Πρώτη σειρά ασκήσεων 2022-2023

Μάθημα: 858 - Στατιστική Μοντελοποίηση

ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Ονοματεπώνυμο: Απόστολος Μουστάκης

Αριθμός Μητρώου: 03400182

Καθηγήτρια: Χρυσηίς Καρώνη – Ρίτσαρντσον

Περιεχόμενα

Άσκηση Α – Μαθηματικές Αποδείξεις 1 έως 4

Έστω ότι έχουμε n ανεξάρτητες παρατηρήσεις (x_i, y_i) με $i=1,\dots,n$

Η εκτίμηση της τιμής της τυχαίας μεταβλητής y_i με βάση το απλό γραμμικό μοντέλο $E(y|x) = E(y_x) = B_0 + B_1 x$, δίνεται από την σχέση $\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$ (1), όπου $\hat{B}_0$ και $\hat{B}_1$ καλούνται εκτιμήτριες των παραμέτρων του μοντέλου.

Για τον υπολογισμό των παραμέτρων B_0, B_1 χρησιμοποιούμε την μέθοδο των ελαχίστων τετραγώνων η οποία έγκειται στην ελαχιστοποίηση της παράστασης:

$$S(B_0,B_1) = \sum_{i=1}^{n} (y_i - E(y_i))^2 = \sum_{i=1}^{n} (y_i - B_0 - B_1 x_i)^2$$

ως προς τα B_0 και B_1

Ακολουθώντας την διαδικασία για τον εντοπισμό ελαχίστων τιμών καταλήγουμε στις εξής κανονικές εξισώσεις:

$$n\hat{B}_0 + \hat{B}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (2)$$

$$\hat{B}_0 \sum_{i=1}^{n} x_i + \hat{B}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (3)$$

από τις οποίες λαμβάνουμε τις εκτιμήτριες των B_0, B_1 ως

$$\hat{B}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\cdot(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (4)$$

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \cdot \bar{x} \quad (5)$$

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

Για να απλοποιήσουμε τους συμβολισμούς ορίζουμε τα εξής:

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

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

$$SST = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})\cdot(y_i - \bar{y})$$

Με βάση τους συμβολισμούς έχουμε $\hat{B}_1 = \frac{S_{xy}}{S_{xx}}$ (6)

(1)

1) Θέλουμε να δείξουμε ότι $R^2 = r_{xy}^2$

Γνωρίζουμε ότι ο συντελεστής προσδιορισμού R^2 δίνεται από την σχέση:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (7)$$

Ο δειγματικός συντελεστής συσχέτισης Pearson δίνεται από την σχέση:

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

άρα $r_{xy}^2 = \dfrac{S_{xy}^2}{S_{xx}\cdot S_{yy}}$ (8)

Από την (7) έχουμε:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{S_{yy}} \overset{(1)}{=} \frac{\sum_{i=1}^{n}(\hat{B}_0+\hat{B}_1 x_i-\bar{y})^2}{S_{yy}} =$$

$$\overset{(5)}{=} \frac{\sum_{i=1}^{n}(\cancel{\bar{y}}-\hat{B}_1\bar{x}+\hat{B}_1 x_i-\cancel{\bar{y}})^2}{S_{yy}} = \hat{B}_1^2 \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{S_{yy}} =$$

$$\overset{(6)}{=} \frac{S_{xy}^2}{S_{xx}^{\cancel{2}}} \cdot \frac{\cancel{S_{xx}}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}\cdot S_{yy}} \overset{(8)}{=} r_{xy}^2$$

2) Θέλουμε να δείξουμε ότι $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$

Από την (2) έχουμε: $n\hat{B}_0 + \hat{B}_1\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i \Rightarrow$

$$\sum_{i=1}^{n}\hat{B}_0+\hat{B}_1 x_i = \sum_{i=1}^{n}y_i \overset{(1)}{\Rightarrow} \sum_{i=1}^{n}\hat{y}_i = \sum_{i=1}^{n}y_i$$

3) Θέλουμε να δείξουμε ότι: $\sum_{i=1}^{n}(y_i-\hat{y}_i)\cdot(\hat{y}_i-\bar{y}) = 0$

Έχουμε: $y_i-\hat{y}_i \overset{(1)}{=} y_i-(\hat{B}_0+\hat{B}_1 x_i) \overset{(5)}{=} y_i-(\bar{y}-\hat{B}_1\bar{x}+\hat{B}_1 x_i) =$

$$= y_i-\bar{y}+\hat{B}_1\bar{x}-\hat{B}_1 x_i = y_i-\bar{y}-\hat{B}_1(x_i-\bar{x}) \quad (9)$$

