Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΙΧΝΕΥΣΗ ΕΚΤΟΠΩΝ ΜΕΤΡΗΣΕΩΝ, ΔΙΑΓΝΩΣΤΙΚΗ ΚΑΙ ΑΝΘΕΚΤΙΚΗ ΕΚΤΙΜΗΣΗ: ΑΝΑΣΚΟΠΗΣΗ

Χριστίνα Ζαχαροπούλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Ιούλιος 2004

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΙΧΝΕΥΣΗ ΕΚΤΟΠΩΝ ΜΕΤΡΗΣΕΩΝ, ΔΙΑΓΝΩΣΤΙΚΗ ΚΑΙ ΑΝΘΕΚΤΙΚΗ ΕΚΤΙΜΗΣΗ: ΑΝΑΣΚΟΠΗΣΗ

Χριστίνα Ζαχαροπούλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Ιούλιος 2004

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Γεώργιος Πιτσέλης (Επιβλέπων)
- Τάκης Παπαϊωάννου
- Μάρκος Κούτρας

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

IDENTIFICATION OF OUTLIERS, DIAGNOSTICS AND ROBUST ESTIMATION: REVIEW

By

Christina Zacharopoulou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
July 2004

Ευχαριστίες

Πριν από την παρουσίαση της διπλωματικής εργασίας θεωρώ υποχρέωσή μου να εκφράσω τις ευχαριστίες μου στον Λέκτορα κ. Γεώργιο Πιτσέλη που δέχτηκε να αναλάβει την επίβλεψη της μελέτης αυτής, για την επιστημονική καθοδήγηση, την ηθική υποστήριξη, καθώς και για τις υποδείξεις του κατά τη διάρκεια της συγγραφής και στο τελικό κείμενο.

Επίσης θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Τάκη Παπαϊωάννου καθώς και τον Καθηγητή κ. Μάρκο Κούτρα που συμμετείχαν στην τριμελή επιτροπή.

Τέλος, νιώθω την ανάγκη να ευχαριστήσω με όλη μου την καρδιά την οικογένειά μου και τους φίλους μου που με στήριξαν αποφασιστικά σε όλη τη διάρκεια των σπουδών μου. Την αδερφή μου Μαριάννα, που με μεγάλη φροντίδα διάβασε το κείμενο και μου υπέδειξε γλωσσικές διορθώσεις, την ευχαριστώ ιδιαίτερα.

Χριστίνα Ζαχαροπούλου

Ιούλιος 2004

Περίληψη

Η ανάλυση παλινδρόμησης είναι ένα σημαντικό στατιστικό εργαλείο που εφαρμόζεται στις περισσότερες επιστήμες. Η πιο γνωστή από όλες τις μεθόδους παλινδρόμησης που υπάρχουν είναι η μέθοδος ελαχίστων τετραγώνων χάρη στην ευκολία υπολογισμού της. Παρόλο αυτά όμως, ιδιαίτερα διαδεδομένος είναι ο κίνδυνος που υπάρχει οφειλόμενος στις έκτοπες παρατηρήσεις. Προκειμένου να αντιμετωπιστεί αυτό το πρόβλημα έχουν αναπτυχθεί νέες στατιστικές μέθοδοι οι οποίες δεν επηρεάζονται από τις έκτοπες παρατηρήσεις. Αναφερόμαστε στις γνωστές ανθεκτικές μεθόδους τα αποτελέσματα των οποίων είναι αξιόπιστα ακόμα και αν ένα μεγάλο μέρος των δεδομένων μας είναι αλλοιωμένο. Η ανθεκτική παλινδρόμηση παρέχει μία εναλλακτική λύση όταν οι βασικές υποθέσεις δεν ικανοποιούνται από το σύνολο των δεδομένων.

Ως εκ τούτου, ο πρωταρχικός στόχος της ανθεκτικής στατιστικής είναι να αναπτύξει ανθεκτικές μεθόδους ενάντια στην ύπαρξη έκτοπων παρατηρήσεων. Ο έλεγχος της ακρίβειας και της σταθερότητας των εκτιμητών μέσω των ανθεκτικών μεθόδων είναι απαραίτητος. Συνεπώς το ενδιαφέρον μας επικεντρώνεται στην αναγκαιότητα των ανθεκτικών εκτιμητών προκειμένου να δείξουμε την αδυναμία της μεθόδου ελαχίστων τετραγώνων να χειριστεί τις έκτοπες παρατηρήσεις. Πιο συγκεκριμένα γίνεται αναφορά στις πιο γνωστές τεχνικές ανθεκτικής παλινδρόμησης όπως M, GM, S, LMS, LTS και MM-εκτιμητές. Τέλος, βάσει παραδειγμάτων υπολογίζονται οι προαναφερθέντες ανθεκτικοί εκτιμητές και γίνεται σύγκριση μεταξύ αυτών.

Abstract

Regression analysis is an important statistical tool that is routinely applied in most sciences. Out of many possible regression techniques, the least squares method has been generally adopted because of tradition and ease of computation. However, there is presently a widespread awareness of the dangers posed by the occurrence of outliers events. To remedy this problem, new statistical techniques have been developed that are not so easily affected by outliers. These are robust (or resistant) methods the results of which remain trustworthy even if a certain amount of data is contaminated. Robust regression analysis provides an alternative to a least regression model when fundamental assumptions are unfulfilled by the nature of the data.

Therefore, a major goal of robust statistics is to develop methods that are robust against the possibility that one or several outliers may occur anywhere in the data. It is therefore important to check the accuracy and stability of estimates using robust estimation methods. Our interest will be focused on the importance of robust estimators in order to present the problem of least squares method to handle outliers. In specific, the most known techniques of robust regression such as L, R, M, GM, S-estimators, LMS, LTS and MM-estimators are referred. Finally by means of two examples, the above-mentioned estimators will be compared.

Περιεχόμενα

Κατάλογος Πινάκων			x
Κατάλογος Σχημάτων			xi
Κατάλογος Συντομογραφιών			xii
1.	Εισαγωγή		1
	1.1	Η αδυναμία της μεθόδου ελαχίστων τετραγώνων	1
	1.1.1	Οι υποθέσεις της μεθόδου ελαχίστων τετραγώνων	2
	1.2	Η φύση της ανθεκτικής στατιστικής	3
	1.2.1	Οι στόχοι της ανθεκτικής στατιστικής	6
	1.2.2	Η σημασία των ανθεκτικών διαδικασιών	7
	1.3	Οι πρώτες προσεγγίσεις	8
2.	Απομονωμένες τιμές		11
	2.1	Ορισμός απομονωμένων τιμών	11
	2.1.1	Ταξινόμηση έκτοπων παρατηρήσεων	12
	2.1.2	Θέση και επίδραση απομονωμένων τιμών	13
	2.2	Απόρριψη έκτοπων παρατηρήσεων	16
	2.3	Συμπεριφορά των κανόνων απόρριψης σχετικά με την ανθεκτική εκτίμηση	17
	2.4	Λόγοι για τους οποίους ο εντοπισμός των έκτοπων παρατηρήσεων και η απομάκρυνσή τους από το δείγμα δεν θεωρείται αρκετός	18
3.	Δύο βασικές προσεγγίσεις αντιμετώπισης έκτοπων παρατηρήσεων		19
	3.1	Εισαγωγή	19
	3.2	Οι εκτιμητές και οι βασικές ιδιότητές τους	20
	3.3	Το κλασικό πρόβλημα παλινδρόμησης	21
	3.4	Διαγνωστική παλινδρόμηση	22
	3.5	Ανθεκτική παλινδρόμηση	23
4.	Μέτρα ανθεκτικότητας		25
	4.1	Εισαγωγή	25
	4.2	Ανθεκτικό συμπέρασμα	25
	4.3	Συναρτησοειδή	26
	4.3.1	Συνάρτηση Επίδρασης (ΣΕ) – Influence Function	26
	4.3.2	Σημείο Κατάρρευσης (ΣΚ) – Breakdown point	28
	4.3.3	Ευαισθησία Γενικού Σφάλματος ΕΓΣ – Gross Error Sensitivity	29
	4.3.4	Καμπύλη Ευαισθησίας ΚΕ- Sensitivity Curve	30
	4.4	Ορισμοί (Rousseeuw και Leory, 1987)	30
	4.5	Συνάρτηση επίδρασης για τα ελάχιστα τετράγωνα	31
5.	Ανθεκτικοί Εκτιμητές		33
	5.1	Οι βασικότεροι ανθεκτικοί εκτιμητές	33
	5.2	L- εκτιμητές	35
	5.3	R-εκτιμητές	36

	5.4	Μ-εκτιμητές	37
	5.5	Μέθοδοι ανθεκτικής παλινδρόμησης	40
	5.5.1	Μ-εκτίμηση παλινδρόμησης του Huber	40
	5.5.2	Ο αλγόριθμος των Μ-εκτιμητών	41
	5.5.3	Γενικευμένοι εκτιμητές (ΓΕ) - GM-estimators	42
	5.5.4	S-εκτιμητές	42
	5.5.5	Ελάχιστα Διάμεσα Τετράγωνα (ΕΔΤ) – Least Median Squares	43
	5.5.6	Ελάχιστα Περικεκομμένα Τετράγωνα (ΕΠΤ) – Least Trimmed Squares	45
	5.5.7	ΜΜ-εκτιμητές	47
	5.6	Εκτιμητές υψηλού σημείου κατάρρευσης	49
6.	Εφαρμογή ανθεκτικών εκτιμητών		52
	6.1	Απλή παλινδρόμηση	52
	6.2	Πολλαπλή παλνδρόμηση	63
	6.3	Συμπεράσματα	67
7.	Παραρτήματα		68
		Παράρτημα Α	68
		Παράρτημα Β	82
8.	Βιβλιογραφία		85

Κατάλογος Πινάκων

6.1.1	Δεδομένα παραδείγματος	52
6.1.2	Συντελεστές παλινδρόμησης χωρίς ΕΠ	54
6.1.3	Συντελεστές παλινδρόμησης με ΑΣΜ	55
6.1.4	Συντελεστές παλινδρόμησης με ΘΣΜ	56
6.1.5	Συντελεστές παλινδρόμησης με ΚΕΠ	57
6.1.6	Συντελεστές παλινδρόμησης με ΑΣΜ & ΚΕΠ	58
6.1.7	Συντελεστές παλινδρόμησης με ΚΕΠ & ΘΣΜ	60
6.1.8	Συντελεστές παλινδρόμησης με ΑΣΜ & ΘΣΜ	61
6.1.9-6.1.43	Αποτελέσματα μέσω του στατιστικού προγράμματος S-PLUS	68-81
6.2.1	Δεδομένα παραδείγματος	64
6.2.2	Συντελεστές παλινδρόμησης ΜΕΤ & ανθεκτικών εκτιμητών	66
6.2.3-6.2.7	Αποτελέσματα μέσω του στατιστικού προγράμματος S-PLUS	82-84

Κατάλογος Σχημάτων

2.1	Επίδραση έκτοπων παρατηρήσεων στη ΜΕΤ	12
2.2	Τρία βασικά σύνολα έκτοπων παρατηρήσεων	15
5.1	Παραδείγματα συναρτήσεων Μ-εκτιμητών	39
6.1.1	Διάγραμμα διασποράς χωρίς ΕΠ	53
6.1.2	Διάγραμμα διασποράς με ΑΣΜ	54
6.1.3	Διάγραμμα διασποράς με τη χρήση ΜΕΤ και ανθεκτικών εκτιμητών	55
6.1.4	Διάγραμμα διασποράς με ΘΣΜ	56
6.1.5	Διάγραμμα διασποράς με ΚΕΠ	57
6.1.6	Διάγραμμα διασποράς με ΑΣΜ & ΚΕΠ	58
6.1.7	Διάγραμμα διασποράς με τη χρήση ΜΕΤ και ανθεκτικών εκτιμητών	59
6.1.8	Διάγραμμα διασποράς με ΚΕΠ & ΘΣΜ	60
6.1.9	Διάγραμμα διασποράς με ΑΣΜ & ΘΣΜ	61
6.1.10	Διάγραμμα διασποράς με τη χρήση ΜΕΤ και ανθεκτικών εκτιμητών	62
6.2.1	Διάγραμμα σφαλμάτων ελαχίστων τετραγώνων	65
6.2.2	Διάγραμμα σφαλμάτων ελαχίστων διαμέσων τετραγώνων	66

Κατάλογος Συντομογραφιών

ΜΕΤ	Μέθοδος Ελαχίστων Τετραγώνων
ΕΠ	Έκτοπες Παρατηρήσεις
ΑΤ	Απομονωμένες Τιμές
ΣΕ	Συνάρτηση Ευαισθησίας
ΣΚ	Σημείο Κατάρρευσης
ΚΕ	Καμπύλη Ευαισθησίας
ΕΓΣ	Ευαισθησία Γενικού Σφάλματος
ΓΕ	Γενικευμένοι εκτιμητές
ΕΔΤ	Ελάχιστα Διάμεσα Τετράγωνα
ΕΠΤ	Ελάχιστα Περικεκομμένα Τετράγωνα
ΚΕΠ	Κάθετη Έκτοπη Παρατήρηση
ΘΣΜ	Θετικό Σημείο Μόχλευσης
ΑΣΜ	Αρνητικό Σημείο Μόχλευσης

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

1.1 Η αδυναμία της μεθόδου ελαχίστων τετραγώνων

Η ανάλυση παλινδρόμησης είναι ένα σημαντικό στατιστικό εργαλείο που εφαρμόζεται στις περισσότερες επιστήμες. Ανάμεσα στις διάφορες μεθόδους παλινδρόμησης που υπάρχουν, η μέθοδος που τελικά επικράτησε χάρη στην ευκολία υπολογισμού της καθώς και στην παράδοση είναι η μέθοδος των ελαχίστων τετραγώνων. Παρόλα αυτά όμως τα τελευταία χρόνια υπάρχει μία ευρέως διαδεδομένη ενημέρωση σχετικά με την ύπαρξη έκτοπων παρατηρήσεων η οποία μπορεί να οφείλεται σε λάθη πληκτρολόγησης, λάθη τοποθέτησης δεκαδικών ψηφίων, σε σπάνια φαινόμενα όπως σεισμοί, απεργίες ή σε μέλη διαφορετικών πληθυσμών που τυχαία μπήκαν στο δείγμα. Οι έκτοπες παρατηρήσεις υπάρχουν πολύ συχνά σε πραγματικά δεδομένα και σχεδόν πάντα περνάνε απαρατήρητες εξαιτίας του γεγονότος ότι η επεξεργασία των δεδομένων γίνεται μέσω υπολογιστών χωρίς προσεκτική επιθεώρηση.

Παρά τα αναμφισβήτητα χαρακτηριστικά η συγκεκριμένη μέθοδος δεν στερείται αδυναμιών. Εύκολα διαπιστώνεται η ευαισθησία της μεθόδου να αντιμετωπίσει ένα μεγάλο ποσό αλλοιωμένων δεδομένων, κάτι το οποίο οφείλεται στο γεγονός ότι οι έκτοπες παρατηρήσεις καθώς και οι υπόλοιπες αποκλίσεις από το μοντέλο τυπικής γραμμικής παλινδρόμησης εμφανίζονται συχνότατα στα πραγματικά δεδομένα. Στη μέθοδο των ελαχίστων τετραγώνων, ο κίνδυνος των έκτοπων παρατηρήσεων, τόσο στην κατεύθυνση των εξαρτημένων όσο και στην κατεύθυνση των ανεξάρτητων μεταβλητών, είναι ότι μπορούν να επηρεάζουν σημαντικά τον εκτιμητή και να περάσουν απαρατήρητα. Οι έκτοπες παρατηρήσεις μπορούν παντελώς να καταστρέψουν την ανάλυση των ελαχίστων τετραγώνων. Πολύ συχνά οι

απομονωμένες τιμές παραμένουν άγνωστες στον χρήστη κάτι το οποίο οφείλεται στο γεγονός ότι δεν εμφανίζονται πάντα στα διαγράμματα σφαλμάτων με τη μέθοδο των ελαχίστων τετραγώνων.

Προκειμένου να αντιμετωπιστεί ριζικά αυτό το πρόβλημα, αναπτύχθηκαν νέες στατιστικές μέθοδοι οι οποίες δεν επηρεάζονται σε μεγάλο βαθμό από τις έκτοπες παρατηρήσεις και είναι σε θέση να τις αντιμετωπίσουν και να τις ανιχνεύσουν. Αναφερόμαστε στις ανθεκτικές μεθόδους τα αποτελέσματα των οποίων είναι αξιόπιστα, ακόμα και αν μία συγκεκριμένη ποσότητα των δεδομένων είναι αλλοιωμένη. Κάποιοι ερευνητές πιστεύουν ότι οι ανθεκτικές μέθοδοι παλινδρόμησης κρύβουν τις έκτοπες παρατηρήσεις κάτι το οποίο φυσικά δεν είναι σωστό. Στην πραγματικότητα οι έκτοπες τιμές βρίσκονται αρκετά μακριά από την ανθεκτική προσαρμογή και ως εκ τούτου μπορούν να ανιχνευθούν από τα μεγάλα σφάλματα από αυτή. Αντιθέτως τα τυποποιημένα σφάλματα μέσω της μεθόδου των ελαχίστων τετραγώνων μπορεί να μην παρουσιάζουν καθόλου έκτοπες παρατηρήσεις.

Ο όρος ανθεκτικός επινοήθηκε στην στατιστική από τον Box (1953). Αυτόν τον όρο μπορούν να αποδώσουν διαφορετικοί ορισμοί μικρότερης ή μεγαλύτερης αξίας. Γενικότερα όμως οποιαδήποτε αναφορά σε ανθεκτικό στατιστικό εκτιμητή ερμηνεύεται ως εξής: ανεπηρέαστος από μικρές αποκλίσεις από τις ιδανικές υποθέσεις για τις οποίες ο εκτιμητής είναι βέλτιστος. Η λέξη μικρές αποκλίσεις μπορεί να έχει δύο διαφορετικές ερμηνείες εξίσου σημαντικές: είτε μικρές αποκλίσεις από το σύνολο των δεδομένων, είτε μεγάλες αποκλίσεις από ένα μικρό αριθμό δεδομένων. Η δεύτερη ερμηνεία, οδηγεί στη θεωρία των έκτοπων παρατηρήσεων, η οποία μπορεί να προκαλέσει ιδιαίτερα προβλήματα στις στατιστικές διαδικασίες.

1.1.1 Οι υποθέσεις της μεθόδου ελαχίστων τετραγώνων

Υπάρχουν ορισμένες υποθέσεις που πρέπει να ικανοποιούνται έτσι ώστε το μοντέλο παλινδρόμησης ελαχίστων τετραγώνων να είναι έγκυρο. Τα σφάλματα, οι μεγάλες δηλαδή διαφορές που υπάρχουν ανάμεσα στις προβλεπόμενες τιμές και στα πραγματικά δεδομένα, μπορούν να δώσουν λανθασμένη πρόβλεψη. Όταν τα σφάλματα είναι εξαιρετικά μεγάλα, αυτό έχει σαν αποτέλεσμα την αύξηση της διασποράς σφάλματος και των τυπικών σφαλμάτων. Τα διαστήματα εμπιστοσύνης θα έχουν μεγαλύτερο εύρος και κατά συνέπεια η εκτίμηση δεν θα είναι ασυμπτωτικά συνεπής. Μια από τις βασικές υποθέσεις της ανάλυσης παλινδρόμησης είναι η

σταθερότητα της διασποράς σφάλματος σε όλο το μήκος της προβλεφθείσας γραμμής, ένας όρος ο οποίος ονομάζεται ομοσκεδαστικότητα. Η ομοσκεδαστικότητα παρέχει μια μικρή ποσότητα ομοιομορφίας στα διαστήματα εμπιστοσύνης. Ακόμα και αν το μέγεθος του δείγματος είναι μεγάλο, η επίδραση του σφάλματος μπορεί να αυξήσει την τοπική και ακόμα τη γενική διασπορά σφάλματος. Αυτή η αύξηση της διασποράς σφάλματος έχει σαν αποτέλεσμα την μείωση της αποτελεσματικότητας της εκτίμησης.

Η ανθεκτική στατιστική κατά μία μη τεχνική έννοια βασίζεται στο γεγονός ότι πολλές από τις υποθέσεις που γίνονται στη στατιστική (όπως η κανονικότητα, η γραμμικότητα και η ανεξαρτησία) προσεγγίζουν την πραγματικότητα. Ένας λόγος είναι η ύπαρξη έκτοπων παρατηρήσεων, οι οποίες βρίσκονται μακριά από τον όγκο των δεδομένων και είναι επικίνδυνες για τις κλασικές στατιστικές διαδικασίες. Το πρόβλημα των έκτοπων παρατηρήσεων είναι πολύ γνωστό και πιθανότατα τόσο παλιό όσο και η στατιστική. Οποιαδήποτε μέθοδος αντιμετώπισής του όμως, όπως για παράδειγμα ο υποκειμενικός κανόνας απόρριψης ανήκει κατά μία ευρεία έννοια στην ανθεκτική στατιστική. Παρά το γεγονός ότι διακεκριμένοι στατιστικοί γνώριζαν το πρόβλημα της ανθεκτικότητας μόνο τις τελευταίες δεκαετίες έγιναν προσπάθειες για την αντιμετώπισή του.

1.2 Η φύση της ανθεκτικής στατιστικής

Η στατιστική συμπερασματολογία βασίζεται κατά ένα μέρος στο δείγμα από ένα πληθυσμό. Σημαντικό ρόλο παίζουν επίσης οι εκ των προτέρων υποθέσεις για το σύνολο των δεδομένων όπως, π.χ. η κανονικότητα, η γραμμικότητα και η ανεξαρτησία. Η χρήση τους γίνεται με την παραδοχή ότι μικρές αποκλίσεις από τις υποθέσεις θα προκαλέσουν μικρό σφάλμα στα τελικά αποτελέσματα και οι στατιστικές διαδικασίες που είναι βέλτιστες κάτω από το ακριβές μοντέλο, θα είναι κατά προσέγγιση βέλτιστες κάτω από το προσεγγιστικό μοντέλο. Δυστυχώς αυτό δεν ισχύει πάντα. Κάποιες από τις πιο δημοφιλείς στατιστικές διαδικασίες (π.χ. η μέθοδος ελαχίστων τετραγώνων, t-τεστ) είναι υπερβολικά ευαίσθητες σε φαινομενικά μικρές αποκλίσεις από τις υποθέσεις.

Οι αποκλίσεις από τις υποθέσεις μπορεί να οφείλονται στην παρουσία γενικών σφαλμάτων (gross errors), που είναι τα σφάλματα που προκύπτουν από κάποια πηγή

που δρα περιστασιακά, αλλά είναι αρκετά ισχυρά. Συνήθως προέρχονται από λάθη αντιγραφής, εναλλαγής τιμών, απρόσεκτη παρατήρηση, παροδικές επιδράσεις, παρερμηνευμένες ερωτήσεις, ελλιπείς απαντήσεις, παροδικές επιδράσεις. Αποτελούν τη συχνότερη αιτία έκτοπων παρατηρήσεων (outliers). Ίσως, εάν επιδειχθεί ιδιαίτερη προσοχή, είναι δυνατή η προστασία από γενικά σφάλματα και πράγματι, υπάρχουν σύνολα χιλιάδων δεδομένων, χωρίς γενικά σφάλματα, αν και αυτό είναι σπάνιο. Πάντως, η αναγκαία προσοχή δεν είναι πάντα εφικτό να υπάρχει. Η ύπαρξη γενικών σφαλμάτων είναι η πιο επικίνδυνη απόκλιση από τις στατιστικές υποθέσεις, δεδομένου ότι ένα μοναδικό τεραστίου μεγέθους γενικό σφάλμα είναι δυνατό να καταστρέψει τη στατιστική ανάλυση. Τα δεδομένα που προέρχονται από πραγματικές μετρήσεις περιέχουν συνήθως 1 μέχρι 10% γενικά σφάλματα.

Κατά την ανάλυση δεδομένων προκύπτουν ερωτήματα της μορφής: Είναι τα δεδομένα ομογενή ή διαφορετικά τμήματα δεδομένων δίνουν διαφορετικές πληροφορίες; Στη δεύτερη περίπτωση τι υποδεικνύει η πλειοψηφία των δεδομένων; Ποιες μειονότητες δεδομένων συμπεριφέρονται διαφορετικά και πώς; Ποια η επίδραση των διαφορετικών τμημάτων στο τελικό αποτέλεσμα; Ποια δεδομένα είναι κρίσιμης σημασίας για την επιλογή του μοντέλου ή για τα τελικά αποτελέσματα και πρέπει να εξεταστούν με ιδιαίτερη προσοχή; Ποια μπορεί να είναι η επίδραση των γενικών σφαλμάτων στα αποτελέσματα; Πόσα γενικά σφάλματα μπορεί να ανεχτεί το μοντέλο; Ποια μέθοδος δίνει τη μεγαλύτερη ασφάλεια και ποιες μέθοδοι είναι ταυτόχρονα ικανοποιητικά ασφαλείς και αποδοτικές; Πόσο αξιόπιστα είναι τα αποτελέσματα, αν οι υποθέσεις του μοντέλου ισχύουν μόνο προσεγγιστικά;

Όλα αυτά τα ερωτήματα μαζί με τις αποκλίσεις από τις υποθέσεις αποτελούν στοιχεία αιτιολόγησης της ανάπτυξης της Ανθεκτικής Στατιστικής (Robust Statistics).

Ανθεκτική Στατιστική είναι το σύνολο των γνώσεων και τεχνικών που σχετίζονται με τις αποκλίσεις από τις υποθέσεις που γίνονται στη στατιστική. Ανθεκτικότητα σημαίνει μη ευαισθησία σε μικρές αποκλίσεις από τις υποθέσεις. Καθώς όλες οι θεωρίες ανθεκτικότητας λαμβάνονται ως αποκλίσεις από τις υποθέσεις παραμετρικών μοντέλων, η ανθεκτική στατιστική είναι η στατιστική των προσεγγιστικών παραμετρικών μοντέλων.

Η ανθεκτική στατιστική αποτελεί επέκταση της κλασικής παραμετρικής στατιστικής. Απαιτεί τη διατύπωση νέων στατιστικών εννοιών για την περιγραφή της συμπεριφοράς στατιστικών διαδικασιών, όχι μόνο κάτω από ακριβή παραμετρικά μοντέλα, αλλά και σε περιοχές τέτοιων μοντέλων.

Η ανθεκτική στατιστική επιδιώκει:

- να συμπληρώσει όλες τις κλασικές μεθοδολογίες της παραμετρικής στατιστικής, προσθέτοντας την έννοια της ανθεκτικότητας
- να μελετήσει και να περιγράψει τη συμπεριφορά στατιστικών διαδικασιών σε περιοχές παραμετρικών μοντέλων
- να προτείνει ανθεκτικές διαδικασίες καλύτερες από τις ήδη υπάρχουσες για γνωστά προβλήματα χωρίς διαθέσιμες διαδικασίες επεξεργασίας τους
- να οδηγήσει σε τεχνικές βαθύτερης μελέτης των δεδομένων
- να παράγει σύστημα αξιολόγησης και σύγκρισης των στατιστικών διαδικασιών με βάση την ανθεκτικότητά τους

Κάθε ανθεκτική διαδικασία πρέπει:

- να προτείνει το μοντέλο που προσαρμόζεται καλύτερα στον κύριο όγκο των δεδομένων
- να είναι ανθεκτική, με την έννοια ότι μικρές αποκλίσεις από το μοντέλο πρέπει να επηρεάζουν ελάχιστα τα αποτελέσματα
- να έχει ικανοποιητική αποδοτικότητα
- να αναγνωρίζει τις ακραίες παρατηρήσεις ή υποσύνολα δεδομένων με κάποια συστηματική μορφή
- να μην επηρεάζεται από μεγαλύτερες αποκλίσεις.

Για υψηλής ποιότητας δεδομένα, χωρίς γενικά σφάλματα, ή τουλάχιστον, για «επεξεργασμένα» δεδομένα, οι ανθεκτικές μέθοδοι δεν είναι απόλυτα αναγκαίες. Η συνεισφορά τους μένει κρυμμένη λόγω της απουσίας ακραίων παρατηρήσεων. Πάντως, ακόμα και για υψηλής ποιότητας δεδομένα, οι καλές ανθεκτικές μέθοδοι δίνουν αξιοσημείωτη βελτίωση έναντι των κλασικών, όμως το μέγεθος αυτής της βελτίωσης διαφέρει κατά περίπτωση και είναι δευτερεύουσας σημασίας.

Σε καταστάσεις όπου υπάρχουν ακραίες παρατηρήσεις, η ανάγκη ανθεκτικών διαδικασιών είναι προφανής. Πρέπει να σημειωθεί ότι οι ακραίες παρατηρήσεις εμφανίζονται ακόμα και κάτω από ιδανικές συνθήκες.

Η ανθεκτικότητα είναι μόνο ένα στοιχείο μιας καλής ανάλυσης δεδομένων, ενώ άλλα σημαντικά στοιχεία είναι η σωστή επιλογή μοντέλου, η ερμηνεία των ακραίων παρατηρήσεων, η κατανόηση και εξήγηση των αποτελεσμάτων κ.λ.π. Για την πρόληψη αποκλίσεων, κυρίως εξαιτίας ακραίων παρατηρήσεων, είναι απαραίτητες κάποιες ανθεκτικές μέθοδοι. Καλές ανθεκτικές μέθοδοι, όπως έχουν αναπτυχθεί τις

τελευταίες δεκαετίες, προλαμβάνουν απώλειες της τάξης 3-30% ή περισσότερο, ως προς την αποδοτικότητα. Όσο μεγαλώνει η διάσταση και η πολυπλοκότητα του συνόλου των δεδομένων, τόσο πιο απαραίτητη και ασφαλής είναι η χρήση μοντέρνων ανθεκτικών μεθόδων.

