ΚΕΦΑΛΑΙΟ 16

ΑΝΘΕΚΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (ROBUST REGRESSION)

Τα μοντέλα παλινδρόμησης αντιμετωπίζουν ιδιόρρυθμα και δύσκολα προβλήματα ανθεκτικότητας. Μία από τις δυσκολίες οφείλεται στο ότι ένα outlier δεν προκαλεί πάντοτε μεγάλο κατάλοιπο στο δικό του σημείο, αλλά επιφέρει αύξηση του μεγέθους των άλλων καταλοίπων. Μπορούμε να χαρακτηρίσουμε μία παρατήρηση $(y_i, x_{1i}, x_{2i}, ..., x_{ki})$ σαν outlier, αν η απομάκρυνσή του από το σύνολο των δεδομένων προκαλεί σημαντικές αλλαγές στις εκτιμήσεις των ελαχίστων τετραγώνων. Η αναγνώριση των outliers σε μία πολλαπλή παλινδρόμηση δεν είναι εύκολη, διότι τα outliers πολλές φορές έλκουν τη γραμμή (ή πολυεπίπεδο) των ελαχίστων τετραγώνων προς το μέρος τους και τα κατάλοιπα στα σημεία αυτά έχουν κανονικό μέγεθος.

Στόχος των ανθεκτικών μεθόδων είναι η τροποποίηση της μεθόδου των ελαχίστων τετραγώνων, έτσι ώστε να περιορισθεί η επίδραση των outliers στις εκτιμήσεις των συντελεστών παλινδρόμησης, και ταυτόχρονα να διατηρηθούν οι σπουδαιότερες ιδιότητές τους.

Πριν προχωρήσουμε στην ανάλυση της ανθεκτικότητας στην παλινδρόμηση, αξίζει να αναφέρουμε ότι γενικά υπάρχουν δύο τρόποι αντιμετώπισης του προβλήματος. Ο πρώτος είναι η **ανίχνευση** (αναγνώριση) των outliers και η **απόρριψή** τους από το δείγμα, και ο δεύτερος η ανθεκτική εκτίμηση. Σκοπός της ανίχνευσης και απόρριψης των outliers δεν είναι αποκλειστικά η εκτίμηση των παραμέτρων, αλλά ενδεχομένως η αναγνώριση του μοντέλου παλινδρόμησης.

Όλες οι ανθεκτικές μέθοδοι εκτίμησης μπορούν να χρησιμοποιηθούν έμμεσα και ως μέθοδοι αναγνώρισης των outliers από τις ελαφρύνσεις που επιβάλλουν στα κατάλοιπα. Στις πιο σύγχρονες εκτιμήσεις προηγείται η ανίχνευση των πιθανών outliers (potential outliers) και ακολουθεί η ανθεκτική εκτίμηση με πιο αξιόπιστες αρχικές παραμέτρους.

16.1 OUTLIERS ΣΤΗΝ ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Η ανάλυση παλινδρόμησης, όπως αναπτύχθηκε και στο κεφάλαιο 13, είναι μία από τις πιο συνηθισμένες τεχνικές στους μηχανικούς, οικονομολόγους και άλλους επιστήμονες. Θεωρούμε το μοντέλο παλινδρόμησης

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + u$$

$$\acute{\eta} \qquad y = \boldsymbol{x}^T \boldsymbol{\beta} + u \tag{16.1}$$

όπου:

- y είναι η εξαρτημένη μεταβλητή,
- $\boldsymbol{x}$ Είναι ένα $p \times 1$ διάνυσμα των ανεξάρτητων μεταβλητών,

$(x_1, x_2, ..., x_p)$

- $\boldsymbol{\beta}$ είναι ένα $p \times 1$ διάνυσμα των αγνώστων παραμέτρων $(\beta_1, \beta_2, ..., \beta_p)$
- u είναι το τυχαίο σφάλμα κανονικής κατανομής, $N(0, \sigma^2)$.

Παρατηρούμε ένα δείγμα $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$ και επιθυμούμε την εκτίμηση για την παράμετρο $\boldsymbol{\beta}$. Η κλασσική μέθοδος στηρίζεται στα ελάχιστα τετράγωνα (*LS*), όπου η εκτίμηση $\hat{\boldsymbol{\beta}}$ υπολογίζεται από τη λύση του προβλήματος.

$$\underset{\hat{\boldsymbol{\beta}}}{\text{ελαχιστοποίηση}} \sum_{i=1}^{n} (y_i - x_i^T \hat{\boldsymbol{\beta}})^2 = \sum_{i=1}^{n} r_i^2 \qquad (16.2)$$

Βασικά, το πρόβλημα εκτίμησης της παραμέτρου β λύνεται με την ελαχιστοποίηση του παραπάνω αθροίσματος των τετραγώνων των καταλοίπων r_i^2, ή ισοδύναμα με τη λύση των κανονικών εξισώσεων που επιτυγχάνονται από το διαφορικό του αθροίσματος ως προς $\hat{\boldsymbol{\beta}}$ και εξισώνοντας τις μερικές παραγώγους ως προς μηδέν,

$$\sum_{i=1}^{n} (y_i - \hat{\boldsymbol{\beta}} x_i) x_{ij} = \sum_{i=1}^{n} (r_i) x_{ij} = 0 \quad \gamma\iota\alpha \quad j = 1, ..., p \qquad (16.3)$$

Όπως φαίνεται και στα Σχήματα 16.1, 16.2, 16.3, 16.4, 16.5 είναι παραδεκτό ότι τα ελάχιστα τετράγωνα είναι πολύ ευαίσθητα στην απόκλιση των βασικών προϋποθέσεων του μοντέλου $u \sim N(0, \sigma^2)$. Αυτές οι αποκλίσεις (outliers) εκφράζονται με ακραίες τιμές των παρατηρήσεων $(\boldsymbol{x}_i, y_i)$, είτε ως προς $\boldsymbol{x}_i$ ή ως προς y_i. Αυτό έχει ως αποτέλεσμα τη δημιουργία μεγάλων καταλοίπων r_i και το τετράγωνό τους αυξάνει εκθετικά την αντικειμενική συνάρτηση (16.2). Για να αποφευχθεί αυτό μεταβάλλεται η τιμή της παραμέτρου $\boldsymbol{\beta}$. Παρόμοια, προκαλείται δυσμενής επίδραση των μεγάλων καταλοίπων και στη λύση του συστήματος 16.3. Ακόμη, ένα μόνο outlier θα μπορούσε να έχει καταστροφική επίδραση στην εκτίμηση των συντελεστών της παλινδρόμησης με ελάχιστα τετράγωνα, και να οδηγήσει σ' ένα μοντέλο, το οποίο δεν αντιπροσωπεύει την πληθώρα των δεδομένων.

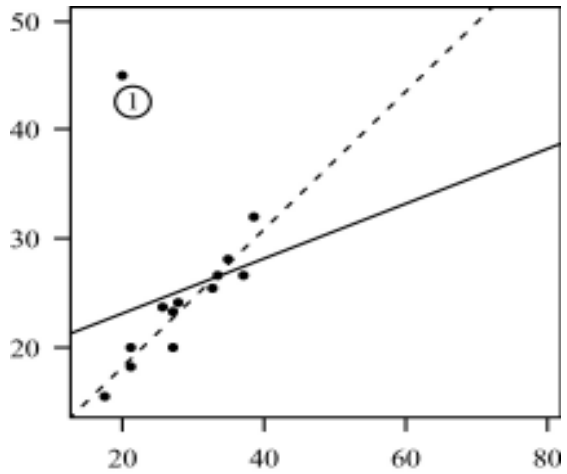
Τύποι outliers

Διακρίνουμε δύο τύπους outliers:

1. ***y-outliers***. Αυτά τα outliers είναι παρατηρήσεις, οι οποίες αποκλίνουν από την πληθώρα των παρατηρήσεων, επειδή παρουσιάζουν μη φυσιολογική τιμή στην εξαρτημένη μεταβλητή. Τα αποκαλούμε outli-

ers στην ***y*-κατεύθυνση** ή απλώς ***y*-outliers**. Αυτός ο τύπος των outliers συχνά παρατηρείται όταν οι τιμές των ανεξάρτητων μεταβλητών έχουν σίγουρα τις φυσιολογικές τους τιμές.

Επίδραση. Τέτοια outliers επιδρούν στην εκτίμηση της παλινδρόμησης, καθώς αυξάνουν και το μέγεθος των καταλοίπων, αλλά δεν επιφέρουν κάποια συνταρακτική μεταβολή στην εκτίμηση των παραμέτρων της. Γιατί, όπως αναφέρουμε παρακάτω, υπάρχουν outliers με δραματικές επιδράσεις. Για παράδειγμα, το *y*-outlier στο Σχήμα 16.1 έχει σημαντική επίδραση στην κλασική εκτίμηση της γραμμής των ελαχίστων τετραγώνων και στην σταθερά και στην κλίση και στην εκτίμηση της τυπικής απόκλισης σ των καταλοίπων.



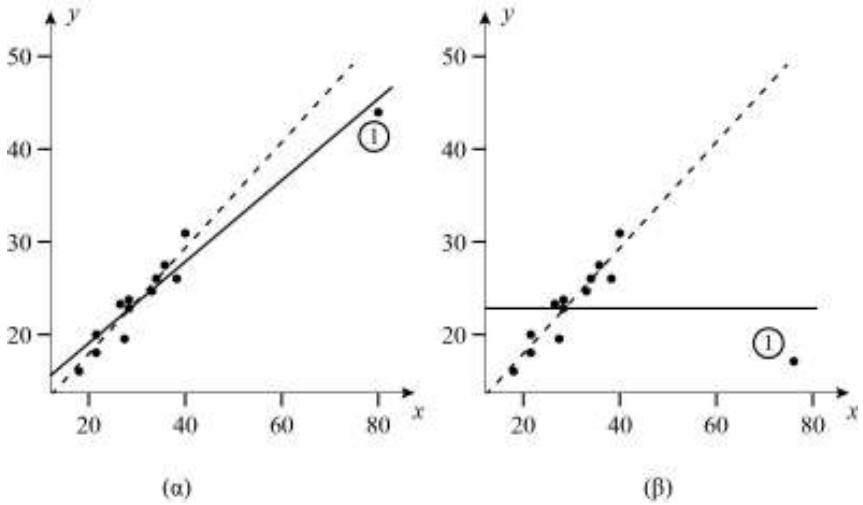
Σχήμα 16.1 Προσαρμογή γραμμής με ελάχιστα τετράγωνα σε 13 παρατηρήσεις με ένα outlier *y*-κατεύθυνσης. Η διακεκομμένη γραμμή είναι χωρίς το outlier

2. ***x*-outliers**. Τέτοια outliers συμβαίνουν όταν η τιμή x μιας παρατήρησης της ανεξάρτητης μεταβλητής απέχει κατά πολύ από την πληθώρα των τιμών της ανεξάρτητης μεταβλητής X, ή η απόσταση μιας παρατήρησης $\boldsymbol{x}$ του διανύσματος των ανεξάρτητων μεταβλητών απέχει κατά πολύ από τον πίνακα των τιμών των ανεξάρτητων μεταβλητών, βλέπε Σχήμα 16.2.

Ένα outlier προς την *x*-κατεύθυνση έχει μεγάλη επίδραση στον εκτιμητή των ελαχίστων τετραγώνων, επειδή πρακτικά προσελκύει τη γραμμή *LS*, γνωστό ως **μοχλός** (leverage point).

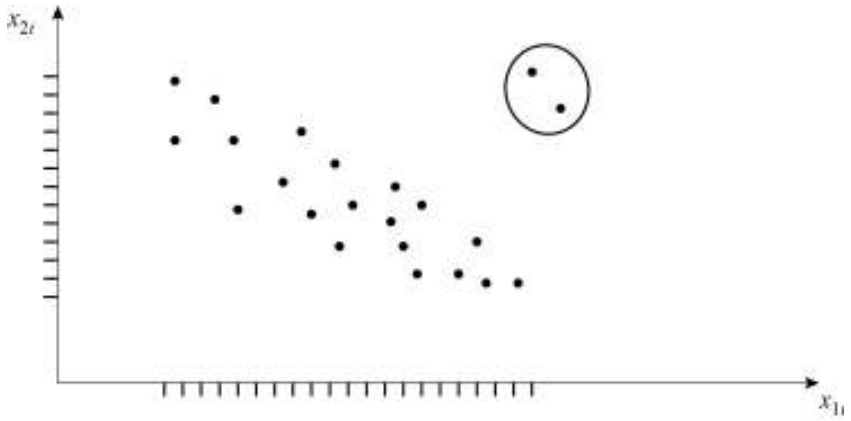
Ένα x-outlier ή leverage point όπως διαφορετικά λέγεται, δεν είναι κατ' ανάγκη ένα outlier παλινδρόμησης. Γι' αυτό τα x-outliers διακρίνονται σε δύο κατηγορίες, τους καλούς μοχλούς (Good leverage) και κακούς μοχλούς (Bad leverage).

- ***Good leverage points.*** Στην περίπτωση αυτή, τα σημεία αυτά συμβάλλουν στην εκτίμηση της γραμμής της παλινδρόμησης, όπως φαίνεται και στο Σχήμα 16.2(α). Το σημείο αυτό συμφωνεί με την πληθώρα των άλλων σημείων ως προς την κατεύθυνση της παλινδρόμησης. Εάν, ένα good leverage point απομακρυνθεί από τις παρατηρήσεις, η γραμμή των ελαχίστων τετραγώνων δεν μεταβάλλεται σημαντικά.
- ***Bad leverage points.*** Αντίθετα με τα good leverage points τα σημεία αυτά δεν συμφωνούν με την πληθώρα των παρατηρήσεων, μεταβάλλουν σημαντικά τη γραμμή, και είναι κακοί μοχλοί. Όπως φαίνεται και στο Σχήμα 16.2(β), η διακεκομμένη γραμμή είναι η εκτίμηση της γραμμής χωρίς το σημείο 1. Εάν στην εκτίμηση συμμετέχει και το 1, η γραμμή (συνεχής) αλλάζει δραματικά κατεύθυνση.



Σχήμα 16.2 Τα ίδια δεδομένα όπως στο Σχήμα 16.1, αλλά στο σχήμα (α) το σημείο 1 είναι good leverage point, στο σχήμα (β) το σημείο 1 είναι bad leverage point. Οι γραμμές προσδιορίζονται από τα ελάχιστα τετράγωνα με όλα τα δεδομένα (συνεχής γραμμή), και μετά την απάλειψη του 1 (διακεκομμένη γραμμή)

Στην πολλαπλή παλινδρόμηση, όπου η ανεξάρτητη μεταβλητή $\boldsymbol{x}$ βρίσκεται σε χώρο p διάστασης, $(x_1, x_2, ..., x_p)$, ένα σημείο μοχλός είναι ένα σημείο $(x_{1i}, x_{2i}, ..., x_{pi})$, το οποίο απέχει πολύ από το κέντρο της μάζας του πίνακα X των ανεξάρτητων μεταβλητών. Όπως και πριν ένα τέτοιο σημείο έχει μεγάλη επίδραση στο πολυεπίπεδο *LS*, η οποία βέβαια εξαρτάται και από την τιμή y_i. Παρόλα αυτά, τέτοια σημεία μοχλούς είναι δύσκολο να τα αναγνωρίσουμε όταν υπάρχουν για παράδειγμα 10 ανεξάρτητες μεταβλητές. Σαν μία απλή παράσταση του προβλήματος βλέπε Σχήμα 16.3, το οποίο είναι



Σχήμα 16.3 Σχεδιάγραμμα των δύο ανεξάρτητων μεταβλητών από ένα σύνολο δεδομένων παλινδρόμησης. Υπάρχουν δύο μοχλοί (μέσα στον κύκλο), οι οποίοι δεν έχουν ακραίες τιμές ως προς έκαστο των αξόνων.

