Εθνικό Μετσόβιο Πολυτεχνείο
Επιστήμη Δεδομένων και Μηχανική Μάθηση

ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Σειρά 2

Νάνος Γεώργιος
03400144
nanosgiwrgos1997@gmail.com

"The classification of facts, the recognition of their sequence and relative significance is the function of science, and the habit of forming a judgment upon these facts unbiassed by personal feeling is characteristic of what may be termed the scientific frame of mind."

– Karl Pearson - Statistician

Άσκηση Α

Στο αρχείο vehicles.txt παρουσιάζονται τα αποτελέσματα μιας μελέτης για 32 τύπους αυτοκινήτων. Η πρώτη στήλη δίνει τη λίστα των τύπων αυτοκινήτων (car). Ακολουθούν οι μεταβλητές:

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

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 και αν τηρούνται οι προϋποθέσεις του μοντέλου με βάση την εξέταση των υπολοίπων. Να γίνει χρήση διαγνωστικών ελέγχων π.χ. για την πιθανή παρουσία άτυπων σημείων ή σημείων επιρροής, αξιοποιώντας μέτρα όπως τα h_{ii}, απόσταση Cook, DFFiTS, DFBETAS .**

   Προσαρμόζουμε το μοντέλο πολλαπλής γραμμικής παλινδρόμησης σχετίζοντας τα $mpg(Y)$ με τις υπόλοιπες 10 επεξηγηματικές μεταβλητές.

```
cars <- read.table('./cars.txt', header = TRUE)
attach(cars)
model <- lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb)
summary(model)
```

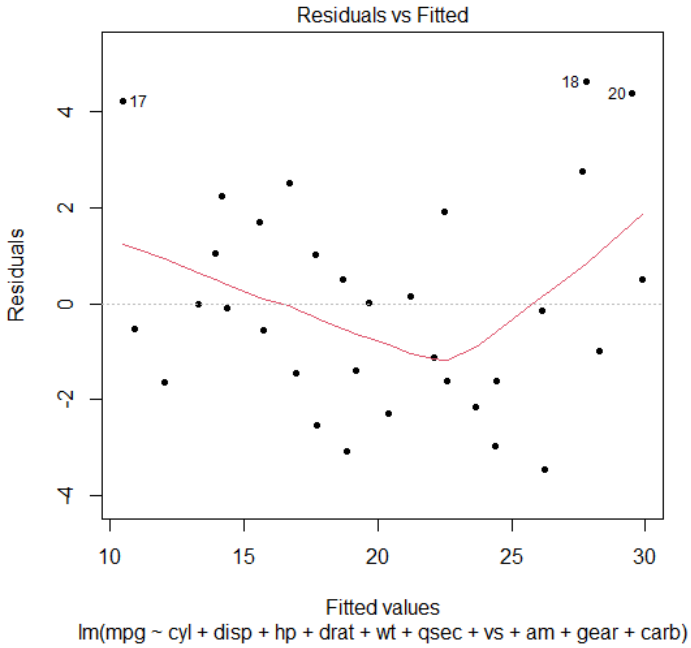
```
Call:
lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +
    am + gear + carb)

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
```

Figure 1: Οι στατιστικές πληροφορίες του μοντέλου

Οι πληροφορίες του μοντέλου παρατηρούμε ότι το p value είναι αρκετά μεγάλο, ενώ οι όλες οι μεταβλητές είναι στατιστικά ασήμαντες. Πρακτικά, αυτό σημαίνει ότι υπάρχει μεγάλη συσχέτιση (corellation) μεταξύ των μεταβλητών συνεπώς δεν αποσπάμε παραπάνω πληροφορίες από αυτές στο μοντέλο μας. Στη συνέχεια, κάνοντας χρήση διαγνωστικών ελέγχων προσπαθούμε να εντοπίσουμε περισσότερες προϋποθέσεις του μοντέλου οι οποίες δεν τηρούνται.

- Διάγραμμα Residuals vs fitted values

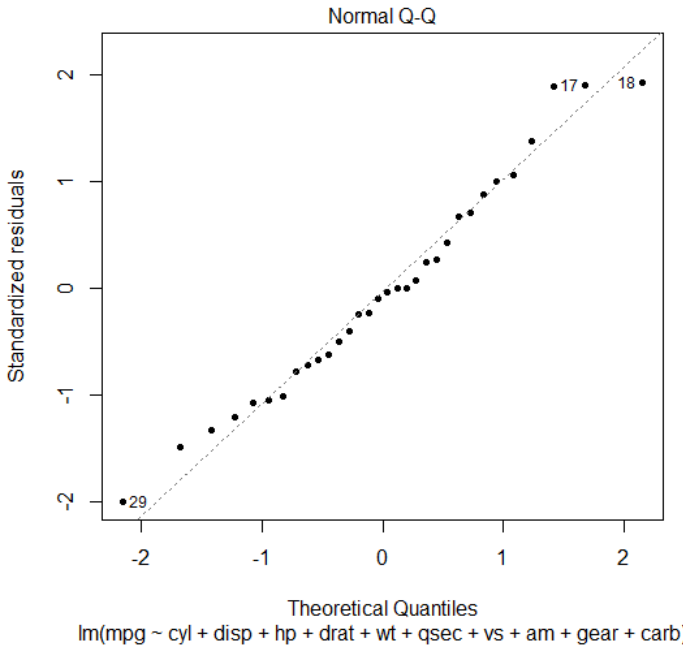
$$\mathbf{plot}(\mathbf{model},\ \mathbf{which}=1, \mathrm{pch}=20)$$



Παρατηρούμε ότι στο διάγραμμα, τα σημεία δεν είναι διασκορπισμένα τυχαία αλλά, τα περισσότερα είναι κάτω από το 0. Συνεπώς, υπάρχει ετεροσκεδαστικότητα στα δεδομένα μας, την οποία προσπαθούμε να αποφύγουμε.

- Normal qq plot

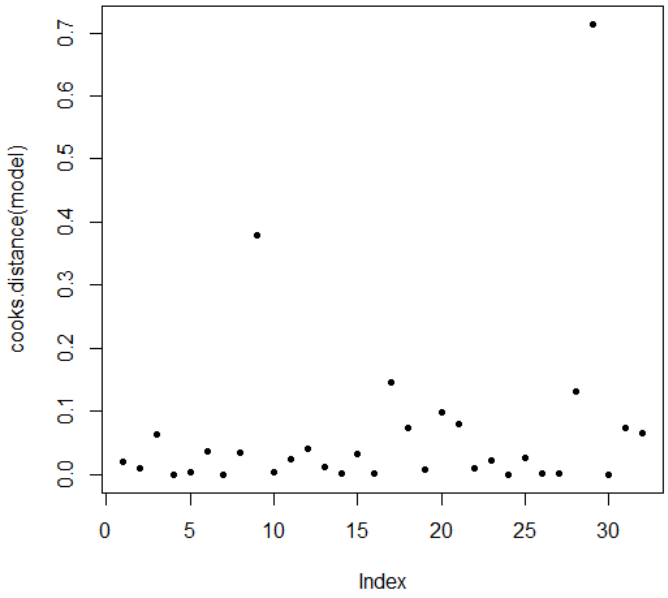
$$\mathbf{plot}(\mathbf{model},\ \mathbf{which}=2, \mathrm{pch}=20)$$



Παρατηρούμε ότι στο διάγραμμα qqplot παραβιάζεται σε ένα βαθμό η κανονικότητα, εφόσον υπάρχουν ορισμένα σημεία που δεν πέφτουν κοντά στην ευθεία.

