Κεφάλαιο 2

Απλό Γραμμικό Μοντέλο

2.1 Εισαγωγή

Όπως είδαμε στο Κεφάλαιο 1, η βασική μέθοδος που ακολουθείται για την ανάλυση ενός στατιστικού μοντέλου ονομάζεται **Ανάλυση Παλινδρόμησης** (Regression Analysis) και έχει ευρείες εφαρμογές λόγω της χρησιμότητάς της.

Το **απλό γραμμικό μοντέλο** περιλαμβάνει δύο μεταβλητές, την **ανεξάρτητη** ή **προβλέπουσα** ή αλλιώς την **επεξηγηματική** μεταβλητή x και την **εξαρτημένη** ή διαφορετικά τη **μεταβλητή απόκρισης** y, οι οποίες συνδέονται μεταξύ τους με τη **γραμμική συνάρτηση παλινδρόμησης**. Σκοπός μας είναι η προσαρμογή μιας ευθείας γραμμής, η οποία επεξηγεί όσο το δυνατόν καλύτερα τη συμπεριφορά των δεδομένων μας. Μια τέτοια ευθεία θα έχει τη μορφή

$$E(y|x) = E(y_x) = \beta_0 + \beta_1 x = \mu_x \,, \tag{2.1}$$

όπου τα β_0 και β_1 αποτελούν τις **παραμέτρους του μοντέλου** ή αλλιώς τους **συντελεστές παλινδρόμησης**. Η παραπάνω σχέση είναι μια στατιστική σχέση που περιγράφει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής y_x, όταν

η ανεξάρτητη μεταβλητή πάρει την τιμή x. Η μεταβλητή y θεωρείται ότι είναι μια τυχαία μεταβλητή, ενώ αντιθέτως η x θεωρείται μη στοχαστική.

Η προσαρμογή της καλύτερης ευθείας, δηλαδή η καλύτερη δυνατή εκτίμηση των παραμέτρων, γίνεται λαμβάνοντας υπόψη τις n ανεξάρτητες παρατηρήσεις (x_i, y_i), $i = 1, \ldots, n$, που έχουμε στη διάθεσή μας προς επεξεργασία, για τις οποίες υποθέτουμε ότι δεν υπόκεινται σε σφάλματα μέτρησης.

Τα σημεία (x_i, y_i), $i = 1, \ldots, n$, είναι πιθανόν να διαφέρουν από τα σημεία $(x_i, \hat{y}_i)$, όπου

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

είναι η **εκτίμηση της τιμής της τυχαίας μεταβλητής** y με βάση το απλό γραμμικό μοντέλο, που προσαρμόσαμε στα δεδομένα μας, και $\hat{\beta}_0$, $\hat{\beta}_1$ οι εκτιμήσεις των παραμέτρων του μοντέλου (ο τρόπος υπολογισμού τους παρουσιάζεται παρακάτω).

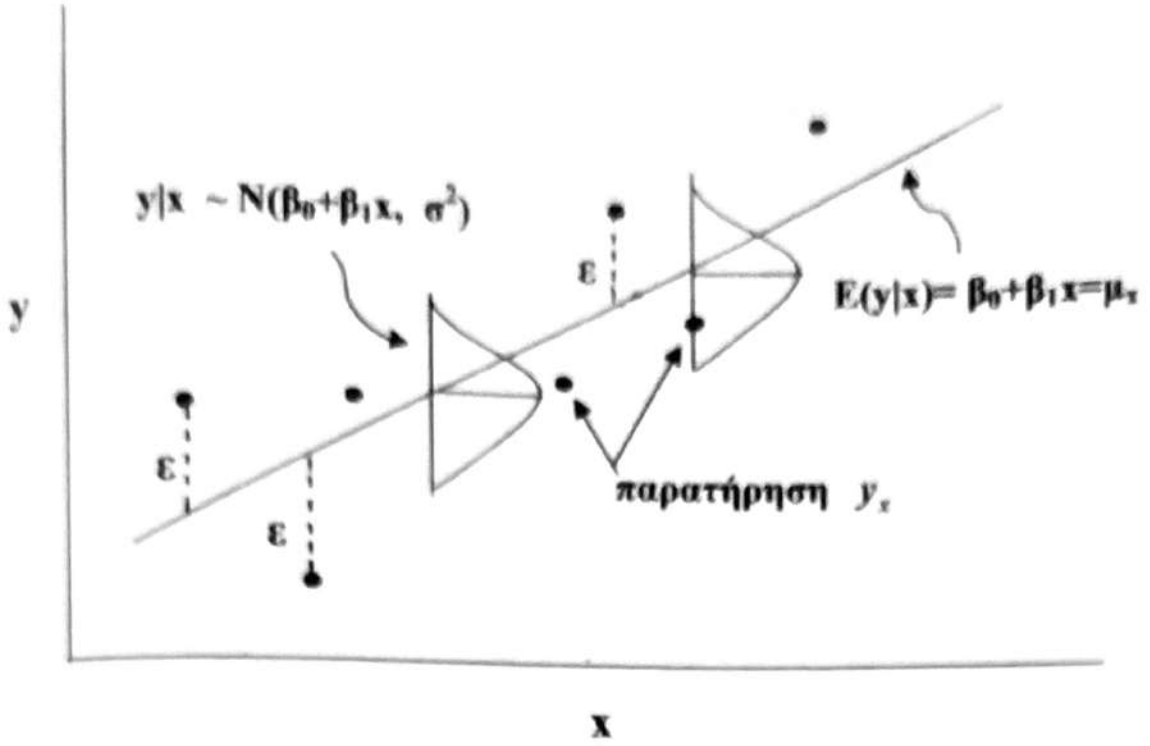
Οι παρατηρήσεις y_i δίνονται από τη σχέση

$$\begin{aligned} y_i =& E(y_{x_i}) + \varepsilon_i \\ =& \beta_0 + \beta_1 x_i + \varepsilon_i \,, \end{aligned}$$

όπου $E(y_i) = E(y_{x_i})$. Το ε_i ονομάζεται **τυχαίο σφάλμα** και παριστάνει για δοθείσα τιμή x_i την (άγνωστη) κατακόρυφη απόκλιση της τιμής y_i από την ευθεία της συνάρτησης παλινδρόμησης, που φυσικά είναι ακόμα άγνωστη (βλ. Σχήμα 2.1). Το τυχαίο σφάλμα δεν πρέπει να συγχέεται με τη διαφορά

$$\begin{aligned} e_i =& y_i - \hat{y}_i \\ =& y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \,, \end{aligned}$$

η οποία αποτελεί την κατακόρυφη απόκλιση του y_i από την ευθεία της **εκτιμημένης ή προσαρμοσμένης** συνάρτησης παλινδρόμησης και ονομάζεται **υπόλοιπο ή κατάλοιπο** (residual). Τα e_i μπορούν να θεωρηθούν ως οι **εκτιμήσεις** των άγνωστων τυχαίων σφαλμάτων ε_i.



Σχήμα 2.1: Το διάγραμμα διασποράς ενός δείγματος τιμών (x_i, y_i) και η γραμμική παλινδρόμηση

Η προσαρμογή του μοντέλου (2.1), δηλαδή η εκτίμηση των παραμέτρων β_0 και β_1 του απλού γραμμικού μοντέλου, μπορεί να γίνει με τη **μέθοδο ελαχίστων τετραγώνων** (ordinary least squares – OLS), η οποία παρουσιάζεται στη συνέχεια.

2.2 Μέθοδος ελαχίστων τετραγώνων

Η μέθοδος ελαχίστων τετραγώνων έγκειται στην ελαχιστοποίηση της παράστασης

$$\begin{aligned} S(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i - E(y_i))^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

ως προς τα β_0 και β_1, δηλαδή στην ελαχιστοποίηση του αθροίσματος των τετραγώνων των αποκλίσεων μεταξύ των y_i και $\hat{y}_i$. Ακολουθώντας τη συνηθισμένη διαδικασία για τον εντοπισμό ελαχίστων τιμών, δηλαδή παραγωγίζοντας το $S(\beta_0, \beta_1)$ ως προς β_0 και ως προς β_1 και θέτοντας στη συνέχεια τις μερικές αυτές παραγώγους ίσες με μηδέν σε ένα σημείο $(\hat{\beta}_0, \hat{\beta}_1)$, καταλήγουμε εύκολα σε ένα ζεύγος εξισώσεων, τις επονομαζόμενες **κανονικές εξισώσεις**, οι οποίες είναι οι εξής

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned} \tag{2.2}$$

από τις όποιες λαμβάνουμε τις εκτιμήσεις των β_0 και β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

και

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\,,$$

όπου $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ και $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Η συνθήκη ώστε η συνάρτηση S να λαμβάνει την ελάχιστη τιμής της στο σημείο $(\hat{\beta}_0, \hat{\beta}_1)$ είναι ο πίνακας των δεύτερων παραγώγων της S να είναι θετικά ορισμένος σε αυτό το σημείο (βλ. και Παράρτημα Α.13). Σε δισδιάστατα προβλήματα ένας πίνακας είναι θετικά ορισμένος, όταν το (1,1)–στοιχείο του είναι θετικό καθώς και όταν η ορίζουσά του είναι θετική. Για την ορίζουσα των δεύτερων παραγώγων έχουμε ότι

$$\left| \begin{array}{cc} \frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0^2} & \frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_1^2} \end{array} \right|_{(\hat{\beta}_0,\hat{\beta}_1)} = \left| \begin{array}{cc} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{array} \right| = 4n \sum_{i=1}^{n} (x_i - \bar{x})^2 > 0,$$

από την οποία επιβεβαιώνεται ότι πράγματι στο σημείο $(\hat{\beta}_0, \hat{\beta}_1)$, που προκύπτει από τις παραπάνω εξισώσεις, η συνάρτηση $S(\beta_0, \beta_1)$ παρουσιάζει ελάχιστο. Οι παραπάνω εκτιμήτριες των συντελεστών β_0 και β_1 ονομάζονται **εκτιμήτριες ελαχίστων τετραγώνων (ε.ε.τ.)**.

Ο έλεγχος και η αξιολόγηση του εκτιμημένου ή προσαρμοσμένου μοντέλου με τη μέθοδο των ελαχίστων τετραγώνων βασίζεται σε ορισμένες υποθέσεις για τα τυχαία σφάλματα ε_i. Αν παραβιάζεται μια ή περισσότερες από αυτές, τότε το μοντέλο, στο οποίο εφαρμόζουμε τη μέθοδο αυτή, δεν είναι το κατάλληλο, για να εξηγήσει τη συμπεριφορά των παρατηρήσεών μας. Οι υποθέσεις αυτές είναι οι ακόλουθες

- $E(\varepsilon_i) = 0$, για κάθε i
- $V(\varepsilon_i) = \sigma^2$, για κάθε i
- $cov(\varepsilon_i, \varepsilon_j) = 0$, για $i \neq j$, δηλαδή τα ε_i είναι ασυσχέτιστα

και πρέπει σε κάθε περίπτωση να εξετάζεται αν ικανοποιούνται (βλ. Παράγραφο 4.2).

Στην επόμενη παράγραφο προσθέτουμε και την υπόθεση ότι η κατανομή των τυχαίων σφαλμάτων ε_i είναι η Κανονική, δηλαδή ότι $\varepsilon_i \sim N(0, \sigma^2)$ ανεξαρτήτως και ισονόμως (βλ. Σχήμα 2.1). Η προϋπόθεση της κανονικότητας δεν απαιτείται για την εφαρμογή της μεθόδου ελαχίστων τετραγώνων αλλά επιτρέπει την εκτέλεση στατιστικών ελέγχων. Αξίζει να παρατηρήσουμε ότι με εξαίρεση την υπόθεση για τη σχέση (2.1), την οποία υιοθετούμε για τη σύνδεση των δύο μεταβλητών και το ότι η μεταβλητή x είναι μη στοχαστική, οι υπόλοιπες προϋποθέσεις του μοντέλου αφορούν τα τυχαία σφάλματα ε_i.

Η γνώση ή έστω η υπόθεση της κατανομής των τυχαίων σφαλμάτων μας επιτρέπει να εφαρμόσουμε και τη **μέθοδο μέγιστης πιθανοφάνειας** για την εκτίμηση των παραμέτρων του μοντέλου (βλ. Παράγραφο 3.7). Όπως θα δούμε στο επόμε-

νο κεφάλαιο, οι εκτιμήσεις των δύο αυτών μεθόδων συμπίπτουν υπό την υπόθεση της κανονικότητας των τυχαίων σφαλμάτων. Επιπροσθέτως, η επιπλέον υπόθεση της κανονικότητας της κατανομής των τυχαίων σφαλμάτων επιτρέπει και την κατασκευή στατιστικών ελέγχων υποθέσεων για τους συντελεστές του μοντέλου. Οι στατιστικοί αυτοί έλεγχοι βασίζονται στην υπόθεση της κατανομής των τυχαίων σφαλμάτων και η χρήση τους δεν μπορεί να δικαιολογηθεί διαφορετικά.

