National and Kapodistrian
UNIVERSITY OF ATHENS

Τεχνικές Ανάλυσης και Πρόβλεψης Τηλεπικοινωνιακών Αγορών

Πολυμεταβλητή Παλινδρόμηση

Πολυμεταβλητή Παλινδρόμηση

- Η πολυμεταβλητή παλινδρόμηση περιλαμβάνει τη χρήση περισσότερων από μία ανεξάρτητων μεταβλητών για την πρόβλεψη μίας εξαρτημένης μεταβλητής.

Πίνακας Συσχέτισης

- 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.

Πίνακας Συσχέτισης

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

Μοντέλο Πολυμεταβλητής Παλινδρόμησης

- Στην απλή παλινδρόμηση, η εξαρτημένη μεταβλητή συμβολίζεται με Υ και η ανεξάρτητη μεταβλητή με Χ.
- Στην ανάλυση πολυμεταβλητής παλινδρόμησης, τα Χ με κάτω δείκτες χρησιμοποιούνται για να συμβολίσουν τις ανεξάρτητες μεταβλητές. Η εξαρτημένη μεταβλητή ακόμα συμβολίζεται με Υ, και οι ανεξάρτητες μεταβλητές με X_1, X_2, …, X_k.
- Όταν καθοριστεί τα αρχικό σύνολο ανεξάρτητων μεταβλητών, η σχέση μεταξύ του Υ και των Χ μπορεί να εκφραστεί σαν μοντέλο πολυμεταβλητής παλινδρόμησης.

Μοντέλο Πολυμεταβλητής Παλινδρόμησης

- Στο μοντέλο πολυμεταβλητής παλινδρόμησης, η μέση παρατηρούμενη τιμή είναι η ακόλουθη γραμμική συνάρτηση των ανεξάρτητων μεταβλητών:

$$\mu_y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

- Αυτή η έκφραση είναι η συνάρτηση πολυμεταβλητής παλινδρόμησης του πληθυσμού.

Δομή Δεδομένων για Πολυμεταβλητή Παλινδρόμηση

Μεταβλητές Πρόβλεψης					Απόκριση
Περίπτωση	X_1	X_2	.....	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
.	.	.	.....	.	.

Στατιστικό Μοντέλο για Πολυμεταβλητή Παλινδρόμηση

- 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:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Στατιστικό Μοντέλο για Πολυμεταβλητή Παλινδρόμηση

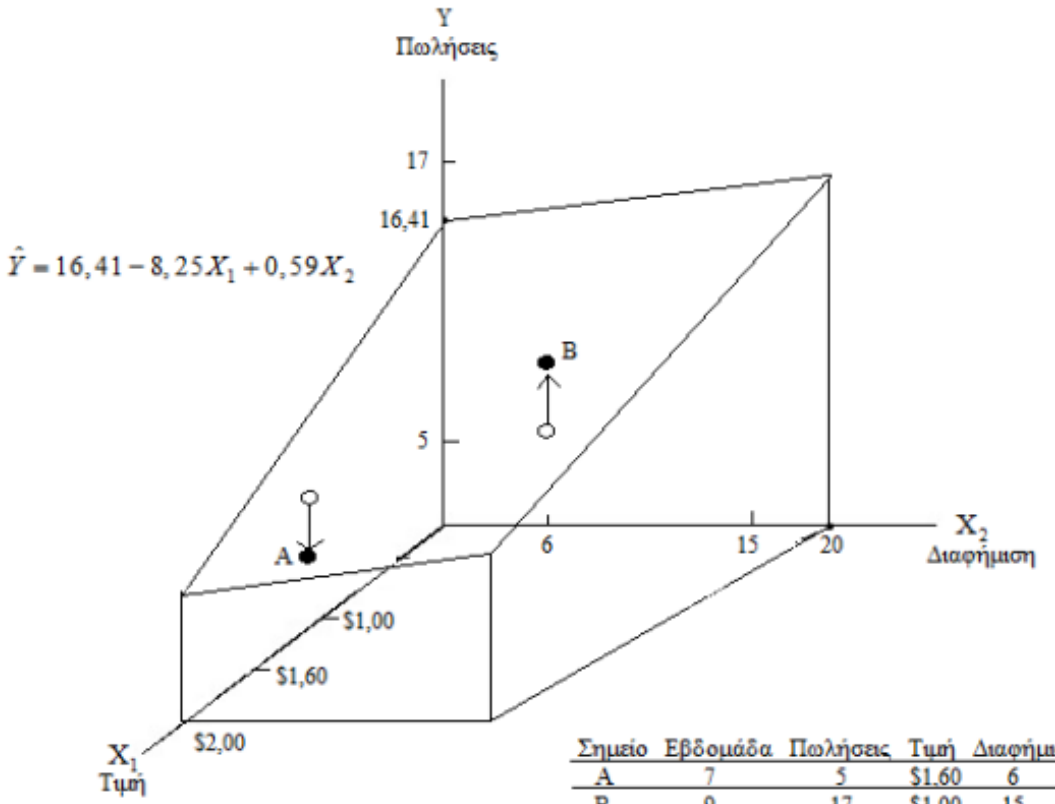
όπου:

1. Για κάθε i παρατήρηση, $Y=Y_i$ και τα $X_1, X_2, \ldots, X_k$ αντιστοιχούν στις τιμές $X_{i1}, X_{i2}, \ldots, X_{ik}$. Για τις $X_1, X_2, \ldots, X_k$ δεν υπάρχουν γραμμικές σχέσεις μεταξύ τους.
2. Τα ε, είναι συντελεστές σφάλματος, που αντιπροσωπεύουν τις αποκλίσεις της προβλεπόμενης τιμής από την αληθή σχέση.
3. Οι συντελεστές παλινδρόμησης, $\beta_0, \beta_1, \ldots, \beta_k$, που μαζί ορίζουν τη συνάρτηση παλινδρόμησης είναι άγνωστοι, όπως συνήθως και η διασπορά.

Στατιστικό Μοντέλο για Πολυμεταβλητή Παλινδρόμηση

$$\hat{Y} = 16{,}41 - 8{,}25\,X_1 + 0{,}59\,X_2$$

$$\hat{Y} = 16{,}41 - 8{,}25(1{,}5) + 0{,}59(10) = 9{,}935$$



Σημείο	Εβδομάδα	Πωλήσεις	Τιμή	Διαφήμιση
A	7	5	$1,60	6
B	9	17	$1,00	15

Ερμηνεία Συντελεστών Παλινδρόμησης

- 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.

Συμπεράσματα Πολυμεταβλητής Παλινδρόμησης

- Τα συμπεράσματα για την πολυμεταβλητή παλινδρόμηση είναι ανάλογα με αυτά για την απλή γραμμική παλινδρόμηση.
- Κάθε παρατήρηση Υ μπορεί να γραφεί ως:

$$\textbf{Observation} = \textbf{Fit} + \textbf{Residual}$$

$$\text{ή} \quad Y = \hat{Y} + (Y - \hat{Y})$$

$$\text{και με} \quad \hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Συμπεράσματα Πολυμεταβλητής Παλινδρόμησης

› Η ανάλυση του αθροίσματος των τετραγώνων και οι σχετικοί βαθμοί ελευθερίας είναι¨

$$\sum (Y-\bar{Y})^2 = \sum (\hat{Y}-\bar{Y})^2 + \sum (Y-\hat{Y})^2$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\text{n-1} = \text{k} + \text{n-k-1}$$

Τυπικό Σφάλμα Εκτίμησης

› 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).

