ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Ε.ΔΕ.Μ.Μ.

ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ

1η ΣΕΙΡΑ ΑΣΚΗΣΕΩΝ

ΧΕΙΜΕΡΙΝΟ ΕΞΑΜΗΝΟ 2021-2022

[redacted]

[redacted]

([redacted])

ΑΣΚΗΣΗ Β

Η συγκεκριμένη υπολογιστική άσκηση εστιάζει στην στατιστική ανάλυση των δεδομένων του αρχείου cholesterol.txt, το οποίο περιέχει πληροφορίες ασθενών σχετικά με τα επίπεδα ολικής χοληστερόλης (mg/ml) – μεταβλητή y , ανάλογα με την ηλικία τους – μεταβλητή x . Το μέγεθος του δείγματος των ασθενών ανέρχεται στους 24 και η επεξεργασία πραγματοποιήθηκε με Rstudio.

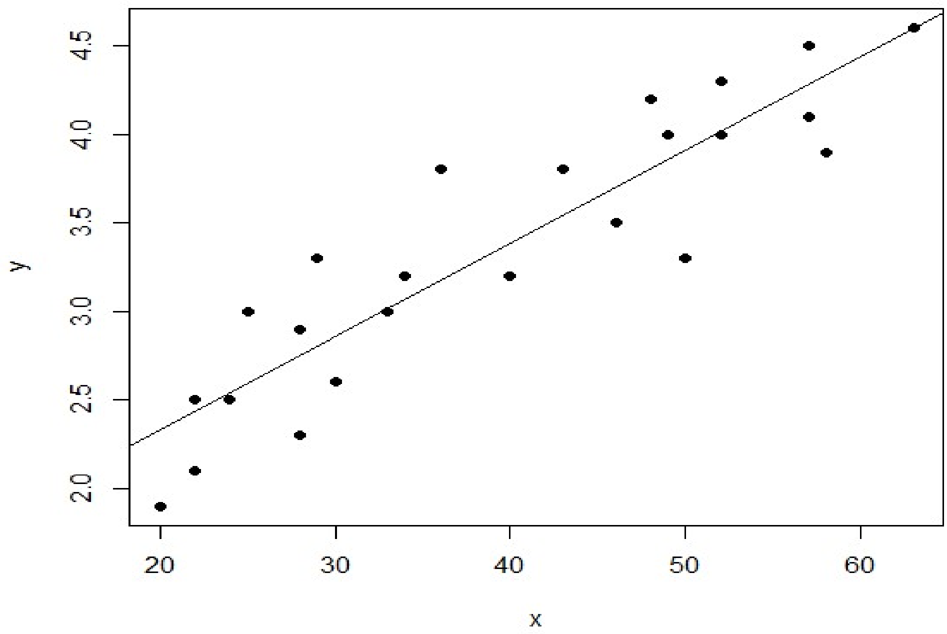
Αρχικά, τα δεδομένα διαβάζονται από το λογισμικό μέσω της εντολής read.table() και ανατίθενται στις αντίστοιχες μεταβλητές x και y μέσω της εντολής attach(). Το πρώτο βήμα της επεξεργασίας στοχεύει στην αναπαράσταση των τιμών y για τα επίπεδα χοληστερόλης ως προς την μεταβλητή x της ηλικίας των ασθενών. Το αντίστοιχο διάγραμμα διασποράς παρουσιάζεται στην Εικόνα 1. Στη συνέχεια, προσαρμόστηκε στα δεδομένα το Απλό Γραμμικό Μοντέλο σύμφωνα με τη σχέση:

$$E(y) = \beta_o + \beta_1 \chi$$

Το μοντέλο προσαρμόζεται από την R μέσω της εντολής mod1 = lm(y~x) και η αντίστοιχη ευθεία απεικονίζεται μεταξύ των δεδομένων στην Εικόνα 1. Οι εκτιμήσεις για τους συντελεστές παλινδρόμησης β_0 και β_1 που προκύπτουν είναι:

$$\hat{\beta}_0 = 1.27987 \pm 0.21570$$

$$\hat{\beta}_1 = 0.05263 \pm 0.00519$$



Εικόνα 1: Διάγραμμα Διασποράς δεδομένων και προσαρμοσμένη ευθεία σύμφωνα με το Απλό Γραμμικό Μοντέλο.