1.2.1 Οι στόχοι της ανθεκτικής στατιστικής

Οι βασικότεροι από τους στόχους της ανθεκτικής στατιστικής είναι οι ακόλουθοι:

- να γίνει περιγραφή της δομής που προσαρμόζει καλύτερα τον όγκο των δεδομένων.

Υποθέτουμε ότι έχουμε ένα παραμετρικό μοντέλο και προσπαθούμε όσο γίνεται καλύτερα να εκτιμήσουμε και να ελέγξουμε τις παραμέτρους του μοντέλου, λαμβάνοντας υπόψη ότι η μειονότητα των δεδομένων μπορεί να μην ανήκει στο μοντέλο.

- να γίνει αναγνώριση των σημείων που αποκλίνουν (έκτοπες παρατηρήσεις) και να βρεθούν μέθοδοι αντιμετώπισής τους.

Τα σφάλματα από την ανθεκτική προσαρμογή αυτόματα δείχνουν τις έκτοπες παρατηρήσεις καθώς και την τυχαία μεταβλητότητα των «καλών» δεδομένων πολύ πιο καθαρά από ότι τα σφάλματα των ελαχίστων τετραγώνων, τα οποία επιδρούν αρνητικά στο σύνολο των δεδομένων. Ενώ μέσω μιας προσεκτικής οπτικής εποπτείας των δεδομένων σε ένα μικρό δείγμα, είναι πολύ πιθανή η ανίχνευση έκτοπων παρατηρήσεων, κάτι τέτοιο είναι αδύνατο να συμβεί όταν το δείγμα μας είναι μεγάλο. Ακόμα και μία επιμελής ανάλυση των σφαλμάτων μπορεί να μη δείξει τα σημεία που αποκλίνουν ή ακόμα μπορεί να χρειαστεί περισσότερος χρόνος απ' ότι συνήθως απαιτείται. Τα τυποποιημένα σφάλματα μπορούν να ανεχτούν μόνο το 10% των έκτοπων παρατηρήσεων στην απλή περίπτωση πριν καταρρεύσει ο εκτιμητής.

- να γίνει αναγνώριση των σημείων μόχλευσης και να αντιμετωπιστούν.

Μία παρατήρηση $(x_κ,y_κ)$ ονομάζεται σημείο μόχλευσης όταν το $x_κ$ βρίσκεται μακριά από τον όγκο των παρατηρούμενων x τιμών του δείγματος. Η μόχλευση της i παρατήρησης ορίζεται ως εξής:

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i-1}^{n} (X_i - \bar{X})^2}.$$

Εάν από το δείγμα μας αφαιρέσουμε ένα σημείο μόχλευσης ενώ στην πραγματικότητα είναι μία κανονική παρατήρηση το αποτέλεσμα θα είναι να χάσουμε αρκετή από την αποτελεσματικότητα. Σε περίπτωση όμως που δεν αφαιρέσουμε αυτή την παρατήρηση ενώ είναι γενικό σφάλμα, αυτό θα έχει σαν αποτέλεσμα να καταστραφεί παντελώς η ανάλυσή μας. Ίσως η καλύτερη συμβουλή που μπορεί να δοθεί στην προκειμένη περίπτωση είναι να εφαρμόσουμε δύο φορές την ανθεκτική παλινδρόμηση, μία με το σημείο μόχλευσης και την άλλη χωρίς αυτό και ακολούθως να συγκρίνουμε τα αποτελέσματα. Αν τα αποτελέσματα διαφέρουν μεταξύ τους τότε το μοντέλο που χρησιμοποιείται δεν είναι σωστό.

Οι ανθεκτικοί εκτιμητές θα πρέπει να ικανοποιούν τους ακόλουθους στόχους:

- συνέπειας, ασυμπτωτικής κανονικότητας και υψηλής αποτελεσματικότητας των εκτιμητών εάν το μοντέλο δεν παραβιάζεται
- εύρεση κατάλληλων μεθόδων για τη διαμόρφωση των διαστημάτων εμπιστοσύνης για τις άγνωστες παραμέτρους και για τον έλεγχο υποθέσεων σχετικά με αυτές
- απλότητας της θεωρίας
- ευκολίας υπολογισμού, δοθέντος ενός τυπικού προγράμματος υπολογιστή.

1.2.2 Η σημασία των ανθεκτικών διαδικασιών

Μια ματιά στη βιβλιογραφία μας βοηθάει να γνωρίσουμε τις αντιφατικές αναλύσεις σχετικά με την αναγκαιότητα και τη σημασία των ανθεκτικών μεθόδων. Μερικοί συγγραφείς όπως ο Stigler (1977) πρότεινε κάποιους ανθεκτικούς εκτιμητές όπως η μέση τιμή μετά από ποσοστιαία αποκοπή (mean trimmed), την οποία δεν θεώρησαν καλύτερη από τον αριθμητικό μέσο. Ο Mallows (1979) και ο Rocke et al. (1982), δίνουν παραδείγματα όπου χρησιμοποιούνται ανθεκτικοί εκτιμητές και οι οποίοι θεωρούνται πολύ ανώτεροι των εκτιμητών ελαχίστων τετραγώνων. Για δεδομένα υψηλής ποιότητας, απαλλαγμένα από την ύπαρξη έκτοπων παρατηρήσεων, η εφαρμογή ανθεκτικών μεθόδων δεν είναι απαραίτητη. Κάποιοι ερευνητές μπορεί να πάρουν το ρίσκο, ελπίζοντας ότι δεν υπάρχουν στα δεδομένα απομονωμένες τιμές και χωρίς απολύτως κανένα έλεγχο να εφαρμόσουν τις κλασικές μεθόδους, κάτι το οποίο θεωρείται εξαιρετικά επικίνδυνο. Ακόμα και στην περίπτωση υψηλής ποιότητας δεδομένων, κάποιες καλές ανθεκτικές μέθοδοι μπορούν να δείξουν μία εμφανή βελτίωση σε σύγκριση με τις κλασικές μεθόδους. Μερικές ανθεκτικές μέθοδοι (όπως

η υποκειμενική ή η αντικειμενική απόρριψη των έκτοπων παρατηρήσεων) θεωρούνται αναγκαίες για να αντιμετωπιστεί οποιαδήποτε καταστροφή οφειλόμενη στις απομονωμένες τιμές. Οι ορθές ανθεκτικές μέθοδοι, όπως αναπτύχθηκαν τις τελευταίες δεκαετίες κρίνονται απαραίτητες για την αντιμετώπιση απώλειας αποτελεσματικότητας. Ως εκ τούτου όσο πιο πολύπλοκο είναι το σύνολο των δεδομένων, τόσο λιγότερο κατάλληλες θεωρούνται οι υποκειμενικές μέθοδοι και τόσο περισσότερο σημαντικές θεωρούνται οι ασφαλείς στατιστικές μέθοδοι.

1.3 Οι πρώτες προσεγγίσεις

Οι ανθεκτικές τεχνικές χρονολογούνται από την προϊστορική περίοδο της στατιστικής. Η διερεύνηση των δεδομένων και ο έλεγχος για παρατηρήσεις με κάποια ιδιαιτερότητα είναι ένα πρώτο βήμα προς την ανθεκτικότητα, ενώ η εξαίρεση τιμών με υψηλή απόκλιση αποτελεί μια άτυπη ανθεκτική διαδικασία. Η διάμεσος είναι ανθεκτικό μέγεθος και η μετάβαση από τη μέση τιμή στη διάμεσο, για δεδομένα με μεγάλες ουρές, είναι η ανθεκτική μέθοδος. Η επικρατούσα τιμή είναι ανθεκτική, όταν η πιο πιθανή τιμή είναι φανερά πιο πιθανή από τη δεύτερη. Όταν οι Έλληνες πολιορκητές της αρχαιότητας μετρούσαν τα στρώματα πλίνθων του τείχους της πολιορκημένης πόλης και μετά όριζαν το απαραίτητο μήκος για τις ανεμόσκαλές τους με βάση την επικρατούσα τιμή χρησιμοποιούσαν ουσιαστικά μια ανθεκτική μέθοδο.

Η συζήτηση για την καταλληλότητα της απόρριψης των ακραίων παρατηρήσεων αρχίζει από τον Daniel Bernoulli, το 1777 και τους Bessel & Baeyer, το 1838, ενώ ο Boscovich, το 1755 είχε ήδη χρησιμοποιήσει απόρριψη ακραίων παρατηρήσεων, κάτι που, κατά τον Bernoulli, ήταν κοινή πρακτική για τους αστρονόμους της εποχής. Επίσης η μέση τιμή μετά από ποσοστιαία αποκοπή χρησιμοποιείται από παλιά (“Anonymous”, το 1821 και Mendeleev, το 1895). Εκτός του χώρου των επιστημών, μία μέση τιμή μετά από ασύμμετρη ποσοστιαία αποκοπή (αγνοούνται μόνο οι καλύτερες ή οι χειρότερες τιμές) χρησιμοποιείται στην αξιολόγηση αθλητικών επιδόσεων, έτσι ώστε να υπάρχει προστασία από μεροληπτικούς κριτές. Ο πρώτος τυπικός κανόνας απόρριψης δόθηκε από τους Peirce, το 1852 και Chauvenet, το 1863, στη συνέχεια από τους Stone, το 1868, Wright, το 1884, Irwin, το 1925, Student, το 1927, Thompson, το 1935, Pearson & Chaudra Sekar, το 1936 και πολλούς άλλους.

Ο Student, το 1927, πρότεινε επαναληπτική δειγματοληψία στην περίπτωση ακραίων παρατηρήσεων, συνδυασμένη πιθανόν με απόρριψη, μία εκλεπτυσμένη τεχνική, που δεν έτυχε ιδιαίτερης προσοχής. Παρόμοιες τεχνικές απαιτούν να υπάρχει δυνατότητα για επιπλέον παρατηρήσεις. Αλλά αν αυτό είναι εφικτό, είναι αξιοσημείωτα καλύτερες για μικρά δείγματα από τους συνηθισμένους κανόνες απόρριψης. Στην απλή της μορφή συνίσταται στο να διαπιστωθεί αν δύο παρατηρήσεις είναι μακριά και αν είναι να υπάρξει μια τρίτη παρατήρηση, οπότε τότε η στατιστική συνάρτηση που προκύπτει είναι η μέση τιμή των δύο κοντινότερων παρατηρήσεων.

Παράλληλα παρουσιάζονται μείξεις μοντέλων και εκτιμητών που υποβιβάζουν μερικώς τις υπερβολικά μακρινές παρατηρήσεις (Glaisher, το 1872-73, E.J. Stone, το 1873, Edgeworth, το 1883, Newcomb, το 1886, Jeffreys, το 1932 και το 1939). Οι προσπάθειες αυτές αντιμετώπισης των ακραίων παρατηρήσεων, κυρίως αίροντας την επικινδυνότητά τους παρά αγνοώντας ή απομονώνοντάς τες, είναι πολύ κοντά στο πνεύμα της μοντέρνας ανθεκτικής θεωρίας.

Με τον κεντρικό ρόλο που κερδίζει η Κανονική κατανομή ως μοντέλο κατανομών τον $19^{ο}$ αιώνα, το σύστημα καμπυλών Pearson μπορεί να θεωρηθεί σχεδιασμός για την διευθέτηση ενός μεγάλου τμήματος των αποκλίσεων από την κανονικότητα στα πραγματικά δεδομένα. Αυτό ισχύει και για την θεωρία Bayes. Παράλληλα δίνεται προσοχή στην παραβίαση της υπόθεσης ανεξαρτησίας.

Ο E.S. Pearson, το 1931, ανακαλύπτει δραστική έλλειψη ανθεκτικότητας τους ακριβείς ελέγχους του Fisher για διασπορές σε μικρά δείγματα. Η μελέτη του μπορεί να θεωρηθεί ως η αρχή συστηματικής έρευνας για την ανθεκτικότητα των ελέγχων υποθέσεων. Κατά τη διάρκεια αυτής της περιόδου οι μη παραμετρικές μέθοδοι γίνονται δημοφιλείς, ιδιαίτερα οι έλεγχοι βαθμίδων (rank tests). Οι τυχαιοποιημένοι έλεγχοι χρησιμοποιούνται κυρίως σε θεωρητικό επίπεδο. Η μελέτη ανθεκτικότητας των ελέγχων αφορούσε σχεδόν αποκλειστικά το επίπεδο σημαντικότητας (robustness of validity), ενώ η ισχύς (robustness of efficiency) αγνοήθηκε. Τα προβλήματα ήταν ήδη αρκετά δύσκολα και χωρίς ισχύ. Πάντως εύκολα διαπιστώνεται ότι οι κλασικοί τυχαιοποιημένοι έλεγχοι, κατά κανόνα, έχουν μικρή ισχύ όταν υπάρχουν ακραίες παρατηρήσεις. Επίσης κάποιοι έλεγχοι βαθμίδων μπορεί να έχουν μεγάλη απώλεια ισχύος, αν προστεθεί αλλοίωση (contamination).

Η συστηματική ενασχόληση με το πρόβλημα της ανθεκτικής εκτίμησης άρχισε αργότερα, επειδή οι έλεγχοι υποθέσεων ήταν πιο δημοφιλείς λόγω των κομψών

μαθηματικών ιδιοτήτων τους, ενώ τα ευρήματα του E.S Pearson προκάλεσαν μεγάλο ενδιαφέρον για τους ελέγχους, χωρίς παράλληλα να υπάρχει εμφανής ανάγκη εκτίμησης. Ο Tukey, το 1960, ανέδειξε τη δραματική αποτυχία της μέσης τιμής στην επίτευξη ανθεκτικότητας και ερεύνησε ενδιαφέρουσες εναλλακτικές λύσεις. Έτσι η ανθεκτική εκτιμητική ανάγεται σε αυτόνομο πεδίο έρευνας και αίρεται η απομόνωσή της. Ακολούθησαν οι Huber & Hampel, που στράφηκαν προς τη θεμελίωση εύχρηστης, ρεαλιστικής και περιεκτικής θεωρίας ανθεκτικότητας.

Πάνω στη θεμελίωση, ανάπτυξη και επιδιώξεις της ανθεκτικής στατιστικής, αναφέρονται εκτεταμένα οι Huber (1981) και Hampel et al (1986), Huber (1972) και Stigler (1973).

Κ Ε Φ Α Λ Α Ι Ο 2

Απομονωμένες τιμές

2.1 Ορισμός απομονωμένων τιμών

Οι Barnett & Lewis (1979) ορίζουν ως απομονωμένη τιμή μία παρατήρηση η οποία φαίνεται να αποκλίνει από το σύνολο των δεδομένων. Ο Kraskel et al. (1982), ο Hampel et al. (1986), ο Rousseeuw & Leory (1987) δεν δίνουν ένα συγκεκριμένο ορισμό της έκτοπης παρατήρησης αλλά κατατάσσουν τις έκτοπες παρατηρήσεις σε διαφορετικές κατηγορίες.

Ένας εναλλακτικός ορισμός που έχει δοθεί από τον F.J. Anscombe (1960) για την απομονωμένη τιμή είναι ο ακόλουθος. Απομονωμένη τιμή μεταξύ των υπολοίπων είναι εκείνη που κατά απόλυτη τιμή είναι αρκετά μεγαλύτερη από τις υπόλοιπες και ίσως βρίσκεται σε απόσταση τριών ή τεσσάρων τυπικών αποκλίσεων από τη μέση τιμή των υπολοίπων. Η απόφαση σχετικά με το αν η παρατήρηση είναι έκτοπη εξαρτάται από την υποκειμενική κρίση του ερευνητή. Η απομονωμένη τιμή αποτελεί ιδιομορφία και υποδεικνύει ένα σημείο των δεδομένων που δεν είναι καθόλου αντιπροσωπευτικό όπως τα άλλα δεδομένα και θα πρέπει να εξεταστεί ιδιαίτερα προσεκτικά για να δούμε εάν η αιτία της ιδιομορφίας του μπορεί να προσδιοριστεί.

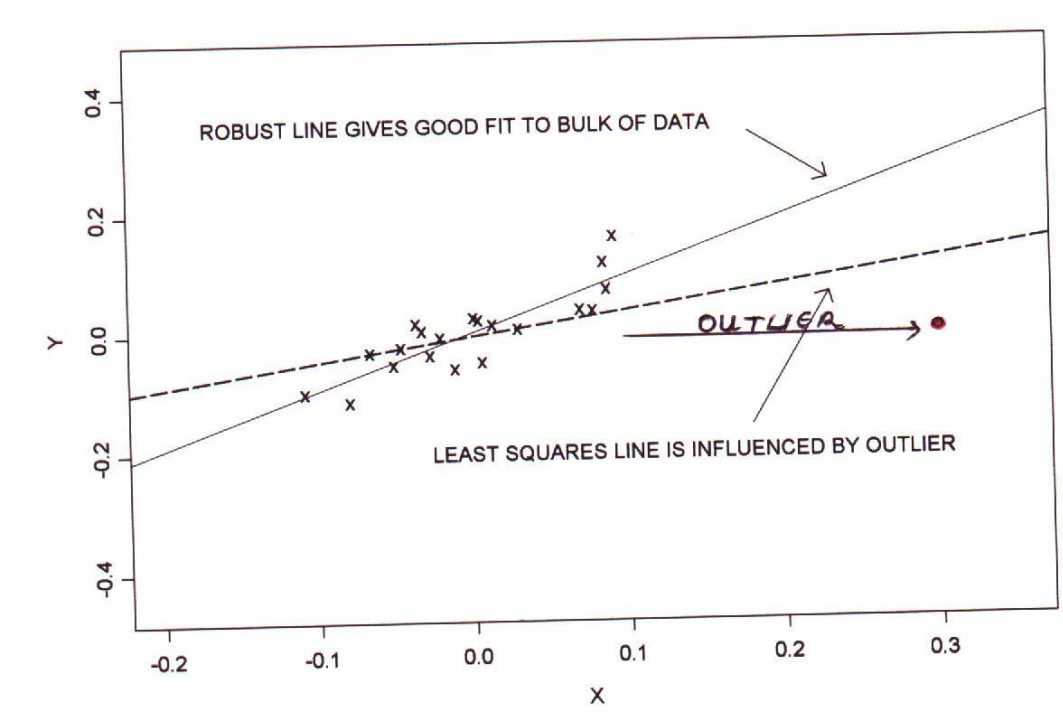
Έχουν προταθεί κανόνες για την απόρριψη των απομονωμένων τιμών (δηλαδή για να αποφασίσουμε να εξαιρέσουμε την (τις) αντίστοιχη(ες) παρατήρηση(εις) από τα δεδομένα, μετά την εξαίρεση των οποίων πρέπει να αναλύσουμε ξανά τα δεδομένα χωρίς τις παρατηρήσεις). Η αυτόματη απόρριψη απομονωμένων τιμών δεν είναι πάντοτε μία σοφή διαδικασία. Μερικές φορές οι απομονωμένες τιμές περιλαμβάνουν πληροφορίες που δεν υπάρχουν στα άλλα δεδομένα διότι προκύπτουν από ένα ασυνήθιστο συνδυασμό περιστάσεων που έχουν ζωτικό ενδιαφέρον και απαιτείται συνεπώς περαιτέρω διερεύνηση. Ως ένα γενικό κανόνα θα λέμε ότι οι απομονωμένες

τιμές θα πρέπει να απορρίπτονται μόνο αν αποτελούν λάθη καταγραφής των δεδομένων ή της λειτουργίας της μηχανής στην οποία ή με την οποία γίνεται η μέτρηση. Διαφορετικά θα πρέπει να διερευνώνται προσεκτικά.

Από το ακόλουθο διάγραμμα εύκολα γίνεται αντιληπτή η αδυναμία της μεθόδου ελαχίστων τετραγώνων να αντιμετωπίσει μία και μόνο έκτοπη 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 μία απομονωμένη τιμή.

Διάγραμμα 2.1

Επίδραση έκτοπων παρατηρήσεων στη ΜΕΤ



2.1.1 Ταξινόμηση έκτοπων παρατηρήσεων

Οι έκτοπες παρατηρήσεις μπορούν να ταξινομηθούν σε δύο κατηγορίες:

α) γενικά σφάλματα

β) έκτοπες παρατηρήσεις ή απομονωμένες τιμές που οφείλονται σε αποτυχημένο μοντέλο.

Πιο συγκεκριμένα :

α) Τα γενικά σφάλματα οφείλονται σε λάθη καταχώρησης, όπως λάθη πληκτρολόγησης, λάθη αντιγραφής, τεχνικές δυσκολίες με τον εξοπλισμό,

ελλιπείς απαντήσεις ερωτηματολογίου, παρερμηνευμένες ερωτήσεις. Οι παρατηρήσεις που είναι γενικά σφάλματα μπορούν να μην ληφθούν υπόψη. Είναι γεγονός ότι η ύπαρξη έκτοπων παρατηρήσεων στα εμπειρικά δεδομένα είναι ο κανόνας και όχι η εξαίρεση. Αυτό έχει σαν αποτέλεσμα ένας μικρός αριθμός έκτοπων παρατηρήσεων να μπορεί να προκαλέσει μεγάλες δυσκολίες στον εκτιμητή ελαχίστων τετραγώνων. Τα γενικά σφάλματα, είναι η πιο συνηθισμένη αιτία για την εμφάνιση απομονωμένων τιμών. Συνήθως, οι έκτοπες τιμές οφείλονται σε μία αυθεντική κατανομή απώλειας (long-tailed distribution). Με κάποια κατάλληλη μέριμνα τα γενικά σφάλματα μπορούν να αντιμετωπιστούν. Είναι όμως λογικό ότι η απαραίτητη μέριμνα δεν είναι πάντα εφικτή. Τα απομακρυσμένα γενικά σφάλματα είναι μια από τις πιο επικίνδυνες αποκλίσεις των συνήθων στατιστικών υποθέσεων. Ο αριθμός των απομακρυσμένων γενικών σφαλμάτων καθώς και των έκτοπων παρατηρήσεων, την ύπαρξη των οποίων οι στατιστικοί παρατηρούν στο δείγμα, συχνά μειώνεται σημαντικά καθώς οι επιστήμονες επεξεργάζονται τα δεδομένα χωρίς να συμβουλευτούν το στατιστικό. Κάθε φορά που υπάρχει στα δεδομένα μία έκτοπη παρατήρηση, η εφαρμογή ανθεκτικών μεθόδων κρίνεται απαραίτητη.

β) Η δεύτερη κύρια αιτία για την ύπαρξη έκτοπων παρατηρήσεων είναι τα ίδια στατιστικά/οικονομετρικά μοντέλα. Η χρήση των ψευδοτυχαίων μεταβλητών στο εμπειρικό οικονομετρικό μοντέλο είναι μια συνήθης πρακτική. Εισάγοντας μια ψευδομεταβλητή που ισούται με τη μονάδα, για μία μεμονωμένη παρατήρηση ισοδυναμεί με το να διώξουμε εντελώς την παρατήρηση από το δείγμα. Το καλύτερο που μπορεί να κάνει κάποιος σε αυτήν την περίπτωση είναι να φτιάξει ένα μοντέλο που περιγράφει την πλειοψηφία των δεδομένων. Αυτός ακριβώς είναι και ο στόχος της ανθεκτικής στατιστικής διαδικασίας.

2.1.2 Θέση και επίδραση των απομονωμένων τιμών

Οι απομονωμένες τιμές μπορούν να περιγραφούν λαμβάνοντας υπόψη τόσο τη θέση όσο και την επίδρασή τους. Αναφορικά με τη θέση τους οι έκτοπες παρατηρήσεις μπορούν να παρατηρηθούν είτε στην κατεύθυνση του άξονα x είτε στην κατεύθυνση του άξονα y, είτε και στις δύο κατευθύνσεις ταυτόχρονα. Μία μοναδική λάθος τιμή στο y μπορεί να οδηγήσει σε εντελώς εσφαλμένα αποτελέσματα.

Αυτό το ζήτημα είναι γνωστό ως απομονωμένη τιμή στην κατεύθυνση του y. Οι τιμές του y, συχνά θεωρούνται ως παρατηρήσεις και οι τιμές του x σαν σταθεροί αριθμοί. Όμως συχνά οι ανεξάρτητες μεταβλητές, οι οποίες ονομάζονται και επεξηγηματικές, είναι παρατηρημένες ποσότητες που υπόκεινται σε τυχαία μεταβλητότητα. Επίσης οι έκτοπες παρατηρήσεις μπορούν να ασκούν μεγάλη επίδραση στην εκτίμηση των συντελεστών της εξίσωσης παλινδρόμησης. Μία έκτοπη παρατήρηση στην κατεύθυνση του άξονα y μπορεί να ασκήσει μια μικρή επίδραση στην εκτίμηση συντελεστών παλινδρόμησης, ενώ μία έκτοπη παρατήρηση στην κατεύθυνση του άξονα x ασκεί μεγαλύτερη επίδραση.

Οι απομονωμένες τιμές μπορούν να κατηγοριοποιηθούν ανάλογα με την επίδρασή τους. Μία παρατήρηση μπορεί να θεωρηθεί ως σημείο μόχλευσης όταν βρίσκεται μακριά από τα υπόλοιπα δεδομένα. Η μόχλευση δεν λαμβάνει υπόψη την κατεύθυνση της απόστασης από τα εναπομείναντα δεδομένα. Το διάγραμμα διασποράς θα μπορούσε λογικά να χρησιμοποιηθεί και να απεικονιστούν οι τιμές των x και y μέσω της γραμμής καλύτερης προσαρμογής. Αυτή η γραμμή σχηματίζεται χρησιμοποιώντας τις προβλεπόμενες τιμές $\hat{y}$ του y. Σε περίπτωση που η έκτοπη παρατήρηση είναι μακριά από την γραμμή καλύτερης προσαρμογής, αυτό μπορεί να έχει σαν αποτέλεσμα την μετακίνηση της γραμμής προς την κατεύθυνση της απομονωμένης τιμής. Εάν όμως η τιμή αυτή πέσει πάνω στην γραμμή καλύτερης προσαρμογής αλλά μακριά από το σύνολο των δεδομένων, τότε αυτή θεωρείται ως ένα καλό σημείο μόχλευσης, δεδομένου ότι βοηθάει στη μείωση του τυπικού σφάλματος.

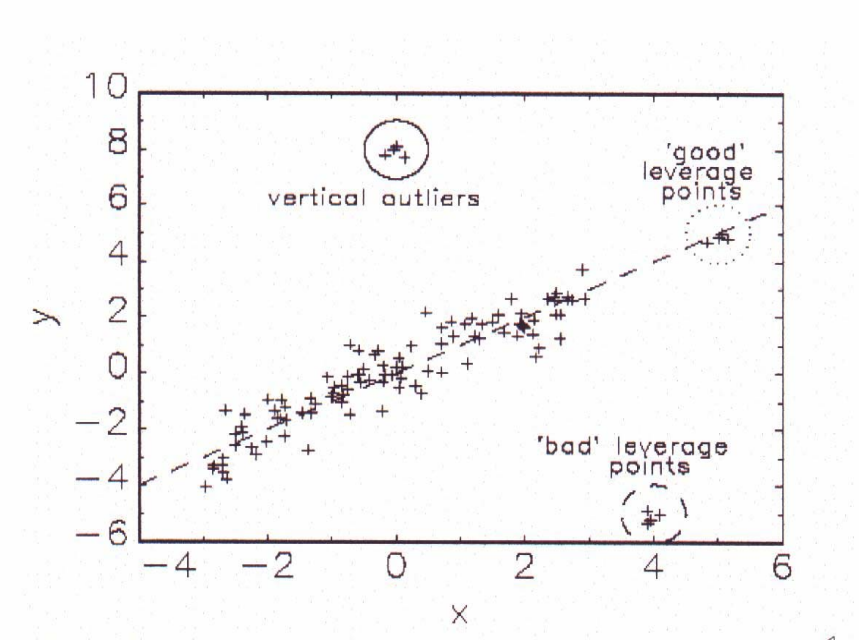
Υπάρχουν γραφικές μέθοδοι για τον εντοπισμό των έκτοπων παρατηρήσεων που βασίζονται στον σχεδιασμό των σφαλμάτων. Μερικές από τις γραφικές παραστάσεις περιλαμβάνουν το ιστόγραμμα, τις προσαρμοσμένες τιμές των σφαλμάτων καθώς και τα διαγράμματα μερικών σφαλμάτων που χρησιμοποιούνται στην πολυμεταβλητή περίπτωση. Οι συγκεκριμένες μέθοδοι είναι χρήσιμες για τον εντοπισμό έκτοπων παρατηρήσεων, αλλά όμως μπορούν να δημιουργήσουν πρόβλημα αν οι απομονωμένες τιμές είναι υψηλής μόχλευσης προς την κατεύθυνση του άξονα x. Αυτές οι τιμές μπορούν να τραβήξουν προς το μέρος τους την γραμμή καλύτερης προσαρμογής, δημιουργώντας μικρά σφάλματα. Συνεπώς οι μέθοδοι που βασίζονται στα σφάλματα μπορεί να αποτύχουν να αναγνωρίσουν αυτά τα δεδομένα σαν έκτοπες τιμές. Επιπρόσθετα, εάν υπάρχουν πολλές ανεξάρτητες μεταβλητές ο σχεδιασμός

κάθε μιας ξεχωριστά ανεξάρτητης μεταβλητής ως προς την τιμή του σφάλματος μπορεί να είναι χρονοβόρος και δύσκολος να πραγματοποιηθεί. Οι ανθεκτικές μέθοδοι παλινδρόμησης δίνουν στον ερευνητή τη δυνατότητα να διαχειρίζεται τις έκτοπες παρατηρήσεις και όχι να τις αγνοεί ή να τις διαγράφει. Επιπρόσθετα, επιτρέπει την αναγνώριση των έκτοπων τιμών και προσαρμόζει την ύπαρξή τους με βάση την ανάλυση παλινδρόμησης.

Παρακάτω παρατίθεται ένα διάγραμμα στο οποίο απεικονίζονται τα τρία βασικά σύνολα έκτοπων παρατηρήσεων.

