Assignment 2

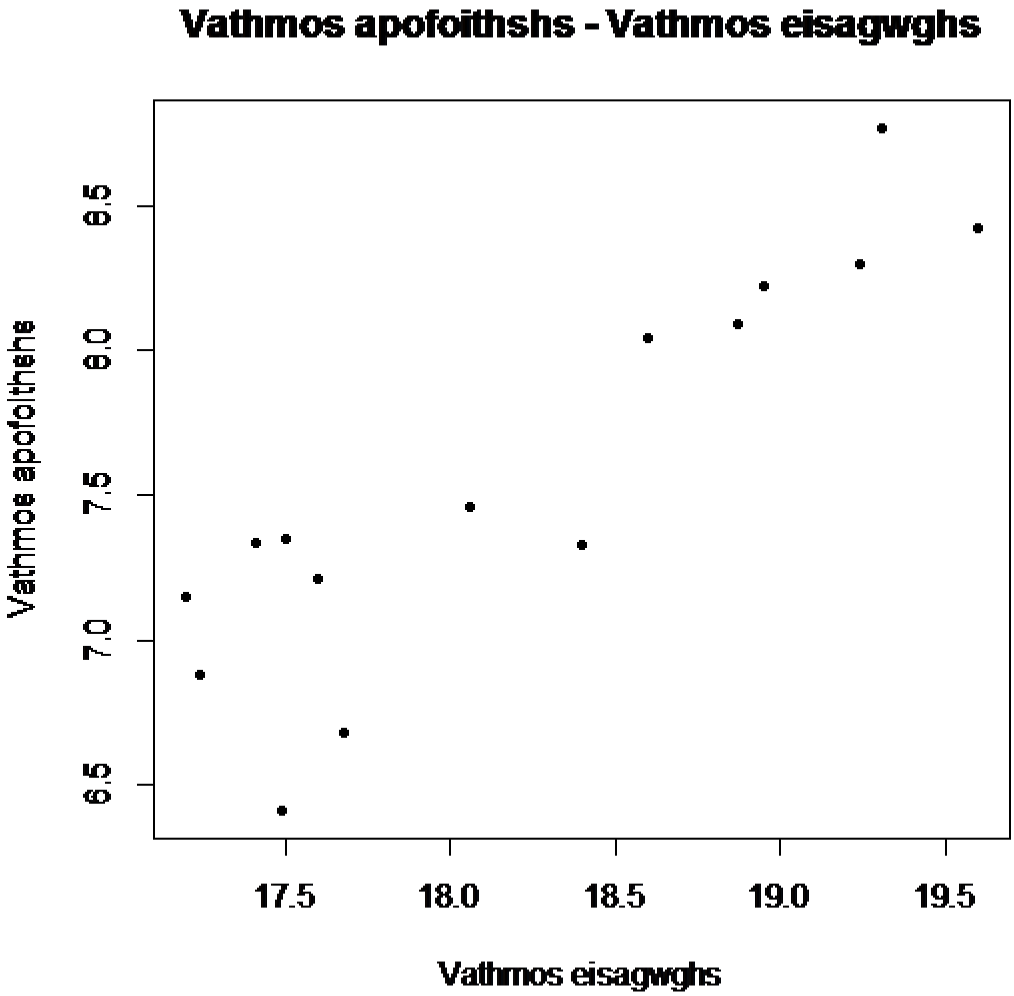
Άσκηση 1

1)

Υποθέτουμε ότι η σχέση μεταξύ του βαθμού εισαγωγής και αποφοίτησης προσεγγίζεται ικανοποιητικά από μία γραμμική συνάρτηση της μορφής $y = ax + b$.

Αρχικά θα σχεδιάσουμε το γράφημα της σχέσης των δύο μεταβλητών:

```
> grade_entry <- c(17.24, 18.06, 17.41, 17.60, 18.95, 19.60, 17.49, 18.60, 17.50, 19.24,
18.87, 17.68, 19.31, 18.40, 17.20)
> grade_grdt <- c(6.88, 7.46, 7.34, 7.21, 8.22, 8.42, 6.41, 8.04, 7.35, 8.30, 8.09, 6.68, 8.77,
7.33, 7.15)
> plot(grade_entry,grade_grdt,main="Vathmos apofoithshs - Vathmos eisagwghs",
xlab="Vathmos eisagwghs", ylab="Vathmos apofoithshs", pch= 20)
```



Στο σημείο αυτό θα πραγματοποιηθεί έλεγχος των συσχετίσεων, ώστε να διαπιστωθεί εάν υπάρχει γραμμική σχέση μεταξύ των δύο μεγεθών.

```
> cor(grade_entry,grade_grdt)
[1] 0.8933966
> cor.test(grade_entry,grade_grdt)
        Pearson's product-moment correlation
data:  grade_entry and grade_grdt
t = 7.1698, df = 13, p-value = 7.263e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7027501 0.9643301
sample estimates:
        cor
0.8933966
```

Ο βαθμός συσχέτισης βρέθηκε $p = 0.8933966$, ενώ η $p - value = 000000.7263 < a = 0.05$.

Άρα, απορρίπτουμε την Null Hypothesis H0, όπου:

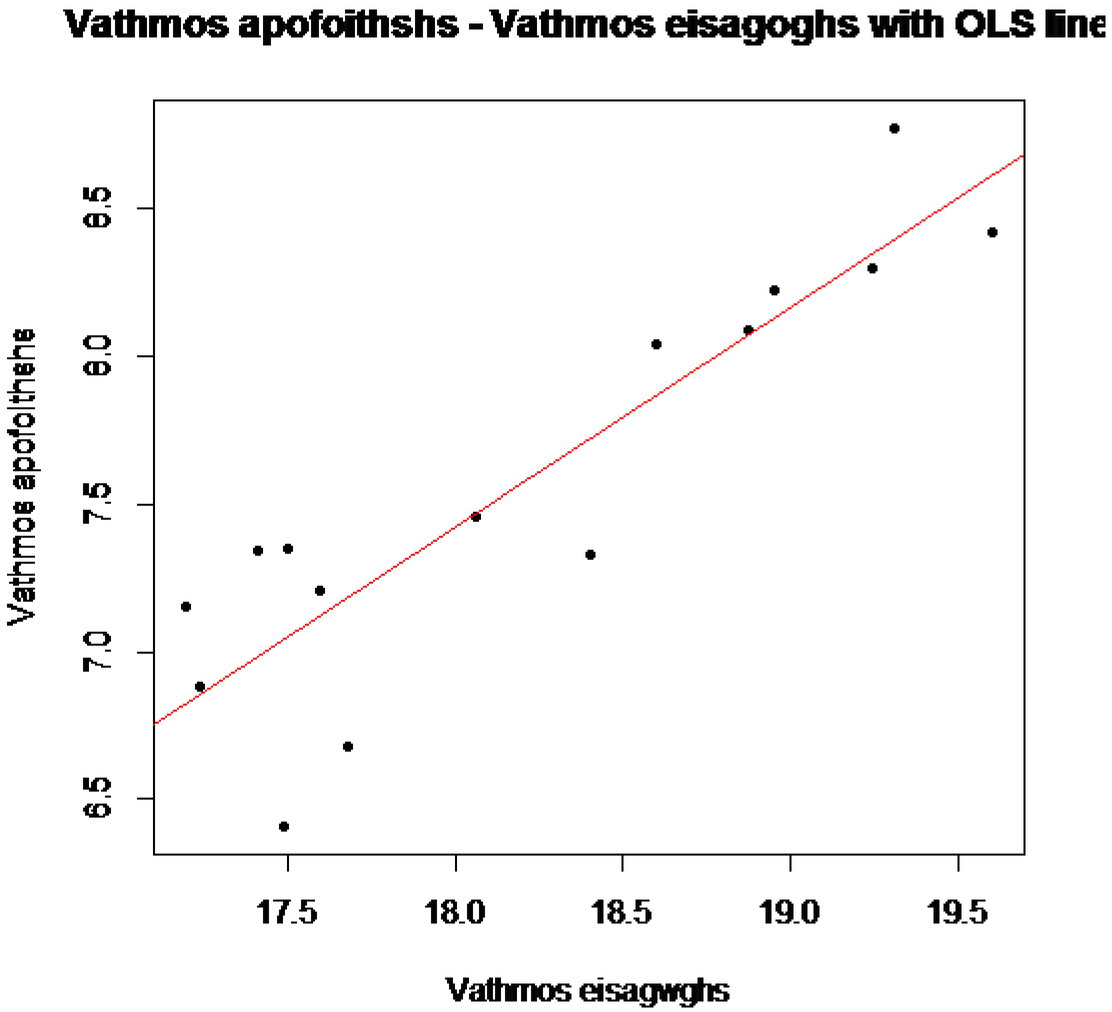
H0: Τα μεγέθη grade_entry και grade_grdt δεν συνδέονται με γραμμική σχέση ($p = 0$).
H1: $p \neq 0$.

