ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

Στατιστική Μοντελοποίηση

2η Σειρά Ασκήσεων

Ονοματεπώνυμο: Ευάγγελος Τσόγκας

Αριθμός Μητρώου: 03400120

A

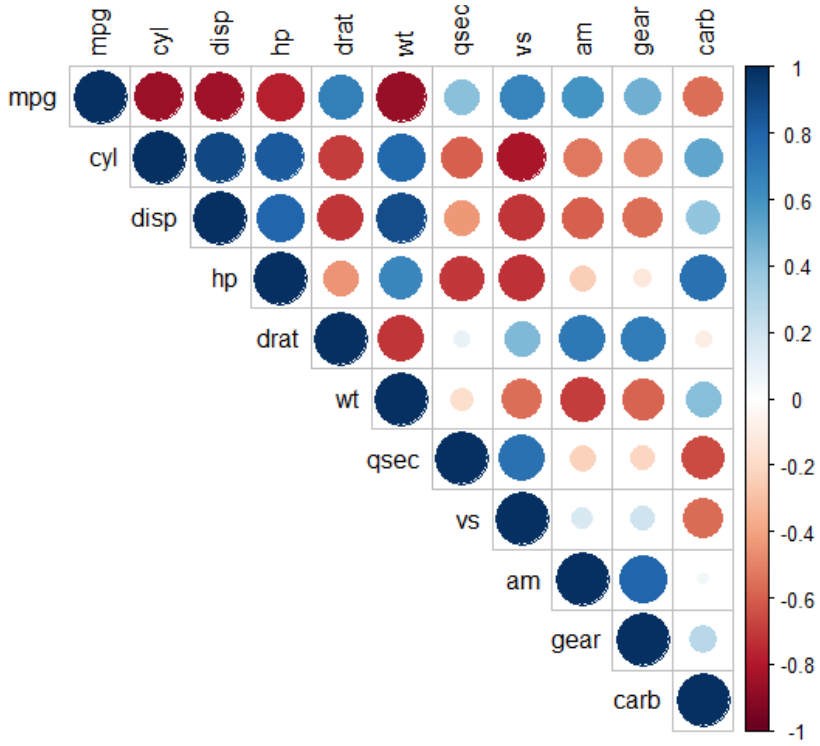
1) Αρχικά προσαρμόζουμε ένα μοντέλο γραμμικής παλινδρόμησης στα δεδομένα παίρνοντας τα εξής αποτελέσματα:

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

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
```

Το μοντέλο φαίνεται πως έχει προσαρμοστεί ικανοποιητικά στα δεδομένα, καθώς η p-value του ελέγχου F είναι πολύ χαμηλή και ο συντελεστής προσδιορισμού είναι R^2 = 86.9%. Παρατηρούμε, όμως, πως υπάρχουν p-values των ελέγχων t των μεταβλητών οι οποίες δεν είναι μικρές. Αυτό σημαίνει πως ίσως κάποιες από αυτές τις μεταβλητές δεν χρειάζονται στο μοντέλο κάτι το οποίο όμως απαιτεί περαιτέρω διερεύνηση.

Εξέταση συσχετίσεων / πολυσυγγραμμικότητας

Στο παρακάτω σχήμα φαίνεται ο πίνακας συσχετίσεων των μεταβλητών από τον οποίο προκύπτει το πρόβλημα πως υπάρχουν αρκετές συσχετίσεις ανάμεσα στις επεξηγηματικές μεταβλητές, ενώ ιδανικά θα θέλαμε να είναι ανεξάρτητες.

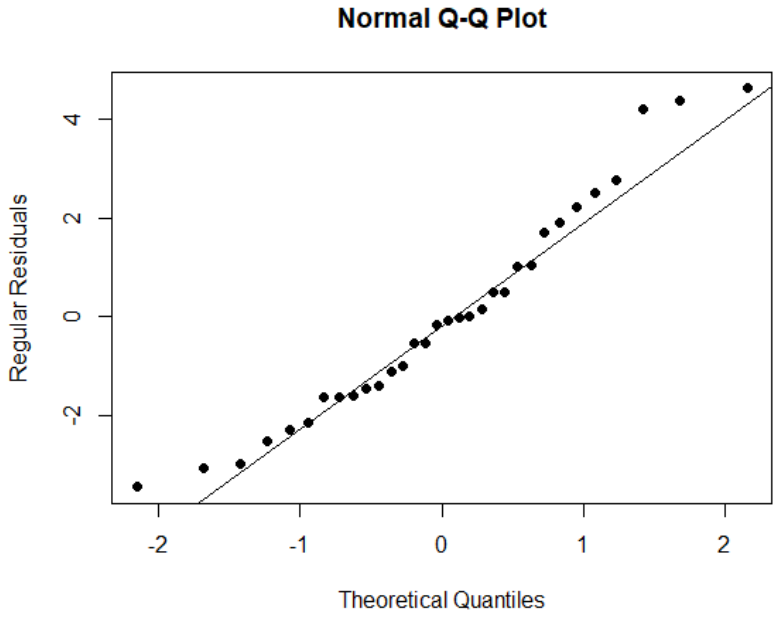


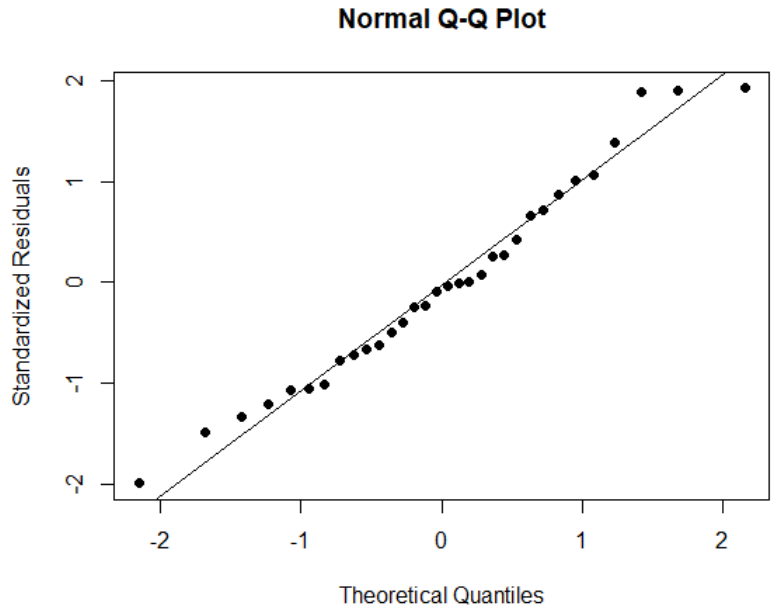
Χρησιμοποιώντας το κριτήριο VIF βλέπουμε πως πράγματι υπάρχει πολυσυγγραμμικότητα, καθώς πολλές τιμές είναι μεγαλύτερες του 5, ειδικά στις περιπτώσεις των μεταβλητών cyl, disp και wt που ξεπερνάει την τιμή 15.

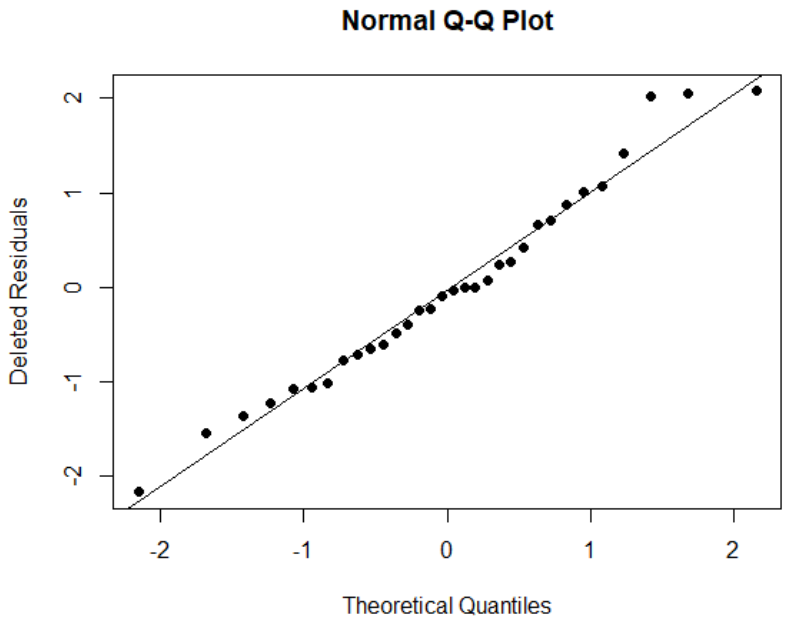
cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
15.373833	21.620241	9.832037	3.374620	15.164887	7.527958	4.965873	4.648487	5.357452	7.908747

Εξέταση υπολοίπων

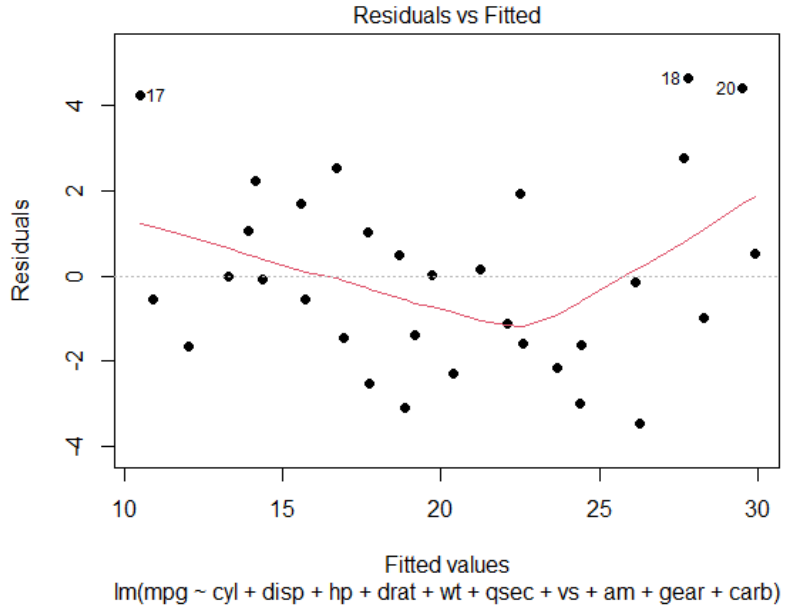
Από τα παρακάτω Normal Q-Q διαγράμματα για τα συνήθη, τυποποιημένα και deleted υπόλοιπα, και προτιμώντας τα δύο τελευταία, παρατηρούμε πως τα σημεία εμφανίζουν μια καλή γραμμικότητα και επομένως συμπεραίνουμε πως η υπόθεση κανονικότητας των υπολοίπων δεν παραβιάζεται αν και υπάρχουν 1-2 ελαφρώς άτυπα σημεία.

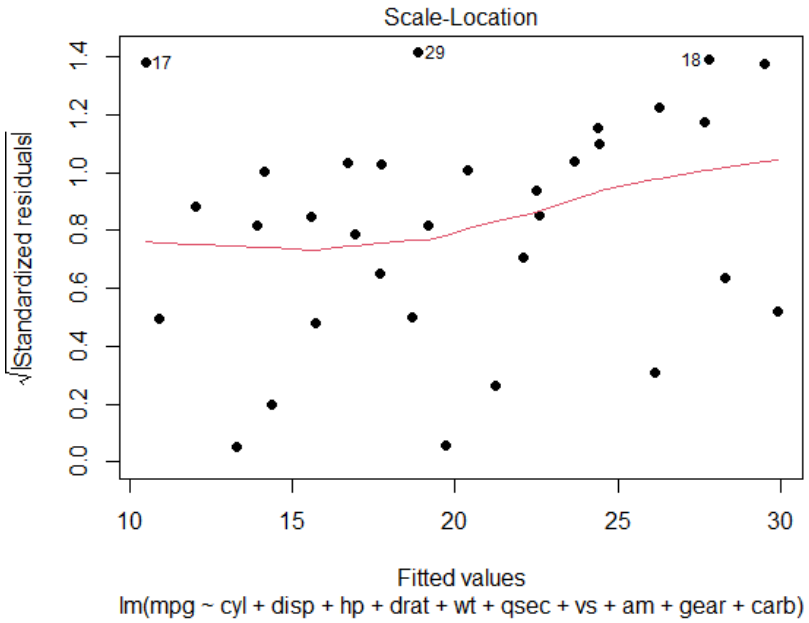






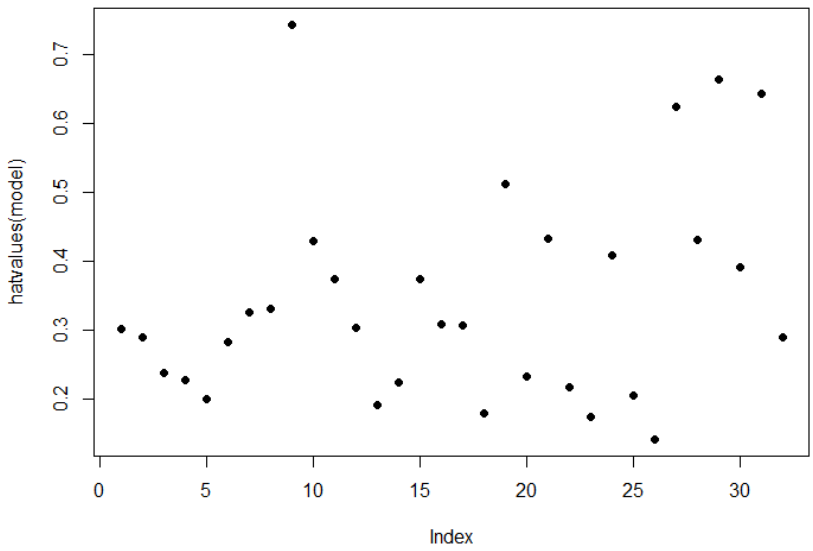
Επίσης, παρατηρώντας τα διαγράμματα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές φαίνεται πως αν και γενικά τηρείται το κριτήριο της ομοσκεδαστικότητας υπάρχουν κάποιες ελαφρώς έκτοπες τιμές και ίσως μια μικρή σχέση μεταξύ των υπολοίπων και των προσαρμοσμένων τιμών.



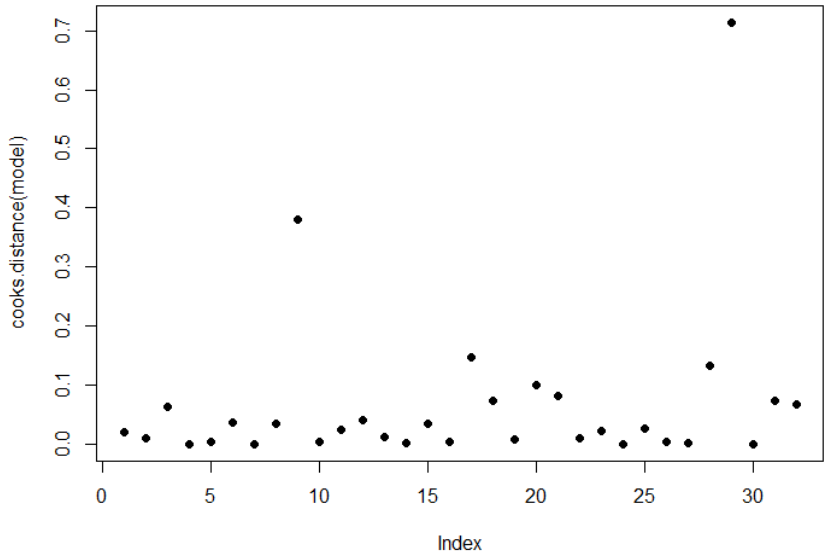


Εξέταση παρουσίας άτυπων σημείων

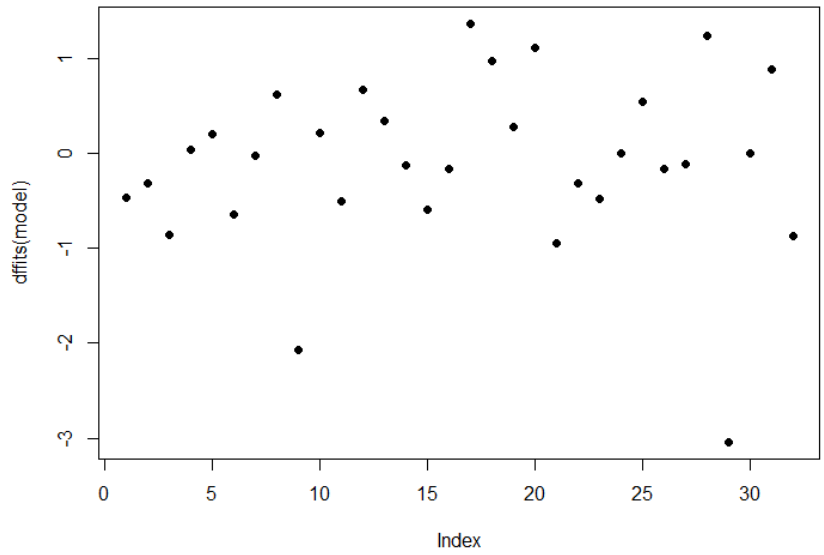
Με βάση το κριτήριο hii, παρατηρούμε τουλάχιστον μια έκτοπη τιμή. Οι μεγαλύτερες τιμές αντιστοιχούν στις παρατηρήσεις 9, 29, 31 και 30.



