# ΠΟΛΛΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ: ΑΣΚΗΣΕΙΣ

1. Σε μελέτη της επίδρασης γεωργικών χημικών στην προσρόφηση ιζημάτων και εδάφους, δίνονται στον πιο κάτω πίνακα 13 δεδομένα για το δείκτη προσρόφησης φωσφορικού άλατος (Y), για το εξαγώγιμο σίδηρο ($X_1$) και το εξαγώγιμο αργίλιο ($X_2$).

| A/A | Σίδηρο ($X_1$) | Αργίλιο ($X_2$) | δεικτης προσρόφησης |
|---|---|---|---|
| 1 | 61 | 13 | 4 |
| 2 | 175 | 21 | 18 |
| 3 | 111 | 24 | 14 |
| 4 | 124 | 23 | 18 |
| 5 | 130 | 64 | 26 |
| 6 | 173 | 38 | 26 |
| 7 | 169 | 33 | 21 |
| 8 | 169 | 61 | 30 |
| 9 | 160 | 39 | 28 |
| 10 | 244 | 71 | 36 |
| 11 | 257 | 112 | 65 |
| 12 | 333 | 88 | 62 |
| 13 | 199 | 54 | 40 |

a) Να εφαρμόσετε το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
   i. Να εκτιμήσετε τα $\beta_i$ σημειακά και με δ.ε. συντελεστού 95%. Ποια η ερμηνεία τους;
   ii. Πως ερμηνεύετε το παρόν μοντέλο;
   iii. Να κάνετε τον έλεγχο $H_0$: $\beta_1 = \beta_2 = 0$ του μοντέλου μέσω του *F*-τεστ. Ποια είναι η εκτίμηση της διασποράς των σφαλμάτων;
   iv. Τι ποσοστό της μεταβλητότητας των $Y_i$ ερμηνεύεται από το μοντέλο;
   v. Να γίνει έλεγχος ορθότητας του μοντέλου:
      a. Εξετάστε αν τα τυποποιημένα κατάλοιπα προέρχονται πράγματι από κανονική κατανομή (ιστόγραμμα, Q-Q plot και K-S τεστ).
      b. Εξετάστε αν υπάρχει σχέση μεταξύ των τυποποιημένων υπόλοιπων και των μεταβλητών Y, $X_1$, $X_2$.

Απάντηση:

**Εισάγω τα δεδομένα και σχηματίζω το μοντέλο στην R:**

```
>iron<-c(61, 175, 111, 124, 130, 173, 169, 169, 160, 244, 257, 333, 199)
>argile<-c(13, 21, 24, 23, 64, 38,  33, 61, 39, 71, 112, 88,  54)
>ind<-c(4 , 18, 14, 18, 26, 26, 21, 30, 28, 36, 65, 62, 40)

>my.data<-data.frame(ind, iron, argile)
>lm.1<-lm(ind~iron + argile, data=my.data)
>summary.lm(lm.1)
>confint(lm.1)
```

**Αποτέλεσμα:**

```
Call:
lm(formula = ind ~ iron + argile, data = my.data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9352 -2.2182  0.4613  3.3448  6.0708

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.35066    3.48467  -2.109 0.061101 .
iron         0.11273    0.02969   3.797 0.003504 **
argile       0.34900    0.07131   4.894 0.000628 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.379 on 10 degrees of freedom
Multiple R-squared:  0.9485,   Adjusted R-squared:  0.9382
F-statistic: 92.03 on 2 and 10 DF, p-value: 3.634e-07
```

Εκτιμητής $\beta_1$

Εκτιμητής $\beta_2$

P-value του $\beta_1$: 0.003504< 0.01 (=1/100) άρα η $H_0$: $\beta_1$=0 απορρίπτεται

P-value του $\beta_2$: 0.000628 < 0.01 (=1/100) άρα η $H_0$: $\beta_2$=0 απορρίπτεται

**Διαστήματα εμπιστοσύνης των $\beta_i$**

```
                  2.5 %    97.5 %
(Intercept) -15.11498341 0.4136638
iron          0.04657763 0.1788882
argile        0.19011968 0.5078809
```

Δ/τα εμπιστοσύνης για τα $\beta_1$ και $\beta_2$. Παρατηρούμε ότι κανένα δ/μα δεν περιέχει το 0.

