1 Δεδομενα Αυτοκινητων

Τα δεδομένα της παρούσας άσκησης προέκυψαν από μια μελέτη 11 χαρακτηριστικών για 32 τύπους αυτοκινήτων (μεταβλητή **car**). Τα χαρακτηριστικά αυτά, μαζί με τα σύμβολά τους, συνοψίζονται ως εξής:

mpg: Κατανάλωση βενζίνης Miles/(US) gallon
cyl: Αριθμός κυλίνδρων
disp: Μετατόπιση (cu.in.)
hp: Μικτή ιπποδύναμη
drat: Αναλογία οπίσθιου άξονα
wt: Βάρος (1000 lbs)
qsec: 1/4 mile time
vs: Διάταξη κινητήρα (0 = V, 1 = straight)
am: Κιβώτιο ταχυτήτων (0 = automatic, 1 = manual)
gear: Αριθμός προς τα εμπρός ταχυτήτων
carb: Αριθμός καρμπυρατέρ

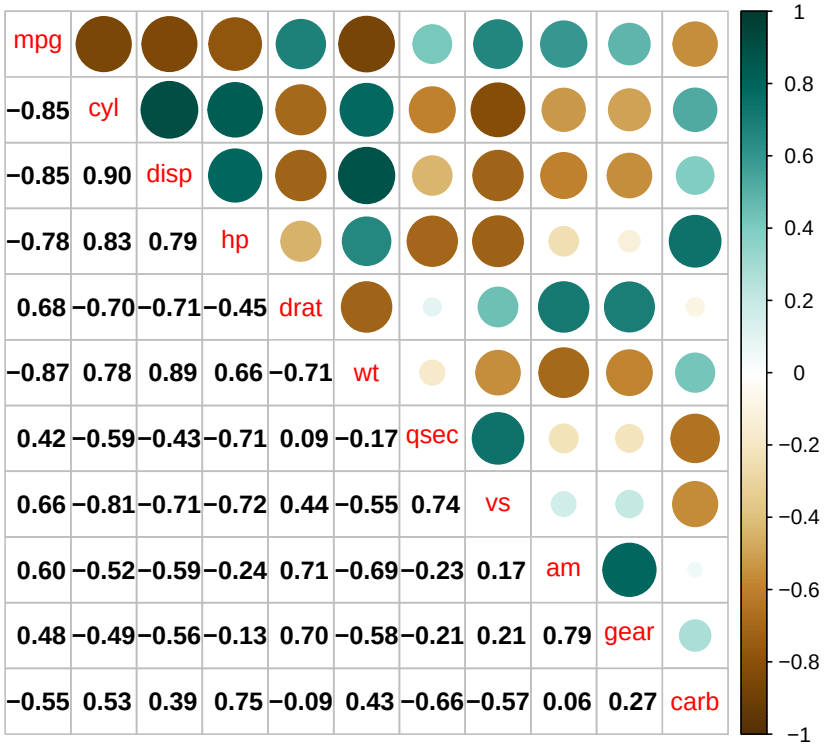
1.1 Αρχικά, τα δεδομένα φορτώνονται στο R Studio. Στη συνέχεια προσαρμόζεται σε αυτά ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης (mod1), όπου η εξαρτημένη μεταβλητή είναι η **mpg** και οι ανεξάρτητες μεταβλητές είναι τα υπόλοιπα 10 χαρακτηριστικά. Ο κώδικας σε R μαζί με τα αποτελέσματά του φαίνεται παρακάτω.

```
> vehicles <- fread('vehicles.txt')
> mod1 <- lm(mpg ~ cyl + disp + hp + drat + wt +
qsec + vs + am + gear + carb, data=vehicles)
> summary(mod1)

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
```

Παρότι η προσαρμογή στα δεδομένα γίνεται με μια σχετικά καλή τιμή για το συντελεστή προσδιορισμού, για τον οποίο ισχύει $R^2 = 0.869$, οι προκύπτουσες p-values για πολλές από τις ανεξάρτητες μεταβλητές είναι αρκετά υψηλές (με τη μεταβλητή **cyl**, συγκεκριμένα, να αντιστοιχεί σε p-value ίσο με 0.9161). Γίνεται, επομένως, αντιληπτό πως το μοντέλο πρέπει να βελτιωθεί και άρα είναι απαραίτητη μια προκαταρκτική διερεύνηση των διάφορων γνωρισμάτων του. Το πρώτο βήμα προς αυτήν την κατεύθυνση είναι ένα διάγραμμα συσχέτισης των 11 χαρακτηριστικών, μέσω της βιβλιοθήκης `corrplot` της R και των εντολών:

```
> library(corrplot)
> feats <- vehicles[,2:12]
> corrplot.mixed(cor(feats), upper.col = COL2('BrBG'),
lower.col = 'Black')
```

Το διάγραμμα απεικονίζεται στην Εικόνα 1.1.



Εικόνα 1.1: Διάγραμμα συσχέτισης των χαρακτηριστικών που αφορούν τα αμάξια της μελέτης.

Όπως γίνεται εμφανές, το χαρακτηριστικό **mpg** (1η γραμμή και 1η στήλη του διαγράμματος συσχέτισης) εμφανίζει έντονη (αρνητική) συσχέτιση με τα χαρακτηριστικά **cyl** (-85%), **disp** (-85%) και **wt** (-87%).

Αξίζει να σημειωθεί πως υψηλή συσχέτιση εμφανίζεται και μεταξύ άλλων ζευγών χαρακτηριστικών, με βασικά παραδείγματα αυτά των **cyl**-**disp** (90%), **cyl**-**hp** (83%) και **disp**-**wt** (89%), κάτι το οποίο ίσως να υποδεικνύει την ύπαρξη πολυσυγγραμμικότητας ανάμεσα στα υπό μελέτη χαρακτηριστικά. Φυσικά, η πολυσυγγραμμικότητα δεν έχει σημαντικές συνέπειες στην αξία του μοντέλου, παρ' όλα αυτά μπορεί να αποτελέσει εμπόδιο στην ερμηνεία του ρόλου κάθε επί μέρους χαρακτηριστικού, λόγω επικαλύψεων. Για παράδειγμα, το χαρακτηριστικό **mpg** εμφανίζει την ίδια συσχέτιση με τα **cyl** και **disp**, τα οποία με τη σειρά τους είναι επίσης υψηλά συσχετισμένα. Ένα πιθανό ενδεχόμενο είναι να υπάρχει κοινή πληροφορία στα **cyl** και **disp** και ως

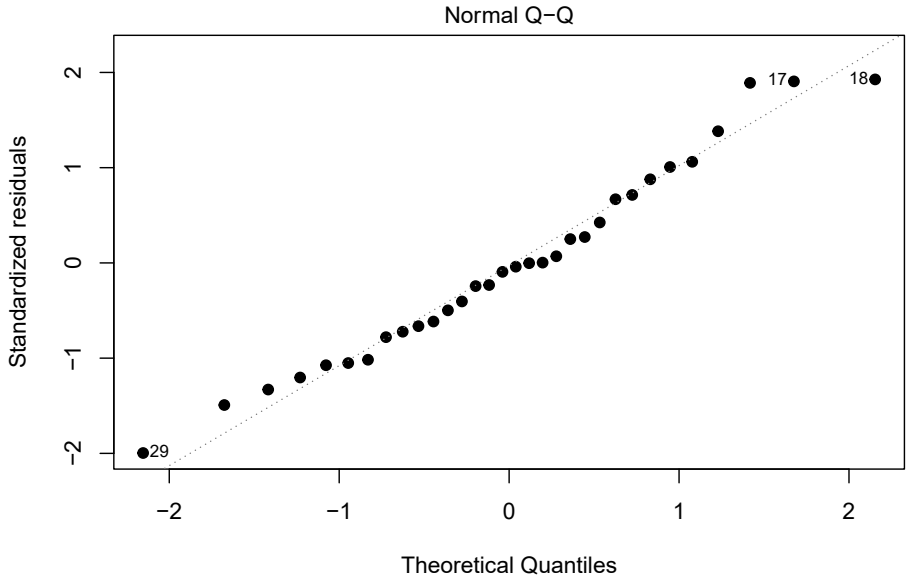
εκ τούτου το ένα εξ αυτών να είναι πρακτικά περιττό για το μοντέλο. Ένας τρόπος να επαληθευτεί η ύπαρξη πολυσυγγραμμικότητας στο μοντέλο είναι μέσω του παράγοντα μεγέθυνσης διασποράς (VIF) [1]. Η εντολή της R `vif(mod1)` υπολογίζει τον παράγοντα μεγέθυνσης για κάθε χαρακτηριστικό του μοντέλου, με τα αποτελέσματα να φαίνονται στον επόμενο πίνακα (Πίνακας 1.1).

Χαρακτηριστικό	**cyl**	**disp**	**hp**	**drat**	**wt**	**qsec**	**vs**	**am**	**gear**	**carb**
VIF	15.37	21.62	9.83	3.37	15.16	7.53	4.97	4.65	5.36	7.91

Πίνακας 1.1: Παράγοντες μεγέθυνσης διασποράς (VIF) για κάθε επεξηγηματική μεταβλητή του μοντέλου.

Τα αποτελέσματα του Πίνακα 1.1 έρχονται σε πλήρη συμφωνία με τις αρχικές παρατηρήσεις που έγιναν βάσει του διαγράμματος συσχέτισης της Εικόνας 1.1, αφού οι υψηλότερες τιμές για τον παράγοντα μεγέθυνσης διασποράς προκύπτουν για τις μεταβλητές **cyl**, **disp** και **wt**. Φυσικά, κι άλλα χαρακτηριστικά εμφανίζουν υψηλές τιμές για τον VIF[1], αφού τιμές υψηλότερες του 5 αποτελούν εν γένει ένδειξη πολυσυγγραμμικότητας. Συμπερασματικά, οι παρατηρήσεις αυτές συνηγορούν υπέρ της ύπαρξης πολυσυγγραμμικότητας στο μοντέλο mod1.

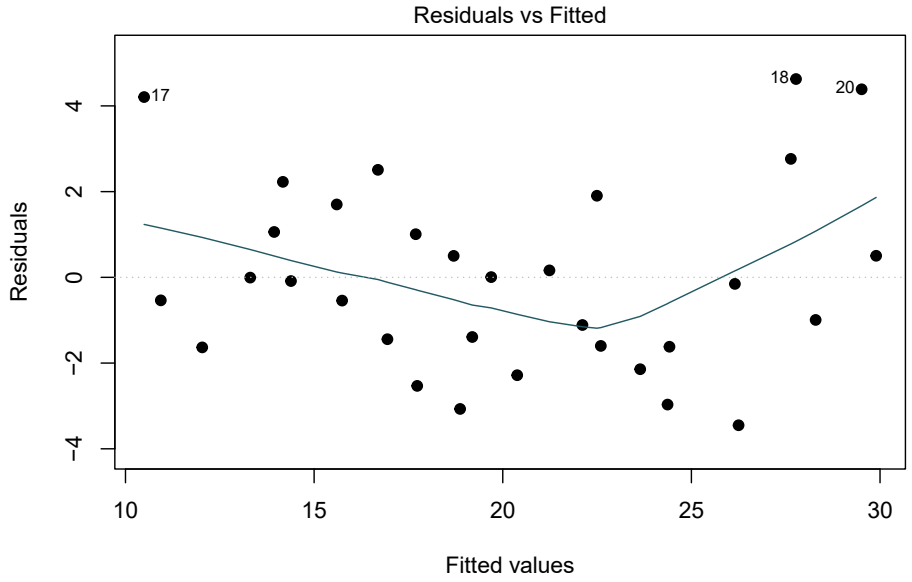
Η προκαταρκτική διερεύνηση του μοντέλου συνεχίζεται με τον έλεγχο τήρησης των προϋποθέσεων του μοντέλου σε ό,τι αφορά τα υπόλοιπα. Αρχικά, παρουσιάζεται το διάγραμμα Q-Q στην Εικόνα 1.2, το οποίο προκύπτει μέσω της εντολής `plot(mod1,which=2,pch=19)` στην R.



