Κεφάλαιο 5

Συσχέτιση και Παλινδρόμηση

Στο προηγούμενο κεφάλαιο μελετήσαμε τη διάδοση του σφάλματος από μια τυχαία μεταβλητή X σε μια τ.μ. Y που δίνεται ως συνάρτηση της X. Σε αυτό το κεφάλαιο θα διερευνήσουμε τη συσχέτιση των δύο μεταβλητών X και Y και θα εκτιμήσουμε τη συνάρτηση που δίνει τη Y γνωρίζοντας τη X. Επίσης θα επεκτείνουμε τη μελέτη για την περίπτωση που θέλουμε να εκτιμήσουμε τη Y γνωρίζοντας περισσότερες από μια μεταβλητές.

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 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 και 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 μας ενδιαφέρει η μεταβολή της μέσης τιμής (και ορισμένες φορές και η διασπορά), για αυτό και 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 πρώτα την απλή γραμμική παλινδρόμηση, δηλαδή όταν η συνάρτηση παλινδρόμησης είναι γραμμική ως προς μια τ.μ. X. Στη συνέχεια θα μελετήσουμε κάποιες μορφές μη-γραμμικής παλινδρόμησης, καθώς και πολλαπλής παλινδρόμησης, όπου η Y εξαρτάται από περισσότερες από μια μεταβλητές.

5.1 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 και οι δύο από άλλες μεταβλητές, όπως η ταχύτητα του κινητήρα. Η συσχέτιση λοιπόν δεν υποδηλώνει απαραίτητα κάποια αιτιακή σχέση των δύο μεταβλητών. Η παρατήρηση ότι σε κάποια περιοχή ο πληθυσμός των πελαργών συσχετίζεται με το πλήθος των γεννήσεων, δε σημαίνει πως οι πελαργοί προκαλούν γεννήσεις.

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. Στην παράγραφο 2.2.4 ορίσαμε τη συνδιασπορά σ_{XY} και το συντελεστή συσχέτισης ρ δύο τ.μ. X και Y με διασπορά σ_X^2 και σ_Y^2 αντίστοιχα (δες τη σχέση (2.16) για το σ_{XY} και (2.17) για το ρ). Ο συντελεστής συσχέτισης $\rho = \mathrm{Corr}(X, Y)$ αποτελεί κανονικοποίηση της συνδιασποράς σ_{XY} και 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. Ο συντελεστής συσχέτισης λέγεται και συντελεστής συσχέτισης Pearson για να το διαχωρίσουμε από άλλους συντελεστές συσχέτισης, όπως του Spearman και του Kendall.

Όταν $\rho = \pm 1$ η σχέση είναι αιτιοκρατική και 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 ισχυρή τη συσχέτιση όταν $|\rho| > 0.9$) ενώ όταν είναι κοντά στο μηδέν οι τ.μ. είναι πρακτικά ασυσχέτιστες.

Όπως φαίνεται από τον ορισμό στη σχέση (2.17), ο συντελεστής συσχέτισης ρ δεν εξαρτάται από τη μονάδα μέτρησης των X και Y και είναι συμμετρικός ως προς τις X και Y.

