ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Ε.ΔΕ.Μ.Μ

ΜΑΘΗΜΑ: ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ
ΧΕΙΜΕΡΙΝΟ ΕΞΑΜΗΝΟ 2021-2022

2η ΣΕΙΡΑ ΑΣΚΗΣΕΩΝ

[redacted]

1 Άσκηση 1

1.1 Ερώτημα 1

Η άσκηση αυτή εστιάζει στην προσαρμογή ενός μοντέλου πολλαπλής γραμμικής παλινδρόμησης (εξισώσεις 1 και 2) σε δεδομένα χαρακτηριστικών ($\mathbf{X}$) και απόδοσης ($\mathbf{y}$) ποικίλων τύπων αυτοκινήτων. Στο συγκεκριμένο πρόβλημα θα εξεταστεί η κατανάλωση βενζίνης των συγκεκριμένων αυτοκινήτων σε μονάδες μιλίων μετακίνησης προς τα αντίστοιχα γαλόνια βενζίνης προς κατανάλωση (miles/gallon) ως προς συγκεκριμένα τεχνικά χαρακτηριστικά των εξαρτημάτων των αυτοκινήτων. Στον Πίνακα 1 αναφέρονται αυτά τα χαρακτηριστικά, τα οποία είναι και οι μεταβλητές στις οποίες θα προσαρμοστεί το γραμμικό μοντέλο με στόχο την πρόβλεψη της μεταβλητής ***mpg***, δηλαδή τα miles/gallon.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad \epsilon_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i \tag{1}$$

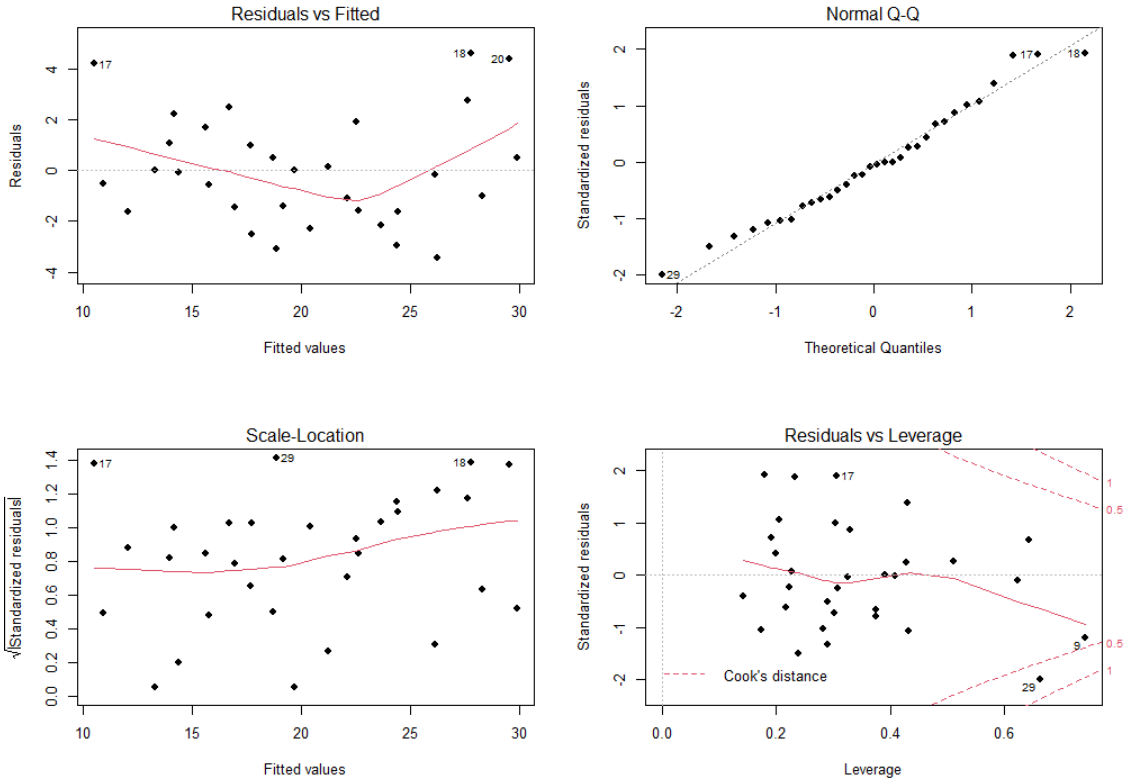
$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y} \tag{2}$$

Όνομα Μεταβλητής	Περιγραφή
cyl	Αριθμός κυλίνδρων
disp	Μετατόπιση (Displacement) (cu.in.)
hp	Μικτή ιπποδύναμη (Gross horsepower)
drat	Αναλογία οπίσθιου άξονα (Rear axle ratio)
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 (forward gears)
carb	Αριθμός καρμπυρατέρ

Πίνακας 1: Μεταβλητές προς προσαρμογή

Αρχικά, προσαρμόζεται το πολλαπλό γραμμικό μοντέλο για την μεταβλητή mpg και ως προς τις 10 επεξηγηματικές μεταβλητές που προαναφέρθηκαν. Εξετάζοντας πρώτα το Q-Q plot κανονικής κατανομής για το συγκεκριμένο μοντέλο, το οποίο παρουσιάζεται στην Εικόνα 1 των διαγνωστικών ελέγχων υπολοίπων, παρατηρείται πως πράγματι τηρούνται οι υποθέσεις κανονικής κατανομής για τα τυχαία σφάλματα, καθώς οι προσαρμοσμένες τιμές των ποσοστιαίων σημείων ακολουθούν τα αντίστοιχα θεωρητικά. Ωστόσο, το συγκεκριμένο μοντέλο δεν είναι και το βέλτιστο είτε λόγο πιθανών outliers ειτε λόγω πολυσυγγραμικότητας. Ειδικότερα, εάν εξεταστούν τα υπόλοιπα ως προς τις προσαρμοσμένες τιμές εμφανίζεται μια ένδειξη καμπυλοειδούς συμπεριφοράς των υπολοίπων το οποίο μπορεί να οφείλεται είτε σε πιθανά outliers (π.χ. σημεία 17, 18 και 20) είτε στο ότι κάποια μεταβλητή πρέπει να υποστεί κατάλληλο μετασχηματισμό. Επίσης, υπόδειξη πιθανών outliers αποτελεί και το διάγραμμα των υπολοίπων ως προς το επίπεδο της μόχλευσης (leverage), στο οποίο παρατηρούνται δυο σημεία που βρίσκονται στα όρια της απόστασης Cook.

Η πρώτη αριθμητική ένδειξη ακαταλληλότητας είναι εμφανής εάν εξεταστούν τα επίπεδα σημαντικότητας των συντελεστών $\hat{\beta}_i$ σύμφωνα με τον έλεγχο t. Ειδικότερα, οι τιμές τις p-value για την πλειοψηφία των συντελεστών εμφανίζεται στο εύρος 0.5 - 0.9, το οποίο δεν είναι στατιστικά σημαντικό για να



Εικόνα 1: Residuals diagnostics plots

αποδεχθούμε και τις 10 μεταβλητές εξαιτίας πιθανών συσχετίσεων.Ομως, ο έλεγχος αυτός δεν είναι ενδεικτικός για να αξιολογηθεί επαρκώς το μοντέλο. Για την αριθμητική προσέγγιση της έκτασης της πολυσυγγραμικότητας του μοντέλου θα αξιοποιηθεί το κριτήριο *VIF(Variance Inflation Factor)*:

