Πριν ξεκινήσω, εγκαθιστώ ορισμένες βιβλιοθήκες που θα χρειαστώ στην γραφική μελέτη των μοντέλων μου.

```
install.packages("olsrr")
library(olsrr)
library(car)
library(ggcorrplot)
library(ggplot2)



file<-read.table('/content/vehicles.txt',header=TRUE)
attach(file)
```

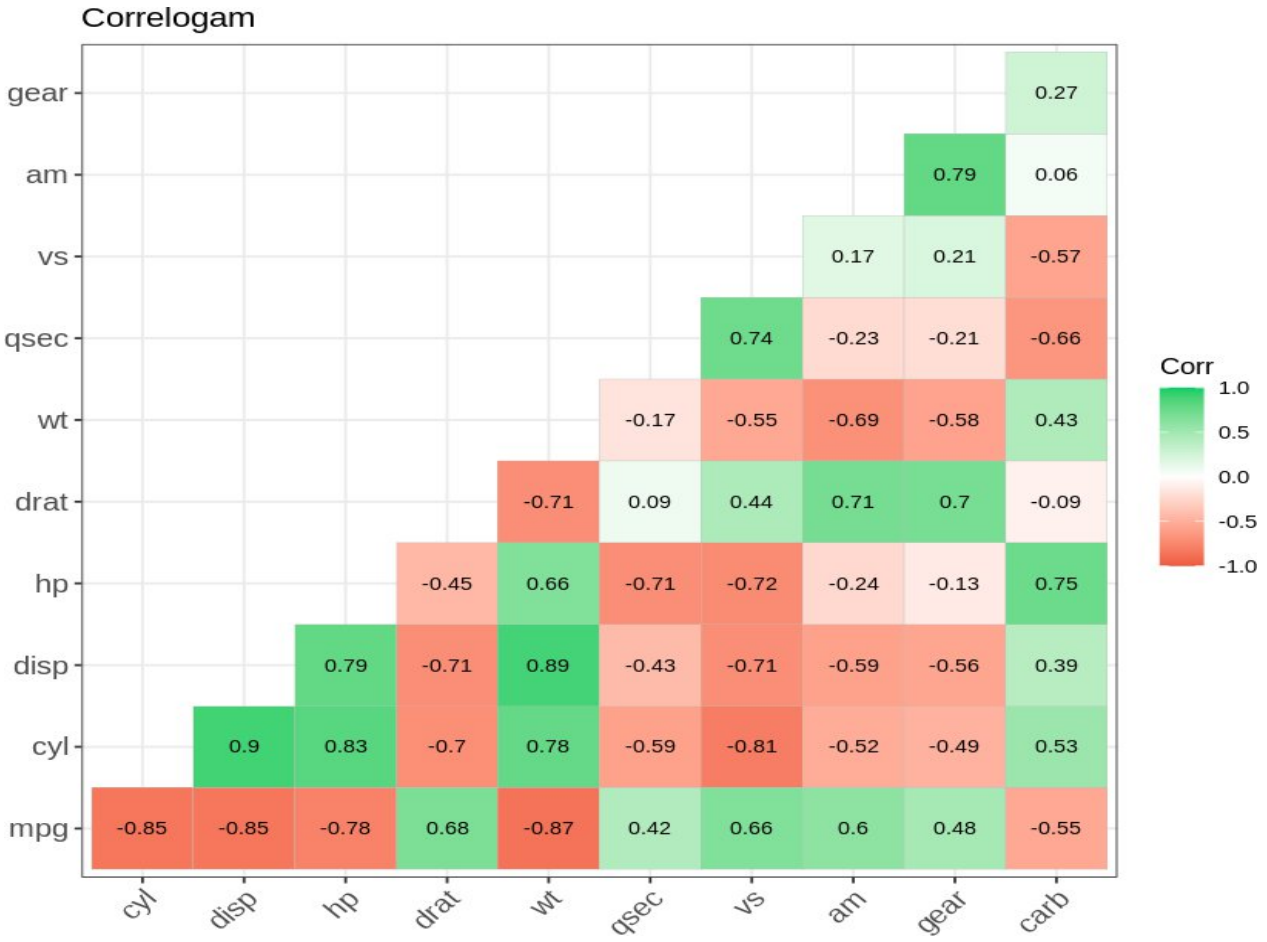
Το αρχικό μου μοντέλο είνα αυτό το οποίο μοντελοποιεί την εξαρτημένη μεταβλητή “mpg” κάνοντας χρήστη όλων των ανεξάρτητων(επεξηγηματικών) μεταβλήτων, δηλαδή το

Mod1=mpg ~ cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb

το οποίο υπολογίζω με τον κώδικα

```
mod1<-lm(mpgcyl+disp+hp+drat+wt+qsec+vs+am+gear+carb)
```

Με μια αρχική εξέταση των correlation (εικόνα 1) μεταξύ των χαρακτηριστικών μου, παρατηρώ καλή γραμμική συσχέτιση ανεξάρτητων και εξαρτημένης μεταβλητής για την πλειοψηφία των χαρακτηριστικών μου, με μερικές εξαιρέσεις όπως τα “qsec”,”gear” κτλπ. Βλέπω μεγάλες τιμές όμως και μεταξύ των επεξηγηματικών μεταβλητών, κάτι που προμηνύει την ύπαρξη πολυσυγγραμμικότητας . Επιβεβαιώνω αυτή την υπόθεση παρατηρώντας πως οι τιμές των VIF είναι κατά βάση αρκετά μεγαλύτερες απο 5.



VIF

cyl	displ	hp	drat	wt	qsec	vs	am	gear	carb
15.4	21.6	9.8	3.4	15.2	7.5	5.0	4.6	5.4	7.9

Εικόνα 1) Correlogram και VIF τα οποία σχεδίασα με τον ακόλουθο κώδικα

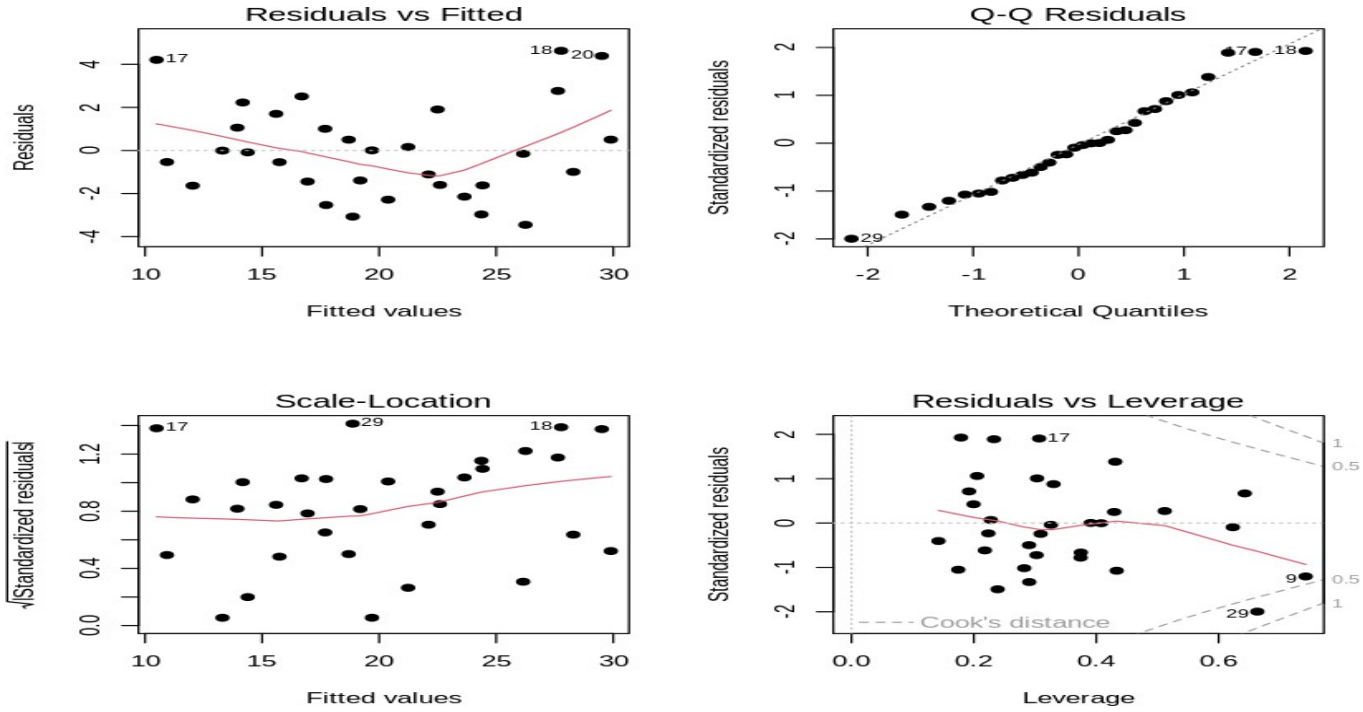
```
corr<-round(cor(subset(file,select=-c(car))),2)
ggcorrplot(corr,type='lower',lab=TRUE,lab_size=3,method='square'
,colors=c("tomato2","white","springgreen3"),title='Correlogam',
ggtheme=theme_bw)
vif(mod1)
```

Για τον έλεγχο των προυποθέσεων του μοντέλου θα εξετάσω τα γραφήματα της εικόνας 2.