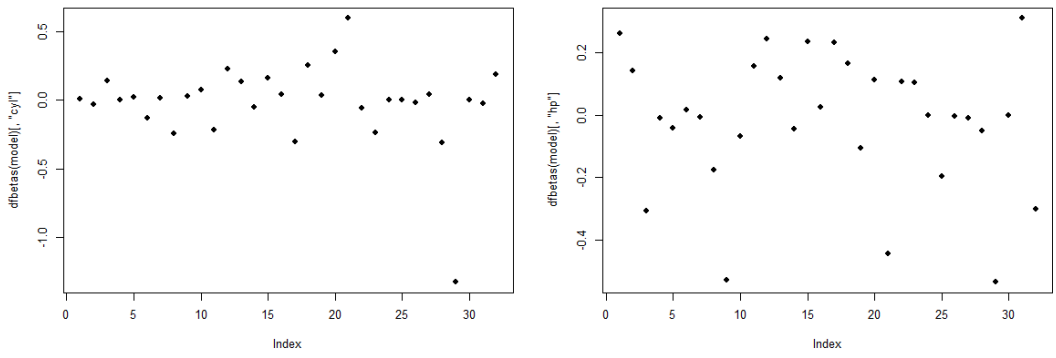
Με βάση το κριτήριο της απόστασης Cook παρατηρούμε δύο έκτοπες τιμές που αντιστοιχούν στις παρατηρήσεις 9 και 29.

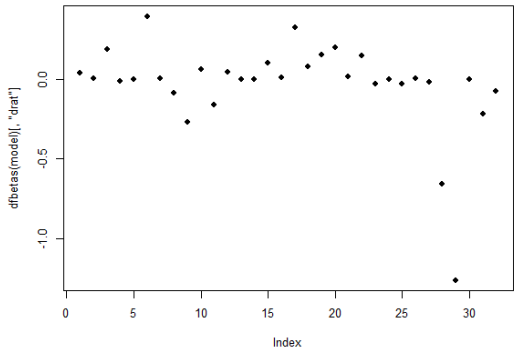


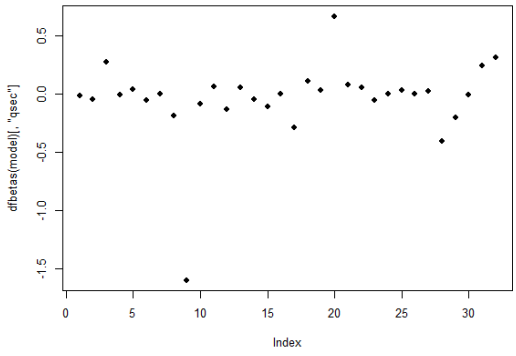
Με βάση το κριτήριο DFFITS και πάλι αναγνωρίζουμε ως έκτοπες παρατηρήσεις τις 9 και 29.

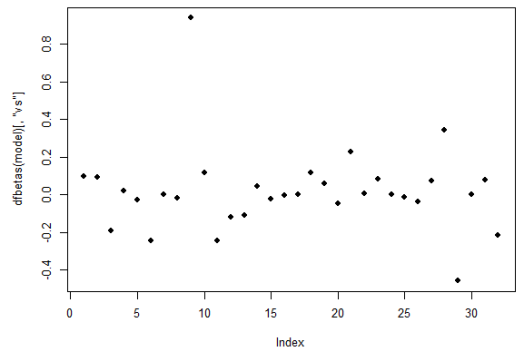


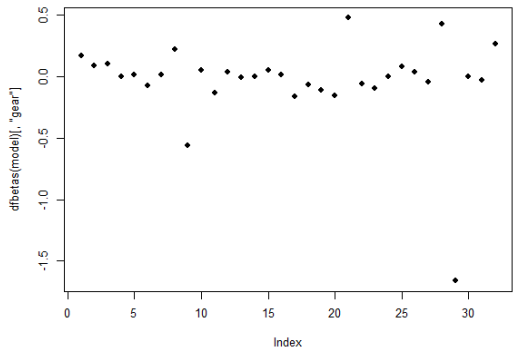
Τέλος, χρησιμοποιώντας το κριτήριο DFBETAS παρατηρούμε ότι στις 7 από τις 10 μεταβλητές (cyl, hp, drat, qsec, vs, gear, carb), όπως φαίνεται στα παρακάτω διαγράμματα, εμφανίζεται σε τουλάχιστον μία από τις δύο αυτές παρατηρήσεις έκτοπη τιμή.

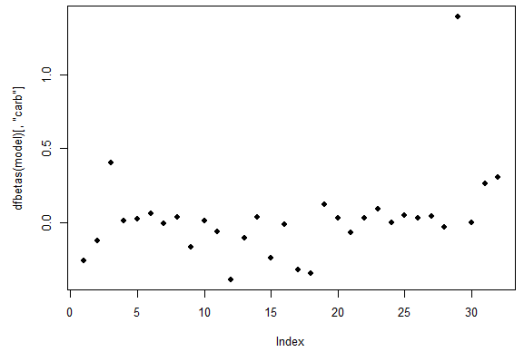












Συμπεραίνουμε, λοιπόν, πως πιθανόν οι παρατηρήσεις 9 και 29 να αποτελούν σημεία επιρροής.

2) Προκειμένου να βρούμε πιο είναι το καλύτερο μοντέλο αρχικά χρησιμοποιούμε τρεις μεθόδους: forward selection, backward elimination και stepwise selection χρησιμοποιώντας ως κριτήριο την τιμή AIC. Τα καλύτερα μοντέλα που προκύπτουν είναι τα εξής:

Forward selection

