

ΕΡΓΑΣΙΑ ΣΤΟ ΜΑΘΗΜΑ ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ

**ΜΑΘΗΜΑ ΣΤΟ ΔΠΜΣ : ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ
ΜΑΘΗΣΗ**



ΣΤΟΙΧΕΙΑ ΦΟΙΤΗΤΗ:

ΟΝΟΜΑΤΕΠΩΝΥΜΟ: ΖΕΡΚΕΛΙΔΗΣ ΔΗΜΗΤΡΙΟΣ
Α.Μ. 003400049

ΑΣΚΗΣΗ 1 ~ ΑΠΟΔΕΙΞΕΙΣ

ΣΕΛΙΔΑ 2

A)

$$\textcircled{1} \quad \text{Παραγγελείτε ότι } SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}' X' Y - n \bar{y}^2$$

$$\text{και } SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n y_i^2 - \hat{\beta}' X' Y$$

$$\text{Έχουμε } SSE = \sum_{i=1}^n e_i^2 = e'e. \quad \textcircled{2}$$

$$\text{Οπως } e = y - X\hat{\beta} \quad \textcircled{2}$$

$$\text{η } \textcircled{1} \xrightarrow{\text{ε}} SSE = (y - X\hat{\beta})' (y - X\hat{\beta}) =$$

$$= y'y - \hat{\beta}' X' Y - Y' X \hat{\beta} + \hat{\beta}' X' X \hat{\beta}$$

$$= y'y - \hat{\beta}' X' Y + \hat{\beta}' X' X \hat{\beta} \quad \textcircled{3}$$

$$\text{Ισχει επίσης ότι } X' X \hat{\beta} = X'y \quad \textcircled{4}$$

$$\text{-Άρα } \textcircled{3} \xrightarrow{\text{ε}} SSE = y'y - \hat{\beta}' X' Y -$$

$$\text{η } \textcircled{4} \Rightarrow SSE = \sum_{i=1}^n y_i^2 - \hat{\beta}' X' Y$$

$$\text{Tipa der Abweichung von } SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{B}' \hat{x}' y - n \bar{y}^2$$

$$\text{Ixiel von } SST = SSR + SSE$$

$$\text{Frwijkoupe, } SSE = y'y - \hat{B}' \hat{x}' y$$

$$SST = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} = y'y - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

$$SSR = SST - SSE = y'y - \frac{(\sum_{i=1}^n y_i)^2}{n} - (y'y - \hat{B}' \hat{x}' y)$$

$$\Leftrightarrow SSR = \hat{B}' \hat{x}' y - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

$$\therefore \boxed{SSR = \hat{B}' \hat{x}' y - n \bar{y}^2}$$

(2) Μα αναδιζετε ότι οποιο τ.Γ.Μ. $\hat{Y} = XB + \epsilon$, 15×26

$$\text{cov}(t_i, \hat{f}_i) = \sigma^2_{\text{hi}}$$

όπου $h_{ii} = X_i'(X'X)^{-1}X_i$ Το διανυσματικό πρώτον γραμμήν του
κινητήρα $H = X(X'X)^{-1}X'$

Εντού, μάλιστα $\text{cov}(\hat{Y}, \hat{f}) = E[(\hat{Y} - E(\hat{Y}))(\hat{f} - E(\hat{f}))^T] \quad (1)$

-Οφεις παραδοχή ότι :

$$e = Y - E(Y) \quad (2)$$

$$\begin{aligned} \hat{Y} - E(\hat{Y}) &= H(Y - E(Y)) = H(Y - E(H) - H(Y - E(Y))) \\ &= He \quad (3) \end{aligned}$$

$$h \quad (1) \xrightarrow[\text{(3)}]{(2)} \text{cov}(Y, \hat{f}) = E(ee^T H) = H E(ee^T) = H \sigma^2$$

~~Επειδή e και \hat{f} είναι ανεξάρτητα~~

• Το i,i πρώτον γραμμήν του διανυσματικού πρώτου γραμμήν του $\text{cov}(Y, \hat{f})$ δίνει το $\text{cov}(t_i, \hat{f}_i)$

• Ενώ το i,i πρώτον γραμμήν της διανυσματικού πρώτου γραμμήν του $HAT(H)$ ματρικού δίνει

το h_{ii}

$\text{cov}(t_i, \hat{f}_i) = \sigma^2 h_{ii}$

ΑΣΚΗΣΗ 2 ~ ΑΝΑΛΥΣΗ ΜΟΝΤΕΛΟΥ ΣΤΟ DATASET MTCARS

ΕΡΩΤΗΜΑ 1

Να προσαρμοστεί ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης στα δεδομένα του αρχείου σχετίζοντας τα μίλια/gallon mpg (Y) με τις δέκα παραπάνω επεξηγηματικές μεταβλητές. Να εξετάσετε αν υπάρχουν συσχετίσεις μεταξύ των μεταβλητών X_j, αν υπάρχει πολυσυγγραμμικότητα, αν τηρούνται οι προϋποθέσεις του μοντέλου καθώς και άλλοι διαγνωστικοί έλεγχοι.

Αρχικά, μας ζητείται να προσαρμόσουμε το μοντέλο και με τις 10 επεξηγηματικές μεταβλητές. Το οποίο και κάνουμε με τις παρακάτω εντολές.

```
file1 <- read.table("/home/zerkes/Desktop/DSML - MASTER/1st semester/STATISTICAL MODELLING/solution/mcars1.txt", header=TRUE)
attach(file1)
mod1 <- lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb)
```

Στη συνέχεια τρέχουμε την εντολή `summary(mod1)` για να δούμε τι πληροφορίες θα μας δώσει για το μοντέλο που δημιουργήσαμε.