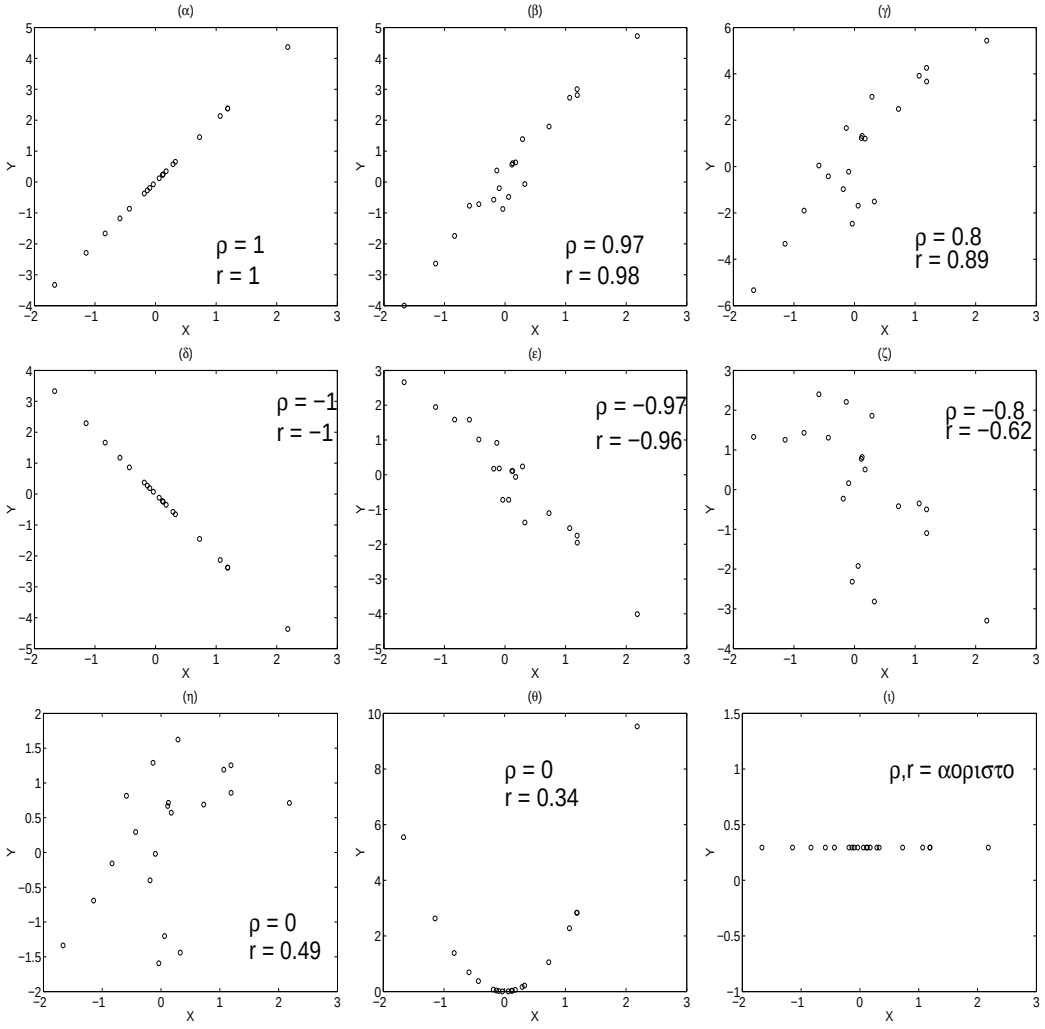
5.1.1 Δειγματικός συντελεστής συσχέτισης

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 πάρουμε μια πρώτη εντύπωση για τη συσχέτιση τους 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 5.1 παρουσιάζονται τυπικά διαγράμματα διασποράς για ισχυρές και ασθενείς συσχετίσεις δύο τ.μ. X και Y. Στα Σχήματα 5.1α και 5.1δ



Σχήμα 5.1: 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 ρ και η δειγματική r. Στο (ι) ο συντελεστής συσχέτισης δεν ορίζεται.

η σχέση είναι τέλεια ($\rho = 1$ και $\rho = -1$ αντίστοιχα), στα Σχήματα 5.1β και 5.1ε είναι ισχυρή (θετική με $\rho = 0.97$ και αρνητική με $\rho = -0.97$ αντίστοιχα)

και στα Σχήματα 5.1γ και 5.1ζ είναι λιγότερο ισχυρή (θετική με $\rho = 0.8$ και αρνητική με $\rho = -0.8$ αντίστοιχα). Στο Σχήμα 5.1η είναι $\rho = 0$ γιατί οι τ.μ. X και Y είναι ανεξάρτητες ενώ στο Σχήμα 5.1θ είναι πάλι $\rho = 0$ αλλά οι X και Y δεν είναι ανεξάρτητες αλλά συσχετίζονται μόνο μη-γραμμικά. Τέλος για το Σχήμα 5.1ι ο συντελεστής συσχέτισης δεν ορίζεται γιατί η Y είναι σταθερή ($\sigma_Y = 0$ στον ορισμό του ρ στην (2.17)).

