ΕΙΣΗΓΗΤΗΣ:Ι.ΚΟΥΓΙΑΣ
ΣΠΟΥΔΑΣΤΕΣ:ΚΥΡΙΑΚΗ ΝΙΚΟΛΕΤΟΥ
ΒΑΣΙΛΕΙΟΣ ΝΑΝΟΣ

ΣΥΣΧΕΤΙΣΗ-ΠΑΛΙΝΔΡΟΜΗΣΗ ΣΤΟΝ ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΚΑΙ ΟΙΚΟΝΟΜΙΚΟ ΤΟΜΕΑ

ΑΘΗΝΑ-ΙΟΥΝΙΟΣ 1998

Τ.Ε.Ι ΜΕΣΟΛΟΓΓΙΟΥ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ

ΕΙΣΗΓΗΤΗΣ ΣΚΟΥΓΙΑΣ
ΣΠΟΥΔΑΣΤΕΣ : ΚΥΡΙΑΚΗ ΝΙΚΟΛΕΤΟΥ
ΒΑΣΙΛΕΙΟΣ ΝΑΝΟΣ

ΣΥΣΧΕΤΙΣΗ-ΠΑΛΙΝΔΡΟΜΗΣΗ ΣΤΟΝ ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΚΑΙ ΟΙΚΟΝΟΜΙΚΟ ΤΟΜΕΑ

ΑΘΗΝΑ - ΙΟΥΝΙΟΣ 1998

ΚΕΦΑΛΑΙΟ ΠΡΩΤΟ

ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ

Εισαγωγή

Για να πραγματοποιήσουν πωλήσεις οι επιχειρήσεις χρειάζονται περιουσιακά στοιχεία. Αν οι πωλήσεις πρόκειται να αυξηθούν, θα απαιτηθούν πρόσθετα περιουσιακά στοιχεία. Οι αναπτυσσόμενες επιχειρήσεις χρειάζονται νέες επενδύσεις-άμεσες επενδύσεις σε κυκλοφορούντα στοιχεία- και καθώς εξαντλείται η δυναμικότητα παραγωγής και νέες επενδύσεις σε πάγιες εγκαταστάσεις. Οι πρόσθετες επενδύσεις πρέπει να χρηματοδοτηθούν και η νέα χρηματοδότηση δημιουργεί δεσμεύσεις και υποχρεώσεις για την εξυπηρέτηση των κεφαλαίων που θα αποκτηθούν. Μια αναπτυσσόμενη κερδοφόρα επιχείρηση πιθανόν να χρειαστεί πρόσθετα χρήματα για επενδύσεις σε αποθέματα και πάγιες εγκαταστάσεις, καθώς και για την δημιουργία εισπρακτέων λογαριασμών. Μια τέτοια επιχείρηση είναι πιθανό να οδηγηθεί σε ταμιακό πρόβλημα, εκτός αν η απαιτούμενη ανάγκη για χρηματοδότηση επισημανθεί έγκαιρα.
Μέ την ανάπτυξη πολύπλοκων υποδειγμάτων για την επιχείρηση που κάνουν χρήση ηλεκτρονικών υπολογιστών, η πρόβλεψη των επενδυτικών αναγκών απαιτεί την εφαρμογή πολύ λεπτομερών διαδικασιών. Όλα πάντως τα υποδείγματα στηρίζονται στις βασικές μεθοδολογίες που θα παρουσιάσουμε εδώ, δηλαδή στην μέθοδο παλινδρόμηση-συσχέτιση.

1.1 Μοντέλο

Στα πλαίσια μιας επιστημονικής ανάλυσης συχνά μας ενδιαφέρει να προσδιορίσουμε τη σχέση που συνδέει μια μεταβλητή, έστω y, μέ μία ή περισσότερες άλλες, έστω τις $\chi_1,\chi_2...,\chi_\kappa$.

Στην περίπτωση αυτή λέμε ότι θέλουμε να προσδιορίσουμε ένα μοντέλο το οποίο να ερμηνεύει τη συμπεριφορά της y μέ βάση τη συμπεριφορά των χ_j, j=1,...,κ. Γενικά, το μοντέλο ή υπόδειγμα είναι ένας

μαθηματικός τύπος ο οποίος παρουσιάζει απλουστευτικά σχέσεις που στην πραγματικότητα μπορεί να είναι πολύπλοκες.
Όταν η ανάλυση είναι θεωρητική το μοντέλο μπορεί να είναι **προσδιοριστικό** (deterministic) δηλαδή να ορίζει πλήρως την y αν είναι γνωστή η χ. Ετσι π.χ το μοντέλο κατανάλωσης

$$y=b_o+b_1\chi$$

προσδιορίζει μια τιμή για την κατανάλωση y ενός ατόμου όταν το εισόδημά του είναι χ. Τα b_o και b_1 είναι οι παράμετροι του μοντέλου και μπορεί να είναι οποισδήποτε σταθερές.

1.2 Ανάλυση Παλινδρόμησης

Το καλύτερο που μπορούμε να κάνουμε για ένα στατιστικό μοντέλο είναι να το εκτιμήσουμε. Αυτό, γενικά, σημαίνει να παρατηρήσουμε τη συμπεριφορά της y και συγχρόνως, των μεταβλητών $\chi_1,...\chi_\kappa$ και μέ τη στατιστική ανάλυση αυτών των παρατηρήσεων, να επιλέξουμε ένα μοντέλο το οποίο θα μπορούσε να τις είχε παράγει.
Ολες οι τεχνικές που χρησιμοποιούνται για το σκοπό αυτό αναφέρονται ως **Ανάλυση Παλινδρόμησης** (Regression Analysis) ή **Μεθοδολογία Προσαρμογής Καμπυλών** (Curve Fitting Methodology) ή **Εκτίμηση ενός Μοντέλου Παλινδρόμησης** (Estimation of a Regression Model).
Γενικά, στην Ανάλυση Παλινδρόμησης θεωρούμε ότι η τυχαία μεταβλητή y μπορεί να εκφραστεί ως άθροισμα δύο μερών:
μιας μέσης σχέσης ξ η οποία είναι συνάρτηση των μεταβλητών χ_j, $j=1,...,\kappa$ μέ παραμέτρους $b_1,b_2,...b_\kappa$ και ενός τυχαίου όρου ε. Υποθέτουμε δηλαδή ότι ισχύει:

$$y=\xi+\varepsilon$$

και

$$\xi=\varepsilon(y)=f(\chi_1,\chi_2,...\chi_\kappa / b_1,b_2,...b_\kappa)$$

Η μεταβλητή y ονομάζεται **εξαρτημένη** (dependent) ή **ενδογενής** (endogenous) και οι χ_j, $j=1,...,\kappa$ ονομάζονται **ανεξάρτητες** (independent) ή **εξωγενείς** (exogenous) ή **ερμηνευτικές** (explanatory) ή **μεταβλητές ελέγχου** (control varioubles).
Οι ερμηνευτικές μεταβλητές θεωρούνται κατ'αρχήν ότι μπορούν να ελεγχθούν (δηλαδή ότι ο ερευνητής μπορεί να επιλέξει τα επίπεδα των χ_j στα οποία θα παρατηρήσει την y) και μετριούνται χωρίς σφάλμα. Όμως η ανάλυση παλινδρόμησης μπορεί να επεκταθεί και σε τυχαίες

ερμηνευτικές μεταβλητές, αρκεί να ισχύουν ορισμένες υποθέσεις, όπως ειδικότερα θα δούμε στη συνέχεια.
Η ξ ονομάζεται **συνάρτηση παλινδρόμησης** της y επι των $\chi_1,...\chi_\kappa$. Όταν οι ερμηνευτικές μεταβλητές είναι περισσότερες από μία ($\kappa>1$) τότε ονομάζεται **πολυμεταβλητή** ή **πολλαπλή** διαφορετικά ονομάζεται **απλή**.

1.3 Γραμμική Συνάρτηση Παλινδρόμησης

Εκτός από τις περιπτώσεις που υπάρχει μια καλά αναπτυγμένη σχετική θεωρία, ο μαθηματικός τύπος της συνάρτησης παλινδρόμησης είναι άγνωστος. Σχεδόν πάντα αρχίζουμε την ανάλυση υιοθετώντας την πιο απλή μορφή που είναι η γραμμική δηλαδή η εξής:

$$\xi=b_o+b_1\cdot\chi_1+b_2\cdot\chi_2+...+b_\kappa\cdot\chi_\kappa$$

όπου $b_o,b_1,...b_\kappa$ είναι άγνωστοι παράμετροι. Η μορφή αυτή είναι γενική και περιλαμβάνει οποιαδήποτε γραμμική ως προς τις παραμέτρους συνάρτηση. Αυτό σημαίνει οποιδήποτε συνάρτηση που έχει το ακόλουθο χαρακτηριστικό: καμιά από τις ερμηνευτικές μεταβλητές δεν εξαρτάται από κάποια άγνωστη παράμετρο και καμιά άγνωστη παράμετρος δεν εξαρτάται από οποιαδήποτε άλλη παράμετρο. Ετσι π.χ οι συναρτήσεις

$$\xi=b_o+b_1\cdot\chi_1+b_2^{b3}\cdot\chi_2$$

ή $$\xi=b_o+b_1\cdot\chi_1+b_2\cdot\chi_2^{b3}$$

δεν είναι γραμμικές ενώ οι

$$\xi=b_o+b_1\cdot ln\cdot\chi_1+b_2\cdot\chi_1^{2}$$

ή $$\xi=b_o+b_1\cdot\chi_1^{1/2}+b_2\cdot\chi_1\cdot\chi_2$$

είναι γραμμικές.

ΚΕΦΑΛΑΙΟ ΔΕΥΤΕΡΟ

ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

2.1 Το απλό γραμμικό μοντέλο παλινδρόμησης

Το απλό γραμμικό μοντέλο παλινδρόμησης είναι αυτό στο οποίο η μέση τιμή της εξαρτημένης μεταβλητής y είναι γραμμική συνάρτηση της ανεξάρτητης μεταβλητής χ. Ετσι, αν $y_1,...,y_n$ οι παρατηρήσεις της y στα επίπεδα, αντίστοιχα, $\chi_1,...,\chi_n$ της χ τότε ισχύει:

$$y_i=b_o+b_1*\chi_i+\varepsilon_i,\ i=1,...n \qquad (2.1.1)$$

όπου ε_j ο όρος σφάλματος. Για να εκτιμήσουμε τις παραμέτρους b_o+b_1 μέ τη μέθοδο των ελαχίστων τετραγώνων κάνουμε τις ακόλουθες υποθέσεις:

I. Για την ερμηνευτική μεταβλητή

i. Οι τιμές χ_i, i=1,...n είναι σταθερές, όχι όλες ίσες μεταξύ τους που επιλέγονται από τον ερευνητή και μετριούνται χωρίς σφάλμα.

Αυτό σημαίνει ότι η χ μπορεί να προκαλεί μεταβολές στην y αλλά όχι το αντίστροφο. Καθώς η y είναι τυχαία μεταβλητή, αν προκαλούσε μεταβολές στην χ, θα ήταν κι αυτή τυχαία. Η συνθήκη αυτή είναι πολύ περιοριστική, ιδιαίτερα για τις οικονομικές μεταβλητές οι οποίες είναι σχεδόν πάντα τυχαίες. Αποδεικνύεται όμως ότι η μέθοδος των ελαχίστων τετραγώνων ορίζει άριστους εκτιμητές και όταν ισχύει η λιγότερο περιοριστική συνθήκη :

i'. Οι χi είναι παρατηρήσεις της τυχαίας μεταβλητής χ η οποία είναι ασυσχέτιστη μέ κάθε διαταρακτικό όρο ε_i, i=1,...,n.

Στην περίπτωση αυτή όλες οι εκτιμήσεις και οι αναγωγές γίνονται υπο συνθήκη των παρατηρήσεων χ_i.

ΙΙ. Για τον όρο σφάλματος

Ισχύει :

i. $E(\varepsilon_i) = 0 \Leftrightarrow E(Y_i) = b_0 + b_1 X_i$ $\quad i = 1, \ldots, n$

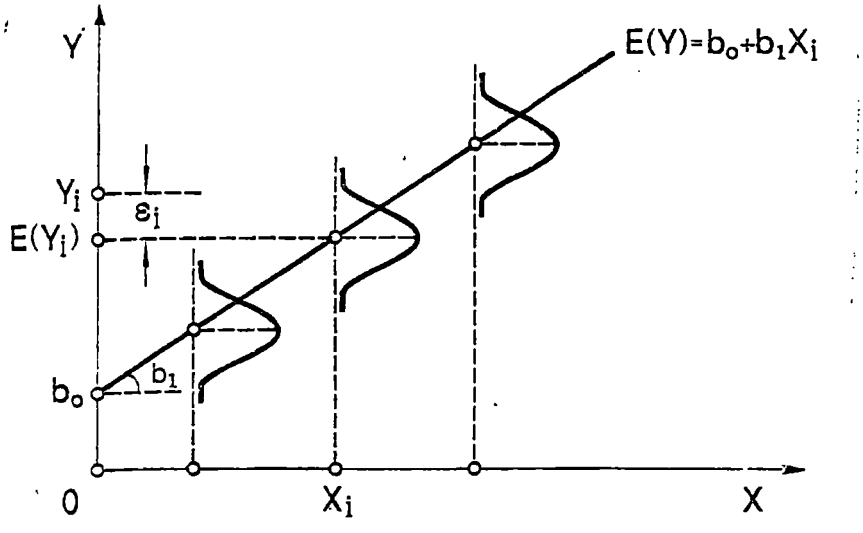
ii. $Var(\varepsilon_i) = \sigma^2 \Leftrightarrow Var(Y_i) = \sigma^2$ $\quad i = 1, \ldots, n$

iii. $Cov(\varepsilon_i, \varepsilon_j) = 0 \Leftrightarrow Cov(Y_i, Y_j) = 0$ $\quad \forall i \neq j$

iv. $\varepsilon_i \sim N(0, \sigma^2) \Leftrightarrow Y_i \sim N(b_0 + b_1 X_i, \sigma^2)$ $\quad i = 1, \ldots, n$

Οι συνθήκες (i)-(iv) μαζί με την (2.1.1) ορίζουν ένα μοντέλο δειγματοληψίας στο οποίο οι παρατηρήσεις yi είναι ανεξάρτητες, έχουν την ίδια κατανομή, ίδια διακύμανση και μέση τιμή που βρίσκεται επάνω σε μια ευθεία με σταθερό όρο b_0 και κλίση b_1.Η ευθεία αυτή ονομάζεται **θεωρητική ή ευθεία παλινδρόμησης πληθυσμού**.

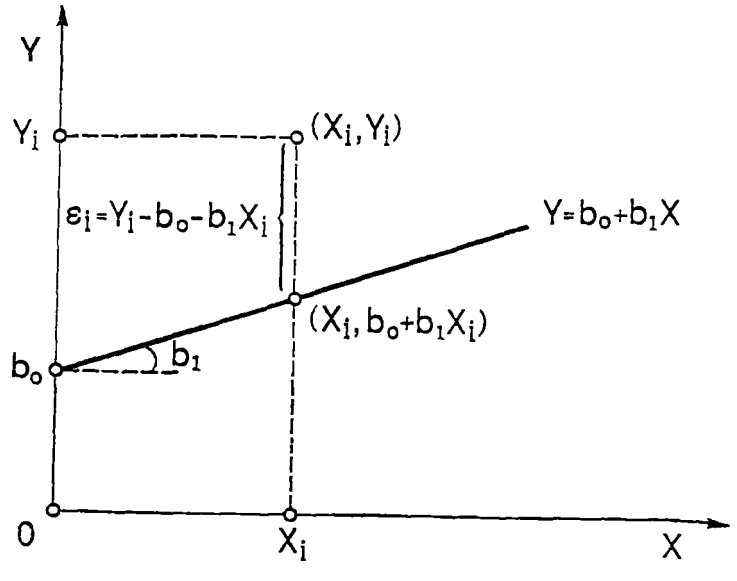
Η γραφική παράσταση ενός τέτοιου μοντέλου δίνεται στο ακόλουθο σχήμα :



2.2 Η μέθοδος των ελαχίστων τετραγώνων

Ας θεωρήσουμε το ακόλουθο σχήμα στο οποίο δίνεται η ευθεία L μέ εξίσωση $y=b_o+b_1 \cdot x$ και ένα σημείο (χi,yi).Η απόσταση του σημείου αυτού από την L είναι η

$$\varepsilon_i = Y_i - (b_o + b_1 \cdot X_i)$$



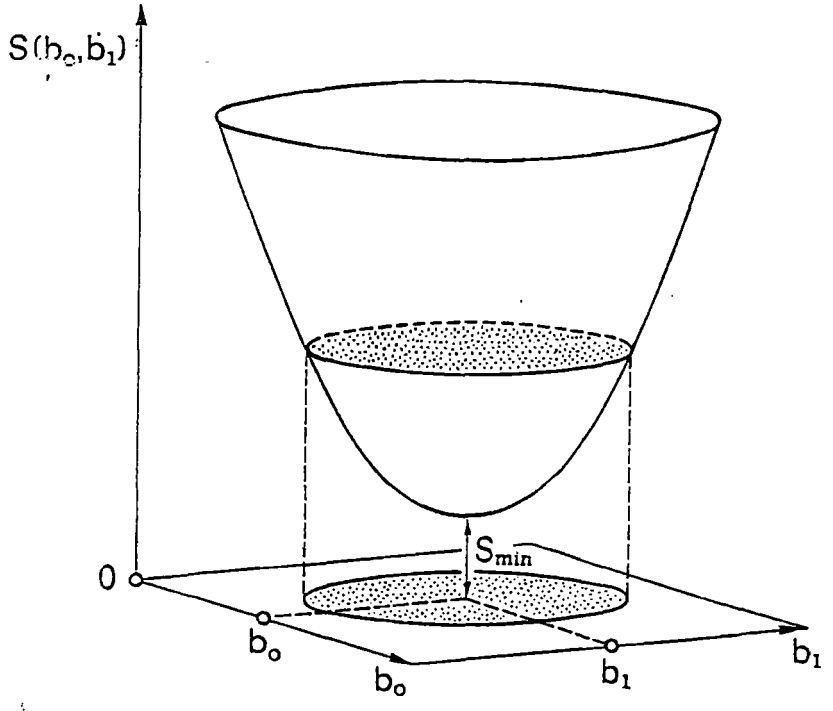
Παρατηρούμε ότι αν θεωρήσουμε μια άλλη ευθεία L' μέ εξίσωση $y=b'_o+b'_1 \cdot x$ τότε η απόσταση του σημείου (χi,yi) από την ευθεία L' θα είναι η

$$\varepsilon'_i = Y_i - (b'_o + b'_1 \cdot X_i)$$

Από όλες τις δυνατές ευθείες στο επίπεδο xy **η ευθεία ελαχίστων τετραγώνων** είναι εκείνη που ελαχιστοποιεί το άθροισμα των τετραγώνων των ή των αποστάσεων ε'i=yi-(b_o+b'$_1$*xi) δηλαδή αυτή για την οποία η συνάρτηση

$$S = S(b_0, b_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i)^2 \qquad (2.2.1)$$

έχει ελάχιστο. Η γραφική παράσταση μιας τέτοιας συνάρτησης είναι μια παραβολοειδής επιφάνεια σχήματος μπωλ όπως αυτή που δίνεται στο σχήμα.



Εστω $\hat{b}_o, \hat{b}_1$ οι τιμές των b_o, b_1 αντίστοιχα, οι οποίες όταν αντικατασταθούν στην εξίσωση (2.2.1) δίνουν την ελάχιστη δυνατή τιμή για την S. Για να προσδιορίσουμε τις $\hat{b}_o$ και $\hat{b}_1$ υπολογίζουμε τις μερικές παραγώγους της S

ως προς b_o και b_1 και εξισώνουμε μέ μηδέν. Εχουμε (οι δείκτες στα Σ παραλείπονται) :

$$\frac{\partial S}{\partial b_o} = -2\Sigma(Y_i - b_o - b_1 \cdot X_i)$$

$$\frac{\partial S}{\partial b_1} = -2\Sigma X_i(Y_i - b_o - b_1 \cdot X_i)$$

Οι τιμές των $\hat{b}_o, \hat{b}_1$ προσδιορίζονται από τη λύση του συστήματος εξισώσεων:

$$\left.\begin{aligned}\Sigma(Y_i - \hat{b}_o - \hat{b}_1 \cdot X_i) = 0\\ \Sigma X_i(Y_i - \hat{b}_o - \hat{b}_1 \cdot X_i) = 0\end{aligned}\right\} \Longleftrightarrow$$

$$\left.\begin{aligned}\Sigma Y_i - n\hat{b}_o - \hat{b}_1 \cdot \Sigma X_i = 0\\ \Sigma X_i \cdot Y_i - \hat{b}_o \Sigma X_i - \hat{b}_1 \Sigma X_i^2 = 0\end{aligned}\right\} \Longleftrightarrow$$

$$\begin{aligned} n\hat{b}_o + \hat{b}_1 \Sigma X_i &= \Sigma Y_i \\ \hat{b}_o \Sigma X_i + \hat{b}_1 \Sigma X_i^2 &= \Sigma X_i \cdot Y_i \end{aligned} \qquad (2.2.2)$$

Οι (2.2.2) ονομάζονται **κανονικές εξισώσεις** και είναι ένα γραμμικό σύστημα μέ αγνώστους τα b_o και b_1. Η λύση του συστήματος αυτού είναι η :

$$\hat{b}_o = \bar{Y} - \hat{b}_1 \cdot \bar{X} \qquad (2.2.3)$$

$$\hat{b}_1 = \frac{\Sigma \bar{X}_i \cdot \bar{Y}_i - n\bar{X}\bar{Y}}{\Sigma X_i^2 - n \cdot \bar{X}^2} \qquad (2.2.4)$$

Χρήσιμη είναι η ισοδύναμη έκφραση για το b_1

$$\hat{b}_1 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} \qquad (2.2.4)'$$

Η ευθεία ελαχίστων τετραγώνων είναι η

$$\hat{y} = \hat{b}_0 + \hat{b}_1 \cdot x$$

και ονομάζεται **δειγματική ή εκτιμηθείσα ευθεία παλινδρόμησης σε αντιδιαστολή μέ τη θεωρητική ή ευθεία παλινδρόμησης στον πληθυσμό.**

2.3 Τα κατάλοιπα ή σφάλματα εκτίμησης

Αν στο εκτιμηθέν μοντέλο αντικαταστήσουμε όπου χ τις τιμές χi, i=1,...,n των δεδομένων μας παίρνουμε τις εκτιμήσεις

$$\hat{y}_i=\hat{b}_o+\hat{b}_1 \cdot x_i, \quad i=1,...,n$$

Οι διαφορές των εκτιμήσεων $\hat{y}_i$ από τις παρατηρήσεις y_i συμβολίζονται μέ $\hat{\varepsilon}_i$ και ονομάζονται **σφάλματα εκτίμησης ή κατάλοιπα.** Δηλαδή έχουμε :

$$\begin{aligned}\hat{\varepsilon}_i &= Yi-\hat{Y}i \\ &= y_i - \hat{b}_o-\hat{b}_1 \cdot X_i, \quad i=1,...,n\end{aligned}$$

Το κατάλοιπο $\hat{\varepsilon}_i$ είναι το μέρος της παρατήρησης Y_i το οποίο μένει ανερμήνευτο από το εκτιμηθέν μοντέλο και αποτελεί μια εκτίμηση του διαταρακτικού όρου ε_i. Επομένως, η συμπεριφορά των καταλοίπων $\hat{\varepsilon}_i$, i=1,...,n μας δίνει χρήσιμες πληροφορίες για τη συμπεριφορά των διαταρακτικών όρων.

Παράδειγμα :

Οι εκτιμήσεις Y_i και τα κατάλοιπα $\hat{\varepsilon}_i$ του παραδείγματος της αξίας των πωλήσεων δίνονται στον παρακάτω πίνακα :

Y_i	X_i	$Y_i =$	$\varepsilon_i = Y_i - Y_i$
1100	170,0	1168,86	-68,86
1180	172,0	1407,22	-187,22
1124	169,5	1574,64	25,36
1350	190,0	1812,43	-52,43
1220	212,0	1835,13	-185,13
1590	228,0	1180,21	0,22
1340	202,0	1498,02	91,98
1570	235,0	1567,26	-17,26
1660	241,5	1694,95	-64,95
1550	240,2	1746,60	73,41
1820	283,3	1166,03	73,97
2040	300,4	1350,47	-10,47
1760	283,4	1811,86	8,14
1630	262,7	1916,28	3,72
1920	301,7	1282,37	67,63
1720	253,2	1537,75	32,25
1650	287,4	1908,90	131,10
1820	271,8	1641,04	78,96

Ιδιότητες των καταλοίπων

Τα κατάλοιπα $\hat{\varepsilon}_i$, i=1,...,n της εκτίμησης ενός μοντέλου έχουν τις ακόλουθες ιδιότητες :

1η ιδιότητα : Το άθροισμα τους ισούται μέ μηδέν, δηλαδή (ο συντελεστής Σ είναι για i από 1 ως n) :

$$\Sigma \hat{\varepsilon}_i = 0$$

Πράγματι έχουμε :

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i = \bar{Y} - \hat{b}_1 \cdot \bar{X} + \hat{b}_1 \cdot X_i = \bar{Y} + \hat{b}_1 (X_i - \bar{X})$$

και

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \bar{Y} - \hat{b}_1 (X_i - \bar{X}) \Leftrightarrow$$
$$\Sigma \hat{\varepsilon}_i = \Sigma (Y_i - \bar{Y}) - \hat{b}_1 \Sigma (X_i - \bar{X}) = 0$$

2η ιδιότητα : Το άθροισμα των γινομένων $\chi_i \cdot \hat{\varepsilon}_i$ ισούται μέ μηδέν δηλαδή

$$\boxed{\Sigma \chi_i \cdot \hat{\varepsilon}_i = 0}$$

Πράγματι έχουμε :

$$\Sigma X_i \cdot \hat{\varepsilon}_i = \Sigma X_i (Y_i - \hat{b}_0 - \hat{b}_1 \cdot X_i)$$
$$= \Sigma X_i Y_i - \hat{b}_0 \Sigma X_i - \hat{b}_1 \cdot \Sigma X_i^2$$
$$= 0 \quad \text{από τη δεύτερη κανονική εξίσωση.}$$

Σημείωση : Η 1η και η 2η ιδιότητα χαρακτηρίζουν την ευθεία ελαχίστων τετραγώνων αρκεί μια ευθεία να έχει τις δύο αυτές ιδιότητες για να είναι ευθεία ελαχίστων τετραγώνων. Τα κατάλοιπα όμως έχουν και την ακόλουθη σημαντική ιδιότητα :

3η ιδιότητα : Το άθροισμα των γινομένων $\hat{y}_i \cdot \hat{\varepsilon}_i$ ισούται μέ μηδέν, δηλαδή

$$\boxed{\Sigma \hat{y}_i \cdot \hat{\varepsilon}_i = 0}$$

Πράγματι έχουμε :

$$\hat{Y} \cdot \hat{\varepsilon}_i = (\hat{b}_0 + \hat{b}_1 \cdot X_i) \cdot (Y_i - \hat{b}_0 - \hat{b}_1 \cdot X_i)$$
$$= \hat{b}_0 (Y_i - \hat{b}_0 - \hat{b}_1 \cdot X_i) + \hat{b}_1 \cdot X_i (Y_i - \hat{b}_0 - \hat{b}_1 X_i)$$

Αθροίζοντας για όλες τις τιμές του i παίρνουμε :

$$\Sigma \hat{Y}_i \cdot \hat{\varepsilon}_i = \hat{b}_0 (\Sigma Y_i - n \hat{b}_0 - \hat{b}_1 \Sigma X_i) + \hat{b}_1 (\Sigma X_i Y_i - \hat{b}_0 \Sigma X_i - \hat{b}_1 \Sigma X_i^2)$$
$$= 0$$

αφού από τις δύο κανονικές εξισώσεις οι παραστάσεις στις δύο παρενθέσεις μηδενίζονται

2.4 Μέση τιμή και διακύμανση των εκτιμητών ελαχίστων τετραγώνων.

Οι εκτιμήσεις των παραμέτρων bo και b1 της θεωρητικής ευθείας παλινδρόμησης βασίζονται σε ένα τυχαίο δείγμα ή ζευγών παρατηρήσεων (xi, Yi). Οι Yi είναι τυχαίες μεταβλητές. Είναι προφανές επομένως, ότι από μια άλλη διαδικασία παρατήρησης στην οποία οι τιμές xi παραμένουν ίδιες θα είχαμε, με πιθανότητα ίση με 1, διαφορετικές τιμές για τις Yi και διαφορετικές εκτιμήσεις $\hat{b}o$ και $\hat{b}1$. Επομένως οι $\hat{b}o$ και $\hat{b}1$ που ορίζονται από τους τύπους (2.2.3) και (2.2.4) ως συναρτήσεις των τυχαίων μεταβλητών Yi, είναι τυχαίες μεταβλητές και θα αναφέρονται ως εκτιμητές των bo και b1. Αντίστοιχα, οι μεμονωμένες τιμές που υπολογίζουμε από ορισμένο δείγμα θα αναφέρονται ως εκτιμήσεις.

Αποδεικνύεται ότι οι εκτιμητές ελαχίστων τετραγώνων $\hat{b}o$ και $\hat{b}1$ είναι αμερόληπτοι εκτιμητές των bo και b1, δηλαδή ισχύει:

$$E(\hat{b}_0) = b_0 \quad \text{και} \quad E(\hat{b}_1) = b_1 \qquad (2.4.1)$$

και οι διακυμάνσεις τους είναι, αντίστοιχα, ίσες με (ο συντελεστής Σ είναι για τιμές από 1 ως η):

$$Var(\hat{b}_0) = \sigma^2_{\hat{b}_0} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(x_i - \bar{X})^2}\right] \qquad (2.4.2)$$

$$Var(\hat{b}_1) = \sigma^2_{\hat{b}_1} = \frac{\sigma^2}{\Sigma(x_i - \bar{X})^2} \qquad (2.4.3)$$

Για να χρησιμοποιήσουμε τους τύπους (2.4.2) και (2.4.3) θα πρέπει να εκτιμήσουμε την σ^2. Αποδεικνύεται ότι, όταν ισχύουν οι υποθέσεις του γραμμικού μοντέλου ο τύπος:

$$s^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-2} = \frac{\sum (y_i - \hat{b}_0 - \hat{b}_1 \cdot x_i)^2}{n-2} \qquad (2.4.4)$$

ορίζει ένα αμερόληπτο εκτιμητή της κοινής διακύμανσης των διαταρακτικών όρων δηλαδή της σ^2. Αν αναπτύξουμε τον αριθμητή της (2.4.4) προκύπτει ο ισοδύναμος τύπος:

$$s^2 = \frac{\sum y_i^2 - \hat{b}_0 \sum y_i - \hat{b}_1 \sum y_i \cdot x_i}{n-2} \qquad (2.4.5)$$

που είναι υπολογιστικά ευκολότερος αφού όλοι οι όροι του έχουν υπολογιστεί προηγουμένως. Το s^2 **μέσο τετραγωνικό σφάλμα**. Η τατραγωνική του ρίζα s ονομάζεται **τυπικό σφάλμα της εκτίμησης** και μετρά τη διασπορά των παρατηρήσεων Yi γύρω από την ευθεία ελαχίστων τετραγώνων σε μονάδες της Y.

Αν στις (2.4.2) και (2.4.3) αντικαταστήσουμε την σ^2 με την εκτίμηση s^2 παίρνουμε τους τους αμερόληπτους εκτιμητές των $\mathrm{Var}(\hat{b}_0)$ και $\mathrm{Var}(\hat{b}_1)$ που συμβολίζονται με $s^2_{\hat{b}_0}$ και $s^2_{\hat{b}_1}$, αντίστοιχα. Οι τετραγωνικές ρίζες αυτών των εκτιμητών είναι τα τυπικά σφάλματα, αντίστοιχα, των εκτιμητών $\hat{b}_0$ και $\hat{b}_1$.

Επομένως, το τυπικό σφάλμα του $\hat{b}_0$ είναι το εξής:

$$s_{\hat{b}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

ενώ το τυπικό σφάλμα του $\hat{b}_1$ το εξής:

$$s_{\hat{b}_1} = s\sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}$$

2.5 Πόσο καλοί είναι οι εκτιμητές ελαχίστων τετραγώνων

Όταν ισχύουν οι υποθέσεις του γραμμικού μοντέλου παλινδρόμησης οι εκτιμητές ελαχίστων τατραγώνων είναι:

Γραμμικοί:δηλαδή είναι γραμμικές συναρτήσεις των παρατηρήσεων Yi. Πράγματι έχουμε:

$$\hat{b}_1 = \frac{\Sigma(x_i-\bar{x})\cdot(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = \frac{\Sigma(x_i-\bar{x})}{\Sigma(x_i-\bar{x})^2}(y_i-\bar{y})$$

$$\hat{b}_0 = \bar{y}-\hat{b}_1\cdot\bar{x}$$

Επειδή τα Xi είναι σταθεροί αριθμοί οι όροι Σ (xi-x)/Σ(xi-x)^2 και $\bar{x}$ είναι επίσης σταθερές.Επομένως στους τύπους των $\hat{b}1$ και $\hat{b}o$ περιέχονται μόνο οι όροι οι οποίοι είναι γραμμικοί ως προς τις Yi. Η σημασία της ιδιότητας αυτής οφείλεται στο ότι, συνήθως, η κατανομή μιας γραμμικής συνάρτησης τυχαίων μεταβλητών προσδιορίζεται εύκολα.

Αμερόληπτοι: δηλαδή $E(\hat{b}_0)=b_0$ και $E(\hat{b}_1)=b_1$

Άριστοι: δηλαδή από όλους τους γραμμικούς αμερόληπτους εκτιμητές έχουν τη μικρότερη διακύμανση .

Έχει καθειερωθεί οι ιδιότητες αυτές να αναφέρονται συνοπτικά ως BLUE (=Best Linear Unbiased Estimators). Οι Άριστοι ,Γραμμικοί ,Αμερόληπτοι εκτιμητές δεν μας εξασφαλίζουν ότι οι εκτιμήσεις που θα πάρουμε από ένα μόνο δείγμα θα είναι κοντά στις πραγματικές τιμές των παραμέτρων , παρά μόνον πιθανοκρατικά . Έτσι , οι εκτιμήσεις που θα προκύψουν από τη μέθοδο των ελαχίστων τετραγώνων θα είναι, με μεγαλύτερη πιθανότητα, κοντά στις πραγματικές τιμές b..και b..από τις εκτιμήσεις που θα προκύψουν από οποιαδήποτε άλλη μέθοδο που χρησιμοποιεί γραμμικές συναρτήσεις των y_i .

Το αποτέλεσμα αυτό είναι γνωστό ως θεώρημα των Gauss-Markov Σημειώνεται επίσης ότι επειδή υποθέσεις του γραμμικού μοντέλου είναι αποφασιστικές για την ισχύ του αποτελέσματος αυτού για αυτό ονομάζονται και συνθήκες Gauss- Markov.

2.6 Διαστήματα εμπιστοσύνης και έλεγχος των υποθέσεων.

Αποδεικνύεται ότι, όταν οι διαταρακτικοί όροι είναι ανεξάρτητες και κανονικές τυχαίες μεταβλητές μέ μέση τιμή μηδέν και διακύμανση σ^2, δηλαδή συνοπτικά , όταν

$$\varepsilon_i \sim IDN(0, \sigma^2), \quad i = 1, \dots, n$$

τότε, οι τυχαίες μεταβλητές

$$\frac{\hat{b}_0 - b_0}{S_{\hat{b}_0}} \quad \text{και} \quad \frac{\hat{b}_1 - b_1}{S_{\hat{b}_1}}$$

ακολουθούν την κατανομή t-student μέ ν=n-2 βαθμούς ελευθερίας. Μέ βάση το αποτέλεσμα αυτό, μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης και να κάνουμε έλεγχο των υποθέσεων για τις παραμέτρους b_0 και b_1 της θεωρητικής ευθείας παλινδρόμησης. Ειδικότερα έχουμε :

Διαστήματα εμπιστοσύνης

Εστω $t_{n-2,\alpha/2}$ η τιμή της κατανομής t-student μέ n-2 βαθμούς ελευθερίας, για την οποία ισχύει $P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$ και επομένως $P(-t_{n-2,\alpha/2} \leq t_{n-2} \leq t_{n-2,\alpha/2}) = 1-\alpha$

Τότε, το διάστημα το οποίο μέ επίπεδο εμπιστοσύνης (1-α) θα περιέχει την παράμετρο b_j, j=0,1 θα είναι το εξής :

$$\hat{b}_j - t_{n-2,\alpha/2} S_{\hat{b}_j} \leq b_j \leq \hat{b}_j + t_{n-2,\alpha/2} S_{\hat{b}_j}$$

Ειδικότερα για j=0 το διάστημα αυτό είναι το

$$\hat{b}_0 - t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{\bar{x}}{\Sigma(x_i - \bar{x})^2}} \leq b_0 \leq \hat{b}_0 + t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{\bar{x}}{\Sigma(x_i - \bar{x})^2}}$$

ενώ για j=1 είναι το

$$\hat{b}_1 - t_{n-2,\alpha/2}\, s\sqrt{\frac{1}{\Sigma(x_i-\bar{x})^2}} \le b_1 \le \hat{b}_1 + t_{n-2,\alpha/2}\, s\sqrt{\frac{1}{\Sigma(x_i-\bar{x})^2}}$$

Αυτό σημαίνει ότι η πιθανότητα για κάθε ένα από τα διαστήματα αυτά να περιέχει την αντίστοιχη παράμετρο ισούται μέ 1-α. Είναι προφανές ότι για μια παράμετρο bj, όσο πιο μικρό είναι το αντίστοιχο διάστημα εμπιστοσύνης τόσο μεγαλύτερη η ακρίβεια της εκτίμησης. Το εύρος του διαστήματος εμπιστοσύνης είναι $2t_{n-2,\alpha/2}\, s_{\hat{b}j}$., επομένως όσο πιο μικρό το $S_{\hat{b}_j}$ τόσο μεγαλύτερη η αξιοπιστία του $\hat{b}_j$

Ελεγχος των υποθέσεων

Όταν εκτιμούμε ένα μοντέλο παλινδρόμησης μας ενδιαφέρει να ελέγξουμε την υπόθεση ότι η συμβολή της χ στην ερμηνεία της εξαρτημένης δεν είναι στατιστικά σημαντική. Αυτό ισοδυναμεί μέ τον έλεγχο της μηδενικής υπόθεσης $H_o: b_1=0$, οπότε η παρατηρούμενη διαφορά του b_1 από το μηδέν μπορεί να αποδοθεί στις διακυμάνσεις της δειγματοληψίας. Ο δίπλευρος έλεγχος της υπόθεσης αυτής θα γίνει ως εξής:

<u>Η μηδενική και η ενναλακτική υπόθεση</u>

$$H_0: b_1 = 0$$

$$H_\varepsilon: b_1 \neq 0$$

<u>Το κριτήριο αποφάσεως</u>

$$\text{Αν } |t_\pi| = \frac{\hat{b}_1}{s_{\hat{b}_1}} > t_{n-2,\alpha/2}$$

Θα απορίψουμε την H_o στο επίπεδο σημαντικότητας α. Αξίζει να σημειωθεί ότι αν α=0,05 και n-2>12, η τιμή $t_{n-2,\alpha/2}$ είναι περίπου ίση μέ 2.Γι'αυτό συχνά χρησιμοποιείται ο ακόλουθος πρακτικός κανόνας: η τιμή του $\hat{b}_1$ θεωρείται στατιστικά σημαντική, όταν είναι τουλάχιστον διπλάσια

από το $s_{\hat{b}_1}$. Πάντως ο κανόνας αυτός, όπως και οποισδήποτε άλλος πρακτικός κανόνας θα πρέπει να χρησιμοποιείται μέ προσοχή.
Είναι δυνατόν, να έχουμε α priori πληροφορίες για το πρόσημο του b_1 οπότε εξειδικέυεται μονόπλευρος έλεγχος ο οποίος είναι πιο ευαίσθητος από το δίπλευρο. Τότε, το στατιστικό του ελέγχου συγκρίνεται μέ την τιμή
$t_{n-2,\alpha}$ ή την $-t_{n-2,\alpha}$ ανάλογα μέ το αν ο έλεγχος είναι δεξιόπλευρος ή αριστερόπλευρος.
Εξάλλου, είναι δυνατό να έχουμε α priori πληροφορίες για την τιμή του b_1. Τότε ο έλεγχος θα γίνει ως εξής:

Η μηδενική και η εναλλακτική υπόθεση

$H_0 : b_1 = b_1^\circ$

H_ε : i) $b_1 \neq b_1^\circ$
ii) $b_1 > b_1^\circ$
iii) $b_1 < b_1^\circ$

Το κριτήριο αποφάσεως

Υπολογίζουμε $$t_n = \frac{\hat{b}_1 - b_1^\circ}{s_{\hat{b}_1}}$$

οπότε απορρίπτουμε μέ την H_o στο επίπεδο σημαντικότητας α, για τα τρία είδη ελέγχου, αντίστοιχα αν:

i) $|t_n| > t_{n-2,\alpha/2}$
ii) $t_n > t_{n-2,\alpha}$
iii) $t_n < -t_{n-2,\alpha}$

Για το σταθερό όρο, συνήθως, γίνεται ο δίπλευρος έλεγχος για τη μηδενική τιμή του. Η διαδικασία είναι ανάλογη μ'αυτήν που περιγράψαμε για τον b_1 όπως και στο παράδειγμα που ακολουθεί.

Σημείωση 1: Όταν οι χ και y δεν συνδέονται γραμμικά, τότε $b_1=0$. Και στην περίπτωση αυτή όμως, λόγω διακυμάνσεων της δειγματοληψίας μπορεί να υπολογίσουμε μια εκτίμηση $\hat{b}_1$ διαφορετική από το μηδέν. Τότε, μέ πιθανότητα (1-α) θα οδηγηθούμε στην αποδοχή της H_o: $b_1=0$. Θα πρέπει να τονιστεί ότι η αποδοχή της H_o δεν σημαίνει αναγκαστικά

ότι οι χ ή y είναι ανεξάρτητες και άρα ότι θα πρέπει να υιοθετήσουμε το βασικό μοντέλο

$$y_i = b_o + \varepsilon_i \qquad i=1,\ldots,n$$

Είναι δυνατόν οι χ και y να συνδέονται μέ μια σχέση η οποία γίνεται γραμμική μέ τον κατάλληλο μετασχηματισμό της χ ή (και) της y.

Σημείωση 2: Είναι προφανές ότι από τους δύο ελέγχους ο πιο σημαντικός είναι ο έλεγχος για την παράμετρο b_1. Ετσι, όταν δεν μπορούμε να απορίψουμε την $H_o: b_1 = 0$ θα κάνουμε νέα εξειδίκευση του μοντέλου ενώ συνήθως αφήνουμε το σταθερό όρο ακόμη και όταν η υπόθεση $H_o: b_o = 0$ δεν μπορέι να απορριφθέι.