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

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
```

1η παρατήρηση : Μεγάλο p-value .

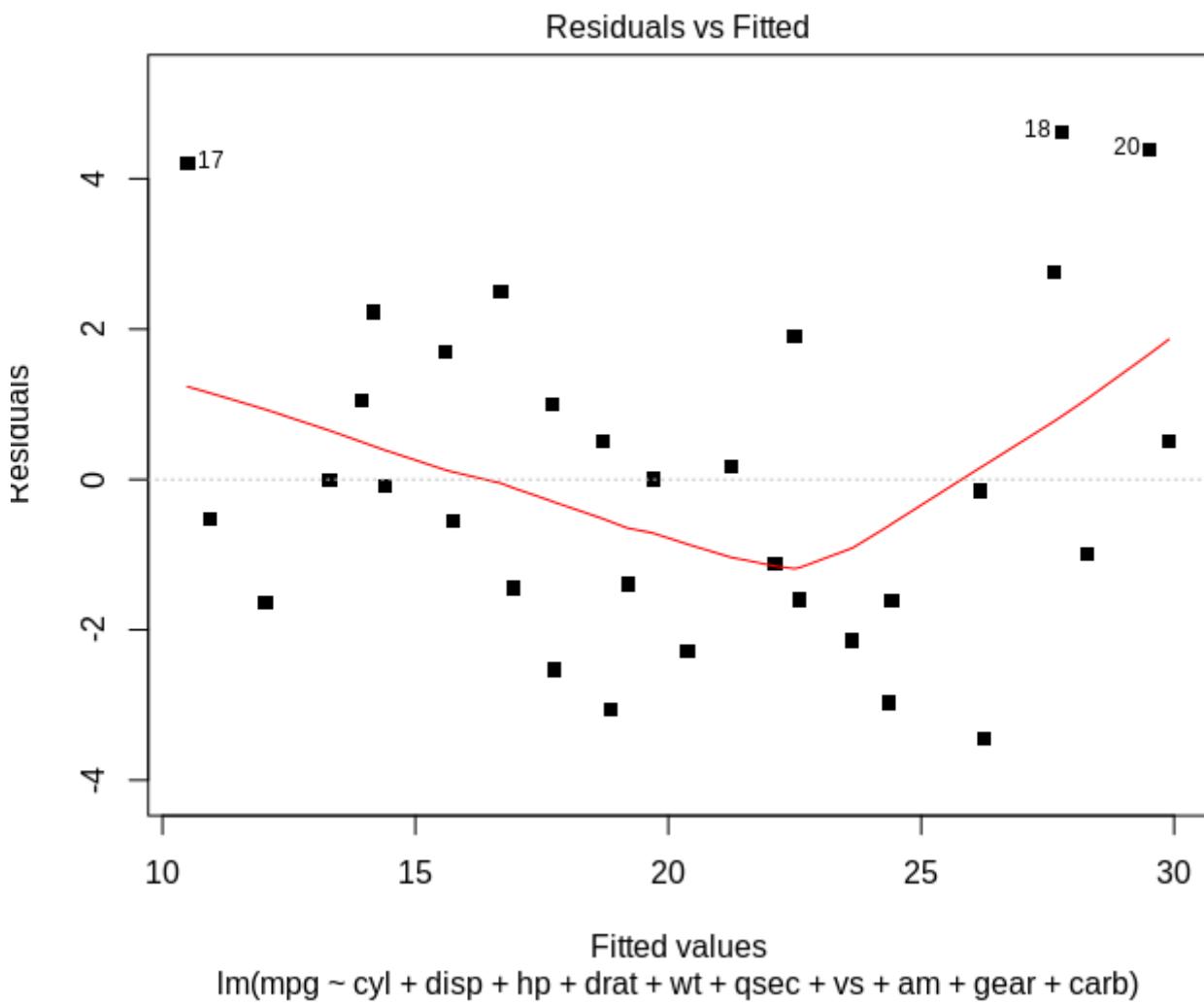
Από την παραπάνω ανάλυση του μοντέλου καταλαβαίνουμε πως υπάρχει μεγάλο πρόβλημα στο μοντέλο μας καθώς όλες οι μεταβλητές ειναι στατιστικά ασήμαντες με p-value αρκετά μεγάλο.

Αυτό σημαίνει πως προφανώς έχουμε μεταβλητές που σχετίζονται μεταξύ τους και έχουν υψηλό correlation και δε δίνουν παραπάνω πληροφορία στο μοντέλο μας.

Τώρα θα προσπαθήσουμε να βρούμε αν υπάρχουν και άλλες προϋποθέσεις του μοντέλου που δεν τηρούνται με τη χρήση διαγνωστικών ελέγχων όπως residual plots ή τη χρήση του αριθμού VIF.

Παρακάτω, δείχνουμε το διάγραμμα residuals vs fitted values

plot(mod1,which=1,pch=15)



Από το διάγραμμα φαίνεται πως υπάρχει ετεροσκεδαστικότητα κάτι που θέλουμε να αποφύγουμε.

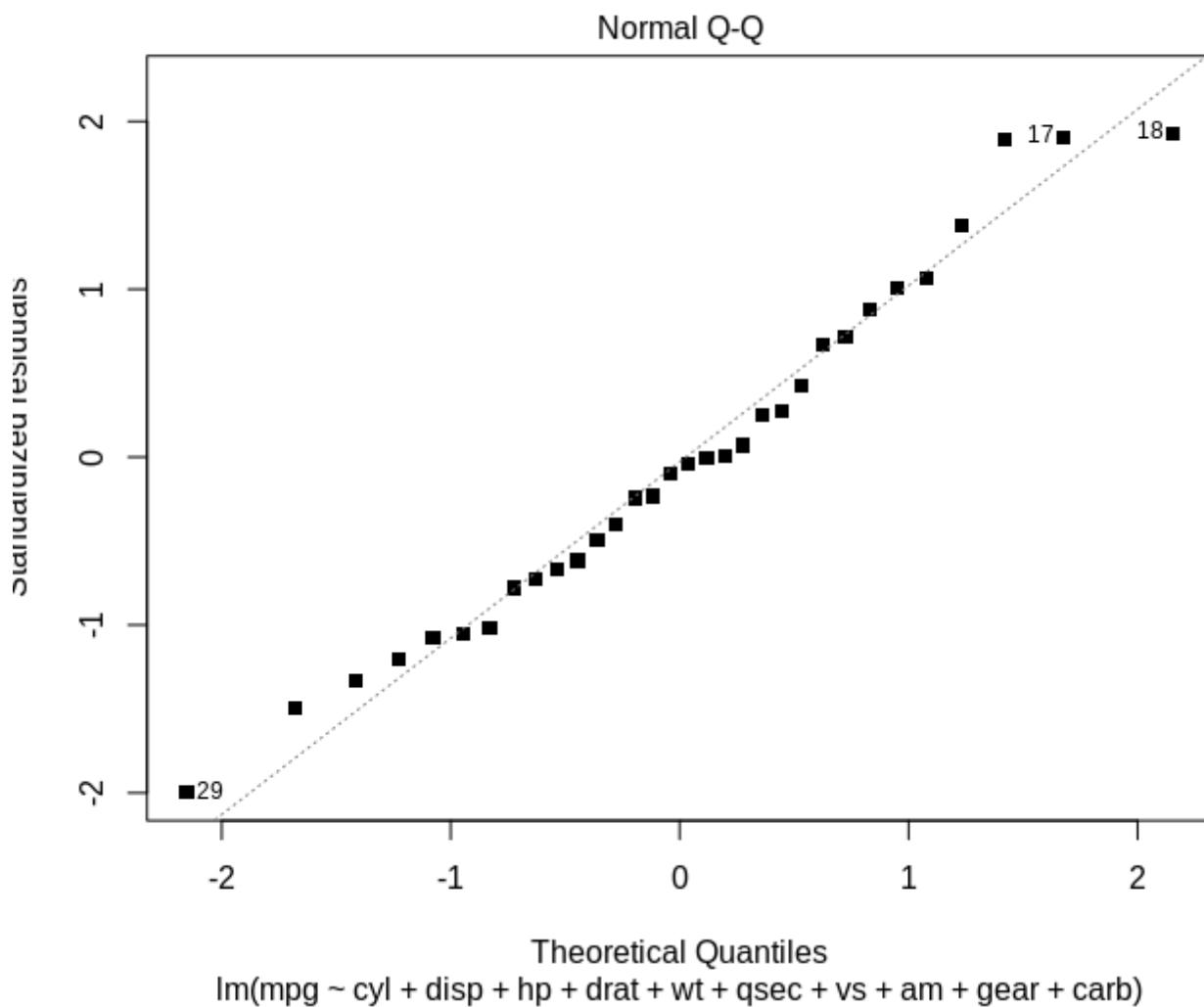
Στο διάγραμμα τα περισσότερα σημεία είναι κάτω από το 0, ενώ θα έπρεπε να είναι όλα διασκορπισμένα τυχαία.

Παρακάτω παρουσιάζουμε το *normal qq plot* που δίνεται με την εντολή

`plot(mod1,which=2,pch=15)`

Παρατηρούμε πως το *qqplot* μας δείχνει υπάρχουν 3-4 σημεία που δεν πέφτουν κοντά στην ευθεία γραμμή και παραβιάζεται σε ένα βαθμό η κανονικότητα. Ωστόσο, όχι σε αρκετά μεγάλο.

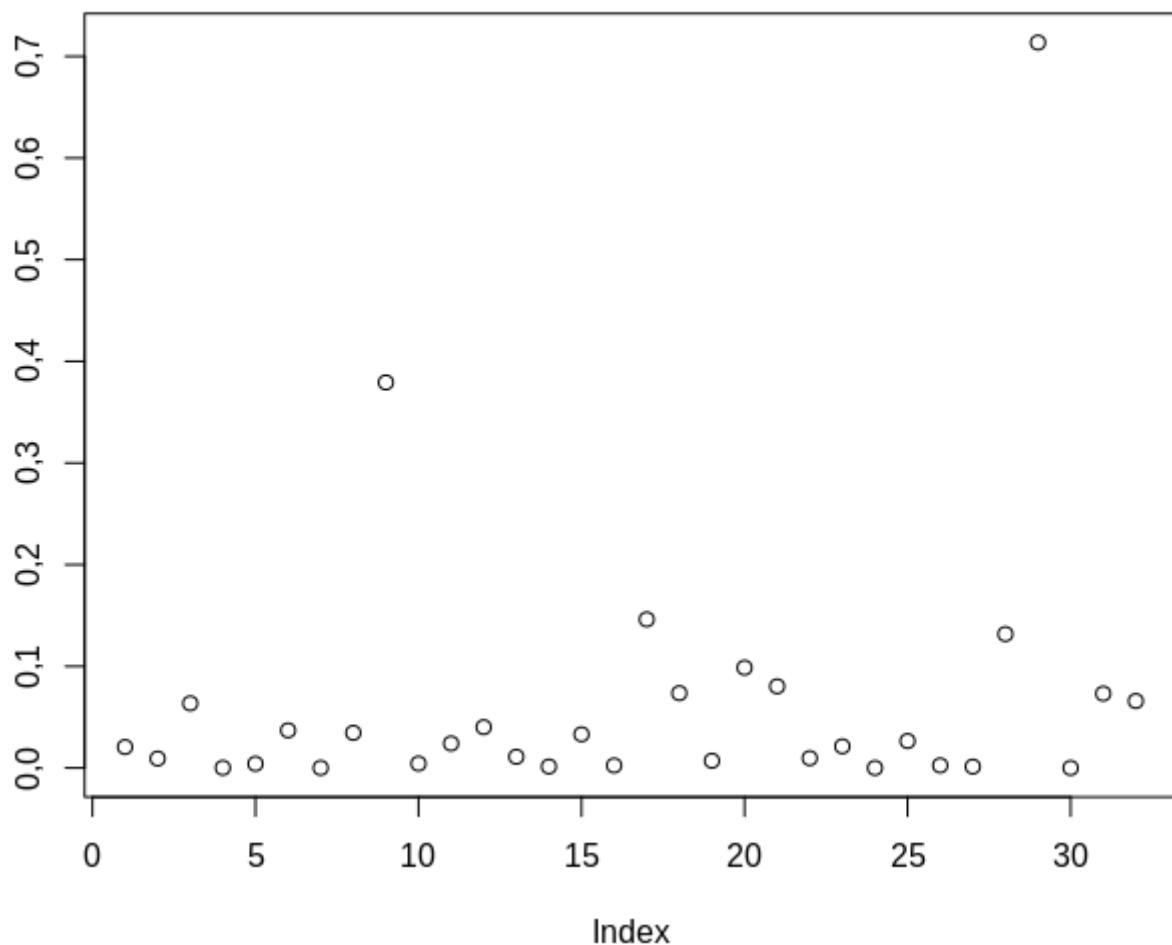
Στην παρακάτω γραφική παράσταση έχει γίνει χρήση των *standarised(studentized) residuals*, τα οποία έχουν κανονικοποιηθεί, ώστε να έχουν την ίδια διασπορά αφαιρώντας την ετεροσκεδαστικότητα.



Παρακάτω δημιουργούμε το διάγραμμα *Cook Distance vs index*.