- Διάγραμμα Cook Distance vs index

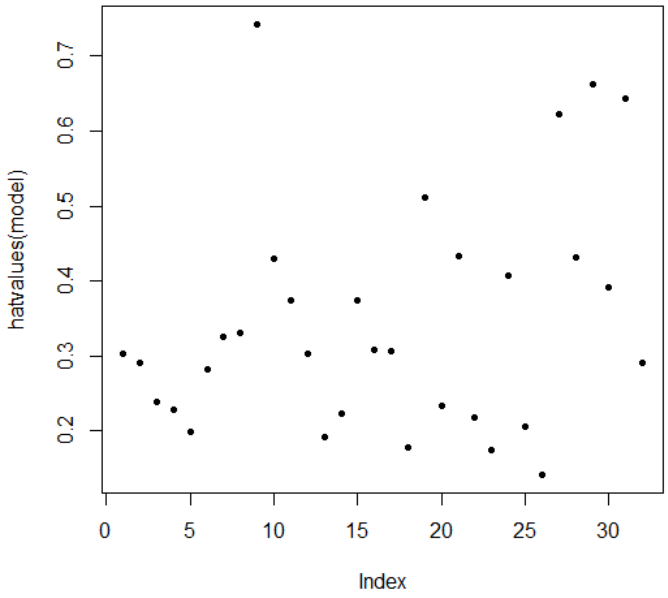
$$\mathbf{plot}\,(\mathbf{cooks.distance}\,(\mathbf{model})\,,\ \ \mathrm{pch}=20)$$



Παρατηρούμε ότι, υπάρχουν ορισμένα outliers, αλλά κανένα σημείο δεν ξεπερνάει το 1 οπότε δε μπορούμε να πούμε ότι έχουμε κάποιο σημείο επιρροής απο την απόσταση Cook.

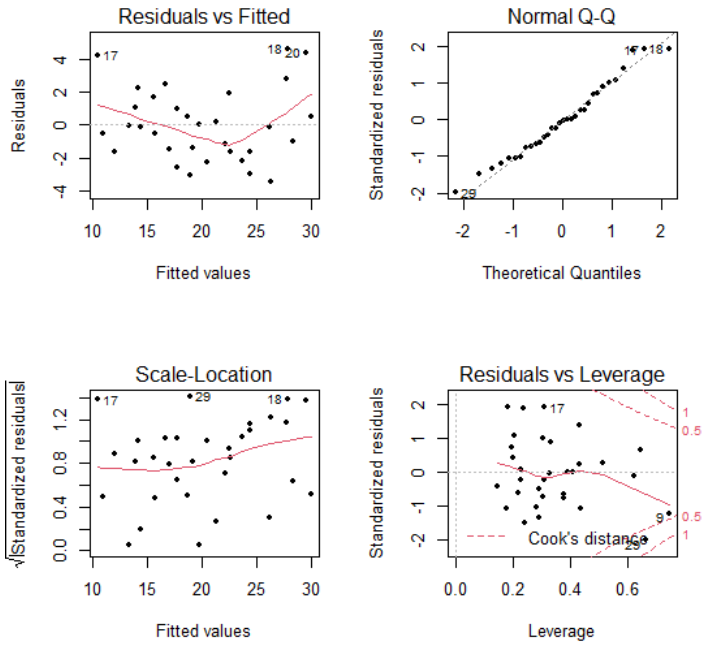
- Leverage plot για hat values H μέθοδος leverage (μόχλευση) χρησιμοποιεί μόνο τις ανεξάρτητες μεταβλητές και όχι και τις εξαρτημένες όπως η μέθοδος της απόστασης Cook που πραγματοποιήθηκε παραπάνω.

plot(hatvalues(**model**), pch=20)



Σύμφωνα με το διάγραμμα leverage για να θεωρήσουμε μια παρατήρηση ως σημείο επιρροής πρέπει να ισχύει, $hii > 2 * 10/32 = 0.625$. Παρατηρούμε πως έχουμε 4 τέτοια σημεία που επηρεάζουν την προσαρμογή του μοντέλου μας.

- Σύνολο γραφικών παραστάσεων.



Έπειτα κάνοντας χρήση της εντολής **dfbetas(model)** ελέγχουμε την επιρροή των παρατηρήσεων σε κάθε ένα coefficient. Αν κάποιο από αυτά έχει μεγάλο DFBETAS VALUE τότε η συγ-

κεκριμένη παρατήρηση ασκεί επιρροή στο μοντέλο. Μία α παρατήρηση i ασκεί επιρροή όταν:

$$|DFBETAS_{ij}| > \frac{2}{\sqrt{n}DFBETAS_{ij}}$$

όπου, για n = 32 $|DFBETAS_{ij}| > 0.35$

```
> dfbetas(model)
   (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 -0.0080168469 -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  0.0612215256  0.0595024864 -0.0800701708  1.152810e-01 -0.0833337299  0.0474909343  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 -1.201149e-01 -0.0083615610  0.0365739106 -0.3873200989
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 -0.0454868094  0.0040360128 -4.740677e-03 -0.0464016613  0.0107269720 -0.0128565290
17  0.0429492736 -0.3060891928 -0.1948915128  0.2324904149  0.3214121655  0.7752259274 -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 -0.0609476021 -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 -0.2209181204 -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
```

Υπάρχουν σημεία τα οποία ξεπερνάνε την τιμή αυτή. Ένα από αυτά είναι η παρατήρηση 29 η οποία παίρνει την τιμή 4.5 στη μεταβλητή "vs" γεγονός που δείχνει ότι ασκεί vs" γεγονός που δείχνει ότι ασκείμεγάλη επιρροή πάνω σε αυτήν. Παρατηρούμε, ότι η παρατήρηση **9** παίρνει την τιμή **0.937** στη μεταβλητή 'vs' συνεπώς της ασκεί αρκετά μεγάλη επιρροή.

Στη συνέχεια για να εντοπίσουμε τυχόν σημεία επιρροής με την εντολή **dffits(model)** ελέγχουμε κατά πόσο θα επηρεαστεί το μοντέλο αν δεν συμπεριλάβουμε κάποια παρατήρηση. Για να επηρεάζει μία παρατήρηση το μοντέλο χρειάζεται:

$$|DFFITSS_{ij}| > 2\sqrt{\frac{p}{n}}$$

όπου στη συγκεκριμένη περίπτωση p = 10, n = 32, συνεπώς για να ασκείται επιρροή από ένα σημείο πρέπει: $|DFFITSS_{ij}| > 1.11$

```
> dffits(model)
          1           2           3           4           5           6
-0.470336460 -0.312613697 -0.862811386  0.037003357  0.207716116 -0.638256415
          7           8           9          10          11          12
-0.026901447  0.612906168 -2.065687952  0.212125552 -0.507249992  0.664699488
         13          14          15          16          17          18
 0.343914128 -0.121788112 -0.597355814 -0.159154958  1.360566891  0.967564035
         19          20          21          22          23          24
 0.272152176  1.115696987 -0.943257304 -0.320186447 -0.484676559 -0.002393136
         25          26          27          28          29          30
 0.541365500 -0.161425967 -0.118037702  1.231964143 -3.037374792  0.002333660
         31          32
 0.884515333 -0.867680270
```

Βλέπουμε ότι οι παρατηρήσεις που ασκούν τη μεγαλύτερη επιρροή είναι οι: 29, 28, 17 και 20.