Διάγραμμα 2.2

Τρία βασικά σύνολα έκτοπων παρατηρήσεων



- Το πρώτο σύνολο των έκτοπων παρατηρήσεων αποτελείται από τις κάθετες έκτοπες παρατηρήσεις. Οι τιμές των x βρίσκονται μέσα στο εύρος του όγκου των δεδομένων. Βέβαια οι παρατηρήσεις αυτές βρίσκονται εμφανώς μακριά από την γραμμική σχέση που καθορίζεται από το σύνολο των δεδομένων.
- Το δεύτερο σύνολο των έκτοπων παρατηρήσεων αποτελείται από τα θετικά σημεία μόχλευσης τα οποία ικανοποιούν τη γραμμική σχέση που ορίζεται από τον όγκο των δεδομένων, αλλά οι τιμές των x βρίσκονται έξω από το συνηθισμένο εύρος. Τέτοιες παρατηρήσεις τείνουν να αυξήσουν την αποτελεσματικότητα του εκτιμητή ελαχίστων τετραγώνων.

- Το τρίτο σύνολο δεδομένων αποτελείται από τα αρνητικά σημεία μόχλευσης. Τα αρνητικά σημεία μόχλευσης έχουν αποκλίνουσες x τιμές που δεν προσαρμόζονται στη γραμμική σχέση όπως αυτή ορίζεται από τα δεδομένα.

Συνδυάζοντας τις γνώσεις σχετικά με τις έκτοπες παρατηρήσεις και τα σημεία μόχλευσης, καταλήγουμε στο συμπέρασμα ότι υπάρχουν τέσσερις τύποι παρατηρήσεων που μπορούν να παρουσιαστούν στα δεδομένα παλινδρόμησης:

- κανονικές παρατηρήσεις με εσωτερικά x και καλά προσαρμοσμένες τιμές y
- κάθετες έκτοπες παρατηρήσεις με εσωτερικά x και όχι καλά προσαρμοσμένες τιμές y
- θετικά σημεία μόχλευσης με έκτοπες παρατηρήσεις στην κατεύθυνση των x και καλά προσαρμοσμένες τιμές y
- αρνητικά σημεία μόχλευσης με έκτοπες παρατηρήσεις στην κατεύθυνση των x και όχι καλά προσαρμοσμένες τιμές y.

2.2 Απόρριψη έκτοπων παρατηρήσεων

Υποθέτοντας ότι 9 εκ των 10 μετρήσεων βρίσκονται κοντά η μία στην άλλη τότε μία συνήθης διαδικασία στην εφαρμογή της στατιστικής είναι η απόρριψη της έκτοπης παρατήρησης, κάτι το οποίο σημαίνει ότι πρέπει η απομονωμένη τιμή να διωχθεί από το δείγμα και ακολούθως να συνεχιστεί η διαδικασία σαν να ήταν οι 9 μετρήσεις μοναδικές του δείγματος. Κάποιες φορές οι έκτοπες τιμές ερμηνεύονται χωριστά και κάποιες άλλες φορές δεν λαμβάνονται καθόλου υπόψη. Η απόφαση σχετικά με το τι σημαίνει «μακριά» από το δείγμα μπορεί να γίνει υποκειμενικά, δηλαδή είτε κοιτάζοντας απλά τα δεδομένα και παίρνοντας κάποια υποκειμενική απόφαση, είτε εφαρμόζοντας κάποιο επίσημο αντικειμενικό κανόνα σχετικά με την απόρριψη έκτοπων παρατηρήσεων το οποίο είναι ένα είδος στατιστικού τεστ. Υπάρχουν επίσης και κάποιες άλλες πιθανότητες για την αντιμετώπιση των έκτοπων παρατηρήσεων. Οι απομονωμένες τιμές συχνά μπορούν να ελεγχθούν παίρνοντας τα γνήσια αρχεία, από τα οποία συχνά προκύπτει ότι οι απομονωμένες τιμές συχνά οφείλονται σε γενικά σφάλματα, τα οποία μπορούν εύκολα να διορθωθούν. Μία μόνο απομονωμένη τιμή μπορεί να είναι επιβλαβής στην στατιστική διαδικασία που ακολουθείται (όπως ο αριθμητικός μέσος, η τυπική απόκλιση, η μέθοδος των ελαχίστων τετραγώνων). Οπωσδήποτε, το επιθυμητό είναι η αναγνώριση έκτοπων

παρατηρήσεων και η απομάκρυνσή τους από το δείγμα με μία καλή δικαιολογία. Παρόλο αυτά όμως κάτι τέτοιο δεν είναι πάντοτε δυνατό, με αποτέλεσμα ο κίνδυνος που προέρχεται από την ύπαρξη έκτοπων παρατηρήσεων να παραμένει και να είναι μεγαλύτερος από τον κίνδυνο απώλειας της αποτελεσματικότητας που προέρχεται από την απόρριψη σε περίπτωση μιας «καλής» παρατήρησης. Επιπλέον, θα ήταν εξαιρετικά απίθανο να έχουμε μία «καλή» παρατήρηση από το μοντέλο κατανομής σε μεγάλη απόσταση. Ως εκ τούτου, όταν μία απομονωμένη τιμή είναι μακριά από τον όγκο των δεδομένων θεωρείται ότι είναι απίθανο να παραμείνει στο δείγμα με αποτέλεσμα να απορρίπτεται από αυτό.

2.3 Συμπεριφορά των κανόνων απόρριψης σχετικά με την ανθεκτική εκτίμηση

- Είναι γεγονός ότι οποιοσδήποτε τρόπος αντιμετώπισης των έκτοπων παρατηρήσεων μπορεί να οδηγήσει στην αποφυγή καταστροφής της μελέτης μας. Τέτοιες μέθοδοι περιλαμβάνουν υποκειμενικούς κανόνες απόρριψης με υψηλό σημείο κατάρρευσης, καθώς και αντικειμενικούς κανόνες ύστερα από οπτική παρακολούθηση των δεδομένων. Ένα μη ανθεκτικό πρόγραμμα υπολογιστή, χωρίς την ύπαρξη μιας προσεκτικής ανάλυσης των σφαλμάτων κρίνεται εντελώς ακατάλληλο.

- Για μεγαλύτερα επίπεδα αποτελεσματικότητας: Οι περισσότερες μέθοδοι χάνουν σε κάποιες ρεαλιστικές περιπτώσεις τουλάχιστον 5-20% της αποτελεσματικότητάς τους. Αυτές οι μέθοδοι περιλαμβάνουν όλους τους υποκειμενικούς καθώς και τους αντικειμενικούς κανόνες απόρριψης.

- Ο ειδικός δεν θα πρέπει μόνο να ανιχνεύσει τις έκτοπες τιμές, αλλά κυρίως θα πρέπει να τις ερμηνεύσει. Οφείλει να ενσωματώνει το μη στατιστικό υπόβαθρο προκειμένου να ερμηνεύσει σωστά τις παρατηρήσεις. Η καλή γνώση είναι τουλάχιστον το ίδιο σημαντική με τους στατιστικούς ισχυρισμούς. Υπάρχουν επίσης γενικές στατιστικές πληροφορίες συνηθισμένων τύπων γενικών σφαλμάτων και άλλων αποκλίσεων από το παραμετρικό μοντέλο το οποίο θα πρέπει να ενσωματωθεί στην ανάλυση των δεδομένων. Η τυπική αναγνώριση ύποπτων παρατηρήσεων μπορεί να

είναι πιο γρήγορη και εύκολη καθώς επίσης και η μοναδική πιθανότητα σε πολύπλοκους σχεδιασμούς.

- Η αναγνώριση έκτοπων παρατηρήσεων μπορεί να γίνεται με μεγαλύτερη ασφάλεια και αποτελεσματικότητα κοιτάζοντας τα σφάλματα της ανθεκτικής προσαρμογής.

- Οι κανόνες απόρριψης και η μεταγενέστερη εκτίμησή τους δεν είναι τίποτα άλλο από τους ειδικούς ανθεκτικούς εκτιμητές. Η εκτίμηση ύστερα από υποκειμενική απόρριψη μπορεί να θεωρηθεί σαν μία υποκειμενική ανθεκτική διαδικασία.

2.4 Λόγοι για τους οποίους ο εντοπισμός των έκτοπων παρατηρήσεων και η απομάκρυνσή τους από το δείγμα δεν θεωρείται αρκετός.

Οι βασικότεροι λόγοι για τους οποίους ο εντοπισμός των ΕΠ και η απομάκρυνσή τους από το δείγμα δεν θεωρείται αρκετός είναι οι ακόλουθοι:

1. Οι χρήστες, ακόμα και ειδικοί στατιστικοί δεν ελέγχουν πάντα τα δεδομένα.
2. Η απόφαση σχετικά με το αν θα πρέπει να κρατήσουμε ή να απορρίψουμε μία παρατήρηση είναι χάσιμο χρόνου. Τις περισσότερες φορές είναι προτιμότερο να αντιμετωπίζουμε τις ύποπτες παρατηρήσεις μέσω κατάλληλων στατιστικών μεθόδων από το να τις απορρίπτουμε.
3. Μπορεί να είναι πάρα πολύ δύσκολος ή και αδύνατος ο εντοπισμός των έκτοπων παρατηρήσεων στην πολυμεταβλητή ανάλυση.
4. Η απόρριψη των απομονωμένων τιμών επηρεάζει την θεωρία της κατανομής.

Κ Ε Φ Α Λ Α Ι Ο 3

Δύο βασικές προσεγγίσεις αντιμετώπισης έκτοπων παρατηρήσεων

3.1 Εισαγωγή

Οι δύο βασικές προσεγγίσεις που επικράτησαν προκειμένου να οδηγηθούμε στην οριστική αντιμετώπιση των έκτοπων παρατηρήσεων είναι οι ακόλουθες:

- Διαγνωστική παλινδρόμηση, όπου συγκεκριμένες ποσότητες του δείγματος υπολογίζονται με σκοπό να εντοπιστούν τα σημεία επιρροής, από τα οποία οι απομονωμένες τιμές μπορούν να διορθωθούν ή να απομακρυνθούν και ακολούθως να εφαρμοστεί η ανάλυση ελαχίστων τετραγώνων στις εναπομείνασες περιπτώσεις.
- Ανθεκτική παλινδρόμηση, η οποία προσπαθεί να δημιουργήσει εκτιμητές που δεν επηρεάζονται από τις έκτοπες παρατηρήσεις. Πριν από οποιαδήποτε εφαρμογή ανθεκτικών μεθόδων, θα πρέπει οπωσδήποτε να εξεταστεί αν οι αποκλίσεις οφείλονται σε αποτυχημένο μοντέλο, κάτι το οποίο μπορεί να αντιμετωπιστεί προσθέτοντας στο μοντέλο μας περισσότερους όρους.

Παρά το γεγονός ότι η τόσο η διαγνωστική όσο και η ανθεκτική παλινδρόμηση έχουν τον ίδιο στόχο, η διαδικασία πραγματοποιείται με διαφορετικό τρόπο. Σε μερικές εφαρμογές και οι δύο προσεγγίσεις καταλήγουν στο ίδιο συμπέρασμα, με αποτέλεσμα οποιαδήποτε διαφορά να είναι μόνο υποκειμενική.

3.2 Οι εκτιμητές και οι βασικές ιδιότητές τους

Είναι ευρέως γνωστό ότι στον τομέα της στατιστικής μπορούν να χρησιμοποιηθούν παραμετρικά και μη παραμετρικά τεστ. Τα μη παραμετρικά τεστ είναι ελεύθερης κατανομής δεδομένου ότι γίνονται λιγότερες υποθέσεις σχετικά με τον πληθυσμό. Στην πραγματικότητα, στην παραμετρική στατιστική ο ερευνητής δεν συμπεραίνει ότι η δειγματική στατιστική συνάρτηση είναι ένας εκτιμητής παραμέτρου πληθυσμού. Σε αντίθεση, η παραμετρική στατιστική χρησιμοποιεί τη δειγματική στατιστική συνάρτηση σαν μία εκτίμηση της πληθυσμιακής παραμέτρου με αποτέλεσμα ο ερευνητής να βγάζει κάποιο συμπέρασμα σχετικά με την τιμή της πληθυσμιακής παραμέτρου. Ως εκ τούτου, ένας εκτιμητής είναι η τιμή της δειγματικής στατιστικής συνάρτησης που δίνει χρήσιμες πληροφορίες σχετικά με την πληθυσμιακή παράμετρο.

Όταν εκτελούνται παραμετρικά στατιστικά τεστ, ο ερευνητής θα πρέπει να ενδιαφέρεται για τις ιδιότητες των εκτιμητών. Οι ιδιότητες ενός εκτιμητή είναι πολύ σημαντικές δεδομένου ότι μας ενδιαφέρει η δειγματική στατιστική συνάρτηση να είναι ένας ακριβής και σταθερός εκτιμητής. Οι ιδιότητες των εκτιμητών είναι η αμεροληψία, η συνέπεια, η αποτελεσματικότητα και η επάρκεια. Συγκεκριμένα:

- Ένας εκτιμητής είναι αμερόληπτος (unbiassed) όταν ο μέσος της δειγματικής κατανομής της στατιστικής συνάρτησης ισούται με την πληθυσμιακή παράμετρο που έχει εκτιμηθεί.

- Ένας εκτιμητής είναι συνεπής (consistent) εάν η δειγματική στατιστική συνάρτηση πλησιάζει ολοένα και περισσότερο την πληθυσμιακή παράμετρο καθώς το μέγεθος του δείγματος αυξάνει.

- Ένας εκτιμητής είναι αποτελεσματικός (efficient) εάν δεν διαφέρει πολύ από δείγμα σε δείγμα (σφάλμα διασποράς ή δειγματικό σφάλμα). Κατά συνέπεια αναφέρεται στην ακρίβεια της εκτίμησης. Η διασπορά σφάλματος της στατιστικής συνάρτησης είναι η διασπορά της δειγματικής κατανομής της στατιστικής συνάρτησης.

- Ένας εκτιμητής είναι επαρκής (sufficient) όταν η δειγματική στατιστική συνάρτηση είναι ο καλύτερος εκτιμητής της πληθυσμιακής παραμέτρου.

3.3 Το κλασικό πρόβλημα παλινδρόμησης

Το κλασικό πρόβλημα παλινδρόμησης ορίζεται ως εξής :

$$\mathbf{Y}=\mathbf{X\beta}+\boldsymbol{\varepsilon},$$

όπου

- $\mathbf{Y} = [Y_1, \ldots, Y_n]^T$ είναι ένα $n \times 1$ διάνυσμα παρατηρήσεων
- $\mathbf{X} = [x_{ij}]$ είναι ένας $n \times (p+1)$ πίνακας σχεδιασμού πλήρους τάξης
- $\boldsymbol{\beta} = [\beta_0,...,\beta_p]^T$ είναι ένα $(p+1) \times 1$ διάνυσμα αγνώστων παραμέτρων
- $\boldsymbol{\varepsilon} = [\varepsilon_1,\ldots,\varepsilon_n]^T$ είναι ένα $n \times 1$ διάνυσμα ανεξάρτητων τυχαίων σφαλμάτων με

(α) $E[\boldsymbol{\varepsilon}] = \mathbf{0}$

(β) $\mathrm{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I_n}$

Η σταθερά $\sigma^2 > 0$ είναι μία άγνωστη παράμετρος.

Η μέθοδος ελαχίστων τετραγώνων έχει ως στόχο την εύρεση εκτιμητριών για τις παραμέτρους β_0, β_1, β_2,… έτσι ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων

$$SSE(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^{\mathrm{T}}\varepsilon = \mathbf{(Y - X\beta)^T (Y - X\beta)}.$$

Δείχνεται ότι η ελαχιστοποίηση του SSE(β) είναι ισοδύναμη με το επόμενο σύστημα p εξισώσεων με p αγνώστους (Κανονικές Εξισώσεις-Normal Equations)

$$\mathbf{X^T X\beta = X^T Y.}$$

Είναι φανερό ότι οι εκτιμητές ελαχίστων τετραγώνων μπορούν να γραφούν πινακικά από τον ακόλουθο τύπο:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Εύκολα μπορεί να επιβεβαιωθεί ότι ο $\widehat{\boldsymbol{\beta}}$ είναι αμερόληπτος και έχει πίνακα συνδιασποράς ανάλογο του πίνακα $(\mathbf{X}^T\mathbf{X})^{-1}$:

$$E[\,\widehat{\boldsymbol{\beta}}\,] = \boldsymbol{\beta}$$

$$\mathrm{Cov}[\,\widehat{\boldsymbol{\beta}}\,] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Ο κλασικός εκτιμητής της παραμέτρου σ^2 είναι ο μέσος τετραγωνικών σφαλμάτων:

$$s^2 = \frac{(\mathbf{Y} - \mathbf{X\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X\beta})}{n - p - 1} = \frac{\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}}{n - p - 1}.$$

3.4 Διαγνωστική παλινδρόμηση

Οι διαγνωστικές των έκτοπων παρατηρήσεων είναι στατιστικές που εστιάζουν το ενδιαφέρον τους σε παρατηρήσεις που ασκούν μεγάλη επίδραση στον εκτιμητή ελαχίστων τετραγώνων. Μερικά διαγνωστικά μέτρα έχουν σχεδιαστεί με σκοπό να ανιχνεύσουν μεμονωμένες περιπτώσεις ή σύνολο περιπτώσεων που διαφέρουν από τον όγκο των δεδομένων ή ασκούν ασυνήθιστα μεγάλη επίδραση στους εκτιμητές καθώς και στις προσαρμοσμένες τιμές. Το πεδίο των διαγνωστιικών μεθόδων αποτελείται από ένα συνδυασμό αριθμητικών και γραφικών εργαλείων.

Πολλές διαγνωστικές βασίζονται στα σφάλματα που απορρέουν από τα ελάχιστα τετράγωνα. Όμως αυτό μπορεί να οδηγήσει σε αναληθή αποτελέσματα εξαιτίας του ακόλουθου λόγου. Εξ ορισμού, η μέθοδος των ελαχίστων τετραγώνων προσπαθεί να αποφύγει τα μεγάλα σφάλματα. Συνεπώς, μία και μόνο έκτοπη παρατήρηση μπορεί να προκαλέσει μια φτωχή προσαρμογή στην πλειοψηφία των δεδομένων, επειδή ο εκτιμητής ελαχίστων τετραγώνων προσπαθεί να προσαρμόσει αυτή την περίπτωση εις βάρος των υπόλοιπων παρατηρήσεων. Ως εκ τούτου, μία έκτοπη παρατήρηση μπορεί να έχει ένα μικρό σφάλμα ελαχίστων τετραγώνων, ειδικά όταν αποτελεί ένα σημείο μόχλευσης. Συνεπώς, οι διαγνωστικές μέθοδοι που βασίζονται σε σφάλματα ελαχίστων τετραγώνων, συχνά αποτυγχάνουν να αποκαλύψουν αυτά τα σημεία.

Μία άλλη κατηγορία διαγνωστικών μεθόδων βασίζεται στην αρχή της διαγραφής μιας περίπτωσης σε κάθε στιγμή. Για παράδειγμα, ορίζουμε ως $\widehat{\boldsymbol{\beta}}$**(i)** τον εκτιμητή του **β** που υπολογίζεται από το δείγμα χωρίς την i-οστη περίπτωση. Στη συνέχεια η διαφορά μεταξύ του $\widehat{\boldsymbol{\beta}}$ και του $\widehat{\boldsymbol{\beta}}$**(i)** δίνει τον βαθμό στον οποίο η παρουσία της i-οστης περίπτωσης επηρεάζει τους συντελεστές παλινδρόμησης. Αυτές είναι οι λεγόμενες διαγνωστικές μονής περίπτωσης, οι οποίες υπολογίζονται για κάθε περίπτωση i και αυτό είναι πιθανό να γενικευθεί στις διαγνωστικές πολλαπλής περίπτωσης. Στην πραγματικότητα, η διαγραφή μιας ή περισσότερων περιπτώσεων φαίνεται ως ένας πολύ λογικός τρόπος για να εξετάσουμε την επίδρασή τους. Όμως δεν είναι πάντοτε φανερό ποιο υποσύνολο περιπτώσεων πρέπει να διαγραφεί. Μπορεί

να ισχύει ότι μερικά σημεία επηρεάζουν από κοινού, ενώ τα μεμονωμένα σημεία δεν ασκούν την ίδια επιρροή.

Μερικές ποσότητες που υπάρχουν συχνά στις κλασικές διαγνωστικές μεθόδους είναι τα διαγώνια στοιχεία των ελαχίστων τετραγώνων του πίνακα προβολής **H**. Αυτός ο πίνακας είναι πολύ γνωστός με το όνομα hat matrix. Αυτό σημαίνει ότι:

$$\widehat{\mathbf{Y}} = \mathbf{YH}$$

όπου $\widehat{\mathbf{Y}}$ είναι η πρόβλεψη των ελαχίστων τετραγώνων για το **Y**.
Ο hat matrix ορίζεται ως ακολούθως :

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t,$$

όπου ο $n\times(p+1)$ πίνακας **H** ονομάζεται πίνακας προβολής επειδή μετασχηματίζει το παρατηρούμενο διάνυσμα **Y** στον εκτιμητή ελαχίστων τετραγώνων $\widehat{\mathbf{Y}} = \mathbf{YH}$. Τα διαγώνια στοιχεία του πίνακα προβολής **H** είναι τα σημεία μόχλευσης.

3.5 Ανθεκτική παλινδρόμηση

Όταν κάνουμε παλινδρόμηση ελαχίστων τετραγώνων χρησιμοποιώντας n παρατηρήσεις, για ένα p-παραμετρικό μοντέλο $\mathbf{Y} = \mathbf{X\beta}+\boldsymbol{\varepsilon}$, κάνουμε ορισμένες ιδανικές υποθέσεις σχετικά με το διάνυσμα των σφαλμάτων **ε**, συγκεκριμένα ότι ακολουθεί $N(0,\mathbf{I}\sigma^2)$. Στην πράξη εμφανίζονται αποκλίσεις από αυτές τις υποθέσεις. Αν οι αποκλίσεις είναι σοβαρές, ελπίζουμε να φανούν στη συμπεριφορά των υπολοίπων και έτσι να αποτελέσουν τον οδηγό ώστε να γίνουν κατάλληλες τροποποιήσεις στο μοντέλο. Συχνά οι αποκλίσεις, αν τελικά υφίστανται, δεν είναι αρκετά σοβαρές για επανορθωτικές ενέργειες και προχωρούμε στην ανάλυση κατά συνήθη τρόπο.

Ορισμένοι τύποι αποκλίσεων από τις ιδανικές υποθέσεις θεωρούνται συχνά ότι είναι πιθανότερο να εμφανιστούν στην πράξη από κάποιους άλλους. Για παράδειγμα, η κατανομή των σφαλμάτων μπορεί να είναι συμμετρική αλλά όχι κανονική. Μπορεί δηλαδή να είναι «οξύτερη στην κορυφή» και με «λεπτότερα άκρα» από ότι η κανονική ή «λιγότερο οξεία στην κορυφή» και με «φαρδύτερα άκρα» από την κανονική. Ή, ακόμα και αν η κατανομή είναι κανονική, μπορεί να υπάρχουν απομονωμένες τιμές, δηλαδή να υπάρχουν μη τυπικές παρατηρήσεις της συνήθους κανονικής κατανομής με διαφορετική ίσως μέση τιμή ή παρατηρήσεις από μία κανονική κατανομή με κάπως μεγαλύτερη διασπορά από την υποτιθέμενη σ^2.

Υπάρχουν διάφορες υποδείξεις όπου, για την αντιμετώπιση αυτών καθώς και άλλων πιθανών μειονεκτημάτων, χρησιμοποιούνται μέθοδοι ανθεκτικής παλινδρόμησης, στη θέση των μεθόδων ελαχίστων τετραγώνων. Οι αρετές αυτών των μεθόδων είναι ότι είναι λιγότερο ευαίσθητες από ότι οι μέθοδοι των ελαχίστων τετραγώνων, στις συνήθεις πρακτικές αποκλίσεις από τις ιδανικές υποθέσεις.

Η ανθεκτική παλινδρόμηση προσπαθεί να επινοήσει εκτιμητές που δεν επηρεάζονται σε σημαντικό βαθμό από τις έκτοπες παρατηρήσεις. Πολλοί στατιστικοί που έχουν αμυδρώς ακούσει για την ανθεκτικότητα πιστεύουν ότι ο στόχος της είναι απλά να αγνοήσει τις έκτοπες παρατηρήσεις. Όμως σίγουρα η συγκεκριμένη αντίληψη δεν είναι σωστή. Αντιθέτως, στην ανθεκτική παλινδρόμηση οι έκτοπες παρατηρήσεις μπορούν να αναγνωριστούν κοιτάζοντας μόνο τα σφάλματα, κάτι το οποίο δεν συμβαίνει στην περίπτωση των σφαλμάτων των ελαχίστων τετραγώνων.

Κ Ε Φ Α Λ Α Ι Ο 4

Μέτρα ανθεκτικότητας

4.1 Εισαγωγή

Για να προσδιοριστεί ο βαθμός ανθεκτικότητας μιας διαδικασίας έχουν εισαχθεί διάφορα μέτρα ανθεκτικότητας. Τα πιο διαδεδομένα είναι η Συνάρτηση Επίδρασης (Influence Function), το Σημείο Κατάρρευσης (Breakdown Point), η Καμπύλη Ευαισθησίας (Sensitivity Curve) και η Ευαισθησία Γενικού Σφάλματος (Gross Error Sensitivity), τα οποία θα χρησιμοποιηθούν σε αυτή την διαδικασία. Καθένα από τα μέτρα αυτά περιγράφει διαφορετικά χαρακτηριστικά της διαδικασίας, συνεπώς λειτουργούν συμπληρώνοντας το ένα το άλλο.

Τόσο το σημείο κατάρρευσης όσο και η συνάρτηση επίδρασης δίνουν σημαντικές πληροφορίες αναφορικά με την ανθεκτικότητα ενός εκτιμητή. Αν η συνάρτηση επίδρασης είναι φραγμένη, τότε η επίδραση ενός μικρού αριθμού έκτοπων παρατηρήσεων στον εκτιμητή είναι επίσης φραγμένη.

4.2 Ανθεκτικό συμπέρασμα

Υποθέτουμε ότι έχουμε ένα μονοδιάστατο δείγμα $x_1,\ldots x_n$ ανεξάρτητων και ομοιόμορφα κατανεμημένων παρατηρήσεων με συνάρτηση κατανομής F_θ το οποίο ανήκει σε μία παραμετρική οικογένεια $\{F_\theta;\ \theta\in\Omega\}$. Στην κλασική στατιστική υποθέτουμε ότι τα x_i είναι κατανεμημένα όπως και η F_θ και αναλαμβάνουμε να εκτιμήσουμε το θ βάσει των δεδομένων. Στην ανθεκτική θεωρία συνειδητοποιούμε ότι το μοντέλο $\{F_\theta;\ \theta\in\Omega\}$ είναι μόνο μια ιδανική προσέγγιση της πραγματικότητας. Πολλές από τις υποθέσεις που γίνονται στη στατιστική όπως η κανονικότητα, η

γραμμικότητα και η ανεξαρτησία προσεγγίζουν την πραγματικότητα. Τις περισσότερες φορές χρησιμοποιούμε βέλτιστες διαδικασίες κάτω από αυτές τις υποθέσεις. Όμως το πρόβλημα που δημιουργείται είναι ότι μια μικρή απόκλιση από το μοντέλο μπορεί να επηρεάσει σοβαρά τα αποτελέσματα. Στόχος της ανθεκτικής στατιστικής είναι να περιγράψει τη δομή στην οποία προσαρμόζεται καλύτερα ο όγκος των δεδομένων, να αναγνωρίσει τις έκτοπες παρατηρήσεις καθώς επίσης να προειδοποιήσει για τα σημεία μόχλευσης [Hampel et al. (1986)].

4.3 Συναρτησοειδή (functionals)

Έστω $\mathbf{x}^{\mathbf{T}} = (x_1, x_2, ..., x_n)$ τυχαίο δείγμα που προέρχεται από την κατανομή F_θ. Η εμπειρική συνάρτηση κατανομής του δείγματος είναι

$$F_n = \frac{1}{n}\sum_{i=1}^{n} \Delta x_i ,$$

όπου Δx_i είναι η συνάρτηση κατανομής που τοποθετεί μάζα πιθανότητας 1 στο σημείο x_i. Η εκτίμηση του θ γίνεται μέσω στατιστικών συναρτήσεων της μορφής

$$T_n = T_n(x_1, x_2, \ldots, x_n) = T_n(F_n).$$

Οι εκτιμητές μπορούν να αντιμετωπιστούν ως συναρτησοειδή, επειδή η θεώρηση αυτή συνεισφέρει στη μελέτη ιδιοτήτων των εκτιμητών, κυρίως σε σχέση με την ασυμπτωτική τους συμπεριφορά.

4.3.1 Συνάρτηση Επίδρασης (ΣΕ) – Influence Function

Ένα από τα βασικότερα μέτρα ανθεκτικότητας είναι η συνάρτηση επίδρασης ενός εκτιμητή T. [βλέπε Hampel et al. (1986)]. Προκύπτει ως η πρώτη παράγωγος του εκτιμητή, που θεωρείται συναρτησοειδές σε κάποια κατανομή. Χρησιμοποιείται για να μελετηθούν ιδιότητες ανθεκτικότητας και για να βρεθούν ασυμπτωτικές διασπορές. Επίσης για την εύρεση εκτιμητών με προκαθορισμένα χαρακτηριστικά ως προς την ανθεκτικότητα. Η συνάρτηση επίδρασης περιγράφει την επίδραση που έχει στον εκτιμητή μια πολύ μικρή αλλοίωση στο σημείο x, μετρώντας την ασυμπτωτική μεροληψία που προκαλείται. Ορίζεται για ένα μεγάλο μέγεθος δείγματος, όμως στην πράξη είναι συχνά χρήσιμη σε δείγματα μεγέθους 20 ή ακόμα μικρότερου.