```
plot(cooks.distance(mod1))
```

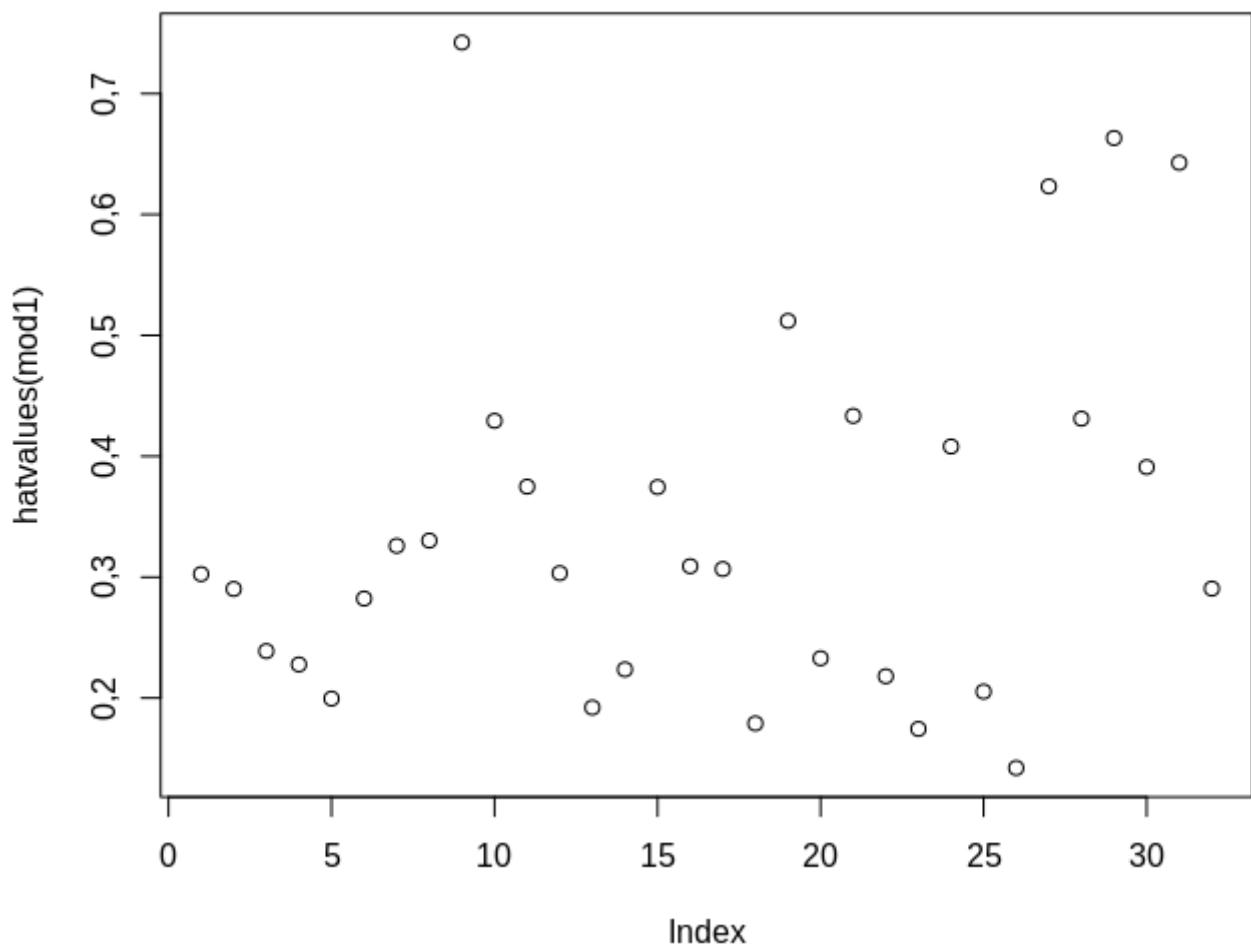
Από το παρακάτω διάγραμμα φαίνεται πως έχουμε κάποια *outliers* τα οποία μπορεί να μας επηρεάζουν στην προσαρμογή του μοντέλου. Ωστόσο κανένα δεν ξεπερνάει το 1, οπότε δε μπορούμε να πούμε ότι έχουμε κάποιο σημείο επιρροής από την απόσταση *Cook*.



Επιπλέον , Θα δοκιμάσουμε και το leverage plot για τις τιμές *hat values* , διότι μας εμφανίζει σημεία επιρροής , τα οποία η cook distance δεν τα υπολογίζει. Η μέθοδος leverage(μόχλευση) χρησιμοποιεί μόνο τις ανεξάρτητες μεταβλητές και όχι και τις εξαρτημένες όπως η μέθοδος της απόστασης Cook.

Με την παρακάτω εντολή σχεδιάζουμε το γράφημα *hat values* vs *index*.

```
plot(hatvalues(mod1))
```

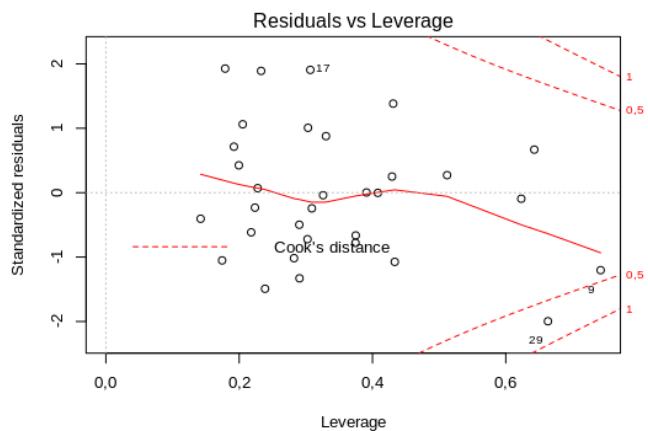
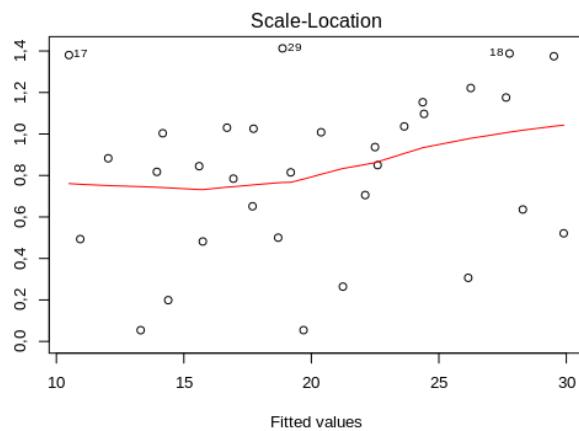
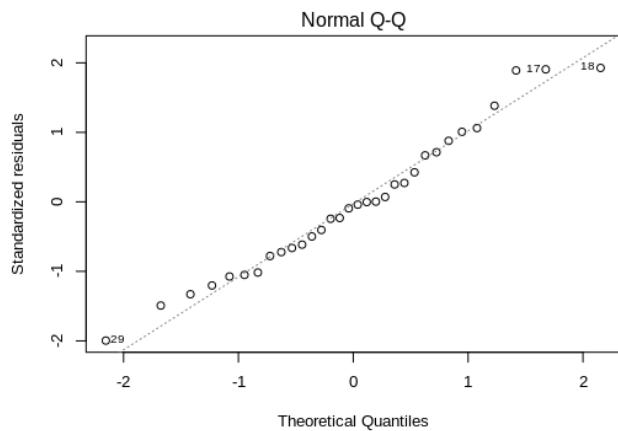
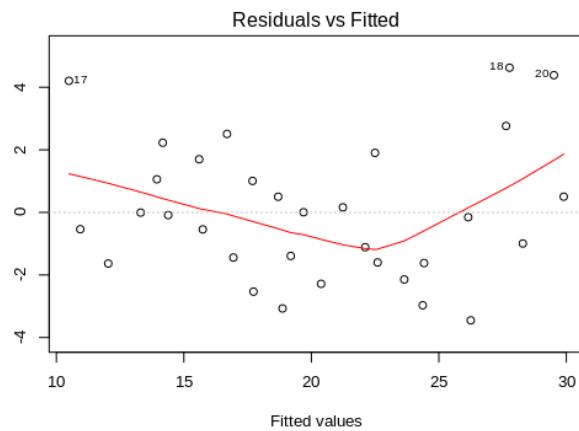


Σύμφωνα με το παραπάνω γράφημα για να θεωρήσουμε μια παρατήρηση ως σημείο επιρροής πρέπει να ισχύει, $h_{ii} > 2 * 10/32 = 0.625$. Παρατηρούμε πως έχουμε 4 τέτοια στοιχεία που επηρεάζουν την προσαρμογή του μοντέλου μας.

Παρακάτω παρουσιάζουμε το σύνολο των γραφικών παραστάσεων που μας παρέχει η R . Τις 2 πρώτες γ.π. στην 1η σειρά τις είδαμε πρηγουμένως. Οι επόμενες 2 της επόμενης σειράς είναι τα :

-Τυποποιημένα residuals vs fitted values που αφαιρούν την ετεροσκεδαστικότητα από τα residuals.

-Residuals vs Leverage.



Στη συνέχεια θα χρησιμοποιήσουμε την εντολή **dfbetas(mod1)** για να δούμε οι παρατηρήσεις τι επιρροή έχουν σε κάθε ένα coefficient του μοντέλου μας. Αν κάποιο από αυτά έχει μεγάλο DFBETAS VALUE τότε η συγκεκριμένη παρατήρηση ασκεί επιρροή στο μοντέλο μας.

Γενικά, κάποια παρατήρηση I ασκεί επιρροή όταν $|DFBETAS_{ij}| > 2/\sqrt{n}$

Στη δική μας περίπτωση $n = 32$ άρα $|DFBETAS_{ij}| > 0.35$

Παρατηρούμε πως υπάρχουν αρκετά σημεία ζεπερνάνε την τιμή 0.35. Ένα από αυτά είναι η παρατήρηση 29 η οποία παίρνει την τιμή 4.5 στη μεταβλητή “vs” γεγονός που δείχνει ότι ασκεί μεγάλη επιρροή πάνω σε αυτήν.

Γενικά η στήλη “vs” φαίνεται να επηρεάζει πολύ το μοντέλο μας καθώς η πλειοψηφία των παρατηρήσεων φαίνεται να αποκτά μεγάλη τιμή σε αυτή τη στήλη.

	> dfbetas(mod1)												
	(Intercept)	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb		
1	-0.0802281838	0.0062625932	-0.0966338918	0.2600486555	0.0354849408	0.1090686613	-0.0157343673	9.862100e-02	-0.1397359321	0.1668142181	-0.2585584508		
2	0.0052328159	-0.0322931507	-0.0097810980	0.1413196427	0.0030085566	0.0136763524	-0.0466591388	9.068081e-02	-0.1293494728	0.0867605925	-0.1260086701		
3	-0.2478182269	0.1380197891	0.2994530767	-0.3054268783	0.1870032533	-0.3859113573	0.2733576071	-1.906396e-01	-0.3712309391	0.1027036364	0.4066847273		
4	0.0055880788	0.0014548382	0.0162420317	-0.0107229679	-0.0146515287	-0.0130389264	-0.0004311089	2.099255e-02	0.0026894351	-0.0016927101	0.0080458630		
5	-0.0148865231	0.0218880243	0.1017261117	-0.0411448924	0.0001760854	-0.1198671265	0.0426644008	-2.877391e-02	-0.0409196316	0.0131446799	0.0210984708		
6	-0.0143219808	-0.1319932444	0.0501885266	0.0172930412	0.3927172853	-0.0310496244	-0.0548181304	-2.424401e-01	-0.1645406990	-0.0760120651	0.0573968931		
7	-0.0116472383	0.0127220434	-0.00880168469	-0.0069470392	0.0017836523	0.0136687266	0.0023130583	1.641978e-03	0.0076646570	0.0141758222	-0.0098181791		
8	0.1809706701	-0.2413738635	0.0610506722	-0.1747670868	-0.0869272866	0.0822111534	-0.1836150430	-1.563950e-02	-0.3713108072	0.2198323189	0.0376607501		
9	1.2621677975	0.0241003847	-0.2615139174	-0.5280186930	-0.2705710277	0.5690921612	-1.5950527663	9.378371e-01	0.4025619443	-0.5608331367	-0.1706675580		
10	-0.0018443256	0.0714213801	-0.0455635431	-0.0668762141	0.0612212526	0.0595024864	-0.0800701708	1.152810e-01	-0.0833337299	0.0474999343	0.0105186294		
11	0.1104050493	-0.2204173928	0.0774670476	0.1562832948	-0.1616551197	-0.0827656630	0.0678245575	-2.449552e-01	0.1757084416	-0.1326324102	-0.0637457212		
12	-0.0095080493	0.2234813540	-0.5673552154	0.2426556337	0.0427879697	0.4346811830	-0.1292769487	-2.01149e-01	-0.0083615610	0.0365739106	-0.3873209989		
13	-0.0565173271	0.1356455995	-0.2097605554	0.1165897350	-0.0027277286	0.0699370271	0.0559280879	-1.100568e-01	-0.0115710423	-0.0073029170	-0.1076173848		
14	0.0396300560	-0.0533580267	0.0678154540	-0.0435573241	-0.0016872618	-0.0177541345	-0.0420954706	4.571447e-02	-0.0048521942	-0.0015719420	0.0340863144		
15	0.0192842107	0.1567043100	-0.3532330924	0.2338593996	0.1005074681	0.0752603707	-0.1031804149	-2.113319e-02	-0.1710874994	0.0509740475	-0.2439605438		
16	-0.0008974285	0.0401614475	-0.0356717210	0.0246545620	0.0068456750	-0.0454868894	0.0040360128	-4.740677e-03	-0.0464016613	0.0107269720	-0.0128565290		
17	0.0429492736	-0.3608891928	-0.1948915128	0.2324904149	0.3214121655	0.7752539274	-0.2843233814	1.930443e-03	0.2330302771	-0.1663461545	-0.3207287566		
18	-0.2107550564	0.2509539824	-0.3050736908	0.1652926417	0.0772650900	0.3088224890	0.1156942135	1.154706e-01	0.5721444850	-0.0681930279	-0.3492065828		
19	-0.0398080182	0.0343816888	0.1155667827	-0.1065518357	0.1544897497	-0.1108498547	0.0325095023	5.833917e-02	0.0308896834	-0.1091580202	0.1213327686		
20	-0.5592002881	0.3509602416	0.1023978375	0.1122960701	0.1982794347	-0.3099084456	0.6683675458	-4.717503e-02	0.5630018916	-0.1600944519	0.0288609544		
21	-0.4234594478	0.5942808502	0.0584352076	-0.4427638538	0.0143209064	0.1709862018	0.0802762574	2.299156e-01	0.3701412076	0.4735304181	-0.0703297039		
22	-0.0858086282	-0.06094746021	-0.0107072656	0.1076241384	0.1489518699	0.0355037155	0.0620033316	4.932487e-03	0.0371467677	-0.0644015845	0.0264518337		
23	0.0832923658	-0.2391054064	0.0301861488	0.1024188732	-0.0330241207	0.0691085979	-0.0473316838	8.385001e-02	0.0784474972	-0.0950109401	0.0902594054		
24	-0.0008290439	0.0006531818	0.0004374677	-0.0009149880	-0.0011951733	-0.0002582488	0.0010231151	-3.195125e-05	0.0008576534	0.0011020800	0.0002596754		
25	0.0010437621	-0.0019604462	0.3260796338	-0.1949922314	-0.0298638528	-0.2235259345	0.0341735385	-1.441958e-02	-0.0530393093	0.0758721542	0.0443070575		
26	-0.0052000786	-0.0191132919	0.0130579934	-0.0036923500	0.0017418874	-0.0101633974	0.0042732041	-3.940118e-02	-0.0810327574	0.0329877789	0.0264918850		
27	-0.0175584836	0.0426272994	0.0125342546	-0.0085660023	-0.0174960048	-0.0249762189	0.0241920466	7.447629e-02	0.0483032571	-0.0462534029	0.0436178419		
28	0.5102015963	-0.3117084737	0.2604600955	-0.0493228786	-0.6604033145	-0.2557366267	-0.3998012490	3.447797e-01	-0.2498187278	0.4233467519	-0.0326844062		
29	1.5031369368	-1.3250467886	-0.0083468216	-0.5335536869	-1.2628586930	-0.2912018391	-0.1955388198	-4.565789e-01	-0.0935737950	-1.6594052827	1.3946783541		
30	0.0007154665	-0.0004883266	-0.0001073003	-0.0004285907	-0.0011904630	-0.0001552195	-0.0002400735	-4.720437e-04	-0.0001419433	0.0002766015	0.0008559461		
31	-0.1181378553	-0.0259328258	0.0435358245	0.3084974640	-0.2299181204	0.2103408144	0.2458690372	7.996855e-02	0.1832247195	-0.0287468353	0.2623689542		
32	-0.2099417251	0.1874010237	0.3419571679	-0.3001747963	-0.0749213726	-0.5194358361	0.3142646375	-2.158247e-01	-0.3717999505	0.2609872777	0.3061125751		

Παρακάτω θα τρέξουμε την εντολή **dffits(mod1)** ,ώστε να δούμε για την κάθε παρατήρηση αν δεν την συμπεριλαμβάναμε στο μοντέλο πόσο θα επηρεάζε τη καινούρια προσαρμογή μας χωρίς αυτήν ή με αυτήν.

Αυτή είναι μια άλλη μέθοδος για να βρούμε τα σημεία επιρροής.