**Απάντηση στο ερώτημα (i):**

Εκτιμητής του $\beta_1$: 0.11273 (σημειακός) με 95% δ/μα εμπιστοσύνης: [0.04657763,0.1788882]. Ερμηνεία: Η αύξηση του σιδήρου κατά μία μονάδα επιφέρει αύξηση στο δείκτη προσρόφησης κατά 0.112, δεδομένου ότι η ποσότητα αργιλίου παραμένει σταθερή.

Εκτιμητής του $\beta_2$: 0.34900 (σημειακός) με 95% δ/μα εμπιστοσύνης: [0.19011968, 0.5078809]. Ερμηνεία: Η αύξηση του αργιλίου κατά μία μονάδα επιφέρει αύξηση στο δείκτη προσρόφησης κατά 0.349, δεδομένου ότι η ποσότητα σιδήρου παραμένει σταθερή.

**Απάντηση στο ερώτημα (ii):**

Παρατηρούμε πως και τα 2 P-value (του $\beta_1$ και του $\beta_2$) συνιστούν απόρριψη των $H_0$: $\beta_1$=0 και $H_0$: $\beta_2$=0. Επιπλέον τα αντίστοιχα διαστήματα εμπιστοσύνης δεν περιέχουν το 0. Ακόμα περισσότερο, το F test για έλεγχο της $H_0$: $\beta_1$=$\beta_2$=0 έχει p-value: 3.634e-07 που συνιστά απόρριψη της υπόθεσης ότι οι 2 μεταβλητές δεν επηρεάζουν ταυτόχρονα το μοντέλο. Για όλους αυτούς του λόγους, η ερμηνεία της ανάλυσης του μοντέλου είναι ότι συμπεραίνουμε ότι το εξαγώγιμο σίδηρο και αργίλιο επηρεάζουν σημαντικά το δείκτη προσρόφησης και καλώς συμπεριλαμβάνονται στο μοντέλο.

**Απάντηση στο ερώτημα (iii):**

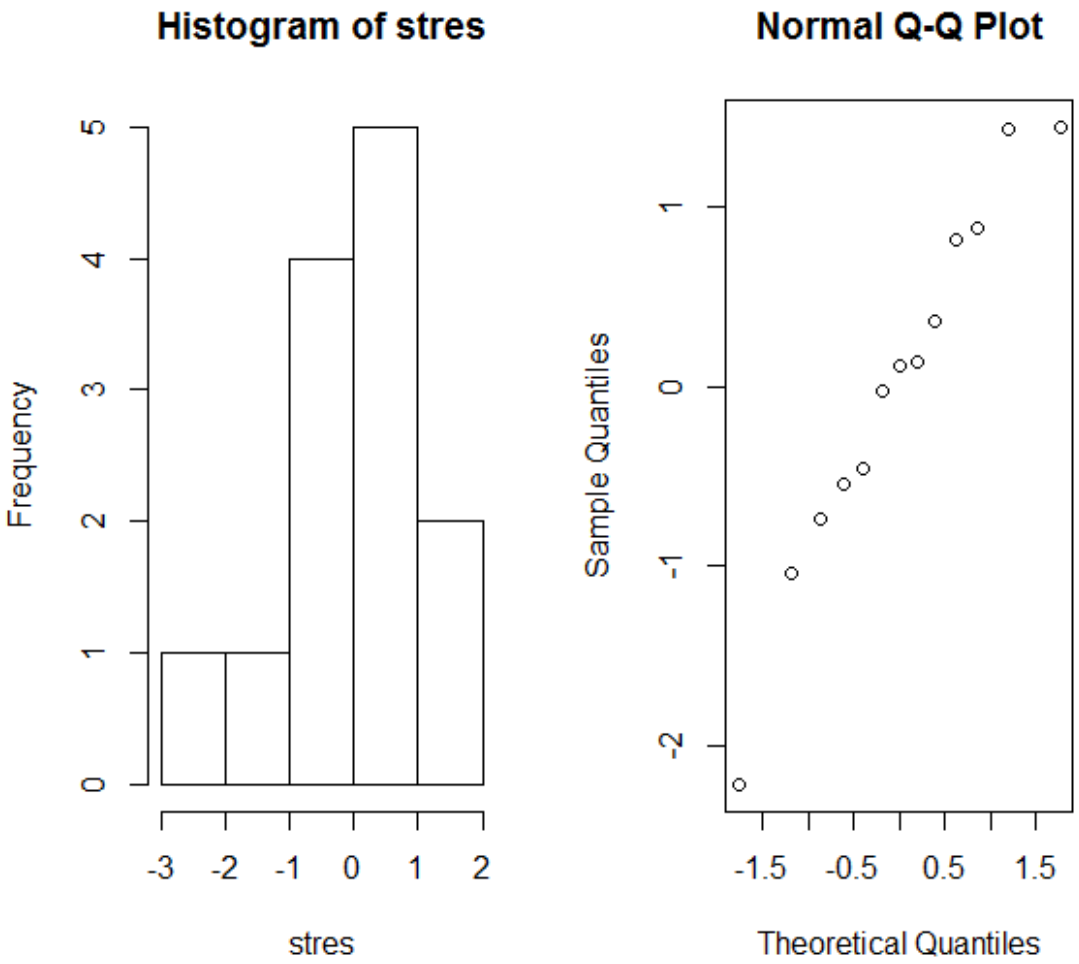
Όπως είπαμε στο ερώτημα (ii), το F test για έλεγχο της $H_0$: $\beta_1$=$\beta_2$=0 έχει p-value: 3.634e-07 που συνιστά απόρριψη της υπόθεσης ότι οι 2 μεταβλητές δεν επηρεάζουν ταυτόχρονα το μοντέλο. Η εκτίμηση της διασποράς των σφαλμάτων είναι : “Residual standard error: 4.379” άρα: 4.379 (σε τί μονάδες;)

