

ΚΕΦΑΛΑΙΟ Ι

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

ΕΝΟΤΗΤΕΣ

1. ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ
2. ΧΡΗΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΠΑΚΕΤΩΝ ΣΤΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ
3. ΣΕΙΡΙΑΚΟΣ ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ
4. ΠΑΛΙΝΔΡΟΜΗΣΗ ΤΑΞΗΣ ΜΕΓΕΘΟΥΣ
5. ΤΟ ΓΕΝΙΚΟ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ
6. ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΣΤΗ ΓΡΑΜΜΙΚΗ ΣΥΣΧΕΤΙΣΗ

1. ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Σε πολλές στατιστικές εφαρμογές συναντάμε το πρόβλημα της μελέτης της σχέσης δύο ή περισσότερων τυχαίων μεταβλητών. Παραδείγματα τέτοιας σχέσης έχουμε στη μελέτη του ύψους και του βάρους μιας ομάδας ανθρώπων, του εισοδήματος και κατανάλωσης εργαζομένων σε μια εταιρεία κλπ. Το πρόβλημα που θέλουμε να λύσουμε είναι αφ' ενός να αποφασίσουμε αν υπάρχει μια τέτοια σχέση και στη συνέχεια να προσδιορίσουμε τη σχέση αυτή με βάση ορισμένες παρατηρήσεις. Ένας από τους κύριους λόγους που η μελέτη αυτή είναι σημαντική, κυρίως σε εφαρμογές που έχουν σχέση με επιχειρήσεις και με την οικονομία, είναι ότι οι σχέσεις αυτές χρησιμοποιούνται συχνά για προβλέψεις. Είναι προφανές ότι συχνά, είτε ιδιωτικές εταιρείες είτε κρατικές μονάδες χρειάζεται να προβλέψουν μεταβλητές όπως η ζήτηση, τα επιτόκια, ο πληθωρισμός, οι τιμές πρώτων υλών, το εργατικό κόστος κλπ.

Ανεξάρτητα, όμως, από τους λόγους για τους οποίους η μελέτη της σχέσης δύο ή περισσότερων μεταβλητών είναι χρήσιμη, το πρώτο βήμα για να πραγματοποιηθεί η μελέτη αυτή είναι η κατασκευή μιας μαθηματικής εξίσωσης (μοντέλου) που περιγράφει τη φύση της σχέσης που υφίσταται μεταξύ των υπό μελέτη μεταβλητών.

Η διαδικασία δημιουργίας μιας μαθηματικής εξίσωσης για την περιγραφή ενός φαινομένου μπορεί να είναι ιδιαίτερα πολύπλοκη. Αυτό οφείλεται στο γεγονός ότι για την κατασκευή του μοντέλου απαιτείται κάποια γνώση της φύσης της σχέσης μεταξύ των μεταβλητών. Για παράδειγμα, ένας επενδυτής στην αγορά χρυσού, προκειμένου να προβεί σε μια μεγάλη αγορά, θα ενδιαφερόταν να προβλέψει την τιμή του χρυσού σε δύο χρόνια από τώρα χρησιμοποιώντας κάποια τεχνική. Παράγοντες που μπορεί να επηρεάσουν την τιμή αυτή είναι τα επιτόκια, ο πληθωρισμός, η τιμή του πετρελαίου, η ζήτηση για χρυσά κοσμήματα, η ζήτηση για βιομηχανικό και εμπορικό χρυσό και ο δείκτης του Χρηματιστηρίου. Έστω, πιο συγκεκριμένα, ότι ο επενδυτής αυτός ενδιαφέρεται να μελετήσει αρχικά τον τρόπο που τα επιτόκια επηρεάζουν την τιμή του χρυσού. Αν θεωρήσει ότι υπάρχει μια γραμμική σχέση, αυτό θα

συνεπάγεται ίσως ότι όταν τα επιτόκια αυξάνουν (ή πέφτουν) η τιμή του χρυσού θα αυξάνει (θα μειώνεται). Μια εξίσωση τετραγωνικής μορφής θα οδηγεί, ενδεχομένως, στο συμπέρασμα ότι η τιμή του χρυσού θα αυξάνει για κάποιο συγκεκριμένο φάσμα τιμών του πληθωρισμού αλλά θα μειώνεται για κάποιο διαφορετικό φάσμα. Ενδεχομένως, οι συγκεκριμένοι συνδυασμοί τιμών του πληθωρισμού και άλλων μεταβλητών επιδρούν στην τιμή του χρυσού με ένα τρόπο, ενώ άλλοι συνδυασμοί επηρεάζουν την τιμή αυτή με ένα άλλο τρόπο. Ο αριθμός των διαφορετικών μαθηματικών μοντέλων που θα μπορούσαν να χρησιμοποιηθούν είναι σχεδόν άπειρος.

Στα περισσότερα από τα πρακτικά προβλήματα που θα μας απασχολήσουν μας ενδιαφέρει να μελετήσουμε πώς μεταβάλλεται μια τυχαία μεταβλητή Y σε σχέση με μια - ή περισσότερες - μεταβλητές $X_1, X_2, \dots, X_n$ των οποίων ο ερευνητής μπορεί να επιλέξει τις τιμές.

Αν το μοντέλο που εξετάζουμε είναι τέτοιας μορφής που η τυχαία μεταβλητή Y είναι γραμμική συνάρτηση των *παραμέτρων* του μοντέλου τότε μιλάμε για ένα *γραμμικό μοντέλο* (*linear model*).

Η απλούστερη μορφή τέτοιας σχέσης είναι η

$$Y = \alpha + \beta X$$

όπου α και β είναι σταθερές.

Σημείωση: Επισημαίνουμε ότι ο όρος "γραμμικό" για το χαρακτηρισμό του μοντέλου αναφέρεται στις παραμέτρους και όχι στις μεταβλητές.

Η μελέτη γραμμικών μοντέλων είναι χρήσιμη και για την πρόβλεψη της τιμής της τυχαίας μεταβλητής Y για μια δοθείσα τιμή x του X , όπως ήδη αναφέραμε, όπως επίσης για τον προσδιορισμό του βαθμού εξάρτησης μεταξύ των τυχαίων μεταβλητών X και Y .

Μοντέλα τέτοιας μορφής συναντώνται συχνά σε διάφορες επιστήμες. Για παράδειγμα, σχέσεις τέτοιας μορφής που συναντάμε στη Φυσική είναι, μεταξύ άλλων και οι εξής:

$$E = m c^2,$$

όπου E = ενέργεια, m = μάζα και c = ταχύτητα του φωτός.

$$F = m a,$$

όπου F = δύναμη, m = μάζα και a = επιτάχυνση.

$$S = \alpha t^2 / 2,$$

όπου S = απόσταση, t = χρόνος και α = επιτάχυνση βαρύτητας.

Σε μαθήματα οικονομικού περιεχομένου συναντώνται επίσης σχέσεις όπως οι εξής:

$$\text{Κέρδος} = \text{Εισπράξεις} - \text{Κόστος}$$

$$\text{Συνολικό κόστος} = \text{Σταθερό κόστος} + (\text{Μεταβλητό κόστος} \times \text{αριθμό μονάδων που παρήχθησαν}).$$

Όλα τα παραπάνω παραδείγματα είναι παραδείγματα ντετερμινιστικών μοντέλων (*deterministic models*). Η ονομασία οφείλεται στο γεγονός ότι εξισώσεις τέτοιας μορφής επιτρέπουν τον καθορισμό της τιμής της εξαρτημένης μεταβλητής (της μεταβλητής στο αριστερό μέρος της εξίσωσης) από την τιμή της ανεξάρτητης μεταβλητής (της μεταβλητής στο δεξί μέρος της εξίσωσης), με εξαίρεση μικρά λάθη μετρήσεων. Η μελέτη σχέσεων τέτοιας μορφής είναι γνωστή από τα Μαθηματικά.

Στην πράξη όμως (στην καθημερινή ζωή) τα πράγματα δεν είναι πάντα ιδεώδη, είναι δηλαδή σχεδόν απίθανο να έχουμε δύο μεγέθη που να έχουν μια τέλεια γραμμική σχέση. Είναι δύσκολο, για παράδειγμα, να πιστέψουμε ότι μπορούμε να καθορίσουμε την τιμή πώλησης ενός διαμερίσματος με βάση αποκλειστικά το μέγεθός του. Είναι βέβαια αναμφίβολο ότι το μέγεθος του διαμερίσματος επηρεάζει την τιμή του, υπάρχουν όμως και άλλες μεταβλητές (μερικές από τις οποίες μπορεί να μην είναι καν μετρήσιμες) που επίσης επηρεάζουν την τιμή του. Στις περισσότερες, επομένως, περιπτώσεις που αναφέρονται σε πρακτικά προβλήματα πρέπει να χρησιμοποιηθούν μοντέλα που να περικλείουν το στοιχείο της τυχαιότητας (*randomness*), στοιχείο που είναι μέρος της καθημερινής ζωής. Τέτοια μοντέλα ονομάζονται *μοντέλα πιθανότητας* (*probabilistic models*).

Προκειμένου να κατασκευάσουμε ένα μοντέλο πιθανότητας ξεκινάμε με ένα ντετερμινιστικό μοντέλο που προσεγγίζει ικανοποιητικά τη σχέση την οποία θέλουμε να μελετήσουμε. Στην συνέχεια, προσθέτουμε ένα τυχαίο όρο που μετρά τις αποκλίσεις (τα λάθη) του ντετερμινιστικού όρου. Ο τυχαίος αυτός όρος αναφέρεται

σε όλες τις μεταβλητές, μετρήσιμες και μη μετρήσιμες, που δεν είναι μέρος του μοντέλου.

Γραμμικά μοντέλα τέτοιας μορφής είναι, για παράδειγμα, αυτά που ακολουθούν:

- i) $Y = \alpha + \beta X + \varepsilon$ (απλή γραμμική παλινδρόμηση, simple linear regression).
- ii) $Y = \alpha + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$ (πολλαπλή παλινδρόμηση, multiple regression).
- iii) $Y = \alpha + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X + \varepsilon$ (πολυωνυμική παλινδρόμηση, multinomial regression).

(Επαναλαμβάνουμε ότι ο όρος "γραμμικό" για το χαρακτηρισμό του μοντέλου αναφέρεται στις παραμέτρους και όχι στις μεταβλητές).

Στη συνέχεια του κεφαλαίου αυτού θα επικεντρώσουμε την προσοχή μας σε μοντέλα που έχουν μια μόνο ανεξάρτητη μεταβλητή. Τα μοντέλα αυτά ονομάζονται *μοντέλα απλής γραμμικής παλινδρόμησης (simple linear regression models)*.

Ας δούμε τώρα ένα παράδειγμα. Ο πίνακας που ακολουθεί, δίνει στοιχεία για την ποσότητα σε νερό που χρησιμοποιήθηκε για πότισμα σε ένα χωράφι (σε εκατοστά) και την παραγωγή τριφυλλιού (σε τόνους ανά στρέμμα) στο χωράφι αυτό (που χρησιμοποιήθηκε πειραματικά).

Νερό (x)	12	18	24	30	36	42	48
Σοδειά (y)	5.27	5.68	6.25	7.21	8.05	8.71	8.42

Ας υποθέσουμε ότι υπάρχει μια γραμμική σχέση

$$Y = \alpha + \beta X \quad (1)$$

ανάμεσα στη σοδειά και το νερό.

Είναι φανερό ότι αν χρησιμοποιήσουμε πάλι 30cm νερού είναι μάλλον απίθανο να πάρουμε πάλι σοδειά 7.21 τόνους/στρέμμα. Για το λόγο αυτό αντί για το ντετερμινιστικό μοντέλο $Y = \alpha + \beta X$ χρησιμοποιούμε ένα στοχαστικό μοντέλο

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (2)$$

όπου το ε_i αντιπροσωπεύει ένα τυχαίο λάθος. (Φυσικά, αν είχαμε μια τέλεια γραμμική σχέση για όλα τα $i = 1, 2, \dots$ τα ε_i θα ήταν μηδέν). Η σχέση (2) συνεπάγεται ότι η τιμή y του Y που παρατηρήσαμε (για ένα συγκεκριμένο $X=x$) είναι μια τυχαία τιμή από τον πληθυσμό των τιμών της τυχαίας μεταβλητής Y που αντιστοιχούν στο συγκεκριμένο x . (Στο παράδειγμα η τιμή 7.21 είναι μια τιμή που θα μπορούσε να έχει η σοδειά Y όταν η ποσότητα νερού που χρησιμοποιήθηκε ήταν 30cm).

Για τα ε_i υποθέτουμε ότι είναι τυχαία και όχι συστηματικά. Επίσης υποθέτουμε ότι

$$E(\varepsilon_i) = 0$$

Τέλος, λόγω των αποκλίσεων των παρατηρουμένων τιμών της τυχαίας μεταβλητής Y και του γεγονότος ότι η τυχαία μεταβλητή Y ακολουθεί μια κατανομή, υποθέτουμε ότι η μέση τιμή της τυχαίας μεταβλητής Y για δοθέν x συνδέεται γραμμικά με το x . Δηλαδή υποθέτουμε ότι

$$E(Y_i | x_i) = \alpha + \beta x_i \quad (3)$$

Η ευθεία (3) ονομάζεται *ευθεία παλινδρόμησης (regression line)* και οι συντελεστές α και β που την προσδιορίζουν ονομάζονται *μοντέλα απλής γραμμικής παλινδρόμησης (regression coefficients)*.

Αν χρησιμοποιήσουμε το συμβολισμό

$$\mu_{Y_i | x_i} = E(Y_i | x_i)$$

Τότε η (3) γράφεται

$$\mu_{Y_i | x_i} = \alpha + \beta x_i$$

Οι υποθέσεις που κάνουμε στο στοχαστικό μοντέλο όπου οι τιμές της μεταβλητής X καθορίζονται από τον ερευνητή, ισοδυναμούν με την υπόθεση ότι, για ένα συγκεκριμένο x , η μέση τιμή του Y (και όχι μια συγκεκριμένη τιμή του Y) είναι σε γραμμική σχέση με το x .

Η μεταβλητή X (οι τιμές της οποίας καθορίζονται από τον ερευνητή) λέγεται *ανεξάρτητη μεταβλητή (independent variable)* ενώ η τυχαία μεταβλητή Y λέγεται *εξαρτημένη (dependent variable)*.

Το πρόβλημά μας συνίσταται στον προσδιορισμό της ευθείας παλινδρόμησης (regression line) (της ευθείας (3)). Επειδή η τιμή y της Y που θα παρατηρήσουμε θα αποκλίνει, εν γένει, από την $\mu_{Y|x}$ δεν

μπορούμε να προσδοκούμε ότι είναι δυνατόν να προσδιορίσουμε την ακριβή ευθεία παλινδρόμησης. Θα προσπαθήσουμε λοιπόν να βρούμε μια *εκτίμηση της ευθείας παλινδρόμησης* έστω $\hat{y} = a + bx$ η οποία θα βασίζεται σε ένα τυχαίο δείγμα παρατηρήσεων (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) . Δοθέντος δηλαδή του τυχαίου δείγματος (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) θα προσπαθήσουμε να προσδιορίσουμε τους συντελεστές a και b , (οι οποίοι θα είναι σημειακές εκτιμήσεις των α και β , δηλαδή τιμές των εκτιμητριών α και β των $\hat{\alpha}$ και $\hat{\beta}$ αντίστοιχα), ώστε η γραμμή

$$\hat{y} = a + bx$$

(η οποία είναι μία σημειακή εκτίμηση της ευθείας (3), δηλαδή μια "τιμή" της "εκτιμήτριας" $\hat{\mu}_{Y_i|X_i} = \hat{\alpha} + \hat{\beta}x_i$) να είναι όσο το δυνατόν πλησιέστερη προς την πραγματική ευθεία παλινδρόμησης της οποίας η συναρτησιακή μορφή είναι

$$\mu_{Y_i|X_i} = \alpha + \beta x_i$$

(Με $\hat{\alpha}$, $\hat{\beta}$ και $\hat{\mu}_{Y_i|X_i}$ συμβολίζουμε τις εκτιμήτριες των α , β και $\mu_{Y_i|X_i}$ αντίστοιχα).

Η ευθεία (3) λέγεται *γραμμική παλινδρόμηση (linear regression)*.

1.1 Εκτίμηση της Ευθείας Παλινδρόμησης με τη Μέθοδο των Ελαχίστων Τετραγώνων

Μια από τις μεθόδους εκτίμησης των α και β είναι η μέθοδος των ελαχίστων τετραγώνων (Ε.Τ.) ή μέθοδος του Gauss (*least-squares method*).

Δοθέντων των τιμών (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) ενός τυχαίου δείγματος (x_1, Y_1) , (x_2, Y_2) , ..., (x_n, Y_n) , από τη (2) έχουμε ότι

$$\varepsilon_i = y_i - (\alpha + \beta x_i), i = 1, 2, \dots, n$$

όπου e_i είναι η τιμή του "σφάλματος" ε για το ζεύγος τιμών (x_i, y_i) , $i = 1, 2, \dots, n$.

Βεβαίως οι τιμές e_i δεν μπορούν να παρατηρηθούν (non-observable) αφού οι συντελεστές a και b δεν μπορούν να παρατηρηθούν. Επομένως, αν με a , b συμβολίσουμε τις εκτιμήσεις των αγνώστων παραμέτρων a και b αντίστοιχα που προκύπτουν από το συγκεκριμένο δείγμα με κάποιο τρόπο, οι παρατηρούμενες τιμές e_i των e_i θα δίνονται από τον τύπο

$$e_i = y_i - (a + bx_i), i = 1, 2, \dots, n$$

Το ερώτημα που αναφύεται είναι πως θα προσδιορισθούν οι τιμές a και b ώστε να θεωρούνται ικανοποιητικές προσεγγίσεις των a και b . Ένα κριτήριο ικανοποιητικής προσέγγισης (εκτίμησης) είναι δυνατόν να στηριχθεί στην ελαχιστοποίηση μιας κατάλληλα επιλεγμένης συνάρτησης της απόκλισης αυτού που παρατηρείται (y_i) από αυτό που εκτιμάται ότι πρέπει να παρατηρηθεί ($a + bx_i$).