Από την προσαρμογή του μοντέλου, όπως αναφέρθηκε και πριν, προκύπτει η προσαρμοσμένη ευθεία (fitted line) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, η οποία διέρχεται από τα σημεία $(x_i, \hat{y}_i)$, $i = 1, 2, \ldots, n$, όπου

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{\mu}_i$$

είναι η εκτίμηση της μεταβλητής y για $x = x_i$ (fitted y value). Η εκτίμηση $\hat{\beta}_1$ εκφράζει την αναμενόμενη μεταβολή της y για μια μονάδα αύξησης της αντίστοιχης επεξηγηματικής μεταβλητής x.

Επίσης λαμβάνονται και τα υπόλοιπα ή κατάλοιπα (residuals)

$$e_i = y_i - \hat{y}_i\,,$$

τα οποία αποτελούν σημαντικές ποσότητες. Τα υπόλοιπα θα μας βοηθήσουν στο να καταλήξουμε στο συμπέρασμα για την ορθότητα ή μη του μοντέλου, δηλαδή αν πραγματικά επαρκεί, για να περιγράψει τη σχέση μεταξύ των y και x. Η ανάπτυξη της σχετικής θεωρίας θα γίνει στο Κεφάλαιο 4, στα πλαίσια της γενίκευσης του μοντέλου.

2.3 Κατανομή των εκτιμητριών και έλεγχος t

Η μέθοδος ελαχίστων τετραγώνων μας δίνει μια σημειακή εκτίμηση των παραμέτρων του μοντέλου και μια σημειακή εκτίμηση της τιμής της εξαρτημένης μεταβλητής για δεδομένη αντίστοιχη τιμή της x. Οι σημειακές αυτές εκτιμήσεις είναι σχεδόν σίγουρο ότι διαφέρουν από τις πραγματικές τιμές. Είναι, λοιπόν, χρήσιμο

για καθεμία ποσότητα που εκτιμούμε, να υπολογίσουμε διαστήματα της μορφής (L, U) τα οποία θα περιέχουν τις πραγματικές τιμές με μεγάλη πιθανότητα. Η πιθανότητα ένα τέτοιο διάστημα να περιέχει την πραγματική τιμή μιας ποσότητας, την οποία εκτιμήσαμε, ορίζεται ως **βαθμός εμπιστοσύνης** (π.χ. 95% ή 99%). Το διάστημα (L, U) καλείται **διάστημα εμπιστοσύνης (δ.ε.)**.

Επιπροσθέτως, είναι χρήσιμο αλλά και απαραίτητο να ελέγξουμε κατά πόσο συμβάλλει η επεξηγηματική μεταβλητή x στο μοντέλο. Για το σκοπό αυτό θα πρέπει να κατασκευάσουμε ένα στατιστικό έλεγχο υποθέσεων για το συντελεστή β_1 της επεξηγηματικής μεταβλητής x.

Προτού προχωρήσουμε στην κατασκευή διαστημάτων εμπιστοσύνης και στατιστικών ελέγχων υποθέσεων, είναι απαραίτητο να βρούμε πρώτα τι κατανομή ακολουθούν οι διάφορες ποσότητες που μας ενδιαφέρουν. Αυτό είναι εφικτό μόνο εφόσον υιοθετήσουμε μια συγκεκριμένη κατανομή για τα τυχαία σφάλματα ε_i. Αν υποθέσουμε ότι αυτή είναι η Κανονική, καθώς και ότι ισχύουν οι προϋποθέσεις που σημειώσαμε στην Παράγραφο 2.2, τότε οι εκτιμήτριες μέγιστης πιθανοφάνειας συμπίπτουν με τις ε.ε.τ. (βλ. και Παράγραφο 3.7).

Από τη σχέση $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ εύκολα προσδιορίζουμε την αναμενόμενη τιμή

$$E(y_i) = \beta_0 + \beta_1 x_i = \mu_i$$

και διασπορά

$$V(y_i) = V(\beta_0 + \beta_1 x_i + \varepsilon_i) = V(\varepsilon_i) = \sigma^2$$

της εξαρτημένης μεταβλητής y.

Επίσης, επειδή αρχικά θεωρήσαμε τα ε_i ασυσχέτιστα μεταξύ τους, προφανώς ισχύει ότι και

$$cov(y_i, y_j) = cov(\varepsilon_i, \varepsilon_j) = 0, \quad \text{για} \quad i \neq j.$$

Για την εκτιμήτρια ελαχίστων τετραγώνων της παραμέτρου β_1 του μοντέλου (2.1)

μπορεί να αποδειχτεί ότι δίνεται και από τη σχέση

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

αφού $\sum_{i=1}^{n}(x_i - \bar{x})\,\bar{y} = 0$. Είναι δηλαδή, ένας γραμμικός συνδυασμός των ανεξάρτητων τυχαίων μεταβλητών y_i, από το οποίο προκύπτει ότι

$$\begin{aligned} E(\hat{\beta}_1) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})\, x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1, \end{aligned}$$

αφού $\sum_{i=1}^{n}(x_i - \bar{x})\, x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Επομένως η $\hat{\beta}_1$ είναι αμερόληπτη εκτιμήτρια της β_1.

Παρομοίως αποδεικνύεται ότι η διασπορά της $\hat{\beta}_1$ είναι

$$\begin{aligned} V(\hat{\beta}_1) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 V(y_i)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}. \end{aligned}$$

Τα παραπάνω αποτελέσματα για τη $E(\hat{\beta}_1)$ και $V(\hat{\beta}_1)$ ισχύουν γενικώς για τις εκτιμήτριες ε.ε.τ. Με την επιπρόσθετη υπόθεση της κανονικότητας των τυχαίων σφαλμάτων, $\varepsilon_i \sim N(0, \sigma^2)$, καταλήγουμε στο συμπέρασμα ότι οι τιμές της εξαρτημένης μεταβλητής y επίσης ακολουθούν την Κανονική κατανομή

$$y_i \sim N(\mu_i, \sigma^2),$$

καθώς και ότι

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

ως γραμμικοί μετασχηματισμοί τ.μ. της Κανονικής κατανομής (βλ. Παράρτημα

Γ.1), και συνεπώς

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim N(0,1). \tag{2.3}$$

Η τυπική απόκλιση σ των τυχαίων σφαλμάτων, η οποία εμφανίζεται στην προηγούμενη σχέση, είναι άγνωστη. Για το λόγο αυτό είναι απαραίτητη η εκτίμησή της μέσω της ποσότητας

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

η οποία αποτελεί την εκτιμήτρια του σ^2 των τυχαίων σφαλμάτων. Η εκτιμήτρια αυτή μπορεί να αποδειχθεί ότι είναι αμερόληπτη της σ^2, δηλαδή ότι $E(S^2) = \sigma^2$ (βλ. Παράγραφο 4.1.1). Επίσης, ονομάζεται και **μέσο άθροισμα τετραγώνων των υπολοίπων** (MSE), όπως θα δούμε στην επόμενη παράγραφο.

Η προηγούμενη σχέση για το S^2 μπορεί να απλοποιηθεί ως

$$S^2 = \frac{1}{n-2} \left\{ S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right\},$$

όπου $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SST$ το **συνολικό άθροισμα τετραγώνων** των y παρατηρήσεων (total sum of squares). Η εκτιμημένη διασπορά του $\hat{\beta}_1$ εκφράζεται ως

$$\widehat{V}(\hat{\beta}_1) = S^2 \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{-1}$$

και το τυπικό σφάλμα ως

$$se(\hat{\beta}_1) = \sqrt{\widehat{V}(\hat{\beta}_1)} = S \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{-1/2} = S\, S_{xx}^{-1/2}.$$

Επιπλέον, μπορεί να αποδειχθεί ότι η κατανομή της στατιστικής συνάρτησης $(n-2)\, S^2/\sigma^2$ είναι η χ^2_{n-2} (βλ. γενική περίπτωση στην Παράγραφο 3.4.1) και ότι οι εκτιμήτριες των συντελεστών β_0 και β_1 είναι ανεξάρτητες του S^2 (βλ. Άσκηση 8).

Εν συνεχεία, αντικαθιστώντας τη σ^2 με την S^2 στη σχέση (2.3), αποδεικνύεται εύκολα (βλ. Παράρτημα Γ.3) ότι

$$\frac{\hat{\beta}_1 - \beta_1}{S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim t_{n-2}\,.$$

Η κατασκευή ενός $100\,(1-\alpha)\,\%$ δ.ε. αλλά και του ελέγχου υποθέσεων για το β_1 βασίζονται στην προηγούμενη σχέση και παρουσιάζονται στη συνέχεια.

- Ένα $100\,(1-\alpha)\,\%$ **διάστημα εμπιστοσύνης** για το β_1 είναι το

$$\left[\hat{\beta}_1 - t_{n-2,\,\alpha/2}S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}},\;\; \hat{\beta}_1 + t_{n-2,\,\alpha/2}S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right],$$

  όπου με $t_{n-2,\,\alpha/2}$ συμβολίζουμε το άνω $100(\alpha/2)$ ποσοστιαίο σημείο της t_{n-2} κατανομής.

- **Στατιστικός έλεγχος υποθέσεων για το συντελεστή** β_1

  Οι υποθέσεις για το συντελεστή β_1 είναι

$$H_0 : \beta_1 = \beta_{1\,(0)} \;\text{ έναντι της }\; H_0 : \beta_1 \neq \beta_{1\,(0)}\,,$$

  όπου $\beta_{1\,(0)}$ μια δεδομένη τιμή. Ο έλεγχος βασίζεται στην ελεγχοσυνάρτηση

$$t = \frac{\hat{\beta}_1 - \beta_{1\,(0)}}{S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim t_{n-2}\,. \tag{2.4}$$

  Τις περισσότερες φορές ελέγχουμε αν $\beta_{1\,(0)} = 0$, δηλαδή εξετάζουμε αν υπάρχει σχέση μεταξύ των μεταβλητών y και x. Σε κάθε περίπτωση όμως ακολουθείται η συνηθισμένη διαδικασία της στατιστικής συμπερασματολογίας που αναπτύχθηκε στην Παράγραφο 1.4, δηλαδή υπολογίζεται αρχικά η

p–τιμή του ελέγχου, η οποία δίνεται από τη σχέση $P(|t_{n-2}| > t)$, όπου t η υπολογισμένη με βάση τα δεδομένα μας τιμή της ελεγχοσυνάρτησης (2.4).

Όπως έχει αναφερθεί και νωρίτερα, η p–τιμή του ελέγχου είναι η πιθανότητα το στατιστικό κριτήριο να πάρει τιμές πιο ακραίες από εκείνη που παρατηρήθηκε, όταν αληθεύει η H_0. Αν η p–τιμή κρίνεται μικρή, τότε η μηδενική υπόθεση απορρίπτεται, αλλιώς γίνεται δεκτή. Το τι αποτελεί «μικρή τιμή» επαφίεται στην κρίση του αναλυτή, λαμβάνοντας υπόψη τη φύση του προβλήματος που αναλύει. Για το λόγο αυτό στην παρουσίαση μιας στατιστικής ανάλυσης δίνεται η p–τιμή του ελέγχου και αφήνεται στον αναγνώστη να αποφασίσει αν συμφωνεί ότι αυτή η τιμή μπορεί να θεωρηθεί μικρή ή όχι. Η συχνότερη επιλογή είναι να θεωρηθούν τιμές κάτω του 0.05 ως μικρές, ή κάτω του 0.01 όταν εκτελούνται πολλοί έλεγχοι. Ωστόσο δεν υπάρχουν γενικοί κανόνες.

Εναλλακτικά, επιλέγουμε ένα επίπεδο σημαντικότητας α (π.χ. 0.05 ή 0.01), προσδιορίζουμε την αντίστοιχη κρίσιμη περιοχή του ελέγχου και εξετάζουμε αν η τιμή της ελεγχοσυνάρτησης πέφτει μέσα στην κρίσιμη περιοχή ή όχι (βλ. Παράγραφο 1.4). Ο προσδιορισμός της κρίσιμης περιοχής γίνεται μέσω του υπολογισμού των ορίων της κατανομής t_{n-2}, όπως στην κατασκευή του παραπάνω διαστήματος εμπιστοσύνης. Για παράδειγμα, για $n-2 = 18$ β.ε. και $\alpha = 0.05$ τα όρια (οι κρίσιμες τιμές του ελέγχου) είναι ± 2.101. Αν η τιμή της ελεγχοσυνάρτησης που προκύπτει από την ανάλυση των δεδομένων, ξεπεράσει τις κρίσιμες τιμές (και άρα θα πέφτει εντός της κρίσιμης περιοχής), δηλαδή αν $|t| > 2.101$, τότε η $H_0 : \beta_1 = \beta_{1\,(0)}$ απορρίπτεται στο επίπεδο στατιστικής σημαντικότητας 0.05. Παραδείγματος χάριν, αν $t = 2.40$, τότε η $H_0 : \beta_1 = \beta_{1\,(0)}$ απορρίπτεται στο επίπεδο στατιστικής σημαντικότητας 0.05, αφού $2.40 > 2.101$. Αν είχαμε υπολογίσει την p–τιμή του ελέγχου, στην προκειμένη περίπτωση την πιθανότητα $P(|t_{18}| > 2.40)$, θα βρίσκαμε ότι αυτή ισούται με 0.027, η οποία μπορεί να θεωρείται μικρή και επομένως θα απορρίπταμε τη μηδενική υπόθεση.

