ΣΤΑΤΙΣΤΙΚΗ ΣΤΗ ΠΛΗΡΟΦΟΡΙΚΗ

ΕΡΓΑΣΙΑ 4

ΜΕΛΗ ΟΜΑΔΑΣ:

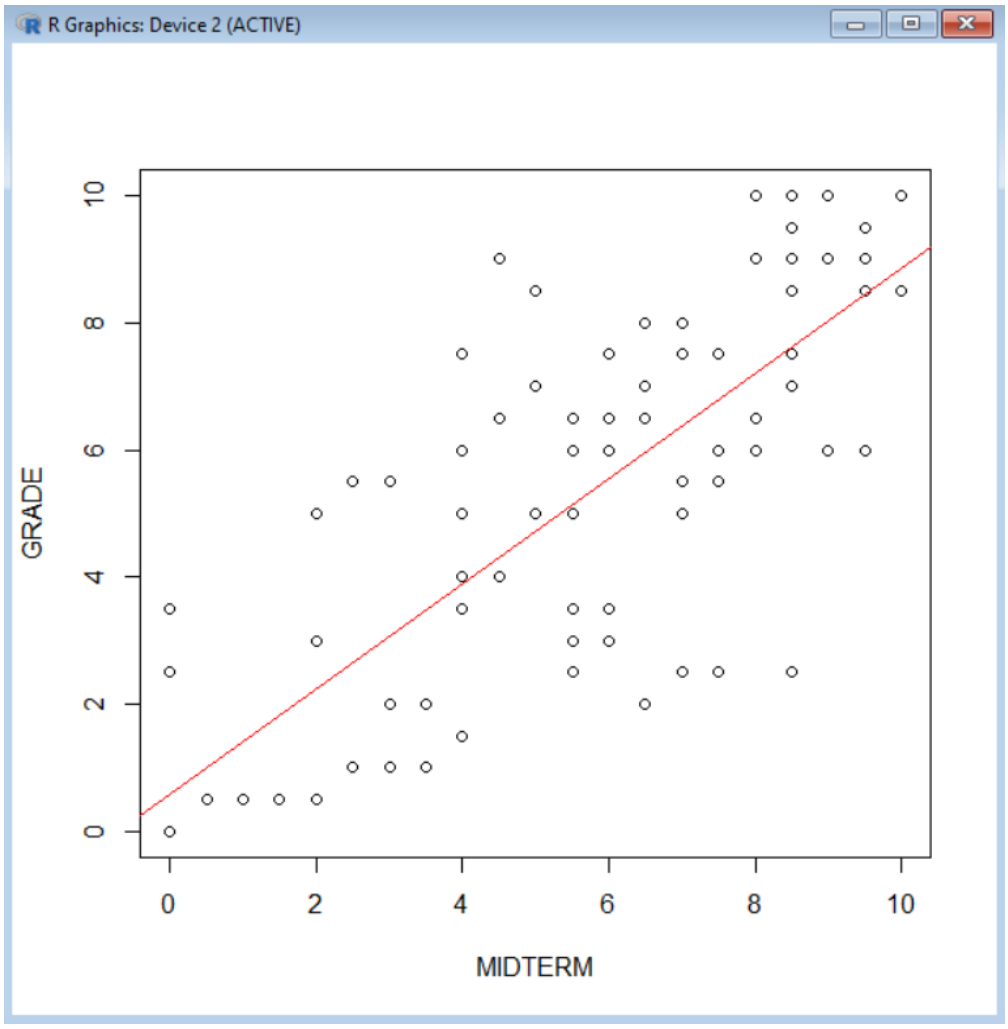
ΑΘΑΝΑΣΙΟΣ ΚΛΕΤΤΑΣ: 3180079

ΙΑΣΩΝ ΖΙΩΓΑΣ: 3180057

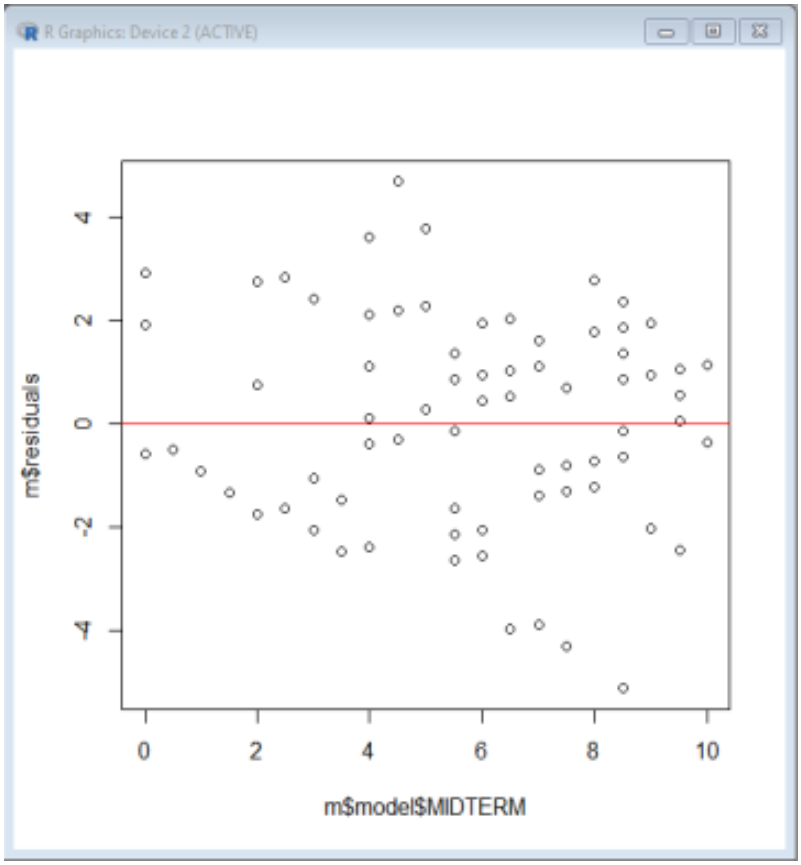