και: $\hat{y}_i-\bar{y} \overset{(1)}{=} \hat{B}_0+\hat{B}_1 x_i-\bar{y} \overset{(5)}{=} \cancel{\bar{y}}-\hat{B}_1\bar{x}+\hat{B}_1 x_i-\cancel{\bar{y}} = \hat{B}_1(x_i-\bar{x})$ (10)

Άρα: $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i) =$

$$\overset{(9),(10)}{=} \sum_{i=1}^{n}\hat{B}_1(x_i-\bar{x})\cdot[y_i-\bar{y}-\hat{B}_1(x_i-\bar{x})] =$$

(2)

$$= \sum_{i=1}^{n} [\hat{B_1}\cdot(x_i-\bar{x})\cdot(y_i-\bar{y}) - \hat{B_1}^2(x_i-\bar{x})^2] =$$

$$= \sum_{i=1}^{n} \hat{B_1}(x_i-\bar{x})(y_i-\bar{y}) - \sum_{i=1}^{n} \hat{B_1}^2(x_i-\bar{x})^2 =$$

$$= \hat{B_1}\cdot S_{xy} - \hat{B_1}^2\cdot S_{xx} \overset{⑥}{=} \frac{S_{xy}}{S_{xx}}\cdot S_{xy} - \frac{S_{xy}^2}{S_{xx}^{\cancel{2}}}\cdot \cancel{S_{xx}} = 0$$

4) Θέλουμε να δείξουμε ότι $\frac{\hat{B_1}}{se(\hat{B_1})} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$

Αρχικά έχουμε:

$$\frac{r_{xy}\cdot\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \overset{⑧}{=} \frac{\frac{S_{xy}\cdot\sqrt{n-2}}{\sqrt{S_{xx}\cdot S_{yy}}}}{\sqrt{1-\frac{S_{xy}^2}{S_{xx}\cdot S_{yy}}}} = \frac{S_{xy}\cdot\sqrt{n-2}}{\sqrt{S_{xx}\cdot S_{yy}}\cdot\sqrt{1-\frac{S_{xy}^2}{S_{xx}\cdot S_{yy}}}} \quad (11)$$

$$\frac{\hat{B_1}}{se(\hat{B_1})} = \frac{\hat{B_1}}{\frac{s}{\sqrt{S_{xx}}}} = \frac{\hat{B_1}\cdot\sqrt{S_{xx}}}{s} = \frac{\hat{B_1}\cdot\sqrt{S_{xx}}}{\sqrt{\frac{1}{n-2}\cdot\sum_{i=1}^{n}(y_i-\hat{y_i})^2}} =$$

$$= \frac{\hat{B_1}\cdot\sqrt{S_{xx}}\cdot\sqrt{n-2}}{\sqrt{\underbrace{\sum_{i=1}^{n}(y_i-\hat{y_i})^2}_{SSE}}} = \frac{\hat{B_1}\cdot\sqrt{S_{xx}}\cdot\sqrt{n-2}}{\sqrt{S_{yy}-\frac{S_{xy}^2}{S_{xx}}}} \overset{⑥}{=} \frac{S_{xy}\cdot\sqrt{S_{xx}}\cdot\sqrt{n-2}}{S_{xx}\cdot\sqrt{S_{yy}\left(1-\frac{S_{xy}^2}{S_{xx}S_{yy}}\right)}} =$$

$$= \frac{S_{xy}\cdot\cancel{\sqrt{S_{xx}}}\cdot\sqrt{n-2}}{(\sqrt{S_{xx}})^{\cancel{2}}\cdot\sqrt{S_{yy}}\cdot\sqrt{1-\frac{S_{xy}^2}{S_{xx}\cdot S_{yy}}}} = \frac{S_{xy}\cdot\sqrt{n-2}}{\sqrt{S_{xx}\cdot S_{yy}}\cdot\sqrt{1-\frac{S_{xy}^2}{S_{xx}\cdot S_{yy}}}} \quad (12)$$

Από (11) και (12) έχουμε: $\frac{\hat{B_1}}{se(\hat{B_1})} = \frac{r_{xy}\cdot\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$

ΠΑΡΑΤΗΡΗΣΕΙΣ για το 4)

⇒ Το τυπικό σφάλμα $se(\hat{B_1})$ ορίζεται ως $\sqrt{\frac{s^2}{S_{xx}}} = \frac{s}{\sqrt{S_{xx}}}$

⇒ Το s^2 (εκτιμήτρια πραγματικής διασποράς σ^2) ορίζεται ως

$$s^2 = \frac{1}{n-2}\cdot\sum_{i=1}^{n}(y_i-\hat{y_i})^2 \quad \text{άρα} \quad s = \frac{1}{\sqrt{n-2}}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\hat{y_i})^2}$$

$$\Rightarrow SSE = SST - SSR = S_{yy} - \hat{B_1}^2 S_{xx} \overset{⑥}{=} S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$SSR = \sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2 \overset{①}{=} \sum_{i=1}^{n}(\hat{B_0}+\hat{B_1}x_i-\bar{y})^2 \overset{⑤}{=} \sum_{i=1}^{n}(\cancel{\bar{y}}-\hat{B_1}\bar{x}+\hat{B_1}x_i-\cancel{\bar{y}})^2 =$$