$$VIF = \frac{1}{1 - R_j^2} \tag{3}$$

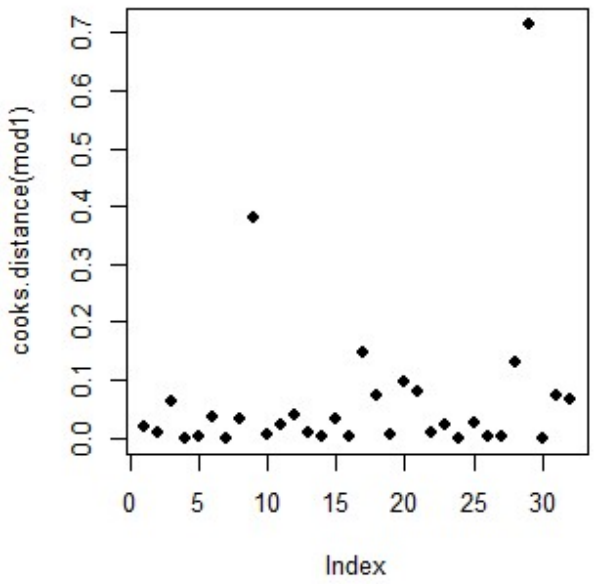
όπου R_j^2 ο συντελεστής προσδιορισμού της παλινδρόμησης της j-οστής μεταβλητής. Το κριτήριο αυτό πιστοποιεί την αύξηση της διασποράς ενός εκτιμημένου συντελεστή $\hat{\beta}_i$ σε περίπτωση ύπαρξης συσχετίσεων μεταξύ των επεξηγηματικών μεταβλητών. Στον Πίνακα 2 παρουσιάζονται οι τιμές του VIF για τις μεταβλητές του μοντέλου. Για τιμές $VIF > 5$ είναι έντονη η πολυσυγραμικότητα, γεγονός το οποίο ισχύει για την πλειοψηφία των μεταβλητών. Συνεπώς, θα πρέπει να πραγματοποιηθεί έλεγχος σε επιπροσθετα κριτήρια για να αφαιρεθούν συσχετίζόμενες μεταβλητές.

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

Πίνακας 2: Τιμές Κριτηρίου VIF

Στη συνέχεια, θα αναζητηθούν οι παρατηρήσεις οι οποίες επηρεάζουν έντονα (influence points) τις τιμές των συντελεστών και αποτελούν πιθανά outliers για το μοντέλο που προσαρμόζεται. Αρχικά

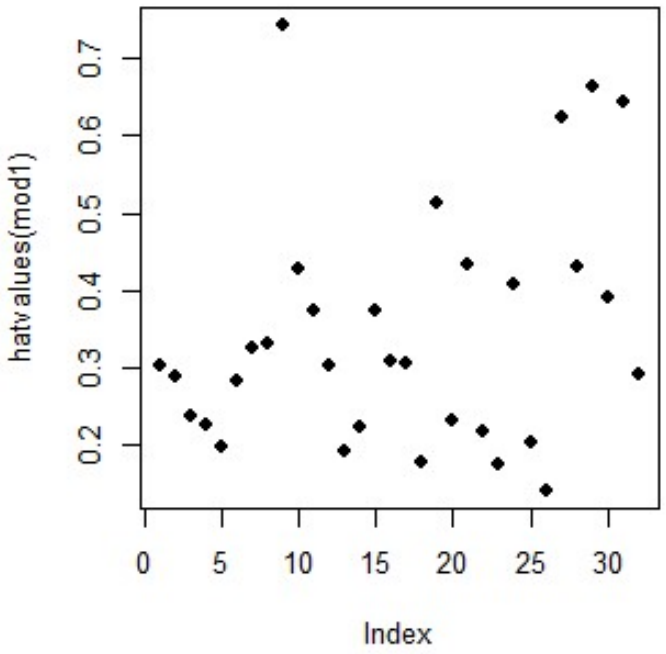
εξετάζονται οι τιμές της απόστασης Cook για την κάθε παρατήρηση, η οποία ορίζει την "απόσταση" μεταξύ των συντελεστών του μοντέλου λαμβάνοντας υπόψη όλες τις παρατηρήσεις, με τους αντίστοιχους συντελεστές αν αφαιρεθεί η i-οστή παρατήρηση. Ειδικότερα, εάν εμφανιστούν τιμές της απόστασης Cook πολύ μεγαλύτερες του 1, τότε το σημείο αποτελεί σημείο επιρροής του μοντέλου παλινδρόμησης. Στην Εικόνα 2 παρουσιάζονται οι αντιστοιχές τιμές και παρατηρείται πως το σύνολο τους έχει απόσταση Cook περίπου μικρότερη του 0.7 και συνεπώς πιθανώς δεν υπάρχουν σημεία έντονης επιρροής για το μοντέλο αυτό. Ωστόσο, οι παρατηρήσεις 17 και 29 επηρεάζουν ισχυρότερα τους συντελεστές σε σχέση με τις υπόλοιπες καθώς οι τιμές αποστάσεων Cook παρουσιάζουν απόκλιση.



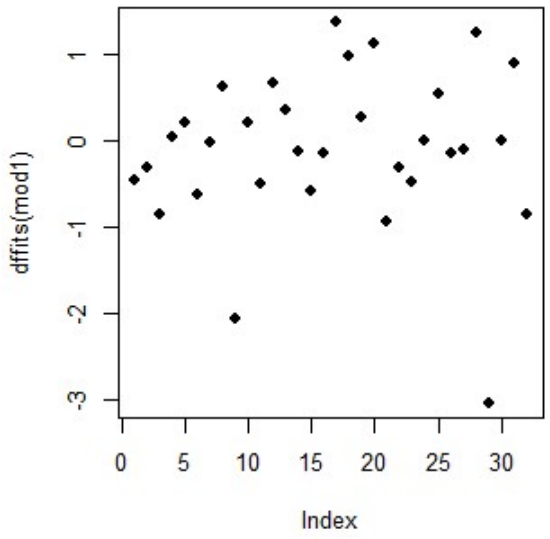
Εικόνα 2: Απόσταση Cook για το σύνολο των παρατηρήσεων του δείγματος

Το επόμενο κριτήριο για τα σημεία επιρροής που θα εξεταστεί αφορά τις τιμές h_{ii}, οι οποίες αποτελούν τα διαγώνια στοιχεία του πίνακα H (εξίσωση 2). Οι τιμές αυτές αποτελούν ένα μέτρο του πόσο απομακρυσμένη είναι η κάθε παρατήρηση από την μέση τιμή της κάθε μεταβλητής, γεγονός το οποίο αποτελεί ένδειξη πιθανού σημείου επιρροής. Οι παρατηρήσεις για τις οποίες ισχύει ότι $h_{ii} > 2p/n = 0.6875$ όπου p=11 ο αριθμός παραμέτρων του μοντέλου και n=32 ο αριθμός των παρατηρήσεων θεωρούνται πιθανά σημεία επιρροής. Σύμφωνα με την Εικόνα 3, το σημείο που ξεπερνά ελάχιστα αυτό το κατώφλι είναι η 9η παρατήρηση με $h_{99} = 0.7422870$, η οποία μπορεί να θεωρηθεί πιθανό σημείο επιρροής.

Το τελευταίο κριτήριο που θα εξεταστεί για πιθανά σημεία επιρροής είναι το $DFFITS_i$, το οποίο εξετάζει την επιρροή της i-οστής παρατηρήσης στην προβλεπόμενη τιμή για την mpg μεταβλητή. Εάν $|DFFITS_i| > 2\sqrt{p/n} = 1.173$ τότε το σημείο μπορεί να θεωρηθεί σημείο επιρροής για το συγκεκριμένο μοντέλο. Όπως παρουσιάζεται στην Εικόνα 4 και στα αποτελέσματα του κώδικα που υπάρχουν στο τέλος της εργασίας, οι παρατηρήσεις που ξεπερνούν αυτό το κατώφλι είναι η 9η, η 17η, η 28η και η 29η. Ειδικότερα, η 17η και η 29η παρατήρηση επηρεάζουν ισχυρότερα την πρόβλεψη καθώς, όπως προαναφέρθηκε, παρουσιάζουν και υψηλότερη απόσταση Cook, αν και όχι μεγαλύτερη του 1. Ωστόσο, η 9η παρατήρηση συμφωνεί πλήρως και με το κριτήριο των h_{ii}.



Εικόνα 3: Hat values για το σύνολο των παρατηρήσεων του δείγματος



Εικόνα 4: Τιμές κριτηρίου DFFITS για το σύνολο των παρατηρήσεων του δείγματος

Συμπερασματικά, παρατηρείται πως ένα σημαντικό σημείο επιρροής είναι η 9η παρατήρηση, η οποία προκύπτει τόσο από το κριτήριο των h_{ii} όσο και από το κριτήριο $|DFFITS_i|$. Επιπλέον, σημεία επιρροής αποτελούν και οι παρατηρήσεις 17, 28 και 29, οι οποίες προκύπτουν μόνο από το κριτήριο $|DFFITS_i|$.

