ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΧΡΗΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΠΑΚΕΤΩΝ

Εργασία 1η

Φοιτητής
Παντελεήμων Μαθιουδάκης
ΜΕΣ20022

Διδάσκων
Δημήτριος Αντζουλάκος

Πανεπιστήμιο Πειραιά
ΠΜΣ Εφαρμοσμένης Στατιστικής

Περιεχόμενα

1 Άσκηση 1

Αρχικά εισάγουμε τα δεδομένα σε ένα διάνυσμα d

```
> d <- c(73, 6, 77, 81, 91, 101, 135, 61, 65, 68, 18, 20, 23, 12,
+     14, 18, 23, 26, 26, 27, 2, 3, 3, 40, 41, 41, 6, 10, 11, 12,
+     37, 38, 38, 6, 73, 6, 51)
```

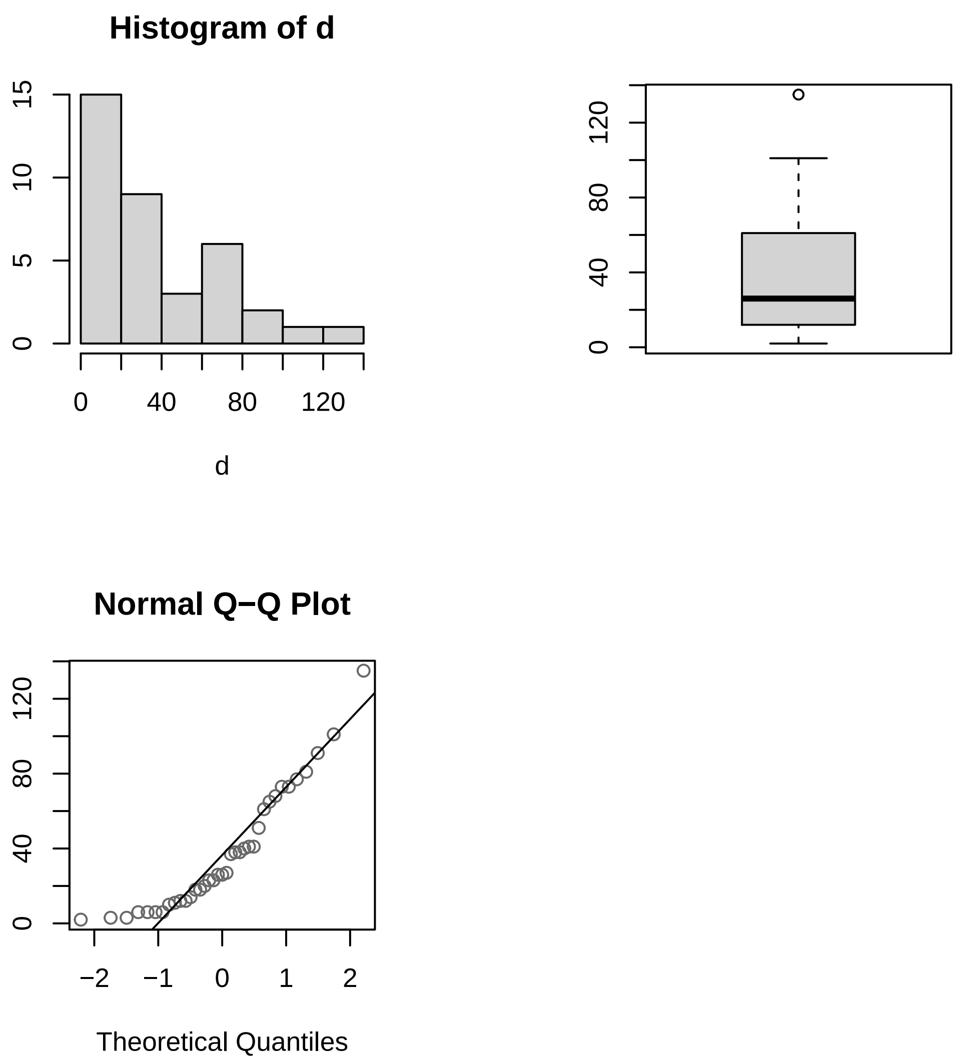
1.1

```
> mesos <- mean(d)
> diamesos <- median(d)
> euros <- max(d) - min(d)
> lo3othta <- mean((d - mean(d))^3)/(sd(d)^3)
> q1 <- quantile(d, 0.25)
> q3 <- quantile(d, 0.75)
> print(c(q1, q3))

25% 75%
 12  61

> max(table(d))

[1] 4

> var(d)

[1] 1044.686

> timh_mode <- table(d)[table(d) == 4]
> timh_mode

6
4
```

