Άσκηση Α

Για το απλό γραμμικό μοντέλο $E(Y|x) = \beta_0 + \beta_1 x$:

1) Δείξτε ότι: $R^2 = r^2_{(x,y)}$, όπου R^2 ο συντελεστής προσδιορισμού και $r_{(x,y)}$ ο δειγματικός συντελεστής συσχέτισης (Pearson).

Ας υποθέσουμε ότι έχουμε n παρατηρήσεις $(x_1, y_1), (x_2, y_2)...(x_n, y_n)$ από ένα απλό γραμμικό μοντέλο:

$$Y_i = \beta_0 + \beta_1 x + \varepsilon_i \tag{1}$$

για $i = 1, 2, ..., n$. Επίσης, θα συμβολίσουμε $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ για $i = 1, 2, ..., n$, όπου $\hat{y}$ υποδεικνύει μια πρόβλεψη του Y, και $\hat{\beta}_0, \hat{\beta}_1$ είναι οι estimators ελαχίστων τετραγώνων των αντίστοιχων παραμέτρων β_0, β_1.

Οι εφράσεις γι'αυτά, μπορούν εύκολα να υπολογιστούν σύμφωνα με τη μέθοδο ελαχίστων τετραγώνων. Το i^{th} residual ορίζεται ως: $e_i = y_i - \hat{y}_i$, δηλαδή η διαφορά της πραγματικής τιμής από αυτή που προβλέπουμε.
Έτσι, βρίσκοντας το ελάχιστο του Residual Sum of Squares:

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

μπορούμε να τα εκτιμήσουμε, εμείς όμως τα θεωρούμε δεδομένα.

Επομένως από την $\frac{\partial RSS}{\partial \hat{\beta}_0} = 0$ και την $\frac{\partial RSS}{\partial \hat{\beta}_1} = 0$, και φυσικά για θετικά ορισμένες τις παραγώγους 2ης τάξης, έχουμε:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx}} \tag{3}$$

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

Αφού έχουμε αυτά που χρειαζόμαστε, ας πάμε στο ερώτημα. Ο συντελεστής προσδιορισμού ορίζεται ως:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} \tag{4}$$

ενώ ο δειγματικός συντελεστής συσχέτισης Pearson:

$$r_{(x,y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

Θα ξεκινήσουμε υπολογίζοντας την παράσταση:

$$\begin{aligned}
\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\
&\overset{(2)}{=} \sum_{i=1}^{n}(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\
&= \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&\overset{(3)}{=} \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \\
&= \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

Αντικαθιστώντας την σχέση που υπολογίσαμε στην σχέση για το R^2:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}\right)^2 \\ &\overset{(5)}{=} \left(r_{(x,y)}\right)^2 \end{aligned}$$

Επομένως, αποδείχτηκε το ζητούμενο.

2) Δείξτε ότι $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$:

Γνωρίζουμε πως ισχύει: $y_i = \hat{y}_i + \hat{\varepsilon}_i$. Επίσης γνωρίζουμε ότι η μέση τιμή του $\hat{\varepsilon}_i$ είναι μηδενική, επομένως έχουμε:

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i + \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i \quad (\to 0)$$

και έτσι έχουμε:

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

3) Δείξτε ότι $\text{cov}(\bar{y}, \hat{\beta}_1) = 0$

Από την σχέση (3) μπορούμε να γράψουμε ισοδύναμα ότι:

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

Όπου $\forall i$ έχουμε $c_i = (x_i - \bar{x})/\text{S}_{xx}$. Άρα έχουμε:

$$\begin{aligned}
\text{Cov}\left(\bar{y}, \hat{\beta}_1\right) &= \text{Cov}\left(\frac{1}{n}\sum y_i, \sum c_i y_i\right) \\
&= \frac{1}{n}\sum c_i \,\text{Cov}\left(y_i, y_i\right) \\
&= \frac{\sigma^2}{n}\sum c_i \\
&= 0
\end{aligned}$$

Για το παραπάνω, θα σχολιάσουμε δύο πράγματα:

►Ο λόγος που το $\sum c_i$ βγήκε έξω από το Cov είναι επειδή θεωρούμε ότι αυτή η ροπή που υπολογίζουμε είναι conditional moment επομένως υπολογίζεται για μια σταθερή τιμή του x, η οποία δεν είναι τυχαία μεταβλητή. Επομένως αφού το c_i εξαρτάται από το x, είναι fixed.

