

**Ε. ΜΠΟΡΑ - ΣΕΝΤΑ**  
Λέκτορας Α.Π.Θ.

**Χ. ΜΩΥΣΙΑΔΗΣ**  
Επικ. Καθηγητής Α.Π.Θ.

# **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

**Πολλαπλή Παλινδρόμηση  
Ανάλυση Διασποράς  
Χρονοσειρές**

 **ΕΚΔΟΣΕΙΣ  
ΖΗΤΗ**  
ΘΕΣΣΑΛΟΝΙΚΗ

## ΠΡΟΛΟΓΟΣ

Η αναζήτηση μοντέλων για την περιγραφή και πρόβλεψη διαφόρων φαινομένων, με τη βοήθεια πραγματικών δεδομένων, ήταν είναι και θα είναι, αντικείμενο μελέτης και έρευνας σε όλες τις επιστήμες.

Η ανάλυση μεγάλου όγκου δεδομένων, απαιτεί πολλούς και πολύπλοκους, υπολογισμούς, που αντιμετωπίζονται μόνο με τη βοήθεια των Η/Υ. Η ραγδαία εξάπλωση των προσωπικών υπολογισμών καθώς και η παράλληλη ανάπτυξη έτοιμων προγραμμάτων, έκανε την ανάλυση δεδομένων, προσιτή σε μεγάλο πλήθος ενδιαφερομένων από όλες τις ειδικότητες.

Η κατανόηση της λειτουργίας των προγραμμάτων, καθώς και η γνώση της θεωρίας των μεθόδων που χρησιμοποιούνται, έγινε έτσι απαραίτητη ώστε να ελαχιστοποιηθεί η πιθανότητα εξαγωγής λαθεμένων συμπερασμάτων.

Το εγχειρίδιο αυτό είναι μια προσπάθεια προς την κατεύθυνση αυτή. Αποτελείται από δύο μέρη με τέσσερα κεφάλαια το καθένα. Στο πρώτο μέρος ασχολούμαστε με την πολλαπλή γραμμική παλινδρόμηση και την ανάλυση διασποράς, ενώ στο δεύτερο με χρονικές σειρές. Σε κάθε κεφάλαιο, αφού αναπτύξουμε τη βασική θεωρία που αντιστοιχεί, σ' αυτό την εφαρμόζουμε σε πρακτικά προβλήματα. Σ' αυτές τις εφαρμογές, δίνουμε ιδιαίτερη έμφαση στην ανάγνωση και ερμηνεία των πληροφοριών που λαμβάνονται από τον υπολογιστή.

Για την επεξεργασία των προγραμμάτων χρησιμοποιήθηκε το στατιστικό πακέτο SPSS μέσω του κεντρικού υπολογιστή IBM του ΑΠΘ. Για τις χρονικές σειρές χρησιμοποιήθηκε το πρόγραμμα TRENDS του SPSS, προσαρμοσμένο για προσωπικούς υπολογιστές. Βεβαίως υπάρχουν και πολλά άλλα στατιστικά πακέτα που επεξεργάζονται τέτοια προβλήματα, που οδηγούν όμως σε ακριβώς ανάλογα συμπεράσματα.

Στη συνέχεια δίνονται δύο παραρτήματα. Στο πρώτο περιέχεται μια σύντομη περίληψη της Άλγεβρας πινάκων και στο δεύτερο, διάφοροι στατιστικοί πίνακες.

Η δακτυλογράφηση του κειμένου και η επιμέλεια της έκδοσης, έγινε από τις εκδόσεις Ζήτη, τις οποίες και ευχαριστούμε.

Μάιος 1990

Ε. Μπόρα-Σέντα  
Χ. Μωϋσιάδης

# ΠΕΡΙΕΧΟΜΕΝΑ

## ΜΕΡΟΣ Ι' ΠΑΛΙΝΔΡΟΜΗΣΗ - ANOVA

### 1 ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

	σελ.
1.1. Εισαγωγή .....	7
1.2. Εκτίμηση των παραμέτρων του γραμμικού μοντέλου .....	9
1.3. Πρόβλεψη και παρεμβολή .....	14
1.4 Το σφάλμα μετά την προσαρμογή του μοντέλου .....	15
1.5. Ανάλυση της διασποράς μετά την παλινδρόμηση .....	17
1.6. Συντελεστής προσδιορισμού .....	25
1.7. Έλεγχοι υποθέσεων .....	26
1.8. Σφάλμα προσαρμογής - Επαναλαμβανόμενες μετρήσεις .....	33
1.9. Το πρόγραμμα REGRESSION .....	38
1.10. Ασκήσεις .....	46

