Παράδοξα Αποτελέσματα κατά τη Βέλτιστη Ανάλυση Δεδομένων με τη Μέθοδο των Ελαχίστων Τετραγώνων

Χ. Κωτσάκης

Τμήμα Αγρονόμων και Τοπογράφων Μηχανικών, Πολυτεχνική Σχολή, ΑΠΘ

Περίληψη

Στην εργασία αυτή παρουσιάζονται μέσω συγκεκριμένων απλών παραδειγμάτων ορισμένα "παράδοξα" αποτελέσματα που μπορούν να προκύψουν κατά τη στατιστική επεξεργασία πειραματικών δεδομένων με τη μέθοδο των ελαχίστων τετραγώνων. Συγκεκριμένα, εξετάζονται οι περιπτώσεις όπου ο κεντροβαρικός μέσος όρος ενός δείγματος επαναλαμβανόμενων μετρήσεων του ίδιου μεγέθους είναι μικρότερος από την ελάχιστη τιμή του δείγματος, ή αντίστοιχα μεγαλύτερος από τη μέγιστη τιμή του δείγματος. Σκοπός της εργασίας είναι να διερευνήσει τα ποιοτικά χαρακτηριστικά που πρέπει να έχουν τα διαθέσιμα δεδομένα ώστε να προκύψουν τα προαναφερθέντα αποτελέσματα κατά την εφαρμογή της μεθόδου των ελαχίστων τετραγώνων.

Some Strange Results Obtained from the Least-Squares Analysis of Experimental Data

C. Kotsakis

Department of Rural and Surveying Engineering, School of Engineering, AUTH

Abstract

The aim of this study is to expose some rather peculiar, or even paradoxical, results that may arise from the implementation of the least-squares method for the adjustment of experimental data. In particular, we examine the case where the weighted sample mean of a data set is smaller than the minimum value in the set, or larger than the maximum value in the data set. The main focus in this paper is to study the qualitative statistical characteristics that the given data should have, in order to obtain the aforementioned "strange" results when using the least-squares estimation method.

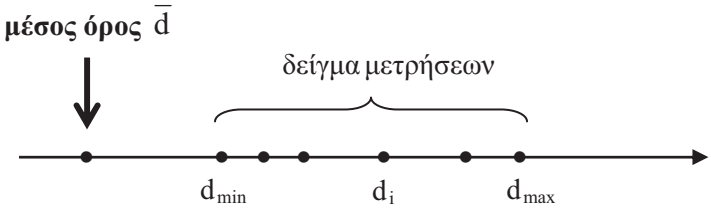
1. Εισαγωγή

Η *μέθοδος των ελαχίστων τετραγώνων* (ΜΕΤ) είναι ένα από τα βασικότερα εργαλεία βέλτιστης επεξεργασίας αριθμητικών δεδομένων για την εξαγωγή ποιοτικής ή/και ποσοτικής πληροφορίας σε σχέση με κάποιο φυσικό σύστημα. Η ΜΕΤ χρησιμοποιείται στα περισσότερα πεδία των εφαρμοσμένων επιστημών που έχουν ως κοινό χαρακτηριστικό τους την ανάγκη αριθμητικού προσδιορισμού ενός συνόλου αγνώστων μεγεθών (παραμέτρων) με βάση τις διαθέσιμες τιμές ενός άλλου συνόλου μεγεθών (μετρήσεων) οι οποίες είναι επηρεασμένες από άγνωστα σφάλματα. Λεπτομέρειες σχετικά με την εφαρμογή και τις δυνατότητες της ΜΕΤ καθώς και για τη σχέση της με τις μεθόδους στατιστικής εκτίμησης γραμμικών μοντέλων στο πλαίσιο της θεωρίας Πιθανοτήτων, μπορούν να βρεθούν σε διάφορα βιβλία της Ελληνικής και ξένης βιβλιογραφίας (βλέπε, για παράδειγμα, Δερμάνης 1987 και τις αναφορές που δίνονται εκεί).

Ένα από τα πιο συνηθισμένα προβλήματα που συχνά αντιμετωπίζεται μέσω της εφαρμογής της ΜΕΤ σε δεδομένα που έχουν προκύψει από μετρήσεις ή άλλες πειραματικές διαδικασίες είναι ο υπολογισμός ενός αντιπροσωπευτικού *μέσου όρου* από τις διαθέσιμες τιμές ενός δείγματος επαναλαμβανόμενων μετρήσεων. Ανεξάρτητα από το εξειδικευμένο αντικείμενο μελέτης κάθε επιστήμονα που επεξεργάζεται αριθμητικά δεδομένα, το παραπάνω πρόβλημα ανακύπτει σε μόνιμη σχεδόν βάση κατά τη στατιστική ανάλυση παρατηρήσεων που προκύπτουν από μετρήσεις φυσικών ή/και γεωμετρικών μεγεθών.