Ορισμός 1: Η συνάρτηση επίδρασης ενός εκτιμητή T της κατανομής F που καταμετρά την επίδραση των απειροστών αναταραχών στον εκτιμητή δίνεται από τη σχέση:

$$\text{IF(x ;T,F)}=\lim_{t\to 0}\frac{\text{T}((1-t)F+t\Delta x)-T(F)}{t},$$

σε αυτά τα σημεία x του δειγματικού χώρου, όπου το όριο υπάρχει. Αν στον τύπο της ΣΕ αντικαταστήσουμε το F με το F_{n-1}, όπου F_n είναι η εμπειρική κατανομή εκτίμησης του F και βάλουμε $t=\frac{1}{n}$, συνειδητοποιούμε ότι η $IF(x; T,F_{n-1})$ μετράει περίπου n φορές την αλλαγή στο T που προκαλείται από μία επιπρόσθετη παρατήρηση στο x όταν ο T εφαρμόζεται σε ένα μεγάλο δείγμα μεγέθους n-1 [βλέπε Hampel et al. (1986)].

Εάν κάποια κατανομή G είναι «κοντά» στο F, τότε το θεώρημα ανάπτυξης του Von Mises στο T γύρω από το F εκτιμώμενο στο G δίνεται από τη σχέση:

$$\text{T(G)} = \text{T(F)}+\int IF(x;T,F)d(G-F)(x)+R_n, \qquad (4.3.1.1)$$

όπου R_n είναι ο υπολειπόμενος όρος. Όταν οι παρατηρήσεις x_i είναι i.i.d σύμφωνα με το F, τότε η εμπειρική κατανομή F_n τείνει στο F σύμφωνα με το θεώρημα του Glivenko-Cantelli. Αν στη σχέση (4.3.1.1) αντικαταστήσουμε το G με το F_n για μεγάλο n και επίσης υποθέσουμε ότι T(F) μπορεί να προσεγγιστεί από την $T_n=T(F_n)$ και λαμβάνοντας υπόψη ότι:

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

τότε καταλήγουμε στο ότι:

$$\text{T}_n=\text{T(F)}+\int IF(x;T,F)dF_n(x)+R_n,$$

με $R_n= o_p(\frac{1}{\sqrt{n}})$.

Αν υπολογίσουμε το ολοκλήρωμα ως προς F_n έχουμε:

$$\sqrt{n}[T_n-T(F)]\approx\frac{1}{\sqrt{n}}\sum IF(X_i;T,F)+R_n.$$

Από τη χρήση του κεντρικού οριακού θεωρήματος προκύπτει:

$$\sqrt{n}[T_n-T(F)]\to N[0,V(T,F)],$$

όπου V(T,F) είναι η ασυμπτωτική διασπορά που ορίζεται ως εξής:

$$V(T,F) = \int IF^2(x;T,F)dF(x).$$

4.3.2 Σημείο Κατάρρευσης (ΣΚ) – Breakdown Point

Μία και μόνο απομονωμένη τιμή στο σύνολο των δεδομένων μπορεί ναι έχει σαν αποτέλεσμα την μετατόπιση του εκτιμητή ελαχίστων τετραγώνων. Κατά συνέπεια, εύκολα προκύπτει ότι ο εκτιμητής ελαχίστων τετραγώνων δεν θεωρείται ανθεκτικός εκτιμητής. Απεναντίας υπάρχουν εκτιμητές οι οποίοι έχουν τη δυνατότητα να αντιμετωπίσουν ένα συγκεκριμένο ποσοστό έκτοπων παρατηρήσεων. Προκειμένου να διατυπωθεί αυτή η άποψη, εισάγεται η έννοια του σημείου κατάρρευσης

Το σημείο κατάρρευσης είναι ένα ολικό μέτρο αξιοπιστίας. Μετράει πόση αλλοίωση μπορεί να αντέξει ο εκτιμητής, υπολογίζοντας το μέγιστο κλάσμα αλλοίωσης που διατηρεί τον εκτιμητή υπό έλεγχο, ενώ ένα μεγαλύτερο κλάσμα αλλοίωσης προκαλεί κατάρρευση του εκτιμητή. Η κατάρρευση σχετίζεται, κατά κύριο λόγο, με μη προβλέψιμη τιμή του εκτιμητή. Ένας εκτιμητής είναι ανθεκτικός μόνο όταν έχει σημείο κατάρρευσης μεγαλύτερο από το μηδέν.

Το σημείο κατάρρευσης μαζί με την συνάρτηση επίδρασης, είναι τα πιο διαδεδομένα μέτρα ανθεκτικότητας. Σημείο κατάρρευσης μπορεί να βρεθεί για όλους τους εκτιμητές, όμως δεν συμβαίνει το ίδιο για τη συνάρτηση επίδρασης.

Έστω ότι παίρνουμε οποιοδήποτε δείγμα από n σημεία δεδομένων

$$X=\{(x_{i1},\ldots,x_{1p},y_1),\ldots,(x_{n1},\ldots x_{np},y_n)\}$$

και ορίζουμε ως T τον εκτιμητή παλινδρόμησης έτσι ώστε

$$\hat{\boldsymbol{\beta}} = T(\quad).$$

Θεωρούμε όλα τα πιθανά αλλοιωμένα δείγματα $'$ τα οποία επιτυγχάνονται αντικαθιστώντας οποιαδήποτε m από τα αρχικά σημεία δεδομένων με αυθαίρετες τιμές. Αφήνουμε τη μέγιστη μεροληψία που μπορεί να προκληθεί από τέτοια αλλοίωση να είναι:

$$\text{bias}(m;T,X)=\sup\left\|T(\quad')-T(\quad)\right\|.$$

Αν η $\text{bias}(m;T,X)\to\infty$, αυτό υποδηλώνει ότι οι m έκτοπες παρατηρήσεις μπορούν να έχουν μία αυθαίρετα μεγάλη επίδραση στον T με αποτέλεσμα ο εκτιμητής να καταρρέει.

Ορισμός 2: Το σημείο κατάρρευσης του εκτιμητή Τ στο δείγμα Χ ορίζεται ως εξής:

$$\mathrm{E}_n^*(T, \quad) = \min\{\frac{m}{n}; bias; T, \quad\} \text{ είναι άπειρο.}$$

Με άλλα λόγια ΣΚ είναι το μικρότερο τμήμα της μόλυνσης που μπορεί να προκαλέσει τη μέθοδο παλινδρόμησης Τ να απομακρυνθεί αυθαίρετα από το Τ(Χ). Στα ελάχιστα τετράγωνα, το ΣΚ ισοδυναμεί με $1/n$ το οποίο τείνει να γίνει μηδέν καθώς το μέγεθος του δείγματος αυξάνεται. Κατά συνέπεια το ΣΚ πρέπει να είναι περισσότερο από 0%. Ένα ΣΚ της τάξης του 50% είναι το καλύτερο που μπορούμε να αναμένουμε. Στη μέθοδο ελαχίστων τετραγώνων μία και μόνο έκτοπη παρατήρηση είναι αρκετή για να καταστρέψει τον εκτιμητή T_0. Το σημείο κατάρρευσης είναι $\mathrm{E}_N^*(T, X) = 1/n$ και ως εκ τούτου $\mathrm{E}^*(T) = 0$. Οι εκτιμητές Τ με $\mathrm{E}^*(T) > 0$ ονομάζονται θετικές μέθοδοι κατάρρευσης.

4.3.3 Ευαισθησία Γενικού Σφάλματος (ΕΓΣ) – Gross Error Sensitivity

Για την επίτευξη ανθεκτικότητας, είναι επιθυμητή μία φραγμένη ΣΕ. Ένα μέτρο ανθεκτικότητας, που προκύπτει από την ΣΕ και προσδιορίζει άμεσα την ανθεκτικότητα ή μη ενός εκτιμητή, είναι η Ευαισθησία Γενικού Σφάλματος.

Ορισμός 3: Η ευαισθησία του γενικού σφάλματος του εκτιμητή Τ στην κατανομή F ορίζεται από τη σχέση

$$\gamma^* = \sup_x \left| IF(x; T, F) \right|.$$

Η ΕΓΣ περιγράφει τη μέγιστη επίδραση που έχει στην τιμή του εκτιμητή μια μικρή αλλοίωση της κατανομής και για την εξασφάλιση ανθεκτικότητας επιδιώκεται να είναι πεπερασμένη.

Οι πεπερασμένες τιμές του γ^* είναι επιθυμητές γιατί έχουν σαν αποτέλεσμα να προκαλούν μικρές αλλαγές στον εκτιμητή σε περίπτωση που προστεθεί ένας μικρός αριθμός έκτοπων παρατηρήσεων

4.3.4 Καμπύλη ευαισθησίας (ΚΕ) - Sensitivity Curve

Η ΚΕ μετρά την επίδραση που θα ασκηθεί στον εκτιμητή από μία και μόνο επιπρόσθετη παρατήρηση x.

Ορισμός 4: Δοθέντος ενός τυχαίου δείγματος $x_1,\ldots,x_{n-1}$ η ΚΕ ενός εκτιμητή T_n σε ένα σημείο x ορίζεται ως εξής:

$$SC_n(X;T_n)=nT_n(x_1,\ldots,x_{n-1},x) - T_{n-1}(x_1,\ldots,x_{n-1})).$$

Όταν $T_n(x_1,x_2,\ldots,x_n)=T(F_n)$ για κάθε n, με την αντίστοιχη εμπειρική κατανομή F_n τότε:

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

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

4.4 Ορισμοί (Rousseeuw & Leory, 1987)

Ο εκτιμητής θέσης T_n που στηρίζεται στο τυχαίο δείγμα $(x_1,\ldots,x_n)$, είναι:

- Ίσης μεταβολής ως προς θέση (location equivariant), όταν
  $T_n(x_1+u, x_2+u,\ldots,x_n+u)=T_n(x_1,x_2,\ldots,x_n)+u,$

για κάθε σταθερά u.

- Ίσης μεταβολής ως προς κλίμακα (scale equivariant), όταν
  $T_n(cx_1,cx_2,\ldots,cx_n)=cT_n(x_1,x_2,\ldots,x_n),$

για κάθε σταθερά $c \neq 0$.

- Ίσης μεταβολής ως προς θέση και κλίμακα (affine equivariant), όταν
  $T_n(cx_1+u,cx_2+u,\ldots,cx_n+u)=cT_n(x_1,x_2,\ldots,x_n)+u,$

για σταθερές u και $c \neq 0$.

- Ίσης μεταβολής ως προς την παλινδρόμηση (regression equivariant), όταν
  $T(\{(x_i, y_i + x_iv); i=1,\ldots n\})=T(\{(x_i, y_i); i=1,\ldots,n\}) + v,$

όπου v είναι ένα οποιοδήποτε διάνυσμα στήλη.

- Εκτιμητής κλίμακας (scale estimator) είναι κάθε στατιστική συνάρτηση S_n τέτοια ώστε

$$S_n(\alpha+bx_1,\alpha+bx_2,\ldots,\alpha+bx_n)=|b|Sn(x_1,x_2,\ldots,x_n)>0.$$

4.5 Συνάρτηση Επίδρασης για τα ελάχιστα τετράγωνα

Υποθέτουμε ότι οι γραμμές του πίνακα σχεδιασμού **X** είναι i.i.d παρατηρήσεις από μία p-διάστατη κατανομή $F_{x.}$ Έστω F είναι η από κοινού κατανομή για το (p+1)-διάστατο διάνυσμα $[\mathbf{x}^T y]$, όπου $\mathbf{x}^T$ είναι μια τυχαία γραμμή του **X** και y είναι η εξαρτημένη μεταβλητή που ορίζεται ως $y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$. Είναι χρήσιμο να θεωρήσουμε την F ότι προσδιορίζεται από την περιθώρια κατανομή του $\mathbf{x}^T$, και την δεσμευμένη κατανομή $F_{y/\mathbf{x}}$ του y δοθέντος του $\mathbf{x}^T$. Το τελευταίο προσδιορίζεται από την κατανομή του ε, έστω F_ε και ο ορισμός είναι:

$$y - \mathbf{x}^T\boldsymbol{\beta} = \varepsilon \sim F_{\varepsilon.}$$

Το πρόβλημα παλινδρόμησης χαρακτηρίζεται από την υπόθεση ότι $y - \mathbf{x}^T\boldsymbol{\beta}$ και $\mathbf{x}^T$ είναι ανεξάρτητα.

Ακολούθως ορίζουμε ως F_n την εμπειρική κατανομή η οποία θέτει μάζα $1/n$ σε κάθε μία από τις n γραμμές $[\mathbf{x_i}^T, y_i]$, i=1,…n, του πίνακα $[\mathbf{X}/\mathbf{Y}]$. Θέλουμε να εκφράσουμε τον εκτιμητή $\widehat{\boldsymbol{\beta}} = \mathbf{T}(F_n)$, όπου **T(**F**)** είναι ένα συναρτησοειδές διάνυσμα σε μία κατάλληλη κατηγορία κατανομών και F_n είναι μία εμπειρική κατανομή αυτής της κατηγορίας. Χρησιμοποιούμε την προσέγγιση του Hinkley (1977) και ορίζουμε το συναρτησοειδές διάνυσμα $\gamma(F) = E_F[y\mathbf{x}]$. Όταν εκτιμάται στο F_n,

$$\boldsymbol{\gamma}(F_n) = \frac{1}{n}\sum_{i=1}^{n} y_i \quad = \frac{1}{n}\mathbf{X}^T\mathbf{Y}\,.$$

Επίσης το **Σ(**F**)** ορίζεται ως $E_F[\mathbf{x}\mathbf{x}^T]$ το οποίο ικανοποιεί τη σχέση:

$$\boldsymbol{\Sigma}(F_n) = \frac{1}{n}\sum_{=} \mathbf{x_i}\mathbf{x_i}^T = \frac{1}{n}\mathbf{X}^T\mathbf{X}.$$

Ως εκ τούτου το συναρτησοειδές των ελαχίστων τετραγώνων είναι το $\mathbf{T}(F) = \boldsymbol{\Sigma}^{-1}(F)\boldsymbol{\gamma}(F)$ από όπου προκύπτει ότι

$$\mathbf{T}(F_n) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Η συνάρτηση επίδρασης των ελαχίστων τετραγώνων μας φανερώνει πως οι έκτοπες παρατηρήσεις τόσο στην εξαρτημένη όσο και στην ανεξάρτητη μεταβλητή

μπορούν να επηρεάσουν τον εκτιμητή $\widehat{\boldsymbol{\beta}}$ των ελαχίστων τετραγώνων. Εξ ορισμού, η συνάρτηση επίδρασης δίνει για κάθε $[\mathbf{x}^{\mathbf{T}}y]$ την κατευθυντήρια παράγωγο του T στο F στην κατεύθυνση του Δ – F, το οποίο τοποθετεί μάζα πιθανότητας 1 στο $(\mathbf{x}^{\mathbf{T}},y)$,

$$\mathbf{IF}(\mathbf{x}^{\mathbf{T}},y;F)=\lim_{t\to 0}\frac{\mathbf{T}[(1-t)F+t\Delta]-\mathbf{T}(F)}{t}.$$

Ακολουθώντας τον Cook Weisberg (1982) η συνάρτηση επίδρασης για τα ελάχιστα τετράγωνα ορίζεται ως:

$$\mathbf{IF}(\mathbf{x}^{\mathbf{T}},\ ;F)=\boldsymbol{\Sigma}^{-1}[\mathbf{x}(\ -\mathbf{x}^{\mathbf{T}}\mathbf{T}(F)].$$

Στη συνέχεια παίρνουμε την ανάπτυξη του von Mises

$$\widehat{\boldsymbol{\beta}}=\mathbf{T}(F_n)=\mathbf{T}(F)+-\sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{IF}(\mathbf{x_i}^{\mathbf{T}},\ ;F)+R_n,$$

όπου R_n είναι ο υπολειπόμενος όρος και F_n είναι η εμπειρική συνάρτηση της F.

Κ Ε Φ Α Λ Α Ι Ο 5

Ανθεκτικοί Εκτιμητές

5.1 Οι βασικότεροι ανθεκτικοί εκτιμητές

Η ανάλυση παλινδρόμησης ασχολείται με την ανακάλυψη εκτιμητών που επηρεάζονται όσο το δυνατόν λιγότερο από την ύπαρξη έκτοπων παρατηρήσεων. Η θεωρία των πολυδιάστατων ανθεκτικών κατανομών αναπτύχθηκε τη δεκαετία του ’70 και οι βασικότεροι εκτιμητές είναι οι L, οι R και οι M. Συγκεκριμένα:

- **Οι L-εκτιμητές** είναι γραμμικοί συνδυασμοί του διατεταγμένου δείγματος. Είναι οι πιο κατάλληλοι εκτιμητές κεντρικής τιμής και κεντρικής τάσης, και ως εκ τούτου μπορούν κατά διαστήματα να εφαρμοστούν σε προβλήματα εκτίμησης παραμέτρων. Οι δύο πιο βασικοί L-εκτιμητές είναι i) η διάμεσος ii) η μέση τιμή μετά από ποσοστιαία αποκοπή του Tukey, που ορίζονται ως τον σταθμικό μέσο του πρώτου, δεύτερου και τρίτου τεταρτημορίου σε μία κατανομή, με βάρη ¼, ½, και ¼ αντίστοιχα.

- **Οι R-εκτιμητές** βασίζονται στους ελέγχους βαθμών. Για παράδειγμα η ισότητα ή η ανισότητα δύο κατανομών μπορεί να υπολογιστεί από το τεστ του Wilcoxon υπολογίζοντας τον μέσο βαθμό της μιας κατανομής σε ένα συνδυασμένο δείγμα και των δύο κατανομών. Η στατιστική συνάρτηση του Kolmogorov-Smirnov και οι βαθμοί διάταξης των συντελεστών συσχέτισης του Spearman είναι στην πραγματικότητα R-εκτιμητές.

- **Οι M-εκτιμητές** θεμελιώθηκαν από τον Huber (1973) και θεωρούνται από τις πιο απλές προσεγγίσεις τόσο θεωρητικά όσο και υπολογιστικά. Παρά το

γεγονός ότι δεν είναι ανθεκτικοί αναφορικά με τα σημεία μόχλευσης εξακολουθούν να χρησιμοποιούνται ευρέως στην ανάλυση δεδομένων για την οποία μπορεί να υποτεθεί ότι η αλλοίωση είναι κυρίως στην κατεύθυνση των y_i.

- **Οι γενικευμένοι Μ-εκτιμητές** εισήχθηκαν από τον Mallows [βλέπε Maronna et al. (1979)] έχοντας ως βασικό σκοπό να φράξουν τις έκτοπες παρατηρήσεις στη κατεύθυνση των x_i χρησιμοποιώντας τη συνάρτηση βάρους w. Οι γενικευμένοι εκτιμητές έχουν ένα σημείο κατάρρευσης του οποίου η τιμή μειώνεται ανάλογα με τον αριθμό των συντελεστών παλινδρόμησης.

- **Η S-εκτίμηση** είναι μία μέθοδος υψηλού σημείου κατάρρευσης η οποία εισήχθηκε από τον Rousseeuw & Yohai (1984). Η S-εκτίμηση βασίζεται στην ελαχιστοποίηση του ανθεκτικού Μ- εκτιμητή της κλίμακας των σφαλμάτων. Με το ίδιο σημείο κατάρρευσης ο εκτιμητής έχει υψηλότερη στατιστική αποτελεσματικότητα σε σχέση με τον εκτιμητή ελαχίστων περικεκομμένων τετραγώνων.

- **Η εκτίμηση ελαχίστων διαμέσων τετραγώνων** εισήχθηκε από τον Rousseeuw (1984). Είναι μια μέθοδος πολύ ανθεκτική με υψηλό σημείο κατάρρευσης. Συγκεκριμένα είναι ο πρώτος εκτιμητής παλινδρόμησης ίσης μεταβολής που κατόρθωσε να φτάσει στο μέγιστο ασυμπτωτικό σημείο κατάρρευσης ½. Η μέθοδος αυτή θεωρείται ότι είναι ένας S-εκτιμητής ο οποίος ελαχιστοποιεί έναν τύπο ανθεκτικής Μ-εκτίμησης κλίμακας επί των σφαλμάτων.

- **Η εκτίμηση ελαχίστων περικεκομμένων τετραγώνων** είναι μια μέθοδος με υψηλό σημείο κατάρρευσης [βλέπε Rousseeuw & Leroy (1987)]. Η εκτίμηση ΕΠΤ -ένας άλλος ευρύτατα διαδεδομένος S-εκτιμητής- προσεγγίζει τη μέθοδο ελαχίστων τετραγώνων αν και στην προκειμένη περίπτωση αφαιρούμε το τμήμα των μεγαλύτερων τετραγωνικών σφαλμάτων.

- **Οι MM-εκτιμητές** που εισήχθηκαν από τον Yohai (1987) συνδυάζουν την εκτίμηση υψηλού σημείου κατάρρευσης και την M-εκτίμηση. Έχουν υψηλότερη στατιστική αποτελεσματικότητα και υψηλότερο σημείο κατάρρευσης από ότι η S-εκτίμηση.

5.2 L- εκτιμητές

Ένας πολύ ικανοποιητικός ανθεκτικός εκτιμητής που ουσιαστικά αντικατέστησε τη μέθοδο ελαχίστων τετραγώνων είναι ο εκτιμητής ελαχίστων απολύτων τιμών. Η συγκεκριμένη μέθοδος είναι ευρύτατα διαδεδομένη και πολύ αναγνωρισμένη ειδικότερα στα σφάλματα κατανομών απώλειας. Ο L_1-εκτιμητής είναι πιθανότατα ο πιο παλιός ανθεκτικός εκτιμητής. Οι ρίζες του εντοπίζονται από την εποχή του Galilei (1632) ο οποίος στην προσπάθειά του να εξακριβώσει την ακριβή θέση ενός νέου άστρου πρότεινε την ελάχιστη πιθανή διόρθωση με σκοπό να πετύχει ένα αξιόπιστο αποτέλεσμα για την επίλυση του συγκεκριμένου προβλήματος. Όμως ο L_1-εκτιμητής πιο αναλυτικά μελετήθηκε από τον Boscovich (1757) και τον Laplace (1793). Σχεδόν ύστερα από 70 χρόνια μετά την έκδοση του βιβλίου του Laplace (1818), ο Edgeworth (1887) παρουσίασε μία μέθοδο για την γραμμική παλινδρόμηση χρησιμοποιώντας τον L_1-εκτιμητή.

Κάποιες από τις βασικότερες μορφές των ελαχίστων απολύτων τιμών είναι οι ακόλουθες:

- Παλινδρόμηση ελαχίστων απολύτων τιμών (L1 παλινδρόμηση)

Προσδιορίζεται από

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

Στην εκτίμηση θέσης μονομεταβλητής περίπτωσης ο L_1-εκτιμητής είναι η δειγματική διάμεσος, του οποίου το σημείο κατάρρευσης θεωρείται ότι είναι 50%.

- α-παλινδρόμηση ποσοστιαίου σημείου (α-regression quintile)

Ο L_1-εκτιμητής γενικεύεται ως εξής:

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

όπου

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

Ο εκτιμητής που προκύπτει ονομάζεται α-παλινδρόμηση ποσοστιαίου σημείου του οποίου το σημείο κατάρρευσης είναι 0%.(Για τον L_1-εκτιμητή θεωρούμε ότι α=0.5)

- α-περικεκομμένος εκτιμητής (a-trimmed estimator)

Σχετικά με την εκτίμηση θέσης, ο α-περικεκομμένος εκτιμητής ορίζεται ως εξής:

$$\hat{\beta} = \frac{1}{n - 2\lfloor n\alpha \rfloor} \sum_{i=\lfloor n\alpha \rfloor + 1}^{n - \lfloor n\alpha \rfloor} y_{i:n}$$

όπου το α κυμαίνεται μεταξύ 0 και 1,το $y_{1:n},\ldots,y_{n:n}$ συμβολίζει το διατεταγμένο δείγμα και τέλος το [.] υποδηλώνει τη στρογγυλοποίηση του ακεραίου. Θα πρέπει επίσης να επισημανθεί ότι

a. για κανονικό δειγματικό μέσο, το α=0
b. αν το α=1/4, ο εκτιμητής αναφέρεται ως κεντρικός μέσος
c. το σημείο κατάρρευσης ισούται με α%

Στην πολλαπλή παλινδρόμηση, ο α-περικεκομμένος μέσος μπορεί να επεκτεθεί σε α-περικεκομμένο εκτιμητή ελαχίστων τετραγώνων ο οποίος είναι ευαίσθητος στα σημεία μόχλευσης και του οποίου το σημείο κατάρρευσης είναι 0%.

5.3 R-εκτιμητές

Οι R-εκτιμητές βασίζονται στους βαθμούς των σφαλμάτων. Αν R_i είναι ο βαθμός του $r_i = y_i - \mathbf{x_i}\boldsymbol{\beta}$, τότε ο στόχος μας είναι να

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

όπου η συνάρτηση των scores $\alpha_n(i)$ είναι μονότονη και ικανοποιεί τη συνθήκη

$$\sum_{i=1}^{n} \alpha_n(i) = 0$$

Μερικές πιθανές εκδοχές των scores είναι:

- Wilcoxon scores

$\alpha_n(i) = i - (n+1)/2$

- Van der Waerden scores

$\alpha_n(i) = \Phi^{-1}(i/(n+1))$

όπου Φ^{-1} είναι η αντίστροφη συνάρτηση της αθροιστικής κανονικής κατανομής.

- median score

$\alpha_n(i) = \text{sgn}(i - (n+1)/2)$

- bounded normal scores

$\alpha_n(i) = \min(c, \max\{\Phi^{-1}(i/(n+1)), -c\})$

όπου c είναι μία σταθερά.

Ένα σημαντικό πλεονέκτημα του R-εκτιμητή έναντι του M-εκτιμητή είναι ότι γίνεται αυτόματα ίσης μεταβολής ως προς τη κλίμακα. Παρόλα αυτά όμως οι R-εκτιμητές δεν μπορούν να αντιμετωπίσουν τις απομονωμένες τιμές στον άξονα x με αποτέλεσμα το σημείο κατάρρευσης να εξακολουθεί να είναι 0%.

5.4 M-εκτιμητές

Οι M-εκτιμητές είναι μια γενίκευση των εκτιμητών μεγίστης πιθανοφάνειας. Αντί να ελαχιστοποιήσουμε το άθροισμα των scores, $\log f(x,\beta)$ όπως και στην εκτίμηση μεγίστης πιθανοφάνειας, χρησιμοποιείται μια πιο γενική συνάρτηση $\rho(x,\beta)$ [Huber(1964, 1981)].

Κάθε εκτιμητής T_n που ορίζεται από το πρόβλημα ελαχιστοποίησης του τύπου:

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

ή από μία έμμεση εξίσωση

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

όπου ρ είναι μια αυθαίρετη συνάρτηση και έχει ως παράγωγο ψ(x,β)=$\frac{\partial}{\partial\beta}\rho(x,\beta)$, ονομάζεται Μ-εκτιμητής ή τύπος εκτίμησης μεγίστης πιθανοφάνειας. Η συνάρτηση επίδρασης ενός Μ-εκτιμητή δίνεται από τη σχέση:

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

με ασυμπτωτική διασπορά:

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

Κάποιοι από τους βασικότερους Μ-εκτιμητές που έχουν προταθεί είναι οι ακόλουθοι:

- **Huber minimax**

$$\psi(t)=\begin{bmatrix} t & \alpha\nu & |t|<b \\ b\,\mathrm{sgn}(t) & \alpha\nu & |t|\geq b \end{bmatrix}$$

όπου b είναι μια σταθερά

- **Descending minimax**

$$\psi(t)=\begin{bmatrix} t & \alpha\nu\,|t|<a \\ b\,\mathrm{sgn}(t)\tanh[\frac{1}{2}b(c-|t|)] & \alpha\nu\,a\leq|t|<c \\ 0 & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{bmatrix}$$

όπου α,b και c είναι σταθερές

- **Hampel**

$$\psi(t)=\begin{bmatrix} t & \alpha\nu\,|t|<\alpha \\ a\,\mathrm{sgn}(t) & \alpha\nu\,a\leq|t|<b \\ \{(c-|t|)/(c-b)\}a\,\mathrm{sgn}(t) & \alpha\nu\,b\leq|t|<c \\ 0 & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{bmatrix}$$

όπου α,b και c είναι σταθερές

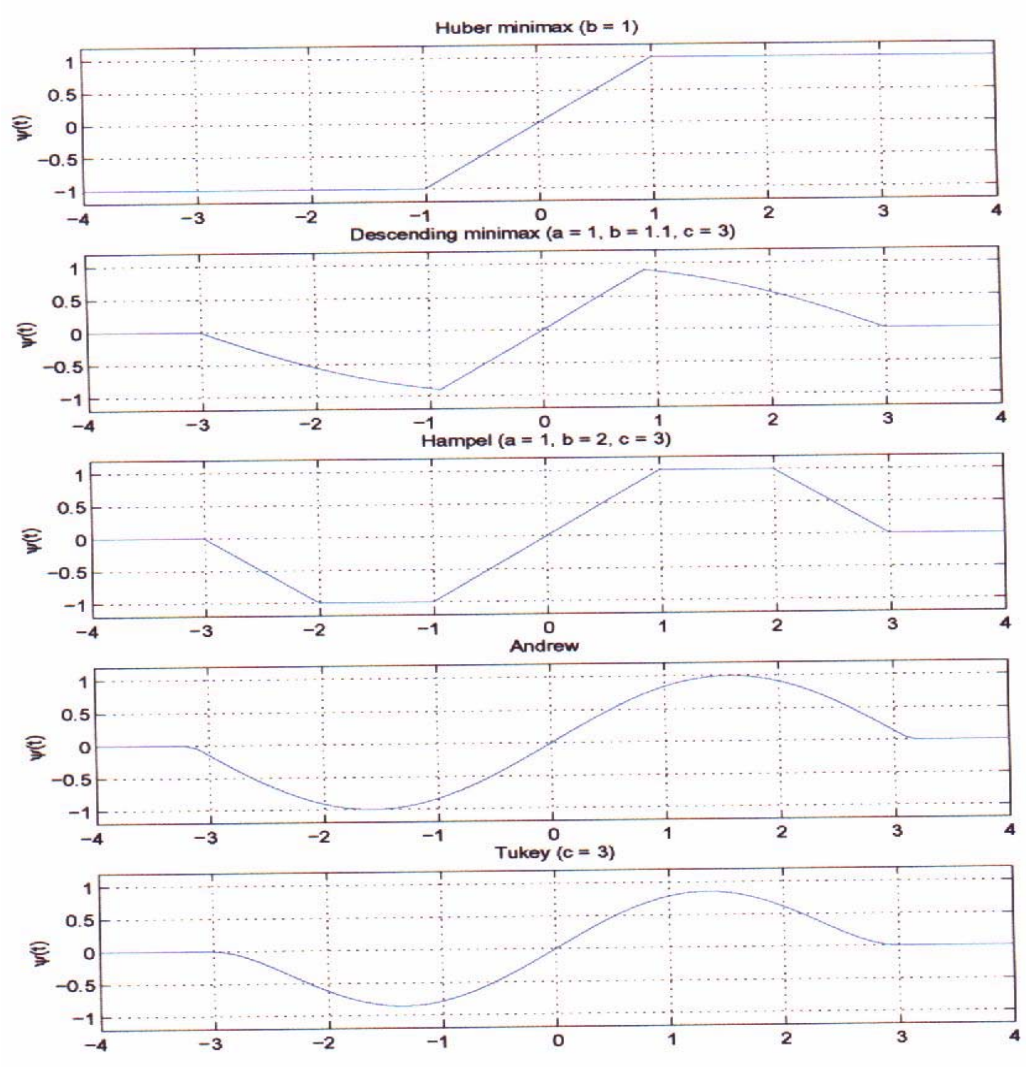
- **Andrew**

$$\psi(t) = \begin{cases} \sin(t) & \alpha\nu \quad -\pi \le t < \pi \\ 0 & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

- **Tukey**

$$\psi(t) = \begin{cases} t(1-(t/c)^2)^2 & \alpha\nu \quad |t| < c \\ 0 & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

Διάγραμμα 5.1
Παραδείγματα συναρτήσεων Μ-εκτιμητών



Οι Μ-εκτιμητές είναι στατιστικά περισσότερο αποτελεσματικοί από την L_1 παλινδρόμηση, ενώ την ίδια στιγμή είναι ακόμα ανθεκτικοί αναφορικά με τις έκτοπες τιμές στην κατεύθυνση των y_i .Το σημείο κατάρρευσης των Μ-εκτιμητών είναι 0% εξαιτίας της ευπάθειας των σημείων μόχλευσης.