Ύστερα πραγματοποιείται ο έλεγχος για την συνεισφορά των δεδομένων της x μεταβλητής στο μοντέλο. Ο έλεγχος αυτός συνεπάγεται με τον έλεγχο της μηδενικής υπόθεσης H_0 έναντι της εναλλακτικής H_1 όσον αφορά τον συντελεστή παλινδρόμησης του x σύμφωνα με τη σχέση:

$$H_0: \hat{\beta}_1 = 0 \quad \text{ενώ} \quad H_1: \hat{\beta}_1 \neq 0$$

Ο έλεγχος βασίζεται στο t-test, όπου η μεταβλητή t υπολογίζεται με βάση την τιμή του $\hat{\beta}_1$ για την μηδενική υπόθεση:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_1(H_o)}{S\sqrt{\frac{1}{\sum_{i=1}^{24}(x_i - \bar{x})}}} = 10.136$$

όπου S η εκτιμήτρια της διασποράς των τυχαίων σφαλμάτων του μοντέλου. Η αντίστοιχη p-value για επίπεδο σημαντικότητας α = 0,05 για το συγκεκριμένο test υπολογίζεται από την R ως:

$$P(|t| > 1.136) = 9.43 * 10^{-1}$$

Καθώς p-value << 0.001 τότε η απόρριψη της H_0 είναι στατιστικά σημαντική και επομένως η συνεισφορά της μεταβλητής x στο μοντέλο αναδεικνύεται αξιοσημείωτη. Ειδικότερα, εάν η ηλικία x ενός ασθενή αυξηθεί κατά ένα έτος, τότε τα mg/ml χοληστερόλης στο αίμα του θα αυξηθούν κατά 0.05263 mg/ml. Το 95% διάστημα εμπιστοσύνης για τον συντελεστή παλινδρόμησης του x δίνεται από την εντολή confint() για n=24 δείγματα στο μοντέλο και προκύπτει ως:

$$\left[\hat{\beta}_1 - t_{n-2,a/2} S\sqrt{\frac{1}{\sum_{i=1}^{24}(x_i - \bar{x})}} \ , \ \hat{\beta}_1 + t_{n-2,a/2} S\sqrt{\frac{1}{\sum_{i=1}^{24}(x_i - \bar{x})}}\ \right] = [\, 0.04185806 \, , 0.06339175]$$

Το επόμενο βήμα της ανάλυσης προδιαγράφει τον προσδιορισμό ενός 99% διαστήματος εμπιστοσύνης για την πρόβλεψη της ποσότητας χοληστερόλης στο αίμα ενός ασθενή με ηλικία x = 35 έτη. Με χρήση της εντολής predict(mod1, newdata=list(x=35), interval="prediction", level=.99) όπου mod1 το γραμμικό μοντέλο που έχει οριστεί, η R υπολογίζει το διάστημα εμπιστοσύνης της πρόβλεψης:

$$[\, 2.158578 \ , 4.084902] \ \textit{[mg/ml]}$$

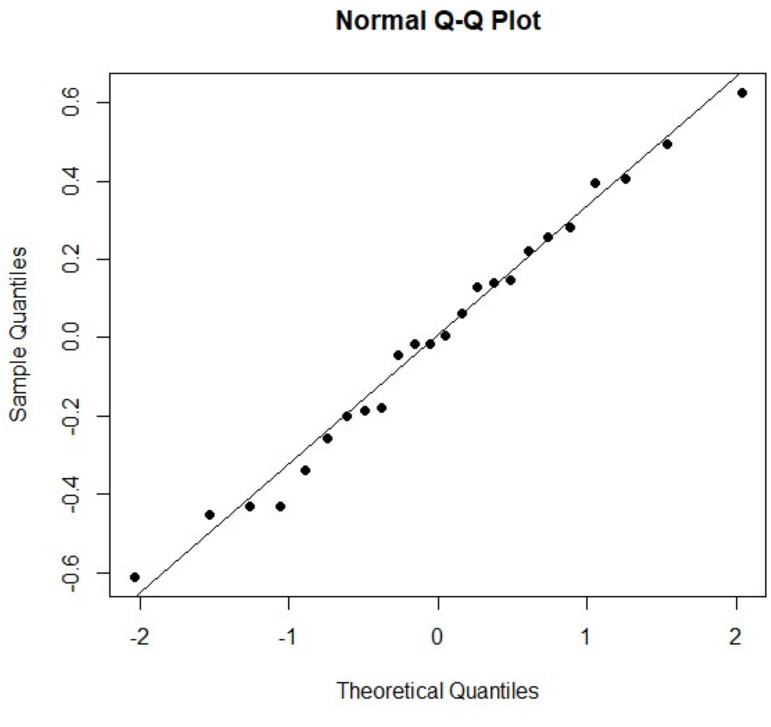
Αντίστοιχα υπολογίζεται το διάστημα εμπιστοσύνης για την μέση τιμή της προαναφερόμενης πρόβλεψης μέσω της εντολής predict(mod1, newdata=list(x=35), interval="confidence", level=.99) . Η R προσδιορίζει το διάστημα εμπιστοσύνης ως:

$$[\, 2.918965 \ , 3.324515] \ \textit{[mg/ml]}$$

Επιπροσθέτως ελέγχεται εάν τα υπόλοιπα που παράγονται μεταξύ των εκτιμήσεων και των πραγματικών τιμών για τα επίπεδα χοληστερόλης y, ακολουθούν την Κανονική Κατανομή. Η παρατήρηση αυτού του ελέγχου γίνεται οπτικά μέσω του Q-Q plot της Εικόνας 2, το οποίο παράγεται από το μοντέλο mod1 σύμφωνα με την στήλη residuals (res), μέσω της σειράς εντολών:

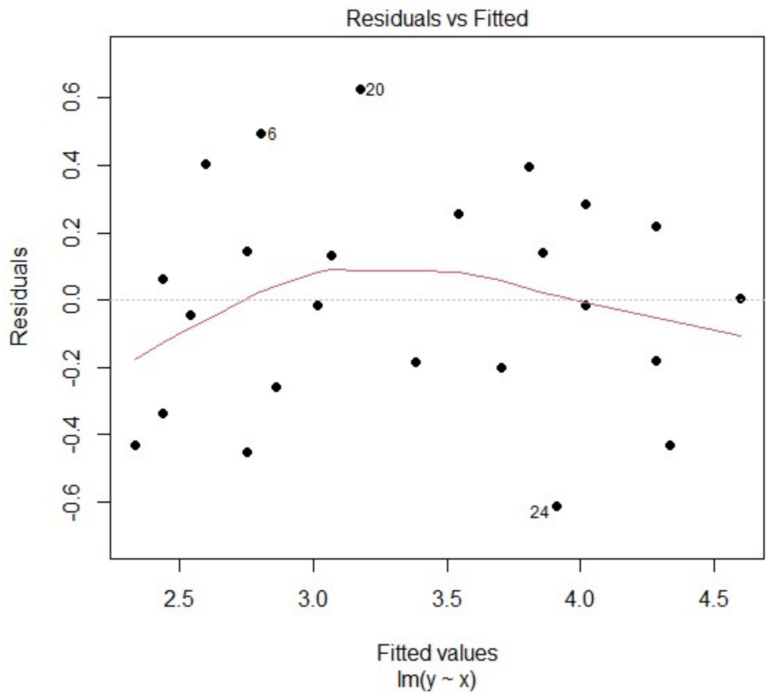
```
qqnorm(mod1$res,pch=19)
qqline(mod1$res)
```

Ειδικότερα, ένα Q-Q plot αναπαριστά ποσοστιαία σημεία κατανομών. Ο άξονας x αναφέρεται στα ποσοστιαία σημεία μιας γνωστής κατανομής ενώ στον άξονα y απεικονίζονται τα αντίστοιχα σημεία για τις πειραματικές τιμές. Εάν τα σημεία που αντιστοιχούν στις παρατηρούμενες τιμές βρίσκονται επί της ευθείας y = x, τότε τα πειραματικά δεδομένα ακολουθούν την κατανομή που δηλώνουν τα ποσοστιαία σημεία του άξονα x. Στο συγκεκριμένο παράδειγμα όπως προαναφέρθηκε, τα ποσοστιαία σημεία του δείγματος αναφέρονται στα υπόλοιπα και βρίσκονται κατά μήκος της ευθείας y = x. Επομένως, ακολουθούν κανονική κατανομή με μέση τιμή 0, γεγονός το οποίο έρχεται και σε συμφωνία με την υπόθεση κανονικότητας των θεωρητικών σφαλμάτων.



