Γραμμική προσαρμογή σε συνθήκες αβεβαιότητας: Μια συγκριτική ανάλυση με εφαρμογή στη μέθοδο των ελαχίστων τετραγώνων

Ε. Μαθιουλάκης, Γ. Πανάρας και Β. Μπελεσιώτης

Εργαστήριο Ηλιακών & άλλων Ενεργειακών Συστημάτων - ΕΚΕΦΕ «ΔΗΜΟΚΡΙΤΟΣ»

15310 Αγ. Παρασκευή Αττικής, e-mail: math@ipta.demokritos.gr

Περίληψη

Η εργασία αυτή έχει ως στόχο την κριτική αξιολόγηση των κυριότερων διαφορετικών προσεγγίσεων που έχουν κατά καιρούς εφαρμοστεί για την προσαρμογή γραμμικού μοντέλου σε εμπειρικά δεδομένα με τη μέθοδο των ελαχίστων τετραγώνων. Εξετάζεται ειδικότερα η συχνά εμφανιζόμενη περίπτωση, στα πλαίσια των συνήθων μετρολογικών εφαρμογών, της συσχέτισης δεδομένων από εμπειρικές παρατηρήσεις με χρήση γραμμικού μοντέλου σε συνθήκες αβεβαιότητας.

Αναλύεται η φιλοσοφία της κάθε προσέγγισης, εστιάζοντας στις κάθε φορά επικαλούμενες υποθέσεις εργασίας και στις συνεπαγόμενες μεθοδολογικές επιλογές για την επίλυση του προβλήματος της συσχέτισης σε πρακτικό επίπεδο. Έμφαση δίδεται στον προσανατολισμό του συστήματος αναφοράς, δηλαδή στα κριτήρια καθορισμού των εξαρτημένων και των ανεξάρτητων μεταβλητών, καθώς και στις επιπτώσεις από την αποδοχή περιοριστικών υποθέσεων σχετικά με μεταβλητότητα των σφαλμάτων σε όλο το εύρος της συσχέτισης.

Επιχειρείται τέλος η σύγκριση στην πράξη διαφόρων μεθοδολογικών προσεγγίσεων, μέσω της εφαρμογής τους σε συγκεκριμένο παράδειγμα και αξιολογούνται οι διαφορές που προκύπτουν. Στα πλαίσια της συγκριτικής αυτής αξιολόγησης διερευνώνται, ως προς την αξιοπιστία και την αποτελεσματικότητα τους, οι συνηθέστερα χρησιμοποιούμενες προσεγγίσεις, όπως αυτές των Σταθμισμένων Ελαχίστων Τετραγώνων (Weighted Least Squares), των Ολικών Ελαχίστων Τετραγώνων (Total Least Squares), των Ορθογώνιων Ελαχίστων Τετραγώνων (Orthogonal Least Squares), καθώς και προσεγγίσεις που βασίζονται στην διάχυση κατανομών πιθανοτήτων με χρήση προσομοίωσης Monte-Carlo.

***Λέξεις Κλειδιά**: Αβεβαιότητα, γραμμικό μοντέλο, ελάχιστα τετράγωνα, Monte-Carlo*

1. Εισαγωγή

Η αποκατάσταση συσχετίσεων ανάμεσα σε διαφορετικές ομάδες ποσοτικών δεδομένων με τη βοήθεια κατάλληλων μοντέλων, συνιστά εξαιρετικά διαδεδομένη δραστηριότητα στα πλαίσια της επικύρωσης ερμηνευτικών σχημάτων που σκοπεύουν στην οργάνωση και επεξεργασία της πληροφορίας που προέρχεται από εμπειρικές παρατηρήσεις. Η χρήση μάλιστα και η πρακτική αξιοποίηση της εύρεσης συσχετίσεων μεταξύ διαφορετικών ομάδων δεδομένων, δεν περιορίζεται στο σύνολο σχεδόν των αμιγώς επιστημονικών δραστηριοτήτων, αλλά επεκτείνεται, συχνά ασυνείδητα, και σε όλους τους τομείς της ανθρώπινης δραστηριότητας, σε μια διαρκή προσπάθεια ανάδειξης υπαρκτών ή υποθετικών σχέσεων μεταξύ αιτίων και αποτελεσμάτων.

Ο σκοπός της επιχειρούμενης συσχέτισης, αν και εξαρτάται από τις ιδιαιτερότητες της κάθε συγκεκριμένης εφαρμογής, συνίσταται συνήθως είτε στην επιβεβαίωση μιας υπόθεσης εργασίας βασιζόμενης σε υφιστάμενη γνώση, είτε στη συγκεκριμενοποίηση μιας θεωρητικά αποδεκτής συσχέτισης για την οποία όμως δεν είναι ακόμα διαθέσιμα συγκεκριμένα ποσοτικά στοιχεία

(π.χ. οι συντελεστές συσχέτισης), είτε τέλος στην απόπειρα εντοπισμού μια άγνωστης έως τώρα συσχέτισης στα πλαίσια της ανάπτυξης ή της επιβεβαίωσης ενός ερμηνευτικού σχήματος για το οποίο δεν είναι ακόμα διαθέσιμη επαρκής θεωρητική τεκμηρίωση. Σε πρακτικό επίπεδο, σκοπός είναι ο προσδιορισμός των τιμών εκείνων των συντελεστών ενός μοντέλου, με τρόπο που το μοντέλο αυτό να εξηγεί κατά το δυνατόν πιστότερα τα διαθέσιμα δεδομένα. Είναι επίσης σημαντικό να τονιστεί ότι βασική υπόθεση εργασίας αποτελεί η αποδοχή της ισχύος του μοντέλου στην εκάστοτε εξεταζόμενη περίπτωση, παρά το ότι είναι εκ των προτέρων γνωστή η γενική μόνο μορφή του.

Η συνήθως υιοθετούμενη προσέγγιση είναι αυτή της παλινδρόμησης ελαχίστων τετραγώνων (least square regression). Η εφαρμογή της συνιστά σημαντική δραστηριότητα και στον τομέα της μετρολογίας, με χαρακτηριστικότερα παραδείγματα:

- τη διακρίβωση, όπου το ζητούμενο είναι ο προσδιορισμός μιας καμπύλης διακρίβωσης η οποία συσχετίζει τις ενδείξεις ενός οργάνου μέτρησης με τις αντίστοιχες ενδείξεις αναφοράς που παρέχονται, για παράδειγμα, από μια διάταξη μέτρησης υψηλότερης μετρολογικής στάθμης ή από ένα υλικό αναφοράς,
- τη δοκιμή ενός υλικού ή μιας διάταξης, για το οποίο είναι διαθέσιμο ένα γενικό θεωρητικό μοντέλο συμπεριφοράς, και επιδιώκεται η εύρεση των τιμών εκείνων των χαρακτηριστικών συντελεστών του μοντέλου αυτού, οι οποίες αντιστοιχούν στο συγκεκριμένο προϊόν.