### 2 ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ

2.1. Πολυσυγγραμμικότητα .....	52
2.2. Απαλειφή μεταβλητών - Περιορισμένο μοντέλο .....	54
2.3. Παλινδρόμηση υπό περιορισμούς - Συμπτηγμένο μοντέλο .....	61
2.4. Μερικός συντελεστής προσδιορισμού - συσχέτισης .....	64
2.5. Όριο ανοχής .....	72
2.6. Επίλογή του καλύτερου μοντέλου απ' όλα τα δυνατά .....	74
2.7. Σταδιακή επιλογή μεταβλητών .....	81
2.8. Ασκήσεις .....	86

### 3 ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΙ ΜΕΤΑΒΛΗΤΩΝ ΣΕ ΠΕΡΙΠΤΩΣΕΙΣ ΑΠΟ- ΚΛΙΣΗΣ ΑΠΟ ΤΙΣ ΥΠΟΘΕΣΕΙΣ

3.1. Εισαγωγή .....	88
3.2. Ποιοτικές μεταβλητές ως προβλέπουσες .....	88
3.3. Επίδραση του χρόνου στα δεδομένα .....	92
3.4. Ετεροσκεδαστικότητα .....	95
3.5. Μέθοδος σταθμισμένων ελαχίστων τετραγώνων .....	99
3.6. Το στατιστικό Durbin - Watson .....	102
3.7. Μεροληψία .....	103
3.8. Ασκήσεις .....	109

## 4 ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

4.1. Εισαγωγή	108
4.2. Ανάλυση διασποράς μ' ένα παράγοντα	108
4.3. Ισοδυναμία ανάλυσης διασποράς και παλινδρόμησης	118
4.4. Ανάλυση διασποράς με δύο παράγοντες	124
4.5. Παραγοντικά πειράματα	137
4.6. BIB-σχεδιασμοί	139
4.7. Λατινικά τετράγωνα	141
4.8. Ελληνολατινικά τετράγωνα	144
4.9. Ασκήσεις	146

## ΜΕΡΟΣ ΙΙ'

### ΧΡΟΝΙΚΕΣ ΣΕΙΡΕΣ

## 1 ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΧΡΟΝΙΚΕΣ ΣΕΙΡΕΣ

1.1. Εισαγωγή	153
1.2. Στατικότητα	156
1.3. Συνάρτηση αυτοσυσχέτισης	158
1.4. Ασκήσεις	165

## 2 ΓΡΑΜΜΙΚΑ ΣΤΑΤΙΚΑ ΜΟΝΤΕΛΑ

2.1. Το γενικό γραμμικό μοντέλο (GLM)	167
2.2. Αυτοπαλινδρομούμενα μοντέλα AR (p)	170
2.3. Συνάρτηση μερικής αυτοσυσχέτισης	176
2.4. Κινούμενου μέσου μοντέλα MA(q)	179
2.5. Το μεικτό μοντέλο ARMA (p, q)	184
2.6. Εύρεση τάξης ενός γραμμικού στατικού μοντέλου. Κριτήριο Akaike	187
2.7. Έλεγχος του μοντέλου	188
2.8. Ασκήσεις	191

## 3 ΟΛΟΚΛΗΡΩΜΕΝΑ - ΕΠΟΧΙΚΑ ΜΟΝΤΕΛΑ

3.1. Εισαγωγή	195
3.2. Τα ολοκληρωμένα μεικτά μοντέλα ARIMA (p, d, q)	195
3.3. Εκτίμηση των παραμέτρων ενός ARIMA μοντέλου	196
3.4. Το πρόγραμμα TRENDS για ARIMA (p, d, q) μοντέλα	199
3.5. Χρονικές σειρές με εποχικότητα (seasonal time series)	207
3.6. Ασκήσεις	

## 4 ΠΡΟΒΛΕΨΗ

4.1. Η μέθοδος πρόβλεψης των Box και Jenkins	224
4.2. Διαστήματα εμπιστοσύνης για τις προβλέψεις	227
4.3. Ασκήσεις	230

## ΠΑΡΑΡΤΗΜΑΤΑ

## Α ΑΛΓΕΒΡΑ ΠΙΝΑΚΩΝ

A.1. Πράξεις πινάκων .....	235
A.2. Ορίζουσα και αντιστροφή πινάκων .....	238
A.3. Τετραγωνικές μορφές .....	241
A.4. Βαθμός πίνακα - Γραμμικά συστήματα .....	241
A.5. Γενικευμένος αντίστροφος .....	243
A.6. Έχνος πίνακα .....	245
A.7. Ειδικά γινόμενα πινάκων .....	245
A.8. Ιδιοτιμές - Ιδιοδιανύσματα .....	246
A.9. Μετασχηματισμοί πινάκων .....	249
A.10. Παραγωγή ως προς διάνυσμα ή πίνακα .....	250

