**Ελληνικό Στατιστικό Ινστιτούτο**
Πρακτικά 18ου Πανελληνίου Συνεδρίου Στατιστικής (2005) σελ.39-48

# ΜΙΑ ΜΕΘΟΔΟΣ "ΤΙΜΩΡΙΑΣ" ΓΙΑ ΤΗΝ ΑΠΟΡΡΙΨΗ ΑΚΡΑΙΩΝ ΤΙΜΩΝ ΣΤΗΝ ΑΝΘΕΚΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

**Αντώνης Αβραμίδης και Γιώργος Ζιούτας**
*Γενικό Τμήμα Πολυτεχνικής Σχολής Α.Π.Θ.*
aavram@gen.auth.gr

**ΠΕΡΙΛΗΨΗ**

Στην ανθεκτική παλινδρόμηση, συχνά χρειάζεται να αποφανθούμε πόσες είναι οι ασυνήθιστες παρατηρήσεις (outliers) στα δεδομένα και ποιες από αυτές πρέπει να απομακρυνθούν με σκοπό την εύρεση της καλύτερης γραμμής παλινδρόμησης στις υπόλοιπες παρατηρήσεις. Μία βασική μέθοδος είναι η **LTS** (least trimmed squares), η οποία αναφέρεται στην προσαρμογή της γραμμής στην πλειοψηφία των δεδομένων (~51%) αναγνωρίζοντας ως ακραίες παρατηρήσεις τα υπόλοιπα σημεία (~49%) που προκαλούν τη μεγαλύτερη ζημιά στην προσαρμογή της ευθείας. Στην προτεινόμενη μέθοδο, εισάγουμε κόστος τιμωρίας για την απόρριψη κάθε ακραίας παρατήρησης (outlier), επομένως, η καλύτερη προσαρμογή της γραμμής στην πλειοψηφία των δεδομένων επιτυγχάνεται απομακρύνοντας μόνο τις παρατηρήσεις που επιδρούν κακώς στην παλινδρόμηση. Η προτεινόμενη ανθεκτική εκτιμήτρια παλινδρόμησης προκύπτει λύνοντας ένα κυρτό μικτό ακέραιο τετραγωνικό πρόβλημα και για την σύγκριση της αποτελεσματικότητας και ανθεκτικότητάς της με άλλες ανθεκτικές εκτιμήτριες, διεξάγουμε μελέτη προσομοίωσης (Monte Carlo Simulation) με κατάλληλα αριθμητικά δεδομένα.

## 1. ΕΙΣΑΓΩΓΗ

Έστω το γραμμικό μοντέλο παλινδρόμησης

$$y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + u\,, \tag{1.1}$$

όπου $y$ η εξαρτημένη μεταβλητή, $\mathbf{x}$ το $p$x1 διάνυσμα των ανεξάρτητων μεταβλητών, $\boldsymbol{\beta}$ το $p$x1 διάνυσμα των αγνώστων παραμέτρων και $u$ το διάνυσμα των τυχαίων σφαλμάτων με μέση τιμή μηδέν και διακύμανση $\sigma^2$. Παρατηρούμε ένα τυχαίο δείγμα $(\mathbf{x}_1,\ y_1),...,(\mathbf{x}_n,\ y_n)$ και επιθυμούμε να βρούμε μία ανθεκτική εκτιμήτρια εννοώντας ότι η επίδραση κάθε παρατήρησης $(\mathbf{x}_i,\ y_i)$ στη δειγματική εκτιμήτρια περιορίζεται. Όπως είναι γνωστό, η εκτιμήτρια ελαχίστων τετραγώνων του $\boldsymbol{\beta}$ δεν ικανοποιεί τις προϋποθέσεις της ανθεκτικότητας.

Οι γενικευμένες Μ εκτιμήτριες (GM estimators), όπως οι προτάσεις των Mallows (1975), Hampel (1978), Krasker and Welsch (1982) και άλλες ανθεκτικές εκτιμήτριες

σχεδιάστηκαν με σκοπό να προστατέψουν την εκτιμήτρια από τις ακραίες παρατηρήσεις. Πιο συγκεκριμένα, έχουν συναρτήσεις επίδρασης περιορισμένες τόσο στα **x** όσο και στα y. Ατυχώς, αυτές οι περιορισμένης επίδρασης εκτιμήτριες έχουν χαμηλά σημεία κατάρρευσης.

Έχουν προταθεί αρκετές εκτιμήτριες υψηλού σημείου κατάρρευσης, γνωστές ως HBP (high breakdown point), οι οποίες επιτυγχάνουν σημεία κατάρρευσης κοντά στο 50%. Ανάμεσα σε αυτές είναι η ελάχιστη διάμεσος τετραγώνων (least median of squares estimator LMS) του Rousseeuw (1984) και η least trimmed squares (LTS) του Rousseeuw (1984), Rousseeuw and Leroy (1987). Η LTS μπορεί να θεωρηθεί ως μία εναλλακτική ανθεκτική μέθοδος παλινδρόμησης των GM-μεθόδων, η οποία ορίζει μια εκτιμήτρια των παραμέτρων ($\alpha$, $\boldsymbol{\beta}$), ελαχιστοποιώντας το άθροισμα τετραγώνων των υπολοίπων στο υποσύνολο των [(n+p+1)/2] καλύτερων σημείων.