Μια ιδιαιτερότητα της αξιοποίησης της συσχέτισης ποσοτικών δεδομένων στα πλαίσια μετρολογικών εφαρμογών, σχετίζεται με την υποχρέωση να λαμβάνεται υπόψη η αβεβαιότητα που χαρακτηρίζει τα κάθε είδους ποσοτικά στοιχεία, είτε αυτά αφορούν τα συσχετιζόμενα δεδομένα, είτε αφορούν την ορθότητα της συσχέτισης και, κατά συνέπεια, την ποιότητα των δεδομένων που θα προκύψουν όταν η συσχέτιση αυτή χρησιμοποιηθεί σε ένα επόμενο στάδιο. Δεδομένου μάλιστα ότι οι σύγχρονες αντιλήψεις στον τομέα της μετρολογίας δεν επιτρέπουν να αγνοείται το ότι κάθε ποσοτικό δεδομένο χαρακτηρίζεται από μικρότερες ή μεγαλύτερες αβεβαιότητες, η υιοθέτηση μεθοδολογικών προσεγγίσεων που να παίρνουν υπόψη την ύπαρξη των αβεβαιοτήτων αυτών είναι επιβεβλημένη.

Από την άποψη αυτή έχει ιδιαίτερο ενδιαφέρον να αξιολογηθεί η αποτελεσματικότητα των διαθέσιμων προσεγγίσεων συσχέτισης με τη μέθοδο των ελαχίστων τετραγώνων, ως προς την καταλληλότητά τους να χειριστούν αβέβαια δεδομένα. Δεν είναι άλλωστε τυχαίο ότι ένα από τα πρώτα συμπληρώματα του 'Guide to the expression of uncertainty in measurement', εν συντομία GUM, η έκδοση του οποίου προβλέπεται για το άμεσο μέλλον, αφορά τις εφαρμογές της προσαρμογής με τη μέθοδο των ελαχίστων τετραγώνων στον τομέα της μετρολογίας (ISO 1995, Bich et al. 2006).

Στα πλαίσια της παρούσας εργασίας εξετάζονται, τόσο η συμβατική προσέγγιση των κοινών ελαχίστων τετραγώνων, όσο και πλέον σύνθετες προσεγγίσεις (μέθοδοι Σταθμισμένων Ελαχίστων Τετραγώνων, Ολικών Ελαχίστων Τετραγώνων, Ορθογώνιων Ελαχίστων Τετραγώνων, προσομοίωσης Monte-Carlo) και προτείνονται ορισμένες επιλογές βελτιστοποίησης της χρήσης τους στις πράξη. Για λόγους που σχετίζονται με την απλούστερη παρουσίαση των διαφόρων πλευρών του προβλήματος, επιλέγεται η περίπτωση που το προσαρμοζόμενο μοντέλο είναι μια ευθεία, τονίζοντας όμως ότι τα ίδια συμπεράσματα έχουν ισχύ και στη γενικότερη περίπτωση αναζήτησης των σταθερών συντελεστών ενός γραμμικού μοντέλου το οποίο συναρτά μια εξαρτημένη μεταβλητή από μία ή περισσότερες ανεξάρτητες μεταβλητές.

3. Η μέθοδος των ελαχίστων τετραγώνων

Το γενικό πρόβλημα της προσαρμογής ενός γραμμικού μοντέλου σε εμπειρικές παρατηρήσεις, συνίσταται στον προσδιορισμό των εκτιμώμενων τιμών z_k, $k=1,..M$, των συντελεστών Z ενός

μοντέλου, με τρόπο που το μοντέλο αυτό να περιγράφει με τον καλύτερο δυνατό τρόπο N ανεξάρτητες παρατηρήσεις $(\mathbf{x}^{\mathbf{T}}, \mathbf{y}^{\mathbf{T}})=(x_i, y_i)$, $i=1,\ldots,N$ των μεγεθών X και Y αντίστοιχα:

$$Y = F(X) = Z_1 f_1(X) + Z_2 f_2(X) + \ldots + Z_M f_M(X) \qquad (1)$$

όπου $f_k(X)$, $k=1,\ldots,M$ μπορούν να είναι οποιαδήποτε είδους γνωστές συναρτήσεις του X. Σημειώνεται ότι η γραμμικότητα του μοντέλου δεν αφορά τις συναρτήσεις $f_k(X)$, οι οποίες μπορούν να είναι και μη γραμμικές, αλλά εντοπίζεται στη γραμμικότητα ως προς τους συντελεστές Z.

Αυτό που πρέπει καταρχήν να γίνει κατανοητό, είναι ότι η επιλογή της γραμμικής συσχέτισης εμπεριέχει ως βασική υπόθεση εργασίας την αποδοχή ότι η συσχέτιση αυτή συνιστά κατάλληλο μοντέλο, όσον αφορά τη δυνατότητά του να περιγράψει την πραγματικότητα που υπάρχει πίσω από τα πειραματικά δεδομένα.

Η συμβατότητα με τις σύγχρονες αντιλήψεις για τη μέτρηση και την αβεβαιότητα, όπως αυτές εκφράζονται από τον GUM, επιβάλει τη θεώρηση όχι μόνο των παρατηρήσεων των μεγεθών X και Y, αλλά και των αβεβαιοτήτων που χαρακτηρίζουν τις τιμές αυτές. Επιπλέον, ενδιαφέρει η εκτίμηση όχι μόνο των πιθανότερων τιμών z των συντελεστών Z, αλλά και της μεταβλητότητας (αβεβαιότητας) και συμμεταβλητότητας $\boldsymbol{u}_z$ που χαρακτηρίζει τις τιμές αυτές. Προκειμένου ειδικότερα για μετρολογικές εφαρμογές, η πληροφορία αυτή είναι απολύτως αναγκαία για την ανάλυση αβεβαιοτήτων σε υπολογισμούς που εμπλέκουν τα αποτελέσματα που θα προκύψουν από τη μελλοντική χρήση του μοντέλου, μετά την προσαρμογή του.