## Β ΣΤΑΤΙΣΤΙΚΟΙ ΠΙΝΑΚΕΣ

B.1. Αθροιστική σ.κ. τυπικής κανονικής κατανομής .....	255
B.2. Τυχαίοι κανονικοί αριθμοί $\mu=0$ $\sigma=1$ .....	256
B.3. Κρίσιμες τιμές της κατανομής $t_n$ , σε σ.σ. $\alpha$ .....	258
B.4. Κρίσιμες τιμές της κατανομής $\chi_n^2$ σε σ.σ. $\alpha$ .....	259
B.5. Κρίσιμες τιμές της κατανομής $F_{m, n}$ σε σ.σ. $\alpha$ .....	260
B.6. Κρίσιμες τιμές για το στατιστικό Durbin - Watson .....	262
B.7. Γραφικές παραστάσεις ACF-PACF, ειδικών μοντέλων .....	264
Ευρετήριο όρων .....	267

## ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

### 1.1. Εισαγωγή

Ένα σημαντικό ερώτημα σε πάρα πολλά προβλήματα, σχεδόν κάθε είδους, όπως παραγωγής (βιομηχανική, αγροτική, κλπ.), εκπαίδευσης (μαθητών, στελεχών, στρατιωτών, κλπ.), πρόβλεψης (εκλογές, καιρός, κλπ.), χωροθέτησης, βελτιστοποίησης και άλλων, είναι αν μπορούμε να εκτιμήσουμε ή να προβλέψουμε την τιμή μιας ή περισσότερων «μεταβλητών» κάτω από ορισμένες συνθήκες. Οι δοσμένες συνθήκες περιγράφονται και αυτές από μεταβλητές, οι τιμές των οποίων είναι δυνατό να ελεγχθούν από τον ερευνητή. Έτσι για παράδειγμα η μεταβλητή  $Y$  που ζητούμε να εκτιμηθεί ή να προβλεφθεί, μπορεί να παριστάνει «ζήτηση κάποιου προϊόντος στην αγορά», «παραγωγή κάποιου γεωργικού προϊόντος», «απόδοση μαθητού», «αύξηση ποσοστού σε εκλογές», κλπ. Ενώ οι μεταβλητές  $X_i$  που περιγράφουν τις συνθήκες και που μπορούν να ελεγχθούν, μπορεί να παριστάνουν «τιμή πώλησης προϊόντος», «συσκευασία», «κόστος διαφήμισης», «ταχύτητα διανομής», «ποικιλία», «λίπανση», «θερμοκρασία», «είδος διδασκαλίας», «φύλο», και πολλά άλλα.

Για την εύρεση του μοντέλου εκτίμησης ή πρόβλεψης χρησιμοποιούνται δεδομένα που έχουν προκύψει από μία σειρά  $n$  παρατηρήσεων και που συχνά δίνονται με τη μορφή του παρακάτω πίνακα:

$$\begin{bmatrix} X_{11} & X_{21} & X_{31} & \dots & X_{k1} & Y_1 \\ X_{12} & X_{22} & X_{32} & \dots & X_{k2} & Y_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_{1n} & X_{2n} & X_{3n} & \dots & X_{kn} & Y_n \end{bmatrix} \quad (1.1)$$

Οι γραμμές του πίνακα παριστάνουν τις παρατηρήσεις, ενώ οι στήλες δίνουν τις τιμές των αντίστοιχων μεταβλητών για κάθε παρατήρηση.

Η μορφή του μοντέλου πρόβλεψης μπορεί να είναι οποιαδήποτε, εδώ όμως θα ασχοληθούμε μόνο με το γραμμικό μοντέλο. Το γενικό γραμμικό μοντέλο είναι το:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon, \quad (1.2)$$

όπου:  $Y$  είναι η **εξαρτημένη** (dependent) μεταβλητή ή **απόκριση** (response).

$X_1, X_2, \dots, X_k$  είναι οι  $k$  «**ανεξάρτητες**» (independent) ή «**προβλέπουσες**» (predictor) μεταβλητές.

$\beta_0, \beta_1, \dots, \beta_k$  είναι  $(k+1)$  άγνωστες παράμετροι (**συντελεστές παλινδρόμησης**) που ζητείται να εκτιμηθούν.

$\varepsilon$  είναι το σφάλμα.

Για τις μεταβλητές  $X_1, X_2, \dots, X_k$  η ονομασία «ανεξάρτητες» δεν σημαίνει ότι είναι πράγματι ανεξάρτητες. Μπορεί για παράδειγμα να ισχύει  $X_2 = X_1^2$  ή  $X_3 = X_1 + X_2$ . Ο λόγος στον οποίο οφείλεται αυτή η ονομασία, είναι, ότι το ζητούμενο συνήθως είναι, πως οι τιμές αυτών των μεταβλητών επηρεάζουν τις τιμές της εξαρτημένης μεταβλητής και μπορούν να ελέγχονται από τον ερευνητή. Από το τελευταίο αυτό προκύπτει και η ονομασία «προβλέπουσες» μεταβλητές. Στην πράξη πολλές φορές οι μεταβλητές εναλλάσσουν ρόλους. Μία μεταβλητή, π.χ. που στο πρώτο μέρος μιας μελέτης είναι εξαρτημένη, μπορεί στο δεύτερο μέρος της μελέτης να είναι ανεξάρτητη. Εκείνο όμως που απαιτείται από τις μεταβλητές αυτές, είναι να είναι ποσοτικές, να περιγράφουν δηλ. μετρήσιμα μεγέθη. Με ποιοτικές προβλέπουσες θ' ασχοληθούμε στα κεφάλαια 3 και 4.