- Σχετικά με την ομοσκεδαστικότητα κοιτώντας το Residuals vs Fitted plot και το Scale-Location που χρησιμοποιεί κανονικοποιημένα residuals δεν φαίνεται να σχηματίζεται κάποιο μοτίβο για τα residuals και επομένως η συνθήκη ικανοποιείται.

- Απο το QQ-plot για τα residuals είναι εμφανές πως ικανοποιείται η συνθήκη για κανονική κατανομή των σφαλμάτων.

- Από το residuals vs Leverage Plot βλέπουμε χαμηλά επίπεδα μόχλευσης, ενώ οι παρατηρήσεις 29 και 9 είναι στα όρια της απόστασης Cook και επομένως είναι πιθανά σημεία επιρροής.



Εικόνα 2) Διαγράμματα για τον έλεγχο των συνθηκών του μοντέλου με τη χρήση του κώδικα

```
par(mfrow = c(2,2) )
plot(mod1, pch=19)
```

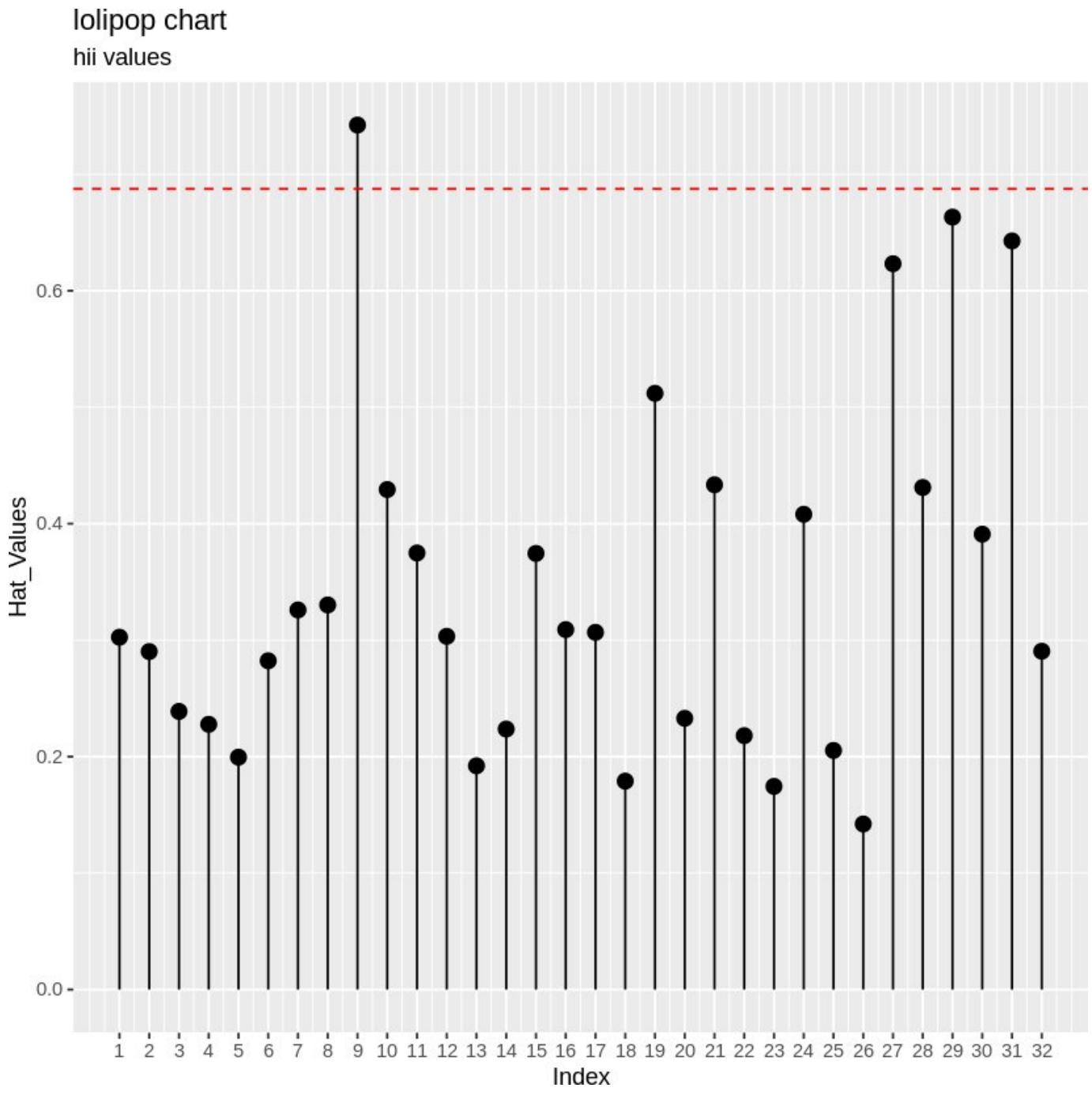
Για τα πιθανά σημεία επιρροής θα χρησιμοποιήσω τα ακόλουθα μέτρα.

- Τιμές hii
- Απόσταση Cook
- DFFiTS
- DFBETAS

- Ξεκινώντας με τα hii, παρατηρώ πως n=32 & p=11 επομένως για hii > $\frac{2p}{n}$ δηλαδή για hii> 0.6875 η παρατήρηση i θα θεωρείται ως πιθανό σημείο επιρροής. Κάνω τον κατάλληλο έλεγχο στην R μέσω της εντολής

```
hats <- hatvalues(mod1)
hats_df <- data.frame(Index = seq_along(hats), Hat_Values =
hats)
ggplot(hats_df, aes(x = Index, y = Hat_Values)) +
geom_segment(aes(x=Index,xend=Index,y=0,yend=hats))+geom_point(s
ize=3)+scale_x_continuous(breaks =
hats_df$Index)+labs(title="lolipop chart",subtitle="hii
values")+
geom_hline(yintercept = 0.6875, linetype = "dashed", color =
"red")
```

και παίρνω την εικόνα 3) απ'όπου βλέπω πως μόνο η παρατήρηση 9 μπορεί να θεωρηθεί σαν πιθανό σημείο επιρροής.