**Απάντηση στο ερώτημα (iv):**

`Adjusted R-squared:  0.9382`: Αυτές οι 2 μεταβλητές ερμηνεύουν το περίπου 94% της μεταβλητότητας του δείκτη προσρόφησης.

**Απάντηση στο ερώτημα (v (α)):**
**Για να πάρουμε το ιστόγραμμα και το qq plot δίνουμε τις εντολές:**
>res<- residuals.lm(lm.1) #απομονώνουμε τα υπόλοιπα σε μια μεταβλητή
>stres<-rstandard(lm.1) #κανονικοποιούμε τα υπόλοιπα
>par(mfrow=c(1,2)) # λέμε στην R να χωρέσει 2 γραφήματα μαζί
>hist(stres) # ιστόγραμμα
>qqnorm(stres) # qqplot

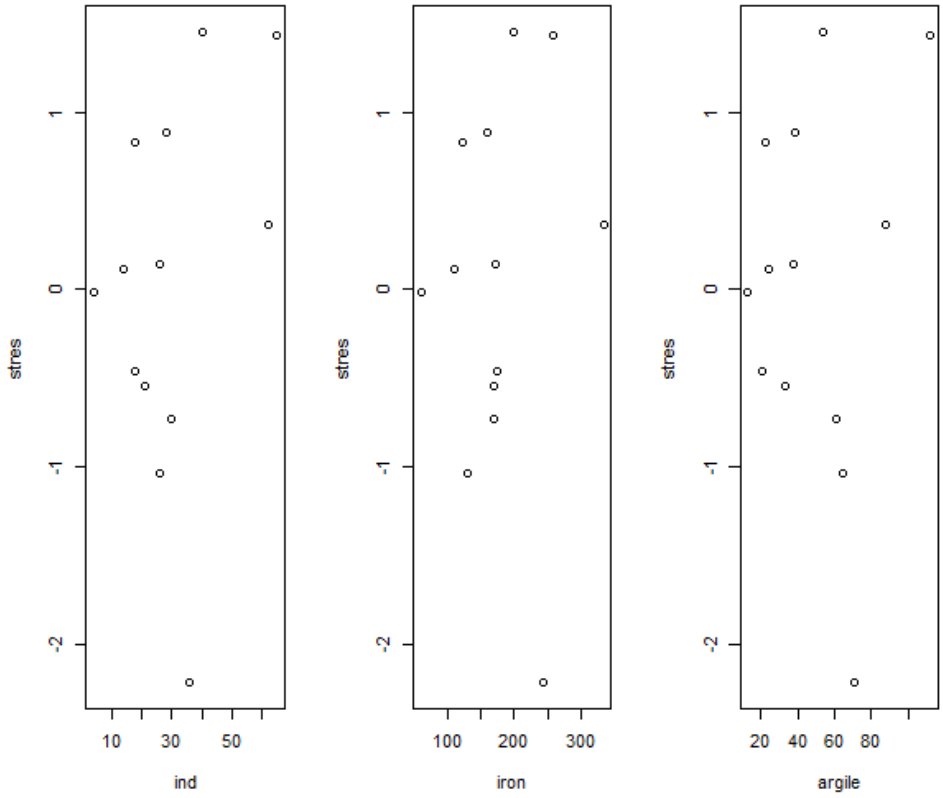
Το αποτέλεσμα που παίρνουμε είναι:




**Ερμηνεία των γραφημάτων:** Το ιστόγραμμα γίνεται δεκτό σαν προσέγγιση της κανονικής κατανομής γιατί έχουμε μόλις 13 παρατηρήσεις οπότε α) ξέρουμε ότι δεν πρόκειται να πάρουμε ακριβώς το ζητούμενο σχήμα β) δεν απέχει πολύ από το σχήμα της κανονικής.
Το qqplot φαίνεται να προσαρμόζει τα σημεία πάνω σε ευθεία (όχι τέλεια) – αλλά δεν φαίνεται και μεγάλη απόκλιση. Οπότε ο γραφικός έλεγχος συνιστά κανονικότητα των υπολοίπων.

**Απάντηση στο ερώτημα (v (β)):**
**Για να πάρουμε τα 3 γραφήματα δίνουμε τις εντολές:**

>par(mfrow=c(1,3)) # λέμε στην R να χωρέσει 3 γραφήματα δίπλα-δίπλα
>plot(ind, stres)
>plot(iron, stres)
>plot(argile, stres)




**Ερμηνεία των γραφημάτων:** Σε σχέση με την Y (πρώτο γράφημα από αριστερά) παρατηρούμε ότι η κατανομή των υπόλοιπων είναι τυχαία, χωρίς να σχηματίζεται κάποια πατέντα ή να αυξάνεται η διακύμανση. Επίσης παρατηρούμε ότι οι τιμές των υπολοίπων (άξονας Y) είναι μέσα στο 95% των ορίων των τιμών της κανονικής κατανομής.

2. Σε κοιλάδες τρίτης τάξης μετρήθηκαν: α) ο αριθμός των ρυακιών πρώτης τάξης (Y) β) η πυκνότητα αποστράγγισης[1] ($X_1$) γ) το εμβαδόν κάθε κοιλάδας ($X_2$), δ) η υψομετρική διαφορά του υψηλότερου και του χαμηλότερου σημείου της λεκάνης κάθε κοιλάδας ($X_3$) και ε) το σχήμα[2] κάθε κοιλάδας ($X_4$). Τα αποτελέσματα των μετρήσεων φαίνονται στον πίνακα που ακολουθεί.

[1] *Η **πυκνότητα αποστράγγισης** της κοιλάδας ορίζεται ως το πηλίκο του συνολικού μήκους όλων των ρυακιών της κοιλάδας προς το εμβαδόν της κοιλάδας.*
[2] *Ως **σχήμα** της κοιλάδας ορίζεται το πηλίκο του πλάτους προς το μήκος της κοιλάδας.*

