# *Κεφάλαιο 7* Ανάλυση δύο τυχαίων μεταβλητών - Εκτιμήσεις ελάχιστων τετραγώνων

Στα κεφάλαια 5 και 6 δώσαμε τη μεθοδολογία για την ανάλυση μίας υδρολογικής μεταβλητής καθώς και τις πιο συνηθισμένες συναρτήσεις κατανομής που χρησιμοποιούνται ως μοντέλα για τις μεταβλητές της τεχνικής υδρολογίας. Η ανάλυση μίας μεταβλητής αποτελεί το θεμέλιο της ανάλυσης διακινδύνευσης των έργων και των διαδικασιών που σχετίζονται με τους υδατικούς πόρους και την προστασία από υδρολογικούς κινδύνους. Σε αυτό το κεφάλαιο θα επεκταθούμε στην ταυτόχρονη ανάλυση δύο (ή και περισσότερων) τυχαίων μεταβλητών. Η ανάλυση δύο μεταβλητών δεν χρησιμοποιείται τόσο στην ανάλυση διακινδύνευσης. Κυρίως χρησιμοποιείται για τη διαπίστωση και τη μαθηματική διατύπωση της σχέσης που συνδέει δύο υδρολογικές μεταβλητές $X$ και $Y$, η οποία αξιοποιείται κυρίως σε εφαρμογές που έχουν σχέση με την επεξεργασία της υδρολογικής πληροφορίας. Για παράδειγμα, οι μεταβλητές $X$ και $Y$ μπορεί να συμβολίζουν το ύψος βροχής σε δύο διαφορετικά σημεία μιας λεκάνης, ή η μεταβλητή $X$ μπορεί να συμβολίζει την επιφανειακή βροχή μιας λεκάνης και η $Y$ την αντίστοιχη απορροή, κοκ. Μια μαθηματική σχέση ανάμεσα στις μεταβλητές $X$ και $Y$ μας επιτρέπει να συμπληρώσουμε τυχόν κενά στη σειρά των μετρήσεων της $Y$, αξιοποιώντας τις μετρήσεις της $X$. Μας επιτρέπει επίσης να ελέγξουμε τη συνέπεια των ταυτόχρονων μετρήσεων των $X$ και $Y$ και να εντοπίσουμε πιθανά σφάλματα στις μετρήσεις.

Η μαθηματική περιγραφή αυτού του είδους των σχέσεων είναι στατιστικού και όχι φυσικού χαρακτήρα, δηλαδή βασίζεται στα διαθέσιμα δείγματα των μεταβλητών και όχι στο φυσικό περιεχόμενο των μεταβλητών και των αλληλοσυνδέσεών τους. Οι σχέσεις που θα μελετήσουμε είναι κατ' αρχήν γραμμικές, δηλαδή της μορφής

$$Y = a + b\,X \tag{7.1}$$

όπου $a$ και $b$ αριθμητικές σταθερές, γνωστές ως *περίλημμα* (intercept) και *κλίση* (slope) της ευθείας, αντίστοιχα. Θα δώσουμε όμως αρκετά στοιχεία και για μη γραμμικές σχέσεις. Επίσης, θα δώσουμε γενικά στοιχεία για την μαθηματική διατύπωση σχέσεων ανάμεσα σε περισσότερες από δύο μεταβλητές.

Στην απλούστερη περίπτωση, το πρόβλημα της ανάλυσης δύο μεταβλητών περιλαμβάνει (α) την προσαρμογή του μοντέλου (7.1), δηλαδή την εκτίμηση των παραμέτρων $a$ και $b$, (β) τον έλεγχο του πόσο ισχυρό είναι το μοντέλο, και (γ) την εφαρμογή του μοντέλου για την εκτίμηση τιμών της $Y$ από τιμές της $X$. Η προσαρμογή και ο έλεγχος του μοντέλου στηρίζονται σε ένα παρατηρημένο δείγμα $n$ ζευγών (ταυτόχρονων) τιμών των μεταβλητών $(x_i, y_i)$. Για την καλύτερη κατανόηση της μεθόδου ξεκινούμε την παρουσίαση δίνοντας την προσδιοριστική (γεωμετρική και αναλυτική) αντιμετώπιση του προβλήματος και στη συνέχεια προχωρούμε στην πιθανοτική προσέγγισή του. Η πρώτη αντιμετωπίζει τα $(x_i, y_i)$ ως ζεύγη γνωστών αριθμών, χωρίς να τα συναρτά με τις τυχαίες μεταβλητές $X$ και $Y$. Όπως θα δούμε, τα αποτελέσματα και των δύο προσεγγίσεων συνδέονται στενά.

Θα πρέπει να πούμε ότι και η πιθανοτική προσέγγιση στο πρόβλημα δεν είναι μονοσήμαντη αλλά έχει διάφορες ερμηνείες και όψεις. Στο κείμενο αυτό περιοριζόμαστε σε μια συνοπτική περιγραφή του αντικειμένου, δίνοντας πάντως στοιχεία από τρεις διαφορετικές όψεις (στις ενότητες 7.1, 7.4 και 7.5, αντίστοιχα). Ο αναγνώστης που ενδιαφέρεται για πληρέστερη παρουσίασή του παραπέμπεται σε συγγράμματα γενικής στατιστικής (π.χ. Papoulis, 1990, σσ. 388-413· Benjamin and Cornell, 1970, σσ. 419-440· Spiegel, 1977, σσ. 258-305) αλλά και σε εξειδικευμένα συγγράμματα (π.χ. Draper and Smith, 1981).

## 7.1 Τυπική γραμμική παλινδρόμηση

### 7.1.1 Προσδιοριστική αντιμετώπιση

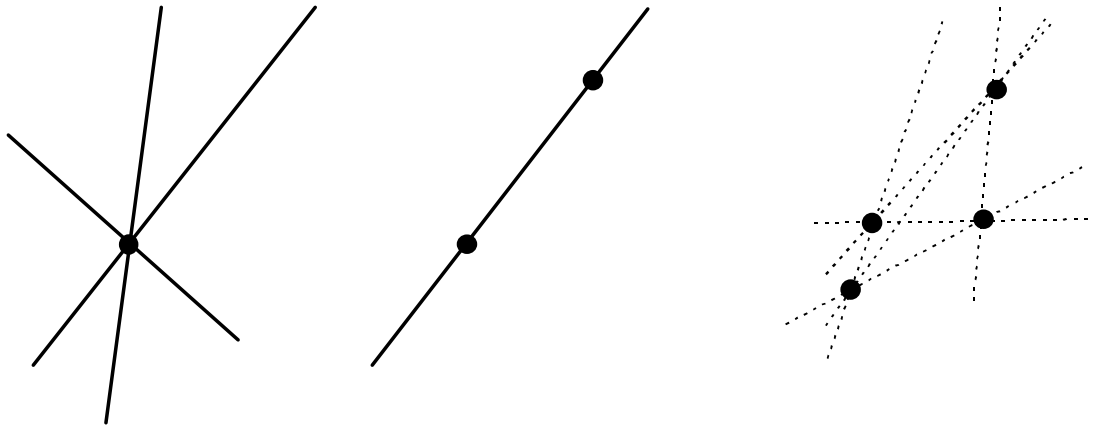
Από προσδιοριστική άποψη το πρόβλημα τίθεται κατ' αρχήν ως εξής: Δίνονται $n$ ζεύγη αριθμών $(x_i, y_i)$ και ζητείται να προσαρμοστεί σε αυτά μια ευθεία της μορφής

$$y = a + b\,x \tag{7.2}$$

Τα ζεύγη αριθμών $(x_i, y_i)$ μπορούν να θεωρηθούν ως σημεία $A_i$ στο επίπεδο, πράγμα που μας επιτρέπει να δώσουμε μια πρώτη γεωμετρική ερμηνεία στο πρόβλημα. Η γεωμετρική ερμηνεία παρατίθεται για λόγους αμεσότερης κατανόησης και όχι για να χρησιμοποιηθεί στην πράξη.

**Γεωμετρική ερμηνεία**

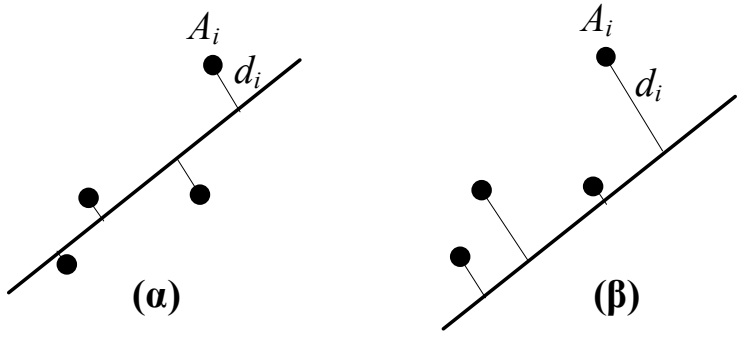
Ο όρος *προσαρμογή ευθείας* δεν είναι ακόμη αρκετά σαφής. Ας θεωρήσουμε κατ' αρχήν ότι σημαίνει την κατασκευή μιας ευθείας που περνά από όλα τα σημεία. Στην περίπτωση αυτή, όπως είναι γνωστό, το πρόβλημα έχει μία και μοναδική λύση για $n = 2$ και μόνο (Σχ. 7.1(β)). Για $n = 1$ υπάρχουν άπειρες λύσεις (Σχ. 7.1(α)), ενώ για $n > 2$ το πρόβλημα δεν έχει λύση, δηλαδή δεν υπάρχει ευθεία που να περνά ταυτόχρονα απ' όλα τα σημεία (Σχ. 7.1(γ)), εκτός από την ειδική περίπτωση που όλα τα σημεία βρίσκονται σε ευθεία.



**(α)** $n = 1$ **(β)** $n = 2$ **(γ)** $n > 2$

**Σχ. 7.1** Ακριβής γεωμετρική λύση στο πρόβλημα της κατασκευής ευθείας που περνά από $n$ σημεία.

Φυσικά, η περίπτωση που ενδιαφέρει από πρακτική άποψη είναι η τελευταία. Για να την αντιμετωπίσουμε θα πρέπει να εγκαταλείψουμε την απαίτηση ακριβούς λύσης, δηλαδή την αναζήτηση ευθείας που να περνά απ' όλα τα σημεία, και να δεχτούμε μια προσεγγιστική λύση, δηλαδή μια ευθεία που να μην απέχει πολύ από τα σημεία. Μπορούμε να συγκεκριμενοποιήσουμε και πάλι το πρόβλημα, θέτοντας ως απαίτηση, το άθροισμα των αποστάσεων των σημείων από την ευθεία να είναι το ελάχιστο δυνατό.




**Σχ. 7.2** Προσεγγιστική γεωμετρική λύση στο πρόβλημα της κατασκευής ευθείας που περνά από $n > 2$ σημεία.

Στο Σχ. 7.2 έχουμε σχεδιάσει δύο τυχαίες ευθείες και τις αποστάσεις τους $d_i$ από τα σημεία $A_i(x_i, y_i)$. Οπτικά διακρίνουμε ότι στην πρώτη ευθεία (Σχ. 7.2(α)) το άθροισμα των αποστάσεων $d_i$ είναι μικρότερο απ' ότι στη δεύτερη (Σχ. 7.2(β)). Εμπειρικά, θα μπορούσαμε να δοκιμάσουμε πολλές τυχαίες ευθείες, να βρούμε για κάθε μια το άθροισμα των αποστάσεων και να επιλέξουμε αυτή με το μικρότερο άθροισμα. Μια ακριβής γεωμετρική κατασκευή της ευθείας, με την αυστηρή έννοια, είναι πολύπλοκη. Άλλωστε δεν έχει και πρακτικό νόημα γιατί στην πραγματικότητα αυτό που στις εφαρμογές ενδιαφέρει είναι η εξίσωση (7.2) και όχι η γεωμετρική απεικόνισή της. Για το λόγο αυτό προχωρούμε στην αναλυτική αντιμετώπιση, κάνοντας και κάποια τροποποίηση ως προς το στόχο του προβλήματος.

**Αναλυτική προσέγγιση**

Σύμφωνα με την αναλυτική προσέγγιση, εφαρμόζοντας την εξίσωση (7.2) για καθένα από τα σημεία $(x_i, y_i)$ παίρνουμε ένα σύστημα από $n$ εξισώσεις τις μορφής

$$y_i = a + b\,x_i \tag{7.3}$$

Οι άγνωστοι του συστήματος των εξισώσεων είναι οι παράμετροι $a$ και $b$. Το σύστημα έχει μοναδική ακριβή λύση μόνο για $n = 2$, ενώ για $n > 2$ γενικά δεν έχει λύση, αλλά είναι ένα *υπερκαθορισμένο σύστημα*. Για να βρούμε τα $a$ και $b$ στην περίπτωση αυτή δεχόμαστε ότι η εξίσωση (7.2) δεν θα επαληθεύεται ακριβώς από τα σημεία $(x_i, y_i)$, αλλά θα υπάρχουν σφάλματα

$$w_i = y_i - (a + b\,x_i) \tag{7.4}$$

τα οποία προσπαθούμε να ελαχιστοποιήσουμε με κάποια έννοια.

Το απλούστερο υπολογιστικό σχήμα ελαχιστοποιεί το άθροισμα των τετραγώνων των επιμέρους σφαλμάτων

$$q = \sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \tag{7.5}^*$$

Η προκύπτουσα μέθοδος ονομάζεται μέθοδος των *ελάχιστων τετραγώνων*. Η ελαχιστοποίηση του μεγέθους $q$ διαφέρει από την ελαχιστοποίηση του αθροίσματος των αποστάσεων των σημείων από την ευθεία που συζητήθηκε στο προηγούμενο εδάφιο.

Η εύρεση των άγνωστων παραμέτρων που ελαχιστοποιούν το $q$ είναι απλή. Παρατηρούμε ότι το $q$ είναι συνάρτηση των $a$ και $b$ (τα $x_i$ και $y_i$ είναι γνωστοί αριθμοί) και κατά συνέπεια παίρνει την ελάχιστη τιμή του όταν

$$\frac{\partial q}{\partial a} = -2 \sum_{i=1}^{n} [y_i - (a + bx_i)] = 0 \tag{7.6}$$

---

* Η άθροιση γίνεται προφανώς για να συνυπολογίσουμε τα σφάλματα σε όλα τα επιμέρους σημεία. Η ύψωση στο τετράγωνο γίνεται για να απαλλαγούμε από το πρόσημο των επιμέρους σφαλμάτων. Εναλλακτικά θα μπορούσαμε να χρησιμοποιήσουμε τις απόλυτες τιμές των σφαλμάτων αλλά αυτό δημιουργεί υπολογιστική δυσκολία.

$$\frac{\partial q}{\partial b} = -2 \sum_{i=1}^{n} [y_i - (a + bx_i)]\, x_i = 0 \tag{7.7}$$

Ισοδύναμα οι εξισώσεις αυτές γράφονται

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{7.8}$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{7.9}$$

Η λύση του συστήματος είναι

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{7.10}$$

$$a = \overline{y} - b\,\overline{x} \tag{7.11}$$

όπου $\overline{x}$ και $\overline{y}$ οι μέσες τιμές των $x_i$ και $y_i$, αντίστοιχα, δηλαδή

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{7.12}$$

**Συντελεστής προσδιορισμού**

Οι εξισώσεις (7.10) και (7.11) δίνουν τις ζητούμενες παραμέτρους $b$ και $a$, χωρίς όμως να δίνουν καμιά πληροφορία σχετικά με το πόσο καλά είναι διατεταγμένα τα σημεία $(x_i, y_i)$ γύρω από την ευθεία που υπολογίστηκε. Αυτή την πληροφορία τη δίνει το μέγεθος $q$, το οποίο με αντικατάσταση των (7.10) και (7.11) στην (7.5) γίνεται μετά από πράξεις

$$q = \sum_{i=1}^{n} (y_i - \overline{y})^2 - \frac{\left[\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})\right]^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{7.13}$$

Συνήθως χρησιμοποιούμε την ακόλουθη αδιαστατοποιημένη μορφή του σφάλματος

$$d = 1 - \frac{q}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{7.14}$$

Το μέγεθος $d$ λέγεται *συντελεστής προσδιορισμού* (determination coefficient). Παρατηρούμε η μέγιστη τιμή του $d$ είναι 1 και αντιστοιχεί σε μηδενική τιμή του σφάλματος $q$. Η ελάχιστη τιμή του είναι 0 και αντιστοιχεί σε τιμή του σφάλματος

$$q = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{7.15}$$

Εύκολα μπορούμε να αποδείξουμε ότι το ελάχιστο τετραγωνικό σφάλμα δεν μπορεί να είναι μεγαλύτερο από την παραπάνω τιμή. Αν ήταν, τότε η ευθεία $y = a + bx$ με $a = \overline{y}$ και $b = 0$ θα έδινε μικρότερο σφάλμα, το οποίο είναι άτοπο, αφού οι (7.10) και (7.11) δίνουν το ελάχιστο σφάλμα. Κατά συνέπεια

$$0 \le d \le 1 \tag{7.16}$$

Αν αντικαταστήσουμε στην (7.14) την (7.13) παίρνουμε την τελική έκφραση του $d$

$$d = \frac{\left( n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right)^2}{\left[ n\sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2 \right] \left[ n\sum_{i=1}^{n} y_i^2 - \left( \sum_{i=1}^{n} y_i \right)^2 \right]} = \frac{\left[ \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}) \right]^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2} \quad (7.17)$$

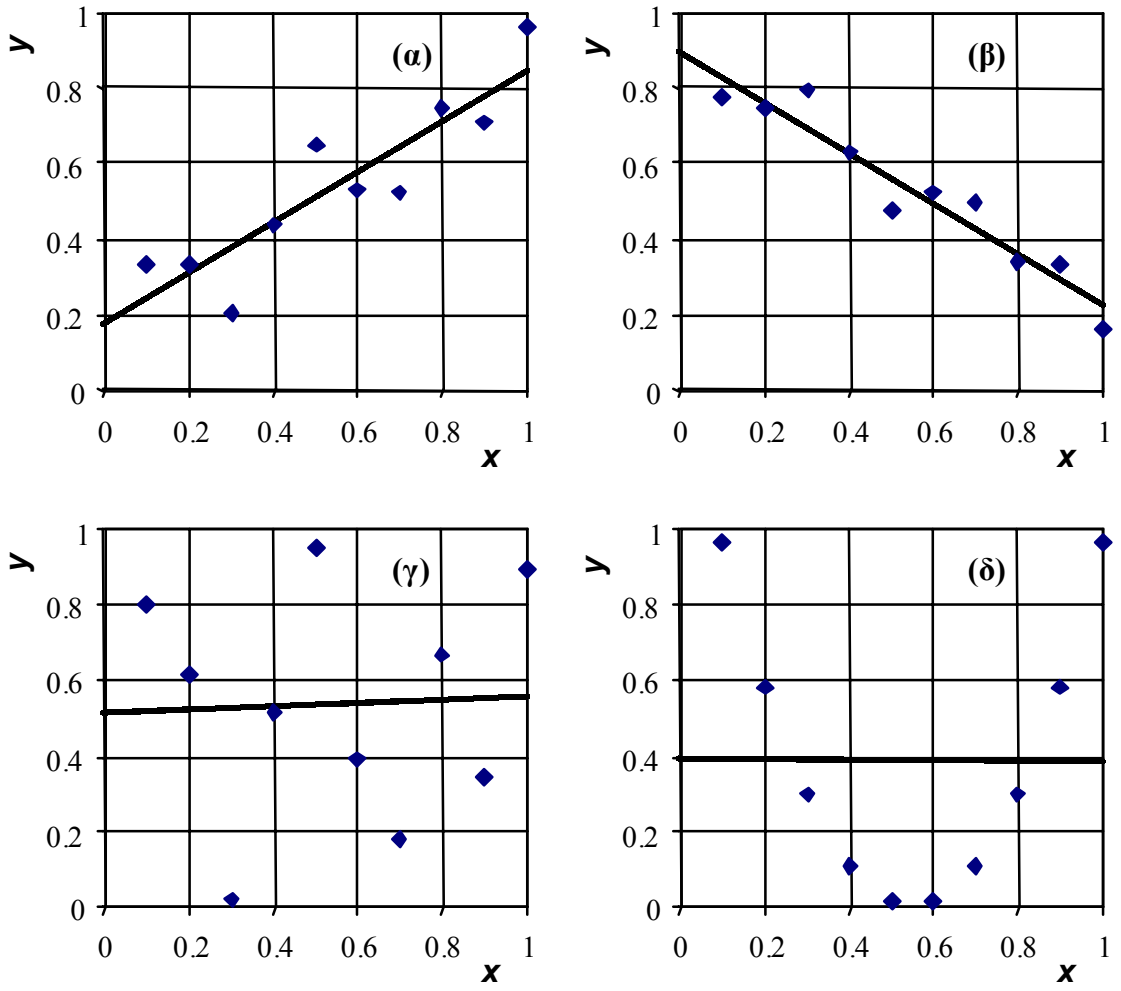
Η τετραγωνική του ρίζα

$$\begin{aligned} r = \sqrt{d} &= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[ n\sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2 \right] \left[ n\sum_{i=1}^{n} y_i^2 - \left( \sum_{i=1}^{n} y_i \right)^2 \right]}} \\ &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \end{aligned} \quad (7.18)$$

είναι ο γνωστός μας συντελεστής συσχέτισης, με τιμές

$$-1 \le r \le 1 \quad (7.19)$$

Συγκεκριμένα, το $r$ παίρνει τιμή κοντά στο 0 όταν η διάταξη των σημείων δεν είναι ευθύγραμμη, οπότε το προκύπτον σφάλμα είναι αρκετά μεγάλο, κοντά στο 1 όταν η διάταξη των σημείων είναι ευθύγραμμη και η κλίση $b$ της ευθείας είναι θετική, και κοντά στο –1 όταν η διάταξη των σημείων είναι ευθύγραμμη αλλά η κλίση $b$ της ευθείας είναι αρνητική (Σχ. 7.3).




**Σχ. 7.3** Παραδείγματα ευθειών ελάχιστων τετραγώνων για 10 σημεία $(x_i, y_i)$ και για τέσσερις περιπτώσεις διάταξης σημείων: (α) περίπου ευθύγραμμη διάταξη σημείων με θετική κλίση ($a = 0.17$, $b = 0.68$, $r = 0.90$)· (β) περίπου ευθύγραμμη διάταξη σημείων με αρνητική κλίση ($a = 0.90$, $b = –0.67$, $r = –0.96$)· (γ) τυχαία μη ευθύγραμμη διάταξη σημείων ($a = 0.52$, $b = 0.03$, $r = 0.03$)· (δ) παραβολική διάταξη σημείων ($a = 0.39$, $b = 0.00$, $r = 0.00$). Παρατηρούμε ότι στις περιπτώσεις (γ) και (δ) ο συντελεστής συσχέτισης είναι πρακτικώς μηδέν παρόλο που υπάρχει ουσιαστική διαφορά στη διάταξη των σημείων. Συγκεκριμένα στην περίπτωση (γ) η διάταξη των σημείων είναι τελείως τυχαία, ενώ στη (δ) τα σημεία υπακούουν σε ένα παραβολικό νόμο.

**Αντιστροφή του γραμμικού νόμου**

Αφού προσδιοριστεί η ευθεία $y = a + bx$, φαίνεται φυσική η άμεση αντιστροφή του νόμου, δηλαδή η έκφραση του $x$ ως προς $y$ με άμεση επίλυση, η οποία δίνει $x = (y – a) / b$. Ωστόσο, η έκφραση αυτή έχει προκύψει με ελαχιστοποίηση του σφάλματος του γραμμικού νόμου ως προς τη

μεταβλητή $y$. Αν θέλουμε την έκφραση του γραμμικού νόμου ως προς $x$, τότε θα πρέπει να ελαχιστοποιήσουμε το σφάλμα του νόμου ως προς $x$, δηλαδή την ποσότητα

$$q' = \sum_{i=1}^{n} w'^2_i = \sum_{i=1}^{n} [x_i - (a' + b'y_i)]^2 \tag{7.20}$$

όπου έχουμε υποθέσει ότι η νέα έκφραση του γραμμικού νόμου είναι η $x = a' + b'y$.

Σε αυτή την περίπτωση οι παράμετροι $b$ και $a$ δίνονται από τις ακόλουθες εξισώσεις, ανάλογες με τις (7.10) και (7.11):

$$b' = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{7.21}$$

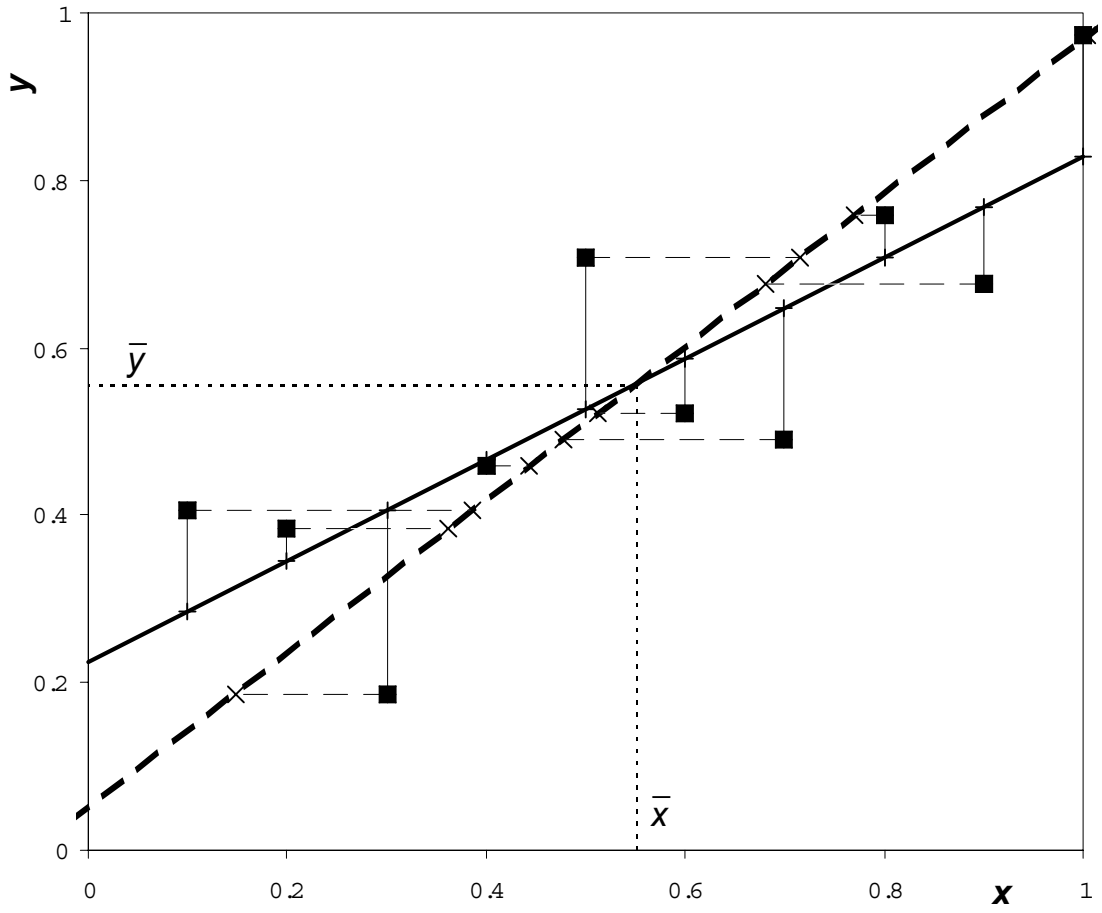
$$a' = \overline{x} - b'\,\overline{y} \tag{7.22}$$

Οι εξισώσεις (7.17) και (7.18) που δίνουν τους συντελεστές προσδιορισμού και συσχέτισης, αντίστοιχα, οι οποίες είναι συμμετρικές ως προς $x$ και $y$, ισχύουν ως έχουν και για αυτή την περίπτωση. Επισημαίνουμε ότι οι παράμετροι που προκύπτουν από τις (7.21) και (7.22) ορίζουν διαφορετική ευθεία από αυτή που ορίζουν οι παράμετροι που προκύπτουν από τις (7.10) και (7.11) (βλ. Σχ. 7.4). Οι δύο ευθείες ταυτίζονται μόνο όταν ο συντελεστής προσδιορισμού είναι ίσος με 1.

