Δεύτερη σειρά ασκήσεων 2022-2023

Μάθημα: 858 - Στατιστική Μοντελοποίηση

ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Ονοματεπώνυμο: Απόστολος Μουστάκης

Αριθμός Μητρώου: 03400182

Καθηγήτρια: Χρυσηίς Καρώνη – Ρίτσαρντσον

Περιεχόμενα

Άσκηση Α

Ερώτημα 1

Τα δεδομένα μας, τα οποία βρίσκονται στο αρχείο vehicles.txt, παρουσιάζουν τα αποτελέσματα μίας έρευνας 32 τύπων αυτοκινήτων. Η επεξήγηση των μεταβλητών που χρησιμοποιούνται δίνεται στο παρακάτω πίνακα:

mpg	Κατανάλωση βενζίνης Miles/(US) gallon
cyl	Αριθμός κυλίνδρων
disp	Μετατόπιση (Displacement) (cu.in.)
hp	Μικτή ιπποδύναμη (Gross horsepower)
drat	Αναλογία οπίσθιου άξονα (Real axle ratio)
wt	Βάρος (1000lbs)
qsec	¼ mile time
vs	Διάταξη κινητήρα (0 = V, 1 = straight)
am	Κιβώτιο ταχυτήτων (0 = automatic, 1 = manual)
gear	Αριθμός προς τα εμπρός ταχυτήτων (forward gears)
carb	Αριθμός καρμπυρατέρ

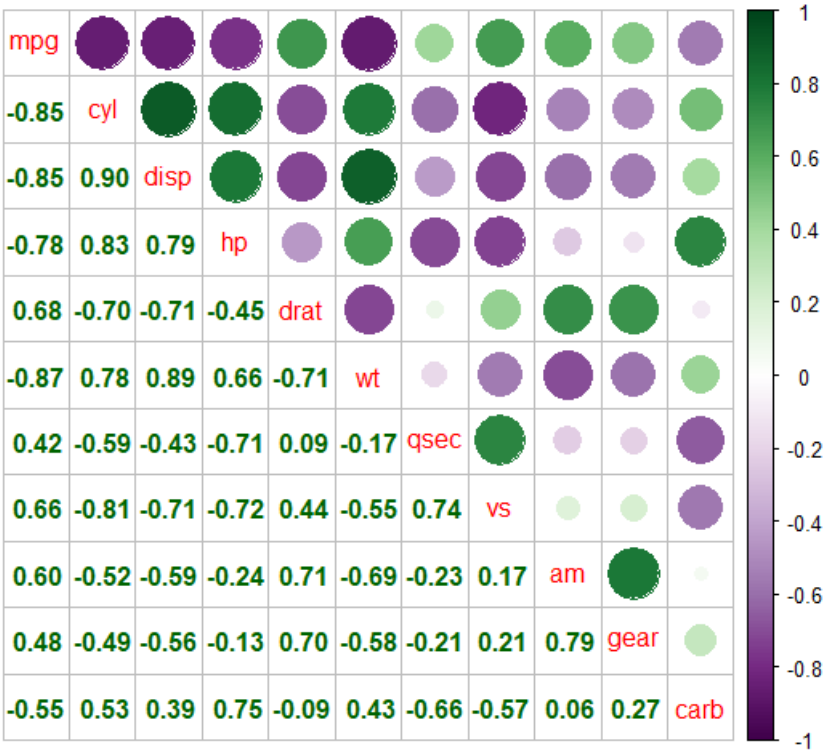
Πίνακας 1: Περιγραφή των δεδομένων της άσκησης

Θα προσαρμόσουμε, με χρήση της R, ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης στα δεδομένα, όπου η εξαρτημένη μεταβλητή y είναι η mpg και οι υπόλοιπες 10 μεταβλητές είναι οι ανεξάρτητες x_j. Τα αποτελέσματα φαίνονται παρακάτω.

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

Coefficients:
(Intercept)         cyl        disp          hp        drat          wt        qsec          vs          am
   12.30337    -0.11144     0.01334    -0.02148     0.78711    -3.71530     0.82104     0.31776     2.52023
       gear        carb
    0.65541    -0.19942


Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Πίνακας 2: Μοντέλο πολλαπλής γραμμικής παλινδρόμησης

Παρατηρούμε πως γενικά η προσαρμογή του μοντέλου είναι καλή καθώς έχουμε υψηλό συντελεστή προσδιορισμού (R^2=0.869) αλλά τα p-values των ανεξάρτητων μεταβλητών x_j είναι υψηλά (ορισμένα πάρα πολύ υψηλά) γεγονός που φανερώνει πως υπάρχουν συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών και πως δεν χρειάζονται όλες στο μοντέλο. Για να διερευνήσουμε τις συσχετίσεις μεταξύ των μεταβλητών x_j δημιουργούμε το παρακάτω διάγραμμα συσχετίσεων.



Πίνακας 3: Διάγραμμα συσχετίσεων των 11 μεταβλητών x_j

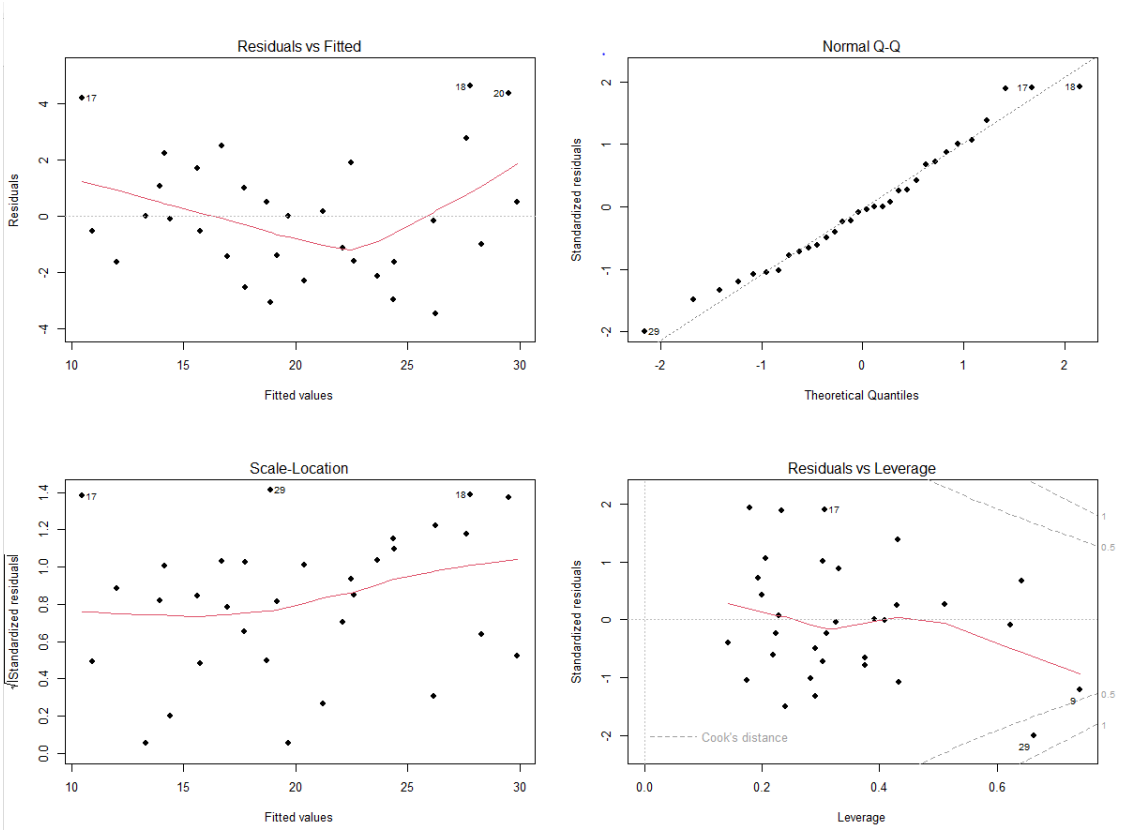
Κοιτώντας το διάγραμμα παρατηρούμε πως η εξαρτημένη μεταβλητή mpg εμφανίζει έντονη (αρνητική) συσχέτιση με τις ανεξάρτητες μεταβλητές cyl (-0.85), disp (-0.85) και wt (-0.87). Ωστόσο, αυτό το οποίο μας ενδιαφέρει είναι αν ανεξάρτητες μεταβλητές έχουν συσχέτιση μεταξύ τους, κάτι το οποίο παρατηρούμε έντονα μεταξύ των μεταβλητών cyl & disp (0.90), cyl & hp (0.83), cyl & vs (-0.81) και disp & wt (0.89). Το γεγονός πως ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους σε μεγάλό βαθμό αποτελεί ένδειξη πολυσυγγραμμικότητας, την οποία θα εξετάσουμε με το κριτήριο VIF. Ο υπολογισμός του VIF (Variance Inflation Factor) για κάθε ανεξάρτητη μεταβλητή του μοντέλου φαίνεται στον παρακάτω πίνακα.

cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
15.373833	21.620241	9.832037	3.374620	15.164887	7.527958	4.965873	4.648487	5.357452	7.908747

Πίνακας 4: Υπολογισμός VIF για τις 10 ανεξάρτητες μεταβλητές x_j

Δεδομένου πως τιμές του VIF > 5 αποτελούν ένδειξη πολυσυγγραμμικότητας, είναι πάρα πολύ πιθανό να υπάρχει πολυσυγγραμμικότητα στο μοντέλο μας. Ειδικότερα για τις μεταβλητές cyl, disp και wt ο VIF είναι πολύ υψηλός.

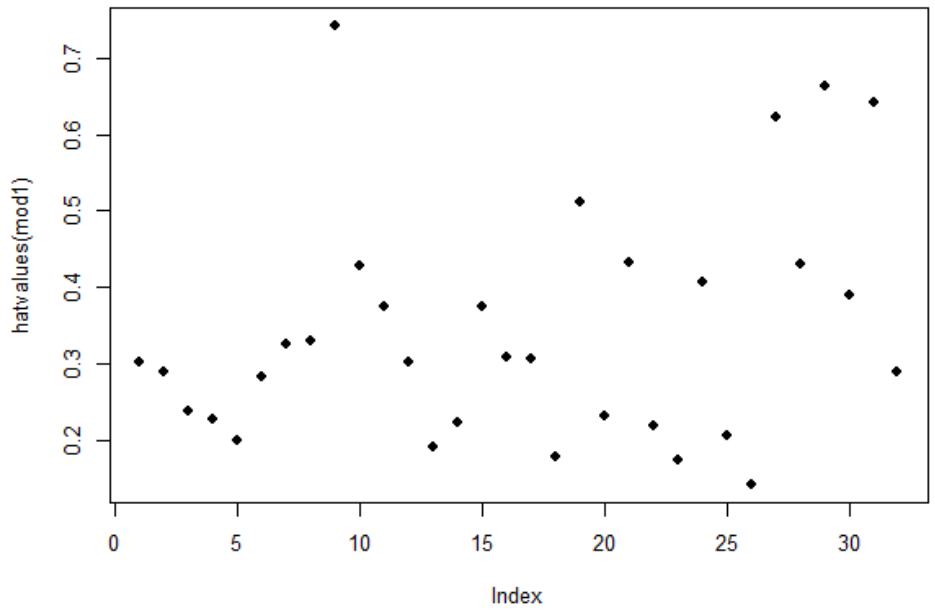
Θα προχωρήσουμε εξετάζοντας αν τηρούνται οι προϋποθέσεις του μοντέλου με βάση τα υπόλοιπα, χρησιμοποιώντας τα τυποποιημένα υπόλοιπα. Τα διαγράμματα Residuals vs Fitted Values, Normal Q-Q, Scale-Location και Residuals vs Leverage φαίνονται στον παρακάτω πίνακα.



Πίνακας 5: Διαγράμματα τυποποιημένων υπολοίπων

Όσον αφορά το διάγραμμα Residuals vs Fitted θέλουμε τα υπόλοιπα να κατανέμονται με τυχαίο τρόπο γύρω από το μηδέν, ώστε να ικανοποιείται η υπόθεση της ομοσκεδαστικότητας, κάτι το οποίο συμβαίνει στο διάγραμμα μας. Παρομοίως στο διάγραμμα Scale-Location θέλουμε τα υπόλοιπα να κατανέμονται τυχαία, κάτι το οποίο συμβαίνει καθώς η γραμμή που προκύπτει είναι σχεδόν οριζόντια. Όσον αφορά το διάγραμμα Normal Q-Q θέλουμε τα υπόλοιπα να κείτονται σε μία ευθεία, κάτι το οποίο επαληθεύεται σε μεγάλο βαθμό. Τέλος, όσον αφορά το διάγραμμα Residuals vs Leverage θέλουμε η διασπορά των υπολοίπων να μην μεταβάλλεται καθώς μεταβάλλεται η μόχλευση, κάτι το οποίο ικανοποιείται σε μεγάλο βαθμό. Αξίζει να σημειωθεί στην συγκεκριμένη περίπτωση πως οι παρατηρήσεις 9 και 29 μπορούν να θεωρηθούν σημεία επιρροής με βάση την απόσταση COOK.

Θα προχωρήσουμε κάνοντας χρήση διαγνωστικών ελέγχων για την πιθανή παρουσία άτυπων σημείων ή σημείων επιρροής. Ξεκινάμε με τον υπολογισμό των h_{ii}, τα οποία είναι τα διαγώνια στοιχεία του hat matrix. Παρακάτω παρατίθεται το διάγραμμα των h_{ii} καθώς και ο πίνακας με τις ακριβείς μετρήσεις των h_{ii}.