Το σφάλμα  $\varepsilon$ , περιέχει κάθε απόκλιση της πραγματικής κατάστασης από το μοντέλο. Έτσι εκτός από τα πιθανά σφάλματα μετρήσεων, περιέχει επίσης και σφάλματα προσαρμογής, που οφείλονται είτε σε παράλειψη μεταβλητών είτε σε χρήση μεταβλητών που δε σχετίζονται με την  $Y$ .

Η δυνατότητα των προβλεπουσών μεταβλητών να συσχετίζονται μεταξύ τους διευρύνει τις περιπτώσεις εφαρμογής του μοντέλου (1.2). Πράγματι, με κατάλληλους μετασχηματισμούς «μη-γραμμικά» ή και «εκθετικά» μοντέλα, ανάγονται στο γενικό γραμμικό μοντέλο όπως φαίνεται παρακάτω. Πράγματι:

Θέτοντας  $X_j = x^j, j = 1, \dots, k$  το πολυωνυμικό μοντέλο

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \varepsilon$$

ανάγεται αμέσως στο γραμμικό μοντέλο (1.2).

Όμοια θέτοντας  $X_1 = x, X_2 = z, X_3 = x^2, X_4 = xz$  το μοντέλο

$$Y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x^2 + \beta_4 xz + \varepsilon$$

ανάγεται πάλι στο (1.2).





τη μέθοδο. Οι εκτιμήσεις των  $\beta_i$  θα συμβολίζονται με  $\hat{\beta}_i$ , και μ' αυτές το μοντέλο παίρνει τη μορφή:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k \quad (1.5)$$

Η τιμή  $\hat{Y}$  είναι η εκτίμηση της πραγματικής τιμής της  $Y$ , όταν δίνονται οι τιμές  $X_1, X_2, \dots, X_k$  και διαφέρει απ' αυτήν κατά ένα σφάλμα. Αν οι  $X_j$  πάρουν τις τιμές  $x_{ji}$  που δίνονται στην  $i$  γραμμή του πίνακα (1.1), τότε η  $\hat{Y}$  συμβολίζεται με  $\hat{Y}_i$  και ισχύει

$$\hat{Y}_i - Y_i = \varepsilon_i, \quad i=1, 2, \dots, n \quad (1.6)$$

όπου  $\sum_{i=1}^n \varepsilon_i^2$  είναι ελάχιστο. Η διαφορά  $Y_i - \hat{Y}_i$  λέγεται **υπόλοιπο** (residual).

Για τον υπολογισμό των  $\hat{\beta}_i$  εργαζόμαστε ως εξής: Θέτουμε

$$\tilde{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \quad \tilde{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \tilde{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

οπότε το σύστημα (1.4) γίνεται

$$\tilde{Y} = X \tilde{\beta} + \tilde{\varepsilon} \quad (1.7)$$

Για το διάνυσμα των σφαλμάτων  $\tilde{\varepsilon}$  υποθέτουμε ότι

$$E(\tilde{\varepsilon}) = 0 \quad \text{και} \quad V(\tilde{\varepsilon}) = \sigma^2 I_n, \quad (1.8)$$

όπου  $E(\tilde{\varepsilon}) = (E\varepsilon_1, E\varepsilon_2, \dots, E\varepsilon_n)'$ , το διάνυσμα των μέσων τιμών, και

$V(\tilde{\varepsilon}) = E(\varepsilon_i - E\varepsilon_i)(\varepsilon_i - E\varepsilon_i)'$ , ο πίνακας διακυμάνσεων-συνδιακυμάνσεων των  $\varepsilon_i$ .

Ο πίνακας  $X$  στην (1.7) περιέχει όλα τα σημεία όπου έγιναν οι παρατηρήσεις, γι' αυτό λέγεται **πίνακας σχεδιασμού** (design matrix).