Κάποιες καλύτερες προτάσεις εκτιμητριών επιτυγχάνουν υψηλά σημεία κατάρρευσης και ταυτόχρονα βελτιώνουν την αποτελεσματικότητα των HBP εκτιμητριών. Ανάμεσα σε αυτές είναι οι S εκτιμήτριες των Rousseeuw and Yohai (1984), οι MM εκτιμήτριες του Yohai (1987) και Yohai and Zamar (1988), οι οποίες συνδυάζουν καλή ασυμπτωτική αποτελεσματικότητα στο κανονικό γραμμικό μοντέλο με HBP. Οι εκτιμήτριες αυτές επιτυγχάνουν καλές ασυμπτωτικές ιδιότητες, μπορεί όμως να έχουν χαμηλή αποτελεσματικότητα σε πεπερασμένα δείγματα εάν το δείγμα περιέχει σημεία υψηλής επίδρασης (high leverage points).

Με σκοπό να επιτευχθούν ταυτόχρονα εκτιμήτριες παλινδρόμησης περιορισμένης επίδρασης, υψηλού σημείου κατάρρευσης και αποτελεσματικότητας, οι Coakley and Hettmansperger (1993), Simpson, Ruppert and Carroll (1992) πρότειναν την εκτιμήτρια ενός βήματος S1S, ξεκινώντας με αρχικές εκτιμήτριες υψηλού σημείου κατάρρευσης και χρησιμοποιώντας το σχήμα του Schweppe με συντελεστές βαρύτητας του Mallows. Οι εκτιμήτριες αυτές βελτιώνουν την αποτελεσματικότητα. Όμως, η απόδοσή τους στην πράξη εξαρτάται κατά μεγάλο ποσοστό από τις αρχικές τιμές του LTS.

Ο σκοπός αυτής της εργασίας είναι να προτείνει μια νέα ανθεκτική προσέγγιση η οποία βασίζεται μόνο σε ανθεκτική εκτιμήτρια κλίμακας καταλοίπων (παρέχεται από την LTS μέθοδο) και σε ανθεκτικά βάρη σχεδιασμού. Για την κατασκευή της νέας εκτιμήτριας παλινδρόμησης NQMIP, η οποία συνδυάζει υψηλό σημείο κατάρρευσης με καλή αποτελεσματικότητα, προτείνουμε νέα αντικειμενική συνάρτηση (loss function), η οποία αποτελείται από το άθροισμα των τετραγώνων των καταλοίπων και του κόστους τιμωρίας για την απομάκρυνση των ακραίων παρατηρήσεων. Στην ουσία, το πέναλτυ κόστος είναι μια συνάρτηση της ανθεκτικής κλίμακας $\sigma$ και των καταλοίπων των δεδομένων.

Η αντικειμενική συνάρτηση της νέας εκτιμήτριας παρουσιάζεται στη $2^{\eta}$ ενότητα. Η ελαχιστοποίηση της προτεινόμενης αντικειμενικής συνάρτησης φορμάρεται ως μικτό ακέραιο τετραγωνικό πρόβλημα προγραμματισμού στην $3^{\eta}$ ενότητα. Τα αποτελέσματα της NQMIP εκτιμήτριας παρουσιάζονται χρησιμοποιώντας προσομοιώσεις Monte-Carlo στην $4^{\eta}$ ενότητα. Τέλος, συμπεράσματα, μελλοντική έρευνα και κάποια υπολογιστικά θέματα αναφέρονται στην $5^{\eta}$ ενότητα.

## 2. Η ΠΡΟΤΕΙΝΟΜΕΝΗ ΕΚΤΙΜΗΤΡΙΑ

### 2.1 ΟΡΙΣΜΟΣ

Θεωρούμε την εκτιμήτρια η οποία ορίζεται ως η λύση στο πρόβλημα

$$\underset{\beta}{\text{ελαχιστοποίηση}} \quad \sum_{i=1}^{n} \rho_{c_i\sigma}(u_i) \tag{2.1}$$

όπου $\rho_{c_i\sigma}$ είναι η αντικειμενική συνάρτηση η οποία ορίζεται ως

$$\rho_{c_i\sigma}(u_i) = \begin{cases} u_i^2 & \text{για} \quad |u_i| < c_i\sqrt{1-h_i}\,\sigma \\ (c_i\sigma)^2 & \text{για} \quad |u_i| \geq c_i\sqrt{1-h_i}\,\sigma \end{cases} \tag{2.2}$$

όπου $\sigma$ είναι κάποια ανθεκτική εκτιμήτρια κλίμακας των καταλοίπων u, $c_i$ τα σημεία αποκοπής για ακραίες παρατηρήσεις τα οποία εξαρτώνται από τα σημεία σχεδιασμού, $c_i = c \cdot (w(x_i))$ , $w(x_i)$ είναι τα βάρη σχεδιασμού και $c$ η γνωστή παράμετρος αποκοπής. Η επιλογή της παραμέτρου $c$ ρυθμίζει την ανθεκτικότητα και την αποτελεσματικότητα της εκτιμήτριας.

