Άσκηση 2

Για την εισαγωγή των δεδομένων της άσκησης δίνουμε την εντολή

> data<-read.csv(file.choose(),header=TRUE,sep='')

και επιλέγουμε το αρχείο “seira2_exercise2”.

Επίσης δημιουργούμε τις πιο κάτω μεταβλητές από τα δεδομένα του αρχείου με τις εντολές:

> x1<-data$x1

> x2<-data$x2

> x3<-data$x3

> x4<-data$x4

> x5<-data$x5

> x6<-data$x6

> y<-data$y

1. Για τον υπολογισμό του συντελεστή συσχέτισης $r_{x_iy_i}$, i≠j, j=1,...,6 των Χ μεταβλητών δίνουμε την εντολή
   > cor(data)

   και παίρνουμε το εξής αποτέλεσμα

```
            x1         x2           x3         x4           x5         x6
x1  1.00000000  0.1035827  0.045848007  0.1639008  0.751999949 -0.8401788
x2  0.10358266  1.0000000 -0.158956233 -0.1906652  0.135396967  0.2249824
x3  0.04584801 -0.1589562  1.000000000  0.2834654 -0.009432581  0.1375657
x4  0.16390083 -0.1906652  0.283465396  1.0000000  0.487903889 -0.2185308
x5  0.75199995  0.1353970 -0.009432581  0.4879039  1.000000000 -0.6490191
x6 -0.84017881  0.2249824  0.137565688 -0.2185308 -0.649019055  1.0000000
y   0.86592697  0.4107667 -0.075829536 -0.1308151  0.618701076 -0.6337331
             y
x1  0.86592697
x2  0.41076665
x3 -0.07582954
x4 -0.13081512
x5  0.61870108
x6 -0.63373305
y   1.00000000
```

   όπου έχουμε ότι:

   $cor(x_1,x_2)=cor(x_2,x_1)=0.104$

   $cor(x_1,x_3)=cor(x_3,x_1)=0.046$

   $cor(x_1,x_4)=cor(x_4,x_1)=0.164$

   $cor(x_1,x_5)=cor(x_5,x_1)=0.752$

   $cor(x_1,x_6)=cor(x_6,x_1)=-0.840$

   $cor(x_2,x_3)=cor(x_3,x_2)=-0.159$

   $cor(x_2,x_4)=cor(x_4,x_2)=-0.191$

   $cor(x_2,x_5)=cor(x_5,x_2)=0.135$

   $cor(x_2,x_6)=cor(x_6,x_2)=0.225$

   $cor(x_3,x_4)=cor(x_4,x_3)=0.283$

   $cor(x_3,x_5)=cor(x_5,x_3)=-0.009$

$cor(x_3,x_6)=cor(x_6,x_3)=0.137$
$cor(x_4,x_5)=cor(x_5,x_4)=0.488$
$cor(x_4,x_6)=cor(x_6,x_4)=-0.219$
$cor(x_5,x_6)=cor(x_6,x_5)=-0.649$

Για την προσαρμογή του μοντέλου χρησιμοποιούμε την εντολή:
> full_model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6, data = data)
Και από την εντολή:
> summary(full_model)
μπορούμε να δούμε τα χαρακτηριστικά του μοντέλου που προσαρμόσαμε.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63405 -0.18306 -0.07142  0.19190  0.71298 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   1.2907     7.1883   0.180   0.8603  
x1            5.6794     1.9458   2.919   0.0120 *
x2            3.1629     1.3432   2.355   0.0349 *
x3            0.2561     1.1594   0.221   0.8286  
x4           -0.8958     0.3904  -2.295   0.0391 *
x5            1.1765     1.6590   0.709   0.4907  
x6           -1.8798     6.5623  -0.286   0.7790  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4586 on 13 degrees of freedom
Multiple R-squared:  0.9037,	Adjusted R-squared:  0.8593 
F-statistic: 20.34 on 6 and 13 DF,  p-value: 6.607e-06
```

2. Για τους ελέγχους H_0: $β_j=0$ με εναλλακτική H_1: $β_j≠0$ μπορούμε να δούμε από τα πιο πάνω αποτελέσματα του t-test ότι για:
   j=1: p-value=0.0120 < 0.05 που είναι το επίπεδο σημαντικότητας άρα απορρίπτω την H_0
   j=2: p-value=0.0349 < 0.05 που είναι το επίπεδο σημαντικότητας άρα απορρίπτω την H_0
   j=3: p-value=0.8246 > 0.05 που είναι το επίπεδο σημαντικότητας άρα αποδέχομαι την H_0
   j=4: p-value=0.0391 < 0.05 που είναι το επίπεδο σημαντικότητας άρα απορρίπτω την H_0
   j=5: p-value=0.4907 > 0.05 που είναι το επίπεδο σημαντικότητας άρα αποδέχομαι την H_0
   j=6: p-value=0.7790 > 0.05 που είναι το επίπεδο σημαντικότητας άρα αποδέχομαι την H_0

   Επίσης, από τα πιο πάνω p-values παρατηρούμε ότι οι μεταβλητές x_1,x_2 και x_4 είναι στατιστικά σημαντικές

   Ένας άλλος τρόπος να ελεγχθεί αυτό είναι μέσα από την εντολή

   > anova(full_model)

   Από τον πιο κάτω πίνακα ANOVA παρατηρώντας τα p-values βλέπουμε πως οι μεταβλητές x_1,x_2 και x_4 είναι στατιστικά σημαντικές αφού p-value< 0.05 που είναι το επίπεδο σημαντικότητας επομένως σε αυτές τις περιπτώσεις απορρίπτεται και η H_0.

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
x1         1 21.3032 21.3032 101.2741 1.672e-07 ***
x2         1  2.9605  2.9605  14.0742   0.00242 ** 
x3         1  0.1139  0.1139   0.5417   0.47480    
x4         1  1.1774  1.1774   5.5975   0.03420 *  
x5         1  0.1038  0.1038   0.4933   0.49485    
x6         1  0.0173  0.0173   0.0821   0.77904    
Residuals 13  2.7346  0.2104                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Για να ελέγξουμε την πολυσυγγραμμικότητα δίνουμε τις εντολές

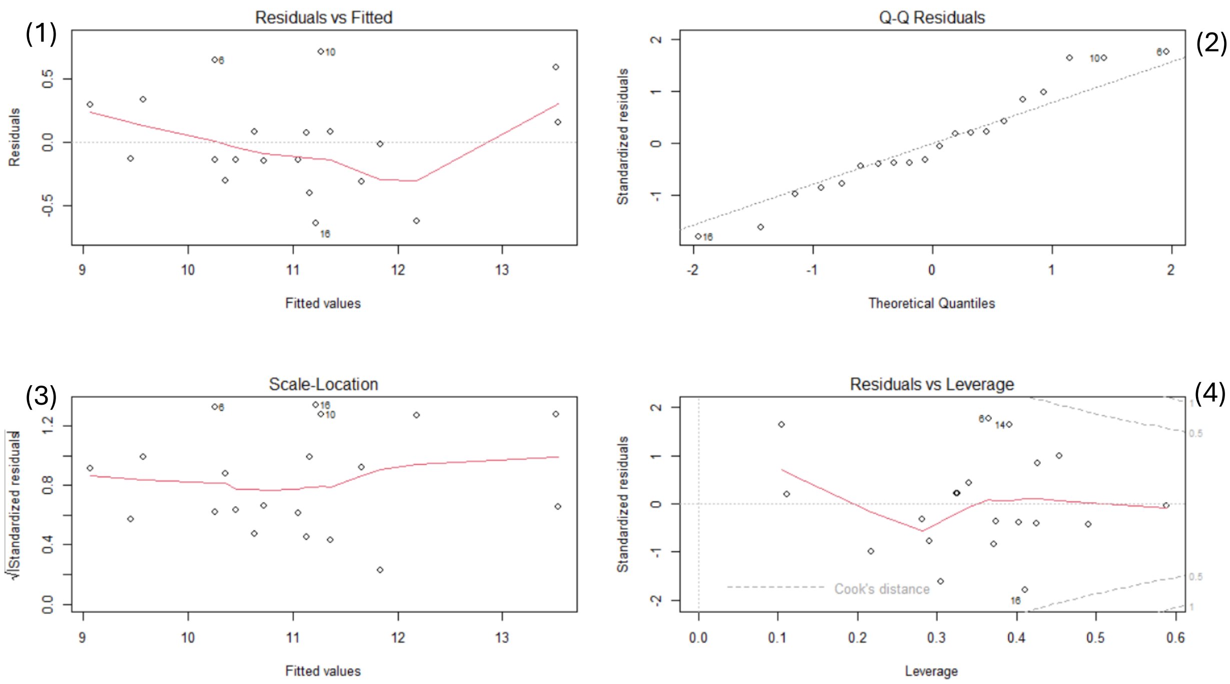
```
> library(car)
> vif(full_model)
```