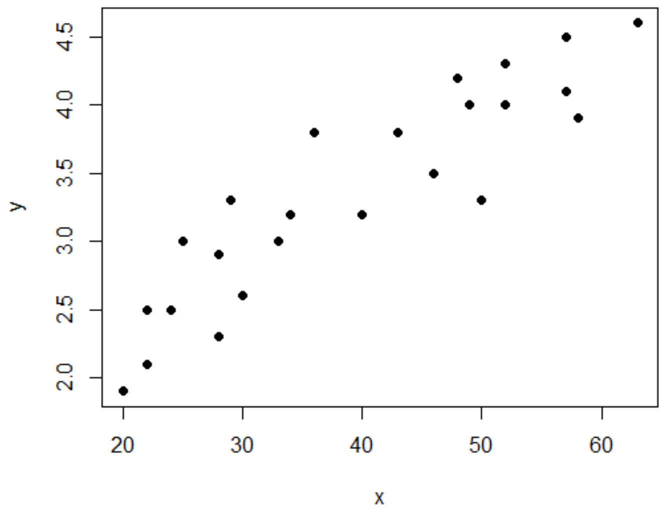
$$= \hat{B_1}^2\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2 = \hat{B_1}^2\cdot S_{xx}$$

(3)

Άσκηση B

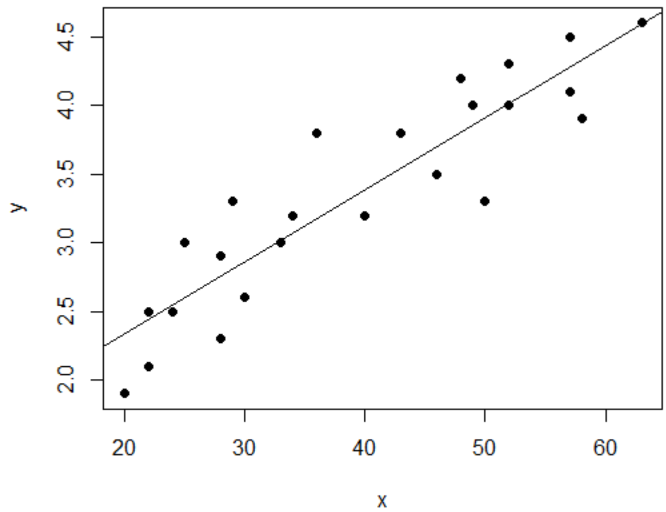
Ερώτημα i)

Για της δημιουργία του διαγράμματος διασποράς μεταξύ των μεταβλητών x (ηλικία ασθενών) και y (επίπεδα ολικής χοληστερόλης ασθενών) χρησιμοποίησα το R studio. Αφού εισήγαγα τα δεδομένα μέσω του αρχείου cholesterol.txt με τις κατάλληλες εντολές σχημάτισα το παρακάτω διάγραμμα διασποράς των x και y.



Διάγραμμα 1: Διασπορά ηλικίας-ολικής χοληστερόλης

Εκ πρώτης όψεως παρατηρούμε πως υπάρχει μία θετική συσχέτιση μεταξύ των x, ηλικία ασθενών, και των y, επίπεδα χοληστερόλης ασθενών, καθώς όσο αυξάνονται τα x αυξάνονται και τα y. Η συσχέτιση αυτή μπορεί να ταιριάζει με το απλό γραμμικό μοντέλο, οπότε θα προσαρμόσουμε το μοντέλο αυτό $E(y|x) = \beta_o + \beta_1 x$ στα δεδομένα μας, το οποίο φαίνεται στο παρακάτω διάγραμμα.



Διάγραμμα 2: Προσαρμογή γραμμικού μοντέλου

Στην περίπτωση μας η ευθεία αυτή δίνεται από τον τύπο $\hat{y}$ = 1.27987 + 0.05262x, όπου $\hat{\beta}_0$= 1.27987 ± 0.215699 και $\hat{\beta}_1$ = 0.05262 ± 0.005192

Ερώτημα ii)