Η προτεινόμενη αντικειμενική συνάρτηση όπως φαίνεται στην (2.2) είναι απλή. Για μεγάλα κατάλοιπα $u_i$ ($|u_i| \geq c_i\sigma$) το άθροισμα των τετραγώνων των καταλοίπων είναι λιγότερο απότομα αυξανόμενο, αφού το μεγάλο τετραγωνικό κατάλοιπο ελαττώνεται στην τιμή $(c_i\sigma)^2$, η οποία εξαρτάται μόνο από τα σημεία σχεδιασμού και την αρχική εκτιμήτρια κλίμακας $\sigma$. Η τιμή αυτή, $(c_i\sigma)^2$, ερμηνεύεται ως το κόστος *τιμωρίας* για την απομάκρυνση μιας ακραίας παρατήρησης.

Μια βέλτιστη εφικτή λύση του προβλήματος (2.1) επιτυγχάνεται χρησιμοποιώντας μαθηματικό προγραμματισμό. Το πρόβλημα (2.1) φορμάρεται ως ένα κυρτό μικτό ακέραιο τετραγωνικό πρόβλημα προγραμματισμού. Η τεχνική αυτή αναπτύσσεται στην 3η ενότητα.

### 2.2 ΙΔΙΟΤΗΤΕΣ

Αν και στην παρούσα εργασία το θεωρητικό υπόβαθρο για τις ιδιότητες της προτεινόμενης εκτιμήτριας δεν αναπτύσσεται ικανοποιητικά, κάποιες καλές ιδιότητές της βασίζονται σε λογικές εξηγήσεις οι οποίες προέρχονται από το πρόβλημα ελαχιστοποίησης (2.1) και την εξίσωση (2.2).

***Συνέπεια (Consistency):*** Η εκτιμήτρια ορίζεται από τη λύση του προβλήματος ελαχιστοποίησης (2.1). Το πρόβλημα αυτό είναι ισοδύναμο με την απόφαση πόσες και ποιες από τις ακραίες παρατηρήσεις πρέπει να απομακρυνθούν ώστε να προκύψει το ελάχιστο άθροισμα τετραγώνων των καταλοίπων και του κόστους τιμωρίας. Μετά την απομάκρυνση των ακραίων παρατηρήσεων, η Εκτιμήτρια Ελαχίστων Τετραγώνων (OLS estimator) μπορεί να εφαρμοστεί στα καθαρά δεδομένα.

***Ανθεκτικότητα (Robustness):*** Στο πρόβλημα ελαχιστοποίησης (2.1), το κόστος τιμωρίας $(c_i\sigma)^2$ για την απομάκρυνση μιας ακραίας παρατήρησης εξαρτάται από τον πίνακα σχεδιασμού και την κλίμακα των καταλοίπων. Στα σημεία υψηλής επίδρασης

το κόστος τιμωρίας μειώνεται επειδή τα βάρη $w(x_i)$ γενικά μικραίνουν στα σημεία μοχλούς. Επομένως, τα σημεία υψηλής επίδρασης τείνουν να απομακρυνθούν εκτός αν τα κατάλοιπά τους είναι μικρά. Επιπλέον, για παρατηρήσεις που είναι y-outliers, μια απομάκρυνση πραγματοποιείται μόνο στην περίπτωση που $|u_i| \geq c_i\sigma$.

***Αποτελεσματικότητα (Efficiency):*** Η προτεινόμενη ανθεκτική διαδικασία απορρίπτει μόνο τις καταστροφικές παρατηρήσεις, αφού υπάρχει κόστος τιμωρίας $(c_i\sigma)^2$ για κάθε απομάκρυνση, όπως φαίνεται στην (2.2). Γενικά, μία παρατήρηση απομακρύνεται αν το τελικό κατάλοιπο $u_i$ είναι μεγάλο ($|u_i| \geq c_i\sigma$). Τελικά, η καλύτερη προσαρμογή μοντέλου αποκτάται στα καθαρά δεδομένα.

***Υψηλό Σημείο Κατάρρευσης (High Break Down-Point)***: Η προτεινόμενη εκτιμήτρια βασίζεται σε εκτιμήτρια των καταλοίπων $\sigma$ υψηλού σημείου κατάρρευσης όπως προκύπτει από τον LTS και σε ανθεκτική εκτίμηση των βαρών $w(x_i)$, τα οποία εξαρτώνται μόνο από τα σημεία σχεδιασμού.

Υπό αυτές τις συνθήκες, η προτεινόμενη εκτιμήτρια κληρονομεί υψηλό σημείο κατάρρευσης όπως παρουσιάζεται μέσω προσομοιώσεων Monte Carlo στην 4η ενότητα.