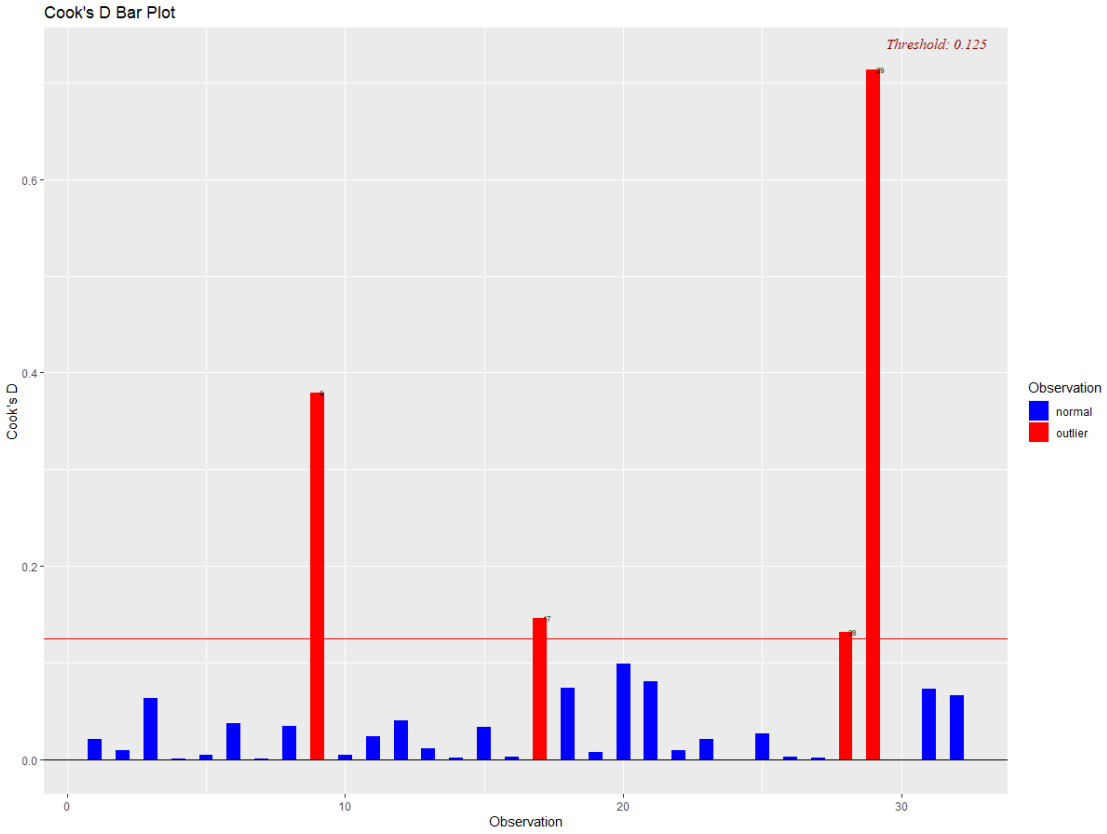
Διάγραμμα 1: h_{ii} για τις 32 παρατηρήσεις

```
> hatvalues(mod1)
        1         2         3         4         5         6         7         8         9        10        11        12
0.3025065 0.2902207 0.2388171 0.2277394 0.1995118 0.2822841 0.3259181 0.3302312 0.7422870 0.4293261 0.3748573 0.3032810
       13        14        15        16        17        18        19        20        21        22        23        24
0.1921148 0.2236587 0.3744802 0.3090439 0.3066962 0.1789510 0.5119321 0.2328717 0.4334135 0.2180100 0.1744450 0.4080732
       25        26        27        28        29        30        31        32
0.2053054 0.1421645 0.6232257 0.4310982 0.6632516 0.3910191 0.6427573 0.2905077
```

Πίνακας 6: h_{ii} για τις 32 παρατηρήσεις

Σύμφωνα με το κριτήριο h_{ii} μία παρατήρηση i θεωρείται σημείο επιρροής αν ισχύει h_{ii} > 2p/n, όπου p = k +1 (k ανεξάρτητες μεταβλητές) και n το πλήθος των παρατηρήσεων. Στη περίπτωση μας έχουμε p = 11 και n = 32 και άρα πρέπει h_{ii} > 0.6875. Κοιτάζοντας τον πίνακα 6 βλέπουμε πως μόνο η παρατήρηση 9 ικανοποιεί την συνθήκη και μπορεί να θεωρηθεί σημείο επιρροής. Ωστόσο, οι παρατηρήσεις 27, 29 και 31 έχουν hat values πολύ κοντά στο όριο και συγκεκριμένα 0.62, 0.66 και 0.64 αντίστοιχα. Καλό θα ήταν να έχουμε αυτά τα σημεία υπόψιν μας καθώς θα δούμε και άλλους ελέγχους.

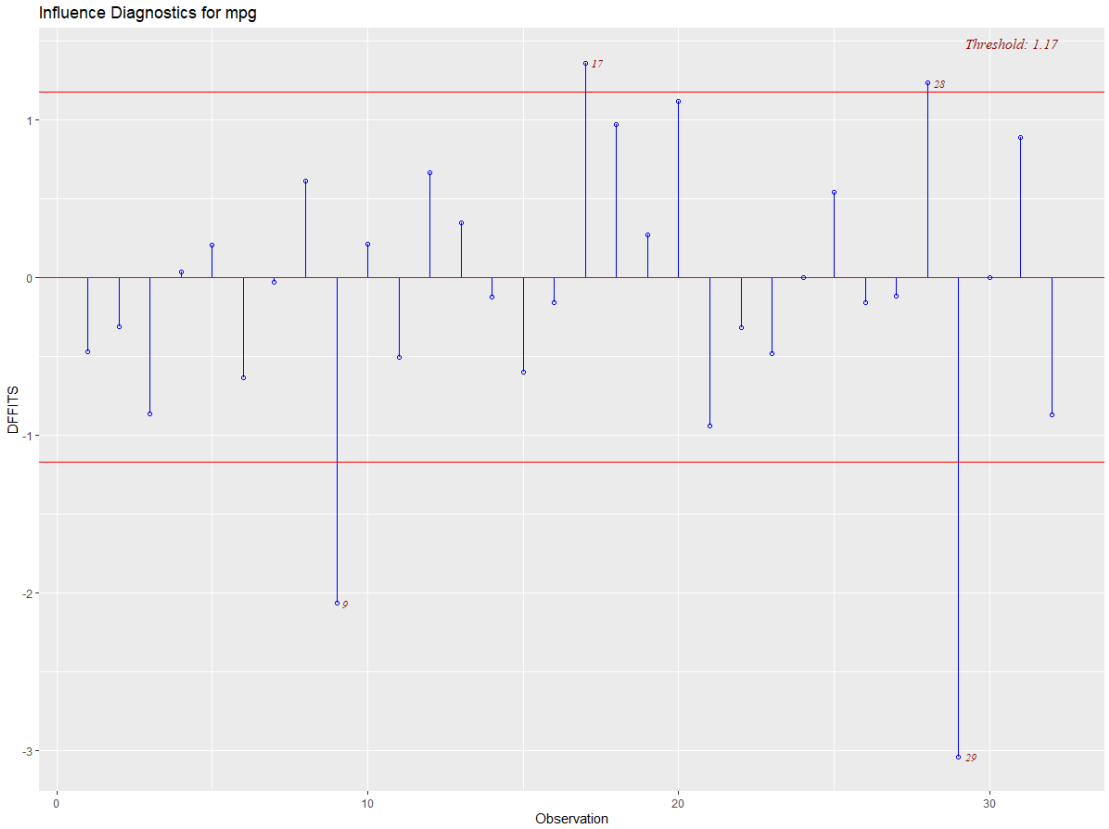
Στην συνέχεια θα εξετάσουμε το κριτήριο της απόστασης COOK, το οποίο μετρά το αποτέλεσμα της διαγραφής μίας δεδομένης παρατήρησης i. Η απόσταση COOK υπολογίζεται ως $D_i = \frac{r_i^2 h_{ii}}{p(1-h_{ii})}$, όπου r_i το τυποποιημένο υπόλοιπο και ισχύει πως αν D_i >> 1 η παρατήρηση i θεωρείται σημείο επιρροής. Παρακάτω παρατίθεται το αντίστοιχο διάγραμμα.



Διάγραμμα 2: Υπολογισμός απόστασης COOK για κάθε παρατήρηση i

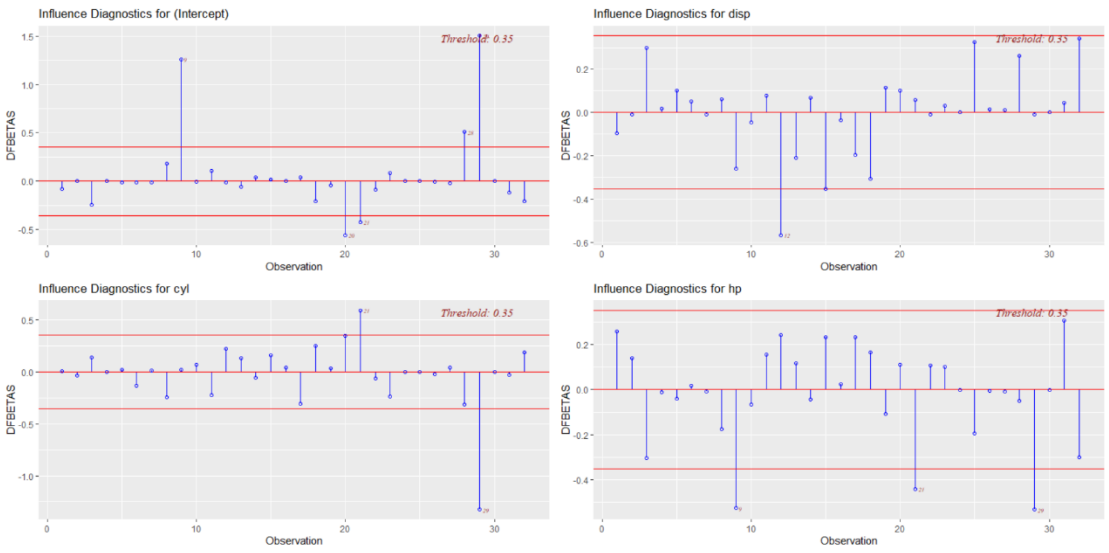
Παρατηρούμε πως καμία παρατήρηση δεν έχει απόσταση μεγαλύτερη του 1, όμως έχει οριστεί ένα όριο 0.125 το οποίο δείχνει ποιες παρατηρήσεις έχουν μεγάλη επιρροή. Σύμφωνα με αυτό το όριο ενδεχόμενα σημεία επιρροής αποτελούν οι παρατηρήσεις 9, 17, 28 και 29, ενώ συγκεκριμένα οι παρατηρήσεις 9 και 29 ξεπερνούν κατά πολύ το όριο.

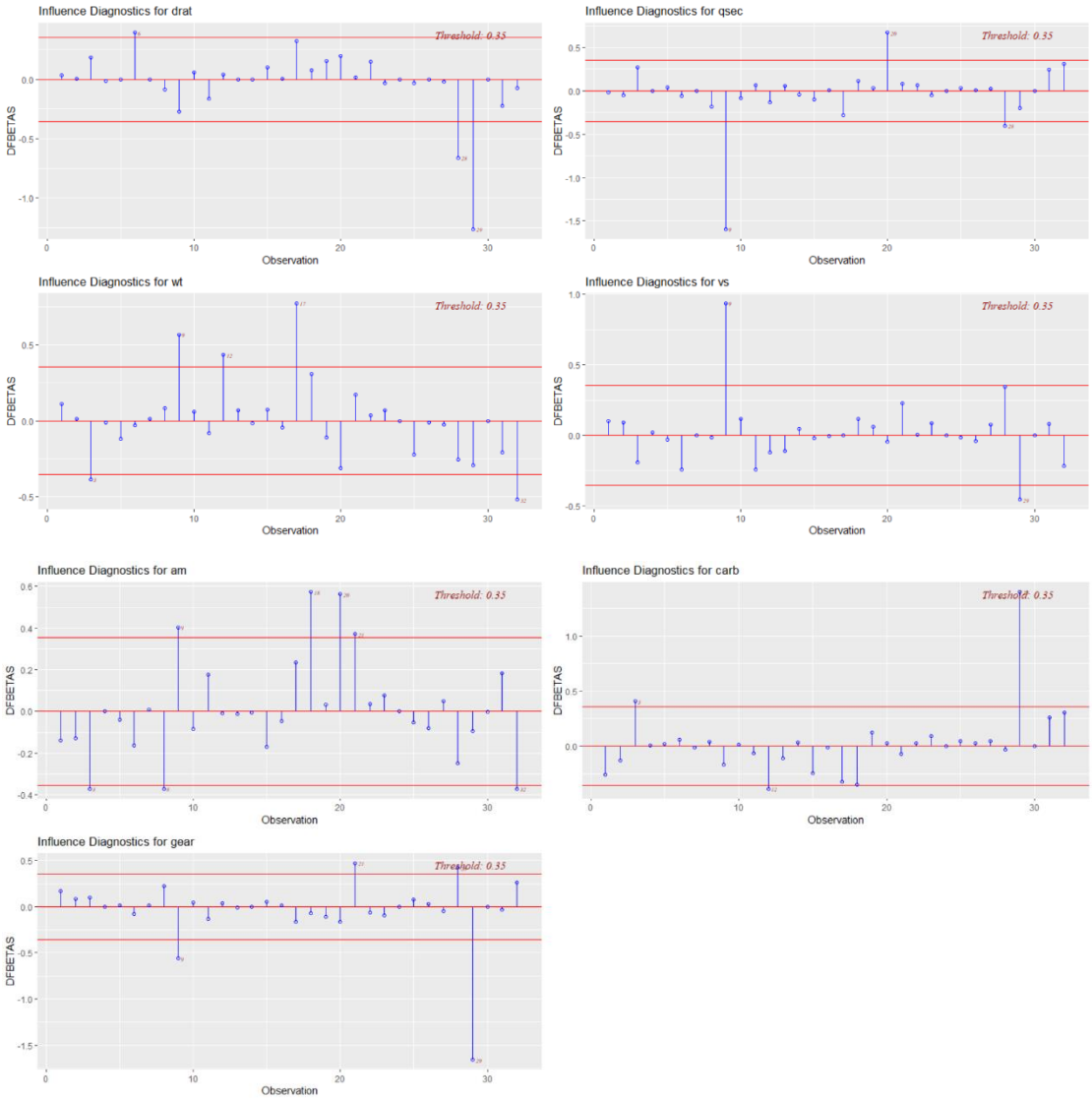
Στην συνέχεια θα εξετάσουμε το κριτήριο DFFITS, το οποίο υπολογίζει την μεταβολή της προβλεπόμενης τιμής y_i για μία παρατήρηση i, που προκύπτει όταν αυτή η παρατήρηση μείνει εκτός παλινδρόμησης. Σύμφωνα με το συγκεκριμένο κριτήριο μια παρατήρηση μπορεί να αποτελεί σημείο επιρροής αν ισχύει $|DFFITS_i|$ > $2\sqrt{p/n}$, δηλαδή σε αυτή την άσκηση $|DFFITS_i|$ > 1.1726. Στο παρακάτω διάγραμμα φαίνεται πως σημεία επιρροής αποτελούν οι παρατηρήσεις 9, 17, 28 και 29 με τις παρατηρήσεις 9 και 29 να απέχουν πολύ από το όριο (threshold), ενώ τις 17 και 28 να είναι πολύ κοντά στο όριο (threshold) .



Διάγραμμα 3: Υπολογισμός DFFITS για κάθε παρατήρηση i

Τέλος, θα εξετάσουμε το κριτήριο DFBETAS, το οποίο είναι παρόμοιο με το κριτήριο DFFITS με μόνη διαφορά πως υπολογίζεται για κάθε χαρακτηριστικό ξεχωριστά. Σε αυτή την περίπτωση μια παρατήρηση μπορεί να αποτελεί σημείο επιρροής αν $|DFBETAS_i| > 2/\sqrt{n}$, δηλαδή σε αυτή την περίπτωση $|DFBETAS_i| > 0.35$. Τα αποτελέσματα φαίνονται παρακάτω, όπου τις περισσότερες φορές εμφανίζονται σαν σημεία επιρροής οι παρατηρήσεις 9 και 29 αλλά και ορισμένες φορές οι παρατηρήσεις 17 και 28.