Όταν ως τέτοια συνάρτηση θεωρηθεί η μέση τετραγωνική απόκλιση του y_i από το $a + bx_i$, η επιλογή των τιμών a και b πρέπει να εξασφαλίζει ελάχιστη τιμή της παράστασης $\sum_{i=1}^n e_i^2 / n$, δηλαδή της παράστασης

$$\frac{1}{n} \sum_{i=1}^n [y_i - (a + bx_i)]^2.$$

Στη διεθνή βιβλιογραφία χρησιμοποιείται συχνά ο συμβολισμός SSE (sum of squares of errors) για το $\sum e_i^2$.

Η απαίτηση αυτή ικανοποιείται τότε και μόνο τότε εάν ικανοποιείται το σύστημα εξισώσεων

$$\frac{\partial P}{\partial a} = 0$$

(κανονικές εξισώσεις)

$$\frac{\partial P}{\partial b} = 0$$

όπου

$$P = \sum_{i=1}^n [y_i - (a + bx_i)]^2.$$

Εύκολα μπορεί να επαληθευτεί ότι το ακρότατο που προκύπτει από την επίλυση του συστήματος των κανονικών εξισώσεων ως προς a και b είναι ελάχιστο, δηλαδή ότι οι τιμές a και b ελαχιστοποιούν την παράσταση P . Επειδή η παραπάνω μέθοδος στηρίζεται στην ελαχιστοποίηση της παράστασης P , ονομάζεται *μέθοδος ελαχίστων τετραγώνων (method of least -squares)*.

Οι κανονικές εξισώσεις γράφονται

$$\bar{y} = a + b\bar{x}$$

$$\sum_i^n x_i y_i = a \sum_i^n x_i + b \sum_i^n x_i^2$$

και επομένως

$$a = \bar{y} - b\bar{x} \quad \text{όπου}$$

(4)

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

Τα a και b ονομάζονται *εκτιμήσεις ελαχίστων τετραγώνων των συντελεστών γραμμικής παλινδρόμησης (least-squares estimates of the coefficients of linear regression)*.

Στην πράξη πολλές φορές για τον υπολογισμό του b χρησιμοποιούμε την παρακάτω ισοδύναμη έκφραση

$$b = \frac{n \sum_i y_i x_i - \sum_i y_i \sum_i x_i}{n \sum_i x_i^2 - (\sum_i x_i)^2} \quad (5)$$

Για ευκολότερη απομνημόνευση των τύπων χρησιμοποιούμε το συμβολισμό

$$S_{xx} = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \quad (6)$$

$$S_{yy} = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - \frac{\left(\sum y_i\right)^2}{n} \quad (7)$$

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x})y_i \quad (8)$$

Ετσι, η σχέση (4) γράφεται με τη μορφή

$$b = \frac{S_{xy}}{S_{xx}}$$

οπότε η εκτίμηση της ευθείας παλινδρόμησης του Y πάνω στο x (η εκτιμήτρια ελαχίστων τετραγώνων της πραγματικής ευθείας παλινδρόμησης) γίνεται

$$\hat{y} = a + bx \quad (9)$$

όπου

$$b = \frac{S_{xy}}{S_{xx}} \quad \text{και} \quad a = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}.$$

Επειδή η τιμή y_i είναι μία πραγματοποίηση (*realization*) της i μεταβλητής της ακολουθίας $Y_1, Y_2, \dots, Y_n$ που παριστάνει το τυχαίο δείγμα επί της εξαρτημένης μεταβλητής Y, καθίσταται φανερό ότι οι τιμές a και b αποτελούν πραγματοποιήσεις (*realizations*) των τυχαίων

$$\text{μεταβλητών } \hat{a} = \bar{Y} - \frac{S_{xY}}{S_{xx}} \bar{x} \quad \text{και} \quad \hat{b} = \frac{S_{xY}}{S_{xx}}$$

αντίστοιχα, όπου προφανώς

$$S_{xY} = \sum_i (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_i (x_i - \bar{x})Y_i.$$

Οι τιμές a και b δηλαδή είναι οι τιμές των εκτιμητριών $\hat{\alpha}$ και $\hat{\beta}$ των παραμέτρων α και β , που αντιστοιχούν στο συγκεκριμένο δείγμα $\{(x_i, y_i), i=1,2,\dots,n\}$. (Αν πάρουμε ένα άλλο δείγμα $(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_n, y'_n)$ θα βρούμε μια άλλη τιμή a' της $\hat{\alpha}$ και μία άλλη τιμή b' της $\hat{\beta}$). Με την ίδια λογική η τιμή $\hat{Y}$ είναι μια τιμή της εκτιμήτριας $\hat{Y} = \hat{\mu}_{Y|X}$.

Εφαρμογή: Στο παράδειγμά μας βρίσκουμε ότι $S_{xx} = 1008$, $S_{xy} = 103.68$, $\bar{x} = 30$, $\bar{y} = 7.08$, $S_{yy} = 11.278$. Συνεπώς $b = 0.10$ και $a = 4.0$ και επομένως η ευθεία ελαχίστων τετραγώνων είναι η

$$\hat{Y} = 4.0 + 0.10x \quad (10)$$

Η ευθεία (10) είναι μια εκτίμηση (η εκτιμήτρια ελαχίστων τετραγώνων) της πραγματικής ευθείας παλινδρόμησης $\mu_{Y_i|X_i} = \alpha + \beta x_i$ (είναι δηλαδή μια τιμή της $\hat{\mu}_{Y_i|X_i} = \hat{\alpha} + \hat{\beta} x_i$) και αντιστοιχεί στο συγκεκριμένο δείγμα. Ένα άλλο δείγμα θα έδινε μια άλλη ευθεία.

Η ευθεία παλινδρόμησης χρησιμοποιείται πολλές φορές για πρόβλεψη. Στο παράδειγμά μας, αν η ποσότητα του νερού ήταν 20cm θα περιμέναμε σοδειά $y = 6$ τόνους/στρέμμα. (Η τιμή αυτή είναι επίσης μια εκτίμηση της μέσης σοδειάς για τις περιπτώσεις που χρησιμοποιήθηκε νερό 20cm).

Στο παράδειγμα που εξετάστηκε οι τιμές της ανεξάρτητης μεταβλητής X είχαν καθορισθεί εκ των προτέρων από τον ερευνητή. Υπάρχουν περιπτώσεις όπου όχι μόνο το Y αλλά και το X είναι τυχαίες μεταβλητές. Η μέθοδος εκτίμησης των α και β είναι η ίδια.

1.2 Συμπερασματολογία στη Γραμμική Παλινδρόμηση

Οι εκτιμήσεις a και b που βρήκαμε για το a και β είναι προφανώς σημειακές εκτιμήσεις. Αν θέλουμε μεγαλύτερη σιγουριά στην εκτίμησή μας θα πρέπει να κατασκευάσουμε διαστήματα εμπιστοσύνης. Επίσης, συχνά, χρειάζεται να κάνουμε ελέγχους υποθέσεων για τις παραμέτρους του μοντέλου. Τέλος, συχνά, χρησιμοποιούμε το μοντέλο για προβλέψεις. Για να γίνουν αυτά χρειαζόμαστε την κατανομή των τυχαίων μεταβλητών $\hat{a}$ και $\hat{\beta}$.

Κάνουμε τις ακόλουθες υποθέσεις:

- i) Τα ε_i είναι $N(0, \sigma^2)$ για όλα τα $i = 1, 2, \dots$ (δηλαδή όλα τα ε_i ακολουθούν την κανονική κατανομή με μέση τιμή μηδέν και την ίδια διασπορά).
- ii) Τα ε_i είναι ανεξάρτητα μεταξύ τους.

Παρατήρηση: Για τις περισσότερες εφαρμογές είναι αρκετό να υποθέσουμε ότι τα ε_i είναι ασυσχέτιστα μεταξύ τους. Δηλαδή ότι $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, για κάθε $i \neq j$, $i, j = 1, 2, \dots$.

Σημείωση: Οι υποθέσεις αυτές είναι οι απλούστερες που γίνονται συνήθως. Διαφορετικές καταστάσεις χρειάζονται διαφορετικές υποθέσεις που, φυσικά, θα οδηγήσουν σε διαφορετικά συμπεράσματα. Κάνοντας χρήση της σχέσης (2) βλέπουμε ότι οι παραπάνω υποθέσεις είναι ισοδύναμες με τις υποθέσεις:

- iii) Τα Y_i είναι $N(\alpha + \beta x_i, \sigma^2)$, $i = 1, 2, \dots$

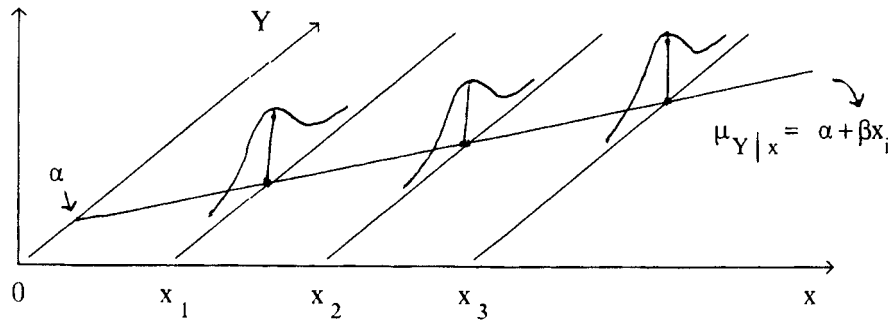
- iv) Τα Y_i είναι ανεξάρτητα μεταξύ τους.

(Αντίστοιχα, όπως προηγουμένως, τα Y_i είναι ασυσχέτιστα).

Επίσης υποθέτουμε ότι τα x_i είναι προκαθορισμένα.

Σημείωση: Επειδή υποθέτουμε ότι τα ε_i (αντίστοιχα τα Y_i), $i = 1, 2, \dots$ έχουν την ίδια διακύμανση, λέμε ότι είναι *ομοσκεδαστικά* (*homoscedastic*). Στην αντίθετη περίπτωση τα ε_i (αντίστοιχα τα Y_i) λέγονται *ετεροσκεδαστικά* (*heteroscedastic*).

Παρατήρηση: Η υπόθεση της κανονικότητας απαιτεί ότι οι τιμές των Y ακολουθούν την κανονική κατανομή σε κάθε τιμή του x . (Βλέπε και το σχήμα που ακολουθεί).



Όπως και στην συμπερασματολογία που στηρίζεται στη στατιστική συνάρτηση t (βλέπε π.χ. Ι. Πανάρετου (1993) *Εκτιμητική-Έλεγχοι Υποθέσεων*) η ανάλυση παλινδρόμησης είναι ανθεκτική (*robust*) σε αποκλίσεις από την υπόθεση της κανονικότητας. Δηλαδή, αν η κατανομή των τιμών της Y_i γύρω από κάθε τιμή του x δεν διαφέρει πάρα πολύ από την κανονική κατανομή η συμπερασματολογία που ακολουθείται για την γραμμική παλινδρόμηση και τους συντελεστές της γραμμικής παλινδρόμησης δεν θα επηρεασθεί σημαντικά.

Η υπόθεση της ομοσκεδαστικότητας απαιτεί ότι η μεταβλητότητα γύρω από τη γραμμή παλινδρόμησης είναι σταθερή για όλες τις τιμές του x . Αυτό σημαίνει ότι το Y αποκλίνει με τον ίδιο τρόπο όταν το x έχει μια μικρή τιμή όπως όταν το x έχει μια υψηλή τιμή. (Βλέπε και το προηγούμενο σχήμα). Η υπόθεση της ομοσκεδαστικότητας είναι σημαντική για την χρησιμοποίηση της μεθόδου των ελαχίστων τετραγώνων που ακολουθείται για τον καθορισμό των συντελεστών της γραμμικής παλινδρόμησης. Αν υπάρχει μεγάλη απόκλιση από την υπόθεση αυτή θα πρέπει ή να γίνει μετασχηματισμός των δεδομένων ή να χρησιμοποιηθεί η μέθοδος των σταθμισμένων ελαχίστων τετραγώνων (*weighted least-squares*) για τη συμπερασματολογία.

Η τρίτη υπόθεση της ανεξαρτησίας των λαθών απαιτεί ότι τα κατάλοιπα θα πρέπει να είναι ανεξάρτητα για κάθε τιμή του X . Η

υπόθεση αυτή συχνά αναφέρεται σε δεδομένα που έχουν συλλεγεί κατά τη διάρκεια μιας χρονικής περιόδου. Για παράδειγμα, σε οικονομικά δεδομένα, οι τιμές για μια συγκεκριμένη χρονική περίοδο είναι συχνά συσχετισμένες με τις τιμές της προηγούμενης χρονικής περιόδου. Τα μοντέλα τέτοιας μορφής ανήκουν στην γενική κατηγορία των χρονολογικών σειρών που θα δούμε αργότερα.

1.2.1 Διαστήματα Εμπιστοσύνης για το β

Προκειμένου να μελετήσουμε την κατανομή της εκτιμήτριας $\hat{\beta}$ της παραμέτρου β ξαναγράφουμε τη σχέση (4) με συμβολισμό τυχαίων μεταβλητών, χρησιμοποιώντας το $\hat{\beta}$ και το Y αντί του b και του y , αντίστοιχα.

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} = \sum_i c_i Y_i$$

όπου

$$c_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2} = \frac{x_i - \bar{x}}{S_{xx}} \quad (12)$$

$$(\text{Προφανώς } \sum_i c_i = \frac{1}{S_{xx}} \sum_i (x_i - \bar{x}) = 0).$$

Τα Y_i όμως ακολουθούν κανονικές κατανομές. Επομένως το $\hat{\beta}$, σαν γραμμικός συνδυασμός τυχαίων μεταβλητών που ακολουθούν κανονική κατανομή, κατανέμεται και αυτό σύμφωνα με την κανονική κατανομή, δηλαδή,

$$\hat{\beta} \sim N(E(\sum_i c_i Y_i), V(\sum_i c_i Y_i)).$$

Έχουμε όμως

$$\begin{aligned} E(\hat{\beta}) &= E(\sum_i c_i Y_i) = \sum_i c_i E(Y_i) \\ &= \sum_i c_i (\alpha + \beta x_i) = \alpha \sum_i c_i + \beta \sum_i c_i x_i. \end{aligned}$$

Από τον ορισμό όμως του c_i έχουμε

$$\sum c_i = \frac{\sum (x_i - \bar{x})}{S_{xx}} = \frac{n\bar{x} - n\bar{x}}{S_{xx}} = 0 \quad (13)$$

Επίσης,

$$\begin{aligned} \sum c_i x_i &= \frac{\sum (x_i - \bar{x}) x_i}{S_{xx}} = \frac{\sum (x_i - \bar{x})(x_i - \bar{x} + \bar{x})}{S_{xx}} \\ &= \frac{\sum (x_i - \bar{x})^2}{S_{xx}} + \frac{\sum (x_i - \bar{x}) \bar{x}}{S_{xx}} = S_{xx} / S_{xx} = 1 \end{aligned} \quad (14)$$

Επομένως,

$$E(\hat{\beta}) = \beta. \quad (15)$$

Όσον αφορά τη διασπορά του $\hat{\beta}$ έχουμε,

$$\begin{aligned} V(\hat{\beta}) &= V\left(\sum c_i Y_i\right) = \sum c_i^2 V(Y_i) \\ &(\text{διότι τα } Y_i \text{ είναι ανεξάρτητα μεταξύ τους}) \\ &= \sum c_i^2 \sigma^2 \end{aligned}$$

Από τη (12) έχουμε,

$$\sum c_i^2 = \frac{\sum (x_i - \bar{x})^2}{S_{xx}^2} = \frac{1}{S_{xx}}$$

Επομένως,

$$V(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}. \quad (16)$$

Δηλαδή τελικά για το $\hat{\beta}$ έχουμε

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right). \quad (17)$$

Συνεπώς,

$$\frac{\hat{\beta} - \beta}{\sigma^2 / \sqrt{S_{xx}}} \sim N(0,1) \quad (18)$$

Η τυπική απόκλιση σ των $Y_i | x_i$ $i=1,2,\dots$ όμως είναι εν γένει άγνωστη και πρέπει να εκτιμηθεί.

Σαν εκτιμήτρια του σ^2 μπορεί να θεωρηθεί η στατιστική συνάρτηση,

$$\hat{\sigma}^2 \equiv S_{Y|x}^2 \equiv S_{\varepsilon}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n} \quad (19)$$

όπου Y_i είναι η παρατηρηθείσα τιμή του Y για το συγκεκριμένο x_i ενώ $\hat{Y}_i$ είναι η εκτιμώμενη τιμή του Y_i για το δεδομένο x_i όπως την παίρνουμε από την ευθεία παλινδρόμησης. (Αποδεικνύεται ότι η εκτιμήτρια αυτή είναι εκτιμήτρια μεγίστης πιθανοφάνειας).

Σημείωση: Η στατιστική συνάρτηση $S_{Y|x}^2$, όπως δίνεται από τη σχέση (19) δεν είναι αμερόληπτη εκτιμήτρια του σ^2 . Εάν θέλουμε αμερόληπτη εκτιμήτρια του σ^2 πρέπει να θεωρήσουμε τη συνάρτηση

$$S_{Y|x}^{*2} \equiv S_{\varepsilon}^{*2} = \frac{nS_{Y|x}^2}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = \frac{nS_{\varepsilon}^2}{n-2}$$

(Βλέπε σχετική άσκηση).

Σημείωση: Από τα προηγούμενα προκύπτει ότι μια αμερόληπτη εκτιμήτρια της διασποράς της κατανομής του $\hat{\beta}$ όταν το σ^2 είναι

άγνωστο και πρέπει να εκτιμηθεί, είναι η συνάρτηση $S_{\hat{\beta}}^{*2} = \frac{S_{Y|x}^{*2}}{S_{xx}}$.