Η **σημειακή εκτίμηση** (point estimation) του συντελεστή συσχέτισης ρ του πληθυσμού από το δείγμα των n ζευγαρωτών παρατηρήσεων των X και Y γίνεται με την αντικατάσταση στη σχέση (2.17) της συνδιασποράς σ_{XY} και των τυπικών αποκλίσεων σ_X και σ_Y από τις αντίστοιχες εκτιμήσεις από το δείγμα

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

Ο αμερόληπτος εκτιμητής s_{XY} του σ_{XY} δίνεται όπως και ο αμερόληπτος εκτιμητής της συνδιασποράς σ_X^2 (δες σχέση 3.3) ως

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right), \tag{5.1}$$

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

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. \tag{5.2}$$

Είναι προφανές πως η παραπάνω σχέση για το r δεν αλλάζει αν θεωρήσουμε τους μεροληπτικούς εκτιμητές των σ_{XY}, σ_X και σ_Y. Το r λέγεται **δειγματικός συντελεστής συσχέτισης (του Pearson** (sample Pearson correlation coefficient).

Καλύτερη φυσική ερμηνεία της συσχέτισης δύο τ.μ. επιτυγχάνεται με το r^2 που λέγεται **συντελεστής προσδιορισμού** (coefficient of determination) (εκφράζεται συνήθως σε ποσοστό, δηλαδή $100r^2$). Ο συντελεστής προσδιορισμού δίνει το ποσοστό μεταβλητότητας των τιμών της Y που υπολογίζεται από τη X (και αντίστροφα) και είναι ένας χρήσιμος τρόπος να συνοψίσουμε τη σχέση δύο τ.μ..

Στο Σχήμα 5.1 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 ο εκτιμητής r όπως δίνεται στη σχέση (5.2) είναι μια τ.μ. που εξαρτάται από τις τιμές και το πλήθος των ζευγών των παρατηρήσεων.