Αξίζει να σημειωθεί ότι ο υπολογισμός "μέσων τιμών" από επαναλαμβανόμενες μετρήσεις αποτελεί ίσως την απλούστερη εφαρμογή συνόρθωσης δεδομένων μέσω της ΜΕΤ. Μία από τις παλαιότερες βιβλιογραφικές αναφορές που υπάρχουν σχετικά με την εφαρμογή της διαδικασίας αυτής είναι η εργασία του Simpson (1755) στην οποία περιγράφεται αναλυτικά η δυνατότητα περιορισμού της επίδρασης των μετρητικών σφαλμάτων σε αστρονομικές παρατηρήσεις μέσω του υπολογισμού αριθμητικών μέσων όρων. Το ίδιο πρόβλημα ήταν μάλιστα αυτό που οδήγησε τον διάσημο μαθηματικό C.F. Gauss στη θεμελίωση μιας αυστηρής *θεωρίας σφαλμάτων* και στη μαθηματική τεκμηρίωση της γνωστής συνάρτησης κατανομής για τη μελέτη τυχαίων μεταβλητών στο πλαίσιο της θεωρίας Πιθανοτήτων (Gauss 1809, Plackett 1972).

Η κοινή λογική που συχνά συνοδεύει τη στατιστική επεξεργασία δεδομένων στις εφαρμοσμένες επιστήμες υπαγορεύει ότι η αντιπροσωπευτική τιμή ενός συνόλου επαναλαμβανόμενων μετρήσεων του ίδιου μεγέθους, θα πρέπει να αναζητηθεί κάπου *ανάμεσα* στις αριθμητικές τιμές των διαθέσιμων δεδομένων. Μια μέση τιμή $\bar{d}$ η οποία είναι μικρότερη από την ελάχιστη τιμή d_{min} του δείγματος των μετρήσεων, ή αντίστοιχα μεγαλύτερη από την μέγιστη τιμή d_{max} του δείγματος, θα πρέπει λογικά να θεωρηθεί ασυνήθιστα περίεργη, αν όχι ως εντελώς λανθασμένη (βλέπε Σχήμα 1).



Σχήμα 1. Μία "μη-συνηθισμένη" τοπολογία για τη θέση του μέσου όρου ενός δείγματος μετρήσεων.

Εντούτοις, οι προηγούμενες περιπτώσεις ενδέχεται να προκύψουν στην πράξη κάτω από συγκεκριμένες (και όχι τόσο σπάνιες) συνθήκες. Η παρούσα εργασία επιχειρεί να δείξει ότι τέτοια σενάρια ανάλυσης δεδομένων δεν είναι στην πραγματικότητα και τόσο παράδοξα ή εξωπραγματικά όσο φαίνονται από μια πρώτη ματιά. Ο βασικός σκοπός της είναι να διερευνήσει, μέσω συγκεκριμένων και απλών παραδειγμάτων, τα ποιοτικά χαρακτηριστικά που πρέπει να έχουν τα διαθέσιμα δεδομένα ώστε κατά την ανάλυσή τους να προκύψουν τα προαναφερθέντα "ασυνήθιστα" αποτελέσματα.

Για καθαρά διδακτικούς λόγους, η εργασία έχει περιοριστεί στην αντιμετώπιση μόνο κάποιων απλών περιπτώσεων όπου ο όγκος των διαθέσιμων δεδομένων είναι ο ελάχιστος δυνατός. Με αυτόν τον τρόπο έχουμε την ευχέρεια να ανιχνεύσουμε και να τεκμηριώσουμε σχετικά εύκολα τις συνθήκες που πρέπει να πληρούν οι παρατηρήσεις ώστε να προκύψουν αποτελέσματα όπως αυτό που φαίνεται στο Σχήμα 1. Η αλγεβρική πολυπλοκότητα που υπεισέρχεται κατά την αλγοριθμική επεξεργασία ενός δείγματος επαναλαμβανόμενων παρατηρήσεων με αυξανόμενο αριθμό τιμών, κάνει εξαιρετικά δύσκολη την *αναλυτική μελέτη* πιο σύνθετων περιπτώσεων. Παρόλα αυτά, η κατανόηση ακόμα και των απλών παραδειγμάτων που εξετάζονται στην εργασία αυτή αποτελεί ένα πολύτιμο εφόδιο, τόσο για τη θεωρητική κατάρτιση όσο και για την πρακτική εξάσκηση που πρέπει να έχουν όσοι ασχολούνται με τη στατιστική επεξεργασία και ανάλυση δεδομένων.

