Στατιστική Μοντελοποίηση

Πρώτη Σειρά ασκήσεων

Κωνσταντίνος Παπαδάκης
ΕΔΕΜΜ 03400149

17 Νοεμβρίου 2021

Α′ Μέρος

Δείξτε ότι για το <u>απλό γραμμικό μοντέλο</u> $\mathrm{E}[Y] = \beta_0 + \beta_1 X$ ισχύουν τα ακόλουθα:

Άσκηση Α′.1. $R^2 = r_{xy}^2$, *R^2 ο συντελεστής προσδιορισμού, r_{xy} ο δειγματικός συντελεστής συσχέτισης (Pearson) των x και y παρατηρήσεων.*

Απόδειξη. Ισχύει ότι

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{1}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \tag{3}$$

$$\mathrm{SSR} = \hat{\beta}_1^2 S_{xx} \tag{4}$$

$$\mathrm{SST} = S_{yy} \tag{5}$$

Έτσι έχουμε

$$\begin{aligned} R^2 &= \frac{\mathrm{SSR}}{\mathrm{SST}} \\ &= \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} \\ &= \frac{S_{xy}^2}{S_{xx}^2} \frac{S_{xx}}{S_{yy}} \\ &= \frac{S_{xy}^2}{S_{xx} S_{yy}} \\ &= r_{xy}^2 \end{aligned}$$

□

Άσκηση Α′.2.

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

Απόδειξη. Για να αποδείξουμε το αποτέλεσμα θα ορίσουμε κάποιους πίνακες και θα αποδείξουμε μερικές ιδιότητές τους.

Ορίζουμε

$$
\begin{aligned}
I &:= \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} \\
J &:= \frac{1}{n} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix} \\
H &:= X(X^\top X)^{-1} X^\top
\end{aligned}
$$

οπου όλοι τους είναι διάστασης $n \times n$.

Και οι τρεις αυτοί πίνακες είναι *προβολές* αφού είναι εύκολο να δει κανείς πως είναι *συμμετρικοί* και *ταυτοδύναμοι*. Οι υπόχωροι που αντιστοιχούν σε αυτές τις προβολές είναι οι εξής:

- Για τον I είναι όλος χώρος στον οποίον ορίζεται.
- Για τον J ειναι ο χώρος που παράγεται από το $\mathbf{1} := \begin{pmatrix} 1 & \dots & 1 \end{pmatrix}^\top$. Αυτό ισχύει αφού $\operatorname{rank} J = 1$ και $J\mathbf{1} = \mathbf{1}$
- Για τον H είναι ο χώρος που παράγεται από της στήλες του X (μία εξ αυτών η $\mathbf{1}$). Αυτό ισχύει αφού

  $$\begin{pmatrix} HX^{(1)} & \dots & HX^{(p)} \end{pmatrix} = HX = X = \begin{pmatrix} X^{(1)} & \dots & X^{(p)} \end{pmatrix}$$

  και $\operatorname{rank} H \leq \operatorname{rank} X = p$

Παρατηρούμε ότι η J είναι *υποπροβολή* της H αφού ο χώρος που παράγεται από το $\mathbf{1}$ είναι υπόχωρος του στηλοχώρου του X. Αυτό σημαίνει ότι $HJ = JH = J$. Σημείωση ότι από αυτό προκύπτει ότι η προβολή $I - J$ αναλύεται στις κάθετες μεταξύ τους προβολές $H - J$ και $I - H$ από όπου προκύπτει η ισότητα $\mathrm{SSE} = \mathrm{SSR} + \mathrm{SST}$.

Θα χρησιμοποίησουμε τον συμβολισμό $\langle \boldsymbol{x} , \boldsymbol{y} \rangle$ για το ευκλείδιο εσωτερικό γινόμενο.