```
      x1       x2       x3       x4       x5       x6 
8.551698 1.972474 1.670979 1.968215 3.860328 7.340394 
```

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι για τις μεταβλητές x_1και x_6 ο παράγοντας μεγέθυνσης διασποράς(VIF) είναι >5 επομένως υπάρχει πολυσυγγραμμικότητα.

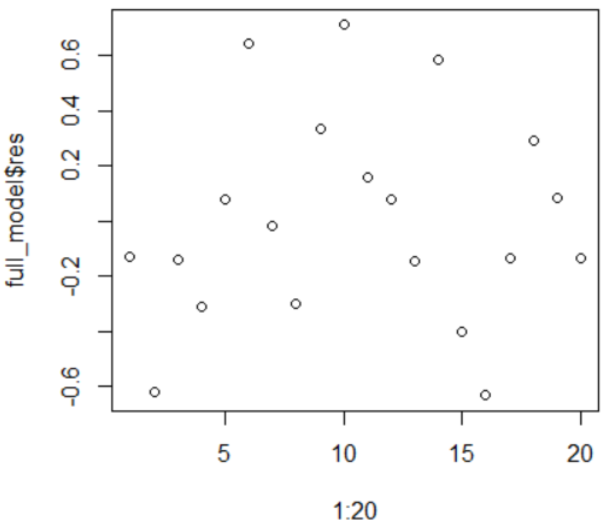
3. Για τη εξέταση των υπολοίπων δίνουμε τις εντολές

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



Στο διάγραμμα (1) βλέπουμε τα υπόλοιπα σε σχέση με τις προβλεπόμενες τιμές. Τα υπόλοιπα εδώ φαίνονται να είναι διασκορπισμένα τυχαία με μια ελαφριά καμπυλότητα, γεγονός που μπορεί να υποδηλώνει ότι το μοντέλο ενδέχεται να μην περιγράφει επαρκώς τη γραμμική σχέση μεταξύ των μεταβλητών. Επίσης, δεν υπάρχει σαφής τάση στην κατανομή των υπολοίπων γεγονός που δείχνει ότι ισχύει η ομοσκεδαστικότητα, ωστόσο υπάρχουν μερικές παρατηρήσεις που δείχνουν αύξηση της διασποράς στις υψηλές

προβλεπόμενες τιμές, αυτό υποδεικνύει πιθανές αποκλίσεις από την υπόθεση της ομοσκεδαστικότητας.
Στο διάγραμμα (2) βλέπουμε την σύγκριση της κατανομής των υπολοίπων με την κανονική κατανομή. Οι περισσότερες τιμές βρίσκονται κοντά στη διαγώνιο γραμμή, υποδεικνύοντας ότι η υπόθεση της κανονικότητας ισχύει γενικά. Ωστόσο, υπάρχουν κάποιες αποκλίσεις στα άκρα που μπορεί να υποδεικνύουν ελαφρά απόκλιση.
Στο διάγραμμα (3) παρουσιάζεται η τετραγωνική ρίζα των τυποποιημένων υπολοίπων σε σχέση με τις προβλεπόμενες τιμές. Η κόκκινη γραμμή παραμένει σχετικά επίπεδη, κάτι που είναι ενδεικτικό ότι η υπόθεση της ομοσκεδαστικότητας γενικά ικανοποιείται, ωστόσο υπάρχουν μερικά σημεία που παρουσιάζουν μεγαλύτερη διασπορά υποδεικνύοντας πιθανή ανομοιογένεια.
Στο διάγραμμα (4) υπολοίπων έναντι μοχλού βλέπουμε αν υπάρχουν παρατηρήσεις που επηρεάζουν σημαντικά το μοντέλο (outliers ή σημεία υψηλής μόχλευσης). Παρατηρούμε πως ορισμένα σημεία με υψηλή μόχλευση πλησιάζουν την καμπύλη του Cook's Distance υποδηλώνοντας ότι ενδέχεται να έχουν σημαντική επιρροή στις εκτιμήσεις του μοντέλου.
Για την εξέταση της ανεξαρτησίας των υπολοίπων δίνουμε την εντολή
> plot(1:20,full_model$res)



Από αυτό το διάγραμμα παρατηρούμε πως τα υπόλοιπα δεν ακολουθούν κάποιο συγκεκριμένο μοτίβο επομένως ισχύει η υπόθεση της ανεξαρτησίας.

4. Για να εξετάσουμε αν υπάρχουν σημεία επιρροής δίνουμε την εντολή
> cooks.distance(full_model)

```
           1            2            3            4            5            6 
0.0060054471 0.1653722846 0.0171442912 0.0608966397 0.0028490023 0.2551654604 
           7            8            9           10           11           12 
0.0005699129 0.0349317809 0.1148727913 0.0449970621 0.0133696260 0.0006013775 
          13           14           15           16           17           18 
0.0267871884 0.2461906479 0.0387078664 0.3220196718 0.0144463215 0.0745438406 
          19           20 
0.0034097026 0.0121981772 
```

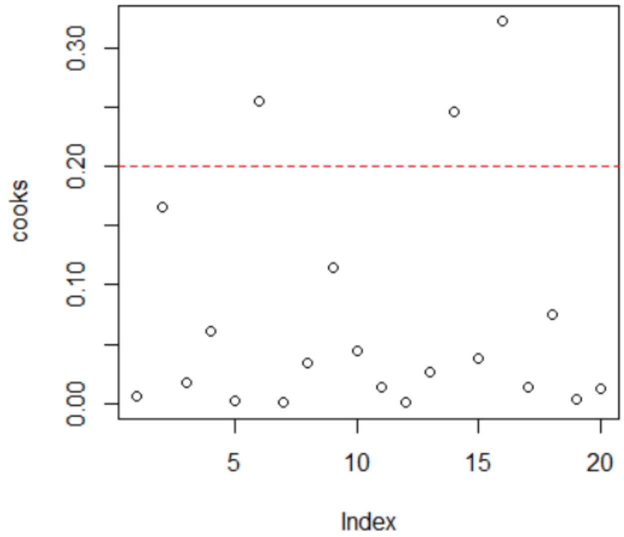
Από τα παραπάνω αποτελέσματα παρατηρούμε ότι οι περισσότερες τιμές είναι <0.5, υποδεικνύοντας ότι οι περισσότερες τιμές δεν είναι ιδιαίτερα επιδραστικές. Ωστόσο, οι παρατηρήσεις 2, 6, 14 και 16 εμφανίζουν σχετικά υψηλές τιμές, γεγονός που δείχνει ότι ενδέχεται να έχουν επιρροή στο μοντέλο.

Για περαιτέρω εξέταση των σημείων επιρροής δίνουμε τις εντολές

> cooks<-cooks.distance(full_model)

> plot(cooks)

> abline(h=4/length(cooks),col="red", lty=2)



Παρατηρούμε ότι οι παρατηρήσει 6, 14 και 16 είναι πάνω από το όριο, επομένως θεωρούνται σημεία επιρροής.

5. α) Για την εκτέλεση του Forward Selection δίνουμε την εντολή
> forward_model <- step(lm(y ~ 1, data = data), scope = formula(full_model), direction = "forward", test='F')

```
Start:  AIC=9.02
y ~ 1

       Df Sum of Sq     RSS      AIC F value    Pr(>F)    
+ x1    1   21.3032  7.1075 -16.6916 53.9509 8.09e-07 ***
+ x6    1   11.4102 17.0005   0.7502 12.0811 0.002698 ** 
+ x5    1   10.8754 17.5353   1.3697 11.1635 0.003635 ** 
+ x2    1    4.7937 23.6170   7.3247  3.6536 0.072004 .  
<none>              28.4107   9.0207                    
+ x4    1    0.4862 27.9245  10.6754  0.3134 0.582508    
+ x3    1    0.1634 28.2473  10.9053  0.1041 0.750681    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Step:  AIC=-16.69
y ~ x1