Εικόνα 2: Q-Q plot για τα ποσοστιαία σημεία της Κανονικής Κατανομής σε σχέση με τα υπόλοιπα μεταξύ των εκτιμήσεων του μοντέλου και των μετρούμενων τιμών.

Στο τέλος δημιουργείται ένα γράφημα των προαναφερόμενων υπολοίπων σύμφωνα με τις εκτιμημένες τιμές με στόχο την εφαρμογή διαγνωστικού ελέγχου στο μοντέλο. Τα υπόλοιπα αυτά θα πρέπει να παρουσιάζουν διακύμανση γύρω από το 0, αλλά δεν θα πρέπει να υπάρχει κάποιο μοτίβο που τα χαρακτηρίζει όπως παραδείγματος χάρη μια γραμμική σχέση. Πιο συγκεκριμένα, θα πρέπει να είναι κατανεμημένα τυχαία κατά μήκος του άξονα των εκτιμημένων τιμών. Το αντίστοιχο γράφημα προκύπτει μέσω της εντολής plot(mod1, which=1, pch=19) και παρουσιάζεται στην Εικόνα 3. Πράγματι η κατανομή των υπολοίπων δεν παρουσιάζει κάποιο μοτίβο και συνεπώς επιβεβαιώνεται και η αρχική υπόθεση του απλού γραμμικού μοντέλου σχετικά με την κανονικότητα των θεωρητικών σφαλμάτων.



Εικόνα 3: Residuals σε σχέση με τις εκτιμημένες τιμές y για το μοντέλο.

Συνολικά, οι εντολές που χρησιμοποιήθηκαν για την άσκηση αυτή ήταν:

```
adat1 = read.table("C:/Users/user/Desktop/dsml/statistical modeling/cholesterol.txt",header=TRUE)
attach(adat1)
plot(y~x, pch=19)
abline(lm(y~x))
mod1 = lm(y~x)
#
summary(mod1)
confint(mod1)
#
predict(mod1, newdata=list(x=35), interval="prediction", level=.99)
predict(mod1, newdata=list(x=35), interval="confidence", level=.99)
#
qqnorm(mod1$res,pch=19)
qqline(mod1$res)
plot(mod1, which=1, pch=19)
```

ΑΣΚΗΣΗ Γ

Η συγκεκριμένη άσκηση στοχεύει στην δημιουργία ενός γραμμικού μοντέλου το οποίο έχει κατασκευαστεί με αναγωγή από μη γραμμικό. Πιο συγκεκριμένα, το μοντέλο που θα μελετηθεί περιλαμβάνει εκθετική εξάρτηση της εξαρτημένης μεταβλητής y από την ανεξάρτητη μεταβλητή x. Τα δεδομένα στα οποία θα προσαρμοστεί το απλό γραμμικό μοντέλο παρουσιάζονται στον Πίνακα 1. Επιπλέον, η σχέση που τα χαρακτηρίζει και εν συνέχεια τροποποιείται σε γραμμική είναι:

$$y = 3 - ae^{\beta x} \Leftrightarrow \ln(3-y) = \ln(a) + \beta x \quad \rightarrow \quad y^* = \beta_o + \beta_1 x$$

Επομένως, τα απλό γραμμικό μοντέλο θα προσαρμοστεί στην νέα μεταβλητή $y^* = \ln(3-y)$, η οποία παρουσιάζει γραμμική εξάρτηση με το x. Για τον λόγο αυτό, τα δεδομένα της μεταβλητής y, θα εκχωρηθούν στο διάνυσμα y_s σύμφωνα με την R ως:

y_s = log(3-y)