5.5 Μέθοδοι ανθεκτικής παλινδρόμησης

5.5.1 Μ-εκτίμηση παλινδρόμησης του Huber

Στην ανάλυση παλινδρόμησης, οι Μ-εκτιμητές μπορούν να προκύψουν σαν μία λύση του ακόλουθου τύπου ελαχιστοποίησης:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\rho\left(\frac{Y_i - f_i(\beta)}{\sigma}\right) + \Lambda\right]\sigma$$

(5.5.1.1)

όπου $f_i(\beta) = \sum_{j=0}^{p} \mathbf{x}_{ij}^{\mathbf{T}} \beta_j$ με ρ(0) = 0 και Λ>0. Διαφορίζοντας την (5.5.1.1) ως προς β και σ παίρνουμε

$$\frac{1}{n}\sum_{i-1}^{n}\psi\left(\frac{r_i}{\sigma}\right)\frac{\vartheta f_i(\beta)}{\vartheta \beta_j} = 0,$$

για j = 0,1,…,p και

$$\frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{r_i}{\sigma}\right) = \Lambda,$$

όπου

$$r_i = Y_i - f_i(\beta),$$

με

$$\psi(t) = \rho'(t) \text{ και } \chi(t) = t\psi(t) - \rho(t).$$

Πρέπει να ληφθεί υπόψη επίσης ότι το χ(t) έχει απόλυτο ελάχιστο στο t = 0, συγκεκριμένα χ(0) = 0. Επίσης υποθέτουμε ότι οι ψ και χ είναι συνεχής. Για να πετύχουμε συνέπεια του εκτιμητή κλίμακας στο κανονικό μοντέλο για την κλασική επιλογή $\rho(t) = \frac{1}{2}t^2$, παίρνουμε

$$\Lambda = \frac{(n-p-1)}{n}E_{\Phi}(\chi),$$

όπου Φ είναι η αθροιστική συνάρτηση της κανονικής κατανομής.

Η συνάρτηση επίδρασης του Huber εκτιμητή T^{Hu} στο μοντέλο κατανομής $F_\beta(x,y)$ με σ.π.π $f_\beta(x,y)=\varphi(y-x^T\beta)k(x)$, (k(x) η σ.π.π. του x) δίνεται από τη σχέση:

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

όπου

$$\mathbf{M}=(E\psi'_c)(E\mathbf{x}\mathbf{x}^T)=\int \psi'_c(r)d\Phi(r)\int \mathbf{x}\mathbf{x}^{\mathbf{T}}dK(x)$$

και K(x) η συνάρτηση κατανομής του x. Για περισσότερες πληροφορίες ο αναγνώστης μπορεί να ανατρέξει στον Hampel et al. (1986) p.313.

5.5.2 Ο αλγόριθμος των Μ-εκτιμητών

Συνήθως η εξίσωση (5.5.1.1) δεν είναι δυνατό να λυθεί άμεσα, οδηγώντας σε κλειστό τύπο για τον Μ-εκτιμητή. Για αυτό ακριβώς το λόγο χρησιμοποιούνται επαναληπτικές μέθοδοι. Η κυριότερη μέθοδος υπολογισμού Μ-εκτιμητών βασίζεται στον αλγόριθμο Newton-Raphson. Ακολούθως παρουσιάζεται ο αλγόριθμος, που εισήχθηκε από τον Huber (1973), υπολογίζοντας ταυτόχρονα τη θέση και την κλίμακα των Μ-εκτιμητών, ο οποίος λειτουργεί ως εξής:

1) Παίρνουμε ως αρχικές τιμές $\beta^{(0)}$ τον εκτιμητή ελαχίστων τετραγώνων και $\sigma^0 = \text{med}(|r_i|)$.
2) Υπολογίζουμε τα σφάλματα
$$r_i^m = y_i - f_i(\beta^m).$$
3) Υπολογίζουμε μια νέα τιμή για το σ μέσω της σχέσης:
$$(\sigma^{(m+1)})^2 = \frac{1}{(n-p-1)\Lambda}\sum \psi(\frac{r_i}{\sigma^{(m)}})^2(\sigma^{(m)})^2,$$
όπου Λ είναι ο παράγοντας διόρθωσης για την μεροληψία.
4) Ανθεκτικοποιούμε τα σφάλματα:
$$r_i^* = \psi(\frac{r_i}{\sigma^{(m+1)}})\sigma^{m+1}.$$
5) Υπολογίζουμε τις μερικές παραγώγους
$$x_{ik} = \frac{\vartheta}{\vartheta\beta} f_i(\beta),$$
με $x_{i0} = 1$ και $y_{i1} = x_i$
6) Λύνουμε για $\hat{\tau}$
$$\mathbf{X}^{\mathrm{T}}\mathbf{X}_{\hat{\tau}} = \mathbf{X}^{\mathrm{T}} r^*.$$
7) Θέτουμε
$$\beta^{(m+1)} = \beta^{(m)} + q\hat{\tau},$$
όπου $0<q<2$ είναι ένας αυθαίρετος σταθερός παράγοντας

8) Διακόπτουμε την επανάληψη εάν οι εκτιμητές αλλάξουν την τυπική τους απόκλιση λιγότερο από ε φορές. Διαφορετικά θέσε m :=m+1 και πήγαινε στο βήμα 2.

5.5.3 Γενικευμένοι εκτιμητές (ΓΕ) - GM-estimators

Έχοντας σαν στόχο τη βελτίωση της ανθεκτικότητας αναφορικά με τις έκτοπες τιμές στον άξονα των x_i, εισάγονται οι γενικευμένοι Μ-εκτιμητές [βλέπε Maronna et al. (1979)] οι οποίοι ικανοποιούν τη σχέση:

$$\sum_{i=1}^{n} w(x_i)\psi(r_i / \hat{\sigma})x_i = 0$$

ή τη σχέση

$$\sum_{i=1}^{n} w(x_i)\psi(r_i /(w(x_i)\hat{\sigma}))x_i = 0,$$

όπου w_i είναι η συνάρτηση βάρους. Οι γενικευμένοι εκτιμητές είναι γνωστοί και ως φραγμένης επίδρασης εκτιμητές. Το σημείο κατάρρευσης των γενικευμένων εκτιμητών δεν μπορεί να είναι καλύτερο από μία συγκεκριμένη τιμή που μειώνεται σαν συνάρτηση του p, όπου p είναι ο αριθμός των συντελεστών συσχέτισης. Αυτό όμως δεν μπορεί να θεωρηθεί καθόλου αποτελεσματικό, δεδομένου ότι συνεπάγεται ότι το σημείο κατάρρευσης μειώνεται καθώς αυξάνεται η διάσταση, με αποτέλεσμα να υπάρχουν περισσότερες πιθανότητες εμφάνισης έκτοπων παρατηρήσεων.

5.5.4 S-εκτιμητές

Οι S-εκτιμητές παλινδρόμησης που εισήχθηκαν από τους Rousseeuw & Yohai (1984) είναι συνεπείς για την αληθή παράμετρο β και ασυμπτωτικά κανονικοί όταν η κατανομή των σφαλμάτων είναι συμμετρική γύρω από το μηδέν. Προκύπτει ότι οι S-εκτιμητές έχουν ουσιαστικά την ίδια ασυμπτωτική συμπεριφορά με τους Μ-εκτιμητές παλινδρόμησης και μπορούν επιπλέον να πετύχουν ένα υψηλό σημείο κατάρρευσης. Όμως το πρόβλημα που εντοπίζεται στους συγκεκριμένους εκτιμητές είναι ότι δεν μπορούν να πετύχουν ταυτόχρονα υψηλό σημείο κατάρρευσης σε συνδυασμό με υψηλή αποτελεσματικότητα. Οι S-εκτιμητές είναι ίσης μεταβολής ως προς τη θέση,

ως προς την κλίμακα και ως προς τη θέση και την κλίμακα με τιμή σύγκλισης $n^{-1/2}$. Ορίζονται από την ελαχιστοποίηση της διαφοράς των σφαλμάτων:

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

με τελική εκτίμηση κλίμακας

$$\widehat{\sigma} = s(r_1(\widehat{\beta}),...,r_n(\widehat{\beta})).$$

Η διασπορά $s(r_1(\widehat{\beta}),...,r_n(\widehat{\beta}))$ ορίζεται ως τη λύση του:

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

όπου το K ισούται με $E_\Phi[\rho]$ και Φ είναι η τυπική κανονική κατανομή.

Η συνάρτηση ρ πρέπει να ικανοποιεί τις ακόλουθες συνθήκες:

1. Η ρ είναι συμμετρική και συνεχώς παραγωγίσιμη, και ρ(0)=0.
2. Υπάρχει c>0 τέτοιο ώστε η ρ να αυξάνεται στο διάστημα [0,c] και να παραμένει σταθερή στο διάστημα $[c,\infty]$.

Το σημείο κατάρρευσης των S-εκτιμητών δίνεται από τον ακόλουθο τύπο:

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

το οποίο μπορεί να φτάσει το 50% εάν γίνει μια κατάλληλη επιλογή σταθερών.

Η ασυμπτωτική διασπορά των πολυμεταβλητών S-εκτιμητών στο μοντέλο του Gauss με $\varepsilon_i \sim N(0,\sigma^2)$ επιτυγχάνεται από:

$$V(\psi,\Phi) = \frac{E[(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}]\sigma^2}{e},$$

όπου e είναι η ασυμπτωτική αποτελεσματικότητα (efficiency) που ορίζεται από

$$e = \frac{\int(\psi' d\Phi)^2}{\int \psi^2 d\Phi}$$

και ψ είναι η παράγωγος του ρ.

5.5.5 Ελάχιστα Διάμεσα Τετράγωνα (ΕΔΤ) – Least Median Squares

Ένα πιο ολοκληρωμένο όνομα για την μέθοδο ελαχίστων τετραγώνων θα ήταν το ελάχιστο άθροισμα τετραγώνων. Όμως προφανώς δεν ήταν λίγοι αυτοί που αντέδρασαν στην αφαίρεση της λέξης άθροισμα, θεωρώντας ότι η πρόσθεση των n

θετικών αριθμών θα ήταν η μόνη συνετή πράξη. Ίσως σαν συνέπεια του ιστορικού του ονόματος, μερικοί ερευνητές προσπάθησαν να κάνουν αυτόν τον εκτιμητή ανθεκτικό αντικαθιστώντας το άθροισμα με τη διάμεσο η οποία είναι πολύ ανθεκτική. Έτσι ο Rousseeuw (1984) πρότεινε τον εκτιμητή διαμέσων τετραγώνων που δίνεται από τη σχέση:

Χάρη στο υψηλό σημείο κατάρρευσης, ο εκτιμητής διαμέσων τετραγώνων μπορεί να χειριστεί μερικές έκτοπες παρατηρήσεις την ίδια στιγμή (σχεδόν πάνω από τις n/2 από αυτές, παρά το γεγονός ότι αυτό σπάνια εφαρμόζεται στην πράξη). Αυτή η αντίσταση είναι ανεξάρτητη από το p, τον αριθμό των επεξηγηματικών μεταβλητών και ως εκ τούτου ο εκτιμητής ελαχίστων τετραγώνων είναι ένα αξιόπιστο αναλυτικό εργαλείο δεδομένων που μπορεί να χρησιμοποιηθεί σε πολυμεταβλητές περιπτώσεις. Η βασική αρχή του εκτιμητή είναι να προσαρμόσει την πλειοψηφία των δεδομένων, μετά από την οποία οι έκτοπες παρατηρήσεις μπορούν να αναγνωριστούν σαν τα σημεία που βρίσκονται μακριά από την ανθεκτική προσαρμογή, συγκεκριμένα δηλαδή στα σημεία με μεγάλα θετικά ή αρνητικά σφάλματα. Παρά το γεγονός ότι η συνάρτηση επίδρασης του εκτιμητή ελαχίστων διαμέσων τετραγώνων δεν ορίζεται πολύ καλά, είναι πολύ πιθανό να πάρουμε κάποια ιδέα από τις ανθεκτικές του ιδιότητες, κατασκευάζοντας τις καμπύλες ευαισθησίας. Προκύπτει ότι αυτός ο εκτιμητής είναι πολύ ανθεκτικός αναφορικά με τις έκτοπες τιμές τόσο στον άξονα των x, όσο και στον άξονα των y.

Μερικές από τις ιδιότητες του εκτιμητή ελαχίστων διαμέσων τετραγώνων είναι οι παρακάτω:

1. Υπάρχει πάντοτε μία λύση για τον εκτιμητή των ελαχίστων διαμέσων τετραγώνων.
2. Ο εκτιμητής ελαχίστων διαμέσων τετραγώνων είναι ίσης μεταβολής ως προς την παλινδρόμηση, ίσης μεταβολής ως προς την κλίμακα και ως προς τη θέση και την κλίμακα.
3. Αν ο αριθμός των συντελεστών παλινδρόμησης είναι μεγαλύτερος της μονάδας τότε το σημείο κατάρρευσης του ΕΔΤ είναι 50% και ορίζεται ως $E_n^*(T,Z)=([n/2]-p+2)/n$.

Το μεγαλύτερο μειονέκτημα της μεθόδου ελαχίστων διαμέσων τετραγώνων είναι η έλλειψη αποτελεσματικότητας όταν τα σφάλματα είναι κανονικά κατανεμημένα. Η

τιμή σύγκλισης της μεθόδου είναι μόνο $n^{-1/3}$. Κατά συνέπεια προκύπτει το συμπέρασμα ότι ο εκτιμητής ελαχίστων διαμέσων τετραγώνων δεν είναι ασυμπτωτικά κανονικός.

Η συνάρτηση επίδρασης της διαμέσου είναι:

$$IF(y;median,\Phi) = sgn(y)/(2\Phi(0)),$$

από όπου προκύπτει ότι

$$V(median,\Phi) = (2\Phi(0))^{-2} = \pi/2 \approx 1{,}571$$

Ένας τρόπος για να βελτιώσουμε την αργή τιμή σύγκλισης του εκτιμητή έγκειται στη χρήση μιας διαφορετικής αντικειμενικής συνάρτησης. Αντί να προσθέσουμε όλα τα τετραγωνικά σφάλματα όπως στη μέθοδο ελαχίστων τετραγώνων, μπορούμε να επικεντρώσουμε το ενδιαφέρον μας στο περικεκομμένο άθροισμα τετραγώνων.

5.5.6 Ελάχιστα Περικεκομμένα Τετράγωνα (ΕΠΤ) – Least Trimmed Squares

Ο εκτιμητής ελαχίστων διαμέσων τετραγώνων συμπεριφέρεται φτωχά από την άποψη ότι δεν είναι ασυμπτωτικά αποτελεσματικός. Αυτό το πρόβλημα μπορεί να ξεπεραστεί χρησιμοποιώντας τον εκτιμητή ελαχίστων περικεκομμένων τετραγώνων. Αντί να προσθέσουμε όλα τα τετραγωνικά σφάλματα, δίνουμε ιδιαίτερη σημασία στα περικεκομμένα αθροίσματα τετραγώνων. Ο συγκεκριμένος εκτιμητής προτάθηκε από τον Rousseeuw (1984) και δίνεται από τη σχέση:

$$\sum_{i=1}^{h} (r^2)i : n, \qquad (5.10.1)$$

όπου $(r)^2_{1:n} \leq \ldots \leq (r)^2_{n::n}$ είναι τα διατεταγμένα τετράγωνα των σφαλμάτων και $h = ½ n+1$ αν το n είναι άρτιος. Η μέθοδος αυτή προσεγγίζει κατά πολύ τη μέθοδο των ελαχίστων τετραγώνων με τη μόνη διαφορά ότι τα μεγαλύτερα τετραγωνικά σφάλματα δεν χρησιμοποιούνται στην άθροιση.

Εύκολα προκύπτει ότι η συγκεκριμένη μέθοδος είναι πολύ περισσότερο ανθεκτική, δεδομένου ότι είναι αδύνατο μια απομονωμένη τιμή να ασκήσει μεγάλη επιρροή στην προσαρμογή. Συγκεκριμένα ο εκτιμητής ελαχίστων περικεκομμένων τετραγώνων μπορεί να αντισταθεί στην επίδραση έκτοπων παρατηρήσεων σε ένα ποσοστό της τάξης του 50%. Όμως όταν υπάρχουν στο δείγμα περισσότερες έκτοπες

παρατηρήσεις απ' ότι περικεκομμένες, τότε η συγκεκριμένη μέθοδος δεν είναι αρκετά αποτελεσματική.

Δεδομένου ότι στη σχέση (5.10.1) δεν υπολογίζονται οι μεγάλες τιμές των τετραγωνικών σφαλμάτων, καταλήγουμε στο συμπέρασμα ότι στη συγκεκριμένη προσαρμογή με βάση την μέθοδο ελαχίστων περικεκομμένων τετραγώνων δεν υπάρχουν έκτοπες παρατηρήσεις. Όταν χρησιμοποιούμε την παλινδρόμηση ελαχίστων περικεκομμένων τετραγώνων, το σ μπορεί να εκτιμηθεί ως εξής:

$$\hat{\sigma} = c_{h,n}\sqrt{\frac{1}{h}\sum_{i=1}^{h}(r^2)_{i:n}},$$

όπου r_i είναι τα σφάλματα από την προσαρμογή ελαχίστων περικεκομμένων τετραγώνων και η σταθερά $c_{h,n}$ κάνει τον $\hat{\sigma}$ συνεπή και αμερόληπτο στις κατανομές σφάλματος του Gauss. Ο εκτιμητής κλίμακας $\hat{\sigma}$ των ελαχίστων περικεκομμένων τετραγώνων είναι από μόνος του πολύ ανθεκτικός, με αποτέλεσμα να μπορούμε να αναγνωρίζουμε τις έκτοπες παρατηρήσεις από τα τυποποιημένα σφάλματα των ελαχίστων περικεκομμένων τετραγώνων $r_i / \hat{\sigma}$.

Η συνάρτηση επίδρασης του εκτιμητή ελαχίστων περικεκομμένων τετραγώνων σε μία συμμετρική κατανομή Φ δίνεται από τον τύπο:

$$\text{IF(y;LTS,}\Phi) = \frac{y1_{[-q,q]}(y)}{2\Phi(q) - 1 - 2q\Phi(q)}.$$

Εναλλακτικά μπορεί να γραφεί ως εξής:

$$\text{IF(y;LTS,}\Phi) = \begin{cases} 14{,}021y, \alpha\nu |y| \le q \\ 0, \delta\iota\alpha\phi\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

όπου $q = \Phi^{-1}(0{,}75) = 0{,}6745$.

Μερικές από τις ιδιότητες του εκτιμητή ελαχίστων διαμέσων τετραγώνων είναι οι παρακάτω:

1. Υπάρχει πάντοτε μία λύση για τον εκτιμητή των ελαχίστων περικεκομμένων τετραγώνων.
2. Ο εκτιμητής ελαχίστων περικεκομμένων τετραγώνων είναι ίσης μεταβολής ως προς την παλινδρόμηση, ίσης μεταβολής ως προς την κλίμακα και ως προς τη θέση και την κλίμακα.
3. Το σημείο κατάρρευσης ισούται με $E_n^*(T,Z)=([(n-p)/2+1)/n$. Κατά συνέπεια υπό τις υπάρχουσες συνθήκες ο εκτιμητής ελαχίστων περικεκομμένων

τετραγώνων διατηρεί το ίδιο σημείο κατάρρευσης, δηλαδή της τάξης του 50% με τον εκτιμητή διαμέσων τετραγώνων.

4. Ο εκτιμητής περικεκομμένων τετραγώνων έχει την ίδια ασυμπτωτική αποτελεσματικότητα στην κανονική κατανομή με τον Μ-εκτιμητή που ορίζεται ως εξής:

$$\psi(t)=\begin{bmatrix} t, & \alpha\nu\ |t|<\Phi^{-1}(1-a/2) \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{bmatrix}.$$

Το βασικό μειονέκτημα της μεθόδου ελαχίστων περικεκομμένων τετραγώνων είναι ότι η αντικειμενική συνάρτηση απαιτεί ταξινόμηση των τετραγωνικών σφαλμάτων, η οποία χρειάζεται πολλές περισσότερες επαναλήψεις συγκριτικά με τη μέθοδο ελαχίστων διαμέσων τετραγώνων.

5.5.7 ΜΜ-εκτιμητές

Ο Yohai (1985) εισήγαγε τους ΜΜ-εκτιμητές οι οποίοι θεωρούνται εκτιμητές με υψηλό σημείο κατάρρευσης και υψηλή αποτελεσματικότητα, όπου η αρχική εκτίμηση επιτυγχάνεται μέσω ενός S-εκτιμητή, ο οποίος στη συνέχεια βελτιώνεται μέσω ενός Μ-εκτιμητή. Συνεπώς οι ΜΜ-εκτιμητές προέρχονται από τους συντελεστές καθώς και από τη σταθερή κλίμακα που παίρνουμε από τον S-εκτιμητή.

Στην πραγματικότητα ο Yohai απέδειξε ότι οι ΜΜ-εκτιμητές έχουν σημείο κατάρρευσης 50% στο πρώτο στάδιο και ότι επίσης κατέχουν ακριβώς την ίδια ιδιότητα προσαρμογής. Επιπλέον, απέδειξε ότι οι ΜΜ-εκτιμητές είναι υψηλά αποτελεσματικοί όταν τα σφάλματα κατανέμονται κανονικά.

Στην ΜΜ-εκτίμηση, ένας ανθεκτικός Μ-εκτιμητής δίνεται από τη σχέση:

$$\sum_{i=1}^{n}\rho\left(\frac{y_i-\mathbf{x_i^T}\beta}{\hat{s}}\right),$$

όπου $\hat{s}$ είναι μία ανθεκτική εκτίμηση κλίμακας των σφαλμάτων και ρ είναι μια πραγματική συνάρτηση η οποία ικανοποιεί τις ακόλουθες υποθέσεις:

(i) $\rho(0)=0$

(ii) $\rho(-u)=\rho(u)$, δηλαδή η ρ είναι άρτια

(iii) $0\le u\le v$ τότε $\rho(u)\le\rho(v)$, δηλαδή η ρ είναι αύξουσα στο R^+ και φθίνουσα στο R^-

(iv) η ρ είναι συνεχής

(v) ορίζουμε ότι αν α=supρ(u), τότε 0<α<∞

(vi) αν ρ(u)<α και 0≤u<v τότε ρ(u)<ρ(v).

Μία εναλλακτική επιλογή για τη συνάρτηση εκτίμησης είναι:

$$\sum_{i=1}^{n} \quad \psi\left(\frac{y_i - \quad \boldsymbol{\beta}}{\widehat{s}}\right),$$

όπου ψ είναι μια μη μονότονη συνάρτηση.

Ορίζουμε ως F_ε την κατανομή των σφαλμάτων ε_i του μοντέλου παλινδρόμησης. Θεωρούμε δύο συναρτήσεις ρ_o, ρ_1: $R \rightarrow R$ τέτοιες ώστε:

$$E_{F\varepsilon}(\rho_o(\varepsilon))=0{,}5 \quad \rho_1(u) \leq \rho_o(u) \quad \forall u \in R, \qquad \sup_{x \in R} \rho_0(x) = \sup_{x \in R} \rho_1(x).$$

Ο MM-εκτιμητής που ορίζεται ως $\mathbf{B_{Tj}}=\mathbf{T^1}$, ορίζεται σε τρία βήματα ως ακολούθως:

1. Υπολογίζουμε έναν αρχικό εκτιμητή $\mathbf{T^0}$ του συντελεστή παλινδρόμησης **β**. Αυτός ο αρχικός εκτιμητής παλινδρόμησης είναι ανθεκτικά συνεπής με ένα υψηλό σημείο κατάρρευσης αλλά δεν είναι υποχρεωτικά αποτελεσματικός. Ο S-εκτιμητής θα χρησιμοποιηθεί σαν ένα αρχικό κομμάτι της συνολικής υπολογιστικής στρατηγικής του MM-εκτιμητή. Ο S-εκτιμητής έχει τις ίδιες ασυμπτωτικές ιδιότητες με τον MM-εκτιμητή και επίσης μπορεί να χειριστεί το 50% των έκτοπων παρατηρήσεων που παρουσιάζονται στα δεδομένα. Ο εκτιμητής κλίμακας μπορεί να επιτευχθεί μέσω του ακόλουθου προβλήματος ελαχιστοποίηση διασποράς:

$$\widehat{\boldsymbol{\beta}}_{\mathbf{j}} = \arg\min_{\beta} S\{r_1(\beta),...,r_n(\beta)\}$$

που υπόκειται σε

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

με $r_i(\beta) = y_i - \beta$ και $K = E_\Phi(\rho)$ το οποίο επιβεβαιώνει την συνέπεια του S στην κανονική κατανομή Φ.

2. Υπολογίζουμε την M-κλίμακα των σφαλμάτων $r_i(\mathbf{T^0})$, S(β) χρησιμοποιώντας τη συνάρτηση ρ_0. Για παράδειγμα, σε αυτό το δεύτερο βήμα υπολογίζεται ένας M-εκτιμητής σφαλμάτων κλίμακας χρησιμοποιώντας σφάλματα που βασίζονται στην αρχική εκτίμηση.

3. Ορίζουμε $\mathbf{T^1}$ ως μία λύση της εξίσωσης

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

όπου ψ = ρ_1' που επίσης ικανοποιεί τη σχέση

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

Αυτοί οι εκτιμητές έχουν υψηλό σημείο κατάρρευσης και υψηλή αποτελεσματικότητα.

Κάτω από τις συνήθεις συνθήκες συμπεριλαμβάνοντας ότι η κατανομή των σφαλμάτων είναι συμμετρική, οι εκτιμητές είναι συνεπείς και ασυμπτωτικά κανονικοί με διασπορά που εξαρτάται μόνο από την περιοριστική τιμή του εκτιμητή κλίμακας S_n .

Οι ΜΜ-εκτιμητές έχουν ταυτόχρονα τις ακόλουθες ιδιότητες:

- υψηλής αποτελεσματικότητας όταν τα σφάλματα έχουν κανονική κατανομή
- το σημείο κατάρρευσης είναι 0,5. Από την ανθεκτική ΜΜ μέθοδο παλινδρόμησης προκύπτει ένα μοντέλο το οποίο είναι στη δομή σχεδόν ίδιο με το τυπικό μοντέλο γραμμικής παλινδρόμησης.

Από τη στιγμή που ο ΜΜ- εκτιμητής είναι ένας Μ-εκτιμητής η συνάρτηση ευαισθησίας του δίνεται από τη σχέση [Yohai (1987)]:

$$\mathrm{IF}(\mathbf{T},\mathrm{F}, \quad ; y) = \psi(r) \;\; S^2[E_F(\psi^T(r/S) \quad (F)]^{-1},$$

με r = y- $\mathbf{x^T\beta}$, $\mathbf{\Sigma}$(F) = $E_F(\mathbf{xx^T})$ και S είναι ο εκτιμητής κλίμακας. Για να φράξουμε τη συνάρτηση επίδρασης του ΜΜ-εκτιμητή είναι πιο σημαντικό να θεωρήσουμε την περίπτωση ενός μικρού αλλά θετικού μεγέθους αλλοίωσης.