Παράδειγμα

Στο παράδειγμα της αξίας των πωλήσεων θα κάνουμε το δίπλευρο έλεγχο των υποθέσεων για μηδενικής τιμές των παραμέτρων b_o και b_1 ως εξής :

Για το σταθερό όρο

$H_o: b_o = 0$

$H_\varepsilon: b_o \neq 0$

Επειδή: $|t_n| = \dfrac{|\hat{b}_o|}{S_{\hat{b}_o}} = \dfrac{204,09}{118,14} = 1,73 < t_{16(0,025)} = 2,12$

δεν μπορούμε να απορίψουμε την H_o

Για την κλίση της ευθείας

$H_o: b_1 = 0$

$H_\varepsilon: b_1 \neq 0$

Επειδή: $|t_n| = \frac{\hat{b}_1}{S_{\hat{b}_1}} = \frac{5,675}{0,486} = 11,68 > 2,12$

απορρίπτουμε την H_o στο $\alpha=0,05$. Στην πραγματικότητα η H_o απορρίπτεται και για μικρότερες τιμές του α.

2.7 Η ερμηνευτική ικανότητα του μοντέλου

Στο τμήμα αυτό θα παρουσιάσουμε ένα μέτρο της ικανότητας του μοντέλου να ερμηνεύσει τη συνολική μεταβλητότητα των παρατηρήσεων y_i γύρω από τη μέση τιμή τους. Κάθε παρατήρηση y_i μπορεί να γραφεί ως εξής:

$$y_i = \hat{y}_i + \hat{\varepsilon}_i \quad i=1,...,n \qquad (2.7.1)$$

όπου $\hat{y}_i = \hat{b}_c + \hat{b}_1 \cdot x_i$ η εκτίμηση που ορίζεται από την ευθεία ελαχίστων τετραγώνων στη τιμή $\chi=\chi_i$ και $\hat{\varepsilon}_i = y_i - \hat{y}_i$, τα κατάλοιπο της εκτίμησης.
Η (2.7.1) μπορεί να γραφεί ισοδύναμη ως εξής:

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

$$\Longleftrightarrow Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \quad i = 1, \ldots, n \qquad (2.7.2)$$

Αν υψώσουμε τις δύο πλευρές στο τετράγωνο και αθροίσουμε για όλες τις τιμές του i τότε, επειδή $\Sigma(\hat{y}_i - \bar{y}) \cdot (y_i - \hat{y}_i)=0$ παίρνουμε :

$$\Sigma(y_i - \bar{y})^2 = \Sigma (\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2 \qquad (2.7.3)$$

Αν θέσουμε :

$\Sigma(y_i - \bar{y})^2$=SST= **Ολικό Αθροισμα Τετραγώνων**
$\Sigma(\hat{y}_i - \bar{y})^2$=SSR= **Αθροισμα Τετραγώνων Παλινδρόμησης**
$\Sigma (Yi-\hat{Y}i)^2$=SSE= **Άθροισμα Τετραγώνων των Σφαλμάτων**

τότε η (2.7.3.) γράφεται ισοδύναμα ως εξής :

SST=SSR+SSE

Σημειώνεται ότι αν για την Y εκτιμηθεί το βασικό μοντέλο τότε εκτιμούμε τη σταθερή μέση τιμή με την $\bar{Y}$ και τα κατάλοιπα είναι οι διαφορές $Yi-\bar{Y}$, i=1, ...,n. Επομένως ο όρος $SST=\Sigma(Yi-\bar{Y})^2$ είναι ένα συνολικό μέτρο σφάλματος όταν για την εκτίμηση των τιμών Yi, i=1,...,n. Χρησιμοποιείται ο αριθμητικός τους μέσος.Ο όρος $SSE=\Sigma(Yi-\hat{Y}i)^2=\Sigma(Yi-\hat{b}o-\hat{b}_1 \cdot xi)^2$ είναι το αντίστοιχο μέτρο όταν χρησιμοποιείται το απλό γραμμικό μοντέλο παλινδρόμησης. Η διαφορά SST-SSE είναι η μείωση του σφάλματος εκτίμησης που επιτυγχάνεται με το μοντέλο παλινδρόμησης. Η **ποσοστιαία μείωση του σφάλματος** δηλαδή η

$$PRE = \frac{SST-SSE}{SST}$$

ονομάζεται **συντελεστής προσδιορισμού** και συμβολίζεται με R^2
Δηλαδή έχουμε:

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

Σημειώνεται ότι ο όρος SST είναι ένα μέτρο της συνολικής μεταβλητότητας των παρατηρήσεων Yi γύρω από τη μέση τιμή τους ενώ ο όρος SSE ένα μέτρο τησυνολικής μεταβλητότητας των Yi γύρω από την ευθεία ελαχίστων τετραγώνων. Επομένως, ο συντελεστής R^2 ερμηνεύεται, ισοδύναμα, ως το ποσοστό της μεταβλητότητας των Yi γύρω από τη μέση τιμή τους το οποίο ερμηνεύεται από το από το εκτιμηθέν μοντέλο παλινδρόμησης.
Από τον ορισμό αυτόν προκύπτει ότι ισχύει.

$$0 \leq R^2 \leq 1$$

Οσο ο όρος SSR είναι μεγαλύτερος από τον SSE, ή, ισοδύναμα, όσο πιο κοντά στη μονάδα είναι ο συντελεστής R^2 , τόσο μεγαλύτερη είναι η ερμηνευτική ικανότητα του μοντέλου παλινδρόμησης.

Παράδειγμα

Για την αξία των πωλήσεων εκτιμήσαμε το μοντέλο.

$$\hat{Y} = 205 + 5{,}66$$

Εχουμε επίσης υπολογίσει τα ακόλουθα.

$$SSE = \Sigma \hat{\varepsilon}_i^2 = 130520$$
$$SST = \Sigma (Yi - \bar{Y})^2 = \Sigma Yi^2 - n\bar{Y}^2$$
$$= 451102000 - 18(1561{,}1)^2$$
$$= 1243602{,}22$$

οπότε υπολογίζουμε το συντελεστή προσδιορισμού ως εξής :

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{130520}{1243602{,}22} = 0{,}895$$

Δηλαδή το απλό μοντέλο παλινδρόμησης στο οποίο παίρνεται υπόψη η X στην ερμηνεία της συμπεριφοράς των παρατηρήσεων Yi επιτυγχάνεται, σε σύγκριση με το βασικό μοντέλο μείωση του σφάλματος εκτίμησης ίση με 89,5%.

2.8 Προβλέψεις

Το μοντέλο που εκτιμήσαμε μπορεί να χρησιμοποιηθεί για προβλέψεις Διακρίνουμε δυο περιπτώσεις :

i) Πρόβλεψη της μέσης τιμής της εξαρτημένης όταν x = xn + 1

Αν στην ευθεία ελαχίστων τετραγώνων αντικαταστήσουμε την τιμή X= Xn + 1 θα πάρουμε την

$$\hat{Y}n + 1 = \hat{b}o + \hat{b}1\, Xn + 1$$

που είναι μια εκτίμηση της bo + b1 Xn + 1 και επομένως είναι μια σημειακή εκτίμηση της μέσης τιμής της εξαρτημένης για X = Xn + 1. Λόγω του σφάλματος δειγματοληψίας η εκτιμούμενη μέση τιμή θα διαφέρει από την πραγματική, με πιθανότητα ίση με 1. Σε

επαναληπτικές όμως προβλέψεις, η μέση τιμή αυτών των διαφορών θα ισούτε με μηδέν διότι

$$E(\hat{Y}n+1) = E(\hat{b}o) + E(\hat{b}1)\,Xn+1$$
$$= bo + b1\,Xn + 1.$$

Οταν κάνουμε μια πρόβλεψη, θα εκτιμήσουμε τη διακύμανση της εκτιμούμενης μέσης τιμής με τον τύπο

$$S^2\left(\frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right) \qquad (2.8.1)$$

όπου $S^2 = \frac{\Sigma \hat{\varepsilon}_i^2}{n-2}$

Αποδεικνύεται ότι, κάτω από την υπόθεση της κανονικότητας των διαταρακτικών όρων, το διάστημα το οποίο, με πιθανότητα 1 - α, θα περιέχει την
E (Yn+1) είναι το

$$\hat{Y}_{n+1} \pm t_{n-2,\alpha/2}\, S \sqrt{\frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}} \qquad (2.8.2)$$

όπου $t_{n-2,\alpha/2}$ η τιμή της κατανομής t - student με ν = n -2 βαθμού ελευθερίας για την οποία ισχύει $P(t_{n-2} > t_{n-2,\alpha/2})$

Το διάστημα αυτό ερμηνεύεται ως εξής : Εστω ότι έχουμε επαναληπτικές λήψεις δειγμάτων ή ζευγών (Xi, Yi) στις οποίες οι τιμές των Xi επιλέγονται ίδιες όπως στο δείγμα μας και από κάθε δείγμα εκτιμούμε την ευθεία παλινδρόμησης. Τότε το (1 - α)% των διαστημάτων τα οποία θα εκτιμήσουμε για X = Xn + 1 θα περιέχουν την E (Yn + 1). Οταν υπολογίζουμε ένα μόνο διάστημα τότε η πιθανότητα να περιέχει την E (Yn + 1) ισούτε με (1 - α).

ii) Πρόβλεψη μιας μεμονομένης τιμής της Y όταν X = Xn + 1

Στην περίπτωση αυτή θα πρέπει να προβλέψουμε και μια τιμή για τον En + 1. Επειδή E (ε + 1) = 0, προβλέπουμε την τιμή μηδέν, οπότε προκύπτει και στην περίπτωση αυτή η ίδια σημειακή εκτίμηση. Η αβεβαιότητα όμως στην πρόβλεψη είναι μεγαλύτερη απ' ότι στην προηγούμενη περίπτωση λόγω της επίδρασης και του διαταρακτικού όρου E . Ετσι θα εκτιμήσουμε τη διακύμανση της προβλεπόμενης τιμής για την Yn + 1 με το άθροισμα της διακύμανσης της μέσης τιμής, συν τη διακύμανση του διαταρακτικού όρου E δηλαδή θα είναι ίση με

$$S^2 + S^2 \frac{1}{n} + \left[\frac{(X_{n+1} - \bar{X})^2}{\Sigma (X_i - \bar{X})^2}\right] = S^2 \left[1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\Sigma (X_i - \bar{X})^2}\right] \quad (2.8.3)$$

Όταν η κατανομή των διαταρακτικών όρων είναι κανονική, τότε το διάστημα στο οποίο θα βρίσκεται η y_{n+1} με πιθανότητα (1-α), είναι το εξής :

$$\hat{Y}_{n+1} \pm t_{n-2,\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\Sigma (X_i - \bar{X})^2}} \quad (2.8.4)$$

Η ερμηνεία του διαστήματος αυτού είναι ανάλογη με την ερμηνεία του διαστήματος (2.8.2). Από τις (2.8.2) και (2.8.4) προκύπτουν μερικές γενικές παρατηρήσεις για το εύρος των αντιστοίχων διαστημάτων εμπιστοσύνης :

- Οσο πιο μεγάλο είναι το μέγεθος του δείγματος η, τόσο πιο μικρό το εύρος του αντιστοίχου διαστήματος εμπιστοσύνης, επομένως τόσο μικρότερη η αβεβαιότητα για την αντίστοιχη εκτίμηση, ceteris paribus.

- Οσο μεγαλύτερη η ποσότητα $(x_{n+1} - \bar{x})^2$ δηλαδή όσο μεγαλύτερη η απόσταση του x_{n+1} από το δειγματικό μέσο της ερμηνευτικής μεταβλητής τόσο μεγαλύτερο το εύρος του αντιστοίχου διαστήματος, ceteris paribus.

• Οσο μεγαλύτερο το πληροφοριακό περιεχόμενο του δείγματος για την εκτίμηση του μοντέλου, δηλαδή ο όρος $\sum_{i=1}^{n} (x_i - \bar{x})^2$ τόσο πιο στενά τα όρια του αντιστοίχου διαστήματος εμπιστοσύνης, ceteris paribus.

• Τέλος, όσο πιο μεγάλη η τιμή του μέσου τετραγωνικού σφάλματος της εκτίμησης s^2 δηλαδή όσο πιο μεγάλη είναι η μεταβλητότητα των παρατηρήσεων γύρω από την εκτιμούμενη ευθεία παλινδρόμησης, τόσο πιο ευρύ είναι το αντίστοιχο διάστημα εμπιστοσύνης.

Παράδειγμα

Εστω ότι θέλουμε να χρησιμοποιήσουμε το μοντέλο που εκτιμήσαμε προηγουμένως για να προβλέψουμε τη μηνιαία αξία των πωλήσεων της επιχείρησης όταν αυτή δαπανά μηνιαίως για διαφήμιση 250 χιλ. δρχ. Η σημειακή εκτίμηση των πωλήσεων για x_{n+1} =250 ισούται με :

$$\hat{y}_{n+1} = 204,9 + 5,675\, x_{n+1}$$
$$= 204,9 + (5,675)250 = 1623,65 \text{ χιλ. δρχ.}$$

Για να υπολογίσουμε διαστήματα εμπιστοσύνης για την αξία των πωλήσεων και τη μέση αξία των πωλήσεων όταν y_{n+1} =250 χρειαζόμαστε τις τυπικές αποκλίσεις των αντιστοίχων εκτιμητών.
Προηγουμένως υπολογίσαμε :

$$\bar{X} = 239,12 \quad \Sigma (x_i - \bar{x})^2 = 34541 \text{ και } S = 90,32$$

οπότε έχουμε

$$S\sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}} = 90,32\sqrt{\frac{1}{18} + \frac{(250 - 239,12)^2}{34541}} = 21,94$$

και

$$S\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}} = 90,32 + 21,94 = 112,26$$

Για επίπεδο εμπιστοσύνης (1- α) = 0,95 βρίσκουμε από τον πίνακα της κατανομής t- student την τιμή $t_{n-2,\alpha/2} = t_{16(0,025)} = 2,12$ οπότε υπολογίζουμε τα ακόλουθα διαστήματα εμπιστοσύνης.

Για τη μέση αξία των πωλήσεων όταν Xn+1= 250

$$1623,65-(2,12)(21,94) \leq E(Y_{n+1}) \leq 1623,65+(2,12)(21,94)$$

$$1577,14 \leq E(Y_{n+1}) \leq 1670,16$$

Το εύρος του διαστήματος αυτού ισούτε με 93,02

Για την αξία των πωλήσεων όταν X n+1= 250

$$1623,65-(2,12)(112,26) \leq Y_{n+1} \leq 1623,65+(2,12)(112,26)$$

$$1385,66 \leq Y_{n+1} \leq 1861,64$$

Το εύρος του διαστήματος αυτού ισούτε με 476 περίπου

Παρατηρούμε ότι όσο απομακρυνόμαστε από τη μέση τιμή των δεδομένων μας τόσο αυξάνει το εύρος των αντίστοιχων διαστημάτων δηλαδή η αβεβαιότητα της αντίστοιχης εκτίμησης και επομένως θα πρέπει να αποφεύγουμε τις προβλέψεις για τιμές της X, οι οποίες είναι μακριά από τα όρια των δεδομένων μας.

2.9 Ανάλυση διακυμάνσεων

Σε κάθε άθροισμα τετραγώνων αντιστοιχεί ένας αριθμός που ονομάζεται βαθμοί ελευθερίας και δείχνει πόσα ανεξάρτητα μέρη πληροφορίας χρησιμοποιούμε από τις η ανεξάρτητες παρατηρήσεις Y1,Y2,....,Yn για να το υπολογίσουμε. Ετσι π.χ για το SST χρησιμοποιούμε τις (n-1) ανεξάρτητες διαφορές (Yi-$\bar{Y}$) οπότε στο SST αντιστοιχούν n-1 βαθμοί ελευθερίας. Αντίστοιχα, επειδή

$$SSR=\Sigma(\hat{Y}i-\bar{Y})^2=\Sigma(\hat{b}o+\hat{b}1x1-\hat{b}o-\hat{b}1\bar{x})^2$$
$$=\Sigma[\hat{b}1(xi-\bar{x})]^2$$
$$=\hat{b}1^2\Sigma(xi-\bar{x})^2$$

δηλαδή μπορούμε να υπολογίσουμε τα SSR από μια μονάχα συνάρτηση των Yi έχουμε 1 βαθμό ελευθερίας. Επομένως, με απλή αφαίρεση, στα SSE αντιστοιχούν (n-1)-1=n-2 βαθμοί ελευθερίας. Αυτό αντικατοπτρίζει το γεγονός ότι τα κατάλοιπα ε1 προκύπτουν από ένα εκτιμούμενο μοντέλο με 2 παραμέτρους.

Ολα τα προγράμματα υπολογιστή για τη γραμμική παλινδρόμηση δίνουν μαζί με τους υπόλοιπους υπολογισμούς έναν πίνακα ανάλυσης διακυμάνσεως (Analysis of variance of Anova+able)
του οποίου η τυπική μορφή στο απλό μοντέλο είναι η εξής :

SOURCE OF VARIATION	SUM OF SQUARES	D.F	MEAN OF SQUARES
REGRESSION	$\Sigma(\hat{Y}_i - \bar{Y})^2$	1	SSR/1
ERROR	$\Sigma(\hat{Y}_i - Y_i)^2$	n-2	SSE/(n-2)
TOTAL	$\Sigma(Y_i - \bar{Y})^2$	n-1	SST/(n-1)

Ετσι στο παράδειγμα της αξίας των πωλήσεων ο πίνακας ANOVA θα δοθεί ως εξής :

SOURCE OF VARIATION	SUM OF SQUARES	D.F	MEAN SQUARE
REGRESSION	1112458,06	1	1112458,06
ERROR	130519,7	16	8157,48
TOTAL	1242977,76	17	73116,34

2.10 Το μοντέλο χωρίς σταθερό όρο.

Όταν θέλουμε η συνάρτηση παλινδρόμησης να περνά από την αρχή των αξόνων τότε το μοντέλο εξειδικεύεται χωρίς σταθερό όρο. Σ' ένα τέτοιο μοντέλο μπορεί να καταλήξουμε :

I. **Υποχρεωτικά** : Οταν τα δεδομένα υπαγορεύουν κάποιον μετασχηματισμό που απαλοίφει το σταθερό όρο. Ετσι π.χ όταν η διακύμανση των διαταρακτικών όρων φαίνεται να αυξάνει με τις τιμές xi τότε στο μοντέλο

$$Yi = bo + b1xi + \varepsilon 1$$

θα δοκιμάσουμε τον ακόλουθο μετασχηματισμό :

$$\frac{Yi}{\sqrt{xi}} = \frac{bo}{\sqrt{xi}} + b1\sqrt{xi} + vi$$

όπου vi =εi /$\sqrt{xi}$. Το μετασχηματισμένο αυτό μοντέλο δεν έχει σταθερό όρο αλλά δυο ερμηνευτικές μεταβλητές, τις X1 =1/$\sqrt{\chi}$ και $X_2=\sqrt{X}$

ii. **Μετά από ανάλυση** : Η σχετική θεωρία μπορεί να υπαγορεύει ότι στη μηδενική τιμή της ερμηνευτικής μεταβλητής αντιστοιχεί μηδενική τιμή της εξαρτημένης. Ετσι π.χ ας θεωρήσουμε την Κενσιανή συνάρτηση κατανάλωσης :

$$E(Y) = bo + b1x$$

όπου Υπ κατανάλωση όταν το εισόδημα ισούτε με x. Η Οριακή Ροπή προς κατανάλωση (ΟΡΚ) ισούτε με b1 ενώ η Μέση Ροπή προς Κατανάλωση (ΜΡΚ) ισούτε με :

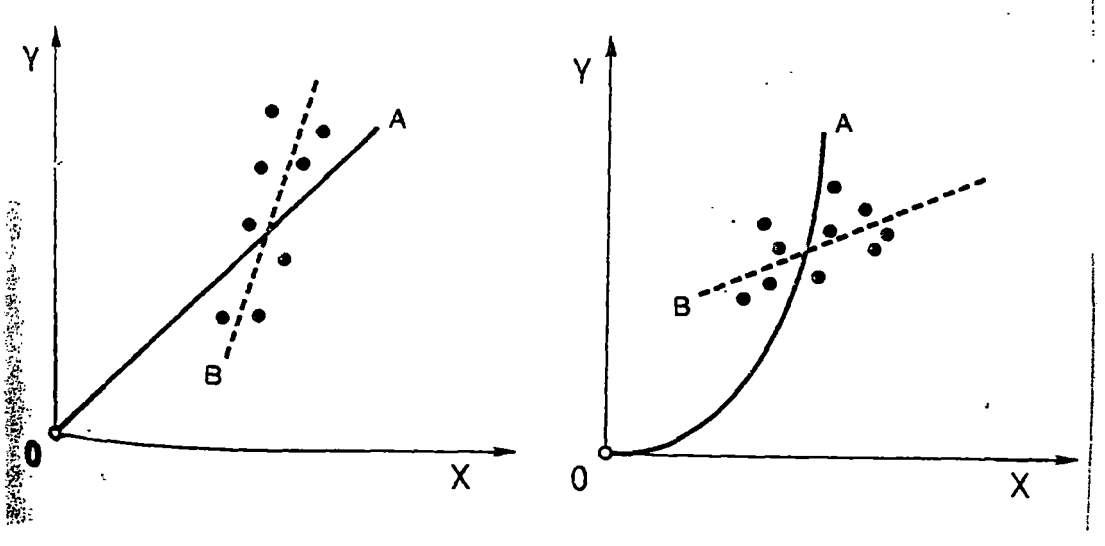
$$\frac{E(Y)}{X} = \frac{bo}{X} + b1$$

Εστω ότι έχουμε λόγους να πιστεύουμε ότι ΟΡΚ=ΜΡΚ δηλαδή ότι

$$b1 = \frac{bo}{X} + b1$$

Αυτό ισχύει μόνο όταν bo=0. Συμπεραίνουμε ότι αν έχουμε λόγο να πιστεύουμε ότι ΟΡΚ=ΜΡΚ τότε θα θεωρήσουμε το Κενσιανικό μοντέλο κατανάλωσης χωρίς σταθερό όρο.

Η σχετική θεωρία πάντως μπορεί να μη ληφθεί υπόψη αν η συμπεριφορά των δεδομένων είναι διαφορετική. Στο σχήμα που ακολουθεί βλέπουμε δυο περιπτώσεις στις οποίες παρόλο που το θεωρητικό μοντέλοι παλινδρόμησης (γραμμή Α) περνά από την αρχή των αξόνων δηλαδή το σημείο (Χ=0, Υ=0) εν τούτοις για τα δεδομένα που διαθέτουμε, η ευθεία με μη μηδενικό σταθερό όρο (γραμμή Β) έχει καλύτερη προσαρμογή.



Τέλος, η επιλογή ανάμεσα σε ένα μοντέλο με και ένα χωρίς σταθερό όρο μπορεί να γίνει με στατιστικά κριτήρια όπως είναι ο t- έλεγχος της υπόθεσης bo=0.
Θα πρέπει πάντως να τονιστεί ότι ακόμη και αν οι σχετικοί στατιστικοί έλεγχοι απορίψουν το σταθερό όρο δεν κρίνεται πάντα σκόπιμη η αφαίρεση του από το μοντέλο. Επισημαίνεται η σχετική παρατήρηση του A. Spanos (1986) : << ο σταθερός όρος θα πρέπει πάντα να περιλαμβάνεται σε ένα μοντέλο παλινδρόμησης για μεταβλητές σε αναλογική κλίμακα.>>
Στη συνέχεια της ανάλυσης επισημαίνονται τα κυριότερα σημεία στα οποία το μοντέλο χωρίς σταθερό όρο διαφέρει από το αντίστοιχο μοντέλο με το σταθερό όρο. Η αναφορά γίνεται στο απλό μοντέλο αλλά η γενίκευση στο πολύ μεταβλητό είναι άμεση.

Η μέθοδος των ελαχίστων τετραγώνων στο μοντέλο χωρίς σταθερό όρο.

Η μέθοδος των ελαχίστων τετραγώνων στο μοντέλο

$$Yi = b1x1 + \varepsilon 1 \qquad i = 1,....,n$$

ελαχιστοποιεί τη συνάρτηση

$$S = S(b1) = \sum_{i=1}^{n} (Yi - b1xi)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

Επειδή

$$\frac{dS}{db_1} = -2 (\Sigma Yi - b1x1)\, xi$$

η εκτίμηση των ελαχίστων τετραγώνων του b1 είναι αυτή που ικανοποιεί την εξίσωση

$$-2 (\Sigma Yi - \hat{b}1Xi)\, Xi = 0$$
$$\Leftrightarrow \Sigma YiXi - \hat{b}1\Sigma Xi^2 = 0$$
$$\Leftrightarrow \hat{b}1 = \frac{\Sigma XiYi}{\Sigma x_i^2} \qquad (2.10.1)$$

Παράδειγμα

Μετρήσαμε την ηλικία X σε λεπτά, ενός ψεκασμού αεροσόλ και την παρατηρούμενη διασπορά του Y σ' αυτόν τον χρόνο. Η διασπορά μετρήθηκε ως το αντίστροφο του αριθμού των σωματιδίων ανά μονάδα όγκου. Σε 9 παρατηρήσεις (Xi, Yi) πήραμε τα ακόλουθα αποτελέσματα :

Xi 8 22 35 40 57 73 78 87 98
Yi 6,16 9,88 14,35 24,06 30,34 32,17 42,18 43,23 48,76

(Τα δεδομένα δίνονται από τους Box - Hunter - Hunter, 1978)

θα εκτιμήσουμε ένα απλό μοντέλο παλινδρόμησης χωρίς σταθερό όρο. Υπολογίζουμε :

$\Sigma XiYi = 17638,61$ $\Sigma xi^2 = 35208$ $\Sigma Yi^2 = 8901,31$

οπότε

$$\hat{b}1 = \frac{\Sigma xi\ Yi}{\Sigma x_i^2} = \frac{17638,61}{35208} = 0,501$$

και το μοντέλο που εκτιμήσαμε είναι το

$$\hat{Y} = 0,501\ X$$

Δηλαδή, εκτιμούμε ότι κάθε επιπλέον λεπτό που περνά, αυξάνει τη διασπορά του ψεκασμού κατά μισό περίπου.

Τα κατάλοιπα της εκτίμησης

Στο μοντέλο χωρίς σταθερό όρο τα κατάλοιπα της εκτίμησης $\hat{e}i = Yi - \hat{Y}i = Yi - \hat{b}1xi$, i= 1,....,n ικανοποιούν τις σχέσεις :

$$\Sigma \hat{Y}i\hat{e}i = 0$$
$$\Sigma Yi\hat{e}i = 0$$

αλλά γενικά

$$\Sigma\hat{\varepsilon}_i \neq 0$$

Παράδειγμα

Υπολογίζουμε τα κατάλοιπα της εκτίμησης στο μοντέλο του προηγούμενου παραδείγματος :

Xi	Yi	Yi=0,501Xi	ε_i=Yi-Yi
8	6,16	4,0079	2,1521
22	9,88	11,0216	-1,1416
35	14,35	17,5344	-3,1844
40	24,06	20,0393	4,0207
57	30,34	28,5660	1,7840
73	32,17	36,5718	-4,4018
78	42,18	39,0767	3,1033
87	43,23	43,5855	-0,3555
98	48,76	49,0963	-0,3363
Άθροισμα 498	251,13		1,6405

Ανάλυση Διακυμάνσεως

Γενικά

$$\Sigma (Yi - \hat{Y}) \neq \Sigma (\hat{Yi} - \bar{Y})^2 + \Sigma\hat{\varepsilon}i^2$$
$$\text{ή } SST \neq SSR + SSE$$

ενώ ισχύει

$$\Sigma Yi^2 = \Sigma\hat{Yi}^2 + \Sigma\hat{\varepsilon}i^2$$
$$(n) \quad (k) \quad (n-k)$$

και οι αριθμοί στις παρενθέσεις είναι οι βαθμοί ελευθερίας του κάθε όρου. Για το απλό μοντέλο κ=1. Από τη σχέση αυτή προκύπτει ότι γενικά

$$\frac{\Sigma (\hat{Yi} - \bar{Y})^2}{\Sigma (Yi - \bar{Y})^2} \neq 1 - \frac{\Sigma\hat{\varepsilon}i^2}{\Sigma (Yi - \bar{Y})^2}$$

Ο συντελεστής προσδιορισμού

Μπορούμε να πάρουμε για το συντελεστή προσδιορισμού R^2 μια τιμή που να ικανοποιεί τη σχέση $0 \leq R^2 \leq 1$ μόνο από το τύπο

$$R^2 = 1 - \frac{\Sigma\hat{\varepsilon}_i^2}{\Sigma Yi^2} \qquad (2.10.2)$$

Στα περισσότερα προγράμματα Η/Υ, πάντως (μεταξύ αυτών και το RATS) ο συντελεστής R^2 υπολογίζεται σε όλες τις περιπτώσεις από τον τύπο

$$R^2 = 1 - \frac{\Sigma \hat{\varepsilon}i^2}{\Sigma (Yi - \bar{Y})^2} \qquad (2.10.3)$$

Ετσι, στα μοντέλα χωρίς σταθερό όρο, μπορεί να πάρουμε αρνητική τιμή για τον R^2 . Ο αρνητικός συντελεστής προσδιορισμού είναι φυσικά δείκτης κακής προσαρμογής του μοντέλου. Θα πρέπει να σημειωθεί ότι αυτό ισχύει για οποιονδήποτε περιορισμό στις παραμέτρους του μοντέλου, ένας λανθασμένος περιορισμός που επιβάλλεται στις παραμέτρους μπορεί να έχει ως αποτέλεσμα το SSE να γίνει μεγαλύτερο από το SST.

Η διακύμανση του $\hat{b}1$

Αποδεικνύεται ότι η διακύμανση του b1 δίνεται από τον τύπο :

$$\text{Var}(\hat{b}1) = \frac{\sigma^2}{\Sigma x_i^2} < \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}$$

Επομένως στο μοντέλο χωρίς σταθερό όρο, η εκτίμηση του $\hat{b}1$ γίνεται, γενικά, με μεγαλύτερη ακρίβεια απ' ότι στο μοντέλο με σταθερό όρο.
Ο αμερόληπτος εκτιμητής της σ^2 για το απλό μοντέλο χωρίς σταθερό όρο υπολογίζεται από τον τύπο

$$S^2 = \frac{\Sigma \hat{\varepsilon}_i^2}{n-1}$$

Αντίστοιχα, ο αμερόληπτος εκτιμητής της Var $(\hat{b}1)$ δίνεται από τον τύπο

$$S^2_{\hat{b}_1} = \frac{S^2}{\Sigma x_i^2}$$

και ισχύει :

$$\frac{\hat{b}_1 - b_1}{S_{\hat{b}_1}} \sim t_{n-1}$$

Ο έλεγχος των υποθέσεων και το διάστημα εμπιστοσύνης για το b1 γίνεται όπως και στο μοντέλο με σταθερό όρο.

Παράδειγμα

Εκτιμούμε την διακύμανση του $\hat{b}1$ στο παράδειγμα της διασποράς του αεροσόλ ως εξής :

$$S^2_{\hat{b}_1} = \frac{S^2}{\Sigma x_i^2} = \frac{8,08}{35208} = 0,000229$$

και το τυπικό σφάλμα της εκτίμησης ως εξής :

$$S_{\hat{b}_1} = \sqrt{0,000229} = 0,015$$

Ο έλεγχος για μηδενική τιμή του b1 θα γίνει ως εξής :

Ho : b1 = 0
Hε : b1 ≠ 0

Επειδή
$t\pi = \frac{\hat{b}1}{\hat{s}_{\hat{b}1}} = \frac{0,501}{0,015} = 33,4 > t_{8(0,025)} = 2,306$
απορρίπτουμε την Ho. Το διάστημα το οποίομε πιθανότητα 95% θα περιέχει την τιμή του b1 είναι το εξής :

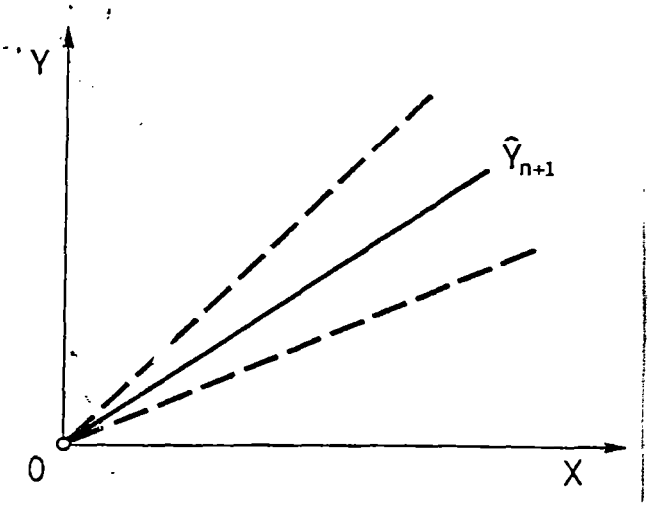
$\hat{b}1 \pm (0,01515)(2,306)$
ή $0,501 \pm (0,01515)(2,306)$
ή $0,466 \leq b1 \leq 0,536$

Προβλέψεις και αντίστοιχες λωρίδες εμπιστοσύνης

Στην τιμή X_{n+1} προβλέπουμε την μέση τιμή E (Yn+1) της εξαρτημένης με την $\hat{Y}n+1 = \hat{b}1\ Xn+1$. Το διάστημα στο οποίο θα βρίσκεται η E(Yn+1) με επίπεδο εμπιστοσύνης (1 - α) είναι το :

$$\hat{Y}_{n+1} \pm t_{n-1,\alpha/2}\, S \sqrt{\frac{X^2_{n+1}}{\sum X_i^2}}$$

Οι λωρίδες εμπιστοσύνης τείνουν στο μηδέν όσο πλησιάζουμε στην αρχή των αξόνων όπως φαίνεται και στο επόμενο σχήμα. Αυτό είναι σύμφωνο με την υπόθεση ότι με πιθανότητα ίση με 1 η E(Yi) ισούτε με μηδέν όταν Xi=0



Λωρίδες εμπιστοσύνης του απλού μοντέλου χωρίς σταθερό όρο.

ΚΕΦΑΛΑΙΟ ΤΡΙΤΟ

Συσχέτιση

3.1 Εισαγωγή

Στην ανάλυση συσχέτισης μας ενδειαφέρει να δούμε αν δυο τυχαίες μεταβλητές Χ και Υ συνδέονται με μια σχέση και πόσο στενή είναι αυτή η σχέση. Η ανάλυση συσχέτισης εφαρμόζεται είτε αυτοτελώς, είτε ως ένα προκαταρκτικό στάδιο πριν από την ανάλυση παλινδρόμησης. Ετσι, όταν έχουμε πολλές υποψήφιες ερμηνευτικές μεταβλητές η ανάλυση συσχέτισης μπορεί, καταρχήν, να δείξει ποιές απ' αυτές συνδέονται περισσότερο με την εξαρτημένη μεταβλητή και θα πρέπει να περιληφθούν στο μοντέλο.

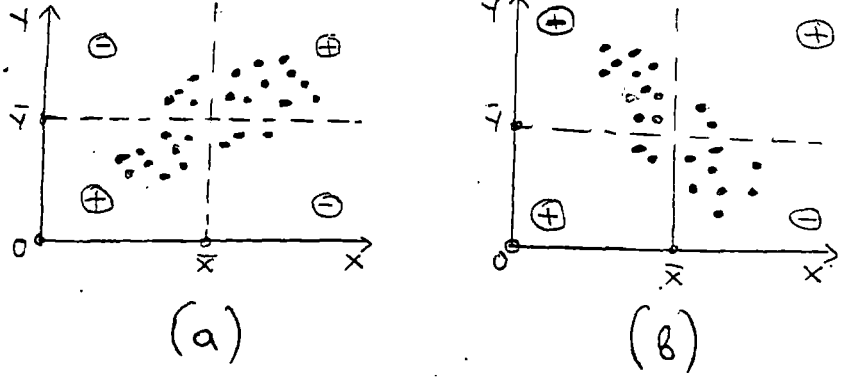
Οι συντελεστές συσχέτισης συνδέονται στενά με τους συντελεστές παλινδρόμησης όμως, το αναλυτικό πλαίσιο, συνεπώς και το πεδίο εφαρμογών των δυο μεθόδων, είναι διαφορετικό. Στην ανάλυση παλινδρόμησης έχουμε μια τυχαία μεταβλητή Υ της οποίας τη συμπεριφορά θέλουμε να ερμηνεύσουμε ή να προβλέψουμε με τη συμπεριφορά της μεταβλητής Χ.
Η Χ είναι ελεγχόμενη δηλαδή οι τιμές της έχουν επιλεγεί με βάση ένα σχέδιο πειράματος. Η Χ μπορεί να είναι και τυχαία μεταβλητή αλλά και στην περίπτωση αυτή αντιμετωπίζεται ως ελεγχόμενη με την έννοια ότι δεν γίνεται καμιά αναφορά στην πιθανοκρατική συμπεριφορά της- οι παρατηρήσεις Χi θεωρούνται ως οι τιμές της Χ στις οποίες επιλέξαμε να παρατηρήσουμε την Υ και τα αποτελέσματα ερμηνεύονται υπό συνθήκη των Χi.
Στην ανάλυση συσχέτισης οι δυο μεταβλητές Χ και Υ είναι τυχαίες και αντιμετωπίζονται συμμετρικά. Εξάλλου, όταν οι Χ και Υ είναι μετρήσιμες μόνο στην ταξική κλίμακα μέτρησης, συνήθως περιοριζόμαστε στον υπολογισμό του αντίστοιχου συντελεστή συσχέτισης. Ετσι δικαιολογείται και η μάλλον υπερβολική αλλά όχι αβάσιμη δήλωση του Tukey (1954) ότι οι συντελεστές συσχέτισης είναι χρήσιμοι σε δυο και μόνον περιπτώσεις : όταν είναι συντελεστές παλινδρόμησης ή όταν η μέτρηση μιας ή περισσοτέρων μεταβλητών δεν μπορεί να γίνει παρά μόνον σε ταξική κλίμακα. Σημειώνεται πάντως ότι τα τελευταία χρόνια έχουν αναπτυχθεί τεχνικές της ανάλυσης παλινδρόμησης που είναι κατάλληλες και για τις μεταβλητές αυτές και είναι γνωστές με τον όρο **στιβαρή παλινδρόμηση**.

3.2 Η συνδιακύμανση

Εστω ότι σε κάθε στοιχείο το οποίο παίρνεται από τον πληθυσμό, με απλή τυχαία δειγματοληψία παρατηρείται η μεταβλητή X και συγχρόνως η μεταβλητή Y και διαθέτουμε τα η ζεύγη παρατηρήσεων (X1, Y1), ..., (Xn, Yn). Κάθε ζεύγος (X1,Y1), i =1,..., n αποτελεί μια διμεταβλητή παρατήρησηκαι παριστάνεται με ένα σημείο στο επίπεδο XY. Η θέση των παρατηρήσεων Xi στον άξονα των X δίνεται από τον αριθμητικό τους μέσο $\bar{X} = \Sigma Xi / n$ ενώ η διασπορά τους από τη διακύμανση τους $S^2x = \Sigma (Xi - \bar{X})^2 / (n - 1)$. Αντίστοιχα, η θέση και διασπορά των Yi δίνεται από τα $\bar{Y} = \Sigma Yi / n$ και $S^2y = \Sigma (Yi - \bar{Y})^2 / n - 1$. Οι παράμετροι $\bar{X}$, $\bar{Y}$, S^2x , S^2y, δεν μας δίνουν καμιά πληροφορία για την κοινή κατανομή των X και Y δηλαδή για τον τρόπο που τα σημεία (Xi, Yi) κατανέμονται στο επίπεδο XY. Για το σκοπό αυτό υπολογίζουμε τη συνδιακύμανση η οποία ορίζεται ως εξής :

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (Xi - \bar{X})(Yi - \bar{Y})$$

και μετρά την γραμμική συμμεταβολή δηλαδή την τάση των σημείων (Xi, Yi) να συγκεντρώνονται κατά μήκος μιας ευθείας. Ανάλογα με το αν η συμμεταβολή των παρατηρήσεων X και Y είναι ευθεία ή αντίστροφη, το πρόσημο της συνδιακύμανσης είναι θετικό ή αρνητικό. Ειδικότερα παρατηρούμε τα εξής :
Αν στο ζεύγος (Xi,Yi) η παρατήρηση Xi είναι μεγαλύτερη (μικρότερη) από τον αριθμητικό μέσο X και η Yi μεγαλύτερη (μικρότερη) από τον $\bar{Y}$ τότε το γινόμενο (Xi - $\bar{X}$) (Yi - $\bar{Y}$) είναι θετικό. Αν η παρατήρηση της μιας μεταβλητής είναι μεγαλύτερη από το μέσο της, ενώ της άλλης είναι μικρότερη τότε το γινόμενο (Xi - $\bar{X}$) (Yi - $\bar{Y}$) είναι αρνητικό. Οταν οι X και Y τείνουν να συμμεταβάλλονται ευθέως τότε τα θετικά γινόμενα (Xi - $\bar{X}$) (Yi - $\bar{Y}$) είναι περισσότερα από τα αρνητικά και η συνδιακύμανση είναι θετική, όπως στο σχήμα (α) που ακολουθεί. Αντίστοιχα, όταν οι X και Y τείνουν να συμμεταβάλλονται αντιστρόφως τότε τα αρνητικά γινόμενα (Xi - $\bar{X}$) (Yi - $\bar{Y}$) είναι περισσότερα από τα θετικά και η συνδιακύμανση αρνητική όπως στο σχήμα (β).



Η μηδενική συνδιακύμανση δείχνει αν οι X και Y δεν συμμεταβάλλονται ούτε ευθέως ούτε αντιστρόφως. Είναι προφανές ότι όταν Xi = Yi, ∀i, τότε :

$$S_{xy} = \frac{1}{1-n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = S_x^2$$

δηλαδή, η συνδιακύμανση των παρατηρήσεων Xi είναι απλώς η διακύμανση τους.