# 3. ΠΡΟΤΥΠΑ ΜΙΚΤΟΥ ΑΚΕΡΑΙΟΥ ΤΕΤΡΑΓΩΝΙΚΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ

## 3.1 Η ΕΚΤΙΜΗΤΡΙΑ LTS

Αφού θεωρούμε ότι η προτεινόμενη εκτιμήτρια είναι μία βελτίωση της LTS, δίνουμε τον ορισμό της εκτιμήτριας LTS και αναπτύσσουμε μία μαθηματική φόρμα για την επίλυσή της. Στη συνέχεια (επόμενη παράγραφο) τροποποιούμε την φόρμα για να επιτύχουμε την προτεινόμενη μέθοδο «τιμωρίας»

Η εκτιμήτρια LTS (Least Trimmed Squares) η οποία προτάθηκε από τον Rousseeuw (1984) ορίζεται σαν το $p$-διάνυσμα

$$\boldsymbol{\beta}_{\mathrm{LTS}} = \underset{\boldsymbol{\beta}}{\text{ελαχιστοποίηση}}\, Q_{\mathrm{LTS}}(\boldsymbol{\beta}) \tag{3.1}$$

όπου $Q_{LTS}(\boldsymbol{\beta}) = \sum_{\mathrm{i}=1}^{\mathrm{h}} \mathrm{u}_{\mathrm{i}}^2$, $\mathrm{u}_1^2 \leq \mathrm{u}_2^2 \leq \mathrm{u}_3^2 \leq \ldots \leq \mathrm{u}_{\mathrm{h}}^2$ είναι σε αύξουσα σειρά τα τετράγωνα των καταλοίπων και h είναι η παράμετρος που ρυθμίζει το σημείο κατάρρευσης και για το υψηλότερο σημείο κατάρρευσης έχει την τιμή $h=(n+p+1)/2$.

Για την επίλυση του προβλήματος (3.1) αναπτύσσουμε την παρακάτω φόρμουλα μικτού ακέραιου τετραγωνικού προγραμματισμού:

$$\underset{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, u_i, \varepsilon_i, \delta_i}{\text{ελαχιστοποίηση}} \sum_{i=1}^{n} u_i^2 \tag{3.2}$$

σύμφωνα με τους περιορισμούς:

$$\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + u_i \geq y_i - \varepsilon_i$$
$$\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 - u_i \leq y_i + \varepsilon_i$$
$$\varepsilon_i \leq \delta_i K$$
$$\sum_{i=1}^{n} \delta_i = n - (n + p + 1)/2$$
$$\delta_i : (0,1) \text{ μεταβλητή}$$
$$\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, u_i, \varepsilon_i \geq 0 \quad \gamma\iota\alpha \quad i = 1,...,n$$

όπου $\varepsilon_i$ παριστάνει το μέγεθος προσέλκυσης των $y_i$ που απαιτείται για να μηδενισθούν τα μεγάλα κατάλοιπα $u_i$ Στην φόρμα αυτή οι μεταβλητές απόφασης είναι όλες θετικές και είναι οι άγνωστοι συντελεστές παλινδρόμησης $\boldsymbol{\beta}_1^{\mathrm{T}}=(\beta_{11},...\beta_{1p})$, $\boldsymbol{\beta}_2^{\mathrm{T}}=(\beta_{21},...,\beta_{2p})$, $\boldsymbol{\beta}=\boldsymbol{\beta}_1-\boldsymbol{\beta}_2$, τα κατάλοιπα $u_i$ και οι αποστάσεις προσέλκυσης $\varepsilon_i$.

Οι πρώτοι δύο περιορισμοί περιγράφουν γεωμετρικά την παλινδρόμηση για την περίπτωση των θετικών ή αρνητικών καταλοίπων αφού τα $u_i$ είναι μόνο θετικά.

Ακόμη, $\delta_i$ είναι μια δυαδική μεταβλητή λήψης αποφάσεων, τέτοια ώστε: Για $\boldsymbol{\delta_i=1}$, ο τρίτος περιορισμός επιτρέπει την απόσταση προσέλκυσης $\varepsilon_i$ να πάρει οποιοδήποτε άνω όριο $K$ στο μέγιστο απόλυτο κατάλοιπο όπου ή $y_i$ προσελκύεται προς την εκτιμώμενο πολυεπίπεδο της παλινδρόμησης έτσι ώστε να μειωθεί το κατάλοιπο στο μηδέν, $u_i$ =0. Για $\boldsymbol{\delta_i=0}$, ο τρίτος περιορισμός θέτει την απόσταση προσέλκυσης $\varepsilon_i$ στο μηδέν, οπότε, η τιμή $y_i$ δεν προσελκύεται προς την εκτιμώμενη γραμμή.

Στον τέταρτο περιορισμό υποδεικνύεται ότι ο αριθμός των διαγραφόμενων σημείων είναι (n-(n+p+1)/2), ο οποίος είναι ο μέγιστος αριθμός των σημείων που απορρίπτονται στην εκτιμήτρια LTS.

Το πρόβλημα (3.1) έχει κυρτή αντικειμενική συνάρτηση και οι περιορισμοί σχηματίζουν κυρτό σύνολο Arthanari and Dodge (1993), γι' αυτό η μέθοδος simplex ψάχνοντας την λύση σταματά στο σημείο $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{u}, \boldsymbol{\varepsilon})$, το οποίο είναι και η μοναδική ολική βέλτιστη λύση.