Τυπικό Σφάλμα Εκτίμησης

- Το $S_{y \cdot x's}$ είναι ένας άριστος εκτιμητής του σ, διότι όπως αποδεικνύεται η αναμενόμενη τιμή του $S_{y \cdot x's}^2$ είναι το σ^2 , η διασπορά των σφαλμάτων. Δηλαδή:

$$E\left(s^2_{y \cdot x's}\right) = \sigma^2 = \frac{\left(\sum Y - \hat{Y}\right)^2}{n}$$

- 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.

Σημαντικότητα Παλινδρόμησης

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.

Σημαντικότητα Παλινδρόμησης

- Ο συντελεστής προσδιορισμού R^2 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.
- Δίνεται από τον τύπο:

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$$

Σημαντικότητα Παλινδρόμησης

- Μία τιμή του $R^2=1$ σημαίνει, ότι όλες οι παρατηρούμενες Y τιμές πέφτουν ακριβώς πάνω στην προσαρμοσμένη συνάρτηση παλινδρόμησης. Όλη η μεταβλητότητα στην απόκριση ερμηνεύεται από την παλινδρόμηση.
- Μία τιμή του $R^2=0$ σημαίνει ότι $\hat{Y}=Y$, (SSR=0) και ότι καμία μεταβολή στο Y δεν ερμηνεύεται από την παλινδρόμηση.
- Η ποσότητα $R=\sqrt{R^2}$ καλείται συντελεστής πολυμεταβλητής συσχέτισης και είναι η συσχέτιση μεταξύ των αποκρίσεων - παρατηρούμενων τιμών και των προσαρμοσμένων τιμών.

Σημαντικότητα Παλινδρόμησης

- Ο συντελεστής προσδιορισμού (R^2) μπορεί πάντα να αυξηθεί προσθέτοντας μία επιπλέον ανεξάρτητη μεταβλητή, Χ, 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.

Σημαντικότητα Παλινδρόμησης

- Ο προσαρμοσμένος συντελεστής προσδιορισμού (adjusted coefficient of determination) 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.

