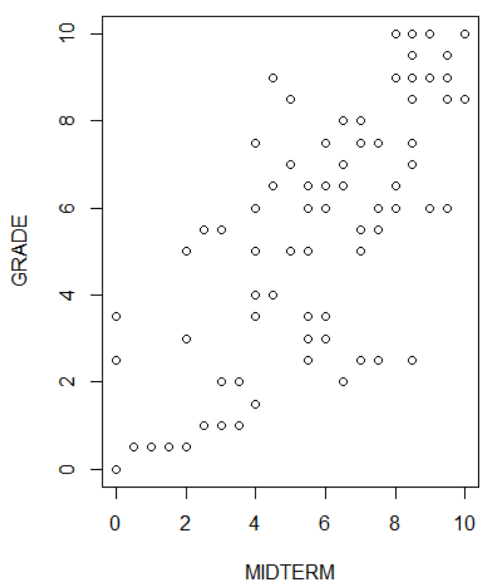
Εκφώνηση:

Εδώ θα χρησιμοποιήσετε τα δεδομένα του αρχείου βαθμών από κάποιο μάθημα (για το ακαδ. έτος 2014-15) για να εξετάσετε τη σχέση μεταξύ του βαθμού στη εξέταση προόδου (μεταβλητή $MIDTERM$) με τον τελικό βαθμό (μεταβλητή $GRADE$) που επιτυγχάνουν οι φοιτητές στο μάθημα αυτό.

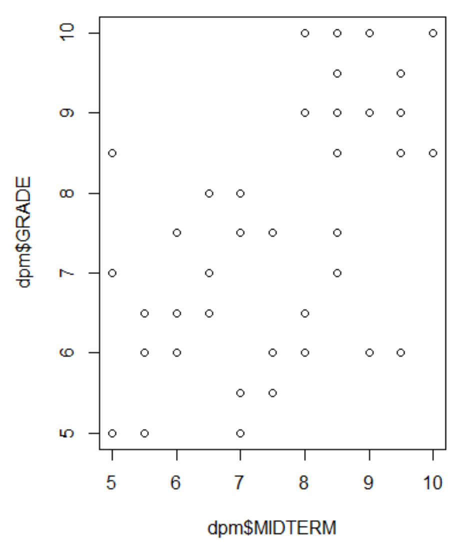
a. Χρησιμοποιώντας διερευνητική ανάλυση, σχολιάστε κατά πόσο η σχέση μεταξύ των δύο μεταβλητών φαίνεται να είναι γραμμική και εάν ικανοποιούνται η ομοσκεδαστικότητα και κανονικότητα.
b. Υποθέτοντας ότι οι μεταβλητές σχετίζονται γραμμικά ως εξής: $GRADE = \beta 1 \times MIDTERM + \beta 0$ + άλλοι παράγοντες, εκτιμήστε τον συντελεστή $\beta 1$ και δώστε ένα 95% διάστημα εμπιστοσύνης για αυτόν.
c. Υπάρχει σχέση μεταξύ των δύο μεταβλητών; Χρησιμοποιήστε έναν έλεγχο σημαντικότητας για να απαντήσετε.
d. Εκτιμήστε το τελικό βαθμό που θα επιτύγχαναν φοιτητές οι οποίοι στην εξέταση προόδου έλαβαν 7. Δώστε ένα 95% διάστημα εμπιστοσύνης.
e. Προβλέψτε τον τελικό που θα επετύγχανε ένας τυχαία επιλεγμένος φοιτητής που πήρε 7 στην πρόοδο, δίνοντας ένα 95% διάστημα πρόβλεψης.

A)

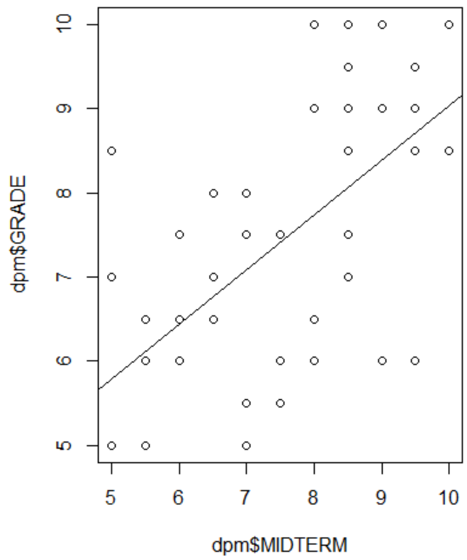
```
> plot(GRADE~MIDTERM)
```




```
> dpm<-grades_2014_data[GRADE>=5 & MIDTERM >=5 & !is.na(GRAD
E) & !is.na(MIDTERM),]
> plot(dpm$GRADE~dpm$MIDTERM)
```

