Αριστοτέλειο
Πανεπιστήμιο
Θεσσαλονίκης

Στατιστική για Χημικούς Μηχανικούς

Συσχέτιση και Γραμμική Παλινδρόμηση

Κουγιουμτζής Δημήτριος
Τμήμα Χημικών Μηχανικών

Θεσσαλονίκη, Μάιος 2015

Άδειες Χρήσης

Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons. Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα. Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.

ανοικτάμαθήματα
opencourses

Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Σημείωμα Ιστορικού Εκδόσεων Έργου

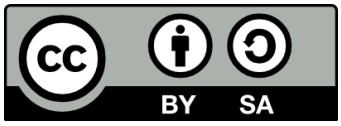
Το παρόν έργο αποτελεί την έκδοση 1.00.

Σημείωμα Αναφοράς

Copyright Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Κουγιουμτζής Δημήτριος, 2015. «Στατιστική για Χημικούς Μηχανικούς. Συσχέτιση και Γραμμική Παλινδρόμηση.». Έκδοση: 1.0. Θεσσαλονίκη 2015. Διαθέσιμο από τη δικτυακή διεύθυνση: http://eclass.auth.gr/courses/OCRS248/.

Σημείωμα Αδειοδότησης

Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά, Παρόμοια Διανομή 4.0[1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».



Διατήρηση Σημειωμάτων

- Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:
- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.

[1] http://creativecommons.org/licenses/by- sa/4.0/

Κεφάλαιο 4

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 ή φυσικού φαινομένου ενδιαφερόμαστε να προσδιορίσουμε τη σχέση μεταξύ δύο μεταβλητών. Για παράδειγμα σε μια μελέτη για τη μόλυνση του αέρα μπορεί να μας ενδιαφέρει η σχέση της συγκέντρωσης όζονος και της συγκέντρωσης διοξειδίου του άνθρακα. 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 εκπομπή νιτρικού οξέος μιας μηχανής είναι τ.μ. με κατανομή που μεταβάλλεται με την παλαιότητα της μηχανής. Συνήθως η μεταβολή αφορά μόνο τη μέση τιμή (και σπανιότερα 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.

4.1 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 η περιεκτικότητα του αέρα σε διοξείδιο του άνθρακα σε μια περιοχή και η κατανάλωση υγρών καυσίσμων (πετρέλαιο θέρμανσης, βενζίνη, κτλ) στην ίδια περιοχή μπορούν να θεωρηθούν σαν δύο τ.μ. που συσχετίζονται, όπου η περιεκτικότητα του αέρα σε διοξείδιο του άνθρακα εξαρτάται από την κατανάλωση υγρών καυσίμων (το αντίθετο δεν έχει πρακτική σημασία). Μπορούμε επίσης να θεωρήσουμε τη συσχέτιση της περιεκτικότητας του αέρα σε διοξείδιο του άνθρακα και την περιεκτικότητα σε όζον, αλλά τώρα δεν εξαρτάται η μια τ.μ. από την άλλη παρά εξαρτιούνται κι οι δύο από παράγοντες μόλυνσης, όπως η κατανάλωση υγρών καυσίμων.

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.

4.1.1 Ο συντελεστής συσχέτισης ρ

Ο βαθμός της *γραμμικής* συσχέτισης δύο τ.μ. X και Y με διασπορά σ_X^2 και σ_Y^2 αντίστοιχα και συνδιασπορά $\sigma_{XY} = \mathrm{Cov}(X, Y) = \mathrm{E}(X, Y) - \mathrm{E}(X)\mathrm{E}(Y)$, μετριέται με τον **συνετελεστή συσχέτισης** (correlation coefficient) ρ που ορίζεται ως

$$\rho = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y}. \tag{4.1}$$

Ο συντελεστής συσχέτισης ρ, όπως και η συνδιασπορά σ_{XY}, εκφράζει το βαθμό και τον τρόπο που οι δύο μεταβλητές συσχετίζονται, δηλαδή πως η μία τ.μ. μεταβάλλεται ως προς την άλλη. Η σ_{XY} παίρνει τιμές που εξαρτώνται από το πεδίο τιμών των X και Y ενώ ο συντελεστής ρ παίρνει τιμές στο διάστημα $[-1, 1]$. Οι χαρακτηριστικές τιμές του ρ ερμηνεύονται ως εξής:

- $\rho = 1$: υπάρχει *τέλεια θετική* συσχέτιση μεταξύ των X και Y,
- $\rho = 0$: δεν υπάρχει καμιά (γραμμική) συσχέτιση μεταξύ των X και Y,
- $\rho = -1$: υπάρχει *τέλεια αρνητική* συσχέτιση 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 συνετελεστής συσχέτισης είναι κοντά στο -1 ή 1 η γραμμική συσχέτιση των δύο τ.μ. είναι ισχυρή (συνήθως χαρακτηρίζουμε ισχυρές τις συσχετίσεις όταν $|\rho| > 0.9$) ενώ όταν είναι κοντά στο 0 οι τ.μ. είναι πρακτικά ασυσχέτιστες.

