Στατιστική Μοντελοποίηση Εργασία 2

Φοιτητής: Αντώνιος Προμπονάς
AM: 03400232

Δεκέμβριος 2023

1 Άσκηση Α

1.1 ΠΡΟΣΑΡΜΟΓΗ ΜΟΝΤΕΛΟΥ ΚΑΙ ΕΛΕΓΧΟΣ ΜΕΤΑΒΛΗΤΩΝ

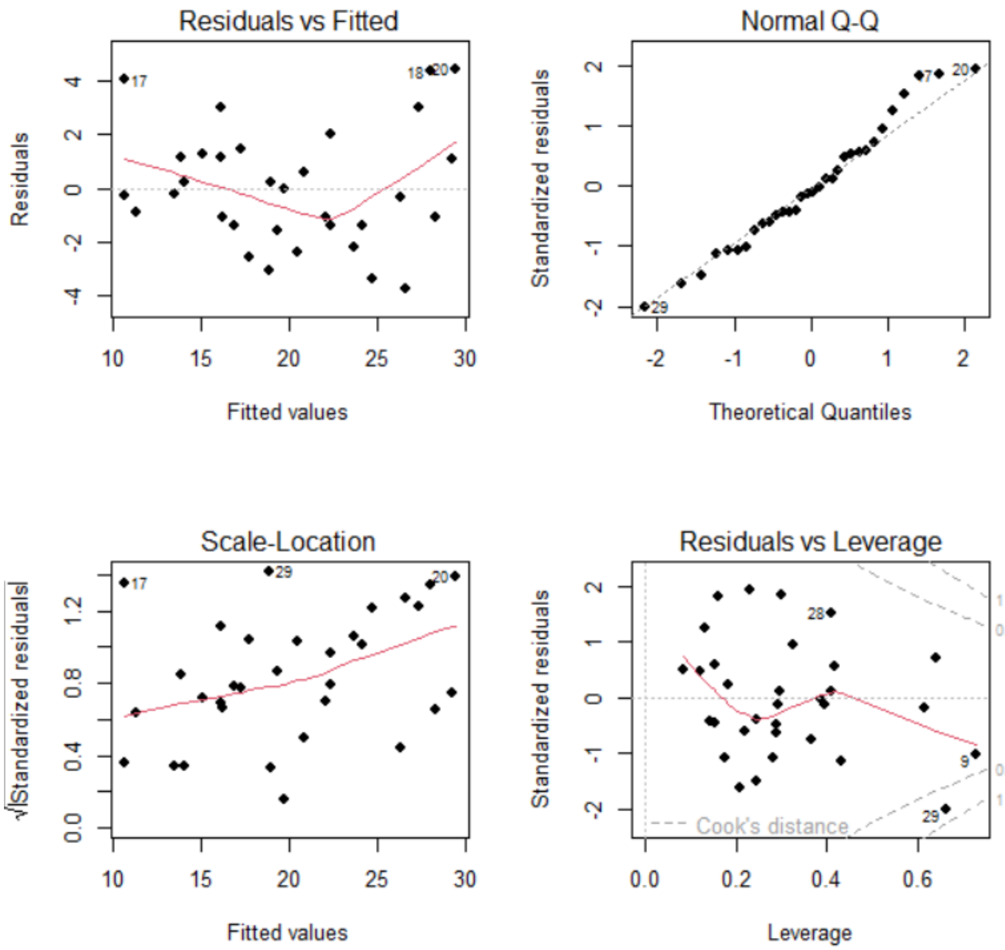
Για να μπορέσουμε να προσαρμόσουμε τα δεδομένα μας σε ένα μοντέλο πολλαπλής γραμμικής παλλινδρόμησης πρέπει να χρησιμοποιήσουμε την εντολή:
$model < -lm(mpg\, cyl + hp + drat + wt + qsec + vs + am + gear + carb)$
Το αποτέλεσμα της συγκεκριμένης εντολής φαίνεται στον παρακάτω πίνακα.

```
lm(formula = mpg ~ cyl + hp + drat + wt + qsec + vs + am + gear +
    carb)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7863 -1.4055 -0.2635  1.2029  4.4753

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.55052   18.52585   0.677   0.5052
cyl          0.09627    0.99715   0.097   0.9240
hp          -0.01295    0.01834  -0.706   0.4876
drat         0.92864    1.60794   0.578   0.5694
wt          -2.62694    1.19800  -2.193   0.0392 *
qsec         0.66523    0.69335   0.959   0.3478
vs           0.16035    2.07277   0.077   0.9390
am           2.47882    2.03513   1.218   0.2361
gear         0.74300    1.47360   0.504   0.6191
carb        -0.61686    0.60566  -1.018   0.3195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.623 on 22 degrees of freedom
Multiple R-squared:  0.8655,    Adjusted R-squared:  0.8105
F-statistic: 15.73 on 9 and 22 DF,  p-value: 1.183e-07
```