Υπολογισμός του $S_{Y|x}^2$

Από τη σχέση (19) έχουμε,

$$\begin{aligned}
S_{Y|x}^2 &= \frac{\sum (Y_i - \hat{Y}_i)^2}{n} \\
&= \frac{1}{n} \sum_i (Y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 \\
&= \frac{1}{n} \sum_i (Y_i - \bar{Y} + \bar{Y} - (\hat{\alpha} + \hat{\beta}x_i))^2 \\
&= \frac{1}{n} \sum_i ((Y_i - \bar{Y}) + (\hat{\alpha} + \hat{\beta}\bar{x} - \hat{\alpha} + \hat{\beta}x_i))^2 \\
&= \frac{1}{n} \left[\sum_i (Y_i - \bar{Y})^2 + \hat{\beta}^2 \sum_i (x_i - \bar{x})^2 - 2\hat{\beta} \sum_i (x_i - \bar{x})(Y_i - \bar{Y}) \right] \\
&= \frac{1}{n} \left\{ S_{YY} + \frac{S_{XY}^2}{S_{XX}} - 2 \frac{S_{XY}^2}{S_{XX}} \right\} \\
&= \frac{1}{n} \left\{ S_{YY} - \frac{S_{XY}^2}{S_{XX}} \right\}
\end{aligned} \tag{20}$$

Μπορεί να αποδειχθεί ότι,

$$\frac{\sum (Y_i - \hat{Y}_i)^2}{\sigma^2} \equiv \frac{nS_{Y|x}^2}{\sigma^2} \equiv \frac{(n-2)S_{Y|x}^{*2}}{\sigma^2} \sim X_{n-2}^2$$

Επομένως,

$$\frac{\hat{\beta} - \beta}{\sqrt{S_{Y|x}^{*2}/S_{xx}}} \sim t_{n-2} \tag{21}$$

Από τα προηγηθέντα προκύπτει ότι σαν ένα $100(1-a)\%$ διάστημα εμπιστοσύνης για την κλίση β της ευθείας παλινδρόμησης μπορούμε να θεωρήσουμε το διάστημα

$$\hat{\beta} \pm t_{n-2, 1-a/2} \sqrt{S_{Y|x}^{*2}/S_{xx}} \tag{22}$$

1.2.2 Διαστήματα εμπιστοσύνης για το α

Έχουμε ήδη δει ότι

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

αλλά

$$\hat{\beta} = \sum c_i Y_i$$

Επομένως,

$$\begin{aligned}\hat{\alpha} &= \bar{Y} - \left(\sum c_i Y_i\right)\bar{x} = \sum (1/n) Y_i - \sum c_i \bar{x} Y_i \\ &= \sum ((1/n) - c_i \bar{x}) Y_i = \sum d_i Y_i\end{aligned}\tag{23}$$

όπου

$$d_i = 1/n - c_i \bar{x} = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum (x_i - \bar{x})^2}$$

Η σχέση (23) δείχνει ότι το $\hat{\alpha}$ μπορεί να γραφεί σαν γραμμικός συνδυασμός ανεξαρτήτων κανονικών τυχαίων μεταβλητών. Επομένως το α ακολουθεί επίσης την κανονική κατανομή. Θα υπολογίσουμε τη μέση τιμή και τη διασπορά του $\hat{\alpha}$. Έχουμε,

$$\begin{aligned}E(\hat{\alpha}) &= E(\bar{Y} - \hat{\beta}\bar{x}) = E(\bar{Y}) - \bar{x}E(\hat{\beta}) \\ &= \alpha + \beta\bar{x} - \bar{x}\beta = \alpha\end{aligned}$$

Επίσης από την (23)

$$V(\hat{\alpha}) = V\left(\sum d_i Y_i\right) = \sum d_i^2 \sigma^2.$$

Αλλά,

$$\begin{aligned}\sum d_i^2 &= \sum \left(1/n^2 + c_i^2 \bar{x}^2 - (2/n)c_i \bar{x}\right) \\ &= 1/n + \bar{x}^2 \sum c_i^2 - (2\bar{x}/n) \sum c_i\end{aligned}$$

Από την (13) όμως $\sum c_i = 0$. Επίσης $\sum c_i^2 = 1/S_{xx}$.

Επομένως,

$$\sum d_i^2 = 1/n + \bar{x}^2 / S_{xx}\tag{24}$$

Δηλαδή τελικά,

$$V(\hat{\alpha}) = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right\} \quad (25)$$

Στην περίπτωση που το σ^2 είναι άγνωστο και πρέπει να εκτιμηθεί, μια αμερόληπτη εκτιμήτρια της διασποράς της κατανομής του $\hat{\alpha}$ είναι η συνάρτηση

$$S_{\hat{\alpha}}^{*2} = S_{Y|x}^{*2} \left\{ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right\}$$

Το τελικό συμπέρασμα για το $\hat{\alpha}$ είναι ότι

$$\hat{\alpha} \sim N \left(a, \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right\} \right) \quad (26)$$

Από την (26) έχουμε,

$$\frac{\hat{\alpha} - \alpha}{\sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim N(0,1)$$

Εάν χρησιμοποιήσουμε την εκτιμήτρια $S_{Y|x}^2$ του σ^2 και το γεγονός ότι

$$\frac{nS_{Y|x}^2}{\sigma^2} \equiv \frac{(n-2)S_{Y|x}^{*2}}{\sigma^2} \sim X_{n-2}^2$$

έχουμε,

$$\frac{\hat{\alpha} - \alpha}{\sqrt{S_{Y|x}^{*2} \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}} \sim t_{n-2}$$

και επομένως, σαν ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το α μπορούμε να πάρουμε το διάστημα

$$\hat{\alpha} \pm t_{n-2, 1-\frac{\alpha}{2}} \sqrt{S_{Y|x}^{*2} \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \quad (27)$$

Παράδειγμα: Στο πρόβλημα της παραγωγής τριφυλλιού να βρεθούν 90% διαστήματα εμπιστοσύνης για το α και το β .

Λύση: Πρέπει καταρχήν να υποθέσουμε ότι η σοδειά Y για κάθε δεδομένη ποσότητα νερού ακολουθεί κανονική κατανομή με μέση τιμή που είναι γραμμική συνάρτηση του x και κοινή διασπορά σ^2 . Επειδή το σ είναι άγνωστο το εκτιμάμε από τα δεδομένα.

$$S_{Y|x}^2 = \frac{1}{7} \left\{ 11278 - \frac{(103.68)^2}{1008} \right\} = 0.087682$$

Ένα 90% διάστημα εμπιστοσύνης για το β είναι το

$$0.10 \pm t_{5,0.95} \sqrt{\frac{7 (0.087682)}{5 (1008)}}$$

ή $0.10 \pm 2.015 (0.0110)$ δηλαδή το $(0.08, 0.12)$.

Το αντίστοιχο 90% διάστημα εμπιστοσύνης για το α είναι (από την (27))

$$4 \pm 2.015 \sqrt{\frac{7 (0.087682)}{5} \left[\frac{1}{7} + \frac{(30)^2}{1008} \right]}$$

ή 4 ± 0.718 δηλαδή το $(3.382, 4.718)$.

Σημείωση: Σε πολλές περιπτώσεις, προκειμένου να εκτιμηθούν οι συντελεστές α και β της γραμμικής παλινδρόμησης με βάση την παρατηρηθείσα τιμή $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ενός τυχαίου δείγματος, δεν χρησιμοποιείται η σχέση

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

δηλαδή η ευθεία παλινδρόμησης

$$\mu_{Y|x} = \alpha + \beta x$$

που χρησιμοποιήθηκε προηγουμένως.

Αυτό οφείλεται στο γεγονός ότι οι εκτιμήτριες α και β για τη συγκεκριμένη μορφή του μοντέλου, δεν είναι ασυσχέτιστες. (Βλέπε τη σχετική άσκηση).

Προκειμένου να αποφευχθεί η αδυναμία αυτή στην προηγούμενη σχέση προσθαφαιρούμε την ποσότητα $\beta \bar{X}$. Έτσι, έχουμε

$$\begin{aligned} Y_i &= \alpha + \beta x_i + \varepsilon_i + \beta \bar{X} - \beta \bar{X} \\ &= (\alpha + \beta \bar{X}) + \beta(x_i - \bar{X}) + \varepsilon_i \\ &= \alpha^* + \beta(x_i - \bar{X}) + \varepsilon_i. \end{aligned}$$

Δηλαδή, τελικά

$$Y_i = \alpha^* + \beta(x_i - \bar{X}) + \varepsilon_i$$

ή

$$\mu_{Y|x} = \alpha^* + \beta(x - \bar{X})$$

όπου

$$\alpha^* = \alpha + \beta \bar{X}$$

Η τελευταία αυτή μορφή γραμμικής παλινδρόμησης ονομάζεται *απλή παλινδρόμηση αποκλίσεων από τον δειγματικό μέσο*.

Από τη σχέση (4) προκύπτει ότι η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\alpha}^*$ του α^* είναι

$$\hat{\alpha}^* = \bar{Y} - \hat{\beta}\bar{X} + \hat{\beta}\bar{X}$$

δηλαδή

$$\hat{\alpha}^* = \bar{Y}$$

με τιμή για ένα συγκεκριμένο δείγμα $(y_1, x_1), \dots, (y_n, x_n)$

$$\alpha^* = \bar{y}$$

Είναι προφανές ότι για τις εκτιμήτριες $\hat{\alpha}^*$ και $\hat{\beta}$ του απλού γραμμικού μοντέλου παλινδρόμησης αποκλίσεων από το δειγματικό μέσο ισχύει ότι

$$E(\hat{\alpha}^*) = \alpha^*, \quad V(\hat{\alpha}^*) = \frac{\sigma^2}{n}$$

και

$$E(\hat{\beta}) = \beta, \quad V(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}.$$

Τέλος (βλέπε σχετική άσκηση) για την παλινδρόμηση αποκλίσεων από το δειγματικό μέσο ισχύει ότι

$$\text{cov}(\hat{\alpha}^*, \beta) = 0.$$

1.2.3 Διαστήματα εμπιστοσύνης για την $\mu_{Y|x}$

Στο αρχικό παράδειγμα για γραμμική παλινδρόμηση είχαμε βρει (σχέση 10) ότι η ευθεία ελαχίστων τετραγώνων είναι

$$\hat{Y} = 4.0 + 0.10x.$$

Είχαμε επίσης δει ότι για ποσότητα νερού $x_0 = 20\text{cm}$ θα περιμέναμε σοδειά $y_0 = 6$ τόνους/στρέμμα. Επειδή οι τιμές 4.0 και 0.10 είναι τιμές των τυχαίων μεταβλητών $\hat{\alpha}$ και $\hat{\beta}$, συνεπάγεται ότι η τιμή $y_0=6$ είναι μια τιμή της τυχαίας μεταβλητής $\hat{Y}$ όπως επίσης και της $\mu_{Y|x=20} = E(Y|x=20)$. Μια και το $\mu_{Y|x=20}$ είναι μια παράμετρος μπορούμε να κατασκευάσουμε γι αυτήν διαστήματα εμπιστοσύνης. Έχουμε, γενικά

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x.$$

Χρησιμοποιώντας τις σχέσεις (12) και (23) παίρνουμε

$$\hat{Y} = \sum d_i Y_i + \sum c_i Y_i x = \sum (d_i + c_i x) Y_i \quad (28)$$

Επομένως το $\hat{Y}$ είναι ένας γραμμικός συνδυασμός ανεξάρτητων τυχαίων μεταβλητών που ακολουθούν την κανονική κατανομή. Δηλαδή,

$$\hat{Y} \sim N E(\hat{Y}), V(\hat{Y}).$$

Προφανώς,

$$E(\hat{Y}) = E(\hat{\alpha} + \hat{\beta}x) = \alpha + \beta x = \mu_{Y|x}.$$

Επίσης,

$$\begin{aligned} V(\hat{Y}) &= V\left[\sum (d_i + c_i x_i) Y_i\right] = \sum (d_i + c_i x_i)^2 \sigma^2 \\ &= \sigma^2 \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right\} \end{aligned}$$

Δηλαδή, τελικά

$$\hat{Y} \sim N \left(\mu_{Y|x}, \sigma^2 \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right\} \right) \quad (29)$$

Χρησιμοποιώντας την εκτιμήτρια $S_{Y|x}^2 \equiv \hat{\sigma}^2$ του σ^2 και το γεγονός ότι

$$\frac{nS_{Y|x}^2}{\sigma^2} = \frac{(n-2)S_{Y|x}^{*2}}{\sigma^2} \sim X_{n-2}^2$$

καταλήγουμε στο συμπέρασμα ότι,

$$\frac{\hat{Y} - \mu_{Y|x}}{\sqrt{S_{Y|x}^{*2} \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right\}}} \sim t_{n-2} \quad (30)$$

Επομένως σαν ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το $\mu_{Y|x}$ μπορούμε να χρησιμοποιήσουμε το διάστημα

$$\hat{Y} \pm t_{n-2, 1-\alpha/2} \sqrt{S_{Y|x}^{*2} \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right\}}. \quad (31)$$

Σημείωση 1: Είναι σημαντικό να προσέξει κανείς ότι, όπως φαίνεται από τη σχέση (28), ο συντελεστής $(d_i + c_i x)$ εξαρτάται από το x , και επομένως δεν πρέπει να μας εκπλήσσει το γεγονός ότι η διασπορά του $\hat{Y}$ εξαρτάται από το x . Επίσης από την (31) φαίνεται ότι όσο περισσότερο διαφέρει η τιμή του x από το $\bar{x}$, τόσο μεγαλύτερο είναι το διάστημα εμπιστοσύνης για το $\mu_{Y|x}$. Το διάστημα εμπιστοσύνης έχει το ελάχιστο μήκος όταν $x = \bar{x}$.

Σημείωση 2: Θα πρέπει να είναι κανείς προσεκτικός όταν προσπαθεί να εκτιμήσει την τιμή του $\mu_{Y|x}$ για τιμές του x που είναι έξω από το πεδίο τιμών του x στο τυχαίο δείγμα. Από τη μαθηματική πλευρά βέβαια δεν υπάρχει πρόβλημα γιατί το μόνο που έχει να κάνει κανείς

είναι να αντικαταστήσει την τιμή του x στον κατάλληλο τύπο. Παρόλα αυτά συμπεραματολογία τέτοιας μορφής θα πρέπει να αποφεύγεται γιατί εγκυμονεί κινδύνους σφαλερών συμπερασμάτων. Αυτό γιατί η αρχική υπόθεση για γραμμικές σχέσεις ανάμεσα στην $E(Y|x)$ και στο x έγινε για τις τιμές του x που έχουμε στο τυχαίο δείγμα. Είναι πολύ πιθανό η γραμμική σχέση να μην εξακολουθεί να ισχύει για τιμές του x που αποκλίνουν πολύ από τις τιμές του x στο δοθέν δείγμα.

1.2.4 Διαστήματα Πρόβλεψης (Prediction Intervals)

Αντί να χρησιμοποιήσει τη μέση τιμή του Y για δοθέν x ο ερευνητής είναι πιθανό να ενδιαφέρεται να προβλέψει μια επόμενη τιμή του Y_{n+1} για δεδομένο x_{n+1} . Προφανώς το να κάνει κανείς μια τέτοια πρόβλεψη με ακρίβεια είναι περισσότερο δύσκολο από το να προβλέψει μια εκτίμηση για την $E(Y_{n+1} | x)$. Αυτό γιατί στην πρόβλεψη μιας συγκεκριμένης τιμής υπάρχουν δύο αιτίες διασποράς. Διασπορά στην εκτίμηση του πραγματικού μέσου $\mu_{Y_{n+1}|x}$ και διασπορά της συγκεκριμένης τιμής γύρω από τη μέση τιμή. Μια σημειακή εκτιμήτρια του Y_{n+1} είναι φυσικά το $\hat{Y}_{n+1} = \hat{\alpha} + \hat{\beta}x_{n+1}$. Το σφάλμα πρόβλεψης $\Delta = \hat{Y}_{n+1} - Y_{n+1}$ αποδεικνύεται ότι ακολουθεί την κανονική κατανομή με μέση τιμή 0 και διασπορά $\sigma^2 \left\{ \frac{n+1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right\}$. Δηλαδή,

$$\hat{Y}_{n+1} - Y_{n+1} \sim N \left(0, \sigma^2 \left\{ \frac{n+1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right\} \right)$$

(Η απόδειξη στηρίζεται στο γεγονός ότι η τυχαία μεταβλητή $\hat{Y}_{n+1}$ είναι ανεξάρτητη από τις τυχαίες μεταβλητές $\hat{\alpha}$ και $\hat{\beta}$).

Επομένως, μια αμερόληπτη εκτιμήτρια της διακύμανσης του σφάλματος πρόβλεψης όταν το σ^2 είναι άγνωστο, δίνεται από την σχέση:

$$S_{\hat{Y}_{n+1}-Y_{n+1}}^{*2} = S_{Y|x}^2 \left\{ \frac{n+1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right\}$$

Αν θεωρήσουμε την τυποποιημένη κατανομή του $\hat{Y}_{n+1} - Y_{n+1}$ και στη συνέχεια διαιρέσουμε την $N(0,1)$ που καταλήξαμε με την τετραγωνική ρίζα του $\frac{nS_{Y|x}^2}{\sigma^2(n-2)} = \frac{S_{Y|x}^{*2}}{\sigma^2}$ θα οδηγηθούμε σε μια κατανομή t με $n-2$ βαθμούς ελευθερίας. Μετά τις απλοποιήσεις θα έχουμε

$$\frac{\hat{Y}_{n+1} - Y_{n+1}}{\sqrt{S_{Y|x}^2 \left\{ \frac{n+1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right\}}} \sim t_{n-2} \quad (32)$$

Αυτό συνεπάγεται ότι σαν ένα $100(1-\alpha)\%$ διάστημα πρόβλεψης για το Y_{v+1} μπορούμε να χρησιμοποιήσουμε το διάστημα

$$\hat{Y}_{n+1} \pm t_{n-2, 1-\alpha/2} \sqrt{S_{Y|x}^{*2} \left\{ \frac{n+1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right\}}. \quad (33)$$

1.3 Έλεγχοι Υποθέσεων

Σε πολλές περιπτώσεις ο ερευνητής ενδιαφέρεται να κάνει έλεγχο υποθέσεων για τις παραμέτρους της γραμμικής παλινδρόμησης.

