ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ ΤΗΣ ΣΧΟΛΗΣ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ»

ΑΝΘΕΚΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΜΕ ΧΡΗΣΗ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R

Διπλωματική Εργασία

του
Μουράτ Βάρνα

Επιβλέπουσα : Χρυσηίς Καρώνη
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Μήνας 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ ΤΗΣ ΣΧΟΛΗΣ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ»

ΑΝΘΕΚΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΣΤΗΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΜΕ ΧΡΗΣΗ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R

Διπλωματική Εργασία

του

ΜΟΥΡΑΤ ΒΑΡΝΑ

Επιβλέπουσα : Χρυσηίς Καρώνη
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Φεβρουαρίου 1999.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

(Υπογραφή)
.....
ΜΟΥΡΑΤ ΒΑΡΝΑ
Διπλωματούχος Δ.Π.Μ.Σ. Εφαρμοσμένων Μαθηματικών Επιστημών Ε.Μ.Π.

Αθήνα, Μήνας 2016

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την Καθηγήτρια κα. Χρυσηίς Καρώνη Ρίτσαρντσον για την πολύτιμη καθοδήγησή της, την κατανοησή της και το ευχάριστο κλίμα συνεργασίας που αναπτύξαμε καθ'όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας.

Επίσης ευχαριστώ τους φίλους μου για τις πολύτιμες υποδείξεις και συμβουλές τους.

Πάνω απ' όλα, είμαι ευγνώμων στους γονείς μου για την αγάπη και υποστήριξή τους όλα αυτά τα χρόνια.

Περίληψη

Η ανάλυση παλινδρόμησης είναι μια ευρέως χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης, που βρίσκει εφαρμογή στις περισσότερες επιστήμες. Μελετάμε τη σχέση δύο ή περισσοτέρων μεταβλητών και στη συνέχεια προσδιορίζουμε τη σχέση αυτή με βάση ορισμένες παρατηρήσεις. Η πιο γνωστή μέθοδος προσαρμογής μίας παλινδρόμησης που υπάρχει είναι η μέθοδος ελαχίστων τετραγώνων. Όμως, ο κίνδυνος που οφείλεται στην ύπαρξη ακραίων παρατηρήσεων στα δεδομένα μας, είναι μεγάλος και μπορεί να διαστρεβλώσει ακόμη και ολόκληρη τη στατιστική ανάλυση. Για την αντιμετώπιση αυτού του προβλήματος έχουν αναπτυχθεί ανθεκτικές (robust) μέθοδοι τα αποτελέσματα των οποίων δεν επηρεάζονται από τις ακραίες παρατηρήσεις.

Στόχος λοιπόν, είναι η χρήση μεθόδων οι οποίες δεν ευαισθητοποιούνται σε μικρές παραβιάσεις των υποθέσεων που επιφέρουν σφάλματα στα τελικά αποτελέσματα. Γίνεται αναφορά στις γνωστότερες ανθεκτικές μεθόδους Huber M-εκτιμήτρια, εκτιμήτρια Ελαχίστων Περικοπτόμενων Τετραγώνων, bisquare MM-εκτιμήτρια, L1 εκτιμήτρια. Τέλος, βάσει παραδειγμάτων γίνεται σύγκριση ανάμεσα στις παραπάνω ανθεκτικές μεθόδους και τη μέθοδο ελαχίστων τετραγώνων, σε συνδυασμό με τη χρήση διαγνωστικών μεθόδων, σε μολυσμένες βάσεις δεδομένων που ακολουθούν την Κανονική κατανομή.

Abstract

In statistical modeling, regression analysis is a very popular statistical technique that is applied in most sciences. The focus is on the relationship between two or more random variables and then determine the relationship based on some observations. The best known method of fitting a regression is the method of Ordinary Least Squares. However, the risk due to the existence of extreme observations in our data, is large and can even distort the entire statistical analysis. To tackle this problem, robust methods have been developed, producing results that are not influenced by outliers.

The aim therefore is to use methods that are not sensitive to minor violations of assumptions that could lead to errors in the final results. Reference is made to the best known robust methods: Huber M-estimator, Least Trimmed Squares estimator, bisquare MM-estimator, L1 estimator. Finally, on the basis of examples, comparisons are made between results obtained with these robust methods and the method of least squares in conjunction with the use of diagnostic methods on perturbed databases that follow the Normal distribution.

Πίνακας περιεχομένων

Κεφάλαιο 1

Πολλαπλή Γραμμική Παλινδρόμηση

1.1 Ορισμός μοντέλου

Το γενικό γραμμικό μοντέλο δίνεται από τη σχέση

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

όπου

- y_i, *i=1, 2,..., n*, οι τιμές των παρατηρήσεων της εξαρτημένης μεταβλητής *y*
- x_{ij}, *i=1, 2,..., n*, *j=1, 2,..., k*, οι τιμές των ανεξάρτητων (ή επεξηγηματικών) μεταβλητών x_j, για την *i*-οστή παρατήρηση
- $\beta_0, \beta_1, \beta_2, \dots, \beta_k$, οι άγνωστες παράμετροι του μοντέλου και
- ε_i, *i=1, 2,..., n*, τα τυχαία σφάλματα, τα οποία υποθέτουμε ότι ικανοποιούν τις παρακάτω υποθέσεις ανάλογες του απλού γραμμικού μοντέλου
  - $E(\varepsilon_i) = 0$, για κάθε *i*
  - $V(\varepsilon_i) = \sigma^2$, για κάθε *i*, δηλαδή τα τυχαία σφάλματα ικανοποιούν την υπόθεση της ομοσκεδαστικότητας
  - $cov(\varepsilon_i, \varepsilon_j) = 0$, για *i≠j*, δηλαδή τα ε_i είναι ασυσχέτιστα μεταξύ τους

Η αρχική σχέση μπορεί να γραφτεί και υπό τη μορφή πινάκων ως εξής:

$$y = X\beta + \varepsilon,$$

όπου $y = (y_1, y_2, \dots, y_n\ )'$ είναι το *n×1* διάνυσμα τιμών της μεταβλητής απόκρισης *y*, *X* ο n×p πίνακας σχεδιασμού με *p=k+1*, $(\beta = \beta_0, \beta_1, \dots, \beta_k)'$ το *p×1* διάνυσμα των παραμέτρων και $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$, *n×1* διάνυσμα των τυχαίων σφαλμάτων.

1.2 Μέθοδος ελαχίστων τετραγώνων

Η μέθοδος των ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων β βασίζεται στην ελαχιστοποίηση της παράστασης

$$S(\beta) = \big(y - E(y)\big)'\big(y - E(y)\big)$$

$$= (y - X\beta)'(y - X\beta).$$

Οπότε παραγωγίζοντας ως προς β έχουμε

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'(y - X\beta)$$

Και θέτοντας την παραπάνω σχέση ίση με το μηδέν, καταλήγουμε στη σχέση

$$X'y = X'X\widehat{\beta}.$$

Αν ο πίνακας $X'X$ αντιστρέφεται, τότε η *εκτιμήτρια ελαχίστων τετραγώνων* (Ordinary Least Squares estimator, OLS) του διανύσματος β δίνεται από τη σχέση

$$\hat{\beta} = (X'X)^{-1}X'y,$$

Οπότε και το μοντέλο που εκτιμούμε δίνεται από τη σχέση

$$\hat{y} = X\hat{\beta}$$

και τα υπόλοιπα (residuals) υπολογίζονται ως

$$r = y - \hat{y}.$$

Καθεμία από τις εκτιμήσεις $\hat{\beta}_j$, *j=1, 2,…,k*, εκφράζει την αναμενόμενη μεταβολή της y για μια μονάδα αύξησης της αντίστοιχης επεξηγηματικής μεταβλητής x_j, δεδομένου ότι οι άλλες επεξηγηματικές μεταβλητές παραμένουν σταθερές.

Υπό την υπόθεση ότι τα τυχαία σφάλματα κατανέμονται σύμφωνα με την $\varepsilon \sim N_n(0, \sigma^2 I)$, η κατανομή της *y* όπως ήδη παρατηρήσαμε είναι

$$y = X\beta + \varepsilon \sim N_n(X\beta, \sigma^2 I).$$

Επιπλέον το διάνυσμα $\hat{\beta}$ των εκτιμητριών ελαχίστων τετραγώνων μπορεί να εκφραστεί ως

$$\hat{\beta} = Ay,$$

όπου ο πίνακας $A = (X'X)^{-1}X'$ είναι ένας μη στοχαστικός πίνακας. Η αναμενόμενη τιμή του $\hat{\beta}$ είναι

$$E(\hat{\beta}) = E(Ay) = AE(y)$$

$$= (X'X)^{-1}X'X\beta = \beta,$$

και ο πίνακας διασποράς συνδιασποράς της $\hat{\beta}$ είναι

$$V(\hat{\beta}) = E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))']$$

$$= V(Ay) = AV(y)A'$$

$$= (X'X)^{-1}X'\sigma^2 I\{(X'X)^{-1}X'\}'$$

$$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1} = \sigma^2 C.$$

1.3 Ανάλυση διασποράς (ANOVA)

Η ανάλυση διασποράς (analysis of variance) εξετάζει τη σχέση της εξαρτημένης με τις ανεξάρτητες μεταβλητές, υπολογίζοντας στην ουσία το αν η μεταβλητότητα των τιμών της εξαρτημένης μεταβλητής y εξηγείται από τη μεταβλητότητα των ανεξάρτητων μεταβλητών x_j. Η ανάλυση διασποράς για το γενικό γραμμικό μοντέλο παρουσιάζεται στον παρακάτω πίνακα.

Πηγή μεταβλητότητας	*Άθροισμα τετραγώνων*	*Βαθμοί ελευθερίας*	*Μέσο άθροισμα τετραγώνων*	*Έλεγχος F*
Παλινδρόμηση (Regression)	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	k	$\text{MSR} = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Υπόλοιπα (Error)	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	n-k-1	$MSE = S^2 = \frac{SSE}{n-k-1} = \frac{SSE}{n-p}$	
Σύνολο (Total)	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$	n-1		

1.4 Πρόβλημα στην παλινδρόμηση

Στην πραγματικότητα, τα δεδομένα που έχουμε συλλέξει σχεδόν ποτέ δεν ακολουθούν ακριβώς, την Κανονική κατανομή. Επιπλέον, πολλές φορές στο δείγμα μας υπάρχουν απομακρυσμένες παρατηρήσεις (*ακραίες παρατηρήσεις, outliers*), οι οποίες επηρεάζουν τον δειγματικό μέσο. Οι παρατηρήσεις αυτές μπορεί να προέρχονται από λάθη πληκτρολόγησης, λάθη στα δεκαδικά ψηφία, από παράλειψη σημαντικών επεξηγηματικών μεταβλητών καθώς και από άλλα σπάνια φαινόμενα. Οι παρατηρήσεις αυτές υπάρχουν σχεδόν πάντα σε πραγματικά δεδομένα. Οπότε, τίθεται το θέμα της επιλογής του πιο ανθεκτικού (robust) μέτρου θέσης, π.χ. της διαμέσου, για την περιγραφή των δεδομένων. Η ανθεκτικότητα (robustness), κατά μια έννοια, αναφέρεται στην μη ευαισθησία της εκτιμήτριας σε ακραίες παρατηρήσεις και σε παραβιάσεις των προϋποθέσεων εφαρμογής ενός στατιστικού ελέγχου. Για να προσδιοριστεί ο βαθμός ανθεκτικότητας μιας διαδικασίας, έχουν εισαχθεί διάφορα μέτρα ανθεκτικότητας, όπως το *σημείο κατάρρευσης* (*breakdown point*). Το σημείο κατάρρευσης ορίζεται σαν ένα τμήμα των δεδομένων το οποίο μπορεί αυθαίρετα να αλλοιωθεί, χωρίς αυτό να έχει ως αποτέλεσμα να επηρεαστεί σημαντικά η εκτίμηση. Για παράδειγμα, αν στον δειγματικό μέσο υπάρχει έστω και μια παρατήρηση x, η οποία είναι μεγάλη συγκριτικά με τις υπόλοιπες, ο μέσος θα επηρεαστεί σημαντικά, θα αυξηθεί δηλαδή σε μεγάλο βαθμό. Αυτό σημαίνει ότι ακόμα και αν υπάρχει μόνο μια ακραία παρατήρηση, ο δειγματικός μέσος δεν θα είναι αντιπροσωπευτικό μέτρο θέσης. Συνεπώς, το σημείο κατάρρευσης του μέσου είναι το 0. Η διάμεσος χρησιμοποιείται συχνά ως μέτρο θέσης, όταν έχουμε λοξές κατανομές ή όταν υπάρχουν ακραίες παρατηρήσεις. Είναι το σημείο εκείνο κάτω από το οποίο βρίσκεται το 50% των (διατεταγμένων) παρατηρήσεων. Διαισθητικά, βλέπουμε ότι αν ένα υποσύνολο των παρατηρήσεων πάρει πολύ μεγάλες τιμές, ακόμα και αν τείνει στο άπειρο, η διάμεσος δεν θα επηρεαστεί. Το σημείο κατάρρευσης για τη διάμεσο είναι το *0,5*. Για να επηρεαστεί δηλαδή η διάμεσος πρέπει το 50% των παρατηρήσεων να αποτελείται από ακραίες τιμές. Από την οπτική λοιπόν της ανθεκτικότητας, ο μέσος είναι αξιόπιστο μέτρο θέσης όταν δεν έχουμε ακραίες παρατηρήσεις ή λοξότητα, στην κατανομή. Μη παραμετρικές μέθοδοι, όπως ο στατιστικός έλεγχος Mann-Whitney χρησιμοποιούν τη διάμεσο και άρα είναι πιο ανθεκτικές. Γενικά, οι

ανθεκτικές μέθοδοι εξάγουν έγκυρα συμπεράσματα ακόμα και αν μια σημαντική ποσότητα των δεδομένων είναι αλλοιωμένη. Μάλιστα είναι σε θέση να εντοπίσουν και να ανιχνεύσουν τις ακραίες παρατηρήσεις.

Άλλος λόγος απόκλισης από τις στατιστικές υποθέσεις είναι ο προσεγγιστικός χαρακτήρας πολλών θεωρητικών μοντέλων. Το κεντρικό οριακό θεώρημα (ΚΟΘ) συχνά χρησιμοποιείται για την υπόδειξη της κατανομής κατά προσέγγιση. Ιδιαίτερη προσοχή στην ανάλυση τους απαιτούν κατανομές με μεγαλύτερες ουρές από την Κανονική κατανομή.

Κατά την ανάλυση δεδομένων προκύπτουν ερωτήματα όπως: Ακολουθούν τα δεδομένα κάποια συγκεκριμένη μορφή ή διαφορετικά τμήματα δεδομένων δίνουν διαφορετικές πληροφορίες; Στη τελευταία περίπτωση τι υποδεικνύει η πλειοψηφία των δεδομένων; Ποια είναι η επίδραση των διαφορετικών τμημάτων στο τελικό αποτέλεσμα έτσι ώστε να εξεταστούν με μεγάλη προσοχή; Πόσα μεγάλα σφάλματα μπορεί να ανεχτεί το επιλεγμένο μοντέλο; Ποια μέθοδος είναι ταυτόχρονα η πιο ασφαλής και αποδοτική; Πόσο αξιόπιστα είναι τα αποτελέσματα αν οι υποθέσεις του μοντέλου ισχύουν μόνο κατά προσέγγιση;

Όλα τα παραπάνω ερωτήματα αποτέλεσαν την αιτία της ανάπτυξής της Ανθεκτικής Στατιστικής (Robust Statistics). Η ανθεκτική στατιστική είναι το σύνολο των γνώσεων και τεχνικών που σχετίζονται με τις αποκλίσεις από τις υποθέσεις που γίνονται στη στατιστική.

1.5 Ανθεκτικές μέθοδοι στη στατιστική

Πολλοί επιστήμονες συχνά χρησιμοποιούν τη στατιστική συμπερασματολογία για την ανάλυση δεδομένων και την εξαγωγή χρήσιμων συμπερασμάτων η οποία όμως βασίζεται μόνο κατά ένα μέρος στο δείγμα των παρατηρήσεων που έχουν συλλεχθεί. Αποκλίσεις από τις αρχικές υποθέσεις για το σύνολο των δεδομένων όπως είναι οι ανεξαρτησία, γραμμικότητα ή κανονικότητα συχνά παρατηρούνται με συνέπεια την αποτυχία γνωστών μεθόδων. Οι τελευταίες είναι βέλτιστες υπό το ακριβές μοντέλο όταν μικρές αποκλίσεις από τις υποθέσεις έχουν σα συνέπεια μικρά σφάλματα στα τελικά αποτελέσματα. Έτσι λοιπόν, στόχος είναι η χρήση μεθόδων που δεν ευαισθητοποιούνται σε μικρές παραβιάσεις υποθέσεων.

Η ανάλυση παλινδρόμησης είναι ένα σημαντικό στατιστικό εργαλείο που εφαρμόζεται στις περισσότερες επιστήμες. Από τις διάφορες μεθόδους παλινδρόμησης που υπάρχουν, η μέθοδος που τελικά επικράτησε λόγω της ευκολίας υπολογισμού της καθώς και λόγω της παράδοσης είναι η μέθοδος ελαχίστων τετραγώνων. Παρόλα αυτά διατρέχεται μεγάλος κίνδυνος με την ύπαρξη ακραίων (απομακρυσμένων) παρατηρήσεων, όπως προαναφέρθηκε. Οι απομακρυσμένες τιμές υπάρχουν πολύ συχνά σε πραγματικά δεδομένα και σχεδόν πάντα περνάνε απαρατήρητες εξαιτίας του γεγονότος ότι, η επεξεργασία των δεδομένων γίνεται μέσω υπολογιστών χωρίς να ελέγχονται επαρκώς.

Παρόλη την ευκολία υπολογισμού και την αποδοτικότητά της, η μέθοδος των ελαχίστων τετραγώνων αδυνατεί να αντιμετωπίσει ένα μεγάλο ποσό αλλοιωμένων δεδομένων. Διαπιστώνουμε λοιπόν την ευαισθησία της μεθόδου, που οφείλεται στο γεγονός ότι οι ακραίες παρατηρήσεις και οι υπόλοιπες αποκλίσεις από το μοντέλο τυπικής γραμμικής παλινδρόμησης εμφανίζονται πολύ συχνά στα πραγματικά δεδομένα. Στη συγκεκριμένη μέθοδο, η σημασία των ακραίων παρατηρήσεων, τόσο στην κατεύθυνση των εξαρτημένων μεταβλητών όσο και στην κατεύθυνση των ανεξάρτητων μεταβλητών μπορεί να επηρεάσει σημαντικά τις εκτιμήσεις των παραμέτρων. Και στις δύο περιπτώσεις οι ακραίες παρατηρήσεις μπορούν να καταστρέψουν τελείως την κλασική μέθοδο ελαχίστων τετραγώνων. Συχνά, τέτοιου είδους τιμές παραμένουν κρυμμένες στο χρήστη, καθώς δεν εμφανίζονται πάντα στα διαγράμματα σφαλμάτων με τη μέθοδο των ελαχίστων τετραγώνων.

Για να αντιμετωπιστεί το πρόβλημα αυτό, αναπτύχθηκαν νέες στατιστικές μέθοδοι οι οποίες δεν επηρεάζονται σε μεγάλο βαθμό από τις ακραίες παρατηρήσεις. Αυτές είναι οι ανθεκτικές μέθοδοι, τα αποτελέσματα των οποίων είναι αξιόπιστα ακόμη και όταν μια συγκεκριμένη ποσότητα των δεδομένων είναι αλλοιωμένη. Κάποιοι ερευνητές πιστεύουν ότι οι ανθεκτικές μέθοδοι παλινδρόμησης "κρύβουν" τις ακραίες τιμές, αλλά ακριβώς το αντίθετο συμβαίνει. Οι ακραίες τιμές βρίσκονται αρκετά μακριά από την ανθεκτική προσαρμογή και συνεπώς μπορούν να ανιχνευθούν από τα μεγάλα σφάλματα από αυτή, αντιθέτως τα τυποποιημένα σφάλματα μέσω της κλασικής μεθόδου των ελαχίστων τετραγώνων μπορεί να μην παρουσιάζουν καθόλου ακραίες τιμές.

Το πρόβλημα της ανθεκτικότητας προφανώς, χρονολογείται από την προϊστορία της στατιστικής. Παρόλο που πλήθος καταξιομένων στατιστικών, όπως ο Newcomb, K. Pearson, Gosset, E. S. Pearson, ήταν ενημερωμένοι επί του θέματος, μόνο τις τελευταίες δεκαετίες έγιναν προσπάθειες για να μοντελοποιηθεί το πρόβλημα και να αντιμετωπιστούν οι ακραίες παρατηρήσεις μέσω της θεωρίας της ανθεκτικότητας. Οι λόγοι γι' αυτή την καθυστερημένη ανάπτυξη δεν είναι ξεκάθαροι. Ωστόσο, η εκτενής εξέλιξη των μαθηματικών επιστημών και η ανάπτυξη των ηλεκτρονικών υπολογιστών έκαναν την αντικατάσταση των κυρίως υποκειμενικών μέχρι τώρα μεθόδων, σε επίσημες μεθόδους τόσο εφαρμόσιμες όσο και απτές, καθώς είναι δύσκολο να εντοπιστούν ακραίες παρατηρήσεις μέσα σε ένα ευρύ και περίπλοκο πλήθος δεδομένων από υποκειμενικές μεθόδους. Ένας άλλος σημαντικός λόγος ήταν προφανώς, η αυξημένη επίγνωση της ανάγκης για ανθεκτικές διαδικασίες κάτι που προκύπτει από το έργο των E. S. Pearson, G. E. P. Box, J. W. Tukey και πολλών άλλων.

1.6 Το πρόβλημα της πολυσυγγραμικότητας

Η ερμηνεία ενός προβλήματος με τη χρησιμοποίηση της μεθόδου αναλύσεως της πολλαπλής παλινδρόμησης επιτυγχάνεται καλύτερα όταν οι ανεξάρτητες μεταβλητές που αποτελούν το μοντέλο είναι μεταξύ τους ασυσχέτιστες. Όταν υφίσταται έντονες συσχετίσεις μεταξύ των μεταβλητών είναι δύσκολο, αν όχι αδύνατο, να αξιολογηθεί η ουσιαστική προσφορά μιας συγκεκριμένης ανεξάρτητης μεταβλητής επί της εξαρτημένης μεταβλητής που οφείλεται αποκλειστικά στη συγκεκριμένη ανεξάρτητη μεταβλητή. Έτσι ενδέχεται να χρησιμοποιήσουμε άλλη εκτιμήτρια από την εκτιμήτρια ελαχίστων τετραγώνων. Σ'αυτή την περίπτωση η προτίμηση για άλλη εκτιμήτρια δεν οφείλεται σε παραβίαση προϋποθέσεων του γραμμικού μοντέλου, αλλά σε μια συγκεκριμένη τεχνική δυσκολία.

Η κατάσταση η οποία δημιουργείται όταν υπάρχουν ισχυρές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών στην πολλαπλή παλινδρόμηση ονομάζεται πολυσυγγραμικότητα (multicollinearity).