Συνεπώς θα εκτιμήσουμε το γραμμικό υπόδειγμα:

```
> fit<-lm(grade_grdt ~ grade_entry)
> summary(fit)
Call:
lm(formula = grade_grdt ~ grade_entry)
Residuals:
     Min       1Q   Median     3Q      Max
-0.63136 -0.11628  0.02451  0.23729  0.37550
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.9621     1.8901  -3.154  0.00761 **
grade_entry   0.7435   0.1037   7.170 7.26e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3225 on 13 degrees of freedom
Multiple R-squared:  0.7982,  Adjusted R-squared:  0.7826
F-statistic: 51.41 on 1 and 13 DF,  p-value: 7.263e-06
```

Λαμβάνουμε το ακόλουθο διάγραμμα ως εξής:

```
> plot(grade_entry, grade_grdt, main="Vathmos apofoithshs - Vathmos eisagoghs with OLS line", xlab="Vathmos eisagwghs", ylab="Vathmos apofoithshs", pch=20)
> abline(fit, col = "red")
```



Πραγματοποιούμε ανάλυση διακύμανσης (ANOVA):

```
> anova(fit)
Analysis of Variance Table
Response: grade_grdt
            Df Sum Sq Mean Sq F value  Pr(>F)
grade_entry  1 5.3471  5.3471  51.407 7.263e-06 ***
Residuals   13 1.3522  0.1040
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Συνεπώς, το απλό γραμμικό μοντέλο περιγράφει ικανοποιητικά τη σχέση μεταξύ των δύο μεταβλητών, καθώς η τιμή του συντελεστή προσδιορισμού είναι μεγάλη, και η κλίση της ευθείας είναι στατιστικά σημαντική. (Επειδή η ανάλυση ANOVA έδωσε ένδειξη “***” δίπλα από το συντελεστή της μεταβλητής).

Το υπόδειγμα της παλινδρόμησης που εκτιμήθηκε είναι:

Βαθμός αποφοίτησης = -5.9621 + 0.7435 * Βαθμός εισαγωγής,

όπου συντελεστής προσδιορισμού = -5.9621 και συντελεστής συσχέτισης = 0.7435.

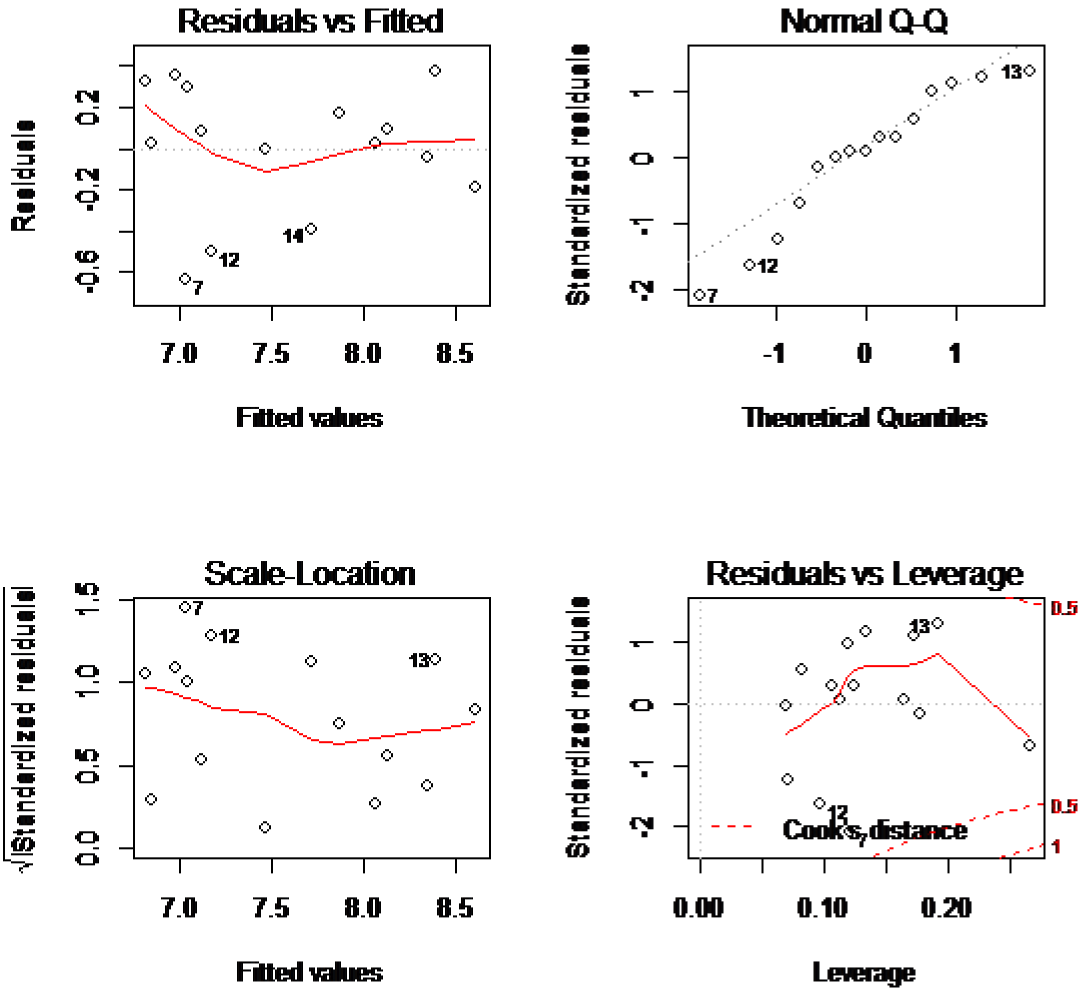
Στο συγκεκριμένο παράδειγμα, η ανάλυση διακύμανσης (ANOVA) είναι ισοδύναμη με τη διενέργεια t-test, καθώς συγκρίνουμε τις μέσες τιμές μεταξύ δύο μεταβλητών (ή ισοδύναμα πραγματοποιούμε ANOVA με k = 2).

2)

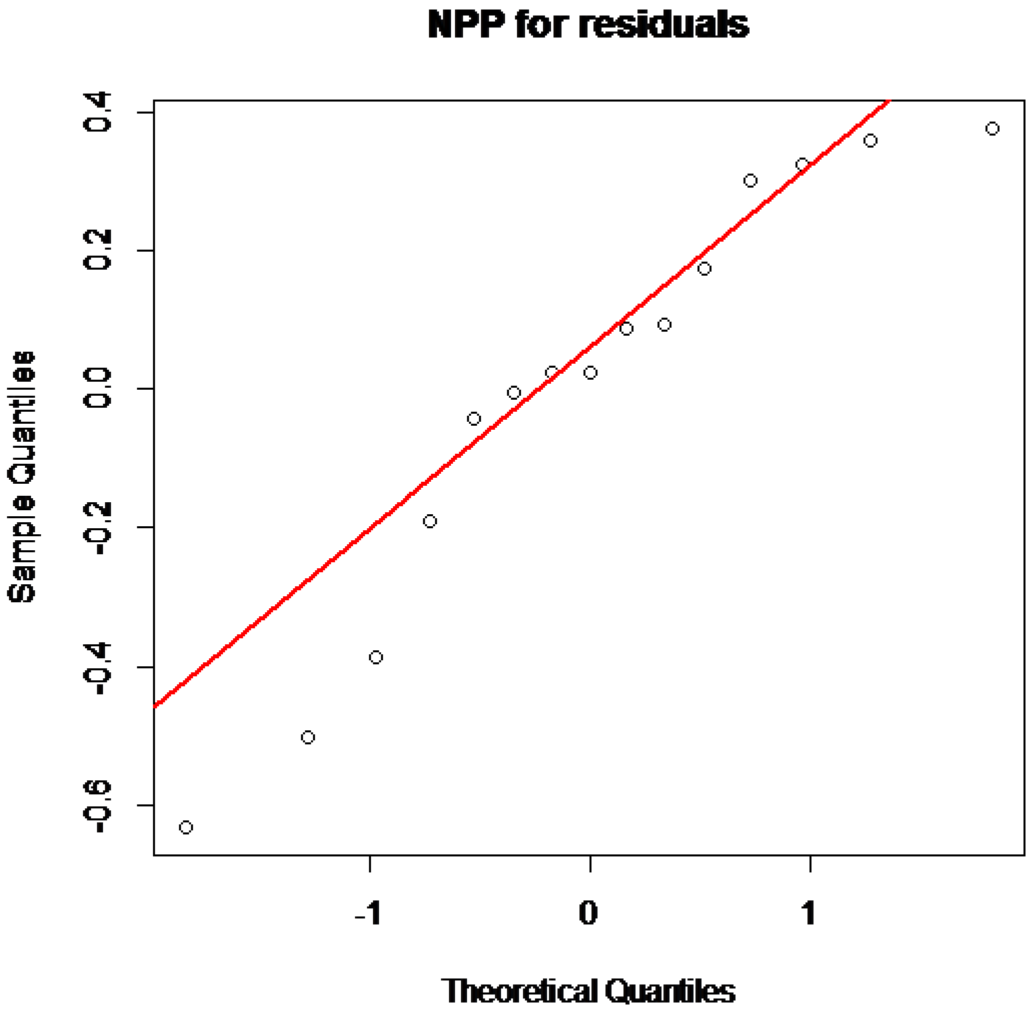
Θα εξεταστεί εάν ικανοποιούνται όλες οι υποθέσεις του μοντέλου (κανονικότητα, ομοσκεδαστικότητα και τυχαιότητα των καταλοίπων) κατασκευάζοντας τα ακόλουθα γραφήματα:
α. Διάγραμμα καταλοίπων/τυποποιημένων καταλοίπων με τις προβλεπόμενες τιμές της εξαρτημένης μεταβλητής
β. Normal probability plot ή QQplot των τυποποιημένων καταλοίπων για να παρατηρήσουμε την κανονικότητα των καταλοίπων.

Κατασκευάζουμε τα ακόλουθα διαγράμματα ώστε να γίνει έλεγχος των υποθέσεων του μοντέλου:

```
> par(mfrow=c(2,2))
> plot(fit)
```




```
> qqnorm(fit$residuals,main="NPP for residuals")
> qqline(fit$residuals,col="red",lty=1,lwd=2)
```



Προς έλεγχο της κανονικότητας, θα πραγματοποιήσουμε ελέγχους Shapiro-Wilk και Lilliefors (Kolmogorov-Smirnov).

```
> shapiro.test(fit$residuals)
        Shapiro-Wilk normality test
data:  fit$residuals
W = 0.91717, p-value = 0.1744

> install.packages("nortest")
> library(nortest)
> lillie.test(fit$residuals)
        Lilliefors (Kolmogorov-Smirnov) normality test
data:  fit$residuals
D = 0.17901, p-value = 0.2195
```

Οι δύο έλεγχοι έδωσαν $p-value = 0.1744$ και $p-value = 0.2195$ αντίστοιχα, τιμές μεγαλύτερες από το a (υποθέτω $a = 0.05$), συνεπώς δεν απορρίπτω την υπόθεση κανονικότητας και με τους δύο ελέγχους.

Επίσης ισχύει ο έλεγχος των καταλοίπων και από τα διαγράμματα ελέγχου.

Ικανοποιούνται όλες οι υποθέσεις του μοντέλου (κανονικότητα, ομοσκεδαστικότητα και τυχαιότητα των καταλοίπων). Συνεπώς δεν υπάρχουν αντενδείξεις της αξιοπιστίας του γραμμικού υποδείγματος.

3)

Για το συντελεστή συσχέτισης:

```
> intercept_p <- 0.7435
> s_d_p <- 0.1037

> min_b_p = intercept_p -( 2.145 * ( (s_d_p)/(length(grade_entry))^0.5))
> min_b_p
[1] 0.6860671
> max_b_p = intercept_p + ( 2.145 * ( (s_d_p)/(length(grade_entry))^0.5))
> max_b_p
[1] 0.8009329
```

Για το συντελεστή προσδιορισμού:

```
> intercept <- -5.9621
> s_d <- 1.8901

> min_b = intercept -( 2.145 * ( (s_d)/(length(grade_entry))^0.5))
> min_b
[1] -7.008907
> max_b = intercept + ( 2.145 * ( (s_d)/(length(grade_entry))^0.5))
> max_b
[1] -4.915293
```

Σημείωση: Η τιμή 2.145 προκύπτει από το t table για $n-1 = 14$.

Συνεπώς ο συντελεστής συσχέτισης βρίσκεται στο διάστημα (0.6860 , 0.8009) και ο συντελεστής προσδιορισμού στο διάστημα (-7.0089 , -4.9152) με συντελεστή εμπιστοσύνης 95%.

Το απλό γραμμικό μοντέλο προβλέπει:

```
> grade <- -5.9621 + 0.7435 * 18.5
> grade
[1] 7.79265
```

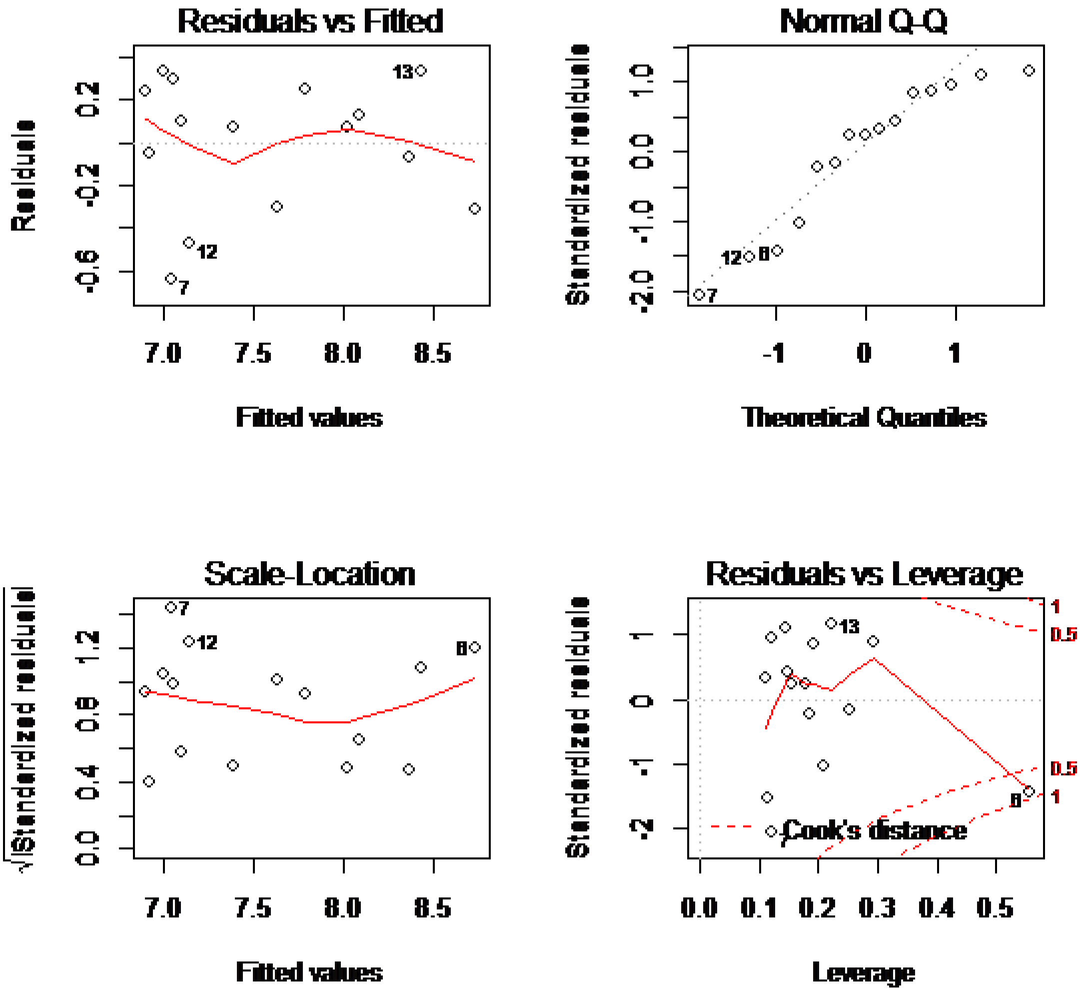
Το ΄95% διάστημα εμπιστοσύνης προκύπτει από την αντικατάσταση των τιμών που βρέθηκαν για το 95% διάστημα εμπιστοσύνης του συντελεστή προσδιορισμού.

```
> grade_min <- -7.0089 + 0.7435 * 18.5
> grade_min
[1] 6.74585
> grade_max
[1] 8.83955
```

Επομένως το 95% διάστημα εμπιστοσύνης είναι το (6.74585, 8.83955).

4)

```
> grade_entry_2 <- grade_entry^2
> fit2 <- lm(grade_grdt ~ grade_entry + grade_entry_2)
> summary(fit2)
Call:
lm(formula = grade_grdt ~ grade_entry + grade_entry_2)
Residuals:
     Min       1Q   Median       3Q      Max
-0.63730 -0.18321  0.07252  0.24683  0.33536
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    38.4211    62.3830   0.616    0.549
grade_entry    -4.1090     6.8180  -0.603    0.558
grade_entry_2   0.1324     0.1860   0.712    0.490
Residual standard error: 0.3288 on 12 degrees of freedom
Multiple R-squared:  0.8063,  Adjusted R-squared:  0.7741
```



Στο σημείο αυτό θέλουμε να συγκρίνουμε τα μοντέλα fit, fit2, τα οποία είναι nested models, καθώς το fit2 πρόκειται για full model σε σχέση με το reduced model fit, καθώς περιέχουν τις ίδιες μεταβλητές, ενώ το fit2 λαμβάνει υπόψιν το τετράγωνο της μεταβλητής αυτής.

Αρχικά παρατηρώ ότι το summary(fit) δίνει Multiple R-squared: 0.7982 και Residual standard error: 0.3225. Αντίστοιχα το summary(fit2) δίνει Multiple R-squared: 0.8063 και Residual standard error: 0.3288, άρα δεν υπάρχει μεγάλη απόκλιση μεταξύ των μοντέλων αυτών για τις συγκεκριμένες παραμέτρους (αμφότερες υποδεικνύουν καλύτερη προσέγγιση καθώς οι αντίστοιχες τιμές “πλησιάζουν” στο μηδέν).

Την οριστική απάντηση θα μας δώσει το Partial F-Test, το οποίο δηλώνει ότι:

H0 Hypothesis: Τα μοντέλα fit και fit2, δεν διαφέρουν ιδιαίτερα.
H1: To full model fit2 είναι σημαντικά καλύτερο από το Reduced model fit.

> anova(fit,fit2)
Analysis of Variance Table
Model 1: grade_grdt ~ grade_entry
Model 2: grade_grdt ~ grade_entry + grade_entry_2

```
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     13 1.3522
2     12 1.2974  1   0.05478 0.5067 0.4902
```

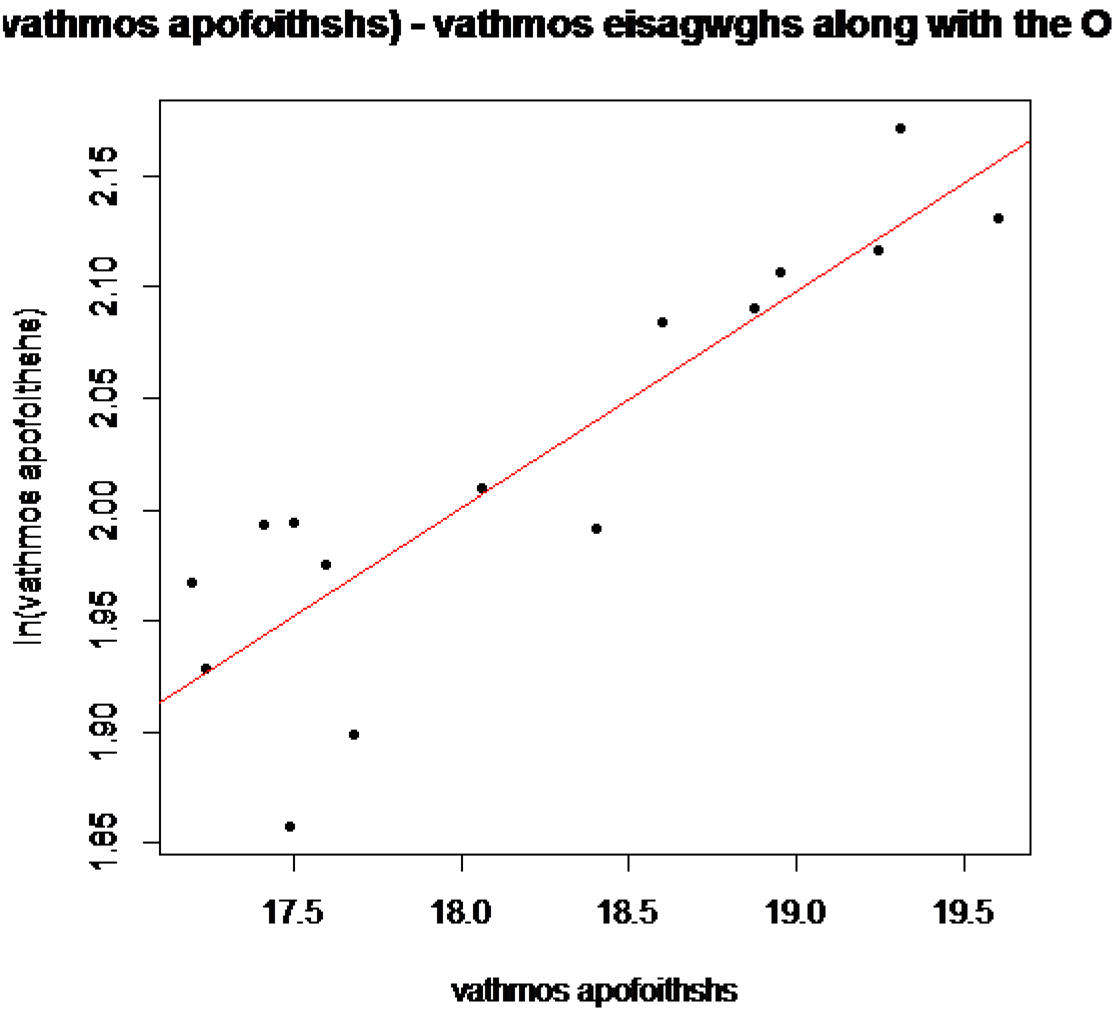
Παρατηρούμε ότι p-value = 0.4902, άρα δεν απορρίπτουμε την Null Hypothesis, και δεν θεωρούμε ότι το full model fit2 βελτιώνει τις προβλέψεις. Συνεπώς, εάν έπρεπε να επιλέξουμε μεταξύ των 2 θα επιλέγαμε το reduced model fit, καθώς είναι πιο απλό. Επιλέγουμε τα full models και γενικά πιο σύνθετα μοντέλα μόνο εάν έχουμε ενδείξεις ότι βελτιώνουν σημαντικά την ακρίβεια της πρόβλεψης.

5)
Έχουμε το μοντέλο lnY=α+βX+ε.