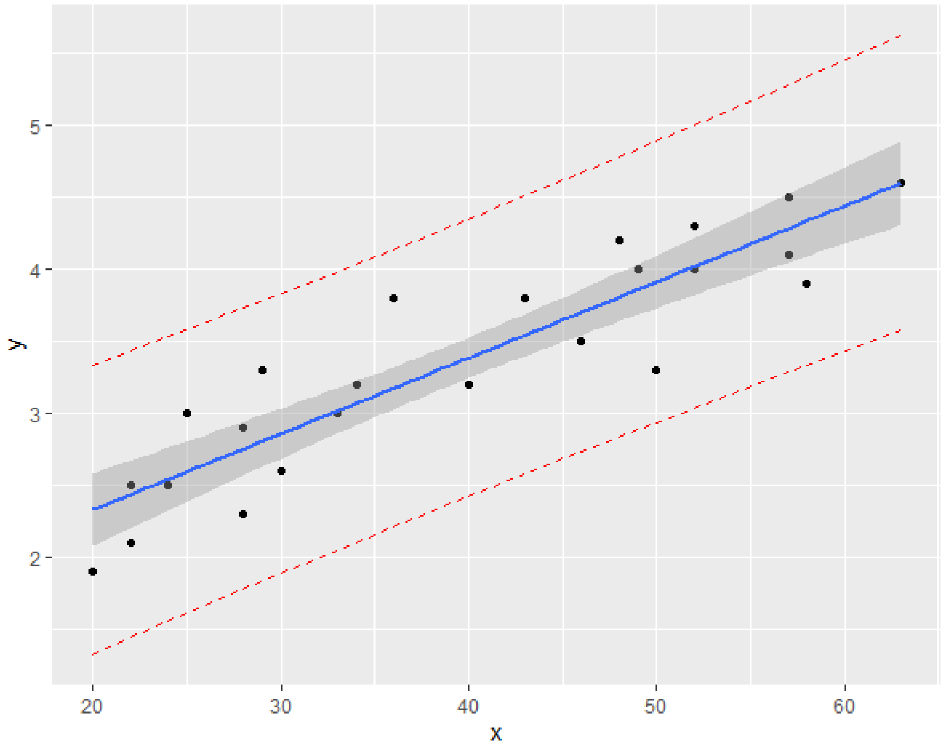
Για να γίνει ο έλεγχος H_0: β_1 = 0 έναντι της H_1: $\beta_1 \neq 0$ χρησιμοποιώ t-test. Στόχος μας είναι να απορρίψουμε την H_0 καθώς αν ίσχυε αυτό θα σήμαινε πως οι μεταβλητές μας δεν συσχετίζονται. Με βάση τον έλεγχο, η τιμή της ελεγχοσυνάρτησης t είναι 10.136 με p-value = 9.428e-10, το οποίο είναι <0.001 και συνεπώς ο έλεγχος είναι στατιστικά σημαντικός και απορρίπτουμε την H_0. Επίσης θα μπορούσαμε να χρησιμοποιήσουμε F-test. Με βάση τον έλεγχο F, ή τιμή της ελεγχοσυνάρτησης F είναι 102.7 με p-value = 9.428e-10, το οποίο είναι <0.001 και συνεπώς και σε αυτή την περίπτωση απορρίπτουμε την H_0. Οι έλεγχοι t και F ταυτίζονται στο απλό γραμμικό μοντέλο και συνεπώς είναι κάτι που περιμέναμε.

Το 0.95-διάστημα εμπιστοσύνης (δ.ε.) για το συντελεστή της x είναι το (0.04185806, 0.06339175). Εφόσον το 0 δεν ανήκει σε αυτό το διάστημα απορρίπτω την H_0 και συνεπώς και αυτός ο έλεγχος ταιριάζει με τα αποτελέσματα των δύο παραπάνω ελέγχων.

Το $\hat{\beta}1$ ερμηνεύεται ως εξής: Αν αυξηθεί η ηλικία του ασθενούς κατά 1 χρόνο, το επίπεδο χοληστερόλης του αναμένεται να αυξηθεί κατά περίπου 0.05 mg/ml.

Ερώτημα iii)

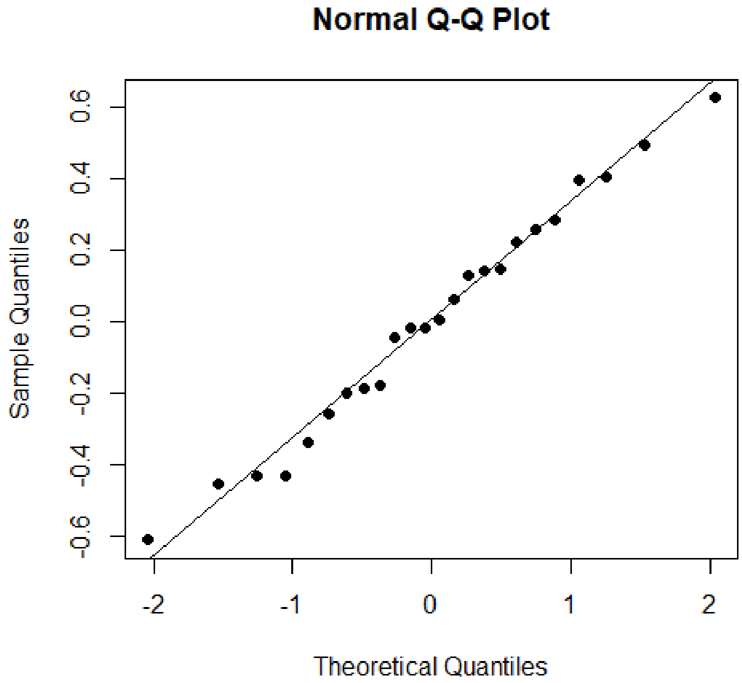
Το 0.99 διάστημα εμπιστοσύνης για το επίπεδο χοληστερόλης ενός ασθενή ηλικίας 35 χρονών (x_0=35) είναι (2.158578, 4.084902), ενώ το 0.99 διάστημα εμπιστοσύνης για την αναμενόμενη τιμή της E(y) ενός ασθενή ηλικίας 35 χρονών είναι (2.918965, 3.324515). Το παρακάτω διάγραμμα δείχνει τα 99% διαστήματα εμπιστοσύνης πρόβλεψης για τις y και E(y) με x_0=35.