Το ζητούμενο αποτέλεσμα είναι ισοδύναμο με το $\sum (y_i - \hat{y}_i) = 0$, το οποίο ισχύει αφού

$$
\begin{aligned}
\sum (y_i - \hat{y}_i) &= \langle (I - H)\boldsymbol{y} , \mathbf{1} \rangle \\
&= \left\langle \boldsymbol{y} , (I - H)^\top \mathbf{1} \right\rangle \\
&= \langle \boldsymbol{y} , (I - H)\mathbf{1} \rangle \\
&= \langle \boldsymbol{y} , \mathbf{1} - H\mathbf{1} \rangle \\
&= \langle \boldsymbol{y} , \mathbf{1} - \mathbf{1} \rangle \\
&= \mathbf{0}
\end{aligned}
$$

$\square$

Άσκηση Α΄.3.

$$\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = 0$$

Απόδειξη. Έχουμε,

$$\begin{aligned}
\mathrm{Cov}(\bar{y}, \hat{\beta}_1) &= \mathrm{Cov}(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}, \hat{\beta}_1) \\
&= \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_1)\bar{x}
\end{aligned}$$

Θα βρούμε τις σχετικές συνδιακυμάνσεις από τον πίνακα διακύμανσης του $\hat{\beta}$, δηλαδή τον $\mathrm{V}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$ Έχουμε,

$$X^\top X = \begin{pmatrix} — & \mathbf{1} & — \\ — & \boldsymbol{x} & — \end{pmatrix} \begin{pmatrix} | & | \\ \mathbf{1} & \boldsymbol{x} \\ | & | \end{pmatrix} = \begin{pmatrix} \langle \mathbf{1} \, , \, \mathbf{1} \rangle & \langle \mathbf{1} \, , \, \boldsymbol{x} \rangle \\ \langle \boldsymbol{x} \, , \, \mathbf{1} \rangle & \langle \boldsymbol{x} \, , \, \boldsymbol{x} \rangle \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \|\boldsymbol{x}\|^2 \end{pmatrix}$$

Άρα,

$$(X^\top X)^{-1} = \frac{1}{n\|\boldsymbol{x}\|^2 - n^2 \bar{x}} \begin{pmatrix} \|\boldsymbol{x}\|^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}$$

Συνεπώς,

$$\begin{aligned}
\mathrm{Cov}(\bar{y}, \hat{\beta}_1) &= \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_1)\bar{x} \\
&\propto (-n\bar{x}) + (n)\,\bar{x} \\
&= 0 \\
\implies \mathrm{Cov}(\bar{y}, \hat{\beta}_1) &= 0
\end{aligned}$$

□

Άσκηση Α΄.4.

$$\sum y_i \hat{y}_i = \sum \hat{y}_i^2$$

Απόδειξη. Το ζητούμενο αποτέλεσμα είναι ισοδύναμο με το $\sum \hat{y}_i (y_i - \hat{y}_i) = 0$.
Έχουμε,

$$\begin{aligned}
\sum \hat{y}_i (y_i - \hat{y}_i) &= \langle \hat{\boldsymbol{y}} \, , \, \boldsymbol{y} - \hat{\boldsymbol{y}} \rangle \\
&= \langle H\boldsymbol{y} \, , \, (I - H)\boldsymbol{y} \rangle \\
&= \left\langle \boldsymbol{y} \, , \, H^\top (I - H)\boldsymbol{y} \right\rangle \\
&= \langle \boldsymbol{y} \, , \, H(I - H)\boldsymbol{y} \rangle \\
&= \langle \boldsymbol{y} \, , \, \mathbf{0} \rangle \\
&= 0
\end{aligned}$$

□

Άσκηση Α΄.5.

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

Απόδειξη.

$$\begin{aligned}\sum\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right) &= \langle \boldsymbol{y}-H\boldsymbol{y}\ ,\ H\boldsymbol{y}-J\boldsymbol{y}\rangle \\ &= \langle (I-H)\boldsymbol{y}\ ,\ (H-J)\boldsymbol{y}\rangle \\ &\stackrel{*}{=} 0\end{aligned}$$

Όπου η τελευταία (*) ισότητα ισχύει αφού η $H-J$ είναι η υποπροβολή της H (προβολή στις στήλες του X πλην της $\mathbf{1}$), και η $I-H$ είναι η προβολή στον κάθετο χώρο που προβάλει η H (δηλαδή στον $\mathrm{Im}(X)^{\perp}$).

Ποιο αναλυτικά,

$$\left(I-H\right)\left(H-J\right)=H-J-H^2+HJ=H-J-H+J=0$$

$\square$

Άσκηση Α΄.6.