μέσος	37.37838
διάμεσος	26
διασπορά	1044.686
εύρος	133
λοξότητα	1.024515
συχνή τιμή	6
1ο τεταρτημόριο	12
3ο τεταρτημόριο	61

1.2

```
> par(mfrow=c(2,2),mai = c(1, 0.35, 0.6, 1.06))
> hist(d)
> boxplot(d)
> qqnorm(d,col='dimgray')
> qqline(d)
```



Διακρίνεται απο το ιστόγραμμα πως υπάρχει μεγάλη πυκνότητα τιμών σε περιοχή που η Normal έχει λεπτή ουρά οπότε δεν φαίνεται να ακολουθούν Normal κατανομή.Αυτό διακρίνεται και στο θηκόγραμμα, καθώς η κατανομή φαίνεται δεξιά ασύμμετρη, θα μπορούσε να φανεί και απο το QQplot κυρίως επειδή διακρίνεται μια 'καμπύλη' των δεδομένων γύρω της γραμμής.

1.3

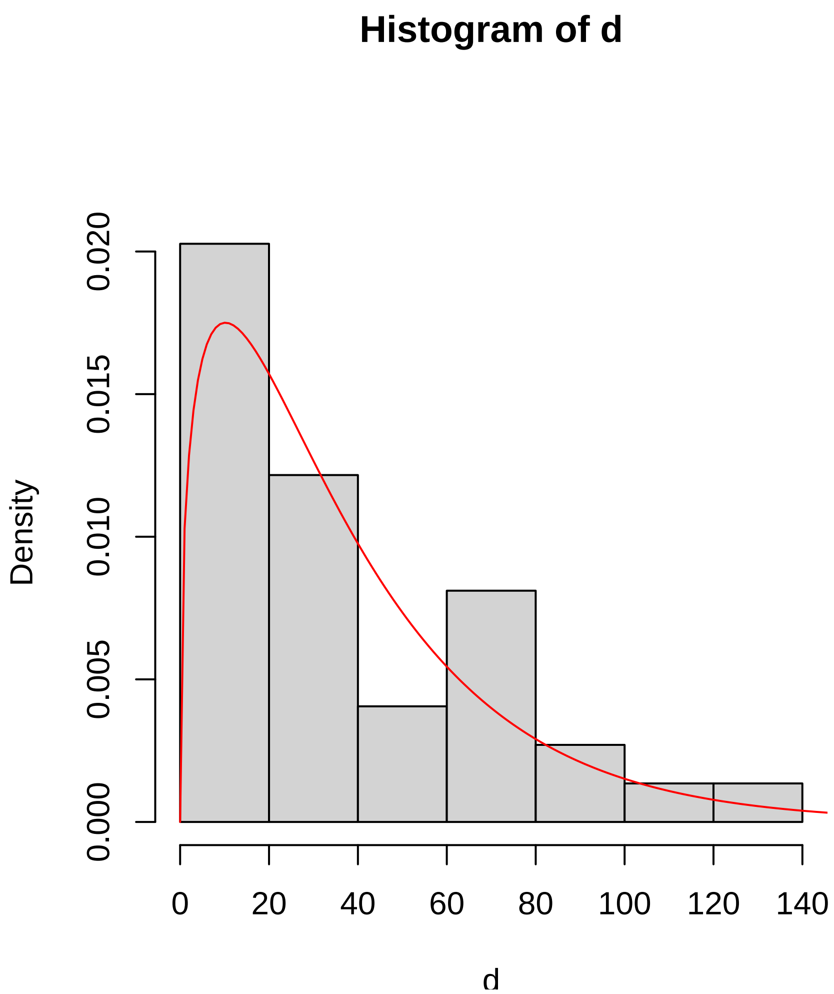
```
> library(nortest)
> lillie.test(d)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  d
D = 0.16647, p-value = 0.01109

> shapiro.test(d)

        Shapiro-Wilk normality test

data:  d
W = 0.88731, p-value = 0.001333

> ks.test(d, "pnorm", mean(d), sd(d))

        One-sample Kolmogorov-Smirnov test

data:  d
D = 0.16647, p-value = 0.2567
alternative hypothesis: two-sided
```

Τα αποτελέσματα των 2 ελέγχων (πέραν του ks.test)συμφωνούν σε $\alpha = 0.05$ οτι απορρίπτεται η υπόθεση κανονικότητας των δεδομένων.