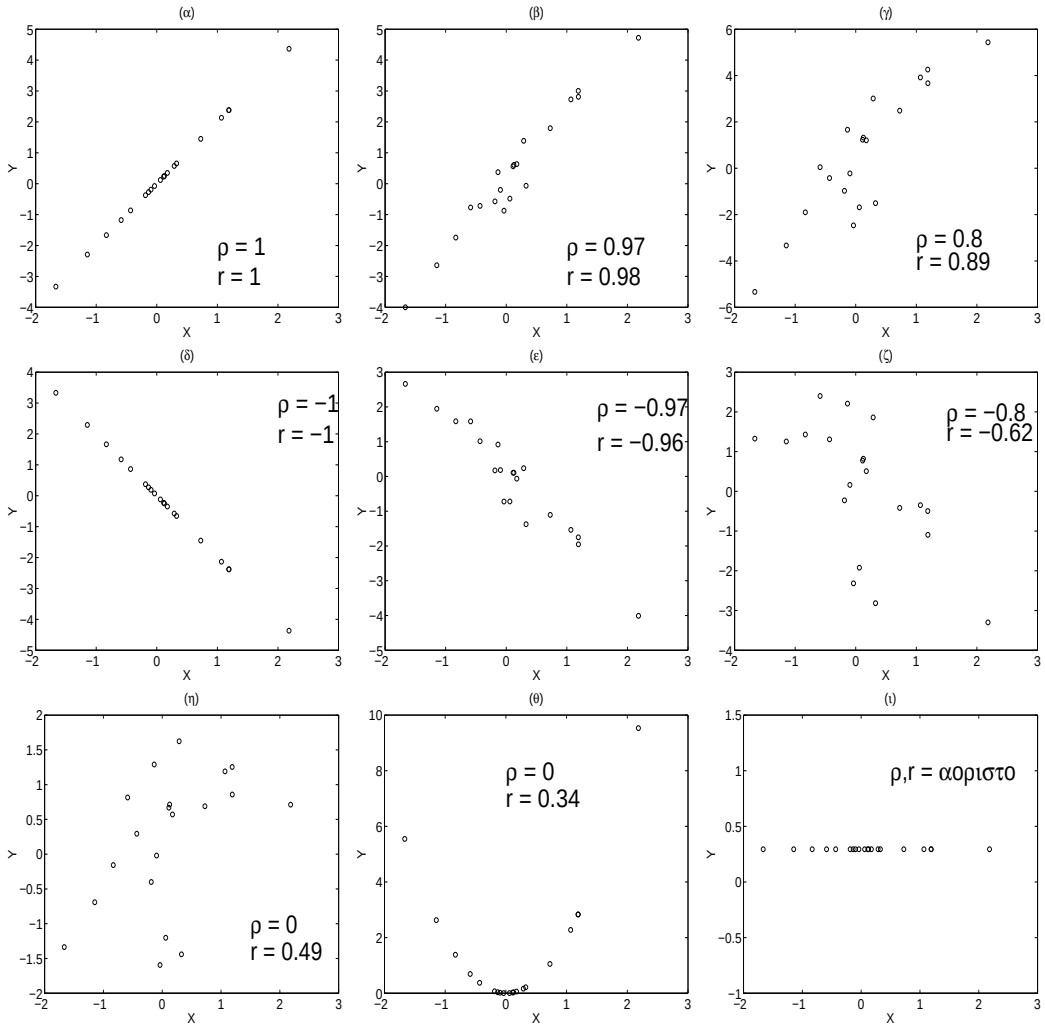
Όπως φαίνεται από τον ορισμό στη σχέση (4.1), ο συντελεστής συσχέτισης ρ δεν εξαρτάται από τη μονάδα μέτρησης των X και Y και είναι συμμετρικός ως προς τις X και Y.