Εύκολα μπορούμε να διαπιστώσουμε χρησιμοποιώντας τις παραπάνω σχέσεις ότι οι εξισώσεις των δύο ευθειών μπορεί να γραφούν εναλλακτικά με την ακόλουθη μορφή

$$y - \overline{y} = b\,(x - \overline{x}) \qquad x - \overline{x} = b'\,(y - \overline{y}) \tag{7.23}$$

απ' όπου προκύπτει ότι και οι δύο ευθείες περνούν από το σημείο $(\overline{x}, \overline{y})$ (Σχ. 7.4).




**Σχ. 7.4** Επεξηγηματικό παράδειγμα για τις δύο ευθείες ελάχιστων τετραγώνων που ορίζονται από μια σημειοσειρά $(x_i, y_i)$. Τα 10 σημεία παριστάνονται με ■. Η έντονη συνεχής ευθεία με εξίσωση $y = 0.22 + 0.60\,x$ αντιστοιχεί στην ελαχιστοποίηση του τετραγωνικού σφάλματος $q$ (εξίσωση (7.5)). Τα επιμέρους σφάλματα $w_i$, δηλαδή οι κατακόρυφες αποστάσεις των σημείων από αυτή την ευθεία έχουν απεικονιστεί με λεπτή συνεχή γραμμή. Η έντονη διακεκομμένη ευθεία με εξίσωση $x = -0.05 + 1.09\,y$ αντιστοιχεί στην ελαχιστοποίηση του τετραγωνικού σφάλματος $q'$ (εξίσωση (7.20)). Τα επιμέρους σφάλματα $w'_i$, δηλαδή οι οριζόντιες αποστάσεις των σημείων από αυτή την ευθεία έχουν απεικονιστεί με λεπτή διακεκομμένη γραμμή. Ο συντελεστής συσχέτισης στο παράδειγμα είναι 0.81. Παρατηρούμε ότι οι δύο ευθείες τέμνονται στο σημείο $(\bar{x}, \bar{y})$.

### 7.1.2 Πιθανοτική αντιμετώπιση

**Εξισώσεις εκτίμησης παραμέτρων**

Σύμφωνα με μία από τις πιθανοτικές προσεγγίσεις τα ζεύγη $(x_i, y_i)$ θεωρούνται πραγματοποιήσεις δύο τυχαίων μεταβλητών $(X, Y)$ και το πρό-

βλημα είναι η εκτίμηση της $Y$ συναρτήσει της $X$. Στην περίπτωση γραμμικού νόμου ανάμεσα στις $X$ και $Y$ θα ισχύει

$$Y = a + bX + W \tag{7.24}$$

όπου $a$ και $b$ άγνωστες αριθμητικές παράμετροι και $W$ τυχαία μεταβλητή που εκφράζει το σφάλμα της εκτίμησης

$$\hat{Y} = a + bX \tag{7.25}$$

δηλαδή $W = Y - \hat{Y} = Y - (a + bX)$. Οι παράμετροι $a$ και $b$ υπολογίζονται έτσι ώστε η (7.25) να αποτελεί την εκτίμηση *ελάχιστου μέσου τετραγωνικού σφάλματος*, με την έννοια ότι ελαχιστοποιείται το μέγεθος

$$\begin{aligned} E[W^2] &= E[(Y - \hat{Y})^2] = E[(Y - (a + bX))^2] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [y - (a + bx)]^2 f_{XY}(x, y)\, dx\, dy \end{aligned} \tag{7.26}$$

όπου με $E[\ ]$ συμβολίζεται η αναμενόμενη τιμή. Παραγωγίζοντας την (7.26) ως προς $a$ και $b$ παίρνουμε μετά από πράξεις (βλ. π.χ. Papoulis, 1990, σ. 150) ένα σύστημα εξισώσεων, το οποίο επιλυόμενο δίνει

$$b = \frac{\sigma_{XY}}{\sigma_X^2} = \rho_{XY} \frac{\sigma_Y}{\sigma_X} \tag{7.27}$$

$$a = \mu_Y - b\, \mu_X \tag{7.28}$$

όπου $\mu_X$ και $\mu_Y$ οι μέσες τιμές των $X$ και $Y$, αντίστοιχα, $\sigma_X$ και $\sigma_Y$ οι αντίστοιχες τυπικές αποκλίσεις και $\sigma_{XY}$ και $\rho_{XY}$ η συνδιασπορά και ο συντελεστής συσχέτισης, αντίστοιχα, των $X$ και $Y$. Τα μεγέθη αυτά συνήθως είναι άγνωστα (παράμετροι πληθυσμού). Αν στη θέση τους τοποθετήσουμε τις εκτιμήσεις τους από το δείγμα των $n$ σημείων $(x_i, y_i)$, τότε παίρνουμε πάλι τις εξισώσεις (7.10) και (7.11), οι οποίες είχαν εξαχθεί με την προσδιοριστική προσέγγιση.

Η εξίσωση (7.25), αν αντικαταστήσουμε τις παραπάνω τιμές των $a$ και $b$, παίρνει την ακόλουθη μορφή

$$\frac{\hat{Y}-\mu_Y}{\sigma_Y}=\rho_{XY}\frac{X-\mu_X}{\sigma_X} \tag{7.29}$$

Κατά ανάλογο τρόπο προκύπτουν και οι εξισώσεις της αντίστροφης ευθείας ελάχιστων τετραγώνων, οι οποίες είναι

$$b'=\frac{\sigma_{XY}}{\sigma_Y^2}=\rho_{XY}\frac{\sigma_X}{\sigma_Y} \tag{7.30}$$

$$a'=\mu_X-b'\,\mu_Y \tag{7.31}$$

**Η έννοια της παλινδρόμησης**

Ο όρος *παλινδρόμηση* (regression) χρησιμοποιείται ως συνώνυμος του όρου *μέθοδος ελάχιστων τετραγώνων*. Ο όρος (αν και ανεπιτυχής) έχει καθιερωθεί από τον Galton*, ο οποίος παρατήρησε ότι "Τα πληθυσμιακά ακρότατα οπισθοχωρούν (regress)† προς τη μέση τιμή τους". Η παρατήρηση αυτή προέρχεται από τη μελέτη του ύψους των παιδιών σχετικά με το ύψος των γονέων. Από τη μελέτη φάνηκε ότι τα παιδιά ψηλών (αντίστροφα, κοντών) γονέων είναι κατά μέσο όρο κοντότερα (αντίστροφα, ψηλότερα) από τους γονείς τους. Αυτό μπορεί να ερμηνευτεί στατιστικά με βάση την ευθεία παλινδρόμησης (7.29), όπου οι μεταβλητές $X$ και $Y$ συμβολίζουν το ύψος των γονέων και των παιδιών, αντίστοιχα. Στη συγκεκριμένη περίπτωση ισχύει $\mu_X=\mu_Y=\mu$ και $\sigma_X=\sigma_Y$, οπότε η εξίσωση παλινδρόμησης παίρνει τη μορφή $\hat{Y}-\mu=\rho_{XY}\,(X-\mu)$. Δεδομένου ότι για τα εξεταζόμενα μεγέθη $0<\rho_{XY}<1$, θα είναι $\hat{Y}-\mu<X-\mu$, πράγμα που δείχνει την τάση μείωσης της διαφοράς από τη μέση τιμή (οπισθοχώρησης προς τη μέση τιμή).

Ας σημειωθεί ότι ο όρος παλινδρόμηση δεν χρησιμοποιείται μόνο για την εξεταζόμενη περίπτωση γραμμικής σχέσης ανάμεσα σε δύο μεταβλη-

---

* Sir Francis Galton, γενετιστής και βιοστατιστικός (1822-1911).

† Οι αγγλικές λέξεις *regress* και *regression* στην ελληνική επιστημονική ορολογία καθιερώθηκε να αποδίδονται ως *παλινδρομώ* και *παλινδρόμηση*, αν και στη συγκεκριμένη περίπτωση έχουν το νόημα της οπισθοχώρησης.

τές, αλλά για οποιασδήποτε μορφής σχέση ανάμεσα σε οποιοδήποτε αριθμό μεταβλητών (βλ. και ενότητες 7.3 και 7.4).

**Ροπές του σφάλματος εκτίμησης**

Σύμφωνα με την πιθανοτική θεώρηση που αναπτύσσουμε εδώ, το σφάλμα $W$ θεωρείται ως τυχαία μεταβλητή. Συνδυάζοντας τις (7.24) και (7.28) βρίσκουμε ότι

$$W = (Y - \mu_Y) - b\,(X - \mu_X) \tag{7.32}$$

Παίρνοντας αναμενόμενες τιμές στην παραπάνω καταλήγουμε ότι

$$\mu_W = 0 \tag{7.33}$$

δηλαδή η αναμενόμενη τιμή του σφάλματος είναι μηδενική.

Εξ άλλου υψώνοντας τα δύο μέλη της (7.32) στο τετράγωνο και παίρνοντας πάλι αναμενόμενες τιμές βρίσκουμε ότι η διασπορά του $W$ είναι

$$\sigma_W^2 = \sigma_Y^2 + b^2\,\sigma_X^2 - 2b\,\sigma_{XY} \tag{7.34}$$

Αντικαθιστώντας στην παραπάνω το $b$ από την (7.27) παίρνουμε

$$E[W^2] = \sigma_W^2 = \sigma_Y^2\,(1 - \rho_{XY}^2) \tag{7.35}$$

Τέλος, πολλαπλασιάζοντας τα δύο μέλη της (7.32) με $(X - \mu_X)$ και παίρνοντας στη συνέχεια αναμενόμενες τιμές βρίσκουμε ότι

$$\sigma_{WX} = \sigma_{YX} - b\,\sigma_X^2 \tag{7.36}$$

Αν αντικαταστήσουμε στην παραπάνω εξίσωση το $b$ από την (7.27) βρίσκουμε ότι

$$\sigma_{WX} = 0 \tag{7.37}$$

πράγμα που σημαίνει ότι τα σφάλματα $W$ είναι ασυσχέτιστα με τα δεδομένα $X$. Το συμπέρασμα αυτό επεκτείνεται (Papoulis, 1990, σ. 409-411) για οποιαδήποτε καμπύλη ελάχιστων τετραγώνων (απλή ή πολλαπλή,

γραμμική ή μη γραμμική). Έτσι αποδεικνύεται π.χ. ότι $E[W\, g(X)] = 0$ όπου $g(X)$ είναι οποιαδήποτε συνάρτηση της $X$.

**Ροπές της εκτιμήτριας**

Οι ροπές της μεταβλητής $\hat{Y} = a + bX$, η οποία αποτελεί την εκτιμήτρια της $Y$ για δεδομένη τιμή της $X$, προκύπτουν εύκολα με αξιοποίηση των παραπάνω σχέσεων και είναι

$$\mu_{\hat{Y}} = \mu_Y \tag{7.38}$$

$$\sigma^2_{\hat{Y}} = \frac{\sigma^2_{XY}}{\sigma^2_X} = \rho^2_{XY}\, \sigma^2_Y \tag{7.39}$$

Παρατηρούμε ότι η εκτιμήτρια $\hat{Y}$ διατηρεί τη μέση τιμή (έχει την ίδια μέση τιμή με την $Y$, αλλά όχι και τη διασπορά. Δεδομένου ότι $\rho_{XY} < 1$, η διασπορά της $\hat{Y}$ είναι πάντα μικρότερη από αυτήν της $Y$.

Συνδυάζοντας τις εξισώσεις (7.39) και (7.35) βρίσκουμε ότι

$$\sigma^2_Y = \sigma^2_{\hat{Y}} + \sigma^2_W \tag{7.40}$$

Τα μεγέθη $\sigma^2_Y$, $\sigma^2_{\hat{Y}}$ και $\sigma^2_W$ συχνά αποκαλούνται *ολική*, *παλινδρομική* και *υπόλοιπη* διασπορά, αντίστοιχα.

**Συντελεστής προσδιορισμού**

Ο συντελεστής προσδιορισμού σύμφωνα με την πιθανοτική προσέγγιση γενικά ορίζεται ως

$$\delta = 1 - \frac{E[W^2]}{\sigma^2_Y} \tag{7.41}$$

Ο ορισμός αυτός βρίσκεται σε αντιστοιχία με αυτόν της προσδιοριστικής προσέγγισης (σχέση (7.14)). Πράγματι, αν στην (7.41) αντικατασταθούν οι θεωρητικές ροπές με τις δειγματικές εκτιμήσεις τους, προκύπτει η (7.14). Στην εξεταζόμενη περίπτωση της γραμμικής παλινδρόμησης συμβαίνει να ισχύει

$$\delta = \rho_{XY}^2 = \frac{\sigma_{\hat{Y}}^2}{\sigma_Y^2} \tag{7.42}$$

δηλαδή ο συντελεστής προσδιορισμού είναι ταυτόχρονα ίσος με το τετράγωνο του συντελεστή συσχέτισης των δύο μεταβλητών, καθώς και με το λόγο της παλινδρομικής προς την ολική διασπορά.

Αυτή η ιδιότητα δεν ισχύει για κάθε παλινδρόμηση (βλ. π.χ. επόμενο εδάφιο). Λόγω της τελευταίας ισότητας ο συντελεστής προσδιορισμού, συχνά εκφρασμένος ως ποσοστό στα %, ονομάζεται και *ποσοστό εξηγούμενης διασποράς* του μοντέλου παλινδρόμησης.

**Εφαρμογή 7.1**

*Η λεκάνη απορροής του Αχελώου ανάντη του φράγματος Κρεμαστών έχει ολική έκταση 3584 km$^2$, από τα οποία τα 80.6 km$^2$ (κατά μέγιστο) καταλαμβάνει ο ταμιευτήρας, ωφέλιμης χωρητικότητας 3300 hm$^3$. Στον Πίν. 7.1 δίνονται για την περίοδο 1967-68 μέχρι 1991-92 (25 χρόνια) σε ετήσια βάση (1) η επιφανειακή βροχόπτωση της λεκάνης, P, (2) το ισοδύναμο ύψος επιφανειακής απορροής της λεκάνης, Q, και (3) η εξάτμιση από την επιφάνεια του ταμιευτήρα κατά Penman, E.*

*α. Να διερευνηθεί στατιστικά με βάση τα δείγματα αν συσχετίζονται γραμμικά τα μεγέθη P-Q και E-Q και να προσδιοριστούν οι εξισώσεις εφόσον έχουν νόημα.*

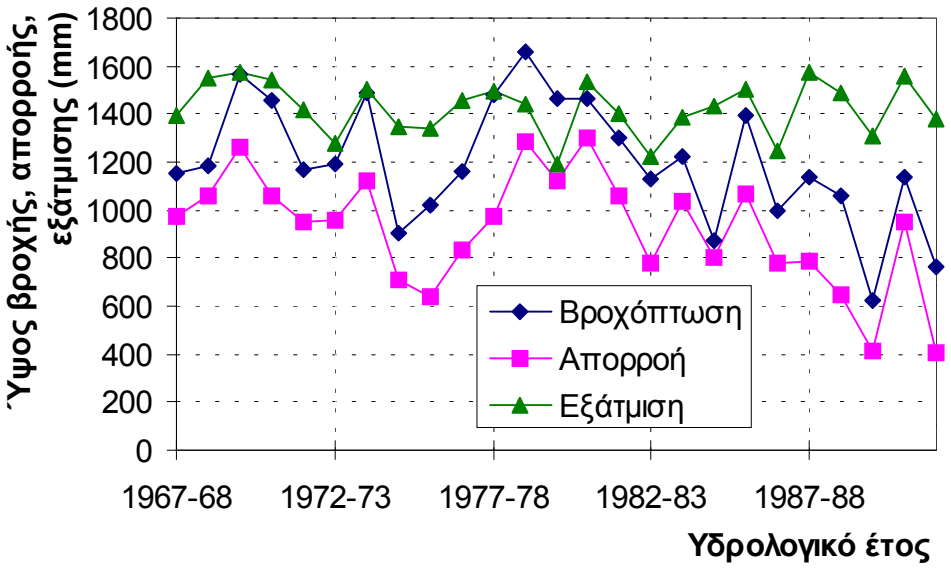
Μια πρώτη εμπειρική εικόνα για την ύπαρξη ή όχι συσχέτισης ανάμεσα στις μεταβλητές *P*, *Q* και *E* παρέχεται στο Σχ. 7.5, όπου έχουν απεικονιστεί οι τρεις χρονοσειρές συναρτήσει του υδρολογικού έτους. Είναι εμφανές ότι οι υπάρχει έντονη συσχέτιση ανάμεσα στη βροχόπτωση και την απορροή της λεκάνης, αλλά όχι ανάμεσα στην εξάτμιση του ταμιευτήρα και την απορροή της λεκάνης. Άλλωστε, αυτό είναι και φυσικά αναμενόμενο και εξηγήσιμο. Σε μια περιοχή σημαντικής υδροφορίας, όπως είναι η βορειοδυτική Ελλάδα, στην οποία ανήκει η λεκάνη του Αχελώου, το μεγαλύτερο τμήμα της βροχόπτωσης απορρέει επιφανειακά, πράγμα που εξηγεί την ισχυρή συσχέτιση ανάμεσα στη βροχόπτωση και την επιφανειακή απορροή. Αντίθετα, δεν περιμένουμε να υπάρχει συσχέτιση ανάμεσα στην απορροή της λεκάνης και στο ύψος εξάτμισης από τον ταμιευτήρα κατά Penman. Το τελευταίο εξαρτάται από μετεωρολογικές μεταβλητές της περιοχής (θερμοκρασία, σχετική υγρασία, ηλιοφάνεια, ταχύτητα ανέμου) και δε συναρτάται με τη βροχόπτωση ή την απορροή.

**Πίν. 7.1** Χρονοσειρές ετήσιου ύψους βροχόπτωσης ($P$) και απορροής ($Q$) της λεκάνης Αχελώου ανάντη των Κρεμαστών, και εξάτμισης ($E$) του ταμιευτήρα Κρεμαστών.

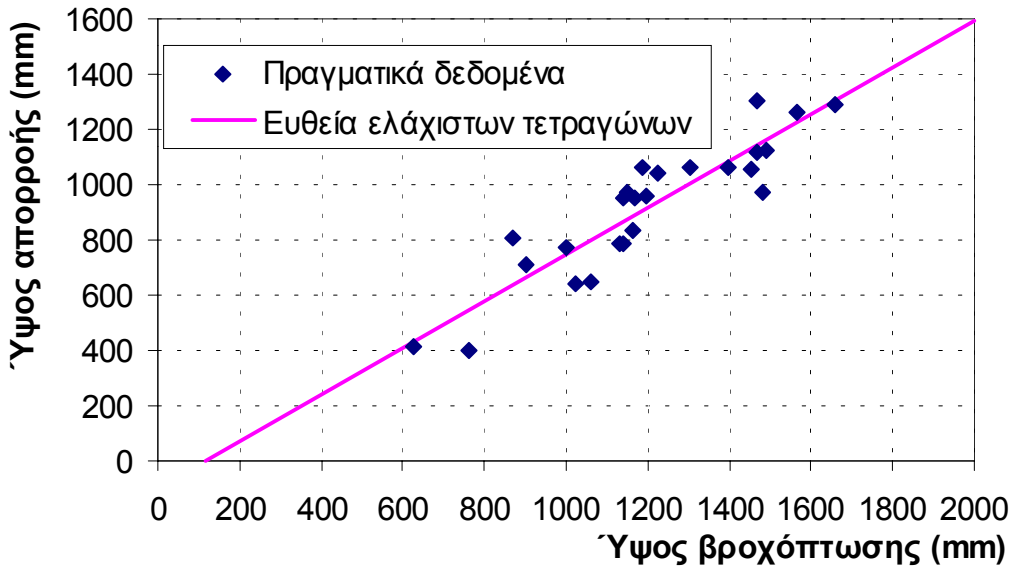
| Υδρολ. έτος | $P$ (mm) | $Q$ (mm) | $E$ (mm) | Υδρολ. έτος | $P$ (mm) | $Q$ (mm) | $E$ (mm) |
|---|---|---|---|---|---|---|---|
| 1967-68 | 1150.6 | 972.8 | 1392.0 | 1980-81 | 1466.3 | 1301.0 | 1535.2 |
| 1968-69 | 1185.3 | 1060.7 | 1549.2 | 1981-82 | 1304.9 | 1063.5 | 1399.9 |
| 1969-70 | 1563.5 | 1260.1 | 1576.2 | 1982-83 | 1131.5 | 782.8 | 1226.3 |
| 1970-71 | 1454.6 | 1057.1 | 1545.5 | 1983-84 | 1223.2 | 1038.1 | 1390.0 |
| 1971-72 | 1165.9 | 948.9 | 1417.3 | 1984-85 | 871.2 | 805.9 | 1434.6 |
| 1972-73 | 1194.6 | 956.4 | 1279.4 | 1985-86 | 1398.2 | 1065.0 | 1501.5 |
| 1973-74 | 1490.7 | 1120.7 | 1504.7 | 1986-87 | 1000.5 | 775.8 | 1245.0 |
| 1974-75 | 901.7 | 710.7 | 1345.9 | 1987-88 | 1141.3 | 787.6 | 1572.7 |
| 1975-76 | 1021.9 | 641.9 | 1342.7 | 1988-89 | 1058.6 | 647.8 | 1488.0 |
| 1976-77 | 1162.9 | 833.9 | 1460.2 | 1989-90 | 626.4 | 411.7 | 1305.7 |
| 1977-78 | 1482.9 | 972.2 | 1492.4 | 1990-91 | 1139.3 | 950.7 | 1560.2 |
| 1978-79 | 1661.2 | 1286.6 | 1440.1 | 1991-92 | 761.3 | 403.2 | 1382.8 |
| 1979-80 | 1467.0 | 1118.5 | 1190.9 | | | | |

Στα ίδια συμπεράσματα καταλήγουμε μελετώντας τα Σχ. 7.6 και Σχ. 7.7. Συγκεκριμένα, στο Σχ. 7.6 έχουμε απεικονίσει το ύψος επιφανειακής απορροής συναρτήσει του ύψους βροχόπτωσης. Είναι εμφανής η συσχέτιση ανάμεσα στα δύο μεγέθη, η οποία μάλιστα φαίνεται να είναι γραμμική, κάτι που είναι επίσης αναμενόμενο και εξηγήσιμο, δεδομένου ότι στις περιοχές μεγάλης υδροφορίας η ετήσια βροχή και απορροή ακολουθούν από κοινού κανονική κατανομή (βλ. ενότητα 7.4). Αντίστοιχα, στο Σχ. 7.7 έχουμε απεικονίσει το ύψος επιφανειακής απορροής συναρτήσει του ύψους εξάτμισης από τον ταμιευτήρα. Εδώ είναι εμφανής η ανυπαρξία συσχέτισης, γραμμικής ή άλλης, ανάμεσα στα δύο μεγέθη.

Βεβαίως οι παραπάνω παρατηρήσεις είναι ποιοτικού χαρακτήρα, αλλά μπορούν εύκολα να ποσοτικοποιηθούν με τη χρήση στατιστικής. Η στατιστική παράμετρος που ενδιαφέρει είναι ο συντελεστής συσχέτισης (ή ο συντελεστής προσδιορισμού της γραμμικής παλινδρόμησης). Σύμφωνα με όσα έχουν αναφερθεί στο εδάφιο 3.5.3 (εξίσωση 3.74), για να είναι στατιστικά σημαντικός (διάφορος του μηδενός) ο συντελεστής συσχέτισης θα πρέπει να είναι μεγαλύτερος του $r_c = 2/\sqrt{n}$ , όπου $n$ το μέγεθος του δείγματος. Στην προκειμένη περίπτωση $r_c = 2/\sqrt{25} = 0.4$.

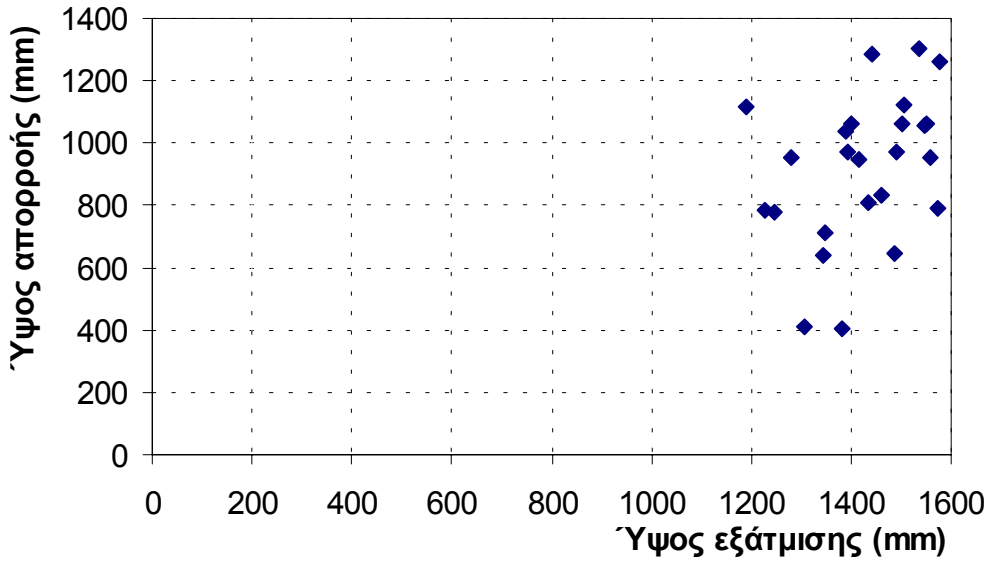



**Σχ. 7.5** Απεικόνιση των χρονοσειρών ετήσιου ύψους βροχόπτωσης και απορροής της λεκάνης Αχελώου ανάντη των Κρεμαστών, και της εξάτμισης του ταμιευτήρα Κρεμαστών.




**Σχ. 7.6** Ετήσιο ύψος επιφανειακής απορροής συναρτήσει του ετήσιου ύψους βροχόπτωσης στη λεκάνη ανάντη Κρεμαστών.

Η εκτίμηση του δειγματικού συντελεστή συσχέτισης δίνεται από την εξίσωση (7.18). Αν και ο υπολογισμός του γίνεται αυτόματα από υπολογιστές ή αριθμομηχανές, εδώ εκθέτουμε για διδακτικούς λόγους τον υπολογισμό του με το χέρι. Στους παρακάτω συμβολισμούς των αθροισμάτων παραλείπουμε για ευκολία τους δείκτες $i$ στα δείγματα $x_i\, y_i$.




**Σχ. 7.7** Ετήσιο ύψος επιφανειακής απορροής στη λεκάνη ανάντη Κρεμαστών συναρτήσει του ετήσιου ύψους εξάτμισης από τον ταμιευτήρα Κρεμαστών.

Για τις μεταβλητές $P$ ($\equiv X$) και $Q$ ($\equiv Y$) έχουμε:

$$n = 25,\ \sum x = 30\,025.5,\ \sum y = 22\,973.6,\ \sum xy = 28\,941\,151,$$

$$\sum x^2 = 37\,655\,556,\ \sum y^2 = 22\,487\,374$$

Κατά συνέπεια από την εξίσωση (7.18) βρίσκουμε

$$r = \frac{25 \times 28\,941\,151 - 30\,025.5 \times 22\,973.6}{\sqrt{(25 \times 37\,655\,556 - 30\,025.5^2)(25 \times 22\,487\,374 - 22\,973.6^2)}}$$

απ' όπου τελικά προκύπτει $r = 0.911$. Η τιμή αυτή είναι μεγαλύτερη από την κρίσιμη τιμή $r_c = 0.4$, πράγμα που μας οδηγεί στην απόρριψη της υπόθεσης ότι ο συντελεστής γραμμικής συσχέτισης ανάμεσα στις δύο μεταβλητές είναι μηδενικός. Επιβεβαιώνεται δηλαδή και στατιστικά η ύπαρξη γραμμικής συσχέτισης ανάμεσα στη βροχόπτωση και την απορροή της λεκάνης.