Ο ασυμπτωτικός πίνακας συνδιασποράς του ΜΜ-εκτιμητή επιτυγχάνεται ως εξής:

$$\begin{aligned}
V(\mathbf{T},\mathrm{F}) &= \mathrm{E}[\mathrm{IF}(\mathbf{T},\mathrm{F},\mathbf{x^T};\mathrm{y})\, IF^T(\mathbf{T},\mathrm{F}, \quad ;\mathrm{y})] \\
&= \mathrm{E}[\{\psi(\mathrm{r})\mathbf{x}\, \mathrm{S}^2\, [\mathrm{E_F}(\psi^T(r/S))\Sigma(F)]^{-1}\{\psi(\mathrm{r})\mathbf{x}\, \mathrm{S}^2\, [\mathrm{E_F}(\psi^T(r/S))\Sigma(F)]^{-1}\}^T] \\
&= \mathrm{S}^2\, [\mathrm{E_F}(\psi^2(r/S)/E_F^2(\psi^T(r/S))][\Sigma(F)^{-1}].
\end{aligned}$$

5.6 Εκτιμητές υψηλού σημείου κατάρρευσης

Υπάρχουν αρκετοί εκτιμητές που στοχεύουν στο να έχουν ένα υψηλό σημείο κατάρρευσης, δηλαδή να βρίσκονται κοντά στο άνω φράγμα. Όμως δεν είναι όλοι οι εκτιμητές ανάμεσα σε αυτούς αρκετά αποτελεσματικοί έτσι ώστε να πετύχουν αυτό το υψηλό ΣΚ, εξαιτίας της ευαισθησίας που δείχνουν σε ένα συγκεκριμένο είδος παρατηρήσεων, δηλαδή στις έκτοπες παρατηρήσεις. Ο βαθμός της ανθεκτικότητας

ενός εκτιμητή όταν υπάρχουν έκτοπες παρατηρήσεις μπορεί να μετρηθεί όπως έχει ήδη αναφερθεί, με τη βοήθεια του ΣΚ το οποίο αρχικά προτάθηκε από τον Hampel (1971). Όμως το πρόβλημα που υπάρχει είναι ότι πολλές από τις προτάσεις για ανθεκτική εκτίμηση στην παλινδρόμηση δεν έχουν υψηλό σημείο κατάρρευσης.

Οι Μ-εκτιμητές με μονότονη ψ-συνάρτηση που προτάθηκαν από τον Huber (1973) έχουν σημείο κατάρρευσης μηδέν. Μια απλή ανθεκτική μέθοδος παλινδρόμησης είναι ο Μ-εκτιμητής ο οποίος μπορεί να είναι πολύ αποτελεσματικός στην αντιμετώπιση των έκτοπων παρατηρήσεων που παρουσιάζονται στα εκτιμημένα σφάλματα. Το πρόβλημα της μεθόδου αυτής είναι ότι μπορεί να μην είναι ευαίσθητες στην κατεύθυνση των ανεξάρτητων μεταβλητών. Εύκολα μπορεί να φανεί ότι οι Μ-εκτιμητές μπορούν να επηρεαστούν από ένα και μόνο σημείο μόχλευσης δεδομένου ότι το σημείο κατάρρευσής τους ισούται με $1/N$, όπου Ν είναι το μέγεθος του δείγματος.

Αυτός ο περιορισμός οδήγησε στην ανάπτυξη του γενικευμένου Μ-εκτιμητή ο οποίος δεν επηρεάζεται ούτε από τις έκτοπες παρατηρήσεις αλλά ούτε και από τα σημεία μόχλευσης. Σε αντίθεση με τους Μ-εκτιμητές η ιδιότητα ίσης μεταβολής ως προς την κλίμακα διατηρείται στους R-, και L-εκτιμητές. Το σημείο κατάρρευσης των Μ-, L-, R-εκτιμητών παλινδρόμησης είναι 0% εξαιτίας της ευπάθειας αναφορικά με τα αρνητικά σημεία μόχλευσης.

Τόσο οι Μ-εκτιμητές όσο και οι γενικευμένοι εκτιμητές μπορούν να υπολογιστούν από τα επαναλαμβανόμενα σταθμισμένα ελάχιστα τετράγωνα ή από τον αλγόριθμο του Newton-Raphson. Ο Maronna, ο Bustos & Yohai (1979) έδειξαν ότι οι GM-εκτιμητές έχουν ένα σημείο κατάρρευσης που τείνει στο μηδέν όταν το ο αριθμός των συντελεστών αυξάνει. Αυτοί οι εκτιμητές περιέχουν τη βέλτιστη φραγμένη επίδραση εκτιμητών που παρουσιάστηκε από τον Krasker (1980) και τον Krasker & Welsch (1982). Ο Rousseeuw (1984) πρότεινε τον εκτιμητή ΕΔΤ και τον εκτιμητή ΕΠΤ. Ο Rousseeuw & Yohai (1984) πρότειναν μια κλάση εκτιμητών που βασίζεται στην ελαχιστοποίηση του ανθεκτικού Μ-εκτιμητή του σφάλματος κλίμακας (S-εκτιμητές). Όμως όλοι αυτοί οι εκτιμητές είναι υψηλά αποτελεσματικοί όταν όλες οι παρατηρήσεις ικανοποιούν το μοντέλο παλινδρόμησης με κανονικά σφάλματα.

Ο Rousseeuw (1984) θέλοντας να συνδυάσει υψηλό σημείο κατάρρευσης με υψηλή αποτελεσματικότητα, πρότεινε να χρησιμοποιηθεί ένας εκτιμητής με υψηλό σημείο κατάρρευσης που ακολουθείται από τον Μ-εκτιμητή ενός βήματος ή από τα

σταθμισμένα ελάχιστα τετράγωνα ενός βήματος. Η συγκεκριμένη διαδικασία εξασφαλίζει ένα υψηλό σημείο κατάρρευσης και βελτιώνει την αποτελεσματικότητα του αρχικού εκτιμητή. Όμως το αρχικό σημείο κατάρρευσης αυτής της διαδικασίας δεν είναι γνωστό και ως εκ τούτου δεν υπάρχει εγγύηση ότι το ΣΚ του αρχικού εκτιμητή διατηρείται αναλλοίωτο. Επιπλέον, ο Μ-εκτιμητής ενός βήματος, δεν έχει την ίδια ασυμπτωτική αποτελεσματικότητα όπως ο πλήρως επαναλαμβανόμενος εκτιμητής. Στην πραγματικότητα η ασυμπτωτική του αποτελεσματικότητα εξαρτάται από τον αρχικό εκτιμητή και είναι αρκετά δύσκολο να υπολογιστεί.

Δύο εκτιμητές με πραγματικά υψηλό σημείο κατάρρευσης είναι ο εκτιμητής ΕΔΤ και ο εκτιμητής ΕΠΤ, ο οποίος προτάθηκε για να αντιμετωπιστεί η χαμηλή ασυμπτωτική αποτελεσματικότητα του εκτιμητή ελαχίστων διαμέσων τετραγώνων. Συγκεκριμένα στην πολλαπλή παλινδρόμηση, τόσο ο εκτιμητής ΕΔΤ όσο και ο εκτιμητής ΕΠΤ ήταν οι πρώτες μέθοδοι ίσης μεταβολής που πέτυχαν τιμή κατάρρευσης της τάξης του 50%.

Κ Ε Φ Α Λ Α Ι Ο 6

6.1 Απλή παλινδρόμηση

Στη συνέχεια παρατίθεται ένα παράδειγμα τα δεδομένα του οποίου (Πίνακας 6.1.1) προέρχονται από το βιβλίο του Rousseeuw & Yohai (1984), μέσα από το οποίο αποδεικνύεται η επιτακτική ανάγκη χρήσης ανθεκτικής παλινδρόμησης. Στο συγκεκριμένο παράδειγμα στόχος μας είναι να δείξουμε την επίδραση των έκτοπων παρατηρήσεων.

Παράδειγμα $1^{ο}$: Η εξαρτημένη μεταβλητή αντιστοιχεί στην περιεκτικότητα οξέως που προσδιορίζεται από τον καθορισμό πυκνότητας μίγματος (y_i) και η ανεξάρτητη μεταβλητή είναι η οργανική ποσότητα οξέως που προσδιορίζεται από την απόσταξη (x_i).

Πίνακας 6.1.1
Δεδομένα παραδείγματος

Παρατήρηση	Απόσταξη	Καθορισμός πυκνότητας
(i)	(x_i)	(y_i)
1	123	76
2	109	70
3	62	55
4	104	71
5	57	55
6	37	48
7	44	50
8	100	66
9	16	41
10	28	43
11	138	82
12	105	68
13	159	88
14	75	58
15	88	64
16	164	88
17	169	89
18	167	88
19	149	84
20	167	88

Εφαρμόζοντας στα δεδομένα μας τόσο την μέθοδο ελαχίστων τετραγώνων όσο και τις ανθεκτικές μεθόδους, διαπιστώνουμε κατά πόσο η ευθεία γραμμικής παλινδρόμησης επηρεάζεται από:

- καμία έκτοπη παρατήρηση
- ένα αρνητικό σημείο μόχλευσης
- ένα θετικό σημείο μόχλευσης
- μία κάθετη έκτοπη παρατήρηση

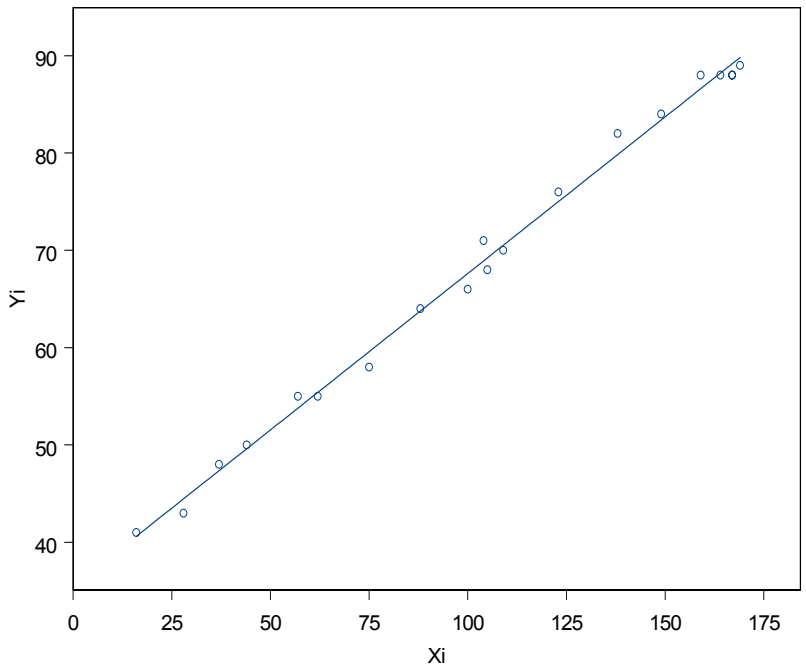
καθώς επίσης και από όλους τους δυνατούς συνδυασμούς των ΕΠ, δηλαδή

- ένα αρνητικό σημείο μόχλευσης & μία κάθετη έκτοπη παρατήρηση
- μία κάθετη έκτοπη παρατήρηση & ένα θετικό σημείο μόχλευσης

ένα αρνητικό σημείο μόχλευσης & ένα θετικό σημείο μόχλευσης.

1. Χωρίς έκτοπη παρατήρηση

Διάγραμμα 6.1.1
Διάγραμμα διασποράς χωρίς ΕΠ



Στον πίνακα 6.1.2 παρουσιάζονται συνοπτικά οι συντελεστές παλινδρόμησης τόσο με τη μέθοδο ελαχίστων τετραγώνων όσο και με την χρήση μεθόδων ανθεκτικής παλινδρόμησης (χωρίς έκτοπη παρατήρηση):

Πίνακας 6.1.2
Συντελεστές παλινδρόμησης χωρίς ΕΠ

	Συντελεστές παλινδρόμησης	
Εκτιμητές	$\hat{\beta}_0$	$\hat{\beta}_1$
ΕΤ (LS)	35,458	0,321
Μ-εκτιμητές	35,458	0,321
ΕΠΤ (LTS)	35,473	0,319
ΕΔΤ (LMS)	35,638	0,315
ΜΜ-εκτιμητές	35,458	0,322

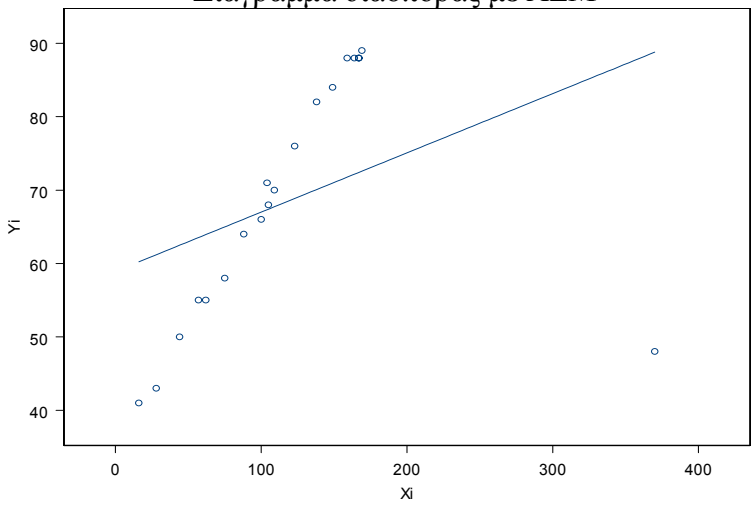
❖ **Συμπεράσματα**

Από το διάγραμμα διασποράς (scatterplot) (Διάγραμμα 6.1.1) εύκολα διαπιστώνουμε ότι υπάρχει μία ισχυρή στατιστική σχέση ανάμεσα στην εξαρτημένη και στην ανεξάρτητη μεταβλητή. Η υπόθεση ενός γραμμικού μοντέλου φαίνεται να είναι πολύ λογική. Παρατηρώντας προσεκτικά το διάγραμμα διαπιστώνουμε ότι δεν υπάρχουν έκτοπες παρατηρήσεις. Όπως αναμένεται σε αυτήν την περίπτωση μεταξύ των ανθεκτικών εκτιμητών και των εκτιμητών ελαχίστων τετραγώνων (Πίνακα 6.1.2) υπάρχουν μόνο πολύ μικρές διαφορές.

2. Αρνητικό σημείο μόχλευσης

Ας υποθέσουμε τώρα ότι η x-τιμή της έκτης παρατήρησης παίρνει την τιμή 370 αντί της τιμής 37. Αυτή η παρατήρηση παράγει ένα αρνητικό σημείο μόχλευσης. (Διάγραμμα 6.1.2)

Διάγραμμα 6.1.2
Διάγραμμα διασποράς με ΑΣΜ



Στον πίνακα 6.1.3 παρουσιάζονται συνοπτικά οι συντελεστές παλινδρόμησης τόσο με τη μέθοδο ελαχίστων τετραγώνων όσο και με την χρήση μεθόδων ανθεκτικής παλινδρόμησης (αρνητικό σημείο μόχλευσης):

Πίνακας 6.1.3

Συντελεστές παλινδρόμησης με ΑΣΜ

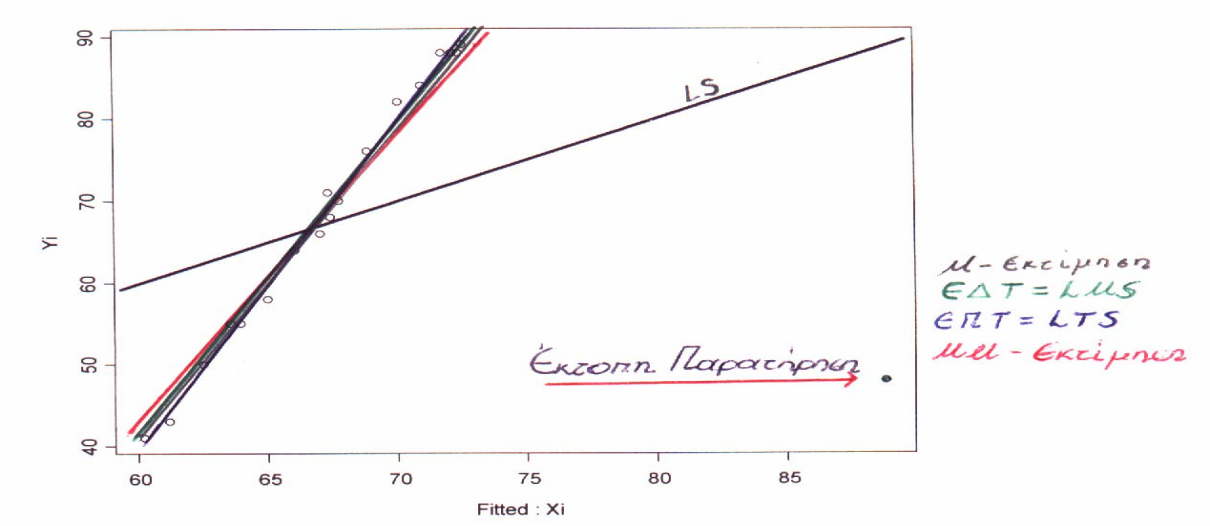
	Συντελεστές παλινδρόμησης	
Εκτιμητές	$\hat{\beta}_0$	$\hat{\beta}_1$
ΕΤ (LS)	58,939	0,081
ΜΜ-εκτιμητές	35,317	0,323
ΕΠΤ (LTS)	35,335	0,321
ΕΔΤ (LMS)	36,343	0,314
Μ-εκτιμητές	36,436	0,312

α. **<u>Συμπεράσματα</u>**

Από το διάγραμμα διασποράς (scatterplot) (Διάγραμμα 6.1.2) εύκολα διαπιστώνουμε ότι η ευθεία έχει επηρεαστεί σε μεγάλο βαθμό από μία και μόνο έκτοπη παρατήρηση με αποτέλεσμα η μελέτη μας να καταστρέφεται. Επιπλέον, η ευθεία παλινδρόμησης με βάση τους ανθεκτικούς εκτιμητές βρίσκεται πολύ κοντά στον εκτιμητή ελαχίστων τετραγώνων που εφαρμόστηκε στα αρχικά δεδομένα. Τα αποτελέσματα που προκύπτουν από το νέο αλλοιωμένο δείγμα (Πίνακας 6.1.3) είναι ότι οι συντελεστές παλινδρόμησης μεταβάλλονται στην περίπτωση των ελαχίστων τετραγώνων.

i. <u>Διάγραμμα διασποράς με τη χρήση μεθόδου ελαχίστων τετραγώνων και ανθεκτικών εκτιμητών</u>

Διάγραμμα 6.1.3



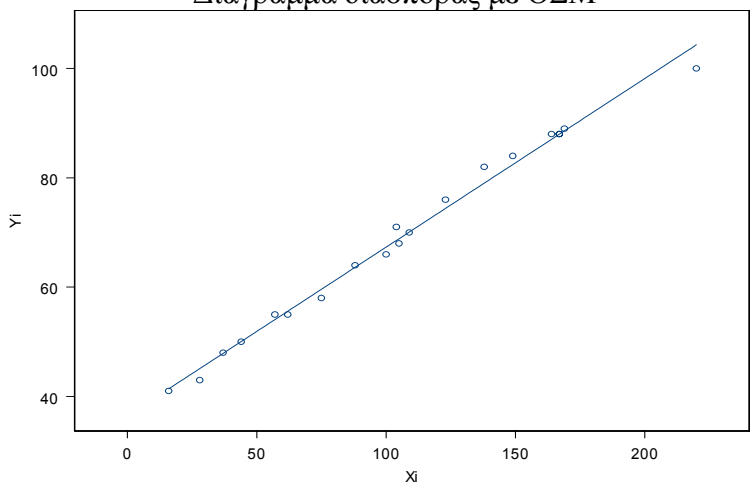
Από το διάγραμμα 6.1.3 διαπιστώνουμε ότι στο συγκεκριμένο παράδειγμα η ευθεία παλινδρόμησης ελαχίστων τετραγώνων μετατοπίζεται προς την κατεύθυνση του αρνητικού σημείου μόχλευσης. Σε αντίθεση οι ευθείες που προκύπτουν από τις ανθεκτικές μεθόδους παραμένουν αναλλοίωτες, βρίσκονται δηλαδή πολύ κοντά μεταξύ τους και σε μεγάλη απόσταση από την ευθεία ελαχίστων τετραγώνων παραμένοντας ανεπηρέαστες από το αρνητικό σημείο μόχλευσης. Πιο συγκεκριμένα, οι ανθεκτικές μέθοδοι δίνουν μια πολύ καλή προσαρμογή στην πλειοψηφία των δεδομένων.

3. <u>Θετικό σημείο μόχλευσης</u>

Ας υποθέσουμε τώρα ότι για παράδειγμα το ζεύγος (x , y)=(159 , 88) παίρνει την τιμή (220 , 100). Αυτή η παρατήρηση παράγει ένα θετικό σημείο μόχλευσης. (Διάγραμμα 6.1.4).

Διάγραμμα 6.1.4

Διάγραμμα διασποράς με ΘΣΜ



Πίνακας 6.1.4

Συντελεστές παλινδρόμησης με ΘΣΜ

	Συντελεστές παλινδρόμησης	
Εκτιμητές	$\hat{\beta}_0$	$\hat{\beta}_1$
ΕΤ (LS)	36,474	0.308
ΜΜ-εκτιμητές	35,565	0,319
ΕΠΤ (LTS)	35,586	0,318
ΕΔΤ (LMS)	35,638	0,315
Μ-εκτιμητές	35,958	0,314

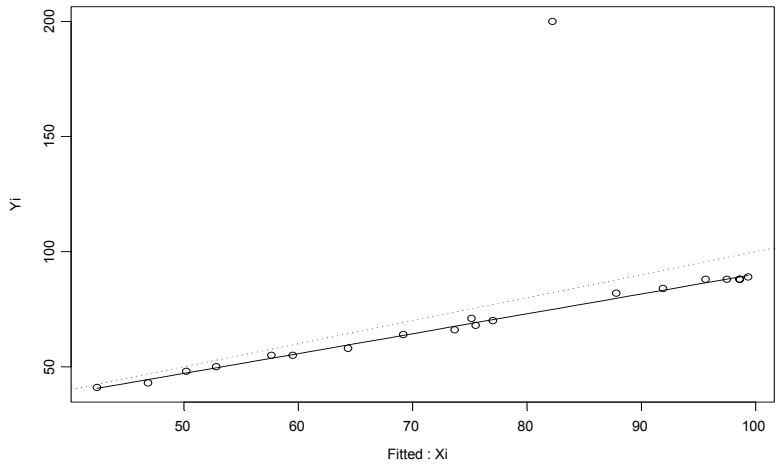
❖ **Συμπεράσματα**

Με βάση το διάγραμμα 6.1.4 παρατηρούμε ότι η ύπαρξη ενός θετικού σημείου μόχλευσης δεν προκαλεί καμία μεταβολή στην ευθεία παλινδρόμησης. Διαπιστώνουμε ότι τα αποτελέσματα που προκύπτουν από το νέο δείγμα (Πίνακας 6.1.4) δεν διαφέρουν σημαντικά μεταξύ τους. Συνεπώς καταλήγουμε στο συμπέρασμα ότι η ύπαρξη ενός θετικού σημείου μόχλευσης στο δείγμα μας αλλοιώνει ελάχιστα το αποτέλεσμα, δηλαδή η ευθεία παλινδρόμησης παραμένει αμετάβλητη.

4. Κάθετη έκτοπη παρατήρηση

Ας υποθέσουμε τώρα ότι για παράδειγμα η y-τιμή της πρώτης παρατήρησης παίρνει την τιμή 200 αντί της τιμής 76. Αυτό το σημείο παράγει μια κάθετη έκτοπη παρατήρηση. (Διάγραμμα 6.1.5)

Διάγραμμα 6.1.5
Διάγραμμα διασποράς με ΚΕΠ



Πίνακας 6.1.5
Συντελεστές παλινδρόμησης με ΚΕΠ

	Συντελεστές παλινδρόμησης	
Εκτιμητές	$\hat{\beta}_0$	$\hat{\beta}_1$
ΕΤ (LS)	36,459	0,372
ΜΜ-εκτιμητές	35,449	0,321
ΕΠΤ (LTS)	35,557	0,319
ΕΔΤ (LMS)	35,639	0,315
Μ-εκτιμητές	35,468	0,322

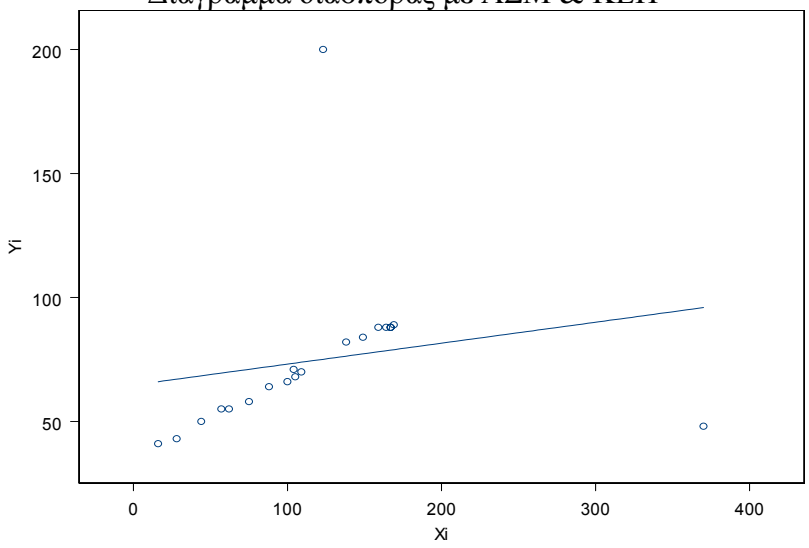
❖ **Συμπεράσματα**

Παρατηρούμε ότι τα αποτελέσματα που προκύπτουν από το νέο δείγμα (Πίνακας 6.1.5) δεν διαφέρουν σημαντικά μεταξύ τους. Συνεπώς καταλήγουμε στο συμπέρασμα ότι η ύπαρξη μιας κάθετης έκτοπης παρατήρησης στο δείγμα μας αλλοιώνει ελάχιστα το αποτέλεσμα.

5. Αρνητικό σημείο μόχλευσης - Κάθετη έκτοπη παρατήρηση

Ας υποθέσουμε τώρα ότι έχουμε δύο έκτοπες παρατηρήσεις. Για παράδειγμα η x-τιμή της έκτης παρατήρησης παίρνει την τιμή 370 αντί της τιμής 37 και η y-τιμή της πρώτης παρατήρησης παίρνει την τιμή 200 αντί της τιμής 76. Αυτές οι δύο παρατηρήσεις παράγουν ένα αρνητικό σημείο μόχλευσης και μία κάθετη έκτοπη παρατήρηση (Διάγραμμα 6.1.6).