διάγραμμα διασποράς x_1 ενάντια x_2 για κάποιο σύνολο δεδομένων. Παρατηρούμε ότι οι τιμές $(x_{21}, x_{22}, x_{23}, ..., x_{2n})$ δεν εμπεριέχουν κανένα outlier, παρόμοια και για τις τιμές $(x_{11}, x_{12}, x_{13}, ..., x_{1n})$ της x_1. Παρόλα αυτά το ζευγάρι της i παρατήρησης (x_{1i}, x_{2i}) μέσα στον κύκλο βρίσκεται μακριά από την πληθώρα των άλλων ζευγαριών, και αποτελεί ένα μοχλό στην παλινδρόμηση. Γενικά, η αναγνώριση μιας παρατήρησης $(x_{1i}, x_{2i}, ..., x_{pi})$ ως x-outlier είναι ένα δύσκολο πρόβλημα, το οποίο αντιμετωπίζεται με τα **διαγνωστικά**, όπως περιγράφονται στην επόμενη παράγραφο. Παρόλα αυτά, σε αυτό το κεφάλαιο αναφερόμαστε κυρίως στα **outliers παλινδρόμησης**, δηλαδή, περιπτώσεις για τις οποίες το $(x_{1i}, x_{2i}, ..., x_{pi}, y_i)$ αποκλίνει από το γραμμικό μοντέλο παλινδρόμησης που ακολουθεί ο κύριος όγκος των δεδομένων,

λαμβάνοντας υπόψη ταυτόχρονα και τα δύο τις ανεξάρτητες μεταβλητές και την εξαρτημένη μεταβλητή.
Συνήθως, το ποσοστό των σημείων σε ένα δείγμα δεδομένων, τα οποία δεν ακολουθούν την υποτιθέμενη κανονική κατανομή, π.χ. την κανονική, ονομάζεται **ποσοστό μόλυνσης**, και επιδρά δυσμενώς στην ανάλυση της παλινδρόμησης. Το ποσοστό μόλυνσης ενδέχεται να φθάσει μέχρι και 50%.

Παρατηρήσεις

Α) Τα outliers είναι συνήθως παρατηρήσεις με μεγάλη επίδραση στη γραμμή (πολυεπίπεδο) παλινδρόμησης (influential points), δηλαδή η διαγραφή τους από το δείγμα προκαλεί σημαντικές αλλαγές στις εκτιμήσεις των συντελεστών παλινδρόμησης. Καθώς οι τιμές και η συχνότητα των outliers διαφέρει σημαντικά από δείγμα σε δείγμα, τα outliers κάνουν αναξιόπιστα τα συμπεράσματα μιας στατιστικής ανάλυσης.

Β) Τα **bad leverage points** είναι πιο **καταστροφικά** από τα *y*-outliers. Πράγματι,

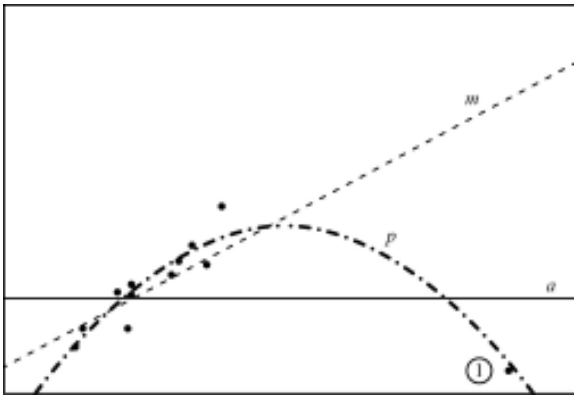
1. Αυτά είναι τυπικά με δυνατότητα μεγάλης επίδρασης.
2. Όταν εφαρμόζουμε τα ελάχιστα τετράγωνα, η προκύπτουσα γραμμή περνά κοντά από αυτά τα σημεία και έτσι είναι δύσκολη η ανίχνευσή τους με την γραφική παράσταση των καταλοίπων. Ειδικά, για τα πολυμεταβλητά μοντέλα, η ανίχνευση των *x*-outliers είναι ακόμη δυσκολότερη. ■

Επιλογή μοντέλου με την παρουσία outliers

Υποτίθεται ότι τα δεδομένα στο Σχήμα παριστάνουν τη συσχέτιση μεταξύ εξόδων διακοπών και εισοδήματος από 13 οικογένειες. Το σημείο 1 αντιστοιχεί σε κάποιον που δουλεύει σκληρά και έχει μεγάλο εισόδημα αλλά δεν ευκαιρεί να διασκεδάσει τις διακοπές του. Μπορούμε να θεωρήσουμε τρεις γραμμές σαν μοντέλο παλινδρόμησης:

- μία ευθεία γραμμή προσαρμοζόμενη σε όλα τα δεδομένα (α)
- μία ευθεία γραμμή προσαρμοζόμενη στα περισσότερα δεδομένα, αγνοώντας το σημείο 1 (m)
- μία parabola προσαρμοζόμενη σε όλα τα δεδομένα (p).

Η επιλογή εξαρτάται από πολλά πράγματα, μεταξύ άλλων και ο σκοπός της περιγραφής και ο βαθμός αξιοπιστίας του 1 σαν αντιπροσωπευτικό από μία ομογενή και υπολογίσιμη μειοψηφία.



Σχήμα 16.4 Επιλογές μοντέλου με την παρουσία ενός x-outlier

Συμπερασματικά, τα outliers στην παλινδρόμηση δημιουργούν σοβαρό πρόβλημα στην *LS* εκτίμηση. Γι' αυτό πρέπει να ανιχνεύσουμε τα outliers σε ένα σύνολο παρατηρήσεων και να περιορίσουμε την επίδρασή τους στην αναγνώριση ή εκτίμηση των παραμέτρων του μοντέλου παλινδρόμησης, είτε διαγράφοντας τα σημεία αυτά από το δείγμα ή προσελκύοντας αυτά προς την γραμμή παλινδρόμησης.

16.2 ΑΝΙΧΝΕΥΣΗ ΤΩΝ OUTLIERS – ΔΙΑΓΝΩΣΤΙΚΑ ΓΙΑ ΤΑ OUTLIERS

Μία πολύ γνωστή μέθοδος ανίχνευσης των outliers στα δεδομένα μιας παλινδρόμησης είναι τα **διαγνωστικά παλινδρόμησης**. Διαγνωστικά είναι συγκεκριμένα στατιστικά, τα οποία υπολογίζονται από τα δεδομένα, με σκοπό την υπόδειξη των σημείων με σημαντική επιρροή (influential points). Στη συνέχεια, αυτά απομακρύνονται (διαγράφονται) και η ανάλυση των ελαχίστων τετραγώνων ακολουθεί στα παραμένοντα δεδομένα. Όταν υπάρχει στα δεδομένα ένα ή λίγα outliers, μερικά από τα διαγνωστικά λειτουργούν πολύ καλά στην αναγνώριση των outliers, διαγράφοντας ένα προς ένα τα σημεία και μετρώντας την επίδρασή τους κάθε φορά. Ατυχώς, όταν υπάρχουν περισσότερα outliers στα δεδομένα και μάλιστα **πολλαπλά outliers**, προς την ίδια κατεύθυνση, τότε μπορεί να είναι **επικαλυπτόμενα** (masked), και τα διαγνωστικά αποτυγχάνουν στην ανίχνευση. Η απομάκρυνση ενός μόνον outlier από μία ομάδα πολλαπλών outliers πολύ πιθανόν να μην προκαλεί σημαντική αλλαγή στα μετρούμενα διαγνωστικά λόγω του προβλήματος επικάλυψης (masking problem).

Παρόλα αυτά, νέες ανθεκτικές τεχνικές έχουν αναπτυχθεί για την εκτίμηση αξιόπιστων διαγνωστικών ακόμη και στην περίπτωση πολλών πολλαπλών outliers. Οι τεχνικές αυτές σχετικά με τα ανθεκτικά διαγνωστικά περιγράφονται στο επόμενο κεφάλαιο.

Πολλές γραφικές ή αριθμητικές διαδικασίες για τα διαγνωστικά της παλινδρόμησης στηρίζονται στην προσαρμογή του μοντέλου παλινδρόμησης στα δεδομένα με τα ελάχιστα τετράγωνα. Πολλά διαγνωστικά στηρίζονται στα κατάλοιπα που προκύπτουν από την *LS* προσαρμογή. Αλλά αυτό μπορεί να οδηγήσει παραπλανητικά αποτελέσματα, αφού τα *LS* αποφεύγουν τα μεγάλα κατάλοιπα και πιθανόν να δημιουργούν φτωχή προσαρμογή στην πληθώρα των δεδομένων. Γι' αυτό, ένα outlier μπορεί να έχει ένα μικρό *LS* κατάλοιπο, ειδικά όταν είναι σημείο μοχλός. Κατά συνέπεια, τα διαγνωστικά που προκύπτουν από τα *LS* κατάλοιπα συχνά αποτυγχάνουν να αναγνωρίσουν τέτοια σημεία.

Μία άλλη κλάση διαγνωστικών στηρίζεται στην αρχή της διαγραφής μιας παρατήρησης κάθε φορά. Για παράδειγμα, $\hat{\boldsymbol{\theta}}(i)$ είναι η εκτίμηση του $\boldsymbol{\theta}$, υπολογιζόμενη από το δείγμα χωρίς την i παρατήρηση. Τότε, το στατιστικό $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(i)$ δίνει το μέγεθος της επίδρασης της i παρατήρησης στους συντελεστές της παλινδρόμησης. Αυτά τα στατιστικά ονομάζονται **απλά διαγνωστικά**, τα οποία υπολογίζονται για κάθε παρατήρηση i του δείγματος. Γενικεύοντας, είναι δυνατόν να υπολογίσουμε τα **πολλαπλά διαγνωστικά** για να ανιχνεύσουμε την επίδραση ομάδας πολλαπλών outliers. Στην περίπτωση αυτή, διαγράφονται όλες οι δυνατές ομάδες για να εκτιμηθεί η αντίστοιχη επίδρασή τους. Ενδέχεται κάποια σημεία ομαδικά να έχουν σημαντική επίδραση, αλλά από μόνα τους να μην έχουν. Η υπολογιστική διαδικασία εδώ είναι δύσκολη λόγω του μεγάλου αριθμού των υποομάδων που πρέπει να διαγράψουμε.

Τα περισσότερα διαγνωστικά που ακολουθούν, είναι συνάρτηση του βαθμού μοχλότητας h_i του κάθε σημείου $(\boldsymbol{x}_i, y_i)$ των δεδομένων. Τα στατιστικά h_i είναι τα διαγώνια στοιχεία του γνωστού πίνακα H (hat matrix) και παριστάνουν τη μοχλότητα (leverage) των $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ..., \boldsymbol{x}_n$, όπως περιγράφεται παρακάτω.

16.2.1 Ο Πίνακας Hat (H) – Διαγνωστικά Μοχλότητας (leverage)

Θεωρούμε την εξίσωση γραμμικής παλινδρόμησης (16.1) και τον πίνακα X των ανεξάρτητων μεταβλητών,

$$X = \begin{bmatrix} x_{11} & x_{12}, & \cdots, & x_{1p} \\ x_{21} & x_{22}, & \cdots, & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2}, & \cdots, & x_{np} \end{bmatrix} \tag{16.4}$$

Ορίζουμε ως πίνακα Hat τον πίνακα H, ο οποίος προκύπτει ως εξής:

$$H = X(X^T X)^{-1} X^T \tag{16.5}$$

με την υπόθεση ότι $X^T X$ είναι αντιστρέψιμο. Ο πίνακας $n \times n$ H, καλείται ο πίνακας Hat (καπέλο), επειδή μετασχηματίζει το παρατηρούμενο διάνυσμα $\boldsymbol{y} = \{y_1, y_2, ..., y_n\}^T$ σε εκτιμούμενο με *LS* (βλέπε εξίσωση (13.34))

$$\begin{aligned} \hat{\boldsymbol{y}} &= X\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{y}} &= X(X^T X)^{-1} X^T \boldsymbol{y} \\ \hat{\boldsymbol{y}} &= H\boldsymbol{y} \end{aligned} \tag{16.6}$$

Από την 16.5 προκύπτει ότι:

$$\begin{aligned} &\textit{ίχνος } H = p \\ \textit{ή} \quad &\sum_{i=1}^{n} h_{ii} = p \end{aligned} \tag{16.7}$$

$$h_{ii} = \sum_{i=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{i \neq j} h_{ij}^2 \quad \textit{για κάθε } i \tag{16.8}$$

Τα h_{ii} είναι τα διαγώνια στοιχεία του πίνακα H και συνήθως γράφονται ως h_i, και υπολογίζονται από την εξίσωση:

$$h_i = \boldsymbol{x}_i^T (X^T X)^{-1} \boldsymbol{x}_i \tag{16.9}$$

Από την εξίσωση (16.8) έπεται ότι το h_i παίρνει τιμές από μηδέν έως 1, $0 \leq h_i \leq 1$. Τα στοιχεία h_{ij}, όπως προκύπτει από την εξίσωση (16.6) παριστάνουν τη συμβολή της j παρατήρησης στην εκτίμηση $\hat{y}_i$. Ειδικότερα, τα διαγώνια στοιχεία του H μπορεί εύκολα να αποδειχθεί ότι προκύπτουν από την παραγώγιση,

$$h_i = \frac{\partial \hat{y}_i}{\partial y_i} \tag{16.10}$$

και ως εκ τούτου παριστάνουν την επιρροή της i παρατήρησης y_i στην εκτίμησή της $\hat{y}_i$, $\hat{y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$. Γι' αυτό η τιμή της h_i ερμηνεύεται και ως μοχλότητα (leverage) της i παρατήρησης. Η μοχλότητα h_i όπως προκύπτει από την (16.9) είναι συνάρτηση μόνο των ανεξάρτητων μεταβλητών x και εκφράζει την απόσταση της x_i από το κέντρο μάζας του πίνακα X. Η μέση τιμή των h_i είναι p/n, γι' αυτό ένα σημείο με $h_i > 2p/n$ (ή $3p/n$) έχει δυνατότητα μεγάλης επίδρασης στην εκτίμηση LS των συντελεστών της παλινδρόμησης.

Παρατήρηση

Στην περίπτωση όπου το μοντέλο παλινδρόμησης (16.1) έχει σταθερό όρο (β_0), χρησιμοποιείται η ίδια άλγεβρα πινάκων, για την εκτίμηση του h_i, όπου ο πίνακας X των ανεξάρτητων μεταβλητών εμπεριέχει την τεχνητή ανεξάρτητη μεταβλητή x_p, η οποία ταυτίζεται με την τιμή 1. Δηλαδή,

$$X = \begin{bmatrix} x_{1,1},\ldots,x_{1,p-1},1 \\ x_{2,1},\ldots,x_{2,p-1},1 \\ \vdots \\ x_{n,1},\ldots,x_{n,p-1},1 \end{bmatrix} \tag{16.11}$$

και ο πίνακας H ορίζεται πάλι από την (16.5) με το νέο X.