       Df Sum of Sq    RSS     AIC F value    Pr(>F)   
+ x2    1   2.96054 4.1470 -25.467 12.1363 0.002843 **
+ x4    1   2.17175 4.9358 -21.985  7.4800 0.014105 * 
+ x6    1   0.84996 6.2576 -17.239  2.3091 0.146999   
<none>              7.1075 -16.692                    
+ x3    1   0.38000 6.7275 -15.790  0.9602 0.340865   
+ x5    1   0.06896 7.0385 -14.887  0.1666 0.688276   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Step:  AIC=-25.47
y ~ x1 + x2

       Df Sum of Sq    RSS     AIC F value  Pr(>F)  
+ x4    1   1.28883 2.8581 -30.911  7.2149 0.01623 *
<none>              4.1470 -25.467                  
+ x5    1   0.17233 3.9747 -24.316  0.6937 0.41717  
+ x3    1   0.11395 4.0330 -24.024  0.4521 0.51095  
+ x6    1   0.00809 4.1389 -23.506  0.0313 0.86187  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-30.91
y ~ x1 + x2 + x4

       Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>               2.8581 -30.911               
+ x5    1  0.105015 2.7531 -29.660  0.5722 0.4611
+ x6    1  0.017118 2.8410 -29.031  0.0904 0.7678
+ x3    1  0.002551 2.8556 -28.929  0.0134 0.9094
```

<u>1ο Βήμα</u>

M_0: H_0: $y=\beta_0+\varepsilon$

M_1: H_1: $y=\beta_0+\beta_j x_j+\varepsilon^*$ για κάθε επεξηγηματική μεταβλητή j=1,2,3,4,5,6

$SSE_0=28.4107$

$q=1$

$n-p=20-5=15$

<u>j=1</u>: M_1: $y=\beta_0+\beta_1 x_1+\varepsilon^*$

$SSE_1=7.1075$

$F_{11}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=53.9509$

$P(F> 539509)=8.09\times10^{-7}<0.001 \Rightarrow x_1$ χρειάζεται στο μοντέλο

<u>j=2</u>: M_2: $y=\beta_0+\beta_2 x_2+\varepsilon^*$

$SSE_1=23.6170$

$F_{12}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=3.6536$

$P(F> 3.6536)=0.072004>0.001 \Rightarrow x_2$ δεν χρειάζεται στο μοντέλο

<u>j=3</u>: M_3: $y=\beta_0+\beta_3 x_3+\varepsilon^*$

$SSE_1=28.2473$

$F_{13}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.1041$

$P(F>0.1041)=0.750681>0.001 \Rightarrow x_3$ δεν χρειάζεται στο μοντέλο

<u>j=4</u>: M_4: $y=\beta_0+\beta_4 x_4+\varepsilon^*$

$SSE_1=27.9245$

$F_{14}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.3134$

P(F> 0.3134)=0.582508>0.001=>x_4 δεν χρειάζεται στο μοντέλο

<u>j=5</u>: M_5: $y=\beta_0+\beta_5x_5+\varepsilon^*$

$SSE_1=17.5353$

$F_{15}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=11.1635$

P(F>11.1635)=0.003635>0.001=>x_5 δεν χρειάζεται στο μοντέλο

<u>j=6</u>: M_6: $y=\beta_0+\beta_6x_6+\varepsilon^*$

$SSE_1=17.0005$

$F_{16}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=12.0811$

P(F>12.0811)=0.002698>0.001=>x_6 δεν χρειάζεται στο μοντέλο

<u>2ο Βήμα</u>

M_0: H_0: $y=\beta_0+ \beta_1x_1 +\varepsilon^*$

M_1: H_1: $y=\beta_0 + \beta_1x_1 +\beta_jx_j+\varepsilon^*$ για κάθε επεξηγηματική μεταβλητή j=2,3,4,5,6

$SSE_0=7.1075$

<u>j=2</u>: $SSE_1=4.1470$

$F_{22}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=12.1363$

P(F> 12.1363)=0.002843=>x_2 χρειάζεται στο μοντέλο

<u>j=3</u>: $SSE_1=6.7275$

$F_{23}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.9602$

P(F>0.9602)=0.340865=>x_3 δεν χρειάζεται στο μοντέλο

<u>j=4</u>: $SSE_1=4.9358$

$F_{24}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=7.4800$

P(F>7.4800)=0.014105=>x_4 δεν χρειάζεται στο μοντέλο

<u>j=5</u>: $SSE_1=7.0385$

$F_{25}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.1666$

P(F>0.1666)=0.688276=>x_5 δεν χρειάζεται στο μοντέλο

<u>j=6</u>: $SSE_1=6.2576$

$F_{26}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=2.3091$

$P(F>2.3091)=0.146999 \Rightarrow x_6$ δεν χρειάζεται στο μοντέλο

<u>3^o Βήμα</u>

M_0: H_0: $y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon^*$

M_1: H_1: $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_jx_j+\varepsilon^*$ για κάθε επεξηγηματική μεταβλητή j=3,4,5,6

$SSE_0=4.1470$

<u>j=3</u>: $SSE_1=4.0330$

$F_{23}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.4521$

$P(F>0.4521)=0.51095 \Rightarrow x_3$ δεν χρειάζεται στο μοντέλο

<u>j=4</u>: $SSE_1=2.8581$

$F_{24}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=7.2149$

$P(F>7.2149)=0.01623 \Rightarrow x_4$ χρειάζεται στο μοντέλο

<u>j=5</u>: $SSE_1=3.9747$

$F_{25}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.6937$

$P(F>0.6937)=0.41717 \Rightarrow x_5$ δεν χρειάζεται στο μοντέλο

<u>j=6</u>: $SSE_1=4.1389$

$F_{26}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.0313$

$P(F>0.0313)=0.86187 \Rightarrow x_6$ δεν χρειάζεται στο μοντέλο

<u>4^o Βήμα</u>

M_0: H_0: $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_4x_4+\varepsilon^*$

M_1: H_1: $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_4x_4+\beta_jx_j+\varepsilon^*$ για κάθε επεξηγηματική μεταβλητή j=3,4,5,6

$SSE_0=2.8581$

<u>j=3</u>: $SSE_1=2.8556$

$F_{23}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.0134$

$P(F>0.0134)=0.9094 \Rightarrow x_3$ δεν χρειάζεται στο μοντέλο

<u>j=5</u>: $SSE_1=2.7531$

$F_{25}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.5722$

$P(F>0.5722)=0.4611 \Rightarrow x_5$ δεν χρειάζεται στο μοντέλο

<u>j=6</u>: $SSE_1=2.8410$

$$F_{26}=\frac{(SSE_0-SSE_1)/q}{SSE_1/(n-p)}=0.0904$$

$P(F> 0.0904)=0.7678 \Rightarrow x_6$ δεν χρειάζεται στο μοντέλο

Άρα τελικά με τη Forward Selection σε ε.σ. 5% καταλήγω στο μοντέλο $y=\beta_0+ \beta_1 x_1 + \beta_2 x_2+ \beta_4 x_4$.

Ουσιαστικά στην Forward Selection ξεκινάμε από το $y=\beta_0+\varepsilon$ και εισάγουμε την πρώτη «καλύτερη» επεξηγηματική μεταβλητή από τις x_1, x_2, x_3, x_4, x_5, x_6 υποψήφιες με βάση τον έλεγχο F, έστω την x_j. Στη συνέχεια εισάγουμε την αμέσως επόμενη «καλύτερη» επεξηγηματική μεταβλητή, έστω x_i , δεδομένου ότι η x_j είναι στο μοντέλο($i\neq j$). Έπειτα συνεχίζουμε έως ότου δεν υπάρχει άλλη μεταβλητή που χρειάζεται στο μοντέλο με βάση τον έλεγχο F.

Με την εντολή

```
> summary(forward_model)
```

παίρνουμε συνοπτικά τα στοιχεία του μοντέλου που προσαρμόστηκε.

