Ενότητα 5: Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)

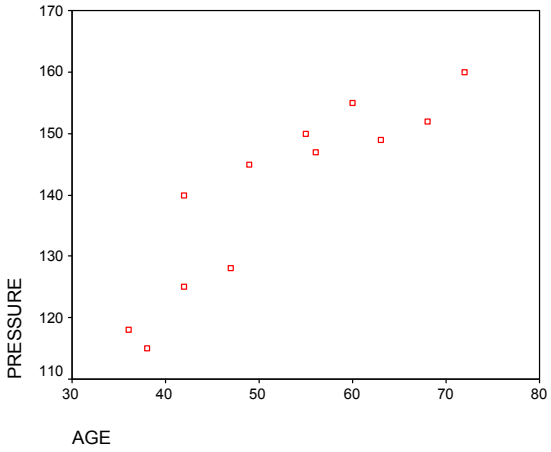
Κύριο πρόβλημα σε αυτή την ενότητα αποτελεί η διερεύνηση της σχέσης μεταξύ δυο (scaled) μεταβλητών X, Y (π.χ. X: ηλικία και Y: πίεση αίματος). Το γενικό πρόβλημα περιγράφεται ως εξής: από έναν (θεωρητικά άπειρο) πληθυσμό λαμβάνουμε ένα δείγμα μεγέθους n και για κάθε άτομο του δείγματος καταγράφουμε τις τιμές δύο μεταβλητών X, Y. Με βάση λοιπόν τα ζεύγη τιμών (X_1,Y_1), (X_2,Y_2), …, (X_n,Y_n) του δείγματος (π.χ. X_i: Ηλικία σε έτη i-ατόμου, Y_i: Πίεση αίματος i-ατόμου) επιθυμούμε να διερευνήσουμε τη σχέση μεταξύ των μεταβλητών X, Y. Επιπλέον θεωρούμε ότι

- Η μεταβλητή X η οποία καλείται *ανεξάρτητη* (independent) ή *ερμηνευτική μεταβλητή* (explanatory variable) δεν θεωρείται τυχαία, ενώ
- Η μεταβλητή Y η οποία καλείται *εξαρτημένη* (dependent) ή *μεταβλητή απόκρισης* (response variable) θεωρείται τυχαία μεταβλητή.

Παράδειγμα. Από $n = 12$ γυναίκες λαμβάνουμε τις ακόλουθες τιμές της πίεσης του αίματος και της αντίστοιχης ηλικίας σε έτη:

Ηλικία (X)	36	38	42	42	47	49	55	56	60	63	68	72
Πίεση αίματος (Y)	118	115	125	140	128	145	150	147	155	149	152	160

(εδώ $(X_1,Y_1) = (36, 118)$, $(X_2,Y_2) = (38, 115)$, κ.ο.κ.) Εισάγουμε τα δεδομένα στο SPSS σε δύο μεταβλητές – στήλες με $n = 12$ cases – γραμμές. Ονομάζουμε τις μεταβλητές Age (ή X) και Pressure (ή Y). Το πρώτο πράγμα που μπορούμε να κάνουμε είναι να δούμε τη «σχέση» των συγκεκριμένων μεταβλητών στο επίπεδο: Εκτελούμε Graphs/ Scatterplot / Simple/ Y Axis: Pressure, X Axis: Age λαμβάνοντας το ακόλουθο γράφημα



Παρατηρούμε ότι όσο αυξάνεται η X (Age) τόσο αυξάνεται και η Y (Pressure). Μάλιστα φαίνεται ότι τα σημεία (X_i,Y_i) βρίσκονται «κοντά» σε μία ευθεία, π.χ. την $y = b_0 + b_1x$, δηλαδή $Y_i \approx b_0 + b_1X_i$, $i = 1, 2, \ldots, n$ για κάποιες σταθερές b_0, b_1. Οι αποκλίσεις $Y_i - b_0 + b_1X_i$, $i = 1, 2,\ldots,n$ των σημείων (X_i,Y_i) από την ευθεία αυτή φαίνονται τυχαίες. Αν ονομάσουμε ε_i, $i = 1, 2,\ldots,n$ τις διαφορές αυτές τότε προκύπτει φυσιολογικά το γνωστό ως απλό γραμμικό μοντέλο που θα περιγράψουμε στη συνέχεια.

Επανερχόμενοι στην γενικότερη περίπτωση, επιθυμούμε να διερευνήσουμε τη σχέση μεταξύ των μεταβλητών X, Y. Θεωρούμε το απλούστερο μοντέλο που θα μπορούσε να ερμηνεύσει μια τέτοια σχέση (και που όπως είδαμε προέκυψε φυσιολογικά στο προηγούμενο παράδειγμα), το απλό γραμμικό μοντέλο. Σύμφωνα με το μοντέλο αυτό θεωρούμε ότι τα X_i, Y_i συνδέονται με τη σχέση

$$Y_i = b_0 + b_1X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

όπου b_0, b_1 είναι δυο άγνωστες σταθερές (καλούνται και *τεταγμένη* ή *intercept* και *κλίση* ή *slope* αντίστοιχα), ενώ οι ε_1, $\varepsilon,\ldots,\varepsilon_n$ είναι *ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν κανονική κατα-*

νομή $N(0,\sigma^2)$ (σ^2 άγνωστο) και συνήθως καλούνται «σφάλματα» των μετρήσεων. Μπορεί να θεωρηθεί ότι τα σφάλματα $\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_n$ εμπεριέχουν όλους τους άλλους παράγοντες (εκτός της X) επηρεάζουν την τιμή της μεταβλητής Y.

Υπογραμμίζεται και πάλι ότι οι τιμές $X_1,X_2,\ldots,X_n$ δεν είναι τυχαίες, αντίθετα με τις $Y_1, Y_2, \ldots, Y_n$ οι οποίες προφανώς είναι τυχαίες και μάλιστα θα ακολουθούν κανονική κατανομή (αφού είναι γραμμικές συναρτήσεις των κανονικών τ.μ. ε_i) με παραμέτρους

$$E(Y_i) = E(b_0 + b_1X_i + \varepsilon_i) = b_0 + b_1X_i + E(\varepsilon_i) = b_0 + b_1X_i,$$
$$V(Y_i) = V(b_0 + b_1X_i + \varepsilon_i) = V(\varepsilon_i) = \sigma^2$$

για $i = 1, 2, \ldots, n$, δηλαδή $Y_i \sim N(b_0+b_1X_i, \sigma^2)$. Επίσης οι τ.μ. $Y_1,Y_2,\ldots,Y_n$ είναι ανεξάρτητες αφού τα σφάλματα $\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_n$ είναι ανεξάρτητα (το «τυχαίο» ενός Y_i οφείλεται αποκλειστικά στο σφάλμα ε_i). Αρχικά, θα πρέπει με βάση τα (X_i,Y_i), i=1,2,…,n, να εκτιμήσουμε τις παραμέτρους b_0, b_1 και σ^2 ενώ φυσικά είναι απαραίτητο να διερευνήσουμε πόσο ικανοποιητικά προσαρμόζονται τα δεδομένα μας στο μοντέλο αυτό.

5.1. Εκτίμηση των παραμέτρων b_0, b_1 και σ^2

Εφόσον $Y_i \sim N(b_0+b_1X_i, \sigma^2)$, η από κοινού συνάρτηση πυκνότητας πιθανότητας των $Y_1, Y_2, \ldots, Y_n$, έστω $f_{\mathbf{Y}}(y_1,y_2,\ldots,y_n; b_0,b_1,\sigma^2)$, θα εξαρτάται από τις παραμέτρους b_0, b_1 και σ^2. Μάλιστα η συνάρτηση πιθανοφάνειας των $Y_1, Y_2, \ldots, Y_n$ θα είναι

$$L(b_0,b_1,\sigma^2) = f_{\mathbf{Y}}(y_1,y_2,...,y_n;b_0,b_1,\sigma^2) = \prod_{i=1}^{n} f_{Y_i}(y_i;b_0,b_1,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-b_0-b_1X_i)^2},$$

από όπου προκύπτει ότι οι εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέτρων (τιμές των παραμέτρων που μεγιστοποιούν την συνάρτηση πιθανοφάνειας) θα είναι:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2} \equiv \frac{S_{XY}}{S_{XX}}, \quad \hat{b}_0 = \overline{Y} - \hat{b}_1\overline{X}.$$

και

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{b}_0 - \hat{b}_1X_i)^2.$$

Από την μορφή της συνάρτησης πιθανοφάνειας είναι προφανές ότι οι εκτιμήτριες μέγιστης πιθανοφάνειας των b_1, b_2 προκύπτουν ισοδύναμα από την ελαχιστοποίηση (ως προς b_1, b_2) του αθροίσματος των τετραγώνων των σφαλμάτων,

$$\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2,$$

για αυτό και οι εκτιμήτριες των b_1, b_2 καλούνται και *εκτιμήτριες ελαχίστων τετραγώνων*. Επομένως, η *εκτιμημένη ευθεία γραμμικής παλινδρόμησης* θα είναι η

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

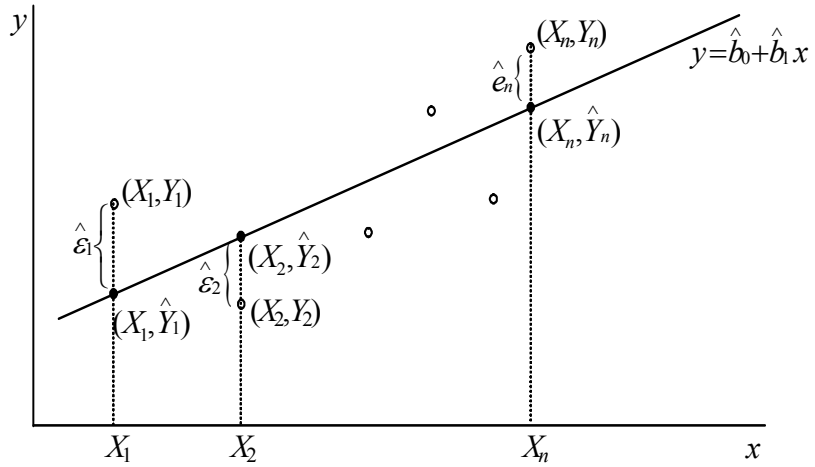
Προβλέψεις των Y_i (Y predicted) ή *προσαρμοσμένες* (πάνω στην εκτιμημένη ευθεία γραμμικής παλινδρόμησης) *τιμές των* Y_i καλούνται οι εκτιμήσεις των $E(Y_i) = b_0+b_1X_i$:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1X_i = \overline{Y} + \hat{b}_1(X_i - \overline{X})$$

ενώ οι διαφορές των προσαρμοσμένων $\hat{Y}_i$ από τις παρατηρούμενες Y_i καλούνται *κατάλοιπα* (residuals) ή *εκτιμημένα σφάλματα* και συμβολίζονται με

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = (Y_i - \overline{Y}) - \hat{b}_1(X_i - \overline{X}).$$

Οι παραπάνω ποσότητες φαίνονται και στο παρακάτω σχήμα,



Υπογραμμίζεται ότι η ευθεία στο παραπάνω σχήμα είναι η *εκτιμημένη* ευθεία γραμμικής παλινδρόμησης και τα κατάλοιπα είναι τα *εκτιμημένα* σφάλματα.

