}=t_{0.975}$.

Δ. Απόφαση: Η απόφαση παίρνεται ανάλογα με τη τιμή της t σε

σχέση με το διάστημα ισχύος της μηδενικής υπόθεσης και τα όρια του διαστήματος.

Παράδειγμα (Τ3-1)

Πωλήσεις VCR για ένα ολόκληρο έτος κάποιας εταιρείας.



Time i	Month	Πωλήσεις	Καθυστέρηση 1	Διαφορές
1	Ιανουάριος	123		
2	Φεβρουάριος	130	123	7
3	Μάρτιος	125	130	-5
4	Απρίλιος	138	125	13
5	Μάϊος	145	138	7
6	Ιούνιος	142	145	-3
7	Ιούλιος	141	142	-1
8	Αύγουστος	146	141	5
9	Σεπέμβριος	147	146	1
10	Οκτώβριος	157	147	10
11	Νοέμβριος	150	157	-7

12	Δεκέμβριος	160	150	10

Συντελεστές αυτοσυσχέτισης της χρονοσειράς:

$$r_1=\frac{\sum_{t=1+1}^{n}(Y_t-\bar{Y})(Y_{t-1}-\bar{Y})}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2}=\frac{843}{1{,}474}=.572$$

$$r_2=\frac{\sum_{t=2+1}^{n}(Y_t-\bar{Y})(Y_{t-2}-\bar{Y})}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2}=\frac{682}{1{,}474}=.463$$

Εάν η μηδενική υπόθεση είναι αληθινή, τότε η στατιστική:

$$t=\frac{r_1-\rho_1}{SE(r_1)}=\frac{r_1-0}{SE(r_1)}=\frac{r_1}{SE(r_1)}$$

ακολουθεί την κατανομή Student με n-1=12-1=11 βαθμούς ελευθερίας και με 95% διάστημα εμπιστοσύνης θα είναι -2.2<t<2.2. Οπότε:

$$SE(r_1)=\sqrt{\frac{1}{12}}=.289$$

Με 95% διάστημα εμπιστοσύνης και εάν είναι t<-2.2 και t>2.2 απορρίπτουμε τη μηδενική υπόθεση. Είναι όμως:

$$t=\frac{r_1}{SE(r_1)}=\frac{.572}{.289}=1.98$$

που σημαίνει, ότι η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

Όμοια για τη χρονοκαθυστέρηση 2 έχουμε:

$$t=\frac{r_2-\rho_2}{SE(r_2)}=\frac{r_2-0}{SE(r_2)}=\frac{r_2}{SE(r_2)}$$

$$SE(r_k)=\sqrt{\frac{1+2\sum_{i=1}^{k-1}r_i^2}{n}}=\sqrt{\frac{1+2\sum_{i=1}^{2-1}r_i^2}{n}}=\sqrt{\frac{1+2(.572)^2}{12}}=.371$$

$$t=\frac{r_2}{SE(r_2)}=\frac{.463}{.371}=1.25$$

Επομένως δεν μπορεί να απορριφθεί ούτε η μηδενική υπόθεση για το lag 2 σε επίπεδο σημαντικότητας 0.05.

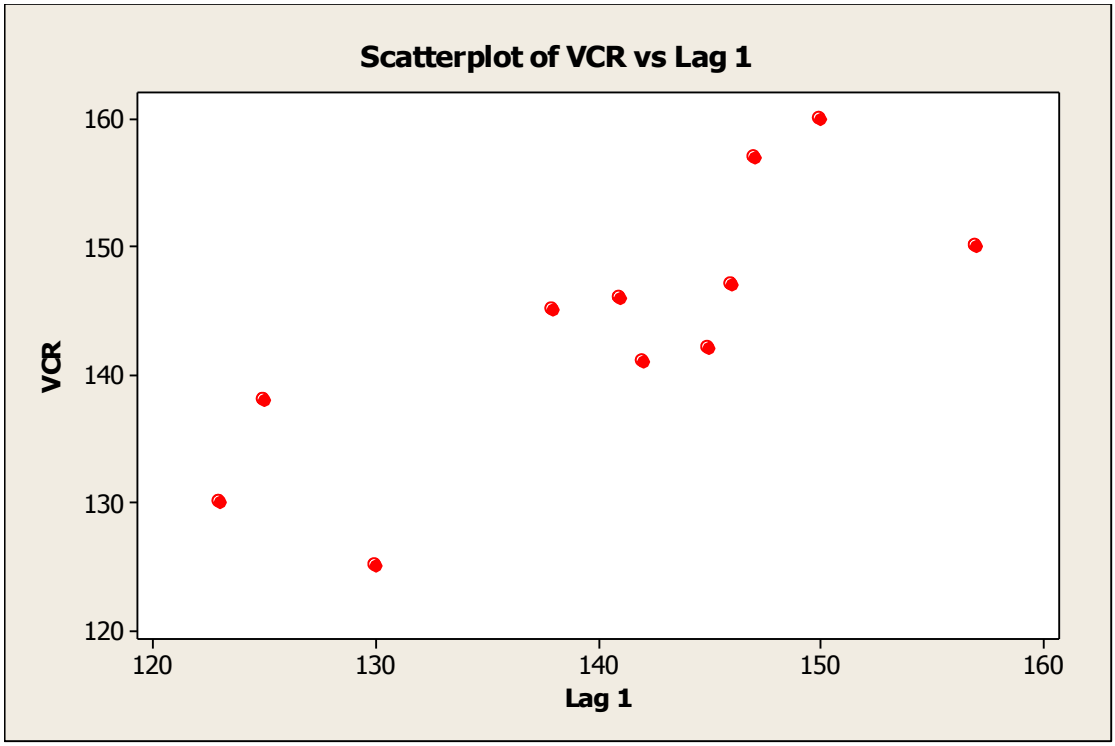
Ένας εναλλακτικός τρόπος για να ελέγξουμε τα όρια εμπιστοσύνης είναι να δημιουργήσουμε το διάστημα εμπιστοσύνης με κέντρο το 0.

Για τις αντίστοιχες χρονοκαθυστερήσεις θα έχουμε:

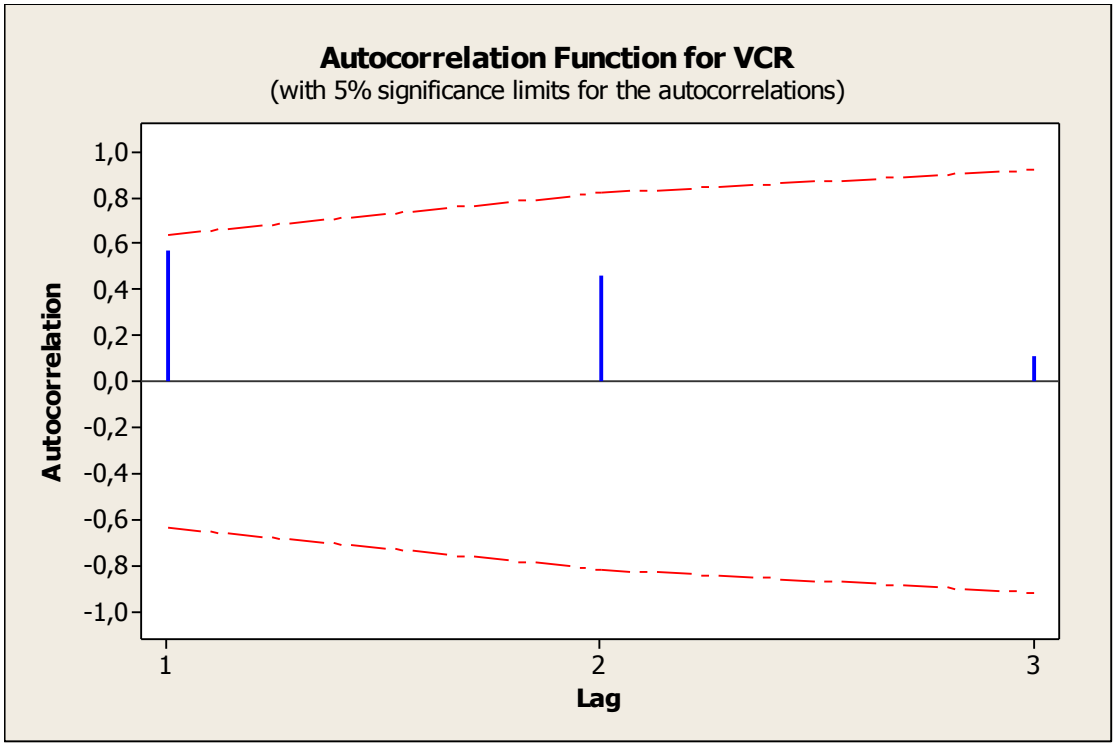
$$lag\,1: 0\pm t_{0.975}*SE(r_1) \;\; \acute{\eta} \;\; 0\pm 2.2*(0.289) \;\; \rightarrow \;\; (-0.636, 0.636)$$

$$lag\,2: 0\pm t_{0.975}*SE(r_2) \;\; \acute{\eta} \;\; 0\pm 2.2*(0.371) \;\; \rightarrow \;\; (-0.816, 0.816)$$

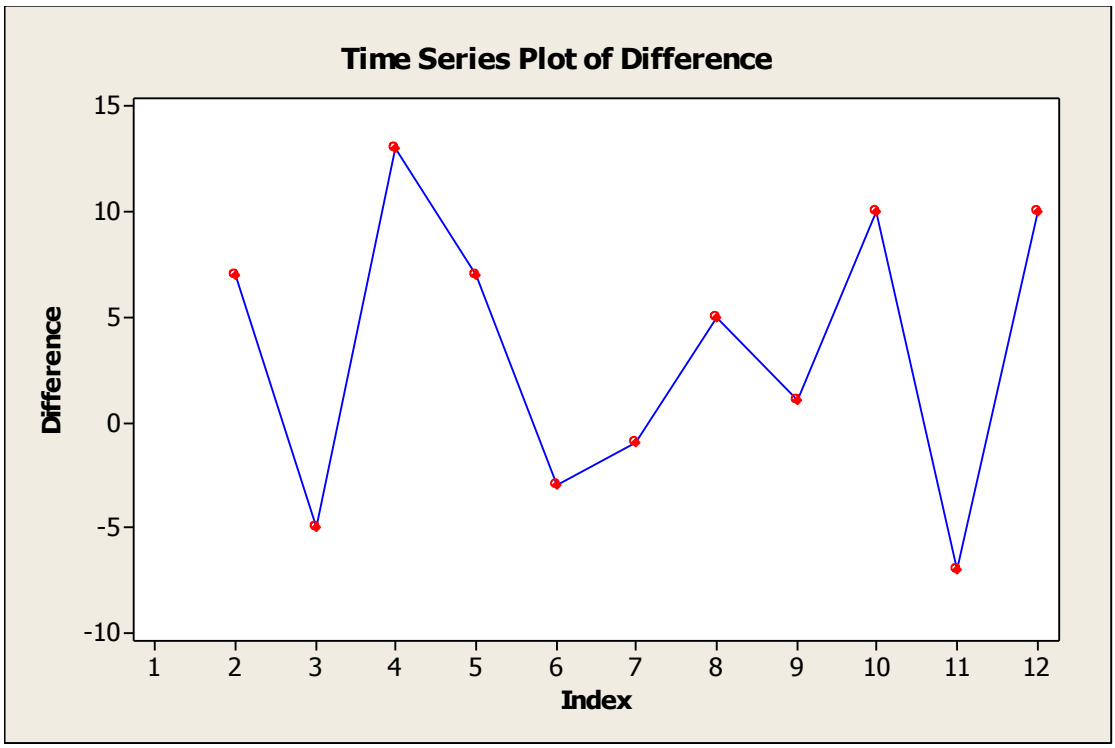
Το διάγραμμα διασποράς όμως ανάμεσα στην χρονοσειρά και το Lag 1 δείχνει κάποιο βαθμό συσχέτισης.



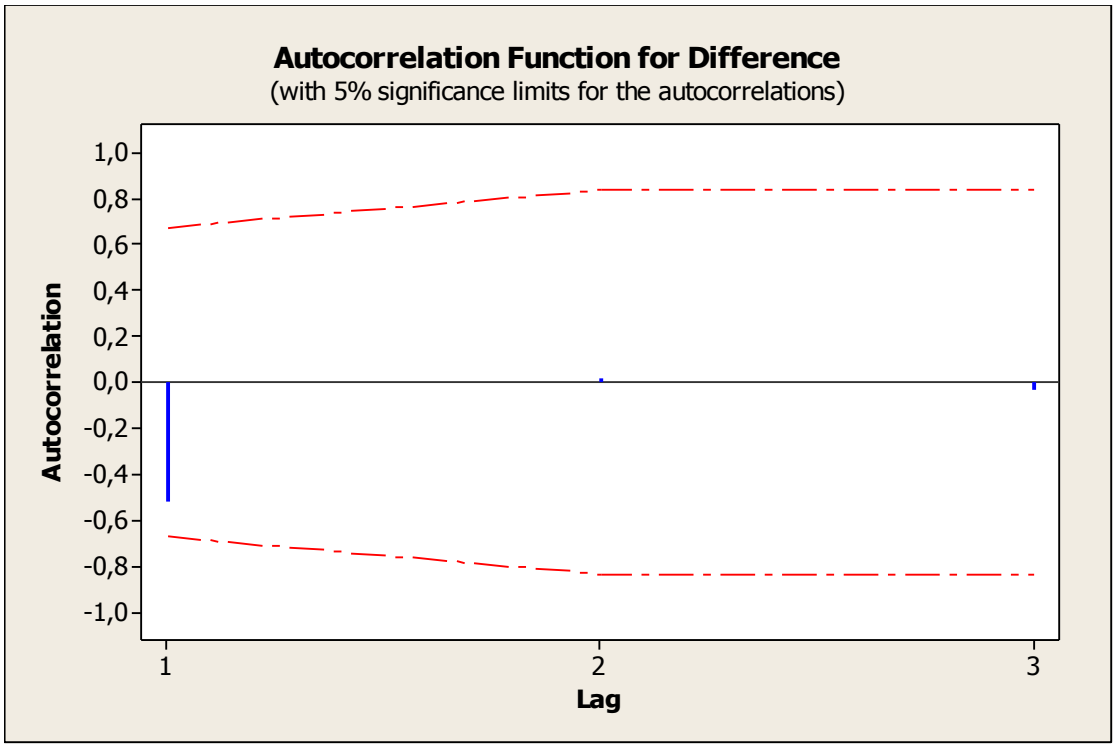
Είναι χρήσιμο λοιπόν να δούμε τι δείχνει και η Q στατιστική. Χρησιμοποιώντας την Q στατιστική παρακάτω υπολογίζουμε τον συντελεστή αυτοσυσχέτισης για την αρχική χρονοσειρά και τις διαφορές της στο Lag 1.



Lag	ACF	T	LBQ
1	0,571913	1,98	5,00
2	0,462687	1,25	8,59
3	0,110583	0,27	8,82



Η χ^2 με 3 βαθμούς ελευθερίας, με 95% διάστημα εμπιστοσύνης και μονόπλευρο έλεγχο είναι 7.81, που είναι μικρότερο από το 8.82. Επομένως η μηδενική υπόθεση απορρίπτεται.

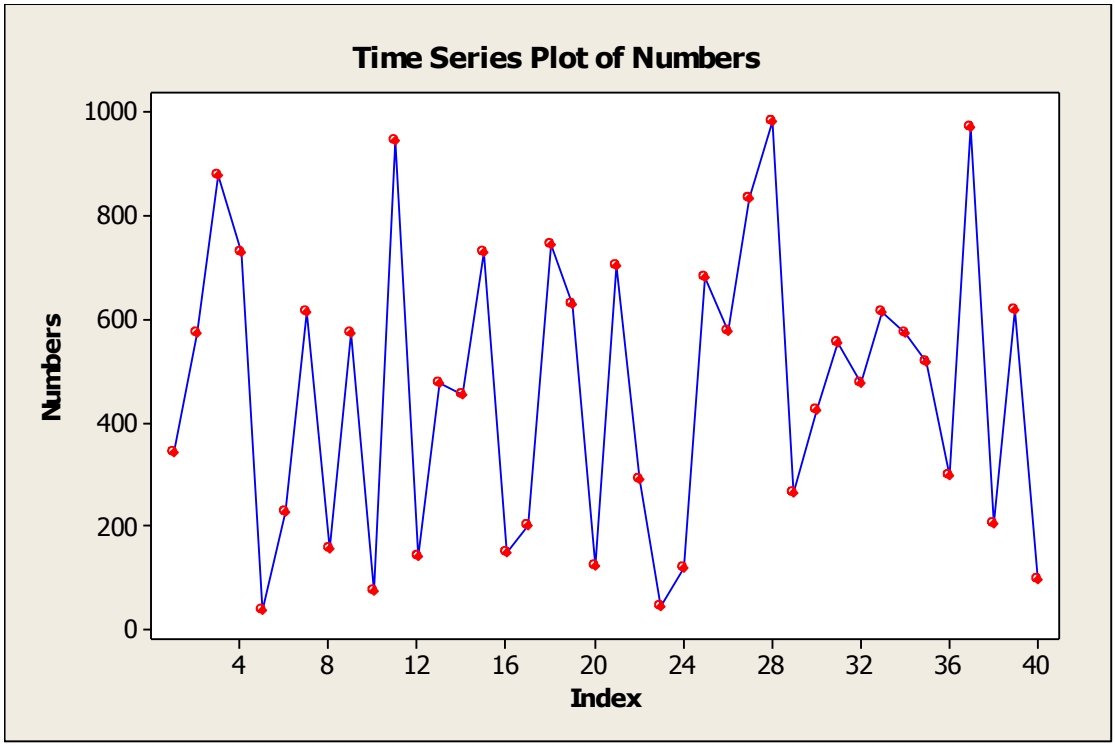


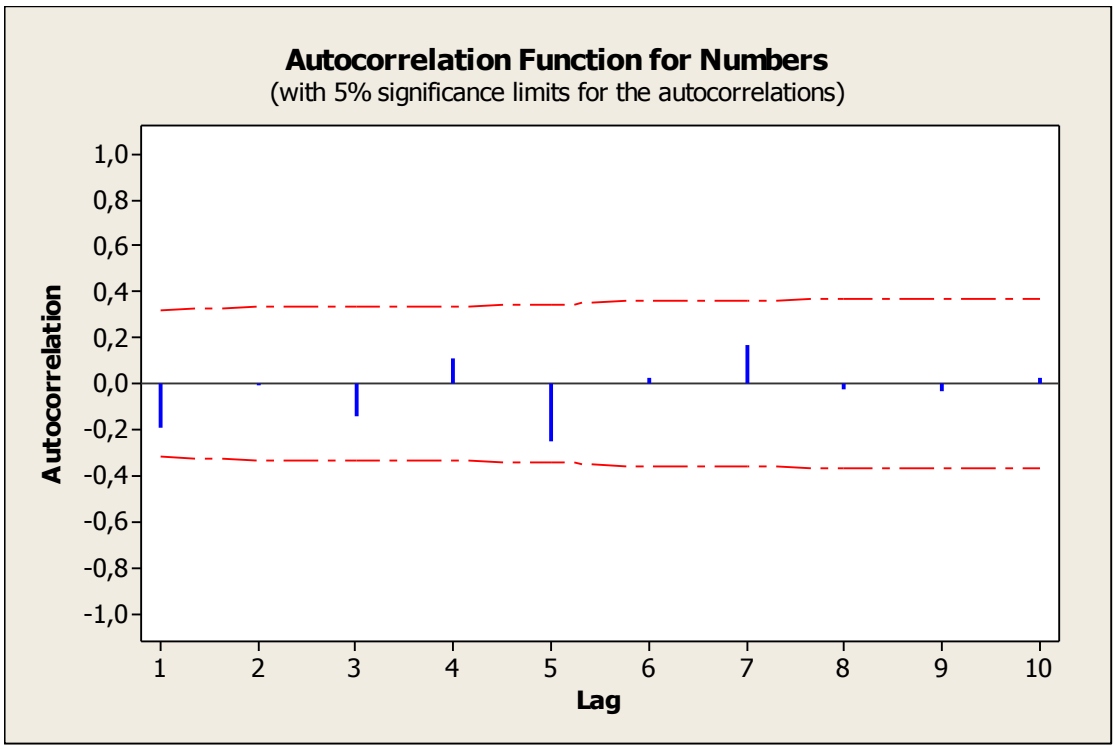
Lag	ACF	T	LBQ
1	-0,520581	-1,73	3,88
2	0,014884	0,04	3,88
3	-0,035830	-0,10	3,90

Υπολογίζοντας τους συντελεστές αυτοσυσχέτισης για τα κατάλοιπα, η χ^2 με 3 βαθμούς ελευθερίας είναι 7.81, το οποίο είναι μεγαλύτερο από το 3.90. Επομένως η μηδενική υπόθεση δεν απορρίπτεται.

Παράδειγμα (Τ3-3) (Τυχαιότητα) Αυτοσυσχέτιση 40 τριψήφιων ψευδοτυχαίων αριθμών.

343, 574, 879, 728, 037, 227, 613, 157, 571, 072, 946, 142, 477, 452, 727, 147, 199, 744, 627, 122, 704, 291, 043, 118, 682, 577, 834, 981, 263, 424, 555, 476, 612, 574, 518, 296, 970, 204, 616, 097.





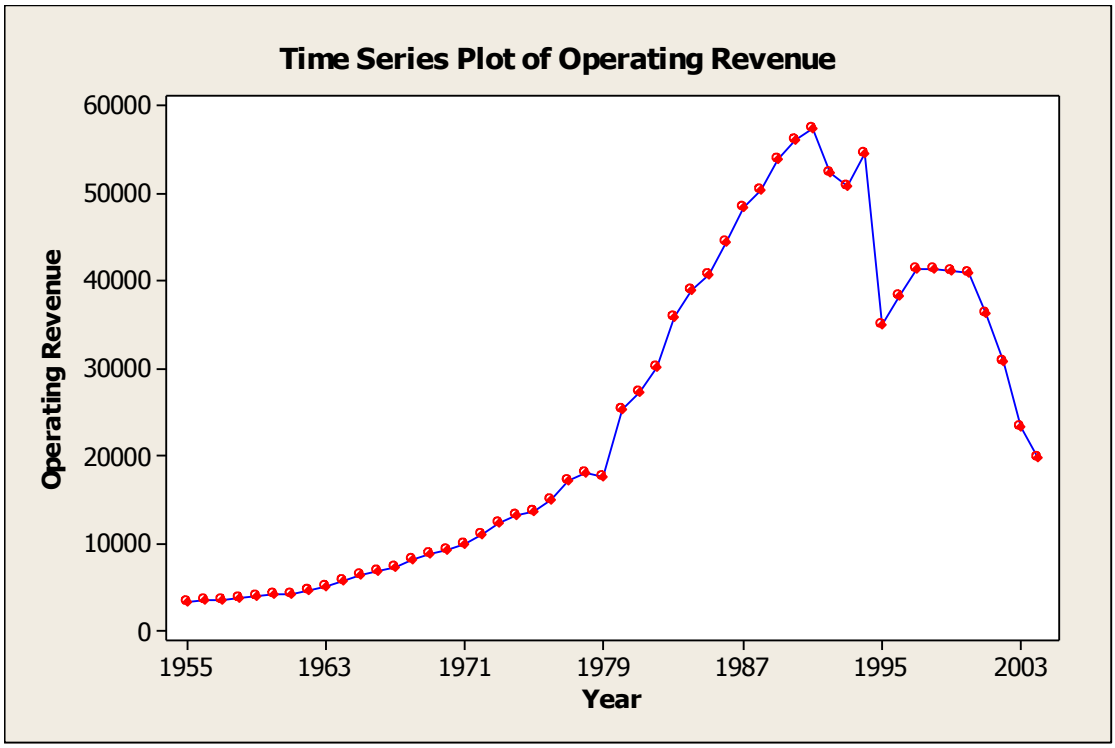
Lag	ACF	T	LBQ
1	-0,191171	-1,21	1,57
2	-0,006293	-0,04	1,58
3	-0,145224	-0,89	2,53
4	0,104475	0,63	3,04

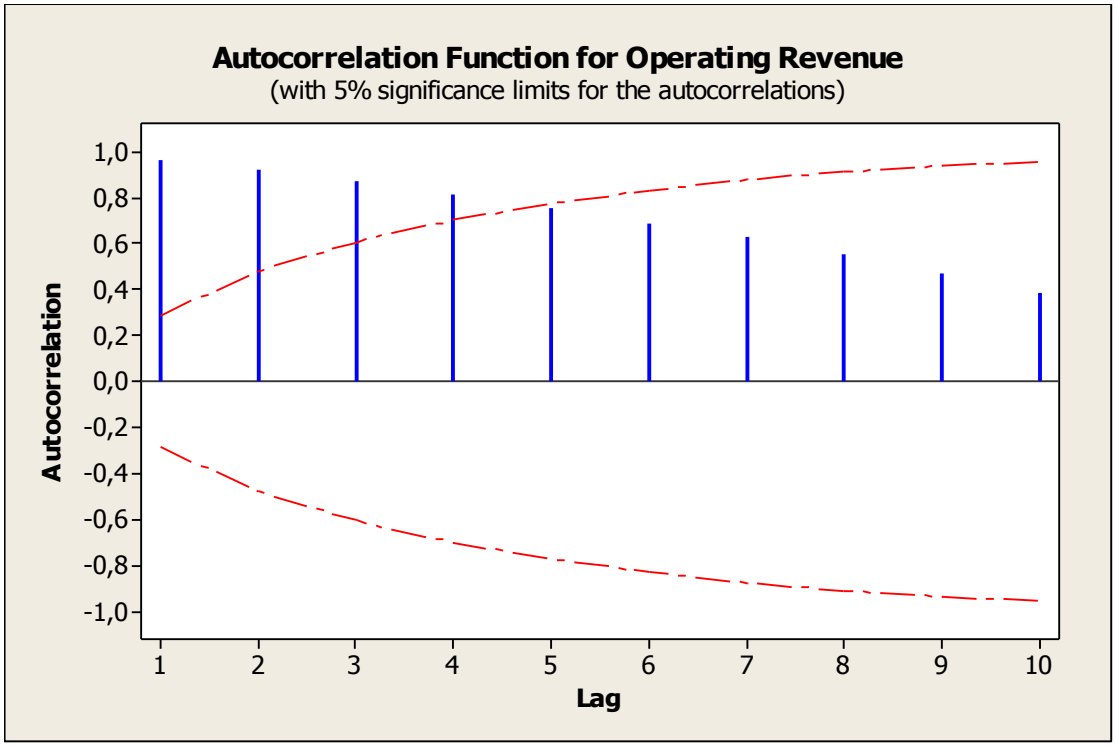
```
 5   -0,253756   -1,50   6,13
 6    0,028640    0,16   6,17
 7    0,169071    0,95   7,63
 8   -0,027390   -0,15   7,67
 9   -0,032551   -0,18   7,73
10    0,021546    0,12   7,75
```

- m = 10 (lags)
- n = 40

Επειδή η LBQ ακολουθεί την χ^2 στατιστική με m=10 βαθμούς ελευθερίας, η τιμή LBQ πρέπει να είναι μικρότερη από την τιμή 18.31 για 95% επίπεδο εμπιστοσύνης. Επειδή το 7.75 είναι μικρότερο του 18.31 δεν απορρίπτουμε τη μηδενική υπόθεση, που σημαίνει, ότι τα δεδομένα είναι τυχαία χωρίς εμφανή τάση.

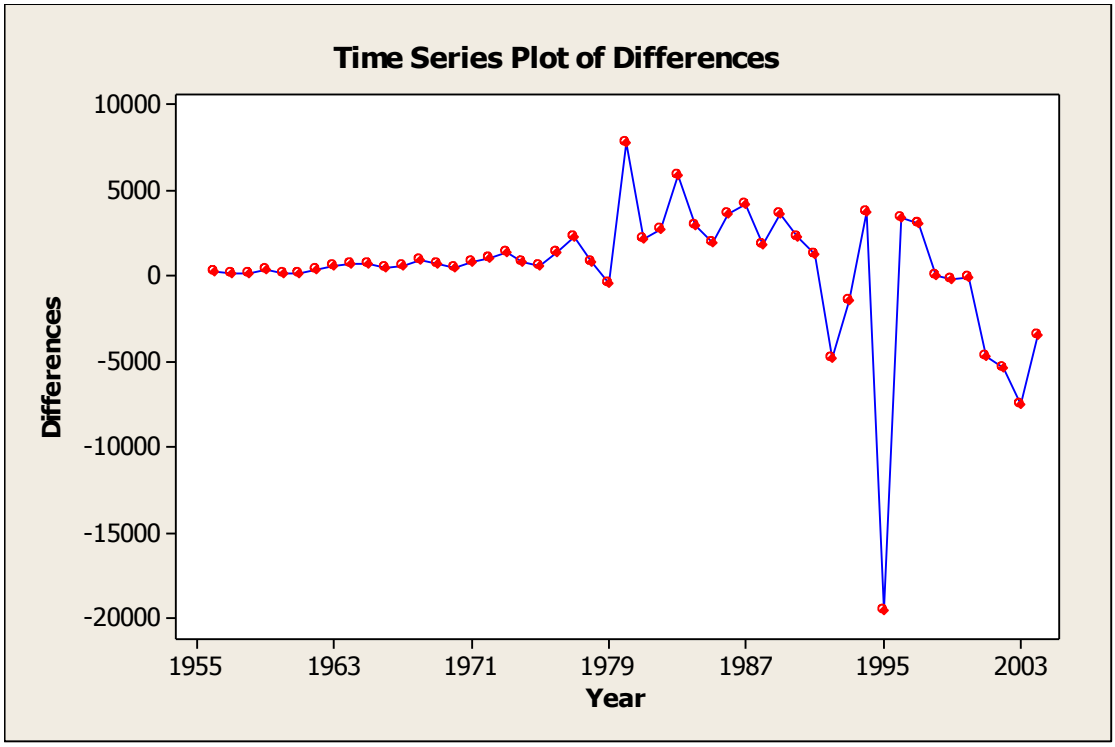
Παράδειγμα (Τ3-4) (Τάση) Ετήσια έσοδα μιας εταιρείας.

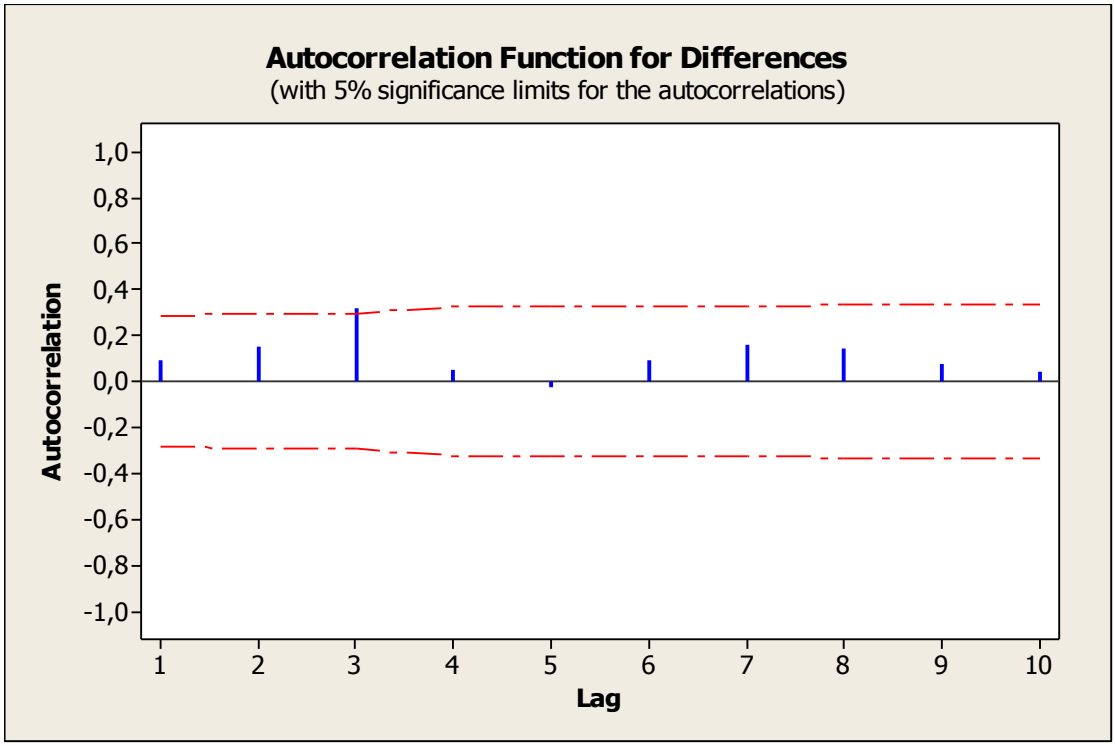




Lag	ACF	T	LBQ
1	0,961601	6,80	49,06
2	0,919842	3,85	94,90
3	0,872940	2,90	137,05
4	0,812378	2,33	174,35
5	0,750592	1,95	206,90
6	0,689363	1,67	234,98
7	0,623393	1,43	258,48
8	0,549571	1,21	277,18
9	0,468019	1,00	291,07
10	0,382058	0,80	300,56

Επειδή η χ^2 κατανομή με 10 βαθμούς ελευθερίας έχει τιμή 18.31 πολύ μικρότερη από την τιμή 300.56, η μηδενική υπόθεση απορρίπτεται. Οι συντελεστές αυτοσυσχέτισης δείχνουν στατιστικά σημαντικά διάφοροι του μηδενός και ο κάθε ένας ξεχωριστά και σαν ομάδα.



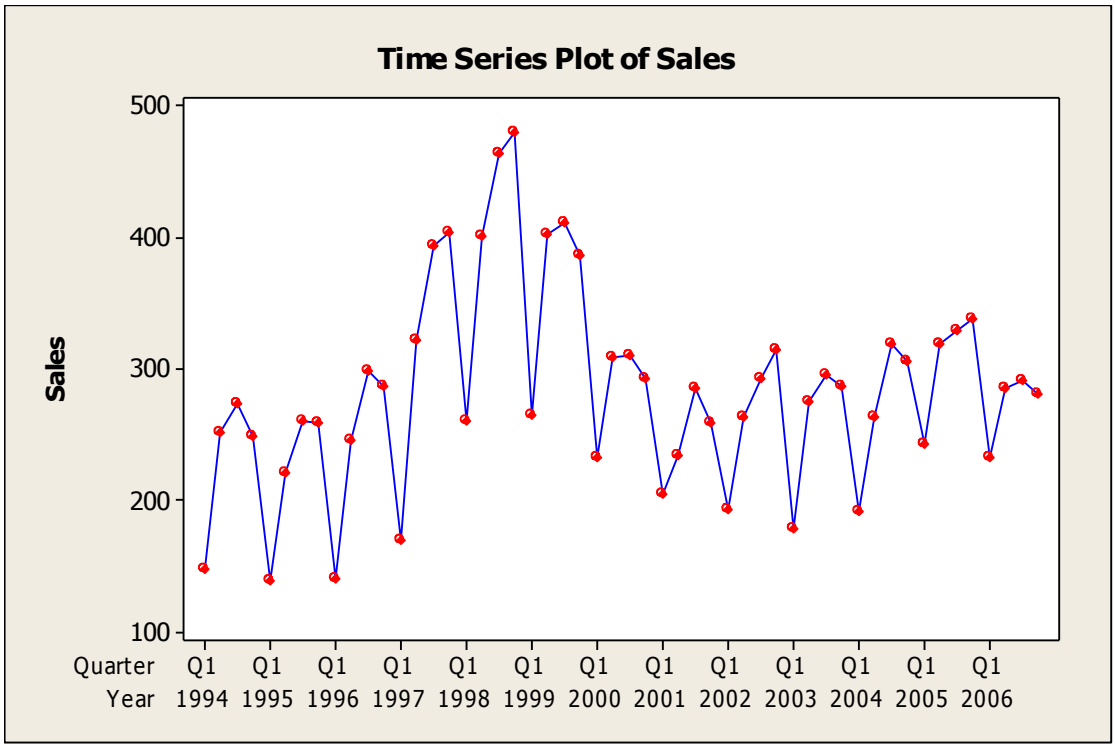


Lag	ACF	T	LBQ
1	0,093467	0,65	0,45
2	0,148351	1,03	1,62
3	0,316708	2,15	7,07
4	0,049511	0,31	7,21

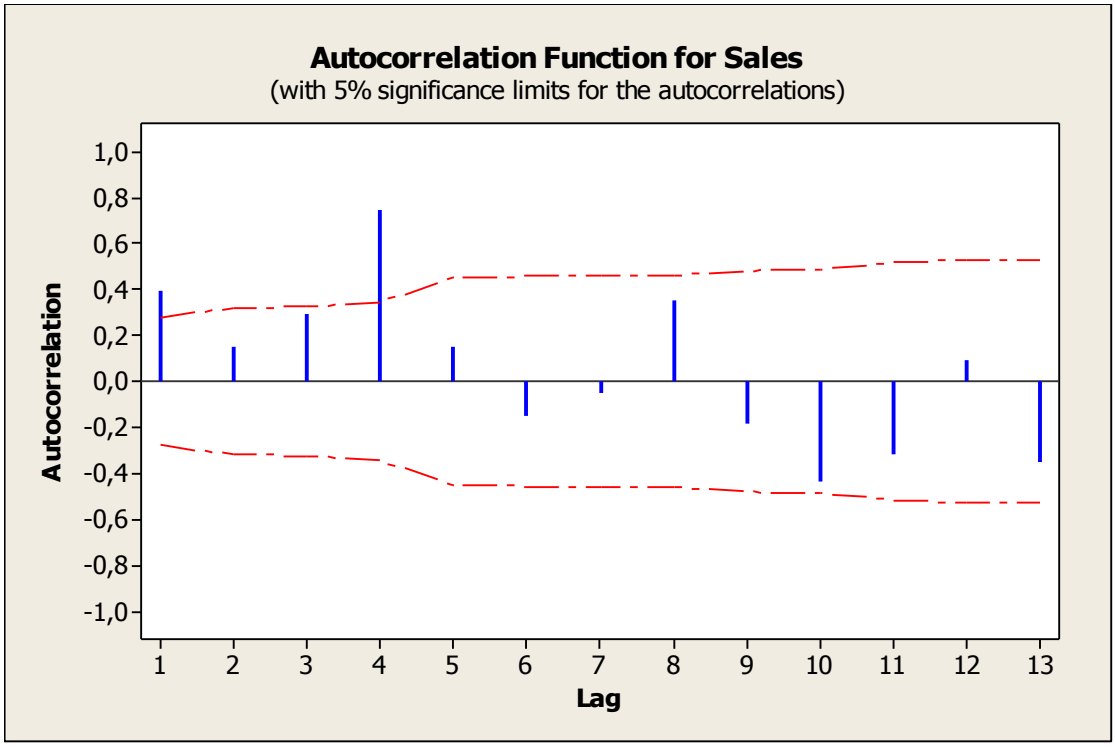
5	-0,025610	-0,16	7,25
6	0,089635	0,56	7,71
7	0,159533	0,99	9,23
8	0,144327	0,87	10,50
9	0,076160	0,45	10,86
10	0,044129	0,26	10,99

Επειδή η χ^2 κατανομή με 10 βαθμούς ελευθερίας έχει τιμή 18.31 μεγαλύτερο από την τιμή 10,99, η μηδενική υπόθεση δεν απορρίπτεται και μπορούμε να θεωρήσουμε, ότι η χρονοσειρά των διαφορών είναι στάσιμη και οι συντελεστές αυτοσυσχέτισης στατιστικά μηδέν.

Παράδειγμα (Τ3-5) (Εποχικότητα) Πωλήσεις από 1994 – 2006 μιας εταιρείας.



Στην προκείμενη περίπτωση συμπεραίνουμε, ότι υπάρχει εποχικότητα των δεδομένων βλέποντας τον συντελεστή αυτοσυσχέτισης στα lag 1, 4 να υπερβαίνει το διάστημα εμπιστοσύνης.



Lag	ACF	T	LBQ
1	0,392843	2,83	8,50
2	0,153948	0,97	9,83
3	0,293739	1,82	14,77
4	0,743533	4,34	47,11
5	0,150619	0,67	48,47
6	-0,153014	-0,67	49,90
7	-0,047049	-0,21	50,04
8	0,346979	1,51	57,72
9	-0,182597	-0,76	59,90
10	-0,434737	-1,80	72,53
11	-0,315090	-1,23	79,33
12	0,091175	0,35	79,92
13	-0,353272	-1,34	88,90

Στην προκείμενη περίπτωση προκειμένου να γίνει η χρονοσειρά στάσιμη θα χρειαστούμε δύο διαφορές μία κανονική και μία εποχική.

3.10 ΒΑΣΙΚΑ ΣΤΑΔΙΑ ΣΤΗ ΔΙΑΔΙΚΑΣΙΑ ΠΡΟΒΛΕΨΗΣ

Τα βασικά στάδια σε μια διαδικασία πρόβλεψης είναι τα εξής:

I. Καθορισμός Προβλήματος και Συγκέντρωση Πρώτων Δεδομένων (Problem Definition και Gathering Information)

Τις περισσότερες φορές το πιο δύσκολο μέρος στη διαδικασία πρόβλεψης και ταυτόχρονα το πιο σημαντικό είναι να καθορίσουμε το πρόβλημα, το πώς θα χρησιμοποιηθούν οι προβλέψεις, από ποιους και ποια δεδομένα θα χρησιμοποιηθούν.

II. Διαχείριση Πληροφοριών και Τελική Επιλογή

Σε αυτό το βήμα απαιτούνται τουλάχιστον δύο πράγματα. Το πρώτο είναι τα αριθμητικά δεδομένα και το δεύτερο η κρίση, η πείρα και η εμπειρία του προσωπικού, που ασχολήθηκαν με τη συλλογή των στοιχείων – δεδομένων για αυτό το χρονικό διάστημα. Επίσης για τις παραπάνω πληροφορίες πρέπει να γίνει η τελική επιλογή, πριν ξεκινήσει η διαδικασία της πρόβλεψης.

III. Προκαταρτική Ανάλυση (Exploratory Analysis)

Στο βήμα αυτό μας απασχολεί το είδος της πληροφορίας που αποκομίζουμε από τα ακατέργαστα ιστορικά δεδομένα. Αρχικά, αναπαριστούμε γραφικά τα δεδομένα και στη συνέχεια, υπολογίζουμε κάποιους βασικούς στατιστικούς δείκτες, όπως η μέση τιμή, η τυπική απόκλιση, ελάχιστο, μέγιστο και τάση. Οι παραπάνω δείκτες αναδεικνύουν κάποια δευτερεύοντα χαρακτηριστικά της χρονοσειράς.

Σκοπός μας είναι να αποκτήσουμε μία αίσθηση των δεδομένων, δίνοντας απαντήσεις σε ερωτήματα, όπως αν υπάρχουν λανθάνοντα πρότυπα, αν υπάρχει σημαντική τάση, ή εποχικότητα και τέλος, αν υπάρχουν ακραίες – ασυνήθιστες τιμές (outliers). Η ανάλυση αυτή μας οδηγεί στην οικογένεια μοντέλων πρόβλεψης, που λογικά αναμένεται να δώσει ικανοποιητικές προβλέψεις.

IV. Επιλογή και Προσαρμογή Μοντέλου (Choosing & Fitting models)

Εδώ γίνεται η επιλογή και ο καθορισμός των παραμέτρων διάφορων ποσοτικών μοντέλων πρόβλεψης, που έχουν επιλεγεί στο προηγούμενο βήμα.

V. Χρήση και αποτίμηση του μοντέλου πρόβλεψης (Using and Evaluating a forecasting model)

Στο τελικό στάδιο, αφού ένα μοντέλο έχει επιλεγεί και έχουν καθοριστεί οι παράμετροι του, παράγονται οι προβλέψεις. Κατά την εξέλιξη της διαδικασίας, γίνεται αποτίμηση των πλεονεκτημάτων και μειονεκτημάτων του μοντέλου και εφόσον κριθεί απαραίτητο, επαναλαμβάνονται κάποια βήματα στη διαδικασία.

3.12 Η ΕΠΙΛΟΓΗ ΜΙΑΣ ΤΕΧΝΙΚΗΣ ΠΡΟΒΛΕΨΗΣ

Ορισμένα από τα ερωτήματα, που πρέπει να ληφθούν υπόψη, προτού αποφασιστεί η κατάλληλη τεχνική πρόβλεψης για ένα συγκεκριμένο πρόβλημα, είναι τα ακόλουθα:

- Γιατί χρειάζεται η πρόβλεψη;
- Ποιος θα χρησιμοποιήσει την πρόβλεψη;
- Ποια είναι τα χαρακτηριστικά των διαθέσιμων δεδομένων;
- Ποια χρονική περίοδο θα γίνει η πρόβλεψη;
- Ποιες είναι οι ελάχιστες απαιτήσεις σε δεδομένα;
- Πόση ακρίβεια είναι επιθυμητή;
- Ποιο είναι το κόστος της πρόβλεψης;

Για να γίνει η επιλογή της κατάλληλης τεχνικής πρόβλεψης σωστά, ο αναλυτής πρέπει να είναι σε θέση να εκτελέσει τα ακόλουθα:

- Να καθορίζει τη φύση του προβλήματος.
- Να εξηγήσει τη φύση των δεδομένων, που αποτελούν αντικείμενο

της έρευνας.

- Να περιγράψει τις δυνατότητες και τους περιορισμούς των δυνητικών τεχνικών πρόβλεψης.
- Να αναπτύξει κάποια προκαθορισμένα κριτήρια.

Ανά είδος δεδομένων ακολουθούμε τις ακόλουθες μεθόδους πρόβλεψης:

Τεχνικές πρόβλεψης για στάσιμα δεδομένα

- Όταν οι συνθήκες, που δημιουργούν τις χρονοσειρές, έχουν σταθεροποιηθεί και το περιβάλλον δημιουργίας παραμένει αμετάβλητο. Π.χ. Ο αριθμός πτώσεων ενός συστήματος γραμμής παραγωγής με ομοιόμορφη παραγωγή, ο αριθμός πωλήσεων ενός προϊόντος στη φάση ωρίμανσης.
- Όταν οι χρονοσειρές μπορούν να μετατραπούν σε σταθερές π.χ. παίρνοντας τον λογάριθμο, ή τη τετραγωνική ρίζα, ή τις διαφορές των δεδομένων.

Σ' αυτή τη περίπτωση χρησιμοποιούμε τη naïve μέθοδο, απλό ή κινητό μέσο όρο, AutoRegressive Moving Average (ARMA) μοντέλα.

Τεχνικές πρόβλεψης για δεδομένα με τάση

- Όταν η αύξηση της παραγωγικότητας και εισαγωγή νέων τεχνολογιών μπορεί να οδηγήσει σε αλλαγές του τρόπου ζωής. Π.χ. Ζήτηση ηλεκτρονικών λόγω αύξησης πωλήσεων Η.Υ., μείωση της χρήσης σιδηροδρόμων, λόγω αύξησης της χρήσης του αεροπλάνου.
- Όταν η αύξηση του πληθυσμού οδηγεί σε αύξηση της ζήτησης των αγαθών και υπηρεσιών. Π.χ. Αύξηση ζήτησης καταναλωτικών αγαθών, αύξηση ζήτησης ενέργειας, η περίοδος αύξησης των

πωλήσεων ενός νέου προϊόντος κατά τα αρχικά στάδια του κύκλου ζωής του.

Σ' αυτή τη περίπτωση χρησιμοποιούμε κινητό μέσο όρο, παλινδρόμηση, εκθετικές εξομαλύνσεις, AutoRegressive Integrated Moving Average (ARIMA) μοντέλα, μοντέλα Gompertz και Bass.

Τεχνικές πρόβλεψης για εποχικά δεδομένα

- Όταν ο καιρός μεταβάλλει το ενδιαφέρον. Π.χ. Κατανάλωση ενέργειας το χειμώνα, ή το καλοκαίρι, ρουχισμός, τρόφιμα, τουρισμός.
- Όταν κάποιες ημέρες μεταβάλλουν το ενδιαφέρον. Π.χ. Αύξηση πωλήσεων κατά τη διάρκεια των διακοπών, ή Σαββατοκύριακων.

Σ' αυτή τη περίπτωση χρησιμοποιούμε Winter – εκθετική εξομάλυνση, Census X12, AutoRegressive Integrated Moving Average (ARIMA) μοντέλα.

Τεχνικές πρόβλεψης για κυκλικά δεδομένα

- Αλλαγές στο Product Life Cycle: Εισαγωγή, αύξηση, ωρίμανση, κορεσμός, πτώση.
- Αλλαγές στον ίδιο το πληθυσμό: Πόλεμοι, φυσικές καταστροφές, επιδημίες.
- Αλλαγές στις προτιμήσεις του πληθυσμού: Μουσική, τροφή, μόδα.

Σ' αυτή τη περίπτωση χρησιμοποιούμε οικονομικούς δείκτες, οικονομετρικά μοντέλα, AutoRegressive Integrated Moving Average (ARIMA) μοντέλα.

Προς θεώρηση παράγοντες για την επιλογή μεθόδου πρόβλεψης

Ο χρονικός ορίζοντας είναι σημαντικός παράγοντας για την επιλογή μιας μεθόδου πρόβλεψης. Για βραχυπρόθεσμες και

μεσοπρόθεσμες προβλέψεις υπάρχουν αρκετές ποσοτικές μέθοδοι. Όταν όμως ο χρονικός ορίζοντας αυξάνει, τότε πολλές μέθοδοι γίνονται λιγότερο ακριβείς. Για παράδειγμα δεν μπορούμε να χρησιμοποιούμε ARIMA μεθόδους, ή μεθόδους εξομάλυνσης για μακροπρόθεσμες προβλέψεις. Σε αυτές τις περιπτώσεις πρέπει να χρησιμοποιούμε οικονομετρικές προβλέψεις. Η πολυμεταβλητή παλινδρόμηση δίνει καλύτερες προβλέψεις, όπως επίσης και οι ποιοτικές μέθοδοι. Όλα όμως εξαρτώνται από την εμπειρία, τη γνώση και τη κρίση του αναλυτή.

TABLE 3-6 Choosing a Forecasting Technique

Method	*Pattern of Data*	*Time Horizon*	*Type of Model*	*Minimal Data Requirement*	
				Nonseasonal	*Seasonal*
Naive	ST, T, S	S	TS	1	
Simple averages	ST	S	TS	30	
Moving averages	ST	S	TS	4–20	
Exponential smoothing	ST	S	TS	2	
Linear exponential smoothing	T	S	TS	3	
Quadratic exponential smoothing	T	S	TS	4	
Seasonal exponential smoothing	S	S	TS		2×s
Adaptive filtering	S	S	TS		5×s
Simple regression	T	I	C	10	
Multiple regression	C, S	I	C	10×V	
Classical decomposition	S	S	TS		5×s
Exponential trend models	T	I, L	TS	10	
S-curve fitting	T	I, L	TS	10	
Gompertz models	T	I, L	TS	10	
Growth curves	T	I, L	TS	10	
Census *X*-12	S	S	TS		6×s
Box-Jenkins	ST, T, C, S	S	TS	24	3×s
Leading indicators	C	S	C	24	
Econometric models	C	S	C	30	
Time series multiple regression	T, S	I, L	C		6×s

Pattern of the data: ST, stationary; T, trending; S, seasonal; C, cyclical
Time horizon: S, short term (less than three months); I, intermediate term; L, long term
Type of model: TS, time series; C, causal
Seasonal: s, length of seasonality
Variable: V, number of variables

Μετά από μία εκτενή ανάλυση των μεθόδων βγήκαν τα εξής συμπεράσματα:

- Οι περίπλοκες μέθοδοι δεν δίνουν αναγκαστικά και καλύτερα

αποτελέσματα.

- Κάποιες μέθοδοι δίνουν καλύτερα αποτελέσματα για βραχυπρόθεσμες προβλέψεις και κάποιες άλλες για μακροπρόθεσμες. Κάποιες άλλες δίνουν καλύτερα αποτελέσματα για ετήσια δεδομένα και κάποιες άλλες για τριμηνιαία, ή μηνιαία δεδομένα.
- Ο συνδυασμός τριών μεθόδων εξομάλυνσης δίνουν πολλές φορές καλύτερα αποτελέσματα από άλλες μεθόδους.

3.13 ΑΞΙΟΛΟΓΗΣΗ ΠΡΟΒΛΕΨΕΩΝ

Ένα σπουδαίο ρόλο στην μελέτη χρονοσειρών, γενικώς, παίζουν τα *κατάλοιπα* (residuals). Το κατάλοιπο είναι η διαφορά ανάμεσα στο πραγματικό αποτέλεσμα και την πρόβλεψη του αποτελέσματος, δηλαδή, η διαφορά των εκτιμήσεων από τις παρατηρήσεις και υπολογίζεται ως εξής:

$$e_t = Y_t - \hat{Y}_t$$

όπου,

- e_t = το σφάλμα πρόβλεψης τη χρονική στιγμή t.
- Y_t = η πραγματική τιμή τη χρονική στιγμή t.
- $\hat{Y}_t$ = η προβλεπόμενη τιμή τη χρονική στιγμή t.

Μπορούμε να συγκρίνουμε τις προβλεπόμενες με τις πραγματικές τιμές για την περίοδο, για την οποία έχουμε στοιχεία. Για τη σύγκριση αυτή, δηλαδή για την αξιολόγηση της προβλεπτικής ικανότητας του υποδείγματος, έχουν προταθεί διάφορα κριτήρια, από τα οποία τα κυριότερα είναι:

Ρίζα του Μέσου του Τετραγώνου του Σφάλματος (Root Mean

Square Error, RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}$$

όπου,

- Y_t = η πραγματική τιμή τη χρονική στιγμή t.
- $\hat{Y}_t$ = η προβλεπόμενη τιμή τη χρονική στιγμή t.
- n = αριθμός χρονικών περιόδων

Μέση Απόλυτη Απόκλιση (Mean Absolute Deviation, MAD):

$$MAD = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \hat{Y}_t|$$

Μέσο Τετραγωνικό Σφάλμα (Mean Square Error, MSE):

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2$$

Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (Mean Absolute Percentage Error, MAPE):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \hat{Y}_t|}{Y_t}$$

Μέσο Ποσοστιαίο Σφάλμα (Mean Percentage Error , MPE):

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{(Y_t - \hat{Y}_t)}{Y_t}$$

Παράδειγμα

Ο παρακάτω πίνακας παρουσιάζει τα στοιχεία με τον καθημερινό αριθμό των πελατών, που ζητάνε επισκευές, Y_t και μια πρόβλεψη $\hat{Y}_t$ για αυτά τα δεδομένα, σε μια επιχείρηση.

Time	Y_t	$\hat{Y}_t$	e_t	$\lvert e_t \rvert$	e_t^2	$\lvert e_t \rvert / Y_t$	e_t / Y_t
1	58	-	-	-	-	-	-

2	54	58	-4	4	16	.074	-.074
3	60	54	6	6	36	.100	.100
4	55	60	-5	5	25	.091	-.091
5	62	55	7	7	49	.113	.113
6	62	62	0	0	0	.000	.000
7	65	62	3	3	9	.046	.046
8	63	65	-2	2	4	.032	-.032
9	70	63	7	7	49	.100	.100
		Totals	12	34	188	.556	.162

Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2} = \sqrt{\frac{188}{8}} = 4.8$$

Mean Absolute Deviation

$$MAD = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \hat{Y}_t| = \frac{34}{8} = 4.3$$

Mean Squared Error

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 = \frac{188}{8} = 23.5$$

Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \hat{Y}_t|}{Y_t} = \frac{0.556}{8} = 0.0695 = 6.95 \ \%$$

Mean Percentage Error

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{(Y_t - \hat{Y}_t)}{Y_t} = \frac{0.162}{8} = 0.0203 = 2.03 \ \%$$

Το MAD μας λέει ότι κατά μέσον όρο θα έχουμε απόκλιση 4.3 πελάτες ανά πρόβλεψη. Τα MSE και MAPE συγκρίνονται με τα αντίστοιχα άλλων μεθόδων, προκειμένου να επιλέξουμε τη πιο ακριβή μέθοδο.

3.14 CASE 3-1A MURPHY BROTHERS FURNITURE

Το 1958 οι αδερφοί Murphy άνοιξαν ένα κατάστημα επίπλων στο κέντρο της πόλης του Dallas. Η κίνηση τους ήταν αρκετά επιτυχημένη και άνοιξαν δύο ακόμη καταστήματα (1. West 2. Midwest). Έως το 1996 η αλυσίδα καταστημάτων τους εξαπλώθηκε σε 36 πολιτείες.

Η κόρη ενός εκ των δύο ιδρυτών της επιχείρησης (Julie Murphy) συμμετείχε στην επιχείρηση. Ο πατέρας της και ο θείος της είχαν εντρυφήσει σε πολλές έννοιες εκτός από τον τομέα των ποσοτικών δεξιοτήτων. Κατάλαβαν ότι δεν θα μπορούσαν να κάνουν ακριβείς προβλέψεις για τις μελλοντικές πωλήσεις των επίπλων τους χρησιμοποιώντας μοντέρνες υπολογιστικές τεχνικές. Για το λόγο αυτό, ανέθεσαν στη Julie τη συγκεκριμένη δουλειά.

Η Julie, αρχικά θεώρησε ότι θα μπορούσε να χρησιμοποιήσει ως μεταβλητή, τις πωλήσεις της εταιρείας σε δολάρια. Ωστόσο, παρατήρησε ότι τα ιστορικά στοιχεία για ορισμένα έτη είχαν χαθεί. Εξήγησε στον πατέρα της πόσο σημαντικά είναι αυτά τα στοιχεία και εκείνος της υποσχέθηκε ότι όλες οι μελλοντικές πωλήσεις πλέον θα αποθηκεύονται.

Η Julie διαπίστωσε ότι οι πωλήσεις των Murphy πιθανόν να σχετίζονται με τις εθνικές πωλήσεις και γι' αυτό αποφάσισε να αναζητήσει την κατάλληλη μεταβλητή σε μία από τις ομοσπονδιακές δημοσιεύσεις. Κατάφερε και βρήκε το ιστορικό των μηνιαίων πωλήσεων όλων των λιανικών καταστημάτων των Ηνωμένων Πολιτειών και αποφάσισε να χρησιμοποιήσει αυτή τη μεταβλητή ως υποκατάστατο της μεταβλητής που την ενδιέφερε (δηλαδή τις πωλήσεις των Murphy σε δολάρια). Ο στόχος της λοιπόν ήταν να βρει ακριβείς προβλέψεις για τις εθνικές πωλήσεις και συσχετίζοντας τες, να βρει τις προβλέψεις που

επιθυμεί για την επιχείρηση τους.

Ο πίνακας 3-8 περιέχει τα δεδομένα που συνέλεξε η Julie, ενώ στο σχήμα 3-22 φαίνεται η γραφική αναπαράσταση τους (time series plot), η οποία παρήχθη από το υπολογιστικό πρόγραμμα. Η ανάλυση ξεκινά με την εύρεση των συντελεστών αυτοσυσχέτισης και τη γραφική τους απεικόνιση. Από τον έλεγχο του σχήματος 3-23 ήταν ξεκάθαρη η άποψη ότι τα δεδομένα περιέχουν τάση. Οι πρώτοι συντελεστές αυτοσυσχέτισης είναι πολύ μεγάλοι και μικραίνουν πολύ αργά με το χρόνο ώσπου να μηδενιστούν. Συνεπώς, η χρονοσειρά θα πρέπει να γίνει στάσιμη απαλείφοντας την τάση. Η πρώτη σκέψη είναι η εύρεση των διαφορών της αρχικής με τη χρονοκαθυστερημένη κατά ένα (differences με lag1), ώστε να δούμε εάν προκύπτει μία μη στάσιμη χρονοσειρά. Οι συντελεστές αυτοσυσχέτισης της νέας χρονοσειράς φαίνονται στο σχήμα 3-24.

ΕΡΩΤΗΣΕΙΣ

1. Τι θα πρέπει να συμπεράνει η Julie για τη χρονοσειρά των λιανικών πωλήσεων;

Αυτό που αρχικά συμπεραίνει η Julie είναι ότι οι λιανικές πωλήσεις αποτελούν μια συνεχή χρονολογική σειρά, μη στάσιμη. Σαν στοιχεία ύπαρξης μη στασιμότητας είναι κυρίως η ύπαρξη τάσης, η ύπαρξη εποχικότητας, η αλλαγή της μεταβλητότητας συναρτήσει του χρόνου. Μέσω της γραφικής απεικόνισης της χρονοσειράς διαπιστώνει πως εμφανίζει μια ανοδική τάση καθώς τα δεδομένα δεν κυμαίνονται γύρω από ένα σταθερό μέσο, αλλά αυξάνονται συνεχώς. Παράλληλα, χρησιμοποιεί και το διάγραμμα αυτοσυσχετίσεων για τον έλεγχο στασιμότητας ως προς τη μέση τιμή. Παρατηρεί ότι οι συντελεστές αυτοσυσχέτισης φθίνουν αργά στο χρόνο, γεγονός που ενισχύει την

αρχική της διαπίστωση για μη στάσιμη χρονοσειρά. Στη συνέχεια συμπεραίνει την ύπαρξη εποχικότητας. Όπως παρατηρεί από το διάγραμμα της χρονοσειράς και κυρίως από την απεικόνιση των καταλοίπων της, υπάρχει εποχικότητα των δεδομένων καθώς ο συντελεστής αυτοσυσχέτισης στα lag 12, 24, 36 εμφανίζει ένα «peak» (και στο lag 12 υπερβαίνει το διάστημα εμπιστοσύνης). Διαπιστώνει λοιπόν ότι κατά τον μήνα Δεκέμβρη, οι λιανικές πωλήσεις κάθε έτους μεγιστοποιούνται, γεγονός που ίσως ερμηνεύεται από την παραδοχή ότι τον μήνα αυτό λόγω των γιορτών αυξάνεται η ζήτηση στην αγορά επίπλων. Στο σημείο αυτό αναφέρουμε ότι η Julie θα έπρεπε αρχικά να υπολογίσει κάποιους βασικούς στατιστικούς δείκτες, όπως η μέση τιμή, η τυπική απόκλιση, ελάχιστο, μέγιστο, προκειμένου να αναδείξει κάποια δευτερεύοντα χαρακτηριστικά της χρονοσειράς.

2. Έχει επιτελέσει η Julie κάποια πρόοδο σχετικά με την εξεύρεση κάποιας μεθόδου πρόβλεψης;

Η δεδομένη χρονοσειρά μας είναι μη στάσιμη, όπως διαπιστώσαμε στο παραπάνω ερώτημα. Για την απομάκρυνση της τάσης από τα μη στάσιμα δεδομένα, η Julie χρησιμοποίησε τη μέθοδο των διαφορών. Η σημαντική πρόοδο που πέτυχε είναι η αφαίρεση της τάσης, γεγονός που παρατηρείται και στο διάγραμμα αυτοσυσχέτισης των καταλοίπων των διαφορών. Πλέον, οι συντελεστές αυτοσυσχέτισης δεν φθίνουν αργά, αλλά στα lag 12, 24, 36 υπερβαίνουν το διάστημα εμπιστοσύνης, όπως και προηγουμένως (ύπαρξη εποχικότητας).

3. Ποιες τεχνικές πρόβλεψης θα μπορούσε να δοκιμάσει να εφαρμόσει η Julie;

Στην περίπτωση που τα δεδομένα είναι απαλλαγμένα από τάση, μπορούμε να χρησιμοποιούμε μεθόδους πρόβλεψης όπως είναι η naive

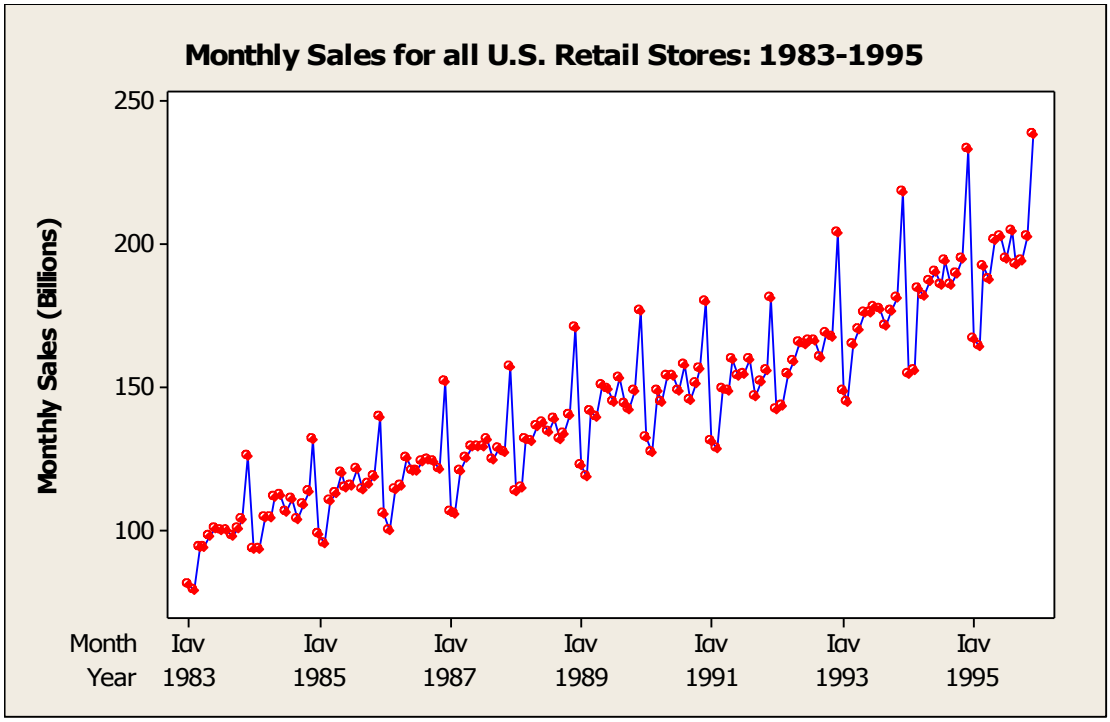
μέθοδος, ο απλός ή κινητός μέσος όρος και τα μοντέλα autoregressive moving average (ARMA) με εποχικότητα.

4. Πως θα μπορούσε η Julie να γνωρίζει ποια τεχνική δουλεύει καλύτερα;

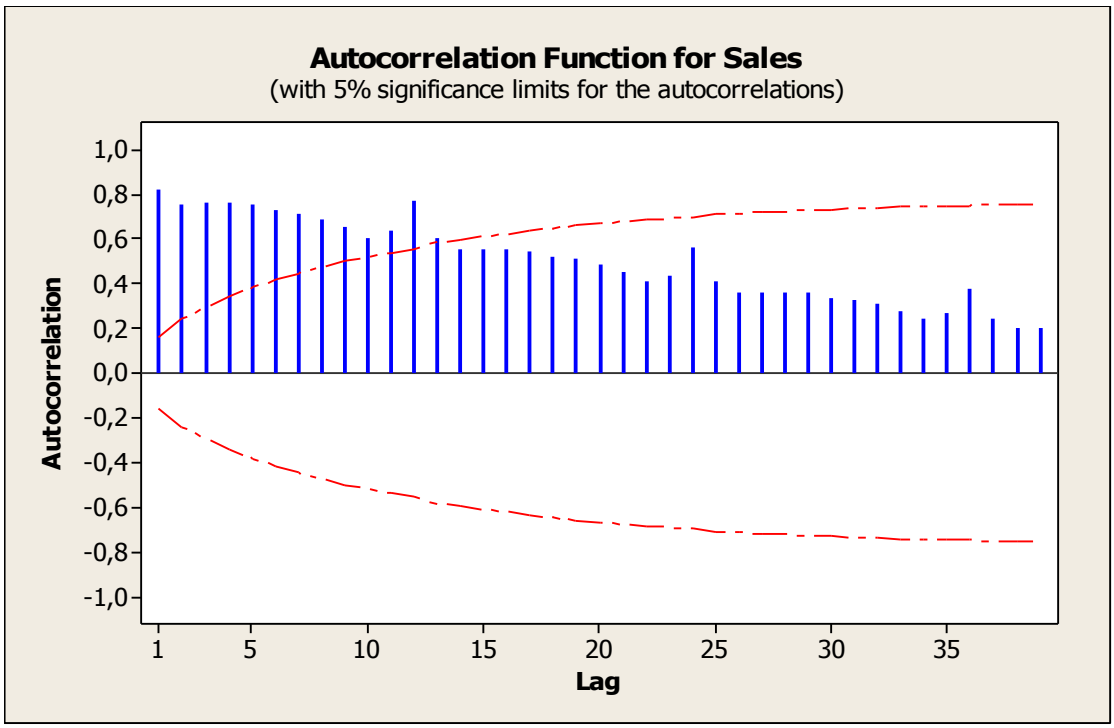
Στην ανάλυση χρονοσειρών μας ενδιαφέρει να ελέγξουμε αν υπάρχει αλληλεξάρτηση μεταξύ διαδοχικών παρατηρήσεων, οπότε παραβιάζεται μία από τις υποθέσεις των τυχαίων δεδομένων που είναι η ανεξαρτησία των παρατηρήσεων. Για τον σκοπό αυτό υπολογίζονται οι συντελεστές αυτοσυσχέτισης. Η Julie προκειμένου να διαπιστώσει ποια τεχνική δουλεύει καλύτερα, θα μπορούσε να εφαρμόσει την Q στατιστική στα αντίστοιχα κατάλοιπα (residuals) της κάθε μεθόδου.

Η στατιστική Q εξετάζει τους συντελεστές αυτοσυσχέτισης σαν ομάδα και όχι ανεξάρτητα τον κάθε ένα. Η στατιστική αυτή ακολουθεί την κατανομή χι τετράγωνο με m βαθμούς ελευθερίας, τον αριθμό δηλαδή των χρονοκαθυστερήσεων. Εάν οι τιμές των συντελεστών αυτοσυσχέτισης για την εκάστοτε μέθοδο είναι στα όρια, που δίνει η Q στατιστική, θεωρούμε ότι τα σφάλματα δεν αυτοσυσχετίζονται, είναι δηλαδή ανεξάρτητα μεταξύ τους, επομένως η αντίστοιχη μέθοδος είναι αξιόπιστη για την πραγματοποίηση προβλέψεων.

Επιπρόσθετα, για να αξιολογήσει, ποια από τις χρησιμοποιούμενες τεχνικές είναι καλύτερη, θα χρησιμοποιήσει τους στατιστικούς δείκτες αξιοπιστίας που βασίζονται στην ανάλυση των καταλοίπων. Τέτοιοι είναι οι δείκτες RMSE (Ρίζα του Μέσου του Τετραγώνου του Σφάλματος), MAD (Μέση Απόλυτη Απόκλιση), MSE(Μέσο Τετραγωνικό Σφάλμα), MAPE (Μέσο Απόλυτο Ποσοστιαίο Σφάλμα). Το μοντέλο που εμφανίζει τις μικρότερες τιμές των δεικτών αυτών (κοντά στο 0), κάνει καλύτερη προσαρμογή στα δεδομένα της χρονοσειράς μας.

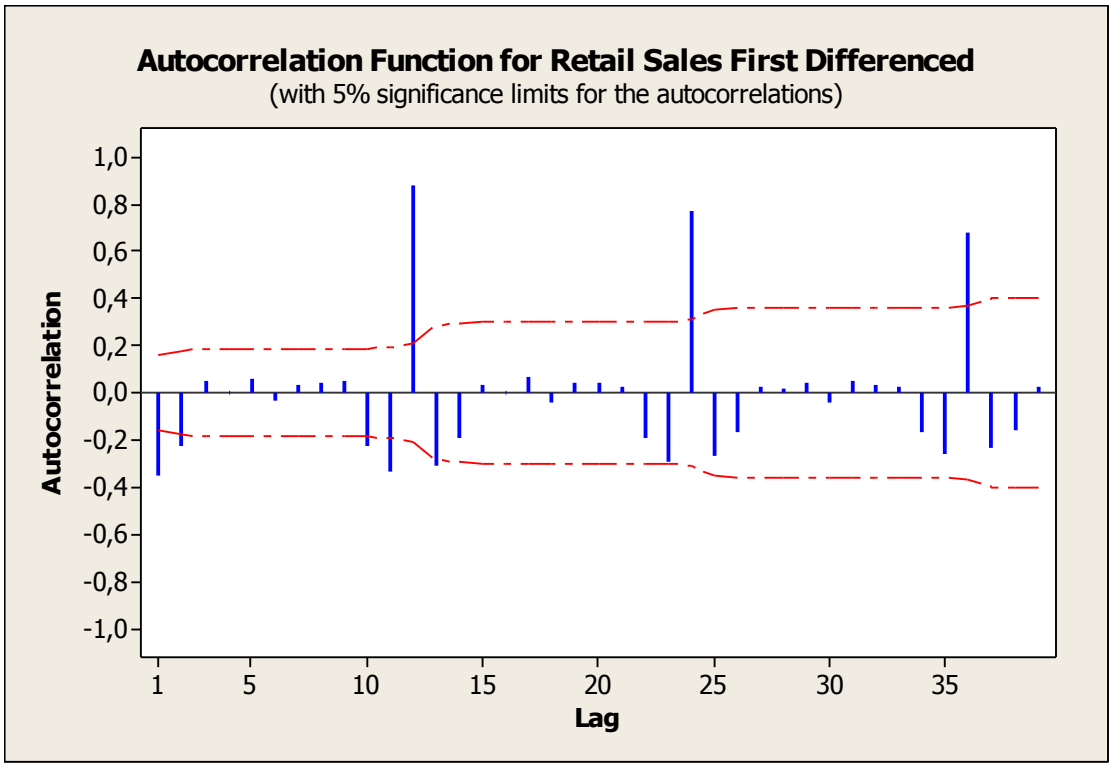


Σχήμα 3-22 Γράφημα χρονοσειράς των μηνιαίων λιανικών πωλήσεων των καταστημάτων επίπλων των Ηνωμένων Πολιτειών.



Σχήμα 3-23 Γραφική αναπαράσταση της συνάρτησης αυτοσυσχέτησης των μηνιαίων λιανικών πωλήσεων όλων των καταστημάτων επίπλων των

Ηνωμένων Πολιτειών.



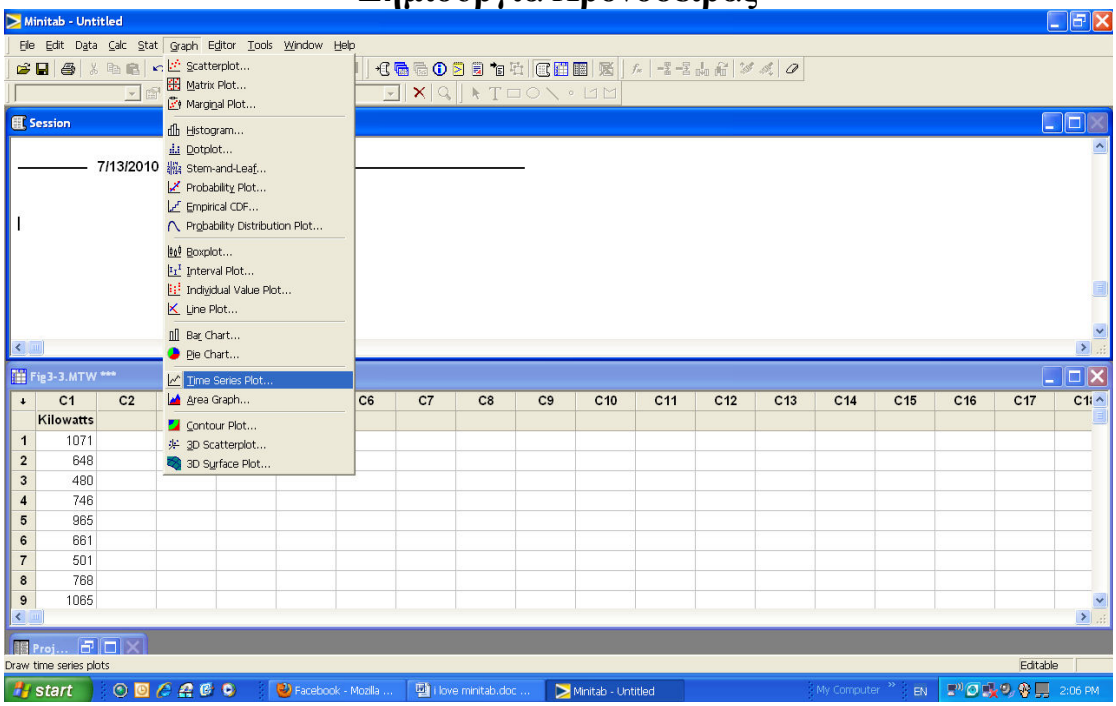
Σχήμα 3-24 Γραφική αναπαράσταση της συνάρτησης αυτοσυσχέτησης στις διαφορές της αρχικής από την χρονοκαθυστερημένη κατά ένα (lag1)

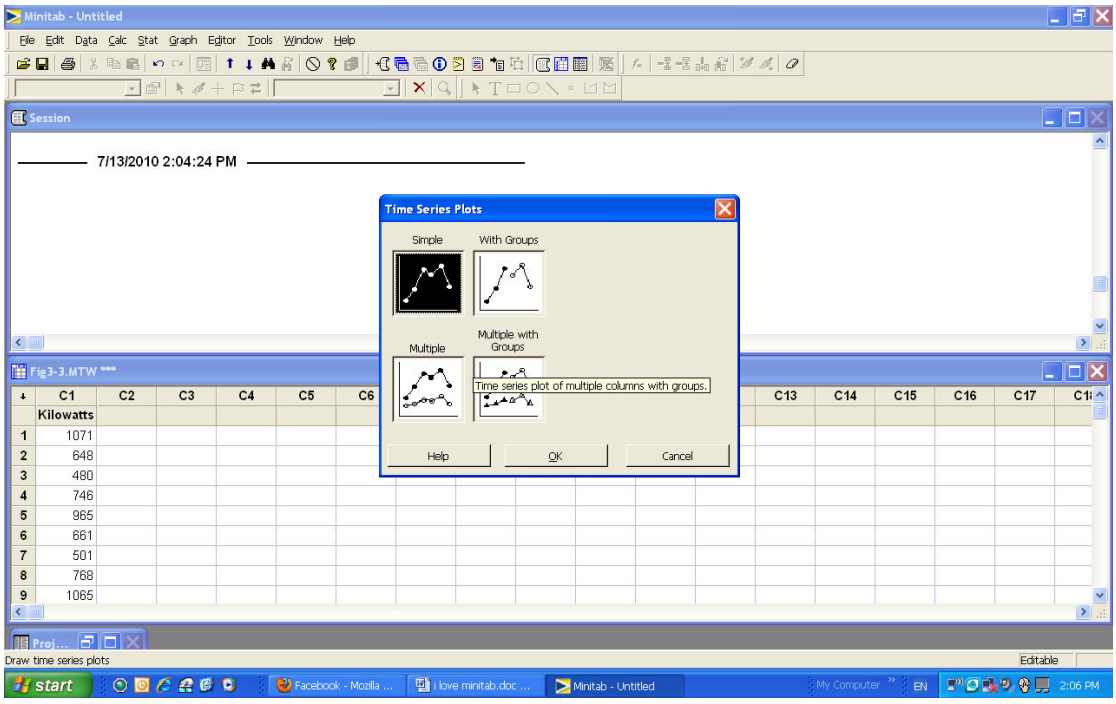
Table 3-8 Μηνιαίες πωλήσεις (σε δισεκατομύρια $) για όλα τα καταστήματα λιανικής 1983-1995												
1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
81,3	93,1	98,8	105,6	106,4	113,6	122,5	132,6	130,9	142,1	148,4	154,6	167
78,9	93,7	95,6	99,7	105,8	115	118,9	127,3	128,6	143,1	145	155,8	164
93,8	104,3	110,2	114,2	120,4	131,6	141,3	148,3	149,3	154,7	164,6	184,2	192,1
94	104,3	113,1	115,7	125,4	130,9	139,8	145	148,5	159,1	170,3	181,8	187,5
97,8	111,3	120,3	125,4	129,1	136	150,3	154,1	159,8	165,8	176,1	187,2	201,4
100,6	112	115	120,4	129	137,5	149	153,5	153,9	164,6	175,7	190,1	202,6
99,6	106,6	115,5	120,7	129,3	134,1	144,6	148,9	154,6	166	177,7	185,8	194,9
100,2	110,7	121,1	124,1	131,5	138,7	153	157,4	159,9	166,3	177,1	193,8	204,2
98	103,9	114,2	124,4	124,5	131,9	144,1	145,6	146,7	160,6	171,1	185,9	192,8
100,7	109,2	116,1	123,8	128,3	133,8	142,3	151,5	152,1	168,7	176,4	189,7	194
103,9	113,3	118,6	121,4	126,9	140,2	148,8	156,1	155,6	167,2	180,9	194,7	202,4

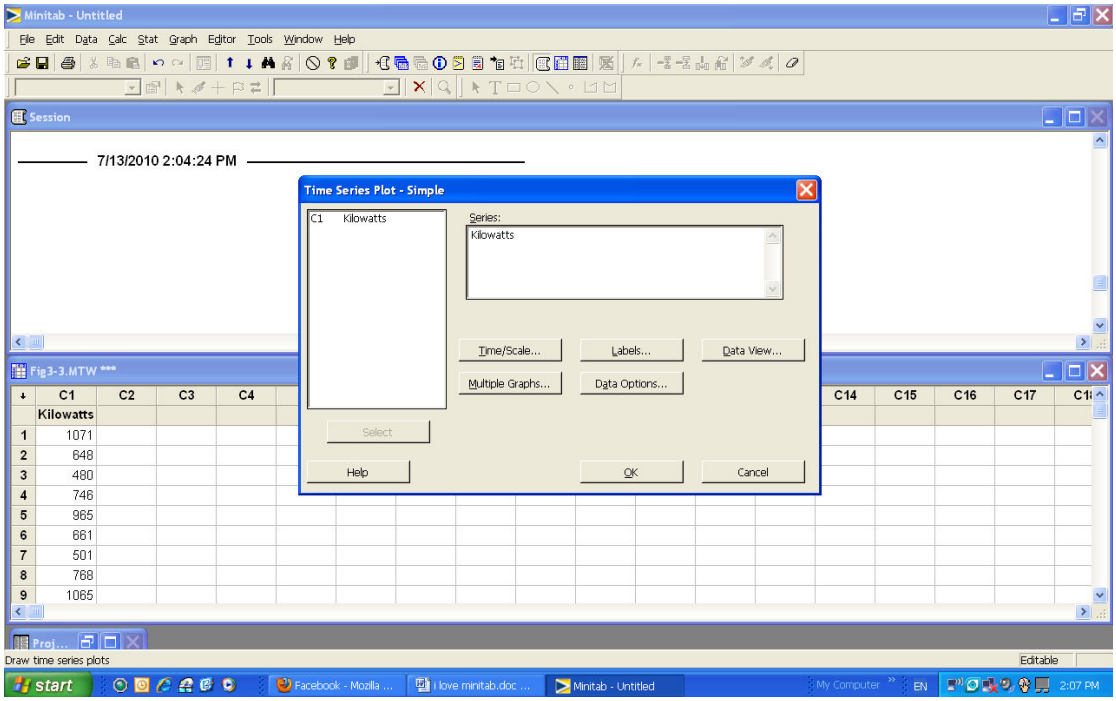
125,7	131,8	139,5	152,1	157,2	171	176,5	179,7	181	204,1	218,3	233,3	238

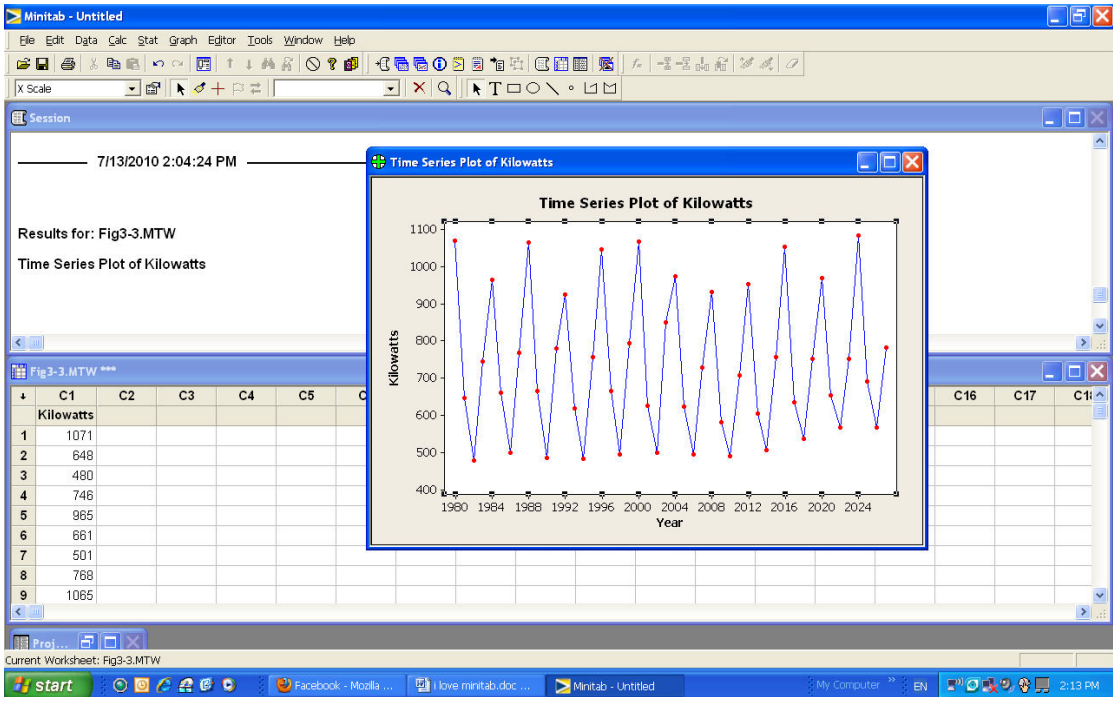
3.15 ΧΡΗΣΗ MINITAB

Δημιουργία Χρονοσειράς

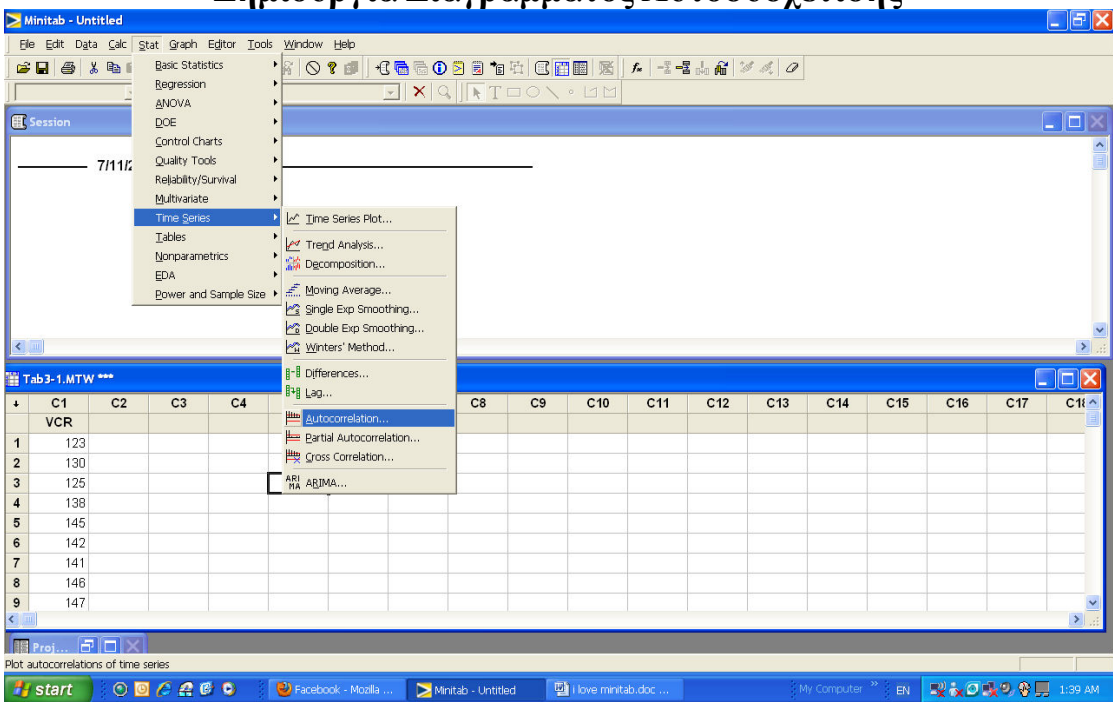


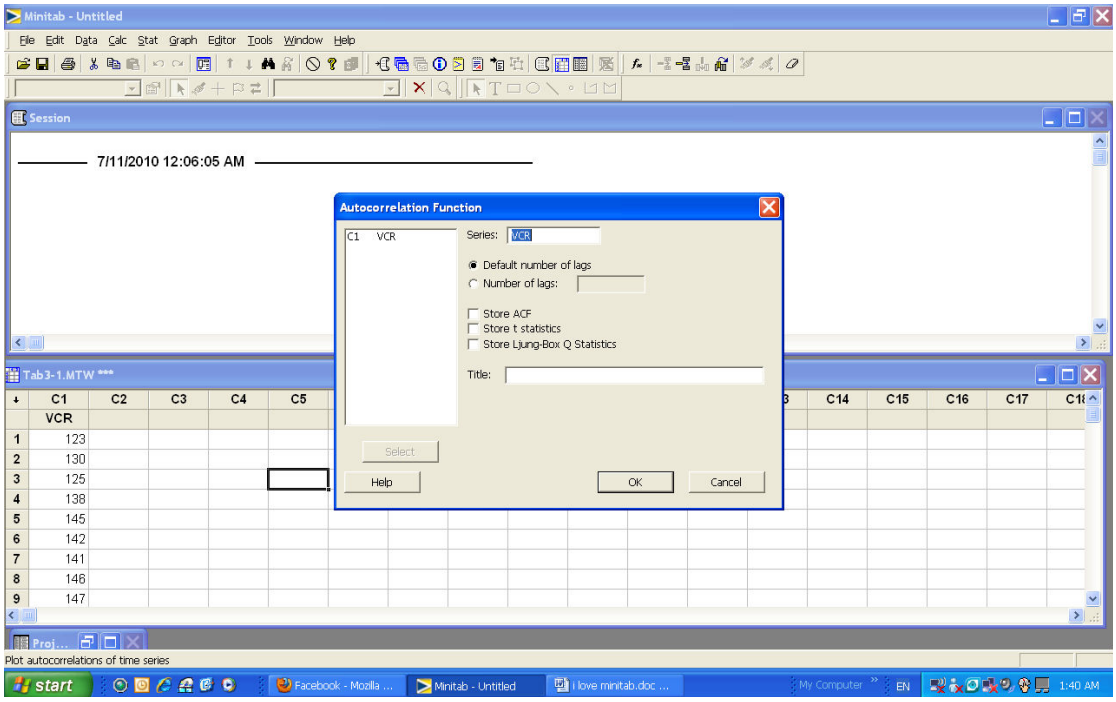


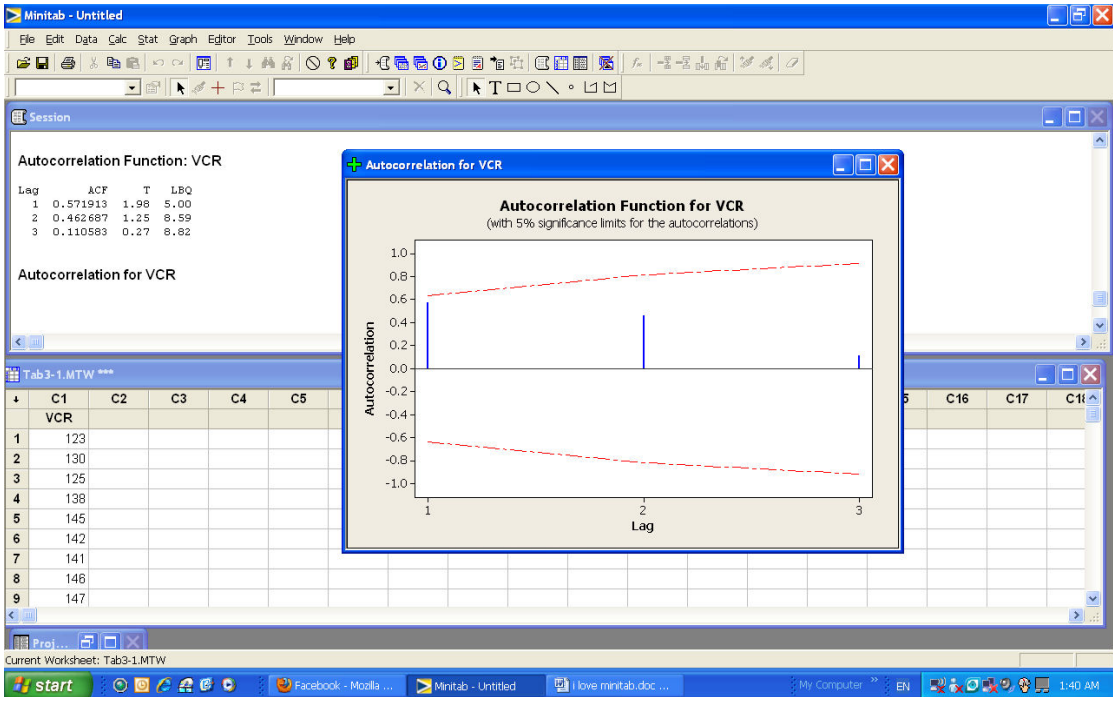

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/13/2010 2:04:24 PM
Time Series Plot - Simple
C1 Kilowatts
Series:
Kilowatts
Time/Scale...
Labels...
Data View...
Multiple Graphs...
Data Options...
Select
Help
OK
Cancel
Fig3-3.MTW ***
C1 C2 C3 C4
Kilowatts
1 1071
2 648
3 480
4 746
5 965
6 681
7 501
8 768
9 1065
Draw time series plots
Editable
start
Facebook - Mozilla ...
I love minitab.doc ...
Minitab - Untitled
My Computer
EN
2:07 PM


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
X Scale
Session
7/13/2010 2:04:24 PM
Results for: Fig3-3.MTW
Time Series Plot of Kilowatts
Time Series Plot of Kilowatts
Time Series Plot of Kilowatts
Kilowatts
1100
1000
900
800
700
600
500
400
1980 1984 1988 1992 1996 2000 2004 2008 2012 2016 2020 2024
Year
Fig3-3.MTW ***
C1 C2 C3 C4 C5
Kilowatts
1 1071
2 648
3 480
4 746
5 965
6 681
7 501
8 768
9 1065
Current Worksheet: Fig3-3.MTW
start
Facebook - Mozilla ...
I love minitab.doc ...
Minitab - Untitled
My Computer
EN
2:13 PM

Δημιουργία Διαγράμματος Αυτοσυσχέτισης






Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Autocorrelation Function: VCR
Lag ACF T LBQ
1 0.571913 1.98 5.00
2 0.462687 1.25 8.59
3 0.110583 0.27 8.82
Autocorrelation for VCR
Autocorrelation for VCR
Autocorrelation Function for VCR
(with 5% significance limits for the autocorrelations)
Autocorrelation
Lag
Tab3-1.MTW ***
C1 C2 C3 C4 C5 C16 C17
VCR
123
130
125
138
145
142
141
146
147
Current Worksheet: Tab3-1.MTW
start
Facebook - Mozilla ...
Minitab - Untitled
i love minitab.doc ...
My Computer
EN
1:40 AM

ΚΕΦΑΛΑΙΟ 4 (ΚΙΝΗΤΟΙ ΜΕΣΟΙ ΚΑΙ ΜΕΘΟΔΟΙ ΕΞΟΜΑΛΥΝΣΗΣ)

4.1 ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ

Για τη διενέργεια προβλέψεων υπάρχει ένας μεγάλος αριθμός μεθόδων, που μπορούν να χρησιμοποιηθούν ανάλογα με την εξεταζόμενη περίπτωση. Οι περισσότερες από αυτές αναπτύχθηκαν πριν από μερικές δεκαετίες και εφαρμόστηκαν σε ευρεία κλίμακα, λόγω της ραγδαίας ανάπτυξης κατάλληλων υπολογιστικών προγραμμάτων. Στόχος μας είναι η χρησιμοποίηση εκείνων των μεθόδων, που είναι εύχρηστες, εφαρμόζονται συχνά και έχουν αποδειχθεί στην πράξη, ότι δίνουν καλά αποτελέσματα. Μάλιστα, αν ληφθεί υπόψη η ευκολία, με την οποία προσδιορίζονται οι προβλέψεις με τη χρήση των Η/Υ, τότε το κύριο βάρος της έρευνας θα πρέπει να εστιάζεται αφενός στην αξιολόγηση των αποτελεσμάτων και αφετέρου στην επιλογή της καταλληλότερης μεθόδου πρόβλεψης.

Σ' αυτό το κεφάλαιο θα περιγράψουμε τρεις απλές προσεγγίσεις για πρόβλεψη μιας χρονοσειράς.

Η πιο απλή μέθοδος είναι η naïve και οι παραλλαγές της, όπου πάντα χρησιμοποιούμε, ή διαμορφώνουμε τη τελευταία τιμή, έτσι ώστε να προβλέψουμε την επόμενη.

Οι μέθοδοι μέσου όρου χρησιμοποιούν το μέσο όρο όλων, ή μερικών από τις προηγούμενες τιμές, προκειμένου να προβλεφθούν οι επόμενες τιμές. Οι μέθοδοι μέσου όρου είναι τρεις:

- Του απλού μέσου όρου
- Του κινητού μέσου όρου

- Του διπλού κινητού μέσου όρου

Η *εκθετική εξομάλυνση* (exponential smoothing) είναι μία μέθοδος πρόβλεψης, η οποία προεκτείνει στοιχεία του προτύπου των ιστορικών δεδομένων στο μέλλον, όπως τάσεις και εποχικούς κύκλους. Οι προβλέψεις υπολογίζονται μετά από εξομάλυνση των δεδομένων, προκειμένου να απομονωθούν τα πραγματικά πρότυπα από τις καθαρά τυχαίες διακυμάνσεις. Μεγαλύτερη βαρύτητα δίδεται στα πρόσφατα δεδομένα, ενώ φθίνει εκθετικά η βαρύτητα για τα δεδομένα, που αντιστοιχούν σε παλαιότερες χρονικές περιόδους.

Οι κυριότερες μέθοδοι εκθετικής εξομάλυνσης είναι οι εξής:

- Η απλή εκθετική εξομάλυνση
- Η διπλή εκθετική εξομάλυνση, ή μέθοδος Holt
- Η τριπλή εκθετική εξομάλυνση, ή μέθοδος Winters

Η στρατηγική που ακολουθούμε είναι η ακόλουθη:

1. Επιλογή μιας μεθόδου πρόβλεψης, η οποία βασίζεται στη διαίσθηση του αναλυτή και στα υπάρχοντα δεδομένα.
2. Τα δεδομένα επιλέγονται σε δύο τμήματα: Το αρχικό ή fitting τμήμα και το δεύτερο τμήμα για το τεστ της πρόβλεψης.
3. Η επιλεγμένη μέθοδος εφαρμόζεται για την προσαρμογή των τιμών (fitting) για το πρώτο τμήμα των δεδομένων.
4. Η τεχνική χρησιμοποιείται για την δοκιμαστική πρόβλεψη και εκτιμώνται τα σφάλματα σύμφωνα με τα αναφερόμενα στο κεφάλαιο 3.
5. Λήψη της απόφασης. Ή χρησιμοποιούμε την τεχνική όπως είναι, ή την τροποποιούμε, ή την αλλάζουμε.

4.2 ΜΕΘΟΔΟΣ NAIVE ΚΑΙ ΟΙ ΠΑΡΑΛΛΑΓΕΣ ΤΗΣ

Η «*απλοϊκή*» (Naïve) μέθοδος είναι η πιο απλή μέθοδος και αποτελεί σημείο αναφοράς για όλες τις άλλες μεθόδους πρόβλεψης. Κάθε μέθοδος, για να θεωρηθεί αποτελεσματική, πρέπει να δίνει αποτελέσματα πιο ακριβή από τα αποτελέσματα της Naïve. Πολλές φορές χρησιμοποιείται ως benchmark για άλλες μεθόδους. Η μέθοδος αυτή δίνει ως πρόβλεψη την τελευταία γνωστή παρατήρηση γιαυτό και λέγεται και "no change forecast". Έχει καλές επιδόσεις για πρόβλεψη μιας περιόδου μπροστά, καθώς η αναμενόμενη τιμή της πρόβλεψης δεν διαφέρει σημαντικά από την τελευταία παρατήρηση, που έχουμε στην διάθεση μας. Το θέμα είναι ότι μεταφέρουμε τις τυχαίες διακυμάνσεις.

Η μέθοδος Naïve περιγράφεται από την ακόλουθη εξίσωση:

$$\hat{Y}_{t+1} = Y_t$$

όπου,

- $\hat{Y}_{t+1}$: η πρόβλεψη για την περίοδο t + 1.
- Y_t: η παρατήρηση την χρονική περίοδο t.

Στην ουσία η χρονοσειρά της πρόβλεψης αντικαθίσταται από το lag 1 αυτής.

Παράδειγμα (Τ4-1)

Το παρακάτω διάγραμμα δείχνει τις τριμηνιαίες πωλήσεις μιας επιχείρησης κατασκευής πριονιών τα έτη 2000 – 2006.

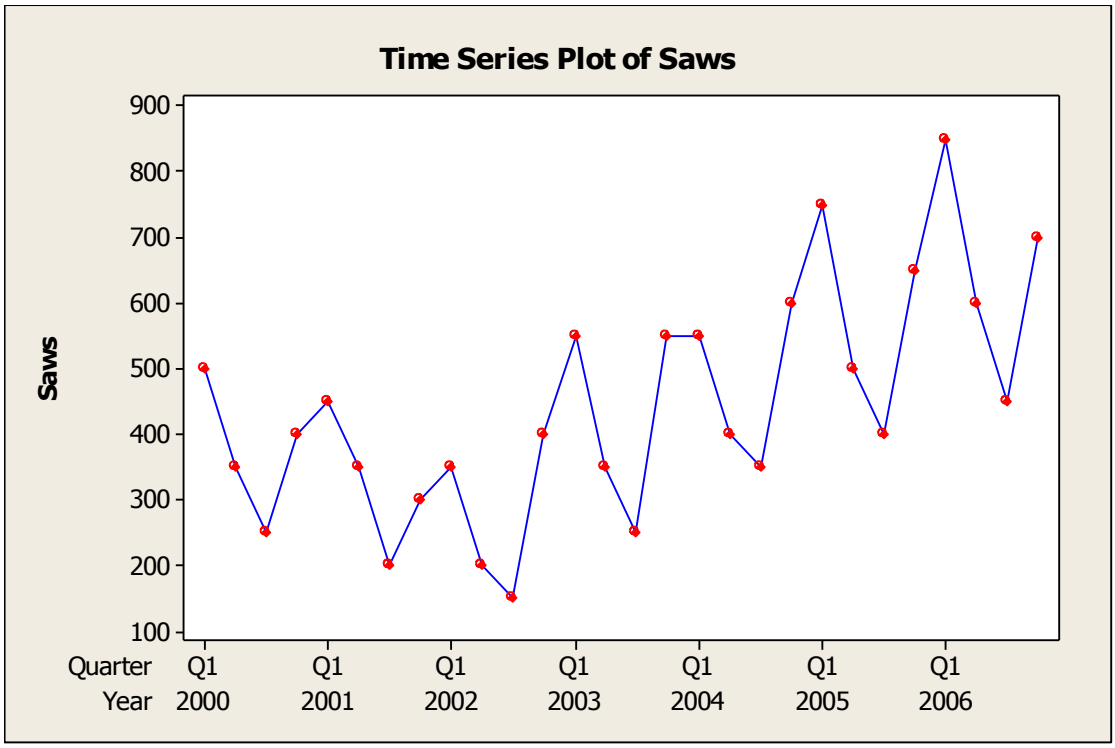
Αν τα δεδομένα, τα έτη 2000 – 2005, χρησιμοποιούνται ως αρχικό μέρος και τα δεδομένα από το 2006 ως μέρος δοκιμής, η πρόβλεψη για το πρώτο τρίμηνο του 2006 είναι:

$$\hat{Y}_{24+1} = Y_{24}$$

$$\hat{Y}_{25} = 650$$

Το σφάλμα για τη περίοδο 25 είναι:

$$e_{25} = Y_{25} - \hat{Y}_{25} = 850 - 650 = 200$$



Η τεχνική αυτή μπορεί να προσαρμοστεί για να ληφθεί υπόψη η μεταβολή των δύο τελευταίων σημείων της χρονοσειράς, προσθέτοντας τη διαφορά μεταξύ αυτής της περιόδου και της τελευταίας περιόδου. Συνεπώς, η πρόβλεψη για το πρώτο τετράμηνο του 2006 είναι:

$$\hat{Y}_{t+1} = Y_t + (Y_t - Y_{t-1})$$

$$\hat{Y}_{24+1} = Y_{24} + (Y_{24} - Y_{24-1})$$

$$\hat{Y}_{25} = Y_{24} + (Y_{24} - Y_{23})$$

$$= 650 + (650 - 400)$$

$$= 650 + 250 = 900$$

Το σφάλμα πρόβλεψης με αυτό το μοντέλο είναι:

$$e_{25} = Y_{25} - \hat{Y}_{25} = 850 - 900 = -50$$

Μπορούμε επίσης να χρησιμοποιήσουμε ποσοστιαία μεταβολή,

αντί να προσθέτουμε διαφορές. Συγκεκριμένα:

$$\hat{Y}_{t+1} = Y_t \frac{Y_t}{Y_{t-1}}$$

Στην περίπτωση που υπάρχει εποχικότητα στα (τριμηνιαία) δεδομένα, τότε μπορούμε να χρησιμοποιήσουμε μία εξίσωση της μορφής:

$$\hat{Y}_{t+1} = Y_{t-3}$$

Μία εξίσωση αυτής της μορφής δεν λαμβάνει υπόψη της καθόλου τα δεδομένα το τελευταίο έτος. Μπορούμε όμως να συνδυάσουμε εποχικά δεδομένα αλλά και δεδομένα τάσης με την ακόλουθη εξίσωση:

$$\hat{Y}_{t+1} = Y_{t-3} + \frac{(Y_t - Y_{t-1}) + \ldots + (Y_{t-3} - Y_{t-4})}{4} = Y_{t-3} + \frac{Y_t - Y_{t-4}}{4}$$

Οι προβλέψεις για το πρώτο τετράμηνο του 2006 και για τις ανωτέρω εξισώσεις είναι:

$$\hat{Y}_{24+1} = Y_{24} \frac{Y_{24}}{Y_{24-1}} \qquad \hat{Y}_{25} = Y_{24} \frac{Y_{24}}{Y_{23}} = 650 \frac{650}{400} = 1{,}056$$

$$e_{25} = Y_{25} - \hat{Y}_{25} = 850 - 1{,}056 = -206$$

$$\hat{Y}_{24+1} = Y_{24-3}$$

$$\hat{Y}_{25} = Y_{21} = 750$$

$$e_{25} = Y_{25} - \hat{Y}_{25} = 850 - 750 = 100$$

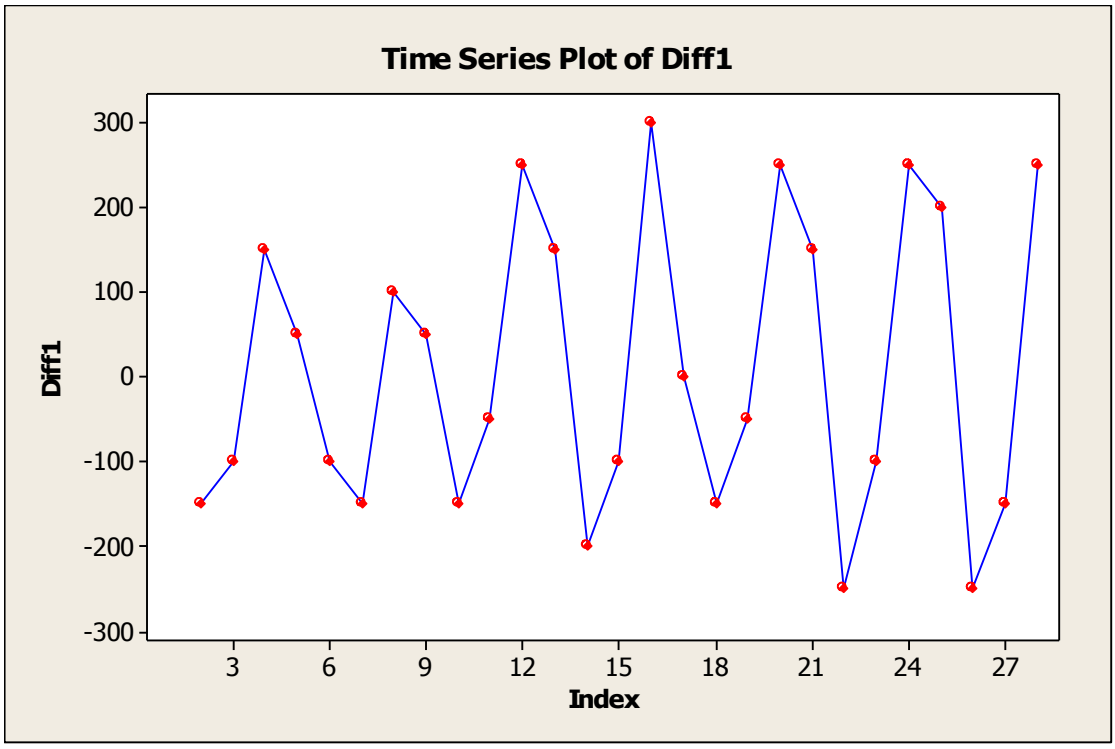
$$\hat{Y}_{24+1} = Y_{24-3} + \frac{Y_{24} - Y_{24-4}}{4}$$

$$\hat{Y}_{25} = Y_{21} + \frac{Y_{24} - Y_{20}}{4} = 750 + \frac{650 - 600}{4} = 750 + 12.5 = 762.5$$

$$e_{25} = Y_{25} - \hat{Y}_{25} = 850 - 762.5 = 87.5$$

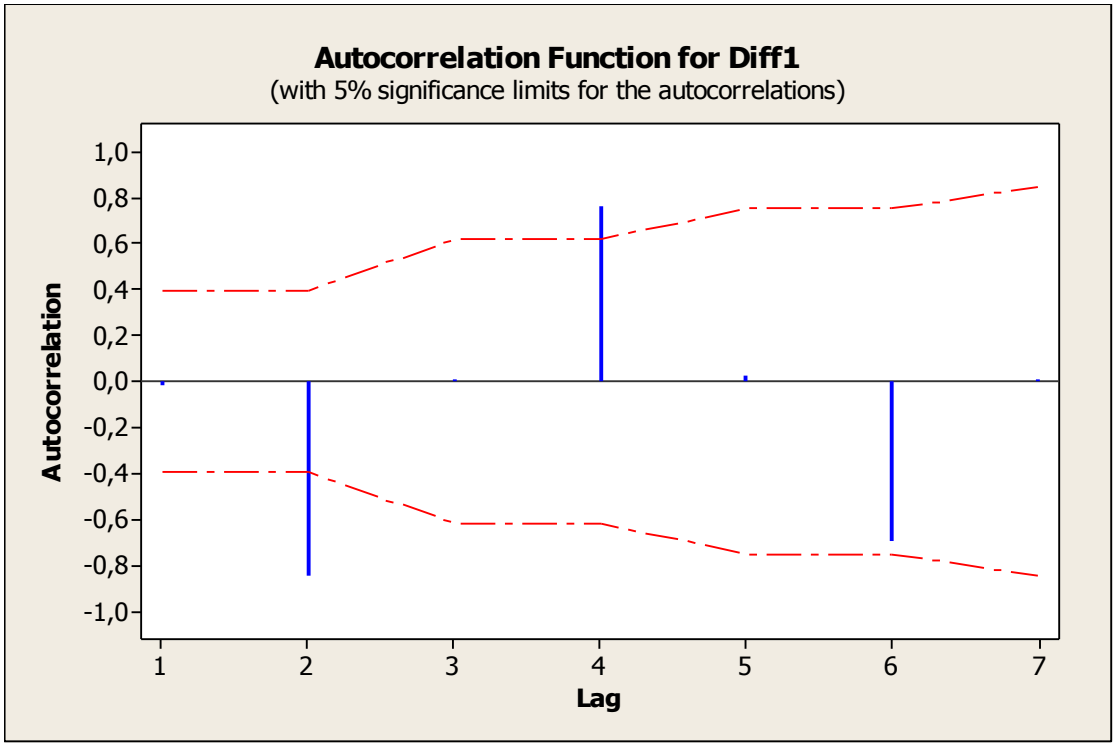
Οι παραλλαγές σε περίπτωση ύπαρξης τάσης μπορούν να δώσουν πιο ακριβή αποτελέσματα, εκτός και αν βρεθούμε σε σημείο μεγίστου ή ελαχίστου της συνάρτησης. Κάτι όμως που δεν είναι πολύ πιθανό να συμβεί, μιας και οι αλλαγές στις χρονοσειρές δεν είναι τόσο συχνές σε σχέση με τον όγκο των δεδομένων.

Στο παρακάτω διάγραμμα δίδονται τα σφάλματα ανάμεσα στο Y_t και Y_{t-1} και οι συντελεστές αυτοσυσχέτισης.

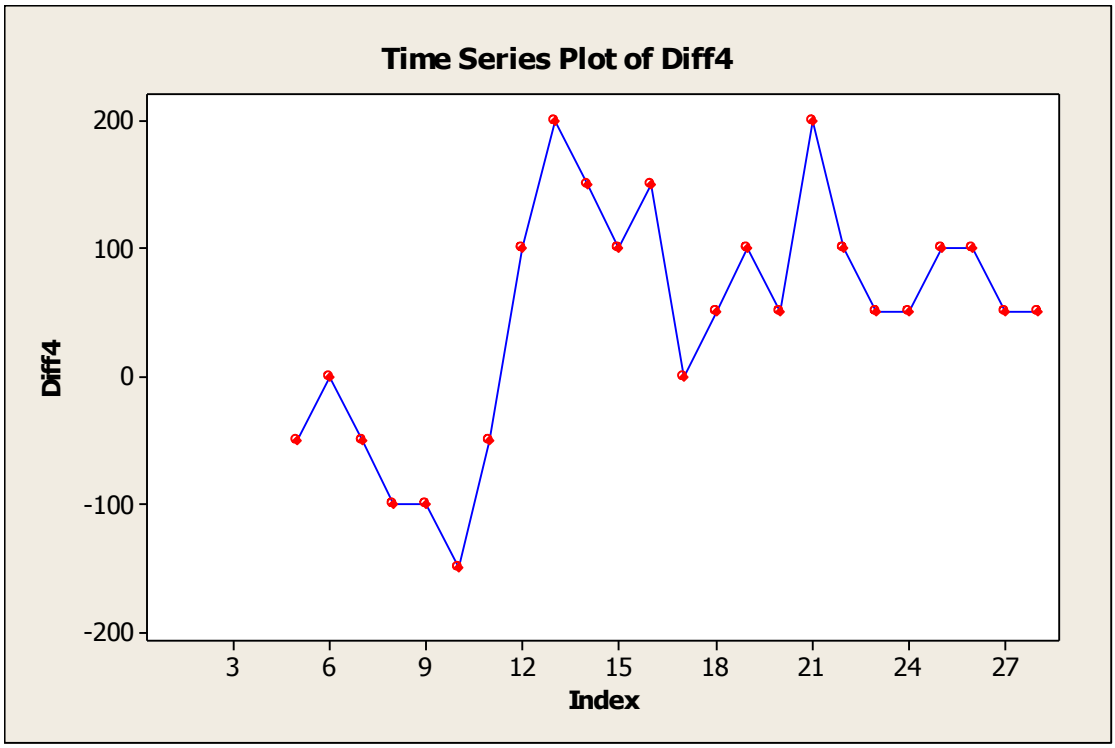


Lag	ACF	T	LBQ
1	-0,016969	-0,09	0,01
2	-0,845803	-4,39	22,41
3	0,005108	0,02	22,42
4	0,760415	2,53	42,10
5	0,024703	0,07	42,12
6	-0,695646	-1,91	60,17
7	0,011209	0,03	60,17

Η τιμή της LBQ είναι μεγαλύτερη από την οριακή τιμή ισχύος της μηδενικής υπόθεσης, που είναι 14,07. Άρα τα σφάλματα αυτοσυσχετίζονται.

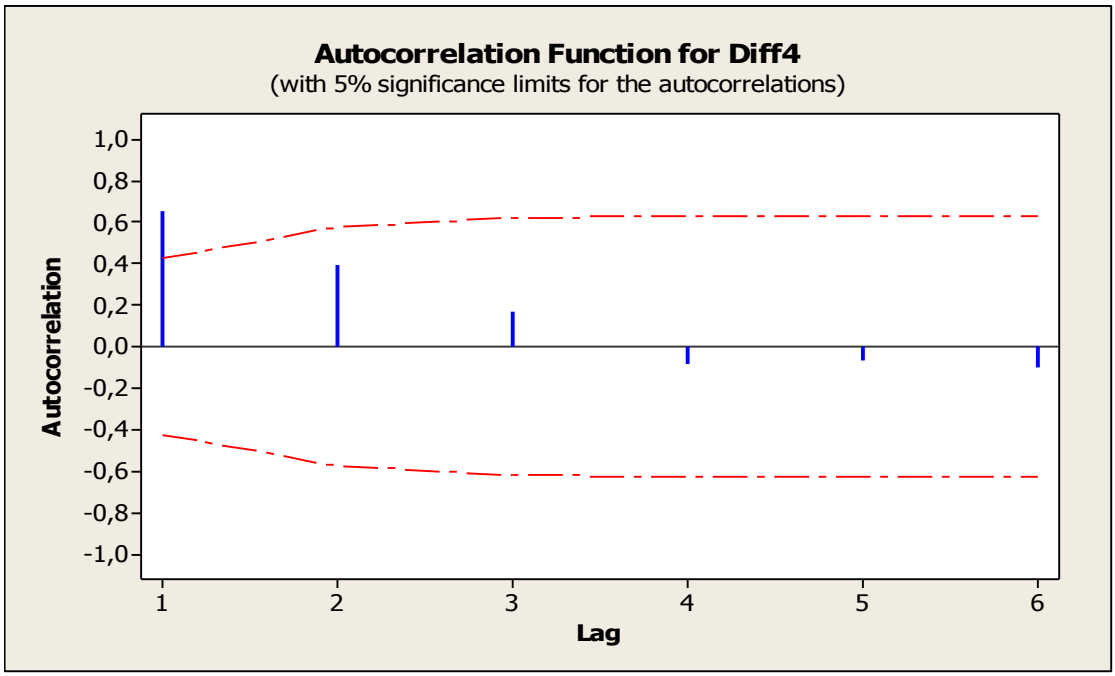


Στο παρακάτω διάγραμμα δίδονται τα σφάλματα ανάμεσα στο Y_t και Y_{t-4} και οι συντελεστές αυτοσυσχέτισης.

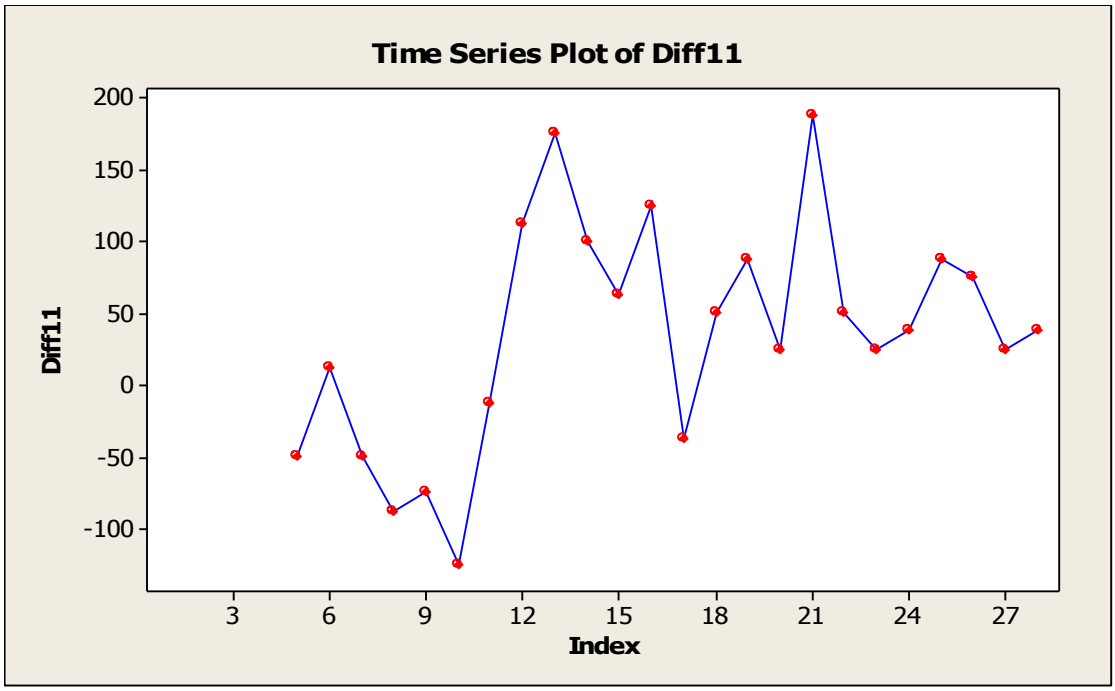


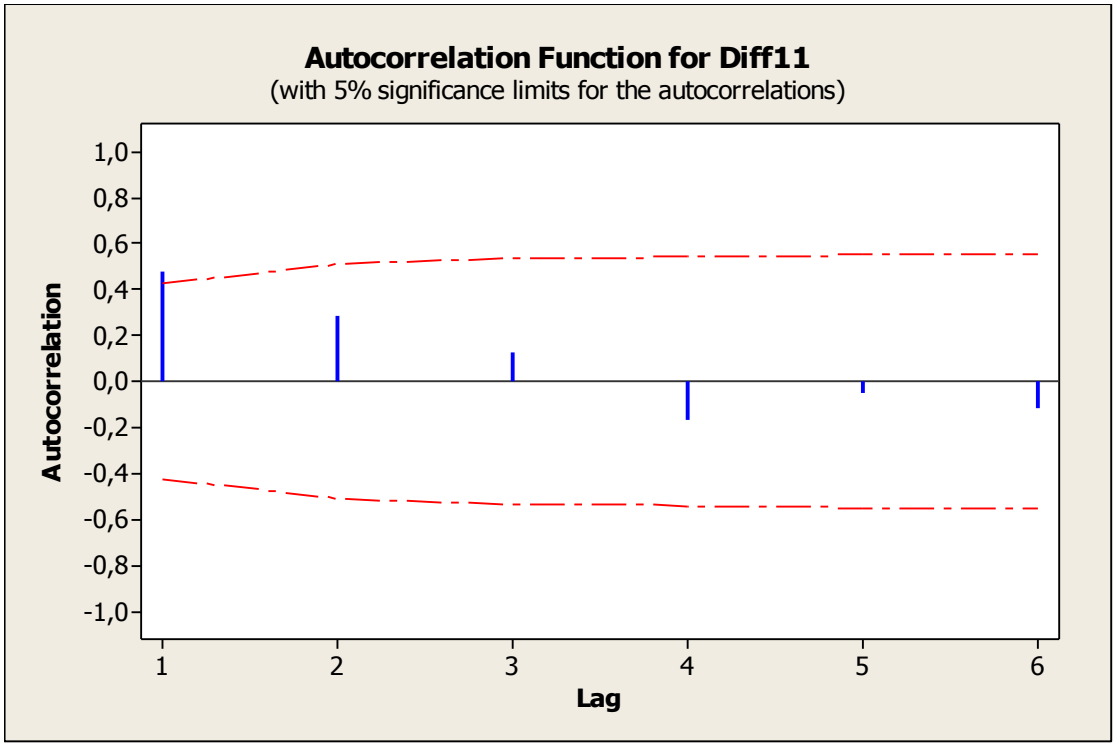
Lag	ACF	T	LBQ
1	0,651270	3,19	11,51
2	0,389179	1,40	15,80

3	0,164666	0,55	16,61
4	-0,083855	-0,28	16,83
5	-0,068285	-0,22	16,98
6	-0,101775	-0,33	17,34



Στο παρακάτω διάγραμμα δίδονται τα σφάλματα ανάμεσα στο Y_t – διορθωμένο με τις μεταβολές του έτους – και Y_{t-4} και οι συντελεστές αυτοσυσχέτισης.





Lag	ACF	T	LBQ
1	0,472406	2,31	6,05
2	0,281214	1,15	8,30
3	0,125808	0,49	8,77
4	-0,164930	-0,63	9,62
5	-0,049805	-0,19	9,70
6	-0,113522	-0,43	10,14

Η τιμή της LBQ είναι μέσα στα όρια ισχύος της μηδενικής υπόθεσης, που είναι 12,59. Άρα τα σφάλματα δεν αυτοσυσχετίζονται, εκτός από την τιμή του πρώτου συντελεστή αυτοσυσχέτισης, που είναι οριακή. Η καλυτέρευση της πρόβλεψης λαμβάνοντας υπόψη τις ετήσιες μεταβολές είναι προφανής.

4.3 ΜΕΘΟΔΟΙ ΜΕΣΟΥ ΟΡΟΥ

4.3.1 ΑΠΛΟΙ ΜΕΣΟΙ

Η πρόβλεψη χρονοσειρών με τη *μέθοδο των μέσων* εμπεριέχει τον

υπολογισμό του μέσου όρου του δείγματος παρατηρήσεων, καθώς και τη χρησιμοποίηση αυτού του μέσου σαν πρόβλεψη για την επόμενη περίοδο. Ο αριθμός των παρατηρήσεων του δείγματος που συμπεριλαμβάνονται στον υπολογισμό του μέσου προσδιορίζεται στην αρχή της διαδικασίας πρόβλεψης. Αλγεβρικά ο μέσος όρος εκφράζεται:

$$\hat{Y}_{t+1} = \frac{1}{t} \sum_{i=1}^{t} Y_i$$

Ο μέσος όρος προκύπτει από την μέθοδο των ελαχίστων τετραγώνων εφαρμοσμένη στο άθροισμα τετραγωνικών σφαλμάτων για κάποια σταθερά $SSE = \sum (Y - \hat{Y})^2$. Ελαχιστοποιώντας αυτή τη σχέση παίρνουμε ως σημείο ελαχιστοποίησης τον μέσο όρο.

Για τον προσδιορισμό της επόμενης προβλεπόμενης τιμής χρησιμοποιούμε τη σχέση:

$$\hat{Y}_{t+2} = \frac{t\,\hat{Y}_{t+1} + Y_{t+1}}{t+1}$$

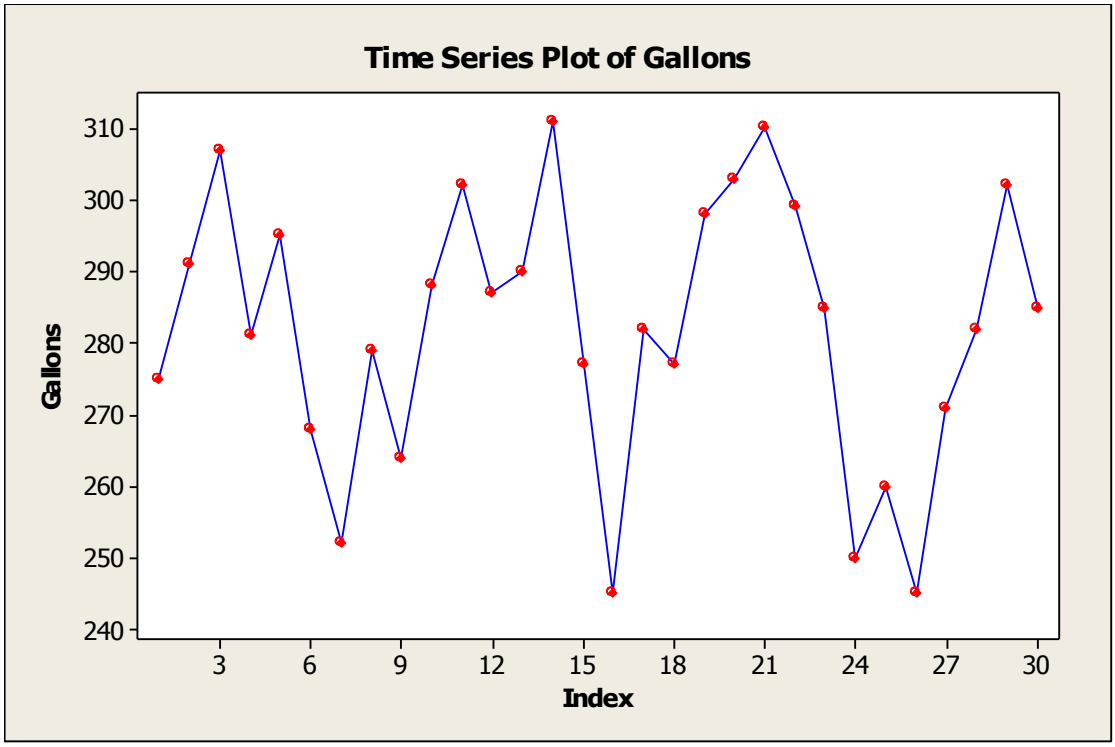
Η μέθοδος πρόβλεψης του απλού μέσου λειτουργεί με καλά αποτελέσματα, όταν οι συνθήκες που παράγουν τις χρονοσειρές έχουν σταθεροποιηθεί και το περιβάλλον παραμένει αμετάβλητο. Ο απλός μέσος χρησιμοποιεί όλα τα παρελθόντα δεδομένα, ώστε να γίνει πρόβλεψη για την επόμενη περίοδο. Μπορούμε να δημιουργήσουμε και ένα διάστημα εμπιστοσύνης με κάποιο επίπεδο σημαντικότητας.

Παράδειγμα Τ(4-2)

Κατανάλωση καυσίμου ανά εβδομάδα για μεταφορά προσώπων. Θα υπολογίσουμε τις δύο τελευταίες τιμές για τις εβδομάδες 29 και 30.

$$\hat{Y}_{28+1} = \frac{1}{28} \sum_{i=1}^{28} Y_i \qquad \hat{Y}_{29} = \frac{7874}{28} = 281.2$$

$$e_{29} = Y_{29} - \hat{Y}_{29} = 302 - 281.2 = 20.8$$



$$\hat{Y}_{28+2} = \hat{Y}_{30} = \frac{28\,\hat{Y}_{28+1} + Y_{28+1}}{28+1} = 281.9$$

$$e_{30} = Y_{30} - \hat{Y}_{30} = 285 - 281.9 = 3.1$$

Για την εβδομάδα 31 η πρόβλεψη θα είναι

$$\hat{Y}_{30+1} = \frac{1}{30}\sum_{i=1}^{30} Y_i = \frac{8{,}461}{30} = 282$$

4.3.2 ΚΙΝΗΤΟΙ ΜΕΣΟΙ

Στον απλό μέσο χρησιμοποιούμε όλα τα δεδομένα. Στην περίπτωση, που θέλουμε να συγκεντρωθούμε περισσότερο στα πρόσφατα γεγονότα, τότε χρησιμοποιούμε τον κινητό μέσο. Η ονομασία κινητός μέσος όρος χρησιμοποιείται για να περιγράψει την διαδικασία, κατά την οποία καθώς μια νέα παρατήρηση γίνεται διαθέσιμη, ένας νέος μέσος όρος μπορεί να υπολογιστεί, στον οποίο παραλείπεται η πιο παλιά παρατήρηση, προκειμένου να συμπεριληφθεί η πιο πρόσφατη. Ο αριθμός των δεδομένων, που χρησιμοποιούνται στον υπολογισμό του μέσου

όρου, παραμένει σταθερός.

Ως κινητός μέσος τάξης k είναι η μέση τιμή των k διαδοχικών παρατηρήσεων και υπολογίζεται από την ακόλουθη σχέση:

$$\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-k+1}}{k}$$

όπου,

$\hat{Y}_{t+1}$ = η πρόβλεψη για την επόμενη περίοδο,

Y_t = η πραγματική τιμή,

k = ο αριθμός των παραγόντων του κινητού μέσου.

Η σχέση αυτή προκύπτει ελαχιστοποιώντας το άθροισμα τετραγωνικών σφαλμάτων:

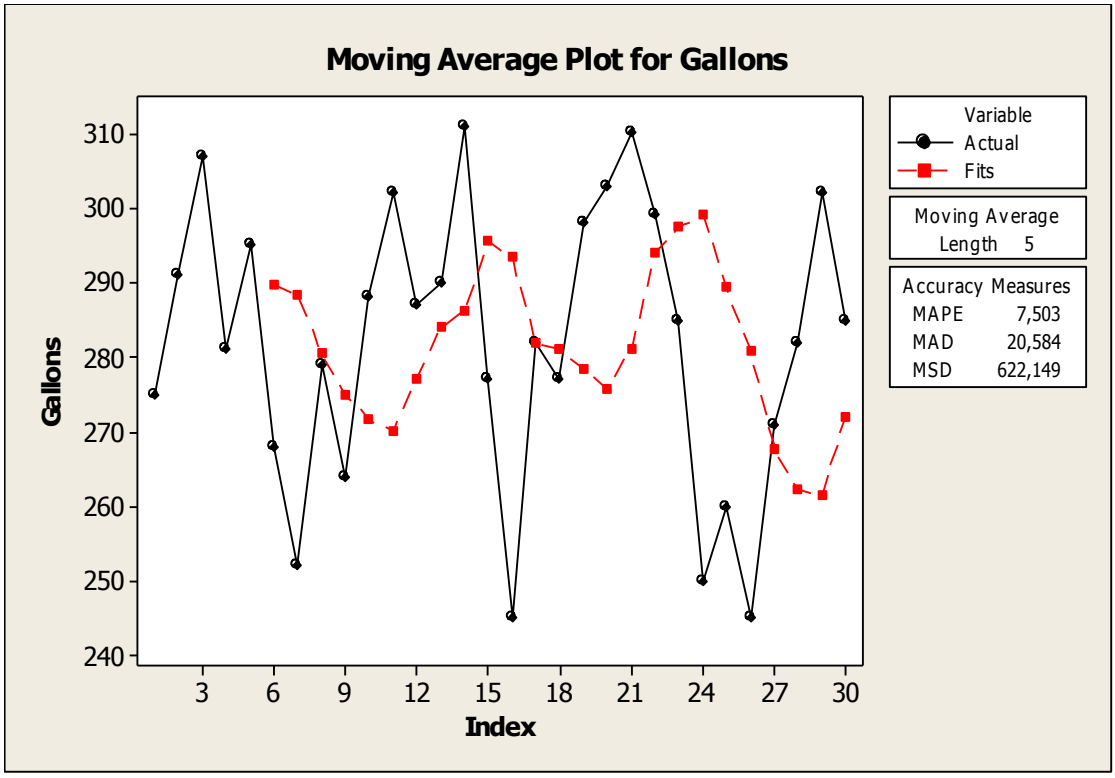
$$SSE = \sum_{i=t-k+1}^{t} \left(Y - \hat{Y}\right)^2$$

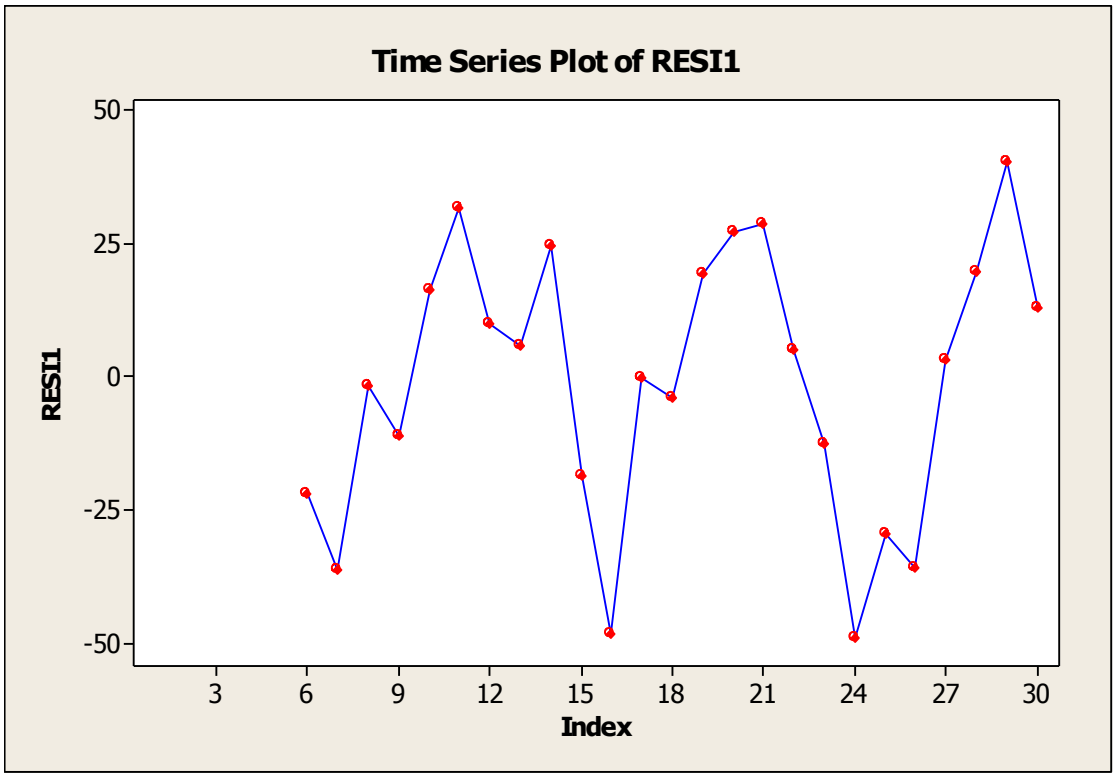
για έναν δεδομένο αριθμό k όρων. Ο αριθμός k μπορεί να ορισθεί, έτσι ώστε το MSE να γίνει ελάχιστο, κάνοντας διάφορες δοκιμές.

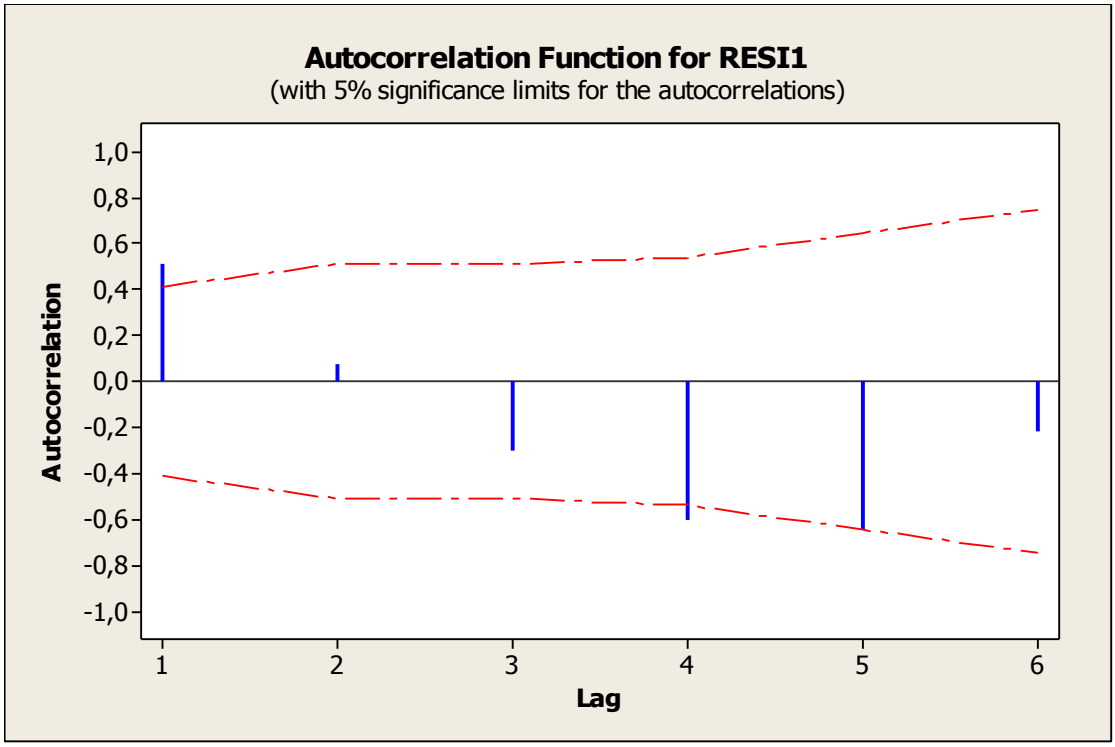
Τη μέθοδο αυτή τη χρησιμοποιούμε, όταν έχουμε τάση, δίνοντας καλύτερα αποτελέσματα από τον απλό μέσο όρο, χωρίς αυτό να σημαίνει ότι είναι και η πλέον κατάλληλη μέθοδος για πρόβλεψη με τάση, γιατί δεν λαμβάνει υπόψη του την επίδραση του κάθε όρου ξεχωριστά αλλά θεωρεί την βαρύτητα του κάθε όρου στη σχέση σταθερή. Μπορούμε να δημιουργήσουμε και ένα διάστημα εμπιστοσύνης με κάποιο επίπεδο σημαντικότητας.

Παράδειγμα (Τ4.2)

Υπολογίζουμε τις προβλεπόμενες τιμές της χρονοσειράς κατανάλωσης καυσίμου στο προηγούμενο παράδειγμα.


Moving Average Plot for Gallons
Gallons
Index
Variable
Actual
Fits
Moving Average
Length 5
Accuracy Measures
MAPE 7,503
MAD 20,584
MSD 622,149


Time Series Plot of RESI1
RESI1
Index



Lag	ACF	T	LBQ
1	0,506290	2,53	7,21
2	0,078551	0,32	7,39
3	-0,298611	-1,21	10,13
4	-0,602830	-2,31	21,81
5	-0,642632	-2,06	35,74
6	-0,219452	-0,61	37,46

Η μηδενική υπόθεση απορρίπτεται επειδή η τιμή 37.46 είναι μεγαλύτερη του άνω ορίου εμπιστοσύνης 12.59.

4.3.3 ΔΙΠΛΟΙ ΚΙΝΗΤΟΙ ΜΕΣΟΙ

Για να προβλέψουμε νέους όρους σε χρονοσειρές, που διαθέτουν γραμμική τάση, χρησιμοποιούμε το διπλό κινητό μέσο. Επειδή με τη μέθοδο των ελαχίστων τετραγώνων κάθε φορά που εμφανίζονται νέες τιμές στις χρονοσειρές χρειάζεται να επανεκτιμήσουμε τις παραμέτρους του γραμμικού μοντέλου για να προχωρήσουμε σε νέες προβλέψεις, δεν είναι αναγκαίο να πάρουμε όλες τις προηγούμενες τιμές του μακρινού παρελθόντος, γιατί δεν έχουν κάποια σχέση με το παρόν. Οπότε

προσπαθούμε να εκτιμήσουμε (προσαρμόσουμε) τη καμπύλη με ευθύγραμμα τμήματα. Εκτός και αν η σχέση (το υπόδειγμα) της ευθείας για όλα τα σημεία λειτουργεί ακόμη.

Στην περίπτωση αυτή εφαρμόζουμε τη μέθοδο των ελαχίστων τετραγώνων αλλά όχι για όλες τις τιμές αλλά για κάποιες k τιμές:

$$SSE = \sum_{i=t-k+1}^{t} (Y - \hat{Y})^2 = \sum_{i=t-k+1}^{t} (Y - bt - a).$$

Οπότε η εφαρμογή της μεθόδου ελαχίστων τετραγώνων δεν οδηγεί σε μία μέση ευθεία αλλά σε ευθύγραμμα τμήματα. Η περιγραφή γίνεται με την ακόλουθη συλλογιστική και τις ακόλουθες σχέσεις.

Στην αρχή παίρνουμε τον πρώτο κινητό μέσο και μετά το κινητό μέσο των πρώτων κινητών μέσων πάντα με τον ίδιο αριθμό όρων. Οι τιμές του πρώτου κινητού μέσου είναι μικρότερες από τις πραγματικές τιμές, ενώ οι τιμές του διπλού κινητού μέσου είναι μικρότερες και από αυτές του πρώτου κινητού μέσου. Στο τέλος με τη βοήθεια μιας ευθείας κάνουμε την πρόβλεψη για p περιόδους μπροστά.

$$M_t = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-k+1}}{k}$$

$$M_t' = \frac{M_t + M_{t-1} + \ldots + M_{t-k+1}}{k}$$

$$\hat{Y}_{t+p} = a_t + b_t p, \text{ με:}$$

$$a_t = M_t + (M_t - M_t') = 2M_t - M_t' \text{ και}$$

$$b_t = \frac{2}{k-1}(M_t - M_t')$$

- M_t = ο πρώτος κινητός μέσος.
- M_t' = ο διπλός κινητός μέσος.

- a_t = ο πρώτος κινητός μέσος συν την διαφορά του με το διπλό.
- b_t = η κλίση της ευθείας πρόβλεψης από σημείο σε σημείο. Η κλίση δεν είναι σταθερή παντού.

Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί με p > 1 για τη διενέργεια προβλέψεων για περισσότερες από μια μελλοντικές περιόδους, ενώ για p = 1 δίνει την πρόβλεψη για την επόμενη περίοδο. Βέβαια, η χρήση της προϋποθέτει την ύπαρξη μεγάλου αριθμού παρατηρήσεων, ιδιαίτερα μάλιστα, όταν η τιμή του k είναι σχετικά μεγάλη. Όπως και στη μέθοδο του απλού κινητού μέσου μπορούμε να επιλέξουμε εκείνη την τιμή του k, που ελαχιστοποιεί την τιμή του κριτηρίου MSE, ή κάποιου άλλου κριτηρίου στα δεδομένα της χρονοσειράς. Μπορούμε επίσης να δημιουργήσουμε και εδώ ένα διάστημα εμπιστοσύνης με κάποιο επίπεδο σημαντικότητας.