```
> ln_grd<-log(grade_grdt)
> fitln_grd<-lm(ln_grd ~ grade_entry)
> summary(fitln_grd)
Call:
lm(formula = ln_grd ~ grade_entry)
Residuals:
      Min       1Q    Median       3Q      Max
-0.093293 -0.015410  0.005273  0.033953  0.049966
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25027    0.26113   0.958      0.355
grade_entry  0.09725 0.01433   6.788 1.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.04456 on 13 degrees of freedom
Multiple R-squared:   0.78,     Adjusted R-squared:  0.763
F-statistic: 46.08 on 1 and 13 DF,  p-value: 1.285e-05
```

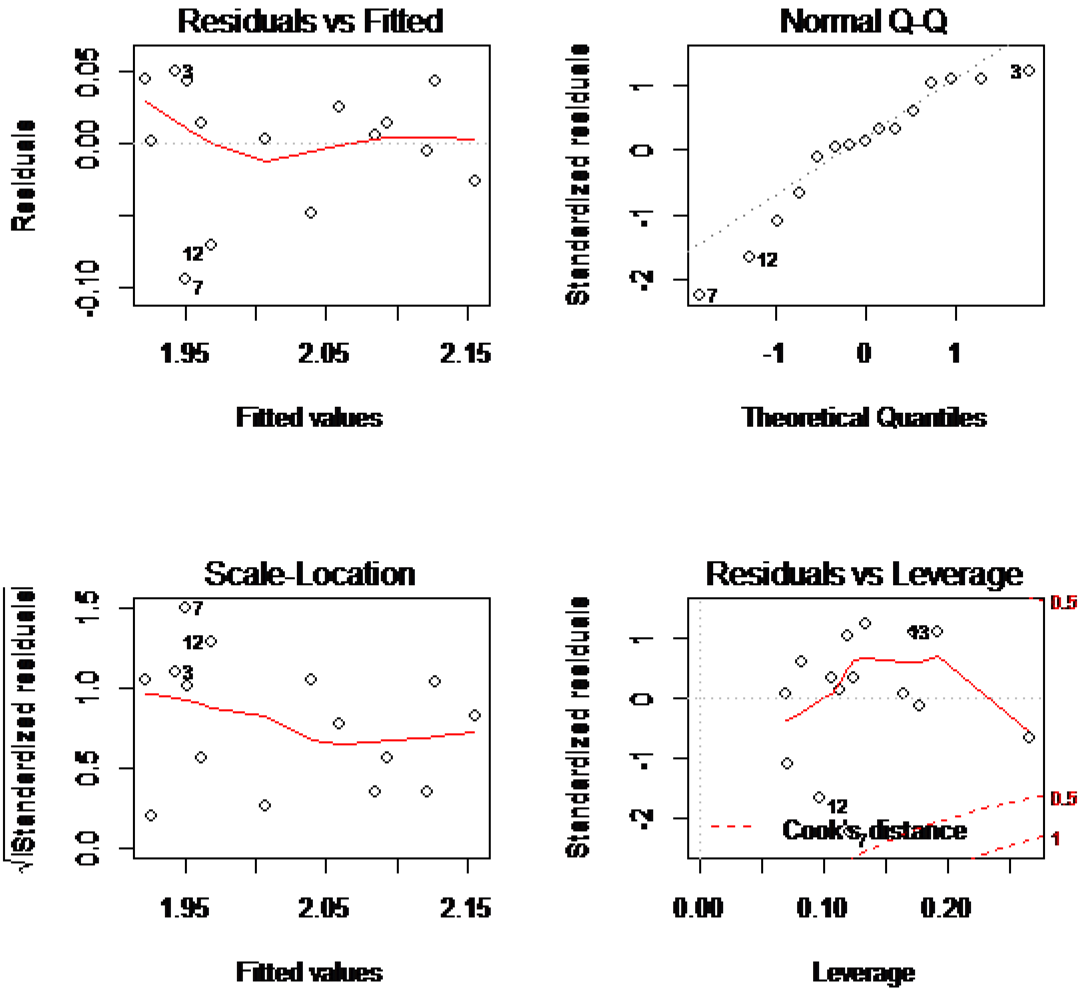
Συνεπώς το υπόδειγμα είναι $ln(grade\_grdt) = 0.25027 + 0.09725 * grade\_entry$
Θα το αναπαραστήσουμε γραφικά:

```
> plot(grade_entry, ln_grd, main="ln(vathmos apofoithshs) - vathmos eisagwghs along with the OLS line", xlab="vathmos apofoithshs",ylab="ln(vathmos apofoithshs)", pch=20)
> abline(fitln_grd, col="red")
```



Λαμβάνουμε τα διαγράμματα ελέγχου:

```
> par(mfrow=c(2,2))
> plot(fitln_grd)
```



Παρατηρώ ότι το μοντέλο αυτό δίνει Multiple R-squared: 0.78, το οποίο είναι κοντά στην τιμή που έδιναν τα προηγούμενα μοντέλα, όμως δίνει Residual standard error: 0.04456, το οποίο είναι σημαντικά μικρότερο από τα προηγούμενα.

Για να αποφανθούμε αν το μοντέλο κάνει fit στα δεδομένα, ελέγχουμε την p-value: $p - value \; = \; 0.0000128 \; < \; a = 0.05$.

Συνεπώς απορριπτουμε την H0, που δηλώνει ότι το μοντέλο αυτό δεν εκτιμά αποτελεσματικά τα δεδομένα.

Ζητείται το ποσοστό της μεταβλητότητας του Y, δηλαδή το r^2.

$$r^2 \; = \; SSR \,/\, SSTO \; = \; SSR \,/\, (SSR \; + \; SSE) \; = \; 0.78 \,/\, (0.78 + 0.04456) \; = \; 94{,}5959\%.$$

Συνεπώς, το 94,5959% της μεταβλητότητας της εξαρτημένης μεταβλητής ερμηνεύεται από το γραμμικό υπόδειγμα.

Άσκηση 2

1)

Αρχικά διαβάζουμε το dataset και δίνουμε στις μεταβλητής ονόματα κατ' αναλογία της εκφώνησης:

```
> dataall<-
read.table("C:/Users/leo/Desktop/AUEB/Probability_and_Statistics_/Assignment_3/Assignm
ent3-askisi2.txt")
> names(dataall) <- c('y','x1','x2','x3','x4','x5')
```

Προς απλοποίηση του κώδικα στο R, θέτουμε:

```
> x1 <- dataall["x1"]
> x2 <- dataall["x2"]
> x3 <- dataall["x3"]
> x4 <- dataall["x4"]
> x5 <- dataall["x5"]
> y <- dataall["y"]
```

Στο σημείο αυτό θα κάνουμε unlist τα y,x1,x2,x3,x4,x5 ώστε να τα διαχειριστούμε σαν vectors και να πραγματοποιήσουμε την ανάλυση:

```
> y <- unlist(y, use.names=FALSE)
> x1 <- unlist(x1, use.names=FALSE)
> x2 <- unlist(x2, use.names=FALSE)
> x3 <- unlist(x3, use.names=FALSE)
> x4 <- unlist(x4, use.names=FALSE)
> x5 <- unlist(x5, use.names=FALSE)

> fit1 <- lm( y ~ x1 + x4 + x5)
> fit2 <- lm( y ~ x1 + x3 + x4)
> fit3 <- lm( y ~ x1 + x3)
```

Κάνουμε ένα πρώτο έλεγχο μελετώντας τα summaries για τα μοντέλα:

```
> summary(fit1)
Call:
lm(formula = y ~ x1 + x4 + x5)
Residuals:
      Min      1Q   Median    3Q       Max
-10.4147  -2.1197   0.9254   1.8273   5.1872
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.4798    3.7553  13.176 5.14e-13 ***
x1            2.1600    0.3374   6.402 8.79e-07 ***
x4            0.2394    0.2483   0.964   0.3439
```

```
x5              3.4168      1.4995   2.279   0.0311 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4.047 on 26 degrees of freedom
Multiple R-squared:  0.7909,  Adjusted R-squared:  0.7667
F-statistic: 32.77 on 3 and 26 DF,  p-value: 5.503e-09

> summary(fit2)
Call:
lm(formula = y ~ x1 + x3 + x4)
Residuals:
     Min      1Q  Median      3Q      Max
-7.3691 -1.3323  0.3038  1.2954  5.7429
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.6912    6.8229   3.179  0.00379 **
x1              0.9573      0.3436   2.786  0.00983 **
x3              6.5856      1.3325   4.942 3.91e-05 ***
x4              0.3764      0.1967   1.913  0.06676 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.183 on 26 degrees of freedom
Multiple R-squared:  0.8706,  Adjusted R-squared:  0.8557
F-statistic: 58.33 on 3 and 26 DF,  p-value: 1.116e-11

> summary(fit3)
Call:
lm(formula = y ~ x1 + x3)
Residuals:
     Min      1Q  Median      3Q      Max
-8.9528 -1.1951  0.6049  1.9241  5.4247
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.2741    7.1441   3.118 0.004296 **
x1              1.3584      0.2854   4.759 5.82e-05 ***
x3              6.2696      1.3858   4.524 0.000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.336 on 27 degrees of freedom
Multiple R-squared:  0.8524,  Adjusted R-squared:  0.8415
F-statistic: 77.97 on 2 and 27 DF,  p-value: 6.054e-12
```

Παρατηρώ ότι όλα τα μοντέλα δίνουν $p - value < a$ για$a = 0.05$, συνεπώς όλα αποτελούν μία ‘καλή’ εκτίμηση των δεδομένων. Επίσης παρατηρώ ότι τα μοντέλα 1 και 3 δεν είναι nested καθώς δεν είναι το ένα υπερσύνολο του άλλου ως προς τις μεταβλητές που χρησιμοποιούν. Όμως αυτό ισχύει στα μοντέλα 2 και 3 (fit2 - fi3). Θα πραγματοποιήσουμε f-test, με τα ακόλουθα δεδομένα:

H0 Hypothesis: Τα μοντέλα fit2 και fit3, δεν διαφέρουν ιδιαίτερα.
H1: To full model fit2 είναι σημαντικά καλύτερο από το Reduced model fit3.

```
> anova(fit2,fit3)
Analysis of Variance Table
Model 1: y ~ x1 + x3 + x4
Model 2: y ~ x1 + x3
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1      26 263.41
2      27 300.51 -1   -37.093 3.6613 0.06676 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Θεωρώ α = 0.1. Τότε $p - value \ = \ 0.06676 \ < \ \alpha$ και απορρίπτω την Null Hypothesis σε επίπεδο σημαντικότητας 90% (a= 10%).

Συνεπώς, το μοντέλο fit2 δίνει καλύτερα αποτελέσματα από το fit3.

Τώρα πρέπει να συγκρίνω τα μοντέλα fit1 και fit2, ώστε να καταλήξω στο πιο κατάλληλο.

Το μοντέλο fit1 δίνει Multiple R-squared: 0.7909 και Residual standard error: 4.047
Το μοντέλο fit2 δίνει Multiple R-squared: 0.8706 και Residual standard error: 3.183.

Συνεπώς ούτε η σύγκριση αυτή προδίδει την καταλληλότητα κάποιου μοντέλου.

Εργάζομαι ως εξής:

Δημιουργώ ένα τέταρτο μοντέλο, το fit4 στο οποίο προσθέτω τη μεταβλητή x3. Με τον τρόπο, αν και δεν μπορώ άμεσα να συγκρίνω τα fit1, fit2, μπορώ όμως να συγκρίνω και τα δύο με το fit4 με το οποίο αμφότερα είναι nested models.

Έτσι, θα είναι δυνατόν να κατανοηθεί ποιας μεταβλητής η πληροφορία βελτιώνει το μοντέλο και ποιας όχι.

```
> fit4 <- lm( y ~ x1 + x4 + x5 + x3)
> anova (fit1,fit4)
Analysis of Variance Table
Model 1: y ~ x1 + x4 + x5
Model 2: y ~ x1 + x4 + x5 + x3
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1      26 425.86
2      25 263.32  1    162.53 15.431 0.0005956 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova (fit2,fit4)
Analysis of Variance Table
Model 1: y ~ x1 + x3 + x4
Model 2: y ~ x1 + x4 + x5 + x3
```

```
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     26 263.41
2     25 263.32  1  0.089658 0.0085 0.9272
```

H ANOVA ανάλυση στα μοντέλα fit1 - fit4 δίνει p-value = 0.0005956 < α για α=0.1, άρα απορρίπτω τη null Hypothesis H0 (ότι τα 2 μοντέλα είναι εξίσου κατάλληλα) και η γνώση της μεταβλητής x3 καθιστά το μοντέλο fit4 καταλληλότερο.

Αντίθετα, η ANOVA ανάλυση στα μοντέλα fit2 - fit4 δίνει p-value 0.9272, άρα δεν απορρίπτω τη Null Hypothesis που δηλώνει ότι το μοντέλο fit4 είναι καταλληλότερο. Συνεπώς, αν έπρεπε να επιλέξω ένα από τα δύο, θα επέλεγα το fit2 ώς απλούστερο.

Άρα, αφού fit2 > fit4 (> σημαίνει καταλληλότερο) και fit4 > fit1 => fit2>fit1.

Συνεπώς το μοντέλο που θα υιοθετούσα για την πρόβλεψη της τιμής πώλησης των κατοικιών θα στηριζόταν σε 3 μεταβλητές:

- X1: Μέγεθος διαμερίσματος (σε τετραγωνικά πόδια)
- X3: Συνολικός αριθμός δωματίων
- X4: Ηλικία διαμερίσματος (σε έτη)

, όπως εισηγήθηκε ο δεύτερος ερευνητής.