```
Call:
lm(formula = vehicles$mpg ~ vehicles$wt + vehicles$cyl + vehicles$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 ***
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
```

Backward elimination

```
Call:
lm(formula = mpg ~ wt + qsec + am, data = vehicles)

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
```

Stepwise selection

```
Call:
lm(formula = vehicles$mpg ~ vehicles$wt + vehicles$cyl + vehicles$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 ***
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
```

Παρατηρούμε πως με forward selection και stepwise selection επιλέχθηκαν τα ίδια μοντέλα, ενώ με backward elimination διαφορετικό. Και τα δύο μοντέλα περιλαμβάνουν συνολικά τρεις επεξηγηματικές μεταβλητές και οι έλεγχοι F και t είναι ικανοποιητικοί. Επομένως, θα επιλέξουμε τελικά ένα από αυτά και όχι το αρχικό μοντέλο με όλες τις μεταβλητές. Προκειμένου να επιλέξουμε τελικά ένα από τα δύο θα εξετάσουμε και τα κριτήρια $R^2, \overline{R}^2, R^2_{predict}, C_p$ και AIC.

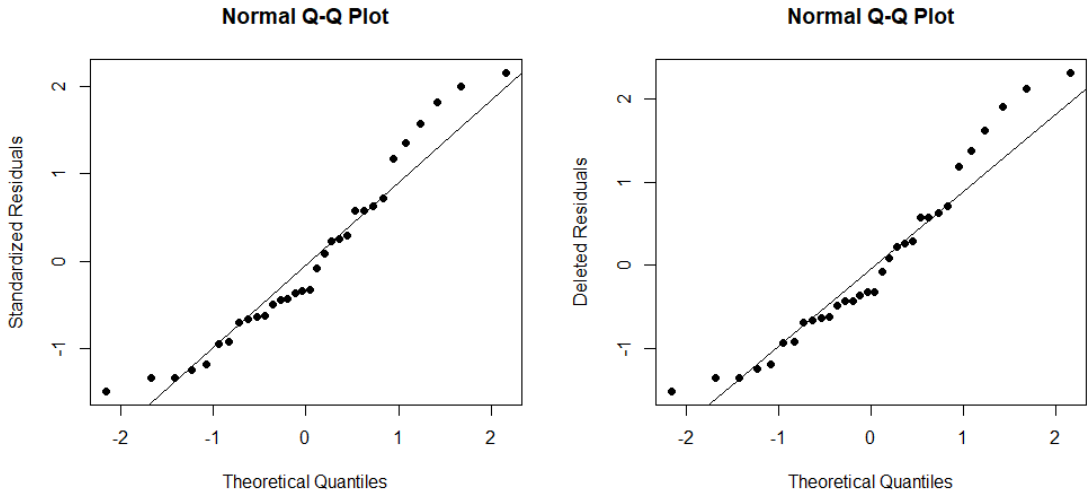
Οι τιμές των μετρικών για τα δύο μοντέλα φαίνονται στον παρακάτω πίνακα.

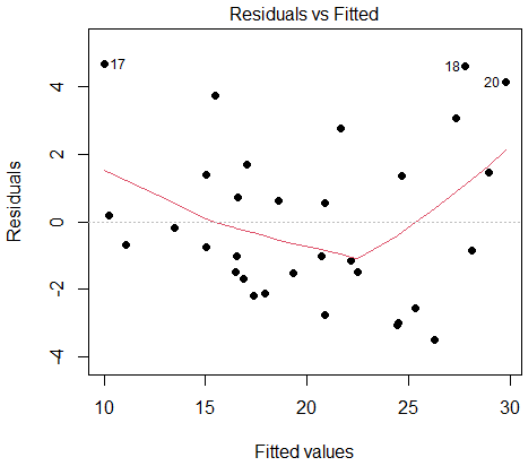
Κριτήρια	**mpg ~ wt + cyl + hp**	**mpg ~ wt + qsec + am**
R^2	0.843	0.85
$\overline{R}^2$	0.826	0.834
$R^2_{predict}$	0.796	0.795
C_p	1.147	0.103
AIC	155.477	154.119

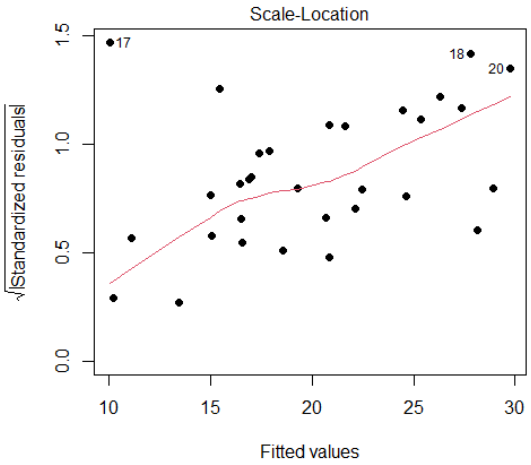
Και με αυτές τις μετρικές τα μοντέλα φαίνεται να έχουν μικρή διαφορά ως προς το πιο είναι καλύτερο. Παρ' όλα αυτά τελικά θα επιλέξουμε το μοντέλο με τις επεξηγηματικές μεταβλητές wt, qsec και am, επειδή έχει καλύτερο συντελεστή προσδιορισμού, καθώς και τον διορθωμένο και μικρότερο AIC.

Επομένως, το <u>καλύτερο μοντέλο</u> είναι το **mpg ~ wt + qsec + am**

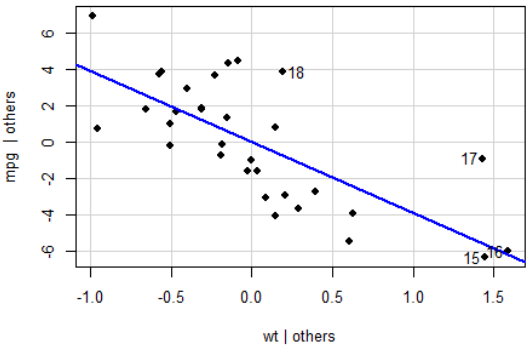