Επίσης, μπορεί να αποδειχτεί ότι η κατανομή της $\hat{\beta}_0$ είναι και αυτή της Κανονικής κατανομής με μέση τιμή

$$E\left(\hat{\beta}_0\right) = \beta_0$$

και διασπορά

$$V\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Όπως αναμένεται, οι εκτιμήτριες $\hat{\beta}_0$ και $\hat{\beta}_1$ δεν είναι ασυσχέτιστες μεταξύ τους και συγκεκριμένα

$$cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Όπως για την β_1, έτσι και για την β_0 μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης και στατιστικούς ελέγχους. Συνήθως όμως η ανάλυση εστιάζεται στη β_1 και σε πολύ μικρότερο βαθμό στη β_0 (βλ. Παράγραφο 2.10).

Από τα παραπάνω προκύπτει ότι και η προσαρμοσμένη $\hat{y}_i$ είναι της Κανονικής κατανομής με αναμενόμενη τιμή

$$E(\hat{y}_i) = E\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = \beta_0 + \beta_1 x_i = \mu_i = E(y_i)$$

και διασπορά

$$\begin{aligned} V(\hat{y}_i) =& V\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \\ =& V\left(\bar{y} + \hat{\beta}_1 (x_i - \bar{x})\right) \\ =& V(\bar{y}) + (x_i - \bar{x})^2 V\left(\hat{\beta}_1\right) + 2(x_i - \bar{x}) cov\left(\bar{y}, \hat{\beta}_1\right) \\ =& \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right), \end{aligned}$$

αφού $cov\left(\bar{y}, \hat{\beta}_1\right) = 0$ (βλ. Άσκηση 5).

2.4 Ανάλυση διασποράς (ANOVA)

Ανάλογη της προσαρμογής του (απλού) γραμμικού μοντέλου είναι η **ανάλυση διασποράς** (analysis of variance), που επίσης εξετάζει τη σχέση της εξαρτημένης με την ανεξάρτητη μεταβλητή, εξετάζοντας στην ουσία αν η **μεταβλητότητα** των τιμών της εξαρτημένης μεταβλητής y εξηγείται από την ανεξάρτητη μεταβλητή x. Η ανάλυση διασποράς του απλού γραμμικού μοντέλου παρουσιάζεται στον Πίνακα 2.1.

Το συνολικό άθροισμα τετραγώνων (SST) μπορεί να αναλυθεί σε άθροισμα δύο συνιστωσών μεταβλητότητας, του **αθροίσματος τετραγώνων λόγω παλινδρόμησης** (SSR – sum of squares due to regression) και του **αθροίσματος τετραγώνων των υπολοίπων ή λόγω σφάλματος** (SSE – residual sum of squares ή sum of squares due to error) ως εξής:

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

αφού $\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ (βλ. Άσκηση 6).

Σε κάθε ένα άθροισμα τετραγώνων αντιστοιχούν ορισμένοι βαθμοί ελευθερίας (degrees of freedom). Η στήλη «Μέσο άθροισμα τετραγώνων» προκύπτει διαιρώντας τα αθροίσματα τετραγώνων με τους αντίστοιχους βαθμούς ελευθερίας. Παρατηρούμε από τον Πίνακα 2.1 ότι το μέσο άθροισμα τετραγώνων των υπολοίπων ή λόγω σφάλματος

$$MSE = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = S^2,$$

αποτελεί την εκτιμήτρια της σ^2 (βλ. προηγούμενη παράγραφο καθώς και Παράγραφο 4.1.1). Παρομοίως, αν υπολογίζαμε ένα μέσο του συνολικού αθροίσματος τετραγώνων (που συνήθως δεν εισάγεται στον πίνακα, αφού δε θα χρησιμοποιηθεί), θα ήταν

Πίνακας 2.1: Πίνακας ανάλυσης διασποράς

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελ/ρίας	Μέσο άθροισμα τετραγώνων	Έλεγχος F
Παλινδρόμηση (Regression)	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Υπόλοιπα (Error)	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-2$	$MSE = S^2 = \frac{SSE}{n-2}$	
Σύνολο (Total)	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$		

$$MST = \frac{SST}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 ,$$

που αποτελεί τη γνωστή εκτιμήτρια της διασποράς του δείγματος τιμών y_i. Γιατί υπολογίζουμε το μέσο των n όρων $(y_i - \bar{y})^2$, διαιρώντας με $n-1$ και όχι με n; Η τελική απάντηση βρίσκεται στη θεωρία των κατανομών αυτών των ποσοτήτων, αλλά η γενική ιδέα είναι η ακόλουθη.

Αν και το SST υπολογίζεται από n όρους $(y_i - \bar{y})^2$, δεν πρόκειται για n ξεχωριστές πληροφορίες λόγω του περιορισμού

$$\sum_{i=1}^{n} (y_i - \bar{y}) = 0.$$

Έτσι, το SST θα έχει $n-1$ «βαθμούς ελευθερίας», διότι, αν γνωρίζουμε $n-1$ αποκλίσεις $(y_i - \bar{y})$, τότε προφανώς και η εναπομείνασα απόκλιση $(y_j - \bar{y})$, $j \neq i$ θα είναι γνωστή και στατιστικά δε θα προσθέτει καμία νέα πληροφορία. Το SSE έχει $n-2$ βαθμούς ελευθερίας, γιατί υπάρχουν δύο περιορισμοί

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0 \quad \text{και} \quad \sum_{i=1}^{n} x_i \, (y_i - \hat{y}_i) = 0 \,,$$

όπως φαίνεται από τις κανονικές εξισώσεις (2.2) (βλ. και Παράγραφο 3.4). Τελικά, το SSR έχει 1 βαθμό ελευθερίας, γιατί μπορεί να υπολογισθεί από μια μόνο πληροφορία τη $\hat{\beta}_1$, αφού

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,.$$

Η ανάλυση διασποράς και η σχέση της με την ανάλυση παλινδρόμησης θα παρουσιαστεί αναλυτικά στο Κεφάλαιο 6.

2.5 Στατιστικός έλεγχος F

Οι στατιστικές συναρτήσεις SSR και SSE είναι ανεξάρτητες μεταξύ τους και διαιρεμένες με το σ^2, ακολουθούν την κατανομή χ^2 με βαθμούς ελευθερίας τους αντίστοιχους των αθροισμάτων τετραγώνων, όταν τα τυχαία σφάλματα ε_i κατανέμονται σύμφωνα με την Κανονική κατανομή (βλ. Παράγραφο 3.4). Κατά συνέπεια, η στατιστική συνάρτηση F του Πίνακα 2.1 ακολουθεί την $F_{1,(n-2)}$ κατανομή και μπορεί να χρησιμοποιηθεί, για να ελέγξουμε τη σημαντικότητα της παλινδρόμησης, δηλαδή της υπόθεσης

$$H_0 : \beta_1 = 0$$

έναντι της

$$H_0 : \beta_1 \neq 0.$$

Η μηδενική υπόθεση H_0 απορρίπτεται, όταν η p–τιμή του ελέγχου, δηλαδή η $P\left(F_{1,(n-2)} > F\right)$ με F την υπολογισμένη τιμή της ελεγχοσυνάρτησης, είναι μικρή. Ουσιαστικά με τον έλεγχο αυτό εξετάζουμε τον ισχυρισμό ότι πράγματι η y σχετίζεται με τη x, όπως περιγράφεται από το γραμμικό μοντέλο.

Παρατήρηση 2.5.1. *Στην περίπτωση του απλού γραμμικού μοντέλου ο έλεγχος t και ο έλεγχος F είναι ισοδύναμοι για την υπόθεση $H_0 : \beta_1 = 0$. Μπορεί να αποδειχθεί ότι η τιμή του F του Πίνακα 2.1 είναι το τετράγωνο της τιμής του t της Παραγράφου 2.3 για αυτή την υπόθεση. Επίσης οι δύο θεωρητικές κατανομές έχουν τη σχέση*

$$F_{1,(n-2)} = t^2_{n-2}\,,$$

επομένως και οι p–τιμές των δύο ελέγχων θα συμπίπτουν (βλ. Παράρτημα Γ.4). Η διαφορετική χρησιμότητα των δύο ελέγχων θα φανεί όταν υπάρχουν περισσότερες από μια ανεξάρτητες μεταβλητές στο μοντέλο.

2.6 Συντελεστής προσδιορισμού R^2

Ένας άλλος σημαντικός τρόπος για να κρίνουμε την αξία του (απλού) γραμμικού μοντέλου, που προσαρμόζουμε στα δεδομένα, είναι ο λόγος

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}\,,$$

ο οποίος ονομάζεται **συντελεστής ή δείκτης προσδιορισμού** (coefficient of determination) και εκφράζει το ποσοστό της μεταβλητότητας της τ.μ. y, που εξηγείται από την x (βλ. και Παραγράφους 3.5 και 5.1.1). Οι τιμές, τις οποίες μπορεί να πάρει ο R^2, είναι ανάμεσα στο μηδέν και το ένα (ή μηδέν έως 100, όταν εκφράζεται ως ποσοστό επί τοις εκατό). Όσο πιο κοντά είναι η τιμή του R^2 στη μονάδα τόσο ισχυρότερη είναι η γραμμική σχέση εξάρτησης των τ.μ. y και x. Στην περίπτωση του απλού γραμμικού μοντέλου και μόνο ισχύει ότι $R^2 = r^2_{xy}$

(βλ. Άσκηση 2), όπου

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

ο δειγματικός συντελεστής συσχέτισης Pearson που εκφράζει το βαθμό της γραμμικής συσχέτισης μεταξύ των δύο μεταβλητών x και y. Οι τιμές που μπορεί να πάρει ο συντελεστής αυτός είναι μεταξύ -1 και 1. Επίσης, εύκολα διαπιστώνεται ότι για το απλό γραμμικό μοντέλο ισχύει και η σχέση

$$S^2 = \frac{1}{n-2} S_{yy}\left\{1 - r_{xy}^2\right\} = \frac{1}{n-2} S_{yy}\left\{1 - R^2\right\},$$

η οποία συνδέει την εκτίμηση της διασποράς των τυχαίων σφαλμάτων με τους προαναφερθέντες συντελεστές (βλ. Άσκηση 3).

Ο έλεγχος της $H_0 : \rho_{xy} = 0$ έναντι της $H_1 : \rho_{xy} \neq 0$ για τον πληθυσμιακό, πραγματικό συντελεστή συσχέτισης ρ_{xy} πραγματοποιείται με την ελεγχοσυνάρτηση

$$\frac{r_{xy}\sqrt{n-2}}{\sqrt{1 - r_{xy}^2}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)},$$

δηλαδή πάλι με την ελεγχοσυνάρτηση (2.4), η οποία υπό την H_0 ακολουθεί την κατανομή t_{n-2} (βλ. και Άσκηση 7). Για την κατασκευή ενός διαστήματος εμπιστοσύνης για το συντελεστή συσχέτισης ρ_{xy} (και κατά συνέπεια για το συντελεστή R^2 του πληθυσμού) χρησιμοποιείται η στατιστική συνάρτηση

$$Z = \frac{\frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right)}{\sqrt{\frac{1}{n-3}}},$$

η οποία προσεγγιστικά ακολουθεί την $N(0,1)$ κατανομή.

2.7 Πρόβλεψη

Η σημειακή πρόβλεψη μιας άγνωστης παρατήρησης y για δοθέν x_0, το οποίο δεν είναι απαραίτητο να περιλαμβάνεται στο αρχικό σύνολο των x_i παρατηρήσεων, δίνεται από τη σχέση

$$\hat{y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \hat{\mu}_{x_0} .$$

Είναι προφανές (βλ. Παράγραφο 2.3, Παράρτημα Γ.1 και Άσκηση 1) ότι, αν ισχύει η υπόθεση της κανονικότητας των τυχαίων σφαλμάτων $\varepsilon_i \sim N(0, \sigma^2)$, τότε η $\hat{y}_{x_0}$ ακολουθεί την Κανονική κατανομή

$$\hat{y}_{x_0} \sim N\left(\mu_{x_0}, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]\right),$$

όπου $\mu_{x_0} = E(\hat{y}_{x_0}) = \beta_0 + \beta_1 x_0 = E(y_{x_0})$.