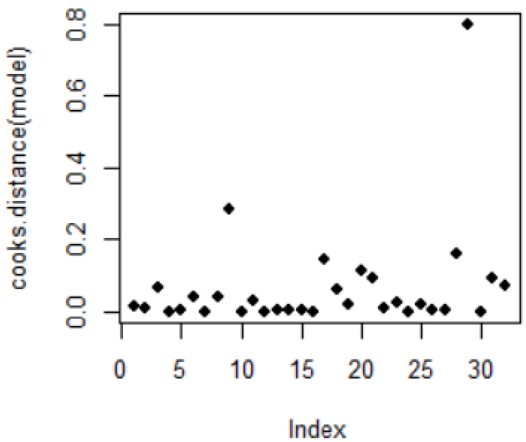
Στο συγκεκριμένο μοντέλο έχουμε ορίσει ως εξαρτημένη μεταβλητή την μεταβλητή mpg. Τη συγκεκριμένη μεταβλητή προσπαθούμε να την προβλέψουμε μέσα από τις 10 υπόλοιπες επεξηγηματικές μεταβλητές. Ωστόσο, πριν ξεκινήσουμε να αναλύουμε το μοντέλο, πρέπει να πραγματοποιήσουμε κάποιους ελέγχους προκειμένου να διαπιστώσουμε τη συσχέτιση και τη πολλυσυγκραμικότητα μεταξύ των μεταβλητών, καθώς και για το αν τα $residuals$ τους, πληρούν τις απαραίτητες προϋποθέσεις.
Σε πρώτη φάση πραγματοποιούμε έλεγχο συσχέτισης του $Pearson$ προκειμένου να εξετάσουμε κατα πόσο τα επίπεδα μιας επεξηγηματικής μεταβλητής επιδρούν στην κατανομή της άλλης.Ο συντελεστής συσχέτισης παίρνει τιμές από [-1,1]. Όταν έχει τιμή κοντά στο 0 , η συσχέτιση μεταξύ των μεταβλητών θεωρείται αδύναμη, ενώ όταν η τιμή ειναι κοντά στο -1 ή 1, υπάρχει δυνατή συσχέτιση.Η συσχέτιση $Pearson$ υλοποιείται μέσω της εντολής $cor(x, y)$
Μέσα λοιπόν, από τον έλεγχο που πραγματοποιήσαμε διαπιστώσαμε ότι οι μεταβλητές $qsec, vs, carb$ έχουν πολύ μικρή συσχέτιση και συνεπώς και επίδραση σε όλες τις άλλες επεξηγηματικές μεταβλητές. Συνεπώς είναι πολύ πιθανό να είναι από τις μεταβλητές που θα αφαιρέσουμε στη συνέχεια προκειμένου να προσδώσουμε μεγαλύτερη αξιοπιστία και εγκυρότητα στο μοντέλο μας.
Ένας άλλος τρόπος να εξετάσουμε αν οι μεταβλητές είναι υψηλά συσχετισμένα μεταξύ τους ή όχι, είναι υπολογίζοντας τη πολυσυγγραμικότητα τους. Αυτό μπορούμε να το πραγματοποιήσουμε μέσα από την εντολή $vif(model)$.Το αποτέλεσμα της συγκεκριμένης εντολής είναι το εξής:

```
> vif(model)
      cyl        hp      drat        wt      qsec        vs        am      gear
14.284737  7.123361  3.329298  6.189050  6.914423  4.916053  4.645108  5.324402
     carb
 4.310597
```

Οι τιμές του vif, οι οποίες αποτελούν ένδειξη πολυσυγγραμικότητας είναι εκείνες για τις οποίες vif >5.
Στη συνέχεια, προκειμένου να ελέγξουμε αν τα $residuals$ του μοντέλου πληρούν τις απαραίτητες προϋποθέσεις , θα εξετάσουμε τα παρακάτω διαγράμματα:

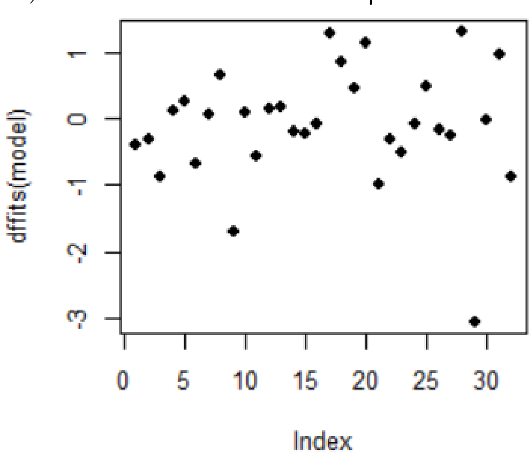


Σε κάθε περίπτωση παρατηρούμε ότι τα υπόλοιπα είναι ομοιόμορφα συμμετρικά κατανεμημένα, γεγονός το οποίο ενισχύει το επιχείρημα ότι το μοντέλο μας πληρεί όλες τις προϋποθέσεις.
Εκτός από τα συγκεκριμένα διαγνωστικά διαγράμματα, υπάρχουν μερικοί ακόμη μέθοδοι προκειμένου να διαπιστώσουμε τη εγκυρότητα των $residuals$. Συγκεκιμένα θα υλοποιήσουμε τις τεχνικές $Cookdistance, DFFiTS$ και hii.
Η τεχνική $cookdistance$ υλοποιείται με την εντολή $plot(cooks.distance(model), pch = 19)$ και απεικονίζεται ως εξής:

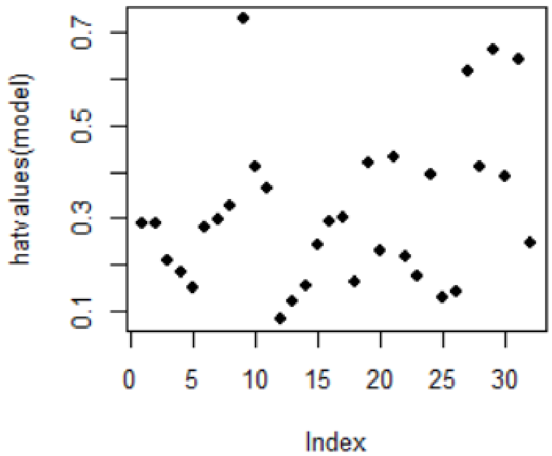


Στη παραπάνω εικόνα παρατηρούμε εύκολα την ύπαρξη μόνο μίας ατυπικής τιμής, ενώ στην πλειοψηφία τους τα $residuals$ είναι ομοιόμορφα κατανεμημένα στην κλίμακα του 0.Καμία τιμή δεν είναι πάνω από 1, άρα δεν υπάρχει σημείο επιρροής στο μοντέλο.
Πάνω κάτω τα ίδια συμπεράσματα βγάζουμε και από τη χρήση των παρακάτω διαγνωστικών τεχνικών. Καμία ύπαρξη σημείου επιρροής και κάποιες μικρές περιπτώσεις, όπου κάπιες τιμές αποκλίνουν από τις συνηθισμένες τιμές των μεταβλητών.