3) Για το μοντέλο που επιλέξαμε θα εξετάσουμε ξανά τις υποθέσεις κανονικότητας και ομοσκεδαστικότητας των υπολοίπων.

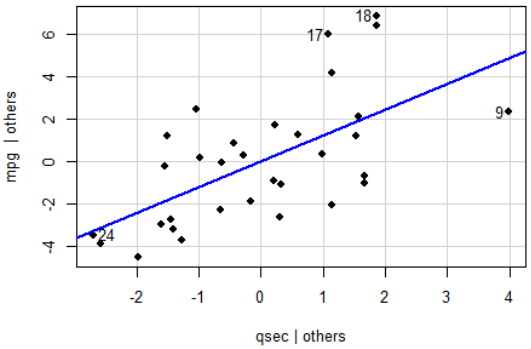


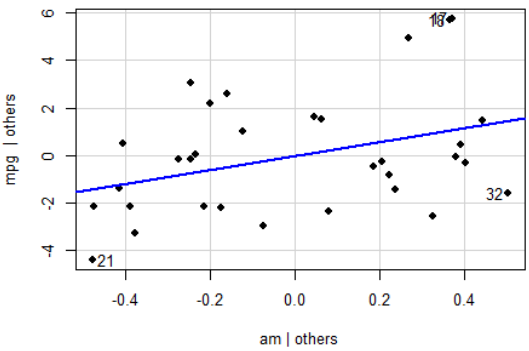




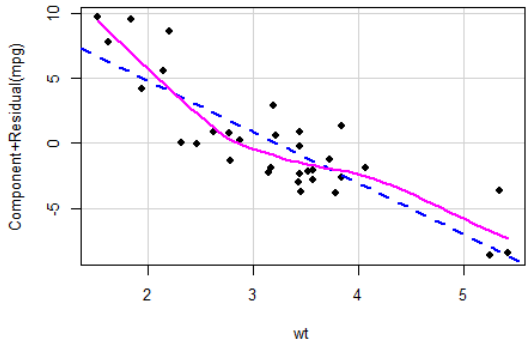
Παρατηρούμε πως και στις δύο περιπτώσεις παρουσιάζεται ένα μικρό πρόβλημα, καθώς τα υπόλοιπα δεν φαίνεται να σχηματίζουν αρκετά καλά ορισμένη ευθεία και επίσης φαίνεται να υπάρχει μια μικρή σχέση με τις προσαρμοσμένες τιμές. Αυτό μας οδηγεί στο συμπέρασμα πως το μοντέλο θα χρειαστεί κάποια βελτίωση. Παρακάτω φαίνονται και οι γραφικές παραστάσεις πρόσθετων μεταβλητών και μερικών υπολοίπων αντίστοιχα.

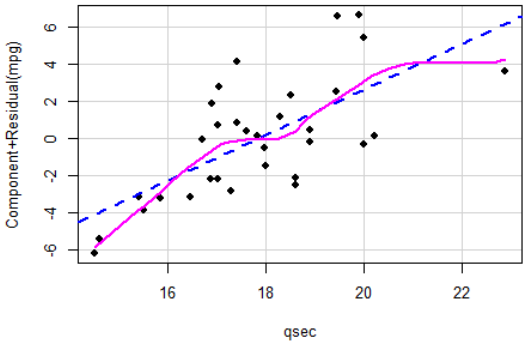


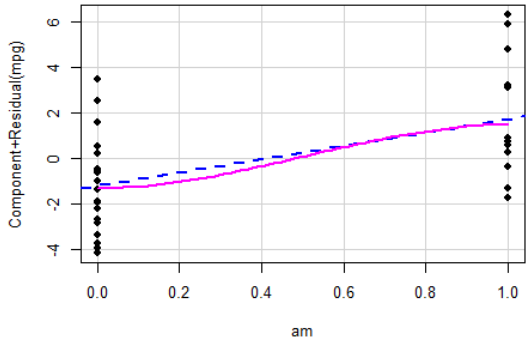




Από το διάγραμμα πρόσθετων μεταβλητών αν και σε καμία μεταβλητή δεν σχηματίζεται αρκετά καλά ορισμένη ευθεία παρατηρούμε πως υπάρχουν συσχετίσεις, επομένως συμπεραίνουμε πως οι μεταβλητές wt και qsec χρειάζονται στο μοντέλο. Μόνο η μεταβλητή am φαίνεται να μην συνεισφέρει αρκετά, επομένως ίσως την αφαιρέσουμε.







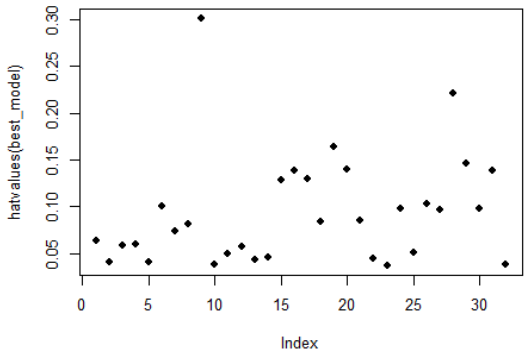
Και από τα διαγράμματα μερικών υπολοίπων συμπεραίνουμε πως οι μεταβλητές wt και qsec μπορούν να παραμείνουν στο μοντέλο, αλλά η μεταβλητή am θα αφαιρεθεί, αφού όπως έχει παρατηρηθεί και από τις μεθόδους επιλογής καλύτερου μοντέλου δεν συνεισφέρει αρκετά στη βελτίωση του μοντέλου. Παρ' όλα αυτά παρατηρούμε κάποιες μικρές μη γραμμικότητες. Μετασχηματίζοντας τη μεταβλητή qsec παίρνοντας το τετράγωνο ή τον λογάριθμο δεν φαίνεται να βελτιώνεται το μοντέλο, αλλά στην περίπτωση της μεταβλητής **wt λογαριθμίζοντάς** την, το μοντέλο βελτιώθηκε, όπως φαίνεται παρακάτω.

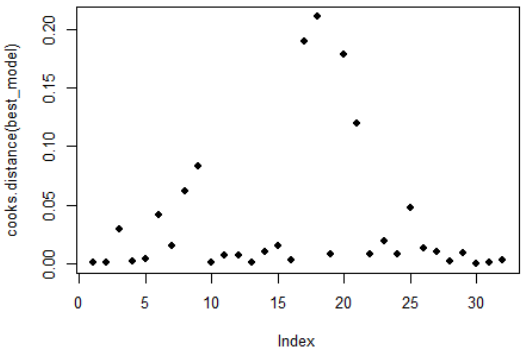
```
Call:
lm(formula = mpg ~ wt_log + qsec, data = vehicles)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0729 -1.3876 -0.4368  0.7493  5.4694