Πρέπει να σημειώσουμε ότι στην εργασία αυτή χρησιμοποιούμε ως μέθοδο βέλτιστης εκτίμησης για τον προσδιορισμό κεντροβαρικών μέσων όρων από τις τιμές ενός δείγματος n επαναλαμβανομένων μετρήσεων (d_1, d_2, …, d_n) το γνωστό γενικευμένο κριτήριο των ελαχίστων τετραγώνων. Η μέθοδος αυτή θεμελιώθηκε από τους Gauss και Legendre πριν από δύο περίπου αιώνες και εξακολουθεί να αποτελεί έως σήμερα ένα από τα σημαντικότερα εργαλεία για την επεξεργασία πειραματικών δεδομένων που προκύπτουν μέσα από μετρητικές διαδικασίες. Η κατά Gauss εκδοχή της ΜΕΤ είναι γνωστή και ως μέθοδος *ανεπηρέαστης γραμμικής εκτίμησης ελάχιστης μεταβλητότητας* (minimum-variance linear unbiased estimation – MVLUE). Για περισσότερες λεπτομέρειες και εφαρμογές, βλέπε Δερμάνης (1987), Δερμάνης και Φωτίου (1992).

2. Υπολογισμός βέλτιστης "μέσης τιμής"

Η χρήση του όρου "*μέση τιμή*" είναι μάλλον μια ατυχής επιλογή που σκοπό έχει να περιγράψει την εκτίμηση μιας άγνωστης παραμέτρου από ένα δείγμα πολλαπλών μετρήσεων της. Ο συγκεκριμένος όρος υπονοεί, κατά κάποιο τρόπο, ότι η βέλτιστη εκτίμηση ενός άγνωστου μεγέθους από επαναλαμβανόμενες παρατηρήσεις θα πρέπει να βρίσκεται κάπου **ανάμεσα** στις τιμές των διαθέσιμων μετρήσεων. Όπως θα διαπιστώσουμε όμως στη συνέχεια αυτής της εργασίας, η παραπάνω αντίληψη μπορεί να μην αντιστοιχεί πάντα στο βέλτιστο αποτέλεσμα που λαμβάνεται από την εφαρμογή της ΜΕΤ σε πειραματικά δεδομένα. Το γεγονός αυτό συμβαίνει επειδή η τιμή της (στατιστικά βέλτιστης) εκτίμησης ενός άγνωστου μεγέθους, σε σχέση με τις επιμέρους τιμές των επαναλαμβανόμενων μετρήσεων που εκτελούμε για τον προσδιορισμό του, επηρεάζεται ποικιλοτρόπως από τη στοχαστική συμπεριφορά (ιδιαίτερα από τις συσχετίσεις) των τυχαίων σφαλμάτων των παρατηρήσεων. Επομένως, ο εμπειρικός όρος "*μέση τιμή*" αντιπροσωπεύει ένα παραπλανητικό σημειολογικό μοντέλο για την περιγραφή ενός απλού, κατά τα άλλα, μαθηματικού/στατιστικού προβλήματος.

Ας θεωρήσουμε την απλή περίπτωση όπου έχουμε διαθέσιμο ένα δείγμα δύο τιμών (d_1, d_2) που αντιστοιχούν σε ξεχωριστές και ανεπηρέαστες μετρήσεις ενός άγνωστου μεγέθους μ. Σε αυτή την περίπτωση, μπορούμε να σχηματίσουμε το ακόλουθο σύστημα των εξισώσεων παρατηρήσεων

$$\mathbf{d} = \mathbf{A}\mu + \mathbf{v} \tag{2.1}$$

ή σε πιο αναλυτική μορφή

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{2.2}$$

Οι ποσότητες v_1 και v_2 εκφράζουν τα τυχαία σφάλματα που εμπεριέχονται στις τιμές των παρατηρήσεων d_1 και d_2, αντίστοιχα. Η στοχαστική/στατιστική συμπεριφορά των τυχαίων σφαλμάτων περιγράφεται μέσω των μηδενικών προσδοκιών τους

$$E\{\mathbf{v}\} = \mathbf{0} \quad \Rightarrow \quad E\{v_1\} = E\{v_2\} = 0 \tag{2.3}$$