3.3 Ο συντελεστής συσχέτισης

Η συνδιακύμανση εκφράζεται σε (μονάδες X) (μονάδες Y). Διαιρώντας την με το γινόμενο Sx SY προκύπτει ένα μέτρο γραμμικής συμμεταβολής που είναι απαλλαγμένο από μονάδες μέτρησης και ονομάζεται **συντελεστής συσχέτισης του Pearson ή του γινομένου των ροπών.** Ο συντελεστής αυτός συμβολίζεται με r_{xy} οπότε έχουμε :

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$$

$$r_{xy} = \frac{\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2} \sqrt{\Sigma (y_i - \bar{y})^2}} = \frac{\Sigma x_i y_i - n \bar{x} \bar{y}}{\sqrt{\Sigma x_i^2 - n \bar{x}^2} \sqrt{\Sigma y_i^2 - n \cdot \bar{y}^2}}$$

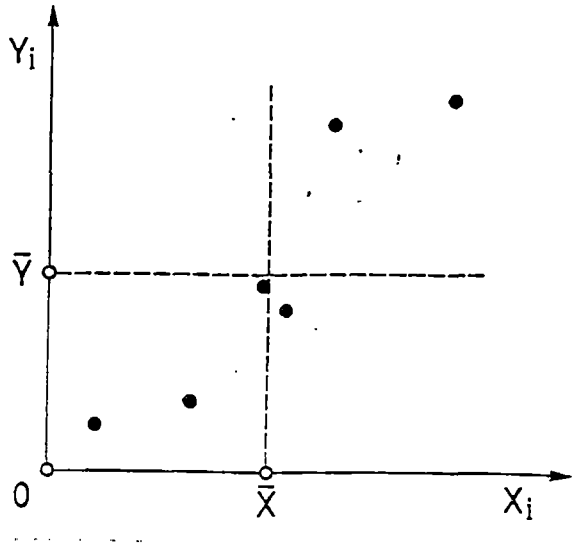
και ο συντελεστής Σ είναι για τιμές του i από 1 ως n.

Ο συντελεστής r_{xy} είναι καθαρός αριθμός επιτρέποντας έτσι τις συγκρίσεις, ενώ παράλληλα έχει ενδιαφέρουσες στατιστικές ιδιότητες. Γι' αυτό χρησιμοποιείται πολύ πιο συχνά από την συνδιακύμανση. Τις βασικές ιδιότητες του r_{xy} θα δούμε στο επόμενο τμήμα. Προηγουμένως θα δούμε την πληροφόρηση που δίνουν οι συντελεστές S_{xy} και r_{xy} για τη συμμεταβολή η διμεταβλητών παρατηρήσεων με το ακόλουθο.

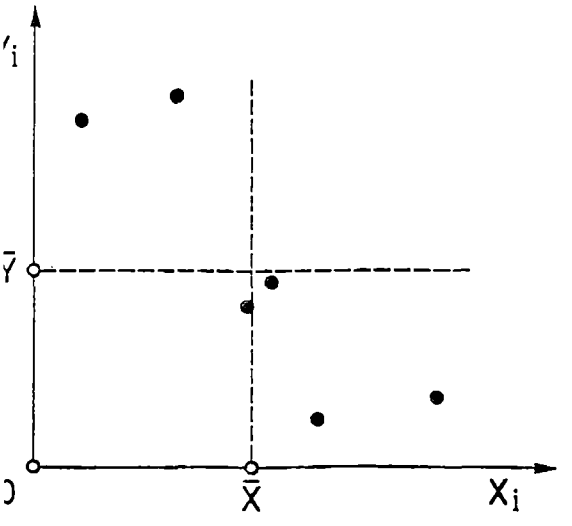
Παράδειγμα

Δίνεται η γραφική παράσταση τριών διμεταβλητών δειγμάτων με n=6. Οι παρατηρήσεις (Xi, Yi) σημειώνονται στον άξονα X και Y αντίστοιχα,

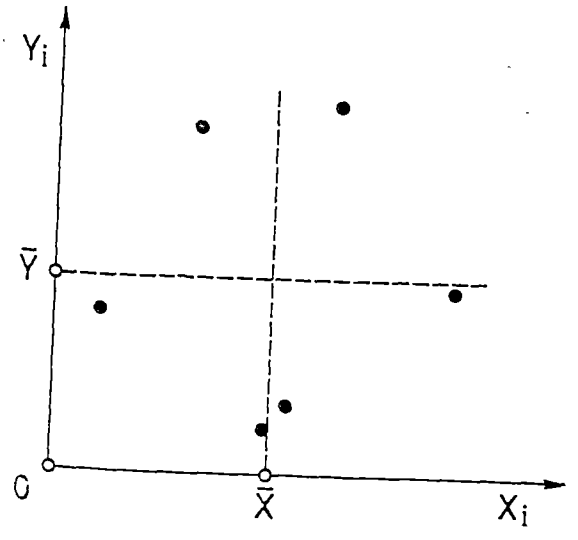
με τις τελείες και είναι ίδιες και στα τρία δείγματα. Η, διαφορετικά, οι περιθώριες κατανομές των Xi και Yi είναι ίδιες και στα τρια δείγματα. Η κοινή κατανομή όμως των 6 σημείων (Xi, Yi) στο επίπεδο XY είναι εντελώς διαφορετική όπως φαίνεται στα αντίστοιχα διαγράμματα διασποράς. Ομοίως, διαφέρουν οι συντελεστές συνδιακύμανσης και συσχέτισης.



Xi 2 6 9 10 12 17
Yi 2 3 8 7 15 16
X = 9,3 Sy = 10,65 Sxy = 28,34
Y = 8,5 Sy = 5,89 rxy = 0,45



Xi 2 6 9 10 12 17
Yi 15 16 7 8 2 3
X = 9,3 Sx = 10,65 Sxy = 28,34
Y = 8,5 Sy = 5,89 rxy = 0,78



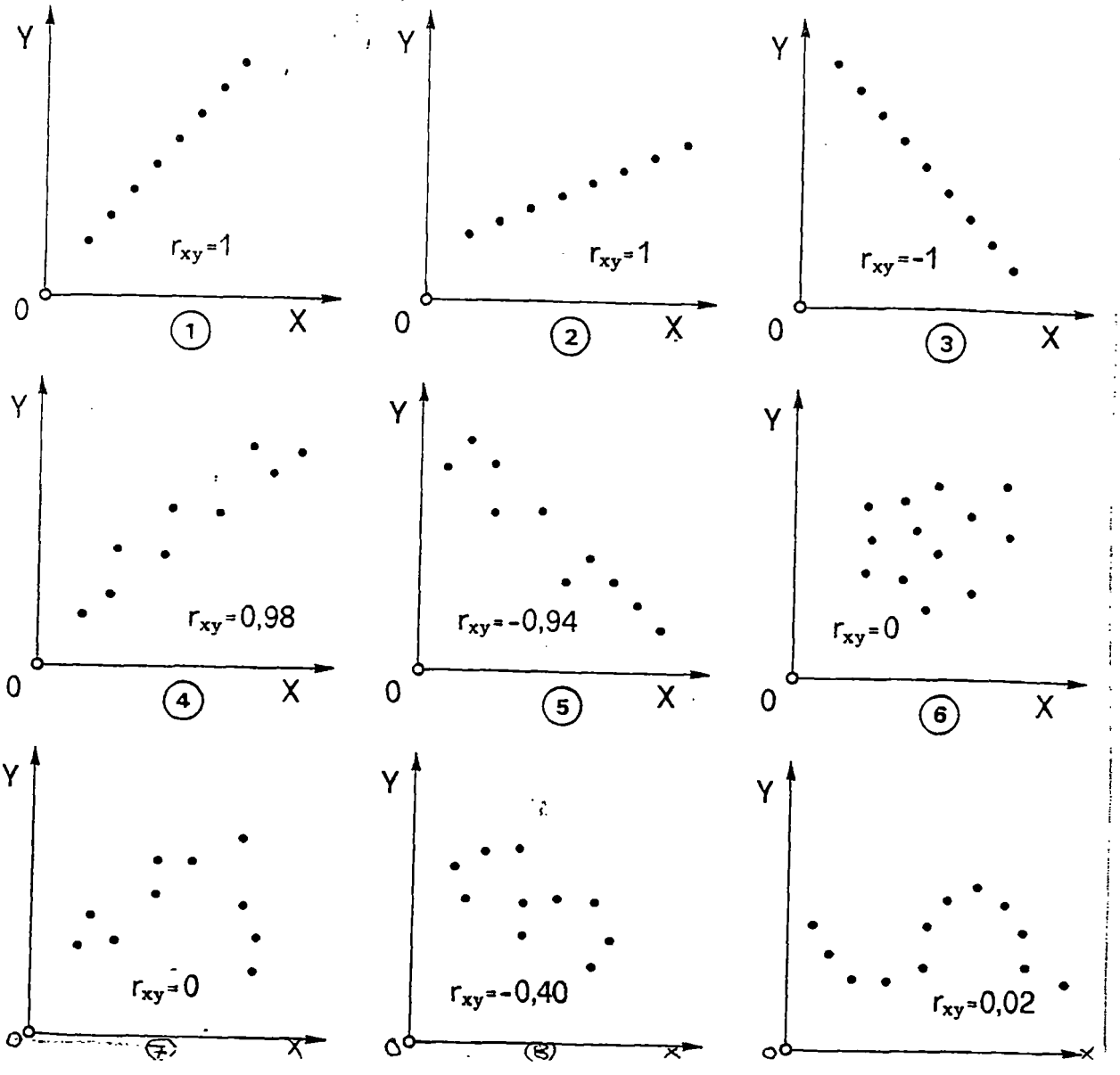
Xi 2 6 9 10 12 17
Yi 7 15 2 3 16 8
X = 9,3 Sx = 10,65 Sxy = 21,81
Y = 8,5 Sy = 5,89 rxy = - 0,35

3.4 **Ιδιότητες του συντελεστή συσχέτισης**

Αποδεικνύεται ότι ο συντελεστής **rxy** έχει τις ακόλουθες ιδιότητες

- Ιδιότητα 1n : Παίρνει τιμές στο κλειστό διάστημα [-1,1]

Οι ακραίες τιμές -1 και 1 αντιστοιχούν στην περίπτωση που όλα τα σημεία (Xi, Yi) , i = 1,....,n βρίσκονται επάνω σε μια ευθεία μεμ αρνητική ή θετική κλίση αντίστοιχα. Οταν οι Xi και Yi είναι παρατηρήσεις τυχαίων μεταβλητών, η πιθανότητα να υπολογίσουμε συντελεστή συσχέτισης ίσο με 1 ή -1 είναι μηδενική (σημειώνεται πάντως το προφανές **rxy** =1). Οι ακραίες τιμές όμως χρησιμεύουν για να ερμηνεύσουμε τις διάμεσες. Οσο πιο κοντά σε μια απ' αυτές βρίσκεται ο **rxy** τόσο πιο έντονη η γραμμική συμμεταβολή των παρατηρήσεων X και Y . Μια τιμή για τον **rxy** ίση ή πολύ κοντά στο μηδέν, δηλώνει απουσία οποιασδήποτε σχέσης. Θα πρέπει να τονιστεί ότι απρόβλεπτη είναι η επίδραση στην τιμή του **rxy** μιας μη γραμμικής όπως και μιας ή περισσοτέρων ακραίων τιμών.



• Ιδιότητα $2^η$: Η τιμή του δεν μεταβάλλεται με τον γραμμικό μετασχηματισμό των Xi, Yi.

Ετσι, αν Zi = α + b Xi, Wi = c + dYi, i = 1,...,n και τα b,d έχουν το ίδιο πρόσημο, τότε

$$r_{xy} = r_{zw}$$

Αντίστοιχα, αν τα b,d έχουν διαφορετικό πρόσημο, τότε

$$r_{xy} = - r_{zw}$$

Η ιδιότητα αυτή είναι ιδιαίτερα χρήσιμη διότι μας επιτρέπει να κάνουμε μετασχηματισμούς κλίμακας και αρχής των δεδομένων μας, έτσι ώστε να διευκολύνουμε τους υπολογισμούς. Ορισμένα μάλιστα υποπρογράμματα μετασχηματίζουν αυτομάτως τα δεδομένα στο διάστημα (0,1) πριν υπολογίσουν τον

Παράδειγμα

Δίνεται ο ετήσιος αριθμός των σκαφών Xi και ο αριθμός των πληρωμάτων Yi της Ελλάδος κατά τα έτη 1968-1981. Οι παρατηρήσεις είναι εκφρασμένες σε δεκάδες.

Έτη	1968	1969	1970	1971	1972	1973	1974	1975	1977	1978
Xi	9,5	9,9	7,0	5,7	6,0	6,2	6,2	6,0	8,3	8
Yi	69,7	84,3	49,0	30,7	33,5	34,9	35,0	29,2	37,9	36,5

1979	1980	1981
8,1	7,1	6,2
37,7	25,5	23,3

Πηγή : ΕΣΥΕ, Στατιστική Επετηρίς της Ελλάδος 1984.

Υπολογίζουμε :

$\Sigma X_i = 102,2$ $\quad \Sigma X_i^2 = 770,72$ $\quad \bar{X} = 7,3$

$\Sigma Y_i = 564,5$ $\quad \Sigma Y_i^2 = 26520,35$ $\quad \bar{Y} = 40,3$

$\Sigma Y_i X_i = 4333,24$

οπότε:

$$r_{xy} = \frac{\Sigma X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\Sigma X_i^2 - n\cdot\bar{X}^2}\ \sqrt{\Sigma Y_i^2 - n\bar{Y}^2}}$$

$$= \frac{4333,24 - 14(7,3)\cdot(40,3)}{\sqrt{770,72-14(7,3)^2}\ \sqrt{26520,35-14(40,3)^2}} = 0,70$$

Αν οι παρατηρήσεις δεν εκφράζονται σε δεκάδες, αλλά σε μονάδες, ο συντελεστής συσχέτισης παραμένει ο ίδιος.

3.5 Ερμηνεία του συντελεστή συσχέτισης

Επαναλαμβάνεται ότι ο συντελεστής συσχέτισης δεν είναι παρά μόνο ένα μέτρο γραμμικής συμμεταβολής δυο τυχαίων μεταβλητών. Επειδή συχνά παρερμηνεύεται από τους μη ειδικούς κρίθηκε σκόπιμο να γίνουν οι ακόλουθες διευκρινίσεις:

Συσχέτιση και αιτιότητα

Η συσχέτιση δεν σημαίνει αιτιότητα. Στη διεθνή βιβλιογραφία αναφέρεται χαρακτηριστικά το εξής παράδειγμα : ο υψηλός συντελεστής συσχέτισης ανάμεσα στις αφίξεις πελαργών σε μια περιοχή και στον αριθμό των γεννήσεων, δεν μας επιτρέπει να συμπεραίνουμε ούτε ότι οι πελαργοί φέρνουν τα μωρά (ούτε πολύ περισσότερο για τον έλεγχο των γεννήσεων να προτείνουμε τη θανάτωση τους) ούτε ότι τα μωρά προσελκύουν τους πελαργούς.
Γενικά, υψηλός συντελεστής συσχέτισης μπορεί να παρατηρηθεί όταν η Χ προκαλεί μεταβολές στην Υ ή η Υ προκαλεί μεταβολές στην Χ ή όταν μια Τρίτη μεταβλητή η οποία ονομάζεται **κεκαλυμένη** προκαλεί μεταβολές στην ίδια ή αντίθετη κατεύθυνση και στις δυο μεταβλητές ή τέλος, μπορεί να οφείλεται στην τύχη. Στις δύο τελευταίες περιπτώσεις η συσχέτιση ονομάζεται **νόθα ή χωρίς νόημα.**
Στις περισσότερες περιπτώσεις αρκεί ο κοινός νους για να συμπεράνουμε αν μια παρατηρούμενη συσχέτιση μπορεί να ερμηνευτεί ή είναι χωρίς νόημα.

Συσχέτιση και ανεξαρτησία

Ο συντελεστής συσχέτισης μετρά την ένταση γραμμικής συμμεταβολής. Ετσι δυο μεταβλητές Χ,Υ μπορεί να έχουν συντελεστή συσχέτισης ίσο με μηδέν και οι μεταβλητές να μην είναι ανεξάρτητες αλλά να συνδέονται με μια σχέση η οποία δεν είναι γραμμική. Μηδενική συσχέτιση δεν συνεπάγεται ανεξαρτησία εκτός από την ακόλουθη περίπτωση : Όταν η κοινή κατανομή των Χ και Υ είναι η κανονική όπως θα δούμε στο τμήμα 8 αυτού του κεφαλαίου.

Συσχέτιση και το εύρος των δεδομένων

Το μικρό εύρος των παρατηρήσεων της μιας ή και των δυο μεταβλητών έχει συνήθως ως αποτέλεσμα να υπολογίσουμε μια μικρή τιμή για τον **rxy**. Αν υποπτευόμστε ότι η μικρή τιμή του **rxy**
οφείλεται σ' αυτόν το λόγο και το επιτρέπουν οι συνθήκες δειγματοληψίας θα πρέπει να διευρύνουμε το δείγμα.

Ετσι π.χ αν μας ενδιαφέραι η συσχέτιση του ύψους Χ ενός παιδιού με το Νο Υ του παπουτσιού του είναι προτιμότερο να μην πάρουμε δείγμα παιδιών μόνον από την ίδια τάξη γιατί τότε το εύρος των δυο μεταβλητών θα είναι πιθανότατα μικρό και συνεπώς η τιμή του **rxy** μικρή ακόμη και αν η τιμή του ρ**xy** είναι μεγάλη.

Συσχέτιση και ακραίες τιμές

Η ύπαρξη μιας ή περισσοτέρων ακραίων τιμών μπορεί να έχει σημαντική επίδραση στην τιμή του **rxy**, η έκταση και η κατεύθυνση της οποίας μπορεί να είναι απρόβλεπτη. Το διάγραμμα διασποράς των τιμών (Xi, Yi) , i = 1,..,n είναι διαφωτιστικό στην περίπτωση αυτή.

3.6. Ο συντελεστής συσχέτισης ομαδοποιημένων παρατηρήσεων.

Η διμεταβλητή κατανομή συχνοτήτων.

Συχνά ι η διμεταβλητές παρατηρήσεις για τις οποίες θέλουμε να υπολογίσουμε τον συντελεστή συσχέτισης , είναι ομαδοποιημένες σε μια διμεταβλητή κατανομή συχνοτήτων όπως στο ακόλουθο ,

Παράδειγμα

Τυχαίο δείγμα 100 γυναικών που γέννησαν μη πρόωρα βρέφη ομαδοποιήθηκε στην ακόλουθη κατανομή ανάλογα με το ύψος Χ της μητέρας και το ύψος Υ του παιδιού, κατά τη γέννηση.

Ύψος παιδιού Υ Yi Ύψος μητέρας Χ Xi	34 - 38 36	38 - 42 40	42 - 46 44	46 - 50 48	50 - 54 52	54 - 58 56	Άθροισμα
154 - 158 156	4	2	1				7
158 - 162	2	5	2				9

160							
162 - 166 164	1	4	6	5			16
166 - 170 168		6	10	7	1		24
170 - 174 172	1		14	12			27
174 - 178 176		2	4	2	1	1	10
178 - 182 180				2	4	1	7
Άθροισμα	8	19	37	28	6	2	100

Υπολογισμός του συντελεστή συσχέτισης

Για να υπολογίσουμε το συντελεστή συσχέτισης ή ομαδοποιημένων διμεταβλητών παρατηρήσεων εργαζόμαστε ως εξής :

- Από τα περιθώρια κατανομή των X υπολογίζουμε τα X , Sx^2 .
- Από τα περιθώρια κατανομή των Y υπολογίζουμε τα Y, Sy^2 .
- Αν Fij η συχνότητα στη θέση ή "κελί " ij , δηλαδή στη θέση που είναι κοινή στην i-γραμμή και στην j-στήλη του διμεταβλητού πίνακα συχνοτήτων και Xi ,Yi οι κεντρικές τιμές των αντίστοιχων Fij Xi Yi και αθροίζουμε για όλα τα i και j .
- Υπολογίζουμε την συνδιακύμανση

$$Sxy = \frac{1}{n} \sum_i \sum_j (x_i - \bar{x}) \cdot (y_j - \bar{y}) f_{ij} \qquad (3.6.1.)$$

$$Sxy = \frac{1}{n} \sum_i \sum_j x_i y_j f_{ij} - \bar{x}\bar{y}$$

και το συντελεστή συσχέτισης

$$r_{xy} = \frac{Sxy}{Sx \cdot Sy} \qquad (3.6.2.)$$

Παράδειγμα

Υπολογίζουμε τον συντελεστή συσχέτισης του ύψους Χ της μητέρας και του ύψους Υ του παιδιού στις 100 διμεταβλητές παρατηρήσεις του προηγούμενου παραδείγματος , σταδιακά , ως εξής :

• Από την περιθώρια κατανομή των Χ υπολογίζουμε την μέση τιμή και την διακύμανση τους . Ειδικότερα έχουμε :

Τάξεις	Xi	Fi	Xi	XiFi	Xi^2Fi
154 - 158	156	7	-3	-21	63
158 - 162	160	9	-2	-18	36
162 - 166	164	16	-1	-16	16
166 - 170	168	24	0	0	0
170 - 174	172	27	1	27	27
174 - 178	176	10	2	20	40
178 - 182	180	7	3	21	63
Άθροισμα		100		13	245

Η βοηθητική μέση τιμή

$$\bar{X}' = \frac{1}{n} \Sigma X_i' F_i = \frac{13}{100} = 0,13$$

και η βοηθητική διακύμανση

$$S'^2_X = \frac{1}{n} \Sigma X_i'^2 F_i - \bar{X}'^2 = \frac{1}{100} 245 - (0,13)^2 = 2,433$$

οπότε

$$\bar{X} = 168 + (0,13)\,4 = 168,52$$
$$S^2_X = 4^2 (2,433) = 38,93 \quad \text{και} \quad S_X = 6,24$$

• Ομοίως , από την περιθώρια κατανομή των Υ υπολογίζουμε

$$\bar{Y} = 44,44$$
$$S^2_Y = 19,166 \quad \text{και} \quad S_Y = 4,38$$

• Για να υπολογίσουμε την συνδιακύμανση υπολογίζουμε αρχικά τους όρους :

$$\sum_i \sum_j X_i Y_j F_{ij} = (156)(36)4 + (156)\cdot(40)\cdot 2 + (156)\cdot 44$$
$$+ (160)(36)2 + (160)\cdot(40)\,5 + (160)\cdot(44)\cdot 2$$
$$+ (164)(36) + (164)\cdot(40)\cdot 4 + (164)(44)6 + (164)\cdot$$
$$\cdot(48)\cdot 5.$$

$$+(168)(40)6+(168)(44)10+(168)(48)7+(168)(52)$$
$$+(172)(36)+(172)(44)\cdot 14+(172)(48)\cdot 12$$
$$+(176)(40)2+(176)(44)4+(176)(48)2+(176)52+$$
$$\cdot(176)56$$
$$+(180)(48)2+(180)(52)\cdot 4+(180)\cdot 56=750608$$

και

$$\bar{X}\bar{Y}=(168,52)(44,44)=7489,03$$

οπότε

$$S_{XY}=\frac{750608}{100}-7489,03=17,05$$

και

$$r_{XY}=\frac{S_{XY}}{S_X\cdot S_Y}=\frac{17,05}{(6,24)(4,38)}=0,624$$

3.7 Ο Θεωρητικός συντελεστής συσχέτισης

Ο συντελεστής συσχέτισης δύο τυχαίων μεταβλητών X και Y με μέση τιμή μx και μy , αντίστοιχα , συμβολίζεται με ρ_{xy} και ορίζεται ως εξής :

$$\rho_{XY}=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}=\frac{E(X-\mu_X)\cdot(Y-\mu_Y)}{\sqrt{E(X-\mu_X)^2}\sqrt{E(Y-\mu_Y)^2}} \qquad (3.7.1.)$$

Όταν οι X και Y είναι διακριτές , τότε η συνδιακύμανση τους δηλαδή ο αριθμός της (3.7.1.) υπολογίζεται ως εξής :

$$Cov(X,Y)=\sum_i\sum_j(X_i-\mu_X)\cdot(Y_j-\mu_Y)\cdot P(X_i\cdot Y_j) \qquad (3.7.2.)$$

ενώ οι τυπικές αποκλίσεις ως εξής :

$$\sigma_X=\sqrt{\sum_i(X_i-\mu_X)^2P(X_i)}$$
$$\sigma_Y=\sqrt{\sum_i(Y_i-\mu_Y)^2P(X_i)} \qquad (3.7.3.)$$

Συγκρίνοντας τις (3.7.1.),(3.7.2.),(3.7.3.) με τους τύπους που εφαρμόσαμε για να υπολογίσουμε τον συντελεστή συσχέτισης ομαδοποιημένων παρατηρήσεων παρατηρούμε ότι αυτοί διαφέρουν μόνο στο ότι οι σχετικές συχνότητες F_{ij}/η έχουν αντικατασταθεί από τις πιθανότητες P(Xi,Yi) και οι X , Y από τις μx , μy .

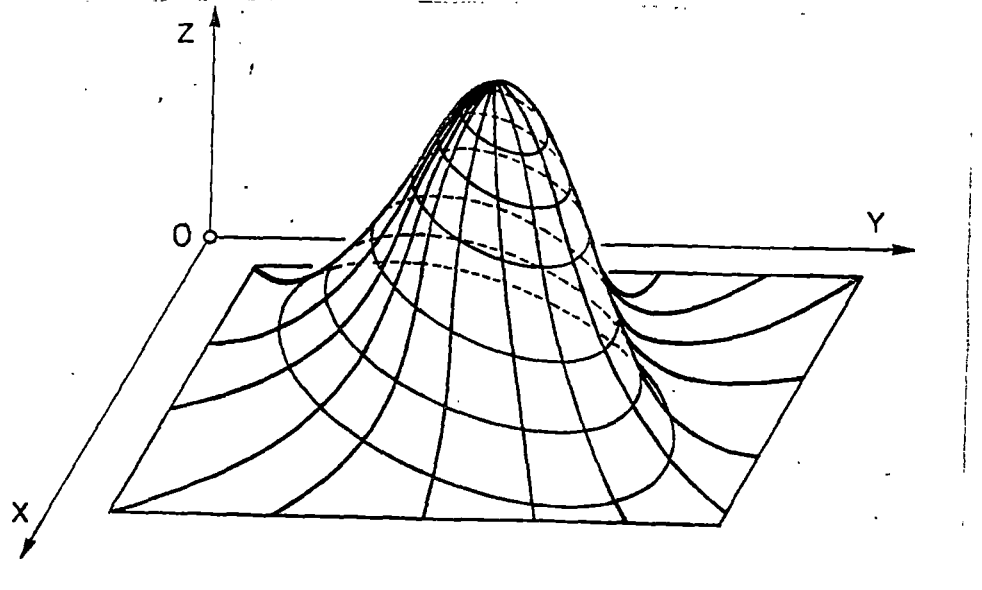
Πράγματι, ο r_{xy} είναι το δειγματικό ανάλογο του ρ_{xy}, όπως τα $\bar{x}$, $\bar{y}$, s^2x, s^2y είναι τα δειγματικά ανάλογα των παραμέτρων μx, μy, σ^2x, σ^2y αντίστοιχα.

Για να υπολογίσουμε τον ρ_{xy} πρέπει να γνωρίζουμε την κοινή κατανομή πιθανοτήτων των X και Y κάτι που στην πράξη δεν συμβαίνει σχεδόν ποτέ .Συνήθως έχουμε ένα δείγμα η παρατηρήσεων των τυχαίων μεταβλητών X και Y για τις οποίες υπολογίζουμε τον συντελεστή συσχέτισης r_{xy} που στην περίπτωση αυτή ονομάζεται **δειγματικός συντελεστής συσχέτισης** ενώ ο ρ_{xy} ονομάζεται **θεωρητικός ή συντελεστής συσχέτισης πληθυσμού .**

Ο r_{xy} είναι εκτιμητής του ρ_{xy} του οποίου η αξιοπιστία εξαρτάται από το μέγεθος του δείγματος αλλά και την κατανομή πληθυσμού δηλαδή την κοινή κατανομή των X και Y . Όταν η κατανομή αυτή είναι η διμεταβλητή κανονική μπορούμε να προσδιορίσουμε την αξιοπιστία του r_{xy} , να υπολογίσουμε διαστήματα εμπιστοσύνης και να κάνουμε τους σχετικούς ελέγχους .

3.8. **Η διμεταβλητή κανονική κατανομή .**

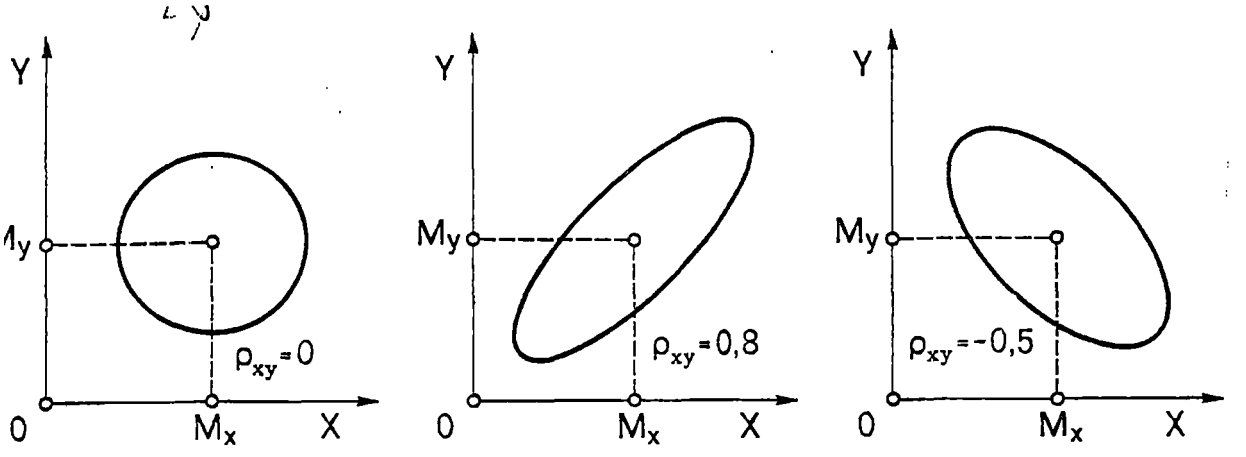
Ένα μοντέλο για την κοινή κατανομή δύο συνεχών τυχαίων μεταβλητών X και Y είναι **η κοινή κανονική κατανομή .**
Οι παράμετροι της διμεταβλητής κανονικής κατανομής είναι οι μέσες τιμές μ_x ,μ_y,
οι διακυμάνσεις σ^2x σ^2y των x και y αντίστοιχα και ο συντελεστής συσχέτισης P_{xy}.
Η γραφική παράσταση της συνάρτησης πυκνότητας πιθανότητας είναι η επιφάνεια ενός λόφου που υψώνεται πάνω στο επίπεδο xy όπως στο ακόλουθο σχήμα και έχει τα εξής χαρακτηριστικά :



• Ο όγκος του χώρου που περικλείεται από την επιφάνεια αυτή και το επίπεδο ΧΥ ισούται με 1.

• Η κορυφή του ΄΄ λόφου ΄΄ βρίσκεται στο σημείο (μx , μy)

• Για δεδομένες τις μέσες τιμές μx μy και τις διακυμάνσεις σ^2x , σ^2y , το σχήμα του ΄΄ λόφου ΄΄ εξαρτάται από τον συντελεστή συσχέτισης Ρxy . Όταν Ρxy=ο , ο λόφος είναι συμμετρικά κυκλικός και η τομή του με το επίπεδο ΧΥ είναι ένας κύκλος (εφ. σχ. α) . Όσο αυξάνει η απόλυτη τιμή του Ρxy , ο λόφος πιέζεται από την μία πλευρά έτσι ώστε η τομή του με το επίπεδο ΧΥ είναι μια έλλειψη με κέντρο στο σημείο (μx , μy) (σχ. β,γ). Αν Ρxy>ο η μεγάλη ακτίνα της έλλειψης έχει θετική κλίση διαφορετικά , έχει αρνητική . Όταν Ρxy =1 ο λόφος δεν έχει καθόλου θετική ή αρνητική κλίση ανάλογα με το αν το πρόσημο του Ρxy είναι (+) ή (-) .

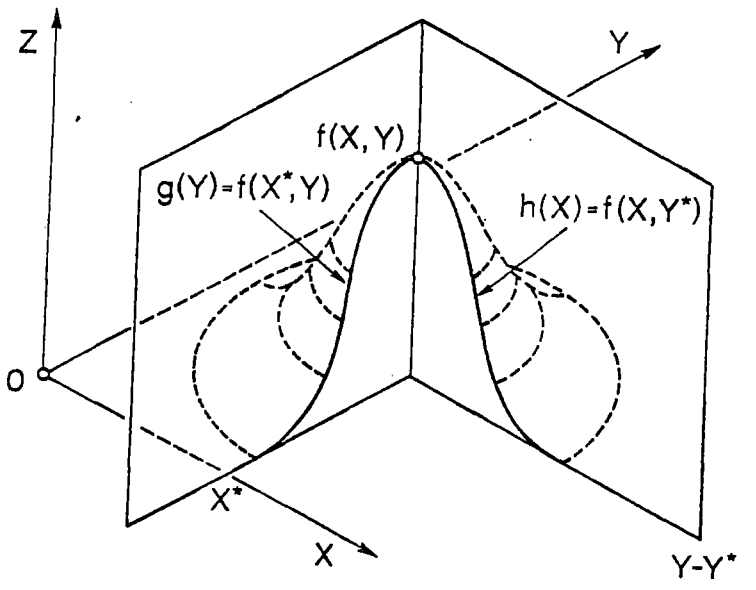


(α) (β) (γ)

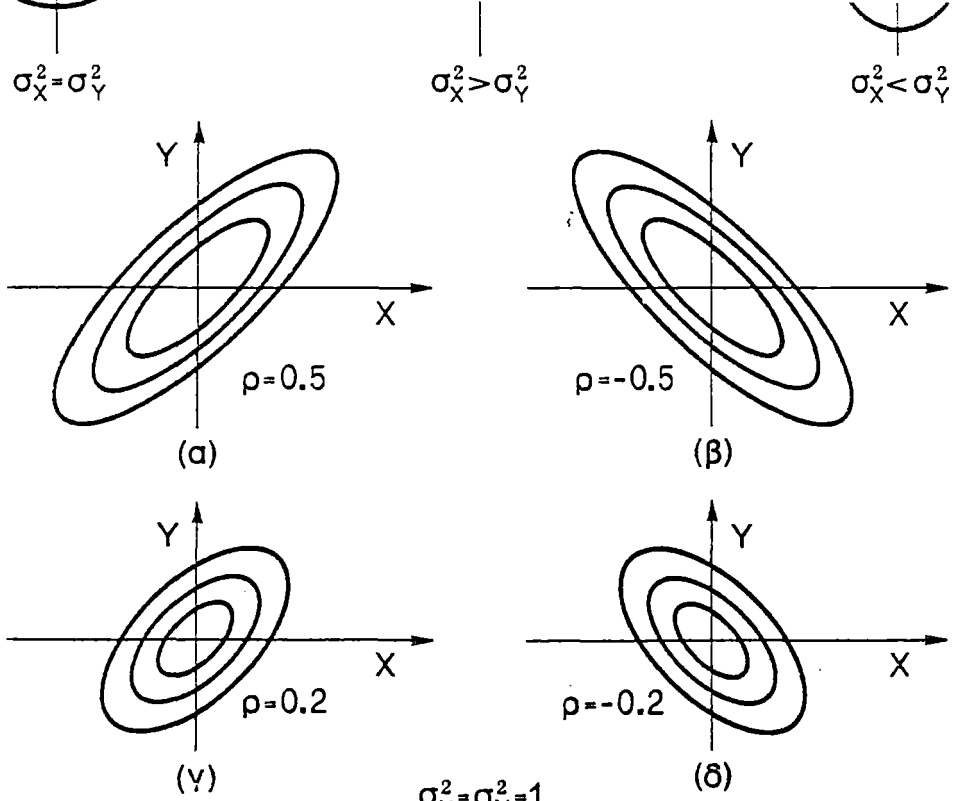
• Μπορούμε να κόψουμε το λόφο που σχηματίζει η διμεταβλητή κανονική συνάρτηση πυκνότητας πιθανότητας με ένα επίπεδο παράλληλο στο ΧΥ . Η τομή του βουνού με το επίπεδο αυτό ονομάζεται **ισοϋψής καμπύλη πιθανότητας** και είναι επίσης έλλειψη όπως και η τομή με το ΧΥ . Ο όγκος που περικλείεται από την επιφάνεια που ορίζει μια ισοϋψής καμπύλη και το υπερκείμενο μέρος του λόφου αντιστοιχεί σε ορισμένη πιθανότητα . Έτσι, αρκεί να γνωρίζουμε τις ισοϋψής καμπύλες

που αντιστοιχούν σε διαφορετικές πιθανότητες για να γνωρίζουμε την διμεταβλητή κατανομή πιθανοτήτων . Στο (σχ. δ) δίνονται οι ισοϋψής καμπύλες που αντιστοιχούν σε πιθανότητες 50% , 75% και 99% για διαφορετικές τιμές των $\sigma^2 x$, $\sigma^2 y$ και Pxy ενώ μx=μy=0 .

• Η τομή της επιφάνειας που ορίζει η διμεταβλητή κανονική συνάρτηση πιθανότητας με ένα επίπεδο κάθετο στο XY είναι μία κανονική καμπύλη . Η διαφορετικά , η δεσμευμένη κατανομή της Y (αντίστοιχα , της X) για οποιαδήποτε τιμή της X (αντίστοιχα , της Y) είναι κανονική (σχ. ε) .



(ε)


$\sigma_X^2=\sigma_Y^2$
$\sigma_X^2>\sigma_Y^2$
$\sigma_X^2<\sigma_Y^2$
Y
X
ρ=0.5
(α)
Y
X
ρ=-0.5
(β)
Y
X
ρ=0.2
(γ)
Y
X
ρ=-0.2
(δ)
$\sigma_X^2=\sigma_Y^2=1$

(δ)

Y
X
Y
X
Y
X
Y
X
Y
X
Y
X
Y
X
Y
X
Y
X
ρ=0.2
ρ=0.5
ρ=0.8
$\sigma_X^2=4$ $\sigma_Y^2=1$

Βασικές Ιδιότητες της Διμεταβλητής Κανονικής Κατανομής

Η διμεταβλητή κανονική κατανομή έχει ορισμένες ιδιότητες οι οποίες θα φανούν χρήσιμες στην συνέχεια της ανάλυσης . Έτσι, όταν η κοινή κατανομή των X,Y είναι κανονική, ισχύουν τα εξής :

1. Οι περιθώριες κατανομές των X και Y είναι επίσης κανονικές δηλαδή X~N ($\mu x, \sigma^2 x$) και Y~N (μy, $\sigma^2 y$).Το αντίθετο όμως δεν ισχύει . Πολλές διμεταβλητές μη κανονικές κατανομές έχουν κανονικές περιθώριες .
2. Κάθε γραμμικός συνδυασμός των X,Y κατανέμεται επίσης κανονικά . Απ'όλες τις διμεταβλητές κατανομές μόνο η κανονική έχει αυτή την ιδιότητα .
3. Αν $Pxy=o \Leftrightarrow$ οι X.,Y είναι ανεξάρτητες . Δηλαδή στην διμεταβλητή κανινική κατανομή και μόνον α'αυτήν οι έννοιες " ανεξάρτατες " και " ασυσχέτιστες " είναι ταυτόσημες .
4. Η συνάρτηση παλινδρόμησης της Y επί της X όπως και η συνάρτηση παλινδρόμησης της X επί της Y είναι γραμμικές .
5. Η δεσμευμένη κατανομή της Y για κάθε Xi έχει σταθερή διακύμανση και ίση με $\sigma^2 y (1-p^2 xy)$.

Πως θα ξέρουμε αν η κοινή κατανομή δύο συνεχών τυχαίων μεταβλητών είναι κανονική .

Ο έλεγχος προσαρμογής της διμεταβλητής κανονικής κατανομής σε ένα δείγμα η ζευγών παρατηρήσεων είναι πολύπλοκος και δεν μπορεί να γίνει χωρίς ηλεκτρονικό υπολογιστή .

Αρχικά γίνεται έλεγχος προσαρμογής της κανονικής κατανομής σε κάθε μια από τις περιθώριες κατανομές των X και Y . Αν αυτές δεν μπορούν να θεωρηθούν κανονικές τότε μπορούμε να απορρίψουμε την υπόθεση της κανονικότητας για την κοινή κατανομή πληθυσμού .Αν όμως οι περιθώριες κατανομές μπορούν να θεωρηθούν κανονικές τότε , σύμφωνα με την ιδιότητα 1 της διμεταβλητής κανονικότητας είναι πολύ πιθανόν αλλά όχι βέβαιο ότι η κοινή κατανομή είναι κανονική . Η διαδικασία αυτή αναφέρεται πληροφοριακά παρόλο που υπάρχει κίνδυνος να χρησιμοποιηθεί λανθασμένα .

3.9 Έλεγχος των υποθέσεων και διαστήματα εμπιστοσύνης .

Έστω (X_1, Y_1), ... (Xn ,Yn) ένα τυχαίο δείγμα η παρατηρήσεων από μια διμεταβλητή κατανομή πιθανοτήτων και r_{xy} ο δειγματικός συντελεστής συσχέτισης .

Έχει αποδειχθεί ότι , ακόμη και όταν η κατανομή πληθυσμού είναι η διμεταβλητή κανονική , η κατανομή δειγματοληψίας του r_{xy} δεν είναι η κανονική. Εξαρτάται από τον συντελεστή συσχέτισης pxy της κατανομής πληθυσμού και το μέγεθος του δείγματος η , προσεγγίζει Δε την κανονική μόνον για πάρα πολύ μεγάλες τιμές του η- μερικές εκατοντάδες . Αν όμως κάνουμε στον rxy τον ακόλουθο μετασχηματισμό :

$$W = \frac{1}{2} \log_e \left(\frac{1 + r_{xy}}{1 - r_{xy}} \right)$$

τότε η κατανομή της W προσεγγίζει κανονική με μέση τιμή του W-μετασχηματισμό του pxy και τυπική απόκλιση ίση με $1/\sqrt{n-3}$. Η προσέγγιση της κατανομής της W στην κανονική είναι ικανοποιητική ακόμη και για μικρά , σχετικά , δείγματα ($n \geq 15$)
Επομένως , ο έλεγχος των υποθέσεων και ο υπολογισμός του διαστήματος εμπιστοσύνης για τον συντελεστή συσχέτισης πληθυσμού θα γίνει , αντίστοιχα , ως εξής :

Έλεγχος των υποθέσεων

α . Υποθέσεις που γίνονται δεκτές .
Η κατανομή πληθυσμού είναι η διμεταβλητή κανονική

β . Η μηδενική και η εναλλακτική υπόθεση .

$$H_o : \rho_{xy} = \rho_o$$

$$H_\varepsilon : \rho_{xy} \neq \rho_o$$

γ . Το κριτήριο αποφάσεως .

Έστω Wr και Wo ο W- μετασχηματισμός του rxy και του ρο αντίστοιχα . Τους μετασχηματισμούς αυτούς βρίσκουμε απ'ευθείας στον πίνακα 7 στο τέλος της εργασίας .
Τότε , αν : $|z_n| = \frac{|W_r - W_o|}{1/\sqrt{n-3}} > z_{1-\alpha/2}$

απορίπτουμε την Ηο στο επίπεδο σημαντικότητας α .