Εικόνα 1.2: Διάγραμμα Q-Q για έλεγχο της κανονικότητας των υπολοίπων.

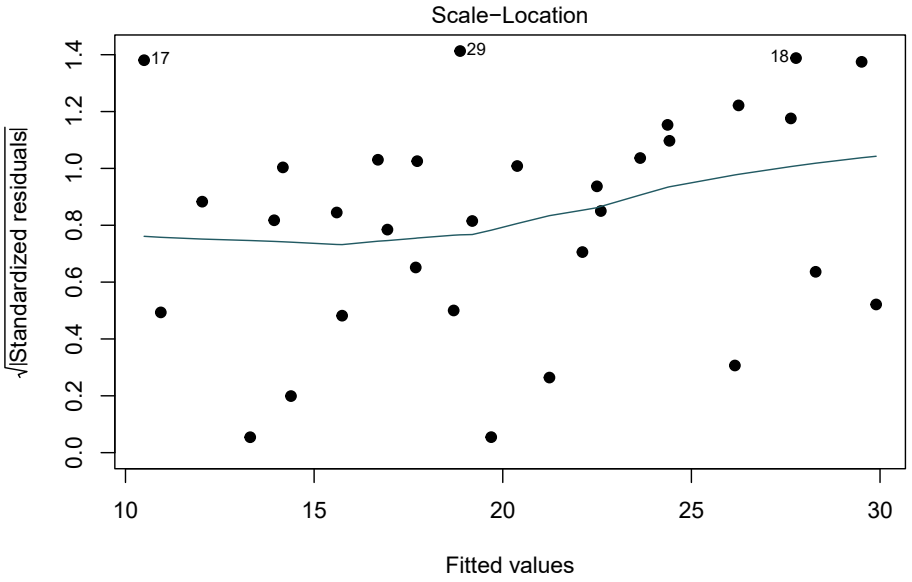
Με εξαίρεση λίγα σημεία, είναι εμφανής η ύπαρξη μιας γραμμικότητας, η οποία υποδεικνύει πως, πράγματι, η κανονικότητα των υπολοίπων επαληθεύεται. Σε ό,τι αφορά το διάγραμμα των υπολοίπων σε σχέση με τις εκτιμημένες τιμές βάσει του μοντέλου, το οποίο προκύπτει μέσω της εντολής `plot(mod1,which=2,pch=19)` στην R, αυτό απεικονίζεται στην Εικόνα 1.3.

[1] Συγκεκριμένα, τα **hp**, **qsec**, **gear**, **carb** και οριακά το **vs**.



Εικόνα 1.3: Διάγραμμα υπολοίπων συναρτήσει των εκτιμημένων τιμών.

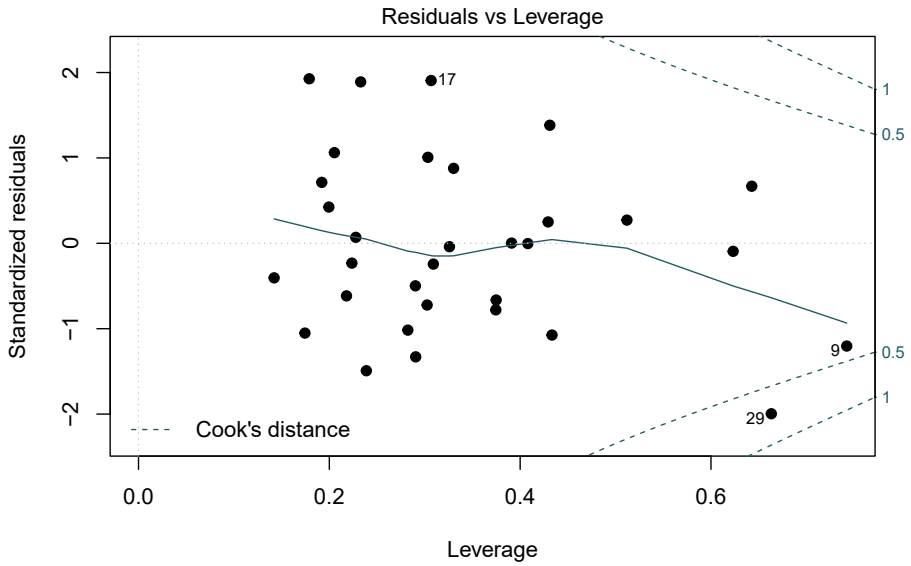
Το διάγραμμα αυτό αντιστοιχεί στον έλεγχο για την υπόθεση της ομοσκεδαστικότητας, η οποία φαίνεται να επαληθεύεται, μιας και η κατανομή των υπολοίπων μοιάζει να μην έχει προτιμητέα διεύθυνση ή εξάρτηση από τις εκτιμημένες τιμές. Προκειμένου αυτό να επαληθευτεί με μεγαλύτερη βεβαιότητα, στην Εικόνα 1.4 φαίνεται το διάγραμμα της τετραγωνικής ρίζας της απόλυτης τιμής των τυποποιημένων υπολοίπων ως συνάρτηση των εκτιμημένων τιμών, το οποίο δίνει μια καθαρότερη εικόνα ως προς το κατά πόσο τα υπόλοιπα είναι ομοιόμορφα μοιρασμένα.



Εικόνα 1.4: Διάγραμμα για περαιτέρω έλεγχο της ομοσκεδαστικότητας.

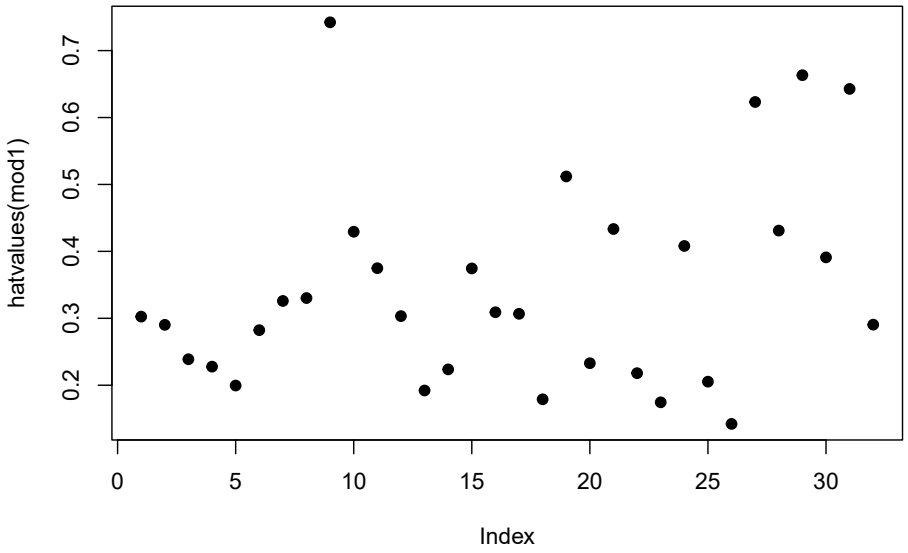
Το διάγραμμα αυτό προκύπτει στην R μέσω της εντολής `plot(mod1,which=3,pch=19)`. Πράγματι, η υπόθεση της ομοσκεδαστικότητας φαίνεται ξανά να επαληθεύεται, αφού η προκύπτουσα γραμμή στο διάγραμμα της Εικόνας 1.4 είναι κατά προσέγγιση οριζόντια. Το τελευταίο διάγραμμα που αφορά τα υπόλοιπα είναι το διάγραμμα των τυποποιημένων υπολοίπων συναρ-

τήσει της μόχλευσης, το οποίο προκύπτει μέσω της εντολής `plot(mod1,which=5,pch=19)` στην R και φαίνεται στην Εικόνα 1.5.



Εικόνα 1.5: Διάγραμμα τυποποιημένων υπολοίπων συναρτήσει της μόχλευσης και απεικόνιση των αποστάσεων Cook.

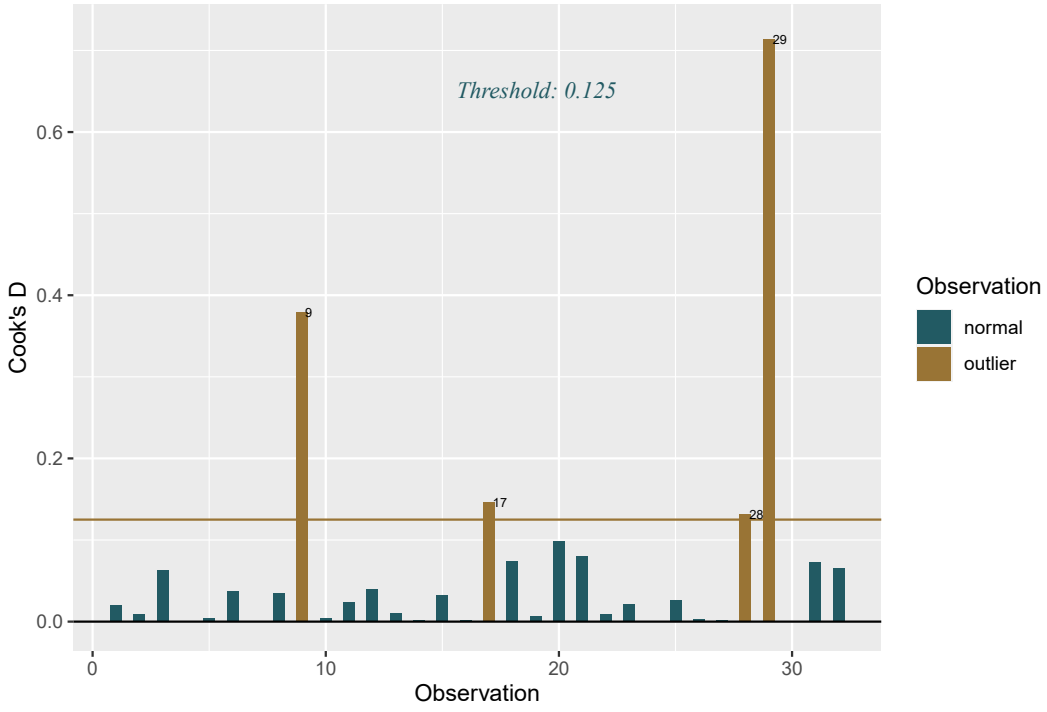
Με εξαίρεση τα σημεία με δείκτες 9 και 29, η διασπορά των τυποποιημένων υπολοίπων δε φαίνεται να μεταβάλλεται καθώς μεταβάλλεται η μόχλευση, το οποίο είναι επιθυμητό. Ειδικά για τα σημεία με δείκτες 9 και 29, αυτά φαίνεται να αποτελούν σημεία επιρροής, αφού βρίσκονται είτε ανάμεσα είτε πολύ κοντά στις διακεκομμένες γραμμές, οι οποίες αντιστοιχούν στην απόσταση Cook. Επιπλέον, σε αυτά οφείλεται η φθίνουσα τάση της γραφικής παράστασης για μόχλευση μεγαλύτερη από 0.6.