►Στην τελευταία ισότητα ισχύει:

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\text{S}_{\text{xx}}} = \frac{n\bar{x} - \bar{x}\sum_{i=1}^{n} 1}{\text{S}_{\text{xx}}} = \frac{n\bar{x} - n\bar{x}}{\text{S}_{\text{xx}}} = 0$$

4) Δείξτε ότι $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$:

Ξεκινώντας έχουμε τα εξής, από αντικατάσταση των αρχικών γνωστών σχέσεων.

$$y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}) \tag{6}$$

$$\hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x}) \Rightarrow \tag{7}$$

$$\boldsymbol{y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)} \tag{8}$$

Η τελευταία σχέση (8), θα μας χρειαστεί σε επόμενο ερώτημα.

Από τις (6) και (7) έχουμε:

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \sum_{i=1}^{n}(\hat{\beta}_1(x_i - \bar{x}))(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})) = \sum_{i=1}^{n}\hat{\beta}_1(\mathrm{S}_{xy} - \hat{\beta}_1\mathrm{S}_{xx}) = 0$$

εφόσον γνωρίζουμε ότι $\hat{\beta}_1 = \mathrm{S}_{xy}/\mathrm{S}_{xx}$.

5) Δείξτε ότι:

$$\mathrm{S}^2_{\mathrm{Y_x}} = \frac{\sum_{i=1}^{n}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2}{\mathrm{n}-2} = \frac{1}{\mathrm{n}-2}\left\{\mathrm{S_{yy}} - \frac{\mathrm{S^2_{xy}}}{\mathrm{S_{xx}}}\right\} = \frac{\mathrm{S_{yy}}}{\mathrm{n}-2}\left\{1 - \mathrm{r^2_{xy}}\right\}$$

►Για το πρώτο σκέλος της ισότητας, αρκεί να δείξουμε ότι:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathrm{S}_{yy} - \frac{\mathrm{S}^2_{xy}}{\mathrm{S}_{xx}}$$

Αν τετραγωνίσουμε και τις δύο πλευρές τις εξίσωσης (8), και αθροίσουμε από $i = 1, 2, ..., n$:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \cancelto{0}{2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)}$$

Για τον Cross Product Term, αποδείξαμε ότι κάνει μηδέν, στο προηγούμενο ερώτημα.

Για τα υπόλοιπα έχουμε:

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 &= \sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\\ &\overset{(7)}{=} \mathrm{S}_{yy}-\sum_{i=1}^{n}\hat{\beta}_1{}^2(x_i-\bar{x})^2\\ &= \mathrm{S}_{yy}-\frac{\mathrm{S}_{xy}^2}{\mathrm{S}_{xx}^2}\cdot\mathrm{S}_{xx}\\ &= \mathrm{S}_{yy}-\frac{\mathrm{S}_{xy}^2}{\mathrm{S}_{xx}}\end{aligned}$$

Επομένως, εισάγοντας και την σταθερά $\frac{1}{n-2}$ που μας λείπει για το ερώτημα, έχουμε το πρώτο σκέλος του ζητήματος:

$$\mathrm{S}_{\mathrm{Y_x}}^2=\frac{\sum_{i=1}^{\mathrm{n}}\left(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i}\right)^2}{\mathrm{n}-2}=\frac{1}{\mathrm{n}-2}\left\{\mathrm{S_{yy}}-\frac{\mathrm{S_{xy}^2}}{\mathrm{S_{xx}}}\right\}$$

►Για το δεύτερο σκέλος της ισότητας, αρκεί να βγάλουμε κοινό παράγοντα το S_{yy} και να παρατηρήσουμε ότι:

$$\mathrm{r}_{xy}=\frac{\mathrm{S}_{xy}}{\sqrt{\mathrm{S}_{xx}\mathrm{S}_{yy}}}$$

Έτσι έχουμε:

$$\frac{1}{\mathrm{n}-2}\left\{\mathrm{S_{yy}}-\frac{\mathrm{S_{xy}^2}}{\mathrm{S_{xx}}}\right\}=\frac{\mathrm{S}_{yy}}{\mathrm{n}-2}\cdot\left\{1-\frac{\mathrm{S}_{xy}^2}{\mathrm{S}_{xx}\mathrm{S}_{yy}}\right\}=\frac{\mathrm{S}_{yy}}{\mathrm{n}-2}\cdot\left\{1-r_{xy}^2\right\}$$

Άρα ισχύει και το δεύτερο μέρος της ισότητας, και το ζητούμενο αποδείχθηκε.