```
> dffits(mod1)
   1      2      3      4      5      6      7      8      9      10     11     12
-0,470336460 -0,312613697 -0,862811386  0,037003357  0,207716116 -0,638256415 -0,026901447  0,612906168 -2,065687952  0,212125552 -0,507249992  0,664699488
   13     14     15     16     17     18     19     20     21     22     23     24
 0,343914128 -0,121788112 -0,597355814 -0,159154958  1,360566891  0,967564035  0,272152176  1,115696987 -0,943257304 -0,320186447 -0,484676559 -0,002393136
   25     26     27     28     29     30     31     32
 0,541365500 -0,161425967 -0,118037702  1,231964143 -3,037374792  0,002333660  0,884515333 -0,867680270
```

Για να επηρεάζει μια παρατήρηση το μοντέλο πρέπει η απόλυτη τιμή του dffits να είναι μεγαλύτερη από το $2 * \text{sqrt}(p/n)$, όπου p είναι οι παράμετροι.

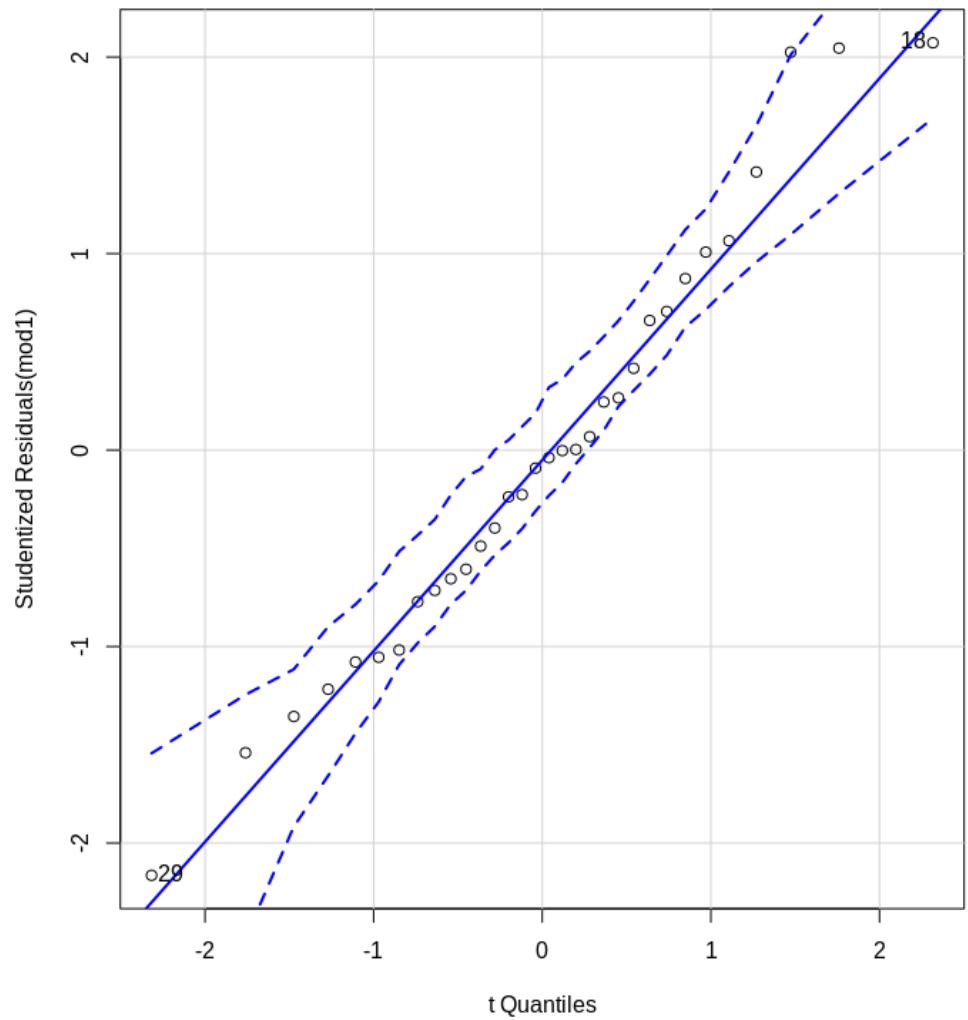
Στη δική μας περίπτωση για $p=10$, $n=32$ τα σημεία που ασκούν επιρροή είναι αυτά για τα οποία ισχύει :

$\text{dffits} > 1.11$

Κάποιες τέτοιες παρατηρήσεις είναι οι 29,28,17 και 20.

Studentised qqplot

Από την παρακάτω γραφική παράσταση όπου σαν residuals έχουμε πάρει τα studentised residuals ,ώστε να αφαιρεθεί η ετεροσκεδαστικότητα, βλέπουμε πως έχουμε μια ευθεία γραμμή, την οποία τα σφάλματα δεν την παραβιάζουν ιδιαιτέρως. Όποτε δεν υπάρχει πρόβλημα ως προς την κανονικότητα των σφαλμάτων.



Πολυσυγγραμμικότητα(Multicollinearity)

Για να δούμε αν υπάρχει πολυσυγγραμμικότητα θα κάνουμε χρήση της συνάρτησης VIF. Η συνάρτηση αυτή θα επιστρέψει για κάθε επεξηγηματική μεταβλητή μια τιμή, η οποία είναι μεγαλύτερη από 5 είναι ένδειξη ότι η επεξηγηματική μεταβλητή X_j εξηγείται από τις άλλες επεξηγηματικές μεταβλητές.

Στην R , με την εντολή **vif(mod1)** παίρνουμε τα εξής δεδομένα.