4.1.2 Σημειακή εκτίμηση του συντελεστή συσχέτισης

Όταν έχουμε παρατηρήσεις των δύο τ.μ. X και Y κατά ζεύγη $(x_1, y_1), \ldots, (x_n, y_n)$, 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 4.1 παρουσιάζονται τυπικά διαγράμματα διασποράς για ισχυρές κι ασθενείς συσχετίσεις δύο τ.μ. X και Y. Στα Σχήματα 4.1α και 4.1δ η σχέση είναι τέλεια ($\rho = 1$ και $\rho = -1$ αντίστοιχα), στα Σχήματα 4.1β και 4.1ε είναι ισχυρή (θετική με $\rho = 0.97$ κι αρνητική με $\rho = -0.97$ αντίστοιχα) και στα Σχήματα 4.1γ και 4.1ζ είναι λιγότερο ισχυρή (θετική με $\rho = 0.8$ κι αρνητική με $\rho = -0.8$ αντίστοιχα). Στο Σχήμα 4.1η είναι $\rho = 0$ γιατί οι τ.μ. X και Y είναι ανεξάρτητες ενώ στο Σχήμα 4.1θ είναι πάλι $\rho = 0$ αλλά οι X και Y δεν είναι ανεξάρτητες αλλά συσχετίζονται μόνο μη-γραμμικά. Τέλος για το Σχήμα 4.1ι ο συντελεστής συσχέτισης δεν ορίζεται γιατί η Y είναι σταθερή ($\sigma_Y = 0$ στον ορισμό του ρ στην (4.1)).

Η σημειακή εκτίμηση του συντελεστή συσχέτισης ρ του πληθυσμού από το δείγμα των n ζευγαρωτών παρατηρήσεων των X και Y γίνεται με την αντικατάσταση στη σχέση (4.1) της συνδιασποράς σ_{XY} και των διασπορών σ_X^2 και σ_Y^2 από τις αντίστοιχες εκτιμήσεις από το δείγμα

$$\hat{\rho} \equiv r = \frac{s_{XY}}{s_X \, s_Y}. \tag{4.2}$$



Σχήμα 4.1: Διάγραμμα διασποράς δύο τ.μ. X και Y από $n = 20$ παρατηρήσεις για θετική συσχέτιση στα σχήματα (α), (β) και (γ), για αρνητική συσχέτιση στα σχήματα (δ), (ε) και (ζ) και για ασυσχέτιστες τ.μ. στα σχήματα (η), (θ) και (ι). Σε κάθε σχήμα δίνεται η πραγματική τιμή του συντελεστή συσχέτισης ρ κι η δειγματική r. Στο (ι) ο συντελεστής συσχέτισης δεν ορίζεται.

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

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \tag{4.3}$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \tag{4.4}$$

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

έκφραση της εκτιμήτρια r

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. \tag{4.5}$$

Είναι προφανές πως η παραπάνω σχέση για το r δεν αλλάζει αν θεωρήσουμε τις μεροληπτικές εκτιμήτριες των σ_{XY}, σ_X και σ_Y.

Στο Σχήμα 4.1 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 ρ. Αυτό συμβαίνει γιατί η εκτιμήτρια r όπως δίνεται στη σχέση (4.5) 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 συσχέτιση δύο τ.μ..

Παράδειγμα 4.1. *Θέλουμε να διερευνήσουμε τη συσχέτιση της περιεκτικότητας σε όζου και σε δευτερογενή άνθρακα στον αέρα κάποιας περιοχής. Γι αυτό έγιναν 20 διαφορετικές μετρήσεις της περιεκτικότητας σε όζου (σε ppm) και της περιεκτικότητας σε δευτερογενή άνθρακα (σε $\mu g/m^3$) που παρουσιάζονται στον Πίνακα 4.1. Για να βρούμε τον συντελεστή συσχέτισης r υπολογίζουμε πρώτα τα παρακάτω*

$$\bar{x} = 0.175 \qquad \bar{y} = 12.95$$

$$\sum_{i=1}^{20} x_i^2 = 0.633 \qquad \sum_{i=1}^{20} y_i^2 = 3860.17 \qquad \sum_{i=1}^{20} x_i y_i = 47.74.$$

Αντικαθιστώντας τα παραπάνω αποτελέσματα στη σχέση (4.5) βρίσκουμε

$$r = \frac{47.74 - 20 \cdot 0.175 \cdot 12.95}{\sqrt{(0.633 - 20 \cdot 0.175^2)\,(3860.17 - 20 \cdot 12.95^2)}} = 0.74.$$

Η τιμή $r = 0.74$ υποδηλώνει ότι η περιεκτικότητα σε όζου και η περιεκτικότητα σε δευτερογενή άνθρακα έχουν γραμμική θετική συσχέτιση αλλά όχι ισχυρή. Αυτό φαίνεται κι από το διάγραμμα διασποράς στο Σχήμα 4.2. Η μεταβλητότητα της μιας τ.μ. (περιεκτικότητα του αέρα σε όζου ή σε δευτερογενή άνθρακα) μπορεί να εξηγηθεί από τη συσχέτιση της με την άλλη κατά ποσοστό που δίνεται από το συντελεστή προσδιορισμού, που είναι $r^2 \cdot 100 = 0.74^2 \cdot 100 = 0.55\%$. Συμπεραίνουμε λοιπόν πως η γνώση της μιας τ.μ. δε μας επιτρέπει να προσδιορίσουμε την άλλη με ακρίβεια. Πρέπει επίσης να σημειωθεί ότι η εκτίμηση r του συνετελεστή συσχέτισης μπορεί να αλλάξει σημαντικά με την πρόσθεση ή αφαίρεση λίγων ζευγών παρατηρήσεων γιατί το μέγεθος του δείγματος είναι μικρό.