Ακόμη, στην περίπτωση παλινδρόμησης με σταθερό όρο, μπορεί κανείς να μετρήσει τη μοχλότητα της κάθε παρατήρησης με την **Mahalanobis απόσταση** MD_i^2,

$$MD_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T C^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{16.12}$$

όπου C είναι ο πίνακας συνδιακύμανσης (covariance) των ανεξάρτητων μεταβλητών $\boldsymbol{x}$,

$$C = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}})$$

όπου $\boldsymbol{x} = (x_1, x_2, \ldots, x_{p-1})$. Κάνοντας πράξεις στην 16.12 προκύπτει τελικά η σχέση μεταξύ της Mahalanobis απόστασης και της μοχλότητας h_i,

$$MD_i^2 = (n-1)\left[h_i - \frac{1}{n}\right] \tag{16.13}$$

Η Mahalanobis απόσταση, όπως φαίνεται και από την (16.12) μετρά την απόσταση της κάθε παρατήρησης $\boldsymbol{x}_i$ από τη μέση τιμή $\bar{\boldsymbol{x}}$.

Σκοπός της μοχλότητας h_i και της Mahalanobis απόστασης MD_i^2 είναι να υποδείξουν ποιες παρατηρήσεις $\boldsymbol{x}_i$ βρίσκονται μακριά από τον κύριο όγκο των δεδομένων των ανεξάρτητων μεταβλητών. Οι τιμές h_i και MD_i^2 λειτουργούν ως διαγνωστικά για την ανίχνευση των x-outliers σε ένα σύνολο δεδομένων. Τα διαγνωστικά αυτά υποδεικνύουν πόσο μοχλότητα (leverage) έχουν οι παρατηρήσεις $\boldsymbol{x}_i$.

Ο πίνακας Hat είναι χρήσιμος όταν τα δεδομένα περιέχουν μόνο ένα ή το πολύ μερικά x-outliers. Παρόλα αυτά, όταν υπάρχουν παραπάνω x-outliers, ενδέχεται η αντίστοιχη τιμή h_i να μην είναι κατάλληλα αρκετή για να υποδείξει «μοχλότητα» στα αντίστοιχα σημεία. Ο λόγος είναι ότι στην (16.12) και (16.9) οι πίνακες συνδιακύμανσης C ή X^TX δεν είναι ανθεκτικοί. Προκειμένου να ανιχνεύσουμε τα x-outliers, θα ήταν καλύτερα να εισαχθούν στις (16.12) και (16.9) ανθεκτικοί πίνακες συνδιακύμανσης C ή X^TX. Περισσότερα για την διαδικασία ανθεκτικοποίησης αυτών των πινάκων ή για την ανθεκτική εκτίμηση των h_i ή MD_i^2 περιγράφονται στο επόμενο κεφάλαιο.

Παράδειγμα 16.1

Στον Πίνακα 16.1 παρατίθενται τα τεχνητά δεδομένα $(y_i, x_{1i}, x_{2i},\ i = 1,...,25)$ για ένα γραμμικό μοντέλο παλινδρόμησης

$$y_i = 1,20x_{1i} - 0,80x_{2i} + u_i$$

όπου οι ανεξάρτητες ακολουθούν κανονική κατανομή, $x_1 \sim N(10,\ 8^2)$, $x_2 \sim N(20,\ 10^2)$, όπως και τα στοχαστικά σφάλματα, $u \sim N(0,\ 16^2)$. Η πρώτη γραμμή του πίνακα (α) των δεδομένων είναι ένα x-outlier το οποίο προφανώς αναγνωρίζεται και από τη μεγάλη τιμή της μοχλότητας h_1. Αυτό όμως δεν συμβαίνει όταν υπάρχει μεγαλύτερη «μόλυνση» με 5 x-outliers στον πίνακα (β). Όπως φαίνεται και από τις τιμές h_1, h_2, h_3, h_4, h_5 δεν διαφέρουν σημαντικά από τη μοχλότητα h_i των υπόλοιπων παρατηρήσεων. Τα πέντε x-outliers δεν ανιχνεύονται από την τιμή της αντίστοιχης μοχλότητας.

(α)					(β)				
i	y_i	x_{1i}	x_{2i}	h_i	i	y_i	x_{1i}	x_{2i}	h_i
1	129,31	-3,65	-1,39	**0,728**	1	143,51	29,33	2,22	**0,155**
2	31,30	57,54	-39,99	0,045	2	143,52	-3,85	37,73	0,147
3	21,39	12,39	-3,15	0,058	3	5,76	105,76	-39,89	0,128
4	21,79	45,53	47,12	0,009	4	243,23	-19,83	307,70	0,552
5	24,87	47,70	-12,83	0,011	5	29,23	189,73	-118,67	0,609
6	16,69	-5,62	16,35	0,181	6	24,24	61,99	-32,61	0,022
7	23,08	14,42	6,56	0,046	7	20,65	-8,20	15,46	0,049
8	37,73	39,40	-15,70	0,009	8	3,88	16,74	-25,22	0,028
9	24,85	68,44	-59,17	0,084	9	25,32	42,50	-3,74	0,004
10	24,21	37,52	47,28	0,002	10	43,23	-,69	60,92	0,023
11	31,32	35,47	-10,25	0,000	11	13,44	46,73	-7,03	0,011
12	26,35	35,06	4,71	0,000	12	19,45	5,65	35,21	0,026
13	40,33	07,43	33,56	0,053	13	29,62	23,42	46,10	0,003
14	35,77	46,16	-14,42	0,018	14	4,48	27,99	00,62	0,018
15	30,43	35,59	00,38	0,000	15	17,33	26,05	-45,03	0,008
16	15,87	63,46	-15,55	0,058	16	-2,40	00,55	1,48	0,060
17	37,93	-12,88	26,87	0,168	17	-1,95	32,70	-40,81	0,023
18	00,81	46,91	-34,21	0,060	18	11,41	20,68	17,05	0,016
19	13,27	31,89	310,91	0,026	19	38,65	40,00	311,37	0,002
20	34,57	42,56	320,88	0,009	20	23,21	9,13	-2,26	0,018
21	24,57	19,62	-3,46	0,025	21	54,60	-20,08	372,24	0,058
22	53,07	38,00	335,46	0,052	22	36,20	42,29	319,08	0,003
23	02,05	23,28	-24,58	0,095	23	21,53	16,10	-16,72	0,012
24	24,67	55,44	-20,52	0,030	24	17,49	42,98	-15,05	0,007
25	10,88	90,96	-36,33	0,222	25	17,04	36,37	-1,49	0,006

Πίνακας 16.1 Τεχνητά δεδομένα (α) με ένα x-outlier, (β) με πέντε x-outliers, και οι τιμές μοχλότητας h_i

16.2.2 Διαγνωστικά Καταλοίπων

Η επίδραση μιας παρατήρησης $(\boldsymbol{x}_i, y_i)$ στην εκτίμηση *LS* εξαρτάται και από τα δύο, από την πολύ μεγάλη ή μικρή τιμή της y_i συγκρινόμενη με y από παρόμοια $\boldsymbol{x}$ και από την μοχλότητα του x_i. Τα διαγνωστικά μοχλότητας όπως h_i από μόνα τους δεν είναι αρκετά για την ανίχνευση των outliers στην παλινδρόμηση, διότι δεν λαμβάνουν υπόψη το y_i. Είναι παραδεκτό ότι τα κατάλοιπα *LS* (ελαχίστων τετραγώνων) κατέχουν όλη τη συμπληρωματική πληροφορία. Επικεντρωνόμαστε περισσότερο στους ακόλουθους τύπους καταλοίπων:

1) τυποποιημένα (standardized) κατάλοιπα
2) Studentized κατάλοιπα

Τα πιο δημοφιλή στατιστικά προγράμματα όπως το SPSS και SAS υπολογίζουν αυτά τα στατιστικά στην ανάλυση παλινδρόμησης.

1) Τα **τυποποιημένα κατάλοιπα** ορίζονται ως

$$\frac{r_i}{s} \tag{16.14}$$

όπου s^2 δίνεται από το άθροισμα:

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 \tag{16.15}$$

Όταν τα σφάλματα στην y (βλέπε εξίσωση 16.1) είναι ανεξάρτητα κανονικής κατανομής με μηδέν μέση τιμή και τυπική απόκλιση σ, είναι γνωστό ότι s^2 είναι μία αμερόληπτη εκτίμηση του σ^2.

2) Το **stutentized κατάλοιπο** είναι επίσης τυποποιημένο κατάλοιπο, όπως στην (16.14), αλλά ως τυπική απόκλιση στον παρονομαστή παίρνουμε την τιμή απόκλιση του κατάλοιπου $\hat{r}_i$ της LS εκτίμησης,

$$\begin{aligned} Var(r_i) &= Var(y_i - \hat{y}_i) \\ &= Var(y_i) - Var(\hat{y}_i) \\ &= \sigma^2 - \sigma^2 h_i \end{aligned}$$

όπως προκύπτει και από την (13.39). Άρα η τυπική απόκλιση του $\hat{r}_i$ είναι

$$\boldsymbol{\sigma}_i = \sqrt{Var(\hat{r}_i)} = \sigma\sqrt{1-h_i} \tag{16.16}$$

και κατά συνέπεια μια δειγματική εκτίμηση για το **studendized κατάλοιπο** είναι

$$t_i = \frac{r_i}{s\sqrt{1-h_i}} \tag{16.17}$$

Παρατήρηση

Είναι πολύ σημαντικό να τονίσουμε ότι, στην εκτίμηση των συντελεστών παλινδρόμησης με ελάχιστα τετράγωνα η τυπική απόκλιση του εκτιμούμενου κατάλοιπου $r_i = y_i - \hat{y}_i$ εξαρτάται και από τη μοχλότητα h_i του σημείου $(\boldsymbol{x}_i, y_i)$. Τα σημεία με μεγάλο h_i έλκουν τη γραμμή παλινδρόμησης προς το

μέρος τους με αποτέλεσμα τα προκύπτοντα r_i να είναι μικρά. Συμπερασματικά, σε παρατηρήσεις με μεγάλη μοχλότητα τα κατάλοιπα πρέπει να είναι μικρότερα από ότι στα άλλα δεδομένα, διαφορετικά οι παρατηρήσεις αυτές ενδέχεται να είναι outliers. Η σύγκριση του μεγέθους των καταλοίπων πραγματοποιείται με το studentized κατάλοιπο t_i όπως στην (16.17).
Ένας αποτελεσματικότερος τύπος του studentized καταλοίπου είναι:

$$t_{(i)} = \frac{r_i}{s_{(i)}\sqrt{1-h_i}} \tag{16.18}$$

όπου $s_{(i)}$ είναι η δειγματική τυπική απόκλιση των καταλοίπων χωρίς την i παρατήρηση $(\boldsymbol{x}_i, y_i)$ στο δείγμα. Κάτω από την υπόθεση της κανονικότητας, το $t_{(i)}$ ακολουθεί την t-student κατανομή με n-1 βαθμούς ελευθερίας. Έτσι ένας έλεγχος για το μέγεθος του καταλοίπου μπορεί να πραγματοποιηθεί από την ανίσωση $\left|t_{(i)}\right| > t_{n-1,(1-a/2)}$.

Ενώ τα διαγνωστικά καταλοίπων $t_{(i)}$ μπορούν να ανιχνεύσουν outliers σε απλές περιπτώσεις, όταν υπάρχουν πολλαπλά outliers στο δείγμα των δεδομένων αποτυγχάνουν. Η αιτία γι' αυτό είναι ότι οι εκτιμήσεις των r_i, h_i και s ενδέχεται να επηρεάζονται δυσμενώς από τα πολλαπλά outliers. Στην επόμενη υποπαράγραφο περιγράφεται μία ασφαλέστερη προσέγγιση διαγνωστικών.

Παράδειγμα 16.2

Διερευνούμε την αποτελεσματικότητα των διαγνωστικών σε μία πολλαπλή παλινδρόμηση,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

Τα δεδομένα δημιουργήθηκαν από τους Hawkins, Bradu και Kass (1984), τα οποία φαίνονται στον Πίνακα 16.2. Είναι γνωστό, ότι οι πρώτες 10 παρατηρήσεις είναι x-outliers (bad leverage), ενώ οι επόμενες 4 περιπτώσεις (11, 12, 13 και 14) είναι σημεία καλής μοχλότητας (good leverage). Ως εκ τούτου τα πραγματικά 10 outliers επικαλύπτονται από την επίδραση των καλών σημείων μοχλότητας. Όπως φαίνεται στον Πίνακα 16.3, ο οποίος περιορίζεται στις 14 πρώτες παρατηρήσεις επειδή αυτές είναι οι πιο ενδιαφέρουσες, οι μοχλότητες h_i των 10 πρώτων παρατηρήσεων επικαλύπτονται από τις άλλες. Μόνον οι παρατηρήσεις καλής μοχλότητας (11, 12, 13 και 14) έχουν μεγάλες τιμές h_i, το ίδιο παρατηρούμε και για τις τιμές MD_i^2.
Οι τρεις τύποι των καταλοίπων έχουν μικρές τιμές παντού εκτός των καλών παρατηρήσεων 11, 12 και 13. Ατυχώς, αυτές οι παρατηρήσεις θα έπρεπε να

έχουν κατάλοιπα επειδή δεν δημιουργήθηκαν σαν outliers. Κανένα από τα διαγνωστικά δεν δείχνει τα αληθινά outliers σε αυτό το σύνολο δεδομένων.

Index	x_1	x_2	x_3	y	Index	x_1	x_2	x_3	y
1	10.1	19.6	28.3	9.7	39	2.1	0.0	1.2	-0.7
2	9.5	20.5	28.9	10.1	40	0.5	2.0	1.2	-0.5
3	10.7	20.2	31.0	10.3	41	3.4	1.6	2.9	-0.1
4	9.9	21.5	31.7	9.5	42	0.3	1.0	2.7	-0.7
5	10.3	21.1	31.1	10.0	43	0.1	3.3	0.9	0.6
6	10.8	20.4	29.2	10.0	44	1.8	0.5	3.2	-0.7
7	10.5	20.9	29.1	10.8	45	1.9	0.1	0.6	-0.5
8	9.9	19.6	28.8	10.3	46	1.8	0.5	3.0	-0.4
9	9.7	20.7	31.0	9.6	47	3.0	0.1	0.8	-0.9
10	9.3	19.7	30.3	9.9	48	3.1	1.6	3.0	0.1
11	11.0	24.0	35.0	-0.2	49	3.1	2.5	1.9	0.9
12	12.0	23.0	37.0	-0.4	50	2.1	2.8	2.9	-0.4
13	12.0	26.0	34.0	0.7	51	2.3	1.5	0.4	0.7
14	11.0	34.0	34.0	0.1	52	3.3	0.6	1.2	-0.5
15	3.4	2.9	2.1	-0.4	53	0.3	0.4	3.3	0.7
16	3.1	2.2	0.3	0.6	54	1.1	3.0	0.3	0.7
17	0.0	1.6	0.2	-0.2	55	0.5	2.4	0.9	0.0
18	2.3	1.6	2.0	0.0	56	1.8	3.2	0.9	0.1
19	0.8	2.9	1.6	0.1	57	1.8	0.7	0.7	0.7
20	3.1	3.4	2.2	0.4	58	2.4	3.4	1.5	-0.1
21	2.6	2.2	1.9	0.9	59	1.6	2.1	3.0	-0.3
22	0.4	3.2	1.9	0.3	60	0.3	1.5	3.3	-0.9
23	2.0	2.3	0.8	-0.8	61	0.4	3.4	3.0	-0.3
24	1.3	2.3	0.5	0.7	62	0.9	0.1	0.3	0.6
25	1.0	0.0	0.4	-0.3	63	1.1	2.7	0.2	-0.3
26	0.9	3.3	2.5	-0.8	64	2.8	3.0	2.9	-0.5
27	3.3	2.5	2.9	-0.7	65	2.0	0.7	2.7	0.6
28	1.8	0.8	2.0	0.3	66	0.2	1.8	0.8	-0.9
29	1.2	0.9	0.8	0.3	67	1.6	2.0	1.2	-0.7
30	1.2	0.7	3.4	-0.3	68	0.1	0.0	1.1	0.6
31	3.1	1.4	1.0	0.0	69	2.0	0.6	0.3	0.2
32	0.5	2.4	0.3	-0.4	70	1.0	2.2	2.9	0.7
33	1.5	3.1	1.5	-0.6	71	2.2	2.5	2.3	0.2
34	0.4	0.0	0.7	-0.7	72	0.6	2.0	1.5	-0.2
35	3.1	2.4	3.0	0.3	73	0.3	1.7	2.2	0.4
36	1.1	2.2	2.7	-1.0	74	0.0	2.2	1.6	-0.9
37	0.1	3.0	2.6	-0.6	75	0.3	0.4	2.6	0.2
38	1.5	1.2	0.2	0.9					

Πίνακας 16.2 Τα τεχνητά δεδομένα των Hawkins – Bradu – Kass

Index i	h_{ii} (0.107)	MD_i^2 (7.82)	r_i/s (2.50)	t_i (2.50)	$t_{(i)}$ (2.50)
1	0.063	3.674	1.50	1.55	1.57
2	0.060	3.444	1.78	1.83	1.86
3	0.086	5.353	1.33	1.40	1.41
4	0.081	4.971	1.14	1.19	1.19
5	0.073	4.411	1.36	1.41	1.42
6	0.076	4.606	1.53	1.59	1.61
7	0.068	4.042	2.01	2.08	2.13
8	0.063	3.684	1.71	1.76	1.79
9	0.080	4.934	1.20	1.26	1.26
10	0.087	5.445	1.35	1.41	1.42
11	0.094	5.986	-3.48	-3.66	-4.03
12	0.144	9.662	-4.16	-4.50	-5.29
13	0.109	7.088	-2.72	-2.88	-3.04
14	0.564	40.725	-1.69	-2.56	-2.67

Πίνακας 16.3 Οι μοχλότητες h_i, Mahalanobis αποστάσεις Md_i, τυποποιημένα κατάλοιπα r_i/s και studentized κατάλοιπα t_i, $t_{(i)}$ για τα πρώτα 14 δεδομένα των Hawking-Bradu-Kass.