Διάγραμμα 6.1.6
Διάγραμμα διασποράς με ΑΣΜ & ΚΕΠ



Πίνακας 6.1.6
Διάγραμμα διασποράς με ΑΣΜ & ΚΕΠ

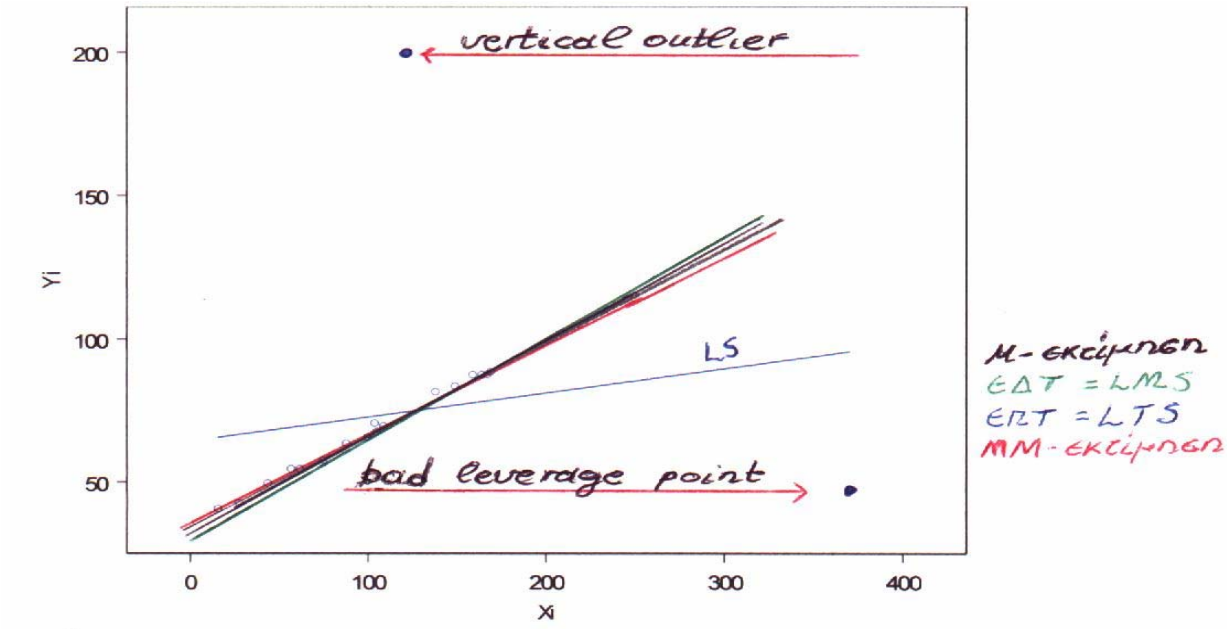
	Συντελεστές παλινδρόμησης	
Εκτιμητές	$\hat{\beta}_0$	$\hat{\beta}_1$
ΕΤ (LS)	64,695	0,084
ΜΜ-εκτιμητές	35,304	0,322
ΕΠΤ (LTS)	35,062	0,324
ΕΔΤ (LMS)	36,343	0,314
Μ-εκτιμητές	36,512	0,311

❖ **Συμπεράσματα**

Από το διάγραμμα 6.1.6 παρατηρούμε ότι η ευθεία έχει επηρεαστεί σε μεγάλο βαθμό από την κάθετη έκτοπη παρατήρηση και το αρνητικό σημείο μόχλευσης με αποτέλεσμα η ευθεία παλινδρόμησης να μετατοπίζεται προς την κατεύθυνση του αρνητικού σημείου μόχλευσης. Τα αποτελέσματα που προκύπτουν από το νέο αλλοιωμένο δείγμα (Πίνακας 6.1.6) είναι ότι οι συντελεστές παλινδρόμησης μεταβάλλονται στην περίπτωση των ελαχίστων τετραγώνων. Αντιθέτως, οι ευθείες που προκύπτουν από τις ανθεκτικές μεθόδους παραμένουν αναλλοίωτες.

i. Διάγραμμα διασποράς με τη χρήση μεθόδου ελαχίστων τετραγώνων και ανθεκτικών εκτιμητών

Διάγραμμα 6.1.7

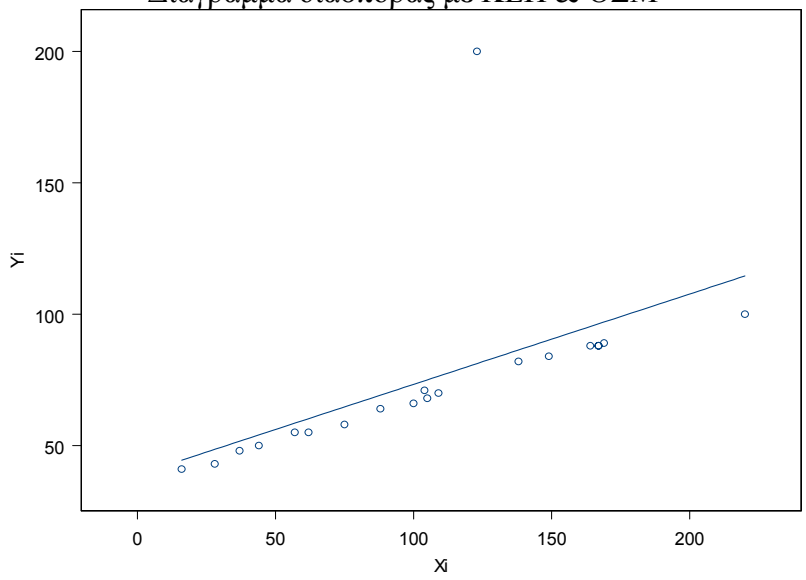


Από το διάγραμμα 6.1.7 διαπιστώνουμε ότι στο συγκεκριμένο παράδειγμα η ευθεία παλινδρόμησης ελαχίστων τετραγώνων μετατοπίζεται προς την κατεύθυνση του αρνητικού σημείου μόχλευσης. Αντιθέτως οι ευθείες των ανθεκτικών εκτιμητών βρίσκονται πολύ κοντά μεταξύ τους και σε μεγάλη απόσταση από την ευθεία ελαχίστων τετραγώνων παραμένοντας ανεπηρέαστες από το αρνητικό σημείο μόχλευσης.

6. Κάθετη έκτοπη παρατήρηση - Θετικό σημείο μόχλευσης

Ας υποθέσουμε τώρα ότι η y-τιμή της πρώτης παρατήρησης παίρνει την τιμή 200 αντί της τιμής 76 και το ζεύγος (x , y)=(159 , 88) παίρνει την τιμή (220 , 100). Αυτές οι δύο παρατηρήσεις παράγουν μία κάθετη έκτοπη παρατήρηση και ένα θετικό σημείο μόχλευσης (Διάγραμμα 6.1.8).

Διάγραμμα 6.1.8
Διάγραμμα διασποράς με ΚΕΠ & ΘΣΜ



Πίνακας 6.1.7
Συντελεστές παλινδρόμησης με ΚΕΠ & ΘΣΜ

	Συντελεστές παλινδρόμησης	
Εκτιμητές	$\hat{\beta}_0$	$\hat{\beta}_1$
ΕΤ (LS)	38,931	0,344
ΜΜ-εκτιμητές	35,562	0,319
ΕΠΤ (LTS)	35,621	0,319
ΕΔΤ (LMS)	35,638	0,315
Μ-εκτιμητές	35,969	0,314

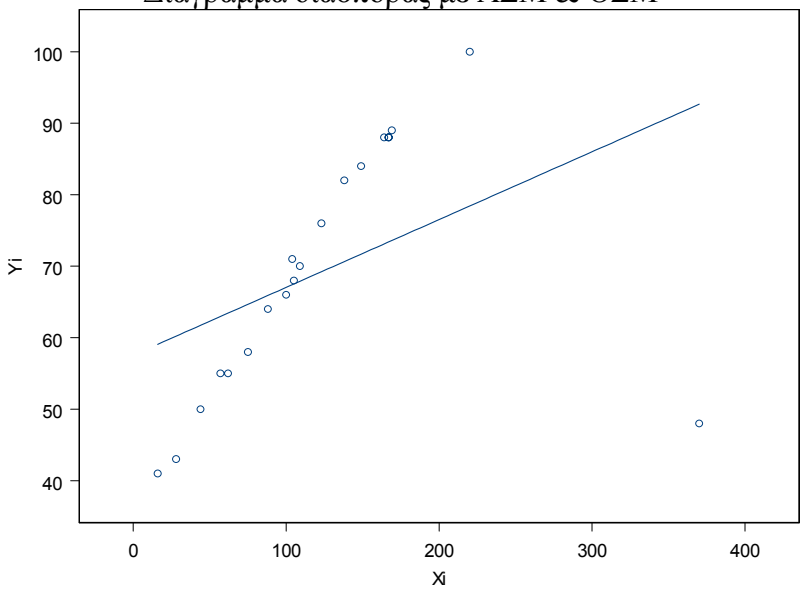
- **Συμπεράσματα**

Παρατηρούμε ότι τα αποτελέσματα που προκύπτουν από το νέο δείγμα (Πίνακας 6.1.7) δεν διαφέρουν σημαντικά μεταξύ τους. Συνεπώς καταλήγουμε στο συμπέρασμα ότι η ύπαρξη μιας κάθετης έκτοπης παρατήρησης και ενός θετικού σημείου μόχλευσης στο δείγμα μας στο συγκεκριμένο παράδειγμα αλλοιώνει ελάχιστα το αποτέλεσμα. Παρόλο αυτά όμως η εφαρμογή μεθόδων ανθεκτικής παλινδρόμησης κρίνεται απαραίτητη.

7. **<u>Αρνητικό σημείο μόχλευσης - Θετικό σημείο μόχλευσης</u>**

Ας υποθέσουμε τώρα ότι η x-τιμή της πρώτης παρατήρησης παίρνει την τιμή 370 αντί της τιμής 37 και το ζεύγος (x , y)=(159 , 88) παίρνει την τιμή (220 , 100). Αυτές οι δύο παρατηρήσεις παράγουν ένα αρνητικό καθώς και ένα θετικό σημείο μόχλευσης (Διάγραμμα 6.1.9).

Διάγραμμα 6.1.9
Διάγραμμα διασποράς με ΑΣΜ & ΘΣΜ



Πίνακας 6.1.8
Συντελεστές παλινδρόμησης με ΑΣΜ & ΘΣΜ

	Συντελεστές παλινδρόμησης	
Εκτιμητές	$\hat{\beta}_0$	$\hat{\beta}_1$
ΕΤ (LS)	57,558	0.095
ΜΜ-εκτιμητές	35,432	0,321
ΕΠΤ (LTS)	35,375	0,321
ΕΔΤ (LMS)	36,343	0,314
Μ-εκτιμητές	37,819	0,295

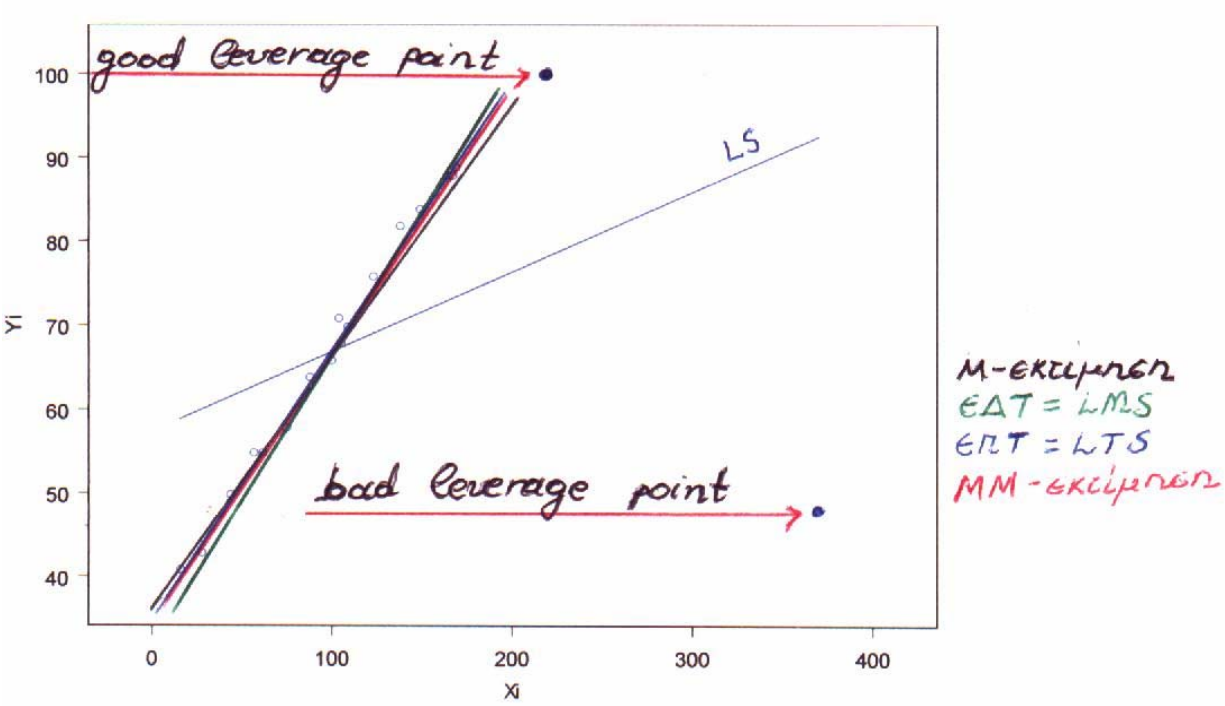
❖ **<u>Συμπεράσματα</u>**

Από το διάγραμμα 6.1.8 παρατηρούμε ότι η ευθεία έχει επηρεαστεί σε μεγάλο βαθμό από το αρνητικό και το θετικό σημείο μόχλευσης με αποτέλεσμα η ευθεία παλινδρόμησης να μετατοπίζεται προς την κατεύθυνση του αρνητικού σημείου μόχλευσης. Συνεπώς καταλήγουμε στο συμπέρασμα ότι τα αποτελέσματα που προκύπτουν από το νέο αλλοιωμένο δείγμα (Πίνακας 6.1.6) είναι ότι οι συντελεστές

παλινδρόμησης μεταβάλλονται στην περίπτωση των ελαχίστων τετραγώνων. Αντιθέτως οι ευθείες που προκύπτουν από τις ανθεκτικές μεθόδους παραμένουν αναλλοίωτες.

- <u>Διάγραμμα διασποράς με τη χρήση μεθόδου ελαχίστων τετραγώνων και ανθεκτικών εκτιμητών</u>

Διάγραμμα 6.1.10



Από το διάγραμμα 6.1.10 διαπιστώνουμε ότι στο συγκεκριμένο παράδειγμα η ευθεία παλινδρόμησης ελαχίστων τετραγώνων μετατοπίζεται προς την κατεύθυνση του αρνητικού σημείου μόχλευσης. Αντιθέτως οι ευθείες των ανθεκτικών εκτιμητών βρίσκονται πολύ κοντά μεταξύ τους και σε μεγάλη απόσταση από την ευθεία ελαχίστων τετραγώνων παραμένοντας ανεπηρέαστες από το αρνητικό σημείο μόχλευσης.

Στο Παράρτημα Α παρουσιάζονται αναλυτικά τα αποτελέσματα συντελεστών παλινδρόμησης ελαχίστων τετραγώνων και ανθεκτικών εκτιμητών μέσω του στατιστικού πακέτου S-PLUS.

6.2 Πολλαπλή παλινδρόμηση

Στην πολλαπλή παλινδρόμηση, η εξαρτημένη μεταβλητή y_i σχετίζεται με τις p επεξηγηματικές μεταβλητές $x_{i1}\beta_1$, $x_{i2}\beta_2$, …$x_{ip}\beta_p$ στο μοντέλο:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p \qquad (i = 1,2,\ldots,n)$$

Όπως και στην περίπτωση της απλής παλινδρόμησης, η μέθοδος ελαχίστων τετραγώνων είναι ιδιαίτερα ευαίσθητη όταν στα δεδομένα μας υπάρχουν έκτοπες παρατηρήσεις. Η αναγνώριση απομονωμένων τιμών γίνεται ακόμα πιο δύσκολη στην περίπτωση της πολλαπλής παλινδρόμησης, εξαιτίας του γεγονότος ότι ο εντοπισμός έκτοπων παρατηρήσεων είναι αδύνατο να πραγματοποιηθεί μέσω του διαγράμματος διασποράς (scatterplot).

Προκειμένου να αντιμετωπιστεί αυτό το πρόβλημα η χρήση των διαγραμμάτων των τυποποιημένων σφαλμάτων κρίνεται απαραίτητη. Αρχικά είναι απαραίτητο να συγκρίνουμε τα αποτελέσματα των τυποποιημένων σφαλμάτων που προκύπτουν τόσο με τη μέθοδο ελαχίστων τετραγώνων όσο και με τις ανθεκτικές μεθόδους. Στη συνέχεια αν τα αποτελέσματα των δύο μεθόδων είναι ίδια, τότε η μέθοδος ελαχίστων τετραγώνων μπορεί να εφαρμοστεί. Σε περίπτωση όμως που τα αποτελέσματα διαφέρουν μεταξύ τους, τότε οι ανθεκτικές μέθοδοι μπορούν να χρησιμοποιηθούν ως ένα αξιόπιστο εργαλείο για τον εντοπισμό έκτοπων παρατηρήσεων.

Παράδειγμα 2^o: Στη συνέχεια παρατίθεται ένα παράδειγμα τα δεδομένα του οποίου (Πίνακας 6.2.1) προέρχονται από το βιβλίο του Rousseeuw & Yohai (1984). Τα δεδομένα περιγράφουν την λειτουργία ενός φυτού για την οξείδωση της αμμωνίας σε νιτρικό οξύ και περιλαμβάνουν 21 παρατηρήσεις. Η απώλεια οξείδωσης (y) αναλύεται από τον βαθμό λειτουργίας (x_1), την εσωτερική θερμοκρασία του νερού (x_2) και την συγκέντρωση οξέως (x_3). Μέσω του συγκεκριμένου παραδείγματος εύκολα προκύπτει ο κίνδυνος που υπάρχει να χρησιμοποιήσουμε το μοντέλο ελαχίστων τετράγωνων βασιζόμενοι αποκλειστικά και μόνο στο διάγραμμα σφαλμάτων ελαχίστων τετραγώνων.

Πίνακας 6.2.1
Δεδομένα παραδείγματος

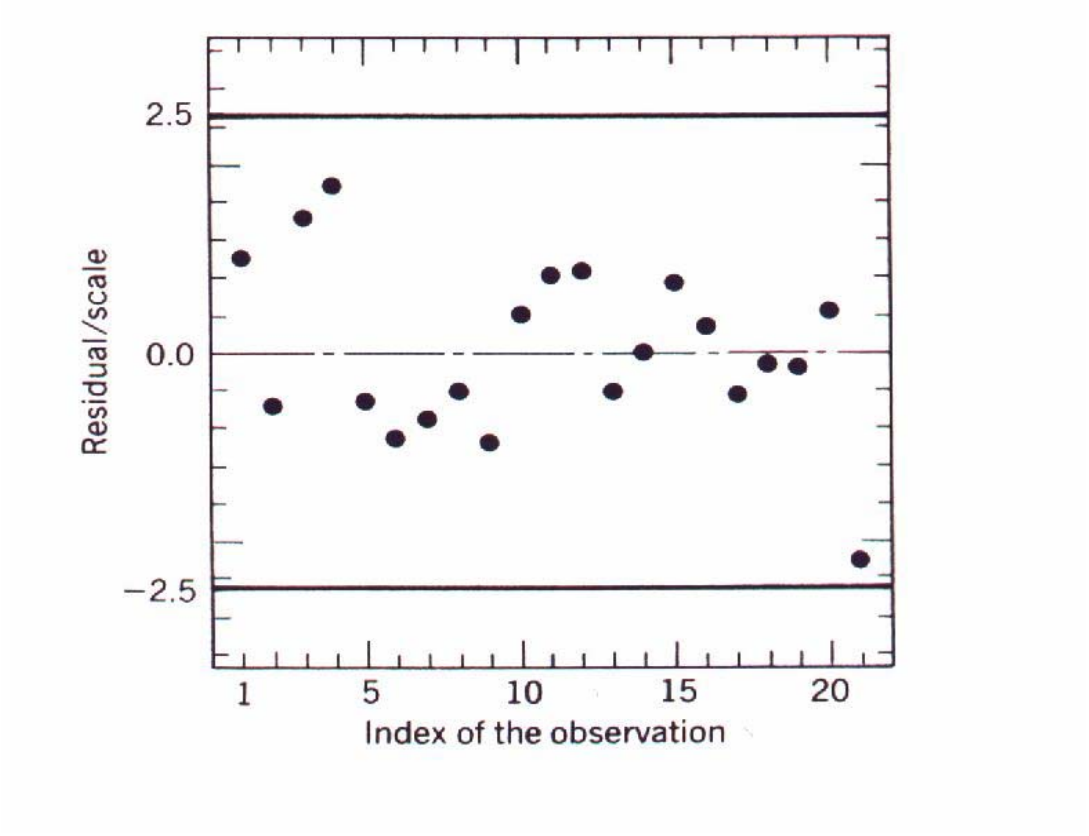
Παρατήρηση	Βαθμός λειτουργίας	Θερμοκρασία	Συγκέντρωση οξέως	Απώλεια οξείδωσης
(i)	(x_1)	(x_2)	(x_3)	(y)
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

Μέσω της μεθόδου ελαχίστων τετραγώνων προκύπτει η ακόλουθη εξίσωση:

$$\widehat{y}_{LS} = -39{,}91967 + 0{,}71564x_1 + 1{,}29528x_2 - 0{,}15212x_3 .$$

Από το διάγραμμα των σφαλμάτων των ελαχίστων τετραγώνων (Διάγραμμα 6.2.1) προκύπτει ότι ανάμεσα στις δύο ευθείες παρουσιάζονται τα τυποποιημένα σφάλματα που παίρνουν τιμές μεταξύ του –2,5 και του 2,5. Καταλήγουμε στο συμπέρασμα ότι τα δεδομένα δεν περιέχουν καμία έκτοπη παρατήρηση δεδομένου ότι όλα τα τυποποιημένα σφάλματα ελαχίστων τετραγώνων βρίσκονται μεταξύ των δύο ευθειών.

Διάγραμμα 6.2.1
Διάγραμμα σφαλμάτων ελαχίστων τετραγώνων



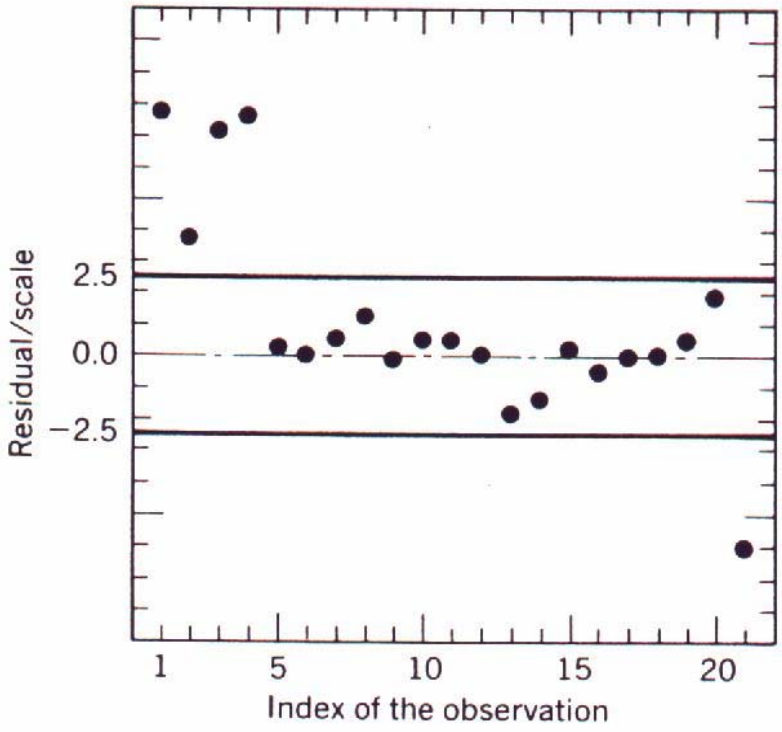
Εφαρμόζοντας όμως τις ανθεκτικές μεθόδους καταλήγουμε σε εντελώς διαφορετικά αποτελέσματα. Συγκεκριμένα μέσω της μεθόδου ελαχίστων διαμέσων τετραγώνων προκύπτει η ακόλουθη εξίσωση:

$$\hat{y}_{LMS} = -34{,}17857 + 0{,}71428x_1 + 0{,}35714x_2 - 0.0000x_3 .$$

Από διάγραμμα 6.2.2 το οποίο βασίζεται στην ανθεκτική προσαρμογή μέσω της μεθόδου ελαχίστων διαμέσων τετραγώνων γίνεται γρήγορα αντιληπτό ότι υπάρχουν κάποιες παρατηρήσεις που επηρεάζουν την μελέτη μας. Συγκεκριμένα οι παρατηρήσεις 1, 2, 3, 4 και 21 θεωρούνται έκτοπες δεδομένου ότι βρίσκονται εκτός των ορίων των δύο ευθειών.

Ακολούθως στον πίνακα 6.2.2 παρουσιάζονται συνοπτικά τα αποτελέσματα των συντελεστών παλινδρόμησης που προκύπτουν μέσω της μεθόδου ελαχίστων τετραγώνων και ανθεκτικών εκτιμητών.

Διάγραμμα 6.2.2
Διάγραμμα σφαλμάτων ελαχίστων διαμέσων τετραγώνων



Πίνακας 6.2.2
Συντελεστές παλινδρόμησης ΜΕΤ & ανθεκτικών εκτιμητών

	Συντελεστές παλινδρόμησης			
Εκτιμητές	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
ΕΤ (LS)	-39,92	0,71564	1,29528	-0,1521
ΕΔΤ (LMS)	-34,179	0,71428	0,35714	0
ΕΠΤ (LTS)	-36,203	0,7398	0,4786	-0,0235
Μ-εκτιμητές	-36,046	0,71351	0,49085	-0,0103
ΜΜ-εκτιμητές	-34,863	0,74688	0,45423	-0,0372

Στο Παράρτημα Β παρουσιάζονται αναλυτικά τα αποτελέσματα συντελεστών παλινδρόμησης ελαχίστων τετραγώνων και ανθεκτικών εκτιμητών μέσω του στατιστικού πακέτου S-PLUS.

6.3 Συμπεράσματα

Με βάση τη μελέτη μας, εύκολα συμπεραίνουμε ότι η ανθεκτική παλινδρόμηση είναι ιδιαίτερα χρήσιμη στην αναγνώριση έκτοπων παρατηρήσεων. Η υπεροχή των ανθεκτικών εκτιμητών έναντι της μεθόδου ελαχίστων τετραγώνων σε δεδομένα όπου υπάρχουν έκτοπες παρατηρήσεις είναι αναμφισβήτητη.

Στην περίπτωση της απλής παλινδρόμησης όπως προκύπτει από το παράδειγμα, ένα και μόνο αρνητικό σημείο μόχλευσης έχει σαν αποτέλεσμα την μετατόπιση της ευθείας παλινδρόμησης προς αυτό και συνεπώς την καταστροφή της μελέτης μας. Αντιθέτως οι ευθείες των ανθεκτικών εκτιμητών βρίσκονται πολύ κοντά μεταξύ τους και σε μεγάλη απόσταση από την ευθεία ελαχίστων τετραγώνων παραμένοντας ανεπηρέαστες από το αρνητικό σημείο μόχλευσης. Συνεπώς η εφαρμογή ανθεκτικών εκτιμητών κρίνεται απαραίτητη.

Στην περίπτωση της πολλαπλής παλινδρόμησης όπου έχουμε μεγάλο αριθμό συντελεστών, παρά το γεγονός ότι οι κάθετες έκτοπες παρατηρήσεις μπορούν να εντοπιστούν μέσω των διαγραμμάτων σφαλμάτων, κάτι τέτοιο δεν συμβαίνει στην περίπτωση όπου στο δείγμα μας υπάρχουν αρνητικά σημεία μόχλευσης. Για αυτό ακριβώς τον λόγο εμπιστευόμαστε τα διαγράμματα σφαλμάτων των ελαχίστων τετραγώνων μόνο όταν αυτά συμπίπτουν με τα αποτελέσματα που προκύπτουν από την εφαρμογή ανθεκτικών μεθόδων. Σε οποιαδήποτε άλλη περίπτωση οφείλουμε να εφαρμόσουμε τις ανθεκτικές μεθόδους.

Η ανθεκτική εκτίμηση παλινδρόμησης πρέπει να έχει τα ακόλουθα χαρακτηριστικά:

- να συμπεριφέρεται εξίσου καλά με την ΜΕΤ όταν η τελευταία είναι η κατάλληλη επιλογή
- να συμπεριφέρεται καλύτερα από την ΜΕΤ όταν δεν πληρούνται οι βασικές υποθέσεις
- να μην είναι ιδιαίτερα δύσκολο να υπολογιστεί και να γίνει κατανοητή.

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα Α

Παρακάτω παρουσιάζονται αναλυτικά τα αποτελέσματα του παραδείγματος 6.1 όπως αυτά προκύπτουν μέσω του στατιστικού προγράμματος S-PLUS.

Α.1) Χωρίς έκτοπη παρατήρηση

Πίνακας 6.1.9

Αποτελέσματα εφαρμόζοντας τη MET χωρίς ΕΠ

```
*** Linear Model ***

Call: lm(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)
Residuals:
   Min     1Q  Median     3Q  Max
 -1.619 -1.167 0.01909 0.7276 2.16

Coefficients:
                Value Std. Error t value Pr(>|t|)
(Intercept) 35.4583  0.6350    55.8355  0.0000
         Xi  0.3216  0.0056    57.8993  0.0000

Residual standard error: 1.23 on 18 degrees of freedom
Multiple R-Squared: 0.9947
F-statistic: 3352 on 1 and 18 degrees of freedom, the p-value is 0
```

Πίνακας 6.1.10

Αποτελέσματα εφαρμόζοντας την ΜΜ-εκτίμηση χωρίς ΕΠ

```
*** Robust MM Linear Regression ***
Final M-estimates.

Call: lmRobMM(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude,
robust.control =
          list(tlo = 0.0001, tua = 1.5e-006, mxr = 50, mxf = 50, mxs = 50, tl =
        1e-006, estim = "Test Based", seed = 1313, level = 0.1, efficiency = 0.85,
      sampling = "Exhaustive", weight = c("Optimal", "Optimal")), genetic.control
       = list(popsize = NULL, mutate.prob = NULL, random.n = NULL, births.n =
         NULL, stock = list(), maxslen = NULL, stockprob = NULL, nkeep = 1))

Residuals:
   Min     1Q  Median     3Q  Max
 -1.619 -1.167 0.01909 0.7276 2.16
Coefficients:
                Value Std. Error t value Pr(>|t|)
(Intercept) 35.4583  0.6989    50.7333  0.0000
         Xi  0.3216  0.0061    52.5996  0.0000

Residual scale estimate: 1.401 on 18 degrees of freedom
```

Πίνακας 6.1.11

Αποτελέσματα εφαρμόζοντας την μέθοδο LTS χωρίς ΕΠ

*** Robust LTS Linear Regression ***

Method:
[1] "Least Trimmed Squares Robust Regression."