1.4

```
> a = (mean(d)^2)/(mean(d^2) - (mean(d))^2)
> b = (mean(d^2) - mean(d)^2)/mean(d)
> a

[1] 1.37453

> b

[1] 27.19357

> f <- function(x, a, b) {
+     Ga <- prod(1:(a - 1))
+     fun <- (1/((b^a) * Ga)) * (x^(a - 1)) * exp(-(x/b))
+     return(fun)
```

```
+ }
> br <- seq(0, 140, 20)
> par(mfrow = c(1, 1), mai = c(0.75, 1, 2, 1.5))
> hist(d, breaks = br, prob = TRUE)
> lines(seq(0, 150), f(seq(0, 150, 1), a, b), col = "red")
```



$$a = \frac{E(x)^2}{var(x)} \simeq 1.37453$$

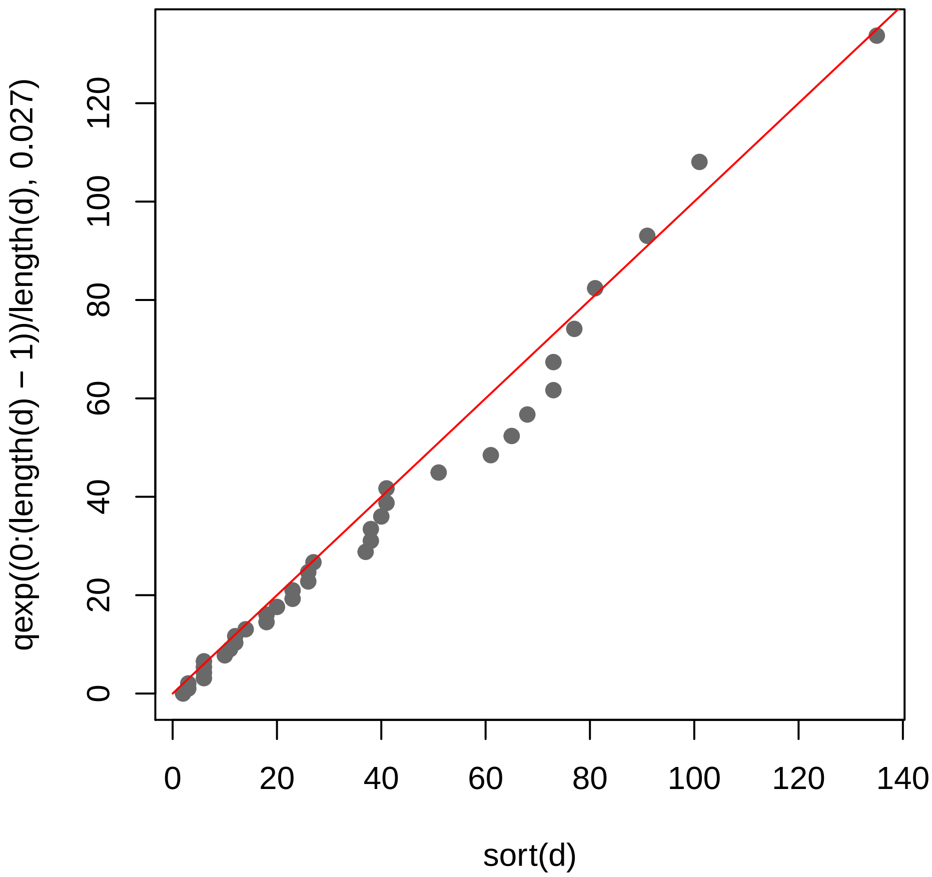
$$b = \frac{var(x)}{E(x)} \simeq 27.19357$$

όπου $var(x) = E(x^2) - E(x)^2$

Ο αλγόριθμος f ← function(x,a,b) ειναι συνάρτηση που υπολογίζει την πυκνότητα της Γάμμα κατονομής (με παραμέτρους εισαγωγής x, a, b)