Στις περιπτώσεις που το πρόβλημα αυτό υφίσταται θα πρέπει κανείς να είναι ιδιαίτερα προσεκτικός στην ερμηνεία όλων των εκτιμητριών που προκύπτουν από το μοντέλο αυτό. Καθώς οδηγεί σε αυξημένα τυπικά σφάλματα των $\hat{\beta}$ και κατά συνέπεια δυσκολεύει την εκτίμηση της επίδρασης της κάθε επεξηγηματικής μεταβλητής στην εξαρτημένη

μεταβλητή, αφού τα διαστήματα εμπιστοσύνης των αντίστοιχων συντελεστών θα είναι μεγάλα σε εύρος (Οικονόμου Π. και Καρώνη Χ., 2010).

Η τιμή $vif = \frac{1}{1-R_j^2}$ όπου R_j^2 ο συντελεστής προσδιορισμού μιας γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή τη x_j και επεξηγηματικές μεταβλητές όλες τις άλλες x_i, $i \neq j$ είναι γνωστή ως παράγοντας μεγέθυνσης διασποράς (variance inflation factor). Η τιμή αυτή δείχνει κατά πόσο αυξάνεται η διασπορά ενός εκτιμημένου συντελεστή παλινδρόμησης $\hat{\beta}_j$ όταν υπάρχουν συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών. Τιμές του $1 - R_j^2 < 0.2$ ή του $vif > 5$ θεωρούνται ως ένδειξη πολυσυγγραμικότητας.

Κεφάλαιο 2

Απομακρυσμένες τιμές στην ανάλυση Παλινδρόμησης

2.1 *Ορισμός απομακρυσμένων τιμών*

Μία απομακρυσμένη ή ακραία ή απομονωμένη παρατήρηση (outlier), είναι «η παρατήρηση που φαίνεται να αποκλίνει σημαντικά από το σύνολο των δεδομένων του δείγματος στο οποίο εμφανίζεται» όπως σημειώνει ο Grubbs (1969) (Barnett and Lewis, 1994).

Σύμφωνα με τον Hawkins (1980): «Έκτοπη είναι μία παρατήρηση που αποκλίνει σημαντικά πολύ από τις άλλες παρατηρήσεις, ώστε να μας υποψιάζει ότι δημιουργήθηκε από διαφορετικό μηχανισμό»

Οι Rousseeuw και Leroy (1987), αναφέρουν πως υπάρχουν τρεις διαφορετικές πηγές μεταβλητότητας και κατατάσσουν τις απομακρυσμένες παρατηρήσεις σε τρεις διαφορετικές κατηγορίες:

1. Εγγενής διακύμανση (inherent variability): Η μεταβλητότητα των τιμών είναι μια φυσιολογική ιδιότητα ενός πληθυσμού. Δεν ελέγχεται. Αντανακλάει την κατανομή ενός ορθού βασικού μοντέλου.
2. Σφάλμα μέτρησης: Ελλείψεις στο όργανο μέτρησης, στρογγυλοποίηση των τιμών, ή λάθη καταγραφής.
3. Σφάλμα εκτέλεσης: Αν απρόσεκτα επιλέξουμε ένα μη αντιπροσωπευτικό δείγμα (μεροληπτικό δείγμα) ή λάβουμε παρατηρήσεις που δεν αντιπροσωπεύουν τον πληθυσμό από τον οποίο προσπαθούμε να αντλήσουμε πληροφορίες, προκύπτει αυτός ο τύπος μεταβλητότητας.

Η απόφαση σχετικά με το αν μία παρατήρηση είναι απομακρυσμένη λαμβάνεται με βάση διάφορα κριτήρια, όπως είδαμε παραπάνω, και την κρίση του ερευνητή. Η απομακρυσμένη τιμή δείχνει ένα σημείο των δεδομένων που δεν είναι καθόλου αντιπροσωπευτικό όπως τα υπόλοιπα δεδομένα και θα πρέπει να εξεταστεί προσεκτικά. Η απευθείας απόρριψη απομακρυσμένων τιμών δεν είναι πάντα μία σωστή διαδικασία. Κάποιες φορές οι απομακρυσμένες τιμές περιέχουν πληροφορίες που δεν μπορούμε να αντλήσουμε από τα άλλα δεδομένα και χρειάζεται λοιπόν

παραπάνω διερεύνηση. Θα πρέπει να απορρίπτονται μόνο όταν είναι λάθη αντιγραφής, απρόσεκτη παρατήρηση, μηχανικού σφάλματος ή άλλα ανθρώπινα λάθη.

Στο διάγραμμα που ακολουθεί βλέπουμε πώς μία και μόνο απομακρυσμένη τιμή μπορεί να επηρεάσει την προσαρμογή με τη μέθοδο *ελαχίστων τετραγώνων.* Αντίθετα το ανθεκτικό μοντέλο δίνει μια πολύ καλή προσαρμογή στον όγκο των δεδομένων.

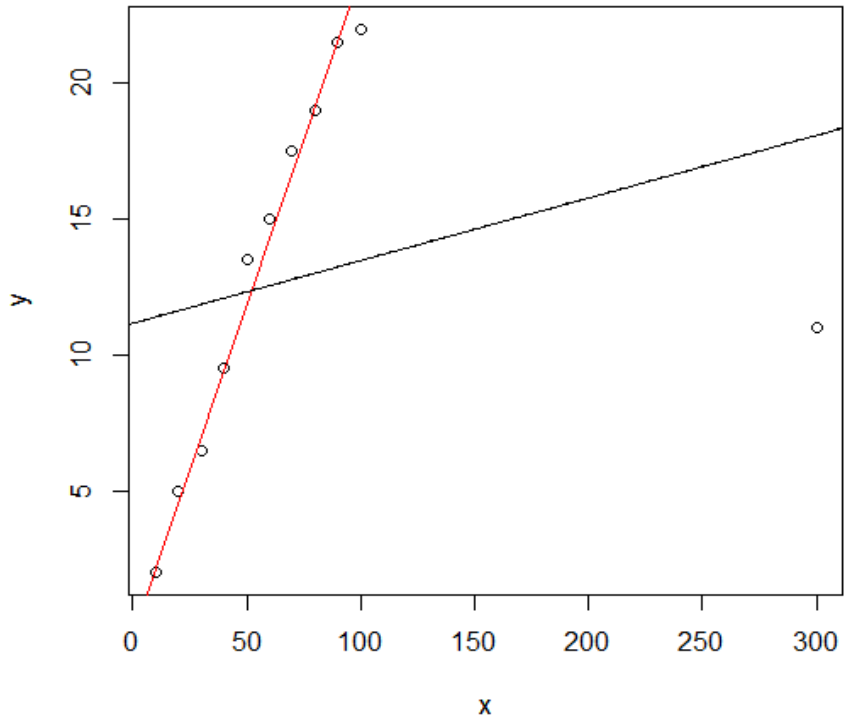
2.2 Ταξινόμηση και θέση των ακραίων παρατηρήσεων

Οι απομακρυσμένες παρατηρήσεις μπορούν να ταξινομηθούν σε:

i. Γενικά σφάλματα (gross errors) τα οποία μπορεί να έχουν προέλθει από λάθη στην αντιγραφή, στα δεκαδικά σημεία, στην εναλλαγή τιμών ή ακόμη να είναι μέλη μιας διαφορετικής κατανομής που εισήλθε στο δείγμα.
ii. Απομακρυσμένες τιμές (outliers) που οφείλονται σε αποτυχημένο μοντέλο.

Οι ακραίες αυτές τιμές μπορούν να συμβούν κατά τύχη σε οποιαδήποτε κατανομή αλλά είναι συχνά ενδεικτικές είτε ενός σφάλματος μέτρησης είτε ότι ο πληθυσμός προέρχεται από μία κατανομή με βαριά ουρά. Στην πρώτη περίπτωση κάποιος έχει τη δυνατότητα να τα απορρίψει ή να χρησιμοποιήσει ανθεκτικές μεθόδους, ενώ στη δεύτερη περίπτωση πρέπει κανείς να είναι πολύ προσεκτικός με τη χρήση μεθόδων που θεωρούν δεδομένη μια Κανονική κατανομή. Όταν λοιπόν υπάρχουν ακραίες παρατηρήσεις στα δεδομένα, οι κλασικές γραμμικές μέθοδοι έχουν συχνά χαμηλή αποτελεσματικότητα οπότε η ανθεκτική στατιστική (Robust statistics) έχει στόχο να παρέχει μεθόδους που περιγράφουν την πλειοψηφία των δεδομένων και δεν επηρεάζονται από ακραίες τιμές.

Από το παρακάτω Διάγραμμα 2.1 αντιλαμβανόμαστε την αδυναμία της μεθόδου ελαχίστων τετραγώνων να αντιμετωπίσει μία και μόνο ακραία παρατήρηση. Πιο συγκεκριμένα παρατηρούμε ότι η ανθεκτική γραμμή δίνει μια πολύ καλή προσαρμογή στον όγκο των δεδομένων, σε αντίθεση με την γραμμή ελαχίστων τετραγώνων η οποία επηρεάζεται από την ακραία παρατήρηση.



Διάγραμμα 2.1 Επίδραση ακραίων τιμών στην μέθοδο Ελαχίστων Τετραγώνων.

Οι απομακρυσμένες τιμές περιγράφονται τόσο από τη θέση τους όσο και από την επιδρασή τους. Όσον αφορά τη θέση τους μπορούν να παρατηρηθούν είτε στην κατεύθυνση του άξονα x είτε στην κατεύθυνση του άξονα y, είτε και στις δύο κατευθύνσεις ταυτόχρονα.

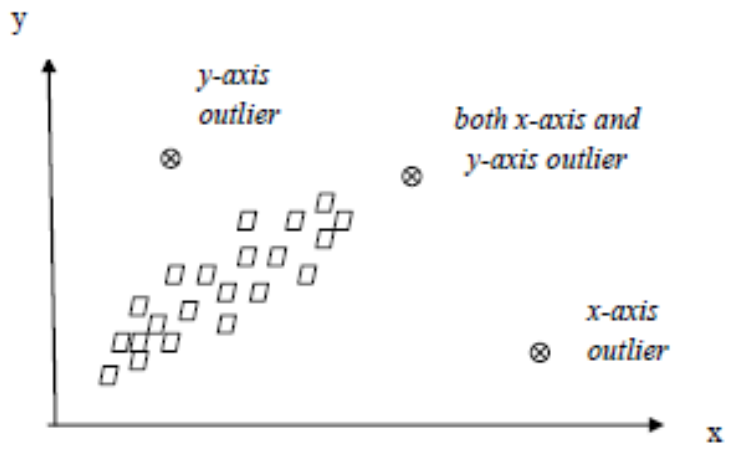
Ακραίες τιμές στην Y-κατεύθυνση

Οι ακραίες τιμές στην παλινδρόμηση μπορούν να καταστρέψουν σε σημαντικό βαθμό μια ανάλυση με την κλασσική μέθοδο *ελαχίστων τετραγώνων*. Οι ακραίες παρατηρήσεις μπορεί να είναι και στις δύο κατευθύνσεις X και Y. Αυτές που είναι στη y-κατεύθυνση θέλουν ιδιαίτερη προσοχή γιατί συχνά κανείς θεωρεί την y_i σαν παρατήρηση και την x_i σαν σταθερό αριθμό (η οποία ισχύει μόνο όταν ο σχεδιασμός έχει δοθεί εκ των προτέρων) και επειδή τέτοιες 'κάθετες' ακραίες παρατηρήσεις (vertical outliers) συχνά έχουν μεγάλα θετικά ή μεγάλα αρνητικά υπόλοιπα. Ακόμη και στην πολλαπλή ανάλυση παλινδρόμησης με ένα μεγάλο πλήθος ανεξάρτητων μεταβλητών, όπου είναι τόσο δύσκολο να απεικονιστούν τα δεδομένα, τέτοιες κάθετες ακραίες παρατηρήσεις μπορούν συχνά να παρατηρηθούν από τις γραφικές παραστάσεις των υπολοίπων (Rousseeuw and Leroy, 1987). Οι παρατηρήσεις αυτές

βρίσκονται εμφανώς μακριά από την γραμμική σχέση που καθορίζεται από το σύνολο των δεδομένων.

Ακραίες τιμές στην Χ-κατεύθυνση

Οι επεξηγηματικές μεταβλητές μπορούν επίσης να έχουν ακραίες τιμές. Σε πολλές εφαρμογές, κάποιος λαμβάνει μία λίστα από μεταβλητές και στη συνέχεια έχει να επιλέξει μια μεταβλητή απόκρισης και κάποιες επεξηγηματικές μεταβλητές. Μια ακραία παρατήρηση στην Χ-κατεύθυνση επηρεάζει σε μεγάλο βαθμό την γραμμή παλινδρόμησης της μεθόδου *ελαχίστων τετραγώνων* διότι τραβάει τη γραμμή της μεθόδου κοντά της. Ως εκ τούτου, καλείται σημείο μόχλευσης (leverage point). Γενικότερα, μια παρατήρηση (x_k, y_k) καλείται σημείο μόχλευσης, όποτε η x_k βρίσκεται αρκετά μακριά από τον κύριο όγκο των x_i στο δείγμα. Όπως φαίνεται από το *Διάγραμμα 2.2*, οι x-ακραίες παρατηρήσεις χωρίζονται σε δύο κατηγορίες, στα *θετικά σημεία μόχλευσης* (good leverage points) και στα *αρνητικά σημεία μόχλευσης (bad leverage points).* Σημειώνεται ότι η τιμή y_k δε λαμβάνεται υπόψη και κατά συνέπεια το σημείο (x_k, y_k) δεν αποτελεί απαραιτήτως ένα απομακρυσμένο σημείο σε σχέση με την ευθεία της παλινδρόμησης. Όταν το σημείο *(x_k , y_k)* είναι κοντά στην γραμμή παλινδρόμησης που προκύπτει από την πλειοψηφία των δεδομένων, τότε είναι ένα 'θετικό' σημείο μόχλευσης. Σ' αυτήν την περίπτωση, μπορεί να προσαρμόζεται τέλεια στην γραμμή παλινδρόμησης και να είναι ακόμη χρήσιμη διότι θα περιορίζει τα όρια εμπιστοσύνης (Rousseeuw and Leroy, 1987). Εάν ένα θετικό σημείο μόχλευσης απομακρυνθεί από τις παρατηρήσεις, η γραμμή των ελαχίστων τετραγώνων δε μεταβάλλεται σημαντικά. Αντίθετα, τα αρνητικά σημεία μόχλευσης δε συμφωνούν με την πληθώρα των παρατηρήσεων και μεταβάλλουν σημαντικά τη γραμμή. Τα τρία σύνολα ακραίων παρατηρήσεων φαίνονται στο Διάγραμμα 2.2.



Διάγραμμα 2.2 Τα τρία βασικά σύνολα ακραίων παρατηρήσεων (Rousseeuw and Leroy, 1987).

Κεφάλαιο 3

Μέτρα ανθεκτικότητας

3.1 Εισαγωγή

Στο κεφάλαιο αυτό θα αναφέρουμε μέτρα ανθεκτικότητας για τον προσδιορισμό του βαθμού ανθεκτικότητας μιας διαδικασίας. Τα πιο γνωστά μέτρα είναι η Συνάρτηση Επιρροής (Influence Function), το Σημείο Κατάρρευσης (Breakdown Point), η Καμπύλη Ευαισθησίας (Sensitivity Curve) και η Ευαισθησία Μεγάλου Σφάλματος (Gross Error Sensitivity). Τα μέτρα αυτά αλληλοσυμπληρώνονται, καθώς το κάθε ένα περιγράφει διαφορετικά χαρακτηριστικά.

Πολλές από τις υποθέσεις που γίνονται στη στατιστική όπως η κανονικότητα, η γραμμικότητα και η ανεξαρτησία, προσεγγίζουν την πραγματικότητα. Παρ' όλο που χρησιμοποιούμαι βέλτιστες διαδικασίες κάτω από αυτές τις υποθέσεις, μια μικρή απόκλιση από το μοντέλο μπορεί να επηρεάσει σοβαρά τα αποτελέσματα. Στόχος λοιπόν, της ανθεκτικής στατιστικής είναι να:

i) Περιγράψει τη δομή στην οποία προσαρμόζεται καλύτερα ο όγκος των δεδομένων.
ii) Να αναγνωρίσει τις ακραίες παρατηρήσεις και να προειδοποιήσει για τα σημεία μόχλευσης (Hampel et al., 1986).

3.2 Συναρτησιακές

Ένα παραμετρικό μοντέλο αποτελείται από μια οικογένεια κατανομών πιθανότητας F_θ στον χώρο του δείγματος, όπου η άγνωστη παράμετρος θ ανήκει στο ανοικτό κυρτό σύνολο ΘR.

Στην θεωρία της ανθεκτικότητας παραδεχόμαστε λοιπόν ότι το μοντέλο F_θ είναι μόνο μια ιδανική προσέγγιση στην πραγματικότητα και προφανώς επιθυμούμε μεθόδους οι οποίες συμπεριφέρονται αρκετά καλά έστω και αν υπάρχουν αποκλίσεις από το υποθετικό μοντέλο. Έστω $(x_1,\ x_2, \dots, x_n)$ τυχαίο δείγμα το οποίο προέρχεται από την κατανομή F_θ και έχει εμπειρική συνάρτηση κατανομής G_n. Η εκτίμηση του θ γίνεται μέσω του στατιστικού

$$T_n = T_n(x_1, x_2, \dots, x_n) = T(G_n).$$

Επιπλέον θεωρούμε ότι οι εκτιμήτριες είναι συναρτησιακές, δηλαδή $T_n(G_n) = T(G_n)$.

Τις εκτιμήτριες μπορούμε να τις αντιμετωπίσουμε ως συναρτησιακές, καθώς αυτό συνεισφέρει στη μελέτη ιδιοτήτων των εκτιμητριών, κυρίως σε σχέση με την ασυμπτωτική τους συμπεριφορά.

3.2.1 Συνάρτηση Επιρροής (Influence Function, IF)

Η συνάρτηση επιρροής καθιερώθηκε από τον Hampel για να διερευνήσει την απειροστή συμπεριφορά της συνάρτησης T(G). Εάν η εκτιμήτρια T για το θ_0 είναι συνεχής στην F_{θ_0} τότε σύμφωνα με τον Huber (1977) και τον Hampel (1974) μπορούμε να θεωρήσουμε το ενδεχόμενο προσέγγισης της T κοντά στην F_{θ_0} με την παραγώγιση

$$\frac{\partial}{\partial t} T\left((1-t)F_{\theta_0} + tG\right)_{t=0}$$

Η κατανομή $\boldsymbol{(1-t)F_{\theta_0} + tG}$ η οποία εμφανίζεται εδώ μπορεί να ερμηνευθεί ως εξής: Υποθέτουμε ότι η παρατήρηση x_i ακολουθεί την κατανομή F_{θ_0} με πιθανότητα 1-t και με πιθανότητα t την κατανομή G. Οπότε η κατανομή που προκύπτει για την x_i είναι $\boldsymbol{(1-t)F_{\theta_0} + tG}$. Για κάθε κατανομή G στο πεδίο ορισμού της T ισχύει

$$\frac{\partial}{\partial t} T\left((1-t)F_{\theta_0} + tG\right)_{t=0} = \int IF(x, T, F_{\theta_0}) dG(x),$$

όπου $IF(T, F_{\theta_0}, x)$ είναι η συνάρτηση επιρροής για την T στην F_{θ_0}.

Εάν δ_x είναι η μονάδα μάζας στο x τότε προκύπτει ότι

$$\frac{\partial}{\partial t}\left((1-t)F_{\theta_0} + t\delta_x\right)_{t=0} = IF(T, F_{\theta_0}, x)$$

και η IF μπορεί να υπολογιστεί αναλυτικότερα απ' αυτόν τον τύπο.

Για έναν απλούστερο ορισμό, υποθέτουμε ότι συνάρτηση $IF: R^P \to R^P$ (η οποία εξαρτάται από την εκτιμήτρια και την πραγματική συνάρτηση κατανομής στο χώρο R^P) ικανοποιεί τη σχέση

$$\sqrt{n}[T(F_n) - T(F) - \frac{1}{n}\sum_{i=1}^{n} IF(x_i, T, F)] \to 0$$

πιθανολογικά καθώς το $n \to \infty$. Εφ' όσον $IF(x, T, F)$ παριστάνει το αποτέλεσμα μιας παρατήρησης στην τιμή θ, η IF καλείται συνάρτηση επίδρασης της εκτιμήτριας. Κάτω από λογικές συνθήκες η κυριότερη ιδιότητα της IF είναι η ακόλουθη

$$\int IF(x, T, F)dF(x) = 0.$$

Η συνάρτηση επίδρασης (IF) είναι η κυριότερη μέτρηση ανθεκτικότητας και διατυπώνει μ' άλλα λόγια τη μεροληψία, που προκαλείται από μια απομακρυσμένη παρατήρηση.

3.2.2 Σημείο Κατάρρευσης (Breakdown Point, BP)

Αρκεί μία και μόνο απομακρυσμένη τιμή για να καταστρέψει την εκτιμήτρια ελαχίστων τετραγώνων. Οπότε, εύκολα προκύπτει ότι η εκτιμήτρια ελαχίστων τετραγώνων δεν θεωρείται ανθεκτική. Όμως θα δούμε παρακάτω ότι υπάρχουν εκτιμήτριες που αντέχουν σε δεδομένα τα οποία περιέχουν ένα συγκεκριμένο ποσοστό απομακρυσμένων τιμών. Προκειμένου να υπολογιστεί το μέγεθος του ποσοστού αυτού εισήχθη το σημείο κατάρρευσης. Ας θεωρήσουμε ένα δείγμα από n σημεία δεδομένων $X = (x_{i1}, x_{i2}, \dots, x_{in}, y_i)$ για $i = 1,2, \dots, n$ και έστω T μια εκτιμήτρια για το διάνυσμα των αγνώστων παραμέτρων θ για το οποίο ισχύει $T(X) = \hat{\theta}$ όταν εφαρμοστεί στο δείγμα X. Ας θεωρήσουμε τώρα όλα τα αλλοιωμένα δείγματα X' τα οποία προκύπτουν αντικαθιστώντας οποιεσδήποτε m παρατηρήσεις με τυχαίες τιμές. Τότε η μέγιστη μεροληψία που μπορεί να προκληθεί από μία τέτοια αλλοίωση σε όλα τα πιθανά X' δίνεται από τη σχέση

$$bias(m; T, X) = sup\|T(X') - T(X)\|.$$

Όταν η μεροληψία $bias(m; T, X)$ τείνει στο άπειρο, αυτό σημαίνει ότι οι m απομακρυσμένες τιμές έχουν αυθαίρετα μεγάλη επίδραση στην εκτιμήτρια T οπότε η εκτιμήτρια καταρρέει.

Έτσι, το σημείο κατάρρευσης της εκτιμήτριας T στο δείγμα X ορίζεται ως

$$\epsilon_n^*(T, X) = \min\{\frac{m}{n}; bias(m; T, X) = \infty\}.$$

Το σημείο κατάρρευσης λοιπόν, είναι με άλλα λόγια το μικρότερο κλάσμα μόλυνσης που προκαλεί την εκτιμήτρια T να πάρει αυθαίρετα μακρινές τιμές από την εκτιμήτρια $T(X)$. Το σημείο κατάρρευσης είναι ίσως το σημαντικότερο μέτρο ανθεκτικότητας. Μπορεί να βρεθεί για όλους τις εκτιμήτριες και ισχύει $\epsilon^* \leq \frac{1}{2}$ αφού η μόλυνση σε ένα δείγμα δεν μπορεί πρακτικά να υπερβεί το 50%, τιμή πάνω από την οποία είναι αδύνατος ο διαχωρισμός των καλών από τις απομακρυσμένες τιμές.