16.2.3 Διαγνωστικά Επίδρασης

Η επίδραση μιας παρατήρησης $(\boldsymbol{x}_i, y_i)$ στην *LS* εκτίμηση εξαρτάται και από τα δύο τη μεγάλη ή μικρή τιμή y_i συγκρινόμενη με την y από παρόμοιο $\boldsymbol{x}$ και πόσο μεγάλο είναι το $\boldsymbol{x}_i$, π.χ. πόσο μοχλότητα έχει το $\boldsymbol{x}_i$. Τα πιο δημοφιλή διαγνωστικά για τη μέτρηση της επίδρασης της $(\boldsymbol{x}_i, y_i)$ στηρίζονται στη σύγκριση της *LS* εκτίμησης βασιζόμενη σε όλο το δείγμα με την *LS* εκτίμηση βασιζόμενη στο δείγμα χωρίς την $(\boldsymbol{x}_i, y_i)$.

Προκειμένου να επιβεβαιώσουμε την επίδραση της i παρατήρησης, είναι χρήσιμο να εκτιμήσουμε την παλινδρόμηση μαζί και χωρίς την παρατήρηση. Γι' αυτό το σκοπό, η απόσταση του Cook (1977) μετρά την αλλαγή στους συντελεστές της παλινδρόμησης που θα μπορούσε να συμβεί με την παράλειψη μιας παρατήρησης. Η απόσταση Cook ορίζεται ως

$$CD_{(i)}^2 = \frac{\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} X^T X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{(i)})}{ps^2} \tag{16.19}$$

όπου $\hat{\boldsymbol{\beta}}$ είναι η *LS* εκτίμηση του $\boldsymbol{\beta}$, και $\hat{\boldsymbol{\beta}}_{(i)}$ είναι η *LS* εκτίμηση του $\boldsymbol{\beta}$ με βάση το δείγμα χωρίς την i παρατήρηση.

Παρατήρηση

Η απόσταση Cook $CD_{(i)}^2$ είναι η τυποποιημένη τετραγωνική απόσταση που διανύει το $\boldsymbol{\beta}$, όταν εκτιμάται χωρίς την i παρατήρηση. Μια μεγάλη τιμή της $CD_{(i)}^2$ δείχνει ότι η i παρατήρηση έχει σημαντική επίδραση στον προσδιορισμό του $\hat{\boldsymbol{\beta}}$. Από την βιβλιογραφία της ανθεκτικής στατιστικής οι Cook και Weisberg (1982), προτείνουν ότι μία απόσταση $CD_{(i)}^2$ γύρω στο 1.0 θεωρείται μεγάλη. ■

Χρησιμοποιώντας τη σχέση $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$, η απόσταση Cook μπορεί να γραφεί ως

$$CD_{(i)}^2 = \frac{(\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(i)})}{ps^2} \tag{16.20}$$

Η εξίσωση (16.20) είναι ένας ισοδύναμος ορισμός, ο οποίος δείχνει ότι η απόσταση $CD_{(i)}^2$ μετρά την επίπτωση της διαγραφής της i παρατήρησης στο διάνυσμα των εκτιμούμενων τιμών $\hat{\boldsymbol{y}}$. Μια άλλη χρήσιμη φόρμα της απόστασης Cook είναι

$$CD_{(i)}^2 = \frac{r_i^2}{s^2(1-h_i)^2} \frac{h_i}{p} \tag{16.21}$$

Από αυτήν την εξίσωση έπεται ότι η απόσταση $CD_{(i)}^2$ εξαρτάται από τρεις ποσότητες, οι οποίες υπολογίζονται με βάση όλα τα δεδομένα. Οι ποσότητες αυτές είναι το κατάλοιπο r_i η τυπική απόκλιση s και η μοχλότητα h_i. Από την (16.21) έπεται ακόμη ότι μία παρατήρηση με μεγάλη μοχλότητα h_i έχει μεγαλύτερη επίδραση από μία παρατήρηση με χαμηλότερη μοχλότητα και το ίδιο κατάλοιπο.

Η απόσταση $CD_{(i)}^2$ είναι παρόμοια και με ένα διαγνωστικό το οποίο προτάθηκε από τον Belsey και άλλους (1980), όπως

$$DFFITS(i) = \frac{r_i}{s(i)} \frac{\sqrt{h_i}}{1-h_i} \tag{16.22}$$

Πρακτικά, το διαγνωστικό $DFFITS_{(i)}$ ξεκινά από την τυποποίηση της i διαφοράς $\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(i)}$, έτσι αυτό μετρά την επίδραση στην πρόβλεψη όταν μία παρατήρηση διαγράφεται. Παρατηρήσεις με $DFFITS$ μεγαλύτερο του $2(p/n)^{1/2}$ διαγράφονται.

Αξίζει να σημειωθεί ότι όλα τα διαγνωστικά είναι συνάρτηση των καταλοίπων από την *LS* εκτίμηση και τα διαγώνια στοιχεία h_i του πίνακα Hat. Ατυχώς, τα διαγνωστικά αυτά δεν είναι αξιόπιστα όταν τα δεδομένα περιέχουν πολλαπλά outliers. Η αιτία είναι ότι τα r_i και h_i δέχονται μεγάλη επίδραση από τα outlier και είναι ευπαθή στην επικάλυψη, επειδή δύο ή περισσότερα outliers μπορούν να ενεργήσουν μαζί με πολύπλοκους τρόπους, όπου δυναμώνουν ή ακυρώνουν το ένα με το άλλο την επίδραση. Σαν τελικό αποτέλεσμα είναι ότι τα διαγνωστικά μονής διαγραφής συχνά δεν επιτυγχάνουν στην αναγνώριση των πολλαπλών outliers.

Παράδειγμα 16.3

Θεωρούμε και πάλι τα τεχνητά δεδομένα του Hawkins-Bradu-Kass του Πίνακα 16.2, όπου εφαρμόζουμε τα διαγνωστικά επίδρασης. Στον Πίνακα 16.4 φαίνεται ότι τα πολλαπλά outliers είναι δύσκολο να ανιχνευθούν με τα διαγνωστικά επίδρασης. Η κρίσιμη τιμή 1,0 για την απόσταση Cook $CD^2_{(i)}$ επιτυγχάνεται μόνο για την παρατήρηση 14, η οποία είναι καλής μοχλότητας. Επίσης το *DFFITS* δεν κατορθώνει να ξεχωρίσει τα κακά σημεία από τα σημεία καλής μόχλευσης.

Δείκτης	$CD^2_{(i)}$ (1.00)	$DFFITS_{(i)}$ (0.462)	Ανθεκτικά διαγνωστικά RD_i
1	0.040	0.407	12.999
2	0.053	0.470	13.500
3	0.046	0.430	13.991
4	0.031	0.352	13.961
5	0.039	0.399	13.982
6	0.052	0.459	13.451
7	0.079	0.575	13.910
8	0.052	0.464	13.383
9	0.034	0.372	13.718
10	0.048	0.439	13.535
11	0.348	-1.300	11.730
12	0.851	-2.168	12.004
13	0.254	-1.065	12.297
14	2.114	-3.030	13.674

Πίνακας 16.4 Διαγνωστικά των outliers για τις πρώτες 14 γραμμές των τεχνητών δεδομένων των Hawkins-Bradu-Kass

16.2.4 Ανθεκτικά Διαγνωστικά των Outliers

Τα διαγνωστικά επίδρασης μονής περίπτωσης της υποπαραγράφου 16.2.3 μπορούν να επεκταθούν προκειμένου να διαπιστώσουμε την αλλαγή που προκαλείται από την ταυτόχρονη διαγραφή περισσότερων από μία παρατηρήσεις. Αυτά τα διαγνωστικά πολλαπλής περίπτωσης είναι λιγότερο εύχρηστα λόγω του μεγάλου χρόνου υπολογισμού τους.

Η επιλογή των παρατηρήσεων που συμπεριλαμβάνονται στο υποσύνολο των δεδομένων δεν είναι προφανής. Ενδέχεται ένα υποσύνολο παρατηρήσεων από κοινού να έχουν μεγάλη επίδραση στην *LS* εκτίμηση, αλλά από μόνες τους όχι. Έτσι, η δοκιμή όλων των δυνατών υποσυνόλων συχνά είναι πολύ δύσκολη. Υπάρχουν αρκετές στρατηγικές για να προσεγγίσουν το υπολογιστικό πρόβλημα, αλλά καμία δεν εγγυάται την αναγνώριση όλων των κακών παρατηρήσεων που παρουσιάζονται στο σύνολο των δεδομένων.

Στην ανάλυση παλινδρόμησης είναι πολύ σημαντικό να ανιχνεύσουμε τα σημεία μοχλότητας, τα οποία είναι $\boldsymbol{x}$-outliers. Έχουμε αναφερθεί ότι τα διαγώνια στοιχεία h_i του πίνακα Hat είναι αδύνατο να αντιμετωπίσουν την δυσκολία των πολλαπλών σημείων μοχλότητας. Η πιο δημοφιλής διαδικασία για την ανίχνευση των $\boldsymbol{x}$-outliers είναι εκείνη η οποία επιλέγει εκείνο το υποσύνολο h παρατηρήσεων έτσι ώστε να ελαχιστοποιείται το άθροισμα των Mahalanobis αποστάσεων MD_i^2, όπως

$$\underset{h}{\text{miniminze}} \sum_{i=1}^{h} (MD_i^2) i:n$$

όπου $(MD_i^2)_{1:n} \le ... \le (MD_i^2)_{n:n}$ είναι οι τετραγωνικές αποστάσεις Mahalanobis σε αύξουσα σειρά. Στη λύση του παραπάνω προβλήματος το βέλτιστο υποσύνολο των h παρατηρήσεων δεν περιέχει τις μεγαλύτερες Mahalanobis αποστάσεις, οι οποίες αντιστοιχούν στα $\boldsymbol{x}$-outliers. Η διαδικασία αυτή ονομάζεται **Ελάχιστη Ορίζουσα Συνδιακύμανσης, "Minimum Covariance Determinant"** (*MCD*), και οδηγεί στην ανθεκτική εκτίμηση του πίνακα συνδιακύμανσης (robust covariance)

$$C^* = X_h^T X_h$$

όπου X_h είναι ο πίνακας των δεδομένων των ανεξάρτητων μεταβλητών, $x_{1i}, x_{2i}, ..., x_{pi}$, i = 1,2,…,h. Περισσότερα για τον ανθεκτικό πίνακα συνδιακύμανσης αναφέρονται στο επόμενο κεφάλαιο.

Με βάση τον ανθεκτικό πίνακα συνδιακύμανσης C^*, εκτιμώνται οι ανθεκτικές Mahanobis αποστάσεις, MD_i^2, από την εξίσωση 16.12.

Στη συνέχεια υπολογίζεται και η ανθεκτική εκτίμηση μοχλότητας h_i, από

την εξίσωση 16.13. Όπως φαίνεται και στον Πίνακα 16.4, τα ανθεκτικά διαγνωστικά MD_i^2 και h_i αναγνωρίζουν και τα 14 σημεία ως $\boldsymbol{x}$-outliers. Τα ανθεκτικά διαγνωστικά μοχλότητας είναι αξιόπιστα, αλλά η πληροφορία ότι το $\boldsymbol{x}_i$ είναι outlier δεν είναι αρκετή, επειδή πρέπει να εξετασθεί και η τιμή της εξαρτημένης y_i, για να επιβεβαιωθεί εάν το σημείο $(\boldsymbol{x}_i, y_i)$ είναι ένα outlier παλινδρόμησης (π.χ. εάν το $(\boldsymbol{x}_i, y_i)$ αποκλίνει από το γραμμικό μοντέλο της πληθώρας των δεδομένων).

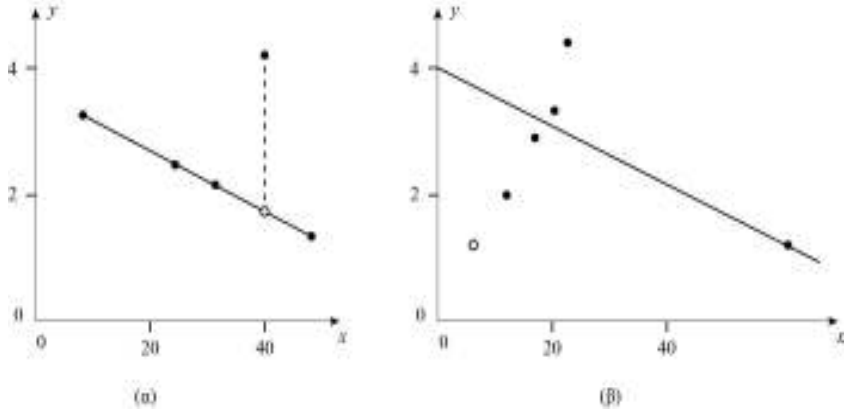
Έτσι, διαγνωστικά για να αναγνωρισθεί εάν το $\boldsymbol{x}$ είναι καλός ή κακός μοχλός είναι τα διαγνωστικά των studentized καταλοίπων t_i, καθώς και τα διαγνωστικά επίδρασης CD_i^2 και $DFFITS_{(i)}$. Τα διαγνωστικά αυτά μπορούν να μετασχηματισθούν σε ανθεκτικά αν οι τιμές των h_i στους αντίστοιχους τύπους αντικατασταθούν με τις ανθεκτικές τιμές των h_i. Εναλλακτικά, ανθεκτικά διαγνωστικά t_i, CD_i^2 και $DFFITS_{(i)}$ μπορούν να επιτευχθούν, εάν τα κατάλοιπα προέρχονται όχι από την *LS* εκτίμηση αλλά από μία πολύ ανθεκτική εκτίμηση παλινδρόμησης. Όμως, η διαδικασία αυτή είναι έμμεση, απαιτεί την ανθεκτική εκτίμηση των συντελεστών του μοντέλου παλινδρόμησης, η οποία περιγράφεται στην επόμενη παράγραφο.