Παράδειγμα (Τ4-4)

Μια εταιρεία ενοικιαζόμενων ταινιών αναπτύσσεται και πρέπει να επεκταθεί για να ανταπεξέλθει στην αυξανόμενη ζήτηση. Ο πρόεδρος της εταιρείας θέλει να προβλέψει τις ενοικιάσεις για τον επόμενο μήνα. Ο απλός κινητός μέσος 3 εβδομάδων είναι:

$$M_{15}=\frac{728+711+712}{3}=717,$$

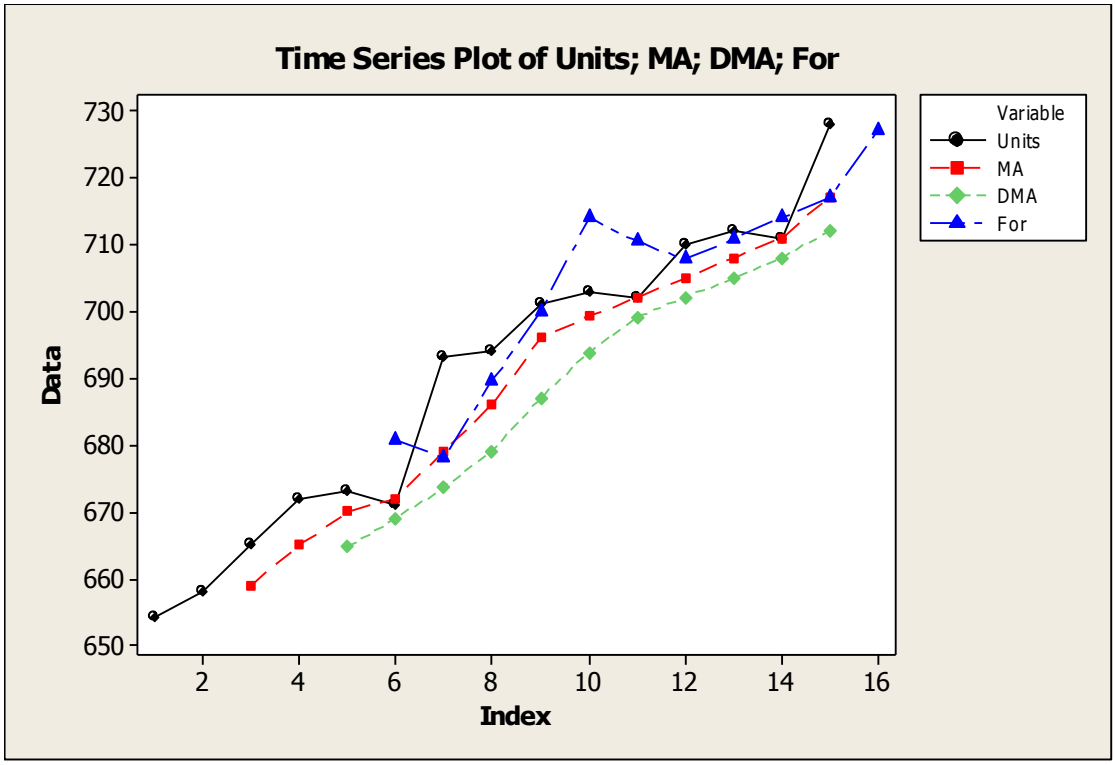
$$M'_{15}=\frac{717+711+708}{3}=712,$$

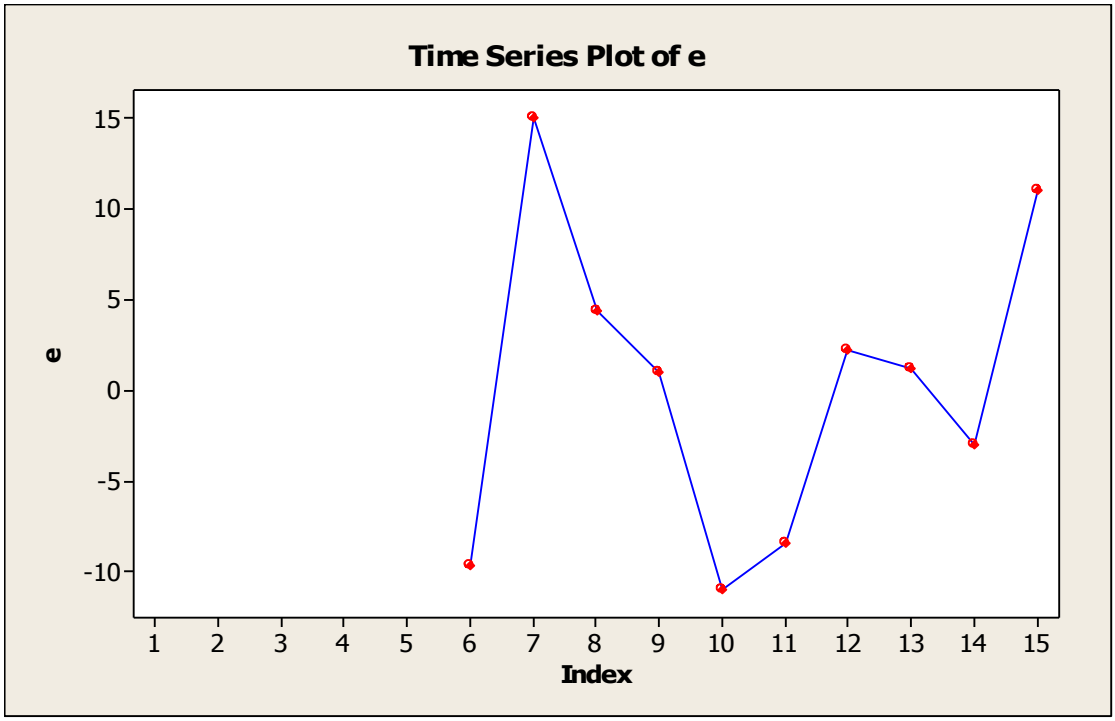
$$a_{15}=717+(717-712)=722,$$

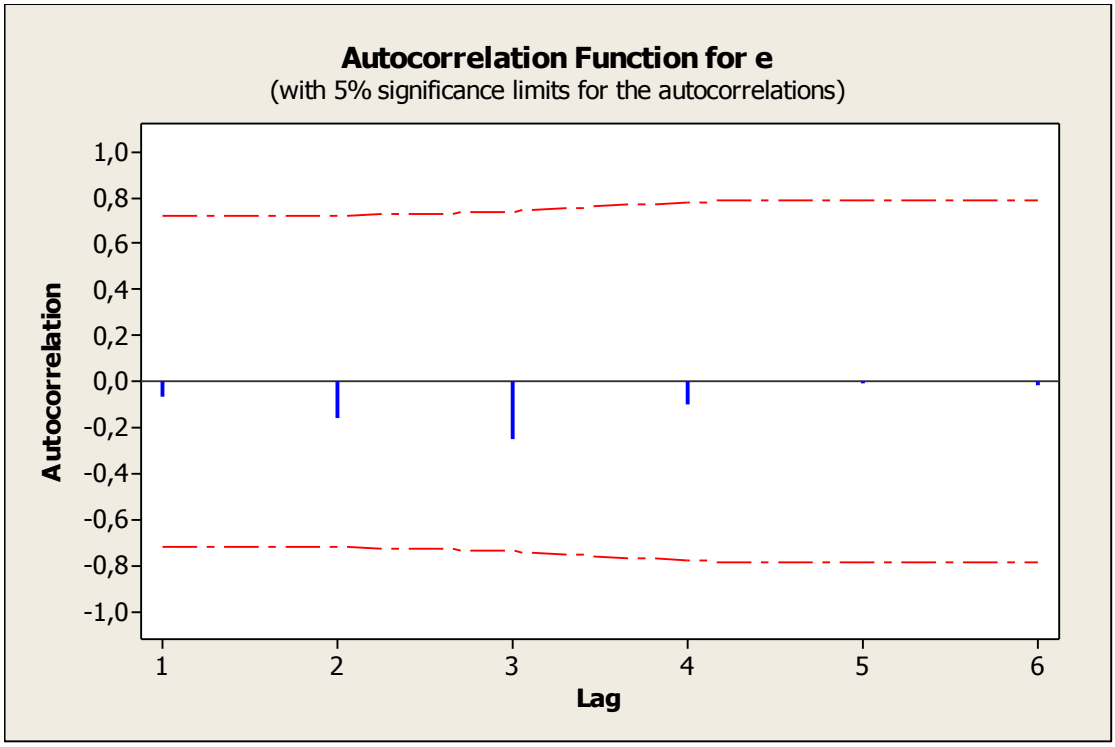
$$b_{15}=\frac{2}{3-1}(717-712)=5,$$

$$\hat{Y}_{15+1}=722+5(1)=727,$$

$$\hat{Y}_{15+4}=722+5(4)=742.$$


Time Series Plot of Units; MA; DMA; For
Data
Index
Variable
Units
MA
DMA
For
730
720
710
700
690
680
670
660
650
2
4
6
8
10
12
14
16


Time Series Plot of e
e
Index
15
10
5
0
-5
-10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15



Το MSE έχει μειωθεί από 133 σε 63.7 σε σχέση με τον απλό κινητό μέσο και επίσης δεν υπάρχει αυτοσυσχέτιση.

4.4 ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ (ΑΠΛΗ)

Ενώ η μέθοδος του κινητού μέσου όρου λαμβάνει υπόψη της μόνο τις τελευταίες παρατηρήσεις, η μέθοδος της *εκθετικής εξομάλυνσης* (exponential smoothing) παίρνει υπόψη της όλες τις τιμές με διαφορετικό όμως βάρος. Η μέθοδος αυτή χρησιμοποιείται για πρόβλεψη στο άμεσο μέλλον και κυρίως σε προβλήματα αποθεματικών.

Η εξίσωση της *απλής εκθετικής εξομάλυνσης* (simple exponential smoothing) είναι:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t,$$

όπου t είναι η τρέχουσα περίοδος, τα $\hat{Y}_{t+1}$ *και* $\hat{Y}_t$, είναι τιμές πρόβλεψης για την επόμενη και την τρέχουσα περίοδο και Y_t είναι η τιμή που παρατηρήθηκε την τρέχουσα περίοδο. Το α ονομάζεται σταθερά

εξομάλυνσης και παίρνει τιμές από 0 έως 1. Αφού η παραπάνω εξίσωση περιλαμβάνει μόνο μία σταθερά, το μοντέλο αυτό είναι μοντέλο εκθετικής εξομάλυνσης μιας παραμέτρου. Η εξίσωση αυτή παράγεται αν ορίσουμε και ελαχιστοποιήσουμε την ακόλουθη σχέση:

$$SSE = \sum_{1}^{t} b^t (Y_t - \hat{Y}_t),\ με\ 0 < b < 1$$

και μετά θέσουμε $1-b=a$ *και* $b^t = 0$ για t πάρα πολύ μεγάλο.

Η εκθετική εξομάλυνση μιας παραμέτρου είναι πολύ απλή μέθοδος, αφού μόνο μια τιμή, η πρόβλεψη της τελευταίας περιόδου, είναι αυτή που πρέπει να διασωθεί. Στην ουσία, ολόκληρη η χρονοσειρά εμπεριέχεται σ' αυτή την πρόβλεψη. Εάν εκφράσουμε το $\hat{Y}_t$, σε όρους της προηγηθείσας παρατήρησης Y_{t-1} και των τιμών της πρόβλεψης $\hat{Y}_{t-1}$, τότε το ισοδύναμο για την πρόβλεψη της επόμενης περιόδου γίνεται:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)[\alpha Y_{t-1} + (1-\alpha)\hat{Y}_{t-1}]$$

ή πιο απλά:

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha) Y_{t-1} + (1-\alpha)^2 \hat{Y}_{t-1}$$

Η νέα αυτή εξίσωση είναι μοντέλο δευτεροβάθμιας εκθετικής εξομάλυνσης μιας παραμέτρου.

Μπορούμε να συνεχίσουμε, έτσι για έναν αριθμό προηγούμενων περιόδων, πράγμα που δείχνει, ότι όλες οι προηγηθείσες τιμές του Y αντανακλώνται στην τρέχουσα πρόβλεψη. Έτσι, το όνομα αυτής της διαδικασίας προέρχεται από τις διαδοχικές σταθμίσεις α, α(1-α), α(1-α)^2, α(1-α)^3,..., οι οποίες μειώνονται εκθετικά.

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha) Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \alpha(1-\alpha)^3 Y_{t-3} + ...$$

Το άθροισμα $S=\alpha+\alpha(1-\alpha)+\alpha(1-\alpha)^2+\alpha(1-\alpha)^3+...=1$ και αυτό υποδηλώνει ότι η πρόβλεψη είναι σταθμισμένος μέσος όρος.

Οι πιο πρόσφατες περίοδοι στη χρονοσειρά λαμβάνουν μεγαλύτερη στάθμιση στον υπολογισμό της πρόβλεψης. Πρακτικά, οι αρκετά παλιές τιμές της Y εξαιρούνται. Η διαδικασία πρόβλεψης μπορεί να τροποποιηθεί οποιαδήποτε στιγμή με τη μεταβολή της τιμής της α.

Μπορούμε να ξαναγράψουμε την εξίσωση ως εξής:

$$\hat{Y}_{t+1}=\hat{Y}_t+\alpha(Y_t-\hat{Y}_t)=\hat{Y}_t+\alpha e_t$$

Δηλαδή η πρόβλεψη στο t+1 είναι η πρόβλεψη στο t συν έναν παράγοντα επί της διαφοράς της πραγματικής τιμής μείον την τιμή της πρόβλεψης στο t, όπου το αe_t το σφάλμα πρόβλεψης για την περίοδο t. Επομένως, βλέπουμε ότι η πρόβλεψη που δίνεται από την εκθετική εξομάλυνση, είναι η παλαιά πρόβλεψη συν μια προσαρμογή για το σφάλμα, που έγινε στην τελευταία πρόβλεψη. Όταν το α βρίσκεται πλησίον του 1, η νέα πρόβλεψη περιέχει μια ουσιώδη προσαρμογή για το σφάλμα της προηγούμενης πρόβλεψης. Αντίθετα, εάν το α βρίσκεται πολύ κοντά στο 0, η νέα πρόβλεψη θα περιέχει μικρή μόνο προσαρμογή για το σφάλμα.

Ένας τρόπος υπολογισμού του α είναι να ελαχιστοποιήσουμε το MSE, το οποίο εξαρτάται από το α. Ο υπολογισμός μπορεί να γίνει και με επαναληπτική μέθοδο.

Ως αρχική τιμή παίρνουμε ή $\hat{Y}_1=Y_1$ ή $\hat{Y}_1=\frac{1}{k}\sum_{i=1}^{k}Y_i$.

Παράδειγμα (Τ4-7)

Η εταιρία με τα πριόνια και για τα έτη 2000 – 2006 προσαρμόζει τις τιμές και προβλέπει με την μέθοδο της εκθετικής εξομάλυνσης, χρησιμοποιώντας τις σταθερές εξομάλυνσης 0.1 και 0.6. Η εξομάλυνση

των τιμών της χρονοσειράς υπολογίζεται καθορίζοντας το Y_1=500. Από τη χρονική περίοδο 2 η πρόβλεψη για την περίοδο 3 με α = 0.1 είναι:

$$\hat{Y}_{2+1}=\alpha Y_2+(1-\alpha)\hat{Y}_2, \quad \hat{Y}_3=0.1(350)+ .9(500)=485$$

Το σφάλμα πρόβλεψης είναι:

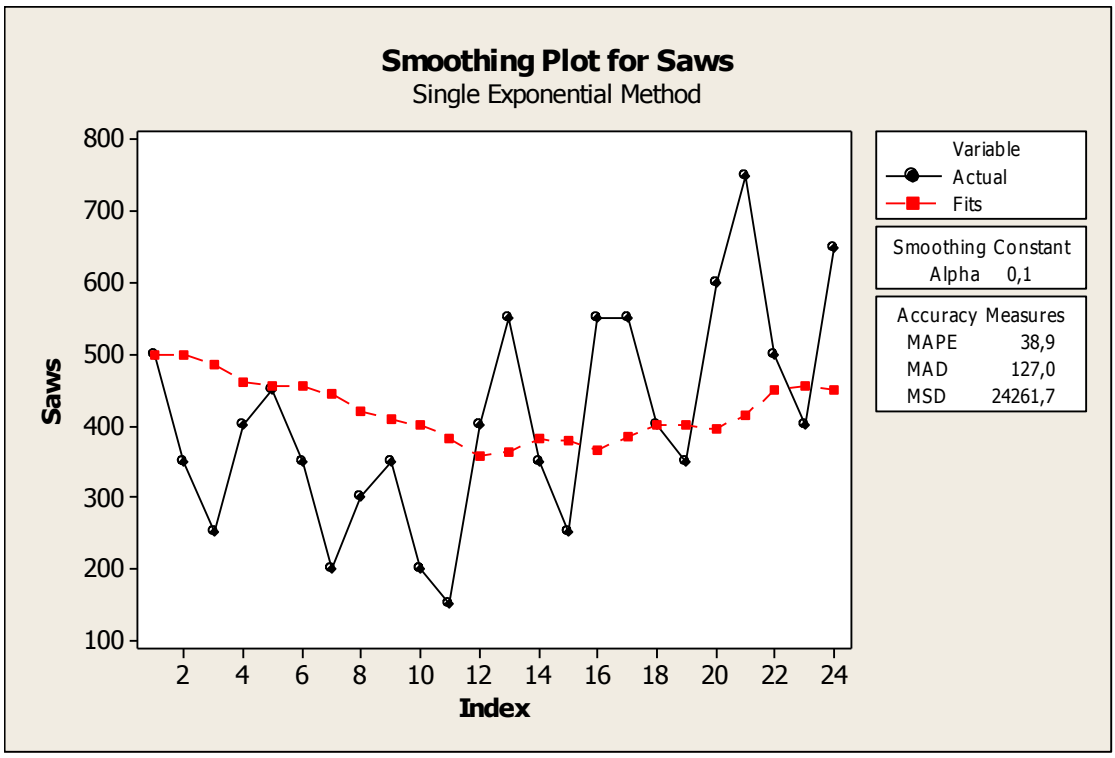
$$e_3=Y_3-\hat{Y}_3=250-485=-235$$

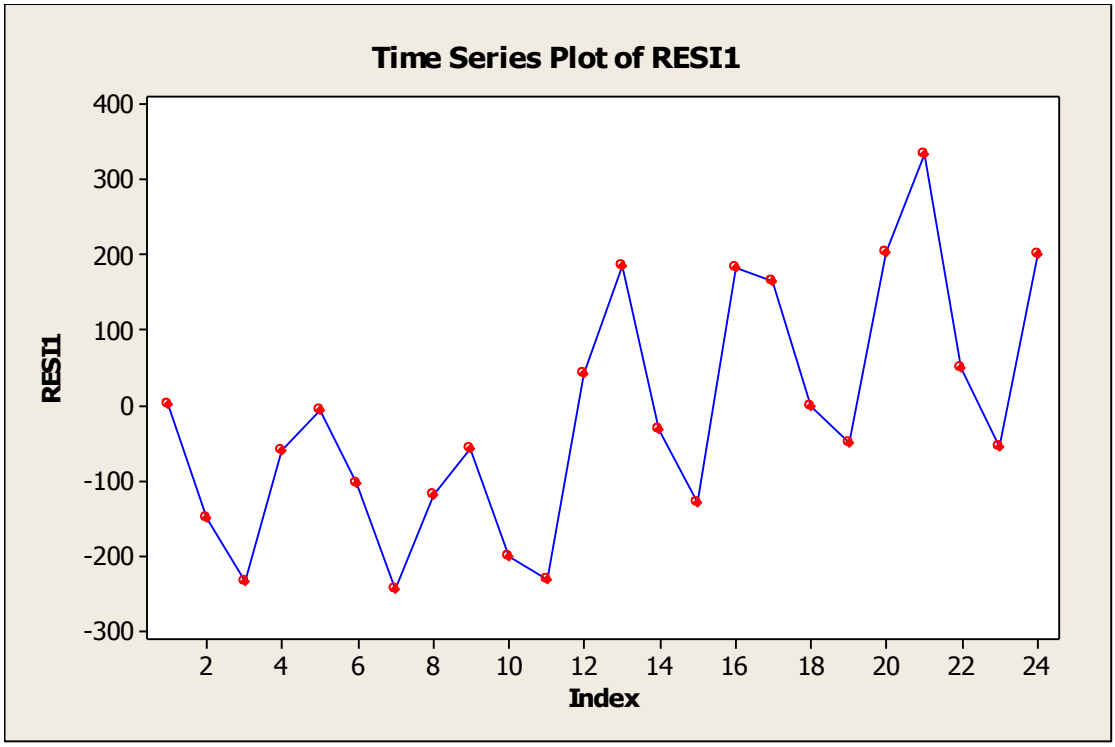
Η πρόβλεψη για την τέταρτη περίοδο είναι:

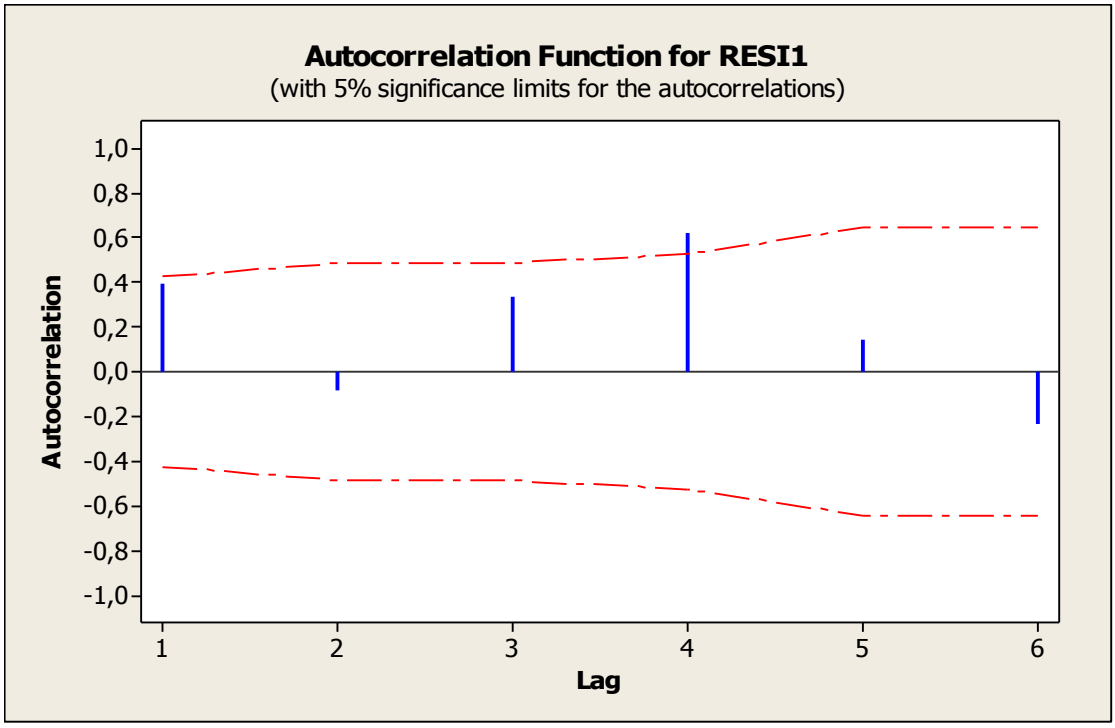
$$\hat{Y}_{3+1}=\alpha Y_3+(1-\alpha)\hat{Y}_3, \quad \hat{Y}_4=0.1(250)+ 0.9(485)=461.5$$

Για α = 0.1 έχουμε: MSE = 24,262, MAPE = 38.9%

Για α = 0.6 έχουμε: MSE = 22,248, MAPE = 36.5%



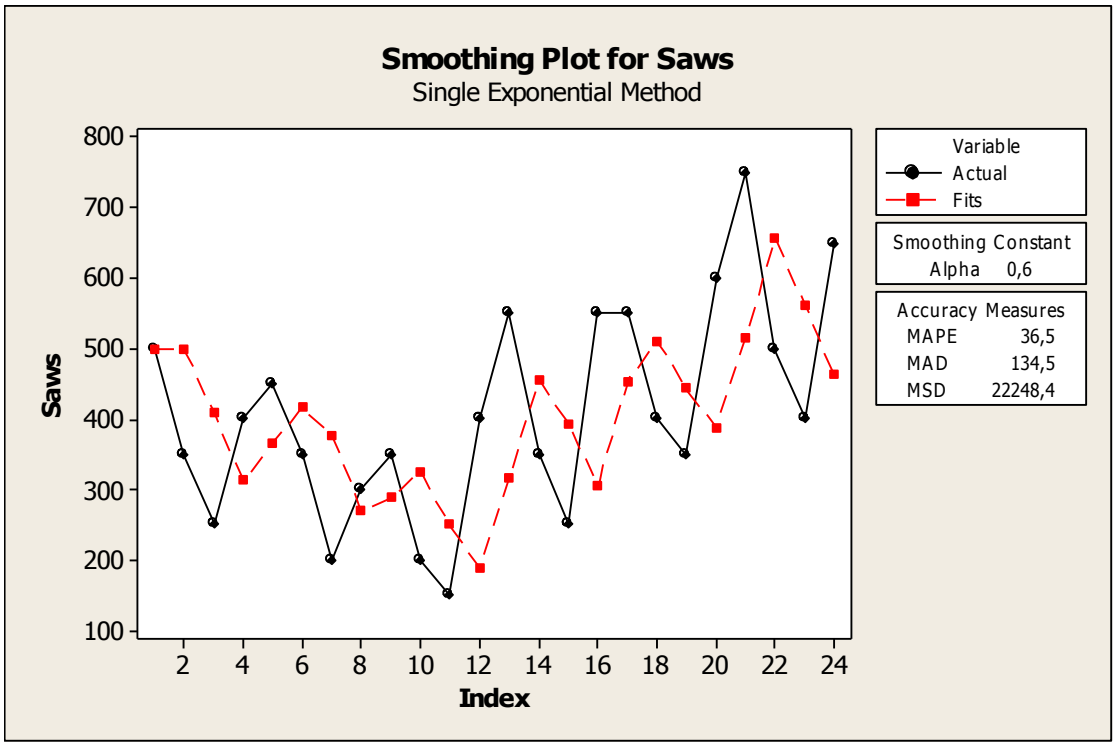


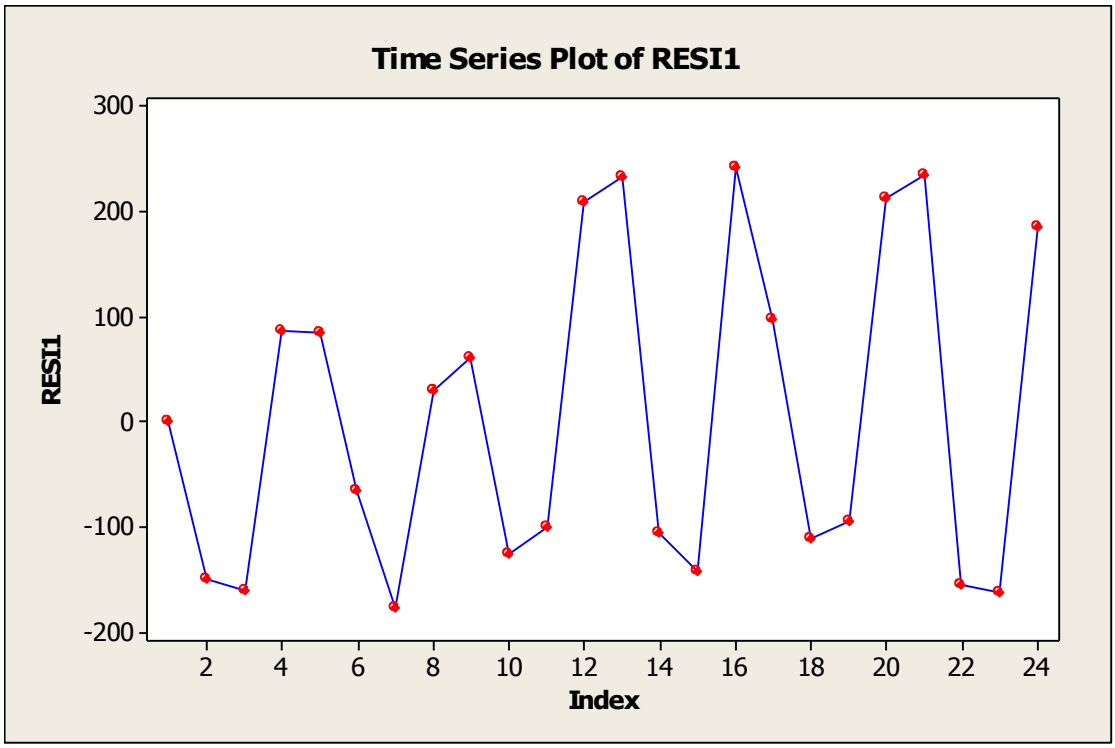


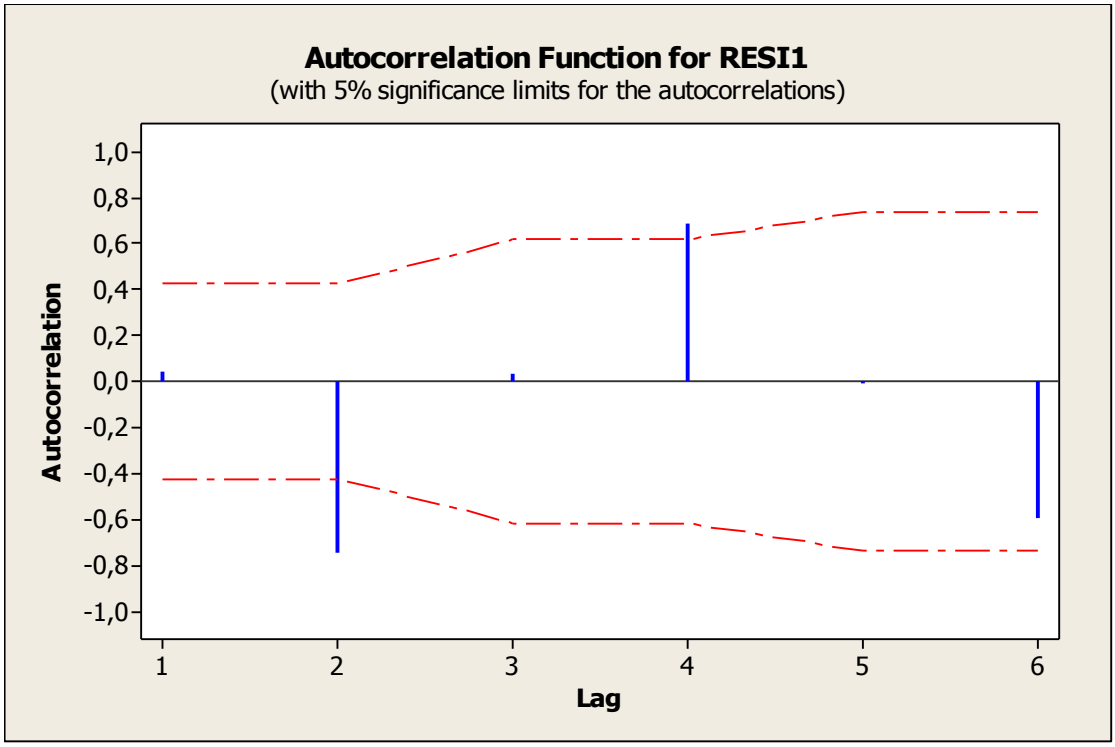
Lag	ACF	T	LBQ
1	0,395377	1,94	4,24
2	-0,084768	-0,36	4,44
3	0,330875	1,41	7,70
4	0,614720	2,42	19,49
5	0,137906	0,45	20,11

6 -0,232801 -0,75 21,99

Η τιμή 21,99 είναι πολύ μεγαλύτερη από το όριο εμπιστοσύνης της μηδενικής υπόθεσης 12.59, οπότε απορρίπτεται και τα σφάλματα αυτοσυσχετίζονται.







Lag	ACF	T	LBQ
1	0,039008	0,19	0,04
2	-0,746934	-3,65	15,87
3	0,036115	0,12	15,90
4	0,682690	2,30	30,45
5	-0,006159	-0,02	30,45
6	-0,595083	-1,67	42,72

Η τιμή 42,72 είναι πολύ μεγαλύτερη από το όριο εμπιστοσύνης της μηδενικής υπόθεσης 12.59, οπότε απορρίπτεται και τα σφάλματα αυτοσυσχετίζονται.

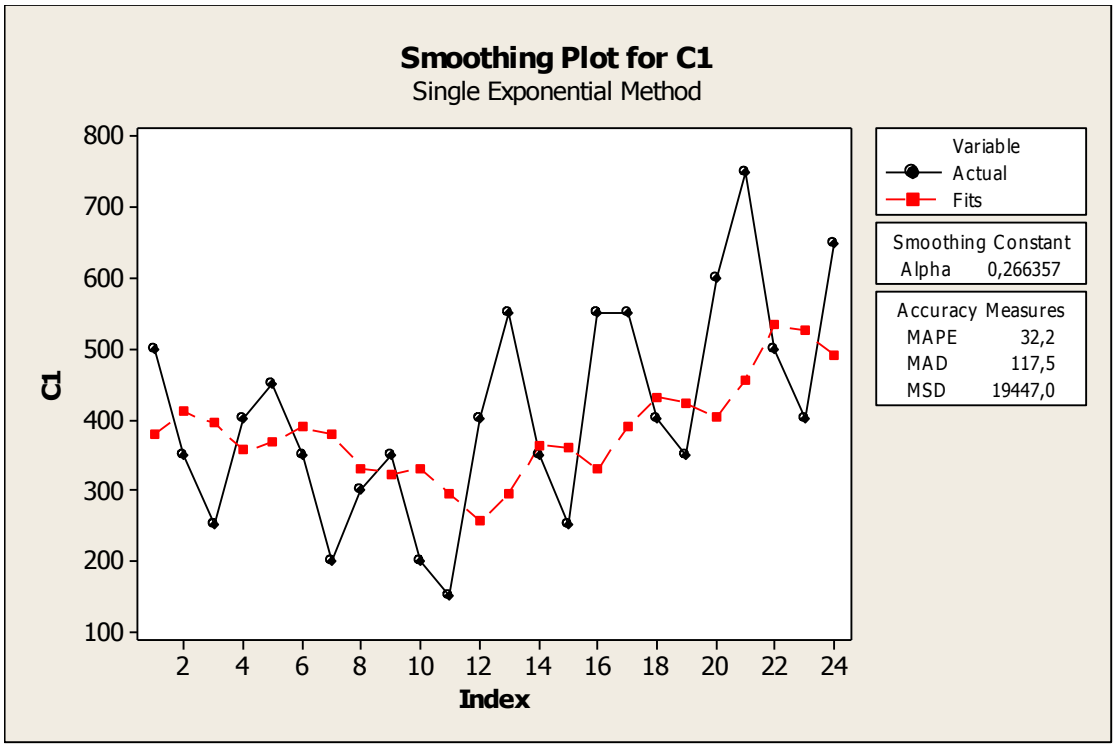
Αν k = 6 τότε έχουμε:

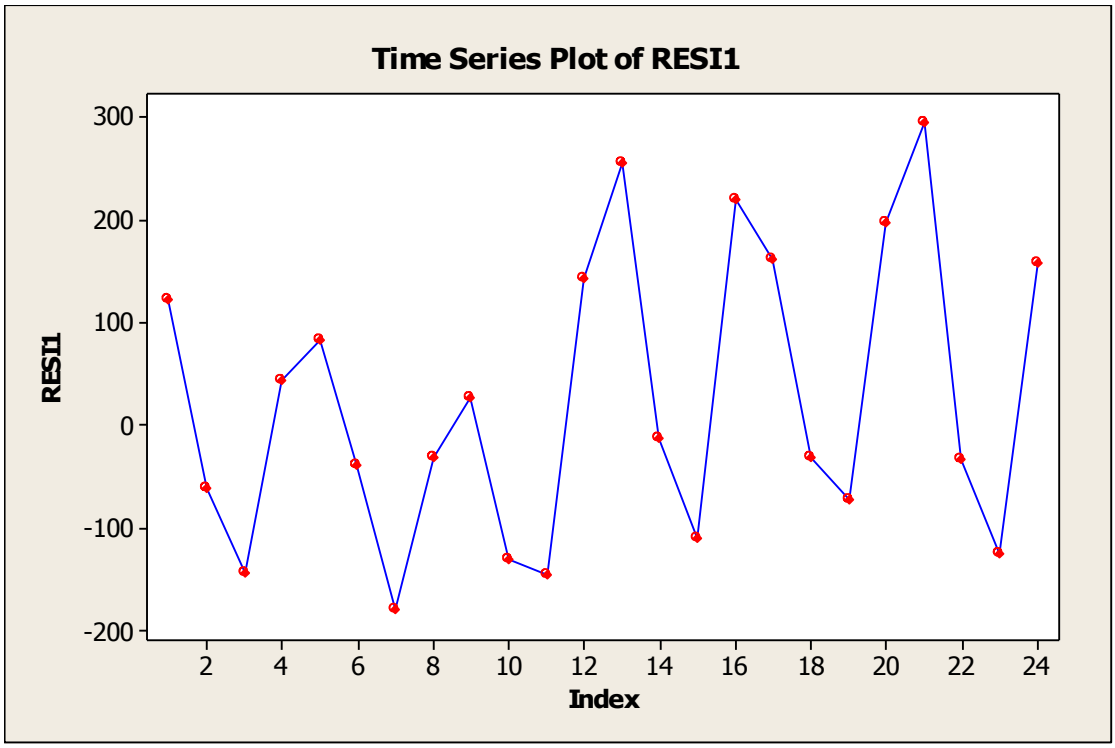
$$\hat{Y}_1 = \frac{1}{6}\sum_{i=1}^{6} Y_i = \frac{1}{6}(500 + 350 + 250 + 400 + 450 + 350) = 383.3$$

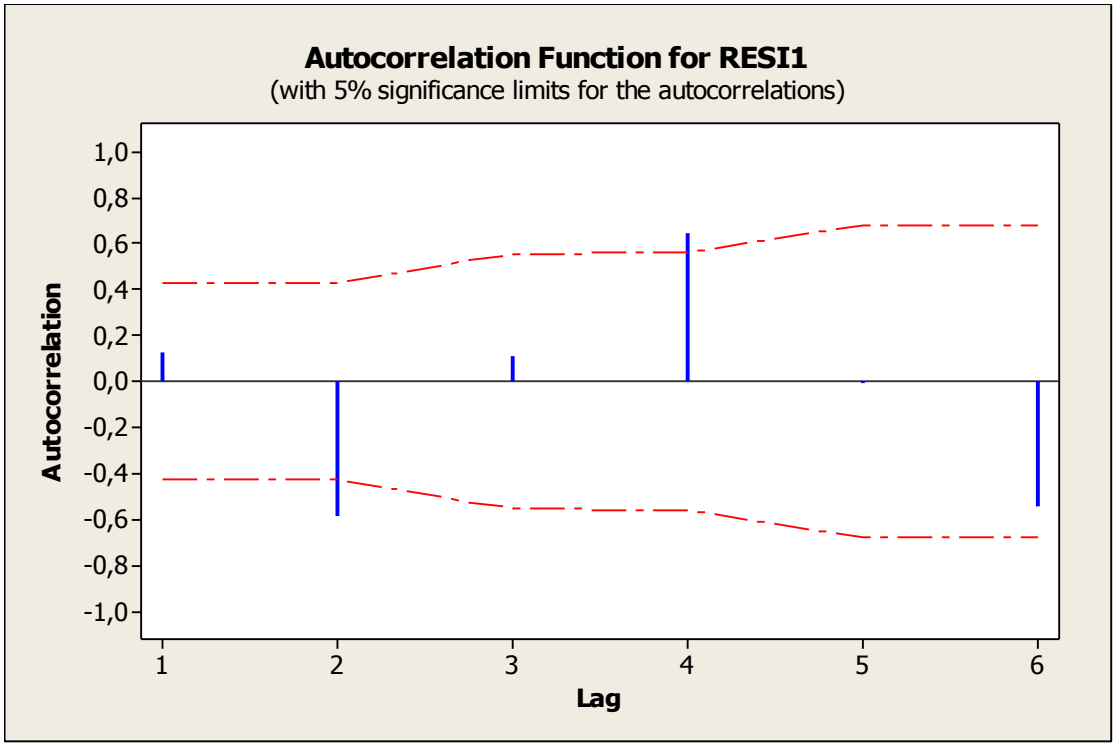
Ξανακάνουμε πάλι τους υπολογισμούς και βρίσκουμε τα ακόλουθα: Για α = 0.1 έχουμε: MSE = 21,091, MAPE = 32.1%, Για α = 0.6 έχουμε: MSE = 22,152, MAPE = 36.7%

Στο ακόλουθο διάγραμμα υπολογίζεται η παράμετρος α αυτόματα

από το MINITAB, έτσι ώστε το MSE να είναι ελάχιστο. Συγκεκριμένα ορίζει ένα ARIMA μοντέλο, το οποίο είναι ισοδύναμο με την απλή εκθετική εξομάλυνση (Κεφάλαιο 9).







Lag	ACF	T	LBQ
1	0,121421	0,59	0,40
2	-0,588941	-2,84	10,24
3	0,109122	0,41	10,59
4	0,646967	2,40	23,65
5	-0,006908	-0,02	23,65
6	-0,542625	-1,65	33,86

Η τιμή 33.86 είναι πολύ μεγάλη σε σχέση με την αντίστοιχη της χ^2 κατανομής 12.59 και σε 3 περιπτώσεις ο συντελεστής αυτοσυσχέτισης είναι μεγαλύτερος από τα όρια εμπιστοσύνης. Άρα τα κατάλοιπα δεν είναι τυχαία.

4.5 ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ ΜΕ ΠΡΟΣΑΡΜΟΓΗ ΣΤΗΝ ΤΑΣΗ (ΔΙΠΛΗ)

Ονομάζεται επίσης μέθοδος Holt και χρησιμοποιείται, όταν υπάρχει τάση στις παρατηρήσεις της χρονοσειράς. Η μέθοδος Holt έχει δύο παραμέτρους εξομάλυνσης, την παράμετρο α για την εξομάλυνση

του επιπέδου των τιμών της χρονοσειράς και την παράμετρο β για την εξομάλυνση της τάσης, σε αντίθεση με τη μέθοδο της απλής εκθετικής εξομάλυνσης που έχει μόνο μία. Η εφαρμογή της μεθόδου βασίζεται στην ακόλουθη διαδικασία:

Η εξομάλυνση του επιπέδου των τιμών (current Level) της χρονοσειράς γίνεται με την ακόλουθη σχέση:

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1}),$$

όπου α είναι η σταθερά για την εξομάλυνση για $0 \leq \alpha \leq 1$, L_t οι εξομαλυνθείσες τιμές της χρονοσειράς, που προκύπτουν από την εξομάλυνση, για t=2, 3, ..., n, ***ενώ για t=1 ορίζεται ως αρχική συνθήκη*** $\boldsymbol{L_1 = Y_1}$. Η σχέση αυτή είναι όμοια με τη σχέση της απλής εκθετικής εξομάλυνσης εκτός από τον παράγοντα T_{t-1}, ο οποίος εισήχθη προκειμένου να ενσωματώσουμε τη τάση στη περίπτωση που υπάρχει. Χωρίς τον παράγοντα αυτό η σχέση είναι ίδια με τη σχέση της απλής εκθετικής εξομάλυνσης. Το Level είναι η εξομαλυνθείσα τιμή των δεδομένων στο τέλος μιας περιόδου.

Η εξομάλυνση της τάσης (Trend estimate) γίνεται ως εξής:

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

όπου β, για $0 \leq \beta \leq 1$, είναι η σταθερά για την εξομάλυνση της τάσης, T_t οι εξομαλυνθείσες τιμές της τάσης, για t=2, 3, ..., n, ***ενώ για t=1 ορίζεται ως αρχική συνθήκη η*** $\boldsymbol{T_1 = 0}$. Η παρούσα τάση σταθμίζεται με δύο παράγοντες: Ο πρώτος προέρχεται από τη διαδοχική διαφορά των εξομαλυνθέντων τιμών της χρονοσειράς ενώ ο δεύτερος από την εξομαλυνθείσα τιμή τάσης της προηγούμενης περιόδου. Το Trend λοιπόν είναι η εξομαλυνθείσα τιμή της αύξησης, ή της μείωσης των δεδομένων στο τέλος μιας περιόδου.

Η πρόβλεψη, για την p μελλοντική περίοδο υπολογίζεται ως:

$$\hat{Y}_{t+p} = L_t + pT_t$$

όπου p = 1, 2, 3, ….

Οι τιμές των παραμέτρων α και β για μια συγκεκριμένη χρονοσειρά είναι αυτές, που ελαχιστοποιούν την τιμή του κριτηρίου MSE, ή κάποιου άλλου κριτηρίου στα δεδομένα της χρονοσειράς. Οι σχέσεις αυτές προκύπτουν αν πάρουμε την αναμενόμενη τιμή της εξομαλυνθείσας τιμής στην απλή εκθετική εξομάλυνση και μετά ορίσουμε δεύτερη εξομάλυνση στην αναμενόμενη τιμή της πρώτης εξομάλυνσης.

Παράδειγμα.

Στο προηγούμενο παράδειγμα η απλή εκθετική εξομάλυνση δεν είχε επιτυχείς προβλέψεις για την εταιρία. Επειδή, ίσως να υπάρχει τάση στα δεδομένα αυτά χρησιμοποιείται η μέθοδος Holt. Έχουμε α = 0.3 και β = 0.1. Η εξομάλυνση των τιμών της χρονοσειράς είναι:

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$L_2 = 0.3Y_2 + (1-0.3)(L_{2-1} + T_{2-1})$$

$$L_2 = 0.3(350) + 0.7(500+0) = 455$$

Η εξομάλυνση της τάσης είναι:

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

$$T_2 = 0.1(L_2 - L_{2-1}) + (1-0.1)T_{2-1}$$

$$T_2 = 0.1(455-500) + 0.9(0) = -4.5$$

Η πρόβλεψη για μία μελλοντική περίοδο είναι:

$$\hat{Y}_{t+p} = L_t + pT_t$$

$$\hat{Y}_3 = L_2 + pT_2 = 455 + 1(-4.5) = 450.5$$

Το σφάλμα πρόβλεψης είναι:

$$e_3 = Y_3 - \hat{Y}_3 = 250 - 450.5 = -200.5$$

Η εξομάλυνση των τιμών της χρονοσειράς είναι:

$$L_{24} = 0.3\,Y_{24} + (1-0.3)(L_{24-1} + T_{24-1})$$

$$L_{24} = 0.3(650) + 0.7(517.6 + 9.8) = 564.2$$

Η εξομάλυνση της τάσης είναι:

$$T_{24} = 0.1(L_{24} - L_{24-1}) + (1-0.1)T_{24-1}$$

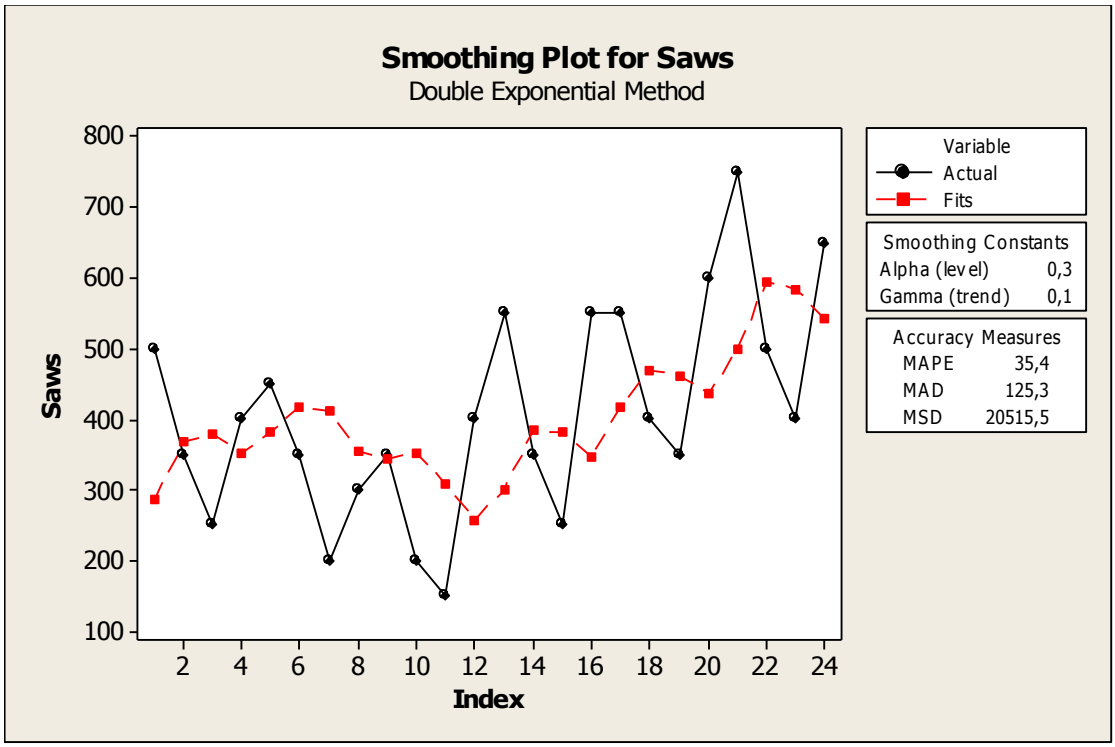
$$T_{24} = 0.1(564.2 - 517.6) + 0.9(9.8) = 13.5$$

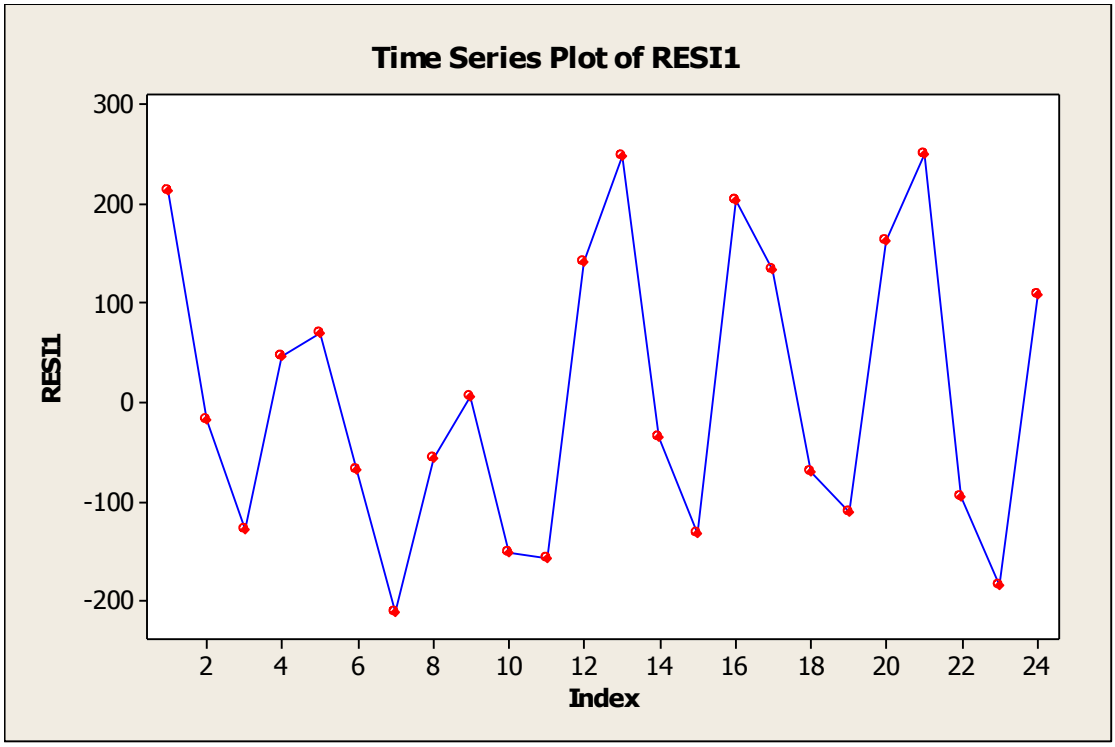
Η πρόβλεψη για μία μελλοντική περίοδο είναι:

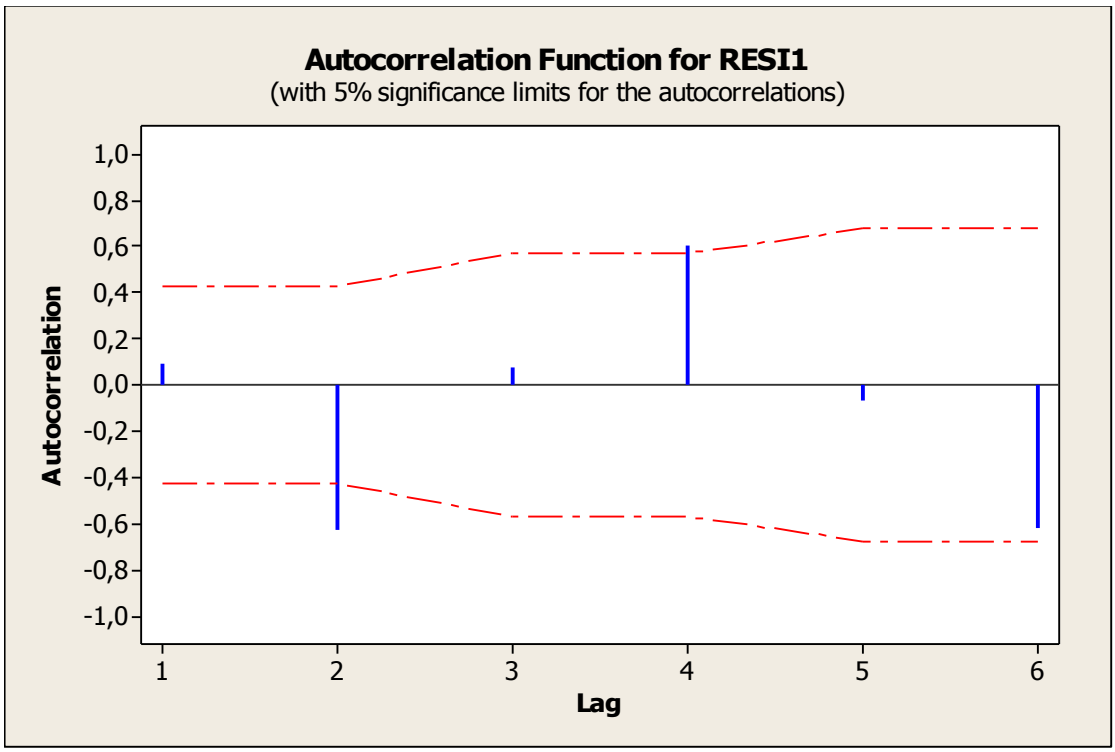
$$\hat{Y}_{24+1} = L_{24} + 1\mathrm{T}_{24} = 564.2 + 1(13.5) = 577.7$$

Για την απλή εκθετική εξομάλυνση και α = 0.266 έχουμε: MSE = 19,447, MAPE = 32.2%.

Για α = 0.3, β = 0.1 έχουμε: MSE = 20,516, MAPE = 35.4%.





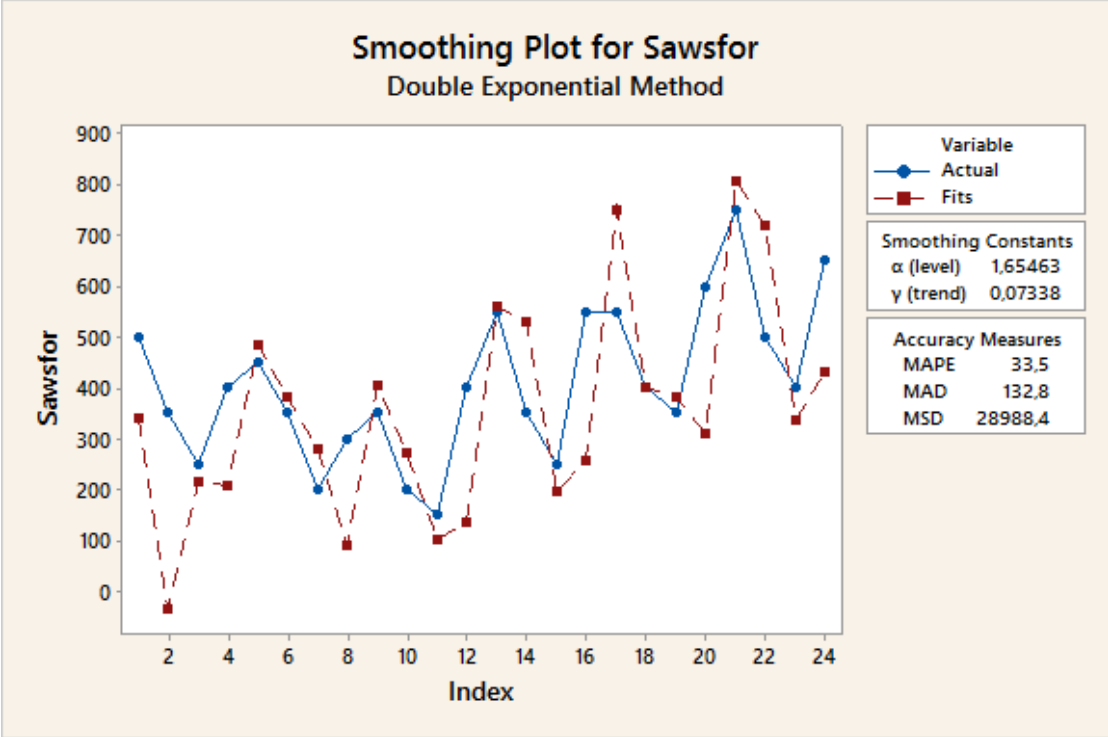


Lag	ACF	T	LBQ
1	0,092547	0,45	0,23
2	-0,629408	-3,06	11,47
3	0,074355	0,27	11,63
4	0,601371	2,18	22,92

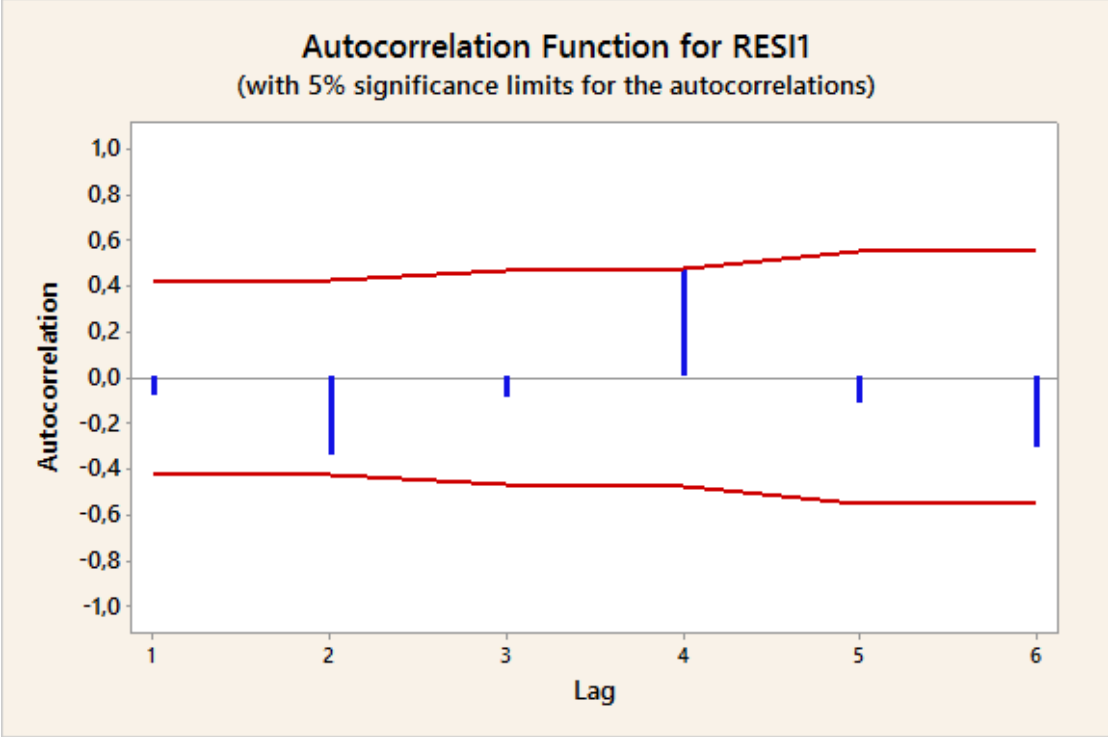
5	-0,070132	-0,22	23,08
6	-0,618267	-1,90	36,33

Η τιμή 36,33 είναι πολύ μεγαλύτερη από το όριο εμπιστοσύνης της μηδενικής υπόθεσης 12.59, οπότε απορρίπτεται και τα σφάλματα αυτοσυσχετίζονται.

Στο ακόλουθο διάγραμμα υπολογίζεται η παράμετρος α και β αυτόματα από το MINITAB, χωρίς να εγγυάται ότι το MSE θα είναι ελάχιστο. Συγκεκριμένα ορίζει ένα ARIMA μοντέλο, το οποίο είναι ισοδύναμο με την διπλή εκθετική εξομάλυνση (Κεφάλαιο 9).



Lag	ACF	T	LBQ
1	-0,078054	-0,38	0,17
2	-0,339556	-1,65	3,44
3	-0,088625	-0,39	3,67
4	0,466857	2,04	10,47
5	-0,117911	-0,44	10,93
6	-0,306342	-1,14	14,18



Οι τιμές των παραμέτρων υπολογιζομένων από τα ARIMA μοντέλα μπορεί να είναι μεταξύ 0 και 2.

4.6 ΕΚΘΕΤΙΚΗ ΕΞΟΜΑΛΥΝΣΗ ΜΕ ΠΡΟΣΑΡΜΟΓΗ ΣΤΗΝ ΤΑΣΗ ΚΑΙ ΣΤΗΝ ΕΠΟΧΙΚΟΤΗΤΑ (ΤΡΙΠΛΗ)

Ο Winters ανέπτυξε μια μέθοδο για την προσαρμογή της εποχικής, ή περιοδικής κίνησης μέσα στο πλαίσιο της γραμμικής εκθετικής εξομάλυνσης με ή χωρίς τάση. Επομένως, η διαδικασία του Winters μπορεί να εφαρμοστεί για προβλέψεις με βάση μια χρονοσειρά, που εμφανίζει και τάση και εποχικό πρότυπο. Η επέκταση συνίσταται στην ύπαρξη επιπλέον εξίσωσης για τον υπολογισμό της εποχικής συνιστώσας της χρονοσειράς.

Για την ανάπτυξη της μεθόδου αυτής θεωρούμε το Multiplicative Model για την αναπαράσταση των δεδομένων μέσω των συνιστωσών.

$$Y = L \times S.$$

Μέσα στο L "κρύβεται" η τάση, η κυκλικότητα και η τυχαιότητα.

Προκειμένου να κάνουμε πρόβλεψη με αυτό το μοντέλο χρειαζόμαστε τέσσερις εξισώσεις:

Η εκθετικά εξομαλυνθείσα σειρά είναι η εξής:

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + T_{t-1})$$

Η εκτίμηση της εποχικότητας είναι η εξής:

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma) S_{t-s}$$

όπου S είναι ο παράγοντας προσαρμογής της εποχικότητας και το s είναι το μήκος της εποχικότητας.

Ο εκτίμηση της τάσης παραμένει η ίδια:

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta) T_{t-1}$$

Προβλέποντας p περιόδους στο μέλλον, έχουμε:

$$\hat{Y}_{t+p} = (L_t + pT_t) S_{t-s+p}$$

Με την πρώτη εξίσωση γίνονται επίκαιρες οι εξομαλυνθείσες τιμές της σειράς. Το L δεν εμπεριέχει την εποχικότητα. Στην εξίσωση αυτή το Y_t διαιρείται δια του S_{t-s} που προσαρμόζει τις αρχικές παρατηρήσεις, Y_t για εποχικότητα και αναιρεί τις επιδράσεις της εποχικότητας, όσο καλύτερα αυτές μπορεί να μετρηθούν από τη χρονολογική σειρά.

Η δεύτερη εξίσωση δίνει την εκτίμηση της εποχικής συνιστώσας, Y_t/L_t πολλαπλασιασμένη επί τη σταθερά γ συν την παλαιά εποχική εκτίμηση, S_{t-s} πολλαπλασιασμένη επί (1–γ). Επομένως, η επικαιροποίηση των εποχικών εκτιμήσεων είναι από μόνη της μια διαδικασία εκθετικής εξομάλυνσης. Επίσης, ο Y_t διαιρείται με L_t, προκειμένου να εκφραστεί η τιμή ως δείκτης παρά ως απόλυτο μέγεθος. Αυτό επιτρέπει την εύρεση

του μέσου όρου των νέων εποχικών εκτιμήσεων με βάση τον εποχικό δείκτη της προηγούμενης περιόδου.

Η τρίτη εξίσωση εκφράζει τη σύγχρονη τιμή της συνιστώσας της τάσης, που επιτυγχάνεται με τη συνηθισμένη διαδικασία εκθετικής εξομάλυνσης. Τέλος, μετά από αυτή την εξίσωση λαμβάνουμε την εξίσωση για τις μελλοντικές περιόδους. Η διαφορά είναι ότι αυτή η εκτίμηση για τη μελλοντική περίοδο, t+p, πολλαπλασιάζεται επί S_{t-s+p}. Αυτός είναι ο τελικός διαθέσιμος εποχικός δείκτης και αποτελεί την προσαρμογή της πρόβλεψης για εποχικότητα. ***Οι αρχικές τιμές είναι*** $S_1 = 1.0$, $T_1 = 0$, $L_1 = Y_1$. Οι παράμετροι α,β,γ ορίζονται με τέτοιο τρόπο, ώστε το MSE να γίνει ελάχιστο κάτι που σ' αυτή την περίπτωση είναι πολύ χρονοβόρο.

Παράδειγμα

α=0.4, β=0.1, γ=0.3

Level estimate:

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$L_{24} = 0.4 \frac{Y_{24}}{S_{24-4}} + (1-0.4)(L_{24-1} + T_{24-1})$$

$$L_{24} = 0.4 \frac{650}{1.39628} + (1-0.4)(501.286 + 9.148) = 492.469$$

Trend estimate:

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

$$T_{24} = 0.1(L_{24} - L_{24-1}) + (1-0.1)T_{24-1}$$

$$T_{24} = 0.1(492.469 - 501.286) + 0.9(9.148) = 7.352$$

Seasonality estimate:

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma) S_{t-s}$$

$$S_{24} = 0.3 \frac{Y_{24}}{L_{24}} + (1-0.3) S_{24-4}$$

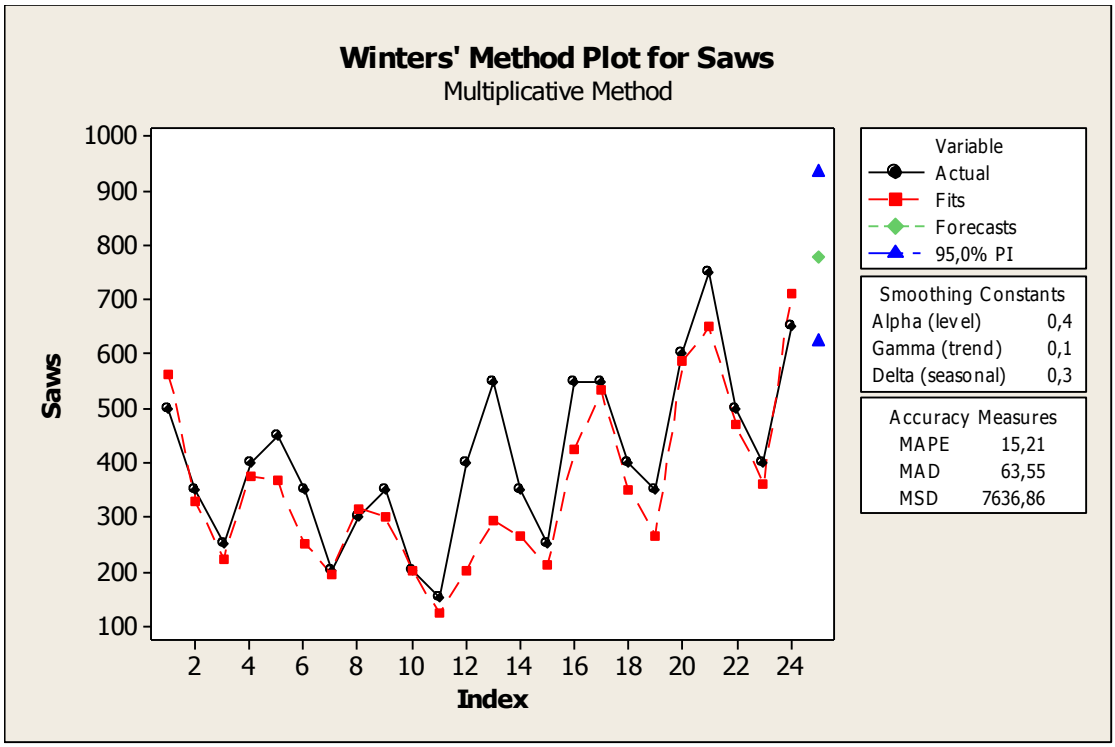
$$S_{24} = 0.3 \frac{650}{492.469} + 0.7(1.39628) = 1.3734$$

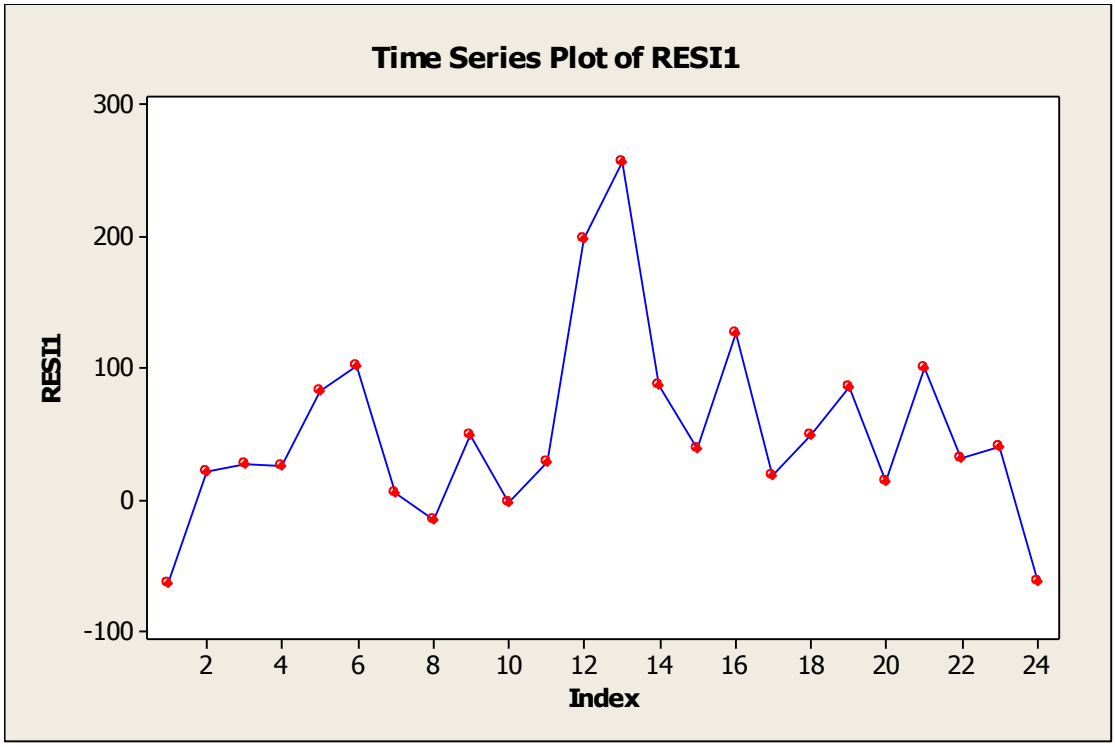
Forecast:

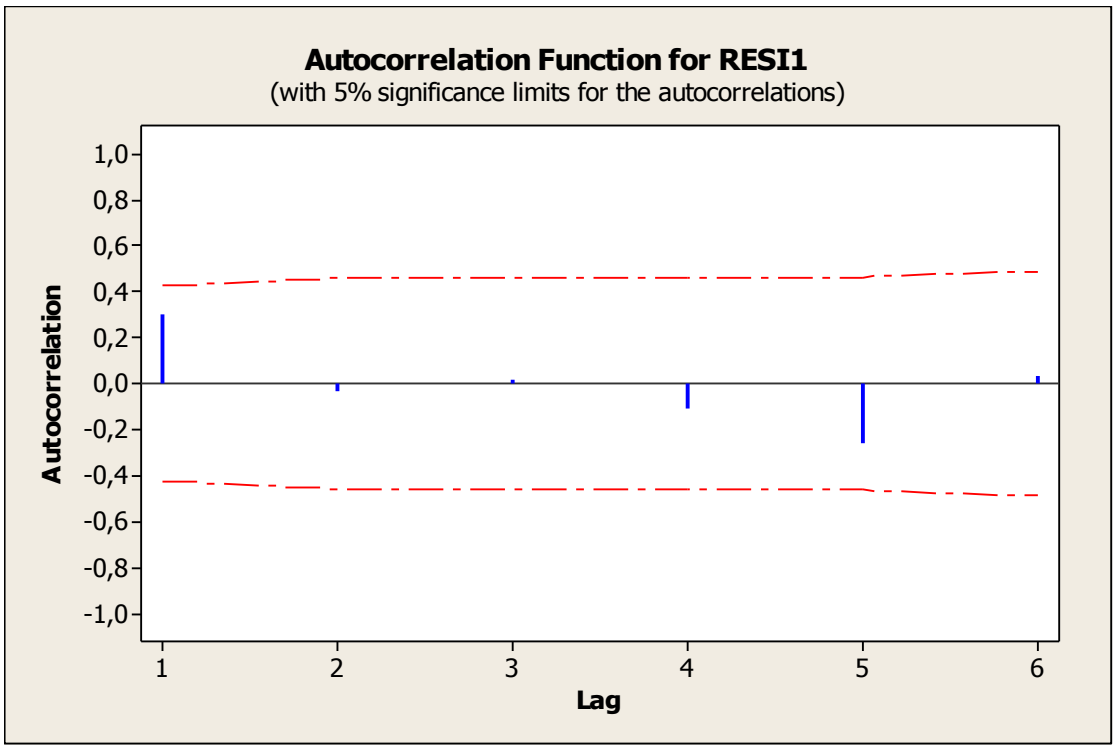
$$\hat{Y}_{t+p} = (L_t + pT_t) S_{t-s+p}$$

$$\hat{Y}_{24+1} = (L_{24} + 1\mathrm{T}_{24}) S_{24-4+1}$$

$$\hat{Y}_{25} = (492.469 + 7.352)\, 1.5569 = 778.17$$







Lag	ACF	T	LBQ
1	0,297749	1,46	2,41
2	-0,031543	-0,14	2,43
3	0,017610	0,08	2,44
4	-0,107084	-0,48	2,80

5	-0,257066	-1,15	4,97
6	0,033280	0,14	5,01

Τα κατάλοιπα δεν αυτοσυσχετίζονται και είναι τυχαία. Η τιμή 5.01 είναι μικρότερη από την τιμή 12.59.

4.7 CASE 4-4 MURPHY BROTHERS FURNITURE

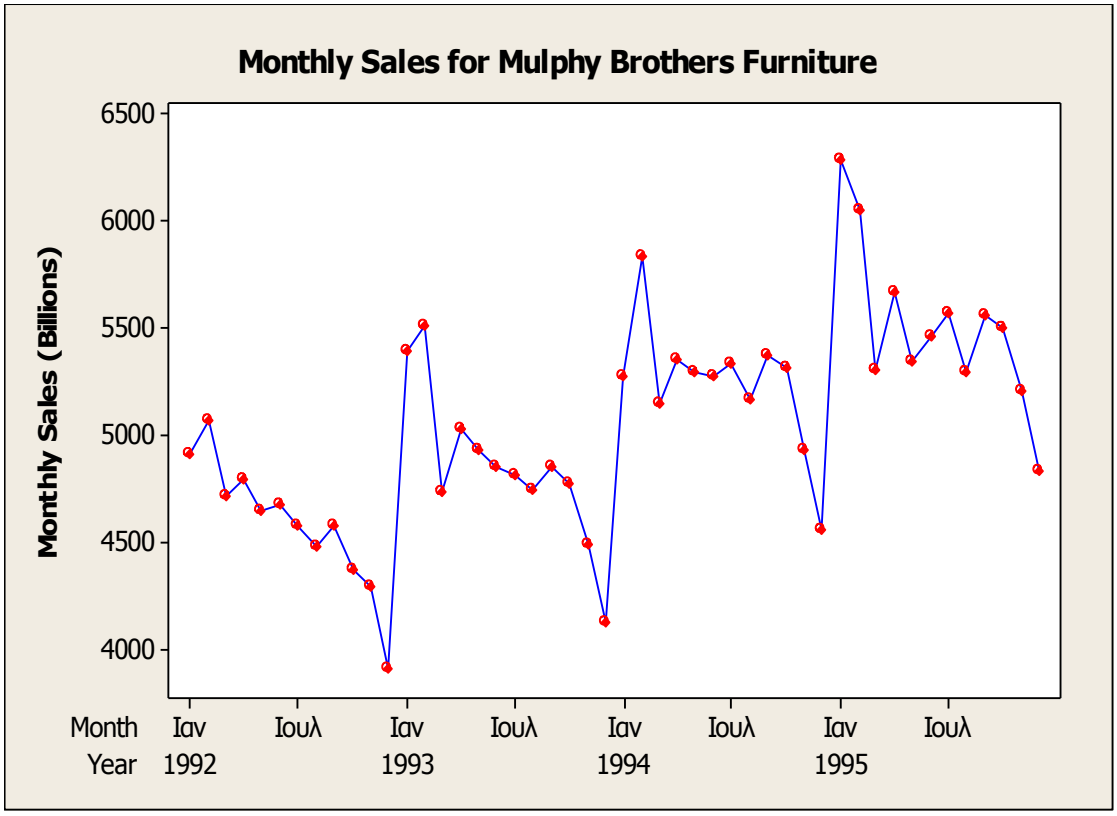
Η Julie γνωρίζει ότι οι περισσότερο σημαντικές λειτουργικές αποφάσεις που πρέπει να παρθούν, σε κάποιο βαθμό, εξαρτώνται από τις προβλέψεις για τις μελλοντικές πωλήσεις. Και αυτό γιατί οι προβλέψεις των πωλήσεων, θα έχουν επιπτώσεις στην προσθήκη νέων γραμμών παραγωγής ή στη διακοπή λειτουργίας των παλαιότερων γραμμών. Επιπλέον, από τις προβλέψεις πωλήσεων εξαρτάται ο σχεδιασμός των αγορών της εταιρείας, καθορίζονται οι ποσοστιαίες προβλέψεις, γίνεται αναδιάταξη του προσωπικού, ενώ βάσει των προβλέψεων παίρνονται χρηματοοικονομικές αποφάσεις και αποφάσεις που αφορούν το μάρκετινγκ (διαφήμιση). Συγκεκριμένα, η Julie ήταν ενήμερη ότι υπάρχουν πολύ σημαντικές τρέχουσες ανάγκες για καλές προβλέψεις. Γνώριζε επίσης ότι το τμήμα παραγωγής έπρεπε να αποφασίσει για τις πρώτες ύλες που θα προμηθευτεί και τους υπαλλήλους που θα χρειαστεί για τους επόμενους μήνες. Συνεπώς, για να προκύψουν οι επενδυτικές ευκαιρίες θα πρέπει να γίνει σωστή πρόβλεψη ζήτησης της νέας παραγωγικής γραμμής επίπλων.

Στο παράδειγμα 3-1Α είδαμε ότι η Julie χρησιμοποίησε τις λιανικές πωλήσεις όλων των καταστημάτων (1983-1995, Πίνακας 3-8) προκειμένου να αναπτύξει ένα «pattern» για τις πωλήσεις της οικογενειακής της επιχείρησης. Στο παράδειγμα 3-1Β είδαμε πως ο πατέρας της κατάφερε να βρει πραγματικά δεδομένα της επιχείρησης

τους μόνο για τα τέσσερα τελευταία χρόνια (1992-1995). Ως εκ τούτου, η Julie είχε το δίλημμα ποιο από τα δύο πακέτα δεδομένων να χρησιμοποιήσει προκειμένου να προβεί σε πρόβλεψη για το έτος 1996. Ωστόσο, ήταν εμφανές ότι τα δύο «πακέτα» δεδομένων ήταν αρκετά όμοια μεταξύ τους.

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
1992	4906	5068	4710	4792	4638	4670	4574	4477	4571	4370	4293	3911
1993	5389	5507	4727	5030	4926	4847	4814	4744	4844	4769	4483	4120
1994	5270	5835	5147	5354	5290	5271	5328	5164	5372	5313	4924	4552
1995	6283	6051	5298	5659	5343	5461	5568	5287	5555	5501	5201	4826

Πίνακας 3-9 Μηνιαίες πωλήσεις επίπλων (των αδερφών Murphy) για τα έτη 1992 – 1995



ΕΡΩΤΗΣΕΙΣ:

1. **Μπορεί κάποιο από τα μοντέλα πρόβλεψης του 4ου κεφαλαίου να λειτουργήσει καλά για τα δεδομένα των εθνικών πωλήσεων;**

Εφόσον η Julie έχει αφαιρέσει την τάση από τα δεδομένα της χρονοσειράς μας με τη μέθοδο των διαφορών, μια πρώτη μέθοδος πρόβλεψης που θα μπορούσε να έχει αξιόπιστα αποτελέσματα είναι η μέθοδος του απλού μέσου όρου. Η μέθοδος πρόβλεψης του απλού μέσου λειτουργεί με καλά αποτελέσματα, όταν οι συνθήκες που παράγουν τις χρονοσειρές έχουν σταθεροποιηθεί και το περιβάλλον παραμένει αμετάβλητο. Παράλληλα, η μέθοδος αυτή χρησιμοποιεί όλα τα παρελθόντα δεδομένα, ώστε να γίνει πρόβλεψη για την επόμενη περίοδο (έτος 1996). Τα δεδομένα που έχει στην κατοχή της η Julie αφορούν 13 χρόνια επομένως με τη μέθοδο αυτή θα είναι εφικτή μια ικανοποιητική πρόβλεψη βασιζόμενη στο μέσο όρο του δείγματος. Επιπλέον, θα μπορούσε να χρησιμοποιηθεί και η απλή μέθοδος εκθετικής εξομάλυνσης, η οποία παίρνει υπόψη της όλες τις τιμές με διαφορετικό όμως βάρος, άρα απαιτεί ένα αρκετά μεγάλο εύρος παρατηρήσεων. Δηλαδή, οι πιο πρόσφατες παρατηρήσεις στη χρονοσειρά λαμβάνουν μεγαλύτερη στάθμιση στον υπολογισμό της πρόβλεψης. Στην περίπτωση που η Julie δεν χρησιμοποιούσε την μέθοδο των διαφορών για να αφαιρέσει την τάση, τότε μια καλή μέθοδος πρόβλεψης θα ήταν η απλοϊκή μέθοδος naïve και αυτή του διπλού κινητού μέσου. Η μέθοδος αυτή χρησιμοποιείται όταν έχουμε τάση στα δεδομένα μας, ενώ ταυτόχρονα απαιτεί ένα μεγάλο αριθμό αρχικών παρατηρήσεων. Παράλληλα, και η μέθοδος Holt χρησιμοποιείται, όταν υπάρχει τάση στις παρατηρήσεις της χρονοσειράς.

2. **Μπορεί κάποιο από τα μοντέλα πρόβλεψης που μελετήθηκαν στο 4ο κεφάλαιο να χρησιμοποιηθεί για τα πραγματικά δεδομένα;**

   Όπως αναφέραμε και παραπάνω, στον απλό μέσο χρησιμοποιούμε

όλα τα παρελθόντα δεδομένα, άρα χρειάζεται να έχουμε αρκετά στοιχεία. Στην περίπτωση που θέλουμε να συγκεντρωθούμε περισσότερο στα πρόσφατα γεγονότα, τότε χρησιμοποιούμε τον κινητό μέσο. Συμπεραίνουμε λοιπόν, πως στην περίπτωση χρήσης των πραγματικών, πιο πρόσφατων και περιορισμένων στοιχείων της εταιρείας των Murphy, το μοντέλο πρόβλεψης του κινητού μέσου θα είναι το πλέον κατάλληλο. Με τη μέθοδο αυτή, υπολογίζεται ένας νέος μέσος όρος στον οποίο παραλείπεται η πιο παλιά παρατήρηση προκειμένου να συμπεριληφθεί η πιο πρόσφατη. Απ' την άλλη πλευρά, ακριβώς επειδή ο αριθμός των πραγματικών παρατηρήσεων δεν είναι μεγάλος, οι μέθοδοι του κινητού και του διπλού κινητού μέσου όρου δεν θα μπορούσαν να χρησιμοποιηθούν. Πραγματοποιώντας μια διαγραμματική απεικόνιση των πραγματικών δεδομένων, διαπιστώνουμε πως πρόκειται για μια μη στάσιμη χρονοσειρά, καθώς οι τιμές αυξάνονται χωρίς να ταλαντεύονται γύρω από ένα σταθερό επίπεδο. Επομένως, η χρονοσειρά μας έχει τάση, άρα η μέθοδος Holt θα μπορούσε να έχει αξιόπιστα αποτελέσματα. Η μέθοδος Holt είναι πιο περίπλοκη από την απλή εκθετική εξομάλυνση καθώς έχει δύο παραμέτρους, ωστόσο δίνει καλύτερες προβλέψεις για δεδομένα με τάση. Ωστόσο, αν παρατηρήσουμε το διάγραμμα των μηνιαίων πωλήσεων, εκτός από τάση στα δεδομένα, έχουμε και εποχικότητα (φανερώνεται από το «peak» στο διάγραμμα). Η διαδικασία του Winters μπορεί να εφαρμοστεί για προβλέψεις με βάση μια χρονοσειρά που εμφανίζει και τάση και εποχικότητα. Ίσως μάλιστα αυτή να είναι η πλέον κατάλληλη μέθοδος για πρόβλεψη.

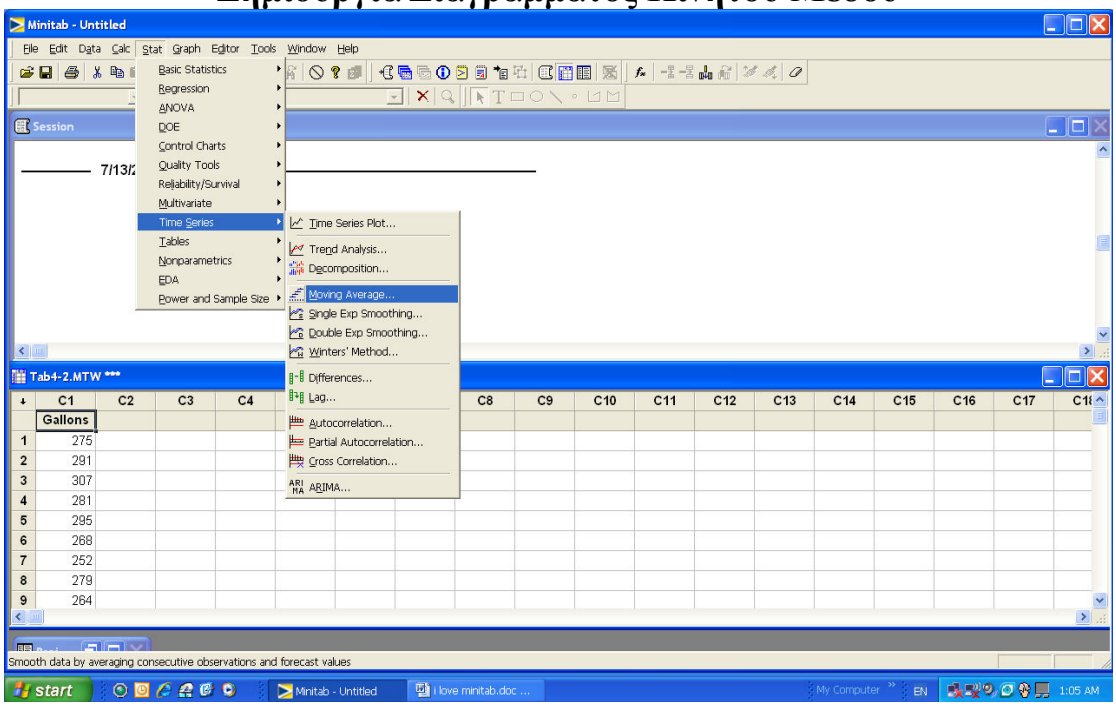
3. **Ποιο «πακέτο» δεδομένων και ποια μέθοδο πρόβλεψης θα προτείνατε στην Julie για πρόβλεψη των πωλήσεων του 1996;**

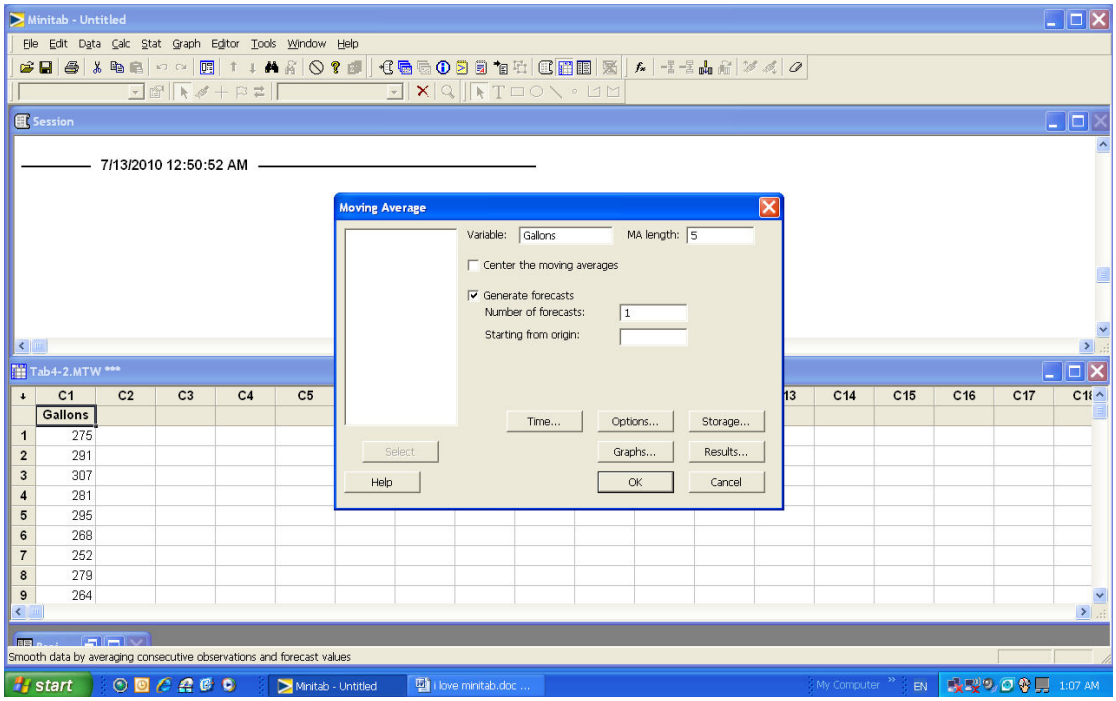
   Από τη διαγραμματική απεικόνιση των δεδομένων, όπως

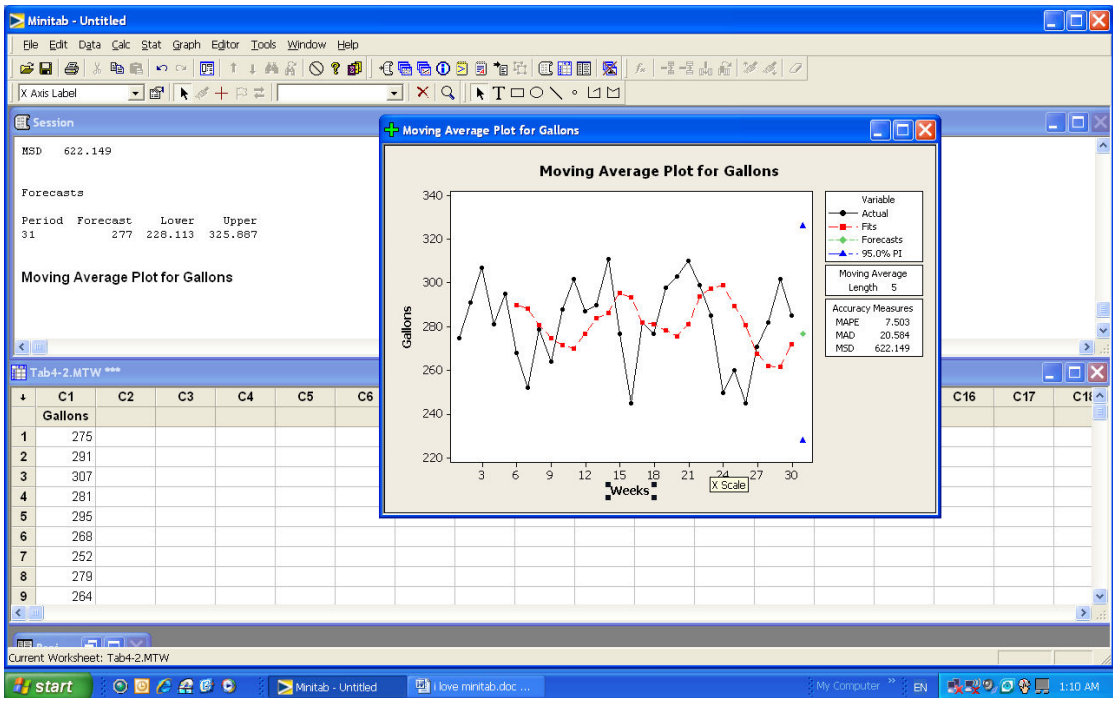
αναφέραμε και παραπάνω, διαπιστώθηκε η ύπαρξη τάσης και εποχικότητας στη χρονοσειρά. Στηριζόμενοι στο γεγονός αυτό, θα προτείναμε να επιλεγεί η μέθοδος πρόβλεψης Winters στο πακέτο δεδομένων που αφορά την επιχείρηση της οικογένειας Murphy. Λαμβάνοντας βέβαια υπόψη το γεγονός ότι αυτή η μέθοδος είναι κάπως περίπλοκη στη χρήση της και απαιτεί κάποιο υπολογιστικό σύστημα προκειμένου να μας δώσει αποτελέσματα, η πιο απλοϊκή λύση θα ήταν η χρήση της naïve μεθόδου με εποχικά δεδομένα.

4.8 ΧΡΗΣΗ MINITAB

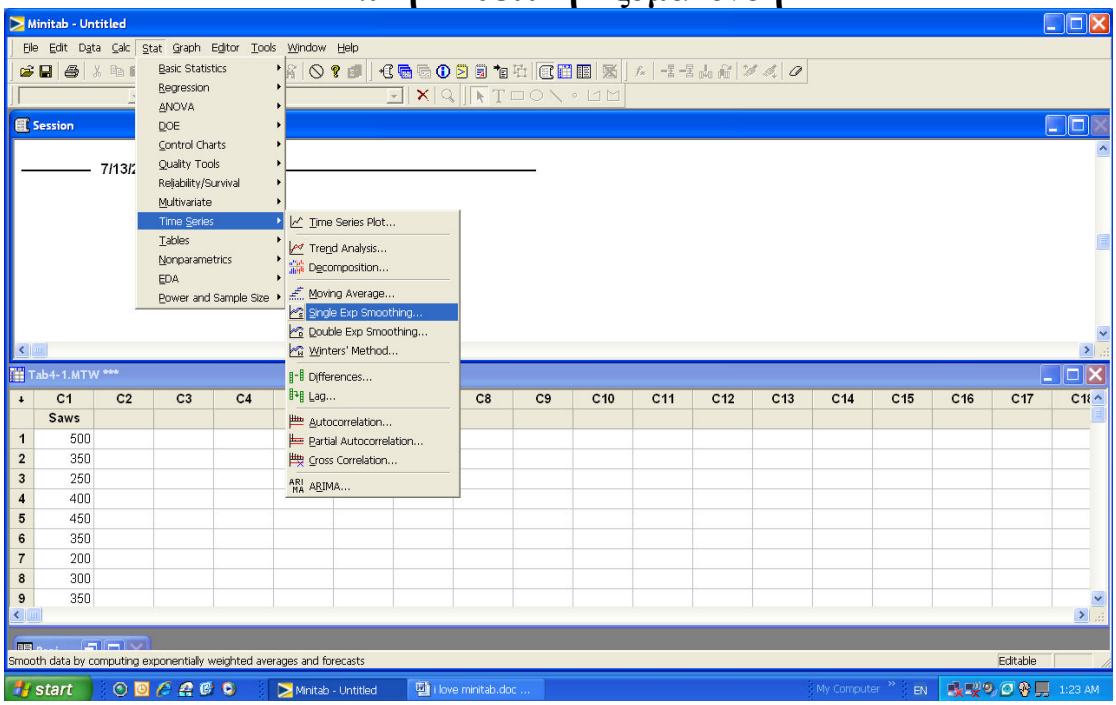
Δημιουργία Διαγράμματος Κινητού Μέσου

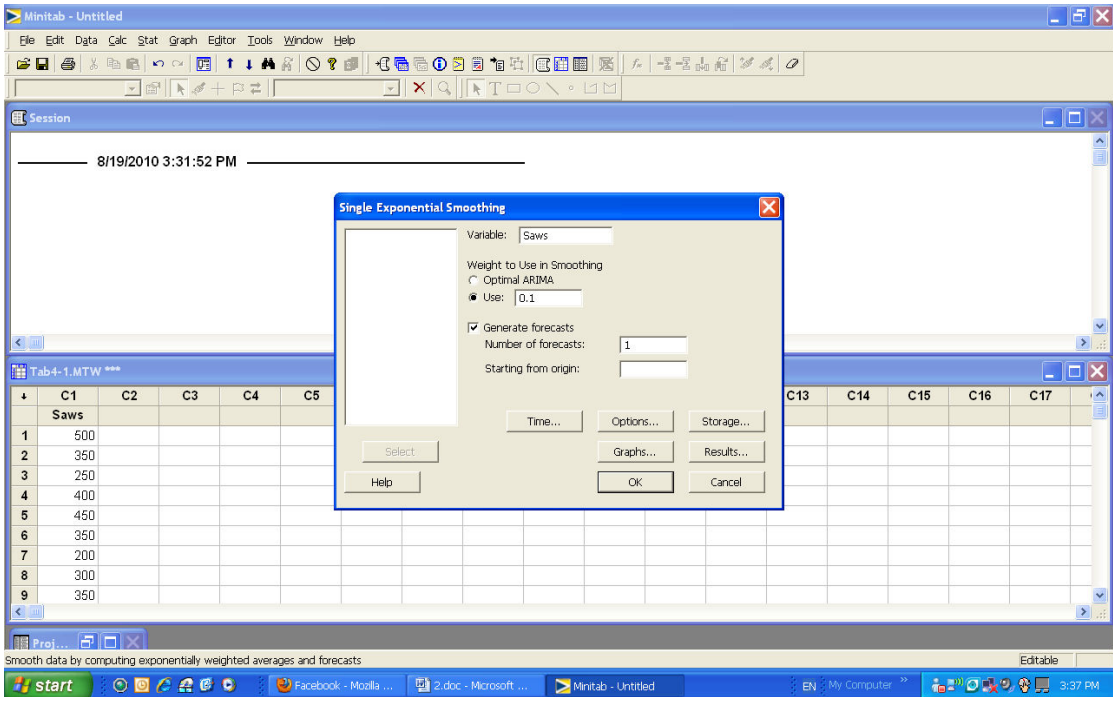


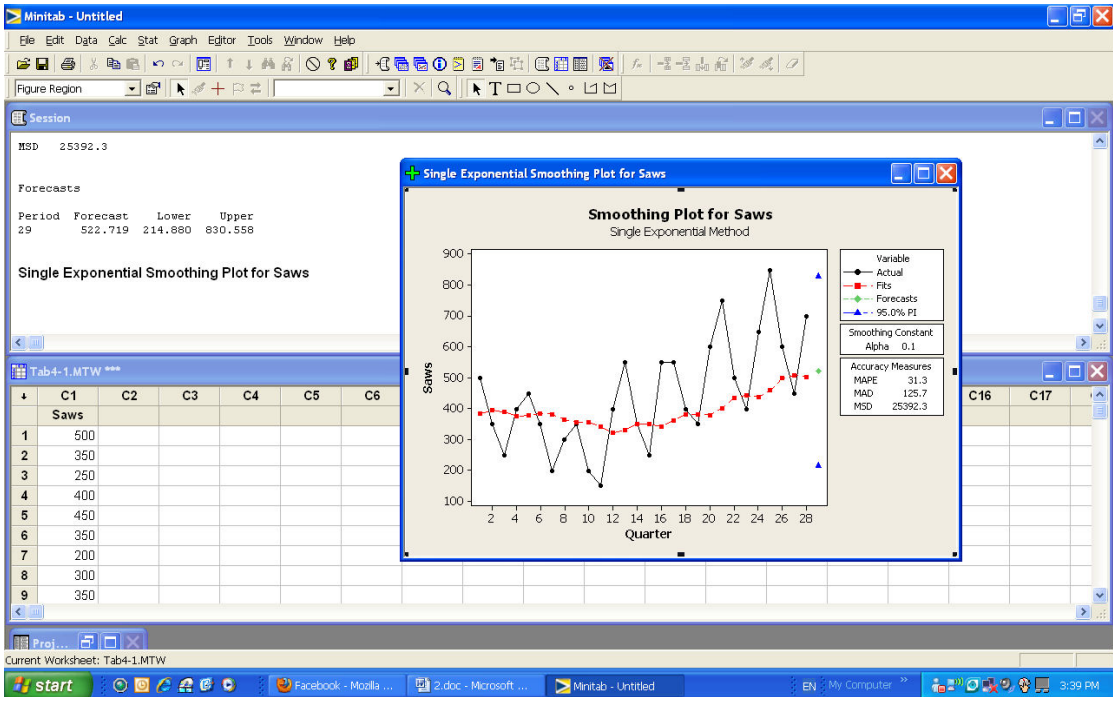

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/13/2010 12:50:52 AM
Moving Average
Variable: Gallons
MA length: 5
Center the moving averages
Generate forecasts
Number of forecasts: 1
Starting from origin:
Time...
Options...
Storage...
Select
Graphs...
Results...
Help
OK
Cancel
Tab4-2.MTW ***
C1 C2 C3 C4 C5
Gallons
1 275
2 291
3 307
4 281
5 295
6 268
7 252
8 279
9 264
Smooth data by averaging consecutive observations and forecast values
start
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
1:07 AM


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
X Axis Label
Session
MSD 622.149
Forecasts
Period Forecast Lower Upper
31 277 228.113 325.887
Moving Average Plot for Gallons
Moving Average Plot for Gallons
Gallons
Weeks
X Scale
Variable
Actual
Fits
Forecasts
95.0% PI
Moving Average
Length 5
Accuracy Measures
MAPE 7.503
MAD 20.584
MSD 622.149
Tab4-2.MTW ***
C1 C2 C3 C4 C5 C6 C16 C17
Gallons
1 275
2 291
3 307
4 281
5 295
6 268
7 252
8 279
9 264
Current Worksheet: Tab4-2.MTW
start
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
1:10 AM

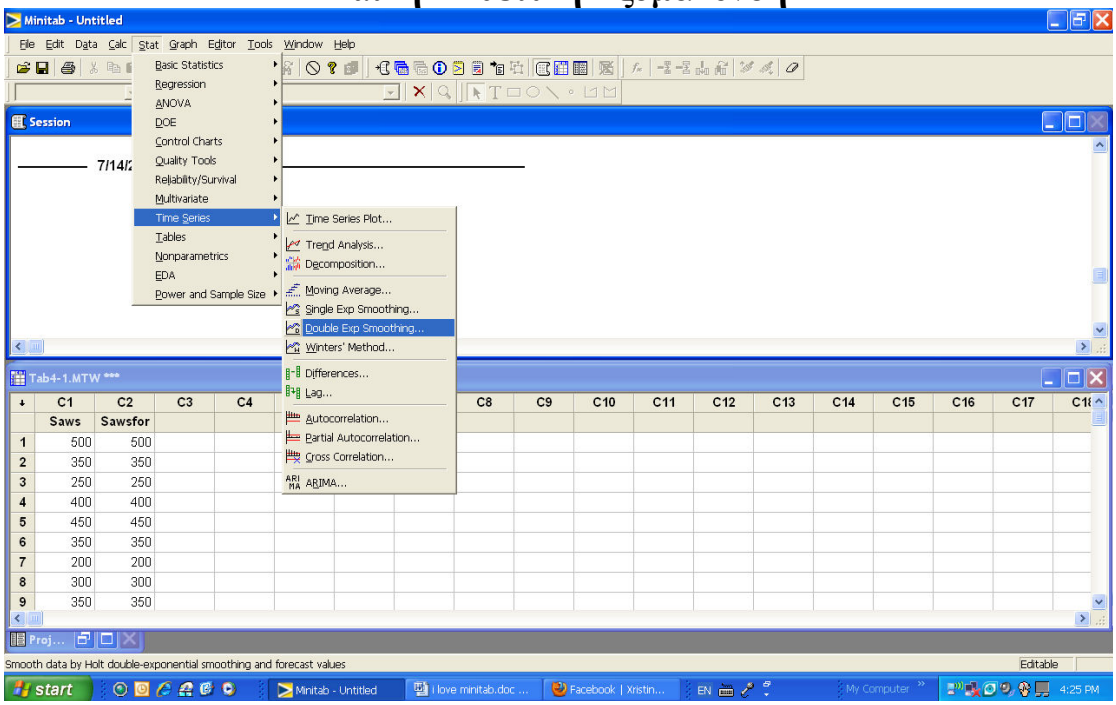
Απλή Εκθετική Εξομάλυνση

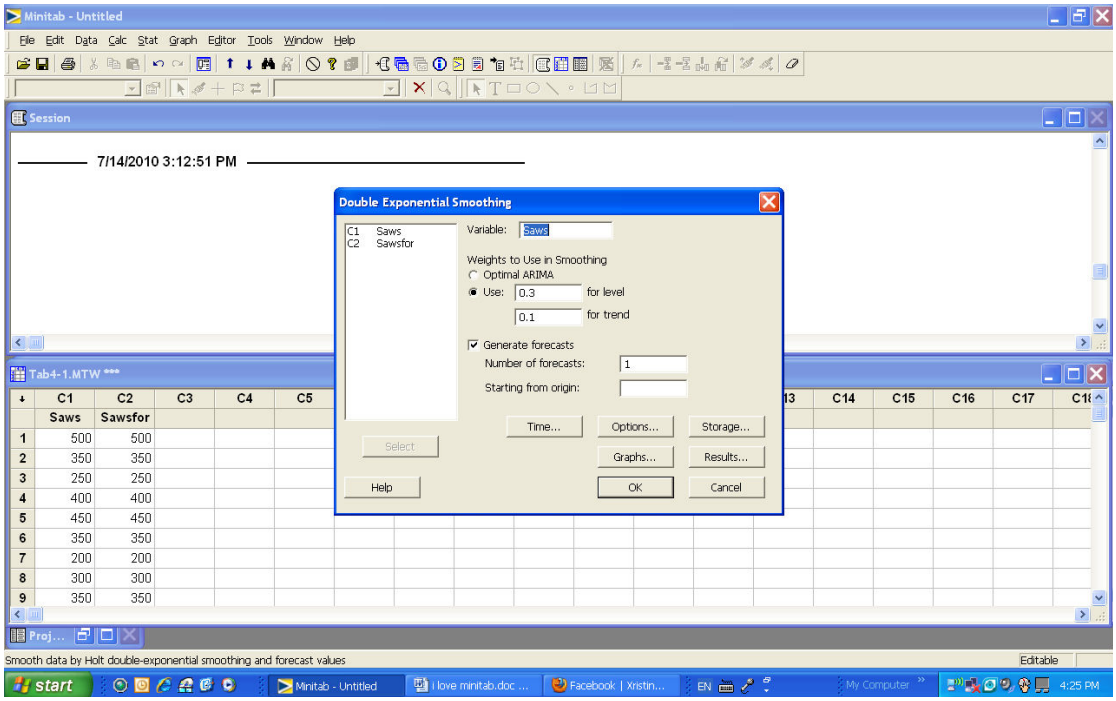


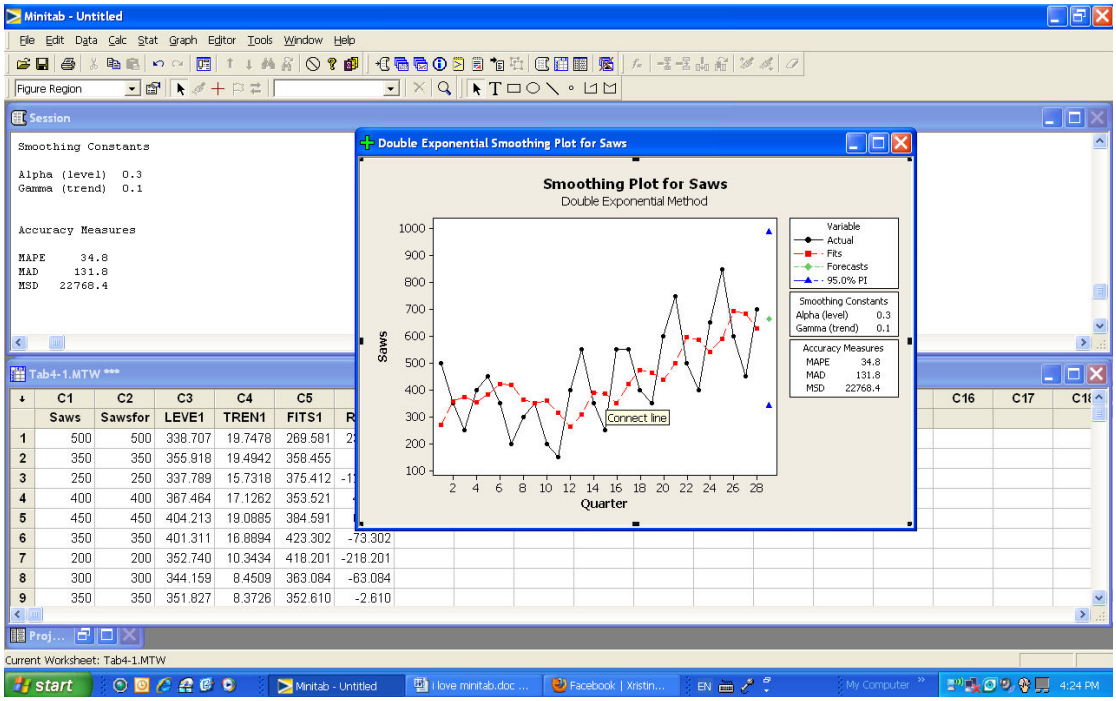




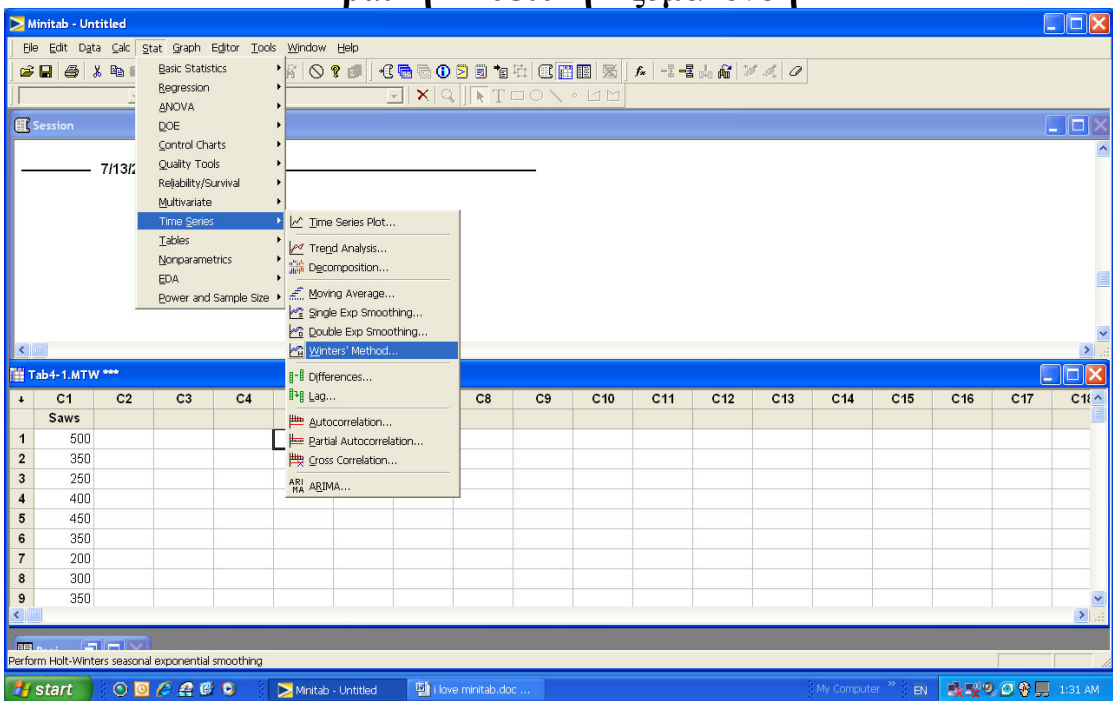
Διπλή Εκθετική Εξομάλυνση

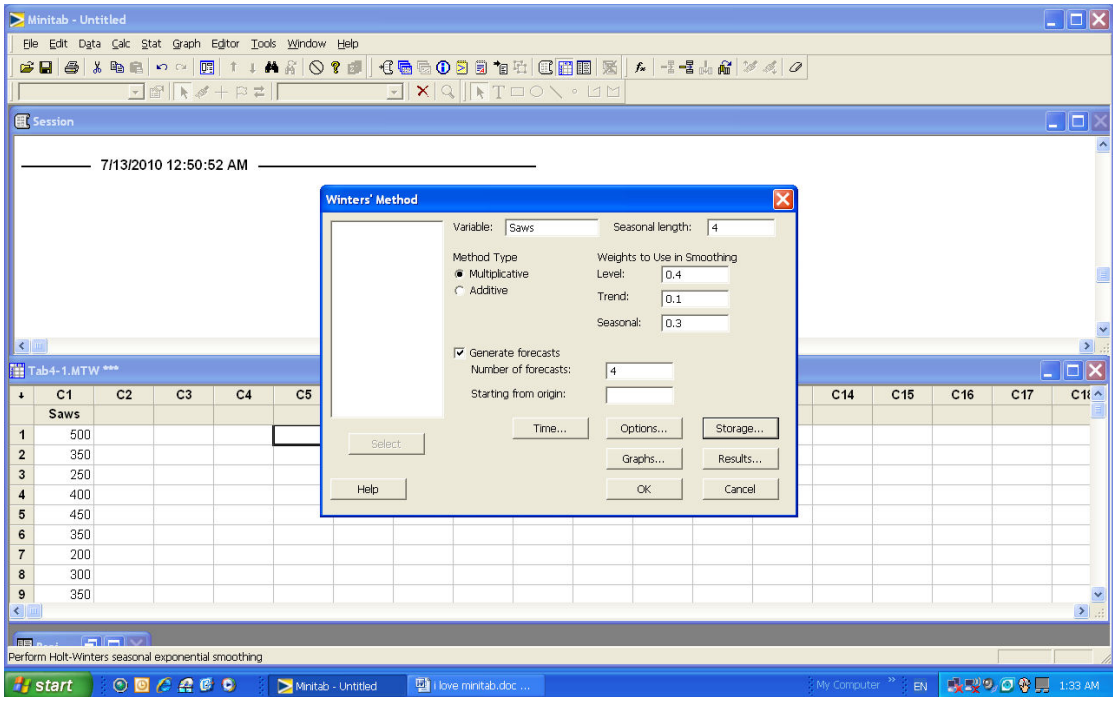


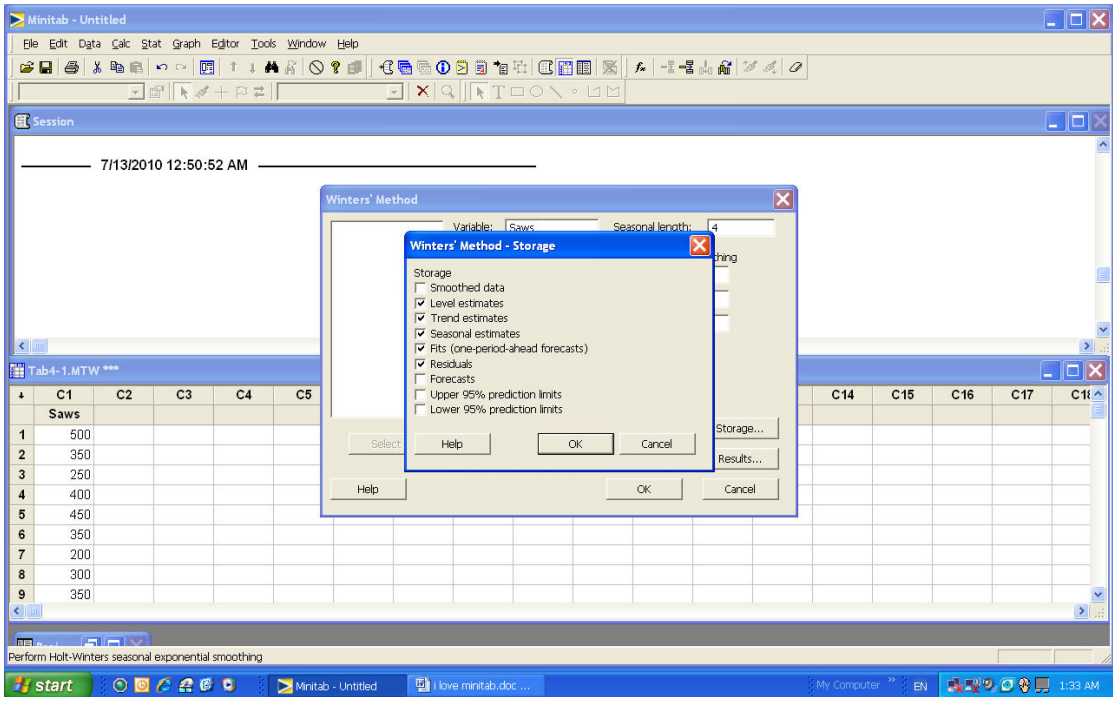

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/14/2010 3:12:51 PM
Double Exponential Smoothing
C1 Saws
C2 Sawsfor
Variable: Saws
Weights to Use in Smoothing
Optimal ARIMA
Use: 0.3 for level
0.1 for trend
Generate forecasts
Number of forecasts: 1
Starting from origin:
Time... Options... Storage...
Graphs... Results...
Select
Help
OK
Cancel
Tab4-1.MTW ***
C1 C2 C3 C4 C5 C16 C17
Saws Sawsfor
1 500 500
2 350 350
3 250 250
4 400 400
5 450 450
6 350 350
7 200 200
8 300 300
9 350 350
Smooth data by Holt double-exponential smoothing and forecast values
Editable
start
Minitab - Untitled
I love minitab.doc ...
Facebook | Xristin...
My Computer
4:25 PM

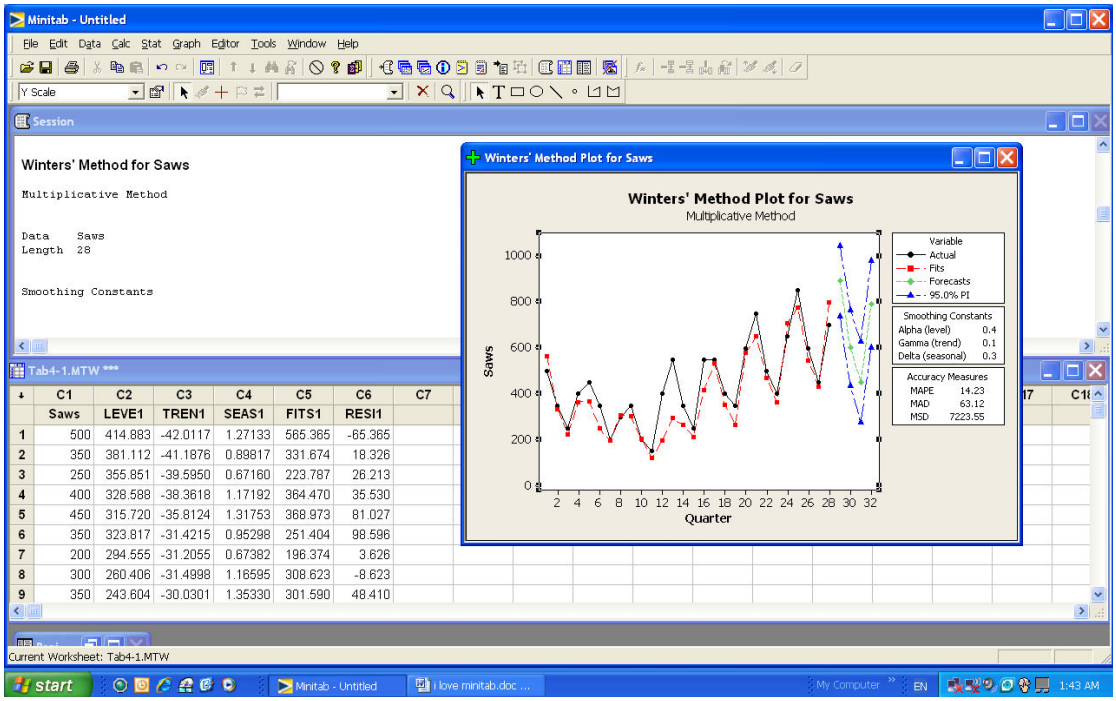

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Figure Region
Session
Smoothing Constants
Alpha (level) 0.3
Gamma (trend) 0.1
Accuracy Measures
MAPE 34.8
MAD 131.8
MSD 22768.4
Double Exponential Smoothing Plot for Saws
Smoothing Plot for Saws
Double Exponential Method
Saws
Quarter
Connect line
Variable
Actual
Fits
Forecasts
95.0% PI
Smoothing Constants
Alpha (level) 0.3
Gamma (trend) 0.1
Accuracy Measures
MAPE 34.8
MAD 131.8
MSD 22768.4
Tab4-1.MTW ***
C1 C2 C3 C4 C5 C16 C17
Saws Sawsfor LEVE1 TREN1 FITS1
1 500 500 338.707 19.7478 269.581
2 350 350 355.918 19.4942 358.455
3 250 250 337.789 15.7318 375.412
4 400 400 367.464 17.1262 353.521
5 450 450 404.213 19.0885 384.591
6 350 350 401.311 16.8894 423.302 -73.302
7 200 200 352.740 10.3434 418.201 -218.201
8 300 300 344.159 8.4509 363.084 -63.084
9 350 350 351.827 8.3726 352.610 -2.610
Current Worksheet: Tab4-1.MTW
start
Minitab - Untitled
I love minitab.doc ...
Facebook | Xristin...
My Computer
4:24 PM

Τριπλή Εκθετική Εξομάλυνση






Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/13/2010 12:50:52 AM
Winters' Method
Variable: Saws
Seasonal length: 4
Winters' Method - Storage
Storage
Smoothed data
Level estimates
Trend estimates
Seasonal estimates
Fits (one-period-ahead forecasts)
Residuals
Forecasts
Upper 95% prediction limits
Lower 95% prediction limits
Storage...
Results...
Help
OK
Cancel
Tab4-1.MTW ***
C1 C2 C3 C4 C5 C14 C15 C16 C17 C18
Saws
1 500
2 350
3 250
4 400
5 450
6 350
7 200
8 300
9 350
Perform Holt-Winters seasonal exponential smoothing
start
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
1:33 AM


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Y Scale
Session
Winters' Method for Saws
Multiplicative Method
Data Saws
Length 28
Smoothing Constants
Winters' Method Plot for Saws
Winters' Method Plot for Saws
Multiplicative Method
Saws
Quarter
Variable
Actual
Fits
Forecasts
95.0% PI
Smoothing Constants
Alpha (level) 0.4
Gamma (trend) 0.1
Delta (seasonal) 0.3
Accuracy Measures
MAPE 14.23
MAD 63.12
MSD 7223.55
Tab4-1.MTW ***
C1 C2 C3 C4 C5 C6 C7
Saws LEVE1 TREN1 SEAS1 FITS1 RESI1
1 500 414.883 -42.0117 1.27133 565.365 -65.365
2 350 381.112 -41.1876 0.89817 331.674 18.326
3 250 355.851 -39.5950 0.67160 223.787 26.213
4 400 328.588 -38.3618 1.17192 364.470 35.530
5 450 315.720 -35.8124 1.31753 368.973 81.027
6 350 323.817 -31.4215 0.95298 251.404 98.596
7 200 294.555 -31.2055 0.67382 196.374 3.626
8 300 260.406 -31.4998 1.16595 308.623 -8.623
9 350 243.604 -30.0301 1.35330 301.590 48.410
Current Worksheet: Tab4-1.MTW
start
Minitab - Untitled
I love minitab.doc ...
My Computer
EN
1:43 AM

ΚΕΦΑΛΑΙΟ 5 (ΧΡΟΝΟΣΕΙΡΕΣ ΚΑΙ ΤΑ ΣΥΣΤΑΤΙΚΑ ΤΟΥΣ)

5.1 ΔΙΑΣΠΑΣΗ ΧΡΟΝΟΣΕΙΡΩΝ

Στα προηγούμενα χρησιμοποιήσαμε μεθόδους παρεμβολής – πρόβλεψης προκειμένου να προβλέψουμε μελλοντικές τιμές μιας χρονοσειράς από παρατηρηθείσες τιμές. Η προσαρμογή γινόταν με τη προσέγγιση της καμπύλης μέσω ευθύγραμμων τμημάτων ενώ η χρονική περίοδος εκφραζόταν σε έτη, τρίμηνα ή μήνες. Οι χρονοσειρές μελετώνται κάτω από την υπόθεση, ότι με βάση τη συμπεριφορά στο παρελθόν μπορεί να προβλεφθεί η μελλοντική πορεία της διακύμανσης της παρατηρούμενης μεταβλητής. Μολονότι η πρόβλεψη μελλοντικών καταστάσεων δεν γίνεται χωρίς σφάλμα, η ανάλυση και ο διαχωρισμός (decomposition) των χρονοσειρών έχει αξία στη διαδικασία της πρόβλεψης, γιατί έτσι μπορεί να μειωθεί το σφάλμα πρόβλεψης.

Η διάσπαση των χρονοσειρών (time series decomposition) στηρίζεται στην υπόθεση, ότι οι τιμές μιας χρονοσειράς σχηματίζονται από τα στοιχεία, που τη συνθέτουν και τα οποία είναι η τάση, η κυκλικότητα, η εποχικότητα και η μη-κανονικότητα.

- Τάση είναι η κείμενη κατά μέσον όρο αύξηση, ή η μείωση των τιμών μιας χρονοσειράς. Η τάση είτε υπολογίζεται ως μεταβολές, όπως στην μέθοδο Holt και Winters και προσέγγιση μέσω ευθύγραμμων τμημάτων, είτε υπολογίζεται σαν μία μέση καμπύλη σε επίπεδο τιμών με τη μέθοδο των ελαχίστων τετραγώνων.
- Κυκλικότητα είναι κυματοειδείς διακυμάνσεις μιας χρονοσειράς με περίοδο μεγαλύτερη του έτους. Στη πράξη η κυκλικότητα μιας

χρονοσειράς είναι δύσκολο να βρεθεί, διότι η περίοδός της δεν είναι σταθερή. Πολλές φορές την αναλύουμε μαζί με την τάση ως τάση – κύκλος.

- Εποχικότητα είναι οι εποχικές διακυμάνσεις μιας χρονοσειράς με μέγιστη περίοδο το ένα έτος. Η εποχικότητα μπορεί εύκολα να απομονωθεί σε αντίθεση με τη κυκλικότητα, διότι παρουσιάζει σταθερή περίοδο, ενώ οι παράγοντες, που την επηρεάζουν, είναι σταθεροί και εξελίσσονται κατά τη διάρκεια του έτους. Η μέγιστη επίδραση στις τιμές μπορεί να είναι λίγο διαφορετική.
- Μη-κανονικές είναι οι διακυμάνσεις, οι οποίες επηρεάζονται από πολλούς απρόβλεπτους παράγοντες: Εκρήξεις ηφαιστείων, σεισμοί, ξαφνικές πολιτικές αλλαγές αποφάσεων κλπ.

Για τη δημιουργία των προβλέψεων με τη μέθοδο, που θα αναπτύξουμε, η χρονοσειρά διασπάται στα τέσσερα προαναφερόμενα συνθετικά συστατικά της και προσδιορίζεται η επιρροή, η φύση και η έκταση, που έχει καθένα από αυτά τα συστατικά στη διαμόρφωση των τιμών της μεταβλητής.

Επισημαίνεται ότι είναι αδύνατο να προβλεφθεί το μέλλον με απόλυτη ακρίβεια και ότι πάντοτε θα υπάρχει, ως ένα βαθμό, σφάλμα στην πρόβλεψη. Για το διαχωρισμό των χρονοσειρών δεχόμαστε, ότι τα δεδομένα εμφανίζουν την εξής εικόνα:

Δεδομένα = πρότυπο (υπόδειγμα) + σφάλμα =

f(συστηματικές συνιστώσες) + σφάλμα

Το σφάλμα αυτό, ή υπολειπόμενη (residual) απόκλιση, υποθέτουμε ότι μπορεί να εκφραστεί ως η διαφορά ανάμεσα στο συνδυασμό των συστηματικών συνιστωσών της χρονοσειράς, για να μας δώσει την προσαρμογή και των παρατηρηθεισών τιμών της. Οι τεχνικές

διαχωρισμού, που χρησιμοποιούνται για τη μελέτη αυτών των συνιστωσών, δεν αποτελούν ενιαίο σώμα της στατιστικής θεωρίας, όπως στις μεθόδους εξομάλυνσης. Επίσης, δεν πληρούνται συνήθως, όλες οι προϋποθέσεις για επιστημονική στατιστική διερεύνηση, όπως, για παράδειγμα, η ύπαρξη τυχαίας δειγματοληψίας. Επομένως, οι διαδικασίες διάσπασης των χρονοσειρών συχνά αξιολογούνται με κριτήρια, που επηρεάζονται από την κρίση, εκείνου που διενεργεί την πρόβλεψη – ανάλυση. Οι τέσσερις συνιστώσες αλληλεπιδρούν, για να εξηγήσουν τη διακύμανση στις τιμές των δεδομένων χρονικά. Οι επιδράσεις αυτές είναι υπεύθυνες για την συχνά ασταθή συμπεριφορά των χρονοσειρών.

5.2 ΜΟΝΤΕΛΑ ΔΙΑΣΠΑΣΗΣ ΧΡΟΝΟΣΕΙΡΩΝ

Στην ανάλυση χρονοσειρών (ιδιαίτερα οικονομικών), υποτίθεται ότι οι τρεις συστηματικές συνιστώσες και η τυχαία συνιστώσα είναι αποτέλεσμα διαφορετικών αιτίων. Διαχωρίζοντας τρεις συστηματικές συνιστώσες, ρίχνουμε κάποιο φως στη φύση των παραγόντων που προκαλούν τις χρονοσειρές. Η γνώση αυτή μας επιτρέπει να προβάλουμε την επίδραση των συστηματικών συνιστωσών στο μέλλον. Στη συνέχεια, κάνοντας εκτιμήσεις, ή υποθέσεις για τη μη – κανονική συνιστώσα, μπορούμε να προχωρήσουμε στην πρόβλεψη. Εάν το Y_t συμβολίζει την πραγματική παρατήρηση σε χρόνο t, εκφράζουμε το Y_t σαν σχέση της τάσης της σειράς, T_t σε χρόνο t, του εποχικού παράγοντα, S_t σε χρόνο t, της κυκλικής συνιστώσας, C_t και του μη – κανονικού παράγοντα με I_t.

Στη συνέχεια υποθέτουμε, ότι η σχέση παίρνει την παρακάτω μορφή:

$$Y_t = T_t * S_t * C_t * I_t$$

Η ισότητα αυτή υποδηλώνει, ότι η τιμή της μεταβλητής σε χρόνο t

είναι το γινόμενο των τεσσάρων συνιστωσών, που αναφέρθηκαν. Ακόμη, η εξίσωση δείχνει, ότι τα T, S, C και I σε επίπεδο τιμών και δεικτών συσχετίζονται, με την έννοια ότι επειδή τα S, C και I είναι δείκτες αυξάνουν την τάση σε κάθε σημείο κατά κάποιο ποσοστό. Εναλλακτικά, η σχέση μπορεί να υποτεθεί, ότι είναι προσθετική ως εξής:

$$Y_t = T_t + S_t + C_t + I_t,$$

όπου,

- Y_t = παρατήρηση κατά τη χρονική περίοδο t
- T_t = συνιστώσα τάσης
- S_t = συνιστώσα εποχικότητας
- C_t = συνιστώσα κύκλου
- I_t = συνιστώσα τυχαιότητας

Σε αυτή την εξίσωση οι συνιστώσες δεν συσχετίζονται. Το μέγεθος της μιας συνιστώσας είναι ασυσχέτιστο με το μέγεθος οποιασδήποτε άλλης. Από τα δυο μοντέλα διαχωρισμού, το προσθετικό είναι πιο δύσκολο για παραπέρα ανάλυση. Επίσης, το προσθετικό μοντέλο υποθέτει, ότι οι συντελεστές μεταξύ τους είναι ανεξάρτητοι. Για παράδειγμα, αν χρησιμοποιήσουμε το προσθετικό μοντέλο, δεχόμαστε, ότι η τάση δεν επηρεάζει τον εποχικό συντελεστή, όσο υψηλός ή χαμηλός κι αν γίνεται και λειτουργεί πολύ καλά, όταν οι διακυμάνσεις είναι σταθερές. Κατά κανόνα, η παραδοχή αυτή δεν είναι ισχυρή, εκτός από την περίπτωση της πολύ βραχυχρόνιας πρόβλεψης. Το πολλαπλασιαστικό μοντέλο λειτουργεί καλά, όταν οι διακυμάνσεις εξαρτώνται από το ύψος των τιμών, κάτι που κατά κανόνα ισχύει. Επομένως, θα δώσουμε έμφαση στο πολλαπλασιαστικό μοντέλο, εφόσον στην πραγματικότητα, όλοι οι χρήστες της διαχωριστικής (decomposition) ανάλυσης το βρίσκουν πιο ευλογοφανές.

Η απαλοιφή της εποχικότητας και της τυχαιότητας οδηγεί στην καμπύλη τάσης – κύκλου, η οποία είναι ιδιαίτερα σημαντική για την πρόβλεψη επιχειρησιακών δεδομένων (business forecasting). Για το λόγο αυτό οι μέθοδοι αποσύνθεσης χρησιμοποιούνται σήμερα ευρέως από τις επιχειρήσεις. Για παράδειγμα η American Airlines χρησιμοποιεί την εποχικότητα στη χρονοσειρά του αριθμού των ταξιδιωτών, που την προτιμούν ανά πτήση, προκειμένου να προβλέψει τις χαμηλές τιμές – προσφορές των εισιτηρίων.

Ανάμεσα στις πιο δημοφιλείς μεθόδους αποσύνθεσης είναι η κλασσική μέθοδος αποσύνθεσης και η μέθοδος Census II (με τις διάφορες παραλλαγές της).

5.3 ΠΡΟΣΑΡΜΟΓΗ ΚΑΙ ΑΝΑΛΥΣΗ ΤΗΣ ΤΑΣΗΣ

Μελετούμε τις τάσεις για να προσδιορίσουμε τη μακρόχρονη κατεύθυνση της χρονοσειράς. Γενικά, αυτές οι εξελίξεις προέρχονται από μια ποικιλία αιτιών π.χ. στις οικονομικές χρονοσειρές τέτοιες αιτίες είναι οι μεταβολές της τεχνολογίας, των τιμών, του πληθυσμού και της παραγωγικότητας. Για παράδειγμα, το σύνολο των αποταμιευτικών καταθέσεων σε μια τράπεζα ταμιευτηρίου μπορεί να αυξάνεται από έτος σε έτος για ένα χρονικό διάστημα, επειδή ο αριθμός των καταθετών συνεχίζει να αυξάνεται. Παραπέρα, ο κάθε καταθέτης τείνει να έχει μεγαλύτερη αποταμίευση σε χρηματικές μονάδες, πράγμα που σχετίζεται με μια γενική αύξηση στα εισοδήματα των καταθετών. Εν μέρει, αυτό προέρχεται από μια γενική αύξηση των τιμών που ονομάζεται πληθωρισμός (τάση). Επίσης, το γενικό ανοδικό πρότυπο των εσόδων ενός νοσοκομείου σχετίζεται με τη γενική αύξηση (τάση) των τιμών, ή των τιμών των νοσοκομειακών υπηρεσιών.

Ένα πρώτο βήμα στην ανάλυση της συνιστώσας της τάσης είναι ο καθορισμός του λόγου, για τον οποίο γίνεται η ανάλυσή της. Μπορεί κάποιος να θέλει να προβάλει την τάση στο μέλλον, ή να την εξαλείψει από την αρχική χρονοσειρά, ώστε να μελετήσει τις διακυμάνσεις και να κατανοήσει τους παράγοντες, που την επηρεάζουν. Φυσικά, ίσως κάποιος να επιθυμεί να πράξει και τα δυο μαζί. Η γνώση του σκοπού βοηθάει τον αναλυτή στην επιλογή της μεθόδου και του εύρους της χρονοσειράς, που θα αναλύσει. Εάν ο σκοπός είναι η προβολή της τάσης στο μέλλον, είναι σημαντικό να γνωρίζει, το πόσο μακριά στο μέλλον πρόκειται να γίνει αυτή η προβολή. Μια προβολή για 15 χρόνια κατά κανόνα απαιτεί τη χρησιμοποίηση χρονοσειράς μακρύτερης διάρκειας, από ότι μια προβολή τριών μόνο χρόνων.

Το πρώτο βήμα για την ανάλυση της τάσης είναι η γραφική παράσταση της χρονοσειράς, σε αριθμητική, σε ημι-λογαριθμική κλίμακα, ή άλλους μετασχηματισμούς πριν την επιλογή της μεθόδου μέτρησης. Αν και υπάρχουν στατιστικά κριτήρια για τον καθορισμό της «σπουδαιότητας» της τάσης, μια εξέταση της γραφικής παράστασης, ως προς τον χρόνο, θα είναι συνήθως επαρκής.

5.4 ΜΟΝΤΕΛΑ ΤΑΣΗΣ

Τα μοντέλα τάσης βασίζονται σε δύο παράγοντες:

- Στην επιλογή της κατάλληλης καμπύλης.
- Η παραδοχή ότι η καμπύλη, που συνδέει με κάποια σχετική ακρίβεια τις παρελθούσες τιμές, είναι ενδεικτική και για το μέλλον.

Υπάρχουν κάποια αντικειμενικά κριτήρια, μέσω των οποίων

μπορεί να επιλεχθεί η κατάλληλη καμπύλη. Αυτά είναι το "Akaike Information Criterion (AIC)" και το "Bayesian Information Criterion (BIC)", τα οποία περικλείουν το MSE και τον αριθμό των παραμέτρων.

5.4.1 ΓΡΑΜΜΙΝΟ ΜΟΝΤΕΛΟ

Η κρίση (judgement) χρησιμοποιείται πολύ συχνά ως κριτήριο για τον προσδιορισμό του μοντέλου με την καλύτερη προσαρμογή. Μερικές φορές ο αναλυτής χαράζει απλώς μια γραμμή διαμέσου της σειράς των παρατηρήσεων, προκειμένου να αποκαλύψει το σχήμα και την κατεύθυνση της τάσης της χρονοσειράς. Ο αναλυτής, χαράζοντας τη γραμμή της τάσης, πρέπει να έχει την ικανότητα να αναγνωρίζει την επαναλαμβανόμενη εποχική διακύμανση και τους κύκλους, μέσω των οποίων διέρχεται η τάση. Με άλλα λόγια, θα πρέπει ο αναλυτής να έχει εξοικειωθεί σημαντικά με αυτές τις συγκεκριμένες σειρές, διαφορετικά η διαδικασία αυτή μπορεί να αποδειχθεί εξαιρετικά δύσκολη. Γι' αυτό οι διενεργούντες προβλέψεις συνήθως εμπιστεύονται την αντικειμενική μέθοδο των ελαχίστων τετραγώνων για να αποφύγουν την υποκειμενικότητα, που υπεισέρχεται στη μέθοδο που βασίζεται στην κρίση τους. Για να αρχίσουμε, θα ορίσουμε την εκτιμηθείσα γραμμική εξίσωση ως εξής:

$$\hat{Y}_t = a + b\,t \quad \text{με}$$

$$a = \frac{\sum y \sum t^2 - \sum t \sum ty}{n\sum t^2 - \left(\sum t\right)^2} \quad \text{και}$$

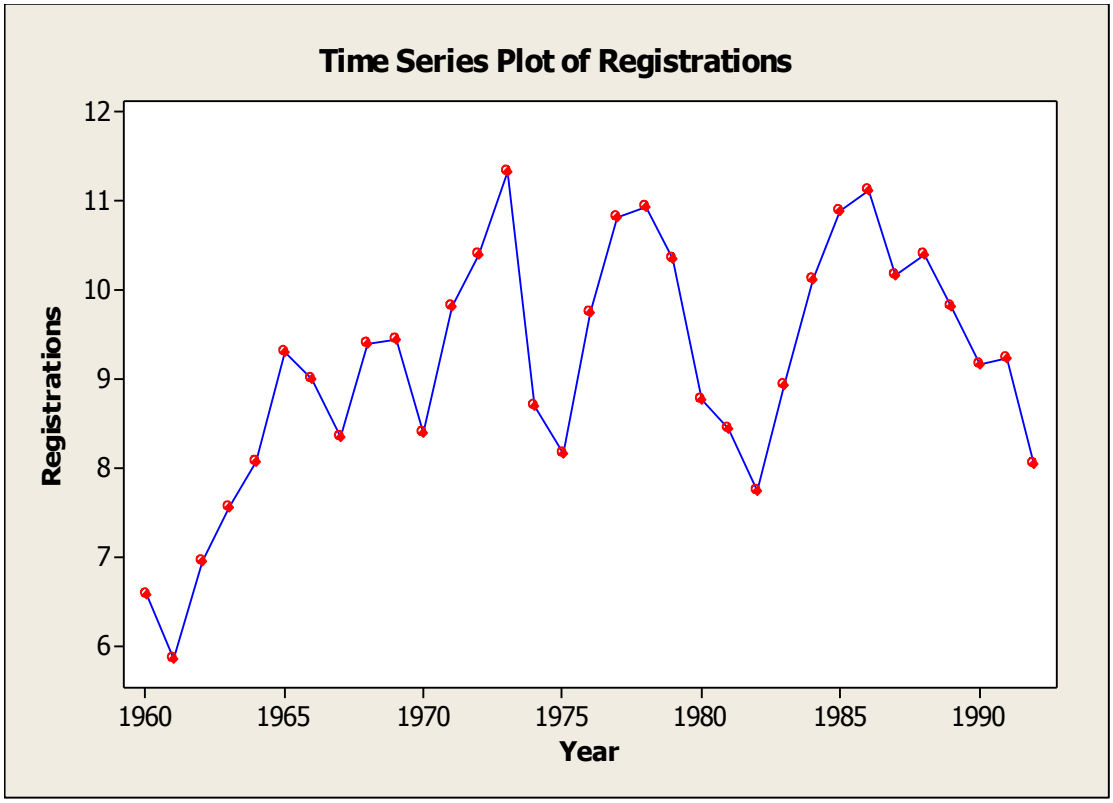
$$b = \frac{n\sum ty - \sum t \sum y}{n\sum t^2 - \left(\sum t\right)^2}$$

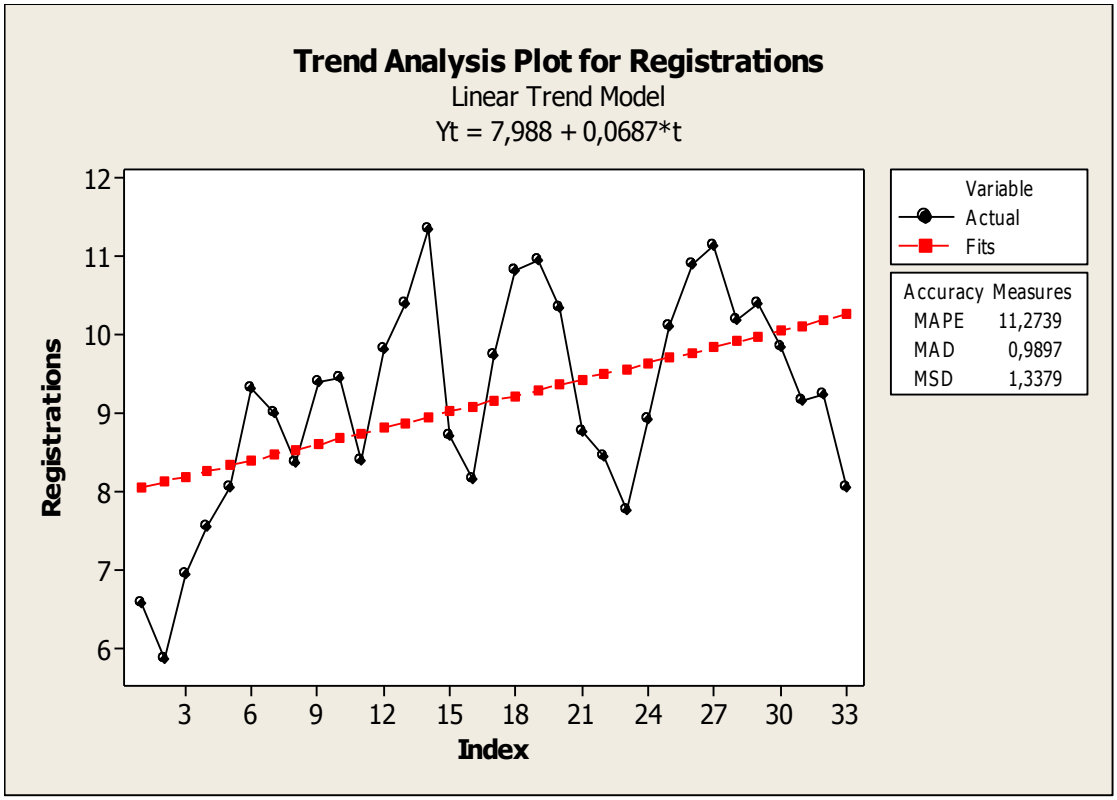
$\hat{Y}_t$ είναι η προβλεφθείσα τιμή τάσης για τη μεταβλητή Y για μια επιλεγείσα χρονική περίοδο t. α είναι η τιμή της τάσης, όταν t=0 (η

διαφορά ύψους). b είναι η κλίση της ευθείας (η μεταβολή στο Y που σχετίζεται με μια μοναδιαία μεταβολή στη t). Το t παριστάνει την παρατήρηση για την χρονική περίοδο.

Παράδειγμα (T5-1)

Έχουμε τα δεδομένα σχετικά με τις ετήσιες αγορές καινούριων αυτοκινήτων στις Ηνωμένες Πολιτείες για τα έτη 1960 – 1992 και στη συνέχεια αυτά τα δεδομένα απεικονίζονται γραφικά.