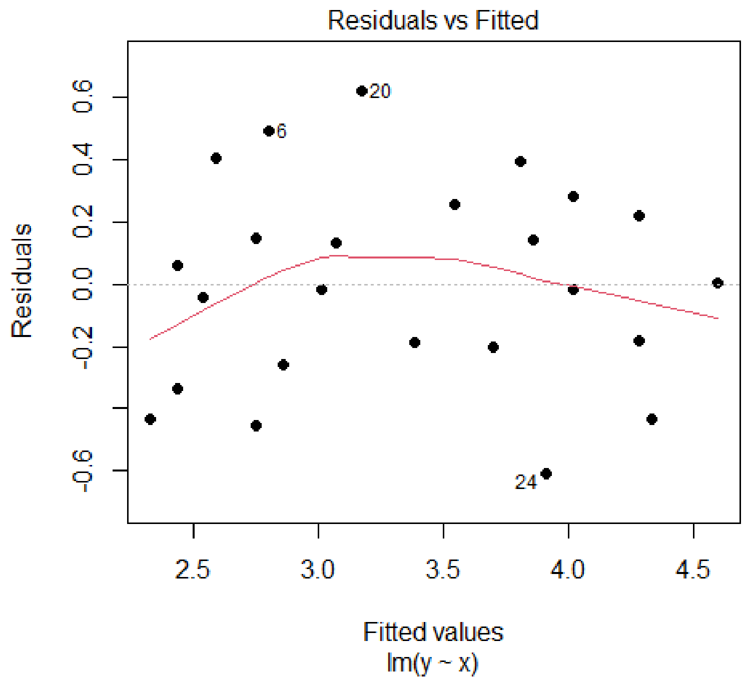
Διάγραμμα 3: 99% διαστήματα εμπιστοσύνης πρόβλεψης y, E(y) για x_0=35

Ερώτημα iv)

Παρακάτω παρουσιάζονται ο γραφικός έλεγχος της Κανονικής κατανομής και η γραφική παράσταση e_i με $\hat{y}_i$, για τα υπόλοιπα e_i.



Διάγραμμα 4: Γραφικός έλεγχος Κανονικής κατανομής



Διάγραμμα 5: Υπόλοιπα σε σχέση με τα εκτιμημένα $\hat{y}_i$

Στο πρώτο διάγραμμα έχουμε σχεδιάσει το Q-Q plot και παρατηρούμε ότι τα σημεία της γραφικής παράστασης κείτονται σε μία ευθεία και συνεπώς μπορούμε να συμπεράνουμε πως το μοντέλο παλινδρόμησης που έχουμε επιλέξει είναι κατάλληλο (δεν υπάρχουν λόγοι αμφισβήτησης).

Στο δεύτερο διάγραμμα έχουμε σχεδιάσει τα υπόλοιπα e_i με τα εκτιμημένα $\hat{y}_i$. Παρατηρούμε ότι τα υπόλοιπα κατανέμονται με τυχαίο και ομοσκεδαστικό τρόπο γύρω από το μηδέν και συνεπώς δεχόμαστε την υπόθεση της ομοσκεδαστικότητας και δεν αμφισβητείται η καταλληλόλητα του μοντέλου.