## 3.2 Η ΠΡΟΤΕΙΝΟΜΕΝΗ ΕΚΤΙΜΗΤΡΙΑ NQMIP

Με σκοπό την απομάκρυνση μόνο των καταστροφικών ακραίων παρατηρήσεων, οι Zioutas and Avramidis (2005) πρότειναν κόστος *τιμωρίας* για κάθε απομάκρυνση μιας ακραίας παρατήρησης (outlier). Έτσι το πρόβλημα ελαχιστοποίησης (3.2) διαμορφώθηκε στην ακόλουθη φόρμουλα :

$$\underset{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, u_i, \varepsilon_i, \delta_i}{\text{ελαχιστοποίηση}} \sum_{i=1}^{n} (u_i^2 + \delta_i (c_i \sigma)^2) \qquad (3.3)$$

σύμφωνα με τους περιορισμούς:

$$\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + u_i \geq y_i - \varepsilon_i$$

$$\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 - u_i \leq y_i + \varepsilon_i$$

$$\varepsilon_i \leq \delta_i K$$

$$\delta_i : (0,1) \text{ μεταβλητή}$$

$$\boldsymbol{\beta}_{1i}, \boldsymbol{\beta}_{2i}, u_i, \varepsilon_i \geq 0 \qquad \gamma\iota\alpha \quad i = 1,..., n$$

Για ***δ*****$_i$*****=1,*** ο τρίτος περιορισμός επιτρέπει την απόσταση προσέλκυσης $\varepsilon_i$ να πάρει οποιαδήποτε τιμή έτσι ώστε να μειωθεί το κατάλοιπο στο μηδέν, $u_i$ =0, η αντικειμενική συνάρτηση όμως αυξάνει κατά ένα σταθερό κόστος τιμωρίας $(c_i\sigma)^2$. Στην περίπτωση αυτή θεωρούμε ότι το σημείο $(\mathbf{x_i}, y_i)$ δεν επηρεάζει πλέον την εκτιμήτρια παλινδρόμησης. Η εκτιμήτρια που προκύπτει από την λύση του προβλήματος (3.3) συμβολίζεται με QMIP.

Στην παρούσα εργασία διερευνήθηκε ένα καλύτερο κόστος τιμωρίας έτσι ώστε να μην απορρίπτονται από το δείγμα σημεία τα οποία είναι καλοί μοχλοί (good leverage points) και συμβάλουν στην ακρίβεια της εκτιμήτριας. Το κόστος τιμωρίας πρέπει να είναι τέτοιο ώστε να απορρίπτονται σημεία των οποίων τα τελικά κατάλοιπα είναι μεγαλύτερα από τρεις φορές το τυποποιημένο σφάλμα $|u_i| > 3\sqrt{1-h_i}\,\sigma$, όπου $h_i$ είναι το *i* διαγώνιο στοιχείο του πίνακα *H* (*Hat*) και η τιμή του δηλώνει αν ένα σημείο είναι μοχλός (*leverage*).

Ατυχώς όμως, όταν υπάρχει μία ομάδα από μοχλούς (leverage points) στα δεδομένα, συμβαίνει το φαινόμενο της επικάλυψης και το $h_i$ αποτυγχάνει να υποδείξει όλα τα σημεία που είναι πράγματι μοχλοί. Για να αποφευχθούν τέτοιες συνέπειες στην παραπάνω εκτιμήτρια QMIP, το κόστος τιμωρίας για παρατηρήσεις με μεγάλη επίδραση (high leverage points) ελαφρύνεται χρησιμοποιώντας τα βάρη $w(x_i)$, όπως υπέδειξε ο Hampel (1978) για τον εκτιμητή Mallows, προκειμένου να περιορίσει την επίδραση των μοχλών σε μία παλινδρόμηση.

Μέχρι τώρα υπάρχουν εναλλακτικές προτάσεις για την ελάφρυνση των δυνατών μοχλών σε μία παλινδρόμηση. Στη νέα προτεινόμενη μέθοδο *τιμωρίας* τίθεται κατώτερο φράγμα στο κόστος απόρριψης μιας παρατήρησης, έτσι ώστε να παραμένει στο δείγμα όταν το τελικό της κατάλοιπο είναι μικρό, μικρότερο του 1,5σ. Τελικά, ένα κόστος τιμωρίας που επιτρέπει ευκολότερα την απόρριψη των σημείων που είναι κακοί μοχλοί (bad leverage) στην παλινδρόμηση αλλά διατηρεί τους καλούς μοχλούς (good leverage) στο δείγμα είναι το ακόλουθο $(c_i\sigma)^2 = \max[(1.5\sigma)^2, (3w_i\sigma)^2]$. Τη νέα εκτιμήτρια συμβολίζουμε με **NQMIP**.