16.3 ΑΝΘΕΚΤΙΚΟΙ Μ-ΕΚΤΙΜΗΤΕΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Ένα πρώτο βήμα προς την ανθεκτική παλινδρόμηση έκανε ο Edgeworth (1887), ο οποίος υποστήριξε ότι τα outliers έχουν μεγάλη επίδραση στην *LS* εκτίμηση επειδή τα κατάλοιπα r_i τετραγωνίζονται στην αντικειμενική συνάρτηση (16.2). Ως εκ τούτου, αυτός πρότεινε την εκτίμηση των **ελαχίστων απόλυτων τιμών** (least absolute values), ο οποίος ορίζεται από τη λύση του προβλήματος

$$\underset{\hat{\boldsymbol{\beta}}}{\text{ελαχιστοποίηση}} \sum_{i=1}^{n} \left| y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i \right| = \sum_{i=1}^{n} |r_i| \tag{16.23}$$

Η τεχνική αυτή είναι γνωστή ως L_1 παλινδρόμηση (επίσης, καλείται Least Absolute Diviation *LAD* εκτίμηση). Η *LAD* εκτίμηση σε μία παλινδρόμηση είναι λιγότερο ευαίσθητη στα y-outliers απ' ότι η *LS* εκτίμηση. Παρόλα αυτά η L_1 παλινδρόμηση δεν προστατεύεται από τα $\boldsymbol{x}$-outliers. Όπως φαίνεται στο Σχήμα 16.5 όταν υπάρχει ένα σημείο με μεγάλη μοχλότητα, η L_1 γραμμή περνά ακριβώς πάνω από αυτό.



Σχήμα 16.5 (α) Ανθεκτικότητα της L_1 παλινδρόμησης ως προς y-outlier. (b) Ευαισθησία της L_1 παλινδρόμησης ως προς ένα $\boldsymbol{x}$-outlier (leverage point)

Γενικά, οι ανθεκτικοί Μ-εκτιμητές συνδυάζουν ανθεκτικότητα και αποτελεσματικότητα.

Ορίζουμε ως **Μ-εκτιμήσεις παλινδρόμησης** κάθε λύση $\hat{\boldsymbol{\beta}}$ του προβλήματος

$$\underset{\hat{\boldsymbol{\beta}}}{\text{ελαχιστοποίηση}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right) \tag{16.24}$$

όπου $\hat{\sigma}$ είναι μία εκτίμηση της κλίμακας των σφαλμάτων u_i. Παραγωγίζοντας την (16.24) προκύπτει το σύστημα των κανονικών εξισώσεων,

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)\boldsymbol{x}_i = 0 \tag{16.25}$$

όπου $\psi = \rho'$. Οι συναρτήσεις ψ και ρ είναι ψ- και ρ- συναρτήσεις όπως ορίστηκαν στο κεφάλαιο 14 ορισμοί 14.1 και 14.2. Εάν η ψ είναι μονότονη, η λύση της (16.25) είναι λύση της (16.24) και είναι μοναδική. Όταν η ψ είναι επανακατερχόμενη, ο Μ-εκτιμητής είναι πιο ανθεκτικός, αλλά η λύση της (16.25) δεν είναι πάντα η ολική βέλτιστη της (16.24). Η επιτυχία των επανακατερχόμενων ψ εξαρτάται από τις αρχικές εκτιμήσεις του αλγόριθμου επίλυσης.

16.3.1 Μ-εκτιμητές Huber

Οι ανθεκτικοί Μ-εκτιμητές του Huber (1981) στηρίζονται στην ιδέα να αντικαταστήσουν τα τετραγωνικά κατάλοιπα r_i^2 από μια ρ-συνάρτηση (γνωστή ως "loss" function), η οποία έχει μικρότερο ρυθμό αύξησης (less rapidly increasing).

Ένας ανθεκτικός Μ-εκτιμητής Huber ορίζεται από τη λύση του προβλήματος ελαχιστοποίησης (16.24) όπου η "loss" συνάρτηση ρ_c είναι

$$\rho_c\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)=\begin{cases}\frac{1}{2}\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)^2 & \text{για} \quad \left|\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right|\le c\\ c\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)-c^2/2 & \text{για} \quad \left|\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right|>c\end{cases} \tag{16.26}$$

Όπου c είναι μια παράμετρος, η οποία ρυθμίζει την ανθεκτικότητα και αποτελεσματικότητα της Μ-εκτίμησης και συνήθως παίρνει τιμές στο διάστημα $1.65 < c < 2$. Η παράμερος c είναι γνωστή ως ρυθμιστική παράμετρος (tuning parameter).

Παραγωγίζοντας τη συνάρτηση (16.26) ως προς $\hat{\boldsymbol{\beta}}$ προκύπτει

$$\sum_{i=1}^{n}\psi_c\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)\boldsymbol{x}_i=0 \tag{16.27}$$

όπου $\psi_c=\rho_c'$ και από την 16.26 επιτυγχάνεται,

$$\psi_c\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)=\begin{cases}\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}} & \text{για} \quad \left|\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right|\le c\\ c & \text{για} \quad \left|\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right|>c\end{cases} \tag{16.28}$$

Η ψ_c είναι μία μονότονη συνάρτηση και η λύση της (16.27) είναι και λύση της (16.26). Επί πλέον, επειδή η ψ_c είναι και αύξουσα η λύση είναι μοναδική. Η εξίσωση (16.27) είναι ένα σύστημα p εξισώσεων, των οποίων η λύση δεν είναι πάντα εύκολο να βρεθεί. Στην πράξη, μπορούμε να χρησιμοποιήσουμε ένα επαναληπτικό αλγόριθμο (reweighted *LS*) παρόμοια με την παράγραφο 14.1.4, ο οποίος περιγράφεται παρακάτω, ή με την τεχνική του τετραγωνικού προγραμματισμού, ο οποίος περιγράφεται στην παράγραφο 16.3.3.

Στις περισσότερες περιπτώσεις το $\hat{\sigma}$ υπολογίζεται εκ των προτέρων, αλλά μπορεί επίσης να υπολογισθεί ταυτόχρονα από μια κλίμακας M-εκτίμηση εξίσωση. Στην περίπτωση όπου το σ είναι γνωστό, ο αναγνώστης μπορεί να επιβεβαιώσει ότι οι εκτιμήσεις είναι ισο-μεταβολής (regression affine equivariant).

Στην εξίσωση 16.25 η συνάρτηση ψ περιορίζει την επίδραση των μεγάλων καταλοίπων r_i αλλά η επίδραση των ακραίων τιμών $\boldsymbol{x}_i$ (x_i-outlier) εξακολουθεί να υπάρχει. Γι' αυτό έχει αποδειχθεί ότι ο M-εκτιμητής Huber έχει μικρό σημείο κατάρρευσης *BP*, $1/n$, και είναι πιο κατάλληλος μόνο για y-outlies.

M-εκτιμήσεις με εκτίμηση κλίμακας εκ των προτέρων

Για ανθεκτικές M-εκτιμήσεις παραμέτρων θέσης στην παράγραφο 14.1.2 εκτιμήθηκε ο σ χρησιμοποιώντας το *MAD*. Εδώ η ισοδύναμη διαδικασία είναι η εκτίμηση L_1 παλινδρόμησης και ακολούθως λαμβάνεται η διάμεσος των απόλυτων καταλοίπων:

$$\hat{\sigma} = \frac{1}{0,675} Med_i(|\, r_i \,|) \tag{16.29}$$

Όπως έχει περιγραφεί στην παράγραφο 14.1.2 η εκτίμηση $\hat{\sigma}$ είναι affine equivariant και ως εκ τούτου η λύση της (16.27) ή (16.26) είναι μία M-εκτίμηση, όπου το $\hat{\boldsymbol{\beta}}$ είναι affine and scale equivariant.

Για μεγάλο αριθμό παρατηρήσεων n, η κατανομή του $\hat{\boldsymbol{\beta}}$ είναι προσεγγιστικά κανονική

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},\ \upsilon(X^T X)^{-1}) \tag{16.30}$$

όπου υ είναι το ίδιο όπως και στην εξίσωση (14.14):

$$\upsilon = \hat{\sigma} \frac{E\psi(u/\sigma)^2}{E(\psi'(u/\sigma))^2}, \tag{16.31}$$

όπως έχει αποδειχθεί από τους Yohai και Marrona (1979).

Ταυτόχρονη εκτίμηση παλινδρόμησης και κλίμακας σ

Μία άλλη προσέγγιση για την εκτίμηση της σ είναι η διαδικασία όπως και στην παράγραφο 14.1.3. Δηλαδή, εισέρχεται μία επιπλέον εξίσωση, στο σύστημα εξισώσεων 16.27) για το $\boldsymbol{\beta}$, για M-εκτίμηση της σ, όπως

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)\boldsymbol{x}_i = 0 \tag{16.32}$$

$$\frac{1}{n}\sum_{i=1}^{n} \rho_{scale}\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right) = \delta \tag{16.33}$$

όπου ρ_{scale} είναι ρ-συνάρτηση. Το παραπάνω σύστημα εξισώσεων προκύπτει αν παραγωγίσουμε την (16.26) ως προς $\boldsymbol{\beta}$ και σ.

16.3.2 Υπολογισμός της Ανθεκτικής M-εκτίμησης

Χρησιμοποιώντας μία επαναληπτική επαναζυγιζόμενη μέθοδο παρόμοια με την παράγραφο 14.1.4, ορίζουμε μία συνάρτηση βάρους W,

$$W\left(\frac{r_i(\hat{\beta})}{\hat{\sigma}}\right) = \begin{cases} \psi_c\left(\frac{r_i}{\hat{\sigma}}\right) \Big/ \left(\frac{r_i}{\hat{\sigma}}\right) & \text{όταν} \quad \left|\frac{r_i}{\hat{\sigma}}\right| > c \\ 1 & \text{όταν} \quad \left|\frac{r_i}{\hat{\sigma}}\right| \le c \end{cases} \tag{16.34}$$

Έτσι, η M-εκτίμηση του $\boldsymbol{\beta}$ από την εξίσωση (16.27) μπορεί να γραφεί:

$$\sum_{i=1}^{n} w_i r_i \boldsymbol{x}_i = \sum_{i=0}^{n} w_i \boldsymbol{x}_i (y_i - \boldsymbol{x}^T \hat{\boldsymbol{\beta}}) = 0 \tag{16.35}$$

με $w_i = W(r_i / \hat{\sigma})$. Αυτές είναι οι «ζυγιζόμενες κανονικές εξισώσεις» και εάν τα w_i ήταν γνωστά, οι εξισώσεις θα μπορούσαν να λυθούν εφαρμόζοντας τα ελάχιστα τετράγωνα *LS* στα δεδομένα $\sqrt{w_i}\, y_i$ και $\sqrt{w_i}\, \boldsymbol{x}_i$. Αλλά τα w_i δεν είναι γνωστά και εξαρτώνται από τα δεδομένα. Έτσι, η διαδικασία η οποία εξαρτάται από μία παράμετρο ανοχής ε, είναι

1. Υπολόγισε μία L_1 εκτίμηση $\hat{\boldsymbol{\beta}}_0$ και $\hat{\sigma}$ από την (16.29).
2. Για $k = 0, 1, 2, \ldots$:
(α) Δεδομένου $\hat{\boldsymbol{\beta}}_k$, για $i = 1,\ldots,n$ υπολόγισε $r_{i,k} = y_i - \boldsymbol{x}^T \hat{\boldsymbol{\beta}}_k$ και $w_{i,k} = W(r_{i,k} / \hat{\sigma})$.
(β) Υπολόγισε $\hat{\boldsymbol{\beta}}_{k+1}$ λύνοντας

$$\sum_{i=1}^{n} w_{i,k} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}) = 0$$

3. Σταμάτα όταν $\max_i \left(\left| r_{i,k} - r_{i,i+1} \right| \right) / \hat{\sigma} < \varepsilon$

Ο αλγόριθμος αυτός συγκλίνει εάν $W(x)$ είναι μη-αύξουσα για $x > 0$. Δεδομένου ότι η ψ είναι μονότονη η λύση είναι βασικά μοναδική, γι' αυτό η αρχική λύση επιδρά στον αριθμό επαναλήψεων και όχι στην τελική λύση. Η διαδικασία καλείται «επαναληπτικά επαναζυγιζόμενα ελάχιστα τετράγωνα» (iteratively reweighted least squares, IRWLS).
Για ταυτόχρονη εκτίμηση του $\boldsymbol{\beta}$ και σ η διαδικασία είναι η ίδια, με τη διαφορά ότι σε κάθε επανάληψη το $\hat{\sigma}$ επίσης αναπροσαρμόζεται,

$$\hat{\sigma}_{k+1} = \sqrt{\frac{1}{n\delta} \sum_{i=1}^{n} W_{k,i} (r_i(\hat{\boldsymbol{\beta}}))^2} \tag{16.36}$$

16.3.3 Πρακτική Ερμηνεία της Μ-εκτίμησης Huber

Πολλαπλασιάζοντας και τις δύο πλευρές της εξίσωσης (16.26) με $2\hat{\sigma}^2$ και μετά από εύκολες πράξεις καταλήγουμε σε μία ισοδύναμη "loss" συνάρτηση,

$$2\hat{\sigma}^2 \rho_c \left(\frac{r_i(\hat{\beta})}{\hat{\sigma}} \right) = \begin{cases} r_i^2 & \text{όταν} \quad |r_i| \le c\hat{\sigma} \\ (c\hat{\sigma})^2 + 2(c\hat{\sigma})\varepsilon_i & \text{όταν} \quad |r_i| > c\hat{\sigma} \end{cases} \tag{16.37}$$

όπου ε_i παριστάνει το μέγεθος ελάττωσης των μεγάλων καταλοίπων όπως

$$\varepsilon_i = \begin{cases} |r_i| - c\hat{\sigma} & \text{όταν} \quad |r_i| > c\hat{\sigma} \\ 0 & \text{όταν} \quad |r_i| \le c\hat{\sigma}_1 \end{cases} \tag{16.38}$$

ή ισοδύναμα, το ε_i μπορεί να ερμηνευθεί ως απόσταση προσέλκυσης της τιμής y_i προς τη γραμμή (ή πολυεπίπεδο) παλινδρόμησης.

Συνδυάζοντας την (16.37) και (16.38), μία Μ-εκτίμηση Huber μπορεί να προκύψει από τη λύση του προβλήματος

$$\underset{\hat{\beta}}{\text{ελαχιστοποίηση}} \sum_{i=1}^{n} r_i^{*2} + 2c\sigma\varepsilon_i \tag{16.39}$$

όπου r_i^* είναι τα μετασχηματισμένα κατάλοιπα, τα οποία είναι γνωστά ως Winsorized κατάλοιπα,

$$r_i^* = \begin{cases} r_i & \text{όταν} \quad |r_i| \le c\sigma \\ c\sigma & \text{όταν} \quad |r_i| > c\sigma \end{cases} \tag{16.40}$$

Παρατήρηση

Στην αντικειμενική συνάρτηση (16.39) ο όρος $2c\sigma\varepsilon_i$ μπορεί να ερμηνευθεί ως το κόστος ποινικοποίησης για την προσέλκυση του y_i προς τη γραμμή παλινδρόμησης για μια απόσταση ε_i. ■

Έτσι, το ισοδύναμο πρόβλημα της Μ-εκτίμησης Huber (16.39), μπορεί κανείς να το λύσει με τετραγωνικό προγραμματισμό όπως ακολουθεί:

$$\underset{\hat{\beta}}{\text{ελαχιστοποίηση}} \sum_{i=1}^{n} r_i^{*2} + 2c\sigma\varepsilon_i$$

με περιορισμούς:

$$\begin{aligned} & \boldsymbol{x}^T\hat{\boldsymbol{\beta}} + r_i^* + \varepsilon_i \ge y_i \\ & \boldsymbol{x}^T\hat{\boldsymbol{\beta}} - r_i^* - \varepsilon_i \le y_i \\ & \text{για} \quad i = 1,\ldots,n \end{aligned} \tag{16.41}$$

Το πρόβλημα αυτό είναι ένα κλασσικό πρόβλημα κυρτού τετραγωνικού προγραμματισμού, το οποίο έχει μία και μοναδική βέλτιστη λύση $(\hat{\beta},\ r_i^*,\ \varepsilon_i)$.