5.2. Έλεγχοι υποθέσεων και δ.ε. για τις παραμέτρους του μοντέλου.

Υποθέτοντας ότι τα σφάλματα είναι ανεξάρτητα και κανονικά ($\varepsilon_i \sim N(0,\sigma^2)$) αποδεικνύεται ότι

$$\hat{b}_1 \sim N(b_1, \frac{\sigma^2}{S_{XX}}), \quad \hat{b}_0 \sim N(b_0, \sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}\right)) \quad \text{με} \quad Cov(\hat{b}_0,\hat{b}_1) = -\sigma^2\frac{\overline{X}}{S_{XX}}$$

και

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \sum_{i=1}^{n}(\frac{Y_i-\hat{b}_0-\hat{b}_1 X_i}{\sigma})^2 \sim \chi^2_{n-2}$$

(χι τετράγωνο κατανομή με n–2 βαθμούς ελευθερίας). Επομένως, $E(\sum\hat{\varepsilon}_i^2) = \sigma^2(n-2)$ και ως εκτιμήτρια του σ^2 χρησιμοποιούμε την αμερόληπτη

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \equiv S^2$$

αντί της εκτιμήτριας μέγιστης πιθανοφάνειας που είδαμε παραπάνω (η μόνο διαφορά είναι ότι η ε.μ.π. διαιρεί το άθροισμα με n αντί n–2). Από τα παραπάνω προκύπτει ότι (υπό τις υποθέσεις του μοντέλου)

$$\frac{\hat{b}_0-b_0}{S\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}} \sim t_{n-2} \quad \text{και} \quad \frac{\hat{b}_1-b_1}{S\sqrt{\frac{1}{S_{XX}}}} \sim t_{n-2}$$

και επομένως τα παρακάτω είναι δ.ε. για τα b_0, b_1 αντίστοιχα, με σ.ε. 1–a:

$$(\hat{b}_0 - S\sqrt{\tfrac{1}{n}+\tfrac{\overline{X}^2}{S_{XX}}}t_{n-2}(\tfrac{a}{2}),\ \hat{b}_0 + S\sqrt{\tfrac{1}{n}+\tfrac{\overline{X}^2}{S_{XX}}}t_{n-2}(\tfrac{a}{2}))\,, \qquad (\hat{b}_1 - S\sqrt{\tfrac{1}{S_{XX}}}t_{n-2}(\tfrac{a}{2}),\ \hat{b}_1 + S\sqrt{\tfrac{1}{S_{XX}}}t_{n-2}(\tfrac{a}{2}))$$

ενώ για τον έλεγχο των υποθέσεων H_0: b_0=0 και H_0: b_1=0 θα έχουμε αντίστοιχες περιοχές απόρριψης (δίπλευροι έλεγχοι σε ε.σ. a):

$$K:\ |T_0| > t_{n-2}(\tfrac{a}{2}) \ \text{ και } \ K:\ |T_1| > t_{n-2}(\tfrac{a}{2})\,, \ \text{όπου} \ \ T_0 = \frac{\hat{b}_0}{S\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}}\,, \ \ T_1 = \frac{\hat{b}_1}{S\sqrt{\frac{1}{S_{XX}}}}$$

με αντίστοιχα p-value (αν από τα δεδομένα βρέθηκε ότι $T_0 = t_0,\ T_1 = t_1$)

$$p-value = P(|T_0| > |t_0|) = 2(1-F_{t_{n-2}}(|t_0|))\,, \quad p-value = P(|T_1| > |t_1|) = 2(1-F_{t_{n-2}}(|t_1|))\,.$$

Από τους δύο παραπάνω ελέγχους σημαντικότερος είναι ο έλεγχος για την «κλίση» της ευθείας γραμμικής παλινδρόμησης H_0: $b_1 = 0$. Αν απορριφθεί αυτή η υπόθεση τότε μπορούμε να πούμε ότι η μεταβλητή Y εξαρτάται από την X (αντίθετα, αν $b_1 = 0$ τότε η ευθεία παλινδρόμησης είναι παράλληλη με τον άξονα των x και επομένως όσο και αν μεταβάλλεται η X, δεν επηρεάζεται η Y).

5.3. Ερμηνεύοντας τη συνολική μεταβλητότητα του μοντέλου

Η δειγματική διασπορά των παρατηρήσεων Y_i αποδεικνύεται ότι χωρίζεται σε δύο αθροίσματα, συγκεκριμένα ισχύει ότι

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

Τα τρία αυτά αθροίσματα συμβολίζονται με *SST* (Sum of Squares Total), *SSE* (Sum of Squares Error) και *SSR* (Sum of Squares Regression) αντίστοιχα, δηλαδή,

$$SST = SSE + SSR.$$

Μπορεί τώρα να θεωρηθεί ότι

- το *SST* εκφράζει τη συνολική παρατηρούμενη μεταβλητότητα των Y_i,
- το *SSR* εκφράζει τη μεταβλητότητα των προσαρμοσμένων τιμών διότι

$$\overline{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n}Y_i = \overline{Y} \quad \text{και άρα} \quad \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{\hat{Y}})^2 .$$

Αυτή η μεταβλητότητα ερμηνεύεται από το μοντέλο αφού, σύμφωνα με αυτό, οι αναμενόμενες προσαρμοσμένες τιμές των Y_i είναι $b_0+b_1X_i$ και επομένως φυσιολογικά διαφέρουν από τον μέσο όρο τους (αφού τα X_i είναι διαφορετικά).

- Το *SSE* εκφράζει τη μεταβλητότητα των Y_i σε σχέση με τις αντίστοιχες προσαρμοσμένες τιμές $\hat{Y}_i$. Η μεταβλητότητα αυτή οφείλεται στην διασπορά σ^2 των σφαλμάτων ε_i τα οποία όπως είπαμε μπορεί να θεωρηθεί ότι «περιέχουν» όλους τους άλλους παράγοντες που επηρεάζουν την τιμή των Y_i (και δεν υπάρχουν στο μοντέλο).

Άρα τελικά παρατηρούμε ότι η συνολική παρατηρούμενη μεταβλητότητα των Y_i (SST) μπορεί να χωριστεί στα δύο, στην μεταβλητότητα που ερμηνεύεται από το μοντέλο (SSR) και στην μεταβλητότητα που οφείλεται σε παράγοντες που δεν έχουν περιληφθεί στο μοντέλο. Συνεπώς, το πηλίκο (συντελεστής προσδιορισμού)

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

μπορεί να θεωρηθεί ότι εκφράζει το ποσοστό της μεταβλητότητας των παρατηρήσεων που ερμηνεύεται από το μοντέλο. Είναι προφανές ότι όσο μεγαλύτερο (πιο «κοντά» στην μονάδα) είναι το R^2 τόσο καλύτερο είναι το μοντέλο που έχουμε θεωρήσει διότι ερμηνεύει μεγαλύτερο μέρος της παρατηρούμενης μεταβλητότητας.

Αξίζει να παρατηρήσουμε ότι

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}((Y_i - \overline{Y}) - \hat{b}_1(X_i - \overline{X}))^2 = S_{YY} + (\hat{b}_1)^2 S_{XX} - 2\hat{b}_1 S_{XY} = S_{YY} - \frac{S_{XY}^2}{S_{XX}}$$

και επομένως

$$SSR = SST - SSE = \frac{S_{XY}^2}{S_{XX}}$$

Επίσης, ο συντελεστής προσδιορισμού R^2 είναι ίσος με

$$R^2 = \frac{SSR}{SST} = \frac{S_{XY}^2}{S_{XX}S_{YY}}$$

που συμπίπτει με το τετράγωνο του δειγματικού συντελεστή συσχέτισης του Pearson (βλ. Εφαρμογή 3 στην παράγραφο 2.3).