```
Call:
lm(formula = y ~ x1 + x2 + x4, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.64524 -0.23678 -0.08526  0.14426  0.76835 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.4652     1.0767   2.290  0.03598 *  
x1            6.7532     0.6277  10.759 9.84e-09 ***
x2            3.0919     0.9066   3.410  0.00358 ** 
x4           -0.7144     0.2660  -2.686  0.01623 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4227 on 16 degrees of freedom
Multiple R-squared:  0.8994,	Adjusted R-squared:  0.8805 
F-statistic: 47.68 on 3 and 16 DF,  p-value: 3.342e-08
```

β) Για την εκτέλεση του Backward Selection δίνουμε την εντολή

```
> backward_model <- step(full_model,y~1, direction = "backward", test='F')
```

```
Start:  AIC=-25.8
y ~ x1 + x2 + x3 + x4 + x5 + x6

       Df Sum of Sq    RSS     AIC F value  Pr(>F)  
- x3    1   0.01026 2.7448 -27.720  0.0488 0.82863  
- x6    1   0.01726 2.7518 -27.669  0.0821 0.77904  
- x5    1   0.10580 2.8404 -27.036  0.5029 0.49073  
<none>              2.7346 -25.795                  
- x4    1   1.10748 3.8420 -20.994  5.2649 0.03905 *
- x2    1   1.16643 3.9010 -20.690  5.5452 0.03491 *
- x1    1   1.79208 4.5266 -17.715  8.5194 0.01197 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-27.72
y ~ x1 + x2 + x4 + x5 + x6

       Df Sum of Sq    RSS     AIC F value   Pr(>F)   
- x6    1   0.00830 2.7531 -29.660  0.0423 0.839970   
- x5    1   0.09619 2.8410 -29.031  0.4906 0.495120   
<none>              2.7448 -27.720                    
- x4    1   1.20980 3.9546 -22.417  6.1706 0.026262 * 
- x2    1   1.26803 4.0129 -22.125  6.4676 0.023427 * 
- x1    1   2.77635 5.5212 -15.743 14.1608 0.002098 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Step:  AIC=-29.66
y ~ x1 + x2 + x4 + x5

       Df Sum of Sq    RSS      AIC F value   Pr(>F)
- x5    1    0.1050 2.8581 -30.9112  0.5722 0.461119
<none>              2.7531 -29.6599
- x4    1    1.2215 3.9747 -24.3159  6.6553 0.020924 *
- x2    1    1.7189 4.4721 -21.9577  9.3653 0.007935 **
- x1    1    6.7594 9.5125  -6.8624 36.8276 2.15e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-30.91
y ~ x1 + x2 + x4

       Df Sum of Sq     RSS      AIC  F value    Pr(>F)
<none>               2.8581 -30.9112
- x4    1    1.2888  4.1470 -25.4671   7.2149  0.016228 *
- x2    1    2.0776  4.9358 -21.9845  11.6306  0.003581 **
- x1    1   20.6776 23.5357   9.2557 115.7539 9.835e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Το Backward Elimination ξεκινάει με το μοντέλο που περιλαμβάνει όλες τις υπό εξέταση μεταβλητές και βγαίνει η «χειρότερη», δηλαδή αυτή που συμβάλλει λιγότερο στον έλεγχο F, έστω την x_j. Αυτή η διαδικασία επαναλαμβάνεται με τις 5 επεξηγηματικές μεταβλητές και βρίσκουμε την «χειρότερη», έστω x_j (i≠j).

Με την εντολή

> summary(backward_model)

παίρνουμε συνοπτικά τα στοιχεία του μοντέλου που προσαρμόστηκε

```
Call:
lm(formula = y ~ x1 + x2 + x4, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.64524 -0.23678 -0.08526  0.14426  0.76835 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.4652     1.0767   2.290  0.03598 *  
x1            6.7532     0.6277  10.759 9.84e-09 ***
x2            3.0919     0.9066   3.410  0.00358 ** 
x4           -0.7144     0.2660  -2.686  0.01623 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4227 on 16 degrees of freedom
Multiple R-squared:  0.8994,	Adjusted R-squared:  0.8805 
F-statistic: 47.68 on 3 and 16 DF,  p-value: 3.342e-08
```

και παρατηρούμε πως είναι τα ίδια αποτελέσματα με τις forward.

g) Για την εκτέλεση του Stepwise δίνουμε την εντολή

> stepwise_model <- step(lm(y ~ 1), y~x1+x2+x3+x4+x5+x6, direction = "both", test='F')

Τα αποτελέσματα φαίνονται στην επόμενη σελίδα.