## 4. ΑΠΟΤΕΛΕΣΜΑΤΑ MONTE CARLO

Για την αξιολόγηση της αποδοτικότητας της ανθεκτικής μας εκτιμήτριας, διεξάγουμε μελέτη προσομοίωσης συγκρίνοντάς την με άλλες γνωστές ανθεκτικές

εκτιμήτριες. Για τη διεξαγωγή ενός πειράματος προχωρούμε ως εξής. Δίνονται οι κατανομές των ανεξάρτητων μεταβλητών και οι τιμές των παραμέτρων. Σφάλματα παράγονται σύμφωνα με μια κατανομή των σφαλμάτων και παρατηρήσεις, $y_i$,προκύπτουν ακολουθώντας το μοντέλο παλινδρόμησης

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

Θεωρούμε πως το δείγμα μπορεί να περιέχει τρεις τύπους ακραίων τιμών, "bad" leverage points (κακός μοχλός), "good" leverage points (καλός μοχλός) και y-outliers.

Επιλέγουμε δείγματα μεγέθους $n$=50 και βαθμό του μοντέλου παλινδρόμησης $p$=2 με συντελεστές $\beta_1$=1.20, $\beta_2$=-0.80 και σταθερό όρο $\beta_0$=0.0, ενώ προτιμούμε την κατανομή Gauss ως την κατανομή των σφαλμάτων, $u \sim N(0, \sigma=16)$.

Θεωρούμε πέντε διαφορετικούς τύπους ανθεκτικών εκτιμητριών: την LTS, την MM, την S1S, την QMIP και τη νέα προτεινόμενη εκτιμήτρια NQMIP (New QMIP).

Οι υπολογισμοί των ανθεκτικών εκτιμητριών προέκυψαν από τη χρήση του στατιστικού πακέτου **S-plus**, έναν κώδικα του MINITAB των Coakley και Hettmansperger (1993) και τον επιλυτή FortMP/QMIP-Fortran Code.

Για να πετύχουμε αξιόπιστα αποτελέσματα στη Monte Carlo προσομοίωση, συνήθως απαιτείται ένας μεγάλος αριθμός επαναλήψεων, π.χ. 1000, Hawkins and Olive (1999). Παρόλα αυτά, για τις προτεινόμενες εκτιμήτριες QMIP και NQMIP, οι 100 επαναλήψεις ήταν αρκετές για να επιτύχουμε ακρίβεια μικρότερη από 10%, δηλαδή $(\beta - \hat{\beta})/\beta < 10\%$, με επίπεδο εμπιστοσύνης τουλάχιστον 90%.

Η απόδοση των ανθεκτικών εκτιμητριών μετρήθηκε από τις εκτιμήσεις της προσομοίωσης σύμφωνα με τα κριτήρια: μέση εκτίμηση του $\hat{\beta}_i$, διακύμανση του $\hat{\beta}_i$, νόρμα μεροληψίας του $\hat{\boldsymbol{\beta}}$, μέσο τετραγωνικό σφάλμα προσαρμογής.

Όλα τα παρακάτω συμπεράσματα προέκυψαν μετά από προσεκτική εξέταση καθεμιάς από τις εκτιμήτριες.

**ΠΙΝΑΚΑΣ 1**

Αρ. των "bad" leverage points 6, "good" leverage points 4, y-outliers 6,
παράμετροι παλινδρόμησης: $\beta_0 = 0.00$, $\beta_1 = 1.20$, $\beta_2 = -0.80$

| εκτιμήτριες | εκτίμηση του $\hat{\beta}_0$ | εκτίμηση του $\hat{\beta}_1$ | εκτίμηση του $\hat{\beta}_2$ | διακύμ. του $\hat{\beta}_0$ | διακύμ. του $\hat{\beta}_1$ | διακύμ. του $\hat{\beta}_2$ | νόρμα μερολ. του $\hat{\boldsymbol{\beta}}$ | μέσο τετραγ. σφάλμα προσαρ. | μέσος χρόνος υπολογ. |
|---|---|---|---|---|---|---|---|---|---|
| LTS | -0.671 | 1.008 | -0.677 | 98.547 | 0.309 | 0.059 | 7.775 | 353 | <1 sec |
| MM | 1.817 | 0.980 | -0.750 | 71.265 | 0.253 | 0.014 | 5.974 | 314 | <1 sec |
| QMIP | -0.352 | 1.177 | -0.784 | 22.702 | 0.059 | 0.005 | 4.071 | 284 | 10 sec |
| S1S | 8.541 | 0.962 | -0.940 | 145.800 | 0.240 | 0.010 | 9.544 | 344 | <1 sec |
| **NQMIP** | **-0.272** | **1.158** | **-0.800** | **14.772** | **0.003** | **0.005** | **3.107** | **272** | **10 sec** |

