# Κεφάλαιο 3

# ΣΥΣΧΕΤΙΣΗ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗ

Στα προηγούμενα κεφάλαια ορίσαμε και μελετήσαμε την τ.μ. με τη βοήθεια της πιθανοθεωρίας (κατανομή, ροπές) και της στατιστικής (εκτίμηση, στατιστική υπόθεση). Σ' αυτό το κεφάλαιο θα μελετήσουμε πως μια τ.μ. μεταβάλλεται όταν αλλάζει μια άλλη μεταβλητή (τυχαία ή μη).

Πρώτα θα μελετήσουμε τη σχέση δύο τ.μ. $X$ και $Y$. Συχνά στη μελέτη ενός τεχνικού συστήματος ενδιαφερόμαστε να προσδιορίσουμε τη σχέση μεταξύ δύο μεταβλητών του συστήματος. Για παράδειγμα στη λειτουργία μια μηχανής μας ενδιαφέρει η σχέση του χρόνου ως την αποτυχία κάποιου στοιχείου της μηχανής και της ταχύτητας του κινητήρα (σε περιστροφές ανά λεπτό). Θα προσδιορίσουμε και θα εκτιμήσουμε το συντελεστή συσχέτισης που μετράει τη γραμμική συσχέτιση δύο τ.μ..

Στη συνέχεια θα μελετήσουμε τη συναρτησιακή σχέση εξάρτησης μιας τ.μ. $Y$ ως προς μια άλλη μεταβλητή $X$. Η σχέση αυτή είναι πιθανοκρατική κι ορίζεται με την κατανομή της $Y$ για κάθε τιμή της $X$. Για παράδειγμα η απόδοση κάποιου εργαστηριακού πειράματος μπορεί να θεωρηθεί τ.μ. με κατανομή που μεταβάλλεται με τη θερμοκρασία στην οποία πραγματοποιείται. Συνήθως η μεταβολή αφορά μόνο τη μέση τιμή (κι ορισμένες φορές και τη διασπορά), γι αυτό κι η περιγραφή της κατανομής της $Y$ ως προς τη $X$ περιορίζεται στη δεσμευμένη μέση τιμή $\mathrm{E}(Y|X)$ και γίνεται με τη λεγόμενη ανάλυση παλινδρόμησης. Θα μελετήσουμε την απλή γραμμική παλινδρόμηση, δηλαδή θα περιοριστούμε να εκτιμήσουμε τη γραμμική σχέση για τη μέση τιμή $\mathrm{E}(Y|X)$ ως προς μια τ.μ. $X$.

## 3.1 Συσχέτιση δύο τ.μ.

Δύο τ.μ. $X$ και $Y$ μπορεί να συσχετίζονται με κάποιο τρόπο. Αυτό συμβαίνει όταν επηρεάζει η μία την άλλη, ή αν δεν αλληλοεπηρεάζονται όταν επηρεάζονται και οι δύο από μια άλλη μεταβλητή. Για παράδειγμα ο χρόνος ως την αποτυχία ενός στοιχείου κάποιας μηχανής και η ταχύτητα του κινητήρα της μηχανής μπορούν να θεωρηθούν σαν δύο τ.μ. που συσχετίζονται, όπου ο χρόνος αποτυχίας εξαρτάται από την ταχύτητα του κινητήρα (το αντίθετο δεν έχει πρακτική σημασία). Μπορούμε επίσης να θεωρήσουμε τη συσχέτιση του χρόνου αποτυχίας και της θερμοκρασίας του στοιχείου της μηχανής, αλλά τώρα δεν εξαρτάται η μια από την άλλη παρά εξαρτιούνται κι οι δύο από άλλες μεταβλητές, όπως η ταχύτητα του κινητήρα.

Στη συνέχεια θα θεωρήσουμε ότι οι δύο τ.μ. $X$ και $Y$ είναι συνεχείς. Για διακριτές τ.μ. μπορούμε πάλι να ορίσουμε μέτρο συσχέτισης τους αλλά δε θα μας απασχολήσει εδώ.

### 3.1.1 Ο συντελεστής συσχέτισης $\rho$

Ας θεωρήσουμε δύο τ.μ. $X$ και $Y$ με διασπορά $\sigma_X^2$ και $\sigma_Y^2$ αντίστοιχα και συνδιασπορά

$$\sigma_{XY} \equiv \mathrm{Cov}(X,Y) = \mathrm{E}(X,Y) - \mathrm{E}(X)\mathrm{E}(Y).$$

Η συνδιασπορά εκφράζει τη γραμμική συσχέτιση δύο τ.μ., δηλαδή την αναλογική μεταβολή (αύξηση ή μείωση) της μιας τ.μ. που αντιστοιχεί σε μεταβολή της άλλης μεταβλητής. Η συνδιασπορά είναι ένα ποσοτικό μέγεθος κι η μονάδα μέτρησης της εξαρτάται από τις μονάδες μέτρησης των δύο τ.μ. $X$ και $Y$. Γι αυτό για να μετρήσουμε καλύτερα τον βαθμό της *γραμμικής* συσχέτισης δύο τ.μ. χρησιμοποιούμε τον **συντελεστή συσχέτισης** (correlation coefficient) $\rho$ που ορίζεται ως

$$\rho = \frac{\sigma_{XY}}{\sigma_X\,\sigma_Y}. \tag{3.1}$$

Ο συντελεστής συσχέτισης $\rho$ παίρνει τιμές στο διάστημα $[-1,1]$:

- $\rho = 1$: υπάρχει *τέλεια θετική* σχέση μεταξύ των $X$ και $Y$,
- $\rho = 0$: δεν υπάρχει καμιά (γραμμική) σχέση μεταξύ των $X$ και $Y$,
- $\rho = -1$: υπάρχει *τέλεια αρνητική* σχέση μεταξύ των $X$ και $Y$.

Όταν $\rho = \pm 1$ η σχέση είναι αιτιοκρατική κι όχι πιθανοκρατική γιατί γνωρίζοντας την τιμή της μιας τ.μ. γνωρίζουμε και την τιμή της άλλης τ.μ. ακριβώς. Όταν ο συντελεστής συσχέτισης είναι κοντά στο $-1$ ή $1$ η γραμμική συσχέτιση των δύο τ.μ. είναι ισχυρή (συνήθως χαρακτηρίζουμε ισχυρές τις σχέσεις όταν $|\rho| > 0.9$) ενώ όταν είναι κοντά στο $0$ οι τ.μ. είναι πρακτικά ασυσχέτιστες.