Η μέθοδος των ελαχίστων τετραγώνων επιχειρεί εν γένει να δώσει απάντηση στο ερώτημα: *με δεδομένο ένα σετ τιμών για τους συντελεστές Z, ποια είναι η πιθανότητα το σετ αυτό να είναι το πλέον κατάλληλο*; Εάν $\mathbf{v}_e=(\mathbf{x},\mathbf{y})^T$ είναι το διάνυσμα μήκους $2N$ που σχηματίζεται από την συνένωση των διανυσμάτων $\mathbf{x}$ και $\mathbf{y}$ των πειραματικών παρατηρήσεων των μεγεθών X και Y αντίστοιχα, και $\mathbf{u}_{\mathbf{V}}^2$ ο αντίστοιχος πίνακας αβεβαιοτήτων, η μέθοδος των ελαχίστων τετραγώνων αποσκοπεί στην εύρεση των τιμών των X και Y που ικανοποιούν τη σχέση (1) και αποκλίνουν το λιγότερο δυνατόν από τις πειραματικές τιμές $\mathbf{v}_e$. Η μεγιστοποίηση της πιθανότητας να είναι το καταλληλότερο σετ συντελεστών μεταφράζεται πρακτικά στην ελαχιστοποίηση του σταθμισμένου γινόμενου (Lira 2000):

$$\chi^2 = \left(\mathbf{v} - \mathbf{v}_\mathbf{e}\right)^T \left(\mathbf{u}_\mathbf{V}^2\right)^{-1} \left(\mathbf{v} - \mathbf{v}_\mathbf{e}\right) \qquad (2)$$

Στη συνέχεια εξετάζονται και αξιολογούνται οι κυριότερες προσεγγίσεις που έχουν αναπτυχθεί κατά καιρούς για την επίλυση του προβλήματος αυτού στην πράξη, εφαρμοζόμενες, για λόγους απλοποίησης της παρουσίασης, στην περίπτωση της προσαρμογής του απλού γραμμικού μοντέλου:

$$y=a\,x+b \qquad (3)$$

Πρέπει βέβαια να τονιστεί και πάλι ότι τα συμπεράσματα που προκύπτουν από τη διερεύνηση της απλής αυτής περίπτωσης έχουν εφαρμογή, χωρίς ιδιαίτερες τροποποιήσεις, και στην γενικότερη περίπτωση του μοντέλου της σχέσης (1), καθώς και στις διάφορες παραλλαγές του.

3. Η μέθοδος των ελαχίστων τετραγώνων σε συνθήκες γενικευμένης αβεβαιότητας

3.1. Κοινά ελάχιστα τετράγωνα και επιλογή του συστήματος αναφοράς

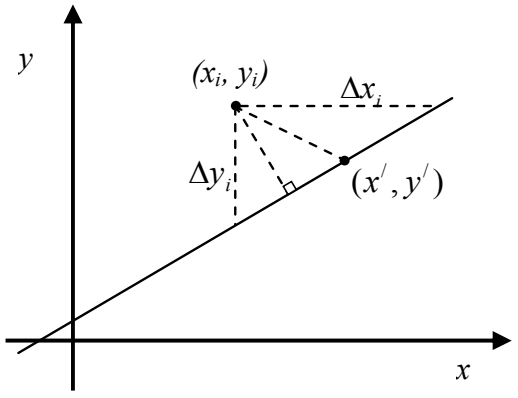
Στις περιπτώσεις που μία μεταβλητή X βρίσκεται υπό απόλυτο έλεγχο σε ένα στατιστικά ελεγχόμενο πείραμα, οι τιμές που παίρνει η μεταβλητή αυτή μπορούν εύλογα να θεωρηθούν ως απαλλαγμένες από σφάλματα. Στις περιπτώσεις αυτές θεωρείται ότι η ποσότητα X είναι επακριβώς γνωστή και ότι οι αποκλίσεις των πειραματικών σημείων από την ευθεία του μοντέλου οφείλονται αποκλειστικά σε σφάλματα στις τιμές της εξαρτημένης μεταβλητής. Υπό την έννοια αυτή, η καταλληλότερη διαμόρφωση του μοντέλου, είναι αυτή για την οποία ελαχιστοποιείται το άθροισμα των τετράγώνων των αποκλίσεων $\Delta y_i=y_i-F(x_i)$ μεταξύ της πειραματικής τιμής y_i και της τιμής $F(x_i)$ που υπολογίζεται από το μοντέλο. Η σχέση (2) απλοποιείται τότε και γίνεται:

$$\chi^2 = \left(\mathbf{y}-\mathbf{y_e}\right)^T \left(\mathbf{u}_y^2\right)^{-1} \left(\mathbf{y}-\mathbf{y_e}\right) \tag{4}$$

Η γενικευμένη θεώρηση της προσέγγισης αυτής, αποκαλούμενης και προσέγγισης των κοινών ελαχίστων τετραγώνων (*Ordinary Least Square* ή *OLS)*, οφείλεται σε μεγάλο βαθμό στην απλότητα που χαρακτηρίζει την εφαρμογή της στην πράξη. Η τεχνική όμως αυτή βασίζεται σε ορισμένες υποθέσεις, η ισχύς των οποίων δεν είναι εύκολο να γίνει πάντα αποδεκτή, ειδικότερα στο ευρύτερο πλαίσιο των μετρολογικών εφαρμογών, και οι οποίες αφορούν την απουσία σφαλμάτων στις τιμές του X και στην ύπαρξη ίδιας μεταβλητότητας για όλες τις παρατηρήσεις του Y (homoscedasticity).

Σχετικό είναι και το ζήτημα της επιλογής του συστήματος αναφοράς, το οποίο δεν στερείται πρακτικής σημασίας στο βαθμό που οδηγεί σε διαφορετικά αποτελέσματα. Πράγματι, η επιλογή του Y ως εξαρτημένη μεταβλητή οδηγεί στις τιμές $\hat{a}$ και $\hat{b}$ ως καλύτερες εκτιμήσεις των συντελεστών a και b της εξίσωσης (3), μέσω της ελαχιστοποίησης του αθροίσματος των τετραγώνων των κατακόρυφων αποστάσεων (Σχήμα 1):

$$\chi^2 = \sum_{i=1}^{N} \frac{\Delta y_i^2}{u_{y,i}^2} = \sum_{i=1}^{N} \left(\frac{y_i - \left(ax_i + b\right)}{u_{y,i}} \right)^2 \tag{5}$$



Σχήμα 1: Αποκλίσεις μοντέλου από τις παρατηρήσεις και προσανατολισμός συστήματος αναφοράς

Αντίθετα, η επιλογή του X ως εξαρτημένη μεταβλητή μεταφράζεται στην αναζήτηση των συντελεστών ενός μοντέλου της μορφής $X = AX + B$. Η προσαρμογή του μοντέλου αυτού στα διαθέσιμα πειραματικά δεδομένα οδηγεί στις τιμές $\hat{A}$ και $\hat{B}$ των συντελεστών, μέσω της ελαχιστοποίησης του αθροίσματος των τετραγώνων των οριζόντιων αποστάσεων (σχήμα 1):

$$\chi^2 = \sum_{i=1}^{N} \frac{\Delta x_i^2}{u_{x,i}^2} = \sum_{i=1}^{N} \left(\frac{x_i - (Ax_i + B)}{u_{x,i}} \right)^2 \quad (5)$$

Το προσαρμοσμένο μοντέλο μπορεί τότε να παρουσιαστεί με τη μορφή:

$$y = \hat{a}^* x + \hat{b}^* \quad (6)$$

θέτοντας $\hat{a}^* = 1/\hat{A}$ και $\hat{b}^* = -\hat{B}/\hat{A}$.