**ΠΙΝΑΚΑΣ 2**

Αρ. των "bad" leverage points 6, "good" leverage points 0,
παράμετροι παλινδρόμησης: $\beta_0 = 0.00$, $\beta_1 = 1.20$, $\beta_2 = -0.80$

| εκτιμήτριες | εκτίμηση του $\hat{\beta}_0$ | εκτίμηση του $\hat{\beta}_1$ | εκτίμηση του $\hat{\beta}_2$ | διακύμ. του $\hat{\beta}_0$ | διακύμ. του $\hat{\beta}_1$ | διακύμ. του $\hat{\beta}_2$ | νόρμα μερολ. του $\hat{\boldsymbol{\beta}}$ | μέσο τετραγ. σφάλμα προσαρ. | μέσος χρόνος υπολογ. |
|---|---|---|---|---|---|---|---|---|---|
| LTS | 4.946 | 0.874 | -0.767 | 229.271 | 0.345 | 0.075 | 11.445 | 378 | <1 sec |
| MM | 1.041 | 1.035 | -0.740 | 48.430 | 0.131 | 0.028 | 5.474 | 298 | <1 sec |
| QMIP | 1.034 | 1.099 | -0.761 | 27.156 | 0.064 | 0.022 | 4.183 | 283 | 10 sec |
| S1S | 3.057 | 0.805 | -0.822 | 102.853 | 0.261 | 0.044 | 6.987 | 327 | <1 sec |
| **NQMIP** | **0.906** | **1.102** | **-0.761** | **22.559** | **0.028** | **0.022** | **4.127** | **282** | **10 sec** |

Στους Πίνακες 1-3 παρουσιάζονται τα αποτελέσματα της προσομοίωσης για τις πέντε εκτιμήτριες όσον αφορά τα κριτήρια που αναφέρθηκαν παραπάνω. Και στις τρεις περιπτώσεις, είναι φανερό ότι η νέα εκτιμήτρια NQMIP πλεονεκτεί έναντι των άλλων ως προς το μέσο τετραγωνικό σφάλμα προσαρμογής, το οποίο είναι και το βασικό από τα παραπάνω κριτήρια ως προς την αποτελεσματικότητα των εκτιμητριών. Πιο συγκεκριμένα, το μέσο τετραγωνικό σφάλμα προσαρμογής, δοθέντων των αληθινών τιμών των συντελεστών της παλινδρόμησης ($\beta_0 = 0.00$, $\beta_1 = 1.20$, $\beta_2 = -0.80$), είναι 256, οπότε με τη νέα προσέγγιση NQMIP υπάρχει σημαντική βελτίωση της τάξης του 20%. Ακολουθεί η προηγούμενη προσέγγιση QMIP, εκτός της περίπτωσης μολύνσεως των δεδομένων μόνο με "good" leverage points του Πίνακα 3, στην οποία η εκτιμήτρια MM υπερισχύει έναντι των υπολοίπων τριών εκτιμητριών.

**ΠΙΝΑΚΑΣ 3**

Αρ. των "bad" leverage points 0, "good" leverage points 6,
παράμετροι παλινδρόμησης: $\beta_0 = 0.00$, $\beta_1 = 1.20$, $\beta_2 = -0.80$

| εκτιμήτριες | εκτίμηση του $\hat{\beta}_0$ | εκτίμηση του $\hat{\beta}_1$ | εκτίμηση του $\hat{\beta}_2$ | διακύμ. του $\hat{\beta}_0$ | διακύμ. του $\hat{\beta}_1$ | διακύμ. του $\hat{\beta}_2$ | νόρμα μερολ. του $\hat{\boldsymbol{\beta}}$ | μέσο τετραγ. σφάλμα προσαρ. | μέσος χρόνος υπολογ. |
|---|---|---|---|---|---|---|---|---|---|
| LTS | 0.531 | 1.170 | -0.771 | 76.880 | 0.029 | 0.019 | 7.551 | 308 | <1 sec |
| MM | -1.160 | 1.204 | -0.754 | 18.004 | 0.004 | 0.013 | 3.027 | 266 | <1 sec |
| QMIP | -2.760 | 1.242 | -0.727 | 15.790 | 0.002 | 0.010 | 3.696 | 269 | 10 sec |
| S1S | -2.259 | 1.202 | -0.909 | 22.855 | 0.010 | 0.017 | 3.661 | 268 | <1 sec |
| **NQMIP** | **-1.166** | **1.214** | **-0.751** | **14.897** | **0.001** | **0.007** | **2.879** | **263** | **10 sec** |

Σε όλες τις περιπτώσεις πάντως, οι εκτιμήτριες MM και S1S βελτιώνουν την εκτιμήτρια LTS, όπως αναμενόταν. Η νέα εκτιμήτρια δίνει επιπλέον τις πιο κοντινές μέσες εκτιμήσεις στις πραγματικές τιμές των παραμέτρων όπως επίσης έχει και τις μικρότερες διακυμάνσεις και στις τρεις εκτιμώμενες παραμέτρους.

**ΠΙΝΑΚΑΣ 4**

Αρ. των "bad" leverage points 6, "good" leverage points 4, y-outliers 6,
παράμετροι παλινδρόμησης: $\beta_0 = 0.00$, $\beta_1 = 1.20$, $\beta_2 = -0.80$

| Εκτιμήτριες | Αριθμός απόρριψης "good" leverage points |
|---|---|
| QMIP | 113 |
| **NQMIP** | **28** |

Ο Πίνακας 4 δίνει τον αριθμό των "good" leverage points που απορρίφθηκαν με την προηγούμενη και τη νέα προσέγγιση QMIP στις 100 επαναλήψεις, στην περίπτωση της μεγαλύτερης μόλυνσης που πραγματοποιήθηκε στα δεδομένα (32%). Η μείωση του αριθμού των σημείων από 113 σε 28 (επί συνόλου 400) δείχνει το μέγεθος της βελτίωσης της εκτιμήτριας, καθώς τα σημεία αυτά συμβάλλουν στην αποτελεσματικότητα της εκτιμήτριας.