Στο stepwise αρχίζουμε από το $y=\beta_0$ και επιλέγουμε ποια μεταβλητή j θα προσθέσουμε στο μοντέλο για κάθε j με βάση τον έλεγχο F και επιλέγουμε έστω την x_i. Στη συνέχεια ελέγχουμε όλες τις j≠i για τα οποία θα προσθέσουμε με βάση τον έλεγχο F και ελέγχουμε αν θα χρειαστεί να αφαιρέσουμε τη x_i.

```
Start:  AIC=9.02
y ~ 1

       Df Sum of Sq     RSS      AIC F value    Pr(>F)    
+ x1    1   21.3032  7.1075 -16.6916 53.9509 8.09e-07 ***
+ x6    1   11.4102 17.0005   0.7502 12.0811 0.002698 ** 
+ x5    1   10.8754 17.5353   1.3697 11.1635 0.003635 ** 
+ x2    1    4.7937 23.6170   7.3247  3.6536 0.072004 .  
<none>              28.4107   9.0207                     
+ x4    1    0.4862 27.9245  10.6754  0.3134 0.582508    
+ x3    1    0.1634 28.2473  10.9053  0.1041 0.750681    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-16.69
y ~ x1

       Df Sum of Sq     RSS      AIC F value    Pr(>F)    
+ x2    1    2.9605  4.1470 -25.4671 12.1363 0.002843 ** 
+ x4    1    2.1717  4.9358 -21.9845  7.4800 0.014105 *  
+ x6    1    0.8500  6.2576 -17.2389  2.3091 0.146999    
<none>               7.1075 -16.6916                     
+ x3    1    0.3800  6.7275 -15.7905  0.9602 0.340865    
+ x5    1    0.0690  7.0385 -14.8866  0.1666 0.688276    
- x1    1   21.3032 28.4107   9.0207 53.9509 8.09e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-25.47
y ~ x1 + x2

       Df Sum of Sq     RSS      AIC F value    Pr(>F)    
+ x4    1    1.2888  2.8581 -30.9112  7.2149  0.016228 *  
<none>               4.1470 -25.4671                      
+ x5    1    0.1723  3.9747 -24.3159  0.6937  0.417169    
+ x3    1    0.1139  4.0330 -24.0243  0.4521  0.510951    
+ x6    1    0.0081  4.1389 -23.5061  0.0313  0.861870    
- x2    1    2.9605  7.1075 -16.6916 12.1363  0.002843 ** 
- x1    1   19.4700 23.6170   7.3247 79.8147 7.867e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-30.91
y ~ x1 + x2 + x4

       Df Sum of Sq     RSS      AIC  F value    Pr(>F)    
<none>               2.8581 -30.9112                       
+ x5    1    0.1050  2.7531 -29.6599   0.5722  0.461119    
+ x6    1    0.0171  2.8410 -29.0313   0.0904  0.767823    
+ x3    1    0.0026  2.8556 -28.9291   0.0134  0.909383    
- x4    1    1.2888  4.1470 -25.4671   7.2149  0.016228 *  
- x2    1    2.0776  4.9358 -21.9845  11.6306  0.003581 ** 
- x1    1   20.6776 23.5357   9.2557 115.7539 9.835e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Με την εντολή

> summary(stepwise_model)

παίρνουμε συνοπτικά τα στοιχεία του μοντέλου που προσαρμόστηκε και παρατηρούμε πως είναι ίδια με της forward και της backward.

Στο τελικό μοντέλο που προσαρμόστηκε από τις τρεις πιο πάνω διαδικασίες παρατηρούμε πως:

- Residual standard error= 0.4227 με 16 βαθμούς ελευθερίας, επομένως έχουμε χαμηλό σφάλμα.
- R^2=0.8994, επομένως το 89,94% της διακύμανσης του y εξηγείται από το μοντέλο.
- R^2-adjusted=0.8805, επομένως το προσαρμοσμένο R^2 είναι επίσης ψηλό και άρα δεν υποπροσαρμόζει τα δεδομένα.

- Το στατιστικό F=47,68 και έχει p-value=3.342×10^{-8}, γεγονός που μας δείχνει ότι το μοντέλο μας είναι στατιστικά σημαντικό.
- AIC=-30.91, το οποίο είναι το χαμηλότερο στη διαδικασία επιλογής, υποδεικνύοντας έτσι το βέλτιστο μοντέλο.

Επομένως, με βάση τις διαδικασίες που εκτελέστηκαν Σύμφωνα (Forward Selection, Backward Elimination και Stepwise Selection), το τελικό μοντέλο που προέκυψε είναι:

$y = 2.4652 + 6.7532x1 + 3.0919x2 - 0.7144x4$

Το μοντέλο αυτό είναι στατιστικά σημαντικό, έχει υψηλή προσαρμογή στα δεδομένα και αποτελεί την καλύτερη εκτίμηση για την μεταβλητή y.

Για το τελικό μοντέλο:

1) Για την κατασκευή του πίνακα ανάλυσης διασποράς δίνουμε τις εντολές
   > final_model <- lm(y ~ x1 + x2+ x4, data = data)
   > anova(final_model)

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
x1         1 21.3032 21.3032 119.2560 7.961e-09 ***
x2         1  2.9605  2.9605  16.5732  0.000889 ***
x4         1  1.2888  1.2888   7.2149  0.016228 *
Residuals 16  2.8581  0.1786
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

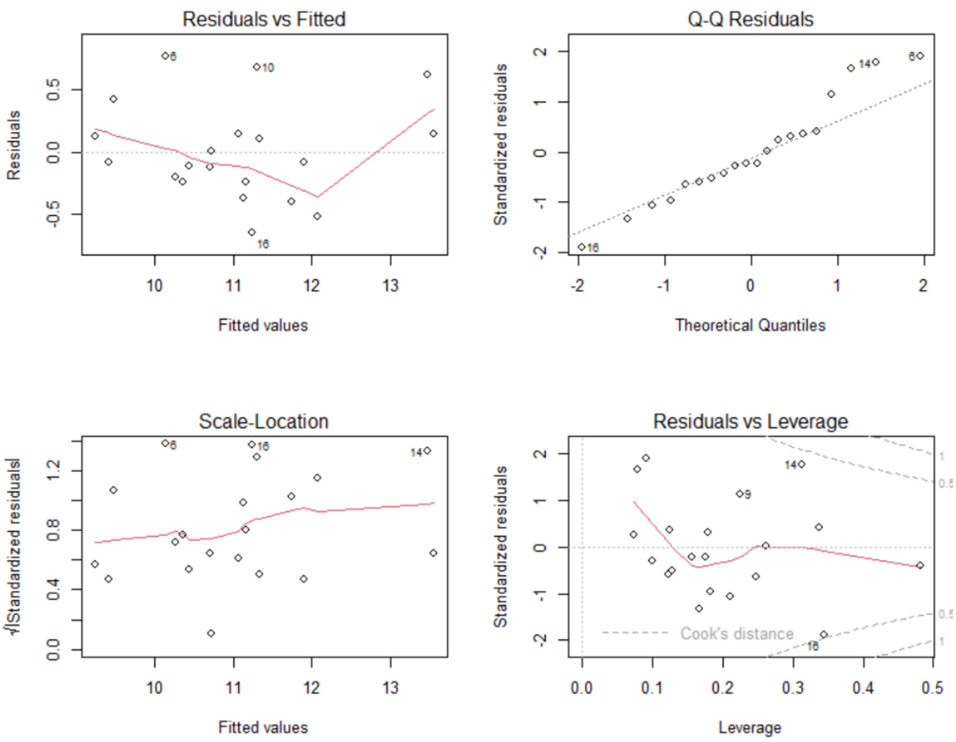
   Με βάση τον παραπάνω πίνακα ANOVA η συνολική παλινδρόμηση είναι στατιστικά σημαντική, αφού όλες οι μεταβλητές έχουν p-value<0.05 που είναι το επίπεδο σημαντικότητας.

2) Για τους ελέγχους H_0: $\beta_j=0$ με εναλλακτική H_1: $\beta_j\neq0$ μπορούμε να δούμε από τα πιο πάνω αποτελέσματα του F-test ότι για:
   j=1: p-value=7.961×10^{-9} < 0.05 που είναι το επίπεδο σημαντικότητας άρα απορρίπτω την H_0
   j=2: p-value=0.000889 < 0.05 που είναι το επίπεδο σημαντικότητας άρα απορρίπτω την H_0
   j=4: p-value=0.016228 < 0.05 που είναι το επίπεδο σημαντικότητας άρα απορρίπτω την H_0
   Από τον παραπάνω έλεγχο παρατηρούμε ότι οι μεταβλητές x_1, x_2, x_4 είναι στατιστικά σημαντικές με την μεγαλύτερη επίδραση στο y να την έχει η x_1 και την μικρότερη η x_4.
3) Για τον έλεγχο της πολυσυγγραμμικότητας δίνουμε την εντολή
   > vif(final_model)

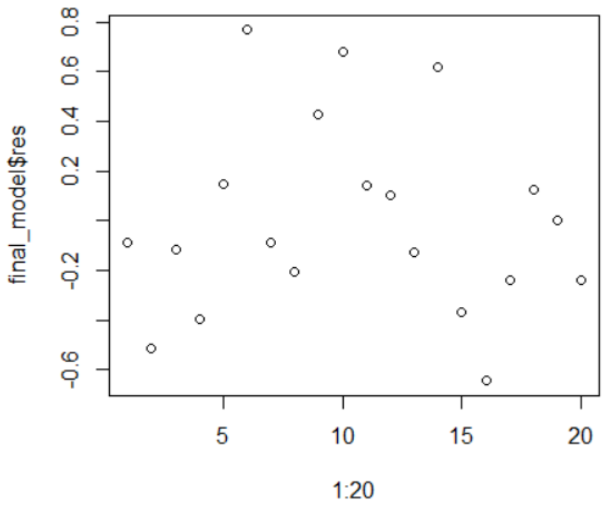
```
      x1       x2       x4
1.047921 1.058240 1.075785
```

   και παρατηρούμε ότι ο παράγοντας μεγέθυνσης διασποράς(VIF) για όλες τις μεταβλητές είναι <5, επομένως δεν υπάρχει πολυσυγγραμμικότητα.

4) Για την εξέταση των προϋποθέσεων του μοντέλου δίνουμε τις εντολές
   > par(mfrow = c(2, 2))
   > plot(final_model)



Στο διάγραμμα (1) βλέπουμε τα υπόλοιπα σε σχέση με τις προβλεπόμενες τιμές. Τα υπόλοιπα εδώ φαίνονται να είναι διασκορπισμένα τυχαία με μια ελαφριά καμπυλότητα, γεγονός που μπορεί να υποδηλώνει ότι το μοντέλο ενδέχεται να μην περιγράφει επαρκώς τη γραμμική σχέση μεταξύ των μεταβλητών. Επίσης, δεν υπάρχει σαφής τάση στην κατανομή των υπολοίπων γεγονός που δείχνει ότι ισχύει η ομοσκεδαστικότητα, ωστόσο υπάρχουν μερικές παρατηρήσεις που δείχνουν αύξηση της διασποράς στις υψηλές προβλεπόμενες τιμές, αυτό υποδεικνύει πιθανές αποκλίσεις από την υπόθεση της ομοσκεδαστικότητας.
Στο διάγραμμα (2) βλέπουμε την σύγκριση της κατανομής των υπολοίπων με την κανονική κατανομή. Οι περισσότερες τιμές βρίσκονται κοντά στη διαγώνιο γραμμή, υποδεικνύοντας ότι η υπόθεση της κανονικότητας ισχύει γενικά. Ωστόσο, υπάρχουν κάποιες αποκλίσεις στα άκρα που μπορεί να υποδεικνύουν ελαφρά απόκλιση.
Στο διάγραμμα (3) παρουσιάζεται η τετραγωνική ρίζα των τυποποιημένων υπολοίπων σε σχέση με τις προβλεπόμενες τιμές. Η κόκκινη γραμμή παραμένει σχετικά επίπεδη, κάτι που είναι ενδεικτικό ότι η υπόθεση της ομοσκεδαστικότητας γενικά ικανοποιείται, ωστόσο υπάρχουν μερικά σημεία που παρουσιάζουν μεγαλύτερη διασπορά υποδεικνύοντας πιθανή ανομοιογένεια.
Στο διάγραμμα (4) υπολοίπων έναντι μοχλού βλέπουμε αν υπάρχουν παρατηρήσεις που επηρεάζουν σημαντικά το μοντέλο (outliers ή σημεία υψηλής μόχλευσης). Παρατηρούμε πως ορισμένα σημεία με υψηλή μόχλευση(ειδικότερα οι παρατηρήσεις 14 και 16) πλησιάζουν την καμπύλη του Cook's Distance ή είναι πέρα από αυτήν υποδηλώνοντας ότι ενδέχεται να έχουν σημαντική επιρροή στις εκτιμήσεις του μοντέλου. Σε σύγκριση με το διάγραμμα του αρχικού μοντέλου εδώ παρατηρούμε ότι η καμπύλη του Cook's Distance είναι πιο κοντά στις παρατηρήσεις.
Για την εξέταση της ανεξαρτησίας των υπολοίπων δίνουμε την εντολή
> plot(1:20,final_model$res)
και από το πιο κάτω διάγραμμα παρατηρούμε πως τα υπόλοιπα δεν ακολουθούν κάποιο συγκεκριμένο μοτίβο επομένως ισχύει η υπόθεση της ανεξαρτησίας.



Για την εξέταση των σημείων επιρροής δίνουμε την εντολή
> cooks.distance(final_model)