Εύκολα εξάλλου αποδεικνύεται ότι, για τη συνήθη περίπτωση που τα σφάλματα στις παρατηρήσεις θεωρούνται ότι ακολουθούν μια κανονική κατανομή με την ίδια για όλα τα σημεία τυπική απόκλιση, ο συντελεστής προσδιορισμού r^2 είναι ο ίδιος και στις δύο περιπτώσεις και ότι $\hat{a} / \hat{a}^* = r^2$. Διαπίστωση που οδηγεί στο συμπέρασμα ότι και οι δύο επιλογές είναι καταρχήν εξίσου αποδεκτές από την άποψη της αποτελεσματικότητάς τους, καθώς και ότι υπάρχει πάντα μια μικρότερη ή μεγαλύτερη διαφορά μεταξύ τους, ανάλογα με το πόσο αποκλίνει η προσαρμογή από την τέλεια συσχέτιση (r^2=1).

Είναι επομένως φανερό ότι η επιλογή της εξαρτημένης μεταβλητής οδηγεί σε δύο διαφορετικά αποτελέσματα, με δυνητικά σημαντικές επιπτώσεις στην μετέπειτα χρήση του μοντέλου, παρά το ότι και τα δύο είναι καταρχήν αποδεκτά. Αν και στη σχετική βιβλιογραφία το ζήτημα αυτό παρουσιάζεται κάπως συγκεχυμένα, κυρίως λόγω των συχνά αντικρουόμενων εκδοχών που προτείνονται, από τη θεωρητική θεμελίωση της OLS προκύπτει ότι η επιλογή του συστήματος αναφοράς, δηλαδή του ποια από τις μεταβλητές ποσότητες *X* ή *Y* θα θεωρηθεί ως εξαρτημένη και ποια ως ανεξάρτητη, δεν μπορεί να είναι αυθαίρετη, αλλά οφείλει να υπακούει σε ένα κριτήριο επιλογής που σχετίζεται με το εύρος των σφαλμάτων στις εμπειρικές παρατηρήσεις. Ποιο συγκεκριμένα, στα πλαίσια της προσέγγισης OLS, ως εξαρτημένη πρέπει να θεωρηθεί η μεταβλητή της οποίας οι εμπειρικές παρατηρήσεις χαρακτηρίζονται από αβεβαιότητες. Αντίθετα, αυτές της ανεξάρτητης μεταβλητής θεωρούνται, και οφείλουν να μπορούν να θεωρηθούν, ως απαλλαγμένες από σφάλματα, προϋπόθεση η οποία σε πολλές περιπτώσεις είναι δύσκολο να γίνει αποδεκτή. Για το λόγο αυτό άλλωστε συνίσταται, στα πλαίσια της OLS προσέγγισης, να επιλέγεται ως εξαρτημένη μεταβλητή αυτή που χαρακτηρίζεται από μεγαλύτερες αβεβαιότητες (classical calibration), αν και, σύμφωνα με ορισμένους συγγραφείς, η αντιστροφή των ρόλων μεταξύ *x* και *y* (inverse calibration), φαίνεται να είναι περισσότερο αποτελεσματική (Cendner et al.).

3.2. Η μέθοδος των ελαχίστων τετραγώνων σε συνθήκες γενικευμένης αβεβαιότητας

Στην πράξη, τα δεδομένα των παρατηρήσεων, είτε αυτά προέρχονται από εμπειρικές παρατηρήσεις, είτε από πειραματικές διεργασίες, είτε από αριθμητικά πειράματα, χαρακτηρίζονται πάντα από κάποια αβεβαιότητα. Η αβεβαιότητα αυτή επηρεάζει εν γένει όχι μόνο την αβεβαιότητα των συντελεστών που θα προκύψουν από την προσαρμογή του μοντέλου, αλλά και την ίδια τους την τιμή. Για το λόγο αυτό, η αξιοποίηση προσεγγίσεων που θα συνυπολογίζουν τις αβεβαιότητες τόσο στην εξαρτημένη όσο και στην ανεξάρτητη μεταβλητή, είναι απολύτως αναγκαίες, ειδικά όταν επιδιώκεται υψηλή μετρολογική στάθμη στα αποτελέσματα από την μελλοντική χρήση του μοντέλου. Χαρακτηριστικά μπορεί να αναφερθεί το παράδειγμα της διακρίβωσης όπου, τόσο η τιμή αναφοράς όσο και η απόκριση του προς διακρίβωση οργάνου, χαρακτηρίζονται από αβεβαιότητες.

Για τον λόγο αυτό προτάθηκαν κατά καιρούς διάφορες προσεγγίσεις, οι περισσότερες από τις οποίες αναπτύχθηκαν στα πλαίσια μετρολογικών εφαρμογών. Ορισμένες από αυτές εξετάζονται

συνοπτικά στις επόμενες παραγράφους, επισημαίνοντας ωστόσο την αδυναμία συστηματικής τους κατάταξης λόγω της συχνά διαφορετικής χρήσης της κατά περίπτωση χρησιμοποιούμενης ορολογίας.

Στην πράξη επίσης, η αβεβαιότητα που συνοδεύει τις πειραματικές παρατηρήσεις δεν υπάρχει κανένας λόγος να είναι η ίδια σε όλα τα σημεία παρατήρησης. Από την άποψη αυτή, η προσαρμογή του μοντέλου σε δεδομένα που χαρακτηρίζονται από μεταβαλλόμενες αβεβαιότητες, τόσο στα *X* όσο και στα *Y* (heteroscedastic errors), συνιστά σημαντική πρόκληση για τον μετρολόγο.