Είδαμε στην προηγούμενη παράγραφο ότι

$$\frac{SSE}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(\frac{Y_i - \hat{b}_0 - \hat{b}_1 X_i}{\sigma})^2 \sim \chi^2_{n-2}.$$

Εξάλλου αν b_1=0 τότε $\hat{b}_1 / \sqrt{\frac{\sigma^2}{S_{XX}}} \sim N(0,1)$ από όπου προκύπτει ότι

$$\frac{SSR}{\sigma^2} = \frac{S^2_{XY}}{S_{XX}\sigma^2} = \frac{(\hat{b}_1)^2 S_{XX}}{\sigma^2} \sim \chi^2_1.$$

Επίσης αποδεικνύεται ότι οι δύο παραπάνω τυχαίες μεταβλητές (*SSE*, *SSR*) είναι ανεξάρτητες και επομένως (αν $b_1 = 0$)

$$\frac{SST}{\sigma^2} = \frac{SSE}{\sigma^2} + \frac{SSR}{\sigma^2} \sim \chi^2_{n-1}$$

κάτι που ήταν αναμενόμενο διότι αν $b_1 = 0$ τότε $Y_i \sim \mathrm{N}(b_0,\sigma^2)$ και σε αυτή την περίπτωση γνωρίζουμε ότι $\frac{1}{\sigma^2}\sum(Y_i - \overline{Y})^2 \sim \chi^2_{n-1}$. Ένα άλλο συμπέρασμα που προκύπτει από τα παραπάνω είναι ότι, αν $b_1 = 0$, τότε το πηλίκο

$$\frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{SSR}{SSE/(n-2)} \sim F_{1,n-2}$$

ακολουθεί κατανομή *F* (ή Snedecor) με 1 και $n - 2$ β.ε. (η $F_{n,k}$ ορίζεται ως η κατανομή του πηλίκου δύο ανεξάρτητων τυχαίων μεταβλητών που ακολουθούν χι-τετράγωνο κατανομή με *n* και *k* β.ε. αντίστοιχα, δια τους β.ε. τους). Από το παραπάνω γεγονός μπορούμε να κατασκευάσουμε έναν έλεγχο για την υπόθεση H_0: $b_1 = 0$. Θα απορρίπτεται η H_0 όταν η παραπάνω στατιστική συνάρτηση λαμβάνει μεγάλες τιμές, δηλαδή (ε.σ. α) όταν

$$\frac{SSR}{SSE/(n-2)} > F_{1,n-2}(a) : \textit{άνω a-σημείο της κατανομής } F \textit{ με } 1 \textit{ και } n-2 \text{ β.ε.}$$

με αντίστοιχο p-value:

$$p - value = 1 - F_{F_{1,n-2}}(\frac{SSR}{SSE/(n-2)})$$

(όπου $F_{F_{1,n-2}}$ είναι η σ.κ. της κατανομής $F_{1,n-2}$). Είναι εύκολο να επαληθεύσουμε ότι ο παραπάνω έλεγχος της H_0: $b_1 = 0$ είναι ισοδύναμος με τον έλεγχο που είδαμε στην προηγούμενη παράγραφο για την ίδια υπόθεση χρησιμοποιώντας την στατιστική συνάρτηση T_1 (διαφορά των δύο αυτών ελέγχων υπάρχει όταν εφαρμόζουμε πολλαπλό γραμμικό μοντέλο).

Όλες οι παραπάνω ποσότητες συνοψίζονται σε έναν πίνακα που είναι γνωστός ως πίνακας ανάλυσης διασποράς (ANOVA):

Model	***SS***	***df***	***MS***	***F***	***Sig.* (p-value)**
Regression	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	1	$MSR = SSR$	$\frac{MSR}{MSE}$	$1 - F_{F_{1,n-2}}(\frac{MSR}{MSE})$
Residuals	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2} = S^2$		
Total	$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$			

5.4. Ατομική και μέση πρόβλεψη της Y

Αφού έχουμε εκτιμήσει τους συντελεστές b_0, b_1 μέσω των $\hat{b}_0, \hat{b}_1$, λαμβάνουμε μια εκτίμηση της ευθείας γραμμικής παλινδρόμησης:

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

και μέσω αυτής μπορούμε να κάνουμε *πρόβλεψη* (prediction) του Y που αντιστοιχεί σε οποιοδήποτε x_0:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 x_0 .$$

Εδώ χρησιμοποιούμε το όρο πρόβλεψη και όχι εκτίμηση γιατί η Y που θέλουμε να «προσδιορίσουμε» είναι τυχαία μεταβλητή και όχι παράμετρος (δηλ. σταθερά). Προφανώς οι προβλέψεις της μεταβλητής Y στα σημεία $X_1, X_2, \ldots, X_n$ είναι οι γνωστές *προσαρμοσμένες τιμές των* Y_i (ή *προβλέψεις των* Y_i)

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i = \overline{Y} + \hat{b}_1 (X_i - \overline{X}) \text{ , } i = 1, 2, \ldots, n$$

Είναι φανερό ότι η πρόβλεψη $\hat{Y} = \hat{b}_0 + \hat{b}_1 x_0$ είναι μια σημειακή πρόβλεψη. Μερικές φορές όμως είναι προτιμότερο να προβλέψουμε ένα Y χρησιμοποιώντας όχι ένα σημείο αλλά ένα διάστημα. Συνήθως χρησιμοποιούμε δυο τέτοια διαστήματα:

(1) Το *διάστημα μέσης πρόβλεψης* (*mean prediction interval*) *του Y στο* x_0, το οποίο είναι ένα δ.ε. (συντ. $1-a$) για το $E(Y) = b_0 + b_1 x_0$:

$$\left(\hat{b}_0 + \hat{b}_1 x_0 - S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{X})^2}{S_{XX}}} t_{n-2}(\tfrac{a}{2}),\ \hat{b}_0 + \hat{b}_1 x_0 + S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{X})^2}{S_{XX}}} t_{n-2}(\tfrac{a}{2}) \right)$$

διότι είναι εύκολο να δούμε ότι $\hat{b}_0 + \hat{b}_1 x_0 \sim N(b_0 + b_1 x_0, \sigma^2 \left(\frac{1}{n} + \frac{1}{S_{XX}} (x_0 - \overline{X})^2 \right))$, ενώ τα $\hat{b}_0, \hat{b}_1$ είναι ανεξάρτητα από το SSE. Από τα παραπάνω προκύπτει ότι αν πάρουμε έναν μεγάλο αριθμό παρατηρήσεων με $X = x_0$ τότε η μέση τιμή της μεταβλητής Y σε αυτές τις παρατηρήσεις θα βρίσκεται μέσα στο διάστημα μέσης πρόβλεψης με σ.ε. 95%.

(2) Το *διάστημα ατομικής πρόβλεψης* (*individual prediction interval*) *του Y στο* x_0, το οποίο είναι ένα διάστημα μέσα στο οποίο βρίσκεται η $Y = b_0 + b_1 x_0 + \varepsilon$ με πιθανότητα $1-a$:

$$\left(\hat{b}_0 + \hat{b}_1 x_0 - S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{X})^2}{S_{XX}}} t_{n-2}(\tfrac{a}{2}),\ \hat{b}_0 + \hat{b}_1 x_0 + S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{X})^2}{S_{XX}}} t_{n-2}(\tfrac{a}{2}) \right)$$

διότι είναι εύκολο να δούμε ότι $\hat{b}_0 + \hat{b}_1 x_0 + \varepsilon \sim N(b_0 + b_1 x_0, \sigma^2 \left(1 + \frac{1}{n} + \frac{1}{S_{XX}} (x_0 - \overline{X})^2 \right))$. Από τα παραπάνω προκύπτει ότι αν πάρουμε μία νέα παρατήρηση με (X,Y) με $X = x_0$ τότε το Y θα βρίσκεται μέσα στο διάστημα ατομικής πρόβλεψης με σ.ε. 95%.

5.5. Εξέταση της ορθότητας του μοντέλου.

Όλα τα παραπάνω έγιναν υπό τις υποθέσεις του γραμμικού μοντέλου:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i , \quad i = 1, 2, \ldots, n$$

όπου τα σφάλματα $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ είναι ανεξάρτητα και ακολουθούν κανονική κατανομή $N(0,\sigma^2)$. Είναι σημαντικό πριν κλείσουμε την ανάλυση (ή καλύτερα πριν την αρχίσουμε) να βεβαιωθούμε ότι οι παρατηρήσεις μας προσαρμόζονται ικανοποιητικά στο παραπάνω μοντέλο ώστε τα συμπεράσματα που προκύπτουν να θεωρούνται αξιόπιστα. Αν διαπιστώσουμε ότι κάτι τέτοιο δεν συμβαίνει τότε θα πρέπει να τροποποιήσουμε κατάλληλα το μοντέλο. Συνήθεις αποκλίσεις που παρατηρούνται είναι:

(1) *Τα σφάλματα δεν είναι κανονικά*

(2) *Τα σφάλματα δεν έχουν σταθερή διασπορά* σ^2

(3) *Τα σφάλματα δεν είναι ανεξάρτητα*

Επειδή τα σφάλματα δεν είναι γνωστά, εξετάζουμε τα παραπάνω χρησιμοποιώντας τα κατάλοιπα. Τα κατάλοιπα δεν είναι ανεξάρτητα, αλλά για μεγάλα δείγματα μπορούν πρακτικά να θεωρηθούν ανεξάρτητα διότι η συνδιασπορά τους είναι της τάξης του $1/n$ (επίσης είναι ανεξάρτητα των προβλέψεων των Y_i). Επίσης τα κατάλοιπα δεν έχουν ούτε σταθερή διασπορά γιατί

$$V(\hat{\varepsilon}_i) = \sigma^2(1-p_{ii}) \quad \text{όπου} \quad p_{ii} = \frac{(X_i - \bar{X})^2}{S_{XX}} + \frac{1}{n},$$

όπου οι ποσότητες p_{ii} καλούνται μόχλευση (leverage). Για το λόγο αυτό βασιζόμαστε στα τυποποιημένα κατάλοιπα (studentized residuals):

$$\hat{\varepsilon}_i^* = \frac{\hat{\varepsilon}_i}{S\sqrt{1-p_{ii}}},\ i = 1,2,\ldots,n. \qquad (S^2 = \hat{\sigma}^2 = SSE/(n-2)))$$

(μερικές φορές χρησιμοποιούμε τα λεγόμενα *κανονικοποιημένα κατάλοιπα – standardized residuals* – τα οποία είναι τα $\hat{\varepsilon}_i / S$). Για μεγάλα λοιπόν δείγματα μπορεί να θεωρηθεί ότι τα κατάλοιπα έχουν την ίδια συμπεριφορά με τα σφάλματα.