1.2 Ερώτημα 2

Το μοντέλο με τις 10 επεξηγηματικές μεταβλητές περιλαμβάνει υψηλές συσχετίσεις μεταξύ των μεταβλητών και συνεπώς θα πρέπει να εφαρμοστεί κατάλληλη μείωσή τους προκειμένου να γίνει πιο αποδοτικό. Για να επιλεχθεί το κατάλληλο μοντέλο θα εφαρμοστούν τεχνικές βηματικής αξιολόγησης μοντέλων που περιέχουν συνδυασμούς μεταβλητών, οι οποίοι θα αντιστοιχούν σε διαφορετικές τιμές πιθανοφάνειας καθώς επίσης και άλλων στατιστικών μέτρων απόδοσης. Τα μέτρα αυτά βασίζονται στην ελαχιστοποίηση του Αθροίσματος των Τετραγωνικών Σφαλμάτων (SSE) και συνεπώς στον στατιστικό έλεγχο F, ο οποίος δίνεται από τη σχέση:

$$F = \frac{(SSE_0 - SSE_1)/q}{(SSE_1/(n-p))} \tag{4}$$

όπου p ο αριθμός των παραμέτρων του μοντέλου SSE_1, n ο αριθμός παρατηρήσεων και q ο αριθμός των επιπρόσθετων παραμέτρων μεταξύ του μοντέλου με σφάλμα SSE_1 και του εμφωλευμένου με σφάλμα SSE_0. Ειδικότερα, τα βασικότερα μέτρα που θα χρησιμοποιηθούν για την επιλογή του καταλληλότερου μοντέλου, είναι:

1. ο συντελεστής προσδιορισμού R^2 για την παλινδρόμηση της εξαρτημένης μεταβλητής ως προς τις επεξηγηματικές, ο οποίος δίνεται από τη σχέση:

   $$R^2 = 1 - \frac{SSE}{SST} \tag{5}$$

   όπου:

   $$SST = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{6}$$

   Ακόμη και όταν προστίθενται μεταβλητές στο μοντέλο οι οποίες δεν είναι στατιστικά σημαντικές, ο δείκτης αυτός θα αυξηθεί εξαιτίας του SSE. Μόνο μεγάλες αυξομοιώσεις είναι επομένως ενδεικτικές της σημαντικότητας αυτής της μεταβλητής που προσθαφαιρείται.

2. ο διορθωμένος συντελεστής προσδιορισμού $\bar{R}^2$, ο οποίος δίνεται από τη σχέση:

   $$\bar{R}^2 = 1 - \frac{(n-1)SSE}{(n-p)SST} \tag{7}$$

   είναι πιο χαρακτηριστικός για την επιλογή του βέλτιστου μοντέλου καθώς αυξάνεται μόνο εάν προστεθεί στο μοντέλο μια μεταβλητή που είναι στατιστικά σημαντική σύμφωνα με τον έλεγχο F.

3. η ποσότητα Cp-Mallows, η οποία εκφράζει τα επίπεδα της μεροληψίας που υφίστανται εντός του μοντέλου. Όσο πιο μικρή είναι αυτή η τιμή και πάντα μικρότερη ή ίση από τον αριθμό των παραμέτρων του μοντέλου, τόσο πιο αμερόληπτο χαρακτηρίζεται το μοντέλο. Οι υψηλές τιμές της Cp-Mallows απορρίπτουν απευθείας τα αντίστοιχα μοντέλα.

4. η ποσότητα AIC, η οποία εκφράζει την πιθανοφάνεια του μοντέλου και συγκεκριμένα για το πολλαπλό γραμμικό μοντέλο δίνεται από τη σχέση:

   $$AIC = n[ln(2\pi SSE/n) + 1] + 2(p+1) \tag{8}$$

   Όσο πιο μικρή είναι η τιμή του AIC, τόσο μικρότερο είναι είναι το SSE και άρα η προσαρμογή του μοντέλου στα δεδομένα.

Τα μέτρα αυτά θα πρέπει να λειτουργήσουν συνδυαστικά για την επιλογή του βέλτιστου μοντέλου.

Για την εποπτική εικόνα των μέτρων αυτών για όλους τους συνδυασμούς μεταβλητών που μπορούν να συμμετέχουν, αξιοποιείται η βιβλιοθήκη της R olsrr και συγκεκριμένα η εντολή ols_step _all _possible(). Παρατηρείται πως η πλειοψηφία των μοντέλων χαρακτηρίζεται από πολύ υψηλά επίπεδα μεροληψίας. Ωστόσο υπάρχει και μια σημαντική μερίδα αυτών για τα οποία το Cp-Mallows είναι μικρότερο του 11, δηλαδή τον μέγιστο αριθμό συντελεστών που μπορεί να δεχτεί το συγκεκριμένο μοντέλο. Για μια συνοπτικότερη εύρεση του βέλτιστου μοντέλου θα χρησιμοποιηθεί η εντολή step() της R, η οποία θα επιλέξει με βάση των στατιστικό έλεγχο F και το κριτήριο AIC. Ο υπολογισμός θα ξεκινήσει από το μοντέλο με μια παράμετρο πόλωσης και θα καταλήξει στις 11 παραμέτρους (forward selection).Επιπλέον, θα πραγματοποιηθεί και ο αντίστροφος έλεγχος, ξεκινώντας δηλαδή από μοντέλο με τις 11 παραμέτρους και καταλήγοντας στο μοντέλο με μια σταθερή παράμετρο (backward selection). Ειδικότερα παρατηρείται ότι οι δύο βηματικές τεχνικές καταλήγουν σε δύο διαφορετικά μοντέλα με τα χαρακτηριστικά που παρουσιάζονται στους Πίνακες 3 και 4. Τα μέτρα καταλληλότητας προέκυψαν από την εντολή ols_step _all _possible().

R^2	0.8431
$\overline{R}^2$	0.8263
R^2_{pred}	0.7957
C_p	1.147
AIC	155.477

Πίνακας 3: εξαρτημένη μεταβλητή: mpg, ανεξάρτητες μεταβλητές: wt, cyl, hp

R^2	0.8497
$\overline{R}^2$	0.8336
R^2_{pred}	0.7946
C_p	0.103
AIC	154.119

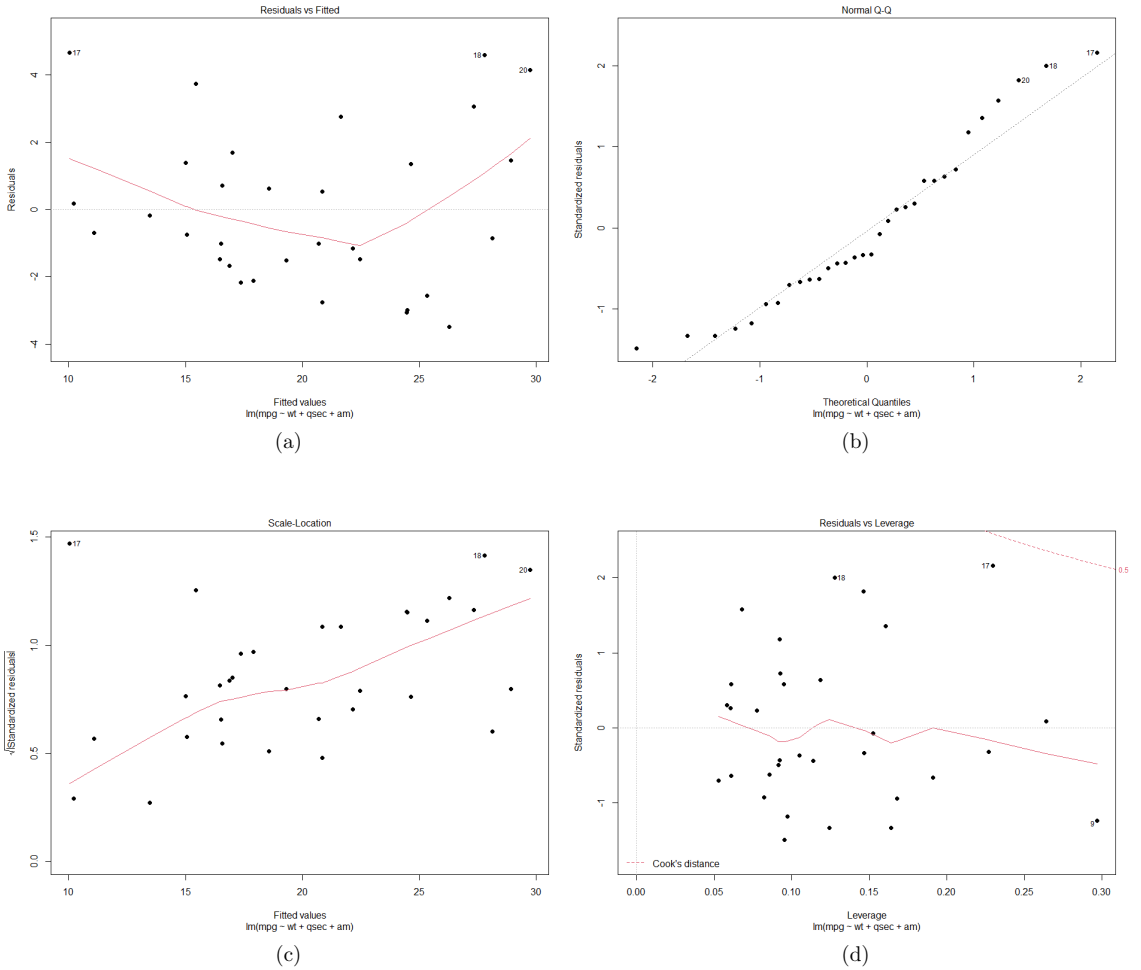
Πίνακας 4: εξαρτημένη μεταβλητή: mpg, ανεξάρτητες μεταβλητές: wt, qsec, am