Εικόνα 1.6: Απεικόνιση των τιμών h_{ii} για καθένα εκ των 32 σημειών που αντιστοιχούν στα δεδομένα.

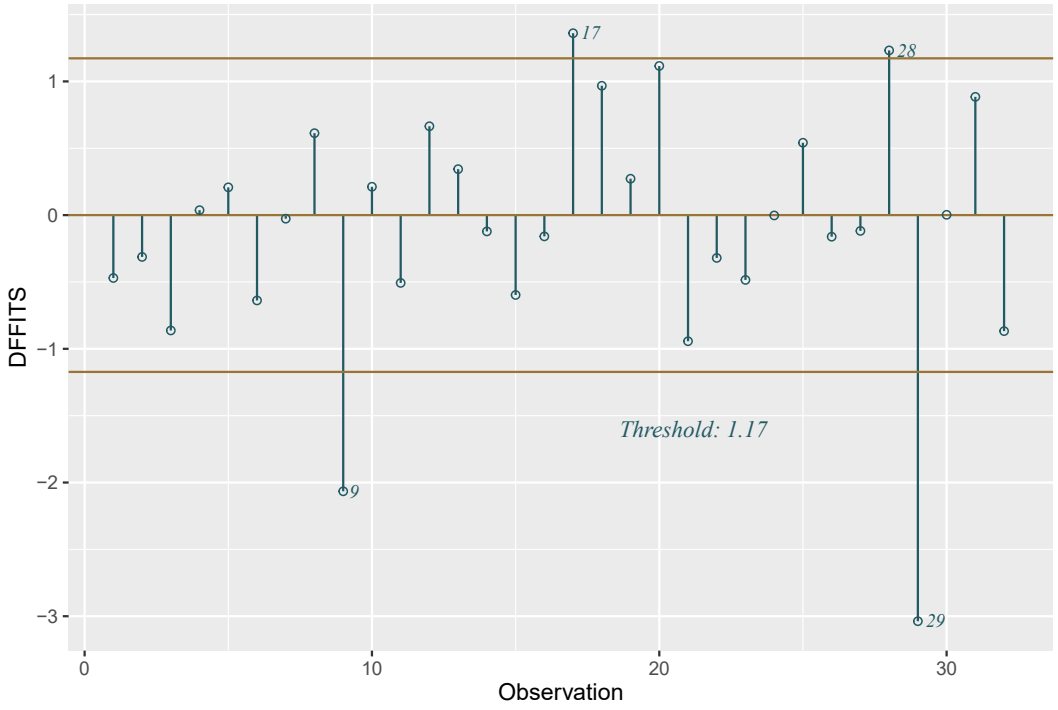
Προκειμένου η εικόνα για τα σημεία επιρροής (καθώς και για τα άτυπα σημεία) να γίνει πιο ξεκάθαρη, παρατίθεται στην Εικόνα 1.6 το διάγραμμα των διαγώνιων στοιχείων του hat matrix (h_{ii}), το οποίο προκύπτει μέσω της εντολής `plot(hatvalues(mod1), pch=19)` στην R. Ο εμπειρικός κανόνας βάσει του οποίου κρίνεται εάν το σημείο με δείκτη i αντιστοιχεί σε σημείο επιρροής είναι η συνθήκη $h_{ii} > 2p/n$, όπου $p = 11$ είναι το πλήθος χαρακτηριστικών και $n = 32$ είναι το συνολικό πλήθος παρατηρήσεων (σημείων). Στην προκείμενη περίπτωση, η συνθήκη ισοδυναμεί με $h_{ii} > 0.6875$. Βάσει του διαγράμματος της Εικόνας 1.6, το μοναδικό σημείο που ικανοποιεί τη συνθήκη αυτή είναι το υπ' αριθμόν 9, ενώ επίσης τα σημεία με δείκτες 29 και 31 είναι πολύ κοντά στο όριο, με $h_{29,29} = 0.66$ και $h_{31,31} = 0.64$. Υπενθυμίζεται εδώ πως τα σημεία με δείκτες 9 και 29 αναγνωρίστηκαν ως σημεία επιρροής και βάσει του ελέγχου του διαγράμματος της Εικόνας 1.5.

Οι διαγνωστικοί έλεγχοι με στόχο την αναγνώριση άτυπων σημείων ή σημείων επιρροής συνεχίζονται με το διάγραμμα της Εικόνας 1.7, το οποίο αντιστοιχεί στην απόσταση Cook κάθε σημείου και προκύπτει μέσω της εντολής `ols_plot_cooksd_bar(mod1)` στην R.



Εικόνα 1.7: Απεικόνιση της απόστασης Cook για καθένα εκ των 32 σημειών που αντιστοιχούν στα δεδομένα.

Όπως φαίνεται και από το διάγραμμα, τα σημεία για τα οποία η απόσταση Cook ξεπερνά το κατώφλι 0.125 είναι αυτά με δείκτες 9 και 29 και, οριακά, αυτά με δείκτες 17 και 28. Γίνεται αντιληπτό πως τα σημεία υπ' αριθμόν 9 και 29 πράγματι αποτελούν σημεία επιρροής, αφού επανεμφανίζονται ως υποψήφια σε όλους τους ελέγχους που έχουν πραγματοποιηθεί ως αυτό το σημείο. Παρ' όλα αυτά, αξίζει να πραγματοποιηθούν δύο επιπλέον έλεγχοι, με πρώτο αυτόν του μέτρου DFFITS. Το μέτρο αυτό ελέγχει την επιρροή κάθε σημείου στην πρόβλεψη της μεταβλητής απόκρισης, ελέγχοντας τη μεταβολή που παρατηρείται σε αυτήν εάν το σημείο παραληφθεί. Το σχετικό διάγραμμα προκύπτει μέσω της εντολής `ols_plot_dffits(mod1)` στην R και απεικονίζεται στην Εικόνα 1.8.



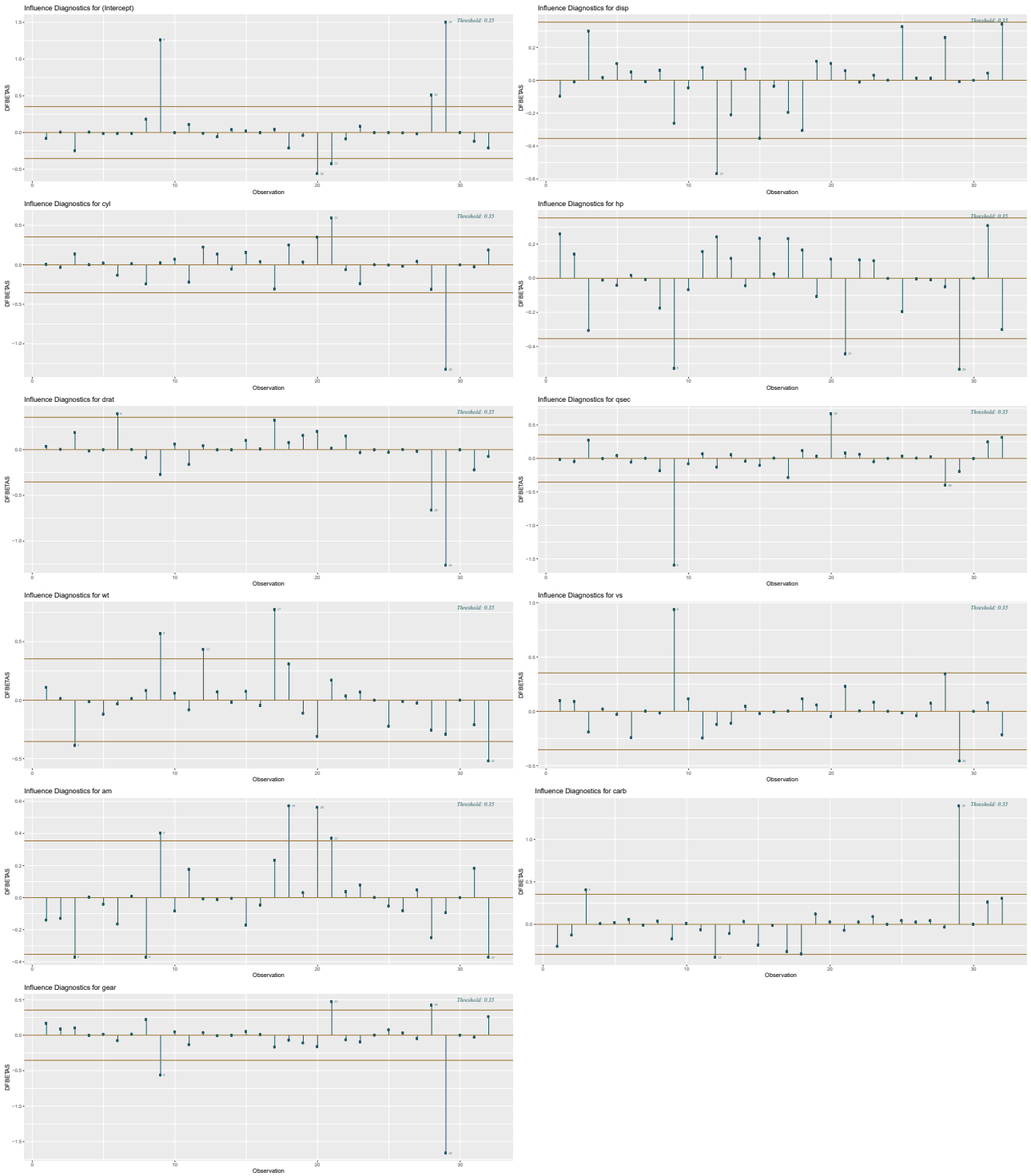
Εικόνα 1.8: Απεικόνιση του μέτρου DFFITS για καθένα εκ των 32 σημειών που αντιστοιχούν στα δεδομένα.

Το κατώφλι στην περίπτωση αυτή οριοθετούν οι τιμές $\pm 2\sqrt{p/n}$, δηλαδή ± 1.17 στην προκείμενη περίπτωση, όπως φαίνεται και στο διάγραμμα. Σε πλήρη συμφωνία με τα αποτελέσματα που προέκυψαν από το διάγραμμα της απόστασης Cook, τα υποψήφια σημεία επιρροής είναι τα υπ' αριθμόν 9, 29 και, οριακά, τα 17 και 28.