Πίνακας 7: Διαγράμματα DFBETAS για τις 11 μεταβλητές του μοντέλου

Ολοκληρώνοντας την εξέταση των μέτρων h_{ii}, απόσταση COOK, DFFITS και DFBETAS βλέπουμε πως οι παρατηρήσεις 9 και 29 εμφανίζονται να αποτελούν σημεία επιρροής με αρκετή διαφορά και συνεπώς πρέπει να προσέξουμε πως θα τις διαχειριστούμε. Επιπλέον, οι παρατηρήσεις 17 και 28 αποτελούν σημεία επιρροής με βάση τα μέτρα απόσταση COOK, DFFITS και DFBETAS, όμως βρίσκονται πιο κοντά στα αντίστοιχα όρια.

Ερώτημα 2

Σε αυτό το ερώτημα θα εξετάσουμε αν το μοντέλο με τις 10 επεξηγηματικές μεταβλητές είναι βέλτιστο χρησιμοποιώντας τεχνικές με ελέγχους F και t και τα κριτήρια R^2 , $R^2_{predict}$, $\bar{R}^2$, C_p και AIC.

Ξεκινώντας θα χρησιμοποιήσουμε το κριτήριο AIC, σύμφωνα με το οποίο το τελικό μοντέλο πρέπει να έχει όσο τον δυνατόν μικρότερο AIC. Το κριτήριο AIC μπορεί να χρησιμοποιηθεί με τρεις διαφορετικούς τρόπους, οι οποίοι θα περιγραφούν παρακάτω μαζί με τα αντίστοιχα αποτελέσματα. Είναι σκόπιμο να αναφέρουμε πως σε κάθε βήμα έχουμε επιλέξει να φαίνεται και ο έλεγχος F, παρόλο που η απόφαση για το τελικό μοντέλο από την R γίνεται μόνο με την χρήση του κριτηρίου AIC.

Ο πρώτος τρόπος να χρησιμοποιήσουμε το AIC είναι η διαδικασία της διαδοχικής πρόσθεσης (Forward AIC) όπου ξεκινάμε από το μοντέλο που έχει μόνο την μεταβλητή mpg και προσθέτουμε κάθε φορά την μεταβλητή που οδηγεί σε μικρότερο AIC. Η διαδικασία σταματά όταν η προσθήκη μίας επιπλέον μεταβλητής οδηγεί σε μεγαλύτερο AIC. Τα αποτελέσματα με την χρήση του Forward AIC φαίνονται παρακάτω:

```
Start:  AIC=115.94
mpg ~ 1

       Df Sum of Sq     RSS     AIC F value    Pr(>F)
+ wt    1    847.73  278.32  73.217 91.3753 1.294e-10 ***
+ cyl   1    817.71  308.33  76.494 79.5610 6.113e-10 ***
+ disp  1    808.89  317.16  77.397 76.5127 9.380e-10 ***
+ hp    1    678.37  447.67  88.427 45.4598 1.788e-07 ***
+ drat  1    522.48  603.57  97.988 25.9696 1.776e-05 ***
+ vs    1    496.53  629.52  99.335 23.6622 3.416e-05 ***
+ am    1    405.15  720.90 103.672 16.8603  0.000285 ***
+ carb  1    341.78  784.27 106.369 13.0736  0.001084 **
+ gear  1    259.75  866.30 109.552  8.9951  0.005401 **
+ qsec  1    197.39  928.66 111.776  6.3767  0.017082 *
<none>              1126.05 115.943
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=73.22
mpg ~ wt

       Df Sum of Sq    RSS    AIC F value   Pr(>F)
+ cyl   1    87.150 191.17 63.198 13.2203 0.001064 **
+ hp    1    83.274 195.05 63.840 12.3813 0.001451 **
+ qsec  1    82.858 195.46 63.908 12.2933 0.001500 **
+ vs    1    54.228 224.09 68.283  7.0177 0.012926 *
+ carb  1    44.602 233.72 69.628  5.5343 0.025646 *
+ disp  1    31.639 246.68 71.356  3.7195 0.063620 .
<none>              278.32 73.217
+ drat  1     9.081 269.24 74.156  0.9781 0.330854
+ gear  1     1.137 277.19 75.086  0.1189 0.732668
+ am    1     0.002 278.32 75.217  0.0002 0.987915
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=63.2
mpg ~ wt + cyl

       Df Sum of Sq    RSS    AIC F value Pr(>F)
+ hp    1   14.5514 176.62 62.665  2.3069 0.1400
+ carb  1   13.7724 177.40 62.805  2.1738 0.1515
<none>              191.17 63.198
+ qsec  1   10.5674 180.60 63.378  1.6383 0.2111
+ gear  1    3.0281 188.14 64.687  0.4507 0.5075
+ disp  1    2.6796 188.49 64.746  0.3980 0.5332
+ vs    1    0.7059 190.47 65.080  0.1038 0.7497
+ am    1    0.1249 191.05 65.177  0.0183 0.8933
+ drat  1    0.0010 191.17 65.198  0.0001 0.9903
```

```
Step:  AIC=62.66
mpg ~ wt + cyl + hp

       Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>              176.62 62.665
+ am    1    6.6228 170.00 63.442  1.0519 0.3142
+ disp  1    6.1762 170.44 63.526  0.9784 0.3314
+ carb  1    2.5187 174.10 64.205  0.3906 0.5372
+ drat  1    2.2453 174.38 64.255  0.3477 0.5603
+ qsec  1    1.4010 175.22 64.410  0.2159 0.6459
+ gear  1    0.8558 175.76 64.509  0.1315 0.7197
+ vs    1    0.0599 176.56 64.654  0.0092 0.9245
```

Πίνακας 8: Αποτελέσματα με την χρήση Forward AIC

```
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
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.2893 -1.5512 -0.4684  1.5743  6.1004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
cyl          -1.5078     0.4147  -3.636 0.001064 **
wt           -3.1910     0.7569  -4.216 0.000222 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.568 on 29 degrees of freedom
Multiple R-squared:  0.8302,    Adjusted R-squared:  0.8185
F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

Πίνακας 9: Summary μοντέλων mpg ~ wt + cyl + hp και mpg ~ cyl + wt

Ξεκινάμε με AIC = 115.94 και προσθέτονται διαδοχικά οι μεταβλητές wt, cyl και hp. Το μοντέλο καταλήγει να περιέχει μόνο τις μεταβλητές wt, cyl και hp και συγκεκριμένα δίνεται από την σχέση mpg = 38.752 -3.167wt -0.942cyl -0.018hp και έχει AIC = 62.665. Ωστόσο, στο τελευταίο βήμα προστίθεται η μεταβλητή hp, η οποία με βάση τον έλεγχο F δεν είναι στατιστικά σημαντική. Η μεταβλητή αυτή προστίθεται καθώς μειώνει το AIC, αλλά η μείωση είναι ελάχιστη (από AIC = 63.2 πηγαίνουμε σε AIC = 62.665). Αξίζει λοιπόν να σημειώσουμε πως με το κριτήριο F το μοντέλο θα περιείχε μόνο τις μεταβλητές wt και cyl και θα δινόταν από την σχέση mpg = 39.6863 -3.191wt – 1.508cyl με AIC = 63.2. Στην συνέχεια θα δούμε και τους υπόλοιπους 2 τρόπους χρήσης του AIC, καθώς και άλλα κριτήρια για να αποφασίσουμε πιο μοντέλο είναι το βέλτιστο.

Ο δεύτερος τρόπος να χρησιμοποιήσουμε το AIC είναι η διαδικασία της διαδοχικής αφαίρεσης (Backward AIC) όπου ξεκινάμε από το μοντέλο που έχει όλες τις μεταβλητές και αφαιρούμε διαδοχικά την μεταβλητή που θα οδηγήσει σε μικρότερο AIC. Η διαδικασία σταματά όταν η αφαίρεση μίας επιπλέον μεταβλητής οδηγεί σε μεγαλύτερο AIC. Το τελευταίο βήμα με την χρήση του Backward AIC φαίνεται παρακάτω:

```
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              169.29 61.307
- am    1    26.178 195.46 63.908  4.3298 0.0467155 *
- qsec  1   109.034 278.32 75.217 18.0343 0.0002162 ***
- wt    1   183.347 352.63 82.790 30.3258 6.953e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Πίνακας10: Τελευταίο βήμα με την χρήση backward AIC

```
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
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.3962 -2.1431 -0.2129  1.4915  5.7486

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.7462     5.2521   3.760 0.000765 ***
qsec          0.9292     0.2650   3.506 0.001500 **
wt           -5.0480     0.4840 -10.430 2.52e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.596 on 29 degrees of freedom
Multiple R-squared:  0.8264,    Adjusted R-squared:  0.8144
F-statistic: 69.03 on 2 and 29 DF,  p-value: 9.395e-12
```

Πίνακας 11: Summary μοντέλων mpg ~ wt + qsec + am και mpg ~ wt + qsec

Παρατηρούμε πως το μοντέλο καταλήγει να περιέχει τις μεταβλητές wt, qsec και am και συγκεκριμένα δίνεται από την σχέση mpg = 9.618 -3.916wt +1.226qsec +2.936am και έχει AIC = 61.307. Ωστόσο, με βάση τον έλεγχο F αξίζει να δούμε τι θα συμβεί στο μοντέλο αν αφαιρεθεί και η μεταβλητή am, η οποία θεωρείται οριακά στατιστικά σημαντική. Αν αφαιρεθεί αυτή η μεταβλητή τότε το μοντέλο θα εξαρτάται μόνο από τις μεταβλητές wt και qsec και συγκεκριμένα θα δίνεται από την σχέση mpg = 19.746 + 0.929qsec -5.048wt και θα έχει AIC = 63.908. Θα κρατήσουμε και αυτό το μοντέλο υπόψιν μας καθώς έχει μία λιγότερη μεταβλητή.

Ο τρίτος, και τελευταίος, τρόπος να χρησιμοποιήσουμε το AIC είναι η κατά βήματα εμπρός-πίσω επιλογή (Both AIC) όπου ξεκινάμε από το μοντέλο που έχει μόνο την μεταβλητή mpg και προσθέτουμε κάθε φορά την μεταβλητή που οδηγεί σε μικρότερο AIC και στην συνέχεια ελέγχουμε αν πρέπει να αφαιρεθεί. Ο τρόπος αυτός συνδυάζει τον Forward και Backward AIC. Τα αποτελέσματα με την χρήση του Both AIC είναι ακριβώς τα ίδια με την χρήση του Forward AIC στην άσκηση μας. Καταλήγει δηλαδή στο μοντέλο mpg = 38.752 -3.167wt -0.942cyl -0.018hp με την μεταβλητή hp στατιστικά μη σημαντική με βάση τον έλεγχο F.

Στην συνέχεια παραθέτω συγκεντρωτικά για τα τέσσερα μοντέλα τις μετρήσεις όσον αφορά τα κριτήρια R^2 , $R^2_{predict}$, $\bar{R}^2$, C_p και AIC. Γενικά για την επιλογή θέλουμε οι συντελεστές R^2 , $R^2_{predict}$ και $\bar{R}^2$ να είναι υψηλοί και κοντά στην μονάδα και τα C_p και AIC να είναι όσο πιο μικρά γίνεται. Τα συγκεντρωτικά αποτελέσματα για τα προαναφερθέντα μοντέλα φαίνονται στον παρακάτω πίνακα, όπου για τιμές του AIC χρησιμοποίησα τα αποτελέσματα της βιβλιοθήκης της R olsrr, που συνυπολογίζει και το σταθερό.

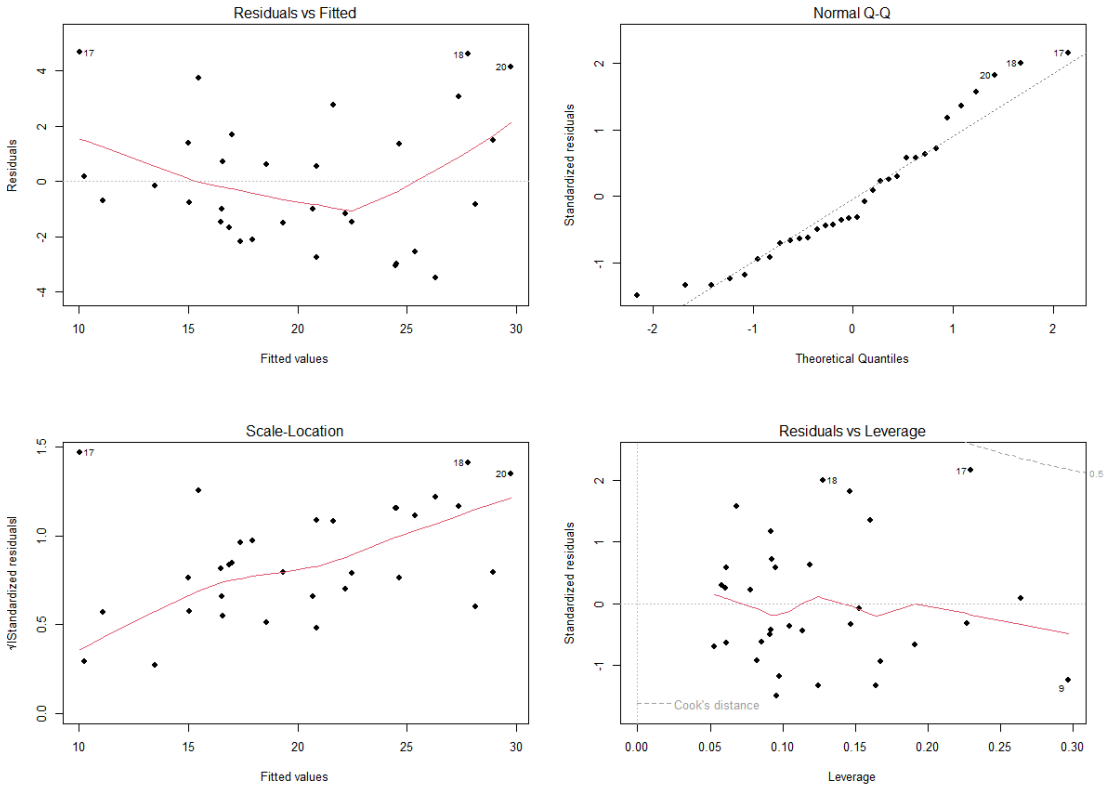
	mpg ~ wt + cyl + hp	mpg ~ cyl + wt	mpg ~ wt + qsec + am	mpg ~ wt + qsec
R^2	0.8432	0.8302	0.8497	0.8264
$\bar{R}^2$	0.8263	0.8185	0.8336	0.8144
$R^2_{predict}$	0.7958	0.7904	0.7946	0.7786
C_p	1.1469	1.2187	0.1026	1.8298
AIC	155.476	156.01	154.119	156.721

Πίνακας 12: Συγκεντρωτικά αποτελέσματα R^2 , $R^2_{predict}$, $\bar{R}^2$, C_p και AIC

Λαμβάνοντας υπόψιν τα τέσσερα αυτά μοντέλα επιλέγω για βέλτιστο το μοντέλο mpg = 9.618 -3.916wt +1.226qsec +2.936am καθώς έχει το μικρότερο AIC, το μικρότερο C_p και το μεγαλύτερο R^2. Σε σχέση με το μοντέλο mpg = 19.746 + 0.929qsec -5.048wt η μεταβλητή am θεωρείται οριακά στατιστικά σημαντική με βάση τον έλεγχο F, ενώ επίσης έχουμε μικρότερα AIC και C_p καθώς και μεγαλύτερο R^2. Για αυτό το λόγο επιλέγω το μοντέλο που έχει την μία παραπάνω μεταβλητή.