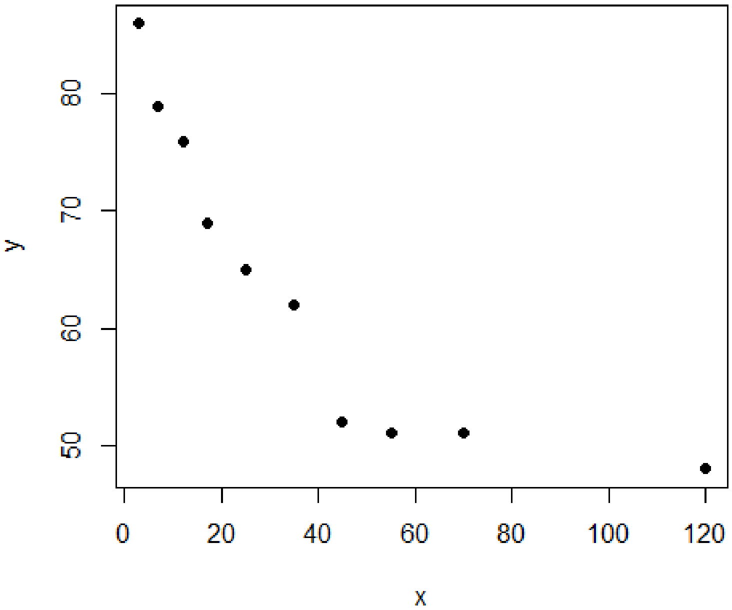
Άσκηση Γ

Τα παρακάτω δεδομένα αφορούν την συμμετοχή y (%) ιδιοκτητών κατοικιών σε δειγματοληπτική έρευνα σε σχέση με την αξία της κατοικίας x:

x	3	7	12	17	25	35	45	55	70	120
y	86	79	76	69	65	62	52	51	51	48

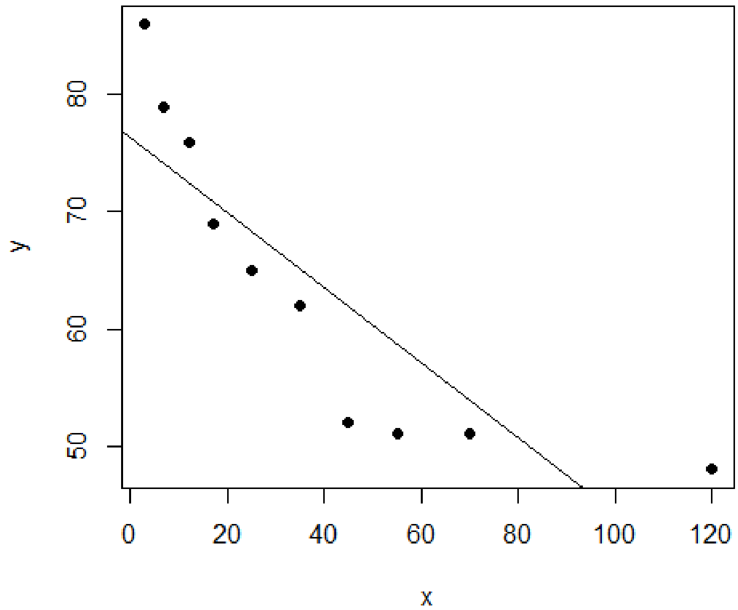
Ερώτημα i)

Για της δημιουργία του διαγράμματος διασποράς μεταξύ των μεταβλητών x και y χρησιμοποίησα το R studio. Αφού εισήγαγα τα δεδομένα μέσω ενός αρχείου που ονόμασα data.txt με τις κατάλληλες εντολές σχημάτισα το παρακάτω διάγραμμα διασποράς των x και y.



Διάγραμμα 6: Διασπορά αξία κατοικίας-συμμετοχή ιδιοκτητών

Εκ πρώτης όψεως φαίνεται πως η συσχέτιση των δεδομένων μας δεν είναι γραμμική. Αν προσπαθήσουμε να προσαρμόσουμε ένα γραμμικό μοντέλο προκύπτει το παρακάτω διάγραμμα που μας διαβεβαιώνει πως η συσχέτιση δεν είναι γραμμική.



Διάγραμμα 7: Ανεπιτυχής προσαρμογή γραμμικού μοντέλου

Το μοντέλο $\frac{100}{100-y}$ = α + $\frac{\beta}{x}$ που μας ζητείται να προσαρμόσουμε στην συγκεκριμένη άσκηση είναι μη γραμμικό αλλά με κάποιες μετατροπές μπορεί να γίνει γραμμικό.

Συγκεκριμένα:

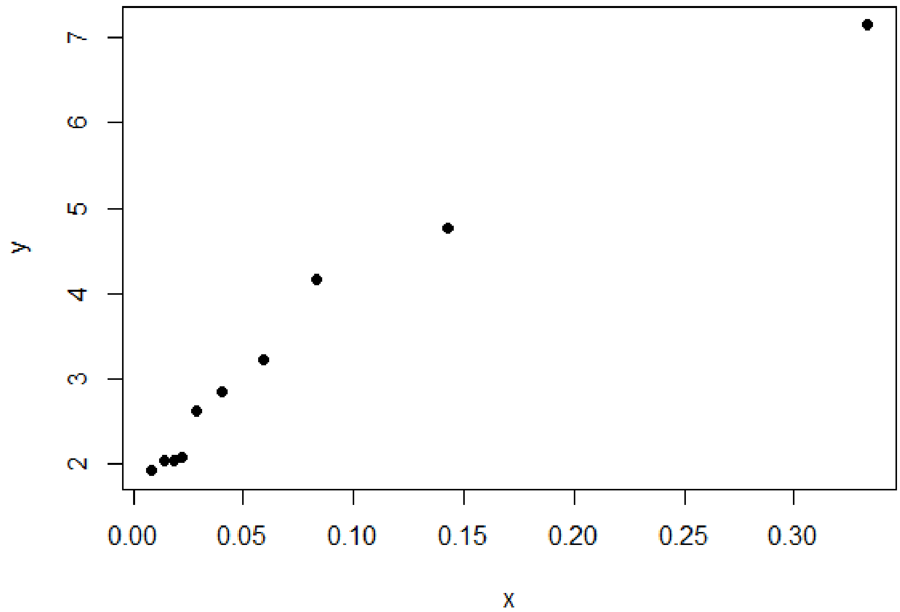
Θέτουμε $y^*= \frac{100}{100-y}$ και $x^*= \frac{1}{x}$

Συνεπώς έχουμε $y^* = \beta_o + \beta_1 x^*$ όπου $\beta_o = \alpha$ και $\beta_1 = \beta$ (1)

Τα νέα μετασχηματισμένα δεδομένα βρίσκονται στον παρακάτω πίνακα

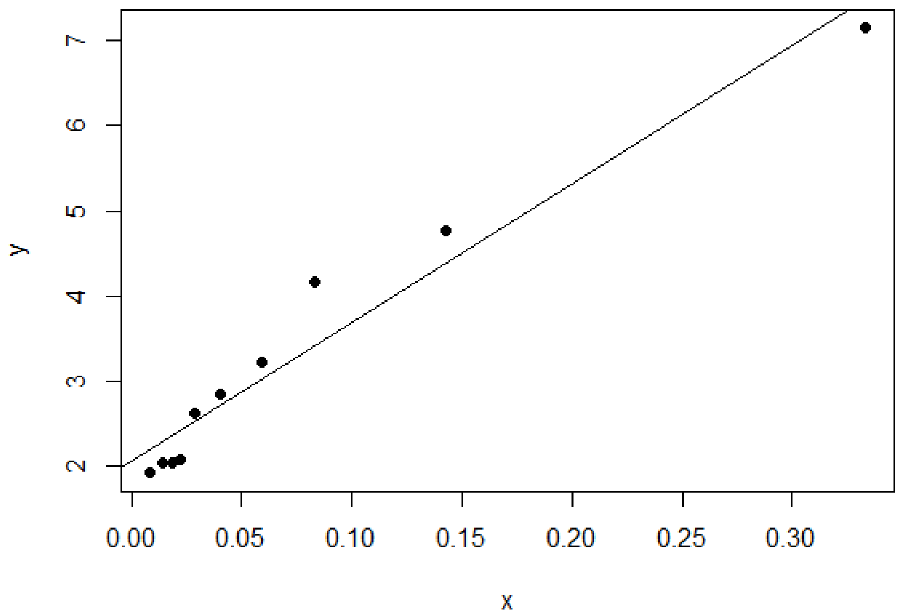
x^*	0.3333	0.1428	0.0833	0.0588	0.04	0.0286	0.0222	0.0181	0.0143	0.0083
y^*	7.1429	4.7619	4.1667	3.2258	2.8571	2.6315	2.0833	2.0408	2.0408	1.9231

Εισήγαγα τα δεδομένα στο R Studio μέσω του αρχείου transformed.txt και με τις κατάλληλες εντολές δημιούργησα το διάγραμμα διασποράς μεταξύ των x^* και y^*.



Διάγραμμα 8: Διασπορά των x^* και y^*

Με αυτή την μετατροπή παρατηρούμε πως τα δεδομένα μας ταιριάζουν περισσότερο στο απλό γραμμικό μοντέλο, οπότε αποφασίζουμε να το εφαρμόσουμε. Στο παρακάτω διάγραμμα φαίνεται και η ευθεία του απλού γραμμικού μοντέλου.



Διάγραμμα 9: Προσαρμογή του γραμμικού μοντέλου στα (x^*, y^*)

Η ευθεία αυτή δίνεται από τον τύπο $\hat{y}^*$= 2.0678 + 16.2677x^*, όπου $\hat{\beta}_0$ = 2.0678 ± 0.1596 και $\hat{\beta}_1$ = 16.2677 ± 1.3232. Με βάση την (1) έχουμε $\alpha = \beta_o$ = 2.0678 και $\beta = \beta_1$ = 16.2677.