1.5

```
> par(mfrow = c(1, 1), mai = c(1.5, 1, 0.8, 1.1))
> plot(sort(d), qexp((0:(length(d) - 1))/length(d), 0.027), pch = 19,
+     col = "dimgray")
> lines(c(0, 140), c(0, 140), col = "red")
> ks.test(d, "pexp", 0.027)

        One-sample Kolmogorov-Smirnov test

data:  d
D = 0.091212, p-value = 0.9179
alternative hypothesis: two-sided
```



Φαίνεται αρκετά καλή προσέγγιση και απο το QQplot (αν εξαιρέσουμε κάποιες αποκλίνουσες τιμές στο κέντρο) παραλλήλως του ελέγχου ks-test με ίδιο συμπέρασμα, p-value $= 0.9179 \Rightarrow$ Δεν απορρίπτεται δεδομένα $\sim exp(\lambda = 0.027)$

1.6

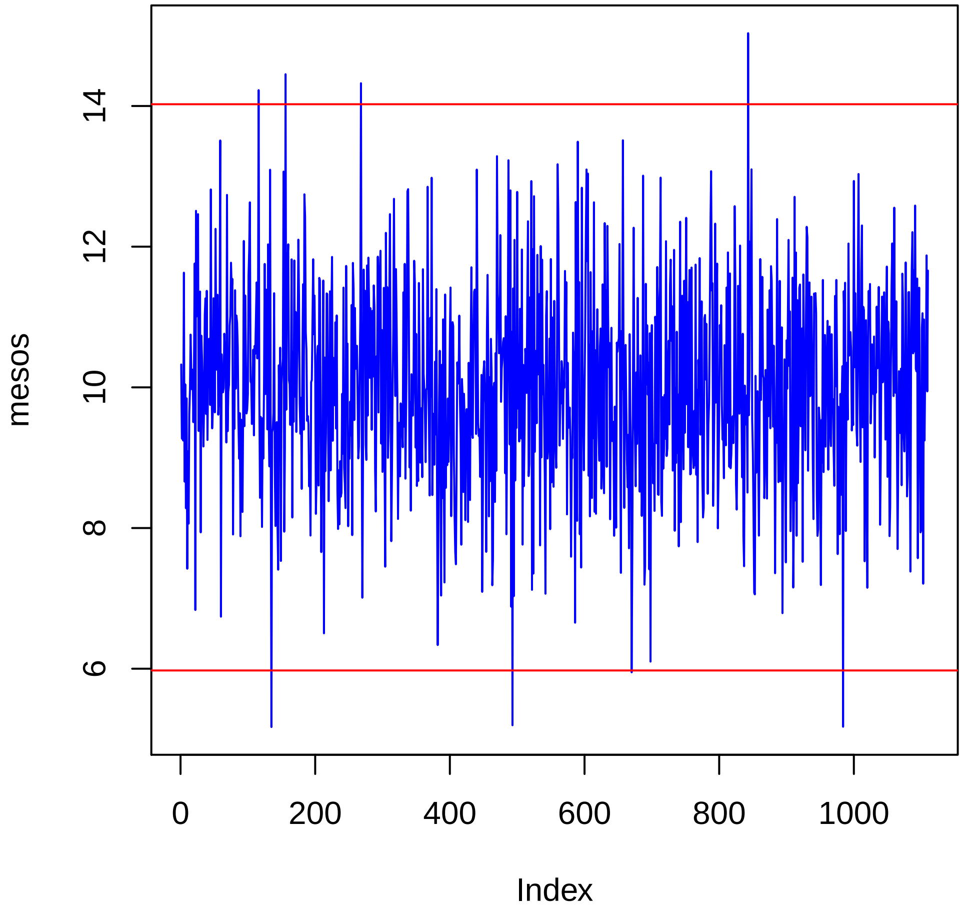
```
> library(PASWR)
> SIGN.test(d, md = 20)

        One-sample Sign-Test

data:  d
```

```
s = 22, p-value = 0.243
alternative hypothesis: true median is not equal to 20
95 percent confidence interval:
 18.00000 40.94273
sample estimates:
median of x
        26

                  Conf.Level L.E.pt  U.E.pt
Lower Achieved CI     0.9011     18 40.0000
Interpolated CI       0.9500     18 40.9427
Upper Achieved CI     0.9530     18 41.0000
```