Ερώτημα 3

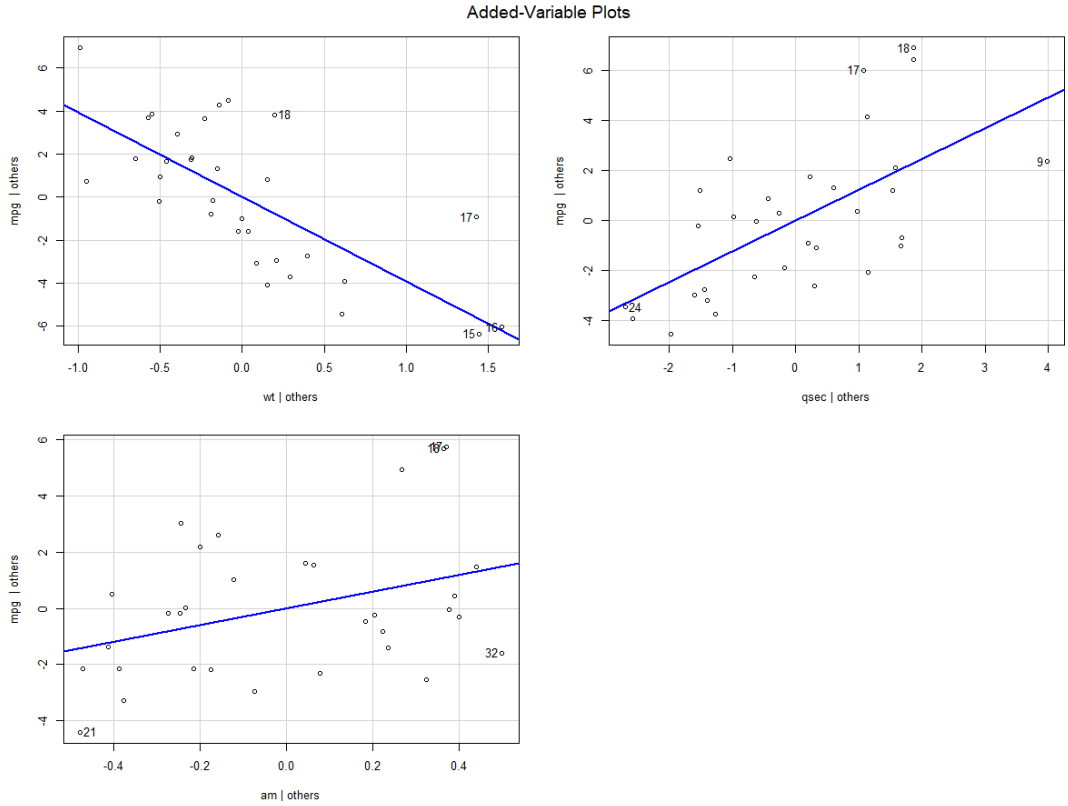
Το τελικό μοντέλο είναι το εξής: mpg = 9.618 -3.916wt +1.226qsec +2.936am δηλαδή mpg ~ wt + qsec + am. Πρώτα θα εξετάσουμε αν τηρούνται οι προϋποθέσεις του μοντέλου με βάση τα υπόλοιπα, όπως εξετάσαμε και στο πρώτο ερώτημα. Τα διαγράμματα Residuals vs Fitted Values, Normal Q-Q, Scale-Location και Residuals vs Leverage είναι τα εξής:



Πίνακας 13: Διαγράμματα τυποποιημένων υπολοίπων για το μοντέλο mpg ~ wt + qsec + am

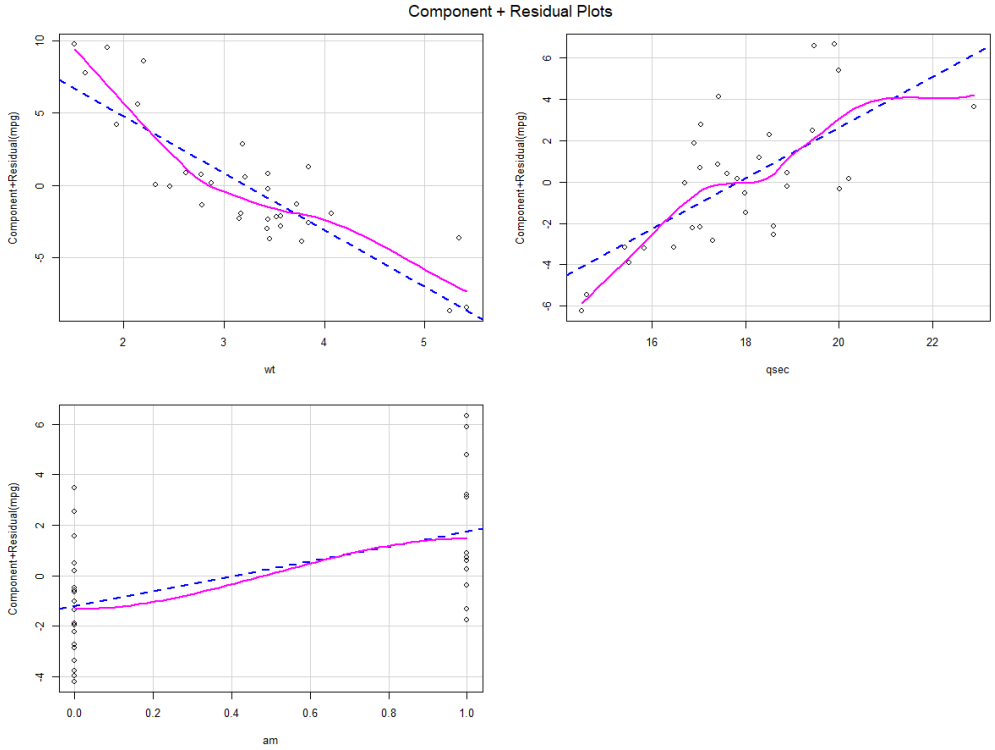
Σε γενικές γραμμές οι προϋποθέσεις τηρούνται αλλά αυτό που μας παραξενεύει είναι πως στο διάγραμμα Scale-Location η γραμμή που προκύπτει δεν παραπέμπει σε οριζόντια αλλά έχει ανοδική κλήση. Αυτό δείχνει πως ίσως χρειάζεται κάποιος μετασχηματισμός στο μοντέλο.

Στη συνέχεια θα σχεδιάσουμε τις added variable plots ώστε να δούμε αν και οι τρεις μεταβλητές wt, qsec και am είναι απαραίτητες. Αφού σχεδιάσουμε τα διαγράμματα που φαίνονται παρακάτω παρατηρούμε πως όλες οι μεταβλητές χρειάζονται στο μοντέλο.



Πίνακας 13: Διαγράμματα added variable plots για τις μεταβλητές wt, qsec, και am

Στη συνέχεια θα σχεδιάσουμε τα διαγράμματα των μερικών υπολοίπων ώστε να δούμε αν χρειάζεται κάποιος μετασχηματισμός στις μεταβλητές που εισέρχονται στο μοντέλο. Τα διαγράμματα φαίνονται παρακάτω:



Πίνακας 14: Διαγράμματα μερικών υπολοίπων για τις μεταβλητές wt, qsec, και am

Παρατηρούμε πως οι μεταβλητές wt και qsec χρειάζονται κάποιο μετασχηματισμό καθώς οι αντίστοιχες ροζ ευθείες δεν προσομοιάζουν αρκετά καλά τις μπλε ευθείες, ενώ η μεταβλητή am δεν χρειάζεται κάποιο μετασχηματισμό. Ύστερα από αρκετές δοκιμές με γνώμονα τα R^2 και AIC κατέληξα πως το καλύτερο μοντέλο προκύπτει όταν mpg ~ log(wt)+log(qsec)+am, καθώς R^2 = 0.886 και AIC = 145.1938. Ωστόσο, στην συγκεκριμένη περίπτωση μπορούμε να παρατηρήσουμε πως το am δεν είναι στατιστικά σημαντικό, το οποίο φαίνεται από το summary του μοντέλου.

```
Call:
lm(formula = mpg ~ wtlog + qseclog + am)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2904 -1.2769 -0.3701  0.9305  4.8255

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -20.884     15.077  -1.385  0.17694
wtlog        -14.029      2.050  -6.843 1.96e-07 ***
qseclog       19.468      4.658   4.179  0.00026 ***
am             1.721      1.320   1.303  0.20303
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.139 on 28 degrees of freedom
Multiple R-squared:  0.8863,    Adjusted R-squared:  0.8741
F-statistic: 72.72 on 3 and 28 DF,  p-value: 2.486e-13
```

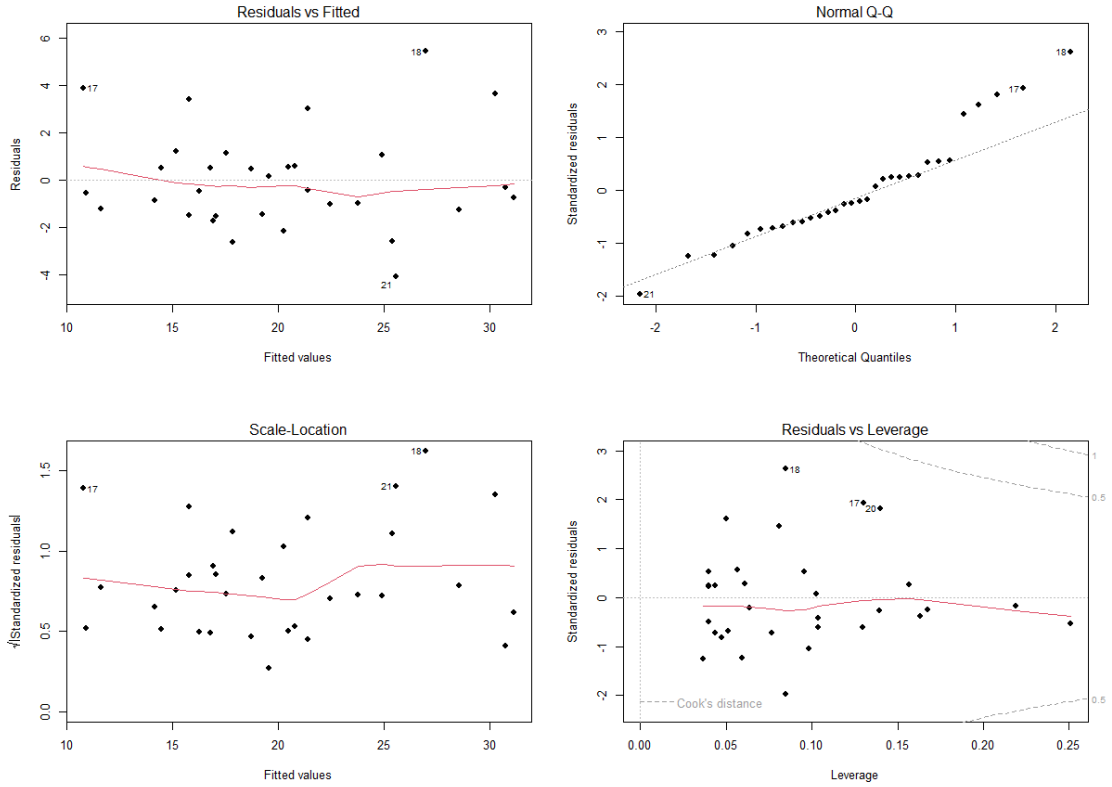
Πίνακας 15: Summary μοντέλου mpg ~ log(wt) + log(qsec) + am

Αν διώξουμε το am και υποθέσουμε πως mpg ~ log(wt)+log(qsec) τότε προκύπτει R^2 = 0.8794 και AIC = 145.079. Ο συντελεστής R^2 μειώνεται ελάχιστα κάτι το οποίο δεν είναι επιθυμητό, ενώ ο συντελεστής AIC μειώνεται ελάχιστα κάτι το οποίο είναι επιθυμητό. Δεδομένου πως το μοντέλο μας περιέχει μια λιγότερη μεταβλητή και προβλέπει εξίσου καλά επιλέγουμε να διώξουμε την am. Σε σχέση με το τελικό μοντέλο mpg ~ wt + qsec + am, αλλά και το αρχικό mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb έχουμε βελτιώσει το R^2 και μειώσει το AIC. Συνεπώς το νέο βελτιωμένο μας μοντέλο είναι το mpg = -8.368 -16.167log(wt) +16.195log(qsec).

```
Call:
lm(formula = mpg ~ wtlog + qseclog)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0690 -1.3081 -0.4533  0.7136  5.4353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.368     11.763  -0.711 0.482520
wtlog        -16.167      1.245 -12.987  1.3e-13 ***
qseclog       16.195      3.971   4.079 0.000323 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.164 on 29 degrees of freedom
Multiple R-squared:  0.8794,    Adjusted R-squared:  0.871
F-statistic: 105.7 on 2 and 29 DF,  p-value: 4.808e-14
```