Σημαντικότητα Παλινδρόμησης

Μεταβλητές	**R^2**
Τιμή	0,75
Τιμή και διαφημιστικά έξοδα	0,93

Μεμονωμένες Μεταβλητές Πρόβλεψης

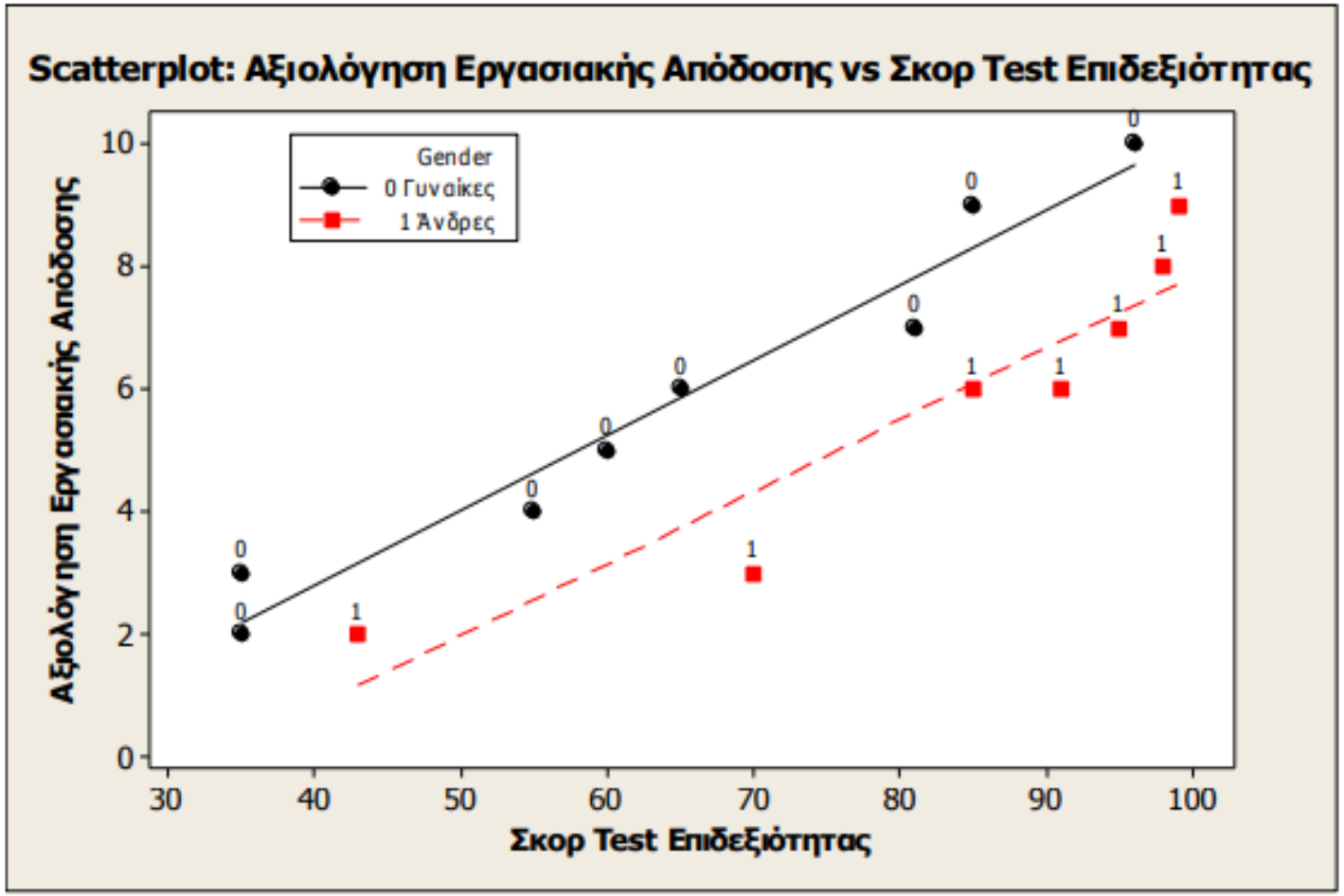
- Ο συντελεστής ενός μεμονωμένου Χ στη συνάρτηση παλινδρόμησης υπολογίζει τη μερική ή οριακή επίδραση εκείνου του Χ στην απόκριση Υτ κρατώντας όλα τα υπόλοιπα Χ στη εξίσωση σταθεράς

- Για να κρίνουμε τη σημαντικότητα του Ш όρουτ Ш ü ẅ ẗ ′ ẗ Ш ẗ στην εξίσωση παλινδρόμησηςτ ο έλεγχος του στατιστικού Ћ τ συγκρίνεται με ένα ποσοστιαίο σημείο της Ћ κατανομής με Ь υ Ш ẅ ü βαθμούς ελευθερίαςς

Ψευδομεταβλητές

- 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.

Ψευδομεταβλητές



Πολυσυγγραμικότητα

- 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.

Πολυσυγγραμικότητα

- Ο βαθμός της πολυσυγγραμμικότητας υπολογίζεται από τον συντελεστή διόγκωσης της διακύμανσης (Variance Inflation Factor), VIF:

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

- Όπου 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.

Πολυσυγγραμικότητα

- 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 $R_j^2=0,90$ $VIF_j=1/(1-0,9)=10$.

Πολυσυγγραμικότητα

- Υπάρχουν αρκετοί τρόποι να ελαχιστοποιήσουμε την πολυσυγραμμικότητα. Μερικοί από αυτούς είναι:
  - Δημιουργία νέων μεταβλητών (εξαρτημένης και ανεξάρτητης) σύμφωνα με τη σχέση

$$\tilde{X}_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{\sum_i (X_{ij} - \bar{X}_j)^2}}$$

  - Αφαίρεση από το μοντέλο της παλινδρόμησης των εξαρτώμενων μεταβλητών.
  - 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.