STUDENTIZED RESIDUALS PLOT NOT DONE

Πολυσυγγραμικότητα (Multicollinearity)

Για τον έλεγχο πολυσυγγραμικότητας κάνουμε χρήση της εντολής **vif(model)**

```
> vif(model)
      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 υπάρχουν πολλές τιμές με υψυλό δείκτη VIF γεγονός που σημαίνει ότι σίγουρα επικρατεί πολυσυγγραμμικότητα. Η μεταβλητή "disp" σύμφωνα με την προσέγγιση VIF, αποτελεί την πρώτη μεταβλητή υποψήφια προς αποχώρηση από το μοντέλο.

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

Αφαιρώντας στη συνέχεια και τη μεταβλητή 'cyl' παρατηρούμε σημαντική μείωση της πολυσυγγραμικότητας.

```
> vif(lm(mpg~hp+drat+wt+qsec+vs+am+gear+carb))
      hp     drat       wt     qsec       vs       am     gear     carb
6.015788 3.111501 6.051127 5.918682 4.270956 4.285815 4.690187 4.290468
```

Figure 2: Γραφική παράσταση για το συνδυασμό κάθε επεξηγηματικής μεταβλητής

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, $\overline{R}$, $R^2_{predict}$, C_p και AIC.**

   - **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. Η διαδικασία θα διακοπεί όταν η εισαγωγή μιας νέας μεταβλητής πάψει να προσφέρει ιδιαίτερη πληροφορία στο μοντέλο.

```
> 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
```

Ο αλγόριθμος forward selection καταλήγει στο μοντέλο wt + cyl + hp.

- **Step Selection - Backward with F test** Αρχικά ξεκινάμε με το μοντέλο χωρίς να περιέχει όλες τις μεταβλητές και στη συνέχεια αφαιρούμε από αυτό μεταβλητές ανάλογα με το πόσο βελτίωση αυτές προσφέρουν.

```
> mod_bw = step(model, 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

Step:  AIC=66.97
mpg ~ disp + hp + drat + wt + qsec + am + gear + carb

       Df Sum of Sq    RSS    AIC F value  Pr(>F)
- carb  1    0.6855 148.53 65.121  0.1066 0.74696
- gear  1    2.1437 149.99 65.434  0.3335 0.56922
- drat  1    2.2139 150.06 65.449  0.3444 0.56301
- disp  1    3.6467 151.49 65.753  0.5673 0.45897
- hp    1    7.1060 154.95 66.475  1.1055 0.30399
<none>              147.84 66.973
- am    1   11.5694 159.41 67.384  1.7999 0.19283
- qsec  1   15.6830 163.53 68.200  2.4398 0.13195
- wt    1   27.3799 175.22 70.410  4.2595 0.05049 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Step:  AIC=65.12
mpg ~ disp + hp + drat + wt + qsec + am + gear

       Df Sum of Sq    RSS    AIC F value   Pr(>F)
- gear  1     1.565 150.09 63.457  0.2529 0.619641
- drat  1     1.932 150.46 63.535  0.3122 0.581508
<none>              148.53 65.121
- disp  1    10.110 158.64 65.229  1.6337 0.213420
- am    1    12.323 160.85 65.672  1.9913 0.171042
- hp    1    14.826 163.35 66.166  2.3956 0.134763
- qsec  1    26.408 174.94 68.358  4.2672 0.049815 *
- wt    1    69.127 217.66 75.350 11.1699 0.002717 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=63.46
mpg ~ disp + hp + drat + wt + qsec + am

       Df Sum of Sq    RSS    AIC F value   Pr(>F)
- drat  1     3.345 153.44 62.162  0.5571 0.462401
- disp  1     8.545 158.64 63.229  1.4233 0.244054
<none>              150.09 63.457
- hp    1    13.285 163.38 64.171  2.2127 0.149381
- am    1    20.036 170.13 65.466  3.3372 0.079692 .
- qsec  1    25.574 175.67 66.491  4.2598 0.049551 *
- wt    1    67.572 217.66 73.351 11.2550 0.002536 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=62.16
mpg ~ disp + hp + wt + qsec + am

       Df Sum of Sq    RSS    AIC F value   Pr(>F)
- disp  1     6.629 160.07 61.515  1.1232 0.298972
<none>              153.44 62.162
- hp    1    12.572 166.01 62.682  2.1303 0.156387
- qsec  1    26.470 179.91 65.255  4.4853 0.043908 *
- am    1    32.198 185.63 66.258  5.4559 0.027488 *
- wt    1    69.043 222.48 72.051 11.6993 0.002075 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=61.52
mpg ~ hp + wt + qsec + am

       Df Sum of Sq    RSS    AIC F value   Pr(>F)
- hp    1     9.219 169.29 61.307  1.5551 0.223088
<none>              160.07 61.515
- qsec  1    20.225 180.29 63.323  3.4115 0.075731 .
- am    1    25.993 186.06 64.331  4.3845 0.045791 *
- wt    1    78.494 238.56 72.284 13.2403 0.001141 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



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
```

Ο αλγόριθμος backward selection with F-test καταλήγει στο μοντέλο 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.

```
> mod_both = step(model, direction = 'both', 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
+ cyl   1    0.0799 147.49 70.898  0.0114 0.91609
- wt    1   27.0280 174.60 72.297  4.0293 0.05716 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Step:  AIC=66.97
mpg ~ disp + hp + drat + wt + qsec + am + gear + carb

       Df Sum of Sq    RSS    AIC F value  Pr(>F)
- carb  1    0.6855 148.53 65.121  0.1066 0.74696
- gear  1    2.1437 149.99 65.434  0.3335 0.56922
- drat  1    2.2139 150.06 65.449  0.3444 0.56301
- disp  1    3.6467 151.49 65.753  0.5673 0.45897
- hp    1    7.1060 154.95 66.475  1.1055 0.30399
<none>              147.84 66.973
- am    1   11.5694 159.41 67.384  1.7999 0.19283
- qsec  1   15.6830 163.53 68.200  2.4398 0.13195
+ vs    1    0.2685 147.57 68.915  0.0400 0.84326
+ cyl   1    0.1883 147.66 68.932  0.0281 0.86852
- wt    1   27.3799 175.22 70.410  4.2595 0.05049 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=65.12
mpg ~ disp + hp + drat + wt + qsec + am + gear

       Df Sum of Sq    RSS    AIC F value   Pr(>F)
- gear  1     1.565 150.09 63.457  0.2529 0.619641
- drat  1     1.932 150.46 63.535  0.3122 0.581508
<none>              148.53 65.121
- disp  1    10.110 158.64 65.229  1.6337 0.213420
- am    1    12.323 160.85 65.672  1.9913 0.171042
- hp    1    14.826 163.35 66.166  2.3956 0.134763
+ carb  1     0.685 147.84 66.973  0.1066 0.746958
+ vs    1     0.434 148.09 67.028  0.0674 0.797502
+ cyl   1     0.414 148.11 67.032  0.0644 0.801995
- qsec  1    26.408 174.94 68.358  4.2672 0.049815 *
- wt    1    69.127 217.66 75.350 11.1699 0.002717 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Step:  AIC=63.46
mpg ~ disp + hp + drat + wt + qsec + am

       Df Sum of Sq    RSS    AIC F value   Pr(>F)
- drat  1     3.345 153.44 62.162  0.5571 0.462401
- disp  1     8.545 158.64 63.229  1.4233 0.244054
<none>              150.09 63.457
- hp    1    13.285 163.38 64.171  2.2127 0.149381
+ gear  1     1.565 148.53 65.121  0.2529 0.619641
+ cyl   1     1.003 149.09 65.242  0.1615 0.691314
+ vs    1     0.645 149.45 65.319  0.1037 0.750269
+ carb  1     0.107 149.99 65.434  0.0171 0.897106
- am    1    20.036 170.13 65.466  3.3372 0.079692 .
- qsec  1    25.574 175.67 66.491  4.2598 0.049551 *
- wt    1    67.572 217.66 73.351 11.2550 0.002536 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=62.16
mpg ~ disp + hp + wt + qsec + am