Η τεχνική $DFFiTS$ υλοποιείται με την εντολή $plot(dffits(model), pch = 19)$ και απεικονίζεται ως εξής:



Στη παραπάνω εικόνα
Η τεχνική hii υλοποιείται με την εντολή $plot(hatvalues(model), pch = 19)$ και απεικονίζεται ως εξής:



1.2 ΕΥΡΕΣΗ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ

```
lm(formula = mpg ~ cyl + hp + drat + wt + qsec + vs + am + gear +
    carb)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7863 -1.4055 -0.2635  1.2029  4.4753

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.55052   18.52585   0.677   0.5052
cyl          0.09627    0.99715   0.097   0.9240
hp          -0.01295    0.01834  -0.706   0.4876
drat         0.92864    1.60794   0.578   0.5694
wt          -2.62694    1.19800  -2.193   0.0392 *
qsec         0.66523    0.69335   0.959   0.3478
vs           0.16035    2.07277   0.077   0.9390
am           2.47882    2.03513   1.218   0.2361
gear         0.74300    1.47360   0.504   0.6191
carb        -0.61686    0.60566  -1.018   0.3195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.623 on 22 degrees of freedom
Multiple R-squared:  0.8655,	Adjusted R-squared:  0.8105
F-statistic: 15.73 on 9 and 22 DF,  p-value: 1.183e-07
```

Στη παραπάνω εικόνα βλέπουμε ξανά τη σύνοψη του μοντέλου πολλαπλής γραμμικής παλλινδρόμησης που δημιουργήσαμε.

Σε μία πρώτη ανάγνωση παρατηρούμε ότι το μοντέλο έχει αρκετά καλά χαρακτηριστικά. Συγκεκριμένα έχει $R-squared$ και $AdjustedR-squared$ 87% και 81% αντίστοιχα.
Έχει πολύ χαμηλούς δείκτες στα $F-statistic$ (15.73) και $pvalue$(˜1%), ενώ σημειώνει πολύ χαμηλή τιμη και στο $residual\ standard\ error$.

```
> AIC(model)
[1] 162.5485
```

Στο ίδο μοντέλο παίρνουμε ως αποτέλεσμα μία σχετικά καλή τιμή AIC. Η τιμή AIC υποδηλώνει τη ποσότητα χαμένης πληροφορίας , που προκύπτει στην προσπάθεια μας να εξηγήσου μία συγκεκριμένη μεταβλητή. Συνεπώς, όσο πιο χαμηλή είναι η τιμή αυτή, τόσο καλύτερη ερμηνεία και εγκυρότητα έχει το μοντέλο που δημιουργήσαμε.
Τώρα όσον αφορά το δικό μας μοντέλο, η αλήθεια είναι ότι μπορεί να γίνει πολύ καλύτερο με τις κατάλληλες προσθαφαιρέσεσι επεξηγηματικών μεταβλητών.Για να το πετύχουμε αυτό, υπάρχουν διάφοροι τρόποι.Συγκεκριμένα, μπορούμε να ακουλουθήσουμε τις παρακάτω τεχνικές:

1. Δημιουργούμε τον πίνακα συσχέτισης του $Pearson$ και μέσω αυτού βρίσκουμε εκείνες τις μεταβλητές που έχουν τη μεγάλυτερη συσχέτιση με τη μεταβλητή (mpg).Στη συνέχεια, δημιουργούμε ένα μοντέλο γραμμικής παλλινδρόμησης, χρησιμοποιώντας ως επεξηγηματικές μεταβλητές, τις μεταβλητές αυτές.Η συγκεκριμένη μέθοδος, έχει τα εξής αποτελέσματα:

```
> correlation_matrix <- cor(data)
> correlation_matrix[,1]
       mpg        cyl       disp         hp       drat         wt       qsec 
 1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594  0.4186840 
        vs         am       gear       carb 
 0.6640389  0.5998324  0.4802848 -0.5509251 
```

```
Call:
lm(formula = mpg ~ cyl + disp + wt + hp)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0562 -1.4636 -0.4281  1.2854  5.8269 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.82854    2.75747  14.807 1.76e-14 ***
cyl         -1.29332    0.65588  -1.972 0.058947 .  
disp         0.01160    0.01173   0.989 0.331386    
wt          -3.85390    1.01547  -3.795 0.000759 ***
hp          -0.02054    0.01215  -1.691 0.102379    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.513 on 27 degrees of freedom
Multiple R-squared:  0.8486,    Adjusted R-squared:  0.8262 
F-statistic: 37.84 on 4 and 27 DF,  p-value: 1.061e-10

> AIC(cor_model)
[1] 156.3376
```

2. Χρησιμοποιούμε τη τεχνική *backward elimination*. Η μέθοδος αυτή, υλοποιείται τυποποιημένα στην R μέσω της εντολής $step(model, direction = "backward")$, όπου $model$ είναι το μοντέλο γραμμικής παλλινδρόμησης που περιλαμβάνει όλες τις μεταβλητές. Αυτό που κάνει η μέθοδος αυτή είναι να παίρνει το μοντέλο με όλες τις μεταβλητές και βήμα βήμα να αφαιρεί αυτές, που δεν είναι στατιστιστικά σημαντικές , καταλήγωντας έτσι σε ένα μοντέλο που περιλαμβάνει μόνο μεταβλητές που είναι ουσιαστικά σημαντικές για την ερμηνεία της εξαρτημένης μεταβλητής.Η συγκεκριμένη μέθοδος, έχει τα εξής αποτελέσματα:

```
> summary(backward_elimination)

Call:
lm(formula = mpg ~ wt + qsec + am)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4811 -1.5555 -0.7257  1.4110  4.6610 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.6178     6.9596   1.382 0.177915    
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
qsec          1.2259     0.2887   4.247 0.000216 ***
am            2.9358     1.4109   2.081 0.046716 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,	Adjusted R-squared:  0.8336 
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11

> AIC(backward_elimination)
[1] 154.1194
```