και μέσω του πίνακα συμ-μεταβλητοτήτων τους

$$\mathbf{C}_v = E\{\mathbf{v}\mathbf{v}^T\} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{2.4}$$

Το σύμβολο $E\{\cdot\}$ στις εξισώσεις (2.3) και (2.4) υποδηλώνει τον τελεστή της μαθηματικής προσδοκίας (expectation operator), ενώ το σύμβολο ρ εκφράζει τον συντε-

λεστή συσχέτισης (correlation coefficient) μεταξύ των τιμών των τυχαίων σφαλμάτων. Οι ποσότητες σ_1 και σ_2 αντιστοιχούν στις τυπικές αποκλίσεις των τυχαίων σφαλμάτων και παρέχουν τα συνήθη μέτρα ακρίβειας για την ποιοτική περιγραφή των διαθέσιμων δεδομένων.

Με βάση τις παραπάνω μετρήσεις, ο προσδιορισμός της άγνωστης παραμέτρου μ σύμφωνα με το κριτήριο της γραμμικής ανεπηρέαστης εκτίμησης ελάχιστης μεταβλητότητας γίνεται μέσω της γνωστής εξίσωσης

$$\hat{\mu} = (\mathbf{A}^T\mathbf{C}_\mathbf{v}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C}_\mathbf{v}^{-1}\mathbf{d} \tag{2.5}$$

Ισοδύναμα, η παραπάνω σχέση αποτελεί τη λύση του συστήματος κανονικών εξισώσεων (normal equations) στο οποίο καταλήγουμε αν εφαρμόσουμε το γενικευμένο κριτήριο βελτιστοποίησης των ελαχίστων τετραγώνων στο αρχικό σύστημα της εξίσωσης (2.1), δηλ.

$$\mathbf{v}^T\mathbf{P}\mathbf{v} = \text{minimum} \Rightarrow (\mathbf{d}-\mathbf{A}\mu)^T\mathbf{P}(\mathbf{d}-\mathbf{A}\mu) = \text{minimum} \tag{2.6}$$

Σε αυτή την περίπτωση, ο πίνακας βάρους **P** λαμβάνεται ίσος με τον αντίστροφο του πίνακα συμ-μεταβλητοτήτων $\mathbf{C}_\mathbf{v}$ των άγνωστων σφαλμάτων (Δερμάνης, 1987). Αν αντικαταστήσουμε τις αναλυτικές εκφράσεις για τον πίνακα σχεδιασμού **A** και τον πίνακα $\mathbf{C}_\mathbf{v}$ στην εξίσωση (2.5), τότε η βέλτιστη εκτίμηση του μ μπορεί να πάρει την ακόλουθη αναλυτική μορφή

$$\hat{\mu} = \frac{(\sigma_2^2 - \rho\sigma_1\sigma_2)\,d_1 + (\sigma_1^2 - \rho\sigma_1\sigma_2)\,d_2}{(\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2)} \tag{2.7}$$

Όπως φαίνεται από την τελευταία σχέση, η βέλτιστη εκτίμηση της άγνωστης παραμέτρου σύμφωνα με το γενικευμένο κριτήριο των ελαχίστων τετραγώνων ακολουθεί τη γενική *κεντροβαρική μορφή*

$$\hat{\mu} = \frac{\sum_i w_i\, d_i}{\sum_i w_i} \tag{2.8}$$

όπου w_i είναι τα βάρη των αντίστοιχων παρατηρήσεων που συμμετέχουν στον προσδιορισμό του μ.

Πριν προχωρήσουμε στη διερεύνηση των προυποθέσεων κάτω από τις οποίες η εξίσωση (2.7) μπορεί να οδηγήσει στα "παράδοξα" αποτελέσματα $\hat{\mu} > max(d_1, d_2)$ ή $\hat{\mu} < min(d_1, d_2)$, αξίζει να σημειώσουμε τις ακόλουθες **ειδικές περιπτώσεις**.