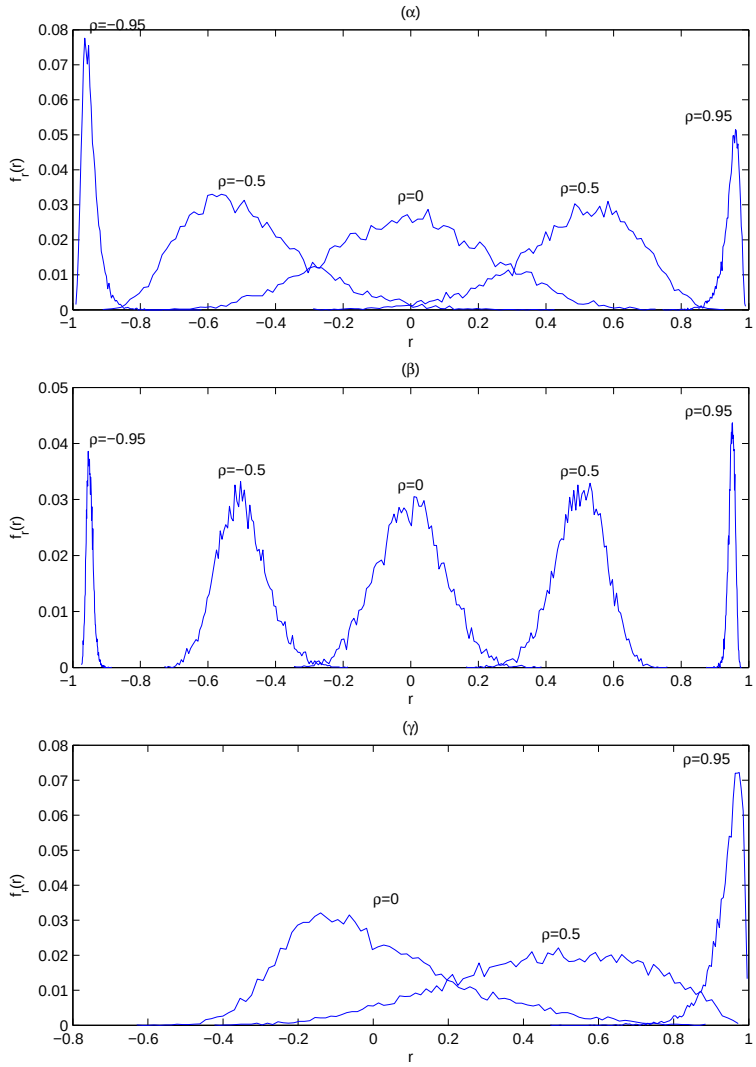
5.1.2 Κατανομή του εκτιμητή r

Η κατανομή του εκτιμητή r δε μπορεί εύκολα να περιγραφτεί καθώς το r δεν έχει την ίδια κατανομή για κάθε τιμή του πραγματικού συντελεστή ρ. Η κατανομή του r δεν είναι κάποια γνωστή, όπως κανονική ή άλλη κατανομή που φθίνει εκθετικά ή με νόμο δύναμης, αφού φράζεται από το -1 και το 1. Για $\rho = 0$, η πυκνότητα των τιμών του r κατανέμεται συμμετρικά γύρω από το 0, και η μέση τιμή του είναι 0. Καθώς όμως το ρ αποκλίνει προς το -1 ή το 1, η κατανομή του r γίνεται θετικά ή αρνητικά ασύμμετρη, αντίστοιχα. Όταν το ρ πλησιάζει το -1 ή το 1, η διασπορά του r μικραίνει για να γίνει 0 όταν το ρ ισούται με -1 ή 1. Η μορφή της κατανομής του r για κάθε τιμή του ρ εξαρτάται από το μέγεθος του δείγματος n αλλά και από την κατανομή των τ.μ. X και Y.

Η κατανομή του r δίνεται αναλυτικά όταν το ζεύγος τ.μ. (X, Y) ακολουθεί διμεταβλητή κανονική κατανομή και υπάρχουν επίσης προσεγγιστικές αναλυτικές εκφράσεις όταν η διμεταβλητή κατανομή δεν είναι κανονική, αλλά όλες αυτές οι αναλυτικές εκφράσεις είναι αρκετά σύνθετες και δεν παρουσιάζονται εδώ. Στο Σχήμα 5.2 δίνεται η εμπειρική κατανομή του r για διαφορετικές τιμές του ρ, που σχηματίστηκε ως εξής. Για σταθερές μέσες τιμές και τυπικές αποκλίσεις των X και Y και για κάθε τιμή του συντελεστή συσχέτισης ρ, πραγματοποιούνται M δείγματα μεγέθους n και σε καθένα από αυτά υπολογίζεται ο δειγματικός συντελεστής συσχέτισης r. Από τις M τιμές του r σχηματίζεται το κανονικοποιημένο ιστόγραμμα (για σχετικές συχνότητες) που προσομοιώνει τη συνάρτηση πυκνότητας πιθανότητας του r, $f_r(r)$. Στο Σχήμα 5.2α οι παράμετροι είναι $\mu_X = 0$, $\mu_Y = 0$, $\sigma_X = 1$, $\sigma_Y = 1$, $M = 10000$ και $n = 20$. Παρατηρούμε πως η μορφή της εμπειρικής $f_r(r)$ διαφέρει για τις διαφορετικές τιμές του ρ και εμφανίζει συμμετρία μόνο για $\rho = 0$. Στο Σχήμα 5.2β, όπου το μέγεθος των δειγμάτων πενταπλασιάστηκε ($n = 100$), η $f_r(r)$ γίνεται πιο συμμετρική και με μικρότερη διασπορά (όπως αναμένεται για μεγαλύτερο δείγμα). Σημειώνεται ότι η σύγκλιση της $f_r(r)$ στην κανονική κατανομή που περιμένουμε σύμφωνα με το Κεντρικό Οριακό Θεώρημα είναι πιο αργή για μεγαλύτερες τιμές του $|\rho|$. Για παράδειγμα για τιμές του ρ κοντά στά όρια -1 και 1, η λοξότητα 'διορθώνεται' όταν η κατανομή γίνεται πολύ στενή, δηλαδή για μεγάλα δείγματα.

Η μορφή της $f_r(r)$ αλλάζει για το ίδιο ρ όταν το δείγμα δεν προέρχεται από διμεταβλητή κανονική κατανομή. Υποθέτοντας τα τετράγωνα των κανονικών τ.μ. X και Y, $X' = X^2$ και $Y' = Y^2$, μπορεί να δειχθεί ότι $\rho' = \mathrm{Corr}(X', Y') = \rho^2$. Σε αυτήν την περίπτωση η μορφή της $f_r(r)$ για την εκτίμηση του ρ' διαφέρει, όπως προκύπτει από τη σύγκριση των Σχημάτων 5.2α και 5.2γ.