δ . Η απόφαση

Παράδειγμα

Επιφανής ανθρωπολόγος δήλωσε ότι ο συντελεστής ανάμεσα στο μήκος του ποδιού X μιας ενήλικης γυναίκας και το μήκος του χεριού της Y, είναι ίσος με 0,80 . Να ελεγχθεί η υπόθεση αυτή αν στις παρατηρήσεις τυχαίου δείγματος μεγέθους n=307 ενηλίκων γυναικών υπολογίσαμε συντελεστή συσχέτισης rxy =0,72 και μπορούμε να υποθέσουμε ότι η κοινή κατανομή των X και Y είναι η κανονική . Ο έλεγχος να γίνει στο επίπεδο σημαντικότητας α=0,05 .

α . Υποθέσεις που γίνονται δεκτές

Το δείγμα των 307 γυναικών είναι τυχαίο, η κοινή κατανομή των X και Y είναι η κανονική .

β . Η μηδενική και η εναλλακτική υπόθεση

$H_0 : \rho_{xy} = 0,80$

$H_\varepsilon : \rho_{xy} \neq 0,80$

γ . Το κριτήριο αποφάσεως και η απόφαση

Από τον πίνακα 7, στο τέλος της εργασίας , βρίσκουμε ότι η τιμή του συντελεστή συσχέτισης 0,80 μετασχηματίζεται στην W- τιμή Wo= 1,0986 , ενώ η τιμή 0,72 μετασχηματίζεται στην Wr = 0,9076.
Υπολογίζουμε : $|z_n| = \frac{|Wr-Wo|}{1/\sqrt{n-3}} = \frac{|0,9076-1,0986|}{1/\sqrt{n-3}} = 3,33$

Επειδή $|z_n| > z_{1-\alpha/2} = z_{0,975} = 1,96$ απορρίπτουμε την Ho στο α= 0,05

Το διάστημα εμπιστοσύνης

Το διάστημα το οποίο με πιθανότητα (1-α) θα περιέχει την pxy είναι το
$\rho_l \leq \rho_{xy} \leq \rho_u$.
Το ρ_l είναι η τιμή του συντελεστή συσχέτισης που στον πίνακα στο τέλος της εργασίας , αντιστοιχεί στην $Wr-Z_{1-\alpha/2}\sqrt{1/n-3}$ όπου Wr είναι ο W-μετασχηματισμός του rxy και $z_{1-\alpha/2}$ η τιμή της τυπικής κανονική

κατανομής που αντιστοιχεί σε αθροιστική συχνότητα 1-α/2 . Ομοίως το ρι είναι η τιμή του p που αντιστοιχεί στην $Wr+Z_{1-\alpha/2}\sqrt{1/n-3}$.

Ο Έλεγχος της υπόθεσης για μηδενική συσχέτιση

Όταν η κατανομή πληθυσμού είναι η διμεταβλητή κανονική και ισχύει ρxy=ο τότε , η τυχαία ματαβλητή

$$t = \frac{r_{xy}}{\sqrt{\frac{1-r^2_{xy}}{n-2}}}$$

ακολουθεί την κατανομή t- student με ν=n-2 βαθμούς ελευθερίας . Επομένως ο έλεγχος για μηδενική συσχέτιση δύο τυχαίων μεταβλητών Χ και Υ θα γίνει σταδιακά ως εξής :

α . Υποθέσεις που γίνονται δεκτές :

Η κοινή κατανομή πληθυσμού είναι η κανονική .

β . Η μηδενική και η εναλλακτική υπόθεση :

$H_0 : \rho_{xy} = 0$

$H_\varepsilon : \rho_{xy} \neq 0$

γ . Το κριτήριο αποφάσεως :

Αν rxy η τιμή του δειγματικού συντελεστή συσχέτισης ,η το μέγεθος του δείγματος και ισχύει :

$$|t_n| = \frac{|r_{xy}|}{\sqrt{\frac{1-r^2_{xy}}{n-2}}} > t_{n-2,\alpha/2}$$

τότε απορρίπτεται η Ηο στο επίπεδο σημαντικότητας α .

δ . Η απόφαση

Παράδειγμα

Σε τυχαίο δείγμα n=4 νεογέννητων μετρήσαμε το βάρος τους X και το ύψος τους Y και υπολογίσαμε δειγματικό συντελεστή συσχέτισης ίσο με rxy =0,76 .Γνωστού όντως ότι η κοινή κατανομή των μεταβλητών ανθρώπινο βάρος και ύψος , σε οποιαδήποτε ομάδα μπορεί να υποτεθεί κανονική , να ελεγχθεί η υπόθεση ότι οι δύο μεταβλητές είναι , στα νεογέννητα , ανεξάρτητες .

Αρκεί να ελέγξουμε αν οι γραμμικώς ανεξάρτητες δηλαδή αν ρxy=ο . Ο ζητούμενος έλεγχος θα γίνει σταδιακά ως εξής :

α . Υποθέσεις που γίνονται δεκτές :

Η κοινή κατανομή των X και Y είναι η κανονική .

β . Η μηδενική και η εναλλακτική υπόθεση :

$H_0 : \rho_{xy} = 0$

$H_\varepsilon : \rho_{xy} \neq 0$

γ . Το κριτήριο αποφάσεως και η απόφαση :

Επειδή $t_{n-2,\,9/2} = t_{12\,(0,025)} = 2,179$ και ισχύει :

$$|t_n| = \frac{|r_{xy}|}{\sqrt{\frac{1-r^2_{xy}}{n-2}}} = \frac{0,76}{\sqrt{\frac{1-(0,76)^2}{14-2}}} = 4,05 > 2,179$$

απορρίπτουμε την Ho .

Σημείωση . Στην κανονική διμεταβλητή κατανομή οι έννοιες ασυσχέτιστες και ανεξάρτητες μεταβλητές είναι συνώνυμες .
Επομένως **στην περίπτωση αυτή και μόνο ο προηγούμενος έλεγχος είναι και έλεγχος ανεξαρτησίας** . Θα πρέπει πάντως να σημειωθεί ότι ο έλεγχος για μηδενική τιμή του ρxy μπορεί να εφαρμοστεί ακόμη και όταν η κοινή κατανομή των X,Y δεν μπορεί να υποτεθεί κανονική , αρκεί να ισχύει $n \geq 30$.

3.10 Παλινδρόμηση και Συσχέτιση

Ο συντελεστής συσχέτισης rxy που υπολογίζεται για τα ζεύγη παρατηρήσεων (X_1 , Y_1),.... , (Xn ,Yn) μετρά την τάση τους να συγκεντρώνονται γύρω από μια ευθεία . Ο rxy είναι συμμετρικός ως προς τις X και Y με την έννοια ότι είναι αδιάφορο αν η ευθεία ορίζεται ως $L = \{ (X,Y) : Y=b_o + b_1 X \}$ ή ως $L' = \{ (Y,X) : X=\alpha_o + \alpha_1 Y \}$.

Στην εκτίμηση ενός μοντέλου παλινδρόμησης οι X και Y δεν είναι συμμετρικές . Έτσι , αν αλλάξουμε τη θέση των X και Y δεν θα πάρουμε τα ίσια αποτελέσματα . Είναι όμως ενδιαφέρον να δούμε πως ο συντελεστής συσχέτισης rxy συνδέεται με τις παραμέτρους και , γενικότερα , τα στατιστικά ενός γραμμικού μοντέλου παλινδρόμησης που εκτιμήσαμε με βάση τις ίδιες παρατηρήσεις . Στη συνέχεια , η αναφορά θα γίνεται στο μοντέλο $\hat{Y}= \hat{b}_o + \hat{b}_1 X$ λόγω συμμετρίας του rxy όμως τα αποτελέσματα ισχύουν και για το μοντέλο $\hat{X}=\hat{\alpha}_o+\hat{\alpha}_1 Y$.

Ο rxy και ο συντελεστής παλινδρόμησης $\hat{b}_1$

Από τους τύπους

$$\hat{b}_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$r_{xy} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2}\sqrt{\Sigma (y_i - \bar{y})^2}}$$

προκύπτει ότι

$$\hat{b}_1 = r_{xy} \frac{\sqrt{\Sigma (y_i - \bar{y})^2}}{\sqrt{\Sigma (x_i - \bar{x})^2}} \qquad (3.10.1)$$

και επειδή

$$\Sigma (x_i - \bar{x})^2 = (n-1) s^2_x \cdot \text{ και } \Sigma (y_i - \bar{y})^2 = (n-1) s^2_y$$

η (3.10.1) γράφεται , ισοδύναμα ως εξής :

$$\hat{b}_1 = r_{xy} \frac{s_y}{s_x} \Longleftrightarrow r_{xy} = \hat{b}_1 \frac{s_x}{s_y} \qquad (3.10.2)$$

Από τη σχέση αυτή προκύπτει ότι ο rxy έχει το ίδιο πρόσημο με το συντελεστή παλινδρόμησης $\hat{b}_1$. Όταν η συμμεταβολή των X και Y είναι προς την ίδια κατεύθυνση , τότε και η κλίση $\hat{b}1$ της ευθείας παλινδρόμησης είναι θετική . Διαφορετικά , είναι αρνητική .

Όταν θέλουμε να ελέγξουμε την υπόθεση ότι οι X και Y δεν συνδέονται γραμμικά , αρκεί να κάνουμε έλεγχο για μηδενική τιμή του b1 είτε έλεγχο για μηδενική τιμή του pxy . Οι δύο έλεγχοι είναι ισοδύναμη .

Παλινδρόμηση προς το μέσο

Στην κεντραρισμένη μορφή του απλού μοντέλου παλινδρόμησης δηλαδή την $\hat{y} - \bar{y} = \hat{b}_1 (x - \bar{x})$

αν αντικαταστήσουμε την $\hat{b}_1 = r_{xy} \frac{s_y}{s_x}$ παίρνουμε : $\hat{y} - \bar{y} = r_{xy} \frac{s_y}{s_x} (x - \bar{x})$

Διαιρώντας και τα δύο σκέλη με Sy προκύπτει η

$$\frac{\hat{y} - \bar{y}}{s_y} = r_{xy} \frac{x - \bar{x}}{s_x} \qquad (3.10.3)$$

Για να εκτιμήσουμε το απλό μοντέλο παλινδρόμησης μπορούμε να εκφράσουμε τις παρατηρήσεις Xi, όπως και τις Yi, σε τυπικές αποκλίσεις από τον αριθμητικό τους μέσο δηλαδή αντίστοιχα , ως :

$$z_{xi} = \frac{x_i - \bar{x}}{s_x}$$

και $i = 1, \ldots, n$ (3.10.4)

$$z_{yi} = \frac{y_i - \bar{y}}{s_y}$$

Είναι εύκολο να δειχτεί ότι , τότε θα πάρουμε το ακόλουθο μοντέλο :

$$\hat{z}_y = r_{xy} z_x \qquad (3.10.5)$$

δηλαδή το (3.10.3) . Ο μετασχηματισμός (3.10.4) αναφέρεται ως τυποποίηση των παρατηρήσεων ή μετασχηματισμός τους σε Z - τιμές (z-scores) . Είναι προφανές ότι οι Zxi, i=1, ... , η έχουν μέση τιμή μηδέν και διακύμανση 1 όπως οι Zyi , i=1, ...,n.

Η εξίσωση (3.10.3) ή η ισοδύναμη της (3.10.5) περιγράφει ένα φαινόμενο το οποίο είναι γνωστό ως **παλινδρόμηση ή επαναφορά προς το μέσο** και μας λέει ότι σε μια τιμή X η οποία απέχει π.χ μία τυπική απόκλιση από το μέσο x εκτιμούμε ότι αντιστοιχεί μια τιμή Y η οποία θα απέχει από τον y όχι μια τυπική αλλά ένα ποσοστό της ίσο με rxy . Η ονομασία οφείλεται στον F. Galton (1822-1911) ο οποίος μελετώντας την ανθρώπινη κληρονομικότητα παρατήρησε ότι οι πατεράδες οι οποίοι είναι πολύ ψηλότεροι από το μέσο όρο, τείνουν να έχουν γιους οι οποίοι είναι λιγότερο ψηλότεροι από τον αντίστοιχο μέσο όρο . Έτσι αν X= το ύψος του πατέρα και Y= το ύψος του γιου και rxy= 0,6 τότε οι πατεράδες που έχουν ύψος 2 τυπικές αποκλίσεις πάνω από το μέσο όρο έχουν γιους

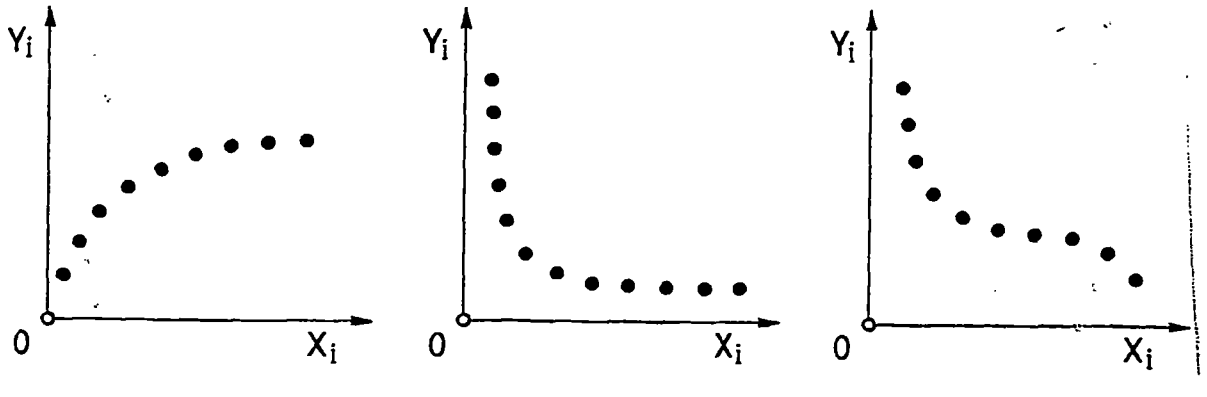
που κατά μέσο όρο έχουν ύψος (0,6)2=1,2 τυπικές αποκλίσεις πάνω από το μέσο όρο .

Στην διεθνή βιβλιογραφία αναφέρονται και άλλα παραδείγματα παλινδρόμησης προς το μέσο . Έτσι π.χ σε δύο διαδοχικές μετρήσεις των ιδίων στοιχείων, εξαιρετικά μεγάλα σφάλματα μέτρησης την πρώτη φορά τείνουν να ακολουθούνται από λιγότερο υψηλά σφάλματα την επόμενη .Εξαιρετικά υψηλές επιδόσεις των αθλητών μία φορά τείνουν να ακολουθούνται από λιγότερο υψηλές επιδόσεις την επόμενη λόγω παλινδρόμησης προς το μέσο . Ομοίως εξαιρετικά χαμηλές επιδόσεις την μία φορά τείνουν να ακολουθούνται από λιγότερο χαμηλές επιδόσεις την άλλη .

Το φαινόμενο της παλινδρόμησης προς το μέσο θα πρέπει να παίρνεται υπόψη στα προβλήματα επαναληπτικών ελέγχων .Έτσι π.χ έστω ότι σε μία ομάδα παιδιών προσχολικής ηλικίας δίνεται ένα IQ test στα 4 χρόνια τους και ένα δεύτερο στα 5 . Τα αποτελέσματα των δύο τεστ φυσικά συνεχίζονται . Αν με βάση τα αποτελέσματα του πρώτου τεστ τα παιδιά με χαμηλό IQ επιλέγηκαν για συμπληρωματική εκπαίδευση η τάση για μεγαλύτερες αποδόσεις στο Β τεστ μπορεί λανθασμένα να αποδοθεί στην συμπληρωματική εκπαίδευση . Στην περίπτωση αυτή η σύγκριση πρέπει να γίνει συγκρίσιμη ομάδα ελέγχου (control group)

3.11 Ο συντελεστής συσχέτισης κατά τάξεις

Είναι δυνατόν οι παρατηρήσεις (Xi, Yi) , i = 1, ..., n να μην προέρχονται από μία κανονική διμεταβλητή κατανομή και να συνδέονται ξεκάθαρα με μία σχέση που όμως δεν είναι γραμμική , όπως στα ακόλουθα σχήματα :



Σχέσεις όπως αυτές ονομάζονται μονοτικές . Γενικά ,

Δύο μεταβλητές X και Y συνδέονται με μία αυστηρά αύξουσα (φθίνουσα) μονοτική σχέση αν και μόνο αν , αυξάνουν (μειώνονται) οι τιμές της μίας με κάθε αύξηση των τιμών της άλλης . Η σχέση δεν είναι αυστηρά μονοτική αν στην αύξηση της μιας η άλλη μπορεί και να μην μεταβάλλεται .

Αν οι (Xi, Yi) συνδέονται με μια μονοτική σχέση και στις τιμές Xi αντίστοιχα τις τάξεις Wi και ομοίως στις Yi τις τάξεις Vi, i= 1,...,n τότε τα σημεία (Wi, Vi) θα βρίσκονται πάνω σε μια ευθεία . Επομένως , αν υπολογίσουμε τον συντελεστή συσχέτισης r_{wv} αυτός θα είναι ίσος με 1 ή -1 .

Ο συντελεστής συσχέτισης που υπολογίζεται για τις τάξεις η ζευγών παρατηρήσεων ονομάζεται **συντελεστής συσχέτισης κατά τάξεις ή συντελεστής συσχέτισης του Spearman** και συμβολίζεται με r_s . Δηλαδή έχουμε (ο τελεστής Σ είναι για όλες τις τιμές i = 1,....,n) :

$$r_s = - \frac{\Sigma (W_i - \bar{W})(V_i - \bar{V})}{\sqrt{\Sigma (W_i - \bar{W})^2}\sqrt{\Sigma (V_i - \bar{V})^2}} \qquad (3.11.1)$$

Επειδή 1+2+....+n=n (n+1) / 2 και

$$\bar{V}^2 = \bar{W}^2 = \bar{V} * \bar{W} = \frac{n(n+1)^2}{4} \equiv C$$

μπορούμε να υπολογίσουμε εύκολα τον r_s υπολογίζοντας το C και στη συνέχεια αντικαθιστώντας στην (3.11.1) η οποία έτσι γίνεται :

$$r_s = \frac{\Sigma W_i V_i - C}{\sqrt{\Sigma W_i^2 - C}\sqrt{\Sigma V_i^2 - C}} \qquad (3.11.2)$$

Αποδεικνύεται , ότι όταν δεν υπάρχουν επαναλαμβανόμενες παρατηρήσεις στα δεδομένα , οι τύποι (3.11.1) και (3.11.2) είναι ισοδύναμοι με τον υπολογιστικά ευκολότερο τύπο : $r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$, όπου $d_i = W_i - V_i$, $i = 1, \ldots, n$

Χρησιμότητα του r_s

Ο συντελεστής συσχέτισης κατά τάξεις είναι ιδιαίτερα χρήσιμος στις κοινωνικές επιστήμες όπου οι παρατηρήσεις είναι πολλές φορές στην τακτική κλίμακα μέτρησης . Είναι συχνές οι περιπτώσεις όπου δύο άτομα αξιολογούν τα ίδια αντικείμενα είτε βαθμολογώντας τα σε ορισμένη κλίμακα είτε συνηθέστερα, κατατάσσοντας τα σε αύξουσα σειρά .

Στην πρώτη περίπτωση, συνήθως οι δυνατές τιμές της κλίμακας είναι λίγες ώστε να μην μπορούμε σε καμία περίπτωση να θεωρήσουμε

ότι οι μεταβλητές είναι συχνές και άρα να υποθέσουμε κανονική κατανομή πληθυσμού. Γι'αυτό θα αντιστοιχήσουμε στις παρατηρήσεις τις τάξης και θα υπολογίσουμε τον συντελεστή rs. Στη δεύτερη περίπτωση, οι παρατηρήσεις είναι τάξεις και ο μόνος συντελεστής που μπορεί να υπολογιστεί είναι ο rs.

Όταν ο συντελεστής rs ανάμεσα σε δύο βαθμολογίες ή σε δύο κατατάξεις είναι υψηλός , τότε μπορούμε να δεχτούμε ότι τα δύο άτομα που βαθμολογούν ή κατατάσσουν χρησιμοποιούν τα ίδια κριτήρια αξιολόγησης .

Σημειώνεται ότι όταν η κλίμακα μέτρησης μας επιτρέπει να υπολογίσουμε και τους δύο συντελεστές συσχέτισης τότε , συνήθως στους συνήθως στους δύο συντελεστές οφείλονται συνήθως σε ακραίες τιμές οι οποίες επηρεάζουν πολύ τον rxy και λιγότερο τον rs .

Παράδειγμα

Ζητήσαμε από δύο προπονητές να δουν το φιλμ του τελευταίου αγώνα μπάσκετ μεταξύ δύο ομάδων και να βαθμολογήσουν την απόδοση των 10 παικτών σε μία κλίμακα από το 1 ως το 20 . Πήραμε τα ακόλουθα αποτελέσματα :

Προπονητής A Xi 4 8 20 12 14 15 18 20 10
Προπονητής B Yi 2 8 4 20 8 16 10 19 18 15

Για να υπολογίσουμε τον συντελεστή συσχέτισης κατά τάξεις αντιστοιχούμε τις τάξεις Wi στις παρατηρήσεις Xi και ομοίως τις τάξεις Vi στις παρατηρήσεις Yi .
Σημειώνεται ότι όταν στις παρατηρήσεις υπάρχουν επαναλαμβανόμενες τιμές αντιστοιχούμε τον αριθμητικό μέσο των τάξεων τις οποίες θα έπρεπε να καταλάβουν.

Wi 1 2 3 9,5 5 6 7 8 9,5 4
Yi 1 3,5 2 10 3,5 7 5 9 8 6

Στη συνέχεια υπολογίζουμε :

$C = n (n+1)^2/4 = 10 \cdot 11^2/4 = 302,5$
$\Sigma Wi^2 = \Sigma Vi^2 = 383,5$
$\Sigma ViWi = 375,5$

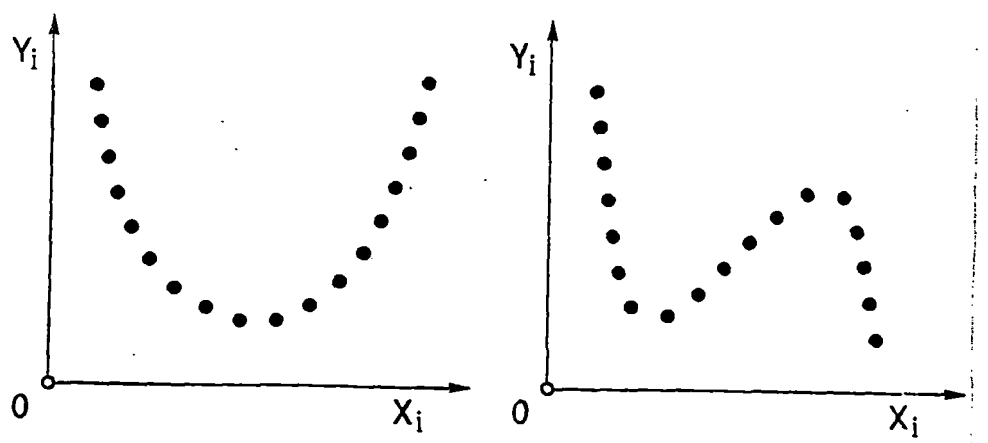
οπότε

$$r_S = \frac{\Sigma w_i v_i - C}{\sqrt{\Sigma w_i^2 - C}\sqrt{\Sigma v_i^2 - C}} = \frac{375,5 - 302,5}{383,5 - 302,5} = \frac{72,5}{81} = 0,895$$

που δείχνει ότι οι δύο προπονητές τείνουν να συμφωνούν στην αξιολόγηση των παικτών .

Έλεγχος των υποθέσεων.

Έστω ότι σε ορισμένο τυχαίο δείγμα η ζευγών παρατηρήσεων δύο μεταβλητών X και Y υπολογίσαμε το συντελεστή κατά τάξεις rs και μας ενδειαφέρει ο έλεγχος σημαντικότητας του rs . Η μηδενική υπόθεση, συνήθως εξειδικεύεται ως υπόθεση ανεξαρτησίας των X και Y παρόλο που ο σχετικός έλεγχος δεν ανιχνεύει μη μονοτονικές σχέσεις . Έτσι, αν οι παρατηρήσεις συνδέονται με μια μη μονοτονική σχέση όπως στα ακόλουθα σχήματα, θα υπολογίσουμε μια τιμή για τον rs ,η οποία δεν είναι στατιστικά σημαντική . Το διάγραμμα διασποράς μπορεί να είναι διαφωτιστικό και στην περίπτωση αυτή .



Αποδεικνύεται ότι όταν οι X και Y είναι ανεξάρτητες , τότε η κατανομή του στατικού

$$t = r_S \sqrt{\frac{n-2}{1-r_S^2}}$$

προσεγγίζει την κατανομή t- student με ν=n-2 βαθμός ελευθερίας . Η προσέγγιση είναι πολύ καλή για n ≥18 . Στον πίνακα 6 στο τέλος της εργασίας δίνονται ποσοστιαία σημεία από την ακριβή κατανομή του rs

για μικρές τιμές του n . Επομένως , ο έλεγχος σημαντικότητας του rs θα γίνει σταδιακά ως εξής :

- Η μηδενική και η εναλλακτική υπόθεση .

Ho : Οι X και Y είναι ανεξάρτητες
Hε : i) Οι Χκαι Y συνδέονται με μια μονοτονική σχέση.
ii) Οι X και Y συνδέονται με μια αύξουσα μονοτονική σχέση .
iii) Οι X και Y συνδέονται με μια φθίνουσα μονοτινική σχέση .

- Το κριτήριο αποφάσεως .

Για $n \geq 18$ Υπολογίζουμε

$$t_n = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

οπότε , απορρίπτουμε την Ho, στο επίπεδο α , για τα τρία είδη ελέγχου , αντίστοιχα , αν
i) $|t n| > t_{n-2}$, 9/2
ii) $t n > t_{n-2}$, α
iii) $t n < -t_{n-2}$, α

3.12 Οι συντελεστές αυτοσυσχέτισης

Οι παρατηρήσεις μιας τυχαίας μεταβλητής που συλλέγονται σε διαδοχικά ισαπέχοντα σημεία του χρόνου ή του χώρου αποτελούν μια **χρονική σειρά .** Στην ανάλυση χρονικών σειρών μας ενδιαφέρει να ελέγξουμε αν υπάρχει αλληλεξάρτηση μεταξύ διαδοχικών παρατηρήσεων οπότε παραβιάζεται μια από τις υποθέσεις του τυχαίου δείγματος όπως ορίστηκε στο πρώτο κεφάλαιο . Για το σκοπό αυτό υπολογίζονται οι συντελεστές συσχέτισης μεταξύ διαδοχικών παρατηρήσεων που ονομάζονται συντελεστές αυτοσυσχέτισης ή σειριακής συσχέτισης .

Ειδικότερα, έστω $X_1,X_2,...X_n$ μια χρονική σειρά . **Ο συντελεστής αυτοσυσχέτισης πρώτης τάξης ή συντελεστής σειριακής συσχέτισης σε υστέρηση 1,** συμβολίζεται με r_1 και είναι ο συντελεστής συσχέτισης που υπολογίζεται στα (n-1) ζεύγη παρατηρήσεων (X_2,X_1),(X_3,X_2),..

(X_n ,X_{n-1}). Συνήθως χρησιμοποιείται ο ακόλουθος , ελαφρώς τροποποιημένος τύπος .

$$r_1 = \frac{\sum_{i=2}^{n} (x_i - \bar{x})(x_{i-1} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (3.12.1)$$

Γενικά , ο συντελεστής αυτοσυσχέτισης κ τάξης συμβολίζεται με r_κ και υπολογίζεται ως εξής :

$$r_\kappa = \frac{\sum_{i=\kappa+1}^{n} (x_i - \bar{x})(x_{i-\kappa} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (3.12.2)$$

Παράδειγμα

Θα υπολογίσουμε τους συντελεστές αυτοσυσχέτισης πρώτης και δεύτερης τάξης για τις καταθέσεις όψεως Xi (σε δρχ) της Ελλάδας των ετών 1977-86.

Έτος	Xi	Xi-1	Xi-2
1997	32	-	-
1978	41	32	-
1979	51	41	32
1980	57	51	41
1981	73	57	51
1982	88	73	57
1983	105	88	73
1984	140	105	88
1985	168	140	105
1986	195	168	140

Πηγή : Μηνιαίος Στατιστικό Δελτίο της Τράπεζας της Ελλάδας , Οκτώβριος 1987.

Υπολογίζουμε : $\sum_{i=1}^{10} X_i = 950 \qquad \bar{X} = 95$

$$\sum_{i=2}^{10} (X_i - \bar{X})(X_{i-1} - \bar{X}) = 19405$$

$$\sum_{i=3}^{10} (X_i - \bar{X})(X_{i-2} - \bar{X}) = 10753$$

$$\sum_{i=1}^{10} (X_i - \bar{X})^2 = 28252$$

οπότε :

$$r_1 = \frac{\sum_{i=2}^{n} (X_i - \bar{X})(X_{i-2} - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{19.405}{28252} = 0,687$$

$$r_2 = \frac{\sum_{i=3}^{n} (X_i - \bar{X})(X_{i-2} - \bar{X})}{\sum (X_i - \bar{X})^2} = \frac{10753}{28252} = 0,38$$

- Χρήσιμη είαι η ακόλουθη παρατήρηση : Αν σε δύο χρονικές σειρές υπολογίζουμε υψηλούς συντελεστές αυτοσυσχέτισης τότε είναι σχεδόν βέβαιο ότι και ο συντελεστής της μεταξύ τους συσχέτισης θα είναι υψηλός .

Έλεγχος σημαντικότητας

Ο συντελεστής αυτοσυσχέτισης r_κ είναι δειγματικό ανάλογο του θεωρητικού συντελεστή αυτοσυσχέτισης κ τάξης ο οποίος συμβολίζεται με ρ_κ και ορίζεται για την στοχαστική διαδικασία παραγωγής των παρατηρήσεων $X_1,...,X_n$.

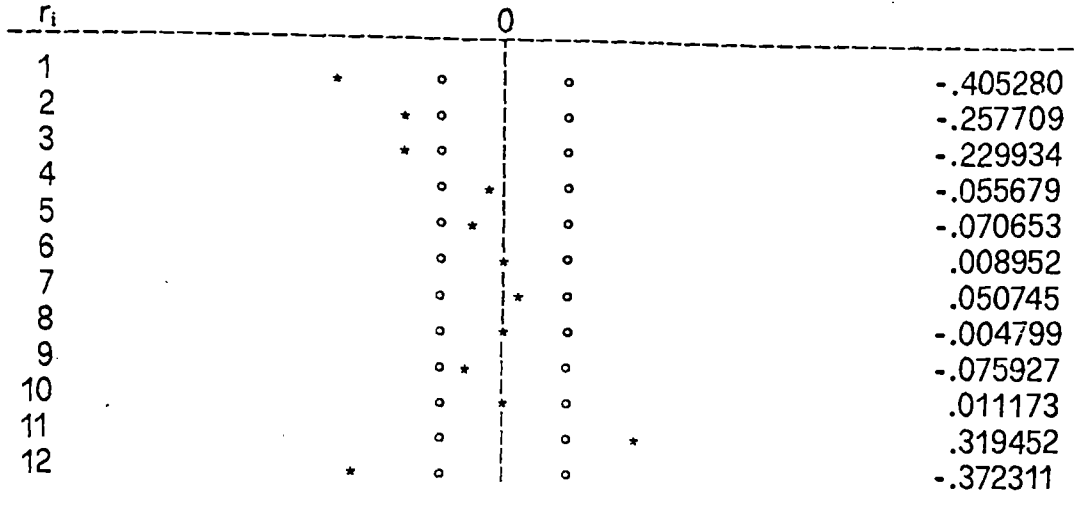
Ο r_κ είναι ένας εκτιμητής του ρ_κ και , αποδεικνύεται ότι , όταν οι διαδοχικές παρατηρήσεις της τυχαίας μεταβλητής X είναι ασυσχέτιστές, δηλαδή ισχύει :

$$\rho_\kappa = 0 \text{ για κάθε } \kappa > 0$$

τότε η διακύμανση του r_κ ισούται με 1/n Αν ,επιπλέον , το n είναι μεγάλο τότε η κατανομή του r_κ προσεγγίζει την κανονική .Ισοδύναμα , η κατανομή της $r_\kappa\sqrt{n}$ προσεγγίζει την N (0,1) . Έτσι ναι τιμή του (r_κ) μεγαλύτερη από το $2\sqrt{n}$ .. θα πρέπει να θεωρείται σημαντικά διαφορετική από το μηδέν .

Η γραφική παράσταση των r_κ , κ=1,2,.... στο διάστημα $\pm 2 / \sqrt{n}$

ονομάζεται **συνάρτηση αυτοσυσχέτισης** και μας επιτρέπει να δούμε αμέσως για ποίες τιμές του κ ο σχετικός συντελεστής είναι έξω από το διάστημα αυτό και άρα είναι στατιστικά σημαντικός . Η τυπική μορφή μιας τέτοιας συνάρτησης είναι όπως στο ακόλουθο σχήμα :



ΚΕΦΑΛΑΙΟ ΤΕΤΑΡΤΟ

Το πολλάπλο μοντέλο παλινδρόμησης

4.1 Εισαγωγή

Το απλό μοντέλο του προηγούμενου κεφαλαίου γενικεύεται στο **πολλαπλό ή πολυμεταβλητό γραμμικό μοντέλο παλινδρόμησης** , στο οποίο η μέση τιμή της εξαρτημένης μεταβλητής Y είναι γραμμική συνάρτηση των ερμηνευτικών μεταβλητών $X_1,...X_\kappa$. Έτσι , σε κάθε δείγμα παρατηρήσεων της Y αντιστοιχεί το n x κ σύστημα εξισώσεων :

$$Yi = b_o+b_1X_{1i}+....+b_kX_{ki}+\varepsilon_i \qquad i=1,...,n \qquad (4.1.1)$$

όπου X_{ij} , i=1,..,κ , j=1,...,n είναι η i-τιμή της ερμηνευτικής μεταβλητής X_j και ε_i ο όρος σφάλματος που αντιστοιχεί στην παρατήρηση Y_i . Μπορούμε να γράψουμε την (4.1.1) ισοδύναμα , ως εξής :

$$Y_i = b_oX_{oi}+b_1X_{1i}+....+b_kX_{ki}+\varepsilon_i \qquad i=1,...,n \qquad (4.1.2)$$

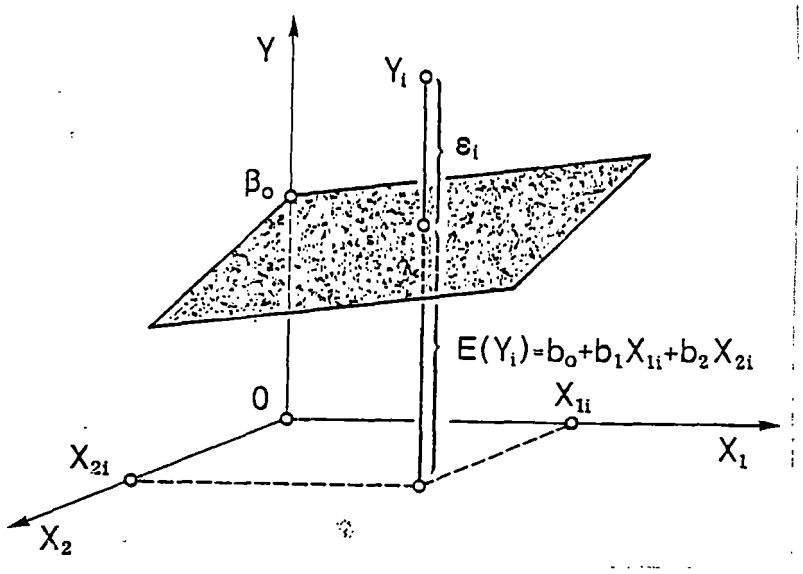
όπου $X_{oi}=1$, i=1,...,n

Σε ένα πρακτικό πρόβλημα , υποθέτοντας ότι ισχύει το μοντέλο αυτό υποθέτουμε ότι σε επαναληπτικές μετρήσεις στις οποίες οι τιμές των ερμηνευτικών μεταβλητών παραμένουν σταθερές και ίσες με $X_{1i},...,X_{\kappa i}$ αντιστοιχεί η τυχαία μεταβλητή Y_i . Η στοχασμική συμπεριφορά της Yi προσδιορίζεται από την αντίστοιχη συμπεριφορά του όρου σφάλματος ε_i ο οποίος είναι μια τυχαία ματαβλητή .

Υποθέτοντας ότι E (εi) =o έχουμε :

$$E (Y_i) =b_o+b_1X_{1i}+....+b_kX_{ki} , \quad i=1,...,n$$

που είναι η συνάρτηση παλινδρόμησης της Y επί των $X_1,X_2,...,X_n$ και ορίζει ένα επίπεδο στο χώρο των κ+1 διαστάσεων . Στο επόμενο σχήμα δίνεται το γράφημα μιας γραμμικής συνάρτησης παλινδρόμησης με δυο ερμηνευτικές μεταβλητές .



Στην ανάλυση παλινδρόμησης η εξαρτημένη ονομάζεται και **μεταβλητή ανταπόκρισης** . Επειδή η συνάρτηση πολλαπλής παλινδρόμησης ορίζει ένα επίπεδο ή επιφάνεια , το γράφημα της ονομάζεται και **επιφάνεια ανταπόκρισης** .

Το διάνυσμα των τιμών (X_{1i} ,.....,$X_{\kappa i}$) στις οποίες παρατηρείται η Yi ονομάζεται το **i- σημείο σχεδιασμού** .Μια τιμή Yi μαζί με το αντίστοιχο σημείο σχεδιασμού σημειώνονται ως (X_{1i},....,X_{ki},Y_i) και αναφέρεται ως το **i- σημείο των δεδομένων ή περίπτωση** .Η εκτίμηση των αγνώστων παραμέτρων b_o, b_1,....,b_k του μοντέλου θα γίνει από ένα σύνολο η (n>k) περιπτώσεων που θα αναφέρεται ως **δείγμα δεδομένων**.

4.2 Η εκτίμηση των παραμέτρων με τη μέθοδο των ελαχίστων τετραγώνων .

Έστω ότι παρατηρήσαμε την εξαρτημένη μεταβλητή Y σε n σημεία σχεδιασμού (X_{1i} ,...., $X_{\kappa i}$) όχι όλα ίδια μεταξύ τους και διαθέτουμε τις n περιπτώσεις (X_{1i} ,..., $X_{\kappa i}$, Y_i) i=1 ,..., n για τις οποίες υποθέτουμε ότι ισχύει :

$$Y_i = b_o+b_1X_{1i}+....+b_kX_{ki} \quad i=1,....,n$$

Όπως και στο μοντέλο, με την μέθοδο των ελαχίστων τετραγώνων ζητούμε να βρούμε τις τιμές bo, b1,....,bκ των παραμέτρων bo,b1,...,bκ αντίστοιχα , οι οποίες για τις n περιπτώσεις των δεδομένων μας ελαχιστοποιούν τη συνάρτηση

$$S=S(b_o,b_1,...,b\kappa)=\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \qquad (4.2.1)$$
$$=\sum_{i=1}^{n}(Y_i-b_o-b_1X_{1i}-,...,-b_kX_{ki})^2$$

Οι τιμές $\hat{b}_o,\hat{b}_1,...,\hat{b}_k$ που ελαχιστοποιούν την $S(b_o,b_1,...,b_k)$ ικανοποιούν το σύστημα εξισώσεων που προκύπτει όταν θέτουμε τις κ+1 μερικές παραγωγούς $\frac{\partial S}{\partial b_j}$, j-o,...,κ ίσες με μηδέν.
Έχουμε δηλαδή το σύστημα :

$$\begin{aligned}
\frac{\partial S}{\partial b_o} &= -2\,\Sigma\,[\,Y_i-(\hat{b}_o+\hat{b}_1X_{1i}+\hat{b}_2X_{2i}+.....+\hat{b}_kX_k)\,]=0\\
\frac{\partial S}{\partial b_1} &= -2\,\Sigma\,[\,X_{1i}\,[\,Y_i-(\hat{b}_o+\hat{b}_1X_{1i}+\hat{b}_2X_{2i}+....+\hat{b}_kX_{ki})]=0\\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad (4.2.2)\\
\frac{\partial S}{\partial b_\kappa} &= -2\,\Sigma\,X_{\kappa i}\,[\,Y_i-(\hat{b}_o+\hat{b}_1X_{1i}+\hat{b}_2X_{2i}+....+\hat{b}_kX_{ki})]=0
\end{aligned}$$

και ο τελεστής Σ είναι για i=1,...,n . Αν πολλαπλασιάσουμε με -½ κάθε εξίσωση του (4.2.2) το σύστημα αυτό γράφεται ισοδύναμα ως :

$$\begin{aligned}
&\hat{b}_{on}+\hat{b}_1\,\Sigma\,X_{1i}+\hat{b}_2\,\Sigma\,X_{2i}+....+\hat{b}_k\,\Sigma\,X_{ki}=\Sigma\,Y_i\\
&\hat{b}_o\,\Sigma\,X_{1i}+\hat{b}_1\,\Sigma\,X^2{}_{1i}+b_2\,\Sigma\,X_{1i}X_{2i}+.....+\hat{b}_k\,\Sigma\,X_{1i}X_{ki}=\Sigma\,X_{1i}Y_i\\
&\vdots \qquad\qquad \vdots \qquad\qquad (4.2.3)\\
&\hat{b}_o\,\Sigma\,X_{ki}+\hat{b}_1\,\Sigma\,X_{ki}X_{1i}+\hat{b}_2\,\Sigma\,X_{ki}X_{2i}+....+\hat{b}_k\,\Sigma\,X^2{}_{ki}=\Sigma\,X_{ki}\,Y_i
\end{aligned}$$

Το (4.2.3) είναι η γενική μορφή των κανονικών εξισώσεων για $\kappa \geq 1$. Για κ=1 είδαμε στο δεύτερο κεφάλαιο ότι είναι οι εξής (οι δέκτες παραλείπονται):

$$n\,\hat{b}_o+\hat{b}_1\,\Sigma X_1=\Sigma\,Y$$
$$b_o\,\Sigma\,X_1+\hat{b}_1\,\Sigma X^2{}_1=\Sigma X_1\,Y$$

ενώ για κ=2 οι κανονικές εξισώσεις είναι οι εξής :

$$n\,\hat{b}_o+\hat{b}_1\,\Sigma X_1+\hat{b}_2\,\Sigma X_2=\Sigma Y$$
$$\Sigma X_1\hat{b}_o+\hat{b}_1\,\Sigma X_1{}^2+\hat{b}_2\,\Sigma X_1{}^*X_2=\Sigma X_1\,Y$$
$$\Sigma X_2\hat{b}_o+\hat{b}_1\,\Sigma X_2X+\hat{b}2\,\Sigma X_2{}^2=\Sigma X_2\,Y$$

Έτσι,

$$X^T=\begin{bmatrix} 1 & 1 & ..... & 1\\ X_{11} & X_{12} & ..... & X_{1n}\\ X_{21} & X_{22} & ..... & X_{2n}\\ & & & \\ X_{k1} & X_{k2} & ..... & X_{kn}\end{bmatrix} \qquad \hat{b}=\begin{bmatrix}\hat{b}o\\ \hat{b}1\\ \\ \hat{b}k\end{bmatrix}$$

γράφουμε το σύστημα των κανονικών εξισώσεων συμπαγώς ως εξής :

$$\mathbf{X^T Y = (X^T X)\,\hat{b}}$$

(kxn)(nx1) (kxn)(nxk)(kx1)

και τη λύση του ως εξής :

$$\mathbf{\hat{b} = (X^T X)^{-1}\, X'Y} \qquad (4.2.4)$$

όπου $(X^TX)^{-1}$ η αντίστροφη του X^TX . Να σημειωθεί ότι η λύση του συστήματος των κανονικών εξισώσεων , δηλαδή το διάνυσμα που ορίζεται από την (4.2.4) υπάρχει αν και μόνον αν η (X^TX) είναι μη ιδιάζουσα δηλαδή έχει αντίστροφη .Αυτό εξασφαλίζεται όταν οι κ+1 στήλες της X είναι γραμμικώς ανεξάρτητες .