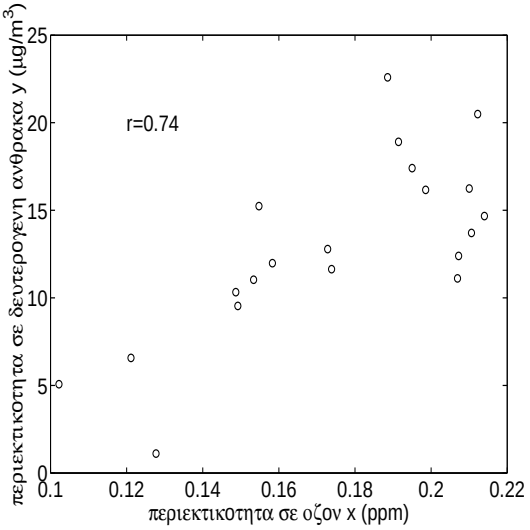
Για τον συντελεστή συσχέτισης ρ μπορούμε να υπολογίσουμε διαστήματα εμπιστοσύνης αν θεωρήσουμε γνωστή την κατανομή της εκτιμήτριας r. Μπορούμε επίσης να κάνουμε στατιστικό έλεγχο για κάποια τιμή r_0, αλλά εδώ δε θα ασχοληθούμε με αυτά τα θέματα.

Α/Α	Περιεκτικότητα σε όζον x_i (ppm)	Περιεκτικότητα σε δευτερογενή άνθρακα y_i ($\mu g/m^3$)
1	0.102	5.07
2	0.121	6.57
3	0.128	1.11
4	0.149	10.32
5	0.149	9.54
6	0.153	11.03
7	0.155	15.23
8	0.158	11.98
9	0.173	12.78
10	0.174	11.64
11	0.189	22.59
12	0.191	18.91
13	0.195	17.41
14	0.199	16.17
15	0.207	11.11
16	0.207	12.39
17	0.210	16.24
18	0.211	13.71
19	0.212	20.49
20	0.214	14.67

Πίνακας 4.1: Δεδομένα περιεκτικότητας σε όζον (x_i) και της περιεκτικότητας σε δευτερογενή άνθρακα (y_i) του αέρα.

4.2 Απλή Γραμμική Παλινδρόμηση

Στη *συσχέτιση* που μελετήσαμε παραπάνω μετρήσαμε με το συντελεστή συσχέτισης τη γραμμική συσχέτιση 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). Η ανεξάρτητη μεταβλητή δε θεωρείται τυχαία αλλά παίρνει καθορισμένες τιμές που διαλέγουμε εμείς ή δίνονται από το πρόβλημα που μελετάμε. 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, όταν αυτές δεν ορίζονται ξεκάθαρα από το πρόβλημα. Για παράδειγμα, όταν μετράμε τη διατμητική αντοχή του αργίλου σε διάφορα βάθη, θέλουμε να μελετήσουμε την (γραμμική) εξάρτηση της διατμητικής αντοχής του αργίλου από το βάθος τους εδάφους και γι αυτό η διατμητική αντοχή του αργίλου είναι η εξαρτημένη μεταβλητή Y και το βάθος του εδάφους η ανεξάρτητη μεταβλητή X.



Σχήμα 4.2: Διάγραμμα διασποράς για το δείγμα παρατηρήσεων περιεκτικότητας του αέρα σε όζον και δευτερογενή άνθρακα του Πίνακα 4.1.

4.2.1 Το πρόβλημα της γραμμικής παλινδρόμησης

Η εξαρτημένη τ.μ. Y ακολουθεί κάποια κατανομή με αθροιστική συνάρτηση κατανομής $F_Y(y|X = x)$, δεσμευμένη για κάθε τιμή x της μεταβλητής X. Περιορίζουμε τη μελέτη του προβλήματος στη μέση τιμή και υποθέτουμε εδώ ότι η εξάρτηση εκφράζεται από μια γραμμική σχέση