ΑΣΚΗΣΗ 1:

a)

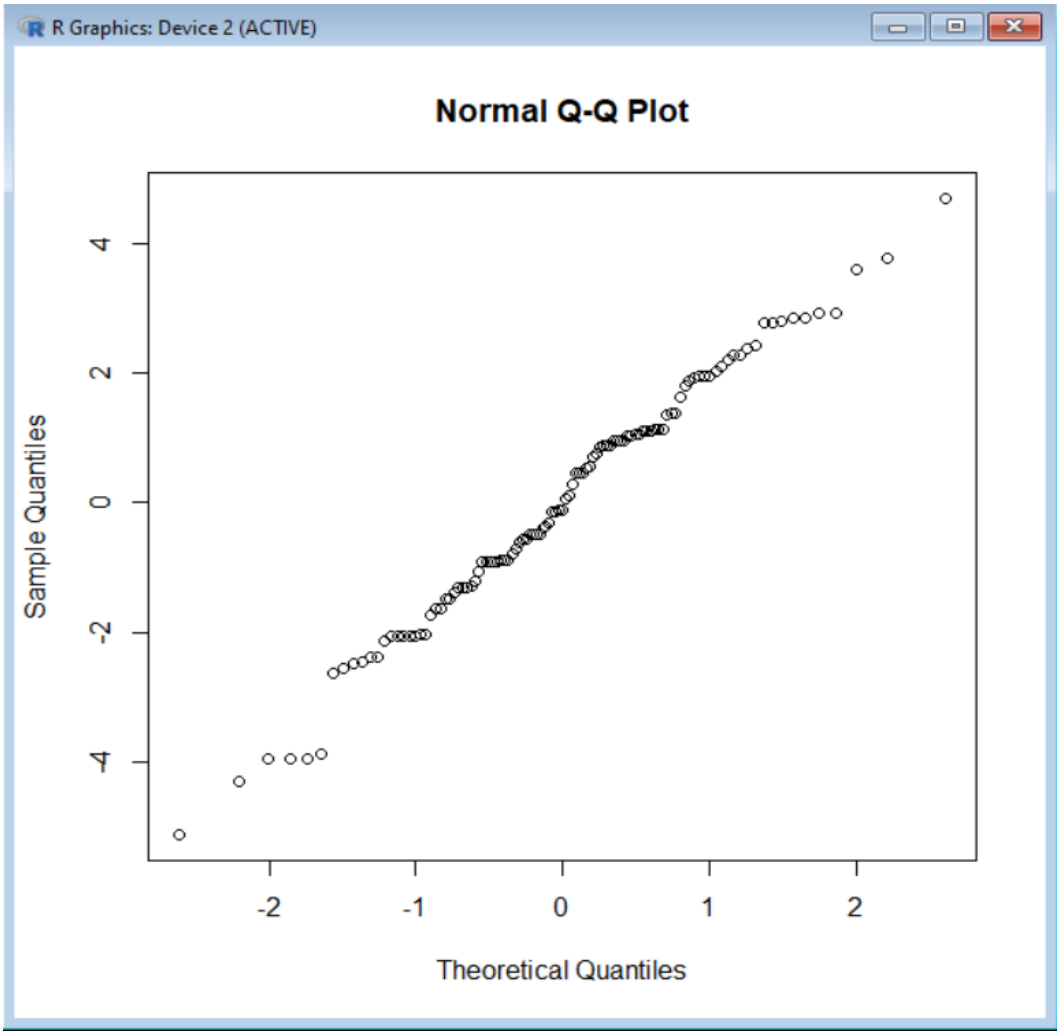
Παρατηρούμε ότι ο η σχέση μεταξύ των δύο μεταβλητών φαίνεται να είναι γραμμική και αυτό το διαπιστώνουμε από το παρακάτω scatterplot.