Η συνηθέστερη απάντηση στο πρόβλημα αυτό συνίσταται στην εφαρμογή της προσέγγισης των Σταθμισμένων Ελαχίστων Τετραγώνων (Weighted Least Square, σε συντομογραφία WLS), η οποία συνιστά βελτιωμένη εκδοχή της OLS. Η γενική ιδέα πίσω από την προσέγγιση WLS είναι ότι επιδιώκεται η ελαχιστοποίηση μια συνάρτησης όπως αυτής της εξίσωσης (5), σταθμίζοντας τις κατακόρυφες αποστάσεις των πειραματικών σημείων από τη γραμμή του μοντέλου, με την αβεβαιότητα που χαρακτηρίζει το κάθε συγκεκριμένο πειραματικό σημείο, με τρόπο που οι περισσότερο αβέβαιες παρατηρήσεις να επηρεάζουν λιγότερο το τελικό αποτέλεσμα, συγκρινόμενες με τις λιγότερο αβέβαιες (σχήμα 1). Η WLS approach παίρνει υπόψη της την heteroscedasticity στον άξονα των *Y*, ενώ, στη συνηθέστερη εκδοχή της, θεωρεί και αυτή τον άξονα των *X* απαλλαγμένο από σφάλματα. Έτσι, η μέθοδος αυτή προκρίνει την ελαχιστοποίηση των κατακόρυφων αποστάσεων Δy των πειραματικών σημείων από την Model line, με κάθε μια από αυτές τις αποστάσεις να σταθμίζεται από την αβεβαιότητα που χαρακτηρίζει τη συγκεκριμένη τιμή (Fuller 1987, Dietrich 1991):

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - (ax_i + b)}{u_{y,i}} \right)^2 \qquad (7)$$

Μια άλλη παραλλαγή, η αποκαλούμενη και προσαρμογή σταθερής μεταβλητότητας (constant variance ration ή CVR), θεωρεί ένα σταθερό λόγο για τις μεταβλητότητες στους δύο άξονες, για όλα τα σημεία παρατήρησης ($u_x / u_y = \text{σταθερό}$) (Mandel 1984). Η υπόθεση ωστόσο του σταθερού λόγου των αβεβαιοτήτων των τιμών των δύο μεταβλητών, αν και διευκολύνει σημαντικά την επίλυση του προβλήματος, δεν είναι ρεαλιστική στο βαθμό που δεν περιγράφει μια συγκεκριμένη υπαρκτή κατάσταση.

Κατά καιρούς έχουν προταθεί τεχνικές προσαρμογής οι οποίες επιχειρούν να υπερβούν τους περιορισμούς των προηγούμενων μεθόδων. Χαρακτηριστικό είναι το παράδειγμα μιας άλλης εκδοχής της WLS, αποκαλούμενης και bivariate least square (BLS). Σύμφωνα με την προσέγγιση αυτή, συνυπολογίζονται, μέσω μιας επαναληπτικής διαδικασίας, οι αβεβαιότητες τόσο στα *x* όσο και στα *y*, με τον παράγοντα στάθμισης είναι η αβεβαιότητα που χαρακτηρίζει συνολικά το κάθε σημείο παρατήρησης (Mathioulakis and Belessiotis 2000, Javier et al. 2001):

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - (ax_i + b))^2}{u_{y,i}^2 + a^2 u_{x,i}^2} \qquad (8)$$

Και στην περίπτωση αυτή επιδιώκεται η ελαχιστοποίηση του αθροίσματος των κατακόρυφων αποκλίσεων των πειραματικών σημείων από το μοντέλο, με τις αποκλίσεις αυτές να σταθμίζονται με τρόπο που οι περισσότερο αβέβαιες παρατηρήσεις να επηρεάζουν λιγότερο το τελικό αποτέλεσμα συγκρινόμενες με τις λιγότερο αβέβαιες.

Σε όλες τις παραπάνω περιπτώσεις επιχειρείται η ελαχιστοποίηση των κατακόρυφων αποστάσεων Δy των πειραματικών σημείων από την fitted line, με ή χωρίς στάθμιση. Γίνεται δηλαδή, έμμεσα ή άμεσα, η παραδοχή ότι η όποια απόκλιση των πειραματικών παρατηρήσεων

από το μοντέλο οφείλεται μόνο στα σφάλματα στις τιμές του *Y*. Από την άποψη αυτή η περιστροφή των αξόνων, δηλαδή η εύρεση των συντελεστών της γραμμικής συσχέτισης $x = a * y + b *$ εν γένει δεν καταλήγει σε ισοδύναμα αποτελέσματα και δεν συνάδει με τη θεωρητική θεμελίωση της μεθόδου των ελαχίστων τετραγώνων.

Στην περίπτωση όμως που και οι δύο μεταβλητές χαρακτηρίζονται από αβεβαιότητες, οι αποκλίσεις των πειραματικών παρατηρήσεων από το μοντέλο είναι δυνατόν να οφείλονται σε σφάλματα και στις δύο μεταβλητές. Από την άποψη αυτή είναι ποιο λογικό να αναζητηθεί μια μεθοδολογία επίλυσης που θα οδηγεί σε αποτελέσματα τα οποία είναι ανεξάρτητα από τον χαρακτηρισμό της μιας ή της άλλης μεταβλητής ως ανεξάρτητης ή εξαρτημένης.

Μια τέτοια προσέγγιση στο γενικό πρόβλημα της προσαρμογής ενός γραμμικού μοντέλου σε δεδομένα με αβεβαιότητες και στους δύο άξονες διατυπώθηκε πρώτα από τον Pearson, ήδη από τις αρχές του 20 αιώνα. Μια γενική λύση προτάθηκε από τον Deming, βασισμένη στην ελαχιστοποίηση της συνάρτησης (Reed 1992):

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{\left(x_i - x_i^{/}\right)^2}{u_{x,i}^2} + \frac{\left(y_i - y_i^{/}\right)^2}{u_{y,i}^2} \right) \qquad (9)$$

