Στατιστική ΙΙ: Γραμμικό Μοντέλο

Λουκία Μελιγκοτσίδου

Εαρινό εξάμηνο, 2006

Αβεβαιότητα και Στατιστική Μοντελοποίηση

Σε ένα πρώτο μάθημα στατιστικής μαθαίνουμε για περιγραφικά μέτρα, αριθμητικά και γραφικά, για συμπερασματολογία για κάποιο πληθυσμό με βάση ένα δείγμα, κάποια στοιχεία θεωρίας πιθανοτήτων και κατανομών.

Το ενδιαφέρον της στατιστικής επιστήμης βρίσκεται στην κατανόηση στοχαστικών φαινομένων και την ποσοτικοποίηση της αβεβαιότητας σχετικά με αυτά.

Στατιστική Μοντελοποίηση: κατασκευή μοντέλων με λίγες παραμέτρους για την περιγραφή στοχαστικών φαινομένων - σχέσεων μεταξύ μεταβλητών.

Στατιστική Συμπερασματολογία: εκτίμηση των άγνωστων παραμέτρων με βάση παρατηρήσεις/δεδομένα και ποσοτικοποίηση της αβεβαιότητας σχετικά με τις εκτιμήσεις.

Γραμμικό Μοντέλο

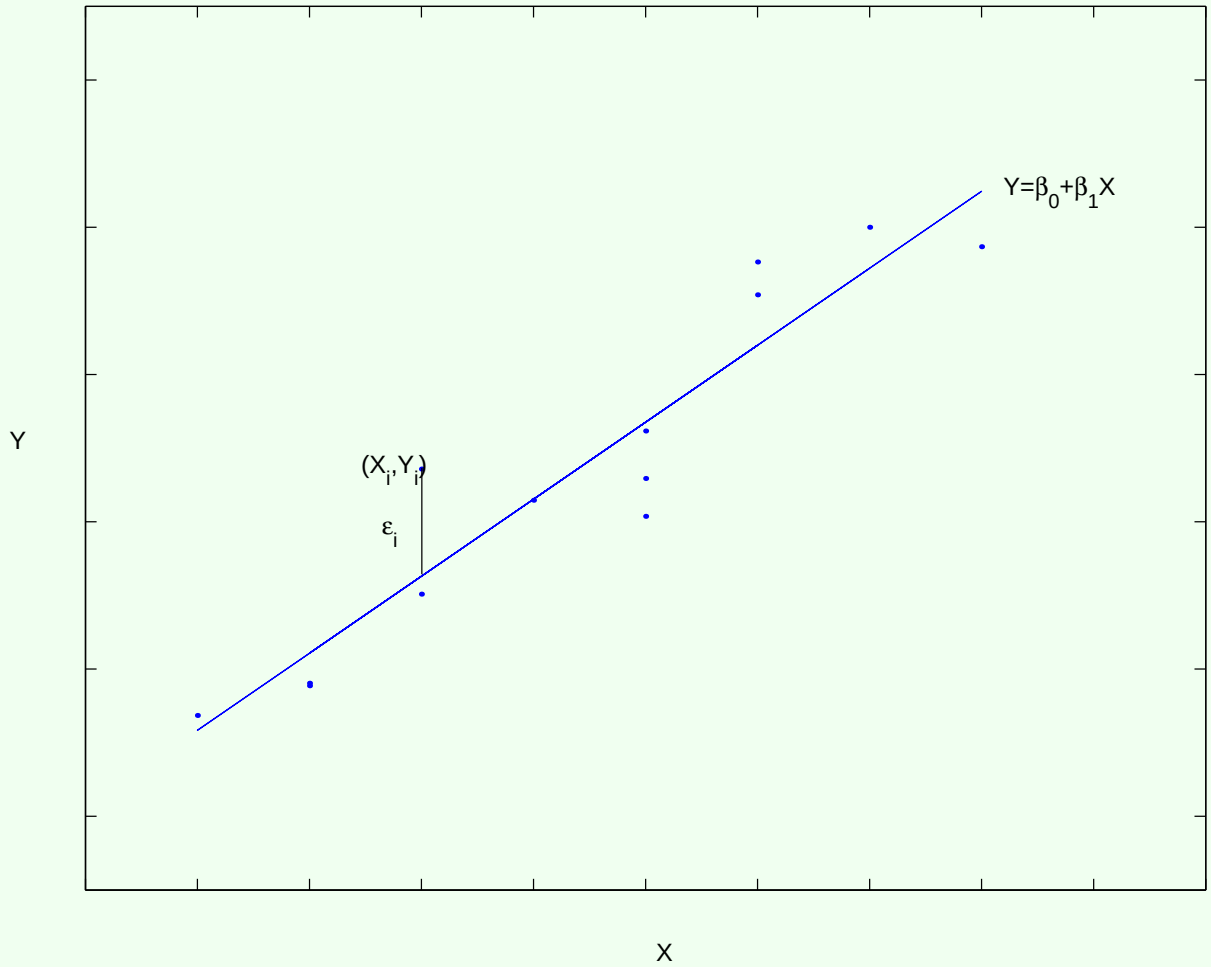
Απλή γραμμική παλινδρόμηση:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Η τυχαία μεταβλητή Y εξαρτάται από τη μεταβλητή X (η οποία συχνά έχει προκαθορισμένες τιμές), αλλά και από κάποιους μη μετρήσιμους παράγοντες. Αυτοί συνοψίζονται στον στοχαστικό όρο ϵ.

Σκοπός μας είναι να εκτιμήσουμε τις άγνωστες παραμέτρους β_0, β_1 χρησιμοποιώντας δείγμα (Y_i, X_i), $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$


Y=β0+β1X
(Xi,Yi)
εi
Y
X

Εκτίμηση

Μέθοδος Ελαχίστων Τετραγώνων

Υποθέσεις: $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$

Ελαχιστοποίηση του $\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$

$\rightarrow$ Αμερόληπτες εκτιμήτριες $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ με ελάχιστη διασπορά ανάμεσα στους αμερόληπτους εκτιμητές.