Επίσης από το παρακάτω σχεδιάγραμμα καταλαβαίνουμε ότι ισχύει η ομοσκεδαστικότητα.



Και τέλος ικανοποιείται και η κανονικότητα εφόσον τα δεδομένα προσεγγίζουν την κανονική κατανομή όπως φαίνεται στο παρακάτω normal quantile plot.



b)

Το 95% διάστημα εμπιστοσύνης για τον β_1 το υπολογίσαμε χρησιμοποιώντας την συνάρτηση confint και είναι το:

[0.7035840, 0.9544545]

```
> confint(m)
                2.5 %     97.5 %
(Intercept) -0.1862308 1.3374309
MIDTERM      0.7035840 0.9544545
```

c)

Έλεγχος σημαντικότητας:

- H_0: $\beta_1 = 0$
- H_a: $\beta_1 \neq 0$

Υπολογίζουμε το p-value όπως φαίνεται παρακάτω και βρίσκουμε ότι ισούνται με 4.042778e-24. Επομένως εφόσον το p-value είναι τόσο μικρό μπορούμε να απορρίψουμε την μηδενική μας υπόθεση η οποία είναι ότι β1 = 0. Άρα υπάρχει σχέση μεταξύ των δύο μεταβλητών.

```
> summary(m)$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 0.5756001 0.38438110  1.497472 1.371607e-01
MIDTERM     0.8290192 0.06328825 13.099102 4.042767e-24
> t = 0.8290192 /0.06328825
> 2*pt(-abs(t), df = 109)
[1] 4.042778e-24
```

d)

Με την βοήθεια της συνάρτησης predict υπολογίσαμε ότι το 95% διάστημα εμπιστοσύνης που θα επιτύγχαναν φοιτητές οι οποίοι στην εξέταση προόδου έλαβαν 7 είναι το: [5.960928, 6.796541].

```
> predict(m,newdata = data.frame(MIDTERM = 7),interval = "confidence")
       fit      lwr      upr
1 6.378735 5.960928 6.796541
```

e)