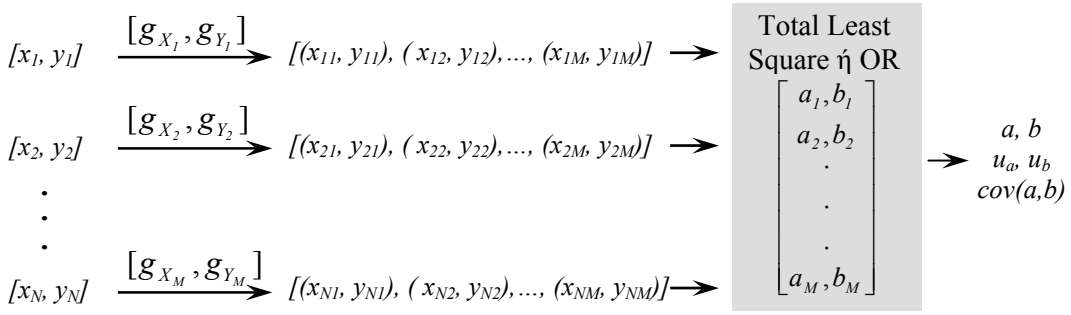
όπου (x_i, y_i) αντιπροσωπεύει ένα σημείο παρατήρησης, $(u_{x,i}, u_{y,i})$ είναι οι αβεβαιότητες που χαρακτηρίζουν τις τιμές που παρατηρήθηκαν και $(x^{/}, y^{/})$ είναι οι αντίστοιχες τιμές που υπολογίζονται από το μοντέλο (σχήμα 1). Η προσέγγιση αυτή, αποκαλούμενη και Orthogonal Least-Square (WOLS), οδηγεί σε λύσεις ανεξάρτητες από το πλαίσιο αναφοράς και δεν κάνει καμία διάκριση μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής, όπως φαίνεται και από τη μορφή της εξίσωσης (9). Η δυσκολία όμως της έγκειται στο ότι η ελαχιστοποίηση της ποσότητας χ^2 επιβάλει την χρήση επαναληπτικών μεθόδων, δεδομένου ότι ο υπολογισμός των $x^{/}$ και $y^{/}$ προϋποθέτει τη γνώση του μοντέλου, πριν ακόμα αυτό γίνει γνωστό. Αργότερα, άλλοι συγγραφείς, βασισμένοι στην προσέγγιση του Deming, πρότειναν προσεγγιστικές ή αναλυτικές λύσεις στο πρόβλημα αυτό, αν και ορισμένες από τις λύσεις αυτές αμφισβητήθηκαν στη συνέχεια ως προς την ορθότητά τους (Reed 1992, York 1966). Ακόμα ποιο πρόσφατα προτάθηκε μια διαφορική υπολογιστική τεχνική, αποκαλούμενη και Weighted Total Least-Square (WTLS) η οποία οδηγεί σε ελαφρά διαφορετικά αποτελέσματα όσον αφορά τις αβεβαιότητες στις τιμές των συντελεστών *a* και *b* του μοντέλου (Krystek 2007).
Μια ειδική περίπτωση είναι αυτή που οι αβεβαιότητες στα *x* και *y* είναι ίσες σε κάθε σημείο, δηλαδή όταν $u_{x,i}/u_{y,i}=1$, $i=1,\ldots,N$. Η εξίσωση (9) καταλήγει τότε στο αναφερόμενο ως πρόβλημα των Ολικών Ελαχίστων Τετραγώνων (Total Least-Squares ή TLS), η επίλυση του οποίου ισοδυναμεί με την ελαχιστοποίηση των κάθετων αποστάσεων από την γραμμή του μοντέλου, προσέγγιση αποκαλούμενη από ορισμένους συγγραφείς και Orthogonal Regression (OR). Στην γενική όμως περίπτωση που οι αβεβαιότητες δεν είναι ίσες μεταξύ τους, οι εν λόγω αποστάσεις δεν είναι κάθετες αλλά κεκλιμένες ως προς την γραμμή του μοντέλου (σχήμα 1).

4. Μέθοδος των ελαχίστων τετραγώνων και προσομοίωση Monte-Carlo

Η χρήση της προσομοίωσης Monte-Carlo για την εκτίμηση των αβεβαιοτήτων, αν και σχετικά πρόσφατη, έχει αναδειχθεί σε μια ενδιαφέρουσα εναλλακτική δυνατότητα, ειδικότερα σε περιπτώσεις που η απευθείας εφαρμογή του νόμου διάδοσης των αβεβαιοτήτων δεν έχει άμεση εφαρμογή, λόγω ύπαρξης μη γραμμικότητας ή αδυναμίας υπολογισμού των μερικών παραγώγων του μοντέλου μέτρησης (Cox and Siebert 2006).

Η προσομοίωση Monte-Carlo επιτρέπει την εκτίμηση της τιμής ενός μεγέθους και της αβεβαιότητας που το χαρακτηρίζει, όταν η τιμή αυτή προκύπτει από τις τιμές άλλων πρωτογενών μεγεθών, με τη βοήθεια κατάλληλων σχέσεων οι οποίες συνιστούν το μοντέλο μέτρησης. Η διαθέσιμη γνώση για κάθε ένα από τα πρωτογενή μεγέθη (τιμή και μεταβλητότητα) παρουσιάζεται με τη μορφή μιας κατανομής πιθανοτήτων, υλοποιούμενης στην πράξη από μεγάλο πλήθος τιμών, η στατιστική συμπεριφορά των οποίων ταυτίζεται με τα χαρακτηριστικά της αντίστοιχης κατανομής πιθανοτήτων. Η τιμή του παράγωγου μεγέθους υπολογίζεται με τη βοήθεια του μοντέλου μέτρησης για όλους τους συνδυασμούς τιμών των πρωτογενών μεγεθών, οδηγώντας στη συγκέντρωση ενός συνόλου τιμών, η στατιστική επεξεργασία του οποίου επιτρέπει τον προσδιορισμό μιας εκτιμώμενης τιμής και της μεταβλητότητας που χαρακτηρίζει την τιμή αυτή (ISO 2008, Hall 2006).

Η χρήση της τεχνικής αυτής, αποκαλούμενης και Monte-Carlo Least Square (MCLS) μέθοδος, προτάθηκε τελευταία και για την επίλυση του γενικού προβλήματος των ελαχίστων τετραγώνων (Mathioulakis et al. 2009). Το ενδιαφέρον της προσέγγισης αυτής συνίσταται στο ότι μπορεί να εφαρμοστεί χωρίς να γίνει καμία υπόθεση σχετικά με το είδος ή το εύρος των αβεβαιοτήτων και στο ότι οδηγεί σε αποτελέσματα που είναι ανεξάρτητα του ορισμού της μιας ή της άλλης μεταβλητής ως ανεξάρτητης ή εξαρτημένης. Επιπλέον, η τεχνική αυτή οδηγεί σε μια ρεαλιστική εκτίμηση της μεταβλητότητας και της συμμεταβλητότητας των συντελεστών του μοντέλου, επιτρέποντας έτσι την πλήρη αξιοποίησή του στα πλαίσια μιας μετέπειτα ανάλυσης αβεβαιοτήτων.



Σχήμα 2: Σχηματικό διάγραμμα προσέγγισης MCLS

Η εφαρμογή της τεχνικής Monte-Carlo στη περίπτωση της προσαρμογής ενός γραμμικού μοντέλου σε αβέβαια δεδομένα παρατηρήσεων, ακολουθεί τα παρακάτω βήματα (σχήμα 2):