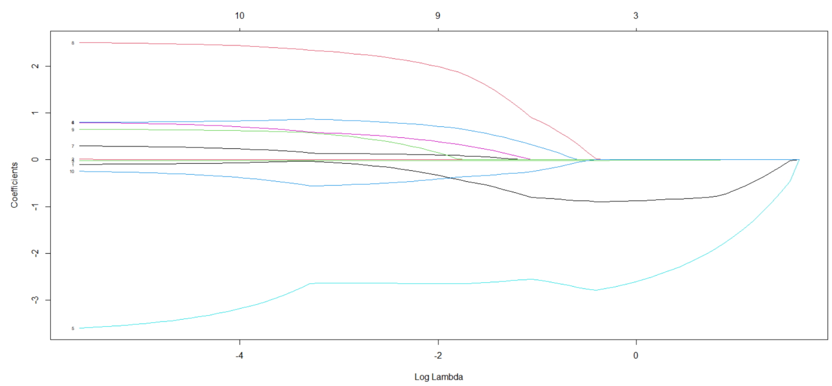
3. Χρησιμοποιούμε τη τεχνική *forward selection*.Η τεχνική αυτή, υλοποιείται τυποποιημένα στην R μέσω της εντολής $step(model, direction = "forward")$.Αυτό που κάνει η μέθοδος αυτή είναι να παίρνει ένα μοντέλο που περιλαμβάνει μόνο την εξαρτώμενη μεταβλητή και βήμα βήμα να προσθέτει εκείνες τις μεταβλητές που έχουν την υψηλότερη συσχέτιση με την μεταβλητή που θέλουμε να επεξηγήσουμε.Η συγκεκριμένη μέθοδος, έχει τα εξής αποτελέσματα:

```
> summary(forward_selection)

Call:
lm(formula = mpg ~ wt + cyl + hp)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9290 -1.5598 -0.5311  1.1850  5.8986 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
cyl         -0.94162    0.55092  -1.709 0.098480 .  
hp          -0.01804    0.01188  -1.519 0.140015    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,	Adjusted R-squared:  0.8263 
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

> AIC(forward_selection)
[1] 155.4766
```

4. Χρησιμοποιούμε τη *both selection* , μία μέθοδος , η οποία συνδυάζει τις 2 προηγούμενες μεθόδους, αφού ξεκινάει από ένα μοντέλο που περιέχει μόνο την εξαρτώμενη μεταβλητή και στη συνέχεια με προσθαφαιρέσεις, καταλήγει σε ένα μοντέλο που περιλαμβάνει μόνο μεταβλητές που είναι ουσιαστικά σημαντικές για την ερμηνεία της εξαρτημένης μεταβλητής.Η συγκεκριμένη μέθοδος, έχει τα εξής αποτελέσματα:

```
> summary(both)

Call:
lm(formula = mpg ~ wt + cyl + hp)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9290 -1.5598 -0.5311  1.1850  5.8986 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
cyl         -0.94162    0.55092  -1.709 0.098480 .  
hp          -0.01804    0.01188  -1.519 0.140015    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,	Adjusted R-squared:  0.8263 
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

> AIC(both)
[1] 155.4766
```

5. Τέλος, υπάρχει και η μέθοδος *lasso*, η οποία λειτουργεί ως εξής: Χρησιμοποιούμε την εντολή αυτή $glmnet$ $(as.matrix(data_no_missing_columns[-1])$, $data_no_missing_columns[,1], standarize = TRUE, alpha = 1)$, προκειμένου να δημιουργήσουμε το παρακάτω διάγραμμα, στο οποίο φαίνεται ότι οι μεταβλητές cyl, am, wt και $carb$,επηρεάζουν περισσότερο από ολες τις άλλες μεταβλητές το μοντέλο, διότι είναι αυτές οι οποίες εισέρχονται πρώτες και με μεγαλύτερη διακύμανση, οπώς φαίνεται και στη παρακάτω γραφική παράσταση.



Συνεπώς , το μοντέλο που προκύπτει από τις μεταβλητές αυτές είναι το εξής:

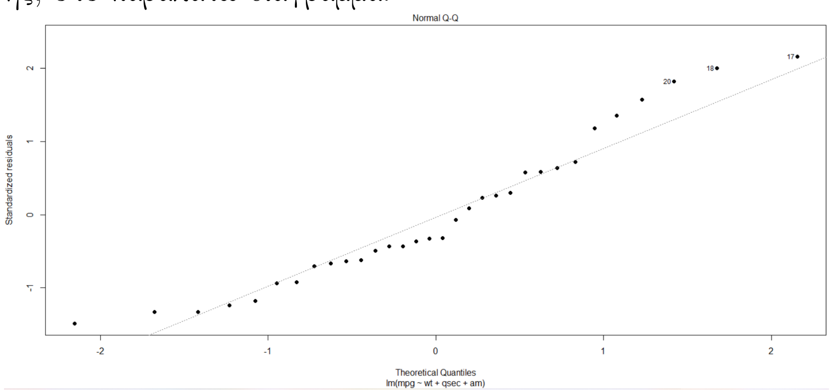
```
Call:
lm(formula = mpg ~ cyl + am + wt + carb)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5451 -1.2184 -0.3739  1.4699  5.3528 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  36.8503     2.8694  12.843 5.17e-13 ***
cyl          -1.1968     0.4368  -2.740   0.0108 *  
am            1.7801     1.5091   1.180   0.2485    
wt           -2.4785     0.9364  -2.647   0.0134 *  
carb         -0.7480     0.3956  -1.891   0.0694 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.5 on 27 degrees of freedom
Multiple R-squared:  0.8502,	Adjusted R-squared:  0.828 
F-statistic:  38.3 on 4 and 27 DF,  p-value: 9.255e-11

> AIC(after_lasso)
[1] 156.0095
```