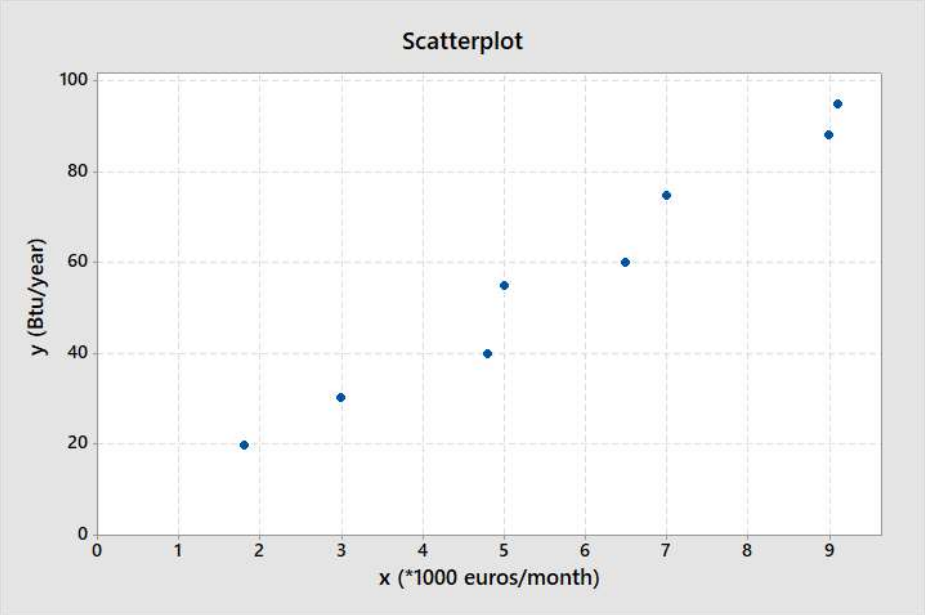
Άσκηση Β

Θέλουμε να εξετάσουμε τη σχέση μεταξύ ενεργειακής κατανάλωσης και των εσόδων ενός νοικοκυριού. Έτσι, στον παρακάτω πίνακα με x συμβολίζεται το εισόδημα (σε 1000 euro/μήνα) και με y η ενεργειακή κατανάλωση (σε Btu/χρόνο).

x	1.8	3.0	4.8	5.0	6.5	7.0	9.0	9.1
y	20.0	30.5	40.0	55.1	60.3	74.9	88.4	95.2

---

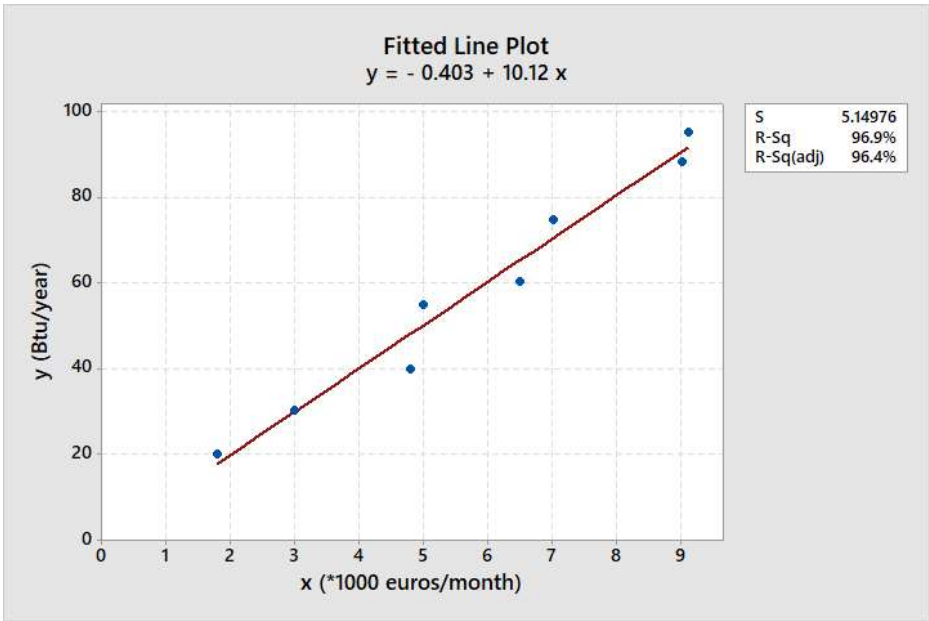
1) Να γίνει το διάγραμμα διασποράς των μεταβλητών y και x.