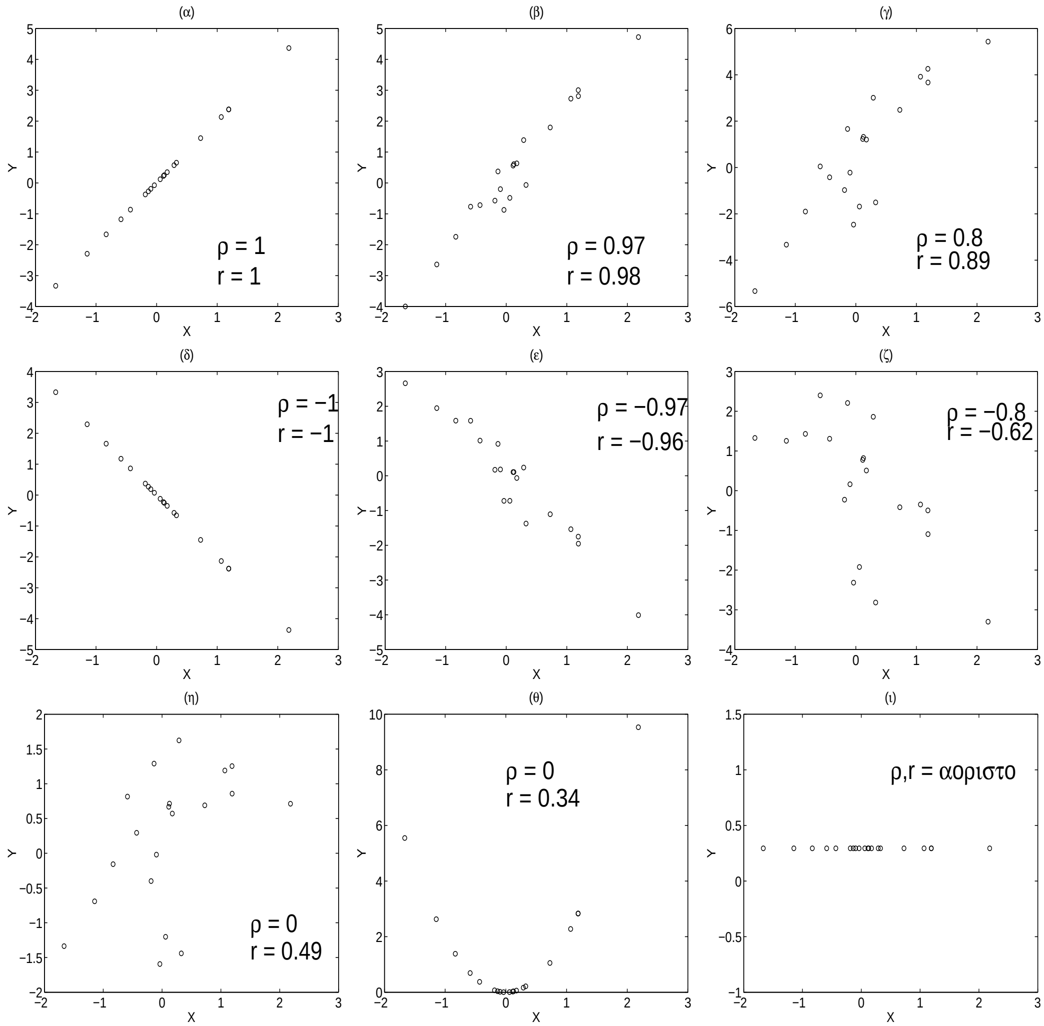
Όπως φαίνεται από τον ορισμό στη σχέση (3.1), ο συντελεστής συσχέτισης $\rho$ δεν εξαρτάται από τη μονάδα μέτρησης των $X$ και $Y$ και είναι συμμετρικός ως προς τις $X$ και $Y$.

### 3.1.2 Σημειακή εκτίμηση του συντελεστή συσχέτισης

Όταν έχουμε παρατηρήσεις των δύο τ.μ. $X$ και $Y$ κατά ζεύγη

$$\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\},$$

μπορούμε να εκτιμήσουμε τη συσχέτιση τους ποιοτικά από το **διάγραμμα διασποράς** (scatter diagram), που είναι η απεικόνιση των σημείων $(x_i,y_i)$, $i=1,\ldots,n$, σε καρτεσιανό σύστημα συντεταγμένων. Στο Σχήμα 3.1 παρουσιάζονται τυπικά διαγράμματα διασποράς για ισχυρές κι ασθενείς συσχετίσεις δύο τ.μ. $X$ και $Y$. Στα Σχήματα 3.1α και 3.1δ η σχέση είναι τέλεια ($\rho = 1$ και $\rho = -1$ αντίστοιχα), στα Σχήματα 3.1β και 3.1ε είναι ισχυρή (θετική με $\rho = 0.97$ κι αρνητική με $\rho = -0.97$ αντίστοιχα) και στα Σχήματα 3.1γ και 3.1ζ είναι λιγότερο ισχυρή (θετική με $\rho = 0.8$ κι αρνητική με $\rho = -0.8$ αντίστοιχα). Στο Σχήμα 3.1η είναι $\rho = 0$ γιατί οι τ.μ. $X$ και $Y$ είναι ανεξάρτητες ενώ στο Σχήμα 3.1θ είναι πάλι $\rho = 0$ αλλά οι $X$ και $Y$ δεν είναι ανεξάρτητες αλλά συσχετίζονται μόνο μη-γραμμικά. Τέλος για το Σχήμα 3.1ι ο συντελεστής συσχέτισης δεν ορίζεται γιατί η $Y$ είναι σταθερή ($\sigma_Y = 0$ στον ορισμό του $\rho$ στην (3.1)).