Μέσα από τις παραπάνω μεθόδους ,παρατηρούμε ότι πάνω κάτω όλα τα μοντέλα μας δίνουν περίπου τις ίδιες τιμές στους δείκτες AIC, $R-squared$ και $Adjusted\ R-squared$. Ωστόσο, η άποψη μου είναι ότι το καλύτερο μοντέλο , το δίνει η μέθοδος $backward\ elimination$. Σύμφωνα με τη μέθοδο αυτή, τη μεταβλητή mpg, μπορούν να την εξηγήσουν καλύτερα , οι μεταβλητές wt, $qsec$ και am. Συγκεκριμένα, δημιουργείται ένα μοντέλο το οποίο έχει τη χαμηλότερη απώλεια πληροφορίας (AIC=154.12), και έχει από τις υψηλότερες τιμές $Rsquared$ και $Adjusted\ R-squared$ με 0.85 και 0.83 αντίστοιχα. Έχοντας στο νου μάλιστα, ότι η τιμή 0.83 είναι η μεγαλύτερη για $Adjusted$

$R - squared$ και τη μεγαλύτερη τιμή $Rsquared$ τη δίνει το μοντέλο που περιλμάνει όλες τις μεταβλητές , με 0.865, καταλαβαίνουμε ότι το μοντέλο που επιλέξαμε , είναι αρκετά αποδοτικό και μπορεί να ερμηνεύσει με πολύ μεγάλη ακρίβεια την εξαρτώμενη μεταβλητή.

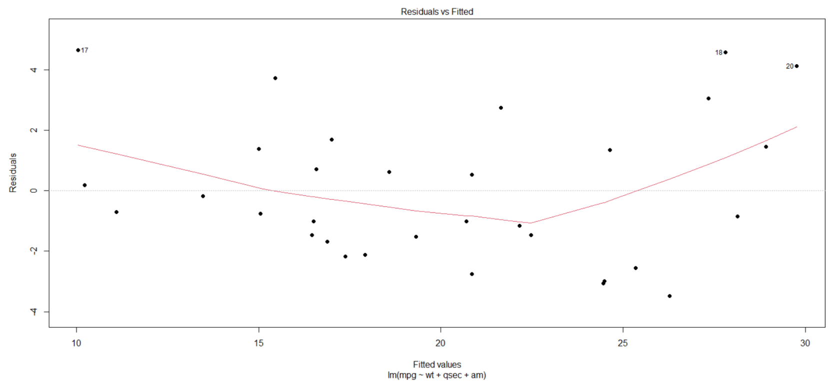
1.3 ΕΛΕΓΧΟΣ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΜΟΝΤΕΛΟΥ

Τώρα αυτό που θα κάνουμε για το συγκεκριμένο μοντέλο είναι να πραγματοποιήσουμε ορισμένα διαγνωστικά $test$.
Αρχικά, εξετάζουμε την εμφάνιση άτυπων σημείων ή σημείων επιρροής, στο παρακάτω διάγραμμα:

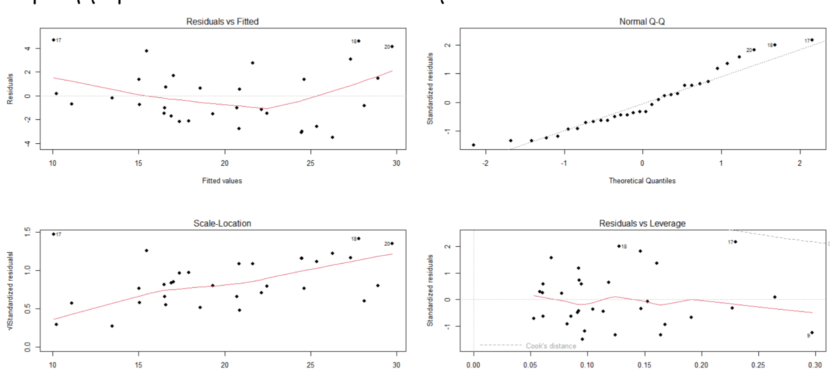


Όπως, γίνεται εύκολα αντιληπτό σχεδόν όλα τα σημεία βρίσκονται πάνω στην ευθεία της κανονικής κατανομής, με εξαίρεση μόνο το πρώτο στοιχείο που δείχνει να βρίσκεται αρκετά μακριά από την ευθεία. Σε γενικές γραμμές, ομώς, το διάγραμμα δείχνει ότι δεν αντιμετωπίζουε καποίο σοβαρό πρόβλημα από εμφάνιση ατυπικών τιμών.