Η προσαρμοσμένη γραμμική τάση έχει την εξίσωση:

$$\overline{Y}_{\bar{t}} = 7.988 + 0.0687\ t.$$

Για παράδειγμα, η εξίσωση της τάσης εκτιμά, ότι οι αγορές καινούριων αυτοκινήτων το 1992 (t = 33) θα είναι:

$$\overline{Y}_{\bar{t}} = 7.988 + 0.0687\ (33) = 10.255$$

Παρόλα αυτά, οι αγορές καινούριων αυτοκινήτων ήταν 8.054 το 1992.

5.4.2 ΠΑΡΑΒΟΛΙΚΟ ΜΟΝΤΕΛΟ

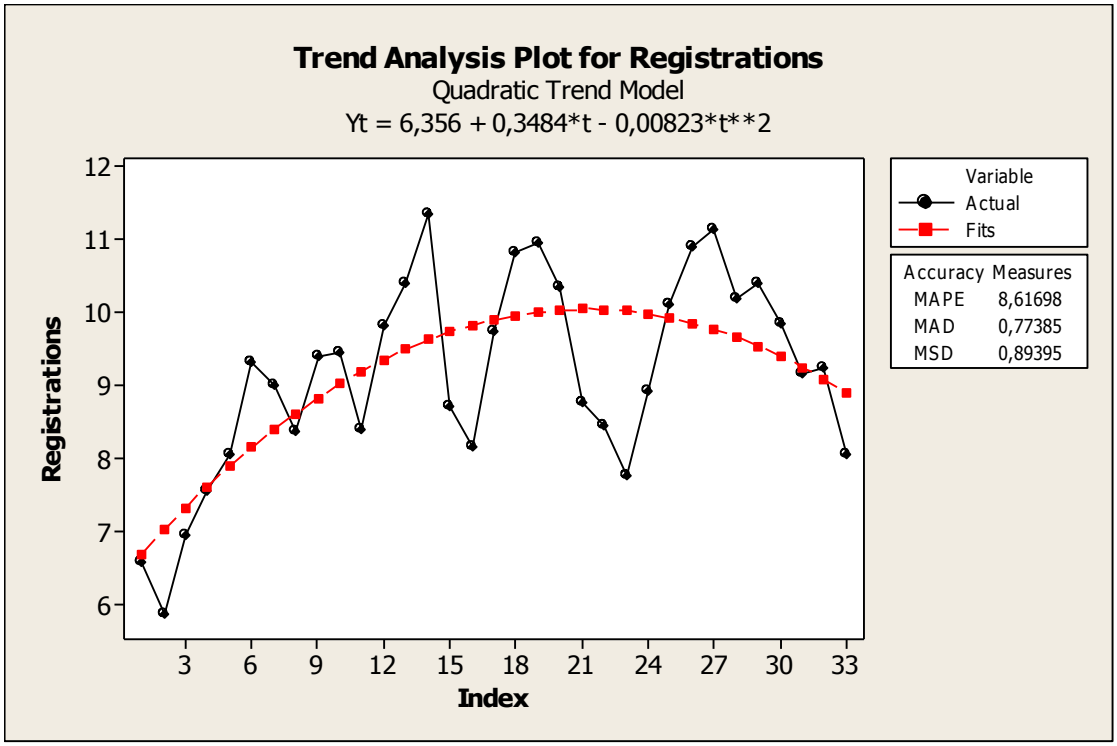
Μία επιπρόσθετη καμπύλη τάσης είναι η παραβολή ελαχίστων τετραγώνων, που ορίζεται από τις σχέσεις:

$$y = a + b\,x + cx^2, \qquad \sum y = na + b\sum x + c\sum x^2,$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \qquad \sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

οπότε οποιαδήποτε τιμή πρόβλεψης με ανεξάρτητη μεταβλητή το χρόνο

δίνεται από τη σχέση: $\hat{Y}_t = a + b\,t + c\,t^2$.



5.4.3 ΕΚΘΕΤΙΚΟ ΜΟΝΤΕΛΟ

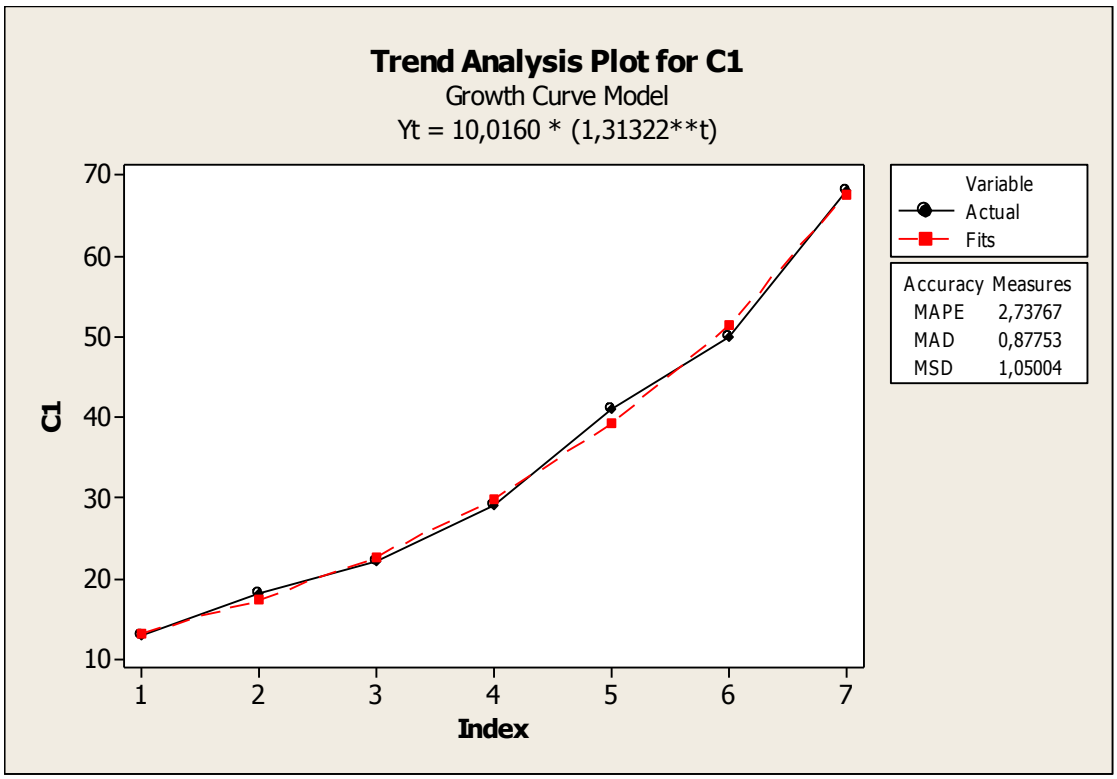
Οι Growth Curves αναπαριστούν τάσεις, σε πολλές βιομηχανίες και γραμμές παραγωγής προϊόντων, που αυξάνονται με μικρότερη ταχύτητα στην φάση ωρίμανσης. Μία από αυτές τις καμπύλες είναι η εκθετική καμπύλη τάσης, για σταθερή την ποσοστιαία αύξηση των τιμών. Η καμπύλη είναι της μορφής: $\hat{Y}_t = b_0 b_1^t$.

Παράδειγμα:

Αριθμός εργαζομένων πωλητών ανά έτος σε μία εταιρεία.

Year	Salespeople
2000	13
2001	18
2002	22
2003	29
2004	41

2005	50
2006	68



Εάν κάναμε γραμμική προσέγγιση στα αρχικά στάδια θα είχαμε υπερεκτίμηση και στα τελικά στάδια υποεκτίμηση των αυξήσεων.

5.4.4 ΜΟΝΤΕΛΟ GOMPERTZ

Μία άλλη καμπύλη – s-type curve – είναι και η καμπύλη Gompertz, που περιγράφεται με την ακόλουθη σχέση:

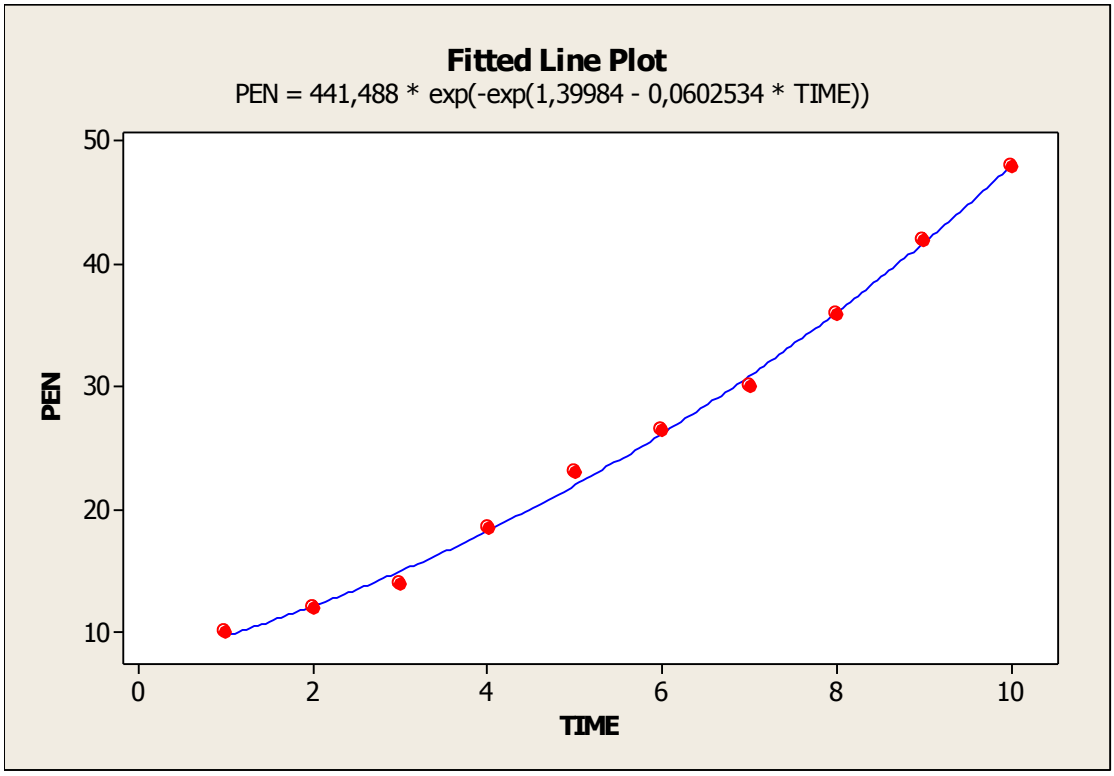
$$\hat{Y}_t = \theta_1 \exp\left(-\exp\left(\theta_2 - \theta_3 t\right)\right)$$

Παράδειγμα: Ευρυζωνικότητα σε όλα τα κράτη μέλη της ΕΕ.

TIME	PENETRATION
1	10
2	12
3	14
4	18,5
5	23

6	26,5
7	30
8	36
9	42
10	48

Καμπύλη Gompertz



MSE = 0,443506.

Επειδή η ασυμπτωτική τιμή είναι 441,488, στην προκείμενη περίπτωση υπάρχει ένα πρόβλημα. Επειδή η ευρυζωνικότητα δεν μπορεί να είναι μεγαλύτερη από το 100%, η προσαρμογή Gompertz οδηγεί σε μη φυσικά αποτελέσματα, όσον αφορά την ασύμπτωτη. Επομένως δεν μπορούμε να τη χρησιμοποιήσουμε έτσι.

Μπορούμε όμως να κάνουμε Lock την τιμή θ_1 στο 100 ή και μικρότερο προκειμένου να έχουμε μια πιο καλή προσέγγιση της πραγματικότητας, αφήνοντας τις άλλες δύο παραμέτρους ελεύθερες. Η

ασύμπτωτη τιμή θα πρέπει να οριστεί κατόπιν πολύ μεγάλης ανάλυσης, σκέψης και κρίσης.

Στις μη γραμμικές προσαρμογές πρέπει να είμαστε πολύ προσεκτικοί, γιατί μπορεί να υπάρχουν διάφορα set παραμέτρων με τοπικό ελάχιστο στα σφάλματα, αλλά εμείς ενδιαφερόμαστε για το απόλυτο ελάχιστο. Η μέθοδος Gauss Newton βρίσκει τα τοπικά ελάχιστα. Οπότε η ανάλυση αυτή μπορεί να είναι ιδιαιτέρως κοπιαστική και πρέπει να χρησιμοποιείται με πολύ προσοχή.

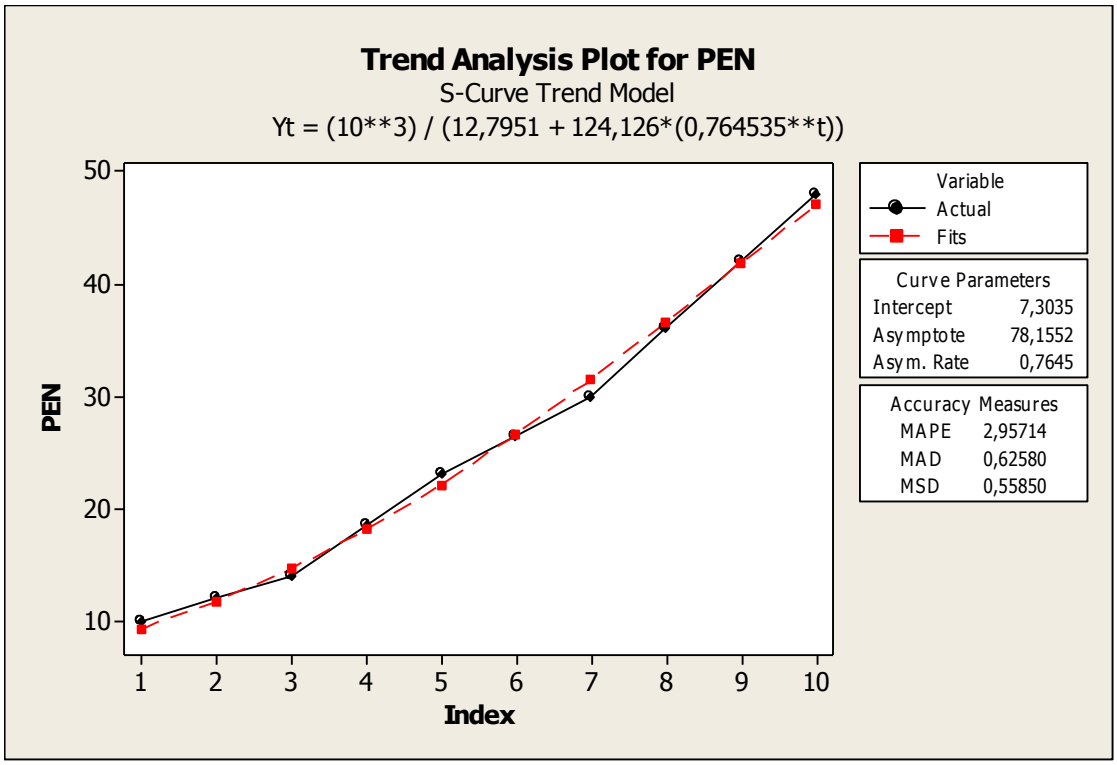
5.4.5 ΜΟΝΤΕΛΟ PEARL REED LOGISTIC

Επίσης άλλη μία καμπύλη είναι η Pearl-Reed logistic:

$$Y_t = \frac{1000}{\beta_0 + \beta_1 \beta_2^t}$$

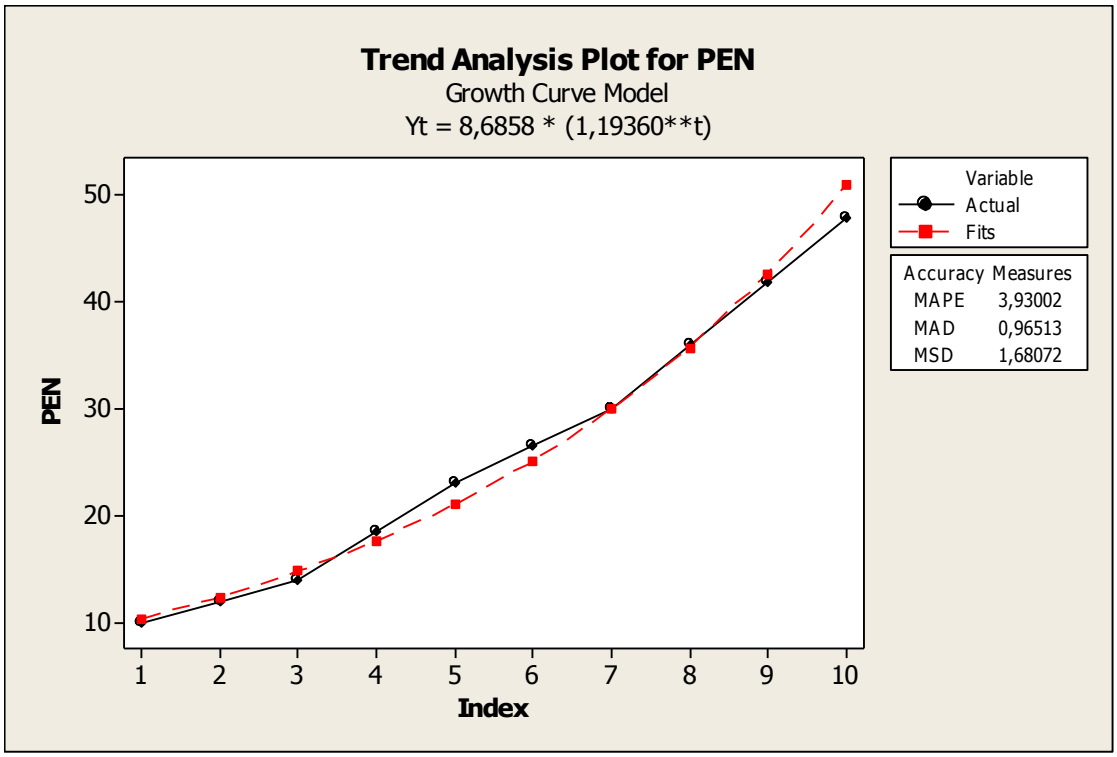
Παράδειγμα: Ευρυζωνικότητα σε όλα τα κράτη μέλη της ΕΕ.

Καμπύλη Pearl-Reed logistic



Το μοντέλο Pearl-Reed logistic δίνει στη συγκεκριμένη περίπτωση

πιο φυσικά αποτελέσματα. Για σύγκριση παραθέτουμε και το εκθετικό μοντέλο.



5.5 ΑΝΑΛΥΣΗ ΚΑΙ ΔΙΑΧΩΡΙΣΜΟΣ ΤΗΣ ΕΠΟΧΙΚΟΤΗΤΑΣ

Η γνώση των εποχικών προτύπων είναι ζωτικός παράγοντας για τον σωστό και έγκυρο σχεδιασμό. Για παράδειγμα, η πληροφορία, ότι οι πωλήσεις λουλουδιών θα είναι υψηλές τον Μάϊο, εξαιτίας της ημέρας της Μητέρας, λέει στον ανθοπώλη, ότι τα αποθέματά του σε λουλούδια θα πρέπει να είναι αυξημένα, προκειμένου να ανταποκριθεί στη ζήτηση.

Εκτός από την σχεδίαση ετήσιου επιχειρηματικού προγράμματος, η γνώση των εποχικών κινήσεων μας επιτρέπει να αναλύουμε τις επιδόσεις του παρελθόντος, να αλλάζουμε το χρονικό πρότυπο παραγωγής και διάθεσης και ίσως ακόμα να μεταβάλλουμε και το ίδιο το εποχικό πρότυπο. Αν και υπάρχουν διάφορες μέθοδοι για τη μέτρηση της εποχικής διακύμανσης, θα εστιάσουμε την προσοχή μας σε μία, στη

μέθοδο του ***λόγου της πραγματικής τιμής προς τον κινητό μέσο (ratio – to – moving average).***

Η ταυτοποίηση του εποχικού παράγοντα σε μία χρονοσειρά διαφέρει από την ανάλυση τάσης σε τουλάχιστον δύο σημεία:

- Η τάση ορίζεται απευθείας από τα δεδομένα, αλλά ο εποχικός παράγοντας ορίζεται άμεσα μετά την εξάλειψη των άλλων παραγόντων και αφήνοντας μόνο τον εποχικό.
- Η τάση αναπαριστάται με ένα "best fitting curve", ενώ η εποχικότητα θα πρέπει να υπολογιστεί για κάθε παρατηρούμενο διάστημα του χρόνου (εβδομάδα, μήνας κλπ) και γράφεται συνήθως σε μορφή ενός δείκτη. (Δείκτες είναι ποσοστά που δείχνουν αλλαγές στο χρόνο).

Για να βρούμε την εποχικότητα, θα πρέπει πρώτα να εκτιμήσουμε τη τάση και ίσως τη κυκλικότητα και μη κανονικότητα, μέσω μίας καμπύλης, όπως αναπτύξαμε στα προηγούμενα και να την αφαιρέσουμε, ή να την εκτιμήσουμε με κινητούς μέσους όρους. Αυτοί οι κινητοί μέσοι όροι θα μπορούσαν να εξαλείψουν τις κυμάνσεις, που οφείλονται σε εποχικότητα και συνυπάρχουν στα δεδομένα. Η βασική παραδοχή, που βρίσκεται πίσω από τη μέθοδο του κινητού μέσου όρου, είναι, ότι το εποχικό πρότυπο επανεμφανίζεται σε κανονική βάση κάθε επόμενο έτος. Μολονότι το εποχικό πρότυπο δεν είναι απόλυτα ακριβές κατά τα διαδοχικά έτη, ωστόσο υπάρχει μέσα σε λογικά όρια και ο κινητός μέσος όρος ελαχιστοποιεί την επίδραση σε έναν τουλάχιστον βαθμό. Θα υποθέσουμε, ότι οι συνιστώσες της χρονοσειράς είναι πολλαπλασιαστικής μορφής. Ο όρος ratio – to – moving – average χρησιμοποιείται επειδή παίρνουμε τη μεταβλητή Y_t της χρονοσειράς και τη διαιρούμε με τον κινητό μέσο όρο ως εξής:

$$\frac{Y_t}{MA} = \frac{S_t \ x \ C_t \ x \ T_t \ x \ I_t}{C_t \ x \ T_t \ x \ I_t}$$

Η ποσότητα που βρίσκουμε, η S_t είναι η καλύτερη δυνατή μέτρηση της εποχικής συνιστώσας στα δεδομένα της χρονοσειράς. Στην ουσία όμως με το κινητό μέσο απαλείφουμε και το Irregular, γιατί η χρονική σκάλα εξέλιξής του είναι μικρότερη του έτους, οπότε μένει το S*I. Λόγω όμως της τεχνικής που εφαρμόζουμε με τους διάμεσους και τη κανονικοποίηση (βήμα 4 στο επόμενο παράδειγμα) παίρνουμε την εποχικότητα.

Παράδειγμα (T5-2)

Ανάλυση της τάσης και της εποχικότητας των πωλήσεων μιας εταιρείας.

Period	Sales	12 Month M.SUM	24 Month M.SUM	12 Month cent. MA	SI
2004 Ιαν.	518				
Φεβρ.	404				
Μαρ.	300				
Απρ.	210				
Μαι.	196				
Ιουν.	186				
Ιουλ.	247	4869	9833	409.7	0.60
Αυγ.	343	4964	9916	413.2	0.83
Σεπ.	464	4952	9877	411.5	1.13
Οκτ.	680	4925	9962	415.1	1.64
Νοε.	711	5037	10067	419.5	1.69
Δεκ.	610	5030	10131	422.1	1.45
2005 Ιαν.	613	5101	10279	428.3	1.43
Φεβρ.	392	5178	10417	434.0	0.90
Μαρ.	273	5239	10691	445.5	0.61

Απρ.	322	5452	11082	461.8	0.70
Μαι.	189	5630	11444	476.8	0.40
Ιουν.	257	5814	11682	486.8	0.53
Ιουλ.	324	5868			
Αυγ.	404				
Σεπ.	677				
Οκτ.	858				
Νοε.	895				
Δεκ.	664				

Βήμα 1: Στη δεύτερη στήλη υπολογίζουμε το 12-μηνο άθροισμα αλλάζοντας κάθε φορά τον πρώτο μήνα και το γράφουμε κεντρικά, δηλαδή το MA του πρώτου δωδεκάμηνου τον γράφουμε τον Ιούλιο.

Βήμα 2: Στη τρίτη στήλη υπολογίζουμε το άθροισμα δύο διαδοχικών τιμών της προηγούμενης στήλης, δηλαδή άθροισμα κινητού 24-μήνου.

Βήμα 3: Διαιρούμε με 24 για να βρούμε το δωδεκάμηνο κινητό μέσο.

Βήμα 4: Διαιρούμε τη κανονική τιμή με το μέσο όρο, ορίζοντας έτσι τον εποχικό δείκτη **SI**. Το άθροισμα των **SI** πρέπει να είναι ίσο με 12. Γιαυτό το λόγο διορθώνουμε με κανονικοποίηση τον εποχικό δείκτη, ώστε να μπορεί να προεκταθεί ως pattern στις τιμές, που λείπουν και με αυτό το τρόπο παίρνουμε τελικά το **S**.

Διαιρώντας τις πραγματικές τιμές με τον εποχικό δείκτη παίρνουμε τα “Seasonal Adjusted Data“. Ο δωδεκάμηνος κινητός μέσος όρος δείχνει τα δεδομένα απαλλαγμένα από εποχικότητα. Για τις τιμές που λείπουν και για να αφαιρέσουμε κυκλικότητα και μη-κανονικές διακυμάνσεις κάνουμε παλινδρόμηση στα “Seasonal Adjusted Data“, οπότε παίρνουμε τη καθαρή τάση.

Στο παράδειγμα αυτό χρησιμοποιήσαμε το πολλαπλασιαστικό μοντέλο. Για το προσθετικό μοντέλο αντί για διαίρεση στο Βήμα 4 θα κάναμε αφαίρεση.

Επειδή υπάρχουν διάφορες εκτιμήσεις για τον εποχικό δείκτη από χρόνο σε χρόνο και για κάθε ίδιο μήνα, θα πρέπει αυτές να κανονικοποιηθούν, ώστε οι τιμές για τον ίδιο μήνα να είναι ίσες για όλα τα χρόνια και το άθροισμα του εποχικού δείκτη ανά 12-μηνο να ισούται με 12.

Για κάθε μήνα και για όλες τις διαθέσιμες τιμές όλων των ετών βρίσκουμε τον διάμεσο, όπως φαίνεται στον παραπάνω πίνακα. Μετά προσθέτουμε τους διάμεσους του εποχικού δείκτη και βρίσκουμε έναν πολλαπλασιαστή ως ακολούθως:

$$Multiplier = \frac{12}{\sum_{i=1}^{12} (Median)_i}$$

Το επόμενο βήμα είναι να πολλαπλασιάσουμε τον κάθε συντελεστή εποχικότητας με τον πολλαπλασιαστή. Στη τελευταία στήλη του προηγούμενου πίνακα έχουμε τις κανονικοποιημένες τιμές του δείκτη εποχικότητας.

Month	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**Median**	**Norm. S**
Ιαν.		1,208	1,202	1,272	1,411	1,431	1,272	1,278
Φεβρ.		0,700	0,559	0,938	1,089	0,903	0,903	0,907
Μαρ.		0,524	0,564	0,785	0,800	0,613	0,613	0,616
Απρ.		0,444	0,433	0,480	0,552	0,697	0,480	0,482
Μαι.		0,424	0,365	0,488	0,503	0,396	0,424	0,426
Ιουν.		0,490	0,459	0,461	0,465	0,528	0,465	0,467
Ιουλ.	0,639	0,904	0,598	0,681	0,603	0,662	0,651	0,654

Αυγ.	1,115	0,913	0,889	0,799	0,830	0,830	0,860	0,864
Σεπ.	1,371	1,560	1,346	1,272	1,128	1,395	1,359	1,365
Οκτ.	1,792	1,863	1,796	1,574	1,638	1,771	1,782	1,790
Νοε.	1,884	2,012	1,867	1,697	1,695	1,846	1,857	1,865
Δεκ.	1,519	1,088	1,224	1,282	1,445		1,282	1,288
							11,948	**12,002**

5.6 ΑΠΟΜΟΝΩΣΗ ΤΗΣ ΚΥΚΛΙΚΟΤΗΤΑΣ

Στη παράγραφο αυτή θα ασχοληθούμε με μια εφαρμογή της μεθόδου του κινητού μέσου όρου προκειμένου να απομονωθεί η κυκλική συνιστώσα. Αυτή η διαδικασία απομόνωσης είναι παρόμοια με τη θεωρία απομόνωσης της τάσης και της εποχικής συνιστώσας, με τις οποίες ασχοληθήκαμε νωρίτερα.

Η απομόνωση μιας κυκλικής συνιστώσας επιτυγχάνεται με την απάλειψη της εποχικότητας από τα αρχικά δεδομένα ως εξής:

$$\frac{Y_t}{\text{εποχικός δείκτης}} = \frac{S_t \; x \; C_t \; x \; T_t \; x \; I_t}{S_t} = C_t \; x \; T_t \; x \; I_t$$

Τα δεδομένα αυτά, τα οποία είναι απαλλαγμένα από την εποχικότητα, μπορούν να απαλλαγούν από την τάση ως εξής:

$$\frac{C_t \; x \; T_t \; x \; I_t}{T_t} = C_t \; x \; I_t$$

Οι σειρές που προκύπτουν περιλαμβάνουν μόνο την επίδραση που σχετίζεται με την κυκλική συνιστώσα και τη μη – κανονική διακύμανση. Η μη – κανονική διακύμανση μπορεί να εξαλειφθεί από τη χρονοσειρά μέσω υπολογισμού του κινητού μέσου όρου, που εξομαλύνει αυτές τις μη – κανονικές διακυμάνσεις από τα δεδομένα. Συνήθως παίρνουμε περιττό αριθμό χρονικών περιόδων για τον κινητό μέσο. Για μηνιαία δεδομένα

αρκεί κινητός μέσος 5, 7, 9 ή και 11 χρονικών περιόδων, ενώ για τριμηνιαία δεδομένα αρκεί κινητός μέσος 3 περιόδων.

Η συνιστώσα τυχαιότητας υπολογίζεται ως εξής:

$$I_t = \frac{C_t \ x \ I_t}{C_t}$$

Παράδειγμα (T5-4)

Στο παράδειγμα αυτό ένας αναλυτής μιας επιχείρησης χρησιμοποιήσε την ανάλυση αυτοσυσχέτισης για να προσδιορίσει, ότι οι πωλήσεις είναι εποχικές σε μια τριμηνιαία βάση (από το 2000 έως και το 2006 σε τρίμηνα). Τώρα θα χρησιμοποιήσει τον διαχωρισμό για να καταλάβει την μεταβλητότητα των τριμηνιαίων πωλήσεων.

Sales	T	SCI	S	TCI	CI	C	I
232,700	262,000	0,888	0,780	298,458	1,139		
309,200	262,759	1,177	1,016	304,434	1,159	1,118	1,036
310,700	263,518	1,179	1,117	278,239	1,056	1,078	0,980
293,000	264,276	1,109	1,088	269,300	1,019	1,022	0,997
205,100	265,035	0,774	0,780	263,059	0,993	0,960	1,034
234,400	265,794	0,882	1,016	230,787	0,868	0,940	0,924
285,400	266,553	1,071	1,117	255,582	0,959	0,906	1,059
258,700	267,312	0,968	1,088	237,775	0,890	0,924	0,962
193,200	268,071	0,721	0,780	247,796	0,924	0,927	1,000
263,700	268,830	0,981	1,016	259,635	0,966	0,954	1,012
292,500	269,589	1,085	1,117	261,940	0,972	1,003	0,969
315,200	270,348	1,166	1,088	289,705	1,072	0,962	1,114
178,300	271,107	0,658	0,780	228,686	0,844	0,970	0,870
274,500	271,866	1,010	1,016	270,269	0,994	0,936	1,062
295,400	272,625	1,084	1,117	264,537	0,970	0,976	0,994

286,400	273,383	1,048	1,088	263,234	0,963	0,942	1,022
190,800	274,142	0,696	0,780	244,718	0,893	0,933	0,957
263,500	274,901	0,959	1,016	259,438	0,944	0,957	0,986
318,800	275,660	1,157	1,117	285,492	1,036	0,998	1,038
305,300	276,419	1,104	1,088	280,605	1,015	1,058	0,960
242,600	277,178	0,875	0,780	311,156	1,123	1,089	1,031
318,800	277,937	1,147	1,016	313,886	1,129	1,104	1,023
329,600	278,696	1,183	1,117	295,164	1,059	1,100	0,963
338,200	279,455	1,210	1,088	310,844	1,112	1,078	1,032
232,100	280,214	0,828	0,780	297,689	1,062	1,059	1,004
285,600	280,973	1,016	1,016	281,198	1,001	0,996	1,005
291,000	281,732	1,033	1,117	260,597	0,925	0,947	0,977
281,400	282,491	0,996	1,088	258,639	0,916		

Ο αλγόριθμος έχει ως ακολούθως: Με τη μέθοδο του “ratio to moving average” υπολογίζουμε τον εποχικό δείκτη. Διαιρώντας τις πραγματικές τιμές με τον εποχικό δείκτη παραμένουν οι τιμές TCI. Ο υπολογισμός της τάσης υπολογίζεται με τη μέθοδο ελαχίστων τετραγώνων στις τιμές TCI δηλαδή στις „seasonal adjustment” τιμές. Για τον υπολογισμό της κυκλικής συνιστώσας C παίρνουμε κινητούς μέσους στις τιμές CI με 3 όρους για τα τριμηνιαία δεδομένα (από 5 έως 11 όρους για τα μηνιαία δεδομένα). Μετά διαιρούμε το CI με το C και βρίσκουμε το I. Το MINITAB δεν διαχωρίζει το CI από το I. Τα residual είναι οι τιμές CI και μάλιστα τις δίνει στο αθροιστικό μοντέλο.

Επίσης πρέπει να λάβουμε υπόψη μας, ότι το MINITAB παίρνει πάντα γραμμική τάση. Εάν η τάση δεν είναι γραμμική, θα πρέπει να τη βρούμε μόνοι μας, αναπτύσσοντας όλη τη διαδικασία: πρώτα βρίσκουμε την εποχικότητα, με τον τρόπο των κινητών μέσων, που είναι

ανεξάρτητος της τάσης, μετά διαιρούμε τις πραγματικές τιμές με τον εποχικό δείκτη, ώστε να εφαρμόσουμε ανάλυση τάσης με την κατάλληλη καμπύλη πάνω στις τιμές TCI κλπ.

Υπολογισμός των παραμέτρων για το δεύτερο τρίμηνο του 2000:

$$\hat{T}_t = 261.24 + 0.759t \qquad SCI = \frac{Y}{T} = \frac{232.7}{262.0} = 0.888$$

$$TCI = \frac{232.7}{.77967} = 298.458$$

$$CI = \frac{Y}{TS} = \frac{232.7}{(262.0)(.77967)} = 1.139$$

$$I = \frac{CI}{C} = \frac{1.159}{1.118} = 1.036$$

Οι εποχικοί δείκτες είναι οι εξής:

Πρώτο τρίμηνο = 0.77967 ---> 78.0 %

Δεύτερο τρίμηνο = 1.01566 ---> 101.6 %

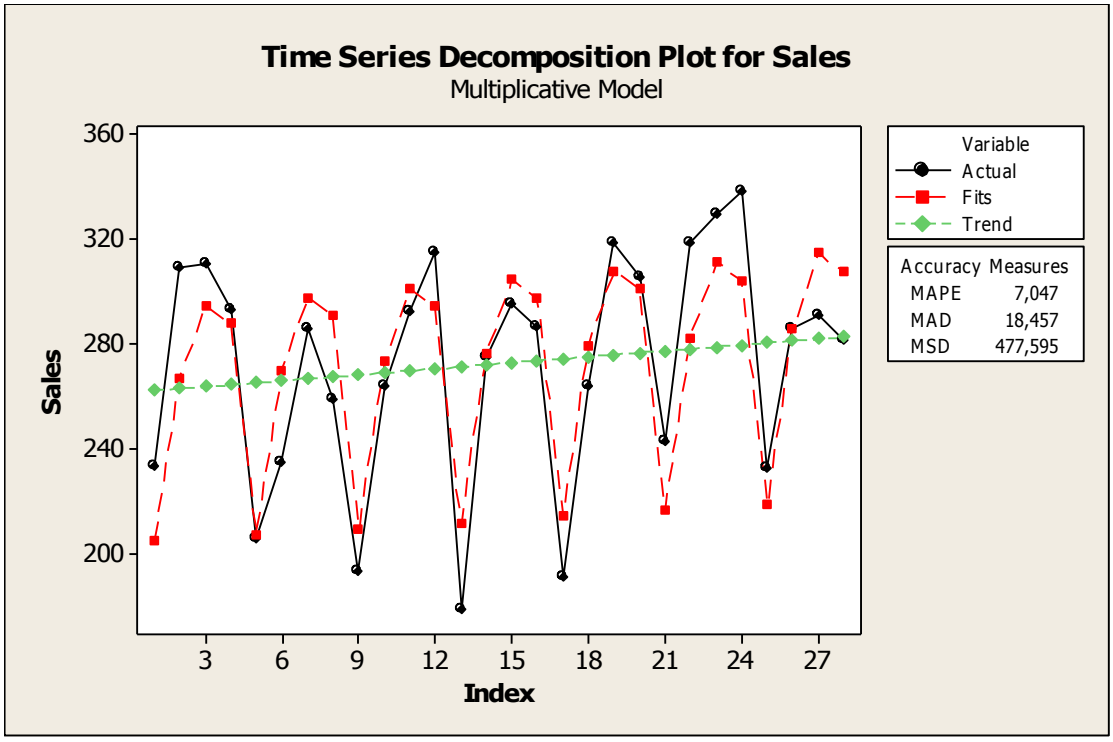
Τρίτο τρίμηνο = 1.11667 ---> 111.7 %

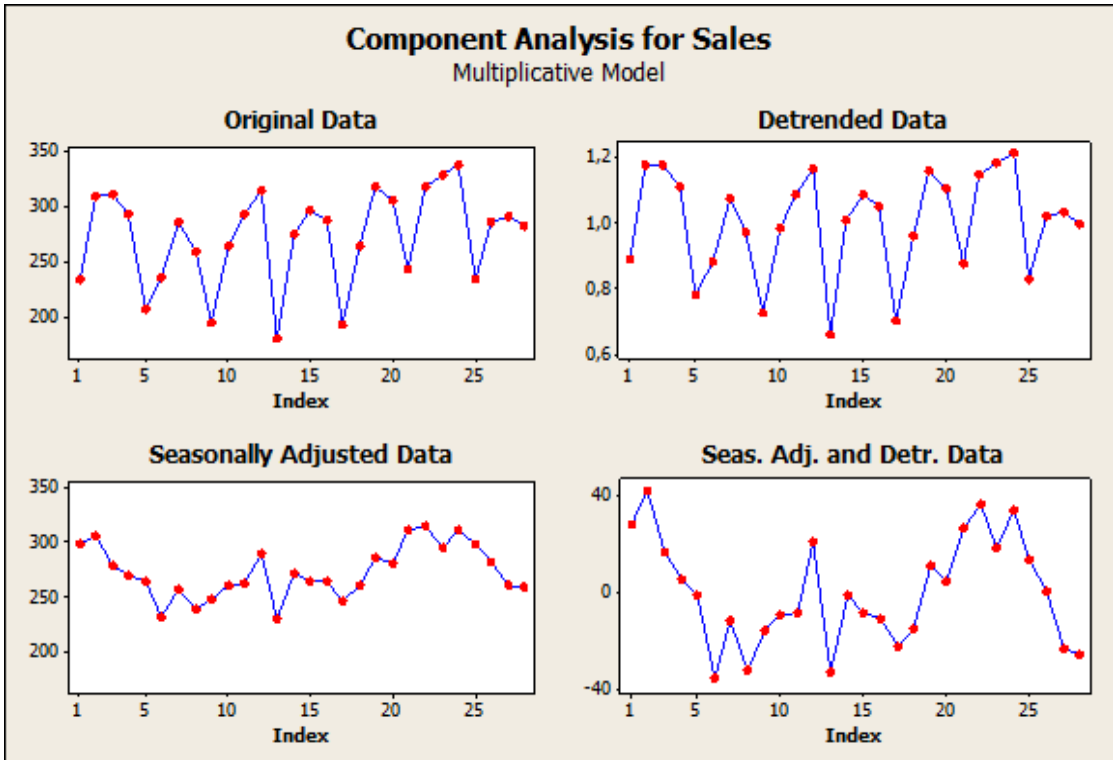
Τέταρτο τρίμηνο = 1.08800 ---> 108.8 %

Για να κάνουμε πρόβλεψη για το πρώτο τρίμηνο του 2007 βρίσκουμε τη τιμή της τάσης από την ευθεία ελαχίστων τετραγώνων, παίρνουμε τη τιμή του εποχικού δείκτη S για το πρώτο τρίμηνο και τα πολλαπλασιάζουμε μαζί με τον κυκλικό δείκτη C, που είναι περίπου 1 στη περίπτωσή μας και με το δείκτη τυχαιότητας I πάλι 1, επειδή δεν ξέρουμε κάτι γι αυτόν.

Οπότε: $\hat{T}_{29} = 261.24 + 0.759(29) = 283.251$

$$\hat{Y}_{29} = T_{29}\ x\ S_{29}\ x\ C_{29}\ x\ I_{29} = (283.251)(.77967)(1.0)(1.0) = 220.842$$





5.7 ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΙ ΔΕΙΚΤΕΣ

Κυκλικοί δείκτες χρησιμοποιούνται συνήθως για να απαντήσουμε

τις ακόλουθες ερωτήσεις:

- Εμφανίζει κυκλικότητα η προς έρευνα χρονοσειρά;
- Εάν ναι πόσο έντονη είναι αυτή;
- Ακολουθεί η χρονοσειρά την γενικότερη κατάσταση της οικονομίας (Επιχειρηματικοί Δείκτες);

Ένας τρόπος να ερευνήσουμε τα κυκλικά σχήματα στις χρονοσειρές είναι να μελετήσουμε και να συγκρίνουμε αυτά με τα σχήματα επιχειρηματικών δεικτών. Ως επιχειρηματικός δείκτης ορίζεται μία χρονοσειρά, συσχετιζόμενη με κάποια επιχειρηματικότητα και χρησιμοποιούνται για την εκτίμηση της γενικής κατάστασης της οικονομίας. Πολλοί επιχειρηματίες και οικονομολόγοι ακολουθούν συστηματικά την συμπεριφορά τέτοιων δεικτών, προκειμένου να πάρουν πληροφορίες, να κατανοήσουν τα αίτια και να μειώσουν τα έξοδα στις επιχειρήσεις τους.

Οι στατιστικές υπηρεσίες δημοσιεύουν συνήθως δείκτες κάθε μήνα. Οι δείκτες αυτοί χωρίζονται σε τρεις κατηγορίες:

1. Leading indicators: Τους χρησιμοποιούμε για να ανιχνεύσουμε σημεία καμπής στην οικονομία και την μελλοντική κατεύθυνση της οικονομίας. Π.χ. Ο δείκτης έκδοσης των οικοδομικών αδειών και ο δείκτης μετοχών του χρηματιστηρίου.
2. Coincident indicators: Τους χρησιμοποιούμε για να κατανοήσουμε την τρέχουσα κατάσταση της οικονομίας. Π.χ. Ο δείκτης βιομηχανικής παραγωγής, ο δείκτης λιανικού εμπορίου κλπ.
3. Lagging indicators: Αυτοί οι δείκτες έχουν μία χρονοκαθυστέρηση ως προς την κατάσταση της οικονομίας. Π.χ. βασικό επιτόκιο, ποσοστό ανεργίας.

Οι Leading indicators αλλάζουν κατεύθυνση στα σημεία καμπής

της γενικής οικονομίας. Οι Coincident indicators αυξομειώνονται ταυτόχρονα με τις αυξήσεις, ή μειώσεις της γενικής οικονομίας και οι Lagging indicators ακολουθούν την πορεία της γενικής οικονομίας με κάποια καθυστέρηση.

Παρόλα αυτά είναι δύσκολο να ταυτοποιήσουμε σημεία καμπής, γιατί δεν αναπτύσσονται όλοι οι κλάδοι της οικονομίας με τον ίδιο τρόπο, ή τον ίδιο ρυθμό. Οι Leading indicators σηματοδοτούν συνήθως τις αλλαγές της οικονομίας, οπότε είναι οι σημαντικότεροι. Οι Coincident και Lagging indicators μας βοηθούν να εκτιμήσουμε την απόδοση της τωρινής και παρελθούσας οικονομίας σε συγκεκριμένους κλάδους και να βοηθήσουν στην ανάπτυξη μελλοντικών πολιτικών.

5.8 ΔΕΙΚΤΕΣ ΤΙΜΩΝ

Πάρα πολλές χρονοσειρές, που αφορούν την παραγωγή, πωλήσεις και άλλες οικονομικές δραστηριότητες δίδονται σε κύκλο εργασιών: Ευρώ, Δολάριο κλπ. Ο πληθωρισμός και γενικότερα αλλαγές τιμών με τη πάροδο του χρόνου μπορεί να είναι αιτία προβλημάτων ανάλυσης. Π.χ. Μία αύξηση πωλήσεων σε κύκλο εργασιών μπορεί να κρύβει μειωμένες πωλήσεις σε όγκο, μονάδες προϊόντων δηλαδή, όταν υπάρχει πληθωρισμός.

Σε αυτές τις περιπτώσεις γίνεται σημαντική η έννοια της αγοραστικής δύναμης. Η αγοραστική δύναμη ορίζεται ως:

$$Αγοραστική\ Δύναμη\ σε\ 1\ Ευρώ = \frac{100}{Δείκτη\ Τιμών\ Καταναλωτή}$$

Αυτό σημαίνει ότι, όταν τον Νοέμβριο του 2006 ο Δ.Τ.Κ. φθάσει το 150, παίρνοντας σταθερή τιμή το 2002 ως 100, τότε η αγοραστική δύναμη τον Νοέμβριο του 2006 θα είναι:

$$Αγοραστική\ Δύναμη\ σε\ 1\ Ευρώ = \frac{100}{150} = 0.67$$

Αυτό σημαίνει, ότι η αγοραστική δύναμη τον Νοέμβριο του 2006 είναι τα 2/3 αυτής του 2002.

Για να εκφράσουμε τώρα τις τιμές Ευρώ σε σταθερό Ευρώ, με βάση κάποιο έτος, χρησιμοποιούμε την ακόλουθη εξίσωση:

$$Αποπληθωρισμένη\ αξία\ σε\ Ευρώ = (Αγοραστική\ Δύναμη\ σε\ 1\ Ευρώ)\ x\ (Πραγματική\ (Τρέχουσα)\ Αξία)$$

Ας υποθέσουμε ότι η αξία ενός αυτοκινήτου αυξήθηκε από 300,000 Ευρώ το 2005 στα 350,000 Ευρώ το 2006, ενώ ο δείκτης τιμών αυτοκινήτων, με σταθερές τιμές 2002, αυξήθηκε από 135 σε 155. Οι αποπληθωρισμένες τιμές για τα έτη 2005 και 2006 είναι:

$$Αποπληθωρισμένη\ αξία\ 2005 = (300,000)(\frac{100}{135}) = 222,222$$

$$Αποπληθωρισμένη\ αξία\ 2006 = (350,000)(\frac{100}{155}) = 225,806$$

Αυτό σημαίνει, ότι η διαφορά τρεχουσών τιμών ήταν 50,000 Ευρώ, ενώ των αποπληθωρισμένων 3,584 Ευρώ.

5.9 Η ΜΕΘΟΔΟΣ ΑΠΟΣΥΝΘΕΣΗΣ CENSUS II

Στις αρχές της δεκαετίας του 1950 ο Julius Shiskin Δ/ντής οικονομικών στατιστικών στο U.S. Bureau of the Census δημιούργησε ένα πρόγραμμα μεγάλης κλίμακας για τη διάσπαση των χρονοσειρών. Η νεότερη version του προγράμματος είναι το X-13-ARIMA. Το πρόγραμμα αυτό είναι διαθέσιμο χωρίς κόστος από το site www.census.gov/srd/www/x12a/.

Σύμφωνα με τη μέθοδο που εφαρμόζεται στο πρόγραμμα γίνονται

επαναλήψεις βημάτων προκειμένου να αποσυνθέσει τις χρονοσειρές στα συστατικά τους: τάση - κυκλικότητα, εποχικότητα και μη κανονικότητα των δεδομένων της χρονοσειράς.

Η μέθοδος περιλαμβάνει 12 βήματα προς το σκοπό αυτό.

Βήμα 1: Εφαρμογή κινητού μέσου s περιόδων με s=12 για μηνιαία και s=4 τριμηνιαία δεδομένα και υπολογίζουμε χονδρικά την τάση – κύκλο.

Βήμα 2: Ο λόγος ανάμεσα στα πραγματικά δεδομένα και τις τιμές του κινητού μέσου σε πολλαπλασιαστική αποσύνθεση υπολογίζεται, όπως στα προαναφερόμενα. Μόνο που εδώ ο κύκλος – τάση είναι μαζί και όχι ξεχωριστά. Διαιρώντας τις πραγματικές τιμές με τη τάση – κύκλο παίρνουμε το SI.

Βήμα 3: Οι λόγοι από το Βήμα 2 περιέχουν εποχικά και μη κανονικά συστατικά. Εφαρμόζοντας κινητούς μέσους για το SI βρίσκουμε το S και διαιρώντας με τα χονδρικά εποχικά συστατικά S παίρνουμε τα μη κανονικά συστατικά. Οι τιμές που λείπουν από την αρχή και το τέλος της χρονοσειράς, λόγω χρήσης των κινητών μέσων, αντικαθίστανται με εκτιμήσεις. Επίσης ακραίες τιμές των μη κανονικών συστατικών στο SI αντικαθίστανται.

Βήμα 4: Οι λόγοι από τα τροποποιημένα δεδομένα "λειαίνονται" πάλι με κινητούς μέσους, προκειμένου να εξαλειφθούν τα μη κανονικά συστατικά. Αυτό δημιουργεί μία προκαταρκτική εκτίμηση των εποχικών συστατικών S.

Βήμα 5: Τα αρχικά δεδομένα διαιρούνται με τα προκαταρκτικά εποχικά προσαρμοσμένα δεδομένα και παίρνουμε το TI. Δηλαδή:

$$\frac{Y_t}{S_t} = \frac{S_t \; x \; T_t \; x \; I_t}{S_t} = T_t \; x \; I_t$$

Βήμα 6: Η τάση – κύκλος εκτιμάται παίρνοντας έναν σταθμισμένο μέσο όρο από την στο βήμα 5 ΤΙ χρονοσειρά, δηλαδή την προκαταρκτικά προσαρμοσμένη χρονοσειρά.

Βήμα 7: Επανάληψη του βήματος 2, ώστε να υπολογιστεί πάλι ο κύκλος τάσης. Οι νέοι λόγοι που περιέχουν μόνο εποχικά και μη κανονικά συστατικά, υπολογίζονται διαιρώντας τα αρχικά δεδομένα με τη τάση από το βήμα 6 βάση της σχέσης:

$$\frac{Y_t}{T_t} = \frac{S_t \; x \; T_t \; x \; I_t}{T_t} = S_t \; x \; I_t$$

Βήμα 8: Επανάληψη του βήματος 3 με τους λόγους του βήματος 7.

Βήμα 9: Επανάληψη του βήματος 4 με τα δεδομένα του βήματος 8.

Βήμα 10: Επανάληψη του βήματος 5 με τα δεδομένα του βήματος 9.

Βήμα 11: Διαιρούμε τα εποχικά προσαρμοσμένα δεδομένα από το βήμα 10 με τη τάση βρίσκοντας, όπως στο βήμα 6, τα μη κανονικά συστατικά.

Βήμα 12: Εδώ ελέγχουμε για ακραίες τιμές των μη κανονικών συστατικών κάνουμε εκτίμηση των τιμών που λείπουν και αν δεν είμαστε ικανοποιημένοι από το αποτέλεσμα ξεκινάμε πάλι από το βήμα 1 με δεδομένα από το βήμα 12.

5.10 CASE 5-4 MURPHY BROTHERS FURNITURE

Στο παράδειγμα 4-4, η Julie Murphy ανέπτυξε ένα μοντέλο Naïve, το οποίο συνδύαζε τις εκτιμήσεις για τάση και εποχικότητα. Ο βασικότερος λόγος για τον οποίο επέλεξε το συγκεκριμένο μοντέλο, ήταν

η απλότητά του. Ήθελε ένα απλό μοντέλο πρόβλεψης, προκειμένου να το αντιλαμβάνεται ο πατέρας της και να πειστεί να το υιοθετήσει για την εταιρεία τους.

Έστω ότι τώρα είναι Οκτώβριος του 2002 και πολλά πράγματα έχουν αλλάξει. Ο Glen Murphy αποσύρθηκε, ενώ η Julie έχει ολοκληρώσει αρκετά μαθήματα διοίκησης επιχειρήσεων στο τοπικό πανεπιστήμιο, συμπεριλαμβανομένων και μαθημάτων επιχειρησιακών προβλέψεων. Αξίζει να σημειωθεί ότι η εταιρεία Murphy Brothers Furniture είχε χτίσει ένα εργοστάσιο στο Dallas και είχε ολοκληρώσει τη δική της βιομηχανική γραμμή παραγωγής από τον Οκτώβριο του 1995.

Τα δεδομένα των μηνιαίων πωλήσεων των Murphy Brothers από το 1996 έως τον Οκτώβριο του 2002 φαίνονται στον πίνακα 5-10. Όπως υποδεικνύεται από το «pattern» που φαίνεται ξεκάθαρα στο σχήμα 5-13, οι πωλήσεις έχουν αυξηθεί σημαντικά από το 1996. Δυστυχώς, το σχήμα 5-13 δείχνει ότι το σημαντικό πρόβλημα της ζήτησης, είναι τα σημάδια εποχικότητας (seasonality) που παρουσιάζει. Η γενική πολιτική που ακολουθεί η εταιρεία είναι να λειτουργεί σε δύο βάρδιες κατά τη διάρκεια του καλοκαιριού και των πρώτων μηνών του φθινοπώρου και για το υπόλοιπο του χρόνου να δουλεύει με μία βάρδια. Έτσι, με τις διπλές βάρδιες δημιουργούνται τα ουσιαστικά αποθέματα που θα χρειαστούν καθώς αυξάνεται η ζήτηση, μέχρι και το «peak» που παρουσιάζει κατά τους μήνες Νοέμβριο και Δεκέμβριο. Εξαιτίας των προαναφερθέντων απαιτήσεων παραγωγής, η Julie είναι πολύ ανήσυχη προκειμένου να ετοιμάσει τις βραχυπρόθεσμες προβλέψεις για τις πωλήσεις της εταιρείας. Οι προβλέψεις θα πρέπει να είναι βασισμένες στις καλύτερες διαθέσιμες πληροφορίες αναφορικά με τη ζήτηση.

Για τους σκοπούς της πρόβλεψης, η Julie αποφάσισε να

χρησιμοποιήσει τα δεδομένα από το 1996 και μετά, από όταν δηλαδή πρωτοεγκατέστησαν τη δική τους βιομηχανική γραμμή παραγωγής επίπλων (δεδομένα Table 5-10). Η Julie παρατήρησε από το σχήμα 5-13 ότι τα δεδομένα τους εμπεριέχουν και τάση και εποχικότητα. Για το λόγο αυτό, αποφάσισε να χρησιμοποιήσει την τεχνική της ανάλυσης χρονοσειράς στα συστατικά της (Decomposition), με σκοπό να προσεγγίσει με τον καλύτερο δυνατό τρόπο τη μεταβλητή των πωλήσεων.

Η Julie αποφάσισε να χρησιμοποιήσει τα στοιχεία της χρονοσειράς με το προσθετικό (additive) μοντέλο για να κάνει την πρόβλεψη που επιθυμεί. Το μοντέλο που θα τρέξει είναι το: Yt = Tt + St + It. Τα αποτελέσματα φαίνονται περιληπτικά στον πίνακα 5-11. Επιπλέον, για να ελέγξει την τυχαιότητα, χρησιμοποίησε το διάγραμμα αυτοσυσχέτισης των σφαλμάτων (Residuals), όπως φαίνεται στο σχήμα 5-14. Όμως, τα σφάλματα δεν ήταν τυχαία και άρα το μοντέλο κρίνεται ως ανεπαρκές.

Μέχρι τώρα είχε χρησιμοποιήσει τη μέθοδο Naïve συνδυασμένη με εκτιμήσεις τάσης και εποχικότητας, τη μέθοδο εκθετικής εξομάλυνσης Winters' και την κλασική αποσύνθεση. Τελικά, αποφάσισε να ρυθμίσει την εποχικότητα από τα δεδομένα με τέτοιο τρόπο ώστε να μπορεί να χρησιμοποιήσει τεχνικές πρόβλεψης που δεν περιλαμβάνουν εποχικότητα. Κατάφερε, λοιπόν, να βγάλει από τα δεδομένα την εποχικότητα προσθέτοντας (ή αφαιρώντας) τον εποχικό δεικτή από τον κατάλληλο μήνα. Για παράδειγμα, πρόσθεσε 674,6 στα δεδομένα για κάθε Ιανουάριο και αφαίρεσε 1.234,63 από τα δεδομένα για κάθε Δεκέμβρη. Ο πίνακας 5-12 αποτελείται από τα δεδομένα, ρυθμισμένα χωρίς εποχικότητα.

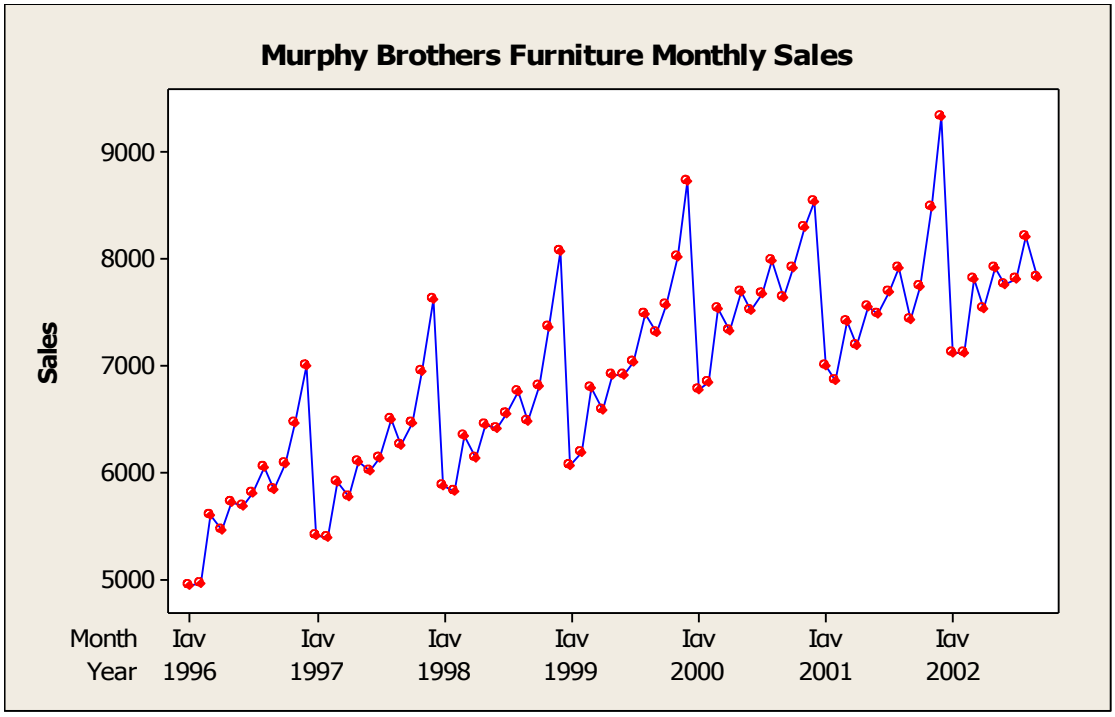
	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**Jun.**	**Jul.**	**Aug.**	**Sep.**	**Oct.**	**Nov.**	**Dec**

1996	4946	4968	5601	5454	5721	5690	5804	6040	5843	6087	6469	7002
1997	5416	5393	5907	5768	6107	6016	6131	6499	6249	6472	6946	7615
1998	5876	5818	6342	6143	6442	6407	6545	6758	6485	6805	7361	8079
1999	6061	6187	6792	6587	6918	6920	7030	7491	7305	7571	8013	8727
2000	6776	6847	7531	7333	7685	7518	7672	7992	7645	7923	8297	8537
2001	7005	6855	7420	7183	7554	7475	7687	7922	7426	7736	8483	9329
2002	7120	7124	7817	7538	7921	7757	7816	8208	7828			

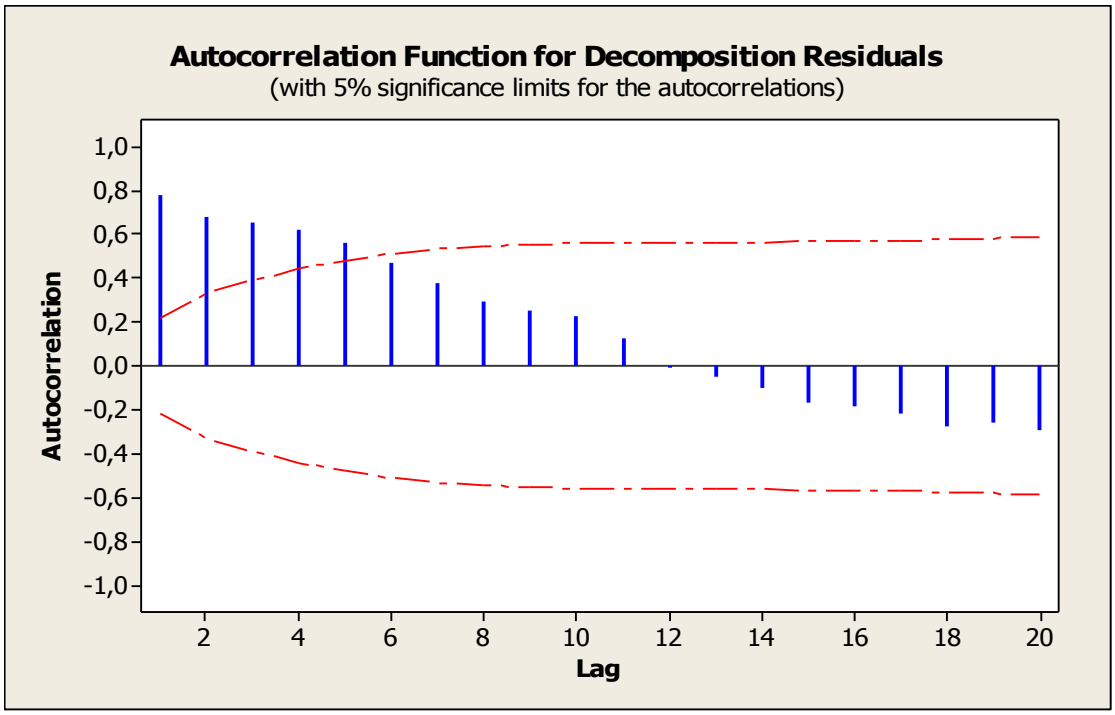
Πίνακας 5-10 Μηνιαίες πωλήσεις της Murphy Brothers Furniture, 1996 - 2002

	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**Jun.**	**Jul.**	**Aug.**	**Sep.**	**Oct.**	**Nov.**	**Dec**
1996	5621	5671	5745	5821	5775	5863	5847	5818	5901	5941	5857	5767
1997	6091	6096	6051	6135	6161	6189	6174	6277	6307	6326	6334	6380
1998	6551	6521	6486	6510	6496	6580	6588	6536	6543	6659	6749	6844
1999	6736	6890	6936	6954	6972	7093	7073	7269	7363	7425	7401	7492
2000	7451	7550	7675	7700	7739	7691	7715	7770	7703	7777	7685	7302
2001	7680	7558	7564	7550	7608	7648	7730	7700	7484	7590	7871	8094
2002	7795	7827	7961	7905	7975	7930	7859	7986	7886			

Πίνακας 5-12 Μηνιαίες πωλήσεις της Murphy Brothers Furniture, χωρίς την εποχική συνιστώσα, 1996-2002



Σχήμα 5-13 Μηνιαίες πωλήσεις της επιχείρησης Murphy Brothers, 1996 - 2002



Σχήμα 5-14 Διάγραμμα αυτοσυσχέτισης καταλοίπων με τη χρήση του προσθετικού μοντέλου αποσύνθεσης χρονοσειράς

Fitted Trend Equation

Yt = 5672,0 + 31,4*t

Seasonal Indices

Period	Index
1	-674,60
2	-702,56
3	-143,72
4	-366,64
5	-53,52
6	-173,27
7	-42,74
8	222,32
9	-57,95
10	145,76
11	612,30
12	1234,63

Accuracy Measures

MAPE	1,9
MAD	135,1
MSD	30965,3

Πίνακας 5-11 Περιγραφή μοντέλου αποσύνθεσης για την επιχείρηση των Murphy Brothers

ΑΣΚΗΣΕΙΣ

1. **Χρησιμοποιώντας τα δεδομένα του 2001 του πίνακα 5-12, αναπτύξτε ένα μοντέλο για να προβλέψετε τις πωλήσεις των εννιά πρώτων μηνών του 2002, απαλλαγμένες από την**

εποχικότητα που περιέχουν τα δεδομένα των πωλήσεων.

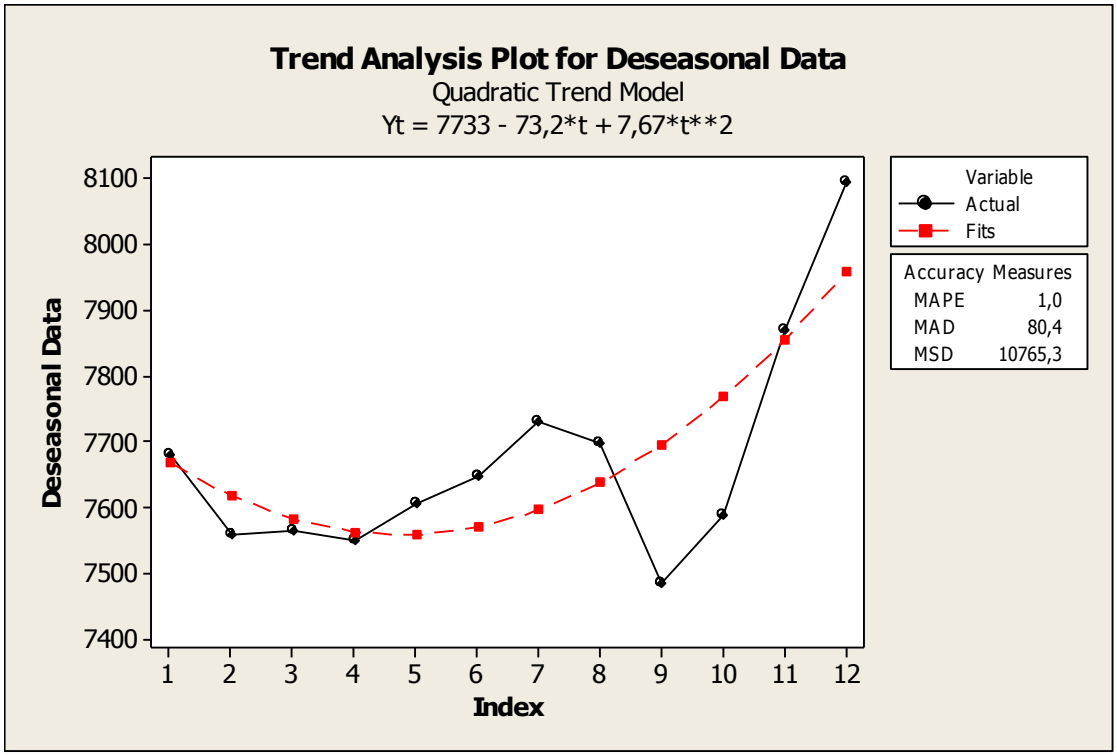
Θεωρούμε δεδομένες τις πωλήσεις του 2001, απαλλαγμένες από την εποχική συνιστώσα, όπως φαίνεται παρακάτω:

time	DESEAS
1	7680
2	7558
3	7564
4	7550
5	7608
6	7648
7	7730
8	7700
9	7484
10	7590
11	7871
12	8094

Προκειμένου να προβλέψουμε τις πωλήσεις για τους εννέα πρώτους μήνες του 2002, θα πρέπει να χρησιμοποιήσουμε την μέθοδο της αποσύνθεσης χρονοσειράς και συγκεκριμένα το προσθετικό μοντέλο $Y_t = T_t + S_t + I_t$. Αρχικά, απομονώσαμε την εποχική συνιστώσα και στη συνέχεια απομονώνουμε την συνιστώσα της τάσης. Για να μπορέσουμε να αναγνωρίσουμε τη μορφή της τάσης, πραγματοποιούμε μια ανάλυση τάσης και για τα 3 μοντέλα (γραμμικό, παραβολικό, εκθετικό). Συγκρίνοντας στη συνέχεια τους δείκτες αξιοπιστίας, διαπιστώνουμε πως το μοντέλο που εμφανίζει καλύτερες τιμές και άρα προσεγγίζει ακριβέστερα την τάση μας είναι το παραβολικό μοντέλο (MSE =

10765,3).

Με βάση λοιπόν την σχέση Yt = 7733 - 73,2*t + 7,67*t**2, βρίσκουμε την τάση μας για t = 1,..,12. Το αποτέλεσμα φαίνεται στο παρακάτω σχήμα. Μπορούμε τώρα να αφαιρέσουμε από τα Deseasonal Data την τάση, ώστε να βρούμε και το δείκτη τυχαιότητας I_t. Τα αποτελέσματα αυτά φαίνονται παρακάτω:



Time	T+I	S	T	I
1	7680	-674,60	7667,47	12,53
2	7558	-702,56	7617,28	-59,28
3	7564	-143,72	7582,43	-18,43
4	7550	-366,64	7562,92	-12,92
5	7608	-53,52	7558,75	49,25
6	7648	-173,27	7569,92	78,08
7	7730	-42,74	7596,43	133,57
8	7700	222,32	7638,28	61,72
9	7484	-57,95	7695,47	-211,47
10	7590	145,76	7768,00	-178,00
11	7871	612,30	7855,87	15,13
12	8094	1234,63	7959,08	134,92

Έχοντας απομονώσει όλα τα χαρακτηριστικά μας (S, T, I),

μπορούμε να κάνουμε τις προβλέψεις μας για το επόμενο έτος θέτοντας time = 13,...21 (9 μήνες) και υπολογίζοντας τις 9 τιμές για την τάση από την παραβολική εξίσωση. Ο εποχικός συντελεστής που απεικονίζεται στον πίνακα 5-11 χρησιμοποιείται και εδώ. Η τιμή του It, δηλαδή η συμβολή της μη κανονικότητας για την εκάστοτε μελλοντική περίοδο, δεν μπορεί να καθοριστεί, αφού εξαρτάται από τυχαίους και απρόσμενους παράγοντες και κατά συνέπεια δεν μπορεί να προσδιοριστεί.

Time	T+I	S	T	I
13	8078,63	-674,60	8077,63	1,00
14	8212,52	-702,56	8211,52	1,00
15	8361,75	-143,72	8360,75	1,00
16	8526,32	-366,64	8525,32	1,00
17	8706,23	-53,52	8705,23	1,00
18	8901,48	-173,27	8900,48	1,00
19	9112,07	-42,74	9111,07	1,00
20	9338,00	222,32	9337,00	1,00
21	9579,27	-57,95	9578,27	1,00

Επομένως, τα προβλεπόμενα δεδομένα (χωρίς την εποχική συνιστώσα) για το πρώτο εννεάμηνο του 2002 φαίνονται στη δεύτερη στήλη.

2. **Χρησιμοποιώντας την προηγούμενη πρόβλεψη των εννιά πρώτων μηνών του 2002, στις οποίες έχουν προστεθεί ή αφαιρεθεί οι κατάλληλοι εποχιακοί δείκτες του πίνακα 5-11. Είναι αυτές οι προβλέψεις ακριβείς, σε σύγκριση με τις πραγματικές τιμές;**

Στο ερώτημα αυτό καλούμαστε να προσθέσουμε (ή να αφαιρέσουμε) τον εποχικό συντελεστή στα παραπάνω δεδομένα που προβλέψαμε. Στον παρακάτω πίνακα παρουσιάζεται το αποτέλεσμα αυτής της ενέργειας. Στη δεύτερη στήλη φαίνονται οι προβλεπόμενες

τιμές, ενώ στην τρίτη στήλη φαίνονται οι πραγματικές τιμές, όπως αυτές παρουσιάστηκαν στον πίνακα 5-10 για το έτος 2002.

Time	Forecasts	Actuals
13	7404,03	7120
14	7509,96	7124
15	8218,03	7817
16	8159,68	7538
17	8652,71	7921
18	8728,21	7757
19	9069,33	7816
20	9560,32	8208
21	9521,32	7828

Παρατηρούμε κάποιες σημαντικές αποκλίσεις μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Το γεγονός αυτό μας οδηγεί στο συμπέρασμα πως η μέθοδος της αποσύνθεσης της χρονοσειράς δεν αποφέρει τα αναμενόμενα αποτελέσματα και δεν θεωρείται μια αξιόπιστη μέθοδος για πρόβλεψη. Αυτό ήταν αναμενόμενο, καθώς στο διάγραμμα αυτοσυσχέτισης καταλοίπων 5-14 παρατηρούμε πως οι αυτοσυσχετίσεις δεν περνάνε την Q στατιστική, άρα τα κατάλοιπα δεν είναι ανεξάρτητα.

3. Προβλέψτε τις πωλήσεις του Οκτωβρίου 2002 χρησιμοποιώντας την ίδια διαδικασία με την ερώτηση 2.

Χρησιμοποιώντας την παραβολική εξίσωση της τάσης, όπως προέκυψε στο παραπάνω ερώτημα, Yt = 7733 - 73,2*t + 7,67*t**2, βρίσκουμε την τάση των πωλήσεων μας τον μήνα Οκτώβριο του 2002 (t = 22). Προκύπτει ότι T = 9834,88. Για να προβλέψουμε τώρα την τιμή των πωλήσεων αρκεί να προσθέσουμε στο παραπάνω νούμερο τον εποχικό συντελεστή για τον μήνα Οκτώβρη (seasonal index = 145,76) και το συντελεστή τυχαιότητας (που τον θεωρούμε ίσο με τη μονάδα). Οι πωλήσεις για τον Οκτώβρη του 2002 θα είναι 9981,64. Ωστόσο, όπως αναφέραμε και παραπάνω, το μοντέλο αυτό δεν είναι αξιόπιστο για την

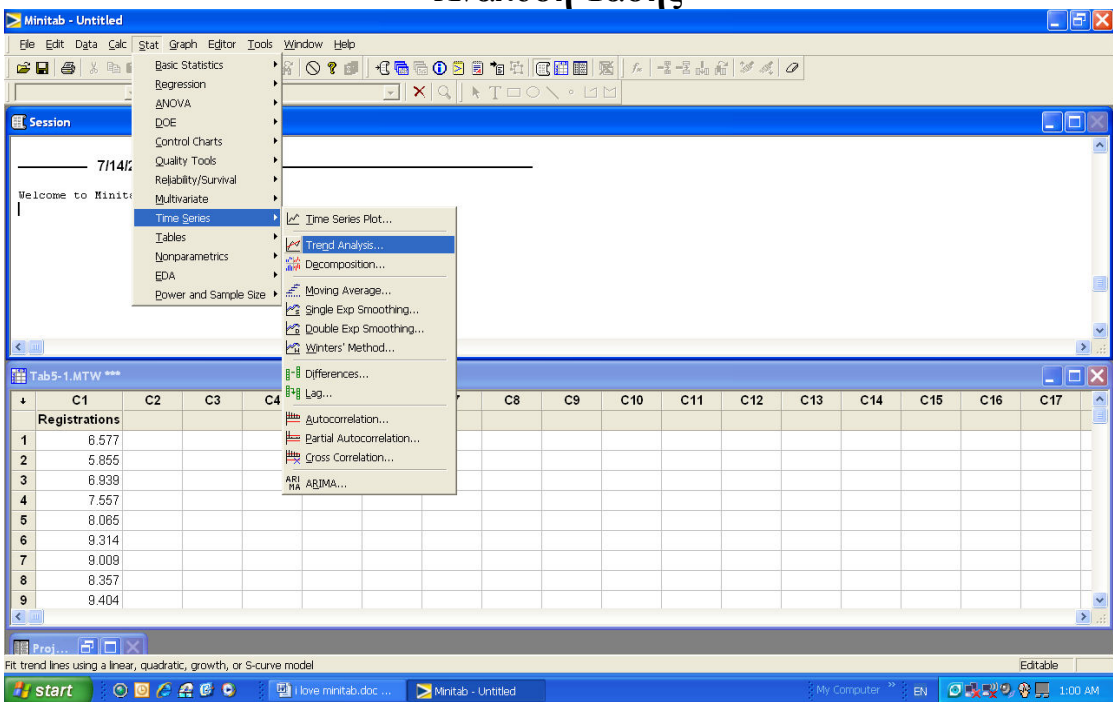
διενέργεια προβλέψεων.

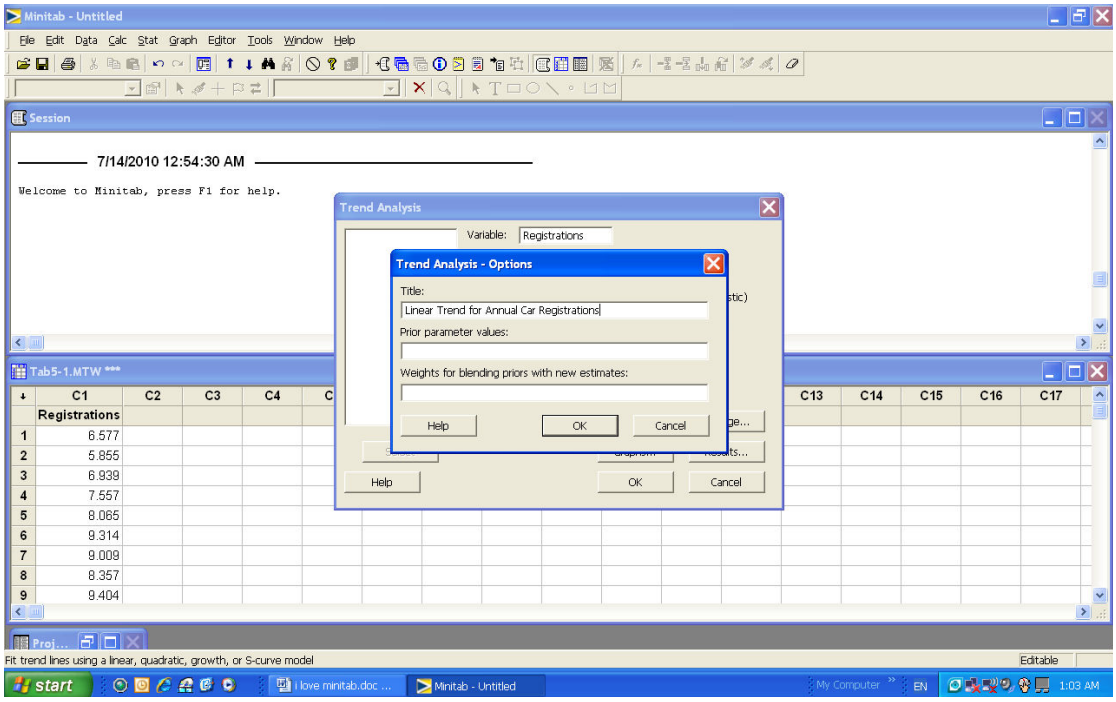
4. **Να συγκρίνετε το «pattern» των λιανικών πωλήσεων που παρουσιάστηκαν στο case 3-1A με το «pattern» των πραγματικών πωλήσεων από το 1992 έως το 1995 που παρουσιάστηκε στο case 4-4 και με το «pattern» των πραγματικών πωλήσεων από το 1996 έως το 2001 που παρουσιάστηκαν στο παρόν case.**

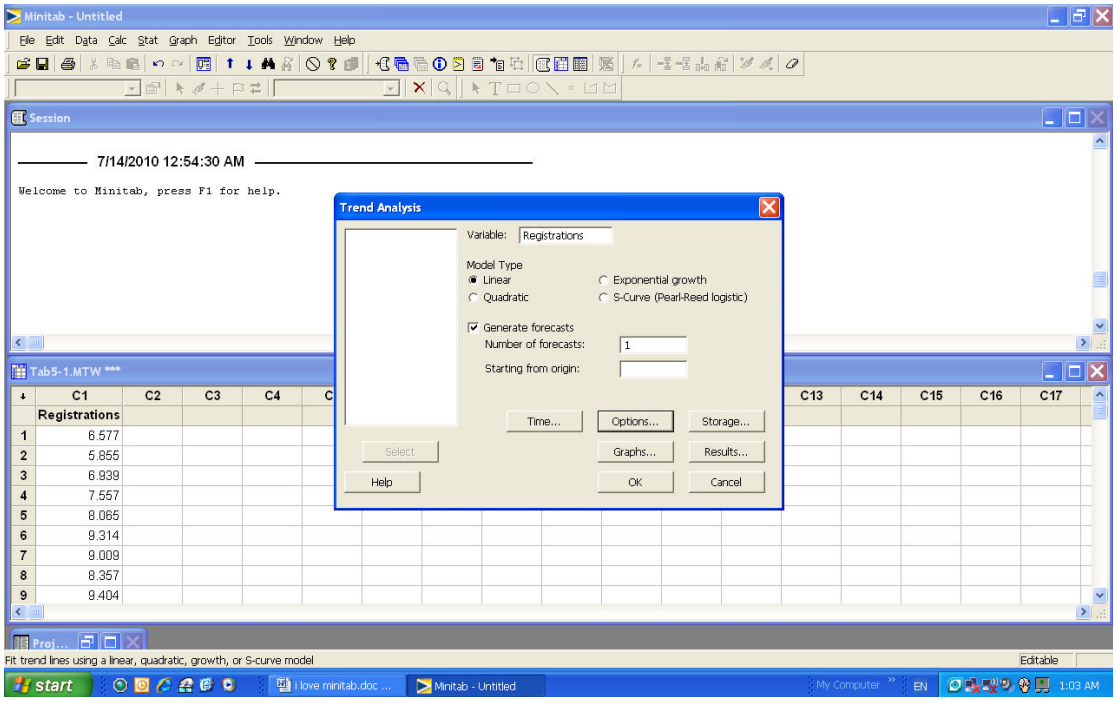
Συγκρίνοντας τα πραγματικά δεδομένα πωλήσεων των ετών 1992 -1995 με αυτά των επομένων ετών 1996 -2002, διαπιστώνουμε μια γενικότερη αύξηση στις πωλήσεις λιανικής. Αυτό ίσως να οφείλεται στις βελτιωμένες κινήσεις και πρακτικές πρόβλεψης που χρησιμοποιήθηκαν από το 1996 και μετά. Η μέθοδος naïve που χρησιμοποιήθηκε ως αρχική μέθοδος πρόβλεψης, λόγω της απλότητας της, δεν ήταν η καταλληλότερη για πρόβλεψη. Οι «συγκρατημένες» πωλήσεις των πρώτων ετών μετά την αλλαγή διοίκησης της εταιρίας, έδωσαν τη θέση τους στις αυξημένες τιμές των επομένων ετών. Φαίνεται πως οι επιχειρηματικές πρακτικές, οι πρακτικές μάρκετινγκ και οι δυνατότητες πρόβλεψης της Julie, συνεισέφεραν σημαντικά σε αυτήν την αλλαγή. Εντούτοις, η αύξηση αυτή θα μπορούσε να προέλθει και από τυχαία γεγονότα όπως αλλαγή (αύξηση) ορισμένων μεταβλητών του άμεσου περιβάλλοντος. Παραδείγματος χάρη, η αύξηση του ακαθάριστου εγχώριου προϊόντος (ΑΕΠ), η αύξηση των εισοδημάτων των πολιτών, ο καταναλωτισμός και ο υλισμός θα μπορούσαν και από μόνες τους να προκαλέσουν αυτή την αύξηση στις πωλήσεις.

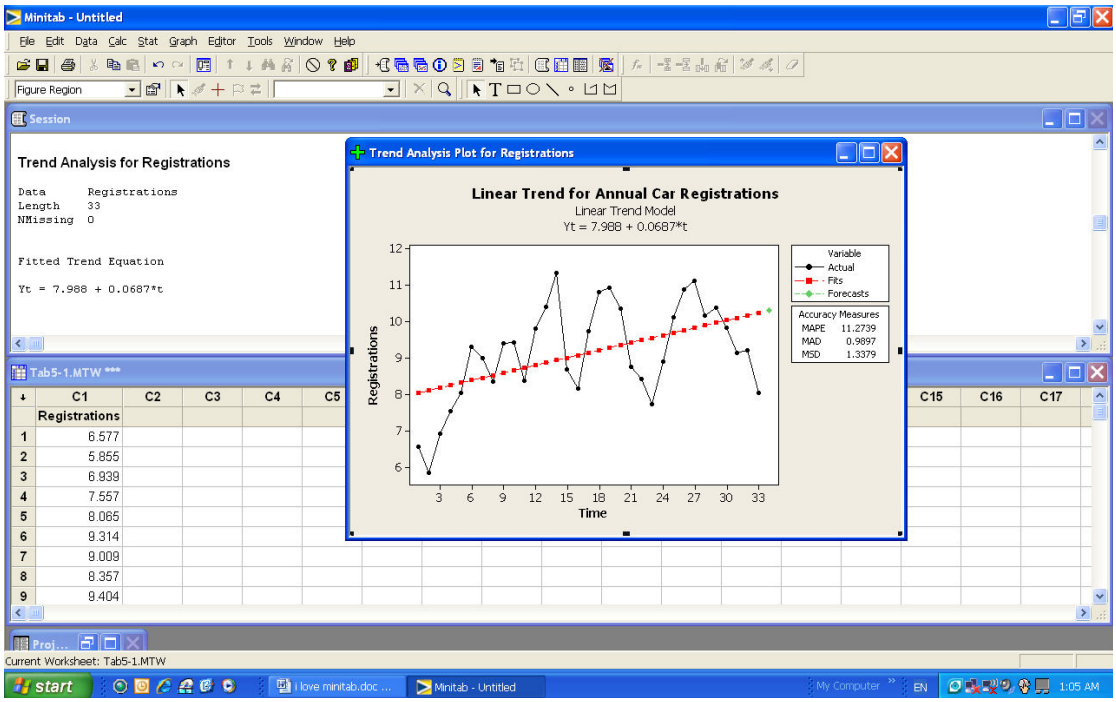
5.11 ΧΡΗΣΗ MINITAB

Ανάλυση Τάσης

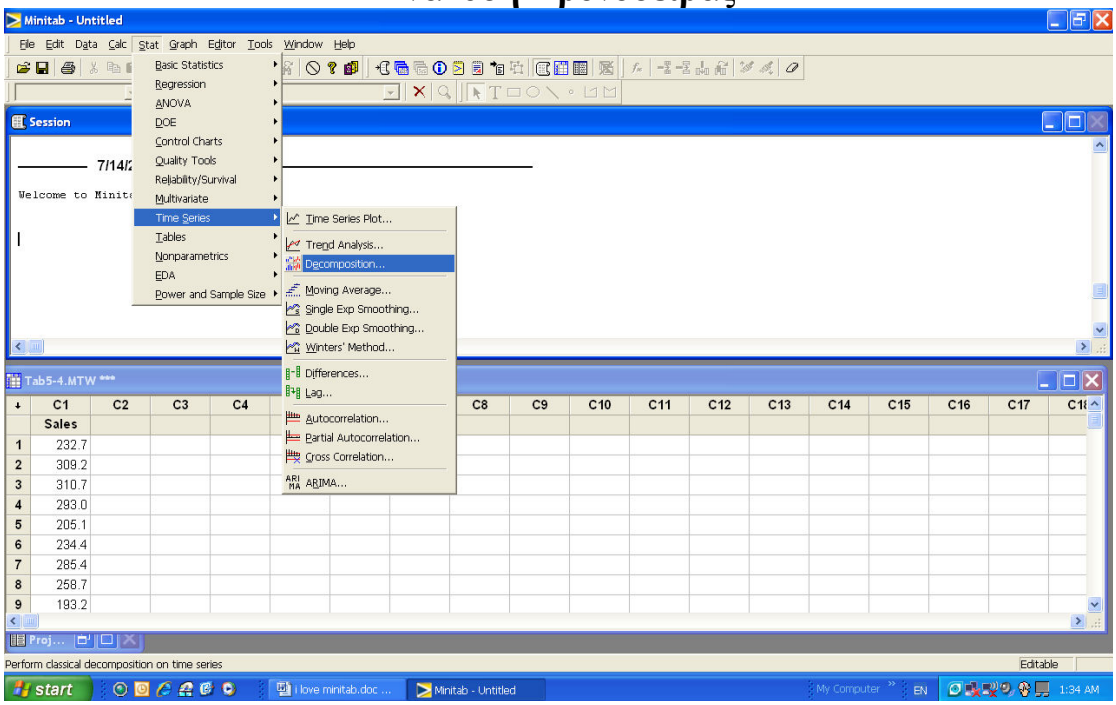


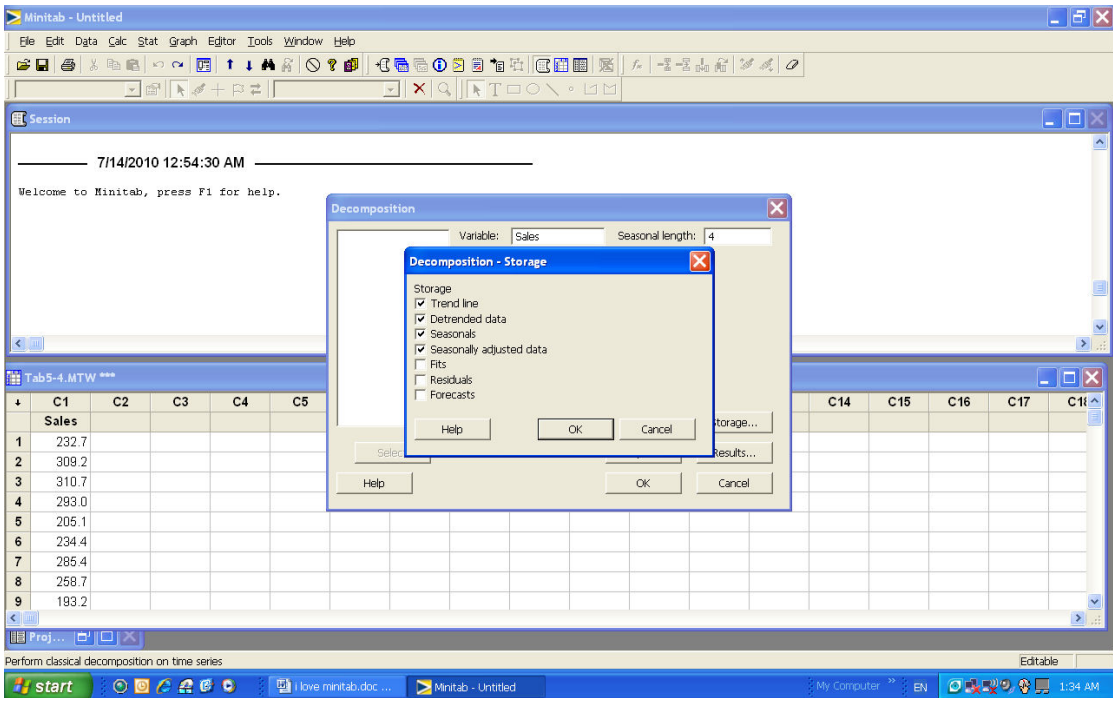


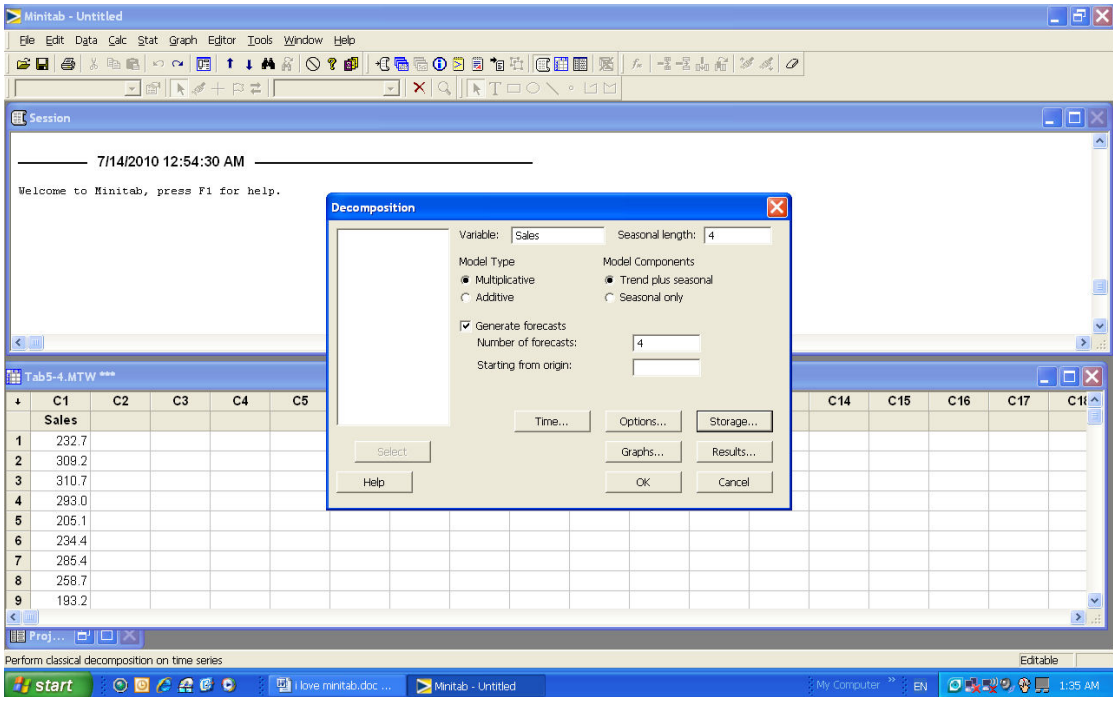

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/14/2010 12:54:30 AM
Welcome to Minitab, press F1 for help.
Trend Analysis
Variable: Registrations
Model Type
Linear
Quadratic
Exponential growth
S-Curve (Pearl-Reed logistic)
Generate forecasts
Number of forecasts: 1
Starting from origin:
Time... Options... Storage...
Select Graphs... Results...
Help OK Cancel
Tab5-1.MTW ***
C1 Registrations
1 6.577
2 5.855
3 6.939
4 7.557
5 8.065
6 9.314
7 9.009
8 8.357
9 9.404
Fit trend lines using a linear, quadratic, growth, or S-curve model
Editable
start
Minitab - Untitled
1:03 AM

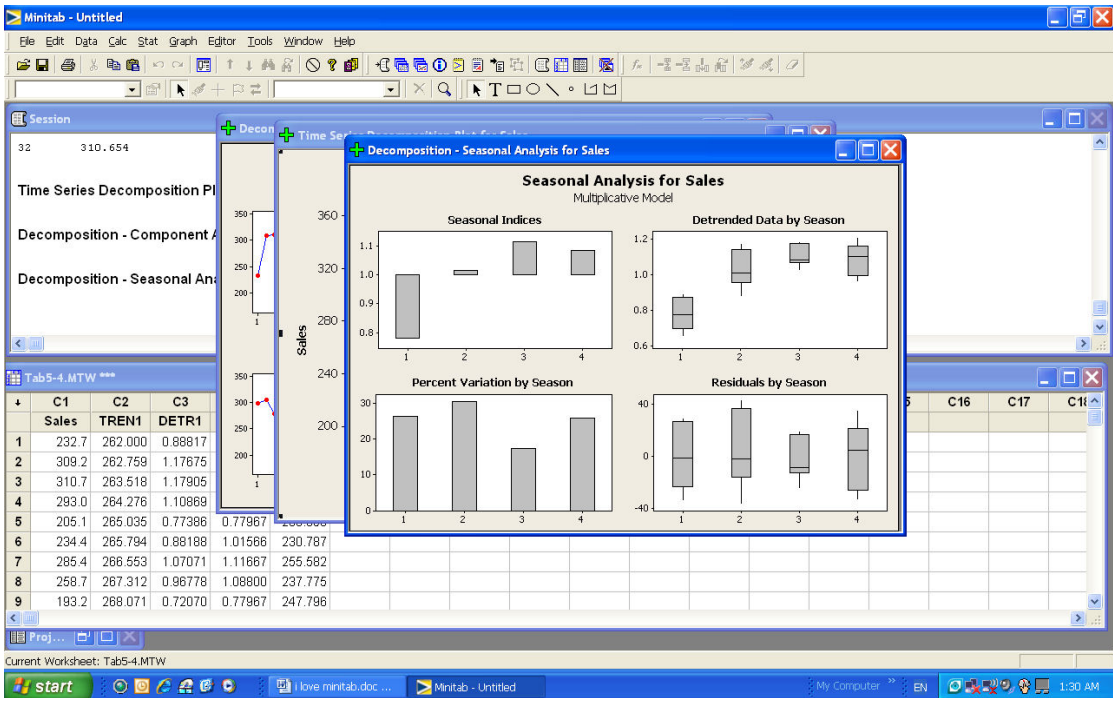

Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Figure Region
Session
Trend Analysis for Registrations
Data Registrations
Length 33
NMissing 0
Fitted Trend Equation
Yt = 7.988 + 0.0687*t
Trend Analysis Plot for Registrations
Linear Trend for Annual Car Registrations
Linear Trend Model
Yt = 7.988 + 0.0687*t
Registrations
Time
Variable
Actual
Fits
Forecasts
Accuracy Measures
MAPE 11.2739
MAD 0.9897
MSD 1.3379
Tab5-1.MTW ***
C1 Registrations
1 6.577
2 5.855
3 6.939
4 7.557
5 8.065
6 9.314
7 9.009
8 8.357
9 9.404
Current Worksheet: Tab5-1.MTW
start
Minitab - Untitled
1:05 AM

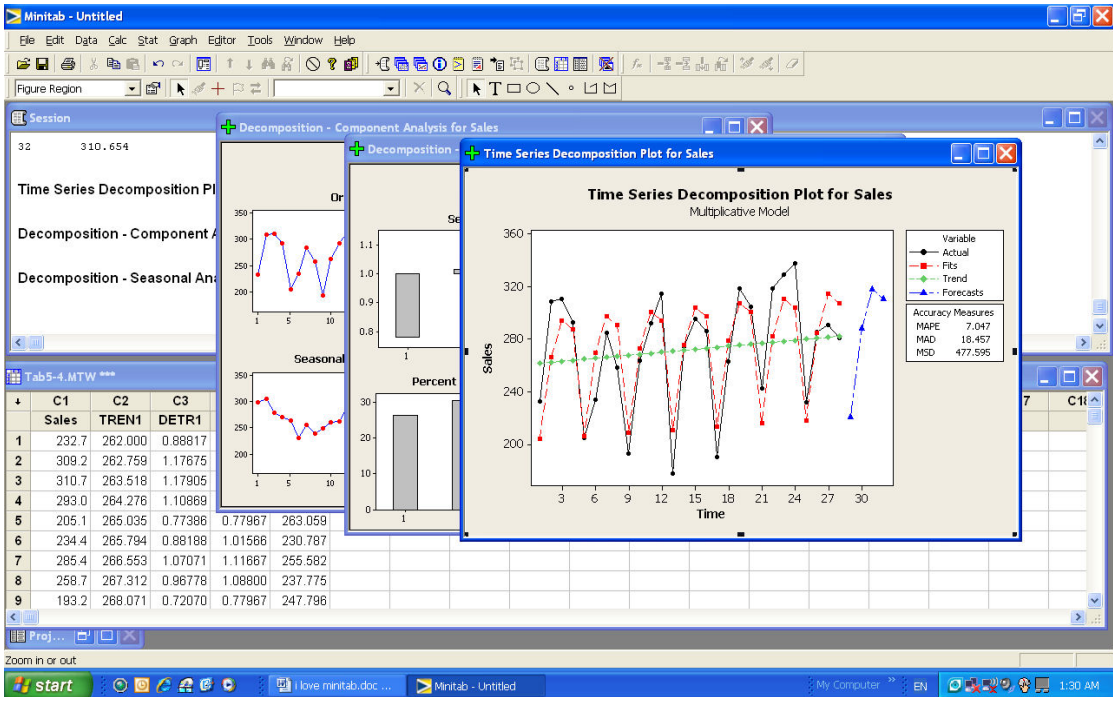
Ανάλυση Χρονοσειράς

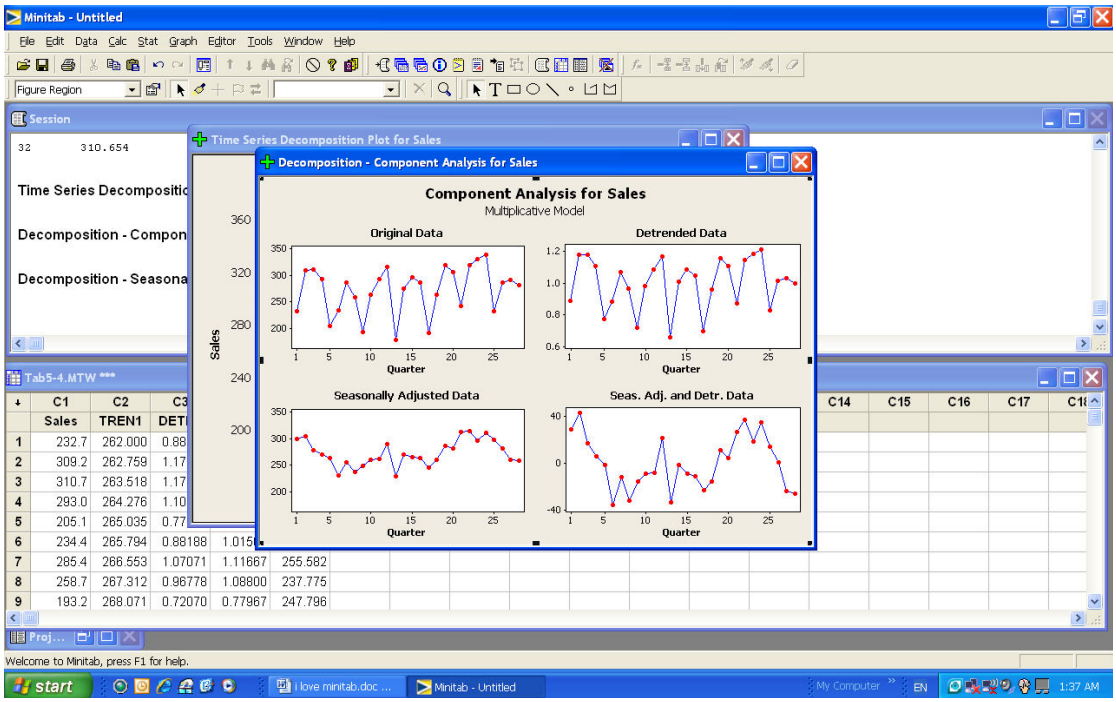





Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
7/14/2010 12:54:30 AM
Welcome to Minitab, press F1 for help.
Decomposition
Variable: Sales
Seasonal length: 4
Model Type
Multiplicative
Additive
Model Components
Trend plus seasonal
Seasonal only
Generate forecasts
Number of forecasts: 4
Starting from origin:
Time...
Options...
Storage...
Select
Graphs...
Results...
Help
OK
Cancel
Tab5-4.MTW ***
C1 C2 C3 C4 C5 C14 C15 C16 C17 C18
Sales
1 232.7
2 309.2
3 310.7
4 293.0
5 205.1
6 234.4
7 285.4
8 258.7
9 193.2
Perform classical decomposition on time series
Editable
start
Minitab - Untitled
My Computer
EN
1:35 AM


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
32 310.654
Time Series Decomposition Pl
Decomposition - Component A
Decomposition - Seasonal Ana
Decomposition - Seasonal Analysis for Sales
Seasonal Analysis for Sales
Multiplicative Model
Seasonal Indices
Detrended Data by Season
Percent Variation by Season
Residuals by Season
Sales
Tab5-4.MTW ***
C1 C2 C3 C16 C17 C18
Sales TREN1 DETR1
1 232.7 262.000 0.88817
2 309.2 262.759 1.17675
3 310.7 263.518 1.17905
4 293.0 264.276 1.10869
5 205.1 265.035 0.77386 0.77967
6 234.4 265.794 0.88188 1.01566 230.787
7 285.4 266.553 1.07071 1.11667 255.582
8 258.7 267.312 0.96778 1.08800 237.775
9 193.2 268.071 0.72070 0.77967 247.796
Current Worksheet: Tab5-4.MTW
start
Minitab - Untitled
My Computer
EN
1:30 AM


Minitab - Untitled
Time Series Decomposition Plot for Sales
Multiplicative Model
Sales
Time
Variable
Actual
Fits
Trend
Forecasts
Accuracy Measures
MAPE 7.047
MAD 18.457
MSD 477.595
Seasonal
Percent
Tab5-4.MTW ***
Zoom in or out


Minitab - Untitled
Time Series Decomposition Plot for Sales
Decomposition - Component Analysis for Sales
Component Analysis for Sales
Multiplicative Model
Original Data
Detrended Data
Seasonally Adjusted Data
Seas. Adj. and Detr. Data
Quarter
Sales
Tab5-4.MTW ***
Welcome to Minitab, press F1 for help.

ΚΕΦΑΛΑΙΟ 6 (ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ)

6.1 ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο 2 συζητήθηκε η γραμμική συσχέτιση μεταξύ δύο αριθμητικών μεταβλητών (correlation), προσαρμόζοντας τις τιμές με τη μέθοδο των ελαχίστων τετραγώνων (method of least squares). Στο κεφάλαιο αυτό, θα συζητηθούν λεπτομερώς μοντέλα απλής γραμμικής παλινδρόμησης και πρόβλεψη της εξαρτημένης μεταβλητής από την ανεξάρτητη ή ερμηνευτική μεταβλητή.

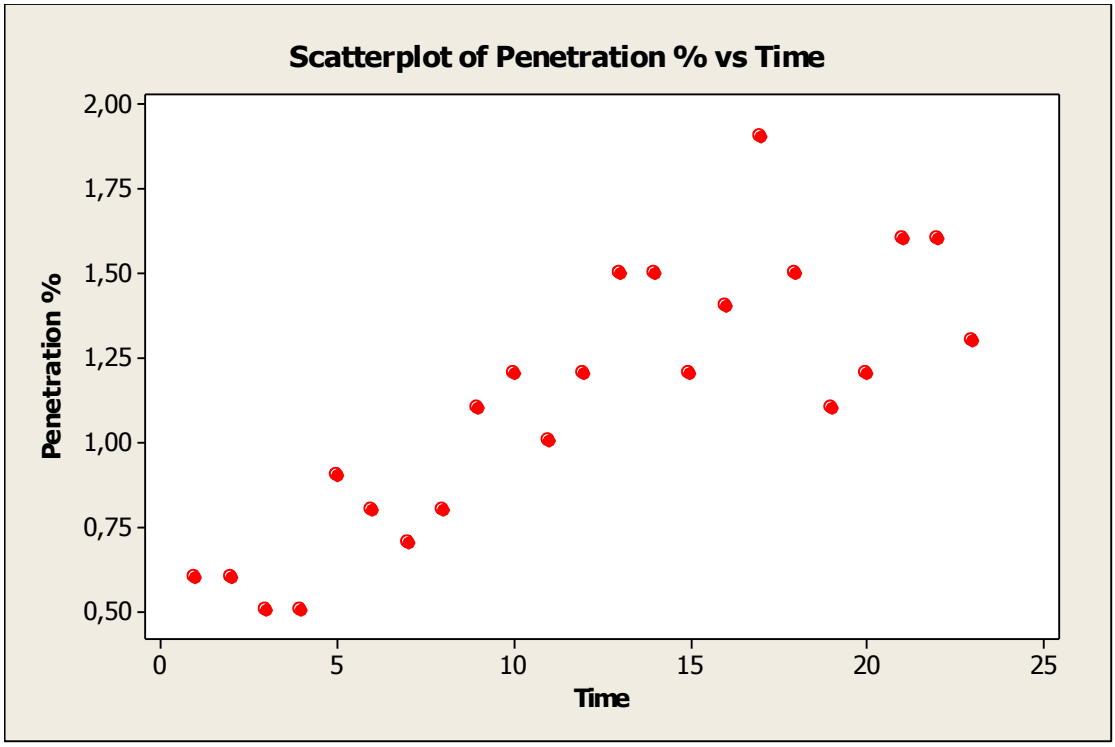
Σαν επανάληψη της ανάλυσης της σχέσης μεταξύ δύο μεταβλητών, που συζητήθηκε στο Κεφάλαιο 2, παρουσιάζεται το ακόλουθο παράδειγμα.

Παράδειγμα

Στο παράδειγμα αυτό συγκεντρώθηκαν τα δεδομένα της επί τοις εκατό αλλαγής της διείσδυσης της ευρυζωνικής πρόσβασης στην Ευρώπη από το 2001 μέχρι το 2007. Η αλλαγή αυτή της διείσδυσης μετρήθηκε ανά τρίμηνο ξεκινώντας από το Σεπτέμβριο του 2001. Τα δεδομένα, που συγκεντρώθηκαν, παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας Τρίμηνο	**Διείσδυση (%)**
1	0,6
2	0,6
3	0,5
4	0,5
5	0,9
6	0,8
7	0,7
8	0,8
9	1,1
10	1,2

11	1
12	1,2
13	1,5
14	1,5
15	1,2
16	1,4
17	1,9
18	1,5
19	1,1
20	1,2
21	1,6
22	1,6
23	1,3



Από το διάγραμμα διασποράς φαίνεται, ότι υπάρχει μία θετική γραμμική σχέση μεταξύ του Y, του ποσοστού αλλαγής της διείσδυσης της ευρυζωνικότητας στην Ευρώπη και του X, του χρόνου. Η διείσδυση της ευρυζωνικότητας στην Ευρώπη από το 2001 έως το 2007 αυξάνει διαρκώς με ορισμένες αυξομειώσεις.

Στη συνέχεια θα υπολογιστεί ο βαθμός αυτής της θετικής σχέσης υπολογίζοντας το συντελεστή συσχέτισης του δείγματος r, (sample

correlation coefficient). Χρησιμοποιώντας την εξίσωση:

$$r=\frac{n\sum XY-\sum X\sum Y}{\sqrt{n\sum X^2-\left(\sum X\right)^2}\sqrt{n\sum Y^2-\left(\sum Y\right)^2}}$$

Ο συντελεστής συσχέτισης του δείγματος (sample correlation coefficient), r είναι 0,829 και δείχνει μία πολύ ισχυρή θετική συσχέτιση μεταξύ του Y και του X, καθώς αυξάνει ο χρόνος.

Το ερώτημα είναι το πόσο αυξάνει το ποσοστό αυτό, καθώς αυξάνει ο χρόνος; Αυτό το ερώτημα μας υποδεικνύει να σχεδιάσουμε-προσαρμόσουμε μία ευθεία γραμμή ανάμεσα στα σημεία δεδομένων (data points), που αναπαριστώνται στο διάγραμμα διασποράς. Αφού έχει σχεδιαστεί αυτή η γραμμή, η κλίση της (slope) θα δείξει τη μέση αύξηση του ποσοστού της διείσδυσης της ευρυζωνικότητας (Y), για κάθε τρίμηνο που περνάει (X).

6.2 ΕΥΘΕΙΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ (ΣΤΑΤΙΣΤΙΚΟ ΜΟΝΤΕΛΟ)

Η ευθεία που προσαρμόζεται καλύτερα σε μία συλλογή X-Y σημείων είναι η γραμμή αυτή, που ελαχιστοποιεί το άθροισμα των τετραγωνικών κάθετων αποστάσεων της γραμμής από τα προς μελέτη σημεία. Αυτή η γραμμή είναι γνωστή ως γραμμή ελαχίστων τετραγώνων ή προσαρμοσμένη γραμμή παλινδρόμησης (fitted regression line) και η εξίσωσή της ονομάζεται προσαρμοσμένη εξίσωση παλινδρόμησης (fitted regression equation).

Η προσαρμοσμένη ευθεία γραμμή είναι της μορφής: $\hat{Y}=b_0+b_1X$. Ο πρώτος όρος, είναι η τομή της ευθείας με τον άξονα Y (Y-intercept) και ο δεύτερος όρος, είναι η κλίση (slope). Υπενθυμίζεται ότι η κλίση αναπαριστά, το πόσο αλλάζει το Y, όταν το X αυξάνει κατά

μία μονάδα. Το άμεσο αντικείμενό μας είναι να καθοριστούν τιμές για τα b_0 και b_1.

Η μέθοδος των ελαχίστων τετραγώνων επιλέγει τις τιμές για τα b_0 και b_1, οι οποίες ελαχιστοποιούν το άθροισμα των τετραγωνικών σφαλμάτων (αποστάσεις).

$$SSE = \sum (Y - \hat{Y})^2 = \sum (Y - b_0 - b_1 X)^2$$

$$b_0 = \frac{\sum Y \sum X^2 - \sum X \sum XY}{n \sum X^2 - (\sum X)^2}$$

$$b_1 = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

Ο συντελεστής κλίσης ελαχίστων τετραγώνων (least squares slope coefficient) σχετίζεται με τον συντελεστή συσχέτισης δείγματος (sample correlation coefficient). Συγκεκριμένα:

$$b_1 = \frac{\sqrt{\sum (Y - \bar{Y})^2}}{\sqrt{\sum (X - \bar{X})^2}} r$$

Έτσι τα b_1 και r είναι ανάλογα το ένα με το άλλο και έχουν το ίδιο πρόσημο.

Οι διαφορές μεταξύ των παρατηρούμενων και προσαρμοσμένων τιμών, τα γνωστά μας κατάλοιπα, είναι οι κάθετες αποστάσεις (θετικές ή αρνητικές) των σημείων δεδομένων από τη γραμμή ελαχίστων τετραγώνων.

Στην παρακάτω σχέση:

Observation = Fit + Residual,

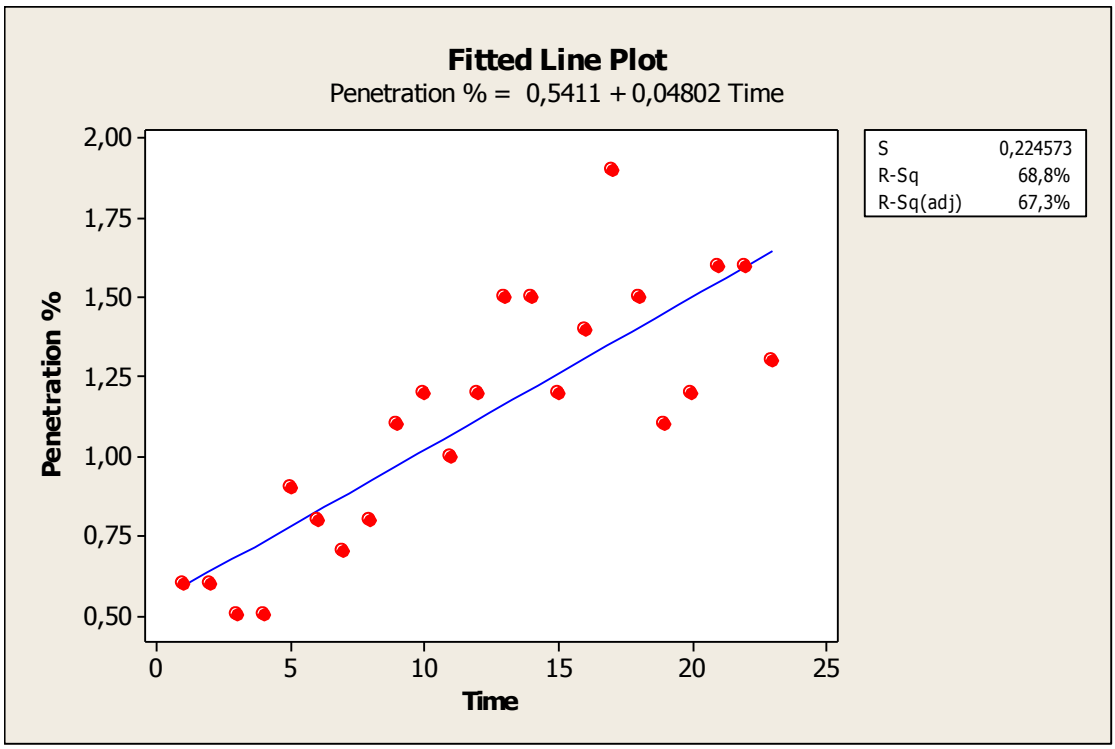
ή με σύμβολα: $Y = \hat{Y} + (Y - \hat{Y})$,

το fit αντιπροσωπεύει το ολικό υπόδειγμα (pattern) των δεδομένων και τα κατάλοιπα, δηλαδή τις αποκλίσεις από το υπόδειγμα αυτό.

Παράδειγμα

Χρησιμοποιώντας τις παραπάνω Εξισώσεις υπολογίστηκαν οι συντελεστές ελαχίστων τετραγώνων για την ευθεία γραμμή, που προσαρμόζεται στα δεδομένα του ποσοστού αλλαγής της διείσδυσης της ευρυζωνικότητας στην Ευρώπη. Έχουμε: $\hat{Y} = 0{,}541 + 0{,}048\,X$.

Συχνά δεν είναι σοφό να προβλέπουμε Y τιμές για κάποιο X, έξω από το εύρος των X, που έχουν συλλεχθεί στα δεδομένα του δείγματος. Μία συνάρτηση παλινδρόμησης πρέπει να ερμηνεύεται σαν μία χρήσιμη προσέγγιση στην πραγματική συμπεριφορά μέσα στην περιοχή, στην οποία υπάρχουν δεδομένα. Η επέκταση πέρα των υπαρχόντων δεδομένων απαιτεί την ισχυρή υπόθεση, ότι η φύση της σχέσης μεταξύ των Y και X δεν αλλάζει.



Στο παράδειγμά μας, το Y αυξάνει κατά μέσο όρο κατά 0,048 (αύξηση του ποσοστού της διείσδυσης της ευρυζωνικότητας κατά 0,048%), όταν το X αυξάνει κατά 1 (όταν δηλαδή περνάει ένα τρίμηνο).

Όταν τα σημεία του διαγράμματος διασποράς (scatter plot) δεν βρίσκονται πάνω στην ευθεία γραμμή, τότε θα πρέπει να εισαχθούν στατιστικές σχέσεις στη μελέτη της σχέσης μεταξύ δύο μεταβλητών,. Θεωρούμε τα δεδομένα του διαγράμματος διασποράς σαν ένα δείγμα από παρατηρήσεις μιας υποκείμενης σχέσης, που ισχύει για τις X-Y τιμές.

Το στατιστικό μοντέλο για την απλή γραμμική παλινδρόμηση υποθέτει, ότι για όλες τις τιμές του X, οι παρατηρούμενες τιμές της εξαρτημένης μεταβλητής Y, είναι κανονικά κατανεμημένες γύρω από ένα μέσο (mean):

$$\mu_y = \beta_0 + \beta_1 X ,$$

ο οποίος εξαρτάται γραμμικά από το X. Αυτό σημαίνει ότι καθώς το X αλλάζει, οι μέσες τιμές των κατανομών των πιθανών Y τιμών βρίσκονται πάνω σε μία ευθεία γραμμή. Αυτή η γραμμή είναι γνωστή ως γραμμή παλινδρόμησης του πληθυσμού (population regression line). Οι παρατηρούμενες τιμές Y θα διαφέρουν γύρω από αυτούς τους μέσους εξαιτίας της επιρροής των μη υπολογισμένων παραγόντων. Το μοντέλο υποθέτει ότι αυτή η απόκλιση, που μετριέται από την τυπική απόκλιση (standard deviation) σ, είναι η ίδια για όλες τις τιμές του X. Τελικά, η απόκλιση (απόσταση) μεταξύ μιας Y τιμής και του μέσου της μ_y είναι γνωστή σαν σφάλμα (error) και αναπαρίσταται με το γράμμα ε.

Στο μοντέλο της απλής γραμμικής παλινδρόμησης η εξαρτημένη μεταβλητή Y είναι το άθροισμα του μέσου της και μίας τυχαίας απόκλισης (ε) από το μέσο. Οι αποκλίσεις (ε) αντιπροσωπεύουν μεταβολή στο Y λόγω άλλων μη παρατηρημένων παραγόντων που εμποδίζουν τις X-Y τιμές από το να βρίσκονται ακριβώς πάνω στην ευθεία γραμμή του διαγράμματος διασποράς.

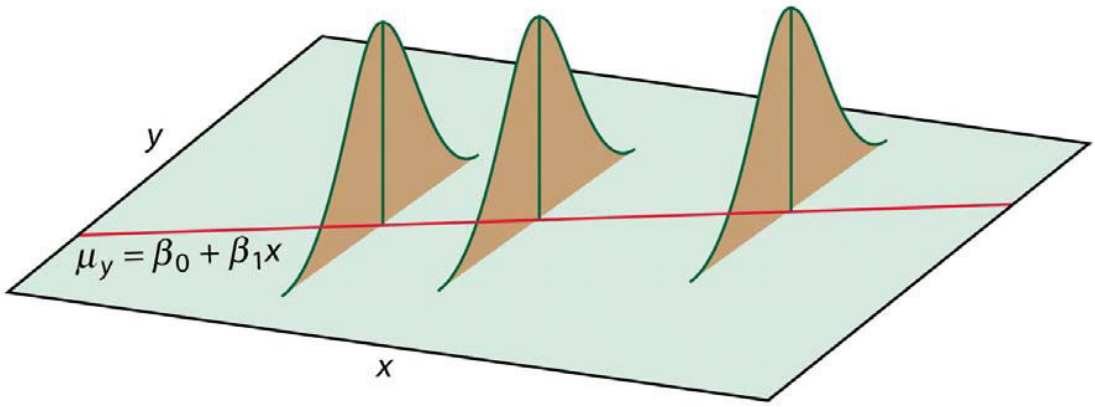
Στατιστικό Μοντέλο για Παλινδρόμηση Ευθείας Γραμμής

Η εξαρτημένη μεταβλητή Y σχετίζεται με την ανεξάρτητη μεταβλητή X, με τη σχέση:

$$Y=\beta_0+\beta_1 X+\varepsilon$$

Το $\beta_0+\beta_1 X$ είναι η μέση τιμή για ένα δεδομένο X. Οι αποκλίσεις (ε) υποθέτουμε, ότι είναι ανεξάρτητες, τυχαίες και κανονικά κατανεμημένες, με μέσο 0 $(E(\varepsilon)=0)$ και τυπική απόκλιση σ $(E(\varepsilon^2)=\sigma^2)$, δηλαδή $\varepsilon=N(0,\sigma^2)$. Οι άγνωστες σταθερές είναι οι β_0, β_1 και σ.

Το στατιστικό μοντέλο για την απλή γραμμική παλινδρόμηση αναπαρίσταται σχηματικά στο παρακάτω Γράφημα.



Η προσαρμοσμένη (στο δείγμα) γραμμή παλινδρόμησης, $\hat{Y}=b_0+b_1 X$ μπορεί να θεωρηθεί σαν μία εκτίμηση της γραμμής παλινδρόμησης του πληθυσμού, $\mu_y=\beta_0+\beta_1 X$ και τα κατάλοιπα, $e=Y-\hat{Y}$, μπορούν να θεωρηθούν σαν εκτιμήσεις των συντελεστών σφάλματος, ε. Αυτό υποδεικνύει την ακόλουθη αντιστοιχία:

$$Population: Y=\beta_0+\beta_1 X+\varepsilon$$

$$Sample: Y=b_0+b_1 X+e$$

με $E(b_0)=\beta_0$ *και* $E(b_1)=\beta_1$

6.3 ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΕΚΤΙΜΗΣΗΣ

Έχοντας υπολογίσει την προσαρμοσμένη ευθεία γραμμή, ενδιαφερόμαστε να υπολογίσουμε το βαθμό, κατά τον οποίο τα σημεία του δείγματος δεδομένων κατανέμονται γύρω από την προσαρμοσμένη συνάρτηση παλινδρόμησης (fitted regression function). Είναι δυνατόν να ορισθεί ένα μέτρο για τη διασπορά (dispersion) ανάλογο της τυπικής απόκλισης του δείγματος. Αυτό το μέτρο ονομάζεται **τυπικό σφάλμα εκτίμησης** (standard error of the estimate) και υπολογίζει την κατανομή των σημείων δεδομένων γύρω από την προσαρμοσμένη γραμμή στην κατεύθυνση Y. Το τυπικό σφάλμα εκτίμησης συμβολίζεται με $s_{y.x}$ και δίνεται από τη σχέση:

$$s_{y\cdot x}=\sqrt{\frac{\sum (Y-\hat{Y})^2}{n-2}}=\sqrt{\frac{\sum Y^2-b_0\sum Y-b_1\sum XY}{n-2}}$$

Το τυπικό σφάλμα εκτίμησης υπολογίζει την ποσότητα, κατά την οποία οι πραγματικές τιμές του Y διαφέρουν από τις εκτιμημένες, ή $\hat{Y}$, τιμές. Για σχετικά μεγάλα δείγματα, περιμένουμε περίπου το 67% των διαφορών $Y-\hat{Y}$ να είναι σε ένα διάστημα $\pm s_{y\cdot x}$ από το 0 και περίπου το 95% αυτών των διαφορών να είναι σε ένα διάστημα 2 x $(\pm s_{y\cdot x})$ από το 0.

Το τυπικό σφάλμα εκτίμησης είναι παρόμοιο με την τυπική απόκλιση του δείγματος. Μπορεί να χρησιμοποιηθεί για να υπολογίσει την τυπική απόκλιση του πληθυσμού (population standard deviation). Στην πράξη, το $s_{y\cdot x}$ μας δίνει την τυπική απόκλιση σ, του σφάλματος ε, στο στατιστικό μοντέλο της απλής γραμμικής παλινδρόμησης. Ισοδύναμα, το $s_{y\cdot x}$ μας δίνει την κοινή τυπική απόκλιση σ, της κανονικής κατανομής των Y τιμών γύρω από τη γραμμή παλινδρόμησης του πληθυσμού (population regression line), $\mu_y=\beta_0+\beta_1 X$ για κάθε X.

Το $s_{y\cdot x}$ είναι ένας άριστος εκτιμητής του σ, διότι όπως αποδεικνύεται η αναμενόμενη τιμή του $s^2_{y\cdot x}$ είναι το σ^2, η διασπορά των σφαλμάτων του πληθυσμού. Δηλαδή:

$$E\left(s^2_{y\cdot x}\right)=\sigma^2=\frac{\sum\left(Y-\hat{Y}\right)^2}{n}.$$

Μία ανάλυση παλινδρόμησης με ένα μικρό τυπικό σφάλμα εκτίμησης σημαίνει, ότι όλα τα σημεία δεδομένων βρίσκονται πολύ κοντά στην προσαρμοσμένη γραμμή παλινδρόμησης. Αν το τυπικό σφάλμα εκτίμησης είναι μεγάλο, τα σημεία δεδομένων είναι ευρέως διασκορπισμένα γύρω από την προσαρμοσμένη γραμμή.

Για το παράδειγμά μας έχουμε:

$$s_{y\cdot x}=\sqrt{\frac{32{,}11-0{,}541*25{,}7-0{,}048*357}{23-2}}=0{,}226$$

6.4 ΠΡΟΒΛΕΠΟΝΤΑΣ ΤΟ Y

Στη συνέχεια, η προσαρμοσμένη γραμμή παλινδρόμησης μπορεί να χρησιμοποιηθεί για να εκτιμήσει την τιμή του Y για μία δεδομένη τιμή του X. Για να βρούμε ένα σημείο πρόβλεψης (point forecast), ή για να προβλέψουμε το Y για μία δεδομένη τιμή του X, απλά υπολογίζουμε την εκτιμώμενη συνάρτηση παλινδρόμησης στο X.

Παράδειγμα

Ας υποθέσουμε ότι θέλουμε να προβλέψουμε το ποσοστό της αλλαγής της διείσδυσης της ευρυζωνικότητας από το δεύτερο στο τρίτο τρίμηνο του 2008, δηλαδή X=30. Από την Εξίσωση:

$$\hat{Y}=0{,}541+0{,}048\,X$$

η πρόβλεψη είναι:

$$\hat{Y} = 0{,}541 + 0{,}048\,X = 0{,}541 + 0{,}048 * 30 = 1{,}981$$

ή ότι η διείσδυση της ευρυζωνικότητας στο τρίτο τρίμηνο του 2008 θα αυξηθεί κατά 1,981%. Να σημειωθεί ότι αυτή η πρόβλεψη είναι μία τιμή του $\hat{Y}$, δηλαδή η πρόβλεψη είναι η Y συντεταγμένη του σημείου με X=30 στην προσαρμοσμένη γραμμή παλινδρόμησης.

Είναι φυσικό ότι οι πραγματικές τιμές του Y για τα X είναι απίθανο να βρίσκονται ακριβώς πάνω στην γραμμή παλινδρόμησης. Στην πραγματικότητα θα απλώνονται γύρω από τη γραμμή, όπως υπολογίστηκε από το $s_{y \cdot x}$. Επιπλέον, η (προσαρμοσμένη) γραμμή παλινδρόμησης του δείγματος είναι μία εκτίμηση της γραμμής παλινδρόμησης του πληθυσμού (population regression line), που βασίζεται στα 23 σημεία δεδομένων. Άλλα 23 τυχαία ζεύγη τιμών θα έδειχναν διαφορετική προσαρμοσμένη γραμμή παλινδρόμησης και για ν δείγματα ν διαφορετικές γραμμές παλινδρόμησης, όπως και στην περίπτωση, στην οποία πολλά διαφορετικά δείγματα από τον ίδιο πληθυσμό έχουν διαφορετικούς μέσους δείγματος.

Υπάρχουν δύο πηγές αβεβαιότητας, που σχετίζονται με ένα σημείο πρόβλεψης, το οποίο παράγεται από μία προσαρμοσμένη εξίσωση παλινδρόμησης:

1) Αβεβαιότητα λόγω της διασποράς των σημείων δεδομένων γύρω από τη γραμμή παλινδρόμησης του δείγματος.

2) Αβεβαιότητα λόγω της διασποράς της γραμμής παλινδρόμησης του δείγματος γύρω από τη γραμμή παλινδρόμησης του πληθυσμού.

Μπορεί να κατασκευαστεί ένα διάστημα πρόβλεψης (interval forecast) του Y, λαμβάνοντας υπόψη τις παραπάνω δύο αβεβαιότητες.

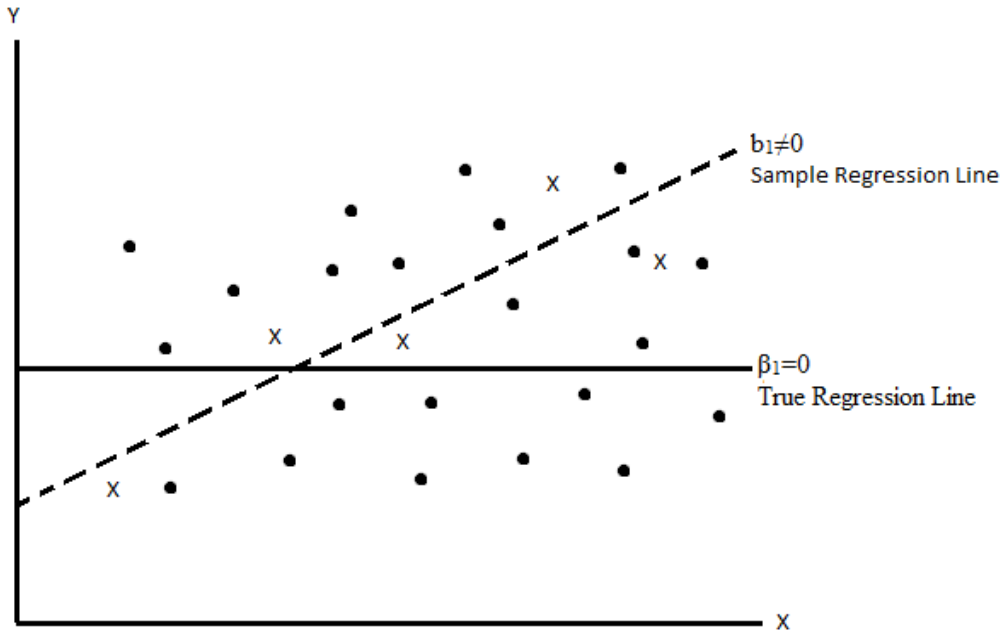
Το τυπικό σφάλμα της πρόβλεψης (standard error of the

forecast) s_f ***υπολογίζει τη μεταβλητότητα της προβλεπόμενης τιμής Y, από την γραμμική σχέση, γύρω από την πραγματική τιμή του Y για μία τιμή X.***

Το τυπικό σφάλμα της πρόβλεψης είναι:

$$s_f = \sqrt{s_{y\cdot x}^2 + s_{y\cdot x}^2 \left(\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum (X-\bar{X})^2} \right)} \quad \text{ή}$$

$$s_f = s_{y\cdot x} \sqrt{1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{\sum (X-\bar{X})^2}}$$



Ο πρώτος όρος κάτω από την ρίζα, $s_{y\cdot x}^2$, υπολογίζει τη διασπορά των σημείων δεδομένων γύρω από τη γραμμή παλινδρόμησης του δείγματος **(πρώτη πηγή αβεβαιότητας)**. Ο δεύτερος όρος κάτω από τη ρίζα υπολογίζει τη διασπορά της γραμμής παλινδρόμησης του δείγματος γύρω από της γραμμή παλινδρόμησης του πληθυσμού **(δεύτερη πηγή αβεβαιότητας)**. Το τυπικό σφάλμα της πρόβλεψης εξαρτάται από το X, την τιμή του X για την οποία επιθυμούμε μία πρόβλεψη του Y. Επίσης,

το s_f έχει ελάχιστη τιμή όταν X=$\overline{X}$, δεδομένου ότι ο αριθμητής στον τρίτο όρο κάτω από τη ρίζα της Εξίσωσης (κάτω εξίσωση) θα είναι $(\overline{X}-\overline{X})^2 = 0$. Αν κρατήσουμε όλα τα άλλα σταθερά, όσο πιο μακριά είναι το X από το $\overline{X}$, τόσο πιο μεγάλο είναι το τυπικό σφάλμα της πρόβλεψης.

Αν το στατιστικό μοντέλο της απλής γραμμικής παλινδρόμησης είναι κατάλληλο, ένα διάστημα πρόβλεψης για το Y δίνεται από τη σχέση:

$$\hat{Y} \pm ts_f ,$$

όπου το t είναι ένα εκατοστιαίο σημείο της t κατανομής με df=n-2. Αν το μέγεθος του δείγματος είναι μεγάλο (n-2≥30), το t εκατοστιαίο σημείο μπορεί να αντικατασταθεί με το αντίστοιχο εκατοστιαίο σημείο Z της τυπικής κανονικής κατανομής. Για παράδειγμα, ένα 95% διάστημα πρόβλεψης μεγάλου δείγματος για το Y είναι:

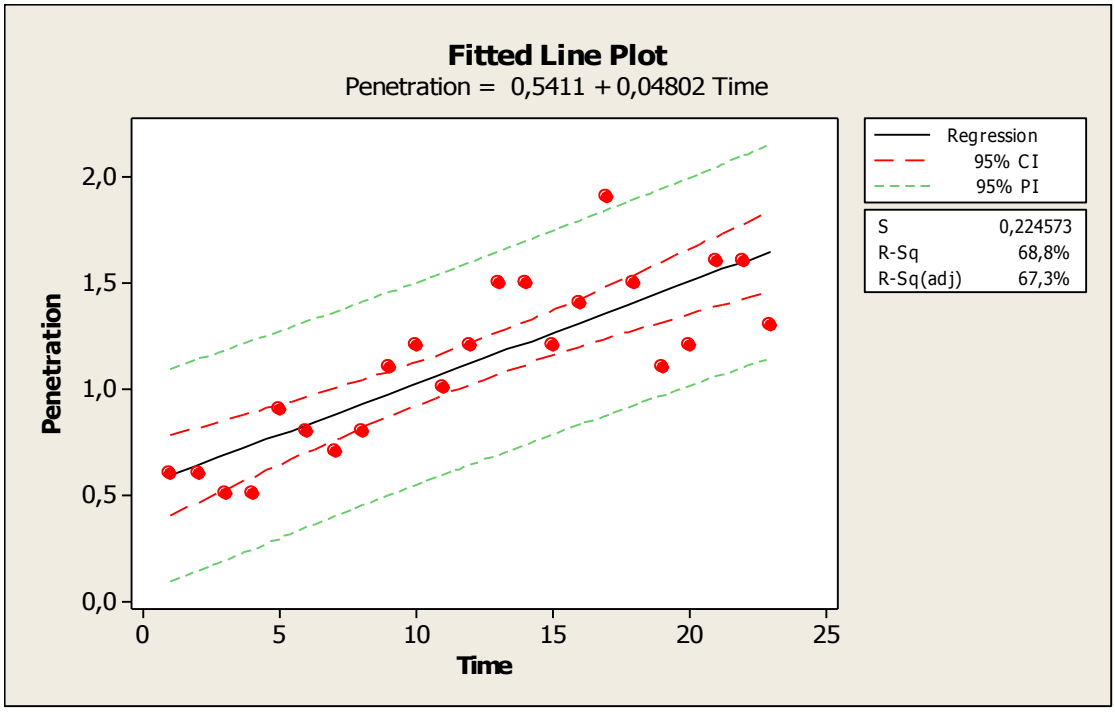
$$\hat{Y} \pm 1.96\, s_f$$

Παράδειγμα

Γραφικά, ένα διάστημα πρόβλεψης 95% για το Y σε διάφορες τιμές του X θα έμοιαζε, όπως το παρακάτω Γράφημα.

X	$(X-X)^2$
1	121
2	100
3	81
4	64
5	49
6	36
7	25
8	16
9	9
10	4
11	1
12	0
13	1

14	4
15	9
16	16
17	25
18	36
19	49
20	64
21	81
22	100
23	121
$\sum X = 276$	$\sum (X - \bar{X})^2 = 1012$



Χρησιμοποιώντας τα αποτελέσματα του πίνακα για την διείσδυση της ευρυζωνικότητας, το τυπικό σφάλμα της πρόβλεψης για X=30 και $\overline{X}$=12 είναι:

$$s_f = 0{,}226\sqrt{1+\frac{1}{23}+\frac{(30-12)^2}{1012}} = 0{,}226\sqrt{1{,}36} = 0{,}31$$

Για X=30 το $\hat{Y} = 1{,}981$ και το διάστημα πρόβλεψης είναι:

$$1{,}981 \pm 2{,}08 \cdot 0{,}31 = 1{,}981 \pm 0{,}645$$

ή το ποσοστό αύξησης της διείσδυσης της ευρυζωνικότητας για X=30 θα

είναι από 1,336 έως 2,626%, δηλαδή από το δεύτερο στο τρίτο τρίμηνο του 2008. Εδώ, $2,08=t_{0,975}$ είναι το άνω 2,5% σημείο της t κατανομής με df=21.

Γενικά, είναι επικίνδυνο να χρησιμοποιείται η προσαρμοσμένη συνάρτηση παλινδρόμησης για να προβλεφθούν Y τιμές πέρα από το εύρος των διαθέσιμων δεδομένων. Όταν προβλέπονται τιμές του Y για τέτοιες τιμές του X, αυτές είναι ιδιαίτερα ύποπτες. Αυτό ισχύει γιατί δεν γνωρίζουμε αν πέρα του εύρους των δεδομένων μας, το μοντέλο ευθείας γραμμής ισχύει ακόμα. Μπορεί να έχουμε καλούς λόγους για να κάνουμε μία τέτοια υπόθεση, αλλά δεν έχουμε σαφείς ενδείξεις για να το υποστηρίξουμε.

Είναι χρήσιμο να τελειώσουμε αυτή την ενότητα ανακεφαλαιώνοντας τις υποθέσεις για το στατιστικό μοντέλο γραμμικής παλινδρόμησης.

1) Για μία δεδομένη τιμή του X, ο πληθυσμός των Y τιμών είναι κανονικά κατανεμημένος γύρω από τη γραμμή παλινδρόμησης του πληθυσμού. Πρακτικά, λογικά ακριβή αποτελέσματα βρίσκονται, όσο οι τιμές του Y είναι περίπου κανονικά κατανεμημένες.
2) Η διασπορά των σημείων δεδομένων του πληθυσμού γύρω από τη γραμμή παλινδρόμησης του πληθυσμού παραμένει σταθερή οπουδήποτε πάνω στη γραμμή. Αυτό ισχύει γιατί η διακύμανση πληθυσμού δεν μεγαλώνει ή μικραίνει καθώς οι X τιμές των σημείων δεδομένων αυξάνουν. Μία παραβίαση αυτής της υπόθεσης καλείται ετεροσκεδαστικότητα (heteroscedasticity).
3) Τα σφάλματα (ε) είναι ανεξάρτητα μεταξύ τους. Αυτή η υπόθεση υποθέτει ένα τυχαίο δείγμα X-Y σημείων δεδομένων. Όταν τα X-Y σημεία δεδομένων καταγράφονται πάνω στο χρόνο, αυτή η

υπόθεση συχνά παραβιάζεται. Αντίθετα από το να είναι ανεξάρτητες, οι διαδοχικές παρατηρήσεις είναι σειριακά συσχετισμένες (serially correlated).

4) Υπάρχει μία γραμμική σχέση μεταξύ των X και Y στον πληθυσμό. Υπάρχουν επεκτάσεις της απλής γραμμικής παλινδρόμησης για να μελετήσουμε X-Y σχέσεις, που δεν είναι γραμμικές.

6.5 ΑΝΑΛΥΣΗ ΤΗΣ ΔΙΑΣΠΟΡΑΣ

Από τις παρατηρούμενες τιμές έχουμε:

$$Y = \hat{Y} + (Y - \hat{Y})$$

ή

$$Y = b_0 + b_1 X + (Y - b_0 - b_1 X)$$

όπου Y είναι η παρατηρούμενη τιμή, ο όρος εκτός παρένθεσης ερμηνεύεται από τη γραμμική σχέση και ο όρος εντός παρένθεσης είναι το κατάλοιπο (residual), ή η απόκλιση από τη γραμμική σχέση.

Σε μία ιδανική κατάσταση, στην οποία όλα τα σημεία βρίσκονται πάνω στην προσαρμοσμένη γραμμή, τα κατάλοιπα είναι όλα μηδέν, και οι τιμές Y είναι πλήρως υπολογισμένες, ή ερμηνεύονται από τη γραμμική σχέση με το X.

Αφαιρώντας και από τις δύο πλευρές της προηγούμενης έκφρασης το Y έχουμε:

$$(Y - \bar{Y}) = (\hat{Y} - \bar{Y}) + (Y - \hat{Y})$$

Μετά από λίγη άλγεβρα (υψώνοντας πρώτα τη σχέση στο τετράγωνο και μετά αθροίζοντας) μπορούμε να αποδείξουμε εύκολα, ότι ισχύει:

$$\sum (Y-\bar{Y})^2=\sum (\hat{Y}-\bar{Y})^2+\sum (Y-\hat{Y})^2$$

ή SST = SSR + SSE ή

Ολική μεταβλητότητα του Y = Εξηγήσιμη μεταβλητότητα από την παλινδρόμηση (γραμμική σχέση) + τα κατάλοιπα.

όπου:

$$SST=\sum (Y-\bar{Y})^2$$

$$SSR=\sum (\hat{Y}-\bar{Y})^2$$

$$SSE=\sum (Y-\hat{Y})^2$$

SS = Sum of squares και T, R, E σημαίνουν total, regression, error.

Οι βαθμοί ελευθερίας των αθροισμάτων αυτών είναι:

df(SST) = n – 1

df(SSR) = 1

df(SSE) = n – 2

ενώ η σχέση των βαθμών ελευθερίας είναι:

n – 1 = 1 + (n – 2)

Στην περίπτωση που δεν υπάρχει γραμμική σχέση του Y από το X, το εκτιμώμενο Y θα είναι ο μέσος όρος και η διασπορά του Y θα περιγράφεται από την δειγματική διασπορά:

$$s_y^2=\frac{\sum (Y-\bar{Y})^2}{n-1}$$

Διαιρώντας με n-1 και τροποποιώντας τη σχέση των αθροισμάτων των τετραγώνων των διαφορών παίρνουμε:

$$\frac{\sum (Y-\bar{Y})^2}{n-1}=\frac{\sum (\hat{Y}-\bar{Y})^2}{n-1}+\frac{(n-2)\sum (Y-\hat{Y})^2}{(n-1)(n-2)}$$ ή

$$s_y^2 = \frac{SSR}{n-1} + \frac{(n-2)}{(n-1)} s_{y\cdot x}^2$$

Τα αθροίσματα τετραγώνων, που σχετίζονται με την ανάλυση της μεταβλητότητας του Y και τους αντίστοιχους βαθμούς ελευθερίας, φαίνονται στον Πίνακα, ο οποίος είναι γνωστός ως ANOVA πίνακας, ή πίνακας ανάλυσης της διασποράς (ANalysis Of VAriance table).

Η τελευταία στήλη στον πίνακα ANOVA είναι η στήλη μέσου τετραγώνου (mean square). Η παλινδρόμηση μέσου τετραγώνου, MSR, είναι το άθροισμα τετραγώνων της παλινδρόμησης διαιρεμένο με το βαθμό ελευθερίας του. Παρόμοια, το σφάλμα μέσου τετραγώνου, MSE, είναι το σφάλμα του αθροίσματος τετραγώνου διαιρεμένο με το βαθμό ελευθερίας του.

Source	Sum of Squares	Df	Mean Square
Regression	SSR	1	MSR=SSR/1
Error	SSE	n – 2	MSE=SSE/(n-2)
Total	SST	n – 1	

Το μέσο τετραγωνικό σφάλμα είναι:

$$MSE = \frac{SSE}{n-2} = \frac{\sum (Y-\hat{Y})^2}{n-2} = s_{y\cdot x}^2$$

Παράδειγμα

Στην ανάλυση για το ποσοστό αύξησης της διείσδυσης της ευρυζωνικότητας ανά τρίμηνο συγκεντρώθηκαν δεδομένα από το 2001 έως το 2007. Αν δεν υπήρχαν άλλες πληροφορίες διαθέσιμες θα χρησιμοποιούσαμε το $\bar{Y} = 1{,}18,$ σαν πρόβλεψη της αύξησης της διείσδυσης της ευρυζωνικότητας στην Ευρώπη ανά τρίμηνο. Τα σφάλματα, ή κατάλοιπα, που σχετίζονται με αυτή την πρόβλεψη είναι $Y - \bar{Y}$ και το άθροισμα των τετραγωνικών σφαλμάτων είναι $\sum (Y-\bar{Y})^2.$ Η τελευταία αυτή ποσότητα είναι το SST, το συνολικό

άθροισμα των τετραγώνων, όπως αναφέρθηκε στα προηγούμενα. Επομένως, το SST υπολογίζει τη μεταβλητότητα του Y γύρω από τη μέση τιμή του δείγματος, που χρησιμοποιεί μόνο τις Y τιμές στους υπολογισμούς του. Διαιρώντας το με n-1 θα έχουμε τη δειγματική διασπορά.