Οι έλεγχοι για την παρουσία σημείων επιρροής ολοκληρώνονται με τον έλεγχο του μέτρου DFBETAS. Το μέτρο αυτό είναι όμοιο του DFFITS, με βασική διαφορά το γεγονός πως υπολογίζεται για καθένα εκ των 11 χαρακτηριστικών του μοντέλου (και όχι συνολικά) βάσει της επιρροής κάθε παρατήρησης σε αυτά, η οποία υπολογίζεται μέσω παράληψης της εκάστοτε παρατήρησης. Στην περίπτωση αυτή, το κατώφλι ισοδυναμεί με τις τιμές $\pm 2/\sqrt{n}$, δηλαδή ± 0.35 για το υπό μελέτη μοντέλο[2]. Το διάγραμμα εξάγεται μέσω της εντολής `ols_plot_dfbetas(mod1)` στην R και απεικονίζεται στην Εικόνα 1.9. Ξανά, τα σημεία που στις περισσότερες περιπτώσεις είναι υποψήφια σημεία επιρροής είναι τα υπ' αριθμόν 9 και 29.

Με το πέρας αυτής της προκαταρκτικής διερεύνησης των χαρακτηριστικών του μοντέλου, γεννάται η ανάγκη για τη διερεύνηση του κατά πόσο το μοντέλο αυτό είναι το βέλτιστο για να περιγράψει και να προβλέψει την κατανάλωση βενζίνης **mpg** αμαξιών, ή εάν υπάρχουν καλύτερες επιλογές στις οποίες αξιοποιούνται λιγότερα χαρακτηριστικά ή συναρτήσεις αυτών.

[2] Είναι αναμενόμενο το κατώφλι στην περίπτωση αυτή να είναι ανεξάρτητο του πλήθους των χαρακτηριστικών, p, αφού το μέτρο DFBETAS αφορά κάθε χαρακτηριστικό ξεχωριστά.



Εικόνα 1.9: Απεικόνιση του μέτρου DFBETAS για καθένα εκ των 32 σημειών που αντιστοιχούν στα δεδομένα και για κάθε χαρακτηριστικό του μοντέλου.

1.2 Η βελτιστοποίηση του μοντέλου πραγματοποιείται σε δύο κατευθύνσεις: η πρώτη είναι η μείωση των χαρακτηριστικών που λαμβάνει υπ' όψιν, ενώ η δεύτερη είναι η συναρτησιακή σχέση με την οποία τα λαμβάνει. Σε ό,τι αφορά την πρώτη κατεύθυνση, χρησιμοποιείται το κριτήριο πληροφορίας Akaike (AIC), το οποίο ορίζεται στη γενική περίπτωση ως

$$\mathrm{AIC} = 2p - 2\ln \mathcal{L}, \tag{1.1}$$

με p το πλήθος παραμέτρων και $\mathcal{L}$ τη μέγιστη πιθανοφάνεια, και στοχεύει στην εύρεση του βέλτιστου μοντέλου με όσο το δυνατό μικρότερο αριθμό παραμέτρων (χαρακτηριστικών). Η ελαχιστοποίηση των παραμέτρων και παράλληλα βελτιστοποίηση του μοντέλου μέσω του AIC πραγματοποιείται με τους εξής τρεις διαφορετικούς τρόπους.

a. Μέσω της *διαδικασίας διαδοχικής αφαίρεσης*, όπου, ξεκινώντας από το πλήρες μοντέλο πολλαπλής παλινδρόμησης (10 επεξηγηματικές μεταβλητές), σε κάθε βήμα αφαιρείται η μεταβλητή η αφαίρεση της οποίας οδηγεί στην ελάχιστη τιμή του AIC. Η σύγκλιση της διαδικασίας επέρχεται όταν πλέον το AIC ελαχιστοποιείται και η αφαίρεση πρόσθετων επεξηγηματικών μεταβλητών οδηγεί σε αύξησή του. Η σχετική εντολή στην R και το αντίστοιχο αποτέλεσμα φαίνονται παρακάτω.

```
> mod1a <- step(mod1, direction = 'backward', test = 'F')

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

> summary(mod1a)

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
```

Προκύπτει πως το μοντέλο στο οποίο συγκλίνει η διαδικασία είναι το

$$\mathbf{mpg} = 9.618 - 3.917\mathbf{wt} + 1.226\mathbf{qsec} + 2.936\mathbf{am}, \tag{1.2}$$

δηλαδή το μοντέλο που λαμβάνει υπ' όψιν ως χαρακτηριστικά μόνο το βάρος, το 1/4 mile time και το κιβώτιο ταχυτήτων. Για το μοντέλο αυτό προκύπτει $R^2 = 0.8497$.

b. Μέσω της *διαδικασίας διαδοχικής πρόσθεσης*, όπου, ξεκινώντας από το τετριμμένο μοντέλο, σε κάθε βήμα προστίθεται η μεταβλητή η προσθήκη της οποίας οδηγεί στην ελάχιστη τιμή του

AIC. Η σύγκλιση της διαδικασίας επέρχεται όταν πλέον το AIC ελαχιστοποιείται και η προσθήκη επιπλέον επεξηγηματικών μεταβλητών οδηγεί σε αύξησή του. Η σχετική εντολή στην R και το αντίστοιχο αποτέλεσμα φαίνονται παρακάτω.

```
> mod1b <- step(lm(vehicles$mpg ~ 1), direction = 'forward',
test = 'F', scope =(~ vehicles$cyl + vehicles$disp +
vehicles$hp + vehicles$drat + vehicles$wt +vehicles$qsec +
vehicles$vs + vehicles$am + vehicles$gear + vehicles$carb))
> summary(mod1b)

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 ***
vehicles$wt  -3.16697    0.74058  -4.276 0.000199 ***
vehicles$cyl -0.94162    0.55092  -1.709 0.098480 .
vehicles$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
```

Προκύπτει πως το μοντέλο στο οποίο συγκλίνει η διαδικασία είναι το

$$\mathbf{mpg} = 38.752 - 3.167\mathbf{wt} - 0.942\mathbf{cyl} - 0.018\mathbf{hp}, \tag{1.3}$$

δηλαδή το μοντέλο που λαμβάνει υπ’ όψιν ως χαρακτηριστικά μόνο το βάρος, τον αριθμό κυλίνδρων και τη μικτή ιπποδύναμη. Για το μοντέλο αυτό προκύπτει $R^2 = 0.8431$.

c. Μέσω της *κατά βήματα εμπρός-πίσω επιλογής*, όπου, ξεκινώντας από το πλήρες μοντέλο πολλαπλής παλινδρόμησης (10 επεξηγηματικές μεταβλητές), σε κάθε βήμα είτε αφαιρείται είτε προστίθεται μια μεταβλητή, η προσθήκη/αφαίρεση της οποίας οδηγεί στην ελάχιστη τιμή του AIC, συνδυάζοντας κατά μία έννοια τις δύο προαναφερθείσες μεθόδους. Η σχετική εντολή στην R και το αντίστοιχο αποτέλεσμα φαίνονται παρακάτω.

```
> mod1c <- step(mod1, direction = 'both', test = 'F')
> summary(mod1c)

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 ***
```

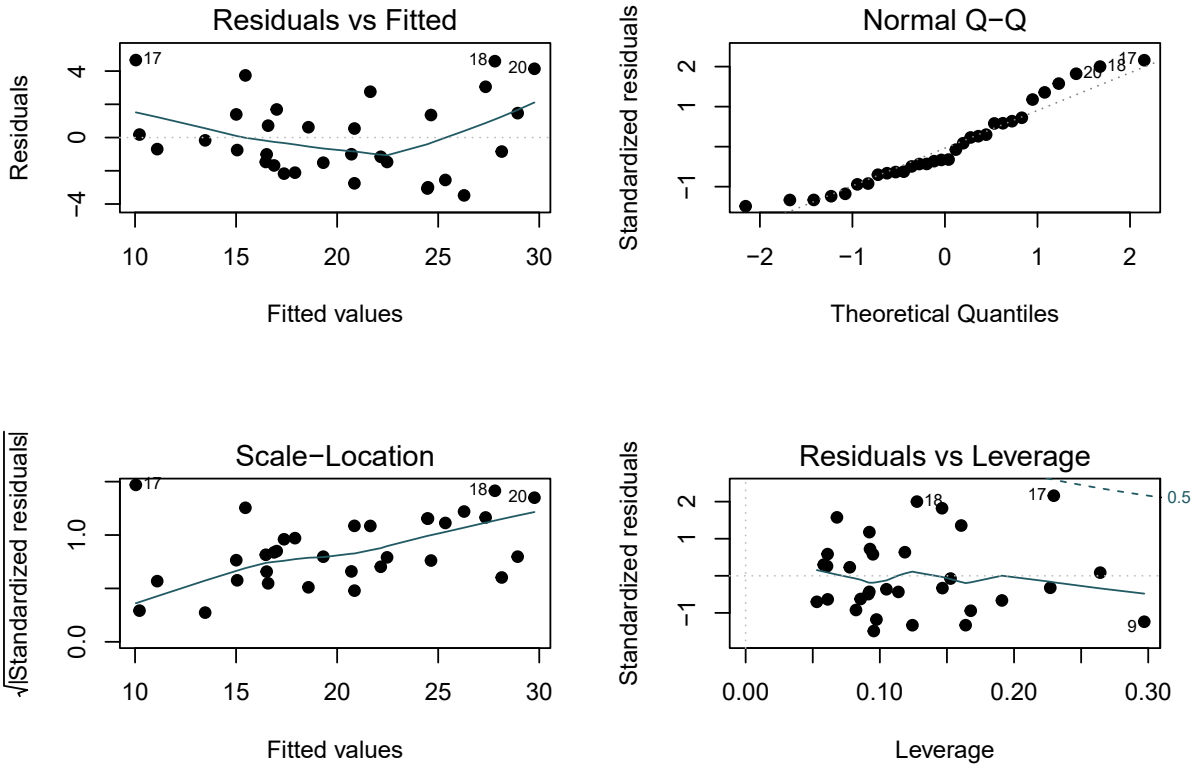
```
am            2.9358     1.4109   2.081 0.046716 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

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
```

Προκύπτει πως το μοντέλο στο οποίο συγκλίνει η διαδικασία είναι το

$$\mathbf{mpg} = 9.618 - 3.917\mathbf{wt} + 1.226\mathbf{qsec} + 2.936\mathbf{am}, \tag{1.4}$$

δηλαδή ακριβώς το μοντέλο που προκύπτει και μέσω του τρόπου **a**.

Δεδομένου πως υπάρχουν δύο υποψήφια μοντέλα, 3 χαρακτηριστικών το καθένα, και με παρόμοιες τιμές για τους δείκτες R^2, η τελική επιλογή μοντέλου θα εξαρτηθεί από τη διερεύνηση του εάν και σε τι βαθμό οι συνθήκες για τα υπόλοιπα ικανοποιούνται σε καθένα από τα δύο νέα μοντέλα, δημιουργώντας εκ νέου τα διαγράμματα των Εικόνων 1.2 έως 1.5.