Μέθοδος Μέγιστης Πιθανοφάνειας

Υπόθεση κανονικότητας για τα σφάλματα, $\epsilon_i \sim N(0, \sigma^2)$

Πιθανοθεωρητικό μοντέλο: $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

Μεγιστοποίηση της πιθανότητας να παρατηρήσουμε τα δεδομένα κάτω από το μοντέλο που υποθέσαμε (συνάρτηση πιθανοφάνειας)

$\rightarrow$ Ίδιες εκτιμήτριες $\hat{\beta}_0$, $\hat{\beta}_1$ με Μ.Ε.Τ, αλλά διαφορετική (όχι αμερόληπτη) για το $\hat{\sigma}^2$.

Κάτω από την υπόθεση της κανονικότητας για τα ϵ_i:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right), \; \hat{\beta}_0 \sim N\left(\beta_0, \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2})\right]\right),$$

όπου $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$ ο Ε.Ε.Τ.

Είναι

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2_{\hat{\beta}_1}}} \sim t_{n-2}$$

$\rightarrow$ Διάστημα Εμπιστοσύνης για το β_1
$\rightarrow$ Έλεγχος Υποθέσεων για το β_1

Αντίστοιχα για το β_0.

Διάστημα Εμπιστοσύνης για το β_1

Από τη θεωρία κατανομών:

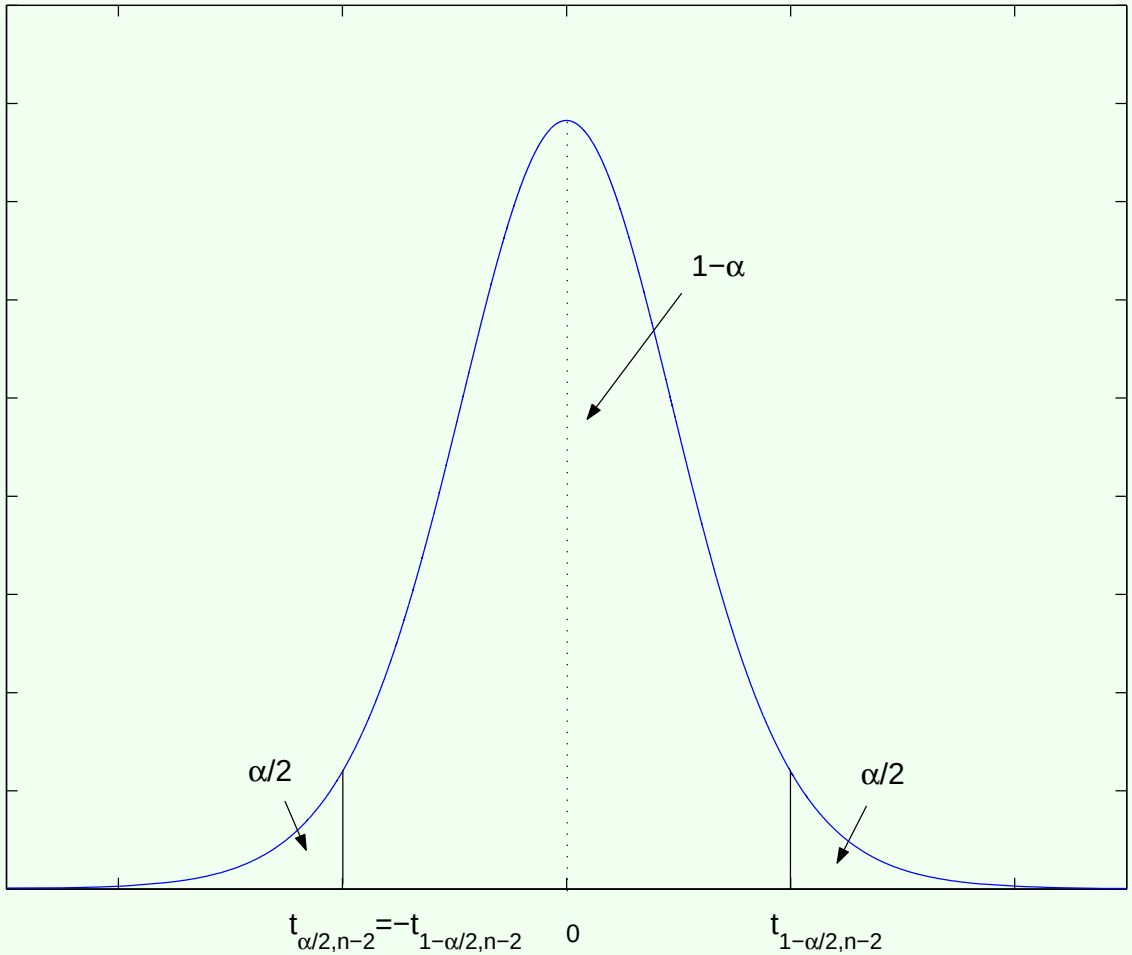
$$\Pr\left(t_{\alpha/2,n-2} < \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} < t_{1-\alpha/2,n-2}\right) = 1 - \alpha$$

Επομένως θέλουμε:

$$-t_{1-\alpha/2,n-2} < \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} < t_{1-\alpha/2,n-2}$$

$$\rightarrow \ \hat{\beta}_1 - \sigma_{\hat{\beta}_1} t_{1-\alpha/2,n-2} < \beta_1 < \hat{\beta}_1 + \sigma_{\hat{\beta}_1} t_{1-\alpha/2,n-2}$$

Εάν το 0 περιέχεται μέσα στο Δ.Ε., έχουμε ένδειξη ότι η παράμετρος β_1 δεν είναι στατιστικα σημαντική (σε επίπεδο στατιστικής σημαντικότητας α).