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

Απόδειξη. Από την Α΄.3 ξέρουμε ότι

$$\begin{aligned}\mathrm{V}[\hat{\beta}_1] &= \sigma^2\frac{1}{n\left\|\boldsymbol{x}\right\|^2-n^2\bar{x}^2} \\ &= \frac{\sigma^2}{\left\|\boldsymbol{x}\right\|^2-n\bar{x}^2}\end{aligned}$$

Θα χρειαστούμε το παρακάτω αποτέλεσμα,

$$\begin{aligned}S_{xx} &= \sum\left(x_i-\bar{x}\right)^2 \\ &= \left\|(I-J)\boldsymbol{x}\right\|^2 \\ &= \boldsymbol{x}^\top(I-J)^\top(I-J)\boldsymbol{x} \\ &= \boldsymbol{x}^\top(I-J)\boldsymbol{x} \\ &= \left\|\boldsymbol{x}\right\|^2-\boldsymbol{x}^\top\left(Jx\right) \\ &= \left\|\boldsymbol{x}\right\|^2-\boldsymbol{x}^\top(\bar{x}\mathbf{1}) \\ &= \left\|\boldsymbol{x}\right\|^2-\bar{x}(\boldsymbol{x}^\top\mathbf{1}) \\ &= \left\|\boldsymbol{x}\right\|^2-\bar{x}(n\bar{x}) \\ &= \left\|\boldsymbol{x}\right\|^2-n\bar{x}^2\end{aligned}$$

Έτσι καταλήγουμε στο ότι

$$\mathrm{V}[\hat{\beta}_1]=\frac{\sigma^2}{S_{xx}}$$

απο το οποίο παίρνουμε

$$\mathrm{se}(\hat{\beta}_1)^2 = \frac{s^2}{S_{xx}}$$

Έχουμε,

$$\begin{aligned} r_{xy}^2 &= R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \\ \implies 1 - r_{xy}^2 &= \frac{\mathrm{SSE}}{\mathrm{SST}} \\ \implies \mathrm{SST}\left(1 - r_{xy}^2\right) &= \mathrm{SSE} \end{aligned} \tag{*}$$

Χρησμοποιόντας τις ισότητες $s^2 = \frac{\mathrm{SSE}}{n-2}$ και $\mathrm{SST} = S_{yy}$, τότε από την (*) συμπεραίνουμε ότι

$$s^2 = \frac{1}{n-2} S_{yy}\left(1 - r_{xy}^2\right)$$

Χρησμοποιόντας την παραπάνω έκφραση έχουμε

$$\begin{aligned} \left(\frac{\hat{\beta}_1}{\mathrm{se}(\hat{\beta}_1)}\right)^2 &= \frac{\hat{\beta}_1^2}{\frac{s^2}{S_{xx}}} \\ &= \frac{S_{xy}^2/S_{xx}^2}{\frac{1}{S_{xx}}\frac{1}{n-2} S_{yy}\left(1 - r_{xy}^2\right)} \\ &= \frac{S_{xy}^2(n-2)}{S_{xx}S_{yy}\left(1 - r_{xy}^2\right)} \\ &= \frac{r_{xy}^2(n-2)}{1 - r_{xy}^2} \end{aligned}$$

Συνεπώς,

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

□

B′ Μέρος

Τα δεδομένα στο αρχείο cholesterol.txt αφορούν επίπεδα ολικής χοληστερόλης (mg/ml) 24 ασθενών (y) και την ηλικία τους (x).