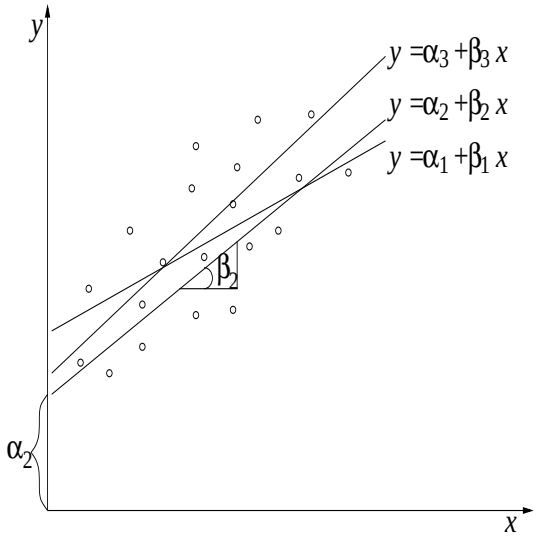
$$\mathrm{E}(Y|X = x) = \alpha + \beta x \tag{4.6}$$

και η σχέση αυτή λέγεται **γραμμική παλινδρόμηση της Y στη X** (linear regression) [οι τιμές της Y για κάθε τιμή $X = x$ παλινδρομούνται γύρω από το σημείο $y = \mathrm{E}(Y|X = x)$ της ευθείας $y = \alpha + \beta x$, δηλαδή οι τιμές της Y για κάθε τιμή της X βρίσκονται πάνω και κάτω από αυτήν την ευθεία]. Το πρόβλημα της παλινδρόμησης είναι η εύρεση των παραμέτρων α και β που εκφράζουν καλύτερα τη γραμμική εξάρτηση της Y από τη X. Κάθε ζεύγος τιμών (α, β) καθορίζει μια διαφορετική γραμμική σχέση που εκφράζεται γεωμετρικά από ευθεία γραμμή και οι δύο παράμετροι ορίζονται ως:

- Ο σταθερός όρος α είναι η τιμή του y για $x = 0$ (intercept).
- Ο συντελεστής β του x 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 4.3.

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 το $X = x_i$, αλλά θα χρησιμοποιήσουμε εδώ τον ίδιο συμβολισμό y_i για την τ.μ. και την παρατήρηση]. Η τ.μ.



Σχήμα 4.3: Ευθείες γραμμικής παλινδρόμησης

y_i για κάποια τιμή x_i της X θα δίνεται κάτω από την υπόθεση της γραμμικής παλινδρόμησης ως

$$y_i = a + \beta x_i + \epsilon_i, \tag{4.7}$$

όπου ϵ_i είναι κι αυτή τ.μ., λέγεται **σφάλμα παλινδρόμησης** (regression error) κι ορίζεται ως η διαφορά της y_i από τη δεσμευμένη μέση τιμή $\mathrm{E}(Y|X = x_i)$ (δες σχέση (4.6)).

Για την ανάλυση της γραμμικής παλινδρόμησης κάνουμε τις παρακάτω υποθέσεις:

- Η μεταβλητή X είναι *ελεγχόμενη* για το πρόβλημα που μελετάμε, δηλαδή γνωρίζουμε τις τιμές της χωρίς καμιά αμφιβολία.
- Η σχέση (4.6) 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 $\mathrm{Var}(Y|X = x) = \sigma^2_{Y|X}$, δηλαδή ότι η διασπορά της εξαρτημένης μεταβλητής Y είναι η ίδια για κάθε τιμή της X και μάλιστα είναι $\sigma^2_{Y|X} = \sigma^2_\epsilon \equiv \sigma^2$, όπως προκύπτει από τη σχέση (4.7), αφού οι παράμετροι a και β είναι σταθερές και το x_i γνωστό. Η ιδιότητα αυτή λέγεται *ομοσκεδαστικότητα* (δες επίσης 2.2.4) και αντίθετα έχουμε *ετεροσκεδαστικότητα* όταν η διασπορά της Y (ή του σφάλματος ϵ) 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

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

4.2.2 Σημειακή εκτίμηση των παραμέτρων της γραμμικής παλινδρόμησης

Λύση στο πρόβλημα της γραμμικής παλινδρόμησης με τις υποθέσεις που ορίστηκαν παραπάνω αποτελεί ο προσδιορισμός της σταθερού όρου της παλινδρόμησης a και του συντελεστή της παλινδρόμησης β για να γνωρίζουμε την ευθεία της παλινδρόμησης αλλά και της διασποράς σ^2 για να γνωρίζουμε το βαθμό μεταβλητότητας γύρω από την ευθεία.