1−α
α/2
α/2
$t_{\alpha/2,n-2}=-t_{1-\alpha/2,n-2}$
0
$t_{1-\alpha/2,n-2}$

΄Ελεγχος Υποθέσεων για το β_1

$$H_0 : \ \beta_1 = 0$$
$$H_1 : \ \beta_1 \neq 0$$

Εάν η παράμετρος β_1 είναι στατιστικά ίση με το 0, τότε δεν υφίσταται η γραμμική σχέση $Y = \beta_0 + \beta_1 X$.

Κάτω απο την H_0: $\frac{\hat{\beta}_1}{\sqrt{\sigma^2_{\hat{\beta}_1}}} \sim t_{n-2}$, δηλαδή

$$-t_{1-\alpha/2,n-2} < \frac{\hat{\beta}}{\sigma_{\hat{\beta}_1}} < t_{1-\alpha/2,n-2}$$

Επομένως, αν $-\sigma_{\hat{\beta}_1} t_{1-\alpha/2,n-2} < \hat{\beta} < \sigma_{\hat{\beta}_1} t_{1-\alpha/2,n-2}$, τότε δεν απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας α. ΄Αρα, δεκτή η H_1 αν

$$| \ \hat{\beta}_1 \ | > \sigma_{\hat{\beta}_1} t_{1-\alpha/2,n-2} \rightarrow | \ \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}} \ | > t_{1-\alpha/2,n-2}$$

ANOVA

Με την Ανάλυση Διακύμανσης βρίσκουμε που οφείλεται περισσότερο η μεταβλητότητα της εξαρτημενής μεταβλητής Y: στην παλινδρόμηση ή στα τυχαία σφάματα;
(Ουσιαστικά σπάμε σε δύο κομμάτια το άθροισμα τετραγώνων και τους βαθμούς ελευθερίας της Y).

Η μεταβλητότητα της Y μετράται ως προς τις αποκλίσεις $Y_i - \bar{Y}$. Η συνολική μεταβητότητα των δεδομένων δίνεται από το συνολικό άθροισμα τετραγώνων: $\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Από την παλινδρόμηση υπολογίζουμε τα κατάλοιπα $e_i = Y_i - \hat{Y}$. Το άθροισμα τετραγώνων των καταλοίπων $\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ εκφράζει τη μεταβλητότητα των δεδομένων που δεν εξηγεί το γραμμικό μοντέλο.

Επίσης από την παλινδρόμηση υπολογίζουμε τις αποκλίσεις $\hat{Y}_i - \bar{Y}$. Είναι

$$Y_i - \bar{Y} = (Y_i - \hat{Y}) + (\hat{Y}_i - \bar{Y})$$

Αποδεικνύεται ότι

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2,$$

όπου το άθροισμα τετραγώνων της παλινδρόμησης SSR $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ εκφράζει τη μεταβλητότητα των δεδομένων που εξηγείται από το γραμμικό μοντέλο.

Πίνακας ANOVA

Πηγή Μεταβλητότητας	SS AT	d.f β.ε	MS=SS/d.f	F	p-value
Παλινδρόμηση	SSR	1	SSR/1	MSR/MSE	$*$
Σφάλμα	SSE	$n-2$	SSE/$(n-2)$		
Σύνολο	SST	$n-1$			

΄Ελεγχος ΄Υπαρξης Γραμμικής Σχέσης

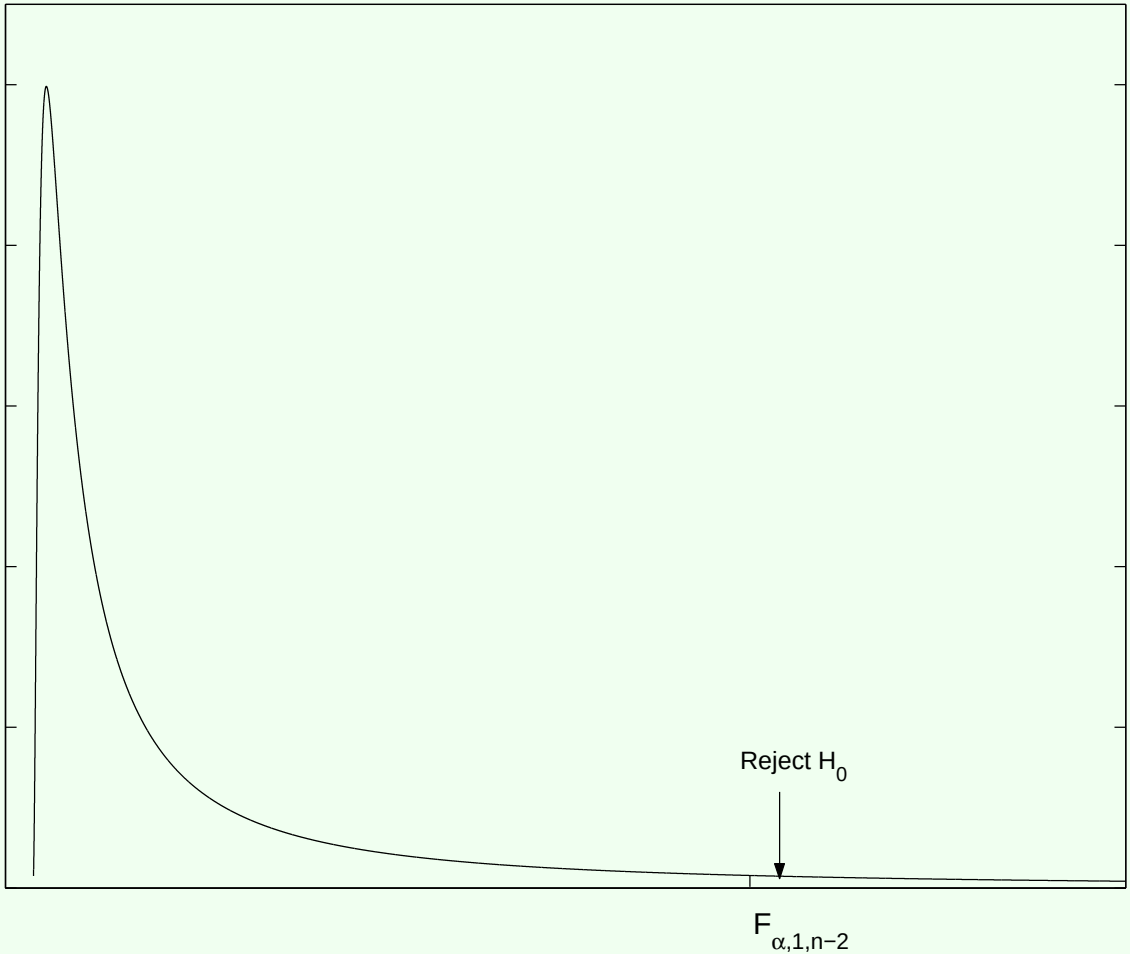
$$H_0 : \nexists \text{ γραμμική σχέση} \qquad Y_i \sim N(\beta_0, \sigma^2)$$
$$H_1 : \exists \text{ γραμμική σχέση} \quad Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