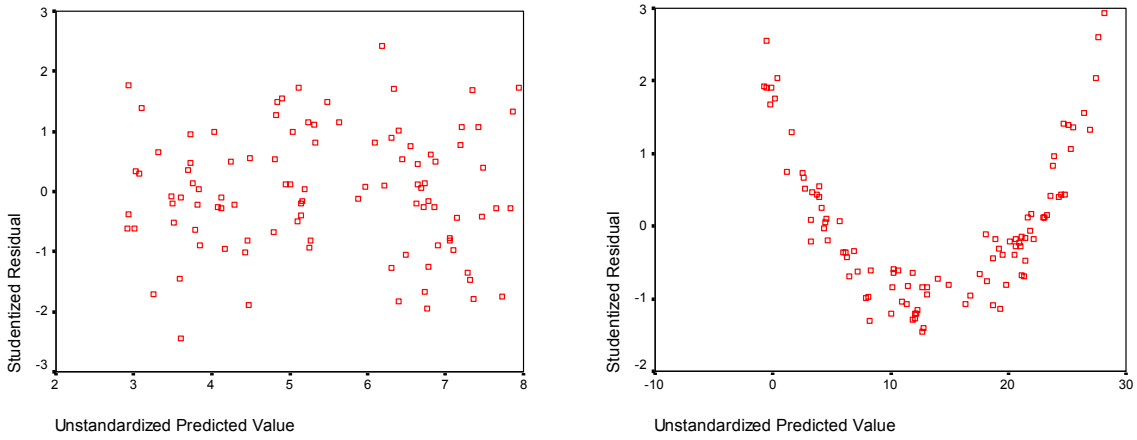
Για να διερευνήσουμε αν το μοντέλο είναι σωστό (δεν ισχύει κάποια από τις παραπάνω αποκλίσεις) συνήθως προχωράμε στους παρακάτω ελέγχους:

(i) Εξετάζουμε αν τα τυποποιημένα κατάλοιπα ακολουθούν πράγματι κανονική κατανομή (χρησιμοποιούμε ιστόγραμμα, Q-Q ή P-P plots και K-S τεστ).

(ii) Εξετάζουμε αν υπάρχει σχέση μεταξύ των προσαρμοσμένων Y_i και των τυποποιημένων καταλοίπων (υπό τις υποθέσεις του γραμμικού μοντέλου είναι ανεξάρτητα), χρησιμοποιώντας το γράφημα των σημείων

$$(\hat{Y}_i, \hat{\varepsilon}_i^*),\ i = 1, 2, \ldots, n$$

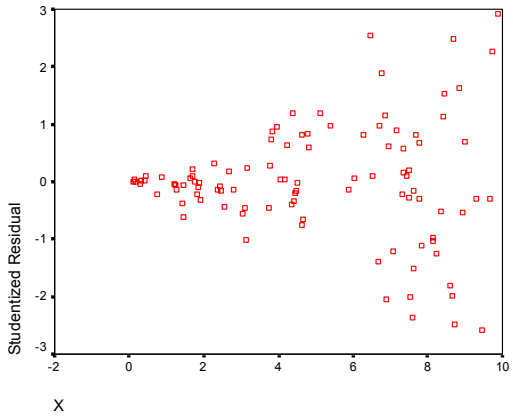
στο επίπεδο. Αν βρεθεί ότι υπάρχει σχέση (όπως π.χ. στο δεξιό γράφημα παρακάτω όπου τα σημεία δεν φαίνεται να βρίσκονται «τυχαία» στο επίπεδο, αντίθετα με το αριστερό γράφημα) τότε θα πρέπει να εκτελέσουμε κατάλληλο μετασχηματισμό (των Y_i ή των X_i) ώστε να εξαλειφθεί αυτή η σχέση (ο μετασχηματισμός αυτός δεν είναι πάντοτε εύκολο να προσδιοριστεί).



(iii) Εξετάζουμε αν υπάρχει σχέση μεταξύ των X_i και των τυποποιημένων καταλοίπων, χρησιμοποιώντας το γράφημα των σημείων

$$(X_i, \hat{\varepsilon}_i^*),\ i = 1, 2, \ldots, n$$

στο επίπεδο. Αν βρεθεί ότι υπάρχει σχέση (κάτι που προκύπτει π.χ. όταν η διασπορά των σφαλμάτων δεν είναι σταθερή) θα πρέπει και πάλι να εκτελέσουμε κατάλληλο μετασχηματισμό (των Y_i ή των X_i) ώστε να εξαλειφθεί αυτή η σχέση. Αν π.χ. φαίνεται ότι η διασπορά των καταλοίπων αυξάνεται με το X, όπως π.χ. στο παρακάτω γράφημα,



τότε προχωράμε σε μια τεχνική που σταθεροποιεί τη διασπορά των σφαλμάτων. Μπορούμε να θεωρήσουμε ότι $V(\varepsilon_i) = \sigma^2 X_i^2$ και αντί του μοντέλου $Y = b_0 + b_1 X + \varepsilon$, μπορούμε να θεωρήσουμε το μοντέλο (διαιρούμε και τα δύο μέλη με X)

$$\frac{Y}{X} = b_0 \frac{1}{X} + b_1 + \frac{\varepsilon}{X} \quad \Leftrightarrow Y' = b_1 + b_0 X' + \varepsilon', \quad \text{όπου } Y' = \frac{Y}{X}, X' = \frac{1}{X}, \varepsilon' = \frac{\varepsilon}{X}$$

όπου τώρα $V(\varepsilon') = \sigma^2$.

(iv) Εξετάζουμε αν τα τυποποιημένα κατάλοιπα είναι ανεξάρτητα από την σειρά με την οποία πήραμε τις παρατηρήσεις (επαναλαμβάνουμε ότι υπό τις υποθέσεις του γραμμικού μοντέλου και για μεγάλα δείγματα θα πρέπει πρακτικά να είναι ανεξάρτητα). Για το σκοπό αυτό χρησιμοποιούμε το γράφημα των σημείων

$$(i, \hat{\varepsilon}_i^*),\ i = 1, 2, \ldots, n \qquad \text{ή το γράφημα των} \qquad (\hat{\varepsilon}_i^*, \hat{\varepsilon}_{i+1}^*),\ i = 1, 2, \ldots, n-1.$$

Επίσης συνήθως χρησιμοποιούμε ένα τεστ ροών (runs test) για τα κατάλοιπα (το οποία εξετάσαμε σε προηγούμενη ενότητα) ή ένα τεστ αυτοπαλινδρόμησης που είναι γνωστό ως Durbin –Watson test. Σύμφωνα με το τεστ αυτό θεωρούμε ότι $\varepsilon_i = \rho\varepsilon_{i-1} + u_i$, ($|\rho|<1$, $u_i \sim N(0,\sigma^2)$) δηλαδή τα σφάλματα ακολουθούν ένα AR (Auto Regressive) μοντέλο και ελέγχουμε αν H_0: $\rho = 0$ (ανεξάρτητα σφάλματα) έναντι της H_1: $\rho > 0$ (θετικά εξαρτημένα σφάλματα). Για τον έλεγχο αυτό χρησιμοποιούμε την στατιστική συνάρτηση:

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

της οποίας η κατανομή (υπό την H_0) έχει μελετηθεί. Απορρίπτεται η H_0: $\rho = 0$ όταν η d λαμβάνει τιμές «κοντά» στο 0. Τα στατιστικά πακέτα συνήθως δίνουν αυτόματα την τιμή του p-value που αντιστοιχεί σε αυτό το τεστ.

(v) Εξετάζουμε αν υπάρχουν «έκτροπες» παρατηρήσεις χρησιμοποιώντας και πάλι τα γραφήματα

$$(\hat{Y}_i, \hat{\varepsilon}_i^*),\ i = 1, 2, \ldots, n \quad \text{και} \quad (X_i, \hat{\varepsilon}_i^*),\ i = 1, 2, \ldots, n$$

(ακόμη και το γράφημα των (X_i, Y_i)). Θεωρούμε ως «ασυνήθιστες» τις παρατηρήσεις με studentized residual μεγαλύτερο του 2 και «έκτροπες» αυτές με studentized residual μεγαλύτερο του 3. Οι έκτροπες παρατηρήσεις είτε προέρχονται από λάθος καταγραφή του ερευνητή (οπότε ελέγχεται αν

μια έκτροπη παρατήρηση έχει καταγραφεί και περαστεί στον Η/Υ σωστά) ή είναι πραγματικές παρατηρήσεις υποδεικνύοντας ότι το μοντέλο μας δεν είναι απόλυτα σωστό.

Ιδιαίτερη προσοχή θα πρέπει να δοθεί στις παρατηρήσεις που έχουν μεγάλη «επιρροή» στο μοντέλο (παρατηρήσεις που αν ληφθούν υπόψη αλλάζουν σημαντικά την εκτίμηση της ευθείας γραμμικής παλινδρόμησης). Τέτοιες παρατηρήσεις είναι αυτές που έχουν X_i αρκετά μακριά από τα υπόλοιπα $X_j, j \neq i$ ή πιο απλά έχουν X_i αρκετά μακριά από το $\overline{X}$. Η «απόσταση» αυτή συνήθως μετράται χρησιμοποιώντας μια ποσότητα που έχει εμφανιστεί και παραπάνω, την μόχλευση (leverage)

$$p_{ii} = \frac{(X_i - \overline{X})^2}{S_{XX}} + \frac{1}{n}$$

ή την λεγόμενη «κεντρική μόχλευση» (centered leverage) που είναι η παραπάνω ποσότητα μείον το $1/n$. Επειδή $V(\hat{\varepsilon}_i) = \sigma^2(1 - p_{ii})$ παρατηρήσεις με μεγάλη μόχλευση θα δίνουν μικρό κατάλοιπο (μη τυποποιημένο). Αυτό συνηγορεί στο γεγονός ότι επηρεάζουν σημαντικά την ευθεία γραμμικής παλινδρόμησης αφού την «αναγκάζουν» να περάσει «κοντά» τους. Αποδεικνύεται ότι παρατηρήσεις με μεγάλη μόχλευση (συνήθως με $p_{ii} > 3 \cdot 2/n$) επηρεάζουν σημαντικά το μοντέλο και επομένως θα πρέπει ή να λαμβάνονται με μεγάλη προσοχή ή να εξαιρούνται του μοντέλου.