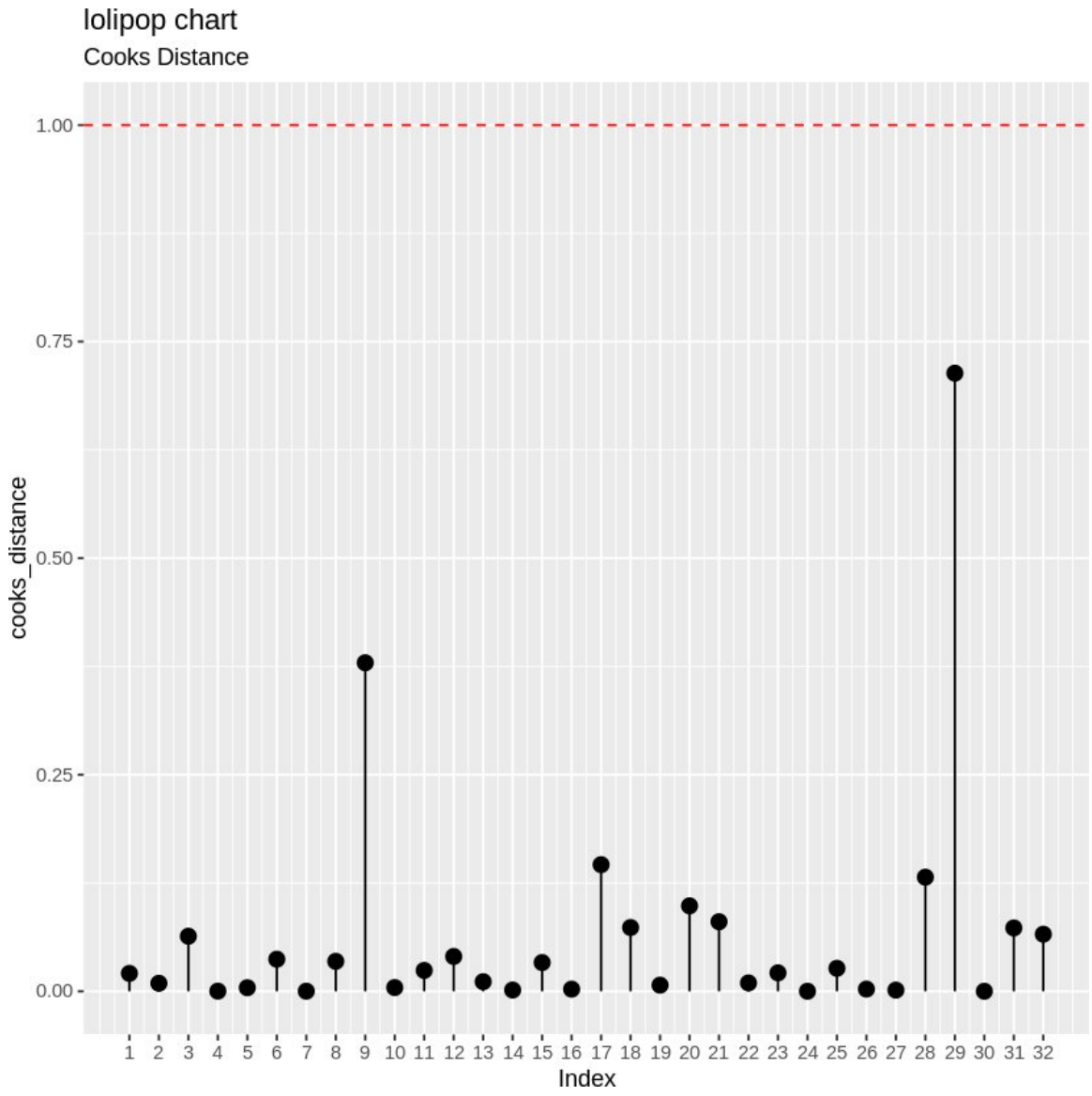


Εικόνα 3) Lollipop Chart για τα hii values

• Για την απόσταση cook πάλι μέσω της χρήσης R με την εντολή

```
cooks_d <- cooks.distance(mod1)
cooks_df <- data.frame(Index = seq_along(cooks_d),
cooks_distance = cooks_d)
ggplot(cooks_df, aes(x = Index, y = cooks_distance)) +
geom_segment(aes(x=Index,xend=Index,y=0,yend=cooks_distance))+ge
om_point(size=3)+scale_x_continuous(breaks =
hats_df$Index)+labs(title="lolipop chart",subtitle="Cooks
Distance")+
geom_hline(yintercept = 1, linetype = "dashed", color = "red")
```

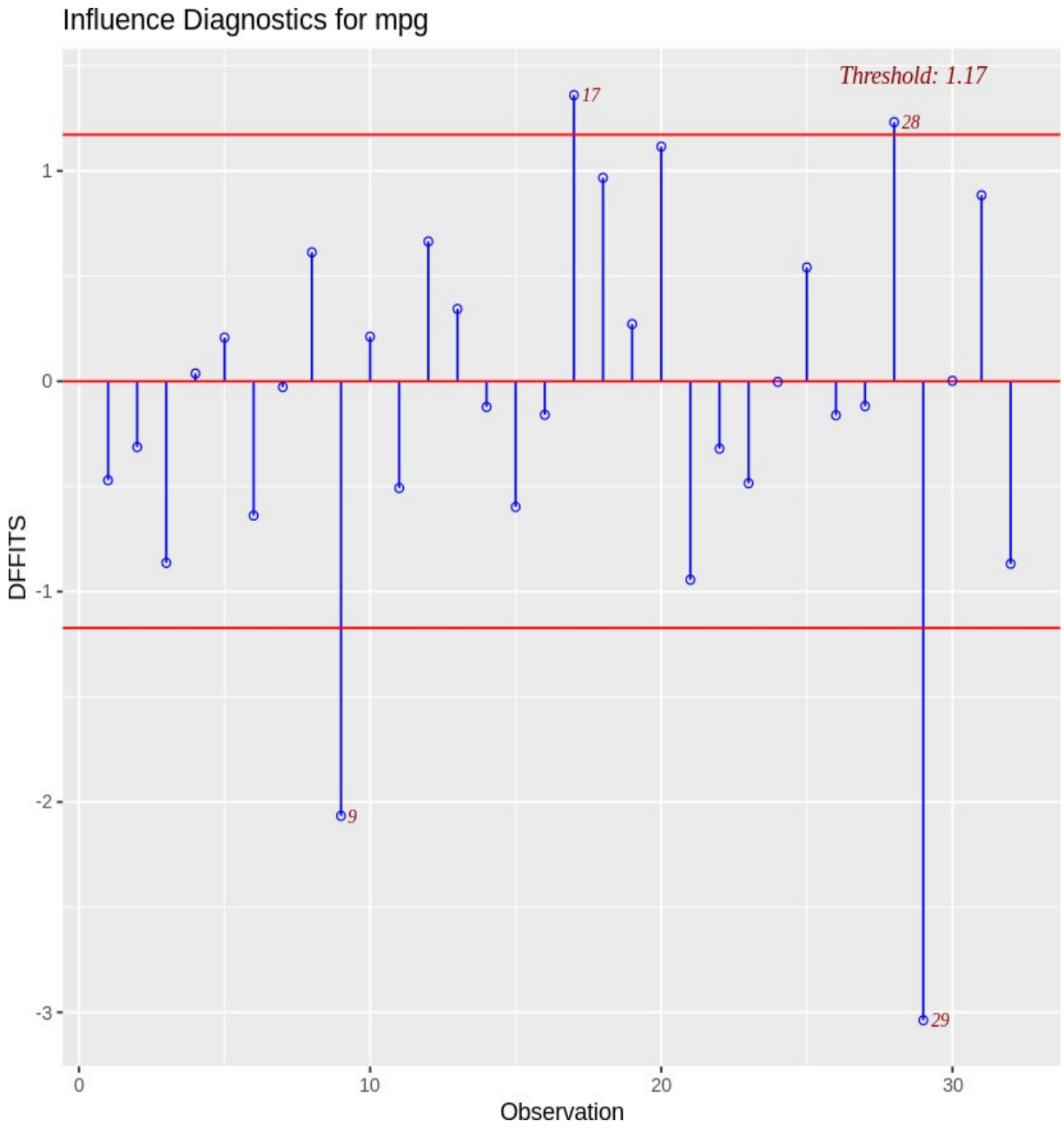
παίρνω το ακόλουθο διάγραμμα και αφού καμία παρατήρηση δεν έχει μεγαλύτερη απόσταση COOK από 1, δεν παίρνω κάποιο πιθανό σημείο επιρροής.



• Στη συνέχεια εξετάζω τα DFFITSi. Για το μοντέλο μου, καθώς n=32 & p=1 έχω πως αν |DFFITSi| > 1.173 τότε μπορώ να θεωρήσω την παρατήρηση i σαν πιθανό σημείο επιρροής. Με χρήση της βιβλιοθήκης olsrr στην R και με την εντολή