Εικόνα 1.10: Σύνοψη των γραφικών ελέγχων για τα υπόλοιπα που αφορούν το μοντέλο της Σχέσης (1.2).

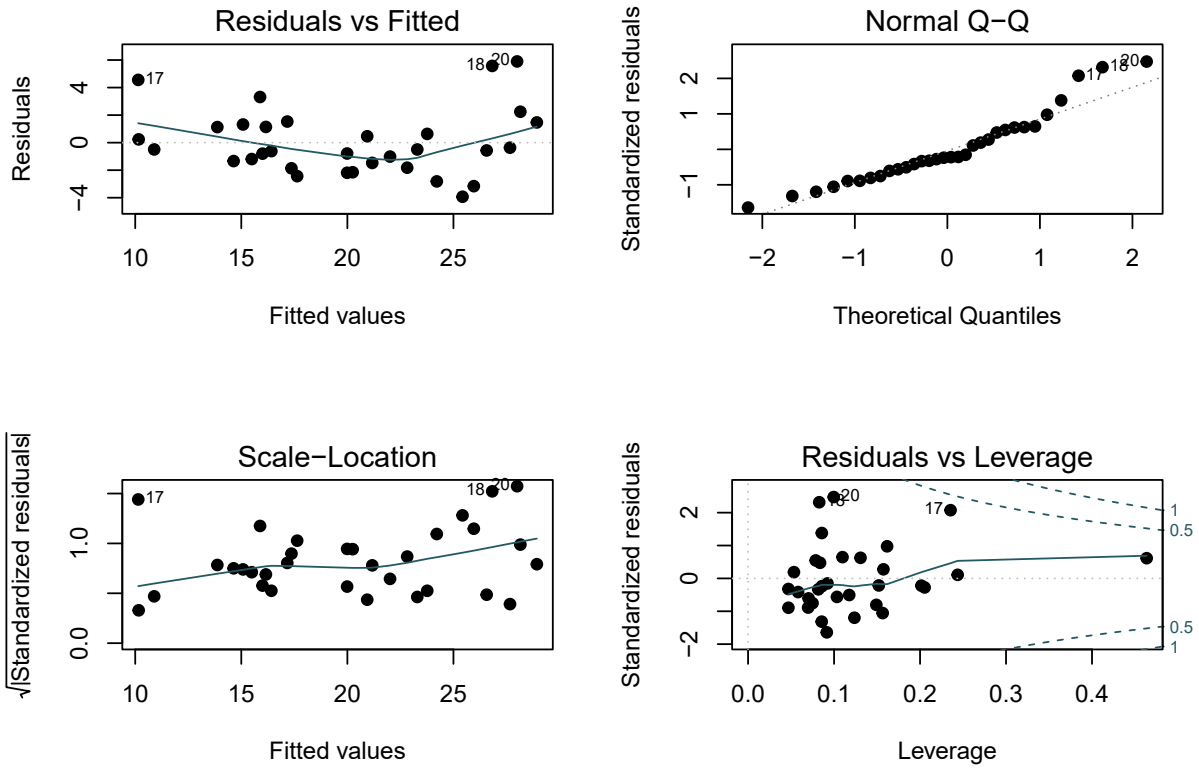
Στην Εικόνα 1.10 απεικονίζεται η σύνοψη των διαγραμμάτων αυτών για το μοντέλο mod1a, δηλαδή το μοντέλο της Σχέσης (1.2), η οποία εξήχθη μέσω των ακόλουθων εντολών της R:

```
> par(mfrow=c(2,2))
> plot(mod1a,pch=19)
```

Με εξαίρεση το διάγραμμα της ρίζας των τυποποιημένων υπολοίπων ως συνάρτηση των εκτιμημένων τιμών, στο οποίο η προκύπτουσα γραμμή απέχει σημαντικά από την ευθεία, οι προϋποθέσεις για τα υπόλοιπα φαίνεται να ικανοποιούνται στον ίδιο βαθμό που ικανοποιούνται και για το αρχικό μοντέλο, mod1. Προχωρώντας με την αντίστοιχη διαδικασία για το μοντέλο mod1b, δηλαδή το μοντέλο της Σχέσης (1.3), οι εντολές στην R είναι οι

```
> par(mfrow=c(2,2))
> plot(mod1b,pch=19)
```

και τα αντίστοιχα διαγράμματα φαίνονται στην Εικόνα 1.11.



Εικόνα 1.11: Σύνοψη των γραφικών ελέγχων για τα υπόλοιπα που αφορούν το μοντέλο της Σχέσης (1.3).

Παρότι και στην περίπτωση αυτή η εικόνα δεν είναι πολύ διαφορετική, γίνεται εμφανής μια καλύτερη προσέγγιση της οριζόντιας ευθείας στο διάγραμμα της ρίζας των τυποποιημένων υπολοίπων σε συνάρτηση με τις εκτιμημένες τιμές. Επιπλέον, το διάγραμμα των τυποποιημένων υπολοίπων συναρτήσει της μόχλευσης επίσης φαίνεται πιο ικανοποιητικό. Έτσι, το μοντέλο με τα μειωμένα χαρακτηριστικά που λαμβάνεται στο εξής υπ' όψιν, με σκοπό περαιτέρω βελτιώσεις, είναι αυτό της Σχέσης (1.3) (mod1b).

1.3 Έχοντας πλέον ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης με λιγότερα χαρακτηριστικά απ' ότι πριν, μπορεί κανείς να προχωρήσει στη δεύτερη κατεύθυνση της βελτιστοποίησης, δηλαδή τη συναρτησιακή σχέση με την οποία υπεισέρχονται οι μεταβλητές στο μοντέλο. Η διαδικασία που ακολουθείται στο σημείο αυτό δε βασίζεται σε κάποιο θεωρητικό υπόβαθρο, παρά αντιστοιχεί σε μια trial & error διαδικασία, όπου διαφορετικές συναρτησιακές μορφές δοκιμάζονται και το μοντέλο αξιολογείται βάσει κάποιου κριτηρίου. Το κριτήριο επιλογής στην προκειμένη περίπτωση αποτελεί ο δείκτης R^2, εφόσον φυσικά το προκύπτον μοντέλο δεν εμφανίζει

χειρότερα αποτελέσματα σε ό,τι αφορά χαρακτηριστικά που έχουν ήδη μελετηθεί και φαίνονται ικανοποιητικά (όπως για παράδειγμα είναι η προσεγγιστική τήρηση των προϋποθέσεων για τα υπόλοιπα). Κατόπιν αρκετών δοκιμών, το τελικό, βελτιστοποιημένο μοντέλο προκύπτει βάσει του ακόλουθου κώδικα στην R:

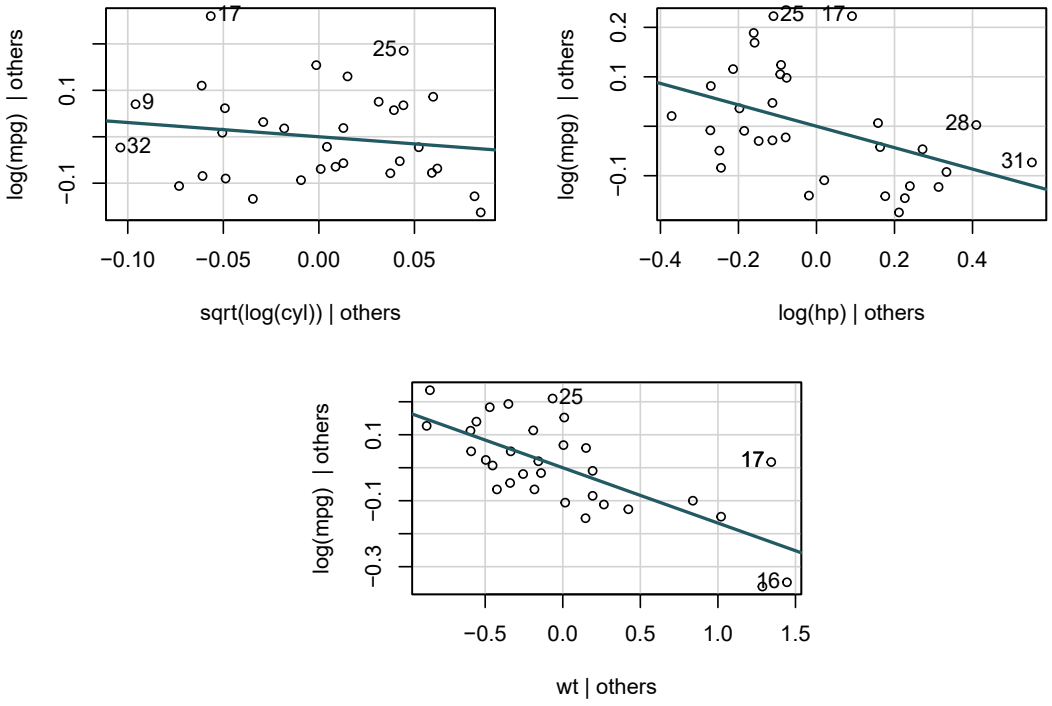
```
> mod1f <- lm(log(mpg) ~ sqrt(log(cyl)) + log(hp) + wt,
data=vehicles)
> summary(mod1f)

Residuals:
Min       1Q   Median       3Q      Max
-0.14439 -0.06764 -0.01354  0.07233  0.24250

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.96160    0.27093  18.313  < 2e-16 ***
sqrt(log(cyl)) -0.30654    0.36271  -0.845   0.4052
log(hp)        -0.21656    0.08120  -2.667   0.0126 *
wt             -0.16774    0.03083  -5.441 8.32e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

Residual standard error: 0.1048 on 28 degrees of freedom
Multiple R-squared:  0.8881,    Adjusted R-squared:  0.8761
F-statistic: 74.05 on 3 and 28 DF,  p-value: 1.989e-13
```