Παρατηρείται πως και στα δύο μοντέλα η ποσότητα Cp-Mallows είναι σημαντικά μικρότερη από τον αριθμό 11, των μέγιστων παραμέτρων, επομένως η αμεροληψία τους είναι αρκετά υψηλή. Ωστόσο, το μοντέλο mpg ∼wt, qsec, am παρουσιάζει μικρότερη τιμή για το AIC καθώς επίσης και για την ποσότητα Cp-Mallows. Επομένως αυτό το μοντέλο συνδυάζει υψηλή αμεροληψία και το ελάχιστο δυνατό μέσο τετραγωνικό σφάλμα και άρα για το συγκεκριμένο πρόβλημα θεωρείται το κατάλληλο. Οι διαφορές στους συντελεστές προσδιορισμούς για τα δύο μοντέλα είναι ελάχιστες.

1.3 Ερώτημα 3

Το μοντέλο που επιλέχθηκε με βάση τις βηματικές τεχνικές είναι κατάλληλο εάν θεωρηθεί πως οι ανεξάρτητες μεταβλητές παρουσιάζουν γραμμική σχέση με την εξαρτημένη μεταβλητή mpg. Για το μοντέλο αυτό, θα πραγματοποιηθούν ορισμένοι διαγνωστικοί έλεγχοι προκειμένου να πιστοποιηθεί εάν είναι απαραίτητος κάποιος μετασχηματισμός. Αρχικά πραγματοποιούνται διαγνωστικοί έλεγχοι με βάση τα υπόλοιπα, οι οποίοι παρουσιάζονται στην Εικόνα 5. Παρατηρείται πως και το νέο μειωμένο μοντέλο ικανοποιεί σε μεγάλο βαθμό τις προϋποθέσεις της κανονικής κατανομής, ωστόσο τα υπόλοιπα ορισμένων παρατηρήσεων όπως η 17η, η 18η και η 20η παρουσιάζουν συστηματική απόκλιση από την ευθεία του αντίστοιχου QQ-plot. Επιπλέον, τα τυποποιημένα υπόλοιπα παρουσιάζουν μια συστηματική

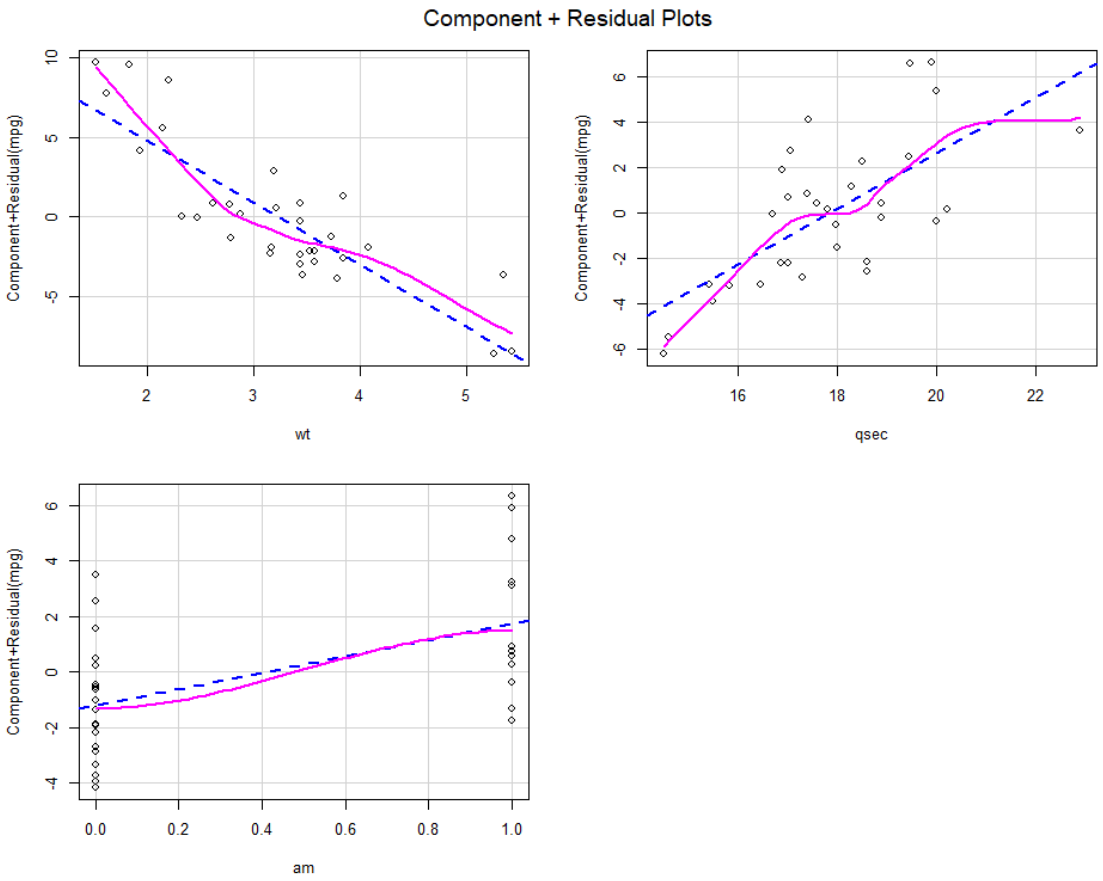
άνοδο σύμφωνα με τις προσαρμοσμένες τιμές και επομένως, η πιθανότητα ενός μετασχηματισμού των ανεξάρτητων μεταβλητών αυξάνεται. Ωστόσο, για το συγκεκριμένο μοντέλο, τα επίπεδα της μόχλευσης διατηρούνται χαμηλά και κανένα από τα σημεία δεν ξεπερνά τα όρια της απόστασης Cook.