       Df Sum of Sq    RSS    AIC F value   Pr(>F)
- disp  1     6.629 160.07 61.515  1.1232 0.298972
<none>              153.44 62.162
- hp    1    12.572 166.01 62.682  2.1303 0.156387
+ drat  1     3.345 150.09 63.457  0.5571 0.462401
+ gear  1     2.977 150.46 63.535  0.4947 0.488331
+ cyl   1     2.447 150.99 63.648  0.4051 0.530249
+ vs    1     1.121 152.32 63.927  0.1840 0.671666
+ carb  1     0.011 153.43 64.160  0.0019 0.965925
- qsec  1    26.470 179.91 65.255  4.4853 0.043908 *
- am    1    32.198 185.63 66.258  5.4559 0.027488 *
- wt    1    69.043 222.48 72.051 11.6993 0.002075 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Step:  AIC=61.52
mpg ~ hp + wt + qsec + am

       Df Sum of Sq    RSS    AIC F value   Pr(>F)
- hp    1     9.219 169.29 61.307  1.5551 0.223088
<none>              160.07 61.515
+ disp  1     6.629 153.44 62.162  1.1232 0.298972
+ carb  1     3.227 156.84 62.864  0.5350 0.471059
+ drat  1     1.428 158.64 63.229  0.2340 0.632611
- qsec  1    20.225 180.29 63.323  3.4115 0.075731 .
+ cyl   1     0.249 159.82 63.465  0.0405 0.842062
+ vs    1     0.249 159.82 63.466  0.0404 0.842196
+ gear  1     0.171 159.90 63.481  0.0278 0.868810
- am    1    25.993 186.06 64.331  4.3845 0.045791 *
- wt    1    78.494 238.56 72.284 13.2403 0.001141 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=61.31
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              169.29 61.307
+ hp    1     9.219 160.07 61.515  1.5551 0.2230879
+ carb  1     8.036 161.25 61.751  1.3456 0.2562120
+ disp  1     3.276 166.01 62.682  0.5328 0.4717085
+ cyl   1     1.501 167.78 63.022  0.2416 0.6270601
+ drat  1     1.400 167.89 63.042  0.2251 0.6390028