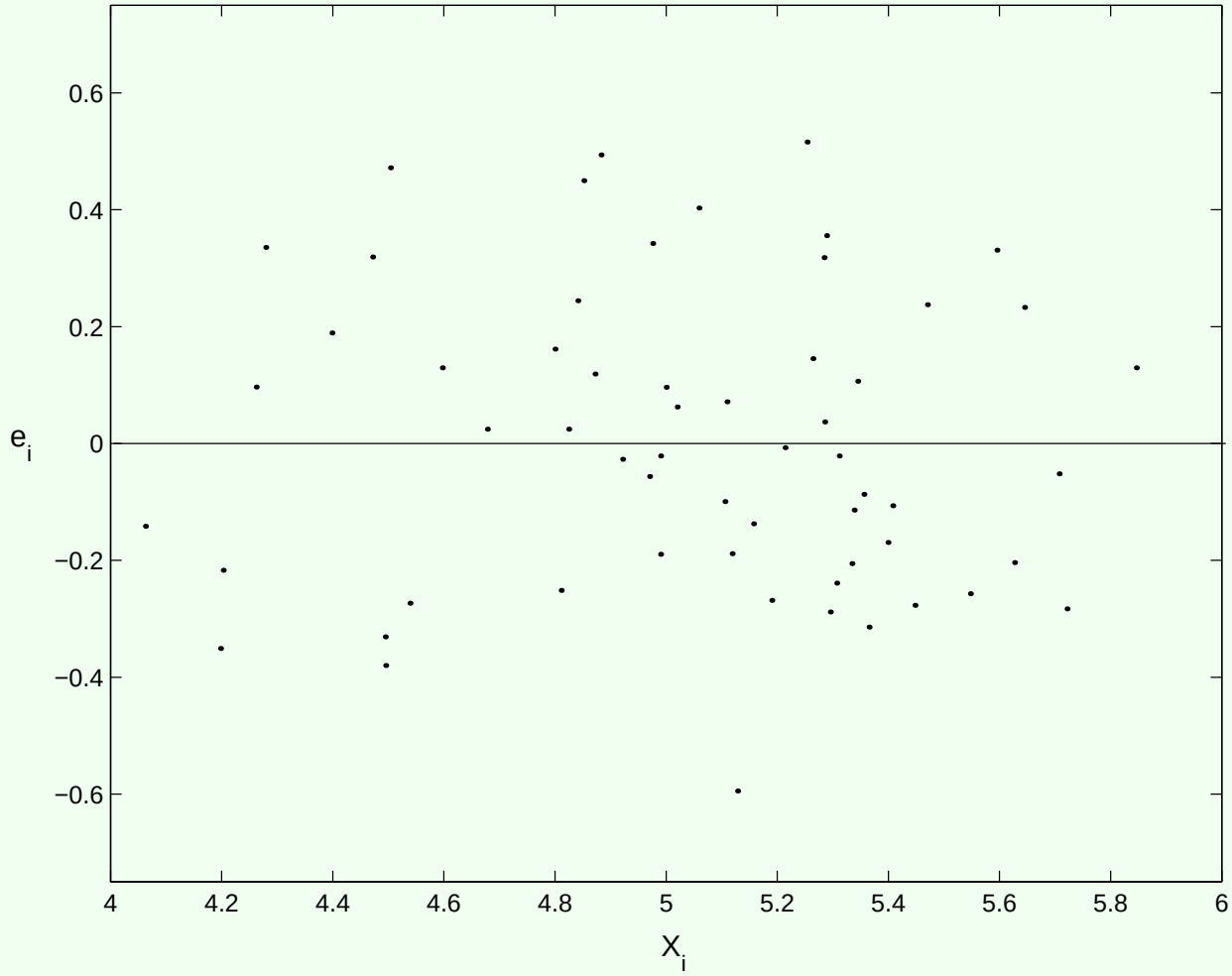
Κάτω απο την H_0 (Θ. Cohran): $\frac{\text{MSR}}{\text{MSE}} = F_0 \sim \mathcal{F}_{1,n-2}$.
Απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας α αν