(a) (b) (c) (d)

Εικόνα 5: Γραφήματα διαγνωστικών ελέγχων τελικού μοντέλου

Στη συνέχεια, πραγματοποιούνται τα γραφήματα των μερικών υπολοίπων για την κάθε μεταβλητή, τα οποία παρουσιάζονται στην Εικόνα 6 καθώς επίσης και τα γραφήματα πρόσθετων μεταβλητών στην Εικόνα 7. Από τα γραφήματα μερικών υπολοίπων προκύπτει πως οι μεταβλητές wt και qsec μπορεί να χρειάζονται κάποιον μετασχηματισμό καθώς δεν ακολουθούν ικανοποιητικά την αναμενόμενη ευθεία. Παρόλα αυτά όμως δεν εμφανίζεται κάποια μη γραμμική σχέση και συνεπώς η υψηλή σκεδαστικότητα κατά μήκος των ευθειών αντικατοπτρίζει πιθανώς την σχετικά μειωμένη πληροφορία της κάθε μεταβλητής στο μοντέλο. Από τα διαγράμματα των πρόσθετων μεταβλητών, προκύπτει πως και οι τρεις μεταβλητές παρουσιάζουν θετική ή αρνητική συσχέτιση με την εξαρτημένη μεταβλητή και συνεπώς συνεισφέρουν στο μοντέλο. Ωστόσο, η συσχέτιση αυτή δεν είναι τόσο ισχυρή καθώς οι παρατηρήσεις δεν ακολουθούν πλήρως τις αναμενόμενες ευθείες. Ωστόσο, δεν εμφανίζεται κάποια μη γραμμική σχέση μεταξύ των παρατηρήσεων και συνεπώς δεν κρίνεται απαραίτητος κάποιος μετασχηματισμός τους. Παρόλα αυτά όμως, οι αποκλίσεις που εμφανίζονται από τις επιθυμητές ευθείες τόσο στα διαγράμματα μερικών υπολοίπων όσο και πρόσθετων μεταβλητών επιβάλλει την δοκιμή ορισμένων μετασχηματισμών των επεξηγηματικών μεταβλητών, σε περίπτωση που βελτιωθούν τα χαρακτηριστικά των μοντέλου.



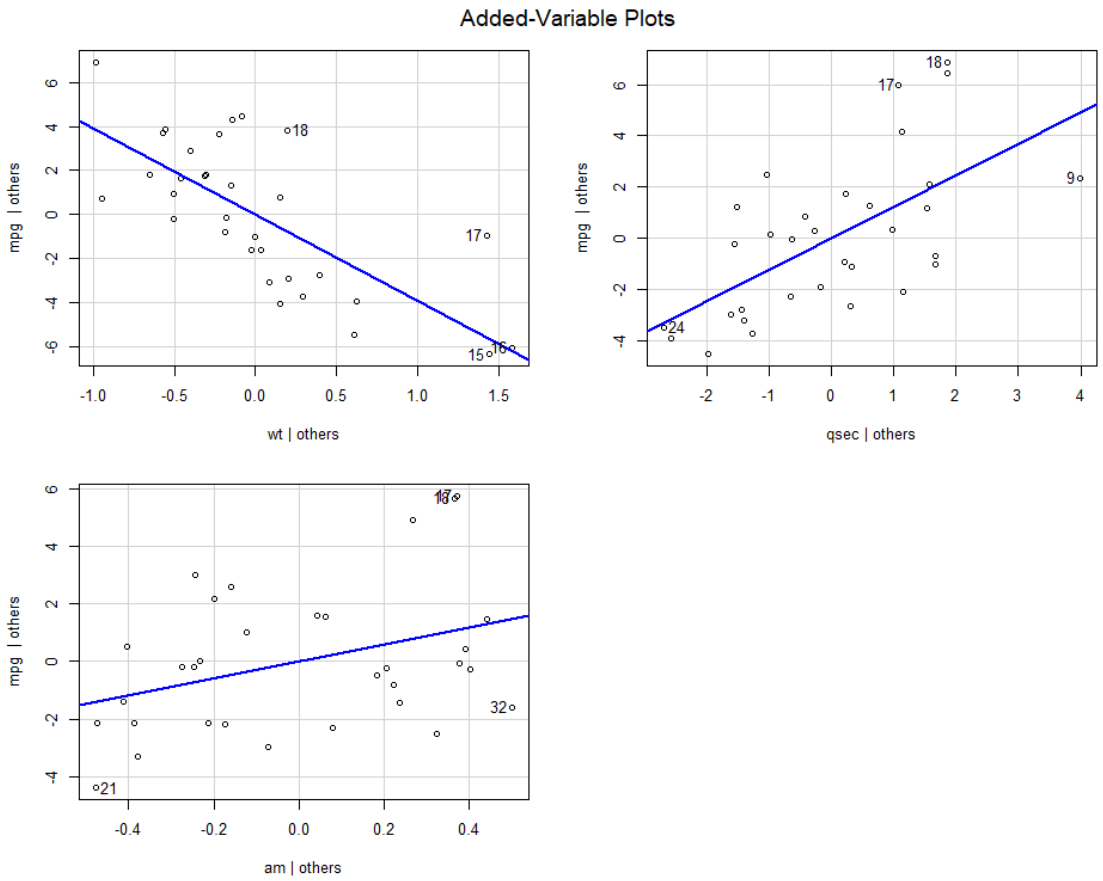
Εικόνα 6: Partial Residual Plots

Για την βελτίωση του υπάρχοντος μοντέλου πραγματοποιήθηκαν μετασχηματισμοί στις μεταβλητές wt και qsec. Δοκιμάστηκαν τα μοντέλα που προστέθηκε το τετράγωνο αυτών των μεταβλητών καθώς επίσης και τα μοντέλα με την αντικατάσταση αυτών των μεταβλητών από τον αντίστοιχο λογάριθμο. Ωστόσο, τα τρία από τα τέσσερα υπό εξέταση μοντέλα εμφάνιζαν στατιστικά μη σημαντικές μεταβλητές σύμφωνα με τον έλεγχο t και συνεπώς απορρίφθηκαν. Το μοναδικό μοντέλο για το οποίο οι μεταβλητές κρίθηκαν στατιστικά σημαντικές είναι το mpg ∼wt, log(qsec), am καθώς οι p-values που προέκυψαν ήταν 0.036346, 0.000126 και 6.3e-06 αντίστοιχα. Επιπλέον, οι ποσότητες αξιολόγησης αυτού του μοντέλου παρουσιάζονται στον Πίνακα 5. Παρατηρείται, πως σε σχέση με το μοντέλο mpg ∼wt, qsec, am, το AIC μειώνεται όπως επίσης και η ποσότητα Cp-Mallows. Επομένως το τετραγωνικό σφάλμα μειώνεται και η αμεροληψία αυξάνεται. Επιπλέον, αυξάνεται ελάχιστα και ο συντελεστής προσδιορισμού του μοντέλου. Άρα το τελικό βέλτιστο μοντέλο είναι το mpg ∼wt, log(qsec), am με συντελεστές που παρουσιάζονται στον Πίνακα 6.

R^2	0.8551
$\overline{R}^2$	0.8395
C_p	-0.7641231
AIC	152.9474

Πίνακας 5: εξαρτημένη μεταβλητή: mpg, ανεξάρτητες μεταβλητές: wt, log(qsec), am

Η παρουσία πιθανών άτυπων σημείων για αυτό το μοντέλο θα αξιολογηθεί με βάση τα κριτήρια DFFITS και DFBETAS. Ειδικότερα, το DFBETAS υπολογίζεται για την κάθε παρατήρηση για κάθε έναν από τους συντελεστές του μοντέλου και συνεπώς εκφράζει την επιρροή της i-οστής παρατήρησης στον j-οστό συντελεστή. Οι παράμετροι αυτού του μοντέλου είναι p=4 και οι παρατηρήσεις n = 32.



Εικόνα 7: Added Variable Plots

$\hat{\beta}_0$	-33.8909
$\hat{\beta}_1$ (wt)	3.0547
$\hat{\beta}_2$ (log(qsec))	22.6619
$\hat{\beta}_3$ (am)	-3.8730