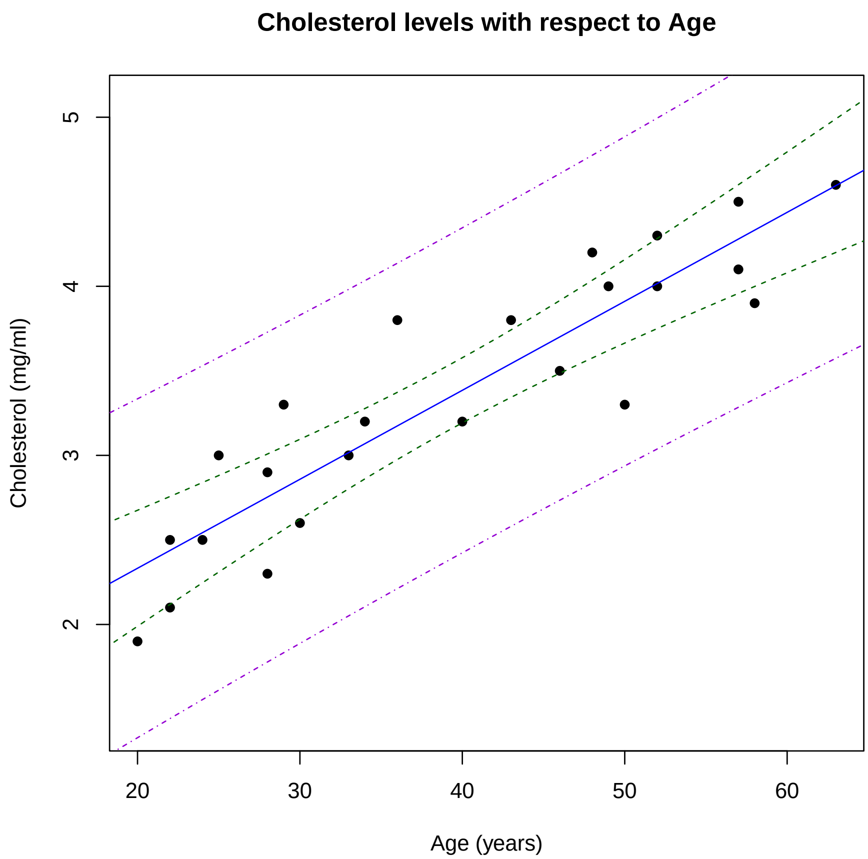
B′.i

Εκφώνηση

Να κατασκευαστεί ένα διάγραμμα διασποράς μεταξύ των δύο μεταβλητών y και x και να προσαρμοστεί το μοντέλο $\mathrm{E}[y] = \hat{\beta}_0 + \hat{\beta}_1 x$.

Απάντηση

Στο Σχήμα 1 βλέπουμε το σχετικό διάγραμμα



Σχήμα 1: Διάγραμμα διασποράς για τα επίπεδα χοληστερόλης μαζί με διαστήματα εμπιστοσύνης και πρόβλεψης

B′.ii

Εκφώνηση

Να γίνει ο έλεγχος $H_0 : \beta_1 = 0$ έναντι της $H_0 : \beta_1 \neq 0$ και επιπλέον να προσδιοριστεί ένα 0.95-διάστημα εμπιστοσύνης (δ.ε.) για το συντελεστή της x στο μοντέλο που προσαρμόστηκε. Πώς ερμηνεύουμε το $\hat{\beta}_1$;

Απάντηση

Η $H_0 : \beta_1 = 0$ γίνεται δεκτή αν και μόνο αν το επίπεδο σημαντικότητας του ελέγχου έχει τεθεί μικρότερο του 0.0000000009428305. Πρακτικά δηλαδή, είμαστε βέβαιοι πως η H_0 δεν ισχύει. Το β_1 εκτιμάται να είναι 0.052625 και το ζητούμενο 0.95-διάστημα εμπιστοσύνης είναι το [0.04185806, 0.06339175]. Η ερμηνεία του β_1 είναι ότι, αν η ηλικία κάποιου αυξηθεί κατά 1 χρόνο, τότε αναμένουμε το επίπεδο χοληστερόλης του να αυξηθεί κατά β_1.

B′.iii

Εκφώνηση

Να κατασκευαστεί ένα 0.99-δ.ε. πρόβλεψης για το επίπεδο χοληστερόλης y ενός ασθενή ηλικίας 35 ετών, καθώς και για την αναμενόμενη τιμή της, $\mathrm{E}[y]$.

Απάντηση

Το $\mathrm{E}[Y]$ εκτιμάται να είναι 3.12174. Το ζητούμενο 0.99-διάστημα πρόβλεψης για το Y είναι το [2.158578, 4.084902], ενώ το ζητούμενο 0.99-διάστημα εμπιστοσύνης για το $\mathrm{E}[Y]$ είναι το [2.918965, 3.324515].