Όπως είναι γνωστό από την στατιστική συμπερασματολογία οι έλεγχοι υποθέσεων για τις παραμέτρους χρησιμοποιούν τις ίδιες στατιστικές συναρτήσεις που χρησιμοποιούνται για την κατασκευή διαστημάτων εμπιστοσύνης.

Είναι αυτονόητο ότι και για τον έλεγχο υποθέσεων είναι απαραίτητο να ισχύουν οι υποθέσεις που οδήγησαν στην κατασκευή των διαστημάτων εμπιστοσύνης. Δηλαδή,

- i) Τα ε_i είναι $N(0, \sigma^2)$ για όλα τα $i = 1, 2, \dots$
- ii) Τα ε_i είναι ανεξάρτητα μεταξύ τους.

Η ισοδύναμα,

iii) Τα Y_i είναι $N(\alpha + \beta x_i, \sigma^2)$, $i = 1, 2, \dots$

iv) Τα Y_i είναι ανεξάρτητα μεταξύ τους.

Επίσης υποθέτουμε ότι τα x_i είναι προκαθορισμένα.

1.3.1 Έλεγχος Υποθέσεων για το β

Έστω η μηδενική υπόθεση

$$H_0 : \beta = \beta_0$$

Ως στατιστική συνάρτηση ελέγχου, κάτω από την H_0 , χρησιμοποιείται η συνάρτηση

$$T = \frac{\hat{\beta} - \beta_0}{\sqrt{S_{Y|x}^{*2} / S_{xx}}} \sim t_{n-2}$$

Η απόφαση που θα πάρουμε για τη μηδενική υπόθεση εξαρτάται από την εναλλακτική υπόθεση. Συγκεκριμένα (για επίπεδο σημαντικότητας α):

i) $H_1 : \beta > \beta_0$. Απορρίπτουμε την H_0 αν $t > t_{n-2, 1-\alpha}$

ii) $H_1 : \beta < \beta_0$. Απορρίπτουμε την H_0 αν $t < t_{n-2, \alpha}$

iii) $H_1 : \beta \neq \beta_0$. Απορρίπτουμε την H_0 αν $t > t_{n-2, 1-\alpha/2}$ ή
 $t < t_{n-2, \alpha/2}$

Σημείωση: Συνήθως μας ενδιαφέρει ο έλεγχος της υπόθεσης $H_0 : \beta = 0$, αν δηλαδή υπάρχει γραμμική σχέση μεταξύ των μεταβλητών X και Y .

Με την ίδια λογική ελέγχουμε υποθέσεις για τις παραμέτρους α , $\mu_{Y|x}$ και σ^2 της γραμμικής παλινδρόμησης. Οι στατιστικές συναρτήσεις που χρησιμοποιούνται κάτω από την H_0 : $\alpha = \alpha_0$, H_0 : $\mu_{Y|x} = \mu_{Y|x_0}$ και H_0 : $\sigma^2_{Y|x} = \sigma^2_{Y|x_0}$ είναι, αντίστοιχα, οι:

$$T = \frac{\hat{\alpha} - \alpha_0}{\sqrt{S_{Y|x}^{*2} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} \sim t_{n-2}$$

$$T = \frac{\hat{Y}_{-0} - \mu_{Y|x}}{\sqrt{S_{Y|x}^{*2} \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)}} \sim t_{n-2}$$

$$T = \frac{\hat{Y}_{n+1} - Y_{n+1}}{\sqrt{S_{Y|x}^{*2} \left(\frac{n+1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right)}} \sim t_{n-2}$$

1.3.2 Μερικά Γενικά Παραδείγματα

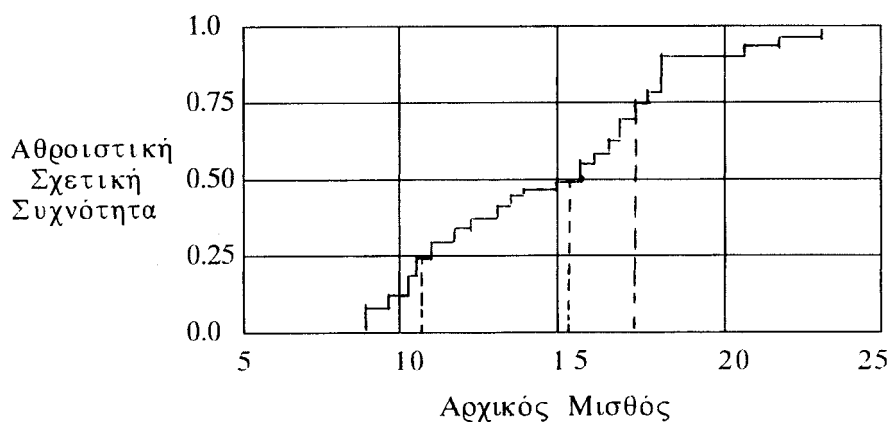
Παράδειγμα 1: Ο υπολογισμός των ασφαλίσεων για διάφορες μορφές ασφάλειας που παρέχουν οι ασφαλιστικές εταιρίες είναι, μία πολύπλοκη διαδικασία που απαιτεί ειδικευμένο προσωπικό. Η επιστήμη που ασχολείται με το θέμα αυτό λέγεται **ασφαλιστική επιστήμη** και οι επιστήμονες που ασχολούνται σ' αυτή ασφαλιστές. Στη διεθνή πρακτική για να ασκήσει κάποιος το επάγγελμα του ασφαλιστή θα πρέπει να είναι μέλος του αντίστοιχου επιμελητηρίου ασφαλιστών. Για να αποκτηθεί το δικαίωμα αυτό στο εξωτερικό θα πρέπει ο υποψήφιος να πετύχει δέκα διαγωνισμούς. Ο πρώτος από αυτούς στα μαθηματικά και τον απειροστικό λογισμό. Ο δεύτερος σε πιθανότητες και στατιστική. Οι εξετάσεις στο τρίτο, το τέταρτο και το πέμπτο αντικείμενο αφορούν διάφορα θέματα στα μαθηματικά τα χρηματοπιστωτικά και τις ασφάλειες. Αν κάποιος υποψήφιος πετύχει στις πέντε από τις εξετάσεις έχει το δικαίωμα να γίνει πάρεδρο μέλος του ασφαλιστικού επιμελητηρίου. Επιτυχία στις εξετάσεις και στα δέκα αντικείμενα δίνει την δυνατότητα σε κάποιον να γίνει τακτικό μέλος του επιμελητηρίου αυτού. Τα προπτυχιακά πανεπιστημιακά προγράμματα σε ασφαλιστικές επιστήμες είναι σχεδιασμένα με τρόπο που να βοηθούν τους φοιτητές να εξετασθούν με επιτυχία σε όσο το δυνατόν περισσότερα από τα πέντε πρώτα αντικείμενα (εξετάσεις) πριν αποκτήσουν το πτυχίο τους.

Ας υποθέσουμε ότι ένας προπτυχιακός φοιτητής ενός τέτοιου προγράμματος αμερικάνικου πανεπιστημίου ενδιαφέρεται να πληροφορηθεί το μισθό που μπορεί να προσδοκά μετά την απόκτηση του πτυχίου του. Το συγκεκριμένο πανεπιστήμιο προκειμένου να

βοηθήσει τους φοιτητές του στην επαγγελματική τους σταδιοδρομία, συγκεντρώνει στοιχεία για τους μισθούς των πτυχιούχων του. Τα στοιχεία για ένα τυχαίο δείγμα 28 φοιτητών που πρόσφατα πήραν το πτυχίο τους που αναφέρονται στον αριθμό των εξετάσεων για ασφαλιστές που πέρασαν και τους αρχικούς μισθούς τους αμέσως μετά την απόκτηση του πτυχίου τους παρατίθεται στον πίνακα που ακολουθεί:

Πτυχ/χοι	Αρχικός Μισθός (σε \$)	Αριθμός επιτυχών εξετάσεων	Πτυχ/χοι	Αρχικό Μισθός (σε \$)	Αριθμός επιτυχών εξετάσεων
1	\$23800	1	15	\$12000	1
2	18000	3	16	15900	2
3	17300	2	17	13000	1
4	18000	2	18	8900	0
5	21300	5	19	17000	3
6	10200	0	20	20500	4
7	11100	0	21	14800	2
8	15900	3	22	10200	0
9	18000	4	23	13300	1
10	13100	2	24	12500	1
11	9400	0	25	17800	4
12	16500	2	26	17000	3
13	22500	5	27	11200	1
14	11100	1	28	16600	4

Μια πρώτη προσέγγιση για τον προσδιορισμό του αναμενόμενου μισθού του συγκεκριμένου φοιτητή είναι δυνατό να δοθεί από το γράφημα της εμπειρικής συνάρτησης κατανομής του δείγματος των αρχικών μισθών, όπως φαίνεται και από το σχήμα που ακολουθεί.



Εμπειρική συνάρτηση κατανομής αρχικών μισθών

Από την εμπειρική συνάρτηση κατανομής προκύπτει ότι η διάμεσος του δείγματος είναι \$15350 ενώ το ανώτερο και το κατώτερο τεταρτημόριο (*lower and upper quartiles*) είναι \$11200 και \$17300, αντίστοιχα. Από αυτό προκύπτει ότι το μέσο 50% των φοιτητών που πρόσφατα κατέστησαν πτυχιούχοι έχει αρχικό μισθό από \$11200 έως \$17300. Αυτό δίνει και μια ιδέα στον συγκεκριμένο φοιτητή για το τι μπορεί να περιμένει κατά προσέγγιση.

Ας υποθέσουμε ότι έχουμε την πρόσθετη πληροφορία ότι ο συγκεκριμένος φοιτητής έχει περάσει και τις πέντε πρώτες εξετάσεις του ασφαλιστικού επιμελητηρίου. Η πρόσθετη αυτή πληροφορία είναι δυνατόν να βοηθήσει τον φοιτητή να εκτιμήσει με μεγαλύτερη προσέγγιση τον αρχικό μισθό που μπορεί να περιμένει με βάση τα υπάρχοντα στοιχεία. Εάν τοποθετήσουμε τα προηγούμενα δεδομένα συνοπτικά σε ένα πίνακα ανάλογα με τον αριθμό των εξετάσεων στις οποίες έχουν πετύχει οι πρόσφατοι πτυχιούχοι έχουμε τον παρακάτω πίνακα.

**Αριθμός επιτυχών εξετάσεων και
οι αντίστοιχοι αρχικοί μισθοί**

$x_0 = 0$	$x_1 = 1$	$x_2 = 2$	$x_3 = 3$	$x_4 = 4$	$x_5 = 5$
\$8.900	\$11.100	\$13.100	\$15.900	\$ 16.600	\$ 21.300
9.400	11.200	14.800	17.000	17.800	22.500
10.200	11.900	15.900	17.000	18.000	
12.200	12.000	16.500	18.000	20.500	
11.100	13.000	17.300			
	13.000	18.000			

$$\bar{y}_0 = \$9.960 \quad \bar{y}_1 = \$12.143 \quad \bar{y}_2 = \$15.933$$

$$\bar{y}_3 = \$16.975 \quad \bar{y}_4 = \$18.225 \quad \bar{y}_5 = \$21.900$$

(Προφανώς, τα $\bar{y}_0, \bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4$ και $\bar{y}_5$ είναι σημειακές εκτιμήσεις των $\mu_{Y_0|x_0}, \mu_{Y_1|x_1}, \mu_{Y_2|x_2}, \mu_{Y_3|x_3}, \mu_{Y_4|x_4}$ και $\mu_{Y_5|x_5}$ αντίστοιχα).

Από τα παραπάνω στοιχεία προκύπτει ότι ο μέσος αρχικός μισθός τείνει να αυξάνεται ανάλογα με τον αριθμό των εξετάσεων που έχουν περάσει οι πτυχιούχοι του δείγματος. Είναι προφανές ότι οι πρόσθετες πληροφορίες για τον αριθμό των εξετάσεων που έχουν περάσει οι φοιτητές είναι χρήσιμες για τον καλύτερο υπολογισμό των αρχικών μισθών. Ένας πτυχιούχος για παράδειγμα που έχει περάσει και τις πέντε εξετάσεις θα πρέπει να περιμένει ένα μισθό πλησιέστερο προς τον μέσο αρχικό μισθό των άλλων πτυχιούχων που έχουν περάσει και τις πέντε εξετάσεις.

Άσκηση: α. Να καθορισθεί η ευθεία γραμμικής παλινδρόμησης για την σχέση των αρχικών μισθών των νέων πτυχιούχων με τον αριθμό των εξετάσεων που έχουν περάσει.

β. Να κατασκευασθούν διαστήματα εμπιστοσύνης για τις παραμέτρους $\alpha, \beta, \mu_{Y|x}$ και διάστημα πρόβλεψης για την τιμή του Y για πτυχιούχους που έχουν περάσει τρεις εξετάσεις.

Παράδειγμα 2: Τα βήτα των μετοχών (*stock betas*) έχουν χρησιμοποιηθεί πολλές φορές από τις αρχές της δεκαετίας του 70, σαν

ένα μέτρο της αστάθειας μιας μετοχής σε σχέση με την υπόλοιπη χρηματιστηριακή αγορά. Κάθε μετοχή έχει τη δική της τιμή για το βήτα η οποία εκτιμάται από παλαιότερα στοιχεία και χρησιμοποιείται σαν μια ένδειξη του κινδύνου (risk) της συγκεκριμένης μετοχής σε σχέση με τον δείκτη αγοράς. (Ο δείκτης αγοράς έχει εξ ορισμού $\beta = 1.00$).

Μια μετοχή της οποίας η αξία τείνει να αυξάνει ή να μειώνεται κατά 20% περισσότερο από ότι η αγορά ως σύνολο ανεβαίνει ή πέφτει θεωρείται ότι έχει τιμή για το βήτα 1.20. Εάν εξάλλου η αύξηση (ή πτώση) της τιμής μιας μετοχής είναι 15% λιγότερο από την αγορά θεωρείται ότι έχει βήτα με τιμή 0.85.

Μερικοί αναλυτές χαρτοφυλακίου (*portfolio analysts*) συγκεντρώνουν τα χαρτοφυλάκια των πελατών τους παρατηρώντας τα βήτα των μετοχών τα οποία διακινούν. Ένα χαρτοφυλάκιο για να θεωρείται συντηρητικό θα πρέπει να μην μεταβάλλεται σε τιμή τόσο όσο η αγορά και επομένως θα πρέπει να έχει ένα μέσο βήτα μικρότερο από 1.00. Ένα πιο ριψοκίνδυνο χαρτοφυλάκιο θα αποτελείται από μετοχές με μέσο βήτα μεγαλύτερο του 1.00. Τα βήτα των μετοχών πολλές φορές αλλάζουν με το χρόνο. Οι μετοχές χρυσού (*gold stocks*) όπως επίσης και οι μετοχές εταιριών εξόρυξης χρυσού συνήθως μεταβάλλονται αντίθετα από την υπόλοιπη αγορά και επομένως έχουν αρνητικά βήτα. Παρόλα αυτά από το 1974 και μέχρι την αρχή της δεκαετίας του 90 τα βήτα της ομάδας των μετοχών χρυσού είχαν αυξηθεί σε μια τιμή που κινείτο γύρω στο 0.65. Η τιμή αυτή κάνει τις μετοχές χρυσού να εξακολουθούν να θεωρούνται σαν μια από τις ελάχιστα ασταθείς ομάδες μετοχών η οποία όμως, παρόλα αυτά, αυξομειώνεται αντίστοιχα με τις υπόλοιπες αυξομειώσεις της αγοράς. Μια άλλη κατηγορία μετοχών μικρότερης αστάθειας είναι αυτή των εταιριών κατανομής φυσικού αερίου με βήτα 0.59, των εταιριών επεξεργασίας ζάχαρης με βήτα 0.69, των εταιριών παραγωγής ηλεκτρικής ενέργειας με βήτα 0.75 και των αποθεμάτων καπνού με βήτα 0.78.