```
           1            2            3            4            5            6 
2.256669e-03 8.868745e-02 2.285850e-03 7.434111e-02 4.959309e-03 9.005788e-02 
           7            8            9           10           11           12 
2.608793e-03 9.870703e-03 9.494447e-02 6.009532e-02 2.203886e-02 1.242524e-03 
          13           14           15           16           17           18 
3.939710e-02 3.585476e-01 5.208888e-02 4.682294e-01 1.250679e-02 5.595925e-03 
          19           20 
1.057381e-05 3.424539e-02 
```

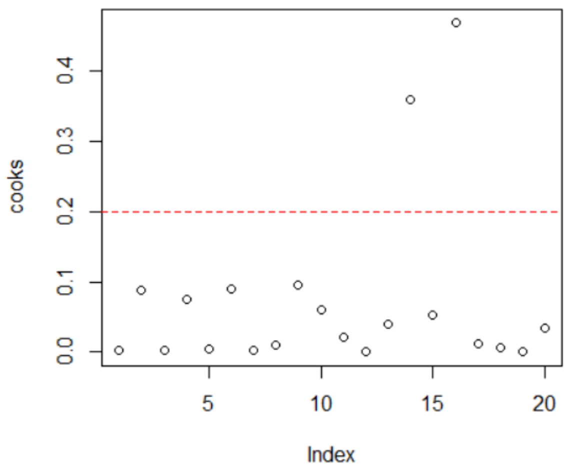
Από τα παραπάνω αποτελέσματα παρατηρούμε ότι οι παρατηρήσεις 14 και 16 εμφανίζουν σχετικά υψηλές τιμές, γεγονός που δείχνει ότι ενδέχεται να έχουν επιρροή στο μοντέλο.

Για περαιτέρω εξέταση των σημείων επιρροής δίνουμε τις εντολές

> cooks<-cooks.distance(final_model)

> plot(cooks)

> abline(h=4/length(cooks),col="red", lty=2)



Παρατηρούμε ότι οι παρατηρήσει 14 και 16 είναι πάνω από το όριο, επομένως θεωρούνται σημεία επιρροής.

Θεωρούμε ότι το καλύτερο μοντέλο είναι το $E(Y)=\beta_0+\beta_1X_1+\beta_2X_2$

1) Για την κατασκευή του 0.95-δ.ε. των συντελεστών β_1,β_2 δίνουμε την εντολή
   > confint(lm(y~x1+x2), level = 0.95)

```
                  2.5 %   97.5 %
(Intercept) -1.070743 3.862345
x1           4.916172 7.956054
x2           1.422646 5.791960
```

   και παρατηρούμε ότι το 95% διάστημα εμπιστοσύνης του β_1 είναι [4.916,7.956] και του β_2 είναι [1.423, 5.792]. Παρατηρούμε επίσης ότι και στα δύο διαστήματα δεν περιλαμβάνεται το 0, επομένως είναι οι μεταβλητές x_1 και x_2 είναι στατιστικά σημαντικές.
2) Για την δημιουργία ενός 0.99-δ.ε πρόβλεψης μιας παρατήρησης Y για ένα νέο $x_0=(1, x_1, x_2)'=(1, 0.9, 1.2)$ δίνουμε τις εντολές
   > new_x <- data.frame(x1 = 0.9, x2 = 1.2)
   > predict(lm(y~x1+x2), newdata = new_x, interval = "prediction", level = 0.99)

```
       fit      lwr      upr
1 11.51707 9.902408 13.13172
```

   Το διάστημα πρόβλεψης της παρατήρησης είναι [9.902, 13.132]. παρατηρούμε πως είναι ευρύτερο από το διάστημα εμπιστοσύνης των συντελεστών, γεγονός που οφείλεται στο ότι λαμβάνει υπόψη την αβεβαιότητα από την κατασκευή του μοντέλου και την τυχαία διακύμανση της εξαρτημένης μεταβλητής.

Άσκηση 3

Για την υλοποίηση της άσκησης 3 κωδικοποιούμε τα δεδομένα με τον πιο κάτω τρόπο

x	y	gender
2.4	3.3	1
2.1	5.3	1
0.5	1.4	1
1.8	4.7	1
2.1	6.6	1
1.5	3.0	1
1.3	5.9	1
0.3	1.8	2
1.0	4.6	2
1.3	3.0	2
2.5	8.1	2
2.5	8.0	2
1.2	3.3	2
1.8	7.5	2

και τα αποθηκεύουμε σε ένα αρχείο με όνομα “seira2_exercise3”.

Για την εισαγωγή των δεδομένων της άσκησης δίνουμε την εντολή

```
> data<-read.csv(file.choose(),header=TRUE,sep='')
```

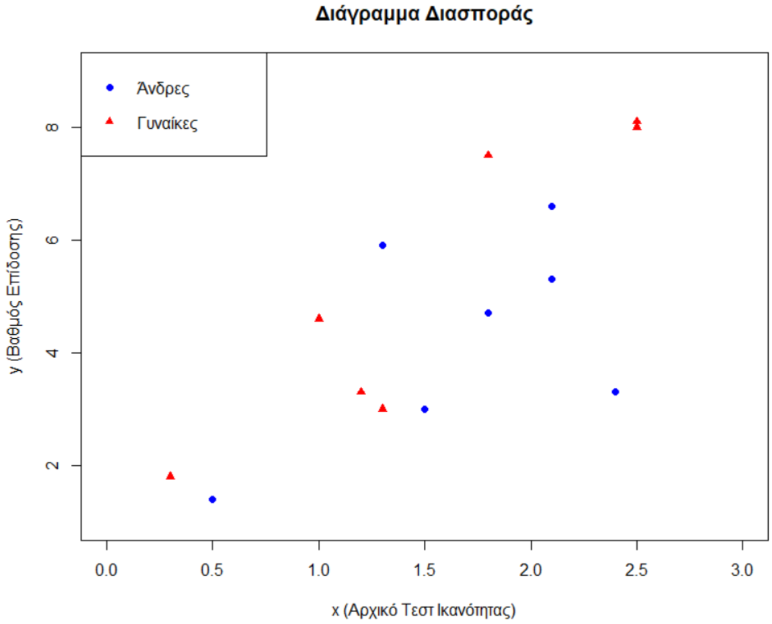
και επιλέγουμε το αρχείο “seira2_exercise3”.

Δημιουργούμε τις παρακάτω μεταβλητές

```
> x_men<-data$x[data$gender==1]
> y_men<-data$y[data$gender==1]
> x_women<-data$x[data$gender==2]
> y_women<-data$y[data$gender==2]
```

1) Για το διάγραμμα διασποράς δίνουμε τις εντολές

```
> plot(x_men, y_men, col = "blue", pch = 16, xlab = "x (Αρχικό Τεστ Ικανότητας)",
+     ylab = "y (Βαθμός Επίδοσης)", main = "Διάγραμμα Διασποράς", xlim =c(0,3), ylim =c(1,9))
> points(x_women, y_women, col = "red", pch = 17)
> legend("topleft", legend = c("Άνδρες", "Γυναίκες"), col = c("blue", "red"), pch = c(16, 17))
```



2) $y=\beta_0+\beta_1x+\beta_2z+\beta_3xz+\varepsilon$
z: δείκτης μεταβλητής(0 για άνδρες, 1 για γυναίκες)
xz: το γινόμενο x*z, περιλαμβάνει την αλληλεπίδραση
Για να ελέγξουμε τα ζητούμενα της άσκησης κάνουμε τους εξής ελέγχους:
(Α)Χρειαζόμαστε δύο διαφορετικές ευθείες: Ελέγχουμε αν οι παράμετροι β_2 (z) και β_3 (xz) είναι στατιστικά σημαντικές.
(Β): Δύο παράλληλες ευθείες: Ελέγχουμε αν β_3 (xz) είναι μηδέν
(Γ): Μία ευθεία: Ελέγχουμε αν β_2 (z) και β_3 (xz) είναι μηδέν.

3) Για τον παραπάνω έλεγχο δίνουμε τις εντολές