Επίσης ισχύει ότι

$$\hat{y}_{x_0} - y_{x_0} \sim N\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]\right)$$

και επομένως

$$\frac{\hat{y}_{x_0} - y_{x_0}}{\sigma \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]^{1/2}} \sim N(0, 1) .$$

Η προηγούμενη σχέση μπορεί να χρησιμοποιηθεί, για να κατασκευάσουμε ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης της πρόβλεψης της y_{x_0} παρατήρησης στο σημείο x_0, αφού αντικατασταθεί η άγνωστη ποσότητα σ^2 με την εκτίμησή της S^2. Όπως έχει σημειωθεί ξανά (βλ. και Παράρτημα Γ.3), αποδεικνύεται ότι

$$\frac{\hat{y}_{x_0} - y_{x_0}}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}} \sim t_{n-2} .$$

Επομένως, ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την **άγνωστη παρατήρηση** y_{x_0} είναι το $\left(L_{y_{x_0}}, U_{y_{x_0}}\right)$, όπου

$$L_{y_{x_0}} = \hat{y}_{x_0} - t_{n-2,\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

και

$$U_{y_{x_0}} = \hat{y}_{x_0} + t_{n-2,\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Στην περίπτωση που θέλουμε να κατασκευάσουμε ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την **άγνωστη μέση τιμή** $\mu_{x_0} = E(y_{x_0})$ στο σημείο x_0, έχουμε ότι

$$\frac{\hat{y}_{x_0} - \mu_{x_0}}{\sigma \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]^{1/2}} \sim N(0,1)$$

και ακολουθώντας παρόμοια διαδικασία με πριν, προκύπτει ότι ένα $100(1-\alpha)\%$ δ.ε. για την μ_{x_0} είναι το $\left(L_{\mu_{x_0}}, U_{\mu_{x_0}}\right)$ με

$$L_{\mu_{x_0}} = \hat{y}_{x_0} - t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

και

$$U_{\mu_{x_0}} = \hat{y}_{x_0} + t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Είναι προφανές ότι το διάστημα εμπιστοσύνης της μ_{x_0} περιέχεται στο διάστημα εμπιστοσύνης της πρόβλεψης της y_{x_0} παρατήρησης, για κάθε x_0 (βλ. Παράδειγμα 2.8.1, Σχήμα 2.3, καθώς και Παράγραφο 3.9).

2.8 Χρήση MINITAB

Η προσαρμογή του απλού γραμμικού μοντέλου, δηλαδή η εκτίμηση των παραμέτρων αλλά και η ανάλυση διασποράς και ο υπολογισμός του συντελεστή προσδιορισμού, γίνεται αυτόματα με τη βοήθεια του MINITAB ακολουθώντας την παρακάτω διαδικασία.

- Επιλέγουμε από τη γραμμή εργαλείων τις εντολές
  Stat→Regression→Regression

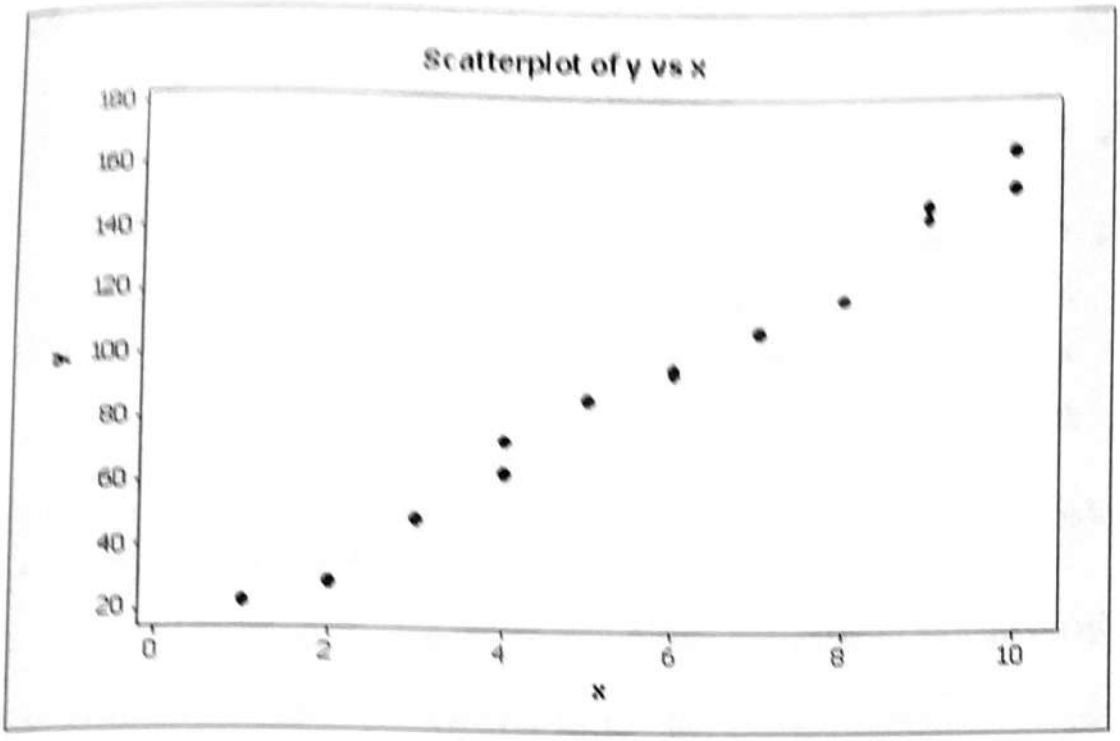
- Στο παράθυρο που ανοίγει, εισάγουμε στο πλαίσιο **Response** τη στήλη που περιέχει την **εξαρτημένη** μεταβλητή y, και στο πλαίσιο **Predictors** την **ανεξάρτητη** μεταβλητή x.

Παράδειγμα 2.8.1. *Σε μια μελέτη καταγράφηκαν ο χρόνος επισκευής και ο αριθμός των προς διόρθωση εξαρτημάτων για 14 ηλεκτρονικά μηχανήματα, με σκοπό τον εντοπισμό της σχέσης που υπάρχει ανάμεσα στον αριθμό x των εξαρτημάτων του μηχανήματος που απαιτούν επισκευή, και στο χρονικό διάστημα y (σε λεπτά) που απαιτείται για την ολοκλήρωση της εργασίας επισκευής. Τα αποτελέσματα των μετρήσεων παρουσιάζονται στον Πίνακα 2.2.*

Ζητείται να κατασκευαστεί το διάγραμμα διασποράς και να προσαρμοστεί το απλό γραμμικό μοντέλο.

Πίνακας 2.2: Πίνακας δεδομένων για το Παράδειγμα 2.8.1

x	1	2	3	4	4	5	6	6	7	8	9	9	10	10
y	23	29	49	64	74	87	96	97	109	119	149	145	154	166



Σχήμα 2.2: Διάγραμμα διασποράς για τα δεδομένα του Παραδείγματος 2.8.1

Η **κατασκευή του διαγράμματος διασποράς** *γίνεται*

- *επιλέγοντας* **Graph→Scatterplot→Simple** *από τη γραμμή εργαλείων*

- *θέτοντας στη συνέχεια στο παράθυρο που ανοίγει, κάτω από το* **Y variables** *τη στήλη που περιέχει τις τιμές της διάρκειας επισκευής, δηλαδή την* **εξαρτημένη μεταβλητή**, *και κάτω από το* **X variables** *τη στήλη με τις τιμές των αριθμών των εξαρτημάτων προς επισκευή, δηλαδή την* **ανεξάρτητη μεταβλητή**.

Το MINITAB κατασκευάζει το διάγραμμα διασποράς του Σχήματος 2.2, από τη μορφή του οποίου παρατηρούμε ότι υπάρχει μια ισχυρή γραμμική σχέση ανάμεσα στις δύο μεταβλητές. Αυτή η σχέση περιμένουμε να αποδοθεί με την προσαρμογή του απλού γραμμικού μοντέλου στα δεδομένα μας.

Για την **προσαρμογή του γραμμικού μοντέλου** επιλέγουμε

- **Stat→Regression→Regression** από τη γραμμή εργαλείων
- Στο παράθυρο που ανοίγει, εισάγουμε στο πλαίσιο **Response** τη στήλη που περιέχει τις τιμές της **εξαρτημένης μεταβλητής** y, και στο πλαίσιο **Predictors** τη στήλη που περιέχει τις τιμές της **ανεξάρτητης μεταβλητής** x.

Ακολουθώντας την παραπάνω διαδικασία το *MINITAB* τυπώνει τα *Αποτελέσματα 2.1*, στα οποία μας παρουσιάζει την **προσαρμοσμένη συνάρτηση παλινδρόμησης** για τα δεδομένα μας, η οποία είναι η

$$\hat{y} = 4.16 + 15.5x$$

δηλαδή η εκτίμηση του συντελεστή β_0 είναι ίση με 4.16, ενώ του β_1 με 15.5.

Επίσης, για κάθε ένα συντελεστή παρουσιάζεται το **τυπικό σφάλμα** (se) και η **τιμή του ελέγχου** t για την υπόθεση ότι ο συντελεστής αυτός ισούται με το μηδέν, καθώς και η p–τιμή για τον έλεγχο αυτό. Παραδείγματος χάρη, το τυπικό σφάλμα του συντελεστή της μεταβλητής x, δηλαδή $se(\hat{\beta}_1)$, είναι ίσο με 0.505, η τιμή της ελεγχοσυνάρτησης t υπό την υπόθεση $H_0 : \beta_1 = 0$ είναι $15.5088/0.5050{=}30.71$ και η αντίστοιχη p–τιμή <0.001[1] της κατανομής t με 12 β.ε., γεγονός που μας υποδεικνύει ότι ο συντελεστής του x είναι στατιστικά διάφορος του μηδενός. Πρόκειται για τον έλεγχο (2.4) με $\beta_{1(0)} = 0$.

Υπό τον ίδιο συλλογισμό μπορούμε να ισχυριστούμε ότι η **σταθερά** (constant) του μοντέλου δε φαίνεται να είναι στατιστικά σημαντική (p–τιμή$=0.239$), δηλαδή δεν μπορεί να απορριφθεί η μηδενική υπόθεση

$$H_0 : \beta_0 = 0\,.$$

[1] Είναι προτιμότερο να γράφουμε p–τιμή <0.001, ακόμα και αν στον πίνακα αποτελεσμάτων του προγράμματος η p–τιμή έχει την τιμή μηδέν, και αυτό γιατί η p–τιμή δεν είναι ποτέ μηδέν, αλλά μπορεί να πάρει τιμές πολύ κοντά στο μηδέν.

Αποτελέσματα 2.1

```
Regression Analysis: y versus x

The regression equation is
y = 4.16 + 15.5 x

Predictor      Coef  SE Coef      T      P
Constant      4.162    3.355   1.24  0.239
x           15.5088   0.5050  30.71  0.000

S = 5.39172   R-Sq = 98.7%   R-Sq(adj) = 98.6%

Analysis of Variance

Source           DF     SS     MS       F      P
Regression        1  27420  27420  943.20  0.000
Residual Error   12    349     29
Total            13  27768

Predicted Values for New Observations

New
Obs     Fit  SE Fit        95% CI             95% PI
  1  174.76    2.91  (168.42; 181.09)  (161.41; 188.10)

Values of Predictors for New Observations

New
Obs     x
  1  11.0
```

Πρέπει όμως να σημειώσουμε ότι συνήθως δε μας ενδιαφέρει ο σταθερός όρος του μοντέλου, αλλά ποιες επεξηγηματικές μεταβλητές και κυρίως με ποιο τρόπο επηρεάζουν την εξαρτημένη μεταβλητή. Για το λόγο αυτό τις περισσότερες φορές δεν εξετάζουμε τη σταθερά του μοντέλου και σχεδόν πάντα τη συμπεριλαμβάνουμε στο μοντέλο ανεξάρτητα από τη στατιστική σημαντικότητά της (βλ. Παράγραφο 2.10). Αν όμως, παρόλα αυτά, επιθυμούμε η σταθερά να αφαιρεθεί από την ανάλυση, το MINITAB μας επιτρέπει την αφαίρεσή της επιλέγοντας την απενεργοποίηση της επιλογής **Fit intercept** *στο* **Options** *του παραθύρου* **Regression**. *(Να σημειώσουμε ότι intercept ονομάζεται στα Αγγλικά η απόσταση*

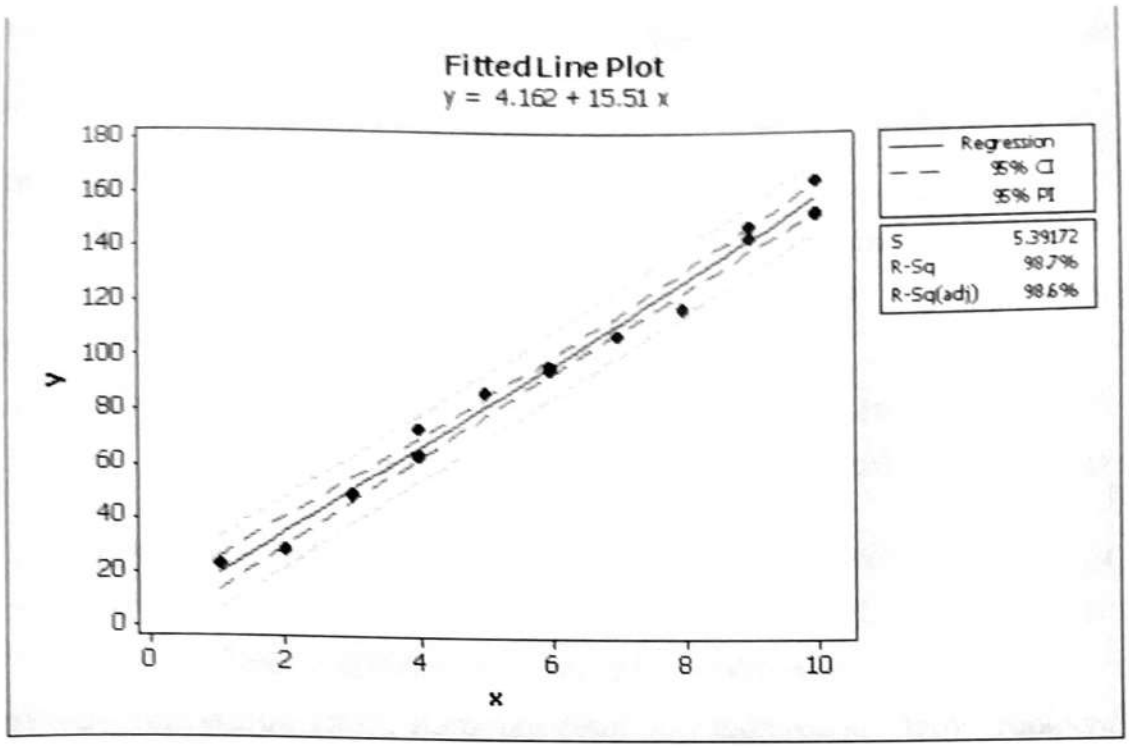
β_0 από την αρχή των αξόνων στο σημείο, όπου η ευθεία κόβει τον άξονα του y).