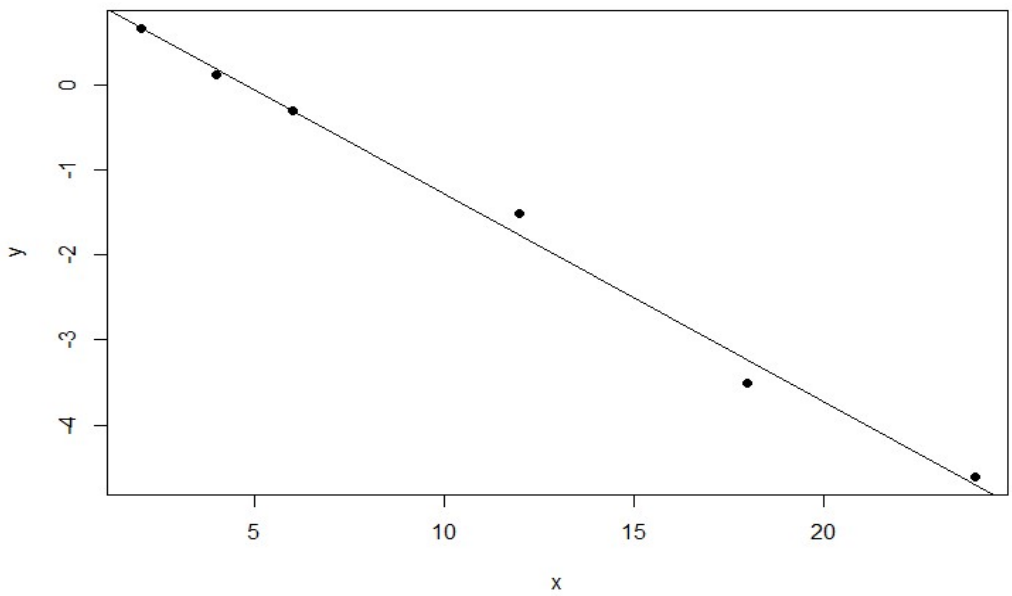
x	2	4	6	12	18	24
y	1.07	1.88	2.26	2.78	2.97	2.99

Πίνακας 1 : Δεδομένα προς επεξεργασία

Το απλό γραμμικό μοντέλο προσαρμόζεται στα μετασχηματισμένα δεδομένα x και y_s μέσω της εντολής mod2 =lm(y_s ~ x). Το διάγραμμα διασποράς των δεδομένων καθώς επίσης και η ευθεία που προκύπτει παρουσιάζονται στην Εικόνα 4. Τα μετασχηματισμένα δεδομένα έχουν αρνητική γραμμική συσχέτιση, καθώς ο εκτιμημένος συντελεστής της κλίσης είναι αρνητικός. Οι εκτιμήσεις για τους συντελεστές παλινδρόμησης είναι:

$$\hat{\beta}_0 = 1.15435 \pm 0.13703$$

$$\hat{\beta}_1 = -0.24367 \pm 0.01012$$



Εικόνα 4: Διάγραμμα διασποράς μετασχηματισμένων δεδομένων y_s ως προς x και αντίστοιχη ευθεία του προσαρμοσμένου γραμμικού μοντέλου

Στη συνέχεια πραγματοποιείται σημειακή πρόβλεψη για x=9 με χρήση της εντολής pred = predict(mod2, newdata=list(x=9), interval="prediction", level=.95). Επιπλέον, προσδιορίζεται ένα 95% διάστημα εμπιστοσύνης για την συγκεκριμένη πρόβλεψη μέσω της εντολής pred_mean = predict(mod2, newdata=list(x=9), interval="confidence", level=.95). Ωστόσο, αυτή η πρόβλεψη αφορά την εκτίμηση της μεταβλητής y_s με αποτέλεσμα, τόσο για την ίδια την πρόβλεψη όσο και για το διάστημα εμπιστοσύνης, θα πρέπει να πραγματοποιηθεί ο μετασχηματισμός προς την μεταβλητή y:

$$\mathrm{y_s} = \log(3 - \mathrm{y}) \quad \Leftrightarrow \quad y = 3 - e^{y_s}$$

Μετά τον μετασχηματισμό, η πρόβλεψη και το αντίστοιχο διάστημα εμπιστοσύνης είναι:

$$\hat{y} = 2.646078 \ \text{ με } \ 95\% \ \text{ διάστημα εμπιστοσύνης } \ [2.361758\,, 2.803741\,]$$

Αντίστοιχη πρόβλεψη πραγματοποιείται και για την μέση τιμή της E(y) μέσω της εντολής pred_mean = predict(mod2, newdata=list(x=9), interval="confidence", level=.95):

$$\widehat{E(y)} = 2.646078\text{με } \ 95\% \ \text{ διάστημα εμπιστοσύνης } \ [2.555063\,, 2.718475\,]$$