Ένας ακόμη τρόπος να εντοπίσουμε έκτροπες παρατηρήσεις βασίζεται στα λεγόμενα διαγραμμένα κατάλοιπα (deleted residuals)

$$\hat{\varepsilon}_{(i)} = Y_i - \hat{Y}_i^* = \frac{\hat{\varepsilon}_i}{1 - p_{ii}}$$

όπου $\hat{Y}_i^*$ είναι η προσαρμοσμένη τιμή της Y_i αν εξαιρέσουμε από τα δεδομένα το ζεύγος (X_i, Y_i).

Όλοι οι παραπάνω έλεγχοι (κυρίως αυτοί που βασίζονται σε γραφήματα) καθώς και προτάσεις για «διόρθωση» των όποιων αποκλίσεων παρατηρηθούν απαιτούν ιδιαίτερη εμπειρία στην ανάλυση καταλοίπων από τον εκάστοτε ερευνητή (που δεν είναι δυνατό να αποκτηθεί στα πλαίσια ενός προπτυχιακού ή ακόμη και μεταπτυχιακού μαθήματος).

5.6. Μετασχηματισμοί.

Αρκετές φορές συμβαίνει οι μεταβλητές X και Y να μην έχουν γραμμική σχέση, κάτι που μπορεί άμεσα να φανεί από το διάγραμμα διασποράς ή από κάποιο γράφημα καταλοίπων. Σε αυτές τις περιπτώσεις δεν μπορούμε να εφαρμόσουμε απευθείας το γραμμικό μοντέλο αλλά θα πρέπει να μετασχηματίσουμε τα δεδομένα $X' = f(X)$ και $Y' = g(Y)$ έτσι ώστε οι X', Y' να έχουν γραμμική σχέση. Συνήθως χρησιμοποιούμε τους μετασχηματισμούς

$$Y' = \sqrt{Y},\ Y' = \ln Y,\ Y' = 1/Y, \quad X' = \sqrt{X},\ X' = \ln X,\ X' = 1/X$$

Ορισμένες σχετικές επισημάνσεις είναι οι ακόλουθες:
(1) Μερικές φορές για την εύρεση του κατάλληλου μετασχηματισμού λαμβάνονται υπόψη και διάφορες a-priori υποθέσεις. Για παράδειγμα, αν X είναι η τιμή ενός αγαθού και Y είναι η ζήτησή του, συχνά προτείνεται ένας λογαριθμικός μετασχηματισμός και για τις δύο μεταβλητές ώστε να επιτευχθεί γραμμικότητα, διότι με τον μετασχηματισμό αυτό, το b_1 εκφράζει την ποσοστιαία αλλαγή στην ζήτηση για κάθε 1% αλλαγής στην τιμή.
(2) Μετασχηματισμοί του X δεν επηρεάζουν την μεταβλητότητα των σφαλμάτων, ενώ αντίθετα μετασχηματισμοί του Y την επηρεάζουν. Μερικές φορές το Y μετασχηματίζεται για αυτόν ακριβώς τον λόγο, αν τα σφάλματα φαίνεται ότι δεν έχουν την ίδια διασπορά. Μετά τον μετασχηματισμό είναι χρήσιμη η εξέταση της διασποράς των καταλοίπων.
(3) Συνήθως προτιμούμε να μετασχηματίζουμε την μεταβλητή που έχει την μεγαλύτερη μεταβλητότητα στις τιμές.

5.7. Ασκήσεις - Παραδείγματα

Άσκηση 1 (συνέχεια προηγούμενου παραδείγματος). Από n = 12 γυναίκες λαμβάνουμε τις ακόλουθες τιμές της πίεσης του αίματος και της αντίστοιχης ηλικίας σε έτη:

Ηλικία (X)	36	38	42	42	47	49	55	56	60	63	68	72
Πίεση αίματος (Y)	118	115	125	140	128	145	150	147	155	149	152	160

1. Να γίνει το διάγραμμα διασποράς (scatterplot) μεταξύ των X, Y. Δικαιολογείται από το γράφημα η εφαρμογή ενός γραμμικού μοντέλου;

2. Να κατασκευάσετε το διάγραμμα διασποράς των δεδομένων (X,Y) μαζί με την εκτιμημένη ευθεία γραμμικής παλινδρόμησης και τις ζώνες εμπιστοσύνης (για την ατομική και μέση πρόβλεψη) με σ.ε. 95%. Ποια είναι η φυσική ερμηνεία των b_0, b_1 στο σχήμα;

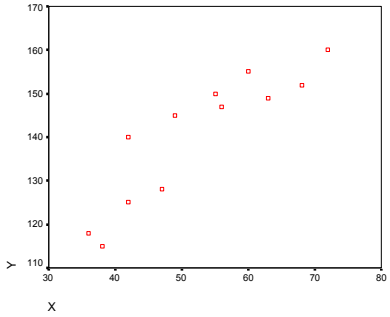
3. Να κάνετε μελέτη του μοντέλου $Y = b_0 + b_1X + \varepsilon$. Συγκεκριμένα: (α) Να εκτιμήσετε τα b_0, b_1 σημειακά και με δ.ε. συντελεστού 95%. (β) Να ελέγξετε (σε ε.σ. 5%) αν H_0: b_1=0, H_1: $b_1\neq0$, και H_0: b_0=0, H_1: $b_0\neq0$. Η μεταβλητή Y εξαρτάται από την X; (γ) Να κατασκευάσετε τον πίνακα ανάλυσης διασποράς (ANOVA) και να κάνετε τον έλεγχο H_0: b_1=0, H_1: $b_1\neq0$ του μοντέλου μέσω του F-τεστ. Ποια είναι η εκτίμηση της διασποράς των σφαλμάτων; (δ) Τι ποσοστό της μεταβλητότητας των Y_i ερμηνεύεται από το μοντέλο;

4. Να δοθούν οι προσαρμοσμένες τιμές των Y_i (προβλέψεις των Y_i) και τα κατάλοιπα.

5. Ποια είναι η πρόβλεψη της πίεσης του αίματος για γυναίκα ηλικίας x_0=52 ετών: (α) Να γίνει σημειακή πρόβλεψη και να δοθούν τα διαστήματα ατομικής και μέσης πρόβλεψης (95%). (β) Εάν επιλεγεί τυχαία μια γυναίκα 52 ετών από τον πληθυσμό, μεταξύ ποιών ορίων θα βρίσκεται η πίεση του αίματός της (σ.ε. 95%). (γ) Εάν επιλέξουμε τυχαία έναν μεγάλο αριθμό από γυναίκες ηλικίας 52 ετών, μεταξύ ποιών ορίων θα βρίσκεται η μέση πίεση του αίματός τους (σ.ε. 95%).

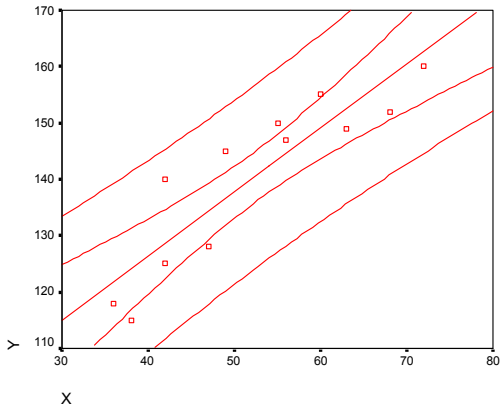
6. Να γίνει έλεγχος ορθότητας του μοντέλου: (α) Εξετάστε αν τα τυποποιημένα κατάλοιπα $\hat{\varepsilon}_i^*$ προέρχονται πράγματι από κανονική κατανομή (ιστόγραμμα, Q-Q ή P-P plots και K-S τεστ). (β) Εξετάστε αν υπάρχει σχέση μεταξύ των προσαρμοσμένων Y_i και των τυποποιημένων καταλοίπων, χρησιμοποιώντας το γράφημα των σημείων $(\hat{Y}_i, \hat{\varepsilon}_i^*)$, i = 1, 2, …, n στο επίπεδο. Υπάρχουν «ασυνήθιστες» ($|\hat{\varepsilon}_i^*|>2$) ή «έκτροπες» παρατηρήσεις ($|\hat{\varepsilon}_i^*|>3$); (γ) Εξετάστε αν τα τυποποιημένα κατάλοιπα είναι ανεξάρτητα από την σειρά με την οποία πήραμε τις παρατηρήσεις χρησιμοποιώντας το γράφημα των $(i, \hat{\varepsilon}_i^*)$, i = 1, 2, …, n. Επίσης, να εκτελέστε και ένα τεστ ροών για το σκοπό αυτό. (δ) Εξετάστε αν υπάρχουν παρατηρήσεις που έχουν μεγάλη «επιρροή» στο μοντέλο (παρατηρήσεις που αν ληφθούν υπόψη αλλάζουν σημαντικά την εκτίμηση της ευθείας γραμμικής παλινδρόμησης συνήθως θεωρούνται αυτές που έχουν centered leverage > 5/n).

Λύση. 1. Αρχικά εισάγουμε τα δεδομένα στο SPSS σε δύο μεταβλητές (στήλες) X, Y και λαμβάνουμε το διάγραμμα διασποράς (Graphs/scatter/simple/Y axis:Y, X axis:X) για να πάρουμε μια αρχική εικόνα για τη σχέση μεταξύ των μεταβλητών

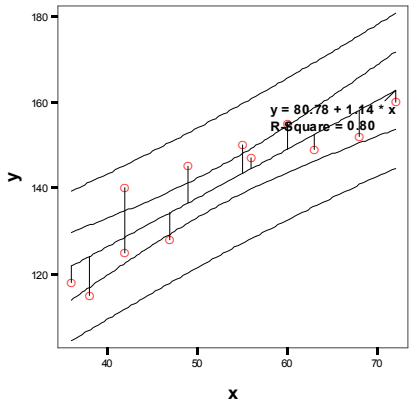


Φαίνεται να υπάρχει γραμμική σχέση μεταξύ των δυο μεταβλητών και επομένως η εφαρμογή του μοντέλου $Y_i = b_0 + b_1 X_i + \varepsilon_i,\ \ i = 1, 2, \ldots, n$, είναι φυσιολογική.