16.4 ΑΝΘΕΚΤΙΚΟΙ GM-ΕΚΤΙΜΗΤΕΣ

Όπως αναφέρθηκε και στην παράγραφο 16.3, οι Μ-εκτιμητές δεν περιορίζουν την επίδραση των $\boldsymbol{x}_i$ με υψηλή μόχλευση (high leverage points), αντιμετωπίζουν μόνο τα y-outliers. Όταν υπάρχουν x-outliers στον πίνακα δεδομένων

X, ενδέχεται η επίδρασή τους να είναι καταστροφική για τους Μ-εκτιμητές. Η επίδραση μιας παρατήρησης $(y_i, \boldsymbol{x}_i)$ μπορεί να είναι ανεξέλεγκτα μεγάλη επειδή το $\boldsymbol{x}_i$ πολλαπλασιάζει την $\psi_c(\cdot)$ στις κανονικές εξισώσεις (16.25) ή (16.27). Ένας απλός τρόπος για να ανθεκτικοποιήσουμε την μονότονη Μ-εκτίμηση είναι να ελαφρύνουμε την επίδραση του $\boldsymbol{x}_i$ στο σύστημα των κανονικών εξισώσεων. Οπότε, μπορούμε να ορίσουμε μια ανθεκτική εκτίμηση από τη λύση της εξίσωσης

$$\sum \psi_c \left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}} \right) \boldsymbol{x}_i W(d(\boldsymbol{x}_i)) = 0 \qquad (16.42)$$

όπου W είναι μία συνάρτηση βάρους και $d(\boldsymbol{x}_i)$ είναι μια μέτρηση της απόστασης του διανύσματος $\boldsymbol{x}_i$ από το κέντρο βάρους του πίνακα των ανεξάρτητων μεταβλητών X. Εδώ η ψ είναι μονότονη και το $\hat{\sigma}$ εκτιμάται ταυτόχρονα όπως και στην (16.33)).

Πιο γενικά, το βάρος για κάθε παρατήρηση $(\boldsymbol{x}_i, y_i)$ μπορεί να εξαρτάται από το μέγεθος του κατάλοιπου r_i αλλά και από τη μοχλότητα του $\boldsymbol{x}_i$ και μια γενικευμένη Μ-εκτίμηση (generalized, GM-estimate) $\hat{\boldsymbol{\beta}}$ ορίζεται από τη λύση

$$\sum_{i=1}^{n} \eta \left(d(\boldsymbol{x}_i), \ \frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}} \right) \boldsymbol{x}_i = 0 \qquad (16.43)$$

όπου για κάθε $d(x_i)$ η συνάρτηση $\eta\left(d(\boldsymbol{x}_i), \ r_i\right)$ είναι μια μη-φθίνουσα και φραγμένη ψ-συνάρτηση των καταλοίπων r_i.

Δύο ειδικές φόρμες της GM-εκτίμησης παρουσιάζουν ιδιαίτερο ενδιαφέρον. Η πρώτη είναι η εκτίμηση της (16.42), η οποία αντιστοιχεί στην επιλογή

$$\eta\left(d(\boldsymbol{x}_i), r_i\right) = W(d\boldsymbol{x}_i)\psi(r_i) \qquad (16.44)$$

και καλείται μία «εκτίμηση Mallows» (Mallows, 1975). Η GM-εκτίμηση του Mallows περιορίζει την επίδραση των x-outliers, αλλά χάνει αποτελεσματικότητα, διότι η συνάρτηση W ελαφρύνει κάθε x-outlier χωρίς να λαμβάνει υπόψη το μέγεθος του κατάλοιπου r. Γι' αυτό όταν το x-outlier είναι ένας «καλός» μοχλός δεν συμβάλλει στην ακρίβεια της εκτίμησης του $\boldsymbol{\beta}$.

Μία δεύτερη φόρμα της GM-εκτίμησης προτεινόμενη από τον Schweppe (1975) έχει τη δυνατότητα να διατηρεί την αποτελεσματικότητά της και εί-

ναι της μορφής

$$\eta\left(d(\boldsymbol{x}_i), r_i\right) = d(\boldsymbol{x}_i)\psi\left(\frac{r_i}{d(\boldsymbol{x}_i)}\right) \tag{16.45}$$

Ο Hill (1977) πρότεινε, για την GM-εκτίμηση του Schweppe, το σύστημα των εξισώσεων

$$\sum_{i=1}^{n} d(\boldsymbol{x}_i)\psi_c\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma} d(\boldsymbol{x}_i)}\right)\boldsymbol{x}_i = 0 \tag{16.46}$$

όπου $d(\boldsymbol{x}_i) = \sqrt{1-h_i}$, και h_i μετρά τη μοχλότητα της i παρατήρησης (διαγώνια στοιχεία του πίνακα H (Hat)). Εάν η παρατήρηση είναι ένα $\boldsymbol{x}_i$-outlier, το h_i έχει αυξημένη τιμή και η μικρή τιμή του $d(\boldsymbol{x}_i)$ συμβάλλει σε δραστικότερη ελάφρυνση. Αλλά, όταν $r_i / \hat{\sigma} d(\boldsymbol{x}_i)$ είναι μικρό τα $d(x_i)$ απαλείφονται.

Στους περισσότερους GM-εκτιμητές προτιμάται η φόρμα (16.46) αλλά με διαφορετική εκτίμηση του μεγέθους του $d(\boldsymbol{x}_i)$ όπως περιγράφεται παρακάτω.

GM-Εκτιμητές Φραγμένης Επίδρασης

Οι GM-εκτιμητές αναπτύχθηκαν για να περιορίσουν την επίδραση της κάθε ακραίας παρατήρησης, της οποίας η επίπτωση μετριέται με τη γνωστή **Συνάρτηση Επίδρασης** (Influence Function) IF ($\boldsymbol{x}_i$, y_i).

Η συνάρτηση επίδρασης IF παριστάνει την επίπτωση της παρατήρησης στην εκτίμηση $\hat{\boldsymbol{\beta}}$, και ορίσθηκε από τον Hampel (1974) ως

$$\sqrt{n}\left[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \frac{1}{n}\sum IF(\boldsymbol{x}_i, y_i)\right] \to 0 \tag{16.47}$$

πιθανολογικά καθώς $n \to \infty$.

Οι Krasker και Welsch (1982) προτείνουν μία εμπειρική μέτρηση της συνάρτησης επίδρασης IF στην παλινδρόμηση των ελαχίστων τετραγώνων χρησιμοποιώντας την ευκλείδεια norm του διανύσματος

$$IF(\boldsymbol{x}_i, y_i) = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \left[r_i(\hat{\boldsymbol{\beta}}) / (1-h_i)\right](X^T X)^{-1}\boldsymbol{x}_i^T \tag{16.48}$$

όπου $\hat{\boldsymbol{\beta}}_{(i)}$ είναι η εκτίμηση του $\hat{\boldsymbol{\beta}}$ χωρίς την παρατήρηση i. Μία πιο εύχρηστη μέτρηση της IF προκύπτει από την τυποποίηση της διαφοράς προσαρμογής

$x_i(\hat{\beta}-\hat{\beta}_{(i)})$, η οποία προσεγγίζει το διαγνωστικό *DFFITS* της εξίσωσης (16.22),

$$\frac{x_i(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)})}{\hat{\sigma}\sqrt{h_i}}=\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\ \frac{\sqrt{h_i}}{1-h_i} \tag{16.49}$$

Έτσι, ένας GM-εκτιμητής φραγμένης επίδρασης προκύπτει από τη λύση της (16.46) αν το $d(\boldsymbol{x}_i)$ είναι

$$d(\boldsymbol{x}_i)=\frac{1-h_i}{\sqrt{h_i}} \tag{16.50}$$

Συνδυάζοντας την (16.50), (16.45) και (16.28) ένας GM-εκτιμητής φραγμένης επίδρασης προκύπτει από τη λύση του συστήματος (16.43) αν τελικά η συνάρτηση η είναι

$$\eta\left(d(x_i),\ \frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)=\psi_{c_i}\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right) \tag{16.51}$$

όπου η ψ_{c_i} είναι μία ψ-συνάρτηση Huber όπως στην (16.28) με $c_i=c\sqrt{h_i}/(1-h_i)$,

$$\psi_{c_i}\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)=\begin{cases}\dfrac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}} & \text{όταν} \quad \dfrac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\le c_i\\[2ex] c_i & \text{όταν} \quad \dfrac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}> c_i\end{cases} \tag{16.52}$$

Παρατήρηση

Είναι αξιοσημείωτο ότι διάφοροι GM-εκτιμητές διαφέρουν από τους M-εκτιμητές Huber μόνο ως προς την επιλογή της παραμέτρου c. Διάφορες τιμές της παραμέτρου c, $c = c_i$, οδηγούν σε διαφορετικούς GM-εκτιμητές. Για παράδειγμα, αν

$$c_i=c\cdot\sqrt{1-h_i} \qquad (Schweppe)$$
$$c_i=c\cdot\sqrt{h_i}/(1-h_i) \qquad (Krasker\ \ Welsch)$$

■

Υπολογιστικά, λαμβάνοντας υπόψη και την παραπάνω παρατήρηση, οι GM-εκτιμητές δεν παρουσιάζουν καμία ιδιαίτερη δυσκολία. Ακολουθούμε την

υπολογιστική διαδικασία των Μ-εκτιμητών.

Ιδιότητες GM-εκτιμητών

Από τα παραπάνω συνεπάγεται ότι οι GM-εκτιμητές έχουν μερικές καλές ιδιότητες:

1. Η συνάρτηση επίδρασης *IF* είναι φραγμένη.
2. Το σημείο κατάρρευσης *BP* είναι > 0.
3. Είναι εύκολο να τους υπολογίσουμε, όπως και τους Μ-εκτιμητές.

Παρ' όλα αυτά οι GM-εκτιμητές έχουν μερικά μειονεκτήματα, όπως:

1. Η αποτελεσματικότητά τους εξαρτάται από τη μόλυνση της κατανομής των $\boldsymbol{x}$. Εάν υπάρχουν πολλά x-outliers δεν μπορούν να είναι ταυτόχρονα πολύ αποτελεσματικοί και πολύ ανθεκτικοί.
2. Το σημείο κατάρρευσης *BP* ελαττώνεται καθώς αυξάνει ο αριθμός p των ανεξάρτητων μεταβλητών της παλινδρόμησης.
3. Ακόμη, το σημείο κατάρρευσης *BP* ελαττώνεται εάν η κλίμακα σ υπολογίζεται ταυτόχρονα, ειδικά για μεγάλο p.

Τέλος, οι GM-εκτιμητές είναι μια καλή επιλογή, όταν ο αριθμός p των ανεξάρτητων μεταβλητών p είναι σχετικά μικρός. Οι GM-εκτιμητές χρησιμοποιούνται σε πολλές εφαρμογές λόγω και της εύκολης υπολογιστικής διαδικασίας. Για περισσότερη ανθεκτικότητα και αποτελεσματικότητα μπορούν να χρησιμοποιηθούν οι ανθεκτικές εκτιμήσεις των h_i, όπως προτείνονται στο επόμενο κεφάλαιο.

16.5 ΑΝΘΕΚΤΙΚΟΙ ΕΚΤΙΜΗΤΕΣ ΥΨΗΛΟΥ ΣΗΜΕΙΟΥ ΚΑΤΑΡΡΕΥΣΗΣ (HIGH-BREAKDOWN POINT, HBP)

Στην παράγραφο αυτή περιγράφουμε τους πιο σύγχρονους ανθεκτικούς εκτιμητές παλινδρόμησης, οι οποίοι είναι υψηλού σημείου κατάρρευσης (*HBP*).

16.5.1 Μ-εκτιμητές με φραγμένη ρ-συνάρτηση

Μια ανθεκτική εκτίμηση παλινδρόμησης, όπου και οι δύο μεταβλητές $\boldsymbol{x}$ και y ενδέχεται να περιέχουν outliers, ορίζεται από την Μ-εκτίμηση $\hat{\boldsymbol{\beta}}$ οριζόμενη από τη λύση του προβλήματος,

$$\min \sum_{i=1}^{n} \rho\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right) \tag{16.53}$$

με μία **φραγμένη** ρ και μία εκ των προτέρων *HBP* εκτίμηση της κλίμακας $\hat{\sigma}$. Η κλίμακα $\hat{\sigma}$ πρέπει να πληροί κάποιες απαιτήσεις ανάλογα με τον εκτιμητή όπως περιγράφεται αναλυτικότερα στους επόμενους ανθεκτικούς εκτιμητές. Εάν η ρ έχει παράγωγο την ψ, έπεται ότι

$$\sum \psi\left(\frac{r_i}{\hat{\sigma}}\right)\boldsymbol{x}_i = 0 \tag{16.54}$$

όπου η ψ είναι επανακατερχόμενη (είναι εύκολο να επιβεβαιωθεί ότι μια συνάρτηση ρ με μονότονη παράγωγο ψ δεν μπορεί να είναι φραγμένη). Κατά συνέπεια η εξίσωση εκτίμησης (16.54) μπορεί να έχει πολλές λύσεις, οι οποίες αντιστοιχούν σε **τοπικά** βέλτιστα του προβλήματος ελαχιστοποίησης (16.53), και γενικά μόνο μία από αυτές (η «καλή» λύση) αντιστοιχεί στην **ολική** ελαχιστοποίηση. Οι ρ και $\hat{\sigma}$ μπορούν να επιλεγούν έτσι ώστε να επιτυγχάνεται *HBP* και αποτελεσματικότητα.
Μία φραγμένη συνάρτηση, η οποία συντελεί στην κανονική αποτελεσματικότητα του εκτιμητή είναι η δημοφιλής ρ-συνάρτηση του Tukey (biweight)

$$\rho_c\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right) = \begin{cases} \dfrac{r_i^2}{2} - \dfrac{r_i^4}{2c^2} + \dfrac{r_i^6}{6c^2} & \text{για} \quad \left|\dfrac{r_i}{\hat{\sigma}}\right| \le c \\ \dfrac{c^2}{6} & \text{για} \quad \left|\dfrac{r_i}{\hat{\sigma}}\right| > c \end{cases} \tag{16.55}$$

Η παράγωγος της ρ οδηγεί στην επανακατερχόμενη ψ-συνάρτηση

$$\psi_c\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right) = \begin{cases} \dfrac{r_i}{\hat{\sigma}}\left(1-(r_i/\hat{\sigma}c)^2\right)^2 & \text{για} \quad \left|\dfrac{r_i}{\hat{\sigma}}\right| \le c \\ 0 & \text{για} \quad \left|\dfrac{r_i}{\hat{\sigma}}\right| > c \end{cases} \tag{16.56}$$

Ιδιότητες της Μ-εκτίμησης με φραγμένη ρ-συνάρτηση

Εάν η εκ των προτέρων εκτίμηση της κλίμακας $\hat{\sigma}$ είναι affine equivariant, τότε και ο εκτιμητής της (16.54) κληρονομεί αυτήν την ιδιότητα. Έχει αποδειχθεί ότι η φραγμένη ρ-συνάρτηση οδηγεί σε ανθεκτικές εκτιμήσεις μέχρι 50% *HBP*. Παρ' όλα αυτά η συνάρτηση επίδρασης *IF* της εξίσωσης (16.48) δεν είναι φραγμένη μόνον όταν το $\boldsymbol{x}$ είναι πολύ μεγάλο. Το γεγονός όμως ότι

η *IF* δεν είναι φραγμένη δεν σημαίνει απαραίτητα ότι η μεροληψία του εκτιμητή είναι πολύ μεγάλη για μεγάλη μόλυνση των δεδομένων.
Κάτω από τις γενικές συνθήκες ο εκτιμητής ο οποίος ορίσθηκε στην (16.53) είναι συνεπής και ασυμπτωτικά κανονικός. Πιο ειδικά,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N_p(0,\ \upsilon V_X^{-1}) \tag{16.57}$$

όπου $V_X = E(\boldsymbol{x}\boldsymbol{x}^T)$ και υ είναι όπως στην (16.31)

$$\upsilon = \sigma^2 \frac{E\psi(u/\sigma)^2}{\left(E\psi'(u/\boldsymbol{\sigma})\right)^2} \tag{16.58}$$

Το αποτέλεσμα αυτό συνεπάγει ότι καθώς το $\boldsymbol{x}$ έχει πεπερασμένη διακύμανση, η αποτελεσματικότητα του $\hat{\boldsymbol{\beta}}$ δεν εξαρτάται από την κατανομή του $\boldsymbol{x}$. Αυτό σημαίνει ότι εάν το x-outlier είναι «καλό» σημείο, συμβάλλει στην ακρίβεια εκτίμησης του $\boldsymbol{\beta}$, εάν είναι «κακό» σημείο περιορίζεται η μεροληψία, η οποία προκαλείται από αυτό.