Ωστόσο, το αποτέλεσμα είναι προσεγγιστικό καθώς:

$$E(\ln Y) \neq lnE(y)$$

Συνολικά, οι εντολές που χρησιμοποιήθηκαν για την άσκηση αυτή ήταν:

```
x = c(2, 4, 6, 12, 18, 24)
y = log(3-c(1.07,1.88,2.26,2.78,2.97,2.99))
mod2=lm(y~x)
summary(mod2)
plot(y~x, pch=19)
abline(mod2)
#
pred = predict(mod2, newdata=list(x=9), interval="prediction", level=.95)
y_pred = 3-exp(pred[1])
up_pred = 3-exp(pred[2])
lw_pred = 3-exp(pred[3])
#
pred_mean = predict(mod2, newdata=list(x=9), interval="confidence", level=.95)
#prosegistika
mean_y_pred=3-exp(pred_mean[1])
mean_up_pred = 3-exp(pred_mean[2])
mean_lw_pred = 3-exp(pred_mean[3])
```

Άσκηση Α

1) Ο συντελεστής προσδιορισμού R^2 ορίζεται μέσω της σχέσης:

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

και εκφράζει το ποσοστό της μεταβλητότητας της τυχαίας μεταβλητής Y που εξηγείται από την X. Για το απλό γραμμικό μοντέλο ισχύει ότι:

$$\hat{y}_i = \hat{b}_1 x_i + \hat{b}_0 \quad (2)$$

όπου $$\hat{b}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad (3)$$ η εκτίμηση του συντελεστή παλλινδρόμησης για το X

και $$\hat{b}_0 = -\hat{b}_1 \bar{x} + \bar{y} \quad (4)$$

Η σχέση (3) μπορεί να τροποποιηθεί ως:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \Rightarrow \boxed{\hat{b}_1 = \frac{S_{xy}}{S_{xx}} \quad (5)}$$

καθώς ισχύει ότι: (από σχέση (3))

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

$$= S_{xy} \quad (6)$$

και

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2$$

$$= \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} (x_i - \bar{x})^2 = S_{xx} \quad (7)$$

Στην σχέση (1), αντικαθίσταται η εκτίμηση στο

το y_i από το απλό γραμμικό μοντέλο (σχέση (2)):

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

$$\overset{(4)}{=} \frac{\sum_{i=1}^{n} (-\hat{b}_1\bar{x} + \cancel{\bar{y}} + \hat{b}_1 x_i - \cancel{\bar{y}})^2}{S_{yy}}$$

$$= \frac{\sum_{i=1}^{n} (-\hat{b}_1\bar{x} + \hat{b}_1 x_i)^2}{S_{yy}}$$

$$= \frac{\sum_{i=1}^{n} \hat{b}_1^2 (-\bar{x} + x_i)^2}{S_{yy}}$$

⇒

$$\Rightarrow R^2 = \frac{\hat{b}_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2}{S_{yy}}$$

$$\overset{(5)}{=} \frac{\frac{S_{xy}^2}{S_{xx}^2} \cdot S_{xx}}{S_{yy}}$$

$$= \frac{S_{xy}^2}{S_{xx} \cdot S_{yy}} = r_{xy}^2$$

όπου $r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$, ο δειγματικός

συντελεστής συσχέτισης Pearson.

2) Στο ερώτημα αυτό θα χρησιμοποιηθούν οι σχέσεις και οι αριθμήσεις τους, όπως ορίσθηκαν στο προηγούμενο ερώτημα:

$$\sum_{i=1}^{n} y_i \overset{(2)}{=} \sum_{i=1}^{n}(\hat{b}_1 x_i + \hat{b}_0) \overset{(4)}{=} \sum_{i=1}^{n}(\hat{b}_1 x_i - \hat{b}_1 \bar{x} + \bar{y})$$

$$= \sum_{i=1}^{n} \hat{b}_1 (x_i - \bar{x}) + n\bar{y} = \hat{b}_1 \sum_{i=1}^{n} x_i - \hat{b}_1 \bar{x} \sum_{i=1}^{n} 1 + n\bar{y}$$

$$= \cancel{\hat{b}_1 n \bar{x}} - \cancel{\hat{b}_1 \bar{x} n} + \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i$$

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

Θα αναλυθεί ξεχωριστά το κάθε μέλος:

1° $\sum_{i=1}^{n} y_i \hat{y}_i \overset{(2)}{=} \sum_{i=1}^{n} y_i (\hat{b}_1 x_i + \hat{b}_0) \overset{(4)}{=} \sum_{i=1}^{n} y_i (\hat{b}_1 x_i - \hat{b}_1 \bar{x} + \bar{y})$