Αντίστοιχα, για τις μεταβλητές $E$ ($\equiv X$) και $R$ ($\equiv Y$) έχουμε:

$$n = 25,\ \sum x = 35\,578.4,\ \sum y = 22\,973.6,\ \sum xy = 32\,935\,017,$$

$$\sum x^2 = 50\ 940\ 067, \sum y^2 = 22\ 487\ 374$$

Κατά συνέπεια από την εξίσωση (7.18) βρίσκουμε, όπως παραπάνω, ότι $r = 0.370$. Η τιμή αυτή είναι μικρότερη από την κρίσιμη τιμή $r_c = 0.4$, πράγμα που μας οδηγεί στην μη απόρριψη της υπόθεσης ότι ο συντελεστής γραμμικής συσχέτισης ανάμεσα στις δύο μεταβλητές είναι μηδενικός.

Για την περίπτωση των μεταβλητών $P$ (≡ $X$) και $Q$ (≡ $Y$), όπου όπως είδαμε έχει νόημα η γραμμική συσχέτιση, υπολογίζουμε τους συντελεστές $b$ και $a$ της γραμμικής παλινδρόμησης, χρησιμοποιώντας τις εξισώσεις (7.10) και (7.11), αντίστοιχα. Έτσι, από την (7.10) παίρνουμε

$$b = \frac{25 \times 28\ 941\ 151 - 30\ 025.5 \times 22\ 973.6}{25 \times 37\ 655\ 556 - 30\ 025.5^2} = 0.846$$

και από την (7.11)

$$a = \frac{22\ 973.6}{25} - 0.846 \times \frac{30\ 025.5}{25} = -97.1$$

Η ευθεία με εξίσωση $y = -97.1 + 0.846\ x$ έχει παρασταθεί γραφικά στο Σχ. 7.6.

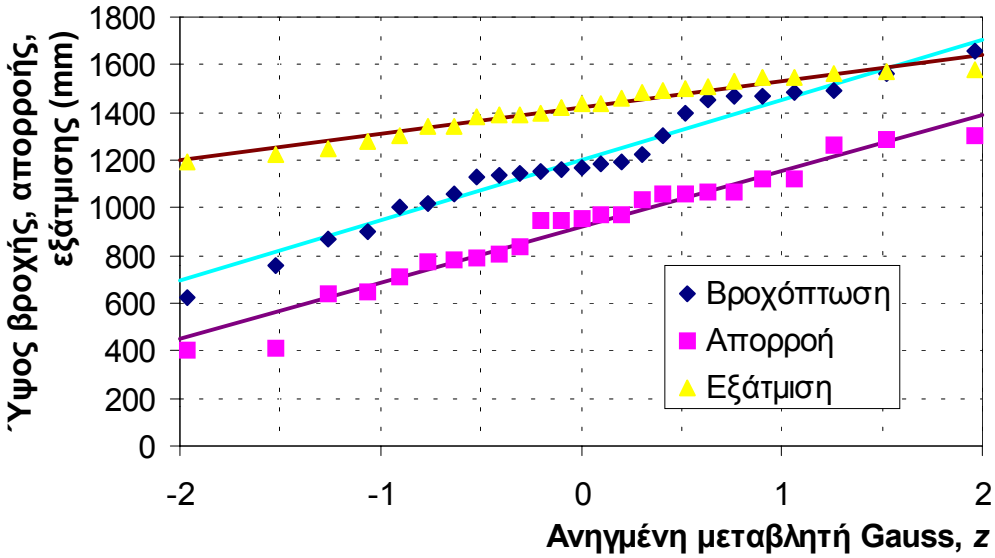
*β. Να υπολογιστεί η περίοδος επαναφοράς των τιμών των τριών μεταβλητών για το υδρολογικό έτος 1978-79, θεωρητικά και εμπειρικά. Αν υποτεθεί ότι έλειπε η τιμή της απορροής του έτους αυτού, ποια θα ήταν η εκτίμησή του και σε ποια περίοδο επαναφοράς θα αντιστοιχούσε;*

Για να μπορέσουμε να υπολογίσουμε θεωρητικές πιθανότητες υπέρβασης είναι απαραίτητο να υποθέσουμε μια συγκεκριμένη συνάρτηση κατανομής σε καθεμιά από τις μεταβλητές. Ήδη έχουμε υπαινιχθεί ότι η βροχόπτωση και η απορροή ακολουθούν κανονικές κατανομές. Το ίδιο θα πρέπει να υποθέσουμε και για την εξάτμιση από τον ταμιευτήρα, δεδομένου ότι αναφερόμαστε σε ετήσια χρονική κλίμακα. Η προσαρμογή (εκτίμηση παραμέτρων) και ο έλεγχος (με τη δοκιμή $\chi^2$) των κανονικών συναρτήσεων κατανομής γίνεται σύμφωνα με όσα έχουμε αναπτύξει στο κεφάλαιο 5. Ειδικά για τη δοκιμή $\chi^2$ επιλέγουμε 5 κλάσεις, οπότε η κρίσιμη τιμή του στατιστικού ελέγχου για επίπεδο σημαντικότητας 5% και για $5 - 2 - 1 = 2$ βαθμούς ελευθερίας είναι $\chi^2_{0.95}(2) = 5.99$. Τα αποτελέσματα της διαδικασίας προσαρμογής και ελέγχου δίνονται συνοπτικά στον Πίν. 7.2. Παρατηρούμε ότι για καμία από τις τρεις μεταβλητές δεν απορρίπτεται η υπόθεση ότι ακολουθεί κανονική κατανομή. Στο Σχ. 7.8 φαίνεται και γραφικά η

προσαρμογή των κανονικών συναρτήσεων κατανομής στις αντίστοιχες εμπειρικές.

**Πίν. 7.2** Παράμετροι κανονικής κατανομής και αποτελέσματα της δοκιμής $\chi^2$ για τις τρεις μεταβλητές της Εφαρμογής 7.1.

| | $\mu$ | $\sigma$ | $\chi^2$ |
|---|---|---|---|
| Βροχόπτωση | 1201.0 | 252.5 | 4.8 |
| Απορροή | 918.9 | 234.6 | 0.0 |
| Εξάτμιση | 1423.1 | 110.8 | 0.4 |




**Σχ. 7.8** Εμπειρικές και θεωρητικές (Gauss) συναρτήσεις κατανομής του ετήσιου ύψους βροχόπτωσης και απορροής της λεκάνης Αχελώου ανάντη των Κρεμαστών, και της εξάτμισης του ταμιευτήρα Κρεμαστών.

Το υδρολογικό έτος 1978-79 η τιμή της βροχόπτωσης είναι 1661.2 mm, και είναι η πρώτη σε μέγεθος τιμή του δείγματος των 25 ετών. Κατά συνέπεια η εμπειρική περίοδος επαναφοράς της, σύμφωνα με τη σχέση Blom (βλ. Πίν. 5.7), είναι

$$T = (25 + 0.25) / (1 - 0.375) = 40.4$$

Η ανηγμένη μεταβλητή είναι

$$z = (1661.2 - 1201.0) / 252.5 = 1.823$$

που, για την υιοθετημένη κανονική κατανομή, αντιστοιχεί σε $F = 0.9658$. Άρα, η θεωρητική περίοδος επαναφοράς είναι

$$T = 1 /(1 - 0.9658) = 29.2.$$

Για το ίδιο υδρολογικό έτος η τιμή της απορροής είναι 1286.6 mm, και είναι η δεύτερη σε μέγεθος τιμή του δείγματος των 25 ετών. Κατά συνέπεια η εμπειρική περίοδος επαναφοράς της είναι

$$T = (25 + 0.25) / (2 - 0.375) = 15.5$$

Η ανηγμένη μεταβλητή είναι

$$z = (1286.6 - 918.9) / 234.6 = 1.567$$

που αντιστοιχεί σε $F = 0.9415$. Άρα, η θεωρητική περίοδος επαναφοράς είναι

$$T = 1 / (1 - 0.9415) = 17.1.$$

Τέλος, για το ίδιο υδρολογικό έτος η τιμή της εξάτμισης είναι 1440.1 mm, και είναι η δωδέκατη σε μέγεθος τιμή του δείγματος των 25 ετών. Κατά συνέπεια η εμπειρική περίοδος επαναφοράς της είναι

$$T = (25 + 0.25) / (12 - 0.375) = 2.17$$

Η ανηγμένη μεταβλητή είναι

$$z = (1440.1 - 1426.1) / 110.8 = 0.153$$

που αντιστοιχεί σε $F = 0.5608$. Άρα, η θεωρητική περίοδος επαναφοράς είναι

$$T = 1 / (1 - 0.5608) = 2.28$$

Αν έλειπε η τιμή της απορροής για το υδρολογικό έτος 1978-79, τότε θα την εκτιμούσαμε από την αντίστοιχη τιμή της βροχόπτωσης, δεδομένου ότι, όπως είδαμε, υπάρχει σημαντική γραμμική συσχέτιση ανάμεσα στα δύο μεγέθη. Βέβαια, στην περίπτωση αυτή οι συντελεστές $b$ και $a$ της γραμμικής εξίσωσης είναι ελαφρά διαφοροποιημένοι, επειδή υπολογίζονται από τα δείγματα των 24 και όχι των 25 ετών. Επαναλαμβάνοντας την ίδια διαδικασία, όπως παραπάνω, βρίσκουμε ότι η εξίσωση της παλινδρόμησης είναι, στην περίπτωση των 24 ετών, $y = -105.7 + 0.854\,x$, οπότε, για $x = 1661.2$ βρίσκουμε

$$y = -105.7 + 0.854 \times 1661.2 = 1313.0 \text{ mm}.$$

Στην τελευταία περίπτωση η εκτιμημένη τιμή του έτους 1978-79 είναι η μεγαλύτερη στο δείγμα των 25 ετών και επομένως η εμπειρική περίοδος επαναφοράς της θα είναι $T = 40.4$. Για τη θεωρητική περίοδο επαναφοράς χρησιμοποιούμε τη μέση τιμή και τυπική απόκλιση των 24 ετών που είναι $\mu = 903.6$ mm και $\sigma = 226.9$ mm. Έτσι, η ανηγμένη μεταβλητή είναι

$$z = (1313.0 - 903.6) / 226.9 = 1.804$$

οπότε $F = 0.9644$ και $T = 28.1$. Παρατηρούμε λοιπόν ότι η εμπειρική περίοδος επαναφοράς της εκτιμημένης τιμής της απορροής ταυτίζεται με αυτήν της βροχής (πρόκειται βέβαια για σύμπτωση, παρά για γενικό κανόνα), ενώ είναι η θεωρητική είναι ελαφρά μικρότερη από την αντίστοιχη της βροχής. Και τα δύο μεγέθη είναι αισθητά μεγαλύτερα από αυτά που υπολογίσαμε για την πραγματική τιμή της απορροής του έτους 1978-79 (1286.6 mm έναντι της εκτιμημένης τιμής 1313.0 mm).

*γ. Να υπολογιστεί η μεταβολή αποθέματος του ταμιευτήρα για ένα έτος στο οποίο (1) πραγματοποιείται ελάχιστη βροχόπτωση περιόδου επαναφοράς 25 ετών· (2) η απορροή και η εξάτμιση έχουν τις τιμές που στατιστικά αναμένονται για την εν λόγω βροχόπτωση· (3) οι ολικές απολήψεις από τον ταμιευτήρα είναι 80% του μέσου ετήσιου όγκου απορροής· (4) η μέση επιφάνεια του ταμιευτήρα είναι ίση με 65% της μέγιστης· και (5) δεν πραγματοποιούνται υπερχειλίσεις από τον ταμιευτήρα.*

Η ελάχιστη βροχόπτωση περιόδου επαναφοράς 25 ετών είναι $1201.0 - 1.7507 \times 252.5 = 758.9$ mm (όπου 1.7507 είναι η τιμή της ανηγμένης κανονικής μεταβλητής $z$ για $F = 1 / T = 1/25 = 0.04$). Αν και η τιμή αυτή αναφέρεται στο σύνολο της λεκάνης, θα θεωρήσουμε ότι την ίδια τιμή έχει και η βροχόπτωση πάνω από τον ταμιευτήρα. Η στατιστικά αναμενόμενη τιμή της απορροής για αυτή την τιμή της βροχής προκύπτει από την εξίσωση παλινδρόμησης του ερωτήματος α. και είναι

$$y = -97.1 + 0.846 \times 758.9 = 544.9 \text{ mm}$$

Η στατιστικά αναμενόμενη τιμή της εξάτμισης είναι ίση με τη μέση τιμή της, δεδομένου ότι δεν υπάρχει συσχέτιση βροχόπτωσης και εξάτμισης, δηλαδή είναι 1423.1 mm.
Η μέση έκταση της επιφάνειας του ταμιευτήρα κατά το εν λόγω έτος είναι

$$A_T = 0.65 \times 80.6 = 52.4 \text{ km}^2$$

Επομένως η έκταση της λεκάνης που δεν καταλαμβάνεται από τον ταμιευτήρα είναι

$$A_\Lambda = 3584 - 52.4 = 3531.6 \text{ km}^2.$$

Κατά συνέπεια ο όγκος απορροής είναι

$$V_Q = 0.5449 \times 3531.6 \times 10^6 = 1924.4 \times 10^6 \text{ m}^3 = 1924.4 \text{ hm3}$$

ο όγκος βροχόπτωσης

$$V_P = 0.7589 \times 52.4 \times 10^6 = 39.8 \times 10^6 \text{ m}^3 = 39.8 \text{ hm}^3$$

και ο όγκος εξάτμισης

$$V_E = 1.4231 \times 52.4 \times 10^6 = 74.6 \times 10^6 \text{ m}^3 = 74.6 \text{ hm}^3$$

Ο μέσος ετήσιος όγκος απορροής είναι

$$E[V_Q] \approx 0.9189 \times 3531.6 \times 10^6 = 3245.2 \times 10^6 \text{ m}^3 = 3245.2 \text{ hm}^3$$

και κατά συνέπεια η απόληψη του υπόψη έτους είναι

$$V_D = 0.80 \times 3245.2 \text{ hm}^3 = 2596.1 \text{ hm}^3$$

Κατά συνέπεια, η μεταβολή του αποθέματος του ταμιευτήρα είναι

$$\Delta S = V_Q + V_P - V_E - V_D =$$

$$= 1924.4 + 39.8 - 74.6 - 2596.1 = -706.5 \text{ hm}^3$$

## 7.2 Άλλες γραμμικές εκτιμήσεις δύο μεταβλητών

### 7.2.1 Ομογενής ευθεία

Σε πολλές περιπτώσεις είναι επιθυμητή η αγνόηση του σταθερού όρου $a$ στην εξίσωση της ευθείας, οπότε οι εξισώσεις (7.24) και (7.25) παίρνουν τη μορφή

$$Y = bX + W \tag{7.43}$$

$$\hat{Y} = bX \tag{7.44}$$

Η τελευταία εξίσωση λέγεται *ομογενής ευθεία*. Η παράμετρος $b$ υπολογίζεται και πάλι έτσι ώστε η (7.44) να αποτελεί την εκτίμηση ελάχιστου μέσου τετραγωνικού σφάλματος, με την έννοια ότι ελαχιστοποιείται το μέγεθος

$$E[W^2] = E[(Y - \hat{Y})^2] = E[(Y - bX)^2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - bx)^2 f_{XY}(x, y)\, dx\, dy \tag{7.45}$$

Για την ελαχιστοποίηση παραγωγίζουμε την παραπάνω ως προς $b$ και εξισώνουμε την παράγωγο με 0, οπότε διαδοχικά βρίσκουμε

$$\frac{\partial E[W^2]}{\partial b} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} -2(y - bx)\, x\, f_{XY}(x, y)\, dx\, dy = 0 \tag{7.46}$$

$$E[(Y - bX)X] = E[XY] - bE[X^2] = 0 \tag{7.47}$$

απ' όπου προκύπτει

$$b = \frac{E[XY]}{E[X^2]} = \frac{\sigma_{XY} + \mu_X\mu_Y}{\sigma_X^2 + \mu_X^2} \tag{7.48}$$

όπου $\mu_X$ και $\mu_Y$ οι μέσες τιμές των $X$ και $Y$, αντίστοιχα, $\sigma_X$ και $\sigma_Y$ οι αντίστοιχες τυπικές αποκλίσεις και $\sigma_{XY}$ και $\rho_{XY}$ η συνδιασπορά και ο συντελεστής συσχέτισης, αντίστοιχα, των $X$ και $Y$. Τα μεγέθη αυτά συνήθως είναι άγνωστα (παράμετροι πληθυσμού). Αν στη θέση των $E[XY]$ και $E[X^2]$ τοποθετήσουμε τις (αμερόληπτες) εκτιμήσεις τους από το δείγμα των $n$ σημείων $(x_i, y_i)$, τότε παίρνουμε την τελική έκφραση

$$b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{7.49}$$

η οποία, όπως παρατηρούμε είναι πολύ απλούστερη από την αντίστοιχη εξίσωση (7.10) της πλήρους γραμμικής μορφής.

Εύκολα μπορεί να αποδειχτεί ότι η ομογενής μορφή της εκτιμήτριας $\hat{Y}_2$ δε διατηρεί ούτε τη μέση τιμή ούτε τη διασπορά, δηλαδή $\mu_{\hat{Y}} \neq \mu_Y$ και $\sigma^2_{\hat{Y}} \neq \sigma^2_Y$. Μάλιστα η διασπορά της $\hat{Y}$ δεν είναι πάντα μικρότερη από αυτήν της $Y$, αλλά μπορεί να είναι και μεγαλύτερη. Ως συνέπεια της μη διατήρησης της μέσης τιμής της $Y$, το μέσο σφάλμα δεν είναι μηδενικό, δηλαδή $\mu_W \neq 0$ και επομένως $\sigma^2_W \neq E[W^2]$. Ωστόσο ισχύει $E[WX] = 0$.

Ο συντελεστής προσδιορισμού στην περίπτωση αυτή ορίζεται από τη γενική σχέση (7.41). Εύκολα μπορεί να αποδειχτεί ότι στην περίπτωση της ομογενούς ευθείας

$$\delta \neq \rho_{XY}^2 \neq \frac{\sigma_{\hat{Y}}^2}{\sigma_Y^2} \tag{7.50}$$

Ακόμη, επειδή είναι δυνατό να ισχύει $\sigma_Y^2 < \sigma_W^2$, ο συντελεστής προσδιορισμού είναι δυνατό να πάρει και αρνητικές τιμές. Γενικά πάντως ισχύει $\delta \leq 1$ και η τιμή του $\delta$ κοντά στο 1 δείχνει καλή προσαρμογή της ομογενούς ευθείας στα δεδομένα.

Με μια εναλλακτική μέθοδο εκτίμησης της παραμέτρου $b$ της ομογενούς ευθείας είναι δυνατό να διατηρηθεί η μέση τιμή της $Y$. Πράγματι, παίρνοντας αναμενόμενες τιμές στην (7.44) και θεωρώντας $\mu_{\hat{Y}} = \mu_Y$, βρίσκουμε

$$b = \frac{E[Y]}{E[X]} = \frac{\mu_Y}{\mu_X} \tag{7.51}$$

Είναι αυτονόητο ότι το τετραγωνικό σφάλμα $E[W^2]$ που προκύπτει με αυτό τον τρόπο εκτίμησης της παραμέτρου $b$ δεν είναι το ελάχιστο δυνατό.

### 7.2.2 Οργανική συσχέτιση

Όπως είδαμε στα προηγούμενα εδάφια, τόσο η πλήρης όσο και η ομογενής γραμμική παλινδρόμηση αποτυγχάνουν στη διατήρηση της διασποράς της εξαρτημένης μεταβλητής $Y$ (ειδικά η ομογενής μπορεί να μη διατηρεί ούτε τη μέση τιμή). Ωστόσο, στην τεχνική υδρολογία είναι συχνά επιθυμητή η διατήρηση της διασποράς, ιδίως στην περίπτωση επέκτασης δειγμάτων (βλ. ενότητα 7.6). Αν εγκαταλείψουμε την απαίτηση του ελάχιστου τετραγωνικού σφάλματος, τότε είναι δυνατό να εκτιμήσουμε τις παραμέτρους της (7.25) σε τρόπο ώστε να διατηρείται και η μέση τιμή και η διασπορά. Αυτός ο τρόπος εκτίμησης των παραμέτρων είναι γνωστός ως *οργανική συσχέτιση* ή και ως *επέκταση διατήρησης διασποράς* (maintenance of variance extension - MOVE.1).

Συγκεκριμένα, η (7.25) σε συνδυασμό με τις απαιτήσεις $\mu_{\hat{Y}} = \mu_Y$ και $\sigma^2_{\hat{Y}} = \sigma^2_Y$ δίνει

$$\mu_Y = a + b\,\mu_X \tag{7.52}$$

$$\sigma^2_Y = b^2\,\sigma^2_X \tag{7.53}$$

απ' όπου προκύπτει

$$b = \mathrm{sgn}(\rho_{XY})\,\frac{\sigma_Y}{\sigma_X} \tag{7.54}$$

$$a = \mu_Y - b\,\mu_X \tag{7.55}$$

όπου $\mathrm{sgn}(\rho_{XY})$ είναι το πρόσημο του συντελεστή συσχέτισης (+1 ή –1). Ο όρος προσήμου έχει τεθεί για να είναι συνεπής η εκτίμηση με την πραγματικότητα, δηλαδή για θετικά συσχετισμένες μεταβλητές να προκύπτει θετική τιμή του $b$ και αντίστροφα (η (7.53) επιτρέπει και τα δύο πρόσημα). Παρατηρούμε ότι η (7.55) ταυτίζεται με την αντίστοιχή της (7.28), ενώ η (7.54) διαφέρει από την (7.27) κατά το ότι το $\rho_{XY}$ έχει αντικατασταθεί από το πρόσημό του. Αν στη θέση των θεωρητικών ροπών του πληθυσμού τοποθετήσουμε τις εκτιμήσεις τους από το δείγμα των $n$ σημείων $(x_i, y_i)$, τότε παίρνουμε

$$b = \mathrm{sgn}(r_{XY})\,\frac{s_Y}{s_X} \tag{7.56}$$

$$a = \overline{y} - b\,\overline{x} \tag{7.57}$$

Οι ροπές του σφάλματος $W = Y - \hat{Y} = Y - (a + bX)$ προκύπτουν με ανάλογο τρόπο όπως προηγουμένως και είναι

$$\mu_W = 0 \tag{7.58}$$

$$\sigma^2_W = 2\,\sigma^2_Y\,(1 - |\rho_{XY}|) \tag{7.59}$$

Όπως αναμέναμε, το μέσο τετραγωνικό σφάλμα είναι γενικά μεγαλύτερο από αυτό της τυπικής γραμμικής παλινδρόμησης. Η μέγιστη τιμή του, ίση με $2\,\sigma^2_Y$, εμφανίζεται για $\rho_{XY} = 0$. Η τιμή αυτή είναι διπλάσια από την

αντίστοιχη της τυπικής γραμμικής παλινδρόμησης. Η ελάχιστη τιμή του, ίση με 0 όπως και στην τυπική γραμμική παλινδρόμηση, εμφανίζεται όταν $|\rho_{XY}| = 1$.

Το σφάλμα $W$ στην περίπτωση της οργανικής παλινδρόμησης δεν είναι ασυσχέτιστο με τη μεταβλητή $X$. Η αντίστοιχη συνδιασπορά είναι

$$\sigma_{WX} = -\text{sgn}(\rho_{XY})\,(1 - |\rho_{XY}|)\,\sigma_X \sigma_Y \tag{7.60}$$

Ο συντελεστής προσδιορισμού είναι

$$\delta = 2|\rho_{XY}| - 1 \tag{7.61}$$

Παρατηρούμε ότι ο συντελεστής προσδιορισμού γίνεται ίσος με 1 για $|\rho_{XY}| = 1$, μηδενίζεται για $|\rho_{XY}| = 0.5$ και παίρνει αρνητικές τιμές για $|\rho_{XY}| < 0.5$. Κατά συνέπεια δεν έχει πρακτικό νόημα η εφαρμογή της οργανικής συσχέτισης για $|\rho_{XY}| < 0.5$.

Η εξίσωση (7.25), αν αντικαταστήσουμε τις τιμές των $a$ και $b$ από τις (7.55) και (7.54), παίρνει την ακόλουθη μορφή (για $\rho_{XY} > 0$)

$$\frac{\hat{Y} - \mu_Y}{\sigma_Y} = \frac{X - \mu_X}{\sigma_X} \tag{7.62}$$

Σε περίπτωση που η $X$ ακολουθεί κανονική κατανομή, την ίδια κατανομή θα ακολουθεί και η $\hat{Y}$, οπότε η παραπάνω εξίσωση ισοδυναμεί με την

$$F_{\hat{Y}}(\hat{y}) = F_X(x), \tag{7.63}$$

η οποία σημαίνει ότι η εκτίμηση $\hat{Y}$ είναι ισοπίθανη με την αντίστοιχη τιμή της $X$.

Τέλος, στην περίπτωση της οργανικής συσχέτισης η αντίστροφη εξίσωση που εκφράζει την εκτίμηση της $\hat{X}$ για δεδομένη $Y$ δεν είναι άλλη από αυτή που προκύπτει με τη συνήθη αλγεβρική αντιστροφή, δηλαδή η

$$\hat{X} = (Y - a) / b \tag{7.64}$$

Δηλαδή, εδώ δεν έχουμε δύο διαφορετικές ευθείες όπως στην τυπική γραμμική παλινδρόμηση.

## Εφαρμογή 7.2

*α. Να βρεθούν οι εξισώσεις ανάμεσα στις μεταβλητές P και Q της Εφαρμογής 7.1, χρησιμοποιώντας την ομογενή ευθεία και την οργανική συσχέτιση.*

Υπενθυμίζουμε ότι τα χαρακτηριστικά αθροίσματα των δειγμάτων $X$ ($\equiv P$) και $Y$ ($\equiv Q$) της Εφαρμογής 7.1 είναι:

$$n = 25,\ \sum x = 30\ 025.5,\ \sum y = 22\ 973.6,\ \sum xy = 28\ 941\ 151,$$

$$\sum x^2 = 37\ 655\ 556,\ \sum y^2 = 22\ 487\ 374$$

Τα στατιστικά χαρακτηριστικά των μεταβλητών είναι

$$\overline{x} = 1201.0,\ \overline{y} = 918.9,\ s_X = 252.5,\ s_Y = 234.6,\ r_{XY} = 0.911,$$

$$s_{XY} = r_{XY}\, s_X\, s_Y = 0.911 \times 252.5 \times 234.6 = 53\ 964$$

Εφαρμόζοντας την εξίσωση (7.49) υπολογίζουμε ότι ο συντελεστής της ομογενούς εξίσωσης ελάχιστου τετραγωνικού σφάλματος είναι

$$b = \sum xy \,/\, \sum x^2 = 28\ 941\ 151 \,/\, 37\ 655\ 556 = 0.769$$

(έναντι 0.846 της πλήρους ευθείας). Εναλλακτικά, ο συντελεστής που διατηρεί τη μέση τιμή στην ομογενή ευθεία δίνεται από την (7.51) και είναι

$$b = \overline{y} \,/\, \overline{x} = 918.9 \,/\, 1201.0 = 0.765$$

Οι συντελεστές $b$ και $a$ της οργανικής συσχέτισης δίνονται από τις εξισώσεις (7.54) και (7.55), και είναι

$$b = +s_Y \,/\, s_X = 234.6 \,/\, 252.5 = 0.929$$

και

$$a = \overline{y} - b\,\overline{x} = 918.9 - 0.929 \times 1201.0 = -196.8$$

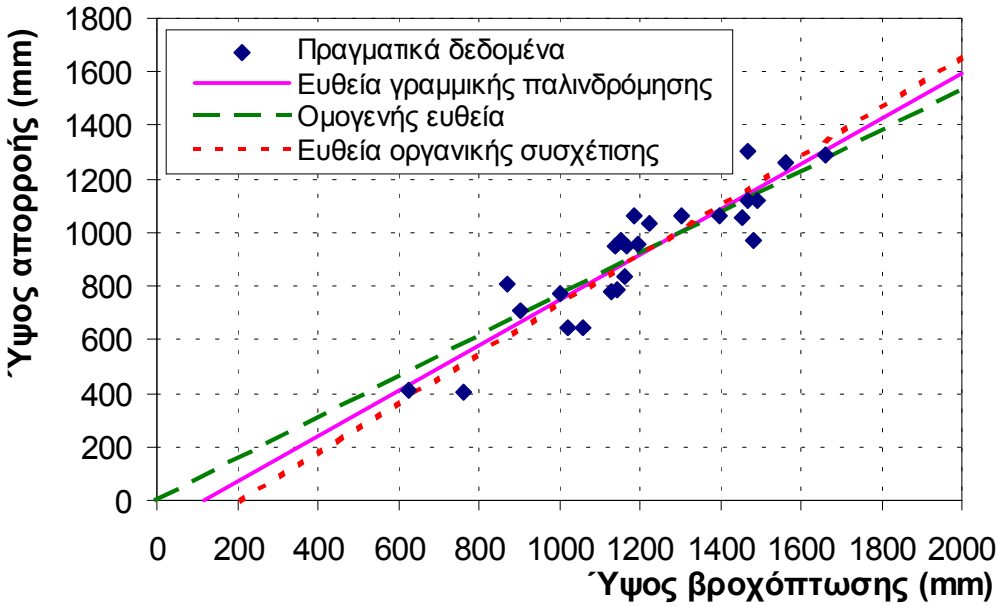
Οι διάφορες εξισώσεις που έχουν προσδιοριστεί παραπάνω καθώς και στην Εφαρμογή 7.1 φαίνονται στο Σχ. 7.9.