| **Κοιλάδα** | **Αριθμός ρυακιών** Y | **Πυκνότητα αποστράγγισης** $X_1$ **(Km/Km²)** | **Εμβαδόν** $X_2$ **(Km²)** | **Υψομετρική διαφορά** $X_3$ **(m)** | **Σχήμα** $X_4$ |
|---|---|---|---|---|---|
| 1 | 25 | 7.16 | 0.968 | 998 | 0.42 |
| 2 | 7 | 8.28 | 0.198 | 562 | 0.53 |
| 3 | 12 | 11.73 | 0.254 | 542 | 0.33 |
| 4 | 59 | 11.47 | 1.018 | 817 | 0.25 |
| 5 | 5 | 14.62 | 0.117 | 635 | 0.17 |
| 6 | 12 | 10.53 | 0.339 | 332 | 0.41 |
| 7 | 6 | 14.76 | 0.126 | 275 | 0.65 |
| 8 | 23 | 10.57 | 0.564 | 786 | 0.73 |
| 9 | 6 | 11.62 | 0.154 | 695 | 0.47 |
| 10 | 7 | 11.28 | 0.218 | 885 | 0.45 |
| 11 | 5 | 7.32 | 0.254 | 690 | 0.71 |
| 12 | 10 | 9.43 | 0.332 | 592 | 0.36 |
| 13 | 9 | 7.76 | 0.595 | 735 | 0.66 |
| 14 | 6 | 7.06 | 0.306 | 548 | 0.42 |
| 15 | 5 | 12.14 | 0.098 | 576 | 0.38 |
| 16 | 9 | 11.76 | 0.272 | 713 | 0.25 |
| 17 | 11 | 12.52 | 0.440 | 805 | 0.31 |
| 18 | 7 | 12.44 | 0.156 | 384 | 0.39 |
| 19 | 17 | 8.46 | 0.766 | 910 | 0.32 |
| 20 | 5 | 9.55 | 0.179 | 507 | 0.42 |

a) Να κατασκευάσετε όλα τα (ανά δύο) γραφήματα διασποράς των μεταβλητών *Y*, $X_1$, $X_2$, $X_3$, $X_4$. Φαίνεται να υπάρχει γραμμική ή άλλη σχέση μεταξύ των μεταβλητών;

b) Να εφαρμόσετε το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$.
   i. Να εκτιμήσετε τα $\beta_i$ σημειακά και με δ.ε. συντελεστού 95%.
   ii. Ποιες ανεξάρτητες μεταβλητές φαίνεται να επηρεάζουν τον αριθμό ρυακιών;
   iii. Να κάνετε τον έλεγχο $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ του μοντέλου μέσω του *F*-τεστ. Ποια είναι η εκτίμηση της διασποράς των σφαλμάτων;
   iv. Τι ποσοστό της μεταβλητότητας των $Y_i$ ερμηνεύεται από το μοντέλο;
   v. Να γίνει έλεγχος ορθότητας του μοντέλου:
      a. Εξετάστε αν τα τυποποιημένα κατάλοιπα προέρχονται πράγματι από κανονική κατανομή (ιστόγραμμα, Q-Q plot και K-S τεστ).
      b. Εξετάστε αν υπάρχει σχέση μεταξύ των τυποποιημένων υπόλοιπων και των μεταβλητών *Y*, $X_1$, $X_2$, $X_3$, $X_4$.

c) Να εφαρμόσετε το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
   vi. Να εκτιμήσετε τα $\beta_i$ σημειακά και με δ.ε. συντελεστού 95%. Ποια η ερμηνεία τους;
   vii. Πως κρίνετε το παρόν μοντέλο σε σχέση με το προηγούμενο;

viii. Να κάνετε τον έλεγχο $H_0$: $\beta_1 = \beta_2 = 0$ του μοντέλου μέσω του *F*-τεστ. Ποια είναι η εκτίμηση της διασποράς των σφαλμάτων;

ix. Τι ποσοστό της μεταβλητότητας των $Y_i$ ερμηνεύεται από το μοντέλο;

x. Να γίνει έλεγχος ορθότητας του μοντέλου:

a. Εξετάστε αν τα τυποποιημένα κατάλοιπα προέρχονται πράγματι από κανονική κατανομή (ιστόγραμμα, Q-Q plot και K-S τεστ).

b. Εξετάστε αν υπάρχει σχέση μεταξύ των τυποποιημένων υπόλοιπων και των μεταβλητών Y, $X_1$, $X_2$.