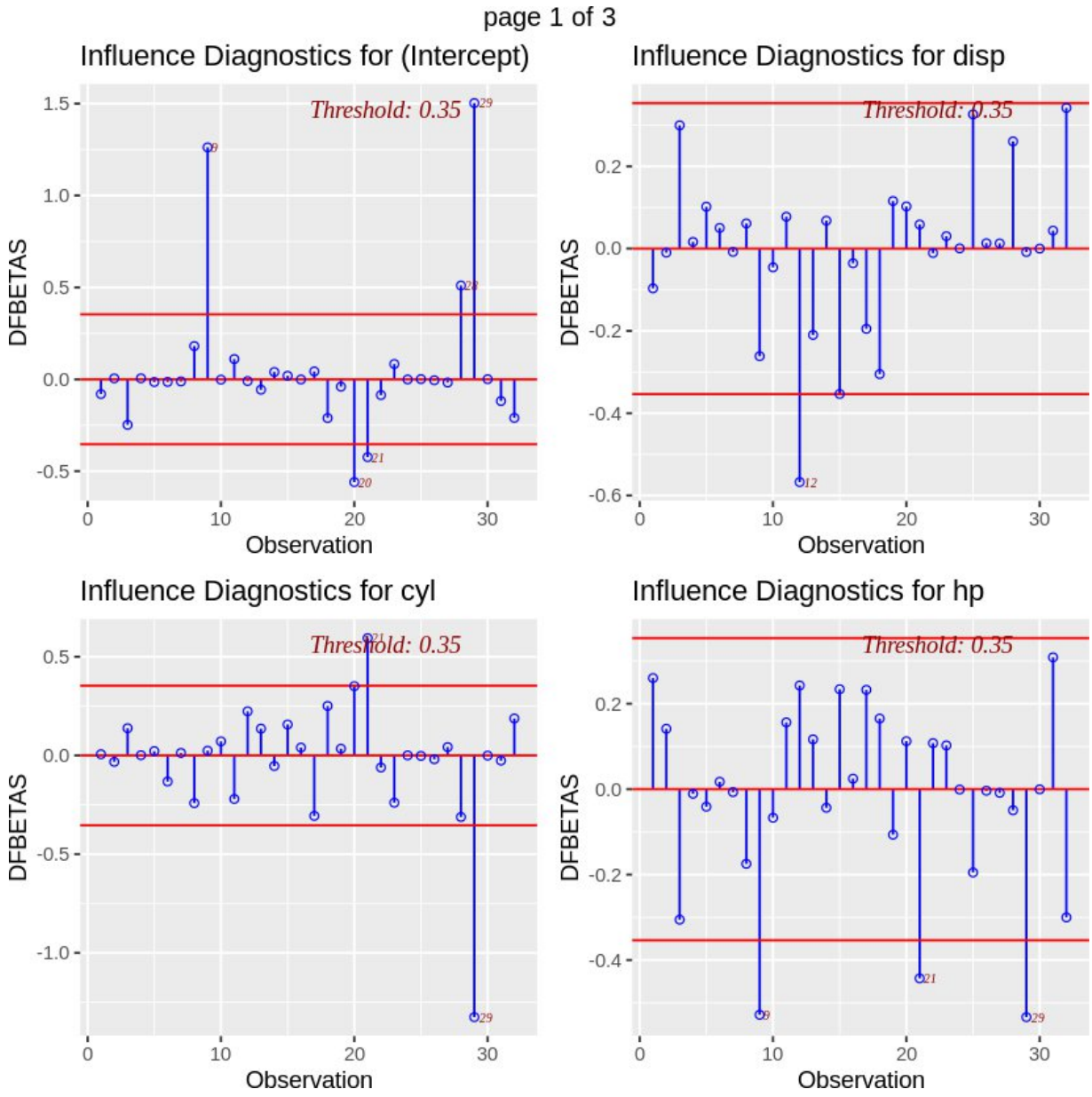
```
ols_plot_dffits(mod1)
```

σχεδιάζω το παρακάτω διάγραμμα, και παρατηρώ πως πιθανά σημεία επιρροής είναι τα 17,28,9,29

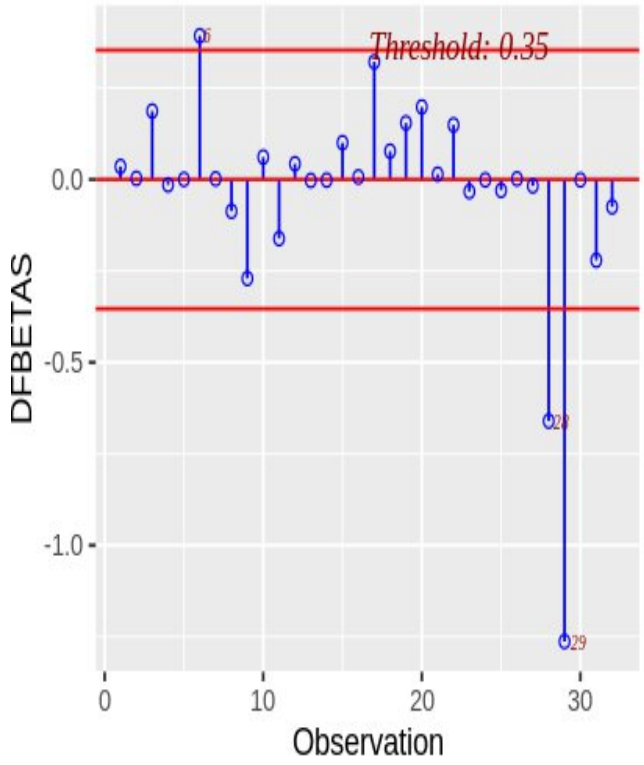


• Τέλος θα κοιτάξω τα DFBETAS. Για το συγκεκριμένο μοντέλο, αν |DFBETASij|>0.353 τότε η i παρατήρηση μπορεί να έχει μεγάλη επιρροή στην εκτιμηση του bj. Πάλι με τη χρήση της βιβλιοθήκης olsrr παίρνω τα γραφήματα για τα DFBETAS μέσω της εντολής