```
> z<-ifelse(data$gender=="1",1,0)
```

> x<-data$x

> mod<-lm(y~x+z+xz)

> summary(mod)

```
Call:
lm(formula = y ~ x + z + xz)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.13132 -1.02319  0.06584  0.80551  2.15518 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   0.6213     1.2299   0.505  0.62442   
x             3.0143     0.7284   4.138  0.00202 **
z             1.1304     2.0419   0.554  0.59200   
xz           -1.4811     1.1728  -1.263  0.23528   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.44 on 10 degrees of freedom
Multiple R-squared:  0.6794,    Adjusted R-squared:  0.5832 
F-statistic: 7.063 on 3 and 10 DF,  p-value: 0.007847
```

Από τα παραπάνω αποτελέσματα βλέπουμε ότι β_2=1.1304, β_3=-1.4811.

Παρατηρούμε ότι τα p-values των β_2, β_3 >0.05 επομένως αποδεχόμαστε την μηδενική υπόθεση H_0: $\beta_2=0$ και $\beta_3=0$ αντίστοιχα επομένως στα δεδομένα μπορούμε να προσαρμόσουμε μια ευθεία όπως αναφέρεται στο (Γ)

$y=\beta_0+\beta_1x+\varepsilon$

> mod2<-lm(y~x)

> summary(mod2)

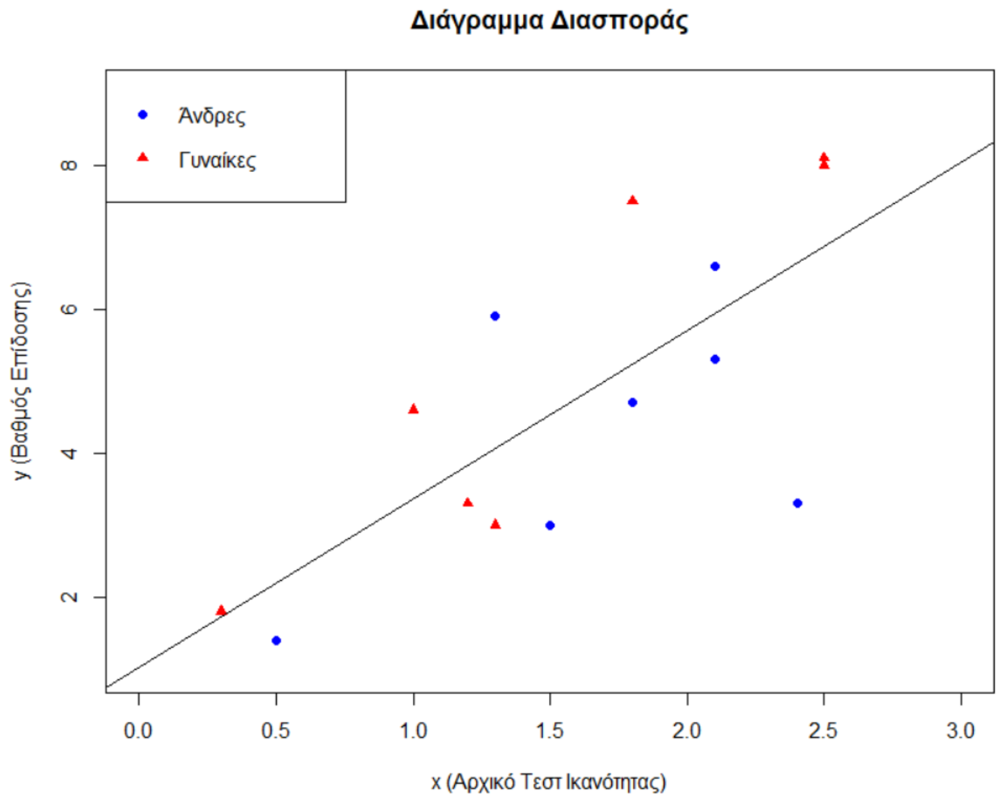
```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3354 -0.7565 -0.2312  1.2060  2.2661 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   1.0292     1.0687   0.963   0.3545   
x             2.3359     0.6172   3.784   0.0026 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.567 on 12 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5061 
F-statistic: 14.32 on 1 and 12 DF,  p-value: 0.002602
```

> abline(mod2)



Άσκηση 4

Για την υλοποίηση της άσκησης 4 κωδικοποιούμε τα δεδομένα με τον πιο κάτω τρόπο

y	A	B
143	1	1
145	1	1
133	1	1
149	1	2
139	1	2
142	1	2
146	1	3
133	1	3
193	1	3
99	2	1
112	2	1
117	2	1
76	2	2
85	2	2

84 2 2

117 2 3

105 2 3

110 2 3

και τα αποθηκεύουμε σε ένα αρχείο με όνομα “seira2_exercise4”.

Για την εισαγωγή των δεδομένων της άσκησης δίνουμε την εντολή

> data<-read.csv(file.choose(),header=TRUE,sep='')

και επιλέγουμε το αρχείο “seira2_exercise2”.

Δημιουργούμε την μεταβλητή y

> y=data$y

Και μετατρέπουμε τις μεταβλητές A και B από το αρχείο σε παράγοντες

> A=as.factor(data$A)

> B=as.factor(data$B)

Στη συνέχεια, προσαρμόζουμε το μοντέλο ανάλυσης διασποράς, συμπεριλαμβανομένου και τις αλληλεπίδρασης με τις εντολές:

> aov.out<-aov(y~A+B+A*B)

> summary(aov.out)

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
A            1   9707    9707  48.094 1.57e-05 ***
B            2   1397     698   3.460   0.0651 .  
A:B          2    705     353   1.748   0.2157    
Residuals   12   2422     202                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Από τον παραπάνω πίνακα ANOVA παρατηρούμε ότι ο παράγοντας A, δηλαδή η παρακολούθηση του σεμιναρίου έχει p-value = 1.57e-05, το οποίο είναι μικρότερο από το 0.05 και επομένως έχει στατιστικά σημαντική επίδραση στο κόστος του έργου. Επίσης, παρατηρούμε ότι ο παράγοντας B, δηλαδή τα χρόνια προϋπηρεσίας είναι οριακά μη σημαντικά αφού έχει p-value=0.0651>0.05. Τέλος, βλέπουμε πως το p-value της αλληλεπίδρασης είναι ίσο με 0.2157>0.05, γεγονός που μας δείχνει ότι δεν είναι στατιστικά σημαντική.

Για την δημιουργία του γραφήματος της αλληλεπίδρασης δίνουμε την εντολή:

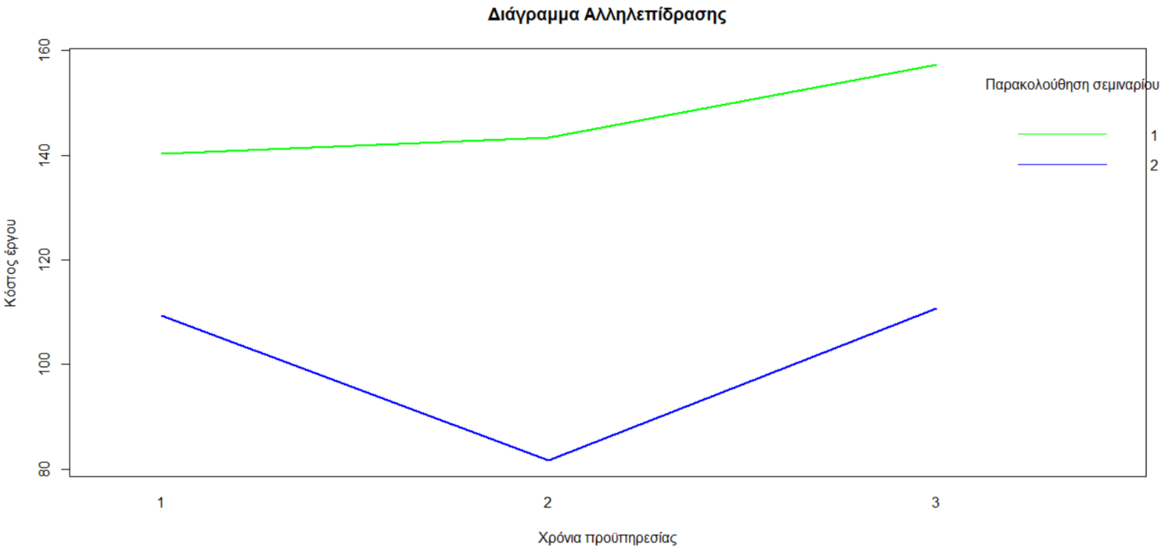
> interaction.plot(data$B, data$A, data$y,

+ xlab = "Χρόνια προϋπηρεσίας",

+ ylab = "Κόστος έργου",

+ trace.label = "Παρακολούθηση σεμιναρίου",

+ col=c("green","blue","red"), lty=1, lwd=2, main="Διάγραμμα Αλληλεπίδρασης")



Από το παραπάνω γράφημα παρατηρούμε πως οι γραμμές δεν είναι παράλληλες, επομένως υπάρχει κάποια αλληλεπίδραση αλλά η ANOVA έδειξε ότι δεν είναι στατιστικά σημαντική. Επίσης, βλέπουμε ότι η πράσινη γραμμή που είναι η παρακολούθηση σεμιναρίου έχει υψηλότερα κόστη σε σχέση με τη μπλε γραμμή που είναι η μη παρακολούθηση, επομένως η παρακολούθηση αυξάνει το κόστος του έργου ανεξαρτήτως προϋπηρεσίας.

Στη συνέχεια προσαρμόζουμε το μοντέλο γραμμικής παλινδρόμησης με τις εντολές:

> lm_model <- lm(y ~ A * B, data = data)

> summary(lm_model)