Στα ελάχιστα τετράγωνα το σημείο κατάρρευσης είναι ίσο με $\frac{1}{n}$ το οποίο τείνει να γίνει μηδέν καθώς το μέγεθος του δείγματος αυξάνεται. Δηλαδή, μία και μόνο απομακρυσμένη τιμή είναι αρκετή για να καταστρέψει την εκτιμήτρια της μεθόδου ελαχίστων τετραγώνων.

3.2.3 Ευαισθησία σε μεγάλα σφάλματα (Gross Error Sensitivity, GES)

Η ευαισθησία σε μεγάλο σφάλμα είναι ένα μέτρο ανθεκτικότητας, που προκύπτει από τη συνάρτηση επίδρασης και προσδιορίζει άμεσα το βαθμό ανθεκτικότητας μιας εκτιμήτριας. Περιγράφει τη μέγιστη επίδραση που προκαλεί στην τιμή μιας εκτιμήτριας, μία μικρή αλλοίωση της κατανομής και θα πρέπει να είναι πεπερασμένη.

Η ευαισθησία σε μεγάλα σφάλματα μιας εκτιμήτριας T για δοσμένη κατανομή F, ορίζεται ως εξής

$$\gamma^*(T, F) = sup|IF(x; T, F)|$$

όπου το $supremum$ ορίζεται για όλα τα x, για τα οποία υπάρχει η συνάρτηση επίδρασης $IF(x; T, F)$. Ο καθορισμός ενός ορίου για το γ^* αποτελεί το πρώτο βήμα για να κάνουμε μια εκτιμήτρια πιο ανθεκτική και αυτό συχνά έρχεται σε σύγκρουση με το στόχο της ασυμπτωτικής αποτελεσματικότητας (βλέπε Hampel et al., 1986).

3.2.4 Καμπύλη Ευαισθησίας (Sensitivity Curve, SC)

Η καμπύλη ευαισθησίας (SC) μετράει την επίδραση που θα έχει στην εκτιμήτρια μία και μόνο επιπρόσθετη παρατήρηση x.

Έστω το τυχαίο δείγμα $x_1, x_2, \dots, x_n$ τότε η SC μιας εκτμήτριας T_n σε ένα σημείο x ορίζεται ως εξής:

$$SC_n(X;T_n) = n[T_n(x_1, x_2, \dots, x_{n-1}, x) - T_{n-1}(x_1, x_2, \dots, x_{n-1})].$$

Στην περίπτωση που $T_n(x_1, x_2, \dots, x_n) = T(F_n)$ για κάθε n, με την αντίστοιχη εμπειρική κατανομή F_n τότε:

$$SC(x) = \frac{T\left(\frac{n-1}{n}F_{n-1} + \frac{1}{n}\Delta x\right) - T(F_{n-1})}{\frac{1}{n}}$$

Όπου F_{n-1} είναι η εμπειρική κατανομή του $(x_1, x_2, \dots, x_{n-1})$.

Κεφάλαιο 4

Ανθεκτικές Εκτιμήτριες Παλινδρόμησης

4.1 Βασικές ανθεκτικές εκτιμήτριες

Η πιο συνηθισμένη διαδικασία στη θεωρία της στατιστικής εκτίμησης είναι ο προσδιορισμός ενός παραμετρικού μοντέλου, το οποίο υποθέτουμε ότι ισχύει, και στη συνέχεια η εύρεση της βέλτιστης εκτιμήτριας για τις άγνωστες παραμέτρους. Η ανάλυση παλινδρόμησης ασχολείται με την ανακάλυψη εκτιμητριών που επηρεάζονται όσο το δυνατόν λιγότερο από την ύπαρξη απομακρυσμένων παρατηρήσεων. Πολλές μέθοδοι έχουν αναπτυχθεί στη θεωρία των πολυδιάστατων κατανομών. Ωστόσο, οι βασικότερες εκτιμήτριες είναι οι Μ-εκτιμήτριες, οι εκτιμήτριες υψηλής κατάρρευσης και οι συνδυασμοί των δύο. Συγκεκριμένα, θα παρουσιαστούν οι ακόλουθες εκτιμήτριες:

- L-εκτιμήτριες
- M-εκτιμήτριες
- R-εκτιμήτριες
- Εκτιμήτριες Ελαχίστων Διαμέσων Τετραγώνων (LMS)
- Εκτιμήτριες Ελαχίστων Αποκοπτόμενων Τετραγώνων (LTS)
- S-εκτιμήτρια
- MM-εκτιμήτρια

4.2 Μέθοδοι ανθεκτικής παλινδρόμησης

4.2.1 L-εκτιμήτριες

Μια εκτιμήτρια που υπολογίζεται από έναν γραμμικό συνδυασμό στατιστικών στοιχείων μπορεί να χαρακτηριστεί L-εκτιμήτρια. Ο πρώτη $L_1 - εκτιμήτρια$ που προτάθηκε και αντικατέστησε την εκτιμήτρια ελαχίστων τετραγώνων, είναι η εκτιμήτρια ελαχίστων απόλυτων τιμών. Επίσης, γνωστή και ως L_1 παλινδρόμηση επειδή ελαχιστοποιεί το $L_1 - norm$ (δηλαδή το άθροισμα απολύτων αποκλίσεων), η *LAD*-εκτιμήτρια αποτελεί την απλούστερη και παλαιότερη προσέγγιση της

ανθεκτικής στατιστικής. Μελετήθηκε πιο αναλυτικά από τον Boscovic (1755), τον Laplace (1793) και στα επόμενα χρόνια που ακολούθησαν χρησιμοποιήθηκε από τον Edgeworth (1887).

Η *LAD* παλινδρόμηση είναι πολύ ανθεκτική σε δεδομένα με ασυνήθιστες τιμές *y*. Οι εκτιμήσεις βρίσκονται ελαχιστοποιώντας το άθροισμα των απολύτων τιμών των υπολοίπων, δηλαδή ελαχιστοποιώντας την ποσότητα

$$\sum_{i=1}^{n} |r_i|$$

όπου $r_i = y_i - \hat{y}_i$.

Επίσης, γενικεύεται σαν α-παλινδρόμηση ποσοστιαίου σημείου (a-regression quantile) και η συνάρτηση που θα πρέπει να ελαχιστοποιηθεί είναι

$$\sum_{i=1}^{n} \rho_\alpha(r_i),$$

όπου

$$\rho_\alpha(r_i) = \begin{cases} \alpha r_i, & \alpha\nu\ r_i \geq 0 \\ (a-1) r_i, & \alpha\nu\ r_i < 0 \end{cases}$$

Η L_1-εκτιμήτρια δίνεται για α=0.5.

Παρά το γεγονός ότι στην εκτίμηση θέσης μονομεταβλητής περίπτωσης η *LAD* εκτιμήτρια επηρεάζεται λιγότερο σε σχέση με την εκτιμήτρια ελαχίστων τετραγώνων από ακραίες τιμές στην κατεύθυνση του *y*, στην πολλαπλή παλινδρόμηση είναι ευαίσθητη και αποτυγχάνει να αντιμετωπίσει τα αρνητικά σημεία μόχλευσης. Το σημείο κατάρρευσής της είναι 0% και έχει χαμηλή αποδοτικότητα.

Για την L1-εκτιμήτρια, γενικώς ισχύουν τα ακόλουθα:

- Η L1-εκτιμήτρια είναι η εκτιμήτρια μέγιστης πιθανοφάνειας, εάν η ακραίες τιμές είναι ανεξάρτητες με μία διπλή-εκθετική κατανομή.
- Οι υπολογισμοί δεν είναι τόσο εύκολοι όπως με την εκτιμήτρια ελαχίστων τετραγώνων, καθώς για την L1 παλινδρόμηση απαιτείται μια γραμμική προγραμματιστική λύση.

- Η L1 είναι ίσης μεταβολής (equivariant), που σημαίνει ότι η αντικατάσταση της **y** με a+b**y** και της **X** με **A**+$\boldsymbol{B}^{-1}$**X** (όπου a, b, **A**, **B** είναι σταθερές) θα αφήσει τη τελική λύση ουσιαστικά αμετάβλητη.
- Το σημείο κατάρρευσης της L1-εκτιμήτριας μπορεί να αποδειχθεί ότι είναι 1-1/$\sqrt{2}$ ≈0.29, έτσι μπορεί να αντέξει το 29% ‘‘κακών’’ δεδομένων (Fox and Weisberg, 2010).

4.2.2 Μ-εκτιμήτριες

Στην ανθεκτική στατιστική, οι *Μ-εκτιμήτριες* αποτελούν μια γενικευμένη κλάση εκτιμητριών που παράγονται από την ελαχιστοποίηση του αθροίσματος διαφόρων συναρτήσεων. Ο Huber (1973) εισήγαγε τη χρήση των *Μ*-εκτιμητριών, οι οποίοι αποτελούν γενίκευση των εκτιμητριών *μέγιστης πιθανοφάνειας* (maximum likelihood estimator) και από τις οποίες προέρχεται το ονομά τους. Επίσης, συνιστούν από τις αρχικές προσπάθειες συνδυασμού της αποτελεσματικότητας της μεθόδου ελαχίστων τετραγώνων και της ανθεκτικότητας των εκτιμητριών *ελαχίστων απολύτων αποκλίσεων* (*Least Absolute Deviations, LAD*), μέθοδοι οι οποίες είναι ειδικές περιπτώσεις *Μ*-εκτιμητριών και οι δύο.

Παρά το γεγονός ότι χρησιμοποιούνται ευρέως στην ανάλυση δεδομένων για την οποία μπορεί να υποτεθεί ότι η αλλοίωση είναι κυρίως στην κατεύθυνση των y_i, οι *Μ*-εκτιμήτριες θεωρούνται ‘‘ευαίσθητες’’ αναφορικά με τα σημεία μόχλευσης. Συνεπώς, το σημείο κατάρρευσης (BP) των *Μ-εκτιμητριών* είναι 0%, στα σημεία αυτά.

Κάθε λοιπόν, εκτιμήτρια T_n που ορίζεται από το πρόβλημα ελαχιστοποίησης της αντικειμενικής συνάρτησης

$$\sum_{i=1}^{n} \rho(X_i, T_n)$$

ή από την έμμεση εξίσωση

$$\sum_{i=1}^{n} \psi(X_i, T_n) = 0,$$

όπου ρ είναι μια αυθαίρετη συνάρτηση και έχει ως παράγωγο $\psi(x,\beta) = \frac{\partial}{\partial\beta}\rho(x,\beta)$, ονομάζεται *M-εκτιμήτρια*.

Έστω G_n η εμπειρική αθροιστική συνάρτηση που προέρχεται από τα δείγμα, τότε η λύση T_n της τελευταίας εξίσωσης μπορεί επίσης να γραφεί σαν $T(G_n)$ και T το συναρτησιακό που δίνεται από τη σχέση

$$\int \psi\big(x, T(G)\big) dG(x) = 0$$

για όλες τις κατανομές G με $F_{t,x} = (1-t)F + t\Delta_x$ και στη συνέχεια θα παραγωγίσουμε ως προς *t*. Οπότε παίρνουμε:

$$\int \psi\big(x, T(F)\big) d(\Delta_x - F) + \int \frac{\partial}{\partial\beta}[\psi(y;\beta)]_{T(F)} dF(x) \frac{\partial}{\partial t}\big[T\big(F_{t,x}\big)\big]_{t=0} = 0.$$

Κάνοντας τώρα χρήση του ορισμού της συνάρτησης επίδρασης και των παραπάνω σχέσεων, η συνάρτηση επίδρασης μιας *M-εκτιμήτριας* δίνεται από τη σχέση

$$IF(x;F,T) = \frac{\psi[x,T(F)]}{-\int \frac{\partial}{\partial\beta}[\psi(y;\beta)]_{T(F)}\, dF(y)}$$

με ασυμπτωτική διασπορά

$$V(T,F) = \frac{\int \psi^2[x,T(F)] dF(x)}{[\int \frac{\partial}{\partial\beta}[\psi(y;\beta)]_{T(F)}\, dF(y)]^2}$$

4.2.2.1 Διτετράγωνη εκτιμήτρια (Bisquare estimator)

Μία συνηθισμένη *M-εκτιμήτρια* είναι η *διτετράγωνη εκτιμήτρια* (*bisquare estimator*), όπου η συνάρτηση *ρ* είναι ανθεκτική και πεπερασμένη,

$$\rho(r) = \min\{1 - (1-r^2)^3, 1\}$$

και *δ=0,5*.

Συχνά χρησιμοποιούμε μια συνάρτηση ρ (less rapidly increasing), η οποία είναι τετραγωνικής μορφής κοντά στην αρχή, π.χ. $\rho'(0) = 0$ και $\rho''(0) > 0$, και σε αυτές

τις περιπτώσεις μια Μ-εκτιμήτρια μπορεί να παριστάνει μια εκτίμηση των μέσων x_i^2 με βάρη.

Η συνάρτηση βάρους ορίζεται ως

$$W(r) = \begin{cases} \rho(r)/r^2 \ \varepsilon\acute{\alpha}\nu\ r \neq 0 \\ \rho''(0) \ \varepsilon\acute{\alpha}\nu\ r = 0 \end{cases}$$

Για τη διτετράγωνη συνάρτηση ρ έχουμε

$$W(r)=min\ 3-3r^2+r^4,\ 1/r^2$$

όπου στα μεγαλύτερα *x* αντιστοιχεί μικρότερο βάρος (Fox, J. and Weisberg, S., 2010).

4.2.2.2 Μοντέλο γραμμικής παλινδρόμησης του Huber

Ο Huber, το 1973, επέκτεινε τα αποτελέσματα του για την ανθεκτική εκτίμηση μίας παραμέτρου θέσης στην περίπτωση της γραμμικής παλινδρόμησης. Συγκεκριμένα, οι *Μ-εκτιμήτριες* μπορούν να προκύψουν σαν μία λύση του ακόλουθου τύπου ελαχιστοποίησης:

$$\sum_{i=1}^{n}[\rho\left(\frac{y_i - x_i^T\beta}{\sigma}\right) + A]\sigma \qquad (4.2.2.1)$$

Παραγωγίζοντας ως προς β και σ και στη συνέχεια θέτοντας ίσο με μηδέν, παίρνουμε αντίστοιχα τις παρακάτω εξισώσεις:

$$\sum_{i=1}^{n}\psi\left(\frac{y_i - x_i^T T_n}{\hat{\sigma}}\right)x_i = 0$$

$$\sum_{i=1}^{n}\chi\left(\frac{y_i - x_i^T T_n}{\hat{\sigma}}\right) = 0$$

όπου $\psi(r)=\rho'(r)$ και $\chi(r)=r\psi(r)-\rho(r)-A$.

Ακολουθώντας τους Hampel et al. (1986), η συνάρτηση επίδρασης της Huber εκτιμήτριας T^{Hu} με συνάρτηση κατανομής $F_\beta(x,y)$ και πυκνότητα $f_\beta(x,y) = \varphi(y - x^T\beta)k(x)$, (όπου *k(x)* η σ.π.π. του *x*), δίνεται από τη σχέση:

$$IF(x, y; T^{Hu}, F_\beta) = \psi_c(y - x^T\beta)M^{-1}x$$

όπου:

$$M = (E\psi_c')(Exx^T) = (\int \psi_c'(r)d\Phi(r))(\int xx^T\, dK(x))$$

και $K(x)$ η συνάρτηση κατανομής του *x*.

4.2.2.3 Ο αλγόριθμος των M-εκτιμητριών

Ο τυπικός αλγόριθμος για την ελαχιστοποίηση της σχέσης (4.2.1.1) έχει σαν πλεονέκτημα ότι είναι απλός, παρόλο που συγκλίνει με αργό ρυθμό. Υποθέτουμε ένα επίπεδο ανεκτικότητας *ε*>0 και αρχικές τιμές, $\sigma^{(0)}$. Αν οι προβλεπόμενες τιμές $x_i^T\beta$ είναι γραμμικές, μπορούμε να χρησιμοποιήσουμε την εκτιμήτρια ελαχίστων τετραγώνων σαν αρχική τιμή. Ακολουθεί ο αλγόριθμος των *Huber* και *Dutter* (1974):

1. Παίρνουμε για αρχικές τιμές $\beta^{(0)}$ την τιμή του εκτιμητή ελαχίστων τετραγώνων.
2. Υπολογίζουμε τα υπόλοιπα $r_i^m = y_i - x_i^T\beta^{(m)}$ όπου *i=1, 2,..., n*.
3. Υπολογίζουμε μια καινούργια τιμή για το *σ*

$$(\sigma^{(m+1)})^2 = \frac{1}{A}\sum \chi(\frac{r_i}{\sigma^{(m)}})(\sigma^{(m)})^2,$$

*ό*που *A* είναι ο διορθωτικός παράγοντας για να πετύχουμε αμεροληψία.

4. Εκτιμούμε τα υπόλοιπα

$$r_i^* = \psi\left(\frac{r_i}{\sigma^{(m+1)}}\right)\sigma^{(m+1)}.$$

5. Υπολογίζουμε τις μερικές παραγώγους

$$x_{ik} = \frac{\partial}{\partial\beta}(x_i^T\beta).$$

6. Λύνουμε ως προς $\hat{\tau}$ την $X^\tau X_{\hat{\tau}} = X^\tau r^*$. Προκύπτει από την λύση της

$$min\sum(r_i - \sum x_{ik}\,\hat{\tau}_k)^2$$

7. Θέτουμε

$$\beta^{(m+1)} = \beta^{(m)} + q\hat{\tau},$$

όπου 0< q <2 ένας αυθαίρετος σταθερός παράγοντας.

8. Σταματάει η επανάληψη, όταν οι εκτιμημένοι συντελεστές συγκλίνουν.
9. Αλλιώς, θέσε $m := m + 1$, πήγαινε στο βήμα 2.

4.2.3 R-εκτιμήτριες

Πολλές ανθεκτικές τεχνικές στατιστικής ανάλυσης βασίζονται στη σειρά κατάταξης (ranks) των δεδομένων. Αυτή η ιδέα δίνει άλλη μία προσέγγιση στην ανθεκτική παλινδρόμηση, μέσω των R-εκτιμητριών. Στην πολλαπλή παλινδρόμηση (Jaeckel, 1972), έστω R_i η σειρά κατάταξης των $r_i = y_i - x_i'\hat{\beta}$. Τότε η μέθοδος αποτελείται από την ελαχιστοποίηση της συνάρτησης

$$\sum_{i=1}^{n} a_n(R_i) r_i \ ,$$

ως προς β, όπου η μονότονη συνάρτηση βαθμολογίας (score function) $a_n(i)$ ικανοποιεί το $\sum_{i=1}^{n} a_n(R_i) = 0$. Συναρτήσεις που έχουν προταθεί είναι:

- Βαθμολογίες Wilcoxon: $a_n(i) = i - (n+1)/2$
- Βαθμολογίες Van der Waerden: $a_n(i) = \Phi^{-1}(i/(n+1))$
- Βαθμολογίες median: $a_n(i) = sgn(i - (n+1)/2))$
- Βαθμολογίες bounded normal: $a_n(i) = \min(c, \max\{\Phi^{-1}(i/(n+1)), -c\})$

(Rousseeuw and Leroy, 1987)

4.2.4 Εκτιμήτρια Ελαχίστων Διαμέσων Τετραγώνων (Least Median of Squares, LMS)

Όπως είδαμε στο πρώτο κεφάλαιο, η εκτιμήτρια ελαχίστων τετραγώνων προκύπτει από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των υπολοίπων. Μερικοί ερευνητές στην ιδέα να κάνουν αυτή την εκτιμήτρια πιο ανθεκτική προσπάθησαν να αντικαταστήσουν το άθροισμα με την πιο ανθεκτική, διάμεσο. Έτσι ο *Rousseeuw* (1984) πρότεινε την εκτιμήτρια της διαμέσου των ελαχίστων τετραγώνων, η οποία χάρη στο υψηλό σημείο κατάρρευσης, μπορεί να αντιμετωπίσει σχεδόν πάνω από τις μισές απομακρυσμένες τιμές, την ίδια στιγμή. Προκύπτει ότι η εκτιμήτρια αυτή είναι πολύ ανθεκτική με τις απομακρυσμένες τιμές στον άξονα των *x*, όσο και στον άξονα των *y* και ορίζεται ως εξής

$$\min median(r_i^2)$$

Μερικές από τις ιδιότητες της μεθόδου ελαχίστων διαμέσων τετραγώνων είναι:

- Υπάρχει πάντα μία λύση για την παραπάνω εκτιμήτρια που αναφέραμε.

- Το σημείο κατάρρευσης του LMS είναι 50% και ορίζεται ως

$$E_n^* = \frac{\left[\frac{n}{2}\right] - p + 2}{n},$$

όπου $[\frac{n}{2}]$ είναι ο μεγαλύτερος ακέραιος μικρότερος του $\frac{n}{2}$.

- Η εκτιμήτρια LMS ικανοποιεί και τις τρεις ιδιότητες ισοδυναμίας. Η εκτιμήτρια της μεθόδου στηρίζεται μόνο στα υπόλοιπα. Είναι δηλαδή, ίσης μεταβολής (equivariant, βλέπε παρ. 4.2.1) ως προς την παλινδρόμηση, ίσης μεταβολής ως προς την κλίμακα και ως προς την θέση και κλίμακα.

Το κύριο μειονέκτημα της μεθόδου LMS είναι ότι, η εκτιμήτρια έχει χαμηλή ασυμπτωτική αποτελεσματικότητα, πράγμα που αποδεικνύεται από τον αργό ρυθμό σύγκλισης της, που είναι $n^{-1/3}$. Συνεπώς, η εκτιμήτρια της μεθόδου δεν είναι ασυμπτωτικά κανονική. Γι' αυτόν τον αργό ρυθμό σύγκλισης ο Rousseeuw (1984), προτείνει τη χρήση μιας διαφορετικής αντικειμενικής συνάρτησης.

4.2.5 Εκτιμήτρια Ελαχίστων Αποκοπτόμενων Τετραγώνων (Least Trimmed sum of Squares, LTS)

Στη μέθοδο *LMS* που είδαμε παραπάνω, αντί να προσθέσουμε όλα τα τετραγωνικά σφάλματα όπως στην κλασική μέθοδο των ελαχίστων τετραγώνων, μπορούμε να επικεντρώσουμε το ενδιαφέρον στο αποκοπτόμενο ή περικεκομμένο άθροισμα τετραγώνων (*LTS*). Έτσι, επιτυγχάνεται η βελτίωση του αργού ρυθμού σύγκλισης της εκτιμήτριας *LMS*. Αρχικά λοιπόν, τα διατεταγμένα κατά αύξουσα σειρά, τετραγωνισμένα υπόλοιπα είναι

$$(r^2)_1 \leq (r^2)_2 \leq \cdots \leq (r^2)_n$$

και η εκτιμήτρια *LTS* δίνεται από τη σχέση

$$min \sum_{i=1}^{h} (r^2)_i$$

όπου $h=\frac{n+1}{2}$ αν n είναι άρτιος και $h=\left[\frac{n}{2}\right]+1$, αν n περιττός. Η μέθοδος αυτή (που αναπτύχθηκε επίσης από τον Rousseeuw, 1984) προσεγγίζει κατά πολύ τη μέθοδο των ελαχίστων τετραγώνων με τη διαφορά ότι τα μεγαλύτερα τετραγωνικά σφάλματα δεν χρησιμοποιούνται στην άθροιση.