Σε πολλά προβλήματα της στατιστικής και ανάλυσης δεδομένων προσπα-



Σχήμα 5.2: (α) Συνάρτηση πυκνότητας πιθανότητας του r για διαφορετικές τιμές του ρ, εκτιμούμενη από το ιστόγραμμα των τιμών του r υπολογισμένο σε 10000 ζευγαρωτά δείγματα μεγέθους $n = 20$ από κανονική κατανομή με μέσες τιμές 0 και διασπορά 1 και για τις δύο τ.μ. X και Y. (β) Όπως στο (α) αλλά για $n = 100$. (γ) Όπως το (α) αλλά μόνο για θετικό συντελεστή ρ των τετραγώνων των X και Y. Σε αυτήν την περίπτωση αυτός ο συντελεστής συσχέτισης είναι το τετράγωνο του συντελεστή συσχέτισης των X και Y.

θούμε να μετασχηματίσουμε την παρατηρούμενη μεταβλητή ώστε να ακολουθεί κάποια γνωστή κατανομή, συνήθως κανονική. Εδώ η μεταβλητή είναι ο

εκτιμητής r. Προτάθηκε από τον Fisher ο μετασχηματισμός

$$z = \tanh^{-1}(r) = 0.5 \ln \frac{1+r}{1-r}, \tag{5.3}$$

ώστε όταν το δείγμα είναι μεγάλο και από διμεταβλητή κανονική κατανομή το z να τείνει προς την κανονική κατανομή με μέση τιμή $\mu_z \equiv \mathrm{E}(z) = \tanh^{-1}(\rho)$ και διασπορά $\sigma_z^2 \equiv \mathrm{Var}(z) = 1/(n-3)$, δηλαδή η διασπορά είναι ανεξάρτητη του ρ. Μπορούμε λοιπόν να υπολογίσουμε διάστημα εμπιστοσύνης και να κάνουμε έλεγχο υπόθεσης χρησιμοποιώντας την προσεγγιστικά κανονική κατανομή του z.

5.1.3 Διάστημα εμπιστοσύνης για το συντελεστή συσχέτισης

Για τον υπολογισμό παραμετρικού διαστήματος εμπιστοσύνης για το ρ, το πρώτο βήμα είναι ο μετασχηματισμός του εκτιμητή r στο z από τη σχέση (5.3). Σύμφωνα και με το διάστημα εμπιστοσύνης για τη μέση τιμή μιας τ.μ., υποθέτοντας ότι το z ακολουθεί προσεγγιστικά κανονική κατανομή, το $(1-a)\%$ διάστημα εμπιστοσύνης για το z δίνεται ως

$$z \pm z_{1-a/2} \sqrt{1/(n-3)}.$$

Το δεύτερο βήμα λοιπόν είναι να υπολογίσουμε από την παραπάνω σχέση αυτό το διάστημα και ας το συμβολίσουμε $[\zeta_l, \zeta_u]$. Τέλος, στο τρίτο βήμα παίρνουμε τον αντίστροφο μετασχηματισμό για τα άκρα του διαστήματος ζ_l και ζ_u, που δίνονται ως

$$r_l = \tanh(\zeta_l) = \frac{\exp(2\zeta_l) - 1}{\exp(2\zeta_l) + 1}, \quad r_u = \frac{\exp(2\zeta_u) - 1}{\exp(2\zeta_u) + 1} \tag{5.4}$$

και ορίζουν το $(1-a)\%$ διαστήματος εμπιστοσύνης για το ρ.

5.1.4 Έλεγχος μηδενικής συσχέτισης

Σε πολλά προβλήματα μας ενδιαφέρει να ελέγξουμε αν δύο τ.μ. συσχετίζονται. Ένας τρόπος να το επιτύχουμε είναι από τον υπολογισμό του διαστήματος εμπιστοσύνης του ρ. Αν το $(1-a)\%$ διαστήματος εμπιστοσύνης του ρ, $[r_l, r_u]$ (δες σχέση (5.4)), δεν περιέχει το 0, τότε μπορούμε να δεχθούμε στο ίδιο επίπεδο εμπιστοσύνης ότι οι δύο τ.μ. συσχετίζονται.

Αν θέλουμε να κάνουμε έλεγχο υπόθεσης για το $\rho = 0$ τότε μπορούμε να χρησιμοποιήσουμε την κατανομή του r κάτω από τη μηδενική υπόθεση H_0:

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 5.1: Δεδομένα αντίστασης x_i και χρόνου αποτυχίας y_i για 20 αντιστάτες.