Σχήμα 3.1: Διάγραμμα διασποράς δύο τ.μ. $X$ και $Y$ από $n = 20$ παρατηρήσεις που παρουσιάζουν θετική σχέση στα σχήματα (α), (β) και (γ), αρνητική σχέση στα σχήματα (δ), (ε) και (ζ) και καμιά συσχέτιση στα σχήματα (η), (θ) και (ι). Σε κάθε σχήμα δίνεται η πραγματική τιμή του συντελεστή συσχέτισης $\rho$ κι η δειγματική $r$. Στο (ι) ο συντελεστής συσχέτισης δεν ορίζεται.

Η σημειακή εκτίμηση του συντελεστή συσχέτισης $\rho$ του πληθυσμού από το δείγμα των $n$ ζευγαρωτών παρατηρήσεων των $X$ και $Y$ γίνεται με την αντικατάσταση στη σχέση (3.1) της συνδιασποράς $\sigma_{XY}$ και των διασπορών $\sigma_X^2$ και $\sigma_Y^2$ από τις αντίστοιχες εκτιμήσεις από το δείγμα

$$\hat{\rho} \equiv r = \frac{s_{XY}}{s_X\, s_Y}.$$

Οι αμερόληπτες εκτιμήτριες $s_{XY}$, $s_X^2$ και $s_Y^2$ δίνονται ως

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \tag{3.2}$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \tag{3.3}$$

όπου $\bar{x}$ και $\bar{y}$ είναι οι δειγματικές μέσες τιμές των $X$ και $Y$. Από τα παραπάνω προκύπτει η έκφραση της εκτιμήτρια $r$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. \tag{3.4}$$

Είναι προφανές πως η παραπάνω σχέση για το $r$ δεν αλλάζει αν θεωρήσουμε τις μεροληπτικές εκτιμήτριες των $\sigma_{XY}$, $\sigma_X$ και $\sigma_Y$.

Στο Σχήμα 3.1 δίνεται ο δειγματικός συντελεστής συσχέτισης $r$ για κάθε περίπτωση. Επειδή το δείγμα είναι μικρό ($n = 20$) η τιμή του $r$ δεν είναι πάντα κοντά στην πραγματική τιμή $\rho$. Αυτό συμβαίνει γιατί η εκτιμήτρια $r$ όπως δίνεται στη σχέση (3.4) είναι μια τ.μ. που εξαρτάται από τις τιμές και το πλήθος των ζευγών των παρατηρήσεων.

Για να εκφράσουμε τη συσχέτιση δύο τ.μ. χρησιμοποιούμε επίσης την ποσότητα $r^2$ που λέγεται και **συντελεστής προσδιορισμού** (coefficient of determination) (εκφράζεται συνήθως σε ποσοστό $\%$, $100r^2$). Ο συντελεστής προσδιορισμού δίνει το ποσοστό μεταβλητότητας των τιμών της $Y$ που υπολογίζεται από τη $X$ (κι αντίστροφα) κι είναι ένας χρήσιμος τρόπος να συνοψίσουμε τη σχέση δύο τ.μ..

**Παράδειγμα 3.1** *Θέλουμε να διερευνήσουμε τη συσχέτιση της αντίστασης και του χρόνου αποτυχίας κάποιου υπερφορτωμένου αντιστάτη. Γι αυτό πήραμε μετρήσεις αντίστασης (σε Ωμ, ohm) και χρόνου αποτυχίας (σε λεπτά, min) από δείγμα 20 αντιστατών, οι οποίες παρουσιάζονται στον Πίνακα 3.1. Για να βρούμε τον συντελεστή συσχέτισης $r$ υπολογίζουμε πρώτα τα παρακάτω*

$$\bar{x} = 38 \qquad\qquad \bar{y} = 33.5$$
$$\sum_{i=1}^{20} x_i^2 = 29634 \qquad \sum_{i=1}^{20} y_i^2 = 23910 \qquad \sum_{i=1}^{20} x_i y_i = 26305.$$

*Από τη σχέση (3.4) βρίσκουμε*

$$r = \frac{26305 - 20 \cdot 38 \cdot 33.5}{\sqrt{(29634 - 20 \cdot 38^2) \cdot (23910 - 20 \cdot 33.5^2)}} = 0.804.$$

*Η τιμή του συντελεστή συσχέτισης $r \simeq 0.8$ υποδηλώνει ότι η αντίσταση κι ο χρόνος αποτυχίας αντιστάτη έχουν γραμμική θετική συσχέτιση αλλά όχι πολύ ισχυρή. Αυτό φαίνεται κι από το διάγραμμα διασποράς στο Σχήμα 3.2. Η μεταβλητότητα της μιας τ.μ. (αντίσταση ή χρόνος αποτυχίας) μπορεί να εξηγηθεί από τη συσχέτιση της με την άλλη κατά ποσοστό που δίνεται από το συντελεστή προσδιορισμού, που είναι*

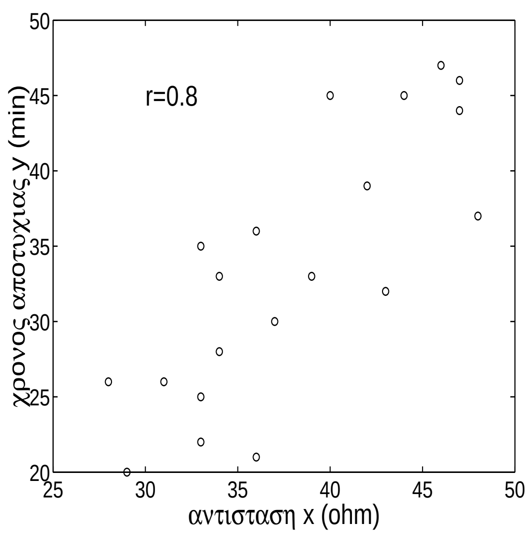
$$r^2 \cdot 100 = 0.804^2 \cdot 100 = 64.64 \quad \rightarrow \quad \simeq 65\%.$$

*Συμπεραίνουμε λοιπόν πως η γνώση της μιας τ.μ. δε μας επιτρέπει να προσδιορίσουμε την άλλη με μεγάλη ακρίβεια. Πρέπει επίσης να σημειωθεί ότι η εκτίμηση $r$ του συντελεστή συσχέτισης μπορεί να αλλάξει σημαντικά με την πρόσθεση ή αφαίρεση λίγων παρατηρήσεων γιατί το μέγεθος του δείγματος είναι μικρό.*