Η λύση αυτή οδηγεί σε μια προφανή διαδικασία υπολογισμού του διανύσματος $\hat{b}$.Υπολογίζουμε αρχικά τη X^TX , την αντίστροφη της $(X^TX)^{-1}$ και το διάνυσμα X^TY .Στη συνέχεια υπολογίζουμε το $\hat{b}$ πολλαπλασιάζοντας την $(X^TX)^{-1}$ με το διάνυσμα X^TY .

Όταν τα n και κυρίως το κ είναι μικρά ($n \leq 15$, $\kappa \leq 4$) αυτή η διαδικασία είναι εύκολη , και εκτός από το μέρος της αντίστροφης της X^TX, κατανοητή για το μέσο αναγνώστη . Η διαδικασία αυτή πάντως έχει σοβαρά υπολογιστικά μειονεκτήματα όπως αριθμητική αστάθεια και αριθμητικά σφάλματα . Τα γνωστά στατιστικά προγράμματα Η/Υ χρησιμοποιούν άλλες μεθόδους υπολογισμού του b . Η $(X^TX)^{-1}$ πάντως είναι πολύ σημαντική στην ανάλυση παλινδρόμησης και υπεισέρχεται σε τύπους που ορίζουν σημαντικές παραμέτρους του εκτιμηθέντος μοντέλου.

4.3 Οι συντελεστές μερικής παλινδρόμησης

Στην εξίσωση παλινδρόμησης

$$Y_i = b_o + b_1 X_{1i} + \ldots\ldots + b_k X_{ki} + \varepsilon_i$$

κάθε Παράμετρος b_j , j=1,...,κ ονομάζεται συντελεστής μερικής ή καθαρής παλινδρόμησης της X_j , j=1,...,κ και μετρά την αναμενόμενη ή μέση μεταβολή της εξαρτημένης μεταβλητής όταν η τιμή της ερμηνευτικής μεταβλητής μεταβληθεί κατά μια μονάδα και οι τιμές των υπολοίπων παραμείνουν σταθερές . Έτσι π.χ αν η τιμή της Χκ αυξηθεί κατά μία μονάδα , δηλαδή γίνει $X'_{ki} = X_{ki}+1$ τότε ισχύει :

$$E(Y_i') = b_o + b_1 X_{1i} + \ldots + b_k (X_{ki} + 1)$$

$$= b_o + b_1 X_{1i} + \ldots + b_k X_{ki} + b_k$$

$$= E(Y_i) + b_k$$

Η ερμηνεία των εκτιμήσεων b_j , $j=1,\ldots,\kappa$ είναι ανάλογη .

Παράδειγμα

Για την πιέση Υ ενός ενήλικα εκτιμήθηκε το ακόλουθο μοντέλο παλινδρόμησης με 20 παρατηρήσεις που έγιναν σε ισάριθμους ενήλικες τους οποίους επιλέξαμε με βάση την ηλικία τους X_1, σε χρόνια και το βάρος τους X_2, σε κιλά, από τα ενήλικα άτομα ίσου περίπου ύψους :

$$\hat{Y} = -65{,}10 + 0{,}425 X_1 + 1{,}077 X_2$$

Οι συντελεστές $\hat{b}_1 = 0{,}425$ και $\hat{b}_2 = 1{,}077$ ερμηνεύονται ως εξής : Στα άτομα ίδιου βάρους, κάθε επιπλέον χρόνο ηλικίας εκτιμούμε ότι αυξάνει την πιέση του αίματος κατά 0,425 μονάδες, ενώ τα άτομα ίδιας ηλικίας κάθε επί πλέον κιλό βάρους, αυξάνει την πιέση κατά 1,077 μονάδες. Η τιμή του σταθερού όρου δεν έχει κανένα νόημα, αφού είναι αδύνατον να υπάρχει στα δεδομένα μας παρατήρηση στο σημείο ($X_1 = 0$, $X_2 = 0$).

Μια άλλη διαδικασία υπολογισμού των $\hat{b}j$

Από τον ορισμό του συντελεστή μερικής παλινδρόμησης προκύπτει ότι στην εκτίμηση π.χ του συντελεστή $\hat{b}_j$.θα πρέπει να υπεισέρχονται μόνον οι μεταβολές της X_j . Αυτό σημαίνει

ότι αν η X_j σχετίζεται με τις υπόλοιπες θα πρέπει να υπεισέρχεται μονόν το μέρος της μεταβλητότητας της X_j το οποίο δεν περιέχεται στις υπόλοιπες. Αποδεικνύεται ότι η εκτίμηση ελαχίστων τετραγώνων $\hat{b}_j$ του συντελεστή b_j μετρά την μεταβολή στην εξαρτημένη από μια μοναδιαία μεταβολή της X_j όταν οι γραμμικές επιδράσεις των υπολοίπων ερμηνευτικών μεταβλητών έχουν αφαιρεθεί και από την X_j και από την Ψ. Έτσι π.χ στο μοντέλο με κ=2 ερμηνευτικές μεταβλητές δηλαδή το

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i \qquad i = 1, \ldots, n \qquad (4.3.1)$$

μπορούμε να εκτιμήσουμε έστω το συντελεστή b_1 ως εξής (ο όρος εκτίμηση αναφέρεται αποκλειστικά στην εκτίμηση ελαχίστων τετραγώνων)

- εκτιμούμε την παλινδρόμηση της Ψ επί της X_2 και έστω $\hat{\varepsilon}_{Y_2}$ τα κατάλοιπα (οι δείκτες για ευκολία παραλείπονται).

- εκτιμούμε την παλινδρόμηση της X_1 έπι της X_2 και έστω $\hat{\varepsilon}_{1_2}$ τα κατάλοιπα.

- εκτιμούμε την παλινδρόμηση των $\hat{\varepsilon}_{Y_2}$ επί των $\hat{\varepsilon}_{1_2}$

επομένως έχουμε :

$$\hat{b}_1 = \frac{\sum \hat{\varepsilon}_{Y_2} \hat{\varepsilon}_{1_2}}{\sum \varepsilon^2_{1_2}} \qquad (4.3.2)$$

Η τιμή του b που θα παρούμε με τον τρόπο αυτό είναι ίδια με την τιμή που θα πάρουμε αν εφαρμόσουμε τη μέθοδο των ελαχίστων τετραγώνων απ'ευθείας στο μοντέλο (4.3.1). Αποδεικνύεται εύκολα ότι η σχέση αυτή είναι ισοδύναμη με την:

$$\hat{b}_1 = \frac{\hat{b}_{y1} - \hat{b}_{y2}\hat{b}_{12}}{1 - r_{12}^2} \qquad (4.3.3)$$

όπου $\hat{b}_{y1}$ και $\hat{b}_{y2}$ η εκτίμηση του συντελεστή απλής παλινδρόμησης της Ψ επί της X_1 και Ψ επί της X_2 αντίστοιχα $\hat{b}_{12}$ η εκτίμηση του συντελεστή παλινδρόμησης της X_1 επί της X_2 και r_{12} ο συντελεστής συσχέτισης των παρατηρήσεων X_{1i} και X_{2i} i=1,....,n. Από τη σχέση (4.3.3) προκύπτουν ορισμένα ενδιαφέροντα συμπεράσματα :

Αν $r_{12} = 0$ οπότε και $\hat{b}_{12} = 0$ τότε $\hat{b}_1 = \hat{b}_{y1}$

Αν $r_{12} = 1$ τότε $\hat{b}_1$ είναι απροσδιόριστο.

Αν $r_{12} \cong 1$ οπότε και $\hat{b}_{21} \cong 1$ τότε $\hat{b}_{y1} \cong \hat{b}_{y2}$ και το $\hat{b}_1$ προσδιορίζεται από την αναλογία δύο πολύ μικρών αριθμών. Στην περίπτωση αυτή μικρές μεταβολές στους όρους $\hat{b}_{y1}$, $\hat{b}_{y2}$, $\hat{b}_{21}$, r_{12}^2 ... που υπεισέρχονται στον ορισμό του $\hat{b}_1$.θα είχε ως αποτέλεσμα μεγάλες μεταβολές στην τιμή του $\hat{b}_1$ Δηλαδή το $\hat{b}_1$ είναι αναξιόπιστος εκτιμητής του b_1 Σημειώνεται ότι όταν $r_{12} \cong 1$ οι X_1, X_2 ονομάζονται **ατελώς συγγραμικές .**

Συμπεραίνουμε ότι η τιμή του $\hat{b}_j$ που υπολογίζουμε στο πολλαπλό μοντέλο θα διαφέρει , γενικά ,από την αντίστοιχη τιμή του στην απλή παλινδρόμηση της Ψ επί της X_j.

4.4 Τα κατάλοιπα της εκτίμησης

Όπως και στο απλό μοντέλο, ορίζουμε τα κατάλοιπα ή σφάλματα εκτίμησης ως τις διαφορές των εκτιμήσεων $\hat{y}_i = \hat{b}_o + \hat{b}_1 x_{1i} + ... + \hat{b}_k x_k$ από τις παρατηρήσεις y_i, δηλαδή

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$
$$= Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \cdots - \hat{b}_k X_{ki} \qquad i = 1, \ldots, n$$

Αν ορίσουμε τα διανύσματα

$$\hat{\varepsilon} = \begin{bmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{bmatrix} \qquad \hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{bmatrix} \qquad \hat{b} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_k \end{bmatrix}$$

τότε ισχύει :

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{b}$$

Οι ιδιότητες των καταλοίπων ελαχίστων τετραγώνων τις οποίες είδαμε στο απλό μοντέλο γενικεύονται ως εξής :

- Το διάνυσμα των καταλοίπων είναι ορθογώνιο στο διάνυσμα των τιμών της κάθε ερμηνευτικής μεταβλητής x_j, $j=1,\ldots,\kappa$ και άρα και στη μήτρα σχεδιασμού. Δηλαδή ισχύει :

$$\sum_{i=1}^{n} x_{ji}\,\hat{\varepsilon}_i = 0 \qquad j = 1, \ldots, k \qquad (4.4.1)$$
$$\Leftrightarrow X^T \hat{\varepsilon} = 0$$

Πράγματι έχουμε :

$$X^T\hat{\varepsilon} = X^T(Y - X\hat{b}) = X^TY - X^TX\hat{b}$$
$$= X^TY - X^TX(X^TX)^{-1}X^TY$$
$$= X^TY - X^TY = 0$$

Ετσι, αν η πρώτη στήλη της μήτρας x είναι το μοναδιαίο διάνυσμα ισχύει

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

- Το διάνυσμα των καταλοίπων είναι ορθογώνιο στο διάνυσμα των εκτιμήσεων $\hat{y}$ δηλαδή ισχύει :

$$\hat{Y}^T \hat{\varepsilon} = \sum_i \hat{Y}_i \hat{\varepsilon}_i = 0 \qquad (4.4.2)$$

Πράγματι από την (4.4.1) έχουμε :

$$\hat{Y}^T \hat{\varepsilon} = \hat{b}^T X^T \hat{\varepsilon} = 0$$

Χρήσιμη είναι η εναλλακτική έκφραση του διανύσματος των καταλοίπων

$$\begin{aligned} \hat{\varepsilon} = Y - \hat{Y} &= Y - X\hat{b} \\ &= Y - X(X^TX)^{-1}X^TY \\ &= [I - X(X^TX)^{-1}X^T]Y \qquad (4.4.3) \\ &= [1 - P]Y \\ &= HY \end{aligned}$$

όπου $H=I-P$ και $P=x(x^Tx)^{-1}x^T$

Είναι εύκολο να διαπιστωθεί ότι οι μήτρες H και P είναι συμμετρικές, δηλαδή ισχύει $H=H^T$ και $P=P^T$ και ταυτοτικές δηλαδή ισχύει H H=H και P P=P. Εξάλλου, ισχύει

$$H\,X = (I - M)X = X - X(X^T \cdot X)^{-1}X^TX = X - X = 0$$

οπότε η (4.4.3) μπορεί να γραφεί ως εξής :

$$\hat{\varepsilon} = HY = H(Xb+\varepsilon) = HXb + H\varepsilon = H\varepsilon \qquad (4.4.4)$$

Σημείωση : Επειδή ισχύει PY=Xb=Y δηλαδή η μήτρα P μετασχηματίζει το διάνυσμα των παρατηρήσεων στο διάνυσμα των προβλέψεων, η **P** ονομάζεται **μήτρα πρόβλεψης**. Εξάλλου, επειδή HY=ε η μήτρα **H** ονομάζεται **μήτρα των καταλοίπων** (επειδή μετασχηματίζει τις παρατηρήσεις y στα κατάλοιπα ε).

Η μήτρα διακυμάνσεων - συνδιακυμάνσεων των καταλοίπων ορίζεται ως εξής :

$$E(\hat{\varepsilon}\hat{\varepsilon}^T) = \begin{bmatrix} Var(\hat{\varepsilon}_1) & Cov(\hat{\varepsilon}_1,\hat{\varepsilon}_2) & \dots & Cov(\hat{\varepsilon}_1,\hat{\varepsilon}_n) \\ Cov(\hat{\varepsilon}_1,\hat{\varepsilon}_2) & Var(\hat{\varepsilon}_1) & \dots & Cov(\hat{\varepsilon}_1,\hat{\varepsilon}_n) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\hat{\varepsilon}_1,\varepsilon_n) & Cov(\hat{\varepsilon}_1,\hat{\varepsilon}_n) & \dots & Var(\hat{\varepsilon}_n) \end{bmatrix}$$

και από την (4.4.4) προκύπτει ότι

$$\begin{aligned} E(\hat{\varepsilon}\hat{\varepsilon}^T) &= E(H\varepsilon\varepsilon^T H^T) \\ &= HE(\varepsilon\varepsilon^T)H^T \\ &= \sigma^2 H H^T \\ &= \sigma^2 H \quad \text{αφού η } H \text{ είναι ταυτοτική} \end{aligned}$$

Προκύπτει επομένως ότι :

$$Var(\hat{\varepsilon}_i) = \sigma^2 h_{ii} = \sigma^2(1-p_{ii}) \quad i=1,\dots,n \qquad (4.4.5)$$

$$Cov(\hat{\varepsilon}_i\varepsilon_j) = \sigma^2 h_{ij} = \sigma^2(1-p_{ij}) \quad i=1,\dots,n \quad j=1,\dots,n$$

όπου h_{ii} ,p_{ii} το i - διαγώνιο στοιχείο της μήτρας H και P, αντίστοιχα. Επειδή η διακύμανση είναι θετική θα πρέπει να ισχύει $p_{ii} < 1$, i=1,...,n. Από τις (4.4.5) προκύπτουν τα εξής ενδιαφέροντα συμπεράσματα :

- Η διακύμανση των καταλοίπων είναι μικρότερη από την διακύμανση των διαταρακτικών όρων και εξαρτάται από την μήτρα x.

- Η συνδιακύμανση των καταλοίπων δεν είναι μηδέν ακόμη και όταν οι διαταρακτικοί όροι έχουν μηδενική συνδιακύμανση.

4.5 Ιδιότητες των εκτιμητών ελαχίστων τετραγώνων

Εστω $\hat{b}$ το διάνυσμα των εκτιμητών ελαχίστων τετραγώνων δηλαδή

$b=(x^Tx)^{-1}x^Ty$

Αποδεικνύεται ότι, κάτω από τις συνθήκες Gauss - Markov το διάνυσμα των εκτιμητών b είναι **Γραμμικό** δηλαδή γραμμική συνάρτηση του τυχαίου διανύσματος $y=(y_1,y_2,...y_n)$ και **Αμερόληπτο** δηλαδή ισχύει

$E(\hat{b}) = b$

Για να προσδιορίσουμε την μεταβλητότητα του $\hat{b}$, σημειώνουμε, κατ' αρχήν τη μήτρα.

$$(\hat{b}-b)(\hat{b}-b)^T=\begin{bmatrix}(\hat{b}_0-b_0)^2 & (\hat{b}_0-b_0)(\hat{b}_1-b_1) & \cdots & (\hat{b}_0-b_0)(\hat{b}_k-b_k)\\ (\hat{b}_0-b_0)(\hat{b}_1-b_1) & (\hat{b}_1-b_1)^2 & \cdots & (\hat{b}_1-b_1)(\hat{b}_k-b_k)\\ \vdots & \vdots & & \vdots\\ (\hat{b}_0-b_0)(\hat{b}_k-b_k) & (\hat{b}_1-b_1)(\hat{b}_k-b_k) & \cdots & (\hat{b}_k-b_k)^2\end{bmatrix}$$

Και στη συνέχεια την :

$$Var(\hat{b})=E(\hat{b}-b)(\hat{b}-b)^T=\begin{bmatrix}Var(\hat{b}_0) & Cov(\hat{b}_0,\hat{b}_1) & \cdots & Cov(\hat{b}_0,\hat{b}_k)\\ Cov(\hat{b}_0,\hat{b}_1) & Var(\hat{b}_1) & \cdots & Cov(\hat{b}_1,\hat{b}_k)\\ \vdots & \vdots & & \vdots\\ Cov(\hat{b}_0,\hat{b}_k) & Cov(\hat{b}_1,\hat{b}_k) & \cdots & Var(\hat{b}_k)\end{bmatrix}$$

δηλαδή :

$$Var(\hat{b})=\sigma^2(X^TX)^{-1} \qquad (4.5.1)$$

και αυτή είναι η μικρότερη μεταβλητότητα την οποία μπορεί να έχει οποιδήποτε διάνυσμα γραμμικών και αμερόληπτων εκτιμητών του b, δηλαδή το διάνυσμα των εκτιμητών ελαχίστων τετραγώνων είναι και **άριστο**. Γι'αυτό όπως αναφέρθηκε ήδη σε προηγούμενο κεφάλαιο, οι εκτιμητές ελαχίστων τετραγώνων χαρακτηρίζονται συνοπτικά ως άριστοι γραμμικοί αμερόληπτοι ή BLUE (Best Linear Unbiased Estimator).
Το θεώρημα των Gauss - Markov γενικεύει τις παραπάνω ιδιότητες σε κάθε γραμμικό συνδυασμό των εκτιμητών ελαχίστων τετραγώνων. Ετσι, π.χ αν $C = 2b_1 + 4b_3 + b_4$ ένας γραμμικός συνδυασμός των παραμέτρων του μοντέλου τότε $C = 2b_1 + 4b_3 + b_4$ είναι άριστος , γραμμικός, αμερόληπτος ή BLUE εκτιμητής του C.
Οι εκτιμητές $\hat{b}_j$, $j = 0, 1, \ldots k$ είναι γραμμικές συναρτήσεις των παρατηρήσεων $y_1, y_2, \ldots, y_n$.Επομένως , αν ισχύει και η συνθήκη $\varepsilon \sim N_n(0, \sigma^2 I)$, δηλαδή αν

$$Y \sim N_n(Xb, \sigma^2 I)$$

τότε και η κατανομή του $\hat{b}$ είναι η κ - μεταβλητή κανονική δηλαδή ισχύει

$$\hat{b} \sim N_k(b, \sigma^2 (X^T X)^{-1})$$

Στην πράξη, η σ^2 είναι άγνωστη και την εκτιμούμε μέ την

$$S^2 = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n-k-1} = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n-k-1} \qquad (4.5.2)$$

όπου $\hat{\varepsilon}^T = (\hat{\varepsilon}_1, \hat{\varepsilon}_2 \ldots, \hat{\varepsilon}_n)$. το διάνυσμα των καταλοίπων.
Η (4.5.2) ορίζει ένα αμερόληπτο εκτιμητή της σ^2 όταν ισχύει η συνθήκη $E(\varepsilon\varepsilon^T) = \sigma^2 I$
Επειδή αντικαθιστούμε την σ^2 μέ την εκτίμηση της s^2 , οι έλεγχοι των υποθέσεων για τις τιμές των b_j , $j = 0, \ldots, 1$ γίνονται μέ το κριτήριο t, όπως θα δούμε στη συνέχεια.

Οι εκτιμητές μεγίστης Πιθανοφάνειας

Εστω ότι ισχύουν οι συνθήκες Cov (εi εj) = 0 $\forall i \neq j$ και $\varepsilon \sim N_n(0, \sigma^2 I)$ .. Η συνάρτηση πυκνότητας πιθανότητας των $Y = (Y_1, Y_2, \dots, Y_n)$ δίνεται από την

$$L(Y/b, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(Y-Xb)^T(Y-Xb)\right\}$$

Η L ως συνάρτηση των b και σ^2 ονομάζεται συνάρτηση πιθανοφάνειας και συμβολίζεται με L(b, σ^2/y). Ο νεπέρειος λογάριθμος της συνάρτησης πιθανοφάνειας ισούται με :

$$L = \ln L(b, \sigma^2/Y) = -\frac{n}{2}\ln 2\pi - \ln\sigma^2 - \frac{(Y-Xb)^T(Y-Xb)}{2\sigma^2} \quad (4.5.3)$$

Επειδή γενικά, ο lnz είναι αύξουσα συνάρτηση του z, οι τιμές των b και σ^2 που μεγιστοποιούν την L, είναι επομένως εκτιμήσεις μέγιστης πιθανοφάνειας . Είναι φανερό ότι η μεγιστοποίηση της **L** ισοδυναμεί με ελαχιστοποίηση της $(Y-Xb)^T(Y-Xb)$ **και** επομένως κάτω από την υπόθεση της κανονικότητας οι εκτιμητές ελαχίστων τετραγώνων $\hat{b}$ είναι και εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων b του μοντέλου
Αποδεικνύεται ότι ο εκτιμητής μέγιστης πιθανοφάνειας της σ^2 ορίζεται από το τύπο

$$\frac{1}{n}(Y-X\hat{b})^T(Y-Xb) = \frac{1}{n}\hat{\varepsilon}^T\hat{\varepsilon} \quad (4.5.4)$$

είναι επομένως μεροληπτικός, αφού ο αμερόληπτος εκτιμητής της σ^2 είναι ο

$$S^2 = \frac{1}{n-k-1}\hat{\varepsilon}^T\hat{\varepsilon}$$

4.6 Ανάλυση διακυμάνσεως

Είδαμε, στο απλό μοντέλο, ότι ισχύει:

$$SSR = \sum(Y_i - \bar{Y})^2 = \hat{b}_1^2\sum(X_i - \bar{X})^2 \quad \therefore \quad (4.6.1)$$

και επειδή

$$\hat{b}_1 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}$$

η (4.6.1) γράφεται, ισοδύναμα, ως εξής:

$$\begin{aligned} SSR &= \hat{b}_1 \Sigma(x_i-\bar{x})(y_i-\bar{y}) \\ &= \hat{b}_1 (\Sigma x_i y_i - n\bar{x}\bar{y}) \\ &= \hat{b}_1 \Sigma x_i y_i - n\bar{y}(\bar{y}-\hat{b}_0) \\ &= \hat{b}_1 \Sigma x_i y_i + \hat{b}_0 \Sigma y_i - n\bar{y}^2 = [\hat{b}_0\ \hat{b}_1] = \left[\frac{\Sigma y_i}{\Sigma x_i y_i}\right] - n\bar{y}^2 \end{aligned}$$

Γενικά αν b είναι το (κ+1)1 διάνυσμα των εκτιμήσεων ελαχίστων τετραγώνων, Ψ το n 1 διάνυσμα των παρατηρήσεων της εξαρτημένης και X η n (κ+1) μήτρα σχεδιασμού, τότε ισχύει :

$$SSR = \hat{b}^T X^T y - n\bar{y}^2 \qquad (4.6.2)$$

Εξάλλου μπορεί εύκολα να διαπιστωθεί ότι ισχύουν :

$$\begin{aligned} y^T y &= \Sigma y_i^2 \\ \hat{\varepsilon}^T \hat{\varepsilon} &= \Sigma \hat{\varepsilon}_i^2 \\ \hat{y}^T y &= \Sigma \hat{y}_i^2 \end{aligned}$$

οπότε γράφουμε

$$SST = \Sigma y_i^2 - n\bar{y}^2 = y^T y - n\bar{y}^2$$

και

$$\begin{aligned} SSE &= SST - SSR \\ &= y^T y - \hat{b}^T X y \\ &= \Sigma y_i^2 - \hat{b}_0 \Sigma y_i - \hat{b}_1 \Sigma y_i x_{1i} - \ldots - b_k \Sigma y_i x_{ki} \end{aligned}$$

Επομένως η γενική μορφή του πίνακα ανάλυσης διακυμάνσεως είναι η εξής:

SOURCE OF VARIATION	SUM OF SQUARES	D.F	MEAN SQUARE
REGRESSION	$SSR=b^TX^TY-_{ny}^2$	K	SSR/K
ERROR	$SSE=Y^TY-B^TX^TY$	n-k-1	SSE/n-k-1
TOTAL	$SST=Y^TY-n\bar{Y}^2$	n-1	SST/n-1

4.7 Οσυντελεστής R^2

Σε ένα μοντέλο μέ κ>1 ερμηνευτικές μεταβλητές ο συντελεστής R^2 ονομάζεται συντελεστής πολλαπλού προσδιορισμού και ορίζεται όπως και στο απλό μοντέλο παλινδρόμησης δηλαδή ως εξής :

$$R^2 = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST} \qquad (4.7.1)$$

όπου
και ο τελεστής Σ είναι για όλες τις τιμές i=1,...,n. Δηλαδή ο συντελεστής R2 είναι η ποσοστιαία μείωση του συνολικού τετραγωνικού σφάλματος εκτίμησης των παρατηρήσεων Yi, η οποία οφείλεται στην εισαγωγή των μεταβλητών x1,...,xκ στο μοντέλο.
Η θετική τετραγωνική ρίζα του R2 ονομάζεται **συντελεστής πολλαπλής συσχέτισης** και ισούται μέ τον συντελεστή

συσχέτισης των παρατηρήσεων Yi και των εκτιμήσεων $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 x_{1i} + \ldots + \hat{b}_k x_{ki}, i = 1, \ldots, n.$
Πράγματι έχουμε :

$$r_{Y\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum (Y_i - \bar{Y})^2}\sqrt{\sum (\hat{Y}_i - \bar{Y})^2}} \qquad (4.7.2)$$

και επειδή

$$\begin{aligned}\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) &= \sum [(\hat{Y}_i - \bar{Y}) + \hat{\varepsilon}_i](\hat{Y}_i - \bar{Y}) \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum \hat{\varepsilon}_i (\hat{Y}_i - \bar{Y}) \\ &= \sum (\hat{Y}_i - \bar{Y})^2\end{aligned}$$

η (4.7.2) γράφεται ως εξής :

$$r_{Y\hat{Y}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sqrt{\sum (Y_i - \bar{Y})^2}\sqrt{\sum (Y_i - \bar{Y})^2}} = \frac{\sqrt{\sum (Y_i - \bar{Y})^2}}{\sqrt{\sum (Y_i - \bar{Y})^2}} = \sqrt{R^2}$$

Σημείωση : Επειδή ο όρος SST και SSE είναι το άθροισμα των τετραγωνικών σφαλμάτων όταν για την εκτίμηση των παρατηρήσεων Yi χρησιμοποιούμε αντίστοιχα, το βασικό μοντέλο και το μοντέλο μέ ερμηνευτικές μεταβλητές τις x1,...,xκ μπορούμε να γράψουμε :
SST=SSEo
SSE=SSEκ

Ο δείκτης κ στο αντίστοιχο SSE δηλώνει το πλήθος των ερμηνευτικών μεταβλητών του μοντέλου. Ετσι ο συντελεστής προσδιορισμού γράφεται,ισοδύναμα,ως εξής :

$$R^2 = \frac{SSE_0 - SSE_k}{SSE_k} \qquad (4.7.2)$$

Ο διορθωμένος συντελεστής προσδιορισμού

Στους δύο όρους που προσδιορίζουν τον R^2 ο $SST=\sum(Y_i-\bar{Y})^2$ είναι σταθερός ενώ ο όρος $SSE=\sum(Y_i-\hat{b}_0-\hat{b}_1X_{1i}-\ldots-\hat{b}_kX_k$ εξαρτάται από το πλήθος των ερμηνευτικών μεταβλητών κ.
Στην πραγματικότητα μέ κάθε επί πλέον ερμηνευτική μεταβλητή στο μοντέλο, ο όρος SSE μείώνεται ή στην χειρότερη περίπτωση δεν μεταβάλλεται. Έτσι, μπορούμε να εκβιάσουμε υψηλή τιμή του R^2 προσθέτοντας απλώς ερμηνευτικές μεταβλητές στο μοντέλο. Αναφέρουμε ως ακραίο παράδειγμα ότι σε 2 ζεύγη τιμών προσαρμόζεται τέλεια ένα πολυώνυμο n-1 βαθμού επιτυγχάνεται δηλαδή $R^2=1$.
Είναι όμως προφανές ότι ένα τέτοιο μοντέλο δεν έχει καμία πραγματική χρησιμότητα.

Για να πάρουμε υπόψη και το πλήθος των παρατηρήσεων σε σχέση μέ τον αριθμό των ερμηνευτικών μεταβλητών του μοντέλου, υπολογίζουμε τον **διορθωμένο συντελεστή προσδιορισμού R** στον οποίο κάθε άθροισμα τετραγώνων διαιρείται μέ τους αντίστοιχους βαθμούς ελευθερίας. Δηλαδή έχουμε :

$$\bar{R}^2 = 1 - \frac{\sum(Y_i-\hat{Y}_i)^2/(n-k-1)}{\sum(Y_i-\bar{Y})^2/(n-1)} = 1 - \frac{S^2}{S^2_Y} \qquad (4.7.3)$$

όπου S η διακύμανση των παρατηρήσεων Ψi γύρω από το εκτιμηθέν μοντέλο και s... η διακύμανση των Ψi γύρω από τη μέση τιμή τους .

Ο διορθωμένος συντελεστής προσδιορισμού μας επιτρέπει να συγκρινούμε διαφορετικά μοντέλα που εκτιμούμε όχι μόνο για τις ίδιες παρατηρήσειςYi αλλά και για διαφορετικές. Θα πρέπει πάντως να σημειωθεί ότι

υπάρχουν επιφυλάξεις για την τελευταία χρήση , οι οποίοι θεωρούν ότι στην σύγκριση μοντέλων που εκτιμήθηκαν μέ διαφορετικά δεδομένα ο $\bar{R}^2$, μπορεί να χρησιμοποιθεί μόνο ως ένας αρχικός, χονδρικός δείκτης και προτείνουν άλλα μέτρα της ερμηνευτικής ικανότητας του μοντέλου.

Είναι εύκολο να δειχθεί ότι οι συντελεστές $\bar{R}^2$ και R^2 συνδέονται.

Πράγματι, αντικαθιστώντας την (4.7.1) στην (4.7.3) παίρνουμε:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1} \qquad (4.7.4)$$

Από τον τύπο αυτό προκύπτει ότι :

- Για $\kappa \geq 1$ ισχύει $\bar{R}^2 \leq R^2$. Η ισότητα ισχύει μόνο όταν υπάρχει τέλεια προσαρμογή, δηλαδή $R^2=1$, ή ασυμπτωματικά, δηλαδή όταν το μέγεθος του δείγματος αυξάνει απεριόριστα.
- Ο $\bar{R}^2$ μπορεί να πάρει και αρνητικές τιμές (ενώ, θυμίζουμε, ότι ισχύει $0 \leq R^2 \leq 1$). Αυτό συμβαίνει όταν

$$R^2 < \frac{k}{n-1}$$

Αποδεικνύεται ότι αν σε ένα μοντέλο προσθέσουμε μία ερμηνευτική μεταβλητή έστω X_j, ενώ όλα τα υπόλοιπα παραμείνουν σταθερά τότε ο $\bar{R}^2$ θα αυξηθεί αν η τιμή του συντελεστή $\hat{b}j$ είναι μεγαλύτερη από ένα τυπικό σφάλμα S_{bj}.

Ερμηνεία του R^2

Είδαμε ότι ο R^2 είναι η ποσοστιαία μείωση του ολικού σφάλματος εκτίμησης των παρατηρήσεων Yi που οφείλεται στην εισαγωγή των μεταβλητών X_1,.....,Χκ στο μοντέλο. Επειδή υπολογίζεται εύκολα (με τη βοήθεια του Η/Υ φυσικά)

συνοδεύει σχεδόν πάντα το εκτιμηθέν μοντέλο μαζί με τα τυπικά σφάλματα των συντελεστών μερικής παλινδρόμισης.

Η ερμηνεία των ακραίων τιμών του R^2 είναι προσφανής. Συγκεκριμένα όταν το εκτιμηθέν μοντέλο προσαρμόζεται τέλεια στα δεδομένα, δηλαδή όταν $Yi=\hat{Y}i$ $i=1,...,n$ τότε SSE=0 και R=1(η περίπτωση αυτή είναι, φυσικά, εντελώς απίθανη, στην πράξη). Όταν η εξαρτημένη δεν συνδέεται με τις ερμηνευτικές , δηλαδή ισχύει $b1 = ... = b\kappa = O$ και στο μοντέλο περιλαμβάνεται μόνο ο σταθερός όρος τότε , καμιά μείωση του ολικού σφάλματος δεν επιτυγχάνεται και ισχύει
SSE = SST οπότε $R^2 = O$. Οι ενδιάμεσες τιμές ερμηνεύονται αναλόγως. Η ευκολία με την οποία ερμηνεύεται η τιμή του R^2 οδηγεί συχνά σε μια μηχανιστική αξιολόγηση του μοντέλου , με βάση τον συντελεστή προδιορισμού. Μπορούμε όμως να αποφύγουμε πολλές από τις παγίδες που στήνει μια τέτοια πρακτική , έχοντας υποψιν τα εξής :

• Ας υποθέσουμε ότι οι X_1 , ... , X_K δεν συνδέονται με την Y, ισχύει δηλαδή $b_1 = ...= b_k = O$ εμείς όμως αυτό δεν το γνωρίζουμε και τις έχουμε περιλάβει στο μοντέλο. Τότε , λόγω της τυχαιότητας των παρατηρήσεων , η τιμή του R μπορεί να μην είναι μηδέν. Αποδεικνύεται ότι , κάτω από ασθενείς υποθέσεις (παρατηρήσεις Y ασυσχέτιστες , με κατανομή συμμετρική) ισχύει:

Έτσι , π.χ για η = 15 και $\kappa = 7$ η αναμενόμενηβ τιμή του R ισούται με 0,5 από καθαρή τύχη ! Γι'αυτό χρήσιμος είναι ο ακόλουθος πρακτικός κανόνας : **ερμηνεύουμε μόνο το μέρος της τιμής του R που είναι μεγαλύτερο από την αναμενόμενη τιμή του .**

• Ο συντελεστής R^2 είμαι μέτρο καλής προσαρμογής του εκτιμηθέντος μοντέλου στα δεδομένα . Δεν μας δίνει όμως καμιά πληροφορία για την προβλεπτική του ικανότητα ιδίως όταν η πρόβλεψη γίνεται για τιμές των ερμηνευτικών μεταβλητών έξω από τα όρια των δεδομένων μας.

• Ο συντελεστής R^2 μπορεί , θεωρητικά , να πάρει τη μέγιστη τιμή του 1 όταν στα δεδομένα δεν υπάρχουν επαναλαμβανόμενες τιμές της ανεξάρτητης τιμής. Όταν έχουμε περισσότερες από μία παρατηρήσεις στο ίδιο σημείο σχεδιασμού (X , ... , X) τότε η μεταβλητότητα των Y δεν οφείλεται στην μεταβλητότητα των X, είναι δηλαδή **καθαρό σφάλμα** οπότε το R δεν μπορεί να πάρει την τιμή 1 .

• Υψηλές τιμές του R^2 μπορούν να προκύψουν σε περιπτώσεις που παραβιάζεται μια ή περισσότερες από τις συνθήκες Gauss - Markov. Έτσι π.χ σχεδόν σε κάθε μοντέλο παλινδρόμησης το οποίο εκτιμάται με δεδομένα χρονικών σειρών , παρατηρείται εξαιρετικά υψηλός συντελεστής προσδιορισμού ενώ συγχρόνως δεν απορρίπτεται η υπόθεση για ύπαρξη αυτοσυσχέτισης των διαταρακτικών όρων. Το ίδιο μπορεί να συμβαίνει λόγω της ύπαρξης μιας ή περισσοτέρων ακραίων τιμών στα δεδομένα μας. Μπορούμε να αποφύγουμε λανθασμένα συμπεράσματα , όταν ερμηνείουμε τον R^2, αν συγχρόνως αναλύουμε τα κατάλοιπα του μοντέλου.

4.8 Οι συντελεστές μερικού προσδιορισμού

Ο συντελεστής πολλαπλού προσδιορισμού R^2 μετρά την ικανότητα όλων μαζί των ερμηνευτικών μεταβλητών X_1 , ... , Χκ να ερμηνεύσουν τη μεταβλητότητα της εξαρτημένης Y. Δεν μας λέει όμως πόση από την ικανότητα αυτή μπορεί να αποδοθεί σε κάθε μια από τις, X_1 , ... , Χκ χωριστά. Για το σκοπό αυτό υπολογίζουμε τους συντελεστές μερικού προσδιορισμού.

Ο συντελεστής μερικού προσδιορισμού της Y επί μιας μεταβλητής έστω της Χκ ή **Τετραγωνικός Συντελεστής Πολλαπλής Συσχέτισης στις Y και Χκ** και συμβολίζεται με $R^2_{Yk/1,2,...,k-1}$.
και υπολογίζεται ως εξής :

α. Εκτιμούμε την παλινδρόμηση της Y επί των X_1, ... , Χκ-1 και έστω

Y_i, $i = 1,...n$ τα κατάλοιπα.

β. Εκτιμούμε την παλινδρόμηση της Χκ επί των X_1 , ... , X_{k-1}
και έστω X_k , $i = 1, ... , n$ τα κατάλοιπα.

γ. Εκτιμούμε την παλινδρόμηση της Y επί της Χκ . Οσυντελεστής προσδιορισμού αυτής της εκτίμησης είναι ο $R^2_{Y_k/1,2,\ldots,k-1}$

Από τον τρόπο υπολογισμού του προκύπτει ότι ο $R^2_{Y_k/1,2,\ldots,k-1}$ μετρά το ποσοστό της μεταβλητότητας της Y η οποία μένει ανερμήνευτη ή κατάλοιπη από τις X_1, ... , Χκ-1 και ερμηνεύευται από την κατάλοιπη μεταβλητότητα της Χκ.
Ο $R^2_{Y_k/1,2,\ldots,k-1}$ υπολογίζεται και διαφορετικά ως εξής:

• Εκτιμούμε την παλινδρόμηση της Y επί των X_1 , ... , Χκ-1 και έστω SSE_{K-1} το άθροισμα των τετραγωνικών σφαλμάτων.

• Εκτιμούμε την παλινδρόμηση της Y επί των X_1 , ... , X_{k-1}, X_k και έστω SSE_K το άθροισμα των τετραγωνικών σφαλμάτων . Η διαφορά $SSE_{k-1} - SSE_k = SSR(X_k/X_1,\ldots,X_{k-1})$ είναι η μείωση του ολικού σφάλματος εκτίμησης λόγω της επί πλέον εισαγωγής της Χκ στο μοντέλο στο οποίο ήδη υπήρχαν οι $X_1,\ldots,X_{k-1}$

• Οσυντελεστής μερικού προσδιορισμού της Y επί της Χκ είναι η ποσοστιαία μείωση του ολικού σφάλματος εκτίμησης λόγω της επιπλέον εισαγωγής της Χκ στο μοντέλο. Δηλαδή

$$R^2_{Y_k/1,2,\ldots,k-1} = \frac{SSE_{k-1} - SSE_k}{SSE_{k-1}} = \frac{SSR(X_k/X_1\ldots X_{k-1})}{SSE_{k-1}} \qquad (4.8.1)$$

Παράδειγμα

Στο ακόλουθο μοντέλο Υ είναι η ζήτηση χρήματος σε δισ. δρχ., Χ το ΑΕΠ σε δισ. δρχ. και Χ το επιτόκιο καταθέσεων ταμιευτηρίου.Η εκτίμηση έγινε με βάση τις 17 ετήσιες παρατηρήσεις της Ελληνικής Οικονομίας για τα έτη 1970-1986 που δημοσιεύτηκαν στο Δελτίο της Τράπεζας της Ελλάδας, το 1987

Y= -0,3716+0,353X +5,345X

Δίνονται ακόμη

SSE=0,2984
SST=1,1146

απ' όπου υπολογίζουμε :SSR=SST-SSE=1,1146-0,2984=0,8162 και

$R^2 = \frac{0,8162}{1,1146} = 0,7323$

$R^2 = 1 - \frac{0,2984/14}{1,1146/16} = 0,694$

Εκτιμήθηκε ακόμη το απλό μοντέλο παλινδρόμησης με ερμηνευτική μεταβλητή την X_1 ως εξής :

$$Y = -1,033+3,54X_1$$

με SSE = 0,392

οπότε υπολογίζουμε :

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{0,392}{1,1146} = 0,6487$$

Χρησιμοποιώντας τον συμβολισμό που υιοθετήθηκε προηγουμένως σημειώνουμε :

$SSE_1 = 0,392$

$SSE_2 = 0,2984$

συντελεστής μερικού προσδιορισμού της Y επί της X_2 υπολογίζεται ως εξής :

$$R^2_{Y2/1} = \frac{SSE_1 - SSE_2}{SSE_1} = \frac{0,392 - 0,2984}{0,392} = 0,239$$

Το απλό μοντέλο της Y επί της X_2 εκτιμήθηκε ως εξής :

$$Y = -0,2876 + 5,85X2$$

με SSE = 0,3121

Ο μερικός συντελεστής προσδιορισμού της Y επί της X_1 υπολογίζεται ως εξής :

$$R^2_{Y1/2} = \frac{0,3121 - 0,2984}{0,3121} = 0,0441$$

Ο συντελεστής μερικής συσχέτισης

Ο συντελεστής μερικής συσχέτισης των μεταβλητών Y και X_κ με σταθερές τις $X_1, X_2, \ldots, X_{\kappa-1}$ συμβολίζεται με $r_{yk/1,\ldots,k-1}$ και ισούται με τη θετική τετραγωνική ρίζα του αντίστοιχου συντελεστή μερικού προσδιορισμού , δηλαδή :

$$r_{Yk/1\ldots k-1} = \sqrt{R^2_{Yk/1\ldots k-1}}$$

Από τον τρόπο που υπολογίζεται προκύπτει ότι ο μερικός συντελεστής συσχέτισης $r_{yk/1,\ldots,k-1}$ μετρά την ένταση της γραμμικής συμμεταβολής ανάμεσα στην εξαρτημένη και την ερμηνευτική μεταβλητή X_κ μετά την αφαίρεση των γραμμικών επιδράσεων των ερμηνευτικών $X_1, \ldots, X_{\kappa-1}$ και στην Y και στην X_κ

4.9 Έλεγχος των στατιστικών υποθέσεων

Όταν εκτιμούμε ένα μοντέλο παλινδρόμησης θέλουμε να ελέγξουμε αν μια ή περισσότερες από τις ερμηνευτικές μεταβλητές είναι πλεονάζουσες , δηλαδή μπορούν να αφαιρεθούν χωρίς σημαντική

μείωση της ερμηνευτικής του ικανότητας . Υπάρχουν δύο περιπτώσεις που μία ερμηνευτική μεταβλητή, έστω η $X_κ$, πλεονάζει :

α. Όταν η $X_κ$ δεν συνδέεται με την Y

β. Όταν η ερμηνευτική ικανότητα της X_j περιέχεται στις υπόλοιπες , οπότε η επί πλέον ερμηνευτική της ικανότητα (επομένως και ο συντελεστής $R^2_{yk/1,...,k-1}$) ισούται με μηδέν ή σχεδόν . Η περίπτωση αυτή θα γίνει ευκολότερα κατανοητή με το ακόλουθο παράδειγμα .