16.5.2 MM-εκτιμητές

Ο Yohai (1985) ανέπτυξε έναν βελτιωμένο ανθεκτικό εκτιμητή με υψηλό σημείο κατάρρευσης και καλή αποτελεσματικότητα, γνωστός ως MM-εκτιμητής. Ο εκτιμητής αυτός ορίζεται σε τρία στάδια:

1. Υπολογίζεται μία αρχική συνεπής εκτίμηση $\hat{\boldsymbol{\beta}}_0$ με *HBP* αλλά ενδεχομένως χαμηλή αποτελεσματικότητα,
2. Υπολογίζεται η ανθεκτική κλίμακα $\hat{\sigma}$ από τα κατάλοιπα $r_i(\hat{\boldsymbol{\beta}}_0)$, χρησιμοποιώντας μία φραγμένη ρ_0-συνάρτηση. Η εκτίμηση της κλίμακας $\hat{\sigma}$ πρέπει να είναι μία M-εκτίμηση

$$\frac{1}{n}\sum \rho_{c_0}\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right) = 0.5$$

   Πρέπει να προσδιορίσουμε τέτοιο c_0, έτσι ώστε η ασυμπτωτική τιμή της σ να συμπίπτει με τυπική απόκλιση, όταν τα σφάλματα u_i ακολουθούν κανονική κατανομή. Για την διτετράγωνη (16.55) η τιμή είναι $c_0 = 1.56$.
3. Τέλος, ο MM-εκτιμητής $\hat{\boldsymbol{\beta}}$ ορίζεται από τη λύση

$$\sum_{i=1}^{n} \psi_c(r_i(\hat{\boldsymbol{\beta}})/\hat{\sigma})\boldsymbol{x}_i = 0 \tag{16.59}$$

η οποία ικανοποιεί

$$S(\hat{\boldsymbol{\beta}}) \leq S(\hat{\boldsymbol{\beta}}_0) \tag{16.60}$$

όπου

$$S(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \rho_c(r_i(\hat{\boldsymbol{\beta}}) / \hat{\sigma}) \tag{16.61}$$

Η συνάρτηση ρ είναι μία ρ-συνάρτηση φραγμένη, το οποίο συνεπάγει ότι η παράγωγός της $\psi = \rho'$ είναι επανακατερχόμενη. Μία καλή επιλογή για την ρ είναι η διτετράγωνη συνάρτηση του Tukey (16.55). Το σημαντικότερο είναι ότι αυτή η ρ πρέπει να είναι διαφορετική από εκείνη που εκτιμά την κλίμακα $\hat{\sigma}$ στο δεύτερο στάδιο, διότι το πρώτο και δεύτερο στάδιο πρέπει να επιτύχουν ανθεκτικότητα (*HBP*) ενώ το τρίτο στάδιο βελτιώνει την αποτελεσματικότητα. Ο Yohai απέδειξε ότι ο MM-εκτιμητής κληρονομεί το 50% σημείο κατάρρευσης, έχει υψηλή αποτελεσματικότητα και είναι ένας από τους πιο ανταγωνιστικούς ανθεκτικούς εκτιμητές παλινδρόμησης.

16.5.3 Εκτιμητές με βάση την Κλίμακα Ανθεκτικών Καταλοίπων

Σε αυτήν την παράγραφο παρουσιάζουμε μία οικογένεια εκτιμητών παλινδρόμησης, οι οποίοι δεν απαιτούν εκ των προτέρων εκτίμηση κλίμακας και γι' αυτό είναι χρήσιμοι ως αρχικές εκτιμήσεις στον MM-εκτιμητή. Ένας εκτιμητής αυτής της οικογένειας προμηθεύει μία καλή αρχική εκτίμηση $\hat{\boldsymbol{\beta}}_0$ και τα αντίστοιχα κατάλοιπα για να προσδιορισθεί η εκ των προτέρων εκτίμηση κλίμακας $\hat{\sigma}$ στη μέθοδο της MM-εκτίμησης.

Εκτίμηση Ελάχιστης Διαμέσου των Τετραγώνων (LMS)

Γνωρίζουμε ότι τα ελάχιστα τετράγωνα (*LS*) και οι απόλυτες τιμές (L_1) ελαχιστοποιούν το μέσο όρο των τετραγώνων και των απολύτων καταλοίπων αντίστοιχα, και ως εκ τούτου ελαχιστοποιούν μία μέτρηση του μεγέθους των καταλοίπων και ενδέχεται να δεχθούν κακή επίδραση ακόμη και από ένα κατάλοιπο ενός μόνον outlier. Μία εναλλακτική ανθεκτική εκτίμηση ελαχιστοποιεί την κλίμακα μέτρησης των καταλοίπων, η οποία δεν ευαισθητοποιείται στις μεγάλες τιμές, και έτσι μία τέτοια δυνατότητα είναι η **διάμεσος** (median) των απόλυτων καταλοίπων. Αυτή είναι η βάση για την εκτίμηση της ελάχιστης διαμέσου των τετραγώνων (Least Median Squares, *LMS*), η οποία εισήχθηκε από τον Hampel (1975) και Rousseeuw (1984).

Ο εκτιμητής *LMS* δίνεται από τη λύση του προβλήματος

$$\hat{\boldsymbol{\beta}}_{LMS} : \arg\min_{\hat{\boldsymbol{\beta}}} med\, r_i(\hat{\boldsymbol{\beta}}) \tag{16.62}$$

Έχει αποδειχθεί ότι ο *LMS*-εκτιμητής είναι ανθεκτικός ως προς y καθώς και $\boldsymbol{x}$-outliers. Το σημείο θραύσης του είναι 50%, και είναι ισο-μεταβολής (equivariant), αφού χρησιμοποιεί μόνο κατάλοιπα. Το μειονέκτημα του *LMS*-εκτιμητή είναι ότι χάνει αποτελεσματικότητα όταν τα δεδομένα ακολουθούν κανονική κατανομή και δεν περιέχουν outliers.

S-εκτιμητής

Οι Rousseeuw και Yohai (1984) όρισαν έναν *HBP*-εκτιμητή (equivariant) από την ελαχιστοποίηση της διασποράς των καταλοίπων:

$$\hat{\boldsymbol{\beta}}_S : \arg\min_{\hat{\boldsymbol{\beta}}} \hat{\sigma}\left(r_1(\hat{\boldsymbol{\beta}}),....,r_n(\hat{\boldsymbol{\beta}})\right) \tag{16.63}$$

Η διασπορά $\hat{\sigma}\left(r_1(\hat{\boldsymbol{\beta}}),...,r_n(\hat{\boldsymbol{\beta}})\right)$ ορίζεται από την λύση της εξίσωσης

$$\frac{1}{n}\sum_{i=1}^{n}\rho_c\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}}\right)=\delta \tag{16.64}$$

όπου δ συνήθως έχει την τιμή $E_{\Phi}[\rho_c]$, όπου Φ είναι η τυπική κανονική κατανομή. Η συνάρτηση ρ είναι μία φραγμένη ρ-συνάρτηση. Εάν η (16.64) έχει περισσότερες από μία λύση, τότε επιλέγουμε

$$\hat{\sigma}\left(r_1(\hat{\boldsymbol{\beta}}),...,r_n(\hat{\boldsymbol{\beta}})\right)=\sup\left\{\hat{\sigma}:1/n\sum_i\rho_c(r_i/\hat{\sigma})=\delta\right\}$$

Ο εκτιμητής $\hat{\boldsymbol{\beta}}_S$ ονομάζεται *S*-εκτιμητής διότι προέρχεται από μια στατιστική κλίμακα με έναν έμμεσο τρόπο. Στην πράξη, η εκτίμηση $\hat{\sigma}$, όπως ορίζεται από την (16.64) είναι μία *M*-εκτίμηση κλίμακας. Προφανώς ο *S*-εκτιμητής είναι ισο-μεταβολής (regression, scale and affine equivariant).
Συνήθως ως ρ-συνάρτηση επιλέγεται η φραγμένη διτετράγωνη το Tukey (16.55), η οποία οδηγεί σε μία εκτίμηση $\hat{\boldsymbol{\beta}}_0$ κατάλληλη και για την αρχική τιμή της *MM*-εκτίμησης. Με αυτήν την επιλογή της ρ επιτυγχάνεται εκτίμηση με σημείο θραύσης 50% και αποτελεσματικότητα όχι μεγαλύτερη του 0.33.

Παρατήρηση

Αφού μια *S*-εκτίμηση προέρχεται από μία ρ-συνάρτηση φραγμένη, η ψ-συνάρτηση είναι επανακατερχόμενη και αποκόπτει τα μεγάλα κατάλοιπα. Γι' αυτό, μπορούμε να θεωρήσουμε ότι $\hat{\boldsymbol{\beta}}_S$ είναι μία *M*-εκτίμηση, αφού διαγράψουμε τα outliers των δεδομένων.

Ο υπολογισμός του $\hat{\boldsymbol{\beta}}_S$ περιγράφεται στη μεθεπόμενη υποπαράγραφο. ■

LTS-εκτιμητής

Ο Rousseeuw (1984) εισήγαγε τον εκτιμητή των ελαχίστων αποκοπτόμενων τετραγώνων (least trimmed squares, *LTS*), ο οποίος ορίζεται ως εξής:

$$\hat{\beta}_{LTS} : \arg\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{h} \left(r(\hat{\boldsymbol{\beta}}) \right)^2_{i:n} \tag{16.65}$$

όπου $(r^2)_{1:n} \leq \ldots\ldots \leq (r^2)_{n:n}$ είναι τα τετράγωνα των καταλοίπων σε αύξουσα σειρά (παρατηρούμε ότι τα κατάλοιπα πρώτα υψώνονται στο τετράγωνο και μετά ταξινομούνται). Η φόρμα (16.65) είναι παρόμοια του *LS*, με τη μόνη διαφορά ότι τα μεγάλα κατάλοιπα δεν συμμετέχουν στο άθροισμα, επιτρέποντας έτσι την προσαρμογή της παλινδρόμησης μακριά από τα outliers. Ο εκτιμητής *LTS* είναι equivariant. Τη μεγαλύτερη ανθεκτικότητα την επιτυγχάνει όταν το *h* προσεγγίζει το *n*/2, με σημείο κατάρρευσης 50%.

Ο *LTS* εκτιμητής μπορεί να αντιμετωπίσει και τους δύο τύπους των outliers ανεξάρτητα και από τον αριθμό *p* των ανεξάρτητων μεταβλητών $\boldsymbol{x}$. Έτσι ο *LTS* είναι από τους πιο αξιόλογους ανθεκτικούς εκτιμητές σε πολυμεταβλητές παλινδρομήσεις. Η βασική αρχή του *LTS* είναι να προσαρμόσει το μοντέλο της παλινδρόμησης στα περισσότερα δεδομένα, τα δε outliers αναγνωρίζονται ως εκείνα τα σημεία, τα οποία βρίσκονται μακριά από την ανθεκτική προσαρμογή, οι περιπτώσεις με μεγάλα θετικά ή μεγάλα αρνητικά κατάλοιπα.

Γενικά, τα y_i μπορεί να έχουν οιαδήποτε μονάδα μέτρησης, και προκειμένου να αποφανθούμε εάν τα κατάλοιπα r_i είναι μεγάλα, χρειαζόμαστε μία εκτίμηση $\hat{\sigma}$ της κλίμακας των σφαλμάτων. Φυσικά, η εκτίμηση αυτή $\hat{\sigma}$ πρέπει να είναι ανθεκτική, και συγκεκριμένα μπορεί να υπολογισθεί από τον τύπο,

$$\hat{\sigma} = C\sqrt{\frac{1}{h}\sum_{i=1}^{h}(r^2)_{i:n}} \tag{16.66}$$

όπου C είναι ένας συντελεστής, ο οποίος επιλέγεται έτσι ώστε η $\hat{\sigma}$ να είναι συνεπής με την Gaussian κατανομή,

$$C = 1/\sqrt{-\frac{2n}{h\cdot a}\Phi(1/\alpha)}, \quad \text{όπου} \quad \alpha = 1/\Phi^{-1}\left(\frac{h+n}{2n}\right)$$

Το μειονέκτημα του *LTS*-εκτιμητή είναι ότι η παράμετρος h, η οποία ρυθμίζει την ανθεκτικότητα του *LTS* προσδιορίζεται εκ των προτέρων. Όπως αναφέρθηκε για να διατηρήσουμε ως μέγιστο σημείο κατάρρευσης 50%, η τιμή h πρέπει να είναι $h \simeq (n+p+1)/2$. Όμως, στην περίπτωση όπου τα δεδομένα δεν εμπεριέχουν outliers και ακολουθούν κανονική κατανομή ο *LTS* χάνει αποτελεσματικότητα.

Είναι γεγονός ότι οι εκτιμητές, οι οποίοι βασίζονται στην ανθεκτική κλίμακα $\hat{\sigma}$ δεν μπορούν να διατηρούν και τα δύο ένα υψηλό σημείο κατάρρευσης *HBP* και υψηλή αποτελεσματικότητα. Οι Rousseeuw και Leroy (1987) πρότειναν μία διαδικασία κατά την οποία αυξάνουν την αποτελεσματικότητα του *LTS* χωρίς να μειώσουν το *BP*. Πιο συγκεκριμένα, πρότειναν τη συνάρτηση βάρους W_i,

$$W_i = \begin{cases} 1 & \text{εάν} \quad |r_i/\hat{\sigma}| \le 2.5 \\ 0 & \text{εάν} \quad |r_i/\hat{\sigma}| > 2.5 \end{cases} \tag{16.67}$$

καλούμενη ως **hard rejection** συνάρτηση, όπου $\hat{\sigma}$ είναι η ανθεκτική κλίμακα της (16.66), ή της κανονικοποιημένης διαμέσου των απόλυτων τιμών των καταλοίπων. Μετά την επίλυση του προβλήματος (16.65) η συνάρτηση W_i ορίζει ποιες παρατηρήσεις τελικά θα παραμείνουν στο δείγμα και ποιες θα διαγραφούν ως outliers. Η επιλογή του 2.5 ως φραγμός των τυποποιημένων καταλοίπων είναι αυθαίρετη, αλλά δικαιολογημένη διότι σε μία κανονική κατανομή θα υπάρχουν μόνο μερικά κατάλοιπα μεγαλύτερα του $2.5\,\hat{\sigma}$.