- *Οι διαθέσιμες δειγματικές μετρήσεις του άγνωστου μεγέθους είναι ίσες μεταξύ τους* ($d_1 = d_2$). Σε αυτή την περίπτωση η βέλτιστη εκτίμηση του μ θα είναι πάντα $\hat{\mu} = d_1 = d_2$, **ανεξάρτητα** από τις τιμές των σ_1, σ_2 και ρ.
- *Οι ακρίβειες των μετρήσεων για την άγνωστη παράμετρο μ είναι ίσες μεταξύ τους* ($\sigma_1 = \sigma_2$). Σε αυτή την περίπτωση η βέλτιστη εκτίμηση θα είναι πάντα ίση με τον αριθμητικό μέσο όρο των δεδομένων, $\hat{\mu} = (d_1 + d_2)/2$, **ανεξάρτητα** από την τιμή του συντελεστή συσχέτισης ρ των τυχαίων σφαλμάτων των μετρήσεων.
- *Ο συντελεστής συσχέτισης μεταξύ των παρατηρήσεων είναι μηδέν* ($\rho = 0$). Σε αυτή την περίπτωση είναι γενικά γνωστό ότι τα βέλτιστα βάρη $\{w_i\}$ που προκύπτουν από την εφαρμογή της ΜΕΤ είναι ίσα με τις αντίστροφες μεταβλητότητες των αντίστοιχων παρατηρήσεων $\{d_i\}$. Από την εξίσωση (2.7) όταν $\rho = 0$, προκύπτει ότι το βάρος w_1 της πρώτης παρατήρησης είναι ίσο με τη μεταβλητότητα της δεύτερης παρατήρησης, ενώ το βάρος w_2 της δεύτερης παρατήρησης είναι ίσο με τη μεταβλητότητα της πρώτης παρατήρησης! Όσο περίεργο και αν δείχνει αυτό το αποτέλεσμα, δεν είναι λανθασμένο μιας και εύκολα αποδεικνύεται ότι

$$\hat{\mu} = \frac{(\sigma_2^2)\, d_1 + (\sigma_1^2)\, d_2}{(\sigma_1^2) + (\sigma_2^2)} = \frac{(1/\sigma_1^2)\, d_1 + (1/\sigma_2^2)\, d_2}{(1/\sigma_1^2) + (1/\sigma_2^2)} \tag{2.9}$$

Οι παραπάνω ειδικές περιπτώσεις μπορούν εύκολα να επεκταθούν και σε εφαρμογές όπου ο αριθμός των διαθέσιμων επαναλαμβανόμενων παρατηρήσεων είναι μεγαλύτερος από δύο.

3. Διερεύνηση της γενικής περίπτωσης

Ας εξετάσουμε τώρα τη γενικότερη περίπτωση όπου $d_1 \neq d_2$, $\sigma_1 \neq \sigma_2$ και $\rho \neq 0$. Η βασική εξίσωση (2.7) της προηγούμενης ενότητας μπορεί να πάρει την εναλλακτική μορφή

$$\hat{\mu} = d_1 + (q-1)(d_1 - d_2) \tag{3.1}$$

ή ισοδύναμα

$$\hat{\mu} = d_2 + q\,(d_1 - d_2) \tag{3.2}$$

όπου ο βοηθητικός συντελεστής q δίνεται από τη σχέση

$$q = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2} \tag{3.3}$$

Με βάση τις σχέσεις (3.1) ή (3.2), μπορούμε εύκολα να διακρίνουμε τις συνθήκες κάτω από τις οποίες ο κεντροβαρικός μέσος όρος $\hat{\mu}$ είναι δυνατόν να βρεθεί έξω από το αριθμητικό διάστημα που καταλαμβάνουν τα διαθέσιμα δεδομένα. Απαραίτητη προϋπόθεση για να συμβεί αυτό είναι να έχουν διαφορετικές τιμές οι μετρήσεις d_1 και d_2 (στην αντίθετη περίπτωση η βέλτιστη εκτίμηση $\hat{\mu}$ ταυτίζεται με τις ίσες τιμές των παρατηρήσεων).

Όταν λοιπόν ισχύει $d_1 \neq d_2$ τότε η δυνατότητα να προκύψουν τα παράδοξα αποτελέσματα $\hat{\mu} > max(d_1, d_2)$ ή $\hat{\mu} < min(d_1, d_2)$ εξαρτάται αποκλειστικά από την αριθμητική τιμή του συντελεστή q. Πιο συγκεκριμένα, εάν η τιμή του q είναι μεγαλύτερη της μονάδας ή μικρότερη του μηδέν, τότε θα έχουμε οπωσδήποτε ένα από τα δύο παράδοξα αποτελέσματα για τον κεντροβαρικό μέσο όρο, όπως μπορεί εύκολα να αποδειχθεί χρησιμοποιώντας τις εξισώσεις (3.1) και (3.2). Στον Πίνακα 1 δίνονται όλοι οι δυνατοί συνδυασμοί που μπορούν να οδηγήσουν σε αυτά τα "παράδοξα" αποτελέσματα για την εκτίμηση της άγνωστης παραμέτρου μ μέσω της συνόρθωσης των επιμέρους μετρήσεων της.

