4η Άσκηση Στατιστική

ΓΕΩΡΓΙΟΥ ΑΛΕΞΙΟΣ ΛΑΖΑΡΟΣ 3180027

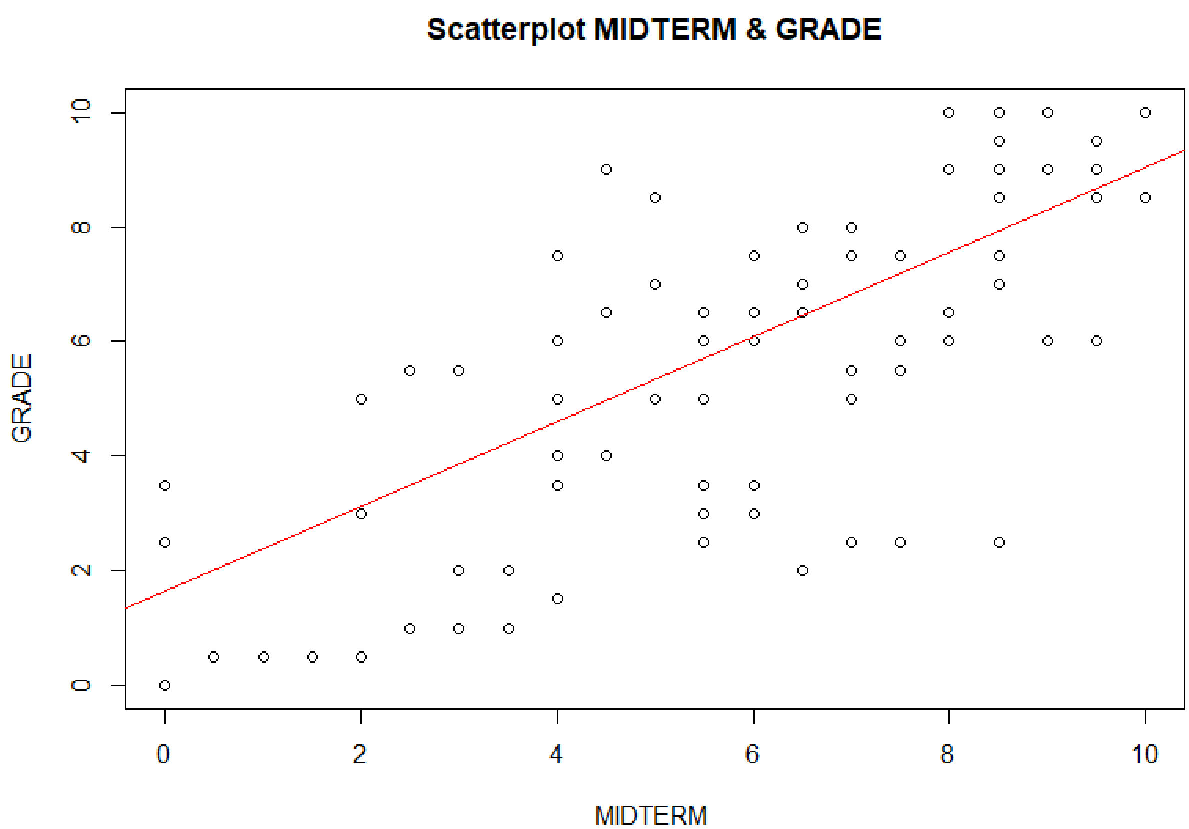
ΒΑΣΙΛΕΙΟΣ-ΕΚΤΩΡ ΚΩΤΣΗΣ-ΠΑΝΑΚΑΚΗΣ 3180094

1)

a.