Δηλαδή, το τελικό μοντέλο είναι το

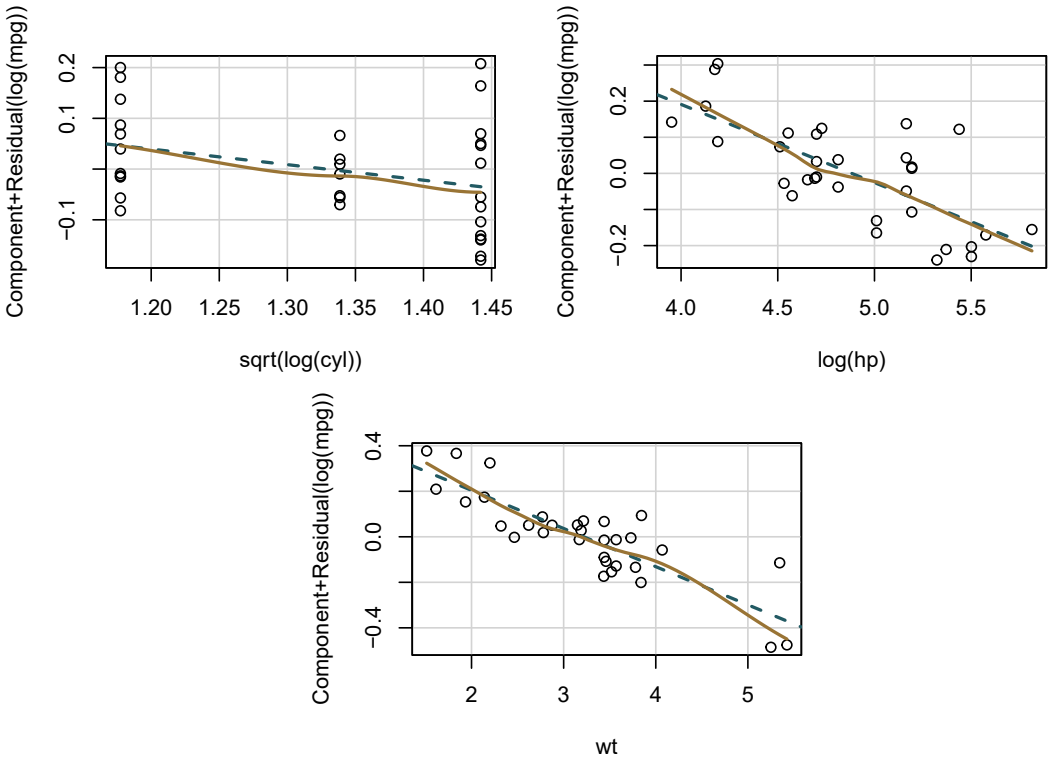
$$\ln(\mathbf{mpg}) = 4.962 - 0.168\mathbf{wt} - 0.307\sqrt{\ln(\mathbf{cyl})} - 0.217\ln(\mathbf{hp}), \tag{1.5}$$

για το οποίο υπολογίζεται $R^2 = 0.888$, δηλαδή η καλύτερη μέχρι στιγμής τιμή για οποιοδήποτε μοντέλο, ακόμα και αυτό των 10 επεξηγηματικών μεταβλητών. Για το τελικό αυτό μοντέλο πρέπει να επαναληφθούν εκ νέου οι έλεγχοι του πρώτου μέρους, προκειμένου να επιβεβαιωθεί πως η αύξηση του δείκτη R^2 δεν έρχεται μαζί με κάποιο κόστος σε άλλα χαρακτηριστικά. Πριν από αυτό, όμως, παρουσιάζονται τα διαγράμματα των πρόσθετων μεταβλητών, ώστε να φανεί εάν οι τρεις μεταβλητές, υπό τη συναρτησιακή σχέση που έχουν εισαχθεί στο μοντέλο, είναι πράγματι απαραίτητες σε αυτό. Καθένα εκ των διαγραμμάτων αφορά κάθε φορά τη μία από τις τρεις μεταβλητές και όσο πιο καλά μπορεί να προσεγγιστεί μέσω μιας ευθείας, τόσο πιο χρήσιμη είναι η αντίστοιχη μεταβλητή για το μοντέλο. Η σχετική εντολή στην R είναι η `avPlots(mod1f)` και τα διαγράμματα παρουσιάζονται στην Εικόνα 1.12.

Πράγματι, η τάση και στις τρεις περιπτώσεις είναι γραμμική, γεγονός που επιβεβαιώνει την ανάγκη ύπαρξης και των τριών μεταβλητών στο μοντέλο. Αξίζει να παρατεθούν στο σημείο αυτό και τα διαγράμματα μερικών υπολοίπων, μέσω της εντολής `crPlots(mod1f)` στην R, τα οποία φαίνονται στην Εικόνα 1.13. Εμφανώς, η συμβολή καθεμιάς εκ των τριών μεταβλητών στο μοντέλο τεκμαίρεται και από τα διαγράμματα αυτά και μάλιστα το γεγονός πως οι χρυσές γραμμές προσεγγίζουν πάρα πολύ καλά τις μπλε διακεκομμένες ευθείες σημαίνει πως η συναρτησιακή μορφή με την οποία οι μεταβλητές υπεισέρχονται στο μοντέλο είναι ικανοποιητική.



Εικόνα 1.12: Διαγράμματα πρόσθετων μεταβλητών για το τελικό μοντέλο.



Εικόνα 1.13: Διαγράμματα μερικών υπολοίπων για το τελικό μοντέλο.

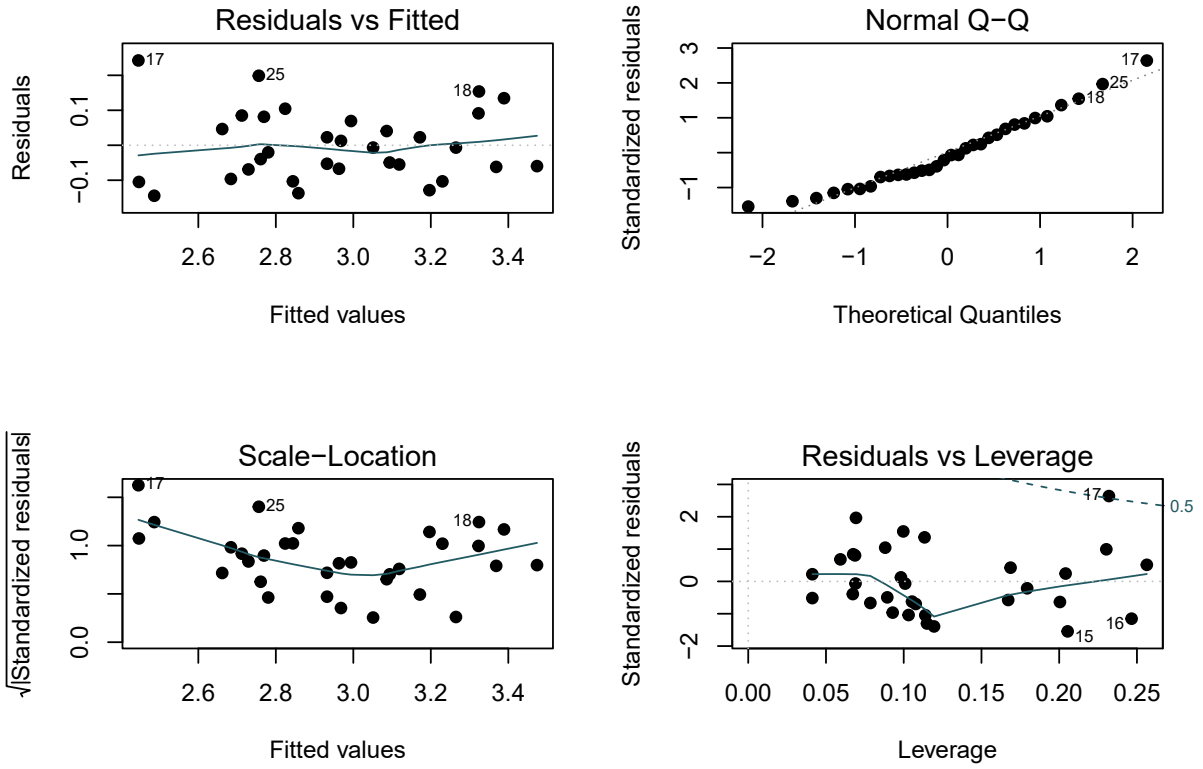
Έχοντας επαληθεύσει την ανάγκη ύπαρξης και των τριών μεταβλητών στο συγκεκριμένο μοντέλο, υπό τη συγκεκριμένη μάλιστα συναρτησιακή μορφή, μπορεί κανείς να προχωρήσει στο τελικό στάδιο της ανάλυσης, το οποίο είναι ο επανέλεγχος των χαρακτηριστικών που ελέγχθηκαν στο αρχικό μοντέλο (των 10 χαρακτηριστικών). Αρχικά, υπολογίζεται για καθένα εκ των

χαρακτηριστικών ο παράγοντας μεγέθυνσης διασποράς, προκειμένου να ελεγχθεί το ποσοστό πολυσυγγραμμικότητας του μοντέλου. Τα αποτελέσματα απεικονίζονται στον Πίνακα 1.2.

Χαρακτηριστικό	$\sqrt{\ln(\mathbf{cyl})}$	$\ln(\mathbf{hp})$	**wt**
VIF	5.17	4.20	2.57

Πίνακας 1.2: Παράγοντες μεγέθυνσης διασποράς (VIF) τις επεξηγηματικές μεταβλητές του τελικού μοντέλου.

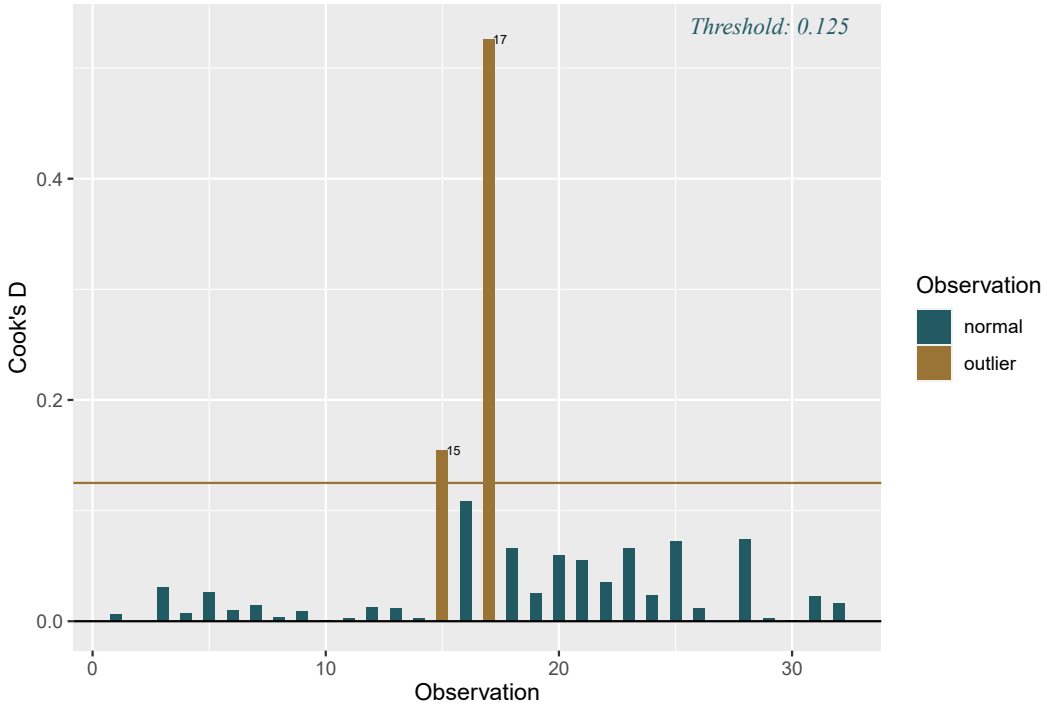
Γίνεται αντιληπτό πως το πρόβλημα της πολυσυγγραμμικότητας που υπήρχε στο αρχικό μοντέλο έχει μειωθεί σημαντικά, αφού μόνο η $\sqrt{\ln(\mathbf{cyl})}$ εμφανίζει VIF > 5, και αυτό οριακά. Στη συνέχεια (Εικόνα 1.14) παρουσιάζεται μια σύνοψη των διαγραμμάτων που αφορούν τους ελέγχους των υπολοίπων.