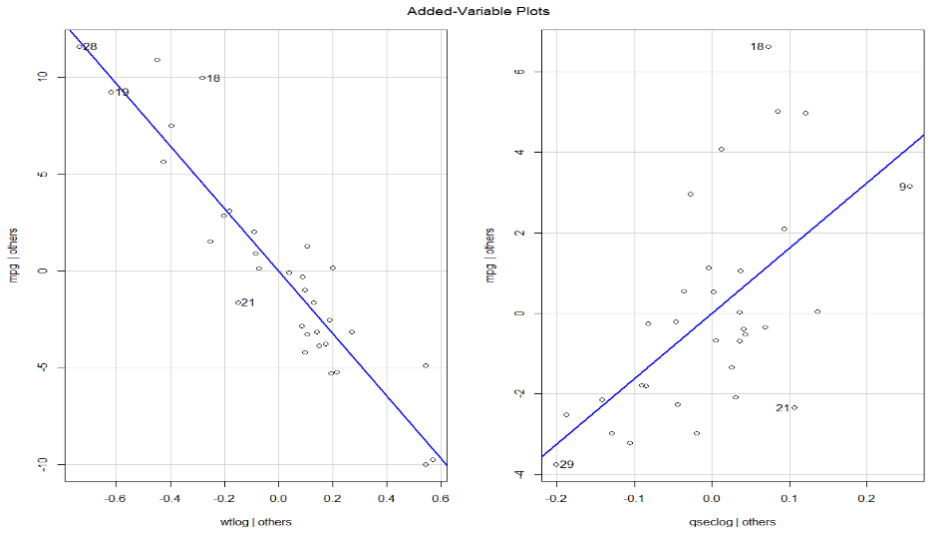
Πίνακας 16: Summary νέου βελτιωμένου μοντέλου mpg ~ log(wt) + log(qsec)

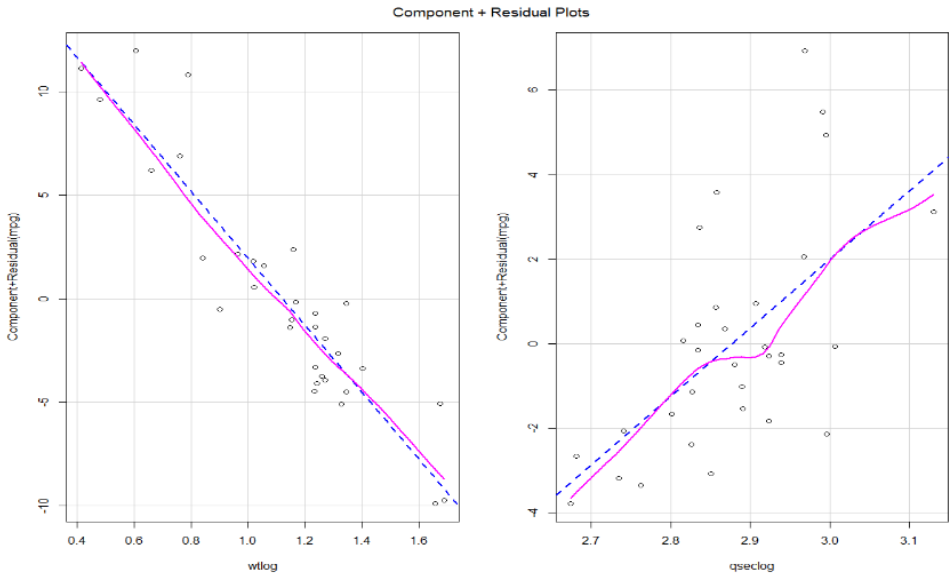
Για το νέο αυτό μοντέλο παρουσιάζονται παρακάτω τα διαγράμματα Residuals vs Fitted Values, Normal Q-Q, Scale-Location και Residuals vs Leverage στα οποία παρατηρούμε πως σε σχέση με προηγουμένως έχουμε μία πολύ καλύτερη κατάσταση στα διαγράμματα Residuals vs Fitted και Scale-Location, ενώ στο διάγραμμα Normal Q-Q ξεφεύγουν 5 παρατηρήσεις όπως και πριν αλλά λίγο περισσότερο.



Πίνακας 17: Διαγράμματα τυποποιημένων υπολοίπων για το μοντέλο mpg ~ log(wt) + log(qsec)

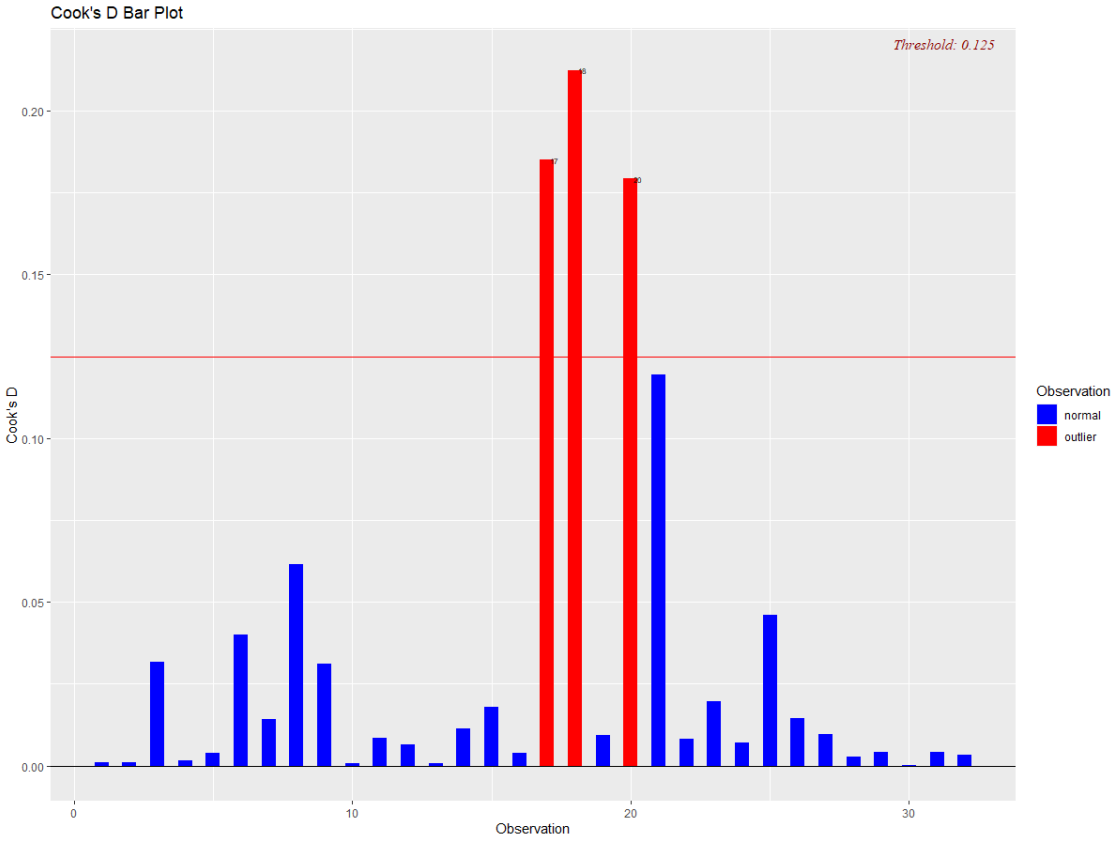
Στην συνέχεια παραθέτονται τα added variable plots και τα component & residual plots, όπου παρατηρούμε πως οι μεταβλητές log(wt) και log(qsec) είναι απαραίτητες για το μοντέλο και πως οι αντίστοιχες ροζ ευθείες τους προσομοιάζουν σε καλύτερο βαθμό τις μπλε, το οποίο σημαίνει πως ο μετασχηματισμός ήταν πετυχημένος.





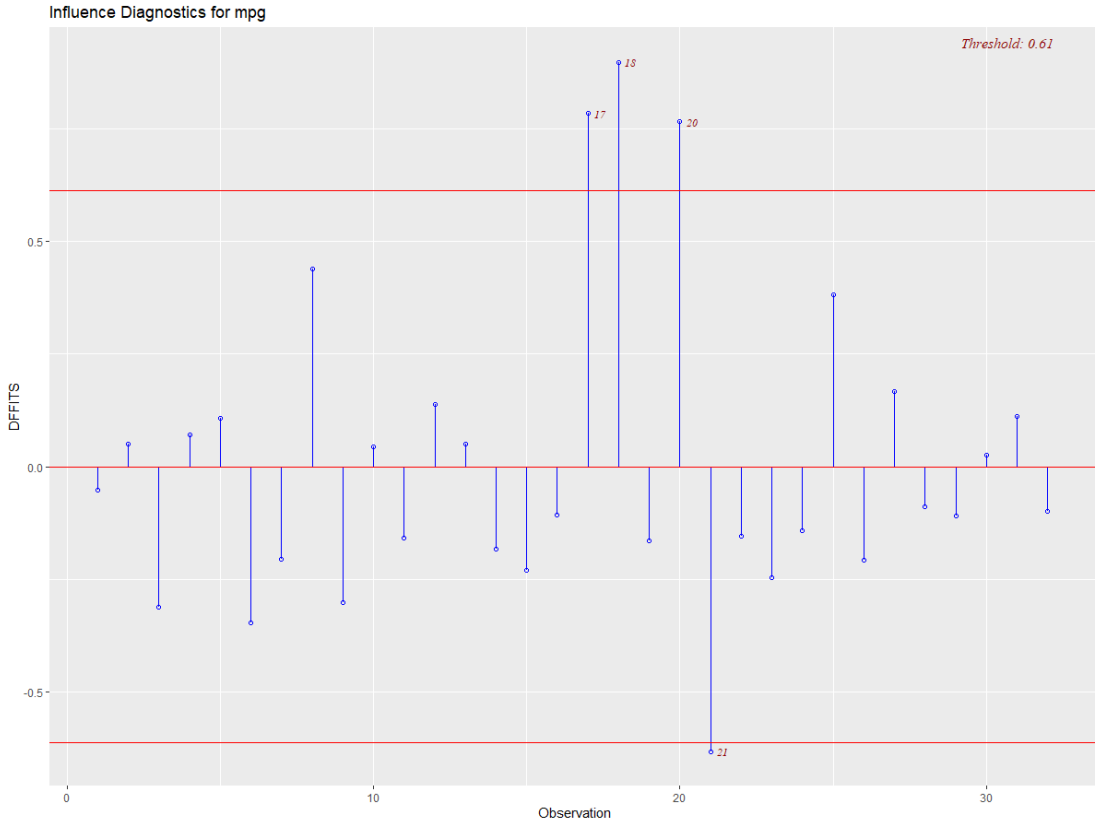
Πίνακας 18: Διαγράμματα added variable plots και μερικών υπολοίπων για τις μεταβλητές log(wt) και log(qsec)

Στην συνέχεια θα εξετάσουμε την πιθανή παρουσία άτυπων σημείων ή σημείων επιρροής με τα μέτρα απόσταση COOK, DFFITS και DFBETAS. Συγκεκριμένα έχουμε τα εξής αποτελέσματα.



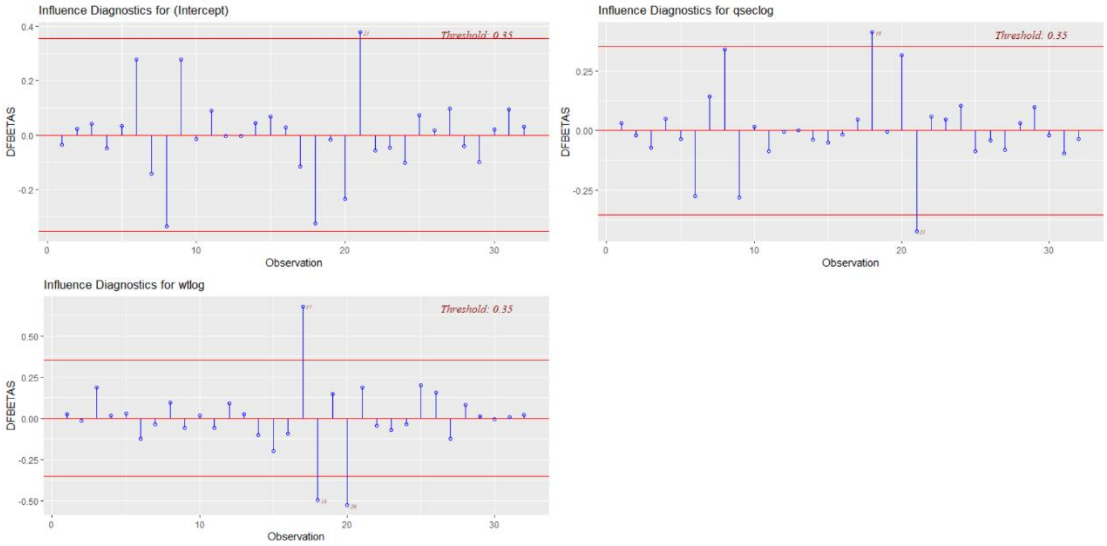
Διάγραμμα 4: Υπολογισμός απόστασης COOK για κάθε παρατήρηση i

Με βάση την απόσταση Cook παρατηρούμε πως και σε αυτή την περίπτωση καμία παρατήρηση δεν έχει απόσταση μεγαλύτερη του 1, όμως έχει οριστεί ένα όριο 0.125 το οποίο δείχνει ποιες παρατηρήσεις έχουν μεγάλη επιρροή. Σύμφωνα με αυτό το όριο ενδεχόμενα σημεία επιρροής αποτελούν οι παρατηρήσεις 17, 18 και 20. Ωστόσο, καμία από αυτές τις παρατηρήσεις δεν ξεπερνά κατά πολύ το όριο όπως γινόταν με τις παρατηρήσεις 9 και 29 του αρχικού μοντέλου των δέκα μεταβλητών στο αντίστοιχο διάγραμμα.



Διάγραμμα 5: Υπολογισμός DFFITS για κάθε παρατήρηση i

Με βάση το κριτήριο DFFITS για το νέο βελτιωμένο μοντέλο έχουμε ως όριο το $|DFFITS_i| > 0.61$. Σύμφωνα με το διάγραμμα σημεία επιρροής μπορούν να θεωρηθούν οι παρατηρήσεις 17, 18, 20 και 21. Ωστόσο, αυτές οι παρατηρήσεις απέχουν πολύ λίγο από το όριο (threshold) σε σχέση με το αρχικό μοντέλο, όπου στο αντίστοιχο διάγραμμα οι παρατηρήσεις 9 και 29 απέχουν πολύ από το αντίστοιχο όριο.



Πίνακας 19: Διαγράμματα DFBETAS για τις 3 μεταβλητές του νέου μοντέλου

Με βάση το κριτήριο DFBETAS για το νέο βελτιωμένο μοντέλο έχουμε ως όριο το $|DFBETAS_i| > 0.35$. Τα αποτελέσματα είναι παρόμοια με τον έλεγχο DFFITS, καθώς οι παρατηρήσεις 17, 18, 20 και 21 εμφανίζονται σαν πιθανά σημεία επιρροής ξεπερνώντας λίγο το όριο.