Παράδειγμα

Έστω η συνάρτηση παλινδρόμησης

$$E (Y_i) = 2X_1+X_2+7X_3$$

και ισχύει $X_3=X_1+X_2$. Τότε η E (Yi) μπορεί να γράφει, ισοδύναμα, με έναν από τους ακόλουθους τρόπους :

$$E (Y_i)=9X_1+8X_2$$

$$E (Y_i)=X_1+8X_3$$

$$E (Y_i)=9X_3-X_2$$

Στο μοντέλο αυτό , οποιοδήποτε ερμηνευτικών μεταβλητών ζεύγος ερμηνευτικών μεταβλητών κάνει την Τρίτη περιττή παρόλο που αυτή συνδέεται με την εξαρτημένη . Οι μεταβλητές που συνδέονται **συγγραμικές** . Δύο ή περισσότερες συγγραμμικές ερμηνευτικές μεταβλητές κάνουν αδύνατη την εκτίμηση των παραμέτρων με τη μέθοδο των ελαχίστων τετραγώνων . Στην πράξη , συνηθέστερη είναι η περίπτωση που δύο η περισσότερες ερμηνευτικές μεταβλητές είναι **σχεδόν συγγραμμικές** δηλαδή συνδέονται με μία σχέση σχεδόν γραμμικής με την έννοια ότι ο αντίστοιχος συντελεστής πολλαπλής συσχέτισης είναι κοντά στο 1 . Τα αποτελέσματα της ατελούς συγγραμμικότητας δεν είναι προφανή γι'αυτό και η ανίχνευση της δεν είναι εύκολη .Στη συνέχεια θα παρουσιάσουμε τη διαδικασία ελέγχου της μηδενικής υπόθεσης ότι μια ή περισσότερες από τις ερμηνευτικές μεταβλητές του μοντέλου πλεονάζουν είτε λόγω συγγραμμικότητας είτε διότι δεν έχουν σχέση με την εξαρτημένη .

Η μηδενική και η ενναλακτική υπόθεση του ελέγχου εξειδικεύονται ως εξής :

$H_o=b_{p+1}=....=b_k=o$

H_ε = μια τουλάχιστον από τις $b_j = 0$, $j = p+1,...,\kappa$

Στην Ηε αντιστοιχεί το μοντέλο

$$Y = b_o+b_1X_1+....+b_pX_p+b_{p+1}X_{p+1}+.....+b_kX_k+\varepsilon$$

το οποίο ονομάζεται **πλήρες ή εναλλακτικό** ενώ στην Ηο αντιστοιχεί το μοντέλο

$$Y=b_o+b_1X_1+.....+b_pX_p+\varepsilon$$

που ονομάζεται **αναγμένο ή μηδενικό ή μοντέλο κάτω από τον γραμμικό περιορισμό** $b_{p+1}=.....= b_k=o$. Ο έλεγχος της μηδενικής υπόθεσης ισοδυναμεί με έλεγχο της υπόθεσης ότι οι μεταβλητές $X_{p+1},....,X_k$ πλεονάζουν, δηλαδή η αφαίρεση τους απ το μοντέλο δεν μειώνει σημαντικα την ερμηνευτική του ικανότητα .

Το κριτήριο αποφάσεως ορίζεται με βάση τα αθροίσματα των τετραγωνικών σφαλμάτων των δύο μοντέλων ως εξής :
Έστω SSEκ και SSEp το άθροισμα των τετραγωνικών σφαλμάτων του πλήρους και του μειωμένου μοντέλου αντίστοιχα . Επειδή p<κ και κάθε επιπλέον μεταβλητή δεν μπορεί να μειώνει το SSE ισχύει SSEp≥SSEκ. Η διαφορά των δύο SSE είναι η αύξηση της ερμηνευτικής ικανότητας του μοντέλου με την επιπλέον εισαγωγή των μεταβλητών Xp+1,....,Xκ και συμβολίζουν με SSR ($X_{p+1},...,X_\kappa / X_1,....,X_p$) Δηλαδή έχουμε :

SSR ($X_{p+1},....,X_\kappa / X_1,....,X_p$)=SSE_p-SSE_κ

Αποδεικνύεται ότι , όταν οι διαταρακτικοί όροι ακολουθούν την κανονική κατανομή έχουν μέση τιμή μηδέν και είναι ανά δύο ασυσχέτιστοι (που λόγω της κανονικότητας των εi , συνεπάγεται ότι είναι και ανεξάρτητες) τότε το στατιστικό

$$F = \frac{SSE_p - SSE_k / (k-p)}{SSE_k / (n-k-1)} \qquad (4.9.1)$$

ακολουθεί την κατανομή $F_{v1,v2}$ με v_1=k-p και v_2=n-k-1 βαθμούς ελευθερίας . Έτσι , $F_{v1,v2,\alpha}$ είναι η τιμή της $F_{v1,v2}$ για την οποία ισχύει :

$$P (F_{v1,v2} > F_{v1,v2,\alpha})= \alpha$$

και F_π η τιμή του στατιστικού F που υπολογίζεται με βάση την (4.9.1) εφαρμόζουμε το ακόλουθο κριτήριο αποφάσεως :
Αν $F_\pi > F_{v1,v2,\alpha}$ απορρίπτουμε την Ηο στο α=0,05

4.10 Ο ολικός έλεγχος σημαντικότητας του μοντέλου

Στον έλεγχο αυτό συγκρίνουμε το πλήρες μοντέλο των κ ερμηνευτικών μεταβλητών με το μειωμένο μοντέλο στο οποιο υπάρχει μόνο ο σταθερός όρος . Ειδικότερα έχουμε :

Η μηδενική και η εναλλακτική υπόθεση

Ho:b_1=....=b_κ=0
Ηε : ένα τουλάχιστον bj ≠ 0, j=1,...,κ

Το μειωμένο είναι το μοντέλο με μονό το σταθερό όρο ή βασικό μοντέλο με αντίστοιχο ολικό άθροισμα τετραγώνων το SSEo=Σ (Yi-Y)^2 το οποίο συμβολίσαμε ως SST.

Κριτήριο αποφάσεως

Αν $$F_n = \frac{(SST-SSE_k)/k}{SSE_k/(n-k-1)} = \frac{SSR(X_1, X_2, \ldots, X_k/X_0)/k}{SSE_k/(n-k-1)} > F_{v_1, v_2, \alpha}$$

απορρίπτουμε την Ho στο επίπεδο σημαντικότητας της α . Σημειώνεται ότι v_1= και v_2=n-κ-1.

Όταν απορρίπτουμε την Ho στον έλεγχο αυτό λέμε ότι εκτιμήσαμε μια στατιστικά σημαντική παλινδρόμηση . Αυτό σημαίνει ότι οι ερμηνευτικές μεταβλητές X1,....,Xκ βελτιώνουν σημαντικά τις εκτμήσεις των παρατηρήσεων Yi, σε σχέση με το βασικό μοντέλο . Η σημαντικότητα εδώ θα πρέπει νε εννοηθεί ως στατιστική σημαντικότητα . Δηλαδή , η αύξηση της ερμηνευτικής ικανότητας του μοντέλου είναι μεγαλύτερη απ εκείνη που θα αποδίδουμε στην τύχη στο 100(1-α)% των ανάλογών δειγμάτων δεδομένων με το ίδιο η και την ίδια μήτρα X .

Η τιμή του στατιστικού F, σ'συτό τον έλεγχο, μπορεί να υπολογιστεί απ'ευθείας από το συντελεστή προσδιορισμού R^2 . Σημειώνεται ότι στον ορισμό του συντελεστή R2 επειδή δεν υπήρχε ανάγκη να διακρίνουμε μεταξύ διαφόρων μοντέλων δεν χρησιμοποιήσαμε δείκτες στα αντίστοιχα αθροίσματα τετραγώνων . Για να κρατήσουμε αντίστοιχα με το συμβολισμό που χρησιμοποιήσαμε στο μέρος αυτό συμβολίζουμε με SSE_κ το άθροισμα των τετραγωνικών

σφαλμάτων. Του μοντέλου με κ>=1 ερμηνευτικές μεταβλητές ενώ τη διαφορά $SST-SSE_\kappa$ συμβολίζουμε ως $SSR_\kappa=SSR(x_1,...x_\kappa/x_0)$.
Έτσι έχουμε:

$$\left.\begin{aligned} R^2 &= \frac{SSR_\kappa}{SST} \\ 1-R^2 &= \frac{SSE_\kappa}{SST} \end{aligned}\right\} \Rightarrow \frac{R^2}{1-R^2} = \frac{SSR_\kappa}{SSE_\kappa} \Rightarrow \frac{R^2/\kappa}{1-R^2/n-\kappa-1} =$$

$$= \frac{SSR_\kappa/\kappa}{SSE_\kappa/n-\kappa-1} = \frac{(SST-SSE_\kappa)/\kappa}{SSE_\kappa/n-\kappa-1} = F_n \qquad (4.10.1)$$

Πάντως, σπανίως θα χρειαστεί να χρησιμοποιήσουμε την (4.10.1) αφού όλα τα προγράμματα Η/Υ τυπώνουν, στον πίνακα ανάλυσης διακυμάνσεως, την F_π ως F-value. Μαζί δίνουν την ρ-value η τιμή της πιθανότητας του ελέγχου που είναι η πιθανότητα, όταν ισχύει η Ho, να πάρουμε μία τιμή για την F όση η F_π ή μεγαλύτερη. Αν η p-value είναι μικρότερη από το α (συνήθως ίσο με 0,05) απορρίπτουμε την Ho.

Προβλήματα του ελέγχου

Ο έλεγχος αυτός, όπως και όλοι οι ολικοί έλεγχοι, είναι ελάχιστα χρήσιμος αφού μας δίνει λίγες πληροφορίες για το αν εξειδικεύσαμε σωστά το μοντέλο. Έτσι, όταν απορρίπτουμε την Ho απλώς δεχόμαστε ότι ένας τουλάχιστον από τους μερικούς συντελεστές παλινδρόμησης είναι στατιστικά σημαντικός αλλά δεν ξέρουμε ποιος. Είναι μάλιστα πιθανό, στον έλεγχο για κάθε bj χωριστά, κανένας να μην είναι στατιστικά διαφορετικός από μηδέν. Απ'αυτό μπορούμε απλώς να συμπεράνουμε ότι υπάρχει κάποια (ατελής) συγγραμμικότητα μεταξύ των ερμηνευτικών μεταβλητών που κάνει την οριακή συμβολή της κάθε μιας στατιστικά ασήμαντη. Κρατώντας τις αναλογίες, είναι σαν να χρησιμοποιήσουμε π.χ. 4 άτομα ίσης σωματικής ισχύος για να μεταφέρουμε ορισμένο βάρος το οποίο μπορούν να μεταφέρουν 2 άτομα.

Θα πρέπει εδώ να σημειωθεί ακόμη ότι η απόρριψη της Ho δεν είναι αρκετή για να πάρουμε από το μοντέλο ικανοποιητικές εκτιμήσεις των Yi. Έχει δειχθεί ότι για να πάρουμε ικανοποιητικές εκτιμήσεις θα πρέπει να ισχύει $F\pi>4F_{\nu_1,\nu_2,\alpha}$.

ΚΕΦΑΛΑΙΟ ΠΕΜΠΤΟ

Ψευδομεταβλητές

5.1 Εισαγωγή

Σε πολλές πρακτικές εφαρμογές μια ή περισσότερες ποιοτικές μεταβλητές ασκούν σημαντική επίδραση στην εξαρτημένη μεταβλητή του μοντέλου. Έτσι π.χ. οι πωλήσεις ενός προιόντος μπορεί να επηρεάζονται σοβαρά από την εποχή του έτους, οι εξαγωγές από την ένταξη της χώρας στην ΕΟΚ, η ζήτηση ενός προιόντος από το αν η παρατήρηση γίνεται σε αγροτική ή μη αγροτική περιοχή.

Μία ποιοτική μεταβλητή με δύο δυνατές τιμές ονομάζεται **δίτιμη ή διχοτομική** διαφορετικά, ονομάζεται **πλειότιμη ή παράγοντας** και οι τιμές της **επίπεδα**. Έτσι, η ύπαρξη ή απουσία ενός χαρακτηριστικού όταν γίνεται η παρατήρηση Yi είναι μια διχοτομική μεταβλητή. Η απάντηση σε μία ερώτηση όταν μπορεί να είναι «συμφωνώ», «διαφωνώ», «δεν απαντώ» είναι μία πλειότιμη μεταβλητή. Η ποιότητα του προιόντος όταν μπορεί να καταταγεί σε μια από τις κατηγορίες «κακή», «μέτρια», «καλή», «εξαιρετική» είναι πλειότιμη μεταβλητή ενώ όταν μπορεί να καταταγεί σε «καλό» ή «ελαττωματικό» είναι δίτιμη μεταβλητή.

Σε ένα μοντέλο για την τυχαία μεταβλητή Y μπορούμε να πάρουμε υπόψη μια δίτιμη ποιοτική μεταβλητή εισάγοντας ως ερμηνευτική μια **ψευδομεταβλητή ή μεταβλητή δείκτη** που ορίζεται ως εξής:

$$D = \begin{cases} 1 \text{ αν η παρατήρηση γίνεται στο } \alpha \\ 0 \text{ διαφορετικά} \end{cases}$$

Ο τρόπος που η D θα εισαχθεί στο μοντέλο εξαρτάται από το μοντέλο αλλά και το είδος των μεταβολών που θέλουμε να συλλάβουμε.

5.2 Μεταβολή του σταθερού όρου στο μοντέλο παλινδρόμησης

Όταν θεωρούμε ότι η μετάβαση από τη μία τιμή της δίτιμης μεταβλητής στην άλλη έχει ως αποτέλεσμα την παράλληλη μετατόπιση

του μοντέλου κατά μία σταθερά, εισάγουμε ως επί πλέον ερμηνευτική μεταβλητή μια ψευδομεταβλητή.

Παράδειγμα

Θέλουμε να εκτιμήσουμε ένα μοντέλο παλινδρόμησης για τη μηνιαία ζήτηση Υ ενός αναψυκτικού με ερμηνευτική μεταβλητή τη μηνιαία δαπάνη για διαφήμιση x. Για να πάρουμε υπόψη την επίδραση του καλοκαιριού στην κατανάλωση του αναψυκτικού εισάγουμε στο μοντέλο την ψευδομεταβλητή D η οποία παίρνει την τιμή 1 για τους καλοκαιρινούς μήνες και μηδέν διαφορετικά. Τα δεδομένα είναι μηνιαίες παρατηρήσεις μιας χρονιάς:

Μήνας	Yi (εκ. δρχ)	Xi (σε χιλ. δρχ)	Di
Ι	1,145	312	0
Φ	1,912	344	0
Μ	1,739	307	0
Α	2,092	435	0
Μ	2,617	307	0
Ι	9,375	376	1
Ι	14,012	443	1
Α	14,812	548	1
Σ	11,001	568	0
Ο	3,710	312	0
Ν	2,561	496	0
Δ	1,912	447	0

Εκτιμούμε το μοντέλο

$$\hat{Y} = -5,363 + 0,022X + 8,16D$$
$$(-1,8) \quad (2,9) \quad (5,1)$$

με $R^2 = 0,842$, $SSE = 46,6$ $s = 2,3$

και οι αριθμοί στις παρενθέσεις είναι οι t-αναλογίες (bj / s_{bj}) για τις αντίστοιχες εκτιμήσεις. Έτσι το μοντέλο της ζήτησης για έναν καλοκαιρινό μήνα (D=1) γίνεται

$$\hat{Y} = (-5,363 + 8,16) + 0,022X$$

ή

$$\hat{Y} = 2,797 + 0,022X \qquad (5.2.1)$$

ενώ για ένα μη καλοκαιρινό μήνα το μοντέλο ζήτησης είναι το εξής:

$$\hat{Y} = -5,363 + 0,022X \qquad (5.2.2)$$

Αγνοώντας την επίδραση του καλοκαιριού στη ζήτηση, εκτιμούμε το μοντέλο

$$\hat{Y} = -8,07 + 0,033X \qquad (5.2.3)$$
$$(-1,4) \quad (2,5)$$

με $R^2 = 0,38$ $s = 4,3$

το οποίο έχει μεγαλύτερη κλίση από τα (5.2.1) και (5.2.2).

Συμπαιρένουμε ότι αν δεν ληφθεί υπόψη η επίδραση του καλοκαιριού τότε έχουμε μια υπερεκτίμηση της συμβολής της διαφήμισης στις μεταβολές της ζήτησης. Ή διαφορετικά, αυτό έχει ως αποτέλεσμα θετικό σφάλμα μεροληψίας στην εκτίμηση του b_1 . Πάντως το μοντέλο (5.2.3) έχει κακή προσαρμογή στα δεδομένα όπως φαίνεται από το χαμηλό συντελεστή προσδιορισμού (αλλά και την χαμηλή τιμή του στατιστικού Durbin-Watson=0,82 , που είναι ένδειξη κακής προσαρμογής όπως θα δούμε στο επόμενο κεφάλαιο) .

5.3 Μεταβολή της κλίσης ή των μερικών κλίσεων του μοντέλου

Αν σε ένα μοντέλο παλινδρόμησης εισάγουμε ως επί πλέον ερμηνευτική μεταβλητή το γινόμενο μιας ψευδομεταβλητής D επί μια ερμηνευτική μεταβλητή, έστω x_j , τότε μπορούμε να συλλάβουμε μεταβολές στο συντελεστή (μερικής) παλινδρόμησης της x_j . Η DXj ονομάζεται πολλαπλαστική ψευδομεταβλητή .

Παράδειγμα

Σε μια έρευνα των παραγόντων που ερμηνεύουν το επίπεδο ανάπτυξης μιας χώρας εκτιμήθηκε το ακόλουθο μοντέλο παλινδρόμησης :

όπου

Y = ο ρυθμός μεγέθυνσης του ΑΕΠ μιας χώρας,
X_1 = ο ρυθμός μεγέθυνσης των εξαγωγών
X_2 = ο ρυθμός μεγέθυνσης του εργατικού δυναμικού

X_3 = ο ρυθμός μεγέθυνσης του κεφαλαίου της χώρας

$$D = \begin{cases} 1 & \text{αν το επίπεδο ανάπτυξης της χώρας είναι πάνω από το θεωρούμενο κριτικό} \\ 0 & \text{διαφορετικά} \end{cases}$$

Αν η κλίση της πολλαπλασιαστικής ψευδομεταβλητής X_1D είναι στατιστικά σημαντική τότε μπορούμε να συμπεράνουμε ότι :
πάνω από το υποτιθέμενο 'κριτικό' επίπεδο ανάπτυξης, η οριακή συμβολή των εξαγωγών στη μεγέθυνση του ΑΕΠ είναι μεγαλύτερη απ' ότι κάτω απ' αυτό.

5.4 Πλειότιμες ποιοτικές μεταβλητές

Για να πάρουμε υπόψη μια ποιοτική μεταβλητή μέ δύο τιμές εισάγουμε στο μοντέλο μια ψευδομεταβλητή. Γενικά, για να πάρουμε υπόψη μια ποιοτική μεταβλητή μέ $\lambda>2$ τιμές εισάγουμε στο μοντέλο λ-1 ψευδομεταβλητές.

Έτσι, π.χ έστω ότι θέλουμε να εκτιμήσουμε το απλό μοντέλο κατανάλωσης

$$Y_i = b_0 + b_1 X_i + \varepsilon_i \qquad i = 1, \ldots, n$$

όπου Y= η κατανάλωση και το X= το διαθέσιμο εισόδημα μιας οικογένειας. Έστω ακόμη ότι διαθέτουμε διαστρωματικά δεδομένα από 3 διαφορετικές περιοχές, έστω Α, Β, Γ. Αν υποπτευόμαστε ότι το μοντέλο μεταβάλλεται στις τρεις περιοχές μπορούμε να χρησιμοποιήσουμε δύο ψευδομεταβλητές, τις D_1 και D_2 που ορίζονται ως εξής :

$D_1 = \begin{cases} 1 \text{ αν η παρατήρηση γίνεται στην περιοχή Α} \\ 0 \text{ διαφορετικά} \end{cases}$

$D_2 = \begin{cases} 1 \text{ αν η παρατήρηση γίνεται στην περιοχή Β} \\ 0 \text{ διαφορετικά} \end{cases}$

Διακρίνουμε τις ακόλουθες περιπτώσεις ανάλογα μέ τις διαφορές του μοντέλου στις τρεις περιοχές.

I. **Διαφορετικός σταθερός όρος**. Το μοντέλο παλινδρόμησης εξειδικεύεται ως εξής :

$$Y_i = b_0 + b_1 X_i + \gamma_1 D_{1i} + \gamma_2 D_{2i} + \varepsilon_i \qquad i = 1, \ldots, n$$

το οποίο είναι ισοδύναμο μέ τρία μοντέλα, ένα για κάθε περιοχή ως εξής:

$Y_i = (b_0 + \gamma_1) + b_1 X_i + \varepsilon_i$ για παρατηρήσεις Y_i στην περιοχή Α

$Y_i = (b_0 + \gamma_2) + b_1 X_i + \varepsilon_i$ για παρατηρήσεις Y_i στην περιοχή Β

$Y_i = b_0 + b_1 X_i + \varepsilon$ για παρατηρήσεις Y_i στην περιοχή Γ.

II. Διαφορετική κλίση ή στο παράδειγμα μας, διαφορετική οριακή ροπή για κατανάλωση. Τότε θα χρησιμοποιήσουμε δύο πολλαπλασιαστές μεταβλητές οπότε το μοντέλο εξειδικεύεται ως εξής :

$$Y_i = b_0 + b_1 X_i + \delta_1 D_{1i} + X_i + \delta_2 D_{2i} X_i + \varepsilon_i \qquad i = 1, \ldots, n$$

το οποίο γίνεται για τις 3 περιοχές, αντίστοιχα, ως εξής :

$Y_i = b_0 + (b_1 + \delta_1) X_i + \varepsilon_i$ για την Α

$Y_i = b_0 + (b_1 + \delta_2) X_i + \varepsilon_i$ για την Β

$Y_i = b_0 + b_1 X_i + \varepsilon_i$ για την Γ

III. Διαφορετικά και ο σταθερός όρος και η κλίση. Το μοντέλο του παραδείγματος μας παίρνει τη μορφή

$$Y_i = b_0 + b_1 X_i + \gamma_1 D_{1i} + \gamma_2 D_{2i} + \delta_1 D_{1i} X_i + \delta_2 X_i + \varepsilon_i$$

που γίνεται, για τις τρεις περιοχές, αντίστοιχα :

$Y_i = (b_0 + \gamma_1) + (b_1 + \delta_1) X_i + \varepsilon_i$ για την Α

$Y_i = (b_0 + \gamma_2) + (b_1 + \delta_2) X_i + \varepsilon_i$ για την Β

$Y_i = b_0 + b_1 X_i + \varepsilon_i$ για την Γ

Παράδειγμα

Για την ημερήσια ζήτηση του γιαουρτιού μάρκας Φ το οποίο βγαίνει σε 4 τύπους, εκτιμήθηκε το ακόλουθο μοντέλο παλινδρόμησης :

Y

όπου Y είναι η ημερήσια ζήτηση του γιαουρτιού Φ, σε χιλιάδες κιβώτια, X είναι ο τηλεοπτικός χρόνος σε sec που διαφημίστηκε η Φ την προηγούμενη μέρα και οι ψευδομεταβλητές D_1, D_2, D_3 ορίζονται στον ακόλουθο πίνακα :

Τύπος γιαουρτιού	D_1	D_2	D_3
Α	1	0	0
Β	0	1	0
Γ	0	0	1
Δ	0	0	0

Έτσι, για τον Δ τύπο γιαουρτιού (D_1= D_2= D_3= 0) το μοντέλο είναι το

$$\hat{Y} = 14,716 + 0,673X \qquad (5.4.1)$$

Για τον Α τύπο το μοντέλο είναι το :

$$\hat{Y} = (14,716 + 0,387) + 0,673X$$

$$= 15,103 + 0,673X \qquad (5.4.2)$$

Έτσι, αν η Φ δεν διαφημιστεί από την τηλεόραση την προηγούμενη μέρα εκτιμούμε ότι η ζήτηση για τον Α τύπο θα είναι μεγαλύτερη από τη ζήτηση του Δ κατά 0,387 χιλιάδες κιβώτια.

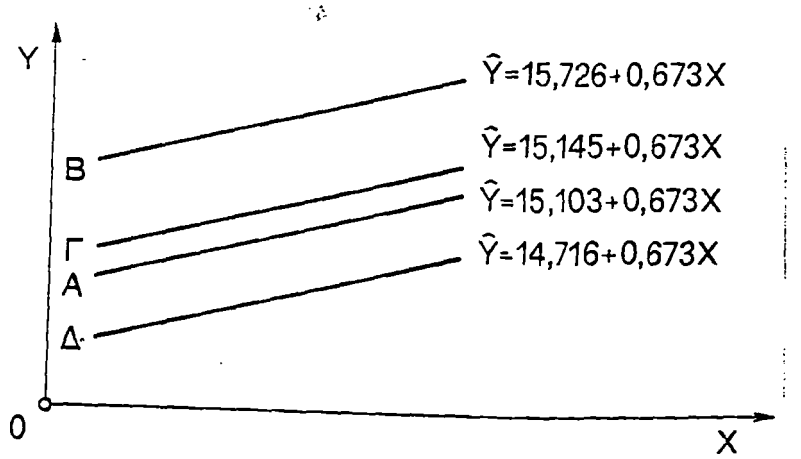
Για τους τύπους Β και Γ το μοντέλο παλινδρόμησης γίνεται αντίστοιχα:

$$\hat{Y} = (14,716 + 1,013) + 0,673\,X$$
$$= 15,729 + 0,673\,X \qquad (5.4.3)$$

και

$$\hat{Y} = (14,716 + 0,429) + 0,673X$$
$$= 15,145 + 0,673\,X \qquad (5.4.4)$$

Ο όρος b_1=0,673 είναι η προβλεπόμενη αύξηση της ζήτησης για κάθε τύπο γιαουρτιού όταν ο διαφημιστικός χρόνος της προηγούμενης μέρας αυξηθεί κατά 1 sec. Η γραφική παράσταση των μοντέλων (5.4.1)-(5.4.4) δίνεται στο ακόλουθο σχήμα.



5.5 Έλεγχος της σταθερότητας του μοντέλου με ψευδομεταβλητές.

Όταν εκτιμούμε ένα μοντέλο παλινδρόμησης υποθέτουμε ότι οι παράμετροι $b_0, b_1, ..., b_k$ είναι σταθερές σε όλες τις παρατηρήσεις Yi. Σε πολλές περιπτώσεις η γραφική παράσταση των καταλοίπων ή πληροφορίες, για τον τρόπο που έχουν συλλεγεί τα δεδομένα μας υπαγορεύουν να εκτιμήσουμε δύο μοντέλα- ένα για τις n_1 παρατηρήσεις και ένα για τις $n_2=n-1$
δηλαδή τα εξής:

$$Y_i = b_0^1 + b_1^1 X_{1i} + \dots + b_k^1 X_{ki} + \varepsilon_i, \quad i = 1, \dots, n$$

$$Y_i = b_0^2 + b_1^2 X_{1i} + \dots + b_k^2 X_{ki} + \varepsilon_i, \quad i = n_1+1, \dots, n \qquad (5.5.1)$$

Με βάση το δείγμα των δεδομένων μας μπορούμε να ελέγξουμε την υπόθεση ότι οι παράμετροι των δύο μοντέλων είναι ίδιες που σημαίνει ότι το μοντέλο είναι κοινό για όλες τις παρατηρήσεις Yi, i=1,..., n. Η μηδενική και εναλλακτική υπόθεση του ελέγχου εξειδικεύονται ως εξής :

$$H_0 : b_j^1 = b_j^2 \quad j = 0, 1, \dots, k$$

$$H_\varepsilon : b_j^1 \neq b_j^2 \quad j = 0, 1, \dots, k$$

κάτω από την H_0 έχουμε το μοντέλο

$$Y_i = b_0 + b_1 X_{1i} + \dots + b_k X_{ki} + \varepsilon_i, \quad i = 1, \dots, n \qquad (5.5.2)$$

Ενώ κάτω από την H_ε έχουμε τα δύο μοντέλα (5.5.1) τα οποία μπορούν να γράφουν ενιαία ως εξής :

$$Y_i = b_0 + b_1 X_{1i} + \dots + b_k X_{ki} + \gamma_0 D + \gamma_1 D X_{1i} + \dots + \gamma_k D X_{ki} + \varepsilon_i$$

$$i = 1, \dots, n \qquad (5.5.3)$$

όπου

$$D_i = \begin{cases} 1 & \text{αν } i = 1, \dots n_1 \\ 0 & \text{αν } i = n_1+1, \dots, n \end{cases}$$

και

$$\gamma_0 = b_0^1 - b_0^2, \ \gamma_1 = b_1^1 - b_1^2, \dots, \gamma_k = b_k^1 - b_k^2$$

Για να ελέγξουμε την H_0 έναντι της H_ε δηλαδή το (5.5.3) και έστω SSE_ε το άθροισμα των τετραγωνικών καταλοίπων το οποίο έχει n- (κ+1) βαθμούς ελευθερίας. Η διαφορά (SSE_0-SSE_ε) έχει [n- (κ+1)] - [n-2 (κ+1)] = κ+1 βαθμούς ελευθερίας.

Αποδεικνύεται ότι το στατικό

$$F = \frac{(SSE_o - SSE_\varepsilon)/(k+1)}{SSE_\varepsilon/(n-2)(k+1)} \qquad (5.5.4)$$

κολουθεί την κατανομή $F_{v1, v2}$ με v_1=κ+1 και v_2=ν-2 (κ+1) βαθμούς ελευθερίας.

Επομένως αν η τιμή που υπολογίζουμε από την (5.5.4) είναι μεγαλύτερη από την $F_{v1, v2, q}$ απορρίπτουμε την H_0 στο α= 0,05.

Ο έλεγχος αυτός είναι γνωστός ως έλεγχος του Chow και μπορεί να εμφαρμοστεί μόνο αν $n_1 > κ+1$ και $n_2 > κ+1$.

Αν $n_1 > κ+1$ αλλά $n_2 < κ+1$ ο Chow (1960) προτείνει την ακόλουθη παραλαγή του :

- Εκτιμούμε το μοντέλο με τις n_1 παρατηρήσεις και έστω SSE_1 το άθροισμα των τετραγωνικών αποκλίσεων το οποίο έχει n_1-(κ+1) βαθμούς ελευθερίας. Η διαφορά (SSE_0-SSE_1) έχει n_2 βαθμούς ελευθερίας. Το στατιστικό :

$$F = \frac{(SSE_o - SSE_1)/n_2}{SSE_1/(n_1 - k - 1)}$$

συγκρίνεται με την τιμή $F_{v1, v2, q}$ της κατανομής F με $v_1=n_2$ και v_2= n_1- κ-1 βαθμούς ελευθερίας.

ΚΕΦΑΛΑΙΟ ΕΚΤΟ

Η αυτοσυσχέτηση των σφαλμάτων

6.1 Εισαγωγή

Ο σκοπός της εκτίμησης ενός μοντέλου παλινδρόμησης είναι να συλλάβει όλη τη συστηματικότητα στη συμπεριφορά της εξαρτημένης έτσι ώστε να μείνουν, ως υπόλοιπη μεταβλητότητα, τυχαία και ανεξάρτητα σφάλματα δηλαδή ποσότητες μη προβλέψιμες. Όταν διαδοχικά σφάλματα εκτίμησης συσχετίζονται λέμε ότι υπάρχει σειριακή συσχέτιση ή αυτοσυσχέτιση στα κατάλοιπα. Ειδικότερα, αν θετικά σφάλματα τείνουν να ακολουθούνται από θετικά, η αυτοσυσχέτιση είναι θετική, διαφορετικά, είναι αρνητική. Είναι προφανές ότι στην περίπτωση αυτή μπορούμε να βελτιώσουμε τις εκτιμήσεις του μοντέλου με νέα εξειδίκευση η οποία θα αρεί αυτή τη συστηματικότητα. Η αυτοσυσχέτιση στα σφάλματα εκτίμησης είναι, γενικά, ένδειξη κακής εξειδίκευσης του μοντέλου. Όταν εκτιμούμε ένα μον τέλο με δεδομένα χρονικών σειρών, η αυτοσυσχέτιση των σφαλμάτων μπορεί, σχεδόν πάντα, να αποδοθεί στο ότι λανθασμένα έγινε η υπόθεση των ασυσχετιστών διαταρακτικών ορών. Σε άλλες περιπτώσεις μπορεί να οφείλεται στο ότι δεν έγινε σωστή εξειδίκευση του μαθηματικού τύπου της συνάρτησης παλινδρόμησης ή στο ότι έχει παραλειφθεί μια σημαντική ερμηνευτική μεταβλητή της οποίας οι τιμές αυτοσυσχετίζονται. Τέλος, η αυτοσυσχέτιση των σφαλμάτων της εκτίμησης μπορεί μα οφείλεται στο ότι ένα ή περισσότερα ακραία σημεία στα δεδομένα προκαλούν μετατόπιση του εκτιμηθέντος μοντέλου.

6.2 Η αυτοσυσχέτιση των διαταρακτικών όρων.

Μια από τις υποθέσεις του μοντέλου παλινδρόμησης είανι ότι οι διαταρακτικοί όροι είναι ανά δύο ασυσχέτιστοι δηλαδή ότι ισχύει :

$$Cov(\varepsilon_i \varepsilon_j) = 0 \qquad \forall i \neq j$$

Με άλλα λόγια υποθέτουμε ότι, για δεδομένο σημείο σχεδιασμού, μια παρατήρηση Yi θα διαφέρει από τη μέση τιμής της κατά ένα ποσό το οποίο είναι ασυσχέτιστο με το μέγεθος της αντίστοιχης διαφοράς σε οποιοδήποτε άλλο σημείο.

Ο διαταρακτικός όρος είναι το αποτέλεσμα της συνδυασμένης δράσης πολλών παραγόντων οι οποίοι είτε δεν είναι αρκετά σημαντικοί για να περιληφθούν στο μοντέλο είτε δεν είναι μετρήσιμοι.
Όταν αναλύσουμε δεδομένα χρονικών σειρών, δηλαδή παρατηρήσεις διατεταγμένες ως προς το χρόνο τότε, συνήθως, η συμπεριφορά των παραγόντων αυτών είναι όμοια με τη συμπεριφορά τους σε προηγούμενες περιόδους. Αυτό έχει ωα αποτέλεσμα οι διαταρακτικοί όροι που αντιστοιχούν σε διαφορετικές παρατηρήσεις, συνήθως γειτονικές, να συσχετίζονται. Στην περίπτωση αυτή λέμε ότι οι διαταρακτικοί όροι συσχετίζονται σειριακά ή αυτοσυσχετίζονται.

6.3 Το μοντέλο αυτοσυσχέτισης AR (1)

Η πιο απλή μορφή αυτοσυσχέτισης είναι αυτή στην οποία διαδοχικοί διαταρακτικοί όροι συνδέονται με το ακόλουθο σχήμα:

$$\varepsilon_i = \rho \varepsilon_{i-1} + u_i \qquad (6.3.1)$$

όπου u_i ασυσχέτιστες τυχαίες μεταβλητές με μηδενική μέση τιμή και σταθερή διακύμανση σ^2_u και ανερξάτητες των $\varepsilon_{i-1}, \varepsilon_{i-2}, \varepsilon_{i-3}, \ldots$. Το μοντέλο αυτοσυσχέτισης που ορίζεται από την (6.3.1) ονομάζεται **αυτοπαλινδρομούμενο πρώτης τάξης** και συμβολίζεται ως AR(1). Ονομάζεται έτσι διότι ορίζει ένα μοντέλο παλινδρόμησης του ε_i επι του ε_i-1.
Θυμίζουμε ότι στο απλό μοντέλο παλινδρόμησης της Y επί της X ισχύει:

$$b_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$$

επομένως και το μοντέλο (6.3.1) ισχύει :

$$b_1 = \rho_1 \frac{\sigma_i^2}{\sigma_{i-1}^2}$$

όπου p_1 ο συντελεστής αυτοσυσχέτησης πρώτης τάξης διαδοχικών διαταρακτικών όρων. Αυτά τα u_i είναι ομοσκεδαστικά τότε $\sigma^2_i = \sigma^2_{i-1}$ και

$$b_1 = \rho_1$$

Δηλάδη ο συντελεστής p του μονέλου (6.3.1) είναι ο συντελεστής αυτοσυσχέτισης πρώτης τάξης των διαταρακτικών όρων.

Είναι εύκολο να δειχθεί ότι ένα σφάλμα στην παρατήρηση i επισρά σε όλες τις επόμενες περιόδους και το μέγεθος της επίδρασης αυτής μειώνεται με το χρόνο. Αρκεί να υπολογίσουμε την συνδιακύμανση του ε_i με το διαταρακτικό όρο οποιαδήποτε προηγούμενης περιόδου :

$$Cov(\varepsilon_i \varepsilon_{i-1}) = E(\varepsilon_i - \varepsilon_{i-1}) = E(\rho\varepsilon_{i-1} + u) \cdot \varepsilon_{i-1} = E(\rho\varepsilon_i^2 + u_i \varepsilon_{i-1})$$

$$\Leftrightarrow Cov(\varepsilon_i \varepsilon_{i-1}) = E(\rho\varepsilon_{i-1} + u_i)\varepsilon_{i-1}$$

$$= E(\rho\varepsilon^2_{i-1} + u_i \varepsilon_{i-1})$$

$$= \rho E(\varepsilon^2_{i-1}) = \rho Var(\varepsilon_{t-1}) = \rho\sigma^2_\varepsilon$$

Ομοίως υπολογίζουμε

$$Cov(\varepsilon_i \varepsilon_{i-2}) = \rho^2\sigma^2_\varepsilon$$

$$Cov(\varepsilon_i \varepsilon_{i-3}) = \rho^3\sigma^2_\varepsilon$$

$$Cov(\varepsilon_i \varepsilon_{i-\lambda}) = \rho^\lambda\sigma^2_\varepsilon$$

Επομένως για το συντελεστή αυτοσυσχέτισης λ-τάξης, λ=1, 2, 3... ισχύει:

$$\rho_\lambda = \frac{Cov(\varepsilon_i \varepsilon_{i-\lambda})}{\sigma_\varepsilon^2} = \frac{\rho^\lambda \sigma_\varepsilon^2}{\sigma_\varepsilon^2} = \rho^\lambda \qquad \lambda = 1, 2, 3 \ldots$$

Συμπεραίνουμε ότι στο AR(1) μοντέλο ολόκληρη η δομή της συσχέτισης των διαταρακτικών όρων προσδιορίζεται από μία παράμετρο- τον συντελεστή ρ.

Συνέπειες των AR(1) σφαλμάτων στην εκτίμηση ελάχιστων τετραγώνων

Η μέθοδος των ελαστικών τετραγώνων σε ένα μοντέλο με AR(1) σφάλματα ορίζει εκτιμητές οι οποίοι εξακολουθούν ζνα είναι αμερόληπτοι. Δεν είναι όμως αμερόληπτες οι διακυμάσεις τους και συνεπώς και τα τυπικά τους σφάλματα. Ειδικότερα αν $\rho > 0$ τότε ισχύει:

$$E(S^2_{\hat{b}_j}) < \sigma^2_{\hat{b}_j} \qquad j = 0, 1, \ldots, K$$

δηλαδή υπάρχει μια συστηματική τάση να υποεκτιμώνται τα τυπικά σφάλματα και συνεπώς να θεωρούνται στατιστικά σημαντικές οι εκτιμήσεις των παραμέτρων ενώ δεν είναι. Αντίστοιχα, όταν $\rho < 0$ τότε τα τυπικά σφάλματα τείνουν να υπερκτιμούνται.
Αν στις ερμηνευτικές μεταβλητές του μοντέλου περιλαμβάνεται και η εξαρτημένη σε υστέρηση τότε τα προβλήματα αυτοσυσχέτισης είναι πιο σοβαρά.

6.4 Το στατιστικό Durbin - Watson.