Ας συμβολίσουμε με  $S_\varepsilon(\tilde{\beta})$  το άθροισμα των τετραγώνων των σφαλμάτων της (1.7). Τότε θα έχουμε

$$S_\varepsilon(\tilde{\beta}) = \sum_{i=1}^n \varepsilon_i^2 = \tilde{\varepsilon}' \tilde{\varepsilon}.$$

Αλλά η (1.7) δίνει

$$\underline{\varepsilon}'\underline{\varepsilon} = (\underline{Y} - \underline{X}\underline{\beta})'(\underline{Y} - \underline{X}\underline{\beta}),$$

οπότε

$$S_e(\underline{\beta}) = \underline{Y}'\underline{Y} - 2\underline{Y}'\underline{X}\underline{\beta} + \underline{\beta}'\underline{X}'\underline{X}\underline{\beta}. \quad (1.9)$$

Μια αναγκαία συνθήκη για την ελαχιστοποίηση του  $S_e(\underline{\beta})$  ως προς το  $\underline{\beta}$  είναι οι μερικές παράγωγοι ως προς τα  $\beta_i$  να είναι ίσες με 0. Επειδή (δες παράρτημα Α.10)

$$\frac{\partial S_e(\underline{\beta})}{\partial \underline{\beta}} = 2(\underline{X}'\underline{X}\underline{\beta} - \underline{X}'\underline{Y})$$

η συνθήκη γίνεται

$$\underline{X}'\underline{X}\underline{\beta} = \underline{X}'\underline{Y} \quad (1.10)$$

Η τελευταία σχέση παριστάνει ένα σύστημα  $k+1$  γραμμικών εξισώσεων ως προς τις παραμέτρους  $\beta_0, \beta_1, \dots, \beta_k$ , που λέγονται **κανονικές εξισώσεις** (normal equations) του μοντέλου (1.7) ή (1.4). Αν ο πίνακας των συντελεστών  $\underline{X}'\underline{X}$  είναι μη-ιδιάζων, τότε ως γνωστόν το σύστημα (1.10) δέχεται μοναδική λύση την:

$$\underline{\beta} = (\underline{X}'\underline{X})^{-1} \underline{X}'\underline{Y} \quad (1.11)$$

ενώ αν  $|\underline{X}'\underline{X}| = 0$ , τότε το  $\underline{\beta}$  μπορεί να εκτιμηθεί μόνο αν το σύστημα (1.10) είναι συνεπές, δηλ. αν  $r(\underline{X}'\underline{X}) = r(\underline{X}'\underline{X}, \underline{X}'\underline{Y})$ , όπου  $r(A)$  ο βαθμός του πίνακα  $A$  (βλέπε και §Α.4).

Στην τελευταία περίπτωση, αν είναι  $(\underline{X}'\underline{X})^-$  ένας γενικευμένος αντίστροφος του  $\underline{X}'\underline{X}$  (§ Α.5), θα είναι

$$\hat{\underline{\beta}} = (\underline{X}'\underline{X})^- \underline{X}'\underline{Y}. \quad (1.12)$$

Για την εκτίμηση των παραμέτρων  $\underline{\beta}$  από την (1.11) ισχύει:

$$E \hat{\underline{\beta}} = \underline{\beta} \quad \text{και} \quad V(\hat{\underline{\beta}}) = \sigma^2 (\underline{X}'\underline{X})^{-1}. \quad (1.13)$$

Πράγματι, έχουμε:

$$E \hat{\underline{\beta}} = E(\underline{X}'\underline{X})^{-1} \underline{X}'\underline{Y} = (\underline{X}'\underline{X})^{-1} \underline{X}'E \underline{Y} = (\underline{X}'\underline{X})^{-1} \underline{X}'\underline{X} \underline{\beta} = \underline{\beta}$$

και επειδή

$$\hat{\underline{\beta}} - \underline{\beta} = (\underline{X}'\underline{X})^{-1} \underline{X}'\underline{Y} - \underline{\beta} = (\underline{X}'\underline{X})^{-1} \underline{X}'(\underline{X}\underline{\beta} + \underline{\varepsilon}) - \underline{\beta} = (\underline{X}'\underline{X})^{-1} \underline{X}'\underline{\varepsilon}$$

συνεπάγεται

$$\begin{aligned}
 V(\hat{\beta}) &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \\
 &= E((X'X)^{-1} X' \tilde{\varepsilon}) ((X'X)^{-1} X' \tilde{\varepsilon})' = \\
 &= (X'X)^{-1} X' [E(\tilde{\varepsilon} \tilde{\varepsilon}')] X (X'X)^{-1} = \\
 &= (X'X)^{-1} X' \sigma^2 I_n X (X'X)^{-1} = \sigma^2 (X'X)^{-1}.
 \end{aligned}$$

Αν θέσουμε

$$C = (X'X)^{-1} = \begin{bmatrix} c_{00} & c_{01} & c_{02} & \dots & c_{0k} \\ c_{10} & c_{11} & c_{12} & \dots & c_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ c_{k0} & c_{k1} & c_{k2} & \dots & c_{kk} \end{bmatrix} \quad (1.14)$$

τότε η (1.13) γράφεται :

$$\left. \begin{aligned} \text{Var}(\hat{\beta}_i) &= \sigma^2 c_{ii} \\ \text{Cov}(\hat{\beta}_i, \hat{\beta}_j) &= \sigma^2 c_{ij}, i \neq j \end{aligned} \right\} i, j = 0, 1, \dots, k \quad (1.15)$$