Εικόνα 1.14: Σύνοψη των γραφικών ελέγχων για τα υπόλοιπα που αφορούν το τελικό μοντέλο.

Και σε αυτήν την περίπτωση η εικόνα είναι βελτιωμένη σε σχέση με τα προηγούμενα μοντέλα, ειδικά στην περίπτωση του διαγράμματος q-q, όπου πρακτικά κανένα σημείο (με μοναδική πιθανή εξαίρεση το υπ' αριθμόν 17) δεν εμφανίζει άτυπη συμπεριφορά. Φυσικά, ο βέλτιστος τρόπος προσδιορισμού άτυπων σημείων ή σημείων επιρροής είναι μέσω του διαγράμματος της απόστασης Cook, το οποίο απεικονίζεται στην Εικόνα 1.15.

Πράγματι, φαίνεται το σημείο υπ' αριθμόν 17 να αποτελεί πιθανό σημείο επιρροής, ενώ το σημείο υπ' αριθμόν 15 επίσης ξεπερνά το κατώφλι, αν και οριακά.



Εικόνα 1.15: Απεικόνιση της απόστασης Cook στα πλαίσια του τελικού μοντέλου για καθένα εκ των 32 σημειών που αντιστοιχούν στα δεδομένα.

Το τελικό μοντέλο φαίνεται να αποδίδει ικανοποιητικά, τόσο ως προς τις συνθήκες που ικανοποιεί σε σχέση με τη θεωρία, όσο και ως προς την απόδοσή του με βάση το κριτήριο R^2. Το τελευταίο στάδιο της παρούσας ανάλυσης ξεκινά με την παράθεση του 95% διαστήματος εμπιστοσύνης για τους συντελεστές του τελικού αυτού μοντέλου, με τον κώδικα στην R και το αποτέλεσμά του να φαίνεται παρακάτω:

```
> confint(mod1f, level=0.95)

                      2.5 %      97.5 %
(Intercept)      4.4066303  5.51656825
sqrt(log(cyl)) -1.0495136  0.43643876
log(hp)        -0.3828973 -0.05021579
wt             -0.2308918 -0.10458648
```

Τα αποτελέσματα αυτά υποδεικνύουν πως η μεταβλητή $\sqrt{\ln(\mathbf{cyl})}$ ενδέχεται να μην είναι απαραίτητη για το τελικό μοντέλο, αφού το 0 ανήκει στο διάστημα 95% εμπιστοσύνης της. Το γεγονός αυτό συνάδει και με το γεγονός πως η p-value της μεταβλητής αυτής ήταν η μεγαλύτερη σε σχέση με τις υπόλοιπες. Παρ' όλα αυτά, η αφαίρεσή της από το μοντέλο δεν οδηγεί σε βελτίωση του R^2, ή σε καλύτερα αποτελέσματα σε ό,τι αφορά άλλα του χαρακτηριστικά, αν και η διαφορά είναι σχετικά μικρή.

Τέλος, αξίζει το τελικό μοντέλο να δοκιμαστεί σε νέα δεδομένα και η απόδοσή του να αξιολογηθεί βάσει της πρόβλεψής του σε αυτά. Για το σκοπό αυτό, χρησιμοποιούνται τα δεδομένα[3] για το μοντέλο 2015 Lexus NX I (AZ10), για το οποίο το βάρος είναι **wt** = 3.99, ο αριθμός των κυλίνδρων είναι 4 και άρα $\sqrt{\ln(\mathbf{cyl})} = 1.17741$, ενώ η μικτή υποδύναμη είναι 238 Hp, και άρα

[3] Η σχετική πηγή είναι η: https://www.auto-data.net/gr/lexus-nx-i-az10-200t-238hp-awd-automatic-21326

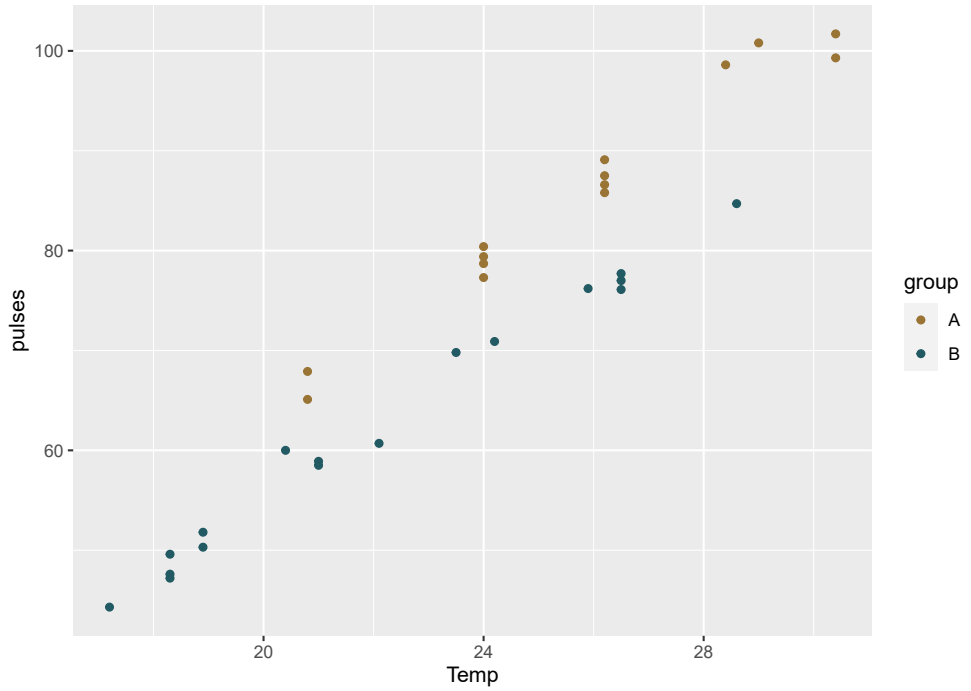
$\ln(\mathbf{hp}) = 5.47$. Χρησιμοποιώντας την εντολή `predict` της R, η πρόβλεψη δίνεται ίση με 3.8, με το κάτω όριο να ισούται με 3.24 και το άνω όριο με 4.37. Φυσικά, οι τιμές αυτές αντιστοιχούν στην $\ln(\mathbf{mpg})$, επομένως παίρνοντας το εκθετικό τους βρίσκουμε προσεγγιστικά το διάστημα $[25.53, 79.04]$, με πρόβλεψη την τιμή 44.7. Η πραγματική τιμή για την κατανάλωση δίνεται ίση με 43.46 mpg, γεγονός που υποδεικνύει πως το τελικό μοντέλο έχει παραπάνω από ικανοποιητική ακρίβεια στην πρόβλεψη τιμών για νέα δεδομένα.

2 Δεδομενα Καναριниων

Τα δεδομένα του Πίνακα Α του Παραρτήματος αφορούν τον αριθμό παλμών ανά δευτερόλεπτο κελαηδήματος (τιμές y) για δύο είδη καναρινιών, Α και Β, σε σχέση με διαφορετικές τιμές της θερμοκρασίας (τιμές x_1 μετρημένες σε ${}^oC$). Ονομάζοντας x_2 τη δείκτρια μεταβλητή που σχετίζεται με το είδος των καναρινιών, ακολουθείται η σύμβαση

$$x_{2i} = \begin{cases} 1, & \text{εάν } y_i \in A \\ 0, & \text{εάν } y_i \in B \end{cases}. \tag{2.1}$$

Τα δεδομένα αυτά απεικονίζονται στο επόμενο διάγραμμα διασποράς (Εικόνα 2.1), όπου με χρυσό χρώμα έχουν σχεδιαστεί τα δεδομένα του είδους Α και με μπλε χρώμα έχουν σχεδιαστεί τα δεδομένα του είδους Β.



Εικόνα 2.1: Διάγραμμα διασποράς δεδομένων του Πίνακα Α.

2.1 Ορίζοντας τη μεταβλητή $x_3 = x_1 x_2$, η οποία περιγράφει την αλληλεπίδραση μεταξύ των x_1 και x_2, μπορεί κανείς να ορίσει το μοντέλο πολλαπλής γραμμικής παλινδρόμησης

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \tag{2.2}$$

βάσει του οποίου θα κριθεί η προσαρμογή των ευθειών στο συγκεκριμένο πρόβλημα. Συγκεκριμένα, εάν ένα σημείο y_{A} αφορά το είδος Α, τότε το μοντέλο της Σχέσης (2.2) θα ισοδυναμεί με

$$E\left(y_{\mathrm{A}}\right) = \left(\beta_0 + \beta_2\right) + \left(\beta_1 + \beta_3\right) x_1, \tag{2.3}$$

ενώ στην περίπτωση όπου ένα σημείο y_{B} αφορά το είδος Β, τότε το μοντέλο θα γράφεται ως

$$E\left(y_{\mathrm{B}}\right) = \beta_0 + \beta_1 x_1. \tag{2.4}$$

Συγκρίνοντας τις Σχέσεις (2.3) και (2.4), γίνεται αντιληπτό πως η προσαρμογή των ευθειών καθορίζεται αποκλειστικά από τις τιμές των συντελεστών β_2 και β_3. Συγκεκριμένα, (Ι) εάν $\beta_3 \neq 0$, τότε οι συντελεστές του x_1 διαφέρουν στις δύο περιπτώσεις σημείων και ως εκ τούτου δύο ευθείες διαφορετικών κλίσεων απαιτούνται για την προσαρμογή στα δεδομένα. (ΙΙ) Εάν $\beta_3 = 0$, αλλά $\beta_2 \neq 0$, τότε ξανά δύο ευθείες θα πρέπει να προσαρμοστούν στα δεδομένα, μία για κάθε είδος, μόνο που στην περίπτωση αυτή η κλίση τους θα είναι κοινή και ίση με β_1 (παράλληλες ευθείες). (ΙΙΙ) Τέλος, εάν $\beta_3 = 0$ και $\beta_2 = 0$, τότε η ευθεία της Σχέσης (2.3) επαρκεί για την προσαρμογή στα δεδομένα και για τα δύο είδη. Κρίνοντας από την απεικόνιση των δεδομένων στην Εικόνα 2.1, φαίνεται πως η περίπτωση που αρμόζει στα δεδομένα είναι η (ΙΙ), όμως κάτι τέτοιο πρέπει να αιτιολογηθεί και βάσει στατιστικών ελέγχων.