Μερικές από τις ιδιότητες της μεθόδου ελαχίστων αποκοπτόμενων τετραγώνων είναι:

- Υπάρχει πάντα μία λύση για την παραπάνω εκτιμήτρια που αναφέραμε.
- Το σημείο κατάρρευσης της *LMS* είναι

$$E_n^* = \frac{\left[\frac{n-p}{2}\right] + 1}{n},$$

  δηλαδή η εκτιμήτρια των ελαχίστων αποκοπτόμενων τετραγώνων μπορεί να αντιμετωπίσει την επίδραση απομακρυσμένων τιμών σε ένα ποσοστό που αγγίζει το 50%.
- Η εκτιμήτρια είναι ίσης μεταβολής ως προς την παλινδρόμηση, ίσης μεταβολής ως προς την κλίμακα και ως προς την θέση και κλίμακα.
- Σε αντίθεση με το χαμηλό ρυθμό σύγκλισης της εκτιμήτριας LMS, η LTS συγκλίνει με ρυθμό $n^{-1/2}$ και έχει μεγαλύτερη ασυμπτωτική αποτελεσματικότητα. Αξίζει να σημειωθεί ότι έχει την ίδια ασυμπτωματική αποτελεσματικότητα στην Κανονική κατανομή με την M-εκτιμήτρια.

Βασικά μειονεκτήματα της παραπάνω μεθόδου είναι ότι:

- Η αντικειμενική συνάρτηση απαιτεί ταξινόμηση των τετραγωνικών σφαλμάτων, η οποία χρειάζεται πολύ περισσότερες επαναλήψεις συγκριτικά με τη μέθοδο LMS.
- Η αποτελεσματικότητα του παραμένει σχετικά χαμηλή, παρόλο που συγκλίνει γρηγορότερα από τη μέθοδο LMS.

4.2.6 S-εκτιμήτριες

Οι *Rousseeuw & Yohai* (1984) δημιούργησαν τις *S*-εκτιμήτριες με υψηλό σημείο κατάρρευσης και με ρυθμό σύγκλισης $n^{-1/2}$. Ορίζονται από την ελαχιστοποίηση της διασποράς των σφαλμάτων:

$$s(r_1(\beta), \dots, r_n(\beta))$$

με τελική εκτίμηση κλίμακας

$$\hat{\sigma} = s\left(r_1(\hat{\beta}), \dots, r_n(\hat{\beta})\right).$$

Η διασπορά $s\left(r_1(\hat{\beta}), \dots, r_n(\hat{\beta})\right)$ ορίζεται από τη λύση της

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i}{s}\right)=K,$$

όπου K είναι μια σταθερά η οποία ισούται με $E_{\Phi}[\rho]$ και Φ είναι η τυποποιημένη Κανονική κατανομή. Η συνάρτηση ρ πρέπει να ικανοποιεί τις ακόλουθες συνθήκες:

1. $\rho(0)=0$.
2. Η ρ είναι συμμετρική με συνεχή παράγωγο
3. Υπάρχει $c>0$ τέτοιο ώστε η ρ να αυξάνεται στο διάστημα *[0,c]* και να παραμένει σταθερή στο *[c,∞]*

Οι *S*-εκτιμήτριες είναι ίσης μεταβολής ως προς την παλινδρόμηση, ως προς κλίμακα και ως προς και την θέση και κλίμακα. Φαίνεται, επίσης, να έχουν την ίδια ασυμπτωτική συμπεριφορά με τις *M*-εκτιμήτριες. Επίσης όμως έχουν υψηλό σημείο κατάρρευσης που φθάνει μέχρι *50%* και δίνεται από τον τύπο

$$E^{*}=\frac{K}{\rho(c)}\ \text{για n}\to\infty$$

Το πρόβλημα που υπάρχει σ' αυτές τις εκτιμήτριες είναι ότι δεν μπορούν να πετύχουν ταυτόχρονα υψηλό σημείο κατάρρευσης μαζί με υψηλή αποτελεσματικότητα.

4.2.7 MM-εκτιμήτριες

Ο *Yohai* (1985) εισήγαγε τις *MM*-εκτιμήτριες, οι οποίες συνδυάζουν υψηλό σημείο κατάρρευσης και υψηλή αποτελεσματικότητα όταν τα σφάλματα προέρχονται από την Κανονική κατανομή. Οι εκτιμήτριες που είδαμε ως τώρα έχουν υψηλό σημείο κατάρρευσης, όμως δεν είναι ικανοποιητικά αποτελεσματικές. Μία MM-εκτιμήτρια $\hat{\beta}$, ορίζεται ως η λύση της σχέσης:

$$\sum_{i=1}^{n}\rho(\frac{y_i-x_i^T\beta}{s_n})$$

Όπου s_n είναι μια εκτίμηση κλίμακας και ρ είναι μια πραγματική συνάρτηση που ικανοποιεί τα παρακάτω:

i. η ρ είναι συμμετρική με συνεχή παράγωγο και $\rho(0)=0$.
ii. υπάρχει $\alpha>0$ τέτοιο ώστε η ρ να είναι γνησίως αύξουσα στο $[0,\alpha]$ και σταθερή στο $[\alpha,\infty)$.

iii. $\rho_1(u) \leq \rho_0(u)$.

Η MM-εκτιμήτρια ορίζεται από μία διαδικασία με τρία στάδια, ως εξής:

- Στο πρώτο στάδιο, υπολογίζουμε μία αρχική εκτιμήτρια T^0 του συντελεστή παλινδρόμησης β, η οποία είναι συνεπής και ανθεκτική με υψηλό σημείο κατάρρευσης αλλά δεν είναι υποχρεωτικά αποτελεσματική.
- Στο δεύτερο στάδιο, υπολογίζουμε τα υπόλοιπα $r_i(T^0)$ και στη συνέχεια η M-εκτιμήτρια κλίμακας $s_n = s(r(T^0))$, η οποία ορίζεται με τις ίδιες συνθήκες για τη συνάρτηση ρ.
- Στο τρίτο, η MM-εκτιμήτρια T^1 ορίζεται από οποιαδήποτε λύση της εξίσωσης

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(T^1)}{S(\beta)}\right) x_i = 0,$$

όπου $\psi=\rho_1'$ η οποία ικανοποιεί την εξίσωση

$$\sum_{i=1}^{n} \rho_1\left(\frac{r_i(T^1)}{S(\beta)}\right) \leq \sum_{i=1}^{n} \rho_1\left(\frac{r_i(T^0)}{S(\beta)}\right).$$

Η συνάρτηση ρ πρέπει να ικανοποιεί τις υποθέσεις που ήδη έχουμε αναφέρει στην S-εκτίμηση. Η συνάρτηση ρ πρέπει λοιπόν κατά πρώτον, να είναι συμμετρική με συνεχή παράγωγο και ρ(0)=0 και κατά δεύτερον, να υπάρχει σταθερά $c>0$ τέτοια ώστε η ρ να είναι γνησίως αύξουσα στο διάστημα $[0,c]$ και σταθερή στο $[\psi,\infty)$.

Κεφάλαιο 5

Στατιστική Ανάλυση με την R

5.1 Εφαρμογή Ι

Στη συνέχεια παρατίθεται το παράδειγμα παλινδρόμησης του Duncan (1961) για το επαγγελματικό κύρος. Μέσα από αυτό αντιλαμβανόμαστε την ανάγκη χρήσης ανθεκτικής παλινδρόμησης. Στο παράδειγμα, σκοπός μας είναι να δείξουμε την αδυναμία της μεθόδου ελαχίστων τετραγώνων να αντιμετωπίσει τις ακραίες ή απόμακρες παρατηρήσεις (outliers) και να παρουσιάσουμε ανθεκτικές (robust) μεθόδους που μπορούν να αντιμετωπίσουν αυτές τις παρατηρήσεις.

Στην περίπτωση της πολλαπλής παλινδρόμησης, η αναγνώριση απομονωμένων τιμών γίνεται ακόμα πιο δύσκολη σε σχέση με την απλή παλινδρόμηση. Αυτό συμβαίνει διότι στην πολλαπλή παλινδρόμηση ο εντοπισμός απομονωμένων παρατηρήσεων είναι αδύνατο να πραγματοποιηθεί μέσω του διαγράμματος διασποράς (scatterplot).

Προκειμένου να αντιμετωπιστεί αυτό το πρόβλημα θα πρέπει αρχικά να χρησιμοποιήσουμε τα διαγράμματα των *τυποποιημένων σφαλμάτων (Studentized residuals)* και κάποια ακόμα διαγνωστικά μέτρα, όπως θα δούμε μετά την περιγραφή των δεδομένων που ακολουθεί.

Περιγραφή

Το πλαίσιο δεδομένων του Duncan αποτελείται από 45 γραμμές και 4 στήλες. Δεδομένα για το κύρος και άλλα χαρακτηριστικά, 45 επαγγελμάτων στην Αμερική το 1950.

Αυτό το πλαίσιο δεδομένων περιλαμβάνει τις παρακάτω στήλες:

- *type*: Τα επαγγέλματα έχουν κατηγοριοποιηθεί ως,
  prof-επαγγελματικό και διευθυντικό,
  wc (white collar)-υπάλληλος γραφείου και
  bc (blue collar)-εργάτης.
- *income*: Ποσοστό ατόμων για το κάθε επάγγελμα που είχαν εισόδημα πάνω από 3,500 δολάρια το χρόνο (το 2008 αντιστοιχεί περίπου σε 31,000 δολάρια το χρόνο)
- *education*: Ποσοστό ατόμων για το κάθε επάγγελμα που ήταν απόφοιτοι

Λυκείου (το 2008 είναι ισοδύναμο με Διδακτορικό τίτλο, PhD).

- *prestige*: Ποσοστό ερωτηθέντων σε μία κοινωνική έρευνα, που βαθμολόγησαν το επάγγελμα ως "καλό" ή "πολύ καλό" σε κύρος.

Ο Duncan χρησιμοποίησε μία γραμμική παλινδρόμηση ελαχίστων τετραγώνων με εξαρτημένη μεταβλητή το επαγγελματικό κύρος και επεξηγηματικές μεταβλητές το εισόδημα και τη μόρφωση, ώστε να προβλεφθεί το κύρος των επαγγελμάτων των οποίων οι τιμές του εισοδήματος και της μόρφωσης ήταν γνωστές, αλλά για το οποίο κύρος δεν υπήρχαν άμεσες αξιολογήσεις. Στην αναλυσή του, δεν χρησιμοποίησε την κατηγορία του επαγγέλματος.

Για να καταστεί δυνατή η ανάλυση των πειραματικών δεδομένων τα εισάγουμε στο στατιστικό πακέτο R και στη συνέχεια σε ένα πλαίσιο δεδομένων, όπως φαίνεται παρακάτω:

```
> Duncan<-read.table(file.choose(), header=TRUE)
> Duncan
                  type income education prestige
accountant        prof     62        86       82
pilot             prof     72        76       83
architect         prof     75        92       90
author            prof     55        90       76
chemist           prof     64        86       90
minister          prof     21        84       87
professor         prof     64        93       93
dentist           prof     80       100       90
reporter            wc     67        87       52
engineer          prof     72        86       88
undertaker        prof     42        74       57
lawyer            prof     76        98       89
physician         prof     76        97       97
welfare.worker    prof     41        84       59
teacher           prof     48        91       73
```

Για τη περιγραφική ανάλυση αρχικά χρησιμοποιώ την εντολή *summary*.

```
> summary(Duncan)
   type        income          education         prestige
 bc   :21   Min.   : 7.00   Min.   :  7.00   Min.   : 3.00
 prof:18   1st Qu.:21.00   1st Qu.: 26.00   1st Qu.:16.00
 wc   : 6   Median :42.00   Median : 45.00   Median :41.00
           Mean   :41.87   Mean   : 52.56   Mean   :47.69
           3rd Qu.:64.00   3rd Qu.: 84.00   3rd Qu.:81.00
           Max.   :81.00   Max.   :100.00   Max.   :97.00
```

Στην εισαγωγή των δεδομένων, η μεταβλητή *type* περιείχε δεδομένα χαρακτήρων, η οποία με την εντολή *read.table* μετατράπηκε σε παράγοντα (factor). Έτσι, η

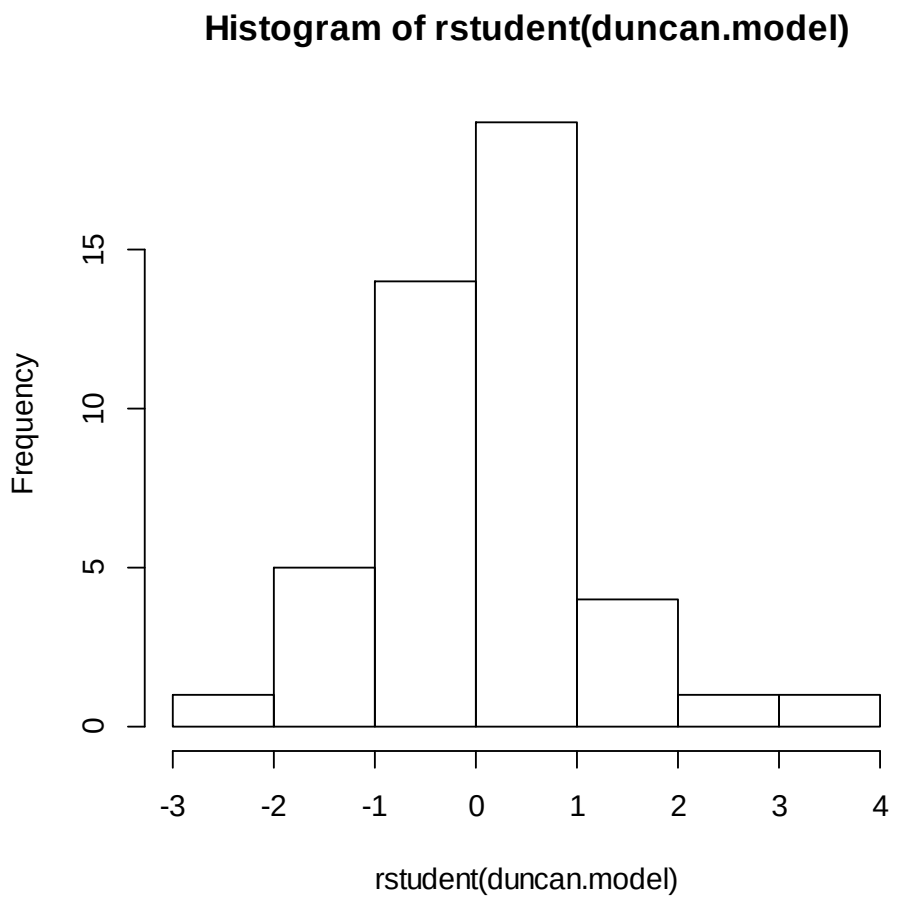
συνάρτηση *summary* αθροίζει το πλήθος των παρατηρήσεων σε κάθε επίπεδο (κατηγορία) του παράγοντα. Οι μεταβλητές *income*, *education* και *prestige* είναι αριθμητικές και από τη συνάρτηση *summary* παίρνουμε τη μικρότερη και τη μεγαλύτερη τιμή, τη δειγματική διάμεσο, το δειγματικό μέσο, το πρώτο και το τρίτο τεταρτημόριο.

Στη συνέχεια, για την παλινδρόμηση με τη μέθοδο *Ελαχίστων Τετραγώνων (OLS)*, του κύρους σε σχέση με το εισόδημα και την εκπαίδευση, πληκτρολογούμε τις παρακάτω εντολές:

```
> mod.ls <- lm(prestige~income+education, data=Duncan)
> summary(mod.ls)

Call:
lm(formula = prestige ~ income + education, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max
-29.538  -6.417   0.655   6.605  34.641

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.06466    4.27194  -1.420    0.163
income       0.59873    0.11967   5.003 1.05e-05 ***
education    0.54583    0.09825   5.555 1.73e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared:  0.8282,    Adjusted R-squared:   0.82
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

Παρατηρούμε ότι και οι δύο μεταβλητές *income* και *education* έχουν στατιστικά σημαντικούς συντελεστές παλινδρόμησης (μικρά p-values).
Τώρα, θα εξετάσουμε τα δεδομένα μας χρησιμοποιώντας διαγνωστικές μεθόδους. Αρχικά, υπολογίζουμε τα *τυποποιημένα υπόλοιπα (Studentized residuals)* με την συνάρτηση *duncan.model* και στη συνέχεια πληκτρολογώντας τις ακόλουθες εντολές, εμφανίζεται το ιστόγραμμα των τυποποιημένων υπολοίπων.

```
> hist(rstudent(duncan.model))
```

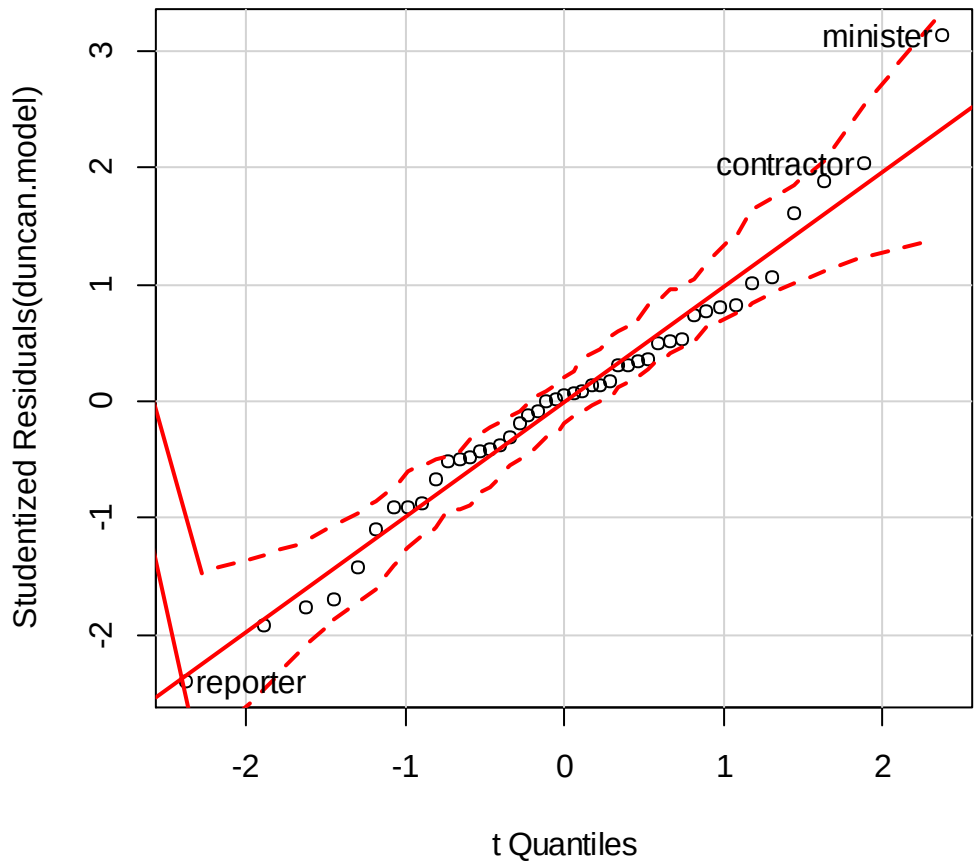


Εικόνα 5.1 Ιστόγραμμα των τυποποιημένων υπολοίπων.

Παρατηρούμε ότι, τα τυποποιημένα υπόλοιπα δείχνουν να ακολουθούν την κατανομή t-Student.

Επιπλέον η συνάρτηση *qqPlot* (*quantile-comparison plots*) από το πακέτο *car*, υπολογίζει τα τυποποιημένα υπόλοιπα και τα παρουσιάζει σε σχέση με τα ποσοστημόρια της κατάλληλης *t-κατανομής* και αποτελεί ένα πιο αξιόπιστο τρόπο εξέτασης της κατανομής αυτής.

```
> qqPlot(duncan.model, labels=row.names(Duncan), id.n=3)
  reporter contractor   minister
         1         44         45
```

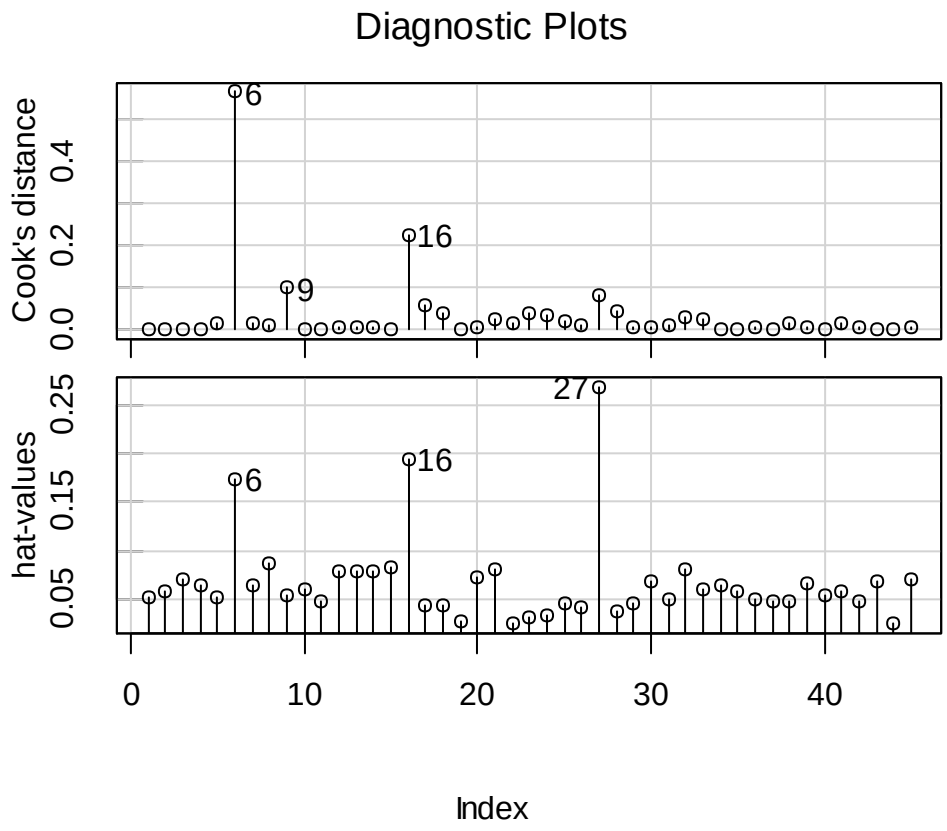


Εικόνα 5.2 95% περιοχή εμπιστοσύνης (confidence envelope) για τα τυποποιημένα υπόλοιπα.

Στην περιπτωσή μας, τα υπόλοιπα ξεφεύγουν ελαφρώς από την γραμμή σύγκρισης και στα δύο άκρα, γεγονός που υποδηλώνει ότι η κατανομή των υπολοίπων έχει σχετικά μεγάλο βάρος στις ουρές (*heavy-tailed*). Εξ ορισμού, η *qqPlot* δημιουργεί μία 95% περιοχή εμπιστοσύνης (*confidence envelope*) για τα τυποποιημένα υπόλοιπα. Τα υπόλοιπα παραμένουν σχεδόν εντός των ορίων και στα δύο άκρα της κατανομής.

Στη συνέχεα ελέγχουμε για σημεία *υψηλής μόχλευσης* (leverage points) και *επιρρεάζουσες παρατηρήσεις* (*influential observations*). Θα χρησιμοποιήσουμε τις γραφικές παραστάσεις *hat-values* και *Cook's distances*. Έχουμε:

```
> influenceIndexPlot(duncan.model, vars=c("Cook", "hat"), id.n=3)
```



Εικόνα 5.3 Γραφικές παραστάσεις hat-values και Cook distances, από την παλινδρόμηση του επαγγελματικού κύρους ως προς το εισόδημα και την εκπαίδευση.

Τα αποτελέσματα παρουσιάζονται στην Εικόνα 5.3. Στα διαγράμματα των Εικόνων 5.2 και 5.3 χρησιμοποιούμε την παράμετρο *id.n=3* για να επισημάνουμε τις τρεις πιο ακραίες παρατηρήσεις. Η προσοχή μας εστιάζεται στις παρατηρήσεις 6 και 16, οι οποίες επισημαίνονται και στα δύο διαγράμματα και τα οποία αντιστοιχούν στα ακόλουθα επαγγέλματα:

```
> rownames(Duncan)[c(6, 16)]
[1] "minister"  "conductor"
```

Θα πρέπει λοιπόν να ανησυχούμε για τα επαγγέλματα του ιερέα (6) και του εισπράκτορα (16), τα οποία μαζί μειώνουν το συντελεστή του *εισοδήματος* (*income*) και αυξάνουν το συντελεστή της *εκπαίδευσης* (*education*). Αυτό το αποτέλεσμα

μπορούμε επίσης να το συμπεράνουμε και παραλείποντας αυτές τις δύο παρατηρήσεις από την ανάλυση. Πληκτρολογούμε λοιπόν:

```
> mod.ls.2 <- update(mod.ls, subset=-c(6,16))
> summary(mod.ls.2)

Call:
lm(formula = prestige ~ income + education, data = Duncan, subset = -c(6,
    16))

Residuals:
    Min      1Q  Median      3Q     Max
-28.612  -5.898   1.937   5.616  21.551

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.40899    3.65263  -1.755   0.0870 .
income       0.86740    0.12198   7.111 1.31e-08 ***
education    0.33224    0.09875   3.364   0.0017 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.42 on 40 degrees of freedom
Multiple R-squared:  0.876,     Adjusted R-squared:  0.8698
F-statistic: 141.3 on 2 and 40 DF,  p-value: < 2.2e-16
```

Παρατηρούμε λοιπόν ότι, παραλείποντας τις δύο παρατηρήσεις, αυξάνεται ο συντελεστής για το εισόδημα κατά περίπου 40% και μειώνεται ο συντελεστής για την εκπαίδευση κατά περίπου 60%. Σε άλλα προβλήματα όμως, η παράλειψη μίας παρατήρησης μπορεί να αλλάξει σημαντικά αποτελέσματα σε μη-σημαντικά.

Εναλλακτικά, ας χρησιμοποιήσουμε την *Μ-εκτιμήτρια* του *Huber*, χρησιμοποιώντας την εντολή *rlm (robust linear model)* μέσα από τη βιβλιοθήκη MASS της R:

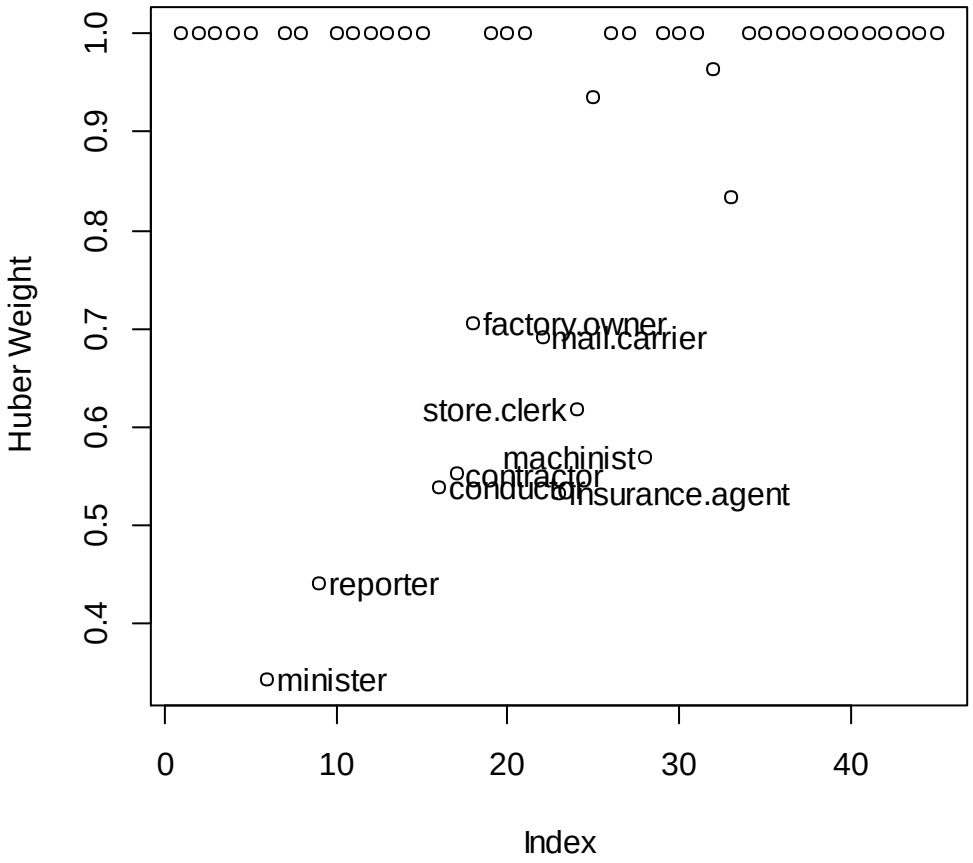
```
> library(MASS)
> mod.huber <- rlm(prestige~income+education, data= Duncan)
> summary(mod.huber)

Call: rlm(formula = prestige ~ income + education, data = Duncan)
Residuals:
    Min      1Q  Median      3Q     Max
-30.120  -6.889   1.291   4.592  38.603

Coefficients:
            Value   Std. Error t value
(Intercept) -7.1107  3.8813    -1.8320
income       0.7014  0.1087     6.4516
education    0.4854  0.0893     5.4380

Residual standard error: 9.892 on 42 degrees of freedom
```

Οι εκτιμήσεις των συντελεστών παλινδρόμησης με τη μέθοδο του *Huber*, είναι ανάμεσα από αυτές που προέκυψαν από την μέθοδο ελαχίστων τετραγώνων και την μέθοδο ελαχίστων τετραγώνων παραλείποντας τα επαγγέλματα του ιερέα και του εισπράκτορα.

Εδώ είναι χρήσιμο να διεξάγουμε τη γραφική παράσταση (Εικόνα 5.4) με τα τελικά βάρη που χρησιμοποιήθηκαν στην παραπάνω ανθεκτική εκτίμηση. Η συνάρτηση *showLabels* από το πακέτο *car* χρησιμοποιείται για την επισήμανση όλων των παρατηρήσεων με βάρη λιγότερο από 0.8:

```
> plot(mod.huber$w, ylab="Huber Weight")
> smallweights <- which(mod.huber$w<0.8)
> showLabels(1:45, mod.huber$w, rownames(Duncan), id.method=smallweights,cex.=.6)
       minister        reporter       conductor      contractor   factory.owner
              6               9              16              17              18
   mail.carrier insurance.agent     store.clerk       machinist
             22              23              24              28
```

Τα επαγγέλματα του *ιερέα* και του *εισπράκτορα* είναι μεταξύ των παρατηρήσεων που λαμβάνουν το μικρότερο βάρος.



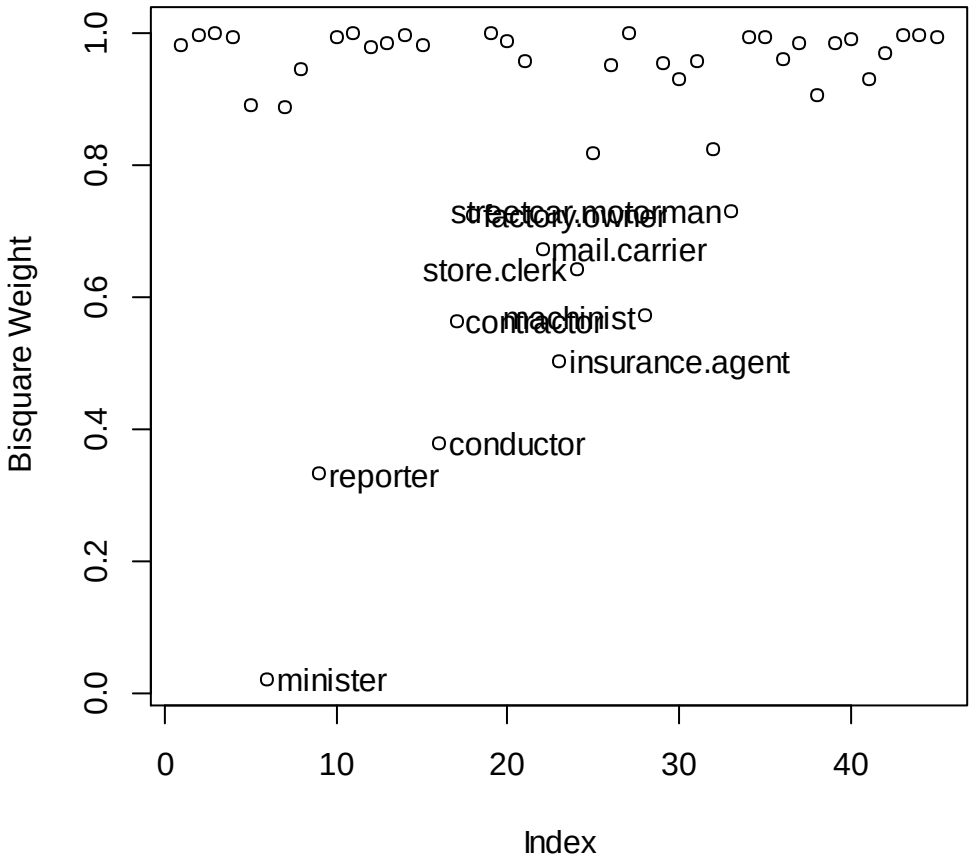
Εικόνα 5.4 Γραφική παράσταση με τα τελικά βάρη που χρησιμοποιήθηκαν στην μέθοδο του *Huber*.

Η εκτιμήτρια *bisquare*, επίσης προσαρμόζεται από την εντολή *rlm*. Οι τιμές εκκίνησης στη ρουτίνα *IRLS* μπορεί να είναι καθοριστικές για αυτή την εκτιμήτρια. Η επιλογή του ορίσματος *method="MM"* στην *rlm* αναζητεί εκτιμήτριες *bisquare* με τιμές εκκίνησης που προκύπτουν από μία αρχική παλινδρόμηση φραγμένης επιρροής.

```
> mod.bisq <- rlm(prestige~income+education, data=Duncan, method="MM")
> summary(mod.bisq)

Call: rlm(formula = prestige ~ income + education, data = Duncan, method = "MM")
Residuals:
    Min      1Q  Median      3Q     Max
-29.871  -6.626   1.444   4.465  42.396

Coefficients:
            Value   Std. Error t value
(Intercept) -7.3886  3.9076    -1.8908
income       0.7825  0.1095     7.1490
education    0.4233  0.0899     4.7102

Residual standard error: 9.792 on 42 degrees of freedom
```

Συγκρίνοντας σε σχέση με τις *Huber εκτιμήτριες*, η *bisquare* εκτιμήτρια του συντελεστή εισοδήματος είναι μεγαλύτερη και η εκτιμήτρια του συντελεστή εκπαίδευσης είναι μικρότερη. Στην Εικόνα 5.5 παρουσιάζεται το γράφημα με τα βάρη της *bisquare* προσαρμογής, αναγνωρίζοντας τις παρατηρήσεις με τα χαμηλότερα βάρη:

```
> plot(mod.bisq$w, ylab="Bisquare Weight")
> showLabels(1:45, mod.bisq$w, rownames(Duncan), id.method=which(mod.bisq$w<0.8), cex.=0.6)
          minister           reporter          conductor         contractor
                 6                  9                 16                 17
     factory.owner       mail.carrier    insurance.agent        store.clerk
                18                 22                 23                 24
         machinist  streetcar.motorman
                28                 33
```



Εικόνα 5.5 Γραφική παράσταση με τα βάρη της *bisquare* προσαρμογής.

Τώρα, θα προσαρμόσουμε το μοντέλο του Duncan με τη μέθοδο *Ελαχίστων Περικοπτόμενων Τετραγώνων (LTS)*. Γι' αυτό, θα χρησιμοποιήσουμε τη συνάρτηση *ltsreg* μέσα από το πακέτο *lqs*. Η *LTS* παλινδρόμηση είναι η προεπιλεγμένη μέθοδος της συνάρτησης *lqs*, η οποία επιπλέον μπορεί να εκτιμήσει κι άλλες εκτιμήτριες *φραγμένης επιρροής* (*bounded-influence*). Έχουμε λοιπόν:

```
> (mod.lts <- ltsreg(prestige~income+education, data=Duncan))
Call:
lqs.formula(formula = prestige ~ income + education, data = Duncan,
    method = "lts")

Coefficients:
(Intercept)       income    education
    -6.8023       0.7975       0.4331

Scale estimates 7.768 7.564
```

Σ’ αυτήν τη περίπτωση, τα αποτελέσματα είναι παρόμοια με εκείνα που προέκυψαν από την *M-εκτιμήτρια.*

Επίσης, μπορούμε να χρησιμοποιήσουμε την *Τεταρτομοριακή Παλινδρόμηση (Quantile Regression)*, στην πολλαπλή ανάλυση παλινδρόμησης. Προχωράμε λοιπόν στην εκτίμηση των συντελεστών του μοντέλου μας, χρησιμοποιώντας την *L1* εκτιμήτρια της τεταρτομοριακής παλινδρόμησης:

```
> library(quantreg)
> mod.quant <- rq(prestige~income+education, data=Duncan)
> summary(mod.quant)

Call: rq(formula = prestige ~ income + education, data = Duncan)

tau: [1] 0.5

Coefficients:
            coefficients lower bd  upper bd
(Intercept)  -6.40826    -12.49552  -3.60027
income        0.74771      0.47194   0.91169
education     0.45872      0.21948   0.66095
```

Βλέπουμε ότι, οι εκτιμήσεις των συντελεστών με την *L1* εκτιμήτρια είναι παρόμοιες με εκείνες της *M-εκτιμήτριας* η οποία βασίζεται στη συνάρτηση βάρους του *Huber*.

Στον πίνακα που ακολουθεί συνοψίζουμε τις τιμές των συντελεστών για την κάθε εκτιμήτρια που χρησιμοποιήσαμε στην αναλυσή μας.

Μέθοδος	b_0	b_1 (income)	b_2 (education)
ET (OLS)	-6.06466	0.54583	0.59873
ET χωρίς τις παρ. 6 & 16	-6.40899	0.86740	0.33224
Huber M-εκτιμήτρια	-7.1107	0.7014	0.4854
Bisquare MM-εκτιμήτρια	-7.3886	0.7825	0.4233
ΕΠΤ (LTS)	-6.8023	0.7975	0.4331
L1-εκτιμήτρια	-6.40826	0.74771	0.45872

Πίνακας 5.1: Διάφορες εκτιμήσεις του μοντέλου παλινδρόμησης επαγγελματικό-κύρος του Duncan.

Τελικά λοιπόν, παρατηρούμε ότι τα αποτελέσματα που προκύπτουν από τις ανθεκτικές μεθόδους, διαφέρουν σημαντικά σε σχέση με τις εκτιμήσεις της *μεθόδου ελαχίστων τετραγώνων*. Οπότε, η ύπαρξη σημείων μόχλευσης και σημείων επιρροής, αλλοιώνει αρκετά το αποτέλεσμα. Συνεπώς, η εφαρμογή μεθόδων ανθεκτικής παλινδρόμησης κρίνεται απαραίτητη.

5.2 Εφαρμογή ΙΙ

Ακολουθεί μία δεύτερη εφαρμογή όπου θα διαπιστώσουμε όπως και στην παραπάνω, την ανάγκη της χρήσης ανθεκτικών μεθόδων. Το ακόλουθο παράδειγμα ανέλυσαν οι Zaman, Rousseeuw και Orhan (2001). Σκοπός τους ήταν να δείξουν πώς αυτές οι εύρωστες τεχνικές που είδαμε ως τώρα, βελτιώνουν αποτελεσματικά τις εκτιμήσεις της κλασσικής μεθόδου *Ελαχίστων Τετραγώνων*, αναλύοντας τη μελέτη των De Long και Summers (1991) για την εθνική ανάπτυξη (national growth).
Οι De Long και Summers μελέτησαν την εθνική ανάπτυξη 61 χωρών από το 1960 έως το 1985.

Περιγραφή
Ο πίνακας δεδομένων του De Long και Summers αποτελείται από 61 γραμμές και τις παρακάτω 6 στήλες:

- *GDP*: ρυθμός ανάπτυξης του ΑΕΠ.
- *LFG*: ρυθμός ανάπτυξης του εργατικού δυναμικού.
- *GAP*: σχετικό κενό ανάπτυξης.
- *EQP*: επενδύσεις στον εξοπλισμό.
- *NEQ*: άλλες επενδύσεις.

Στην μελέτη τους, έκαναν εκτίμηση με τη μέθοδο Ελαχίστων Τετραγώνων και χρησιμοποιήσαν την συνάρτηση παλινδρόμησης

$$GDP = \beta_0 + \beta_1 LFG + \beta_2 GAP + \beta_3 EQP + \beta_4 NEQ + \varepsilon$$

όπου η μεταβλητή απόκρισης είναι η *(GDP)* και οι ανεξάρτητες μεταβλητές είναι οι *(LFG), (GAP), (EQP)* και *(NEQ).*

Αρχικά, εισάγουμε τα δεδομένα στο στατιστικό πακέτο R και στη συνέχεια σε ένα πλαίσιο δεδομένων, όπως φαίνεται παρακάτω:

```
> growth <- read.table(file.choose(), header=TRUE)
> growth
             GDP    LFG    EQP    NEQ    GAP
Argentin  0.0089 0.0118 0.0214 0.2286 0.6079
Austria   0.0332 0.0014 0.0991 0.1349 0.5809
Belgium   0.0256 0.0061 0.0684 0.1653 0.4109
Bolivia   0.0124 0.0209 0.0167 0.1133 0.8634
Botswana  0.0676 0.0239 0.1310 0.1490 0.9474
Brazil    0.0437 0.0306 0.0646 0.1588 0.8498
Cameroon  0.0458 0.0169 0.0415 0.0885 0.9333
Canada    0.0169 0.0261 0.0771 0.1529 0.1783
Chile     0.0021 0.0216 0.0154 0.2846 0.5402
Colombia  0.0239 0.0266 0.0229 0.1553 0.7695
CostaRic  0.0121 0.0354 0.0433 0.1067 0.7043
Denmark   0.0187 0.0115 0.0688 0.1834 0.4079
Dominica  0.0199 0.0280 0.0321 0.1379 0.8293
Ecuador   0.0283 0.0274 0.0303 0.2097 0.8205
```

Για την περιγραφική ανάλυση χρησιμοποιούμε την εντολή *summary*:

```
> summary(growth)
    countries        GDP                LFG                EQP
 Argentin: 1   Min.   :-0.01100   Min.   :0.00140   Min.   :0.01380
 Austria : 1   1st Qu.: 0.01210   1st Qu.:0.01180   1st Qu.:0.02670
 Belgium : 1   Median : 0.02310   Median :0.02390   Median :0.04330
 Bolivia : 1   Mean   : 0.02238   Mean   :0.02113   Mean   :0.05232
 Botswana: 1   3rd Qu.: 0.03010   3rd Qu.:0.02800   3rd Qu.:0.07110
 Brazil  : 1   Max.   : 0.06760   Max.   :0.03780   Max.   :0.13100
 (Other) :55
      NEQ              GAP
 Min.   :0.0267   Min.   :0.0000
 1st Qu.:0.0957   1st Qu.:0.5809
 Median :0.1356   Median :0.8015
 Mean   :0.1399   Mean   :0.7258
 3rd Qu.:0.1790   3rd Qu.:0.8850
 Max.   :0.2846   Max.   :0.9805
```

Η παραπάνω συνάρτηση αθροίζει το πλήθος των παρατηρήσεων σε κάθε επίπεδο (κατηγορία) του παράγοντα *countries*. Ενώ οι μεταβλητές *GDP*, *LFG*, *GAP*, *EQP* και *NEQ* είναι αριθμητικές και από τη συνάρτηση *summary* παίρνουμε τη μικρότερη και τη μεγαλύτερη τιμή, τη δειγματική διάμεσο, το δειγματικό μέσο, το πρώτο και το τρίτο τεταρτομόριο.

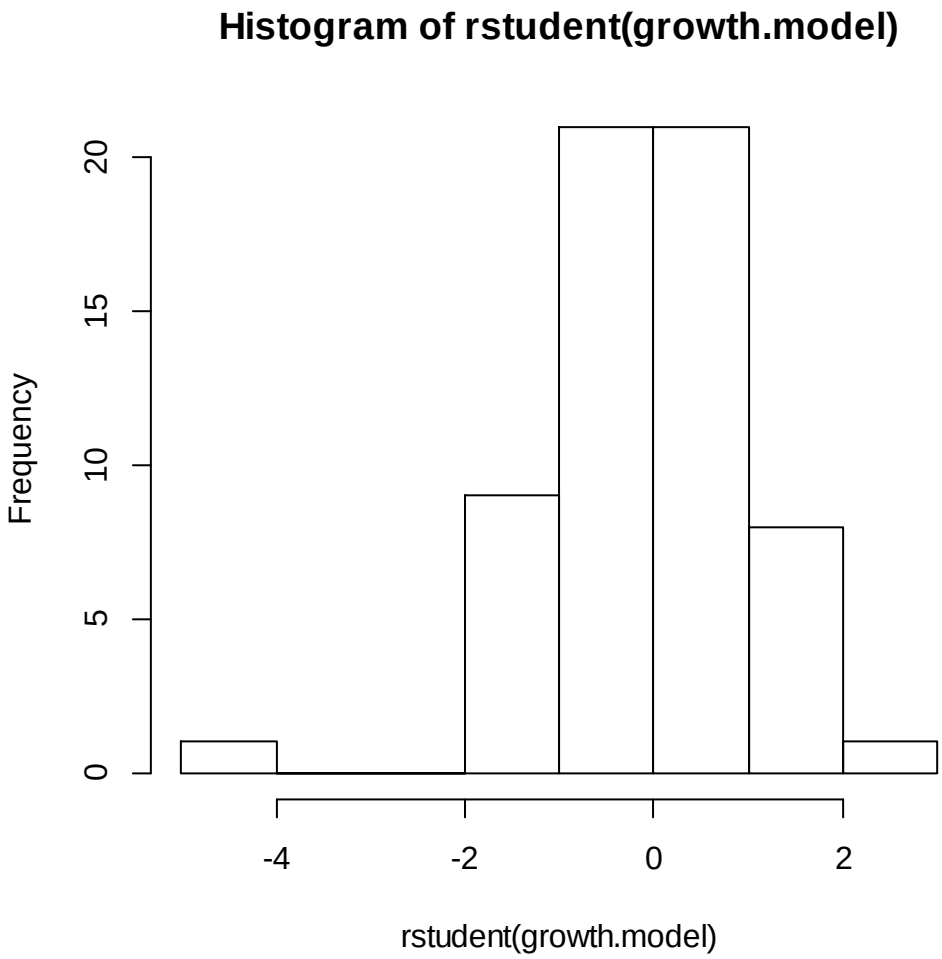
Στη συνέχεια, παρουσιάζονται κάποια αποτελέσματα με τη μέθοδο *Ελαχίστων Τετραγώνων (OLS)*:

```
> mod.ls <- lm(GDP~LFG+GAP+EQP+NEQ, data=growth)
> summary(mod.ls)

Call:
lm(formula = GDP ~ LFG + GAP + EQP + NEQ, data = growth)

Residuals:
      Min        1Q    Median        3Q       Max
-0.044675 -0.004961 -0.000121  0.008983  0.025161

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.014298   0.010278  -1.391 0.169700
LFG         -0.029805   0.198376  -0.150 0.881110
GAP          0.020260   0.009174   2.208 0.031332 *
EQP          0.265376   0.065294   4.064 0.000152 ***
NEQ          0.062363   0.034823   1.791 0.078724 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01306 on 56 degrees of freedom
Multiple R-squared:  0.3388,    Adjusted R-squared:  0.2916
F-statistic: 7.175 on 4 and 56 DF,  p-value: 9.755e-05
```

Παρατηρούμε ότι από την ανάλυση *OLS*, οι μεταβλητές *GAP (p=0.031)* και *EQP (p<0.001)* επιδρούν σημαντικά στο ρυθμό ανάπτυξης *GDP*.

Τώρα, θα εξετάσουμε τα δεδομένα μας χρησιμοποιώντας διαγνωστικές μεθόδους. Αρχικά, υπολογίζουμε τα *τυποποιημένα υπόλοιπα (Studentized residuals)* με την συνάρτηση *growth.model* και στη συνέχεια πληκτρολογώντας τις ακόλουθες εντολές, εμφανίζεται το ιστόγραμμα των τυποποιημένων υπολοίπων:

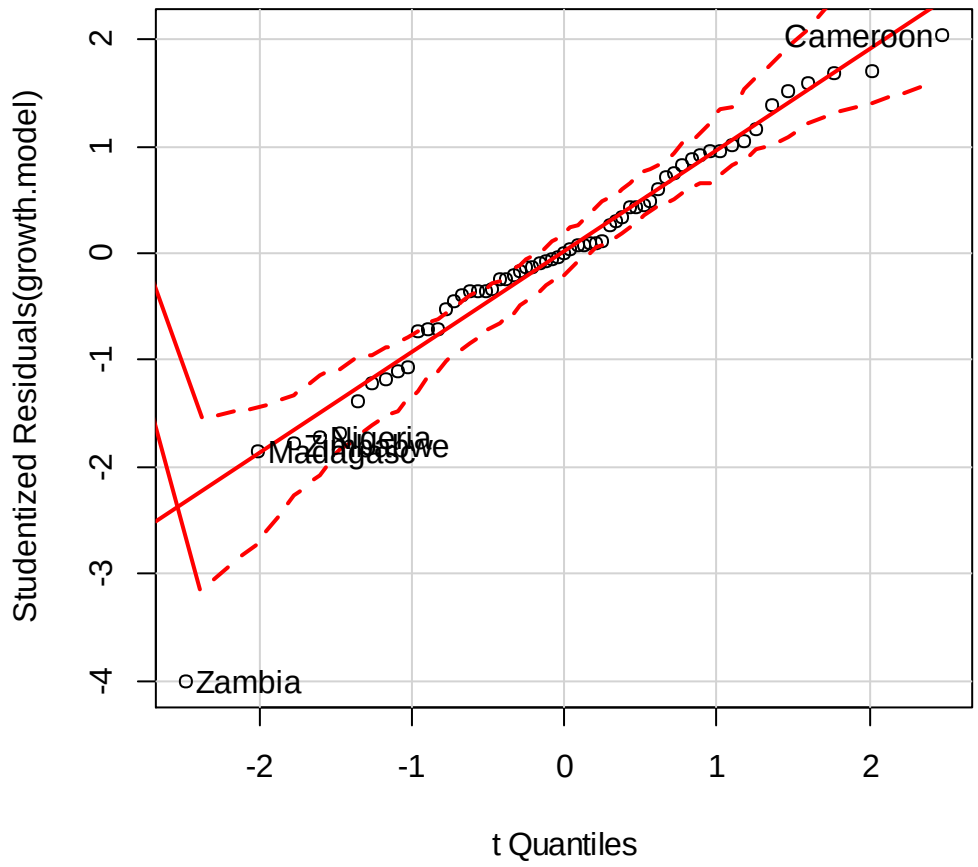
```
> hist(rstudent(growth.model))
```



Εικόνα 5.6 Ιστόγραμμα των τυποποιημένων υπολοίπων.

Παρατηρούμε ότι, τα τυποποιημένα υπόλοιπα ακολουθούν την *t-Student* κατανομή. Η συνάρτηση *qqPlot* από το πακέτο *car*, υπολογίζει τα τυποποιημένα υπόλοιπα και τα παρουσιάζει σε σχέση με τα ποσοστημόρια της κατάλληλης *t-κατανομής* (*quantile-comparison plots*).

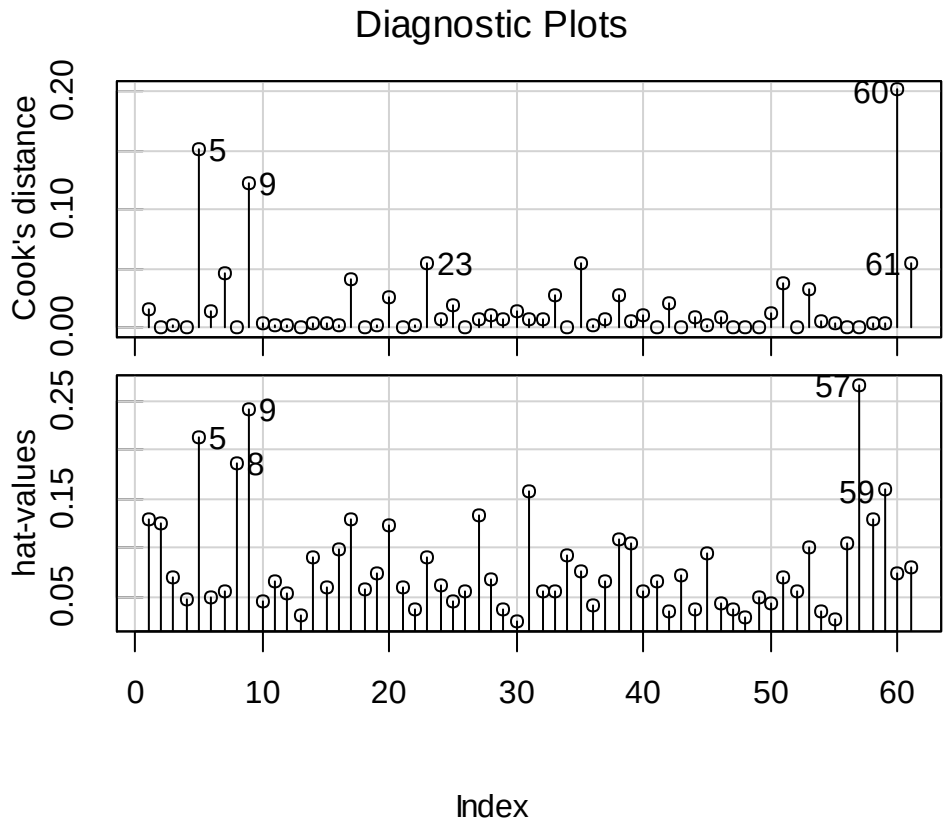
```
> qqPlot(growth.model, labels=row.names(growth), id.n=5)
  Zambia Madagasc Zimbabwe  Nigeria Cameroon
       1        2        3        4       61
```



Εικόνα 5.7 95% περιοχή εμπιστοσύνης (confidence envelope) για τα τυποποιημένα υπόλοιπα.

Στη συνέχεα ελέγχουμε για σημεία *υψηλής μόχλευσης* και *επιρρεάζουσες παρατηρήσεις*. Θα χρησιμοποιήσουμε τις γραφικές παραστάσεις *hat-values* και *Cook's distances*. Πληκτρολογούμε τις παρακάτω εντολές και έχουμε:

```
> influenceIndexPlot(growth.model, vars=c("Cook", "hat"), id.n=5)
```



Εικόνα 5.8 Γραφικές παραστάσεις *hat-values* και *Cook's distances*, από την παλινδρόμηση της πληθυσμιακής ανάπτυξης ως προς τον ρυθμό ανάπτυξης του ΑΕΠ, τον ρυθμό ανάπτυξης του εργατικού δυναμικού, το σχετικό παραγωγικό κενό, επενδύσεις σε εξωτερικό και άλλες επενδύσεις.

Τα παραπάνω διαγράμματα δείχνουν ότι η *Zambia*, η 60η χώρα στα δεδομένα, είναι *απομακρυσμένη τιμή* (*outlier*). Επίσης εμφανίζονται *σημεία μόχλευσης*. Ωστόσο, δεν παρατηρούνται σημαντικά υψηλά σημεία μόχλευσης.

Θα πρέπει λοιπόν να ανησυχούμε για την 60η παρατήρηση (Zambia). Όπως προαναφέραμε, η παράλειψη μίας παρατήρησης μπορεί να αλλάξει σημαντικά αποτελέσματα, σε μη-σημαντικά. Παραλείποντας λοιπόν την παρατήρηση αυτή, πάλι με την μέθοδο *Ελαχίστων Τετραγώνων* έχουμε τα ακόλουθα αποτελέσματα:

```
> mod.ls.2 <- update(mod.ls, subset=-c(60)) #fit without 60th observation (Zambia)
> summary(mod.ls.2)

Call:
lm(formula = GDP ~ LFG + GAP + EQP + NEQ, data = growth, subset = -c(60))

Residuals:
       Min         1Q     Median         3Q        Max
-0.0243986 -0.0053737 -0.0001369  0.0071619  0.0251511

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.022194   0.009337  -2.377  0.02096 *
LFG          0.044584   0.177124   0.252  0.80220
GAP          0.024486   0.008214   2.981  0.00427 **
EQP          0.282428   0.058134   4.858 1.02e-05 ***
NEQ          0.084920   0.031430   2.702  0.00915 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0116 on 55 degrees of freedom
Multiple R-squared:  0.4444,    Adjusted R-squared:  0.404
F-statistic:    11 on 4 and 55 DF,  p-value: 1.263e-06
```

Παρατηρούμε λοιπόν ότι, παραλείποντας την 60η παρατήρηση ο συντελεστής του ρυθμού ανάπτυξης του εργατικού δυναμικού *(LFG)* έχει αλλάξει πρόσημο (από αρνητικό σε θετικό). Ο έλεγχος για το συντελεστή του *GAP* δίνει μικρότερη *p*-τιμή και ο συντελεστής των άλλων επενδύσεων (*NEQ)* έχει αυξηθεί κατά 25% περίπου και συμβάλει στατιστικά σημαντικά στο μοντέλο.

Εναλλακτικά, ας χρησιμοποιήσουμε την *Μ-εκτιμήτρια* του *Huber*, χρησιμοποιώντας την συνάρτηση *rlm (robust linear model)* μέσα από το στατιστικό πακέτο *MASS*:

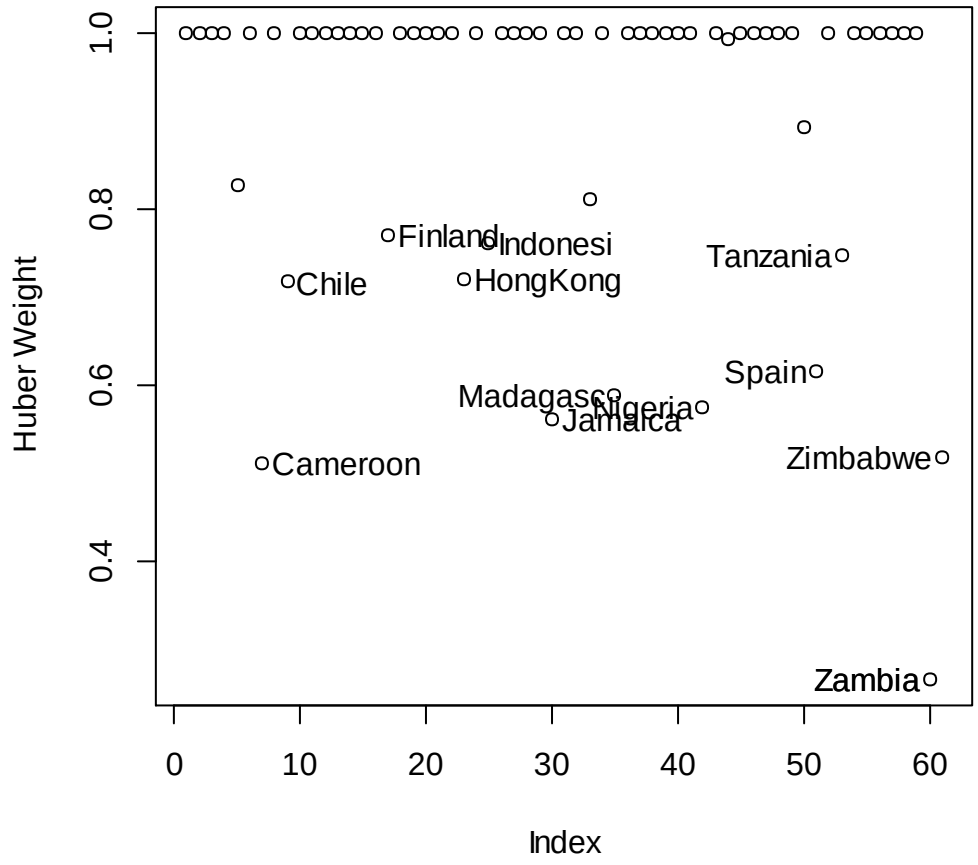
```
> mod.huber <- rlm(GDP~LFG+GAP+EQP+NEQ, data=growth)
> summary(mod.huber)

Call: rlm(formula = GDP ~ LFG + GAP + EQP + NEQ, data = growth)
Residuals:
       Min         1Q     Median         3Q        Max
-0.0482356 -0.0064699 -0.0001412  0.0063549  0.0251736

Coefficients:
            Value   Std. Error t value
(Intercept) -0.0217  0.0098    -2.2049
LFG          0.0735  0.1898     0.3874
GAP          0.0236  0.0088     2.6880
EQP          0.2865  0.0625     4.5856
NEQ          0.0808  0.0333     2.4259

Residual standard error: 0.009592 on 56 degrees of freedom
```

Η παραπάνω μέθοδος δείχνει ότι εκτός από τις μεταβλητές *GAP* και *EQP*, η ανθεκτική ανάλυση παρουσιάζει και την *NEQ* ότι έχει σημαντική επίδραση στην εξαρτημένη μεταβλητή *GDP*.

Εδώ είναι χρήσιμο να διεξάγουμε τη γραφική παράσταση (Εικόνα 5.9) με τα τελικά βάρη που χρησιμοποιήθηκαν στην παραπάνω ανθεκτική εκτίμηση. Η συνάρτηση *showLabels* από το πακέτο *car* χρησιμοποιείται για την επισήμανση όλων των παρατηρήσεων με βάρη λιγότερο από *0.8*:

```
> plot(mod.huber$w, ylab="Huber Weight")
> smallweights <- which(mod.huber$w<0.8)
> showLabels(1:61, mod.huber$w, rownames(growth), id.method=smallweights,cex.=.6)
Cameroon    Chile  Finland HongKong Indonesi  Jamaica Madagasc  Nigeria
       7        9       17       23       25       30       35       42
   Spain Tanzania   Zambia Zimbabwe
      51       53       60       61
```

Η *Zambia* είναι μεταξύ των παρατηρήσεων που λαμβάνουν το μικρότερο βάρος.



Εικόνα 5.9 Γραφική παράσταση με τα τελικά βάρη που χρησιμοποιήθηκαν στην μέθοδο του *Huber*.

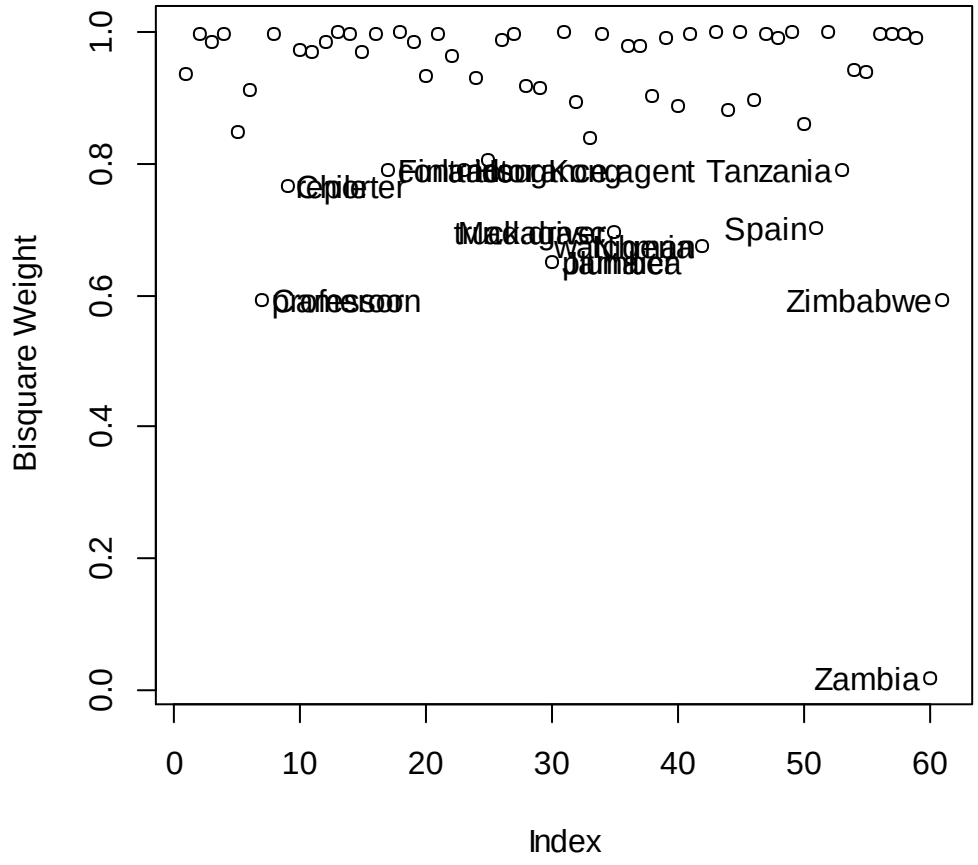
Στη συνέχεια θα χρησιμοποιήσουμε την εκτιμήτρια *bisquare*, που προσαρμόζεται από την εντολή *rlm*. Θα έχουμε λοιπόν:

```
> mod.bisq <- rlm(GDP~LFG+GAP+EQP+NEQ, data=growth, method="MM")
> summary(mod.bisq)

Call: rlm(formula = GDP ~ LFG + GAP + EQP + NEQ, data = growth, method = "MM")
Residuals:
       Min         1Q     Median         3Q        Max
-0.0490537 -0.0064768 -0.0003844  0.0061924  0.0252941

Coefficients:
            Value   Std. Error t value
(Intercept) -0.0238  0.0096    -2.4683
LFG          0.0869  0.1862     0.4667
GAP          0.0247  0.0086     2.8651
EQP          0.2917  0.0613     4.7610
NEQ          0.0872  0.0327     2.6686

Residual standard error: 0.01124 on 56 degrees of freedom
```

Συγκρίνοντας σε σχέση με τις *Huber εκτιμήτριες*, οι τιμές της *bisquare* εκτίμησης των συντελεστών είναι ελαφρώς μεγαλύτερες.

Στην Εικόνα 5.5 παρουσιάζεται το γράφημα με τα βάρη της *bisquare* προσαρμογής, αναγνωρίζοντας τις παρατηρήσεις με τα χαμηλότερα βάρη:

```
> plot(mod.bisq$w, ylab="Bisquare Weight")
> showLabels(1:61, mod.bisq$w, rownames(growth), id.method=which(mod.bisq$w<0.8), cex.=0.6)
Cameroon    Chile  Finland HongKong  Jamaica Madagasc  Nigeria    Spain
       7        9       17       23       30       35       42       51
Tanzania   Zambia Zimbabwe
      53       60       61
```



Εικόνα 5.10 Γραφική παράσταση με τα βάρη της *bisquare* προσαρμογής.

Τώρα, θα προσαρμόσουμε το μοντέλο μας με την μέθοδο *Ελαχίστων Περικοπτόμενων Τετραγώνων (LTS)*. Γι' αυτό, θα χρησιμοποιήσουμε την συνάρτηση *ltsreg* μέσα από

το πακέτο *lqs*. Η *LTS* παλινδρόμηση είναι η προεπιλεγμένη μέθοδος της συνάρτησης *lqs*, η οποία επιπλέον μπορεί να εκτιμήσει κι άλλες εκτιμήτριες *φραγμένης επιρροής* (*bounded-influence*). Έχουμε λοιπόν:

```
> mod.lts <- ltsreg(GDP~LFG+GAP+EQP+NEQ, data=growth)
> (mod.lts <- ltsreg(GDP~LFG+GAP+EQP+NEQ, data=growth))
Call:
lqs.formula(formula = GDP ~ LFG + GAP + EQP + NEQ, data = growth,
    method = "lts")

Coefficients:
(Intercept)          LFG          GAP          EQP          NEQ
   -0.02772      0.40277      0.03341      0.37008      0.02649

Scale estimates 0.008090 0.008819
```

Σ' αυτήν τη περίπτωση, τα αποτελέσματα διαφέρουν αισθητά από τις άλλες εκτιμήτριες.

Επίσης, μπορούμε να χρησιμοποιήσουμε την *Τεταρτομοριακή Παλινδρόμηση (Quantile Regression)* στην πολλαπλή ανάλυση παλινδρόμησης. Προχωράμε λοιπόν στην εκτίμηση των συντελεστών του μοντέλου μας, χρησιμοποιώντας την *L1* εκτιμήτρια της τεταρτομοριακής παλινδρόμησης:

```
> mod.quant <- rq(GDP~LFG+GAP+EQP+NEQ, data=growth)
> summary(mod.quant)

Call: rq(formula = GDP ~ LFG + GAP + EQP + NEQ, data = growth)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) -0.02479     -0.03076 -0.01422
LFG          0.13213     -0.08848  0.32729
GAP          0.02280      0.01433  0.02794
EQP          0.30948      0.02604  0.34765
NEQ          0.08860      0.03739  0.14191
```

Βλέπουμε ότι, οι εκτιμήσεις των συντελεστών με την *L1* εκτιμήτρια είναι παρόμοιες με εκείνες της *M-εκτιμήτριας*, η οποία βασίζεται στη συνάρτηση βάρους του *Huber*.

Στον πίνακα που ακολουθεί συνοψίζουμε τις τιμές των συντελεστών για την κάθε εκτιμήτρια που χρησιμοποιήσαμε στην αναλυσή μας.

Μέθοδος	b_0	b_1 (*LFG*)	b_2 (*GAP*)	b_3 (*EQP*)	b_4 (*NEQ*)
1) *ET (OLS)*	-0.014298	-0.029805	0.020260	0.265376	0.062363
2) *ET χωρίς την παρ. 61*	-0.044584	0.0282428	0.024486	0.282428	0.084920
3) *Huber M-εκτιμήτρια*	-0.0217	0.0735	0.0236	0.2865	0.0808
4) *Bisquare MM- εκτιμήτρια*	-0.0238	0.0869	0.0247	0.2917	0.0872
5) *ΕΠΤ (LTS)*	-0.02772	0.040277	0.03341	0.37008	0.02649
6) *L1-εκτιμήτρια*	-0.02479	0.13213	0.02280	0.30948	0.08860

Πίνακας 5.2 Διάφορες εκτιμήτριες του μοντέλου παλινδρόμησης της πληθυσμιακής ανάπτυξης 61 χωρών από το 1960 έως το 1985 (De Long και Summers).

Τελικά, παρατηρούμε ότι τα αποτελέσματα που προκύπτουν από τις ανθεκτικές μεθόδους, διαφέρουν σημαντικά σε σχέση με τις εκτιμήσεις της αρχικής μεθόδου *ελαχίστων τετραγώνων*. Οπότε η ύπαρξη σημείων μόχλευσης και σημείων επιρροής, αλλοιώνει αρκετά το αποτέλεσμα. Συνεπώς, η εφαρμογή μεθόδων ανθεκτικής παλινδρόμησης επιβάλλεται.

Βέβαια, παρατηρούμε ότι με εξαίρεση τη μέθοδο *LTS* οι εκτιμήσεις των συντελεστών με βάση τις προσεγγίσεις 2,3,4 και 6 ελάχιστα διαφοροποιούνται μεταξύ τους.

5.3 Εφαρμογή ΙΙΙ

Περιγραφή

Το αρχείο “HERS 549.dat” περιέχει *n=549* παρατηρήσεις.

Οι μεταβλητές είναι οι ακόλουθες:

1. Ηλικία
2. BMI (Body mass index) = δείκτης μάζας σώματος
3. LDL = χοληστερίνη – (Η εξαρτημένη μεταβλητή)
4. Φυλή
5. Κάπνισμα
6. Κατανάλωση αλκοόλ

Οι μεταβλητές 4, 5, 6 κωδικοποιούνται ως 0=όχι, 1=ναι.

Για να καταστεί δυνατή η ανάλυση των πειραματικών δεδομένων τα εισάγουμε στο στατιστικό πακέτο R. Στη συνέχεια, τα εισάγουμε σε ένα πλαίσιο δεδομένων αφού πρώτα αλλάξουμε τα ονόματα της κάθε στήλης, ώστε να τις αναγνωρίζουμε πιο εύκολα. Έχουμε με τη σειρά, τις παρακάτω εντολές:

```
> Hers549 <- read.table(file.choose(), header=TRUE)
> names(Hers549)
[1] "AGE"     "BMI"     "LDL"     "NW"      "SMOKE"   "ALCOHOL"
> Hers549
    AGE   BMI   LDL NW SMOKE ALCOHOL
1    70 23.69 122.4  1     0       0
2    65 21.90 150.6  0     0       0
3    61 30.26 133.0  0     0       1
4    62 45.68 220.0  0     1       1
5    72 22.18 172.8  1     0       0
6    73 25.27 123.8  0     0       0
7    57 28.12 153.4  0     1       0
8    58 20.48 165.2  0     1       1
9    66 28.52 134.8  0     0       0
10   60 35.30 198.2  1     0       0
11   67 29.96 140.4  0     0       1
12   61 20.92 190.2  0     1       0
13   53 23.73 247.0  0     0       0
14   56 29.13 124.2  0     1       0
15   68 25.74 164.6  0     0       0
```

Για τη περιγραφική ανάλυση, αρχικά χρησιμοποιώ την εντολή *summary*.