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.2967     4.4603   4.999 2.54e-05 ***
wt_log      -16.1783     1.2519 -12.923 1.47e-13 ***
qsec          0.8932     0.2224   4.016 0.000384 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.177 on 29 degrees of freedom
Multiple R-squared:  0.878,     Adjusted R-squared:  0.8696
F-statistic: 104.3 on 2 and 29 DF,  p-value: 5.661e-14

> AIC(model_new)
[1] 145.4393
```

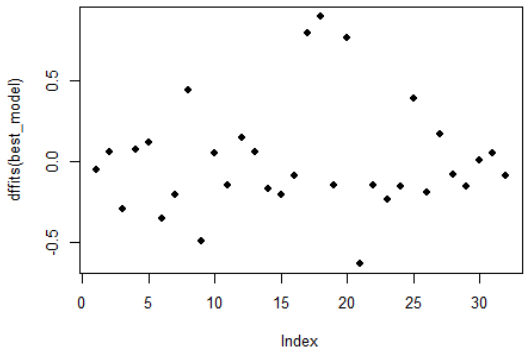
Βλέπουμε πως έχουν βελτιωθεί οι συντελεστές προσδιορισμού, οι έλεγχοι F και t και επίσης η τιμή AIC. Επομένως, ως καλύτερο μοντέλο με τις τελικές βελτιώσεις επιλέγουμε το

mpg ~ wt_log + qsec

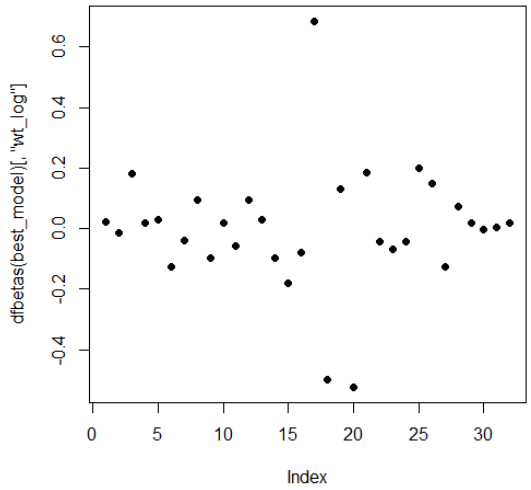
Στη συνέχεια εξετάζουμε ξανά για ύπαρξη άτυπων σημείων.

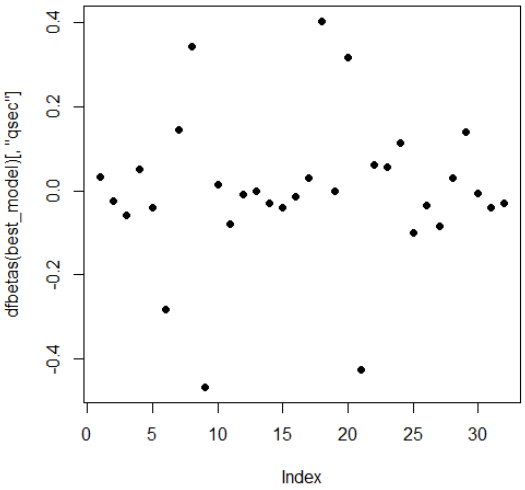






Με βάση τα hii παρουσιάζονται έκτοπες τιμές στις παρατηρήσεις 9 και 28, ενώ με βάση την απόσταση cook και το κριτήριο DFFITS παρουσιάζονται έκτοπες τιμές στις παρατηρήσεις 17, 18, 20 και 21. Επίσης, όπως φαίνεται παρακάτω με το κριτήριο DFBETAS για την μεταβλητή wt_log έχουμε έκτοπες τιμές στις παρατηρήσεις 17, 18 και 20, ενώ για τη μεταβλητή qsec σε πολλαπλές παρατηρήσεις, μεταξύ τους και οι 9, 18, 20 και 21. Κάποιες από αυτές τις παρατηρήσεις ενδεχομένως να είναι σημεία επιρροής.





Διαστήματα εμπιστοσύνης

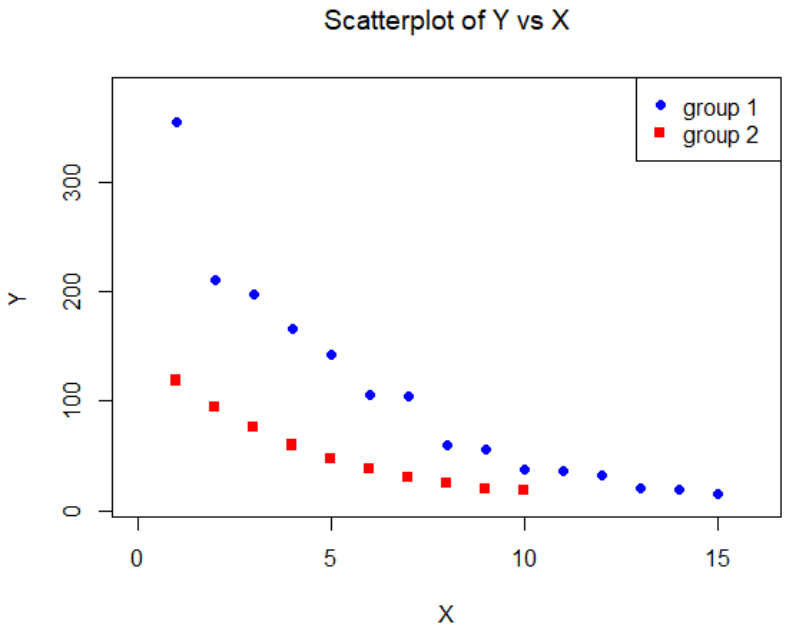
Παρατηρούμε πως δεν περιέχεται το 0 στα διαστήματα εμπιστοσύνης, το οποίο συμφωνεί με του στατιστικούς ελέγχους t, των οποίων τα p-values είναι πολύ μικρά και άρα απορρίπτουμε τις υποθέσεις H_0 για τις μεταβλητές wt_log και qsec.

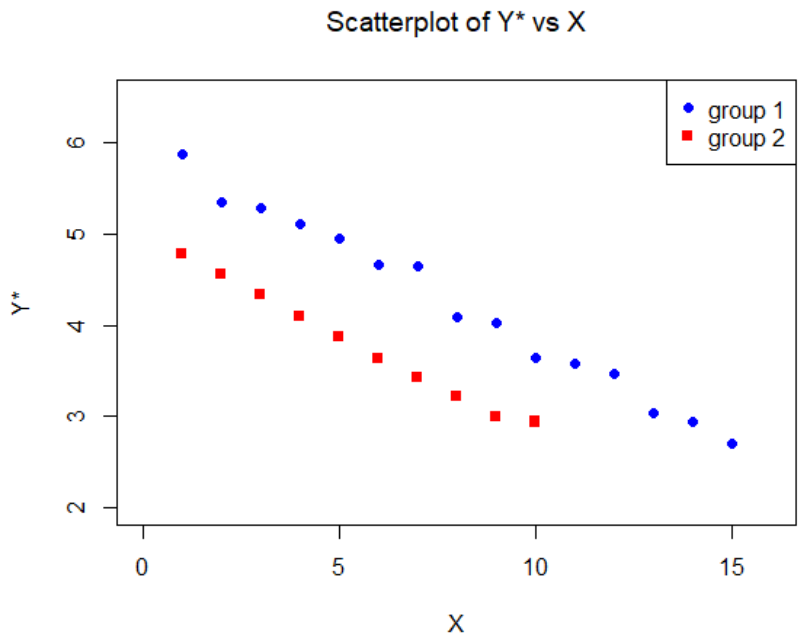
	2.5 %	97.5 %
(Intercept)	13.1743858	31.418988
wt_log	-18.7386729	-13.617999
qsec	0.4382485	1.348074

Επίσης, για μια καινούρια παρατήρηση όπου wt_log = 0.5 και qsec = 18, το διάστημα εμπιστοσύνης για την πρόβλεψη της μέση τιμής mpg είναι (28.519, 32.05) και το αντίστοιχο για την τιμή της είναι (25.496, 35.074).

B

1) Στα παρακάτω διαγράμματα φαίνονται τα scatter plots για τις δύο ομάδες βακτηρίων πριν και μετά το μετασχηματισμό της μεταβλητής Y, αντίστοιχα.





Είναι εμφανές πως χρειάζεται ο μετασχηματισμός Y* = log(Y) προκειμένου να παρουσιαστεί γραμμική σχέση μεταξύ των μεταβλητών Y και X. Παρατηρούμε επίσης, πως δεν αρκεί μια ευθεία για να προσαρμοστεί στα δεδομένα των δύο group.

2) Έστω ότι έχουμε μια εξαρτημένη μεταβλητή y και μια επεξηγηματική μεταβλητή x1. Αν έχουμε μια δείκτρια μεταβλητή Z (ή x2) η οποία παίρνει την τιμή 1 για μια ομάδα δεδομένων και την τιμή 0 για μια δεύτερη ομάδα δεδομένων, κατασκευάζουμε μια τρίτη μεταβλητή x3 που αποτελεί το γινόμενο x1*x2. Έτσι ορίζουμε ένα γενικό γραμμικό μοντέλο y = β0 + β1x1 + β2x2 +β3x3 και μπορούμε να διακρίνουμε τρεις καταστάσεις:

- (Α) Απορρίπτεται η υπόθεση β3=0 και άρα οι ομάδες πρέπει να περιγραφθούν με δύο διαφορετικές ευθείες.
- (Β) β3=0 και δεν απορρίπτεται η υπόθεση β2=0, άρα οι ομάδες περιγράφονται με δύο παράλληλες ευθείες.
- (Γ) β3=0 και β2=0, άρα και οι δύο ομάδες περιγράφονται από μια μόνο ευθεία.

Αρχίζοντας με το γενικό μοντέλο με τις μεταβλητές x1, x2, x3 (κατάσταση Α) και εφαρμόζοντας backward elimination συγκρίνουμε καταστάσεις είτε με t-test είτε με τη ελεγχοσυνάρτηση F που σε αυτήν την περίπτωση είναι ισοδύναμα. Αν βρεθούμε σε κάποια κατάσταση όπου η p-value < 0.001, δηλαδή η υπόθεση H0 για τον αντίστοιχο συντελεστή απορρίπτεται, τότε παραμένουμε σε εκείνη την κατάσταση.

3) Στην περίπτωση των δεδομένων που διαθέτουμε η μεταβλητή group χωρίζει τα δεδομένα σε δύο ομάδες, αλλά παίρνει τις τιμές 1 και 2, επομένως αντίστοιχα τις μετατρέπουμε σε 0 και 1, ώστε να τη χρησιμοποιήσουμε ως δείκτρια μεταβλητή. Έτσι ορίζουμε τρεις μεταβλητές: τη x1 που αντιστοιχεί στην αρχική X, τη x2 που αντιστοιχεί στη δείκτρια και τη x3 = x1*x2.

Αρχικά, εφαρμόζουμε backward elimination ξεκινώντας από την κατάσταση Α και ελέγχοντας τα t-test.