16.5.4 Υπολογιστική Διαδικασία του LTS-εκτιμητή

Η ελαχιστοποίηση της αντικειμενικής συνάρτησης (16.65) για τον *LTS*, (16.63) για τον *S* και (16.62) για *LMS* είναι δύσκολη, διότι υπάρχουν πολλά τοπικά ελάχιστα. Ακόμη, οι "loss" συναρτήσεις των *LMS* και *LTS* δεν είναι διαφορίσιμες, και ως εκ τούτου δεν μπορούν να εφαρμοσθούν οι gradient μέθοδοι.

Πιο συγκεκριμένα, οι υπολογιστικές διαδικασίες των *S* και *LTS* εκτιμητών είναι επαναληπτικές, όπου η αντικειμενική συνάρτηση ελαττώνεται σε κάθε επανάληψη και έτσι οδηγεί σε τοπικό ελάχιστο. Αφού υπάρχουν αρκετά τέτοια ελάχιστα, ο υπολογιστικός αλγόριθμος αρχίζει από έναν μεγάλο α-

ριθμό τυχαίων υποσυνόλων ($N = 500$) των δεδομένων και επιλέγει στο τέλος το μικρότερο από τα τοπικά ελάχιστα. Η λύση αυτή είναι πιθανολογικά κοντά στην ολική βέλτιστη λύση του εκτιμητή.
Αξίζει να αναφέρουμε εδώ, ένα τμήμα του Fast-LTS αλγόριθμου που προτάθηκε από τον Rousseeuw και Driessen (2000) καλούμενο "concentration step" (C-step). Επιλέγοντας τυχαία ένα μικρό υποσύνολο παρατηρήσεων ($> p+1$) εκτιμάται το $\hat{\boldsymbol{\beta}}$ με *LS* και στη συνέχεια δοκιμάζονται μία μία οι υπόλοιπες παρατηρήσεις και επιλέγουμε αυτές με τα μικρότερα κατάλοιπα. Καθώς αυξάνεται ο αριθμός δεδομένων στο υποσύνολο, ανανεώνεται η εκτίμηση $\hat{\boldsymbol{\beta}}$, μέχρις ότου συμπληρωθούν h παρατηρήσεις. Το C-step εφαρμόζεται σε όλα τα αρχικά τυχαία υποσύνολα και τελικά επιλέγεται εκείνο το τοπικό βέλτιστο με το μικρότερο $\hat{\sigma}$.

PTS-εκτιμητής

Για έναν βέλτιστο συνδυασμό μεταξύ ανθεκτικότητας και αποτελεσματικότητας έχουν προταθεί αρκετοί άλλοι ανθεκτικοί εκτιμητές παλινδρόμησης. Ένας τέτοιος εκτιμητής, ο *PTS*, προτάθηκε από τους Pitsoulis και Zioutas (2009), ο οποίος ποινικοποιεί την "loss" συνάρτηση σε περίπτωση διαγραφής μιας παρατήρησης (outlier).

Πιο συγκεκριμένα, η *PTS*-εκτίμηση (Penalized Trimmed Squares) ορίζεται από τη λύση του προβλήματος,

$$\hat{\boldsymbol{\beta}}_{PTS} : \arg\min_{\hat{\boldsymbol{\beta}},h} \sum_{i=1}^{h} (r_i(\hat{\boldsymbol{\beta}})^2_{i:n} + \sum_{i=h+1}^{n} (c_i\hat{\sigma})^2 \tag{16.68}$$

όπου η παράμετρος h δεν δίνεται εκ των προτέρων αλλά προσδιορίζεται από τη βέλτιστη λύση της (16.68). Το «κόστος» διαγραφής μιας παρατήρησης με μεγάλο κατάλοιπο είναι $(c_i\hat{\sigma})^2$, όπου $\hat{\sigma}$ είναι μια ανθεκτική εκτίμηση κλίμακας, όπως και στην (16.66), το δε c_i εξαρτάται από τη μοχλότητα της κάθε παρατήρησης h_i,

$$c_i = c\sqrt{1-h_i} \tag{16.69}$$

Η βασική αρχή του *PTS* είναι η διαγραφή μιας παρατήρησης, εάν η βελτίωση στο πρώτο άθροισμα της (16.68) είναι μεγαλύτερη από το «κόστος» ποινικοποίησης

$$\frac{\left(r_i(\hat{\boldsymbol{\beta}})\right)^2}{1-h_i} > (c\hat{\sigma})^2 \qquad (16.70)$$

όπου $r_i / \sqrt{1-h_i}$ είναι το "adjusted" κατάλοιπο του οποίου το τετράγωνο ισοδυναμεί με την ελάττωση των τετραγώνων των καταλοίπων χωρίς την i παρατήρηση $(\boldsymbol{x}_i, y_i)$. Η παράμετρος αποκοπής c για μία καλή ανθεκτικότητα και αποτελεσματικότητα επιλέγεται ως $c = 2.5$.

Ο εκτιμητής *PTS* έχει υψηλό σημείο κατάρρευσης, *HBP*, αφού έχει τη δυνατότητα να διαγράψει όλα τα outliers δεδομένου ότι το $\hat{\sigma}$ είναι μία ανθεκτική εκτίμηση κλίμακας. Ταυτόχρονα, διατηρεί υψηλή αποτελεσματικότητα, αφού μια παρατήρηση διαγράφεται μόνο εάν προκαλεί σοβαρή αύξηση στα ελάχιστα τετράγωνα.

Το μεγαλύτερο πλεονέκτημα του *PTS* είναι η ικανότητά του να ανιχνεύει πολλαπλά outliers. Αυτό επιτυγχάνεται με την προϋπόθεση, ότι το $\hat{\sigma}$ και h_i στην (16.68) είναι ανθεκτικές εκτιμήσεις κλίμακας και μοχλότητας. Ανθεκτικές εκτιμήσεις μοχλότητας περιγράφονται αναλυτικότερα στο τελευταίο κεφάλαιο.

Παρατήρηση

Το «κόστος» ποινικοποίησης $(c_i\hat{\sigma})^2$ για τη διαγραφή μιας παρατήρησης $(\boldsymbol{x}_i, y_i)$ είναι σχετικά μικρό για σημεία μεγάλης μοχλότητας, όπως προκύπτει από την (16.69). Αυτό προφανώς σημαίνει ότι τα x-outliers διαγράφονται ευκολότερα, δεδομένου ότι η εκτίμηση $\hat{\boldsymbol{\beta}}_{PTS}$ ορίζεται από την ελαχιστοποίηση της αντικειμενικής συνάρτησης (16.68). ■

Υπολογιστικά, η λύση της (16.68) επιτυγχάνεται με τετραγωνικό μικτό ακέραιο προγραμματισμό (QMIP). Η λύση είναι μοναδική και ολικά βέλτιστη αλλά υπολογιστικά είναι χρονοβόρος, συνιστάται για μικρού μεγέθους προβλήματα $n < 100$ και $p < 3$. Παρόμοια και με τους προηγούμενους εκτιμητές έχει αναπτυχθεί ένας ταχύτατος αλγόριθμος, ο οποίος οδηγεί προσεγγιστικά στην ολική βέλτιστη λύση της (16.68).

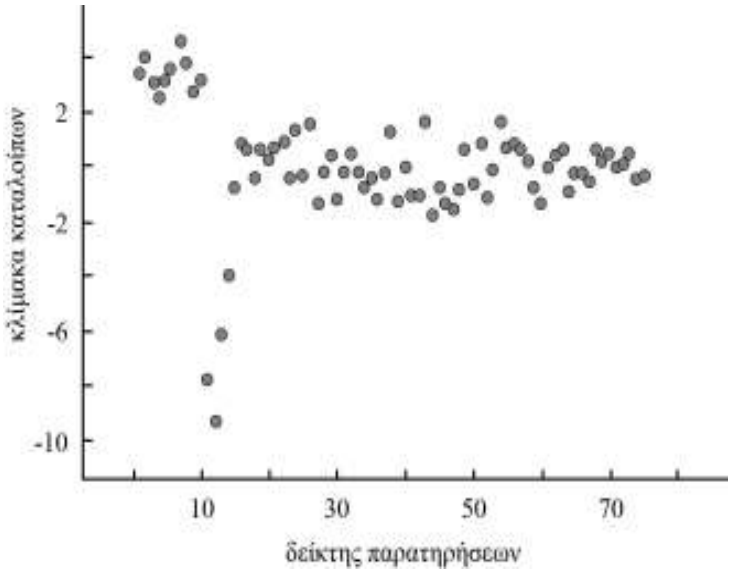
Παράδειγμα 16.4

Θα χρησιμοποιήσουμε τα τεχνητά δεδομένα των Hawkins-Bradu-Kass (1984) για να επιδείξουμε τα προσόντα των ανθεκτικών εκτιμητών. Τα τεχνητά δεδομένα προσφέρουν το πλεονέκτημα ότι γνωρίζουμε καλά ποια είναι τα «κακά» σημεία και έτσι μπορούμε να αξιολογήσουμε την αποτελεσματικότητα των διαφορετικών μεθόδων. Τα δεδομένα αυτά είναι 75, $(x_{1i}, x_{2i}, x_{3i}, y_i,\ i = 1,...,75)$, και περιέχονται στον Πίνακα 16.2. Οι πρώτες 10 παρατηρήσεις είναι σημεία «κακής» μοχλότητας ("bad" leverage points), και τα επόμενα 4 σημεία είναι «καλής μοχλότητας» ("good" leverage points, τα $\boldsymbol{x}_i$ είναι outliers, αλλά οι τιμές y_i προσαρμόζονται πολύ καλά στο μοντέλο παλινδρόμησης).

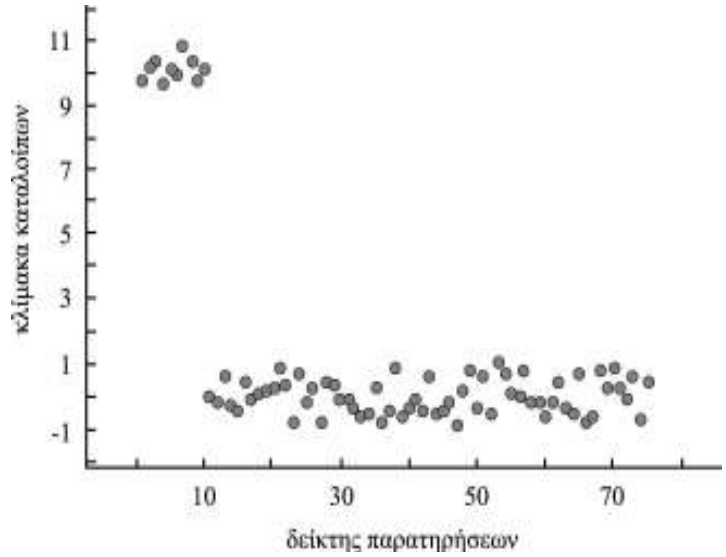
Χρησιμοποιώντας έναν *M*-εκτιμητή ή *GM*-εκτιμητή δεν οδηγούμαστε στα αναμενόμενα αποτελέσματα, επειδή τα outliers (τα «κακής» μόχλευσης σημεία) είναι επικαλυπτόμενα και τα 4 «καλής» μόχλευσης εμφανίζονται ως outliers επειδή σ' αυτά αντιστοιχούν μεγάλα κατάλοιπα κατά την *M* ή *GM* προσαρμογή. Αυτό δεν αποτελεί έκπληξη διότι όπως έχουμε αναφέρει οι *M* ή *GM*-εκτιμητές δεν μπορούν να αντιμετωπίσουν μεγαλύτερο ποσοστό μόλυνσης των δεδομένων.

Αν συγκεντρωθούμε τώρα στα αποτελέσματα των *HBP* ανθεκτικών εκτιμητών, όπως *LMS*, *LTS*, *M*, *S* και *PTS* παρατηρούμε τα ακόλουθα. Από το Σχήμα 16.6, όπου εφαρμόσθηκε η *LS* εκτίμηση, εμφανίζονται οι παρατηρήσεις 11, 12 και 13 ως outliers επειδή βρίσκονται εκτός του διαστήματος ±2.5. Ατυχώς, από τη δημιουργία των δεδομένων γνωρίζουμε ότι αυτές είναι καλές παρατηρήσεις. Τα σημεία «κακής» μόχλευσης έχουν προσελκύσει την *LS* προσαρμογή προς την κατεύθυνσή τους, και ως εκ τούτου τα 10 πρώτα σημεία έχουν μικρά τυποποιημένα *LS* κατάλοιπα. Παρόμοια αποτελέσματα έδειξαν και οι *M* ή *GM*-εκτιμητές.

Από την άλλη πλευρά στο Σχήμα 16.7 αναδεικνύεται ότι οι *HBP* ανθεκτικοί εκτιμητές (*LMS*, *LTS*, *M*, *S* και *PTS*) αναγνωρίζουν τα πρώτα 10 σημεία σαν παρατηρήσεις μεγάλης επίδρασης. Τα τέσσερα σημεία «καλής» μόχλευσης πέφτουν στην γειτονική περιοχή του μηδέν. Αυτό σημαίνει ότι αυτά τα σημεία ταιριάζουν με την προσαρμογή των *HBP* εκτιμητών. Τέλος, οι *HBP* εκτιμήσεις συμφωνούν με την κατασκευή των δεδομένων.



Σχήμα 16.6 Δεδομένα Hawkins – Bradu – Kass, παλινδρόμηση με *LS* εκτίμηση



Σχήμα 16.7 Δεδομένα Hawkins – Bradu – Kass. Γραφική παράσταση παλινδρόμησης με *HBP* εκτίμηση

16.6 ΣΥΜΠΕΡΑΣΜΑ

Σε αυτό το κεφάλαιο περιγράφηκαν διάφοροι τρόποι αντιμετώπισης των outliers στην ανάλυση παλινδρόμησης. Η μεγαλύτερη δυσκολία αντιμετώπισής τους παρουσιάζεται στα $\boldsymbol{x}$-outliers διότι στην πολλαπλή παλινδρόμηση συνήθως εμφανίζεται το πρόβλημα επικάλυψης και η ανίχνευσή τους δεν είναι εύκολη. Ακόμη, $\boldsymbol{x}$-outliers «καλής» μοχλότητας συμβάλλουν στην ακρίβεια εκτίμησης του μοντέλου παλινδρόμησης. Από τις πιο σύγχρονες ανθεκτικές και αποτελεσματικές μεθόδους ξεχωρίζουν οι *GM*, *MM*, *LTS*, *S* και *PTS*. Οι περισσότερες από αυτές είναι διαθέσιμες στα γνωστικά στατιστικά πακέτα προγραμμάτων, όπως *SPLUS*, *SAS*, και επίσης μπορεί κανείς να χρησιμοποιήσει ελεύθερα τις αντίστοιχες ρουτίνες από τα στατιστικά προγράμματα *R* του διαδικτύου.