*β. Να επανεξεταστεί το ερώτημα β. της Εφαρμογής 7.1, χρησιμοποιώντας την οργανική συσχέτιση.*

Αν δε συμπεριλάβουμε το έτος 1978-79 στα δείγματα της βροχής και απορροής τότε τα στατιστικά χαρακτηριστικά των δειγμάτων γίνονται

$$\overline{x} = 1181.8,\ \overline{y} = 903.6,\ s_X = 239.2,\ s_Y = 226.9$$

και η εξίσωση οργανικής συσχέτισης γίνεται $y = -217.0 + 0.948\ x$. Εφαρμόζοντας αυτή την εξίσωση για $x = 1661.2$ βρίσκουμε $y = 1358.2$ (έναντι 1313.0 που έχει δώσει η τυπική γραμμική παλινδρόμηση).




**Σχ. 7.9** Ετήσιο ύψος επιφανειακής απορροής συναρτήσει του ετήσιου ύψους βροχόπτωσης στη λεκάνη ανάντη Κρεμαστών: Πραγματικά δεδομένα και εναλλακτικές γραμμικές εξισώσεις.

Με τα παραπάνω στατιστικά χαρακτηριστικά των 24 ετών η ανηγμένη μεταβλητή για την απορροή είναι

$$z = (1358.2 - 903.6) / 226.9 = 2.00$$

που αντιστοιχεί σε $F = 0.9774$ και $T = 1 / (1 - 0.9774) = 44.3$. Την ίδια τιμή της περιόδου επαναφοράς βρίσκουμε και για τη βροχόπτωση, αν χρησιμοποιήσουμε τα στατιστικά χαρακτηριστικά των 25 ετών, δεδομένου ότι

$$z = (1661.2 - 1181.8) / 239.2 = 2.00$$

Αν χρησιμοποιήσουμε τα στατιστικά χαρακτηριστικά των 25 ετών, που για τη βροχόπτωση είναι

$$\overline{x} = 1201.0,\ s_X = 252.5$$

ενώ για την απορροή γίνονται (με την τιμή 1358.2 για το έτος 1978-79)

$$\overline{y} = 921.8,\ s_Y = 239.5$$

η ανηγμένη μεταβλητή για την απορροή είναι

$$z = (1358.2 - 921.8) / 239.5 = 1.82$$

που αντιστοιχεί σε $F = 0.9657$ και $T = 1 / (1 - 0.9657) = 29.2$. Την ίδια τιμή της περιόδου επαναφοράς βρίσκουμε και για τη βροχόπτωση, δεδομένου ότι

$$z = (1661.2 - 1201.0) / 252.5 = 29.2$$

(βλ. και Εφαρμογή 7.1). Έτσι, επιβεβαιώνουμε και αριθμητικά την ταύτιση των περιόδων επαναφοράς του αρχικού και εκτιμημένου μεγέθους, στην περίπτωση που χρησιμοποιείται η οργανική συσχέτιση και η κατανομή των δύο μεταβλητών είναι κανονική. Βεβαίως αυτό δεν σημαίνει ότι η οργανική συσχέτιση είναι πιο κοντά στην πραγματικότητα από την τυπική γραμμική παλινδρόμηση. Αντίθετα, στο παράδειγμα που εξετάσαμε η εκτίμηση της οργανικής συσχέτισης διαφέρει περισσότερο από την πραγματική τιμή της απορροής του έτους 1978-79, απ' ότι διαφέρει η αντίστοιχη εκτίμηση της τυπικής γραμμικής παλινδρόμησης.

## 7.3 Γενίκευση της γραμμικής παλινδρόμησης

### 7.3.1 Γραμμική παλινδρόμηση πολλών μεταβλητών

Η γραμμική παλινδρόμηση δύο μεταβλητών που εξετάστηκε στην προηγούμενη ενότητα γενικεύεται εύκολα για περισσότερες από δύο μεταβλητές. Θεωρούμε ότι οι $m + 1$ μεταβλητές $X_1, X_2, \ldots, X_m, Y$, συνδέονται με το γραμμικό νόμο

$$Y = a_0 + a_1X_1 + \cdots + a_mX_m + W \tag{7.65}$$

όπου $a_0, a_1, \ldots, a_m$ άγνωστες αριθμητικές παράμετροι και $W$ τυχαία μεταβλητή που εκφράζει το σφάλμα της εκτίμησης

$$\hat{Y} = a_0 + a_1X_1 + \cdots + a_mX_m \tag{7.66}$$

δηλαδή $W = Y - \hat{Y} = Y - (a_0 + a_1X_1 + \cdots + a_mX_m)$. Οι παράμετροι υπολογίζονται έτσι ώστε να ελαχιστοποιείται το μέσο τετραγωνικό σφάλμα, δηλαδή το μέγεθος

$$E[W^2] = E[(Y - \hat{Y})^2] = E[(Y - (a_0 + a_1X_1 + \cdots + a_mX_m))^2]$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [y - (a_0 + a_1X_1 + \cdots + a_mX_m)]^2 f_{XY}(x, y)\, dx\, dy \tag{7.67}$$

Παραγωγίζοντας την (7.26) ως προς $a_i$ παίρνουμε μετά από πράξεις (βλ. π.χ. Papoulis, 1990, σ. 450) ένα γραμμικό σύστημα $m + 1$ εξισώσεων, το οποίο υπό μορφή μητρώων γράφεται

$$\mathbf{c}_{XX}\,\mathbf{a} = \mathbf{c}_{XY} \tag{7.68}$$

όπου το $\mathbf{c}_{XX}$ είναι συμμετρικό τετραγωνικό μητρώο συντελεστών διάστασης $(m + 1) \times (m + 1)$ και το $\mathbf{c}_{XY}$ είναι το διάνυσμα σταθερών όρων διάστασης $(m + 1)$:

$$\mathbf{c}_{XX} = \begin{bmatrix} 1 & E[X_1] & \dots & E[X_m] \\ E[X_1] & E[X_1^2] & \dots & E[X_1X_m] \\ \vdots & \vdots & \ddots & \vdots \\ E[X_m] & E[X_mX_1] & \dots & E[X_m^2] \end{bmatrix} \quad \mathbf{c}_{XY} = \begin{bmatrix} \mathrm{E}[Y] \\ E[X_1Y] \\ \vdots \\ E[X_mY] \end{bmatrix} \tag{7.69}$$

ενώ **a** είναι το διάνυσμα των αγνώστων διάστασης $(m + 1)$:

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} \tag{7.70}$$

Η επίλυση του συστήματος δίνει τις τιμές των άγνωστων παραμέτρων. Για την εφαρμογή αντικαθιστούμε τις θεωρητικές αναμενόμενες τιμές με τις εκτιμήσεις τους από το δείγμα των $n$ σημείων $(x_{1i}, x_{2i}, \dots, x_{mi},, y_i)$:

$$E[X_s] = \bar{x}_s\,,\ E[Y] = \bar{y},\ E[X_sX_r] = \frac{1}{n}\sum_{i=1}^{n} x_{si}\,x_{ri},\ E[X_sY] = \frac{1}{n}\sum_{i=1}^{n} x_{si}\,y_i \tag{7.71}$$

όπου $s, r = 1, \dots, m$.

Στην περίπτωση της ομογενούς εξίσωσης $\hat{Y} = a_1X_1 + \cdots + a_mX_m$ (χωρίς το σταθερό όρο $a_0$) η επίλυση είναι παρόμοια. Η μαθηματική έκφραση παραμένει ίδια, με τη διαφορά ότι διαγράφεται η πρώτη σειρά και η πρώτη στήλη του μητρώου $\mathbf{c}_{XX}$ και η πρώτη σειρά των διανυσμάτων $\mathbf{c}_{XY}$ και $\mathbf{a}$.

Για τον υπολογισμό του συντελεστή προσδιορισμού, ο οποίος δείχνει το μέτρο της προσαρμογής της γραμμικής εξίσωσης προς τα δεδομένα εφαρμόζεται η γενικευμένη εξίσωση (7.41) ή η αντίστοιχή της (7.14), αφού προηγουμένως υπολογιστεί το τετραγωνικό σφάλμα.

### 7.3.2 Μη γραμμική παλινδρόμηση με γραμμικούς συντελεστές

Το παραπάνω γενικευμένο γραμμικό μοντέλο μπορεί να εφαρμοστεί άμεσα και για μη γραμμικό ως προς $X_i$ νόμο $Y = \varphi(X_1, X_2, \ldots) + W$, με την προϋπόθεση ότι ο νόμος είναι γραμμικός ως προς τις άγνωστες παραμέτρους $a_0$, $a_1$, … Για παράδειγμα η πολυωνυμική παλινδρόμηση μιας μεταβλητής βαθμού $m$:

$$Y = a_0 + a_1X + a_2X^2 + \cdots + a_mX^m + W \qquad (7.72)$$

ανάγεται άμεσα στην γραμμική παλινδρόμηση $m$ μεταβλητών (7.65) αν τεθεί $X_1 = X$, $X_2 = X^2$, …, $X_m = X^m$.

### 7.3.3 Γραμμικοποίηση με μετασχηματισμούς μεταβλητών

Η παραπάνω τεχνική δεν μπορεί να εφαρμοστεί όταν η συνάρτηση $\varphi(\ )$ δεν είναι γραμμική ως προς τις παραμέτρους. Σε τέτοιες περιπτώσεις είναι μερικές φορές δυνατό να γίνει γραμμικοποίηση με κατάλληλο μετασχηματισμό των μεταβλητών. Παραδείγματα αυτού του τύπου αποτελούν η εκθετική συνάρτηση και η συνάρτηση δύναμης μιας ή περισσότερων μεταβλητών. Στον Πίν. 7.3 δίνονται οι απαιτούμενοι μετασχηματισμοί μεταβλητών και παραμέτρων για τις συναρτήσεις αυτές. Οι μετασχηματισμοί προκύπτουν με λογαρίθμιση και των δύο μελών των αρχικών εξισώσεων.

Τα στατιστικά χαρακτηριστικά που απαιτούνται για τον υπολογισμό των παραμέτρων αναφέρονται στις μετασχηματισμένες μεταβλητές και όχι στις αρχικές. Επίσης το μέσο τετραγωνικό σφάλμα και ο συντελεστής προσδιορισμού υπολογίζονται για τις μετασχηματισμένες μεταβλητές.

**Πίν. 7.3** Τυπικές συναρτήσεις που επιδέχονται γραμμικοποίηση με μετασχηματισμό μεταβλητών και αντίστοιχοι μετασχηματισμοί.

| Τύπος αρχικής συνάρτησης | Έκφραση αρχικής συνάρτησης | Γραμμικοποιημένη συνάρτηση | Μετασχηματισμοί μεταβλητών | Μετασχηματισμοί παραμέτρων |
|---|---|---|---|---|
| Εκθετική δύο μεταβλητών | $\hat{Y} = \kappa\, \mathrm{e}^{\lambda X}$ | $\hat{Y}' = a + bX'$ | $\hat{Y}' = \ln \hat{Y}$<br>$X' = X$ | $a = \ln \kappa$<br>$b = \lambda$ |
| Δύναμης δύο μεταβλητών | $\hat{Y} = \kappa\, X^{\lambda}$ | $\hat{Y}' = a + bX'$ | $\hat{Y}' = \ln \hat{Y}$<br>$X' = \ln X$ | $a = \ln \kappa$<br>$b = \lambda$ |
| Εκθετική πολλών μεταβλητών | $\hat{Y} = \kappa\, \mathrm{e}^{\lambda_1 X_1 + \lambda_2 X_2 + \cdots}$ | $\hat{Y}' = a_0 +$<br>$a_1 X_1' +$<br>$a_2 X_2' + \cdots$ | $\hat{Y}' = \ln \hat{Y}$<br>$X_1' = X_1,$<br>$X_2' = X_2, \ldots$ | $a = \ln \kappa$<br>$b_1 = \lambda_1,$<br>$b_2 = \lambda_2, \ldots$ |
| Δύναμης πολλών μεταβλητών | $\hat{Y} = \kappa\, X_1^{\lambda_1} X_2^{\lambda_2} \cdots$ | $\hat{Y}' = a_0 +$<br>$a_1 X_1' +$<br>$a_2 X_2' + \cdots$ | $\hat{Y}' = \ln \hat{Y}$<br>$X_1' = \ln X_1,$<br>$X_2' = \ln X_2, \ldots$ | $a = \ln \kappa$<br>$b_1 = \lambda_1,$<br>$b_2 = \lambda_2, \ldots$ |

**Εφαρμογή 7.3.3**

*Στον υδρομετρικό σταθμό Αχλαδόκαστρο του ποταμού Ευήνου έχουν γίνει συστηματικές μετρήσεις στάθμης και παροχής του ποταμού. 35 από αυτές που αναφέρονται σε διάρκεια 2.5 ετών, κατά την οποία δεν υπήρχαν ουσιαστικές μεταβολές της γεωμετρίας και των χαρακτηριστικών της κοίτης του ποταμού, φαίνονται στον Πίν. 7.4. Με βάση τις μετρήσεις αυτές, να περιγραφεί μαθηματικά η σχέση στάθμης και παροχής του ποταμού για την υπόψη περίοδο.*

Η εφαρμογή αυτή αποτελεί ένα από τα πιο τυπικά προβλήματα της τεχνικής υδρολογίας. Η κατασκευή καμπύλης στάθμης-παροχής είναι απαραίτητη προϋπόθεση για την εξαγωγή της χρονοσειράς της παροχής σε κάθε θέση υδρομέτρησης, με βάση την αντίστοιχη χρονοσειρά της στάθμης.

**Πίν. 7.4** Ταυτόχρονες μετρήσεις στάθμης και παροχής του ποταμού Ευήνου στη θέση Αχλαδόκαστρο, για την περίοδο από Μάρτιο 1974 μέχρι Αύγουστο 1976.

| Α/Α | Ημερο-μηνία | Στάθμη (m) | Παροχή ($m^3/s$) | Α/Α | Ημερο-μηνία | Στάθμη (m) | Παροχή ($m^3/s$) |
|---|---|---|---|---|---|---|---|
| 1 | 19/03/74 | 1.65 | 32.174 | 19 | 19/07/75 | 0.58 | 1.447 |
| 2 | 29/03/74 | 1.51 | 21.537 | 20 | 31/07/75 | 0.68 | 2.448 |
| 3 | 29/04/74 | 1.37 | 23.413 | 21 | 07/08/75 | 0.77 | 3.646 |
| 4 | 27/05/74 | 1.12 | 17.157 | 22 | 12/08/75 | 0.66 | 2.200 |
| 5 | 24/06/74 | 0.79 | 4.252 | 23 | 18/09/75 | 0.49 | 1.030 |
| 6 | 16/07/74 | 0.63 | 1.894 | 24 | 24/09/75 | 0.48 | 0.920 |
| 7 | 02/09/74 | 0.56 | 1.067 | 25 | 08/01/76 | 1.05 | 12.066 |
| 8 | 20/09/74 | 0.54 | 0.891 | 26 | 16/02/76 | 1.35 | 25.899 |
| 9 | 15/10/74 | 0.69 | 1.544 | 27 | 19/03/76 | 1.08 | 14.655 |
| 10 | 13/11/74 | 1.35 | 25.314 | 28 | 13/04/76 | 1.49 | 40.192 |
| 11 | 29/11/74 | 1.42 | 34.201 | 29 | 21/05/76 | 0.99 | 7.484 |
| 12 | 16/01/75 | 1.13 | 15.769 | 30 | 02/06/76 | 0.87 | 5.350 |
| 13 | 21/01/75 | 1.06 | 10.907 | 31 | 11/06/76 | 0.83 | 4.296 |
| 14 | 07/03/75 | 1.19 | 18.037 | 32 | 26/06/76 | 0.75 | 3.491 |
| 15 | 17/04/75 | 1.05 | 10.494 | 33 | 13/07/76 | 0.67 | 2.566 |
| 16 | 28/04/75 | 0.99 | 8.080 | 34 | 30/07/76 | 0.63 | 2.076 |
| 17 | 09/06/75 | 0.81 | 4.600 | 35 | 18/08/76 | 0.54 | 1.243 |
| 18 | 10/07/75 | 0.60 | 1.872 | | | | |

Στο Σχ. 7.10 έχει απεικονιστεί το σύνολο των μετρήσεων του Πίν. 7.4 υπό μορφή διαγράμματος παροχής συναρτήσει της στάθμης. Έχει καθιερωθεί (για λόγους εποπτικότερης παρουσίασης) σε τέτοια διαγράμματα η στάθμη $z$ να απεικονίζεται στον κατακόρυφο άξονα, παρόλο που αποτελεί την ανεξάρτητη μεταβλητή του προβλήματος (που κατά κανόνα τοποθετείται στον οριζόντιο άξονα). Παρατηρούμε ότι υπάρχει σαφής συσχέτιση στάθμης και παροχής, η οποία όμως δεν είναι γραμμική. Άλλωστε, από την υδραυλική γνωρίζουμε ότι η σχέση στάθμης-παροχής περιγράφεται κατά προσέγγιση από μια εξίσωση δύναμης, δηλαδή της μορφής $q = \kappa z^{\lambda}$. Όπως είδαμε στον Πίν. 7.3, η εξίσωση αυτής της μορφής γραμμικοποιείται με λογαριθμικό μετασχηματισμό και των δύο μεταβλητών. Ισοδύναμα, το σημειοσύνολο ευθειοποιείται αν παρασταθεί σε διπλό λογαριθμικό χαρτί. Πράγματι, αυτό έχει γίνει στο Σχ. 7.11, όπου πράγματι διαπιστώνουμε την γραμμικότητα της σχέσης ανάμεσα στους λογαρίθμους των μεταβλητών.

Χρησιμοποιώντας τις μετασχηματισμένες μεταβλητές $X \equiv \ln Z$ και $Y \equiv \ln Q$, υπολογίζουμε τα αθροίσματα που θα μας επιτρέψουν τον υπολογισμό των παραμέτρων της γραμμικής σχέσης. Έτσι έχουμε:

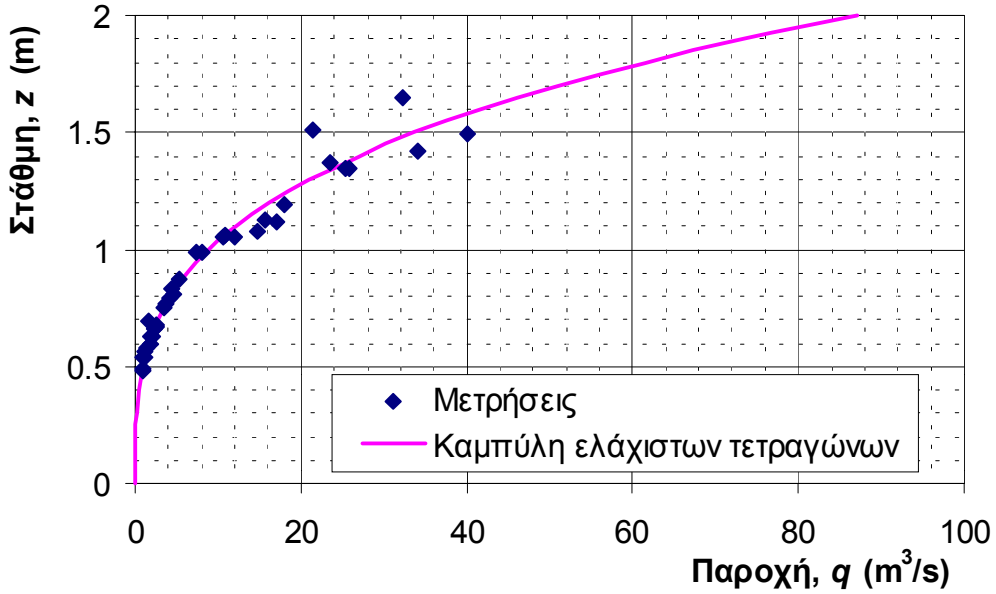
$$n = 35,\ \sum x = -4.933,\ \sum y = 60.107,\ \sum xy = 6.065,$$

$$\sum x^2 = 5.107,\ \sum y^2 = 152.323$$

Κατά συνέπεια από την εξίσωση (7.18) βρίσκουμε $r = 0.988$. Η τιμή αυτή είναι μεγαλύτερη από την κρίσιμη τιμή

$$r_c = 2/\sqrt{35} = 0.34$$

πράγμα που επιβεβαιώνει και στατιστικά η ύπαρξη ισχυρής γραμμικής συσχέτισης ανάμεσα στους λογαρίθμους της στάθμης και της παροχής.




**Σχ. 7.10** Παροχή συναρτήσει της στάθμης στη θέση Αχλαδόκαστρο του ποταμού Ευήνου: μετρήσεις και προσαρμοσμένη με τη μέθοδο ελάχιστων τετραγώνων καμπύλη δύναμης.

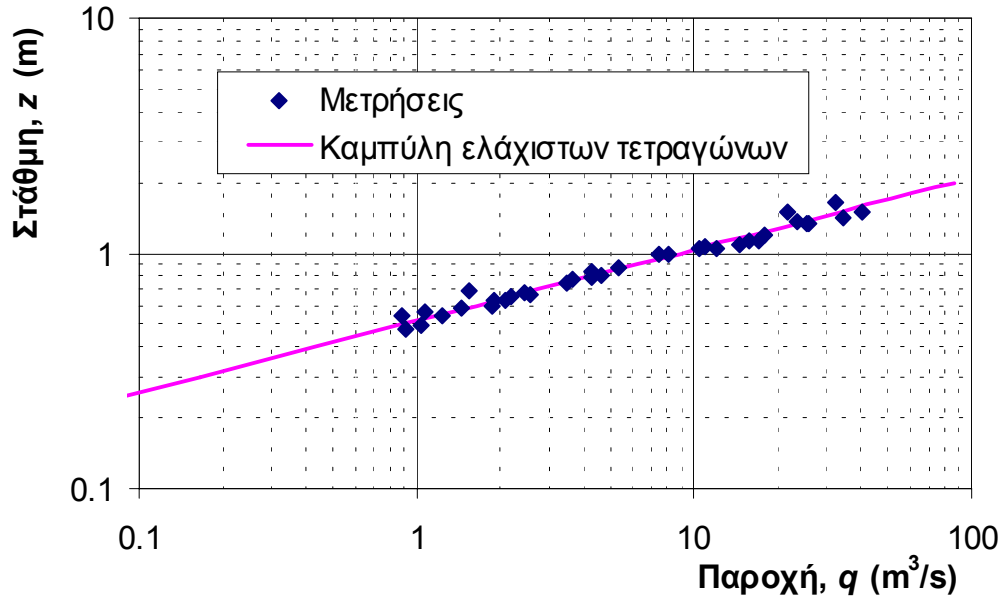
Στη συνέχεια υπολογίζουμε τους συντελεστές $b$ και $a$ της γραμμικής παλινδρόμησης $y = a + b\,x$, χρησιμοποιώντας τις εξισώσεις (7.10) και (7.11), αντίστοιχα. Έτσι, από την (7.10) παίρνουμε $b = 3.296$ και από την (7.11) $a = 2.182$, οπότε η ζητούμενη ευθεία ελάχιστων τετραγώνων είναι η

$$y = 2.182 + 3.296\,x$$

Υψώνοντας τα δύο μέλη της στην e, παίρνουμε την τελική έκφραση

$$q = 8.86\,z^{3.296}$$

η οποία έχει παρασταθεί γραφικά στο Σχ. 7.10 καθώς και στο Σχ. 7.11.




**Σχ. 7.11** Παροχή συναρτήσει της στάθμης στη θέση Αχλαδόκαστρο του ποταμού Ευήνου σε διπλό λογαριθμικό χαρτί: μετρήσεις και προσαρμοσμένη με τη μέθοδο ελάχιστων τετραγώνων καμπύλη δύναμης.

Τελειώνοντας το παράδειγμά μας, θα πρέπει να σημειώσουμε ότι η κατάρτιση καμπυλών στάθμης-παροχής δεν είναι τόσο απλό πρόβλημα όσο φαίνεται στην εν λόγω εφαρμογή. Οι δυσκολίες προκύπτουν από τις μεταβαλλόμενες ιδιότητες της διατομής του ποταμού (γεωμετρία, τραχύτητα, κλίση), που οδηγούν σε μετατοπιζόμενες με το χρόνο καμπύλες στάθμης-παροχής. Μια επιπρόσθετη δυσκολία εισάγουν τα συχνά σφάλματα των υδρομετρήσεων.

## 7.4 Γενική παλινδρόμηση

### 7.4.1 Μεταβλητές με τυχούσα συνάρτηση κατανομής

Στις προηγούμενες ενότητες μελετήθηκαν διάφορα μοντέλα παλινδρόμησης, τα οποία στην περίπτωση των δύο μεταβλητών γράφονται με τη γενική έκφραση

$$\hat{Y} = \phi(X) \tag{7.73}$$

Η συνάρτηση $\phi(X)$ μπορεί να είναι π.χ. γραμμική, παραβολική, πολυωνυμική $m$ βαθμού κτλ. Σε όλες τις περιπτώσεις που εξετάστηκαν η μορφή

της συνάρτησης είχε προεπιλεγεί πριν από την εφαρμογή της μεθόδου ελάχιστων τετραγώνων, κάτι το οποίο γίνεται συχνά στην πράξη, μετά από μια αρχική εξερεύνηση των δεδομένων. Εύλογα, λοιπόν, τίθεται το ερώτημα εάν υπάρχει βέλτιστη συνάρτηση $\phi(X)$ και πώς αυτή μπορεί να προσδιοριστεί. Η βέλτιστη συνάρτηση, αν υπάρχει, θα είναι αυτή για την οποία το μέσο τετραγωνικό σφάλμα είναι μικρότερο από ό,τι σε κάθε άλλη συνάρτηση.