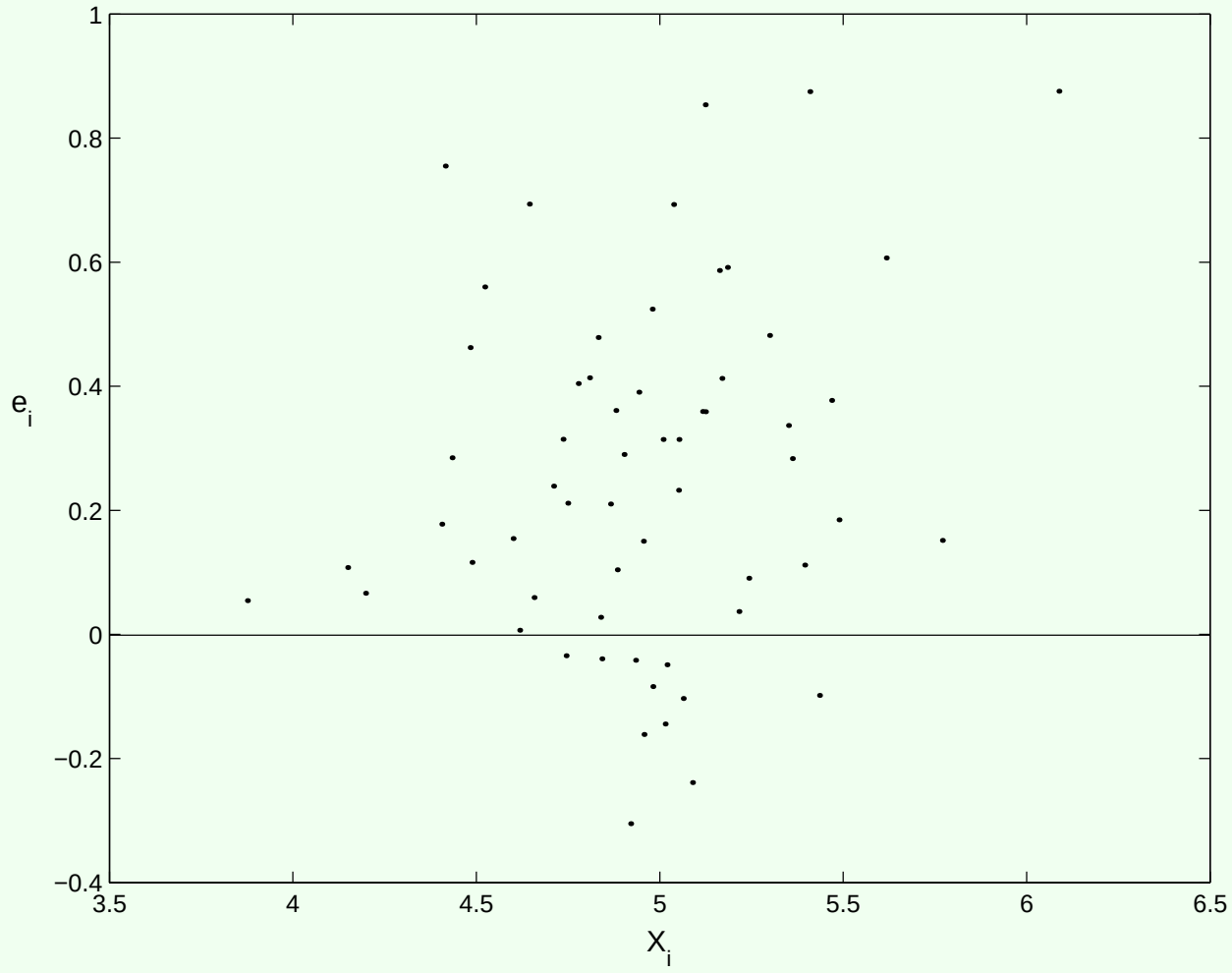
$$F_0 > F_{\alpha,1,n-2}$$

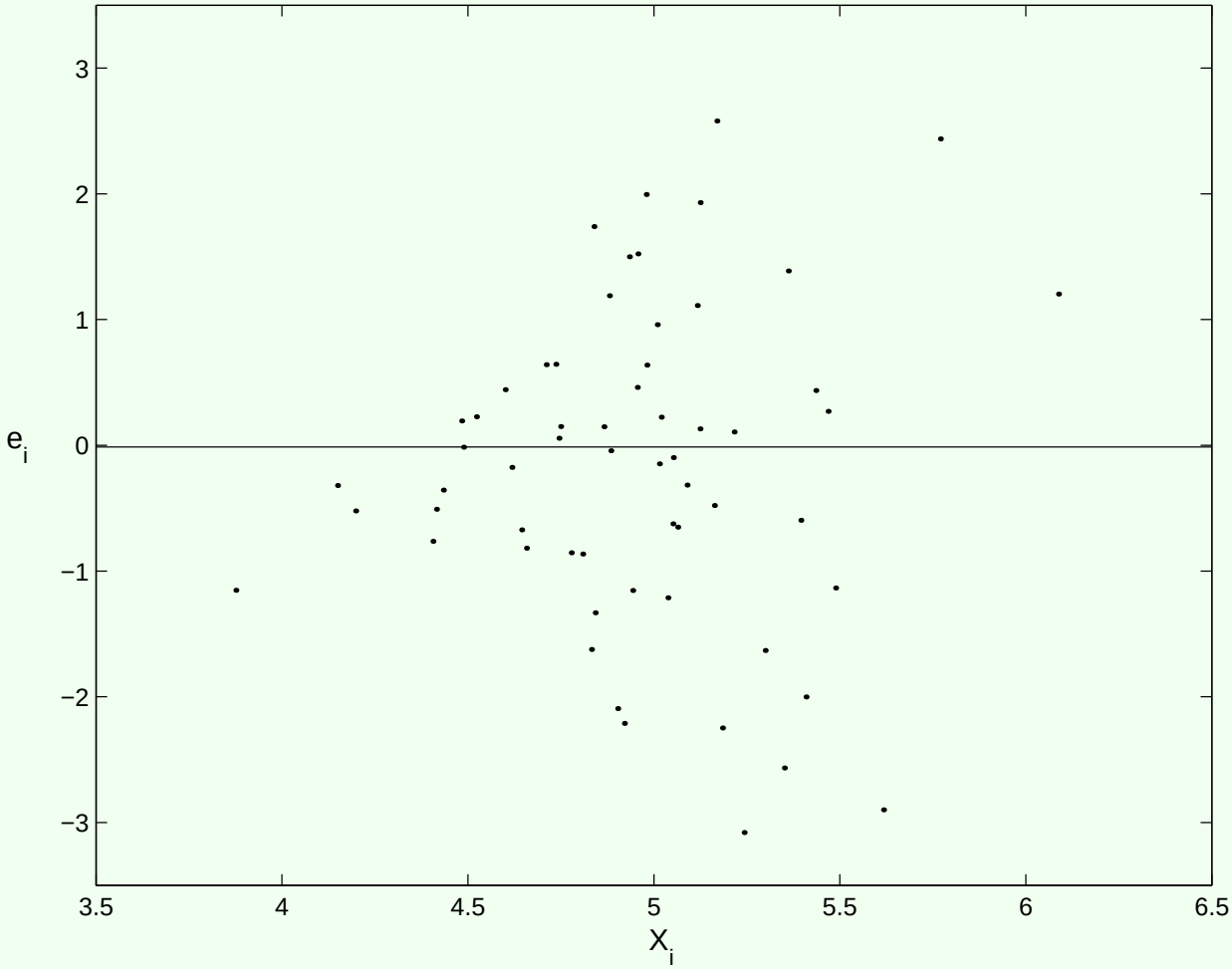
Στην απλή γραμμική παλινδρόμηση το F-test είναι ισοδύναμο με το t-test. Το F-test όμως είναι πιο γενικό (ελέγχει ύπαρξη γραμμικής σχέσης και στην πολλαπλή παλινδρόμηση.)