$\rho = 0$. Για $\rho = 0$, ισχύει

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}, \tag{5.5}$$

δηλαδή η τ.μ. t που προκύπτει από τον παραπάνω μετασχηματισμό του r ακολουθεί κατανομή Student με $n-2$ βαθμούς ελευθερίας. Άρα η απόφαση του ελέγχου γίνεται με βάση το στατιστικό t της σχέσης (5.5) και η p-τιμή του ελέγχου μπορεί να υπολογισθεί από την αθροιστική κατανομή Student για την τιμή του στατιστικού από το δείγμα.

Παράδειγμα 5.1. Θέλουμε να διερευνήσουμε τη συσχέτιση της αντίστασης και του χρόνου αποτυχίας κάποιου υπερφορτωμένου αντιστάτη. Για αυτό πήραμε μετρήσεις αντίστασης (σε Ωμ, ohm) και χρόνου αποτυχίας (σε λεπτά, min) από δείγμα 20 αντιστατών, οι οποίες παρουσιάζονται στον Πίνακα 5.1.

Για να βρούμε τον συντελεστή συσχέτισης r υπολογίζουμε πρώτα τα πα-

ρακάτω

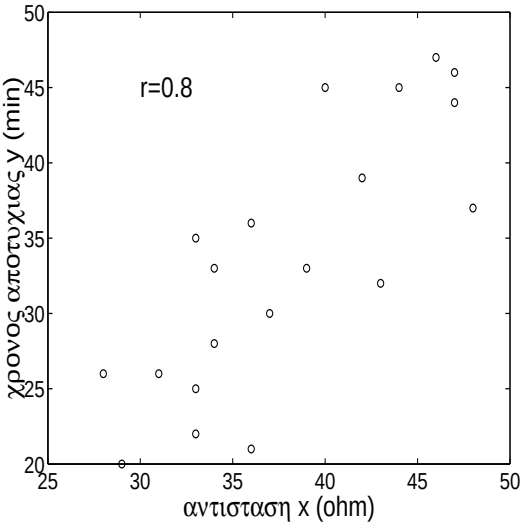
$$\bar{x} = 38 \qquad \bar{y} = 33.5$$

$$\sum_{i=1}^{20} x_i^2 = 29634 \qquad \sum_{i=1}^{20} y_i^2 = 23910 \qquad \sum_{i=1}^{20} x_i y_i = 26305.$$

Από τη σχέση (5.2) 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 και ο χρόνος αποτυχίας αντιστάτη έχουν γραμμική θετική συσχέτιση αλλά όχι πολύ ισχυρή. Αυτό φαίνεται και από το διάγραμμα διασποράς στο Σχήμα 5.3. Η μεταβλητότητα της μιας τ.μ. (αντίσταση ή χρόνος αποτυχίας) μπορεί να



Σχήμα 5.3: Διάγραμμα διασποράς για το δείγμα παρατηρήσεων αντίστασης και χρόνου αποτυχίας 20 αντιστατών του Πίνακα 5.1.

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.

Για τον υπολογισμό του 95% διαστήματος εμπιστοσύνης του συντελεστή συσχέτισης ρ υπολογίζουμε πρώτα το μετασχηματισμό Fisher του r από τη

σχέση (5.3), $z = 1.110$. Τα άκρα του 95% διαστήματος εμπιστοσύνης του z είναι $[0.634, 1.585]$ και ο αντίστροφος μετασχηματισμός στα άκρα αυτού του διαστήματος δίνει το 95% διάστημα εμπιστοσύνης του ρ ως $[0.561, 0.919]$. Τα όρια αυτά δηλώνουν ότι για τόσο μικρό δείγμα δε μπορούμε να εκτιμήσουμε με ακρίβεια το συντελεστή συσχέτισης.

Το κάτω άκρο του 95% διαστήματος εμπιστοσύνης του ρ είναι 0.561 και είναι πολύ μεγαλύτερο του 0, άρα μπορούμε να ισχυριστούμε με μεγάλη σιγουριά πως η αντίσταση και ο χρόνος αποτυχίας συσχετίζονται. Ο έλεγχος της υπόθεσης $\rho = 0$ δίνει το Student στατιστικό $t = 5.736$ (δες σχέση (5.5)) και $p = 0.0000194$, δηλαδή είναι εντελώς απίθανο η αντίσταση και ο χρόνος αποτυχίας να είναι ανεξάρτητες τ.μ. με βάση το δείγμα των 20 ζευγαρωτών παρατηρήσεων.