Για την γραμμικότητα των δυο μεταβλητών MIDTERM, FINAL φτιάχνουμε scatterplot με γραμμική παλινδρόμηση.

```
> A1data <- grades_2014_data
> attach(A1data)
> plot(MIDTERM, GRADE, main="Scatterplot MIDTERM & GRADE", xlab="MIDTERM ",
ylab="GRADE")
> abline(lm(MIDTERM~GRADE), col="red")
```

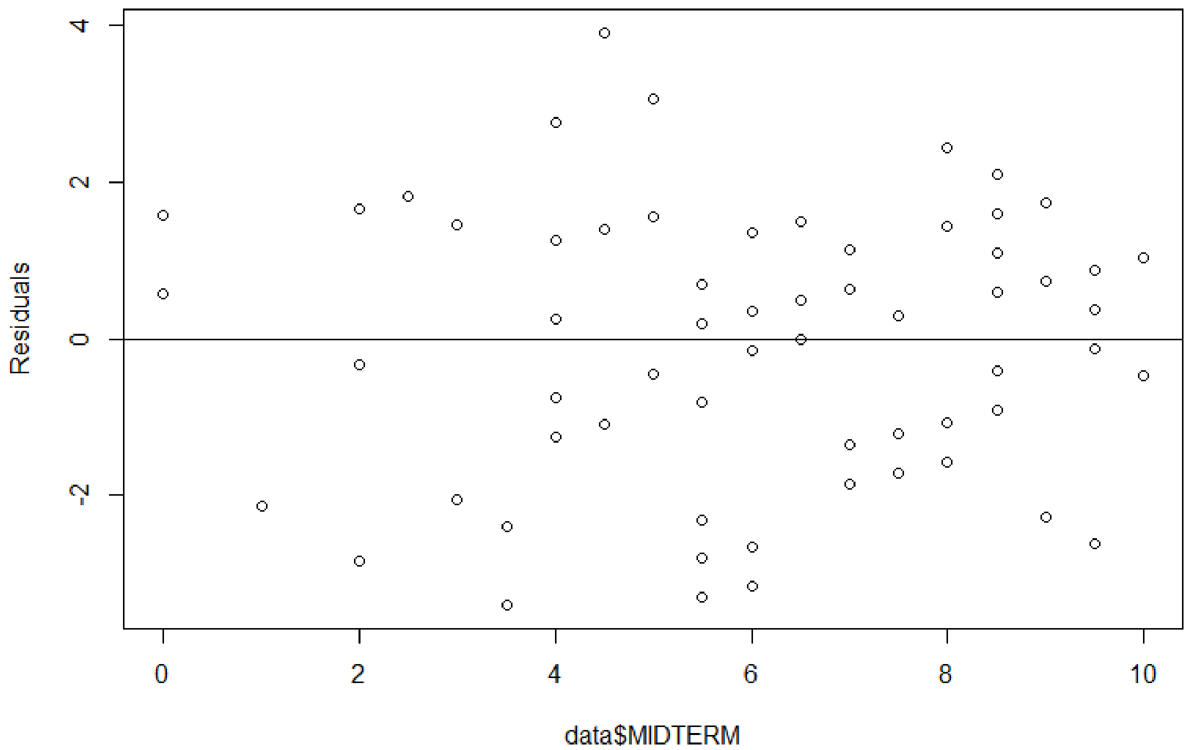


Φαίνεται η σχέση των δυο μεταβλητών να είναι γραμμική αφού τα στοιχεία δεν απέχουν πολύ από την κόκκινη γραμμή γραμμικής παλινδρόμησης ελαχίστων τετραγώνων και σχέση αύξουσα λόγω της κλίσης της.

```
> data <- A1data[complete.cases(A1data), ]
> scores.lm = lm(data$GRADE ~ data$MIDTERM)
> scores.res = resid(scores.lm)
```

```
> plot(data$MIDTERM, scores.res, ylab = "Residuals")
> abline(h  = 0)
```