Από την τελευταία εξίσωση βρίσκουμε και την τυπική απόκλιση των τυχαίων μεταβλητών  $\hat{\beta}_i$ , που είναι

$$\sigma_{\hat{\beta}_i} = \sqrt{\text{Var} \hat{\beta}_i} = \sigma \sqrt{c_{ii}}, \quad i = 0, 1, \dots, k \quad (1.16)$$

Ας θεωρήσουμε τώρα την ειδική περίπτωση όταν  $k=1$ , όταν δηλ. το μοντέλο έχει τη μορφή

$$Y = \alpha + \beta X + \varepsilon \quad (1.17)$$

Τα δεδομένα στην περίπτωση αυτή δίνονται ως εξής

$X$	$x_1$	$x_2$	$x_3$	$\dots$	$x_n$
$Y$	$y_1$	$y_2$	$y_3$	$\dots$	$y_n$

Θέτοντας

$$\tilde{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \tilde{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad \tilde{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

το μοντέλο (1.17) ανάγεται στο (1.7), οπότε σύμφωνα με τα προηγούμενα θα ισχύει

$$\hat{\beta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X'X)^{-1} X'Y.$$

Αλλά

$$X'X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}, \quad X'Y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix},$$

οπότε

$$(X'X)^{-1} = \frac{1}{n(\sum x_i^2 - n\bar{x}^2)} \begin{pmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix},$$

και

$$\hat{\beta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{1}{\sum x_i^2 - n\bar{x}^2} \begin{pmatrix} \bar{y} \sum x_i^2 - \bar{x} \sum x_i y_i \\ \sum x_i y_i - n\bar{x}\bar{y} \end{pmatrix}.$$

Η τελευταία δίνει

$$\hat{\alpha} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}, \quad \hat{\beta} = \frac{\bar{y} \sum x_i^2 - \bar{x} \sum x_i y_i}{\sum x_i^2 - n\bar{x}^2} = \bar{y} - \hat{\beta}\bar{x}, \quad (1.18)$$

σχέσεις που μπορούν επίσης να βρεθούν με αλγεβρικές πράξεις (βλέπε «Εισαγωγή στη Στατιστική» των κ.κ. Κουνιά-Κολυβά-Μπόρα-Μπαγιάτη, Κεφ. 8), και το μοντέλο γράφεται:

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X. \quad (1.19)$$

Από την (1.15) έχουμε:

$$\text{Var}(\hat{\alpha}) = \sigma^2 \frac{\sum x_i^2}{n(\sum x_i^2 - n\bar{x}^2)} = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) \quad (1.20)$$

και

$$\text{Var}(\hat{\beta}) = \sigma^2 \frac{n}{n(\sum x_i^2 - n\bar{x}^2)} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

σχέσεις που επίσης μπορούν να βρεθούν και απευθείας.

### 1.3. Πρόβλεψη και παρεμβολή

Έστω ότι για τις μεταβλητές  $Y, X_1, X_2, \dots, X_k$  έχουμε κάνει ένα σύνολο παρατηρήσεων που δίνονται με ένα πίνακα της μορφής (1.1). Ενδιαφερόμαστε να εκτιμήσουμε την τιμή  $Y_0$  της  $Y$ , όταν οι τιμές των  $X_i$  είναι δοσμένες, π.χ.  $x_{10}, x_{20}, \dots, x_{k0}$  αντίστοιχα. Μια τέτοια πρόβλεψη της τιμής της  $Y$ , είναι αξιόπιστη όταν οι δοσμένες τιμές των  $X_i$  είναι μέσα στην περιοχή που «καλύπτεται» από τα δεδομένα. Αν  $k=1$ , η πρόβλεψη σ' αυτήν την περίπτωση, λέγεται παρεμβολή (interpolation). Για  $k > 1$  χρησιμοποιούμε γενικά τον όρο πρόβλεψη (prediction - forecasting).

Με την υπόθεση ότι το μοντέλο που προσαρμόζεται στα δεδομένα είναι  $Y = X\beta + \varepsilon$  και ότι τα σφάλματα ικανοποιούν τις συνθήκες (1.8), εκτιμούμε το διάνυσμα των παραμέτρων  $\hat{\beta}$  από τη σχέση (1.11). Αν τώρα θέσουμε  $\underline{x}_0 = (1, x_{10}, x_{20}, \dots, x_{k0})'$ , θα έχουμε για την πρόβλεψη  $y_0$  τη σχέση

$$\hat{Y}_0 = \underline{x}_0' \hat{\beta} \quad (1.21)$$

Για την  $\hat{Y}_0$  ισχύουν:

$$E \hat{Y}_0 = E Y_0 \quad \text{και} \quad \text{Var } \hat{Y}_0 = \sigma^2 \underline{x}_0' C \underline{x}_0 \quad (1.22)$$

Πράγματι,

$$E \hat{Y}_0 = E \underline{x}_0' \hat{\beta} = \underline{x}_0' E \hat{\beta} = \underline{x}_0' \beta = E Y_0$$