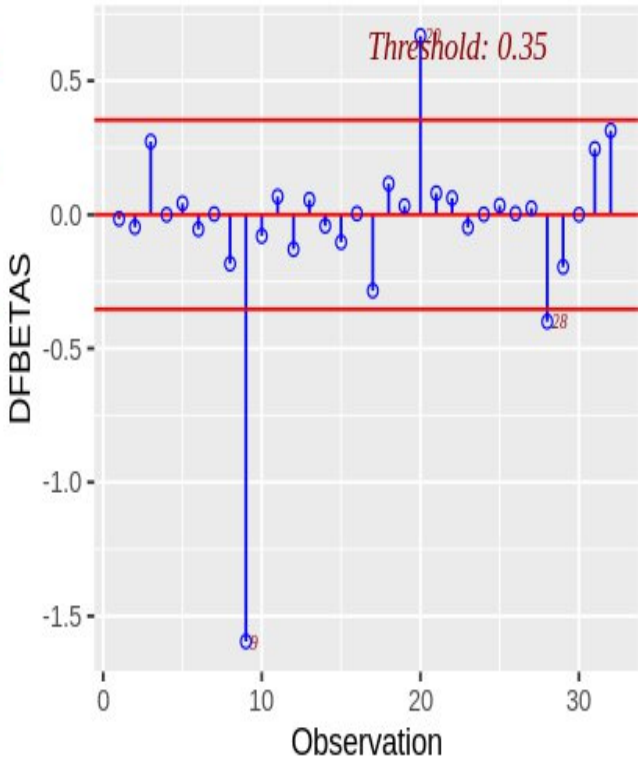
```
ols_plot_dfbetas(mod1)
```



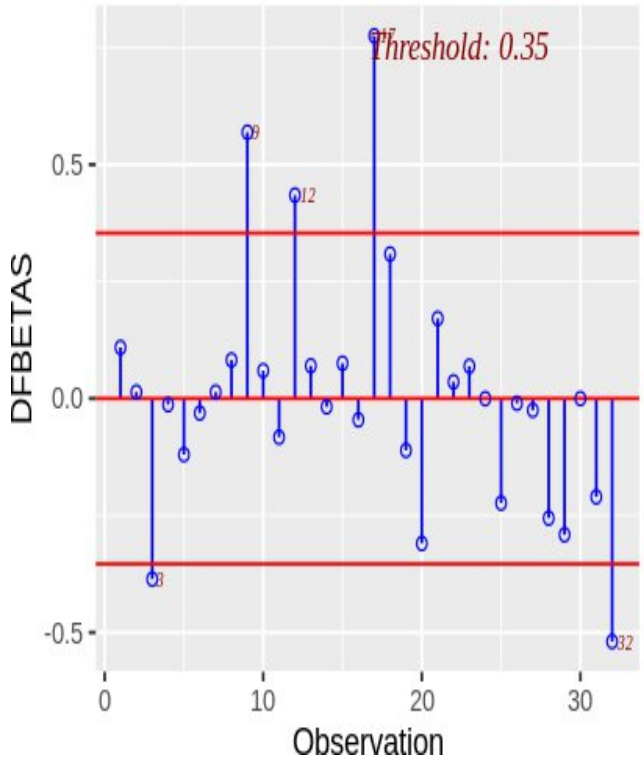
Influence Diagnostics for drat



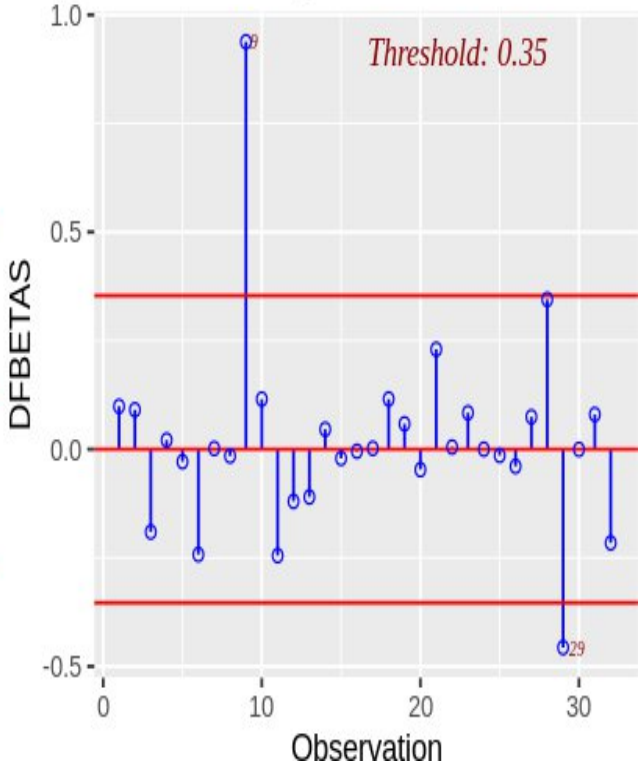
Influence Diagnostics for qsec

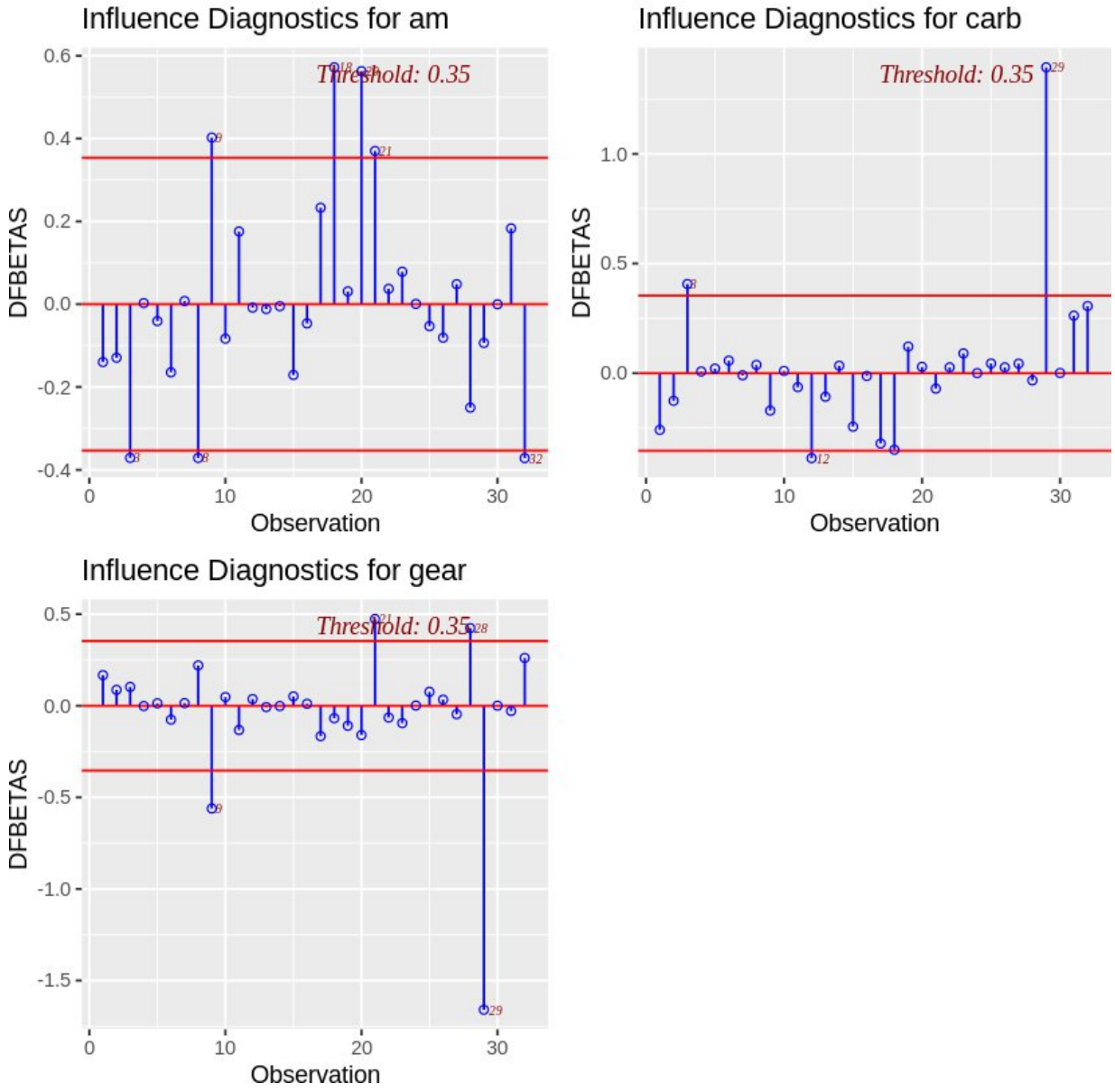


Influence Diagnostics for wt



Influence Diagnostics for vs





Κοιτώντας τα παραπάνω γραφήματα, και συνδιάζωντας όλα τα προηγούμενα, βλέπω οτι στις περισσότερες περιπτώσεις, οι παρατηρήσεις που αποτελούν πιθανά σημεία επιρροής είναι οι 9,28,29.

Μια ένδειξη ότι το αρχικό μου μοντέλο δεν είναι κατάλληλο, είναι πως είδαμε μεγάλες συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών και μεγάλες τιμές VIF δείχνωντας μας την ύπαρξη πολυσυγγραμμικότητας. Ελέγχωντας και τα p- values των t-test(με την εντολή summary(mod1)) για κάθε χαρακτηριστικό βλέπω πως στην πλειοψηφία είναι αρκετά μεγάλα οπότε θα πρέπει να γίνουν ορισμένες αφαιρέσεις στις μεταβλητές που θα κρατήσω για να βελτιώσω το μοντέλο.

```
summary(mod1)
```

variable	cyl	displ	hp	drat	wt	qsec	vs	am	Gear	Carb
p-value	0.9161	0.4635	0.3350	0.6353	0.0633	0.2739	0.8814	0.2340	0.6652	0.8122

Προκειμένου να πετύχω το παραπάνω, θα χρησιμοποιήσω τις τεχνικές Forward Selection όπου ξεκινάω με το τετριμμένο μοντέλο που περιέχει μόνο το intercept και σε κάθε βήμα προσθέτων την πιο σημαντική μεταβλήτη(βάση AIC ή Cp-Mallows), την τεχνική Backward Elimination όπου ξεκινάω με το πλήρες μοντέλο και σε κάθε βήμα με την ίδια λογική αφαιρώ την κατάλληλη μεταβλητή και τέλος την τεχνική Stepwise Selection που είναι συνδιασμός των παραπάνω.

Ξεκινάω με την Backward Elimination.

• Backward Elimination

```
bwd<-step(mod1,direction='backward')
mod_bwd=lm(formula=mpg~wt+qsec+am)
summary(mod_bwd)
ols_mallows_cp(mod_bwd,mod1)
AIC(mod_bwd)
```

```
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
0.10263573946057
154.119370868901
```

Το μοντέλο που παίρνω με αυτη τη τεχνική, βάση των παραπάνω εντολών και αποτελεσμάτων στην R, είναι το mpg=9.62-3.92wt+1.22qsec+2.93am με R-squared=0.8497, Adjusted R-squared=0.8336, AIC=154.12, Cp-Mallows=0.103.

Συνεχίζω με την μέθοδο Forward-Selection

• Forward Selection

```
fwd<-step(lm(mpg~1),y~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb,direction='forward')
summary(mod_fwd)
ols_mallows_cp(mod_fwd,mod1)
AIC(mod_fwd)
```

```

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
1.14692198042015
155.476628510258
```

Με τη μέθοδο αυτή συγκλίνουμε στο μοντέλο mpg=38.75-3.16wt-0.94cyl-0.02hp με R-squared=0.8431, Adjusted R-squared=0.8263, Cp-Mallows=1.147 και AIC=155.48