Μερικοί αναλυτές χαρτοφυλακίου (*portfolio analysis*) συγκεντρώνουν τα χαρτοφυλάκια των πελατών τους παρατηρώντας τα βήτα των μετοχών τα οποία διακινούν. Ένα χαρτοφυλάκιο για να θεωρείται

συντηρητικό θα πρέπει να μην μεταβάλλεται σε τιμή τόσο όσο η αγορά και επομένως θα πρέπει να έχει ένα μέσο βήτα μικρότερο από 1.00. Ένα πιο ριψοκίνδυνο χαρτοφυλάκιο θα αποτελείται από μετοχές με μέσο βήτα μεγαλύτερο του 1.00. Τα βήτα των μετοχών πολλές φορές αλλάζουν με τον χρόνο. Οι μετοχές χρυσού (*gold stocks*) όπως επίσης και οι μετοχές εταιριών εξόρυξης χρυσού συνήθως μεταβάλλονται αντίθετα από την υπόλοιπη αγορά και επομένως έχουν αρνητικά βήτα. Παρόλα αυτά από το 1974 και μέχρι την αρχή της δεκαετίας του 90 τα βήτα της ομάδας των μετοχών χρυσού είχαν αυξηθεί σε μία τιμή που κινείτο γύρω στο 0.65. Η τιμή αυτή κάνει τις μετοχές χρυσού να εξακολουθούν να θεωρούνται σαν μία από τις ελάχιστα ασταθείς ομάδες μετοχών η οποία όμως, παρόλα αυτά, αυξομειώνεται αντίστοιχα με τις υπόλοιπες αυξομειώσεις της αγοράς. Μια άλλη κατηγορία μετοχών μικρότερης αστάθειας είναι αυτή των εταιριών κατανομής φυσικού αερίου με βήτα 0.59, των εταιριών επεξεργασίας ζάχαρης με βήτα 0.69, των εταιριών παραγωγής ηλεκτρικής ενέργειας με βήτα 0.75 και των αποθεμάτων καπνού με βήτα 0.78. Οι εταιρίες κατασκευής φαρμάκων και ζυθοποιίας εθεωρούντο μικρότερης αστάθειας με βήτα γύρω στο 0.79 μέχρι το 1973. Η κατάσταση αυτή όμως από το 1974 άρχισε να αλλάζει και μάλιστα το 1979 το αντίστοιχο βήτα των εταιριών αυτών έχει αυξηθεί σε 1.24 κάνοντας τις μετοχές αυτές να περιλαμβάνονται στον όμιλο μετοχών με την μεγαλύτερη αστάθεια. Οι πιο ασταθείς ομάδες στην αρχή της δεκαετίας του 80 ήταν αυτές λυόμενων σπιτιών (με βήτα περίπου 2.30) και των εταιριών κατασκευής σπιτιών με βήτα 2-21. Το μέσο βήτα για έξι χρόνια από το 1974 έως το 1979 για τις δύο αυτές τελευταίες κατηγορίες μετοχών ήταν 2.40 και 2.16 αντίστοιχα.

Ένας αναλυτής χρηματιστηρίου ενδιαφέρεται να συμβουλευεί έναν πελάτη του για την τιμή του βήτα μιας τοπικής επιχείρησης κοινής ωφέλειας. Ο αναλυτής δεν γνωρίζει την τιμή του βήτα για την συγκεκριμένη εταιρία έχει όμως την δυνατότητα να προσφύγει σε στοιχεία για τις τριμηνιαίες μεταβολές της αξίας των μετοχών της εταιρίας για τα προηγούμενα τρία χρόνια. Οι μεταβολές αυτές μπορούν να συγκριθούν με τις μεταβολές της τιμής ενός τυπικού δείκτη 500 μετοχών (*standard and Poor's 500 stock index* ή αλλιώς *S*

& *P index*) που χρησιμοποιείται σαν ένας δείκτης της αγοράς. Ο πίνακας που ακολουθεί δίνει τις μεταβολές εκφρασμένες ως εκατοστιαίες μεταβολές για τα 12 τρίμηνα των δεδομένων.

Τρίμηνο	Δείκτης S & P	Τοπική Εταιρία
1	-4.3%	-7.0%
2	-9.9	-11.4
3	-5.4	-5.3
4	-0.2	-4.3
5	+4.9	+1.4
6	+1.8	+1.6
7	+1.3	+4.6
8	-1.3	+0.1
9	+5.1	+7.3
10	-1.6	+1.8
11	-9.7	-7.6
12	-12.8	-16.2

Εκείνο που ενδιαφέρει στο συγκεκριμένο πρόβλημα είναι πώς ο αναλυτής μπορεί χρησιμοποιώντας τα παραπάνω στοιχεία, να εκτιμήσει το βήτα της συγκεκριμένης επιχείρησης κοινής ωφέλειας.

Ένας τρόπος για να απαντήσει κανείς στο ερώτημα αυτό είναι να χρησιμοποιήσει τη θεωρία της γραμμικής παλινδρόμησης εφαρμόζοντας μια ευθεία γραμμή στις παρατηρήσεις που δίνονται στο παραπάνω πίνακα. Ο ρυθμός μεταβολής των τιμών της επιχείρησης σε σχέση με τον δείκτη S & P που θα προκύψει από την εκτιμηθείσα ευθεία μπορεί να χρησιμοποιηθεί σαν μια εκτίμηση για την τιμή βήτα της συγκεκριμένης επιχείρησης. Φυσικά θα πρέπει να υποθέσουμε ότι η καμπύλη παλινδρόμησης είναι ευθεία όπως επίσης και να κάνουμε όλες τις υπόλοιπες υποθέσεις που είναι απαραίτητες για τη χρησιμοποίηση της θεωρίας της γραμμικής παλινδρόμησης.

Κάνοντας τους σχετικούς υπολογισμούς βρίσκουμε ότι $b = 1.09$. Η σημειακή αυτή εκτίμηση του βήτα, για τη συγκεκριμένη μετοχή αποτελεί μια ένδειξη ότι οι μετοχές της επιχείρησης αυτής τείνουν να είναι περισσότερο ασταθείς (ή, διαφορετικά, έχουν

μεγαλύτερη επικινδυνότητα) από την γενική αγορά. Σύμφωνα με τα προαναφερθέντα το βήτα μιας μετοχής με τιμή 1.09 δίνει μια ένδειξη ότι αύξηση (ή πτώση) κατά 1% στην τιμή του δείκτη S & P τείνει να επιφέρει μια αύξηση (ή πτώση) της τάξης του 1.09% στην τιμή της μετοχής της συγκεκριμένης επιχείρησης.

Η αντίστοιχη εκτίμηση για το a είναι $a = -0.021$. Επομένως η εκτιμήτρια της ευθείας παλινδρόμησης για το συγκεκριμένο πρόβλημα είναι

$$\hat{\mu}_{Y|x} = -0.021 + 1.093x$$

Αν μας ενδιαφέρει ένα 95% διάστημα εμπιστοσύνης για το βήτα βλέπουμε ότι από τα δεδομένα μας προκύπτουν οι εξής τιμές:

$$S_{xx}^2 = 372.7097 \quad S_{yy} = 525.8770 \quad S_{xy} = 407.2904 \quad S_{Y|x}^{*2} = 8.0620$$

Από τα στοιχεία προκύπτει ότι ένα 95% διάστημα εμπιστοσύνης για το βήτα είναι το διάστημα (0.765, 1.421).

Έστω στη συνέχεια ότι κάποια οικονομική πρόβλεψη για το δείκτη S & P διαβλέπει μια πτώση τιμής του κατά 7%. Ένα ερώτημα είναι πώς αναμένεται η πτώση αυτή να επηρεάσει την τιμή της μετοχής της συγκεκριμένης επιχείρησης η οποία μας ενδιαφέρει. Υπάρχουν δύο προσεγγίσεις σαν απάντηση του προβλήματος αυτού. Η πρώτη είναι να κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης για την αναμενόμενη τιμή $\mu_{Y|x}$ της μετοχής το επόμενο τρίμηνο με την υπόθεση ο δείκτης S & P παρουσιάζει απόδοση -7%. Για την περίπτωση αυτή με τα συγκεκριμένα στοιχεία του δείγματος το 95% διάστημα εμπιστοσύνης για το $\mu_{Y|x}$ είναι το (-10.0%, -5.4%). Το διάστημα αυτό εμπιστοσύνης αναφέρεται στη μέση απόδοση της μετοχής αυτής για το επόμενο τρίμηνο. Εάν κάποιος αναλυτής ενδιαφέρεται για την απόκλιση που αναφέρεται σε μια συγκεκριμένη παρατήρηση τότε θα πρέπει να κατασκευάσει ένα 95% διάστημα πρόβλεψης για την τιμή της μετοχής αυτής Y για το επόμενο τρίμηνο με την προϋπόθεση ότι ο δείκτης S & P αυτός έχει μία πτώση 7%. Για τα δεδομένα μας στο διάστημα αυτό προκύπτει ότι θα είναι (-14.4%, -1%).

Παράδειγμα 3: (συνάρτηση κατανάλωσης του Keynes (Keynes's consumption function)).

Ο Άγγλος οικονομολόγος John Maynard Keynes ήθελε να εξηγήσει τις αποκλίσεις που υπάρχουν στην καταναλωτική συμπεριφορά. Η άποψη του Keynes ήταν ότι η καταναλωτική συμπεριφορά είναι ένα από τα βασικά στοιχεία για την κατανόηση των λόγων που προκαλούν οικονομική ύφεση. Για το λόγο αυτό, ήθελε να μελετήσει τις αιτίες που προκαλούν τις μεταβολές στην καταναλωτική συμπεριφορά. Ο Keynes υπέθεσε ότι το οικογενειακό εισόδημα (household income) είναι ο κύριος παράγοντας που καθορίζει την οικογενειακή κατανάλωση. Η εξήγηση που έδωσε είναι ότι όταν το εισόδημα αυξάνεται, οι άνθρωποι ξοδεύουν περισσότερο. Όταν το εισόδημα ελαττώνεται ξοδεύουν λιγότερο. Ο Keynes ισχυρίστηκε (βλ. Keynes (1936)) ότι "... αυτός είναι ένας βασικός ψυχολογικός κανόνας πάνω στον οποίο μπορούμε να στηριζόμαστε και εκ των προτέρων από τις γνώσεις μας της ανθρώπινης φύσης και εκ των υστέρων από λεπτομερή στοιχεία και εμπειρία". (... "it is a fundamental psychological law upon which we are entitled to depend both a priori from our knowledge of human nature and from the detailed facts of experience").

Με βάση την υπόθεση αυτή, η θεωρία του Keynes εκφράζεται από τη σχέση $E(Y|X=x) = \alpha + \beta x$. Στο μοντέλο αυτό το εισόδημα είναι η επεξηγηματική μεταβλητή (explanatory variable), διότι μεταβολές στο εισόδημα ερμηνεύουν μεταβολές στην κατανάλωση. Η κατανάλωση είναι η εξαρτημένη μεταβλητή (dependent variable), δεδομένου ότι η κατανάλωση εξαρτάται από το εισόδημα. (Στη συγκεκριμένη περίπτωση, το μοντέλο αυτό εμπεριέχει την υπόθεση της σχέσης αιτίου και αιτιατού. Ο ερευνητής δηλαδή πιστεύει ότι μεταβολές στην επεξηγηματική μεταβλητή προκαλούν μεταβολές στην εξαρτημένη μεταβλητή).

Η γραμμική αυτή προσέγγιση είναι αρκετά ακριβής εφόσον περιορίζουμε την ερμηνεία μας σε μικρές μεταβολές του εισοδήματος. Από την γραμμική αυτή υπόθεση, προκύπτει ότι όσο το εισόδημα αυξάνεται ή ελαττώνεται, υπό τα παρόντα δεδομένα, η κατανάλωση αυξάνεται ή μειώνεται κατά περίπου β φορές την αντίστοιχη

μεταβολή του εισοδήματος. Βέβαια, η προσέγγιση αυτή δεν ισχύει για πολύ μεγάλες μεταβολές στο εισόδημα. Για παράδειγμα, το σημείο τομής α δίνει μια πολύ ανακριβή πρόβλεψη της κατανάλωσης όταν το εισόδημα γίνει μηδέν. Επειδή ακριβώς η πραγματική σχέση της κατανάλωσης με το εισόδημα δεν είναι ακριβώς γραμμική παρά μόνο κατά προσέγγιση γραμμική, όταν το εισόδημα γίνει μηδέν ($X=0$) δεν θα πρέπει να ερμηνεύουμε το α ως το επίπεδο την κατανάλωσης για $X=0$. Αντίθετα, θα πρέπει να ερμηνεύουμε το α ως το σημείο τομής μιας κατά προσέγγιση ευθείας παλινδρόμησης με τον άξονα των Y . Αυτή η κατά προσέγγιση ευθεία παλινδρόμησης θα μπορεί να δώσει ακριβή πρόβλεψη της κατανάλωσης για τιμές του εισοδήματος πλησίον του ισχύοντος επιπέδου. Αυτό γιατί, όπως ήδη είπαμε, οι γραμμικές προσεγγίσεις είναι πολύ χρήσιμες και συνήθως πολύ ικανοποιητικές εφόσον δεν προσπαθούμε να βγάλουμε συμπεράσματα για καταστάσεις για τις οποίες έχουμε πολύ μικρές γνώσεις ή ενδείξεις.

Προκειμένου να ελεγχθεί η υπόθεση του Keynes, ας θεωρήσουμε μερικά δεδομένα που αναφέρονται στην πραγματική κατανάλωση ανά νοικοκυριό και στο διαθέσιμο εισόδημα (μετά τους φόρους) για τα έτη 1977 έως και 1981 στις Η.Π.Α.. (Ο οικονομικός όρος "πραγματικό" σημαίνει ότι έχει γίνει μία προσαρμογή ώστε να ληφθεί υπόψη ο πληθωρισμός. Τα δεδομένα που ακολουθούν αναφέρονται σε τιμές του 1967, δηλαδή, τι θα μπορούσε να έχει αγοράσει με ένα δολάριο κανείς το 1967. Το 1981 τα αγαθά κόστιζαν, κατά μέσο όρο, τρεις φορές περίπου όσο κόστιζαν το 1967. Επομένως, 9000 δολάρια το 1981 έχουν περίπου την ίδια αξία όπως 3000 δολάρια το 1967).

Έτος	Διαθέσιμο εισόδημα X (σε χιλ \$)	Οικογενειακή κατανάλωση Y (σε χιλ \$)
1977	3319	3042
1978	3421	3124
1979	3404	3108
1980	3276	2994
1981	3271	2971

Χρησιμοποιώντας τη μέθοδο των ελαχίστων τετραγώνων, βρίσκουμε ότι οι τιμές των εκτιμητριών $\hat{\alpha}$ και $\hat{\beta}$ για τα συγκεκριμένα δεδομένα είναι

$$b = 0.9474 \text{ και } a = -114.7.$$

Επομένως, η εκτιμώμενη με βάση τα δεδομένα αυτά ευθεία παλινδρόμησης είναι, κατά προσέγγιση η

$$\hat{y} = -115 + 0.947x.$$

Η παράμετρος β του μοντέλου, την οποία ο Keynes ονόμασε περιθωριακή τάση για κατανάλωση (marginal propensity to consume) μας δίνει μια ένδειξη για τον τρόπο που μεταβολές στο εισόδημα επηρεάζουν την κατανάλωση. Ο Keynes ισχυρίστηκε ότι το β είναι θετικό αλλά χωρίς αμφιβολία μικρότερο από 1. Η ευθεία παλινδρόμησης που εκτιμήσαμε για τα συγκεκριμένα δεδομένα μας οδηγεί στο συμπέρασμα ότι πρόσθετο εισόδημα ενός δολαρίου αυξάνει την κατανάλωση κατά περίπου 95 cents ενώ, αντίθετα, μείωση του εισοδήματος κατά ένα δολάριο μειώνει την κατανάλωση κατά 95 cents.

Υπολογισμοί όπως οι παραπάνω μπορούν να οδηγήσουν σε διάφορων μορφών προβλέψεις αισιόδοξες ή απαισιόδοξες ή να χρησιμοποιηθούν για να καθορίσουν τα αποτελέσματα που θα έχουν διάφορες κυβερνητικές αποφάσεις. Αν, για παράδειγμα, η αμερικανική κυβέρνηση αποφάσιζε μία μείωση φόρων και αν ίσχυε το οικονομικό μοντέλο το οποίο εκτιμήσαμε με βάση τα δεδομένα, θα μπορούσε κανείς να οδηγηθεί στο συμπέρασμα ότι μια ελάττωση της

φορολογίας κατά 100 δολάρια, θα οδηγούσε σε αύξηση της κατανάλωσης κατά περίπου 95 δολάρια.

Αν θέλουμε να προχωρήσουμε σε στατιστική συμπερασματολογία, από τα δεδομένα βρίσκουμε ότι $S_{xx} = 19938.8$ και, επομένως, $S_{\hat{\beta}}^* = 0.078$ και $S_{\hat{\alpha}}^* = 260$.

Ένας προφανής στατιστικός έλεγχος υπόθεσης για την επιβεβαίωση της υπόθεσης του Keynes είναι ο έλεγχος

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0.$$

Η τιμή της στατιστικής συνάρτησης T για την παράμετρο $\hat{\beta}$ κάτω από την μηδενική υπόθεση είναι

$$t_0 = \frac{0.947}{0.078} = 12.2$$

Η τιμή αυτή οδηγεί σε απόρριψη της μηδενικής υπόθεσης $\beta=0$ σε επίπεδο σημαντικότητας 0.05.

Όσο αφορά την παράμετρο α , η τιμή της στατιστικής συνάρτησης T κάτω από την μηδενική υπόθεση

$$H_0: \alpha = 0$$

$$\text{είναι } t_0 = \frac{-115}{260} = -0.442$$

Με βάση την τιμή αυτή, δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση, δηλαδή $\alpha=0$, στο 5% επίπεδο σημαντικότητας.

Ένας τρόπος να ελέγξουμε τον άλλο ισχυρισμό του Keynes, ότι δηλαδή η κλίση β είναι σχεδόν πάντοτε μικρότερη από 1 είναι να εξετάσουμε αν η τιμή 1 βρίσκεται μέσα στο διάστημα εμπιστοσύνης που κατασκευάζεται για την παράμετρο β . Για τα συγκεκριμένα δεδομένα που έχουμε διαθέσιμα, βλέπουμε ότι, πράγματι η τιμή $\hat{\beta}=1$ της εκτιμήτριας $\hat{\beta}$ βρίσκεται μέσα στο 95% διάστημα εμπιστοσύνης του β και επομένως, με βάση τα συγκεκριμένα δεδομένα, είναι ενδεχόμενο η τιμή του β να υπερβαίνει την τιμή του 1. Αυτό συμβαίνει παρότι η τιμή της εκτιμήτριας είναι $b = 0.947$, τιμή που είναι μικρότερη από το 1 λόγω του ότι το 95% διάστημα

εμπιστοσύνης για το β είναι το (0.791, 1.103), διάστημα που περιέχει τιμή 1.

Τέλος, ο πίνακας που ακολουθεί δίνει το 95% διάστημα πρόβλεψης για διάφορες τιμές X_0 του εισοδήματος.