2. Μπορούμε στο σχήμα του προηγούμενου ερωτήματος να εμφανίσουμε και την (εκτιμημένη) ευθεία γραμμικής παλινδρόμησης καθώς και τις καμπύλες που δείχνουν τα όρια της μέσης και της ατομικής πρόβλεψης. Αυτό μπορεί να γίνει κάνοντας διπλό κλικ στο συγκεκριμένο γράφημα (στο Output του SPSS) και επιλέγοντας Chart/Options/Fit line:total, Fit options: linear regression, regression prediction lines: Mean, Individual (95%).



Τα παραπάνω μπορούν να γίνουν και από το Graphs/Interactive/Scatterplot (Assign vars X,Y), Fit: method regression, Prediction Lines, Chart



Linear Regression with
95.00% Mean Prediction Interval and
95.00% Individual Prediction Interval

Οι κάθετες αποστάσεις των σημείων από την εκτιμημένη ευθεία γραμμικής παλινδρόμησης είναι τα κατάλοιπα (απεικονίζονται επιλέγοντας spikes: Fit Line). Σε αυτό το γράφημα δίνεται και η εκτίμηση του b_0 (80.78) και του b_1 (1.14).

Το b_0 είναι το σημείο που τέμνει η ευθεία τον κάθετο άξονα, ενώ το b_1 είναι η κλίση της ευθείας (η εφαπτομένη της γωνίας που σχηματίσει η ευθεία με τον οριζόντιο άξονα).

3. Εκτελούμε Regression/Linear/Dependent: Y, Independent: X, Statistics: Confidence Intervals λαμβάνοντας 3 πίνακες. Ο πρώτος πίνακας που δίνεται στο output του SPSS είναι ο ακόλουθος

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,896[a]	,803	,783	7,02

a. Predictors: (Constant), X

Ο πίνακας αυτός περιέχει τις ποσότητες:

R	R^2	Adj. R^2	Std. Error of Est.
$R=\sqrt{\frac{SSR}{SST}}$	$R^2=1-\frac{SSE}{SST}=\frac{SSR}{SST}$	$R^2_{adj}=1-\frac{SSE/(n-2)}{SST/(n-1)}$	$S=\sqrt{\frac{SSE}{n-2}}$

Στη συνέχεια δίνεται ο πίνακας ANOVA:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2008,200	1	2008,200	40,778	,000[a]
	Residual	492,467	10	49,247		
	Total	2500,667	11			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Ο πίνακας ANOVA περιέχει (όπως έχουμε δει και παραπάνω) τις ποσότητες:

Model	***SS***	***df***	***MS***	***F***	***Sig.* (p-value)**
Regression	$SSR=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$	1	$MSR=SSR$	$\frac{MSR}{MSE}$	$1-F_{F_{1,n-2}}(\frac{MSR}{MSE})$
Residuals	$SSE=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	$n-2$	$MSE=\frac{SSE}{n-2}=S^2$		
Total	$SST=\sum_{i=1}^{n}(Y_i-\overline{Y})^2$	$n-1$			

Και τέλος δίνεται από το πακέτο και ο πίνακας

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	80,778	9,544		8,464	,000	59,513	102,043
	X	1,138	,178	,896	6,386	,000	,741	1,535

a. Dependent Variable: Y

που περιέχει τις ποσότητες

	B	**Std.Error**	**t**	**Sig (p-value)**	**LB, UB**
b_0	$\hat{b}_0=\overline{Y}-\hat{b}_1\overline{X}$	$\sqrt{\hat{V}(\hat{b}_0)}=S\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}$	$t_0=\frac{\hat{b}_0}{\sqrt{\hat{V}(\hat{b}_0)}}$	$P(\lvert T\rvert>\lvert t_0\rvert \mid T\sim t_{n-2})$	$\hat{b}_0\pm\sqrt{\hat{V}(\hat{b}_0)}\,t_{n-2}(\frac{a}{2})$
b_1	$\hat{b}_1=\frac{S_{XY}}{S_{XX}}$	$\sqrt{\hat{V}(\hat{b}_1)}=\frac{S}{\sqrt{S_{XX}}}$	$t_1=\frac{\hat{b}_1}{\sqrt{\hat{V}(\hat{b}_1)}}$	$P(\lvert T\rvert>\lvert t_1\rvert \mid T\sim t_{n-2})$	$\hat{b}_1\pm\sqrt{\hat{V}(\hat{b}_1)}\,t_{n-2}(\frac{a}{2})$

Το standardized coefficients Beta είναι η εκτίμηση του b_1 όταν εφαρμοστεί το μοντέλο $Y_i = b_0 + b_1X_i' + \varepsilon_i$, όπου X_i' είναι οι τυποποιημένες τιμές των X_i (αυτό έχει μεγαλύτερη χρησιμότητα στο πολλαπλό μοντέλο όπου έχουμε πολλές ανεξάρτητες μεταβλητές και θέλουμε να δούμε τις εκτιμή-

σεις των b_i όταν οι μεταβλητές αυτές μετρώνται στην ίδια κλίμακα). Ας απαντήσουμε τώρα στα ερωτήματα που τίθενται από την άσκηση.

α. Όπως φαίνεται και από τον παραπάνω πίνακα, οι σημειακές εκτιμήσεις των b_0, b_1 είναι 80.778 και 1.138 αντίστοιχα, ενώ τα αντίστοιχα δ.ε. είναι (59.513, 102.043) και (0.741, 1.535).
β. το p-value για τους δυο αυτούς ελέγχους είναι σχεδόν 0 και επομένως απορρίπτουμε ότι $b_0 = 0$, $b_1 = 0$ και άρα η Y εξαρτάται από την X (υπενθυμίζεται ότι αν $b_1 = 0$ τότε η μεταβλητή Y είναι ανεξάρτητη της X).

γ. Ο πίνακας ανάλυσης διασποράς (ANOVA) δίνεται απευθείας από το πακέτο όπως είδαμε παραπάνω. Το p-value για τον έλεγχο H_0: b_1=0, H_1: b_1≠0 δίνεται στον πίνακα ANOVA και είναι ίσο με 0. Όπως αναφέρεται και παραπάνω, στο απλό γραμμικό μοντέλο ο έλεγχος της συγκεκριμένης υπόθεσης μέσω της F τιμής στον πίνακα ANOVA είναι ισοδύναμος με τον έλεγχο που γίνεται μέσω της t_1 (στον τρίτο πίνακα παραπάνω). Η εκτίμηση της διασποράς των σφαλμάτων ως γνωστό είναι η SSE/(n–2) και από τον πίνακα ANOVA βλέπουμε ότι είναι ίση με 49.247.

δ. Το ποσοστό της μεταβλητότητας των Y_i που ερμηνεύεται από το μοντέλο δίνεται από το R^2 = 0.803.

4. Αυτό μπορεί να γίνει εκτελώντας και πάλι την ίδια ανάλυση Regression/Linear/Dependent: Y, Independent: X, επιλέγοντας save : unstandardized predicted values, unstandardized Residuals. Με αυτή την επιλογή προστίθενται στον πίνακα δεδομένων (Data editor) δύο νέες στήλες που έχουν τις ζητούμενες ποσότητες:

X	*Y*	*predicted*	*residuals*
36	118	121,7459	-3,74592
38	115	124,0219	-9,02193
42	125	128,5739	-3,57395
42	140	128,5739	11,42605
47	128	134,2640	-6,26397
49	145	136,5400	8,46002
55	150	143,3680	6,63199
56	147	144,5060	2,49398
60	155	149,0580	5,94196
63	149	152,4721	-3,47206
68	152	158,1621	-6,16208
72	160	162,7141	-2,71410

5. α. Η σημειακή πρόβλεψη για την πίεση αίματος γυναίκας με ηλικία x_0=52 ετών θα είναι (σύμφωνα με το μοντέλο)

$$Y = \hat{b}_0 + \hat{b}_1 x_0 = 80.778 + 1.138 \cdot 52 = 139.954 .$$

Τα διαστήματα ατομικής και μέσης πρόβλεψης του Y όταν X=52 μπορούν να υπολογιστούν χρησιμοποιώντας τους αντίστοιχους τύπους που δόθηκαν παραπάνω. Μπορούμε όμως να τα πάρουμε απευθείας από το πακέτο ως εξής: προσθέτουμε μία ακόμη ($13^η$) παρατήρηση στο SPSS data editor εισάγοντας στην 13 γραμμή της στήλης X το 52 (το Y στην στην $13^η$ γραμμή αφήνεται κενό). Στη συνέχεια εκτελούμε και πάλι τη διαδικασία της παλινδρόμησης Analyze / Regression / Linear επιλέγοντας στο save τώρα τα Unstandardized predicted values, Prediction Intervals (Mean και Individual). Στην $13^η$ στήλη λαμβάνονται τα αποτελέσματα:

Αναμενόμενη τιμή του Y (πίεση): 139.9540
Δ.ε. 95% για την μέση πρόβλεψη: (135.4383, 144.4697)
Δ.ε. 95% για την ατομική πρόβλεψη: (123.6788, 156.2292)

(με αυτή την διαδικασία προστίθενται στον data editor και τα διαστήματα μέσης και ατομικής πρόβλεψης του Y για όλα τα X_i).