Τέλος κοιτάμε τη μέθοδο Stepwise Selection

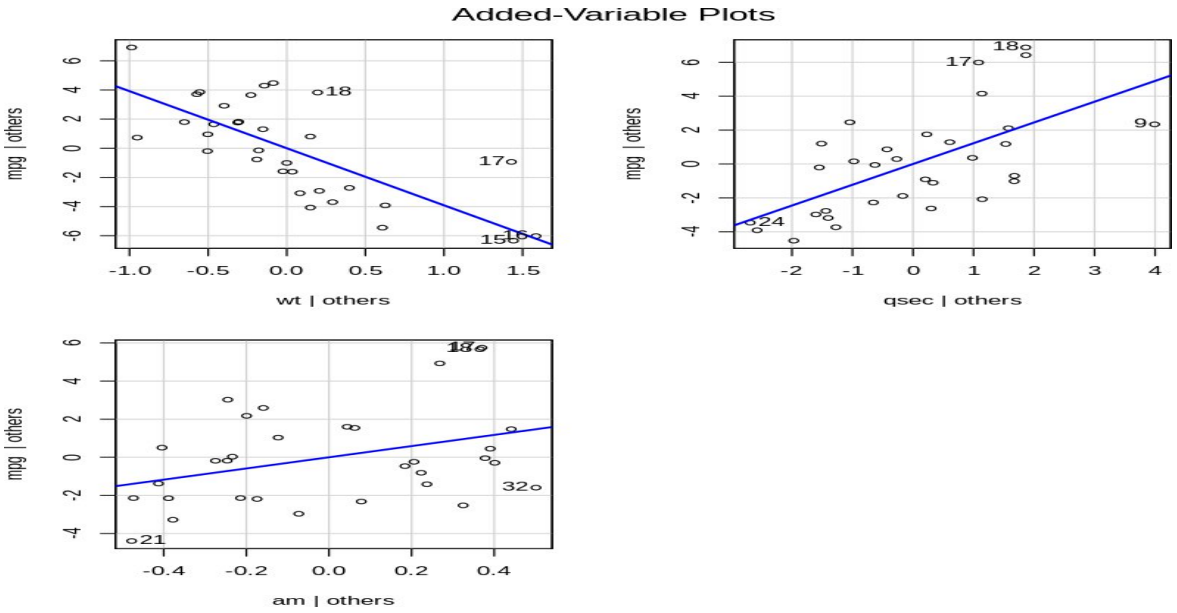
• Stepwise Selection

```
fwd_bwd<-step(mod1,direction='both')
```

Η μέθοδος αυτή συγκλίνει στο ακριβώς ίδιο μοντέλο με την Backward Elimination.

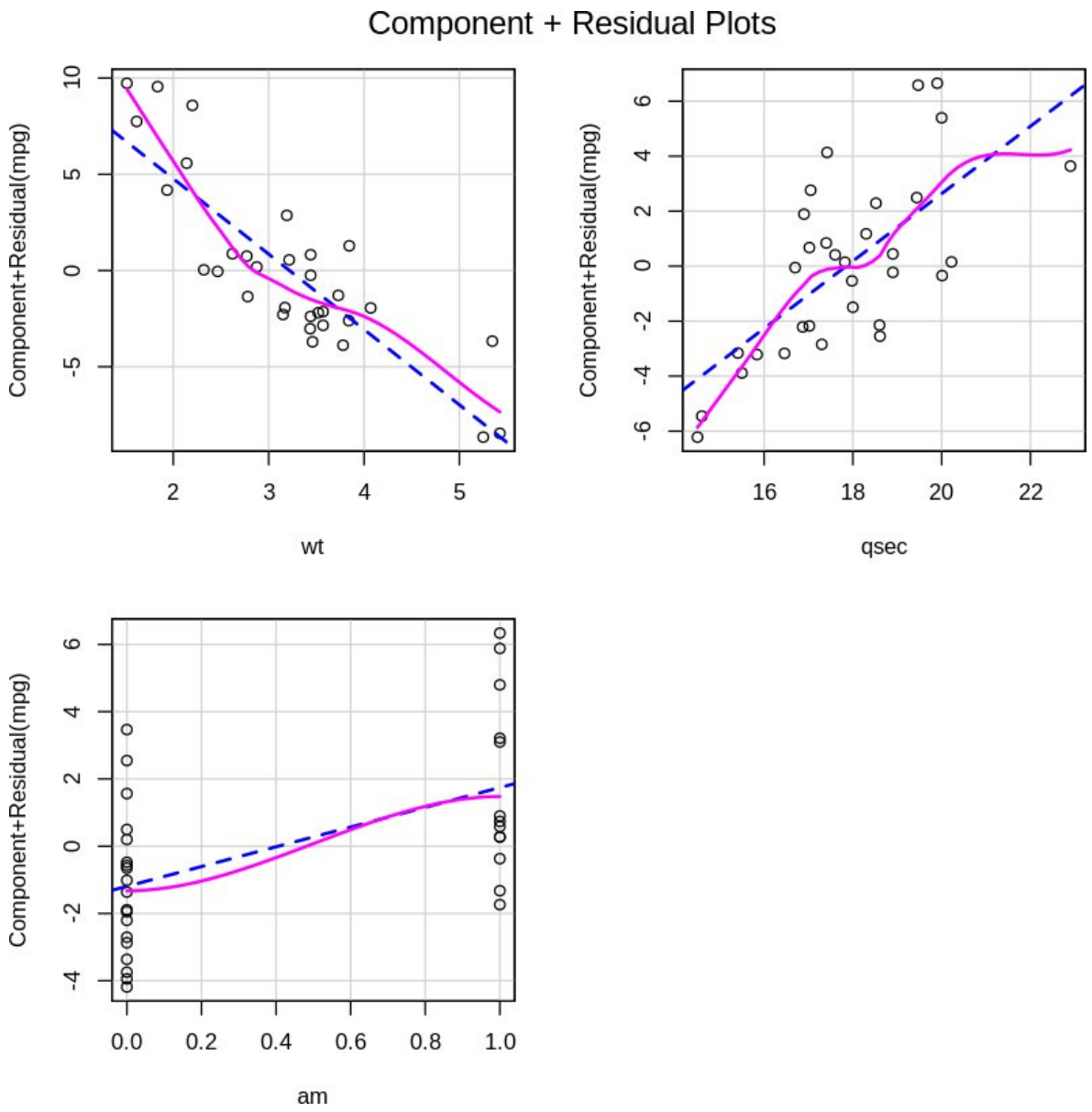
Συγκρίνωντας τώρα τα μοντέλα που έχω βρεί μεταξύ τους, το μοντέλο της Backward Elimination έχει υψηλότερο R-squared, Adjusted R-squared και χαμηλότερο AIC και Cp-Mallows. Συνδιάζει δηλαδή μεγαλύτερη αμεροληψία και R-squared με μικρότερο AIC και συνεπώς μικρότερο SSE από το μοντέλο της διαδικασίας Forward Selection και επομένως είναι αυτό που θα προτιμήσω ανάμεσα σε αυτά τα 2.

Έχωντας πλέον επιλέξει ένα μοντέλο απλούστερο και καλύτερο από το αρχικό, προχωράμε στο να ελέγξουμε αν χρειάζεται αυτό το μοντέλο κάποια βελτιστοποίηση. Αρχικά ελέγχουμε τα διαγράμματα πρόσθετων μεταβλητών για να επιβεβαιώσουμε πως δεν χρείαζεται κάποια περαιτέρω αφαίρεση μεταβλητής. Από τα διαγράμματα, βλέπουμε πως ενώ υπάρχει γραμμικη συσχέτιση μεταξύ κάθε ανεξάρτητης μεταβλητής με την εξαρτημένη (οταν θεωρήσουμε τις υπόλοιπες σταθερές) η σχέση αυτή δεν φαίνεται να είναι τόσο ισχυρή είδικα για την am.



Κοιτώντας και τα Partial Residual Plots για να ελέγξω την επιρροή κάθε επεξηγηματικής μεταβλητής ξεχωριστά στο μοντέλο βλέπω πως για τις μεταβλητές wt και qsec, ξεφεύγω αρκετά απο τις ευθείες και επομένως θα δοκιμάσω διάφορους συνδιασμούς μετασχηματισμών αυτών των 2 για να βελτιώσω το μοντέλο.

```
crPlots(mod_bwd)
```



Μετά από αρκετές δοκιμές μετασχηματισμών, κυρίως τετραγωνίζοντας και λογαριθμόντας, κατέληξα στο μοντέλο

Mpg ~ wt+wt^2+qsec+qsec^2+am

```
wtsq<-wt^2
qsecsq<-qsec^2
mod_transformed<-lm(formula=mpg~wt+wtsq++qsec+qsecsq+am)
summary(mod_2)
ols_mallows_cp(mod_transformed,mod1)
AIC(mod_transformed)
```

```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6436 -1.4743 -0.2155  0.9284  4.7051 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)   6.98176   33.66443   0.207  0.83732   
wt          -10.51153    2.94524  -3.569  0.00142 **
wtsq          0.86380    0.37434   2.308  0.02924 * 
qsec          3.08783    3.31910   0.930  0.36076   
qsecsq       -0.05735    0.08959  -0.640  0.52767   
am            1.27833    1.50495   0.849  0.40340   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.226 on 26 degrees of freedom
Multiple R-squared:  0.8856,   Adjusted R-squared:  0.8636 
F-statistic: 40.25 on 5 and 26 DF,  p-value: 1.966e-11
-1.65785907306279
149.379530680098
```