+ gear  1     0.123 169.16 63.284  0.0196 0.8897368
+ vs    1     0.000 169.29 63.307  0.0001 0.9931865
- 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
```

Ο αλγόριθμος both selection with F-test καταλήγει στο μοντέλο wt + qsec + am, το οποίο είναι ίδιο με αυτό του αλγορίθμου backward selection with F-test.

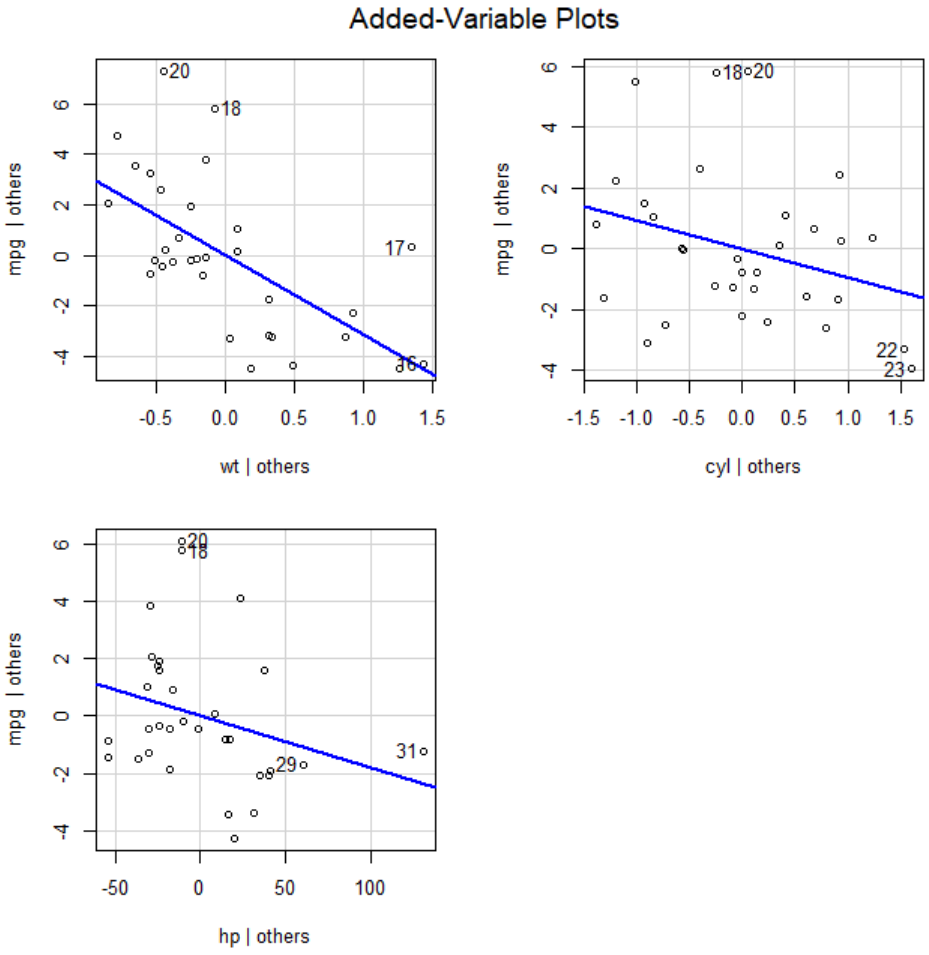
	model : WT + CYL +HP	model : WT + QSEC + AM
AIC	155.4766	154.1194
Rsq	0.8431	0.8497

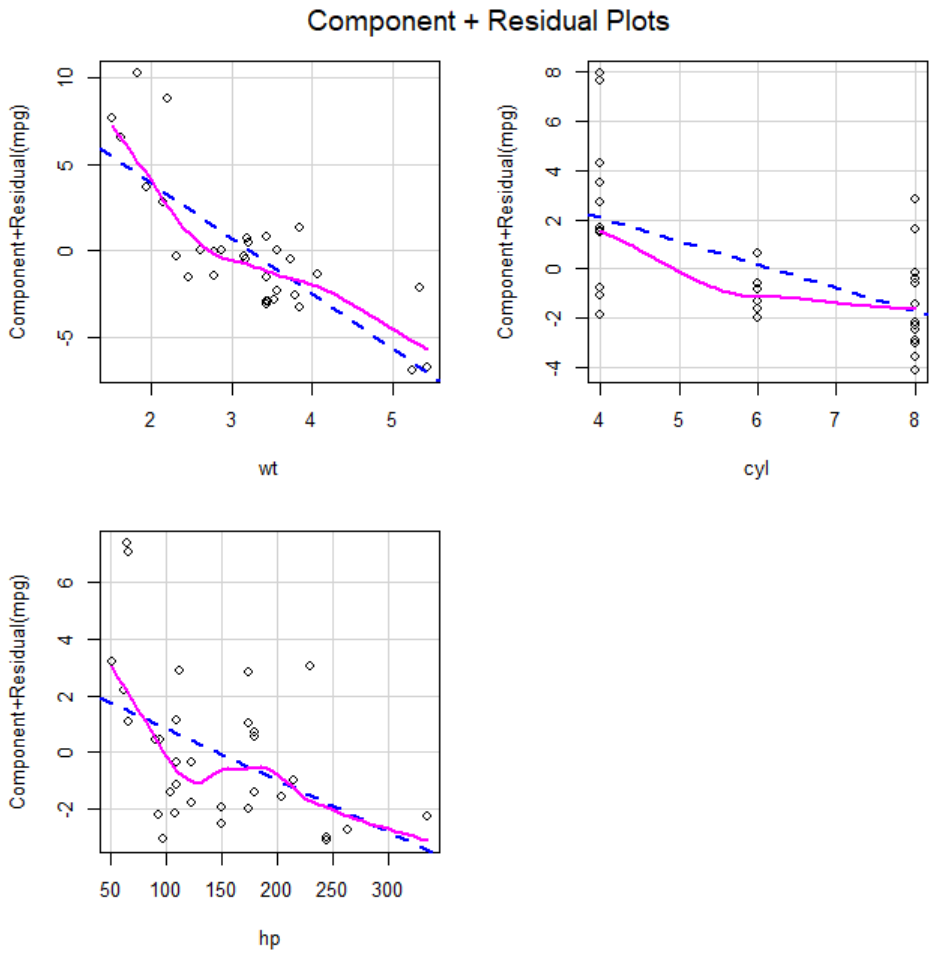
Radj	0.8263	0.8336
Press	230.0767	231.3025
Cp - mallows	1.146922	0.1026357
Num. of ind. var.	3	3

Από τα παραπάνω ισχύουν τα εξής:

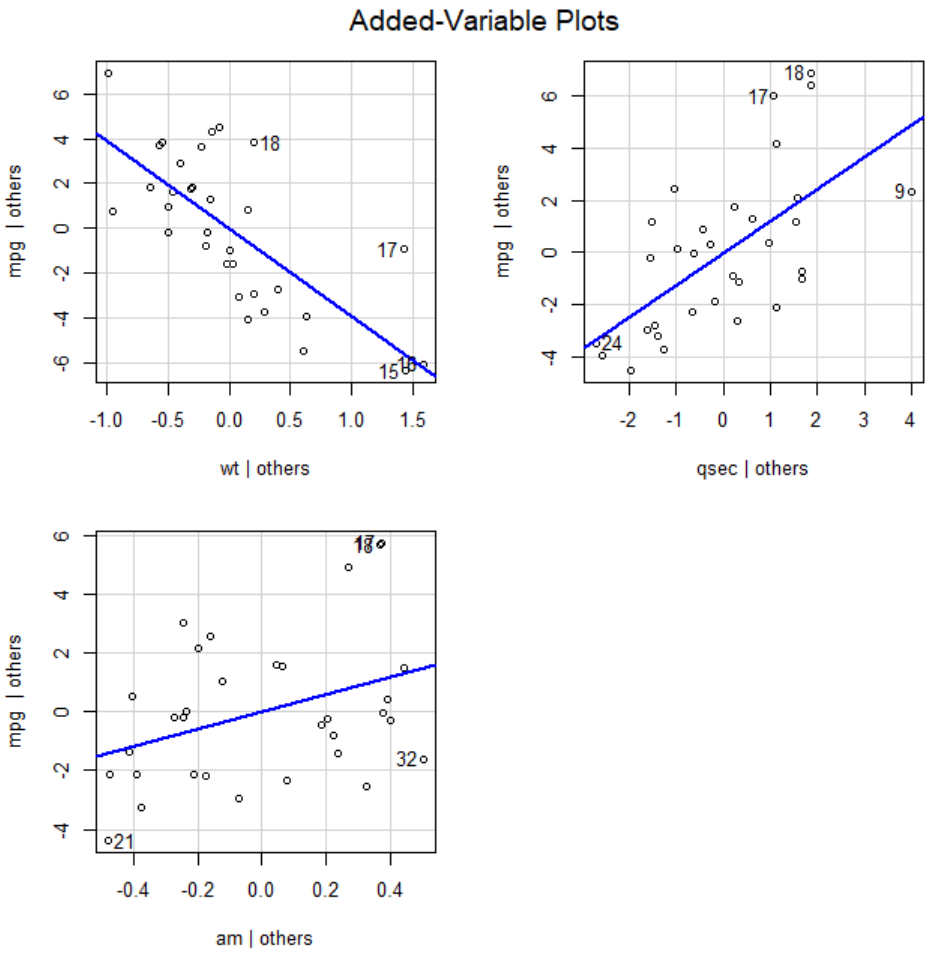
- Θέλουμε το μοντέλο με το χαμηλότερο AIC, επομένως αυτό το κριτήριο αποφασίζει το 2ο μοντέλο.
- Τα Rsq είναι πολύ κοντά όμως τα Rsq-adj έχουν διαφορά 0,1 μεταξύ τους. Το Rsq-adjusted είναι καλύτερο κριτήριο επιλογής καθώς όσες περισσότερες μεταβλητές βάζουμε το Rsq αυξάνεται. Το adjusted Rsq όμως προσπαθεί 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.
- Το PRESS χρησιμοποιείται για σύγκριση μοντέλων ως προς την ικανότητα πρόβλεψης νέων παρατηρήσεων. Γενικά, προτιμάται το μοντέλο με το μικρότερο PRESS, το οποίο εκφράζει το Rsq-pred όπου, ισούται με Rsq-pred= 1 – PRESS/SST και στην περίπτωση αυτή διαλέγουμε το μεγαλύτερο Rsq-pred. Σύμφωνα με αυτό το κριτήριο διαλέγουμε το μοντέλο 1.
- Το cp-mallows διαλέγει το μοντέλο 1, γιατί έχει μικρότερη τιμή.

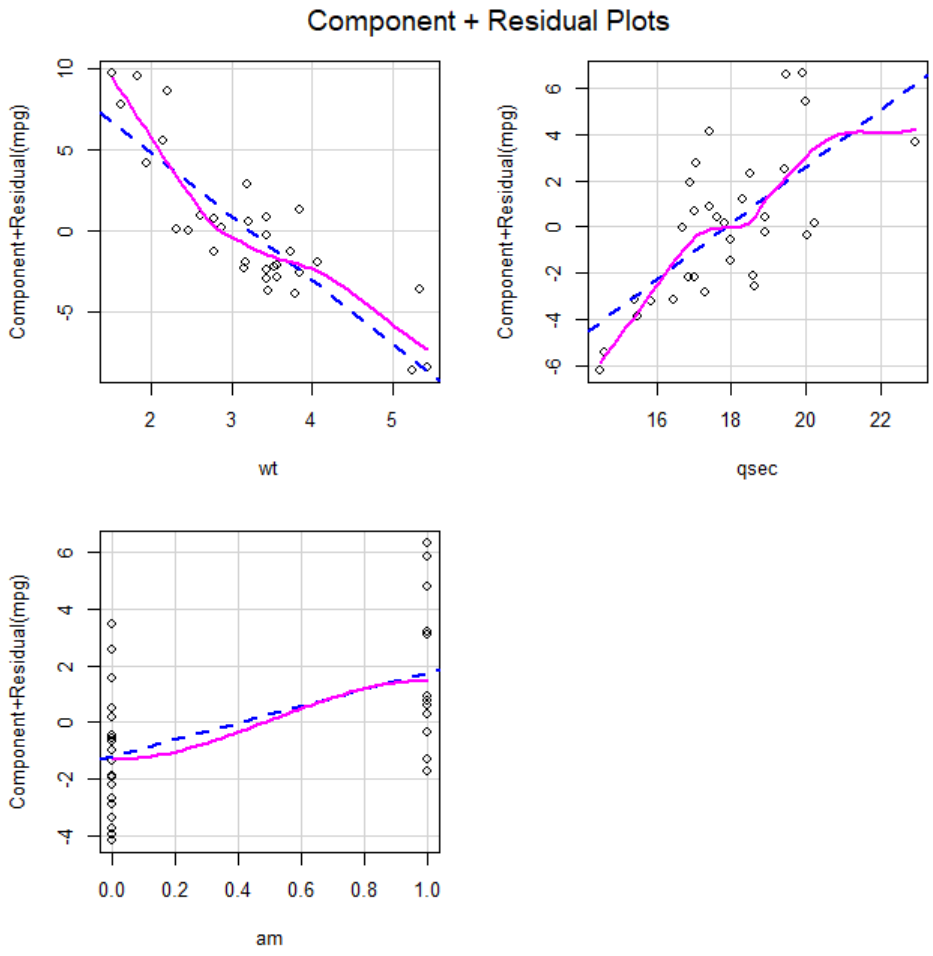
Χρήση added variable και component residual plots για το Forward selection μοντέλο





Χρήση added variable και component residual plots για το Backward selection μοντέλο

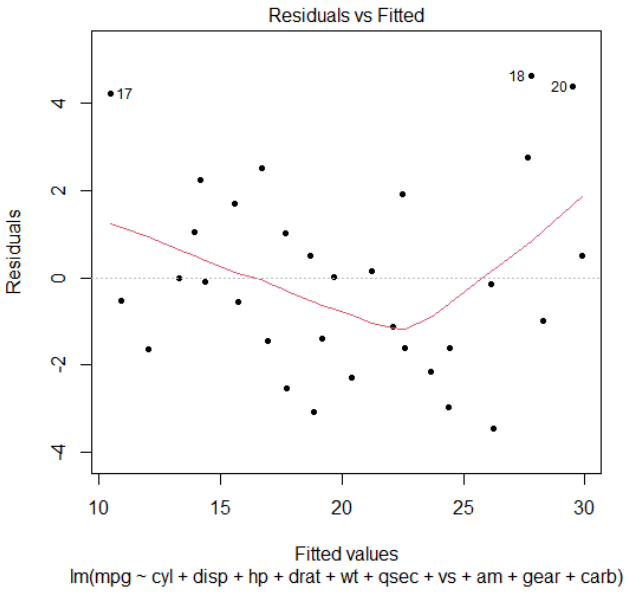