Εν κατακλείδι, δεν απορρίπτεται η παραμετρική διάμεσος να ισούται με 20

2 Άσκηση 2

```
> set.seed(20022)
> deigma <- matrix(rep(0, 1110 * 5), c(5, 1110))
> mesos <- c()
> for (i in 1:dim(deigma)[2]) {
+     deigma[, i] <- rnorm(5, 10, 3)
+     mesos[i] <- mean(deigma[, i])
+ }
> anwgrammh <- 10 + (3 * 3/sqrt(5))
> katwgrammh <- 10 - (3 * 3/sqrt(5))
> par(mfrow = c(1, 1), mai = c(1.5, 0.8, 0.6, 1))
> plot(mesos, type = "l", col = "blue")
> abline(anwgrammh, 0, col = "red")
> abline(katwgrammh, 0, col = "red")
> thesi <- which((mesos < katwgrammh) | (mesos > anwgrammh))
> outliers <- mesos[thesi]
```




```
false alarms = 8
```

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,1110} \\ \vdots & \vdots & \vdots & \vdots \\ x_{5,1} & x_{5,2} & \cdots & x_{5,1110} \end{bmatrix}_{5\times 1110}$$

$$\begin{array}{cccc} \downarrow & \downarrow & \cdots & \downarrow \\ mean(x_{1:5,1}) & mean(x_{1:5,2}) & \cdots & mean(x_{1:5,1110}) \end{array}$$

Πίνακας 1: Ο αλγόριθμος υπολογίζει 5 τιμές απο $Normal$, 1110 φορές, τα αποθηκεύει σε πίνακα 5x1110 και απο αυτό υπολογίζει το μέσο όρο κάθε στήλης.

3 Άσκηση 3

```
> data <- read.table("/home/user/Downloads/ANAL-DED/Ergasia/babies.txt",
+     header = T)
> j <- 0
> count <- c()
> sthlh <- c(5, 10, 12, 13, 15, 17, 18)
> del <- c(999, 99, 99, 999, 99, 99, 999)
> for (i in 1:dim(data)[1]) {
+     if (sum(data[i, sthlh] == del)) {
+         j <- j + 1
+         count[j] <- i
+     }
+ }
> d_new <- data[-c(count), ]
```

Ο παραπάνω αλγόριθμος ελέγχει κάθε γραμμή του πίνακα αν περιέχει τουλάχιστον 1 τιμή όπως ορίζεται απο το διάνυσμα del
δηλαδή εξαιρούνται (X1,..,X7)=(999,99,99,999,99,99,999) κατα σειρά.
Το αποτέλεσμα, πίνακας d_new με διαστάσεις :

```
> dim(d_new)

[1] 701  23
```

3.1

```
> library(rockchalk)
> attach(d_new)
> model <- lm(wt ~ gestation + age + ht + wt1 + dage + dht + dwt)
> model

Call:
lm(formula = wt ~ gestation + age + ht + wt1 + dage + dht + dwt)

Coefficients:
(Intercept)    gestation          age           ht          wt1         dage
 -101.90746      0.45031      0.13497      1.22304      0.03078      0.06032
        dht          dwt
   -0.07833      0.07831
```

	model	
	Estimate	(S.E.)
(Intercept)	-101.90746***	(23.29181)
gestation	0.45031***	(0.03909)
age	0.13497	(0.18811)
ht	1.22304***	(0.28517)
wt1	0.03078	(0.03427)
dage	0.06032	(0.16551)
dht	-0.07833	(0.27056)
dwt	0.07831*	(0.03307)
N	701	
RMSE	16.42882	
R^2	0.21172	
adj R^2	0.20376	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

εκτιμήσεις των παραμέτρων :

constant	-101.907
gestation	0.45031
age	0.13497
ht	1.22304
wt1	0.03078
dage	0.06032
dht	-0.078
dwt	0.07831

3.2

ποσοστό ερμηνείας :

$$R^2 = 0.2117$$

3.3

```
> library(xtable)
> models <- c('constant','model')
> xtable(as.data.frame(cbind(models,anova(lm(wt~1),model))))
```

	models	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	constant	700.00	237282.01				
2	model	693.00	187044.92	7.00	50237.09	26.59	0.00

$$P\_value(F\_test) < 2.2 \cdot 10^{-16}$$

Συνεπώς απορρίπτω ταυτόχρονα οι παράμετροι να είναι μηδέν ⇒ στατιστικά σημαντικό μοντέλο