Εκτίμηση των παραμέτρων της ευθείας παλινδρόμησης Η εκτίμηση των παραμέτρων a και β 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 a και β δίνονται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

$$\min_{a,\beta} \sum_{i=1}^{n} \epsilon_i^2 \qquad \text{ή} \qquad \min_{a,\beta} \sum_{i=1}^{n} (y_i - a - \beta x_i)^2. \tag{4.8}$$

Για να λύσουμε αυτό το πρόβλημα θέτουμε τις μερικές παραγώγους ως προς τα a και β ίσες με το μηδέν και καταλήγουμε στο σύστημα δύο εξισώσεων με δύο αγνώστους

$$\left.\begin{array}{l} \dfrac{\partial \sum_{i=1}^{n} (y_i - a - \beta x_i)^2}{\partial a} = 0 \\ \dfrac{\partial \sum_{i=1}^{n} (y_i - a - \beta x_i)^2}{\partial \beta} = 0 \end{array}\right\} \quad \begin{array}{l} \sum_{i=1}^{n} y_i = na + \beta \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 \end{array}$$

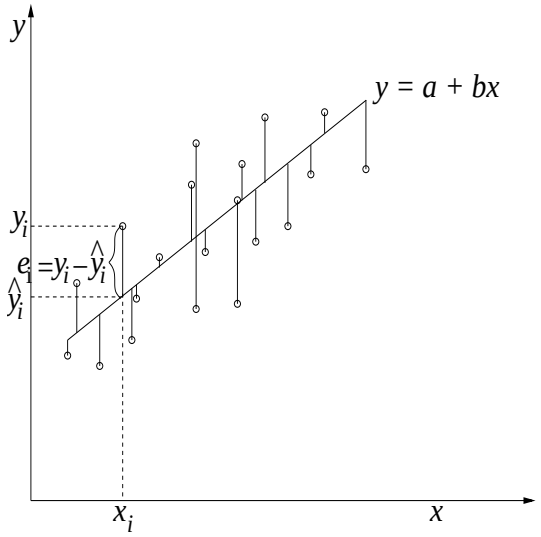
από το οποίο παίρνουμε τις εκτιμήσεις των a και β

$$b = \frac{s_{XY}}{s_X^2}, \qquad a = \bar{y} - b\bar{x}, \tag{4.9}$$

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 (4.3) και (4.4) αντίστοιχα. Τα a και b ορίζουν την ευθεία, $\hat{y} = a + bx$, που λέγεται κι **ευθεία ελαχίστων τετραγώνων**.

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 $e_i = y_i - \hat{y}_i$ είναι η κατακόρυφη απόσταση της πραγματικής τιμής από την ευθεία ελαχίστων τετραγώνων και λέγεται σφάλμα ελαχίστων τετραγώνων ή απλά **υπόλοιπο** (residual). Στο Σχήμα 4.4 απεικονίζονται τα υπόλοιπα της παλινδρόμησης.

Το υπόλοιπο e_i είναι η εκτίμηση του σφάλματος παλινδρόμησης ϵ_i αντικαθιστώντας απλά τις παραμέτρους παλινδρόμησης a και β με τις εκτιμήσεις ελαχίστων τετραγώνων a και b στον ορισμό του σφαλμάτος $\epsilon_i = y_i - a + \beta x_i$. Άρα η εκτίμηση της διασποράς σ^2 του σφάλματος (που



Σχήμα 4.4: Ευθεία ελαχίστων τετραγώνων και υπόλοιπα

είναι κι η δεσμεύμενη διασπορά της Y ως προς X) δίνεται από τη δειγματική διασπορά s^2 των υπολοίπων e_i

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{4.10}$$

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 (4.9) στην παραπάνω σχέση (όπου θέτουμε $\hat{y}_i = a + bx_i$)

$$s^2 = \frac{n-1}{n-2}\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right) = \frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2) \tag{4.11}$$

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 a και β 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 δεσμευμένης κατανομής της Y αλλάζει με τη X, τότε η διαδικασία των ελαχίστων τετραγώνων πρέπει να διορθωθεί έτσι ώστε να δίνει περισσότερο βάρος στις παρατηρήσεις που αντιστοιχούν σε μικρότερη διασπορά.

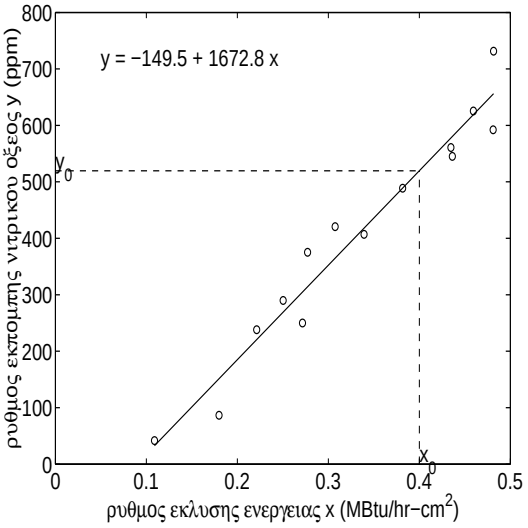
4. Για κάθε τιμή x_0 της X μπορούμε να *προβλέψουμε* την αντίστοιχη τιμή y_0 της Y από την ευθεία ελαχίστων τετραγώνων, $y_0 = a + bx_0$. Εδώ πρέπει να προσέξουμε ότι η τιμή x_0 πρέπει να ανήκει στο εύρος τιμών της X που έχουμε από το δείγμα. Για τιμές έξω από αυτό το διάστημα η πρόβλεψη είναι επικίνδυνη. Επίσης οι προβλέψεις έχουν νόημα όταν η διασπορά του σφάλματος είναι σχετικά μικρή.