Added-Variable Plots
mpg | others
wt | others
qsec | others
am | others
18
17
15
9
24
32
21

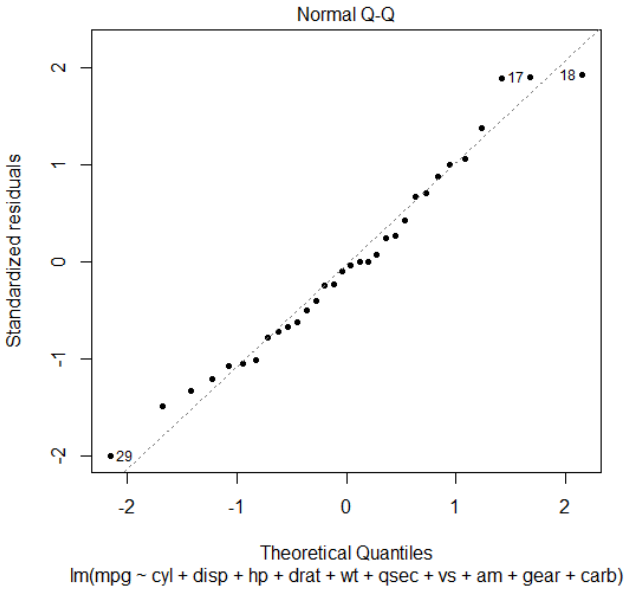


Στο 2ο μοντέλο έχουμε για τις 2 μεταβλητές αρκετά καλοσχηματισμένη ευθεία σε αντίθεση με το 1ο μοντέλο το οποίο μόνο στην 1η μεταβλητή ορίζει μια καλή ευθεία. Ταυτόχρονα στα διαγράμματα Component και Residual Plots φαίνεται να υπερέχει το μοντέλο backward selection, καθώς τα δεδομένα επίσης, τείνουν σε καλύτερες ευθείες από ότι τα διαγράμματα του μοντέλου forward selection. Συνεπώς το μοντέλο που θα επιλεγεί είναι:

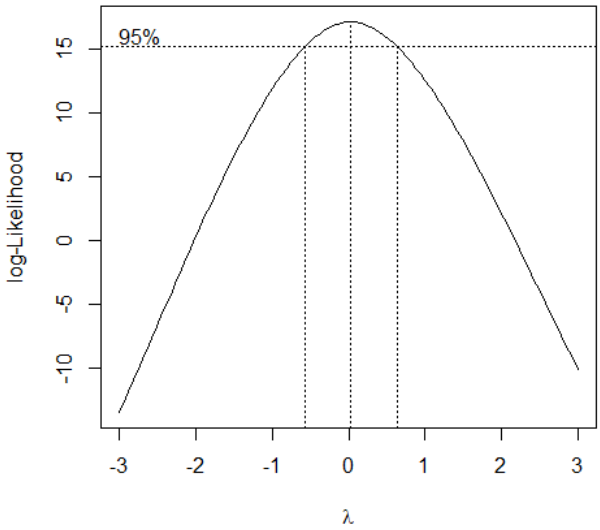
$$WT + QSEC + AM$$

3. **Με χρήση διαγνωστικών τεχνικών, καθώς και με γραφικές παραστάσεις των πρόσθετων μεταβλητών και μερικών υπολοίπων, να εξεταστεί η καταλληλότητα του τελικού μοντέλου, αν χρειάζεται μετασχηματισμούς ή περαιτέρω βελτιώσεις. Εξετάζεται πάλι η πιθανή παρουσία άτυπων σημείων ή σημείων επιρροής. Επίσης να βρεθούν 95% Δ.Ε. για τους συντελεστές του τελικού μοντέλου, καθώς και για τη πρόβλεψη μιας άγνωστης παρατήρησης Υ (με τιμές του x_0 της επιλογής σας) και τέλος να δοθούν ερμηνείες.**

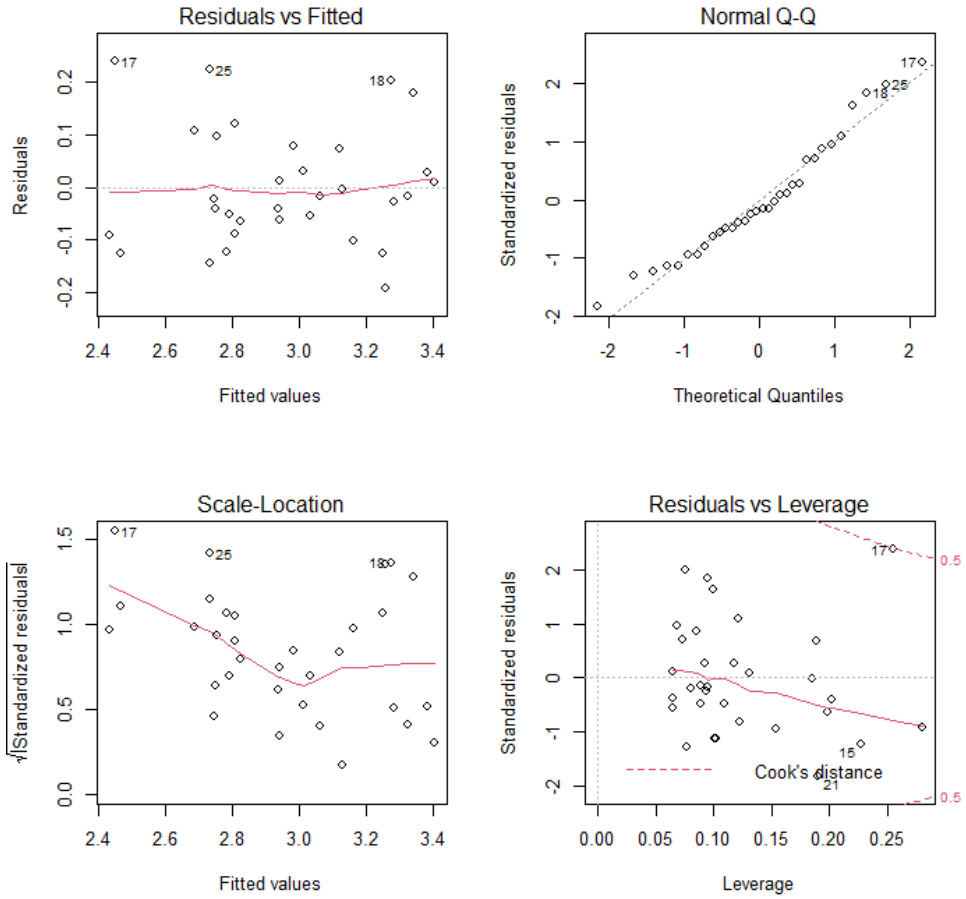




Παρατηρούμε πως τα διαγράμματα που πέρνουμε δεν ειναι ικανοποιητικά. Δηλαδή τα σφάλματα δεν τηρούν την ομοσκεδαστικότητα και το δεύτερο διάγραμμα δείχνει πως δεν ακολουθούν την κανονική κατανομή. Συνεπώς, πραγματοποιούμε μετασχηματισμό box-cox στο μοντέλο μας.



Από το οποίο βλέπουμε πως η πιθανοφάνεια μεγιστοποιείται για λ=0, επομένως θα κάνουμε ένα μετασχηματισμό της μορφής log(mpg).



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.inv)
       wt       cyl        am
3.609011 2.584066 1.924955
```