Έχουμε επίσης πληροφόρηση σχετικά με μία μεταβλητή X, το τρίμηνο, που σχετίζεται με το Y, το ποσοστό της αλλαγής της ευρυζωνικότητας στην Ευρώπη. (Να θυμηθούμε ότι r=0,829). Με αυτή την επιπλέον μεταβλητή, αναμένουμε να είμαστε σε θέση να εξηγήσουμε κάποια από τη διασπορά (διαφορές) στις Y τιμές πέρα από αυτή που εξηγείται από το $\bar{Y}$.

Στο Παράδειγμα μας, η γραμμή προσαρμόζεται στο διάγραμμα διασποράς των X-Y παρατηρήσεων και έχει την εξίσωση:

$$\hat{Y} = 0.541 + 0.048\, X .$$

Μία σύγκριση των τετραγωνικών σφαλμάτων δείχνει, ότι η χρήση του $\hat{Y}$ σαν πρόβλεψη του Y οδηγεί σε γενικές γραμμές σε μικρότερα κατάλοιπα (σε απόλυτη τιμή) και σε ένα αρκετά μικρότερο κατάλοιπο (σφάλμα) αθροίσματος τετραγώνων, από το να χρησιμοποιηθεί το $\bar{Y}$ για πρόβλεψη. Η χρήση της σχετικής μεταβλητής X μειώνει τα σφάλματα της πρόβλεψης. Αυτό σημαίνει ότι η γνώση του X βοηθάει στην περαιτέρω ερμηνεία των διαφορών στα Y. Πόσο βοηθά το X; Η ανάλυση της μεταβλητότητας δίνει μία απάντηση σε αυτή την ερώτηση.

Από τους πίνακες έχουμε:

$$SST = \sum (Y - \bar{Y})^2 = 3{,}3932$$

$$SSE = \sum (Y - \hat{Y})^2 = 1{,}059$$

$$SSR = \sum (\hat{Y} - \bar{Y})^2 = SST - SSE = 3{,}3932 - 1{,}059 = 2{,}3342$$

Η ανάλυση της μεταβλητότητας είναι:

SST	**=**	**SSR**	**+**	**SSE**
3,3932	**=**	**2,3342**	**+**	**1,059**
Συνολική		**Εξηγήσιμη**		**Μη εξηγήσιμη**
Μεταβολή		**Μεταβολή**		**Μεταβολή**

Από τη μεταβλητότητα που παραμένει μετά την πρόβλεψη του Y με το Y, ένα ποσοστό αυτής:

$$\frac{SSR}{SST}=\frac{2,3342}{3,3932}=0,69$$

έχει ερμηνευθεί από τη σχέση του Y με το X. Ένα ποσοστό, 1-0,69=0,31, της μεταβλητότητας του Y γύρω από το $\bar{Y}$ παραμένει μη ερμηνεύσιμο. Από αυτή την οπτική γωνία, η γνώση της σχετικής μεταβλητής X οδηγεί σε καλύτερες προβλέψεις του Y, από αυτές που μπορούν να γίνουν από το $\bar{Y}$, μία ποσότητα που δεν εξαρτάται από το X.

Η ανάλυση της μεταβλητότητας για τα δεδομένα μπορεί να καθοριστεί από μία ανάλυση του πίνακα διασποράς, όπως δείχνει το επόμενο παράδειγμα.

Παράδειγμα

Κατασκευάζουμε ένα πίνακα ANOVA για τα δεδομένα χρησιμοποιώντας τη δομή του προηγούμενου Πίνακα και τους υπολογισμούς των αθροισμάτων των τετραγώνων από το Παράδειγμα, όπως φαίνεται στον Πίνακα.

Η ανάλυση της μεταβλητότητας για τα δεδομένα μας φαίνεται καθαρά στη στήλη του αθροίσματος των τετραγώνων. Να σημειωθεί, μέσα στο σφάλμα στρογγυλεύσεως, $MSE=0,05=(0,223)^2=s^2_{y\cdot x}$.

Πηγή	**Άθροισμα τετραγώνων**	**Df**	**Τετράγωνο Μέσου**

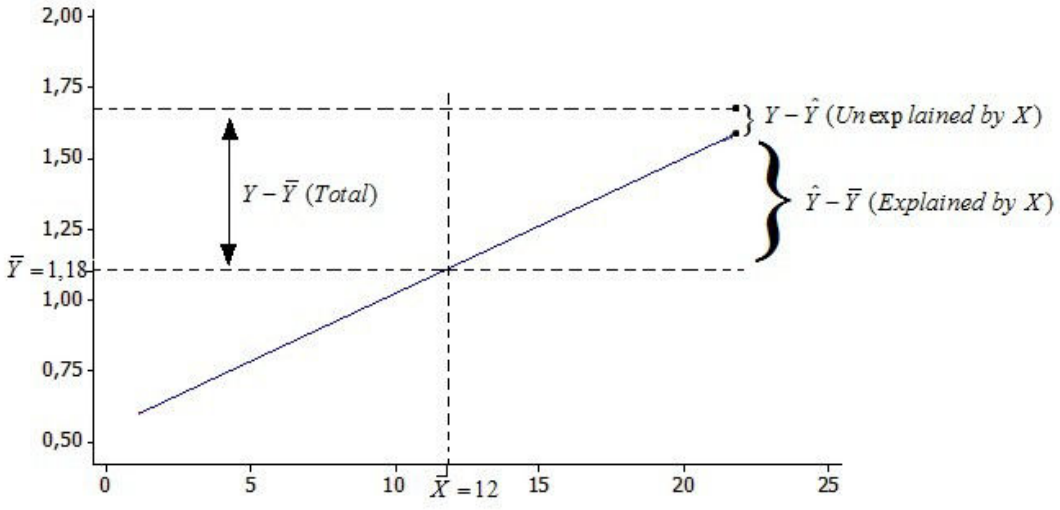
Παλινδρόμηση	2,33	1	2,33
Σφάλμα	1,06	21	0,05
Σύνολο	**3,39**	**22**	

6.6 ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ

Η ταυτότητα,

$$(Y-\bar{Y})=(\hat{Y}-\bar{Y})+(Y-\hat{Y}),$$

η οποία οδηγεί στην ανάλυση της διασποράς και που δίνεται από τα αθροίσματα τετραγώνων, φαίνεται γραφικά για την προσαρμοσμένη γραμμή παλινδρόμησης και ένα υποθετικό σημείο δεδομένων στο παρακάτω διάγραμμα.



Αν το Y δεν εξαρτάται από το X, θα περιμέναμε τα Y να είναι κοντά στο $\overline{Y}$ και η απόκλιση $Y-\bar{Y}$ απλά να αντανακλά τυχαία διασπορά. Ωστόσο, το Y εξαρτάται από το X με έναν τρόπο που προτείνεται από την προσαρμοσμένη συνάρτηση παλινδρόμησης. Στο γράφημα, η παρατηρημένη τιμή X είναι μεγαλύτερη από το $\overline{X}$, και είναι γνωστό ότι το X και το Y έχουν μία αρκετά ισχυρή θετική συσχέτιση

μεταξύ τους (r=0.829). Από τη συνολική κάθετη απόσταση, $Y-\bar{Y}$ η ποσότητα $\hat{Y}-\bar{Y}$ ερμηνεύεται από την κίνηση στο X, ενώ η εναπομείνασα κάθετη απόσταση, $Y-\hat{Y}$, δεν ερμηνεύεται από την κίνηση στο X.

Όπως επισημαίνεται και στην προηγούμενη ενότητα, το SST υπολογίζει την συνολική διασπορά γύρω από το $\bar{Y}$, και εκείνο το μέρος του συνόλου που ερμηνεύεται από την κίνηση στο X είναι το SSR. Η υπόλοιπη ή μη ερμηνευμένη διασπορά είναι το SSE. Ο λόγος της ερμηνευμένης προς τη συνολική μεταβολή καλείται ***συντελεστής προσδιορισμού*** του δείγματος και συμβολίζεται με r^2.

$$r^2=\frac{Explained\ Variation}{Total\ Variation}=\frac{SSR}{SST}=\frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2}$$

$$r^2=1-\frac{Unexplained\ Variation}{Total\ Variation}=1-\frac{SSE}{SST}=1-\frac{\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2}$$

Ο συντελεστής προσδιορισμού υπολογίζει το ποσοστό της μεταβλητότητας στο Y, το οποίο μπορεί να ερμηνευθεί μέσω της γνώσης της μεταβλητότητας (διαφορές) στην ανεξάρτητη μεταβλητή X.

Επίσης υπάρχει και ο προσαρμοσμένος συντελεστής προσδιορισμού $\bar{r}^2$, προσαρμοσμένος με τους βαθμούς ελευθερίας:

$$\bar{r}^2=r^2(adj)=1-\frac{SSE/(n-2)}{SST/(n-1)}=1-\frac{\sum(Y-\hat{Y})^2/(n-2)}{\sum(Y-\bar{Y})^2/(n-1)}=1-\frac{s_{y\cdot x}^2}{s_y^2}$$

Παράδειγμα

Ο συντελεστής προσδιορισμού για τα δεδομένα μας υπολογίστηκε στο προηγούμενο Παράδειγμα, αν και τότε δεν είχε ονομαστεί έτσι. Ο συντελεστής προσδιορισμού είναι επίσης άμεσα διαθέσιμος από τον πίνακα ANOVA. Από το παράδειγμά μας έχουμε:

SST = 3,3932

SSR = 2,3342

SSE = 1,059

και

$$r^2 = \frac{2,3342}{3,3942} = 0,69$$

Εναλλακτικά, το r^2 μπορεί να υπολογιστεί ως:

$$r^2 = 1 - \frac{1,059}{3,3942} = 1 - 0,31 = 0,69$$

Σχεδόν το 70% της μεταβλητότητας στο ποσοστό της διείσδυσης της ευρυζωνικότητας (Y) μπορεί να ερμηνευθεί από τη διαφορά του χρόνου (τριμήνου) (X). Σχεδόν το 30% της μεταβλητότητας αυτής δεν μπορεί να ερμηνευθεί από το χρόνο. Αυτό το κομμάτι της μεταβλητότητας πρέπει να ερμηνευθεί από παράγοντες, οι οποίοι δεν έχουν ληφθεί υπόψη σε αυτή την ανάλυση παλινδρόμησης (για παράδειγμα, διαφήμιση, τιμή κλπ).

Για απλή γραμμική παλινδρόμηση, το r^2, είναι το τετράγωνο του r, δηλαδή του συντελεστή συσχέτισης.

Coefficient of determination = (Coefficient of correlation)^2 ή $r^2 = (r)^2$

Έτσι, για τα δεδομένα μας και αγνοώντας το σφάλμα στρογγύλευσης,

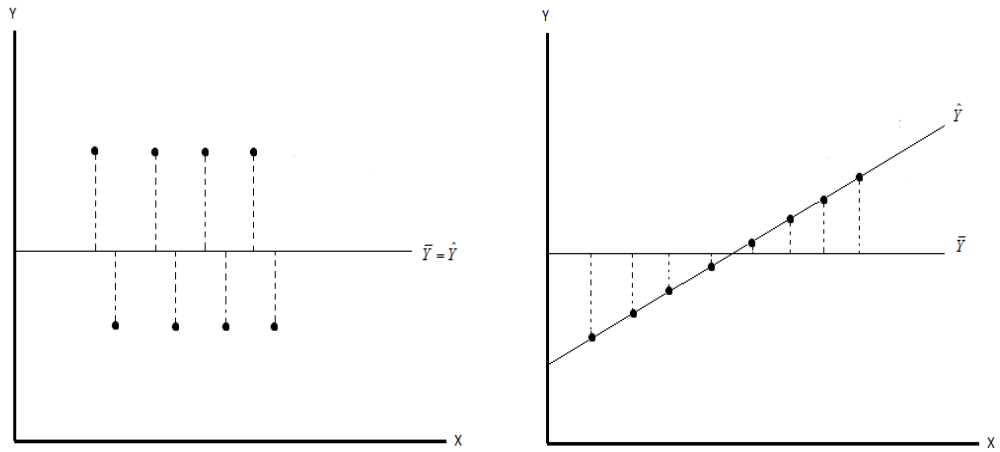
$0,69 = (0,829)^2$.

Γιατί είναι απαραίτητο να αναγνωρίσουμε και το r και το r^2 σε μία ανάλυση παλινδρόμησης; Η απάντηση είναι, ότι περιέχουν διαφορετικές πληροφορίες.

Ο συντελεστής συσχέτισης αποκαλύπτει τη κλίση και την

κατεύθυνση της γραμμικής σχέσης. Στην περίπτωση των δεδομένων μας υπάρχει θετική σχέση (r=0,829). Σε άλλες περιπτώσεις, μπορεί να υποδεικνύεται μία αρνητική σχέση από την τιμή του r. Όπως θα δούμε και στο επόμενο κεφάλαιο, είναι χρήσιμο να αναγνωρίσουμε τη φύση της σχέσης που υπάρχει μεταξύ ενός ζεύγους μεταβλητών, όταν έχουμε να κάνουμε με μία μεγάλη συλλογή μεταβλητών. Να σημειωθεί ότι όταν ο συντελεστής συσχέτισης υψωθεί στο τετράγωνο, η τιμή είναι πάντα θετική, έτσι η φύση της σχέσης χάνεται.

Ο συντελεστής προσδιορισμού, r^2, υπολογίζει τη σχέση μεταξύ του Y και του X με έναν διαφορετικό τρόπο από ότι κάνει το r. Η τιμή του r^2 υπολογίζει το ποσοστό της μεταβλητότητας στο Y, το οποίο ερμηνεύεται από τις διαφορές στο X. Αυτή η χρήσιμη ερμηνεία μπορεί να γενικευτεί σε σχέσεις μεταξύ του Y και περισσότερων του ενός X.



Το παραπάνω Γράφημα απεικονίζει τις δύο ακραίες περιπτώσεις για το r^2, $r^2=0$ και $r^2=1$. Στην πρώτη περίπτωση, δεν ερμηνεύεται τίποτα από την μεταβλητότητα του Y από το X: Το διάγραμμα διασποράς (scatter diagram) δείχνει, ότι δεν υπάρχει γραμμική σχέση μεταξύ του X

και του Υ. Όταν το $r^2=1$, όλη η μεταβλητότητα του Υ μπορεί να ερμηνευθεί από το Χ, όταν είναι γνωστό, ότι τα σημεία δεδομένων του δείγματος βρίσκονται όλα πάντα στην προσαρμοσμένη γραμμή παλινδρόμησης.

6.7 ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ

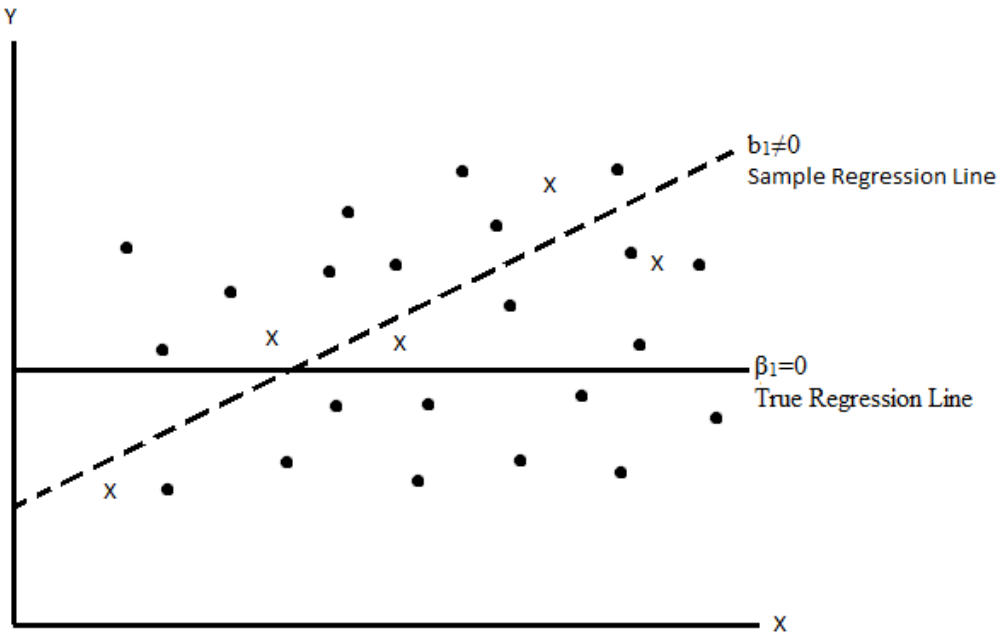
Η προσαρμοσμένη γραμμή παλινδρόμησης παράγεται από ένα δείγμα ζευγών Χ-Υ τιμών. Το στατιστικό μοντέλο για την απλή γραμμική παλινδρόμηση προϋποθέτει, ότι η γραμμική σχέση μεταξύ του Υ και του Χ ισχύει για όλες τις επιλογές των Χ-Υ ζευγών του πληθυσμού. Αυτό γίνεται, γιατί υπάρχει – αν υπάρχει – μία αληθής σχέση μεταξύ του Χ και του Υ της μορφής $\mu_y=\beta_0+\beta_1 X$. Εδώ τίθεται το εξής ερώτημα: δεδομένης της ένδειξης του δείγματος, μπορούμε εμείς να συμπεράνουμε, ότι η αληθινή αυτή σχέση ισχύει για όλα τα Χ και Υ;

Λαμβάνοντας υπόψη την παρακάτω υπόθεση ως μηδενική υπόθεση:

$$H_0:\beta_1=0,$$

όπου το β_1 είναι η κλίση της γραμμής παλινδρόμησης του πληθυσμού, παρατηρούμε, ότι αν αυτή η υπόθεση είναι αληθής, δεν υπάρχει σχέση μεταξύ του Υ και του Χ στον πληθυσμό. Αν το H_0 δεν απορριφθεί σημαίνει, ότι παρά το γεγονός, ότι το δείγμα έχει παράξει μία προσαρμοσμένη γραμμή με μία μη μηδενική τιμή για το b_1, εμείς πρέπει να συμπεράνουμε, ότι δεν υπάρχουν επαρκείς ενδείξεις, που να υποδεικνύουν, ότι το Υ σχετίζεται με το Χ. Αυτό συμβαίνει γιατί δεν μπορούμε να αποκλείσουμε την πιθανότητα, η γραμμή παλινδρόμησης του πληθυσμού να είναι οριζόντια.

Μία οριζόντια γραμμή παλινδρόμησης (β_1=0) είναι ισοδύναμη με την δήλωση H_0:r=0, όπου r είναι ο συντελεστής συσχέτισης του πληθυσμού.



Πώς μπορεί το β_1 να είναι μηδέν, όταν το b_1 είναι μη μηδενικό; Στο Γράφημα, όπου από ένα πληθυσμό δεδομένων ένα δείγμα από πέντε δεδομένα (τα επιλεγμένα σημεία δεδομένων συμβολίζονται με x). Όπως δείχνει αυτό το διάγραμμα διασποράς, αν έχουν επιλεχθεί αρκετά σημεία δεδομένων, θα γίνει εμφανές, ότι ο πληθυσμός των σημείων δεδομένων έχει μία γραμμή παλινδρόμησης με μηδενική κλίση. Ωστόσο, τα πέντε τυχαία επιλεγμένα σημεία δεδομένων βρίσκονται αρκετά κοντά σε μία γραμμή παλινδρόμησης, που τείνει προς τα πάνω. Μπορεί να καταλήξει κάποιος λανθασμένα από αυτή την ένδειξη στο συμπέρασμα, ότι το X και το Y σχετίζονται με έναν θετικά γραμμικό τρόπο. Ωστόσο, αν ελέγξουμε την υπόθεση β_1=0 με τα δεδομένα του δείγματος, ο αλγόριθμος πρόβλεψης πιθανότητα δεν θα είναι σε θέση να το απορρίψει.

Αν η μηδενική υπόθεση H_0:β_1=0 είναι αληθής, τότε η

στατιστική t με τιμή $t=(b_1-\beta_1)/s_{b_1}=b_1/s_{b_1}$ **ακολουθεί την t κατανομή με n-2 βαθμούς ελευθερίας. Το** s_{b_1} **είναι η τυπική απόκλιση του** b_1 **και δίδεται από τη σχέση:**

$$s_{b_1}=\frac{s_{y\cdot x}}{\sqrt{\sum (X-\bar{X})^2}}$$

Παράδειγμα

Διαμορφώνουμε τη μηδενική υπόθεση, ώστε να μην υπάρχει γραμμική σχέση, με την εναλλακτική να υπάρχει γραμμική σχέση στη περίπτωση, που απορριφθεί η μηδενική.

$H_0:\beta_1=0$ Δεν υπάρχει γραμμική σχέση,

ως προς:

$H_1:\beta_1\neq 0$ Υπάρχει γραμμική σχέση.

Υπολογίζουμε:

$$s_{b_1}=\frac{s_{y\cdot x}}{\sqrt{\sum (X-\bar{X})^2}}=\frac{0,226}{31,81}=0,007$$

Η test statistic είναι:

$$t=\frac{b_1}{s_{b_1}}=\frac{0,048}{0,007}=6,86$$

Για να ισχύει η μηδενική υπόθεση πρέπει η τιμή, t=6,86, να είναι μέσα στο διάστημα εμπιστοσύνης $[t_{0,025}, t_{0,975}]=[-2,08, 2,08]$ για df = n-2=21. Επειδή όμως η τιμή αυτή είναι εκτός διαστήματος, απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε την εναλλακτική, που σημαίνει ότι η παλινδρόμηση είναι σημαντική.

Για πολύ μεγάλα μεγέθη δειγμάτων, είναι πιθανό να απορρίψουμε την H_0 και να συμπεράνουμε, ότι υπάρχει γραμμική σχέση μεταξύ του Y και του X ακόμα και αν το r^2 είναι μικρό – ας πούμε 10%. Παρομοίως,

για μικρά δείγματα και ένα πολύ μεγάλο r^2, ας πούμε 95% – είναι πιθανό να συμπεράνουμε, ότι η παλινδρόμηση είναι επίσης σημαντική. Ένα μικρό r^2 σημαίνει, ότι η προσαρμοσμένη εξίσωση παλινδρόμησης, δεν είναι πιθανό να έχει μεγάλη ισχύ πρόβλεψης. Από την άλλη μεριά, ένα μεγάλο r^2 με ένα μικρό μέγεθος δείγματος μπορεί να απαιτήσει περισσότερες ενδείξεις από το δείγμα, πριν η προσαρμοσμένη συνάρτηση χρησιμοποιηθεί για πρόβλεψη. Μπορεί να υπάρξει διαφορά μεταξύ της στατιστικής και της πρακτικής σημασίας.

Ένας εναλλακτικός έλεγχος για την μηδενική υπόθεση $H_0 : \beta_1 = 0$ είναι επίσης διαθέσιμος από τον πίνακα ANOVA. Αν οι υποθέσεις του στατιστικού μοντέλου για τη γραμμική παλινδρόμηση είναι κατάλληλες και αν η μηδενική υπόθεση $H_0 : \beta_1 = 0$ είναι αληθής, ο λόγος:

$$F = \frac{MSR}{MSE},$$

ακολουθεί την F κατανομή με df = 1, n-2. Όταν η H_0 είναι αληθής, το MSE είναι εκτιμητής του σ^2 και της διασποράς του σφάλματος (ε) στο στατιστικό μοντέλο για παλινδρόμηση ευθείας γραμμής. Από την άλλη μεριά, αν η εναλλακτική υπόθεση $H_1 : \beta_1 \neq 0$ είναι αληθής, ο αριθμητής του F-λόγου τείνει να είναι μεγαλύτερος από τον παρονομαστή. Στην περίπτωση αυτή οι μεγάλοι F-λόγοι είναι συνεπείς με την εναλλακτική υπόθεση.

Στο μοντέλο παλινδρόμησης ευθείας γραμμής, ο έλεγχος της μηδενικής υπόθεσης $H_0 : \beta_1 = 0$ ως προς την εναλλακτική υπόθεση $H_1 : \beta_1 \neq 0$ μπορεί να βασιστεί στον F-λόγο, F=MSR/MSE με df = 1, n-2. Σε επίπεδο σημαντικότητας α, η περιοχή απόρριψης είναι η $F > F_\alpha$.

Παράδειγμα

Από τον Πίνακα ANOVA για τα δεδομένα μας έχουμε:

$$F = MSR/MSE = 2,3342/0,05 = 46,684$$

και με $\delta_1 = 1$ και $\delta_2 = 21$ βαθμούς ελευθερίας,

$$F_{0,05} = 4,32$$

$$F_{0,01} = 8,01$$

Δεδομένου ότι F=46,684>8,01, η μηδενική υπόθεση $H_0 : \beta_1 = 0$ απορρίπτεται και στο 1% επίπεδο σημαντικότητας. Άρα η παλινδρόμηση είναι σημαντική.

Δεν είναι τυχαίο, ότι τα αποτελέσματα του F ελέγχου είναι συνεπή με τα αποτελέσματα του t ελέγχου στο Παράδειγμα μας. Στην πραγματικότητα, μέσα στο σφάλμα στρογγυλεύσεων,

$$F = 46,684 = (6,86)^2 = t^2$$

Επιπλέον,

$$F_{0,01} = 8,01 = (2,83)^2 = (t_{0,005})^2$$

Επομένως, για ένα δεδομένο επίπεδο σημαντικότητας, ο t έλεγχος απορρίπτει την μηδενική υπόθεση $H_0 : \beta_1 = 0,$ όποτε και ο F-έλεγχος την απορρίπτει και αντιστρόφως. Αυτή η σχέση μεταξύ του t- και του F-ελέγχου ισχύει μόνο για το μοντέλο παλινδρόμησης ευθείας γραμμής.

Η F στατιστική μπορεί επίσης να εκφραστεί σε όρους του συντελεστή προσδιορισμού, r^2:

$$F = \frac{r^2 (n-2)}{1-r^2}$$

Με όλα τα υπόλοιπα σταθερά, η τιμή του F αυξάνει καθώς το r^2 αυξάνει. Καθώς το μεγαλύτερο μέρος της διασποράς του Y ερμηνεύεται από την προσαρμοσμένη γραμμή παλινδρόμησης, γίνεται πιο πιθανό ότι

η παλινδρόμηση θα κριθεί σημαντική από τον F-έλεγχο.

6.8 ΑΝΑΛΥΣΗ ΚΑΤΑΛΟΙΠΩΝ

Το να προσαρμόζεις μία ευθεία με τη μέθοδο ελαχίστων τετραγώνων, το να δημιουργείς διαστήματα εμπιστοσύνης για προσαρμογή και πρόβλεψη και το να ελέγχεις υποθέσεις, δεν ολοκληρώνει μία μελέτη παλινδρόμησης. Αυτά τα βήματα είναι τα συμπεράσματα, που μπορούν να προκύψουν, όταν το μοντέλο που χρησιμοποιείται είναι επαρκές. Στις περισσότερες μελέτες, δεν είναι προφανές, ότι ένα συγκεκριμένο μοντέλο είναι σωστό.

Τα συμπεράσματα μπορεί να είναι παραπλανητικά, αν οι υποθέσεις, που έγιναν στη διαμόρφωση του μοντέλου, είναι συνολικά ασυμβίβαστες με τα δεδομένα. Είναι απαραίτητο να ελέγξουμε τα δεδομένα προσεκτικά για ενδείξεις κάποιας παραβίασης των υποθέσεων. Να θυμηθούμε ξανά, ότι οι υποθέσεις για το μοντέλο παλινδρόμησης ευθείας γραμμής, είναι οι ακόλουθες:

1. Η υποκείμενη σχέση είναι γραμμική.
2. Τα σφάλματα είναι ανεξάρτητα.
3. Τα σφάλματα έχουν σταθερή διασπορά.
4. Τα σφάλματα είναι κανονικά κατανεμημένα.

Η πληροφόρηση στην διασπορά, που δεν μπορεί να ερμηνευθεί από την προσαρμοσμένη συνάρτηση παλινδρόμησης, περιέχεται στα κατάλοιπα, $e = Y - \hat{Y}$. Για να ελέγξουμε τις μετρικές ενός δοκιμαστικού μοντέλου, μπορούμε να εξετάσουμε τα κατάλοιπα, χαράζοντάς τα με διάφορους τρόπους:

1. Γράφημα ιστογράμματος καταλοίπων.

2. Γράφημα καταλοίπων σαν συνάρτηση των προσαρμοσμένων τιμών.
3. Γράφημα καταλοίπων σαν συνάρτηση της ανεξάρτητης μεταβλητής.
4. Γράφημα καταλοίπων στο χρόνο αν τα δεδομένα είναι χρονολογικά.

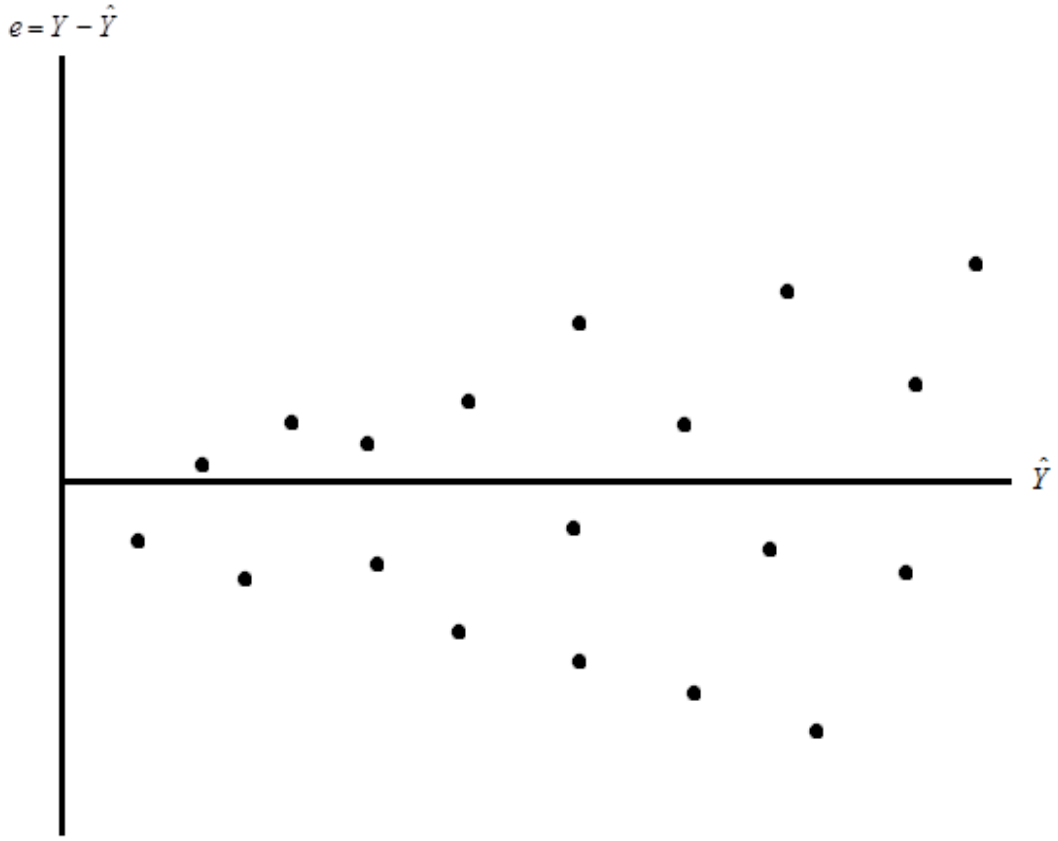
Ένα ιστόγραμμα των καταλοίπων παρέχει έναν έλεγχο της κανονικότητας της υπόθεσης. Τυπικά, μέτριες αποκλίσεις από μία καμπύλη σε σχήμα καμπάνας δεν αποδυναμώνουν τις υποθέσεις από τους ελέγχους, ή τα διαστήματα πρόβλεψης, που βασίζονται στην t κατανομή, ειδικά αν το σετ δεδομένων είναι μεγάλο. Η παραβίαση της υπόθεσης της κανονικότητας μόνο, συνήθως δεν είναι τόσο σοβαρή, όσο η παραβίαση κάποιας από τις υπόλοιπες υποθέσεις.

Αν μία γραφική παράσταση των καταλοίπων με τις προσαρμοσμένες τιμές υποδεικνύει, ότι η γενική φύση της σχέσης μεταξύ του Y και του X διαμορφώνει μία καμπύλη αντί για μία ευθεία γραμμή, ένας κατάλληλος μετασχηματισμός των δεδομένων μπορεί να μετατρέψει μία μη γραμμική σχέση σε μία άλλη, που είναι σχεδόν γραμμική. Ένας μετασχηματισμός μπορεί ακόμα να σταθεροποιήσει τη διασπορά.

Το παρακάτω Γράφημα δείχνει μία γραφική παράσταση των καταλοίπων, που υποδεικνύει, ότι η διασπορά των καταλοίπων αυξάνει, καθώς το μέγεθος των προσαρμοσμένων τιμών αυξάνει. Αυτό σημαίνει, ότι η μεταβλητότητα των σημείων δεδομένων γύρω από τη γραμμή των ελαχίστων τετραγώνων είναι μεγαλύτερη για μεγάλες τιμές, από ότι είναι για μικρές. Αυτό υπονοεί ότι η υπόθεση της σταθερής διασποράς μπορεί να μην ισχύει. Σε αυτή την περίπτωση, το να συσχετίσουμε το λογάριθμο του Y με το X μπορεί να παράγει διασπορά καταλοίπων, που είναι πιο

συνεπής με τη συνεχή διασπορά.

Η υπόθεση της ανεξαρτησίας είναι η πιο κρίσιμη. Η έλλειψη ανεξαρτησίας μπορεί δραστικά να διαστρεβλώσει τα συμπεράσματα, που βγήκαν από τους t ελέγχους. Η υπόθεση ανεξαρτησίας είναι ιδιαίτερα επικίνδυνη για τις χρονοσειρές δεδομένων, τα οποία προκύπτουν συχνά σε επιχειρηματικά και οικονομικά προβλήματα πρόβλεψης.



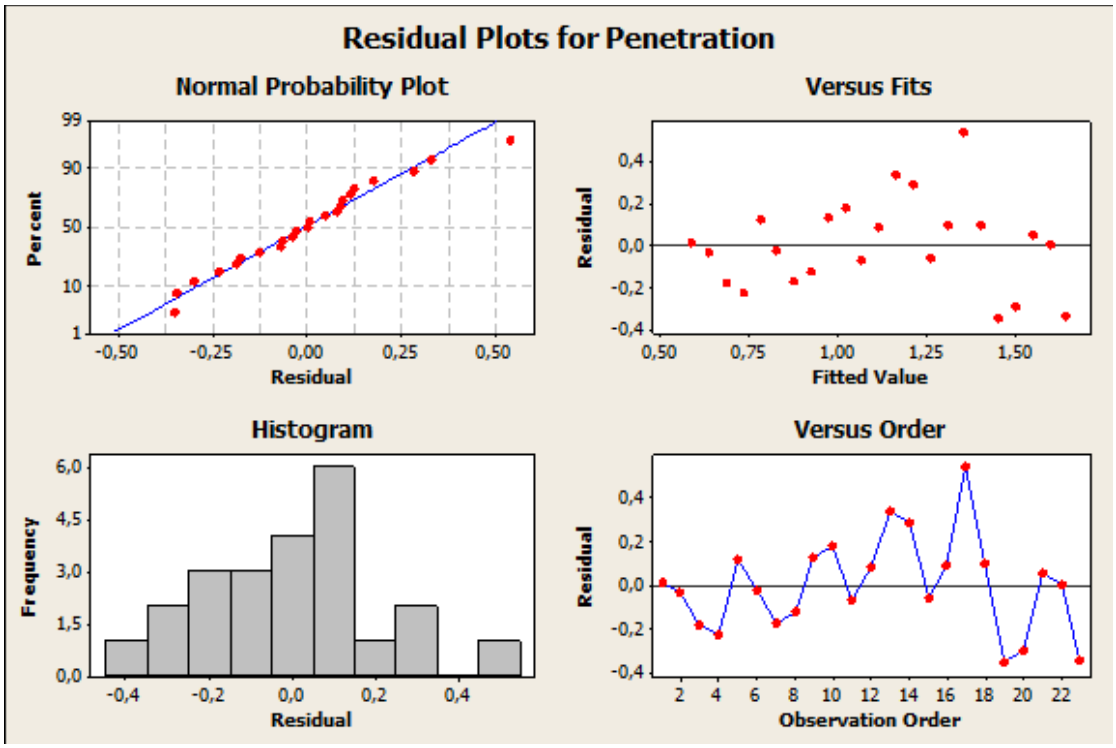
Για κατάλοιπα σειρών δεδομένων – το οποίο σημαίνει, για κατάλοιπα, που παράγονται χρησιμοποιώντας μεθόδους παλινδρόμησης σε χρονο-διατεταγμένα δεδομένα – η ανεξαρτησία μπορεί να ελεγχθεί με ένα διάγραμμα των καταλοίπων με το χρόνο. Δεν θα έπρεπε να υπάρχει συστηματικό σχέδιο, όπως μία σειρά από υψηλές τιμές ακολουθούμενη από μία σειρά με χαμηλές τιμές. Επιπλέον, οι αυτοσυσχετίσεις δείγματος των καταλοίπων (sample autocorrelations of the residuals):

$$r_k(e)=\frac{\sum_{t=k+1}^{n} e_t e_{t-k}}{\sum_{t=1}^{n} e_t^2},\quad k=1,2,\ldots,K$$

όπου n είναι ο αριθμός των καταλοίπων και K είναι χαρακτηριστικά n/4, θα έπρεπε να είναι όλες μικρές. Συγκεκριμένα, η ανεξαρτησία γίνεται εμφανής, αν ο κάθε ένας από τους συντελεστές αυτοσυσχέτισης των καταλοίπων είναι μέσα στο διάστημα $0\pm 2/\sqrt{n}$ για όλα τα k.

Παράδειγμα

Χρησιμοποιώντας τις προσαρμοσμένες τιμές και τα κατάλοιπα του παραδείγματος της ευρυζωνικότητας, κατασκευάσαμε ένα ιστόγραμμα των καταλοίπων και ένα διάγραμμα των καταλοίπων ενάντια στις προσαρμοσμένες τιμές. Αυτά και άλλα αποτελέσματα φαίνονται στο παρακάτω Γράφημα.



Τα σημεία στο Normal Probability Plot στο Γράφημα, βρίσκονται πολύ κοντά στην ευθεία γραμμή. Αυτή η συμπεριφορά υποδεικνύει μία

καλή προσαρμογή μεταξύ των δεδομένων (εδώ, των καταλοίπων) και της κανονικής κατανομής. Το Normal Probability Plot δείχνει, ότι δεν υπάρχει λόγος να αμφισβητήσουμε την κανονική υπόθεση.

Το δεύτερο γράφημα στην πρώτη γραμμή φαίνεται επίσης καλό. Η διασπορά γύρω από το μηδέν στην κάθετη κατεύθυνση πρέπει να είναι περίπου η ίδια για όλες τις τιμές στον οριζόντιο άξονα. Αυτό σημαίνει, ότι τα μεγέθη των καταλοίπων για μικρές προσαρμοσμένες τιμές πρέπει να είναι περίπου τα ίδια με τα μεγέθη των καταλοίπων για μεσαίες προσαρμοσμένες τιμές και μεγάλες προσαρμοσμένες τιμές. Αυτή η ιδανική συμπεριφορά σημαίνει δύο πράγματα: (α) Η υποκείμενη σχέση μεταξύ του Y και του X είναι γραμμική και (β) η μεταβλητότητα των σφαλμάτων είναι σταθερή (τα Y για τις διαφορετικές τιμές του X έχουν την ίδια διασπορά στη γραμμή παλινδρόμησης).

Είναι καλό το ότι δεν έχουμε μία σειρά θετικών καταλοίπων ακολουθούμενη από μία σειρά αρνητικών καταλοίπων, η οποία ακολουθείται επίσης από μία σειρά θετικών καταλοίπων. Αυτή η συμπεριφορά θα υποδείκνυε μία μη γραμμική σχέση μεταξύ Y και X.

Η ανάλυση των καταλοίπων είναι ικανοποιητική. Το μοντέλο παλινδρόμησης ευθείας γραμμής μπορεί να περιγράψει ικανοποιητικά τη σχέση μεταξύ του Y και του X.

6.9 ΥΠΟΛΟΓΙΣΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Το πρόβλημα ανάλυσης παλινδρόμησης (χρησιμοποιώντας τα δεδομένα του παραδείγματος ευρυζωνικότητας) τα τρέξαμε στο Minitab και τα αποτελέσματα παρουσιάζονται παρακάτω.

Για να εξηγήσουμε την ορολογία που χρησιμοποιείται στα

υπολογιστικά αποτελέσματα, οι ορισμοί και οι υπολογισμοί παρουσιάζονται στην ακόλουθη λίστα.

Correlations: Penetration (%); Time

```
Pearson correlation of Penetration (%) and Time
= 0,829    P-Value = 0,000
```

Regression Analysis: Penetration (%) versus Time

```
The regression equation is
Penetration (%) = 0,541 + 0,0480 Time

Predictor  Coef   SE Coef   T     P
Constant 0,54111  0,09679  5,59 0,000
Time     0,048024 0,007059 6,80 0,000

S = 0,224573 R-Sq = 68,8% R-Sq(adj) = 67,3%

Analysis of Variance
Source          DF   SS     MS      F    P
Regression       1 2,3340 2,3340 46,28 0,000
Residual Error 21 1,0591 0,0504
Total          22 3,3930
```

Predictor: Σταθερά και ανεξάρτητη μεταβλητή.
Coef.: Οι τιμές των b_0, b_1.
SE Coef.: Τυπική απόκλιση των b_0, b_1.
T: Οι τιμές της στατιστικής t.
P: Οι P-τιμές της στατιστικής t.
S: Η τυπικό σφάλμα εκτίμησης της παλινδρόμησης $s_{y \cdot x}$.
Από κάτω δίδεται ο πίνακας ANOVA

6.10 ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΙ ΜΕΤΑΒΛΗΤΩΝ

Το μοντέλο της απλής γραμμικής παλινδρόμησης υποθέτει μία σχέση ευθείας γραμμής μεταξύ του Y και του X. Όσο η συνάρτηση παλινδρόμησης είναι γραμμική ως προς τους συντελεστές β, οι ερμηνευτικές μεταβλητές μπορούν να πάρουν διάφορες μορφές και η

μεθοδολογία τυπικής παλινδρόμησης να είναι ακόμα κατάλληλη. Τα μοντέλα παλινδρόμησης μπορούν τότε να χρησιμοποιηθούν, για να διαμορφώσουμε περίπλοκες σχέσεις μεταξύ του Y και του X, ή για να διαμορφώσουμε σχέση ευθείας γραμμής μεταξύ του Y και κάποιας μετασχηματισμένης συνάρτησης του X.

Όταν ένα διάγραμμα διασποράς υποδεικνύει ότι υπάρχει μία μη γραμμική σχέση μεταξύ του Y και του X, υπάρχουν δύο βασικές προσεγγίσεις για να αντιμετωπίσουμε την περίπτωση αυτή. Η πρώτη είναι να προσαρμόσουμε τα δεδομένα με μία καμπύλη – μη γραμμική – συνάρτηση παλινδρόμησης και να χρησιμοποιήσουμε τη προσαρμοσμένη σχέση για σκοπούς πρόβλεψης. Η δεύτερη προσέγγιση περιλαμβάνει το μετασχηματισμό της μεταβλητής X σε άλλη μορφή, έτσι ώστε η προκύπτουσα σχέση με το Y να είναι γραμμική.

Τέσσερις από τους πιο κοινούς μετασχηματισμούς (συναρτήσεις), που χρησιμοποιούνται για να παραχθούν νέες μεταβλητές πρόβλεψης, είναι οι: αντίστροφη, λογαριθμική, τετραγωνική ρίζα και τετράγωνο:

$$\frac{1}{X},\ \log X,\ \sqrt{X},\ X^2$$

Όταν για όλες αυτές τις μεταβλητές δημιουργηθούν γραφικές παραστάσεις με το Y, η ελπίδα είναι ότι η μη γραμμική σχέση μεταξύ του Y και του X θα γίνει μία γραμμική σχέση μεταξύ του Y και ενός εκ των μετασχηματισμών του X. Αν γίνει αυτό, το Y και η νέα μεταβλητή μπορούν να χρησιμοποιηθούν με το μοντέλο ευθείας γραμμής, που συζητήθηκε, συμπεριλαμβάνοντας τον συντελεστή αυτοσυσχέτισης και την εξίσωση προσαρμοσμένης παλινδρόμησης.

Στο ακόλουθο παράδειγμα, χρησιμοποιείται το Minitab για να παραχθεί η γραφική παράσταση μιας απλής σχέσης X και Y που φαίνεται

να είναι μη γραμμική. Το πρόγραμμα στη συνέχεια χρησιμοποιείται για να υπολογίσει τέσσερις μετασχηματισμούς που περιγράφηκαν προηγουμένως. Για αυτές τις μεταβλητές στη συνέχεια δημιουργούνται γραφικές παραστάσεις με το Υ για να παραχθούν οι γραφικές παραστάσεις των δεδομένων που φαίνονται παρακάτω.

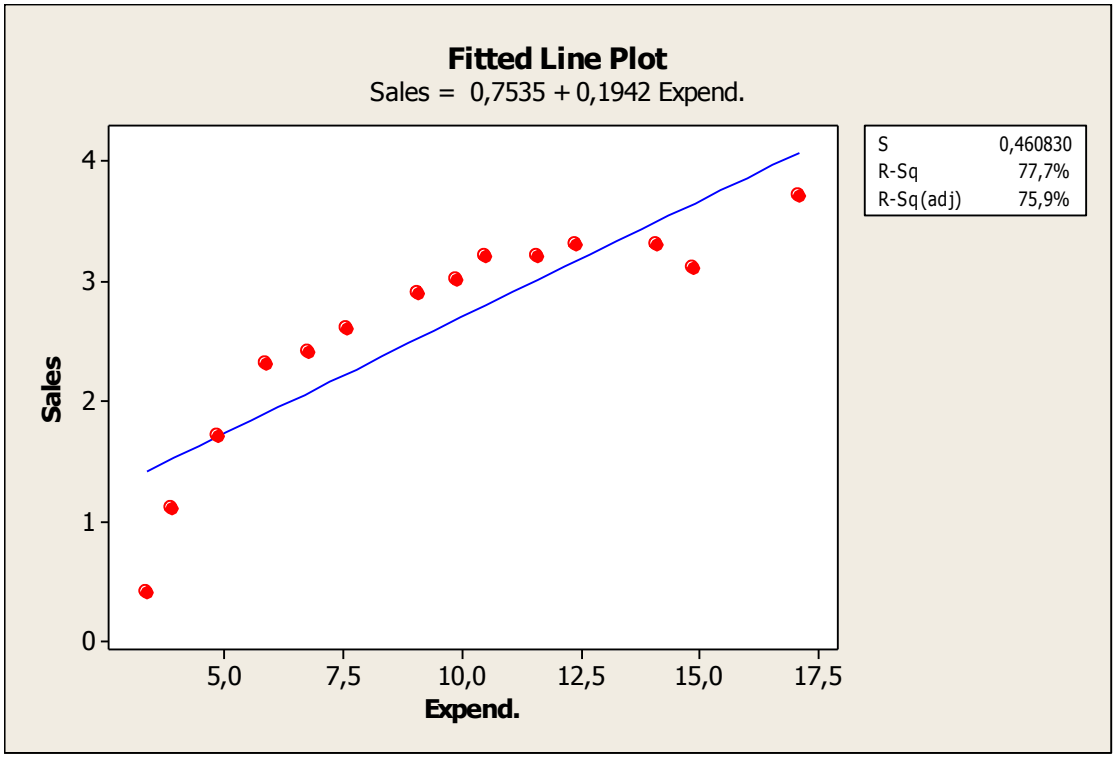
Παράδειγμα (F.6-10)

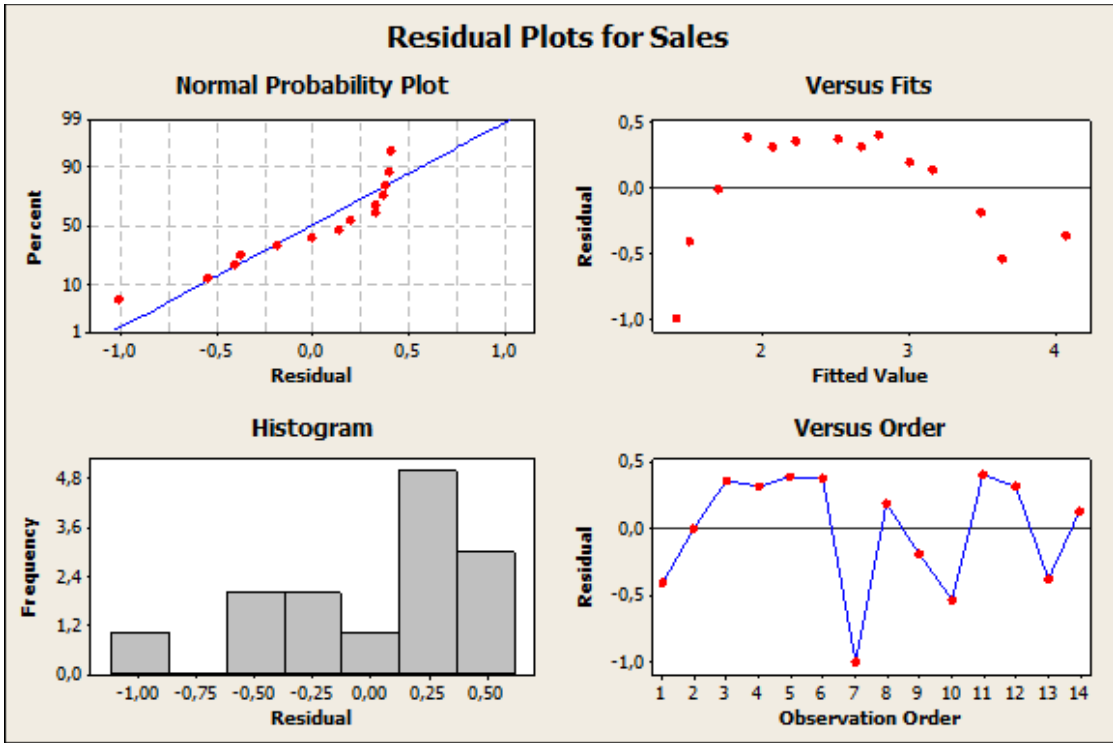
Ο κ. Gilbert Garcia έχει στην ιδιοκτησία του μία αλυσίδα καταστημάτων hardware στο Σικάγο. Ενδιαφέρεται να προβλέψει τις μηνιαίες του πωλήσεις χρησιμοποιώντας τη γνώση των αντίστοιχων μηνιαίων εξόδων του για διαφήμιση. Ο κ. Garcia πιστεύει ότι οι πωλήσεις θα αυξηθούν, αν αυξήσει τις δαπάνες του για διαφήμιση. Ωστόσο, επίσης πιστεύει ότι μετά από ένα σημείο πωλήσεων, οι πωλήσεις του θα αυξάνουν με χαμηλότερο ρυθμό. Ο κ. Garcia πιστεύει ότι αφού ξοδέψει ένα ποσό στη διαφήμιση, θα φτάσει σε ένα σημείο, όπου θα έχει λίγα να κερδίσει με περαιτέρω διαφήμιση των καταστημάτων του.

Ο κ. Garcia επέλεξε ένα τυχαίο δείγμα δεδομένων 14 μηνών από τα αρχεία της εταιρείας του. Ο κ. Garcia σημειώνει ότι οι πωλήσεις φαίνεται να μειώνονται μετά από ένα συγκεκριμένο σημείο, που ξοδεύεται για διαφήμιση. Προσαρμόζει την εξίσωση γραμμικής παλινδρόμησης, η οποία φαίνεται στο Γράφημα και σημειώνει, ότι η εξίσωση ερμηνεύει το 77,7% της μεταβλητότητας στις πωλήσεις.

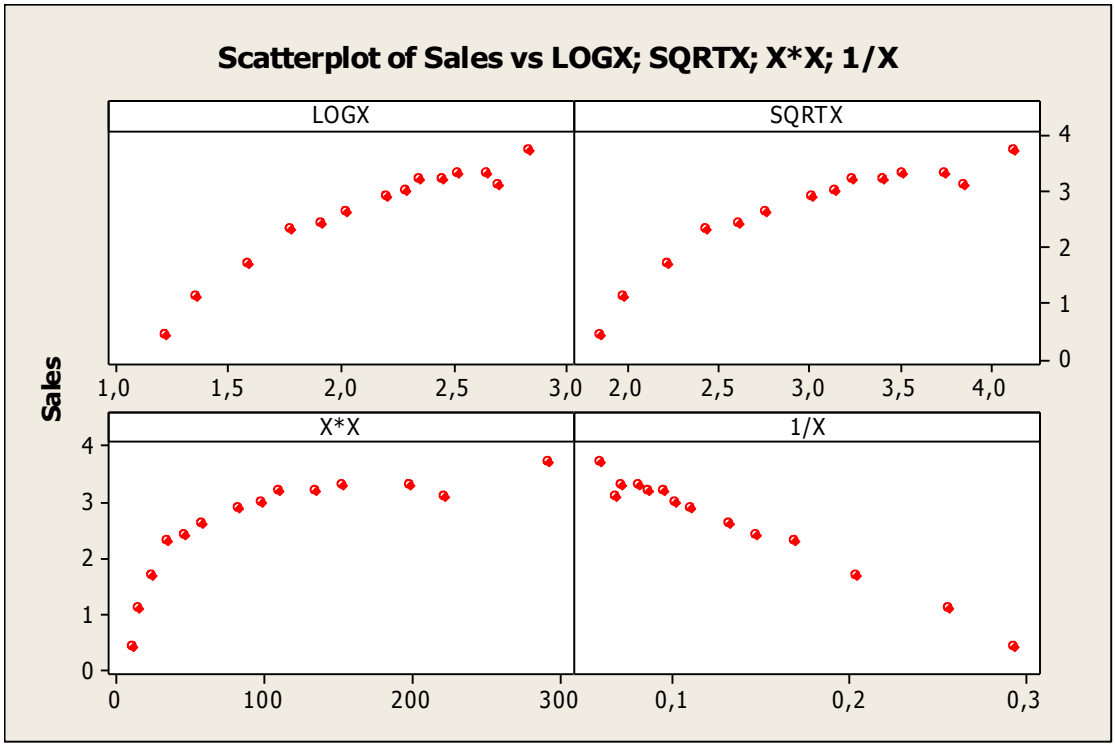
Το διάγραμμα των καταλοίπων σε σχέση με τις προσαρμοσμένες από ευθεία τιμές υποδεικνύει, ότι μία ευθεία γραμμή δεν περιγράφει επαρκώς τη σχέση μεταξύ των πωλήσεων και των διαφημιστικών εξόδων. Τα κατάλοιπα είναι όλα αρνητικά για μικρές προβλεφθείσες τιμές, είναι όλα θετικά για μεσαίες προβλεφθείσες τιμές και γίνονται πάλι

αρνητικά για μεγάλες προβλεφθείσες τιμές. Τα κατάλοιπα δεν είναι ομοιόμορφα κατανεμημένα στην εκτιμηθείσα γραμμή παλινδρόμησης. Είναι ξεκάθαρο, ότι μία ευθεία γραμμή δεν προσαρμόζει την καμπυλότητα των σημείων.





Στη συνέχεια, ο κ. Garcia θεωρεί μερικούς μετασχηματισμούς του X. Κατασκευάζει ένα πολλαπλό διάγραμμα διασποράς με γραφήματα των πωλήσεων (Y) έναντι των λογαρίθμων των διαφημιστικών εξόδων (logX), των πωλήσεων (Y) έναντι της ρίζας των διαφημιστικών εξόδων $(\sqrt{X})$, των πωλήσεων (Y) έναντι του τετραγώνου των διαφημιστικών εξόδων (X^2), και τέλος των πωλήσεων (Y) έναντι του αντιστρόφου των διαφημιστικών εξόδων (1/X). Το αποτέλεσμα φαίνεται παρακάτω.



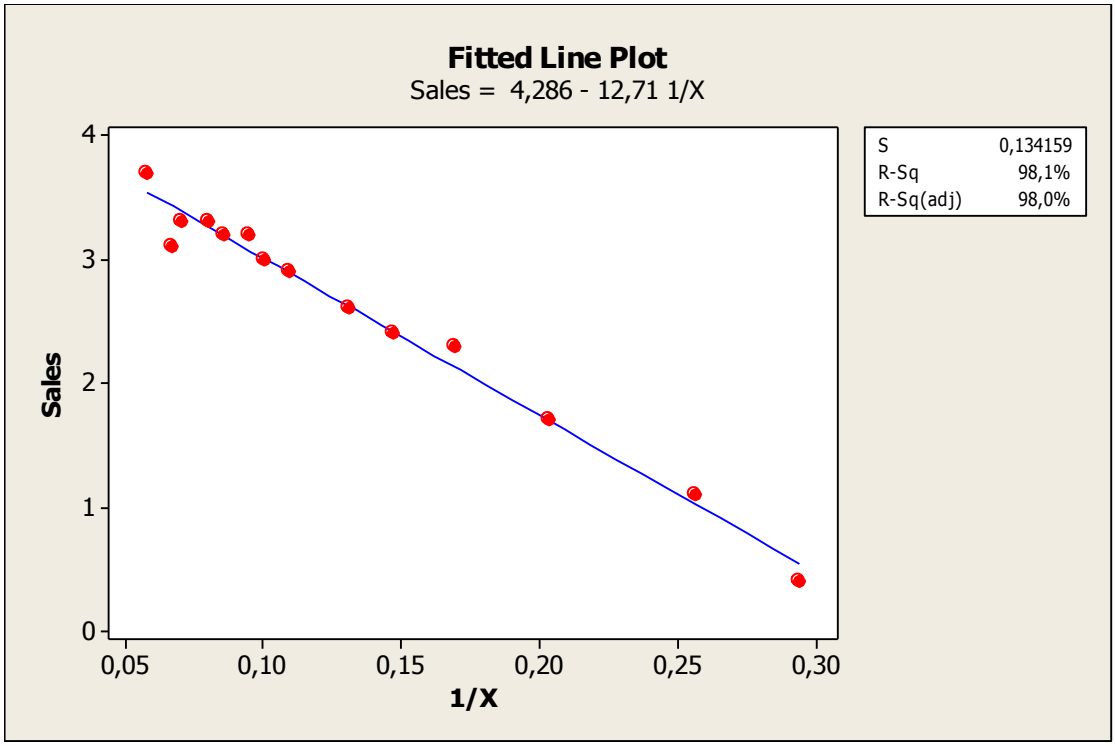
Μετά την εξέταση του πολλαπλού διαγράμματος διασποράς, ο κ. Garcia συμπεραίνει, ότι οι σχέσεις μεταξύ Y και logX, μεταξύ Y και ρίζα X, μεταξύ Y και X^2 και μεταξύ Y και X και βλέπει κάποιο βαθμό καμπυλότητας. Αυτές οι σχέσεις δεν είναι γραμμικές. Ωστόσο, η σχέση μεταξύ των πωλήσεων και του αντιστρόφου των διαφημιστικών εξόδων φαίνεται να είναι γραμμική με αρνητική κλίση. Χρησιμοποιώντας το Minitab, ο κ. Garcia προσαρμόζει ένα μοντέλο απλής γραμμικής παλινδρόμησης στα δεδομένα των πωλήσεων και του αντιστρόφου των

διαφημιστικών εξόδων.

Αυτός καθορίζει την προσαρμοσμένη εξίσωση,

$$\hat{Y} = 4{,}29 - 12{,}7\frac{1}{X}$$

και βλέπει, ότι το r^2 ισούται με 98,1%. Μία ανάλυση των καταλοίπων υποδεικνεύει, ότι το γραμμικό μοντέλο είναι κατάλληλο.



Regression Analysis: Sales versus 1/X

The regression equation is

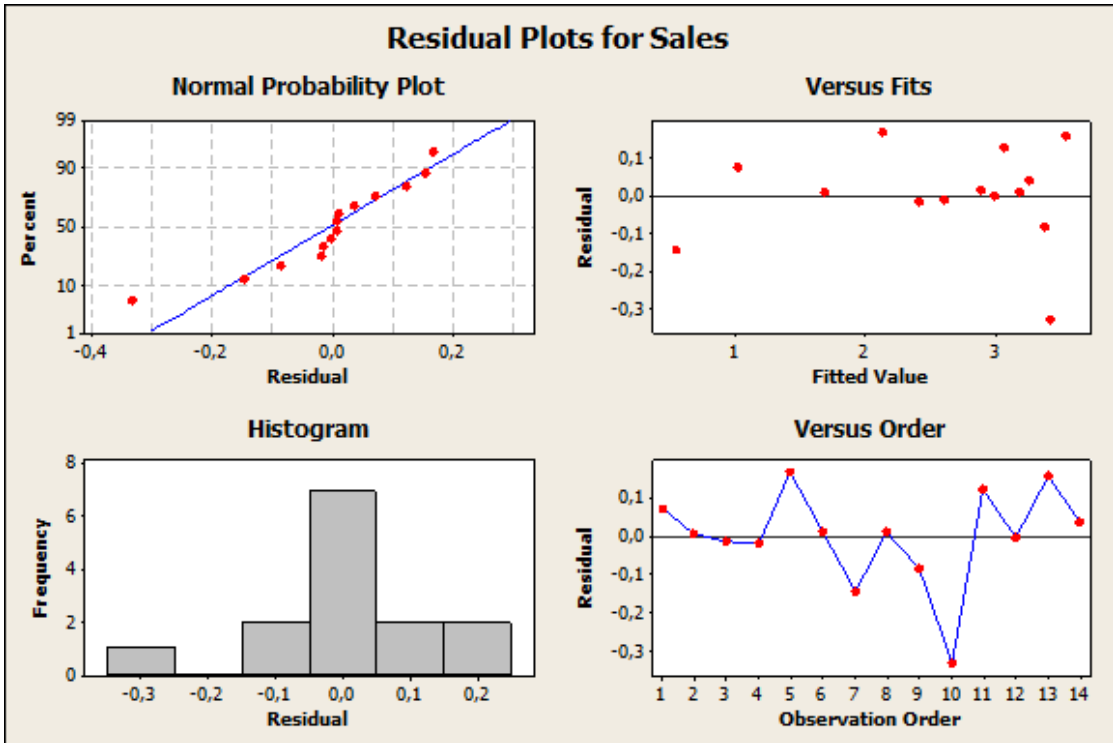
Sales = 4,29 - 12,7 1/X

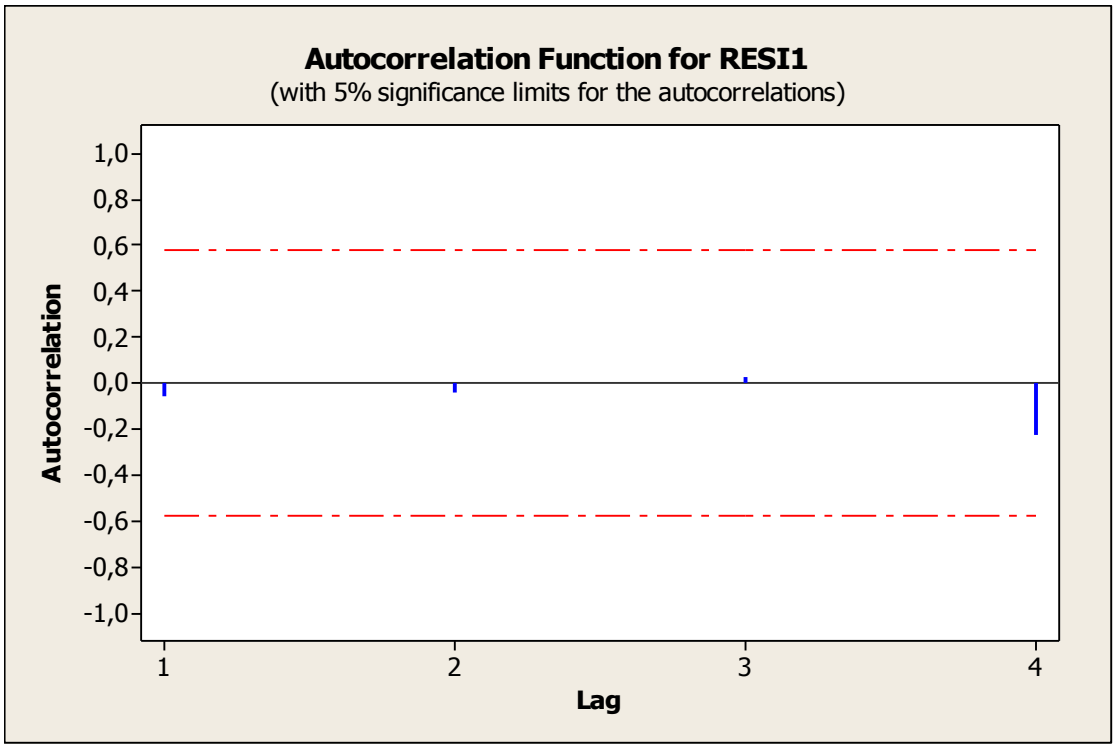
Predictor	Coef	SE Coef	T	P
Constant	4,28587	0,07695	55,69	0,000
1/X	-12,7132	0,5092	-24,97	0,000

S = 0,134159 R-Sq = 98,1% R-Sq(adj) = 98,0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	11,221	11,221	623,44	0,000
Residual Error	12	0,216	0,018		
Total	13	11,437			





Lag	ACF	T	LBQ
1	-0,055110	-0,21	0,05
2	-0,043994	-0,16	0,09
3	0,027625	0,10	0,10
4	-0,223745	-0,83	1,23

6.11 ΚΑΜΠΥΛΕΣ ΑΝΑΠΤΥΞΗΣ (GROWTH CURVES)

Οι καμπύλες ανάπτυξης εισήχθησαν στο Κεφάλαιο 5. Πολύ συχνά η τάση – trend – (η αλλαγή σε μεγάλο χρονικό διάστημα) έχει μεγάλο ενδιαφέρον. Ένας ειδικός στις προβλέψεις για παράδειγμα μπορεί να ενδιαφέρεται, στο να προβλέψει το κόστος ανά υπολογιστική συναλλαγή στο μέλλον, χωρίς να εξαρτάται από την τεχνολογία, που μπορεί να φέρει αλλαγές στο κόστος. Ένα άλλο παράδειγμα μπορεί να είναι ένας αναλυτής ασφάλειας ζωής, ο οποίος μπορεί να ενδιαφέρεται σε μακροπρόθεσμες προβολές του προσδόκιμου ζωής και στις γεννήσεις των ΗΠΑ: μη οικονομικοί και περιβαλλοντικοί παράγοντες, που μπορεί να είναι υπεύθυνοι για αλλαγές στο προσδόκιμο της ζωής του πληθυσμού.

Οι καμπύλες ανάπτυξης είναι καμπυλόγραμμες σχέσεις μεταξύ της μεταβλητής του ενδιαφέροντος και του χρόνου. Οι καμπύλες ανάπτυξης είναι χαρακτηριστικά κατάλληλες για ετήσια δεδομένα, επειδή απαιτούνται μακροπρόθεσμες προβλέψεις.

Αν και μη ακριβείς προβλέψεις είναι πιθανές, όταν οι καμπύλες ανάπτυξης, που προσαρμόζονται σε ιστορικά δεδομένα, επεκτείνονται για να προβλέψουν το μέλλον, αυτή η μέθοδος πρόβλεψης μπορεί να είναι καλή για τους managers, καθώς επικεντρώνεται στις μακροπρόθεσμες πτυχές τις επιχείρησης. Επιπλέον, οι καμπύλες ανάπτυξης δείχνουν τον ετήσιο ρυθμό ανάπτυξης, που πρέπει να διατηρείται, έτσι ώστε να επιτυγχάνονται τα προσχεδιασμένα μελλοντικά επίπεδα τιμών. Αυτός ο ετήσιος ρυθμός ανάπτυξης μπορεί να είναι, ή να μην είναι λογικός.

Αν μία μεταβλητή, που υπολογίζεται με την πάροδο του χρόνου, αυξάνει στο ίδιο ποσοστό σε κάθε χρονική περίοδο, λέγεται ότι

παρουσιάζει εκθετική ανάπτυξη (exponential growth). Αν η μεταβλητή αυξάνει κατά την ίδια ποσότητα κάθε χρονική περίοδο, λέγεται ότι παρουσιάζει γραμμική ανάπτυξη (linear growth). Μερικές φορές ένας απλός μετασχηματισμός θα μετατρέψει μία μεταβλητή, που έχει εκθετική ανάπτυξη, σε μία μεταβλητή, που έχει γραμμική ανάπτυξη. Αν αυτή είναι η περίπτωση, οι μέθοδοι παλινδρόμησης, που συζητήθηκαν σε αυτό το κεφάλαιο, μπορούν να χρησιμοποιηθούν για να μοντελοποιήσουμε την εκθετική ανάπτυξη.

Για παράδειγμα υποθέστε μία μεταβλητή Y, που υπολογίζεται ετήσια, και ακολουθεί τάση εκθετικής καμπύλης.

$$Y = b_0 b_1^t$$

Εδώ, το $100(b_1 - 1)$ % είναι το ποσοστό της αύξησης στο Y από έτος σε έτος. Λογαριθμίζοντας και τις δύο πλευρές αυτής της εξίσωσης εκθετικής τάσης έχουμε:

$$\log Y = \log(b_0 b_1^t) = \log b_0 + t \log b_1 = \bar{b}_0 + \bar{b}_1 t$$

Κατά συνέπεια, η γραφική παράσταση του logY με το χρόνο (t) είναι μία ευθεία γραμμή με κλίση $\bar{b}_1$ και τομή με τον άξονα Y, $\bar{b}_0$.

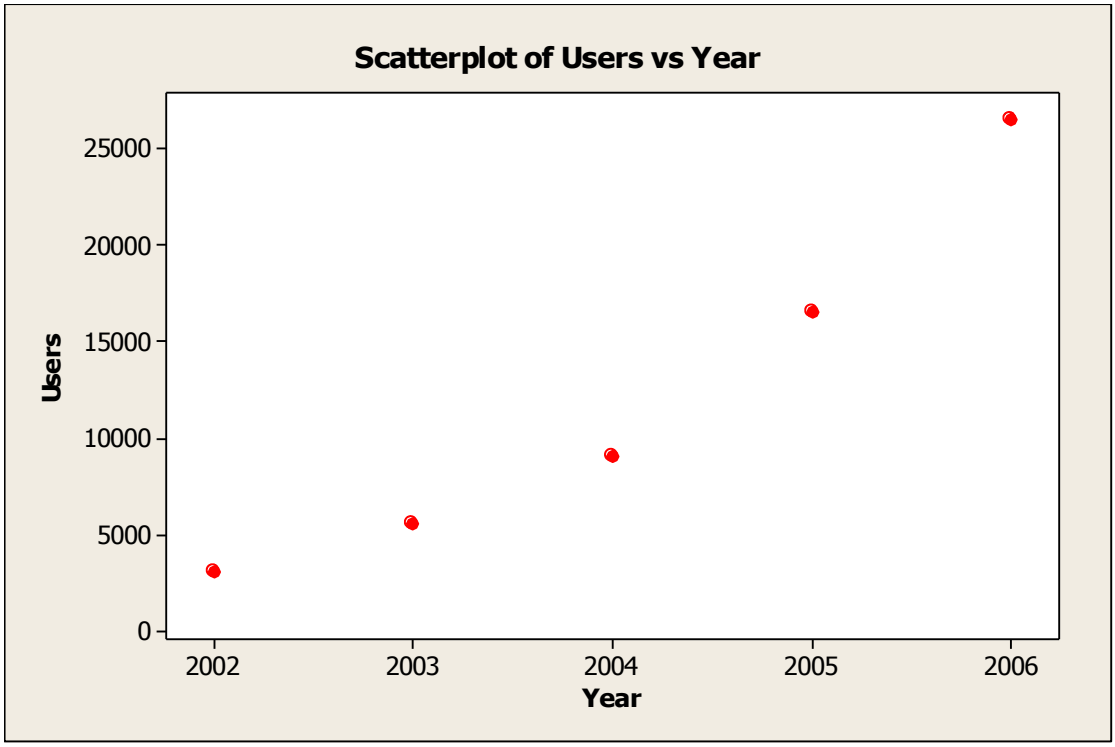
Παράδειγμα (Τ.6-10)

Η Jill Johnson, μία διοικητική υπάλληλος σε ένα μεγάλο πανεπιστήμιο, είναι ικανοποιημένη με την αποδοχή ενός διαδικτυακού συστήματος εγγραφής στα μαθήματα. Έχει συλλέξει τα δεδομένα και βλέπει ότι η ανάπτυξη χρήσης του συστήματος είναι πρωτοφανής. Αν ο τωρινός ρυθμός ανάπτυξης συνεχιστεί, μέσα σε ένα χρόνο όλοι οι σπουδαστές θα εγγράφονται διαδικτυακά για τα μαθήματα και το παλιό τηλεφωνικό σύστημα εγγραφών θα έχει απαξιωθεί πλήρως.

Τα δεδομένα στον παρακάτω Πίνακα δείχνουν τον αριθμό των

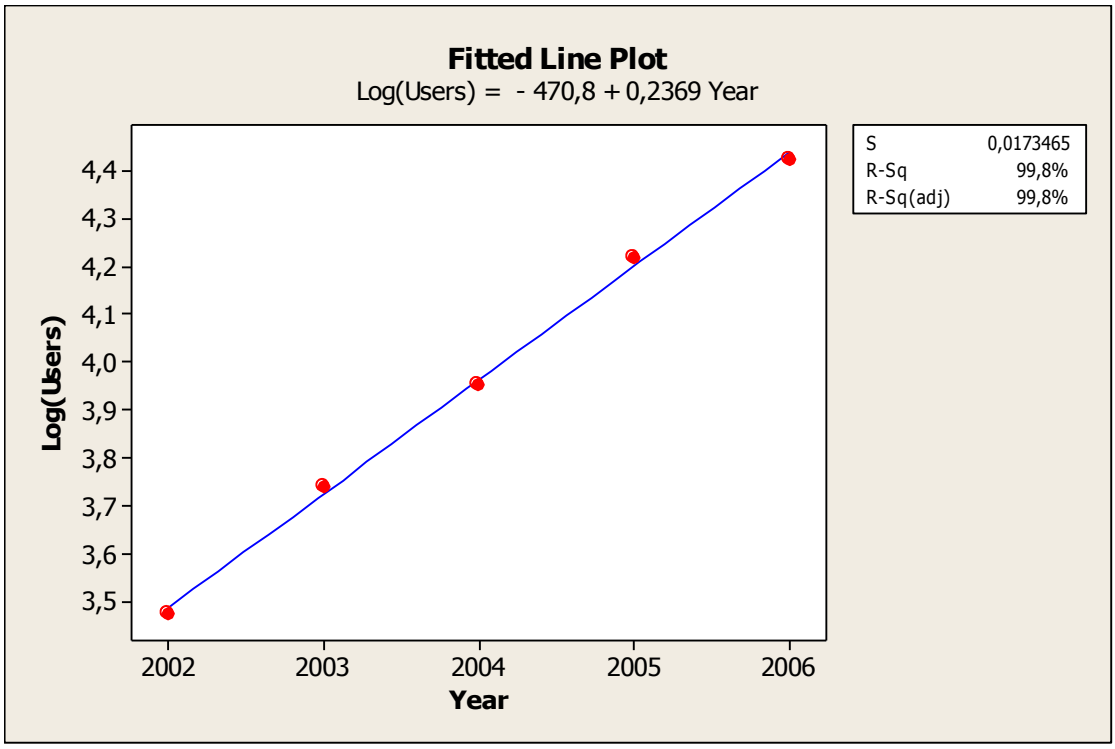
σπουδαστών που χρησιμοποίησαν το διαδικτυακό σύστημα εγγραφής για την περίοδο 2002-2006. Η τελευταία στήλη του πίνακα περιέχει τους λογαρίθμους με βάση το δέκα των αριθμών των χρηστών.

Year	Time	Users	Log (Users)
2002	1	3.000	3,477
2003	2	5.500	3,740
2004	3	9.000	3,954
2005	4	16.500	4,217
2006	5	26.500	4,423



Ένα γράφημα του αριθμού των χρηστών σαν συνάρτηση του χρόνου δείχνει, ότι ένα μοντέλο απλής γραμμικής παλινδρόμησης δεν είναι κατάλληλο για τα δεδομένα της. Η ανάπτυξη στην χρήση φαίνεται να είναι εκθετική, με μία ετήσια ποσοστιαία ανάπτυξη της τάξης του 70% ανά έτος. Η αύξηση του αριθμού των χρηστών από χρόνο σε χρόνο, ωστόσο, είναι αρκετά διαφορετική. Οι αυξήσεις κυμαίνονται από χαμηλές (2.500 επιπλέον χρήστες) στην περίοδο 2002-2003 σε υψηλές (10.000 επιπλέον χρήστες) την περίοδο 2005-2006.

Λογαριθμίζοντας τους αριθμούς των χρηστών, που έκαναν διαδικτυακή εγγραφή, παίρνουμε το γράφημα του λογαρίθμου των χρηστών σαν συνάρτηση του χρόνου. Όπως φαίνεται στο διάγραμμα η νέα σχέση είναι γραμμική. Η αύξηση στο μέγεθος του λογαρίθμου των χρηστών από χρόνο σε χρόνο είναι περίπου η ίδια, περίπου 0,24 το χρόνο. Προσαρμόζοντας μία ευθεία γραμμή στα δεδομένα με τον λογάριθμο των χρηστών σαν την εξαρτημένη μεταβλητή Y και το χρόνο (t) σαν την ανεξάρτητη μεταβλητή. Τα αποτελέσματα φαίνονται στα παρακάτω διαγράμματα.



Χρησιμοποιώντας την προσαρμοσμένη εξίσωση,

$$\log \hat{Y}_{2007} = -470{,}8 + 0{,}2369\, Year.$$

Για το έτος 2007 υπολογίζουμε την πρόβλεψη:

$$\log \hat{Y}_{2007} = -470{,}8 + 0{,}2369\, x\, 2007 = 4{,}6583,$$

ότι περίπου 45.500 σπουδαστές θα χρησιμοποιήσουν το διαδικτυακό σύστημα εγγραφής μέσα στο 2007. Δεδομένου ότι το πανεπιστήμιο

υπολογίζει να έχει περίπου 45.000 σπουδαστές, η Jill συμπεραίνει, ότι όλοι οι σπουδαστές θα χρησιμοποιήσουν πιθανότατα το νέο σύστημα εγγραφής στα μαθήματα και το παλιό τηλεφωνικό σύστημα εγγραφής μπορεί να σταματήσει να χρησιμοποιείται.

6.12 CASE 6-2 BUTCHER PRODUCTS, INC.

Ο Gene Butcher είναι ιδιοκτήτης και πρόεδρος της εταιρείας "Butcher Products, INC", η οποία παράγει οπτικές ίνες για ηλεκτρικές εγκαταστάσεις καλωδίων. Ο Gene μελέτησε τον αριθμό των αγωγών που παράχθηκαν ανά μέρα, τα τελευταία 2,5 χρόνια και εστίασε στην ευρεία μεταβλητότητα που παρατηρείται. Προκειμένου ο Gene να προβλέψει την ποσότητα της παραγωγής, τα κόστη και τα έσοδα, χρειάζεται να ανακαλύψει μία σχέση μεταξύ της παραγωγής και ορισμένων άλλων μεταβλητών.

Βασιζόμενος στη μέχρι τώρα εμπειρία του, αδυνατεί να καταλήξει στις αιτίες της μεταβλητότητας της παραγωγής. Δείχνει να το καταφέρνει όμως, από την στιγμή που άρχισε να συσχετίζει τις καιρικές συνθήκες. Η αιτιολόγηση του είναι, ότι η εξωτερική θερμοκρασία έχει πιθανότατα κάποια επίδραση στην παραγωγικότητα, στην εργατικότητα και κατ' επέκταση στην ημερήσια παραγωγή που επιτυγχάνεται.

Επέλεξε τυχαία, διάφορες μέρες από τα δεδομένα που είχε αποθηκεύσει και βρήκε την παραγωγή της καθεμίας από αυτές τις μέρες. Στη συνέχεια πήγε στην τοπική μετεωρολογική υπηρεσία και ζήτησε την υψηλότερη θερμοκρασία που καταγράφηκε για καθεμία από τις μέρες που είχε προ-επιλέξει. Πλέον είναι έτοιμος να τρέξει μία μελέτη συσχέτισης μεταξύ των μεταβλητών, που αντιλήφθηκε ότι επηρεάζουν την παραγωγή. Η παραγωγή ενδεχομένως να συσχετίζεται περισσότερο

με την απόκλιση της θερμοκρασίας από την ιδανική παρά με την ίδια την τιμή της θερμοκρασίας. Και αυτό γιατί μια πολύ ζεστή ή μια πολύ κρύα ημέρα επιδρά αρνητικά στην παραγωγικότητα, εφόσον αυτή συγκριθεί με μία μέρα ιδανικής θερμοκρασίας. Αποφάσισε να θεωρήσει ως ιδανική τη θερμοκρασία των 18 βαθμών Κελσίου (65 Fahrenheit), η οποία προκαλεί τη μέγιστη αποδοτικότητα των εργατών. Όλες οι αποκλίσεις από τις ιδανικές θερμοκρασίες θα πρέπει να επικαλυφθούν. Στον παρακάτω πίνακα, το Y αναπαριστά τον αριθμό των μονάδων που παράγονται, ενώ το X την απόλυτη διαφορά μεταξύ της υψηλότερης θερμοκρασίας ημέρας και της ιδανικής των 65 Fahrenheit.

Y	**X**	**Y**	**X**
485	12	327	15
512	10	308	25
625	3	603	8
585	4	321	35
318	27	426	5
405	10	410	12
379	18	515	2
497	12	498	7
316	27	357	17
351	20	429	8
525	4	401	12
395	11		

Ο Gene εφάρμοσε το μοντέλο της απλής γραμμικής παλινδρόμησης χρησιμοποιώντας το λογισμικό Minitab από τον υπολογιστή της εταιρείας. Έλαβε τα αποτελέσματα της παλινδρόμησης (πίνακας 6-14) και παρατήρησε, ότι οι t-τιμές που αντιστοιχούν στις εκτιμώμενες τιμές και οι κλίσεις (slope coefficients) είναι κατ' απόλυτη τιμή μεγάλες. Επιπλέον, οι p-τιμές είναι πολύ μικρές και άρα η εξίσωση της παλινδρόμησης είναι ξεκάθαρα σημαντική.

Προχωρώντας στο r^2, ο Gene απογοητεύτηκε, καθώς η τιμή του δεν ήταν τόσο υψηλή όσο, επιθυμούσε (64,2%). Ωστόσο, αποφάσισε ότι ήταν αρκετά μεγάλη, ώστε να ξεκινήσει να σκέφτεται τρόπους για να αυξήσει τα επίπεδα της ημερήσιας παραγωγής.

Regression Analysis: Y versus X

```
The regression equation is
Y = 552 - 8,91 X

Predictor     Coef  SE Coef      T      P
Constant   552,04    22,85  24,16  0,000
X          -8,911    1,453  -6,13  0,000
S = 59,4052   R-Sq = 64,2%   R-Sq(adj) = 62,5%

Analysis of Variance
Source           DF      SS      MS      F      P
Regression        1  132758  132758  37,62  0,000
Residual Error   21   74109    3529
Total            22  206866
```

Πίνακας 6-14 Απλή γραμμική παλινδρόμηση για την Butcher Products, INC

Ερωτήσεις

1. **Πόσες μονάδες παραγωγής θα προβλέπατε για μία μέρα με υψηλότερη θερμοκρασία 89 βαθμούς Fahrenheit;**

Σύμφωνα με την ανάλυση παλινδρόμησης που πραγματοποιήθηκε παραπάνω, η ευθεία που προσαρμόζεται καλύτερα στη συλλογή των Y και X δεδομένων, θα έχει τη μορφή Y = 552 – 8,91*X. Προκειμένου να βρούμε τις μονάδες παραγωγής τη μέρα με υψηλότερη θερμοκρασία τους 89 βαθμούς Fahrenheit (άρα με απόλυτη διαφορά μεταξύ του 89 και του 65, που αποτελεί το σημείο αναφοράς των βαθμών Fahrenheit, τους X = |89-65| = 24 βαθμούς), θα αντικαταστήσουμε: Y = 552 – 8,91*24 =

338,16 μονάδες παραγωγής.

2. **Πόσες μονάδες παραγωγής θα προβλέπατε για μία μέρα με υψηλότερη θερμοκρασία 41 βαθμούς Fahrenheit (διαφορά X = 24);**

Η ανεξάρτητη μεταβλητή που χρησιμοποιείται στο παράδειγμα 6-2, είναι η απόλυτη διαφορά μεταξύ της θερμοκρασίας της συγκεκριμένης μέρας και της θερμοκρασίας των 65 βαθμών Fahrenheit, που θεωρείται η ιδανική για συνθήκες εργασίας. Επομένως, η ανεξάρτητη μεταβλητή X θα πάρει σε αυτήν την περίπτωση την τιμή: X = |41-65| = 24. Όπως και στο ερώτημα 1, οι μονάδες παραγωγής θα είναι Y = 338,16.

3. **Βασιζόμενοι στα αποτελέσματα της ανάλυσης παλινδρόμησης που παρουσιάστηκε προηγουμένως, τι ενέργειες θα προτείνατε να κάνει ο Gene με σκοπό να αυξήσει την ημερήσια παραγωγή;**

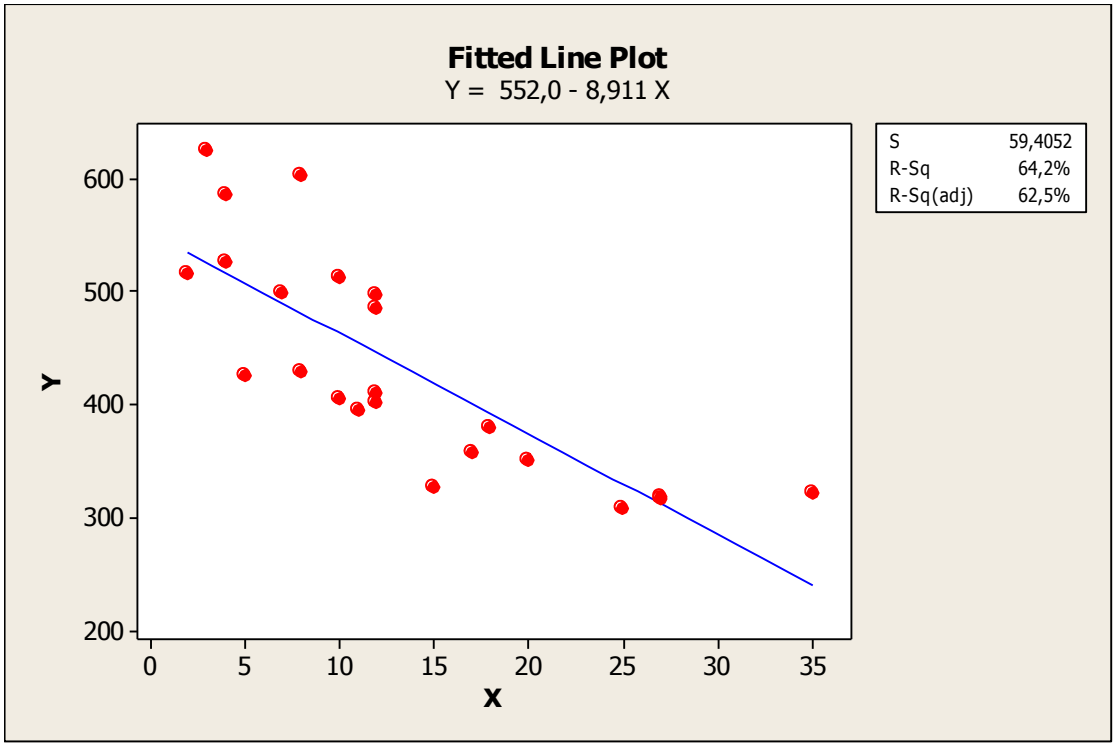
Αν και η επεξηγησιμότητα της παλινδρόμησης είναι περίπου 64%, μπορούμε να καταλάβουμε, ότι η σταθερότητα των εργασιακών περιβαλλοντικών συνθηκών είναι σημαντική, για την σταθερότητα της παραγωγής και μάλιστα σε υψηλότερα επίπεδα. Προτείνουμε την εγκατάσταση κλιματιστικών συστημάτων για την σταθεροποίηση της θερμοκρασίας στους χώρους του εργοστασίου, μετά από μία μελέτη απόδοσης της επένδυσης.

4. **Πιστεύετε ότι ο Gene ανέπτυξε ένα αποτελεσματικό εργαλείο πρόβλεψης;**

Μάλλον χρειάζονται περισσότερες από μία ανεξάρτητες μεταβλητές, έτσι ώστε να προβλέψει την εξαρτημένη μεταβλητή με ακρίβεια (αυτό φαίνεται και από το χαμηλό ποσοστό μεταβλητότητας, 35,8%). Πρέπει να γίνει έρευνα για να ανακαλυφθεί μία ακόμα ανεξάρτητη μεταβλητή που να σχετίζεται με τις μονάδες παραγωγής,

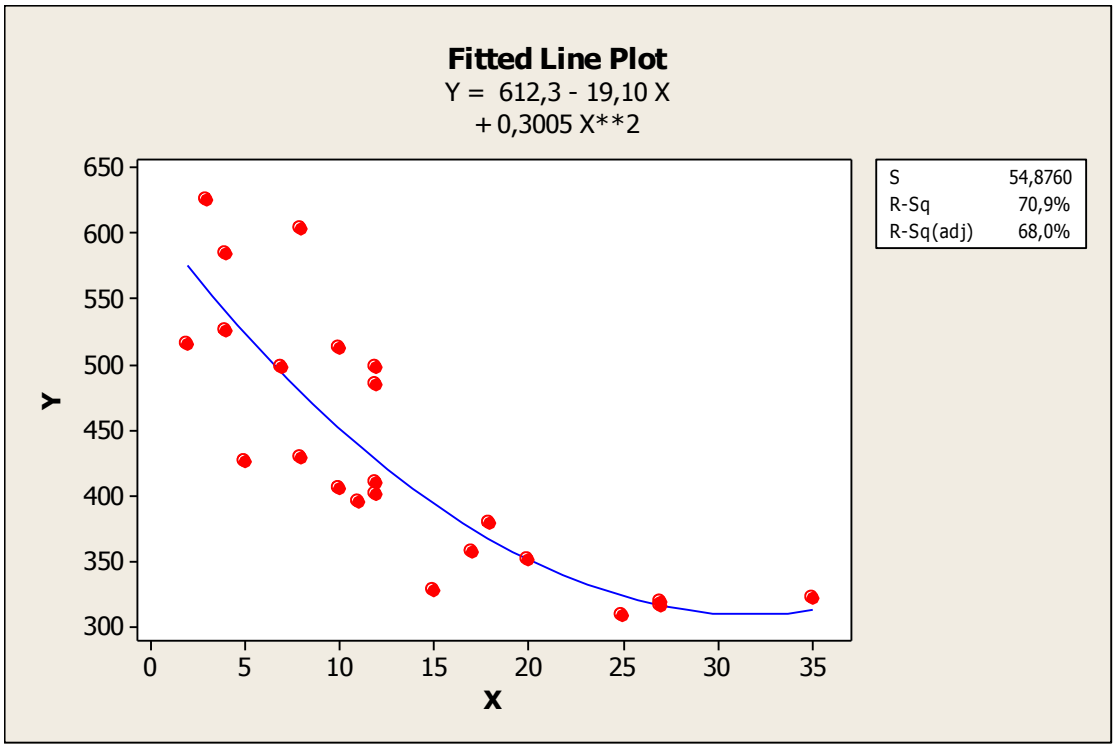
αλλά να μην σχετίζεται πολύ με την υπάρχουσα ανεξάρτητη μεταβλητή.

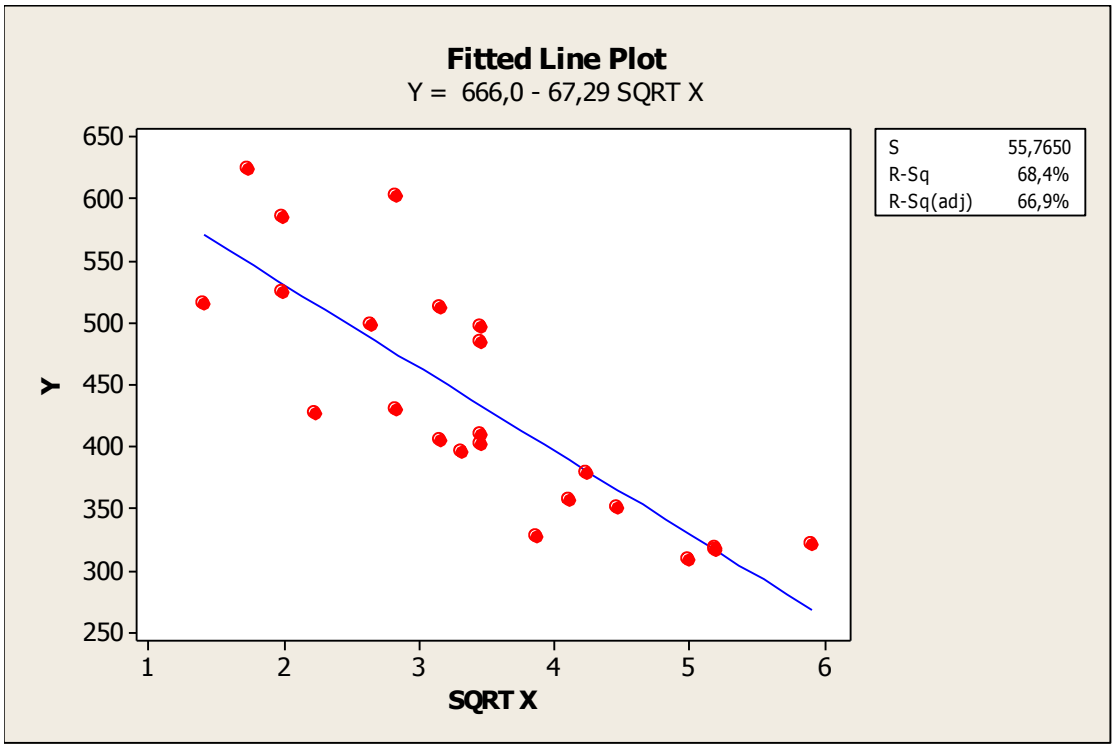
Από την ανάλυση παλινδρόμησης, που πραγματοποίησε ο Gene, διαπιστώνουμε πως μόνο το 64,2% της μεταβλητότητας στις μονάδες παραγωγής ερμηνεύεται από τη διαφορά θερμοκρασίας, ενώ ένα ποσοστό της τάξεως του 100 – 64,2 = 35,8% δεν μπορεί να ερμηνευτεί. Δηλαδή, γνωρίζοντας μόνο τη διαφορά θερμοκρασίας, μπορούμε να αξιοποιήσουμε το 35,8% αυτών που χρειάζεται, προκειμένου να προβλέψουμε σωστά την ποσότητα των πωλήσεων. Το ποσοστό αυτό είναι σημαντικό, περιορίζει τις δυνατότητες πρόβλεψης και δεν μπορεί να παραβλεφθεί. Το συμπέρασμα που εξάγεται είναι πώς άλλοι παράγοντες, που δεν έχουν ληφθεί υπόψη σε αυτή την ανάλυση παλινδρόμησης, ερμηνεύουν το ποσοστό αυτό. Άρα, για να γίνουν πιο ακριβείς οι προβλέψεις, πρέπει να βρεθεί και άλλη μεταβλητή πρόβλεψης, που θα επιτρέψει να ερμηνευθεί περισσότερη από τη συνολική μεταβολή.



Μπορούμε όμως να βελτιώσουμε το υπάρχον μοντέλο με

μετασχηματισμούς ή και να πάρουμε μη γραμμική σχέση ανάμεσα στα Y και X. Τέτοιοι μετασχηματισμοί του X είναι οι 1/X, logX, √X , X^2.

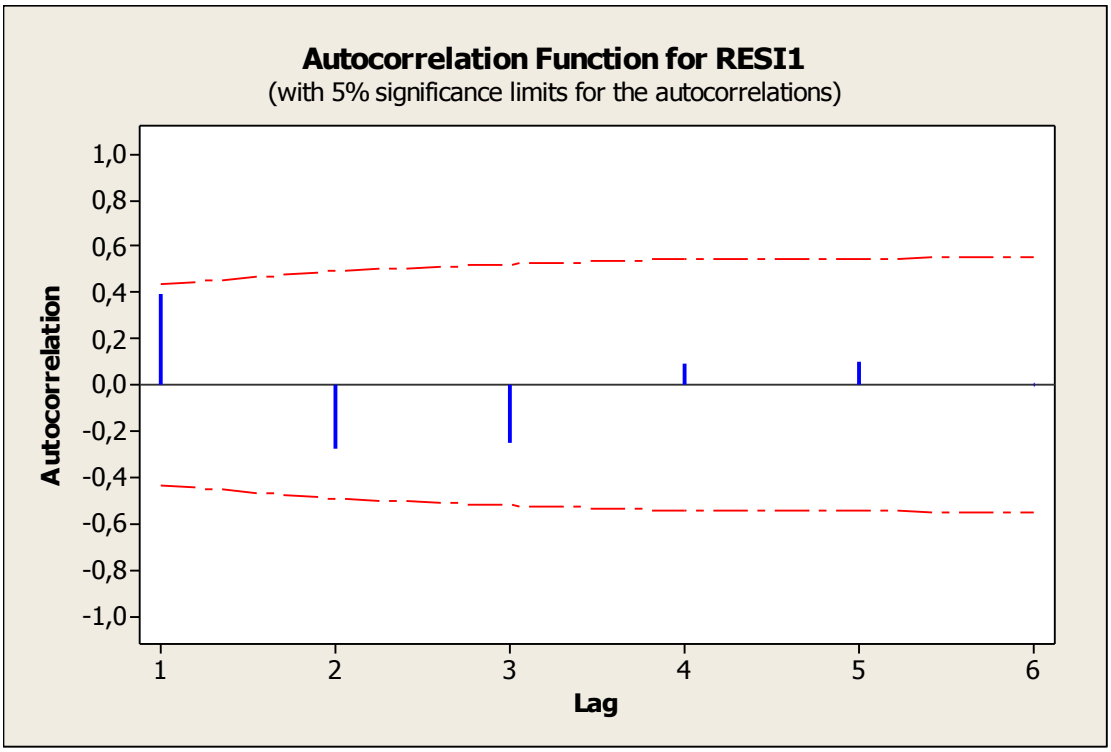




Μία μη γραμμική σχέση ανάμεσα στις μεταβλητές βελτιώνει αρκετά τον συντελεστή προσδιορισμού. Παραμένει όμως ένα ποσοστό

29% μη εξηγήσιμο.

Επίσης αν αντί Χ πάρουμε την τετραγωνική ρίζα έχουμε επίσης βελτίωση της επεξηγησιμότητας.



Διάγραμμα αυτοσυσχέτισης καταλοίπων της απλής γραμμικής παλινδρόμησης.

Lag	ACF	T	LBQ
1	0,389213	1,87	3,96
2	-0,277196	-1,16	6,06
3	-0,247982	-0,99	7,83
4	0,094175	0,36	8,10
5	0,099983	0,38	8,42
6	0,002771	0,01	8,42

Παράλληλα, μελετώντας τα κατάλοιπα της μεθόδου, διαπιστώνουμε, πως συνολικά είναι ελάχιστα παραπάνω από τα όρια εμπιστοσύνης, που θέτει η Q στατιστική και άρα τα κατάλοιπα δεν είναι ανεξάρτητα. Από τα παραπάνω συμπεραίνουμε, πως το εργαλείο πρόβλεψης, που χρησιμοποίησε ο Gene, δεν κρίνεται το πλέον

αποτελεσματικό.

ΚΕΦΑΛΑΙΟ 7 (ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΠΑΛΙΝΔΡΟΜΗΣΗ)

Στην απλή γραμμική παλινδρόμηση, ερευνήθηκε η σχέση μεταξύ μίας μόνο ανεξάρτητης μεταβλητής και μίας εξαρτημένης μεταβλητής. Η σχέση μεταξύ δύο μεταβλητών συχνά επιτρέπει σε κάποιον να προβλέψει με ακρίβεια την εξαρτημένη μεταβλητή, όταν έχει γνώση της ανεξάρτητης. Δυστυχώς, πολλές πραγματικές καταστάσεις πρόβλεψης δεν είναι τόσο απλές. Περισσότερες από μία ανεξάρτητες μεταβλητές είναι συχνά απαραίτητες, έτσι ώστε να προβλέψει κανείς την εξαρτημένη μεταβλητή με ακρίβεια. Τα μοντέλα παλινδρόμησης με περισσότερες από μία ανεξάρτητες μεταβλητές ονομάζονται μοντέλα πολυμεταβλητής παλινδρόμησης. Οι περισσότερες από τις έννοιες που παρουσιάστηκαν στην απλή γραμμική παλινδρόμηση μεταφέρονται και στην πολυμεταβλητή παλινδρόμηση. Ωστόσο, εισάγωνται και μερικές νέες ιδέες, επειδή χρησιμοποιούνται περισσότερες από μία ανεξάρτητες μεταβλητές για την πρόβλεψη της εξαρτημένης.

Η πολυμεταβλητή παλινδρόμηση περιλαμβάνει τη χρήση περισσότερων από μία ανεξάρτητων μεταβλητών για την πρόβλεψη μίας εξαρτημένης μεταβλητής.

7.1 ΜΕΡΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ ΠΡΟΒΛΕΨΗΣ

Ας θεωρήσουμε ένα παράδειγμα, κατά το οποίο μία εταιρία παραγωγής γάλακτος στην Αμερική θέλει να προβλέψει την ποσότητα γαλονιών γάλακτος, που πουλάει, ανάλογα με την τιμή, που χρεώνει το γαλόνι. Εφαρμόζοντας απλή γραμμική παλινδρόμηση στα δεδομένα που είχε συγκεντρώσει κατέληξε στο συμπέρασμα, ότι το πρόβλημα της πρόβλεψης δεν είναι ολοκληρωτικά ακριβές. Μπορεί να ερμηνευθεί μόνο

το 75% (r^2) των διαφορών στα γαλόνια του γάλακτος, που πωλούνται, χρησιμοποιώντας μόνο μία ανεξάρτητη μεταβλητή. Το 25% ($1-r^2$) της συνολικής μεταβλητότητας δεν μπορεί να ερμηνευθεί. Με άλλα λόγια, γνωρίζοντας μόνο την τιμή του προϊόντος, γνωρίζει κανείς το 75%, αυτών που χρειάζεται, προκειμένου να προβλέψει σωστά την ποσότητα των πωλήσεων. Για να γίνουν πιο ακριβείς οι προβλέψεις, πρέπει να βρεθεί και άλλη μεταβλητή πρόβλεψης, που θα επιτρέψει να ερμηνευθεί περισσότερη από τη συνολική μεταβλητότητα. Αν επιτευχθεί η μείωση της μη ερμηνευμένης μεταβλητότητας, οι προβλέψεις θα περιέχουν λιγότερη αβεβαιότητα και θα είναι πιο ακριβείς.

Πρέπει να γίνει έρευνα, για να ανακαλυφθεί μία ακόμα ανεξάρτητη μεταβλητή, που να σχετίζεται με την ποσότητα των πωλήσεων γαλονιών γάλακτος. Ωστόσο, αυτή η νέα ανεξάρτητη μεταβλητή (predictor) δεν μπορεί να σχετίζεται τόσο πολύ με την ανεξάρτητη μεταβλητή (τιμή ανά γαλόνι), που χρησιμοποιείται ήδη. Αν οι δυο ανεξάρτητες μεταβλητές σχετίζονται πολύ μεταξύ τους, θα ερμηνεύουν την ίδια μεταβολή και η προσθήκη της δεύτερης μεταβλητής δεν θα βελτιώσει την πρόβλεψη.

Η απλή λύση στο πρόβλημα των δύο υψηλά συσχετισμένων ανεξάρτητων μεταβλητών είναι μερικώς το να μη χρησιμοποιούνται και οι δύο μαζί. Το πρόβλημα αυτό (multicollinearity) θα συζητηθεί αργότερα σε αυτό το κεφάλαιο.

7.2 ΠΙΝΑΚΑΣ ΣΥΣΧΕΤΙΣΗΣ

Η εταιρία αποφασίζει, ότι και τα διαφημιστικά έξοδα μπορεί να βοηθήσουν στο να βελτιωθεί η πρόβλεψη των εβδομαδιαίων πωλήσεων. Η διερεύνηση της σχέσης μεταξύ των διαφημιστικών εξόδων, του όγκου των πωλήσεων και της τιμής ανά γαλόνι, γίνεται μέσω ενός πίνακα

συσχέτισης (correlation matrix). Ο πίνακας συσχέτισης κατασκευάζεται υπολογίζοντας τους συντελεστές απλής συσχέτισης για κάθε συνδυασμό ζευγαριών των μεταβλητών.

Ένα παράδειγμα του πίνακα συσχέτισης παρουσιάζεται στον ακόλουθο Πίνακα. Ο συντελεστής συσχέτισης που υποδεικνύει τη σχέση μεταξύ των μεταβλητών 1 και 2 συμβολίζεται με r_{12}. Να σημειωθεί ότι το πρώτο ψηφίο 1, επίσης αναφέρεται και στη σειρά ενώ το δεύτερο 2, αναφέρεται στη στήλη του πίνακα. Αυτή η προσέγγιση επιτρέπει σε κάποιον να καθορίσει με μία ματιά, τη σχέση μεταξύ δύο μεταβλητών. Φυσικά, η συσχέτιση μεταξύ, για παράδειγμα, των μεταβλητών 1 και 2 είναι ακριβώς η ίδια με τη συσχέτιση των μεταβλητών 2 και 1. Γι' αυτό ισχύει: $r_{12}=r_{21}$. Έτσι, το μισό του πίνακα είναι απαραίτητο. Επιπλέον, η συσχέτιση μιας μεταβλητής με τον εαυτό της είναι πάντα 1, έτσι για παράδειγμα, $r_{11}=r_{22}=r_{33}=1$.

Η εταιρία τρέχει τα δεδομένα της στον υπολογιστή, και ο πίνακας συσχέτισης φαίνεται στον παρακάτω Πίνακα. Μία διερεύνηση της σχέσης μεταξύ των διαφημιστικών εξόδων, του όγκου των πωλήσεων και της τιμής ανά γαλόνι, δείχνει ότι η νέα ανεξάρτητη μεταβλητή πρέπει να συνεισφέρει σε βελτιωμένη πρόβλεψη. Ο πίνακας συσχέτισης δείχνει ότι τα διαφημιστικά έξοδα έχουν υψηλή θετική σχέση ($r_{13}=0,89$) με την εξαρτημένη μεταβλητή, τον όγκο των πωλήσεων, και μία μέτρια αρνητική σχέση ($r_{23}=-0,65$) με την ανεξάρτητη μεταβλητή, τη τιμή ανά γαλόνι. Αυτός ο συνδυασμός σχέσεων πρέπει να επιτρέψει στα διαφημιστικά έξοδα να ερμηνεύουν κάποια από τη συνολική μεταβολή η οποία δεν ερμηνεύεται ήδη από την τιμή ανά γαλόνι. Όπως θα φανεί και στη συνέχεια, όταν χρησιμοποιηθούν ταυτόχρονα και η τιμή ανά γαλόνι και τα διαφημιστικά έξοδα για την εκτίμηση του όγκου των πωλήσεων,

το R^2 αυξάνει στο 93,2%.

Η ανάλυση του πίνακα συσχέτισης είναι ένα σημαντικό αρχικό βήμα στη λύση κάθε προβλήματος που περιλαμβάνει πολλαπλές ανεξάρτητες μεταβλητές.

Πίνακας Συσχέτισης

Μεταβλητές	**1**	**2**	**3**
1	r_{11}	r_{12}	r_{13}
2	r_{21}	r_{22}	r_{23}
3	r_{31}	r_{32}	r_{33}

Πίνακας Συντελεστή Συσχέτισης για τα Δεδομένα

Μεταβλητές	**Πωλήσεις (1)**	**Αξία (2)**	**Διαφήμιση (3)**
Πωλήσεις (1)	1,000	-0,863	0,891
Αξία (2)		1,000	-0,654
Διαφήμιση (3)			1,000

7.3 ΜΟΝΤΕΛΟ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Στην απλή παλινδρόμηση, η εξαρτημένη μεταβλητή συμβολίζεται με Y και η ανεξάρτητη μεταβλητή με X. Στην ανάλυση πολυμεταβλητής παλινδρόμησης, τα X με κάτω δείκτες χρησιμοποιούνται για να συμβολίσουν τις ανεξάρτητες μεταβλητές. Η εξαρτημένη μεταβλητή ακόμα συμβολίζεται με Y, και οι ανεξάρτητες μεταβλητές με X_1, X_2, ..., X_k. Όταν καθοριστεί τα αρχικό σύνολο ανεξάρτητων μεταβλητών, η σχέση μεταξύ του Y και των X μπορεί να εκφραστεί σαν **μοντέλο πολυμεταβλητής παλινδρόμησης**.

Στο μοντέλο πολυμεταβλητής παλινδρόμησης, η μέση παρατηρούμενη τιμή είναι η ακόλουθη γραμμική συνάρτηση των

ανεξάρτητων μεταβλητών:

$$\mu_{\gamma} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Αυτή η έκφραση είναι η **συνάρτηση πολυμεταβλητής παλινδρόμησης του πληθυσμού**. Όπως και με την απλή γραμμική παλινδρόμηση, δεν μπορούμε απ' ευθείας να παρατηρήσουμε τη συνάρτηση παλινδρόμησης πληθυσμού, επειδή οι παρατηρούμενες τιμές του Y διαφέρουν από τη μέση τιμή. Υποθέτουμε ότι τα Y είναι κανονικά κατανεμημένα γύρω από τους μέσους τους με την ίδια τυπική απόκλιση, σ.

Τα δεδομένα για την απλή γραμμική παλινδρόμηση αποτελούνται από τις παρατηρήσεις (X_i, Y_i) των δύο μεταβλητών. Στην πολυμεταβλητή παλινδρόμηση, τα δεδομένα για κάθε περίπτωση αποτελούνται από μία παρατήρηση της εξαρτημένης μεταβλητής και μία παρατήρηση για κάθε μία από τις ανεξάρτητες μεταβλητές. Η i παρατήρηση της j μεταβλητής πρόβλεψης συμβολίζεται με X_{ij}. Με αυτό το συμβολισμό, τα δεδομένα για την πολυμεταβλητή παλινδρόμηση έχουν τη μορφή του παρακάτω Πίνακα. Είναι βολικό να αναφερόμαστε στα δεδομένα για την περίπτωση i ως απλά την i παρατήρηση. Με αυτό τον τρόπο, n θα είναι ο αριθμός των παρατηρήσεων και k ο αριθμός των μεταβλητών πρόβλεψης.

Δομή Δεδομένων για πολυμεταβλητή παλινδρόμηση

Μεταβλητές Πρόβλεψης					**Απόκριση**
Περίπτωση	$\mathbf{X_1}$	$\mathbf{X_2}$	**…….**	$\mathbf{X_k}$	**Y**
1	X_{11}	X_{12}	…….	X_{1k}	Y_1
2	X_{21}	X_{22}	…….	X_{2k}	Y_2
.	.	.	…….	.	.
.	.	.	…….	.	.
i	X_{i1}	X_{i2}	…….	X_{ik}	Y_i
.	.	.	…….	.	.

.	.	.	…….	.	.
n	X_{n1}	X_{n2}	.....	X_{nk}	Y_n

7.3.1 ΣΤΑΤΙΣΤΙΚΟ ΜΟΝΤΕΛΟ ΓΙΑ ΤΗΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η παρατηρούμενη τιμή (response), Y, είναι η τυχαία μεταβλητή, που σχετίζεται με τις ανεξάρτητες (predictor) μεταβλητές, X_1, X_2, ..., X_k με τη σχέση:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

όπου:

1. Για κάθε i παρατήρηση, $Y=Y_i$ και τα X_1, X_2, ..., X_k αντιστοιχούν στις τιμές X_{i1}, X_{i2}, ..., X_{ik}. Για τις X_1, X_2, ..., X_k δεν υπάρχουν γραμμικές σχέσεις μεταξύ τους.
2. Τα ε, είναι συντελεστές σφάλματος, που αντιπροσωπεύουν τις αποκλίσεις της προβλεπόμενης τιμής από την αληθή σχέση. Είναι μη παρατηρημένες τυχαίες μεταβλητές, οι οποίες υπολογίζουν άλλους παράγοντες, που έχουν επίδραση στην απόκριση. Τα σφάλματα υποτίθεται, ότι είναι ανεξάρτητα, τυχαία και κανονικά κατανεμημένα με μέσο 0, $(E(\varepsilon)=0)$ και άγνωστη σταθερή τυπική απόκλιση σ, $(E(\varepsilon^2)=\sigma^2)$ ή $\varepsilon=N(0,\sigma^2)$.
3. Οι συντελεστές παλινδρόμησης, β_0, β_1,..., β_k, που μαζί ορίζουν τη συνάρτηση παλινδρόμησης είναι άγνωστοι, όπως συνήθως και η διασπορά.

Αν κάποιος έχει στη διάθεσή του τα δεδομένα, μπορεί να υπολογίσει τους συντελεστές παλινδρόμησης χρησιμοποιώντας την αρχή των ελαχίστων τετραγώνων μηδενίζοντας τις μερικές παραγώγους. Οι εκτιμήσεις των ελαχίστων τετραγώνων συμβολίζονται με b_0, b_1,…, b_k και η εκτιμώμενη συνάρτηση παλινδρόμησης με:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

Τα κατάλοιπα, $e = Y - \hat{Y}$, είναι εκτιμήσεις του συντελεστή σφάλματος και παρόμοια με την περίπτωση της απλής γραμμικής παλινδρόμησης. Η αντιστοίχιση μεταξύ πληθυσμού και δείγματος είναι:

$$\textit{Πληθυσμός} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

$$\textit{Δείγμα} \quad Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + e.$$

Η αναμενόμενες τιμές των συντελεστών θα είναι $E(b_i) = \beta_i$. Οι υπολογισμοί στην ανάλυση πολυμεταβλητής παλινδρόμησης συνήθως γίνονται με τη χρήση προγραμμάτων, όπως είναι το Minitab, SAS, SPSS, EVIEWS και το Excel.

Παράδειγμα (Τ.7-4)

Για τα δεδομένα θεωρούμε ένα μοντέλο πολυμεταβλητής παλινδρόμησης που συσχετίζει τον όγκο των πωλήσεων (Y σε χιλιάδες γαλόνια) με την τιμή (X_1 σε $ ανά γαλόνι) και τη διαφήμιση (X_2 σε εκατοντάδες $):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Καθορίζουμε την προσαρμοσμένη συνάρτηση παλινδρόμησης:

$$\hat{Y} = 16{,}41 - 8{,}25 X_1 + 0{,}59 X_2$$

Οι τιμές των ελαχίστων τετραγώνων, (b_0=16,41, b_1=-8,25 και b_2=0,59), ελαχιστοποιούν το άθροισμα των τετραγωνικών σφαλμάτων:

$$SSE = \sum_i (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2})^2 = \sum_i (Y_i - \hat{Y}_i)^2$$

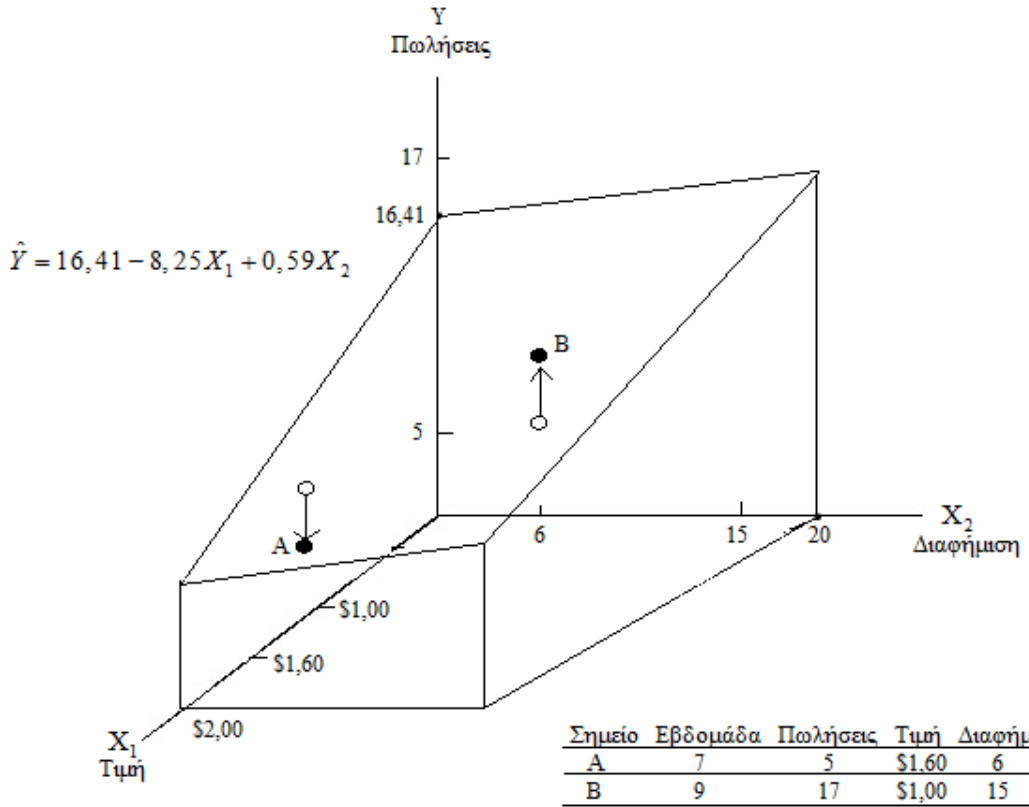
για όλες τις πιθανές επιλογές των b_0, b_1 και b_2. Εδώ, η συνάρτηση που προσαρμόζεται καλύτερα φαίνεται στο Γράφημα. Τα σημεία δεδομένων βρίσκονται στις 3 διαστάσεις στους άξονες Y, X_1 και X_2. Τα σημεία βρίσκονται πάνω και κάτω από το επίπεδο με τέτοιο τρόπο, έτσι ώστε το

SSE να είναι ελάχιστο.

Η προσαρμοσμένη συνάρτηση παλινδρόμησης μπορεί να χρησιμοποιηθεί για να προβλέψει τις πωλήσεις της επόμενης εβδομάδας. Αν η τιμή ανά γαλόνι είναι \$1,50 και τα έξοδα διαφήμισης \$1.000 (στη συνάρτηση στη θέση του X_2 βάζουμε τα έξοδα διαφήμισης σε εκατοντάδες), η πρόβλεψη είναι 9,935 χιλιάδες γαλόνια γάλακτος. Δηλαδή,

$$\hat{Y} = 16{,}41 - 8{,}25\,X_1 + 0{,}59\,X_2$$

$$\hat{Y} = 16{,}41 - 8{,}25(1{,}5) + 0{,}59(10) = 9{,}935$$



Σημείο	Εβδομάδα	Πωλήσεις	Τιμή	Διαφήμιση
A	7	5	\$1.60	6
B	9	17	\$1,00	15

7.4 ΕΡΜΗΝΕΙΑ ΣΥΝΤΕΛΕΣΤΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Στη συνέχεια θα ερμηνεύσουμε τα b_0, b_1, και b_2 στην προσαρμοσμένη συνάρτηση παλινδρόμησης. Η τιμή b_0 είναι εδώ πάλι η

Y-τομή (τεταγμένη), δηλαδή μπορεί να ερμηνευθεί σαν τη τιμή του $\hat{Y}$ όταν τα X_1 και X_2 είναι μηδέν. Οι συντελεστές b_1 και b_2 αναφέρονται ως συντελεστές μερικής παλινδρόμησης (partial or net regression coefficients). Ο κάθε ένας υπολογίζει τη μέση αλλαγή στο Y για κάθε μονάδα αλλαγής στις σχετικές ανεξάρτητες μεταβλητές. Ωστόσο, επειδή η ταυτόχρονη επίδραση όλων των ανεξάρτητων μεταβλητών στο Y υπολογίζεται από τη συνάρτηση παλινδρόμησης, η μερική (partial or net) επίδραση του X_1 (ή οποιουδήποτε άλλου X) πρέπει να υπολογιστεί ξεχωριστά από την επίδραση των υπόλοιπων μεταβλητών. Έτσι, λέγεται ότι το b_1 υπολογίζει τη μέση αλλαγή στο Y για κάθε μονάδα αλλαγής στο X_1, κρατώντας όλες τις άλλες ανεξάρτητες μεταβλητές σταθερές.

Ο συντελεστής μερικής παλινδρόμησης (partial, or net, regression coefficient) υπολογίζει τη μέση αλλαγή στην εξαρτημένη μεταβλητή για κάθε μονάδα αλλαγής στην αντίστοιχη ανεξάρτητη μεταβλητή, κρατώντας όλες τις υπόλοιπες ανεξάρτητες μεταβλητές σταθερές.

Στο συγκεκριμένο παράδειγμα, το b_1 έχει τιμή -8,25 και υποδεικνύει, ότι κάθε αύξηση 1 cent στην τιμή του γαλονιού γάλακτος, όταν τα διαφημιστικά έξοδα μένουν σταθερά, μειώνει τον όγκο των πωλήσεων κατά μέσο όρο 82,5 γαλόνια. Παρομοίως, το b_2 έχει τιμή 0,59, το οποίο σημαίνει, ότι αν αυξηθούν κατά $100 τα διαφημιστικά έξοδα και η τιμή ανά γαλόνι μείνει σταθερή, τότε ο όγκος των πωλήσεων θα αυξηθεί κατά μέσο όρο 590 γαλόνια.

Παράδειγμα (Τ.7-4)

Για να εξηγήσουμε την οριακή επίδραση των μεμονωμένων X στην απόκριση, θεωρούμε την κατάσταση, στην οποία η τιμή είναι $1 ανά γαλόνι και $1000 ξοδεύονται για διαφήμιση. Τότε:

$$\hat{Y} = 16{,}41 - 8{,}25\, X_1 + 0{,}59\, X_2$$

$$=16{,}41-8{,}25(1)+0{,}59(10)$$

$$=16{,}41-8{,}25+5{,}9=14{,}06$$

Η πρόβλεψη είναι, ότι οι πωλήσεις θα είναι 14.060 γαλόνια γάλακτος.

Ποια είναι η επίδραση στις πωλήσεις αν η τιμή αυξηθεί κατά 1 cent και αν τα διαφημιστικά έξοδα παραμείνουν σταθερά;

$$\hat{Y} = 16{,}41 - 8{,}25\, X_1 + 0{,}59\, X_2$$

$$=16{,}41-8{,}25(1{,}01)+0{,}59(10)=16{,}41-8{,}3325+5{,}9=13{,}9775$$

Οι πωλήσεις μειώνονται κατά 82,5 γαλόνια (14,06-13,9775=0,0825).

Ποια είναι η επίδραση στις πωλήσεις μιας αύξησης $100 στα διαφημιστικά έξοδα, αν η τιμή ανά γαλόνι παραμείνει σταθερά στο $1;

$$\hat{Y}=16{,}41-8{,}25(1)+0{,}59(11)=16{,}41-8{,}25+6{,}49=14{,}65.$$

Να σημειωθεί ότι οι πωλήσεις αυξήθηκαν κατά 590 γαλόνια (14,65-14,06=0,59).

7.5 ΣΥΜΠΕΡΑΣΜΑΤΑ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Τα συμπεράσματα για την πολυμεταβλητή παλινδρόμηση είναι ανάλογα με αυτά για την απλή γραμμική παλινδρόμηση. Οι εκτιμήσεις ελαχίστων τετραγώνων των παραμέτρων του μοντέλου, τα εκτιμώμενα τυπικά σφάλματα, η στατιστική t, που χρησιμοποιήθηκαν για να διερευνήσουν τη σημαντικότητα των μεμονωμένων όρων στο μοντέλο παλινδρόμησης και η στατιστική F, που χρησιμοποιήθηκε για να ελέγξει τη σημαντικότητα της παλινδρόμησης, όλα παρέχονται στην ανάλυση

από τα πακέτα τυπικού στατιστικού λογισμικού. Το να καθορίσει κανείς αυτές τις ποσότητες στο χέρι για μία ανάλυση πολυμεταβλητής παλινδρόμησης οποιουδήποτε μεγέθους δεν είναι πρακτικό και πρέπει να χρησιμοποιηθεί υπολογιστής για τους υπολογισμούς.

Όπως και στο Κεφάλαιο 6, κάθε παρατήρηση Y μπορεί να γραφεί ως:

Observation = Fit + Residual

$$\text{ή} \quad Y = \hat{Y} + (Y - \hat{Y})$$

$$\text{και με} \quad \hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

να είναι η προσαρμοσμένη εξίσωση παλινδρόμησης και με το $\hat{Y}$ να είναι μία εκτίμηση της συνάρτησης παλινδρόμησης του πληθυσμού και αντιπροσωπεύει το μέρος του Y, που ερμηνεύεται από τη σχέση του Y με τα X. Τα κατάλοιπα, $Y - \hat{Y}$, είναι η εκτίμηση των συντελεστών σφάλματος του μοντέλου και αντιπροσωπεύουν εκείνο το ποσοστό μεταβλητότητας του Y, που δεν ερμηνεύεται από τις μεταβλητές πρόβλεψης.

Η ανάλυση του αθροίσματος των τετραγώνων και οι σχετικοί βαθμοί ελευθερίας είναι:

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})^2$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\text{n-1} = \text{k} + \text{n-k-1}$$

Η συνολική μεταβολή στην απόκριση, SST, αποτελείται από δύο συστατικά: το SSR, τη μεταβολή που ερμηνεύεται από τις μεταβλητές πρόβλεψης μέσω της εκτιμώμενης συνάρτησης παλινδρόμησης και το SSE, την μη ερμηνευμένη ή μεταβολή σφάλματος. Οι πληροφορίες στην προηγούμενη εξίσωση μπορούν να φανούν στην ανάλυση του πίνακα

διασποράς (ANOVA), που θα συζητηθεί στη συνέχεια.

7.5.1 ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΕΚΤΙΜΗΣΗΣ

Το τυπικό σφάλμα εκτίμησης είναι η τυπική απόκλιση των καταλοίπων. Υπολογίζει την τυπική διασπορά των Y τιμών γύρω από τη προσαρμοσμένη συνάρτηση παλινδρόμησης. Το τυπικό σφάλμα εκτίμησης είναι:

$$s_{y\cdot x's}=\sqrt{\frac{\sum (Y-\hat{Y})^2}{n-k-1}}=\sqrt{\frac{SSE}{n-k-1}}=\sqrt{MSE}$$

όπου:

- n = ο αριθμός των παρατηρήσεων.
- k = ο αριθμός των ανεξάρτητων μεταβλητών στη συνάρτηση παλινδρόμησης.
- SSE = $\sum (Y-\hat{Y})^2$ το άθροισμα τετραγώνων των καταλοίπων.
- MSE = SSE/(n-k-1) ο τετραγωνικός μέσος των καταλοίπων (the residual mean square).

Το τυπικό σφάλμα εκτίμησης είναι η τυπική απόκλιση των καταλοίπων. Υπολογίζει την ποσότητα, κατά την οποία οι πραγματικές τιμές (Y) διαφέρουν από τις εκτιμούμενες τιμές $(\hat{Y})$. Για σχετικά μεγάλα δείγματα, περιμένουμε περίπου το 67% των διαφορών $(Y-\hat{Y})$ να είναι σε ένα διάστημα $0\pm s_{y\cdot x's}$ από το 0 και περίπου το 95% αυτών των διαφορών να είναι σε ένα διάστημα $0\pm 2\,x\,s_{y\cdot x's}$ από το 0.

Το $s_{y\cdot x's}$ είναι ένας άριστος εκτιμητής του σ, διότι όπως αποδεικνύεται η αναμενόμενη τιμή του $s^2_{y\cdot x's}$ είναι το σ^2, η διασπορά των σφαλμάτων. Δηλαδή:

$$E\left(s^2_{y\cdot x's}\right)=\sigma^2=\frac{\left(\sum Y-\hat{Y}\right)^2}{n}$$

Γενικά για τον άριστο εκτιμητή της διασποράς, σ, μπορούμε να εφαρμόσουμε τον κάτωθι κανόνα:

Ενώ ο παρονομαστής για τη διασπορά είναι n, ο αντίστοιχος αριθμός για τον εκτιμητή είναι πάντα οι βαθμοί ελευθερίας του συστήματος δηλαδή του αριθμού των μεταβλητών (παρατηρήσεων) μείον των σχέσεων, που δυνητικά τις συνδέουν.

Για τη δειγματική διασπορά η σχέση είναι ο μέσος των τετραγωνικών διαφορών και οι βαθμοί ελευθερίας θα είναι n-1. Για την απλή γραμμική παλινδρόμηση έχουμε δύο σχέσεις: b_0 και b_1 και οι βαθμοί ελευθερίας θα είναι n-2. Για την πολυμεταβλητή παλινδρόμηση οι σχέσεις είναι ο αριθμός των ανεξάρτητων μεταβλητών k συν τον σταθερό όρο, άρα k+1. Οπότε οι βαθμοί ελευθερίας θα είναι εδώ n-k-1.

Παράδειγμα (Τ.7-4)

Το τυπικό σφάλμα εκτίμησης για το προηγούμενο παράδειγμα είναι:

$$s_{y\cdot x's}=\sqrt{\frac{15,9}{10-2-1}}=\sqrt{2,27}=1,51$$

Με μία μόνο μεταβλητή πρόβλεψης, X_1=τιμή, το τυπικό σφάλμα εκτίμησης υπολογίστηκε να είναι 2,72. Με την επιπλέον μεταβλητή πρόβλεψης, X_2=διαφήμιση, το τυπικό σφάλμα εκτίμησης έχει μειωθεί κατά 50% περίπου. Οι διαφορές μεταξύ των πραγματικών ποσοτήτων γάλακτος, που πουλήθηκαν και των προβλέψεων, που αποκτήθηκαν από την προσαρμοσμένη συνάρτηση παλινδρόμησης, είναι σημαντικά μικρότερες με δύο μεταβλητές πρόβλεψης, από ότι ήταν με μία μόνο μεταβλητή πρόβλεψης. Αυτό σημαίνει, ότι η εξίσωση με δύο μεταβλητές

πρόβλεψης έρχεται πολύ πιο κοντά στο να παράγει τα πραγματικά Y, από ότι η εξίσωση με μία μεταβλητή πρόβλεψης.

7.5.2 ΣΗΜΑΝΤΙΚΟΤΗΤΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Ο πίνακας ANOVA βασίζεται στην ανάλυση της συνολικής μεταβολής του Y (SST) σε ερμηνευμένο (SSR) και μη ερμηνευμένο (SSE) μέρος και δίνεται στον παρακάτω Πίνακα.

ANOVA πίνακας για πολυμεταβλητή παλινδρόμηση

Source	**Sum of Squares**	**df**	**Mean Square**	**F Ratio**
Regression	SSR	k	MSR=SSR/k	*F=MSR/MSE*
Error	SSE	n-k-1	MSE=SSE/(n-k-1)	
Total	SST	n-1		

<u>F έλεγχος για τη σημαντικότητα της παλινδρόμησης</u>

Στο μοντέλο πολυμεταβλητής παλινδρόμησης, οι υποθέσεις:

H_0: $\beta_1=0, \beta_2=0, \ldots, \beta_k=0$ (να μην υπάρχει παλινδρόμηση) και

H_1: τουλάχιστον ένα $\beta_j\neq 0$ (να υπάρχει παλινδρόμηση)

ελέγχονται από τον F λόγο:

$$F=\frac{MSR}{MSE},$$

που ακολουθεί την κατανομή F με df=k, n-k-1. Στο επίπεδο σημαντικότητας α, η περιοχή απόρριψης είναι:

$$F>F_\alpha$$

όπου F_α είναι το άνω α-εκατοστιαίο σημείο (ποσοστό) της F-κατανομής με $\delta_1=k$, $\delta_2=n-k-1$ βαθμούς ελευθερίας.

Ο συντελεστής προσδιορισμού R^2 δίνεται από τον τύπο:

$$R^2=\frac{SSR}{SST}=\frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2}=1-\frac{SSE}{SST}=1-\frac{\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2}$$

και έχει την ίδια μορφή και ερμηνεία, όπως το r^2 στην απλή γραμμική παλινδρόμηση. Αντιπροσωπεύει την αναλογία της μεταβλητότητας της απόκρισης Y, που ερμηνεύεται από τη σχέση του Y με τα X.

Μία τιμή του $R^2=1$ σημαίνει, ότι όλες οι παρατηρούμενες Y τιμές πέφτουν ακριβώς πάνω στην προσαρμοσμένη συνάρτηση παλινδρόμησης, δηλαδή στην επιφάνεια. Όλη η μεταβλητότητα στην απόκριση ερμηνεύεται από την παλινδρόμηση. Μία τιμή του $R^2=0$ σημαίνει ότι $\hat{Y}=\bar{Y}$, $(SSR=0)$ και ότι καμία μεταβολή στο Y δεν ερμηνεύεται από την παλινδρόμηση. Στην πράξη, $0\leq R^2\leq 1$, και η τιμή του R^2 πρέπει να ερμηνευθεί σχετικά με τις ακραίες τιμές, 0 και 1.

Η ποσότητα $R=\sqrt{R^2}$ καλείται **συντελεστής πολυμεταβλητής συσχέτισης** και είναι η συσχέτιση μεταξύ των αποκρίσεων-παρατηρούμενων τιμών Y και των προσαρμοσμένων τιμών $\hat{Y}$.

Για πολυμεταβλητή παλινδρόμηση,

$$F=\frac{R^2}{1-R^2}\left(\frac{n-k-1}{k}\right)$$

και έτσι, όταν n και k είναι σταθερά, σημαντικές παλινδρομήσεις (μεγάλοι F λόγοι) σχετίζονται με σχετικά υψηλές τιμές του R^2.

Ο συντελεστής προσδιορισμού (R^2) μπορεί πάντα να αυξηθεί προσθέτοντας μία επιπλέον ανεξάρτητη μεταβλητή, X, στη συνάρτηση παλινδρόμησης, ακόμα κι αν η επιπλέον μεταβλητή δεν είναι στατιστικά σημαντική. Για αυτό το λόγο, μερικοί αναλυτές προτιμούν να ερμηνεύουν το R^2 προσαρμοσμένο για τον αριθμό των όρων στη συνάρτηση παλινδρόμησης. Ο προσαρμοσμένος συντελεστής προσδιορισμού (adjusted coefficient of determination), $\overline{R}^2$, δίνεται από τον τύπο:

$$\bar{R}^2 = R^2(adj) = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

Στην περίπτωση που το k είναι 0 ή το n είναι πολύ μεγάλο, τότε οι δύο συντελεστές συγκλίνουν.

Παράδειγμα(Τ.7-4)

Για το προηγούμενο παράδειγμά μας έχουμε:

$$SST = SSR + SSE$$

$$233,6 = 217,7 + 15,9$$

Οπότε:

$$R^2 = \frac{217,7}{233,6} = 1 - \frac{15,9}{233,6} = 0,932$$

και ο συντελεστής πολυμεταβλητής συσχέτισης είναι R = 0,965.

Εδώ, περίπου το 93% της μεταβολής στον όγκο των πωλήσεων ερμηνεύεται από την παλινδρόμηση, από τη σχέση των πωλήσεων με τη τιμή και τα διαφημιστικά έξοδα. Επιπλέον, η συσχέτιση μεταξύ των πωλήσεων και των προσαρμοσμένων πωλήσεων είναι περίπου 0,965, το οποίο υποδεικνύει στενή συμφωνία μεταξύ των πραγματικών και των προβλεφθέντων τιμών. Μία σύνοψη της ανάλυσης των δεδομένων φαίνεται στον ακόλουθο Πίνακα:

Μεταβλητές	**R^2**
Τιμή	0,75
Τιμή και διαφημιστικά έξοδα	0,93

7.5.3 ΜΕΜΟΝΩΜΕΝΕΣ ΜΕΤΑΒΛΗΤΕΣ ΠΡΟΒΛΕΨΗΣ

Ο συντελεστής ενός μεμονωμένου X στη συνάρτηση παλινδρόμησης υπολογίζει τη μερική, ή οριακή επίδραση εκείνου του X στην απόκριση, Y, κρατώντας όλα τα υπόλοιπα X στη εξίσωση σταθερά. Αν η παλινδρόμηση κριθεί σημαντική, τότε έχει ενδιαφέρον να ερευνηθεί

η σημαντικότητα των μεμονωμένων μεταβλητών πρόβλεψης. Το ζήτημα είναι το εξής: Αν μας δοθούν τα υπόλοιπα X, είναι η επίδραση του συγκεκριμένου X σημαντική, ή μπορεί αυτός ο X όρος να φύγει από την εξίσωση παλινδρόμησης; Αυτή η ερώτηση μπορεί να απαντηθεί εξετάζοντας μία κατάλληλη t τιμή.

Αν η $H_0: \beta_j = 0$ είναι αληθής, η στατιστική t, με την τιμή $t = b_j / s_{bj}$, ακολουθεί την t κατανομή με df=n-k-1, με s_{bj} τις τυπικές αποκλίσεις των συντελεστών b_j.

Για να κρίνουμε τη σημαντικότητα του j όρου, j=0, 1,…, k, στην εξίσωση παλινδρόμησης, ο έλεγχος του στατιστικού, t, συγκρίνεται με ένα ποσοστιαίο σημείο της t κατανομής με n-k-1 βαθμούς ελευθερίας.

Για ένα α επίπεδο σημαντικότητας των υποθέσεων:

$$H_0: \beta_j = 0 \quad \text{και}$$

$$H_1: \beta_j \neq 0$$

απορρίπτεται η H_0 αν το $|t| > t_{\alpha/2}$. Εδώ, το $t_{\alpha/2}$ είναι το άνω α/2 ποσοστιαίο σημείο μιας t κατανομής με df=n-k-1.

Λίγη προσοχή πρέπει να δοθεί στη διαγραφή από την εξίσωση παλινδρόμησης εκείνων των μεταβλητών πρόβλεψης, που έχουν κριθεί μη σημαντικές από τον t έλεγχο (η $H_0: \beta_j = 0$ δεν μπορούσε να απορριφθεί). Αν τα X συσχετίζονται (multicollinear), οι συντελεστές ελαχίστων τετραγώνων και οι αντίστοιχες t τιμές μπορούν να αλλάξουν, μερικές φορές αρκετά, αν ένα μόνο X διαγραφεί από τη συνάρτηση παλινδρόμησης. Για παράδειγμα ένα X, που προηγουμένως ήταν μη σημαντικό, μπορεί να γίνει σημαντικό. Κατά συνέπεια, αν υπάρχουν μερικές μικρές (μη σημαντικές) t τιμές, οι μεταβλητές πρόβλεψης πρέπει να διαγράφονται μία κάθε φορά (ξεκινώντας με την μεταβλητή με τη

μικρότερη t τιμή) παρά σε ομάδες. Η διαδικασία σταματάει, όταν η παλινδρόμηση είναι σημαντική και όλες οι μεταβλητές πρόβλεψης έχουν μεγάλα (σημαντικά) στατιστικά t.

7.5.4 ΠΡΟΒΛΕΨΗ ΜΕΛΛΟΝΤΙΚΩΝ ΤΙΜΩΝ

Μία μελλοντική πρόβλεψη του $\hat{Y}'$ για νέες τιμές των X, ας πούμε

$$X_1 = X'_1,\ X_2 = X'_2, \ldots,\ X_k = X'_k,$$

δίνεται υπολογίζοντας την προσαρμοσμένη συνάρτηση παλινδρόμησης:

$$\hat{Y}' = b_0 + b_1 X'_1 + b_2 X'_2 + \ldots + b_k X'_k$$

Με διάστημα εμπιστοσύνης 1-α, ένα διάστημα πρόβλεψης του Y παίρνει τη μορφή:

$$\hat{Y}' \pm t_{\alpha/2}\, x\, (\textit{Τυπικό Σφάλμα Πρόβλεψης}),$$

με t την κατανομή Student και n-k-1 βαθμούς ελευθερίας.

Επειδή το τυπικό σφάλμα πρόβλεψης είναι μία περίπλοκη έκφραση, μπορούμε να χρησιμοποιήσουμε το τυπικό σφάλμα εκτίμησης $s_{y \cdot x's}$ σαν προσέγγιση, οπότε ένα διάστημα πρόβλεψης θα είναι:

$$(\hat{Y}' - t_{\alpha/2} s_{y \cdot x's},\ \hat{Y}' + t_{\alpha/2} s_{y \cdot x's})$$

7.6 ΥΠΟΛΟΓΙΣΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Τα υπολογιστικά αποτελέσματα του προβλήματος παρουσιάζονται στον παρακάτω Πίνακα. Η ανάλυση των αποτελεσμάτων αυτών οδηγεί στις ακόλουθες παρατηρήσεις.

1. Οι συντελεστές παλινδρόμησης (regression coefficients) είναι -8,25 για την τιμή γάλακτος και 0,585 για τα έξοδα διαφήμισης. Η προσαρμοσμένη εξίσωση παλινδρόμησης είναι

$$\hat{Y} = 16{,}4 - 8{,}25\, X_1 + 0{,}585\, X_2 .$$

2. Η εξίσωση παλινδρόμησης ερμηνεύει το 93,2% της μεταβολής στον όγκο των πωλήσεων.
3. Το τυπικό σφάλμα της εκτίμησης είναι 1,5072 γαλόνια. Αυτή η τιμή είναι ένας υπολογισμός της ποσότητας, κατά την οποία οι πραγματικές τιμές διαφέρουν από τις προσαρμοσμένες τιμές.
4. Στο Κεφάλαιο 6, ο συντελεστής κλίσης της παλινδρόμησης ελέγχθηκε για να καθοριστεί αν ήταν διαφορετικός από το μηδέν. Στην τωρινή κατάσταση, η μεγάλη τιμή της στατιστικής t (-3,76) για τη μεταβλητή της τιμής, X_1, και η μικρή p-τιμή (0,007) υποδεικνύουν, ότι ο συντελεστής της τιμής είναι σημαντικά διαφορετικός από το μηδέν (απόρριψη της H_0:β_1=0). Δεδομένης της μεταβλητής της διαφήμισης, X_2 στην συνάρτηση παλινδρόμησης, η τιμή δεν μπορεί να διαγραφεί από την συνάρτηση παλινδρόμησης. Παρόμοια, η μεγάλη τιμή της στατιστικής t (4,38) για την μεταβλητή διαφήμισης, X_2 και η μικρή της p-τιμή (0,003) υποδεικνύουν, ότι ο συντελεστής της διαφήμισης είναι σημαντικά διαφορετικός από το μηδέν (απόρριψη της H_0:β_2=0). Δεδομένης της μεταβλητής της τιμής, X_1, στη συνάρτηση παλινδρόμησης, η μεταβλητή της διαφήμισης δεν μπορεί να διαγραφεί από την συνάρτηση παλινδρόμησης. (Σαν ένα σημείο αναφοράς για το μέγεθος των t τιμών, με 7 βαθμούς ελευθερίας $t_{0,01}$=2,998). Οι συντελεστές και των δύο μεταβλητών πρόβλεψης είναι σημαντικά διαφορετικοί από το μηδέν.
5. Η p-τιμή 0,007 είναι η πιθανότητα που αντιστοιχεί στην τιμή t= -3,76 αν η υπόθεση H_0:β_1=0 είναι αληθής. Μιας και η πιθανότητα είναι εξαιρετικά μικρή, η H_0 είναι απίθανο να είναι αληθής, και

απορρίπτεται. Ο συντελεστής της τιμής είναι σημαντικά διαφορετικός από το 0. Η p-τιμή 0,003 είναι η πιθανότητα να αποκτηθεί μία t τιμή τουλάχιστον τόσο μεγάλη, όσο το 4,38 αν η H_0:β_2=0 είναι αληθής. Δεδομένου ότι η t τιμή ενός τέτοιου μεγέθους είναι εξαιρετικά απίθανη, η H_0 απορρίπτεται. Ο συντελεστής της διαφήμισης είναι σημαντικά διαφορετικός από το μηδέν.

6. Ο πίνακας συσχέτισης φάνηκε στον αντίστοιχο Πίνακα.
7. Η ανάλυση του αθροίσματος των τετραγώνων, SST=SSR+SSE, δόθηκε προηγουμένως.
8. Η υπολογισμένη F τιμή (47,92) χρησιμοποιήθηκε για να ελέγξει τη σημαντικότητα της παλινδρόμησης. Ο μεγάλος F λόγος και η μικρή του p-τιμή (0,000) δείχνουν ότι η παλινδρόμηση είναι σημαντική (απόρριψη της H_0:β_1=β_2=0). Ο F λόγος υπολογίζεται:

$$F=\frac{MSR}{MSE}=\frac{108,85}{2,27}=47,92.$$

Σαν μία αναφορά του μεγέθους του F λόγου, το άνω 1% σημείο της F κατανομής με 2 και 7 βαθμούς ελευθερίας είναι $F_{0,01}$=9,55. Η συνάρτηση παλινδρόμησης ερμηνεύει μία σημαντική ποσότητα της μεταβλητότητας στις πωλήσεις, Y.