- Για κάθε σημείο παρατήρησης, η πληροφορία που σχετίζεται με μια δεδομένη τιμή x_i κωδικοποιείται από μια Συνάρτηση Κατανομής Πιθανοτήτων (ΣΚΠ), g_{x_i}, παίρνοντας υπόψη και την αβεβαιότητα u_{x_i} που χαρακτηρίζει τη συγκεκριμένη τιμή. Ο τύπος της ΣΚΠ (κανονική, ορθογώνια, τριγωνική κλπ), εξαρτάται από το είδος της διαθέσιμης πειραματικής πληροφορίας. Χρησιμοποιείται γεννήτρια δεδομένων για την παραγωγή ενός σετ M αριθμών $x_{i,j}$, $j=1,...,M$, των οποίων η κατανομή είναι η g_{x_i} (Matsumoto and Nishimura, 1998).
- Η ίδια διαδικασία επαναλαμβάνεται για όλα τα σημεία, με τρόπο που να είναι τελικά διαθέσιμα $j=1,...,M$ σετ δεδομένων, καθένα από τα οποία περιέχει N σημεία $(x_{1,j}, y_{1,j}),..., (x_{N,j}, y_{N,j})$.
- Για κάθε ένα από τα M σετ $(x_{1,j}, y_{1,j}), (x_{2,j}, y_{2,j}),..., (x_{N,j}, y_{N,j})$, υπολογίζονται οι αντίστοιχοι συντελεστές a_j και b_j με τη μέθοδο Total Least Square ή Orthogonal Regression. Η

Orthogonal Regression επιλέχθηκε ως η πλέον κατάλληλη, δεδομένου ότι μεταχειρίζεται ισότιμα τις δύο μεταβλητές. Πράγματι, στη συγκεκριμένη περίπτωση, δεν υπάρχει λόγος επιλογής της μιας ή της άλλης μεταβλητής ως περισσότερο αβέβαιης.

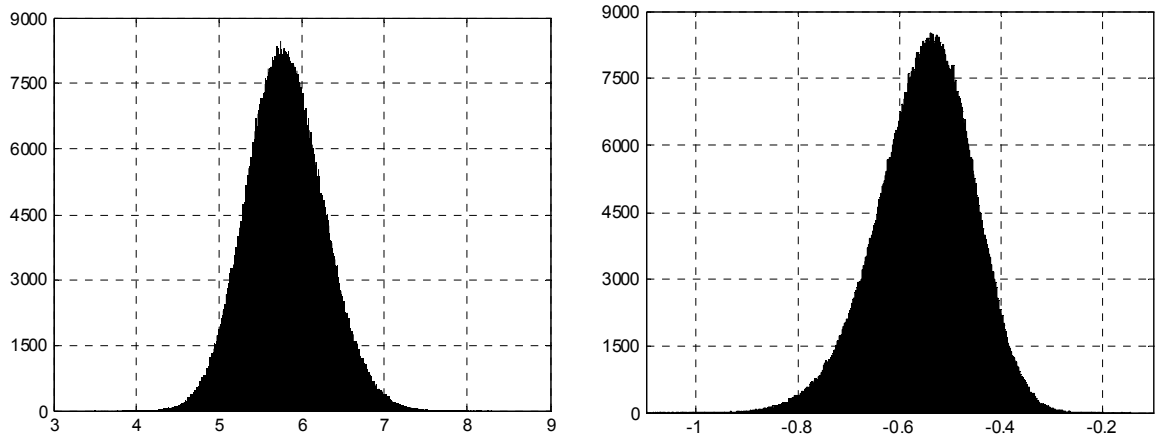
- Με βάση τη διαδικασία αυτή, παράγονται M τιμές των a και b. Η επεξεργασία των τιμών αυτών επιτρέπει την εύρεση μιας προσεγγιστικής διακριτοποιημένης μορφής της ΣΚΠ των a και b, από την οποία υπολογίζονται οι μέσες τιμές $\tilde{a}$ και $\tilde{b}$ (καλύτερες εκτιμήσεις), οι αβεβαιότητες που χαρακτηρίζουν τις μέσες αυτές τιμές, καθώς και τη συμμεταβλητότητά τους $cov(\tilde{a}, \tilde{b})$.

5. Παράδειγμα εφαρμογής

Ως παράδειγμα εφαρμογής των όσων αναλύθηκαν στα προηγούμενα κεφάλαια, επιλέχθηκαν τα γνωστά από τη σχετική βιβλιογραφία δεδομένα του Pearson (Πίνακας 1), με τις αβεβαιότητες που προτάθηκαν από τον York, δεδομένα τα οποία έχουν χρησιμοποιηθεί κατά καιρούς και από άλλους συρραφείς [10, 11, 12]. Για την παραγωγή των τιμών που υλοποιούν την ΣΚΠ των x_i και y_i σε κάθε σημείο παρατήρησης i, χρησιμοποιήθηκε η γεννήτρια Mersenne Twister η οποία έχει επιδείξει ικανοποιητικές στατιστικές επιδόσεις [17].

Πίνακας 1:Δεδομένα Pearson (με αβεβαιότητες από York)

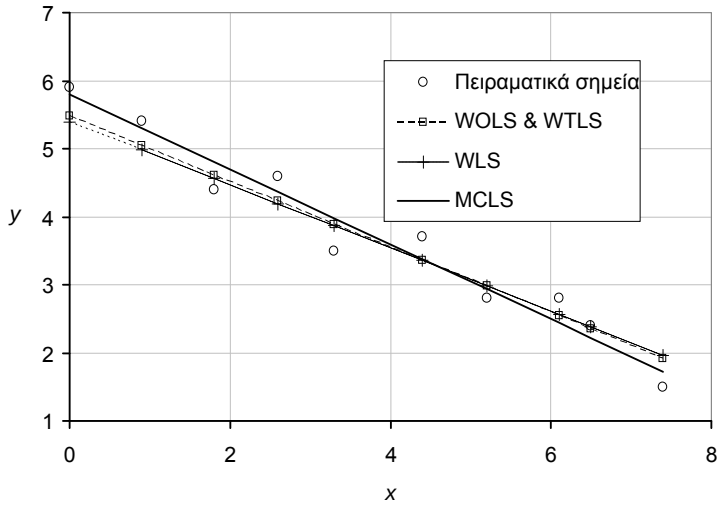
i	x_i	y_i	$u_{x,i}$	$u_{y,i}$
1	0,000	5,900	0,032	1,000
2	0,900	5,400	0,032	0,745
3	1,800	4,400	0,045	0,500
4	2,600	4,600	0,035	0,354
5	3,300	3,500	0,071	0,224
6	4,400	3,700	0,112	0,224
7	5,200	2,800	0,129	0,120
8	6,100	2,800	0,224	0,120
9	6,500	2,400	0,745	0,100
10	7,400	1,500	1,000	0,045



Σχήμα 3: Ιστογράμματα των τιμών των συντελεστών a και b του γραμμικού μοντέλου (αριστερά και δεξιά αντίστοιχα)

Στο σχήμα 3 δίδονται τα ιστογράμματα των τιμών των συντελεστών *a* και *b* που υπολογίστηκαν με τη μέθοδο MCLS, με την κάθε μία από αυτές να αντιστοιχεί σε ένα από τους $M=10^6$ συνδυασμούς των πρωτογενών δεδομένων. Ενδιαφέρον παρουσιάζει το ιστόγραμμα του συντελεστή *b*, στο βαθμό που αναδεικνύει μια σαφή απομάκρυνση από την κανονική κατανομή, αντίθετα με την συνήθως υιοθετούμενη προσέγγιση για κανονικότητα.