Πίνακας 6: Συντελεστές παλινδρόμησης μοντέλου mpg ∼wt, log(qsec), am

Επομένως, το κατώφλι για τα δύο κριτήρια προκειμένου ένα σημείο να θεωρηθεί σημείο επιρροής είναι:

$$|DFFITS_i| > 2\sqrt{p/n} = 0.71 \tag{9}$$

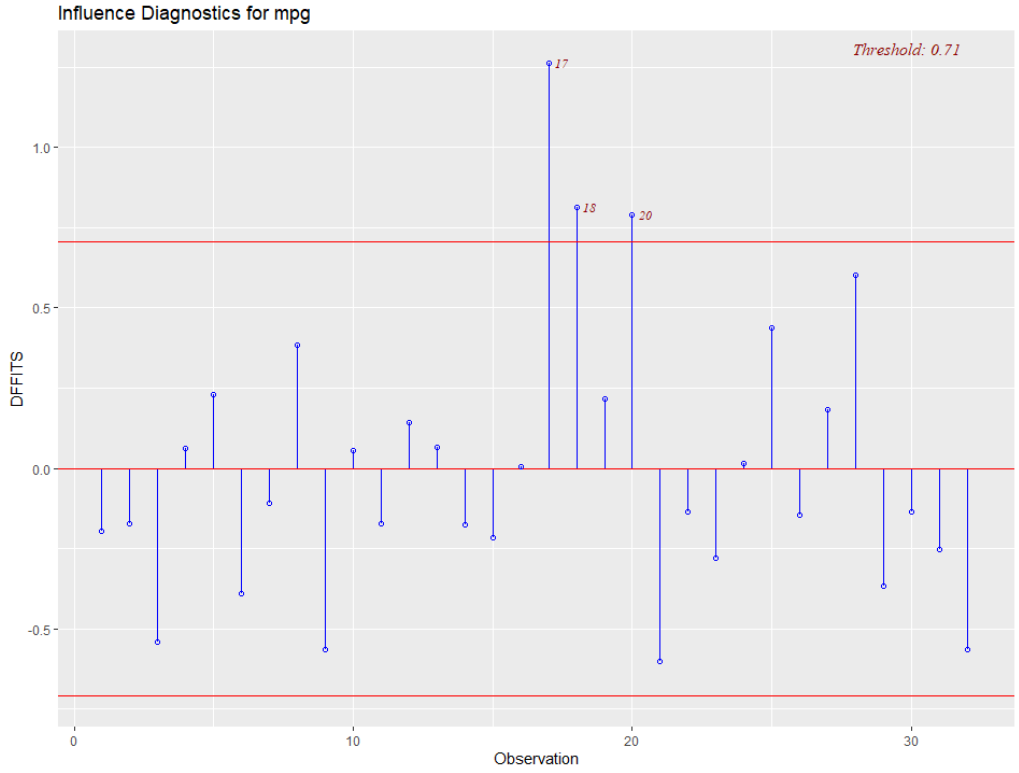
$$|DFBETAS_{ij}| > 2/\sqrt{n} = 0.35 \tag{10}$$

Για την συνοπτικότερη παρουσίαση των αποτελεσμάτων χρησιμοποιήθηκε η βιβλιοθήκη olsrr και πιο συγκεκριμένα οι εντολές ols_plot_dfbetas() και ols_plot_dffits(). Τα αποτελέσματα παρουσιάζονται στα Γραφήματα 8 και 9. Σύμφωνα με το κριτήριο DFFITS, τα βασικά σημεία που επηρεάζουν ισχυρά την τιμή της πρόβλεψης για την mpg είναι οι παρατηρήσεις 17, 18 και 20 και επιπλέον επηρεάζουν ισχυρά την τιμή του συντελεστή για την μεταβλητή log(qsec). Επίσης, η 17η παρατήρηση επηρεάζει ισχυρά και τους συντελεστές των μεταβλητών wt και am και συνεπώς ανάγεται σε ένα σημείο υψηλής επιρροής για το μοντέλο. Επιπρόσθετο σημείο επιρροής τόσο για την συντελεστή της μεταβλητής log(qsec) όσο και για τον συντελεστή πόλωσης είναι η 9η παρατήρηση.

Επιπρόσθετα για το συγκεκριμένο μοντέλο υπολογίζονται τα 95% διαστήματα εμπιστοσύνης για τους συντελεστές των μεταβλητών, τα οποία παρουσιάζονται στον Πίνακα 7. Παρατηρείται πως ειδικά στην περίπτωση της πόλωσης όπως επίσης και στον συντελεστή της μεταβλητής log(qsec), τα διαστήματα εμπιστοσύνης είναι αρκετά ευρεία. Αυτό πιθανώς οφείλεται στην ύπαρξη σημείων επιρροής που αποσταθεροποιούν την προσαρμογή του μοντέλου.

Πόλωση (Intercept)	-67.0736786	-0.7080747
am	0.2084962	5.9009238
log(qsec)	12.2167820	33.1069944
wt	-5.3042945	-2.4417862

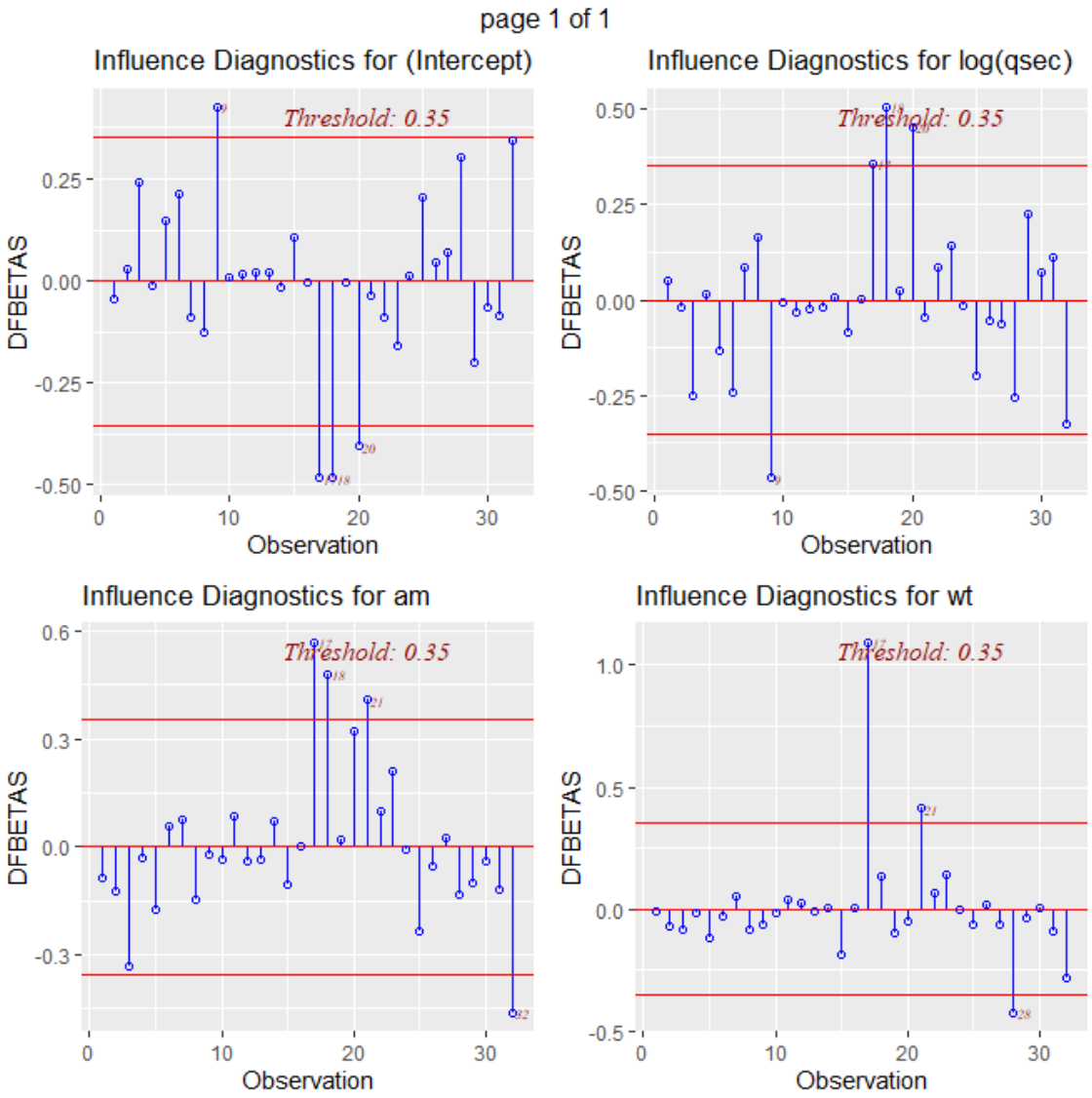
Πίνακας 7: 95% διαστήματα εμπιστοσύνης για τους συντελεστές του μοντέλου mpg ∼wt, log(qsec), am



Εικόνα 8: Κριτήριο DFFITS για κάθε παρατήρηση

Επιπρόσθετα, θα πραγματοποιηθεί και ο υπολογισμός του διαστήματος εμπιστοσύνης για μια πρόβλεψη για συγκεκριμένες τιμές των μεταβλητών του μοντέλου. Ειδικότερα, για βάρος αυτοκίνητου 3000 lbs (wt=3.0), για 21 sec χρόνο που χρειάζεται για να διανύσει το αυτοκίνητο το 25% από ένα μίλι (qsec=21.0) και αυτόματο κιβώτιο ταχυτήτων (am=0), προκύπτει πως διανύονται $\hat{mpg} = 23.48463$ miles/gallon με 95% διάστημα εμπιστοσύνης [18.2189 28.75036].