Στη συνέχεια παρουσιάζεται η εκτίμηση της **τυπικής απόκλισης** σ του τυχαίου σφάλματος ε, $S = 5.392 = \sqrt{MSE}$ και ο **συντελεστής προσδιορισμού** R^2, ο οποίος στην περίπτωση αυτή είναι ίσος με 98.7%, ένδειξη της ισχυρής γραμμικής σχέσης μεταξύ των δύο μεταβλητών.

Επίσης, στα αποτελέσματα που παρέχει το MINITAB παρουσιάζεται και η **ανάλυση διασποράς (Analysis of Variance)** για τα δεδομένα μας. Όπως παρατηρήσαμε στην Παράγραφο 2.5, ο έλεγχος F και ο έλεγχος t για την $H_0 : \beta_1 = 0$ προσφέρουν την ίδια πληροφορία στην περίπτωση του απλού γραμμικού μοντέλου. Πράγματι $\sqrt{F} = \sqrt{943.2} = 30.71 = t$. Για την παράμετρο β_1 του μοντέλου υπολογίζουμε και ένα 95% διάστημα εμπιστοσύνης ως (14.41, 16.61), όπου $se(\hat{\beta}_1) = S/(S_{xx})^{1/2} = 5.39172/\sqrt{114} \simeq 0.505$ είναι το τυπικό σφάλμα του $\hat{\beta}_1$, και $t_{0.975} = 2.179$ είναι το 0.025–άνω ποσοστιαίο σημείο της κατανομής t με βαθμούς ελευθερίας $n - 2 = 12$ (βλ. Παράρτημα Θ). Παρατηρούμε ότι το μηδέν δεν περιέχεται στο διάστημα εμπιστοσύνης, το οποίο συμφωνεί με τους στατιστικούς ελέγχους για την υπόθεση $H_0 : \beta_1 = 0$.

Αν τώρα θέλουμε να προσθέσουμε τη συνάρτηση παλινδρόμησης $\hat{y} = 4.16 + 15.5x$ στο διάγραμμα διασποράς, πρέπει να ακολουθήσουμε την εξής διαδικασία:

- Επιλέγουμε **Stat→Regression→Fitted Line Plot** από τη γραμμή εργαλείων
- θέτουμε στο παράθυρο που ανοίγει, στο πλαίσιο **Response(Y)** τη στήλη που περιέχει τις τιμές της διάρκειας επισκευής **(εξαρτημένη μεταβλητή)** και στο πλαίσιο **Predictor(X)** τη στήλη με τους αριθμούς των εξαρτημάτων προς επισκευή **(ανεξάρτητη μεταβλητή)**
- επιλέγουμε **Type of Regression Model: Linear**.



Σχήμα 2.3: Η προσαρμοσμένη ευθεία παλινδρόμησης για τα δεδομένα του Παραδείγματος 2.8.1

Το MINITAB κατασκευάζει το διάγραμμα του Σχήματος 2.3, στο οποίο βλέπουμε τη σχέση της συνάρτησης παλινδρόμησης με τις παρατηρήσεις μας. Σε αυτό προσθέτουμε διακεκομμένες γραμμές των ορίων (L, U) των 95% διαστημάτων εμπιστοσύνης πρόβλεψης για τις άγνωστες παρατηρήσεις y_x, καθώς και για τις μέσες τιμές των, μ_x (βλ. Παράγραφο 2.7).

Σημειωτέον ότι το MINITAB καλεί το διάστημα εμπιστοσύνης για μια άγνωστη παρατήρηση **Prediction Interval** *και το αντίστοιχο για τη μέση τιμή της,* **Confidence Interval**. *Τα δ.ε. κατασκευάζονται επιλέγοντας* **Display confidence interval** *και* **Display prediction interval** *στο* **Options** *του* **Fitted Line Plot**.

Παρατηρούμε, όπως είχαμε τονίσει στην Παράγραφο 2.7, ότι το διάστημα εμπιστοσύνης για ένα μ_x περιέχεται στο αντίστοιχο για την y_x. Αξίζει να σημειωθεί – το οποίο ωστόσο δεν είναι εμφανές στο συγκεκριμένο παράδειγμα – ότι οι γραμμές αυτές δεν είναι ευθείες. Λόγω του όρου $(x_0 - \bar{x})^2$ στις εκφράσεις για τα δ.ε.,

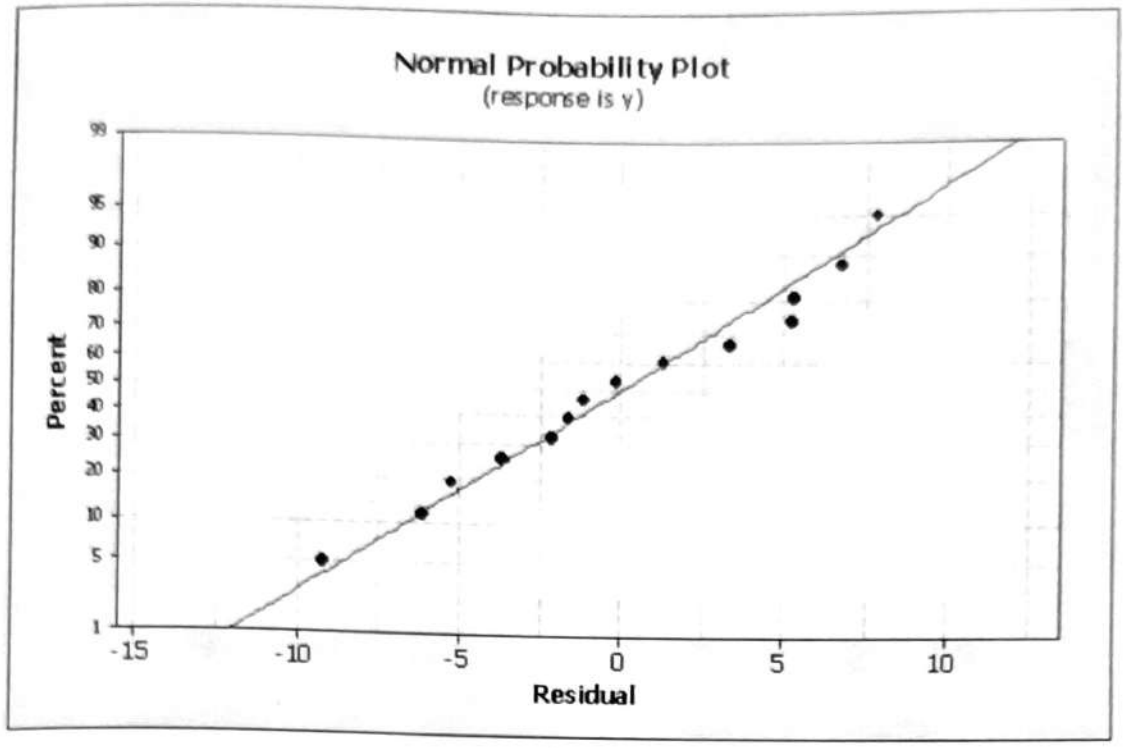
το πλάτος του δ.ε. αυξάνεται καθώς η τιμή της x_0 απομακρύνεται από τη μέση τιμή $\bar{x}$.

Επιπρόσθετα, στα Αποτελέσματα 2.1, κάτω από τον τίτλο **Predicted Values for New Observations** παρουσιάζονται τα 95% διαστήματα εμπιστοσύνης της πρόβλεψης για την άγνωστη παρατήρηση y_{x_0}, (161.41, 188.10), καθώς και για τη μέση τιμή της, $E(y_{x_0})$, (168.42, 181.09), όταν $x_0 = 11$ (**Values of Predictors for New Observations**). Αυτά τα διαστήματα εμπιστοσύνης λαμβάνονται μέσω της επιλογής **Options** του παραθύρου **Regression**.

Για να είμαστε σίγουροι ότι το μοντέλο που προσαρμόσαμε είναι κατάλληλο για την περιγραφή των δεδομένων μας, πρέπει να ελέγξουμε ότι οι προϋποθέσεις για τα τυχαία σφάλματα αληθεύουν. Αν μια ή περισσότερες από τις προϋποθέσεις δεν ισχύουν, τότε το μοντέλο που προσαρμόσαμε βάσει αυτών των υποθέσεων δεν είναι το σωστό. Δηλαδή η ανάλυση της παλινδρόμησης δε σταματά με την προσαρμογή ενός μοντέλου παλινδρόμησης, αλλά επεκτείνεται με την ανάλυση των υπολοίπων μέσω γραφικών παραστάσεων. Με αυτές τις γραφικές παραστάσεις είναι δυνατόν να εξεταστεί η καταλληλότητα προσαρμογής του μοντέλου. Όπως θα δούμε διεξοδικά και στην Παράγραφο 4.2, υπάρχει μια σειρά γραφικών παραστάσεων, οι οποίες εξετάζουν αν ισχύουν αυτές οι υποθέσεις. Εδώ θα περιοριστούμε σε μια απλή παρουσίαση δύο τέτοιων γραφικών παραστάσεων, με τη βοήθεια των οποίων εξετάζουμε τις προϋποθέσεις της **κανονικότητας** και της **ομοσκεδαστικότητας**[2] (homoscedasticity) των τυχαίων σφαλμάτων.

Η πρώτη γραφική παράσταση (Σχήμα 2.4) αποτελεί ένα γραφικό έλεγχο της υπόθεσης της κανονικότητας των υπολοίπων (Normal Probability Plot). Αν τα υπόλοιπα (residuals) ακολουθούν την Κανονική κατανομή, τότε τα σημεία της γραφικής παράστασης πρέπει να κείτονται σε μια ευθεία, όπως συμβαίνει στη συγκεκριμένη περίπτωση. Σε διαφορετική περίπτωση η καταλληλότητα του μοντέλου παλινδρόμησης αμφισβητείται (βλ. Παράρτημα Δ).

[2] Με τον όρο ομοσκεδαστικότητα εννοούμε την υπόθεση ότι τα τυχαία σφάλματα έχουν κοινή διασπορά $V(\varepsilon_i) = \sigma^2$

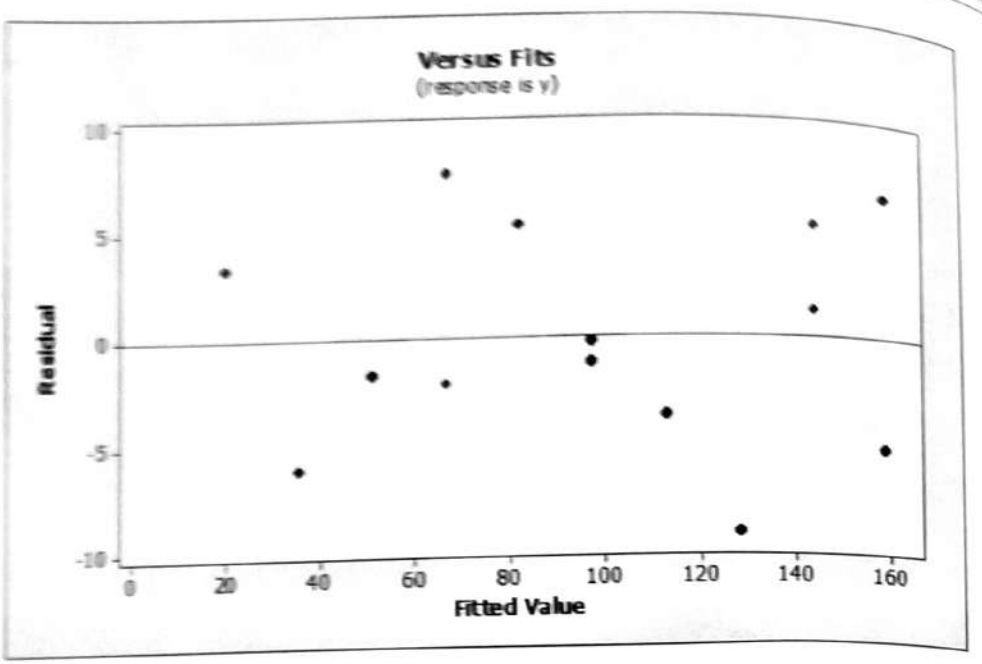


Σχήμα 2.4: Γραφική παράσταση της Κανονικής κατανομής για τα δεδομένα του Παραδείγματος 2.8.1

Στη δεύτερη γραφική παράσταση (Σχήμα 2.5) έχουμε σχεδιάσει τα υπόλοιπα σε σχέση με τα εκτιμημένα $\hat{y}_i$. Λόγω της τυχαίας και ομοσκεδαστικής μορφής της κατανομής των υπολοίπων γύρω από το μηδέν, καταλήγουμε στο να δεχτούμε την υπόθεση της ομοσκεδαστικότητας. Σε αντίθετη περίπτωση, δηλαδή στην περίπτωση όπου τα υπόλοιπα κατανέμονται με μη τυχαίο ή με μη ομοσκεδαστικό τρόπο, αμφισβητείται πάλι η καταλληλότητα του μοντέλου.