Αρχικά διαπιστώνουμε ότι το ερώτημα αυτό δεν έχει ουσιαστικό νόημα στην περίπτωση που αντιμετωπίζουμε το πρόβλημα με την προσδιοριστική προσέγγιση. Πράγματι, αν δοθεί ένα σύνολο $n$ σημείων $(x_i, y_i)$ τότε υπάρχουν άπειρες καμπύλες (δηλαδή συναρτήσεις) που περνούν από όλα τα σημεία $(x_i, y_i)$, δίνοντας έτσι μηδενικό τετραγωνικό σφάλμα. Μια απ' αυτές είναι το ταυτοτικό πολυώνυμο βαθμού $n - 1$. Μια άλλη είναι η τεθλασμένη που συνδέει όλα τα σημεία κατά τέτοιο τρόπο ώστε να αποτελεί συνάρτηση (δηλαδή κατά σειρά μεγέθους της τετμημένης $x$).

Το ερώτημα αποκτά νόημα όταν αντιμετωπιστεί με την πιθανοτική προσέγγιση, οπότε η απάντηση δεν μπορεί να εξαρτάται από το συγκεκριμένο σημειοσύνολο που δίνεται, αλλά από την από κοινού συνάρτηση πυκνότητας πιθανότητας των δύο μεταβλητών $f_{XY}(x, y)$. Αποδεικνύεται, λοιπόν, (π.χ. Papoulis, 1990, σ. 183) ότι υπάρχει μια μοναδική (βέλτιστη) συνάρτηση που ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα, η οποία δίνεται από τη σχέση

$$\phi(x) = E[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(x, y)\, dy \tag{7.74}$$

Υπενθυμίζουμε ότι $E[Y|X = x]$ είναι η δεσμευμένη μέση τιμή της $Y$ για δεδομένη τιμή της $X = x$ (βλ. εξ. (2.37)), ενώ $f_{Y|X}(x, y) = f_{XY}(x, y) / f_X(x)$ είναι η δεσμευμένη συνάρτηση πυκνότητας πιθανότητας της $Y$ για δεδομένη τιμή της $X = x$ (βλ. εξ. (2.32)).

Ένα επεξηγηματικό σκαρίφημα για τη γενική παλινδρόμηση δίνεται στο Σχ. 7.12.

### 7.4.2 Μεταβλητές με κανονική κατανομή

**Εισαγωγικές έννοιες για τη διδιάστατη κανονική κατανομή**

Δύο μεταβλητές $X$ και $Y$ λέμε ότι έχουν από κοινού κανονική συνάρτηση κατανομής (ή διδιάστατη κανονική κατανομή), όταν κάθε γραμμικός συνδυασμός των μεταβλητών $Z = \alpha X + \beta Y$ ακολουθεί κανονική κατανομή. Αποδεικνύεται (βλ. π.χ. Papoulis, 1990, σ. 162) ότι η από κοινού συνάρτηση πυκνότητας πιθανότητας των μεταβλητών σε αυτή την περίπτωση είναι

$$f_{XY}(x,y) = \frac{\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right]}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \quad (7.75)$$

Η κατανομή έχει πέντε παραμέτρους, τις $\mu_X$, $\mu_Y$, $\sigma_X$, $\sigma_Y$ και $\rho$.

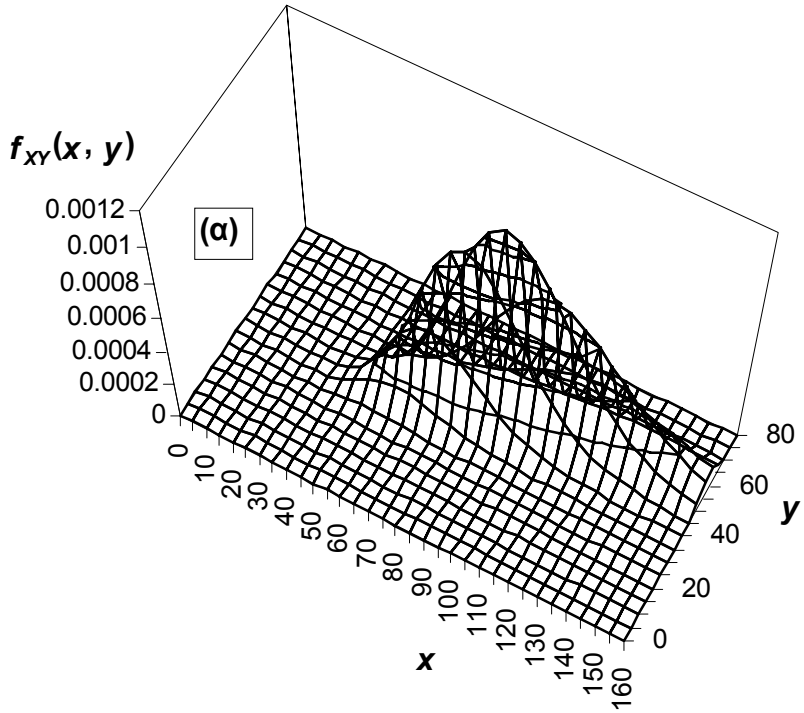
Όπως προκύπτει άμεσα από τον ορισμό οι περιθώριες κατανομές των $X$ και $Y$ είναι κανονικές και οι δύο. Ειδικότερα, αποδεικνύεται ότι η μέση τιμή της $X$ είναι $\mu_X$ και η τυπική απόκλισή της είναι $\sigma_X$. Επίσης, η μέση τιμή της $Y$ είναι $\mu_Y$ και η τυπική απόκλισή της $\sigma_Y$. Τέλος ο συντελεστής συσχέτισης των δύο μεταβλητών, σύμφωνα με τον ορισμό της (2.35), είναι $\rho$.

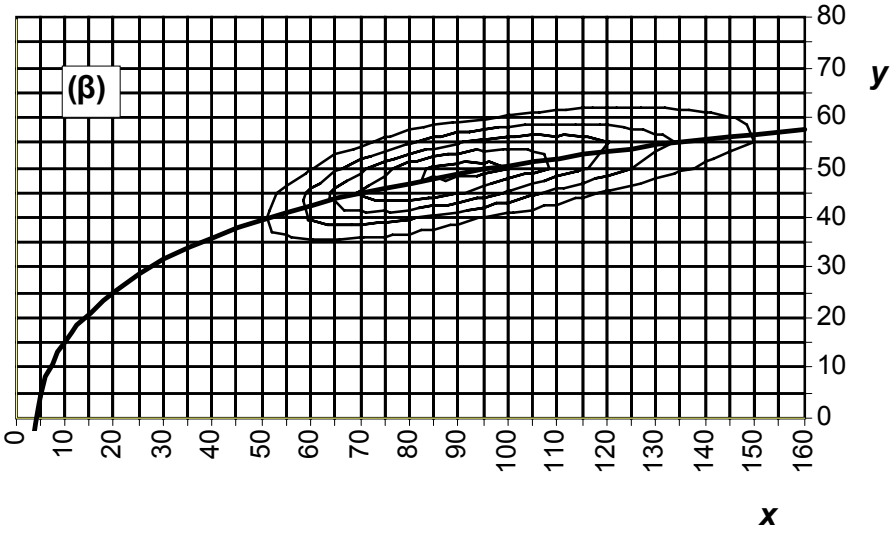
Η διδιάστατη κανονική κατανομή έχει ορισμένες ενδιαφέρουσες ιδιότητες σχετικά με τη συνεπαγόμενη βέλτιστη παλινδρόμηση, οι οποίες μελετώνται στο επόμενο εδάφιο.

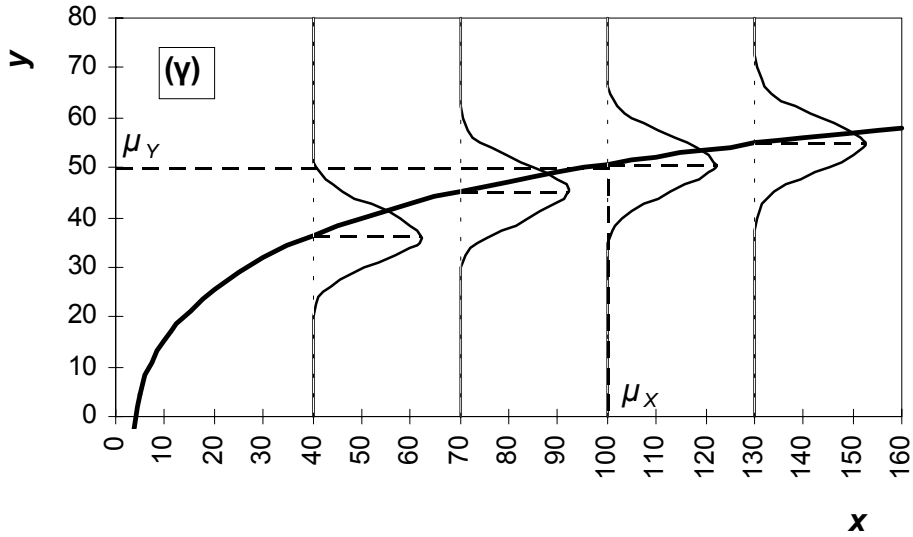
**Η γραμμική παλινδρόμηση ως συνέπεια της γενικής παλινδρόμησης**

Συνδυάζοντας την εξίσωση (7.75) και τον ορισμό της δεσμευμένης συνάρτησης πυκνότητας πιθανότητας (εξίσωση (2.32)) βρίσκουμε εύκολα ότι η τελευταία για μεταβλητές $X$, $Y$ που ακολουθούν διδιάστατη κανονική κατανομή έχει την έκφραση

$$f_{Y|X}(x,y) = \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}}\exp\left\{-\frac{\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2}{2\sigma_Y^2(1-\rho^2)}\right\} \quad (7.76)$$


$f_{XY}(x, y)$
(α)
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
x
0 20 40 60 80
y


(β)
0 10 20 30 40 50 60 70 80
y
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
x


(γ)
y
0 10 20 30 40 50 60 70 80
$\mu_Y$
$\mu_X$
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
x

1. Η δεσμευμένη συνάρτηση κατανομής της $Y$ για δεδομένη τιμή της $X = x$ είναι κανονική. Αυτό διαπιστώνεται από τη σύγκριση της παραπάνω εξίσωσης με την πυκνότητα πιθανότητας της κανονικής κατανομής (εξίσωση (2.60)).

2. Η δεσμευμένη μέση τιμή της $Y$ για δεδομένη τιμή της $X = x$ είναι

$$\mu_{Y|X} = E[Y|X = x] = \mu_Y - \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) \tag{7.77}$$

   Παρατηρούμε ότι η δεσμευμένη μέση τιμή είναι γραμμική συνάρτηση της $x$. Εξ άλλου, από την (7.74) προκύπτει ότι η γενική (βέλτιστη) παλινδρόμηση $\phi(x)$ ταυτίζεται με την παραπάνω γραμμική συνάρτηση. Κατά συνέπεια στην κανονική κατανομή η βέλτιστη παλινδρόμηση είναι η γραμμική. Τέλος, συγκρίνοντας τις παραπάνω εξισώσεις με τις (7.54) και (7.55) που είχαν εξαχθεί για τη γραμμική παλινδρόμηση χωρίς καμιά υπόθεση για την κατανομή που ακολουθούν οι μεταβλητές, διαπιστώνουμε ότι υπάρχει ταύτιση.

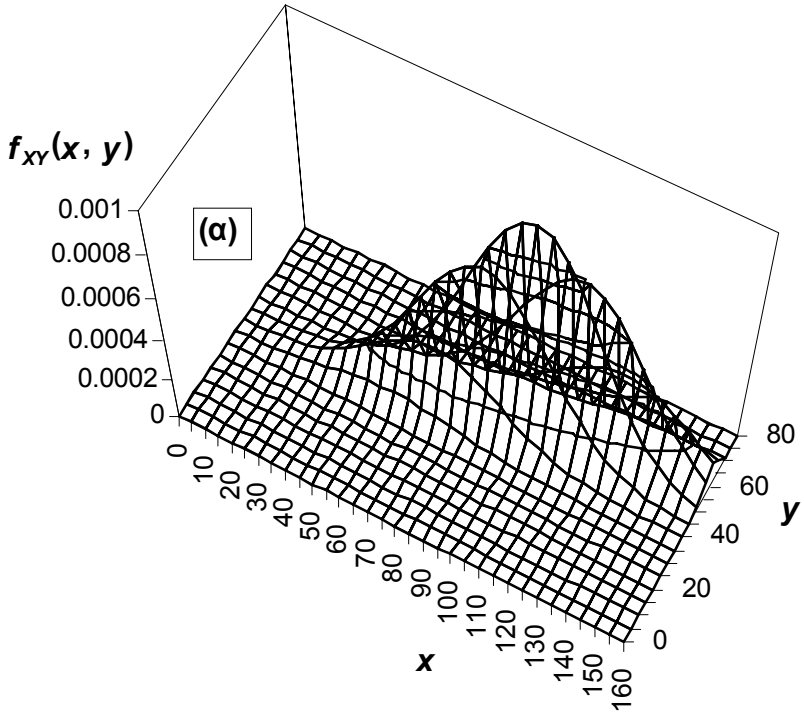
3. Η δεσμευμένη διασπορά της $Y$ για δεδομένη τιμή της $X = x$ είναι

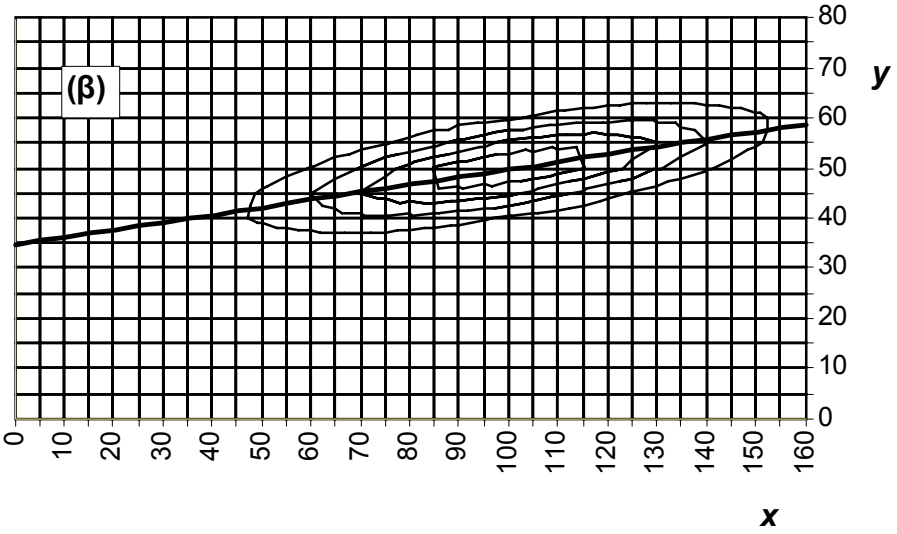
$$\sigma^2_{Y|X} = \mathrm{Var}[Y|X = x] = \sigma^2_Y (1 - \rho^2) \tag{7.78}$$

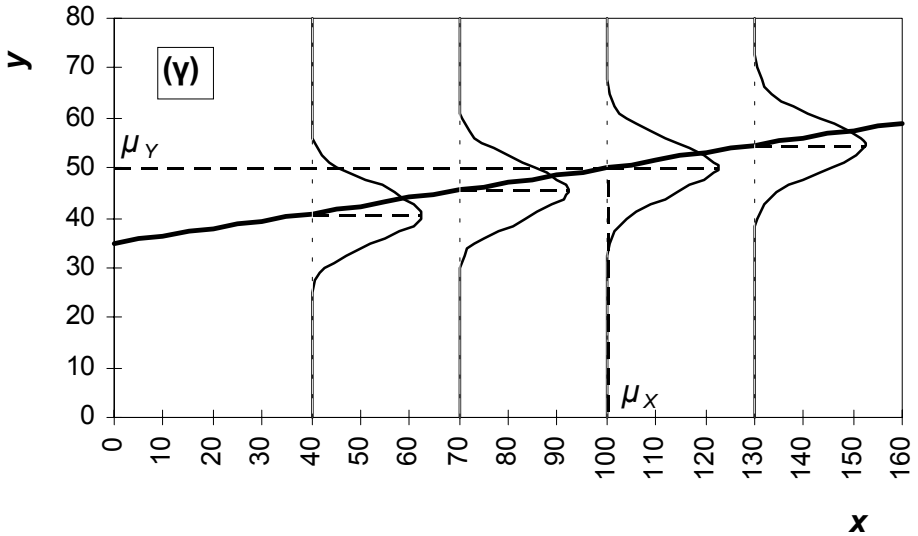
   Παρατηρούμε ότι η διασπορά αυτή είναι ανεξάρτητη της τιμής $x$, δηλαδή είναι σταθερή για όλα τα $x$. Το μέγεθος αυτό δεν είναι άλλο από το μέσο τετραγωνικό σφάλμα της εκτίμησης, όπως επιβεβαιώνουμε συγκρίνοντας με την εξίσωση (7.35).

---

**Σχ. 7.12** (Απέναντι σελίδα) Εποπτικό σκαρίφημα για τη γενική παλινδρόμηση. Στο (α) απεικονίζεται σε προοπτικό διάγραμμα η από κοινού συνάρτηση πυκνότητας πιθανότητας των μεταβλητών $X$ και $Y$. Στο (β) απεικονίζεται επίσης η ίδια συνάρτηση υπό μορφή καμπυλών ίσης πυκνότητας (ισοδιάσταση 0.0002), καθώς και η γενική καμπύλη παλινδρόμησης, η οποία ορίζεται από την εξίσωση (7.74) (παχιά συνεχής γραμμή). Στο (γ) απεικονίζεται και πάλι η γενική καμπύλη παλινδρόμησης (παρατηρούμε ότι περνά από το σημείο $(\mu_X, \mu_Y)$) και η δεσμευμένη συνάρτηση πυκνότητας πιθανότητας $f_{Y|X}(y, x)$ για τέσσερις διαφορετικές τιμές του $x$ (κλίμακα αυθαίρετη, μέγιστη τεταγμένη 0.075 και στις τέσσερις θέσεις). Οι παράμετροι των περιθώριων κατανομών των $X$ και $Y$ για το παράδειγμα που απεικονίζεται είναι: $\mu_X = 100$, $\sigma_X = 30$, $\mu_Y = 50$, $\sigma_Y = 7$.


$f_{XY}(x, y)$
(α)
0.001
0.0008
0.0006
0.0004
0.0002
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
x
0 20 40 60 80
y


(β)
0 10 20 30 40 50 60 70 80
y
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
x


(γ)
y
0 10 20 30 40 50 60 70 80
$\mu_Y$
$\mu_X$
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
x

### Ομοσκεδαστικότητα

Η ιδιότητα του παραπάνω σημείου 3, σύμφωνα με την οποία το τετραγωνικό σφάλμα είναι σταθερό για όλα τις τιμές του $x$ λέγεται *ομοσκεδαστικότητα*. Στη γραμμική παλινδρόμηση μεταξύ μεταβλητών που ακολουθούν κανονική κατανομή η ομοσκεδαστικότητα είναι δεδομένη. Σε περίπτωση που οι μεταβλητές ακολουθούν άλλες κατανομές, η ομοσκεδαστικότητα δεν είναι εξασφαλισμένη, αλλά ωστόσο αποτελεί μια επιθυμητή ιδιότητα. Πρακτικά η ομοσκεδαστικότητα μπορεί να ελεγχθεί με τη γραφική απεικόνιση του σφάλματος εκτίμησης $w$ συναρτήσει των τιμών της μεταβλητής $x$. Αν τα σημεία $(x_i, w_i)$ κατανέμονται με τυχαίο τρόπο σχηματίζοντας ένα νέφος γύρω από την οριζόντια $w = 0$, τότε θεωρούμε ότι υπάρχει ομοσκεδαστικότητα. Αν υπάρχει συστηματικότητα στη διάταξη των σημείων (π.χ. τα σημεία σχηματίζουν κάποια καμπύλη, ή τα σημεία αποκλίνουν λιγότερο από την οριζόντια $w = 0$ για μικρές τιμές του $x$ και περισσότερο για μεγάλες τιμές του $x$), τότε μιλούμε για *ετεροσκεδαστικότητα*. Η ετεροσκεδαστικότητα δείχνει ακαταλληλότητα του μοντέλου παλινδρόμησης που έχει επιλεγεί. Για το λόγο αυτό, αν διαπιστωθεί ετεροσκεδαστικότητα, τότε επαναλαμβάνεται η παλινδρόμηση με άλλο μοντέλο, π.χ. προσθέτοντας ένα μη γραμμικό όρο στην εξίσωση παλινδρόμησης ή κάνοντας κατάλληλους μετασχηματισμούς των μεταβλητών, σαν αυτούς που συζητήθηκαν στο εδάφιο 7.3.3.

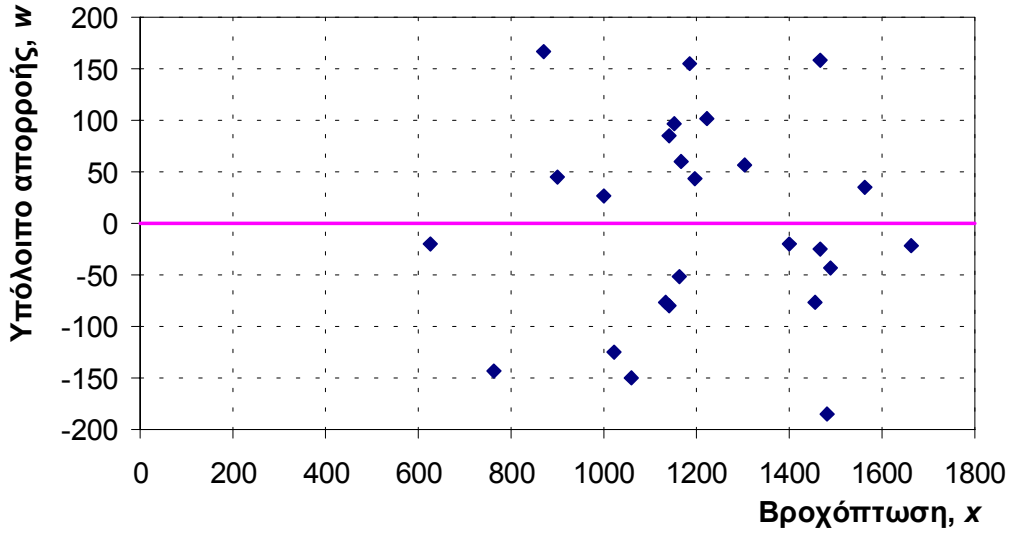
#### Εφαρμογή 7.4.2

*Να εξεταστεί η ομοσκεδαστικότητα των υπολοίπων στις Εφαρμογές 7.1 και7.3.3.*

Υπενθυμίζεται ότι στην Εφαρμογή 7.1 ασχοληθήκαμε (μεταξύ άλλων) με τη διατύπωση μιας γραμμικής σχέσης ανάμεσα στη βροχόπτωση και την απορροή της λεκάνης ανάντη Κρεμαστών του Αχελώου, ενώ στην Εφαρμογή 7.3.3 εξερευνήσαμε τη σχέση ανάμεσα στη στάθμη και την παροχή στη θέση Αχλαδόκαστρο του Ευήνου.
Ο έλεγχος της ομοσκεδαστικότητας των υπολοίπων παλινδρόμησης

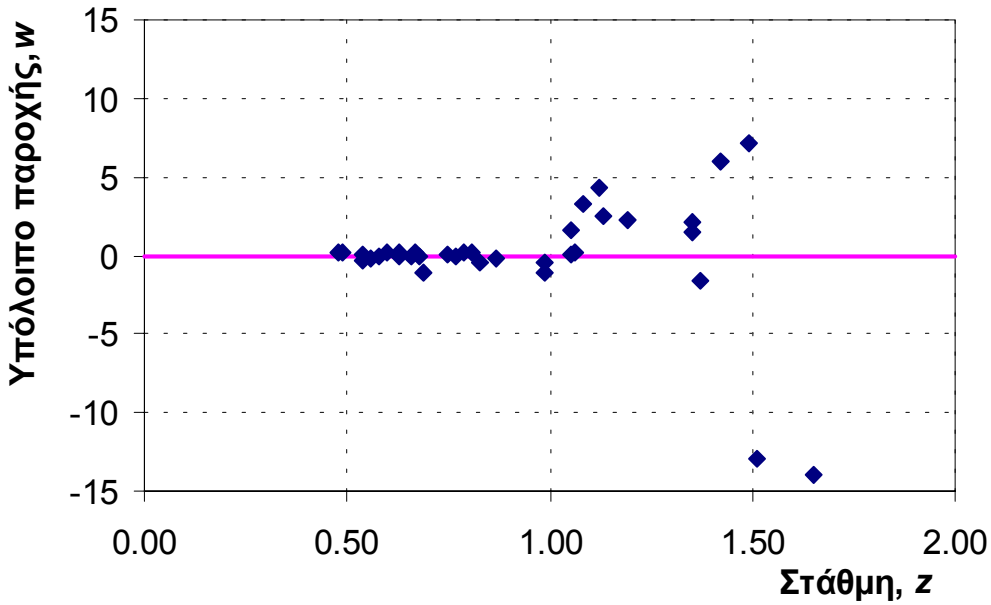
**Σχ. 7.13** (Απέναντι σελίδα) Επεξηγηματικό σκαρίφημα για τη γραμμική παλινδρόμηση ως συνέπεια της διδιάστατης κανονικής κατανομής. Το σχήμα είναι παρόμοιο με το Σχ. 7.12, με τη μόνη διαφορά ότι εδώ η από κοινού συνάρτηση κατανομής των $X$ και $Y$ είναι κανονική. Οι παράμετροι της κατανομής των $X$ και $Y$ για το παράδειγμα που απεικονίζεται είναι: $\mu_X = 100$, $\sigma_X = 30$, $\mu_Y = 50$, $\sigma_Y = 7$ (όπως και στο Σχ. 7.12) και $\rho_{XY} = 0.65$.

γίνεται με γραφικό τρόπο. Στο Σχ. 7.14 φαίνονται τα υπόλοιπα παλινδρόμησης της Εφαρμογής 7.1. Τα υπόλοιπα αυτά υπολογίζονται από την εξίσωση $w_i = y_i - (a + b\ x_i)$ και απεικονίζονται συναρτήσει της ανεξάρτητης μεταβλητής (στην προκειμένη περίπτωση της βροχόπτωσης) $x$. Η τυχαία διάταξη των σημείων γύρω από τη γραμμή $w = 0$ δείχνει ότι υπάρχει ομοσκεδαστικότητα των υπολοίπων.

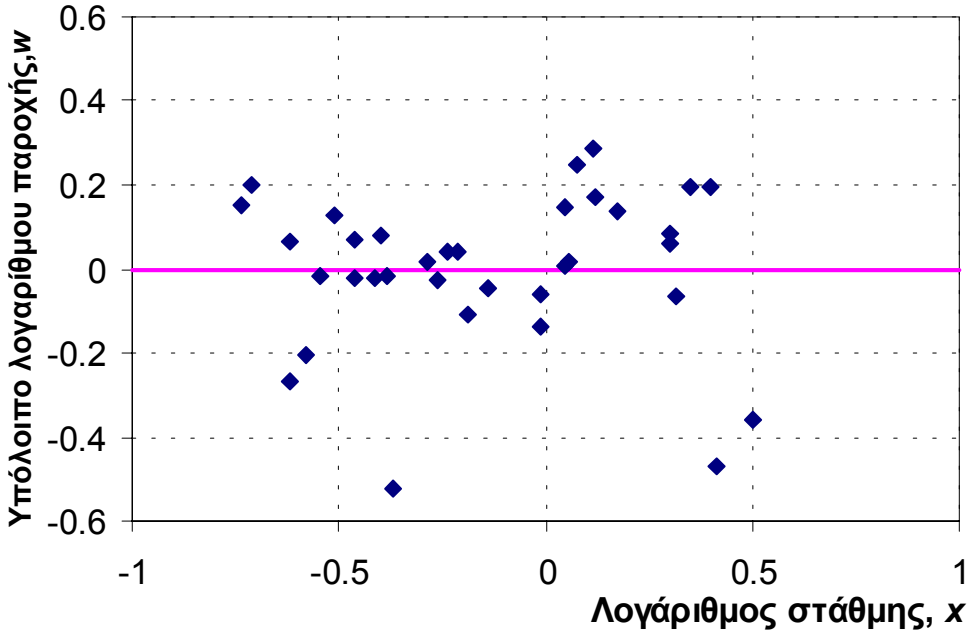



**Σχ. 7.14** Διάγραμμα για τον έλεγχο της ομοσκεδαστικότητας των υπολοίπων παλινδρόμησης στην εκτίμηση της απορροής συναρτήσει της βροχόπτωσης στα Κρεμαστά (βλ. Εφαρμογή 7.1).