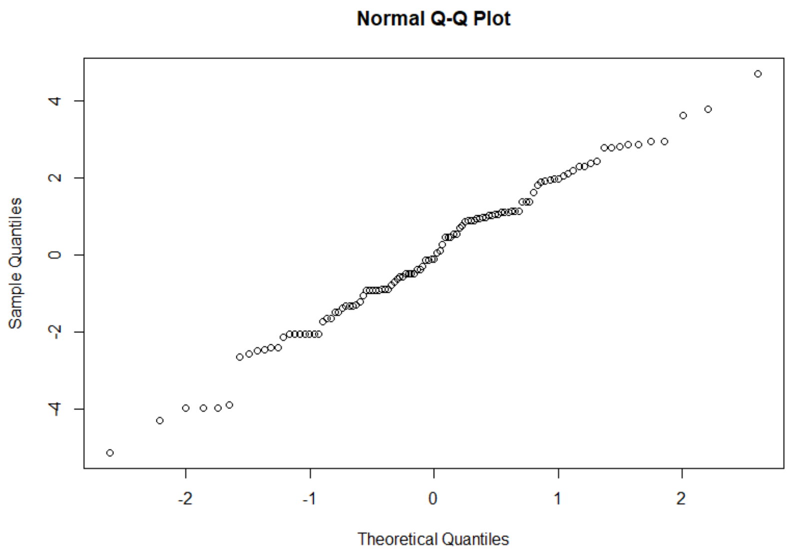
Αφαιρέσαμε τις περιπτώσεις με τα NAs.



Κάνουμε plot τα residuals (απόσταση από την κόκκινη γραμμή) των δεδομένων από το διάγραμμα της γραμμικής παλινδρόμησης με τον βαθμό MIDTERM.

Φαίνεται ότι δεν απέχουν αρκετά από την y = 0 οπότε θα υποθέσουμε ότι ικανοποιείται η ομοσκεδαστικότητα.

```
> qqnorm(lm(MIDTERM~GRADE)$residuals)
```



Τα σημεία στο Q-Q Plot είναι αρκετά συγγραμικά οπότε οι δυο μεταβλητές κατανέμονται κανονικά.

b.

```
> m <- lm(GRADE ~ MIDTERM)
> b0 <- m$coefficients[1]
> b1 <- m$coefficients[2]
> SEB0 <- summary(m)$coefficients[1,2]
> SEB1 <- summary(m)$coefficients[2,2]
> b1
  MIDTERM
0.8290192
> t <- -qt(0.025, df = 109)
> ci <- b1 + c(-1,1) * t * SEB1
> ci
[1] 0.7035840 0.9544545
```

Διάστημα εμπιστοσύνης b1 95% = [0.7035840, 0.9544545]

c.

Για να υπάρχει μια σχέση των GRADE και MIDTERM πρέπει το b1 να μην είναι 0, αφού

(Grade = b1 * Midterm + ..)

Παίρνουμε τις υποθέσεις

H0: b1 = 0

H1: b1 != 0

```
summary(m)

Call:
lm(formula = GRADE ~ MIDTERM)

Residuals:
```

```
    Min      1Q  Median      3Q     Max 
-5.1223 -1.3191 -0.1223  1.1342  4.6938 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.57560    0.38438   1.497    0.137    
MIDTERM      0.82902    0.06329  13.099   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.926 on 109 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.6115,	Adjusted R-squared:  0.608 
F-statistic: 171.6 on 1 and 109 DF,  p-value: < 2.2e-16
```

Το pvalue είναι πολύ μικρό άρα δεχόμαστε την εναλλακτική υπόθεση. Δηλαδή το b1 είναι διάφορο το 0 και τελικά υπάρχει σχέση μεταξύ των δύο μεταβλητών.

d.

Κάνουμε εκτίμηση του τελικού βαθμού με πρόοδο εφτά χρησιμοποιώντας τον τύπο γραμμικής συσχέτισης.

```
> newGrade <- b1 * 7 + b0
> newGrade
 MIDTERM 
6.378735 
```

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

Με διάστημα εμπιστοσύνης [5.960928, 6.796541]

e.

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

Με διάστημα εμπιστοσύνης [2.537905, 10.21956]