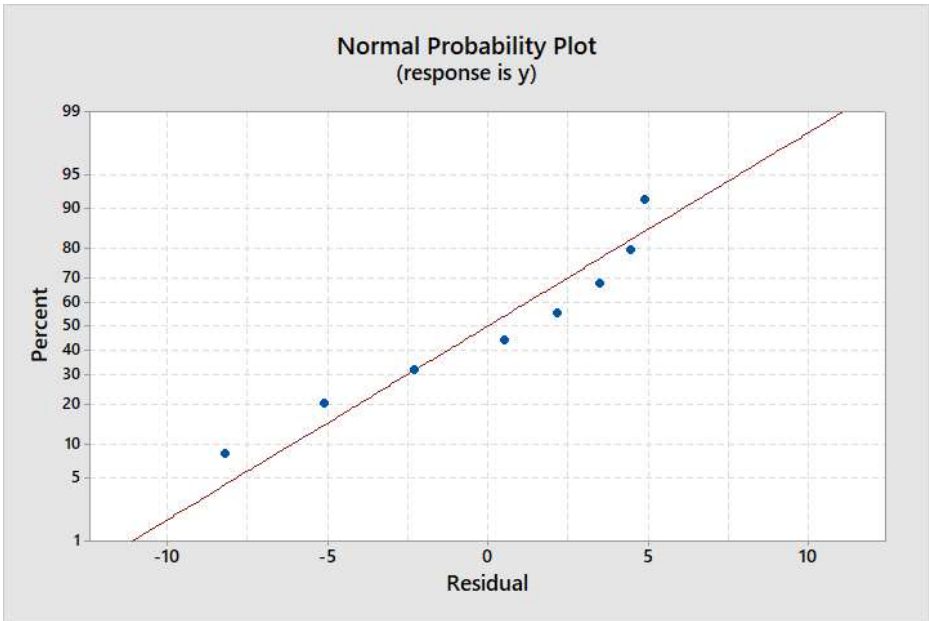
2) Εκτίμηση της εξίσωσης παλινδρόμησης $E(Y|x) = \beta_0 + \beta_1 x$:

Όπως φαίνεται και στο παρακάτω σχήμα, η εκτίμησης της εξισώσεως παλινδρόμησης είναι:

$$y = -0.403 + 10.12 \cdot x \tag{9}$$



3) Να γίνει ο γραφικός έλεγχος της Κανονικής Κατανομής για τα υπόλοιπα e.



Γνωρίζουμε ότι για να ακολουθούν την κανονική κατανομή τα residuals, θα πρέπει να βρίσκονται σε μια ευθεία γραμμή. Στο διάγραμμά μας παρατηρούμε ότι έχουν μια γραμμική τάση, όμως εμφανίζουν κάποια καμπυλότητα στις ουρές. Αυτό, δεδομένου ότι έχουμε μόνο 6 σημεία για τα δεδομένα μας, δεν σημαίνει ότι δεν ακολουθούν την κανονική κατανομή, καθώς είναι πολύ μικρός ο αριθμός των σημείων. Όσο ο αριθμός των παρατηρήσεων μας μειώνεται, τόσο μπορεί να εμφανιστεί variation και μη-γραμμικότητα στο probability plot μας.