3. Στον πίνακα που ακολουθεί φαίνονται οι μετρήσεις της συστολικής πίεσης του αίματος *SBP*, του δείκτη *QUET* = βάρος/υψος$^2$, της ηλικίας *AGE* και του ιστορικού σχετικά με το κάπνισμα *SMK* (*SMK = 0* για μη καπνιστές, *SMK = 1* για καπνιστές ή πρώην καπνιστές) 32 ανδρών ηλικίας άνω των 40 ετών από μια συγκεκριμένη περιοχή.

| Άτομο | SBP (Y) | QUET ($X_1$) | AGE ($X_2$) | SMK ($X_3$) |
|---|---|---|---|---|
| 1 | 135 | 2.876 | 45 | 0 |
| 2 | 122 | 3.251 | 41 | 0 |
| 3 | 130 | 3.1 | 49 | 0 |
| 4 | 148 | 3.768 | 52 | 0 |
| 5 | 146 | 2.979 | 54 | 1 |
| 6 | 129 | 2.79 | 47 | 1 |
| 7 | 162 | 3.668 | 60 | 1 |
| 8 | 160 | 3.612 | 48 | 1 |
| 9 | 144 | 2.368 | 44 | 1 |
| 10 | 180 | 4.637 | 64 | 1 |
| 11 | 166 | 3.877 | 59 | 1 |
| 12 | 138 | 4.032 | 51 | 1 |
| 13 | 152 | 4.116 | 64 | 0 |
| 14 | 138 | 3.673 | 56 | 0 |
| 15 | 140 | 3.562 | 54 | 1 |
| 16 | 134 | 2.998 | 50 | 1 |
| 17 | 145 | 3.36 | 49 | 1 |
| 18 | 142 | 3.024 | 46 | 1 |
| 19 | 135 | 3.171 | 57 | 0 |
| 20 | 142 | 3.401 | 56 | 0 |
| 21 | 150 | 3.628 | 56 | 1 |
| 22 | 144 | 3.751 | 58 | 0 |
| 23 | 137 | 3.296 | 53 | 0 |
| 24 | 132 | 3.21 | 50 | 0 |
| 25 | 149 | 3.301 | 54 | 1 |

| | | | | |
|---|---|---|---|---|
| 26 | 132 | 3.017 | 48 | 1 |
| 27 | 120 | 2.789 | 43 | 0 |
| 28 | 126 | 2.956 | 43 | 1 |
| 29 | 161 | 3.8 | 63 | 0 |
| 30 | 170 | 4.132 | 63 | 1 |
| 31 | 152 | 3.962 | 62 | 0 |
| 32 | 164 | 4.01 | 65 | 0 |

a) Να κατασκευάσετε όλα τα (ανά δύο) γραφήματα διασποράς των μεταβλητών SBP, QUET, AGE, SMK. Φαίνεται να υπάρχει γραμμική ή άλλη σχέση μεταξύ των μεταβλητών;

b) Να εφαρμόσετε το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.
   i. Να εκτιμήσετε τα $\beta_i$ σημειακά και με δ.ε. συντελεστού 95%.
   ii. Ποιες ανεξάρτητες μεταβλητές φαίνεται να επηρεάζουν τον αριθμό ρυακιών;
   iii. Να κάνετε τον έλεγχο $H_0$: $\beta_1 = \beta_2 = \beta_3 = 0$ του μοντέλου μέσω του *F*-τεστ. Ποια είναι η εκτίμηση της διασποράς των σφαλμάτων;
   iv. Τι ποσοστό της μεταβλητότητας των $Y_i$ ερμηνεύεται από το μοντέλο;
   v. Να γίνει έλεγχος ορθότητας του μοντέλου:
      a. Εξετάστε αν τα τυποποιημένα κατάλοιπα προέρχονται πράγματι από κανονική κατανομή (ιστόγραμμα, Q-Q plot και K-S τεστ).
      b. Εξετάστε αν υπάρχει σχέση μεταξύ των τυποποιημένων υπόλοιπων και των μεταβλητών *Y*, $X_1$, $X_2$, $X_3$.

c) Να εφαρμόσετε το μοντέλο $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$.
   vi. Να εκτιμήσετε τα $\beta_i$ σημειακά και με δ.ε. συντελεστού 95%. Ποια η ερμηνεία τους;
   vii. Πως κρίνετε το παρόν μοντέλο σε σχέση με το προηγούμενο;
   viii. Να κάνετε τον έλεγχο $H_0$: $\beta_1 = \beta_2 = 0$ του μοντέλου μέσω του *F*-τεστ. Ποια είναι η εκτίμηση της διασποράς των σφαλμάτων;
   ix. Τι ποσοστό της μεταβλητότητας των $Y_i$ ερμηνεύεται από το μοντέλο;
   x. Να γίνει έλεγχος ορθότητας του μοντέλου:
      a. Εξετάστε αν τα τυποποιημένα κατάλοιπα προέρχονται πράγματι από κανονική κατανομή (ιστόγραμμα, Q-Q plot και K-S τεστ).
      b. Εξετάστε αν υπάρχει σχέση μεταξύ των τυποποιημένων υπόλοιπων και των μεταβλητών Y, $X_2$, $X_3$.