Στο σχήμα 4 απεικονίζονται συγκριτικά οι γραφικές παραστάσεις των διαφόρων γραμμικών μοντέλων που προσδιορίστηκαν ακολουθώντας τις κυριότερες από τις μεθόδους που εξετάστηκαν. Τα αποτελέσματα όσον αφορά τους υπολογιζόμενους συντελεστές του μοντέλου φαίνονται συγκεντρωτικά στον Πίνακα 2.



Σχήμα 4: Συγκριτική απεικόνιση των αποτελεσμάτων των διαφορετικών προσεγγίσεων υλοποίησης της προσαρμογής με τη μέθοδο των ελαχίστων τετραγώνων

Πίνακα; 2: Αποτέλεσμα προσαρμογής με τη μέθοδο των ελαχίστων τετραγώνων για διάφορες προσεγγίσεις

Method	$\tilde{a}$	$\tilde{b}$	$u_{\tilde{a}}$	$u_{\tilde{b}}$	$cov(\tilde{a}, \tilde{b})$
MCLS	5.800	-0.551	0.466	0.094	-0.0421
OLS	5.761	-0.539	-	-	-
WLS ή BLS	5.396	-0.464	0.296	0.058	0.0165
WTLS	5.480	-0.480	0.292	0.057	-0.0162
WOLS	5.480	-0.480	0.355	0.070	-0.0162

6. Συμπεράσματα

Από τα αποτελέσματα που παρουσιάστηκαν στις προηγούμενες ενότητες προκύπτει ότι υπάρχουν, όπως είναι αναμενόμενο, εμφανείς διαφορές ανάμεσα στις διάφορες προσεγγίσεις, γεγονός που συνηγορεί υπέρ μιας τεκμηριωμένης επιλογής αυτής που επιλέγεται τελικά, ανάλογα με τα ιδιαίτερα χαρακτηριστικά των εξεταζόμενων δεδομένων. Στις περιπτώσεις δεδομένων τα οποία χαρακτηρίζονται από απουσία σφαλμάτων ή από συγκριτικά ασήμαντες αβεβαιότητες στην μία από τις δύο μεταβλητές, η κλασική μέθοδος των κοινών ελαχίστων τετραγώνων (OLS) είναι επαρκής. Προϋπόθεση όμως είναι να επιλεγεί ως ανεξάρτητη

μεταβλητή αυτή που παρατηρήθηκε σε συνθήκες στιβαρού στατιστικού ελέγχου και μπορεί να θεωρείται ως επακριβώς γνωστή.

Στις συνήθεις περιπτώσεις που όλα τα δεδομένα χαρακτηρίζονται από την ύπαρξη λιγότερο ή περισσότερο σημαντικών σφαλμάτων, η στάθμιση των αποστάσεων των πειραματικών δεδομένων από το μοντέλο με τις αντίστοιχες αβεβαιότητες κρίνεται απαραίτητη. Η στάθμιση ωστόσο αυτή, όταν περιορίζεται στην μία μόνο από τις δύο μεταβλητές, δεν διασφαλίζει την ανεξαρτησία από τον προσανατολισμό του πλαισίου αναφοράς, οδηγώντας έτσι σε δύο διαφορετικές λύσεις οι οποίες είναι καταρχήν εξίσου αποδεκτές. Οι προσεγγίσεις αντίθετα που διασφαλίζουν την ισοδυναμία των δύο μεταβλητών (εξαρτημένης και ανεξάρτητης), εμφανίζονται να έχουν μεγαλύτερη σταθερότητα, στο βαθμό που τα αποτελέσματά τους είναι ανεξάρτητα από τον προσανατολισμό του πλαισίου αναφοράς. Στην κατηγορία αυτή ανήκει η μέθοδος MCLS η οποία συνδυάζει την προσέγγιση των ορθογώνιων ελαχίστων τετραγώνων και της προσομοίωσης Monte-Carlo. Η μέθοδος MCLS είναι απλή στην εφαρμογή και μπορεί να ανταποκριθεί σε όλες τις δυνατές καταστάσεις, όσον αφορά το εύρος και τη μεταβλητότητα των αβεβαιοτήτων που χαρακτηρίζουν τα πειραματικά δεδομένα.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

Bich W. Cox M., Harris P., "Evolution of the 'Guide to the Expression of Uncertainty in Measurement'", Metrologia 43 (2006), 161-166.

Centner V., Massart D. L., de Jong S., "Inverse calibration predicts better than classical calibration", Fresenius J Anal Chem 361 (1998), 2–9.

Cox M. G., Siebert B. R. L., "The use of a Monte Carlo method for evaluating uncertainty and expanded uncertainty", Metrologia 43 (2006), 178–188.

Dietrich C.F., "Uncertainty, Calibration and Probability", 2nd edn, Bristol: Hilger 1991.

Fuller W.A., "Measurement error models", John Wiley, New York, 1987.

ISO, "Guide to the expression of uncertainty in measurements", ISO ed., Switzerland, 1995.

Hall B. D., "Evaluating methods of calculating measurement uncertainty", Metrologia 45 (2006), L5–L8.

ISO, "Evaluation of Measurement Data - Supplement 1 to the 'Guide to the Expression of Uncertainty in Measurement'—Propagation of distributions using a Monte Carlo method", Sevres, France: BIPM, 2008.

Javier del Rıo F., Riu J., Xavier Rius F., "Prediction intervals in linear regression taking into account errors on both axes", J. Chemometrics, 15 (2001), 773–788.

Krystek M., Anton M., "A weighted total least-squares algorithm for fitting a straight line", Meas. Sci. Technol., 18 3438–3442, 2007.

Lira I., "Curve adjustment by the least-squares method", Metrologia 37 (2000), 677–681.

Mandel J., "Fitting straight line when both variables are subject to error", J. Qual. Techno 16 (1984), 1–14.

Mathioulakis E., Belessiotis V., "Uncertainty and traceability in calibration by comparison", Meas. Sci. Technol. 11 (2000), 771–775.

Mathioulakis E., Belessiotis V., Panaras G., "A Monte-Carlo approach for fitting a straight line to data characterized by uncertainties", 14th International Congress of Metrology, Paris, 2009.

Matsumoto M., Nishimura T., Mersenne Twister: "A 623-dimensionally equidistributed uniform pseudorandom number generator", ACM Trans. on Modeling and Computer Simulation 8(1) (1998), 3-30.

Reed B. C., "Least square fits with errors in both coordinates", Am. J. Phys. 60 (1992), 59-62.

York D., "Least square fitting of a straight line", Can. J. Phys 44 (1966), 1079–86.