Πίνακας 1. *Οι διάφοροι συνδυασμοί που οδηγούν στο "παράδοξο" αποτέλεσμα όπου ο κεντροβαρικός μέσος όρος βρίσκεται έξω από το διάστημα που ορίζεται από τα διαθέσιμα δεδομένα d_1 και d_2.*

	$d_1 < d_2$	$d_1 > d_2$
$q > 1$	$\hat{\mu} < d_1$	$\hat{\mu} > d_1$
$q < 0$	$\hat{\mu} > d_2$	$\hat{\mu} < d_2$

Τι σημαίνουν όμως πρακτικά οι συνθήκες $q > 1$ ή $q < 0$ σε σχέση με τις τιμές που μπορούν να πάρουν τα μεγέθη σ_1, σ_2 και ρ; Τι είδους στατιστική συμπεριφορά, με άλλα λόγια, πρέπει να ακολουθούν τα διαθέσιμα δεδομένα προκειμένου να ικανοποιείται κάποια από αυτές τις δύο ανισότητες για τον συντελεστή q;

Χρησιμοποιώντας την εξίσωση (3.3) αποδεικνύεται εύκολα ότι, αν ισχύει η σχέση

$$\rho > min\left(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1}\right) \tag{3.4}$$

τότε άμεσα θα έχουμε είτε $q > 1$, είτε $q < 0$ (ανάλογα με το αν ισχύει $\sigma_1 < \sigma_2$ ή $\sigma_1 > \sigma_2$, αντίστοιχα). Με απλά λόγια λοιπόν, όταν οι μετρήσεις έχουν διαφορετικές μεταβλητότητες και αρκετά δυνατή *θετική* συσχέτιση στα τυχαία σφάλματα τους, τότε ο κεντροβαρικός μέσος όρος τους (σύμφωνα με το γενικευμένο κριτήριο των

ελαχίστων τετραγώνων) είναι δυνατό να είναι είτε μεγαλύτερος, είτε μικρότερος, από *όλες* τις διαθέσιμες τιμές των μετρήσεων.

Σε αυτό το σημείο αξίζει να σημειωθεί ότι η συνθήκη (3.4) δεν αντιβαίνει τη γνωστή ανισότητα που διέπει, εξ' ορισμού, τον συντελεστή συσχέτισης μεταξύ δύο τυχαίων μεταβλητών, δηλαδή ότι $-1 \le \rho \le 1$. Πράγματι, η μικρότερη τιμή ανάμεσα σε δύο θετικούς λόγους της μορφής α/β και β/α είναι πάντα μικρότερη της μονάδας.

Παρά το γεγονός ότι τα αποτελέσματα $\hat{\mu} > max(d_1, d_2)$ ή $\hat{\mu} < min(d_1, d_2)$ είναι σχετικά ασυνήθιστα στην πράξη, δεν παύουν να είναι θεωρητικώς σωστά και στατιστικώς βέλτιστα (κάτω από τις συνθήκες που προαναφέρθηκαν) σύμφωνα με το γενικευμένο κριτήριο των ελαχίστων τετραγώνων. Προκειμένου να μπορέσουμε να αποδεχτούμε τέτοια αποτελέσματα και με την καθημερινή εμπειρική λογική, θα ήταν ίσως χρήσιμο να αντιμετωπίσουμε το ακόλουθο απλό παράδειγμα.