3.4

```
> df <- data.frame(gestation = 270, age = 29, ht = 63, wt1 = 220,
+      dage = 36, dht = 74, dwt = 230)
> predict(model, newdata = df, interval = "predict", level = 0.99)

       fit      lwr      upr
1 121.8012 78.34987 165.2526
```

Κάτω φράγμα	Εκτίμηση	Άνω φράγμα
78.349	121.8012	165.252

3.5

```
> model1 <- lm(wt ~ gestation + age + ht)
> model2 <- lm(wt ~ gestation + age + ht + wt1 + dage + dht + dwt)
> anova <- anova(model1, model2)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	697	189284.49				
2	693	187044.92	4	2239.57	2.07	0.0825

Σε α $= 0.1$ $Pr(> F) = 0.0825 < \alpha$ απορρίπτουμε τα 2 μοντέλα να έχουν την ίδια επεξηγηματικότητα, συνεπάγεται το μεγαλύτερης διάστασης μοντέλο είναι καλύτερο

3.6

```
Start:  AIC=3932.2
wt ~ gestation + age + ht + wt1 + dage + dht + dwt

            Df Sum of Sq    RSS    AIC
- dht        1        23 187068 3930.3
- dage       1        36 187081 3930.3
- age        1       139 187184 3930.7
- wt1        1       218 187263 3931.0
<none>                   187045 3932.2
- dwt        1      1513 188558 3935.9
- ht         1      4965 192010 3948.6
- gestation  1     35821 222866 4053.0

Step:  AIC=3930.29
wt ~ gestation + age + ht + wt1 + dage + dwt

            Df Sum of Sq    RSS    AIC
- dage       1        46 187114 3928.5
- age        1       130 187197 3928.8
- wt1        1       224 187291 3929.1
<none>                   187068 3930.3
+ dht        1        23 187045 3932.2
- dwt        1      1785 188853 3934.9
- ht         1      5091 192159 3947.1
- gestation  1     35817 222885 4051.1

Step:  AIC=3928.46
wt ~ gestation + age + ht + wt1 + dwt

            Df Sum of Sq    RSS    AIC
- wt1        1       252 187366 3927.4
<none>                   187114 3928.5
- age        1       877 187991 3929.7
+ dage       1        46 187068 3930.3
+ dht        1        33 187081 3930.3
- dwt        1      1756 188870 3933.0
- ht         1      5051 192165 3945.1
- gestation  1     35816 222930 4049.2

Step:  AIC=3927.41
wt ~ gestation + age + ht + dwt
```

```
            Df Sum of Sq    RSS    AIC
<none>                   187366 3927.4
+ wt1        1       252 187114 3928.5
+ dage       1        74 187291 3929.1
+ dht        1        45 187321 3929.2
- age        1      1140 188506 3929.7
- dwt        1      1919 189284 3932.5
- ht         1      7068 194434 3951.4
- gestation  1     36498 223864 4050.2
```

Απο $AIC = 3932.2 \longmapsto AIC = 3927.41$ με παραμέτρους :

```
> modelaic

Call:
lm(formula = wt ~ gestation + age + ht + dwt)

Coefficients:
(Intercept)    gestation          age           ht          dwt
 -109.19459      0.45319      0.21561      1.30161      0.07562
```

3.7

$$0.45031 \pm t_{\frac{a=0.1}{2},693} \cdot se_{\beta_1}$$

$$\Delta.\text{E}.(\beta_1) = [0.386, 0.514]$$

3.8