Για τον συντελεστή συσχέτισης $\rho$ μπορούμε να υπολογίσουμε διαστήματα εμπιστοσύνης κάτω από την υπόθεση κανονικής κατανομής των $X$ και $Y$, αλλά εδώ δε θα ασχοληθούμε με αυτά τα θέματα.

| Α/Α ($i$) | Αντίσταση $x_i$ (ohm) | Χρόνος αποτυχίας $y_i$ (min) |
|---|---|---|
| 1 | 28 | 26 |
| 2 | 29 | 20 |
| 3 | 31 | 26 |
| 4 | 33 | 22 |
| 5 | 33 | 25 |
| 6 | 33 | 35 |
| 7 | 34 | 28 |
| 8 | 34 | 33 |
| 9 | 36 | 21 |
| 10 | 36 | 36 |
| 11 | 37 | 30 |
| 12 | 39 | 33 |
| 13 | 40 | 45 |
| 14 | 42 | 39 |
| 15 | 43 | 32 |
| 16 | 44 | 45 |
| 17 | 46 | 47 |
| 18 | 47 | 44 |
| 19 | 47 | 46 |
| 20 | 48 | 37 |

Πίνακας 3.1: Δεδομένα αντίστασης $x_i$ και χρόνου αποτυχίας $y_i$ για 20 αντιστάτες.




Σχήμα 3.2: Διάγραμμα διασποράς για το δείγμα παρατηρήσεων αντίστασης και χρόνου αποτυχίας 20 αντιστατών του Πίνακα 3.1.

## 3.2 Απλή Γραμμική Παλινδρόμηση

Στη *συσχέτιση* που μελετήσαμε παραπάνω μετρήσαμε με το συντελεστή συσχέτισης τη γραμμική σχέση δύο τ.μ. $X$ και $Y$. Στην *παλινδρόμηση* που θα μελετήσουμε τώρα σχεδιάζουμε την εξάρτηση μιας τ.μ. $Y$, που την ονομάζουμε **εξαρτημένη μεταβλητή** (dependent variable), από κάποια άλλη μεταβλητή $X$ που την ονομάζουμε **ανεξάρτητη μεταβλητή** (independent variable). Ενώ λοιπόν η συσχέτιση είναι συμμετρική ως προς τα $X$ και $Y$, στην παλινδρόμηση η εξαρτημένη μεταβλητή $Y$ 'καθοδηγείται' από την ανεξάρτητη μεταβλητή $X$. Γι αυτό και στην ανάλυση που κάνουμε παίζει ρόλο ποιόν από τους δύο παράγοντες που μετράμε ορίζουμε σαν ανεξάρτητη μεταβλητή και ποιόν σαν εξαρτημένη. Για παράδειγμα, σε μια μονάδα παραγωγής ηλεκτρικής ενέργειας από λιγνίτη, για να προσδιορίσουμε το κόστος της παραγωγής ενέργειας, μελετάμε την εξάρτηση του από το κόστος του λιγνίτη. Είναι φυσικό λοιπόν ως εξαρτημένη μεταβλητή $Y$ να θεωρήσουμε το κόστος παραγωγής ηλεκτρικής ενέργειας κι ως ανεξάρτητη μεταβλητή $X$ το κόστος του λιγνίτη.

### 3.2.1 Το πρόβλημα της απλής γραμμικής παλινδρόμησης

Η εξαρτημένη τ.μ. $Y$ ακολουθεί κάποια κατανομή. Επειδή μας ενδιαφέρει η συμπεριφορά της $Y$ για κάθε δυνατή τιμή της ανεξάρτητης μεταβλητής $X$ θέλουμε να μελετήσουμε τη δεσμευμένη κατανομή της $Y$ για κάθε τιμή $x$ της $X$. Με αναφορά στη δεσμευμένη αθροιστική συνάρτηση κατανομής θέλουμε να προσδιορίσουμε την $F_Y(y|X=x)$ για κάθε τιμή $x$ της $X$. Αυτό είναι αρκετά περίπλοκο πρόβλημα που στην πράξη συχνά δε χρειάζεται να λύσουμε. Περιορίζουμε λοιπόν τη μελέτη του προβλήματος της παλινδρόμησης στη δεσμευμένη μέση τιμή $\mathrm{E}(Y|X=x)$. Υποθέτοντας ότι η εξάρτηση εκφράζεται από γραμμική σχέση έχουμε

$$\mathrm{E}(Y|X=x) = \alpha + \beta x \tag{3.5}$$

και η σχέση αυτή λέγεται **απλή γραμμική παλινδρόμηση της $Y$ στη $X$** (linear regression). [Λέγεται *απλή* γιατί η εξάρτηση είναι προς μόνο μια μεταβλητή.] Το πρόβλημα της παλινδρόμησης είναι η εύρεση των παραμέτρων $\alpha$ και $\beta$ που εκφράζουν καλύτερα τη γραμμική εξάρτηση της $Y$ από τη $X$. Κάθε ζεύγος τιμών $(\alpha, \beta)$ καθορίζει μια διαφορετική γραμμική σχέση που εκφράζεται γεωμετρικά από ευθεία γραμμή. Ο σταθερός όρος $\alpha$ είναι η τιμή του $y$ για $x=0$ και λέγεται **διαφορά ύψους** (intercept) κι ο συντελεστής του $x$, $\beta$, είναι η **κλίση** (slope) της ευθείας ή αλλιώς ο **συντελεστής παλινδρόμησης** (regression coefficient). Αν θεωρήσουμε τις παρατηρήσεις $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ και το διάγραμμα διασποράς που τις απεικονίζει σαν σημεία, μπορούμε να σχηματίσουμε πολλές τέτοιες ευθείες που προσεγγίζουν την υποτιθέμενη γραμμική εξάρτηση της $\mathrm{E}(Y|X=x)$ ως προς $X$, όπως φαίνεται στο Σχήμα 3.3.