Στο νέο μοντέλο δεν υπάρχει πολυσυγγραμμικότητα καθώς όλες οι τιμές είναι κάτω του 5.

Σημεία επιρροής Για n=32 , για να ασκεί μια παρατήρηση επιρροή σε κάποια επεξηγηματική μεταβλητή πρέπει να έχει μεγαλύτερι τιμή απο 0.35.

```
> dfb <- dfbetas(mod_bw.inv)
> print(dfb[dfb > 0.35])
[1] 0.3721751 1.3126018 0.3508274 0.5480169 0.6946978
```

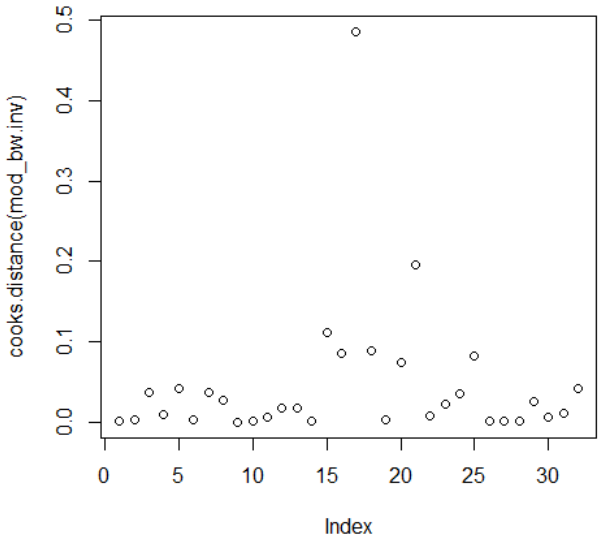
Παρατηρείται ένα σημείο επιρροής στο μοντέλο.

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)
```

Δεν υπάρχει σημείο επιρροής που πληρεί αυτές τις προϋποθέσεις.

Cook's Distance Φαίνεται πως η απόσταση cook δεν βρίσκει κάποιο σημείο επιρροής, καθώς καμία παρατήρηση D_i δεν ξεπερνάει το 1.

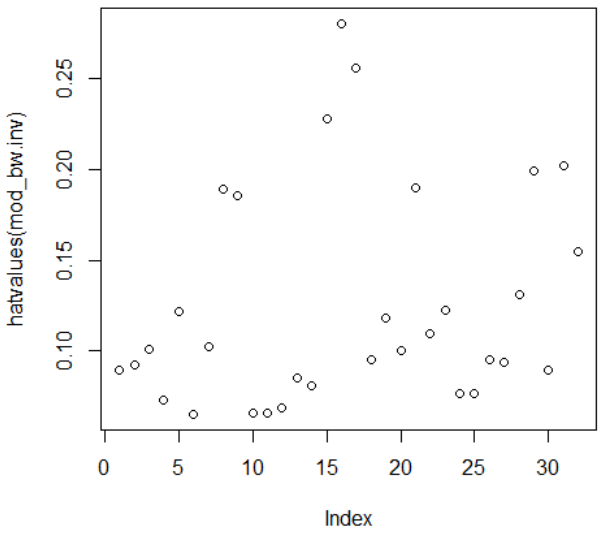


Στην γραφική παράσταση φαίνεται πως υπάρχει ένα σημείο επιρροής, το οποίο μπορεί να επηρεάσει στην προσαρμογή του μοντέλου. Παρόλα αυτά η συνολική εικόνα είναι αρκετά βελτιωμένη.

Hat-values plot Για να θεωρηθεί κάποιο σημείο ως σημείο μόχλευσης πρέπει να ισχύει:

$$h_{ii} > \frac{2p}{n}$$

όπου για n=32 και p=3, $hii > 0.1875$



Παρατηρούνται 3 σημεία που επηρεάζουν το μοντέλο, τα 16, 17, 18. Έχοντας πραγματοποιήσει την ανάλυση που προηγήθηκε, το καλύτερο μοντέλο περιέχει τις μεταβλητές, wt + cyl + am σε συνδιασμό με τον μετασχηματισμό box cox για λ=0,(log(mpg)), ώστε να τηρούνται οι προϋποθέσεις της κανονικότητας του γραμμικού μοντέλου.

Άσκηση Β

Εξετάζεται ο αριθμός παλμών/1" κελαηδήματος Υ δύο ειδών καναρινιών Α και Β σε σχέση με διαφορετικές τιμές θερμοκρασίας 1 x (σε °C). Στόχος της ανάλυσης είναι να εξετάσουμε αν τα δύο είδη καναρινιών διαφοροποιούνται μεταξύ τους ως προς τους ρυθμούς των παλμών. Έστω δείκτρια μεταβλητή x_2 **(** $x_2 = 0$ **-αν ομάδα Β,** $x_2 = 1$ **-αν ομάδα Α).**

1. Περιγράψτε πώς μέσω του μοντέλου παλινδρόμησης $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, μπορούμε να ελέγξουμε αν χρειάζεται να προσαρμοστούν (I) δύο διαφορετικές ευθείες, (II) δύο παράλληλες ευθείες, ή (III) μια κοινή ευθεία και για τα δύο είδη καναρινιών, όπου $x_3 = x_1 x_2$, η μεταβλητή που εκφράζει την αλληλεπίδραση μεταξύ των μεταβλητών x_1 και x_2.