Επιλέγοντας την καλύτερη εξίσωση παλινδρόμησης

Επιλογή συνόλου πιθανών μεταβλητών πρόβλεψης

Απαλοιφή μη βασικών ανεξάρτητων μεταβλητών

Περιορισμός λίστας μεταβλητών πρόβλεψης βάσει στατιστικής σημαντικότητας

Επιλογή Συνόλου Πιθανών Μεταβλητών Πρόβλεψης

- 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.

Απαλοιφή μη Βασικών Ανεξάρτητων Μεταβλητών

› 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).

Περιορισμός Λίστας Μεταβλητών Πρόβλεψης Βάσει Στατιστικής Σημαντικότητας

- Αποσκοπεί στην απόκτηση της καλύτερης επιλογής των ανεξάρτητων μεταβλητών.
- Καμία από τις διαδικασίες έρευνας, δεν μπορεί να ειπωθεί, ότι οδηγεί στο καλύτερο σύνολο ανεξάρτητων μεταβλητών.
- 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.

Περιορισμός Λίστας Μεταβλητών Πρόβλεψης Βάσει Στατιστικής Σημαντικότητας

› Παρουσιάζονται τέσσερις διαδικασίες:

1. Όλες οι πιθανές παλινδρομήσεις (all possible regressions ή best subset)
2. Βήμα-προς-βήμα παλινδρόμηση (stepwise regression)
3. Ανάστροφη παλινδρόμηση (backward regression) και
4. Εμπρόσθια παλινδρόμηση (forward regression)

Όλες οι Πιθανές Παλινδρομήσεις

- 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.

Όλες οι Πιθανές Παλινδρομήσεις - Τεχνική R^2

- Το 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.

Όλες οι Πιθανές Παλινδρομήσεις - Τεχνική R^2

- 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.
- 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.

Βήμα προς Βήμα Παλινδρόμηση

- 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.

Βήμα προς Βήμα Παλινδρόμηση – Αλγόριθμος

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.

Βήμα προς Βήμα Παλινδρόμηση – Αλγόριθμος

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.

Βήμα προς Βήμα Παλινδρόμηση

- Η τεχνική της βήμα-προς-βήμα παλινδρόμησης είναι εξαιρετικά εύκολη στη χρήση.
- Είναι επίσης εξαιρετικά εύκολο να χρησιμοποιηθεί λανθασμένα:
  - Σφάλματα κατά τα t test.
  - Παράλειψη σημαντικών μεταβλητών κατά την αρχική επιλογή.

Ανάστροφη Παλινδρόμηση

- 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.

Εμπρόσθια Παλινδρόμηση

- 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.

Αναγνώριση Ακραίων Τιμών

- 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.

Αναγνώριση Ακραίων Τιμών

- Ένας υπολογισμός της επιρροής των i σημείων δεδομένων στη προσαρμοσμένη συνάρτηση παλινδρόμησης παρέχεται από το **leverage**.
- 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:

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}$$

Αναγνώριση Ακραίων Τιμών

- Αν το i σημείο δεδομένων έχει υψηλό leverage (h_{ii} είναι κοντά στο 1), η προσαρμοσμένη απόκριση σε εκείνο το X καθορίζεται σχεδόν ολοκληρωτικά από το X_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 hii είναι αρκετά μεγάλο, αν $h_{ii} \geq 3(k+1)/n$.