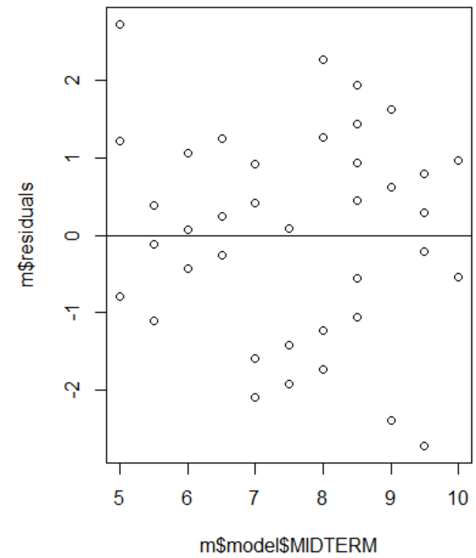


```
> m <- lm(GRADE~MIDTERM, data=dpm)
> abline(m)
```



Βλέπουμε πως η γραμμική σχέση είναι μάλλον καλή προσέγγιση και πως η σχέση των μεταβλητών είναι αύξουσα

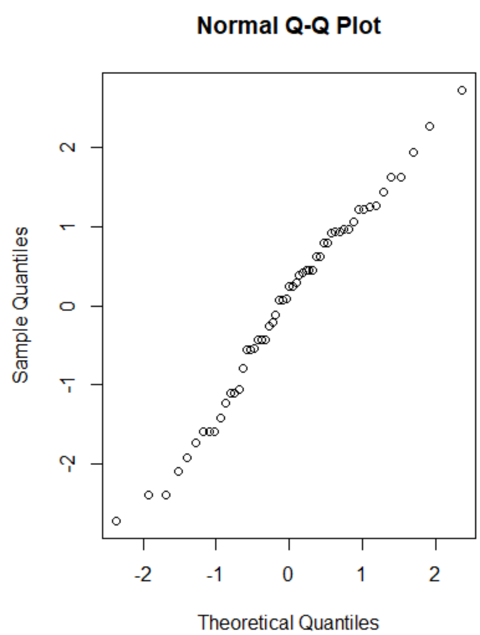
```
plot(m$model$MIDTERM,m$residuals)
abline(0,0)
```



Θα υποθέσουμε ότι ισχύει η ομοσκεδαστικότητα παρόλο που η διασπορά των τιμών δεν είναι ομοιογενής γιατί δεν φαίνεται να είναι πολύ κακή προσέγγιση

Τέλος για τον έλεγχο της κανονικότητας έχουμε:

```
qqnorm(m$residuals)
```



Η κατανομή των υπολοίπων φαίνεται να είναι πάρα πολύ κοντά στην κανονική άρα μπορούμε να συνεχίσουμε.

B)

Το β1 και τα παρακάτω προέκυψαν από τον παρακάτω κώδικα:

```
> summary(m)$coefficients["MIDTERM","Estimate"]->b1
> b1
[1] 0.6503292
> summary(m)$coefficients["MIDTERM","Std. Error"]->SEb1
>
> SEb1
[1] 0.1142137
> n<- length(m$residuals)
> n
[1] 55
> t<-abs(qt(df=n-2,0.025))
> t
[1] 2.005746
> b1+t*SEb1*c(-1,1)
[1] 0.4212456 0.8794128
```

Το 95% διάστημα εμπιστοσύνης είναι
b1+t*SEb1*c(-1,1)=0.4591223, 0.8415362
όπου $SEb1$ = 0.1142137

Γ)

Από το πρώτο ερώτημα έχουμε πως η σχέση των 2 μεταβλητών είναι αύξουσα άρα και υποθέτουμε ως Η0 πως δεν έχουν σχέση και ως Ηα πως έχουν

```
> summary(m)

Call:
lm(formula = GRADE ~ MIDTERM, data = dpm)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7111 -0.9227  0.2399  0.9392  2.7154 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.5330     0.8826   2.870  0.00589 ** 
MIDTERM       0.6503     0.1142   5.694 5.53e-07 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.266 on 53 degrees of freedom
Multiple R-squared:  0.3795,	Adjusted R-squared:  0.3678 
F-statistic: 32.42 on 1 and 53 DF,  p-value: 5.531e-07
```

Το p value είναι 5.531^-7

Άρα από το αποτέλεσμα θεωρούμε πως υπάρχει σχέση

Δ)

```
> predict(m, newdata=data.frame(MIDTERM=7), interval="confidence")
       fit      lwr      upr
1 7.085263 6.717819 7.452707
```

Η πρόβλεψη με 95% διάστημα εμπιστοσύνης δίνεται παραπάνω

Ε)

```
> predict(m, newdata=data.frame(MIDTERM=7), interval="prediction")
       fit      lwr      upr
1 7.085263 4.519378 9.651148
```

Η πρόβλεψη με 95% διάστημα πρόβλεψης δίνεται παραπάνω