Οι δύο αυτές γραφικές παραστάσεις εκτελούνται επιλέγοντας **Normal plot of residuals** *και* **Residuals versus fits** *στο πλαίσιο* **Graphs** *του* **Regression** *ή του* ***Fitted Line Plot***.

Σημειωτέον ότι στη διάθεσή μας δεν έχουμε μονάχα αυτούς τους γραφικούς ελέγχους αλλά και ένα πλήθος άλλων γραφικών παραστάσεων, οι οποίες μας βοηθούν να ελέγξουμε το σύνολο των υποθέσεων μας. Όλοι αυτοί οι έλεγχοι, ο τρόπος με τον οποίο κατασκευάζονται και ερμηνεύονται, παρουσιάζονται αναλυτικά αργότερα.



Σχήμα 2.5: Υπόλοιπα σε σχέση με τα εκτιμημένα $\hat{y}_i$ για τα δεδομένα του Παραδείγματος 2.8.1

Παράδειγμα 2.8.2. *Ένας τηλεοπτικός σταθμός ζητάει να βρει αν επηρεάζεται και με ποιο τρόπο η θεαματικότητα y των διαφόρων προγραμμάτων του από τη θεαματικότητα x του προγράμματος που προηγείται. Για το λόγο αυτό έγιναν 30 μετρήσεις σε διάφορες ώρες. Τα δεδομένα παρουσιάζονται στον Πίνακα 2.3.*

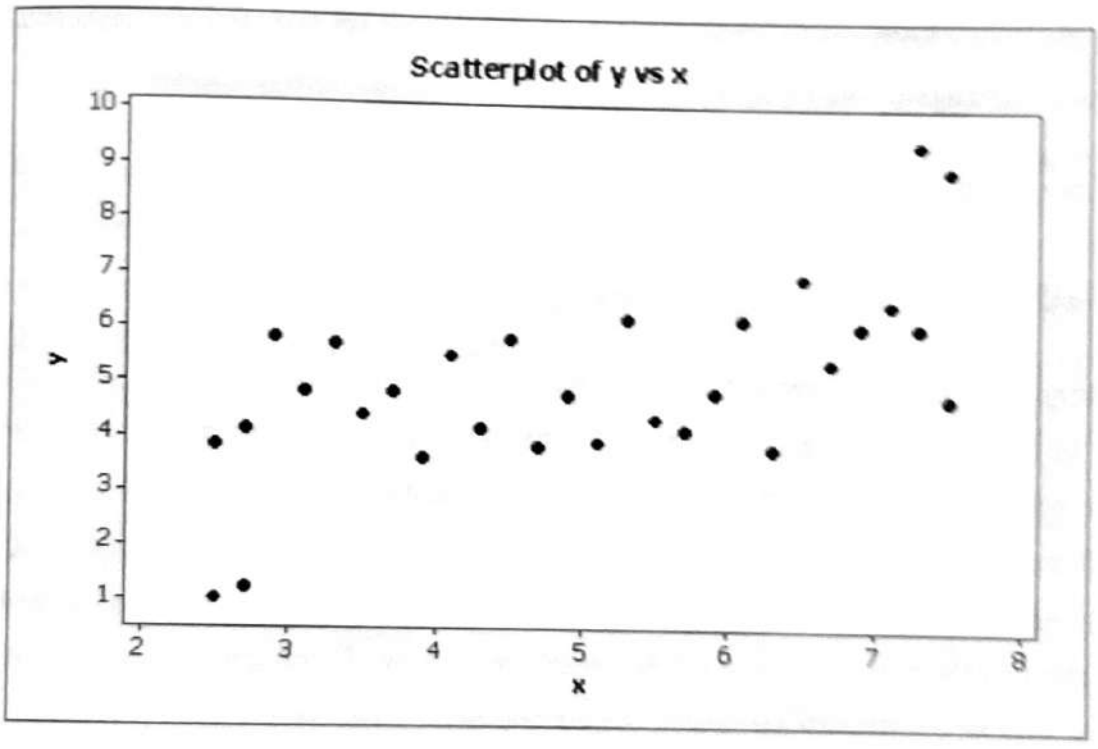
Το διάγραμμα διασποράς των δεδομένων παρουσιάζεται στο Σχήμα 2.6. Στο διάγραμμα διακρίνουμε μια γραμμική, ελαφρά αύξουσα σχέση μεταξύ των μεταβλητών. Στο ίδιο διάγραμμα παρατηρούμε επίσης και την παρουσία τεσσάρων παρατηρήσεων (σημείων), που απέχουν αισθητά από τα υπόλοιπα.

Προσαρμόζοντας το απλό γραμμικό μοντέλο στα δεδομένα μας με τη βοήθεια του MINITAB με τη διαδικασία που παρουσιάστηκε νωρίτερα, λαμβάνουμε τα Αποτελέσματα 2.2.

Παρατηρούμε ότι περίπου 40% της μεταβλητότητας της θεαματικότητας ενός προγράμματος εξηγείται από τη θεαματικότητα του προηγούμενου προγράμμα-

Πίνακας 2.3: Πίνακας δεδομένων για το Παράδειγμα 2.8.2

y	3.8	4.1	5.8	4.8	5.7	4.4	4.8	3.6	5.5	4.15
x	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3
y	5.8	3.8	4.75	3.9	6.2	4.35	4.15	4.85	6.2	3.8
x	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1	6.3
y	7.0	5.4	6.1	6.5	6.1	4.75	1.0	1.2	9.5	9.0
x	6.5	6.7	6.9	7.1	7.3	7.5	2.5	2.7	7.3	7.5



Σχήμα 2.6: Διάγραμμα διασποράς για τα δεδομένα του Παραδείγματος 2.8.2

Αποτελέσματα 2.2

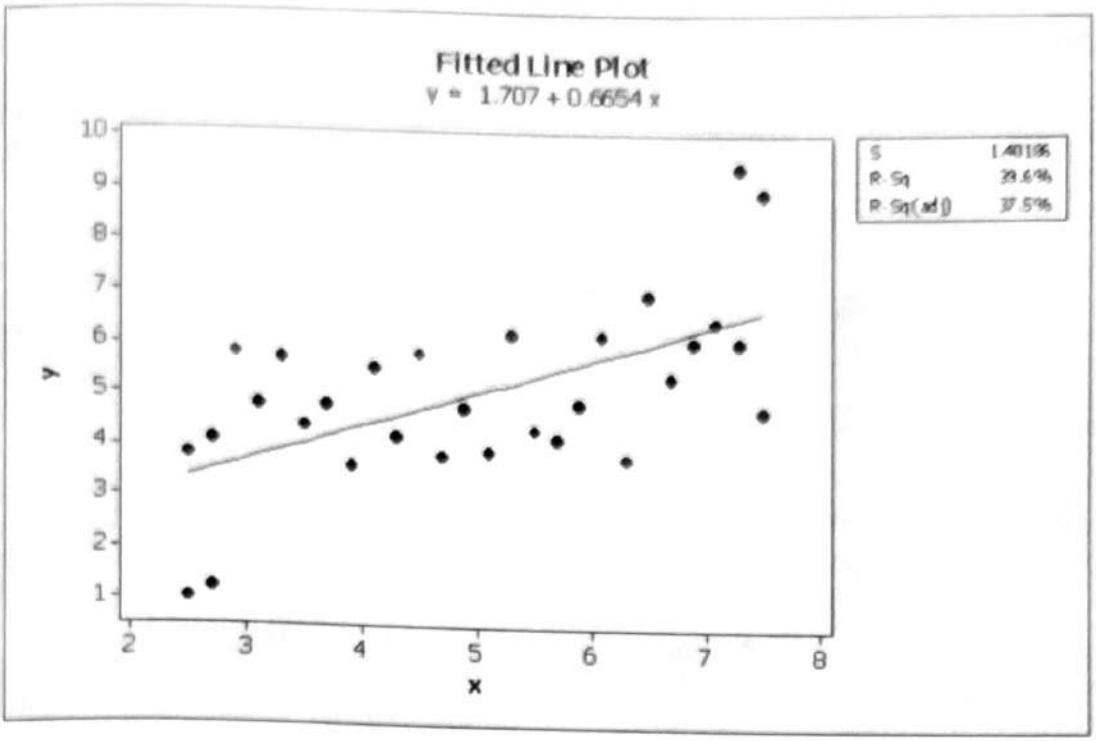
```
Regression Analysis: y versus x

The regression equation is
y = 1.71 + 0.665 x

Predictor     Coef  SE Coef     T      P
Constant    1.7065   0.8172  2.09  0.046
x           0.6654   0.1552  4.29  0.000

S = 1.40186   R-Sq = 39.6%   R-Sq(adj) = 37.5%

Analysis of Variance
Source           DF      SS      MS      F      P
Regression        1  36.116  36.116  18.38  0.000
Residual Error   28  55.026   1.965
Total            29  91.142
```

τος (R^2=39.6%). Επίσης, από τη συνάρτηση παλινδρόμησης μπορούμε να πούμε ότι η θεαματικότητα ενός προγράμματος αναμένεται να αυξηθεί περίπου 0.67 για κάθε επιπλέον μονάδα τηλεθέασης του προηγούμενου προγράμματος ($\hat{\beta}_1$=0.67).

Ακολουθώντας τη διαδικασία που παρουσιάσαμε νωρίτερα, μπορούμε να σχεδιάσουμε την εκτιμημένη ευθεία παλινδρόμησης πάνω στο διάγραμμα διασποράς (βλ. Σχήμα 2.7).

Παρατηρούμε ότι η προσαρμοσμένη ευθεία δεν περνάει κοντά από τα τέσσερα σημεία, τα οποία σχολιάσαμε νωρίτερα. Επιπλέον, φαίνεται ότι το σύνολο των άλλων 26 σημείων θα περιγράφεται καλύτερα από διαφορετική ευθεία με χαμηλότερη κλίση. Η παρουσία των τεσσάρων σημείων μπορεί να ασκήσει μεγάλη επιρροή στη θέση της ευθείας. Γεννιέται λοιπόν εύλογα το ερώτημα πόσο θα μεταβληθεί η ευθεία παλινδρόμησης, αν αφαιρεθούν αυτά τα σημεία; Πόσο επηρεάζει η αφαίρεση αυτών των σημείων την καταλληλότητα του μοντέλου;

Για να απαντήσουμε στα ερωτήματα αυτά, μπορούμε να αφαιρέσουμε τις παρατηρήσεις αυτές και να προσαρμόσουμε εκ νέου ένα γραμμικό μοντέλο. Οι παρατη-



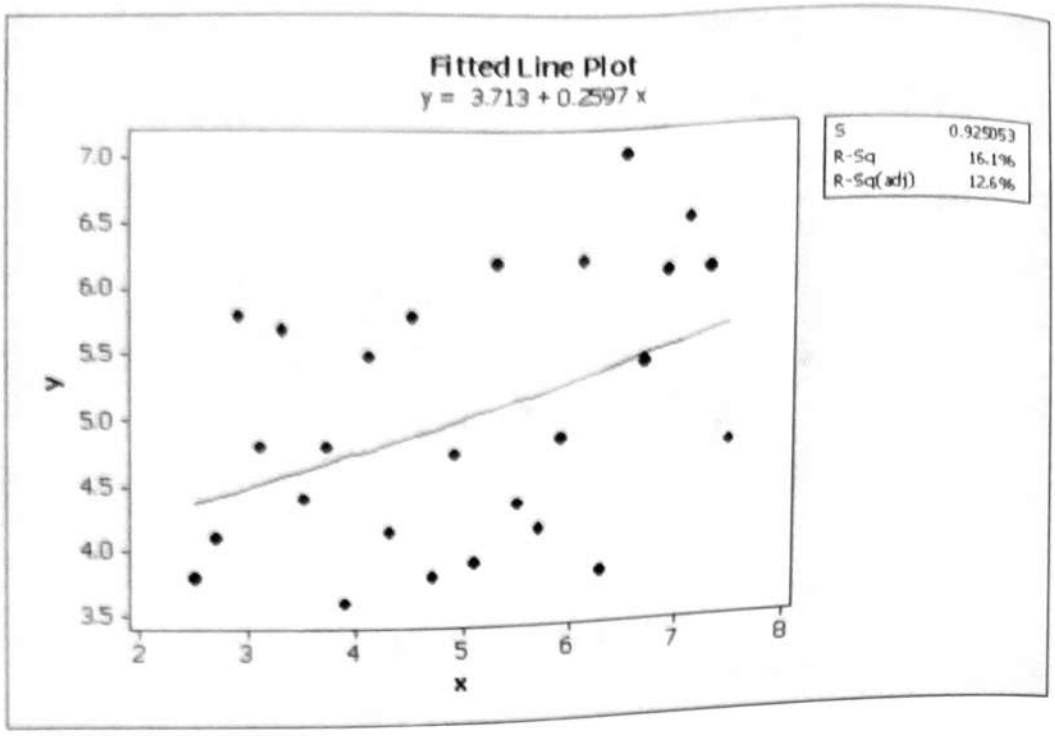
Σχήμα 2.7: Η προσαρμοσμένη ευθεία παλινδρόμησης για τα δεδομένα του Παραδείγματος 2.8.2

ρήσεις αυτές βρίσκονται στην $27^η$, $28^η$, $29^η$ και $30^η$ γραμμή των δεδομένων μας. (Σημείωση: στο MINITAB, αν πάτε με το βελάκι σε οποιοδήποτε σημείο ενός γραφήματος, θα δείτε σε ποια στατιστική μονάδα αντιστοιχεί).