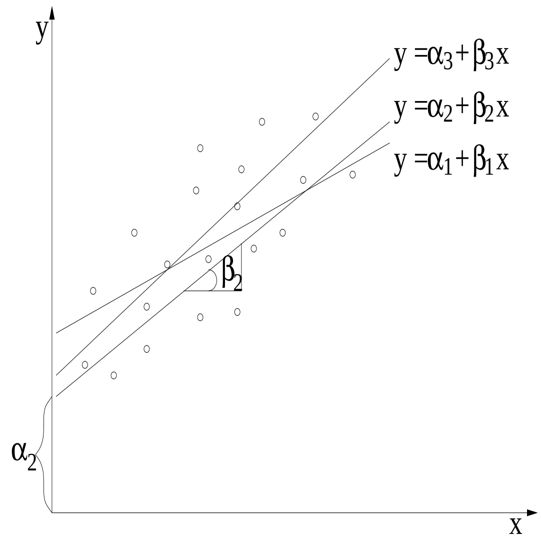
Για κάποια τιμή $x_i$ της $X$ αντιστοιχούν διαφορετικές τιμές $y_i$ της $Y$, σύμφωνα με κάποια κατανομή πιθανότητας $F_Y(y_i|X=x_i)$, δηλαδή μπορούμε να θεωρήσουμε την $y_i$ σαν τ.μ. [θα ήταν σωστότερο να χρησιμοποιούσαμε το συμβολισμό $Y_i$ αντί $y_i$, όπου ο δείκτης $i$ ορίζει την εξάρτηση από $X=x_i$, αλλά θα χρησιμοποιήσουμε εδώ τον ίδιο συμβολισμό $y_i$ για την τ.μ. και την παρατήρηση]. Η τ.μ. $y_i$ για κάποια τιμή $x_i$ της $X$ θα δίνεται κάτω από την υπόθεση της γραμμικής παλινδρόμησης ως

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{3.6}$$

όπου $\epsilon_i$ είναι κι αυτή τ.μ., λέγεται **σφάλμα παλινδρόμησης** (regression error) κι ορίζεται ως η διαφορά της $y_i$ από τη δεσμευμένη μέση τιμή της

$$\epsilon_i = y_i - \mathrm{E}(Y|X=x_i).$$

Για την ανάλυση της γραμμικής παλινδρόμησης κάνουμε τις παρακάτω υποθέσεις:




Σχήμα 3.3: Ευθείες απλής γραμμικής παλινδρόμησης

- Η μεταβλητή $X$ είναι *ελεγχόμενη* για το πρόβλημα που μελετάμε, δηλαδή γνωρίζουμε τις τιμές της χωρίς καμιά αμφιβολία.

- Η σχέση (3.5) ισχύει, δηλαδή η εξάρτηση της $Y$ από τη $X$ είναι γραμμική.

- $\mathrm{E}(\epsilon_i) = 0$ και $\mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2$ για κάθε τιμή $x_i$ της $X$, δηλαδή το σφάλμα παλινδρόμησης έχει μέση τιμή μηδέν για κάθε τιμή της $X$ και η διασπορά του είναι σταθερή και δεν εξαρτάται από τη $X$.

Η τελευταία συνθήκη είναι ισοδύναμη με τη συνθήκη

$$\mathrm{Var}(Y|X=x) \equiv \sigma_{Y|X}^2 = \sigma^2,$$

δηλαδή η διασπορά της εξαρτημένης μεταβλητής $Y$ είναι η ίδια για κάθε τιμή της $X$ και μάλιστα είναι $\sigma_{Y|X}^2 = \sigma_\epsilon^2 \equiv \sigma^2$. Η τελευταία σχέση προκύπτει από τη σχέση (3.6), αφού οι παράμετροι $\alpha$ και $\beta$ είναι σταθερές και το $x_i$ γνωστό. Η ιδιότητα αυτή λέγεται *ομοσκεδαστικότητα* (δες επίσης 2.2.3) και αντίθετα έχουμε *ετεροσκεδαστικότητα* όταν η διασπορά της $Y$ (ή του σφάλματος $\epsilon$) μεταβάλλεται με τη $X$.

Γενικά για να εκτιμήσουμε τις παραμέτρους της γραμμικής παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων, όπως θα δούμε παρακάτω, δεν είναι απαραίτητο να υποθέσουμε κάποια συγκεκριμένη δεσμευμένη κατανομή $F_Y(y_i|X=x_i)$ της $Y$ ως προς τη $X$. Αν θέλουμε όμως να υπολογίσουμε διαστήματα εμπιστοσύνης για τις παραμέτρους θα χρειαστούμε να υποθέσουμε κανονική δεσμευμένη κατανομή για τη $Y$. Επίσης οι παραπάνω υποθέσεις για γραμμική σχέση και σταθερή διασπορά αποτελούν χαρακτηριστικά πληθυσμών με κανονική κατανομή. Συνήθως λοιπόν σε προβλήματα γραμμικής παλινδρόμησης υποθέτουμε ότι η δεσμευμένη κατανομή της $Y$ είναι κανονική

$$Y|X=x \sim \mathrm{N}(\alpha+\beta x, \sigma^2).$$

### 3.2.2 Σημειακή εκτίμηση παραμέτρων της απλής γραμμικής παλινδρόμησης

Το πρόβλημα της απλής γραμμικής παλινδρόμησης με τις υποθέσεις που ορίστηκαν παραπάνω συνίσταται στην εκτίμηση των τριών παραμέτρων της παλινδρόμησης:

1. της διαφοράς ύψους της ευθείας παλινδρόμησης $\alpha$,

2. της κλίσης της ευθείας παλινδρόμησης $\beta$,

3. της διασποράς σφάλματος της παλινδρόμησης $\sigma^2$.

Τα $\alpha$ και $\beta$ προσδιορίζουν την ευθεία παλινδρόμησης κι άρα καθορίζουν τη γραμμική σχέση εξάρτησης της τ.μ. $Y$ από τη μεταβλητή $X$. Η παράμετρος $\sigma^2$ προσδιορίζει το βαθμό μεταβλητότητας γύρω από την ευθεία παλινδρόμησης κι εκφράζει την αβεβαιότητα της γραμμικής σχέσης.