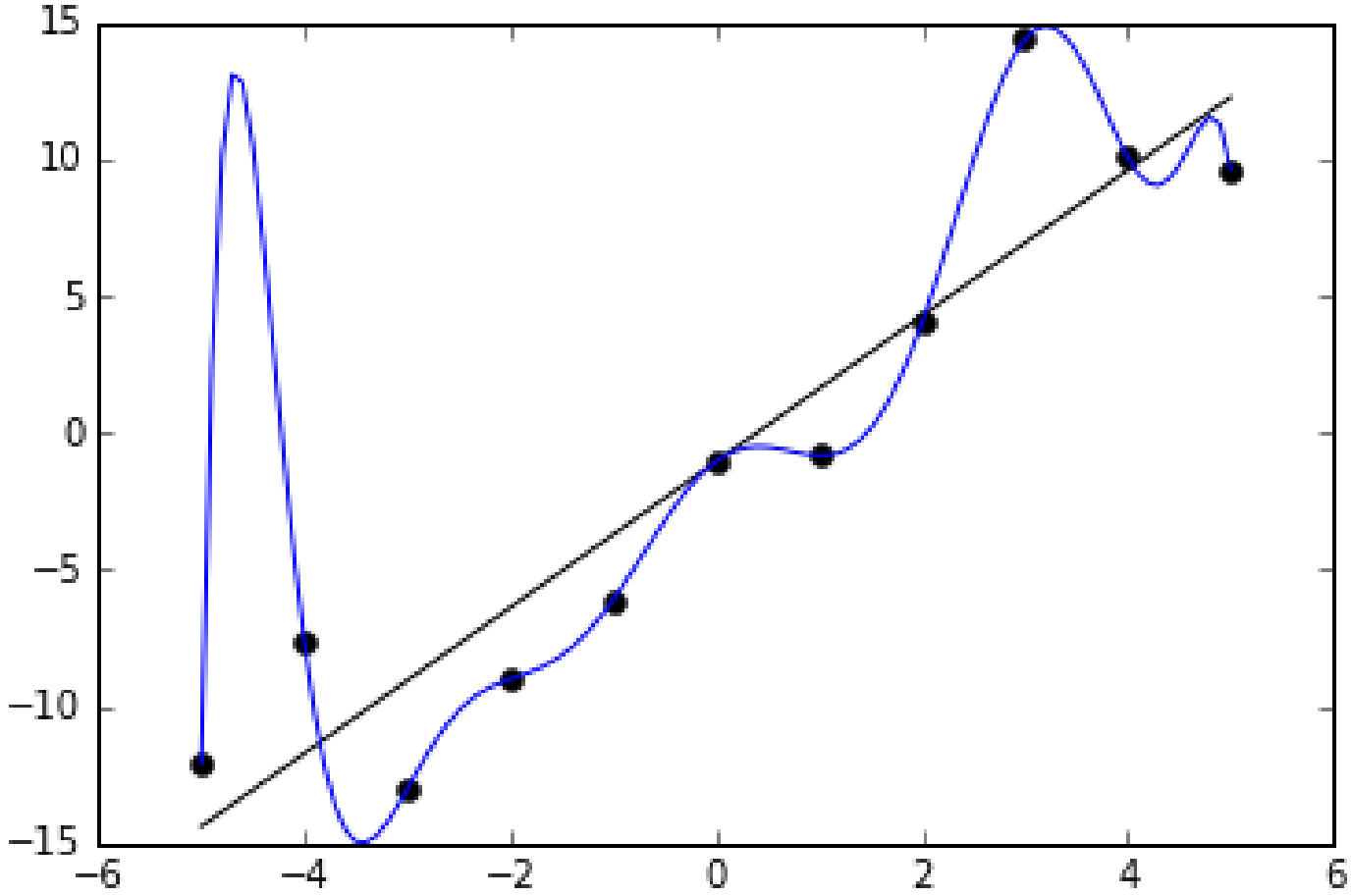
Αναγνώριση Ακραίων Τιμών

- Η ανίχνευση των ακραίων Υ τιμών βασίζεται στο μέγεθος των καταλοίπων.
- Μεγάλα κατάλοιπα υποδεικνύουν, ότι μία Υ τιμή είναι «μακριά» από την προσαρμοσμένη ή προβλεφθείσα τιμή.
- 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.

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.

Overfitting



Χρήσιμες Παλινδρομήσεις

- 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 τουλάχιστον τέσσερις φορές μεγαλύτερος της F-τιμής, που αντιστοιχεί στο επίπεδο σημαντικότητας α%, πριν χρησιμοποιήσουμε τη παλινδρόμηση για προβλέψεις.

Ερωτήσεις???