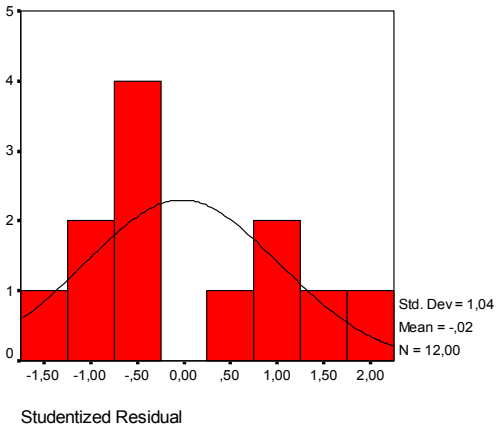
β. Εάν επιλεγεί τυχαία μια γυναίκα ετών 52 από τον πληθυσμό, η πίεση του αίματός της θα βρίσκεται (με σ.ε. 95%) μεταξύ των τιμών 123.6788 και 156.2292 (αυτό ακριβώς εκφράζει το διάστημα ατομικής πρόβλεψης)

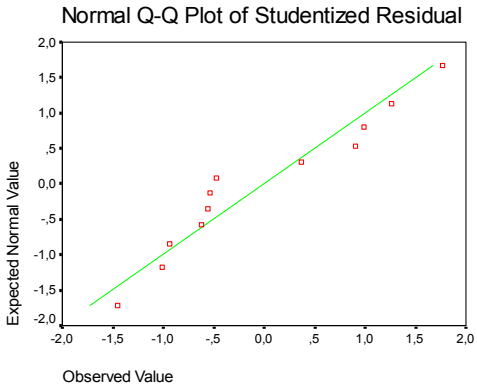
γ. Εάν επιλέξουμε τυχαία έναν μεγάλο αριθμό από γυναίκες ηλικίας 52 ετών, η μέση πίεση του αίματός τους θα βρίσκεται (με σ.ε. 95%) μεταξύ των τιμών 135.4383 και 144.4697 (αυτό ακριβώς εκφράζει το διάστημα μέσης πρόβλεψης)

6. Θα πρέπει πρώτα να αποθηκεύσουμε στον data editor τις τιμές των studentized residuals και των centered leverages. Αυτό γίνεται και πάλι χρησιμοποιώντας την επιλογή save στην ανάλυση regression: εκτελούμε και πάλι τη διαδικασία της παλινδρόμησης Analyze / Regression / Linear επιλέγοντας στο save τα unstandardized predicted values, τα studentized residuals και τα leverages. Στον πίνακα δεδομένων (Data editor) προστίθενται νέες στήλες που έχουν τις ζητούμενες ποσότητες:

X	*Y*	*predicted*	*Studentized residuals*	*Centered leverage values*
36	118	121,7459	-,61859	,17204
38	115	124,0219	-1,45179	,13249
42	125	128,5739	-,55311	,06886
42	140	128,5739	1,76831	,06886
47	128	134,2640	-,94177	,01834
49	145	136,5400	1,26410	,00717
55	150	143,3680	,98955	,00459
56	147	144,5060	,37296	,00867
60	155	149,0580	,90325	,03790
63	149	152,4721	-,53878	,07337
68	152	158,1621	-1,00831	,15828
72	160	162,7141	-,47347	,24943

α. Το ιστόγραμμα και το Q-Q plot των studentized residuals θα είναι (Graphs/histogram, Graphs/Q-Q plot)





από τα οποία δεν μπορούμε να αποφανθούμε (διότι οι παρατηρήσεις είναι λίγες). Το Kolmogorov – Smirnov τεστ δίνει (Analyze/nonparametric tests/1-sample K-S test)

One-Sample Kolmogorov-Smirnov Test

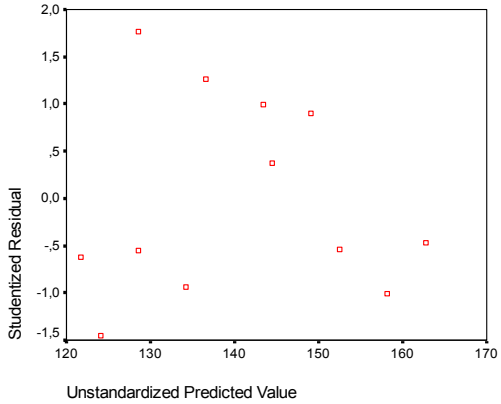
		Studentized Residual
N		12
Normal Parameters[a,b]	Mean	-2,3971E-02
	Std. Deviation	1,0386395
Most Extreme Differences	Absolute	,251
	Positive	,251
	Negative	-,147
Kolmogorov-Smirnov Z		,869
Asymp. Sig. (2-tailed)		,437

a. Test distribution is Normal.

b. Calculated from data.

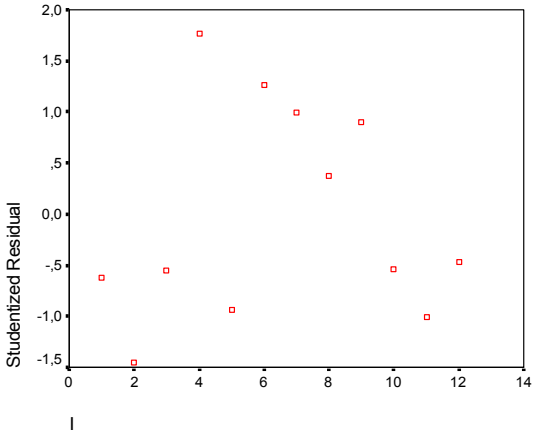
και επομένως (p-value=0.437) δεν μπορούμε να απορρίψουμε ότι τα τυποποιημένα κατάλοιπα προέρχονται από την κανονική κατανομή.

β. Κατασκευάζουμε το γράφημα (scatterplot) των σημείων $(\hat{Y}_i, \hat{\varepsilon}_i^*)$, i = 1, 2, …, n (predicted, studentized residuals):



Οι παρατηρήσεις φαίνεται ότι βρίσκονται τυχαία στο επίπεδο πράγμα που υποδηλώνει ότι δεν πρέπει να υπάρχει κάποια σχέση μεταξύ των δυο μεταβλητών (εξάλλου με τόσες λίγες παρατηρήσεις δεν είναι εύκολο να ανακαλύψουμε κάτι τέτοιο). Επίσης παρατηρούμε ότι δεν υπάρχουν έκτροπες παρατηρήσεις (όλα τα studentized residuals είναι απόλυτα μικρότερα του 3).

γ. Προσθέτουμε άλλη μια μεταβλητή i που δείχνει το αύξοντα αριθμό κάθε παρατήρησης και στη συνέχεια κατασκευάζουμε το γράφημα (scatterplot) των σημείων (i, studentized residuals):



(στην συγκεκριμένη περίπτωση είναι όμοιο με το προηγούμενο γράφημα, κάτι που δεν συμβαίνει γενικά). Και πάλι οι παρατηρήσεις φαίνεται ότι βρίσκονται τυχαία στο επίπεδο. Επίσης, εκτελούμε και ένα τεστ ροών (για τον έλεγχο της τυχαιότητας των σφαλμάτων) με Analyze/non-parametric tests/runs, test variable:studentized residual, cut point=0 (βασιζόμαστε στο πλήθος των ροών θετικών και αρνητικών καταλοίπων):

Runs Test

	Studentized Residual
Test Value[a]	0
Total Cases	12
Number of Runs	5
Z	-,833
Asymp. Sig. (2-tailed)	,405

a. User-specified.

Με βάση το παραπάνω p-value δεν μπορούμε να απορρίψουμε ότι τα κατάλοιπα είναι τυχαία.

δ. Για να εξετάσουμε αν υπάρχουν παρατηρήσεις που έχουν μεγάλη «επιρροή» στο μοντέλο ελέγχουμε ποιες έχουν centered leverage > $5/n = 5/12 = 0.416$. Βλέπουμε ότι καμία παρατήρηση δεν έχει από μόνη της μεγάλη επιρροή στο μοντέλο (όπως έχει σχολιασθεί και παραπάνω, τέτοιες παρατηρήσεις πρέπει να λαμβάνονται με προσοχή).

Άσκηση 2. Στον παρακάτω πίνακα δίνονται οι τιμές πώλησης 14 ειδών θαλασσινών (cents/pound) τα έτη 1970 και 1980 (βλ. Moore, David and McCabe (1989) *Introduction to the Practice of Statistics*).

Είδους Θαλασσινού	Τιμή 1970 X	Τιμή 1980 Y
COD	13.1	27.3
FLOUNDER	15.3	42.4
HADDOCK	25.8	38.7
MENHADEN	1.8	4.5
OCEAN PERCH	4.9	23
SALMON, CHINOOK	55.4	166.3
SALMON, COHO	39.3	109.7

Είδους Θαλασσινού	Τιμή 1970 X	Τιμή 1980 Y
TUNA, ALBACORE	26.7	80.1
CLAMS, SOFT-SHELLED	47.5	150.7
CLAMS, BLUE HARD-SHELLED	6.6	20.3
LOBSTERS, AMERICAN	94.7	189.7
OYSTERS, EASTERN	61.1	131.3
SEA SCALLOPS	135.6	404.2
SHRIMP	47.6	149

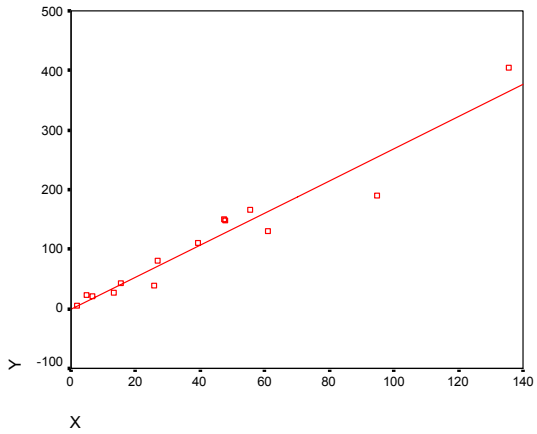
1. Να εφαρμοστεί το απλό γραμμικό $Y = b_0+b_1X+\varepsilon$: (*α*) Να κατασκευάσετε το διάγραμμα διασποράς των δεδομένων (X,Y) μαζί με την εκτιμημένη ευθεία γραμμικής παλινδρόμησης. (*β*) Να εκτιμήσετε τα b_0, b_1 σημειακά και να ελέγξετε (σε ε.σ. 5%) αν H_0: $b_1=0$, H_1: $b_1\neq 0$. Τι ποσοστό της μεταβλητότητας των Y_i ερμηνεύεται από το μοντέλο;

2. Εξετάστε αν η διασπορά των καταλοίπων φαίνεται να είναι σταθερή. Χρησιμοποιείστε το γράφημα των σημείων $(\hat{Y}_i, \hat{\varepsilon}_i^*)$, $i = 1, 2, \ldots, n$. στο επίπεδο. Υπάρχουν «ασυνήθιστες» παρατηρήσεις;

3. Να εξετάσετε αν οι λογάριθμοι των X, Y προσαρμόζονται καλύτερα στο απλό γραμμικό μοντέλο. Με άλλα λόγια εξετάστε αν το (πολλαπλασιαστικό) μοντέλο $Y = cX^{b_1}\cdot e^{\varepsilon}$, $\varepsilon \sim N(0,\sigma^2)$ προσαρμόζει καλύτερα τα δεδομένα (δηλ. δίνει μεγαλύτερο R^2, ενώ τα κατάλοιπα έχουν σταθερή διασπορά).