και

$$\begin{aligned} \text{Var } \hat{Y}_0 &= E (\underline{x}_0' \hat{\beta} - \underline{x}_0' \beta)^2 = E (\underline{x}_0' \hat{\beta} - \underline{x}_0' \beta) (\underline{x}_0' \hat{\beta} - \underline{x}_0' \beta)' \\ &= E (\underline{x}_0' (\hat{\beta} - \beta) (\hat{\beta} - \beta)' \underline{x}_0) = \underline{x}_0' V(\hat{\beta}) \underline{x}_0 = \\ &= \sigma^2 \underline{x}_0' (X'X)^{-1} \underline{x}_0 \end{aligned}$$

Αν  $k=1$ , τότε η πρόβλεψη στο σημείο  $x$  είναι  $\hat{Y} = \hat{\alpha} + \hat{\beta} x$ , και η (1.22) δίνει για την  $\hat{Y}$  ότι:

$$E \hat{Y} = \alpha + \beta x \quad \text{και} \quad \text{Var } (\hat{Y}) = \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \sigma^2 \quad (1.23)$$

σχέσεις που επίσης μπορούν να βρεθούν και με αλγεβρικές πράξεις.

### 1.4. Το σφάλμα μετά την προσαρμογή του μοντέλου

Μετά την προσαρμογή του μοντέλου (1.5) στα δεδομένα μας, το άθροισμα των τετραγώνων των σφαλμάτων  $\varepsilon_i = Y_i - \hat{Y}_i$  είναι ελάχιστο και θα το συμβολίζουμε SSE. Ισχύει:

$$\begin{aligned} \text{SSE} &= \underline{\varepsilon}' \underline{\varepsilon} = (\underline{Y} - \underline{\hat{Y}})' (\underline{Y} - \underline{\hat{Y}}) = \\ &= (\underline{Y} - \underline{X} \underline{\hat{\beta}})' (\underline{Y} - \underline{X} \underline{\hat{\beta}}). \end{aligned}$$

Επειδή  $\underline{Y}' \underline{X} \underline{\hat{\beta}} = \underline{\hat{\beta}}' \underline{X}' \underline{X} \underline{\hat{\beta}} = \underline{\hat{\beta}}' \underline{X}' \underline{Y}$  έχουμε ακόμη

$$\text{SSE} = \underline{Y}' \underline{Y} - \underline{\hat{\beta}}' \underline{X}' \underline{Y}. \quad (1.24)$$

Αντικαθιστώντας την τιμή του  $\underline{\hat{\beta}}$ , από την (1.11) βρίσκουμε

$$\text{SSE} = \underline{Y}' (\underline{I}_n - \underline{X} (\underline{X}' \underline{X})^{-1} \underline{X}') \underline{Y}, \quad (1.25)$$

δηλ. το SSE είναι μια τετραγωνική μορφή  $\underline{Y}' \underline{A} \underline{Y}$  με πίνακα  $\underline{A} = \underline{I}_n - \underline{X} (\underline{X}' \underline{X})^{-1} \underline{X}'$ .

Είναι φανερό ότι όταν η  $\underline{Y}$  είναι πολυδιάστατη τυχαία μεταβλητή, τότε και η τετραγωνική μορφή  $\underline{Y}' \underline{A} \underline{Y}$  είναι τυχαία μεταβλητή. Οι ιδιότητες αυτών των τυχαίων μεταβλητών μελετώνται στη μαθηματική στατιστική (βλέπε «Μαθηματική Στατιστική», τόμος I, των κ.κ. Κ. Μπαγιάτη, Φ. Κολυβά, κεφ. 0). Θα αναφέρουμε παρακάτω, ορισμένες από τις ιδιότητες αυτές:

(I) Αν  $\underline{Y}$  τυχαίο διάνυσμα (τ.δ.) με  $E \underline{Y} = \underline{\mu}$ ,  $V(\underline{Y}) = \sigma^2 \underline{I}_n$ , τότε

$$E(\underline{Y}' \underline{A} \underline{Y}) = \underline{\mu}' \underline{A} \underline{\mu} + \sigma^2 \text{Tr } \underline{A} \quad (1.26)$$

(II) Αν  $\underline{Y} \sim N(\underline{\mu}, \sigma^2 \underline{I}_n)$ , τότε η τετραγωνική μορφή  $\frac{1}{\sigma^2} \underline{Y}' \underline{A} \underline{Y}$ , έχει κατανομή μη-κεντρική  $\chi^2(m)$ , με παράμετρο  $\lambda = \frac{1}{2\sigma^2} \underline{\mu}' \underline{A} \underline{\mu}$ , αν και μόνον αν ο πίνακας  $\underline{A}$  είναι ταυτοδύναμος με βαθμό  $m$ . (Σημειώνουμε ότι η μη-κεντρική κατανομή  $\chi^2(m)$  με παράμετρο  $\lambda$  προκύπτει ως η κατανομή του αθροίσματος τετραγώνων  $X_1^2 + X_2^2 + \dots + X_m^2$ ,  $m$  κανονικών τ.μ. με κατανομές  $N(\mu_i, \sigma^2)$  και η παράμετρος  $\lambda$  δίνεται από τη σχέση  $\lambda = \frac{1}{2} (\mu_1^2 + \mu_2^2 + \dots + \mu_m^2)$ .