Αφαιρώντας τις παρατηρήσεις αυτές λαμβάνουμε το διάγραμμα διασποράς του Σχήματος 2.8 με την ευθεία παλινδρόμησης και τα Αποτελέσματα 2.3 για την ανάλυση παλινδρόμησης.

Από τα αποτελέσματα αυτά παρατηρούμε ότι με την αφαίρεση των παρατηρήσεων αυτών, το μοντέλο που λαμβάνουμε, ναι μεν μπορεί να ικανοποιεί τις υποθέσεις των υπολοίπων, όπως φαίνεται με μια πρώτη ματιά στο διάγραμμα, αλλά η τιμή του συντελεστή R^2 είναι αισθητά μικρότερη, 16%.

Παρατηρούμε, δηλαδή, ότι η αφαίρεση των παρατηρήσεων αυτών επηρέασε σημαντικά την προσαρμογή του μοντέλου. Τέτοιες παρατηρήσεις ονομάζονται **σημεία επιρροής** *(influential observations). Στην Παράγραφο 4.6 αναπτύσσονται*



Σχήμα 2.8: Η προσαρμοσμένη ευθεία παλινδρόμησης για τα δεδομένα του Παραδείγματος 2.8.2 μετά την αφαίρεση 4 παρατηρήσεων

Αποτελέσματα 2.3 (Χωρίς τις παρατηρήσεις 27, 28, 29, 30)

```
Regression Analysis: y versus x

The regression equation is
y = 3.71 + 0.260 x

Predictor      Coef  SE Coef     T      P
Constant     3.7132   0.6314  5.88  0.000
X            0.2597   0.1209  2.15  0.042

S = 0.925053   R-Sq = 16.1%   R-Sq(adj) = 12.6%

Analysis of Variance

Source           DF       SS       MS     F      P
Regression        1   3.9442   3.9442  4.61  0.042
Residual Error   24  20.5373   0.8557
Total            25  24.4815
```

μέθοδοι εντοπισμού σημείων επιρροής, τα οποία στην περίπτωση πολλών επεξηγηματικών μεταβλητών δεν αποκαλύπτονται συνήθως από τα απλά διαγράμματα διασποράς.

2.9 Μη γραμμικό μοντέλο που μετατρέπεται σε γραμμικό

Πολλές φορές οι μεταβλητές για τις οποίες αναζητούμε κάποια σχέση μεταξύ τους δε συνδέονται γραμμικά, αλλά μέσω μίας μη γραμμικής συνάρτησης. Σε αρκετές από αυτές τις περιπτώσεις, αν καταφύγουμε σε κάποιο μετασχηματισμό της ανεξάρτητης ή ακόμα και της εξαρτημένης μεταβλητής, είναι δυνατόν να αποκτήσουμε την επιθυμητή γραμμικότητα.

Ένα χαρακτηριστικό παράδειγμα είναι, όταν η ανεξάρτητη μεταβλητή y συνδέεται με την εξαρτημένη μεταβλητή μέσω της σχέσης

$$y_i = \gamma e^{\beta_1 x_i} \varepsilon_i \,, \tag{2.5}$$

οπότε παίρνοντας λογάριθμους έχουμε ότι

$$\begin{aligned} \ln y_i &= \ln \gamma + \beta_1 x_i + \ln \varepsilon_i \\ &= \beta_0 + \beta_1 x_i + \varepsilon_i^* . \end{aligned}$$

Συνεπώς, αν τα ε_i^* ακολουθούν την Κανονική κατανομή, τότε μπορούμε να προσαρμόσουμε το απλό γραμμικό μοντέλο για τη «νέα» μεταβλητή μας $y^* = \ln y$ φυσικά σε σχέση με τη x.

Το ερώτημα είναι πώς γνωρίζουμε εκ των προτέρων τη σχέση (2.5); Αυτό μπορεί να γίνει είτε παρατηρώντας το διάγραμμα διασποράς των δεδομένων (έχοντας βέβαια και την κατάλληλη εμπειρία) είτε από παλαιότερες μελέτες πάνω στο συγκεκριμένο πρόβλημα. Μια τέτοια περίπτωση αφορά και το παράδειγμα που ακολουθεί.

Παράδειγμα 2.9.1. *Το Michaelis–Menten μοντέλο κινητικής χημείας σχετίζει την αρχική ταχύτητα y μιας αντίδρασης ενζύμων με τη συγκέντρωση x των ενζύμων αυτών. Το μοντέλο αυτό έχει ως εξής*

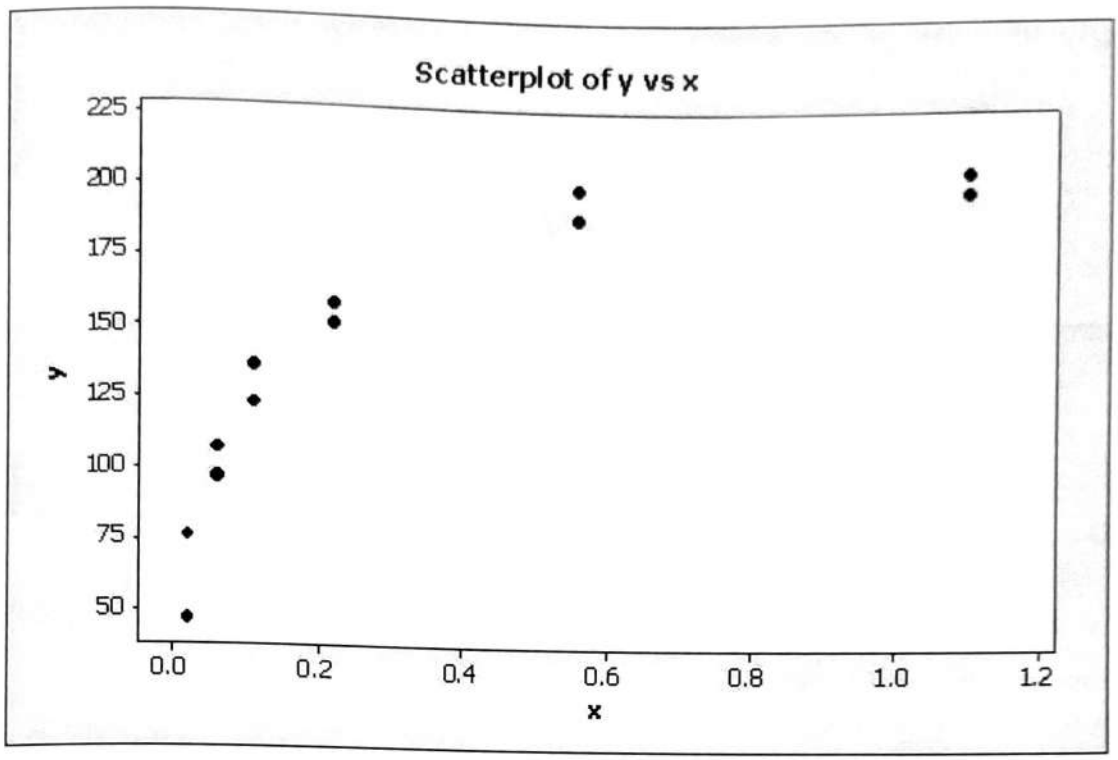
$$E(y_i) = \frac{\theta_1 x_i}{x_i + \theta_2}. \tag{2.6}$$

Τα δεδομένα ενός πειράματος για την εκτίμηση των παραμέτρων του μοντέλου Michaelis–Menten παρουσιάζονται στον Πίνακα 2.4, για τα οποία δεδομένα έχουμε το διάγραμμα διασποράς του Σχήματος 2.9.

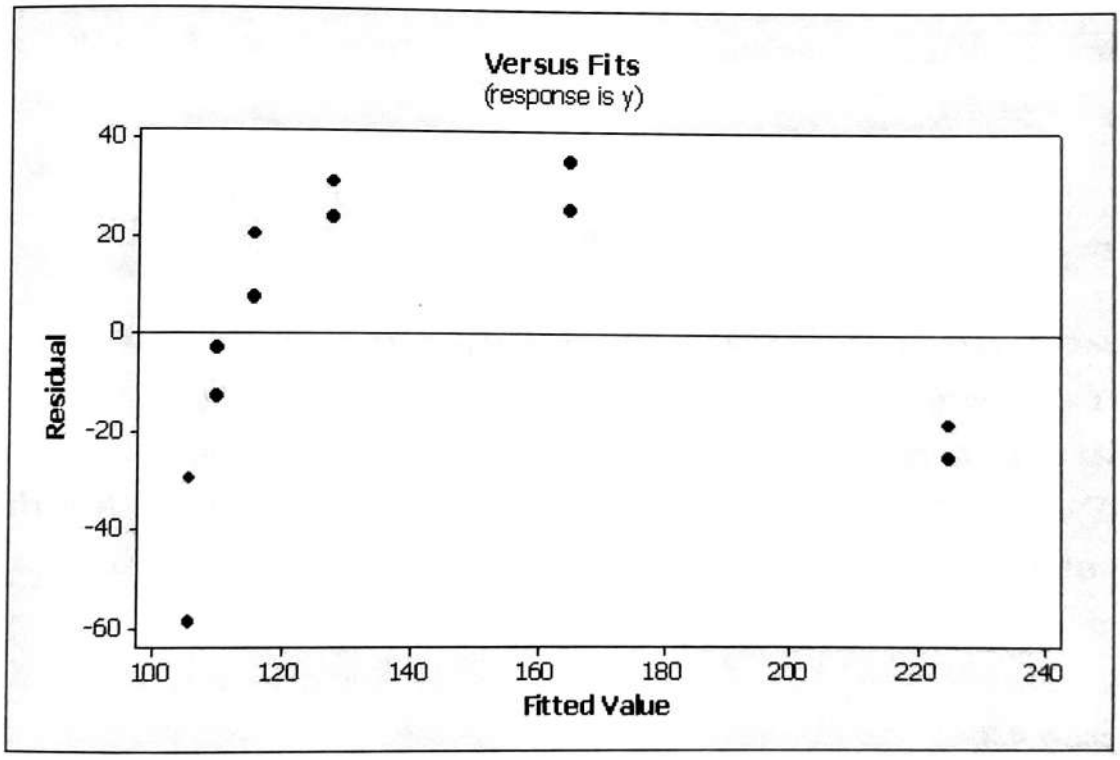
Πίνακας 2.4: Πίνακας δεδομένων για το Παράδειγμα 2.9.1

y	47	76	97	107	123	136
x	0.02	0.02	0.06	0.06	0.11	0.11
y	152	159	191	201	200	207
x	0.22	0.22	0.56	0.56	1.10	1.10

Όπως εύκολα μπορούμε να δούμε, από το διάγραμμα διασποράς δεν προκύπτει γραμμική σχέση μεταξύ των δύο μεταβλητών. Αν είχαμε παραλείψει την εξέταση του διαγράμματος διασποράς και είχαμε εφαρμόσει το απλό γραμμικό μοντέλο, το γράφημα **Residuals versus fits** *του* **Regression** *(βλ. Παράδειγμα 2.8.1) παίρνει τη μορφή του Σχήματος 2.10. Καταλαβαίνουμε από εκεί ότι τα υπόλοιπα e_i δεν κατανέμονται τυχαία, άρα παραβιάζονται οι υποθέσεις του μοντέλου. Επειδή το γραμμικό μοντέλο δεν περιγράφει τα δεδομένα μας, οδηγούμαστε στο να αναζητήσουμε κάποιο μετασχηματισμό των μεταβλητών, που θα γραμμικοποιήσει τη σχέση μεταξύ τους.*



Σχήμα 2.9: Διάγραμμα διασποράς για τα δεδομένα του Παραδείγματος 2.9.1



Σχήμα 2.10: Υπόλοιπα σε σχέση με τα εκτιμημένα $\hat{y}_i$ από την προσαρμογή απλού γραμμικού μοντέλου στα δεδομένα του Παραδείγματος 2.9.1

Εύκολα παρατηρούμε ότι αντιστρέφοντας το συστηματικό μέρος της σχέσης (2.6)

$$\begin{aligned}\frac{x_i+\theta_2}{\theta_1 x_i} &= \frac{1}{\theta_1}+\frac{\theta_2}{\theta_1}\frac{1}{x_i}\\ &= \beta_0+\beta_1 u_i\,,\end{aligned}$$

αποκτούμε το γραμμικό μοντέλο για $y_i^* = 1/y_i$,

$$y_i^* = \beta_0+\beta_1 u_i+\varepsilon_i\,,$$

όπου $u_i = 1/x_i$.