Στο Σχ. 7.15 έχει γίνει ένα παρόμοιο διάγραμμα για την Εφαρμογή 7.3.3. Εδώ η παροχή δίνεται συναρτήσει της στάθμης από την εξίσωση δύναμης $q = \kappa\ z^{\lambda}$, οπότε τα υπόλοιπα προκύπτουν από την εξίσωση $w_i = q_i - \kappa\ z_i^{\lambda}$. Στο διάγραμμα είναι έκδηλη η ετεροσκεδαστικότητα των υπολοίπων: τα σημεία αποκλίνουν λιγότερο από την οριζόντια γραμμή $w = 0$ για μικρές τιμές της στάθμης και περισσότερο για μεγάλες τιμές. Ωστόσο, υπενθυμίζουμε ότι, στην εν λόγω εφαρμογή, η παλινδρόμηση δεν έγινε απ' ευθείας ανάμεσα στη στάθμη και την παροχή, αλλά ανάμεσα στους λογαρίθμους τους. Η εξίσωση που δίνει το υπόλοιπο παλινδρόμησης σε αυτή την περίπτωση είναι $w_i = y_i - (a + b\ x_i)$, όπου $x$ και $y$ είναι οι (νεπέριοι) λογάριθμοι της στάθμης και παροχής, αντίστοιχα. Τα τελευταία υπόλοιπα παλινδρόμησης απεικονίζονται στο Σχ. 7.16, όπου είναι εμφανής η ομοσκεδαστικότητα των υπολοίπων. Το γεγονός αυτό επικυρώνει την ορθότητα του λογαριθμικού μετασχηματισμού, ο οποίος προηγήθηκε της παλινδρόμησης.




**Σχ. 7.15** Διάγραμμα για τον έλεγχο της ομοσκεδαστικότητας των υπολοίπων στην εκτίμηση της παροχής συναρτήσει της στάθμης στο Αχλαδόκαστρο του Ευήνου (βλ. Εφαρμογή 7.3.3).




**Σχ. 7.16** Διάγραμμα για τον έλεγχο της ομοσκεδαστικότητας των υπολοίπων παλινδρόμησης στην εκτίμηση του λογαρίθμου της παροχής συναρτήσει του λογαρίθμου της στάθμης στο Αχλαδόκαστρο του Ευήνου (βλ. Εφαρμογή 7.3.3).

## 7.5 Όρια εμπιστοσύνης και δοκιμές σημαντικότητας στη γραμμική παλινδρόμηση

Όλα όσα εκτέθηκαν στις προηγούμενες ενότητες αναφέρονται ουσιαστικά στην εξαγωγή σημειακών εκτιμήσεων της μεταβλητής $Y$ για δεδομένη τιμή της $X$. Επανερχόμενοι στην γραμμική παλινδρόμηση, στην ενότητα αυτή θα δώσουμε τα απαραίτητα στοιχεία για την εξαγωγή εκτιμήσεων διαστήματος (ορίων εμπιστοσύνης). Διευκρινίζεται ότι όσα αναφέρονται σε αυτή την ενότητα αφορούν στην τυπική γραμμική παλινδρόμηση δύο μεταβλητών και δεν εφαρμόζονται στις ειδικές μορφές (ομογενής ευθεία ή οργανική συσχέτιση) ούτε σε μη γραμμικές μορφές παλινδρόμησης.

Οι πηγές αβεβαιότητας στις εκτιμήσεις της παλινδρόμησης, οι οποίες παίρνονται υπόψη για την εξαγωγή των εκτιμήσεων διαστήματος είναι δύο. Η πρώτη είναι η εγγενής αβεβαιότητα, που εισάγεται από τον όρο σφάλματος $W$ της εξίσωσης παλινδρόμησης (7.24). Η πηγή αυτή αναιρείται μόνο όταν η διασπορά του σφάλματος (εξίσωση (7.35)) είναι μηδενική ή, ισοδύναμα, ο συντελεστής προσδιορισμού είναι 1. Η δεύτερη πηγή είναι η στατιστική αβεβαιότητα, η οποία προκύπτει από το πεπερασμένο μέγεθος του δείγματος $(x_i, y_i)$. Η αβεβαιότητα αυτή εισάγεται στην εκτίμηση των παραμέτρων $a$ και $b$ της εξίσωσης παλινδρόμησης και μεταφέρεται και στην εκτίμηση της $Y$. Αίρεται μόνο αν είναι γνωστές οι παράμετροι $\mu_X$, $\mu_Y$, $\sigma_X$, $\sigma_Y$ και $\sigma_{XY}$ του πληθυσμού, οι οποίες υπεισέρχονται στην εκτίμηση των $a$ και $b$ (εξισώσεις (7.28) και (7.27)).

Η εξαγωγή των εκτιμήσεων διαστήματος απαιτεί μια διαφορετική θεώρηση του προβλήματος της παλινδρόμησης, η οποία είναι γνωστή ως στατιστική θεώρηση. Σύμφωνα με αυτή το μοντέλο παλινδρόμησης γράφεται με τη μορφή

$$Y_i = a + bx_i + W \tag{7.79}$$

η οποία διαφέρει από την μορφή (7.24) στο ότι οι τιμές $x_i$ του δείγματος δεν θεωρούνται ως πραγματοποιήσεις μιας τυχαίας μεταβλητής $X$, αλλά ως γνωστές αριθμητικές τιμές. Στο πρόβλημα αυτό αναζητούνται οι εκτιμήτριες $\hat{A}$ και $\hat{B}$ των παραμέτρων του μοντέλου $a$ και $b$. Οι εκτιμήτριες αυτές είναι συναρτήσεις των τυχαίων μεταβλητών $Y_i$.

Δεν θα περιγράψουμε αναλυτικά τη διαδικασία εύρεσης των παραπάνω εκτιμητριών, την οποία ο ενδιαφερόμενος αναγνώστης μπορεί να βρει σε βιβλία στατιστικής (π.χ. Papoulis, 1990, σσ. 402-407· Benjamin and Cornell, 1970, σσ. 428-439). Θα δώσουμε μόνο τα αποτελέσματα της ανάλυσης σε πινακοποιημένη μορφή (Πίν. 7.5) και συγκεκριμένα τις μέσες τιμές και διασπορές των εκτιμητριών $\hat{A}$ και $\hat{B}$ καθώς και των μεταβλητών $\hat{Y}_x$ και $Y_x$, που ορίζονται από τις σχέσεις

$$\hat{Y}_x = \hat{A} + \hat{B}\,x \qquad Y_x = \hat{Y}_x + W = \hat{A} + \hat{B}\,x + W \tag{7.80}$$

Η $\hat{Y}_x$ εκφράζει τη μέση τιμή της εκτίμησης για δεδομένη τιμή $x$, ενώ η $Y_x$ εκφράζει μια απλή εκτίμηση της παλινδρόμησης, πάλι για δεδομένη τιμή $x$.

**Πίν. 7.5** Στατιστικά χαρακτηριστικά των εκτιμητριών της γραμμικής παλινδρόμησης.

| Μέγεθος προς εκτίμηση | Στατιστική συνάρτηση ή μεταβλητή | Μέση τιμή | Διασπορά |
|---|---|---|---|
| Κλίση ευθείας, $b$ | $\hat{B}$ | $\hat{b}$, όπως το $b$ στην εξίσωση (7.10) | $\sigma^2_{\hat{B}} = \dfrac{\sigma^2_W}{n\,s^2_X}$ |
| Περίλημμα ευθείας, $a$ | $\hat{A}$ | $\hat{a}$, όπως το $a$ στην εξίσωση (7.11) | $\sigma^2_{\hat{A}} = \dfrac{\sigma^2_W}{n}\left(1 + \dfrac{\overline{x}^2}{s^2_X}\right)$ |
| Μέση εκτίμηση (= τεταγμένη της ευθείας) για δεδομένο $x$ | $\hat{Y}_x$ | $\hat{y}_x = \hat{a} + \hat{b}\,x$ | $\sigma^2_{\hat{Y}_x} = \dfrac{\sigma^2_W}{n}\left[1 + \dfrac{(x-\overline{x})^2}{s^2_X}\right]$ |
| Απλή εκτίμηση για δεδομένο $x$ | $Y_x$ | $y_x = \hat{y}_x = \hat{a} + \hat{b}\,x$ | $\sigma^2_{Y_x} = \dfrac{\sigma^2_W}{n}\left[n + 1 + \dfrac{(x-\overline{x})^2}{s^2_X}\right]$ |

Στις εξισώσεις του Πίν. 7.5 τα μεγέθη $\overline{x}$ και $s_X^2$ ορίζονται ως εξής:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \tag{7.81}$$

δηλαδή ταυτίζονται αριθμητικά με τη δειγματική μέση τιμή και τη (μεροληπτική) δειγματική διασπορά των $x_i$.* Εξ άλλου το μέγεθος $\sigma_W^2$ είναι η θεωρητική διασπορά του σφάλματος $W$. Δεδομένου ότι το μέγεθος αυτό δεν είναι γνωστό, χρησιμοποιείται στη θέση του η αμερόληπτη εκτίμησή του από το δείγμα, η οποία είναι (Papoulis, 1990, σ. 405):

$$\hat{\sigma}_W^2 = \frac{1}{n-2}\sum_{i=1}^{n}[y_i - (\hat{a} + \hat{b}\, x_i)]^2 = \frac{n}{n-2}\, s_Y^2\,(1 - r_{XY}^2) \tag{7.82}$$

Στην τελευταία εξίσωση τα $s_Y^2$ και $r_{XY}$ εννοούνται ως μεροληπτικές εκτιμήσεις.

Ο υπολογισμός των ορίων εμπιστοσύνης για οποιοδήποτε από τα μεγέθη του Πίν. 7.5 (συμβολικά $\Xi$) βασίζεται στη συμβολική σχέση

$$\left(\hat{\xi} - t_{(1+\gamma)/2}\,\sigma_\Xi,\ \hat{\xi} + t_{(1+\gamma)/2}\,\sigma_\Xi\right) \tag{7.83}$$

όπου τα $\hat{\xi}$ και $\sigma_\Xi$ δίνονται, για κάθε μέγεθος, στην τρίτη και τέταρτη στήλη του Πίν. 7.5, αντίστοιχα, ενώ $t_{(1+\gamma)/2}$ είναι το $[(1+\gamma)/2]$-ποσοστημόριο της κατανομής Student για $n-2$ βαθμούς ελευθερίας και $\gamma$ είναι ο συντελεστής εμπιστοσύνης.

Οι παραπάνω τύποι μπορούν να χρησιμοποιηθούν και για την εκτέλεση στατιστικών ελέγχων, σχετικά με τις τιμές των παραμέτρων ή των εκτιμήσεων. Η εφαρμογή είναι άμεση και ακολουθεί τη μεθοδολογία της ενότητας 3.5. Ωστόσο, ο βασικότερος έλεγχος αφορά στη σημαντικότητα του συντελεστή γραμμικής συσχέτισης $\rho_{XY}$. Ο έλεγχος αυτός έχει περιγραφεί αναλυτικά στο εδάφιο 3.5.3.

---

* Γι' αυτό και χρησιμοποιούνται τα σύμβολα $\overline{x}$ και $s_X^2$, παρόλο που, όπως αναφέρθηκε παραπάνω, τα $x_i$ δεν εκλαμβάνονται στην εξεταζόμενη στατιστική προσέγγιση ως πραγματοποιήσεις μιας τυχαίας μεταβλητής $X$.

## Εφαρμογή 7.5

*Να υπολογιστούν τα όρια εμπιστοσύνης των συντελεστών a και b, και των εκτιμήσεων της απορροής με βάση τη βροχόπτωση για τα δεδομένα της Εφαρμογής 7.1, για συντελεστή εμπιστοσύνης 98%.*

Υπενθυμίζεται ότι τα στατιστικά χαρακτηριστικά των δειγμάτων της βροχόπτωσης ($X$) και της απορροής ($Y$) της Εφαρμογής 7.1 ήταν:

$$n = 25,\ \bar{x} = 1201.0,\ \bar{y} = 918.9,\ s_X = 252.5,\ s_Y = 234.6,\ r_{XY} = 0.911,$$

$$s_{XY} = r_{XY}\, s_X\, s_Y = 0.911 \times 252.5 \times 234.6 = 53\,964$$

Κατά συνέπεια, η τυπική απόκλιση των υπολοίπων (βλ. εξίσωση (7.82)) είναι

$$\hat{\sigma}_W = \sqrt{n/(n-2)}\, s_Y \sqrt{1 - r_{XY}^2}$$

$$= \sqrt{25/23} \times 234.6 \times \sqrt{1 - 0.911^2} = 100.9 \text{ mm}$$

Οι εκτιμήσεις των παραμέτρων της γραμμικής παλινδρόμησης είναι

$$\hat{b} = 0.846,\ \hat{a} = -97.1$$

Το $[(1+\gamma)/2]$ ποσοστημόριο της κατανομής Student για συντελεστή εμπιστοσύνης $\gamma = 0.98$ και βαθμούς ελευθερίας $25 - 2 = 23$ είναι 2.50 (βλ. Πίν. Π3 Παραρτήματος πινάκων).
Για την κλίση $b$ έχουμε

$$\hat{b} = 0.846,\ \sigma_{\hat{B}} = \sigma_W / (\sqrt{n}\, s_X) = 100.9 / (\sqrt{25} \times 252.5) = 0.0799$$

και επομένως τα όρια εμπιστοσύνης είναι

$$(0.846 - 2.50 \times 0.0799,\ 0.846 + 2.50 \times 0.0799) = (0.646,\ 1.045)$$

Για το περίλημμα $a$ έχουμε

$$\hat{a} = -97.1,\ \sigma_{\hat{A}} = (\sigma_W/\sqrt{n}) \sqrt{1 + \bar{x}^2 / s_X^2} =$$

$$= (100.9 / \sqrt{25}) \times \sqrt{1 + 1201.0^2 / 252.5^2} = 98.1$$

και επομένως τα όρια εμπιστοσύνης είναι

$$(-97.1 - 2.50 \times 98.1,\ -97.1 + 2.50 \times 98.1) = (-342.4,\ 148.2)$$

Για τη μέση εκτίμηση (= τεταγμένη της ευθείας παλινδρόμησης) για δεδομένο $x$ έχουμε

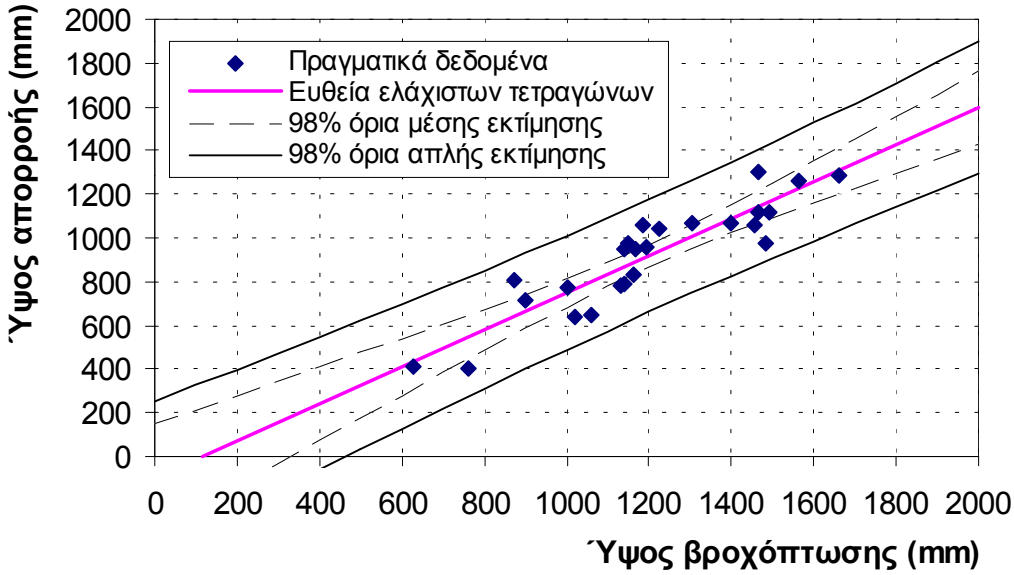
$$\hat{y}_x = \hat{a} + \hat{b}\, x = -97.1 + 0.846\, x$$

και

$$\sigma_{\hat{Y}_x} = (\sigma_W/\sqrt{n})\sqrt{1 + (x - \bar{x}^2)/s_X^2}$$
$$= (100.9/\sqrt{25}) \times \sqrt{1 + (x - 1201.0)^2/252.5^2}$$
$$= 20.2 \times \sqrt{1 + 1.568 \times 10^{-5} \times (x - 1201.0)^2}$$

και επομένως τα όρια εμπιστοσύνης είναι

$$-97.1 + 0.846\,x \pm 2.50 \times 20.2 \times \sqrt{1 + 1.568 \times 10^{-5} \times (x - 1201.0)^2}$$
$$= -97.1 + 0.846\,x \pm 50.5 \times \sqrt{1 + 1.568 \times 10^{-5} \times (x - 1201.0)^2}$$




**Σχ. 7.17** Όρια εμπιστοσύνης 98% της εκτίμησης του ετήσιου ύψους επιφανειακής απορροής συναρτήσει του ετήσιου ύψους βροχόπτωσης στη λεκάνη ανάντη Κρεμαστών.

Για την απλή εκτίμηση για δεδομένο $x$ έχουμε

$$y_x = \hat{a} + \hat{b}\,x = -97.1 + 0.846\,x$$

και

$$\sigma_{\hat{Y}_x} = (\sigma_W/\sqrt{n})\sqrt{n + 1 + (x - \bar{x}^2)/s_X^2}$$
$$= (100.9/\sqrt{25}) \times \sqrt{26 + (x - 1201.0)^2/252.5^2}$$
$$= 20.2 \times \sqrt{26 + 1.568 \times 10^{-5} \times (x - 1201.0)^2}$$

και επομένως τα όρια εμπιστοσύνης είναι

$$-97.1 + 0.846\,x \pm 50.5 \times \sqrt{26 + 1.568 \times 10^{-5} \times (x - 1201.0)^2}$$

Τα παραπάνω όρια εμπιστοσύνης της μέσης και της απλής εκτίμησης, τα οποία είναι συναρτήσεις της ανεξάρτητης μεταβλητής $x$, έχουν παρασταθεί γραφικά στο Σχ. 7.17. Παρατηρούμε ότι, όσο απομακρυνόμαστε από τις μέσες τιμές, τόσο διευρύνονται τα διαστήματα εμπιστοσύνης των εκτιμήσεων, πράγμα που σημαίνει ότι αυξάνει η αβεβαιότητα των εκτιμήσεων.

## 7.6 Συμπλήρωση και επέκταση δειγμάτων

Η τυπική χρήση της παλινδρόμησης στην τεχνική υδρολογία αφορά στη *συμπλήρωση* των ελλείψεων ενός υδρολογικού δείγματος με βάση ένα πληρέστερο δείγμα, το οποίο συσχετίζεται με το πρώτο, ή και την *επέκταση* του πρώτου δείγματος σε μια μεγαλύτερη διάρκεια που καλύπτεται από το δεύτερο δείγμα. Το πρώτο δείγμα θα το λέμε *δείγμα μελέτης* και το δεύτερο *δείγμα αναφοράς* ή *βάσης*. Στην ενότητα αυτή αναφερόμαστε σε ένα μοναδικό δείγμα αναφοράς, και κατά συνέπεια σε απλή παλινδρόμηση, αν και η μεθοδολογία που εφαρμόζεται μπορεί γενικά να περιλαμβάνει και περισσότερα από ένα δείγματα αναφοράς με πολλαπλή παλινδρόμηση.

Τα δείγματα μελέτης και αναφοράς μπορεί να είναι περιέχουν πρωτογενείς σημειακές μετρήσεις μιας φυσικής διεργασίας. Για παράδειγμα μπορεί να αποτελούνται από τις μετρήσεις της σημειακής βροχόπτωσης σε δύο γειτονικούς σταθμούς. Μπορεί επίσης να περιέχουν επεξεργασμένα δεδομένα της ίδιας φυσικής διεργασίας, όπως για παράδειγμα τα επιφανειακά ύψη βροχής δύο κοντινών υδρολογικών λεκανών. Τέλος, μπορεί να περιέχουν δεδομένα από διαφορετικές φυσικές διεργασίες, για τις οποίες υπάρχουν βάσιμοι λόγοι να συσχετίζονται. Για παράδειγμα μπορεί το δείγμα μελέτης να αποτελείται από την ετήσια απορροή μιας λεκάνης με σημαντική υδροφορία και το δείγμα αναφοράς να περιέχει την επιφανειακή βροχόπτωση της ίδιας λεκάνης, ή τη σημειακή βροχόπτωση ενός αντιπροσωπευτικού σταθμού της λεκάνης.

Η συμπλήρωση και η επέκταση είναι έννοιες πολύ κοντινές και δεν διακρίνονται με αυστηρό τρόπο μεταξύ τους. Συνήθως χρησιμοποιούμε τον πρώτο όρο σε περιπτώσεις που το δείγμα μελέτης παρουσιάζει λίγα και σποραδικά κενά (π.χ. 1-3), τα οποία πρέπει να συμπληρωθούν, και το δεύτερο για περιπτώσεις που το δείγμα μελέτης εμφανίζει συστηματικές,

συνήθως πολυετείς, ελλείψεις κατά τη διάρκεια της λειτουργίας του, ή η λειτουργία του εκτείνεται σε σημαντικά μικρότερο διάστημα σε σχέση με αυτό του σταθμού αναφοράς. Αυτό που επιδιώκουμε με την επέκταση και συμπλήρωση είναι η απόκτηση ενός δείγματος μελέτης με μεγαλύτερο μήκος από το αρχικό και κατά συνέπεια μεγαλύτερη αξιοπιστία εκτιμήσεων. Θα πρέπει όμως από τώρα να ξεκαθαρίσουμε ότι η αξιοπιστία του διευρυμένου δείγματος δεν αντιστοιχεί στο ονομαστικό διευρυμένο μήκος του, αλλά σε ένα ενδιάμεσο μήκος ανάμεσα στο αρχικό και το διευρυμένο (μόνο οι άμεσες μετρήσεις δίνουν αξιοπιστία αντίστοιχη με το μήκος). Για παράδειγμα αν έχουμε ένα παρατηρημένο δείγμα μελέτης με 20 ετήσιες τιμές και το επεκτείνουμε, με βάση ένα δείγμα αναφοράς, για άλλα 10 χρόνια, τότε η αξιοπιστία του νέου δείγματος δεν είναι ίδια με αυτή που θα είχε ένα παρατηρημένο (με μετρήσεις) δείγμα 30 ετών, αλλά θα αντιστοιχεί σε ένα παρατηρημένο δείγμα ισοδύναμου μήκους $n'$ όπου $20 \le n' \le 30$. Είναι προφανές ότι, όσο πιο έντονη είναι η συσχέτιση του δείγματος μελέτης με το δείγμα αναφοράς, τόσο πιο πολύ κοντά στο 30 θα είναι το ισοδύναμο μήκος.

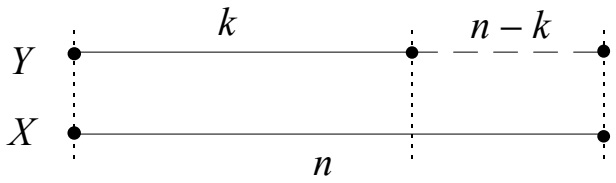
Για τον ίδιο λόγο, οι εκτιμήσεις των στατιστικών χαρακτηριστικών ενός διευρυμένου δείγματος δεν μπορεί να γίνονται με τους απλούς τύπους που χρησιμοποιούμε για ένα παρατηρημένο δείγμα. Το θέμα απαιτεί μεγάλη προσοχή, ιδίως στις περιπτώσεις σημαντικού μήκους επέκτασης δείγματος, οπότε και τα σφάλματα, αν η εκτίμηση γίνεται με τους τύπους του παρατηρημένου δείγματος, είναι μεγάλα.

Η ενότητα αυτή καλύπτει κατά βάση την περίπτωση που η συμπλήρωση η επέκταση των τιμών της $Y$ γίνεται με μια γραμμική εξίσωση ως προς $X$, σε μία από τις διάφορες μορφές που έχουν εξεταστεί προηγουμένως. Μπορούμε να θεωρήσουμε ότι έμμεσα καλύπτει και τις περιπτώσεις γραμμικοποιημένων εξισώσεων μέσω μετασχηματισμών των μεταβλητών (εδάφιο 7.3.3), δεδομένου ότι αυτές ανάγονται σε γραμμικές εξισώσεις. Η γραμμική μορφή εξίσωσης είναι η πιο διαδεδομένη, όχι μόνο λόγω της απλότητάς της. Όπως είδαμε στο εδάφιο 7.4.2, η γραμμική παλινδρόμηση είναι η βέλτιστη παλινδρόμηση για μεταβλητές που ακολουθούν κανονική κατανομή. Μάλιστα, όπως είδαμε και σε άλλα κεφάλαια, στην τεχνική υδρολογία η κανονική κατανομή είναι πολύ συχνά κατάλληλη για μεταβλητές που αναφέρονται σε μεγάλες χρονικές κλίμακες, ώστε σε κάθε χρονικό διάστημα να αντιστοιχεί μεγάλος αριθμός

υδρολογικών επεισοδίων. Η καταλληλότητα της κανονικής κατανομής σε αυτές τις περιπτώσεις εξηγείται από το κεντρικό οριακό θεώρημα.

Ωστόσο, η χρήση της γραμμικής παλινδρόμησης δεν περιορίζεται στην κανονική κατανομή. Δεν υπάρχει λόγος που να αποκλείει τη γραμμική παλινδρόμηση από μη κανονικές μεταβλητές. Αντίθετα, και σε μη κανονικές μεταβλητές, αν δεν υπάρχει θεωρητικός λόγος που να μας οδηγεί σε άλλο τύπο παλινδρόμησης, ξεκινάμε την ανάλυσή μας δοκιμάζοντας την απλούστερη γραμμική παλινδρόμηση. Προηγουμένως, ελέγχουμε το συντελεστή γραμμικής συσχέτισης, ο οποίος θα πρέπει να είναι σημαντικά (με τη στατιστική έννοια) διάφορος του μηδενός. Εκ των υστέρων ελέγχουμε την ομοσκεδαστικότητα των υπολοίπων (βλ. αντίστοιχο εδάφιο), η οποία, αν υπάρχει, αποτελεί ένδειξη για την επιτυχία της επιλογής του συγκεκριμένου τύπου παλινδρόμησης.

Στα επόμενα εδάφια θα αναφερθούμε διεξοδικά στις διάφορες μεθόδους συμπλήρωσης, στα αντίστοιχα προβλήματα και στον τρόπο αντιμετώπισής τους, ενώ στην Εφαρμογή 7.6 δίνουμε ένα πλήρες παράδειγμα για τη χρήση των διάφορων μεθόδων. Σε όλες τις περιπτώσεις χρησιμοποιούμε το συμβολισμό που διευκρινίζεται στο Σχ. 7.18. Θεωρούμε, δηλαδή, ότι το δείγμα της μεταβλητής $Y$ διαθέτει $k$ παρατηρημένα δεδομένα, ενώ το δείγμα της μεταβλητής $X$ διαθέτει $n > k$ παρατηρημένα δεδομένα. Επίσης θεωρούμε ότι για τις $k$ μετρήσεις της μεταβλητής $Y$ υπάρχουν οι ταυτόχρονες μετρήσεις της $X$. Το πρόβλημα που μελετάμε είναι η εκτίμηση των $(n - k)$ τιμών της $Y$ με βάση τις αντίστοιχες της $X$.