**Εκτίμηση των παραμέτρων της ευθείας παλινδρόμησης** Η εκτίμηση των παραμέτρων $\alpha$ και $\beta$ γίνεται με τη μέθοδο των **ελαχίστων τετραγώνων** (method of least squares). Η μέθοδος λέγεται έτσι γιατί βρίσκει την ευθεία παλινδρόμησης με παραμέτρους $a$ και $b$ έτσι ώστε το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων από την ευθεία να είναι το ελάχιστο. Οι εκτιμήσεις των $\alpha$ και $\beta$ δίνονται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

$$\min_{\alpha,\beta}\sum_{i=1}^{n}\epsilon_i^2 \qquad \text{ή} \qquad \min_{\alpha,\beta}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2. \tag{3.7}$$

Για να λύσουμε αυτό το πρόβλημα θέτουμε τις μερικές παραγώγους ως προς τα $\alpha$ και $\beta$ ίσες με το μηδέν και καταλήγουμε στο σύστημα δύο εξισώσεων με δύο αγνώστους

$$\left.\begin{array}{l} \frac{\partial\sum(y_i-\alpha-\beta x_i)^2}{\partial\alpha}=0 \\ \frac{\partial\sum(y_i-\alpha-\beta x_i)^2}{\partial\beta}=0 \end{array}\right\} \begin{array}{l} \sum_{i=1}^{n}y_i=n\alpha+\beta\sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_iy_i=\alpha\sum_{i=1}^{n}x_i+\beta\sum_{i=1}^{n}x_i^2 \end{array}$$

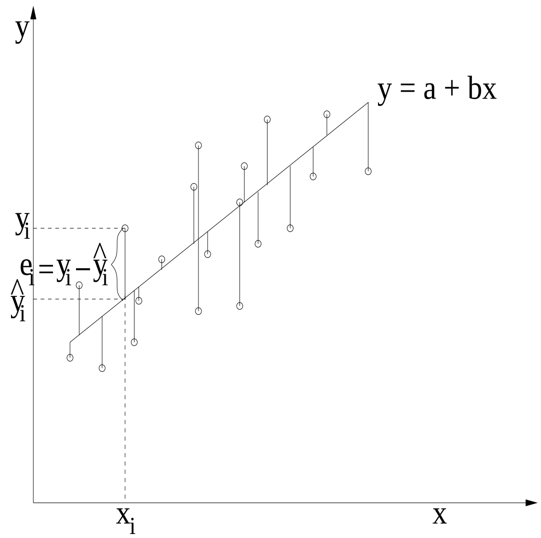
από το οποίο παίρνουμε τις εκτιμήσεις των $\alpha$ και $\beta$

$$b=\frac{s_{XY}}{s_X^2}, \qquad a=\bar{y}-b\bar{x}, \tag{3.8}$$

όπου $s_{XY}$ και $s_X^2$ είναι η δειγματική συνδιασπορά των $X$ και $Y$ και η δειγματική διασπορά της $X$ που ορίστηκαν στις σχέσεις (3.2) και (3.3) αντίστοιχα. Τα $a$ και $b$ ορίζουν την ευθεία

$$\hat{y}=a+bx,$$

που λέγεται κι **ευθεία ελαχίστων τετραγώνων**.




Σχήμα 3.4: Ευθεία ελαχίστων τετραγώνων και υπόλοιπα

**Εκτίμηση της διασποράς των σφαλμάτων παλινδρόμησης** Για κάθε δοθείσα τιμή $x_i$ με τη βοήθεια της ευθείας ελαχίστων τετραγώνων εκτιμούμε την τιμή $\hat{y}_i$ που γενικά είναι διαφορετική από την πραγματική τιμή $y_i$. Η διαφορά

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i$$

είναι η κατακόρυφη απόσταση της πραγματικής τιμής από την ευθεία ελαχίστων τετραγώνων και λέγεται σφάλμα ελαχίστων τετραγώνων ή απλά **υπόλοιπο** (residual). Στο Σχήμα 3.4 απεικονίζονται τα υπόλοιπα της παλινδρόμησης.

Το υπόλοιπο $e_i$ είναι η εκτίμηση του σφάλματος παλινδρόμησης $\epsilon_i$ αντικαθιστώντας απλά τις παραμέτρους παλινδρόμησης με τις εκτιμήσεις ελαχίστων τετραγώνων στον ορισμό του σφάλματος $\epsilon_i = y_i - \alpha - \beta x_i$. Άρα η εκτίμηση της διασποράς $\sigma^2$ του σφάλματος (που είναι κι η δεσμευμένη διασπορά της $Y$ ως προς $X$) δίνεται από τη δειγματική διασπορά $s^2$ των υπολοίπων $e_i$

$$s^2 \equiv \sigma_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{3.9}$$

όπου διαιρούμε με $n-2$ γιατί από τους βαθμούς ελευθερίας $n$ του μεγέθους του δείγματος αφαιρούμε δύο για τις δύο παραμέτρους που έχουν ήδη εκτιμηθεί. Η δειγματική διασπορά $s^2$ μπορεί να εκφραστεί ως προς τις δειγματικές διασπορές των $X$ και $Y$ και της συνδιασποράς τους, αν αντικαταστήσουμε τις εκφράσεις των $a$ και $b$ από την (3.8) στην παραπάνω σχέση (όπου θέτουμε $\hat{y}_i = a + bx_i$)

$$s^2 = \frac{n-1}{n-2}\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right) = \frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2) \tag{3.10}$$

όπου και πάλι υποθέτουμε τις αμερόληπτες εκτιμήτριες για τις διασπορές.

**Παρατηρήσεις**

1. Η ευθεία ελαχίστων τετραγώνων περνάει από το σημείο $(\bar{x}, \bar{y})$ γιατί

$$a + b\bar{x} = \bar{y} - b\bar{x} + b\bar{x} = \bar{y}.$$

   Άρα η ευθεία ελαχίστων τετραγώνων μπορεί επίσης να οριστεί ως

$$y_i - \bar{y} = b(x_i - \bar{x}).$$

2. Η εκτίμηση των $\alpha$ και $\beta$ με τη μέθοδο των ελαχίστων τετραγώνων δεν προϋποθέτει σταθερή διασπορά και κανονική κατανομή της εξαρτημένης μεταβλητής $Y$ για κάθε τιμή της ανεξάρτητης μεταβλητής $X$. Όταν όμως ισχύουν οι δύο αυτές συνθήκες οι εκτιμήτριες ελαχίστων τετραγώνων $a$ και $b$ είναι οι εκτιμήτριες μεγίστης πιθανοφάνειας (κι άρα έχουν και τις επιθυμητές ιδιότητες εκτιμητριών).