Η ερμηνεία των συντελεστών για το συγκεκριμένο μοντέλο είναι εφικτή για τους συντελεστές των μεταβλητών wt και am. Ειδικότερα εάν το βάρος wt του αυτοκινήτου αυξηθεί κατά μια μονάδα, διατηρώντας τις άλλες μεταβλητές σταθερές, τότε η πρόβλεψη για τα miles/gallon mpg θα μεταβληθεί κατά 3.0547 miles/gallon. Επιπλέον, εάν μεταβληθεί η κατηγορία του κιβώτιου ταχυτήτων από αυτόματο (am=0) σε χειροκίνητο (am=1), τότε η πρόβλεψη για την mpg θα μεταβληθεί κατά -3.8730 miles/gallon.



Εικόνα 9: Κριτήριο DFBETAS για κάθε παρατήρηση και για κάθε συντελεστή

2 Άσκηση 2

2.1 Ερώτημα 1

Τα δεδομένα που θα εξεταστούν στην συγκεκριμένη άσκηση αφορούν τον αριθμό των παλμών κελαηδήματος Υ δύο διαφορετικών ειδών καναρινιού σε σχέση με διαφορετικές τιμές της θερμοκρασίας περιβάλλοντος X_1 σε βαθμούς Κελσίου. Το είδος του καναρινιού προσδιορίζεται από την δυαδική μεταβλητή X_2, η οποία παίρνει μηδενική τιμή για την ομάδα B και την τιμή 1 για την ομάδα A των καναρινιού. Επίσης, εισάγεται μια επιπλέον μεταβλητή $X_3 = X_1X_2$, η οποία αποτελεί τον όρο αλληλεπίδρασης μεταξύ των X_1 και X_2. Θα προσαρμοστεί το πολλαπλό γραμμικό μοντέλο παλινδρόμησης:

$$E\left(y_x\right) = \beta_0 + \beta x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{11}$$

Ωστόσο, όπως παρατηρείται στο Γράφημα 10, τα δύο είδη καναρινιού παρουσιάζουν διαφορετική γραμμική συμπεριφορά και συνεπώς είναι αρκετά πιθανό να πρέπει να προσαρμοστούν δύο διαφορετικές ευθείες. Για να εξεταστεί ο αριθμός των ευθειών που απαιτούνται θα πρέπει να διερευνηθεί η σχέση 11 για τα δύο είδη ξεχωριστά. Ειδικότερα για $X_2 = 1$, δηλαδή για το είδος A, η σχέση 11 μετατρέπεται στην σχέση:

$$E\left(y_x\right) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1 \tag{12}$$

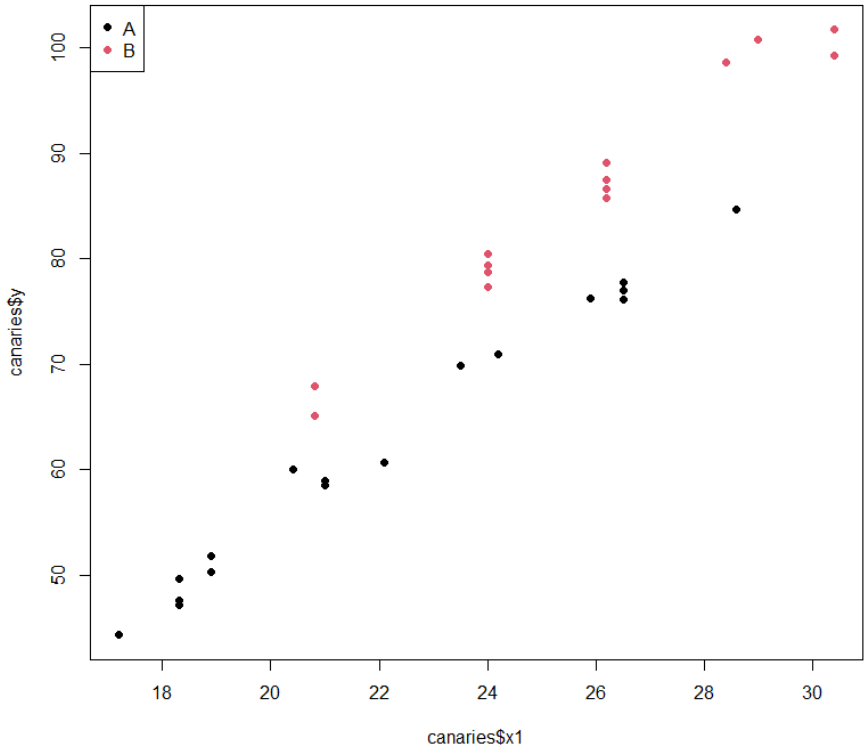
Αντίστοιχα για $X_2 = 0$, δηλαδή για το είδος B, προκύπτει η σχέση:

$$E\left(y_x\right) = \beta_0 + \beta_1 x_1 \tag{13}$$

Από τις σχέσεις 12 και 13 προκύπτει πως οι συντελεστές που θα καθορίσουν το πόσες ευθείες χρειάζονται είναι οι β_2 και β_3. Ο τρόπος για να ελεγχθεί στατιστικά η αναγκαιότητά τους είναι μέσω του ελέγχου F. Αρχικά, θα πραγματοποιηθεί η μηδενική υπόθεση $_0$ στην οποία τίθεται πως $\beta_3 = 0$. Εάν η υπόθεση απορριφθεί τότε χρειάζονται δύο μη παράλληλες ευθείες. Εναλλακτικά, εάν δεν απορριφθεί τότε χρειάζεται είτε να προσαρμοστούν δύο παράλληλες ευθείες, καθώς η κλίση της X_1 δεν θα αλλάξει μεταξύ των 12 και 13, είτε να προσαρμοστεί μια ευθεία και στα δύο είδη. Η επιλογή μεταξύ των δυο παράλληλων ευθειών και της μιας μοναδικής ευθείας συντελείται μέσω του στατιστικού ελέγχου F για τον συντελεστή β_2 και της μηδενικής υπόθεσης H_0, όπου $\beta_2 = 0$. Εάν η υπόθεση απορριφθεί τότε προκύπτουν δύο παράλληλες ευθείες, μια για το κάθε είδος καναρινιού, ενώ εάν δεν απορριφθεί, η ευθεία που θα προσαρμοστεί στα δεδομένα θα είναι μια. Ο έλεγχος για τον συντελεστή β_3 θα πραγματοποιηθεί μέσω της προσαρμογής του γραμμικού μοντέλου στις μεταβλητές X_1, X_2 και X_3, ενώ ο έλεγχος για τον συντελεστή β_2 πραγματοποιείται μέσω της προσαρμογής στις X_1 και X_2.