4) Αν $x_0 = 8$, να εκτιμηθεί η αντίστοιχη ενεργειακή κατανάλωση Y. Επίσης, να κατασκευαστεί ένα 95% δ.ε. για την παρατήρηση Y καθώς και για την μέση τιμή της, $E(Y)$. Με την χρήση του minitab (έκδοση 17), επιλέγοντας: *Stat* $\rightarrow$ *Regression* $\rightarrow$ *Regression* $\rightarrow$ *Predict* και για $x_0 = 8$ έχουμε:

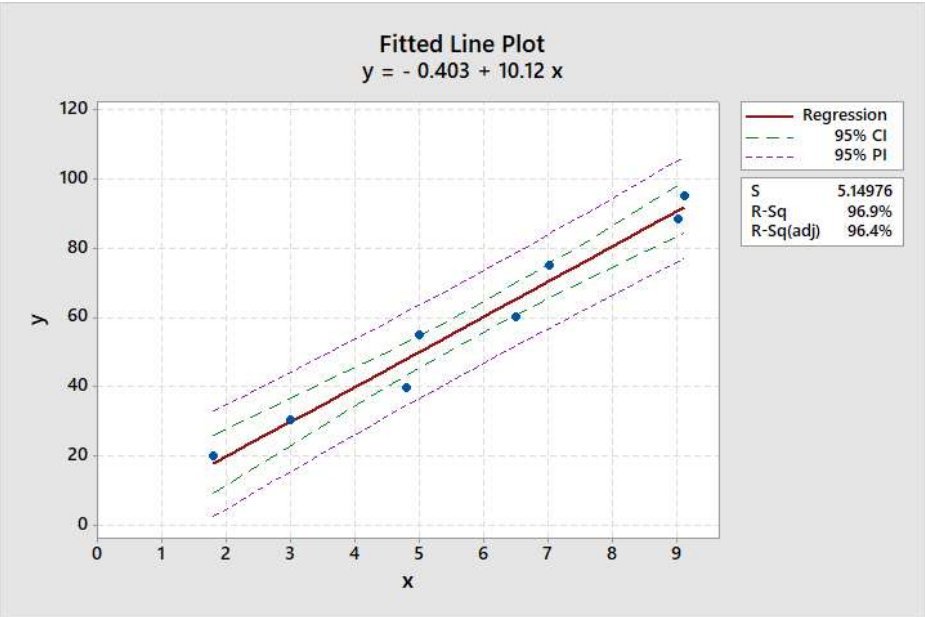
```
Prediction for y
Regression Equation
y = -0.403 + 10.122x

Variable  Setting
x               8
    Fit   SE Fit        95% CI              95% PI
80.5709  2.45359  (74.5672, 86.5746)  (66.6128, 94.5291)
```

Ας εξηγήσουμε λίγο τι σημαίνουν αυτά τα αποτελέσματα:

►Με βάση τα δεδομένα και την εξίσωση παλινδρόμησης, το minitab προβλέπει ότι για έσοδα x = 8000 euros/μήνα, η (fitted) μέση τιμή για την ενεργειακή κατανάλωση είναι 80.5709 Btu/χρόνο, με standar error 2.45359 Btu/χρόνο.

►Με βάση τα δεδομένα της ενεργειακής κατανάλωσης, μπορούμε να εκτιμήσουμε με 95% confidence ότι η **μέση ενεργειακή κατανάλωση** είναι ανάμεσα στα 74.5672 με 86.5746 Btu/χρόνο, για νοικοκυριά με έσοδα x = 8000 euros/μήνα.

►Με βάση τα δεδομένα της ενεργειακής κατανάλωσης, μπορούμε να εκτιμήσουμε με 95% confidence ότι **η προβλεπόμενη ενεργειακή κατανάλωση για μια μεμονωμένη νέα παρατήρηση** βρίσκεται ανάμεσα στα 66.6128 με 94.5291 Btu/χρόνο για νοικοκυριά με έσοδα x = 8000 euros/μήνα.