3. Αν η διασπορά της $Y$ αλλάζει με το $X$, τότε η διαδικασία των ελαχίστων τετραγώνων πρέπει να διορθωθεί έτσι ώστε να δίνει περισσότερο βάρος στις παρατηρήσεις που αντιστοιχούν σε μικρότερη διασπορά.

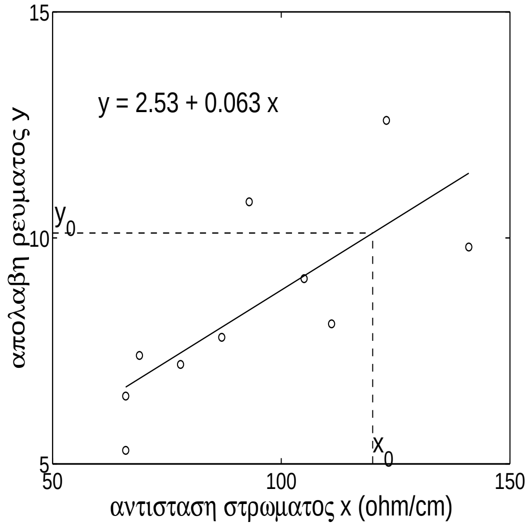
4. Για κάθε τιμή $x_0$ της $X$ μπορούμε να *προβλέψουμε* την αντίστοιχη τιμή $y_0$ της $Y$ από την ευθεία ελαχίστων τετραγώνων, $\hat{y}_0 = a + bx_0$. Εδώ πρέπει να προσέξουμε ότι η τιμή $x_0$ πρέπει να ανήκει στο εύρος τιμών της $X$ που έχουμε από το δείγμα. Για τιμές έξω από αυτό το διάστημα η πρόβλεψη δεν είναι αξιόπιστη.

**Παράδειγμα 3.2** *Θέλουμε να μελετήσουμε σ' ένα ολοκληρωμένο κύκλωμα την εξάρτηση της απολαβής ρεύματος κρυσταλλολυχνίας (τρανζίστορ) από την αντίσταση του στρώματος της κρυσταλλολυχνίας. Στον Πίνακα 3.2 παρουσιάζονται 10 μετρήσεις της απολαβής ρεύματος για αντίστοιχες τιμές της αντίστασης στρώματος της κρυσταλλολυχνίας.*

| Α/Α ($i$) | Αντίσταση στρώματος $x_i$ (ohm/cm) | Απολαβή ρεύματος $y_i$ |
|---|---|---|
| 1 | 66 | 5.3 |
| 2 | 66 | 6.5 |
| 3 | 69 | 7.4 |
| 4 | 78 | 7.2 |
| 5 | 87 | 7.8 |
| 6 | 93 | 10.8 |
| 7 | 105 | 9.1 |
| 8 | 111 | 8.1 |
| 9 | 123 | 12.6 |
| 10 | 141 | 9.8 |

Πίνακας 3.2: Δεδομένα απολαβής ρεύματος τρανζίστορ ($y_i$) για διαφορετικές τιμές της αντίστασης στρώματος κρυσταλλολυχνίας ($x_i$).

*Υποθέτουμε πως η απολαβή ρεύματος της κρυσταλλολυχνίας εξαρτάται γραμμικά από την αντίσταση του στρώματος της και το διάγραμμα διασποράς από το δείγμα στο Σχήμα 3.5 επιβεβαιώνει αυτήν την υπόθεση. Για να εκτιμήσουμε τις παραμέτρους $a$ και $b$ της ευθείας ελαχίστων τετραγών-*




Σχήμα 3.5: Διάγραμμα διασποράς για τα δεδομένα του Πίνακα 3.2 κι ευθεία ελαχίστων τετραγώνων.

*ων υπολογίζουμε πρώτα τα παρακάτω*

$$\bar{x} = 93.9 \qquad \bar{y} = 8.46$$

$$\sum_{i=1}^{10} x_i^2 = 94131 \qquad \sum_{i=1}^{10} y_i^2 = 757.64 \qquad \sum_{i=1}^{10} x_i y_i = 8320.2$$

*και χρησιμοποιώντας τις σχέσεις (3.2) και (3.3) για τη δειγματική συνδιασπορά και διασπορά αντίστοιχα, βρίσκουμε*

$$s_{XY} = 41.81 \qquad s_X^2 = 662.1 \qquad s_Y^2 = 4.66.$$

*Οι εκτιμήσεις $b$ και $a$ είναι*

$$\begin{aligned} b &= \frac{41.81}{662.1} = 0.063 \\ a &= 8.46 - 0.063 \cdot 93.9 = 2.53. \end{aligned}$$

*Από τη σχέση (3.10) υπολογίζουμε την εκτίμηση διασποράς των σφαλμάτων παλινδρόμησης*

$$s^2 = \frac{9}{8}(4.66 - 0.063^2 \cdot 41.81) = 5.056.$$

*Τα αποτελέσματα αυτά ερμηνεύονται ως εξής:*

1. *$b$: Για αύξηση της αντίστασης στρώματος κατά μία μονάδα μέτρησης (1 ohm/cm) η απολαβή του ρεύματος της κρυσταλλολυχνίας αυξάνεται κατά* $0.063$.