Έστω το μοντέλο $Yi = b_0 + b_1X_{1i} +...+b_kX_{ki}+ \varepsilon_i \quad i= 1,..., n$ του οποίου τις παραμέτρους εκτιμήσαμε με την μέθοδο των ελάχιστων τετραγώνων και $\varepsilon_i= Y_i - Y_i$, $i= 1,...,n$ τα κατάλοιπα αυτής της εκτίμησης. Το στατιστικό Durbin - Watson υπολογίζεται ως εξής :

(6.4.1) $$D-W \equiv d = \frac{\sum_{i=2}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

Ο συντελεστής αυτοσυσχέτισης πρώτης τάξης των καταλόγων υπολογίζεται, συμφωνα με τον τύπο (3.12.1) ως εξής :

(6.4.2) $$r_1 = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

Αναπτύσσοντας τον αριθμητή της (6.4.1) έχουμε :

$$d = \frac{\sum \hat{\varepsilon}_i^2}{\sum \varepsilon_i^2} + \frac{\sum \hat{\varepsilon}_{i-1}^2}{\sum \hat{\varepsilon}_i^2} - 2\frac{\sum \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum \hat{\varepsilon}_i^2}$$

Όταν το μέγεθος του δείγματος είναι μεγάλο τότε

$$\sum_{i=2}^{n} \hat{\varepsilon}_i^2 \simeq \sum_{i=1}^{n-1} \hat{\varepsilon}_{i-1} \simeq \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

οπότε χρησιμοποιώντας και

την (6.4.2) το στατιστικό Durbin - Watson γίνεται

$$d = 2\,(I - r_1)$$

Είναι προφανές οι ακόλουθες σχέσεις :

$$r_1 = 0 \Longleftrightarrow d = 2$$

$$r_1 = -1 \Longleftrightarrow d = 4$$

$$r_1 = 1 \Longleftrightarrow d = 0$$

Ο έλεγχος Durbin - Watson για AR(1) σφάλματα

Όταμ=ν οι διαταρακτικοί όροι ή σφάλματα ενός μοντέλου είναι AR(1) είναι συντελεστής αυτοσυσχέτισης πρώτης τάξης ρ είναι η μοναδική παράμετρος που προσδιορίζει τη δομή αυτόσυσχέτισης των διαταρακτικών όρων. Στα μεγάλα δείγματα ο συντελεστής Γ_1 αποτελεί μια ικανοποιητική εκτίμηση του ρ.
Επομένως μια τιμή Γ_1 κοντά στο 1 ή στο -1 και συνεπώς μια τιμή για το D -W κοντά στο 0 ή στο 4 αποτελούν σοβαρή ένδειξη AP(1) διαταρακτικών όρων. Η κατανομή δειγματοληψίας
του d εξαρτάται από τις τιμές των ερμηνευτικών μεταβλητών του μοντέλου. Οι Durbin και Watson (1950) υπολόγισαν, για επιλεγμένα επίπεδα σημαντικότητας δύο κριτικές τιμές d_L και d_u που εξαρτώνται μόνο από το πλήθος κ των ερμηνευτικών μεταβλητών και όχι από τις τιμές τους. Έτσι, η μηδενική υπόθεση των ασυσχέτιστων διαταρακτικών όρων ελέγχεται ως προς την ενναλακτική υπόθεση ότι είναι AR(1), με $\rho > 0$ ως εξής :

1. Αν $d < d_L$ η H_0 απορρίπτεται
2. Αν $d > d_u$ η H_0 δεν απορρίπτεται
3. Αν $d_L < d < d_u$ δεν μπορούμε να αποφανθούμε

Για $\rho > 0$ το στατιστικό συγκρίνουμε το στατιστικό 4-d με τις κριτικές τιμές d_L και d_u σαν να κάναμε έλεγχο για θετική αυτοσυσχέτιση.

Η περιοχή αβεβαιότητας είναι σοβαρό μειονέκτημα του ελέγχου αυτού. Αποδεικνύεται πάντως ότι, όσο αυξάνει το μέγεθος του δείγματος το εύρος της περιοχής αβεβαιότητας τείνει να μειώνεται. Για τα μοντέλα τα οποία εκτιμώνται με δεδομένα οικονομικών χρονικών σειρών η τακτική που συνήθως ακολουθούνται είναι να θεωρείται και η περιοχή αβεβαιότητας ως περιοχή απόρριψης της H_0.

6.5 Εκτίμηση ενός μοντέλου με AR(1) διαταρακτικούς τύπους

Για να εφαρμόσουμε τη μέθοδο των Ελαχίστων Τετραγώνων στην εκτίμηση του γραμμικού μοντέλου παλινδρόμησης όταν οι διαταρακτικοί όροι συνδέονται με μια AR(1) διαδικασία πρέπει να εμφαρμόσουμε προηγουμένως έναν μετασχηματισμό των δεδομέων, έτσι ώστε να αφαιρεθεί η αυτοσυσχέτιση των διαταρακτικών όρων. Έστω λοιπόν το μοντέλο

$$Y_i = b_0 + b_1 X_{1i} + \dots + b_k X_{kp} + \varepsilon_i \qquad (6.5.1)$$

με

$$\varepsilon_i = \rho \varepsilon_{i-1} + u_i \quad |\rho| < 1 \quad i = 2, \dots, n \qquad (6.5.2)$$

και u_i $(0\text{-}\sigma_u^2)$, E (ui uj)=0 για κάθε j=i και
E (ui ε_{i-1}) = E (ui ε_{t-2}) = ... = 0 . Το μοντέλο αυτό ισχύει και για την ποαρατήρηση Y_{i-1} δηλαδή έχουμε :

$$Y_{i-1} = b_0 + b_1 X_{1i-1} + \dots + b_k X_{ki-1} + \varepsilon_{i-1} \qquad (6.5.3)$$

Πολλαπλασιάζουμε με ρ και παίρνουμε :

$$\rho Y_{i-1} = \rho b_0 + b_1 \rho X_{1i-1} + \dots + b_k \rho X_{ki-1} + \rho \varepsilon_{i-1} \qquad (6.5.4)$$

Αφαιρούμε την (6.5.4) από την χρησιμοποιούμε την (6.5.2) οπότε προκύπτει το μοντέλο :

$$Y_i - \rho Y_{i-1} = b_0(1-\rho) + b_1(X_{1i} - \rho X_{1i-1}) + \dots + b_k(X_{ki} - \rho X_{ki}) + u_i \qquad (6.5.5)$$

στο οποίο οι διαταρακτικοί λοροι είναι ασυσχέτιστοι. Επομένως για να εκτιμήσουμε τις παραμέτρους του αρχικού μοντέλου αρκεί να εκτιμήσουμε τις παραμέτρους του (6.5.5). Αυτό υπαγορεύει τους ακόλουθους μετασχηματισμούς των δεδομένων :

$$\begin{aligned} Y_i^* &= Y_i - \rho Y_{i-1} \\ X_{1i}^* &= X_{1i} - \rho X_{1i-1} \\ &\vdots \\ X_{ki}^* &= X_{ki} - \rho X_{ki-1} \qquad i = 2, \dots, n \end{aligned} \qquad (6.5.6)$$

που ονομάζεται **γενικευμένες Διαφορές.** Για ρ=1 οι (6.5.6) γίνονται :

$$Y_i^* = Y_i - Y_{i-1}$$
$$X_{1i}^* = X_{1i} - X_{1i-1}$$
$$\vdots$$
$$X_{ki}^* = X_{ki} - X_{ki-1} \qquad i = 2, \ldots, n$$

και ονομάζεται **πρώτες Διαφορές** των Y και $X_1, \ldots, X_k$, αντίστοιχα. Για $[\rho] < 1$ οι μετασχηματισμοί (6.5.6) ορίζουν **οιονεί Διαφορές.**

Με τους μετασχηματισμούς (6.5.6) χάνεται η πρώτη παρατήρηση γεγονός που μπορεί να είναι πολύ σημαντικό για μικρά δείγματα. Η λύση είναι να πάρουμε :

$$Y_1^* = \sqrt{1-\rho^2}\, Y_1 \quad \text{και} \quad X_{j1}^* = \sqrt{1-\rho^2}\, X_{j1} \quad j = 1, \ldots, k$$

Αυτός ο μετασχηματισμός έχει ως αποτέλεσμα η διακύμανση του διαταρακτικού όρου $\varepsilon_1^* = \sqrt{1-\rho^2}\ \varepsilon_1$ να ισούται με την διακύμανση όλων των άλλων διαταρακτικών όρων. Πράγματι έχουμε :

$$Var(\varepsilon_1^*) = (1-\rho^2)\, Var(\varepsilon_i) = Var(\varepsilon_i) - \rho^2 Var(\varepsilon_i) = \sigma_u^2$$

Επομένως, το πρόβλημα της εκτίμησης των παραμέτρων ενός μοντέλου με AR(1) διαταρακτικούς όρους καταλήγει στο πρόβλημα της εκτίμησης του ρ. Όλες οι μεθόδοι εκτίμησης του ρ που χρησιμοποιούνται στην πράξη είναι επαναληπτικές και επομένως δεν μπορούν να εφαρμοστούν παρά μόνο με Η/Υ. Εδώ θα παρουσιάσουμε τις πιο βασικές.

Μέθοδος Cochrane - Orcutt

Αυτή είναι η πιο παλιά μέθοδος εκτίμησης και περιλαμβάνει τα ακόλουθα στάδια :

• Εκτιμούμε τις παραμέτρους του μοντέλου με τη μέθοδο των Ελάχιστων Τετραγώνων (OLS).

• Υπολογίζουμε τα κατάλοιπα $\varepsilon_i = Y_i - Y_i$, i= 1, ..,n και στη συνέχεια τον συντελεστή αυτοσυσχέτισης πρώτης τάξης απ΄το τον τύπο

$$\Gamma_1 = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum_{i=1}^{n} \varepsilon_i^2}$$

ή απ' ευθείας, από το στατιστικό Durbin - Watson d ως εξής :

$$\Gamma_1 = 1 - \frac{d}{2}$$

• Παίρνουμε τις οιονεί διαφορές των δεδομένων με $\rho=\Gamma_1$ δηλαδή ως εξής :

$$Y_i^* = Y_i - \hat{\rho} Y_{i-1},\ X_{1i}^* = X_{1i} - \hat{\rho} X_{1i-1},\ \ldots,\ X_{ki}^* = X_{ki} - \hat{\rho} X_{ki-1}$$
$$i = 1, \ldots, n$$

• Εκτιμούμε τις παραμέτρους του μοντέλου χρησιμοποιώντας τα μετασχηματισμένα δεδομένα (Y_i^*, X_{1i}^*, ..., X_{ki}^*), i=1,..., n

Η διαδικασία επαναλαμβάνεται μέχρι ότου η τιμή του ρ από δύο διαδοχικές επαναλήψεις Δε διαφέρει περισσότερο από μια τιμή πολύ μικρή συνήθως ίση με 0,0001.

Να σημειωθεί ότι η διαδικάσια αυτή ισοδυναμεί με ελαχιστοποίηση του SSE του μοντέλου (6.5.5) που είναι μη γραμμική συνάρτηση των παραμέτρων $\rho, b_0, b_1, \ldots, b_k$ και μπορεί να έχει περισσότερα από ένα ελάχιστα. Έτσι, είναι δυνατόν η μέθοδος Cohrane - Orcutt να οδηγήσει ελάχιστα σε τοπικό και όχι σε ολικό ελάχιστο.

Η μέθοδος Hildred - Lu

Για όλες τις δυνατές τιμές του ρ στο διάστημα $-1 \leq \rho \leq 1$, με μήκος βήματος, συνήθως , ίσο με 0,01 υπολογίζουμε τις οιονεί διαφορές Y με τις οποίες, στη συνέχεια εκτιμούμε το μοντέλο παλινδρόμησης και υπολογίζουμε το SSE.
Τέλος, επιλέγεται η τιμή του ρ για την οποία το SSE είναι ελάχιστο.

Το μειονέκτημα αυτής της μεθόδου είναι το μήκος βήματος που συνήθως επιλέγεται είναι μικρό, διαφορετικά έχει τεράστιο υπολογιστικό κόστος. Αυτό έχει ως αποτέλεσμα ότι μπορεί να προσπερνάται η τιμή του ρ η οποία ελαχιστοποιεί το SSE.

6.6 Αυτοσυσχέτιση των καταλοίπων λόγω κακής εξειδίκευσης της συνάρτησης παλινδρόμησης.

Θα επισημάνουμε ορισμένες χαρακτηριστικές περιπτώσεις κακής εξειδίκευσης που έχει ως αποτέλεσμα την αυτοσυσχέτιση των καταλοίπων και συνακόλουθα μια τιμή για το στατιστικό Durbin - Watson κοντά στο μηδέν ή στο 4.

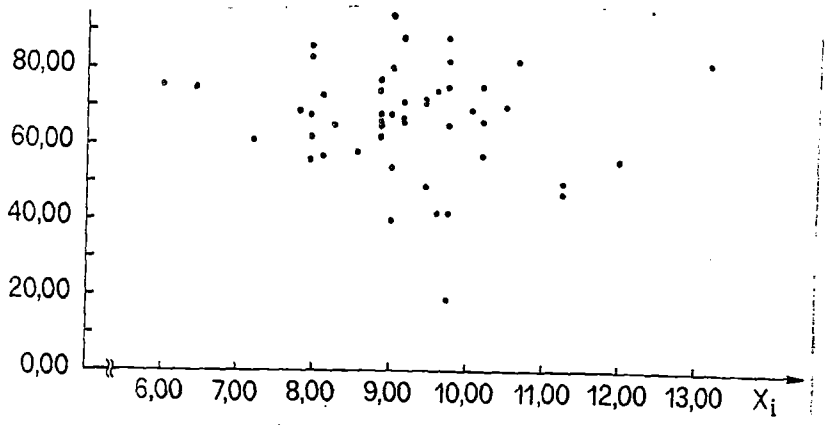
• Αν υποθέσουμε ότι θέλουμε να ερμηνεύσουμε την Y της οποίας Οι διαδοχικές παρατηρήσεις συσχετίζονται,με την X η οποία είναι ασυσχέτιστη με την Y. Είναι προφανές ότι τα κατάλοιπα αυτής της εκτίμησης αυτοσυσχετίζονται και η πληροφορία αυτή ενσωματώνεται στην πολύ μικρή ή πολύ μεγάλη τιμή του στατιστικού Durbin-Watson όπως στο ακόλουθο.

Παράδειγμα: Το μήκος της ανθρώπινης ζωής

Πολλοί άνθρωποι πιστεύουν ότι το μήκος Y της ζωής ενός ανθρώπου που μπορεί να προβλεφθεί από το μήκος X της λεγόμενης ΄΄γραμμή της ζωής΄΄στην παλάμη του χεριού του. Στα ακόλουθα Ζεύγη τιμών Xi είναι το μήκος της γραμμής της ζωής σε εκατοστόμετρα,στρογγυλεμένο στο πλησιέστερο 0,15cm Yi είναι η ηλικία θανάτου του ατόμου στρογγυλεμένη στο πλησιέστερο έτος.Τα δεδομένα δίνονται από τους wilson M.E-Mather L.E : 'Life expectancy' Journal of the American Medical Association,Vol 229,Nr 11,1974.Δίνονται επίσης από τους Draper-Smith,1981,σελ.67-68.

Από το διάγραμμα διασποράς των (Xi,Yi) που δίνεται στο σχήμα προκύπτει ότι δεν υπάρχει συστηματική συμμεταβολή των Xi και Yi.

Τα αποτελέσματα της εκτίμησης του απλού γραμμικού μοντέλου με σταθερό όρο είναι τα ακόλουθα (οι αριθμοί στις παρενθέσεις είναι οι t-αναλογίες):



Y= 79,23-1,37X

(5,34) (0,86)

με R =0,015 D-W=0,00796

SSE=9608 S=14,115

Παρατηρούμε ότι η εκτίμηση του συντελεστή b_1 δεν είναι στατιστικά σημαντική αφού

$t\pi=b_1/S\ b_1= -0,86<2$

Έτσι ,η μηδενική υπόθεση $b_1=0$ δεν μπορεί να απορριφθεί για τα συνηθισμένα επίπεδα σημαντικότητας.

Δοκιμάζουμε ακόμη το μοντέλο χωρίς τον σταθερό όρο και παίρνουμε τα ακόλουθα αποτελέσματα:

Y =7,09 X

(26,32)

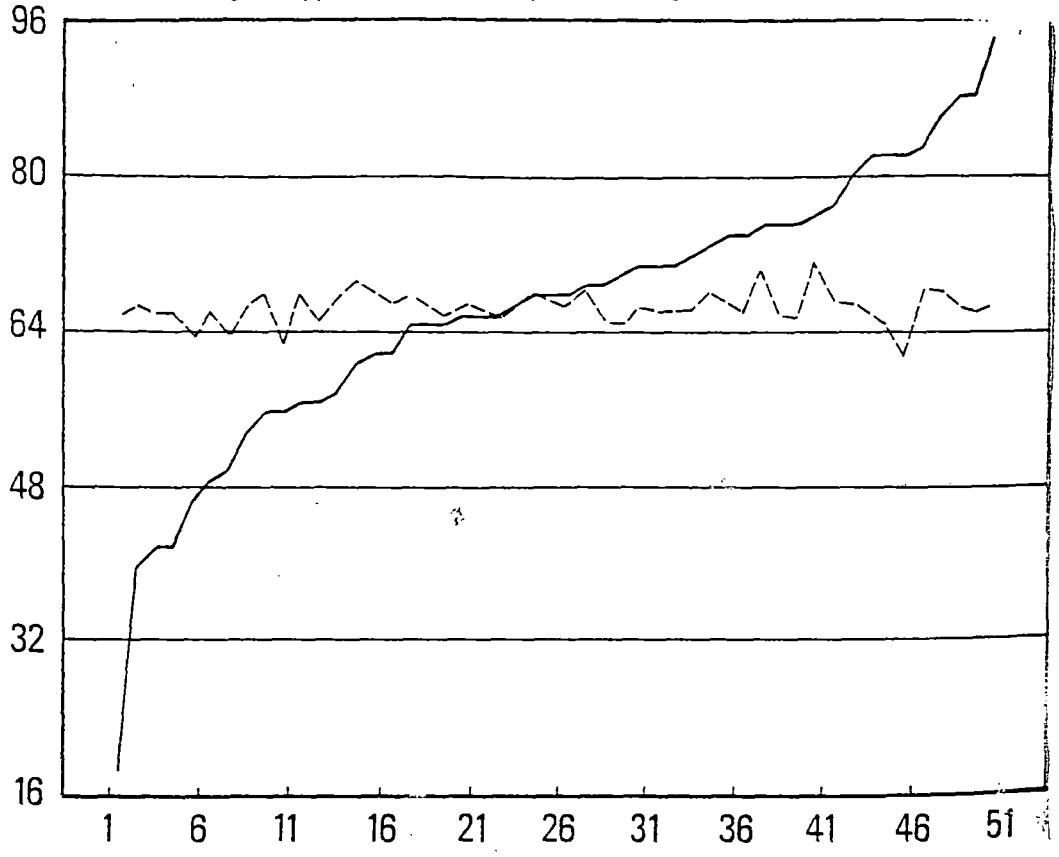
με $R^2 = -0,57$ D-W= 0,609

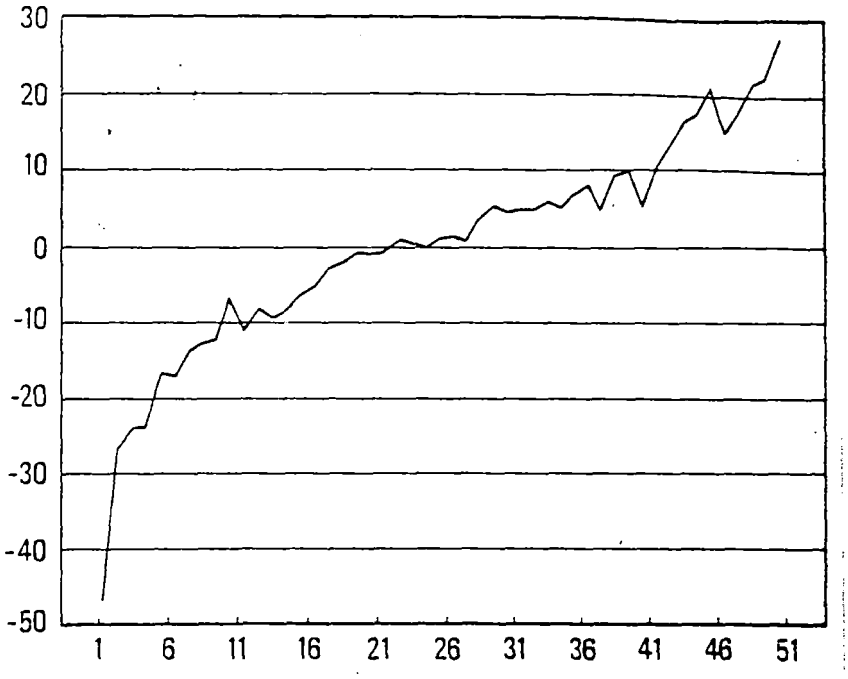
SSE= 15321 s= 17,68

Σημειώνεται ότι η αρνητική τιμή του R^2 στο δεύτερο μοντέλο οφείλεται στο ότι εφαρμόσαμε τον ίδιο τύπο για τα δύο μοντέλα .

Στα μοντέλα αυτά η χαμηλή τιμή του Durbin-Watson ερμηνεύεται ως εξής: Οι παρατηρήσεις Yi αυτοσυσχετίζονται (ειδικότερα $\Gamma_1=0,807$) ενώ οι Xi όχι. Εξάλλου,ο συντελεστής συσχέτισης των (Xi ,Yi) είναι χαμηλός επομένως η X δεν ερμηνέυει την συμπεριφορά των παρατηρήσεων Xi. Αυτό έχει ως αποτέλεσμα τα κατάλοιπα της εκτίμησης να αυτοσυσχετίζονται οπότε και η τιμή του D-W είναι μικρή .

Οι ακόλουθες γραφικές παραστάσεις είναι διαφωτιστικός.





Γραφική παράσταση των $\hat{\varepsilon}_i$ από το μοντέλο $Y_i=b_o+b_1X_i+\varepsilon_i$

ΕΦΑΡΜΟΓΗ

Εξετάζουμε την εξής εταιρεία:

V.A.X A.E

Προιόν : υδρύαλος

Αποτελείται από άμμο (χαλαζιακή) και σόδα (ανθρακκή)

και τα οποία αποτελούν τις πρώτες ύλες.Βρίσκονται σε θερμοκρασία κλιβάνου

1200-1500 C^0 και δημιουργούν μια άμορφη μάζα γυαλιού τη βάζουμε σε κάποια μηχανήματα που λέγονται κλίβανοι υψηλής ατμοσφαιρικής πίεσης,με αποτέλεσμα από τη μεγάλη θερμοκρασία να λιώνει και να γίνεται υγρό.

Αυτό λέγεται υδρύαλος και έχει τις εξείς εφαρμογές :

1. Χρησιμοποιείται σαν κόλλα σε χαρτοβιομηχανίες
2. Χρησιμοποιείται σε βιομηχανίες απορρυπαντικών (δημιουργία πράσινων και μπλέ κόκκων)

3.Χρησιμοποιείται σε τεχνικές κατασκευές (φράγματα -τούνελ).

Πωλείται σε βυτία ή σε βαρέλια 25-30 τόννων και η δυναμικότητα της εταιρείας είναι 80-90.000 κιλά την ημέρα .

Δηλαδή η δυναμικότητα του υδρίαλου κυμαίνεται σε 31.025 τόννους ετησίως

(85 χ 365 ημέρες).

Η απορροφητικότητα (ζήτηση) της αγοράς στην Ελλάδα είναι γύρω στους 15.000

τόννους που σημαίνει ότι με βάση τις 20 εργάσιμες ημέρες χ 12 μήνες βγάζουμε με

20.0 τόννους .
Βέβαια ,ανάλογα με τις βάρδιες εξαρτάται και η παραγωγή.

Επίσης ,υπάρχει πάντα ένα στόκ (stock) που είναι 500-1.000 τόννους το στοκ (stock)
είναι 1 ½ εβδομάδας.
Φυσικά ,η παραγωγή εξαρτάται από την ζήτηση .
Το δυναμικό της εταιρείας τα τελευταία 5 χρόνια είναι σταθερό.Στο τομέα της παραγωγής απασχολούνται 7 εργάτες ,οι οποίει είναι :ένας χημικός ,ένας θερμαστής ,

ένας ηλεκτροτεχνίτης ,ένας χειριστής γερανογέφυρες και τρείς (3) ανειδικευτοι εργάτες.

Εξετάζουμε την παραγόμενη ποσότητα στα έτη 1991-1995.

Δηλαδή:

1991: Η παραγωγή είναι 14.000 τόννοι
1992: Η παραγωγή είναι 15.000 τόννοι
1993: Η παραγωγή είναι 16.000 τόννοι
1994: Η παραγωγή είναι 16.000 τόννοι
1995: Η παραγωγή είναι 17.000 τόννοι

Η συντήρηση γίνεται από εξωτερικά συνεργεία.

Το Κόστος Παραγωγής αναλύεται ως εξής:

1. Κόστος Α' υλών: 100 κιλά υδρύαλου περιέχουν 40-42 κιλά στέρεη υδρύαλος το οποίο 80% είναι άμμος και 50% είναι σόδα (υπάρχει κάποια φύρα). Το κόστος
της σόδας είναι 48 δρχ. το κιλό και της άμμου είναι 12 δρχ. το κιλό .

2. Γ.Β.Ε είναι:

α. Εργατικά = 7 εργάτες / κόστος = 2.200.000 δρχ. το μήνα χ 14 μήνες.

β. Μαζούτ. = κόστος ,57 δρχ. το κιλό /χρειάζεται 12 κιλά μαζούτ ο τόνος του προιόντος.

γ. Δ.Ε.Η = 400.000 δρχ. το μήνα

δ. Η συντήρηση είναι 200.000 δρχ. το μήνα

3. Η μεταφορά : 2,5 δρχ το κιλό
4. Δοικητικά έξοδα: 1.200.000 δρχ. το μήνα
5. Χρηματοοικονομικά έξοδα: 8 δρχ. το κιλό
6. Αποσβέσεις (βιομηχανικό κόστος): 1,5 δρχ. στα Γ.Β.Ε και 0,5 στα Διαθέσεως.

Η τιμή πώλησης στις βιομηχανίες απορρυπαντικών είναι 27 δρχ. το κιλό, στις κατασκευαστικές εταιρείες (επειδή υπάρχει μεγάλη απόσταση) είναι 35 δρχ. το κιλό και για τις κόλλες σε χαρτοβιομηχανίες είναι 40 δρχ. το κιλό.

Στις βιομηχανίες απορρυπαντικών αποτελεί το 60% της παραγωγής ,στις κατασκευαστικές 30% και στις χαρτοβιομηχανίες (κόλλες) το 10%.

Η τιμή του προιόντος
Το 1991 ήταν κατά μέσο όρο 40 δρχ. το κιλό
Το 1992 43 δρχ. το κιλό
Το 1993 45 δρχ. το κιλό
Το 1994 40 δρχ. το κιλό
Το 1995 35 δρχ. το κιλό

Στον πίνακα που ακολουθεί δίνονται το κόστος παραγωγής και η τιμή πώλησης σε σταθερές τιμές:

Ζητείται

i. Να εκτιμηθεί ένα γραμμικό μοντέλο για την τιμή πώλησης με ερμηνευτική μεταβλητή το κόστος παραγωγής.
ii. Να ερμηνευτούν οι παράμετροι του μοντέλου
iii. Να υπολογοστούν τα τυπικά σφάλματα των b_0 και b_1 καθώς και ο συντελεστής προσδιορισμού.
iv. Να ελεγχθεί η υπόθεση ότι οι μεταβλητές X και Y ισούται γραμμικά με 0,7 στο επίπεδο σημαντικότητας $\alpha = 0,05$
v. Να ελεγχθεί η υπόθεση ότι οι μεταβλητές Χκαι Y είναι ίσες στο επίπεδο σημαντικότητας $\alpha = 0,01$
vi. Να υπολογιστεί ο συντελεστής αυτοσυσχέτισης.

ΕΤΟΣ	ΕΤΗΣΙΟ ΚΟΣΤΟΣ ΠΑΡΑΓΩΓΗΣ X_i	ΕΤΗΣΙΑ ΤΙΜΗ ΠΩΛΗΣΗΣ Y_i
1991	146,200	116,529
1992	147,468	121,826
1993	162,485	128,463
1994	174,825	138,954
1995	190,871	149,849

ΛΥΣΗ

i. Συμβολίζουμε με Y την τιμή πώλησης και με X το κόστος παραγωγής και θα εκτιμήσουμε το μοντέλο.

$Yi = b_0 + b_1 Xi + Ei$

Για το σκοπό αυτό υπολογίζουμε:

$SXi = 3580,765$	$Syi = 2827,419$	$SXi\ Yi = 727848,1$
$\bar{X} = 238,718$	$\bar{Y} = 188,465$	$SXi^2 - n\bar{X}^2 = 70393,46$
$SXi^2 = 925187,71$	$SYi^2 = 572826,52$	$SYi^2 - n\bar{Y}^2 = 39871,04$

οπότε

$$\hat{b}_1 = \frac{SXi - nXY}{\Sigma X_i^2 - n\bar{X}^2} = \frac{727848,1 - 15(238,718).(188,495)}{70393,46} = 0,75137$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} = 188,495 - (0,75137)(238,718) = 9,1297$$

ii. Στο μοντέλο αυτό ο συντελεστής $\hat{b}_1 = 0,75137$ είναι η μέση τιμή πώλησης.

Δηλαδή για κάθε αύξηση του ετήσιου κόστους παραγωγής κατά μία μονάδα ή αναμενόμενη αύξηση της τιμής πώλησης ισούται με 0,75137 ή 751,37 εκ.δρχ του 1993.Ο συντελεστής $\hat{b}_0 = 9,1297$ θα μπορούσε να ερμηνευτεί ως αυτόνομο τμήμα της τιμής της πώλησης δηλαδή το ύψος της όταν το κόστος παραγωγής

ισούτε με μηδέν.Όμως η τιμή X= 0 είναι πολύ μακριά από το εύρος των δεδομένων μας και επομένως η προέκταση του μοντέλου ως εκεί είναι παρακινδυνευμένη.

iii. Υπολογίζουμε

$$S^2 = \frac{1}{n-2} S\hat{e}i^2$$

$$= \frac{1}{n-2} (SYi^2 - b_0 S\,Yi - b_1 SXi\,Yi)$$

$$= \frac{1}{3} (572826,52) - (9,1297)(2827,419) -$$

$$-(0,75137) .(727848,1)$$

$$= \frac{130,98}{3} = 10,075$$

και

$$S = 10,075 = 3,174$$

Επομένως

$$S\hat{b}_1 = S \sqrt{\frac{1}{SXi^2 - nX^2}} = 3,174 \sqrt{\frac{1}{70393,46}} = 0,01196$$

$$S\hat{b}_0 = S \sqrt{\frac{1}{n} + \frac{X^2}{SXi - nX^2}} = 3,174 \sqrt{\frac{1}{3} + \frac{(238,718)^2}{70393,46}} = 2,97$$

$$R^2 = 1 - \frac{S\,\varepsilon i^2}{SYi^2 - nY^2} = 1 - \frac{130,98}{39871,04} = 0,9934$$

Τώρα μπορούμε να παρουσιάσουμε τα αποτελέσματα της εκτίμησης σύμφωνα εεεμε την πάγια τακτική, δηλ. ως εξής:

$$Y= 9,1297+ 0,75137X$$
$$(3,07) \quad (62,80)$$

μέ R^2=9,1297 d=1,191597

SSE=130,98 S=3,17

Για n=15 και κ=1 βρίσκουμε στον πίνακα 8 στο τέλος της εργασίας τις τιμές d_l=1,08 και d_u=1,36. Ισχύει:

$$d_l<d<d_u$$

δηλαδή βρισκόμαστε στην περιοχή αβεβαιότητας του ελέγχου.
Ακολουθούμε την καθιερωμένη για τις οικονομικές χρονικές στιγμές τακτική και θεωρούμε ότι και αυτή η περιοχή είναι η περιοχή απόρριψης της μηδενικής υπόθεσης H_o : ρ=0 υπέρ της $H_ε$:ρ>0. Η μέθοδος Cochrane-Orcutt που περιγράψαμε εφαρμόζετε σταδιακά ως εξής:

- Υπολογίζουμε $\hat{\rho}$=1-d/2=1-1,191597/2=0,4042

- Παίρνουμε τις οιονές διαφορές των δεδομένων , δηλαδή τις

  Yi*=Yi-0,4042Yi-1, Xi*=Xi-0,4042Xi-1, i=2,....,5
- και έχουμε:

Έτος	$Y_i^*=Y_i-0,4042Y_{i-1}$	$X_i^*=X_i-0,4042X_{i-1}$
1992	74.72498	88.3796
1993	79.22093	102.8784
1994	87.02925	109.1486
1995	93.683779	120.2067

• Εκτιμούμε ξανα το μοντέλο χρησιμοποιώντας ως δεδομένα τις μετασχηματισμένες τιμές (Yi*, Xi*) i=1

Επαναληπτική εφαρμογή της διαδικασίας αυτής έδωσε τα ακόλουθα αποτελέσματα (οι αριθμοί στις παρενθέσεις είναι t- αναλογίες των αντίστοιχων εκτιμήσεων):

$$Y= 10,88367 + 0,7463073\chi$$
$$(2,09) \qquad (37,63)$$

$R^2= 0,9968$ $\rho= 0,3756$
SSE=108,6 $t\rho=1,24$
s=3,14 D-W=1,49

Ο ζητούμενος έλεγχος θα γίνει σταδιακά ως εξής:

Ho: $b_1 = 0,7$
Hε: $b_1 =0$

Επειδή
$|t_n| = \hat{b}_1-0,7/Sb_1 = 0,75137-0,7/0,01196=4,29>t_{n-1,\alpha/2}=$

$=t_{13(0,025)}= 2,16$ απορρίπτουμε την Ho

Στο μοντέλο

$$Yi= bo+b_1x_i+\varepsilon_i$$

μέ $E(Yi) = bo+b1X\iota$

η τιμή πώλησης ισούται μέ b1 και το ετήσιο κόστος παραγωγής στο σημείο $\chi=\chi_i$ είναι η

$E(Yi)/Xi= bo/Xi +b1$

Αν bo=0 τότε τιμή πώλησης = κόστος παραγωγής= b1.
Επομένως ο ζητούμενος έλεγχος είναι ισοδύναμος μέ τον έλεγχο της μηδενικής υπόθεσης $b_0=0$ που γίνεται ως εξής:

Ho:bo=0
Hε:bo>0

επειδή

$tn=bo/sbo=9,1297/2,97=3,072>t_{13(0,01)}=2,65$

η Ho απορρίπτεται.

Για να υπολογίσουμε τον συντελεστή αυτοσυσχέτισης πρώτης τάξης στις τιμές Xi καταστρώνουμε τον ακόλουθο πίνακα:
απ' όπου υπολογίζουμε

$$\bar{X}=\sum_{i=1}^{5} Xi/5=151,32$$

$$\sum_{i=1}^{1} (Xi- \bar{X})^2= 2959,4$$

$$\sum_{i=2}^{5} (Xi- \bar{X}) (X_{i-1}-\bar{X})^2=7487,1$$

$$\sum_{i=3}^{5}(Xi-\bar{X}) (X_{I-2}-\bar{X})^2= 1733,59$$

$$\Gamma 1= 7487,1/2959,4= 0,759$$

οπότε

$$\Gamma 2= 1733,59/2959,4= 0,586$$

Μέ τρόπο ανάλογο υπολογίζουμε τους ακόλουθους δύο συντελεστές για την χρονική σειρά Yi , i=1,.....,5

ΕΤΟΣ	Xi	Xi-1
1992	115,11	100,00
1993	124,00	115,11
1994	128,17	124,00
1995	133,52	128,17

ΕΤΟΣ	Yi	Yi-1
1992	174,64	156,80
1993	163,38	174,64
1994	173,67	163,38
1995	177,01	173,67

Γ_1= 0,792

Γ_2= 0,624

Είναι έυκολο να ελεγθεί ότι ισχύει :

$$\Gamma_{\chi\psi}=0,83$$

Έτσι, υπολογίσαμε τους συντελεστές αυτοσυσχέτισης πρώτης και δέυτερης τάξης για:

- τον αριθμό της ετήσιας παραγωγής Γ_1=0,759 και Γ_2=0,586
- τον αριθμό της τιμής πώλησης Γ_1= 0,792 και Γ_2= 0,624

Επειδή n=5 το διάστημα μέσα στο οποίο κρίνεται στατιστικά μη σημαντική η τιμή των Γi είναι το

(-0,23 0,23)

και άρα όλοι οι συντελεστές που υπολογίσαμε είναι στατιστικά σημαντικοί στο επίπεδο σημαντικότητας 2%.

Συμπερασματικά θα μπορούσαμε να πούμε ότι η πρόβλεψη για τα επόμενα 3-5 έτη, είναι ότι η εταιρία όχι μόνο θα είναι βιώσιμη αλλά θα έχει ανοδική πορεία και αν λάβουμε υπόψη τους Ισολογισμούς των τελευταίων 5 ετών, το καθαρό κέρδος της επιχείρησης θα κυμαίνεται (και στα επόμενα 3-5 έτη) από 57.200.000 δρχ έως 62.350.000 δρχ.
Επομένως, κάτω από αυτούς τους υπολογισμούς χρησιμοποιώντας την στατιστική μέθοδο της συσχέτισης -παλινδρόμησης μπορούμε να μιλάμε για μια υγιή και κερδοφόρα επιχείρηση ορθά στεκούμενη στην ανταγωνιστική μας κοινωνία .

Παρακάτω παραθέτονται οι Ισολογισμοί της εταιρίας ΑΦΟΙ Δ. ΣΗΜΙΤΗ Α.Ε κατά τα έτη 1991-1995.