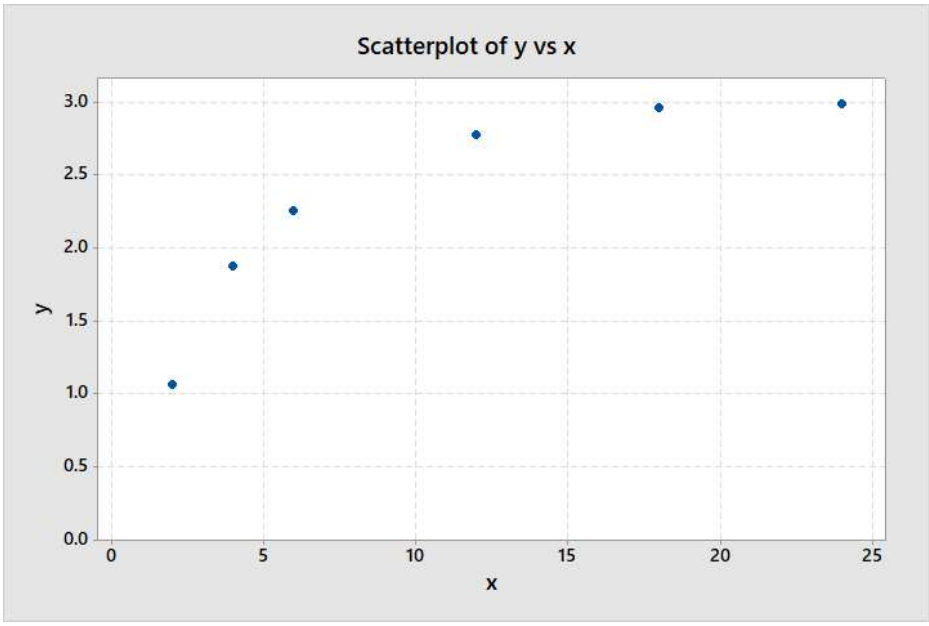


Άσκηση Γ

Για τα παρακάτω δεδομένα:

x	2	4	6	12	18	24
y	1.07	1.88	2.26	2.78	2.97	2.99

1) Να κατασκευαστεί το διάγραμμα διασποράς μεταξύ Y και X.



Παρατηρούμε πως η σχέση μεταξύ των δυο μεταβλητών δεν είναι γραμμική.

2) Μετά από κατάλληλο μετασχηματισμό να προσαρμοστεί ένα μοντέλο της μορφής $Y = 3 - \alpha e^{\beta x}$ και να κατασκευαστεί η γραφική παράσταση των υπολοίπων e επί των εκτιμηθέντων $\hat{y}$.

Ξεκινάμε το μετασχηματισμό του μοντέλου μας:

$$Y = 3 - \alpha e^{\beta x} \Rightarrow 3 - Y = \alpha e^{\beta x} \Rightarrow ln(3 - Y) = ln(\alpha e^{\beta x})$$
$$ln(3 - Y) = ln(\alpha) + \beta x \tag{10}$$

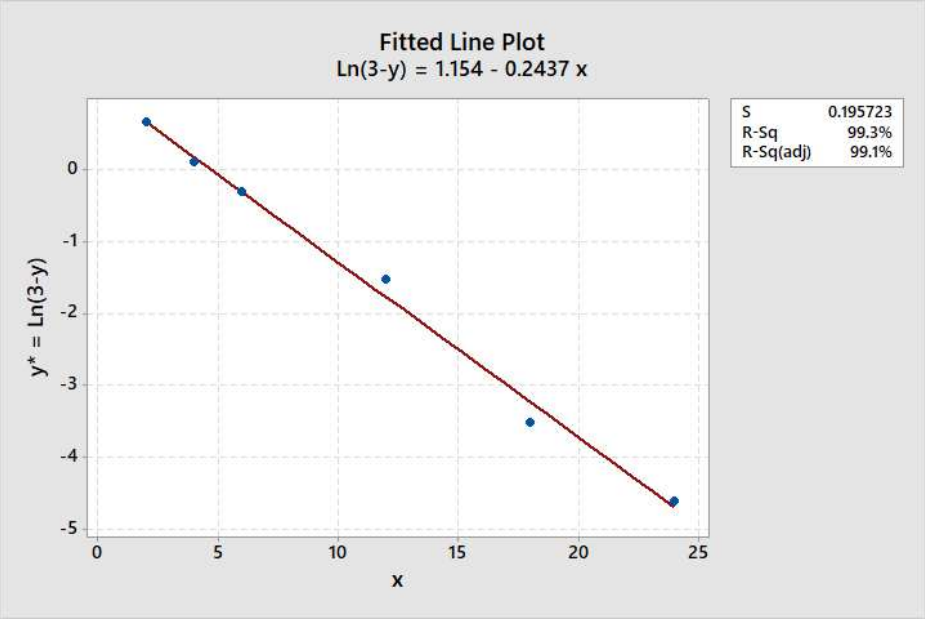
Όπου θέτοντας $y^* = 3 - Y$ (και όχι για παράδειγμα $Y - 3$ γιατί παρατηρώντας τα δεδομένα μας $y_i < 3 \; \forall \; i$ άρα το όρισμα του λογαρίθμου θα ήταν αρνητικό) καθώς και $\beta_0 = ln(\alpha)$ και $\beta_1 = \beta$, έχουμε ανάξει το μοντέλο μας στο απλό γραμμικό της μορφής:

$$y^* = \beta_0 + \beta_1 x \tag{11}$$

Επομένως, για κάθε y που έχουμε στα δεδομένα μας πρέπει να υπολογίσουμε και το αντίστοιχο $ln(3-y)$.

x	2	4	6	12	18	24
y	1.07	1.88	2.26	2.78	2.97	2.99
$\mathbf{ln(3-y)}$	0.6575	0.1133	-0.3011	-1.5141	-3.5065	-4.6051

Είμαστε πλέον έτοιμοι, να κάνουμε fit την καινούρια μας εξίσωση και να βρούμε την εξίσωση παλινδρόμησης:

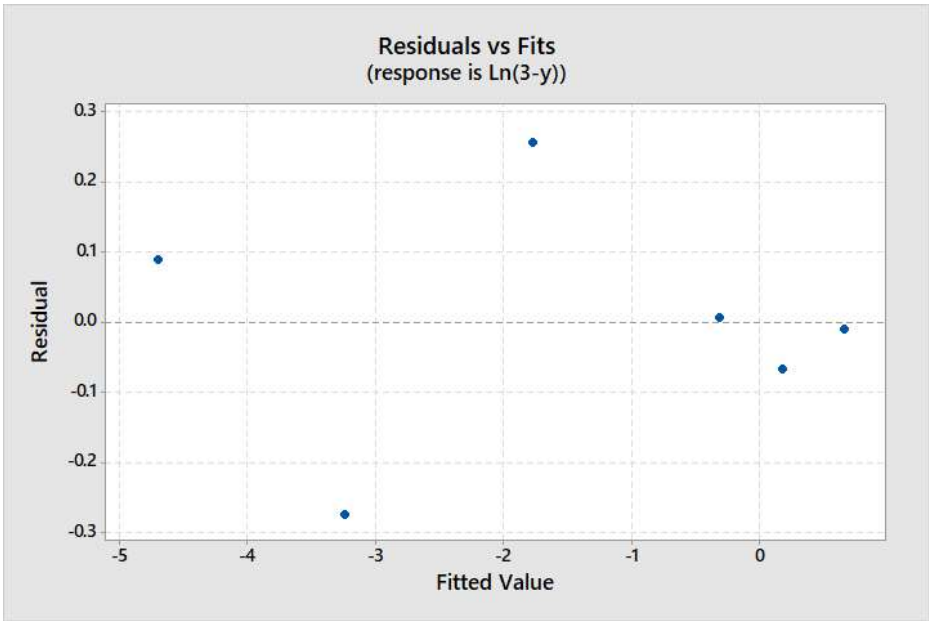


Έχοντας από την εξίσωση παλινδρόμησης τους συντελεστές $\beta_0 = 1.154$ και $\beta_1 = -0.2437$, μπορούμε να υπολογίσουμε τους συντελεστές του αρχικού μοντέλου: $\alpha = e^{\beta_0} = e^{1.154} = 3.171$ και $\beta = \beta_1 = -0.244$. Άρα το αρχικό μοντέλο γίνεται:

$$Y = 3 - 3.171 \cdot e^{-0.244 \cdot x} \tag{12}$$

Για την γραφική παράσταση των υπολοίπων e επί των εκτιμηθέντων $\hat{y}$, στο minitab επιλέγουμε το γράφημα Residuals versus fits.

Με μια πρώτη ματιά, τα υπόλοιπα φαίνεται να είναι τυχαία κατανεμημένα γύρω από το μηδέν, αλλά όπως και στην προηγούμενη άσκηση, είναι πολύ μικρός ο αριθμός των παρατηρήσεων προκειμένου να ελέγξουμε αν υπάρχει κάποιο μοτίβο, ή για παράδειγμα αν υπάρχουν outliers.