Συγκεντρωτικά όσον αφορά τα σημεία επιρροής τα αποτελέσματα είναι καλύτερα σε σχέση με το αρχικό μοντέλο των 10 ανεξάρτητων μεταβλητών καθώς οι παρατηρήσεις 17, 18, 20 και 21 μπορεί να θεωρηθούν σημεία επιρροής αλλά βρίσκονται πολύ κοντά στα όρια των αντίστοιχων μέτρων. Στο αρχικό μοντέλο οι παρατηρήσεις 9 και 29 ξεπερνούσαν κατά πολύ τα όρια των μέτρων ώστε να θεωρηθούν σημεία επιρροής, ενώ τα αντίστοιχα μέτρα αναδείκνυαν και τις παρατηρήσεις 17 και 28 ως πιθανά σημεία επιρροής, οι οποίες ξεπερνούσαν για λίγο τα όρια.

Στην συνέχεια θα υπολογίσουμε τα 95% διαστήματα εμπιστοσύνης για τους συντελεστές του νέου βελτιωμένου μοντέλου mpg ~ log(wt) + log(qsec), τα οποία είναι τα εξής:

	2.5 %	97.5 %
(Intercept)	-32.427003	15.69022
wtlog	-18.713320	-13.62102
qseclog	8.073995	24.31555

Πίνακας 20: 95% Διαστήματα εμπιστοσύνης για τους συντελεστές του mpg ~ log(wt) + log(qsec)

Αξίζει να σημειωθεί πως αν είχαμε κρατήσει την am θα πρόκυπτε διάστημα εμπιστοσύνης που περιέχει το 0 και συνεπώς η am δεν είναι στατιστικά σημαντική, όπως προέκυψε και από το έλεγχο t. Αυτό φαίνεται παρακάτω:

	2.5 %	97.5 %
(Intercept)	-51.7674897	9.999409
wtlog	-18.2289430	-9.829782
qseclog	9.9258228	29.010800
am	-0.9835927	4.425810

Πίνακας 21: 95% Διαστήματα εμπιστοσύνης για τους συντελεστές του mpg~log(wt)+log(qsec)+am

Στην συνέχεια θα εκτιμήσουμε την πρόβλεψη μίας παρατήρησης y με το νέο βελτιωμένο μοντέλο. Έστω ότι έχουμε δεδομένα για ένα μοντέλο αυτοκινήτου, το οποίο έχει wt = 3.642 και qsec = 17.28 και θέλουμε να εκτιμήσουμε το mpg. Αρχικά κάνουμε μετατροπή του 3.642 σε 1.2925 (ln(wt)) και του qsec σε 2.8495 (ln(qsec)), καθώς αυτά έχουμε χρησιμοποιήσει στο μοντέλο μας, και έπειτα με χρήση της R έχουμε τα εξής αποτελέσματα.

```
> newdata = data.frame(wtlog=1.2925,qseclog=2.8495)
> predict(final, newdata, interval="predict", level=.95)
       fit      lwr      upr
1 16.88254 12.36458 21.40049
```

Πίνακας 22: 95% Διαστήματα εμπιστοσύνης πρόβλεψης mpg για wt = 3.642 και qsec = 17.28

Προκύπτει λοιπόν πως το 95% διάστημα εμπιστοσύνης για την τιμή του mpg είναι (12.36458, 21.40049)

Τέλος, όσον αφορά την ερμηνεία των συντελεστών του νέου βελτιωμένου μοντέλου έχουμε το εξής:

```
Call:
lm(formula = mpg ~ wtlog + qseclog)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0690 -1.3081 -0.4533  0.7136  5.4353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.368     11.763  -0.711 0.482520
wtlog        -16.167      1.245 -12.987  1.3e-13 ***
qseclog       16.195      3.971   4.079 0.000323 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.164 on 29 degrees of freedom
Multiple R-squared:  0.8794,    Adjusted R-squared:  0.871
F-statistic: 105.7 on 2 and 29 DF,  p-value: 4.808e-14
```

Πίνακας 23: Summary νέου βελτιωμένου μοντέλου mpg ~ log(wt) + log(qsec)

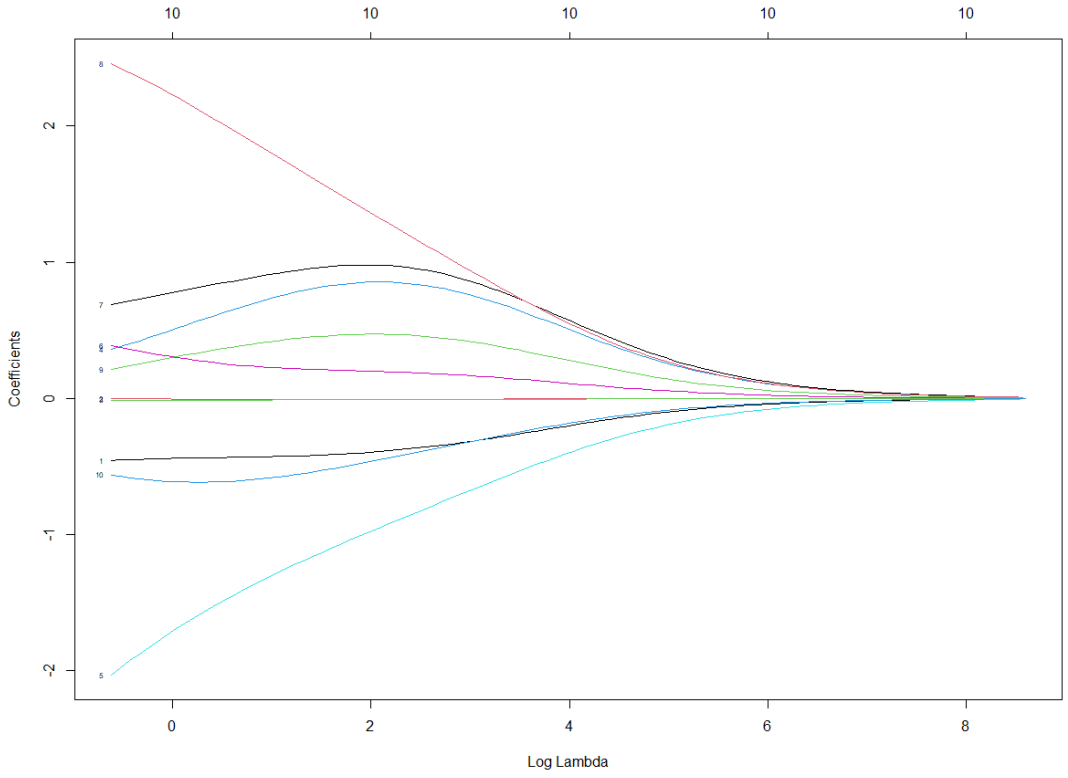
Αν αυξηθεί το wt κατά 10%, δηλαδή ή νέα του τιμή γίνει 1.1wt, και η τιμή του qsec παραμείνει σταθερή τότε η αναμενόμενη μεταβολή του mpg είναι $\hat{\beta}$ln(1.1) = -1.541 ($\hat{\beta}$ = -16.167).

Αν αυξηθεί το qsec κατά 10%, δηλαδή ή νέα του τιμή γίνει 1.1qsec, και η τιμή του wt παραμείνει σταθερή τότε η αναμενόμενη μεταβολή του mpg είναι $\hat{\beta}$ln(1.1) = 1.544 ($\hat{\beta}$ = 16.195).

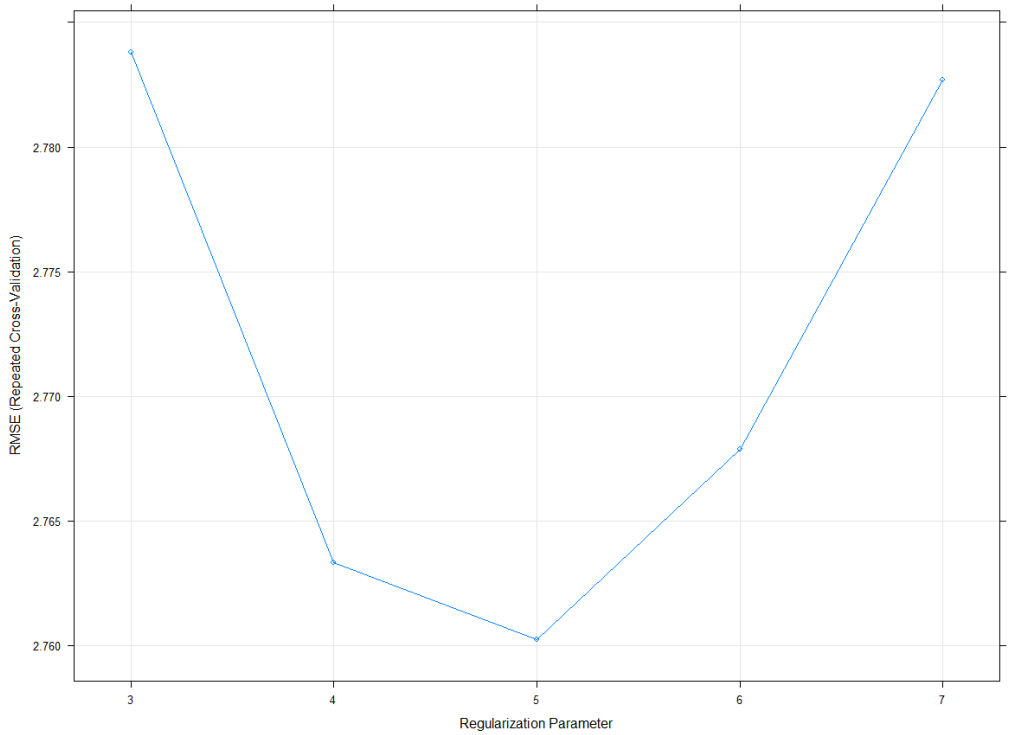
Προαιρετικό ερώτημα- Ridge και Lasso

Όπως είδαμε σε προηγούμενα ερωτήματα της εργασίας το αρχικό μας μοντέλο mpg ~ cyl + disp + hp + drat + wt + qsec + vs+ am + gear + carb έχει έντονο πρόβλημα πολυσυγγραμμικότητας. Ένας ακόμη τρόπος για να δώσουμε λύση στο πρόβλημα της πολυσυγγραμμικότητας είναι να χρησιμοποιήσουμε τις τεχνικές Ridge και Lasso. Η τεχνική Ridge συρρικνώνει τις τιμές των συντελεστών β σε μη μηδενικές τιμές κρατώντας όλες τις μεταβλητές στο μοντέλο, ενώ η τεχνική Lasso υπερτερεί της Ridge, καθώς πέρα από συρρίκνωση των τιμών των συντελεστών β μηδενίζει ορισμένες από αυτές υποδεικνύοντας με αυτό τον τρόπο το υποσύνολο των μεταβλητών που δε συμβάλλουν στο μοντέλο. Συγκεκριμένα, για τις τεχνικές Ridge και Lasso έχουμε τα εξής: $SSE_{Ridge} = \Sigma(y - \hat{y})^2 + \lambda\Sigma\beta^2$ και $SSE_{Lasso} = \Sigma(y - \hat{y})^2 + \lambda\Sigma|\beta|$ όπου το λ καλείται και ποινή (penalty).

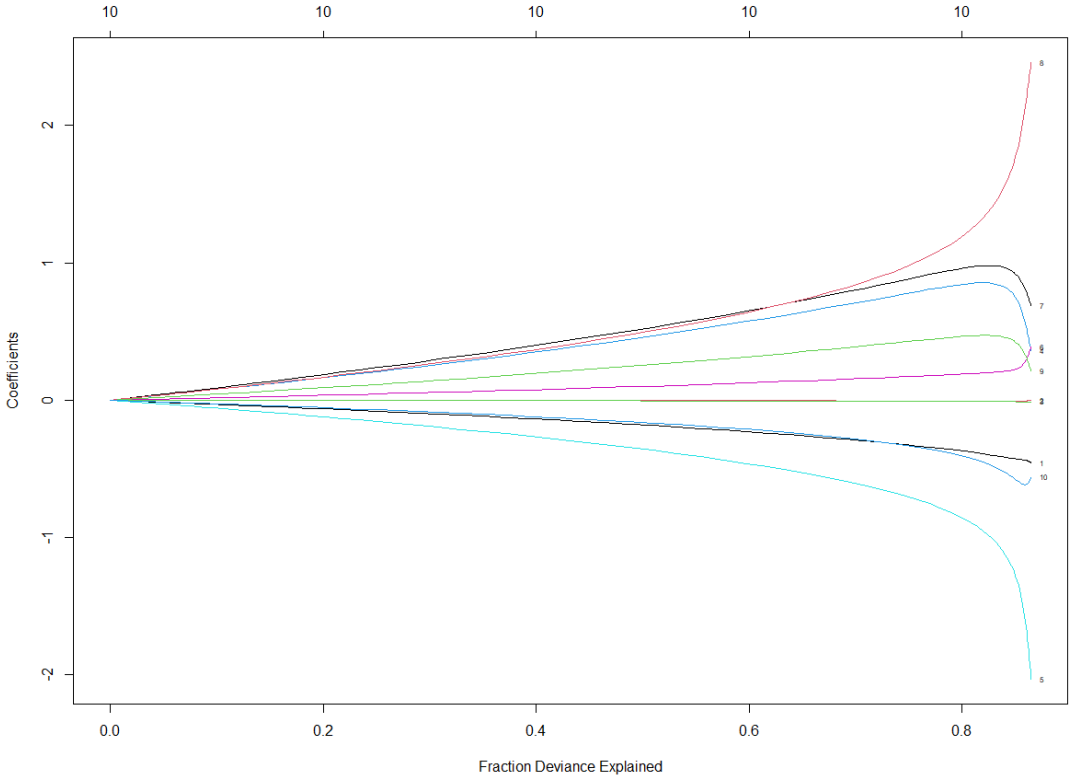
Αφού εφαρμόσουμε την τεχνική Ridge στα δεδομένα μας δημιουργούμε τα παρακάτω διαγράμματα:



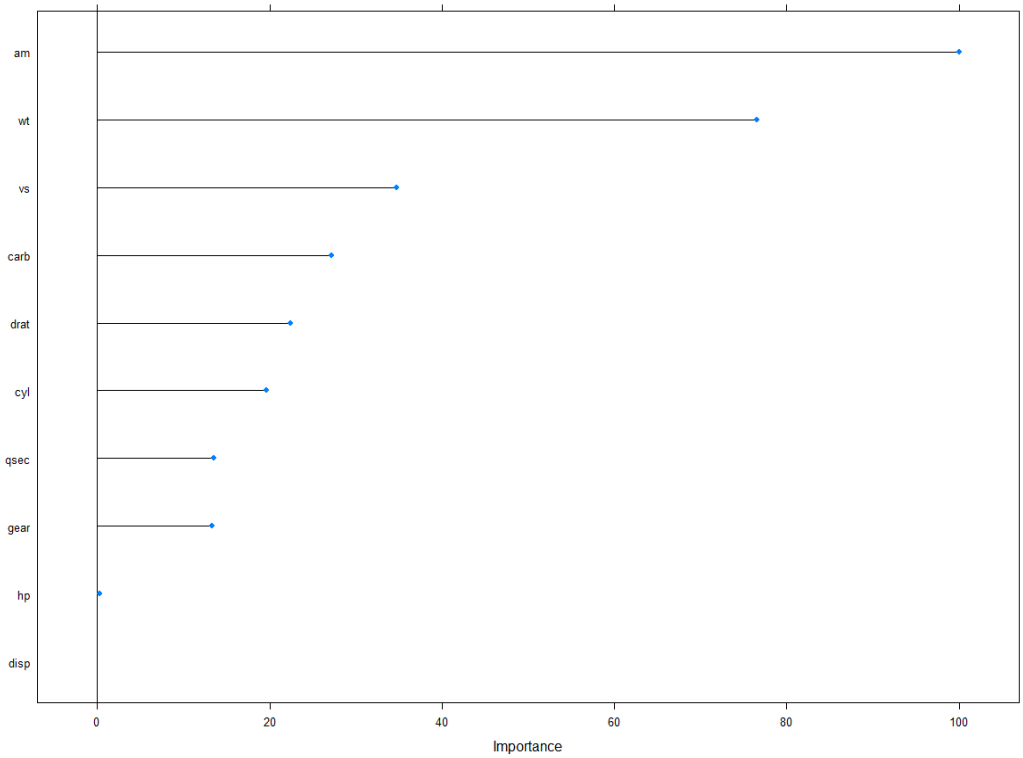
Διάγραμμα 5: Coefficients VS Log Lambda



Διάγραμμα 6: RMSE VS Regularization Parameter



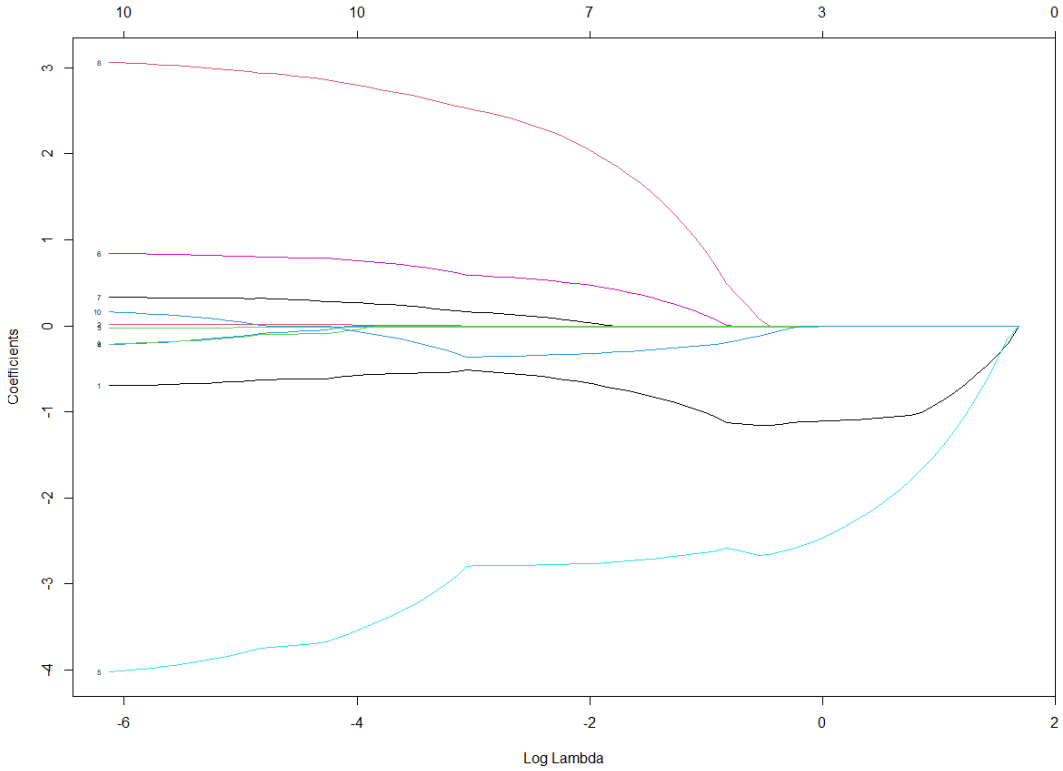
Διάγραμμα 7: Coefficients VS Fraction Deviance Explained



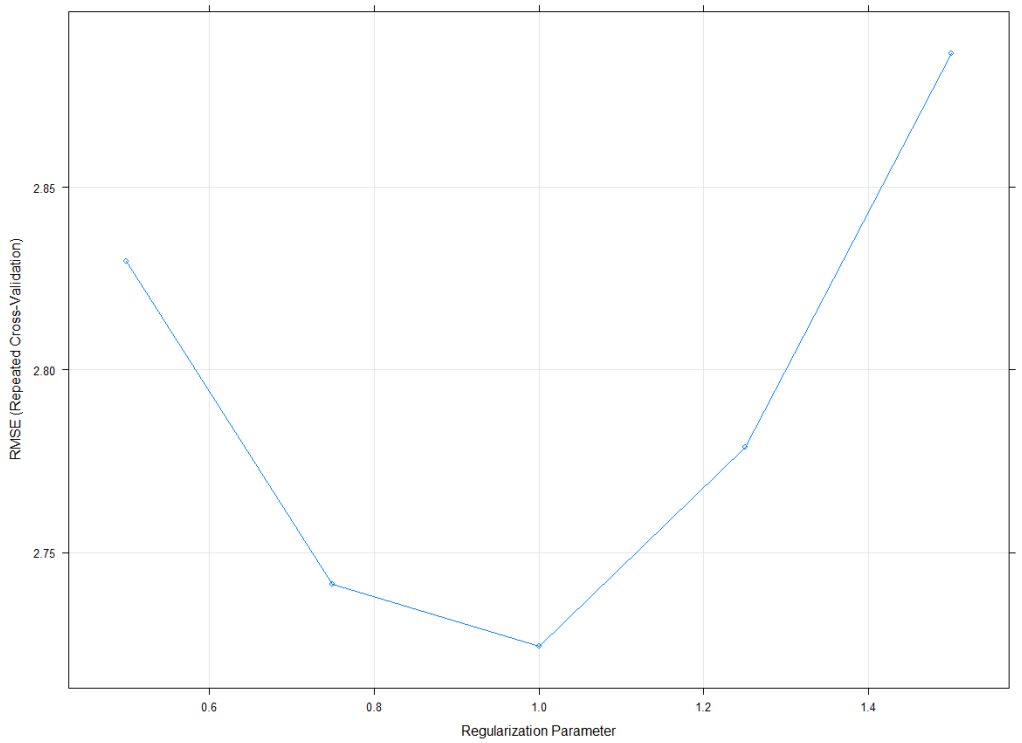
Διάγραμμα 8: Importance για κάθε μεταβλητή

Το διάγραμμα 5 μας βοηθάει να εντοπίσουμε που περίπου βρίσκεται το βέλτιστο λ και στην συνέχεια με το διάγραμμα 6 εντοπίζουμε το βέλτιστο λ, το οποίο σε αυτή την περίπτωση είναι ίσο με 5. Το βέλτιστο λ προκύπτει όταν ελαχιστοποιείται το RMSE, το οποίο όπως φαίνεται στο διάγραμμα 6 έχει καθοδική πορεία μέχρι το λ να γίνει 5 και ύστερα ανοδική πορεία. Στο διάγραμμα 7 παρατηρούμε πως και οι 10 μεταβλητές παραμένουν στο μοντέλο, καθώς όπως αναφέραμε η τεχνική Ridge οδηγεί μόνο στην συρρίκνωση τιμών συντελεστών β, ενώ στο διάγραμμα 8 βλέπουμε την σημασία που έχει κάθε μεταβλητή στο τελικό μοντέλο. Αυτό σημαίνει πως οι μεταβλητές am και wt έχουν υψηλούς συντελεστές β, ενώ οι μεταβλητές hp και disp έχουν πάρα πολύ χαμηλούς συντελεστές β και πρακτικά δεν παίζουν κανέναν ρόλο. Όλες οι υπόλοιπες μεταβλητές είναι κάπου ενδιάμεσα.

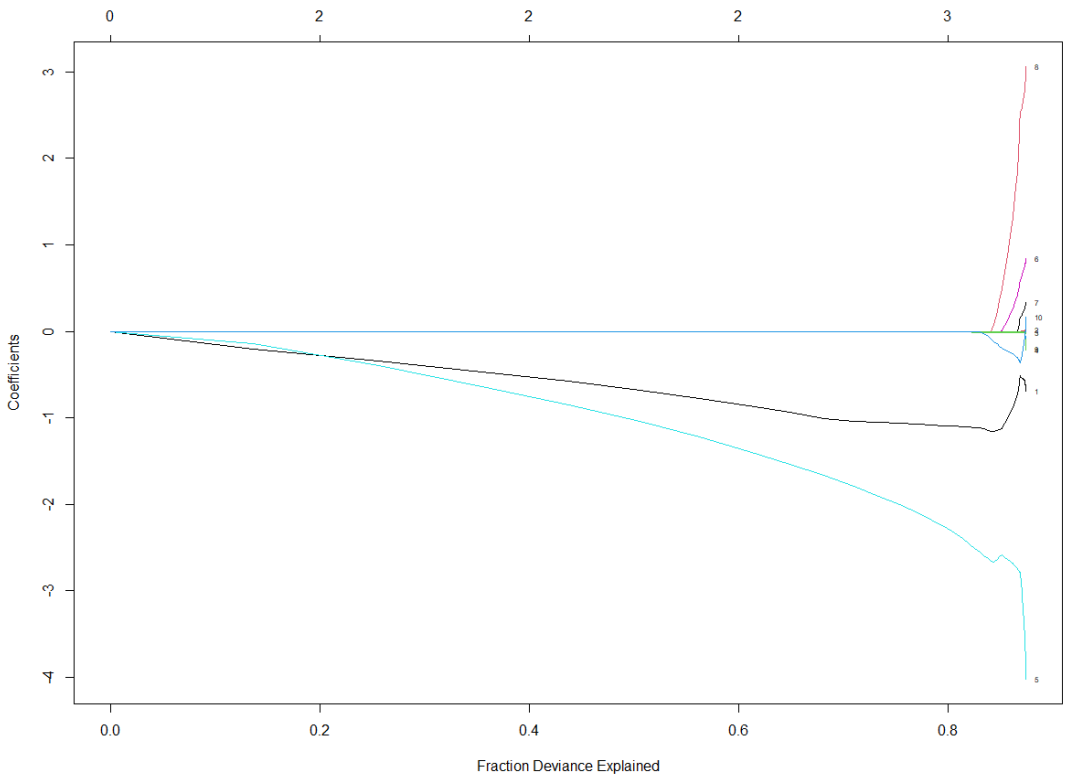
Αφού εφαρμόσουμε την τεχνική Lasso στα δεδομένα μας δημιουργούμε τα παρακάτω διαγράμματα:



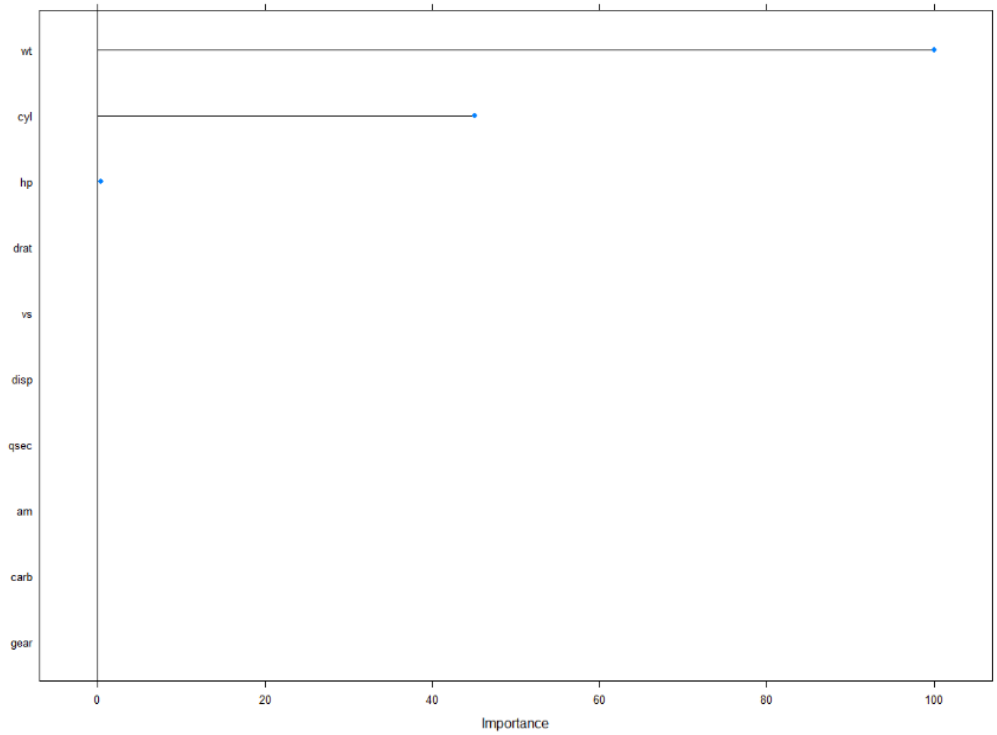
Διάγραμμα 9: Coefficients VS Log Lambda



Διάγραμμα 10: RMSE VS Regularization Parameter



Διάγραμμα 11: Coefficients VS Fraction Deviance Explained



Διάγραμμα 12: Importance για κάθε μεταβλητή

Παρομοίως, το διάγραμμα 9 μας βοηθάει να εντοπίσουμε που περίπου βρίσκεται το βέλτιστο λ και στην συνέχεια με το διάγραμμα 10 εντοπίζουμε το βέλτιστο λ, το οποίο σε αυτή την περίπτωση είναι ίσο με 1. Το βέλτιστο λ προκύπτει όταν ελαχιστοποιείται το RMSE, το οποίο όπως φαίνεται στο διάγραμμα 10 έχει καθοδική πορεία μέχρι το λ να γίνει 1 και ύστερα ανοδική πορεία. Στο διάγραμμα 11 παρατηρούμε πως 60% deviance μπορεί να εξηγηθεί μόνο με 2 μεταβλητές και 80% deviance μόνο με 3 μεταβλητές. Αυτό πρακτικά αποτελεί και την διαφορά της τεχνικής Lasso και Ridge καθώς τώρα οι συντελεστές β των υπόλοιπων 7 μεταβλητών έχουν μηδενιστεί. Στο διάγραμμα 12 αυτό γίνεται εμφανές καθώς πλέον στο μοντέλο παίζουν πολύ σημασία οι μεταβλητές wt και cyl, πάρα πολύ λίγο σημασία η μεταβλητή hp, ενώ όλες οι υπόλοιπες μεταβλητές έχουν αφαιρεθεί. Είναι αξιοσημείωτο να παρατηρήσουμε πως το αποτέλεσμα της τεχνικής Lasso ταιριάζει πάρα πολύ με το αποτέλεσμα του Forward AIC που είδαμε στο ερώτημα 2, όπου το μοντέλο που είχε προκύψει ήταν το mpg = 38.752 -3.167wt -0.942cyl -0.018hp.