B′.iv

Εκφώνηση

Να γίνει ο γραφικός έλεγχος της Κανονικής κατανομής και η γραφική παράσταση e_i με $\hat{y}_i$, για τα υπόλοιπα e_i. Τι συμπεραίνετε;

Απάντηση

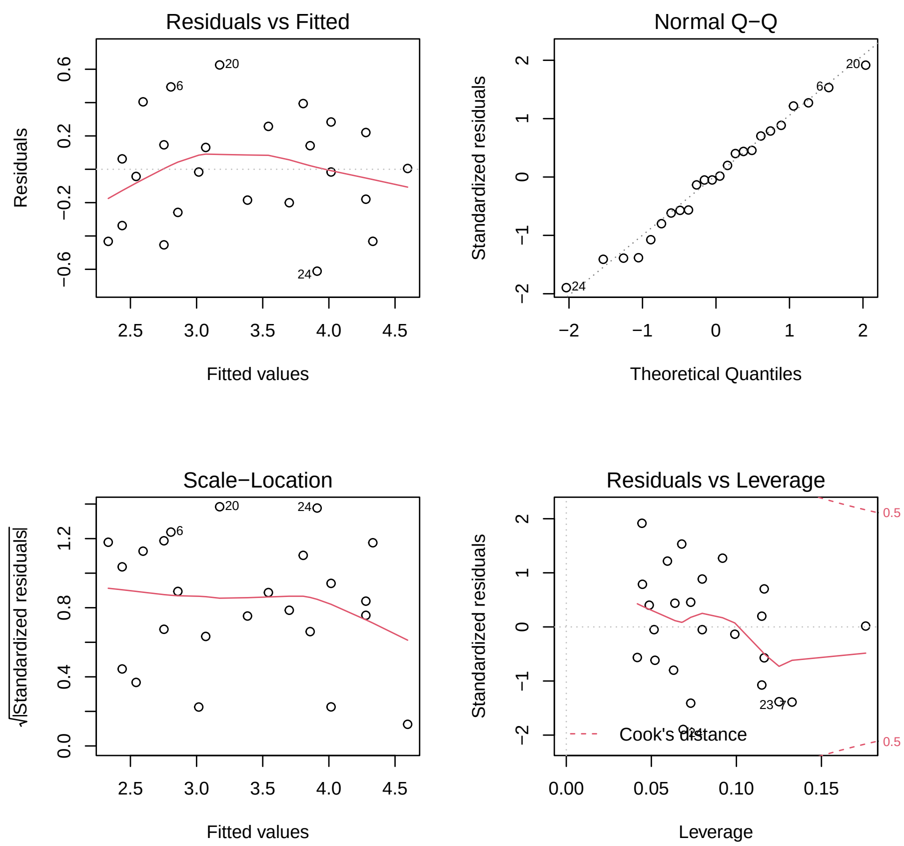
Στο Σχήμα 2 βλέπουμε το σχετικό διάγραμμα. Τα συμπεράσματα είναι

- Η ομοσκεδαστικότητα φαίνεται να ισχύει για τα υπόλοιπα, αφού για όλα τα y η απόκλιση από την 0-γραμμή είναι πάνω κάτω η ίδια.
- Στο QQ-plot η αντιστοίχιση είναι σχεδόν τέλεια με τα θεωρητικά ποσοστημόρια, που σημαίνει πως μπορούμε με ασφάλεια να υποθέσουμε ότι το y ακολουθεί κανονική κατανομή.

B′.v

Εκφώνηση

Να κατασκευαστεί ένα διάγραμμα διασποράς μεταξύ των δύο μεταβλητών y και x και να προσαρμοστεί ένα μοντέλο της μορφής $y = 3 - ae^{\beta x}$.