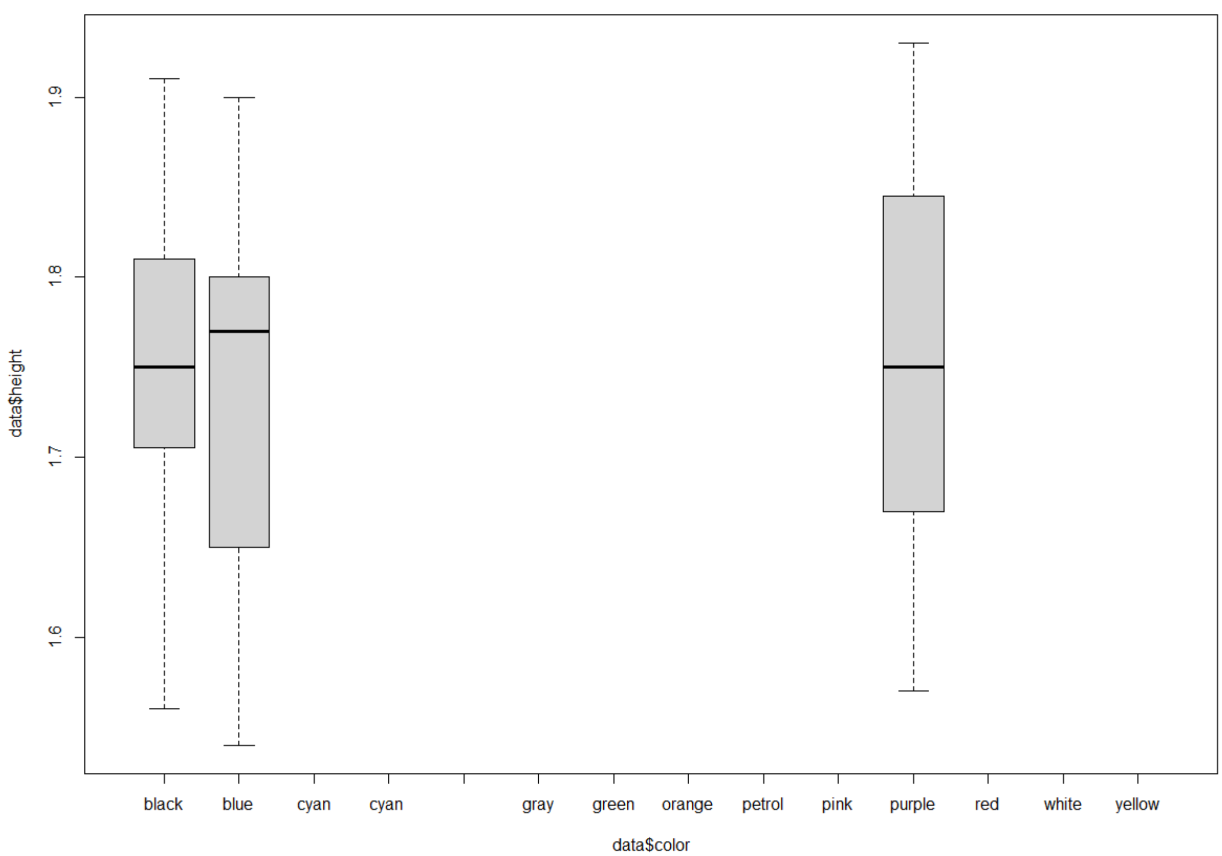
2)

a. Παρατηρούμε ότι τα τρία δημοφιλέστερα χρώματα είναι το μαύρο, μπλε και μοβ.

```
> summary(data)
             color
black           :28
blue            :24
purple          :19
cyan            : 0
cyan            : 0
cyangreen/petrol: 0
(Other)         : 0
```

πλάϊ – πλαϊ boxplots

```
A2data$height[94] = 1.76 #outlier replacement
> plot(A2data$height ~ A2data$color)
> plot(data$height ~ data$color)
```



Δεν φαίνεται να υπάρχει συσχέτιση του χρώματος με το ύψος, τα boxplots είναι αρκετά όμοια μεταξύ τους, οι φοιτητής που επέλεξαν το μαύρο έχουν μικρότερο Q1-Q3 range ύψους, πράγμα που μπορεί να οφείλεται στο γεγονός ότι από τους φοιτητές που επέλεξαν μαύρο η πλειοψηφία είναι Male. Αλλά σε γενικές γραμμές η σύνοψη των πέντε αριθμών δεν έχει μεγάλες διαφορές για τα τρία χρώματα.

*Δεν λάβαμε υπόψη μας το κόκκινο αν και ήταν στην 3η θέση με ισοβαθμία.

b.