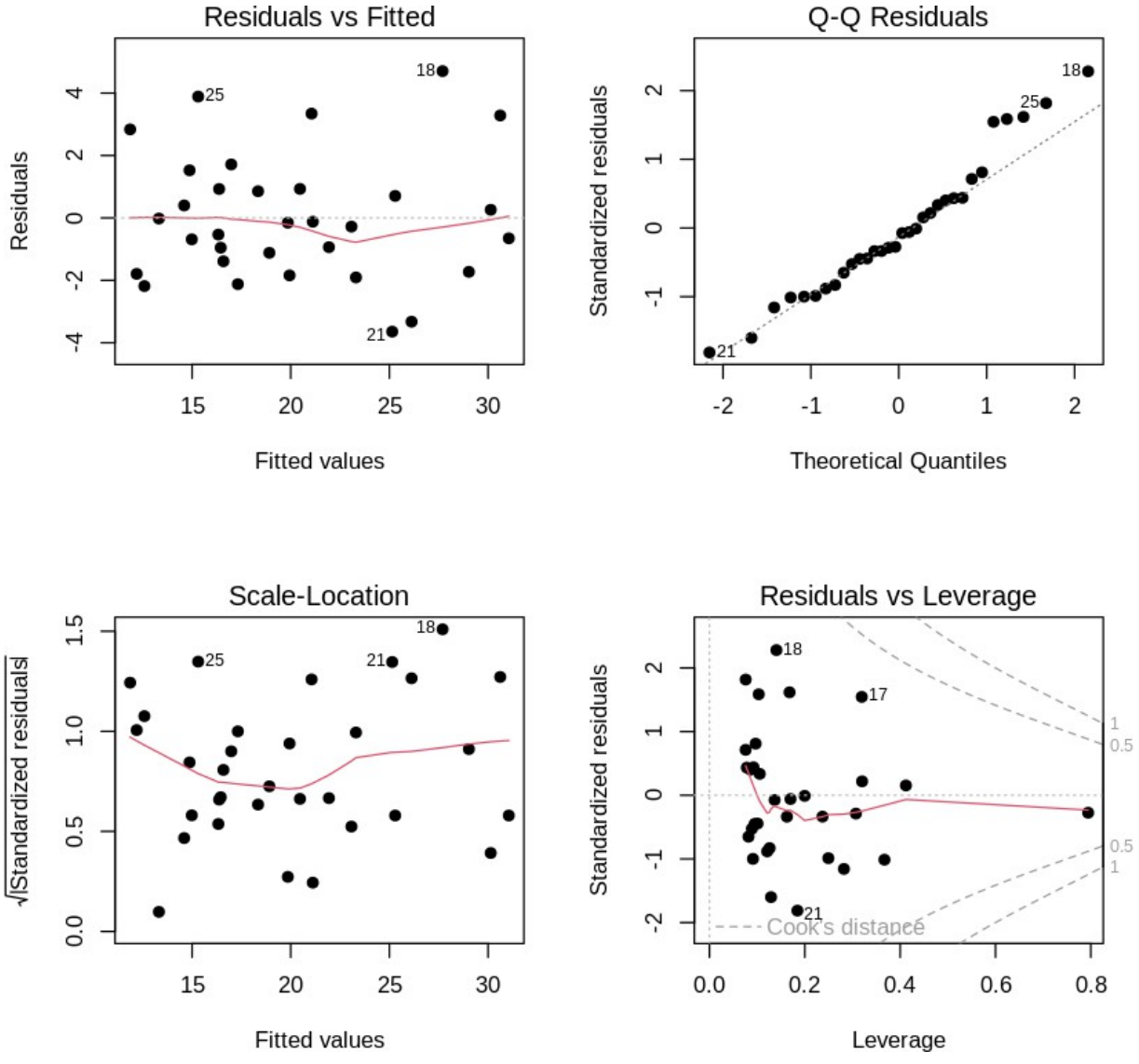
Το νέο μου μοντέλο έχει υψηλότερο R-squared, Adjusted R-squared και μικρότερο AIC και πολύ καλό Cp-Mallows

Κρατάω αυτό λοιπόν σαν τελικό μοντέλο και προχωράω στους τελευταίους μου ελέγχους.

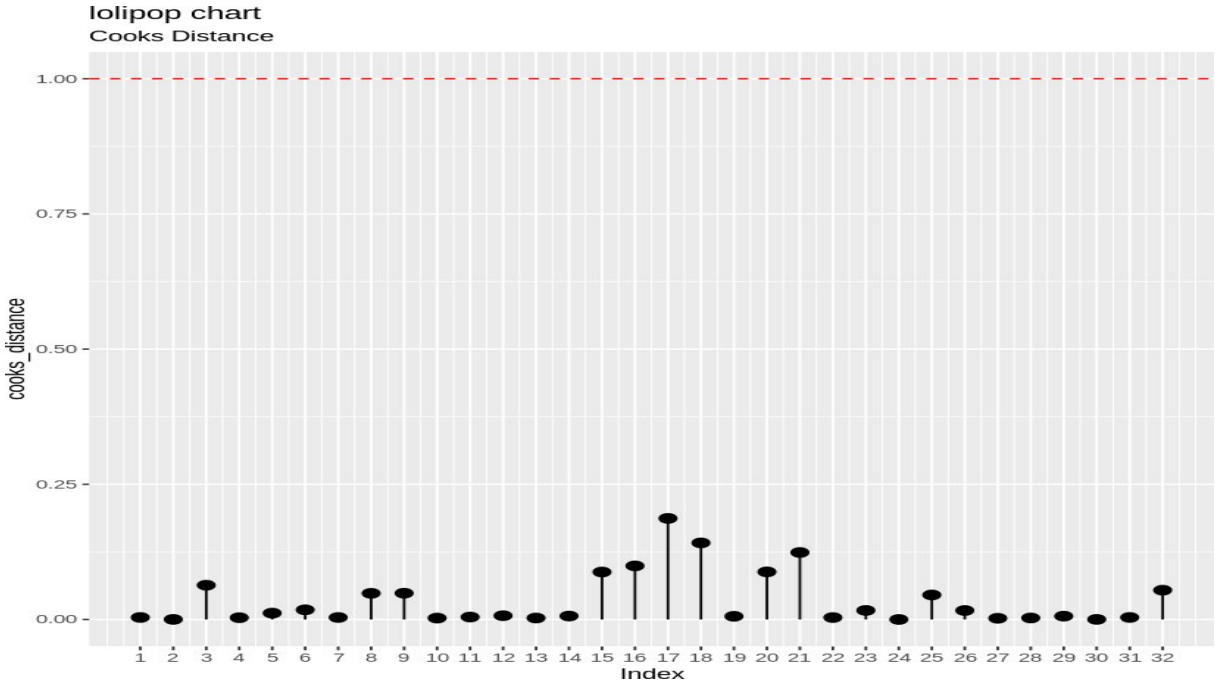
Ξεκινάω ελέγχωντας πως ικανοποιούνται οι συνθήκες του μοντέλου.

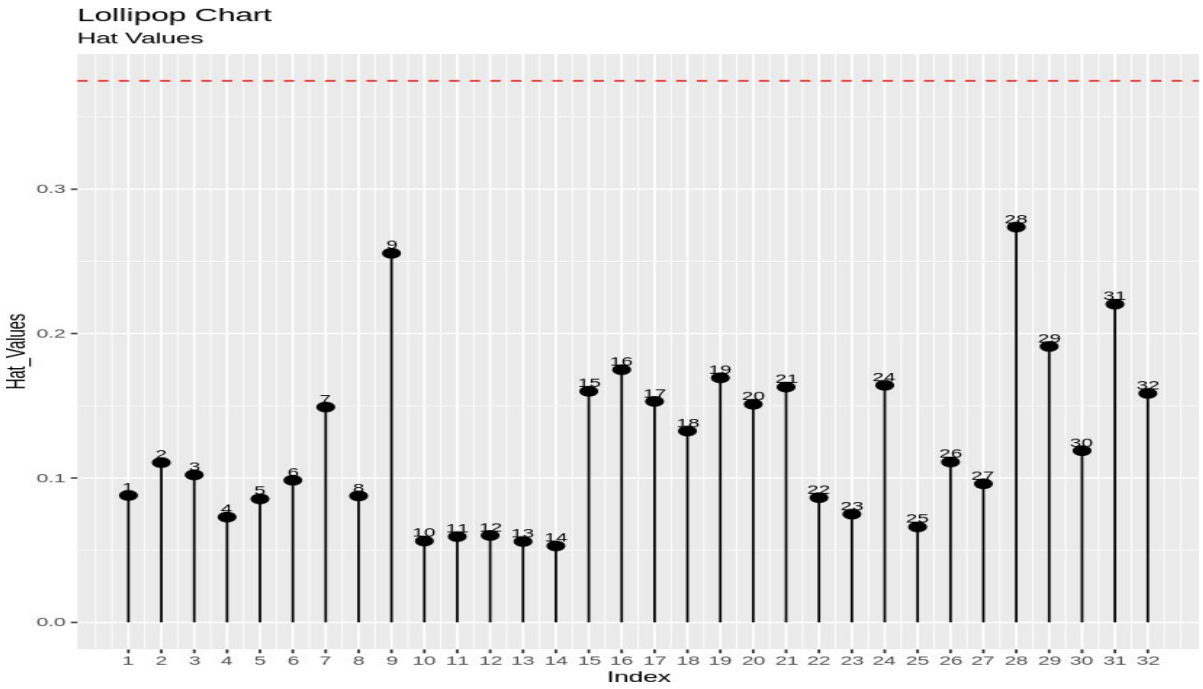
Από το QQ-plot η συνθήκη κανονικής κατανομής των υπολοίπων ικανοποιείται. Το Residuals vs Fitted και το Scale-Location δεν εμφανίζουν κάποιο μοτίβο επομένως έχουμε και τη συνθήκη της ομοσκεδαστικότητας. Τέλος το Residuals vs Leverage δείχνει χαμηλά επίπεδα μόχλευσης και δεν εμφανίζει κάποιο πιθανό σημείο επιρροής. Μέχρι στιγμής όλα είναι θετικά.