Εν συνεχεία, εφαρμόζοντας στα δεδομένα μας αυτό το απλό γραμμικό μόντέλο παίρνουμε την ακόλουθη ευθεία

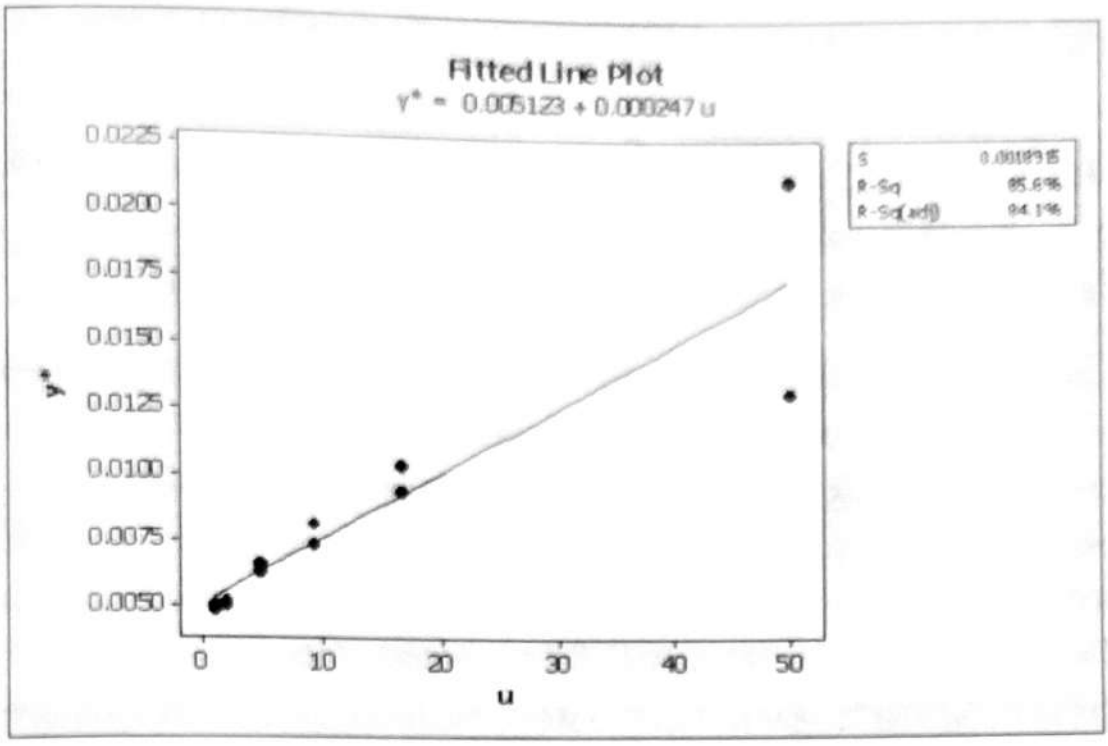
$$\hat{y}^* = 0.00512 + 0.000247u$$

που παρουσιάζεται στο διάγραμμα διασποράς των τ.μ. y^* και u στο Σχήμα 2.11. Βέβαια, αν θέλουμε, μπορούμε να υπολογίσουμε τις τιμές των θ_1 και θ_2, αφού γνωρίζουμε ότι

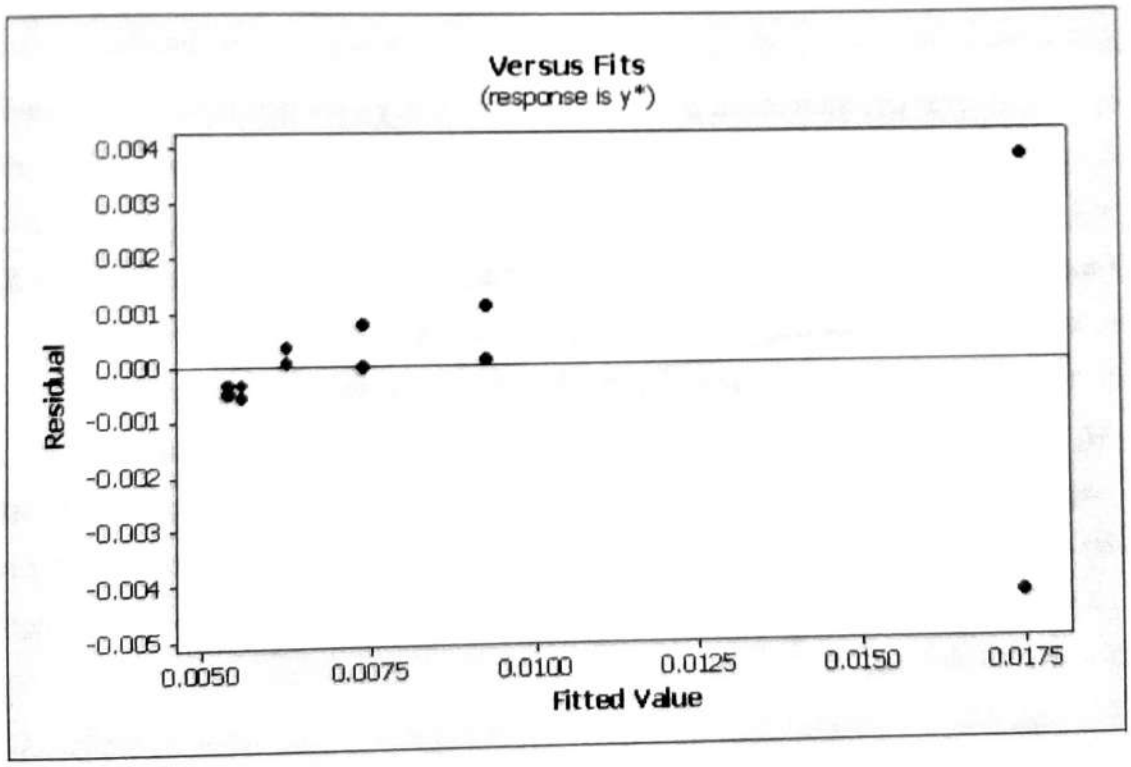
$$\beta_0 = 1/\theta_1 \quad \text{και} \quad \beta_1 = \theta_2/\theta_1\,,$$

συνεπώς $\hat{\theta}_1 = 195.313$ και $\hat{\theta}_2 = 0.0482$.

Επιπλέον, από τα αποτελέσματα της ανάλυσης παλινδρόμησης το μοντέλο μας έχει έναν υψηλό δείκτη προσαρμογής, δηλαδή της τάξεως του 85.6% και μια εξαιρετικά μικρή p–τιμή (<0.001) του F–ελέγχου. Ωστόσο παραμένει ένα πρόβλημα. Το Σχήμα 2.12 δείχνει το διάγραμμα που προκύπτει από την εντολή **Residuals versus fits**. Φαίνεται ότι η διασπορά των υπολοίπων είναι πολύ μεγαλύτερη για μεγάλες τιμές της $\hat{y}^*$ από ότι για μικρές. Αυτό παραβιάζει την προϋπόθεση της ομοσκεδαστικότητας. Μία λύση τέτοιων προβλημάτων παρουσιάζεται στην Παράγραφο 4.3.



Σχήμα 2.11: Η προσαρμοσμένη ευθεία παλινδρόμησης για τα μετασχηματισμένα δεδομένα του Παραδείγματος 2.9.1



Σχήμα 2.12: Υπόλοιπα σε σχέση με τα εκτιμημένα $\hat{y}_i^*$ του Παραδείγματος 2.9.1

2.10 Ο σταθερός όρος β_0

Στο Παράδειγμα 2.8.1 είχε γίνει ειδική αναφορά για την παρουσία του σταθερού όρου β_0 στο απλό γραμμικό μοντέλο (βλ. και Παραγράφους 3.2 και 3.5 στη συνέχεια για παρόμοια σχόλια σχετικά με το γενικό γραμμικό μοντέλο).

Επαναλαμβάνουμε εδώ ότι, κατά τη δική μας άποψη και σε συμφωνία με τη γενική πρακτική της Στατιστικής, ο όρος αυτός πρέπει να συμπεριληφθεί στο γραμμικό μοντέλο σχεδόν χωρίς εξαίρεση. Παρόλο που τα αποτελέσματα μιας ανάλυσης παλινδρόμησης περιλαμβάνουν έναν έλεγχο t για την υπόθεση $\beta_0 = 0$, συνήθως τον αγνοούμε και δεν το χρησιμοποιούμε σαν οδηγό για την ενδεχόμενη αφαίρεση της β_0 από το μοντέλο. Ο έλεγχος F, βέβαια, αφορά μόνο το συντελεστή β_1 (ή, στη γενική περίπτωση, όλους τους συντελεστές $\beta_1, \beta_2, \ldots, \beta_k$ των ανεξάρτητων μεταβλητών) και όχι τη σταθερά β_0.

Αν τελικά παραλειφθεί ο συντελεστής β_0 από το μοντέλο, αυτό θα πρέπει να γίνει μόνο λόγω της φύσης του πρακτικού ζητούμενου. Μια ερμηνεία του β_0 είναι ως η αναμενόμενη τιμή της y όταν η εξαρτημένη μεταβλητή $x = 0$. Μπορεί αρκετές φορές να φαίνεται ότι απαιτείται $y = 0$ όταν $x = 0$, π.χ., αν δεν εισάγεται καθόλου υλικό προς επεξεργασία ($x = 0$) και άρα λογικά η παραγωγή θα είναι επίσης μηδέν ($y = 0$). Ωστόσο αυτό το σκεπτικό ευσταθεί μόνο εφόσον η τιμή $x = 0$ περιλαμβάνεται στο εύρος των τιμών που εξετάζεται. Πιο συχνά, διαθέτουμε άλλο εύρος τιμών και εκεί προσαρμόζουμε το μοντέλο $\beta_0 + \beta_1 x$ για την περιγραφή των τιμών της y, γνωρίζοντας καλά ότι η όλη σχέση μεταξύ y και x ενδέχεται να έχει τη μορφή καμπύλης $f(x)$ που απλώς προσεγγίζεται από την ευθεία $\beta_0 + \beta_1 x$ στο περιορισμένο εύρος τιμών που αναλύεται. Σε μια τέτοια περίπτωση το να επιβάλουμε $\beta_0 = 0$ στο προσεγγιστικό μοντέλο $\beta_0 + \beta_1 x$ λόγω του ότι $f(0) = 0$, δεν είναι λογικό και θα περιόριζε σημαντικά τη δυνατότητα προσαρμογής του μοντέλου στα δεδομένα.

Αν παρόλα αυτά απαιτείται ένα μοντέλο με $\beta_0 = 0$, η προσαρμογή του είναι απλή (και γίνεται μέσα από το MINITAB, όπως είδαμε στο Παράδειγμα 2.8.1). Χωρίς

το β_0 η μόνη κανονική εξίσωση είναι

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

και συνεπώς

$$\hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$$

αντί των εξισώσεων (2.2), όπου περιλαμβάνεται και ο β_0 όρος.

Γενικώς, τα μοντέλα με και χωρίς το β_0 έχουν διαφορετική συμπεριφορά. Ένα παράδειγμα είναι το ότι ο συντελεστής προσδιορισμού R^2 δε θεωρείται καλό μέτρο για την προσαρμογή του μοντέλου, όταν από αυτό παραλείπεται η σταθερά β_0 (Eisenhauer, 2003).

Στο παρόν βιβλίο όλα τα γραμμικά μοντέλα περιλαμβάνουν το σταθερό όρο β_0 με ελάχιστες εξαιρέσεις κατά τις οποίες η παράλειψη δηλώνεται ρητά. Στα στατιστικά προγράμματα με χρήση υπολογιστή η παρουσία του β_0 στο μοντέλο είναι πάντοτε η προεπιλογή.