Είναι ενδιαφέρον να παρατηρήσουμε ότι το διάστημα πρόβλεψης είναι μικρότερο στην τιμή $\bar{X}$ και στη συνέχεια μεγαλώνει όσο απομακρυνόμαστε από την μέση τιμή των δεδομένων μας.

95% Διάστημα Πρόβλεψης για την Κατανάλωση

Εισόδημα X_0	Προβλεπόμενη Κατανάλωση $\hat{Y}_0$	Τυπική Απόκλιση Πρόβλεψης $S_{Y_0 - \hat{Y}_0}^*$	95% Διάστημα Πρόβλεψης
3038.2	2763.6	26.3	(2711.1, 2816.1)
3138.2	2858.3	19.7	(2819.0, 2897.7)
3238.2	2953.1	14.3	(2924.4, 2981.7)
3338.2	3047.8	12.0	(3023.7, 3071.9)
3438.2	3142.5	14.3	(3113.9, 3171.2)
3538.2	3237.3	19.7	(3197.9, 3276.6)
3638.2	3332.0	26.3	(3279.5, 3384.5)

1.4 Χρησιμοποίηση των Καταλοίπων (RESIDUALS) για την Εξέταση των Υποθέσεων του Γραμμικού Μοντέλου

Από την μέχρι τώρα ανάπτυξη της θεωρίας των γραμμικών μοντέλων έχουμε δει πώς μπορεί κανείς να εκτιμήσει την ευθεία παλινδρόμησης με βάση ορισμένα δεδομένα και στη συνέχεια να τη χρησιμοποιήσει. Είδαμε επίσης ότι, προκειμένου να αναπτυχθεί η θεωρία για την κατασκευή διαστημάτων εμπιστοσύνης των παραμέτρων του γραμμικού μοντέλου είναι απαραίτητο να κάνουμε ορισμένες υποθέσεις. **(Σημείωση:** Οι υποθέσεις αυτές δεν χρειάζονται για την κατασκευή της εκτιμώμενης ευθείας παλινδρόμησης αφού η μέθοδος των ελαχίστων τετραγώνων είναι μια γενικά μαθηματική μέθοδος).

Με την μεγάλη εξέλιξη της τεχνολογίας και την ύπαρξη πλήθους στατιστικών πακέτων είναι πολύ εύκολο να κατασκευάσει κανείς την ευθεία παλινδρόμησης και να προχωρήσει σε στατιστική συμπερασματολογία της μορφής που προαναφέραμε. Θα πρέπει όμως να τονισθεί ο κίνδυνος από τη χρησιμοποίηση των στατιστικών πακέτων για την εξαγωγή στατιστικών συμπεράσματος χωρίς να ελέγχονται οι υποθέσεις οι οποίες είναι απαραίτητες για την ανάπτυξη της θεωρίας. Είναι ενδεχόμενο να καταλήξει κανείς σε εξαιρετικά λανθασμένα συμπεράσματα με τη χρησιμοποίηση της μεθόδου της γραμμικής παλινδρόμησης εάν δεν έχει προηγουμένως ελέγξει ότι οι υποθέσεις που απαιτούνταν εξασφαλίζονται.

Ορισμός: Αν $y_1, y_2, \dots, y_n$ είναι οι παρατηρηθείσες τιμές της εξαρτημένης μεταβλητής και $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ είναι οι εκτιμηθείσες τιμές της εξαρτημένης μεταβλητής που προκύπτουν από την ευθεία παλινδρόμησης για τις τιμές $x_1, x_2, \dots, x_n$, αντίστοιχα, οι διαφορές $(y_i - \hat{y}_i)$, $i=1,2,\dots,n$ ονομάζονται **δειγματικά κατάλοιπα** (*Sample residuals*) και συμβολίζονται με $e_i = y_i - \hat{y}_i$, $i=1,2,\dots,n$. Οι ποσότητες αυτές είναι παρατηρούμενες τιμές των τυχαίων μεταβλητών (του τυχαίου δείγματος) $\hat{e}_i = Y_i - \hat{Y}_i$, $i=1,2,\dots,n$.

Οι αντίστοιχες διαφορές $e_i = Y_i - \mu_{Y|X}$ ονομάζονται **κατάλοιπα του πληθυσμού** (*population residuals*).

Προφανώς τα $\hat{e}_i$, $i=1,2,\dots,n$ είναι ένα τυχαίο δείγμα της τυχαίας μεταβλητής $\varepsilon = Y - \alpha - \beta X$.

Παράδειγμα: Τα στοιχεία που παρατίθενται στον πίνακα που ακολουθεί αντιπροσωπεύουν την ταχύτητα ενός αυτοκινήτου (σε χιλιόμετρα ανά ώρα) και αποτελούν τιμές της ανεξάρτητης μεταβλητής, ενώ ως εξαρτημένη μεταβλητή θεωρούμε την απόσταση (σε μέτρα) που απαιτείται για να σταματήσει πλήρως το αυτοκίνητο αυτό όταν φρενάρει.

Απόσταση φρεναρίσματος

x	y	$\hat{y}$	$y - \hat{y}$
10	3,4	-8,0	11,4
20	7,3	4,0	3,3
30	15,3	16,0	-0,7
40	21,0	28,0	-7,0
50	31,8	40,0	-8,2
60	46,5	52,0	-5,5
70	59,2	64,0	-4,8
80	72,4	76,0	-3,6
90	90,5	88,0	2,5
100	112,6	100,0	12,6

Η ευθεία ελαχίστων τετραγώνων που δίνει την καλύτερη προσαρμογή στα δεδομένα αυτά είναι

$$\hat{y} = a + b x = -20 + 1.20x$$

(Η σχέση αυτή χρησιμοποιήθηκε για να υπολογισθούν οι τιμές του $\hat{y}$ οι οποίες προκύπτουν από την εκτιμώμενη ευθεία παλινδρόμησης του πίνακα όπως επίσης και τα δειγματικά κατάλοιπα $y - \hat{y}$).

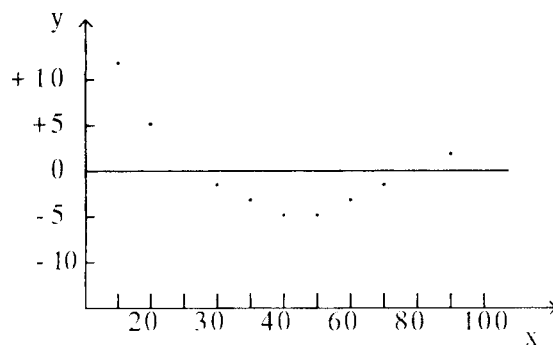
Τα δειγματικά κατάλοιπα (*residuals*) μπορούν να χρησιμοποιηθούν σαν μια πρώτη οπτική προσέγγιση για το κατά πόσο οι υποθέσεις που κάνουμε στο γραμμικό μοντέλο ισχύουν ή όχι. (Υπάρχουν βέβαια και εναλλακτικές μαθηματικές μέθοδοι οι οποίες ελέγχουν τις υποθέσεις αυτές).

Η πρώτη υπόθεση που κάναμε είναι αυτή της **γραμμικότητας**. Όπως είναι φυσικό για οποιοδήποτε τυχαίο δείγμα τιμών έχουμε είναι δυνατόν να υπολογίσουμε με την μέθοδο που προαναφέραμε (των ελαχίστων τετραγώνων) την καλύτερη εκτιμώμενη ευθεία παλινδρόμησης. Αυτό όμως δεν σημαίνει ότι υπάρχει πράγματι γραμμική σχέση μεταξύ των δύο μεταβλητών ώστε να χρησιμοποιήσουμε την αντίστοιχη θεωρία για στατιστική συμπερασματολογία.

Πριν επομένως κάνουμε οποιαδήποτε προσπάθεια για εξαγωγή στατιστικών συμπερασμάτων από την ευθεία παλινδρόμησης θα πρέπει να ελέγξουμε τις υποθέσεις μας.

Ας επανέλθουμε στην υπόθεση της γραμμικότητας. Είναι η πραγματική γραμμή παλινδρόμησης για το συγκεκριμένο υπό εξέταση πρόβλημα ευθεία; Ισχύει δηλαδή ότι $\mu_{Y|x} = \alpha + \beta x$;

Μια πρώτη προσέγγιση για το αν κάτι τέτοιο ισχύει ή όχι παίρνουμε αν κατασκευάσουμε το διάγραμμα της ευθείας παλινδρόμησης των ελαχίστων τετραγώνων ή, όπως κάνουμε συνήθως, το διάγραμμα των **καταλοίπων** (*residuals*). Στον άξονα των X τοποθετούμε τις τιμές του x και στον άξονα των Y τις τιμές των καταλοίπων για κάθε τιμή του x αντίστοιχα. Η γραφική αυτή παράσταση ονομάζεται **διάγραμμα καταλοίπων** (*residuals plot*).



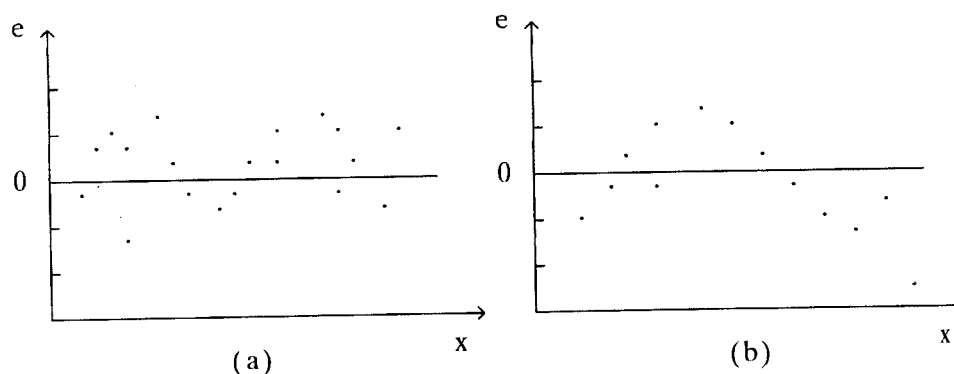
Διάγραμμα καταλοίπων για τις αποστάσεις φρεναρίσματος

Εξετάζοντας το γράφημα αυτό παρατηρούμε ότι η υπόθεση της γραμμικότητας δεν φαίνεται να επαληθεύεται. Αντίθετα, το συγκεκριμένο γράφημα οδηγεί στο συμπέρασμα ότι, ενδεχομένως, μια παραβολή θα εξέφραζε καλύτερα την αληθινή παλινδρόμηση που αντιστοιχεί στα δεδομένα από ότι μια ευθεία γραμμή. (Αυτό εξηγείται από το γεγονός ότι, όπως ξέρουμε από τη φυσική, η απόσταση που απαιτείται για να σταματήσει ένα όχημα και η αρχική ταχύτητα συνδέονται μέσω μιας εξίσωσης δευτέρου βαθμού).

Όπως είναι φυσικό η εξέταση με το μάτι ενός διαγράμματος των καταλοίπων αποτελεί μια υποκειμενική μέθοδο για τον έλεγχο της υπόθεσης της γραμμικότητας. Εάν έχουμε επαναλαμβανόμενες παρατηρήσεις σε μια ή περισσότερες τιμές του x μπορούμε να χρησιμοποιήσουμε ένα στατιστικό έλεγχο για να εξετάσουμε την

υπόθεση της γραμμικότητας. Ο έλεγχος αυτός ονομάζεται έλεγχος έλλειψης προσαρμοστικότητας (*lack-of-fit*) και στηρίζεται στην κατανομή F.

Η γραφική παράσταση των καταλοίπων μπορεί να χρησιμοποιηθεί όχι μόνο σαν μια αρχική τεχνική για τον έλεγχο της υπόθεσης της γραμμικότητας αλλά επίσης μπορεί να χρησιμοποιηθεί και για τον έλεγχο των υπολοίπων υποθέσεων. Για παράδειγμα, για να ισχύει η υπόθεση ότι τα y είναι ανεξάρτητα, (δοθέντος ότι η υπόθεση της γραμμικότητας ισχύει), θα πρέπει η γραφική παράσταση των καταλοίπων να δείχνει μια τυχαία τοποθέτηση των καταλοίπων αυτών γύρω από τη γραμμή που αντιστοιχεί σε κατάλοιπο "0". Εάν εμφανίζεται να υπάρχει μια προφανής έλλειψη τυχειότητας στο γράφημα των καταλοίπων, αυτό ίσως αποτελεί ένδειξη εξάρτησης μεταξύ των μεταβλητών y . Στο σχήμα (a) έχουμε μια περίπτωση όπου το διάγραμμα των καταλοίπων ενισχύει την υπόθεση της τυχειότητας ενώ το (b) αποτελεί ένδειξη έλλειψης τυχειότητας.



Διάγραμμα καταλοίπων για τον έλεγχο της ανεξαρτησίας

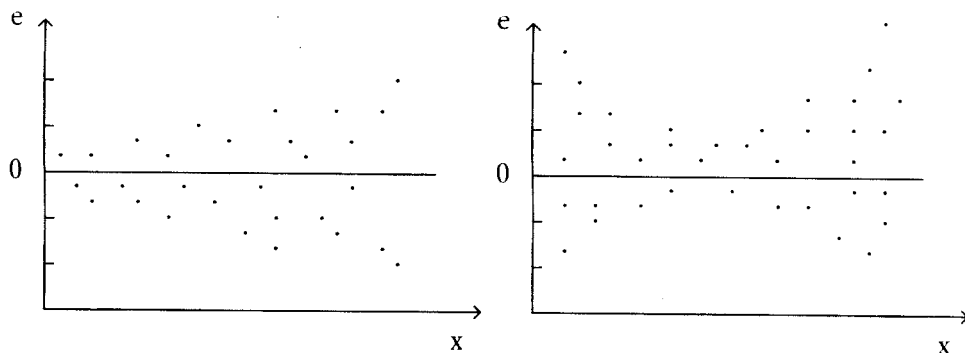
Βέβαια, ο γραφικός έλεγχος είναι πάντοτε υποκειμενικός. Ένας αντικειμενικός στατιστικός έλεγχος της υπόθεσης της ανεξαρτησίας μπορεί να γίνει με τη χρήση του **ελέγχου των ροών** (*runs test*).

Θα πρέπει να σημειώσουμε ότι η υπόθεση της ανεξαρτησίας θεωρείται, εν γένει, ως η πιο σημαντική προκειμένου να προχωρήσει κανείς σε συμπεράσματα μέσω της γραμμικής παλινδρόμησης. Ένας άλλος έλεγχος για την ανεξαρτησία των y μπορεί να γίνει μέσω του

ισοδύναμου ελέγχου ανεξαρτησίας των ε ο οποίος στηρίζεται στη μέθοδο της **σειριακής συσχέτισης** (*serial correlation*).

Μια άλλη υπόθεση που κάναμε στη γραμμική παλινδρόμηση είναι αυτή των ίσων διασπορών, ή καλύτερα της κοινής διασποράς. Στην πράξη, η υπόθεση αυτή πολλές φορές παραβιάζεται. Η υπόθεση των ίσων διασπορών ονομάζεται πολλές φορές υπόθεση της **ομοιογένειας των διασπορών** (*homogeneity of variances*) ή, εναλλακτικά, **ομοσκεδαστικότητα** (*homoscedasticity*). Η παραβίαση της ισότητας των διασπορών ονομάζεται επίσης **ανομοιογένεια των διασπορών** (*non-homogeneity of variances*) ή **ετεροσκεδαστικότητα** (*heteroscedasticity*).

Όταν υπάρχει ανισότητα διασπορών το "άπλωμα" του y τείνει, συχνά, να αυξάνεται όσο μεγαλύτερες είναι οι τιμές της μεταβλητής x . Περιπτώσεις όπου η διασπορά του y φαίνεται να εξαρτάται από το x αποτελούν αυτές που εμφανίζονται στο σχήμα που ακολουθεί.



Διάγραμμα καταλοίπων με διασπορά εξαρτώμενη από το x

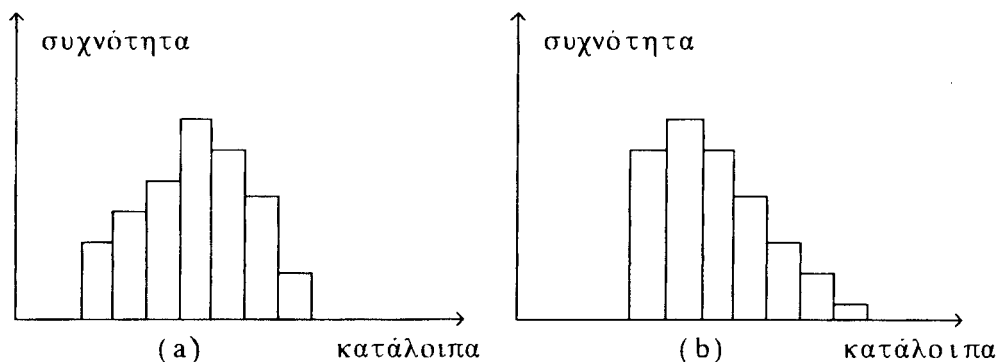
Επομένως, ένας τρόπος ελέγχου της ισότητας των διασπορών είναι να χωρίσουμε τα δειγματικά κατάλοιπα e σε δυο ομάδες η μία από τις οποίες θα προέρχεται από τις "μικρές" τιμές του x και η άλλη από τις "μεγάλες" τιμές του x . Στη συνέχεια υπολογίζουμε τις δειγματικές διασπορές S_1^2 και S_2^2 που αντιστοιχούν στις δύο ομάδες των καταλοίπων. Τέλος χρησιμοποιούμε τον έλεγχο F για την ισότητα δύο διασπορών ώστε να ελέγξουμε αν τα δύο σύνολα των καταλοίπων

έχουν ίσες διασπορές στους αντίστοιχους πληθυσμούς. Απόρριψη της μηδενικής υπόθεσης για ίσες διασπορές θα αποτελεί ισχυρή ένδειξη ότι η υπόθεση των ίσων διασπορών δεν ικανοποιείται στο συγκεκριμένο πρόβλημα. Θα πρέπει βέβαια να θυμόμαστε ότι ο έλεγχος F είναι πάρα πολύ ευαίσθητος στην υπόθεση της κανονικότητας. Επομένως, αν τα κατάλοιπα δεν ακολουθούν την κανονική κατανομή το αποτέλεσμα για τον έλεγχο F δεν μπορεί να είναι αξιόπιστο.