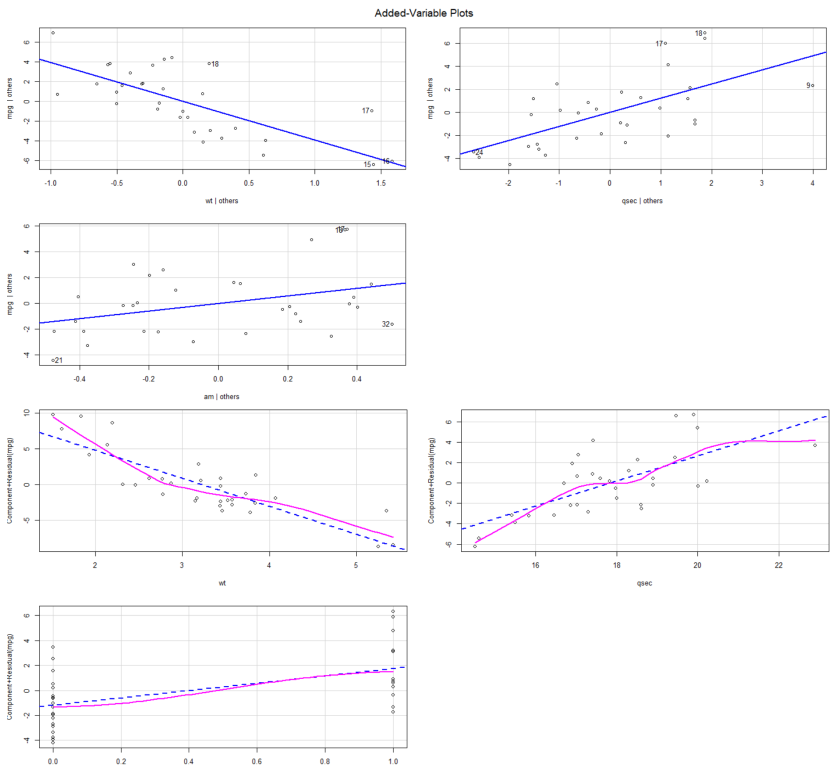
Ένας ακόμη τρόπος για να ελέγξουμε την εγκυρότητα του μοντέλου μας,είναι μέσα από το παρακάτω διάγραμμα:



Στο συγκεκριμένο διάγραμμα , ελέγχουμε τη σχέση μεταξύ των υπολοίπων και των προβλεπόμενων τιμών. Παρατηρούμε ότι δεν υπάρχει πρότυπο στη διακύμανση των υπολοίπων σε σχέση με τις προβλεπόμενες τιμές και δεν υπάρχει μορφή που επαναλαμβάνεται στη διακύμανση των υπολοίπων καθώς αλλάζει η προβλεπόμενη τιμή. Συνεπώς έχουμε πολύ καλές ενδείξεις ότι το μοντέλο μας προσδιορίζει τη δομή των δεδομένων με έναν αρκετά καλό τρόπο.
Παρακάτω, παραθέτω τα 2 παραπάνω διαγράμματα με 2 ακόμη διαγνωστικά $tests$, τα οποία επιβεβαιώνουν όσα μόλις ανεφέραμε σχετικά με την ερμηνεία των $residuals$ του μοντέλου.



Επιπλέον, θα ελέγξουμε για κάθε μία επεξηγηματική μεταβλητή του μοντέλου , το κατά πόσο είναι σημαντική για την ερμηνεία της εξαρτώμενης μεταβλητής. Αυτό θα συμβεί, μέσα από τον έλεγχο των παρακάτω διαγραμμάτων:



Στα διαγράμματα των μεταβλητών, wt και $qsec$, βλέπουμε ότι οι τιμές είναι αρκετά κοντά στην ευθεία της κανονικής κατανομής. Ωστόσο, στο διάγραμμα της μεταβλητής am, τα δεδομένα φαίνεται να αποκλίνουν λίγο ή και πολύ από την ευθεία αυτή. Ωστόσο, αυτό δεν είναι κάτι που πρέπει να μας προβληματίζει καθώς, η μεταβλητή am παίρνει τιμές 0 ή 1. Συνεπώς είναι λογικό να αποκλίουν οι τιμές της, από την ευθεία της κανονικής κατανομής.

Επόμενο βήμα είναι να υπολογίσουμε ένα διάστημα εμπιστοσύνης επιπέδου 95% για τους συντελεστές του μοντέλου αυτού.

```
> confidence_interval <- confint(backward_elimination)
> confidence_interval
                 2.5 %     97.5 %
(Intercept) -4.63829946 23.873860
wt          -5.37333423 -2.459673
qsec         0.63457320  1.817199
am           0.04573031  5.825944
```

Μέσα από την παραπάνω εικόνα, μπορουν να βγουν χρήσιμα συμπεράσματα. Συγκεκριμένα:

1. Το διάστημα εμπιστοσύνης για τον συντελεστή της σταθεράς ($Intercept$) είναι από -4.64 έως 23.87. Αυτό σημαίνει ότι, με επίπεδο εμπιστοσύνης 95%, η πραγματική τιμή του συντελεστή μπορεί να βρίσκεται εντός αυτού του διαστήματος. Η τιμή 0 βρίσκεται εντός του διαστήματος, επομένως δεν μπορούμε να αποκλείσουμε το μηδέν ως μια πιθανή τιμή για το $Intercept$-mpg.

2. Το διάστημα εμπιστοσύνης για τον συντελεστή του βάρους (wt) είναι από -5.37 έως -2.46. Αυτό σημαίνει ότι, με επίπεδο εμπιστοσύνης 95%, ο πραγματικός συντελεστής μπορεί να βρίσκεται εντός αυτού του διαστήματος. Η τιμή 0 δεν βρίσκεται εντός του διαστήματος, επομένως μπορούμε να αποκλείσουμε το μηδέν ως πιθανή τιμή για το wt.

3. Το διάστημα εμπιστοσύνης για τον συντελεστή του $qsec$ είναι από 0.63 έως 1.82. Αυτό σημαίνει ότι, με επίπεδο εμπιστοσύνης 95%, ο πραγματικός συντελεστής μπορεί να βρίσκεται εντός αυτού του διαστήματος. Η τιμή 0 δεν βρίσκεται εντός του διαστήματος, επομένως μπορούμε να αποκλείσουμε το μηδέν ως πιθανή τιμή για το $qsec$.

4. Το διάστημα εμπιστοσύνης για τον συντελεστή του am είναι από 0.05 έως 5.83. Αυτό σημαίνει ότι, με επίπεδο εμπιστοσύνης 95%, ο πραγματικός συντελεστής μπορεί να βρίσκεται εντός αυτού του διαστήματος. Η τιμή 0 δεν βρίσκεται εντός του διαστήματος, επομένως μπορούμε να αποκλείσουμε το μηδέν ως πιθανή τιμή για το am.