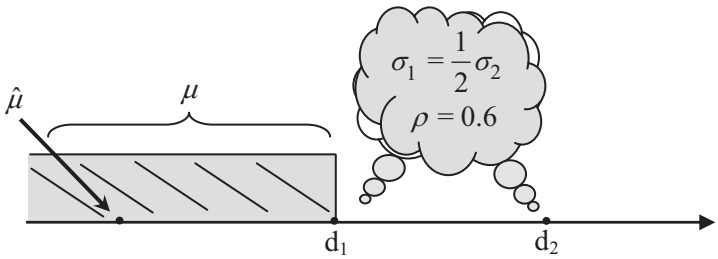
Παράδειγμα

Ας θεωρήσουμε ότι $\sigma_2 = 2\sigma_1$ και $\rho = 0.6$. Αν το διαθέσιμο δείγμα των μετρήσεων για τον υπολογισμό της άγνωστης παραμέτρου μ είναι τέτοιο ώστε $d_1 < d_2$, τότε το αποτέλεσμα που θα υπολογιστεί από τη βασική σχέση (2.7) θα ικανοποιεί την ανισότητα $\hat{\mu} < min(d_1, d_2)$. Αυτό συμβαίνει αφού η τιμή του συντελεστή συσχέτισης ικανοποιεί την αναγκαία συνθήκη (3.4), δηλαδή

$$\rho > min\left(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1}\right) = 0.5$$

Ας προσπαθήσουμε τώρα να εκλογικεύσουμε αυτό το αποτέλεσμα με βάση τα στοχαστικά χαρακτηριστικά των σφαλμάτων των διαθέσιμων μετρήσεων. *Εφόσον τα σφάλματα έχουν αρκετά δυνατή θετική συσχέτιση, είναι λογικό να υποθέσουμε ότι και οι δύο παρατηρήσεις θα βρίσκονται στην ίδια (αριθμητικά) μεριά του* μ. Επίσης, λόγω του γεγονότος ότι η πρώτη παρατήρηση έχει μικρότερη μεταβλητότητα (δηλαδή μεγαλύτερη ακρίβεια) από τη δεύτερη παρατήρηση, η πιθανότητα να βρίσκεται η τιμή d_1 πιο κοντά στην άγνωστη τιμή του μ είναι μεγαλύτερη από την αντίστοιχη πιθανότητα της τιμής d_2. Με αυτόν τον απλό περιγραφικό τρόπο, καταλήγουμε τελικά στο συμπέρασμα ότι η πιο πιθανή (ή λογικοφανής) εκτίμηση για την άγνωστη παράμετρο θα πρέπει να αναζητηθεί κάπου "στα αριστερά" της πρώτης παρατήρησης d_1 (εφόσων βέβαια δίνεται εξαρχής ότι $d_1 < d_2$). Μία εικονογραφική απεικόνιση της παραπάνω συλλογιστικής διαδικασίας δίνεται στο Σχήμα 2.

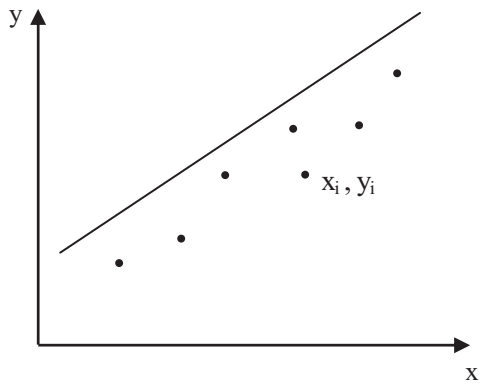
Βλέπουμε λοιπόν ότι το αποτέλεσμα στο οποίο μας οδηγεί η μέθοδος των ελαχίστων τετραγώνων στην περίπτωση του συγκεκριμένου παραδείγματος, δηλαδή ότι $\hat{\mu} < min(d_1, d_2)$, μπορεί πράγματι να θεωρηθεί λογικοφανές παρά την αρχική του παραδοξότητα.



Σχήμα 2. Υπολογισμός κεντροβαρικού μέσου όρου από δύο θετικά συσχετισμένες παρατηρήσεις. Το γραμμοσκιασμένο διάστημα δείχνει την περιοχή όπου είναι πιο πιθανό να βρίσκεται η αληθινή τιμή μ, όταν ξέρουμε ότι $d_1 < d_2$, $\sigma_1 < \sigma_2$ και $\rho > (\sigma_1 / \sigma_2)$.

4. Επίλογος

Στην εργασία αυτή παρουσιάστηκε μια διδακτικού τύπου προσέγγιση σε ορισμένα παράδοξα αποτελέσματα που μπορούν να προκύψουν κατά τη συνόρθωση παρατηρήσεων με τη μέθοδο των ελαχίστων τετραγώνων. Η μελέτη μας περιορίστηκε σε ένα σχετικά απλό πρόβλημα προσδιορισμού του κεντροβαρικού μέσου όρου από ένα δείγμα επαναλαμβανόμενων μετρήσεων.