```
> vif(mod1)
   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
> |
```

Βλέπουμε ότι στο μοντέλο μας έχουμε πολλές τιμές με πάρα πολύ υψηλό VIF γεγονός που σημαίνει ότι σίγουρα υπάρχει πολυσυγγραμμικότητα.

Η μεταβλητή “**disp**” σύμφωνα με τη συνάρτηση VIF ,θα ήταν η πρώτη μεταβλητή υποψήφια προς αποχώρηση , από το μοντέλο μας.

Πράγματι αν αφαιρέσω αυτήν την μεταβλητή το μοντέλο που θα πάρω θα έχει πολύ καλύτερες τιμές VIF.

```
> mod2 <- lm(mpg~cyl+hp+drat+wt+qsec+vs+am+gear+carb)
> vif(mod2)
   cyl      hp      drat      wt      qsec      vs      am      gear      carb
14,284737  7,123361  3,329298  6,189050  6,914423  4,916053  4,645108  5,324402  4,310597
```

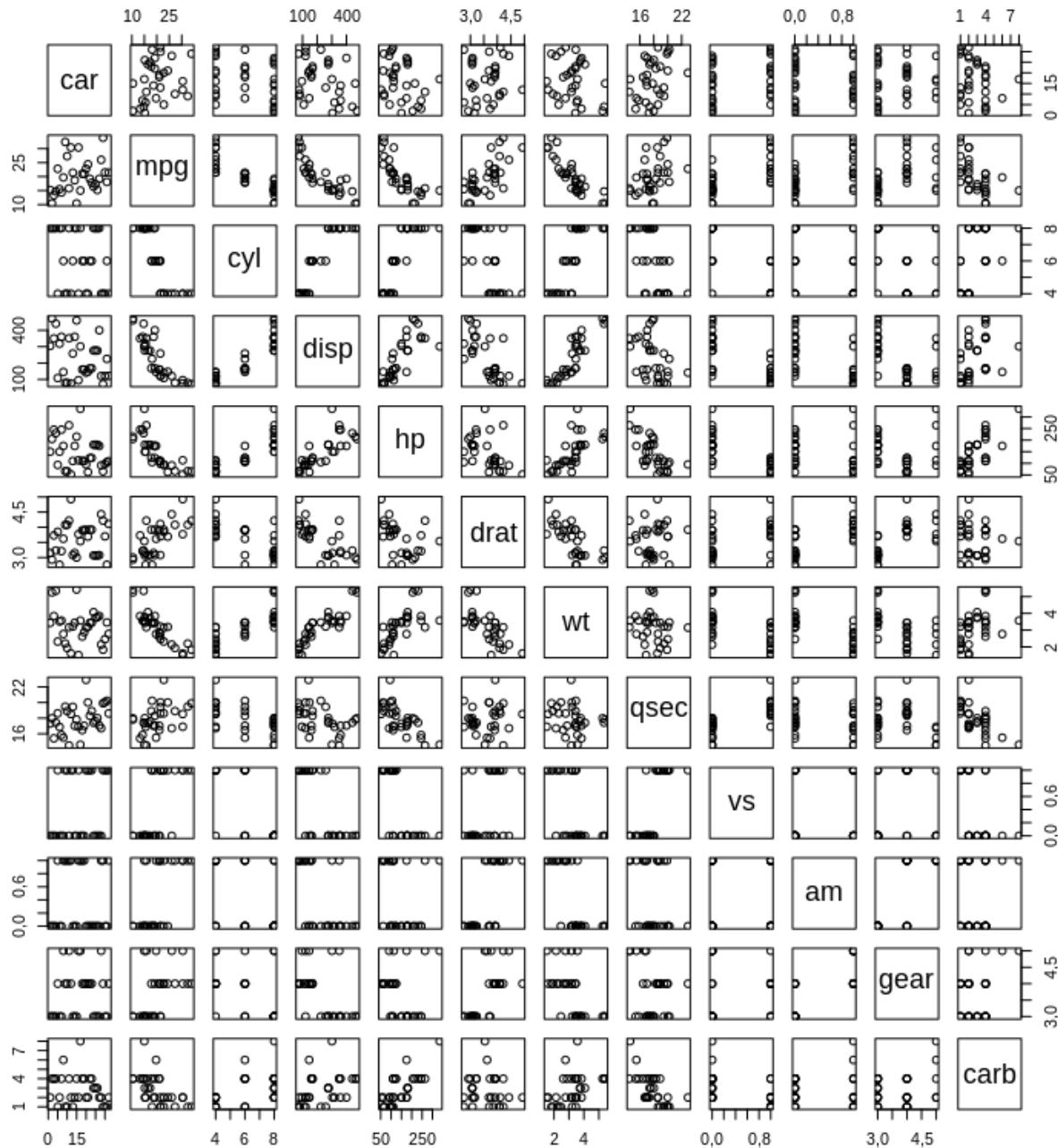
Τώρα η επόμενη μεταβλητή με μεγάλη VIF τιμή είναι η cyl. Ας δοκιμάσουμε να την αφαιρέσουμε , ώστε να δούμε τι τιμές θα πάρουμε με το μοντέλο χωρίς αυτή την μεταβλητή.

```
> mod3 <- lm(mpg~hp+drat+wt+qsec+vs+am+gear+carb)
> vif(mod3)
   hp      drat      wt      qsec      vs      am      gear      carb
6,015788 3,111501 6,051127 5,918682 4,270956 4,285815 4,690187 4,290468
```

Όπως φαίνεται μειώθηκε αρκετά η πολυσυγγραμμικότητα σύμφωνα με την τιμή VIF.

Θα μπορούσαμε να εξάγουμε και μια γραφική παράσταση για το κάθε συνδυασμό επεξηγηματικής μεταβλητής ώστε να δούμε και γραφικά αν έχουν μεταξύ τους γραμμικές σχέσεις.

```
>plot(file1)
```



ΕΡΩΤΗΜΑ 2

Να εξεταστεί αν το μοντέλο με τις δέκα επεξηγηματικές μεταβλητές είναι το βέλτιστο και αν όχι, να επιλέξετε ανάμεσα σε όλα τα δυνατά μοντέλα το βέλτιστο (να αξιοποιηθούν τεχνικές με βήματα με ελέγχους F και t , τα κριτήρια R^2 , Cp , $Press$ και AIC , καθώς και οι γραφικές παραστάσεις των πρόσθετων και μερικών υπολοίπων).

Όπως, είδαμε και στο 1ο υποερώτημα το μοντέλο αυτό δεν είναι βέλτιστο καθώς υπάρχει το πρόβλημα της πολυσυγγραμμικότητας, ενώ για τα σφάλματα υπάρχει το πρόβλημα της ετεροσκεδαστικότητας. Επιπλέον, υπάρχουν κάποια σημεία επιρροής, τα οποία επηρεάζουν το μοντέλο μας.

Εξαιτίας των παραπάνω θα προχωρήσουμε στην εύρεση ενός βέλτιστου μοντέλου σύμφωνα με κάποια κριτήρια όπως $Rsq-adj$ Cp -mallows, $press$, aic και γραφικές παραστάσεις πρόσθετων και μερικών υπολοίπων.

Step Selection ~ Forward with F test

Αρχικά ξεκινάμε με το μοντέλο χωρίς να περιέχει καμία μεταβλητή και στη συνέχεια προσθέτουμε σε αυτό μεταβλητές ανάλογα με το πόσο βελτιώνεται το SSE. Το μοντέλο θα σταματήσει όταν δει πως η εισαγωγή μιας νέας μεταβλητής δε προσφέρει κάτι ιδιαίτερο στο μοντέλο μας.

```
RMS introduced by coercion
> mod_fw = step(lm(mpg~1),mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb, direction = "forward",test="F")
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
..

Όπως φαίνεται καταλήγαμε στο μοντέλο mod_fw = wt + cyl + hp με τον αλγόριθμο forward selection.

Step Selection ~ Backward with F test

Ο αλγόριθμος αυτός ξεκινάει με το να περιέχει όλες τις μεταβλητές και στη συνέχεια αφαιρεί αυτές όπως και στον forward αλγόριθμο, που δεν προσφέρουν πολλά στη βελτίωση του SSE.

```
> mod_bw = step(mod1, direction = "backward", test="F")
Start: AIC=70,9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

      Df Sum of Sq    RSS    AIC F value    Pr(>F)
- cyl   1  0,0799 147,57 68,915  0,0114 0,91609
- vs    1  0,1601 147,66 68,932  0,0228 0,88142
- carb  1  0,4067 147,90 68,986  0,0579 0,81218
- gear  1  1,3531 148,85 69,190  0,1926 0,66521
- drat  1  1,6270 149,12 69,249  0,2317 0,63528
- disp  1  3,9167 151,41 69,736  0,5576 0,46349
- hp    1  6,8399 154,33 70,348  0,9739 0,33496
- qsec  1  8,8641 156,36 70,765  1,2621 0,27394
<none>          147,49 70,898
- am    1  10,5467 158,04 71,108  1,5016 0,23399
- wt    1  27,0144 174,51 74,280  3,8463 0,06325 .
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Step: AIC=68,92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

      Df Sum of Sq    RSS    AIC F value    Pr(>F)
- vs    1  0,2685 147,84 66,973  0,0400 0,84326
- carb  1  0,5201 148,09 67,028  0,0775 0,78326
- gear  1  1,8211 149,40 67,308  0,2715 0,60754
- drat  1  1,9826 149,56 67,342  0,2956 0,59214
- disp  1  3,9009 151,47 67,750  0,5815 0,45381
- hp    1  7,3632 154,94 68,473  1,0977 0,30615
<none>          147,57 68,915
- qsec  1  10,0933 157,67 69,032  1,5047 0,23292
- am    1  11,8359 159,41 69,384  1,7645 0,19768
- wt    1  27,0280 174,60 72,297  4,0293 0,05716 .
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1
```

και έπειτα από κάποια βήματα καταλήγει στο εξής μοντέλο.

```
Step: AIC=61,31
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
```

mod_bw = wt + qsec + am

Step Selection ~ Both with F test

Με τη χρήση της παραμέτρου both , αναγκάζουμε το μοντέλο πριν αποφασίσει να εισάγει μια μεταβλητή να εξετάσει αν οδηγεί στην εξασθένηση της σημαντικότητας μιας μεταβλητής που είχε εισαχθεί νωρίτερα.