Επίσης, υπολογίζουμε τη πρόβλεψη μιας άγνωστης παρατήρησης Y ως εξής:

```
> new_data <- data.frame(wt=2.620,qsec=16.46,am=1)
> predict(backward_elimination, new_data, interval="predict")
       fit      lwr      upr
1 22.47046 17.22244 27.71849
> predict(backward_elimination, new_data, interval="confidence"
       fit      lwr      upr
1 22.47046 20.99627 23.94465
> 
```

Συμπεραίνουμε, λοιπόν, ότι:

- η πρόβλεψη για την μεταβλητή mpg είναι ότι παίρνει τιμές με κατώτατο όριο 17.22244 και ανώτατο όριο 27.71849. Με βάση τις τιμές που δώσαμε στις επεξηγηματικές μεταβλητές, εκτιμώμενη τιμής της εξαρτώμενης μεταβλητής είναι 22.47046

- το διάστημα εμπιστοσύνης για την εκτίμηση του μέσου της εξαρτημένης μεταβλητής, έχει κατώτατο όριο 20.99627 και ανώτατο όριο 23.94465.Αυτό το διάστημα δείχνει περίπου πού βρίσκεται το μέσο της κατανομής των εκτιμήσεων.Και η πρόβλεψη είναι ότι αυτή η τιμή είναι 22.47046

2 Άσκηση B

Στο συγκεκριμένο ερώτημα, θα προσπαθήσουμε να δημιουργήσουμε και να προσαρμόσουμε ένα βέλτιστο μοντέλο γραμμικής παλλινδρόμησης, το οποίο θα προσπαθεί να ερμηνεύσει το βάρος Y (kg), ανδρών (M) και γυναικών (F) σε σχέση με το ύψος τους (m).Το μοντέλο αυτό είναι της μορφής $\mathrm{E}(y)= b0 + b1x1 + b2x2 + b3x3$. Πάνω σε αυτό το μοντέλο θα προσπαθήσουμε να βρούμε αν χρειάζεται να προσαρμοστούν (I) δύο διαφορετικές ευθείες, (II) δύο παράλληλες ευθείες, ή (III) μια κοινή ευθεία και για τις δύο ομάδες. Συνεπώς, χρειάζεται να δημιουργήσουμε και στη συνέχεια να αξιολογήσουμε 3 διαφορετικά μοντέλα. Στο 1ο μοντέλο θα εξετάσουμε την αλληλεπίδραση που έχουν μεταξύ τους, όλα τα στοιχεία ($x1$,$x2$,$x3$), στο 2ο μοντέλο θα εξετάσουμε την αλληλεπίδραση μεταξύ των στοιχείων ($x1$ και $x2$) και στο 3ο μοντέλο θα εξετάσουμε μόνο το στοιχείο $x1$. Τα μοντέλα αυτά, θα δημιουργηθούν με βάση τα δεδομένα από το συγκεκρμένο *dataset*:

```
> data
   id gender height  weight
1   1      F 1.4224  53.118
2   2      F 1.5240  56.750
3   3      F 1.6256  60.382
4   4      F 1.7272  64.014
5   5      F 1.8288  67.646
6   6      F 1.3716  49.486
7   7      F 1.5748  58.112
8   8      F 1.6510  59.474
9   9      F 1.6510  59.474
10 10      F 1.7780  65.830
11 11      M 1.6256  95.794
12 12      M 1.7272 101.242
13 13      M 1.8288 106.690
14 14      M 1.9304 112.138
15 15      M 2.0320 117.586
16 16      M 1.5748  91.254
17 17      M 1.7526 103.512
18 18      M 1.8796 111.230
19 19      M 1.9050 109.414
20 20      M 2.0828 122.126
```

Το συγκεκριμένο σύνολο δεδομένων περιέχει στοιχεία για 20 διαφορετικά άτομα, τα οποία έχουν διαφορετική τιμή id. Η μεταβλητή $gender$ είναι δίτιμη, κάθως είτε περιέχει F, αν το άτομο είναι γυναίκα, είτε περιέχει M, αν το άτομο είναι άντρας. Η μεταβλητή $height$ δείχνει το ύψος του ατόμου, ενώ η μεταβλητή $weight$, δείχνει το βάρος.
Συμβατικά, για να ορίσουμε τα παρακάτω μοντέλα κάνουμε τις εξής παραδοχές: $y < -weight$, $x1 < -height$, $x2 < -gender$ και $x3 < -height * gender$

Το 1ο μοντέλο που δημιουργείται και στο οποίο προσαρμόζονται 2 διαφορετικές ευθείες είναι το εξής:

```
> y<- weight
> x1<- height
> x2<- ifelse(gender == "M", 1, 0)
> x3<-x1*x2
> my_model<- lm(y~x1+x2+x3)
> summary(my_model)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7394 -0.8080  0.2251  0.6163  1.5248 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.088      3.637  -0.299    0.769    
x1            37.462      2.243  16.700 1.51e-11 ***
x2             3.632      5.162   0.703    0.492    
x3            19.552      2.999   6.520 7.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9875 on 16 degrees of freedom
Multiple R-squared:  0.9987,	Adjusted R-squared:  0.9985 
F-statistic:  4250 on 3 and 16 DF,  p-value: < 2.2e-16

> AIC(my_model)
[1] 61.79283
> aov(my_model)
Call:
   aov(formula = my_model)

Terms:
                      x1       x2       x3 Residuals
Sum of Squares  7931.194 4461.733   41.453    15.604
Deg. of Freedom        1        1        1        16

Residual standard error: 0.9875323
Estimated effects may be unbalanced
```

Η εξίσωση παλλινδρόμησης που προκύπτει από το συγκεκριμένο μοντέλο είναι: $y = -1.1 + 37.5x1 + 3.6x2 + 19.55x3$