Σχήμα 3. Μία "ασυνήθιστη" βέλτιστη γραμμή παλινδρόμησης μέσω ΜΕΤ.

Αξίζει να σημειωθεί ότι ανάλογα παράδοξα αποτελέσματα μπορούν να ληφθούν και στο γνωστό πρόβλημα προσδιορισμού μιας βέλτιστης ευθείας $y = ax + b$ από δεδομένα ζεύγη μετρήσεων (x_i, y_i). Όπως και στην περίπτωση προσδιορισμού "μέσων

τιμών", το πρόβλημα αυτό είναι από τα πλέον συνηθισμένα στη στατιστική επεξεργασία δεδομένων που έχουν προκύψει από παρατηρήσεις ή άλλες πειραματικές διαδικασίες. Ανάλογα με το στοχαστικό μοντέλο περιγραφής των τυχαίων σφαλμάτων στις παρατηρήσεις των συντεταγμένων $\{x_i\}$ ή/και $\{y_i\}$, η εφαρμογή του γενικευμένου κριτηρίου των ελαχίστων τετραγώνων είναι δυνατό να δώσει ως βέλτιστη λύση μια "ασυνήθιστη" ευθεία που να έχει τη μορφή που φαίνεται στο Σχήμα 3.

Είναι γεγονός ότι τα αποτελέσματα που περιγράφηκαν στην παρούσα εργασία είναι αρκετά σπάνια σε πρακτικές εφαρμογές. Παρόλα αυτά, όπως ήδη εξηγήσαμε στις προηγούμενες ενότητες, οι περιπτώσεις που εμφανίζονται στα Σχήματα 1 και 3 είναι αναμενόμενες όταν οι διαθέσιμες παρατηρήσεις έχουν διαφορετικές ακρίβειες και αρκετά δυνατή θετική συσχέτιση στα τυχαία σφάλματα τους. Βέβαια στις περισσότερες εφαρμογές συνόρθωσης δεδομένων με τη μέθοδο των ελαχίστων τετραγώνων, οι τυχόν συσχετίσεις των παρατηρήσεων συχνά αγνοούνται και ο πίνακας συμμεταβλητοτήτων τους λαμβάνεται και χρησιμοποιείται σε απλή διαγώνια μορφή. Η πρακτική αυτή, που οφείλεται κυρίως στην έλλειψη επαρκούς στατιστικής πληροφορίας για τον ακριβή προσδιορισμό των πραγματικών συσχετίσεων των τυχαίων σφαλμάτων, καθιστά αδύνατη την εμφάνιση των παράδοξων αποτελεσμάτων που προαναφέρθηκαν. Εντούτοις πρέπει να έχουμε υπόψη μας ότι η υπόθεση των ασυσχέτιστων παρατηρήσεων είναι μάλλον η εξαίρεση, παρά ο κανόνας, στην προσπάθεια μας να μοντελοποιήσουμε τη διαδικασία συλλογής και καταγραφής δεδομένων από κάποιο φυσικό σύστημα. Κατά αυτόν τον τρόπο, το περιεχόμενο της εργασίας εκτός από το γενικότερο θεωρητικό του ενδιαφέρον, έχει σίγουρα και κάποια πρακτική αξία για τις περιπτώσεις όπου τα διαθέσιμα δεδομένα ακολουθούν τη στατιστική συμπεριφορά που υποδείχθηκε στην Ενότητα 3 της παρούσας μελέτης.

Βιβλιογραφία

1. Δερμάνης, Α., 1987. *Συνορθώσεις παρατηρήσεων και θεωρία εκτίμησης* (τόμος 1 και 2), Εκδόσεις Ζήτη, Θεσσαλονίκη.
2. Δερμάνης, Α., Φωτίου Α., 1992. *Μέθοδοι και εφαρμογές συνόρθωσης παρατηρήσεων*, Εκδόσεις Ζήτη, Θεσσαλονίκη.
3. Gauss, C.F., 1809. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Perthes and Besser, Hamburg. English translation (originally in 1857 by C.H. Davis) reprinted as *Theory of the Motions of the Heavenly Bodies Moving about the Sun in Conic Sections*, Dover, New York, 1963.
4. Plackett, R.L., 1972. *The discovery of the method of least squares*. Biometrika, 59(2): 239-251.
5. Simpson, T., 1755. On the advantage of taking the mean of a number of observations in practical astronomy. Philos. Trans. Royal Soc., London, 46: 82-102.