Θα εφαρμόσουμε F έλεγχο σημαντικότητας, τα δεδομένα μας είναι επαρκή σε αριθμό τυχαία και ικανοποιούν σε αρκετά μεγάλο βαθμό το κριτήριο της ομοσκεδαστικότητας.

H0: Οι μέσοι όροι των υψών για τα τρία χρώματα είναι ίδιοι

H1: Οι μέσοι όροι των υψών ανά χρώμα διαφέρουν για κάποια χρώματα

```
> summary(res.aov)
            Df Sum Sq  Mean Sq F value Pr(>F)
data$color   2 0.0059 0.002935   0.342  0.712
Residuals   67 0.5754 0.008588
1 observation deleted due to missingness
```

Το pvalue του ελέγχου είναι αρκετά μεγάλο ~71% όποτε δεχόμαστε την μηδενική υπόθεση ότι το ύψος ενός φοιτητή δεν εξαρτάται από την επιλογή χρώματος.

3)

a.

Γενικά παρατηρούμε μια αύξουσα σχέση μεταξύ του βαθμού της προόδου και το αν ο τελικός βαθμός του φοιτητή είναι προβιβάσιμος, το οποίο είναι λογικό αν σκεφτούμε ότι ο βαθμός της προόδου είναι μέρος του τελικού βαθμού. Τα δεδομένα μας είναι πολύ εύκολο να τα διαχωρίσουμε για συμπερασματολογία μέσω λογιστικής παλινδρόμησης με επιτυχία τελικό βαθμό μεγαλύτερου ή ίσου του 5.

Δημιουργούμε τον πίνακα με τις επιτυχίες, φτιάχνουμε το μοντέλο της λογιστικής παλινδρόμησης με βάση τις επιτυχίες και στην συνέχεια δημιουργούμε το plot με βάση την πρόβλεψη του y από το μοντέλο σε συγκεκριμένα διαστήματα του x.

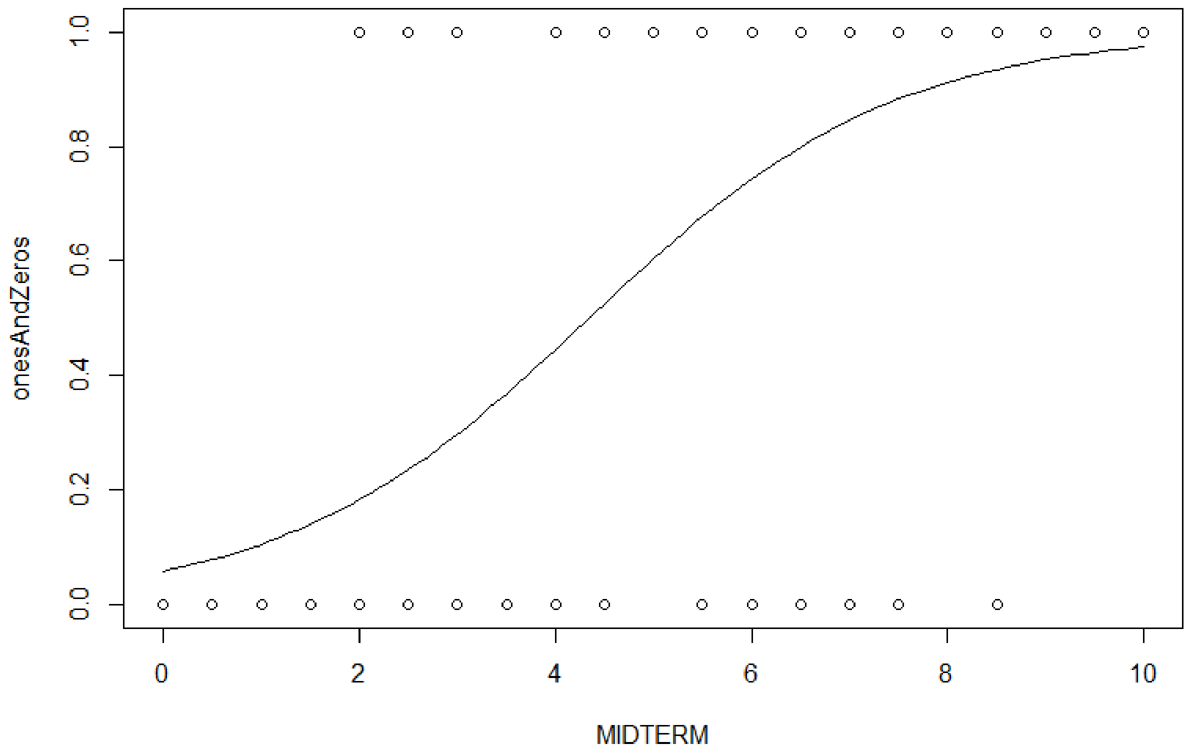
```
> onesAndZeros <- ifelse(GRADE >= 5, 1, 0)
> onesAndZeros
  [1] 1 0 0 1 0 1 0 0 0 0 1 1 1 1 1 1 0 1 1 1 1 0 1 0 0 1 1 1 0 1 0 1 0 1 0 1
0 0 1 1 0 0 0 1 0 1 1 1 0 0 0 1 1 0 1 0 1 0 1 1 0 0
 [63] 1 1 0 0 1 0 0 0 1 1 1 1 0 0 1 0 1 1 0 1 0 0 1 1 1 1 1 0 0 0 0 1 0 0 1 0
1 1 0 0 1 1 1 1 1 1 1 0 1 1 0 0 1 0 1 0 1 1 0 1 0 0
[125] 0 0 0
> plot(MIDTERM,onesAndZeros)
> model <- glm(onesAndZeros~MIDTERM, family = binomial("logit"))
> model

Call:  glm(formula = onesAndZeros ~ MIDTERM, family = binomial("logit"))

Coefficients:
(Intercept)      MIDTERM
    -2.7771       0.6397

Degrees of Freedom: 110 Total (i.e. Null);  109 Residual
  (16 observations deleted due to missingness)
Null Deviance:      149.1
Residual Deviance: 96.6     AIC: 100.6

#Creating 101 numbers 0.0, 0.1 ... 10.0
#Calculating prediction of y in the logistic regression line and plotting it
> x <- seq(from = 0, to = 10, by = 0.1)
> y <- predict(model, newdata = data.frame(MIDTERM = x), type = "response")
> lines(x,y)
```