Επίσης, αν γίνει ο έλεγχος H_0: β_1 = 0 έναντι της H_1: $\beta_1 \neq 0$ με τη χρήση t-test η τιμή της ελεγχοσυνάρτησης t είναι 12.29 με p-value = 1.78e-06, το οποίο είναι <0.001 και συνεπώς ο έλεγχος είναι στατιστικά σημαντικός και απορρίπτουμε την H_0, το οποίο δικαιολογεί τη χρήση του απλού γραμμικού μοντέλου.

Ερώτημα ii)

Η άγνωστη παρατήρηση y μπορεί να εκτιμηθεί σημειακά από την σχέση:

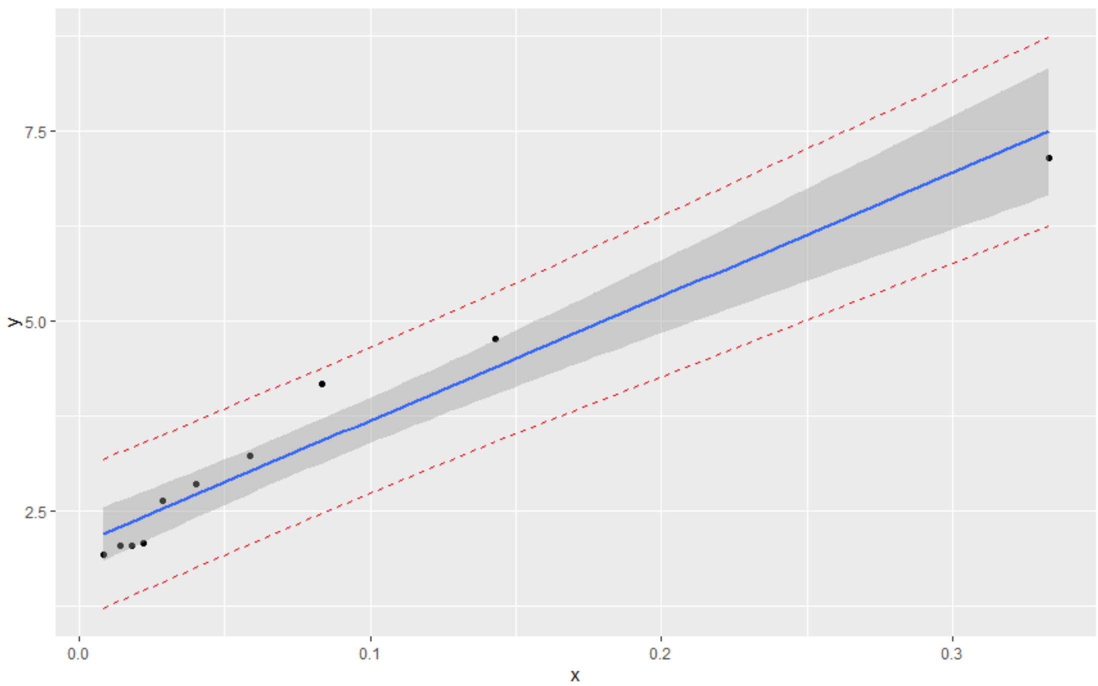
$$\frac{100}{100-y} = 2.0678 + \frac{16.2677}{x}$$

Για x = 20, $\hat{y} \cong 65.29$

Για να κατασκευάσουμε διαστήματα εμπιστοσύνης για την πρόβλεψη της παρατήρησης y και για την μέση τιμή της, E(y), όταν x_0=20 θα χρησιμοποιήσουμε το γραμμικό μοντέλο του παραπάνω ερωτήματος ($\hat{y}^*$= 2.0678 + 16.2677x^*).

Για x_0= 20 έχουμε $x_0{}^* = \frac{1}{20}$ = 0.05 καθώς $x^* = \frac{1}{x}$, και $\hat{y}^* \cong 2.89$

Το 95% διάστημα εμπιστοσύνης για την τιμή y^* για $x_0{}^*$= 0.05 είναι (1.922433, 3.83994), ενώ το 95% διάστημα εμπιστοσύνης για την αναμενόμενη τιμή της E(y^*) για $x_0{}^*$ = 0.05 είναι (2.583123, 3.175925). Το παρακάτω διάγραμμα δείχνει τα 95% διαστήματα εμπιστοσύνης πρόβλεψης για τις y^* και E(y^*) με $x_0{}^*$ =0.05.



Διάγραμμα 10: 95% διαστήματα εμπιστοσύνης πρόβλεψης y^*, E(y^*) με $x_0{}^*$ = 0.05

Επομένως το 95% διάστημα εμπιστοσύνης για την πρόβλεψη της παρατήρησης y όταν x_0= 20 είναι (47.982,73.958) ενώ το προσεγγιστικό 95% διάστημα εμπιστοσύνης για την μέση τιμή της, E(y), όταν x_0= 20 είναι (61.287,68.513).