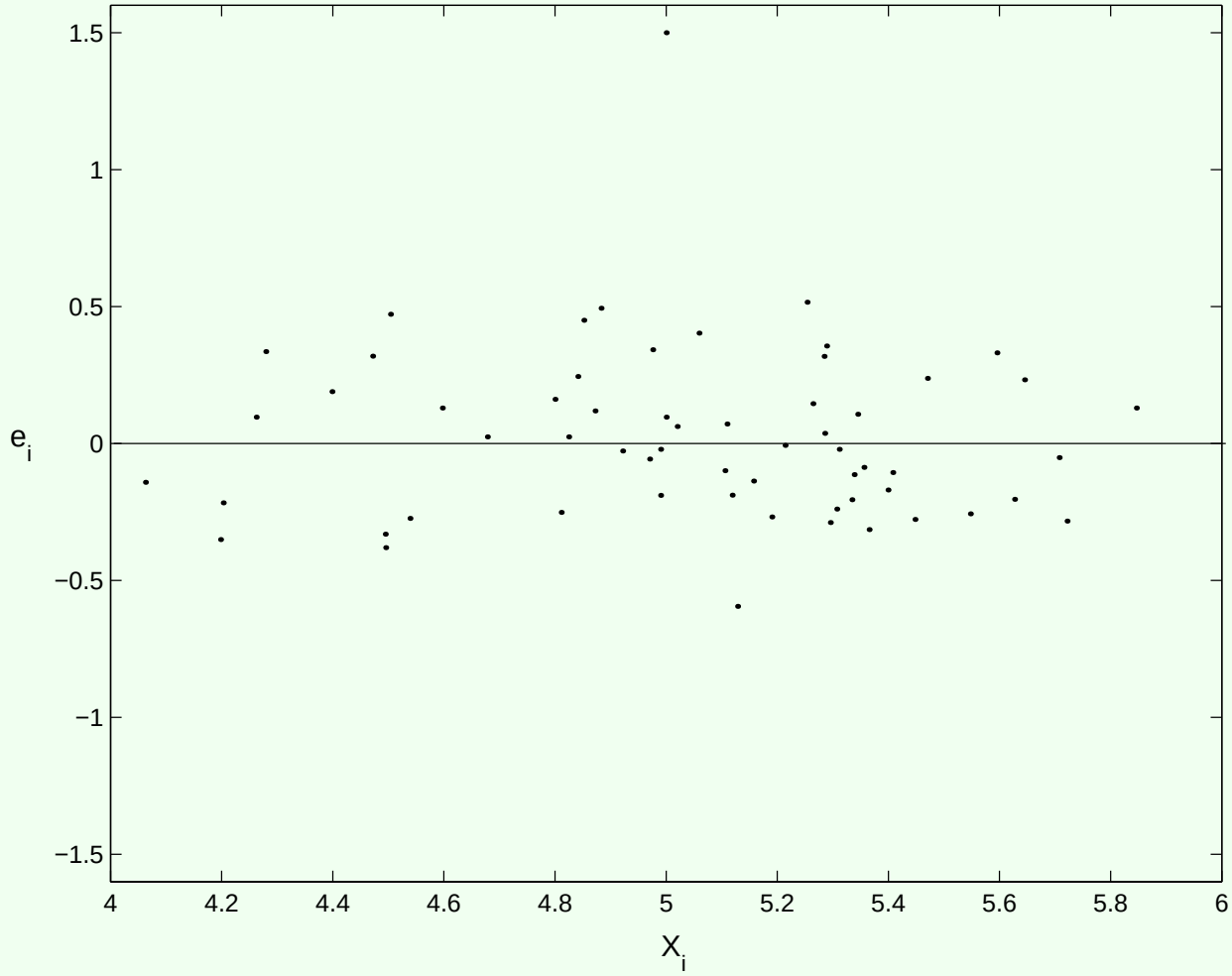
p-value: Η πιθανότητα μια τυχαία μεταβλητή που ακολουθεί $\mathcal{F}_{1,n-2}$ κατανομή να πάρει τιμή τόσο ακραία ή περισσότερο ακραία από αυτή που παρατηρήσαμε (δηλ. από F_0). Απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας α όταν p-value$<\alpha$.