b.

```
> y
         1          2          3          4          5          6          7
8          9         10         11
0.05857234 0.06220100 0.06603869 0.07009546 0.07438159 0.07890757 0.08368404
0.08872180 0.09403171 0.09962467 0.10551155
        12         13         14         15         16         17         18
19         20         21         22
0.11170313 0.11821003 0.12504262 0.13221091 0.13972452 0.14759250 0.15582329
0.16442454 0.17340305 0.18276460 0.19251384
        23         24         25         26         27         28         29
30         31         32         33
0.20265417 0.21318762 0.22411467 0.23543421 0.24714337 0.25923741 0.27170969
0.28455153 0.29775217 0.31129874 0.32517625
        34         35         36         37         38         39         40
41         42         43         44
0.33936756 0.35385344 0.36861264 0.38362196 0.39885637 0.41428912 0.42989198
0.44563538 0.46148863 0.47742015 0.49339775
        45         46         47         48         49         50         51
52         53         54         55
0.50938885 0.52536075 0.54128094 0.55711729 0.57283837 0.58841368 0.60381386
0.61901091 0.63397839 0.64869157 0.66312759
```

```
        56          57          58          59          60          61          62
63          64          65          66
0.67726553 0.69108653 0.70457383 0.71771280 0.73049094 0.74289787 0.75492526
0.76656679 0.77781808 0.78867657 0.79914142
        67          68          69          70          71          72          73
74          75          76          77
0.80921341 0.81889479 0.82818921 0.83710152 0.84563769 0.85380467 0.86161027
0.86906302 0.87617210 0.88294719 0.88939838
        78          79          80          81          82          83          84
85          86          87          88
0.89553609 0.90137096 0.90691381 0.91217550 0.91716694 0.92189898 0.92638239
0.93062780 0.93464565 0.93844620 0.94203943
        89          90          91          92          93          94          95
96          97          98          99
0.94543511 0.94864269 0.95167135 0.95452997 0.95722710 0.95977099 0.96216956
0.96443042 0.96656086 0.96856786 0.97045808
       100         101
0.97223789 0.97391335
```