```
> vcov(model)
```

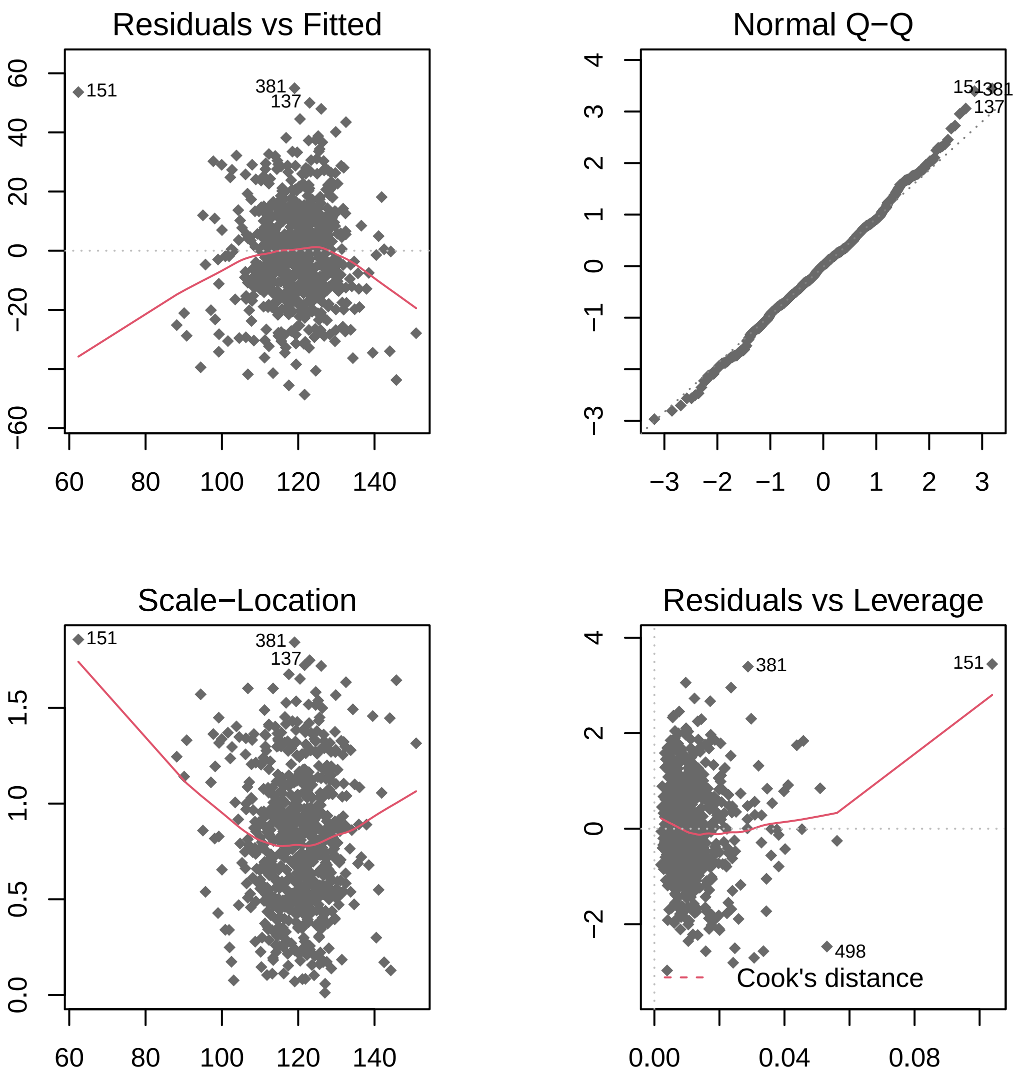
	(Intercept)	gestation	age	ht
(Intercept)	542.50832398	-4.261985e-01	2.107333e-02	-3.435501e+00
gestation	-0.42619854	1.527904e-03	3.324104e-04	7.760986e-05
age	0.02107333	3.324104e-04	3.538604e-02	6.098232e-04
ht	-3.43550145	7.760986e-05	6.098232e-04	8.132185e-02
wt1	0.14462257	-1.026970e-04	-1.875185e-04	-3.856770e-03
dage	-0.58617986	-9.029876e-07	-2.554275e-02	1.000474e-03
dht	-3.32816540	-2.635794e-05	-4.780946e-03	-1.821213e-02
dwt	0.17190445	4.672835e-06	9.003908e-05	-3.953171e-04

	wt1	dage	dht	dwt
(Intercept)	0.1446225654	-5.861799e-01	-3.328165e+00	1.719044e-01
gestation	-0.0001026970	-9.029876e-07	-2.635794e-05	4.672835e-06
age	-0.0001875185	-2.554275e-02	-4.780946e-03	9.003908e-05
ht	-0.0038567698	1.000474e-03	-1.821213e-02	-3.953171e-04
wt1	0.0011742012	-6.198238e-04	3.749124e-04	-1.316390e-04
dage	-0.0006198238	2.739318e-02	6.995701e-03	-1.352113e-04
dht	0.0003749124	6.995701e-03	7.320095e-02	-4.503819e-03
dwt	-0.0001316390	-1.352113e-04	-4.503819e-03	1.093715e-03

Πίνακας 2: Πίνακας Διακύμανσης-Συνδιακύμανσης των β_i παραμέτρων

3.9