Προβλέπουμε τον τελικό που θα επετύγχανε ένας τυχαία επιλεγμένος φοιτητής που πήρε 7 στην πρόοδο, δίνοντας ένα 95% διάστημα με την βοήθεια της συνάρτησης predict και παίρνουμε το εξής διάστημα:

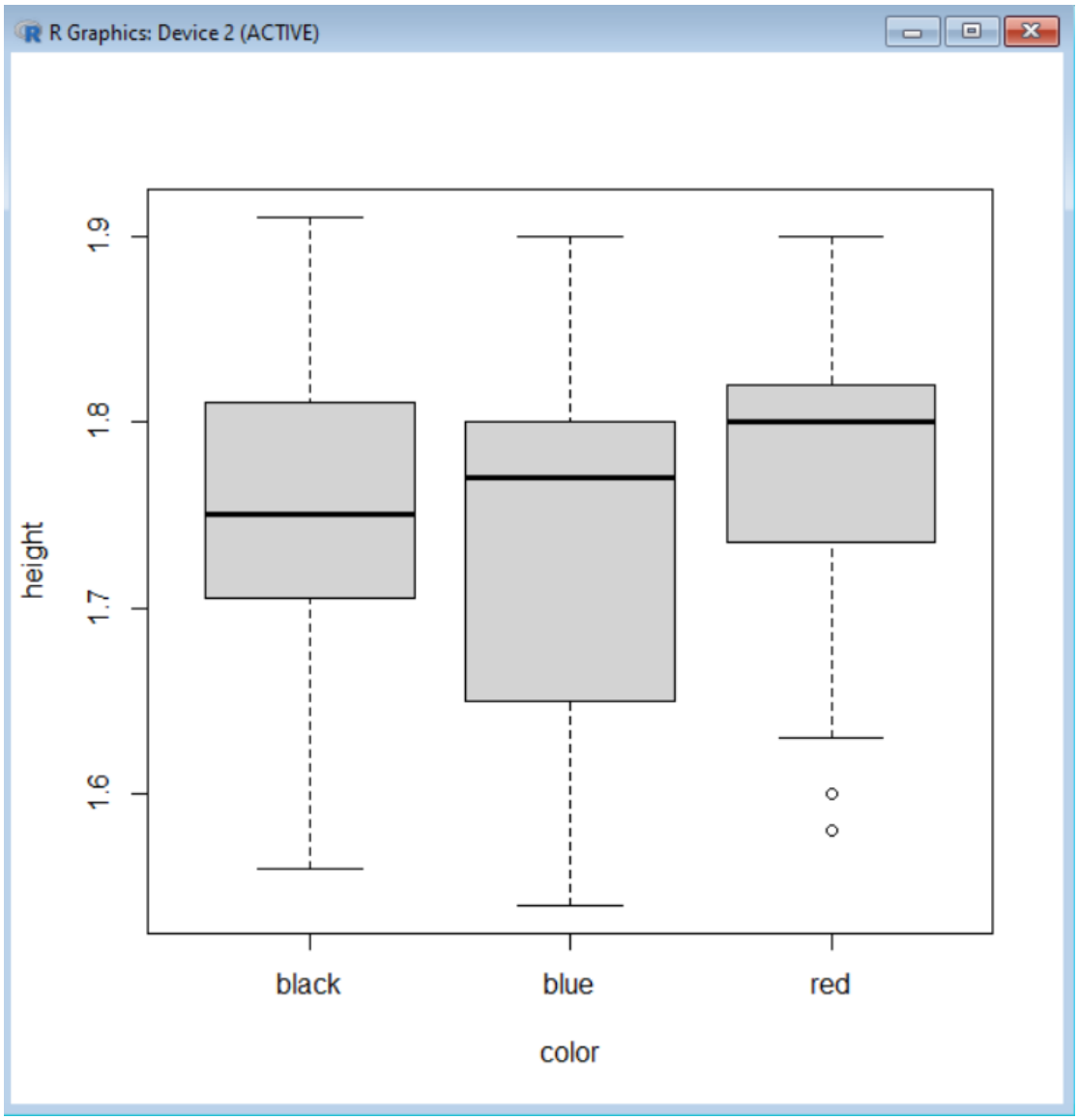
[2.537905, 10.21956]

```
> predict(m,newdata = data.frame(MIDTERM = 7),interval = "prediction")
       fit      lwr      upr
1 6.378735 2.537905 10.21956
```