**Σχ. 7.18** Διευκρινιστικό σκαρίφημα για το συμβολισμό στην επέκταση δείγματος.

Επίσης, τα στατιστικά χαρακτηριστικά που αναφέρονται στην κοινή περίοδο, μεγέθους $k$, συμβολίζονται με άτονα γράμματα, π.χ. τα $\overline{x}$ και $\overline{y}$ συμβολίζουν τις δειγματικές μέσες τιμές των $X$ και $Y$, εκτιμημένες από τα δείγματα μεγέθους $k$, και $s^{*X}_2$ και $s^{*Y}_2$ τις αντίστοιχες αμερόληπτες δειγματικές διασπορές. Τα αντίστοιχα μεγέθη για την ολική περίοδο μεγέθους $n$

συμβολίζονται με τονούμενα γράμματα, π.χ. με $\bar{x}'$ και $\bar{y}'$ συμβολίζονται οι εκτιμήσεις των μέσων τιμών από τα δείγματα μεγέθους $n$, με $s'^2_{*X}$ και $s'^2_{*Y}$ οι αντίστοιχες εκτιμήσεις διασπορών κτλ.

### 7.6.1 Χρήση της γραμμικής παλινδρόμησης χωρίς όρο σφάλματος

Η συμπλήρωση ή επέκταση ενός δείγματος $Y$, με βάση το αντίστοιχο πληρέστερο δείγμα της μεταβλητής $X$, γίνεται τις περισσότερες φορές με βάση την εξίσωση (7.25) ($\hat{Y} = a + bX$), στην οποία δεν παίρνεται υπόψη ο όρος σφάλματος $W$. Η μέθοδος αυτή οδηγεί, όπως είδαμε, στην ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος και γι' αυτό είναι η πιο διαδεδομένη. Ωστόσο έχει σημαντικά μειονεκτήματα. Το διευρυμένο δείγμα που παράγεται με εφαρμογή της μεθόδου αυτής δίνει αμερόληπτη εκτίμηση της μέσης τιμής αλλά μεροληπτική εκτίμηση της διασποράς. Συγκεκριμένα, η χρήση του τύπου της δειγματικής διασποράς παρατηρημένου δείγματος (εξίσωση 3.15) υπεκτιμά τη διασπορά της $Y$. Η διασπορά αυτή θα πρέπει να διορθωθεί με πολλαπλασιασμό επί το συντελεστή

$$\varphi = \frac{n-1}{k + (n-k)\,\rho_{XY}^2 - 1} \tag{7.84}$$

ο οποίος είναι μεγαλύτερος ή ίσος του 1. Συγκεκριμένα, ο συντελεστής διόρθωσης παίρνει την τιμή 1 μόνο όταν $|\rho_{XY}| = 1$, ενώ για $\rho_{XY} = 0$ παίρνει τη μέγιστη τιμή του $(n - 1) / (k - 1)$. Η παραγωγή της εξίσωσης (7.84) δίνεται αναλυτικά στο Παράρτημα 7Α. Παρατηρούμε ότι στην εξίσωση υπεισέρχεται η θεωρητική τιμή του συντελεστή συσχέτισης $\rho_{XY}$, η οποία όμως στην πραγματικότητα είναι άγνωστη. Στην εφαρμογή μπορεί να χρησιμοποιηθεί στη θέση της η δειγματική τιμή $r_{XY}$ που υπολογίζεται από τις $k$ ταυτόχρονες μετρήσεις των $X$ και $Y$. Δεν είναι σωστό να χρησιμοποιηθεί η τιμή του συντελεστή συσχέτισης που προκύπτει από τα $n$ δεδομένα, μετά τη συμπλήρωση των $n - k$ τιμών της $X$, γιατί και σε αυτή την τιμή υπάρχει σοβαρή μεροληψία. Συγκεκριμένα η τιμή αυτή είναι σημαντικά υπερεκτιμημένη.

Η μέση τιμή και η διασπορά του διευρυμένου δείγματος μπορούν να εκτιμηθούν και από τις ακόλουθες εξισώσεις, οι οποίες δεν κάνουν χρήση των εκτιμημένων $n - k$ δεδομένων της $Y$:

$$\bar{y}' = \bar{y} + b(\bar{x}' - \bar{x}) \tag{7.85}$$

$$s'^{2}_{*Y} = s^{2}_{*Y} + b^2 \left( s'^{2}_{*X} - s^{2}_{*X} \right) \tag{7.86}$$

Εύκολα διαπιστώνουμε ότι οι παραπάνω εκτιμήσεις είναι αμερόληπτες εφόσον είναι γνωστή η θεωρητική τιμή της παραμέτρου $b = \rho_{XY}\, \sigma_Y / \sigma_X$. Στην πραγματικότητα, βέβαια, αυτή δεν είναι γνωστή, οπότε χρησιμοποιείται στη θέση της η αντίστοιχη εκτίμηση από τα δείγματα μεγέθους $k$, $\hat{b} = r_{XY}\, s_{*Y} / s_{*X}$. Αυτό εισάγει κάποια μεροληψία στις εκτιμήσεις, η οποία πάντως δεν είναι σημαντική.

Αντίστοιχα, μια κατά προσέγγιση αμερόληπτη εκτίμηση για το συντελεστή συσχέτισης, η οποία αναφέρεται στην περίοδο μήκους $n$, είναι

$$r'_{XY} = r_{XY} \frac{s_{*Y}}{s_{*X}} \frac{s'_{*X}}{s'_{*Y}} \tag{7.87}$$

όπου η τυπική απόκλιση $s'_{*Y}$ θα πρέπει να υπολογιστεί με βάση την (7.86). Η εξαγωγή των εξισώσεων (7.85) έως (7.87) δίνεται αναλυτικά στο Παράρτημα 7Α.

Προκειμένου να αποκτήσουμε μια εικόνα του βαθμού βελτίωσης της εκτίμησης της μέσης τιμής στο διευρυμένο δείγμα σε σχέση με την αρχική, μας χρειάζεται η διασπορά της εκτιμήτριας της μέσης τιμής του διευρυμένου δείγματος. Στο Παράρτημα 7Α αποδεικνύεται ότι αυτή δίνεται από τον τύπο

$$\mathrm{Var}[\bar{Y}'] = \frac{\sigma_Y^2}{k} \left( 1 - \frac{n-k}{n} \rho_{XY}^2 \right) \tag{7.88}$$

Παίρνοντας υπόψη ότι η διασπορά της αρχικής εκτίμησης είναι

$$\mathrm{Var}[\bar{Y}] = \frac{\sigma_Y^2}{k} \tag{7.89}$$

η βελτίωση στην εκτίμηση δίνεται από το λόγο

$$\frac{\mathrm{Var}[\bar{Y}']}{\mathrm{Var}[\bar{Y}]} = 1 - \frac{n-k}{n}\rho_{XY}^2 \tag{7.90}$$

ο οποίος παίρνει τη βέλτιστη (ελάχιστη) τιμή του $k / n$ για $|\rho_{XY}| = 1$, ενώ είναι ίσος με 1 (που σημαίνει μηδενική βελτίωση) για $\rho_{XY} = 0$.

Στην εξίσωση (7.88), όπως και σε προηγούμενες εξισώσεις, υποτίθεται ότι είναι γνωστή η θεωρητική τιμή του συντελεστή συσχέτισης $\rho_{XY}$. Στην πραγματικότητα είναι γνωστή η εκτίμηση $r_{XY}$. Για αυτή την περίπτωση ο τύπος είναι ελαφρά διαφοροποιημένος (βλ. Yevjevich, 1972, σ. 260· Salas, 1993, σ. 19.44):

$$\mathrm{Var}[\bar{Y}'] = \frac{\sigma_Y^2}{k}\left[1 - \frac{n-k}{n}\left(r_{XY}^2 - \frac{1-r_{XY}^2}{k-3}\right)\right] \tag{7.91}$$

οπότε η βελτίωση στην εκτίμηση δίνεται από το λόγο

$$\frac{\mathrm{Var}[\bar{Y}']}{\mathrm{Var}[\bar{Y}]} = 1 - \frac{n-k}{n}\left(r_{XY}^2 - \frac{1-r_{XY}^2}{k-3}\right) \tag{7.92}$$

Γενικά θεωρείται ότι το ισοδύναμο μήκος του διευρυμένου δείγματος $n'$ δίνεται από τη σχέση

$$\frac{k}{n'} = \frac{\mathrm{Var}[\bar{Y}']}{\mathrm{Var}[\bar{Y}]} \tag{7.93}$$

Από την εξίσωση (7.92) προκύπτει ότι, για να είναι ο λόγος $\mathrm{Var}[\bar{Y}'] / \mathrm{Var}[\bar{Y}]$ μικρότερος από 1 και, κατά συνέπεια, για να υπάρχει βελτίωση στην εκτίμηση της μέσης τιμής, θα πρέπει

$$|r_{XY}| > \frac{1}{\sqrt{k-2}} \tag{7.94}$$

Η τελευταία ανισότητα δίνει μια εκτίμηση του κάτω ορίου του συντελεστή συσχέτισης για να έχει νόημα η επέκταση του δείγματος με γραμμική παλινδρόμηση.

Κατά μία άλλη λογική, την οποία συστήνουμε, για να έχει νόημα η παλινδρόμηση, θα πρέπει ο συντελεστής συσχέτισης $r_{XY}$ να είναι σημαντικά διάφορος του μηδενός. Θεωρώντας επίπεδο σημαντικότητας $\alpha = 5\%$, η λογική αυτή οδηγεί στην ισχυρότερη ανισότητα

$$|r_{XY}| \gtrsim \frac{2}{\sqrt{k}} \tag{7.95}$$

Η τελευταία ανισότητα έχει παραχθεί στο εδάφιο 3.5.3 (βλ. εξίσωση 3.74).

### 7.6.2 Χρήση της οργανικής συσχέτισης

Η χρήση της οργανικής συσχέτισης είναι ίδια όπως της γραμμικής παλινδρόμησης χωρίς όρο σφάλματος. Η διαφορά εδώ είναι ότι η παράμετρος $b$ εκτιμάται από την (7.54) αντί της (7.27).

Όπως έχουμε αναφέρει προηγουμένως, η μέθοδος της οργανικής συσχέτισης δεν οδηγεί σε μεροληψία στις εκτιμήσεις των μέσων τιμών και διασπορών της $Y$ και κατά συνέπεια αυτές μπορούν να υπολογίζονται κανονικά από το διευρυμένο δείγμα, χωρίς να απαιτούνται διορθώσεις. Ωστόσο, η μέθοδος εισάγει σοβαρή μεροληψία στην εκτίμηση του συντελεστή συσχέτισης και γι' αυτό ο τελευταίος δεν πρέπει να υπολογίζεται άμεσα από το διευρυμένο δείγμα. Μπορεί να χρησιμοποιηθεί για τον προσεγγιστικό υπολογισμό του η εξίσωση (7.87).

Κατ' αναλογία με όσα αναφέρθηκαν στο προηγούμενο εδάφιο (εξίσωση (7.95)) και παίρνοντας υπόψη και όσα αναφέρθηκαν στο εδάφιο 7.2.2 (εξίσωση (7.61)) οδηγούμαστε στο συμπέρασμα ότι για να έχει νόημα η εφαρμογή της οργανικής συσχέτισης θα πρέπει

$$|r_{XY}| \gtrsim \max\left\{0.5, \frac{2}{\sqrt{k}}\right\} \tag{7.96}$$

### 7.6.3 Χρήση της ομογενούς ευθείας

Η χρήση της ομογενούς ευθείας για επέκταση δείγματος θα πρέπει γενικά να αποφεύγεται, διότι, όπως ήδη αναφέρθηκε, οδηγεί σε μεροληψία και ως προς τη μέση τιμή και ως προς τη διασπορά.

Ωστόσο, σε περιπτώσεις συμπλήρωσης λίγων τιμών, οι οποίες δεν επηρεάζουν τα στατιστικά χαρακτηριστικά ενός δείγματος, είναι πιθανό η ομογενής ευθεία να είναι πλεονεκτική. Ας θεωρήσουμε σαν παράδειγμα το πρόβλημα της συμπλήρωσης λίγων ελλείψεων ενός πολυετούς δείγματος ημερήσιων βροχοπτώσεων. Αν χρησιμοποιηθεί η τυπική γραμμική παλινδρόμηση ή η οργανική συσχέτιση, τότε είναι δυνατό να προκύψουν μερικές αρνητικές τιμές, επειδή ο σταθερός όρος $a$ της γραμμικής εξίσωσης μπορεί να είναι αρνητικός. Με την ομογενή ευθεία αποφεύγουμε αυτό το πρόβλημα. Πάντως και σε αυτή την περίπτωση είναι εξεταστέα και η εφαρμογή της εξίσωσης δύναμης ($Y = \kappa X^{\lambda}$, βλ. εδάφιο 7.3.3), η οποία αποφεύγει επίσης τις αρνητικές τιμές, διατηρώντας παράλληλα τη μέση τιμή του λογαρίθμου της μεταβλητής.

### 7.6.4 Χρήση της γραμμικής παλινδρόμησης με όρο σφάλματος

Όπως είδαμε στα προηγούμενα εδάφια, τόσο η τυπική γραμμική παλινδρόμηση, όσο και η οργανική συσχέτιση, παρουσιάζουν διάφορα προβλήματα, τα οποία οδηγούν σε μεροληπτικές εκτιμήσεις διάφορων στατιστικών χαρακτηριστικών του διευρυμένου δείγματος. Η "πηγή" όλων αυτών των προβλημάτων είναι το γεγονός ότι τα σημεία $(x_i, y_i)$ του διευρυμένου δείγματος βρίσκονται όλα πάνω στην ευθεία παλινδρόμησης (ή οργανικής συσχέτισης, βλ. Σχ. 7.19), γεγονός που αντίκειται στη φυσική πραγματικότητα. Τα σημεία $(x_i, y_i)$ που προέρχονται από μετρήσεις ποτέ δεν είναι συνευθειακά (εκτός της ακραίας περίπτωσης όπου $|\rho_{XY}| = 1$). Φυσικά αυτό το πρόβλημα δεν οφείλεται στη γραμμικότητα της εξίσωσης. Και σε μη γραμμική εξίσωση παλινδρόμησης το πρόβλημα θα υπάρχει, μόνο που αντί να είναι συνευθειακά τα σημεία της επέκτασης θα ανήκουν σε μια λεία καμπύλη. Το πρόβλημα προκύπτει από την αγνόηση του σφάλματος γύρω από το γραμμικό (ή μη) νόμο, που στη φύση πάντα υπάρχει.

Οι διάφορες μέθοδοι που εκτέθηκαν παραπάνω για την άρση των σφαλμάτων στον υπολογισμό των στατιστικών χαρακτηριστικών, είναι ικανοποιητικές μεν σε ότι αφορά αυτά τα ίδια τα στατιστικά χαρακτηριστικά, αλλά δεν λύνουν κάθε συνέπεια της μη ρεαλιστικής διάταξης των σημείων. Πολύ συχνά ένα δείγμα μιας μεταβλητής επεκτείνεται προκειμένου να χρησιμοποιηθεί ως (πιο επαρκής) είσοδος σε ένα άλλο μοντέλο, το οποίο πραγματοποιεί κάποιο μετασχηματισμό της εισόδου αυτής (π.χ.

μοντέλο υδατικού ισοζυγίου, μοντέλο βροχής-απορροής, μοντέλο λειτουργίας ταμιευτήρα κτλ.). Στην περίπτωση αυτή δεν ενδιαφέρουν μόνο τα συγκεκριμένα στατιστικά χαρακτηριστικά της μεταβλητής, αλλά όλη η στατιστική δομή του δείγματος, η οποία ασφαλώς έχει διαταραχθεί με τη διεύρυνση.

Η μόνη ρεαλιστική λύση για αυτές τις περιπτώσεις, που ενδιαφέρει η όλη δομή του δείγματος, είναι η χρήση της πλήρους μορφής της γραμμικής εξίσωσης (7.24) που συμπεριλαμβάνει και τον όρο σφάλματος $W$ ($Y = a + bX + W$). Στην εφαρμογή αυτής της εξίσωσης ο όρος σφάλματος *δεν εκτιμάται* (η εκτίμησή του θα ήταν ίση π.χ. με τη μέση τιμή του, δηλαδή μηδέν) αλλά *γεννάται* ή *προσομοιώνεται*. Χρησιμοποιείται, δηλαδή, μια *γεννήτρια τυχαίων αριθμών*, η οποία γεννά τις απαραίτητες $n - k$ τυχαίες τιμές του σφάλματος.

Μια ακολουθία αριθμών $x_i$ λέγεται *ακολουθία τυχαίων αριθμών δεδομένης κατανομής* $F(x)$ αν αποτελεί δείγμα της τυχαίας μεταβλητής $X$, η οποία έχει συνάρτηση κατανομής $F(x)$. Για κάθε συνάρτηση κατανομής μπορεί να κατασκευαστεί μία γεννήτρια τυχαίων αριθμών (ή και περισσότερες). Η γεννήτρια είναι ένας αλγόριθμος, συνήθως αναδρομικός, ο οποίος μπορεί να παράγει διαδοχικά οσουσδήποτε όρους της τυχαίας ακολουθίας. Στην πράξη βέβαια υπάρχει ένα αρκετά μεγάλο, αλλά πάντως πεπερασμένο, όριο τυχαίων αριθμών που μπορεί να δώσει η ακολουθία. Πάνω από αυτό το όριο γίνεται επανάληψη των ίδιων αριθμών, δηλαδή η ακολουθία γίνεται περιοδική. Πάντως σε κάθε περίπτωση οι τυχαίοι αριθμοί δεν γεννώνται στην τύχη ο ένας μετά τον άλλον, αλλά βάσει ενός αυστηρά προσδιοριστικού αλγορίθμου, ο οποίος οδηγεί στην ίδια ακολουθία αριθμών, αν ξεκινήσει με τις ίδιες αρχικές συνθήκες. Για το λόγο αυτό τους τυχαίους αριθμούς μερικοί τους ονομάζουν ψευδοτυχαίους. Πάντως, αν αλλάξουμε τις αρχικές συνθήκες παίρνουμε άλλη τυχαία ακολουθία (ακριβέστερα άλλο τμήμα της ίδιας περιοδικής ακολουθίας). Στο επόμενο ένθετο εδάφιο δίνονται γεννήτριες τυχαίων αριθμών με ομοιόμορφη και με κανονική κατανομή. Η γεννήτρια της κανονικής κατανομής μπορεί να καλύψει τη συχνότερη περίπτωση επέκτασης δειγμάτων που ακολουθούν κανονική κατανομή (αν οι μεταβλητές $X$ και $Y$ ακολουθούν κανονική κατανομή τότε και η $W$ ακολουθεί κανονική κατανομή). Ο αναγνώστης που ενδιαφέρεται για γεννήτριες άλλων

συναρτήσεων κατανομής παραπέμπεται π.χ. στον Papoulis, 1990, σσ. 251-372.

Ανακεφαλαιωτικά, στην περίπτωση που στη διεύρυνση ενός δείγματος λαμβάνουμε υπόψη το τυχαίο σφάλμα, δεν κάνουμε εκτίμηση των (βέλτιστων) τιμών της $Y$ αλλά προσομοίωσή τους. Αυτή η διαδικασία έχει ένα πολύ σοβαρό μειονέκτημα: δεν δίνει μοναδικές τιμές $y_i$ για τις μεταβλητές που προσομοιώνει. Με εφαρμογή άλλης γεννήτριας τυχαίων αριθμών ή της ίδιας γεννήτριας με διαφορετικές αρχικές συνθήκες θα προκύψουν άλλες τιμές $y_i$. Ωστόσο, έχει το σοβαρό πλεονέκτημα να μη διαταράσσει τη στατιστική δομή του δείγματος. Έτσι, οι εκτιμήσεις της μέσης τιμής, της διασποράς, του συντελεστή συσχέτισης κτλ., μπορούν να γίνουν από το διευρυμένο δείγμα με εφαρμογή των κοινών τύπων του παρατηρημένου δείγματος, χωρίς καμιά διόρθωση. Βέβαια, ανακύπτει και στην εκτίμηση αυτών των στατιστικών χαρακτηριστικών το πρόβλημα ότι αυτά δεν ορίζονται με μονοσήμαντο τρόπο.

Συμπερασματικά, δεν υπάρχει καμιά τέλεια μέθοδος για τη συμπλήρωση και επέκταση ενός δείγματος. Οι τρεις μέθοδοι που εξετάστηκαν αναλυτικά παραπάνω, δηλαδή η εκτίμηση με τυπική γραμμική παλινδρόμηση χωρίς τυχαίο σφάλμα, η εκτίμηση με οργανική συσχέτιση, και η προσομοίωση με γραμμική παλινδρόμηση και συνυπολογισμό του τυχαίου σφάλματος, έχουν πλεονεκτήματα και μειονεκτήματα. Η επιλογή της κατάλληλης κατά περίπτωση μεθόδου διεύρυνσης του δείγματος εναπόκειται στο μελετητή υδρολόγο.

> **Γεννήτριες τυχαίων αριθμών ομοιόμορφης και κανονικής κατανομής**
>
> Παραθέτουμε εδώ χωρίς απόδειξη τους πιο διαδεδομένους αλγορίθμους για τη γέννηση τυχαίων αριθμών ομοιόμορφης και κανονικής κατανομής. Ο αναγνώστης που ενδιαφέρεται για την τεκμηρίωση των αλγορίθμων αυτών παραπέμπεται στον Papoulis, 1990, σσ. 251-372. Οι τυχαίοι αριθμοί ομοιόμορφης κατανομής αποτελούν τη βάση για τη γέννηση τυχαίων αριθμών από άλλες κατανομές. Σημειώνουμε ότι όλοι οι μεταφραστές (compilers) προγραμμάτων ηλεκτρονικών υπολογιστών διαθέτουν ενσωματωμένη γεννήτρια τυχαίων αριθμών ομοιόμορφης κατανομής, η οποία μπορεί να χρησιμοποιηθεί αντί αυτής που δίνεται εδώ. Επίσης, πολλά έτοιμα προγράμματα υπολογισμών (λογιστικά φύλλα κτλ.) διαθέτουν ενσωματωμένη τέτοια γεννήτρια, καθώς και γεννήτριες για άλλες κατανομές.

*α. Γεννήτρια τυχαίων αριθμών ομοιόμορφης κατανομής*
Οι ακέραιοι αριθμοί $q_i$ που υπολογίζονται από τον αναδρομικό τύπο

$$q_i = (k\, q_{i-1} + c) \bmod m \tag{7.97}$$

όπου $k$, $c$ και $m$ κατάλληλες ακέραιες σταθερές, αποτελούν τυχαίους αριθμούς ομοιόμορφα κατανεμημένους στο διάστημα [1, $m$ – 1]. Στην παραπάνω εξίσωση ο συμβολισμός $\alpha = \beta \bmod \gamma$ σημαίνει ότι ο αριθμός α είναι το υπόλοιπο της διαίρεσης του ακέραιου αριθμού $\beta$ με τον ακέραιο αριθμό $\gamma$. Κατά συνέπεια η ακολουθία αριθμών

$$u_i = q_i / m \tag{7.98}$$

αποτελεί πρακτικά ακολουθία τυχαίων αριθμών συνεχούς τύπου στο διάστημα (0, 1).
Οι ακέραιοι αριθμοί $k$ και $m$ επιλέγονται έτσι ώστε η περιοδικότητα της ακολουθίας των $q_i$ να είναι κατά το δυνατό μεγαλύτερη (δηλαδή $m$ – 1). Μια επιτυχημένη επιλογή είναι η

$$k = 16\,807,\ c = 0,\ \ m = 2^{31} - 1 = 2\,147\,483\,647 \tag{7.99}$$

Ωστόσο, η επιλογή αυτή δύσκολα υλοποιείται σε συνήθεις υπολογιστές, οι οποίοι εκτελούν πράξεις ακεραίων μεγέθους μέχρι 32 bits. Μια άλλη επιλογή, που υλοποιείται σε υπολογιστές με αριθμητική ικανότητα 32 bits, είναι η

$$k = 4096,\ c = 150\,889,\ \ m = 714\,025 \tag{7.100}^*$$

Η αρχική τιμή $q_0$ που χρειάζεται για να λειτουργήσει η γεννήτρια μπορεί να είναι οποιοσδήποτε ακέραιος μεταξύ του 1 και του $m$ – 1.

*α. Γεννήτρια τυχαίων αριθμών κανονικής κατανομής*
Αν οι αριθμοί $u_i$ και $v_i$ είναι διαδοχικοί όροι ακολουθίας τυχαίων αριθμών με ομοιόμορφη κατανομή στο διάστημα (0, 1), τότε αποδεικνύεται ότι οι αριθμοί

$$w_i = \sqrt{-2 \ln v_i}\, \cos \pi(2 - u_i),\ z_i = \sqrt{-2 \ln v_i}\, \sin \pi(2 - u_i) \tag{7.101}$$

αποτελούν διαδοχικούς όρους ακολουθίας τυχαίων αριθμών με κανονική κατανομή $N$(0, 1). Κατά συνέπεια οι αριθμοί

$$w'_i = w_i\, \sigma + \mu, \qquad z'_i = z_i\, \sigma + \mu \tag{7.102}$$

---

* Για περισσότερες επιλογές καθώς και για αλγορίθμους κατάλληλους για υπολογιστές με μικρότερη αριθμητική ακρίβεια βλ. Press et al., 1987, σ. 198.

αποτελούν διαδοχικούς όρους ακολουθίας τυχαίων αριθμών με κανονική κατανομή $N(\mu, \sigma)$. Έτσι μια οποιαδήποτε ακολουθία τυχαίων αριθμών ομοιόμορφης κατανομής μπορεί να μετασχηματιστεί εύκολα σε (ισοπληθή) ακολουθία τυχαίων αριθμών κανονικής κατανομής.