2.2 Ερώτημα 2

Αρχικά, μέσω της εντολής lm() στην R προσαρμόζεται το πολλαπλό γραμμικό μοντέλο στο σύνολο των δεδομένων για τις μεταβλητές X_1, X_2 και X_3 και ο κώδικας παρουσιάζεται στο τέλος της εργασίας. Θεωρώντας την μηδενική υπόθεση H_0, όπου ο συντελεστής της X_3 β_3 είναι μηδέν,παρατηρείται πως δεν μπορεί να απορριφθεί καθώς ο έλεγχος δεν είναι στατιστικά σημαντικός. Ειδικότερα μέσω της εντολής anova() για στατιστικό έλεγχο F, η p-value που αντιστοιχεί είναι 0.254, η οποία δεν είναι επαρκώς μικρή. Επομένως ισχύει ότι $\beta_3 = 0$ και θα πρέπει να ελεγχθεί και ο συντελεστής της μεταβλητής



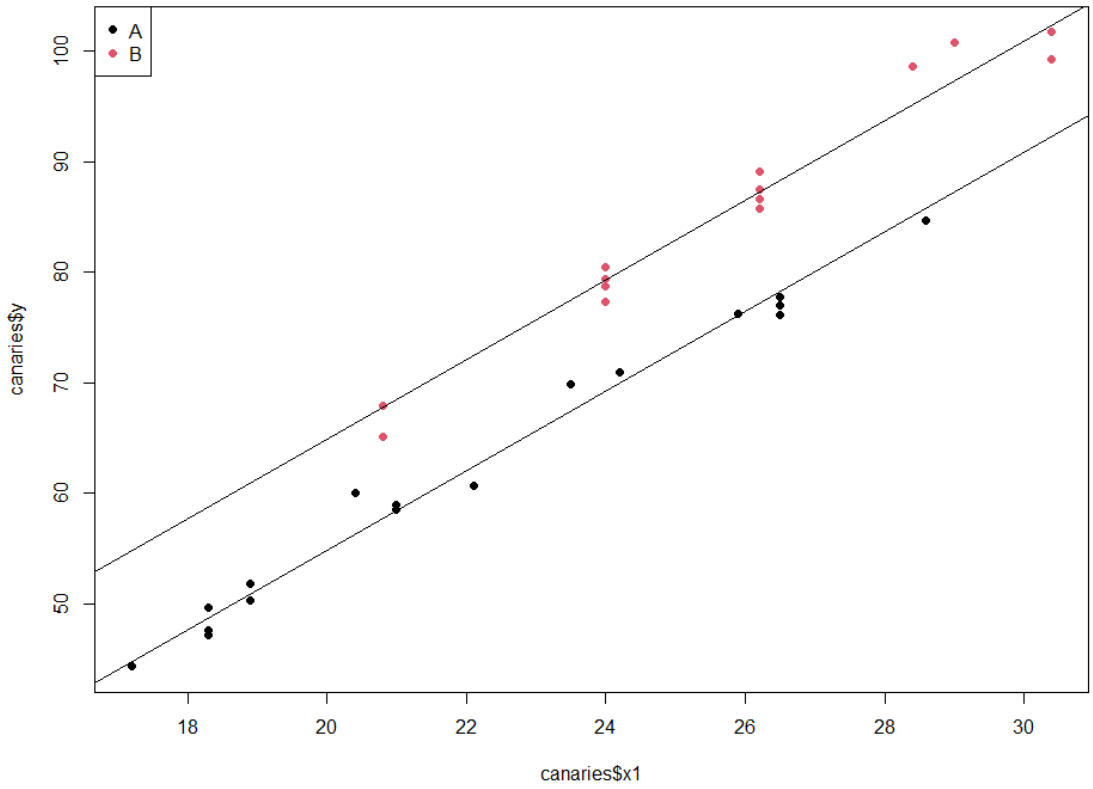
Εικόνα 10: Παλμοί Κελαηδήματος Υ σύμφωνα με την θερμοκρασία περιβάλλοντος X1

X_2, β_2. Προσαρμόζοντας, λοιπόν το πολλαπλό γραμμικό μοντέλο στις μεταβλητές X_1 και X_2, προκύπτει πως για την μηδενική υπόθεση H_0, όπου $\beta_2 = 0$, ο έλεγχος είναι στατιστικά σημαντικός. Ειδικότερα, η p-value του ελέγχου είναι 6.27e-14 και άρα ισχύει ότι $\beta_2 \neq 0$. Συνεπώς, θα πρέπει να προσαρμοστούν στα δεδομένα δύο παράλληλες ευθείες με κλίση την τιμή του συντελεστή $\beta_1 = 3.60275$ σύμφωνα με το μοντέλο $Y \sim X_1 + X_2$. Επομένως, οι σχέσεις 12 και 13 για τιμές συντελεστών $\beta_2 = 10.06529$ και $\beta_0 = -17.27620$ θα περιγράφουν τους παλμούς κελαηδήματος σύμφωνα με τη θερμοκρασία περιβάλλοντος για κάθε είδος καναρινιού σύμφωνα με τις σχέσεις:

$$E\left(y_x\right)_A = -7.21091 + 3.60275x_1 \tag{14}$$

$$E\left(y_x\right)_B = -17.27620 + 3.60275x_1 \tag{15}$$

Οι γραφικές παραστάσεις των αντίστοιχων ευθειών παρουσιάζονται στο Γράφημα 11. Ο συντελεστής $\beta_0 = -17.27620$ εκφράζει την αναμενόμενη τιμή των παλμών κελαηδήματος Y στην περίπτωση της μηδενικής θερμοκρασίας $X_1 = 0$ και του B είδους καναρινιού, $X_2 = 0$. Ο συντελεστής $\beta_1 = 3.60275$ εκφράζει την αναμενόμενη μεταβολή της Υ, σε περίπτωση που η θερμοκρασία αυξηθεί κατά μια μονάδα. Ο συντελεστής $\beta_2 = 10.06529$ εκφράζει την αναμενόμενη μεταβολή της Υ σε περίπτωση που το είδος του καναρινιού μεταβληθεί από A σε B.



Εικόνα 11: Ευθείες παλινδρόμησης για τους Παλμούς Κελαηδήματος Υ για τα δύο groups

Code – Exercise 1

```
file1<- read.table("C:/Users/user/Desktop/dsml/statistical
modeling/hw2/vehicles.txt",header=TRUE)

attach(file1)

file1


mydata<-subset(file1, select = -c(car))

mydata


mod1 = lm(mpg~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb)

summary(mod1)


library(car)


vif(mod1)

plot(mod1, pch=19)


cormat <- cor(mydata)

library(corrplot)

testcor=cor.mtest(mydata)

corrplot(cormat, type = "upper", tl.col = "black", p.mat = testcor$p, insig = 'p-value',sig.level
= 0)


plot(rstudent(mod1), pch=19)

qqnorm(rstudent(mod1), pch=19)

abline(0,1)


plot(hatvalues(mod1), pch=19)
```

```
hatvalues(mod1)


plot(cooks.distance(mod1), pch=19)

cooks.distance(mod1)


plot(dffits(mod1), pch=19)

dffits(mod1)




dfbetas(mod1)


rstandard(mod1)

rstudent(mod1)


##erwthma 2


library(olsrr)

ols_step_all_possible(mod1)

data.frame(ols_step_all_possible(mod1))


mod2 = step(lm(mpg~1), y~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb,
direction = "forward", test="F")

mod3 = step(mod1, direction = "backward", test="F")


mydata$wt_sq = mydata$wt^2

mydata$qsec_sq = mydata$qsec^2


mod4 = lm(mpg~ am + qsec + wt + mydata$wt_sq)
```

```
mod5 = lm(mpg~ am + qsec + wt + mydata$qsec_sq)

mod6 = lm(mpg~ am + log(qsec) + wt)

mod7 = lm(mpg~ am + qsec + log(wt))

summary(mod4)

summary(mod5)

summary(mod6)

summary(mod7)


##erwthma 3

#par(mfrow = c(2,2))

plot(mod3, pch=19)


avPlots(mod3)

crPlots(mod3)


ols_plot_dfbetas(mod6)

ols_plot_dffits(mod6)

confint(mod6)

predict(mod6, newdata=list(wt=3.0, qsec=21.0, am =0 ), interval="prediction", level=.95)
```

Code – Exercise 2

```
data = read.table('C:/Users/user/Desktop/dsml/statistical modeling/hw2/canary.txt',
header=TRUE)

library(data.table)

canaries = data.table(NULL)


canaries$x1 = data$Temp

canaries$x2 = ifelse(data$group=="A",1,0)

canaries$x3 = canaries$x1 * canaries$x2

canaries$y = data$pulses

mod1 = lm(canaries$y~ canaries$x1 + canaries$x2 + canaries$x3)

summary(mod1)

mod2 = lm(canaries$y~ canaries$x1 + canaries$x2)

summary(mod2)

mod3 = lm(canaries$y~ canaries$x1)

summary(mod3)


anova(mod1,mod2, test="F")

anova(mod3,mod2, test="F")


plot(canaries$x1, canaries$y,

    pch = 19,

    col = factor(canaries$x2))

abline(-7.21091,3.60275)

abline(mod2)


legend("topleft",

      legend = levels(factor(data$group)),
```

```
pch = 19,

col = factor(levels(factor(data$group))))
```