3) Να εκτιμηθεί σημειακά η άγνωστη παρατήρηση Y και να κατασκευαστεί ένα 99% δ.ε. για την πρόβλεψη της παρατήρησης Y_{x_0}, για δοθέν $x_0 = 8$.

Η σημειακή άγνωστη παρατήρηση για $x_0 = 8$, μπορεί να βρεθεί με απλή αντικατάσταση είτε στην σχέση (12) είτε με την εξίσωση παλινδρόμησης, λύνοντας ως προς y. Παρόμοια με την (Β) άσκηση, κάνουμε predict με την χρήση του minitab:

```
Prediction for Ln(3-y)

Regression Equation
Ln(3-y) = 1.154 -0.2437x

Variable  Setting
x               8

Fit         SE Fit          99% CI                99% PI
-0.795008  0.0854778  (-1.18856, -0.401460)  (-1.77833, 0.188310)
```

Όμως εδώ θέλει προσοχή, δεν έχουμε υπολογίστει το αρχικό y αλλά το $Ln(3-y)$.

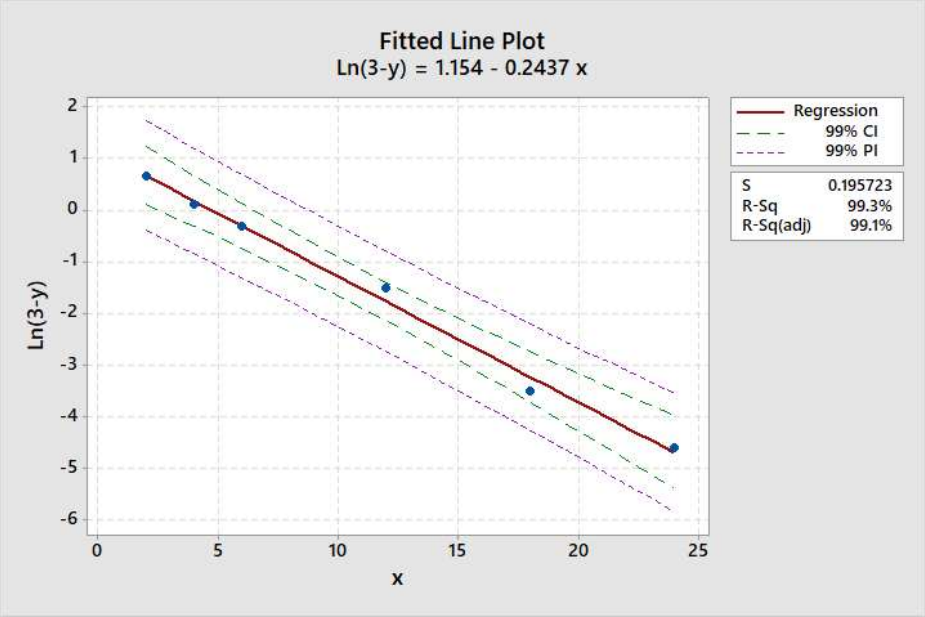
►Επομένως έχουμε για την πρόβλεψη:

$$Ln(3-y) = -0.795008 \Rightarrow y = 3 - e^{-0.795008} = 2.548422$$

Το ίδιο ισχύει για όλες τις τιμές που έχουν υπολογιστεί εδώ.

►Άρα, για να υπολογίσουμε το 99% **διάστημα εμπιστοσύνης**, εφαρμόζουμε τον αντίστροφο μετασχηματισμό στις τιμές για το 99% Prediction Interval που βρήκαμε παραπάνω, και αντιστρέφουμε το διάστημα διότι η συνάρτηση $f(x) = ln(3-x)$ είναι γνησίως φθίνουσα που σημαίνει ότι για $x_1 < x_2 \Rightarrow f(x_1) > f(x_2)$.

Επομένως, το το 99% **διάστημα εμπιστοσύνης** είναι $\mathbf{(1.7928, 2.8311)}$.



►Μπορούμε να υπολογίσουμε και το Confidence interval για την μέση τιμή του Y, όμως εφαρμόζοντας τον αντίστροφο μετασχηματισμό σε αυτό το διάστημα έχουμε μόνο μια **προσεγγιστική τιμή**, το διάστημα για αυτό είναι: $\mathbf{(2.330, 2.695)}$.

♠ ♠