Σχήμα 2: Διάγραμματα ελέγχου

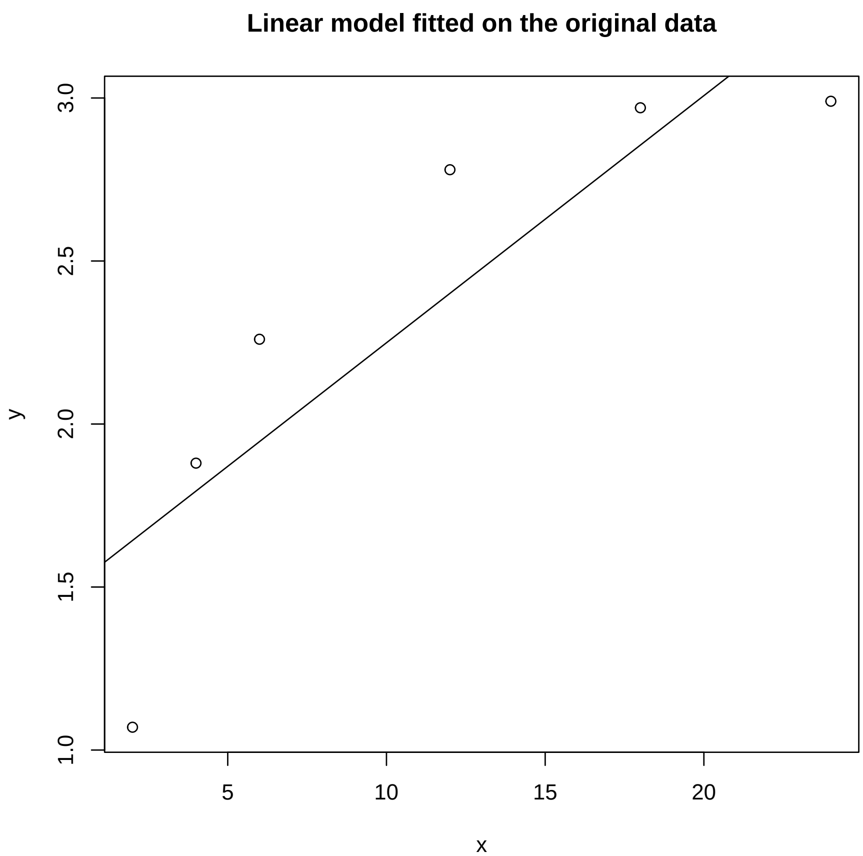
Απάντηση

Στα Σχήματα 3, 4 και 5 βλέπουμε τα σχετικά διαγράμματα διασποράς.

Β΄.vi

Εκφώνηση

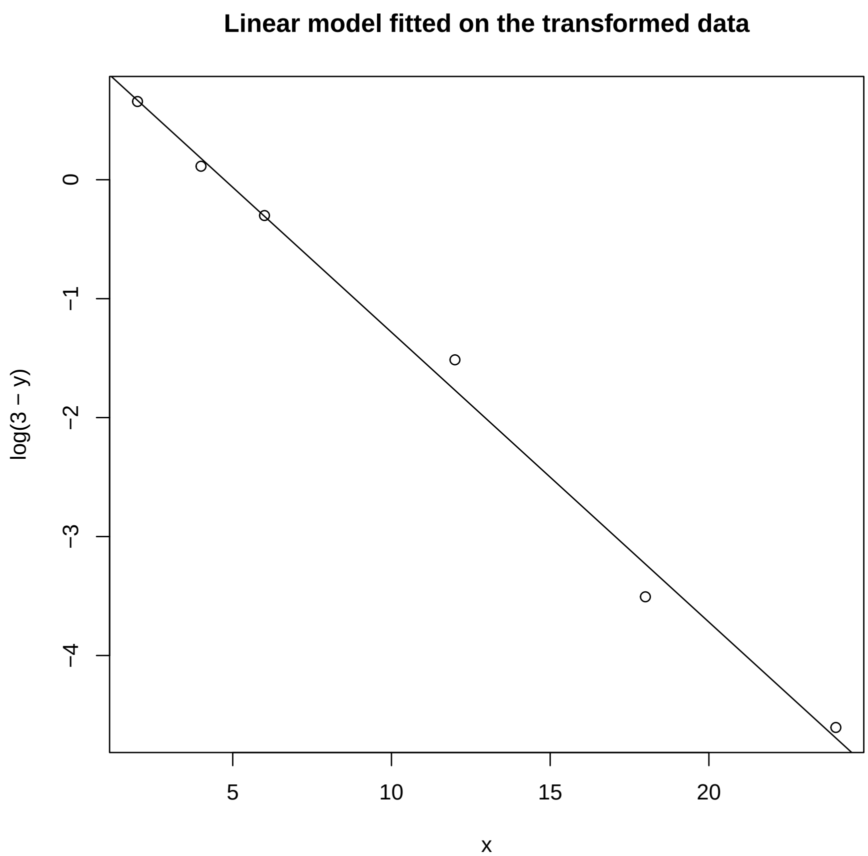
Να εκτιμηθεί σημειακά η άγνωστη παρατήρηση y και να κατασκευαστεί ένα 95% διάστημα εμπιστοσύνης (δ.ε.) για την πρόβλεψη της παρατήρησης y, καθώς και ένα προσεγγιστικό 95% δ.ε. για τη μέση τιμή της, $\mathrm{E}[y]$, όταν $x = 9$.