```
Call:
lm(formula = y ~ A * B, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-24.556  -9.125   0.556   8.208  37.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  160.778     33.698   4.771 0.000298 ***
A            -30.778     21.313  -1.444 0.170714
B             16.333     15.599   1.047 0.312814
A:B           -7.833      9.866  -0.794 0.440455
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.09 on 14 degrees of freedom
Multiple R-squared:  0.7127,	Adjusted R-squared:  0.6512
F-statistic: 11.58 on 3 and 14 DF,  p-value: 0.0004387
```

Από τα πιο πάνω αποτελέσματα παρατηρούμε πως η αναμενόμενη τιμή του κόστους όταν έχει γίνει η παρακολούθηση του σεμιναρίου και τα χρόνια προϋπηρεσίας είναι από 0 έως 5, είναι 160.778 μονάδες. Επιπλέον, ο παράγοντας A έχει συντελεστή -30.778 με p-value=0.170714, το οποίο δείχνει ότι δεν είναι στατιστικά σημαντικός στη γραμμική παλινδρόμηση όπως ήταν στην

ANOVA. Από αυτό έχουμε ότι το κόστος μειώνεται κατά 30.778 μονάδες όταν δεν έχει γίνει η παρακολούθηση του σεμιναρίου σε σχέση με όταν έχει γίνει και τα χρόνια προϋπηρεσίας είναι από 0 έως 5. Ο παράγοντας B έχει συντελεστή 16.333 και πάλι p-value>0.05, επομένως πάλι δεν είναι στατιστικά σημαντικός και μας δείχνει ότι το κόστος αυξάνεται κατά 16.333 μονάδες όταν τα χρόνια προϋπηρεσίας είναι μεγαλύτερα από 5 και έχει γίνει η παρακολούθηση σεμιναρίου. Το ίδιο και η αλληλεπίδραση πάλι δεν είναι στατιστικά σημαντική αφού p-value>0.05 και έχουμε ότι το κόστος μειώνεται κατά 7.833 μονάδες όταν δεν έχει γίνει παρακολούθηση σεμιναρίου και τα χρόνια προϋπηρεσίας είναι περισσότερα από 5. Επίσης, παρατηρούμε ότι ο R^2=0.7127, το οποίο εξηγεί το 71.27% της συνολικής διακύμανσης του κόστους.

Με τις εντολές:

> model_no_interaction <- aov(y~A+B, data = data)

> anova(model_no_interaction, aov.out)

παίρνουμε την ανάλυση σύγκρισης των μοντέλων με και χωρίς αλληλεπίδραση

```
Analysis of Variance Table

Model 1: y ~ A + B
Model 2: y ~ A + B + A * B
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
1     15 4272.1                              
2     12 2422.0  3    1850.1 3.0556 0.06966 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

και παρατηρούμε πως το p-value=0.06966, το οποίο είναι μη σημαντικό και αυτό επιβεβαιώνει ότι η προσθήκη του όρου αλληλεπίδρασης δεν βελτιώνει σημαντικά το μοντέλο.

Ένας άλλος τρόπος για να εκτελέσουμε όσα προαναφέρθηκαν είναι:

> data$A.f=factor(data$A)

> data$B.f=factor(data$B)

> a<-contrasts(data$A.f)

> b<-contrasts(data$B.f)

> contrasts(data$A.f)<-contr.sum(2)

> contrasts(data$B.f)<-contr.sum(3)

> summary(lm(y~data$A.f+data$B.f+data$A.f*data$B.f,data=data))

```
Call:
lm(formula = y ~ data$A.f + data$B.f + data$A.f * data$B.f, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.333  -5.667   0.833   4.333  35.667 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           123.778      3.349  36.964 9.86e-14 ***
data$A.f1              23.222      3.349   6.935 1.57e-05 ***
data$B.f1               1.056      4.736   0.223   0.8274    
data$B.f2             -11.278      4.736  -2.381   0.0347 *  
data$A.f1:data$B.f1    -7.722      4.736  -1.631   0.1289    
data$A.f1:data$B.f2     7.611      4.736   1.607   0.1340    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.21 on 12 degrees of freedom
Multiple R-squared:  0.8298,	Adjusted R-squared:  0.7589 
F-statistic:  11.7 on 5 and 12 DF,  p-value: 0.0002813
```

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι αυτό το μοντέλο με επίπεδα των κατηγορηματικών μεταβλητών(dummies variables) έχει R^2=0.8298, το οποίο εξηγεί το 82.98% της διακύμανσης στο κόστος. Το μέσο σφάλμα πρόβλεψης είναι 14.21 με 12 βαθμούς ελευθερίας και το p-value του μοντέλου είναι <0.001, γεγονός που δείχνει ότι το μοντέλο είναι στατιστικά σημαντικό. Η μέση τιμή του κόστους του έργου όταν έχει γίνει η παρακολούθηση του σεμιναρίου και τα χρόνια προϋπηρεσίας είναι από 0 έως 5, είναι 123.778 μονλαδες. Επίσης, παρατηρούμε ότι το κόστος αυξάνεται κατά 23.222 μονάδες όταν δεν έχει γίνει η παρακολούθηση του σεμιναρίου σε σύγκριση με όταν έχει γίνει. Επιπρόσθετα, το κόστος αυξάνεται κατά 1.056 μονάδες όταν τα χρόνια προϋπηρεσίας είναι από 5 έως 10, παρά όταν είναι από 0 έως 5 και μειώνεται κατά 11.278 μονάδες όταν τα χρόνια προϋπηρεσίας είναι από 10 έως 15, παρά όταν είναι από 0 έως 5. Η αλληλεπίδραση μεταξύ της μη παρακολούθησης σεμιναρίου και όταν τα χρόνια προϋπηρεσίας είναι από 5 έως 10 μειώνει το κόστος κατά 7.722 μονάδες σε σχέση με όταν τα χρόνια προϋπηρεσίας είναι από 0 έως 5 και έχει γίνει η παρακολούθηση σεμιναρίου. Η αλληλεπίδραση μεταξύ της μη παρακολούθησης σεμιναρίου και όταν τα χρόνια προϋπηρεσίας είναι από 10 έως 15 αυξάνει το κόστος κατά 7.611 μονάδες σε σχέση με όταν τα χρόνια προϋπηρεσίας είναι από 0 έως 5 και έχει γίνει η παρακολούθηση σεμιναρίου. Από τα p-values των παραγόντων συμπεραίνουμε ότι η επίδραση της παρακολούθησης σεμιναρίου είναι ο πιο σημαντικός παράγοντας που επηρεάζει το κόστος, τα χρόνια προϋπηρεσίας είναι λιγότερο σημαντικά, με το επίπεδο 2, δηλαδή όταν είναι από 10 έως 15, να έχει μια μικρή αλλά οριακά σημαντική επίδραση και οι αλληλεπιδράσεις μεταξύ προϋπηρεσίας και παρακολούθησης σεμιναρίου δεν είναι σημαντικές, υποδηλώνοντας ότι αυτοί οι δύο παράγοντες δεν αλληλοεπιδρούν στο κόστος.