Τα κατάλοιπα περικλείουν το αποτέλεσμα που προκύπτει από την παράλειψη ανεξαρτήτων μεταβλητών στο μοντέλο. Όπως ήδη είπαμε, η γραμμική παλινδρόμηση είναι η απλούστερη μορφή παλινδρόμησης. Σε πολλά προβλήματα, όμως, μια ανεξάρτητη μεταβλητή δεν είναι αρκετή για να ερμηνεύσει τις μεταβολές της εξαρτημένης μεταβλητής. Έτσι, όταν, χάριν ευκολίας, χρησιμοποιούμε μια μόνο ανεξάρτητη μεταβλητή, το μέγεθος και η τυπική απόκλιση των μεταβλητών που έχουν παραληφθεί αλλά επηρεάζουν την εξαρτημένη μεταβλητή ίσως δεν είναι σταθερά. Συγκεκριμένα, όταν τα δεδομένα τα οποία προσπαθούμε να αναλύσουμε αναφέρονται σε μια χρονολογική σειρά, το μέγεθος και η τυπική απόκλιση των μεταβλητών που έχουν παραληφθεί είναι ενδεχόμενο να αυξάνονται με το χρόνο και να δημιουργούν πρόβλημα ετεροσκεδαστικότητας. Η λύση στο πρόβλημα αυτό είναι να περιλαμβάνονται στο μοντέλο οι μεταβλητές αυτές και να χρησιμοποιείται στη συνέχεια η μέθοδος της πολλαπλής παλινδρόμησης που θα δούμε αργότερα. Μια δεύτερη λύση είναι να επανακαθορίσουμε το μοντέλο, επαναπροσδιορίζοντας τις κλίμακες των μεταβλητών, έτσι ώστε τα κατάλοιπα να έχουν σταθερή τυπική απόκλιση.

Αν όλες οι προηγούμενες υποθέσεις ικανοποιούνται η τελευταία απαραίτητη προϋπόθεση για τη χρησιμοποίηση της ευθείας παλινδρόμησης για στατιστική συμπερασματολογία είναι η υπόθεση ότι τα λάθη (και επομένως, αντίστοιχα, τα y) ακολουθούν κανονική κατανομή. Μια πρώτη προσέγγιση για τον έλεγχο της κανονικότητας των y επιτυγχάνεται και πάλι με τη χρησιμοποίηση των καταλοίπων. Στην περίπτωση αυτή κατασκευάζουμε ιστόγραμμα των καταλοίπων.

Το ιστόγραμμα αυτό μας δίνει μια ένδειξη για το κατά πόσο η υπόθεση της κανονικότητας ισχύει ή όχι. Για παράδειγμα, το ιστόγραμμα (a) του σχήματος που ακολουθεί αποτελεί ένδειξη ότι η υπόθεση της κανονικότητας είναι λογική, ενώ αντίθετα το ιστόγραμμα (b) είναι ένδειξη ότι η υπόθεση αυτή δεν ισχύει. Στατιστικός έλεγχος για την υπόθεση της κανονικότητας μπορεί να γίνει με έναν από τους **ελέγχους καλής προσαρμογής** (*goodness-of-fit*). Άλλος στατιστικός έλεγχος είναι ο έλεγχος Lilliefors.



Ιστόγραμμα Καταλοίπων

Ο έλεγχος αυτός είναι προσεγγιστικός γιατί τα δειγματικά κατάλοιπα $e_1, e_2, \dots, e_n$ δεν είναι ανεξάρτητα μεταξύ τους. Ο έλεγχος του Lilliefors επιτυγχάνεται με κατασκευή του διαγράμματος της εμπειρικής συνάρτησης κατανομής των καταλοίπων τα οποία τυποποιούνται ως ακολούθως:

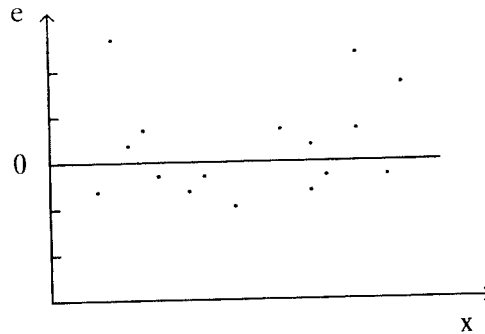
$$e_i^* = e_i / \hat{\sigma}$$

(Για περισσότερες λεπτομέρειες της μεθόδου βλέπε Ι. Πανάρετος: *Εκτιμητική-Ελεγχοι Υποθέσεων*. Αθήνα 1993).

Εάν τα κατάλοιπα δεν κατανέμονται σύμφωνα με την κανονική κατανομή θα πρέπει να χρησιμοποιηθεί ένας άλλος στατιστικός έλεγχος.

Σημείωση: Μια άλλη περίπτωση που χρησιμοποιούνται τα κατάλοιπα είναι στον έλεγχο των **ακραίων τιμών** (*outliers*). Ως ακραία τιμή

χαρακτηρίζουμε μια παρατήρηση η οποία βρίσκεται απροσδόκητα μακριά από τις άλλες παρατηρήσεις. Στο σχήμα που ακολουθεί φαίνονται δύο τέτοιες ακραίες τιμές.



Διάγραμμα καταλοίπων με δύο ακραίες τιμές

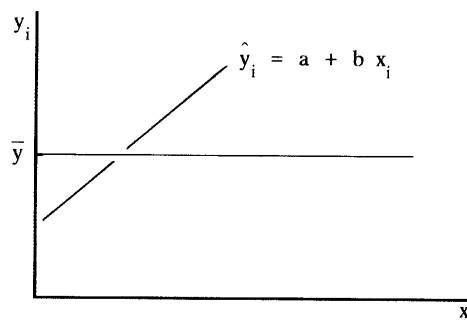
Όταν στο δείγμα μας περιλαμβάνεται μια τέτοια ακραία τιμή θα πρέπει να ελέγξουμε με προσοχή τα αρχικά δεδομένα για να δούμε αν πράγματι η τιμή αυτή αντιπροσωπεύει ένα πραγματικό σημείο του δείγματος. Αυτό γιατί υπάρχει ενδεχόμενο να έχει γίνει κάποιο λάθος στη διεξαγωγή του πειράματος που να μην είναι δειγματικό λάθος. (Για παράδειγμα μπορεί αυτός που καταγράφει τις τιμές να καταγράψει μια τιμή λανθασμένα). Σε μια τέτοια περίπτωση το λάθος θα πρέπει να διορθωθεί ή η τιμή αυτή θα πρέπει να διαγραφεί από τα δεδομένα. Αντίθετα, αν δεν υπάρχει εξήγηση της ακραίας αυτής τιμής που να προκύπτει από τον τρόπο καταγραφής του πειράματος είναι λάθος να αλλάξουμε ή να απαλείψουμε την τιμή αυτή. Κάτι τέτοιο θα ακυρώσει την οποιαδήποτε στατιστική συμπερασματολογία κάνουμε στη συνέχεια.

Καταλήγοντας, θα πρέπει να επαναλάβουμε ότι το σύνολο των δειγματικών καταλοίπων $e_1, e_2, \dots, e_n$ δεν είναι ανεξάρτητα. Αυτό προκύπτει από το γεγονός ότι το άθροισμα των καταλοίπων θα πρέπει να είναι ίσο με το μηδέν. Δηλαδή: $\sum e_i = 0$.

Παρόλα αυτά, εάν τα y είναι μεταξύ τους ανεξάρτητα και αν το μέγεθος του δείγματος είναι σχετικά μεγάλο η εξάρτηση των e είναι μικρή.

1.5 Συντελεστής Προσδιορισμού (Coefficient of Determination)

Μέχρι τώρα ασχοληθήκαμε με την κατασκευή της ευθείας παλινδρόμησης και τη σχετική στατιστική συμπερασματολογία για τους συντελεστές, την $\mu_{Y|x}$ και το Y . Δεν έχουμε εξετάσει το πόσο καλά η ευθεία παλινδρόμησης εξηγεί τα δεδομένα του δείγματος. Για το σκοπό αυτό χρειάζεται ένα κριτήριο που να καθορίζει πόσο καλά μια γραμμή παλινδρόμησης προσαρμόζει (εξηγεί) τα δεδομένα. Ποιό δηλαδή είναι το ποσοστό της συνολικής διασποράς του Y που είναι δυνατόν να αποδοθεί (να εξηγηθεί) στη γνώση του x . Γενικά, έχουμε ένα δείγμα από n ζευγάρια παρατηρήσεων $\{(x_i, y_i): i=1,2,\dots,n\}$ στις οποίες προσαρμόσαμε, με τη μέθοδο των ελαχίστων τετραγώνων την ευθεία $\hat{Y} = a + b x$



Σχήμα 1.5.1

Ας θεωρήσουμε το μέσο $\bar{y}$ των τιμών y_i ($\bar{y} = 1/n \sum y_i$). Η συνολική απόκλιση των παρατηρήσεων y από το $\bar{y}$ μπορεί να εκφραστεί από το άθροισμα $\sum (y_i - \bar{y})^2 = \Sigma AT \equiv SAT$ (συνολικό άθροισμα τετραγώνων, *total sum of squares*). Είναι φανερό ότι

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Η σχέση αυτή δείχνει ότι, για κάθε i η απόκλιση της παρατηρηθείσας τιμής από το μέσο $\bar{y}$ μπορεί να γραφεί σαν το άθροισμα της διαφοράς της παρατηρηθείσας τιμής από την τιμή που παίρνουμε από την ευθεία παλινδρόμησης και της διαφοράς της τιμής αυτής από το μέσο $\bar{y}$. Παίρνοντας τα τετράγωνα και των δύο μελών και αθροίζοντας ως προς i βρίσκουμε

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

[διότι εύκολα επιβεβαιώνεται ότι $\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$]

Δηλαδή τελικά

$$\Sigma AT = \Lambda AT + \Pi AT \quad (34)$$

όπου $\Lambda AT = \sum (y_i - \hat{y}_i)^2 \equiv \text{SSE}$ (άθροισμα τετραγώνων των λαθών,
error sum of squares)

και $\Pi AT = \sum (\hat{y}_i - \bar{y})^2 \equiv \text{SSR}$ (άθροισμα τετραγώνων παλινδρόμησης,
regression sum of squares).

Η σχέση (34) δείχνει ότι το άθροισμα των τετραγώνων ΣAT μπορεί να αποδοθεί σε δύο παράγοντες. Συνεπώς, το ποσοστό της διασποράς του y γύρω από το $\bar{y}$ που μπορεί να εξηγήσει από την παλινδρόμηση πάνω στο x είναι το

$$\frac{\Pi AT}{\Sigma AT} \quad (\text{διότι } \frac{\Pi AT}{\Sigma AT} + \frac{\Lambda AT}{\Sigma AT} = 1)$$

Από εδώ βγαίνει το συμπέρασμα ότι μια καλή προσαρμογή της ευθείας παλινδρόμησης συνεπάγεται $\Pi AT / \Sigma AT \rightarrow 1$ (διότι το ΛAT είναι ασήμαντο).

Αντίθετα, αν η ευθεία γραμμικής παλινδρόμησης δεν δίνει καλή προσαρμογή στα δεδομένα ο λόγος $\Pi AT / \Sigma AT$ θα έχει τιμή κοντά στο μηδέν (γιατί τότε μόνο ένα μικρό ποσοστό της συνολικής διασποράς των τιμών του y θα μπορεί να αποδοθεί στην ευθεία παλινδρόμησης).

Βρήκαμε λοιπόν μια συνάρτηση των παρατηρήσεων $\{(x_i, y_i), i=1,2,\dots,n\}$ που μπορεί να χρησιμοποιηθεί σαν μέτρο καταλληλότητας της ευθείας παλινδρόμησης. Για τη συνάρτηση αυτή $\Pi AT / \Sigma AT$, που συνήθως τη συμβολίζουμε με r^2 , έχουμε

$$r^2 = \frac{\text{ΠΑΤ}}{\text{ΣΑΤ}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (a + bx_i - (a + b\bar{x}))^2}{S_{yy}}$$

$$= \frac{b^2 \sum (x_i - \bar{x})^2}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

Δηλαδή, τελικά

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} \quad (35)$$

Ο συντελεστής r^2 λέγεται **συντελεστής προσδιορισμού** (*coefficient of determination*). Είναι φανερό ότι

$$0 \leq r^2 \leq 1$$

Μεγάλες τιμές του r^2 αποτελούν ένδειξη καλής προσαρμογής της ευθείας παλινδρόμησης στα δεδομένα του δείγματος. Αντίθετα, τιμές του r^2 κοντά στο 0 αποτελούν ένδειξη κακής προσαρμογής.

Σημείωση: Πολλοί ερευνητές θεωρούν ότι ο συντελεστής προσδιορισμού, όπως ορίστηκε προηγουμένως, πρέπει να παίρνει υπόψη του τον αριθμό των παρατηρήσεων στο δείγμα σε σχέση με τον αριθμό των ανεξαρτήτων μεταβλητών που χρησιμοποιούνται. Αυτό έχει ιδιαίτερη σημασία στην περίπτωση της πολλαπλής παλινδρόμησης που θα δούμε αργότερα. Για το λόγο αυτό ορίζεται ο **συντελεστής προσδιορισμού προσαρμοσμένος για τους βαθμούς ελευθερίας** (*coefficient of determination adjusted for degrees of freedom*).

Στην απλή γραμμική παλινδρόμηση που εξετάζουμε εδώ, ο συντελεστής αυτός έχει τη μορφή

$$r_{\text{adj}}^2 = 1 - \left[(1 - r^2) \frac{n-1}{n-2} \right].$$

Την τιμή του συντελεστή αυτού δίνουν πολλά από τα στατιστικά πακέτα που κυκλοφορούν στην αγορά.

Σημείωση: Είναι δυνατόν να γίνει χρήση της θεωρίας υποθέσεων και να κατασκευαστεί έλεγχος υποθέσεων που να καθορίζει, βάσει των πληροφοριών του δείγματος αν η υπόθεση της γραμμικής παλινδρόμησης είναι βάσιμη ή όχι. Ο έλεγχος αυτός βασίζεται στην ανάλυση διασποράς και την κατανομή F.

Στο παράδειγμα που ακολουθεί δίνεται η ευκαιρία να διαπιστώσουμε την σημασία που έχει για την εφαρμογή της μεθόδου της γραμμικής παλινδρόμησης η ικανοποίηση των υποθέσεων που πρέπει να ισχύουν.

Παράδειγμα: Μια εκδοτική εταιρία ενδιαφέρεται να μελετήσει την μεταβλητότητα που παρατηρείται στις ετήσιες πωλήσεις ενός από τα εβδομαδιαία περιοδικά που εκδίδει. Θεωρώντας ότι ο συνολικός πληθυσμός Y των πωλήσεων σε μία πόλη εξαρτάται κυρίως από το συνολικό εισόδημα που έχουν οι κάτοικοι της πόλης αυτής, η διεύθυνση του περιοδικού συγκέντρωσε τα ακόλουθα στοιχεία από ένα δείγμα 10 πόλεων.

Πόλη	Συνολικό εισόδημα X (σε εκατοντ. Εκατ. δρχ.)	Συνολικός αριθμός πωλήσεων Y ανά έτος (σε χιλ. τεύχη)
1	620	964
2	340	462
3	7400	10440
4	3000	4200
5	1050	1310
6	2100	2610
7	4300	5220
8	660	948
9	2300	2830
10	1200	1560

Η εκτιμώμενη ευθεία παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων για τα δεδομένα αυτά είναι η

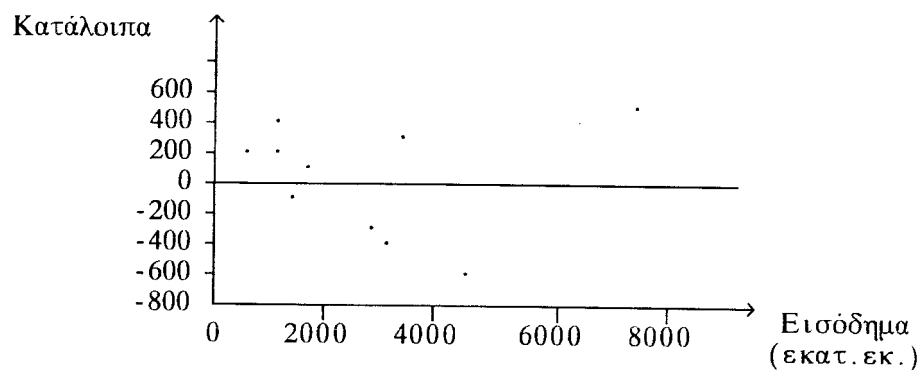
$$\hat{\mu}_{Y|x} \equiv \hat{y} = -109.1 + 1.377x$$

Οι τιμές των στατιστικών συναρτήσεων T για τις μηδενικές υποθέσεις $\alpha=0$ και $\beta=0$ είναι, αντίστοιχα 0.8 και 31.0. Τέλος, η τιμή του συντελεστή προσδιορισμού r^2 και $r^2 = 0.992$.

Τα αποτελέσματα αυτά είναι πολύ ικανοποιητικά. Η συσχέτιση μεταξύ εισοδήματος και πωλήσεων είναι, όπως αναμενόταν, θετική και όπως θα ήλπιζε ο εκδότης, στατιστικά σημαντικά. Επίσης, η τιμή του r^2 είναι σχεδόν 1, γεγονός που θα μπορούσε να οδηγήσει στο συμπέρασμα ότι η προσαρμογή είναι πάρα πολύ ικανοποιητική.