(III) Αν  $\underline{Y} \sim N(\underline{\mu}, I)$ , τότε οι τετραγωνικές μορφές  $\underline{Y}'\underline{A}\underline{Y}$ ,  $\underline{Y}'\underline{B}\underline{Y}$  είναι ανεξάρτητες, τότε και μόνον όταν  $\underline{AB} = 0$ .

(IV) (Θεώρημα Cochran). Αν είναι  $\underline{Y} \sim N(\underline{\mu}, I_n)$  και

$$\underline{Y}'\underline{Y} = \underline{Y}'\underline{A}_1\underline{Y} + \dots + \underline{Y}'\underline{A}_k\underline{Y}$$

όπου  $\underline{A}_1, \underline{A}_2, \dots, \underline{A}_k$   $n \times n$  συμμετρικοί πίνακες με βαθμούς  $n_1, n_2, \dots, n_k$ , τότε κάθε μία από τις παρακάτω πέντε συνθήκες είναι ικανή και αναγκαία συνθήκη για τις άλλες τέσσερις:

- (i)  $n_1 + n_2 + \dots + n_k = n$ .
- (ii) Καθένας από τους  $\underline{A}_i$  είναι ταυτοδύναμος.
- (iii)  $\underline{A}_i \underline{A}_j = 0$  για κάθε  $i \neq j$ .
- (iv)  $\underline{Y}'\underline{A}_i\underline{Y} \sim \chi^2(n_i)$ , μη-κεντρική με  $\lambda_i = \frac{1}{2} \underline{\mu}'\underline{A}_i\underline{\mu}$ ,  $i = 1, 2, \dots, k$ .
- (v)  $\underline{Y}'\underline{A}_i\underline{Y}$ ,  $\underline{Y}'\underline{A}_j\underline{Y}$  είναι ανεξάρτητες τυχαίες μεταβλητές για κάθε  $i \neq j$ .

Για να εφαρμόσουμε την ιδιότητα (I) στην τετραγωνική μορφή SSE, όπως εκφράζεται με την (1.25), παρατηρούμε ότι από τις συνθήκες (1.8) που ικανοποιούν τα σφάλματα έχουμε

$$\underline{E} \underline{Y} = \underline{X} \underline{\beta} \quad \text{και} \quad \underline{V}(\underline{Y}) = \underline{V}(\underline{\varepsilon}) = \sigma^2 I_n.$$

Άρα η (1.26) ισχύει, δηλ.

$$\underline{E}(\text{SSE}) = (\underline{X} \underline{\beta})' \underline{A} (\underline{X} \underline{\beta}) + \sigma^2 \text{Tr } \underline{A} = \sigma^2 (n - k - 1)$$

διότι

$$\begin{aligned} (\underline{X} \underline{\beta})' \underline{A} (\underline{X} \underline{\beta}) &= \underline{\beta}' \underline{X}' (I_n - \underline{X} (\underline{X}' \underline{X})^{-1} \underline{X}') \underline{X} \underline{\beta} = \\ &= \underline{\beta}' \underline{X}' \underline{X} \underline{\beta} - \underline{\beta}' \underline{X}' \underline{X} (\underline{X}' \underline{X})^{-1} \underline{X}' \underline{X} \underline{\beta} = 0, \end{aligned}$$

και

$$\begin{aligned} \text{Tr } \underline{A} &= \text{Tr } I_n - \text{Tr } (\underline{X} (\underline{X}' \underline{X})^{-1} \underline{X}') = n - \text{Tr } ((\underline{X}' \underline{X})^{-1} \underline{X}' \underline{X}) = \\ &= n - \text{Tr } (I_{k+1}) = n - (k+1) = n - k - 1. \end{aligned}$$

Η τελευταία σχέση γράφεται

$$\underline{E}(\text{SSE}/(n - k - 1)) = \sigma^2 \quad (1.27)$$

που σημαίνει ότι το στατιστικό  $s^2 = \frac{\text{SSE}}{n - k - 1}$ , εκτιμά αμερόληπτα τη διασπορά των σφαλμάτων  $\sigma^2$ . Έτσι αν δεν είναι γνωστό το  $\sigma^2$ , η διασπορά των παραμέτρων  $\underline{\beta}_1$ , η τυπική τους απόκλιση, όπως και η διασπορά της πρόβλεψης  $\hat{Y}_0$  βρίσκεται, αν στους τύπους (1.15), (1.16) και (1.22) αντικα-