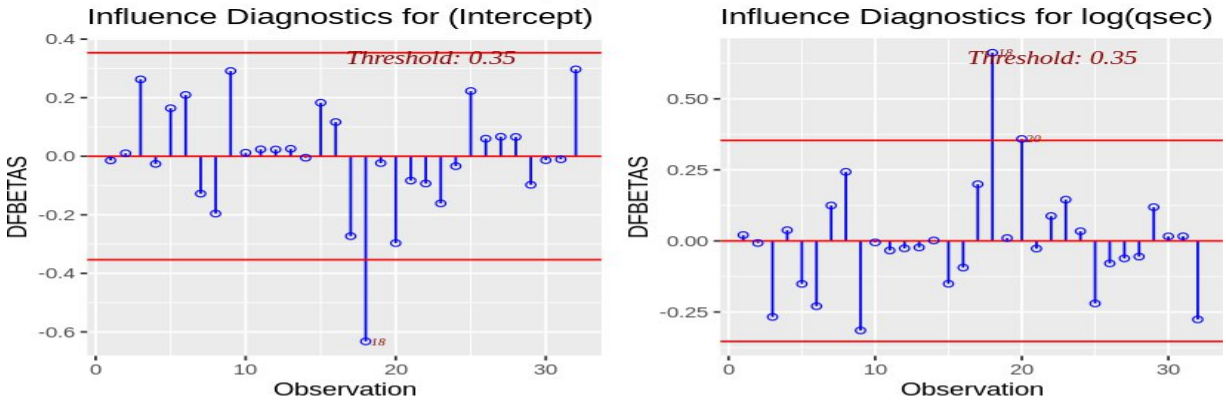
```
Par(mfrow = c(2,2) )
plot(final_model, pch=19)
```



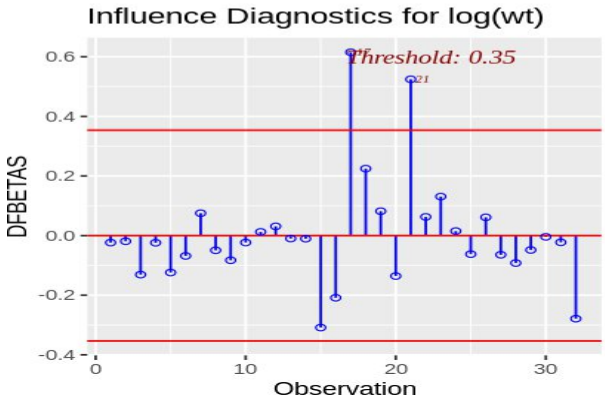
Αναφορικά με τα σημεία επιρροής προχωράω σε εντελώς παρόμοιο έλεγχο με αυτόν στο ερώτημα (1) και καταλήγω με τα παρακάτω διαγράμματα

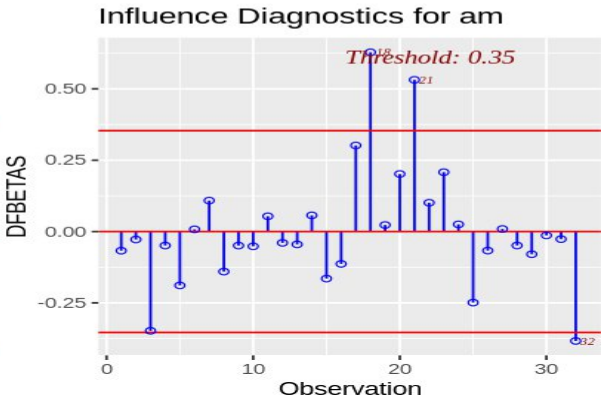


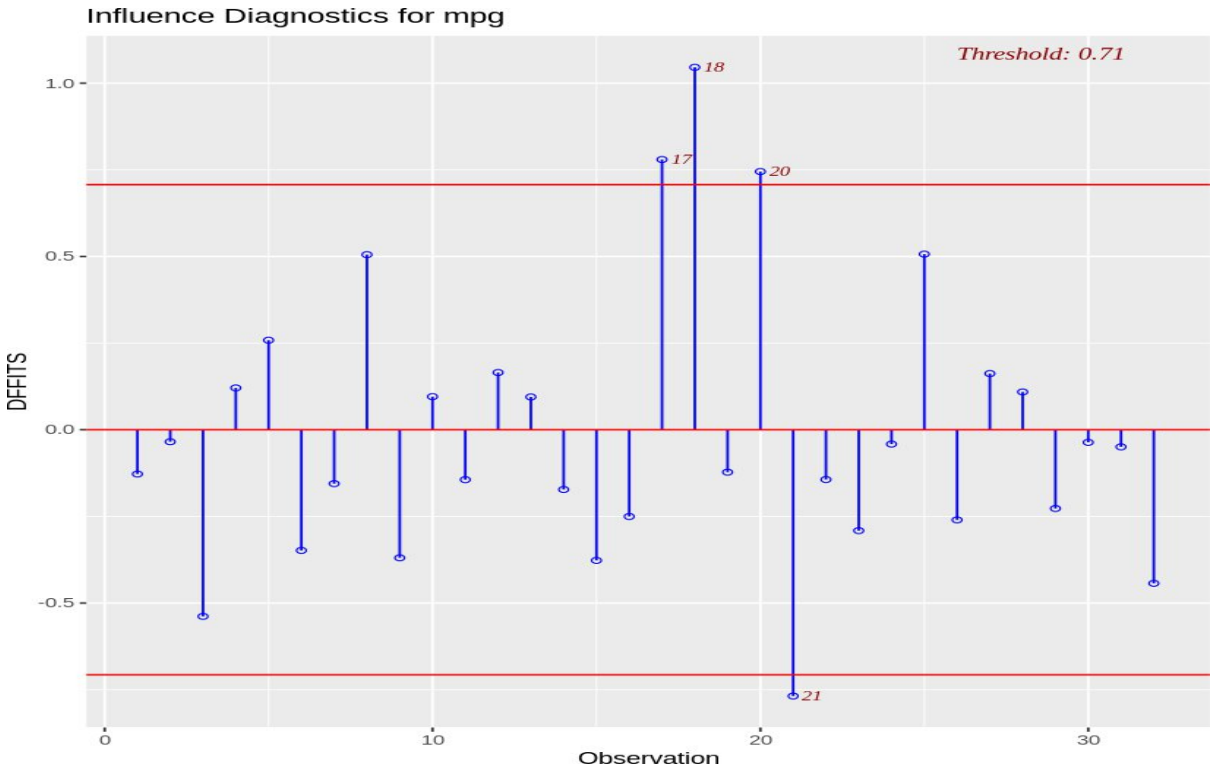



Influence Diagnostics for (Intercept)
Threshold: 0.35
DFBETAS
Observation

Influence Diagnostics for log(qsec)
Threshold: 0.35
DFBETAS
Observation


Influence Diagnostics for log(wt)
Threshold: 0.35
DFBETAS
Observation


Influence Diagnostics for am
Threshold: 0.35
DFBETAS
Observation


Influence Diagnostics for mpg
Threshold: 0.71
DFFITS
Observation

Από τα Cooks Distance και hii δεν βλέπουμε πιθανά σημεία επιρροής. Να σημειωθεί πως για τα hatvalues το νεο threshold προέκυψε με βάση φυσικά το τελικό μοντέλο δηλαδή n=32 και p=6.

Από τα αποτελέσματα των DFFITSi και DFBETAS βλέπουμε πως τα πιο πιθανά σημεία επιρροής είναι τα 18 και 21.