Call:
ltsreg.formula(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)

Coefficients:
Intercept Xi
35.4727 0.3193

Scale estimate of residuals: 1.258

Πίνακας 6.1.12

Αποτελέσματα εφαρμόζοντας την μέθοδο LMS χωρίς ΕΠ

*** Robust LMS Linear Regression ***

$coefficients:
Intercept x
35.6378 0.3149606

$scale:
Y
1.384616

$residuals:
[1] 1.62204724 0.03149606 -0.16535433 2.60629921 1.40944882 0.70866142
0.50393701 -1.13385827 0.32283465 -1.45669291 2.89763780
[12] -0.70866142 2.28346457 -1.25984252 0.64566929 0.70866142 0.13385827 -
0.23622047 1.43307087 -0.23622047

Πίνακας 6.1.13

Αποτελέσματα εφαρμόζοντας την M-εκτίμηση χωρίς έκτοπη παρατήρηση

M-estimator, c=1.2

$coefficients:
(Intercept) x
35.45479 0.321403

$residuals:
[1] 1.0126434 -0.4877144 -0.3817727 2.1193007 1.2252423 0.6533026
0.4034815 -1.5950872 0.4027659 -1.4540703 2.1915982 -1.2021023
[13] 1.4421350 -1.5600119 0.2617489 -0.1648801 -0.7718952 -1.1290891
0.6561651 -1.1290891
$fitted.values:
[1] 74.98736 70.48771 55.38177 68.88070 53.77476 47.34670 49.59652 67.59509
40.59723 44.45407 79.80840 69.20210 86.55787 59.56001 63.73825
[16] 88.16488 89.77190 89.12909 83.34383 89.12909

Α.2) <u>Αρνητικό σημείο μόχλευσης (ΑΣΜ)</u>

Πίνακας 6.1.14

Αποτελέσματα εφαρμόζοντας την ΜΕΤ όταν υπάρχει ΑΣΜ

```
*** Linear Model ***

Call: lm(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)
Residuals:
   Min    1Q Median    3Q   Max
 -40.8 -8.64  1.425 13.67 16.42

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 58.9388  6.6142     8.9110  0.0000
         Xi  0.0807  0.0469     1.7191  0.1027

Residual standard error: 15.6 on 18 degrees of freedom
Multiple R-Squared: 0.141
F-statistic: 2.955 on 1 and 18 degrees of freedom, the p-value is 0.1027
```

Πίνακας 6.1.15

Αποτελέσματα εφαρμόζοντας την ΜΜ-εκτίμηση όταν υπάρχει ΑΣΜ

```
*** Robust MM Linear Regression ***
Final M-estimates.

Call: lmRobMM(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude,
robust.control = list(tlo
      = 0.0001, tua = 1.5e-006, mxr = 50, mxf = 50, mxs = 50, tl = 1e-006, estim =
  "Test Based", seed = 1313, level = 0.1, efficiency = 0.85, sampling = "Exhaustive",
       weight = c("Optimal", "Optimal")), genetic.control = list(popsize = NULL,
 mutate.prob = NULL, random.n = NULL, births.n = NULL, stock = list(), maxslen =
                              NULL, stockprob = NULL, nkeep = 1))

Residuals:
    Min      1Q  Median      3Q   Max
 -106.7 -1.194 -0.2727 0.7102 2.162

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 35.3174  0.7534    46.8783  0.0000
         Xi  0.3226  0.0064    50.0974  0.0000

Residual scale estimate: 1.552 on 18 degrees of freedom
```

Πίνακας 6.1.16

Αποτελέσματα εφαρμόζοντας την μέθοδο LTS όταν υπάρχει ΑΣΜ

```
*** Robust LTS Linear Regression ***
Method:
[1] "Least Trimmed Squares Robust Regression."
```

Call:
ltsreg.formula(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)

Coefficients:
Intercept Xi
35.3348 0.3212

Scale estimate of residuals: 1.265

Πίνακας 6.1.17

Αποτελέσματα εφαρμόζοντας την μέθοδο LMS όταν υπάρχει ΑΣΜ

*** Robust LMS Linear Regression ***

$coefficients:
Intercept x
36.34286 0.3142857

$scale:
Y
1.33278

$residuals:
[1] 1.0000000 -0.6000000 -0.8285714 1.9714286 0.7428571 -104.6285714
[7] -0.1714286 -1.7714286 -0.3714286 -2.1428571 2.2857143 -1.3428571
[13] 1.6857143 -1.9142857 0.0000000 0.1142857 -0.4571429 -0.8285714
[19] 0.8285714 -0.8285714

Πίνακας 6.1.18

Αποτελέσματα εφαρμόζοντας την Μ-εκτίμηση όταν υπάρχει ΑΣΜ

M-estimator, c=1.2

$coefficients:

(Intercept) x
36.43643 0.3104658

$residuals:
[1] 1.37628501 -0.27719437 -0.68530372 2.27513442 0.86702507
[6] -103.30875734 -0.09692007 -1.48300255 -0.40387883 -2.12946794
[11] 2.71929863 -1.03533134 2.19951770 -1.72135859 0.24258655
[16] 0.64718891 0.09486012 -0.28420837 1.30417529 -0.28420837

$fitted.values:
[1] 74.62371 70.27719 55.68530 68.72487 54.13297 151.30876 50.09692
67.48300
[9] 41.40388 45.12947 79.28070 69.03533 85.80048 59.72136 63.75741
87.35281
[17] 88.90514 88.28421 82.69582 88.28421

Α.3) Θετικό σημείο μόχλευσης (ΘΣΜ)

Πίνακας 6.1.19

Αποτελέσματα εφαρμόζοντας την MET όταν υπάρχει ΘΣΜ

```
*** Linear Model ***

Call: lm(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)
Residuals:
   Min      1Q  Median    3Q   Max
 -4.332 -0.6634 0.01582 0.942 2.961

Coefficients:
                Value Std. Error t value Pr(>|t|)
(Intercept) 36.4742  0.8210    44.4279  0.0000
         Xi  0.3084  0.0069    44.8126  0.0000

Residual standard error: 1.677 on 18 degrees of freedom
Multiple R-Squared: 0.9911
F-statistic: 2008 on 1 and 18 degrees of freedom, the p-value is 0
```

Πίνακας 6.1.20

Αποτελέσματα εφαρμόζοντας την MM-εκτίμηση όταν υπάρχει ΘΣΜ

```
*** Robust MM Linear Regression ***
Final M-estimates.

Call: lmRobMM(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude,
robust.control = list(tlo = 0.0001, tua = 1.5e-006, mxr = 50, mxf = 50, mxs =
    50, tl = 1e-006, estim = "Test Based", seed = 1313, level = 0.1, efficiency = 0.85,
sampling = "Exhaustive", weight = c("Optimal",
 "Optimal")), genetic.control = list(popsize = NULL, mutate.prob = NULL, random.n
= NULL, births.n = NULL, stock = list(), maxslen =
                                NULL, stockprob = NULL, nkeep = 1))

Residuals:
   Min     1Q  Median     3Q   Max
 -5.921 -1.015 -0.2027 0.6476 2.302

Coefficients:
            Value Std. Error t value Pr(>|t|)
(Intercept) 35.5655  0.6820    52.1498  0.0000
         Xi  0.3198  0.0061    52.2654  0.0000

Residual scale estimate: 1.401 on 18 degrees of freedom
```

Πίνακας 6.1.21

Αποτελέσματα εφαρμόζοντας την μέθοδο LTS όταν υπάρχει ΘΣΜ

```
*** Robust LTS Linear Regression ***
Method:
[1] "Least Trimmed Squares Robust Regression."
```

Call:
ltsreg.formula(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)

Coefficients:
Intercept Xi
35.5857 0.3184

Scale estimate of residuals: 1.222

Πίνακας 6.1.22

Αποτελέσματα εφαρμόζοντας την M-εκτίμηση όταν υπάρχει ΘΣΜ

M-estimator, c=1.2

$coefficients:
(Intercept) x
35.95856 0.3142661

$residuals:
[1] 1.38671087 -0.21356332 -0.44305524 2.35776733 1.12827541 0.41359800 0.21373509
[8] -1.38516815 0.01318672 -1.75800684 2.67271893 -0.95649880 -5.09710368 -1.52851492
[15] 0.38602540 0.50179956 -0.06953108 -0.44099882 1.21579150 -0.44099882

$fitted.values:
[1] 74.61329 70.21356 55.44306 68.64223 53.87172 47.58640 49.78626 67.38517
[9] 40.98681 44.75801 79.32728 68.95650 105.09710 59.52851 63.61397 87.49820
[17] 89.06953 88.44100 82.78421 88.44100

Πίνακας 6.1.23

Αποτελέσματα εφαρμόζοντας την μέθοδο LMS όταν υπάρχει ΘΣΜ

*** Robust LMS Linear Regression ***

$coefficients:
Intercept x
35.6378 0.3149606
$scale:
Y
1.312714

$residuals:
[1] 1.62204724 0.03149606 -0.16535433 2.60629921 1.40944882 0.70866142 0.50393701
[8] -1.13385827 0.32283465 -1.45669291 2.89763780 -0.70866142 -4.92913386 -1.25984252
[15] 0.64566929 0.70866142 0.13385827 -0.23622047 1.43307087 -0.23622047

Α.4) Κάθετη έκτοπη παρατήρηση (ΚΕΠ)

Πίνακας 6.1.24

Αποτελέσματα εφαρμόζοντας την ΜΕΤ όταν υπάρχει ΚΕΠ

*** Linear Model ***

Call: lm(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)
Residuals:

Min	1Q	Median	3Q	Max
-10.59	-7.723	-6.084	-3.615	117.8

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	**36.4592**	14.7816	2.4665	0.0239
	0.3721	0.1293	2.8777	0.0100

Residual standard error: 28.63 on 18 degrees of freedom
Multiple R-Squared: 0.3151
F-statistic: 8.281 on 1 and 18 degrees of freedom, the p-value is 0.01002

Πίνακας 6.1.25

Αποτελέσματα εφαρμόζοντας την ΜΜ-εκτίμηση όταν υπάρχει ΚΕΠ

*** Robust MM Linear Regression ***
Final M-estimates.

Call: lmRobMM(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude, robust.control = list(tlo = 0.0001, tua = 1.5e-006, mxr = 50, mxf = 50, mxs = 50, tl = 1e-006, estim = "Test Based", seed = 1313, level = 0.1, efficiency = 0.85, sampling = "Exhaustive", weight = c("Optimal", "Optimal")), genetic.control = list(popsize = NULL, mutate.prob = NULL, random.n = NULL, births.n = NULL, stock = list(), maxslen = NULL, stockprob = NULL, nkeep = 1))
Residuals:

Min	1Q	Median	3Q	Max
-1.568	-1.087	0.08108	0.8311	125

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	**35.4498**	0.7632	46.4477	0.0000
Xi	**0.3212**	0.0067	47.8642	0.0000

Residual scale estimate: 1.441 on 18 degrees of freedom

Πίνακας 6.1.26

Αποτελέσματα εφαρμόζοντας την μέθοδο LTS όταν υπάρχει ΚΕΠ

*** Robust LTS Linear Regression ***
Method:
[1] "Least Trimmed Squares Robust Regression."

Call:
ltsreg.formula(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)

Coefficients:
Intercept Xi
35.5566 0.3191

Scale estimate of residuals: 1.251

Πίνακας 6.1.27

Αποτελέσματα εφαρμόζοντας την μέθοδο LMS όταν υπάρχει ΚΕΠ

*** Robust LMS Linear Regression ***

$coefficients:
Intercept x
35.6378 0.3149606

$scale:
Y
1.369368

$residuals:
[1] 125.62204724 0.03149606 -0.16535433 2.60629921 1.40944882
0.70866142 0.50393701 -1.13385827 0.32283465
[10] -1.45669291 2.89763780 -0.70866142 2.28346457 -1.25984252
0.64566929 0.70866142 0.13385827 -0.23622047
[19] 1.43307087 -0.23622047

Πίνακας 6.1.28

Αποτελέσματα εφαρμόζοντας την Μ-εκτίμηση όταν υπάρχει ΚΕΠ

M-estimator, c=1.2

$coefficients:
(Intercept) x
35.46842 0.3221199

$residuals:
[1] 124.9108275 -0.5794938 -0.4398582 2.0311058 1.1707413 0.6131395
0.3583001 -1.6804146 0.3776575 -1.4877814
[11] 2.0790289 -1.2910141 1.3145109 -1.6274170 0.1850243 -0.2960886 -
0.9066881 -1.2624483 0.5357100 -1.2624483

$fitted.values:
[1] 75.08917 70.57949 55.43986 68.96889 53.82926 47.38686 49.64170 67.68041
40.62234 44.48778 79.92097 69.29101 86.68549 59.62742 63.81498
[16] 88.29609 89.90669 89.26245 83.46429 89.26245

Α.5)<u>Κάθετη έκτοπη παρατήρηση (ΚΕΠ) – αρνητικό σημείο μόχλευσης (ΑΣΜ)</u>

Πίνακας 6.1.29

Αποτελέσματα εφαρμόζοντας την ΜΕΤ όταν υπάρχει ΚΕΠ & ΑΣΜ

*** Linear Model ***

Call: lm(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)
Residuals:
Min 1Q Median 3Q Max
-47.93 -14.61 -4.728 9.207 124.9

Coefficients:
Value Std. Error t value Pr(>|t|)
(Intercept) 64.6951 14.3983 4.4932 0.0003
Xi 0.0844 0.1022 0.8260 0.4196

Residual standard error: 33.96 on 18 degrees of freedom
Multiple R-Squared: 0.03652
F-statistic: 0.6823 on 1 and 18 degrees of freedom, the p-value is 0.4196

Πίνακας 6.1.30

Αποτελέσματα εφαρμόζοντας την ΜΜ-εκτίμηση όταν υπάρχει ΚΕΠ & ΑΣΜ

*** Robust MM Linear Regression ***
Final M-estimates.

Call: lmRobMM(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude,
robust.control =
list(tlo = 0.0001, tua = 1.5e-006, mxr = 50, mxf = 50, mxs = 50, tl =
1e-006, estim = "Test Based", seed = 1313, level = 0.1, efficiency = 0.85,
sampling = "Exhaustive", weight = c("Optimal", "Optimal")), genetic.control
= list(popsize = NULL, mutate.prob = NULL, random.n = NULL, births.n =
NULL, stock = list(), maxslen = NULL, stockprob = NULL, nkeep = 1))

-106.5 -1.12 -0.2142 0.847 125.1
Coefficients:
Value Std. Error t value Pr(>|t|)
(Intercept) 35.3035 0.7853 44.9551 0.0000
Xi 0.3222 0.0067 47.8585 0.0000
Residual scale estimate: 1.576 on 18 degrees of freedom

Πίνακας 6.1.31

Αποτελέσματα εφαρμόζοντας την μέθοδο LTS όταν υπάρχει ΚΕΠ & ΑΣΜ

*** Robust LTS Linear Regression ***
Method:
[1] "Least Trimmed Squares Robust Regression."

Call:
ltsreg.formula(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)

Coefficients:
Intercept Xi
35.0615 0.3242

Scale estimate of residuals: 1.272

Πίνακας 6.1.32
Αποτελέσματα εφαρμόζοντας την μέθοδο LMS όταν υπάρχει ΚΕΠ & ΑΣΜ

*** Robust LMS Linear Regression ***

$coefficients:
Intercept x
36.34286 0.3142857

$scale:
Y
1.350869

$residuals:
[1] 125.0000000 -0.6000000 -0.8285714 1.9714286 0.7428571 -104.6285714
[7] -0.1714286 -1.7714286 -0.3714286 -2.1428571 2.2857143 -1.3428571
[13] 1.6857143 -1.9142857 0.0000000 0.1142857 -0.4571429 -0.8285714
[19] 0.8285714 -0.8285714

Πίνακας 6.1.33
Αποτελέσματα εφαρμόζοντας την Μ-εκτίμηση όταν υπάρχει ΚΕΠ & ΑΣΜ

M-estimator, c=1.2

$coefficients:
(Intercept) x
36.5118 0.3101217

$residuals:
[1] 125.34323166 -0.31506408 -0.73934266 2.23554458 0.81126601
[6] -103.25683625 -0.15715147 -1.52396849 -0.47374296 -2.19520375
[11] 2.69140568 -1.07457715 2.17884930 -1.77092518 0.19749230
[16] 0.62824063 0.07763197 -0.30212456 1.28006662 -0.30212456

$fitted.values:
[1] 74.65677 70.31506 55.73934 68.76446 54.18873 151.25684 50.15715
67.52397
[9] 41.47374 45.19520 79.30859 69.07458 85.82115 59.77093 63.80251
87.37176
[17] 88.92237 88.30212 82.71993 88.3021

Α.6)<u>Κάθετη έκτοπη παρατήρηση (ΚΕΠ) – Θετικό σημείο μόχλευσης (ΘΣΜ)</u>

Πίνακας 6.1.34

Αποτελέσματα εφαρμόζοντας την ΜΕΤ όταν υπάρχει ΚΕΠ & ΘΣΜ

*** Linear Model ***

Call: lm(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)
Residuals:

Min	1Q	Median	3Q	Max
-14.55	-7.302	-5.85	-3.961	118.8

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	**38.9307**	14.1179	2.7575	0.0130
Xi	**0.3437**	0.1184	2.9040	0.0095

Residual standard error: 28.85 on 18 degrees of freedom
Multiple R-Squared: 0.319
F-statistic: 8.433 on 1 and 18 degrees of freedom, the p-value is 0.009463

Πίνακας 6.1.35

Αποτελέσματα εφαρμόζοντας την ΜΜ-εκτίμηση όταν υπάρχει ΚΕΠ & ΘΣΜ

*** Robust MM Linear Regression ***
Final M-estimates.

Call: lmRobMM(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude,
robust.control
= list(tlo = 0.0001, tua = 1.5e-006, mxr = 50, mxf = 50, mxs = 50, tl
= 1e-006, estim = "Test Based", seed = 1313, level = 0.1, efficiency
= 0.85, sampling = "Exhaustive", weight = c("Optimal", "Optimal")),
genetic.control = list(popsize = NULL, mutate.prob = NULL, random.n =
NULL, births.n = NULL, stock = list(), maxslen = NULL, stockprob =
NULL, nkeep = 1))
Residuals:

Min	1Q	Median	3Q	Max
-5.789	-0.9228	-0.1332	0.6888	125.2

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	**35.5625**	0.7342	48.4403	0.0000
Xi	**0.3192**	0.0066	48.1188	0.0000

Residual scale estimate: 1.441 on 18 degrees of freedom

Πίνακας 6.1.36

Αποτελέσματα εφαρμόζοντας την μέθοδο LTS όταν υπάρχει ΚΕΠ & ΘΣΜ

*** Robust LTS Linear Regression ***
Method:
[1] "Least Trimmed Squares Robust Regression."

Call:
ltsreg.formula(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)

Coefficients:
Intercept Xi
35.6214 0.3186

Scale estimate of residuals: 1.218

Πίνακας 6.1.37

Αποτελέσματα εφαρμόζοντας την μέθοδο LMS όταν υπάρχει ΚΕΠ & ΘΣΜ

*** Robust LMS Linear Regression ***

$coefficients:
Intercept x
35.6378 0.3149606

$scale:
Y
1.290921

$residuals:
[1] 125.62204724 0.03149606 -0.16535433 2.60629921 1.40944882
[6] 0.70866142 0.50393701 -1.13385827 0.32283465 -1.45669291
[11] 2.89763780 -0.70866142 -4.92913386 -1.25984252 0.64566929
[16] 0.70866142 0.13385827 -0.23622047 1.43307087 -0.23622047

Πίνακας 6.1.38

Αποτελέσματα εφαρμόζοντας την Μ-εκτίμηση όταν υπάρχει ΚΕΠ & ΘΣΜ

M-estimator, c=1.2

$coefficients:
(Intercept) x
35.96923 0.31435

$residuals:
[1] 125.365715624 -0.233383953 -0.458932535 2.338366197 1.112817616
[6] 0.399818220 0.199368008 -1.404233682 0.001168853 -1.771031509
[11] 2.650465171 -0.975983833 -5.126237304 -1.545482927 0.367966680
[16] 0.477364387 -0.094385764 -0.465685704 1.192614839 -0.465685704
$fitted.values:
[1] 74.63428 70.23338 55.45893 68.66163 53.88718 47.60018 49.80063
67.40423
[9] 40.99883 44.77103 79.34953 68.97598 105.12624 59.54548 63.63203
87.52264
[17] 89.09439 88.46569 82.80739 88.46569

Α.7)<u>Θετικό σημείο μόχλευσης (ΘΣΜ) – Αρνητικό σημείο μόχλευσης (ΑΣΜ)</u>

Πίνακας 6.1.39

Αποτελέσματα εφαρμόζοντας την MET όταν υπάρχει ΘΣΜ & ΑΣΜ

*** Linear Model ***

Call: lm(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)
Residuals:

Min	1Q	Median	3Q	Max
-44.65	-8.082	1.294	12.88	21.58

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	**57.5576**	6.7483	8.5292	0.0000
Xi	**0.0948**	0.0466	2.0369	0.0566

Residual standard error: 16.04 on 18 degrees of freedom
Multiple R-Squared: 0.1873
F-statistic: 4.149 on 1 and 18 degrees of freedom, the p-value is 0.05664

Πίνακας 6.1.40

Αποτελέσματα εφαρμόζοντας την MM-εκτίμηση όταν υπάρχει ΘΣΜ & ΑΣΜ

*** Robust MM Linear Regression ***
Final M-estimates.

Call: lmRobMM(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude,
robust.control =
list(tlo = 0.0001, tua = 1.5e-006, mxr = 50, mxf = 50, mxs = 50, tl =
1e-006, estim = "Test Based", seed = 1313, level = 0.1, efficiency = 0.85,
sampling = "Exhaustive", weight = c("Optimal", "Optimal")), genetic.control
= list(popsize = NULL, mutate.prob = NULL, random.n = NULL, births.n =
NULL, stock = list(), maxslen = NULL, stockprob = NULL, nkeep = 1))

Residuals:

Min	1Q	Median	3Q	Max
-106.1	-1.187	-0.3573	0.5343	2.303

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	**35.4320**	0.7330	48.3377	0.0000
Xi	**0.3208**	0.0064	49.9755	0.0000

Residual scale estimate: 1.552 on 18 degrees of freedom

Πίνακας 6.1.41

Αποτελέσματα εφαρμόζοντας την μέθοδο LTS όταν υπάρχει ΘΣΜ & ΑΣΜ

*** Robust LTS Linear Regression ***
Method:

[1] "Least Trimmed Squares Robust Regression."

Call:
ltsreg.formula(formula = Yi ~ Xi, data = SDF1, na.action = na.exclude)

Coefficients:
Intercept Xi
35.3751 0.3210

Scale estimate of residuals: 1.24

Πίνακας 6.1.42

Αποτελέσματα εφαρμόζοντας την μέθοδο LMS όταν υπάρχει ΘΣΜ & ΑΣΜ

*** Robust LMS Linear Regression ***
$coefficients:
Intercept x
36.34286 0.3142857
$scale:
Y
1.307572
$residuals:
[1] 1.0000000 -0.6000000 -0.8285714 1.9714286 0.7428571 -104.6285714
[7] -0.1714286 -1.7714286 -0.3714286 -2.1428571 2.2857143 -1.3428571
[13] -5.4857143 -1.9142857 0.0000000 0.1142857 -0.4571429 -0.8285714
[19] 0.8285714 -0.8285714

Πίνακας 6.1.43

Αποτελέσματα εφαρμόζοντας την Μ-εκτίμηση όταν υπάρχει ΘΣΜ & ΑΣΜ

M-estimator, c=1.2
$coefficients:
(Intercept) x
37.81878 0.2946386
$residuals:
[1] 1.94067860 0.06561848 -1.08636907 2.53881129 0.38682374 -98.83504640
[7] -0.78287494 -1.28263446 -1.53299519 -3.06865794 3.52110016 -0.75582727
[13] -2.63926199 -1.91667039 0.25302830 1.86049753 1.38730471 0.97658184
[19] 2.28007597 0.97658184

$fitted.values:
[1] 74.05932 69.93438 56.08637 68.46119 54.61318 146.83505 50.78287 67.28263
[9] 42.53300 46.06866 78.47890 68.75583 102.63926 59.91667 63.74697 86.13950
[17] 87.61270 87.02342 81.71992 87.02342

Παράρτημα Β

Παρακάτω παρουσιάζονται αναλυτικά τα αποτελέσματα του παραδείγματος 6.2 όπως αυτά προκύπτουν μέσω του στατιστικού προγράμματος S-PLUS.

Πίνακας 6.2.3
Αποτελέσματα μέσω της MET

*** Linear Model ***

Call: lm(formula = y ~ x1 + x2 + x3, data = P3, na.action = na.exclude)
Residuals:
   Min   1Q Median   3Q   Max
 -7.238 -1.712 -0.4551 2.361 5.698

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	**-39.9197**	11.8960	-3.3557	0.0038
x1	**0.7156**	0.1349	5.3066	0.0001
x2	**1.2953**	0.3680	3.5196	0.0026
x3	**-0.1521**	0.1563	-0.9733	0.3440

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-Squared: 0.9136
F-statistic: 59.9 on 3 and 17 degrees of freedom, the p-value is 3.016e-009

Πίνακας 6.2.4
Αποτελέσματα μέσω της μεθόδου ΕΔΤ

*** Robust LTS Linear Regression ***

$coefficients:

Intercept	**x1**	**x2**	**x3**
-34.17857	**0.7142857**	**0.3571429**	**3.741138e-017**

$scale:
       Y
 0.496801

$residuals:

5	6	7	8	9	10	11
0.03571429	-0.3214286	0.3214286	1.321429	-0.4642857	0.3214286	0.3214286

12	13	14	15	16	17	18
-0.3214286	-2.678571	-2.035714	0.03571429	-0.9642857	-0.3214286	-0.3214286

19	20
0.3214286	2.035714

Πίνακας 6.2.5

Αποτελέσματα μέσω της μεθόδου ΕΠΤ

```
        *** Robust LTS Linear Regression ***
Method:
[1] "Least Trimmed Squares Robust Regression."

Call:
ltsreg.formula(formula = y ~ x1 + x2 + x3, data = P3, na.action = na.exclude)
Coefficients:
 Intercept      x1      x2      x3
 -36.2032   0.7398  0.4786  -0.0235

Scale estimate of residuals: 0.998
```

Πίνακας 6.2.6

Αποτελέσματα μέσω της Μ-εκτίμησης

```
                    ***M-estimator, c=1.2***
$coefficients:
 (Intercept)        x1        x2           x3
  -36.04583 0.7135145 0.4908539 -0.01027372

$residuals:
 [1] -0.09703607 -0.58788999 -0.01710159  0.98289841 -0.73383218  0.64852140
 [7]  0.74098489  0.22156509 -2.33093116 -1.70877415  0.44910052 -0.58172064
[13] -0.21640666 -0.14449061  0.37492918  2.11438990

$fitted.values:
 [1] 18.097036 18.587890 19.017102 19.017102 15.733832 13.351479 13.259015
 [8] 12.778435 13.330931 13.708774  7.550899  7.581721  8.216407  8.144491
[15]  8.625071 12.885610
```

Πίνακας 6.2.7

Αποτελέσματα μέσω της ΜΜ-εκτίμησης

```
                *** Robust MM Linear Regression ***
Final M-estimates.
Call: lmRobMM(formula = y ~ x1 + x2 + x3, data = P3, na.action = na.exclude,
        robust.control = list(tlo = 0.0001, tua = 1.5e-006, mxr = 50, mxf = 50,
        mxs = 50, tl = 1e-006, estim = "Test Based", seed = 1313, level = 0.1,
        efficiency = 0.85, sampling = "Exhaustive", weight = c("Optimal",
        "Optimal")), genetic.control = list(popsize = NULL, mutate.prob = NULL,
        random.n = NULL, births.n = NULL, stock = list(), maxslen = NULL,
        stockprob = NULL, nkeep = 1))

Residuals:
   Min     1Q   Median     3Q   Max
-2.583 -0.5081 -0.03945 0.4703 2.003

Coefficients:
             Value Std. Error  t value Pr(>|t|)
(Intercept) -34.8629  26.9372    -1.2942   0.2199
```

x1	**0.7469**	0.5402	1.3827	0.1919
x2	**0.4542**	0.9451	0.4806	0.6395
x3	**-0.0372**	0.3932	-0.0946	0.9262

Residual scale estimate: 0.8772 on 12 degrees of freedom

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

[1] Δάσιου Δ. (1999). Ανθεκτικές μέθοδοι εκτίμησης σε σημείο και σε διάστημα. Διδακτορική διατριβή

[2] Ε.Α. Χατζηκωνσταντινίδης, Α.Γ.Καλαματιανού, *Εφαρμοσμένη πολυμεταβλητή ανάλυση*. Εκδόσεις Παπαζήση (1997).

Ξένη

[1] Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, **2**, 123-147.

[2] Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data*. Wiley, New York.

[3] Box, G. E. P. (1953). Non-normality and tests on variances, *Biometrica*, **40**, 318-335.

[4] Cook, R. D and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, New York.

[5] Davis, P. L. (1993). Aspects of Robust Linear Regression. *The Annals of Statistics*, **21**, 1843-1899.

[6] Hampel, F. R. (1971). A general qualitative definition of robustness, *The Annals of Mathematical Statistics*, **42**, 1887-1896.

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

[8] Hodges, J. L. (1967). *Efficiency in normal samples and tolerance of extreme values for some estimates of location*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, Calif., 163-186.

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

[10] Huber, P.J. (1972). Robust Statistics: A review. *The Annals of Mathematical Statistics*, **43**, 1041-1067.

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

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

[13] Krasker, W. S. (1980). Estimation in linear regression models.

[14] Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *Journal of the American Statistical Association*, **77**, 595-604.

[15] Mallows, C. L. (1979). Robust methods-some examples of their use. *Journal of the American Statistical Association*, **33**, 179-184.

[16] Maronna, R. A., Bustos, O., and Yohai, V. (1979). Bias- and efficiency-robustness of general M-estimators for regression with random carriers, in *Smoothing Techniques for Curve Estimation*, edited by T. Gasser and M. Rosenblatt, Spinger Verlag, New York, pp.91-116.

[17] Rocke, D. M., Downs, G. W., and Rocke, A. J. (1982). Are robust estimators really necessary? *Technometrics* **24**, 95-101.

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

[19] Rousseeuw, P., and Yohai, V. (1984). Robust regression by means of S-estimators, in J. Franke, W. Haerdle, and D. Martin, eds, *Robust and Non linear Time Series Analysis.* Springer-Verlag (Berlin; New-York), 256-272.

[20] Rousseeuw, P. J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

[21] Stigler, S.M. (1973). Simon Newcomb, Percy Daniell and the History of Robust Estimation 1885-1920. *Journal of the American Statistical Association*, **68**, 872-879.

[22] Stigler, S. M. (1977). Do robust estimators work with real data?, *Journal of the American Statistical Association*, **5**, 1055-1098.

[23] Yohai, V. J. (1987). *High breakdown-point and high efficiency robust estimates for regression*. The Annals of statistics., 20,642-656.