```
> summary(Hers549)
      AGE              BMI              LDL              NW
 Min.   :46.00   Min.   :15.21   Min.   : 41.2   Min.   :0.0000
 1st Qu.:62.00   1st Qu.:24.46   1st Qu.:118.0   1st Qu.:0.0000
 Median :67.00   Median :27.88   Median :138.4   Median :0.0000
 Mean   :66.56   Mean   :28.65   Mean   :143.4   Mean   :0.1494
 3rd Qu.:72.00   3rd Qu.:31.96   3rd Qu.:162.8   3rd Qu.:0.0000
 Max.   :79.00   Max.   :51.47   Max.   :345.8   Max.   :1.0000
     SMOKE           ALCOHOL
 Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000
 Mean   :0.1494   Mean   :0.3607
 3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000
```

Στην εισαγωγή των δεδομένων, οι μεταβλητές *NW, SMOKE* και *ALCOHOL* περιείχαν δεδομένα χαρακτήρων, οι οποίες με την εντολή *read.table* μετατράπηκαν σε παράγοντες (factor). Έτσι, η συνάρτηση *summary* αθροίζει το πλήθος των παρατηρήσεων σε κάθε επίπεδο (κατηγορία) του παράγοντα. Οι μεταβλητές *AGE* και *BMI* είναι αριθμητικές και με τη συνάρτηση *summary* παρατηρούμε παραπάνω τη μικρότερη και τη μεγαλύτερη τιμή, τη δειγματική διάμεσο, το δειγματικό μέσο, το πρώτο και το τρίτο τεταρτημόριο, για την κάθε μεταβλητή.

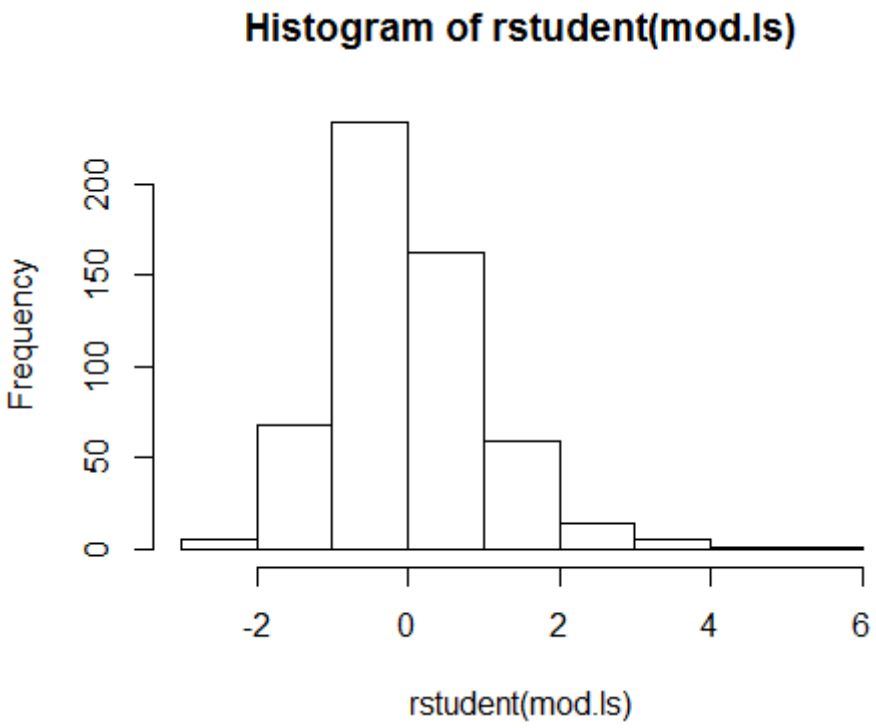
Στη συνέχεια, για την παλινδρόμηση με τη μέθοδο *Ελαχίστων Τετραγώνων (OLS)*, της χοληστερίνης σε σχέση με την ηλικία, το δείκτη μάζας σώματος, τη φυλή, το κάπνισμα και το αλκοόλ, πληκτρολογούμε τις παρακάτω εντολές:

```
> mod.ls <- lm(LDL~AGE+BMI+NW+SMOKE+ALCOHOL, data=Hers549)
> summary(mod.ls)

Call:
lm(formula = LDL ~ AGE + BMI + NW + SMOKE + ALCOHOL, data = Hers549)

Residuals:
    Min      1Q  Median      3Q     Max
-99.603 -24.186  -5.302  20.798 193.801

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 144.7655    20.8383   6.947 1.07e-11 ***
AGE          -0.2314     0.2528  -0.915   0.3604
BMI           0.3792     0.2869   1.322   0.1868
NW           10.9118     4.6908   2.326   0.0204 *
SMOKE        11.6736     4.7949   2.435   0.0152 *
ALCOHOL      -0.4273     3.4281  -0.125   0.9008
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.86 on 543 degrees of freedom
Multiple R-squared:  0.0286,    Adjusted R-squared:  0.01966
F-statistic: 3.198 on 5 and 543 DF,  p-value: 0.007433
```

Η μέθοδος *Ελαχίστων Τετραγώνων (OLS)*, δείχνει ότι οι μεταβλητές *NW* (*p=0.0174*) και *SMOKE* (*p=0.0157*) επιδρούν σημαντικά στη χοληστερίνη *LDL*.

Τώρα, θα εξετάσουμε τα δεδομένα μας χρησιμοποιώντας διαγνωστικές μεθόδους. Αρχικά, υπολογίζουμε τα *τυποποιημένα υπόλοιπα (Studentized residuals)* και στη συνέχεια πληκτρολογώντας τις ακόλουθες εντολές, εμφανίζεται το ιστόγραμμα των τυποποιημένων υπολοίπων.

```
> hist(rstudent(mod.ls))
```



Εικόνα 5.11 Ιστόγραμμα των τυποποιημένων υπολοίπων.

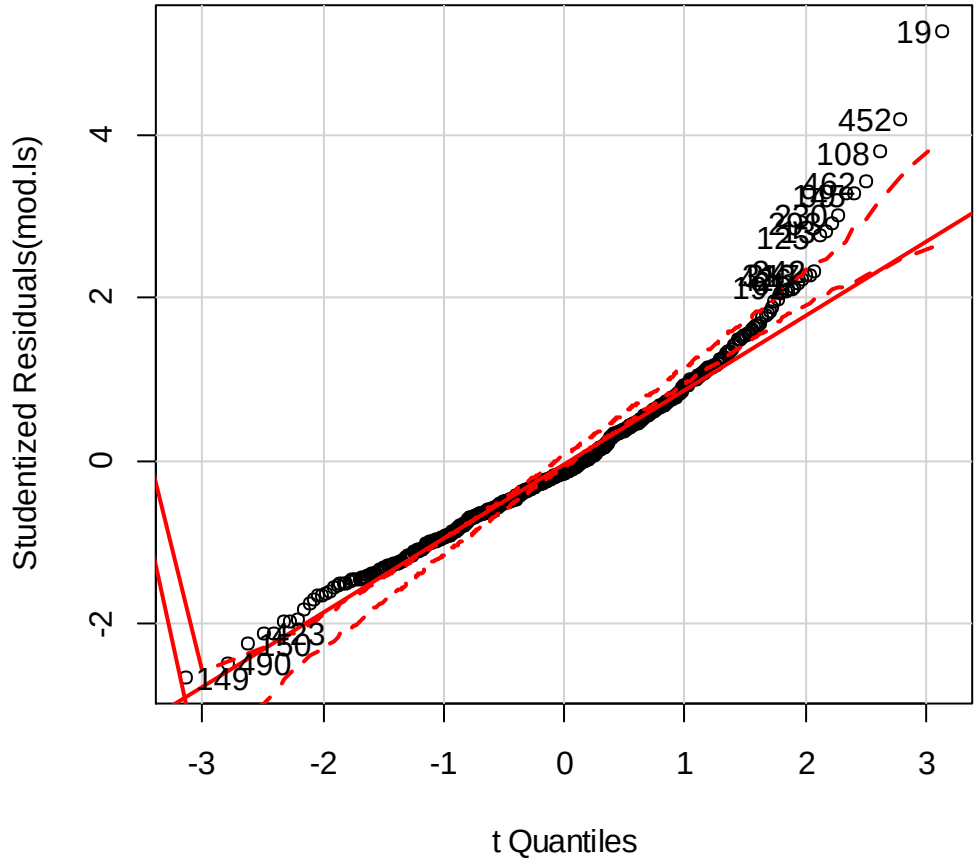
Από την Εικόνα 5.11 παρατηρούμε ότι, τα τυποποιημένα υπόλοιπα ακολουθούν την κατανομή t-Student.

Η συνάρτηση *qqPlot* από το πακέτο *car*, υπολογίζει τα τυποποιημένα υπόλοιπα και τα παρουσιάζει σε σχέση με τα αναμενόμενα ποσοστιαία σημεία της κατανομής *t*. (*quantile-comparison plots*).

Παρατηρούμε ότι, με εξαίρεση κάποιων παρατηρήσεων στη δεξιά πλευρά της κατανομής, τα τυποποιημένα υπόλοιπα ακολουθούν την κατανομή t-Student.

Αυτό επιβεβαιώνεται και από τον πιο κάτω γραφικό έλεγχο *qqplot* (Εικόνα 5.12) της κατανομής *t*, όπου είναι εμφανής η απόκλιση αρκετών σημείων από την προσαρμοσμένη ευθεία.

```
> qqPlot(mod.ls, labels=row.names(Hers549), id.n=20)
149 490 150 423 191  61 426 313 217 342 125  13 293 230  99 145 462 108 452  19
  1   2   3   4 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549
```



Εικόνα 5.12 95% περιοχή εμπιστοσύνης (confidence envelope) για τα τυποποιημένα υπόλοιπα.

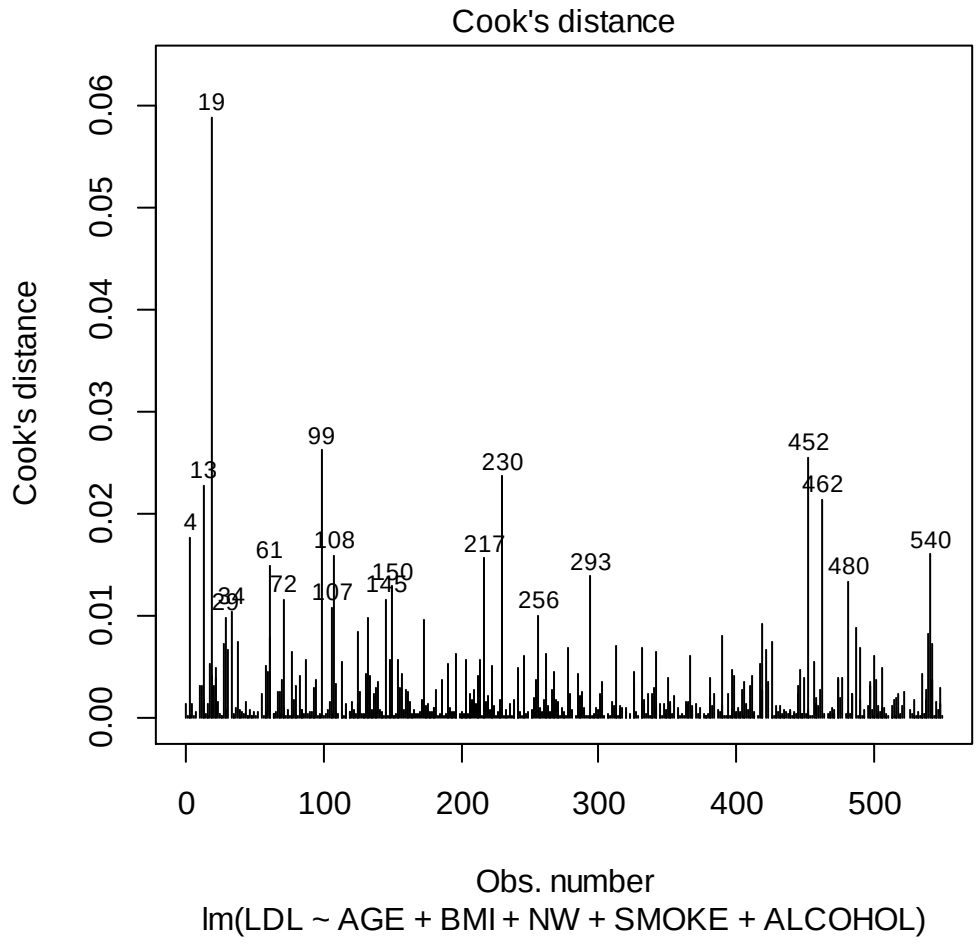
Στην περίπτωσή μας, μία ομάδα υπολοίπων ξεφεύγει αρκετά από την γραμμή σύγκρισης και στα δύο άκρα. Το γεγονός αυτό υποδηλώνει ότι η κατανομή των υπολοίπων έχει σχετικά μεγάλο βάρος στις ουρές (*heavy-tailed*). Εξ ορισμού, η *qqPlot* δημιουργεί μία 95% περιοχή εμπιστοσύνης (*confidence envelope*) για τα τυποποιημένα υπόλοιπα.

Τώρα, ελέγχουμε αν οι μεταβλητές συσχετίζονται μεταξύ τους έντονα, δηλαδή αν υπάρχει πολυσυγγραμικότητα. Στη συνέχεια, ελέγχουμε για σημεία *υψηλής μόχλευσης* (leverage points) και *επιρρεάζουσες παρατηρήσεις* (*influence observations*). Θα χρησιμοποιήσουμε τις γραφικές παραστάσεις *hat-values* και *Cook's distances*. Έχουμε:

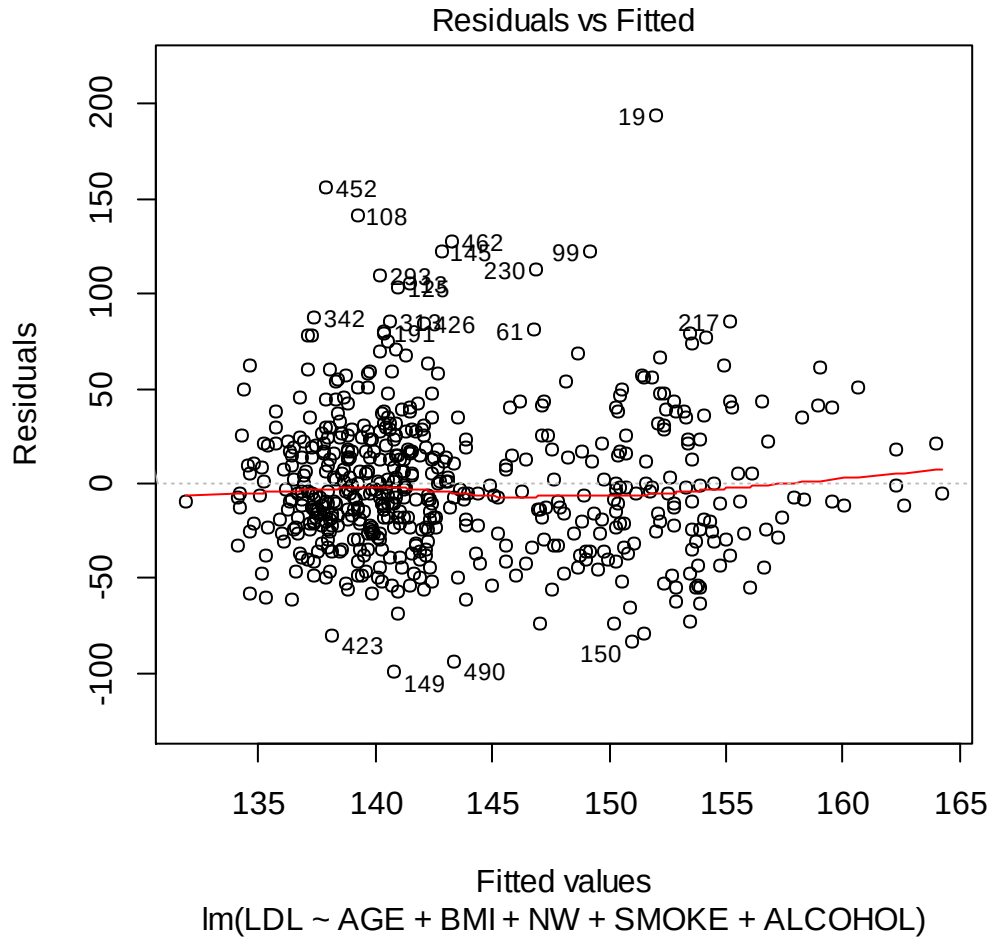
```
> vif(mod.ls)
     AGE      BMI       NW    SMOKE  ALCOHOL
1.127853 1.116357 1.070481 1.118532 1.037613
> cutoff <- 4/((nrow(Hers549)-length(mod.ls$coefficients)-2))
> plot(mod.ls, which=4, cook.levels=cutoff, id.n=20)
```

Φαίνεται να μην υπάρχει πρόβλημα πολυσυγγραμικότητας, καθώς οι τιμές της *vif* για την κάθε μεταβλητή, είναι μικρότερες από 5, δηλαδή *vif*<5 (βλέπε σελίδα 7).

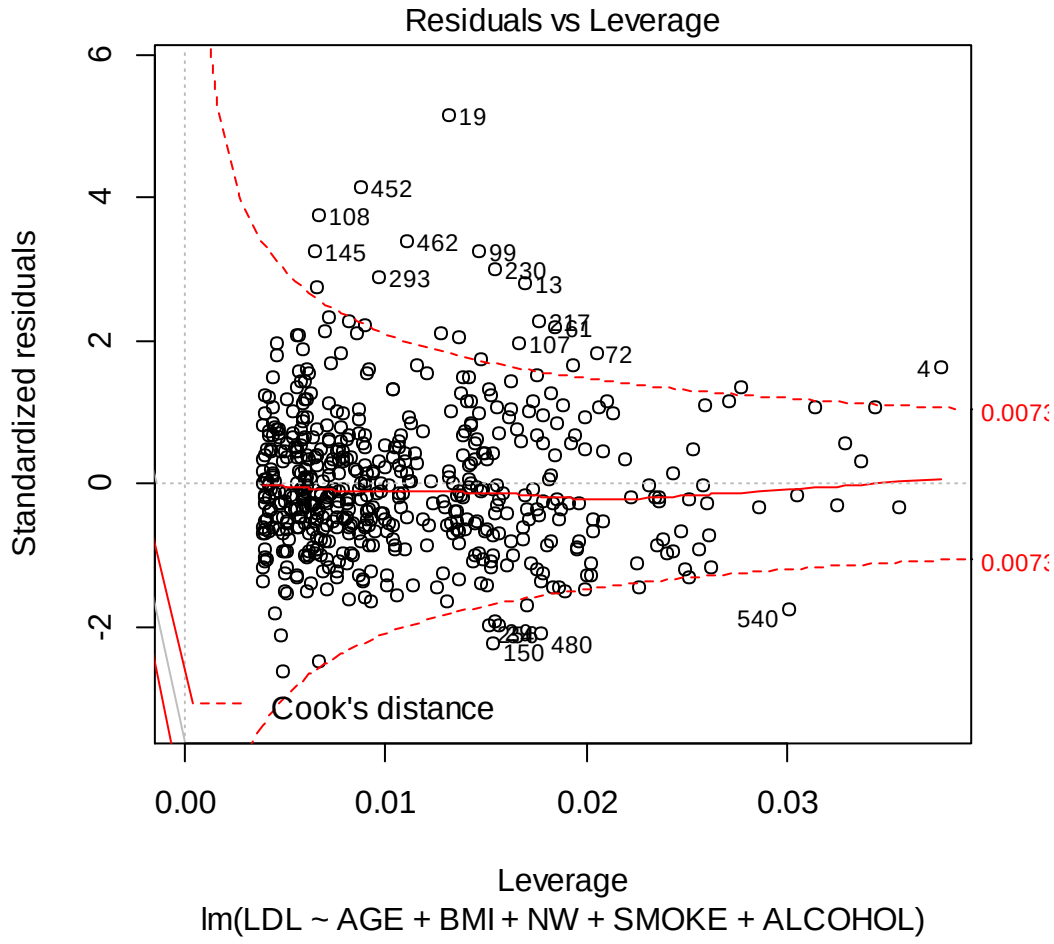
Παρακάτω έχουμε τις γραφικές παραστάσεις που προέκυψαν από τις παραπάνω εντολές.