4. Στο αρχείο BGSgirls.csv περιέχονται τα δεδομένα 70 κοριτσιών που γεννήθηκαν στο Berkeley, California, ανάμεσα στον Ιανουάριο του 1928 και τον Ιούνιο του 1929. Τα στοιχεία μετριόντουσαν περιοδικά μέχρι την ηλικία των 18. Οι μεταβλητές είναι: WT2: βάρος (kg) σε ηλικία 2 ετών, HT2: ύψος (cm) σε ηλικία 2 ετών, WT9: βάρος (kg) σε ηλικία 9 ετών, HT9: ύψος (cm) σε ηλικία 9 ετών, LG9: περίμετρος ποδιού (cm) σε ηλικία 9 ετών, ST9: δύναμη(kg) σε ηλικία 9 ετών, WT18: βάρος (kg) σε ηλικία 18 ετών, HT18: ύψος (cm) σε ηλικία 18 ετών, LG18:

περίμετρος ποδιού (cm) σε ηλικία 18 ετών, ST18: δύναμη (kg) σε ηλικία 18 ετών
Soma: σωματότυπος, από 1 (πολύ λεπτή), μέχρι 7, (πολύ παχιά).

a) Να κατασκευάσετε όλα τα (ανά δύο) γραφήματα διασποράς των μεταβλητών του αρχείου. Φαίνεται να υπάρχει γραμμική ή άλλη σχέση μεταξύ των μεταβλητών;
b) Να εφαρμόσετε το μοντέλο $Y = \beta_0 + \beta_1 WT9 + \beta_2 LG9 + \varepsilon$.
   i. Να εκτιμήσετε τα $\beta_i$ σημειακά και με δ.ε. συντελεστού 95%.
   ii. Κατά πόσο φαίνεται να εξηγούν αυτές οι μεταβλητές τη μεταβλητότητα της Y;
   iii. Να κάνετε τον έλεγχο $H_0$: $\beta_1 = \beta_2 = 0$ του μοντέλου μέσω του *F*-τεστ. Ποια είναι η εκτίμηση της διασποράς των σφαλμάτων;
   iv. Να γίνει έλεγχος ορθότητας του μοντέλου:
      a. Εξετάστε αν τα τυποποιημένα κατάλοιπα προέρχονται πράγματι από κανονική κατανομή (ιστόγραμμα, Q-Q plot και K-S τεστ).
      b. Εξετάστε αν υπάρχει σχέση μεταξύ των τυποποιημένων υπόλοιπων και των μεταβλητών *Y*, HT9, WT9.
c) Να εφαρμόσετε το μοντέλο $Y = \beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 HT9 + \beta_4 WT9 + \beta_5 ST9 + \varepsilon$.
   v. Να εκτιμήσετε τα $\beta_i$ σημειακά και με δ.ε. συντελεστού 95%.
   vi. Κατά πόσο φαίνεται να εξηγούν αυτές οι μεταβλητές τη μεταβλητότητα της Y;
   vii. Να κάνετε τον έλεγχο $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ του μοντέλου μέσω του *F*-τεστ. Ποια είναι η εκτίμηση της διασποράς των σφαλμάτων;
   viii. Να γίνει έλεγχος ορθότητας του μοντέλου:
      a. Εξετάστε αν τα τυποποιημένα κατάλοιπα προέρχονται πράγματι από κανονική κατανομή (ιστόγραμμα, Q-Q plot και K-S τεστ).
      b. Εξετάστε αν υπάρχει σχέση μεταξύ των τυποποιημένων υπόλοιπων και των μεταβλητών *Y*, HT2, WT2, HT9, WT9, ST9.