Το αποτέλεσμα είναι το εξής μοντέλο

```
Step: AIC=62,66
mpg ~ wt + cyl + hp

Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>          176,62 62,665
- hp   1     14,551 191,17 63,198 2,3069 0,1400152
+ am   1      6,623 170,00 63,442 1,0519 0,3141799
+ disp 1      6,176 170,44 63,526 0,9784 0,3313856
- cyl   1     18,427 195,05 63,840 2,9213 0,0984801 .
+ carb  1      2,519 174,10 64,205 0,3906 0,5372269
+ drat  1      2,245 174,38 64,255 0,3477 0,5603447
+ qsec   1      1,401 175,22 64,410 0,2159 0,6459140
+ gear   1      0,856 175,76 64,509 0,1315 0,7197354
+ vs    1      0,060 176,56 64,654 0,0092 0,9244706
- wt    1     115,354 291,98 76,750 18,2873 0,0001995 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

$$\text{mod_both} = \text{wt} + \text{cyl} + \text{hp}$$

Συνοπτικά έχουμε ότι τα μοντέλα από το both και forward selection είναι ίδια.

Άρα έχουμε το 1 μοντέλο της μορφής

$$\text{fw_both} = \text{WT} + \text{CYL} + \text{HP}$$

και το μοντέλο από το backward selection

$$\text{bw} = \text{WT} + \text{QSEC} + \text{AM}$$

BEST MODEL ANALYSIS

	MODEL 1 = WT + CYL + HP	MODEL 2 = WT + QSEC + AM
AIC	155	154
Rsq	0,8431	0,8497
Radj	0,8263	0,8336
PRESS	230,0767	231,3035
Cp - mallows	1,146922	0,1026357
Number of Independ.Var.	3	3

Από τα παραπάνω ισχύουν τα εξής.

1) Θέλουμε το μοντέλο με το χαμηλότερο AIC, επομένως αυτό το κριτήριο αποφασίζει το 2o μοντέλο.

2) Τα Rsq είναι πολύ κοντά όμως τα Rsq-adj είχουν διαφορά 0,1 μεταξύ τους. Το Rsq-adjusted είναι καλύτερο κριτήριο επιλογής καθώς όσο περισσότερες μεταβλητές βάζουμε το Rsq αυξάνεται. Το adjusted Rsq όμως προσπαθεί να μειώσει το ρυθμό αύξησης του Rsq, λόγω του ότι λαμβάνει υπ'οψιν και τον αριθμό των παραμέτρων.

Επομένως κοιτώντας τα Rsq, Rsq-adj διαλέγουμε το μοντέλο 2.

3) To PRESS χρησιμοποιείται για σύγκριση μοντέλων ως προς την ικανότητα πρόβλεψης νέων παρατηρήσεων. Γενικά, προτιμάται το μοντέλο με το μικρότερο PRESS. To PRESS εκφράζει ουσιαστικά το Rsq-pred το οποίο ισούται με:

$$\text{Rsq-pred} = 1 - \text{PRESS/SST} \text{ και στην περίπτωση αυτή διαλέγουμε το μεγαλύτερο}$$

Rsq-pred.

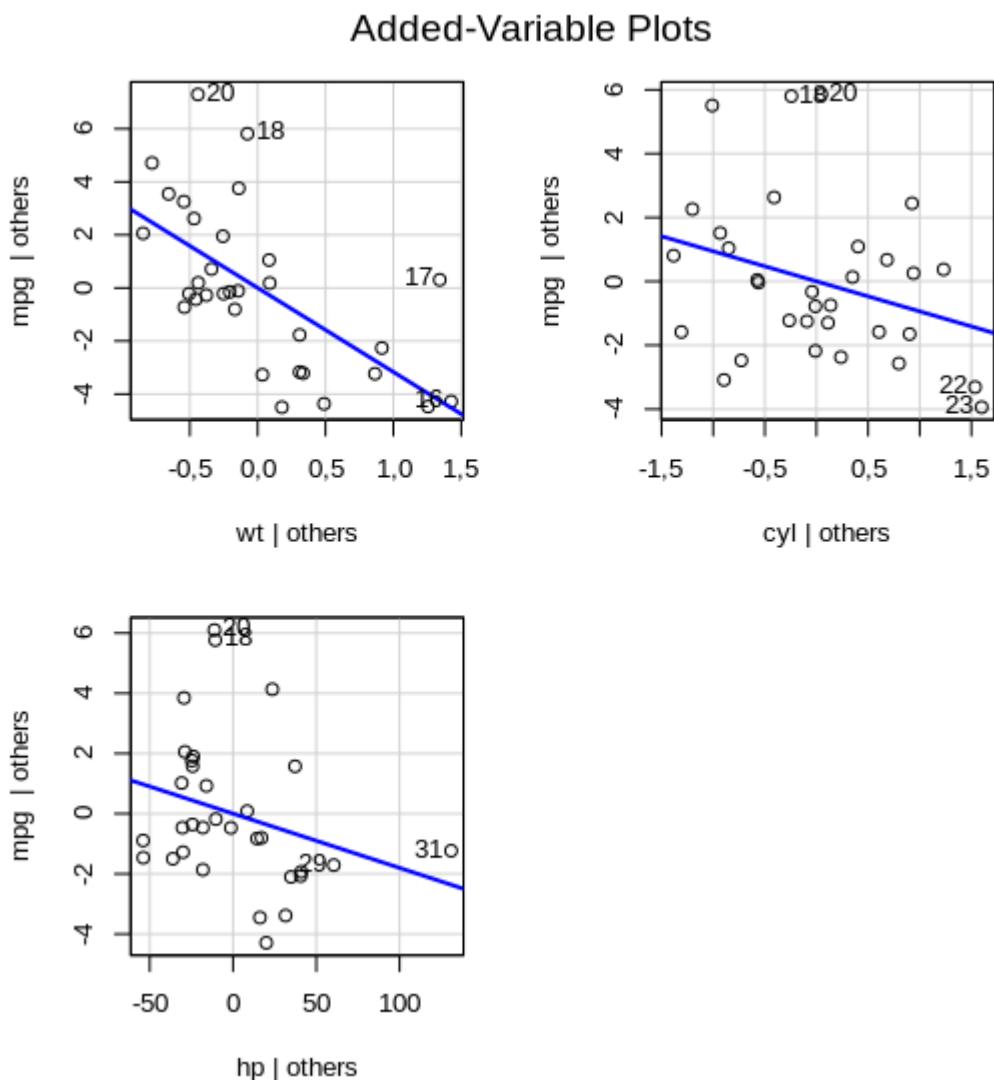
Σύμφωνα με αυτό το κριτήριο διαλέγουμε το μοντέλο 1.

4) Τέλος το cp-mallows διαλέγει το μοντέλο 1, γιατί έχει μικρότερη τιμή.

Χρήση ADDED VARIABLE PLOTS

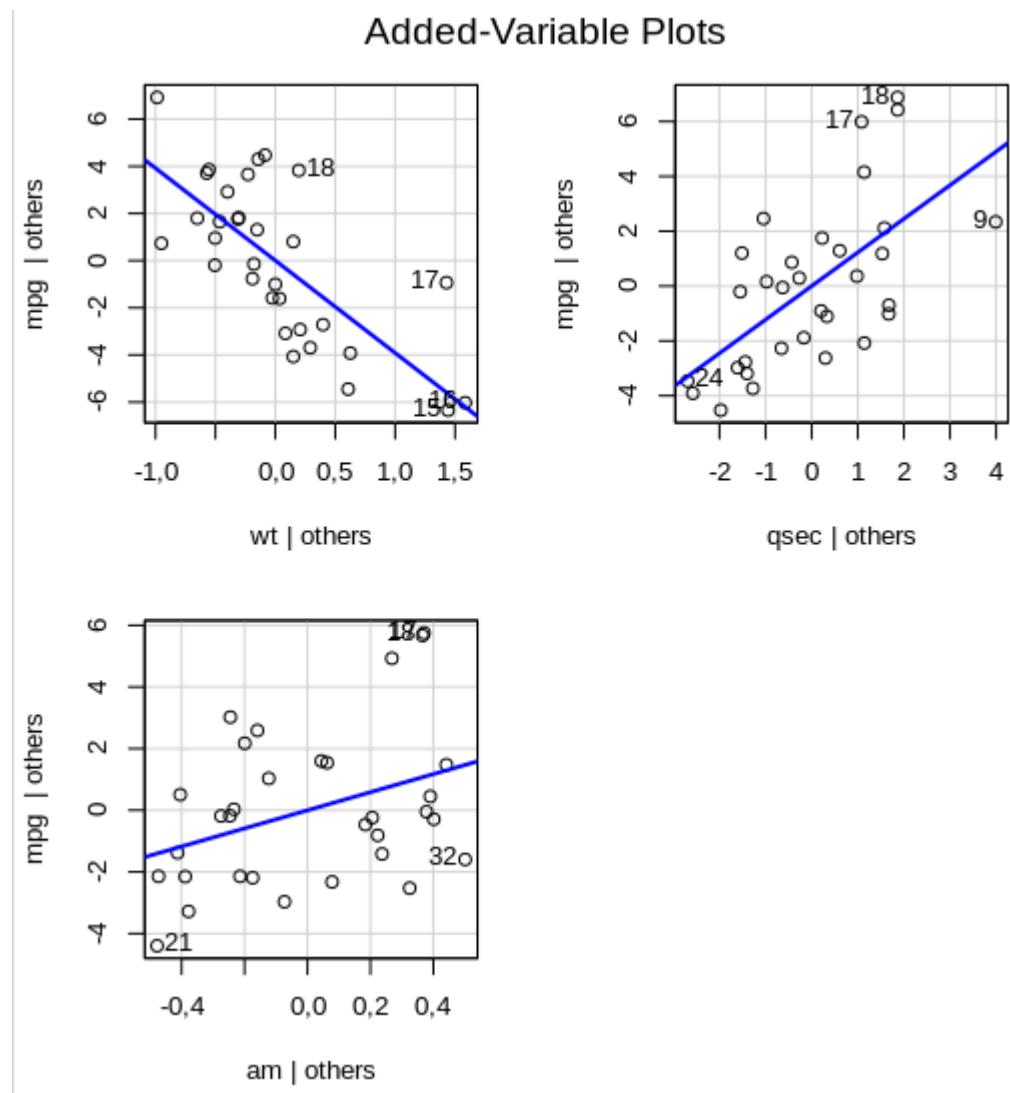
Για το μοντέλο 1

avPlots(mod_fw)



Για το 2o μοντέλο έχουμε

avPlots(mod_bw)



Συμπεράσματα:

Στο 2o μοντέλο έχουμε για τις 2 μεταβλητές αρκετά καλοσχηματισμέμη ευθεία σε αντίθεση με το 1o μοντέλο το οποίο μόνο στην 1η μεταβλητή ορίζει μια καλή ευθεία.

Επομένως από αυτά τα διαγράμματα θα επιλέγαμε το μοντέλο 2.

Component Residual Plots

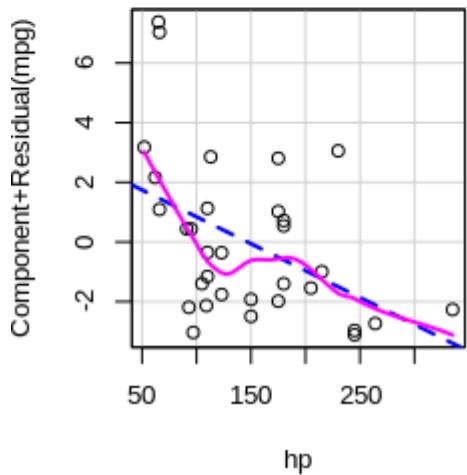
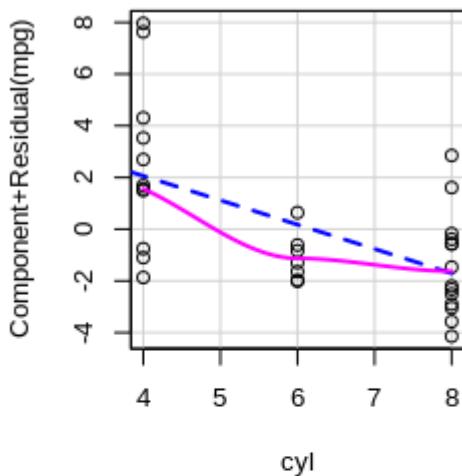
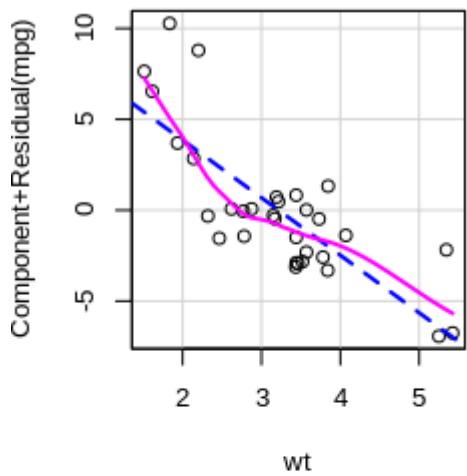
Επίσης έχουμε και τις γραφικές παραστάσεις μερικών υπολούπων που δίνονται με την εντολή