9. Ο υπολογισμός του διορθωμένου ή προσαρμοσμένου R^2, $\overline{R^2}$, είναι:

$$\bar{R}^2=1-(1-R^2)\left(\frac{n-1}{n-k-1}\right)=0,912.$$

Correlations: Y; X1; X2

	Y	X1
X1	-0,863	
X2	0,891	-0,654

Regression Analysis: Y versus X1; X2

```
The regression equation is
Y = 16,4 - 8,25 X1 + 0,585 X2

Predictor   Coef    SE Coef    T         P
Constant   16,406    4,343    3,78    0,007
X1         -8,248    2,196   -3,76    0,007
X2          0,5851   0,1337   4,38    0,003

S = 1,50720 R-Sq = 93,2% R-Sq(adj) = 91,2%

Analysis of Variance
Source        DF    SS      MS      F      P
Regression     2  217,70  108,85  47,92  0,000
Residual Error 7   15,90    2,27
Total          9  233,60
```

7.7 ΨΕΥΔΟΜΕΤΑΒΛΗΤΕΣ

Θεωρούμε το ακόλουθο παράδειγμα:

Παράδειγμα (Τ.7-9)

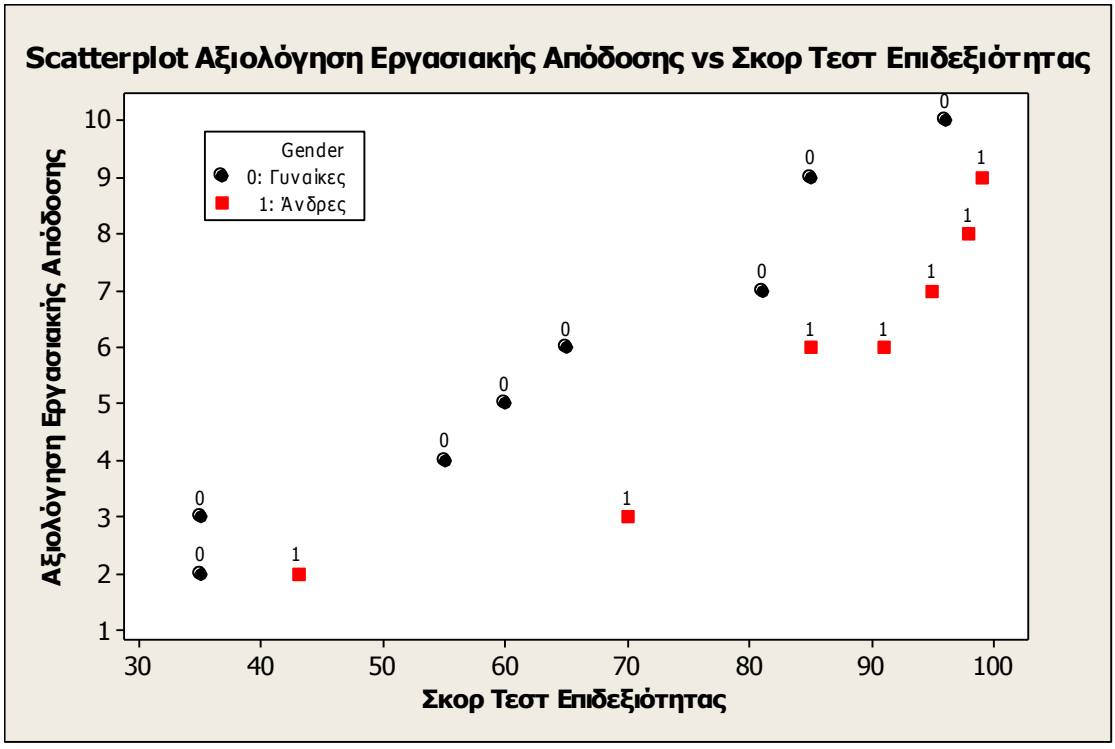
Υποθέστε έναν αναλυτή, ο οποίος θέλει να ερευνήσει, πόσο καλά ένα συγκεκριμένο τεστ ικανότητας προβλέπει την εργασιακή απόδοση. Οχτώ γυναίκες και επτά άντρες (subjects) έκαναν αυτό το τεστ και η παραγωγικότητά τους αξιολογήθηκε από έναν δείκτη, ο οποίος παίρνει τιμές από το 0 μέχρι το 10 (το 0 σημαίνει μη παραγωγικός).

Μερικές φορές είναι απαραίτητο να καθοριστεί, με ποιον τρόπο μία εξαρτημένη μεταβλητή σχετίζεται με μία ανεξάρτητη μεταβλητή, όταν ένας ποιοτικός παράγοντας εισέρχεται στο σύστημα και επηρεάζει την κατάσταση. Αυτή η σχέση ολοκληρώνεται δημιουργώντας μία ψευδομεταβλητή. Υπάρχουν πολλοί τρόποι, για να αναγνωρίσει κάποιος ποσοτικά τις τάξεις μία ποιοτικής μεταβλητής. Οι τιμές 0 και 1 χρησιμοποιούνται σε αυτό το παράδειγμα.

Ψευδομεταβλητές (dummy or indicator variables) χρησιμοποιούνται για να καθορίσουν τις σχέσεις μεταξύ ποιοτικών ανεξάρτητων μεταβλητών και μίας εξαρτημένης μεταβλητής.

Ένα διάγραμμα διασποράς παρουσιάζεται στο Γράφημα. Κάθε γυναίκα εργαζόμενη αντιπροσωπεύεται με 0 και κάθε άντρας με 1.

Είναι αμέσως φανερό παρατηρώντας το Γράφημα, ότι η σχέση του τεστ ικανότητας με την εργασιακή απόδοση ακολουθεί δύο ξεχωριστά μοτίβα, ένα που σχετίζεται με τις γυναίκες και ένα που σχετίζεται με τους άντρες.



Η τεχνική ψευδομεταβλητής παρουσιάζεται στο παραπάνω Γράφημα. Τα σημεία δεδομένων για τις γυναίκες φαίνονται σαν 0, ενώ με 1 αντιπροσωπεύουν τους άντρες. Επίσης κατασκευάζονται δύο παράλληλες γραμμές για το διάγραμμα διασποράς. Η πάνω γραμμή προσαρμόζεται στα δεδομένα για τις γυναίκες, και η κάτω στα δεδομένα για τους άντρες.

Κάθε μία από αυτές τις γραμμές αποκτήθηκε από μία προσαρμοσμένη συνάρτηση παλινδρόμησης της μορφής:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

όπου:

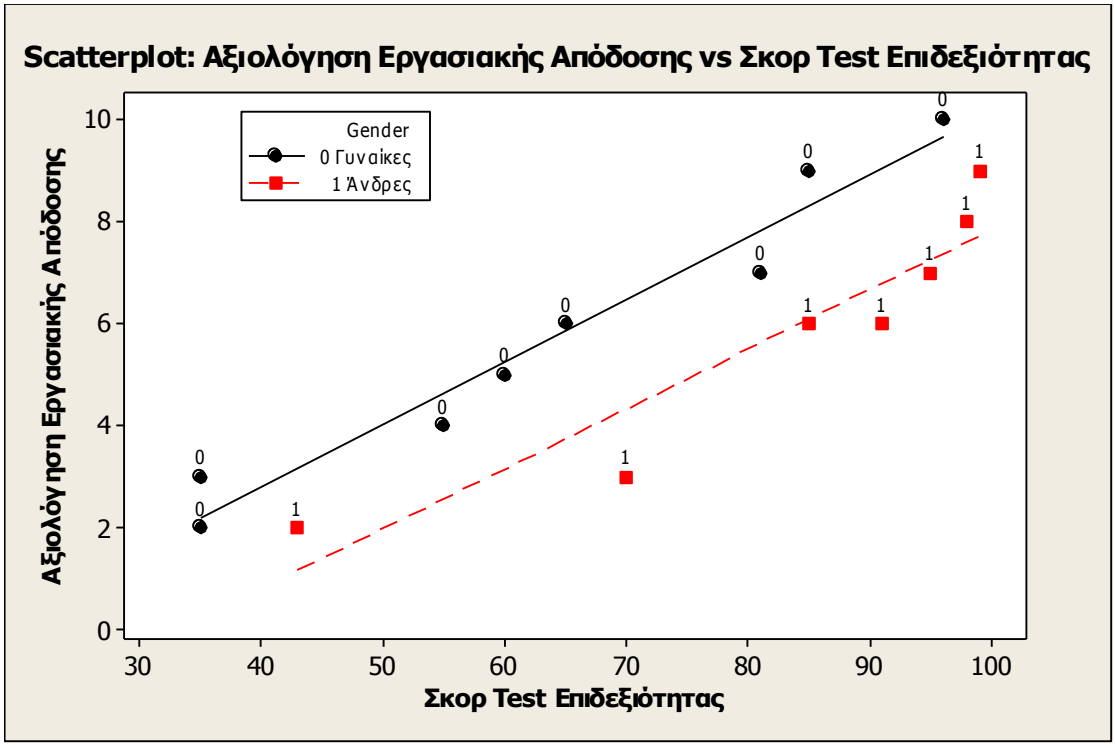
- X_1 = το σκορ στο τεστ,
- X_2 = 0 για τις γυναίκες,
- X_2 = 1 για τους άνδρες.

Η εξίσωση είναι ισοδύναμη με τις δύο ακόλουθες εξισώσεις:

$$\hat{Y} = b_0 + b_1 X_1 \quad \text{για τις γυναίκες και}$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 = (b_0 + b_2) + b_1 X_1 \quad \text{για τους άνδρες}$$

Να σημειωθεί ότι το b_2 αντιπροσωπεύει την επίδραση ενός άντρα στην εργασιακή απόδοση και το b_1 αντιπροσωπεύει την επίδραση των διαφορών στα σκορ του τεστ ικανότητας (η τιμή b_1 θεωρείται ότι είναι ίδια και για τους άντρες και για τις γυναίκες).



Το σημαντικό σημείο είναι, ότι μία εξίσωση πολυμεταβλητής παλινδρόμησης θα δημιουργήσει δύο εκτιμώμενες γραμμές, όπως φαίνεται στο Γράφημα. Η πάνω γραμμή είναι η εκτιμώμενη σχέση για τις γυναίκες, και η κάτω γραμμή είναι η εκτιμώμενη σχέση για τους άντρες. Κάποιος μπορεί να θεωρήσει το X_2 σαν μία μεταβλητή, που εναλλάσσεται και η οποία είναι «On», όταν μία παρατήρηση γίνεται σε άντρα και είναι «Off», όταν μία παρατήρηση γίνεται σε γυναίκα.

Παράδειγμα

Η εκτιμώμενη εξίσωση συνάρτησης παλινδρόμησης για τα δεδομένα του Παραδείγματός μας φαίνεται στα υπολογιστικά αποτελέσματα του Minitab στον ακόλουθο Πίνακα. Έχουμε:

$$\hat{Y} = -1{,}96 + 0{,}12\,X_1 - 2{,}18\,X_2$$

Correlations: Rating; Test; Gender

	Rating	Test
Test	0,876	
Gender	0,021	0,428

Regression Analysis: Rating versus Test; Gender

The regression equation is

Rating = - 1,96 + 0,120 Test - 2,18 Gender

Predictor	Coef	SE Coef	T	P
Constant	-1,9565	0,7068	-2,77	0,017
Test	0,12041	0,01015	11,86	0,000
Gender	-2,1807	0,4503	-4,84	0,000

S = 0,786263 R-Sq = 92,1% R-Sq(adj) = 90,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	86,981	43,491	70,35	0,000
Residual Error	12	7,419	0,618		
Total	14	94,400			

$\hat{Y} = -1{,}96 + 0{,}12\,X_1$ για τις γυναίκες και

$$\hat{Y} = -4{,}14 + 0{,}12\, X_1 \quad \text{για τους άνδρες.}$$

Αυτές οι εξισώσεις μπορούν να ερμηνευθούν με τον ακόλουθο τρόπο: Η τιμή του συντελεστή παλινδρόμησης b_1=0,12, ο οποίος είναι η κλίση για κάθε μία από τις γραμμές, είναι η εκτιμώμενη μέση αύξηση στην αξιολόγηση απόδοσης για κάθε μία μονάδα αύξησης στο σκορ του τεστ ικανότητας. Αυτός ο συντελεστής ισχύει και για τους άντρες και για τις γυναίκες.

Ο άλλος συντελεστής παλινδρόμησης, b_2=-2,18, ισχύει μόνο στους άντρες. Για έναν άντρα, που κάνει το τεστ, η εκτιμώμενη αξιολόγηση της εργασιακής απόδοσης μειώνεται, σε σχέση με τις γυναίκες, που κάνουν το τεστ, κατά 2,18 μονάδες, όταν το σκορ ικανότητας είναι σταθερό.

Μία εξέταση των μέσων των Y και X_1 μεταβλητών, κατηγοριοποιώντας τους με βάση το φύλο, βοηθάει κάποιον να κατανοήσει το αποτέλεσμα. Οι μέσες αξιολογήσεις της εργασιακής απόδοσης είναι περίπου ίσες για τους άντρες (5,86) και για τις γυναίκες (5,75). Ωστόσο, οι άντρες έκαναν κατά μέσον όρο καλύτερο σκορ (83) στο τεστ ικανότητας από ότι οι γυναίκες (64). Ως εκ τούτου, αν δύο υποψήφιοι, ένας άντρας και μία γυναίκα, έκαναν το τεστ ικανότητας και οι δύο συγκέντρωσαν σκορ 70, η γυναικεία εκτιμώμενη εργασιακή απόδοση θα ήταν 2,18 μονάδες υψηλότερη από του άντρα, δεδομένου ότι:

$$\hat{Y} = -1{,}96 + 0{,}12(70) = 6{,}44 \quad \text{για τις γυναίκες και}$$

$$\hat{Y} = -4{,}14 + 0{,}12(70) = 4{,}26 \quad \text{για τους άνδρες}$$

Μία ματιά στον πίνακα συσχέτισης στον Πίνακα Ανάλυσης παρέχει μερικές ενδιαφέρουσες ιδέες. Μία ισχυρή γραμμική σχέση υπάρχει μεταξύ της εργασιακής απόδοσης και του τεστ ικανότητας επειδή r_{12}=0,876. Αν το σκορ στο τεστ ικανότητας χρησιμοποιούνταν

μόνο του για να προβλέψει την απόδοση, θα ερμήνευε περίπου το 77% ($0,876^2$=0,767) της μεταβολής στα σκορ της εργασιακής απόδοσης.

Ο συντελεστής απόδοσης r_{13}=0,02 υποδεικνύει, ότι δεν υπάρχει σχέση μεταξύ του φύλου και της εργασιακής απόδοσης. Αυτό το συμπέρασμα είναι επίσης προφανές από το γεγονός, ότι η μέση εργασιακή απόδοση για τις γυναίκες και τους άντρες είναι σχεδόν ίδια (5,86 και 5,75). Με μία ματιά, κάποιος μπορεί να συμπεράνει, ότι η γνώση του αν κάποιος υποψήφιος είναι άντρας ή γυναίκα, δεν είναι χρήσιμη πληροφορία. Ωστόσο, η μέτρια σχέση, r_{23}=0,428, μεταξύ του φύλου και του σκορ στο τεστ επιδεξιότητας υποδεικνύει, ότι το τεστ μπορεί να κάνει διακρίσεις μεταξύ των φύλων. Οι άντρες φαίνεται να τα καταφέρνουν καλύτερα στο τεστ από ότι οι γυναίκες (83 και 64). Ίσως κάποιος άλλος παράγοντας να απαιτείται στο τεστ, που δεν απαιτείται στη δουλειά.

Όταν χρησιμοποιούνται για την πρόβλεψη και τα αποτελέσματα του τεστ και το φύλο, ερμηνεύεται το 92% της μεταβολής. Αυτό το αποτέλεσμα προτείνει, ότι και οι δύο μεταβλητές έχουν σημαντική συνεισφορά στην απόδοση που προβλέπεται. Τα σκορ στο τεστ ικανότητας ερμηνεύουν το 77% της διασποράς, και το φύλο όταν χρησιμοποιείται σε συνδυασμό με τα σκορ στο τεστ ικανότητας ερμηνεύει ένα ακόμα 15%. Το υπολογισμένο στατιστικό t, 11,86 (p-τιμή = 0,000) και -4,84 (p-τιμή=0,000), για το σκορ στο τεστ ικανότητας και το φύλο, αντίστοιχα, υποδεικνύει, ότι και οι δύο μεταβλητές πρόβλεψης πρέπει να συμπεριλαμβάνονται στην τελική συνάρτηση παλινδρόμησης.

7.8 ΠΟΛΥΣΥΓΓΡΑΜΙΚΟΤΗΤΑ

Σε πολλά προβλήματα παλινδρόμησης, τα δεδομένα συνήθως

καταγράφονται παρά παράγονται από τις προεπιλεγμένες ρυθμίσεις των ανεξάρτητων μεταβλητών. Σε αυτές τις περιπτώσεις, οι ανεξάρτητες μεταβλητές είναι συχνά γραμμικά εξαρτημένες, ή πολυσυγγραμικές (multicollinear). Για παράδειγμα, η τιμή πώλησης ενός σπιτιού μπορεί να σχετίζεται με μεταβλητές, όπως η ηλικία, τα τετραγωνικά, ο αριθμός των λουτρών, ο αριθμός των υπόλοιπων δωματίων και ένας δείκτης της ποιότητας κατασκευής. Ο αριθμός των λουτρών και ο αριθμός των υπόλοιπων δωματίων κινούνται μαζί. Αν μία από αυτές τις μεταβλητές αυξηθεί, θα αυξηθεί και η άλλη.

Αν αυτή η γραμμική εξάρτηση δεν είναι τέλεια, μπορούν να βρεθούν ακόμη οι εκτιμήσεις των ελαχίστων τετραγώνων των συντελεστών του μοντέλου παλινδρόμησης. Ωστόσο, αυτές οι εκτιμήσεις τείνουν να μην είναι σταθερές και οι τιμές τους να είναι μεγαλύτερες από το αναμενόμενο. Συγκεκριμένα, μεμονωμένοι συντελεστές μπορεί να έχουν το λάθος πρόσημο και τα στατιστικά t, που κρίνουν τη σημαντικότητα των μεμονωμένων όρων, μπορεί όλα να είναι μη σημαντικά, αλλά το F τεστ μπορεί να υποδείξει, ότι η παλινδρόμηση είναι σημαντική.

Η πολυσυγγραμικότητα είναι η κατάσταση, κατά την οποία οι ανεξάρτητες μεταβλητές σε μία εξίσωση πολυμεταβλητής παλινδρόμησης είναι υψηλά συσχετισμένες. Αυτό γίνεται, όταν υπάρχει μία γραμμική σχέση μεταξύ δύο ή περισσότερων ανεξάρτητων μεταβλητών.

Ο βαθμός της πολυσυγγραμμικότητας υπολογίζεται από τον συντελεστή διόγκωσης της διακύμανσης (Variance Inflation Factor), VIF:

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

Εδώ, το R_j^2, είναι ο συντελεστής προσδιορισμού, που προκύπτει από την παλινδρόμηση της j ανεξάρτητης μεταβλητής σε σχέση με τις υπόλοιπες k-1 ανεξάρτητες μεταβλητές. Για k=2 ανεξάρτητες μεταβλητές, το R_j^2 είναι το τετράγωνο του συντελεστή συσχέτισης των δύο ανεξάρτητων μεταβλητών r από το δείγμα τους.

Αν η j μεταβλητή πρόβλεψης, X_j, δεν σχετίζεται με τα εναπομένοντα X, το $R_j^2=0$ και το $VIF_j=1$. Αν υπάρχει σχέση, τότε $VIF_j>1$. Για παράδειγμα, όταν το $R_j^2=0,90 \quad VIF_j=1/(1-0,9)=10.$

Ένα VIF κοντά στο 1 δηλώνει, ότι η πολυσυγγραμικότητα δεν είναι πρόβλημα για εκείνη την ανεξάρτητη μεταβλητή. Ο εκτιμώμενος συντελεστής και η σχετική t τιμή δεν θα αλλάξουν πολύ, καθώς οι άλλες ανεξάρτητες μεταβλητές προστίθενται, ή διαγράφονται από την εξίσωση παλινδρόμησης. Ένα VIF πολύ μεγαλύτερο από το 1, δηλώνει ότι ο εκτιμώμενος συντελεστής της ανεξάρτητης μεταβλητής δεν είναι σταθερός. Η τιμή του και το αντίστοιχο στατιστικό t μπορεί να αλλάξουν σημαντικά καθώς άλλες ανεξάρτητες μεταβλητές προστίθενται, ή διαγράφονται από την εξίσωση παλινδρόμησης. Ένα μεγάλο VIF σημαίνει ουσιαστικά, ότι υπάρχει περιττή πληροφορία μεταξύ των μεταβλητών πρόβλεψης. Η πληροφόρηση, που μεταβιβάζεται από μία μεταβλητή με μεγάλο VIF, ερμηνεύεται ήδη από τις εναπομένουσες μεταβλητές πρόβλεψης. Έτσι, η πολυσυγγραμικότητα κάνει δύσκολα κατανοητή την επίδραση μίας μεμονωμένης μεταβλητής πρόβλεψης στην απόκριση (εξαρτημένη μεταβλητή).

Υπάρχουν αρκετοί τρόποι να ελαχιστοποιήσουμε την πολυσυγραμμικότητα. Μερικοί από αυτούς περιγράφονται παρακάτω:

- Δημιουργία νέων μεταβλητών (εξαρτημένης και ανεξάρτητης)

σύμφωνα με τη σχέση $\tilde{X}_{ij}=\frac{X_{ij}-\bar{X}_j}{\sqrt{\sum_i (X_{ij}-\bar{X}_j)^2}}$, όπου το j αναφέρεται στη μεταβλητή και το i στην παρατήρηση. Αυτές οι νέες μεταβλητές έχουν μέση τιμή 0 και την ίδια διασπορά 1. Από μόνος του όμως αυτός ο μετασχηματισμός δεν αρκεί για την άρση της πολυσυγγραμικότητας.

- Αφαίρεση από το μοντέλο της παλινδρόμησης των εξαρτώμενων μεταβλητών (βλ. επόμενο παράδειγμα).
- Καλό είναι από την αρχή της κατασκευής του μοντέλου να δοθεί ιδιαίτερη μνεία, στο να βρεθούν ανεξάρτητες μεταβλητές, που να μην είναι συγγραμμικές.
- Εφαρμογή της μεθόδου "Principal Components" για την εύρεση νέων ανεξάρτητων και ορθοκανονικών μεταβλητών ως γραμμικός συνδυασμός των πρώτων.

Παράδειγμα (ΕΧ.7-7)

Οι εκδότες εφημερίδων ενδιαφέρονται για τους παράγοντες που καθορίζουν την ετήσια κατανάλωση εντύπων. Σε μία μελέτη που έγινε, συλλέχθηκαν τα δεδομένα για την κατανάλωση εντύπων (Y), τον αριθμό των εφημερίδων στην πόλη (X_1), τον λογάριθμο του αριθμού των οικογενειών στην πόλη (X_2) και το λογάριθμο του συνολικού αριθμού λιανικών πωλήσεων στην πόλη (X_3) για n=15 πόλεις. Η σειρά συσχέτισης για τις μεταβλητές πρόβλεψης και τα αποτελέσματα του Minitab από την ανάλυση παλινδρόμησης, που συσχετίζει την κατανάλωση εφημερίδων με τις μεταβλητές πρόβλεψης παρουσιάζονται στον Πίνακα Ανάλυσης.

Correlations: Papers; LnFamily; LnRetSales

	Papers	LnFamily
LnFamily	0,600	

```
LnRetSales 0,643   0,930
```

Regression Analysis: Newsprint versus Papers; LnFamily; LnRetSales

```
The regression equation is
Newsprint = - 56388 + 2385 Papers + 1859 LnFamily
+ 3455 LnRetSales
Predictor   Coef SE Coef   T      P     VIF
Constant  -56388  13206  -4,27  0,001
Papers      2385   1410   1,69  0,119  1,704
LnFamily    1859   2346   0,79  0,445  7,410
LnRetSales  3455   2590   1,33  0,209  8,082

S = 1849,36 R-Sq = 83,5% R-Sq(adj) = 79,0%

Analysis of Variance
Source          DF     SS       MS      F     P
Regression       3 190239371 63413124 18,54 0,000
Residual Error 11  37621478 3420134
Total          14 227860849
```

Το στατιστικό F (18,54) και η p-τιμή του (0,000) υποδεικνύουν ξεκάθαρα, ότι η παλινδρόμηση είναι σημαντική. Το στατιστικό t για κάθε μία από τις ανεξάρτητες μεταβλητές είναι μικρό με μία σχετικά μεγάλη p-τιμή. Πρέπει να συμπεράνουμε, για παράδειγμα, ότι η μεταβλητή LnFamily δεν είναι σημαντική, δεδομένου ότι οι υπόλοιπες μεταβλητές πρόβλεψης παραμένουν στη συνάρτηση παλινδρόμησης. Αυτό υποδεικνύει, ότι ο όρος β_2X_2 μπορεί να απαλειφθεί από τη συνάρτηση παλινδρόμησης αν οι εναπομείναντες όροι, β_1X_1 και β_3X_3 διατηρηθούν. Παρομοίως, ο β_3X_3 μπορεί να απαλειφθεί αν ο β_1X_1 και ο β_2X_2 παραμείνουν στη συνάρτηση παλινδρόμησης. Η t τιμή (1,69) που σχετίζεται με την Papers είναι οριακά σημαντική, αλλά ο όρος β_1X_1 μπορεί επίσης να απαλειφθεί, αν οι άλλες μεταβλητές πρόβλεψης παραμείνουν στην εξίσωση. Εδώ, η παλινδρόμηση είναι σημαντική, αλλά κάθε μία από τις μεταβλητές πρόβλεψης δεν είναι σημαντικές.

Την απάντηση στο γιατί τη δίνει η VIF στήλη του Πίνακα. Δεδομένου ότι το VIF=1,7 για την Papers, η μεταβλητή πρόβλεψης είναι πολύ λίγο συσχετισμένη (VIF κοντά στο 1) με τις υπόλοιπες μεταβλητές πρόβλεψης, LnFamily και LnRetSales. Ο VIF=7,4 για το LnFamily είναι σχετικά μεγάλος και υποδεικνύει, ότι η μεταβλητή αυτή σχετίζεται γραμμικά με τις εναπομένουσες μεταβλητές πρόβλεψης. Επίσης, ο VIF=8,1 για την LnRetSales υποδεικνύει, ότι η LnRetSales σχετίζεται με τις εναπομένουσες μεταβλητές πρόβλεψης. Δεδομένου ότι η Papers είναι λίγο συσχετισμένη με τις LnFamily και LnRetSales, η σχέση μεταξύ των μεταβλητών πρόβλεψης είναι ουσιαστικά η σχέση μεταξύ LnFamily και LnRetSales. Στην πραγματικότητα, ο συντελεστής συσχέτισης δείγματος μεταξύ της LnFamily και της LnRetSales είναι r=0,93, και δείχνει ισχυρή γραμμική συσχέτιση.

Οι μεταβλητές LnFamily και LnRetSales είναι παρόμοιες στην ικανότητά τους να ερμηνεύουν την κατανάλωση εφημερίδων. Χρειαζόμαστε μόνο μία, και όχι και τις δύο, στη συνάρτηση παλινδρόμησης. Τα αποτελέσματα του Minitab από την ανάλυση παλινδρόμησης με την LnFamily (μικρότερο στατιστικό t) διαγραμμένη από τη συνάρτηση παλινδρόμησης φαίνονται στον παρακάτω Πίνακα.

Regression Analysis: Newsprint versus Papers; LnRetSales

```
The regression equation is
Newsprint = - 59766 + 2393 Papers + 5279
LnRetSales

Predictor    Coef SE Coef    T      P     VIF
Constant  -59766   12304  -4,86  0,000
Papers      2393    1388   1,72  0,110  1,704
LnRetSales  5279    1171   4,51  0,001  1,704

S = 1820,49 R-Sq = 82,5% R-Sq(adj) = 79,6%
```

```
Analysis of Variance
Source          DF    SS         MS       F      P
Regression       2 188090489 94045244 28,38 0,000
Residual Error 12  39770360  3314197
Total          14 227860849
```

Ο συντελεστής της Papers είναι περίπου ο ίδιος και για τις δύο παλινδρομήσεις. Οι συντελεστές της LnRetSales, ωστόσο, είναι σημαντικά διαφορετικοί (3,455 για k=3 μεταβλητές πρόβλεψης και 5,279 για k=2 μεταβλητές πρόβλεψης). Επίσης, για τη δεύτερη παλινδρόμηση, η μεταβλητή LnRetSales είναι ξεκάθαρα σημαντική (t=4,51 και p-τιμή=0,001). Με την Papers στο μοντέλο, η LnRetSales είναι μία επιπρόσθετη σημαντική μεταβλητή πρόβλεψης της κατανάλωσης εφημερίδων. Το R^2 για τις δύο παλινδρομήσεις είναι περίπου το ίδιο, περίπου 0,83 και το ίδιο ισχύει για τα τυπικά σφάλματα εκτίμησης $s_{y \cdot x's} = 1.849$ και $s_{y \cdot x's} = 1.820$, αντίστοιχα. Τέλος, το κοινό VIF=1,7 για τις δύο μεταβλητές πρόβλεψης στο δεύτερο μοντέλο υποδεικνύει, ότι η πολυσυγγραμικότητα δεν είναι πλέον πρόβλημα. Όπως μία ανάλυση καταλοίπων επιβεβαιώνει, για τις μεταβλητές που θεωρούνται, η παλινδρόμηση της Newsprint, στις Papers και LnRetSales είναι ολοκληρωτικά επαρκής.

7.9 ΕΠΙΛΕΓΟΝΤΑΣ ΤΗΝ ΚΑΛΥΤΕΡΗ ΕΞΙΣΩΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Πώς μπορεί κάποιος να αναπτύξει την καλύτερη εξίσωση πολυμεταβλητής παλινδρόμησης, για να προβλέψει μία μεταβλητή του ενδιαφέροντός του; Για τον σκοπό αυτό ακολουθούμε τρία βήματα:

Το **πρώτο βήμα** περιλαμβάνει την επιλογή ενός ολοκληρωμένου συνόλου πιθανών μεταβλητών πρόβλεψης. Κάθε μεταβλητή, που μπορεί

να προσθέσει ακρίβεια στην πρόβλεψη, πρέπει να συμπεριληφθεί. Στην επιλογή της τελικής εξίσωσης, κάποιος συνήθως αντιμετωπίζει το δίλημμα της παροχής της πιο ακριβούς πρόβλεψης με το χαμηλότερο κόστος. Με άλλα λόγια, όταν επιλέγει μεταβλητές πρόβλεψης για να συμπεριλάβει στην τελική εξίσωση, ο αναλυτής πρέπει να τις αξιολογήσει χρησιμοποιώντας τα ακόλουθα δύο αντίθετα κριτήρια:

1. Ο αναλυτής θέλει η εξίσωση να περιέχει, όσο το δυνατόν περισσότερες χρήσιμες μεταβλητές πρόβλεψης, ώστε να υπάρχει μεγαλύτερη επεξηγησιμότητα και μεγαλύτερη διασπορά κινδύνου σε περίπτωση αστοχίας στην πρόβλεψη μιας μεταβλητής πρόβλεψης.
2. Δεδομένου ότι το να αποκτηθούν και να καταγραφούν πληροφορίες για μεγάλο αριθμό Χ κοστίζει χρήματα, η εξίσωση πρέπει να περιέχει, όσο το δυνατόν λιγότερες μεταβλητές. Η απλούστερη εξίσωση είναι συνήθως και η καλύτερη εξίσωση.

Η επιλογή της καλύτερης εξίσωσης παλινδρόμησης συνήθως περιλαμβάνει ένα συμβιβασμό μεταξύ των δύο αυτών ακραίων προτάσεων και η σωστή κρίση είναι απαραίτητο μέρος της λύσης.

Μετά από τη σύνθεση μίας φαινομενικά ολοκληρωμένης λίστας των πιθανών μεταβλητών πρόβλεψης, το **δεύτερο βήμα** είναι να βγάλουμε έξω από την παλινδρόμηση τις ανεξάρτητες μεταβλητές, που δεν φαίνονται κατάλληλες για το πρόβλημα, δεν είναι βασικές. Μία ανεξάρτητη μεταβλητή:

1. μπορεί να μην είναι απαραίτητη στο πρόβλημα (πρέπει να υπάρχει εύλογη σχέση μεταξύ της εξαρτημένης μεταβλητής και μίας ανεξάρτητης μεταβλητής),
2. μπορεί να ευθύνεται για μεγάλα υπολογιστικά σφάλματα,

3. μπορεί να είναι ίδια με άλλες ανεξάρτητες μεταβλητές (πολυσυγγραμικότητα), ή
4. μπορεί να είναι δύσκολο να υπολογιστεί με ακρίβεια (τα ακριβή δεδομένα είναι μη διαθέσιμα, ή ακριβά να αποκτηθούν).

Το **τρίτο βήμα** είναι να μειωθεί η λίστα των μεταβλητών πρόβλεψης σύμφωνα με τη στατιστική σημαντικότητά τους, έτσι ώστε να αποκτηθεί η καλύτερη επιλογή των ανεξάρτητων μεταβλητών. Οι τεχνικές, που συνήθως χρησιμοποιούνται, συζητούνται παρακάτω. Καμία από τις διαδικασίες έρευνας, δεν μπορεί να ειπωθεί, ότι οδηγεί στο καλύτερο σύνολο ανεξάρτητων μεταβλητών. Πράγματι, συνήθως δεν υπάρχει ένα μοναδικό «καλύτερο» σύνολο. Μπορεί να υπάρχουν περισσότερα του ενός και η επιλογή του εξαρτάται από τη φύση του προβλήματος. Οι διάφορες τεχνικές συνήθως δεν οδηγούν στην ίδια τελική εξίσωση πρόβλεψης. Η όλη διαδικασία επιλογής μεταβλητών είναι πολύ υποκειμενική. Το κύριο πλεονέκτημα των διαδικασιών αυτόματης εύρεσης, είναι ότι οι αναλυτές μπορούν, τότε να εστιάσουν στην κρίση των κεντρικών περιοχών του προβλήματος.

Για να παρουσιάσουμε μερικές διαδικασίες έρευνας, παρουσιάζεται ένα απλό παράδειγμα, που έχει πέντε πιθανές ανεξάρτητες μεταβλητές.

Παράδειγμα (Τ.7-13)

Ο manager μιας εταιρείας, ενδιαφέρεται να προβλέψει, αν ένας συγκεκριμένος υπάλληλος θα γίνει καλός πωλητής. Αποφασίζει να χρησιμοποιήσει τις πωλήσεις του πρώτου μήνα σαν την εξαρτημένη μεταβλητή (Y) (Sales) και επιλέγει να αναλύσει τις ακόλουθες ανεξάρτητες μεταβλητές:

- X_1: το σκορ στο τεστ ικανότητας στις πωλήσεις (Aptitude).

- X_2: την ηλικία, σε χρόνια (Age).
- X_3: το σκορ στο τεστ άγχους (Anxiety).
- X_4: την εμπειρία, σε χρόνια (Exp.).
- X_5: το βαθμό απολυτηρίου στο λύκειο (GPA).

Συγκεντρώνει τα δεδομένα και αναθέτει στον αναλυτή του την εργασία του, να καταλήξει στο «καλύτερο» σύνολο ανεξάρτητων μεταβλητών για την πρόβλεψη της ικανότητας στις πωλήσεις.

Το πρώτο πράγμα είναι να δημιουργήσει ένα πίνακα συσχέτισης για όλες τις μεταβλητές από ένα πρόγραμμα στον υπολογιστή. Ο πίνακας θα παρέχει την απαραίτητη γνώση για τις βασικές σχέσεις μεταξύ των μεταβλητών.

Η εξέταση του πίνακα συσχέτισης αποκαλύπτει, ότι το σκορ στο τεστ επιδεξιότητας στις πωλήσεις, η ηλικία, η εμπειρία και ο βαθμός είναι θετικά συσχετισμένοι με την ικανότητα στις πωλήσεις και έχουν πιθανότητες σαν καλές μεταβλητές πρόβλεψης. Το σκορ στο τεστ άγχους δείχνει μία χαμηλή αρνητική συσχέτιση με τις πωλήσεις, και πιθανότατα δεν είναι σημαντική μεταβλητή πρόβλεψης. Περαιτέρω ανάλυση υποδεικνύει ότι η ηλικία σχετίζεται μέτρια με το βαθμό και την εμπειρία.

Correlations: Sales; Aptitude; Age; Anxiety; Exp.; GPA

	Sales	Aptitude	Age	Anxiety	Exp.
Aptitude	0,676				
Age	0,798	0,228			
Anxiety	-0,296	-0,222	-0,287		
Exp.	0,550	0,350	0,540	-0,279	
GPA	0,622	0,318	0,695	-0,244	0,312

Παρακάτω παρουσιάζονται τέσσερις διαδικασίες: Όλες οι πιθανές παλινδρομήσεις (all possible regressions, ή best subset), η βήμα-προς-βήμα παλινδρόμηση (stepwise regression), η ανάστροφη παλινδρόμηση (backward regression) και η εμπρόσθια παλινδρόμηση (foreward

regression).

7.9.1 ΟΛΕΣ ΟΙ ΠΙΘΑΝΕΣ ΠΑΛΙΝΔΡΟΜΗΣΕΙΣ

Η διαδικασία αυτή ερευνά όλες τις πιθανές εξισώσεις παλινδρόμησης, που περιλαμβάνουν τις πιθανές ανεξάρτητες μεταβλητές. Ο αναλυτής ξεκινά με μία εξίσωση, που δεν περιέχει καμία ανεξάρτητη μεταβλητή και μετά συνεχίζει στην ανάλυση κάθε πιθανού συνδυασμού, έτσι ώστε να επιλέξει το καλύτερο σύνολο μεταβλητών.

Διαφορετικά κριτήρια για σύγκριση των διαφόρων εξισώσεων παλινδρόμησης μπορεί να χρησιμοποιηθούν με την προσέγγιση όλων των πιθανών παλινδρομήσεων. Μόνο η R^2 τεχνική, η οποία περιλαμβάνει τέσσερα βήματα, θα συζητηθεί εδώ.

Αυτή η διαδικασία στο **πρώτο βήμα** απαιτεί την προσαρμογή κάθε δυνατού μοντέλου παλινδρόμησης, που περιλαμβάνει την εξαρτημένη μεταβλητή και κάθε αριθμό ανεξάρτητων μεταβλητών. Κάθε ανεξάρτητη μεταβλητή μπορεί να είναι, ή να μην είναι στην εξίσωση (δύο δηλαδή πιθανά αποτελέσματα) και αυτό ισχύει για κάθε ανεξάρτητη μεταβλητή. Κατά συνέπεια, συνολικά έχουμε 2^k εξισώσεις (όπου το k ισούται με τον αριθμό των ανεξάρτητων μεταβλητών). Έτσι, αν υπάρχουν οκτώ ανεξάρτητες μεταβλητές (k=8) πρέπει να θεωρηθούν 2^8=256 εξισώσεις, που πρέπει να εξεταστούν.

Το **δεύτερο βήμα** στη διαδικασία είναι να διαχωριστούν οι εξισώσεις σε σύνολα (ομάδες) σύμφωνα με τον αριθμό των παραμέτρων, που πρέπει να εκτιμηθούν.

Το **τρίτο βήμα** περιλαμβάνει την επιλογή της καλύτερης ανεξάρτητης μεταβλητής (ή μεταβλητών) για κάθε γκρουπ παραμέτρων, τον καλύτερο συνδυασμό. Η εξίσωση με το υψηλότερο R^2 θεωρείται η καλύτερη. Χρησιμοποιώντας τα αποτελέσματα από το προηγούμενο

Παράδειγμα, η καλύτερη εξίσωση από κάθε σύνολο φαίνεται στον ακόλουθο Πίνακα.

Το **τέταρτο βήμα** περιλαμβάνει την υποκειμενική απόφαση: «Ποια εξίσωση είναι η καλύτερη;» Από τη μία, ο αναλυτής επιθυμεί το υψηλότερο δυνατό R^2. Από την άλλη πλευρά θέλει την απλούστερη δυνατή εξίσωση. Η προσέγγιση των όλων πιθανών παλινδρομήσεων υποθέτει, ότι ο αριθμός των σημείων δεδομένων n, υπερβαίνει τον αριθμό των παραμέτρων k+1.

Παράδειγμα (Τ.7-13)

Τα αποτελέσματα από όλες τις πιθανές παλινδρομήσεις για την εταιρία παρουσιάζονται στον ακόλουθο Πίνακα. Παρατηρήστε ότι ο Πίνακας χωρίζεται σε έξι σύνολα με αποτελέσματα εξισώσεων παλινδρόμησης. Αυτός ο διαχωρισμός συμπίπτει με τον αριθμό των παραμέτρων, που περιέχονται σε κάθε εξίσωση.

Ανεξάρτητες Μεταβλητές	**Αριθμός Παραμέτρων**	**Βαθμοί Ελευθερίας Σφάλματος**	**R^2**
Καμία	1	29	0,0000
X_1	2	28	0,4570
X_2	2	28	0,6370
X_3	2	28	0,0880
X_4	2	28	0,3020
X_5	2	28	0,3870
X_1,X_2	3	27	0,8948
X_1,X_3	3	27	0,4790
X_1,X_4	3	27	0,5690
X_1,X_5	3	27	0,6410
X_2,X_3	3	27	0,6420

X_2,X_4	3	27	0,6570
X_2,X_5	3	27	0,6460
X_3,X_4	3	27	0,3240
X_3,X_5	3	27	0,4090
X_4,X_5	3	27	0,5270
X_1,X_2,X_3	4	26	0,8951
X_1,X_2,X_4	4	26	0,8948
X_1,X_2,X_5	4	26	0,8953
X_1,X_3,X_4	4	26	0,5750
X_1,X_3,X_5	4	26	0,6460
X_1,X_4,X_5	4	26	0,7010
X_2,X_3,X_4	4	26	0,6590
X_2,X_3,X_5	4	26	0,6500
X_2,X_4,X_5	4	26	0,6690
X_3,X_4,X_5	4	26	0,5310
X_1,X_2,X_3,X_4	5	25	0,8951
X_1,X_2,X_3,X_5	5	25	0,8955
X_1,X_2,X_4,X_5	5	25	0,8953
X_1,X_3,X_4,X_5	5	25	0,7010
X_2,X_3,X_4,X_5	5	25	0,6710
X_1,X_2,X_3,X_4,X_5	6	24	0,8955

Όλες οι δυνατές παλινδρομήσεις

Αριθμός Παραμέτρων	**Ανεξάρτητες Μεταβλητές**	**Βαθμοί Ελευθερίας Σφάλματος**	**R^2**
1	Καμία	29	0,0000
2	X_2	28	0,6370
3	X_1,X_2	27	0,8948

4	X_1,X_2,X_5	26	0,8953
5	X_1,X_2,X_3,X_5	25	0,8955
6	X_1,X_2,X_3,X_4,X_5	24	0,8955

Καλύτερες εξισώσεις παλινδρόμησης

Τα αποτελέσματα του προβλήματος υποδεικνύουν, ότι το να προστεθούν μεταβλητές πέρα των X_1 και X_2, δεν είναι απαραίτητο, διότι επιπλέον μεταβλητές δεν βελτιστοποιούν περισσότερο το αποτέλεσμα.

7.9.2 ΒΗΜΑ ΠΡΟΣ ΒΗΜΑ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η διαδικασία της βήμα-προς-βήμα παλινδρόμησης (stepwise regression) προσθέτει κάθε φορά μία ανεξάρτητη μεταβλητή στο μοντέλο, ένα βήμα κάθε φορά. Η βήμα-προς-βήμα παλινδρόμηση μπορεί να περιγραφεί καλύτερα παραθέτοντας τα κύρια βήματα (αλγόριθμος), που περιλαμβάνονται στους υπολογισμούς.

1. Θεωρούνται όλες οι δυνατές απλές γραμμικές παλινδρομήσεις. Η μεταβλητή πρόβλεψης, που ερμηνεύει το μεγαλύτερο σημαντικό ποσοστό της μεταβολής στο Y (έχει το μεγαλύτερο r άρα και r^2), είναι η πρώτη μεταβλητή, που εισάγεται στην εξίσωση παλινδρόμησης.
2. Η επόμενη μεταβλητή, που εισάγεται, είναι αυτή που έχει τη μεγαλύτερη σημαντική συνεισφορά στο άθροισμα των τετραγώνων της παλινδρόμησης (SSR). Η σημαντικότητα της συνεισφοράς καθορίζεται από τον F ή t έλεγχο. Η τιμή του στατιστικού F, που πρέπει να είναι η ελάχιστη, πριν η συνεισφορά μιας μεταβλητής κριθεί σημαντική, συχνά καλείται **F to enter**.
3. Όταν μία επιπλέον μεταβλητή εισαχθεί στην εξίσωση, οι μεμονωμένες συνεισφορές στο άθροισμα των τετραγώνων της παλινδρόμησης (SSR) των υπόλοιπων μεταβλητών, που είναι ήδη

στην εξίσωση ελέγχονται για τη σημαντικότητά τους χρησιμοποιώντας τους F ή t ελέγχους. Αν η στατιστική F είναι μικρότερη από μία τιμή, που λέγεται **F to remove**, η μεταβλητή διαγράφεται από την εξίσωση παλινδρόμησης.

4. Τα βήματα 2 και 3 επαναλαμβάνονται μέχρι όλες οι δυνατές προσθήκες να είναι μη σημαντικές και όλες οι δυνατές διαγραφές να είναι σημαντικές. Σε αυτό το σημείο, η επιλογή σταματάει.

Η βήμα – προς – βήμα παλινδρόμηση επιτρέπει στις μεταβλητές πρόβλεψης να εισαχθούν, ή να διαγραφούν στη συνάρτηση παλινδρόμησης σε διαφορετικά στάδια της εξέλιξή της. Μία ανεξάρτητη μεταβλητή αφαιρείται από το μοντέλο αν δεν συνεχίζει να έχει σημαντική συνεισφορά, όταν μία νέα μεταβλητή προστίθεται.

Ο χρήστης ενός προγράμματος βήμα-προς-βήμα παλινδρόμησης παρέχει τις τιμές, που αποφασίζουν, πότε μία μεταβλητή επιτρέπεται να εισαχθεί και πότε μία μεταβλητή αφαιρείται. Δεδομένου ότι τα στατιστικά F, που χρησιμοποιήθηκαν στην βήμα-προς-βήμα παλινδρόμηση είναι τέτοια, έτσι ώστε $F=t^2$, όπου το t είναι το στατιστικό t για τον έλεγχο της σημαντικότητας μίας μεταβλητής πρόβλεψης, το F=4 (αντιστοιχεί σε $|t|=2$) είναι μία κοινή επιλογή για το **F to enter** και το **F to remove**. Ένα **F to enter** με τιμή 4 είναι απαραίτητα ισοδύναμο με τον έλεγχο της σημαντικότητας μίας μεταβλητής πρόβλεψης στο επίπεδο σημαντικότητας 5%. Το βήμα-προς-βήμα πρόγραμμα του Minitab επιτρέπει στο χρήστη να επιλέξει ένα α επίπεδο για να εισάγει και να αφαιρεί μεταβλητές ή την F τιμή για την εισαγωγή και διαγραφή μεταβλητών. Χρησιμοποιώντας μία α τιμή 0,05 είναι περίπου ισοδύναμο με το να χρησιμοποιείς F=4. Οι δεδομένες τιμές στο Minitab είναι α=0,15 και F=4.

Το αποτέλεσμα μίας βήμα-προς-βήμα διαδικασίας είναι ένα μοντέλο, που περιέχει μόνο ανεξάρτητες μεταβλητές με t τιμές, που είναι σημαντικές στο συγκεκριμένο επίπεδο σημαντικότητας. Ωστόσο, εξαιτίας της βήμα-προς-βήμα εξέλιξης, δεν υπάρχει εγγύηση, ότι η βήμα-προς-βήμα παλινδρόμηση θα επιλέξει, για παράδειγμα, τις τρεις καλύτερες μεταβλητές για πρόβλεψη. Επιπλέον, μία αυτόματη μέθοδος επιλογής δεν είναι ικανή να κάνει υποδείξεις, όταν μετασχηματισμοί των μεταβλητών είναι χρήσιμοι, ούτε αποφεύγει απαραίτητα ένα πρόβλημα πολυσυγγραμικότητας. Τέλος, η βήμα-προς-βήμα παλινδρόμηση δεν μπορεί να δημιουργήσει σημαντικές μεταβλητές, που δεν παρέχονται από το χρήστη. Είναι απαραίτητο να σκέφτεται κανείς προσεκτικά σχετικά με την συλλογή των ανεξάρτητων μεταβλητών που παρέχονται σε ένα πρόγραμμα βήμα – προς – βήμα παλινδρόμησης. Η βήμα – προς – βήμα διαδικασία παρουσιάζεται στο ακόλουθο Παράδειγμα.

Παράδειγμα

Ας λύσουμε πάλι το πρόβλημά χρησιμοποιώντας την βήμα-προς-βήμα παλινδρόμηση.

Εξετάζουμε τον πίνακα συσχέτισης και αποφασίζουμε, ότι όταν τρέχουμε την βήμα-προς-βήμα ανάλυση, η μεταβλητή της ηλικίας θα εισαχθεί πρώτα στο μοντέλο, επειδή έχει τη μεγαλύτερη συσχέτιση με τις πωλήσεις ($r_{1,3}$=0,798) και ερμηνεύει το 63,7% ($0,798^2$) της μεταβολής στις πωλήσεις.

Correlations: Sales; Aptitude; Age; Anxiety; Exp.; GPA

	Sales	Aptitude	Age	Anxiety	Exp.
Aptitude	0,676				
Age	0,798	0,228			
Anxiety	-0,296	-0,222	-0,287		
Exp.	0,550	0,350	0,540	-0,279	
GPA	0,622	0,318	0,695	-0,244	0,312

Παρατηρούμε ότι το σκορ στο τεστ ικανότητας πιθανότατα θα εισαχθεί στο μοντέλο δεύτερο επειδή σχετίζεται ισχυρά με τις πωλήσεις ($r_{1,2}$=0,676), αλλά δεν σχετίζεται υψηλά με τη μεταβλητή της ηλικίας ($r_{2,3}$=0,228), η οποία είναι ήδη στο μοντέλο.

Παρατηρούμε επίσης, ότι οι υπόλοιπες μεταβλητές πιθανότατα, δεν θα αξιολογηθούν σαν καλές μεταβλητές πρόβλεψης. Το σκορ στο τεστ άγχους δεν θα είναι μία καλή μεταβλητή πρόβλεψης, επειδή δεν σχετίζεται καλά με τις πωλήσεις ($r_{1,4}$=-0,296). Οι μεταβλητές της εμπειρίας και του βαθμού μπορεί να έχουν πιθανότητες σαν καλές μεταβλητές πρόβλεψης ($r_{1,5}$=0,550 και $r_{1,6}$=0,622, αντίστοιχα). Ωστόσο, και οι δύο αυτές μεταβλητές πρόβλεψης έχουν ένα πιθανό πρόβλημα πολυσυγγραμικότητας με τη μεταβλητή της ηλικίας ($r_{3,5}$=0,540 και $r_{3,6}$=0,695 αντίστοιχα).

Τα αποτελέσματα της βήμα-προς-βήμα παλινδρόμησης φαίνονται στον ακόλουθο Πίνακα. Η βήμα-προς-βήμα ανάλυση γίνεται σύμφωνα με τα βήματα που ακολουθούν:

Stepwise Regression: Sales versus Aptitude; Age; Anxiety; Exp.; GPA

```
Alpha-to-Enter: 0,05 Alpha-to-Remove: 0,05
Response is Sales on 5 predictors, with N = 30

Step          1       2
Constant  -100,85 -86,79

Age           6,97   5,93
T-Value       7,01  10,60
P-Value       0,000  0,000

Aptitude             0,200
T-Value              8,13
P-Value              0,000
```

```
S             6,85   3,75
R-Sq         63,70  89,48
R-Sq(adj)    62,41  88,70
```

Βήμα 1: Το μοντέλο μετά το Βήμα 1 είναι:

$$\text{Sales} = -100,85 + 6,97\ (\text{Age})$$

Όπως είχαμε σκεφτεί η μεταβλητή της ηλικίας εισάγεται πρώτα στο μοντέλο και ερμηνεύει το 63,7% της διασποράς στις πωλήσεις. Δεδομένου ότι η p-τιμή 0,000 είναι μικρότερη από την α τιμή 0,05, η ηλικία προστίθεται στο μοντέλο. Η απόφαση του Minitab δείχνει, ότι διάλεξε να εισάγει μία μεταβλητή, που η p-τιμή της είναι μικρότερη από το α=0,05.

Να σημειωθεί ότι t=7,01>2,048, το άνω 0,25 σημείο της t κατανομής με 28 (n-k-1=30-1-1) βαθμούς ελευθερίας. Έτσι, στο 0,05 επίπεδο σημαντικότητας, η υπόθεση $H_0:\beta_1=0$ απορρίπτεται για χάρη της $H_1:\beta_1\neq 0$. Δεδομένου ότι $t^2=F$ ή $2,048^2=4,19$, ένα **F to enter** με τιμή 4 είναι επίσης απαραίτητα ισοδύναμο για έλεγχο της σημαντικότητας μίας μεταβλητής πρόβλεψης στο 5% επίπεδο. Σε αυτή την περίπτωση, δεδομένου ότι ο συντελεστής της μεταβλητής της ηλικίας είναι σημαντικά διαφορετικός από το μηδέν, η ηλικία εισάγεται στην εξίσωση παλινδρόμησης και η διαδικασία προχωράει στο βήμα 2.

Βήμα 2: Το μοντέλο μετά το Βήμα 2 είναι:

$$\text{Sales} = -86,79 + 5,93(\text{Age}) + 0,200(\text{Aptitude})$$

Αυτό το μοντέλο ερμηνεύει το 89,48% της μεταβολής στις πωλήσεις.

Η μηδενική υπόθεση και η εναλλακτική υπόθεση, που καθορίζουν, αν ο συντελεστής παλινδρόμησης του σκορ στο τεστ ικανότητας είναι σημαντικά διαφορετικός από το μηδέν, είναι:

$$H_0:\beta_2=0$$

$H_1: \beta_2 \neq 0$

Ξανά, η p-τιμή 0,000 είναι μικρότερη από την α τιμή 0,05, και το σκορ στο τεστ ικανότητας προστίθεται στο μοντέλο. Ο συντελεστής του σκορ στο τεστ ικανότητας στην παλινδρόμηση είναι σημαντικά διαφορετικός από το μηδέν, και η πιθανότητα ότι αυτό συνέβη εξαιτίας του σφάλματος τυχαίου δείγματος είναι σχεδόν μηδέν. Αυτό το αποτέλεσμα σημαίνει ότι το σκορ στο τεστ ικανότητας είναι μία σημαντική μεταβλητή, όταν χρησιμοποιείται σε συνδυασμό με την ηλικία.

Το κρίσιμο στατιστικό t βασίζεται σε 27 (n-k-1=30-2-1) βαθμούς ελευθερίας και είναι 2,052. Χρησιμοποιώντας το t τεστ, η κενή υπόθεση απορρίπτεται επίσης. Να σημειωθεί ότι η p-τιμή για το στατιστικό t της μεταβλητής της ηλικίας, 0,000, παραμένει πολύ μικρό. Η ηλικία είναι ακόμα σημαντική μεταβλητή πρόβλεψης για τις πωλήσεις. Η διαδικασία τώρα προχωράει στο Βήμα 3.

Βήμα 3: Ο υπολογιστής τώρα σκέφτεται να προσθέσει μία τρίτη μεταβλητή πρόβλεψης, δεδομένου ότι η X_1 (ηλικία) και η X_2 (σκορ στο τεστ ικανότητας) είναι στην εξίσωση παλινδρόμησης. Καμία από τις υπόλοιπες ανεξάρτητες μεταβλητές δεν είναι σημαντικές (έχουν p-τιμή μεγαλύτερη από 0,05) σε συνδυασμό με το X_1 και το X_2, και έτσι η βήμα-προς-βήμα διαδικασία ολοκληρώνεται.

Το τελικό μοντέλο, που επιλέγεται από την βήμα-προς-βήμα διαδικασία, είναι το μοντέλο δύο μεταβλητών πρόβλεψης, που δίνεται στο Βήμα 2.

7.9.3 ΑΝΑΣΤΡΟΦΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Κατά την ανάστροφη παλινδρόμηση εισέρχονται από την αρχή όλες οι ανεξάρτητες μεταβλητές και αφαιρούνται μία κάθε φορά οι μη

στατιστικά σημαντικές ανεξάρτητες μεταβλητές, ξεκινώντας με αυτή που έχει την μεγαλύτερη p-τιμή. Μετά από κάθε αφαίρεση ανεξάρτητης μεταβλητής γίνεται παλινδρόμηση για τον υπολογισμό των συντελεστών των εναπομεινάντων μεταβλητών και των στατιστικών τους και επαναλαμβάνεται η διαδικασία.

Stepwise Regression: Sales versus Aptitude; Age; Anxiety; Exp.; GPA

Backward elimination. Alpha-to-Remove: 0,1
Response is Sales on 5 predictors, with N = 30

Step	1	2	3	4
Constant	-90,63	-90,49	-89,42	-86,79
Aptitude	0,203	0,203	0,202	0,200
T-Value	7,19	7,67	7,87	8,13
P-Value	0,000	0,000	0,000	0,000
Age	6,15	6,15	6,12	5,93
T-Value	6,68	7,72	7,94	10,60
P-Value	0,000	0,000	0,000	0,000
Anxiety	0,11	0,11		
T-Value	0,22	0,23		
P-Value	0,825	0,819		
Exp.	-0,01			
T-Value	-0,02			
P-Value	0,987			
GPA	-0,6	-0,6	-0,6	
T-Value	-0,33	-0,34	-0,36	
P-Value	0,742	0,735	0,725	
S	3,97	3,89	3,82	3,75
R-Sq	89,55	89,55	89,53	89,48
R-Sq(adj)	87,38	87,88	88,32	88,70

7.9.4 ΕΜΠΡΟΣΘΙΑ ΠΑΛΙΝΔΡΟΜΗΣΗ

Κατά αυτήν μόνο εισάγονται νέες ανεξάρτητες μεταβλητές, που είναι σημαντικές. Δεν αφαιρεί όμως ανεξάρτητες μεταβλητές στην περίπτωση, που με την εισαγωγή μιας νέας χαλάσει η στατιστική κάποιας από τις προηγούμενες, που έχει εισάγει.

Stepwise Regression: Sales versus Aptitude; Age; Anxiety; Exp.; GPA

```
Forward selection.  Alpha-to-Enter: 0,25
Response is Sales on 5 predictors, with N = 30

Step              1       2
Constant    -100,85  -86,79

Age            6,97    5,93
T-Value        7,01   10,60
P-Value       0,000   0,000

Aptitude              0,200
T-Value                8,13
P-Value               0,000

S              6,85    3,75
R-Sq          63,70   89,48
R-Sq(adj)     62,41   88,70
```

7.9.5 ΤΕΛΙΚΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ ΓΙΑ ΤΗ ΒΗΜΑ – ΠΡΟΣ – ΒΗΜΑ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η τεχνική της βήμα-προς-βήμα παλινδρόμησης είναι εξαιρετικά εύκολη στη χρήση. Δυστυχώς, είναι επίσης εξαιρετικά εύκολο να χρησιμοποιηθεί λανθασμένα. Οι αναλυτές, που αναπτύσσουν ένα μοντέλο παλινδρόμησης. συχνά παράγουν ένα μεγάλο σύνολο από πιθανές ανεξάρτητες μεταβλητές και μετά αφήνουν τη διαδικασία της βήμα-προς-βήμα παλινδρόμησης να καθορίσει, ποιες από αυτές είναι σημαντικές. Το πρόβλημα είναι ότι, όταν αναλύεται ένα μεγάλο σύνολο

από ανεξάρτητες μεταβλητές, πραγματοποιούνται πολλά t τεστ και είναι πιθανό να προκύψει ένα σφάλμα (μία μη σημαντική μεταβλητή να προστίθεται). Αυτό σημαίνει, ότι το τελικό μοντέλο μπορεί να περιέχει μία μεταβλητή, η οποία να μην σχετίζεται γραμμικά με την εξαρτημένη μεταβλητή και η οποία εισήχθη στο μοντέλο κατά τύχη.

Όπως αναφέρθηκε και προηγουμένως, ένα ακόμη πρόβλημα περιλαμβάνει την αρχική επιλογή των πιθανών ανεξάρτητων μεταβλητών. Όταν αυτές οι μεταβλητές επιλέγονται, παραλείπονται συχνά όροι υψηλότερης-τάξης (καμπυλόγραμμοι, μη γραμμικοί), για να κρατηθεί ο αριθμός των μεταβλητών εύκολος στη διαχείριση. Κατά συνέπεια, μερικές σημαντικές μεταβλητές μπορεί αρχικά να παραλειφθούν από το μοντέλο. Γίνεται φανερό, ότι η ενστικτώδης επιλογή του αναλυτή των αρχικών ανεξάρτητων μεταβλητών είναι κρίσιμη στην ανάπτυξη ενός επιτυχημένου μοντέλου παλινδρόμησης.

7.10 ΔΙΑΓΝΩΣΤΙΚΕΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΑΙ ΑΝΑΛΥΣΗ ΚΑΤΑΛΟΙΠΩΝ

Μία ανάλυση παλινδρόμησης δεν είναι ολοκληρωμένη μέχρι κάποιος να πειστεί, ότι το μοντέλο αντιπροσωπεύει επαρκώς τα δεδομένα. Είναι αναγκαίο λοιπόν να εξεταστεί η συνέπεια του μοντέλου, προτού χρησιμοποιηθεί στη λήψη αποφάσεων.

Μία εξέταση των καταλοίπων είναι ένα σημαντικό συστατικό του καθορισμού της επάρκειας του μοντέλου. Μερικά διαγράμματα καταλοίπων, που σχεδιάζονται για να ελέγξουν τις υποθέσεις του μοντέλου παλινδρόμησης, παρουσιάστηκαν στο Κεφάλαιο 6. Επίσης, αν μοντέλα παλινδρόμησης χρησιμοποιούνται με δεδομένα χρονοσειρών είναι σημαντικό να υπολογίζονται οι αυτοσυσχετίσεις καταλοίπων για να

ελεγχθεί η υπόθεση ανεξαρτησίας. Αναφορές (και αποφάσεις) που γίνονται με μοντέλα, που δεν συμφωνούν με τις υποθέσεις παλινδρόμησης, μπορεί να είναι παραπλανητικές. Για παράδειγμα, μπορούμε να συμπεράνουμε, ότι η αλλαγή μιας ανεξάρτητης μεταβλητής πρόβλεψης θα προκαλέσει μία συγκεκριμένη αλλαγή στην απόκριση, όταν στην πραγματικότητα κάτι τέτοιο δεν θα γίνει. Μπορεί να συμπεράνουμε, ότι μία πρόβλεψη είναι πολύ πιθανό (95% εμπιστοσύνη) να είναι μέσα στο 2% της μελλοντικής απόκρισης, όταν στην πραγματικότητα η πραγματική εμπιστοσύνη είναι πολύ μικρότερη.

Σε αυτή την ενότητα, θα συζητηθούν μερικά επιπρόσθετα εργαλεία, τα οποία μπορούν να χρησιμοποιηθούν για να αξιολογήσουν ένα μοντέλο παλινδρόμησης. Αυτά τα εργαλεία σχεδιάζονται για να αναγνωρίζουν παρατηρήσεις απομακρυσμένες ή ακραίες (αυτές οι παρατηρήσεις είναι καλά ξεχωρισμένες από τα υπόλοιπα δεδομένα). Ακραίες παρατηρήσεις είναι συχνά κρυμμένες από την διαδικασία προσαρμογής και μπορεί να μην είναι εύκολο να εντοπιστούν από μία εξέταση των διαγραμμάτων καταλοίπων. Ωστόσο μπορεί να έχουν σημαντικό ρόλο στο καθορισμό της προσαρμοσμένης συνάρτησης παλινδρόμησης. Είναι σημαντικό να μελετήσει κανείς ακραίες παρατηρήσεις, για να αποφασίσει αν πρέπει να διατηρηθούν, ή να αποκλειστούν και αν τελικά διατηρηθούν, να αποφασίσει για το αν η επιρροή τους θα έπρεπε να μειωθεί στην αναθεώρηση της διαδικασίας προσαρμογής, ή της συνάρτησης παλινδρόμησης.

Ένας υπολογισμός της επιρροής των i σημείων δεδομένων στη προσαρμοσμένη συνάρτηση παλινδρόμησης παρέχεται από το **leverage** $\mathbf{h_{ii}}$**.** Το **leverage** εξαρτάται μόνο από τις μεταβλητές πρόβλεψης (ανεξάρτητες μεταβλητές). Δεν εξαρτάται από την εξαρτημένη

μεταβλητή. Για την απλή γραμμική παλινδρόμηση με μία μεταβλητή πρόβλεψης, X, έχουμε:

$$h_{ii}=\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2}$$

Με k μεταβλητές πρόβλεψης για το i **leverage** είναι πιο πολύπλοκος ο τύπος. Ωστόσο, μπορεί να δειχθεί, ότι $0<h_{ii}<1$ και ότι το μέσο **leverage** είναι $\bar{h}=(k+1)/n$.

Αν το i σημείο δεδομένων έχει υψηλό **leverage** (h_{ii} είναι κοντά στο 1), η προσαρμοσμένη απόκριση $\hat{Y}$, σε εκείνο το X καθορίζεται σχεδόν ολοκληρωτικά από το Y_i, με τα υπόλοιπα δεδομένα να έχουν πολύ μικρή επιρροή. Το υψηλό **leverage** σημείο δεδομένων είναι επίσης ακραίο ανάμεσα στα X (μακριά από τους άλλους συνδυασμούς των X τιμών). Ένας εμπειρικός κανόνας υποδεικνύει, ότι το h_{ii} είναι αρκετά μεγάλο, αν $h_{ii}\geq 3(k+1)/n$.

Η ανίχνευση των ακραίων Y τιμών βασίζεται στο μέγεθος των καταλοίπων $e=Y-\hat{Y}$. Μεγάλα κατάλοιπα υποδεικνύουν, ότι μία Y τιμή είναι «μακριά» από την προσαρμοσμένη ή προβλεφθείσα τιμή, $\hat{Y}$. Ένα μεγάλο κατάλοιπο θα φανεί σε ένα ιστόγραμμα των καταλοίπων σαν μία τιμή μακριά από το μηδέν (προς οποιαδήποτε κατεύθυνση). Ένα μεγάλο κατάλοιπο θα φανεί σε ένα διάγραμμα των καταλοίπων με τις προσαρμοσμένες τιμές σαν ένα σημείο μακριά πάνω ή κάτω από τον οριζόντιο άξονα.

Τα πακέτα λογισμικού, όπως το Minitab, "σημαδεύουν" τα σημεία δεδομένων με ακραίες Y τιμές, υπολογίζοντας τα «τυποποιημένα» κατάλοιπα (Standardized Residuals) και αναγνωρίζοντας σημεία με μεγάλα τυποποιημένα κατάλοιπα.

Μία τυποποίηση βασίζεται στο γεγονός, ότι τα κατάλοιπα έχουν

τυπικές αποκλίσεις της μορφής:

$$s_{e_i} = s_{y \cdot x's} \sqrt{1 - h_{ii}}$$

όπου, $s_{y \cdot x's} = \sqrt{MSE}$ είναι το τυπικό σφάλμα εκτίμησης και το h_{ii} είναι το **leverage,** που σχετίζεται με το i σημείο δεδομένων. Το τυποποιημένο κατάλοιπο είναι τότε:

$$\frac{e_i}{s_{e_i}} = \frac{e_i}{s_{y \cdot x's} \sqrt{1 - h_{ii}}}$$

Τα τυποποιημένα κατάλοιπα όλα έχουν διασπορά 1. Ένα τυποποιημένο κατάλοιπο θεωρείται μεγάλο (ακραία απόκριση) αν:

$$\left| \frac{e_i}{s_{e_i}} \right| > 2$$

Οι Y τιμές, που αντιστοιχούν στα σημεία δεδομένων με μεγάλα τυποποιημένα κατάλοιπα, μπορούν να επηρεάσουν σημαντικά τη θέση της προσαρμοσμένης συνάρτησης παλινδρόμησης.

Παράδειγμα

Ο μισθός του διευθύνοντα συμβούλου για διάφορες εταιρείες διαφέρει σημαντικά από εταιρία σε εταιρία, αλλά τα δεδομένα υποδεικνύουν, ότι ο μισθός μπορεί να ερμηνευθεί με βάση 1) τις πωλήσεις της εταιρίας, 2) την εμπειρία του διευθύνοντα συμβούλου, 3) το επίπεδο μόρφωσης και 4) το μερίδιο ιδιοκτησίας στην εταιρία. Σε μία μελέτη, χρησιμοποιήθηκαν 50 εταιρίες για να αναπτυχθεί ένα μοντέλο πολυμεταβλητής παλινδρόμησης, το οποίο συνέδεε την αποζημίωση των διευθυνόντων συμβούλων με μερικές μεταβλητές πρόβλεψης, όπως οι πωλήσεις, τα κέρδη, η ηλικία, η εμπειρία, το επαγγελματικό υπόβαθρο, το επίπεδο μόρφωσης και το μερίδιο ιδιοκτησίας.

Αφού παραλείφθηκαν οι μη σημαντικές μεταβλητές πρόβλεψης, η

τελική προσαρμοσμένη συνάρτηση παλινδρόμησης ήταν:

$$\hat{Y} = 5{,}52 - 0{,}467\, X_1 + 0{,}263\, X_2$$

όπου,

- Y=ο λογάριθμος της αποζημίωσης του διευθύνοντος συμβούλου.
- X_1=η μεταβλητή δείκτης για το επίπεδο μόρφωσης.
- X_2=ο λογάριθμος των πωλήσεων της εταιρίας.

Το Minitab αναγνώρισε τρεις παρατηρήσεις από την ανάλυση παλινδρόμησης, που είχαν, είτε μεγάλα τυποποιημένα κατάλοιπα, είτε μεγάλο **leverage**.

Παρατή-ρηση	**Educate**	**LnComp**	**Fit**	**StDev Fit**	**Residual**	**St Resid**
14	1,00	6,0568	7,0995	0,0949	1,0427	2,09R
25	0,00	8,1342	7,9937	0,2224	0,1405	0,31X
33	0,00	6,3969	7,3912	0,2032	0,9943	2,13R

Το R δείχνει μία παρατήρηση με μεγάλο τυποποιημένο κατάλοιπο.

Το X δείχνει μία παρατήρηση, της οποίας η X τιμή της, της δίνει μεγάλη επιρροή.

Οι παρατηρήσεις 14 και 33 έχουν μεγάλα τυποποιημένα κατάλοιπα. Η προσαρμοσμένη συνάρτηση παλινδρόμησης προβλέπει την αποζημίωση (log), που είναι πολύ μεγάλη για αυτούς τους δύο διευθύνοντες συμβούλους. Μία εξέταση ολόκληρου του συνόλου δεδομένων δείχνει, ότι εκείνοι οι διευθύνοντες σύμβουλοι έχουν στην κατοχή τους αρκετά μεγάλα ποσοστά των μετοχών της εταιρίας. Στην περίπτωση 14 έχει στην κατοχή του περισσότερο από το 10% των μετοχών της εταιρίας, και στην περίπτωση 33 έχει στην κατοχή του περισσότερο από το 17% των μετοχών της εταιρίας. Δεδομένου ότι η ποσότητα των μετοχών, που έχει ο καθένας στην κατοχή του, δεν περιλαμβάνεται σαν μεταβλητή στη συνάρτηση παλινδρόμησης, δεν

μπορεί να χρησιμοποιηθεί για να προσαρμόσει την πρόβλεψη της αποζημίωσης, που καθορίζεται από τη μόρφωση και τις πωλήσεις της εταιρείας. Αν και η μόρφωση και οι (log) πωλήσεις δεν προβλέπουν την αποζημίωση εκείνων των δύο ατόμων τόσο καλά, όσο προβλέπουν των υπολοίπων, δεν φαίνεται να υπάρχει λόγος να τις παραλείψουμε από τη μελέτη.

Η παρατήρηση 25 ξεχωρίζει γιατί το **leverage** για αυτό το σημείο δεδομένων είναι μεγαλύτερο από το 3(k+1)/n=3(3)/50=0,18. Αυτός ο διευθύνων σύμβουλος δεν έχει πτυχίο (Educate=0), αλλά βρίσκεται σε μία εταιρία με υψηλές πωλήσεις (LnSales=9,394). Ο συνδυασμός αυτός (0, 9,394) απέχει πολύ από το σημείο $(\bar{X}_{1,}\bar{X}_{2})$. Έτσι είναι ακραίο μεταξύ των ζευγαριών των X. Η απόκριση, που σχετίζεται με αυτά τα X, θα έχει μεγάλη επιρροή στον καθορισμό της προσαρμοσμένης συνάρτησης παλινδρόμησης. (Να σημειωθεί ότι το τυποποιημένο κατάλοιπο για αυτό το σημείο δεδομένων είναι μικρό, το οποίο υποδεικνύει, ότι η προβλεφθείσα ή προσαρμοσμένη (log) αποζημίωση είναι κοντά στην πραγματική τιμή). Αυτός ο συγκεκριμένος διευθύνων σύμβουλος έχει 30 χρόνια εμπειρίας, περισσότερη εμπειρία από όλους εκτός από έναν διευθύνοντα σύμβολο στο σύνολο δεδομένων. Αυτή η παρατήρηση είναι ισχυρή, οπότε δεν υπάρχει λόγος να τη διαγράψουμε.

Το leverage μας λέει αν μία παρατήρηση έχει ασυνήθιστες τιμές μεταβλητών πρόβλεψης, και το τυποποιημένο κατάλοιπο μας λέει αν μία παρατήρηση έχει ασυνήθιστη απόκριση σε σχέση με την προσαρμογή. Αυτές οι ποσότητες μπορούν να συνδυαστούν σε ένα συνολικό υπολογισμό της επιρροής, γνωστής σαν Cook's distance.

7.11 ΠΡΟΕΙΔΟΠΟΙΗΣΕΙΣ ΠΡΟΒΛΕΨΗΣ

Τελειώνουμε τη συζήτηση του κεφαλαίου αυτού με κάποια γενικά σχόλια πρακτικού ενδιαφέροντος.

7.11.1 OVERFITTING

Το overfitting αναφέρεται στην προσθήκη ανεξάρτητων μεταβλητών στην συνάρτηση παλινδρόμησης, οι οποίες σε μεγάλο βαθμό, επεξηγούν όλες τις "εκκεντρικότητες-ιδιαιτερότητες" των δεδομένων του υπό ανάλυση δείγματος. Ή με άλλα λόγια χρησιμοποιούμε περισσότερες ανεξάρτητες μεταβλητές, από όσες χρειαζόμαστε.

Όταν ένα τέτοιο μοντέλο εφαρμόζεται σε άλλα δεδομένα του ίδιου πληθυσμού, δίνει χειρότερα αποτελέσματα από τα αρχικά της προσαρμογής.

Το overfitting εμφανίζεται συνήθως σε μικρά δείγματα. Αν έχουμε n ανεξάρτητες μεταβλητές θα χρειαστούμε τουλάχιστον nx10 τιμές. Για να ελέγξουμε το μοντέλο μας για overfitting, εφαρμόζουμε το μοντέλο σε άλλο μέρος του δείγματος και αν τα σφάλματα είναι μεγαλύτερα από αυτά της προσαρμογής, τότε λέμε ότι λαμβάνει χώρα overfitting.

7.11.2 ΧΡΗΣΙΜΕΣ ΠΑΛΙΝΔΡΟΜΗΣΕΙΣ – ΜΕΓΑΛΟΙ F-ΛΟΓΟΙ

Μία στατιστικά σημαντική παλινδρόμηση δεν σημαίνει, ότι είναι αναγκαία και χρήσιμη. Σε ένα σχετικά μεγάλο δείγμα δεν είναι ασύνηθες να έχουμε μεγάλους F-λόγους και μικρό R^2. Σε αυτή τη περίπτωση έχουμε σημαντική παλινδρόμηση αλλά με χαμηλή εξήγηση της μεταβλητότητας. Ένας πρακτικός κανόνας υποδεικνύει, ότι ο υπολογισμένος F-λόγος πρέπει να είναι τουλάχιστο τέσσερις φορές μεγαλύτερος της F-τιμής, που αντιστοιχεί στο επίπεδο σημαντικότητας α

%, πριν χρησιμοποιήσουμε τη παλινδρόμηση για προβλέψεις.

Το κριτήριο τέσσερις φορές έρχεται από το επιχείρημα, ότι το διάστημα πρόβλεψης για όλα τα X, πρέπει να είναι τέσσερις φορές το σφάλμα πρόβλεψης $(0 \pm 2\, x\, s_{y \cdot x' s})$. Οπότε από τη σχέση, που δίνει την F με τον συντελεστή προσδιορισμού, υπολογίζουμε τον ελάχιστο συντελεστή προσδιορισμού.

7.12 CASE 7-1 THE BOND MARKET

Η Judie Johnson, Αντιπρόεδρος Οικονομικών μιας μεγάλης ιδιωτικής επενδυτικής εταιρίας κοινής ωφελείας στη Βορειοδυτική Ευρώπη, αντιμετώπισε ένα πρόβλημα χρηματοδότησης. Προκειμένου να συνεχιστεί η παραγωγή καύσιμου άνθρακα, η εταιρία χρειαζόταν χρήματα για να αποπληρώσει τα βραχυπρόθεσμα χρέη της.

Η Judie επικεντρώθηκε στην εκτίμηση της αγοράς ομολόγων κατά τα επόμενα 10 ή 30 χρόνια. Το βασικό δίλημμα ήταν εάν θα χρησιμοποιηθούν ίδια κεφάλαια ή μακροχρόνιος δανεισμός. Για να παρθεί η εν λόγω απόφαση, ήταν απαραίτητη μία αξιόπιστη πρόβλεψη του επιτοκίου που θα αποπληρωνόταν με το πέρασμα του χρόνου.

Η Judie συγκάλεσε σε σύσκεψη το προσωπικό που ασχολούνταν με τα οικονομικά, με σκοπό να συζητηθεί το πρόβλημα της αγοράς ομολόγων. Ένα μέλος της ομάδας της, ο Ron Peterson, πρόσφατα απόφοιτος ενός M.B.A προγράμματος, πρότεινε ένα μοντέλο πολυμεταβλητής παλινδρόμησης, το οποίο θα μπορούσε να χρησιμοποιηθεί για την πρόβλεψη των επιτοκίων των ομολόγων. Επειδή όμως η αντιπρόεδρος δεν ήταν εξοικειωμένη με την πολυμεταβλητή παλινδρόμηση, κατηύθυνε αλλού τη συζήτηση. Έπειτα από μία ώρα

αντιπαραγωγικής διάσκεψης, η Judie ζήτησε από τον Ron να της ετοιμάσει μία αναφορά την ερχόμενη δευτέρα.

Ο Ron γνώριζε ότι η λύση για την ανάπτυξη ενός καλού μοντέλου πρόβλεψης είναι η αναγνώριση των ανεξάρτητων μεταβλητών που σχετίζονται με το επιτόκιο που πληρώνεται από τη στιγμή που εκδίδεται το ομόλογο. Έπειτα από συζήτηση με διάφορους ανθρώπους αποφάσισε να αναπτύξει τις ακόλουθες μεταβλητές: τη χρησιμότητα της αξιολόγησης ομολόγων (utility's bond rating - Moody's), τη χρησιμότητα των αποδόσεων των σταθερών χρεώσεων (utility's ratio of earnings to fixed charges), τις αποδόσεις των ομολόγων του Υπουργείου Οικονομικών των ΗΠΑ, τον χρόνο αποπληρωμής των ομολόγων (the bond maturity – 10 or 30 years) και το βασικό επιτόκιο δανεισμού τη στιγμή της έκδοσης (prime leading rate at the time of issuance).

Ο Ron συγκέντρωσε δεδομένα, τα οποία πίστευε ότι πιθανόν να συσχετίζονται με τα επιτόκια των ομολόγων από την έκδοσή τους και για τα δύο προηγούμενα χρόνια. Αρχικά, είχε μία αβεβαιότητα για το πως θα χειριστεί τη χρησιμότητα της αξιολόγησης των ομολόγων. Τελικά αποφάσισε να θεωρήσει ως απαραίτητες, μόνο τις χρησιμότητες των οποίων οι αξιολογήσεις τους είναι ίδιες ή υψηλότερες από τις αξιολογήσεις της επιχείρησης για την οποία εργάζεται. Η απόφαση αυτή του παρείχε ένα δείγμα από 93 παρατηρήσεις προς ανάλυση. Όμως, και πάλι ήταν προβληματισμένος για την εγκυρότητα της χρήσης της αξιολόγησης των ομολόγων ως διάστημα δεδομένων. Έτσι, ρώτησε έναν καθηγητή στατιστικών από τον οποίο έμαθε ότι οι ψευδομεταβλητές ήταν αυτές που θα έλυναν το πρόβλημα. Για το λόγο αυτό, κωδικοποίησε τις αξιολογήσεις ομολόγων ακολούθως:

- X1 = 1, εάν η χρησιμότητα της αξιολόγησης του ομολόγου είναι A,

αλλιώς $X_1 = 0$.

- X2 = 1, εάν η χρησιμότητα της αξιολόγησης του ομολόγου είναι AA, αλλιώς $X_2 = 0$.

Εάν η χρησιμότητα της αξιολόγησης του ομολόγου είναι BAA, τότε $X_1 = 0$ και $X_2 = 0$. Το επόμενο βήμα για τον Ron ήταν να επιλέξει ένα πρόγραμμα πολυμεταβλητής παλινδρόμησης στον υπολογιστή, το οποίο θα ταίριαζε στα δεδομένα εισόδου. Οι παρακάτω μεταβλητές θα περιέχονται στην πλήρη εξίσωση του μοντέλου:

- Μεταβλητή 1: Y = το επιτόκιο το οποίο πληρώνεται τη στιγμή της έκδοσης ομολόγου.
- Μεταβλητή 2: $X_1 = 1$ εάν η χρησιμότητα της αξιολόγησης του ομολόγου είναι A.
- Μεταβλητή 3: $X_2 = 1$ εάν η χρησιμότητα της αξιολόγησης του ομολόγου είναι AA.
- Μεταβλητή 4: X_3 = η χρησιμότητα των αποδόσεων των σταθερών χρεώσεων.
- Μεταβλητή 5: X_4 = τα επιτόκια ομολόγων του Υπουργείου Οικονομικών των ΗΠΑ (10 και 30 χρόνια) από τη στιγμή της έκδοσης του ομολόγου.
- Μεταβλητή 6: X_5 = Ο χρόνος αποπληρωμής του ομολόγου (the bond maturity - 10 ή 30 χρόνια).
- Μεταβλητή 7: X_6 = το βασικό επιτόκιο δενεισμού τη στιγμή της έκδοσης.

Τα πραγματικά δεδομένα παρουσιάζονται στο παράρτημα Α.

Ο Ron αποφάσισε να αναλύσει τον πίνακα συσχέτισης μεταβλητών (Πίνακας 7-18). Από την ανάλυση προέκυψε μια ισχυρή θετική συσχέτιση μεταξύ του ρυθμού επιτοκίου (interest rate) που καταβάλλεται από την κοινότητα τη στιγμή έκδοσης του ομολόγου και τους ρυθμούς των ομολόγων του Υπουργείου Οικονομικών των ΗΠΑ, της τάξεως του r = 0,883. Επιπλέον, διαπίστωσε μια αρκετά υψηλή συσχέτιση μεταξύ της εξαρτημένης μεταβλητής και του βασικού επιτοκίου δανεισμού (prime, r = 0,596). Παράλληλα παρατήρησε πως αυτές οι δύο μεταβλητές πρόβλεψης σχετίζονται μεταξύ τους (ισχυρή πολυσυγγραμικότητα, r = 0,713). Η αρνητική συσχέτιση μεταξύ της εξαρτημένης μεταβλητής και του χρόνου αποπληρωμής του ομολόγου (maturity,10 ή 30 χρόνια), r = -0,221 είναι ένα αποτέλεσμα που έκανε αίσθηση στον Ron.

Στη συνέχεια ο Ron έτρεξε ένα παλινδρομικό μοντέλο που περιλάμβανε όλες τις μεταβλητές πρόβλεψης. Εξετάζοντας τις t και τις p τιμές, οι οποίες παρουσιάζονται στον πίνακα 7-19, ανέφερε πως ίσως η ανεξάρτητη μεταβλητή του ρυθμού των κερδών των παγίων τελών και σίγουρα η μεταβλητή του βασικού επιτοκίου δεν είχαν καμία συνεισφορά στην πρόβλεψη του επιτοκίου που καταβάλλεται τη στιγμή της έκδοσης ομολόγων.

Προκειμένου ο Ron να ελέγξει τα συμπεράσματα της πολυμεταβλητής παλινδρόμησης, αποφάσισε να πραγματοποιήσει μια βήμα-προς-βήμα παλινδρόμηση. Τα αποτελέσματα φαίνονται στον πίνακα 7- 20.

Παρόλο που η p τιμή (0,035) για το t στατιστικό του ρυθμού των κερδών των σταθερών χρεώσεων (ratio of earnings) είναι μικρότερη του 0,05, ο Ron αποφάσισε να αφαιρέσει αυτήν τη μεταβλητή από την εξίσωση παλινδρόμησης. Η απόφαση του αυτή ενισχύεται και από τη

μικρή συσχέτιση που παρουσιάζει με την εξαρτημένη μεταβλητή του ρυθμού επιτοκίου (r = 0,097). Τα αποτελέσματα που δίνει το Minitab σε αυτήν την περίπτωση φαίνονται στον πίνακα 7-21.

Ο Ron έγραψε μια αναφορά στην Judy που περιλάμβανε τις ακόλουθες παρατηρήσεις:

- Το καλύτερο μοντέλο, με εξίσωση παλινδρόμησης Επιτόκιο = – 1,28 – 0,929*Α – 1,18*ΑΑ + 1,23*Ρυθμοί ομολόγων + 0,0615*Χρόνος αποπληρωμής, ερμηνεύει το 90,6% της μεταβλητότητας του ρυθμού επιτοκίου.
- Το τυπικό σφάλμα της εκτίμησης είναι 0,53 στο σύνολο των n = 93 παρατηρήσεων. Αυτή η τιμή δηλώνει πως οι πραγματικές τιμές διαφέρουν από τις προσαρμοσμένες τιμές κατά την ποσότητα $\hat{Y}$± 2* 0,53 ή $\hat{Y}$± 1,06.
- Οι συντελεστές των ανεξάρτητων μεταβλητών είναι όλοι σημαντικοί (καθώς εμφανίζουν μικρές p τιμές) και φαίνονται να είναι αξιόπιστοι. Η πολυσυγγραμικότητα δεν αποτελεί πρόβλημα.

Ο Ron ήταν ικανοποιημένος με τα αποτελέσματα και αισθανόταν ότι η Judy θα ήταν το ίδιο ευχαριστημένη.

ΕΡΩΤΗΣΗ

1. **Τι ερωτήσεις πιστεύετε πως θα έκανε η Judy στον Ron;**

Ο Ron αφαιρώντας από το μοντέλο της πολυμεταβλητής παλινδρόμησης την ανεξάρτητη μεταβλητή του ρυθμού των κερδών των σταθερών χρεώσεων, πέτυχε να εξαλείψει το φαινόμενο της πολυσυγραμμικότητας, καθώς όλες οι τιμές του VIF και για τις τέσσερις μεταβλητές είναι πλησίον της μονάδας (1,313 1,338 1,460 και 1,438 αντίστοιχα) και επομένως δεν υπάρχει πολυσυγγραμμικότητα για τις μεταβλητές αυτές. Όπως άλλωστε και ο ίδιος υποστήριξε, φαίνεται το

αποτέλεσμα να είναι ικανοποιητικό. Άλλωστε, ερμηνεύει το 90,6% της μεταβλητότητας του ρυθμού επιτοκίου, ποσοστό που στο αρχικό μοντέλο ήταν στο 91,1% (ναι μεν υψηλότερο, αλλά με αμελητέα διαφορά).

Η Judy, ως μη γνώστης της συγκεκριμένης τεχνικής πρόβλεψης, αρχικά θα αναρωτιόταν για την επιλογή των ανεξαρτήτων μεταβλητών της μεθόδου. Επιπλέον, θα είχε ενστάσεις ως προς την απαλοιφή της ανεξάρτητης μεταβλητής, εφόσον η p τιμή της ήταν μικρότερη από την τιμή 0,05. Τα αποτελέσματα δικαιώνουν την επιλογή του Ron. Παράλληλα, παρατηρεί ότι διατηρούνται οι υψηλές τιμές t σημαντικότητας των ανεξάρτητων μεταβλητών, όπως και στο αρχικό μοντέλο. Η σημαντικότητα είναι υψηλή για τις ανεξάρτητες μεταβλητές των ρυθμών ομολόγων και του χρόνου αποπληρωμής ομολόγων.

Correlations: Int rate; A; AA; Ratio; Bondrate; Maturity; Prime

	Int rate	A	AA	Ratio	Bondrate	Maturity
A	-0,347					
AA	0,173	-0,399				
Ratio	0,097	0,037	0,577			
Bondrate	0,883	-0,256	0,291	0,253		
Maturity	-0,221	0,278	0,010	0,094	-0,477	
Prime	0,596	-0,152	0,342	0,255	0,713	-0,314

Πίνακας 7-18 Πίνακας Συσχέτισης μεταβλητών για την αγορά ομολόγων

Regression Analysis: Int rate versus A; AA; ...

The regression equation is

Int rate = - 1,03 - 0,829 A - 0,889 AA - 0,242 Ratio + 1,26 Bondrate + 0,0628 Maturity - 0,0031 Prime

Predictor	Coef	SE Coef	T	P	VIF
Constant	-1,0263	0,6572	-1,56	0,122	
A	-0,8285	0,1342	-6,17	0,000	1,524
AA	-0,8894	0,2248	-3,96	0,000	2,192

```
Ratio       -0,2417    0,1135  -2,13  0,036  1,797
Bondrate    1,25753   0,05964  21,08  0,000  2,526
Maturity   0,062839  0,006589   9,54  0,000  1,453
Prime      -0,00313   0,02720  -0,12  0,909  2,158

S = 0,523483   R-Sq = 91,1%   R-Sq(adj) = 90,5%

Analysis of Variance
Source          DF       SS      MS       F      P
Regression       6  240,907  40,151  146,52  0,000
Residual Error  86   23,567   0,274
Total           92  264,474

Unusual Observations
Obs     A  Int rate   Fit    SE Fit  Resid.  St Resid
 64  1,00   15,5000 13,7645 0,1627  1,7355    3,49R
 66  1,00   11,3000 13,4780 0,135  -2,1780   -4,31R
 67  1,00   12,3800 13,7897 0,1297 -1,4097   -2,78R
 68  1,00   12,1500 13,7097 0,0947 -1,5597   -3,03R

R denotes an observation with a large standardized
residual.
```

Πίνακας 7-19 Μοντέλο πολυμεταβλητής παλινδρόμησης για την αγορά ομολόγων

Stepwise Regression: Int rate versus A; AA; ...

```
Alpha-to-Enter: 0,05  Alpha-to-Remove: 0,05
Response is Int rate on 6 predictors, with N = 93

Step         1        2      3      4       5
Constant  1,9818 -0,7641 -0,7372 -0,2046  -1,0244
Bondrate  1,029   1,173   1,267   1,231    1,253
T-Value  17,95   20,44   23,43   24,78    27,21
P-Value   0,000   0,000   0,000   0,000    0,000

Maturity          0,0439  0,0537  0,0588   0,0629
T-Value           5,25    7,02    8,37     9,60
P-Value           0,000   0,000   0,000    0,000

Ratio                    -0,547  -0,518   -0,241
T-Value                  -5,08   -5,29    -2,14
P-Value                   0,000   0,000    0,035
```

A				-0,56	-0,83
T-Value				-4,49	-6,28
P-Value				0,000	0,000
AA					-0,89
T-Value					-4,10
P-Value					0,000
S	0,800	0,704	0,623	0,565	0,521
R-Sq	77,97	83,14	86,93	89,37	91,09
R-Sq(adj)	77,73	82,76	86,49	88,88	90,58
Mallows Cp	123,6	75,7	41,2	19,6	5,0

Πίνακας 7-20 Βήμα – προς – βήμα παλινδρόμηση για μελέτη της αγοράς ομολόγων

Regression Analysis: Int rate versus A; AA; Bondrate; Maturity

The regression equation is

Int rate = - 1,28 - 0,929 A - 1,18 AA + 1,23 Bondrate + 0,0615 Maturity

Predictor	Coef	SE Coef	T	P	VIF
Constant	-1,2765	0,6554	-1,95	0,055	
A	-0,9293	0,1264	-7,35	0,000	1,313
AA	-1,1751	0,1781	-6,60	0,000	1,338
Bondrate	1,23308	0,04600	26,81	0,000	1,460
Maturity	0,061474	0,006649	9,25	0,000	1,438

S = 0,530959 R-Sq = 90,6% R-Sq(adj) = 90,2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	239,665	59,916	212,53	0,000
Residual Error	88	24,809	0,282		
Total	92	264,474			

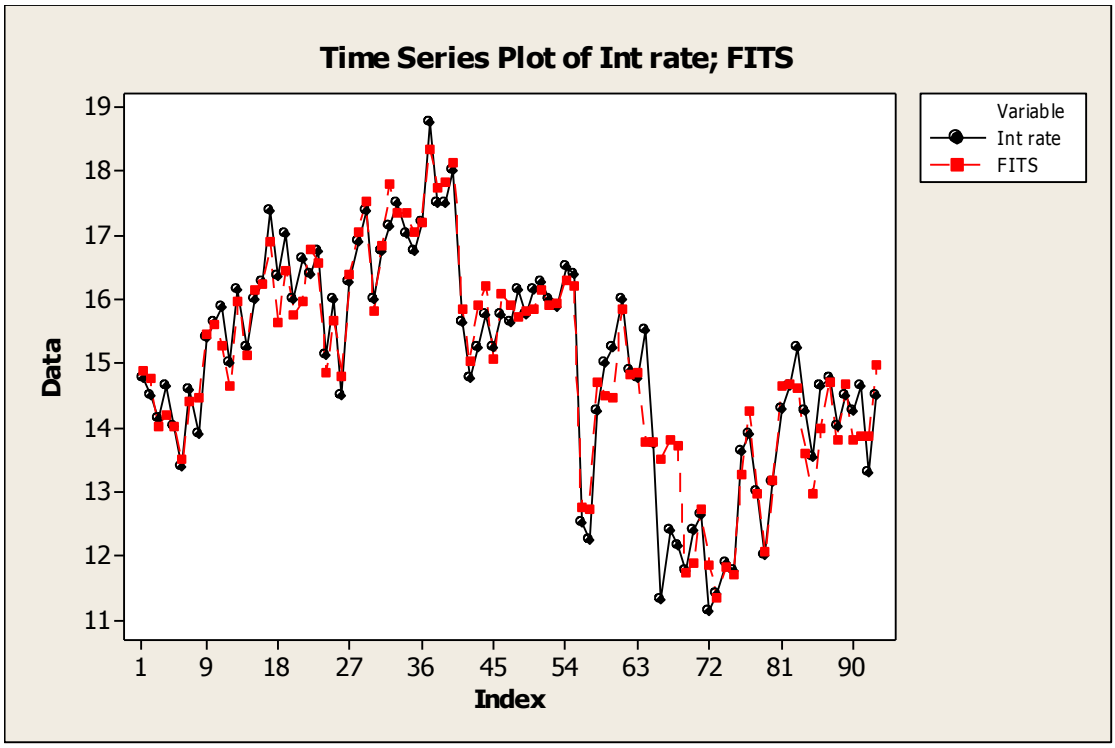
Unusual Observations

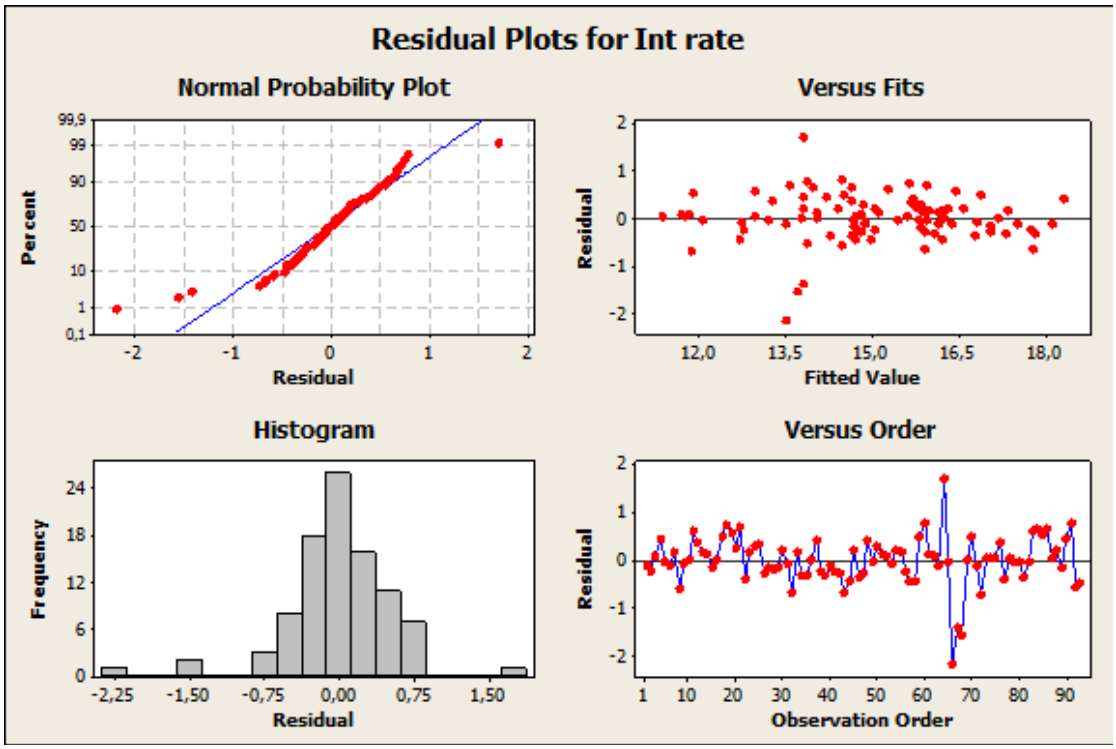
Obs	A	Int rate	Fit	SE Fit	Residual	St Resid
64	1,00	15,5000	13,6956	0,0894	1,8044	3,45R
66	1,00	11,3000	13,6462	0,0898	-2,3462	-4,48R
67	1,00	12,3800	13,6462	0,0898	-1,2662	-2,42R
68	1,00	12,1500	13,6462	0,0898	-1,4962	-2,86R

R denotes an observation with a large standardized

residual.

Πίνακας 7-21 Τελικό μοντέλο παλινδρόμησης για την αγορά ομολόγων





ΚΕΦΑΛΑΙΟ 8 (ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕΤΑΞΥ ΧΡΟΝΟΣΕΙΡΩΝ)

Πολλές επιχειρησιακές και οικονομικές εφαρμογές πρόβλεψης περιλαμβάνουν δεδομένα, που προέρχονται από χρονοσειρές. Διάφορα μοντέλα παλινδρόμησης μπορούν να προσαρμοστούν σε μηνιαία, τριμηνιαία, ή ετήσια δεδομένα, χρησιμοποιώντας τεχνικές, που έχουν περιγραφεί σε προηγούμενα κεφάλαια. Ωστόσο, επειδή τα δεδομένα, που συλλέγονται στο χρόνο, τείνουν να παρουσιάζουν τάσεις, εποχικά μοτίβα και ούτω καθ' εξής, οι παρατηρήσεις σε διαφορετικές χρονικές περιόδους σχετίζονται ή αυτοσυσχετίζονται. Έτσι για δεδομένα χρονοσειρών, το δείγμα παρατηρήσεων δεν μπορεί να θεωρηθεί σαν τυχαίο δείγμα και έτσι μπορεί να προκύψουν προβλήματα ερμηνείας, όταν εφαρμόζονται τυπικές μέθοδοι παλινδρόμησης σε παρατηρήσεις, οι οποίες σχετίζονται μεταξύ τους στο χρόνο. Η προσαρμογή μοντέλων παλινδρόμησης σε δεδομένα χρονοσειρών πρέπει να γίνεται πολύ προσεκτικά.

8.1 ΔΕΔΟΜΕΝΑ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΑΥΤΟΣΥΣΧΕΤΙΣΗ

Τα μοντέλα παλινδρόμησης, που συζητήθηκαν στα Κεφάλαια 6 και 7, υποθέτουν, ότι τα σφάλματα ε, είναι ανεξάρτητες (ή μη συσχετισμένες) και τυχαίες μεταβλητές. Αυτό σημαίνει, ότι οι διαφορετικές τιμές της ερμηνευμένης μεταβλητής Υ, μπορούν να συσχετιστούν με τις τιμές των μεταβλητών πρόβλεψης Χ, αλλά όχι μεταξύ τους. Οι συνήθεις ερμηνείες των αποτελεσμάτων μίας ανάλυσης παλινδρόμησης βασίζονται σημαντικά στην υπόθεση ανεξαρτησίας.

Με δεδομένα χρονοσειρών, η υπόθεση ανεξαρτησίας ισχύει σπάνια. Θεωρείστε την ετήσια βασική τιμή για ένα συγκεκριμένο μοντέλο ενός καινούριου αυτοκινήτου. Μπορείτε να φανταστείτε το

χάος, που θα επικρατούσε, αν οι τιμές του νέου αυτοκινήτου από τη μια χρονιά στην επόμενη δεν σχετίζονταν μεταξύ τους (ανεξάρτητες); Η γνώση της τιμής της μίας χρονιάς δεν θα σήμαινε τίποτα απολύτως για την τιμή της επόμενης χρονιάς. Στον πραγματικό κόσμο, η τιμή τον ένα χρόνο σχετίζεται με την τιμή του επόμενου χρόνου και ίσως με την τιμή δύο χρόνια πριν και ούτω καθεξής. Αυτό σημαίνει, ότι οι τιμές σε διαφορετικούς χρόνους αυτοσυσχετίζονται, δεν είναι ανεξάρτητες.

Η αυτοσυσχέτιση μπορεί να συμβεί, επειδή η επίδραση της αλλαγής της μεταβλητής πρόβλεψης στην ερμηνευμένη μεταβλητή κατανέμεται στο χρόνο. Για παράδειγμα, μία αύξηση στο μισθό μπορεί να επηρεάσει την κατανάλωση (ή τις οικονομίες), όχι μόνο σε αυτή την περίοδο, αλλά επίσης και σε μερικές μελλοντικές περιόδους. Μία αλλαγή στην τιμή ενός προϊόντος μπορεί να επηρεάσει τις πωλήσεις, όχι μόνο στην τωρινή περίοδο, αλλά και στις μελλοντικές περιόδους. Στο χρόνο οι σχέσεις τείνουν να είναι δυναμικές (να εξελίσσονται), όχι να είναι σταθερές.

Από την πλευρά της πρόβλεψης, η αυτοσυσχέτιση δεν είναι κακή. Αν οι τιμές μίας απόκρισης Y, σε μία χρονική περίοδο σχετίζονται με τις Y τιμές προηγούμενων χρονικών περιόδων, τότε τα προηγούμενα Y μπορούν να χρησιμοποιηθούν για να προβλεφθούν μελλοντικά Y. Αυτό το παρουσιάσαμε στα κεφάλαια 4 και 5. Θα το δούμε αυτό και στο κεφάλαιο 9.

Στα πλαίσια της παλινδρόμησης, η αυτοσυσχέτιση αντιμετωπίζεται κατά την κατασκευή του τυπικού μοντέλου παλινδρόμησης. Για να προσαρμοστεί η αυτοσυσχέτιση, μερικές φορές είναι απαραίτητο να αλλάξουμε το συνδυασμό των μεταβλητών πρόβλεψης ή / και τη μορφή της συνάρτησης παλινδρόμησης. Ωστόσο, τις περισσότερες φορές, η

αυτοσυσχέτιση αντιμετωπίζεται αλλάζοντας τη φύση του όρου σφάλματος.

8.2 ΣΕΙΡΙΑΚΗ ΑΥΤΟΣΥΣΧΕΤΙΣΗ

Ένα κοινό είδος αυτοσυσχέτισης, που μερικές φορές καλείται πρώτης τάξης σειριακή συσχέτιση (first-order serial correlation), είναι αυτό, στο οποίο ο όρος σφάλματος την τρέχουσα χρονική περίοδο σχετίζεται απ' ευθείας με τον όρο σφάλματος της προηγούμενης χρονικής περιόδου. Σε αυτή την περίπτωση, με τον κάτω δείκτη t να αναπαριστά το χρόνο, το μοντέλο της απλής γραμμικής παλινδρόμησης παίρνει τη μορφή:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$$

Η εξίσωση αυτή μας λέει, ότι το σφάλμα στο χρόνο t εξαρτάται γραμμικά από το σφάλμα στο χρόνο t-1 με κλίση τον συντελεστή αυτοσυσχέτισης των σφαλμάτων στο lag 1. (Η κλίση είναι ανάλογη του συντελεστή συσχέτισης. Επειδή εδώ έχουμε να κάνουμε με την ίδια μεταβλητή ο συντελεστής συσχέτισης είναι ο συντελεστής αυτοσυσχέτισης και η σταθερά αναλογίας είναι πολύ κοντά στο 1). ν_t είναι το ανεξάρτητο σφάλμα κανονικά κατανεμημένο με μέση τιμή 0 και διασπορά σ^2. Το μέγεθος του συντελεστή αυτοσυσχέτισης ρ, με $-1 \leq \rho \leq 1$, υποδεικνύει το μέτρο της σειριακής συσχέτισης. Αν ρ είναι μηδέν, τότε δεν υπάρχει σειριακή συσχέτιση, τα σφάλματα είναι ανεξάρτητα και $\varepsilon_t = \nu_t$.

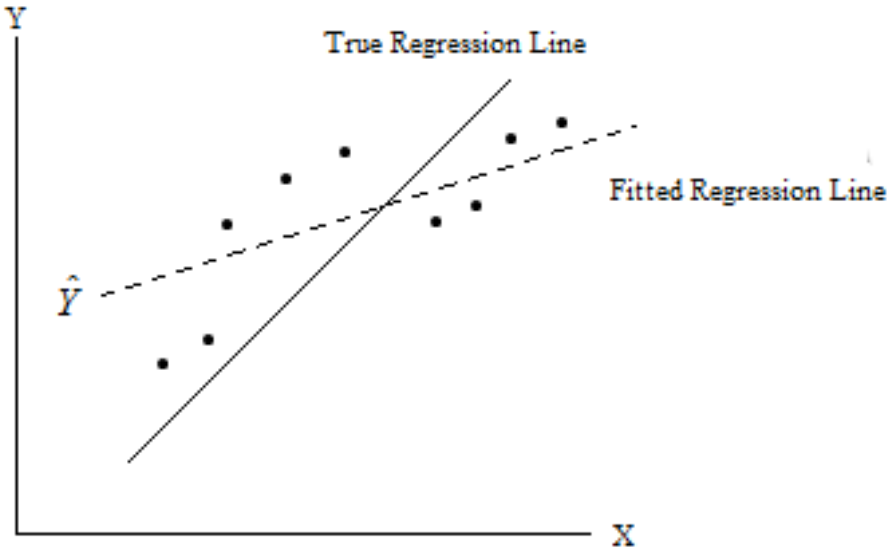
Αντικαθιστώντας το ε_{t-1} με την ίδια σχέση θα έχουμε, όπως το κάναμε στη σειριακή εκθετική εξομάλυνση:

$$\varepsilon_t = \rho(\rho\,\varepsilon_{t-2} + \nu_{t-1}) + \nu_t = \rho^2\varepsilon_{t-2} + \rho\,\nu_{t-1} + \nu_t = \ldots = \rho^t\varepsilon_0 + \sum_{s=0}^{t-1}\rho^{t-(s+1)}\nu_{s+1}$$

Που σημαίνει, ότι όσο πιο πίσω πάμε στο χρόνο, τόσο λιγότερο συνεισφέρουν στη σχέση τα προηγούμενα σφάλματα, διότι το ρ είναι κατά απόλυτη τιμή μικρότερο του 1. Μία δεύτερης τάξης σειριακή συσχέτιση των διαταρακτικών όρων (σφαλμάτων) θα ήταν της μορφής:

$$\varepsilon_t = \rho_1\,\varepsilon_{t-1} + \rho_2\,\varepsilon_{t-2} + \nu_t$$

Το παρακάτω γράφημα απεικονίζει τη συσχέτιση σε ένα μοντέλο απλής γραμμικής παλινδρόμησης. Αν η πρώτη Y τιμή είναι πάνω από την πραγματική γραμμή παλινδρόμησης, τότε μερικές από τις επόμενες Y τιμές θα είναι πάνω από τη γραμμή του πληθυσμού εξαιτίας της θετικής αυτοσυσχέτισης (αν το πρώτο σφάλμα είναι θετικό, το δεύτερο σφάλμα είναι πιθανό να είναι θετικό και ούτω καθεξής). Τελικά, μπορεί να υπάρχει μία σειρά από Y κάτω από την πραγματική γραμμή παλινδρόμησης (ένα αρνητικό σφάλμα είναι πιθανό να ακολουθείται από ένα αρνητικό σφάλμα).



Τα δεδομένα «γέρνουν» ως προς την πραγματική Χ, Υ σχέση. Ωστόσο, η γραμμή ελαχίστων τετραγώνων θα περάσει μέσα από τις παρατηρήσεις, όπως υποδεικνύεται από την διακεκομμένη γραμμή στο γράφημα. Το να χρησιμοποιήσουμε τη διακεκομμένη γραμμή για να βγάλουμε συμπεράσματα για τη συνεχή γραμμή, ή το να χρησιμοποιήσουμε τη διακεκομμένη γραμμή για να δημιουργήσουμε προβλέψεις για τα μελλοντικά Υ μπορεί να είναι πολύ παραπλανητικό.

Πρέπει επίσης να είναι σαφές ότι, η διασπορά των Υ γύρω από τη γραμμή ελαχίστων τετραγώνων είναι μικρότερη, από ότι είναι γύρω από την πραγματική γραμμή παλινδρόμησης. Ως εκ τούτου, το τυπικό σφάλμα εκτίμησης, $s_{y \cdot x}$, θα υποτιμήσει τη μεταβλητότητα των Υ γύρω από την πραγματική γραμμή παλινδρόμησης, ή ισοδύναμα, θα υποτιμήσει τη τυπική απόκλιση σ, του όρου σφάλματος, ε.

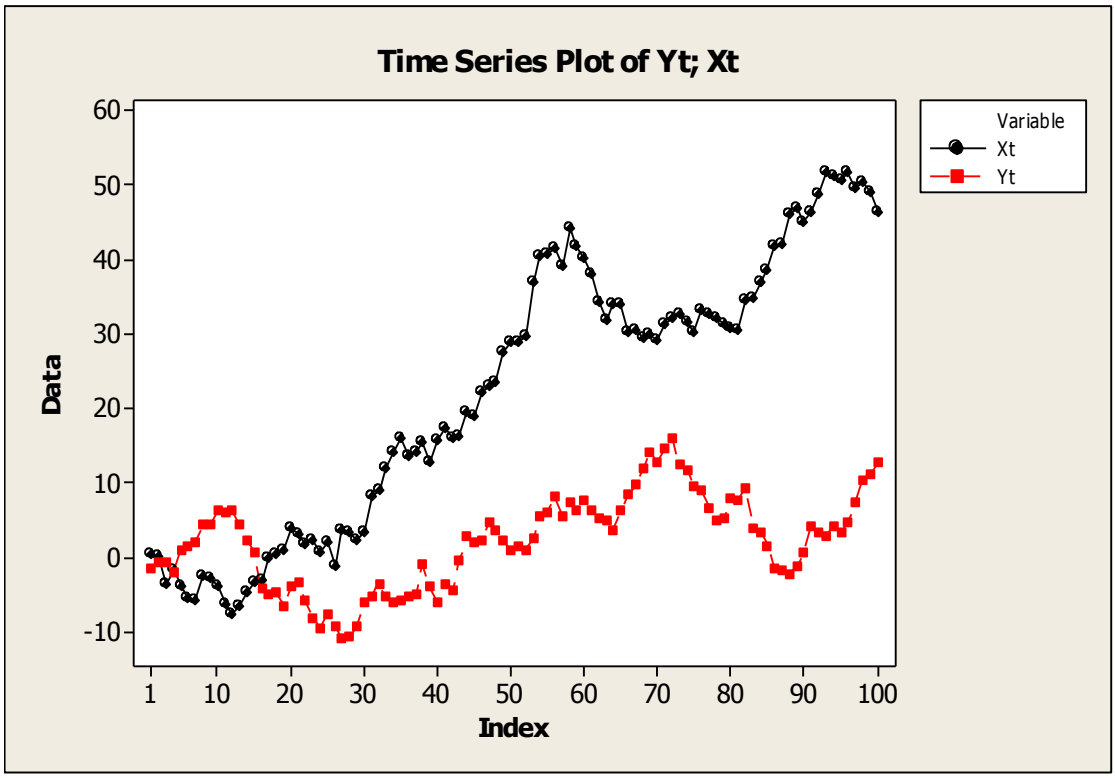
8.3 ΦΑΙΝΟΜΕΝΗ ΑΥΤΟΣΥΣΧΕΤΙΣΗ – ΛΑΝΘΑΝΟΥΣΑ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η ισχυρή αυτοσυσχέτιση μπορεί να κάνει δύο μεταβλητές, που δεν σχετίζονται μεταξύ τους, να μοιάζουν σαν να σχετίζονται. Τυπικές διαδικασίες παλινδρόμησης, που εφαρμόζονται σε παρατηρήσεις εκείνων των μεταβλητών, μπορεί να παράγουν μία σημαντική παλινδρόμηση. Σε αυτή την περίπτωση, η εκτιμώμενη σχέση είναι λανθάνουσα και μία εξέταση των καταλοίπων φυσιολογικά θα αποκαλύψει το πρόβλημα. Ωστόσο, με μία άκριτη εφαρμογή των τυπικών διαδικασιών, η λανθάνουσα παλινδρόμηση δεν θα ανιχνευθεί, οδηγώντας σε σοβαρές παρερμηνείες των αποτελεσμάτων.

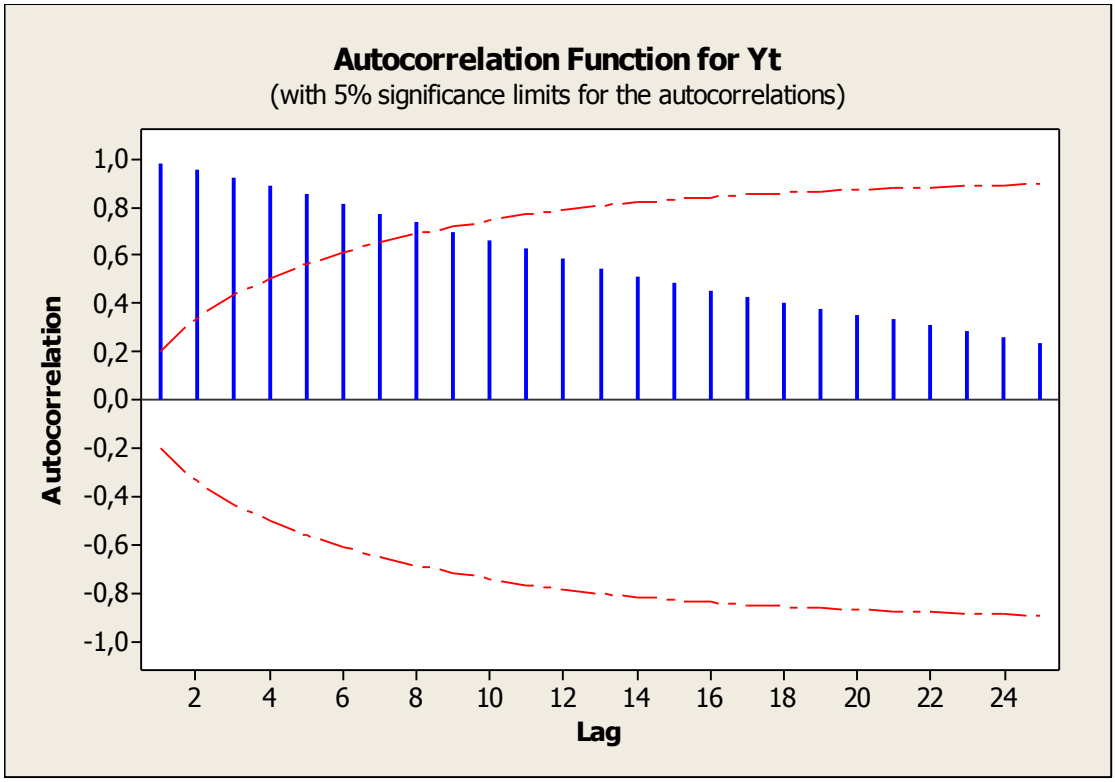
Παράδειγμα

Το Γράφημα περιέχει τα διαγράμματα δύο παραγόμενων από τον

υπολογιστή χρονοσειρών, Y_t και X_t. Αυτές οι δύο σειρές δημιουργήθηκαν με τέτοιο τρόπο, έτσι ώστε οι πρώτες σειρές (Y_t) να μην σχετίζονται με τις δεύτερες σειρές (X_t). Εδώ χρησιμοποιήθηκε κανονική κατανομή με μέσο 0 και διασπορά 1 για την εύρεση 100 τυχαίων τιμών. Οι τιμές της πρώτης χρονοσειράς δημιουργήθηκαν από μερικά αθροίσματα ως εξής: Ο πρώτος αριθμός ισούται με τον πρώτο τυχαίο αριθμό. Ο δεύτερος με το άθροισμα των δύο πρώτων, ο τρίτος με το άθροισμα των τριών πρώτων κ.ο.κ. Παρόμοια δημιουργήθηκε και η δεύτερη χρονοσειρά.



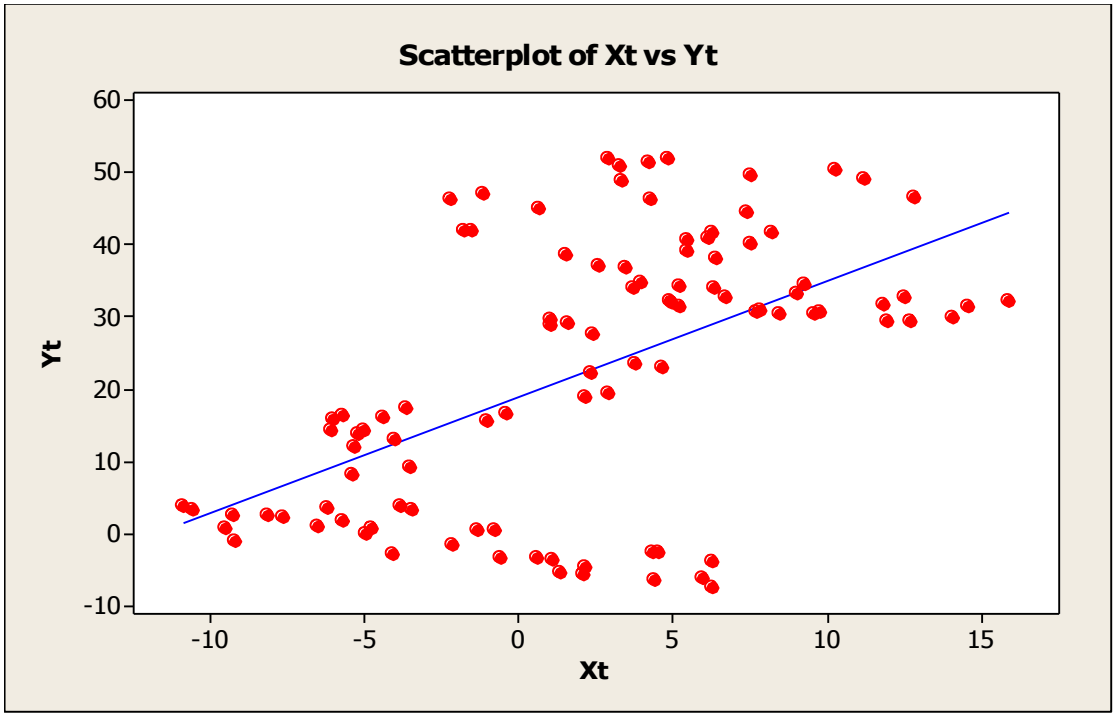
Την ίδια στιγμή, κάθε ακολουθία παρατηρήσεων είναι υψηλά αυτοσυσχετιζόμενη. Οι αυτοσυσχετίσεις για τις πρώτες σειρές φαίνονται στο αντίστοιχο γράφημα. Οι αυτοσυσχετίσεις για τις δεύτερες σειρές είναι παρόμοιες.



Lag	ACF	T	LBQ
1	0,976964	9,77	98,34
2	0,951146	5,58	192,50
3	0,920005	4,24	281,50
4	0,888415	3,51	365,36
5	0,851142	3,01	443,15
6	0,812595	2,64	514,80
7	0,771622	2,35	580,10
8	0,731447	2,12	639,42
9	0,695106	1,93	693,57
10	0,659542	1,76	742,87
11	0,623189	1,62	787,38
12	0,584496	1,48	826,98
13	0,547248	1,35	862,09
14	0,513645	1,25	893,38
15	0,481785	1,15	921,24
16	0,451777	1,07	946,02
17	0,425214	0,99	968,24
18	0,400615	0,93	988,21
19	0,374605	0,86	1005,88
20	0,353138	0,80	1021,78
21	0,330743	0,75	1035,90

22	0,306526	0,69	1048,19
23	0,283520	0,64	1058,84
24	0,257326	0,57	1067,72

Το παραπάνω γράφημα υποδεικνύει, ότι οι δύο χρονοσειρές εμφανίζονται να κινούνται μαζί. Στην πραγματικότητα, είναι πιθανό να συσχετίσουμε τις πάνω σειρές με τις κάτω σειρές, χρησιμοποιώντας ένα μοντέλο απλής γραμμικής παλινδρόμησης. Ένα διάγραμμα διασποράς των δεδομένων φαίνεται στο παρακάτω Γράφημα μαζί με τη γραμμή ελαχίστων τετραγώνων.



Η εκτιμώμενη παλινδρόμηση είναι σημαντική, με το X_t να ερμηνεύει περίπου το 30% της μεταβλητότητας στο Y_t. Ωστόσο οι Y_t σειρές δημιουργήθηκαν ανεξάρτητα από τις X_t σειρές. Αυτό σημαίνει ότι τα X δεν χρησιμοποιήθηκαν για να παραχθούν τα Y. Η εκτιμώμενη παλινδρόμηση σε αυτή την περίπτωση είναι λανθάνουσα.

Μία εξέταση των καταλοίπων θα αποκαλύψει προβλήματα με αυτή την ανάλυση παλινδρόμησης. Για παράδειγμα, οι αυτοσυσχετίσεις των

καταλοίπων είναι μεγάλες για μερικά lags, το οποίο υποδεικνύει ότι η υπόθεση των ανεξάρτητων σφαλμάτων είναι λανθασμένη και ότι το αρχικό μοντέλο παλινδρόμησης θα έπρεπε να τροποποιηθεί. Σε αυτή την περίπτωση, αν η τροποποίηση γίνει σωστά, η σχέση μεταξύ του Y_t και του X_t θα εξαφανιστεί.

Αν χρησιμοποιούνται μοντέλα παλινδρόμησης με αυτοσυσχετιζόμενα δεδομένα (χρονοσειρές), είναι πολύ σημαντική η εξέταση των καταλοίπων. Αν αυτή δεν γίνει, είναι πιθανό να φτάσουμε σε συμπεράσματα, που δεν μπορούν να αιτιολογηθούν. Το πρόβλημα δεν βρίσκεται στη διαδικασία ή μέθοδο ελαχίστων τετραγώνων. Το πρόβλημα βρίσκεται στην εφαρμογή του τυπικού μοντέλου παλινδρόμησης, όταν τα σφάλματα δεν ικανοποιούν τις βασικές υποθέσεις της παλινδρόμησης. Τα τεχνικά προβλήματα, που προκύπτουν, περιλαμβάνουν τα ακόλουθα:

1. Το τυπικό σφάλμα εκτίμησης μπορεί να υποτιμήσει σοβαρά τη μεταβλητότητα των σφαλμάτων.
2. Τα τυπικά σφάλματα των συντελεστών παλινδρόμησης υποτιμούν τη μεταβλητότητα των εκτιμώμενων συντελεστών παλινδρόμησης.
3. Τα συνήθη συμπεράσματα, που βασίζονται στα στατιστικά t και F, δεν είναι πλέον αυστηρά εφαρμόσιμα.
4. Έτσι μπορεί να προκύψουν λανθάνουσες παλινδρομήσεις.

8.4 ΑΥΤΟΣΥΣΧΕΤΙΣΗ ΚΑΙ ΕΛΕΓΧΟΣ DURBIN WATSON

Όπως έχει συζητηθεί ξανά, η αυτοσυσχέτιση μπορεί να ερευνηθεί κατασκευάζοντας τους συντελεστές αυτοσυσχέτισης και συγκρίνοντάς τους με τα τυπικά τους σφάλματα. Επιπλέον, ένας ολικός έλεγχος για σημαντικότητα της αυτοσυσχέτισης παρέχεται από τη στατιστική Ljung-

Box Q. Τώρα θα εφαρμόσουμε αυτές τις διαδικασίες στα κατάλοιπα από μία ανάλυση παλινδρόμησης με δεδομένα χρονοσειρών για να ελέγξουμε για την τυχαιότητα τους. Η σχέση:

$$r_k(e)=\frac{\sum_{t=k+1}^{n} e_t e_{t-k}}{\sum_{t=1}^{n} e_t^2}, \qquad k=1,2,\ldots,K$$

δίνει τους συντελεστές αυτοσυσχέτισης καταλοίπων.

Τα προγράμματα, όπως το Minitab, παρέχουν την επιλογή του απευθείας υπολογισμού μιας στατιστικής, που είναι χρήσιμη για ανίχνευση πρώτης τάξης σειριακής συσχέτισης, ή lag 1 αυτοσυσχέτισης, γνωστής ως στατιστική των Durbin-Watson (DW). Ένας έλεγχος για σημαντική πρώτης-τάξης σειριακή συσχέτιση, που βασίζεται σε αυτή τη στατιστική, είναι γνωστός ως έλεγχος των Durbin-Watson. Ο έλεγχος εξετάζει, το αν ο συντελεστής αυτοσυσχέτισης ρ στην εξίσωση του σφάλματος, είναι μηδέν ή όχι.

Θεωρούμε ότι:

$$\varepsilon_t=\rho\varepsilon_{t-1}+v_t \quad \text{και} \quad e_t=r_1 e_{t-1}$$

Οι υποθέσεις, που θα ελεγχθούν, είναι:

$$H_0:\rho = 0,$$

$$H_1:\rho > 0.$$

Η εναλλακτική υπόθεση είναι ρ>0, δεδομένου ότι οι επιχειρησιακές και οικονομικές χρονοσειρές συνήθως τείνουν να δείχνουν θετική αυτοσυσχέτιση.

Αν ένα μοντέλο παλινδρόμησης δεν λαμβάνει υπόψη του πιθανή αυτοσυσχέτιση, τα κατάλοιπα πιθανώς θα αυτοσυσχετίζονται. Έτσι ελέγχουμε μέσω του Durbin-Watson έλεγχου τα κατάλοιπα από την

ανάλυση παλινδρόμησης για σειριακή συσχέτιση πρώτης τάξης.

Η στατιστική Durbin-Watson ορίζεται ως:

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

όπου:

$$e_t = Y_t - \hat{Y}_t \quad και \quad e_{t-1} = Y_{t-1} - \hat{Y}_{t-1}$$

είναι τα κατάλοιπα για την περίοδο t και t-1.

Για θετική σειριακή συσχέτιση το άθροισμα των τετραγωνικών διαφορών στον αριθμητή της Durbin-Watson στατιστικής θα είναι σχετικά μικρό. Μικρές τιμές για τη Durbin-Watson στατιστική είναι συνεπείς με θετική σειριακή συσχέτιση.

Ο συντελεστής αυτοσυσχέτισης ρ, μπορεί να υπολογιστεί από τον συντελεστή αυτοσυσχέτισης των καταλοίπων $r_1(e)$ στο lag 1 και μετά από μερικές μαθηματικές πράξεις, η Durbin-Watson στατιστική μπορεί να συσχετιστεί με το $r_1(e)$. Για μέτρια έως μεγάλα δείγματα, **$DW=2(1-r_1(e))$**. Επειδή το $r_1(e)$ είναι μεταξύ -1 και 1 το DW θα βρίσκεται μεταξύ του 0 και του 4. Θετική lag 1 αυτοσυσχέτιση καταλοίπων σχετίζεται με DW τιμές κάτω από 2, και η αρνητική lag 1 αυτοσυσχέτιση καταλοίπων σχετίζεται με DW τιμές πάνω από 2.

Ένας χρήσιμος αλλά μερικές φορές όχι καθοριστικός έλεγχος για σειριακή συσχέτιση μπορεί να πραγματοποιηθεί συγκρίνοντας τις υπολογισμένες τιμές της Durbin-Watson στατιστικής με ένα κατώτερο (d_L) και ένα ανώτερο (d_U) όριο. Οι κανόνες απόφασης είναι οι ακόλουθοι:

1. Όταν η Durbin-Watson στατιστική είναι μεγαλύτερη από το ανώτερο όριο, ο συντελεστής αυτοσυσχέτισης ρ, είναι ίσος με

μηδέν (στατιστικά δεν υπάρχει θετική αυτοσυσχέτιση).

2. Όταν η Durbin-Watson στατιστική είναι μικρότερη από το κατώτερο όριο, ο συντελεστής αυτοσυσχέτισης ρ, είναι μεγαλύτερος από το μηδέν (υπάρχει θετική αυτοσυσχέτιση).
3. Όταν η Durbin-Watson στατιστική βρίσκεται ανάμεσα στο ανώτερο και κατώτερο όριο, ο έλεγχος είναι αναποτελεσματικός (δεν γνωρίζουμε αν υπάρχει θετική αυτοσυσχέτιση).

Ο Durbin-Watson έλεγχος μπορεί να χρησιμοποιηθεί για να καθορίσει αν υπάρχει θετική lag 1 αυτοσυσχέτιση.

Αν $DW>d_U$, συμπεραίνουμε ότι $H_0:\rho=0$

Αν $DW<d_L$, συμπεραίνουμε ότι $H_1:\rho>0$

Αν το DW βρίσκεται μέσα στο ανώτερο και κατώτερο όριο ($d_L \leq DW \leq d_U$), ο έλεγχος είναι αναποτελεσματικός.

Για αρνητική αυτοσυσχέτιση ισχύει:

Αν $DW<4-d_U$, συμπεραίνουμε ότι $H_0:\rho=0$

Αν $DW>4-d_L$, συμπεραίνουμε ότι $H_1:\rho<0$

Αν $4-d_U<DW<4-d_L$ ο έλεγχος είναι αναποτελεσματικός. Τα κρίσιμα όρια για το d_U και το d_L δίνονται στον Πίνακα Β (στο Παράρτημα στο τέλος του βιβλίου). Για να βρεθεί το κατάλληλο, d_L και d_U, ο αναλυτής πρέπει να γνωρίζει το μέγεθος του δείγματος, το επίπεδο σημαντικότητας και τον αριθμό των ανεξάρτητων μεταβλητών. Στον Πίνακα Β, το μέγεθος του δείγματος δίνεται στην αριστερή στήλη και ο αριθμός των ανεξάρτητων μεταβλητών καθορίζεται από την κορυφή κάθε στήλης. Αν χρησιμοποιηθούν 3 ανεξάρτητες μεταβλητές, τότε πρέπει να δούμε τη στήλη για k=3.

Από τα όρια του DW μπορούμε να συμπεράνουμε το πρόσημο και

το μέγεθος του lag 1 συντελεστή αυτοσυσχέτισης καταλοίπων. Έτσι, για καταστάσεις, κατά τις οποίες ο Durbin-Watson έλεγχος είναι αναποτελεσματικός, η σημαντικότητα της σειριακής συσχέτισης μπορεί να ερευνηθεί συγκρίνοντας το $r_1(e)$ με $\pm 2/\sqrt{n}$. Αν το $r_1(e)$ βρίσκεται στο διάστημα $0 \pm 2/\sqrt{n}$, συμπεραίνουμε, ότι η αυτοσυσχέτιση είναι μικρή και μπορεί να αγνοηθεί. Επίσης μπορούμε να χρησιμοποιήσουμε την στατιστική Q, ή την παλινδρόμηση ανάμεσα σε e_t *και* e_{t-1} χωρίς τον σταθερό όρο και την χρήση της στατιστικής t.

Παράδειγμα (Τ.8-1)

Θεωρήστε έναν αναλυτή, που θέλει να δημιουργήσει μία ποσοτική βάση για πρόβλεψη μελλοντικών πωλήσεων μιας εταιρείας. Δεδομένου ότι η εταιρία πουλάει τοπικά, ένας υπολογισμός του διαθέσιμου προσωπικού εισοδήματος για την περιοχή θα έπρεπε να σχετίζεται στενά με τις πωλήσεις. Ο παρακάτω Πίνακας δείχνει τις πωλήσεις και το εισόδημα για την περίοδο από το 1986 μέχρι το 2006. Επίσης φαίνονται στον πίνακα οι στήλες, που είναι απαραίτητες για τον υπολογισμό της DW στατιστικής. Τα κατάλοιπα έρχονται από μία προσαρμοσμένη γραμμή ελαχίστων τετραγώνων στα δεδομένα, όπως φαίνεται στο Γράφημα.

Πριν χρησιμοποιήσει την ευθεία ελαχίστων τετραγώνων για πρόβλεψη, ο αναλυτής πραγματοποιεί τον Durbin-Watson έλεγχο για θετική σειριακή συσχέτιση. Οι υπολογισμοί για τις τρεις τελευταίες στήλες για το 1987 είναι ως εξής:

$$e_t - e_{t-1} = -47{,}53 - (-76{,}53) = 28{,}83$$

$$(e_t - e_{t-1})^2 = 28{,}83^2 = 831{,}17$$

$$e_t^2 = (-47{,}53)^2 = 2.259{,}1$$

Η DW στατιστική υπολογίζεται από:

$$DW = \frac{\sum_{t=2}^{21} (e_t - e_{t-1})^2}{\sum_{t=1}^{21} e_t^2} = \frac{1.926.035,14}{2.210.641,78} = 0,87$$

Χρησιμοποιώντας ένα 0,01 επίπεδο σημαντικότητας για ένα δείγμα με 21 στοιχεία και k=1 (μία ανεξάρτητη μεταβλητή), έχουμε:

d_L=0,97

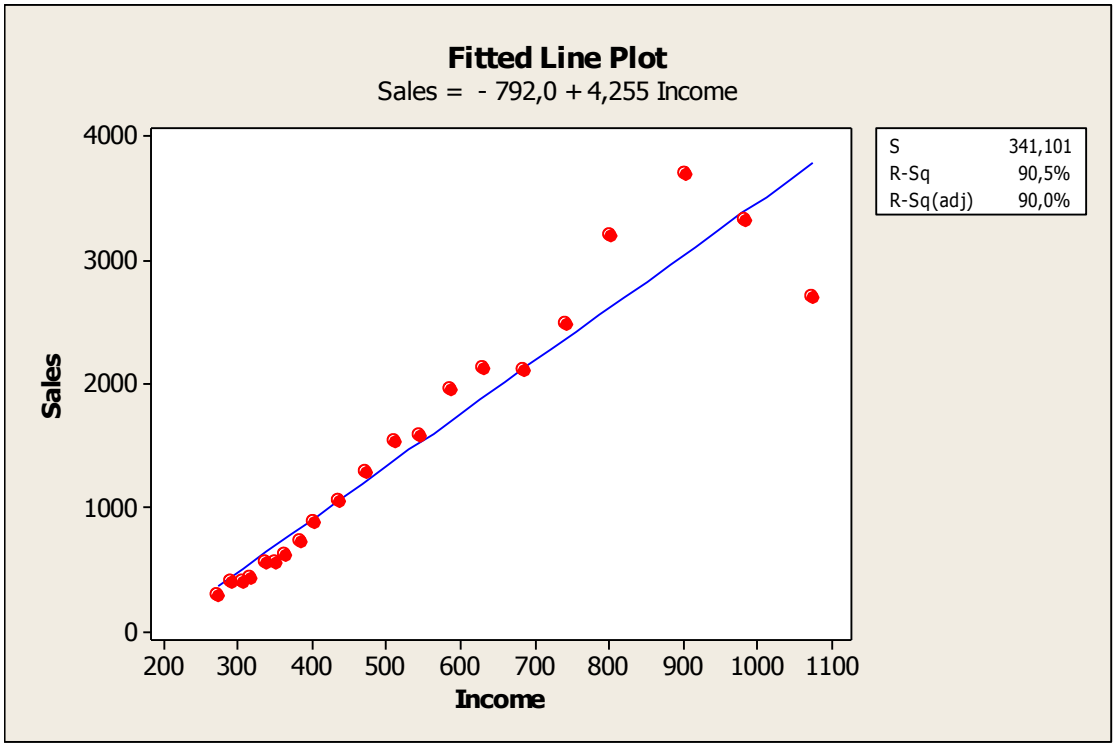
d_U=1,16

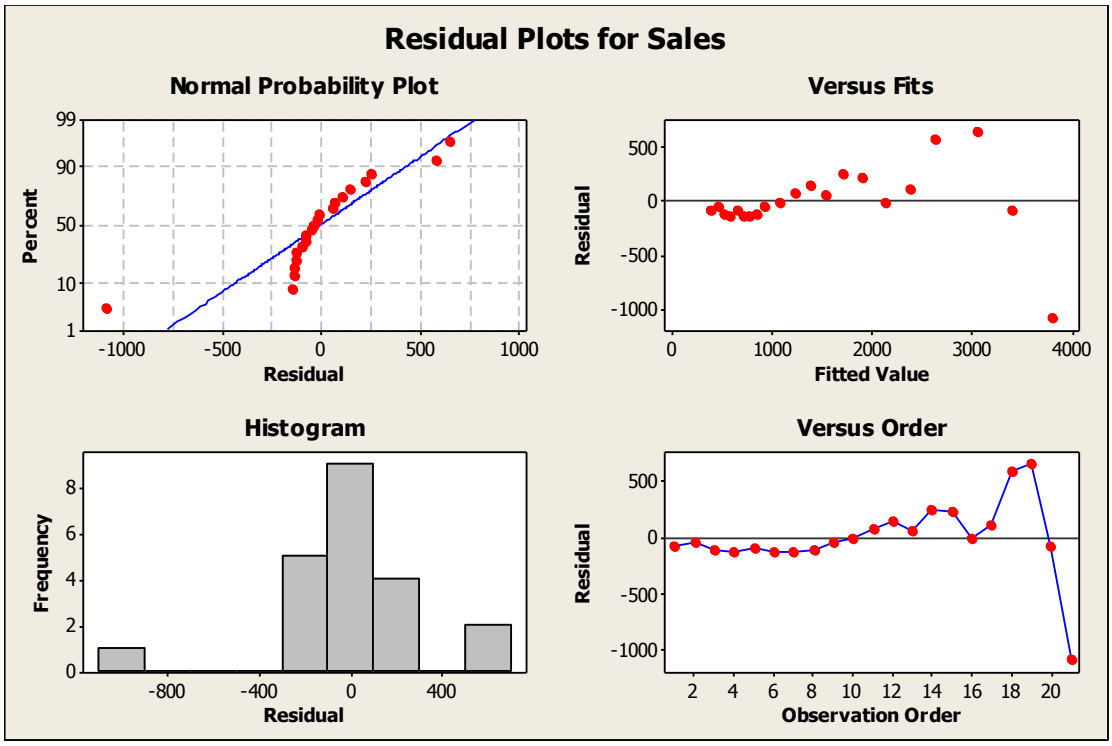
Επειδή το DW=0,87 βρίσκεται κάτω από το d_L=0,97, η μηδενική υπόθεση H_0:ρ=0 απορρίπτεται και συμπεραίνουμε, ότι τα σφάλματα συσχετίζονται θετικά (ρ>0). Το μοντέλο παλινδρόμησης πρέπει να τροποποιηθεί, πριν χρησιμοποιηθεί για πρόβλεψη.

Year	**Sales Y_t**	**Income X_t**	**Residuals e_t**	**e_t-e_{t-1}**	**$(e_t-e_{t-1})^2$**	**e_t^2**
1986	295	273,4	-76,36	-	-	5.830,85
1987	400	291,3	-47,53	28,83	831,17	2.259,10
1988	390	306,9	-123,91	-76,38	5.833,90	15.353,69
1989	425	317,1	-132,32	-8,41	70,73	17.508,58
1990	547	336,1	-91,16	41,16	1.694,15	8.310,15
1991	555	349,4	-139,76	-48,60	2.361,96	19.532,86
1992	620	362,9	-132,20	7,56	57,15	17.476,84
1993	720	383,9	-121,56	10,64	113,21	14.776,83
1994	880	402,8	-41,98	79,58	6.332,98	1.762,32
1995	1.050	437,0	-17,51	24,47	598,78	306,60
1996	1.290	472,2	72,71	90,22	8.139,65	5.286,74
1997	1.528	510,4	148,16	75,45	5.692,70	21.951,39
1998	1.586	544,5	61,06	-87,10	7.586,41	3.728,32
1999	1.960	588,1	249,53	188,47	35.520,94	62.265,22
2000	2.118	630,4	227,54	-21,99	483,56	51.774,45
2001	2.116	685,9	-10,62	-238,16	56.720,19	112,78
2002	2.477	742,8	108,26	118,88	14.132,45	11.720,23
2003	3.199	801,3	581,33	473,07	223.795,22	337.944,57
2004	3.702	903,1	651,16	69,83	4.876,23	424.009,35

2005	3.316	983,6	-77,38	-728,54	530.770,23	5.987,66
2006	2.702	1.076,7	-1.087,54	-1.010,16	1.020.423,23	1.182.743,25
		Totals			1.926.035,14	2.210.641,78

Δεδομένα και Durbin-Watson υπολογισμοί





Regression Analysis: Sales versus Income

```
The regression equation is
Sales = - 792 + 4,26 Income

Predictor    Coef  SE Coef      T      P
Constant   -792,0    187,1  -4,23  0,000
Income     4,2552   0,3163  13,45  0,000

S = 341,101   R-Sq = 90,5%   R-Sq(adj) = 90,0%

Analysis of Variance
Source          DF    SS       MS       F      P
Regression       1 21061160 21061160 181,02 0,000
Residual Error  19  2210646   116350
Total           20 23271806

Durbin-Watson statistic = 0,871253
```

8.5 ΛΥΣΕΙΣ ΓΙΑ ΠΡΟΒΛΗΜΑΤΑ ΑΥΤΟΣΥΣΧΕΤΙΣΗΣ

Αφού εντοπιστεί αυτοσυσχέτιση σε μία παλινδρόμηση δεδομένων χρονοσειρών, είναι απαραίτητο να αφαιρεθεί, ή να μοντελοποιηθεί, προτού η συνάρτηση παλινδρόμησης να μπορεί να αξιολογηθεί για την αποδοτικότητά της. Η κατάλληλη μέθοδος για το χειρισμό της αυτοσυσχέτισης εξαρτάται, από το τι την προκάλεσε εξαρχής. Η αυτοσυσχέτιση μπορεί να προκύψει εξαιτίας ενός σφάλματος προσδιορισμού (specification error), όπως μία μεταβλητή, που έχει παραληφθεί, ή μπορεί να προκύψει επειδή οι όροι των ανεξάρτητων σφαλμάτων συσχετίζονται σε ένα κατά τα άλλα σωστά προσδιορισμένο μοντέλο.

Η λύση στο πρόβλημα της αυτοσυσχέτισης ξεκινάει με μία αξιολόγηση του προσδιορισμού του μοντέλου. Είναι η λειτουργική μορφή σωστή; Παραλήφθηκαν σημαντικές μεταβλητές; Υπάρχουν

επιδράσεις, που μπορεί να περιέχουν μοτίβα στο χρόνο και τα οποία μπορεί να δημιουργήσουν αυτοσυσχέτιση στα σφάλματα;

Επειδή η κύρια αιτία αυτοσυσχετιζόμενων σφαλμάτων σε ένα μοντέλο παλινδρόμησης είναι η παράλειψη μίας ή περισσότερων σημαντικών μεταβλητών, η καλύτερη προσέγγιση, για να λυθεί το πρόβλημα αυτό, είναι να τις βρούμε. Αυτή η προσπάθεια συχνά αναφέρεται σαν βελτίωση του προσδιορισμού του μοντέλου. Ο προσδιορισμός του μοντέλου δεν περιλαμβάνει μόνο την εύρεση των σημαντικών μεταβλητών πρόβλεψης αλλά και την εισαγωγή αυτών των μεταβλητών στη συνάρτηση με το σωστό τρόπο. Δυστυχώς, δεν είναι πάντα δυνατή η βελτίωση του προσδιορισμού του μοντέλου, επειδή μία σημαντική μεταβλητή, που λείπει, μπορεί να μην μπορεί να ποσοτικοποιηθεί, ή ακόμα και να μπορεί, τα δεδομένα μπορεί να μην είναι διαθέσιμα, ή να μην είναι και ακριβή.

Μόνο αφού ο προσδιορισμός της εξίσωσης έχει επιθεωρηθεί προσεκτικά θα έπρεπε να θεωρηθεί η πιθανότητα προσαρμογής. Μία προσέγγιση λοιπόν για την εξάλειψη αυτοσυσχέτισης είναι να προστεθεί στη συνάρτηση παλινδρόμησης μία μεταβλητή, που έχει παραληφθεί και η οποία ερμηνεύει στην απόκριση τη σχέση από τη μία περίοδο στην επόμενη.

Μία άλλη προσέγγιση για την εξάλειψη της αυτοσυσχέτισης περιλαμβάνει τη γενική αντίληψη της μεθόδου των διαφορών. Σε αυτή την προσέγγιση, το μοντέλο παλινδρόμησης προσδιορίζεται σε όρους *αλλαγών – διαφορών των τιμών* και όχι *επιπέδων των τιμών*. Κατά περίπτωση, οι πρωτότυπες μεταβλητές μπορεί να εκφράζονται σε όρους λογαρίθμων οπότε χρησιμοποιούνται οι διαφορές των λογαρίθμων στην παλινδρόμηση. Αυτή η διαδικασία είναι ισοδύναμη με την παλινδρόμηση

της αλλαγής ποσοστού στην απόκριση, στις αλλαγές ποσοστού στις μεταβλητές πρόβλεψης. Τελικά, αντί να γίνεται χρήση απλών ή πρώτων διαφορών (simple or first differences) στο μοντέλο παλινδρόμησης, μπορεί να βρισκόμαστε στην περίπτωση, όπου απαιτούνται, γενικευμένες διαφορές για την εξάλειψη της αυτοσυσχέτισης.