Παράδειγμα 4.2. *Έγινε ένα πείραμα σε λέβητες παραγωγής θερμότητας για τη μελέτη της εξάρτησης της παραγωγής νιτρικού οξέος από το 'ρυθμό έκλυσης σε επιφάνεια καύσης' (burner area liberation rate) (είναι ένα μέτρο της ενέργειας που παράγεται από τη μονάδα ανά τετραγωνικό εκατοστό της επιφάνειας του καυστήρα). Τα αποτελέσματα δίνονται στον Πίνακα 4.2.*

Α/Α	Ρυθμός έκλυσης από καυστήρα x_i (MBtu/hr-cm^2)	Ρυθμός εκπομπής νιτρικού οξέος y_i (ppm)
1	0.109	41.7
2	0.180	86.5
3	0.221	238.1
4	0.250	289.9
5	0.272	250.1
6	0.277	375.2
7	0.307	420.7
8	0.339	407.0
9	0.382	488.8
10	0.435	560.8
11	0.436	545.1
12	0.459	625.4
13	0.481	592.3
14	0.482	731.5

Πίνακας 4.2: Δεδομένα ρυθμού έκλυσης από καυστήρα (x_i) και ρυθμού εκπομπής νιτρικού οξέος (y_i).

Υποθέτουμε πως ο ρυθμός εκπομπής νιτρικού οξέος εξαρτάται γραμμικά από την παραγωγή ενέργειας και το διάγραμμα διασποράς στο Σχήμα 4.5 από το δείγμα των 14 μονάδων παραγωγής θερμότητας επιβεβαιώνει αυτήν την υπόθεση. Για να εκτιμήσουμε τις παραμέτρους a και b της



Σχήμα 4.5: Διάγραμμα διασποράς για τα δεδομένα του Πίνακα 4.2 κι ευθεία ελαχίστων τετραγώνων.

ευθείας ελαχίστων τετραγώνων υπολογίζουμε πρώτα τα παρακάτω

$$\bar{x} = 0.331 \qquad \bar{y} = 403.8$$

$$\sum_{i=1}^{14} x_i^2 = 1.716 \qquad \sum_{i=1}^{14} y_i^2 = 2823556.9 \qquad \sum_{i=1}^{14} x_i y_i = 2177.42$$

και χρησιμοποιώντας τις σχέσεις (4.3) και (4.4) για τη δειγματική συνδιασπορά και διασπορά αντίστοιχα, βρίσκουμε

$$s_{XY} = 23.66 \qquad s_X^2 = 0.014 \qquad s_Y^2 = 41603.9$$

Οι εκτιμήσεις b και a είναι

$$\begin{aligned} b &= \frac{23.66}{0.014} = 1672.85 \\ a &= 403.8 - 1672.85 \cdot 0.331 = -149.50. \end{aligned}$$

Από τη σχέση (4.11) υπολογίζουμε την εκτίμηση διασποράς των σφαλμάτων παλινδρόμησης

$$s^2 = \frac{13}{12}(41603.9 - 1672.85^2 \cdot 0.014) = 2186.0.$$

Τα αποτελέσματα αυτά ερμηνεύονται ως εξής:

1. $b = 1672.85$*: Για αύξηση του ρυθμού έκλυσης ενέργειας κατά μία μονάδα μέτρησης (1 MBtu/hr-cm^2) ο ρυθμός εκπομπής νιτρικού οξέος αυξάνεται κατά* 1672.85 *ppm.*

2. $a = -149.5$*: Όταν δεν εκλύεται ενέργεια από την επιφάνεια του καυστήρα ($x = 0$), η εκπομπή του νιτρικού οξέος είναι* -149.5 *ppm, που φυσικά είναι αδύνατον. Αυτό συμβαίνει γιατί η τιμή $x = 0$ δεν ανήκει στο διάστημα τιμών του ρυθμού έκλυσης ενέργειας που μετρήθηκαν στο δείγμα. Δε θα πρέπει λοιπόν να επιχειρήσουμε προβλέψεις για τιμές του ρυθμού έκλυσης ενέργειας μικρότερες του* 0.1 *MBtu/hr-cm^2 και μεγαλύτερες του* 0.5 *MBtu/hr-cm^2 (προσεγγιστικά).*