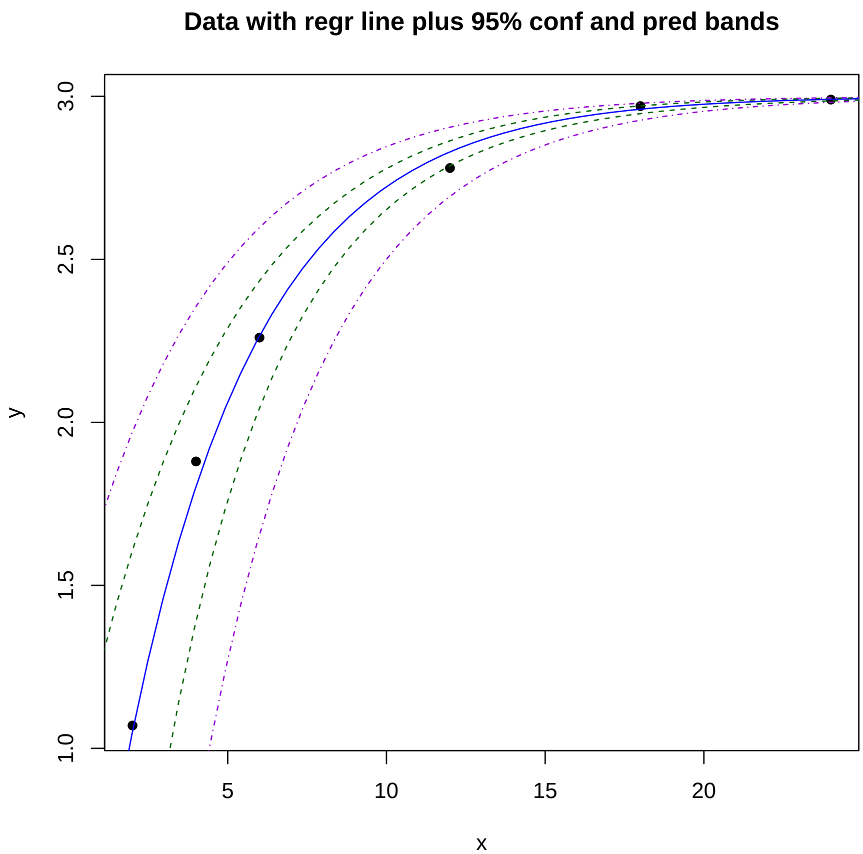
Σχήμα 3: Αρχικό διάγραμμα διασπορά, χωρίς μετασχηματισμό.

Απάντηση

Η εκτίμηση για το y είναι 2.646078. Το 0.95-διάστημα πρόβλεψης για το y είναι το [2.361758, 2.803741]. Το 0.95-διάστημα εμπιστοσύνης για το $\mathrm{E}[y]$ είναι το [2.555063 2.718475]. Στο σχήμα 5 φαίνονται και οι 'λωρίδες' πρόβλεψης και εμπιστοσύνης.



Σχήμα 4: Τελικό διάγραμμα διασποράς.



Σχήμα 5: Τελικό διάγραμμα διασποράς, μετά από αντίστροφο μετασχηματισμό. Συμπεριλαμβάνονται διαστήματα πρόβλεψης και εμπιστοσύνης