## 5. ΤΕΛΙΚΑ ΣΧΟΛΙΑ - ΠΑΡΑΤΗΡΗΣΕΙΣ

Η προτεινόμενη εκτιμήτρια NQMIP περιορίζει σημαντικά τον αριθμό των απορριπτόμενων "good" leverage σημείων, για αυτό και παρουσιάζει αξιόλογη αποτελεσματικότητα.

Βασιζόμενοι στα παραπάνω κριτήρια και αποτελέσματα, συμπεραίνουμε πως η νέα NQMIP εκτιμήτρια συμπεριφέρεται καλά και είναι αποδοτική σε όλα τα είδη ακραίων παρατηρήσεων, και υπάρχει όφελος από την προτεινόμενη προσέγγιση στην ανθεκτική παλινδρόμηση έναντι των γνωστών μεθόδων. Παρόλα αυτά, ο μέσος χρόνος υπολογισμού της επίλυσης του προβλήματος NQMIP ενός δείγματος μεγέθους n=50 πλησιάζει τα 10 δευτερόλεπτα. Είναι πλέον αποδεκτό ότι στα δείγματα μικρού ή μεσαίου μεγέθους, το όφελος από τη χρήση της προτεινόμενης εκτιμήτριας υπερκαλύπτουν τον επιπρόσθετο χρόνο υπολογισμού.

Η μέθοδος του μικτού ακέραιου τετραγωνικού προγραμματισμού προσφέρει μια συστηματική προσέγγιση στη βελτίωση της συμπεριφοράς μικρού ή μεσαίου μεγέθους δειγμάτων στην ανθεκτική παλινδρόμηση. Αφού ο αριθμός των ακραίων παρατηρήσεων σε ένα δείγμα είναι άγνωστος, προτείνουμε τη χρήση της εκτιμήτριας NQMIP, η οποία παρέχει την ικανότητα διαχωρισμού των "bad" και "good" leverage σημείων, σημαντικό γεγονός στην ανθεκτική παλινδρόμηση.


### ABSTRACT

In robust regression we often have to decide how many are the unusual observations, which should be removed from the sample in order to obtain better fitting for the rest of the observations. Generally, we use the basic principle of LTS, which is to fit the majority of the data, identifying as outliers those points that cause the biggest damage to the robust fit. However, in the LTS regression method the choice of default values for high break down-point affects seriously the efficiency of the estimator. In the proposed approach we introduce penalty cost for discarding an outlier, consequently, the best fit for the majority of the data is obtained by discarding only catastrophic observations. The robust estimation is obtained by solving a convex quadratic mixed integer programming problem. Finally, we conduct a simulation study to compare other robust estimators with our approach in terms of their efficiency and robustness.


### ΑΝΑΦΟΡΕΣ


Arthanari, T. S. and Dodge, Y. (1993), *Mathematical Programming in Statistics*, John Wiley& Sons, Inc.

Coakley, C. W. and Hettmansperger, T. P. (1993), "A Bounded Influence, High Breakdown, Efficient Regression Estimator", *J.A.S.A.*, 88, 872-880.

Hampel, F. R. (1978), "Optimally bounding the gross error sensitivity and influence of position in factor space", *Proceedings of the ASA Statistical Computing Section*, ASA, Wasinghton, D.C.,pp. 59-64.

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

Hawkins, D.M. and Olive, D.J. (1999), "Improved Feasible Solution Algorithms for High Breakdown Estimation", *Computational Statistics and Data Analysis*, 30, 1-11.

Krasker, W. S. and Welsch, R. E. (1982), "Efficient Bounded-Influence Regression Estimation", *J.A.S.A.*, 77, 595-604.

Mallows, C. L. (1975), "On Some Topics in Robustness", *unpublished memorandum*, Bell Telephone Laboratories, Murray Hill, New Jersey.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression", *J.A.S.A.*, 79, 871-880.

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

Rousseeuw, P. J. and Yohai, V.J. (1984), "Robust Regression by Means of S Estimators", *in Robust and Nonlinear Time Series Analyses* (Lecture Notes in Statistics No. 26), eds. J. Franke, W. Hardle, and R. D. Martin, New York: Springer – Verlag pp. 256-272.

Simpson, D.J., Ruppert, D. and Carroll, R.J. (1992), "On One Step GM Estimates and Stability of Inferences in Linear Regression", *J.A.S.A.*, 87, 439-450.

Yohai, V.J. (1987), "High Breakdown-point and High Efficiency Robust Estimates for Regression", *Annals of Statistics* 15, 642-656.

Yohai, V.J. and Zamar, R.Z. (1988), "High Breakdown-point Estimates of Regression by Means of Minimization of an Efficient Scale", *J.A.S.A.*, 83, 406-413.

Zioutas, G. and Avramidis, A. (2005), "Deleting Outliers in Robust Regression with Mixed Integer Programming", *Acta Mathematicae Applicatae Sinica, English Series,* 21, 323-334.