Τέλος αξίζει να αναφέρουμε πως υπάρχει και η τεχνική Elastic Net, η οποία αποτελεί συνδυασμός των Ridge και Lasso. Συγκεκριμένα ισχύει $SSE_{EN} = \Sigma(y - \hat{y})^2 + \lambda[(1-\alpha)\Sigma\beta^2 + \alpha\Sigma|\beta|]$ όπου αν α = 0 τότε έχουμε Ridge και αν α = 1 τότε έχουμε Lasso. Χρησιμοποιώντας αυτή την τεχνική θα καταλήξουμε σε ένα βέλτιστο μοντέλο, το οποίο μπορεί να είναι ένας οποιοσδήποτε συνδυασμός μεταξύ Ridge και Lasso. Για παράδειγμα 20% Ridge και 80% Lasso.

Άσκηση 2

Ερώτημα 1

Τα δεδομένα του αρχείου groupsAB.txt αφορούν το βάρος ανδρών (M:Males) και γυναικών (F:Females) σε σχέση με το ύψος τους. Στόχος της ανάλυσης μας είναι να εξετάσουμε αν το βάρος (εξαρτημένη μεταβλητή y) μεταξύ των δύο ομάδων διαφοροποιείται ως προς το ύψος τους (ανεξάρτητη μεταβλητή x_1).

Ορίζουμε την δείκτρια μεταβλητή x_2, η οποία δείχνει το φύλο ως εξής: x_2 = 1 αν η ομάδα είναι M (άνδρες) και x_2 = 0 αν η ομάδα είναι F (γυναίκες). Επιπλέον ορίζουμε την μεταβλητή x_3 ως $x_3 = x_1x_2$, η οποία εκφράζει την αλληλεπίδραση μεταξύ των μεταβλητών x_1 και x_2. Τα δεδομένα μας μετασχηματίζονται ως εξής:

	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

	Y	X1	X2	X3
1:	53.118	1.4224	0	0.0000
2:	56.750	1.5240	0	0.0000
3:	60.382	1.6256	0	0.0000
4:	64.014	1.7272	0	0.0000
5:	67.646	1.8288	0	0.0000
6:	49.486	1.3716	0	0.0000
7:	58.112	1.5748	0	0.0000
8:	59.474	1.6510	0	0.0000
9:	59.474	1.6510	0	0.0000
10:	65.830	1.7780	0	0.0000
11:	95.794	1.6256	1	1.6256
12:	101.242	1.7272	1	1.7272
13:	106.690	1.8288	1	1.8288
14:	112.138	1.9304	1	1.9304
15:	117.586	2.0320	1	2.0320
16:	91.254	1.5748	1	1.5748
17:	103.512	1.7526	1	1.7526
18:	111.230	1.8796	1	1.8796
19:	109.414	1.9050	1	1.9050
20:	122.126	2.0828	1	2.0828

Πίνακας 24: Μετατροπή δεδομένων στην κατάλληλη μορφή

Με αυτό τον τρόπο μπορούμε να ορίσουμε το μοντέλο πολλαπλής γραμμικής παλινδρόμησης ως E(y) = $\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$, όπου αν μία παρατήρηση αφορά γυναίκες (F) μπορεί να γραφτεί ως E(y_F) = $\beta_0 + \beta_1x_1$, ενώ αν αφορά άνδρες (M) μπορεί να γραφτεί ως E(y_M) = $(\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_3$

Το πως θα προσαρμοστούν οι ευθείες εξαρτάται από τις τιμές των συντελεστών β_2 και β_3. Συγκεκριμένα αν $\beta_3 \neq 0$ τότε έχουμε δύο διαφορετικές ευθείες καθώς έχουμε δύο διαφορετικές κλίσεις (Περίπτωση I), αν $\beta_3 = 0$ και $\beta_2 \neq 0$ τότε έχουμε δύο ευθείες παράλληλες καθώς έχουν την ίδια κλίση β_1 (Περίπτωση II) και τέλος αν $\beta_3 = 0$ και $\beta_2 = 0$, τότε αναφερόμαστε στην ίδια ευθεία (Περίπτωση III).

Για την σύγκριση μεταξύ των περιπτώσεων, ξεκινώντας από την περίπτωση I, μπορεί να γίνει η χρήση της ελεγχοσυνάρτησης F για την σύγκριση δύο εμφωλευμένων μοντέλων. Αρχικά θα πραγματοποιηθεί ο έλεγχος E(y) = $\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$ έναντι του E(y) = $\beta_0 + \beta_1x_1 + \beta_2x_2$. Αν ο έλεγχος βγει στατιστικά σημαντικός τότε βρισκόμαστε στην Περίπτωση I. Αν βγει στατιστικά μη σημαντικός τότε συνεχίζουμε με τον έλεγχο E(y) = $\beta_0 + \beta_1x_1 + \beta_2x_2$ έναντι του E(y) = $\beta_0 + \beta_1x_1$. Αν ο έλεγχος βγει

στατιστικά σημαντικός τότε βρισκόμαστε στην Περίπτωση ΙΙ, ενώ αν βγει στατιστικά μη σημαντικός τότε βρισκόμαστε στην Περίπτωση ΙΙΙ.

Θα μπορούσαν να χρησιμοποιηθούν και άλλοι έλεγχοι, όπως ο AIC, σύμφωνα με τον οποίο θα συγκρίναμε τα μοντέλα E(y) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, E(y) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, E(y) = $\beta_0 + \beta_1 x_1$ και θα επιλέγαμε εκείνο με το μικρότερο AIC. Αντίστοιχα θα βρισκόμασταν σε κάποια από τις περιπτώσεις.

Ερώτημα 2

Ξεκινάμε με χρήση της ελεγχοσυνάρτησης F μεταξύ των εμφωλευμένων μοντέλων $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ και $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Τα αποτελέσματα με την χρήση R φαίνονται παρακάτω.

```
> anova(mod1,mod2, test="F")
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3
Model 2: Y ~ X1 + X2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     16 15.604
2     17 57.056 -1   -41.453 42.506 7.065e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Πίνακας 25: Σύγκριση μοντέλων $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ και $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Παρατηρούμε πως ο έλεγχος F είναι στατιστικά σημαντικός και συνεπώς βρισκόμαστε στην Περίπτωση Ι, έχοντας δύο διαφορετικές ευθείες.

Αντίστοιχα όπως περιγράψαμε θα χρησιμοποιήσουμε και τον έλεγχο AIC, όπου τα αποτελέσματα φαίνονται παρακάτω.

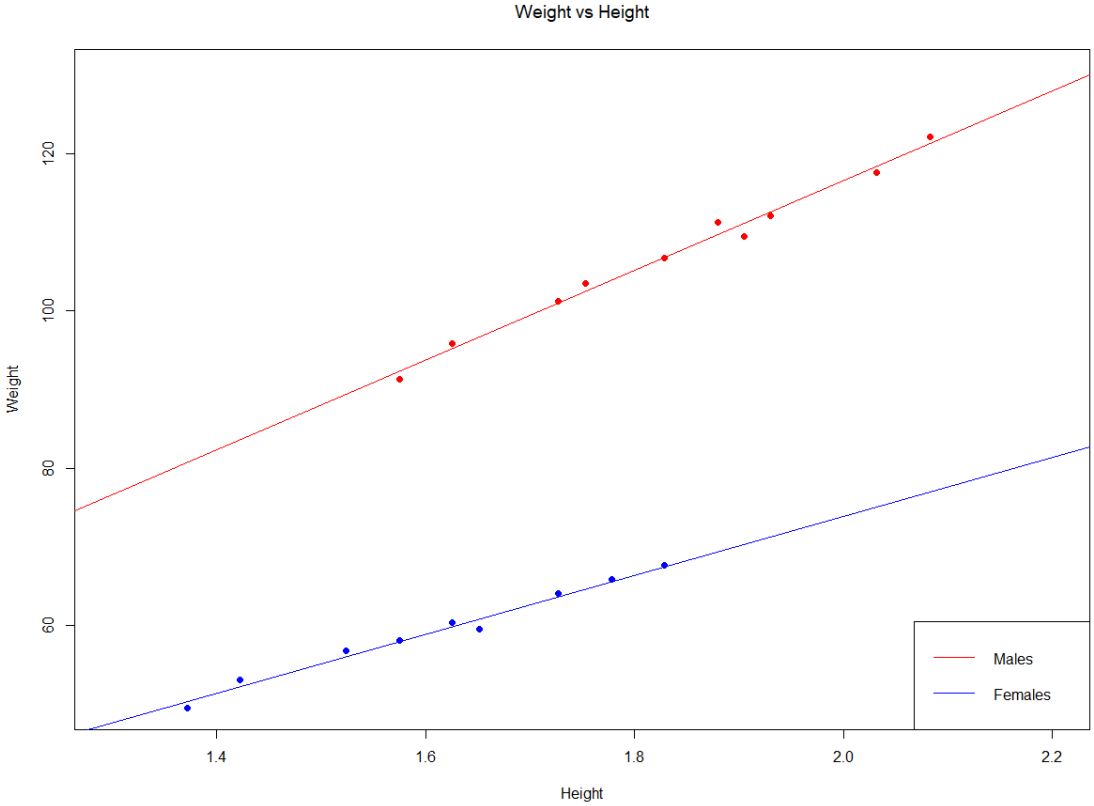
```
> mod1 <- lm(Y ~ X1 + X2 +X3, data = groups)
> summary(mod1)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = groups)

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(mod1)
[1] 61.79283
```

```
> mod2<-lm(Y~X1 +X2, data = groups)
> summary(mod2)

Call:
lm(formula = Y ~ X1 + X2, data = groups)

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(mod2)
[1] 85.72372

> mod3 <- lm(Y~X1, data = groups)
> summary(mod3)

Call:
lm(formula = Y ~ X1, data = groups)

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(mod3)
[1] 171.1629
```

Πίνακας 26: Πολλαπλή γραμμική παλινδρόμηση για τα εμφωλευμένα μοντέλα

Παρατηρούμε πως και σε αυτή την περίπτωση το καλύτερο μοντέλο δίνεται από την σχέση E(y) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ καθώς αυτό έχει το μικρότερο AIC, που ισούται με 61.79. Τα υπόλοιπα δύο μοντέλα έχουν AIC ίσο με 85.72 και 171.16 αντίστοιχα.

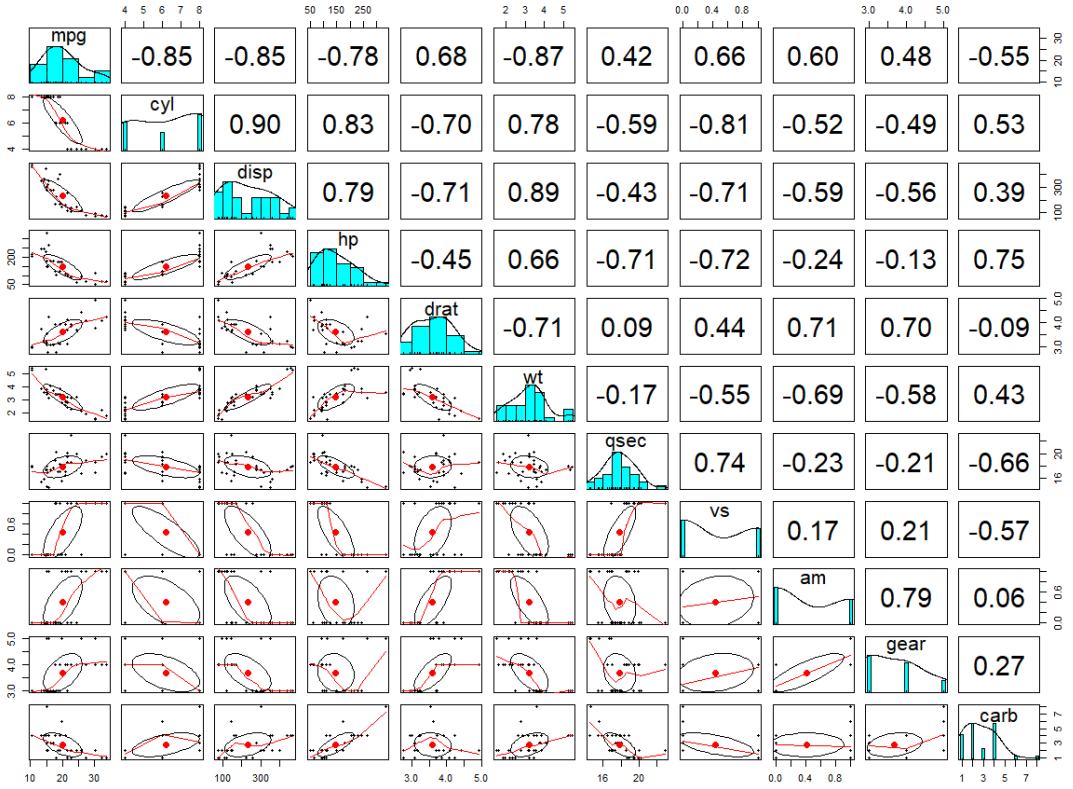
Το τελικό μοντέλο λοιπόν δίνεται από την σχέση E(y) = -1.088 + $37.462x_1$ + $3.632x_2$ + $19.552x_3$, όπου E(y_M) = 2.544 + $57.014x_3$ και E(y_F) = -1.088 + $37.462x_1$ με πτώση της κλίσης κατά 19.552. Η γραφική παράσταση φαίνεται παρακάτω.



Διάγραμμα 13: Βάρος σε σχέση με το ύψος ανάλογα τις ομάδες Males και Females προσαρμόζοντας τις ευθείες $E(y_M) = 2.544 + 57.014x_3$ και $E(y_F) = -1.088 + 37.462x_1$

Η ερμηνεία για τους συντελεστές των δύο αυτών ευθειών δίνεται ως εξής: Αν το ύψος αυξηθεί κατά 0.1 και γνωρίζουμε πως ανήκει στην ομάδα Males τότε το βάρος θα αυξηθεί κατά περίπου 57.014*0.1 = 5.7014, ενώ αν ανήκει στην ομάδα Females τότε το βάρος θα αυξηθεί κατά περίπου 37.462*0.1 = 3.7462.

Παράρτημα



Πίνακας 27: Διάγραμμα συσχετίσεων των 11 μεταβλητών x_j του αρχικού μοντέλου με χρήση της βιβλιοθήκης psych