Στην πραγματικότητα όμως, τα πράγματα δεν είναι ακριβώς έτσι. Και αυτό γιατί οι υποθέσεις που είναι απαραίτητες για την εφαρμογή της μεθόδου παλινδρόμησης δεν ισχύουν στη συγκεκριμένη περίπτωση. Συγκεκριμένα, αν κατασκευάσουμε το διάγραμμα των

καταλοίπων, με το εισόδημα στον οριζόντιο άξονα, βλέπουμε, στο σχήμα που ακολουθεί, ότι τα κατάλοιπα για πόλεις με μεγαλύτερα εισοδήματα έχουν μεγαλύτερες τιμές από τα κατάλοιπα για πόλεις με χαμηλότερα εισοδήματα. Μια τέτοια σχέση είναι, βέβαια, φυσιολογική. Είναι λογικό να περιμένει κανείς ότι τα λάθη στην πρόβλεψη των συνολικών πωλήσεων του περιοδικού σε μία μεγάλη πόλη θα είναι μεγαλύτερα από τα αντίστοιχα λάθη των συνολικών πωλήσεων σε μία μικρή πόλη. Αν κάτι τέτοιο συμβαίνει, τότε και η τυπική απόκλιση είναι επίσης μεγαλύτερη για τις μεγαλύτερες πόλεις από ότι για τις μικρότερες με αποτέλεσμα η υπόθεση ότι η τυπική απόκλιση για τα λάθη είναι σταθερή να μην ικανοποιείται. Είναι επομένως, προφανές ότι στη συγκεκριμένη περίπτωση έχουμε πρόβλημα ετεροσκεδαστικότητας όπως φαίνεται και από το σχήμα.



Είναι εμφανές ότι το πρόβλημα της ετεροσκεδαστικότητας προκαλείται από το γεγονός ότι μερικές πόλεις έχουν πολύ μεγαλύτερους πληθυσμούς από άλλες. Ένας από τους τρόπους, όπως προαναφέρθηκε, για να υπερπηδηθεί το πρόβλημα αυτό είναι ο επανακαθορισμός του μοντέλου. Για να εκφράσουμε τα δεδομένα σε μια συγκρίσιμη κλίμακα μπορούμε να επανακαθορίσουμε το μοντέλο ώστε να εκφράζεται σε κατά κεφαλή εισόδημα και σε κατά κεφαλή πωλήσεις περιοδικών με την χρησιμοποίηση των παρακάτω δεδομένων:

Πόλη	Πληθ. P (σε χιλ.)	Κατά κεφαλή εισόδημα $X^*=X/P$	Κατά κεφαλή πωλήσεις $Y^*=Y/P$
1	40	15.5	24.1
2	20	17.0	23.1
3	400	18.5	26.1
4	150	20.0	28.0
5	50	21.0	26.2
6	100	21.0	26.1
7	200	21.5	26.1
8	30	22.0	31.6
9	100	23.0	28.3
10	50	24.0	31.2

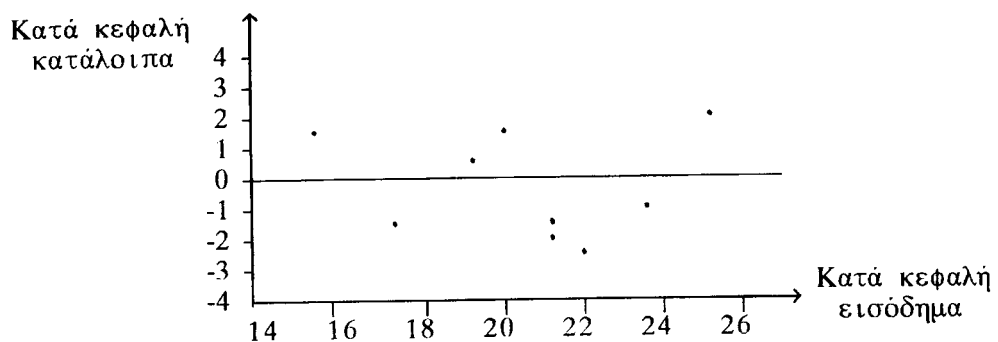
Η ευθεία παλινδρόμησης που προκύπτει με την μέθοδο των ελαχίστων τετραγώνων για τα δεδομένα αυτά είναι η

$$\hat{\mu}_{Y^*|X^*} \equiv \hat{Y}^* = 10.207 + 0.829x^*$$

Οι τιμές της στατιστικής συνάρτησης T κάτω από τις μηδενικές υποθέσεις $\alpha = 0$ και $\beta = 0$ είναι, αντίστοιχα 2.3 και 3.8. Τέλος, ο συντελεστής προσδιορισμού είναι $r^2 = 0.641$.

Η γραφική παράσταση των καταλοίπων στο σχήμα που ακολουθεί δεν δίνει πια κάποια ένδειξη ύπαρξης ετεροσκεδαστικότητας. Είναι επομένως, προφανές ότι το πρόβλημα της ετεροσκεδαστικότητας οφειλόταν στην χρήση των συνολικών δεδομένων για πόλεις με πολύ διαφορετικούς πληθυσμούς και αντιμετωπίσθηκε με την χρησιμοποίηση κλίμακας που αναφερόταν σε δεδομένα κατά κεφαλή. Είναι ενδιαφέρον να παρατηρήσουμε ότι η μετατροπή της κλίμακας επέφερε μία σημαντική μεταβολή στην εκτιμώμενη κλίση της σχέσης. Η αρχική παλινδρόμηση οδηγούσε στο συμπέρασμα ότι η αύξηση στο συνολικό εισόδημα μιας πόλης κατά μια εκατοντάδα εκατομμύρια δραχμές θα επέφερε αύξηση στις ετήσιες πωλήσεις περιοδικών κατά 1.377 χιλιάδες τεύχη (Ισοδύναμα, αύξηση του συνολικού ετησίου εισοδήματος κατά 100 χιλ. δρχ. θα

επέφερε αύξηση των συνολικών πωλήσεων στο έτος αυτό κατά 1.377 τεύχη). Η παλινδρόμηση που αναφέρεται στη νέα κλίμακα δίνει μία εκτίμηση που είναι 40% χαμηλότερη: μια αύξηση πωλήσεων κατά 0.829 περίπου περιοδικά ανά 100 χιλιάδες δραχμές συνολικού εισοδήματος. Αν, επομένως, οι εκτιμήσεις που βασίζονται στη νέα κλίμακα είναι σωστές, οι αρχικές εκτιμήσεις οδηγούνται σε πολύ λανθασμένα συμπεράσματα.



Σχήμα 1.5.3

Κάποιοι αναλυτές ενδεχομένως θα προβληματιζόταν από την παρατήρηση ότι η μετατροπή της κλίμακας οδήγησε σε μία σημαντική πτώση της τιμής της στατιστικής συνάρτησης ελέγχου T για το β , όπως επίσης και σε σημαντική ελάττωση της τιμής του συντελεστή προσδιορισμού r^2 , κάτι που δεν θα μπορούσε να θεωρηθεί με κανένα τρόπο ως βέλτιστη. Η απάντηση στην παρατήρηση αυτή είναι ότι δεν είναι δυνατόν να συγκρίνουμε αυτά τα στατιστικά μέτρα απευθείας, δεδομένου ότι οι επεξηγηματικές μεταβλητές στα δύο μοντέλα (οι ανεξάρτητες μεταβλητές) είναι διαφορετικές. Η πρώτη τιμή του r^2 , για παράδειγμα μετρά την ακρίβεια στην πρόβλεψη των συνολικών πωλήσεων. Η δεύτερη τιμή του r^2 μετρά την ακρίβεια στην πρόβλεψη των κατά κεφαλή πωλήσεων. Αν, όπως συμβαίνει εδώ, οι συνολικές πωλήσεις έχουν μια μεγαλύτερη μεταβλητότητα από ότι οι κατά κεφαλή πωλήσεις, τότε το κατά κεφαλή μοντέλο είναι δυνατόν να είναι πιο ακριβές και παρόλα αυτά να έχει χαμηλότερη τιμή για το r^2 !

Αν θέλουμε να συγκρίνουμε την προβλεπτική ικανότητα των δύο αυτών μοντέλων, θα πρέπει να συγκεντρώσουμε την προσοχή μας σε μια συγκεκριμένη μεταβλητή: ή αυτή που αναφέρεται σε συνολικές πωλήσεις ή αυτή που αναφέρεται σε κατά κεφαλή πωλήσεις. Για το μοντέλο που αναφέρεται σε συνολικούς αριθμούς, οι προβλέψεις δίνονται από την εξίσωση παλινδρόμησης. Για το μοντέλο που αναφέρεται σε κατά κεφαλή στοιχεία, μπορούμε να πολλαπλασιάσουμε τις προβλεπόμενες κατά κεφαλή πωλήσεις με τον πληθυσμό της πόλης: Έτσι θα έχουμε

$$\text{“αθροιστικό μοντέλο”} : \hat{y} = -109.1 + 1377x$$

$$\begin{aligned} \text{“κατά κεφαλή μοντέλο”} : \hat{y} &= P(10.207 + 0.829x^*) \\ &= P10.207 + 0.829x. \end{aligned}$$

Ο πίνακας που ακολουθεί δίνει τις προβλέψεις αυτές.

Αθροιστικές προβλέψεις για το “αθροιστικό” και το “κατά κεφαλή” μοντέλο

Πραγματικές πωλήσεις	“Αθροιστικό” μοντέλο		“Κατά κεφαλή” μοντέλο	
	Πρόβλεψη $\hat{Y}$	Λάθος	Πρόβλεψη $\hat{Y}$	Λάθος
964	744.8	219.2	922.3	41.7
462	359.2	102.8	486.0	-24.0
10440	10082.2	357.8	10218.3	221.7
4200	4022.5	177.5	4018.4	181.6
1310	1337.0	-27.0	1380.9	-70.9
2610	2783.1	-173.1	2761.9	-151.9
5220	5812.9	-592.9	5606.6	-386.6
948	799.9	148.1	853.4	94.6
2830	3058.5	-228.5	2927.7	-97.7
1560	1543.6	16.4	1505.3	54.7

Από τον πίνακα αυτό παρατηρούμε ότι το “κατά κεφαλή” μοντέλο έχει, εν γένει, μικρότερα λάθη πρόβλεψης. Τα λάθη αυτά πολλές φορές είναι σημαντικά μικρότερα από ότι του “αθροιστικού”

μοντέλου. Η τυπική απόκλιση της εκτίμησης (η τετραγωνική ρίζα του αθροίσματος των τετραγώνων των λαθών πρόβλεψης διαιρεμένα με $n-2$) είναι 290.4 για το “αθροιστικό” μοντέλο και 1883. Για το “κατά κεφαλή” μοντέλο.

Το παράδειγμα αυτό παρουσιάζει ενδιαφέρον όχι μόνο γιατί δίνει μια πιθανή τεχνική για την αντιμετώπιση του προβλήματος της ετεροσκεδαστικότητας, αλλά επίσης γιατί επισημαίνει τους κινδύνους που περιέχονται στην απόδοση υπερβολικής σημασίας στην τιμή της στατιστικής συνάρτησης T και του συντελεστή προσδιορισμού r^2 για την αξιολόγηση της ποιότητας του μοντέλου. Ένα πετυχημένο μοντέλο θα πρέπει να οδηγεί σε μια λογικοφανή εκτίμηση για το β , ακριβείς σχετικά προβλέψεις, αλλά και κατάλοιπα που να ικανοποιούν τις υποθέσεις του μοντέλου.

1.6 Συντελεστής Συσχέτισης (Correlation Coefficient)

Μέχρι τώρα στη θεωρία της γραμμικής παλινδρόμησης έχουμε θεωρήσει δύο εξαρτημένες μεταβλητές X και Y . Υποθέτοντας ότι η ανεξάρτητη μεταβλητή X παίρνει τιμές που καθορίζονται από τον ερευνητή μελετήσαμε τη συμπεριφορά της τυχαίας μεταβλητής $Y|x$. Ας υποθέσουμε τώρα ότι δεν είναι εύκολο να ελέγξουμε (να προκαθορίσουμε) τις τιμές της μεταβλητής X (X και Y δηλαδή είναι τυχαίες μεταβλητές). Τότε, το διάνυσμα (X,Y) έχει μια κατανομή που εν γένει, είναι άγνωστη. Παρόλα αυτά είναι δυνατόν να μελετήσουμε μια παράμετρο της κατανομής αυτής που μπορεί να χρησιμοποιηθεί σαν μέτρο της ισχυρότητας της γραμμικής εξάρτησης των X και Y . Η παράμετρος αυτή είναι ο γνωστός **συντελεστής συσχέτισης** (*correlation coefficient*)

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} \quad (36)$$

όπου $\text{Cov}(X,Y) = E[(X - \mu_x)(Y - \mu_y)]$ (συνδιακύμανση των X και Y) και μ_x , μ_y , σ_x , σ_y οι μέσες τιμές και οι τυπικές αποκλίσεις των κατανομών των X και Y αντίστοιχα.

Όπως είναι γνωστό

i) $-1 \leq \rho \leq 1$

ii) $\rho = \pm 1 \Rightarrow X, Y$ είναι γραμμικά εξαρτημένα.

Θα πρέπει να τονισθεί ότι το ρ είναι μέτρο **γραμμικής εξάρτησης μόνο** και όχι μέτρο εξάρτησης οποιασδήποτε μορφής. (Αν δηλαδή $\rho = 0$ αυτό δεν συνεπάγεται ότι τα X και Y είναι ανεξάρτητα. Είναι μεν ασυσχέτιστα αλλά μπορεί να έχουν άλλη μορφή εξάρτησης).

Μια εκτιμήτρια του ρ μπορεί να κατασκευασθεί αν κανείς θεωρήσει ένα τυχαίο δείγμα $\{(X_i, Y_i): i=1,2,\dots,n\}$ από τη διμεταβλητή κατανομή του (X,Y) . Η εκτιμήτρια αυτή λέγεται **δειγματικός συντελεστής συσχέτισης** (*sample correlation coefficient*) και δίνεται από τον τύπο

$$\hat{\rho} = R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}$$

και για ένα συγκεκριμένο δείγμα (x_i, y_i) , $i=1,2,\dots,n$ η τιμή r της εκτιμήτριας R είναι

$$\begin{aligned} r &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \\ &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{[\sum x_i^2 - n\bar{x}^2]} \sqrt{[\sum y_i^2 - n\bar{y}^2]}} \quad (37) \end{aligned}$$

ή, χρησιμοποιώντας το γνωστό συμβολισμό,

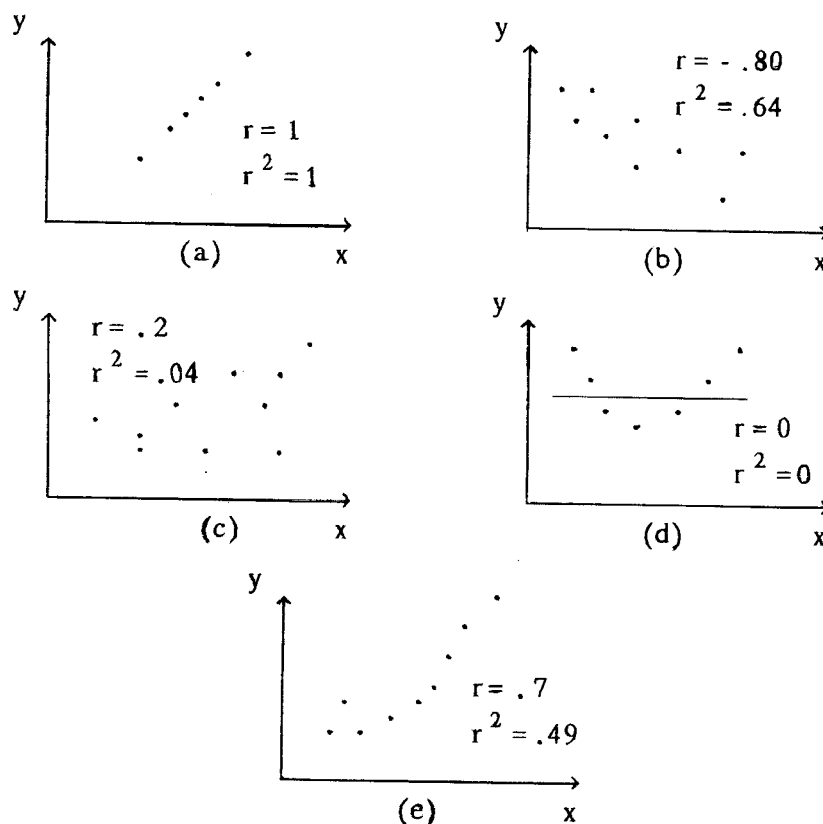
$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}} \quad (38)$$

Προφανώς $r \leq 1$.

Σημείωση: Συγκρίνοντας τις σχέσεις (35) και (38) βλέπουμε ότι ο συντελεστής γραμμικής συσχέτισης r του δείγματος, μπορεί να θεωρηθεί σαν η θετική τετραγωνική ρίζα του συντελεστή προσδιορισμού r^2 . Χρειάζεται όμως προσοχή στη χρησιμοποίηση του

r και του r^2 . Το r^2 μπορεί να χρησιμοποιηθεί σαν μέτρο ισχυρότητας της γραμμικής σχέσης ανεξάρτητα από το αν το x παίρνει καθορισμένες τιμές ή αν είναι τυχαία μεταβλητή. Αντίθετα, το r μπορεί να χρησιμοποιηθεί μόνο αν το Y και το X είναι τυχαίες μεταβλητές.

Στα σχήματα που ακολουθούν δίνονται οι ευθείες παλινδρόμησης και οι αντίστοιχες τιμές του r και του r^2 για μια σειρά από διαφορετικές περιπτώσεις.



Διαγράμματα διασποράς και οι αντίστοιχες τιμές r & r^2