Προκειμένου να εξαχθεί το συμπέρασμα σχετικά με το ποια από τις τρεις περιπτώσεις αρμόζει στο μοντέλο, προτείνονται οι ακόλουθοι έλεγχοι: αρχικά, ο έλεγχος της υπόθεσης $H_0 : \beta_3 = 0$ έναντι της $H_1 : \beta_3 \neq 0$, η οποία ισοδυναμεί με την υπόθεση απουσίας αλληλεπίδρασης μεταξύ των μεταβλητών x_1 και x_2. Εάν η υπόθεση αυτή απορριφθεί, τότε είναι βέβαιο πως μελετάται η περίπτωση (Ι). Εάν η υπόθεση αυτή δεν απορριφθεί, τότε ακολουθεί ο έλεγχος της υπόθεσης $H_0 : \beta_2 = 0$ έναντι της $H_1 : \beta_2 \neq 0$. Απόρριψη της υπόθεσης αυτής ισοδυναμεί με την περίπτωση (ΙΙ), ενώ αποδοχή της με την περίπτωση (ΙΙΙ).

2.2 Προκειμένου να πραγματοποιηθούν οι προτεινόμενοι έλεγχοι, αφότου τα δεδομένα φορτωθούν στο R Studio, πραγματοποιείται πολλαπλή γραμμική παλινδρόμηση. Αρχικά, λαμβάνονται υπ' όψιν και οι τρεις μεταβλητές, x_1, x_2 και x_3 (μοντέλο mod1 - υπόθεση H_1 για τη β_3) και κατόπιν λαμβάνονται υπ' όψιν μόνο τις x_1 και x_2 (μοντέλο mod2 - υπόθεση H_0 για τη β_3). Ο κώδικας σε R και τα αποτελέσματά του παρουσιάζονται παρακάτω.

```
> canary <- fread("canary.txt")
> cdata <- data.table(NULL)
> cdata$Y <- canary$pulses
> cdata$X_1 <- canary$Temp
> cdata$X_2 <- ifelse(canary$group == "A", 1, 0)
> cdata$X_3 <- cdata$X_1 * cdata$X_2
> mod1 <- lm(Y ~ X_1 + X_2 + X_3, data = cdata)
> summary(mod1)

Residuals:
Min      1Q  Median      3Q     Max
-3.7031 -1.3417 -0.1235  0.8100  3.6330

Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -15.3893     2.7173  -5.664 5.16e-06 ***
X_1           3.5175     0.1213  29.005  < 2e-16 ***
X_2           4.3484     4.9617   0.876    0.389
X_3           0.2340     0.2009   1.165    0.254
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

Residual standard error: 1.775 on 27 degrees of freedom
Multiple R-squared:  0.9901,    Adjusted R-squared:  0.989
F-statistic: 898.9 on 3 and 27 DF,  p-value: < 2.2e-16

> mod2 <- lm(Y ~ X_1 + X_2, data = cdata)
> summary(mod2)

Residuals:
Min      1Q  Median      3Q     Max
-3.0128 -1.1296 -0.3912  0.9650  3.7800

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.27620    2.19553  -7.869 1.43e-08 ***
X_1           3.60275    0.09729  37.032  < 2e-16 ***
X_2          10.06529    0.73526  13.689 6.27e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

Residual standard error: 1.786 on 28 degrees of freedom
Multiple R-squared:  0.9896,    Adjusted R-squared:  0.9888
F-statistic:  1331 on 2 and 28 DF,  p-value: < 2.2e-16
```

Ο έλεγχος της υπόθεσης γίνεται μέσω του F-test [1]

$$F = \frac{\mathrm{SSE}_2 - \mathrm{SSE}_1}{\mathrm{SSE}_1/(n-k-1)}, \tag{2.5}$$

όπου τα SSE_1 και SSE_2 είναι τα αθροίσματα τετραγώνων που αντιστοιχούν στα μοντέλα mod1 και mod2, αντίστοιχα, $n = 31$ είναι το πλήθος δεδομένων και $k = 3$, αφού ο δείκτης αφορά το μοντέλο mod1 (για το μοντέλο mod2 ισχύει $k = 2$). Ο υπολογισμός των SSE γίνεται σύμφωνα με τη σχέση

$$\mathrm{SSE} = (\text{Residual standard error})^2 \cdot (n-k-1), \tag{2.6}$$

όπου το Residual standard error είναι το αποτέλεσμα του κώδικα στην R. Βάσει της Σχέσης (2.6) προκύπτουν SSE_1 = 85.06687 και SSE_2 = 89.31429, επομένως η Σχέση (2.5) δίνει $F = 1.34811$. Τέλος, μέσω της εντολής `pf(1.34811, 1, 27, lower.tail = FALSE)` στην R, προκύπτει η p-value του συγκεκριμένου ελέγχου, ίση με 0.254. Αξίζει στο σημείο αυτό να αναφερθεί πως η προκύπτουσα τιμή ταυτίζεται με την p-value που δίνει αυτόματα η R για το συντελεστή του x_3 στο μοντέλο mod1. Αφού 0.254 > 0.05, η υπόθεση H_0 δε μπορεί να απορριφθεί, επομένως, πράγματι, το συμπέρασμα βάσει στατιστικής είναι πως $\beta_3 = 0$, όπως ήταν αναμενόμενο και μέσω απλής παρατήρησης του διαγράμματος διασποράς της Εικόνας 2.1.

Θεωρητικά, μια αντίστοιχη διαδικασία ακολουθείται και για τον έλεγχο της $H_0 : \beta_2 = 0$ έναντι της $H_1 : \beta_2 \neq 0$, δηλαδή η κατασκευή ενός μοντέλου mod3 στο οποίο παρούσα είναι μόνο η x_1 και ο υπολογισμός των (2.5) και (2.6) εκ νέου. Παρ' όλα αυτά, όπως φάνηκε παραπάνω, το αποτέλεσμα της μελέτης αυτής θα ταυτίζεται με την p-value που δίνεται για το συντελεστή της x_2 στη σύνοψη των αποτελεσμάτων για το μοντέλο mod2. Η τιμή αυτή είναι η $6.27 \cdot 10^{-14}$, συνεπώς η υπόθεση H_0 απορρίπτεται και άρα η x_2 είναι σημαντική, αφού ο συντελεστής της είναι μη μηδενικός.

Το τελικό μοντέλο είναι το

$$E\,(y) = -17.276 + 3.603x_1 + 10.065x_2. \tag{2.7}$$

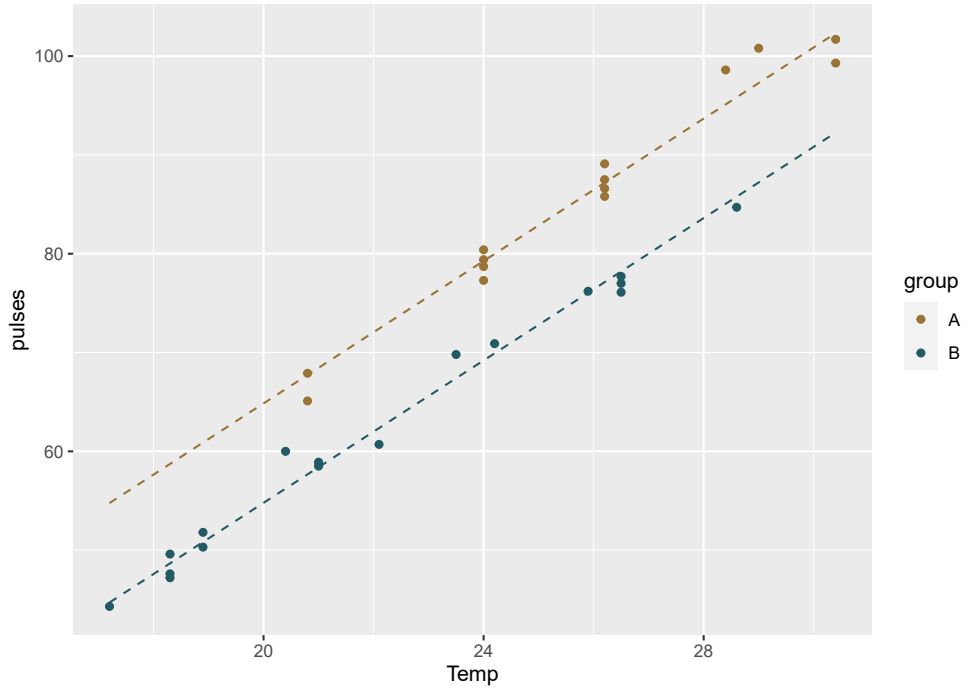
Η ποσότητα $\beta_0 = -17.276$ αντιστοιχεί στην αναμενόμενη τιμή της y εάν όλες οι μεταβλητές του μοντέλου (x_1,x_2) είναι ίσες με μηδέν. Η ποσότητα $\beta_1 = 3.603$ αντιστοιχεί στην αύξηση που θα παρατηρηθεί στην αναμενόμενη τιμή της y, εάν η μεταβλητή x_1 αυξηθεί κατά μία μονάδα. Τέλος, η ποσότητα $\beta_2 = 10.065$ αντιστοιχεί στην αύξηση που παρατηρείται στην αναμενόμενη τιμή της y όταν η παρατήρηση προέρχεται από την κατηγορία A, σε σύγκριση με μια παρατήρηση που προέρχεται από την κατηγορία B (κατηγορία αναφοράς - $x_2 = 0$). Οι ευθείες

$$E\,(y_{\mathrm{A}}) = -7.211 + 3.603x_1 \tag{2.8}$$

και

$$E\,(y_{\mathrm{B}}) = -17.276 + 3.603x_1, \tag{2.9}$$

που αντιστοιχούν στο μοντέλο της Σχέσης (2.7), απεικονίζονται προσαρμοσμένες στα δεδομένα κάθε είδους στη γραφική παράσταση της Εικόνας 2.2.



Εικόνα 2.2: Προσαρμογή του μοντέλου της Σχέσης (2.7) στα δεδομένα του Πίνακα Α.

Παραρτημα

Ακολουθεί ο πίνακας με τα δεδομένα που χρησιμοποιήθηκαν στη 2η Άσκηση.

x_2	x_1	y		x_2	x_1	y
1	20.8	67.9		0	17.2	44.3
1	20.8	65.1		0	18.3	47.2
1	24.0	77.3		0	18.3	47.6
1	24.0	78.7		0	18.3	49.6
1	24.0	79.4		0	18.9	50.3
1	24.0	80.4		0	18.9	51.8
1	26.2	85.8		0	20.4	60.0
1	26.2	86.6		0	21.0	58.5
1	26.2	87.5		0	21.0	58.9
1	26.2	89.1		0	22.1	60.7
1	28.4	98.6		0	23.5	69.8
1	29.0	100.8		0	24.2	70.9
1	30.4	99.3		0	25.9	76.2
1	30.4	101.7		0	26.5	76.1
				0	26.5	77.0
				0	26.5	77.7
				0	28.6	84.7

Πίνακας Α: Δεδομένα παλμών καναρινιών συναρτήσει της θερμοκρασίας.

Αναφορες

[1] Χ. Καρώνη και Π. Οικονόμου, *Στατιστικά Μοντέλα Παλινδρόμησης με χρήση MINITAB και R*. Εκδόσεις Συμεών, 2020.