```
> par(mfrow=c(2,2),mai=c(0.5,0.3,0.5,0.8))
> plot(model,col='dimgray',pch=18)
```



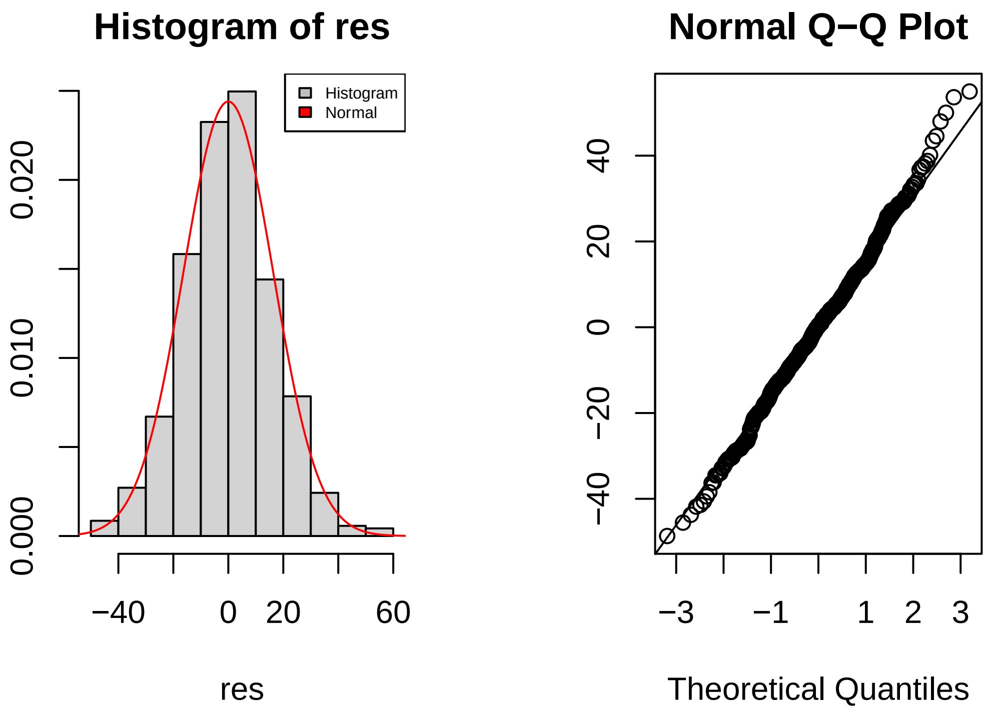
Διακρίνεται πως η παρατήρηση 151 απέχει πολύ απο το κύριο νέφος των δεδομένων (βλέπε Residuals vs Fitted και Residuals vs Leverage)
Συγκεκριμένα πρόκειται για σημείο επιρροής, το οποίο είναι ιδιαίτερα προβληματικό καθώς τέτοιου είδους σημεία (εν αντιθέσει με τα κοινά outliers) επηρεάζουν σημαντικά την εκτίμηση των συντελεστών παλινδρόμησης

Παρατήρηση 151 :

```
> d_new[151,c('gestation','age','ht','wt1','dage','dht','dwt')]

    gestation age ht wt1 dage dht dwt
261       148  28 66 135   36  68 155
```

3.10

```
> res <- resid(model)
> par(mfrow=c(1,2),mai=c(3,0.5,0.5,0.8))
> hist(res,prob=TRUE)
> lines(seq(-70,70),dnorm(seq(-70,70),mean(res),sd(res)),col='red')
> legend("topright", c("Histogram", "Normal"),cex=0.5, fill=c("gray", "red"))
> qqnorm(res)
> qqline(res)
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.99668, p-value = 0.155

> lillie.test(res)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.02832, p-value = 0.1883
```



Τόσο γραφικά (πολύ καλη προσαρμογή ιστογράμματος και ποσοστιαίου διαγράμματος με κανονική κατανομή) όσο και με ελέγχους

shapiro.test	0.155
lillie.test	0.1883

δεν απορρίπτεται κανονικότητα καταλοίπων.