```
> plot(mod.ls, which=1, cook.levels=cutoff, id.n=20)
```




```
> plot(mod.ls, which=5, cook.levels=cutoff, id.n=20)
```



Εικόνα 5.13 Γραφικές παραστάσεις hat-values και Cook distances.

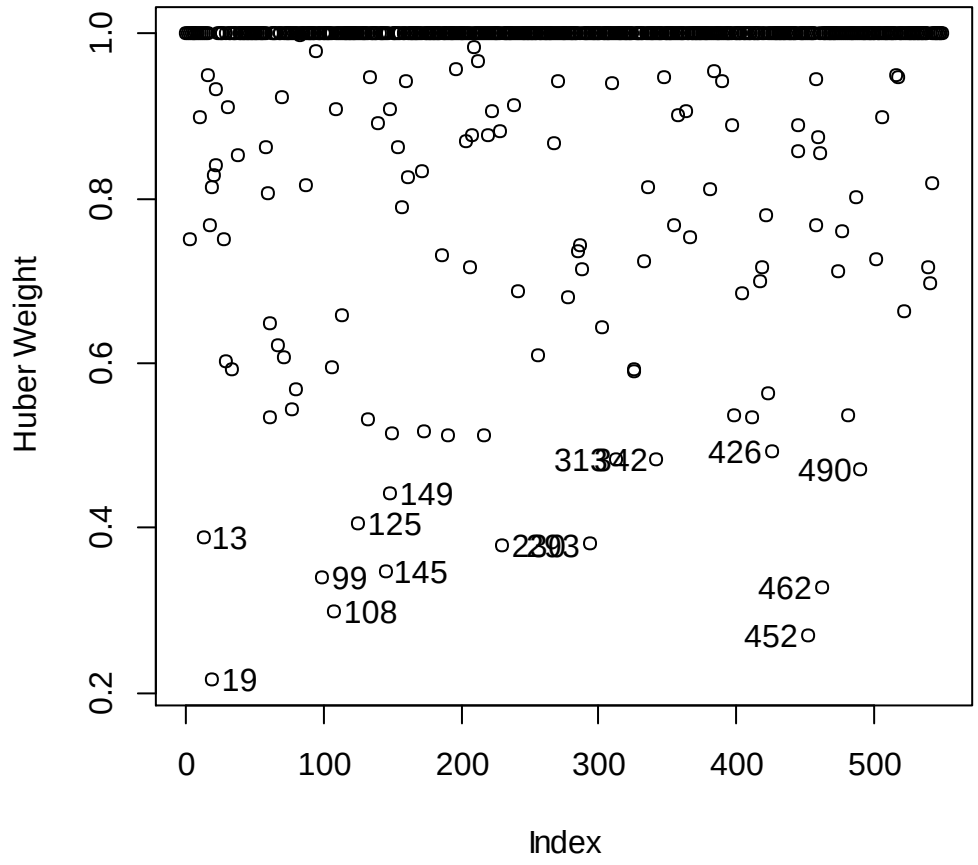
Εναλλακτικά, ας χρησιμοποιήσουμε την *M-εκτιμήτρια* του *Huber*, χρησιμοποιώντας την εντολή *rlm (robust linear model)* μέσα από τη βιβλιοθήκη MASS, έχουμε:

```
> mod.huber <- rlm(LDL~AGE+BMI+NW+SMOKE+ALCOHOL, data=Hers549)
> summary(mod.huber)

Call: rlm(formula = LDL ~ AGE + BMI + NW + SMOKE + ALCOHOL, data = Hers549)
Residuals:
    Min      1Q  Median      3Q     Max
-98.016 -21.326  -3.065  22.586 199.527

Coefficients:
            Value    Std. Error t value
(Intercept) 126.5259  19.6030     6.4544
AGE          -0.0671   0.2378    -0.2822
BMI           0.5390   0.2699     1.9970
NW            8.4842   4.4127     1.9227
SMOKE        15.4396   4.5106     3.4229
ALCOHOL      -0.2081   3.2249    -0.0645

Residual standard error: 32.19 on 543 degrees of freedom
```

Η παραπάνω μέθοδος δείχνει ότι εκτός από τις μεταβλητές *NW* και *SMOKE*, η ανθεκτική ανάλυση παρουσιάζει και την *BMI* ότι έχει σημαντική επίδραση στην εξαρτημένη μεταβλητή *LDL*.

Εδώ είναι χρήσιμο να διεξάγουμε τη γραφική παράσταση (Εικόνα 5.14) με τα τελικά βάρη που χρησιμοποιήθηκαν στην παραπάνω ανθεκτική εκτίμηση. Η συνάρτηση *showLabels* από το πακέτο *car* χρησιμοποιείται για την επισήμανση όλων των παρατηρήσεων με βάρη λιγότερο από *0.5*:

```
> plot(mod.huber$w, ylab="Huber Weight")
> smallweights <- which(mod.huber$w<0.5)
> showLabels(1:549, mod.huber$w, rownames(Hers549), id.method=smallweights, cex.=.5)
 13  19  99 108 125 145 149 230 293 313 342 426 452 462 490
 13  19  99 108 125 145 149 230 293 313 342 426 452 462 490
```



Εικόνα 5.14 Γραφική παράσταση με τα τελικά βάρη που χρησιμοποιήθηκαν στην μέθοδο του *Huber*.

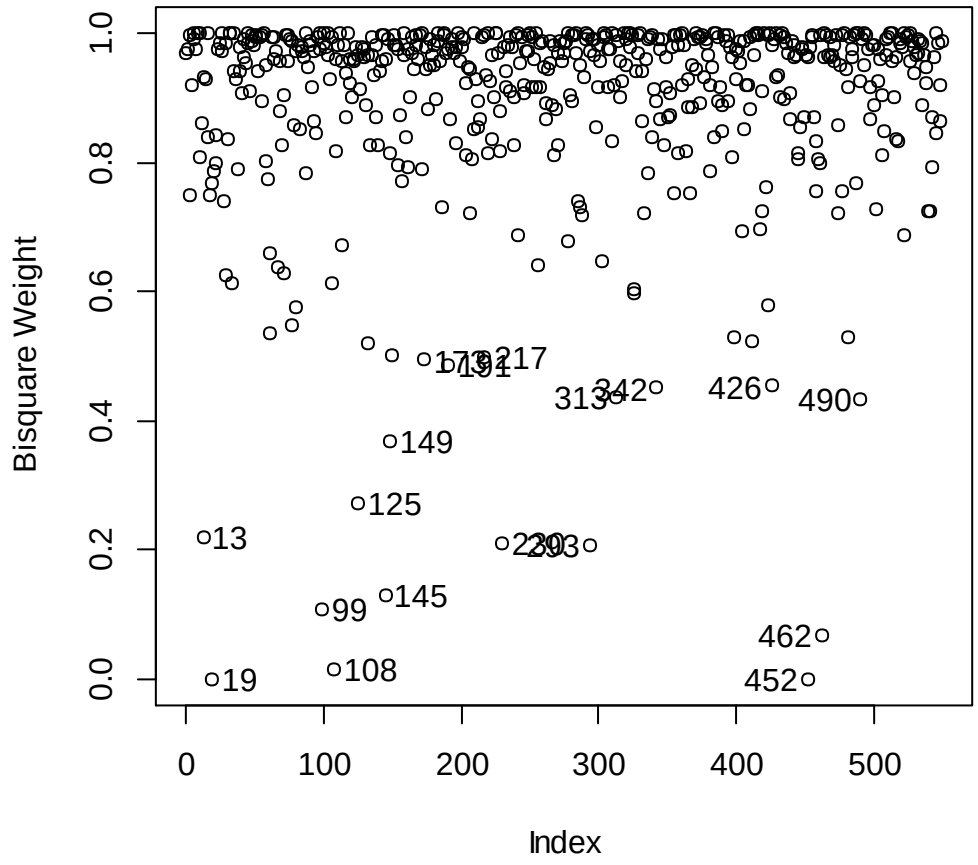
Στη συνέχεια θα χρησιμοποιήσουμε την εκτιμήτρια *bisquare*, που προσαρμόζεται από την εντολή *rlm*. Θα έχουμε λοιπόν:

```
> mod.bisq <- rlm(LDL~AGE+BMI+NW+SMOKE+ALCOHOL, data=Hers549, method="MM")
> summary(mod.bisq)

Call: rlm(formula = LDL ~ AGE + BMI + NW + SMOKE + ALCOHOL, data = Hers549,
    method = "MM")
Residuals:
    Min      1Q  Median      3Q     Max
-97.709 -20.638  -2.523  22.894 200.853

Coefficients:
            Value    Std. Error t value
(Intercept) 118.6095  19.3894     6.1172
AGE           0.0237   0.2352     0.1008
BMI           0.5778   0.2670     2.1643
NW            8.2375   4.3646     1.8873
SMOKE        16.4373   4.4615     3.6843
ALCOHOL       0.1813   3.1897     0.0568

Residual standard error: 33.24 on 543 degrees of freedom
```

Συγκρίνοντας σε σχέση με τις *Huber εκτιμήτριες*, οι τιμές της *bisquare* εκτίμησης των συντελεστών είναι ελάχιστα μεγαλύτερες.

Στην Εικόνα 5.15 παρουσιάζεται το γράφημα με τα βάρη της *bisquare* προσαρμογής, αναγνωρίζοντας τις παρατηρήσεις με τα χαμηλότερα βάρη:

```
> plot(mod.bisq$w, ylab="Bisquare Weight")
> showLabels(1:549, mod.bisq$w, rownames(Hers549), id.method=which(mod.bisq$w<0.5), cex.=.5)
 13  19  99 108 125 145 149 173 191 217 230 293 313 342 426 452 462 490
```



Εικόνα 5.10 Γραφική παράσταση με τα βάρη της *bisquare* προσαρμογής.

Τώρα, θα κάνουμε χρήση της μεθόδου *Ελαχίστων Τετραγώνων*, αλλά παραλείποντας την ομάδα των παρατηρήσεων που αποκλίνουν. Η ομάδα σημείων που υποδεικνύονται από τις διαγνωστικές μεθόδους και κυρίως από τις ανθεκτικές τεχνικές που χρησιμοποιήσαμε παραπάνω αποτελείται από τις παρακάτω παρατηρήσεις. Πληκτρολογούμε τις ακόλουθες εντολές:

```
> mod.ls.2 <- update(mod.ls, subset=-c(4,13,19,99,108,125,145,149,191,217,230,293,313,342,426,452,462,490,540))
> summary(mod.ls.2)

Call:
lm(formula = LDL ~ AGE + BMI + NW + SMOKE + ALCOHOL, data = Hers549,
    subset = -c(4, 13, 19, 99, 108, 125, 145, 149, 191, 217,
        230, 293, 313, 342, 426, 452, 462, 490, 540))

Residuals:
    Min      1Q  Median      3Q     Max
-81.505 -21.314  -2.909  20.618  82.895

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 114.89467   18.08821   6.352 4.62e-10 ***
AGE           0.09654    0.21882   0.441  0.65928
BMI           0.55793    0.25009   2.231  0.02611 *
NW            9.97050    4.01980   2.480  0.01344 *
SMOKE        14.80893    4.13177   3.584  0.00037 ***
ALCOHOL      -1.30457    2.94137  -0.444  0.65757
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.92 on 524 degrees of freedom
Multiple R-squared:  0.04133,   Adjusted R-squared:  0.03218
F-statistic: 4.518 on 5 and 524 DF,  p-value: 0.0004836
```

Παρατηρούμε λοιπόν ότι, παραλείποντας την παραπάνω ομάδα των ακραίων παρατηρήσεων, ο συντελεστής της μεταβλητής *AGE* έχει τώρα θετικό πρόσημο, ο συντελεστής *BMI* συμβάλει σημαντικά στην επεξήγηση της χοληστερίνης *LDL*. Ο συντελεστής της φυλής *(NW)* έχει μειωθεί κατά περίπου 20% αλλά σχετίζεται σημαντικά με την *LDL*. Ο συντελεστής του αν καπνίζουνε, αυξήθηκε κατά περίπου 30% και βεβαία είναι στατιστικά σημαντικός. Παρότι ο συντελεστής της μεταβλητής *ALCOHOL* μειώθηκε κατά περίπου 35%, παραμένει στατιστικά μη-σημαντικός.

Όμως, όπως αναφέραμε εξ αρχής στη παρούσα διπλωματική, η παράλειψη ακόμη και μίας παρατήρησης μπορεί να αλλάξει σημαντικά αποτελέσματα, σε μη-σημαντικά.

Στη συνέχεια, θα προσαρμόσουμε το μοντέλο μας με την μέθοδο *Ελαχίστων Περικοπτόμενων Τετραγώνων (LTS)*. Γι' αυτό, θα χρησιμοποιήσουμε την συνάρτηση *ltsreg* μέσα από το πακέτο *lqs*. Η *LTS* παλινδρόμηση είναι η προεπιλεγμένη μέθοδος της συνάρτησης *lqs*, η οποία επιπλέον μπορεί να εκτιμήσει κι άλλες εκτιμήτριες *φραγμένης επιρροής* (*bounded-influence*). Έχουμε λοιπόν:

```
> (mod.lts <- ltsreg(LDL~AGE+BMI+NW+SMOKE+ALCOHOL, data=Hers549))
Call:
lqs.formula(formula = LDL ~ AGE + BMI + NW + SMOKE + ALCOHOL,
    data = Hers549, method = "lts")

Coefficients:
(Intercept)          AGE          BMI           NW        SMOKE      ALCOHOL
   84.44394     -0.03868      1.76730     -7.05547     65.43839     -0.16033

Scale estimates 32.47 32.55

There were 31 warnings (use warnings() to see them)
```

Επίσης, μπορούμε να χρησιμοποιήσουμε την *Τεταρτομοριακή Παλινδρόμηση (Quantile Regression)* στην πολλαπλή ανάλυση παλινδρόμησης. Προχωράμε λοιπόν στην εκτίμηση των συντελεστών του μοντέλου μας, χρησιμοποιώντας την *L1* εκτιμήτρια της τεταρτομοριακής παλινδρόμησης:

```
> mod.quant <- rq(LDL~AGE+BMI+NW+SMOKE+ALCOHOL, data=Hers549)
> summary(mod.quant)

Call: rq(formula = LDL ~ AGE + BMI + NW + SMOKE + ALCOHOL, data = Hers549)

tau: [1] 0.5

Coefficients:
            coefficients lower bd  upper bd
(Intercept) 125.75004     88.16643 169.66084
AGE          -0.11957     -0.69721   0.33959
BMI           0.60860      0.22343   1.12271
NW            4.17879     -3.88272  14.50081
SMOKE        15.23673      9.03672  21.65768
ALCOHOL      -0.53262     -6.35444   4.34910
```

Βλέπουμε ότι, οι εκτιμήσεις των συντελεστών με την *L1* εκτιμήτρια είναι παρόμοιες με εκείνες της *M-εκτιμήτριας* η οποία βασίζεται στη συνάρτηση βάρους του *Huber*.

Στον πίνακα που ακολουθεί συνοψίζουμε τις τιμές των συντελεστών για την κάθε εκτιμήτρια που χρησιμοποιήσαμε στην αναλυσή μας.

Μέθοδος	b_0	b_1 (*AGE*)	b_2 (*BMI*)	b_3 (*NW*)	b_4 (*SMOKE*)	b_5 (*ALCOHOL*)
1) *ET (OLS)*	144.7655	-0.2314	0.3792	10.9118	11.6736	-0.4273
2) *ET χωρίς τα outlier*	112.07814	0.11680	0.54877	9.20733	17.94632	-0.01447
3) *Huber M-εκτιμήτρια*	126.5259	-0.0671	0.5390	8.4842	15.4396	-0.2081
4) *Bisquare MM- εκτιμήτρια*	118.6095	0.0237	0.5778	8.2375	16.4373	0.1813
5) *ΕΠΤ (LTS)*	84.44394	-0.03868	1.76730	-7.05547	65.43839	-0.16033
6) *L1-εκτιμήτρια*	125.75004	-0.11957	0.60860	4.17879	15.23673	-0.53262

Πίνακας 5.2 Διάφορες εκτιμήτριες του μοντέλου παλινδρόμησης της χοληστερίνης σε σχέση με την ηλικία, το δείκτη μάζας σώματος, τη φυλή, το κάπνισμα και το αλκοόλ.

Τελικά, όπως και στις δύο προηγούμενες εφαρμογές, τα αποτελέσματα που προκύπτουν από τις ανθεκτικές μεθόδους, διαφέρουν σημαντικά σε σχέση με τις εκτιμήσεις της *μεθόδου ελαχίστων τετραγώνων*. Οπότε η ύπαρξη σημείων μόχλευσης και σημείων επιρροής, αλλοιώνει αρκετά το αποτέλεσμα. Συνεπώς, η εφαρμογή μεθόδων ανθεκτικής παλινδρόμησης κρίνεται απαραίτητη.

Επιπλέον, όπως και με τις προηγούμενες δύο εφαρμογές παρατηρούμε ότι οι εκτιμήσεις των συντελεστών με την μέθοδο *LTS* διαφοροποιούνται αισθητά από τις εκτιμήσεις των μεθόδων 2, 3, 4 και 6.

Γενικά Συμπεράσματα

Στόχος της ανθεκτικής στατιστικής είναι να περιγράψει τη δομή στην οποία προσαρμόζεται καλύτερα το πλήθος των δεδομένων, να αναγνωρίσει τις ακραίες παρατηρήσεις καθώς επίσης να προειδοποιήσει για τα σημεία μόχλευσης. Η υπεροχή των ανθεκτικών εκτιμητριών έναντι της μεθόδου ελαχίστων τετραγώνων σε αλλοιωμένα δεδομένα είναι αναμφισβήτητη.

Κάτω από ορισμένες συνθήκες, οι *M-εκτιμήτριες* μπορεί να είναι ευάλωτες σε σημεία υψηλής μόχλευσης (high leverage points). Στη μελέτη μας, παρουσιάσαμε εκτιμήτριες με πολύ-υψηλό σημείο κατάρρευσης φραγμένης επιρροής (very-high breakdown bounded-influence estimators) και αντίστοιχες συναρτήσεις του στατιστικού πακέτου *R*. Θα πρέπει να αποφεύγονται οι εκτιμήτριες πολύ υψηλού σημείου κατάρρευσης, εκτός αν πιστεύουμε ότι το μοντέλο που προσαρμόζουμε είναι κατάλληλο, διότι αυτές οι εκτιμήτριες δεν επιτρέπουν τη διάγνωση εσφαλμένου προσδιορισμού του μοντέλου (Cook et al., 1992).

Μία εκτιμήτρια φραγμένης επιρροής είναι η *Ελαχίστων Περικοπτόμενων Τετραγώνων (LTS)*, η οποία όμως έχει περίπλοκο μηχανισμό στην προσαρμογή της (Rousseeuw and Leroy, 1987). Επιπλέον, οι εκτιμήτριες φραγμένης επιρροής μπορεί να δώσουν παράλογα αποτελέσματα σε ορισμένες περιπτώσεις (Stefanski, 1991) και δεν υπάρχει απλή φόρμουλα για τα τυπικά σφάλματα των συντελεστών.

Μια εφαρμογή των εκτιμητριών φραγμένης επιρροής είναι η παροχή τιμών εκκίνησης για την *M-εκτιμήτρια*. Η διαδικασία αυτή, σε συνδυασμό με τη χρήση της εκτιμήτριας φραγμένης-επιρροής της διακύμανσης σφάλματος, παράγει τη λεγόμενη *MM-εκτιμήτρια*. Η εκτιμήτρια αυτή διατηρεί το υψηλό σημείο κατάρρευσης της εκτιμήτριας φραγμένης επιρροής. Ακόμη, υπό την κανονική κατανομή μοιράζεται την σχετικά υψηλή απόδοση της παραδοσιακής *M-εκτιμήτριας*. Οι εκτιμήτριες-*MM* είναι ιδιαίτερα ελκυστικές όταν συνδυάζονται με συναρτήσεις όπως την *bisquare* οι οποίες είναι ευάλωτες στις τιμές εκκίνησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Ζιούτας Γ. (1992). *Ανθεκτικοί Εκτιμητές Παλινδρόμησης. Προσέγγιση με Μαθηματικό Προγραμματισμό,* ΑΠΘ, Θεσσαλονίκη. Διδακτορική διατριβή.

Οικονόμου Π. και Καρώνη Χ. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης*. Εκδόσεις Συμεών.

Ξένη

Andersen, R. (2008). *Modern Methods for Robust Regression*. Sage, Thousand Oaks, CA.

Atkinson, A. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer, New York.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley, Chichester. 3rd ed.

Chen, C. (2002). Robust Regression and Outlier Detection with the ROBUSTREG Procedure, 1-13. (*www2.sas.com/proceedings/sugi27/p265-27.pdf).*

Cook, R. D., Hawkins, D. M. and Weisberg, S. (1992). **Comparison of model misspecification diagnostics using residuals from OLS and high breakdown estimates.** *Journal of the American Statistical Association*, **87**, 419-424.

De Long, J.B. and Summers, L.H. (1991). Equipment investment and economic growth. *Quarterly Journal of Economics*, **106**, 445-501.

Duncan, O.D. (1961). A socioeconomic index for all occupations. In J. Reiss, Jr. (Ed.), *Occupations and Social Status*. Free Press of Glencoe, New York, pp. 109–138.

Fox, J. (2002). A*n R and S-PLUS Companion to Applied Regression.* Sage, Thousand Oaks, CA.

Fox, J. (2002). Robust Regression: Appendix to an R and S-PLUS Companion to Applied Regression*,* 1-8*. (*www.saedsayad.com/docs/**RobustRegression**.pdf*)*

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. Sage, Thousand Oaks, CA. 2nd ed.

Fox, J. and Weisberg, S. (2010). Robust Regression in R, 1-17. (https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Robust-Regression.pdf.)

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage, Thousand Oaks, CA. 2nd ed.

Grubbs, F. (1969). *Procedures for detecting outlying observations in samples. Technometrics, **11**, 1-21.*

*Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. Annals of Mathematical Statistics, **5**, 1449-1458.*

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383-393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The approach based on Influence Functions*, Wiley, New York.

Hawkins, D. M. (1980). *Identification of Outliers.* Chapman and Hall, London.

Huber, P. J. (1964). Robust estimation of a location parameter*. Annals of Mathematical Statistics,* **35**, 73-101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo, *The Annals of Statistics*, **1**, 799-821.

Huber, P. J. (1977). Huber, P. J. (1977), Robust covariances, in *Statistical Decision Theory and*
Related Topics, Vol. 2., edited by *S. S.* Gupta and D. *S.* Moore, Academic Press, New York, pp. 165-191.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Huber, P. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley, Hoboken NJ. 2nd ed.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press, Cambridge.

Mallows, C. L. (1979). Robust methods-some examples of their use. *American Statistician*, **33**, 179-184.

Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871-880.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley, Hoboken, NJ.

Stefanski, L. A. (1991). *A note on high-breakdown estimators. Statistics and Probability Letters*, **11**, 353-358.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, **15**, 642-656.

Zaman, A., Rousseeuw, P. J. and Orhan, M. (2001). Econometric applications of high breakdown robust regression techniques. *Econometrics Letters*, **71**, 1-8.