```
adat<-read.table("canary.csv", header = TRUE)
attach(adat)
adat
plot(pulses~Temp, pch=20, col = ifelse( group == 'A', "blue",
    "black")
```

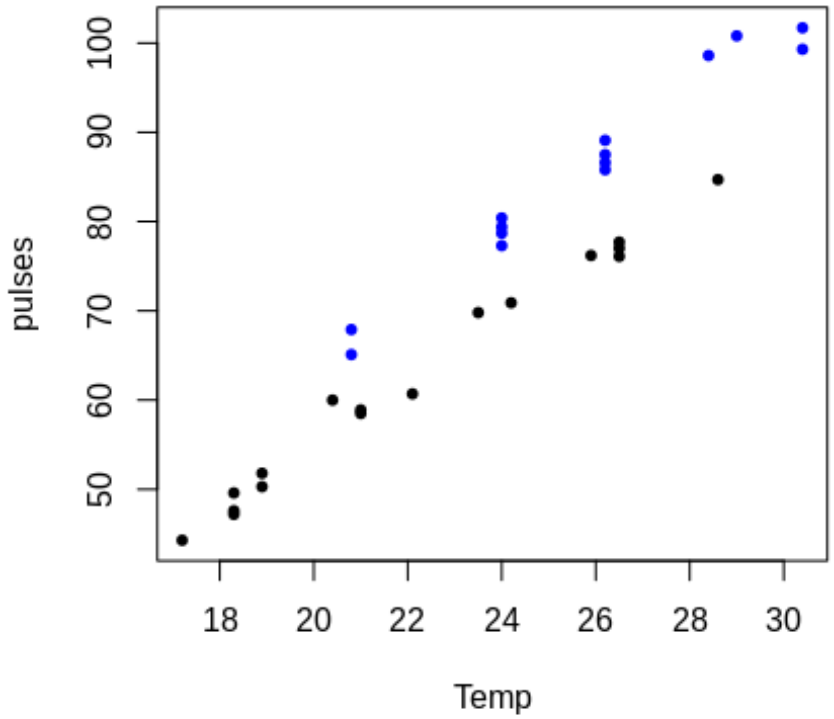


Figure 3: Το διάγραμμα διασποράς όπου με μπλε χρώμα έχουν σχεδιαστεί τα δεδομένα του είδους A και με μαύρο χρώμα τα δεδομένα του είδους B.

Για την επιλογή της καταλληλότερης περίπτωσης στο μοντέλο, αρχικά πρέπει να γίνει ο έλεγχος μηδενικής υπόθεσης $H_0 : \beta_3 = 0$ έναντι της υπόθεσης $H_1 : \beta_3 \neq 0$. Στην περίπτωση που απορριφθεί η υπόθεση H_0 τότε μελετάται η υπόθεση (I). Σε περίπτωση που δεν απορριφθεί η H_0 τότε γίνεται έλεγχος της υπόθεσης $H_0 : \beta_2 = 0$ έναντι της $H_1 : \beta_2 \neq 0$. Σε περίπτωση που απορριφθεί η παραπάνω υπόθεση τότε μελετάται η περίτπωση (II) ενώ αποδοχή της ερμηνεύεται με την περίπτωση (III). Η μεταβλητή $x_3 = x_1x_2$, περιγράφει την αλληλεπίδραση των μεταβλητών x_1, x_2. Στην περίπτωση που το σημείο y_A ανήκει στο A είδος τότε μοντέλο $E(y) = \beta_0 + \beta_1 + x_1 + \beta_2x_2 + \beta_3x_3$ γίνεται της μορφής: $E(y_A) = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x_1$, ενώ στην περίπτωση που το σημείο y_B ανήκει στο B είδος τότε μοντέλο έχει τη μορφή: $E(y_A) = \beta_0 + \beta_1x_1$. Συνεπώς, φαίνεται ότι οι τιμές των συντελεστών β_1 και β_3 καθορίζουν και προσαρμόγοζουν ανάλογα τις ευθείες. (I) αν $\beta_3 \neq 0$, τότε οι συντελεστές του x_1 διαφέρουν στις 2 περιπτώσεις. Εάν (II) $\beta_3 = 0$ και $\beta_2 \neq 0$, τότε η κλίση θα είναι κοινή και ίση με β_1, συνεπώς οι ευθείες είναι παράλληλες. Στην (III) περίπτωση όπου $\beta_3 = \beta_2 = 0$ τότε προκύπτει η ευθεία $E(y_{A)=\beta_0+\beta_2+(\beta_1+\beta_3)x_1}$ για να καλύψει και τα 2 μοντέλα.

2. Να γίνουν αυτοί οι έλεγχοι, η γραφική παράσταση και να δοθούν ερμηνείες για το τελικό μοντέλο.

```
library(data.table)
canary <- fread('canary.csv')
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
model_1 <- lm(Y ~ X_1 + X_2 + X_3, data = cdata)
summary(model_1)
```

```
Call:
lm(formula = Y ~ X_1 + X_2 + X_3, data = cdata)

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
```

```
model_2 <- lm(Y ~ X_1 + X_2, data = cdata)
summary(model_2)
```

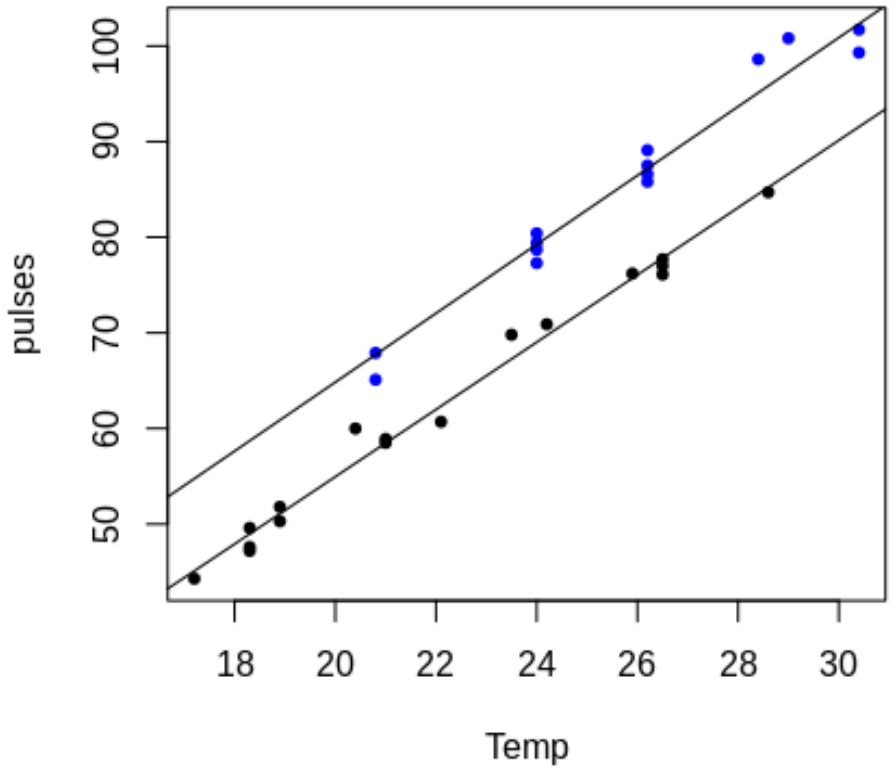
```
Call:
lm(formula = Y ~ X_1 + X_2, data = cdata)

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
```

Έλεγχος F-test:

$$F = \frac{SSE_2 - SSE_1}{SSE_1/(n-k-1)}$$

όπου n = 31, k = 3 και SSE_1 και SSE_2 τα αθροίσματα τετραγώνων των μοντέλων model 1 και model 2 αντίστοιχα.

$$SSE_1 = 85.06687$$
$$SSE_2 = 89.31429$$
$$F = 1.34811$$
$$E(y) = -17.276 + 3.603x_1 + 10.065x_2$$
$$E(y_A) = -7.211 + 3.603x_1$$
$$E(y_B) = -17.276 + 3.603x_1$$



References

[1] Οικονόμου Π. Καρώνη Χ. *Στατιστικά μοντέλα παλινδρόμησης*. 2020.