Εφόσον έχουμε φτιάξει ήδη την καμπύλη της λογιστικής παλινδρόμησης αρκεί να βρούμε το αντίστοιχο y, δηλαδή την πρόβλεψη πιθανότητας επιτυχίας του μαθήματος του μοντέλου για συγκεκριμένο βαθμό midterm. Για midterm = 5 βρισκόμαστε στο x = 5.0 το οποίο είναι το 51^{ο} σημείο της υπολογισμένης καμπύλης (αφού ξεκινάμε από το 0.0, 0.1 .. 5.0 .. 10.0) άρα η πιθανότητα που μας ζητείτε είναι το

51^{ο} y = 0.60381386

c.

```
c.
> summary(model)

Call:
glm(formula = onesAndZeros ~ MIDTERM, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3358  -0.5486   0.3148   0.6696   1.8437

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7771     0.6171  -4.500 6.80e-06 ***
MIDTERM       0.6397     0.1166   5.488 4.06e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149.078  on 110  degrees of freedom
Residual deviance:  96.598  on 109  degrees of freedom
  (16 observations deleted due to missingness)
AIC: 100.6

Number of Fisher Scoring iterations: 5
```

Για να εξετάσουμε άμα υπάρχει σχέση μεταξύ των δύο μεταβλητών κάνουμε έναν z έλεγχο για το b1 της καμπύλης. Άμα το b1 = 0 τότε δεν υπάρχει σχέση αφού η κλίση της καμπύλης σχέσης θα είναι οριζόντια.

H0: b1 = 0

H1: b1 != 0

Ο έλεγχος μπορεί να υπολογιστεί εύκολα από την r η οποία δίνει πολύ μικρό pvalue που σημαίνει ότι απορρίπτουμε την μηδενική υπόθεση και δεχόμαστε την εναλλακτική. Άρα ο βαθμός της προόδου σχετίζεται (και μάλιστα θετικά) με την επιτυχία του φοιτητή.

d.

Μπορούμε να προβλέψουμε ότι θα περάσει αφού έχουμε υπολογίσει ότι η πιθανότητα του φοιτητή να περάσει είναι περίπου 60% > 50%. Αυτό δεν σημαίνει ότι η πρόβλεψη μας θα γίνει πραγματικότητα, απλά είναι πιο πιθανή η επιτυχία από την αποτυχία.

4)

Σύμφωνα με την Αρχή της Πιθανοφάνειας το πιο πιθανό σενάριο για το νόμισμα που το έχουμε ρίξει 100 φορές και οι 44 έχουν έρθει κορώνα είναι να έχει 44% πιθανότητα να έρθει κορώνα. Θα μπορούσε να νόμισμα να είναι δίκαιο, αλλά το πιο πιθανό είναι να έχει ο πληθυσμός των ρίψεων ίση πιθανότητα με το δείγμα (100 ρίψεις).

Μπορούμε να το αποδείξουμε βρίσκοντας το

$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \lambda(\theta)$, όπου $\lambda(\theta) = \log L(\theta)$ **συνάρτηση λογαριθμικής πιθανοφάνειας** (log-likelihood)

$$L(\theta) = \binom{100}{44}\theta^{44}(1-\theta)^{56}$$

$$\hat{\theta}_{MLE} = \max\big(\lambda(\theta)\big)\forall\theta \in \Theta$$

$$\lambda(\theta) = \log\big(L(\theta)\big) \iff$$

$$\lambda(\theta) = 44log\left(\binom{100}{44}\theta\right) + (100-44)\log(1-\theta)$$

$$Για\ να\ βρούμε\ το\ \theta\ στο\ οποίο\ μέγιστοποιείτε\ η\ καμπύλη\ πρέπει\ \theta{:}\ \ \lambda'(\theta) = 0 \iff$$

$$\frac{44}{\theta} - \frac{56}{1-\theta} = 0 \iff$$

$$\theta = 0.44$$

Άρα $\hat{\theta}_{MLE} = 0.44$