3. $s^2 = 2186.0$: *Η διασπορά γύρω από την ευθεία παλινδρόμησης για κάθε τιμή του X (στο διάστημα τιμών του πειράματος) εκτιμάται να είναι 2186, ή αλλιώς το τυπικό σφάλμα της εκτίμησης της παλινδρόμησης είναι 46.75 ppm, που είναι μικρό σε σχέση με το επίπεδο τιμών της Y και άρα το μοντέλο της γραμμική παλινδρόμησης εξηγεί ικανοποιητικά τη σχέση του ρυθμού εκπομπής νιτρικού οξέος και ρυθμού έκλυσης ενέργειας από την επιφάνεια του καυστήρα και επιτρέπει καλές προβλέψεις.*

Με βάση το μοντέλο παλινδρόμησης που εκτιμήσαμε μπορούμε να προβλέψουμε το ρυθμό εκπομπής νιτρικού οξέος για κάθε τιμή του ρυθμού έκλυσης ενέργειας στο διάστημα $[0.1, 0.5]$ ppm (προσεγγιστικά). Στο Σχήμα 4.5 απεικονίζεται η πρόβλεψη του ρυθμού εκπομπής νιτρικού οξέος για ρυθμό έκλυσης ενέργειας $x_0 = 0.4$ ppm και είναι

$$y_0 = -149.5 + 1672.85 \cdot 0.4 = 519.6.$$

Για τις παραμέτρους a και β, καθώς και για τη διασπορά σ^2, μπορούμε να εκτιμήσουμε διαστήματα εμπιστοσύνης και να κάνουμε στατιστικούς ελέγχους.

4.2.3 Σχέση του συντελεστή συσχέτισης και παλινδρόμησης

Η παλινδρόμηση ορίζεται θεωρώντας την ανέξαρτητη μεταβλητή X καθορισμένη 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 ρ όπως και πριν. Η σχέση μεταξύ του r (της εκτιμήτριας του ρ από το δείγμα) και του συντελεστή της παλινδρόμησης b (της εκτιμήτριας του β από το δείγμα) δίνεται ως εξής (συνδιάζοντας τις σχέσεις $r = \frac{s_{XY}}{s_X s_Y}$ και $b = \frac{s_{XY}}{s_X^2}$)

$$r = b\frac{s_X}{s_Y} \quad \text{ή} \quad b = r\frac{s_Y}{s_X}. \tag{4.12}$$

Και τα δύο μεγέθη, r και b, εκφράζουν ποιοτικά τη γραμμική συσχέτιση των μεταβλητών X και Y, αλλά το b εξαρτάται από τη μονάδα μέτρησης των X και Y ενώ το r παίρνει τιμές στο διάστημα $[-1, 1]$. Έτσι 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 είναι $b < 0$ και αν οι μεταβλητές X και Y δε συσχετίζονται ($r = 0$) τότε η ευθεία παλινδρόμησης είναι οριζόντια ($b = 0$).

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

$$s^2 = \frac{n-1}{n-2}s_Y^2(1-r^2) \quad \text{ή} \quad r^2 = 1 - \frac{n-2}{n-1}\frac{s^2}{s_Y^2}. \tag{4.13}$$

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

Παράδειγμα 4.3. *Στο παραπάνω παράδειγμα, ο συντελεστής συσχέτισης του ρυθμού της εκπομπής νιτρικού οξέος και του ρυθμού έκλυσης ενέργειας του καυστήρα είναι*

$$r = \frac{s_{XY}}{s_X\, s_Y} = \frac{23.66}{\sqrt{0.014 \cdot 41603.9}} = 0.975$$

που θα μπορούσαμε να υπολογίσουμε κι από τις σχέσεις (4.12) ή (4.13). Ο συντελεστής συσχέτισης δηλώνει την πολύ ισχυρή θετική συσχέτιση του ρυθμού της εκπομπής νιτρικού οξέος και και του ρυθμού έκλυσης ενέργειας, που δε μπορούμε όμως εύκολα να συμπεράνουμε από την τιμή του συντελεστή παλινδρόμησης $b = 1672.85$ *ή την τιμή της διασποράς των σφαλμάτων* $s^2 = 2186.0$ *γιατί οι τιμές αυτές ορίζονται σε σχέση με το πεδίο τιμών του ρυθμού της εκπομπής νιτρικού οξέος και και του ρυθμού έκλυσης ενέργειας.*

Σ' αυτό το κεφάλαιο ασχοληθήκαμε μόνο με την απλή γραμμική παλινδρόμηση. Η μελέτη της παλινδρόμησης επεκτείνεται στη μή-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 μή-γραμμική).