4. Να ελέγξετε αν $b_1 = 1$ (σε ε.σ. 5%).

Λύση. **1.** Εισάγουμε στο SPSS μόνο τα δεδομένα της δεύτερης (X) και τρίτης στήλης (Y). Με τον ίδιο τρόπο που αυτό έγινε στην προηγούμενη άσκηση λαμβάνουμε το γράφημα:



Από το γράφημα φαίνεται ότι υπάρχει σχέση μεταξύ των δύο μεταβλητών.
β. Εκτελούμε Regression/Linear/Dependent: Y, Independent: X, λαμβάνοντας τους πίνακες:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,967[a]	,935	,930	27,8775

a. Predictors: (Constant), X

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	134512,4	1	134512,387	173,084	,000[a]
	Residual	9325,833	12	777,153		
	Total	143838,2	13			

a. Predictors: (Constant), X

b. Dependent Variable: Y

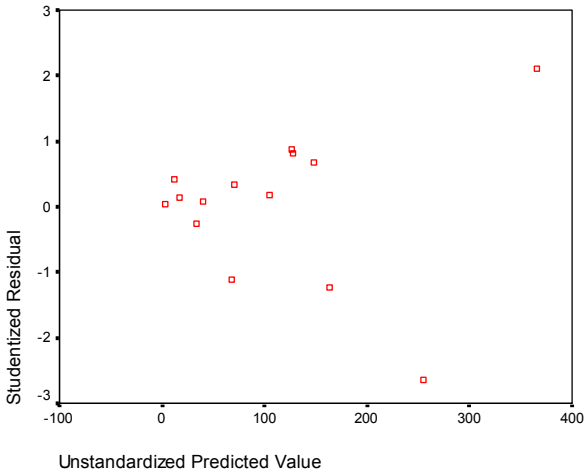
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1,234	11,258		-,110	,915
	X	2,702	,205	,967	13,156	,000

a. Dependent Variable: Y

Οι εκτιμήσεις των b_0 και b_1 είναι -1.234 και 2.702 αντίστοιχα ενώ απορρίπτουμε την υπόθεση H_0: b_1=0, διότι το αντίστοιχο p-value είναι σχεδόν 0. Επίσης, το ποσοστό της μεταβλητότητας των Y_i που ερμηνεύεται από το μοντέλο είναι $R^2 = 0.935$.

2. Εκτελούμε και πάλι Regression/Linear/Dependent: Y, Independent: X, επιλέγοντας save τα unstandardized predicted values και studentized residuals και στη συνέχεια κατασκευάζουμε το ζητούμενο γράφημα:



Από το παραπάνω γράφημα αλλά και από τον πίνακα δεδομένων παρατηρούμε ότι υπάρχουν δύο «ασυνήθιστες» παρατηρήσεις (absolute studentized residual > 2). Από το γράφημα επίσης φαίνεται ότι η διασπορά των studentized residuals δεν πρέπει να είναι σταθερή. Συγκεκριμένα παρατηρούμε ότι όσο αυξάνεται το (προσαρμοσμένο) Y, τόσο αυξάνεται και η διασπορά των studentized residuals. Μάλιστα αυτό φαίνεται να δικαιολογεί και τα «μεγάλα» κατάλοιπα στις δύο ασυνήθιστες παρατηρήσεις. Τα παραπάνω υποδηλώνουν ότι το μοντέλο μας δεν πρέπει να είναι σωστό (αν και το R^2 είναι αρκετά μεγάλο).

3. Εδώ ουσιαστικά θα πρέπει να εξετάσουμε το γραμμικό μοντέλο (λογαριθμούμε κατά μέλη το πολλαπλασιαστικό μοντέλο $Y = cX^{b_1} \cdot e^{\varepsilon}$)

$$\ln Y = b_0 + b_1 \ln X + e \quad \Leftrightarrow \quad Y' = b_0 + b_1 X' + e$$

($Y' = \ln Y, X' = \ln X, b_0 = \ln c$). Αρχικά λοιπόν θα πρέπει να μετασχηματίσουμε τα δεδομένα κατασκευάζοντας δύο νέες μεταβλητές LOGX, LOGY (με compute LOGX=Ln(Y), LOGX=ln(X)). Στη συνέχεια εκτελούμε την ανάλυση Regression/Linear/Dependent: LOGY, Independent: LOGX, επιλέγοντας save τα unstandardized predicted values και studentized residuals. Λαμβάνουμε τους πίνακες:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,974[a]	,950	,945	,2776

a. Predictors: (Constant), LOGX

b. Dependent Variable: LOGY

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17,416	1	17,416	226,070	,000[a]
	Residual	,924	12	7,704E-02		
	Total	18,340	13			

a. Predictors: (Constant), LOGX

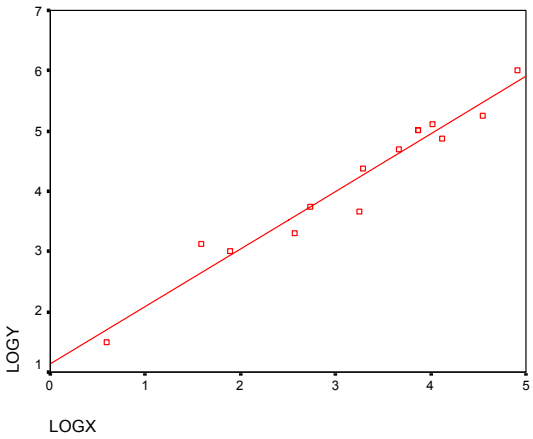
b. Dependent Variable: LOGY

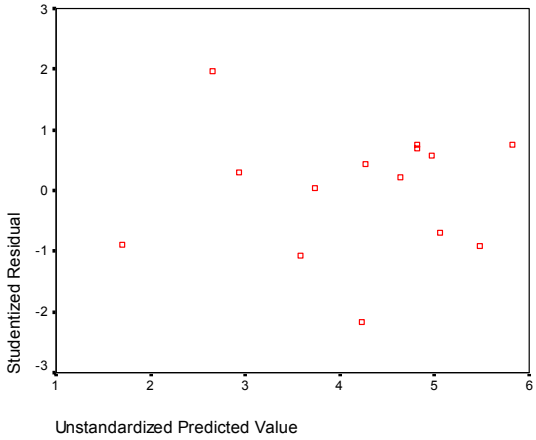
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,132	,217		5,227	,000
	LOGX	,955	,063	,974	15,036	,000

a. Dependent Variable: LOGY

Παρατηρούμε ότι το R^2 είναι 0.950, ελάχιστα μεγαλύτερο αυτού που πήραμε πριν τον μετασχηματισμό, αλλά το σημαντικότερο είναι ότι τώρα τα κατάλοιπα φαίνεται να έχουν σταθερή διασπορά:





4. Η εκτίμηση του b_1 στο πολλαπλασιαστικό μοντέλο βλέπουμε ότι είναι 0.955 και απορρίπτεται η υπόθεση H_0: b_1=0 (p-value=0.000). Το πακέτο δεν προσφέρει άμεσα την δυνατότητα ελέγχου της υπόθεσης H_0: b_1=1 για αυτό και θα κάνουμε τον έλεγχο εμμέσως χρησιμοποιώντας έναν μετασχηματισμό. Θέτουμε $b_1' = b_1 - 1$ (ώστε να ελέγξουμε στη συνέχεια αν $b_1' = 0$). Επομένως $b_1 = b_1' + 1$ και αντικαθιστώντας στο μοντέλο $Y' = b_0 + b_1 X' + e$ θα έχουμε:

$$Y' = b_0 + (b_1' + 1)X' + e \Leftrightarrow Y' - X' = b_0 + b_1' X + e \Leftrightarrow Y^* = b_0 + b_1' X + e$$

Κατασκευάζουμε λοιπόν μια νέα μεταβλητή Ystar = LOGY–LOGX και εφαρμόζουμε το μοντέλο της γραμμικής παλινδρόμησης ανάμεσα στις μεταβλητές Ystar (dependent) και LOGX (independent). Ανάμεσα στα αποτελέσματα μας ενδιαφέρει ο πίνακας που αφορά τον έλεγχο $b_1' = 0$ που είναι ισοδύναμος με τον $b_1 = 1$:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,132	,217		5,227	,000
	LOGX	-4,53E-02	,063	-,202	-,714	,489

a. Dependent Variable: YSTAR

από τον πίνακα αυτό βλέπουμε ότι το p-value για τον έλεγχο της υπόθεσης $H_0: b_1' = 0 \Leftrightarrow H_0$: b_1=1 είναι ίσο με 0.489 και επομένως δεν μπορούμε να απορρίψουμε ότι b_1=1.