ΑΦΟΙ Δ. ΣΗΝΙΤΗ ΑΝΩΝΥΜΗ ΕΤΑΙΡΕΙΑ

ΙΣΟΛΟΓΙΣΜΟΣ 31ης ΔΕΚΕΜΒΡΙΟΥ 1991–16η ΕΤΑΙΡΙΚΗ ΧΡΗΣΗ
(1 ΙΑΝΟΥΑΡΙΟΥ – 31 ΔΕΚΕΜΒΡΙΟΥ 1991)
ΕΔΡΑ:ΥΠΑΤΟ ΘΗΒΑΣ ΑΡΙΘ.ΜΗΤΡΩΟΥ ΝΟΜΑΡΧΙΑΣ ΒΟΙΩΤΙΑΣ :
6662 / 11 /78 / 86 /0

ΕΝΕΡΓΗΤΙΚΟ	Ποσά προηγούμενης χρήσεως 1990 Αξία κτήσ.	Αποσβ.-Αναπόσβεστη αξία		Ποσά κλειομένης χρήσεως 1991 αξία κτήσ.	Αποσβ.-Αναπόσβεστη αξία	
ΕΞΟΔΑ ΕΓΚ/ΣΕΩΣ						
Εξοδα ιδρύσ.κ' πρώτης εγκ.σεως	16.369.084	6.123.582	10.245.502	16.369.084	7.760.490	8.608.594
Τόκοι δανείων κατασκευαστικής περιόδου	164.813.307	49.620.090	115.193.217	317.428.950	82.582.750	234.846.200
	181.182.391	55.734.672	125.438.719	333.798.034	90.343.240	243.454.794
ΠΑΓΙΟ ΕΝΕΡΓΗΤΙΚΟ						
Ενσώμ.ακιν/σεις						
Γήπεδα-Οικόπεδα	59.494.050	–	59.494.050	59.494.050	–	59.494.050
Κτίρια κ' τεχν.έργα	110.423.009	45.487.023	64.935.986	118.953.008	50.548.857	68.404.151
Μηχαν.-τεχν.εγκ. λοιπ.μηχ.εξοπλισμ.	367.103.749	187.136.913	179.966.836	432.958.014	219.116.068	213.841.957
Μεταφορ.μέσα	24.904.708	14.397.574	10.507.134	23.038.578	16.243.802	6.794.776
Επιπλ.κ'λοιπ.εξοπλ.	14.691.164	7.504.624	7.186.540	15.430.832	9.763.921	5.666.911
Ακιν.υπό εκτέλ. κ' προκαταβολές	1.882.353	225.082	1.656.471	1.882.353	451.764	1.430.589
Σύνολο ακιν/σεων	578.499.033	254.752.016	323.747.017	651.756.845	296.124.411	355.632.434
ΣΥΝΟΛΟ ΠΑΓΙΟΥ ΕΝΕΡΓΗΤΙΚΟΥ			323.747.017			355.632.434
ΚΥΚΛΟΦ.ΕΝΕΡΓΗΤΙΚΟ						
Αποθέματα						
Προϊόντα έτοιμα κ' ημιτελή						
Υποπροϊόντα κ' υπολείμματα			70.365.555			58.466.082
Α' κ' βοηθ.ύλες-αναλώσιμα						
Ανταλλ/κά κ' είδη συσκ/σίας			24.589.794			14.741.951
Προκ/λές για αγορ.αποθεμ.			5.940.786			4.761.900
			100.896.135			75.969.933
Απαιτήσεις						
Πελάτες			58.214.308			134.927.832
Γραμ/α εισπρακτέα						
Χαρτοφυλακίου		20.254.850			4.211.645	
Σε τραπ.για εγγύηση		50.115.048	70.469.898		4.662.660	8.874.305
Επιταγές καθυστέρηση			3.766.719			625.000
Βραχυπρ.απαιτ.κατά συνδεδεμένων επιχειρήσεων			40.527.492			47.843.609
Απαιτ.κατά οργάνων διοικήσεως			51.594.529			12.507.401
Χρεώστες διάφοροι			118.355.466			164.574.432
Λογ/σμοί διαχ/σεως προκατ.κ' πιστώσεων			6.887.174			6.818.185
			350.4.5.586			376.270.765
ΔΙΑΘΕΣΙΜΑ						
Ταμείο			3.770.542			2.525.438
Καταθέσεις όψεως κ' προθεσμίας			543.661			246.393
			4.314.203			2.771.831
ΣΥΝΟΛΟ ΚΥΚΛΟΦΟΡΟΥΝΤΟΣ ΕΝΕΡΓΗΤΙΚΟΥ			455.625.924			455.012.529
ΓΕΝΙΚΟ ΣΥΝΟΛΟ ΕΝΕΡΓΗΤΙΚΟΥ			904.811.660			1.054.090.757

ΠΑΘΗΤΙΚΟ	Σε δρχ.	Σε Ξ.Ν.	Ποσά προηγ.χρησ. 1990	Ποσά κλειομ. χρησ.1991
A.ΙΔΙΑ ΚΕΦΑΛΑΙΑ				
I.Μετοχικό Κεφάλαιο (195.000 μετ.των 1000 δρχ)			195.000.000	195.000.000
III.Διαφορές αναπροσαρμογής				
3.Επιχορ.παγ.ενεργητικού			70.097.470	70.097.470
			265.097.470	265.097.470
IV.Αποθεματικά κεφάλαια				
1.Τακτικό αποθεματικό			2.459.319	2.459.3 9
3.Ειδικά αποθεματικά			40.000	40.000
4.Εκτακτα αποθεματικά			–	–
5.Αφορολόγητα αποθεματικά ειδικών διατάξεων νόμων			60.004.530	60.094.530
			62.503.849	62.593.849
V.Αποτελέσματα εις νέο				
Υπόλοιπο ζημιών χρήσεως εις νέο			(135.927.609	(55.784.012)
Υπόλοιπο ζημιών προηγ. χρήσεων			(36.793.514)	(172.721.123)
			(172.72...231	228.505.135
Σύνολο ιδίων κεφαλαίων			154.280.196	99.096.184
Γ.ΥΠΟΧΡΕΩΣΕΙΣ				
I.Μακροπρόθεσμες υποχρεώσεις				
2.Δάνεια Τραπεζών			386.001.524	538.617.167
			386.001.524	538.617.167
II.Βραχυπρόθεσμες υποχρεώσεις				
1.Προμηθευτές			21.065.811	90.274.980
2.Γραμ.πληρ.	25.188.800	61.762.540	86.951.340	90.589.568
2a.Επιτ.πληρωτ.			9.554.819	32.938.478
3.Τράπεζες λογ/σμοί βραχυπρόθ.υποχρ.			18?.568.964	134.353.881
4.Προκαταβολές πελατών			4.226.223	8.573.899
5.Υποχρεώσεις από φόρους-τέλη			8.876.058	9.528.636
6.Ασφαλιστικοί οργανισμοί			5.450.530	11.969.366
11.Πιστωτές διάφοροι			12.249.771	16.926.462
	25.188.800	61.762.540	329.943.016	413.153.218
Σύνολο υποχρεώσεων			715.944.540	951.770.385
Δ.ΜΕΤΑΒ/ΚΟΙ ΛΟΓ/ΣΜΟΙ ΠΑΘΗΤΙΚΟΥ				
2.Εξοδα χρήσ.δουλευμένα			1.788.379	–
3.Λοιποί μεταβατικοί λ/σμοί			32.198.545	3.233.188
			33.986.924	3.233.189
ΓΕΝΙΚΟ ΣΥΝΟΛΟ ΠΑΘΗΤΙΚΟΥ			904.811.660	1.054.090.757

ΑΝΑΛΥΣΗ ΛΟΓΑΡΙΑΣΜΟΥ ΑΠΟΤΕΛΕΣΜΑΤΑ ΧΡΗΣΕΩΣ

ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΚΜΕΤΑΛΛΕΥΣΕΩΣ		Ποσ.προηγ.χρ.1990		Ποσ.κλ.χρ.1991
Πωλήσεις		454.346.616		491.301.025
Κόστος πωλήσεων		[illegible]		[illegible]
Μικτά αποτελέσματα εκμ/σεως		[illegible]		[illegible]
Πλέον: Αλλα έσοδα εκμετ/σεως		[illegible]		[illegible]
		82.786.292		72.220.441
Μείον: 1.Εξοδα διοικ.λειτουργ. 3.Εξοδα λειτουργ.διαθέσεως	[illegible]	104.084.264	[illegible]	78.270.099
Μερικά αποτ.(Ζημιές-κέρδη)εκμ/σεως		(21.298.972)		(6.050.252)
Πλέον: Εσοδα συμμετοχών. Μείον: Χρεωστ.τόκοι κ' συναφή έξοδα	109.3?2.7??	(106.734.?34)	44.155.1??	44.155.154
Ολικά αποτ/τα(Ζημιές) εκμ/σεως		(128.030.406)		(50.213.406)
ΠΛΕΟΝ:ΕΚΤΑΚΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ				
Εκτακτα κ' ανόργανα έξοδα		7.897.203		5.570.606
Οργανικά κ' έκτακτα αποτ/τα (Ζημιές)		(135.927.609)		(55.784.012)
Σύνολο αποσβ.παγίων στοιχείων	77.028.212		76.289.205	
Μείον: Οι από αυτές ενσωματωμένες στο λειτουργικό κόστος	77.028.212	–	76.289.205	–
ΚΑΘΑΡΑ ΑΠΟΤΕΛΕΣΜΑΤΑ (ΖΗΜΙΕΣ) ΧΡΗΣΕΩΣ ΠΡΟ ΦΟΡΩΝ		(135.927.609)		(55.784.012)

ΠΙΝΑΚΑΣ ΔΙΑΘΕΣΕΩΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

	Ποσ.προηγ.χρ.1990	Ποσ.κλ.χρ.1991
Καθαρά αποτ/τα (Ζημιές)χρήσεως	135.927.609	55.781.014
Υπόλοιπ.αποτ/των (Ζημιών) προηγ. χρήσεων	36.793.514	172.721.123
Υπόλοιπο ζημιών εις νέο	172.721.123	228.505.135

ΥΠΑΤΟ ΘΗΒΑΣ 28 ΦΕΒΡΟΥΑΡΙΟΥ 1992

ΠΡΟΕΔΡΟΣ ΤΟΥ Δ.Σ.	Ο ΥΠΕΥΘΥΝΟΣ ΛΟΓΙΣΤΗΡΙΟΥ
ΣΤΑΘΗΣ ΣΗΝΙΤΗΣ ΤΟΥ ΔΗΜΗΤΡΙΟΥ	ΑΝΔΡΕΑΣ ΜΑΡΚΟΣ ΤΟΥ ΝΙΚΟΛΑΟΥ
Α.Δ.Τ. 678640/ΙΣΤ΄ΑΘΗΝΩΝ	Α.Δ.Τ. Σ 120606/ΚΣΤ΄ΑΘΗΝΩΝ

ΑΦΟΙ Δ. ΣΗΜΙΤΗ ΑΝΩΝΥΜΗ ΕΤΑΙΡΕΙΑ

ΕΔΡΑ: ΥΠΑΤΟ ΘΗΒΑΣ ΑΡΙΘ. ΜΗΤΡΩΟΥ ΝΟΜΑΡΧΙΑΣ ΒΟΙΩΤΙΑΣ: 6662/11/Β/86/0

ΙΣΟΛΟΓΙΣΜΟΣ ΤΗΣ 31ης ΔΕΚΕΜΒΡΙΟΥ 1993 - 10η ΕΤΑΙΡΙΚΗ ΧΡΗΣΗ (1 ΙΑΝΟΥΑΡΙΟΥ - 31 ΔΕΚΕΜΒΡΙΟΥ 1993)

(ΠΟΣΑ ΣΕ ΔΡΑΧΜΕΣ)

ΕΝΕΡΓΗΤΙΚΟ

	Ποσά κλειομένης χρήσεως 1993 Αξία Κτήσεως	Αποσβέσεις	Αναπόσβεστη Αξία	Ποσά προηγούμενης χρήσεως 1992 Αξία Κτήσεως	Αποσβέσεις	Αναπόσβεστη Αξία
Β. ΕΞΟΔΑ ΕΓΚΑΤΑΣΤΑΣΕΩΣ						
1 Έξοδα ιδρύσεως και πρώτης εγκ/σεως	16.369.084	11.034.306	5.334.778	16.369.084	[illegible]	[illegible]
3 Τόκοι δανείων κατασκ. περιόδου	394.553.671	154.900.777	239.652.894	394.553.671	[illegible]	[illegible]
	410.922.755	165.935.083	244.987.672	410.922.755	[illegible]	[illegible]
Γ. ΠΑΓΙΟ ΕΝΕΡΓΗΤΙΚΟ						
ΙΙ. Ενσώματες ακινητοποιήσεις						
1 Γήπεδα-Οικόπεδα	109.072.425	-	109.072.425	109.072.425		[illegible]
3 Κτίρια και τεχνικά έργα	223.663.248	86.743.368	136.919.880	214.561.2[illegible]	[illegible]	[illegible]
4 Μηχανήματα-τεχνικές εγκ/σεις και λοιπός μηχανολ. εξοπλισμός	432.858.024	296.569.551	136.288.473	432.858.024	[illegible]	[illegible]
5 Μεταφορικά μέσα	21.783.517	17.587.632	4.195.885	21.783.517	[illegible]	[illegible]
6 Έπιπλα και λοιπός εξοπλισμός	15.980.832	14.474.521	1.506.311	15.980.832	[illegible]	[illegible]
7 Ακινητοποιήσεις υπό εκτέλεση και προκαταβολές	1.882.353	903.528	978.825	1.882.353	677.646	1.204.707
Σύνολο ακινητοποιήσεων (ΓΙΙ)	805.340.399	416.278.600	389.061.799	785.[illegible]	[illegible]	[illegible]

	1993		1992	
Δ. ΚΥΚΛΟΦΟΡΟΥΝ ΕΝΕΡΓΗΤΙΚΟ				
Ι. Αποθέματα				
2 Προϊόντα έτοιμα και ημιτελή υποπροϊόντα και υπολείμματα		7.796.632		[illegible]
4 Α' και βοηθητικές ύλες-αναλώσιμα υλικά-ανταλλακτικά και είδη συσκευασίας		51.865.478		[illegible]
5 Προκαταβολές για αγορές αποθεμάτων		3.583.882		[illegible]
		63.247.992		[illegible]
ΙΙ. Απαιτήσεις				
1. Πελάτες		30.987.674		[illegible]
2. Γραμμάτια Εισπρακτέα				
- Χαρτοφυλακίου	-		[illegible]	
- Στις Τράπεζες για εγγύηση	5.516.093	5.516.093	[illegible]	[illegible]
3 Γραμμάτια σε καθυστέρηση		825.000		[illegible]
5 Βραχυπρόθεσμες απαιτήσεις κατά συνδεδεμένων επιχειρήσεων		14.025.937		[illegible]
11 Χρεώστες διάφοροι		217.306.147		[illegible]
12 Λογαριασμοί διαχειρίσεως προκαταβολών και πιστώσεων		4.909.184		[illegible]
		281.836.845		[illegible]
IV. Διαθέσιμα				
1. Ταμείο		1.837.184		[illegible]
3 Καταθέσεις όψεως και προθεσμίας		833.618		[illegible]
		2.670.802		[illegible]
Σύνολο Κυκλοφορούντος Ενεργητικού (ΔΙ+ΔΙΙ+ΔΙV)		347.548.639		[illegible]
ΓΕΝΙΚΟ ΣΥΝΟΛΟ ΕΝΕΡΓΗΤΙΚΟΥ (Β+Γ+Δ)		981.598.310		[illegible]
ΛΟΓΑΡΙΑΣΜΟΙ ΤΑΞΕΩΣ ΧΡΕΩΣΤΙΚΟΙ				
1 Αλλότρια περιουσιακά στοιχεία		805.182		-

ΠΑΘΗΤΙΚΟ

	Ποσά κλειομένης χρήσεως 1993	Ποσά προηγούμενης χρήσεως 1992
Α ΙΔΙΑ ΚΕΦΑΛΑΙΑ		
Ι Μετοχικό Κεφάλαιο (185.000 μετοχές των 1.000 δρχ.)		
1 Καταβλημένο	185.000.000	185.000.000
ΙΙΙ Διαφορές αναπροσαρμογής - Επιχορηγήσεις επενδύσεων		
3 Επιχορηγήσεις επενδύσεων πάγιου ενεργητικού	73.574.470	73.574.470
IV. Αποθεματικά κεφάλαια		
1 Τακτικό αποθεματικό	2.459.319	2.459.319
3 Ειδικά αποθεματικά	40.000	40.000
5 Αφορολόγητα αποθεματικά ειδικών διατάξεων νόμων	60.004.530	60.004.530
	62.503.849	62.503.849
V Αποτελέσματα εις νέο		
Υπόλοιπο ζημιών χρήσεως εις νέο	29.002.075	(126.635.758)
Υπόλοιπο ζημιών προηγ. χρήσεων	263.071.131	(136.435.373)
	292.073.206	(263.071.131)
Σύνολο Ιδίων Κεφαλαίων	39.005.113	68.007.188
Β ΠΡΟΒΛΕΨΕΙΣ ΓΙΑ ΚΙΝΔΥΝΟΥΣ & ΕΞΟΔΑ		
1 Προβλέψεις για αποζημ. προσωπικού λόγω εξόδου από την υπηρεσία	1.600.000	2.500.000
Γ. ΥΠΟΧΡΕΩΣΕΙΣ		
Ι. Μακροπρόθεσμες υποχρεώσεις		
2 Δάνεια Τραπεζών	708.680.045	708.680.045
ΙΙ Βραχυπρόθεσμες υποχρεώσεις		
1 Προμηθευτές	27.524.358	53.063.344
2 Γραμμάτια πληρωτέα	31.715.211	42.209.659
2α Επιταγές πληρωτέες	9.010.709	6.212.677
3 Τράπεζες λ/βραχυπρ. υποχρ.	107.326.675	123.026.487
4 Προκαταβολές πελατών	6.629.100	5.583.787
5 Υποχρεώσεις από φόρους-τέλη	7.811.449	10.670.216
6 Ασφαλιστικοί οργανισμοί	9.776.508	9.880.976
11 Πιστωτές διάφοροι	13.162.817	29.578.711
	214.953.910	278.024.017
Σύνολο Υποχρεώσεων (ΓΙ + ΓΙΙ)	923.633.955	986.704.062
Δ ΜΕΤΑΒΑΤΙΚΟΙ ΛΟΓΑΡΙΑΣΜΟΙ ΠΑΘΗΤΙΚΟΥ		
3 Λοιποί μεταβατικοί λογαριασμοί	17.359.242	-
ΓΕΝΙΚΟ ΣΥΝΟΛΟ ΠΑΘΗΤΙΚΟΥ (Α+Β+Γ+Δ)	981.598.310	1.057.211.250
ΛΟΓΑΡΙΑΣΜΟΙ ΤΑΞΕΩΣ ΠΙΣΤΩΤΙΚΟΙ		
1 Δικαιούχοι αλλότριων περιουσ. στοιχείων	805.182	-

ΣΗΜΕΙΩΣΕΙΣ:
1 Επί των ακινήτων της εταιρείας έχουν εγγραφεί υποθήκες και προσημειώσεις [illegible] δρχ. 81.074.541 και Β. Σε ασφάλεια Τραπεζικών δανείων ποσού δρχ. 510.082.000.
2 Αποθέματα δρχ. 10.136.307 είναι ενεχυριασμένα υπέρ της ΠΑΕΓΑΕ.

ΑΝΑΛΥΣΗ ΛΟΓΑΡΙΑΣΜΟΥ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΧΡΗΣΕΩΣ

	Ποσά κλειομένης χρήσεως 1993		Ποσά προηγούμενης χρήσεως 1992	
Ι. Αποτελέσματα εκμεταλλεύσεως				
Κύκλος εργασιών		428.305.233		[illegible]
Μείον: Κόστος πωλήσεων		380.875.240		[illegible]
Μικτά αποτελέσματα (κέρδη) εκμεταλλεύσεως		47.429.993		[illegible]
Πλέον Άλλα έσοδα εκμεταλλεύσεως		2.989.474		[illegible]
Σύνολο		50.419.467		[illegible]
Μείον: 1 Έξοδα διοικητικής λειτουργίας	45.597.870		[illegible]	
3 Έξοδα λειτουργίας διαθέσεως	22.376.132	67.974.002	[illegible]	[illegible]
Μερικά αποτελέσματα (ζημιές) εκμεταλλεύσεως		17.554.535		[illegible]
Πλέον 3 Χρεωστικοί τόκοι και συναφή έξοδα		5.286.395		[illegible]
Ολικά αποτελέσματα (ζημιές) εκμεταλλεύσεως		22.840.930		[illegible]
ΙΙ ΠΛΕΟΝ: Έκτακτα αποτελέσματα				
1 Έκτακτα και ανόργανα έξοδα		6.161.145		[illegible]
Οργανικά και έκτακτα αποτελέσματα (ζημιές)		29.002.075		[illegible]
ΜΕΙΟΝ: Σύνολο αποσβέσεων παγίων στοιχείων	94.067.381		[illegible]	
Μείον: Οι από αυτές ενσωματωμένες στο λειτουργικό κόστος	94.067.381	-	[illegible]	-
ΚΑΘΑΡΑ ΑΠΟΤΕΛΕΣΜΑΤΑ (ΖΗΜΙΕΣ) ΧΡΗΣΕΩΣ ΠΡΟ ΦΟΡΩΝ		29.002.075		[illegible]

ΠΙΝΑΚΑΣ ΔΙΑΘΕΣΕΩΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

	Ποσά κλειομένης χρήσεως 1993	Ποσά προηγούμενης χρήσεως 1992
Καθαρά αποτελέσματα (ζημιές) χρήσεως	29.002.075	126.635.758
Υπόλοιπο αποτελεσμάτων (ζημιών) προηγούμενων χρήσεων	263.071.131	136.435.373
Υπόλοιπο ζημιών εις νέο	292.073.206	263.071.131

ΥΠΑΤΟ ΘΗΒΑΣ 28 ΦΕΒΡΟΥΑΡΙΟΥ 1994

Ο ΠΡΟΕΔΡΟΣ ΤΟΥ Δ.Σ.	ΤΟ ΜΕΛΟΣ ΤΟΥ Δ.Σ.	Ο ΠΡΟΪΣΤΑΜΕΝΟΣ ΟΙΚΟΝΟΜΙΚΩΝ ΥΠΗΡΕΣΙΩΝ
ΜΙΛΤΙΑΔΗΣ ΣΗΜΙΤΗΣ ΤΟΥ ΔΗΜΗΤΡΙΟΥ	ΔΗΜΗΤΡΙΟΣ ΣΗΜΙΤΗΣ ΤΟΥ [illegible]	ΑΝΔΡΕΑΣ ΜΑΡΚΟΣ ΤΟΥ ΝΙΚΟΛΑΟΥ

ΑΦΟΙ Δ.ΣΗΜΙΤΗ ΑΝΩΝΥΜΗ ΕΤΑΙΡΕΙΑ

ΕΔΡΑ: ΥΠΑΤΟ ΘΗΒΑΣ ΑΡΙΘ.ΜΗΤΡ.ΝΟΜΑΡΧΙΑΣ ΒΟΙΩΤΙΑΣ 9992/11/Β/86/0 - ΙΣΟΛΟΓΙΣΜΟΣ 31/12/1995 - 20Η ΕΤΑΙΡΙΚΗ ΧΡΗΣΗ (1/1-31/12/95) (ΠΟΣΑ ΣΕ ΔΡΧ.)

ΙΚΟ	Ποσά Κλειομένης Χρήσεως 1995			Ποσά Προηγούμενης Χρήσεως 1994		
. ΕΓΚΑΤΑΣΤΑΣΕΩΣ	Αξία Κτήσης	Αποσβέσεις	Αναπ. Αξία	Αξία Κτήσης	Αποσβέσεις	Αναποσβ. Αξία
ιδρύσεως & πρώτης εγκατ.	16.369.084	11.034.306	5.334.778	16.369.084	11.034.306	5.334.778
δανείων κατασκ.περιόδου	1.372.505.582	154.900.777	1.217.604.805	394.553.671	154.900.777	239.652.894
	1.388.874.666	165.935.083	1.222.939.583	410.922.755	165.935.083	244.987.672
ΕΝΕΡΓΗΤΙΚΟ						
ατες ακινητοποιήσεις						
ι-οικόπεδα	109.072.425	---	109.072.425	109.072.425	---	109.072.425
& τεχνικά έργα	228.093.248	102.148.479	125.944.769	228.093.248	86.743.368	141.349.880
ηματα-τεχν.εγκαταστάσεις	432.958.024	331.366.701	101.591.323	432.958.024	296.569.551	136.388.473
ορικά μέσα	21.783.517	17.587.632	4.195.885	21.783.517	17.587.632	4.195.855
ι & λοιπός εξοπλισμός	15.980.832	14.474.521	1.506.311	15.980.832	14.474.521	1.506.311
οποιήσεις υπό εκτέλεση						
καταβολές	1.882.353	903.528	978.825	1.882.353	903.528	978.825
ινητοποιήσεων (ΓΙΙ)	809.770.399	466.480.861	343.289.538	809.770.399	416.278.600	393.491.799
ΦΟΡΟΥΝ ΕΝΕΡΓΗΤΙΚΟ						
στα						
τα έτοιμα & ημιτελή						
οιόντα & υπολείμματα			-----			167.700
βοηθητικές ύλες-αναλώσιμα						
νταλ/κά & είδη συσκευασίας			21.694.537			29.700.781
			21.694.537			29.868.481
εις						
ς			73.626.846			55.073.915
ντια Εισπρακτέα						
υλακίου			240.000			1.000.000
άπεζες για εγγύηση			-----			----
ντια σε καθυστέρηση						80.000
τες διάφοροι			281.004.418			282.789.000
ασμοί διαχειρίσεως						
αβολών & πιστώσεων			-5.263.478			-5.453.403
			349.607.786			333.489.512
ιμα						
			6.208.072			7.258.423
σεις όψεως & προθεσμίας			10.952.018			788.969
			17.160.090			9.647.392
λοφορούντος ενεργητικού(ΔΙ + ΔΙΙ + ΔΙV)			388.462.413			372.005.385
ΝΟΛΟ ΕΝΕΡΓΗΤΙΚΟΥ (Β+Γ+Δ)			1.954.691.534			1.010.484.856

ΠΑΘΗΤΙΚΟ	Ποσά Κλειομένης Χρήσεως 1995	Ποσά Προηγ/νης Χρήσεως 1994
Α.ΙΔΙΑ ΚΕΦΑΛΑΙΑ		
Ι.Μετοχικό Κεφάλαιο		
(195.000 μετοχές των 1.000δρχ.)		
1.Καταβλημένο	195.000.000	195.000.000
ΙΙΙ.Διαφορές αναπροσαρμογής		
Επιχορηγήσεις επενδύσεων		
3.Επιχορηγήσεις επενδύσεων παγίου ενεργ.	73.574.470	73.574.470
ΙV.Αποθεματικά Κεφάλαια		
1.Τακτικό αποθεματικό	2.459.319	2.459.319
3.Ειδικά αποθεματικά	40.000	40.000
5.Αφορολόγητα αποθεματικά ειδ.διατ.νόμων	60.004.530	60.004.530
	62.503.849	62.503.849
V.Αποτελέσματα εις νέο		
Διαφορά φορολ.ελέγχου	-28.243.366	---
Υπόλοιπο ζημιών προηγ.χρήσεων	-243.896.019	(292.073.206)
Μείον κέρδη χρήσεως	+ 51.897.465	48.177.184
Υπόλοιπο ζημιών προηγ.χρήσεων	-214.957.472	243.896.019
Σύνολο Ιδίων Κεφαλαίων	116.120.847	87.182.300
Γ.ΥΠΟΧΡΕΩΣΕΙΣ		
Ι.Μακροπρόθεσμες Υποχρεώσεις		
2.Δάνεια Τραπεζών (ΕΤΒΑ)	1.616.631.956	708.680.045
ΙΙ.Βραχυπρόθεσμες Υποχρεώσεις		
1.Προμηθευτές	26.503.219	51.966.665
2.Γραμμάτια πληρωτέα	5.439.355	15.721.385
2α.Επιταγές πληρωτέες	----	3.312.824
3.Τράπεζες λ/βραχυπρ.υποχρ.	89.719.760	89.764.204
5.Υποχρεώσεις από φόρους - τέλη	36.614.866	16.960.332
6.Ασφαλιστικοί οργανισμοί	13.368.218	27.044.960
11.Πιστωτές διάφοροι	22.411.073	9.852.141
	194.056.491	214.622.511
Σύνολο Υποχρεώσεων (ΓΙ+ΓΙΙ)	1.810.688.447	923.302.556
Δ.ΜΕΤΑΒΑΤΙΚΟΙ ΛΟΓΑΡΙΑΣΜΟΙ ΠΑΘΗΤΙΚΟΥ		
3.Λοιποί μεταβατικοί λογαριασμοί	27.882.240	---
ΓΕΝΙΚΟ ΣΥΝΟΛΟ ΠΑΘΗΤΙΚΟΥ (Α+Γ+Δ)	1.954.691.534	1.010.484.856

ΙΣ: 1.Επί των ακινήτων της εταιρείας έχουν εγγραφεί υποθήκες και προσημειώσεις : α)Υπέρ του Ελληνικού Δημοσίου δρχ. 61.974.541 και β) Σε ασφάλεια Τραπεζικών δανείων ποσού δρχ. 510.062.000

ΑΝΑΛΥΣΗ ΛΟΓΑΡΙΑΣΜΟΥ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΧΡΗΣΕΩΣ

ισματα εκμεταλλεύσεως	Ποσά Κλειομένης Χρήσεως 1995		Ποσά Προηγούμενης Χρήσεως 1994	
πωλήσεων		511.139.587		480.687.928
στος πωλήσεων		410.218.715		370.775.864
ελέσματα (κέρδη) εκμεταλλεύσεως		100.920.872		109.912.064
α έσοδα εκμεταλλεύσεως		----		1.600.000
		100.920.872		111.512.064
έξοδα διοικητικής λειτουργίας	18.887.216		38.554.560	
έξοδα λειτουργίας διαθέσεως	26.697.872	45.585.088	21.817.285	60.371.845
τελέσματα (ζημίες) εκμεταλλεύσεως		55.335.784		51.140.219
εωστικοί τόκοι & συναφή έξοδα		1.674.633		2.108.212
λέσματα (ζημίες) εκμεταλλεύσεως		53.661.151		49.032.007
έκτακτα αποτελέσματα				
Έκτακτα & ανόργανα έσοδα		3.520.762		(854.820)
έκτακτα αποτελέσματα (ζημίες)		57.181.913		48.177.187
ολο αποσβέσεων παγίων στοιχείων	50.202.261		---	
ν: Οι από αυτές ενσωματωμένες στο λειτουργικό κόστος	50.202.261		---	
ΟΤΕΛΕΣΜΑΤΑ (ΖΗΜΙΕΣ) ΧΡΗΣΕΩΣ ΠΡΟ ΦΟΡΩΝ		57.181.913		48.177.187

ΠΙΝΑΚΑΣ ΔΙΑΘΕΣΕΩΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

	Ποσά Κλειομένης Χρήσεως 1995	Ποσά Προηγ. Χρήσεως 1994
Διαφορά φορολ.ελέγχου	-28.243.366	-----
Ζημίες προηγούμ.χρήσεως	-243.896.019	-292.073..206
Μείον Κέρδη χρήσεως	57.181.913	48.177.187
Ζημίες χρήσεων εις νέο	-214.957.472	-243.896.019

Ο ΠΡΟΕΔΡΟΣ ΤΟΥ Δ.Σ.
ΜΙΛΤΙΑΔΗΣ ΣΗΜΙΤΗΣ ΤΟΥ ΔΗΜΗΤΡΙΟΥ
Η 678640/ΙΣΤ' ΑΘΗΝΩΝ

ΕΝΑ ΜΕΛΟΣ ΤΟΥ Δ.Σ.
ΔΗΜΟΚΡΙΤΟΣ ΣΗΜΙΤΗΣ ΤΟΥ ΣΠΥΡΙΔΩΝΟΣ
Μ 055518/ΙΣΤ' ΑΘΗΝΩΝ

Ο ΠΡΟΪΣΤΑΜΕΝΟΣ ΟΙΚΟΝΟΜΙΚΩΝ ΥΠΗΡΕΣΙΩΝ
ΑΝΔΡΕΑΣ ΜΑΡΚΟΣ ΤΟΥ ΝΙΚΟΛΑΟΥ
Ξ 120607/ΚΣΤ' ΑΘΗΝΩΝ

ΑΦΟΙ Δ.ΣΗΜΙΤΗ ΑΝΩΝΥΜΗ ΕΤΑΙΡΕΙΑ

ΔΡΑ: ΥΠΑΤΟ ΘΗΒΑΣ ΑΡΙΘ.ΜΗΤΡ.ΝΟΜΑΡΧΙΑΣ ΒΟΙΩΤΙΑΣ 9992/11/Β/86/0 - ΙΣΟΛΟΓΙΣΜΟΣ 31/12/1995 - 20Η ΕΤΑΙΡΙΚΗ ΧΡΗΣΗ (1/1-31/12/95) (ΠΟΣΑ ΣΕ ΔΡΧ.)

	Ποσά Κλειομένης Χρήσεως 1995			Ποσά Προηγούμενης Χρήσεως 1994		
ΚΑΤΑΣΤΑΣΕΩΣ	Αξία Κτήσης	Αποσβέσεις	Αναπ. Αξία	Αξία Κτήσης	Αποσβέσεις	Αναποσβ. Αξία
ώσεως & πρώτης εγκατ.	16.369.084	11.034.306	5.334.778	16.369.084	11.034.306	5.334.778
ίων κατασκ.περιόδου	1.372.505.582	154.900.777	1.217.604.805	394.553.671	154.900.777	239.652.894
	1.388.874.666	165.935.083	1.222.939.583	410.922.755	165.935.083	244.987.672
ΡΓΗΤΙΚΟ						
ακινητοποιήσεις						
τόπεδα	109.072.425	---	109.072.425	109.072.425	---	109.072.425
γνικά έργα	228.093.248	102.148.479	125.944.769	228.093.248	86.743.368	141.349.880
α-τεχν.εγκαταστάσεις	432.958.024	331.366.701	101.591.323	432.958.024	296.569.551	136.388.473
ά μέσα	21.783.517	17.587.632	4.195.885	21.783.517	17.587.632	4.195.855
οιπός εξοπλισμός	15.980.832	14.474.521	1.506.311	15.980.832	14.474.521	1.506.311
ήσεις υπό εκτέλεση						
βολές	1.882.353	903.528	978.825	1.882.353	903.528	978.825
οποιήσεων (ΓΙΙ)	809.770.399	466.480.651	343.289.538	809.770.399	416.278.600	393.491.799
ΟΥΝ ΕΝΕΡΓΗΤΙΚΟ						
κοιμα & ημιτελή						
α & υπολείμματα			-----			167.700
ητικές ύλες-αναλώσιμα						
κά & είδη συσκευασίας			21.694.537			29.700.781
			21.694.537			29.868.481
			73.626.846			55.073.915
ιοπρακτέα						
ίου			240.000			1.000.000
ις για εγγύηση			-----			----
ς καθυστέρηση						80.000
ίαφοροι			281.004.418			282.789.000
διαχειρίσεως						
ιν & πιστώσεων			-5.263.478			-5.453.403
			349.607.786			333.489.512
			6.208.072			7.858.423
ψεως & προθεσμίας			10.952.018			788.969
			17.160.090			8.647.392
ιούντος ενεργητικού(ΔΙ+ΔΙΙ+ΔΙV)			388.462.413			372.005.385
ΕΝΕΡΓΗΤΙΚΟΥ (Β+Γ+Δ)			1.954.691.534			1.010.484.856

ΠΑΘΗΤΙΚΟ	Ποσά Κλειομένης Χρήσεως 1995	Ποσά Προηγ/νης Χρήσεως 1994
Α.ΙΔΙΑ ΚΕΦΑΛΑΙΑ		
I.Μετοχικό Κεφάλαιο		
(195.000 μετοχές των 1.000δρχ.)		
1.Καταβλημένο	195.000.000	195.000.000
III.Διαφορές αναπροσαρμογής		
Επιχορηγήσεις επενδύσεων		
3.Επιχορηγήσεις επενδύσεων παγίου ενεργ.	73.574.470	73.574.470
IV.Αποθεματικά Κεφάλαια		
1.Τακτικό αποθεματικό	2.459.319	2.459.319
3.Ειδικά αποθεματικά	40.000	40.000
5.Αφορολόγητα αποθεματικά ειδ.διατ.νόμων	60.004.530	60.004.530
	62.503.849	62.503.849
V.Αποτελέσματα εις νέο		
Διαφορά φορολ.ελέγχου	-28.243.366	---
Υπόλοιπα ζημιών προηγ.χρήσεων	-243.896.019	(292.073.206)
Μείον κέρδη χρήσεως	+ 51.897.465	48.177.184
Υπόλοιπα ζημιών προηγ.χρήσεων	-214.957.472	243.896.019
Σύνολο Ιδίων Κεφαλαίων	116.120.847	87.182.300
Γ.ΥΠΟΧΡΕΩΣΕΙΣ		
I.Μακροπρόθεσμες Υποχρεώσεις		
2.Δάνεια Τραπεζών (ΕΤΒΑ)	1.616.631.956	708.680.045
II.Βραχυπρόθεσμες Υποχρεώσεις		
1.Προμηθευτές	26.503.219	51.966.665
2.Γραμμάτια πληρωτέα	5.439.355	15.721.385
2α.Επιταγές πληρωτέες	----	3.312.824
3.Τράπεζες λ/βραχυπρ.υποχρ.	89.719.760	89.764.204
5.Υποχρεώσεις από φόρους - τέλη	36.614.866	16.960.332
6.Ασφαλιστικοί οργανισμοί	13.368.218	27.044.960
11.Πιστωτές διάφοροι	22.411.073	9.852.141
	194.056.491	214.622.511
Σύνολο Υποχρεώσεων (ΓΙ+ΓΙΙ)	1.810.688.447	923.302.556
Δ.ΜΕΤΑΒΑΤΙΚΟΙ ΛΟΓΑΡΙΑΣΜΟΙ ΠΑΘΗΤΙΚΟΥ		
3.Λοιποί μεταβατικοί λογαριασμοί	27.882.240	---
ΓΕΝΙΚΟ ΣΥΝΟΛΟ ΠΑΘΗΤΙΚΟΥ (Α+Γ+Δ)	1.954.691.534	1.010.484.856

ή των ακινήτων της εταιρείας έχουν εγγραφεί υποθήκες και προσημειώσεις : α)Υπέρ του Ελληνικού Δημοσίου
ρχ. 61.974.541 και β) Σε ασφαλεία Τραπεζικών δανείων ποσού δρχ. 510.062.000

ΑΝΑΛΥΣΗ ΛΟΓΑΡΙΑΣΜΟΥ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΧΡΗΣΕΩΣ

	Ποσά Κλειομένης Χρήσεως 1995		Ποσά Προηγουμενης Χρήσεως 1994	
ρεταλλεύσεως		511.139.587		460.687.928
λήσεων		410.218.715		370.775.864
τα (κέρδη) εκμεταλλεύσεως		100.920.872		109.912.064
ι εκμεταλλεύσεως		---		1.600.000
		100.920.872		111.512.064
ικητικής λειτουργίας	18.887.218		38.554.560	
τουργίας διαθέσεως	26.697.872	46.585.088	21.817.285	60.371.845
τα (ζημίες) εκμεταλλεύσεως		55.335.784		51.140.219
ά τόκοι & συναφή έξοδα		1.674.633		2.108.212
ι (ζημίες) εκμεταλλεύσεως		53.661.151		49.032.007
αποτελέσματα				
& ανόργανα έσοδα		3.520.762		(854.820)
αποτελέσματα (ζημίες)		57.181.913		48.177.187
σβέσεων παγίων στοιχείων	50.202.261		---	
ό αυτές ενσωματωμένες				
λειτουργικό κόστος	50.202.261		---	
ΜΑΤΑ (ΖΗΜΙΕΣ) ΧΡΗΣΕΩΣ ΠΡΟ ΦΟΡΩΝ		57.181.913		48.177.187

ΠΙΝΑΚΑΣ ΔΙΑΘΕΣΕΩΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

	Ποσά Κλειομένης Χρήσεως 1995	Ποσά Προηγ. Χρήσεως 1994
Διαφορά φορολ ελέγχου	-28.243.366	----
Ζημίες προηγούμενης χρήσεως	-243.896.019	-292.073..206
Μείον Κέρδη χρήσεως	57.181.913	48.177.187
Ζημίες χρήσεων εις νέο	-214.957.472	-243.896.019

Ο ΠΡΟΕΔΡΟΣ ΤΟΥ Δ.Σ.
ΜΙΛΤΙΑΔΗΣ ΣΗΜΙΤΗΣ ΤΟΥ ΔΗΜΗΤΡΙΟΥ
Η 678640/ΙΣΤ ΑΘΗΝΩΝ

ΕΝΑ ΜΕΛΟΣ ΤΟΥ Δ.Σ.
ΔΗΜΟΚΡΙΤΟΣ ΣΗΜΙΤΗΣ ΤΟΥ ΣΠΥΡΙΔΩΝΟΣ
Μ 055518/ΙΣΤ ΑΘΗΝΩΝ

Ο ΠΡΟΪΣΤΑΜΕΝΟΣ ΟΙΚΟΝΟΜΙΚΩΝ ΥΠΗΡΕΣΙΩΝ
ΑΝΔΡΕΑΣ ΜΑΡΚΟΣ ΤΟΥ ΝΙΚΟΛΑΟΥ
Ξ 120607/ΚΣΤ ΑΘΗΝΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

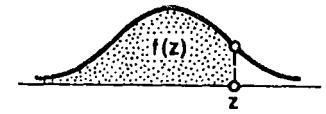
1. Ζαχαροπούλου Χ. (1990), Στατιστική,Μέθοδοι- Εφαρμογές
2. Λουκάκης Μ. (1993), Μαθηματικά Οικονομικών Επιστημών
3. Κεβόρκ Γ. (1991), Συσχέτιση- Παλινδρόμηση, Θεωρία
4. Χρήστου Γ. (1982), Εισαγωγή στην Οικονομετρία

Ξένη

1. Anderson- Sweeney- Williams (1990), Statistics For Business and Economics
2. Anscomble F.J (1973), Graphs in Statistical Analysis
3. Becker W.- Kermedy P. (1992), A Lesson in Least Squares and R Squared

ΠΙΝΑΚΕΣ

ΠΙΝΑΚΑΣ 1: Αθροιστικές πιθανότητες της τυπικής κανονικής κατανομής

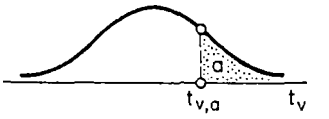


z/	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

Πηγή: Olkin I.-Gleser L.-Derman C. (1980), "Probability Models and Applications", McMillan Publishing Co., Inc.

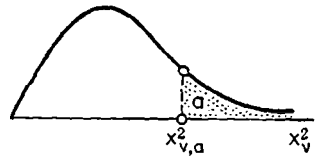
ΠΙΝΑΚΑΣ 2: Κριτικές τιμές της κατανομής t-student



	:100	.050	.025	.010	.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
∞	1.282	1.645	1.960	2.326	2.576

Πηγή: P. Newbold (1984), "Statistics for Business and Economics", Prentice-Hall Inc.

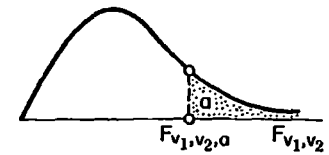
ΠΙΝΑΚΑΣ 3: Κριτικές τιμές της κατανομής χ^2



ν \ α	.995	.990	.975	.950	.900	.100	.050	.025	.010	.005
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

Πηγή: P. Newbold (1984), "Statistics for Business and Economics", Prentice-Hall Inc.

ΠΙΝΑΚΑΣ 4: Κριτικές τιμές της κατανομής F για α=0,05



v2 \ v1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07

16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Πηγή: Maddala GS. (1992), "Introduction to econometrics, second Edition", McMillan

ΠΙΝΑΚΑΣ 5: Κριτικές τιμές της κατανομής F για α=0,01

v2 \ v1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4.052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87

16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Πηγή: Maddala G.S. (1992), "Introduction to Econometrics, Second Edition", McMillan.

ΠΙΝΑΚΑΣ 6: Κριτικές τιμές του συντελεστή συσχέτισης κατά τάξεις του Spearman

Μέγεθος δείγματος n	α 10%	α 5%	α 1%
4	1.000		
5	0.900	1.000	
6	0.829	0.886	1.000
7	0.714	0.786	0.929
8	0.643	0.738	0.885
9	0.600	0.683	0.845
10	0.564	0.648	0.808
11	0.520	0.620	0.773
12	0.496	0.591	0.741
13	0.475	0.566	0.716
14	0.456	0.544	0.695
15	0.438	0.524	0.676
16	0.425	0.506	0.657
17	0.411	0.490	0.639
18	0.399	0.475	0.622
19	0.388	0.462	0.606
20	0.377	0.450	0.591
21	0.368	0.438	0.576
22	0.359	0.428	0.562
23	0.351	0.418	0.549
24	0.343	0.409	0.537
25	0.336	0.400	0.526
26	0.329	0.392	0.515
27	0.323	0.384	0.505
28	0.317	0.377	0.496
29	0.311	0.370	0.487
30	0.305	0.364	0.478

Πηγή: Κουνιάς, Κολυβά-Μαχαίρα, Μπαγιάτης, Μπόρα-Σέντα (1985), "Εισαγωγή στην Στατιστική", Εκδ. Υπηρεσίας Δημοσιευμάτων ΑΠΘ, Θεσσαλονίκη.

ΠΙΝΑΚΑΣ 7: Ο Μετασχηματισμός $W = \frac{1}{2}\log_e\left(\frac{1+r_{XY}}{1-r_{XY}}\right)$ του συντελεστή συσχέτισης

r	z	r	z	r	z
0.00	0.0000	0.45	0.4847	0.90	1.4722
0.01	0.0100	0.46	0.4973	0.91	1.5275
0.02	0.0200	0.47	0.5101	0.92	1.5890
0.03	0.0300	0.48	0.5230	0.93	1.6584
0.04	0.0400	0.49	0.5361	0.94	1.7380
0.05	0.0500	0.50	0.5493	0.95	1.8318
0.06	0.0601	0.51	0.5627	0.96	1.9459
0.07	0.0701	0.52	0.5763	0.961	1.9588
0.08	0.0802	0.53	0.5901	0.962	1.9721
0.09	0.0902	0.54	0.6042	0.963	1.9857
0.10	0.1003	0.55	0.6184	0.964	1.9996
0.11	0.1104	0.56	0.6328	0.965	2.0139
0.12	0.1206	0.57	0.6475	0.966	2.0287
0.13	0.1307	0.58	0.6625	0.967	2.0439
0.14	0.1409	0.59	0.6777	0.968	2.0595
0.15	0.1511	0.60	0.6931	0.969	2.0756
0.16	0.1614	0.61	0.7089	0.970	2.0923
0.17	0.1717	0.62	0.7250	0.971	2.1095
0.18	0.1820	0.63	0.7414	0.972	2.1273
0.19	0.1923	0.64	0.7582	0.973	2.1457
0.20	0.2027	0.65	0.7753	0.974	2.1649
0.21	0.2132	0.66	0.7928	0.975	2.1847
0.22	0.2237	0.67	0.8107	0.976	2.2054
0.23	0.2342	0.68	0.8291	0.977	2.2269
0.24	0.2448	0.69	0.8480	0.978	2.2494
0.25	0.2554	0.70	0.8673	0.979	2.2729
0.26	0.2661	0.71	0.8872	0.980	2.2976
0.27	0.2769	0.72	0.9076	0.981	2.3235
0.28	0.2877	0.73	0.9287	0.982	2.3507
0.29	0.2986	0.74	0.9505	0.983	2.3796
0.30	0.3095	0.75	0.9730	0.984	2.4101
0.31	0.3205	0.76	0.9962	0.985	2.4427
0.32	0.3316	0.77	1.0203	0.986	2.4774
0.33	0.3428	0.78	1.0454	0.987	2.5147
0.34	0.3541	0.79	1.0714	0.988	2.5550
0.35	0.3654	0.80	1.0986	0.989	2.5987
0.36	0.3769	0.81	1.1270	0.990	2.6467
0.37	0.3884	0.82	1.1568	0.991	2.6996
0.38	0.4001	0.83	1.1881	0.992	2.7587
0.39	0.4118	0.84	1.2212	0.993	2.8257
0.40	0.4236	0.85	1.2562	0.994	2.9031
0.41	0.4356	0.86	1.2933	0.995	2.9945
0.42	0.4477	0.87	1.3331	0.996	3.1063
0.43	0.4599	0.88	1.3758	0.997	3.2504
0.44	0.4722	0.89	1.4219	0.998	3.4534

Πηγή: Owen B. (1962), "Handbook of Statistical tables", Addison-Wesley.

ΠΙΝΑΚΑΣ 8: Οριο του στατιστικού Durbin-Watson

	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
36	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.60	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Πηγή: Sen-Srivastava (1990), "Regression Analysis", Springer-Verlag.