Άσκηση 3

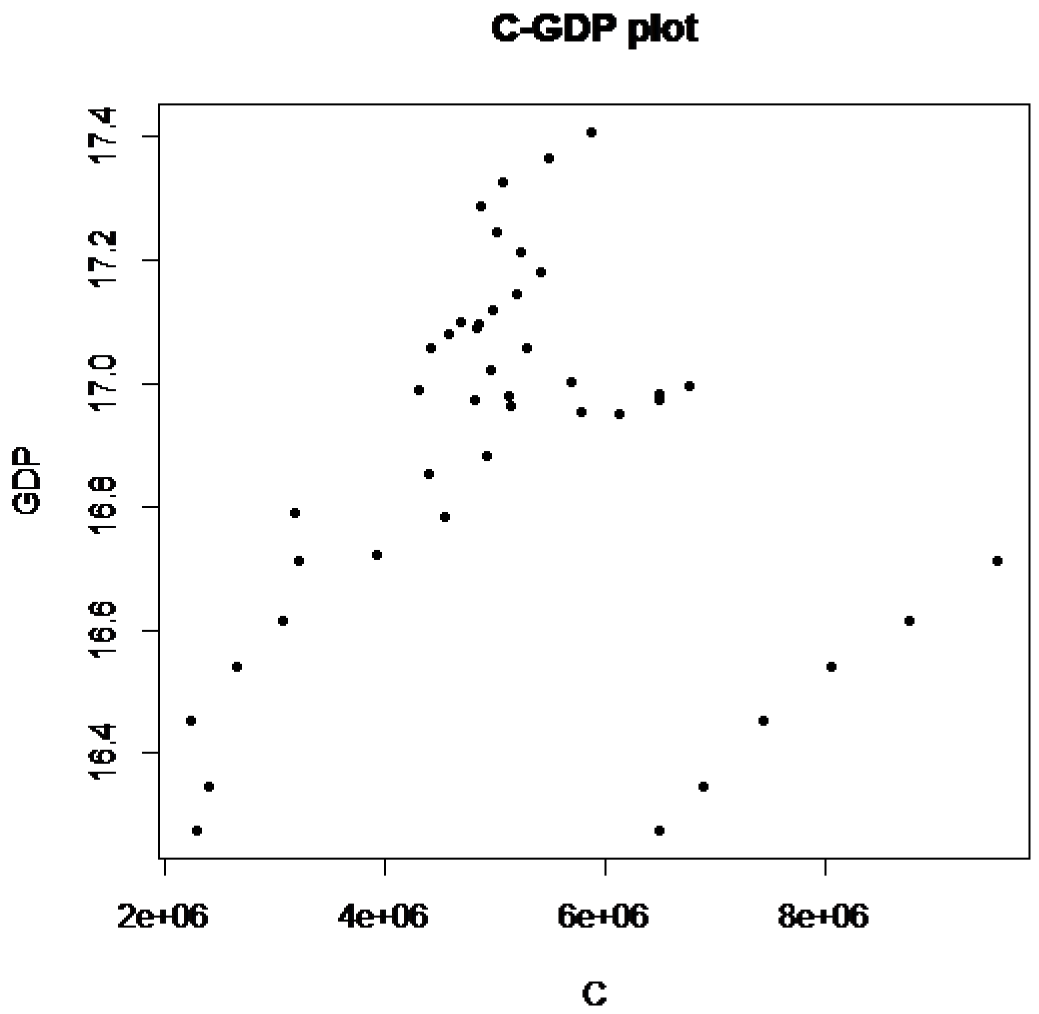
A)

1)

```
> dt<-
read.table("C:/Users/leo/Desktop/AUEB/Probability_and_Statistics_/Assignment_3/Assignm
ent3-askisi3.txt")
> names(dt) <- c('year','GDP','C','L')

> year <- dt["year"]
> GDP <- dt["GDP"]
> C <- dt["C"]
> L <- dt["L"]

> year <- unlist(year, use.names=FALSE)
> GDP <- unlist(GDP, use.names=FALSE)
> C <- unlist(C, use.names=FALSE)
> L <- unlist(L, use.names=FALSE)

> ln_gdp<-log(GDP)
```

Θα αναπαραστήσουμε γραφικά τη σχέση μεταξύ των δύο μεγεθών:

```
> plot(C,ln_gdp, main="C-GDP plot", xlab="C", ylab="GDP", pch=20)
```



Και διαισθητικά καταλαβαίνουμε ότι τα μεγέθη ln_gdp και C δεν έχουν γραμμική εξάρτηση.

Προς επιβεβαίωση, πραγματοποιούμε correlation test :

```
> cor.test(ln_gdp,C)

        Pearson's product-moment correlation

data:  ln_gdp and C
t = 0.52803, df = 41, p-value = 0.6003
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2236792  0.3733149
sample estimates:
        cor
0.08218593
```

Παρατηρούμε ότι ο συντελεστής συσχέτισης είναι 0.08218593, άρα δεν υπάρχει ισχυρή γραμμική συσχέτιση, κάτι που επιβεβαιώνεται από την τιμή$p - value = 0.6003 > a = 0.05$.

2)

Ακολουθεί η προσπάθεια να γίνει fit κάποιο γραμμικό μοντέλο:

```
> fitln_gdp<-lm(ln_gdp ~ C)
> summary(fitln_gdp)

Call:
lm(formula = ln_gdp ~ C)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63977 -0.19936  0.06822  0.20231  0.50172

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.681e+01  1.589e-01 105.838   <2e-16 ***
C           1.558e-08  2.950e-08   0.528      0.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3042 on 41 degrees of freedom
Multiple R-squared:  0.006755,  Adjusted R-squared:  -0.01747
F-statistic: 0.2788 on 1 and 41 DF,  p-value: 0.6003
```

Ο συντελεστής β_oείναι ίσος με 16.81, ενώ ο β_1είναι ίσος με $1.558 * 10^{-8}$.

Εργάζομαι με συντελεστή εμπιστοσύνης 95% συνεπώς απορρίπτω τη Null Hupothesis αν η τιμή ενός εκ των δύο συντελεστών βρίσκεται εκτός του $(T_{0.025}, T_{0.975})$διαστήματος.

Για το συντελεστή συσχέτισης προκύπτει $T_{0.025} < T > T_{0.975}$,όπου $T = -2.018$ άρα δεν απορρίπτω τη Null Hypothesis και ο συντελεστής δεν είναι στατιστικά σημαντικός. Αντίθετα, ο συντελεστής προσδιορισμού είναι στατιστικά σημαντικός.

```
> confint(fitln_gdp, level = 0.95)
                    2.5 %       97.5 %
(Intercept)  1.649351e+01 1.713519e+01
C           -4.400631e-08 7.516510e-08
```

Συνεπώς ο συντελεστής συσχέτισης κυμαίνεται από -4.400631e-08 ως 7.516510e-08 και ο συντελεστής προσδιορισμού από 16.4935 ως17.1351 με συντελεστή εμπιστοσύνης 95%.