Η προσέγγιση του αυτοπαλινδρομικού μοντέλου για την εξάλειψη της αυτοσυσχέτισης παράγει μεταβλητές πρόβλεψης χρησιμοποιώντας την μεταβλητή απόκρισης Y, με καθυστέρηση μίας ή περισσότερων περιόδων. Στην περίπτωση του αυτοπαλινδρομικού μοντέλου πρώτης τάξης, η μόνη μεταβλητή πρόβλεψης είναι η Y μεταβλητή με καθυστέρηση μίας χρονικής περιόδου.

Παρακάτω θα συζητηθούν μερικές τεχνικές για την εξάλειψη αυτοσυσχέτισης. Τα ακόλουθα παραδείγματα απεικονίζουν αυτές τις μεθόδους για την εξάλειψη της αυτοσυσχέτισης.

8.5.1 MODEL SPECIFICATION ERROR

Το παρακάτω Παράδειγμα δείχνει, το πώς η ενσωμάτωση μίας μεταβλητής, που λείπει, μπορεί να εξαλείψει τη σειριακή συσχέτιση.

Παράδειγμα (Τ.8-2)

Μία εταιρία (NOVAK) επιθυμεί να αναπτύξει ένα μοντέλο πρόβλεψης για προβολή των μελλοντικών πωλήσεων. Δεδομένου ότι έχει καταστήματα σε όλη την περιοχή, επιλέγεται το διαθέσιμο προσωπικό εισόδημα με βάση την περιοχή σαν δυνατή μεταβλητή πρόβλεψης. Ο παρακάτω πίνακας δείχνει τις πωλήσεις από το 1990 έως το 2006. Ο πίνακας επίσης δείχνει το διαθέσιμο προσωπικό εισόδημα και την ανεργία της περιοχής.

Row	Year	Sales ($ millions)	Income ($millions)	Unemploy-ment Rate	Y-Lagged

		(Y)			
1	1990	8,0	336,1	5,5	-
2	1991	8,2	349,4	5,5	8,0
3	1992	8,5	362,9	6,7	8,2
4	1993	9,2	383.9	5,5	8,5
5	1994	10,2	402,8	5,7	9,2
6	1995	11,4	437,0	5,2	10,2
7	1996	12,8	472,2	4,5	11,4
8	1997	13,6	510,4	3,8	12,8
9	1998	14,6	544,5	3,8	13,6
10	1999	16,4	588,1	3,6	14,6
11	2000	17,8	630,4	3,5	16,4
12	2001	18,6	685,9	4,9	17,8
13	2002	20,0	742,8	5,9	18,6
14	2003	21,9	801,3	5,6	20,0
15	2004	24,9	903,1	4,9	21,9
16	2005	27,3	983,6	5,6	24,9
17	2006	29,1	1076,7	8,5	27,3

Η Durbin-Watson στατιστική είναι 0,72 και χρησιμοποιώντας ένα επίπεδο σημαντικότητας 0,01 με n=17 και k=1, έχουμε:

$$d_L=0,87$$

$$d_U=1,10$$

Δεδομένου ότι DW=0,72<d_L=0,87, υποδεικνύει θετική σειριακή συσχέτιση. Μία σημαντική μεταβλητή, που ευθύνεται για την εναπομένουσα συσχέτιση των πωλήσεων από τον ένα χρόνο στον άλλο, μπορεί να λείπει από το μοντέλο. Αυτό το αποτέλεσμα μπορεί να είναι αληθές ακόμα και αν το αποτέλεσμα του Minitab, εκφράζει ότι το διαθέσιμο προσωπικό εισόδημα ερμηνεύει το 99,5% της μεταβλητότητας στις πωλήσεις.

Regression Analysis: Sales (Y) versus Income

```
The regression equation is
Sales (Y) = - 1,50 + 0,0292 Income

Predictor      Coef  SE Coef       T    P
Constant    -1,5046 0,3290    -4,57 0,000
```

```
Income    0,0291916 0,0005129 56,92 0,000

S = 0,476669 R-Sq = 99,5% R-Sq(adj) = 99,5%

Analysis of Variance
Source          DF    SS      MS      F        P
Regression       1   736,15 736,15 3239,89 0,000
Residual Error 15     3,41   0,23
Total          16   739,56

Durbin-Watson statistic = 0,721806
```

Το ποσοστό ανεργίας, που λείπει, μπορεί να είναι μία σημαντική μεταβλητή πρόβλεψης των πωλήσεων. Στον παρακάτω Πίνακα εμφανίζονται τα αποτελέσματα της ανάλυσης παλινδρόμησης, όταν το ποσοστό της ανεργίας προστίθεται στο μοντέλο. Το προσαρμοσμένο μοντέλο τώρα ερμηνεύει το 99,9% της μεταβλητότητας στις πωλήσεις. Αν και η σταθερά b_0 δεν είναι σημαντική, η επαναπροσαρμογή του μοντέλου χωρίς την σταθερά αφήνει το Durbin-Watson στατιστικό του 1,98 ουσιαστικά αμετάβλητο.

Με 0,01 επίπεδο σημαντικότητας, n=17, και k=2, ο Πίνακας Β δίνει:

$$d_L=0,77$$

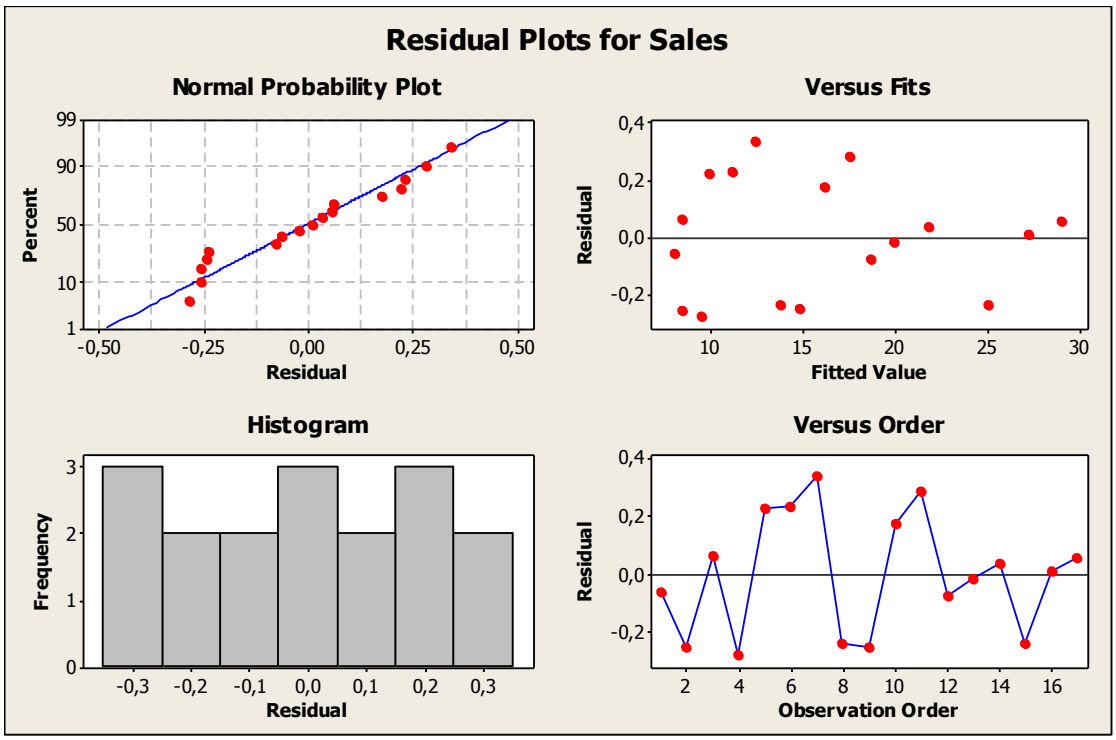
$$d_U=1,25$$

Δεδομένου ότι DW=1,98>d_U=1,25 δεν υπάρχει ένδειξη πρώτης-τάξης σειριακής συσχέτισης. Επίσης, η lag 1 αυτοσυσχέτιση καταλοίπων είναι $r_1(e)=0,005$, το οποίο είναι μέσα στο διάστημα $\pm 2/\sqrt{n}=\pm 2/\sqrt{17}=\pm 0,49$. Παρατηρούμε επίσης ότι DW=1,98≈2(1-$r_1(e)$)=2(1-0,005)=1,99.

Η συνάρτηση $\hat{Y}=-0,014+0,03X_1-0,35X_2$ μπορεί να χρησιμοποιηθεί για να προβλέψει τις πωλήσεις με τη γνώση, ότι τα

σφάλματα είναι ανεξάρτητα. Ειδικές εκτιμήσεις του διαθέσιμου προσωπικού εισοδήματος ($1.185 million) και του ποσοστού ανεργίας (7,8%) για την περιοχή χρησιμοποιήθηκαν για να προβλέψουν τις πωλήσεις για το 2007. Η πρόβλεψη είναι:

$$\hat{Y} = -0{,}014 + 0{,}03(1.185) - 0{,}35(7{,}8) = 32{,}7$$



Regression Analysis: Sales (Y) versus Income, Rate

The regression equation is
Sales (Y) = - 0,014 + 0,0297 Income - 0,350 Rate

Predictor	Coef	SE Coef	T	P
Constant	-0,0140000	0,2498	-0,06	0,956
Income	0,0297492	0,0002480	119,96	0,000
Rate	-0,3498700	0,04656	-7,51	0,000

S = 0,219930 R-Sq = 99,9% R-Sq(adj) = 99,9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	738,88	369,44	7637,91	0,000
Residual Error	14	0,68	0,05		
Total	16	739,56			

```
Durbin-Watson statistic = 1,98003
```

8.5.2 ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΔΙΑΦΟΡΕΣ

Για υψηλά αυτοσυσχετιζόμενα δεδομένα, το να μοντελοποιούμε διαφορές αντί για επίπεδα τιμών, μπορεί συχνά να εξαλείψει τη σειριακή συσχέτιση. Αυτό σημαίνει, ότι αντί να μορφοποιούμε την εξίσωση παλινδρόμησης σε όρους του Y και των X_1, X_2, ... X_K, η εξίσωση παλινδρόμησης γράφεται σε όρους των διαφορών $Y'_t = Y_t - Y_{t-1}$ και $X'_{t,1} = X_{t,1} - X_{t-1,1}$, $X'_{t,2} = X_{t,2} - X_{t-1,2}$ και ούτε καθεξής. Οι διαφορές πρέπει να θεωρηθούν και να χρησιμοποιηθούν, όταν η τιμή της Durbin-Watson στατιστικής, που σχετίζεται με την παλινδρόμηση και περιέχει τις πρωτότυπες μεταβλητές είναι κοντά στο μηδέν και όταν ο συντελεστής αυτοσυσχέτισης είναι κοντά στο 1.

Αν ο συντελεστής αυτοσυσχέτισης δεν είναι κοντά στο 1, τότε πραγματευόμαστε την κατάσταση ως ακολούθως:

Έστω ότι ισχύουν οι εξισώσεις συσχέτισης των χρονοσειρών;

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$$

όπου,

- ρ = η συσχέτιση μεταξύ παρακείμενων σφαλμάτων.
- ν_t = το τυχαίο σφάλμα.
- $\nu_t = \varepsilon_t$ όταν ρ=0.

Το μοντέλο ισχύει για κάθε χρονική περίοδο, έτσι:

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1}$$

Πολλαπλασιάζοντας την ανωτέρω εξίσωση και στις δύο πλευρές

με ρ και αφαιρώντας κατά μέλη τις δύο παρακάτω εξισώσεις έχουμε:

$$\rho Y_{t-1} = \rho \beta_0 + \rho \beta_1 X_{t-1} + \rho \varepsilon_{t-1}$$

$$Y_t - \rho Y_{t-1} = (\beta_0 - \rho \beta_0) + (\beta_1 X_t - \rho \beta_1 X_{t-1}) + (\varepsilon_t - \rho \varepsilon_{t-1}) \quad \text{ή}$$

$$Y_t' = \beta_0 (1-\rho) + \beta_1 X_t' + \nu_t \quad \text{με}$$

$$Y_t' = Y_t - \rho Y_{t-1} \quad \text{και} \quad X_t' = X_t - \rho X_{t-1}$$

Τα σφάλματα ν_t είναι ανεξάρτητα κατανεμημένα με μέσο όρο το μηδέν και σταθερή διασπορά. Οπότε μπορούμε να εφαρμόσουμε τις μεθόδους παλινδρόμησης για τις συσχετίσεις μεταξύ των διαφορών. Στην περίπτωση, που το ρ είναι κοντά στο 1, που σημαίνει μεγάλη συσχέτιση μεταξύ διαδοχικών σφαλμάτων, τότε $Y_t' = Y_t - Y_{t-1}$ και $X_t' = X_t - X_{t-1}$ και ο όρος $\beta_0(1-\rho)$ μηδενίζεται, οπότε έχουμε αναλογική σχέση.

Παράδειγμα (Τ.8-5)

Μια εταιρία (Sears) επιθυμούσε τη πρόβλεψη των πωλήσεων της σε χιλιάδες δολάρια. Επέλεξε το διαθέσιμο προσωπικό εισόδημα για την περιοχή σαν την ανεξάρτητη μεταβλητή. Η εταιρία ήθελε να συσχετίσει τις πωλήσεις με το διαθέσιμο προσωπικό εισόδημα χρησιμοποιώντας ένα λογαριθμικό γραμμικό μοντέλο παλινδρόμησης, μιας και αυτό θα επέτρεπε την εκτίμηση του ποσοστού αλλαγής του εισοδήματος στις πωλήσεις. Οι εκθετικές καμπύλες περιγράφονται από σταθερό ποσοστό αύξησης, ενώ οι γραμμικές σταθερό ποσό.

Το λογαριθμικό γραμμικό μοντέλο παλινδρόμησης υποθέτει, ότι το εισόδημα σχετίζεται με τις πωλήσεις σύμφωνα με την εξίσωση:

$$Sales = \gamma (Income)^{\beta_1}$$

Παίρνοντας τους φυσικούς λογαρίθμους και στις δύο πλευρές της εξίσωσης έχουμε:

$$\ln Sales = \ln \gamma + \beta_1 \ln(Income)$$

Προσθέτοντας έναν όρο σφάλματος για την επιρροή μεταβλητών άλλων από το εισόδημα στις πωλήσεις, η προηγούμενη έκφραση γίνεται ένα λογαριθμικό γραμμικό μοντέλο παλινδρόμησης (log linear regression model) της μορφής:

$$\ln Y_t = \beta_0 + \beta_1 \ln X_t + \varepsilon_t$$

Οι παρακάτω Πίνακες δείχνουν τις πωλήσεις, το διαθέσιμο εισόδημα, τους λογαρίθμους τους και τις διαφορές στους λογαρίθμους των πωλήσεων και του διαθέσιμου εισοδήματος για την περίοδο 1976-1996, όπως και τα αποτελέσματα από την ανάλυση παλινδρόμησης.

	Sales ($1.000)		**Income ($millions)**		**Differences**	
Year	$\mathbf{Y_t}$	$\mathbf{LnY_t}$	$\mathbf{X_t}$	$\mathbf{LnX_t}$	$\mathbf{Y'_t}$	$\mathbf{X'_t}$
1976	3.307	8,1038	273,4	5,6109	-	-
1977	3.556	8,1764	291,3	5,6744	0,0726	0,0634
1978	3.601	8,189	306,9	5,7265	0,0126	0,0522
1979	3.721	8,2218	317,1	5,7592	0,0328	0,0327
1980	4.036	8,303	336,1	5,8174	0,0813	0,0582
1981	4.134	8,327	349,4	5,8562	0,024	0,0388
1982	4.268	8,3589	362,9	5,8941	0,0319	0,0379
1983	4.578	8,429	383,9	5,9504	0,0701	0,0563
1984	5.093	8,5356	402,8	5,9984	0,1066	0,0481
1985	5.716	8,651	437	6,0799	0,1154	0,0815
1986	6.357	8,7573	472,2	6,1574	0,1063	0,0775
1987	5.769	8,8201	510,4	6,2352	0,0628	0,0778
1988	7.296	8,8951	544,5	6,2999	0,075	0,0647
1989	8.178	9,0092	588,1	6,3769	0,1141	0,077
1990	8.844	9,0875	630,4	6,4464	0,0783	0,0695
1991	9.251	9,1325	685,9	6,5307	0,045	0,0844
1992	10.006	9,2109	742,8	6,6104	0,0785	0,0797
1993	11.200	9,3237	801,3	6,6862	0,1127	0,0758

1994	12.500	9,4335	903,1	6,8058	0,1098	0,1196
1995	13.101	9,4804	983,6	6,8912	0,047	0,0854
1996	13.604	9,5208	1.076,70	6,9817	0,0403	0,0904

Regression Analysis: Ln(Sales) versus Ln(Income)

The regression equation is
Ln(Sales) = 1,82 + 1,12 Ln(Income)

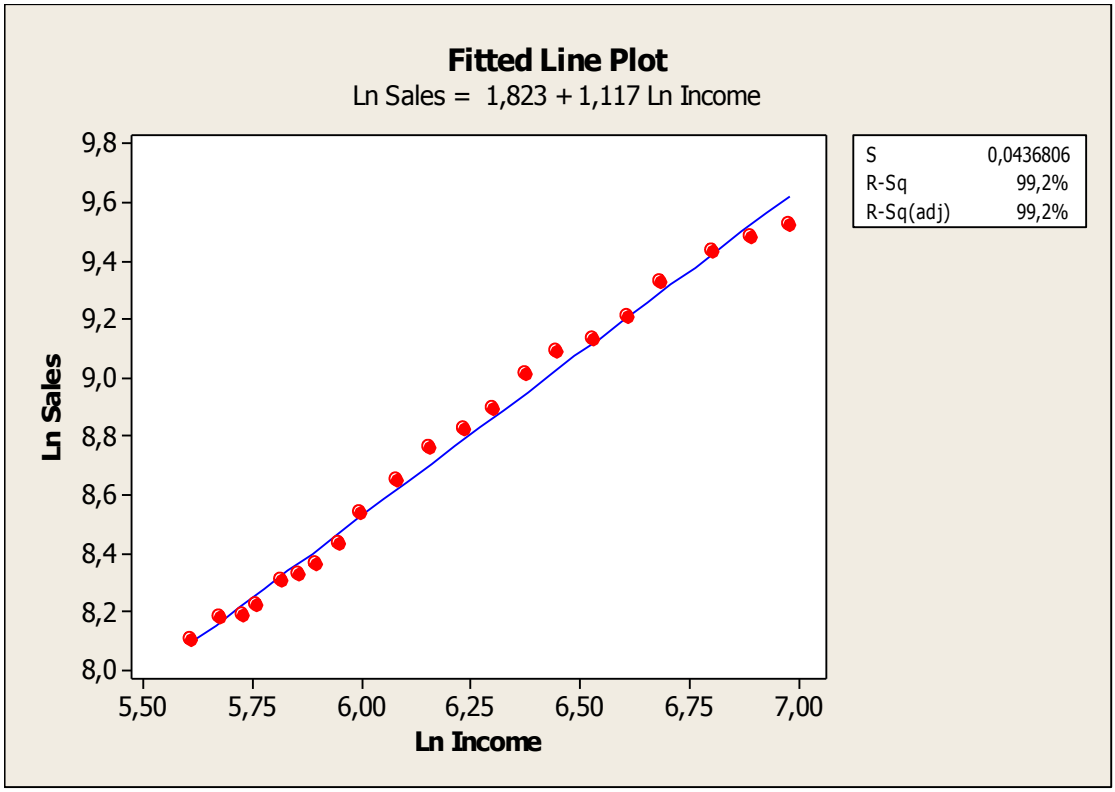
Predictor	Coef	SE Coef	T	P
Constant	1,8232	0,1434	12,71	0,000
Ln(Income)	1,11727	0,02305	48,47	0,000

S = 0,0436806 R-Sq = 99,2% R-Sq(adj) = 99,2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4,4821	4,4821	2349,13	0,000
Residual Error	19	0,0363	0,0019		
Total	20	4,5184			

Durbin-Watson statistic = 0,496346



Από τον πίνακα αποτελεσμάτων ανάλυσης παλινδρόμησης παρατηρούμε, ότι το 99,2% της μεταβλητότητας στο λογάριθμο των πωλήσεων για τη δυτική περιοχή μπορεί να ερμηνευθεί από τη σχέση με το λογάριθμο του διαθέσιμου εισοδήματος για την ίδια περιοχή. Η παλινδρόμηση είναι πολύ σημαντική. Επίσης, η αλλαγή του ποσοστού του εισοδήματος εκτιμάται να είναι b_1=1,117, με τυπικό σφάλμα $s_{b_1}=0{,}023$. Ωστόσο, το DW = 0,5 είναι μικρό και μικρότερο από το d_L=0,97, τη χαμηλότερη του 0,01 επιπέδου κρίσιμη τιμή για n=21 και k=1. Συμπεραίνεται λοιπόν, ότι η συσχέτιση μεταξύ παρακείμενων σφαλμάτων είναι θετική και μεγάλη (κοντά στο 1).

Εξαιτίας της μεγάλης σειριακής συσχέτισης, η εταιρία αποφάσισε να μοντελοποιήσει τις αλλαγές ή διαφορές στους λογάριθμους των πωλήσεων και του εισοδήματος, αντίστοιχα. Ήξερε ότι ο συντελεστής κλίσης στο μοντέλο για τις διαφορές είναι ο ίδιος συντελεστής κλίσης με αυτόν στο αρχικό μοντέλο, που περιλαμβάνει τους λογάριθμους, λόγω της σχέσης $\Delta(\ln Y_t)=\beta_1\Delta(\ln X_t)+v_t$. Επομένως, θα μπορούσε ακόμη να εκτιμήσει απευθείας την ελαστικότητα του εισοδήματος. Ο σταθερός συντελεστής (Y-intercept) στο μοντέλο παλινδρόμησης για τις διαφορές είναι μικρός και παραλήφθηκε. Τα αποτελέσματα του Minitab για τις αλλαγές φαίνονται στον παρακάτω Πίνακα.

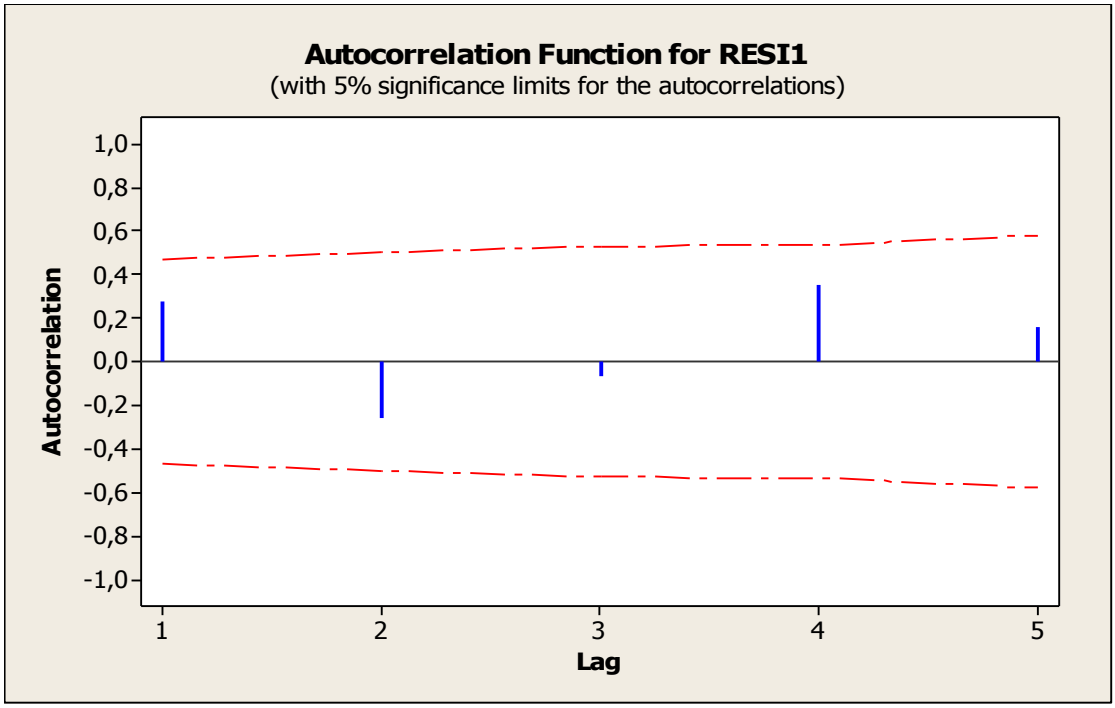
Regression Analysis: Change in Ln(Sales) versus Change in Ln(Income)

```
The regression equation is
Change in Ln(Sales) = 1,01 Change in Ln(Income)

Predictor          Coef  SE Coef   T     P
Noconstant
Change in
Ln(Income)       1,00989 0,09300 10,86 0,000

S = 0,0297390
```

```
Analysis of Variance
Source          DF    SS      MS       F        P
Regression       1 0,10430 0,10430 117,93 0,000
Residual Error 19 0,01680 0,00088
Total          20 0,12110

Durbin-Watson statistic = 1,27967
```

Ο Πίνακας δείχνει ότι η παλινδρόμηση είναι σημαντική. Η ελαστικότητα του εισοδήματος εκτιμάται να είναι b_1=1,010, με τυπικό σφάλμα $s_{b_1}=0{,}093$. Η εκτίμηση ελαστικότητας, b_1, δεν άλλαξε πολύ από την αρχική παλινδρόμηση (μία κατά 1% αύξηση στο διαθέσιμο εισόδημα οδηγεί σε μία περίπου 1% αύξηση στις ετήσιες πωλήσεις και στις δύο περιπτώσεις), αλλά το παρών τυπικό σφάλμα $(s_{b_1}=0{,}093)$ είναι περίπου τέσσερις φορές μεγαλύτερο από το προηγούμενο τυπικό σφάλμα $(s_{b_1}=0{,}023)$. Το προηγούμενο τυπικό σφάλμα είναι πιθανόν να υποτιμήσει το αληθινό τυπικό σφάλμα εξαιτίας της σειριακής συσχέτισης.

Ελέγχοντας με τη Durbin-Watson στατιστική για n=20, k=1 και ένα επίπεδο σημαντικότητας 5%, βρίσκουμε ότι d_L=1,20<DW=1,28<d_U=1,41, έτσι ο έλεγχος για θετική σειριακή συσχέτιση είναι αναποτελεσματικός. Ωστόσο, ένας έλεγχος για αυτοσυσχετίσεις καταλοίπων, που φαίνεται στο παρακάτω Γράφημα, υποδεικνύει ότι όλα είναι μεταξύ των δύο ορίων τυπικών σφαλμάτων (οι διακεκομμένες γραμμές στο γράφημα) για τα πρώτα λίγα lags. Η εταιρία συμπέρανε ότι η σειριακή συσχέτιση εξαλείφθηκε και χρησιμοποίησε την προσαρμοσμένη εξίσωση για πρόβλεψη.



Lag	ACF	T	LBQ
1	0,274432	1,23	1,74
2	-0,257852	-1,08	3,37
3	-0,065264	-0,26	3,48
4	0,349965	1,38	6,85
5	0,160779	0,58	7,61

Για να χρησιμοποιήσει το τελικό μοντέλο για πρόβλεψη, έγραψε:

$$\hat{Y}_t' = b_1 X_t'$$

όπου:

$$\hat{Y}_t' = \ln \hat{Y}_t - \ln \hat{Y}_{t-1}$$

$$X_t' = \ln X_t - \ln X_{t-1}$$

αντικαθιστώντας στη σχέση των διαφορών έχουμε:

$$\ln \hat{Y}_t = \ln \hat{Y}_{t-1} + b_1 (\ln X_t - \ln X_{t-1})$$

$$\ln \hat{Y}_{22} = \ln \hat{Y}_{21} + 1{,}01 (\ln X_{22} - \ln X_{21})$$

Θέτοντας τις τιμές για το t=21, $\hat{Y}_{21} = Y_{21}$, στην προαναφερόμενη σχέση και κάνοντας μία πρόβλεψη για την τιμή του εισοδήματος X_{22}

στην τιμή των 1.185 million $ και μετά από μερικές πράξεις έχουμε:

$$\ln \hat{Y}_{22} = 9,6176 \quad \textit{και} \quad \hat{Y}_{22} = e^{9,6176} = 15.027$$

Η πρόβλεψη της εταιρίας για τις πωλήσεις του 1997 για τη δυτική περιοχή ήταν $15,027 σε εκατομμύρια. Η εταιρία μπορούσε να χρησιμοποιήσει την Εξίσωση και τη διαδικασία, που περιγράφηκε νωρίτερα, για να γενικεύσει προβλέψεις για τις χρονιές 1998, 1999 και ούτε καθεξής, αλλά για να το κάνει αυτό, θα χρειαζόταν εκτιμήσεις του διαθέσιμου προσωπικού εισοδήματος για εκείνες τις χρονιές.

8.5.3 ΑΥΤΟΣΥΣΧΕΤΙΖΟΜΕΝΑ ΣΦΑΛΜΑΤΑ-ΓΕΝΙΚΕΥΜΕΝΕΣ ΔΙΑΦΟΡΕΣ

Το ζητούμενο είναι να περιγραφεί επαρκώς η φύση της σχέσης μεταξύ των μεταβλητών X και Y, όταν υπάρχει σειριακή συσχέτιση σφαλμάτων με $0<\rho<1$.

Θεωρούμε ξανά το μοντέλο παλινδρόμησης με σειριακά συσχετιζόμενα σφάλματα:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$$

Με τη χρήση των OLS (Ordinary Least Squares), δεν εξαλείφουμε τη σειριακή συσχέτιση των σφαλμάτων. Οπότε χρησιμοποιούμε τα Nonlinear Least Squares (NLS) με τον ακόλουθο τρόπο:

1. Υπολογίζουμε το μοντέλο με OLS (Ordinary Least Squares) και υπολογίζουμε τα κατάλοιπα $\hat{\varepsilon}_t$.
2. Εφαρμόζουμε την παλινδρόμηση στα κατάλοιπα με την υστέρησή τους $\hat{\varepsilon}_t = \rho \hat{\varepsilon}_{t-1} + \nu_t$. Έτσι υπολογίζουμε τον συντελεστή $\hat{\rho}$ ανάμεσα στα σφάλματα και την υστέρησή τους $\hat{\varepsilon}_t = \hat{\rho} \hat{\varepsilon}_{t-1}$.
3. Υπολογίζοντας το $\hat{\rho}$ σχηματίζουμε τις γενικευμένες διαφορές

$Y_t^{'} = Y_t - \hat{\rho}\, Y_{t-1}$ και $X_t^{'} = X_t - \hat{\rho}\, X_{t-1}$. Με αυτό τον τρόπο χάνουμε την πρώτη παρατήρηση.

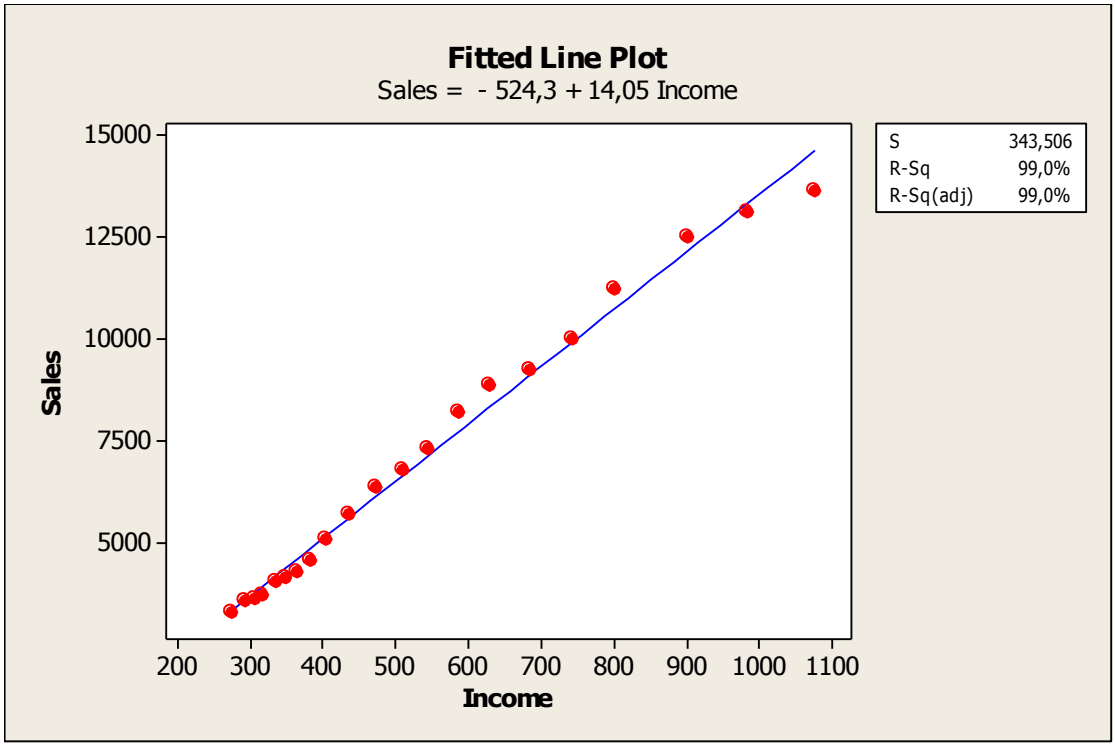
4. Μετά εφαρμόζουμε την παλινδρόμηση στα $Y_t^{'}$ *και* $X_t^{'}$ και βρίσκουμε τον συντελεστή, που θα είναι $\beta_0(1-\hat{\rho})$. Μετά υπολογίζουμε τα $\hat{Y}_t$ και τα κατάλοιπα από τις διαφορές $Y_t - \hat{Y}_t$.
5. Εφαρμόζουμε ξανά παλινδρόμηση στα κατάλοιπα και υπολογίζουμε ένα νέο $\hat{\rho}$.
6. Πηγαίνουμε πίσω στο βήμα 3 και επαναλαμβάνουμε τη διαδικασία, μέχρις ότου έχουμε σύγκλιση.

Στο ακόλουθο παράδειγμα θα εφαρμόσουμε τη μέθοδο των μη γραμμικών ελαχίστων τετραγώνων. Επειδή η μέθοδος αυτή δεν συμπεριλαμβάνεται στο λογισμικό Minitab, χρησιμοποιούμε τα αποτελέσματα από το λογισμικό E-Views.

Παράδειγμα (Τ.8-5)

Τα δεδομένα δίνονται στον Πίνακα του προηγούμενου παραδείγματος. Η συνήθης μέθοδος ελαχίστων τετραγώνων σε επίπεδο τιμών χρησιμοποιείται για να προσαρμόσει ένα μοντέλο απλής γραμμικής παλινδρόμησης, που συσχετίζει τις πωλήσεις με το διαθέσιμο προσωπικό εισόδημα. Τα παρακάτω αποτελέσματα προέρχονται από το E-views. Η προσαρμοσμένη εξίσωση παλινδρόμησης είναι:

$$\hat{Y}_t = -524{,}33 + 14{,}05\, X_t$$



Όπου:

- b_1=14,05,
- $s_{b_1}=0{,}319$,
- $t=b_1/s_{b_1}=44{,}11$,
- r^2=0,99,
- DW=0,63.

Επειδή το DW είναι κοντά στο μηδέν, υποθέτουμε ότι υπάρχει σειριακή συσχέτιση και εφαρμόζουμε το μοντέλο γενικευμένων διαφορών με NLS.

$$Y_t^{'}=\beta_0(1-\rho)+\beta_1 X_t^{'}+\varepsilon_t \quad \text{με}$$

$$Y_t^{'}=Y_t-\rho Y_{t-1} \quad \text{και} \quad X_t^{'}=X_t-\rho X_{t-1}$$

Οπότε η συνάρτηση δίνεται από τη σχέση:

$$\hat{Y}_t^{'}=54{,}483(1-0{,}997)+9{,}26 X_t^{'}$$

Όπου:

- b_1=9,26 και b_0=54,483
- $s_{b_1} = 7{,}241$
- $t = b_1 / s_{b_1} = 1{,}28$
- $\hat{\rho} = 0{,}997$
- και DW=1,12

Επειδή η συσχέτιση των σφαλμάτων είναι κοντά στο ένα, υποδεικνύεται, ότι η σχέση ανάμεσα στα Y και X μπορεί να παρασταθεί και από απλή παλινδρόμηση απλών διαφορών.

8.5.4 ΜΟΝΤΕΛΑ ΑΥΤΟΠΑΛΙΝΔΡΟΜΗΣΗΣ

Η αυτοσυσχέτιση υποδεικνύει, ότι οι τιμές της εξαρτημένης μεταβλητής σε μία χρονική περίοδο σχετίζονται γραμμικά με τις τιμές της εξαρτημένης μεταβλητής σε άλλη χρονική περίοδο. Επομένως, ένας τρόπος, για να λυθεί το πρόβλημα της σειριακής συσχέτισης, είναι να μοντελοποιηθεί η συσχέτιση απευθείας σε διαφορετικές χρονικές περιόδους. Αυτό μπορεί να γίνει στα πλαίσια παλινδρόμησης, χρησιμοποιώντας την εξαρτημένη μεταβλητή χρονο-υστερημένη κατά μία ή περισσότερες χρονικές περιόδους, όπως η ανεξάρτητη μεταβλητή ή μεταβλητή πρόβλεψης. Το μοντέλα παλινδρόμησης που μορφοποιούνται με αυτό τον τρόπο καλούνται αυτοπαλινδρομικά μοντέλα. Το <u>πρώτης τάξης μοντέλο αυτοπαλινδρόμησης AR(1) γράφεται:</u>

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t ,$$

με τα σφάλματα να έχουν τις συνήθεις ιδιότητες της παλινδρόμησης. Με τη συνήθη μέθοδο ελαχίστων τετραγώνων η παλινδρόμηση δίνεται από τη σχέση:

$$\hat{Y}_t = b_0 + b_1 Y_{t-1} .$$

Ένα αυτοπαλινδρομικό μοντέλο εκφράζει μία πρόβλεψη σαν μία

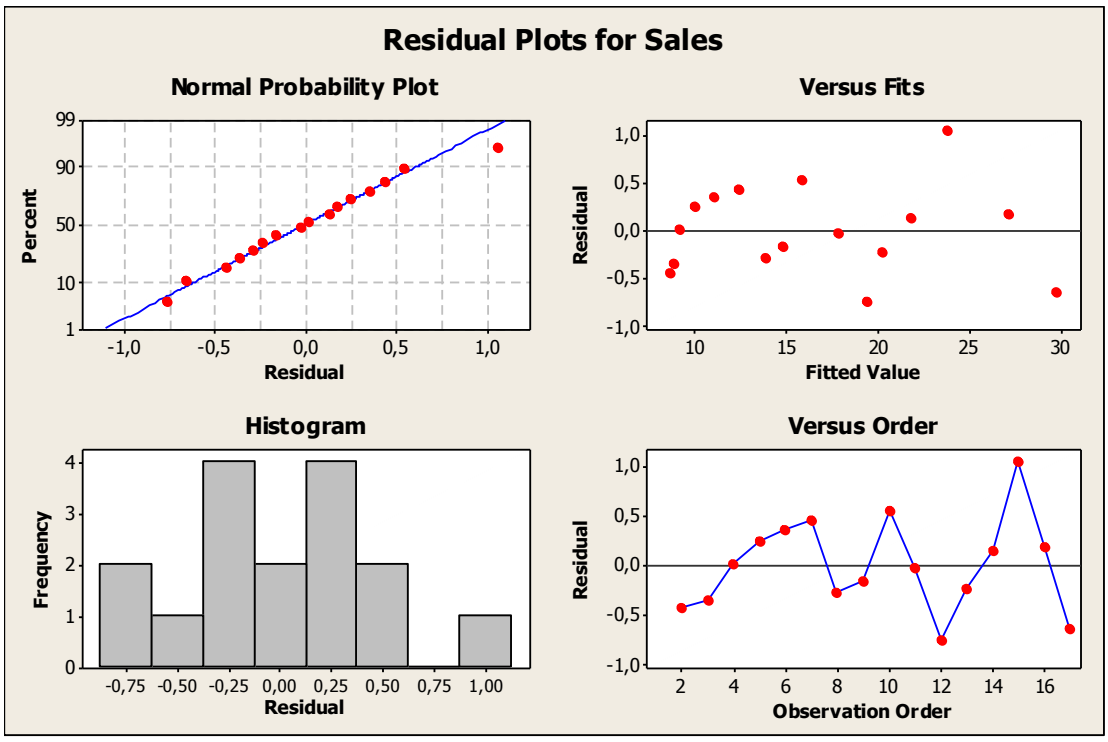
συνάρτηση των προηγούμενων τιμών των χρονοσειρών.

Τα αυτοπαλινδρομικά μοντέλα είναι ένα υποσύνολο των ARIMA μοντέλων, που θα συζητηθούν λεπτομερώς στο Κεφάλαιο 9.

Παράδειγμα (Τ.8-2)

Τα δεδομένα πωλήσεων της εταιρείας NOVAK. που παρουσιάστηκαν προηγουμένως, θα χρησιμοποιηθούν για να παρουσιάσουν την εξέλιξη ενός αυτοπαλινδρομικού μοντέλου. Το μέγεθος του δείγματος είναι n=16 αντί για n=17. Ένα πρώτης-τάξης αυτοπαλινδρομικό μοντέλο αναπτύσσεται με τις πωλήσεις να καθυστερούν ένα χρόνο (Y-lagged) σαν μία μεταβλητή πρόβλεψης. Εφαρμόζοντας παλινδρόμηση ανάμεσα στα Y_t *και* Y_{t-1}, παίρνουμε την ακόλουθη σχέση:

$$\hat{Y}_t = -0{,}109 + 1{,}094\, Y_{t-1}$$



Η πρόβλεψη των πωλήσεων για το 2007:

$$\hat{Y}_{18} = -0{,}109 + 1{,}094\, Y_{18-1} = -0{,}109 + 1{,}094\,(29{,}1) = 31{,}726.$$

Η τιμή που βρήκαμε προηγουμένως ήταν 32,7.

Στην περίπτωση της αυτοπαλινδρόμησης δεν μπορεί να χρησιμοποιηθεί η Durbin Watson στατιστική, η οποία στατιστική χρησιμοποιείται μόνο σε σειριακή συσχέτιση των καταλοίπων μεταξύ διαφορετικών μεταβλητών. Στην περίπτωση αυτή χρησιμοποιείται η στατιστική h του Durbin, που δίδεται από τη σχέση:

$$h = r_1(e)\sqrt{\frac{n}{1-ns_{b_1}}},$$ όπου:

- n: το μέγεθος του δείγματος.
- r_1: ο συντελεστής αυτοσυσχέτισης.
- s_{b_1}: η τυπική απόκλιση του b_1.

Για μεγάλα δείγματα και για μηδενική υπόθεση, ότι ρ=0, η στατιστική h ακολουθεί την κανονική κατανομή και κατά τα γνωστά θεωρείται, ότι η μηδενική υπόθεση δεν απορρίπτεται, όταν $Z_{-\alpha/2}<h<Z_{\alpha/2}$. Η στατιστική αυτή έχει νόημα μόνο, όταν ο παρονομαστής είναι θετικός.

8.5.5 ΣΥΝΟΨΗ

Όταν η ανάλυση παλινδρόμησης εφαρμόζεται σε δεδομένα χρονοσειρών, τα κατάλοιπα είναι συχνά αυτοσυσχετιζόμενα. Μιας και η ανάλυση παλινδρόμησης υποθέτει, ότι τα σφάλματα είναι ανεξάρτητα, δημιουργούνται προβλήματα. Το R^2 για μία παλινδρόμηση με δεδομένα, που περιέχουν αυτοσυσχέτιση, μπορεί να είναι σημαντικά υψηλό. Επιπλέον, τα τυπικά σφάλματα των συντελεστών παλινδρόμησης μπορούν να υποτιμηθούν σημαντικά και τα αντίστοιχα t και F στατιστικά να διογκωθούν.

Κάτι, που προκαλεί την αυτοσυσχέτιση των καταλοίπων, είναι η παράλειψη μίας ή περισσότερων σημαντικών μεταβλητών. Αυτή η

παράλειψη συχνά σημαίνει, ότι ένα σημαντικό μέρος της μεταβολής της εξαρτημένης μεταβλητής δεν έχει ερμηνευθεί επαρκώς. Μία λύση σε αυτό το πρόβλημα είναι να ψάξουμε για τις μεταβλητές, που λείπουν και να τις συμπεριλάβουμε στο μοντέλο. Άλλες λύσεις στα προβλήματα που προκαλούνται από αυτοσυσχέτιση είναι να θεωρηθούν μοντέλα παλινδρόμησης με δεδομένα της μεθόδου διαφορών, γενικευμένων διαφορών, ή / και μοντέλα αυτοπαλινδρόμησης.

8.6 ΔΕΔΟΜΕΝΑ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΕΤΕΡΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑΣ

Η μεταβλητότητα – διακύμανση σε μερικές χρονοσειρές τείνει να αυξάνει με το επίπεδο των τιμών των σειρών. Η μεταβλητότητα μπορεί να αυξηθεί αν μία μεταβλητή αυξάνει με σταθερό ρυθμό αντί για μία σταθερή ποσότητα στο χρόνο. Μη σταθερή μεταβλητότητα ονομάζεται ετεροσκεδαστικότητα (heteroscedasticity).

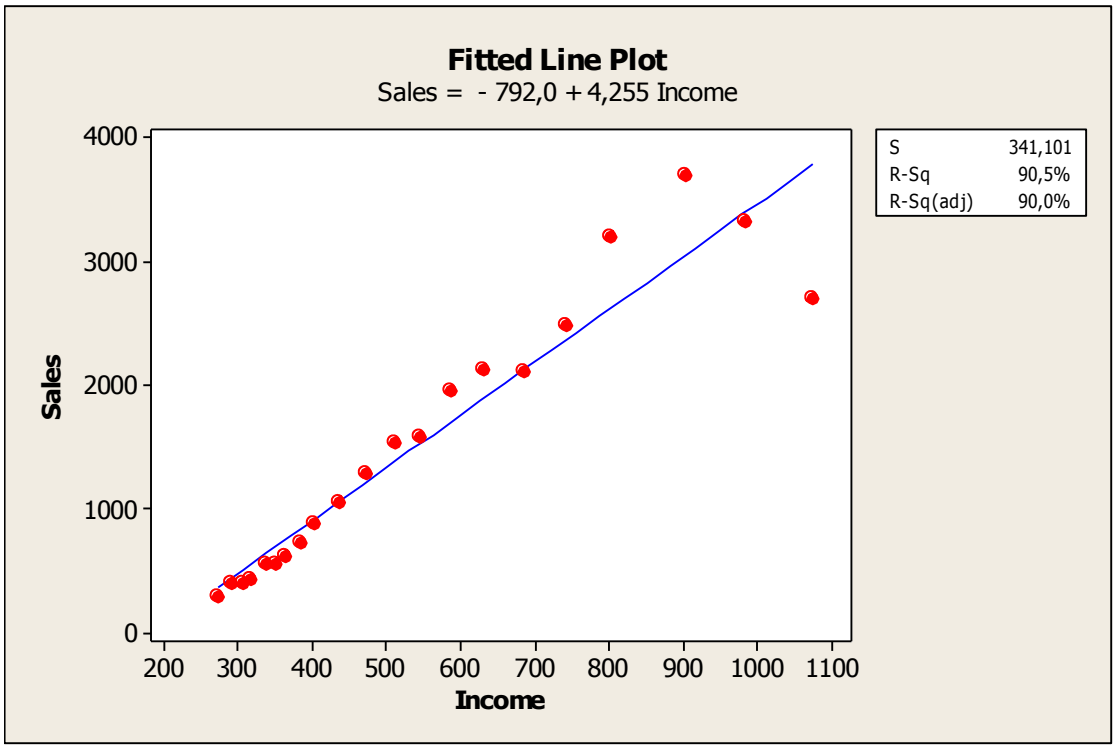
Σε ένα πλαίσιο παλινδρόμησης, η ετεροσκεδαστικότητα συμβαίνει, αν η διασπορά του όρου σφάλματος ε, δεν είναι σταθερή. Αν η μεταβλητότητα για πρόσφατες χρονικές περιόδους είναι μεγαλύτερη από ότι ήταν για προηγούμενες χρονικές περιόδους, τότε το τυπικό σφάλμα εκτίμησης $s_{x \cdot y}$, υποτιμά τη τωρινή τυπική απόκλιση του όρου σφάλματος. Αν χρησιμοποιηθεί η τυπική απόκλιση της εκτίμησης, τότε για τη δημιουργία ορίων πρόβλεψης για μελλοντικές παρατηρήσεις, αυτά τα όρια μπορεί να είναι πολύ στενά για το καθορισμένο επίπεδο εμπιστοσύνης.

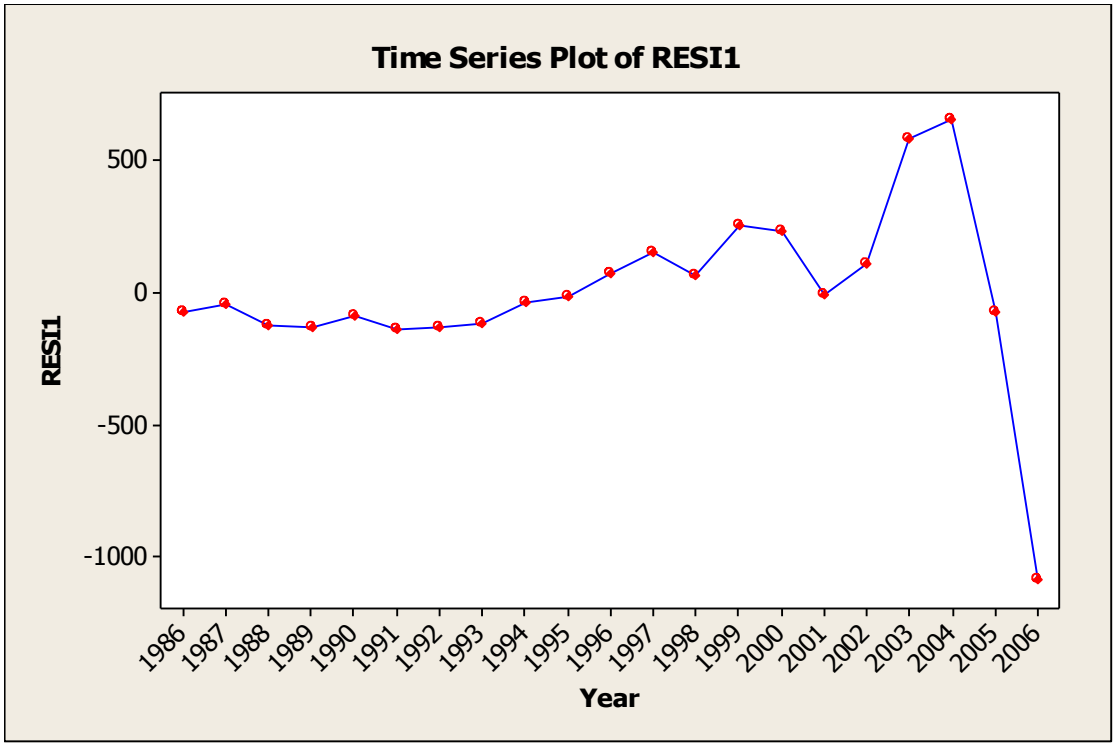
Μερικές φορές το πρόβλημα της ετεροσκεδαστικότητας μπορεί να λυθεί με απλό μετασχηματισμό των δεδομένων. Για παράδειγμα, στην περίπτωση δύο μεταβλητών, μπορεί να χρησιμοποιηθεί το λογαριθμικό

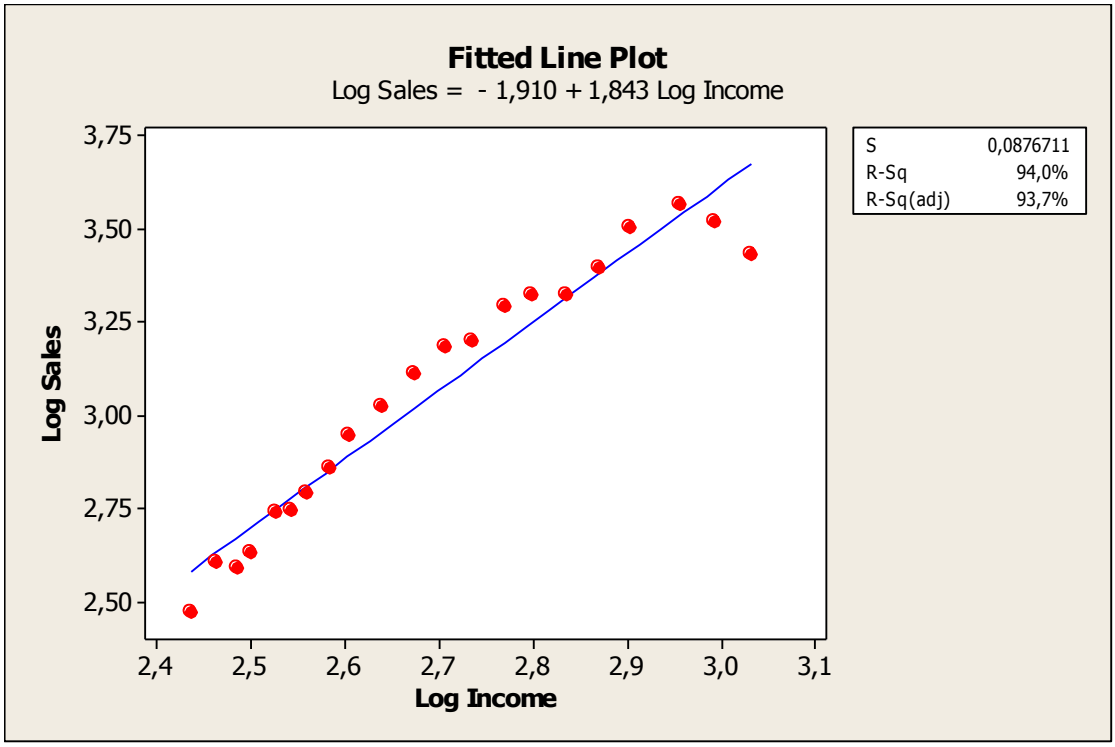
γραμμικό μοντέλο για να μειώσει την ετεροσκεδαστικότητα, όπως επίσης και το μη γραμμικό μοντέλο. Επίσης, αν οι μεταβλητές εκφράζονται σαν ποσότητες νομισμάτων (ευρώ, δολάρια κλπ), η μετατροπή των τρεχουσών τιμών νομισμάτων σε σταθερά νομίσματα (αποπληθωρισμένα) μπορεί να οδηγήσει σε λύση του προβλήματος της αύξησης της μεταβλητότητας σφάλματος.

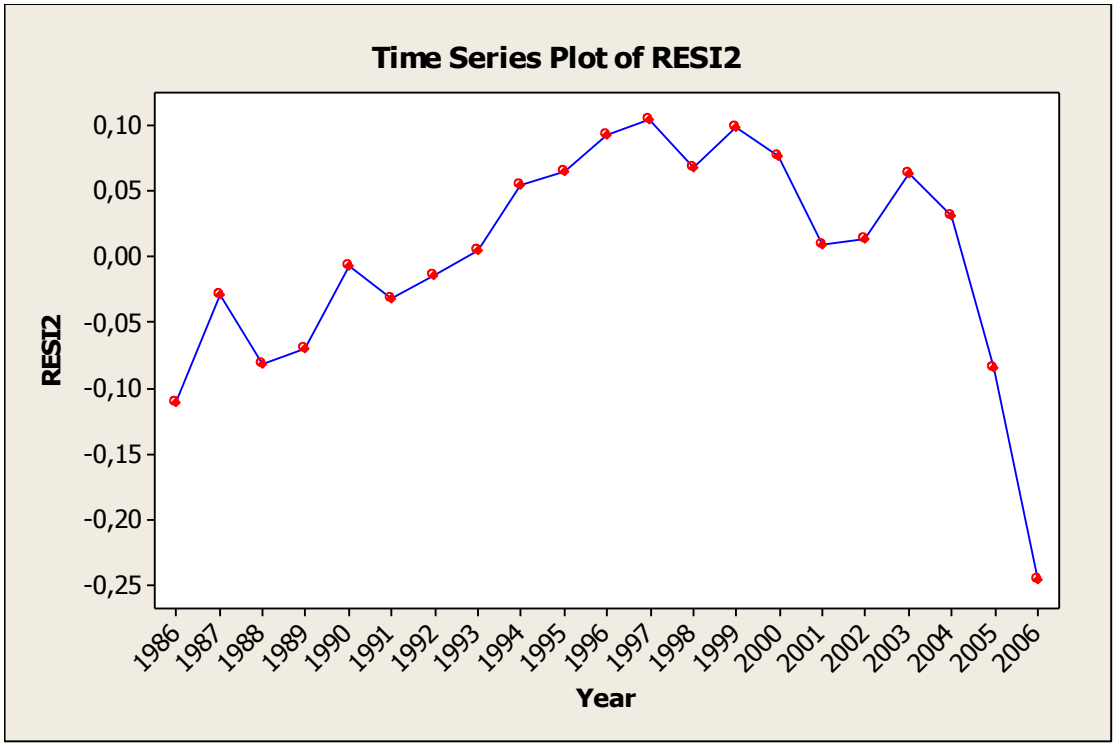
Παράδειγμα (Τ.8-1)

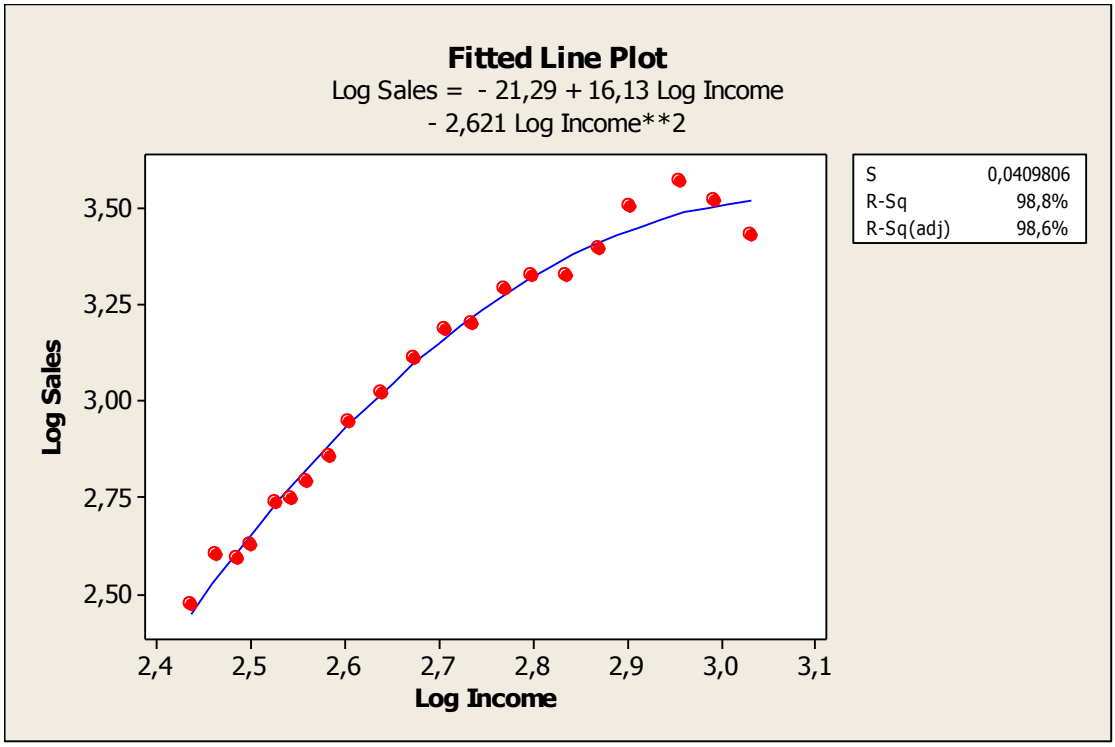
Πωλήσεις μιας εταιρείας ως συνάρτηση των εισοδημάτων του πληθυσμού στη περιοχή. Με την λογαρίθμιση των μεταβλητών και μη γραμμική παλινδρόμηση μειώσαμε το σφάλμα και την συσχέτισή τους.

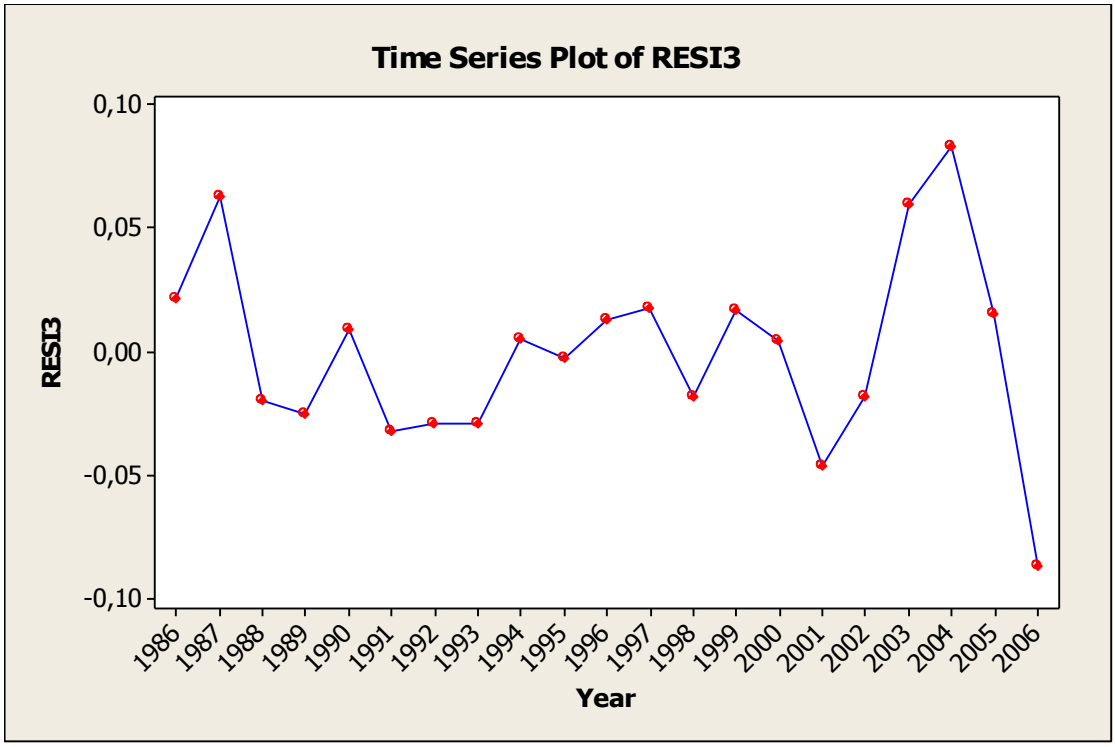



Time Series Plot of RESI1
RESI1
500
0
-500
-1000
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
Year


Fitted Line Plot
Log Sales = - 1,910 + 1,843 Log Income
S 0,0876711
R-Sq 94,0%
R-Sq(adj) 93,7%
Log Sales
3,75
3,50
3,25
3,00
2,75
2,50
2,4
2,5
2,6
2,7
2,8
2,9
3,0
3,1
Log Income


Time Series Plot of RESI2
RESI2
0,10
0,05
0,00
-0,05
-0,10
-0,15
-0,20
-0,25
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
Year


Fitted Line Plot
Log Sales = - 21,29 + 16,13 Log Income
- 2,621 Log Income**2
S 0,0409806
R-Sq 98,8%
R-Sq(adj) 98,6%
Log Sales
3,50
3,25
3,00
2,75
2,50
2,4
2,5
2,6
2,7
2,8
2,9
3,0
3,1
Log Income



Regression Analysis: Log Sales versus Log Income; (Log Income)**2

The regression equation is
Log Sales = - 21,3 + 16,1 Log Income - 2,62 (Log Income)**2

Predictor	Coef	SE Coef	T	P
Constant	-21,294	2,338	-9,11	0,000
Log Income	16,130	1,721	9,37	0,000
(Log Income)**2	-2,6207	0,3156	-8,30	0,000

S = 0,0409806 R-Sq = 98,8% R-Sq(adj) = 98,6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	2,4160	1,2080	719,31	0,000
Residual Error	18	0,0302	0,0017		
Total	20	2,4463			

Durbin-Watson statistic = 1,34855

8.7 ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΕΠΟΧΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Στο Κεφάλαιο 5 συζητήθηκαν τα μοντέλα αποσύνθεσης για χρονοσειρές με εποχικά μοτίβα. Ένα μοντέλο παλινδρόμησης, που αναπαριστά εποχικά δεδομένα και προσεγγίζει στενά την προσθετική αποσύνθεση (additive decomposition), δίνεται παρακάτω. Σε αυτό το μοντέλο, η εποχικότητα πραγματεύεται χρησιμοποιώντας ψευδομεταβλητές στη συνάρτηση παλινδρόμησης.

Ένα εποχικό μοντέλο για τριμηνιαία δεδομένα με μία χρονική τάση είναι:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \varepsilon_t .$$

- Y_t = η προς πρόβλεψη μεταβλητή.
- t = ο χρόνος.
- $S_{2,3,4}$ = Ψευδομεταβλητή για το δεύτερο, τρίτο και τέταρτο τρίμηνο. αντίστοιχα.
- ε_t = τα σφάλματα κανονικά κατανεμημένα και ανεξάρτητα.
- $\beta_{1,2,3,4}$ = οι προς προσδιορισμό συντελεστές.

Το πρώτο τρίμηνο συμπεριλαμβάνεται στη τιμή β_0. Για το τρίμηνο i παίρνουμε το $S_i = 1$ και τα άλλα 0. Για τα δεδομένα του πρώτου τριμήνου είναι:

$$E(Y_t) = \beta_0 + \beta_1 t.$$

Για τα δεδομένα του δεύτερου τριμήνου, $S_2 = 1, S_3 = S_4 = 0$ και το αναμενόμενο επίπεδο είναι:

$$E(Y_t) = \beta_0 + \beta_1 t + \beta_2 = (\beta_0 + \beta_2) + \beta_1 t.$$

Παράδειγμα (Τ.8-9)

Ο υπεύθυνος μιας εταιρίας ήθελε να προβλέψει τη κατανάλωση

του ηλεκτρικού ρεύματος για το τρίτο και τέταρτο τρίμηνο του 1996. Ήξερε ότι τα δεδομένα ήταν εποχικά και αποφάσισε να χρησιμοποιήσει το εποχικό μοντέλο για να αναπτύξει μία εξίσωση πρόβλεψης. Χρησιμοποίησε τριμηνιαία δεδομένα από το 1980 μέχρι το δεύτερο τρίμηνο του 1996. Τα δεδομένα για τη χρήση του ηλεκτρικού υπολογίστηκαν σε εκατομμύρια κιλοβατώρες.

Ο υπεύθυνος αυτός δημιούργησε ψευδομεταβλητές $S_{2,3,4}$, για να αναπαραστήσουν το δεύτερο, τρίτο και τέταρτο τρίμηνο αντίστοιχα. Τα δεδομένα για τα τέσσερα τρίμηνα του 1980 δίνονται στον παρακάτω Πίνακα.

Y_t	S_2	S_3	S_4
1071	0	0	0
648	1	0	0
480	0	1	0
746	0	0	1

Το προσαρμοσμένο μοντέλο εποχικής παλινδρόμησης είναι:

Regression Analysis: Hours versus Time; 2nd Qt.; 3rd Qt.; 4th Qt.

```
The regression equation is
Hours = 968 + 0,938 Time - 342 2nd Qt. - 472 3rd
Qt. - 230 4th Qt.

Predictor  Coef   SE Coef    T      P
Constant    968,39 16,88   57,38 0,000
Time         0,9383 0,3377   2,78 0,007
2nd Qt.    -341,94 17,92  -19,08 0,000
3rd Qt.    -471,60 18,20  -25,91 0,000
4th Qt.    -230,23 18,20  -12,65 0,000

S = 52,2488 R-Sq = 92,4% R-Sq(adj) = 91,9%

Analysis of Variance
Source          DF    SS     MS     F      P
Regression       4 2012975 503244 184,34 0,000
Residual Error 61 166526  2730
```

```
Total           65 2179502

Durbin-Watson statistic = 1,47765
```

Η πρόβλεψη για το τρίτο τρίμηνο του 1996 υπολογίστηκε σε:

Hours = 968 + 0,938(67) – 342(0) – 472(1) – 230(0) = 559

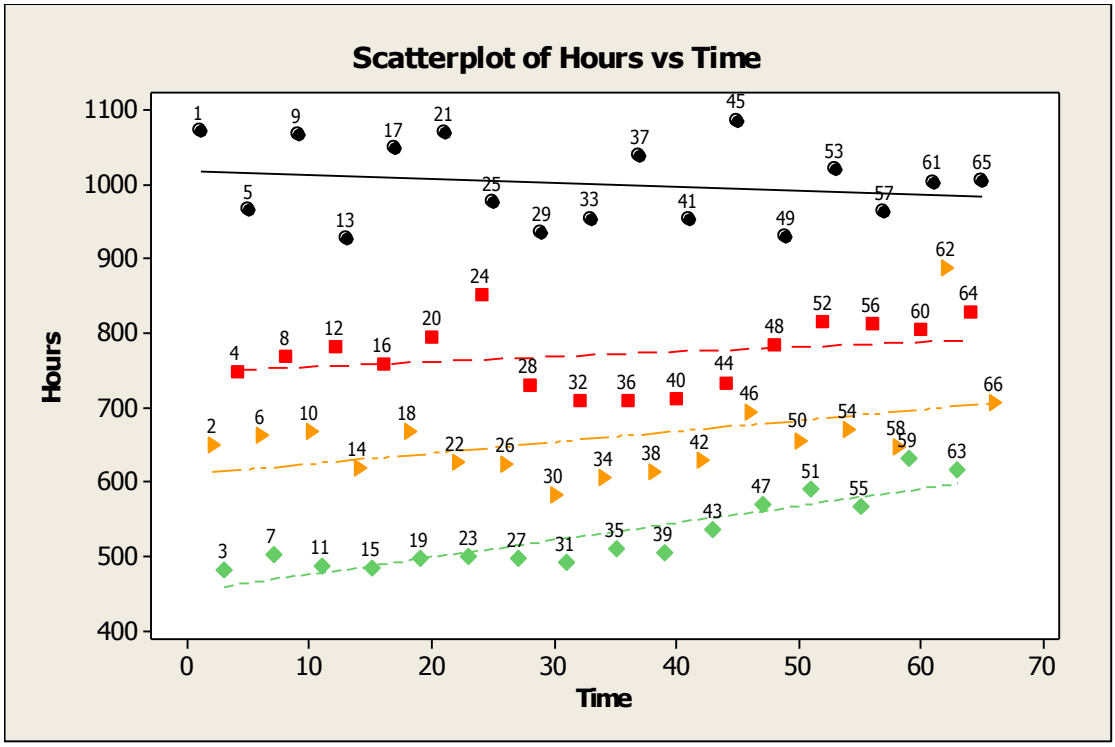
Η πρόβλεψη για το τέταρτο τρίμηνο του 1996 ήταν:

Hours = 968 + 0,938(68) – 342(0) – 472(0) – 230(1) = 802

Ο υπεύθυνος είδε, ότι οι προβλέψεις για τα διαφορετικά τρίμηνα βρίσκονται πάνω σε τέσσερις ευθείες γραμμές. Οι γραμμές έχουν όλες την ίδια κλίση (0,938), αλλά οι σταθεροί όροι διαφέρουν και εξαρτώνται από το εκάστοτε τρίμηνο. Η πρόβλεψη του πρώτου τριμήνου βρίσκεται πάνω σε μία γραμμή με σταθερό όρο 968. Η πρόβλεψη για το δεύτερο τρίμηνο βρίσκεται πάνω σε μία γραμμή με σταθερό όρο 968-342=626. Ο σταθερός όρος για το τρίτο τρίμηνο είναι 968-472=496, και ο σταθερός όρος για την πρόβλεψη του τέταρτου τριμήνου είναι 968-230=738. Ο υπεύθυνος ήταν ικανοποιημένος, που το μοντέλο πρόβλεψης συνέλαβε το εποχικό μοτίβο και παρατηρήθηκε ελαφριά ανοδική τάση στις σειρές.

Μέσα σε ένα δεδομένο χρόνο, οι προβλέψεις για τη χρήση ηλεκτρικού είναι υψηλότερες για το πρώτο τρίμηνο, χαμηλότερες για το δεύτερο τρίμηνο, οι χαμηλότερες για το τρίτο τρίμηνο, και οι δεύτερες υψηλότερες για το τέταρτο τρίμηνο.

Στο παρακάτω διάγραμμα φαίνονται οι ευθείες παλινδρόμησης για κάθε τρίμηνο ανεξάρτητα και με τη σειρά 1ο, 4ο, 2ο και 3ο. Οι κλίσεις των ευθειών εδώ δεν είναι ίδιες.



8.8 ΟΙΚΟΝΟΜΕΤΡΙΚΕΣ ΠΡΟΒΛΕΨΕΙΣ

Η εφαρμογή της μεθόδου παλινδρόμησης σε οικονομικά δεδομένα για παραγωγή προβλέψεων αναφέρεται ως οικονομικές προβλέψεις. Η οικονομική θεωρία απαιτεί πολλές φορές την δημιουργία και εξέταση ταυτόχρονων εξισώσεων. Αυτό οδηγεί στην κατασκευή ***οικονομετρικών μοντέλων ταυτόχρονων εξισώσεων***. Επειδή σε ταυτόχρονα συστήματα οι εξισώσεις αλληλοσυσχετίζονται, τα οικονομετρικά μοντέλα απαιτούν τον συλλογικό προσδιορισμό ενός συνόλου ανεξάρτητων μεταβλητών. Αυτό έρχεται σε αντίθεση με την μέθοδο της παλινδρόμησης, η οποία απαιτεί η εξαρτημένη μεταβλητή να εξαρτάται από ανεξάρτητες μεταβλητές και όχι από άλλες εξαρτημένες μεταβλητές άλλων εξισώσεων.

Όλες οι εξαρτημένες μεταβλητές, οι οποίες προσδιορίζονται από διαφορετικές εξισώσεις, είναι προς υπολογισμό και εμπλέκονται σε

άλλες, καλούνται από τους οικονομέτρες ***ενδογενείς μεταβλητές***. Οι ανεξάρτητες μεταβλητές καλούνται ***εξωγενείς μεταβλητές***. Οι εξωγενείς μεταβλητές προσδιορίζουν τις ενδογενείς αλλά όχι αντίστροφα. Σε ένα τέτοιο σύστημα ο αριθμός των εξισώσεων είναι ίδιος με τον αριθμό των ενδογενών μεταβλητών. Παρακάτω δίδεται ένα τέτοιο οικονομετρικό σύστημα εξισώσεων:

$$Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 I_t + \alpha_3 T_t + \varepsilon_t$$

$$P_t = \beta_0 + \beta_1 Q_t + \beta_2 L_t + \nu_t$$

όπου:

- Q_t: είναι η ζήτηση ενός προϊόντος.
- P_t: η αξία του προϊόντος.
- I_t: το κατά κεφαλήν εισόδημα.
- T_t: η θερμοκρασία.
- L_t: κόστος μισθοδοσίας.
- ε_t και ν_t: τα ανεξάρτητα και μη συσχετιζόμενα μεταξύ τους σφάλματα.

Στις εξισώσεις αυτές η ζήτηση και η αξία του προϊόντος είναι οι ενδογενείς μεταβλητές και εμφανίζονται και στις δύο εξισώσεις. Δηλαδή η ζήτηση του προϊόντος εξαρτάται από την αξία του και η αξία του από την ζήτηση. Πριν εφαρμόσουμε τη μέθοδο της παλινδρόμησης στο σύστημα αυτό, πρέπει πρώτα να επιλύσουμε αναλυτικά (μετασχηματίσουμε τις αρχικές εξισώσεις) ως προς τις ενδογενείς μεταβλητές, έτσι ώστε οι ενδογενείς μεταβλητές να εξαρτώνται μέσω νέων εξισώσεων μόνο από τις εξωγενείς ανεξάρτητες μεταβλητές. Οπότε οι συντελεστές (νέοι συντελεστές μετά το μετασχηματισμό) κάθε εξίσωσης θα μπορούν να βρεθούν μέσω παλινδρόμησης ανεξάρτητα η

μία από την άλλη.

8.9 CASE 8-2 BUSINESS ACTIVITY INDEX FOR SPOKANE COUNTY

Πριν από το 1973, το Spokane County στη Washington δεν είχε κάποια μέτρηση της γενικής επιχειρηματικής δραστηριότητας. Δεν υπήρχαν στοιχεία για το τι θα συμβεί σε αυτόν τον τομέα συνολικά, για το πως επηρεάζεται κάθε τοπική επιχείρηση, πως οι κυβερνητικές υπηρεσίες και πως οι μεμονωμένες υπηρεσίες. Σχέδια και πολιτικές που ακολουθούνται από μια οικονομική μονάδα θα ήταν ελλιπή χωρίς κάποια αξιόπιστη γνώση σχετικά με τις πρόσφατες επιδόσεις της οικονομίας. Διότι, η οικονομική μονάδα αποτελεί συστατικό ολόκληρης της οικονομίας. Ένας δείκτης επιχειρηματικής δραστηριότητας για το Spokane, θα μπορούσε να χρησιμοποιηθεί ως μια ζωτικής σημασίας είσοδος για τη διαμόρφωση στρατηγικών και αποφάσεων, τόσο σε ιδιωτικούς, όσο και σε δημόσιους οργανισμούς.

Ένας δείκτης επιχειρηματικής δραστηριότητας αναφέρεται στις σχετικές αλλαγές του συνόλου της επιχειρηματικής δραστηριότητας εντός συγκεκριμένης περιοχής. Σε εθνικό επίπεδο, το ακαθάριστο εθνικό προϊόν (υπολογιζόμενο από το Υπουργείο Εμπορίου) και ο δείκτης βιομηχανικής παραγωγής (υπολογιζόμενος από την Ομοσπονδιακή Τράπεζα των ΗΠΑ), θεωρούνται ως εξαιρετικοί δείκτες απόδοσης. Καθεμία από αυτές τις χρονοσειρές βασίζεται σε τεράστιες ποσότητες πληροφορίας – η συλλογή τους, η καταγραφή τους, ο υπολογισμός και η επεξεργασία τους είναι πολύ δαπανηρές και χρονοβόρες διαδικασίες. Για μία τοπική περιοχή, όπως το Spokane County, μία απλοποιημένη εκδοχή, η οποία είναι σε θέση να παρέχει λογικά ακριβείς και επίκαιρες

πληροφορίες σε μέτριο κόστος, είναι πολύ επιθυμητή.

Η πολυμεταβλητή παλινδρόμηση θα μπορούσε να χρησιμοποιηθεί για να δομήσει έναν δείκτη επιχειρηματικής δραστηριότητας. Υπάρχουν όμως τρεις σημαντικές ερωτήσεις, προκειμένου να ολοκληρωθεί η επεξεργασία αυτού του δείκτη:

- Ποια είναι τα συστατικά του δείκτη;
- Μπορούν αυτά τα συστατικά να αναπαραστήσουν επαρκώς όλες τις αλλαγές στις συνθήκες του επιχειρησιακού περιβάλλοντος;
- Τι «βάρος» θα πρέπει να δοθεί σε καθένα από τα επιλεγμένα συστατικά;

Οι απαντήσεις σε αυτές τις ερωτήσεις θα μπορούσαν να περιέχονται στην ανάλυση παλινδρόμησης.

Ο καθηγητής οικονομικών του Ανατολικού Πανεπιστημίου της Washington, Dr. Shik Chun Young, προσπαθεί να αναπτύξει έναν δείκτη επιχειρηματικής δραστηριότητας για το Spokane County. Ο Young επέλεξε το προσωπικό εισόδημα ως εξαρτημένη μεταβλητή. Σε επίπεδο νομού, το προσωπικό εισόδημα κρίνεται ως ο καλύτερος διαθέσιμος δείκτης των τοπικών επιχειρησιακών συνθηκών. Αναφερόμαστε σε προσωπικό εισόδημα, εννοώντας το συνολικό εισόδημα ενός νοικοκυριού χωρίς να έχουν αφαιρεθεί οι φόροι. Δεδομένου ότι οι παραγωγικές δραστηριότητες αμείβονται κατά κανόνα με νομισματικά μέσα, το προσωπικό εισόδημα μπορεί πράγματι να θεωρηθεί ως μεσολαβητής για τη γενική οικονομική απόδοση. Γιατί λοιπόν είναι απαραίτητο να δομηθεί ένας άλλος δείκτης, εφόσον το εισόδημα κρίνεται ως ένας καλός δείκτης επιχειρηματικής δραστηριότητας; Δυστυχώς, τα προσωπικά δεδομένα όπως τα εισοδήματα σε επίπεδο νομού εκτιμώνται από το Υπουργείο Εμπορίου των ΗΠΑ σε ετήσια βάση και στην περίπτωση μας

δημοσιεύθηκαν με 16 μήνες καθυστέρησης. Κατά συνέπεια, τα στοιχεία αυτά είναι ελάχιστα χρήσιμα για τον βραχυπρόθεσμο σχεδιασμό. Έτσι, το καθήκον του Young είναι να ορίσει ένα νέο δείκτη επιχειρηματικής δραστηριότητας.

Οι ανεξάρτητες μεταβλητές που προέρχονται από τα τοπικά δεδομένα είναι άμεσα διαθέσιμες σε μηνιαία βάση. Επί του παρόντος, υπάρχουν περίπου 50 χρονοσειρές διαθέσιμες με μηνιαία στοιχεία, τα οποία κυμαίνονται από τις δραστηριότητες της απασχόλησης, τις τραπεζικές συναλλαγές και τις κτηματομεσιτικές συναλλαγές έως την κατανάλωση ηλεκτρικής ενέργειας. Εάν καθεμία από αυτές τις χρονοσειρές μπορούσε να συμπεριληφθεί στην ανάλυση παλινδρόμησης, το αποτέλεσμα δεν θα ήταν παραγωγικό, διότι μόνο μία από αυτές τις χρονοσειρές θα πρέπει να είναι στατιστικά σημαντική. Επομένως, η όποια γνώση της σχέσης μεταξύ του προσωπικού εισοδήματος και των διαθέσιμων δεδομένων είναι αναγκαία, για να καθορίσει, ποιες είναι ανεξάρτητες μεταβλητές, που πρέπει να περιλαμβάνονται στην εξίσωση παλινδρόμησης. Από την εμπειρία και τη γνώση του Young για την οικονομία του Spokane, επιλέγονται οι ακόλουθες 10 χρονοσειρές:

- X_1, η συνολική απασχόληση
- X_2, η απασχόληση στον τομέα της μεταποίησης
- X_3, η απασχόληση στον κατασκευαστικό τομέα
- X_4, η απασχόληση στο χονδρικό και λιανικό εμπόριο
- X_5, η απασχόληση σε υπηρεσίες
- X_6, οι τραπεζικές χρεώσεις
- X_7, οι τραπεζικές καταθέσεις όψεως

- X_8, οι εκδιδόμενες οικοδομικές άδειες
- X_9, οι υποθήκες ακινήτων
- X_{10}, η συνολική κατανάλωση ηλεκτρικής ενέργειας

Το πρώτο βήμα της ανάλυσης είναι ο προσδιορισμός του μοντέλου:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{10} X_{10}$$

όπου

- Y = προσωπικό εισόδημα
- β_0 = ο σταθερός συντελεστής
- β_1, β_2, ……., β_{10} = οι συντελεστές των αντίστοιχων ανεξάρτητων μεταβλητών

Όταν το παραπάνω μοντέλο προσαρμόζεται στα δεδομένα, το προσαρμοσμένο R^2 είναι 0,96, γεγονός που σημαίνει, πως όταν οι δέκα ανεξάρτητες μεταβλητές χρησιμοποιηθούν ταυτόχρονα ερμηνεύουν το 96% της μεταβλητότητας της εξαρτημένης μεταβλητής (προσωπικό εισόδημα). Ωστόσο, άλλα στατιστικά στοιχεία παλινδρόμησης υποδεικνύουν προβλήματα. Αρχικά, από αυτές τις δέκα ανεξάρτητες μεταβλητές, μόνο οι τρεις – η συνολική απασχόληση, η υπηρεσιακή απασχόληση και οι τραπεζικές χρεώσεις – έχουν μια υπολογισμένη σημαντική t τιμή στο επίπεδο του 0,05. Δεύτερον, ο πίνακας συσχέτισης δείχνει έναν υψηλό βαθμό αλληλεξάρτησης μεταξύ των διαφόρων ανεξάρτητων μεταβλητών (πολυσυγγραμμικότητα). Για παράδειγμα, η συνολική απασχόληση και οι τραπεζικές χρεώσεις έχουν συντελεστή αυτοσυσχέτισης 0,88, η συνολική κατανάλωση ηλεκτρικής ενέργειας και οι τραπεζικές καταθέσεις όψεως έχουν συντελεστή αυτοσυσχέτισης 0,76,

οι οικοδομικές άδειες που εκδόθηκαν και οι υποθήκες ακινήτων έχουν συντελεστή αυτοσυσχέτισης 0,68. Τρίτον, ένας έλεγχος αυτοσυσχέτισης με χρήση της Durbin Watson στατιστικής (DW = 0,91) δείχνει ότι οι διαδοχικές τιμές της εξαρτημένης μεταβλητής είναι θετικά συσχετιζόμενες. Άλλωστε η αυτοσυσχέτιση είναι κοινό φαινόμενο σε δεδομένα χρονοσειρών: γενικά, οι παρατηρήσεις της ίδιας χρονοσειράς τείνουν να σχετίζονται μεταξύ τους.

Μια από τις βασικές παραδοχές της ανάλυσης παλινδρόμησης είναι η τυχαιότητα των παρατηρήσεων της εξαρτημένης μεταβλητής. Έτσι, ο Young επέλεξε να αντιμετωπίσει αρχικά το πρόβλημα αυτοσυσχέτισης. Αποφάσισε να υπολογίσει τις πρώτες διαφορές (ή μεταβολές) σε μια προσπάθεια να ελαχιστοποιήσει την αλληλεξάρτηση μεταξύ των παρατηρήσεων σε κάθε χρονοσειρά. Οι δέκα ανεξάρτητες μεταβλητές προκύπτουν τώρα από τη διαφορά μεταξύ των χρονικών περιόδων αντί της απόλυτης τιμής κάθε περιόδου. Έτσι, χρησιμοποιείται μια νέα ονομασία για τις ανεξάρτητες μεταβλητές:

- ΔX_1, η μεταβολή στη συνολική απασχόληση
- ΔX_2, η μεταβολή στη βιομηχανική απασχόληση
- ΔX_3, η μεταβολή στην οικοδομική απασχόληση
- ΔX_4, η μεταβολή στο χονδρικό και λιανικό εμπόριο
- ΔX_5, η μεταβολή στην απασχόληση σε υπηρεσίες
- ΔX_6, η μεταβολή στις τραπεζικές χρεώσεις
- ΔX_7, η μεταβολή στις τραπεζικές καταθέσεις όψεως
- ΔX_8, η μεταβολή στις εκδιδόμενες οικοδομικές άδειες

- ΔX_9, η μεταβολή στις υποθήκες ακινήτων
- ΔX_{10}, η μεταβολή στη συνολική κατανάλωση ηλεκτρικής ενέργειας

Το παλινδρομικό μοντέλο γίνεται:

$$E(\Delta Y) = \beta_0 + \beta_1 \Delta X_1 + \beta_2 \Delta X_2 + \ldots.. + \beta_{10} \Delta X_{10}$$

όπου

- ΔY = η μεταβολή στο προσωπικό εισόδημα
- β_0 = ο σταθερός συντελεστής
- $\beta_1, \beta_2, \ldots..$, β_{10} = οι συντελεστές των αντίστοιχων ανεξάρτητων μεταβλητών

Εκτελώντας το παλινδρομικό μοντέλο των πρώτων διαφορών, βρίσκουμε το Durbin-Watson στατιστικό ίσο με 1,71. Αυτό φανερώνει, πως παραμένει η χαμηλή αυτοσυσχέτιση.

Το επόμενο βήμα είναι να καθοριστεί, ποιες από τις δέκα ανεξάρτητες μεταβλητές είναι σημαντικές για την πρόβλεψη της εξαρτημένης μεταβλητής. Η εξαρτημένη μεταβλητή, ΔΥ, παλινδρομείται με διαφορετικούς συνδυασμούς των δέκα πιθανών μεταβλητών πρόβλεψης προκειμένου να επιλεγεί η καλύτερη εξίσωση. Τα κριτήρια επιλογής είναι τα ακόλουθα:

- Ένα ικανοποιητικά υψηλό προσαρμοσμένο R^2
- Χαμηλούς συντελεστές συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών
- Σημαντικούς συντελεστές (0,05) για κάθε ανεξάρτητη μεταβλητή

Μετά από προσεχτικό έλεγχο των αποτελεσμάτων της παλινδρόμησης, ο Young κατέληξε, πως η εξίσωση, που περιλαμβάνει τις

ανεξάρτητες μεταβλητές ΔX_4, ΔX_5, ΔX_{10}, προσεγγίζει καλύτερα τα παραπάνω κριτήρια.

Ωστόσο, ο Young εξήγησε, πως (παράλληλα με εμπορικές και βιομηχανικές χρήσεις) η συνολική κατανάλωση ηλεκτρικής ενέργειας περιλαμβάνει και την οικιακή κατανάλωση, η οποία δεν έχει σημαντική σχέση με την επιχειρησιακή δραστηριότητα στο εγγύς μέλλον. Για να ελέγξει αυτήν την υπόθεση, ο Young υποδιαίρεσε τη συνολική κατανάλωση ηλεκτρικής ενέργειας σε τέσσερις μεταβλητές:

- ΔX_{11}, η μεταβολή στην οικιακή χρήση της ηλεκτρικής ενέργειας.
- ΔX_{12}, η μεταβολή στην εμπορική χρήση της ηλεκτρικής ενέργειας.

Εξίσωση	**Ανεξάρτητες Μεταβλητές**	**Εξαρτημένη Μεταβλητή**
A	ΔX_4 , ΔX_5 , ΔX_{11}	ΔY
B	ΔX_4 , ΔX_5 , ΔX_{12}	ΔY
C	ΔX_4 , ΔX_5 , ΔX_{13}	ΔY
D	ΔX_4 , ΔX_5 , ΔX_{14}	ΔY

Πίνακας 8-12 Μεταβλητές Παλινδρόμησης κατά Young

- ΔX_{13}, η μεταβολή στη βιομηχανική χρήση της ηλεκτρικής ενέργειας.
- ΔX_{14}, η μεταβολή στην εμπορική και βιομηχανική χρήση της ηλεκτρικής ενέργειας.

Αυτές οι τέσσερις μεταβλητές, συνδυαζόμενες με την ΔX_4 και την ΔX_5, χρησιμοποιούνται για να παράξουν τέσσερις νέες εξισώσεις παλινδρόμησης (βλέπε πίνακα 8-12).

Η στατιστική ανάλυση υπέδειξε, πως η εξίσωση D στον πίνακα 8-

12 είναι η καλύτερη δυνατή. Σε σύγκριση με την προηγούμενη εξίσωση, η οποία περιλάμβανε τις ανεξάρτητες μεταβλητές ΔX_4, ΔX_5, ΔX_{10}, η εξίσωση Α είναι η μόνη, που φανερώνει μια αλλοίωση στη σημαντικότητα. Το αποτέλεσμα αυτό επιβεβαιώνει την ιδέα του Young, πως η χρήση της οικιακής και της βιομηχανικής ηλεκτρικής ενέργειας είναι πιο αξιόπιστες για την πρόβλεψη του προσωπικού εισοδήματος από την συνολική κατανάλωση ηλεκτρικής ενέργειας (η οποία περιλαμβάνει και τη χρήση της οικιακής ηλεκτρικής ενέργειας).

Έτσι, η εξίσωση D επιλέχτηκε ως η τελική εξίσωση παλινδρόμησης και τα αποτελέσματα φαίνονται παρακάτω:

$$\Delta Y = -1{,}86 + 17{,}10\,\Delta X_4 + 23{,}01\,\Delta X_5 + 0{,}007\Delta X_{14}$$

(4,07) (5,61) (0,002)

n = 15 F = 26,26

DW = 1,77 R^2(adj) = 0,835

Οι αριθμοί στις παρενθέσεις κάτω από τους συντελεστές παλινδρόμησης είναι τα τυπικά σφάλματα των εκτιμώμενων συντελεστών. Οι t τιμές των συντελεστών είναι 4,20, 4,10 και 3,50 για τις μεταβλητές ΔX_4, ΔX_5, ΔX_{14} αντίστοιχα. Το R^2(adj) φανερώνει, πως περίπου το 84% της μεταβλητότητας της διαφοράς στο προσωπικό εισόδημα ερμηνεύεται από τις τρεις ανεξάρτητες μεταβλητές. Το Durbin Watson στατιστικό δείχνει πως η αυτοσυσχέτιση δεν αποτελεί πρόβλημα. Επιπλέον, ο πίνακας συσχέτισης μεταβλητών (πίνακας 8-13) καταδεικνύει μια χαμηλή αλληλεξάρτηση μεταξύ των τριών ανεξάρτητων μεταβλητών.

Για την κατασκευή του δείκτη, οι ανεξάρτητες μεταβλητές στην

τελική εξίσωση παλινδρόμησης χρησιμοποιούνται ως συστατικά του. Τα βάρη των συστατικών του μπορούν να καθοριστούν από τους συντελεστές παλινδρόμησης.

	ΔX_4	**ΔX_5**	**ΔX_{14}**
ΔX_4	1		
ΔX_5	0,45	1	
ΔX_{14}	0,11	0,12	1

Πίνακας 8-13 Πίνακας συσχέτισης μεταβλητών

(Υπενθυμίζουμε, ότι οι συντελεστές παλινδρόμησης αντιπροσωπεύουν τη συνολική μεταβολή στην εξαρτημένη μεταβλητή για την αύξηση μιας μονάδας της ανεξάρτητης μεταβλητής). Ωστόσο, επειδή οι μεταβλητές στην εξίσωση παλινδρόμησης δεν έχουν τις ίδιες μονάδες μέτρησης (για παράδειγμα, το ΔΥ μετριέται σε χιλιάδες δολλάρια και το ΔX_{14} σε χιλιάδες κιλοβατώρες), οι συντελεστές παλινδρόμησης πρέπει να μετατραπούν σε ίδιες μονάδες. Αυτή η μετατροπή επιτυγχάνεται με τον υπολογισμό των Β συντελεστών:

$$B = b\ (s_x/s_y) \qquad \textbf{(8.14)}$$

όπου:

- b = ο συντελεστής παλινδρόμησης της ανεξάρτητης μεταβλητής.
- s_x = η τυπική απόκλιση της ανεξάρτητης μεταβλητής.
- s_y = η τυπική απόκλιση της εξαρτημένης μεταβλητής.

Οι τιμές αυτών των στατιστικών προκύπτουν από την επίλυση της παλινδρόμησης. Ως εκ τούτου, οι τυπικοί συντελεστές των τριών ανεξάρτητων μεταβλητών είναι:

$$B_4 = 0,4959$$

$$B_5 = 0,4833$$

$$B_{14} = 0,3019$$

$$\text{Total} = 1,2811$$

Επειδή το άθροισμα των βαρών σε έναν δείκτη πρέπει να είναι 100%, οι τυπικοί συντελεστές κανονικοποιούνται, όπως φαίνεται στον πίνακα 8-14.

Μετά τον προσδιορισμό των συστατικών και των αντίστοιχων σταθμισμένων συντελεστών τους, τα ακόλουθα βήματα δίνουν τον δείκτη:

1. Υπολογισμός της ποσοστιαίας μεταβολής κάθε συστατικού μετά τη βασική περίοδο.
2. Πολλαπλασιασμός της ποσοστιαίας μεταβολής με το κατάλληλο βάρος.
3. Άθροισμα των σταθμισμένων ποσοστιαίων μεταβολών που προκύπτουν από το βήμα 2.

Συστατικό	**Βάρος**
ΔX_4	0,4959 / 1,2811 = 0,3871
ΔX_5	0,4833 / 1,2811 = 0,3772
ΔX_{14}	0,3019 / 1,2811 = 0,2357
	Total = 1,000

Πίνακας 8-14 Τυπικοί συντελεστές

Ο ολοκληρωμένος δείκτης δραστηριότητας της κοινότητας Spokane, για την δεδομένη χρονική περίοδο, συγκρίνεται με τον δείκτη του Α.Ε.Π. των Ηνωμένων Πολιτειών σε σταθερή τιμή δολαρίου

(1967=100), όπως φαίνεται στο διάγραμμα 8-12.

ΕΡΩΤΗΣΕΙΣ

1. **Γιατί ο Young επέλεξε να επιλύσει πρώτα το πρόβλημα της αυτοσυσχέτισης;**

Ο Young επιλέγει να επιλύσει πρώτα το πρόβλημα της αυτοσυσχέτισης, το οποίο εμφανίζεται, όταν παραβιάζεται η υπόθεση της μη συσχέτισης των καταλοίπων. Οδηγείται σε αυτήν την απόφαση, καθώς την θεωρεί σημαντικότερη για την μετέπειτα ορθή έκβαση των προβλέψεων του. Οι συνήθεις ερμηνείες των αποτελεσμάτων μίας ανάλυσης παλινδρόμησης βασίζονται σημαντικά στην υπόθεση ανεξαρτησίας. Άλλωστε, με δεδομένα χρονοσειρών, η υπόθεση ανεξαρτησίας ισχύει σπάνια. Μια από τις βασικότερες συνέπειες αυτής είναι πως η μέθοδος των ελαχίστων τετραγώνων δεν δίνει πλέον τους πιο αποτελεσματικούς εκτιμητές και προβλέψεις. Υπάρχουν άλλοι αμερόληπτοι εκτιμητές με μικρότερες διακυμάνσεις. Το αποτέλεσμα είναι να οδηγούμαστε σε λάθος συμπεράσματα. Η αυτοσυσχέτιση μπορεί να συμβεί επειδή η επίδραση της μεταβλητής πρόβλεψης στην απάντηση κατανέμεται στο χρόνο. Η απασχόληση στον κατασκευαστικό τομέα εξαρτάται από την εποχή, από την οικοδομική δραστηριότητα στη συγκεκριμένη περιοχή κτλ.

2. **Μήπως θα ήταν καλύτερο να εξαλειφθεί πρώτα η πολυσυγγραμικότητα και στη συνέχεια να αντιμετωπιστεί η αυτοσυσχέτιση;**

Οι συνέπειες, που έχει η ύπαρξη πολυσυγραμμικότητας, είναι εμφανώς μικρότερες από αυτές τις αυτοσυσχέτισης. Οπότε η κίνηση του Young να επιλύσει πρώτα αυτή αποδεικνύεται σωστή. Η ύπαρξη πολυσυγγραμικότητας διατηρεί την αμεροληψία, αποτελεσματικότητα

και συνέπεια της πρόβλεψης. Ταυτόχρονα, όλοι οι στατιστικοί έλεγχοι ισχύουν. Το κυριότερο όμως είναι, πως οι προβλέψεις δεν επηρεάζονται από την ύπαρξη υψηλής πολυσυγραμμικότητας.

3. **Πως το μικρό μέγεθος του δείγματος επηρεάζει την ανάλυση;**

Το μικρό μέγεθος του δείγματος μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Είναι άλλωστε γνωστό, πως τα σφάλμα εκτίμησης εξαρτάται από το μέγεθος του δείγματος και μάλιστα αντιστρόφως ανάλογα. Όσο μικρότερο είναι το μέγεθος του δείγματος, τόσο μεγαλύτερη είναι η πιθανότητα τυπικού σφάλματος (μικρότερη ακρίβεια). Όπως γνωρίζουμε, όσο μεγαλύτερη είναι η μεταβλητότητα της εξεταζόμενης μεταβλητής (εδώ του επιτοκίου), τόσο μεγαλύτερο είναι το δείγμα, που χρειαζόμαστε για τη δημιουργία προβλέψεων. Επειδή λοιπόν, ο συνολικός ρυθμός επιτοκίου μεταβάλλεται σημαντικά, απαιτείται και η αντίστοιχη ύπαρξη σημαντικού δείγματος δεδομένων (ανεξάρτητων μεταβλητών).

4. **Θα μπορούσε η παλινδρόμηση των πρώτων διαφορών να έχει καλύτερα αποτελέσματα από την αρχική παλινδρόμηση;**

Είδαμε πως για την εξάλειψη της αυτοσυσχέτισης, δύο προσεγγίσεις μπορούν να ακολουθηθούν. Η μία αφορά την προσθήκη στη συνάρτηση παλινδρόμησης μίας μεταβλητής που έχει παραληφθεί και η οποία ερμηνεύει στην απάντηση τη σχέση από τη μία περίοδο στην επόμενη. Μία άλλη προσέγγιση για την εξάλειψη της αυτοσυσχέτισης περιλαμβάνει τη γενική αντίληψη της μεθόδου των διαφορών. Φυσικά και θα μπορούσε η παλινδρόμηση των πρώτων διαφορών να έχει καλύτερα αποτελέσματα από την αρχική παλινδρόμηση. Όπως άλλωστε παρατηρούμε και στη συγκεκριμένη μελέτη, η παλινδρόμηση των πρώτων διαφορών επιτυγχάνει μια τιμή DW τέτοια ώστε να δηλώνει μια

χαμηλή αυτοσυσχέτιση (ας θυμηθούμε πως με την αρχική παλινδρόμηση είχαμε θετικά συσχετιζόμενες τιμές της εξαρτημένης μεταβλητής). Δηλαδή, η μέθοδος των πρώτων διαφορών εξάλειψε τη σειριακή συσχέτιση. Επομένως, χρησιμοποιείται η προσαρμοσμένη εξίσωση για πρόβλεψη.

5. **Υπάρχει κάποια δυνατότητα για τη χρήση των χρονοκαθυστερημένων δεδομένων;**

Η ερώτηση αυτή αναφέρεται στην περίπτωση των αυτοπαλινδρομικών δεδομένων. Ένας τρόπος για να λυθεί το πρόβλημα της σειριακής συσχέτισης είναι να μοντελοποιηθεί η συσχέτιση απευθείας σε διαφορετικές χρονικές περιόδους. Τα δεδομένα αυτά λοιπόν μπορούν να χρησιμοποιηθούν για τη δημιουργία προβλέψεων και για την εξαγωγή αναλόγων συμπερασμάτων με τις άλλες μεθόδους πρόβλεψης. Η εξαρτημένη μεταβλητή χρονοκαθυστερεί κατά ένα χρονικό διάστημα και αποτελεί μια μεταβλητή πρόβλεψης.

6. **Ποια συμπεράσματα προκύπτουν από τη σύγκριση του δείκτη επιχειρησιακής δραστηριότητας της κοινότητας Spokane και του δείκτη Α.Ε.Π.;**

Παρατηρούμε πως μέχρι το 1969, οι δύο δείκτες ακολουθούν μια σχεδόν παράλληλη πορεία. Στη συνέχεια και μέχρι το 1972, παρατηρούμε μια διαφοροποίη*ση. Ο δείκτης επιχειρησιακής δραστηριότητας της κοινότητας Spokane μάλλον υπερεκτιμά τις πραγματικές τιμές, που δίνει ο δείκτης Α.Ε.Π., καθώς οι προβλέψεις, που μας δίνει είναι μάλλον διαφοροποιημένες. Οι αποκλίσεις βέβαια αυτές ισχυροποιούνται το έτος 1975.*

ΚΕΦΑΛΑΙΟ 9 (ΜΕΘΟΔΟΛΟΓΙΑ BOX – JENKINS (ARIMA))

Μέχρι τώρα έχουν συζητηθεί αρκετές προσεγγίσεις για ανάλυση και πρόβλεψη χρονοσειρών. Στο κεφάλαιο 4 είδαμε μεθόδους εξομάλυνσης και κινητούς μέσους. Στο κεφάλαιο 5 αναλύσαμε χρονοσειρές στα συστατικά τους. Αυτό το κεφάλαιο εισάγει μία κατηγορία μοντέλων, που μπορούν να παράγουν ακριβείς προβλέψεις, βασιζόμενα σε μία περιγραφή ιστορικών μοτίβων στα δεδομένα. Παρουσιάζονται οι μέθοδοι αυτοπαλινδρόμησης – κινητού μέσου (ARMA) στάσιμων χρονοσειρών. Αυτές οι μέθοδοι πρόβλεψης περιλαμβάνουν γενικευμένες μεθόδους **αναπροσαρμοζόμενου φιλτραρίσματος και τεχνικές Box – Jenkins**.

Τα ARIMA (autoregressive integrated moving average) μοντέλα είναι μία κατηγορία γραμμικών μοντέλων, που είναι ικανά να αναπαραστήσουν στάσιμες και μη στάσιμες χρονοσειρές. Ας θυμηθούμε, ότι οι στάσιμες διαδικασίες κινούνται γύρω από ένα σταθερό επίπεδο και ότι οι μη στάσιμες διαδικασίες δείχνουν αύξηση ή μείωση των τιμών στο χρόνο. Τα αυτοπαλινδρομικά και κινητού μέσου (ARMA) μοντέλα είναι στην πραγματικότητα ένα υποσύνολο των ARIMA μοντέλων, που είναι χρήσιμα για μοντελοποίηση στάσιμων χρονοσειρών.

Τα ARIMA μοντέλα δεν περιλαμβάνουν ανεξάρτητες μεταβλητές στην κατασκευή τους. Αντί για αυτό κάνουν χρήση της διαθέσιμης πληροφορίας, που κρύβουν οι χρονοσειρές, για τη δημιουργία προβλέψεων. Για παράδειγμα, ένα ARIMA μοντέλο για μηνιαίες πωλήσεις θα χρησιμοποιούσε το ιστορικό μοτίβο των πωλήσεων για να παράγει προβλέψεις για τις πωλήσεις του επόμενου μήνα.

Τα ARIMA μοντέλα βασίζονται πολύ σε μοτίβα αυτοσυσχέτισης

των δεδομένων. Η μεθοδολογία για αναγνώριση, προσαρμογή και έλεγχο καταλοίπων των ARIMA μοντέλων εξελίχθηκε ιδιαίτερα από τη δουλειά των G.E.P. Box και G.M. Jenkins. Για αυτό το λόγο, η μοντελοποίηση και πρόβλεψη ARIMA συχνά αναφέρεται ως Box-Jenkins μεθοδολογία.

Τα ARIMA μοντέλα συνδυάζουν τις ιδιότητες τριών διαφορετικών υπο- μοντέλων (υπο- συστημάτων):

- Αυτοπαλινδρόμησης (autoregression).
- Ολοκλήρωσης (integration).
- Εξομάλυνσης με κινητό μέσο (moving average).

Ένα ARIMA μοντέλο μπορεί να διαθέτει και τα τρία υποσυστήματα.

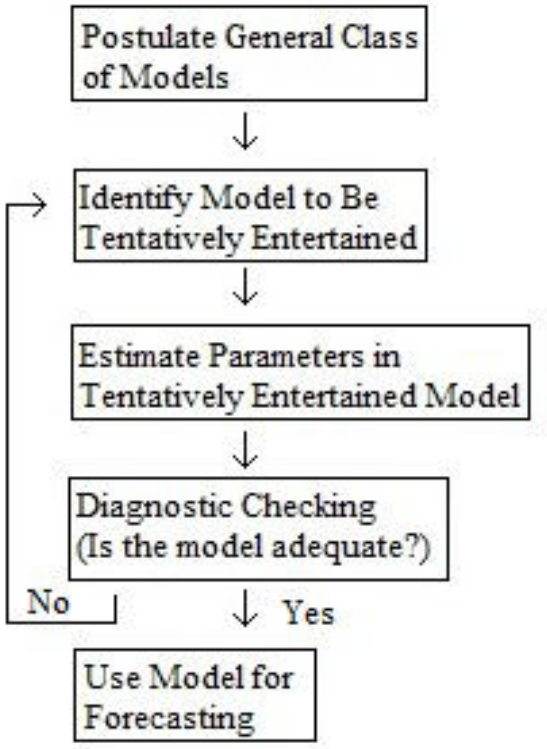
9.1 ΜΕΘΟΔΟΛΟΓΙΑ BOX – JENKINS

Η μεθοδολογία Box-Jenkins για πρόβλεψη είναι διαφορετική από τις περισσότερες μεθόδους, επειδή δεν προϋποθέτει κάποιο συγκεκριμένο μοτίβο στα ιστορικά δεδομένα των χρονοσειρών, που προβλέπει. Χρησιμοποιεί μία επαναληπτική προσέγγιση ξεκινώντας με την αναγνώριση ενός πιθανού μοντέλου, από μία γενική κατηγορία μοντέλων.

Αυτό το επιλεγμένο μοντέλο τότε ελέγχεται στα ιστορικά δεδομένα, για να βρεθεί, αν περιγράφει επαρκώς τις σειρές. Το μοντέλο προσαρμόζεται καλά αν τα κατάλοιπα μετά την προσαρμογή είναι γενικά μικρά και τυχαία κατανεμημένα και δεν περιέχουν χρήσιμες πληροφορίες. Αν το καθορισμένο μοντέλο δεν είναι ικανοποιητικό, η διαδικασία επαναλαμβάνεται χρησιμοποιώντας ένα νέο μοντέλο, που σχεδιάζεται για να βελτιώσει το πρωτότυπο, ή αναπροσαρμόζεται το

πρωτότυπο.

Αυτή η επαναληπτική διαδικασία συνεχίζεται μέχρι να βρεθεί ένα ικανοποιητικό μοντέλο. Από εκείνο το σημείο και μετά, το μοντέλο μπορεί να χρησιμοποιηθεί για πρόβλεψη. Το παρακάτω Γράφημα αναπαριστά τη στρατηγική Box-Jenkins (Box-Jenkins model-building strategy).



Η αρχική επιλογή ενός ARIMA μοντέλου βασίζεται στην εξέταση ενός διαγράμματος των χρονοσειρών σαν συνάρτηση του χρόνου και στην εξέταση της αυτοσυσχέτισής τους για μερικές χρονικές καθυστερήσεις (lags). Συγκεκριμένα, το μοτίβο των αυτοσυσχετίσεων δείγματος, το οποίο υπολογίζεται από τις χρονοσειρές, συνδυάζεται **με το γνωστό μοτίβο αυτοσυσχέτισης**, που αντιστοιχεί σε ένα συγκεκριμένο ARIMA μοντέλο. Αυτός ο συνδυασμός γίνεται και για τις αυτοσυσχετίσεις και για τις **μερικές αυτοσυσχετίσεις**. Θεωρητικοί

συντελεστές αυτοσυσχέτισης για μερικά από τα κοινά ARIMA μοντέλα φαίνονται στα παρακάτω Γραφήματα. Η σημασία των μοτίβων, που φαίνεται σε αυτά τα γραφήματα, θα γίνει φανερή αργότερα.

Η Box-Jenkins μεθοδολογία αναφέρεται σε ένα σύνολο διαδικασιών για αναγνώριση, προσαρμογή και έλεγχο ARIMA μοντέλων με δεδομένα χρονοσειρών. Οι προβλέψεις έπονται απευθείας από τη μορφή του προσαρμοσμένου μοντέλου.

Στην επιλογή του μοντέλου, να έχουμε υπόψη, ότι οι αυτοσυσχετίσεις, οι οποίες υπολογίζονται από τα δεδομένα, πρώτον δεν ταιριάζουν ακριβώς με κάποιο μοτίβο από το σύνολο των θεωρητικών αυτοσυσχετίσεων ARIMA μοντέλων και δεύτερον οι αυτοσυσχετίσεις υπόκεινται στη μεταβολή του δείγματος.

Ωστόσο, θα πρέπει να είμαστε σε θέση, να ταιριάζουμε επαρκώς τα περισσότερα δεδομένα χρονοσειρών με ένα ARIMA μοντέλο. Αν η αρχική επιλογή δεν είναι ακριβώς σωστή, θα εμφανιστούν ανεπάρκειες σε μία ανάλυση των καταλοίπων (έλεγχος μοντέλου) και το αρχικό μοντέλο θα τροποποιηθεί. Καθώς αποκτάται εμπειρία, το χτίσιμο του επαναληπτικού μοντέλου θα γίνεται ευκολότερο.

9.1.1 ΜΕΡΙΚΗ ΑΥΤΟΣΥΣΧΕΤΙΣΗ

Όλες οι αυτοπαλίνδρομες διαδικασίες έχουν συναρτήσεις αυτοσυσχετίσεως, οι οποίες βαίνουν φθίνουσες καθώς αυξάνει το μήκος της υστέρησης s, με συνέπεια να είναι πολλές φορές δύσκολο να καθοριστεί η τάξη του υποδείγματος ARIMA, το οποίο περιγράφει τη χρονοσειρά με βάση τη συνάρτηση αυτοσυσχέτισης. Ως ένα πρόσθετο κριτήριο για το σκοπό αυτό χρησιμοποιείται **η συνάρτηση μερικής αυτοσυσχέτισης** (partial autocorrelation function).

Η μερική αυτοσυσχέτιση ανάμεσα στα Y_t και Y_{t-s} αναφέρεται στη

συσχέτιση ανάμεσα στην Y_t και στην Y_{t-s}, όταν έχουν αφαιρεθεί οι γραμμικές επιδράσεις όλων των ενδιάμεσων μεταβλητών Y_{t-1}, Y_{t-2}, ..., $Y_{t-(s-1)}$. (Εδώ εννοείται, ότι οι ενδιάμεσες μεταβλητές έχουν εισέλθει στην εξίσωση, οπότε έχει αφαιρεθεί η επίδρασή τους). Αν παραστήσουμε με ρ_{ss} το συντελεστή μερικής αυτοσυσχέτισης s τάξης, δηλαδή το συντελεστή αυτοσυσχέτισης ανάμεσα στην Y_t και στην Y_{t-s}, για s=1, 2, …, τότε, σύμφωνα με τον παραπάνω ορισμό, το ρ_{ss} είναι ο **μερικός συντελεστής παλινδρόμησης της μεταβλητής y_{t-s}** στο παρακάτω υπόδειγμα (χωρίς τον σταθερό όρο):

$$y_t = \rho_{1s} y_{t-1} + \rho_{2s} y_{t-2} + \rho_{3s} y_{t-3} + ... + \rho_{ss} y_{t-s} + \varepsilon_t$$

όπου:

$$y_t = Y_t - \mu$$

Για παράδειγμα ο μερικός συντελεστής αυτοσυσχέτισης πρώτης τάξης ρ_{11} προκύπτει από την παλινδρόμηση:

$$y_t = \rho_{11} y_{t-1} + \varepsilon_t .$$

Είναι προφανές, ότι ο μερικός συντελεστής αυτοσυσχέτισης ρ_{11} συμπίπτει με τον συντελεστή αυτοσυσχέτισης r_1. **Δηλαδή $r_1 = \rho_{11}$.**

Ο μερικός συντελεστής αυτοσυσχέτισης δεύτερης τάξης ρ_{22} προκύπτει από την παλινδρόμηση:

$$y_t = \rho_{12} y_{t-1} + \rho_{22} y_{t-2} + \varepsilon_t .$$

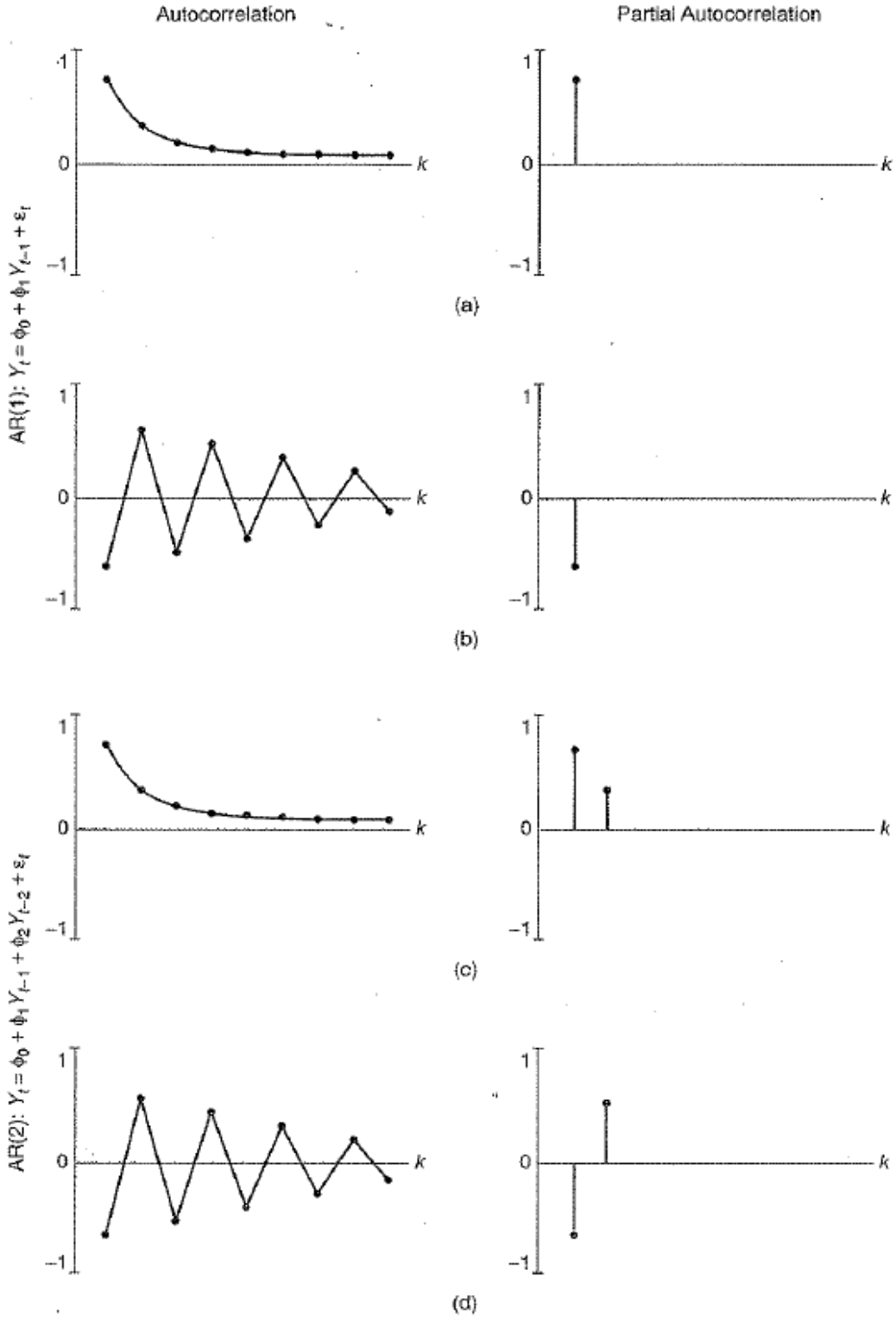
Ο συντελεστής ρ_{12} δεν είναι ίδιος με τον ρ_{11}. Ο μερικός συντελεστής αυτοσυσχέτισης τρίτης τάξης ρ_{33} προκύπτει από την παλινδρόμηση:

$$y_t = \rho_{13} y_{t-1} + \rho_{23} y_{t-2} + \rho_{33} y_{t-3} + \varepsilon_t .$$

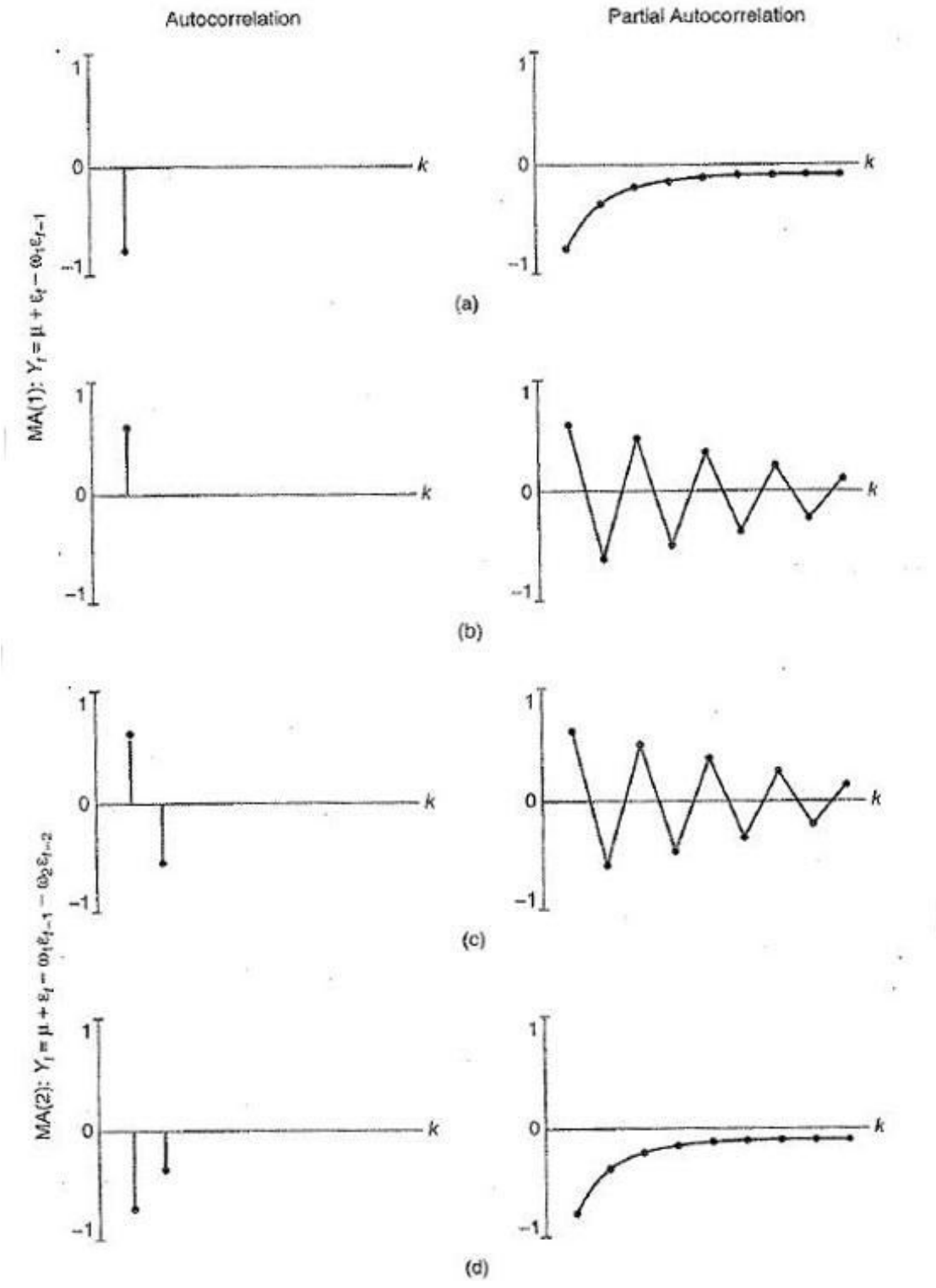
Ο συντελεστής ρ_{13} δεν είναι ίδιος με τον ρ_{12} και ο συντελεστής ρ_{23} δεν είναι ίδιος με τον ρ_{22}.

Με άλλα λόγια, οι μερικοί συντελεστές αυτοσυσχέτισης προκύπτουν από διαδοχικές παλινδρομήσεις ανάμεσα στην y_t και y_{t-s}, αρχίζοντας από το y_{t-1} και προσθέτοντας κάθε φορά και μία υστέρηση. Οι μερικοί συντελεστές αυτοσυσχέτισης μπορούν επίσης να γραφούν και ως συναρτήσεις των συντελεστών αυτοσυσχέτισης.

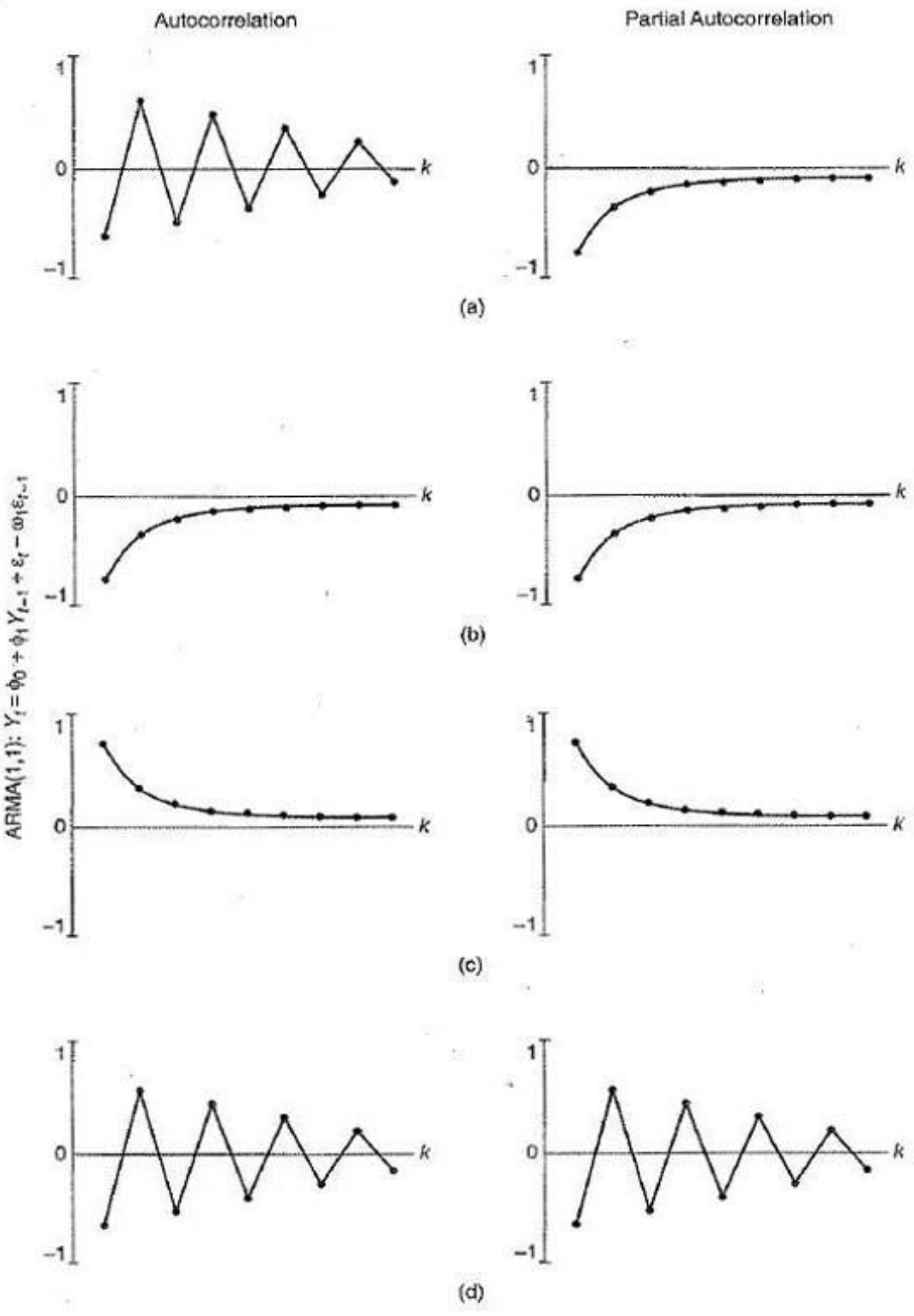
Στα παρακάτω γραφήματα δίδονται τα μοτίβα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης για κάποια AR, MA και ARMA μοντέλα. Μέσω δηλαδή των συναρτήσεων αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, μπορούμε να ταυτοποιήσουμε το ARMA μοντέλο, το οποίο θα χρησιμοποιήσουμε για προσαρμογή και πρόβλεψη.



Γράφημα 1: Συντελεστές αυτοσυσχέτισης και μερικής αυτοσυσχέτισης των AR(1) και AR(2) μοντέλων



Γράφημα 2: Συντελεστές αυτοσυσχέτισης και μερικής αυτοσυσχέτισης των MA(1) και MA(2) μοντέλων



Γράφημα 3: Συντελεστές αυτοσυσχέτισης και μερικής αυτοσυσχέτισης ενός mixed ARMA(1,1) μοντέλου

9.1.2 ΑΥΤΟΠΑΛΙΝΔΡΟΜΙΚΑ ΜΟΝΤΕΛΑ

Ένα πρώτης τάξης αυτοπαλινδρομικό μοντέλο (autoregressive) εισήχθη στο Κεφάλαιο 8. Ένα αυτοπαλινδρομικό μοντέλο p τάξης – AR(p) – παίρνει τη μορφή:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \varepsilon_t$$

όπου,

- Y_t = Η εξαρτημένη μεταβλητή στον χρόνο t.
- Y_{t-1}, Y_{t-2},..., Y_{t-p} = Η εξαρτημένη μεταβλητή στους χρόνους για lag t-1, t-2, …, t-p.
- φ_0, φ_1, φ_2, ..., φ_p, = οι συντελεστές, που πρέπει να εκτιμηθούν.
- ε_t = ο όρος σφάλματος στο χρόνο t, ο οποίος αναπαριστά τις επιδράσεις των μεταβλητών, που δεν ερμηνεύονται από το μοντέλο. Οι υποθέσεις σχετικά με τον όρο σφάλματος είναι ίδιες με εκείνες για το τυπικό μοντέλο παλινδρόμησης.

Η Εξίσωση του αυτοπαλίνδρομου μοντέλου έχει την εμφάνιση ενός μοντέλου παλινδρόμησης με καθυστερημένες (lagged) τιμές της εξαρτημένης μεταβλητής στις θέσεις της ανεξάρτητης μεταβλητής, ως εκ τούτου και το όνομα αυτοπαλινδρομικό μοντέλο. Τα αυτοπαλινδρομικά μοντέλα είναι κατάλληλα και για **στάσιμες χρονοσειρές** και ο συντελεστής φ_0 σχετίζεται με το σταθερό επίπεδο των σειρών. Αν τα δεδομένα διαφέρουν γύρω από το **μηδέν** ή εκφράζονται σαν αποκλίσεις από το μέσο $Y_t - \bar{Y}$, ο συντελεστής φ_0 δεν απαιτείται.

Οι εξισώσεις ενός AR μοντέλου τάξης 1, ή AR(1) μοντέλου και ενός AR μοντέλου τάξης 2, ή AR(2) μοντέλου, φαίνονται στο Γράφημα 1. Τα Γραφήματα 1 (a) και (b) απεικονίζουν τη συμπεριφορά των θεωρητικών συναρτήσεων αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, αντίστοιχα, για ένα AR(1) μοντέλου. Παρατηρήστε, πόσο διαφορετικά

συμπεριφέρονται οι συναρτήσεις αυτοσυσχέτισης και μερικής αυτοσυσχέτισης. Οι συντελεστές αυτοσυσχέτισης εξασθενούν στο μηδέν σταδιακά, ενώ οι συντελεστές μερικής αυτοσυσχέτισης πέφτουν στο μηδέν μετά την πρώτη χρονική καθυστέρηση. Τα Γραφήματα 1 (c) και (d) δείχνουν τις αυτοσυσχετίσεις και τις μερικές αυτοσυσχετίσεις αντίστοιχα για ένα AR(2) μοντέλο. Ξανά, οι συντελεστές αυτοσυσχέτισης εξασθενούν στο μηδέν, ενώ οι συντελεστές μερικής αυτοσυσχέτισης πέφτουν στο μηδέν μετά τη δεύτερη χρονική καθυστέρηση. Αυτός ο τύπος μοτίβου θα ισχύει γενικά για κάθε AR(p) μοντέλο. Πρέπει να θυμόμαστε πάντα ότι οι συναρτήσεις αυτοσυσχέτισης δείγματος θα διαφέρουν από αυτές τις θεωρητικές συναρτήσεις εξαιτίας της μεταβολής δείγματος.

Παράδειγμα

Η πρόβλεψη με ένα AR(2) μοντέλο παρουσιάζεται χρησιμοποιώντας το σύνολο δεδομένων, που αποτελείται από 75 αναγνώσεις μιας διαδικασίας (process readings), που φαίνονται στον παρακάτω Πίνακα. Μόνο οι πέντε τελευταίες παρατηρήσεις στον Πίνακα θα χρησιμοποιηθούν για να κατασκευαστούν οι προβλέψεις.

Period	**Time** **t**	**Values** $\mathbf{Y_t}$	**Forecasts** $\hat{Y}_t$	**Residuals** $\mathbf{e_t}$
t-5	71	90	-	-
t-4	72	78	-	-
t-3	73	87	74	13
t-2	74	99	69,1	29,9
t-1	75	72	62,7	9,3
t	76		77,2	

Ένα AR(2) μοντέλο επιλέγεται για τη διαδικασία δεδομένων, και το Minitab υπολογίζει της εκτιμήσεις ελαχίστων τετραγώνων: $\hat{\varphi}_0 = 115{,}2, \quad \hat{\varphi}_1 = -0{,}535, \quad \hat{\varphi}_2 = 0{,}0055.$ Υποθέστε ότι απαιτείται στο

χρόνο t-1=75 μία πρόβλεψη της παρατήρησης για την επόμενη περίοδο, t=76. Δεδομένου ότι η καλύτερη πρόβλεψη του όρου σφάλματος είναι η μέση τιμή μηδέν, η πρόβλεψη για την περίοδο t=76 είναι:

$$\hat{Y}_t = \hat{\varphi}_0 + \hat{\varphi}_1 Y_{t-1} + \hat{\varphi}_2 Y_{t-2}$$

$$\hat{Y}_{76} = 115{,}2 - 0{,}535\, Y_{75} + 0{,}0055\, Y_{74}$$

$$\hat{Y}_{76} = 115{,}2 - 0{,}535(72) + 0{,}0055(99) = 77{,}2$$

Για αυτοπαλινδρομικά μοντέλα, οι προβλέψεις εξαρτώνται από παρατηρούμενες τιμές σε προηγούμενες χρονικές περιόδους. Για AR(2) μοντέλα, οι προβλέψεις της επόμενης τιμής εξαρτώνται από τις παρατηρούμενες τιμές στις δύο προηγούμενες χρονικές περιόδους. Για AR(3) μοντέλα, οι προβλέψεις για την επόμενη τιμή εξαρτώνται από τις παρατηρούμενες τιμές στις τρεις προηγούμενες χρονικές περιόδους και ούτω καθεξής.

9.1.3 ΜΟΝΤΕΛΑ ΚΙΝΗΤΟΥ ΜΕΣΟΥ

Ένα μοντέλο κινούμενου μέσου (moving average) q τάξης – MA(q) – παίρνει τη μορφή:

$$Y_t = \mu + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \ldots - \omega_q \varepsilon_{t-q}$$

όπου:

- Y_t = Η εξαρτημένη μεταβλητή στον χρόνο t.
- μ = Ο σταθερός μέσος της διαδικασίας.
- $\omega_0, \omega_1, \omega_2, \ldots, \omega_p$, = Οι συντελεστές που πρέπει να εκτιμηθούν.
- ε_t = Ο όρος σφάλματος στο χρόνο t, ο οποίος αναπαριστά τις επιδράσεις των μεταβλητών, που δεν ερμηνεύονται από το μοντέλο. Οι υποθέσεις σχετικά με τον όρο σφάλματος είναι ίδιες με εκείνες για το τυπικό μοντέλο παλινδρόμησης.

- ε_0, ε_1, ε_2, ..., $\varepsilon_{t=q}$ = Σφάλματα από προηγούμενες χρονικές περιόδους.

Η παραπάνω εξίσωση είναι παρόμοια με την Εξίσωση του AR μοντέλου, εκτός από το ότι η εξαρτημένη μεταβλητή Y_t εξαρτάται από τις προηγούμενες τιμές των σφαλμάτων, παρά από την ίδια τη μεταβλητή. Τα μοντέλα κινούμενου μέσου (MA) παρέχουν προβλέψεις βασιζόμενα σε ένα γραμμικό συνδυασμό ενός πεπερασμένου αριθμού παρελθοντικών σφαλμάτων, ενώ τα αυτοπαλινδρομικά μοντέλα (AR) προβλέπουν το Y_t σαν μία γραμμική συνάρτηση ενός πεπερασμένου αριθμού παρελθοντικών τιμών του Y_t. Επίσης είναι κατάλληλα για **στάσιμες χρονοσειρές**.

Ο όρος κινούμενος μέσος για το μοντέλο στην προαναφερόμενη εξίσωση είναι ιστορικός και δεν θα πρέπει να μπερδεύεται με τις διαδικασίες κινούμενου μέσου, που συζητήθηκαν στο Κεφάλαιο 4. Εδώ, ο όρος κινούμενος μέσος αναφέρεται στο γεγονός, ότι η απόκλιση της απόκρισης από το μέσο, $Y_t - \mu$, είναι ένας γραμμικός συνδυασμός των τωρινών και παρελθοντικών σφαλμάτων και ότι καθώς ο χρόνος προχωράει μπροστά, τα σφάλματα, που περιλαμβάνονται σε αυτό το γραμμικό συνδυασμό, προχωράνε επίσης μπροστά.

$$Y_t - \mu = \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - ... - \omega_q \varepsilon_{t-q}$$

$$Y_{t+1} - \mu = \varepsilon_{t+1} - \omega_1 \varepsilon_t - \omega_2 \varepsilon_{t-1} - ... - \omega_q \varepsilon_{t-q+1}$$

Τα βάρη ω_1, ω_2, …, ω_q δεν έχουν αναγκαστικά άθροισμα 1, πρέπει το άθροισμα να είναι όμως μικρότερο του απείρου και μπορεί να είναι θετικά ή αρνητικά αν και πριν από κάθε ένα από αυτά μπορεί να υπάρχει ένα αρνητικό πρόσημο στον προσδιορισμό του μοντέλου. Το MA μοντέλο αποτελεί και το θεώρημα διαχωρισμού του Wold.

Το Γράφημα 2 δείχνει τις εξισώσεις ενός μοντέλου κινούμενου

μέσου τάξης 1, ή MA(1) μοντέλο, και ενός μοντέλου κινούμενου μέσου τάξης 2, ή MA(2) μοντέλο. Οι όροι σφάλματος τάξης q μπορούν να αθροιστούν, ώστε να πάρουμε ένα MA(q) μοντέλο, όπου q είναι ο αριθμός των περασμένων όρων σφαλμάτων, που συμπεριλαμβάνονται στην πρόβλεψη της επόμενης παρατήρησης. Τα Γραφήματα 2 (a) και (b) της σελίδας 6 επίσης απεικονίζουν τη συμπεριφορά των θεωρητικών συντελεστών αυτοσυσχέτισης και μερικής αυτοσυσχέτισης για το MA(1) μοντέλο. Να σημειωθεί, ότι οι συναρτήσεις αυτοσυσχέτισης και μερικής αυτοσυσχέτισης των AR και MA μοντέλων συμπεριφέρονται πολύ διαφορετικά. Αυτό είναι ευτυχές επειδή τα AR και MA μοντέλα μπορούν, τότε να διαχωριστούν από τα διαφορετικά τους μοτίβα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης.

Οι συντελεστές αυτοσυσχέτισης για το MA(1) μοντέλο τείνουν στο μηδέν μετά την πρώτη χρονική υστέρηση, ενώ οι συντελεστές μερικής αυτοσυσχέτισης εξασθενούν στο μηδέν σταδιακά. Επίσης, οι συντελεστές αυτοσυσχέτισης για το MA(2) μοντέλο είναι μηδέν μετά τη δεύτερη χρονική καθυστέρηση, ενώ στη μερική αυτοσυσχέτιση εξασθενούν σταδιακά (βλ. Γραφήματα 2 (c) και (d)). Πρέπει να σημειωθεί ξανά, ότι οι συναρτήσεις αυτοσυσχέτισης δείγματος θα διαφέρουν από αυτές τις θεωρητικές συναρτήσεις εξαιτίας της μεταβολής δείγματος.

Προκειμένου να υπολογίσουμε τους συντελεστές δεν εφαρμόζουμε απλή μέθοδο των ελαχίστων τετραγώνων, όπως στα μοντέλα αυτοπαλινδρόμησης AR(p), διότι η σχέση, που προκύπτει, είναι μη γραμμική ως προς τους συντελεστές. Θεωρούμε για παράδειγμα και για λόγους απλότητας το μοντέλο MA(1) και το γράφουμε στην ακόλουθη μορφή:

$$Y_t - \mu = y_t = \varepsilon_t - \omega_1 \varepsilon_{t-1} \quad \text{ή}$$

$$\varepsilon_t = y_t + \omega_1 \varepsilon_{t-1} \quad \text{ή}$$

$$\varepsilon_t = y_t + \omega_1 (y_{t-1} + \omega_1 \varepsilon_{t-2}) = y_t + \omega_1 y_{t-1} + \omega_1^2 \varepsilon_{t-2} \quad \text{ή} \ldots$$

$$\varepsilon_t = y_t + \omega_1 y_{t-1} + \omega_1^2 y_{t-2} + \omega_1^3 y_{t-3} + \omega_1^4 y_{t-4} \ldots$$

Η σχέση αυτή είναι μη γραμμική ως προς τους συντελεστές και δεν είναι δυνατή η εφαρμογή της μεθόδου των ελαχίστων τετραγώνων. Οι εξισώσεις λύνονται αριθμητικά. Η προς ελαχιστοποίηση συνάρτηση δίδεται από τη σχέση:

$$\sum \varepsilon_t^2 = \sum (y_t + \omega_1 y_{t-1} + \omega_1^2 y_{t-2} + \omega_1^3 y_{t-3} + \omega_1^4 y_{t-4} \ldots)^2$$

Παράδειγμα

Οι 75 αναγνώσεις διαδικασίας, που δίνονται στον παρακάτω Πίνακα, μπορούν να χρησιμοποιηθούν για να παρουσιάσουν προβλέψεις με ένα MA(2) μοντέλο. Η πρόβλεψη παρουσιάζεται χρησιμοποιώντας μόνο τις πέντε τελευταίες παρατηρήσεις και τα κατάλοιπα.

Period	**Time t**	**Values Y_t**	**Forecasts $\hat{Y}_t$**	**Residuals e_t**
t-5	71	90	-	13,9
t-4	72	78	-	8,9
t-3	73	87	75,3	11,7
t-2	74	99	72	27
t-1	75	72	64,3	7,7
t	76		80,6	

Ένα MA(2) μοντέλο επιλέγεται, και το Minitab υπολογίζει τις εκτιμήσεις ελαχίστων τετραγώνων των συντελεστών του μοντέλου, παίρνοντας $\hat{\mu} = 75{,}4$, $\hat{\omega}_1 = 0{,}5667$, $\hat{\omega}_2 = -0{,}3560$. Ας υποθέσουμε, ότι στο χρόνο t-1=75 απαιτείται μία πρόβλεψη της παρατήρησης στην περίοδο t=76. Δεδομένου ότι στο χρόνο t-1 η καλύτερη εικασία για τον όρο σφάλματος στην επόμενη περίοδο είναι η μέση τιμή της, μηδέν, και οι καλύτερες εικασίες των σφαλμάτων στην τωρινή και τις προηγούμενες χρονικές περιόδους είναι τα αντίστοιχα κατάλοιπα, η πρόβλεψη για την

περίοδο t=76 είναι:

$$\hat{Y}_t = \hat{\mu} - \hat{\omega}_1 e_{t-1} - \hat{\omega}_2 e_{t-2}$$

$$\hat{Y}_{76} = 75{,}4 - 0{,}5667\, e_{75} + 0{,}3560\, e_{74}$$

$$\hat{Y}_{76} = 75{,}4 - 0{,}5667\,(7{,}7) + 0{,}3560\,(27) = 80{,}6$$

Εδώ πρέπει να παρατηρήσουμε, ότι τα δύο κατάλοιπα, e_{75} και e_{74} αντικαθίστανται για τα σφάλματα, ε_{75} και ε_{74}, στον υπολογισμό της πρόβλεψης για την περίοδο 76. Για τον υπολογισμό προβλέψεων από τα μοντέλα κινούμενου μέσου τα σφάλματα, που αντιστοιχούν σε χρονικές περιόδους που έχουν ήδη συμβεί, αντικαθίστανται από τα κατάλοιπα για εκείνες τις χρονικές περιόδους. Ο αριθμός των καταλοίπων, που περιλαμβάνονται στην πρόβλεψη της επόμενης παρατήρησης, είναι ίσος με την τάξη του μοντέλου κινούμενου μέσου.

9.1.4 ΑΥΤΟΠΑΛΙΝΔΡΟΜΙΚΑ ΜΟΝΤΕΛΑ ΚΙΝΗΤΟΥ ΜΕΣΟΥ

Ένα μοντέλο με αυτοπαλινδρομικούς όρους μπορεί να συνδυαστεί με ένα μοντέλο, που έχει όρους κινούμενου μέσου και να οδηγήσει σε ένα μεικτό αυτοπαλινδρομικό μοντέλο κινούμενου μέσου (autoregressive-moving average model). Είναι βολικό να γίνει χρήση του συμβολισμού ARMA(p,q), όπου το p είναι η τάξη του αυτοπαλινδρομικού μέρους και q είναι η τάξη του μέρους κινούμενου μέσου, για να αναπαραστήσουμε το μοντέλο. Ένα ARMA(p,q) μοντέλο έχει τη γενική μορφή:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \ldots - \omega_q \varepsilon_{t-q}$$

Τα ARMA(p,q) μοντέλα μπορούν να περιγράψουν μία ευρεία ποικιλία συμπεριφορών για **στάσιμες χρονοσειρές**. Προβλέψεις που παράγονται από ένα ARMA(p,q) μοντέλο, θα εξαρτώνται από τις πρόσφατες και περασμένες τιμές της απόκρισης, Y, καθώς επίσης από τις

πρόσφατες και περασμένες τιμές των σφαλμάτων (καταλοίπων), e.

Το Γράφημα 3 δείχνει την εξίσωση ενός ARMA(1,1) μοντέλου και τις πιθανές συμπεριφορές των θεωρητικών αυτοσυσχετίσεων και μερικών αυτοσυσχετίσεων. Σε αυτή την περίπτωση – και οι δύο φθίνουν – δεν διακόπτονται απότομα.

Προκειμένου να υπολογίσουμε τους συντελεστές δεν εφαρμόζουμε απλή μέθοδο των ελαχίστων τετραγώνων, διότι η σχέση που προκύπτει είναι μη γραμμική ως προς τους συντελεστές. Θεωρούμε για παράδειγμα και για λόγους απλότητας το μοντέλο ARMA(1,1) και το γράφουμε στην ακόλουθη μορφή:

$$y_t = \varphi_1 y_{t-1} + \varepsilon_t - \omega_1 \varepsilon_{t-1} \quad \text{ή}$$

$$\varepsilon_t = y_t - \varphi_1 y_{t-1} + \omega_1 \varepsilon_{t-1} \quad \text{ή}$$

$$\varepsilon_t = y_t - \varphi_1 y_{t-1} + \omega_1 (y_{t-1} - \varphi_1 y_{t-2} + \omega_1 \varepsilon_{t-2}) \quad \text{ή}$$

$$\varepsilon_t = y_t - \varphi_1 y_{t-1} + \omega_1 y_{t-1} - \omega_1 \varphi_1 y_{t-2} + \omega_1^2 \varepsilon_{t-2} = \ldots$$

Η προς ελαχιστοποίηση συνάρτηση είναι:

$$\sum \varepsilon_t^2 = \sum (y_t - \varphi_1 y_{t-1} + \omega_1 y_{t-1} - \omega_1 \varphi_1 y_{t-2} + \omega_1^2 y_{t-2} \ldots)^2,$$

η οποία λύνεται αριθμητικά.