ΑΣΚΗΣΗ 2:

a)

πλάι-πλάι boxplots για τα 3 περισσότερο δημοφιλή χρώματα:



Παρατηρούμε ότι οι διαφορές των δειγματικών διαμέσων τιμών είναι μικρή σε σχέση με τη διακύμανση των τιμών εσωτερικά σε κάθε πληθυσμό, γεγονός που μας προϊδεάζει ότι δεν υπάρχει σχέση μεταξύ του ύψους και του χρώματος.

```
> with(na.omit(data[,c("height","color")]),tapply(height,color,mean))
   black     blue      red
1.757143 1.735652 1.766842
> with(na.omit(data[,c("height","color")]),tapply(height,color,sd))
     black       blue        red
0.07901510 0.09476411 0.08768964
> with(na.omit(data[,c("height","color")]),tapply(height,color,length))
black  blue   red
   28    23    19
```

b)

```
> with(na.omit(data[,c("height","color")]),aov(height~color))-> var
> anova(var)
Analysis of Variance Table

Response: height
           Df  Sum Sq   Mean Sq F value Pr(>F)
color       2 0.01104 0.0055186  0.7328 0.4844
Residuals  67 0.50455 0.0075306
```

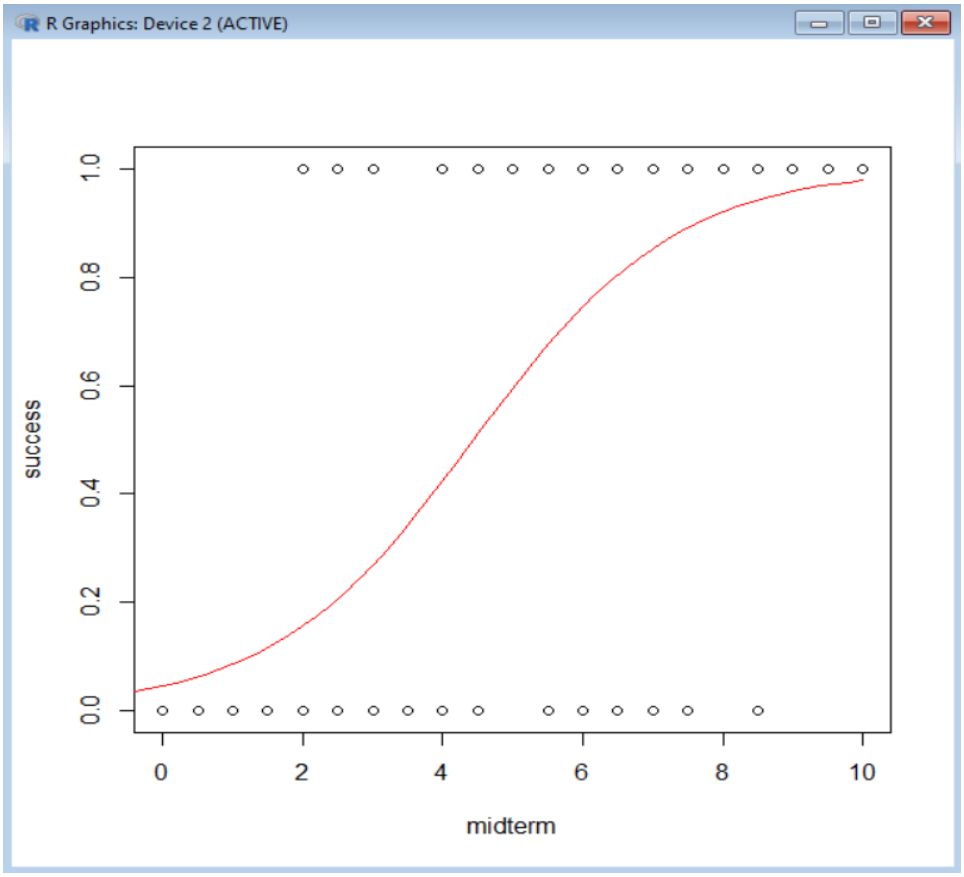
με $H_0: \mu_1=\mu_2=\mu_3$ | Ηα : $\mu_i \neq \mu_j$ για κάποιο i,j:

παρατηρούμε πως το p-value = 0.4844 γεγονός που δεν μας επιτρέπει να απορρίψουμε την H_0. Άρα το ύψος δεν σχετίζεται με το χρώμα

ΑΣΚΗΣΗ 3:

a)

Θέλουμε να εξετάσουμε σχέση μεταξύ βαθμού προόδου και επιτυχίας χρησιμοποιώντας λογιστική παλινδρόμηση. Για να το κάνουμε αυτό χρειαζόμαστε μια ποσοτική μεταβλητή ως επεξηγηματική και μια κατηγορική ως μεταβλητή απόκρισης. Στην περίπτωση μας η επεξηγηματική είναι ο βαθμός της προόδου και η μεταβλητή απόκρισης είναι η επιτυχία η οποία παίρνει 2 τιμές ανάλογα αν κάποιος έχει προβιβάσιμο βαθμό ή όχι.



Παρατηρούμε παραπάνω μια αύξουσα σχέση μεταξύ βαθμού εξαμήνου και επιτυχία στις εξετάσεις επομένως μπορούμε να εφαρμόσουμε λογιστική παλινδρόμηση.

b)

Για να βρούμε το ποσοστό επιτυχίας των φοιτητών όταν παίρνουν βαθμό 5 στην πρόοδο βάσει του υποδείγματος χρησιμοποιήσαμε την συνάρτηση predict και βρήκαμε ότι είναι 0.5943144.

```
> predict(m,newdata=data.frame(midterm=5),type='response')
        1
0.5943144
```

c)

Έλεγχος σημαντικότητας:

- H_0: $\beta_1 = 0$
- H_a: $\beta_1 \neq 0$

```
> summary(m)

Call:
glm(formula = success ~ midterm, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3895  -0.4214   0.2457   0.5632   1.9271

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.0665     0.5789  -5.297 1.17e-07 ***
midterm       0.6897     0.1117   6.173 6.69e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.673  on 126  degrees of freedom
Residual deviance:  98.266  on 125  degrees of freedom
AIC: 102.27

Number of Fisher Scoring iterations: 5
```

Παρατηρούμε ότι το p-value = 6.69e-10. Το p-value αυτό είναι πολύ μικρό επομένως μπορούμε να απορρίψουμε την μηδενική μας υπόθεση. Άρα σχετίζεται ο βαθμός προόδου με την επιτυχία.

d)

Σε αυτήν την ερώτηση θα πρέπει να απαντήσουμε με ένα ναι ή όχι. Εάν αντικαταστήσουμε στην συνάρτηση p(x) το χ = 5 και το p(χ=5) > 0.5 τότε ναι προβλέπουμε πως ο φοιτητής θα περάσει το μάθημα αλλιώς δεν θα το περάσει:

```
> 1/(1+exp(-(b1*5+b0)))
  MIDTERM
0.5943144
```

Αφού το p(χ=5)>0.5 προβλέπουμε ότι ο φοιτητής θα περάσει το μάθημα.

ΑΣΚΗΣΗ 4:

Από την θεωρία γνωρίζουμε πως ο εκτιμητής μέγιστης πιθανοφάνειας για παραμετρικό χώρο Θ=[0,1] είναι ίσος με :

$$\hat{\theta}_{MLE} = \frac{X}{n}$$

Όπου X το πλήθος των “1” στο δείγμα(σε αυτό το παράδειγμα οι κορώνες) και n το πλήθος του δείγματος.

Άρα ο εκτιμητής μέγιστης πιθανοφάνειας είναι ίσος με : 44/100.