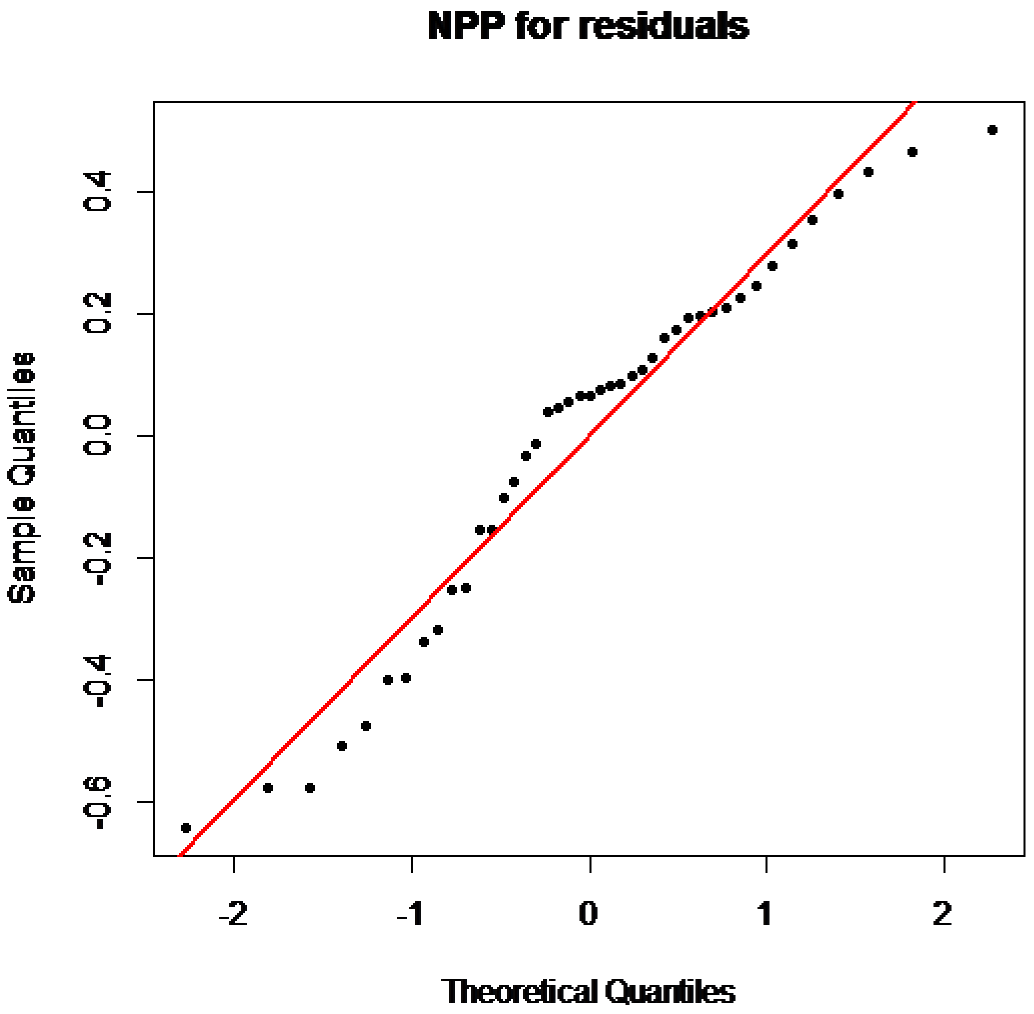
3)

Θα εξεταστεί εάν ικανοποιούνται όλες οι υποθέσεις του μοντέλου (κανονικότητα, ομοσκεδαστικότητα και τυχαιότητα των καταλοίπων) κατασκευάζοντας τα ακόλουθα γραφήματα:

α. Διάγραμμα καταλοίπων/τυποποιημένων καταλοίπων με τις προβλεπόμενες τιμές της εξαρτημένης μεταβλητής

β. Normal probability plot ή QQplot των τυποποιημένων καταλοίπων για να παρατηρήσουμε την κανονικότητα των καταλοίπων.

```
> qqnorm(fitln_gdp$residuals,main="NPP for residuals",pch=20)
> qqline(fitln_gdp$residuals,col="red",lty=1,lwd=2)
```




```
> lillie.test(fitln_gdp$residuals)
```

Lilliefors (Kolmogorov-Smirnov) normality test

data: fitln_gdp$residuals
D = 0.15876, p-value = 0.008165

Ο έλεγχος Kolmogorov-Smirnov δίνει p-value = 0.008<0.05.

Επομένως δεν ικανοποιούνται όλες οι υποθέσεις του μοντέλου, άρα απορρίπτουμε την Η0 και θεωρούμε ότι το υπόδειγμα δεν είναι αξιόπιστο.

B)

1)

```
> lnp <- log(GDP)
> lnc <- log(C)
> lnl <- log(L)
> fitM1 <- lm(lnp ~ lnc)
> fitM2 <- lm(lnp ~ lnl)
> fitM3 <- lm(lnp ~ lnc + lnl)
```

Θα εφαρμόσω ANOVA μεταξύ των μοντέλων 1 fitM1 και fitM3 , αφού είναι nested (fitM3=full, fitM1= Reduced).

H0 Hypothesis: Τα μοντέλα fitM1 και fitM3, δεν διαφέρουν ιδιαίτερα.
H1: To full model fitM3 είναι σημαντικά καλύτερο από το Reduced model fitM1.

```
> anova(fitM1,fitM3)
Analysis of Variance Table

Model 1: lnp ~ lnc
Model 2: lnp ~ lnc + lnl
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     41 3.5866
2     40 3.5719  1  0.014721 0.1649 0.6869
```

H p-value είναι 0.6869, άρα σε επίπεδο σημαντικότητας 5% δεν απορρίπτω την H0.

H0 Hypothesis: Τα μοντέλα fitM2 και fitM3, δεν διαφέρουν ιδιαίτερα.
H1: To full model fitM3 είναι σημαντικά καλύτερο από το Reduced model fitM2.

```
> anova(fitM2,fitM3)
Analysis of Variance Table

Model 1: lnp ~ lnl
Model 2: lnp ~ lnc + lnl
```

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     41 3.8046
2     40 3.5719  1   0.23268 2.6057 0.1143
```

Η p-value είναι 0.1143 , άρα σε επίπεδο σημαντικότητας 5% δεν απορρίπτω την H0.

Συνεπώς εάν έπρεπε να επιλέξω μεταξύ fitM1 - fitM3 θα επέλεγα τη fitM1 ως απλούστερη.
Όμοια, εάν έπρεπε να επιλέξω μεταξύ fitM2 - fitM3 θα επέλεγα τη fitM2 ως απλούστερη.

Συνεπώς θα πρέπει να επιλέξω μεταξύ των μοντέλων fitM1 και fitM2 .

```
> summary(fitM1)

Call:
lm(formula = lnp ~ lnc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68243 -0.14692  0.02907  0.20204  0.47211

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.4110     2.1288   6.300 1.62e-07 ***
lnc           0.2261     0.1382   1.637    0.109
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2958 on 41 degrees of freedom
Multiple R-squared:  0.06133,   Adjusted R-squared:  0.03844
F-statistic: 2.679 on 1 and 41 DF,  p-value: 0.1093

> summary(fitM2)

Call:
lm(formula = lnp ~ lnl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64994 -0.19019  0.09458  0.19044  0.49252

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.9387     9.4158   1.374    0.177
lnl           0.2625     0.6248   0.420    0.677

Residual standard error: 0.3046 on 41 degrees of freedom
Multiple R-squared:  0.004287,  Adjusted R-squared:  -0.02
```

F-statistic: 0.1765 on 1 and 41 DF, p-value: 0.6766

Το πρώτο μοντέλο fitM1 δίνει$p - value \ = \ 0.1093$. Συνεπώς δεν την απορρίπτω σε επίπεδο σημαντικότητας 5%.

Το δεύτερο μοντέλο fitM2 δίνει$p - value \ = \ 0.6766$. Προφανώς δεν δέχομαι την H0, αλλά η τιμή αυτή αποτελεί μια πρώτη ένδειξη πως ίσως το μοντέλο αυτό κάνει καλύτερο fit στα data.

H fitM1 δίνει Multiple R-squared = 0.06133 και Residual standard error = 0.2958.
H fitM2 δίνει Multiple R-squared = 0.004287 Residual standard error = 0.3046.

Παρατηρώντας ότι τα τυπικά σφάλματα είναι περίπου ίσα και ότι η ποσότητα Multiple R-squared του fitM2 είναι μία τάξη μεγέθους μικρότερη από του fitM1, θα επιλέξω το fitM2 ως υπόδειγμα.

2)

```
> result = exp(12.9387 + 0.2625 * log (3500000))
> result
[1] 21728206
```

Συνεπώς το μοντέλο δεν λαμβάνει υπόψιν την τιμή του C. Βάσει του L και του μοντέλου fitM2 δίνει πρόβλεψη για το ΑΕΠ : result = 21728206.