```
Call:
lm(formula = log_Y ~ x1 + x2 + x3, data = bacteria)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18445 -0.03468  0.01127  0.02923  0.20021

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.973160   0.050229 118.918  < 2e-16 ***
x1          -0.218425   0.005524 -39.538  < 2e-16 ***
x2          -1.014065   0.080690 -12.567 3.09e-11 ***
x3           0.005133   0.011580   0.443    0.662
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09244 on 21 degrees of freedom
Multiple R-squared:  0.9903,    Adjusted R-squared:  0.9889
F-statistic: 712.7 on 3 and 21 DF,  p-value: < 2.2e-16
```

Παρατηρούμε πως η p-value για τον έλεγχο t του συντελεστή της x3 είναι μεγαλύτερη από 0.001, άρα συμπεραίνουμε ότι β3=0 και προχωράμε στην κατάσταση Β.

```
Call:
lm(formula = log_Y ~ x1 + x2, data = bacteria)

Residuals:
      Min        1Q    Median        3Q       Max
-0.177442 -0.028839  0.003091  0.017202  0.201376

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.963814   0.044750  133.27   <2e-16 ***
x1          -0.217257   0.004766  -45.59   <2e-16 ***
x2          -0.982911   0.038913  -25.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09074 on 22 degrees of freedom
Multiple R-squared:  0.9902,    Adjusted R-squared:  0.9893
F-statistic:  1110 on 2 and 22 DF,  p-value: < 2.2e-16
```

Σε αυτή την περίπτωση βλέπουμε πως η p-value για τον έλεγχο t του συντελεστή της x2 είναι μικρότερη από 0.001, επομένως απορρίπτουμε την υπόθεση H0: β2=0, και παραμένουμε σε αυτή την κατάσταση, δηλαδή, συμπεραίνουμε ότι οι δύο ομάδες δεδομένων μπορούν να περιγραφούν από δύο παράλληλες ευθείες.

Είναι φανερό πως αποτελεί το καλύτερο μοντέλο, καθώς αν προχωρήσουμε στην κατάσταση Γ βλέπουμε πως χαλάνε πολύ οι τιμές του συντελεστή προσδιορισμού.

```
Call:
lm(formula = log_Y ~ x1, data = bacteria)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6933 -0.4967  0.1289  0.4000  0.7399

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3126     0.1960  27.113  < 2e-16 ***
x1           -0.1804     0.0243  -7.423 1.51e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4861 on 23 degrees of freedom
Multiple R-squared:  0.7055,    Adjusted R-squared:  0.6927
F-statistic: 55.09 on 1 and 23 DF,  p-value: 1.513e-07
```

Τελικά πράγματι, αν προσαρμόσουμε δύο ευθείες στις δύο ομάδες δεδομένων φαίνονται παράλληλες και περιγράφουν πολύ καλά τις τιμές της εξαρτημένης μεταβλητής.