Το 2ο μοντέλο που δημιουργείται και στο οποίο προσαρμόζονται 2 παράλληλες ευθείες, είναι το εξής:

```
> parallel_model<- lm(y~x1+x2)
> summary(parallel_model)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3048 -1.2844  0.0924  1.1104  3.0328 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -18.761      4.499   -4.17 0.000642 ***
x1            48.401      2.762   17.52 2.56e-12 ***
x2            37.097      1.017   36.46  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.832 on 17 degrees of freedom
Multiple R-squared:  0.9954,	Adjusted R-squared:  0.9949 
F-statistic:  1846 on 2 and 17 DF,  p-value: < 2.2e-16

> AIC(parallel_model)
[1] 85.72372
> aov(parallel_model)
Call:
   aov(formula = parallel_model)

Terms:
                      x1       x2 Residuals
Sum of Squares  7931.194 4461.733    57.056
Deg. of Freedom        1        1        17

Residual standard error: 1.832011
Estimated effects may be unbalanced
```

Η εξίσωση παλλινδρόμησης που προκύπτει από το συγκεκριμένο μοντέλο είναι: $y = -18.76 + 48.4x1 + 37.1x2$

Το 3ο μοντέλο που δημιουργείται και στο οποίο προσαρμόζεται 1 ευθεία, είναι το εξής:

```
> one_line<- lm(y~x1)
> summary(one_line)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.876 -13.086   1.814  11.454  24.192 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -103.19      33.36  -3.093  0.00627 ** 
x1            108.11      19.23   5.621 2.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.84 on 18 degrees of freedom
Multiple R-squared:  0.637,	Adjusted R-squared:  0.6169 
F-statistic: 31.59 on 1 and 18 DF,  p-value: 2.473e-05

> AIC(one_line)
[1] 171.1629
> aov(one_line)
Call:
   aov(formula = one_line)

Terms:
                      x1 Residuals
Sum of Squares  7931.194  4518.790
Deg. of Freedom        1        18

Residual standard error: 15.84436
Estimated effects may be unbalanced
```

Η εξίσωση παλλινδρόμησης που προκύπτει από το συγκεκριμένο μοντέλο είναι: $y = -103.19 + 108.11x1$

Και στα 3 παραπάνω μοντέλα, έχουμε συγκεντρώσει μερικά πολύ σημαντικά στοιχεία, τα οποία θα μας βοηθήσουν να αξιολογήσουμε το κάθε μοντέλο ξεχωριστά και να καταλήξουμε ποιο από αυτά είναι το πιο αποδοτικό στην ερμηνεία της εξαρτώμενης μεταβλητής.Επίσης , και στα 3 μοντέλα , έχουμε εφαρμόσει τη μέθοδο $anova$. Με τη συγκεκριμένη τεχνική συγκρίνουμε τις μέσες τιμές διαφόρων δειγμάτων και βασιζόμαστε στις αποκλίσεις από τη μέση τιμή.
Το μοντέλο 1 έχει τιμές $R-squared$ και $Adjusted\ R-squared$ πολύ κοντά στο 1. Αυτό ωστόσο δεν σημαίνει αναγκαστικά ότι το μοντέλο αυτό είναι το τέλειο. Υψήλες τιμές σε αυτούς τους δεικτές δεν εξασφαλίζουν την υψηλή απόδοση ενός μοντέλου. Παρατηρούμε , ότι η σταθερά και η μεταβλητή $Q2$ δεν επηρεάζουν στατιστικά σημαντικά την ερμήνεια του μοντέλου. Παρ΄όλα αυτά, είναι θετικό στοιχείο η χαμηλή τιμή AIC, που δείχνει χαμηλή απώλεια πληροφορίας, οπώς και η χαμηλή τιμή στα $residuals$, που προκύπτει μέσα από την ανάλυση $anova$.

Το μοντέλο 2 έχει και αυτό τιμές $R - squared$ και $Adjusted\ R - squared$ πολύ κοντά στο 1. Όπως ανέφερα και πριν ωστόσο,υψήλες τιμές σε αυτούς τους δεικτές δεν εξασφαλίζουν την υψηλή απόδοση ενός μοντέλου.Παρ΄ όλα αυτά, αξίζει να επισημάνουμε ότι στο μοντέλο αυτό, όλες οι μεταβλητές είναι στατιστικά σημαντικές. Αυτό το στοιχείο είναι πολύ σημαντικό , διότι υποδυκνύει ότι όλες οι μεταβλητές συμμετέχουν ενεργά στην ερμηνεία του μοντέλου.Μπορεί εδώ, η τιμή AIC να είναι λίγο υψηλότερη, ωστόσο όμως δεν είναι κακή (AIC=85.72). Επίσης , το μοντέλο αυτό, έχει καλή τιμή στα $residuals$ με $residuals\ standard\ error$=1.03

Το μοντέλο 3 , είναι με διαφορά το χειρότερο από τα 3 μοντέλα που δημιουργήσαμε , διότι έχει πολύ πιο υψηλή τιμή χαμένης πληροφορίας (AIC=171), πολύ πιο υψηλές τιμές στα $residuals\ standard\ errors$ και επίσης έχει πολύ χαμηλές τιμές $R - squared$ και $Adjusted\ R - squared$, με 0,63 και 0,62 αντίστοιχα.

Με βάση ,λοιπόν, τους παραπάνω συλλογισμούς, καταλήγω ότι το καλύτερο μοντέλο, είναι το μόντελο 2 με συνάρτηση παλλινδρόμησης $y = 18.76 + 48.4x1 + 37.1x2$. Στο συγκεκριμένο μοντέλο, όλες οι μεταβλητές είναι στατιστικά σημαντικές και συμβάλουν ενεργά στην ερμηνεία της εξαρτώμενης μεταβλητής.Συνεπώς, το αρχικό μοντέλο προσαρμόζεται σε 2 παράλληλες ευθείες. Παρακάτω, παρουσιάζω τη γραφική παράσταση της συνάρτησης.