2. *$a$: Όταν δεν υπάρχει καθόλου αντίσταση στρώματος ($x = 0$), η απολαβή του ρεύματος είναι $2.53$ μονάδες αλλά βέβαια είναι αδύνατο να θεωρήσουμε στρώμα χωρίς αντίσταση. Δε θα πρέπει λοιπόν να επιχειρήσουμε προβλέψεις για τιμές της αντίστασης στρώματος μικρότερης του $66$ ohm/cm και μεγαλύτερης του $141$ ohm/cm, που είναι οι ακραίες τιμές της αντίστασης του δείγματος.*

3. *$s^2$: Η εκτίμηση της διασποράς γύρω από την ευθεία παλινδρόμησης για κάθε τιμή της $X$ (στο διάστημα τιμών του πειράματος) είναι $5.056$, ή αλλιώς το τυπικό σφάλμα της εκτίμησης της παλινδρόμησης είναι $2.249$ μονάδες, που είναι μεγάλο σε σχέση με το επίπεδο τιμών της $Y$.*

*Με βάση το μοντέλο παλινδρόμησης που εκτιμήσαμε μπορούμε να προβλέψουμε την απολαβή ρεύματος για κάθε αντίσταση στρώματος κρυσταλλολυχνίας στο διάστημα $[66, 141]$ ohm/cm. Στο Σχήμα 3.5 απεικονίζεται η πρόβλεψη της απολαβής ρεύματος για αντίσταση στρώματος $x_0 = 120$ ohm/cm και είναι*

$$y_0 = 2.53 + 0.063 \cdot 120 = 10.11.$$

Για τις παραμέτρους $\alpha$ και $\beta$, καθώς και για τη διασπορά $\sigma^2$, μπορούμε να εκτιμήσουμε διαστήματα εμπιστοσύνης και να κάνουμε στατιστικούς ελέγχους αλλά δε θα ασχοληθούμε με αυτά τα θέματα.

### 3.2.3 Σχέση του συντελεστή συσχέτισης και παλινδρόμησης

Η παλινδρόμηση ορίζεται θεωρώντας την ανεξάρτητη μεταβλητή $X$ σταθερή και την εξαρτημένη μεταβλητή $Y$ τυχαία, ενώ για τη συσχέτιση θεωρούμε και τις δύο μεταβλητές $X$ και $Y$ τυχαίες. Για τις μεταβλητές $X$ και $Y$ της παλινδρόμησης, μπορούμε να αγνοήσουμε ότι η $X$ δεν είναι τ.μ. και να ορίσουμε το συντελεστή συσχέτισης $\rho$ όπως και πριν. Η σχέση μεταξύ του $r$ (της εκτιμήτριας του $\rho$ από το δείγμα) και του συντελεστή της παλινδρόμησης $b$ δίνεται ως εξής (συνδυάζοντας τις σχέσεις $r = \frac{s_{XY}}{s_X s_Y}$ και $b = \frac{s_{XY}}{s_X^2}$)

$$r = b\frac{s_X}{s_Y} \quad \text{ή} \quad b = r\frac{s_Y}{s_X}. \tag{3.11}$$

Και τα δύο μεγέθη, $r$ και $b$, εκφράζουν ποσοτικά τη γραμμική συσχέτιση των μεταβλητών $X$ και $Y$, αλλά το $b$ εξαρτάται από τη μονάδα μέτρησης των $X$ και $Y$ ενώ το $r$, επειδή προκύπτει από το λόγο της συνδιασποράς προς τις τυπικές αποκλίσεις των $X$ και $Y$, δεν εξαρτάται από τη μονάδα μέτρησης των $X$ και $Y$ και δίνει τιμές στο διάστημα $[-1, 1]$. Η σχέση των $r$ και $b$ περιγράφεται ως εξής:

- Αν η συσχέτιση είναι θετική ($r > 0$) τότε η κλίση της ευθείας παλινδρόμησης $b$ είναι επίσης θετική.
- Αν η συσχέτιση είναι αρνητική ($r < 0$) τότε η κλίση της ευθείας παλινδρόμησης $b$ είναι επίσης αρνητική.
- Αν οι μεταβλητές $X$ και $Y$ δε συσχετίζονται ($r = 0$) τότε η ευθεία παλινδρόμησης είναι οριζόντια ($b = 0$).

Επίσης μπορούμε να εκφράσουμε το $r^2$ ως προς τη δειγματική διασπορά του σφάλματος $s^2$ και αντίστροφα

$$s^2 = \frac{n-1}{n-2}s_Y^2(1-r^2) \quad \text{ή} \quad r^2 = 1 - \frac{n-2}{n-1}\frac{s^2}{s_Y^2}. \tag{3.12}$$

Η παραπάνω σχέση δηλώνει πως όσο μεγαλύτερο είναι το $r^2$ (ή το $|r|$) τόσο μειώνεται η διασπορά του σφάλματος της παλινδρόμησης, δηλαδή τόσο ακριβέστερη είναι η πρόβλεψη που βασίζεται στην ευθεία παλινδρόμησης.

**Παράδειγμα 3.3** *Στο παραπάνω παράδειγμα, ο συντελεστής συσχέτισης του κέρδους ρεύματος της κρυσταλλολυχνίας και της αντίστασης στρώματος είναι*

$$r = \frac{s_{XY}}{s_X\, s_Y} = \frac{41.81}{\sqrt{662.1 \cdot 4.66}} = 0.753$$

*που θα μπορούσαμε να υπολογίσουμε κι από τις σχέσεις (3.11) ή (3.12). Ο συντελεστής συσχέτισης δηλώνει την ασθενή θετική συσχέτιση του κέρδους ρεύματος και της αντίστασης στρώματος, που δε μπορούμε όμως να συμπεράνουμε από την τιμή του συντελεστή παλινδρόμησης* $b = 0.063$ *ή την τιμή της διασποράς των σφαλμάτων* $s^2 = 5.056$ *γιατί εξαρτιούνται από τις μονάδες μέτρησης των δύο μεταβλητών.*

Σ' αυτό το κεφάλαιο ασχοληθήκαμε μόνο με την απλή γραμμική παλινδρόμηση. Η μελέτη της παλινδρόμησης επεκτείνεται στη μη-γραμμική παλινδρόμηση, που αποτελεί γενικότερη και πιο ρεαλιστική (αλλά και πιο πολύπλοκη) προσέγγιση για τον προσδιορισμό της εξάρτησης της $Y$ από τη $X$. Επίσης η τ.μ. $Y$ μπορεί να εξαρτάται από περισσότερες από μια μεταβλητές που είναι το πρόβλημα της πολλαπλής παλινδρόμησης (γραμμική ή μη-γραμμική).