$= \sum_{i=1}^{n} y_i (\hat{b}_1 (x_i - \bar{x}) + \bar{y}) = \hat{b}_1 \sum_{i=1}^{n} y_i (x_i - \bar{x}) + \bar{y} \sum_{i=1}^{n} y_i$

$\overset{(5)}{=} \hat{b}_1^2 \cdot S_{xx} + n\bar{y}^2$

2° $\sum_{i=1}^{n} \hat{y}_i^2 \overset{(2)}{=} \sum_{i=1}^{n} (\hat{b}_1 x_i + \hat{b}_0)^2 \overset{(5)}{=} \sum_{i=1}^{n} (\hat{b}_1 x_i - \hat{b}_1 \bar{x} + \bar{y})^2$

$= \sum_{i=1}^{n} (\hat{b}_1 (x_i - \bar{x}) + \bar{y})^2$

$= \hat{b}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} 2\bar{y}\hat{b}_1 (x_i - \bar{x}) + \bar{y}^2 \sum_{i=1}^{n} 1$

$= \hat{b}_1^2 \cdot S_{xx} + \cancel{2\bar{y}\hat{b}_1 \cdot n\bar{x} - 2\bar{y}\hat{b}_1 \bar{x} n} + n\bar{y}^2$

$= \hat{b}_1^2 \cdot S_{xx} + n\bar{y}^2$

Το 1° προκύπτει ίσο με το 2° άρα αποδεικνύεται ότι $\sum_{i=1}^{n} y_i \hat{y}_i = \sum_{i=1}^{n} \hat{y}_i^2$

5) $$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) =$$ (3)

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

$$\sum_{i=1}^{n} y_i\hat{y}_i - \bar{y}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{y}_i^2 + \bar{y}\sum_{i=1}^{n}\hat{y}_i \overset{*}{=}$$

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

$$-n\bar{y}^2 + n\bar{y}^2 = 0$$

(*) καθώς δείξαμε ότι: $\sum_{i=1}^{n} y_i\hat{y}_i = \sum_{i=1}^{n}\hat{y}_i^2$

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

6) Θέλουμε να δείξουμε ότι $\dfrac{\hat{b}_1}{se(\hat{b}_1)} = \dfrac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$

όπου το τυπικό σφάλμα προκύπτει

ως $\boxed{se(\hat{b}_1) = \sqrt{\hat{V}(\hat{b}_1)} = \dfrac{S}{S_{xx}^{1/2}} \quad (8)}$ όπου

$S^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ η αμερόληπτη εκτιμήτρια για την διασπορά σ^2, η οποία εναλλακτικά μπορεί να γραφτεί ως:

$$\boxed{S^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) \quad (9)}$$

Επίσης από την σχέση (5) ισχύει ότι $\boxed{\hat{b}_1 = \frac{S_{xy}}{S_{xx}}}$

Αντικαθιστούμε :

$$\boxed{\frac{\hat{b}_1}{se(\hat{b}_1)}} \overset{(5)}{\underset{(8)}{=}} \frac{\frac{S_{xy}}{S_{xx}}}{\frac{S}{\sqrt{S_{xx}}}} = \frac{S_{xy}\sqrt{S_{xx}}}{S \cdot S_{xx}} \overset{(8)}{=}$$

$$\frac{S_{xy}\sqrt{S_{xx}}}{\frac{1}{\sqrt{n-2}}\sqrt{S_{yy} - \frac{S_{xy}^2}{S_{xx}}} \cdot S_{xx}} =$$

$$\frac{\sqrt{n-2}\sqrt{S_{xx}}\, S_{xy}}{\sqrt{S_{yy}}\sqrt{1 - \frac{S_{xy}^2}{S_{xx} S_{yy}}} \cdot S_{xx}} =$$

$$\sqrt{n-2}\,\frac{S_{xy}}{\sqrt{S_{yy} \cdot S_{xx}}}\,\frac{1}{\sqrt{1 - \frac{S_{xy}^2}{S_{xx} \cdot S_{yy}}}} \overset{*}{=} \boxed{\frac{\sqrt{n-2}\, r_{xy}}{\sqrt{1 - r_{xy}^2}}}$$

* όπου $\boxed{r_{xy} = \frac{S_{xy}}{\sqrt{S_{yy} \cdot S_{xx}}}}$