9.1.5 ΟΛΟΚΛΗΡΩΜΕΝΗ ΔΙΑΔΙΚΑΣΙΑ

Οι περισσότερες αν όχι όλες οι οικονομικές χρονοσειρές δεν έχουν τα χαρακτηριστικά στασιμότητας. Μπορούν όμως να μετατραπούν σε στάσιμες παίρνοντας τις πρώτες, τις δεύτερες κ.λ.π διαφορές. Γενικά μας ενδιαφέρει να είναι στάσιμες για να μην έχουμε το φαινόμενο της λανθάνουσας παλινδρόμησης.

Όταν μία σειρά μετατρέπεται σε στάσιμη παίρνοντας τις πρώτες διαφορές, λέμε ότι η σειρά είναι *ολοκληρωμένη πρώτης τάξης (integrated*

first order) και παριστάνεται με I(1). Αν πάρουμε τις δεύτερες διαφορές τότε η σειρά είναι *ολοκληρωμένη δεύτερης τάξης* και παριστάνεται με I(2). Αν πάρουμε d διαφορές, τότε η σειρά είναι *ολοκληρωμένη d τάξης*.

Ο όρος ολοκληρωμένη προέρχεται από τον τρόπο, που μία μη στάσιμη διαδικασία προκύπτει από μία στάσιμη. Δηλαδή αθροίζοντας ή ολοκληρώνοντας d φορές.

9.1.6 ΣΥΝΟΨΗ

Τα μοτίβα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης για αυτοπαλινδρομικές διαδικασίες κινούμενου μέσου μπορούν να συνοψιστούν ως εξής:

ΣΥΜΒΟΛΟ	**ΑΥΤΟΣΥΣΧΕΤΙΣΕΙΣ**	**ΜΕΡΙΚΕΣ ΑΥΤΟΣΥΣΧΕΤΙΣΕΙΣ**
MA(q)	Διακόπτονται μετά την τάξη q της διαδικασίας	Φθίνουν
AR(p)	Φθίνουν	Διακόπτονται μετά την τάξη p της διαδικασίας
ARMA(p,q)	Φθίνουν	Φθίνουν

9.2 ΣΤΡΑΤΗΓΙΚΗ ΔΗΜΙΟΥΡΓΙΑΣ ΜΟΝΤΕΛΟΥ

Όπως φαίνεται στο πρώτο γράφημα με το λογικό διάγραμμα, η Box-Jenkins προσέγγιση χρησιμοποιεί μία επαναληπτική στρατηγική δημιουργίας μοντέλου, που συνίσταται από την επιλογή ενός αρχικού μοντέλου (model identification), την εκτίμηση των συντελεστών του μοντέλου (parameter estimation) και την ανάλυση των καταλοίπων (model checking). Αν είναι απαραίτητο, το αρχικό μοντέλο τροποποιείται και η διαδικασία επαναλαμβάνεται μέχρι τα κατάλοιπα να δείξουν, ότι δεν είναι απαραίτητη περαιτέρω τροποποίηση. Σε αυτό το σημείο, το

προσαρμοσμένο μοντέλο μπορεί να χρησιμοποιηθεί για πρόβλεψη.

Τα βήματα στη στρατηγική "χτισίματος" μοντέλου εξετάζονται λεπτομερώς στη συνέχεια.

9.2.1 ΒΗΜΑ 1: ΑΝΑΓΝΩΡΙΣΗ ΜΟΝΤΕΛΟΥ

1. Το πρώτο βήμα στην αναγνώριση του μοντέλου είναι ο καθορισμός της στασιμότητας των χρονοσειρών, το οποίο σημαίνει αν οι χρονοσειρές εμφανίζονται να διαφέρουν γύρω από ένα σταθερό επίπεδο. Είναι χρήσιμο να κοιτάξουμε σε ένα διάγραμμα τις χρονοσειρές μαζί με τη συνάρτηση αυτοσυσχέτισης δείγματος. Μη στάσιμες χρονοσειρές αυξάνονται ή μειώνονται στο χρόνο και οι αυτοσυσχετίσεις δείγματος δεν εξασθενούν γρήγορα.
   Αν οι χρονοσειρές δεν είναι στάσιμες, μπορεί συχνά, όχι πάντα, να μετατραπούν σε στάσιμες με τη μέθοδο των διαφορών. Αυτό σημαίνει, ότι οι πρωτότυπες σειρές αντικαθιστώνται από τις σειρές των διαφορών. Τότε προσδιορίζεται ένα ARMA μοντέλο και μοντελοποιούνται οι αλλαγές-διαφορές παρά τα επίπεδα τιμών.

   Ως ένα παράδειγμα, υποθέτουμε ότι οι πρωτότυπες σειρές, Y_t, γενικά αυξάνουν στο χρόνο, αλλά οι πρώτες διαφορές, $\Delta Y_t=Y_t-Y_{t-1}$, διαφέρουν γύρω από ένα σταθερό επίπεδο. Ας υποθέσουμε, ότι είναι κατάλληλο να μοντελοποιηθούν οι στάσιμες διαφορές χρησιμοποιώντας ένα ARMA μοντέλο, ας πούμε τάξης p=1 και q=1. Σε αυτή την περίπτωση το μοντέλο είναι:

$$\Delta Y_t=\varphi_1 \Delta Y_{t-1}+\varepsilon_t-\omega_1 \varepsilon_{t-1} \quad \text{ή}$$

$$Y_t-Y_{t-1}=\varphi_1(Y_{t-1}-Y_{t-2})+\varepsilon_t-\omega_1 \varepsilon_{t-1}$$

   Σε ορισμένες περιπτώσεις, μπορεί να είναι απαραίτητο να πάρουμε δεύτερες διαφορές, πριν αποκτηθούν στάσιμα δεδομένα. Η απλή

μέθοδος των διαφορών λαμβάνει χώρα δύο φορές και τα στάσιμα δεδομένα είναι:

$$\Delta^2 Y_t = \Delta(\Delta Y_t) = \Delta(Y_t - Y_{t-1}) = \Delta Y_t - \Delta Y_{t-1} = Y_t - 2Y_{t-1} + Y_{t-2}.$$

Η μέθοδος των διαφορών εφαρμόζεται, μέχρις ότου το διάγραμμα των δεδομένων να υποδείξει, ότι οι σειρές των διαφορών μεταβάλλονται γύρω από ένα σταθερό επίπεδο και ότι οι αυτοσυσχετίσεις δείγματος φθίνουν αρκετά γρήγορα. Ο αριθμός των διαφορών, που απαιτείται, για να επιτευχθεί η στασιμότητα ορίζεται με το d. Ο αριθμός των διαφορών συμβολίζει την τάξη της "παραγώγου", που πρέπει να πάρουμε, μέχρι να γίνει σταθερά.

Τα μοντέλα για μη στάσιμες σειρές ονομάζονται ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητού μέσου (autoregressive integrated moving average models) και συμβολίζονται με ARIMA(p,d,q). Εδώ, το p συμβολίζει τη τάξη του αυτοπαλινδρομικού μέρους, το d συμβολίζει τον αριθμό των διαφορών, που πρέπει να λάβουν χώρα, ώστε η χρονοσειρά να γίνει στάσιμη και το q συμβολίζει την τάξη του μέρους κινούμενου μέσου. Αν οι πρωτότυπες σειρές είναι στάσιμες, τότε d=0, και τα ARIMA μοντέλα μειώνονται στα ARMA μοντέλα. Κατά συνέπεια, από αυτό το σημείο, η σημειογραφία ARIMA(p,d,q) θα χρησιμοποιείται για να υποδείξει μοντέλα για στάσιμες (d=0) και μη στάσιμες (d>0) χρονοσειρές.

Αν και τα ARIMA μοντέλα περιλαμβάνουν διαφορές, μπορούν πάντα να υπολογιστούν οι προβλέψεις για τις πρωτότυπες σειρές απευθείας από το προσαρμοσμένο μοντέλο.

2. Όταν οι χρονοσειρές γίνουν στάσιμες, πρέπει να αναγνωρίσουμε τη μορφή του μοντέλου, που θα χρησιμοποιηθεί.

Η αναγνώριση της μορφής του μοντέλου πραγματοποιείται συγκρίνοντας τις αυτοσυσχετίσεις και τις μερικές αυτοσυσχετίσεις, που υπολογίστηκαν από τα δεδομένα, με τις θεωρητικές αυτοσυσχετίσεις και τις θεωρητικές μερικές αυτοσυσχετίσεις για τα διάφορα ARIMA μοντέλα. Οι θεωρητικές συσχετίσεις για μερικά από τα πιο κοινά ARIMA μοντέλα φαίνονται στα Γραφήματα 1, 2, και 3 και βοηθούν στην επιλογή του κατάλληλου μοντέλου.

Κάθε ARIMA μοντέλο έχει ένα μοναδικό σύνολο αυτοσυσχετίσεων και μερικών αυτοσυσχετίσεων και ο αναλυτής πρέπει να είναι σε θέση να ταιριάξει τις αντίστοιχες τιμές δείγματος με ένα από τα θεωρητικά μοτίβα.

Μπορεί να είναι διφορούμενος ο καθορισμός ενός κατάλληλου ARIMA μοντέλου από τα μοτίβα των αυτοσυσχετίσεων δείγματος και των μερικών αυτοσυσχετίσεων. Έτσι, η επιλογή του αρχικού μοντέλου πρέπει να θεωρείται ως διερευνητική. Αναλύσεις μπορούν να γίνουν κατά τη διάρκεια των βημάτων 2 και 3 για να προσδιορίσουν αν το μοντέλο είναι επαρκές. Αν δεν είναι, τότε πρέπει να δοκιμαστεί ένα εναλλακτικό μοντέλο. Με λίγη εξάσκηση, ο αναλυτής μπορεί να γίνει πιο ειδικός στην αναγνώριση ενός επαρκούς μοντέλου.

Να θυμόμαστε ότι, αν οι αυτοσυσχετίσεις δείγματος φθίνουν εκθετικά προς το μηδέν και οι μερικές αυτοσυσχετίσεις δείγματος διακόπτονται, το μοντέλο θα απαιτεί αυτοπαλινδρομικούς όρους. Αν οι αυτοσυσχετίσεις δείγματος διακόπτονται και οι μερικές αυτοσυσχετίσεις δείγματος φθίνουν, το μοντέλο θα απαιτεί όρους κινούμενου μέσου. Αν και οι αυτοσυσχετίσεις δείγματος και οι

μερικές αυτοσυσχετίσεις δείγματος φθίνουν, υποδεικνύονται και αυτοπαλινδρομικοί όροι και όροι κινούμενου μέσου. Στην περίπτωση που διακόπτονται και τα δύο, τότε θα διερευνήσουμε και το αυτοπαλινδρομικό και το κινούμενου μέσου μοντέλο και θα διαλέξουμε για πρόβλεψη το πλέον επαρκές. Μετρώντας τον αριθμό των σημαντικών αυτοσυσχετίσεων δείγματος και των μερικών αυτοσυσχετίσεων, μπορούν να καθοριστούν οι τάξεις των MA και AR μερών. Για να κριθεί η σημασία τους και οι αυτοσυσχετίσεις δείγματος και οι μερικές αυτοσυσχετίσεις δείγματος συχνά συγκρίνονται με $\pm 2/\sqrt{n}$, από το μηδέν, όπου n είναι ο αριθμός των παρατηρήσεων στις χρονοσειρές. Αυτά τα όρια λειτουργούν σωστά, όταν το n είναι μεγάλο.

Με την επάρκεια των μοντέλων δεδομένη, τα απλά μοντέλα προτιμώνται των σύνθετων μοντέλων. Αυτό είναι γνωστό σαν την **αρχή της οικονομίας** (principle of parsimony). Με μία περιορισμένη ποσότητα δεδομένων, είναι σχετικά εύκολο να βρεθεί ένα μοντέλο με ένα μεγάλο αριθμό παραμέτρων, που προσαρμόζεται σωστά στα δεδομένα. Ωστόσο, οι προβλέψεις από ένα τέτοιο μοντέλο ενδέχεται να είναι φτωχές, επειδή το μεγαλύτερο μέρος της μεταβλητότητας των δεδομένων επεξηγείται από το τυχαίο σφάλμα, που μοντελοποιείται. Ο στόχος είναι η ανάπτυξη του απλούστερου μοντέλου, που παρέχει μία επαρκή περιγραφή των κύριων χαρακτηριστικών των δεδομένων.

Η αρχή της οικονομίας αναφέρεται στην προτίμηση για απλά μοντέλα αντί για σύνθετα.

9.2.2 ΒΗΜΑ 2: ΕΚΤΙΜΗΣΗ ΜΟΝΤΕΛΟΥ

1. Όταν ένα διερευνητικό μοντέλο έχει επιλεχθεί, πρέπει να

εκτιμηθούν οι παράμετροι για το συγκεκριμένο μοντέλο.

Οι παράμετροι στα ARIMA μοντέλα υπολογίζονται ελαχιστοποιώντας το άθροισμα των τετραγώνων των προσαρμοσμένων σφαλμάτων. Αυτές οι εκτιμήσεις ελαχίστων τετραγώνων, πρέπει γενικά να αποκτώνται με τη χρήση μίας μη γραμμικής διαδικασίας ελαχίστων τετραγώνων. Μία μη γραμμική διαδικασία ελαχίστων τετραγώνων είναι ένας αλγόριθμος, που βρίσκει το ελάχιστο από τη συνάρτηση αθροίσματος τετραγωνικών σφαλμάτων. Όταν καθοριστούν οι εκτιμήσεις των συντελεστών με χρήση της μεθόδου των ελαχίστων τετραγώνων και τα τυπικά σφάλματά τους, κατασκευάζονται οι t τιμές και ερμηνεύονται με το συνήθη τρόπο. Οι παράμετροι, που κρίνονται σημαντικά διαφορετικές από το μηδέν, διατηρούνται στο προσαρμοσμένο μοντέλο. Οι παράμετροι, που δεν είναι σημαντικές, απορρίπτονται από το μοντέλο.

Για παράδειγμα ας υποθέσουμε, ότι ένα ARIMA(1,0,1) έχει προσαρμοστεί σε μία χρονοσειρά 100 παρατηρήσεων και η προσαρμοσμένη εξίσωση μετά την αναγνώριση του μοντέλου από τους συντελεστές συσχέτισης και μερικής αυτοσυσχέτισης, είναι:

$$\hat{Y}_t = 33{,}4 + 0{,}25\, Y_{t-1} - 0{,}5\, \hat{\varepsilon}_{t-1}$$

$$(7{,}02) \quad (0{,}17) \qquad (0{,}21)$$

όπου οι αριθμοί των παρενθέσεων κάτω από τις υπολογισμένους συντελεστές είναι τα τυπικά σφάλματά τους. Μιας και ο t λόγος για το συντελεστή του αυτοπαλινδρομικού όρου είναι t=0,25/0,17=1,47 με μία p-τιμή 0,14, η υπόθεση H_0:φ_1=0 δεν απορρίπτεται, και αυτός ο όρος μπορεί να διαγραφεί από το μοντέλο. Ένα ARIMA(0,0,1) μοντέλο, που είναι ένα MA(1)

μοντέλο, μπορεί τότε να προσαρμοστεί στα δεδομένα.

2. Υπολογίζεται το μέσο τετραγωνικό σφάλμα καταλοίπων, μία εκτίμηση δηλαδή της διασποράς του σφάλματος, ε_t. Το μέσο τετραγωνικό σφάλμα καταλοίπων ορίζεται ως:

$$s^2 = \frac{\sum_{t=1}^{n} e_t^2}{n-r} = \frac{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}{n-r}$$

όπου,

- $e_t = Y_t - \hat{Y}_t$ = Το κατάλοιπο στο χρόνο t.
- n = Ο αριθμός των καταλοίπων.
- r = Ο συνολικός αριθμός των εκτιμώμενων παραμέτρων.
- Ο Παρονομαστής είναι οι βαθμοί ελευθερίας και υπολογίζεται από τη σχέση n-d-p-q-1 ή r=d+p+q+1 (το 1 είναι για τον σταθερό όρο φ_0).

3. Το μέσο τετραγωνικό σφάλμα καταλοίπων είναι χρήσιμο για αξιολόγηση προσαρμογής και σύγκριση διαφορετικών μοντέλων. Επίσης χρησιμοποιείται για υπολογισμό ορίων σφαλμάτων πρόβλεψης.

9.2.3 ΒΗΜΑ 3: ΕΛΕΓΧΟΣ ΜΟΝΤΕΛΟΥ

Πριν τη χρήση του μοντέλου για πρόβλεψη, πρέπει να ελεγχθεί για την επάρκεια του. Βασικά, ένα μοντέλο είναι επαρκές, αν τα κατάλοιπα δεν μπορούν να χρησιμοποιηθούν για να βελτιώσουν τις προβλέψεις, ή αν στα σφάλματα δεν κρύβεται χρήσιμη πληροφορία. Αυτό σημαίνει, ότι τα κατάλοιπα είναι τυχαία και ανεξάρτητα.

1. Πολλά από τα ίδια διαγράμματα καταλοίπων, που είναι χρήσιμα στην ανάλυση παλινδρόμησης, μπορούν να αναπτυχθούν για τα

κατάλοιπα από ένα ARIMA μοντέλο. Ένα ιστόγραμμα, ένα διάγραμμα κανονικής κατανομής (για έλεγχο κανονικότητας) και ένα διάγραμμα χρονικής αλληλουχίας (time sequence plot) είναι ιδιαίτερα χρήσιμα.

2. Οι μεμονωμένες αυτοσυσχετίσεις καταλοίπων πρέπει να είναι μικρές και γενικά μέσα στο διάστημα $\pm 2/\sqrt{n}$ από το 0. Σημαντικές αυτοσυσχετίσεις καταλοίπων σε χαμηλές υστερήσεις (lags) ή εποχικές υστερήσεις υποδεικνύουν, ότι το μοντέλο είναι ανεπαρκές και ότι πρέπει να επιλεχθεί ένα νέο ή τροποποιημένο μοντέλο.
3. Οι αυτοσυσχετίσεις καταλοίπων σαν σύνολο πρέπει να είναι συνεπείς με εκείνες, που παράγονται από τυχαία σφάλματα.

Ένας συνολικός έλεγχος της επάρκειας του μοντέλου παρέχεται από ένα chi-square (χ^2) έλεγχο, που βασίζεται στη Ljung-Box Q στατιστική. Αυτός ο έλεγχος φαίνεται στα μεγέθη των αυτοσυσχετίσεων καταλοίπων σαν ένα σύνολο. Ο έλεγχος της στατιστικής Q είναι:

$$Q=(n-d)(n-d+2)\frac{\sum_{k=1}^{m} r_k^2(e)}{n-d-k}$$

όπου,

- $r_k(e)$ = είναι οι συντελεστές αυτοσυσχέτισης.
- n = ο αριθμός των παρατηρήσεων της χρονοσειράς.
- m = ο αριθμός των χρονο-υστερήσεων, που πρέπει να ελεγχθούν.
- k = η εκάστοτε χρονο-υστέρηση (lag).
- d = ο αριθμός των διαφορών

Η στατιστική Q ακολουθεί την chi-square κατανομή με **m-r βαθμούς ελευθερίας**, όπου r εδώ είναι ο αριθμός των παραμέτρων

(p+q+1), που εκτιμώνται στο ARIMA μοντέλο.

Αν η p-τιμή, που σχετίζεται με τη Q στατιστική δεν είναι μικρή, (ας πούμε, p-τιμή>0,05 για επίπεδο σημαντικότητας 5%), τότε το μοντέλο θεωρείται ανεπαρκές. Ο αναλυτής πρέπει να θεωρήσει ένα νέο ή τροποποιημένο μοντέλο και να συνεχίσει την ανάλυση μέχρι να καθοριστεί ένα ικανοποιητικό μοντέλο.

Η κρίση παίζει σημαντικό ρόλο στην προσπάθεια δημιουργίας μοντέλου. Δύο απλά μοντέλα μπορεί να περιγράφουν επαρκώς τα δεδομένα και η επιλογή μπορεί να γίνει στη βάση της φύσης της πρόβλεψης. Επίσης, μπορεί να αγνοηθούν μερικά μεγάλα κατάλοιπα, αν μπορούν να ερμηνευθούν από ασυνήθιστες καταστάσεις και το μοντέλο να είναι επαρκές για τις υπόλοιπες παρατηρήσεις.

9.2.4 ΒΗΜΑ 4: ΠΡΟΒΛΕΠΟΝΤΑΣ ΜΕ ΤΟ ΜΟΝΤΕΛΟ

1. Όταν βρεθεί ένα επαρκές μοντέλο, μπορούν να γίνουν προβλέψεις για μία ή περισσότερες περιόδους στο μέλλον.

   Μπορούν επίσης να κατασκευαστούν διαστήματα εμπιστοσύνης πρόβλεψης, που βασίζονται στο χρονικό διάστημα των προβλέψεων. Γενικά, για ένα δεδομένο επίπεδο εμπιστοσύνης, όσο μεγαλύτερο είναι το χρονικό διάστημα της πρόβλεψης, τόσο μεγαλύτερο είναι το διάστημα εμπιστοσύνης της πρόβλεψης. Αυτό είναι λογικό, δεδομένου ότι η αβεβαιότητα αναμένεται να είναι μεγαλύτερη για μία πρόβλεψη πιο μακρινή μελλοντικά, από ότι για μία πρόβλεψη, ας πούμε, της επόμενης παρατήρησης. Ο υπολογισμός προβλέψεων και διαστημάτων πρόβλεψης είναι κουραστικός και είναι καλύτερο να αφήνεται στον υπολογιστή. Τα υπολογιστικά προγράμματα, που προσαρμόζουν ARIMA μοντέλα, παράγουν προβλέψεις και διαστήματα πρόβλεψης στο αίτημα του

αναλυτή.

2. Καθώς περισσότερα δεδομένα γίνονται διαθέσιμα, το ίδιο ARIMA μοντέλο μπορεί να χρησιμοποιηθεί για να παράγει αναθεωρημένες προβλέψεις για άλλη χρονική περίοδο.
3. Αν το μοτίβο των σειρών φαίνεται να αλλάζει στο χρόνο, τα νέα δεδομένα μπορεί να χρησιμοποιηθούν για να επανεκτιμήσουν τις παραμέτρους του μοντέλου ή, αν είναι απαραίτητο, να αναπτύξουν ένα ολοκληρωτικά νέο μοντέλο.

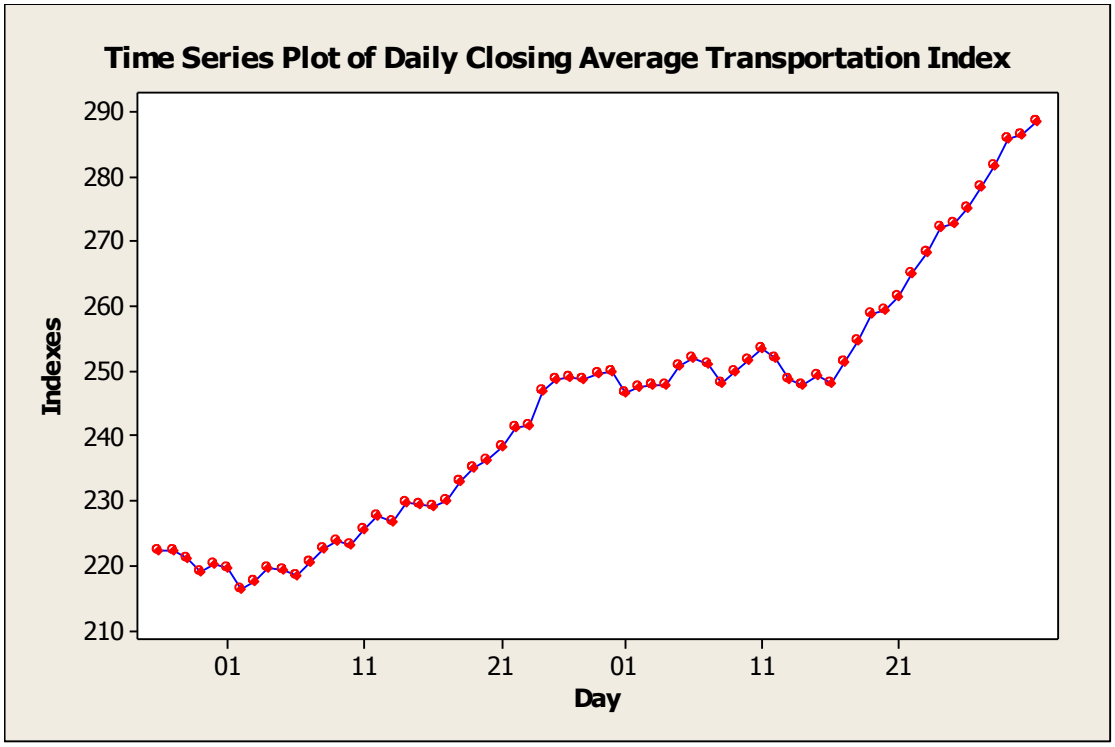
Είναι καλή ιδέα να παρακολουθούνται τα σφάλματα πρόβλεψης. Αν τα πιο πρόσφατα σφάλματα τείνουν να είναι συνεχώς μεγαλύτερα από τα προηγούμενα σφάλματα, ίσως πρέπει να επαναξιολογηθεί το μοντέλο.

Σε αυτό το σημείο, μπορεί να απαιτείται άλλη μία επανάληψη της στρατηγικής δημιουργίας του μοντέλου. Τα ίδια ισχύουν αν τα πρόσφατα σφάλματα πρόβλεψης τείνουν να είναι συνεχώς θετικά (underpredicting) ή αρνητικά (overpredicting).

9.2.5 ΠΑΡΑΔΕΙΓΜΑΤΑ

Παράδειγμα (Τ.9-3)

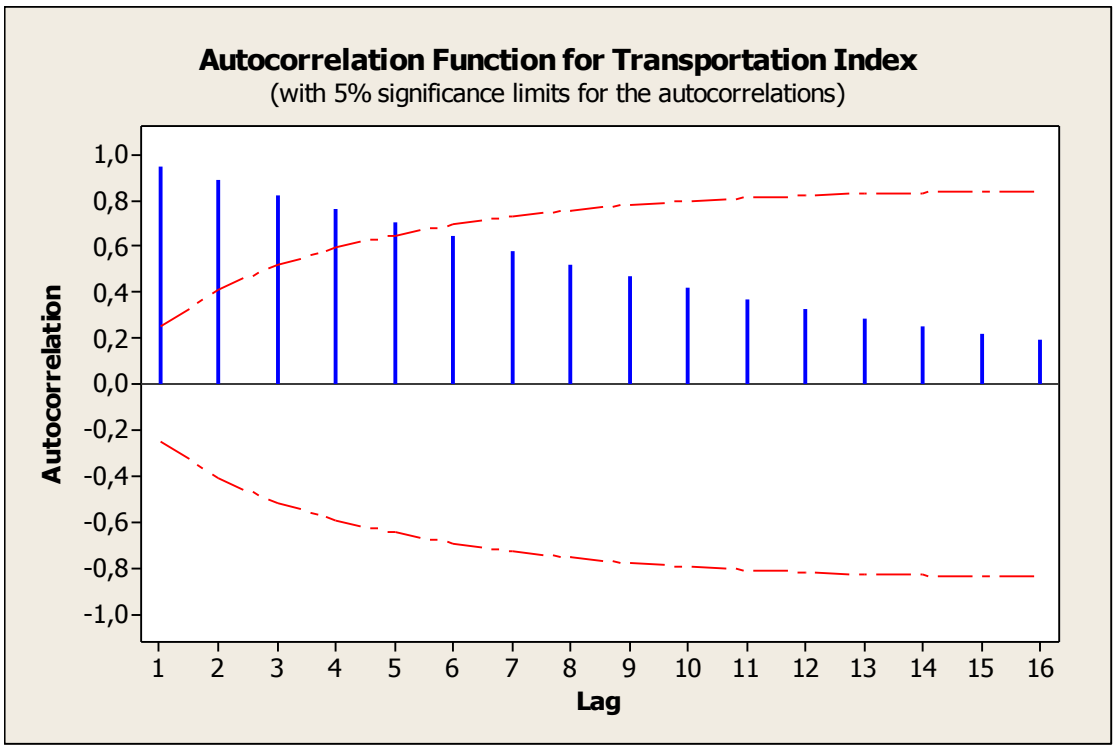
Στον αναλυτή μιας εταιρίας ανατέθηκε η εργασία να αναπτύξει πιο περίπλοκες μεθόδους για πρόβλεψη των μέσων του Dow Jones. Ο αναλυτής αποφάσισε να δοκιμάσει τη τεχνική της Box Jenkins μεθοδολογίας στο δείκτη μεταφορών του Dow Jones (Dow Jones Transportation Index). Το παρακάτω γράφημα αναπαριστά 65 καθημερινούς μέσους κλεισίματος του δείκτη μεταφορών κατά τη διάρκεια των καλοκαιρινών μηνών για 65 χρονικές περιόδους.

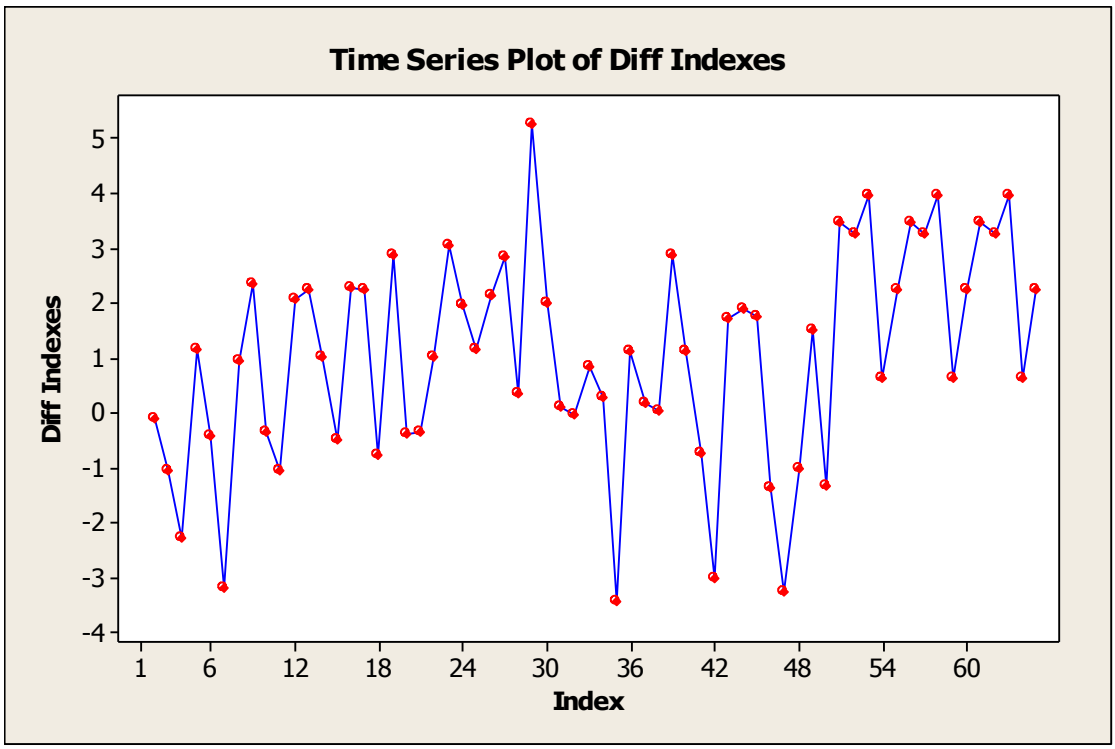


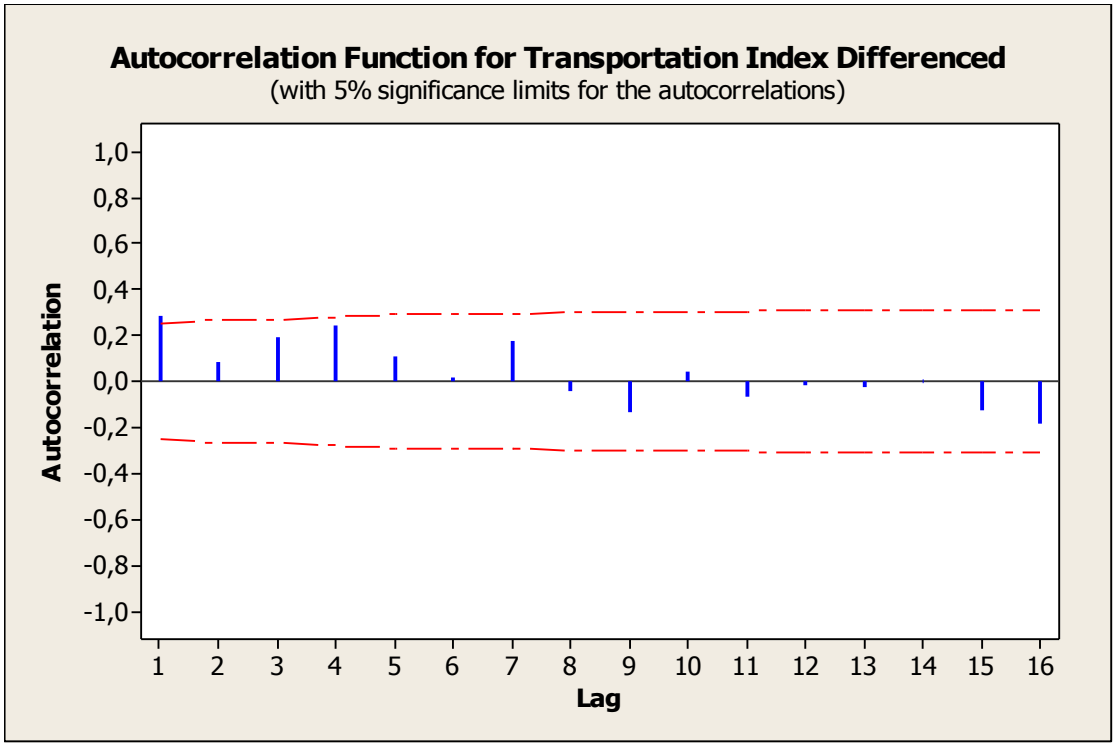
Ο αναλυτής ξεκίνησε την ανάλυση κοιτάζοντας ένα διάγραμμα της χρονοσειράς, που φαίνεται στο παραπάνω Γράφημα. Εκεί εμφανίζεται να υπάρχει μία ανοδική τάση στις σειρές. Το επόμενό του βήμα στην αναγνώριση ενός διερευνητικού μοντέλου ήταν να κοιτάξει τις αυτοσυσχετίσεις δείγματος των δεδομένων, οι οποίες φαίνονται στο Γράφημα της συνάρτησης αυτοσυσχέτισης. Όταν παρατήρησε ότι οι πρώτες αυτοσυσχετίσεις ήταν διαρκώς μεγάλες και ότι έφθιναν στο μηδέν σχετικά αργά, αναγνώρισε ότι οι αρχικές του παρατηρήσεις ήταν σωστές: Αυτή η χρονοσειρά ήταν μη στάσιμη και δεν διέφερε γύρω από ένα σταθερό επίπεδο.

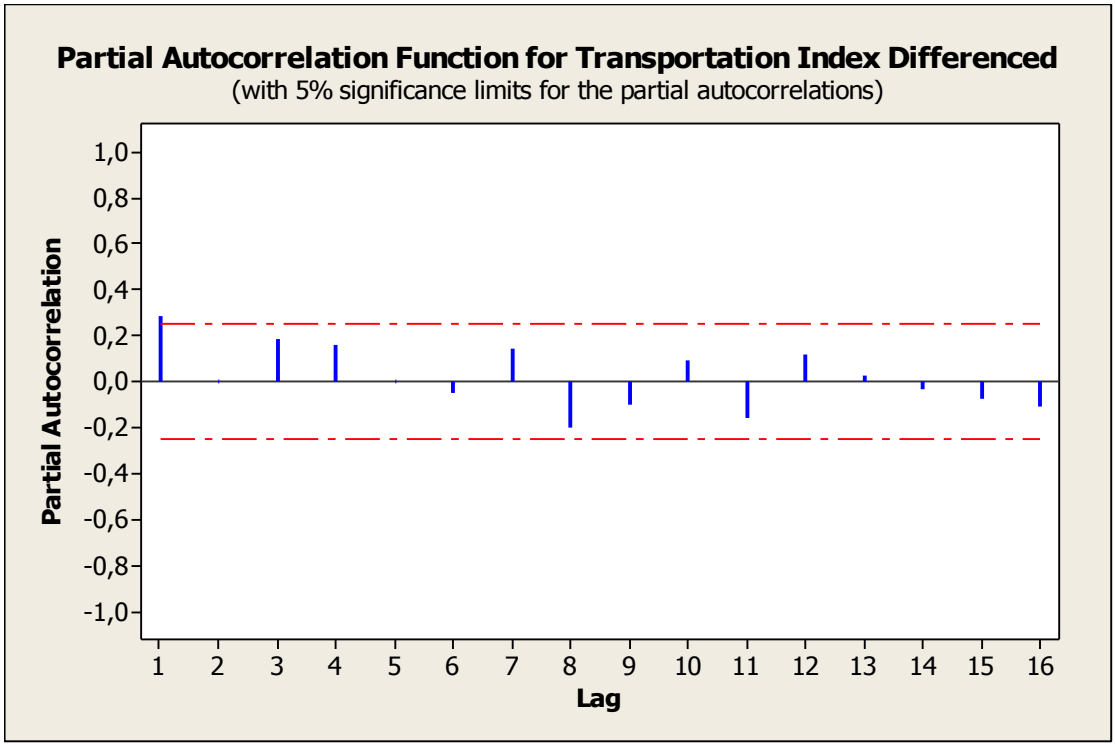
Ο αναλυτής αποφάσισε να εφαρμόσει τη μέθοδο των διαφορών στα δεδομένα για να δει αν θα μπορούσε να εξαλείψει την τάση και να δημιουργήσει στάσιμες σειρές. Ένα διάγραμμα των δεδομένων μετά τη μέθοδο των διαφορών τα εμφανίζει να διαφέρουν γύρω από ένα σταθερό επίπεδο. Στην πραγματικότητα, ο μέσος δείγματος των διαφορών ήταν 1,035. Οι αυτοσυσχετίσεις και μερικές αυτοσυσχετίσεις δείγματος για τις

διαφορές φαίνονται στα αντίστοιχα Γραφήματα.









Ο αναλυτής είχε μπερδευτεί. Συγκρίνοντας τις αυτοσυσχετίσεις με τα όρια σφαλμάτων τους, η μόνο σημαντική αυτοσυσχέτιση ήταν στην καθυστέρηση 1 (lag 1). Παρόμοια, μόνο η lag 1 μερική αυτοσυσχέτιση ήταν σημαντική. Οι αυτοσυσχετίσεις εμφανίζονται να διακόπτονται μετά

τη lag 1, υποδεικνύοντας MA(1) συμπεριφορά. Την ίδια στιγμή, οι μερικές αυτοσυσχετίσεις εμφανίζονται να διακόπτονται μετά τη lag 1, υποδεικνύοντας AR(1) συμπεριφορά. Κανένα μοτίβο δεν εμφανίζεται να φθίνει με ένα πτωτικό τρόπο σε μικρές καθυστερήσεις. Ο αναλυτής αποφάσισε να προσαρμόσει και τα δύο μοντέλα, ARIMA(1,1,0) και ARIMA(0,1,1), στο δείκτη μεταφορών. Αυτός αποφάσισε να συμπεριλάβει ένα σταθερό όρο σε κάθε μοντέλο για να επιτρέψει το γεγονός, ότι οι σειρές των διαφορών εμφανίζεται να διαφέρουν γύρω από ένα επίπεδο μεγαλύτερο από το μηδέν. Αν το Y_t, συμβολίζει το δείκτη μεταφοράς, τότε οι σειρές των διαφορών είναι $\Delta y_t=Y_t-Y_{t-1}$ και τα μοντέλα είναι:

$$\text{ARIMA}(1,1,0): \Delta Y_t=\varphi_0+\varphi_1\Delta Y_{t-1}+\varepsilon_t$$

$$\text{ARIMA}(0,1,1): \Delta Y_t=\mu+\varepsilon_t-\omega_1\varepsilon_{t-1}$$

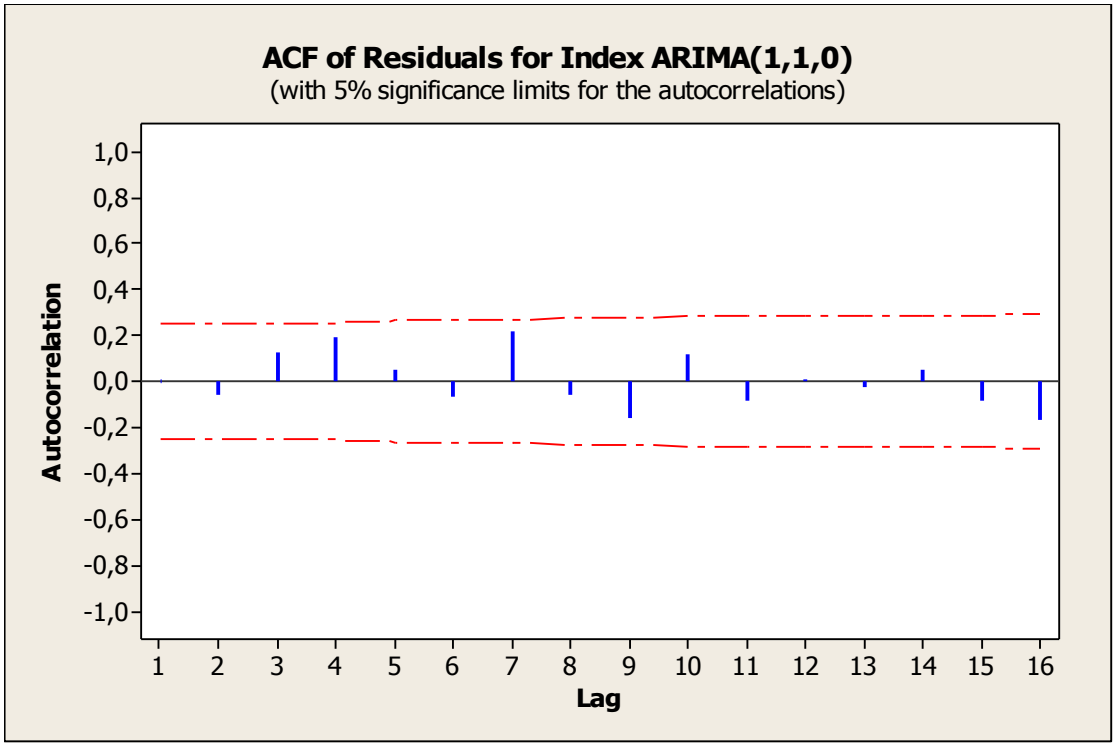
Τα αποτελέσματα του Minitab για τα μοντέλα του αναλυτή φαίνονται στον παρακάτω Πίνακα. Οι αυτοσυσχετίσεις καταλοίπων για την προσαρμογή του μοντέλου ARIMA(1,1,0) φαίνονται στο ακόλουθο Γράφημα.

Και τα δύο μοντέλα προσαρμόζονται στα δεδομένα εξίσου καλά. Οι τετραγωνικοί μέσοι καταλοίπων (MS) είναι:

$$\text{ARIMA}(1,1,0): s^2=3,536$$

$$\text{ARIMA}(0,1,1): s^2=3,538$$

Ο αναλυτής επίσης παρατήρησε, ότι ο σταθερός όρος στο ARIMA(0,1,1) μοντέλο εκτιμάται να είναι $\hat{\mu}=1,0381$, (ουσιαστικά) ο μέσος όρος του δείγματος των διαφορών.



Lag	ACF	T	LBQ
1	-0,002164	-0,02	0,00
2	-0,054880	-0,44	0,21
3	0,126703	1,01	1,32
4	0,192990	1,52	3,94
5	0,053912	0,41	4,15
6	-0,069976	-0,53	4,50
7	0,215477	1,62	7,94
8	-0,062525	-0,45	8,24
9	-0,155310	-1,12	10,09
10	0,119656	0,85	11,21
11	-0,087686	-0,61	11,82
12	0,005717	0,04	11,83
13	-0,024945	-0,17	11,88
14	0,052536	0,37	12,11
15	-0,083232	-0,58	12,71
16	-0,170800	-1,18	15,28

Το Γράφημα αυτοσυσχέτισης δείχνει ότι δεν υπάρχει σημαντική αυτοσυσχέτιση καταλοίπων για το ARIMA(1,1,0) μοντέλο. Αν και η συνάρτηση αυτοσυσχέτισης καταλοίπων για το ARIMA(0,1,1) μοντέλο δεν φαίνεται, τα αποτελέσματα είναι παρόμοια. Τα Ljung-Box Q

στατιστικά, που υπολογίστηκαν για τις ομάδες των lags m=12, 24, 36 και 48, δεν είναι σημαντικά, όπως υποδεικνύεται από τις μεγάλες p-τιμές για κάθε μοντέλο. Ο αναλυτής αποφάσισε ότι κάθε μοντέλο είναι επαρκές. Επίσης, οι προβλέψεις ενός βήματος μπροστά. που παρήχθησαν από τα δύο μοντέλα, ήταν σχεδόν οι ίδιες.

Για να λυθεί το δίλημμά του, ο αναλυτής διάλεξε το ARIMA(1,1,0) μοντέλο, γιατί προσαρμοζόταν ελάχιστα καλύτερα. Έλεγξε την πρόβλεψη για την περίοδο 66 για αυτό το μοντέλο ως εξής:

$$Y_t - Y_{t-1} = \varphi_0 + \varphi_1 (Y_{t-1} - Y_{t-2}) + \varepsilon_t$$

Με $\hat{\varphi}_0 = 0{,}741$ *και* $\hat{\varphi}_1 = 0{,}284$, η εξίσωση πρόβλεψης γίνεται:

$$\hat{Y}_{66} = Y_{65} + 0{,}741 + 0{,}284 (Y_{65} - Y_{64}) = 289{,}947.$$

Η πρόβλεψη συμφωνεί με το αποτέλεσμα του κάτωθι Πίνακα. Το διάστημα πρόβλεψης για την πραγματική τιμή στην περίοδο 66, που υπολογίστηκε από το Minitab, είναι (286,3 , 293,6).

ARIMA(1,1,0): Model for Transportation Index

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
AR 1	0,2844	0,1221	2,33	0,023
Constant	0,7408	0,2351	3,15	0,003

Differencing: 1 regular difference
Number of observations: Original series 65, after differencing 64

Residuals: SS = 219,223 (backforecasts excluded)
MS = 3,536 DF = 62

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	11,8	29,1	37,1	48,1

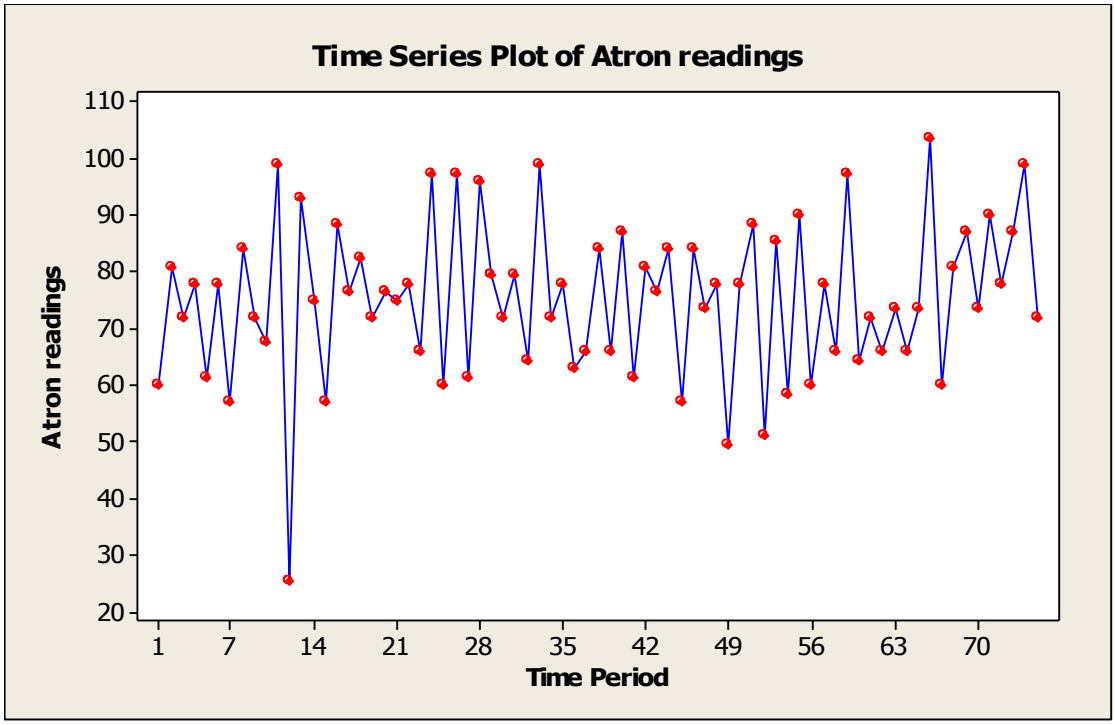
```
DF            10     22     34     46
P-Value    0,297  0,141  0,328  0,389

Forecasts from period 65 95% Limits
Period  Forecast    Lower    Upper   Actual
66       289,948  286,262  293,634
```

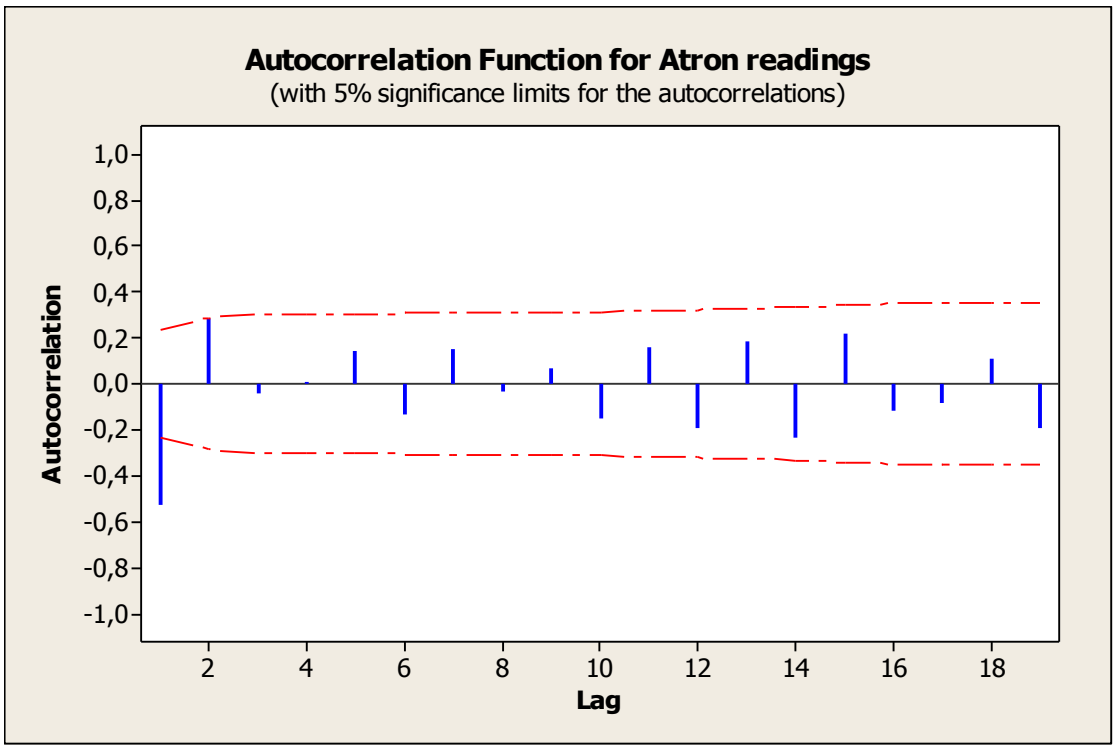
ARIMA(0,1,1): Model for Transportation Index

```
Final Estimates of Parameters

Type        Coef  SE Coef     T     P
MA 1     -0,2913   0,1226 -2,38 0,021
Constant  1,0381   0,3035  3,42 0,001

Differencing: 1 regular difference
Number of observations: Original series 65, after
differencing 64

Residuals: SS = 219,347 (backforecasts excluded)
           MS = 3,538 DF = 62

Modified Box-Pierce (Ljung-Box) Chi-Square
statistic

Lag             12     24     36     48
Chi-Square    11,6   32,0   41,0   51,4
DF              10     22     34     46
P-Value       0,310  0,077  0,189  0,270

Forecasts from period 65 95% Limits
Period  Forecast   Lower    Upper    Actual
66       290,053  286,366  293,740
```

Παράδειγμα (Τ.9-5)

Ένας αναλυτής μίας εταιρίας είχε στη διάθεσή του χρονοσειρές αναγνώσεων για μία εργοστασιακή διαδικασία, που χρειαζόταν να προβλέψει. Ο αναλυτής πίστεψε, ότι η Box-Jenkins μεθοδολογία μπορεί να δουλεύει καλύτερα για αυτά τα δεδομένα. Στο παρακάτω διάγραμμα απεικονίζεται η χρονοσειρά.

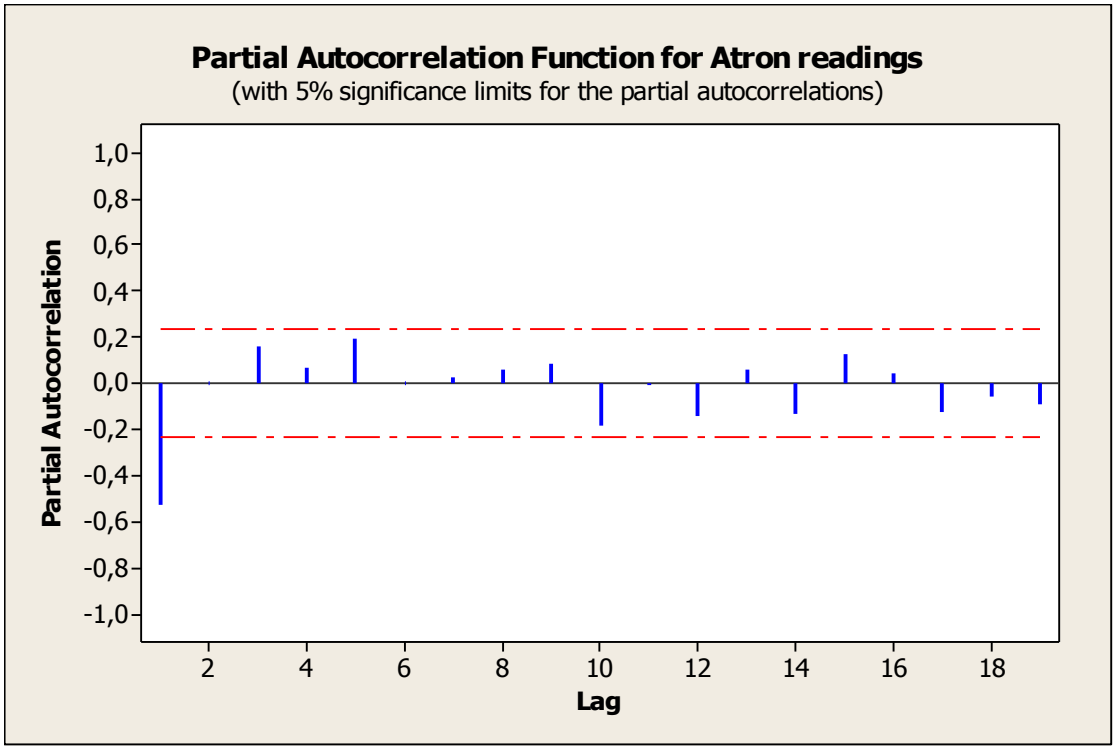




Lag	ACF	T	LBQ
1	-0,528475	-4,58	21,80
2	0,281574	1,95	28,07
3	-0,038337	-0,25	28,19
4	0,007939	0,05	28,19

Ο αναλυτής ξεκίνησε να σκιαγραφεί ένα διερευνητικό μοντέλο κοιτάζοντας το διάγραμμα των δεδομένων και τις αυτοσυσχετίσεις δείγματος, που φαίνονται στα αντίστοιχα Γραφήματα. Η χρονοσειρά των δεδομένων φαίνεται να κινείται γύρω από ένα σταθερό επίπεδο (περίπου 80) και οι αυτοσυσχετίσεις φθίνουν γρήγορα στο μηδέν. Ο αναλυτής συμπέρανε ότι η χρονοσειρά είναι στάσιμη. Ο πρώτος συντελεστής αυτοσυσχέτισης δείγματος (-0,53) είναι σημαντικά διαφορετικός από το μηδέν στο 5% επίπεδο, μιας και βρίσκεται στο εύρος:

$$0 \pm 2\frac{1}{\sqrt{n}} = 0 \pm 2\frac{1}{\sqrt{75}} = 0 \pm 2(0{,}115) = 0 \pm 0{,}23.$$



Lag	PACF	T
1	-0,528475	-4,58
2	0,003175	0,03
3	0,154951	1,34
4	0,065018	0,56

Η αυτοσυσχέτιση στο lag 2 είναι κοντά στο να είναι σημαντικό στο 5% επίπεδο και αντίθετο στο πρόσημο από την lag 1 αυτοσυσχέτιση, r_1.

Οι υπόλοιπες αυτοσυσχετίσεις είναι μικρές και καλά μέσα στο όρια ισχύος της μηδενικής υπόθεσης. Το μοτίβο των αυτοσυσχετίσεων προτείνει, είτε ένα AR(1) μοντέλο (βλ. Γράφημα 2(b)), είτε ίσως ένα MA(2) μοντέλο, αν οι αυτοσυσχετίσεις διακόπτονται (είναι μηδέν) μετά το lag 2. Ο αναλυτής αποφάσισε να εξετάσει και τις μερικές αυτοσυσχετίσεις δείγματος που φαίνονται στο αντίστοιχο Γράφημα.

Ο αναλυτής παρατήρησε ότι η πρώτη μερική αυτοσυσχέτιση (-0,53) ήταν σημαντικά διαφορετική από το μηδέν αλλά καμία από τις άλλες μερικές αυτοσυσχετίσεις δεν προσέγγισε, το να είναι σημαντική. Ο αναλυτής ένιωσε ότι οι αυτοσυσχετίσεις και οι μερικές αυτοσυσχετίσεις προτείνουν ένα AR(1) (ή, ισοδύναμα, ένα ARIMA(1,0,0)) μοντέλο, αλλά για να είναι ασφαλής, αποφάσισε να προσαρμόσει ένα MA(2) (ARIMA(0,0,2)) μοντέλο επίσης. Αν και τα δύο μοντέλα είναι επαρκή, θα αποφασίσει χρησιμοποιώντας την αρχή της οικονομίας.

Ο ακόλουθος Πίνακας δείχνει τα αποτελέσματα χρησιμοποιώντας το Minitab για να προσαρμοστούν τα AR(1) και MA(2) μοντέλα στις αναγνώσεις της διαδικασίας. Ένας σταθερός όρος συμπεριλαμβάνεται και στα δύο μοντέλα για να επιτρέψει και στα δύο μοντέλα να διαφέρουν γύρω από ένα επίπεδο άλλο από το μηδέν.

Οι μελλοντικές προβλέψεις θα συγκλίνουν στον δειγματικό μέσο της χρονοσειράς, επειδή είναι στάσιμη.

ARIMA(0,0,2) Model: Readings

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
MA 1	0,5667	0,1107	5,12	0,000
MA 2	-0,3560	0,1146	-3,11	0,003
Constant	75,410	1,061	71,08	0,000
Mean	75,410	1,061		

```
Number of observations: 75
Residuals: SS = 9724,97
           MS =  135,07 DF = 72

Modified Box-Pierce (Ljung-Box) Chi-Square
statistic
Lag          12    24     36     48
Chi-Square 7,0   23,8   31,8   46,9
DF            9    21     33     45
P-Value    0,638  0,305  0,526  0,394

Forecasts from period 75  95% Limits
Period   Forecast   Lower   Upper   Actual
76        80,648   57,864  103,431
77        78,169   51,982  104,356
78        75,410   47,996  102,825
```

ARIMA(1,0,0) Model: Readings

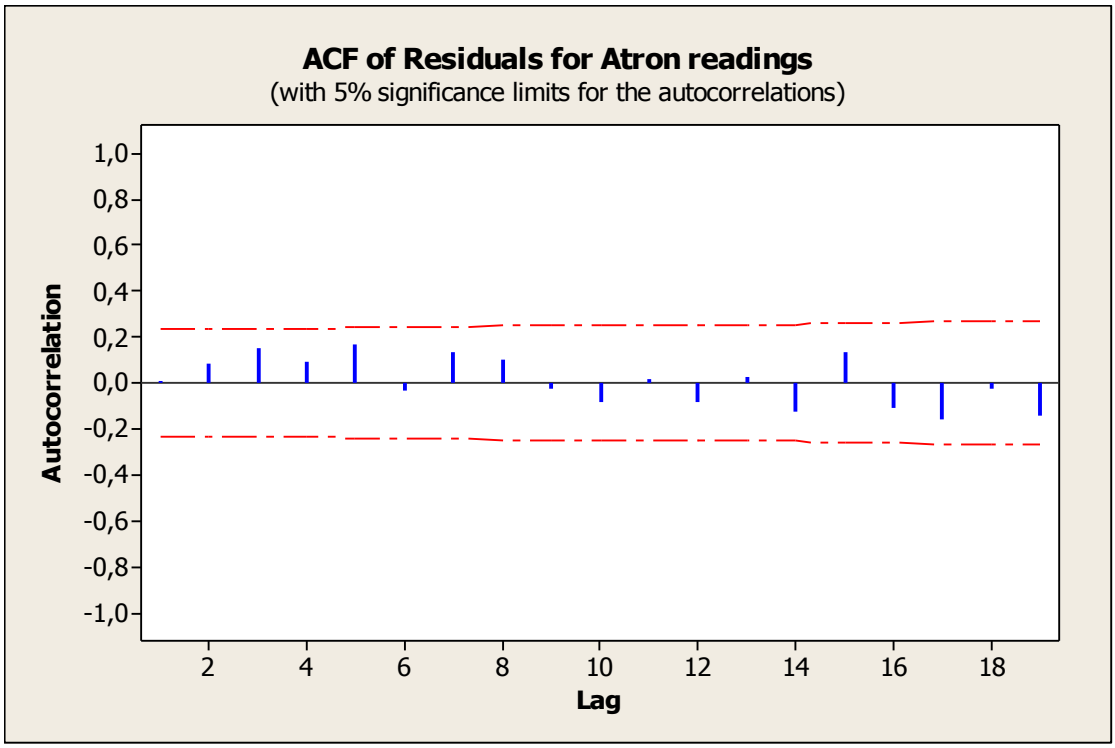
```
Final Estimates of Parameters

Type        Coef   SE Coef   T     P
AR 1      -0,5379  0,0986  -5,46 0,000
Constant 115,842   1,356   85,44 0,000
Mean      75,3269  0,8817

Number of observations: 75
Residuals: SS = 10065,2
           MS =   137,9 DF = 73

Modified Box-Pierce (Ljung-Box) Chi-Square
statistic
Lag          12     24     36     48
Chi-Square  9,3   29,9   37,3   58,3
DF           10     22     34     46
P-Value     0,503  0,121  0,319  0,105

Forecasts from period 75 95% Limits
Period   Forecast   Lower   Upper   Actual
76        77,116   54,097  100,136
77        74,364   48,227  100,502
78        75,845   48,872  102,817
```

Και τα δύο μοντέλα εμφανίζεται να προσαρμόζονται καλά στα

δεδομένα. Οι εκτιμώμενοι συντελεστές είναι σημαντικά διαφορετικοί από το μηδέν. Τα μέσα τετραγωνικά σφάλματα είναι παρόμοια:

MA(2): s^2=135,1

AR(1): s^2=137.9



Οι προβλέψεις για ένα και δύο βήματα μπροστά από τα δύο μοντέλα είναι διαφορετικές, αλλά οι προβλέψεις για τρία βήματα μπροστά (περίοδος 78) είναι σχεδόν οι ίδιες. Οι προβλέψεις για στάσιμες διαδικασίες θα ισούνται τελικά με το εκτιμώμενο μέσο επίπεδο. Σε αυτή την περίπτωση, το μέσο επίπεδο εκτιμάται να είναι περίπου $\hat{\mu}=75$ και για τα δύο μοντέλα.

Το Ljung-Box Q στατιστικό δεν είναι σημαντικό για συλλογές των καθυστερήσεων lags m=12, 24, 36 και 48 για κάθε μοντέλο. Οι αυτοσυσχετίσεις καταλοίπων για το AR(1) μοντέλο φαίνονται στο προηγούμενο Γράφημα. Οι μεμονωμένες αυτοσυσχετίσεις καταλοίπων είναι μικρές και κατά μέσον όρο στα όρια σφάλματος. Η συνάρτηση

αυτοσυσχέτισης καταλοίπων για το MA(2) μοντέλο είναι παρόμοια. Για κάθε μοντέλο, δεν υπάρχει λόγος να πιστέψουμε, ότι τα σφάλματα δεν είναι τυχαία.

Μιας και το AR(1) μοντέλο έχει δύο παραμέτρους (μαζί με το σταθερό όρο) και το MA(2) μοντέλο έχει τρεις παραμέτρους (μαζί με το σταθερό όρο), ο αναλυτής χρησιμοποίησε την αρχή της οικονομίας και αποφάσισε να χρησιμοποιήσει το απλούστερο AR(1) μοντέλο για να προβλέψει τις μελλοντικές αναγνώσεις.

Η AR(1) εξίσωση πρόβλεψης είναι:

$$\hat{Y}_t = 115{,}842 + (-0{,}538)\, Y_{t-1}$$

έτσι, για την περίοδο 76:

$$\hat{Y}_{76} = 115{,}842 + (-0{,}538)\, Y_{75} = 77{,}11.$$

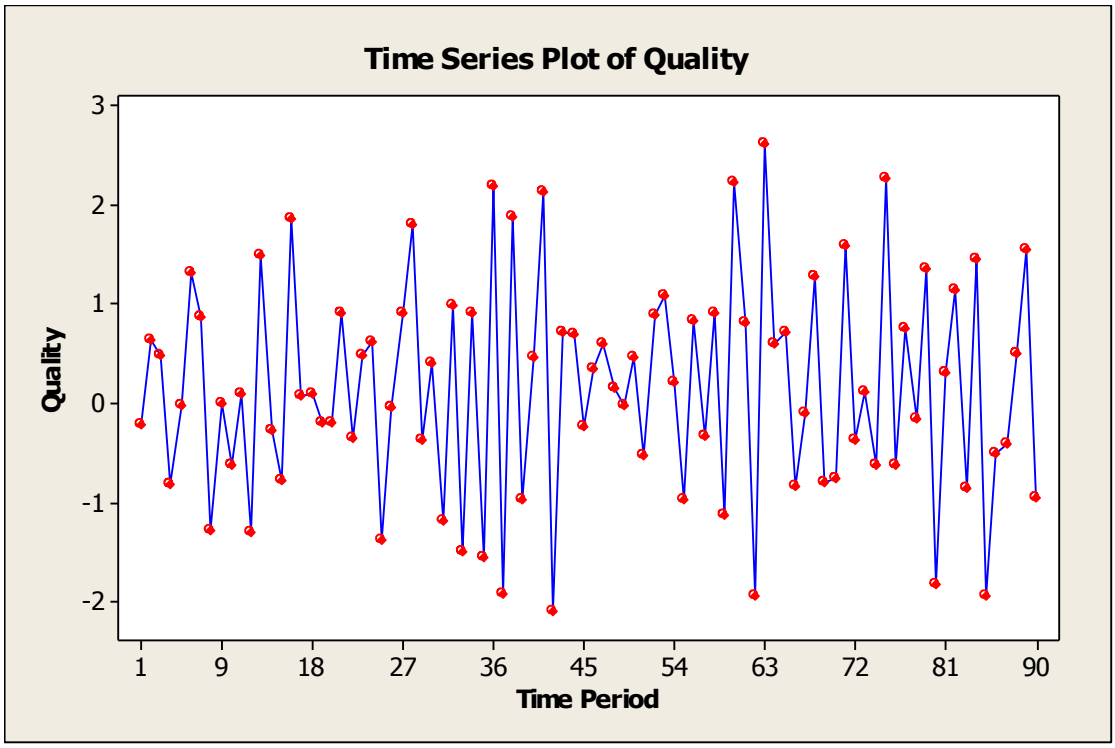
Επίσης, η δύο-βήματα-μπροστά πρόβλεψη είναι:

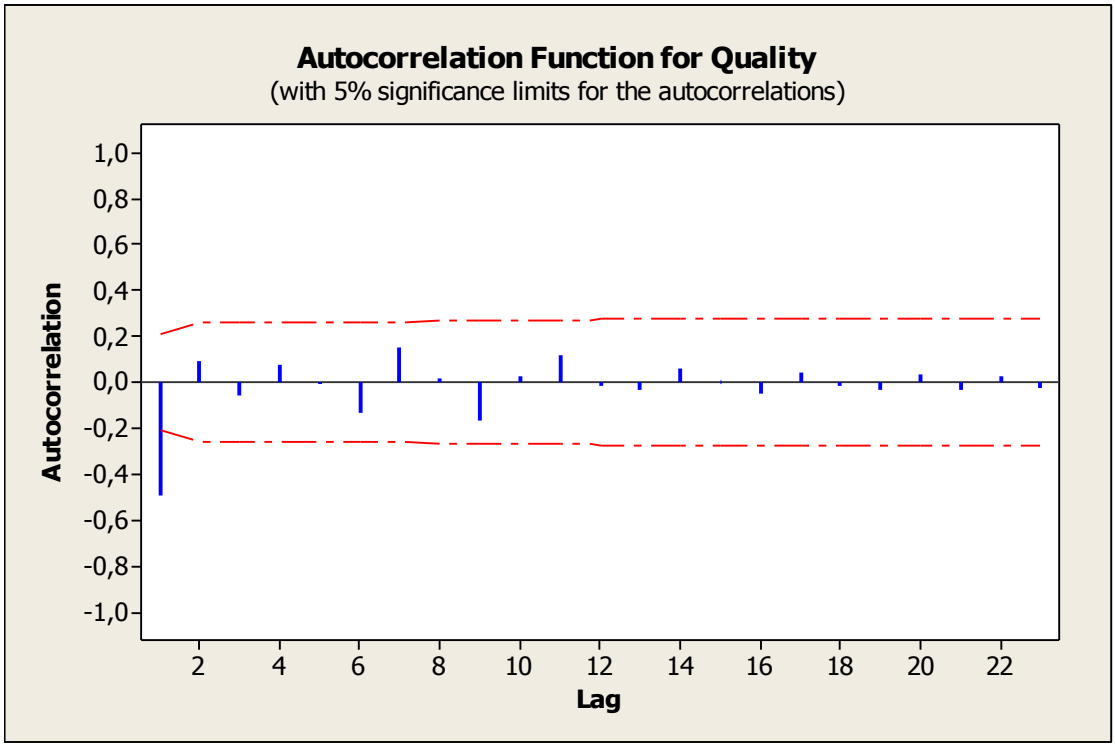
$$\hat{Y}_{77} = 115{,}842 + (-0{,}538)\, \hat{Y}_{76} = 74{,}36.$$

Παράδειγμα (Τ.9-7)

Ο αναλυτής ήταν χαρούμενος με τα αποτελέσματα των προβλέψεών του για τις χρονοσειρές των αναγνώσεων, που φαίνονται στο προηγούμενο παράδειγμα. Αποφάσισε να χρησιμοποιήσει τη Box-Jenkins μεθοδολογία για να προσπαθήσει να προβλέψει τα σφάλματα (αποκλίσεις από το στόχο), που προκύπτουν από τον έλεγχο ποιότητας μίας κατασκευαστικής διαδικασίας, η οποία είναι κάτω από την επιτήρησή του. Η χρονοσειρά των δεδομένων (σφαλμάτων) αναπαρίσταται στο ακόλουθο Γράφημα. Ξεκινάμε τη διαδικασία αναγνώρισης εξετάζοντας το διάγραμμα της χρονοσειράς σφαλμάτων και ελέγχοντας τις αυτοσυσχετίσεις και τις μερικές αυτοσυσχετίσεις που φαίνονται στα Γραφήματα. Το διάγραμμα χρονοσειράς και οι

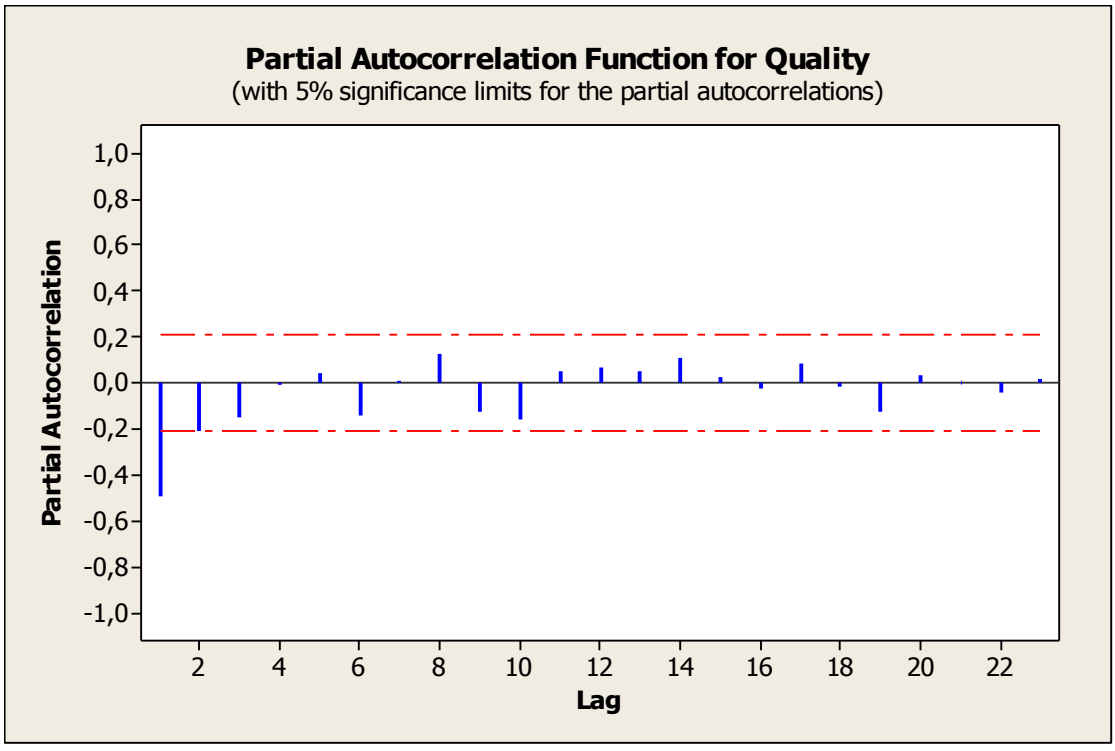
συναρτήσεις αυτοσυσχέτισης υποδεικνύουν ότι η σειρά είναι στάσιμη.





Lag	ACF	T	LBQ
1	-0,495654	-4,70	22,86
2	0,088540	0,69	23,59

3	-0,055353	-0,43	23,89
4	0,073234	0,56	24,40
5	-0,006740	-0,05	24,41
6	-0,132292	-1,02	26,13
7	0,148588	1,13	28,33
8	0,014550	0,11	28,35
9	-0,163422	-1,22	31,08
10	0,023981	0,18	31,14
11	0,116420	0,86	32,56
12	-0,020286	-0,15	32,61
13	-0,032661	-0,24	32,72
14	0,055488	0,41	33,06
15	0,000339	0,00	33,06
16	-0,049988	-0,36	33,34
17	0,040314	0,29	33,52
18	-0,019007	-0,14	33,56
19	-0,029457	-0,21	33,67
20	0,030425	0,22	33,77
21	-0,035033	-0,25	33,92
22	0,022290	0,16	33,98
23	-0,021615	-0,16	34,04



Lag	PACF	T
1	-0,495654	-4,70
2	-0,208309	-1,98

3	-0,146308	-1,39
4	-0,011216	-0,11
5	0,043804	0,42
6	-0,145627	-1,38
7	0,009354	0,09
8	0,121786	1,16
9	-0,125903	-1,19
10	-0,156293	-1,48
11	0,051968	0,49
12	0,067712	0,64
13	0,051781	0,49
14	0,108218	1,03
15	0,028373	0,27
16	-0,026551	-0,25
17	0,081170	0,77
18	-0,020338	-0,19
19	-0,125144	-1,19
20	0,029676	0,28
21	-0,000043	-0,00
22	-0,040495	-0,38
23	0,012561	0,12

Υπάρχει μία σημαντική αυτοσυσχέτιση του 0,50 στο lag 1 και οι αυτοσυσχετίσεις στις εναπομένουσες καθυστερήσεις είναι μικρές και καλά μέσα στα όρια σφάλματος, οι αυτοσυσχετίσεις δείγματος φαίνεται να διακόπτονται μετά το lag 1. Ξεκινώντας με μία σημαντική μερική αυτοσυσχέτιση στο lag 1, οι πρώτες τρεις μερικές αυτοσυσχετίσεις δείγματος είναι όλες αρνητικές και φθίνουν στο μηδέν. Οι συμπεριφορές των αυτοσυσχετίσεων δείγματος και των μερικών αυτοσυσχετίσεων δείγματος μοιάζουν σαν τις θεωρητικές ποσότητες για μία MA(1) (ARIMA(0,0,1)) διαδικασία, που φαίνεται στο Γράφημα 3(a). Ο αναλυτής πείστηκε ότι οι σειρές θα μπορούσαν να αντιπροσωπευθούν από το MA(1) μοντέλο.

Ο αναλυτής χρησιμοποίησε το Minitab για να προσαρμόσει ένα MA(1) μοντέλο στα δεδομένα του. Τα αποτελέσματα παρουσιάζονται

στον παρακάτω Πίνακα. Οι παράμετροι για το MA(1) μοντέλο εκτιμώνται να είναι $\hat{\mu}=0,1513$ *και* $\hat{\omega}_1=0,5875.$ Το κάθε ένα είναι σημαντικά διαφορετικό από το μηδέν. Οι αυτοσυσχετίσεις καταλοίπων φαίνονται στο παρακάτω Γράφημα και οι Ljung-Box Q στατιστικές υποδεικνύουν, ότι τα σφάλματα είναι τυχαία.

Η MA(1) εξίσωση πρόβλεψης είναι:

$$\hat{Y}_t=\hat{\mu}-\hat{\omega}_1\,\varepsilon_{t-1}=0,1513-0,5875\,\varepsilon_{t-1}$$

όπου το ε_{t-1} υπολογίζεται από το αντίστοιχο κατάλοιπο e_{t-1}. Για να γίνει πρόβλεψη του σφάλματος (απόκλιση από το στόχο) για την περίοδο 91, απαιτείται το κατάλοιπο για την περίοδο 90, e_{90} =-0,4804. Οπότε:

$$\hat{Y}_{91}=0,1513-0,5875(-0,4804)=0,4335.$$

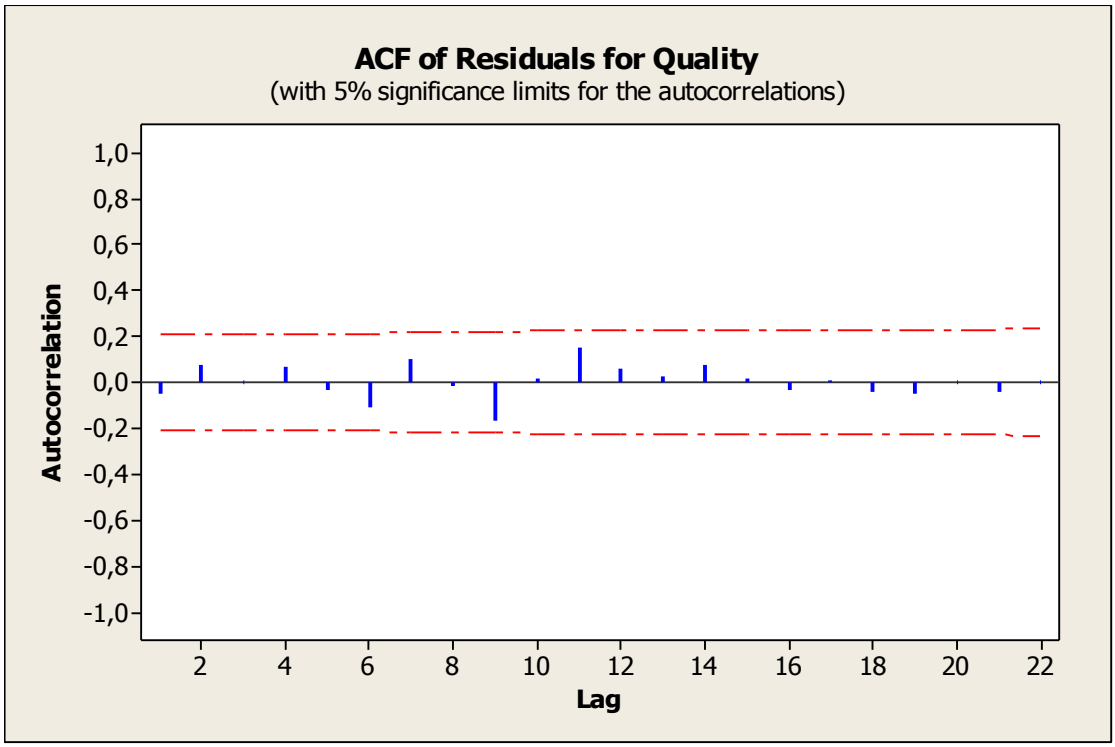
Η πρόβλεψη του σφάλματος ελέγχου ποιότητας την περίοδο 92 είναι απλά ο εκτιμώμενος μέσος των σειρών, επειδή στο προβλεπόμενο σημείο, t=91, η καλύτερη εικασία του όρου σφάλματος, που μπορούμε να κάνουμε για την περίοδο 91 ε_{91}, είναι μηδέν. Επομένως,

$$\hat{Y}_{92}=0,1513-0,5875(0)=0,1513.$$

ARIMA(0,0,1) Model: Quality Control

```
Final Estimates of Parameters

Type       Coef   SE Coef    T      P
MA 1     0,5875  0,0864   6,80 0,000
Constant 0,15129 0,04022  3,76 0,000
Mean     0,15129 0,04022

Number of observations: 90
Residuals: SS = 74,4933 (backforecasts excluded)
           MS =  0,8465 DF = 88

Modified Box-Pierce (Ljung-Box) Chi-Square
statistic
Lag         12     24   36    48
```

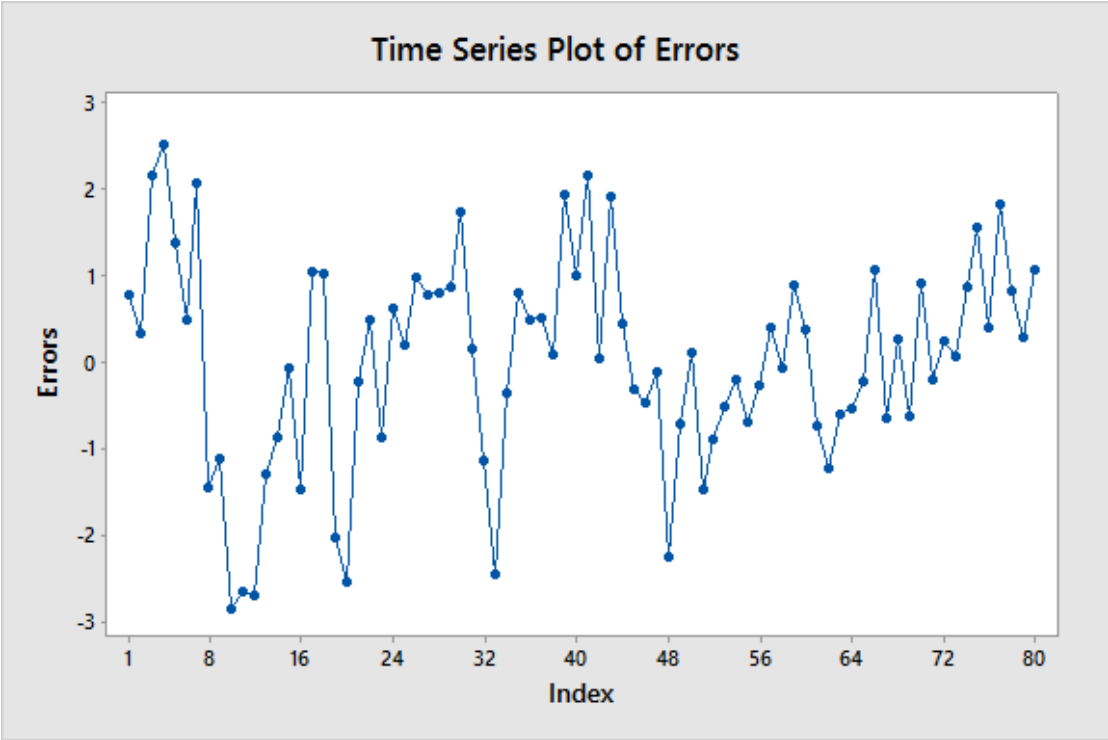
```
Chi-Square 9,1   10,8 17,3  31,5
DF          10     22   34    46
P-Value    0,524 0,977 0,992 0,950

Forecasts from period 90 95% Limits
Period   Forecast   Lower   Upper   Actual
91        0,43350 -1,37018 2,23719
92        0,15129 -1,94064 2,24322
```



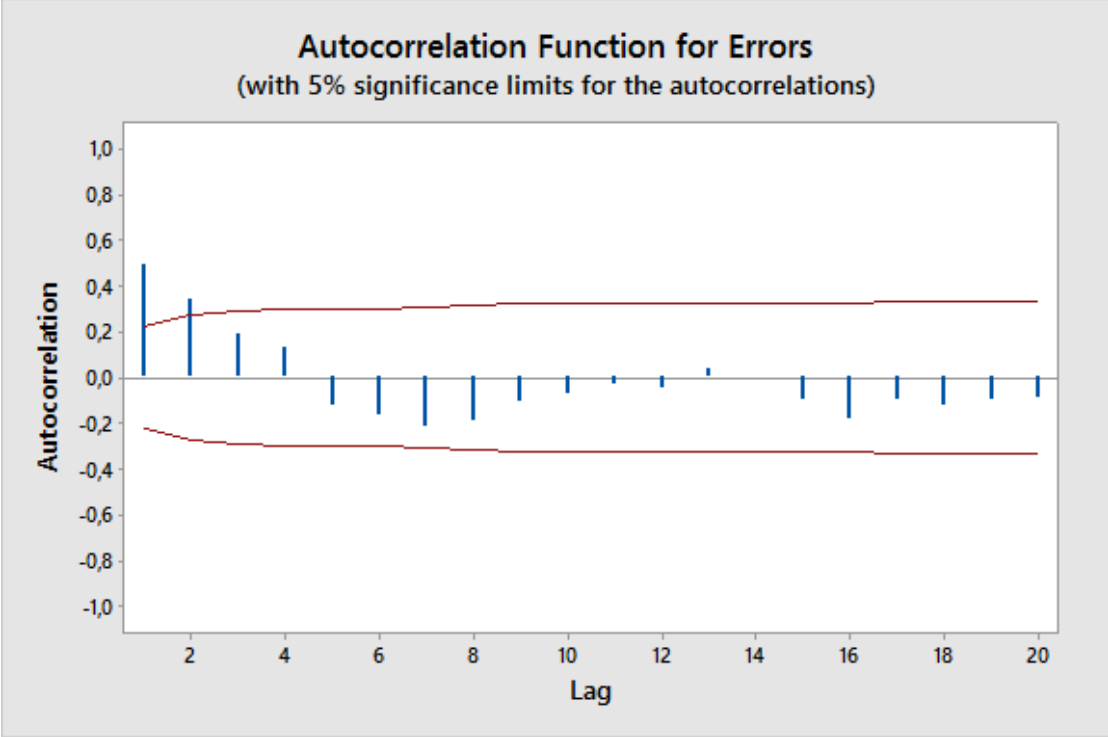
Lag	ACF	T	LBQ
1	-0,049843	-0,47	0,23
2	0,076340	0,72	0,78
3	0,003799	0,04	0,78
4	0,063375	0,60	1,17
5	-0,029768	-0,28	1,25
6	-0,112420	-1,05	2,50
7	0,101193	0,94	3,52
8	-0,017494	-0,16	3,55
9	-0,169055	-1,55	6,47
10	0,015437	0,14	6,50
11	0,146365	1,31	8,74
12	*0,056395*	*0,49*	*9,08*

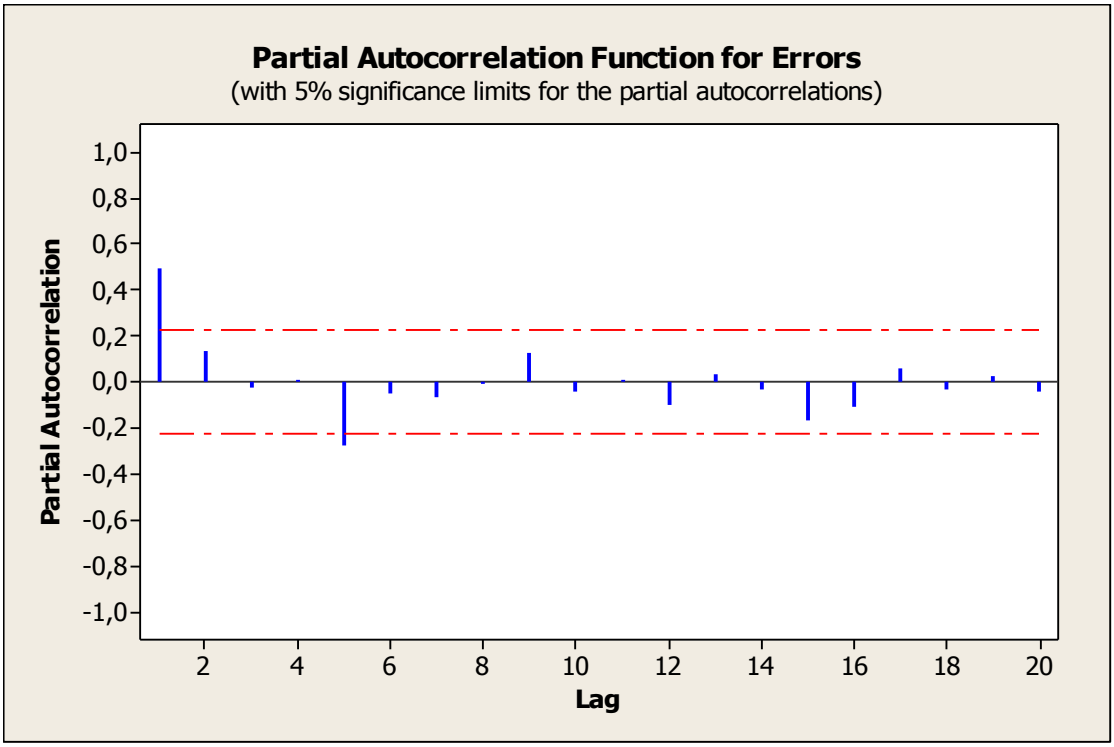
Παράδειγμα(Τ.9-9)

Ένας αναλυτής μίας άλλης εταιρίας αποφάσισε να δοκιμάσει τη Box-Jenkins μεθοδολογία για σφάλματα ποιότητας στη δική του εταιρεία. Τα δεδομένα δίνονται στον Πίνακα, και ένα διάγραμμα χρονοσειρών των δεδομένων φαίνεται στο Γράφημα.



Lag	ACF	T	LBQ
1	0,493582	4,41	20,23
2	0,346020	2,54	30,30
3	0,193314	1,32	33,48
4	0,128236	0,85	34,90
5	-0,126079	-0,83	36,29
6	-0,163715	-1,07	38,67
7	-0,219147	-1,41	42,98
8	-0,191453	-1,21	46,32
9	-0,102329	-0,63	47,29
10	-0,073527	-0,45	47,80





Lag	PACF	T
1	0,493582	4,41
2	0,135379	1,21
3	-0,028518	-0,26
4	0,010798	0,10
5	-0,274199	-2,45

6	-0,054177	-0,48
7	-0,067558	-0,60
8	-0,009173	-0,08
9	0,127521	1,14
10	-0,043168	-0,39

Το διάγραμμα των σειρών και οι αυτοσυσχετίσεις δείγματος, που φαίνονται στα αντίστοιχα Γραφήματα, υποδεικνύουν, ότι οι σειρές των σφαλμάτων ελέγχου ποιότητας είναι στάσιμες. Τα σφάλματα εμφανίζονται να διαφέρουν γύρω από ένα σταθερό επίπεδο του μηδενός, και οι αυτοσυσχετίσεις φθίνουν σχετικά γρήγορα.

Ο αναλυτής αποφάσισε, ότι οι δύο πρώτες αυτοσυσχετίσεις ήταν σημαντικά διαφορετικές απ' το μηδέν και ότι οι αυτοσυσχετίσεις για τις πρώτες μερικές χρονικές καθυστερήσεις (lags) τείνουν στο μηδέν περίπου, όπως το θεωρητικό μοτίβο για μία AR(1) διαδικασία (βλ. Γράφημα 2(a)). Επίσης εξέτασε τις μερικές αυτοσυσχετίσεις δείγματος, που φαίνονται στο αντίστοιχο Γράφημα. Όπως είχε υποπτευθεί, οι μερικές αυτοσυσχετίσεις έμοιαζαν να διακόπτονται μετά τη σημαντική μερική αυτοσυσχέτιση στο lag 1. Λαμβάνοντάς τα μαζί, τα μοτίβα αυτοσυσχέτισης δείγματος και μερικής αυτοσυσχέτισης δείγματος ήταν συνεπή με μία AR(1) διαδικασία, και ο αναλυτής ένιωσε άνετος να προσαρμόσει ένα AR(1) (ARIMA(1,0,0)) μοντέλο στις δικές του χρονοσειρές σφαλμάτων (αποκλίσεις απ' το στόχο).

Τα αποτελέσματα του Minitab για την αρχική προσπάθεια με τη Box-Jenkins μεθοδολογία δίνονται στον Πίνακα. Μιας και ο μέσος δείγματος των σειρών σφαλμάτων είναι εξαιρετικά μικρός (κοντά στο μηδέν) συγκρινόμενος με τη τυπική του απόκλιση, δεν συμπεριέλαβε ένα σταθερό όρο στο AR(1) μοντέλο του. Η παράμετρος στο AR(1) μοντέλο εκτιμάται να είναι $\hat{\varphi}_1 = 0{,}501$ και είναι σημαντικά διαφορετική από το μηδέν (t=5,11). Το μέσο τετραγωνικό σφάλμα των καταλοίπων είναι

s^2=1,1136. Διαγράμματα των καταλοίπων, τα Ljung-Box chi-square στατιστικά και οι αυτοσυσχετίσεις καταλοίπων που παρουσιάζονται στο αντίστοιχο Γράφημα προτείνουν ότι το AR(1) μοντέλο είναι επαρκές. Δεν υπάρχει λόγος για να αμφισβητηθούν οι συνήθεις υποθέσεις όρων σφάλματος.

ARIMA(1,0,0) Model: Quality Control Errors
Final Estimates of Parameters

Type	Coef	SE Coef	T	P
AR 1	0,5009	0,0986	5,08	0,000
Constant	0,0204	0,1180	0,17	0,863
Mean	0,0408	0,2365		

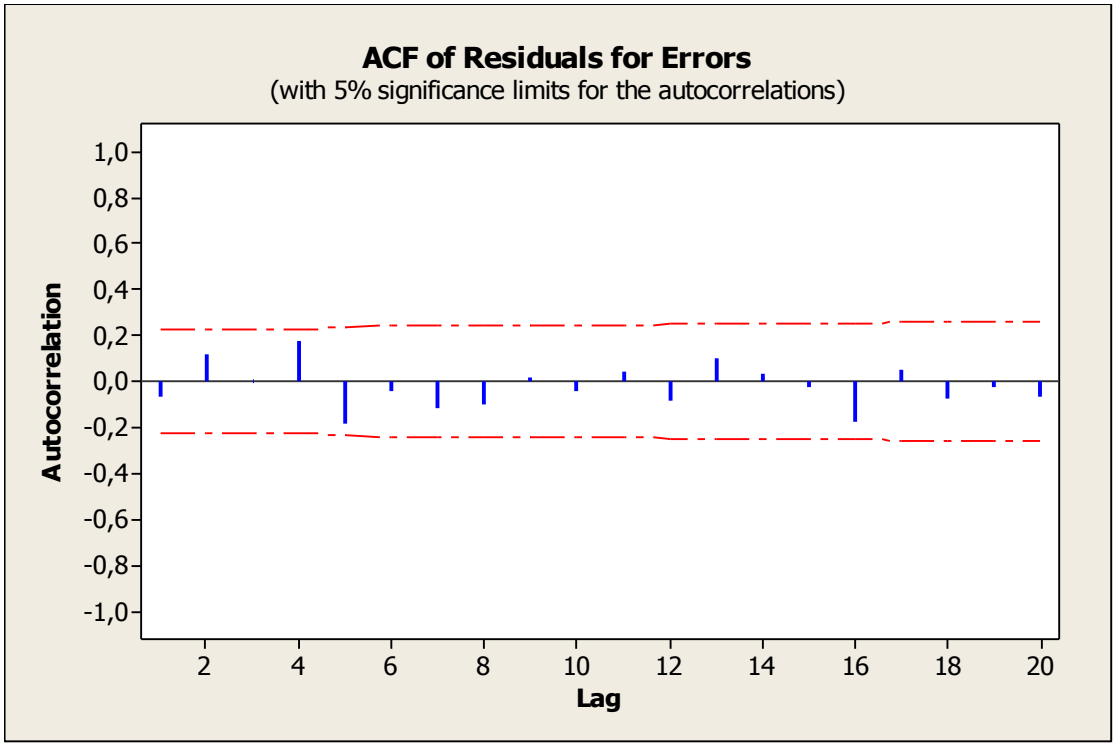
Number of observations: 80
Residuals: SS = 86,8582 (backforecasts excluded)
MS = 1,1136 DF = 78

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	10,7	19,5	36,2	44,1
DF	10	22	34	46
P-Value	0,381	0,612	0,364	0,551

Forecasts from period 80		95% Limits		
Period	Forecast	Lower	Upper	Actual
81	0,55131	-1,51741	2,62003	
82	0,29652	-2,01721	2,61024	

Lag	ACF	T	LBQ
1	-0,069795	-0,62	0,40
2	0,113142	1,01	1,48
3	0,002175	0,02	1,48
4	0,178268	1,57	4,22
5	-0,187692	-1,60	7,31
6	-0,039199	-0,32	7,44
7	-0,117368	-0,97	8,68
8	-0,103656	-0,85	9,66
9	0,013443	0,11	9,68
10	-0,037854	-0,31	9,81



Η AR(1) εξίσωση πρόβλεψης είναι:

$$\hat{Y}_t = 0{,}501\, Y_{t-1}$$

έτσι οι προβλέψεις για τις χρονικές περιόδους 81 και 82 είναι:

$$\hat{Y}_{81} = 0{,}501\, Y_{80} = 0{,}501\,(1{,}06) = 0{,}531$$

$$\hat{Y}_{82} = 0{,}501\, Y_{81} = 0{,}501\,(0{,}531) = 0{,}266$$

Οι υπολογισμένες προβλέψεις του αναλυτή συμφωνούν με τις προβλέψεις ένα-βήμα και δύο-βήματα-μπροστά, που παρήχθησαν από το Minitab.

Ο αναλυτής αποφάσισε να προσαρμόσει ένα πιο περίπλοκο μοντέλο για να δει αν τα αποτελέσματα θα επιβεβαίωναν την επιλογή του για ένα AR(1) μοντέλο για τα δεδομένα του. Αποφάσισε να προσθέσει μία επιπλέον παράμετρο και να προσαρμόσει ένα ARMA(1,1) (ARIMA(1,0,1)) μοντέλο στα σφάλματα ελέγχου ποιότητας. Αιτιολόγησε, ότι αν το AR(1) μοντέλο του ήταν σωστό, η παράμετρος κινούμενου μέσου στο ARIMA(1,0,1) μοντέλο δεν θα ήταν σημαντική.

Ο παρακάτω Πίνακας δείχνει τα αποτελέσματα για την προσαρμογή του ARIMA(1,0,1) μοντέλου στις σειρές του. Ο αναλυτής ήταν ικανοποιημένος, που είδε, ότι η MA(1) παράμετρος δεν ήταν σημαντικά διαφορετική από το μηδέν (t=1,04), υπονοώντας ότι αυτός ο όρος δεν χρειαζόταν. Φυσικά, μιας και είναι ένα ελαφρώς πιο γενικό μοντέλο από το AR(1) μοντέλο, ακόμα αναπαριστά επαρκώς τα δεδομένα όπως υποδεικνύεται από το s^2=1,1097 και τη τυχαία συμπεριφορά των καταλοίπων. **(Χωρίς τον σταθερό όρο η διασπορά των καταλοίπων είναι ελαφρώς μικρότερη και στις δύο περιπτώσεις. Θα πρέπει να επανεκτιμηθεί το μοντέλο χωρίς τον σταθερό όρο γιατί είναι στατιστικά μη σημαντικός.)**

ARIMA(1,0,1) Model: Quality Control Errors
Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	0,6679	0,1636	4,08	0,000
MA	1	0,2215	0,2133	1,04	0,302
Constant		0,01758	0,09180	0,19	0,849
Mean		0,0529	0,2764		

Number of observations: 80
Residuals: SS = 85,4476 (backforecasts excluded)
MS = 1,1097 DF = 77

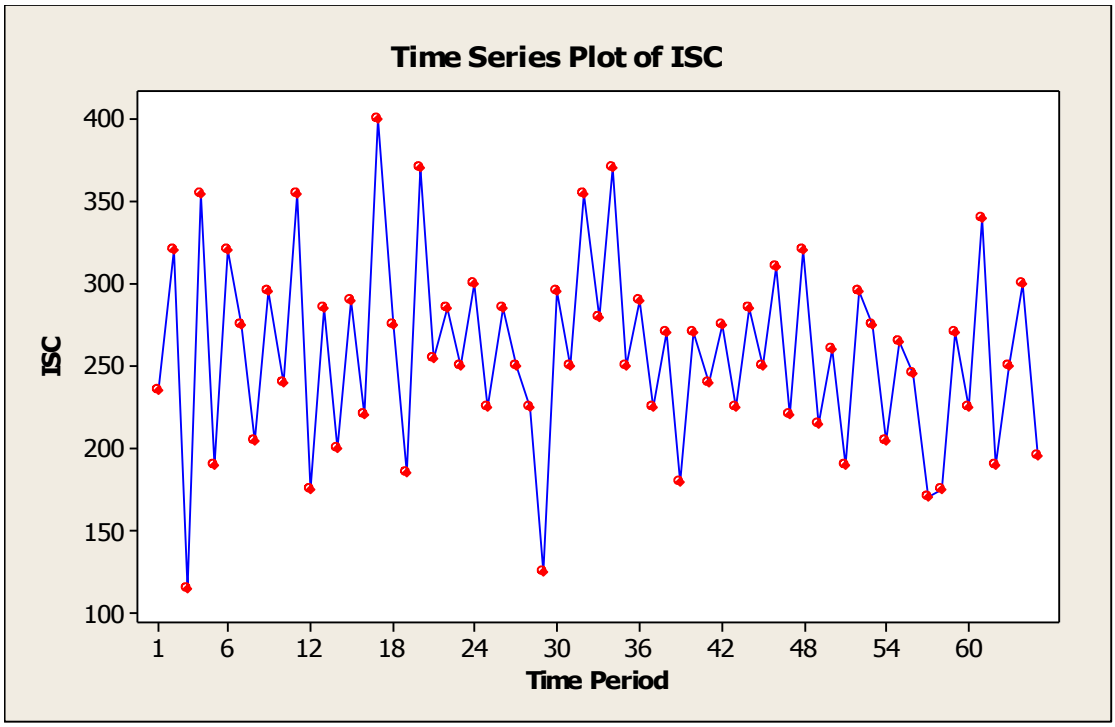
Modified Box-Pierce(Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	8,8	17,5	32,2	38,6
DF	9	21	33	45
P-Value	0,453	0,682	0,506	0,737

Forecasts from period 80		95% Limits		
Period	Forecast	Lower	Upper	Actual
81	0,55086	-1,51428	2,61599	
82	0,38548	-1,87608	2,64704	

Παράδειγμα(Τ.9-12)

Μία εταιρία ζήτησε από τον αναλυτή της να κάνει προβλέψεις για τις τιμές κλεισίματος της μετοχής της. Οι τιμές των μετοχών φαίνονται στο παρακάτω Γράφημα.

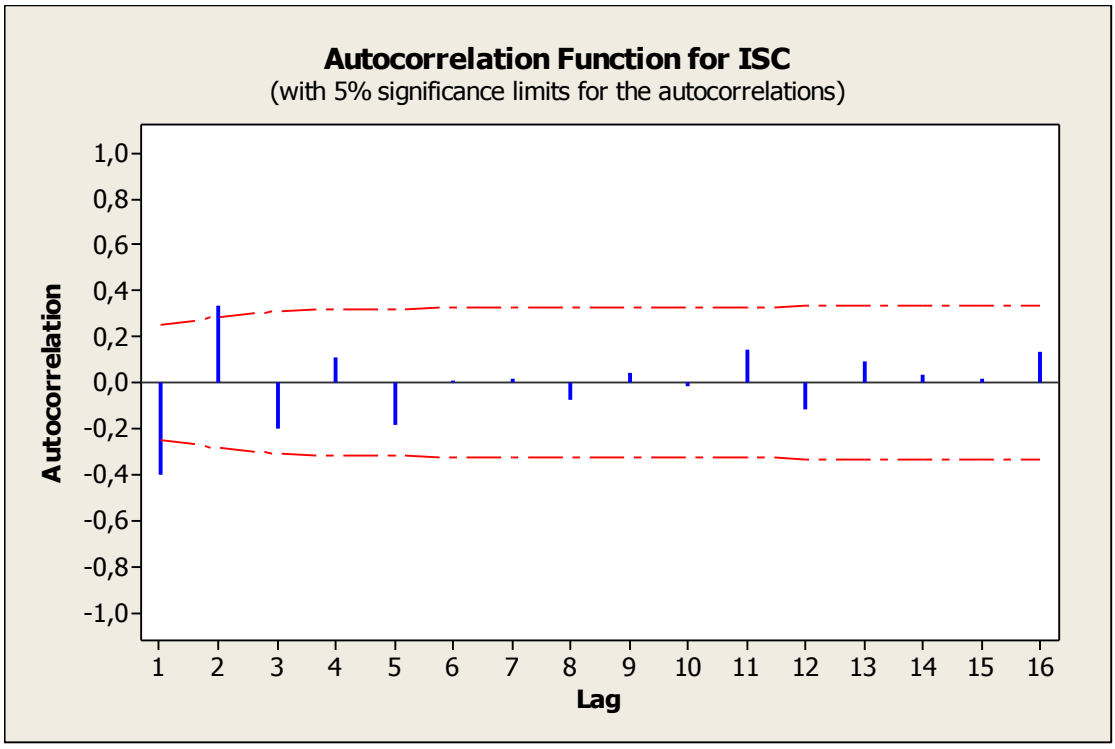


Κάτι που προκαλεί έκπληξη, μιας και οι τιμές των μετοχών τείνουν να είναι μη στάσιμες, είναι ότι το διάγραμμα των τιμών της μετοχής τείνει να διαφέρει γύρω από ένα σταθερό επίπεδο περίπου 250 (βλ. Γράφημα). Ο αναλυτής συμπέρανε, ότι οι τιμές της σειράς είναι στάσιμες.

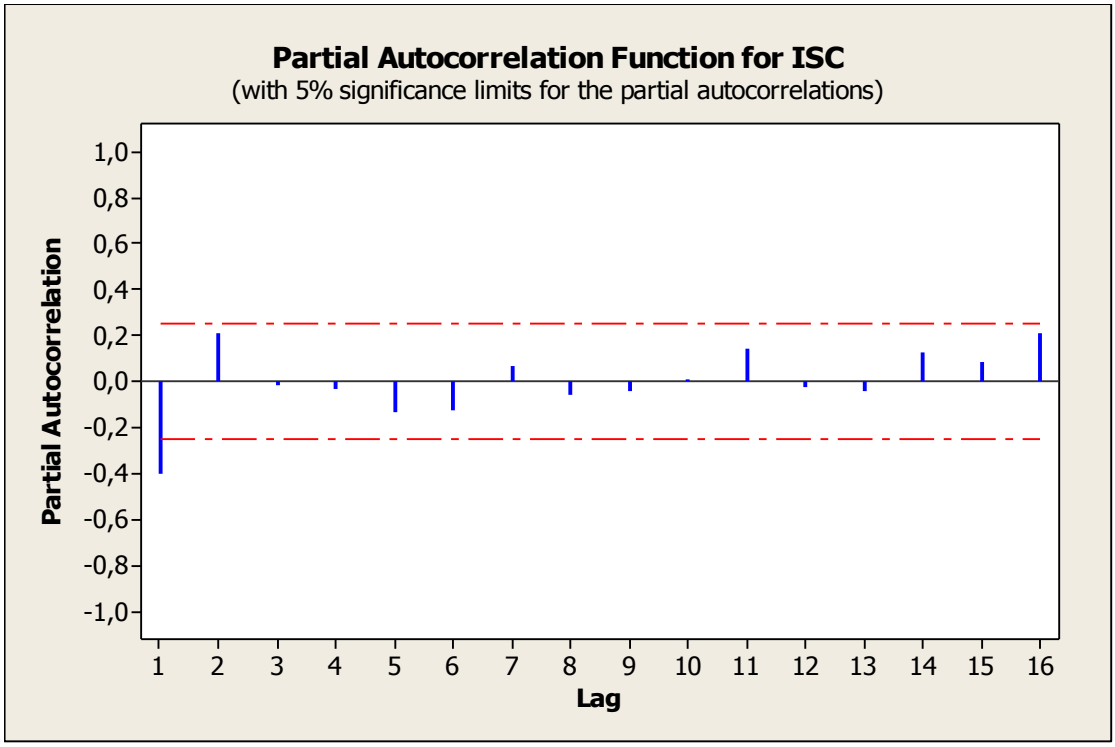
Ο αναλυτής πιστεύει ότι η Box-Jenkins μεθοδολογία μπορεί να ταιριάζει με το σκοπό του και αμέσως παρήγαγε τις αυτοσυσχετίσεις δείγματος και τις μερικές αυτοσυσχετίσεις δείγματος για τις σειρές των τιμών της μετοχής.

Ο αναλυτής πρόσεξε ότι οι αυτοσυσχετίσεις δείγματος αλλάζουν πρόσημο και μειώνονται στο μηδέν για μικρές καθυστερήσεις (lags). Οι μερικές αυτοσυσχετίσεις δείγματος είναι παρόμοιες αλλά ξεκάθαρα

διακόπτονται (είναι μηδέν) μετά το lag 2. Ο αναλυτής θυμάται τα πιθανά μοτίβα για τις θεωρητικές αυτοσυσχετίσεις και τις μερικές αυτοσυσχετίσεις για μία AR(2) διαδικασία (βλ. Γράφημα 9-2(d)) και αποφασίζει ότι οι αντίστοιχες ποσότητες δείγματος είναι συνεχείς με κάποιο από αυτά τα μοτίβα. Αναγνωρίζει ένα AR(2) (ARIMA(2,0,0)) μοντέλο για τα δεδομένα του.



Lag	ACF	T	LBQ
1	-0,401525	-3,24	10,97
2	0,333244	2,34	18,65
3	-0,204772	-1,33	21,59
4	0,111059	0,70	22,47
5	-0,182873	-1,15	24,90
6	0,010786	0,07	24,91



Lag	PACF	T
1	-0,401525	-3,24
2	0,205086	1,65
3	-0,019988	-0,16
4	-0,034493	-0,28
5	-0,136220	-1,10
6	-0,126544	-1,02
7	0,069713	0,56
8	-0,056123	-0,45
9	-0,044177	-0,36
10	0,008517	0,07

Ο αναλυτής χρησιμοποίησε το Minitab για να εκτιμήσει τις παραμέτρους στο μοντέλο του. Συμπεριέλαβε και ένα σταθερό όρο για να επιτρέψει ένα μη μηδενικό επίπεδο. Οι εκτιμήσεις παραμέτρων είναι: $\hat{\varphi}_0=284{,}9$, $\hat{\varphi}_1=-0{,}324$ *και* $\hat{\varphi}_2=0{,}219$. Ο εκτιμώμενος συντελεστής δεν είναι σημαντικός (t=1,75) στο 5% επίπεδο, αλλά είναι σημαντικός στο 10% επίπεδο. Ο αναλυτής αποφάσισε να κρατήσει αυτή την παράμετρο στο μοντέλο του. Ο τετραγωνικός μέσος καταλοίπων είναι $s^2=2.808$ και οι αυτοσυσχετίσεις καταλοίπων, που δίνονται στο

αντίστοιχο Γράφημα, είναι καλά μέσα στα δύο όρια τυπικών σφαλμάτων. Επιπλέον, οι τιμές των Ljung-Box Q στατιστικών για m=12, 24, 36 και 48 lags είναι όλες μικρές. Ο αναλυτής αποφάσισε ότι το μοντέλο του είναι επαρκές.

Ο αναλυτής χρησιμοποιεί το μοντέλο του για να δημιουργήσει προβλέψεις για τις περιόδους 66 και 67, όπως ακολουθεί. Από την τιμή t=65, η πρόβλεψη για την περίοδο 66 παράγεται από την εξίσωση:

$$\hat{Y}_t = \hat{\varphi}_0 + \hat{\varphi}_1 Y_{t-1} + \hat{\varphi}_2 Y_{t-2}$$

οπότε:

$$\hat{Y}_{66} = 284{,}9 - 0{,}324\, Y_{65} + 0{,}219\, Y_{64} = 287{,}4$$

$$\hat{Y}_{67} = 284{,}9 - 0{,}324\, Y_{66} + 0{,}219\, Y_{65} = 234{,}5$$

Τα αποτελέσματα συμφωνούν με εκείνα που φαίνονται στα αποτελέσματα του Minitab. Για στάσιμες σειρές, τα 95% όρια πρόβλεψης είναι περίπου $\hat{Y} \pm 1{,}96\, s$, όπου το $\hat{Y}$ είναι η πρόβλεψη και το s είναι η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος. Για παράδειγμα, τα 95% όρια πρόβλεψης περίπου για την περίοδο 66 είναι:

$$287{,}4 \pm 1{,}96\sqrt{2.808} = 287{,}4 \pm 103{,}4 \text{ ή } (183{,}0, 390{,}8)$$

Το διάστημα είναι κοντά στο 95% διάστημα που δίνεται από το Minitab για την περίοδο 66 στον Πίνακα. **(Κανονικά θα έπρεπε να πάρουμε ένα AR(1) μοντέλο διότι ο συντελεστής του AR(2) είναι στατιστικά μη σημαντικός. Όπως είδαμε και στη συνάρτηση του μερικού συντελεστή αυτοσυσχέτισης, ο μερικός συντελεστής αυτοσυσχέτισης δεύτερης τάξης είναι μέσα στα όρια ισχύος της μηδενικής υπόθεσης).**

ARIMA(2,0,0) Model: Quality Control Errors

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	-0,3243	0,1246	-2,60	0,012
AR	2	0,2192	0,1251	1,75	0,085
Constant		284,903	6,573	43,34	0,000
Mean		257,828	5,949		

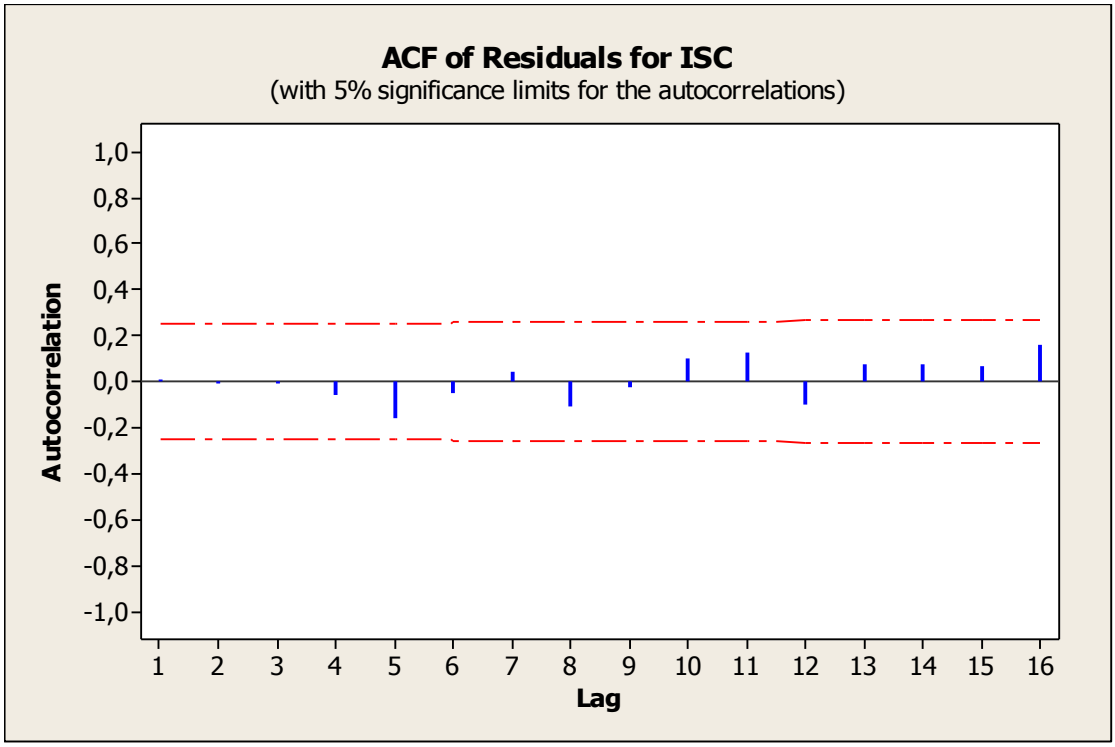
Number of observations: 65
Residuals: SS = 174093 (backforecasts excluded)
MS = 2808 DF = 62

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	6,3	13,3	18,2	29,1
DF	9	21	33	45
P-Value	0,707	0,899	0,983	0,969

Forecasts from period 65 95% Limits

Period	Forecast	Lower	Upper	Actual
66	287,446	183,565	391,328	
67	234,450	125,244	343,656	



Lag	ACF	T	LBQ
1	0,010050	0,08	0,01

2	-0,010389	-0,08	0,01
3	-0,004837	-0,04	0,02
4	-0,060200	-0,49	0,27
5	-0,160907	-1,29	2,15
6	-0,049459	-0,39	2,33
7	0,041586	0,32	2,46
8	-0,108359	-0,85	3,36
9	-0,023229	-0,18	3,40
10	0,100117	0,77	4,20
11	0,124524	0,95	5,45
12	-0,103614	-0,78	6,33

9.2.6 ΠΕΡΙΟΡΙΣΤΙΚΟΙ ΟΡΟΙ

Στην ARIMA μοντελοποίηση, δεν είναι καλή τακτική να συμπεριλαμβάνουμε AR και MA παραμέτρους, για να καλύπτουμε όλες τις πιθανότητες, που προτείνονται από τις συναρτήσεις αυτοσυσχέτισης δείγματος και μερικής αυτοσυσχέτισης δείγματος. Αυτό σημαίνει, όταν υπάρχει αμφιβολία, ξεκινάμε με ένα μοντέλο, που περιέχει μερικές αντί για πολλές παραμέτρους. Η ανάγκη για επιπλέον παραμέτρους θα γίνει προφανής από μία εξέταση των αυτοσυσχετίσεων καταλοίπων και των μερικών αυτοσυσχετίσεων. Αν η MA συμπεριφορά είναι προφανής στις αυτοσυσχετίσεις καταλοίπων και τις μερικές αυτοσυσχετίσεις, προσθέτουμε μία MA παράμετρο και προσαρμόζουμε το αναθεωρημένο μοντέλο. Αν οι αυτοσυσχετίσεις καταλοίπων μοιάζουν με εκείνες της AR διαδικασίας, προσθέτουμε έναν AR όρο και μετά αναπροσαρμόζουμε το μοντέλο.

Οι εκτιμήσεις ελαχίστων τετραγώνων των αυτοπαλινδρομικών παραμέτρων και των παραμέτρων κινούμενου μέσου στα ARIMA μοντέλα τείνουν να είναι υψηλά συσχετισμένες. Όταν υπάρχουν περισσότερες παράμετροι απ’ όσες είναι απαραίτητες, αυτό οδηγεί σε ασταθή μοντέλα, που παράγουν φτωχές προβλέψεις.

Για να συνοψίσουμε, είναι καλή πρακτική να ξεκινάμε με ένα μικρό αριθμό ξεκάθαρα αιτιολογημένων παραμέτρων και να προσθέτουμε μία παράμετρο τη φορά, όταν χρειάζεται. Από την άλλη πλευρά, αν οι παράμετροι σε ένα προσαρμοσμένο ARIMA μοντέλο δεν είναι σημαντικές (όπως αυτό κρίνεται από τους t λόγους), διαγράφουμε μία παράμετρο τη φορά και αναπροσαρμόζουμε το μοντέλο. Εξαιτίας της υψηλής συσχέτισης μεταξύ των εκτιμώμενων παραμέτρων, μπορεί να έχουμε την περίπτωση, κατά την οποία μία προηγουμένως μη σημαντική παράμετρος γίνεται σημαντική.

9.2.7 ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ

Τα ARIMA μοντέλα αναγνωρίζονται (επιλέγονται) κοιτάζοντας ένα διάγραμμα των σειρών και ταιριάζοντας μοτίβα αυτοσυσχέτισης δείγματος και μερικής αυτοσυσχέτισης δείγματος με τα γνωστά θεωρητικά μοτίβα των ARIMA διαδικασιών. Ωστόσο, υπάρχει κάποια αντικειμενικότητα σε αυτή τη διαδικασία, και είναι πιθανό, ότι δύο (ή περισσότερα) αρχικά μοντέλα να είναι συνεπή με τα μοτίβα της αυτοσυσχέτισης δείγματος και της μερικής αυτοσυσχέτισης δείγματος. Επιπλέον, μετά την εκτίμηση και τον έλεγχο, μπορεί και τα δύο μοντέλα να αναπαριστούν επαρκώς τα δεδομένα. Αν τα μοντέλα περιέχουν τον ίδιο αριθμό παραμέτρων, το μοντέλο με το μικρότερο μέσο τετραγωνικό σφάλμα, s^2, συνήθως προτιμάται. Αν τα μοντέλα περιέχουν διαφορετικό αριθμό παραμέτρων, η αρχή της οικονομίας μας οδηγεί στην επιλογή του απλούστερου μοντέλου. Ωστόσο, το μοντέλο με τις περισσότερες παραμέτρους μπορεί να έχει αισθητά μικρότερο μέσο τετραγωνικό σφάλμα.

Έχει αναπτυχθεί μία προσέγγιση στην επιλογή μοντέλου, που θεωρεί και την προσαρμογή μοντέλου και τον αριθμό των παραμέτρων.

Το κριτήριο πληροφόρησης του Akaike (1974), ή AIC, επιλέγει το καλύτερο μοντέλο από ένα σύνολο υποψήφιων μοντέλων σαν αυτό που ελαχιστοποιεί:

$$AIC = \ln \hat{\sigma}^2 + \frac{2}{n} r$$

όπου:

- $\hat{\sigma}^2$ = το άθροισμα των τετραγώνων των καταλοίπων **διαιρεμένο** με τον αριθμό των **παρατηρήσεων n** και **όχι με τους βαθμούς ελευθερίας**.
- n = ο αριθμός των παρατηρήσεων.
- r = ο συνολικός αριθμός των παραμέτρων μαζί με τον σταθερό όρο στο μοντέλο ARIMA.

Το Bayesian κριτήριο πληροφόρησης, που αναπτύχθηκε από τον Schwarz (1978), ή BIC, επιλέγει το μοντέλο, που ελαχιστοποιεί το:

$$BIC = \ln \hat{\sigma}^2 + \frac{\ln n}{n} r$$

Ο δεύτερος όρος στο AIC και στο BIC είναι ένας παράγοντας που «τιμωρεί» τη χρησιμοποίηση μιας επιπλέον παραμέτρου στο μοντέλο. Δεδομένου ότι το BIC κριτήριο επιβάλλει μία μεγαλύτερη τιμωρία για τον αριθμό των παραμέτρων από το AIC κριτήριο, η χρήση του ελάχιστου BIC για την επιλογή μοντέλου θα οδηγήσει σε ένα μοντέλο, του οποίου ο αριθμός των παραμέτρων δεν είναι μεγαλύτερος από αυτόν, που επιλέχθηκε από το AIC. Συχνά τα δύο κριτήρια παράγουν το ίδιο αποτέλεσμα. Για n=8 τα δύο κριτήρια περίπου συμπίπτουν.

Τα AIC και BIC θα πρέπει να αντιμετωπίζονται σαν επιπρόσθετες διαδικασίες, που βοηθούν στην επιλογή μοντέλου. Δεν πρέπει να χρησιμοποιούνται σαν υποκατάστατα για μία προσεκτική εξέταση των

αυτοσυσχετίσεων και μερικών αυτοσυσχετίσεων δείγματος.

Παράδειγμα

Στο Παράδειγμα (Τ.9-5), βρήκαμε, ότι δύο ARIMA μοντέλα εμφανίζονταν να παρέχουν μία επαρκή περιγραφή των αναγνώσεων για τη διαδικασία. Το ένα ήταν το AR(1) μοντέλο με r = 2 εκτιμώμενους παραμέτρους (μαζί με τον σταθερό όρο) και $\hat{\sigma}^2$ = 10.065/75 = 134,2. Το δεύτερο μοντέλο ήταν ένα MA(2) μοντέλο με r = 3 παραμέτρους (μαζί με τον σταθερό όρο) και $\hat{\sigma}^2$ = 9.725/75 = 129,7. Ο αναλυτής υπολόγισε τα AIC και BIC κριτήρια ως εξής:

AR(1):

$$AIC = \ln 134,2 + \frac{2}{75}(2) = 4,9526.$$

$$BIC = \ln 134,2 + \frac{\ln 75}{75}(2) = 5,0144.$$

MA(2):

$$AIC = \ln 129,7 + \frac{2}{75}(3) = 4,9452.$$

$$BIC = \ln 129,7 + \frac{\ln 75}{75}(3) = 5,0379.$$

Τα AIC και BIC δίνουν αντικρουόμενα αποτελέσματα. Το AIC είναι μικρότερο για το MA(2) μοντέλο, ενώ το BIC είναι μικρότερο για το AR(1) μοντέλο. Ο αναλυτής θυμήθηκε ότι το BIC δεν θα επιλέξει ένα μοντέλο με περισσότερες παραμέτρους από το AIC εξαιτίας της μεγαλύτερης τιμωρίας για επιπλέον παραμέτρους. Ο αναλυτής ήταν χαρούμενος να δει, ότι η αρχική του επιλογή του AR(1) μοντέλου, βασιζόμενη στην αρχή της οικονομίας, υποστηρίχτηκε από το BIC κριτήριο.

9.2.8 ARIMA ΜΟΝΤΕΛΑ ΓΙΑ ΕΠΟΧΙΚΑ ΔΕΔΟΜΕΝΑ

Τα εποχικά δεδομένα έχουν ένα ευδιάκριτο μοτίβο, που επαναλαμβάνεται κάθε χρόνο. Για μηνιαία δεδομένα με ένα ετήσιο εποχικό μοτίβο, πρέπει να συσχετίζονται οι παρατηρήσεις για τους ίδιους μήνες σε διαφορετικές χρονιές. Αυτό σημαίνει, ότι το μέγεθος του Ιανουαρίου τον ένα χρόνο πρέπει να είναι παρόμοιο με αυτόν του Ιανουαρίου του επόμενου χρόνου, του Φεβρουαρίου τον ένα χρόνο πρέπει να είναι παρόμοιο με αυτόν του Φεβρουαρίου του επόμενου χρόνου και ούτω καθεξής. Έτσι, όχι μόνο υπάρχουν παρατηρήσεις μέσα στο χρόνο, που σχετίζονται μεταξύ τους (correlated), αλλά υπάρχουν επίσης παρατηρήσεις μεταξύ χρόνων, που σχετίζονται μεταξύ τους. Αν το μήκος της εποχικής περιόδου είναι S, έτσι ώστε S=12 για μηνιαία δεδομένα και S=4 για τριμηνιαία δεδομένα, οι αυτοσυσχετίσεις και μερικές αυτοσυσχετίσεις για εποχικές διαδικασίες είναι μη μηδενικές για χαμηλά lags (within-year association) και για lags, που είναι πολλαπλάσια της εποχικής περιόδου S (between-year association). Οι ερμηνείες της αυτοσυσχέτισης και της μερικής αυτοσυσχέτισης για εποχικές καθυστερήσεις (lags) είναι οι ίδιες με τις ερμηνείες της αυτοσυσχέτισης και της μερικής αυτοσυσχέτισης για χαμηλές καθυστερήσεις (lags).

Εποχικά ARIMA μοντέλα περιέχουν κανονικούς αυτοπαλινδρομικούς όρους και όρους κινούμενου μέσου, που εξηγούν τη συσχέτιση σε χαμηλά lags και εποχικούς αυτοπαλινδρομικούς όρους και όρους κινούμενου μέσου, που εξηγούν τη συσχέτιση σε εποχικά lags. Επιπλέον, για μη στάσιμες εποχικές σειρές, μία επιπλέον εποχική διαφορά απαιτείται συχνά για να καθοριστεί ολοκληρωτικά το μοντέλο.

Τα εποχικά ARIMA μοντέλα σημειογραφικά αποτυπώνονται με έξι

παραμέτρους:

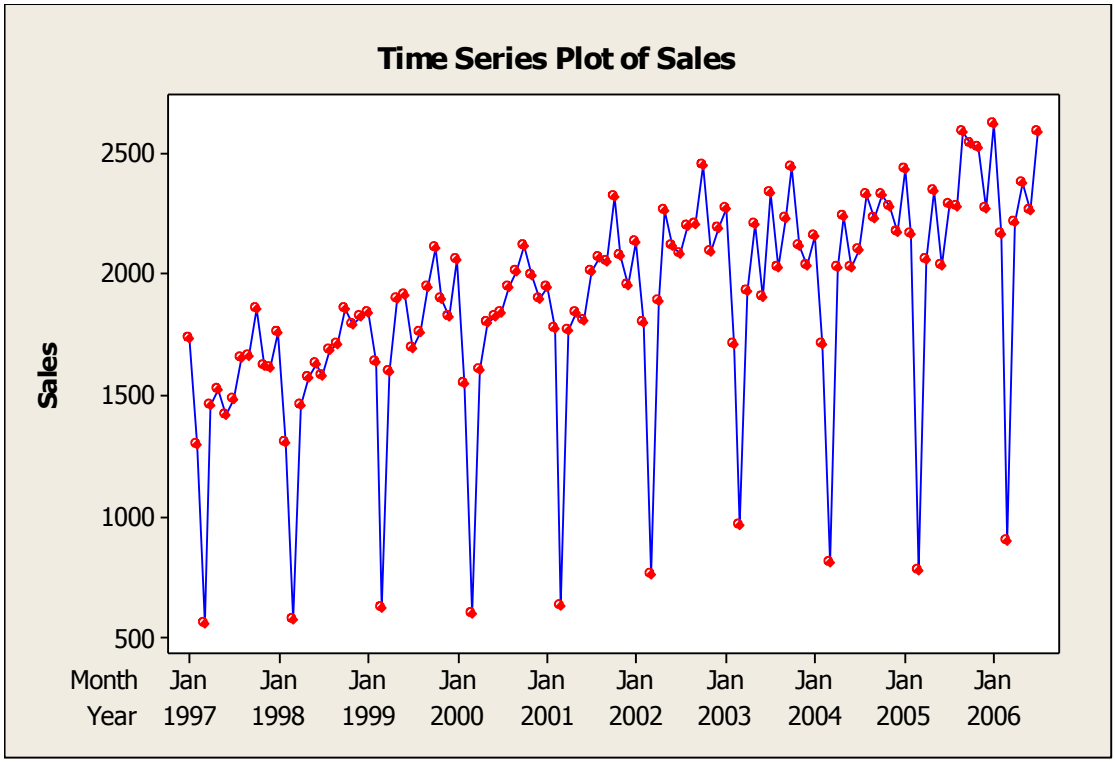
- p: κανονικοί αυτοπαλινδρομικοί όροι
- d: κανονικές διαφορές
- q: κανονικοί όροι κινούμενου μέσου
- P: εποχικοί αυτοπαλινδρομικοί όροι
- D: εποχικές διαφορές
- Q: εποχικοί όροι κινούμενου μέσου

Οπότε ο συμβολισμός είναι ARIMA(p,d,q)$(P,D,Q)_s$ με s την εποχικότητα.

Το επόμενο παράδειγμα απεικονίζει την ανάπτυξη των εποχικών ARIMA μοντέλων. Η ίδια στρατηγική χτισίματος-μοντέλου, που χρησιμοποιήθηκε για μη εποχικά μοντέλα, χρησιμοποιείται τώρα για εποχικά μοντέλα. Όταν ένα επαρκές μοντέλο εδραιωθεί, μπορεί να χρησιμοποιηθεί για πρόβλεψη.

Παράδειγμα (Τ.9-14)

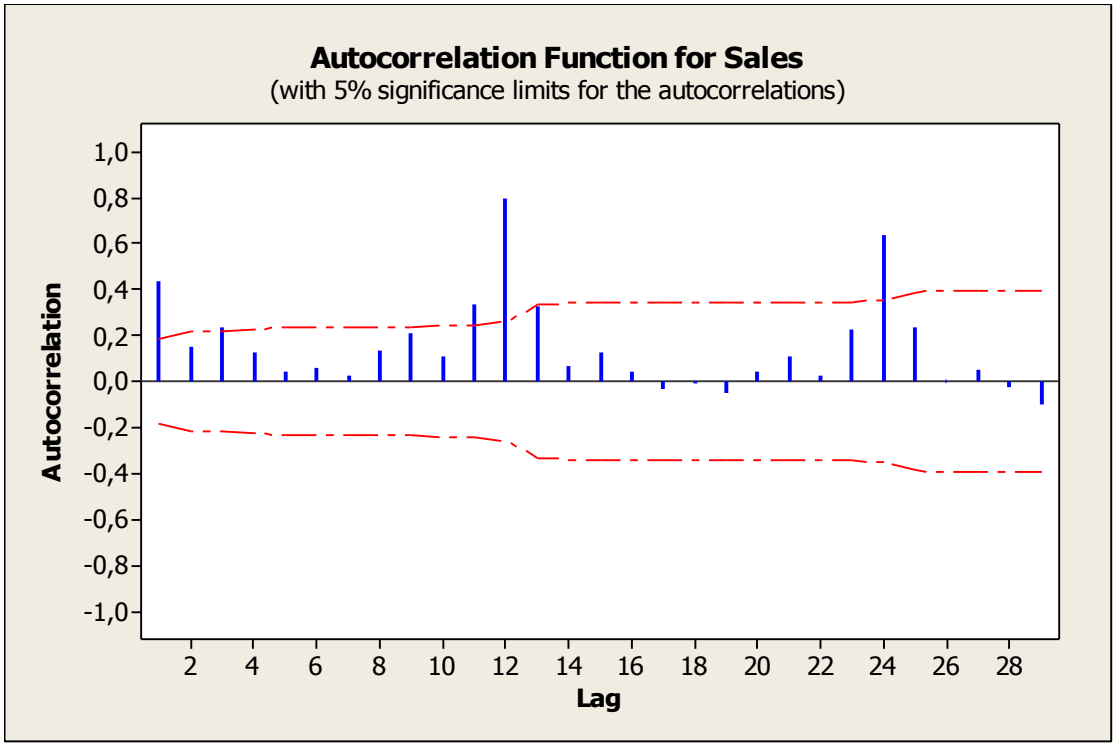
Μία εταιρία ανέθεσε σε έναν υπάλληλό της να προβλέψει τις πωλήσεις της. Ο υπάλληλος κατάφερε να συγκεντρώσει 115 συνεχόμενους μήνες δεδομένων πωλήσεων.



Ο υπάλληλος εξέτασε το διάγραμμα χρονοσειρών, που φαίνεται στο παραπάνω Γράφημα και παρατήρησε το ευδιάκριτο εποχικό μοτίβο μαζί με μία γενική ανοδική τάση. Είδε ότι οι σειρές ήταν μη στάσιμες και αποφάσισε να αναπτύξει ένα εποχικό ARIMA μοντέλο για τα δεδομένα του.

Ξεκίνησε την αναγνώριση μοντέλου κοιτάζοντας τις αυτοσυσχετίσεις δείγματος, που φαίνονται στο αντίστοιχο Γράφημα. Είδε ότι οι αυτοσυσχετίσεις σε μικρές καθυστερήσεις (lags) εμφανίζονταν να διακόπτονται μετά το lag 1, αν και υπήρχε μία μικρή αλλαγή στο lag 3. Επίσης είδε ότι οι αυτοσυσχετίσεις σε εποχικά lags – 12, 24 και 36 – ήταν μεγάλες και δεν έφθιναν γρήγορα. Αυτό υποδείκνυε ότι οι σειρές ήταν μη στάσιμες και επιβεβαίωσαν την εντύπωση, που είχε από το διάγραμμα χρονοσειρών. Πριν συνεχίσει, αποφάσισε να εφαρμόσει τη μέθοδο διαφορών στις σειρές λαμβάνοντας υπ' όψιν τις εποχικές χρονικές καθυστερήσεις για να δει αν μπορούσε να μετατρέψει τις μη

στάσιμες σειρές σε στάσιμες.



Lag	ACF	T	LBQ
1	0,431371	4,63	21,96
2	0,153911	1,41	24,78
3	0,237841	2,14	31,58
4	0,128583	1,11	33,58
5	0,041549	0,36	33,79
6	0,061352	0,53	34,26
7	0,028549	0,24	34,36
8	0,136014	1,16	36,69
9	0,212966	1,80	42,44
10	0,105905	0,87	43,88
11	0,332651	2,71	58,20
12	0,791934	6,08	140,12
13	0,324223	1,94	153,99
14	0,067091	0,39	154,59
15	0,127275	0,74	156,77
16	0,039851	0,23	156,99
17	-0,031030	-0,18	157,12
18	-0,012252	-0,07	157,14
19	-0,049556	-0,29	157,48
20	0,041444	0,24	157,73
21	0,107932	0,62	159,39

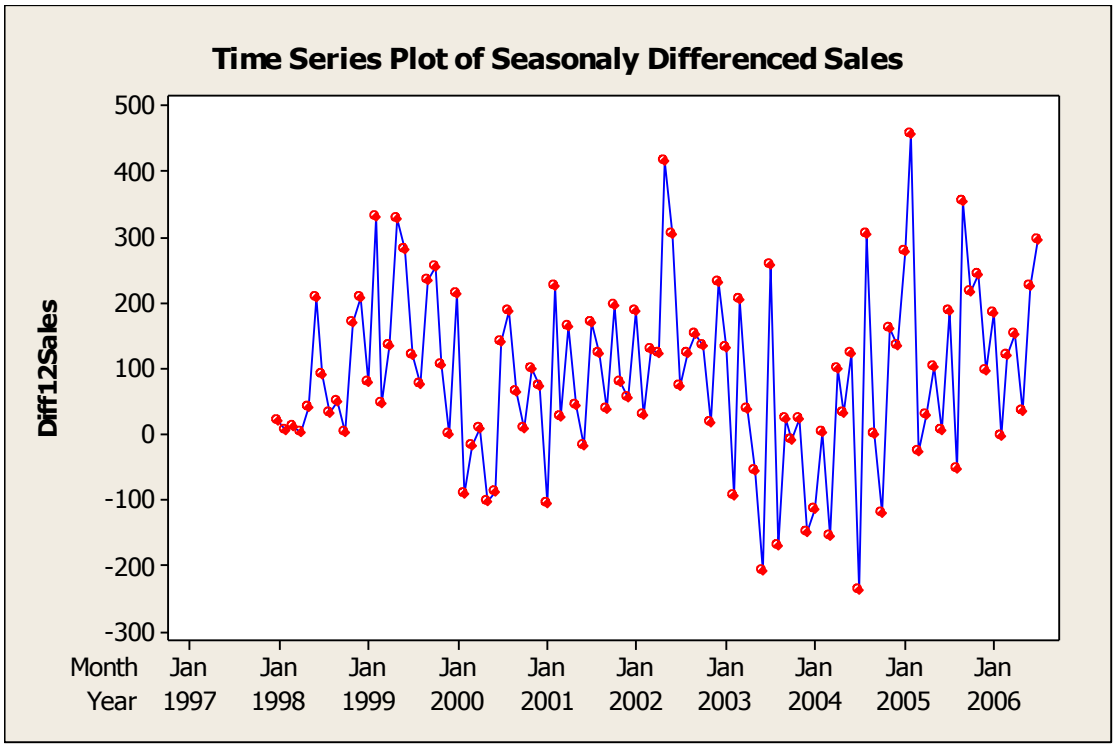
22	0,026250	0,15	159,49
23	0,227137	1,30	167,04
24	0,635500	3,59	226,75
25	0,238127	1,22	235,23
26	0,000420	0,00	235,23
27	0,050577	0,26	235,62
28	-0,025580	-0,13	235,72
29	-0,100489	-0,51	237,30

Η εποχική διαφορά για την περίοδο S=12 καθορίζεται ως:

$$\Delta_{12}=Y_t-Y_{t-12}$$

Για τις πωλήσεις, η πρώτη εποχική διαφορά που μπορεί να υπολογιστεί είναι:

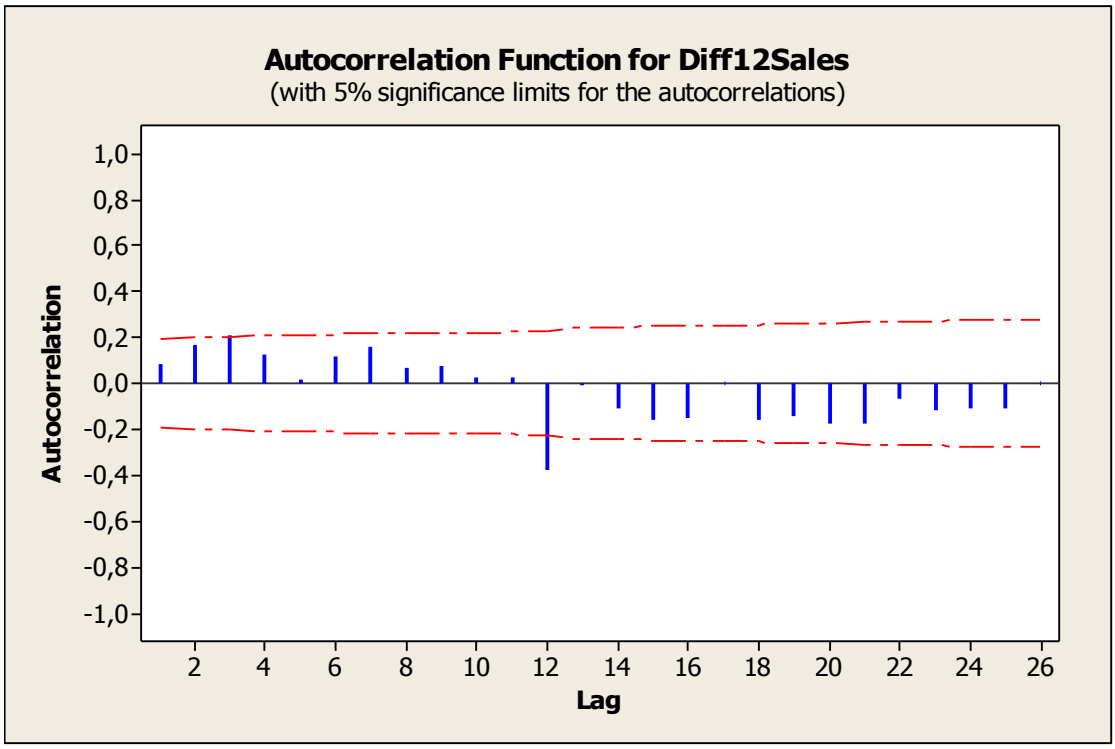
$$Y_{13}-Y_1=1757,6 - 1736,8=20,8$$



Οι εναπομένουσες εποχικές διαφορές υπολογίζονται, και οι σειρές, που αποτελούνται από εποχικές διαφορές των πωλήσεων, φαίνονται στο παρακάτω Γράφημα.