Reject H_0
$F_{\alpha,1,n-2}$


0.6
0.4
0.2
0
−0.2
−0.4
−0.6
e_i
4
4.2
4.4
4.6
4.8
5
5.2
5.4
5.6
5.8
6
X_i


1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
e_i
3.5
4
4.5
5
5.5
6
6.5
X_i


3
2
1
0
−1
−2
−3
e_i
3.5
4
4.5
5
5.5
6
6.5
X_i


1.5
1
0.5
0
−0.5
−1
−1.5
e_i
4
4.2
4.4
4.6
4.8
5
5.2
5.4
5.6
5.8
6
X_i

Πολλαπλή Γραμμική Παλινδρόμηση

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i \rightarrow \mathbf{Y} = \mathbf{XB} + \epsilon.$$

Πλήρης αντιστοιχία με την απλή γραμμική παλινδρόμηση: Ελαχιστοποίηση του $\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}))^2$. Ε.Ε.Τ: $\hat{\mathbf{B}} = (X'X)^{-1}X'Y$.

Πίνακας ANOVA $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

Πηγή	SS	d.f	MS=SS/d.f	F	p-value
Παλινδρόμηση	SSR	$k-1$	SSR$/(k-1)$	MSR/MSE	$*$
Σφάλμα	SSE	$n-k$	SSE$/(n-k)$		
Σύνολο	SST	$n-1$			

Κάτω απο την H_0: $\dfrac{\text{MSR}}{\text{MSE}} = F_0 \sim \mathcal{F}_{k-1,n-k}$. Απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας α αν

$$F_0 > F_{\alpha,k-1,n-k}$$

Συντελεστής Προσδιορισμού

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Εκφράζει το ποσοστό της συνολικής μεταβλητότητας του Y που οφείλεται στην παλινδρόμηση.

Προσαρμοσμένος Συντελεστής Προσδιορισμού:

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-k}\right)\frac{\text{SSE}}{\text{SST}} < R^2.$$

Ο R^2_{adj} λαμβάνει υπόψη του και το πλήθος των παρατηρήσεων σε συνδυασμό με το πλήθος των ανεξάρτητων μεταβλητών.

Αν προσθέσουμε ανεξάρτητες μεταβλητές στο μοντέλο το R^2 πάντα αυξάνεται, ενώ το R^2_{adj} όχι απαραίτητα.

Παρατήρηση

Οι επεξηγηματικές μεταβλητές του γραμμικού μοντέλου X_j, $j = 1, \ldots, k$, μπορεί να παίρνουν προκαθορισμένες τιμές, μπορεί όμως να ειναι και τυχαίες μεταβλητές. Στη δεύτερη περίπτωση ισχύουν όλα όσα και στην πρώτη αν

- Τα X_{ji} είναι ανεξάρτητες τυχαίες μεταβλητές με κατανομές που δεν εξαρτόνται από τις παραμέτρους του μοντέλου.
- Η δεσμευμένη κατανομή του Y_i δοθέντων των $X_{1i}, \ldots, X_{ki}$ έχει αναμενόμενη τιμή $\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$ και διασπορά σ^2.

Πολυσυγγραμμικότητα

Το πρόβλημα της πολυσυγγραμμικότητας παρουσιάζεται όταν οι ερμηνευτικές μεταβλητές X_j δεν ειναι γραμμικώς ανεξάρτητες.

Για παραδειγμα, έστω το γραμμικό μοντέλο

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,$$

όπου για κάποια $\lambda_1 \neq \lambda_2 \neq 0$ ισχύει

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} = 0 \Rightarrow X_{1i} = cX_{2i},\ c = -\lambda_2/\lambda_1.$$

Σε ένα υπόδειγμα με πρόβλημα πολυσυγγραμμικότητας

$$Y_i = \beta_0 + (\beta_1 c + \beta_2) X_{2i} + \epsilon_i,$$

δεν μπορούμε να ξεχωρίσουμε τη συμβολή κάθε μεταβλητής X_{ji} στην ερμηνεία του Y_i. Πρακτικά, οι εκτιμήσεις των β_1, β_2 δεν είναι στατιστικά σημαντικές και έχουν μεγάλη διασπορά.

Εφαρμογές στην Οικονομετρία

Οι μαθηματικές σχέσεις της Οικονομικής θεωρίας αφορούν στις συσχετίσεις διαφόρων οικονομικών μεγεθών. Η ποσοτικοποίηση αυτών των συσχετίσεων γίνεται με τη βοήθεια στατιστικών εργαλείων και αποτελεί αντικείμενο της Οικονομετρίας.

Γραμμικα Μοντελα για Χρονολογικές Σειρές:
D_t: η ζήτηση ενός αγαθού κατά τη χρονική στιγμή t, P_t: η τιμή, C_t: η κατανάλωση, Y_t: το εισόδημα τη χρονική στιγμή t.

$D_t = \beta_0 + \beta_1 P_t + \epsilon_t$
$C_t = \alpha + \beta Y_t + \epsilon_t$, $\beta = \frac{dC_t}{dY_t}$: ροπή προς κατανάλωση
$C_t = \alpha + \beta Y_t + \gamma C_{t-1} + \epsilon_t$: δυναμική σχέση
$C_t = \alpha + \beta Y_t + \delta Y_{t-1} + \omega r_t + \epsilon_t$

Y_t: το προϊόν που παράχθηκε το έτος t, X_{1t}: η εργασία (αριθμός εργατοωρών), X_{2t}: το κεφάλαιο (χιλ. ευρώ).

$$\log Y_t = \beta_0 + \beta_1 \log X_{1t} + \beta_2 \log X_{2t} + \epsilon_t$$

Γραμμικα Μοντελα για Διαστρωματικά Στοιχεία:

$C_i = \alpha + \beta Y_i + \epsilon_i$, εδώ μελετάμε τα στοιχεία μιας συγκεκριμένης χρονιάς για διάφορα στρώματα, π.χ. νοικοκυριά.