```
>crPlots(mod_fw)
```

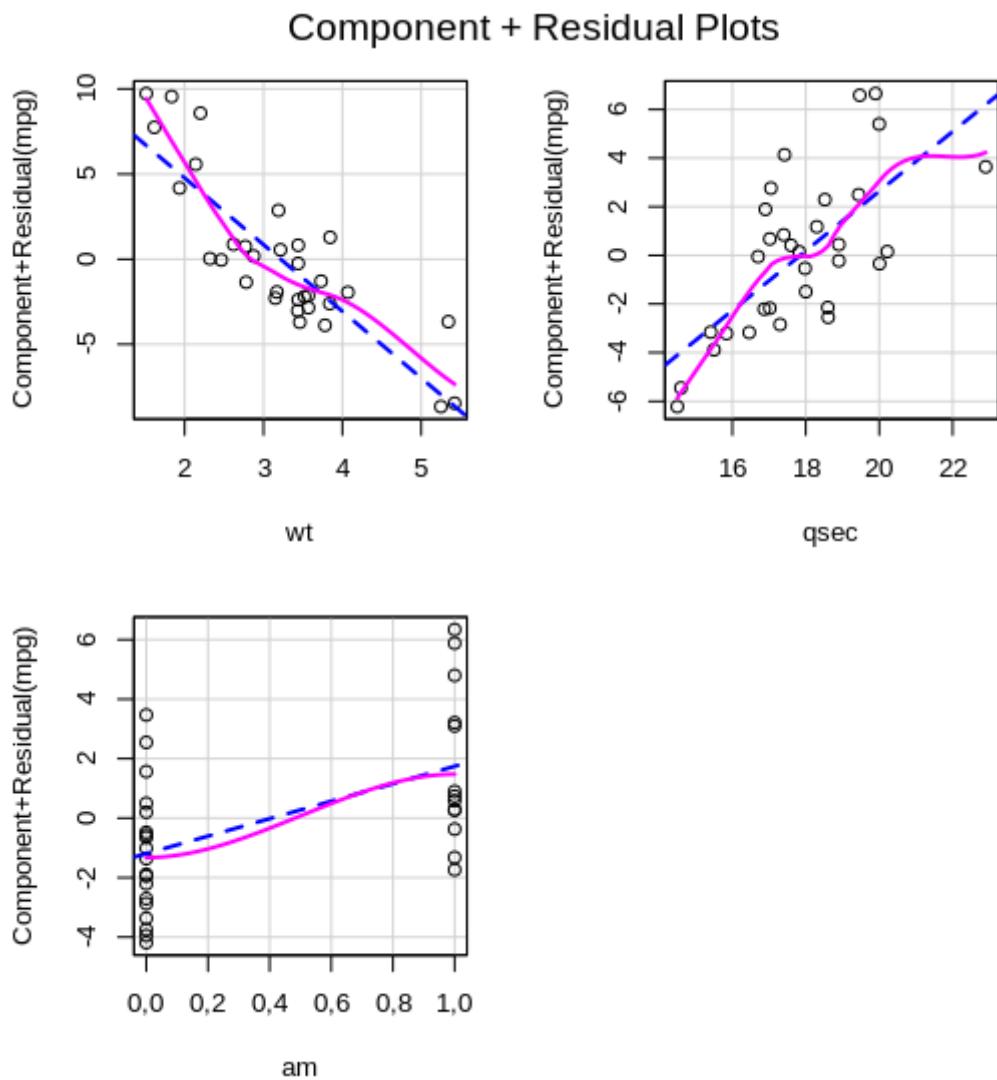
```
>crPlots(mod_bw)
```

Για το 1o μοντέλο

Component + Residual Plots



Για το 2o μοντέλο



Και σε αυτά τα διαγράμματα φαίνεται να υπερέχει το μοντέλο 2, καθώς πάλι τα δεδομένα του τείνουν σε καλύτερες ευθείες από ότι τα διαγράμματα του μοντέλου 1.

Θα επιλέξουμε μετά από όλα αυτά ως καλύτερο μοντέλο , το μοντέλο 2 με τις μεταβλητές
WT + QSEC + AM.

Το οποίο είναι αυτό που λάβαμε από το backward elimination.

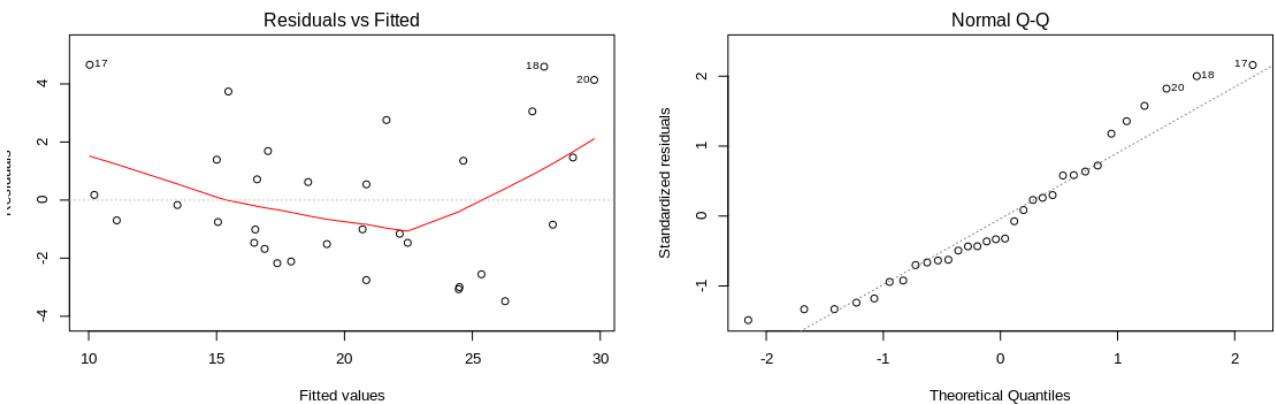
ΕΡΩΤΗΜΑ 3

Με βάση το τελικό μοντέλο να αξιοποιηθούν διαγνωστικές τεχνικές για την εξέταση της καταλληλότητάς του, την πιθανή παρουσία άτυπων σημείων ή σημείων επιρροής, αν χρειάζονται μετασχηματισμοί ή περαιτέρω βελτιώσεις του μοντέλου και να δωθούν ερμηνείες.

Το τελικό μας μοντέλο είναι το $WT + QSEC + AM = mod_bw$ και πάνω σε αυτό θα αξιοποιήσουμε τις διαγνωστικές τεχνικές.

ΑΠΑΝΤΗΣΗ:

Θα ξεκινήσουμε τις διαγνωστικές τεχνικές για το πόσο κατάλληλο είναι το μοντέλο που διαλέξαμε, μέσω διαγραμμάτων υπολοίπων.



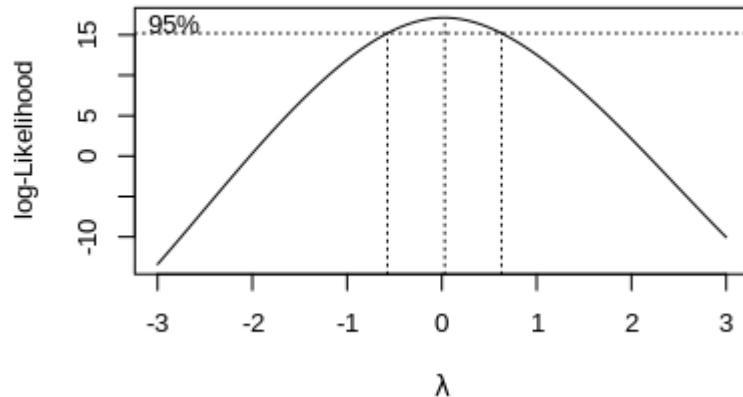
Παρατηρούμε πως τα διαγράμματα που πέρνουμε δεν είναι ικανοποιητικά. Δηλαδή τα λαθη δε τηρουν την ομοσκεδαστικότητα και το δεύτερο διάγραμμα δείχνει πως δεν ακολουθούν την κανονική κατανομή.

Επομένως θα κάνουμε ένα μετασχηματισμό box-cox στο μοντέλο μας. Με τις παρακάτω εντολές:

```
library(MASS)
```

```
bc = boxcox(mod_bw,lambda = seq(-3,3))
```

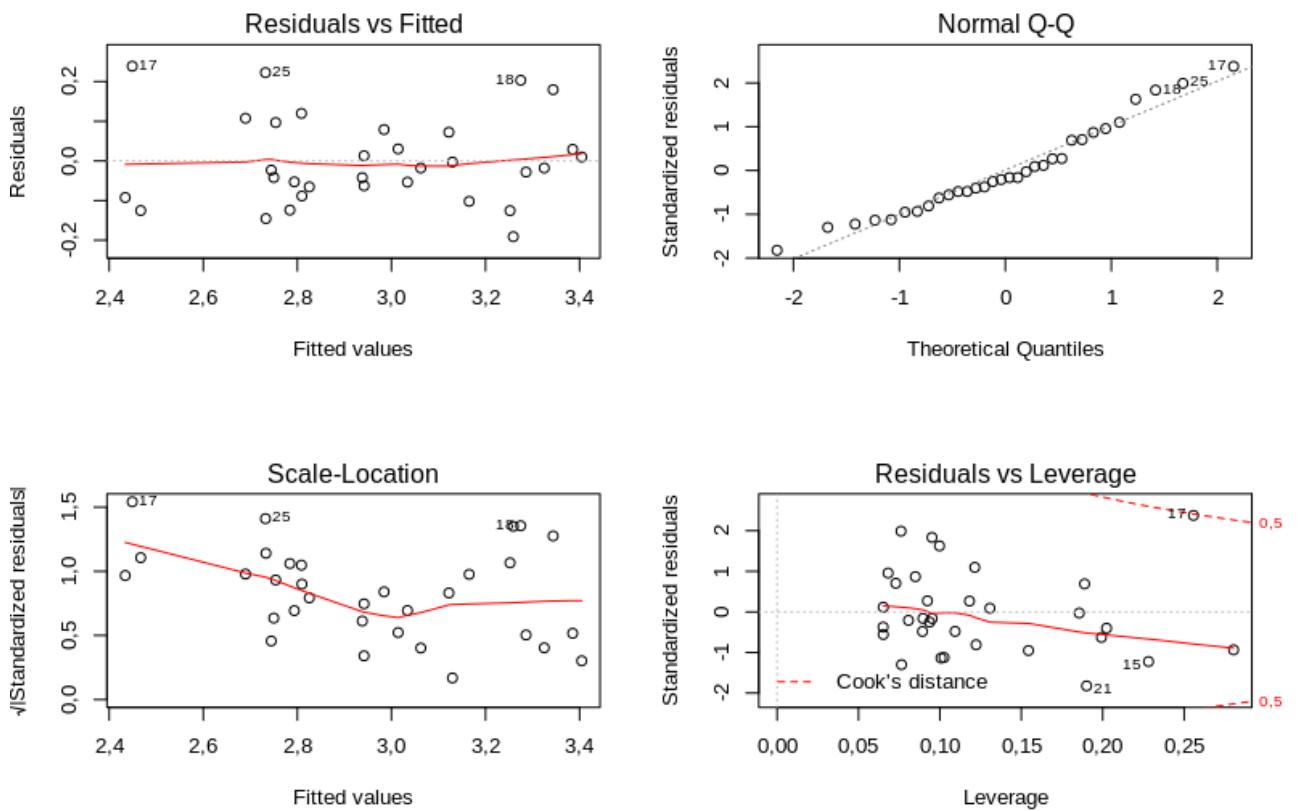
παίρνουμε το εξής γράφημα:



Από το οποίο βλέπουμε πως η πιθανοφάνεια μεγιστοποιείται για $\lambda=0$, επομένως θα κάνουμε ένα μετασχηματισμό της μορφής $\log(\text{mpg})$. Με την παρακάτω εντολή , πετυχάινουμε τον μετασχηματισμό μας.

```
mod_bw.inv = lm(log(mpg) ~ wt+cyl+am)
```

Από αυτό το μοντέλο λαμβάνουμε τα εξής διαγράμματα για τα residuals.



Παρατηρούμε πως πλέον τηρείται η ομοσκεδαστικότητα σε μεγαλύτερο βαθμό στο 1ο διάγραμμα. Επιπλέον τα residuals πλέον ακολουθούν την κανονική κατανομή και δε ξεφεύγουν πολύ από την ευθεία.

Όσον αφορά την πολυσυγγραμμικότητα έχουμε τα εξής συμπεράσματα από την εντολή VIF.

```
> vif(mod_bw.linv)
      wt      cyl      am
 3,609011 2,584066 1,924955
```

Παρατηρούμε πως στο νέο μας μοντέλο δεν υπάρχει πολυσυγγραμμικότητα καθώς όλες οι τιμές είναι κάτω του 5.

Σημεία Επιρροής:

DFBETAS

Για $n=32$, για να ασκεί μια παρατήρηση επιρροή σε κάποια επεξηγηματική μεταβλητή πρέπει να έχει μεγαλύτερη τιμή από 0.35.

```
> dfb <- dfbetas(mod_bw.inv)
> print(dfb[dfb > 0.35])
[1] 0,3721751 1,3126018 0,3508274 0,5480169 0,6946978
```

Παρακάτω βλέπουμε πως μόνο μια τέτοια παρατήρηση υπάρχει.

DFFITS

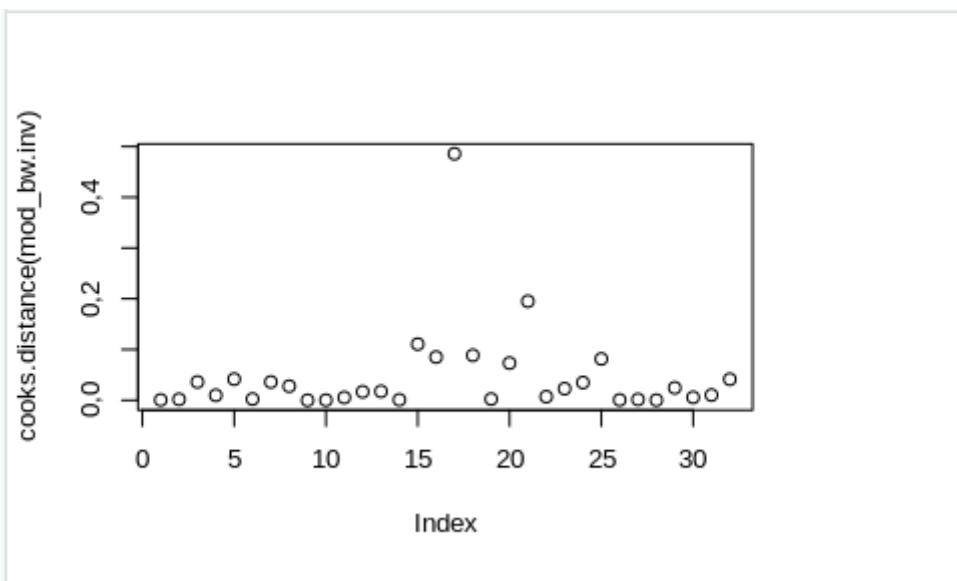
Για 3 μεταβλητές και 32 παρατηρήσεις για να ασκεί ένα σημείο επιρροή πρέπει να έχει τιμή μεγαλύτερη από 6.5.

```
> dff <- dffits(mod_bw.inv)
> print(dff[dff > 6.5])
named numeric(0)
```

Βλέπουμε πως δεν υπάρχει τέτοια τιμή.

Cook's Distance

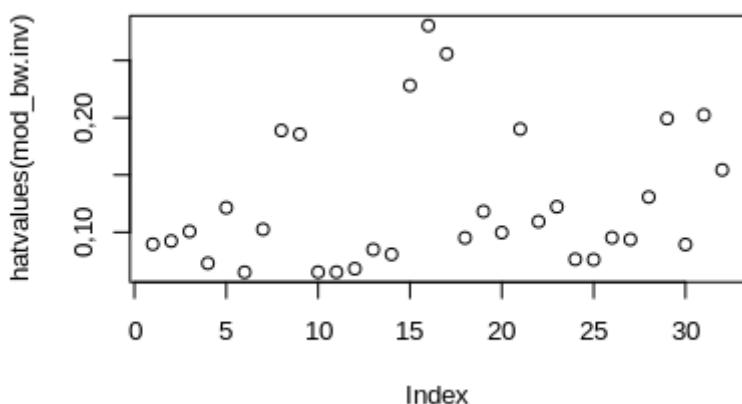
Η απόσταση cook δε μας βρίσκει κάποιο σημείο επιρροής , καθώς δεν ξεπερνάει κάποια τιμή D_i το 1.



Στην παραπάνω Γ.Π. βλέπουμε πως έχουμε ένα σημείο επιρροής , το οποίο μπορεί να μας επηρεάσει στην προσαρμογή του μοντέλου. Παρόλα αυτά η συνολική εικόνα είναι αρκετά βελτιωμένη.

Hat-values plot

Για να θεωρήσουμε κάποιο σημείο ως σημείο μόχλευσης πρέπει να ισχύει, $h_{ii} > 2p / n$. Για $p=3$ και $n = 32$ έχουμε, $h_{ii} > 0.1875$. Παρατηρούμε ότι έχουμε περίπου 3 σημεία που επηρεάζουν το μοντέλο μας και είναι τα 16,17,18.



Μετά από όλη αυτήν την ανάλυση, καταλήξαμε ότι το καλύτερο μοντέλο μας περιέχει τις μεταβλητές, $wt + cyl + am$ και έχουμε κάνει ένα μετασχηματισμό box cox για $\lambda=0$, άρα $\log(mpg)$, ώστε να τηρούντε οι υποθέσεις για την κανονικότητα του γραμμικού μας μοντέλου.