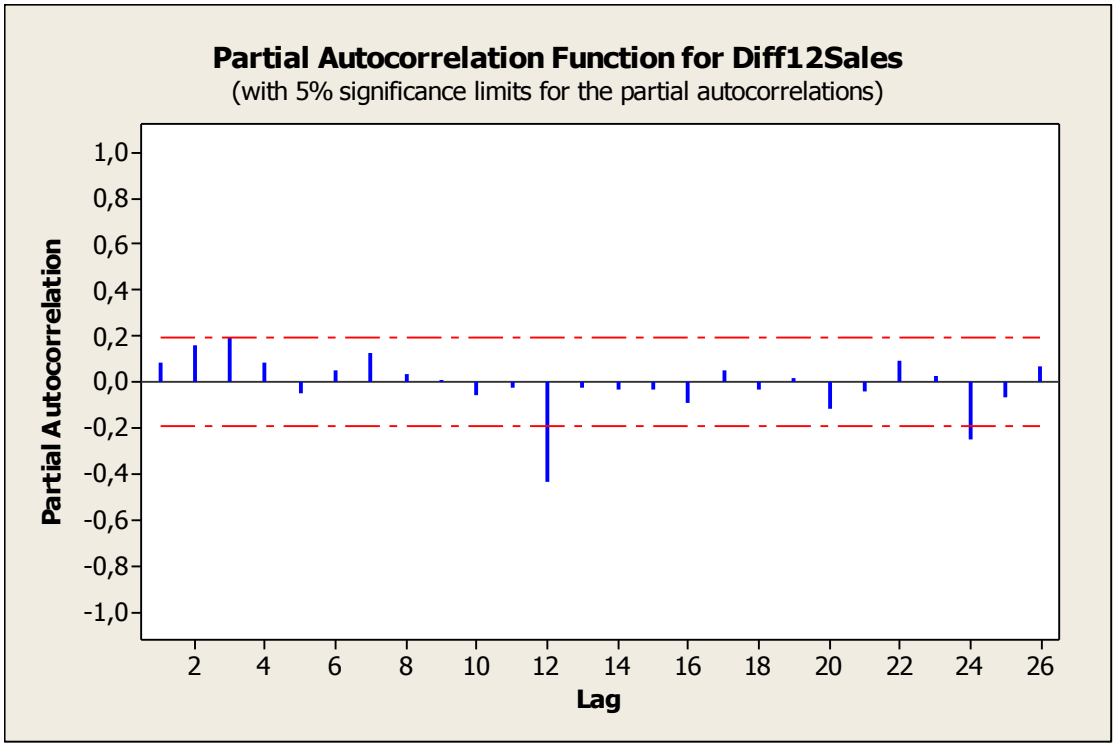
Οι αυτοσυσχετίσεις δείγματος και οι μερικές αυτοσυσχετίσεις

δείγματος για τις διαφορές των σειρών δίνονται στα αντίστοιχα Γραφήματα. Όπως φαίνεται από το Γράφημα οι εποχικές διαφορές των δεδομένων είναι στάσιμες και μοιάζουν να διαφέρουν γύρω από ένα επίπεδο περίπου 100. Οι αυτοσυσχετίσεις έχουν μία σημαντική κορυφή στο lag 22 (διακόπτονται) και οι μερικές αυτοσυσχετίσεις δείγματος έχουν σημαντικές κορυφές στα lags 12 και 24 που γίνονται προοδευτικά μικρότερες (φθίνουν). Αυτή η συμπεριφορά υποδεικνύει έναν MA(1) όρο στο εποχικό lag 12.



Lag	ACF	T	LBQ
1	0,082290	0,84	0,72
2	0,164373	1,66	3,61
3	0,206938	2,03	8,24
4	0,128010	1,21	10,03
5	0,018390	0,17	10,07
6	0,117004	1,09	11,60
7	0,156824	1,44	14,37
8	0,064878	0,59	14,85
9	0,074454	0,67	15,48
10	0,028748	0,26	15,58

11	0,023747	0,21	15,65
12	-0,379941	-3,40	32,80
13	-0,009173	-0,07	32,81
14	-0,111889	-0,90	34,33
15	-0,159221	-1,28	37,45
16	-0,146828	-1,16	40,13
17	0,000677	0,01	40,13
18	-0,161582	-1,26	43,45
19	-0,144594	-1,11	46,14
20	-0,178727	-1,36	50,31



Lag	PACF	T
1	0,082290	0,84
2	0,158675	1,61
3	0,188491	1,91
4	0,086124	0,87
5	-0,053533	-0,54
6	0,052428	0,53
7	0,125409	1,27
8	0,033422	0,34
9	0,007631	0,08
10	-0,056774	-0,58
11	-0,024356	-0,25
12	-0,436449	-4,43

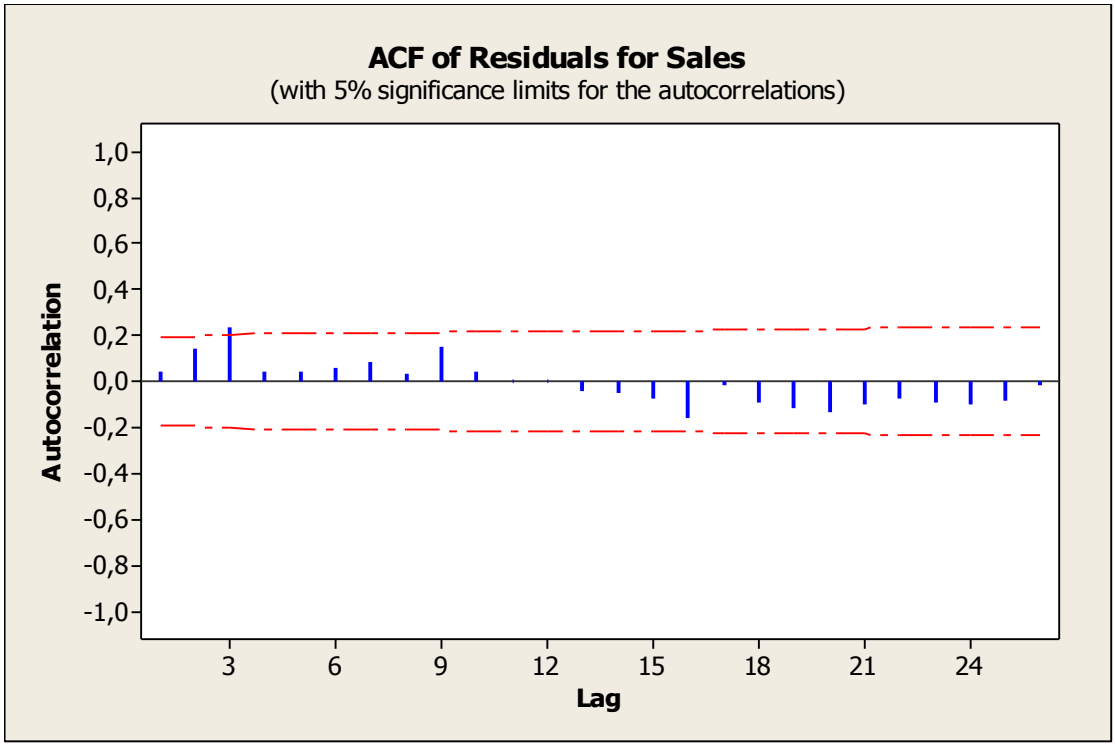
```
 13  -0,022329  -0,23
 14  -0,036679  -0,37
 15  -0,036390  -0,37
 16  -0,091175  -0,93
 17   0,048016   0,49
 18  -0,036175  -0,37
 19   0,012596   0,13
 20  -0,113143  -1,15
```

Ο υπάλληλος αναγνώρισε ένα ARIMA$(0,0,0)(0,1,1)_{12}$ μοντέλο για τα δεδομένα πωλήσεών του. Δεδομένου ότι οι εποχικές διαφορές των σειρών διαφέρουν γύρω από ένα επίπεδο, ο υπάλληλος συμπεριέλαβε ένα σταθερό όρο στο μοντέλο του. Το μοντέλο έχει τη μορφή:

$$Y_t - Y_{t-12} = \mu + \varepsilon_t - \omega_1 \varepsilon_{t-12}$$

ARIMA(0,0,0)(0,1,1)$_{12}$ Model: Sales

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
SMA  12   0,8180   0,0881   9,28  0,000
Constant  85,457    2,910  29,36  0,000

Differencing: 0 regular, 1 seasonal of order 12
Number of observations:  Original series 115,
after differencing 103
Residuals: SS =  1172652 (backforecasts excluded)
           MS =  11610  DF = 101

Modified Box-Pierce (Ljung-Box) Chi-Square
statistic
Lag             12     24     36     48
Chi-Square   12,5   27,1   34,6   46,4
DF             10     22     34     46
P-Value     0,250  0,209  0,439  0,456


Forecasts from period 115 95% Limits
Period  Forecast    Lower    Upper  Actual
   116   2419,69  2208,46  2630,93
   117   2504,31  2293,07  2715,54
```



Ο υπάλληλος χρησιμοποίησε το Minitab για να εκτιμήσει τις παραμέτρους του μοντέλου, υπολόγισε τις αυτοσυσχετίσεις καταλοίπων και παρήγαγε προβλέψεις. Το αποτέλεσμα φαίνεται στον αντίστοιχο Πίνακα και οι αυτοσυσχετίσεις καταλοίπων δίνονται στο αντίστοιχο Γράφημα.

Ο αναλυτής ήταν ικανοποιημένος, που είδε ότι το αρχικό του μοντέλο προσαρμόστηκε καλά στα δεδομένα. Οι Ljung-Box και chi-square στατιστικές, για σύνολα των lags m=12, 24, 36 και 48, ήταν ξεκάθαρο ότι δεν ήταν σημαντικά, όπως φάνηκε από τις μεγάλες p-τιμές. Οι αυτοσυσχετίσεις καταλοίπων ήταν ομοιόμορφα μικρές με όχι κάποιο προφανές μοτίβο. Ο αναλυτής ήταν έτοιμος να χρησιμοποιήσει το προσαρμοσμένο του μοντέλο για να επιβεβαιώσει τις προβλέψεις, που παρήχθησαν από το Minitab.

Οι εκτιμώμενες παράμετροι ήταν $\hat{\mu}=85{,}457$ και $\omega_1=0{,}818$. Χρησιμοποιώντας αυτές τις τιμές, η προηγούμενη Εξίσωση μπορεί να

λυθεί για το Y_t για να δώσει:

$$\hat{Y}_t = Y_{t-12} + 85{,}457 + \varepsilon_t - 0{,}818\,\varepsilon_{t-12}$$

Για να προβλέψει τις πωλήσεις για την περίοδο 116, έθεσε t=116 και σημείωσε, ότι τη χρονική στιγμή που έκανε την πρόβλεψη, η καλύτερη εικασία του ε_{116} (το σφάλμα για την επόμενη περίοδο) ήταν μηδέν. Επομένως, η εξίσωση πρόβλεψης είναι:

$$\hat{Y}_{116} = Y_{104} + 85{,}457 - 0{,}818\,e_{104}$$

$$2.272 + 85{,}457 - 0{,}818\,(-72{,}418) = 2.419{,}7$$

όπου e_{104} = -72,418 είναι το κατάλοιπο (εκτιμώμενο σφάλμα) για την περίοδο 104.

Όμοια:

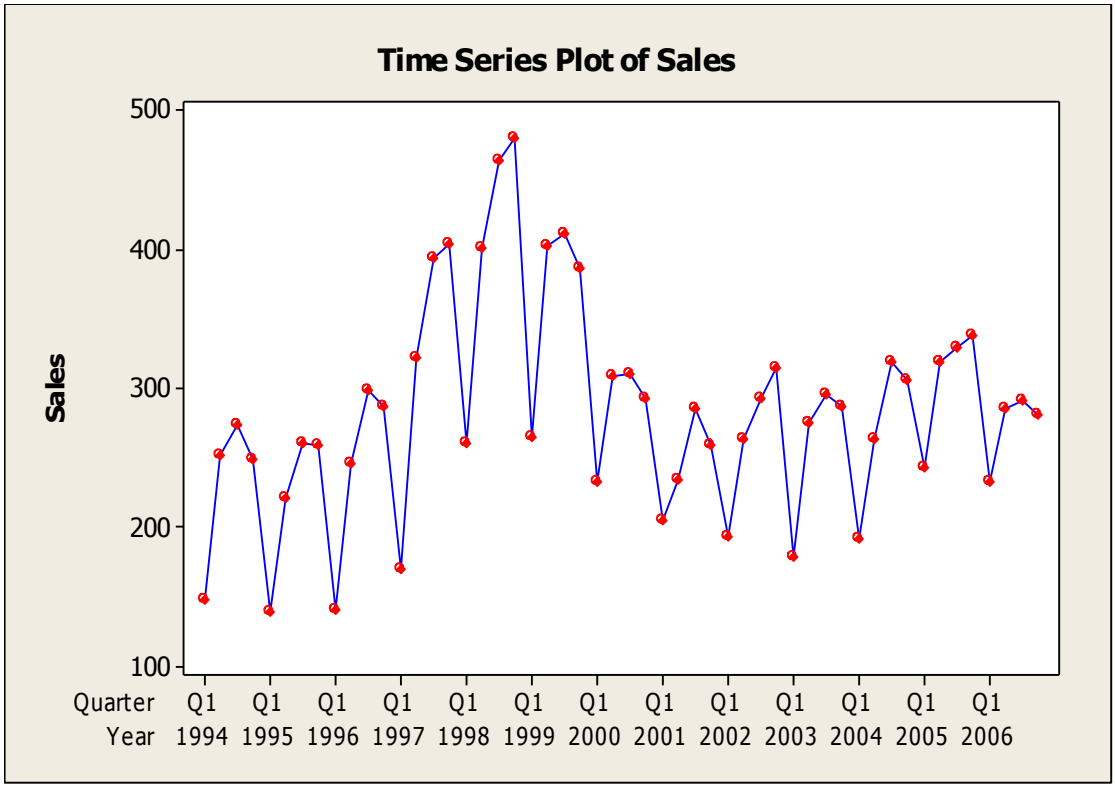
$$\hat{Y}_{117} = Y_{105} + 85{,}457 - 0{,}818\,e_{105}$$

$$2.581{,}8 + 85{,}457 - 0{,}818\,(199{,}214) = 2.504{,}3$$

Οι προβλέψεις του συμφωνούσαν με εκείνες από τα αποτελέσματα του Minitab.

Παράδειγμα (EX.9-10)

Ο αναλυτής μίας εταιρίας χρησιμοποίησε ανάλυση αυτοσυσχέτισης για να συμπεράνει, ότι οι τριμηνιαίες πωλήσεις ήταν εποχικές. Αποφάσισε να προβλέψει τις πωλήσεις για το 2007 χρησιμοποιώντας την Box-Jenkins μεθοδολογία. Το διάγραμμα των δεδομένων φαίνεται στο παρακάτω Γράφημα. Το διάγραμμα χρονοσειρών δείχνει ένα ξεκάθαρο μοτίβο με ίσως μία ελαφρά ανοδική τάση.



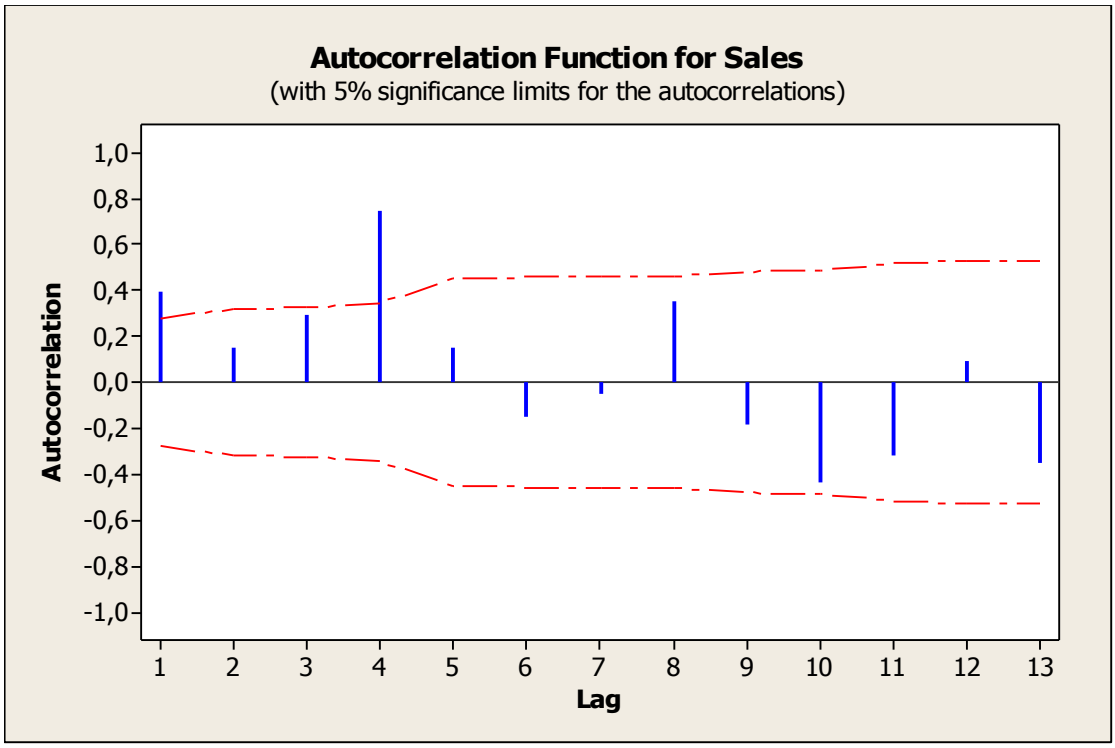
Ο αναλυτής ξεκίνησε την ανάλυσή του κοιτάζοντας τις αυτοσυσχετίσεις δείγματος για τις πρωτότυπες σειρές και τις ποικίλες διαφορές των σειρών. Όταν το απλούστερο μοντέλο, που προτείνεται από ένα μοτίβο αυτοσυσχέτισης, έχει διερευνητικά αναγνωριστεί, η αναγνώριση μπορεί να επιβεβαιωθεί εξετάζοντας τις μερικές αυτοσυσχετίσεις δείγματος. Το Minitab υπολόγισε τις συναρτήσεις αυτοσυσχέτισης για τις ακόλουθες σειρές:

Y_t	οι πρωτότυπες σειρές πωλήσεων
$\Delta Y_t=Y_t-Y_{t-1}$	οι σειρές των πρώτων κανονικών διαφορών
$\Delta_4 Y_t=Y_t-Y_{t-4}$	οι σειρές των εποχικών διαφορών τάξης S=4
$[\Delta\Delta_4 Y_t=\Delta(\Delta_4 Y_t)=\Delta(Y_t-Y_{t-4})=$ $\Delta Y_t-\Delta Y_{t-4}=Y_t-Y_{t-1}-Y_{t-4}+Y_{t-5}]$	οι σειρές που αποτελούνται από μία εποχική διαφορά ακολουθούμενη από κανονική διαφορά, ή αντίστροφα

Κατά περίπτωση, είναι απαραίτητο να παίρνουμε μία εποχική διαφορά και μία κανονική διαφορά, ώστε να γίνουν οι σειρές, που

προκύπτουν, στάσιμες.

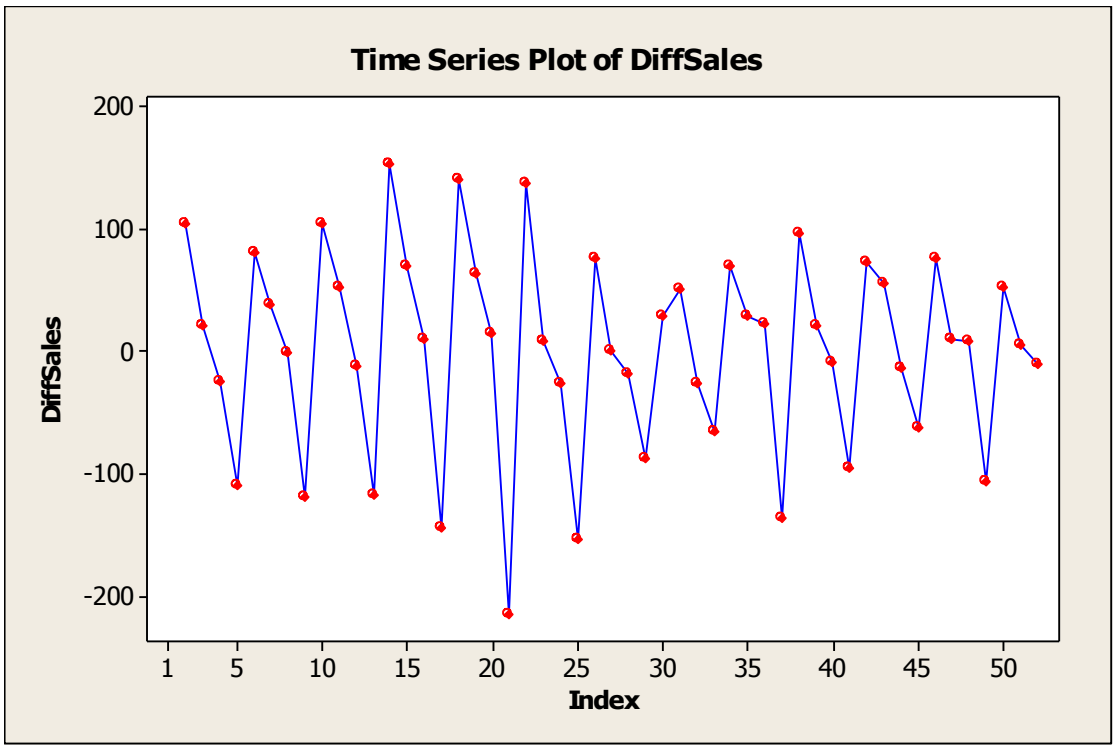
Οι αυτοσυσχετίσεις δείγματος για τις πρωτότυπες σειρές, για τις εποχικές διαφορές των σειρών, και για τις σειρές με μία κανονική και μία εποχική διαφορά φαίνονται στα παρακάτω Γραφήματα.



Lag	ACF	T	LBQ
1	0,392843	2,83	8,50
2	0,153948	0,97	9,83
3	0,293739	1,82	14,77
4	0,743533	4,34	47,11
5	0,150619	0,67	48,47
6	-0,153014	-0,67	49,90
7	-0,047049	-0,21	50,04
8	0,346979	1,51	57,72
9	-0,182597	-0,76	59,90
10	-0,434737	-1,80	72,53
11	-0,315090	-1,23	79,33
12	0,091175	0,35	79,92
13	-0,353272	-1,34	88,90

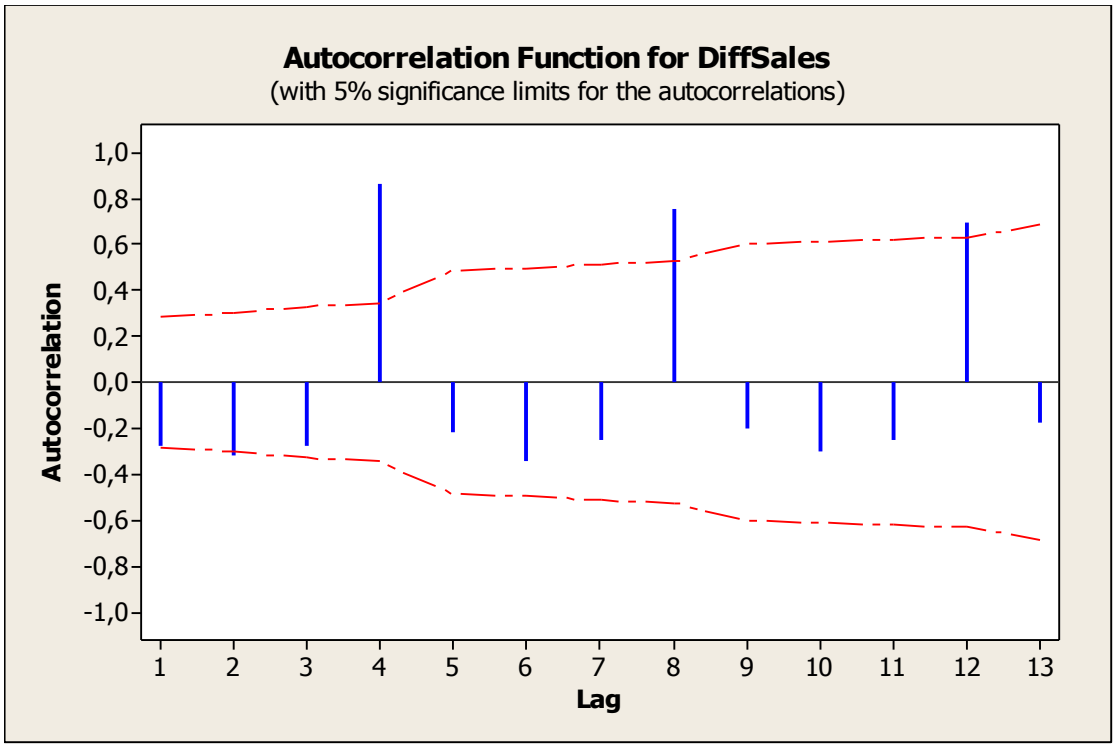
Η συνάρτηση αυτοσυσχέτισης των πρωτότυπων σειρών έχει (σημαντικά) μεγάλες κορυφές στα lags 1 και 4. Ωστόσο, οι

αυτοσυσχετίσεις στα εποχικά lags 4, 8 και 12 φθίνουν στο μηδέν. Αυτό μπορεί να υποδεικνύει μη στάσιμες μη γραμμικές σειρές και την ανάγκη για μία εποχική διαφορά και ίσως μία κανονική διαφορά. Η προσπάθειά του να προσαρμόσει ένα ARIMA μοντέλο στις πρωτότυπες σειρές με ένα σταθερό όρο, ένα κανονικό όρο κινούμενου μέσου, και έναν εποχικό αυτοπαλινδρομικό όρο δεν ήταν επιτυχής. Η εκτιμώμενη εποχική αυτοπαλινδρομική παράμετρος αποδείχτηκε να είναι κοντά στο ένα (το οποίο υπονοεί ότι μία εποχική διαφορά χρειαζόταν), και οι αυτοσυσχετίσεις καταλοίπων ήταν γενικά μεγάλες και όχι συνεπείς με τυχαία σφάλματα. Στο Γράφημα "for Diff4sales" φαίνεται η συνάρτηση αυτοσυσχέτισης των εποχικών διαφορών των σειρών.

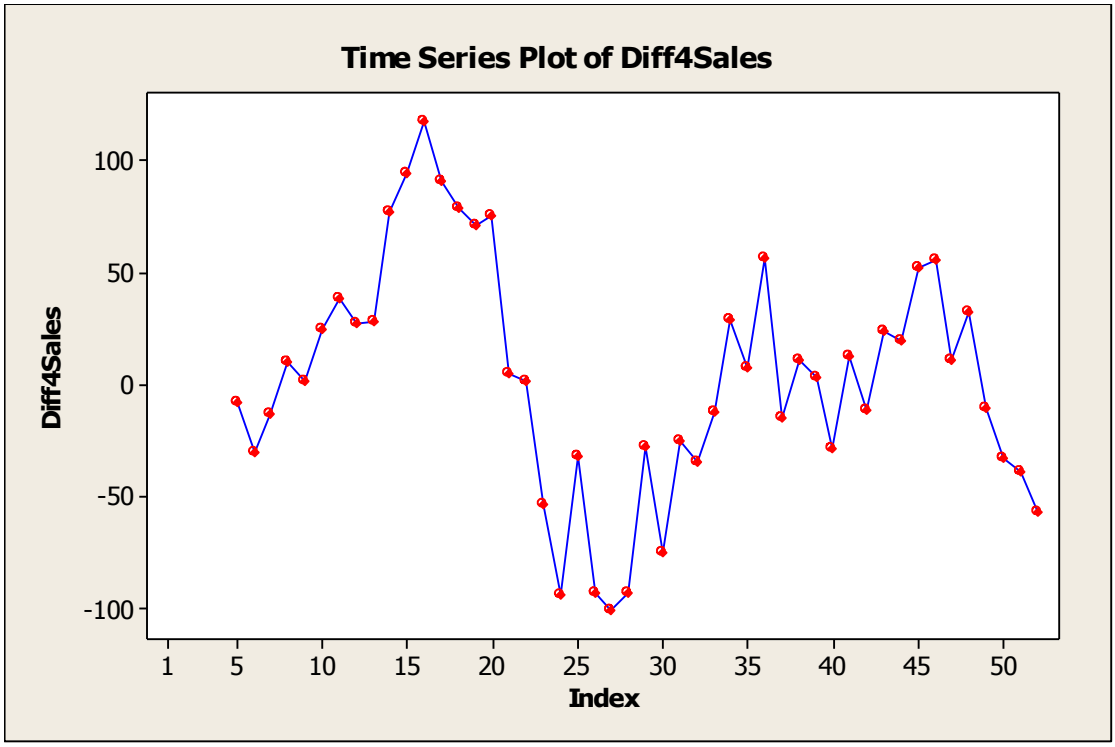


Οι αυτοσυσχετίσεις για τις εποχικές διαφορές των δεδομένων είναι μεγάλες για μικρές καθυστερήσεις (lags) και μειώνονται αρκετά αργά στη μορφή ενός μοτίβου κύματος. Ο αναλυτής αποφάσισε ότι οι σειρές ακόμη δεν ήταν στάσιμες και ένιωσε ότι μία κανονική διαφορά μαζί με εποχική διαφορά μπορεί να απαιτείται, για να επιτευχθεί η στασιμότητα.

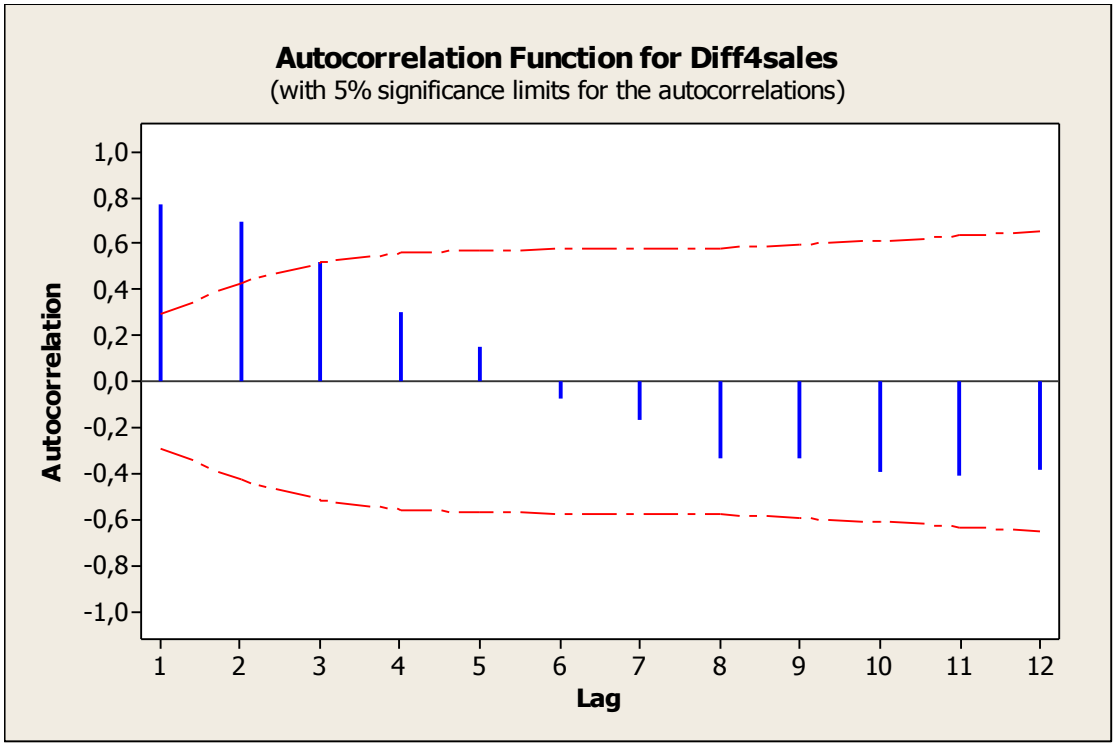
Κατά συνέπεια, το Minitab παρήγαγε αυτοσυσχετίσεις δείγματος για τις πρώτες διαφορές και τις εποχικές διαφορές των πωλήσεων στο Γράφημα "DiffDiff4sales".

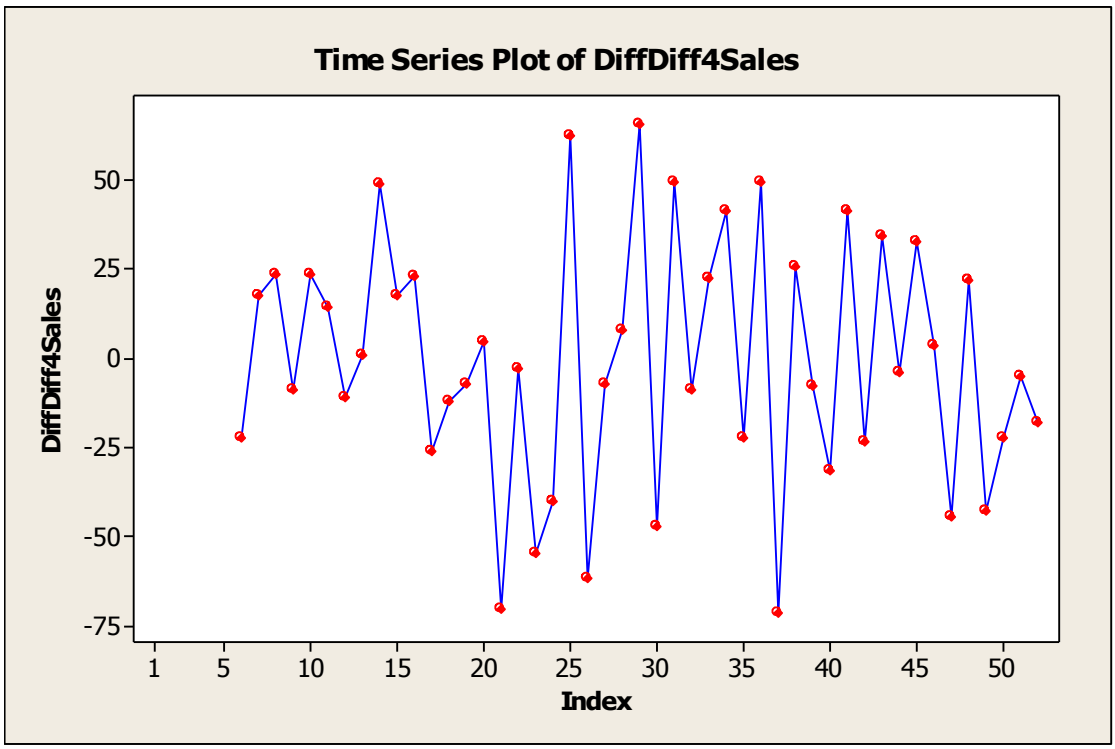


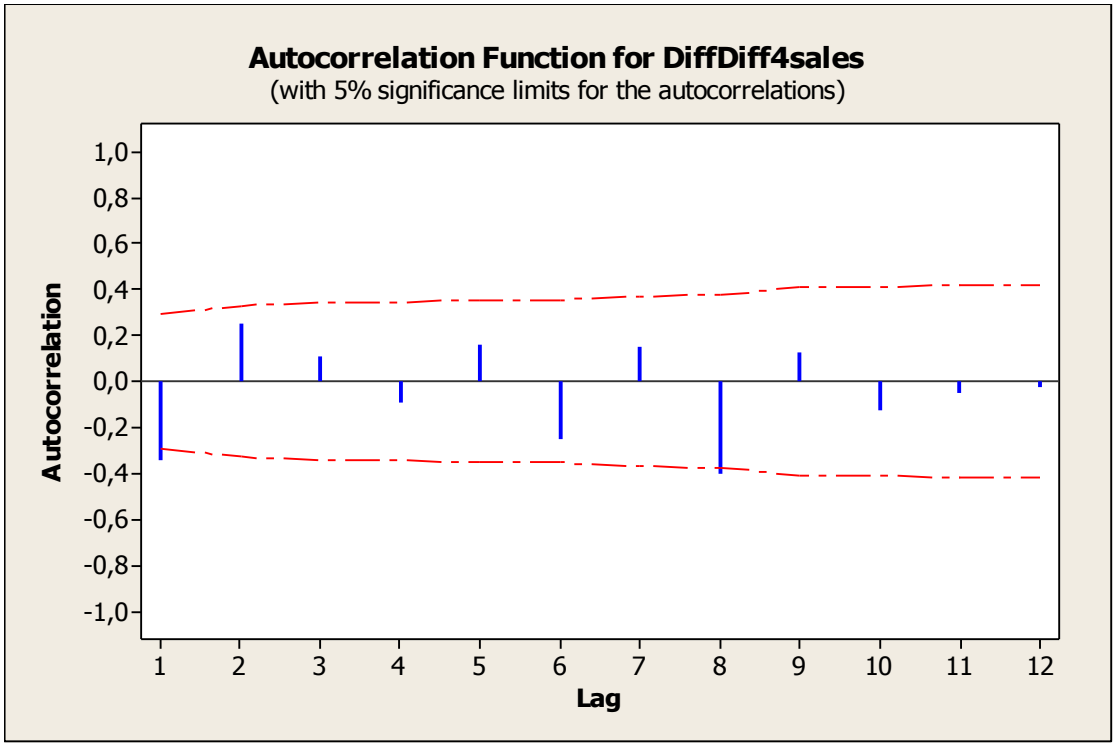
Lag	ACF	T	LBQ
1	-0,278396	-1,99	4,19
2	-0,320930	-2,13	9,87
3	-0,279660	-1,71	14,28
4	0,861621	5,00	56,97
5	-0,219477	-0,90	59,80
6	-0,341133	-1,38	66,79
7	-0,251392	-0,98	70,67
8	0,754176	2,90	106,43
9	-0,199476	-0,66	108,99
10	-0,302129	-1,00	115,01
11	-0,254066	-0,82	119,37
12	0,691622	2,21	152,52
13	-0,179642	-0,53	154,82



Lag	ACF	T	LBQ
1	0,768338	5,32	30,15
2	0,693397	3,25	55,23
3	0,514832	2,01	69,37
4	0,303268	1,10	74,38
5	0,152356	0,54	75,68
6	-0,075501	-0,26	76,00
7	-0,171203	-0,60	77,72
8	-0,333102	-1,16	84,38
9	-0,338302	-1,14	91,42
10	-0,392471	-1,29	101,15
11	-0,407624	-1,30	111,93
12	*-0,387378*	*-1,19*	*121,93*


Autocorrelation Function for Diff4sales
(with 5% significance limits for the autocorrelations)
Autocorrelation
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
-1,0
1
2
3
4
5
6
7
8
9
10
11
12
Lag


Time Series Plot of DiffDiff4Sales
DiffDiff4Sales
50
25
0
-25
-50
-75
1
5
10
15
20
25
30
35
40
45
50
Index

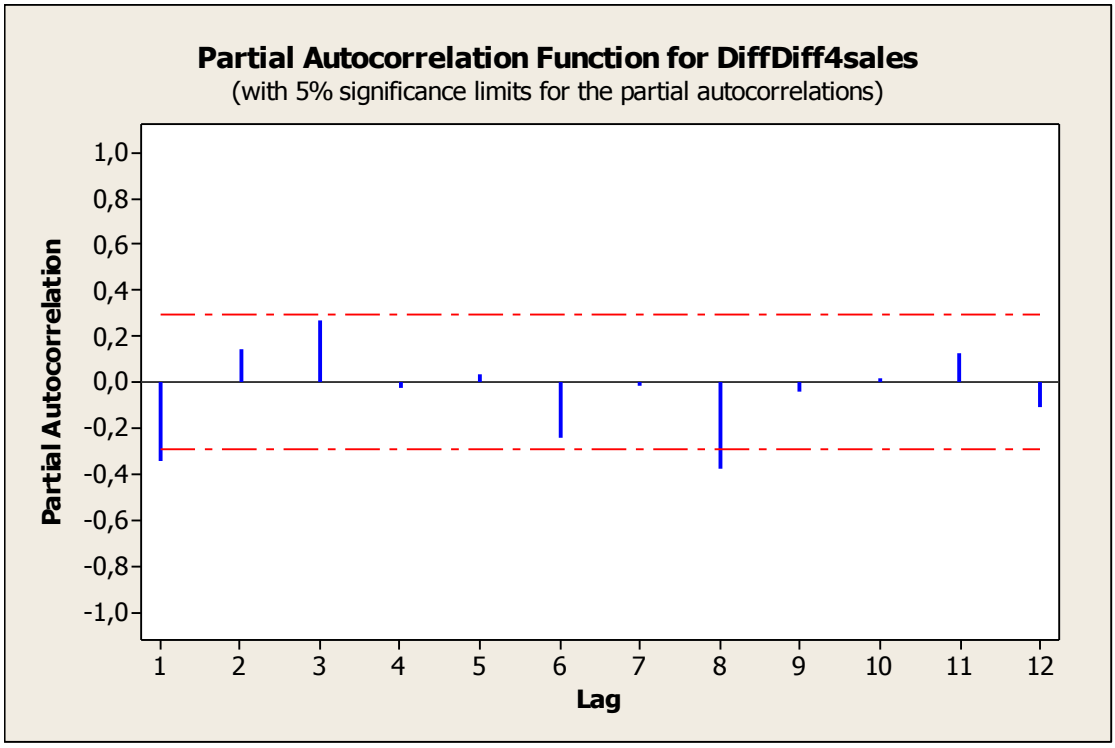


Lag	ACF	T	LBQ
1	-0,346864	-2,38	6,02
2	0,248733	1,53	9,19
3	0,105922	0,62	9,78
4	-0,093787	-0,55	10,25
5	0,158157	0,91	11,62
6	-0,252737	-1,44	15,21
7	0,146952	0,80	16,45
8	-0,398270	-2,14	25,82
9	0,127245	0,63	26,80
10	-0,126554	-0,62	27,80
11	-0,053071	-0,26	27,98
12	*-0,021962*	*-0,11*	*28,01*

Εξετάζοντας τις αυτοσυσχετίσεις του DiffDiff4sales, ο αναλυτής παρατήρησε μόνο δύο σημαντικές τιμές, στα lags 1 και 8. Επιπλέον, είδε ότι οι αυτοσυσχετίσεις για τα πρώτα δύο lags άλλαξαν πρόσημο. Ο αναλυτής είδε, ότι ένα ARIMA μοντέλο με έναν κανονικό αυτοπαλινδρομικό όρο και ίσως έναν εποχικό όρο κινούμενου μέσου στο lag 8 μπορεί να είναι κατάλληλο. Πριν συνεχίσει, ήθελε να επιβεβαιώσει την αρχική του επιλογή εξετάζοντας τις μερικές αυτοσυσχετίσεις για τις

πρώτες διαφορές και τις εποχικές διαφορές των σειρών.

Οι μερικές αυτοσυσχετίσεις μοιάζουν να διακόπτονται μετά το πρώτο lag και υπάρχει μία σημαντική μερική αυτοσυσχέτιση στο lag 8.

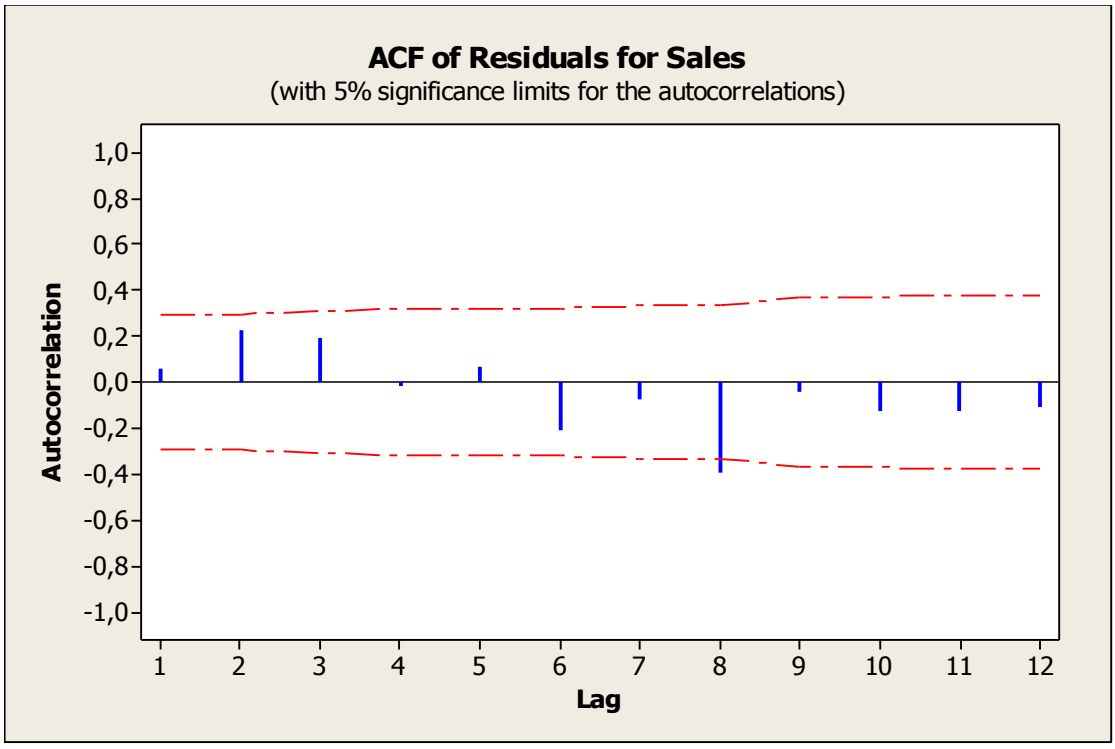


Ο αναλυτής δεν ήταν σίγουρος για το πώς να χειριστεί τις εποχικές παραμέτρους, έτσι αποφάσισε να το κρατήσει απλό και να προσαρμόσει ένα ARIMA(1,1,0)(0,1,0)$_4$ μοντέλο. Αυτό το μοντέλο επιτρέπει δύο διαφορές μία κανονική και μία εποχιχή και έναν κανονικό αυτοπαλινδρομικό όρο αλλά όχι εποχικούς αυτοπαλινδρομικούς συντελεστές ή συντελεστές κινούμενου μέσου. Ένα διάγραμμα των σειρών μετά τη μέθοδο των διαφορών δείχνει ότι διαφέρουν γύρω από το μηδέν, έτσι δεν συμπεριλαμβάνεται σταθερός όρος. Ο αναλυτής αιτιολόγησε, ότι αν χρειάζονταν εποχικές παράμετροι, οι ενδείξεις θα φαίνονταν στις αυτοσυσχετίσεις καταλοίπων από το αρχικό μοντέλο.

ARIMA(1,1,0)(0,1,0)$_4$ Model: Sales

Final Estimates of Parameters

```
Type          Coef  SE Coef      T      P
AR   1    -0,3519   0,1399  -2,51  0,016
Constant   -1,104    4,755  -0,23  0,817

Differencing: 1 regular, 1 seasonal of order 4
Number of observations:  Original series 52,
after differencing 47
Residuals: SS =  47821,3 (backforecasts excluded)
           MS =  1062,7  DF = 45

Modified Box-Pierce (Ljung-Box) Chi-Square
statistic
Lag             12     24     36  48
Chi-Square    19,7   23,7   32,1   *
DF              10     22     34   *
P-Value      0,032  0,364  0,563   *
```



Το αποτέλεσμα για το ARIMA $(1,1,0)(0,1,0)_4$ μοντέλο φαίνεται στον παραπάνω πίνακα και οι αυτοσυσχετίσεις καταλοίπων φαίνονται στο παραπάνω Γράφημα. Το μοντέλο μοιάζει να προσαρμόζεται αρκετά καλά με έναν εκτιμώμενο AR(1) συντελεστή $\hat{\varphi}_1 = 0{,}352$ και ένα μέσο

τετραγωνικό σφάλμα καταλοίπων s^2=1.040,7. Ωστόσο, μία σημαντική κορυφή συμβαίνει στο lag 8 στις αυτοσυσχετίσεις καταλοίπων και το Ljung-Box στατιστικό για τα πρώτα m=12 lags είναι Q=19,7 με μία p-τιμή 0,05. Το τελευταίο προτείνει, ότι σαν σύνολο, οι πρώτες 12 αυτοσυσχετίσεις καταλοίπων είναι μεγαλύτερες, από ότι αναμενόταν για τυχαία σφάλματα.

Ο αναλυτής αποφάσισε να τροποποιήσει το αρχικό μοντέλο, για να συμπεριλάβει όρους κινούμενου μέσου, που αντιστοιχούν στα εποχικά lags 4 και 8. Η διαδικασία υποδεικνύει, ότι μόνο ο όρος στο lag 8 είναι απαραίτητος. Ωστόσο, το Minitab απαιτεί να συμπεριλάβει τίς εποχικές AR ή MA παραμέτρους στο μοντέλο και να συμπεριλάβει το υψηλότερο πολλαπλάσιο του εποχικού lag που χρειάζεται. Σε αυτή την περίπτωση, το εποχικό lag είναι S=4 και απαιτείται μία παράμετρος κινούμενου μέσου στο lag 2*4=8.

ARIMA(1,1,0)(0,1,2)$_4$ Model: Sales

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	-0,3497	0,1438	-2,43	0,019
SMA	4	0,2413	0,1439	1,68	0,101
SMA	8	0,6574	0,1474	4,46	0,000
Constant		-0,7057	0,8855	-0,80	0,430

Differencing: 1 regular, 1 seasonal of order 4
Number of observations: Original series 52, after differencing 47
Residuals: SS = 31181,6 (backforecasts excluded)
MS = 725,2 DF = 43

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

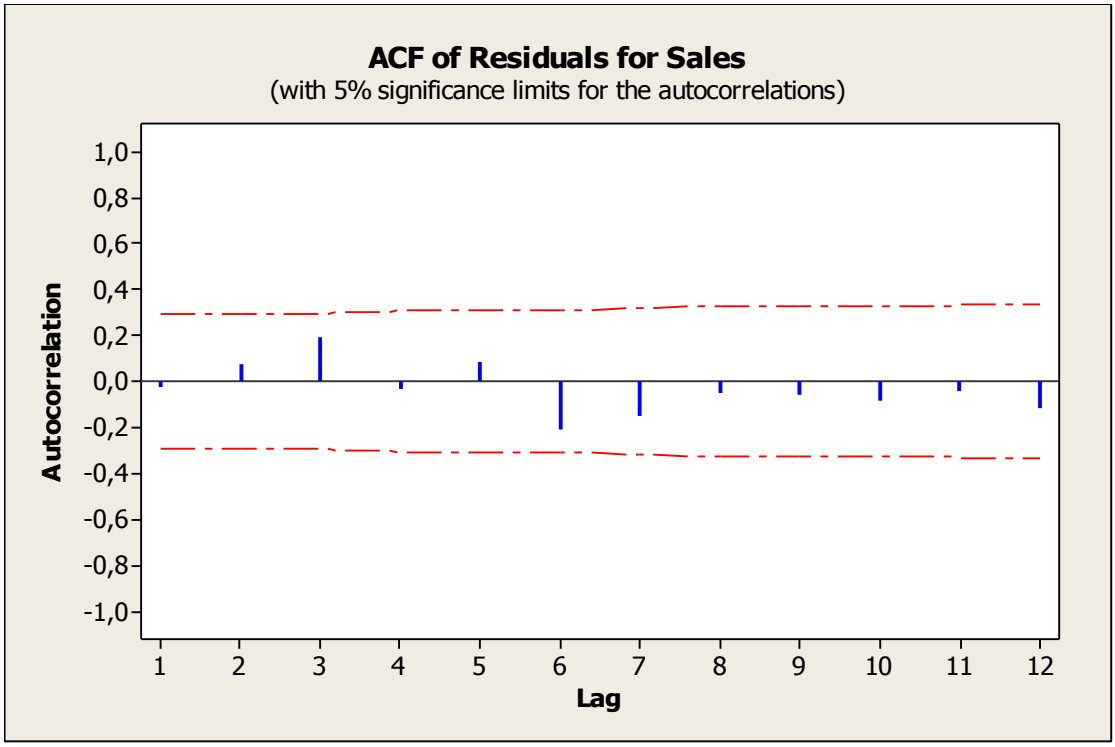
Lag	12	24	36	48
Chi-Square	8,4	10,7	21,5	*
DF	8	20	32	*

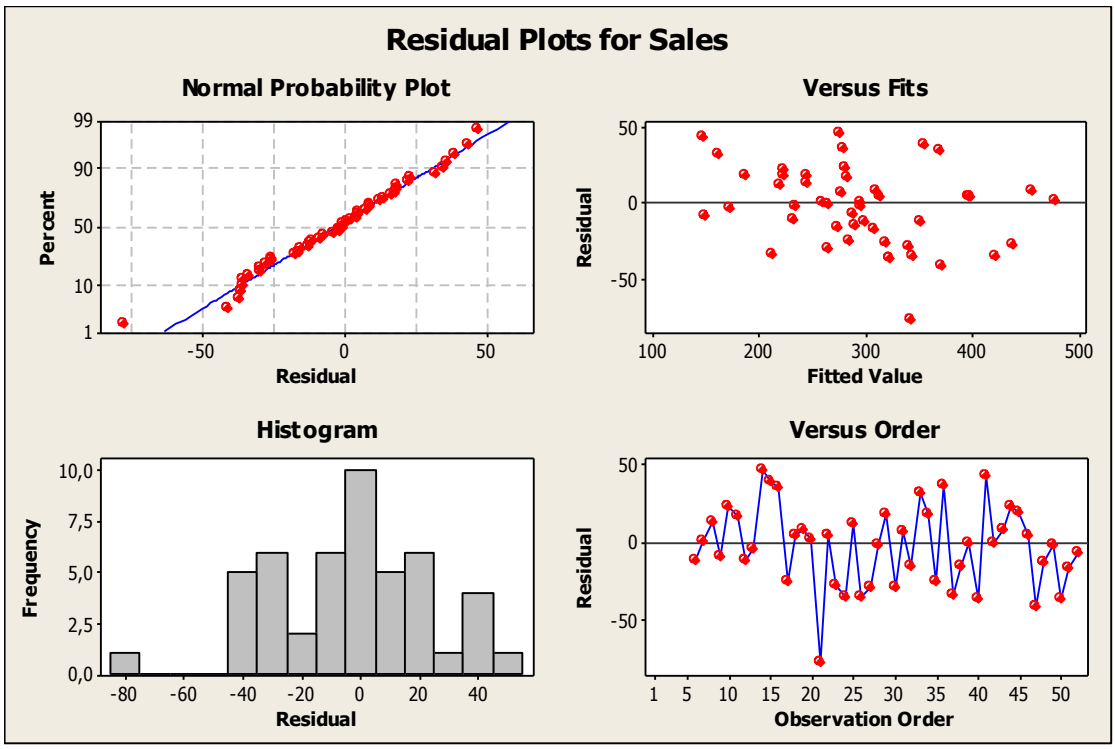
```
P-Value     0,399  0,954  0,920   *

Forecasts from period 52  95% Limits
Period  Forecast    Lower    Upper  Actual
    53   166,292  113,501  219,083
    54   226,290  163,317  289,263
```

Αφαιρούμε τον σταθερό όρο γιατί δεν είναι σημαντικός, οπότε το μοντέλο είναι το ακόλουθο:

ARIMA(1,1,0)(0,1,2)$_4$ Model: Sales

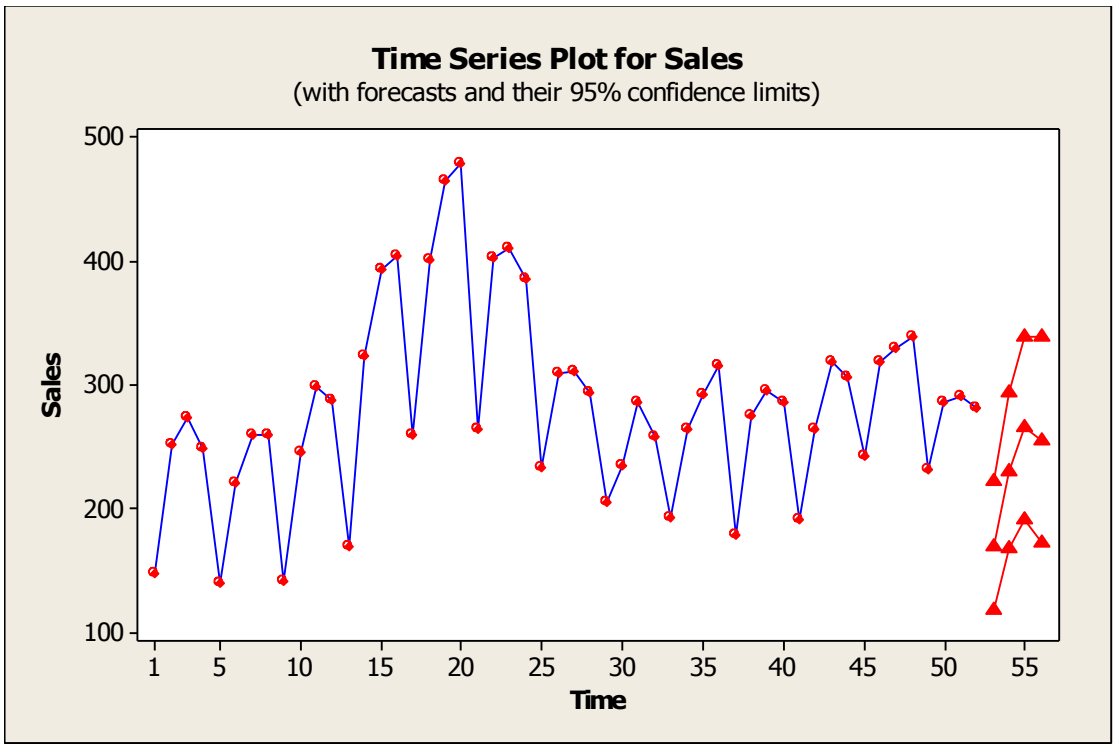
```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
AR   1  -0,3505   0,1423  -2,46  0,018
SMA  4   0,2392   0,1340   1,78  0,081
SMA  8   0,6716   0,1404   4,78  0,000

Differencing: 1 regular, 1 seasonal of order 4
Number of observations:  Original series 52,
after differencing 47
Residuals: SS = 31525,4 (backforecasts excluded)
           MS =  716,5  DF = 44

Modified Box-Pierce (Ljung-Box) Chi-Square
statistic
Lag            12     24     36  48
Chi-Square    8,4   10,7   22,0   *
DF              9     21     33   *
P-Value     0,493  0,969  0,928   *

Forecasts from period 52 95% Limits
Period  Forecast    Lower    Upper  Actual
    53   169,401  116,926  221,875
    54   230,617  168,047  293,188
    55   265,107  190,558  339,656
    56   255,493  171,695  339,291
```





Ο αναλυτής χρησιμοποίησε το Minitab για να προσαρμόσει ένα ARIMA$(1,1,0)(0,1,2)_4$ μοντέλο στα δεδομένα του. Το αποτέλεσμα φαίνεται στον παραπάνω Πίνακα και οι αυτοσυσχετίσεις καταλοίπων δίνονται στο παραπάνω Γράφημα. Ήταν ευχαριστημένος με τα

αποτελέσματα. Το μέσο τετραγωνικό σφάλμα καταλοίπων έχει μειωθεί στο s^2=716,5 και οι αυτοσυσχετίσεις καταλοίπων, όπως κρίνονται από τις μεμονωμένες τιμές και τα Ljung-Box chi-square στατιστικά, είναι συνεπή με τυχαία σφάλματα. Τα διαγράμματα καταλοίπων (δεν φαίνονται) προτείνουν ότι οι άλλες υποθέσεις όρων σφαλμάτων είναι κατάλληλες.



Για να αποκτήσει περισσότερη εμπειρία με την Box-Jenkins μεθοδολογία, ήθελε να ελέγξει τις προβλέψεις των πωλήσεων για τα επόμενα δύο τρίμηνα που φαίνονται στο αποτέλεσμα του Minitab. Χρησιμοποιώντας εκτιμήσεις παραμέτρων στη θέση των αληθινών τιμών, το τελικό μοντέλο είναι:

$$W_t = -0{,}351\, W_{t-1} + \varepsilon_t - 0{,}239\, \varepsilon_{t-4} - 0{,}672\, \varepsilon_{t-8} \quad \text{με}$$

$$W_t = \Delta\Delta_4 Y_t = Y_t - Y_{t-1} - Y_{t-4} + Y_{t-5}$$

είναι οι κατάλληλες διαφορές των σειρών. Η εξίσωση πρόβλεψης γραμμένη σε όρους του Y, παίρνει τη μορφή:

$$\hat{Y}_t = Y_{t-1} + Y_{t-4} - Y_{t-5} - 0{,}351\left(Y_{t-1} - Y_{t-2} - Y_{t-5} + Y_{t-6}\right)$$

$$-0{,}239\, e_{t-4} - 0{,}672\, e_{t-8}$$

Όπου e_{t-4} και e_{t-8} είναι τα κατάλοιπα (εκτιμώμενα σφάλματα) για τις περιόδους t-4 και t-8, αντίστοιχα. Οπότε $\hat{Y}_{53} = 169{,}4$ *και* $\hat{Y}_{54} = 230{,}6$.

9.2.9 ΣΧΕΣΗ ΕΚΘΕΤΙΚΗΣ ΕΞΟΜΑΛΥΝΣΗΣ ΚΑΙ ARIMA ΜΟΝΤΕΛΩΝ

Μερικά ARIMA μοντέλα παράγουν προβλέψεις, που είναι ίδιες, ή περίπου ίδιες με εκείνες, που προκύπτουν από τις μεθόδους εξομάλυνσης, οι οποίες έχουν συζητηθεί. Για να αναπαραστήσουμε μία περίπτωση, θεωρείστε ένα ARIMA (0,1,1) μοντέλο (χωρίς τον σταθερό όρο):

$$Y_t - Y_{t-1} = \varepsilon_t - \omega_1 \varepsilon_{t-1}$$

Υποθέτουμε ότι η τιμή στο t είναι γνωστή και απαιτείται μία πρόβλεψη του Y_{t+1}. Αντικαθιστώντας t+1 για t στην προηγούμενη Εξίσωση και λύνοντας ως προς Y_{t+1} δίνει:

$$Y_{t+1} = Y_t + \varepsilon_{t+1} - \omega_1 \varepsilon_t$$

Επειδή, στο χρόνο t, η καλύτερη εικασία του ε_{t+1} είναι μηδέν και το ε_t εκτιμάται από το κατάλοιπο $\varepsilon_t = Y_t - \hat{Y}_t$, η εξίσωση πρόβλεψης είναι:

$$\hat{Y}_{t+1} = Y_t - \omega_1 (Y_t - \hat{Y}_t) = (1 - \omega_1) Y_t + \omega_1 \hat{Y}_t$$

Θέτουμε $\alpha = 1 - \omega_1$ και η προηγούμενη Εξίσωση είναι ταυτόσημη με την απλή εξίσωση εκθετικής εξομάλυνσης:

$$\hat{Y}_{t+1} = \alpha Y_t + (1 - \alpha) \hat{Y}_t$$

Η πρόβλεψη χρησιμοποιώντας απλή εκθετική εξομάλυνση είναι ισοδύναμη με την παραγωγή προβλέψεων χρησιμοποιώντας ένα ARIMA (0,1,1) μοντέλο με παράμετρο $\omega_1 = 1 - \alpha$. Να σημειωθεί ότι το ARIMA (0,1,1) μοντέλο αναπαριστά μία μη στάσιμη διαδικασία. Η απλή εκθετική εξομάλυνση πρέπει να δουλεύει καλά για εκείνες τις χρονοσειρές, που μπορούν να περιγραφούν επαρκώς από ένα ARIMA (0,1,1) μοντέλο. Για

εκείνες τις χρονοσειρές, που δεν μπορούν να περιγραφούν επαρκώς από αυτό το μοντέλο, οι προβλέψεις, που παράγονται από εκθετική εξομάλυνση, μπορεί να μην είναι συγκεκριμένα καλές.

Για απλή εκθετική εξομάλυνση, η παράμετρος α συχνά περιορίζεται στο εύρος $0<\alpha<1$. Επειδή η παράμετρος κινούμενου μέσου ω_1, στο ARIMA (0,1,1) μοντέλο περιορίζεται στο εύρος $-1<\omega_1<1$, έτσι μιλώντας αυστηρά, οι δύο μέθοδοι πρόβλεψης είναι ισοδύναμες για θετικές τιμές των παραμέτρων α και ω_1. Στην περίπτωση που το $\omega_1=-1$ το $\alpha=2$.

Για τη διπλή εκθετική εξομάλυνση το Minitab χρησιμοποιεί το ARIMA(0,2,2) μοντέλο χωρίς σταθερό όρο. Οι συντελεστές της μεθόδου Holt αντιστοιχούν σε:

$$\alpha=\omega_2+1, \quad \beta=\frac{(1-\omega_1-\omega_2)}{(1+\omega_2)}.$$

Οπότε σύμφωνα με τους περιορισμούς τα α και β δίνονται από τις ακόλουθες σχέσεις:

$$0<\alpha<2 \ \textit{και} \ \beta=\frac{4}{\alpha}-2.$$

9.2.10 ΠΛΕΟΝΕΚΤΗΜΑΤΑ – ΜΕΙΟΝΕΚΤΗΜΑΤΑ ARIMA ΜΟΝΤΕΛΩΝ

Η Box-Jenkins προσέγγιση στην ανάλυση χρονοσειρών είναι ένα πολύ ισχυρό εργαλείο για παροχή ακριβών βραχυπρόθεσμων-μεσοπρόθεσμων προβλέψεων. Τα ARIMA μοντέλα είναι αρκετά ευέλικτα και μπορούν να αναπαριστούν ένα ευρύ φάσμα των χαρακτηριστικών των χρονοσειρών, που μπορούν να συμβούν στην πράξη. Οι τυπικές διαδικασίες για έλεγχο της επάρκειας του μοντέλου είναι διαθέσιμες. Επίσης, προβλέψεις και διαστήματα πρόβλεψης έπονται απευθείας από

το προσαρμοσμένο μοντέλο.

Ωστόσο, η μοντελοποίηση ARIMA έχει μερικά μειονεκτήματα.

1. Απαιτείται μία σχετικά μεγάλη ποσότητα δεδομένων. Πρέπει να αναγνωριστεί ότι, αν τα δεδομένα είναι εποχικά, με μία ας πούμε εποχική περίοδο S=12, τότε οι μηνιαίες παρατηρήσεις για ένα έτος αποτελούν απαραίτητα ένα σημείο δεδομένων, όχι 12. Μιλώντας γενικά, για μη εποχικά δεδομένα, απαιτούνται περίπου 40 παρατηρήσεις ή περισσότερες, για να αναπτυχθεί ένα ARIMA μοντέλο. Για εποχικά δεδομένα, απαιτούνται περίπου 6 έως 10 χρόνια δεδομένων – εξαρτάται και από το μήκος της εποχικής περιόδου – για να κατασκευαστεί ένα ARIMA μοντέλο.
2. Δεν υπάρχουν εύκολοι τρόποι, για να αναβαθμιστούν οι παράμετροι ενός ARIMA μοντέλου, καθώς νέα δεδομένα γίνονται διαθέσιμα, όπως υπάρχουν σε μερικές μεθόδους εξομάλυνσης. Το μοντέλο πρέπει περιοδικά να προσαρμόζεται ολοκληρωτικά και μερικές φορές πρέπει να αναπτύσσεται ένα νέο μοντέλο.
3. Η κατασκευή ενός ικανοποιητικού ARIMA μοντέλου συχνά απαιτεί μεγάλο χρόνο επένδυσης και άλλες πηγές. Τα κόστος για διαχείριση δεδομένων και ανάπτυξη μοντέλου μπορεί να είναι σημαντικά υψηλότερο για ARIMA μοντέλα, από ότι είναι για τις πιο παραδοσιακές τεχνικές πρόβλεψης, όπως η εξομάλυνση.

9.3 CASE 9-4 THE LYDIA E. PINKHAM MEDICINE COMPANY

Η φαρμακευτική εταιρία Lydia E. Pinkham ήταν μια οικογενειακή επιχείρηση, της οποίας το κύριο εισόδημα προερχόταν από την πώληση των συμπληρωμάτων λαχανικών της Lydia Pinkham. Πιθανόν στις μέρες

μας, πολλοί μαθητές και φοιτητές να μπορούσαν να χρησιμοποιούν αυτά τα συμπληρώματα για ανακούφιση από το άγχος. Δυστυχώς όμως σταμάτησε η κυκλοφορία τους στην αγορά. Η «καμπάνια» που προωθήθηκε αφορούσε τις γυναίκες. Οι διαφημίσεις για τα συμπληρώματα περιλάμβαναν πρόσκληση σε όλες τις γυναίκες να γράψουν απολύτως ελεύθερα στην κυρία Pinkham, στο Lynn Mass και να εξασφαλίσουν τις συμβουλές που προσφέρει δωρεάν και εχέμυθα. Αυτές οι συμβουλές, συνήθως, έφερναν ηρεμία σε νοικοκυριά στα οποία επικρατούσε η νευρικότητα και οι οξύθυμες συμπεριφορές. Στην πραγματικότητα, η εταιρία εξασφάλιζε ότι μια γυναίκα εργαζόμενη θα απαντούσε σε κάθε γράμμα. Ορισμένα από τα γράμματα γυναικών που εστάλησαν, έλεγαν: χωρίς τα προϊόντα και τις συμβουλές της κυρίας Pinkham, θα ήμουν πολύ χειρότερα, θα ήμουν παράφρων, θα είχα φτάσει ακόμα και στο θάνατο... Είχα παραδοθεί στον εαυτό μου: δοκίμαζα πολλά πράγματα, αλλά ήμουν πεπεισμένη ότι τίποτα πλέον δεν θα πήγαινε καλά. Χίλια ευχαριστώ για τις ιατρικές και ψυχολογικές συμβουλές σας, τώρα είμαι καλά και δυνατή, κυριολεκτικά είμαι άλλος άνθρωπος. Αυτή η μαρτυρία και άλλες παρόμοιες, αναπαράγονται σε έντυπες διαφημίσεις των προϊόντων της Lydia Pinkham.

Η φύση της εταιρίας που χαρακτηριζόταν από μοναδικότητα - το κυρίαρχο προϊόν που αντιπροσώπευε το μεγαλύτερο μέρος των πωλήσεων της εταιρίας, χωρίς το προσωπικό πωλήσεων, καθώς και το μεγάλο μέρος των εσόδων από πωλήσεις που επενδυόταν σε διαφήμιση οδήγησε την επιτροπή καθορισμού των τιμών, του Εθνικού Γραφείου Οικονομικών Ερευνών να επισημάνει το 1943 ότι τα δεδομένα πρέπει να υποβάλλονται σε λεπτομερή ανάλυση. Η έρευνα δεν πραγματοποιήθηκε για πολλά χρόνια (Palda 1964). Οι αναλυτές έχουν μελετήσει τα

δεδομένα χρησιμοποιώντας αιτιολογικά μοντέλα τα οποία περιλαμβάνουν τα δεδομένα διαφήμισης και άλλες οικονομικές μεταβλητές (παρόμοιες με αυτές που παρουσιάζονται στο κεφάλαιο 8). Ωστόσο, διάφορες μελέτες προτείνουν πως η Box-Jenkins προσέγγιση χρησιμοποιεί μόνο τα δεδομένα πωλήσεων για να παρέχει συγκρίσιμες, ή ακόμα και καλύτερες, προβλέψεις συγκριτικά με τις αιτιολογικές προσεγγίσεις (βλέπε Kyle 1978). Τα δεδομένα πωλήσεων είναι ιδανικά για μελέτη για δύο λόγους:

1. Το ίδιο το προϊόν παρέμεινε αμετάβλητο στο δεδομένο διάστημα. Αυτό σημαίνει πως δεν υπάρχουν μεταβολές στη χρονοσειρά που να οφείλονται στις αλλαγές του προϊόντος.
2. Δεν υπήρξε καμία αλλαγή στη δύναμη των πωλήσεων στο δεδομένο διάστημα και το ποσό των εσόδων που δαπανήθηκε στη διαφήμιση ήταν σχεδόν σταθερό. Επιπλέον, δεν υπάρχουν μεταβολές στα δεδομένα που να οφείλονται σε ειδικές προωθητικές ενέργειες η σε άλλες στρατηγικές μάρκετιγκ.

Συνήθως, τα πραγματικά δεδομένα δεν ακολουθούν αυτήν την άποψη όσον αφορά τη συνέχεια των προϊόντων και των αγορών.

Το καθήκον είναι ο υπολογισμός του καταλληλότερου ARIMA μοντέλου για αυτά τα δεδομένα. Το μοντέλο που θα αναπτυχθεί, χρησιμοποιεί τα δεδομένα από το 1907 έως το 1948 και ελέγχεται με τη χρήση των δεδομένων από το 1949 έως το 1960, όπως φαίνεται στον πίνακα 9-18.

ΑΝΑΓΝΩΡΙΣΗ ΜΟΝΤΕΛΟΥ

Ένα υπολογιστικό πρόγραμμα, ικανό στη μοντελοποίηση ARIMA, χρησιμοποιήθηκε για να εξετάσει τα δεδομένα από το 1907 έως το 1948. Τα δεδομένα από το 1949 έως το 1960 χρησιμοποιήθηκαν για να

ελέγξουν την ικανότητα πρόβλεψης του επιλεγμένου μοντέλου. Προκαταρκτικές δοκιμές υποδεικνύουν πως τα δεδομένα είναι στάσιμα (κι αυτό γιατί δεν υπάρχει εμφανής τάση), έτσι δεν απαιτείται η χρήση της μεθόδου των διαφορών. Μετά από μελέτη των αυτοσυσχετίσεων και των μερικών αυτοσυσχετίσεων, διαπιστώθηκε πως ένα AR μοντέλο ήταν το πλέον κατάλληλο για τα δεδομένα. (Το σύνολο αυτοσυσχετίσεων (ACF) και μερικών αυτοσυσχετίσεων (PACF) για 10 χρονικές καθυστερήσεις δίνεται στον πίνακα 9-19). Οι αυτοσυσχετίσεις και οι μερικές αυτοσυσχετίσεις φαίνονται να είναι συμβατές με αυτές που ορίζει ένα AR(2) μοντέλο. Για την επαλήθευση της τάξης p του AR μοντέλου, χρησιμοποιήθηκε το κριτήριο πληροφόρησης του Akaike's (AIC) (βλέπε σχέση 9.7) για αυτοπαλινδρομικά μοντέλα p = 1, 2 και 3 τάξης. Το κριτήριο αυτό οδηγεί στην επιλογή του μοντέλου AR(2) για τα συγκεκριμένα δεδομένα.

Year	***Sales (thousands of current dollars)***	***Year***	***Sales (thousands of current dollars)***	***Year***	***Sales (thousands of current dollars)***
1907	1016	1925	3438	1943	2602
1908	921	1926	2917	1944	2518
1909	934	1927	2359	1945	2637
1910	976	1928	2240	1946	2177
1911	930	1929	2196	1947	1920
1912	1052	1930	2111	1948	1910
1913	1184	1931	1806	1949	1984
1914	1089	1932	1644	1950	1787
1915	1087	1933	1814	1951	1689
1916	1154	1934	1770	1952	1866
1917	1330	1935	1518	1953	1896
1918	1980	1936	1103	1954	1684
1919	2223	1937	1266	1955	1633
1920	2203	1938	1473	1956	1657
1921	2514	1939	1423	1957	1569
1922	2726	1940	1767	1958	1390

1923	3185	1941	2161	1959	1397
1924	3352	1942	2336	1960	1289

Πίνακας 9-18 Δεδομένα της φαρμακευτικής εταιρίας Lydia E. Pinkham

Lag k	ACF	PACF
1	0,915	0,915
2	0,765	-0,440
3	0,597	-0,061
4	0,424	-0,124
5	0,243	-0,168
6	0,059	-0,141
7	-0,0099	0,024
8	-0,240	-0,170
9	-0,356	-0,036
10	-0,440	-0,040

Πίνακας 9-19

ΕΚΤΙΜΗΣΗ ΚΑΙ ΕΛΕΓΧΟΣ ΜΟΝΤΕΛΟΥ

Ένα υπολογιστικό πρόγραμμα χρησιμοποιείται για να εκτιμήσει τους συντελεστές (συμπεριλαμβανομένου και του σταθερού όρου) του μοντέλου AR(2) για τα δεδομένα από το 1907 έως το 1948. Χρησιμοποιώντας τις εκτιμώμενες παραμέτρους, η προσαρμοσμένη εξίσωση είναι:

$$Y_t = 178,6 + 1,423Y_{t-1} - 0,521Y_{t-2} + \varepsilon_t$$

$$(0,137) \qquad (0,136)$$

όπου οι αριθμοί των παρενθέσεων κάτω από τους υπολογισμένους συντελεστές είναι τα τυπικά σφάλματά τους. Κάθε αυτοπαλινδρομικός συντελεστής είναι σημαντικά διαφορετικός από το μηδέν για κάθε επίπεδο σημαντικότητας. Οι αυτοσυσχετίσεις των καταλοίπων ήταν χαμηλές, και καθεμία ήταν κατά 95% εντός των ορίων σφάλματος. Το Ljung-Box Q στατιστικό είχε p τιμές ίσες με 0,63, 0,21 και 0,64 για τις ομάδες χρονοκαθυστερήσεων m = 12, 24, 36 αντίστοιχα. Το AR(2)

μοντέλο μοιάζει να είναι επαρκές για τα δεδομένα μας.

ΠΡΟΒΛΕΠΟΝΤΑΣ ΜΕ ΤΟ ΜΟΝΤΕΛΟ

Το τελικό βήμα στην ανάλυση του AR(2) μοντέλου για τα δοσμένα δεδομένα είναι να προβλέψει τις τιμές από το 1949 έως το 1960, χρησιμοποιώντας τις παρατηρούμενες τιμές στις δύο προηγούμενες χρονικές περιόδους κάθε φορά. (Για παράδειγμα, τα δεδομένα του 1958 χρησιμοποιούνται στον υπολογισμό της πρόβλεψης για το 1959). Η εξίσωση πρόβλεψης είναι

$$Y_t = 178,6 + 1,423Y_{t-1} - 0,521Y_{t-2}$$

και οι προβλέψεις και τα τυπικά σφάλματα αυτών φαίνονται στον πίνακα 9-20.

Year	***Actual***	***Forecast***	***Error***
1949	1984	1896,4	87,6
1950	1787	2006,7	-219,7
1951	1689	1687,8	1,2
1952	1866	1651	215
1953	1896	1953,9	-57,9
1954	1684	1904,4	-220,4
1955	1633	1587,1	45,9
1956	1657	1625	32
1957	1569	1685,7	-116,7
1958	1390	1548	-158
1959	1397	1339,1	47,9
1960	1289	1428,1	-139,1

Πίνακας 9-20

Παράλληλα, έχουν υπολογιστεί και κάποιοι στατιστικοί δείκτες αξιοπιστίας. Οι προβλέψεις για το μοντέλο AR(2) έχουν Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (MAPE) ίσο με 6,9% και Μέση Απόλυτη Απόκλιση (MAD) ίση με $112 (σε χιλιάδες δολάρια). Τα στοιχεία αυτά συγκρίνονται με τους δείκτες αξιοπιστίας των υπολοίπων αιτιολογικών μοντέλων που αναπτύσσονται σε άλλες έρευνες.

ΠΕΡΙΛΗΨΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Ένα απλό (με τον μικρότερο αριθμό παραμέτρων) AR(2) μοντέλο έχει προσαρμοστεί στα δεδομένα μας για τα χρόνια 1907-1948. Αυτό το μοντέλο παρέχει αρκετά ακριβείς προβλέψεις για τα χρόνια 1949-1960.

ΕΡΩΤΗΣΕΙΣ

1. **Μετά την ολοκλήρωση της ανάλυσης, έγιναν διαθέσιμα τα στοιχεία των πωλήσεων για το έτος 1961 ($1426). Ποια είναι η πρόβλεψη του μοντέλου για το έτος αυτό; Αν αυτό το έτος προσθεθεί στο δοσμένο πακέτο δεδομένων, ποιες θα είναι οι καινούριες τιμές MAPE και MAD;**

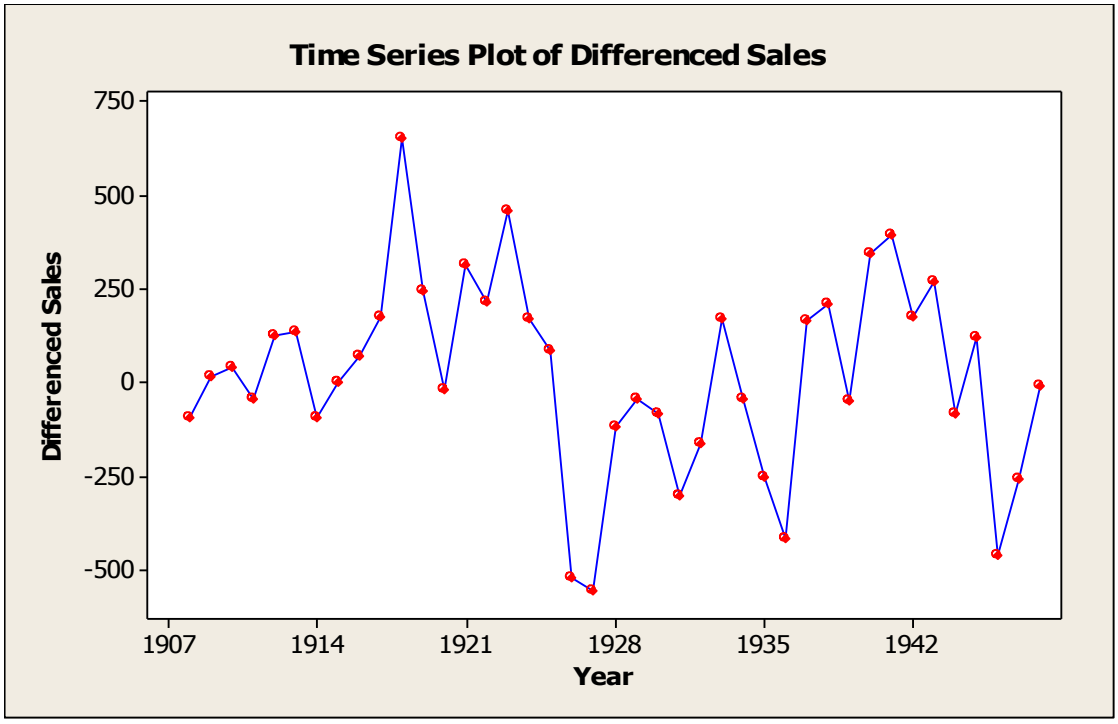
Η εξίσωση πρόβλεψης είναι η εξής $Y_t = 178,6 + 1,423Y_{t-1} – 0,521Y_{t-2}$. Για να βρούμε την πρόβλεψη του μοντέλου ARIMA για το έτος 1961, και με βάση τα στοιχεία των προηγουμένων ετών, θα έχουμε: $Y_{1961-1}=Y_{1960}=1289$ και $Y_{1961-2}=Y_{1959}=1397$. Άρα η προβλεπόμενη τιμή των πωλήσεων για το έτος 1961 θα είναι: $Y_{1961}=178,6+1,423*1289-0,521*1397=\$1285,01$. Η τιμή αυτή αποκλίνει από την πραγματική τιμή του έτους, που είναι ίση με $1,426. Με δεδομένα τα στοιχεία μέχρι το έτος 1960, ο στατιστικός δείκτης MAPE προκύπτει ίσος με 0,07 και ο στατιστικός δείκτης MAD ίσος με $113,82. Αν τα στοιχεία του έτους 1961 προστεθούν στο υπό μελέτη πακέτο δεδομένων, οι νέες τιμές των δεικτών θα είναι:

Year	Actual	Forecast	Error	APE
1949	1984	1896,21	87,79	0,044248992
1950	1787	2006,722	-219,722	0,122955792
1951	1689	1687,837	1,163	0,000688573
1952	1866	1651,02	214,98	0,115209003
1953	1896	1953,949	-57,949	0,030563819
1954	1684	1904,422	-220,422	0,130891924

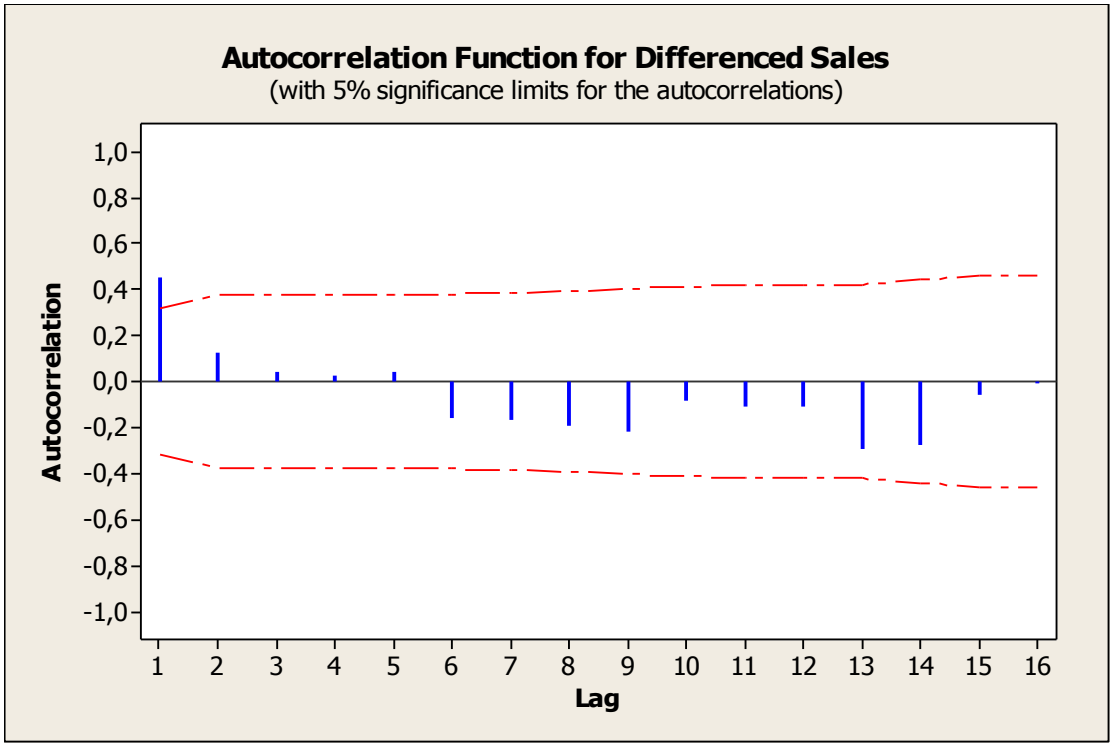
1955	1633	1587,116	45,884	0,028097979
1956	1657	1624,995	32,005	0,019315027
1957	1569	1685,718	-116,718	0,074390057
1958	1390	1547,99	-157,99	0,113661871
1959	1397	1339,121	57,879	0,041430923
1960	1289	1442,341	-153,341	0,11896121
1961	1426	1285,01	140,99	0,098870968

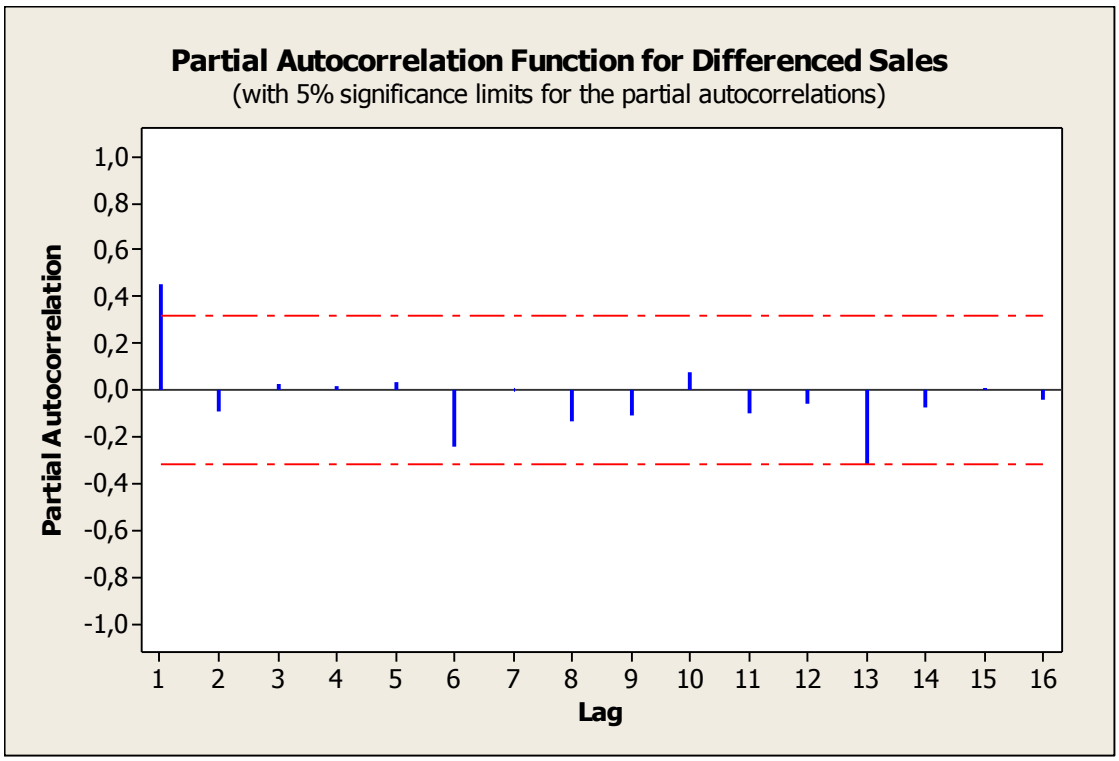
Με βάση τον παραπάνω πίνακα, ο δείκτης MAPE προκύπτει ίσος με MAPE=Σ APE/13=0,07225, ελάχιστα αυξημένος συγκριτικά με την προηγούμενη τιμή του (0,07). Αντίστοιχα, ο δείκτης MAD θα είναι ίσος με 115,91.

2. **Υπάρχουν κάποιες ενδείξεις πως τα διαθέσιμα δεδομένα δεν είναι στάσιμα. Για παράδειγμα, οι αυτοσυσχετίσεις δείγματος τείνουν να είναι μεγάλες για διαφορετικά lags. Χρησιμοποιήστε τη μέθοδο των διαφορών στα δεδομένα. Κατασκεύαστε ένα διάγραμμα των σειρών των διαφορών. Χρησιμοποιήστε το Minitab ή κάποιο άλλο υπολογιστικό πρόγραμμα για να προσαρμόσετε ένα ARIMA(1,1,0) μοντέλο στις ετήσεις πωλήσεις από το 1907 έως το 1948. Δημιουργήστε σταδιακές προβλέψεις για τα έτη 1949-1960. Ποιο μοντέλο εκ των AR(2), ARIMA(1,1,0) πιστεύετε πως είναι καλύτερο;**



Επειδή υπάρχουν ενδείξεις ότι η χρονοσειρά μας είναι μη στάσιμη (οι αυτοσυσχετίσεις φθίνουν σχετικά αργά στο μηδέν), καλούμαστε να εφαρμόσουμε τη μέθοδο των διαφορών στα δεδομένα για να δουμε αν θα μπορούσαμε να εξαλείψουμε την τάση και να δημιουργήσουμε στάσιμες σειρές. Οι διαφορές των χρονοκαθυστερημένων δεδομένων φαίνονται στον παρακάτω πίνακα. Το διάγραμμα των δεδομένων μετά τη μέθοδο των διαφορών, τα εμφανίζει να διαφέρουν γύρω από ένα σταθερό επίπεδο. Παράλληλα, τα διαγράμματα των αυτοσυσχετίσεων και των μερικών αυτοσυσχετίσεων φαίνονται παρακάτω:





Συμπεράνουμε ότι πλέον η χρονοσειρά μας έχει γίνει στάσιμη και άρα μπορούμε να προχωρήσουμε στην αναγνώριση του μοντέλου μας. Οι αυτοσυσχετίσεις εμφανίζονται να διακόπτονται μετά τη lag 1, υποδεικνύοντας MA(1) συμπεριφορά. Την ίδια στιγμή, οι μερικές

αυτοσυσχετίσεις εμφανίζονται να διακόπτονται μετά τη lag 1, υποδεικνύοντας AR(1) συμπεριφορά. Όμως, από την υπόθεση καλούμαστε να εφαρμόσουμε ένα ARIMA (1,1,0) μοντέλο. Τα αποτελέσματα του Minitab για το μοντέλο αυτό φαίνονται παρακάτω:

ARIMA Model: Sales

```
Estimates at each iteration
Iteration      SSE    Parameters
        0  2464192  0,100  19,714
        1  2243690  0,250  16,222
        2  2142995  0,400  12,519
        3  2136053  0,448  10,794
        4  2136022  0,451  10,435
        5  2136022  0,451  10,400
        6  2136022  0,451  10,398
Relative change in each estimate less than 0,0010
Final Estimates of Parameters
Type         Coef  SE Coef     T      P
AR   1    0,4511   0,1429  3,16  0,003
Constant   10,40    36,53  0,28  0,777

Differencing: 1 regular difference
Number of observations:  Original series 42,
after differencing 41
Residuals:    SS =  2133918 (backforecasts
excluded)
              MS =  54716  DF = 39

Modified Box-Pierce (Ljung-Box) Chi-Square
statistic
Lag             12     24     36  48
Chi-Square     5,5   22,1   25,3   *
DF              10     22     34   *
P-Value     0,858  0,456  0,860   *
```

Το τελικό βήμα στην ανάλυση του ARIMA(1,1,0) μοντέλου για τα δοσμένα δεδομένα είναι να προβλέψει τις τιμές από το 1949 έως το 1960, χρησιμοποιώντας τις παρατηρούμενες τιμές στις δύο προηγούμενες χρονικές περιόδους κάθε φορά. (Για παράδειγμα, τα δεδομένα του 1958

χρησιμοποιούνται στον υπολογισμό της πρόβλεψης για το 1959). Η εξίσωση πρόβλεψης είναι

$$\Delta Y_t = 10,4 + 0,4511\Delta Y_{t-1}$$

Θα πρέπει τώρα να δημιουργήσουμε σταδιακές προβλέψεις για τα έτη 1949 – 1960. Για παράδειγμα, η προβλεπόμενη τιμή για το έτος 1949 δίνεται από τη σχέση:

$\Delta Y_t = 10,4 + 0,4511\Delta Y_{t-1} \rightarrow Y_t - Y_{t-1} = 10,4 + 0,4511(Y_{t-1} - Y_{t-2}) \rightarrow$

$Y_{1949} = Y_{1948} + 10,4 + 0,4511(Y_{1948} - Y_{1947}) = 1915,89.$

Year	***Actual***	***Forecast***	***Error***	***APE***
1949	1984	1915,89	68,11	0,034329637
1950	1787	2027,78	-240,78	0,134739787
1951	1689	1708,53	-19,53	0,011563055
1952	1866	1655,19	210,81	0,112974277
1953	1896	1956,24	-60,24	0,031772152
1954	1684	1919,93	-235,93	0,14010095
1955	1633	1598,77	34,23	0,020961421
1956	1657	1620,39	36,61	0,022094146
1957	1569	1678,23	-109,23	0,069617591
1958	1390	1539,7	-149,7	0,107697842
1959	1397	1319,65	77,35	0,055368647
1960	1289	1410,56	-121,56	0,094305663

Με βάση τα παραπάνω στοιχεία, ο δείκτης MAPE προκύπτει ίσος με MAPE = Σ APE / 12 = 0,06962, δηλαδή 6,96%. Αντίστοιχα, ο δείκτης MAD θα είναι ίσος με 113,67. Παρατηρούμε ότι οι στατιστικοί δείκτες αξιοπιστίας έχουν σχεδόν ίδιες τιμές στα δύο ARIMA μοντέλα. Επομένως, θα στραφούμε στην αρχή της οικονομίας προκειμένου να αποφασίσουμε πιο μοντέλο είναι καλύτερο. Σύμφωνα με αυτήν, τα απλά μοντέλα προτιμώνται από τα σύνθετα. Άρα θα επιλεγεί το μοντέλο AR(2) για τη δημιουργία προβλέψεων.

3. Τα δεδομένα της Lydia Pinkham είναι ενδιαφέροντα λόγω της

μοναδικότητας της φύσης του προϊόντος και του μάρκετιγκ για την 54ετή πορεία που αναφέρεται. Ποιοι παράγοντες θα μπορούσαν να επηρεάσουν τις ετήσιες πωλήσεις στα αυτοκίνητα την ίδια χρονική περίοδο; Γιατί;

Τα αυτοκίνητα δεν έχουν μείνει σταθερά και μοναδικά την ίδια χρονική περίοδο. Σε τακτά χρονικά διαστήματα ανακοινώνονται από τις εταιρείες καινούργια μοντέλα. Αυτό σημαίνει ανάπτυξη και έρευνα. Ο ανταγωνισμός είναι μεγάλος, οπότε η διαφήμιση δεν είναι σταθερή και παίζει σημαντικό ρόλο. Το σέρβις των αυτοκινήτων επίσης παίζει σημαντικό ρόλο, όπως η αξία τους και η αξιοπιστία τους. Επομένως η ARIMA μεθοδολογία μπορεί να χρησιμοποιηθεί ως πρώτη προσέγγιση για πρόβλεψη, αλλά η αιτιοκρατική μέθοδος (multivariate regression) πρέπει να είναι σημαντικότερη στην πρόβλεψη των πωλήσεων, γιατί έτσι αποκαλύπτονται οι παράγοντες, που είναι σημαντικοί για την επικράτηση ενός μοντέλου μέσα στον ανταγωνισμό ή όχι.

ΚΕΦΑΛΑΙΟ 10 (ΘΕΩΡΙΑ ΥΙΟΘΕΤΗΣΗΣ ΚΑΙ ΔΙΑΧΥΣΗΣ – ADOPTION AND DIFFUSION THEORY)

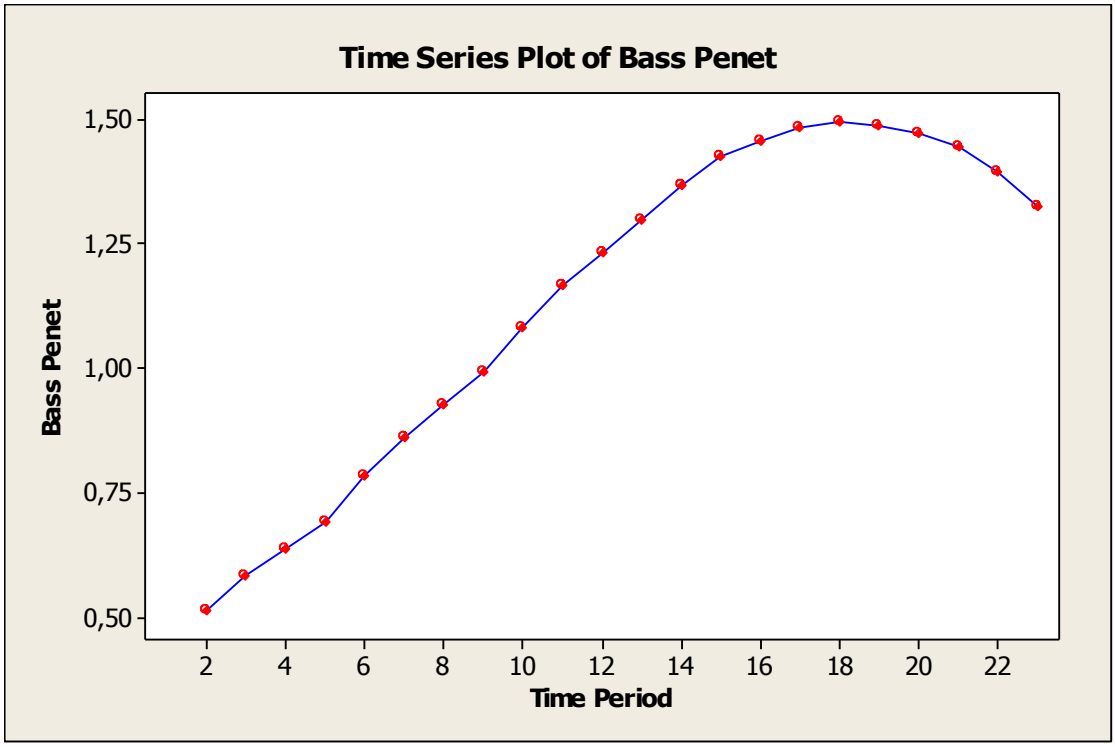
BASS 1969

MANAGEMENT SCIENCE

Vol. 15, No 5, p. 215

1.1 ΕΙΣΑΓΩΓΗ

Η θεωρία αυτή έχει να κάνει με την εκλογή του χρόνου εισαγωγής διαρκών καταναλωτικών αγαθών στην αγορά. Η θεωρία αυτή όμως εφαρμόστηκε και για την περιγραφή της αύξησης των πωλήσεων κατά τη διάρκεια του χρόνου, μετά την εισαγωγή των προϊόντων στην αγορά.



Διάφοροι ερευνητές ανέπτυξαν μοντέλα περιγραφής της αύξησης πωλήσεων θεωρώντας εκθετικές συναρτήσεις προς κάποια ασύμπτωτη. Στη θεωρία Bass το αυξητικό μοντέλο είναι παρόμοιο με το παραπάνω σχήμα αύξησης. Οι τιμές δηλαδή αυξάνονται μέχρι μία μέγιστη τιμή και μετά μειώνονται και σταθεροποιούνται σε κάποιο επίπεδο χαμηλότερο

του μεγίστου.

Ο σταθεροποιητικός παράγοντας είναι συνάρτηση της αύξησης του παράγοντα του κανονικού ρυθμού πωλήσεων και της μείωσης του παράγοντα της αρχικής φάσης των πωλήσεων. Εδώ θα ασχοληθούμε με την αρχική φάση των πωλήσεων.

1.2 ΘΕΩΡΙΑ ΥΙΟΘΕΤΗΣΗΣ ΚΑΙ ΔΙΑΧΥΣΗΣ

Στη συζήτηση, που θα ακολουθήσει, γίνεται προσπάθεια να περιγραφούν οι γενικές γραμμές της αρχικής υιοθέτησης ενός αγαθού από την αγορά, λαμβάνοντας υπόψη κοινωνικές συμπεριφορές.

Σε ένα κοινωνικό σύστημα κάποια άτομα αποφασίζουν να υιοθετήσουν μία καινοτομία ανεξάρτητα από κάποια άλλα άτομα. Αυτοί ονομάζονται "Innovators". Σύμφωνα με τον Rogers υπάρχουν οι ακόλουθες κλάσεις ατόμων, που υιοθετούν μία καινοτομία: 1) "Innovators", 2) "Early Adopters", 3) "Early Majority", 4) "Late Majority" και 5) "Laggards".

Εκτός από τους "Innovators" οι υπόλοιποι επηρεάζονται ως προς το χρόνο υιοθέτησης μιας καινοτομίας από την πίεση του κοινωνικού συστήματος και η πίεση αυτή αυξάνει ολοένα και περισσότερο για τους καθυστερημένους. Στην μαθηματική επεξεργασία του παρόντος εξειδικεύουμε δύο ομάδες: τους "Innovators" και τους τους υπόλοιπους ως "Imitators". Οι Innovators αλληλεπιδρούν μεταξύ τους, δεν επηρεάζονται όμως από τους υπόλοιπους με την έννοια, ότι δεν υιοθετούν κάποια καινούργια τεχνολογία, επειδή αυξάνονται οι πωλήσεις ενός καινοτόμου προϊόντος. Μάλλον το αντίθετο συμβαίνει: όταν αρχίζουν να το αγοράζουν πολλοί, αυτοί έχουν μειωμένο ενδιαφέρον. Ο Rogers υιοθέτησε αυθαίρετα, ότι οι Innovators είναι το 1,5% των

Adopters.

Για την μαθηματική περιγραφή των πωλήσεων προϊόντων χρησιμοποιούμε μοντέλα διάχυσης. Το θεμελιώδες Diffusion Model μπορεί να εκφρασθεί από την ακόλουθη διαφορική εξίσωση:

$$\frac{d\,Y(T)}{dT} = P(T)(m - Y(T))$$

Με την αρχική συνθήκη: Για T=0, Y(0) = 0.

Όπου:

- $Y(t)$: ο αριθμός (αθροιστικά) των μελών, που υιοθετούν σε χρόνο t.
- m: είναι ο αριθμός των δυνητικών μελών, που μπορούν να υιοθετήσουν, ή με άλλα λόγια ο συνολικός αριθμός της αγοράς, αλλά κάτι μικρότερο από το σύνολο του πληθυσμού και αυτό γιατί δεν υιοθετούν όλα τα μέλη, αλλά κάποια μένουν «άνοσα».
- $\frac{d\,Y(T)}{dT}$: η μεταβολή του αριθμού υιοθέτησης.
- P(T): ο συντελεστής της διάχυσης.

Για τον συντελεστή αναλογίας P(T) υπάρχουν τρεις περιπτώσεις:

1. P(T) = a = ct. Με αυτή τη μορφή η εξίσωση διάχυσης λέγεται External – Influence Diffusion Model, δηλαδή έχουμε μόνο εξωγενείς παράγοντες.
2. P(T) = b Y(T). Με αυτή τη μορφή η εξίσωση διάχυσης λέγεται Internal – Influence Diffusion Model, δηλαδή έχουμε ενδογενείς παράγοντες. Στην προκείμενη περίπτωση δεχόμαστε, ότι οι Adopters εξαρτώνται από τον αριθμό προηγούμενων Adopters. Υπάρχει δηλαδή μιμητισμός.
3. P(T) = a + b Y(T): Ένα δηλαδή μεικτό μοντέλο.

1.3 ΥΠΟΔΕΙΓΜΑ BASS

Για να αναπτύξουμε τη θεωρία BASS δεχόμαστε μία βασική υπόθεση: **"*Η πιθανότητα μία αρχική αγορά να γίνει στο χρόνο T, δεδομένου ότι δεν έχει προηγηθεί κάποια άλλη αγορά, είναι μία γραμμική συνάρτηση των προηγούμενων αγοραστών". Ή με λίγα λόγια ο Bass ορίζει την αναλογία, κατά την οποία οι υποψήφιοι αγοραστές, οι οποίοι δεν έχουν υιοθετήσει ακόμη, κατά τη χρονική περίοδο μεταξύ T και T+ΔT με ΔT --> 0, αυξάνονται αναλογικά με τον πληθυσμό, ο οποίος έχει ήδη υιοθετήσει. Η αναλογία είναι ο συντελεστής διάχυσης.***

Οπότε ο συντελεστής P(T) στην εξίσωση διάχυσης λαμβάνει τη μορφή:

$$P(T) = p + \frac{q}{m} Y(T)$$

όπου: p, q/m είναι σταθερές και Y(T) είναι ο αριθμός των προηγούμενων αγοραστών αθροιστικά. Επειδή το Y(0)=0, το p θα είναι η πιθανότητα μιας αρχικής αγοράς στο χρόνο T=0. Με λίγα λόγια η αρχική πιθανότητα να αγοραστεί ένα προϊόν με την εισαγωγή του στην αγορά δεν είναι μηδέν, αλλά συγκεκριμένη p. Το p αντανακλά το ποσοστό όλων των adopters, που είναι Innovators, με την έννοια που ορίστηκε από τον Rogers. Το γινόμενο q/m με το Y(T) αντανακλά τις εξασκούμενες πιέσεις στους Imitators, ως ο αριθμός των προηγούμενων αγοραστών.

1.4 ΒΑΣΙΚΕΣ ΥΠΟΘΕΣΕΙΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Το μοντέλο χαρακτηρίζεται από τις ακόλουθες υποθέσεις:

1. Κατά τη διάρκεια της περιόδου ενδιαφέροντος (product life), θα υπάρχουν m αρχικές αγορές του προϊόντος. Οι πωλήσεις αντικατάστασης αποκλείονται. Άρα λοιπόν κατά την ανάλυση περιοριζόμαστε στο

χρονικό διάστημα, κατά το οποίο οι πωλήσεις αντικατάστασης λόγω αστοχίας υλικών, ή κλοπών, ή άλλων βλαβών αποκλείονται.

2. Αντικαθιστώντας στην εξίσωση διάχυσης το P(T) της προηγούμενης παραγράφου, έχουμε:

$$\frac{d\,Y(T)}{dT}=\left(p+\frac{q}{m}Y(T)\right)(m-Y(T))$$

Αντικαθιστώντας το $Y(T)=m\,F(T)$, όπου F(T) είναι το αθροιστικό ποσοστό διείσδυσης (αθροιστική πιθανότητα):

$$m\frac{d\,F(T)}{dT}=(p+q\,F(T))(m-m\,F(T))\Rightarrow$$

$$f(T)=(p+q\,F(T))(1-F(T))\Rightarrow$$

$$\frac{f(T)}{[1-F(T)]}=P(T)=p+q\,F(T)$$

Το f(T) είναι η πυκνότητα πιθανότητας αγοράς στο χρόνο T και

$$F(T)=\int_0^T f(t)\,dt\,,\ \ \mu\varepsilon\ \ F(0)=0$$

Επίσης μπορούμε να ορίσουμε και τις σχετικές πωλήσεις S ως ακολούθως:

$$Y(T)=m\,F(T)=\int_0^T S(t)\,dt=m\int_0^T f(t)\,dt$$

Y(T) είναι ο συνολικός αριθμός αγορών στο διάστημα (0,T), με m τον αριθμό αγορών στο διάστημα που κατασκευάστηκε η πυκνότητα πιθανότητας. Οπότε:

$$S(T)=m\,f(T)=P(T)[m-Y(T)]=\left[p+\frac{q}{m}\int_0^T S(t)\,dt\right]\left[m-\int_0^T S(t)\,dt\right]$$

και μετά από κάποιες πράξεις θα έχουμε:

$$S(T)=p\,m+(q-p)\,Y(T)-\frac{q}{m}[Y(T)]^2$$

Περιληπτικά οι αρχές συμπεριφοράς αυτών των υποθέσεων είναι:

1. Αρχικές αγορές ενός προϊόντος γίνονται τόσο από τους Innovators, όσο και από τους Imitators. Η σημαντική διαφορά τους είναι η αγοραστική επίδραση. Οι Innovators δεν επηρεάζονται στο χρόνο της αρχικής αγοράς από τους προηγούμενους αγοραστές, όπως συμβαίνει με τους Imitators. Με κάποια έννοια οι Imitators μαθαίνουν από τους προηγούμενους αγοραστές.
2. Η σημαντικότητα των Innovators είναι μεγάλη στην αρχή, αλλά μετά φθίνει μονότονα με το χρόνο.
3. Αναφερόμαστε στο p ως ο συντελεστής του Innovation και στο q ως τον συντελεστή του Imitation.

Επειδή:

$$f(T)=[1-F(T)][p+q\,F(T)]=p+(q-p)\,F(T)-q[F(T)]^2=\frac{dF(T)}{dT}$$

για να βρούμε το F(T) λύνουμε την πιο πάνω μη γραμμική διαφορική εξίσωση πρώτης τάξης. Η λύση της είναι:

$$F(T)=\frac{q-p\,e^{-(T+C)(p+q)}}{q\left(1+e^{-(T+C)(p+q)}\right)}$$

Η σταθερά ολοκλήρωσης C υπολογίζεται από την αρχική συνθήκη F(0) = 0. Οπότε:

$$-C=\left(\frac{1}{p+q}\right)\ln\left(\frac{q}{p}\right)\quad \kappa\alpha\iota$$

$$F(T)=\frac{1-e^{-(p+q)T}}{\frac{q}{p}e^{-(p+q)T}+1}$$

και

$$f(T)=\frac{(p+q)^2}{p}\frac{e^{-(p+q)T}}{\left(\frac{q}{p}e^{-(p+q)T}+1\right)^2}$$

$$S(T)=m\frac{(p+q)^2}{p}\frac{e^{-(p+q)T}}{\left(\frac{q}{p}e^{-(p+q)T}+1\right)^2} \quad \text{και}$$

$$Y(T)=m\frac{1-e^{-(p+q)T}}{\frac{q}{p}e^{-(p+q)T}+1}$$

Στο ακόλουθο σημείο, σημείο μηδενισμού της πρώτης παραγώγου της S ή f, η συνάρτηση παρουσιάζει μέγιστο με q>p:

$$T_{max}=\left(\frac{1}{p+q}\right)\ln\left(\frac{q}{p}\right)$$

1.5 ΙΔΙΟΤΗΤΕΣ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ

Το υπόδειγμα Bass λοιπόν υποθέτει, ότι η πιθανότητα ένας καταναλωτής να αγοράσει το νέο προϊόν εξαρτάται από δύο παράγοντες: Πρώτα από τη τάση του να υιοθετεί νέα προϊόντα, *p* (intrinsic tendency) και δεύτερον από μιμητισμό (social contagion), *q*. Το υπόδειγμα Bass έχει ωστόσο τρεις άγνωστες παραμέτρους το *p*, το *q* και το *m* δηλαδή το συντελεστή καινοτομίας, το συντελεστή μιμητισμού και το αρχικό μέγεθος της αγοράς αντίστοιχα. Όταν η τιμή της παραμέτρου *q* είναι μεγαλύτερη από αυτή της *p*, τότε ο αθροιστικός αριθμός των καταναλωτών *Y*(*T*), ακολουθεί μια χαρακτηριστική s-καμπύλη, η οποία παρατηρείται σε πραγματικά καινοτόμα προϊόντα.

Οι παράμετροι *q*, *p* μας πληροφορούν για τη ταχύτητα διάχυσης του νέου προϊόντος στην αγορά. Υψηλές τιμές της παραμέτρου *p*

δείχνουν, ότι η διάχυση του προϊόντος στην αγορά εκδηλώθηκε στην αρχή της διαδικασίας αλλά γρήγορα μειώθηκε. Οι υψηλές αυτές τιμές της *p* μετατοπίζουν τη καμπύλη διάχυσης αριστερά, ενώ χαμηλές τιμές τη μετατοπίζουν δεξιά. Οι εφαρμογές έχουν δείξει, πως οι τιμές της *p* κυμαίνονται μεταξύ 0 και 0.3. Η μέση τιμή της *p* στις μελέτες του Μάρκετινγκ έχει βρεθεί ίση με 0.03, διπλάσιο από αυτό, που υιοθέτησε αυθαίρετα ο Rogers. Βιομηχανικά διαρκή αγαθά τείνουν να έχουν μικρότερη τιμή *p,* έναντι των καταναλωτικών αγαθών.

Υψηλές τιμές της παραμέτρου *q* συνεπάγεται, ότι η διαδικασία διάχυσης είναι αρχικά αργή, αλλά ύστερα βαίνει αυξανόμενη. Οι υψηλές τιμές της δημιουργούν λεπτόκυρτη καμπύλη διάχυσης, ενώ οι χαμηλές τιμές της δημιουργούν πλατύκυρτη καμπύλη διάχυσης. Οι εφαρμογές έχουν δείξει, πως οι τιμές της *q* κυμαίνονται μεταξύ 0.3 και 0.9. Η μέση τιμή στις μελέτες Μάρκετινγκ έχει βρεθεί ίση με 0.38, ενώ τα βιομηχανικά διαρκή αγαθά τείνουν να έχουν υψηλότερη τιμή για τη παράμετρο *q,* έναντι των καταναλωτικών αγαθών.

Ωστόσο το υπόδειγμα του Bass παρουσιάζει ορισμένους περιορισμούς. Ο βασικότερος περιορισμός είναι ο απαιτούμενος αριθμός παρατηρήσεων (πωλήσεων ανά παρατήρηση). Επίσης έχει δειχθεί, ότι το υπόδειγμα αυτό προσαρμόζεται καλύτερα, όταν το πλήθος των παρατηρήσεων (πωλήσεις), που διαθέτουμε, καλύπτουν και τη περίοδο κορυφής τους, ώστε να είναι σταθερή η τιμή της παραμέτρου *m.* Σημειώνουμε βέβαια, ότι για πραγματικά νέα προϊόντα δεν υπάρχουν ιστορικά δεδομένα για τις πωλήσεις. Ύστερα δεν είναι εύκολο να προσδιοριστεί με ακρίβεια το μέγεθος της αγοράς *m.*

Κάποιες από τις υποθέσεις του υποδείγματος είναι απλουστευμένες όπως:

1. Οι καταναλωτές μπορούν να επηρεαστούν μόνο από τη διαφήμιση ή τις διαπροσωπικές σχέσεις τους.
2. Το δυνητικό μέγεθος της αγοράς είναι διαχρονικά σταθερό για τη καινοτομία.
3. Η διαδικασία διάχυσης δεν επηρεάζεται από τη διαδικασία διάχυσης άλλων καινοτομιών.
4. Οι ιδιότητες του προϊόντος, η αγορά και τα χαρακτηριστικά της πολιτικής Μάρκετινγκ δεν επηρεάζουν τη διαδικασία διάχυσης του προϊόντος.
5. Αντιστοιχεί ένα προϊόν σε κάθε καταναλωτή.
6. Δεν υπάρχουν περιορισμοί στη προσφορά.

Διάφορες μελέτες έχουν προτείνει επεκτάσεις του υποδείγματος Bass, όπως η εκτίμηση της παραμέτρου του μεγέθους της αγοράς *m* με τη χρήση εξωτερικών πληροφοριών, η χρήση μεταβλητών του Μάρκετινγκ, η εκτίμηση των παραμέτρων υποδείγματος με τεχνικές μεγίστης πιθανοφάνειας, ή μη γραμμικές διαδικασίες εκτίμησης, τη γενίκευση του υποδείγματος Bass κ.λ.π. Ωστόσο η αρχική έκφρασή του χρησιμοποιείται ακόμη και σήμερα.

1.6 ΤΟ ΔΙΑΚΡΙΤΟ ΑΝΑΛΟΓΟ

Το βασικό μοντέλο, όσον αφορά τη διείσδυση (σχετικές πωλήσεις), είναι:

$$S(T)=p\,m+(q-p)\,Y(T)-\frac{q}{m}[Y(T)]^2$$

Για να υπολογίσουμε τα p, q και m γράφουμε το διακριτό ανάλογο ως:

$$S_T = a + b\,Y_{T-1} + c\,Y_{T-1}^2$$

Από τα δεδομένα, που έχουμε για τη διείσδυση, εφαρμόζουμε δευτεροβάθμια παλινδρόμηση ανάμεσα στις σχετικές πωλήσεις και αθροιστικές πωλήσεις και βρίσκουμε τη καμπύλη αύξησης της S ενός προϊόντος. Η συνάρτηση S δίνεται στη χρονική περίοδο T, ενώ η συνάρτηση Y δίδεται στη χρονική περίοδο T-1 δηλαδή στο lag 1. Θα μπορούσε όμως η δεξιά πλευρά να θεωρηθεί και στη χρονική περίοδο T.

Τα p, q, m δίδονται μετά από τη λύση της δευτεροβάθμιας εξίσωσης από τις ακόλουθες σχέσεις:

$$m = \frac{-b \pm \sqrt{b^2 - 4ac}}{2c},$$

$$p = \frac{a}{m}, \quad q = p + b.$$

1.7 ΠΑΡΑΔΕΙΓΜΑ ΕΥΡΥΖΩΝΙΚΟΤΗΤΑΣ

Στο κεφάλαιο 6 είχαμε δει τη διείσδυση της ευρυζωνικότητας ανά τρίμηνο από το 2001 μέχρι το 2007 στην Ευρώπη και προσεγγίσαμε τη καμπύλη με γραμμική παλινδρόμηση ως προς το χρόνο. Εδώ θα εφαρμόσουμε δευτεροβάθμια παλινδρόμηση ανάμεσα στις σχετικές πωλήσεις ανά μονάδα χρόνου και τον αριθμό των προηγούμενων adopters και μετά θα δούμε τις χρονοσειρές στο αντίστοιχο διάγραμμα. Στο παράδειγμα αυτό έγινε η παραδοχή, ότι η ΕΕ έχει πληθυσμό 500 εκατομμύρια και ανά νοικοκυριό υπάρχουν 4 άτομα, επομένως 125 εκατομμύρια νοικοκυριά.

Time	Pen. %	Pen. (f)	Cum Pen. (F)	S(T)	Y(T)	Y(T-1)	S-Fit
1	0,6000	0,0060	0,0060	0,7500	0,7500		
2	0,6000	0,0060	0,0120	0,7500	1,5000	0,7500	0,6411

3	0,5000	0,0050	0,0170	0,6250	2,1250	1,5000	0,7271
4	0,5000	0,0050	0,0220	0,6250	2,7500	2,1250	0,7964
5	0,9000	0,0090	0,0310	1,1250	3,8750	2,7500	0,8635
6	0,8000	0,0080	0,0390	1,0000	4,8750	3,8750	0,9788
7	0,7000	0,0070	0,0460	0,8750	5,7500	4,8750	1,0754
8	0,8000	0,0080	0,0540	1,0000	6,7500	5,7500	1,1554
9	1,1000	0,0110	0,0650	1,3750	8,1250	6,7500	1,2415
10	1,2000	0,0120	0,0770	1,5000	9,6250	8,1250	1,3509
11	1,0000	0,0100	0,0870	1,2500	10,8750	9,6250	1,4583
12	1,2000	0,0120	0,0990	1,5000	12,3750	10,8750	1,5383
13	1,5000	0,0150	0,1140	1,8750	14,2500	12,3750	1,6227
14	1,5000	0,0150	0,1290	1,8750	16,1250	14,2500	1,7107
15	1,2000	0,0120	0,1410	1,5000	17,6250	16,1250	1,7792
16	1,4000	0,0140	0,1550	1,7500	19,3750	17,6250	1,8199
17	1,9000	0,0190	0,1740	2,3750	21,7500	19,3750	1,8516
18	1,5000	0,0150	0,1890	1,8750	23,6250	21,7500	1,8674
19	1,1000	0,0110	0,2000	1,3750	25,0000	23,6250	1,8577
20	1,2000	0,0120	0,2120	1,5000	26,5000	25,0000	1,8382
21	1,6000	0,0160	0,2280	2,0000	28,5000	26,5000	1,8049
22	1,6000	0,0160	0,2440	2,0000	30,5000	28,5000	1,7411
23	1,3000	0,0130	0,2570	1,6250	32,1250	30,5000	1,6550

Οι στήλες είναι οι ακόλουθες:

- Time: Ο χρόνος.
- Pen. %: Η διείσδυση επί τις εκατό ως προς τον πληθυσμό.
- Pen. (f): Η διείσδυση ως πυκνότητα πιθανότητας ως προς τον πληθυσμό με μέγιστο 1.
- Cum Pen. (F): Η αθροιστική πιθανότητα.
- S(T): Οι σχετικές πωλήσεις στο χρόνο T και T+ΔT. Το S(T) υπολογίζεται πολλαπλασιάζοντας το Pen. (f) με τον αριθμό των νοικοκυριών του πληθυσμού.
- Y(T): Οι αθροιστικές πωλήσεις στο χρόνο T.

- Y(T-1): Το Lag 1 του Y(T).
- S-Fit: Το εκτιμώμενο S από την παλινδρόμηση στο διακριτό μοντέλο.

Time	BASS Pen. Fit (f)	BASS F Fit	S-Formula	BASS Pen. Formula	BASS F Formula	Y(T) Formula	Gom. Fit
1		0,0051	0,5520	0,0044	0,0044	0,5520	1,1965
2	0,0051	0,0103	0,6220	0,0050	0,0094	1,1740	1,6390
3	0,0058	0,0161	0,6983	0,0056	0,0150	1,8723	2,1890
4	0,0064	0,0224	0,7811	0,0062	0,0212	2,6534	2,8560
5	0,0069	0,0294	0,8698	0,0070	0,0282	3,5232	3,6471
6	0,0078	0,0372	0,9640	0,0077	0,0359	4,4872	4,5663
7	0,0086	0,0458	1,0627	0,0085	0,0444	5,5499	5,6145
8	0,0092	0,0550	1,1645	0,0093	0,0537	6,7144	6,7892
9	0,0099	0,0650	1,2679	0,0101	0,0639	7,9823	8,0847
10	0,0108	0,0758	1,3707	0,0110	0,0748	9,3530	9,4926
11	0,0117	0,0874	1,4707	0,0118	0,0866	10,8237	11,0023
12	0,0123	0,0997	1,5651	0,0125	0,0991	12,3888	12,6012
13	0,0130	0,1127	1,6512	0,0132	0,1123	14,0400	14,2752
14	0,0137	0,1264	1,7261	0,0138	0,1261	15,7660	16,0096
15	0,0142	0,1406	1,7871	0,0143	0,1404	17,5531	17,7895
16	0,0146	0,1552	1,8321	0,0147	0,1551	19,3852	19,5997
17	0,0148	0,1700	1,8592	0,0149	0,1700	21,2444	21,4259
18	0,0149	0,1850	1,8673	0,0149	0,1849	23,1117	23,2545
19	0,0149	0,1998	1,8561	0,0148	0,1997	24,9678	25,0730
20	0,0147	0,2145	1,8261	0,0146	0,2144	26,7939	26,8698
21	0,0144	0,2290	1,7784	0,0142	0,2286	28,5722	28,6351
22	0,0139	0,2429	1,7149	0,0137	0,2423	30,2872	30,3601
23	0,0132	0,2561	1,6381	0,0131	0,2554	31,9252	32,0373
24			1,5505	0,0124	0,2678	33,4757	33,6609
25			1,4550	0,0116	0,2794	34,9307	35,2259
26			1,3544	0,0108	0,2903	36,2851	36,7288
27			1,2513	0,0100	0,3003	37,5364	38,1669
28			1,1480	0,0092	0,3095	38,6844	39,5385

29			1,0466	0,0084	0,3178	39,7310	40,8430
30			0,9486	0,0076	0,3254	40,6796	42,0800
31			0,8552	0,0068	0,3323	41,5348	43,2503
32			0,7674	0,0061	0,3384	42,3022	44,3549
33			0,6857	0,0055	0,3439	42,9879	45,3952
34			0,6103	0,0049	0,3488	43,5982	46,3730
35			0,5414	0,0043	0,3531	44,1396	47,2905
36			0,4788	0,0038	0,3569	44,6185	48,1500
37			0,4223	0,0034	0,3603	45,0408	48,9539
38			0,3717	0,0030	0,3633	45,4125	49,7048
39			0,3264	0,0026	0,3659	45,7388	50,4052
40			0,2861	0,0023	0,3682	46,0249	51,0578
41			0,2504	0,0020	0,3702	46,2753	51,6652
42			0,2188	0,0018	0,3720	46,4941	52,2299
43			0,1910	0,0015	0,3735	46,6852	52,7546
44			0,1666	0,0013	0,3748	46,8517	53,2415
45			0,1451	0,0012	0,3760	46,9968	53,6931

Οι στήλες είναι οι ακόλουθες:

- BASS Pen. Fit (f): S-Fit δια τον αριθμό των νοικοκυριών του πληθυσμού, δηλαδή η πυκνότητα πιθανότητας από το Fit.
- BASS F Fit: Η αθροιστική πιθανότητα από το Fit.
- S-Formula: Οι σχετικές πωλήσεις χρησιμοποιώντας την αναλυτική σχέση του μοντέλου Bass μετά τον προσδιορισμό των παραμέτρων από την παλινδρόμηση.
- BASS Pen. Formula: Η πυκνότητα πιθανότητας ως το πηλίκο του S-Formula με τον αριθμό των νοικοκυριών του πληθυσμού.
- Y(T) Formula: Οι αθροιστικές πωλήσεις από το S-Formula.
- Gom. Fit: Η προσαρμοσμένη καμπύλη Gompertz από τα αρχικά δεδομένα του Y(T).

Time	**f-Formula**	**F-Formula**

1	0,0115766	0,000000
2	0,0130435	0,012299
3	0,0146456	0,026132
4	0,0163808	0,041635
5	0,0182421	0,058936
6	0,0202169	0,078157
7	0,0222857	0,099401
8	0,0244216	0,122751
9	0,0265896	0,148255
10	0,0287467	0,175926
11	0,0308431	0,205728
12	0,0328231	0,237574
13	0,0346279	0,271316
14	0,0361985	0,306751
15	0,0374793	0,343617
16	0,0384223	0,381597
17	0,0389903	0,420336
18	0,0391605	0,459445
19	0,0389257	0,498522
20	0,0382956	0,537164
21	0,0372957	0,574990
22	0,0359650	0,611646
23	0,0343531	0,646826
24	0,0325162	0,680277
25	0,0305137	0,711803
26	0,0284039	0,741268
27	0,0262419	0,768594
28	0,0240764	0,793751
29	0,0219491	0,816759
30	0,0198937	0,837673

31	0,0179359	0,856579
32	0,0160942	0,873584
33	0,0143800	0,888810
34	0,0127995	0,902389
35	0,0113541	0,914454
36	0,0100417	0,925141
37	0,0088573	0,934580
38	0,0077943	0,942896
39	0,0068447	0,950207
40	0,0059999	0,956621
41	0,0052510	0,962238
42	0,0045892	0,967152
43	0,0040059	0,971443
44	0,0034930	0,975187
45	0,0030430	0,978450

Οι στήλες είναι οι ακόλουθες:

- Τα f-FORMULA και F-FORMULA δίνονται εδώ ως ποσοστά του πλήθους των αγοραστών και όχι του συνολικού πληθυσμού.

Στο παρακάτω διάγραμμα εφαρμόσαμε δευτεροβάθμια παλινδρόμηση και προσδιορίσαμε το Fit σύμφωνα με τη θεωρία Bass. Στα παρακάτω διαγράμματα φαίνονται οι χρονοσειρές, τόσο του αριθμού πωλήσεων στη μονάδα του χρόνου, όσο και το Fit αυτής σύμφωνα με τη θεωρία Bass και οι αθροιστικές πωλήσεις. Επίσης στο τελευταίο διάγραμμα φαίνεται και η πρόβλεψη από την καμπύλη Gompertz.

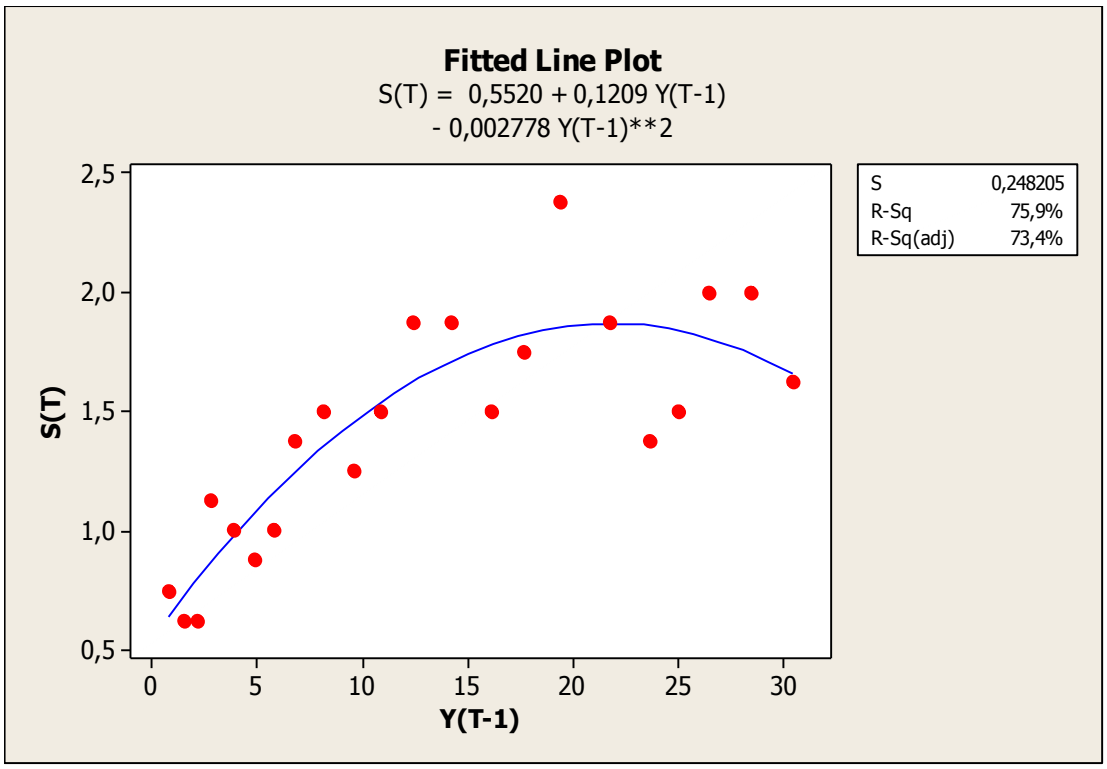
Για να κάνουμε πρόβλεψη αντικαθιστούμε τις τιμές των σταθερών p και q στην αναλυτική λύση, που δίνει το S(T) και προβλέπουμε τη ζήτηση-διείσδυση ως ποσοστό του πληθυσμού στο ζητούμενο χρονικό διάστημα. Οι αναλυτικές σχέσεις για το f και F δίνουν ποσοστά σε σχέση με το m και όχι με τον πληθυσμό.

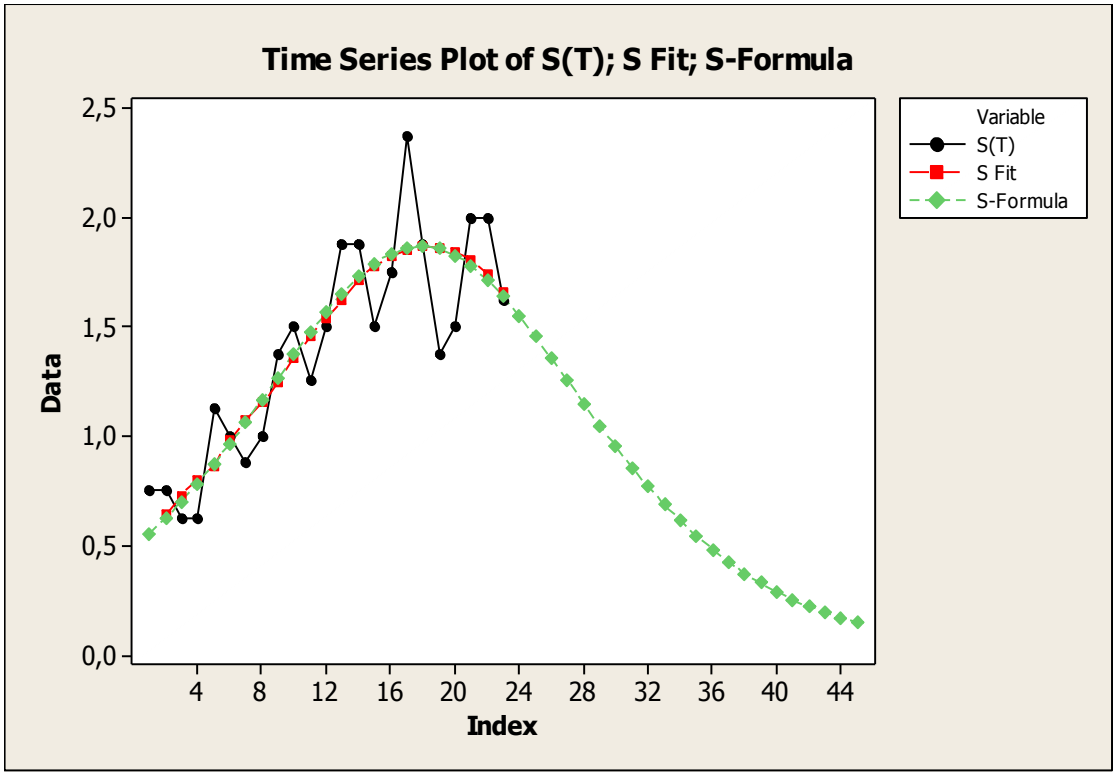
a=	0,5520071123	Δ=	0,144058879	m=	-4,166648197	47,68307232
b=	0,12090571			p=		0,011576584
c=	-2,77839E-03			q=		0,132482294

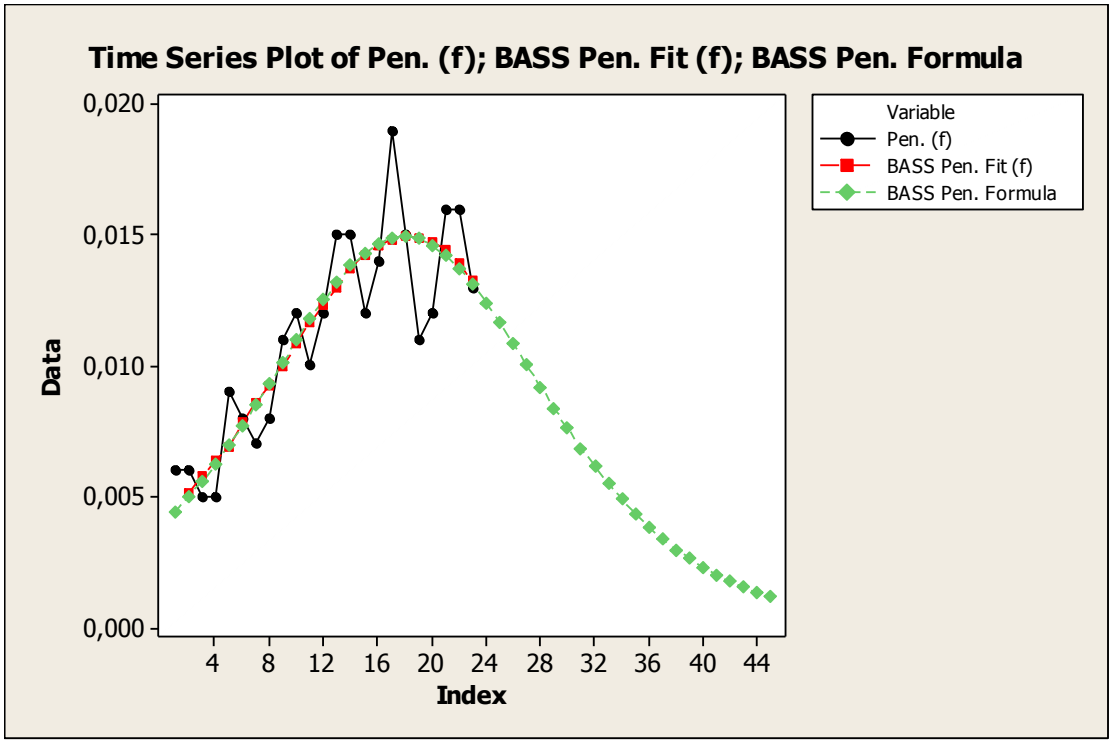
Οι τιμές των παραμέτρων είναι:

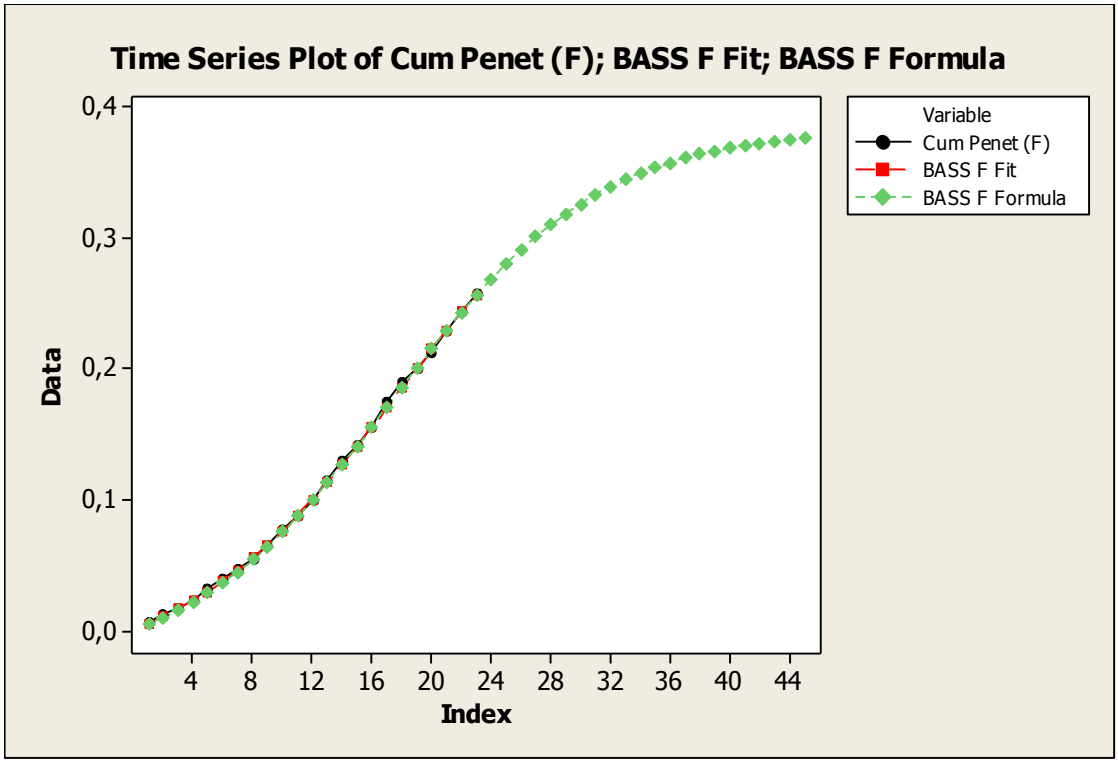
$$m=47,6831\,,\quad p=0,0116\,,\quad \kappa\alpha\iota\quad q=0,1325.$$

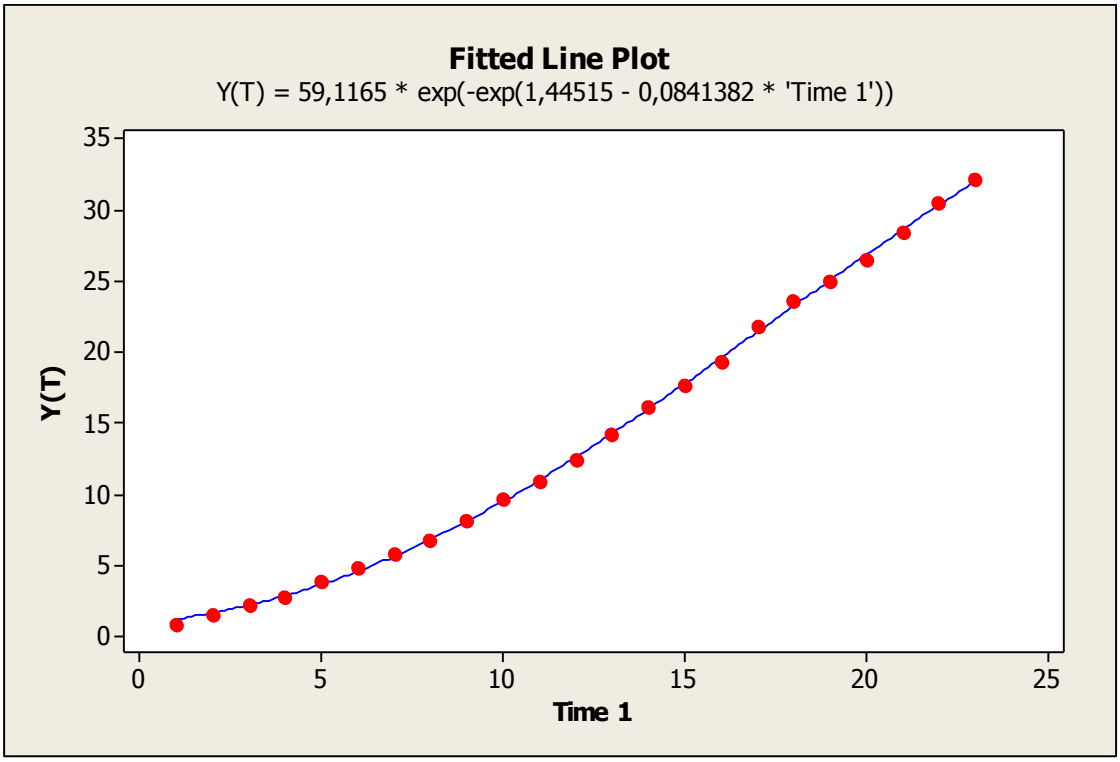
Το m αντιστοιχεί στον αριθμό των νοικοκυριών, που θα υιοθετήσουν την καινοτομία. Ο αριθμός αυτός είναι πολύ μικρότερος από τον αριθμό των νοικοκυριών του πληθυσμού. Στην προκείμενη περίπτωση δεν έχουμε λάβει υπόψη μας τις αρχές ωρίμανσης των πωλήσεων, αλλά μόνο αυτές της αρχικής εισαγωγής της καινοτομίας στην αγορά.

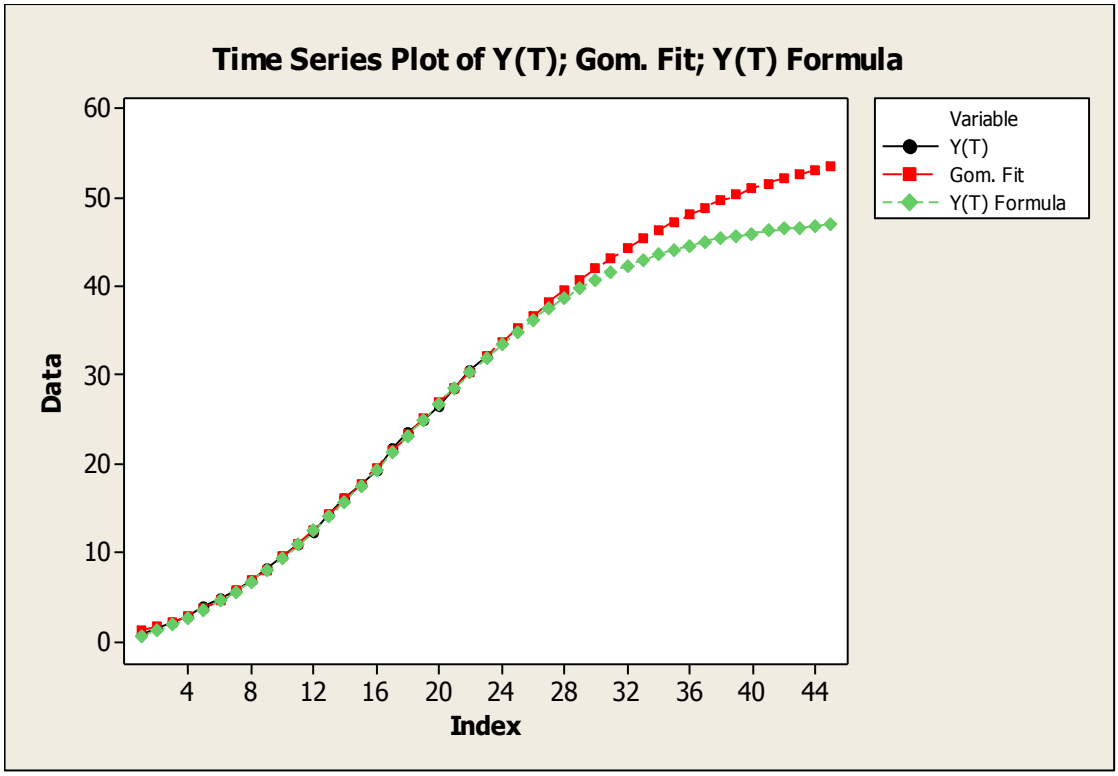



Time Series Plot of S(T); S Fit; S-Formula
Variable
S(T)
S Fit
S-Formula
Data
Index
2,5
2,0
1,5
1,0
0,5
0,0
4
8
12
16
20
24
28
32
36
40
44


Time Series Plot of Pen. (f); BASS Pen. Fit (f); BASS Pen. Formula
Variable
Pen. (f)
BASS Pen. Fit (f)
BASS Pen. Formula
Data
Index
0,020
0,015
0,010
0,005
0,000
4
8
12
16
20
24
28
32
36
40
44


Time Series Plot of Cum Penet (F); BASS F Fit; BASS F Formula
Variable
Cum Penet (F)
BASS F Fit
BASS F Formula
Data
0,4
0,3
0,2
0,1
0,0
4
8
12
16
20
24
28
32
36
40
44
Index


Fitted Line Plot
Y(T) = 59,1165 * exp(-exp(1,44515 - 0,0841382 * 'Time 1'))
Y(T)
35
30
25
20
15
10
5
0
0
5
10
15
20
25
Time 1


Time Series Plot of Y(T); Gom. Fit; Y(T) Formula
Data
Index
Variable
Y(T)
Gom. Fit
Y(T) Formula
0
10
20
30
40
50
60
4
8
12
16
20
24
28
32
36
40
44