Παράδειγμα

Y_i: η ετήσια κατανάλωση ενός προιόντος (σε κιλά) από το νοικοκυριό i, X_{1i}: το ετήσιο εισόδημα (σε χιλ. ευρώ) του νοικοκυριού i, X_{2i}: ο αριθμός μελών στην οικογένεια i.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

΄Εστω δείγμα $n = 8$ νοικοκυριών:

Y_i	18	9	23	16	7	26	22	15
X_{1i}	14	7	17	13	6	19	16	12
X_{2i}	3	1	5	2	1	6	4	2

Εκτιμηθείσα γραμμική παλινδρόμηση:

$$\hat{Y}_i = -0.412 + 1.173 X_{1i} + 0.721 X_{2i}$$

Ερμηνεία των συντελεστών παλινδρόμησης

$\hat{\beta}_1 = 1.173$: Για αύξηση της μεταβλητής X_1 κατά μια μονάδα, με τη X_2 σταθερή, η εξαρτημένη μεταβλητή Y αυξάνεται κατά 1.173 μονάδες. Δηλαδή, για αύξηση του εισοδήματος μιας οικογένειας κατά 1000 ευρώ, με τον αριθμό των μελών της σταθερό, η κατανάλωσή της αυξάνει κατά 1.173 κιλά.
$\hat{\beta}_2 = 0.721$: Για κάθε μέλος που προστίθεται στην οικογένεια, με σταθερό εισόδημα, η κατανάλωση αυξάνεται κατά 0.721 κιλά.

Οι συντελεστές της παλινδρόμησης δεν εκφράζονται στην ίδια μονάδα (β_1: κιλά/χιλ. ευρώ, β_2: κιλά/αριθ. ατόμων), άρα δεν υπάρχει συγκρισιμότητα.
$\rightarrow$ Δείκτες (απαλλαγμένοι από μονάδες).

Ελλαστικότητες

$$\eta_{Y|X_1} = \hat{\beta}_1 \frac{\bar{X}_1}{\bar{Y}} = 1.173 \frac{13}{17} = 0.897, \quad \eta_{Y|X_2} = \hat{\beta}_2 \frac{\bar{X}_2}{\bar{Y}} = 0.721 \frac{3}{17} = 0.127$$

Για 10% αύξηση του εισοδήματος έχουμε 8.97% αύξηση της κατανάλωσης.

Output: Απλή Γραμμική Παλινδρόμηση ως προς X_1.

```
Coefficients:
               Value Std. Error  t value  Pr(>|t|)
(Intercept)  -1.7973    0.7363   -2.4410    0.0504
         X1   1.4459    0.0538   26.8904    0.0000

R-Squared: 0.9918 F-statistic: 723.1 on 1 and 6 df

The p-value is 1.747e-007

Analysis of Variance Table
Response: Y

           Df Sum of Sq  Mean Sq  F Value   Pr(F)
       X1   1  309.432  309.432  723.094 1.75e-007
Residuals   6    2.567    0.427
```

Output: Απλή Γραμμική Παλινδρόμηση ως προς X_2.

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 6.6250 1.4422     4.5938  0.0037
         X2 3.4583 0.4163     8.3069  0.0002

R-Squared: 0.92 F-statistic: 69.01 on 1 and 6 df

The p-value is 0.000165

Analysis of Variance Table
Response: Y

           Df Sum of Sq  Mean Sq  F Value    Pr(F)
       X2   1  287.041   287.041   69.005 0.000165
Residuals   6   24.958     4.159
```

Output: Πολλαπλή Γραμμική Παλινδρόμηση.

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  -0.4135    0.7549   -0.5477   0.6075
         X1   1.1731    0.1130   10.3853   0.0001
         X2   0.7212    0.2805    2.5710   0.0500

R-Squared: 0.9965 F-statistic: 702.9 on 2 and 5 df

The p-value is 7.478e-007

Analysis of Variance Table
Response: Y

           Df Sum of Sq  Mean Sq  F Value  Pr(F)
       X1   1  309.432  309.432  1399.17  0.000
       X2   1    1.461    1.461     6.61  0.050
Residuals   5    1.105    0.221
```

Απλή Γραμμική Παλινδρόμηση

(Y, X_1): $\hat{Y}_i = -1.787 + 1.445^* X_{1i}$, $R^2 = 0.992$, $R^2_{adj} = 0.991$.

(Y, X_2): $\hat{Y}_i = \ \ 6.625 + 3.458^* X_{2i}$, $R^2 = 0.920$, $R^2_{adj} = 0.907$.

Πολλαπλή Γραμμική Παλινδρόμηση

(Y, X_1, X_2): $\hat{Y}_i = -0.412 + 1.173^* X_{1i} + 0.721 X_{2i}$, $R^2 = 0.996$, $R^2_{adj} = 0.994$.

Όλα τα F-test (για τις 3 παλινδρομήσεις) δέχονται την H_0 (ύπαρξη γραμμικής σχέσης). Ωστόσο, στην πολλαπλή γραμμική παλινδρόμηση ο δευτερος συντελεστής είναι στατιστιστικά ασήμαντος.

Σύμφωνα με την πρώτη απλή γραμμική παλινδρόμηση η X_1 ερμηνεύει την μεταβλητότητα της Y κατά 99.2%, αφήνοντας ανερμήνευτο το 0.8%. Η πολλαπλή παλινδρόμηση μειώνει το ανερμήνευτο ποσοστο στο 0.4%.