## Εφαρμογή 7.6

*Στη δεύτερη και τρίτη στήλη του Πίν. 7.6 δίνονται τα ετήσια ύψη βροχής Χ και Υ δύο κοντινών βροχομετρικών σταθμών της Δυτικής Ελλάδας για περίοδο 40 ετών. Για λόγους διερεύνησης και σύγκρισης, να αγνοηθούν τα παρατηρημένα δεδομένα της δεύτερης εικοσαετίας του δείγματος Υ και στη συνέχεια, με βάση τα δεδομένα της πρώτης εικοσαετίας να γίνει επέκταση του δείγματος Υ με τις μεθόδους (α) γραμμικής παλινδρόμησης χωρίς τυχαίο σφάλμα, (β) οργανικής συσχέτισης και (γ) γραμμικής παλινδρόμησης με τυχαίο σφάλμα. Να εκτιμηθούν τα στατιστικά χαρακτηριστικά της Υ και ο συντελεστής συσχέτισης των Χ και Υ για τις διάφορες περιπτώσεις επέκτασης.*

Τα στατιστικά χαρακτηριστικά των δειγμάτων της πρώτης εικοσαετίας είναι:

$$k = 20,\ \overline{x} = 1492.4,\ \overline{y} = 1116.5,\ s_{*X} = 357.3,\ s_{*Y} = 301.2,\ r_{XY} = 0.644$$

Τα στατιστικά χαρακτηριστικά των συνολικών δειγμάτων των 40 ετών είναι:

$$n = 40,\ \overline{x}' = 1513.8,\ \overline{y}' = 1163.4,\ s'_{*X} = 359.6,\ s'_{*Y} = 317.3,\ r'_{XY} = 0.691$$

Εφαρμόζουμε κατ' αρχήν την τυπική γραμμική παλινδρόμηση. Οι παράμετροι της γραμμικής εξίσωσης $y = a + b\,x$, υπολογισμένες από τα δείγματα της πρώτης εικοσαετίας (εξ. (7.27) και (7.28)), είναι:

$$b = 0.644 \times 301.2 / 357.3 = 0.543$$

$$a = 1116.5 - 0.543 \times 1492.4 = 306.1$$

Εφαρμόζοντας τη γραμμική εξίσωση αυτή για τα δεδομένα $x$ της δεύτερης εικοσαετίας, παίρνουμε τις εκτιμήσεις που φαίνονται στη στήλη $Y^{(1)}$ του Πίν. 7.6). Αν τώρα υπολογίσουμε τα στατιστικά χαρακτηριστικά των 40 ετών της στήλης $Y^{(1)}$ βρίσκουμε

$$\overline{y}' = 1128.1,\ s'_{*Y} = 253.0,\ r'_{XY} = 0.772.$$

**Πίν. 7.6** Δεδομένα και αποτελέσματα των υπολογισμών της Εφαρμογής 7.6. Όλες οι στήλες εκτός της πρώτης περιέχουν ύψη βροχής σε mm. Η δεύτερη και η τρίτη στήλη περιέχουν τα πραγματικά δείγματα και οι υπόλοιπες τα δείγματα που έχουν προκύψει με επέκταση στη δεύτερη εικοσαετία του δείγματος $Y$ της πρώτης εικοσαετίας με βάση τις εξεταζόμενες μεθόδους.

| A/A | $X$ | $Y$ | $Y^{(1)}$ | $Y^{(2)}$ | $Y^{(3)}$ |
|---|---|---|---|---|---|
| 1 | 1767.7 | 1334.1 | 1334.1 | 1334.1 | 1334.1 |
| 2 | 1280.3 | 606.8 | 606.8 | 606.8 | 606.8 |
| 3 | 1618.2 | 1652.1 | 1652.1 | 1652.1 | 1652.1 |
| 4 | 1143.1 | 983.9 | 983.9 | 983.9 | 983.9 |
| 5 | 1964.1 | 1620.7 | 1620.7 | 1620.7 | 1620.7 |
| 6 | 1288.9 | 1022.6 | 1022.6 | 1022.6 | 1022.6 |
| 7 | 1979.6 | 1290.5 | 1290.5 | 1290.5 | 1290.5 |
| 8 | 1196.0 | 942.2 | 942.2 | 942.2 | 942.2 |
| 9 | 1354.1 | 1002.2 | 1002.2 | 1002.2 | 1002.2 |
| 10 | 968.3 | 972.5 | 972.5 | 972.5 | 972.5 |
| 11 | 1670.9 | 907.0 | 907.0 | 907.0 | 907.0 |
| 12 | 2052.9 | 1237.1 | 1237.1 | 1237.1 | 1237.1 |
| 13 | 910.5 | 506.9 | 506.9 | 506.9 | 506.9 |
| 14 | 980.4 | 792.7 | 792.7 | 792.7 | 792.7 |
| 15 | 2006.1 | 1086.9 | 1086.9 | 1086.9 | 1086.9 |
| 16 | 1616.0 | 1186.3 | 1186.3 | 1186.3 | 1186.3 |
| 17 | 1490.7 | 1090.9 | 1090.9 | 1090.9 | 1090.9 |
| 18 | 1511.6 | 1238.1 | 1238.1 | 1238.1 | 1238.1 |
| 19 | 1357.0 | 1473.3 | 1473.3 | 1473.3 | 1473.3 |
| 20 | 1691.1 | 1383.9 | 1383.9 | 1383.9 | 1383.9 |
| 21 | 1307.9 | 763.7 | 1016.3 | 961.0 | 1376.7 |
| 22 | 2078.1 | 1565.2 | 1434.5 | 1610.2 | 1694.6 |
| 23 | 1657.3 | 1241.8 | 1206.0 | 1255.5 | 970.6 |
| 24 | 659.0 | 815.8 | 663.9 | 413.9 | 362.2 |
| 25 | 1752.7 | 1278.4 | 1257.8 | 1335.9 | 987.7 |
| 26 | 1437.0 | 1021.9 | 1086.4 | 1069.8 | 740.1 |
| 27 | 1762.8 | 844.2 | 1263.3 | 1344.4 | 1126.2 |
| 28 | 1132.2 | 682.9 | 920.9 | 812.8 | 949.4 |
| 29 | 847.1 | 475.7 | 766.1 | 572.5 | 655.6 |
| 30 | 1629.8 | 1306.6 | 1191.1 | 1232.3 | 1243.9 |
| 31 | 1449.3 | 1560.0 | 1093.1 | 1080.2 | 1542.5 |
| 32 | 1434.3 | 1360.4 | 1084.9 | 1067.5 | 1166.8 |
| 33 | 1444.0 | 1279.2 | 1090.2 | 1075.7 | 1395.3 |
| 34 | 1880.6 | 1360.2 | 1327.3 | 1443.7 | 980.2 |
| 35 | 1438.6 | 1379.5 | 1087.3 | 1071.1 | 1104.8 |
| 36 | 1352.5 | 1187.0 | 1040.5 | 998.6 | 1051.8 |
| 37 | 2111.6 | 1601.1 | 1452.7 | 1638.5 | 1264.7 |
| 38 | 1760.2 | 1432.4 | 1261.9 | 1342.2 | 1319.5 |
| 39 | 1884.4 | 1403.5 | 1329.3 | 1446.9 | 1414.4 |
| 40 | 1685.7 | 1646.1 | 1221.4 | 1279.4 | 1216.0 |

Παρατηρούμε ότι υπάρχει σημαντική υπεκτίμηση της τυπικής απόκλισης (κατά 20% περίπου σε σχέση με την πραγματική τιμή 317.3) και υπερεκτίμηση του συντελεστή συσχέτισης (κατά 12% περίπου σε σχέση με την πραγματική τιμή 0.691).
Η τυπική απόκλιση μπορεί να διορθωθεί με πολλαπλασιασμό επί τον συντελεστή $\sqrt{\varphi}$ της εξίσωσης (7.84), όπου

$$\varphi = (40 - 1) / [20 + (40 - 20) \times 0.644^2 - 1] = 1.429$$

οπότε βρίσκουμε

$$s'_{*Y} = 253.0 \times \sqrt{1.429} = 302.4$$

τιμή πολύ κοντύτερα στην πραγματική. Εναλλακτικά, μπορούμε να εφαρμόσουμε την εξίσωση (7.86), οπότε βρίσκουμε

$$s'^2_{*Y} = 301.2^2 + 0.543^2 \times (359.6^2 - 357.3^2) = 91\ 208$$

και $s'_{*Y} = 302.0$. Εξ άλλου, η περίπου αμερόληπτη εκτίμηση του συντελεστή συσχέτισης του διευρυμένου δείγματος δίνεται από την εξίσωση (7.87) και είναι

$$r'_{XY} = 0.644 \times (301.2 / 357.3) \times (359.6 / 302.0) = 0.646$$

Τέλος, το ισοδύναμο μήκος του διευρυμένου δείγματος δίνεται από την (7.92), ήτοι

$$k / n' = 1 - (20/40) \times [0.644^2 - (1 - 0.644^2) / (20 - 3)] = 0.81$$

άρα

$$n' = 20 / 0.810 \approx 25$$

(έναντι $n = 40$).
Στη συνέχεια εφαρμόζουμε τη μέθοδο της οργανικής συσχέτισης. Οι παράμετροι της γραμμικής εξίσωσης $y = a + b\,x$, υπολογισμένες από τα δείγματα της πρώτης εικοσαετίας(εξ. (7.54) και (7.55)), είναι:

$$b = 301.2 / 357.3 = 0.843,\ a = 1116.5 - 0.843 \times 1492.4 = -141.6$$

Εφαρμόζοντας τη γραμμική εξίσωση αυτή για τα δεδομένα $x$ της δεύτερης εικοσαετίας, παίρνουμε τις εκτιμήσεις που φαίνονται στη στήλη $Y^{(2)}$ του Πίν. 7.6. Αν τώρα υπολογίσουμε τα στατιστικά χαρακτηριστικά των 40 ετών της στήλης $Y^{(2)}$ βρίσκουμε

$$\bar{y}' = 1134.6,\ s'_{*Y} = 303.1,\ r'_{XY} = 0.829.$$

Παρατηρούμε ότι η εκτίμηση της τυπικής απόκλισης είναι στην περίπτωση αυτή ικανοποιητική και δεν χρειάζεται αναγωγή. Αντίθετα, υπάρχει σημαντική υπερεκτίμηση του συντελεστή συσχέτισης (κατά 20% περίπου σε σχέση με την πραγματική τιμή 0.691). Η περίπου αμερόληπτη εκτίμηση του συντελεστή συσχέτισης του διευρυμένου δείγματος μπορεί να υπολογιστεί και πάλι από την εξίσωση (7.87), οπότε θα είναι

$$r'_{XY} = 0.644 \times (301.2 / 357.3) \times (359.6 / 303.1) = 0.643$$

Τέλος εφαρμόζουμε τη μέθοδο της γραμμικής παλινδρόμησης με όρο σφάλματος. Χρησιμοποιούμε την εξίσωση $y = a + b\,x + w$, όπου οι παράμετροι έχουν τις τιμές που έχουν υπολογιστεί στην τυπική γραμμική παλινδρόμηση, δηλαδή $b = 0.543$ και $a = 306.1$. Ο όσος σφάλματος έχει μέση τιμή 0 και τυπική απόκλιση που δίνεται από την εξίσωση (7.82), δηλαδή

$$\hat{\sigma}_W = \sqrt{n/(n-2)}\, s_Y \sqrt{1 - r_{XY}^2} =$$

$$= \sqrt{20/18} \times (301.2 \times \sqrt{19/20}) \times \sqrt{1 - 0.644^2} = 236.7$$

Πριν εφαρμόσουμε την πιο πάνω γραμμική εξίσωση γεννούμε 20 τυχαίους αριθμούς από την τυποποιημένη κανονική κατανομή. Εδώ έχουμε χρησιμοποιήσει τη βοήθεια ρουτίνας ηλεκτρονικού υπολογιστή και οι 20 τυχαίοι αριθμοί που γεννήσαμε είναι οι ακόλουθοι:

+1.52281 +1.09894 –0.99478 –1.27496 –1.14112
–1.46306 –0.57916 +0.12054 –0.46660 +0.22329
+1.89879 +0.34577 +1.28891 –1.46615 +0.07426
+0.04751 –0.79445 +0.24338 +0.35920 –0.02287

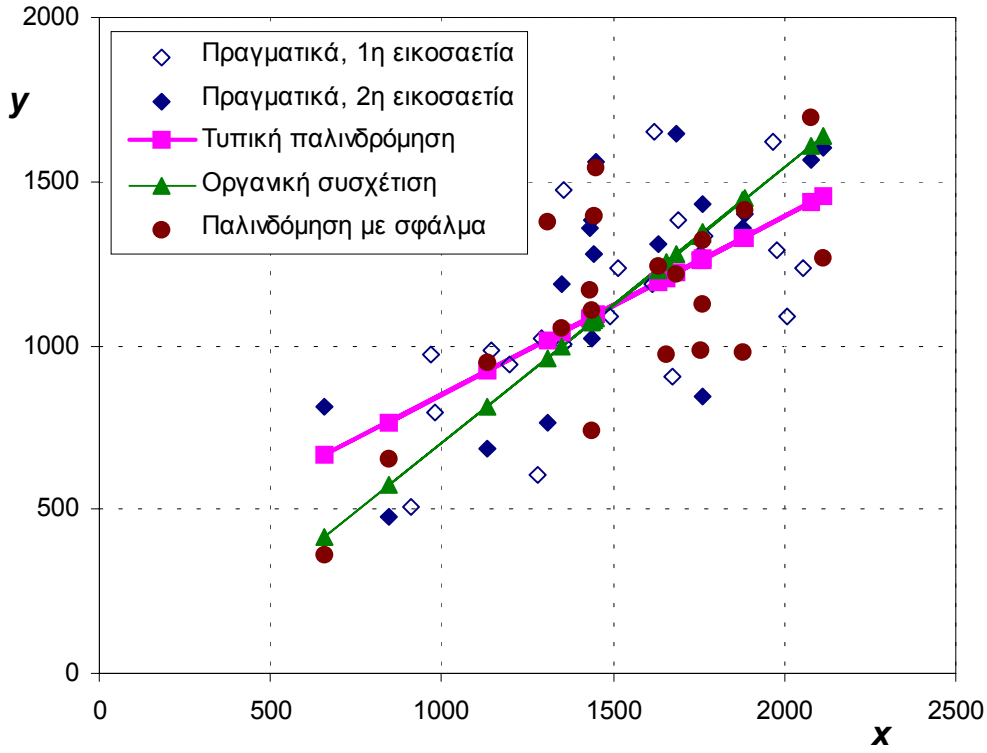
Για να βρούμε τους αριθμούς $w_i$ που είναι απαραίτητοι, πολλαπλασιάζουμε τους παραπάνω τυχαίους αριθμούς επί την τυπική απόκλιση $\hat{\sigma}_W = 236.7$. Εφαρμόζοντας τη γραμμική εξίσωση $y = a + b\,x + w$, για τα δεδομένα $x_i$ της δεύτερης εικοσαετίας, παίρνουμε τις τιμές που φαίνονται στη στήλη $Y^{(3)}$ του Πίν. 7.6). Αν υπολογίσουμε τα στατιστικά χαρακτηριστικά των 40 ετών της στήλης $Y^{(3)}$ βρίσκουμε

$$\bar{y}' = 1122.3,\ s'_{*Y} = 302.8,\ r'_{XY} = 0.654$$

Παρατηρούμε ότι αυτά τα στατιστικά χαρακτηριστικά βρίσκονται κοντά στις πραγματικές τιμές και δεν χρειάζονται αναγωγές ή διορθώσεις.

Τα τρία διευρυμένα δείγματα της μεταβλητής $Y$, που έχουμε υπολογίσει παραπάνω, τα έχουμε απεικονίσει στο Σχ. 7.19, συναρτήσει των

τιμών της $X$. Είναι εμφανές ότι το δείγμα της τρίτης μεθόδου είναι το πιο ρεαλιστικό, με την έννοια ότι τα σημεία του γενικά διατάσσονται με παρόμοιο τρόπο, όπως οι πραγματικές μετρήσεις. Αυτό βέβαια δεν σημαίνει ότι κάθε σημείο πλησιάζει στην πραγματική του θέση περισσότερο απ' ότι τα σημεία των άλλων δύο μεθόδων. Το βασικό μειονέκτημα της τρίτης μεθόδου είναι ότι τα σημεία της δεν ορίζονται με μονοσήμαντο τρόπο, αφού αν χρησιμοποιούσαμε άλλους τυχαίους αριθμούς, θα είχαμε διαφορετικά αποτελέσματα.




**Σχ. 7.19** Γραφική απεικόνιση των τιμών της $Y$ συναρτήσει των τιμών της $X$ στα δεδομένα της Εφαρμογής 7.6.

## Παράρτημα 7Α: Παραγωγή των εκτιμήσεων του διευρυμένου δείγματος και σχετικές αποδείξεις

Θα εξαγάγουμε πρώτα την εκτιμήτρια της μέσης τιμής του διευρυμένου δείγματος. Από την εξίσωση (7.24), παίρνοντας υπόψη και την (7.28), για το τυχόν ζεύγος μεταβλητών ($X_i$, $Y_i$) έχουμε

$$Y_i = \mu_Y + b(X_i - \mu_X) + W_i \tag{7.103}$$

Γράφοντας την (7.103) για όλα τα ζεύγη ($X_i$, $Y_i$), αθροίζοντας και διαιρώντας διά του μεγέθους του δείγματος $k$, βρίσκουμε

$$\frac{1}{k}\sum_{i=1}^{k} Y_i = \mu_Y + b\left(\frac{1}{k}\sum_{i=1}^{k} X_i - \mu_X\right) + \frac{1}{k}\sum_{i=1}^{k} W_i \tag{7.104}$$

ή, ισοδύναμα,

$$\bar{Y} = \mu_Y + b(\bar{X} - \mu_X) + \bar{W} \tag{7.105}$$

Όμοια, για το διευρυμένο δείγμα, μεγέθους $n$, έχουμε

$$\bar{Y}' = \mu_Y + b(\bar{X}' - \mu_X) + \bar{W}' \tag{7.106}$$

Αφαιρώντας τις (7.104) και (7.105) παίρνουμε

$$\bar{Y}' = \bar{Y} + b(\bar{X}' - \bar{X}) + \bar{W}' - \bar{W} \tag{7.107}$$

Αν στην τελευταία εξίσωση παραλείψουμε τους όρους σφάλματος, για καθέναν από τους οποίους η αναμενόμενη τιμή είναι μηδενική, και στη θέση των εκτιμητριών βάλουμε τις εκτιμήσεις, παίρνουμε την (7.85).

Ερχόμαστε τώρα στην εκτιμήτρια της διασποράς. Αφαιρώντας από την (7.105) την (7.103) έχουμε

$$Y_i - \bar{Y} = b(X_i - \bar{X}) + W_i - \bar{W} \tag{7.108}$$

και υψώνοντας στο τετράγωνο παίρνουμε

$$(Y_i - \bar{Y})^2 = b^2(X_i - \bar{X})^2 + (W_i - \bar{W})^2 + 2b\,(X_i - \bar{X})\,(W_i - \bar{W}) \tag{7.109}$$

Γράφοντας την παραπάνω για όλα τα ζεύγη $(X_i, Y_i)$, αθροίζοντας και διαιρώντας διά $k-1$, βρίσκουμε

$$S^2_{*Y} = b^2 S^2_{*X} + S^2_{*W} + 2b S_{*XW} \tag{7.110}$$

Γράφοντας την ίδια εξίσωση για το διευρυμένο δείγμα και στη συνέχεια αφαιρώντας από αυτή την (7.110) παίρνουμε

$$S'^2_{*Y} = S^2_{*Y} + b^2\left(S'^2_{*X} - S^2_{*X}\right) + \left(S'^2_{*W} - S^2_{*W}\right) + 2b\left(S'_{*XW} - S^2_{*XW}\right) \tag{7.111}$$

Αν στην τελευταία εξίσωση παραλείψουμε τους όρους σφάλματος και στη θέση των εκτιμητριών βάλουμε τις εκτιμήσεις, παίρνουμε την (7.86). Η παράλειψη αυτή δεν εισάγει μεροληψία δεδομένου ότι

$$E[S'^2_{*W}] = E[S^2_{*W}] = \sigma^2_W, \qquad E[S'_{*XW}] = E[S^2_{*XW}] = 0 \tag{7.112}$$

και επομένως η συνολική συνεισφορά τους στην αναμενόμενη τιμή του δεξιού μέλους της (7.111) είναι μηδενική.

Στη συνέχεια θα εξετάσουμε την εκτιμήτρια του συντελεστή συσχέτισης του διευρυμένου δείγματος. Πολλαπλασιάζοντας τα δύο μέλη της (7.108) επί $(X_i - \bar{X})$ παίρνουμε

$$(Y_i - \bar{Y})(X_i - \bar{X}) = b(X_i - \bar{X})^2 + (W_i - \bar{W})(X_i - \bar{X}) \tag{7.113}$$

και εργαζόμενοι όπως προηγουμένως καταλήγουμε ότι

$$S_{*XY} = b S^2_{*X} + S_{*XW} \tag{7.114}$$

Παραλείποντας τον τελευταίο όρο, ο οποίος έχει μηδενική συνεισφορά σε αναμενόμενη τιμή, δεδομένου ότι $E[S_{*XW}] = 0$, η (7.114) απλοποιείται, χωρίς εισαγωγή μεροληψίας, και γίνεται

$$S_{*XY} = b S^2_{*X} \tag{7.115}$$

Υπό αυτές τις συνθήκες, η εκτιμήτρια του συντελεστή συσχέτισης είναι

$$R_{XY} = \frac{S_{*XY}}{S_{*X}\, S_{*Y}} = b\,\frac{S_{*X}}{S_{*Y}} \tag{7.116}$$

Γράφοντας την ίδια εξίσωση για το διευρυμένο δείγμα και στη συνέχεια διαιρώντας τη με τη (7.116) παίρνουμε

$$R'_{XY} = R_{XY}\,\frac{S_{*Y}}{S_{*X}}\,\frac{S'_{*X}}{S'_{*Y}} \tag{7.117}$$

Αν στην (7.117) αντικαταστήσουμε τις εκτιμήτριες με τις εκτιμήσεις παίρνουμε την (7.87).

Θα υπολογίσουμε τώρα τη διασπορά της εκτιμήτριας της μέσης τιμής του διευρυμένου δείγματος, όταν η μέση τιμή αυτή δίνεται από την εξίσωση (7.85). Η εξίσωση (7.85) γράφεται ισοδύναμα

$$\begin{aligned}(\bar{Y}' - \mu_Y) &= (\bar{Y} - \mu_Y) + b(\bar{X}' - \mu_X) - b(\bar{X} - \mu_X) \\ &= \frac{1}{k}\sum_{i=1}^{k}(Y_i - \mu_Y) + \frac{b}{n}\sum_{i=1}^{n}(X_i - \mu_X) - \frac{b}{k}\sum_{i=1}^{k}(X_i - \mu_X)\end{aligned} \tag{7.118}$$

Υψώνουμε την παραπάνω στο τετράγωνο και στη συνέχεια παίρνουμε αναμενόμενες τιμές. Παρατηρούμε ότι λόγω της ανεξαρτησίας των διάφορων ζευγών μεταβλητών ($X_i$, $Y_i$) ισχύουν τα ακόλουθα:

$$\mathrm{E}[(Y_i - \mu_Y)(Y_j - \mu_Y)] = \begin{cases} 0 & i \neq j \\ \sigma_Y^2 & i = j \end{cases} \tag{7.119}$$

$$\mathrm{E}[(X_i - \mu_X)(X_j - \mu_X)] = \begin{cases} 0 & i \neq j \\ \sigma_X^2 & i = j \end{cases} \tag{7.120}$$

$$\mathrm{E}[(X_i - \mu_X)(Y_j - \mu_Y)] = \begin{cases} 0 & i \neq j \\ \sigma_{XY} & i = j \end{cases} \tag{7.121}$$

Έτσι βρίσκουμε

$$\begin{aligned}\mathrm{Var}[\bar{Y}'] = &\frac{1}{k^2}\, k\, \sigma_Y^2 + \frac{b^2}{n^2}\, n\, \sigma_X^2 + \frac{b^2}{k^2}\, k\, \sigma_X^2 \\ &+ 2\frac{1}{k}\frac{b}{n}\, k\, \sigma_{XY} - 2\frac{1}{k}\frac{b}{k}\, k\, \sigma_{XY} - 2\frac{b}{n}\frac{b}{k}\, k\, \sigma_X^2\end{aligned} \tag{7.122}$$

Παίρνοντας υπόψη ότι $b^2\, \sigma_X^2 = b\, \sigma_{XY} = \rho^2\, \sigma_Y^2$, βρίσκουμε ότι

$$\mathrm{Var}[\bar{Y}'] = \frac{\sigma_Y^2}{k}\left[1 + \rho_{XY}^2\left(\frac{k}{n} + 1 + 2\,\frac{k}{n} - 2 - 2\,\frac{k}{n}\right)\right] \tag{7.123}$$

και τελικά, κάνοντας τις πράξεις παίρνουμε την εξίσωση (7.88).

Τέλος, θα δώσουμε την απόδειξη για την εξίσωση (7.84), η οποία δίνει το συντελεστή διόρθωσης της μεροληψίας για την περίπτωση που η διασπορά υπολογίζεται από το διευρυμένο δείγμα, χρησιμοποιώντας τον τύπο της δειγματικής διασποράς παρατηρημένου δείγματος. Η δειγματική διασπορά αυτή γράφεται

$$S_{*Y}^{\prime\prime\,2} = \frac{1}{n-1}\left[\sum_{i=1}^{k}(Y_i - \bar{Y}')^2 + \sum_{i=k+1}^{n}(\hat{Y}_i - \bar{Y}')^2\right] \tag{7.124}$$

όπου ο πρώτος όρος μέσα στην τετραγωνική αγκύλη αναφέρεται στις παρατηρημένες τιμές της μεταβλητής $Y$ και ο δεύτερος στις εκτιμημένες μέσω της μεταβλητής $X$. Για τις πρώτες ισχύει

$$Y_i - \mu_Y = (Y_i - \bar{Y}') + (\bar{Y}' - \mu_Y) \tag{7.125}$$

και κατά συνέπεια

$$(Y_i - \mu_Y)^2 = (Y_i - \bar{Y}')^2 + (\bar{Y}' - \mu_Y)^2 + 2\,(Y_i - \bar{Y}')(\bar{Y}' - \mu_Y) \tag{7.126}$$

Για τις δεύτερες ισχύει

$$Y_i - \mu_Y = (\hat{Y}_i - \bar{Y}') + (\bar{Y}' - \mu_Y) + W_i \tag{7.127}$$

και κατά συνέπεια

$$(Y_i - \mu_Y)^2 = (\hat{Y}_i - \bar{Y}')^2 + (\bar{Y}' - \mu_Y)^2 + 2\,(\hat{Y}_i - \bar{Y}')(\bar{Y}' - \mu_Y) + W_i^2 + 2\,W_i\,(\hat{Y}_i - \mu_Y) \tag{7.128}$$

Σχηματίζουμε την (7.126) για $i = 1$ μέχρι $k$ και την (7.128) για $i = k + 1$ μέχρι $n$, προσθέτουμε τις $n$ αυτές εξισώσεις και παίρνουμε αναμενόμενες τιμές. Στο αριστερό μέλος της εξίσωσης θα προκύψει $n$ φορές η διασπορά της $Y$, δεδομένου ότι

$$E[(Y_i - \mu_Y)^2] = \sigma_Y^2 \tag{7.129}$$

Στο δεξιό μέλος της εξίσωσης θα προκύψει κατ' αρχήν το άθροισμα που βρίσκεται μέσα στην αγκύλη της εξίσωσης (7.124) και οι υπόλοιποι όροι για τους οποίους ισχύουν

$$E\left[\sum_{i=1}^{n} (\bar{Y}' - \mu_Y)^2\right] = n\ \mathrm{Var}[\bar{Y}'] = \frac{n}{k}\left(1 - \frac{n-k}{n}\rho_{XY}^2\right) \tag{7.130}$$

(λόγω της (7.88)),

$$\sum_{i=1}^{k} (Y_i - \bar{Y}') + \sum_{i=k+1}^{n} (\hat{Y}_i - \bar{Y}') = 0 \tag{7.131}$$

και

$$E[W_i^2] = (1 - \rho_{XY}^2)\sigma_Y^2, \quad E[W_i\,(\hat{Y}_i - \mu_Y)] = 0 \tag{7.132}$$

Κατά συνέπεια βρίσκουμε

$$n\sigma_Y^2 = (n-1)E[S_{*Y}^{\prime\prime\,2}] + \frac{n}{k}\left(1 - \frac{n-k}{n}\rho_{XY}^2\right)\sigma_Y^2 + (n-k)\,(1 - \rho_{XY}^2)\,\sigma_Y^2 \tag{7.133}$$

Λύνοντας την τελευταία ως προς $E[S_{*Y}^{\prime\prime\,2}]$ παίρνουμε

$$E[S_{*Y}^{''2}] = \sigma_Y^2 \frac{k + (n-k)\,\rho_{XY}^2 - 1 - (1-\rho_{XY}^2)\,\dfrac{n-k}{k}}{n-1} \tag{7.134}$$

Ο τελευταίος όρος στον αριθμητή του δεξιού μέλους της (7.134) μπορεί να παραλειφθεί χωρίς ουσιαστικό σφάλμα, δεδομένου ότι είναι πολύ μικρότερος από το άθροισμα των άλλων όρων. Παραλείποντας τον όρο αυτό και αντιστρέφοντας το κλάσμα παίρνουμε το συντελεστή διόρθωσης της (7.84).