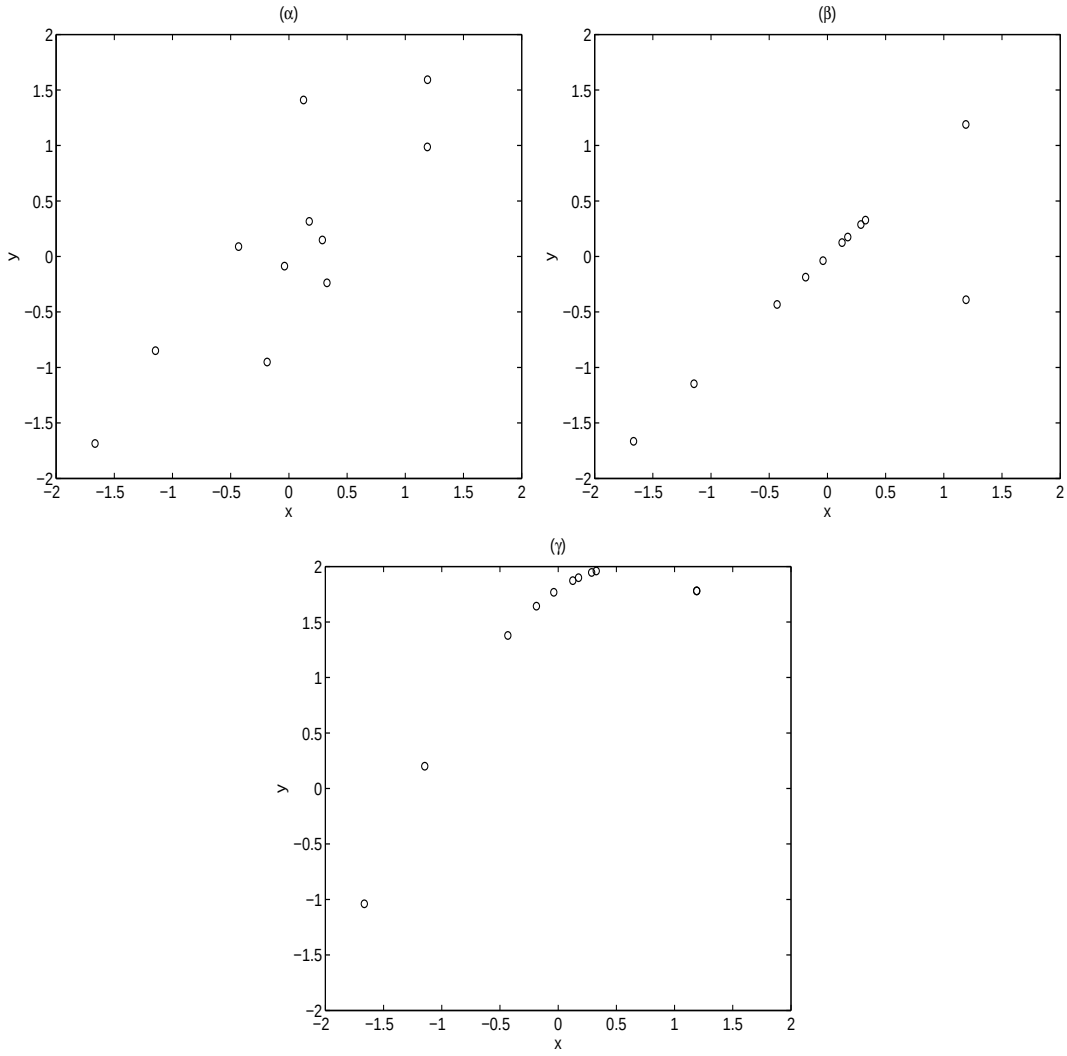
Πρέπει τέλος να σημειωθεί ότι η εκτίμηση r του συντελεστή συσχέτισης μπορεί να αλλάξει σημαντικά με την πρόσθεση ή αφαίρεση λίγων παρατηρήσεων γιατί το μέγεθος του δείγματος είναι μικρό.

5.1.5 Συσχέτιση και γραμμικότητα

Ο συντελεστής συσχέτισης Pearson ρ καθώς και ο εκτιμητής του r αναφέρονται στο μέγεθος της γραμμικής συσχέτισης μεταξύ δύο τ.μ. X και Y. Για ένα δείγμα του ζεύγους (X, Y), η τιμή του r δεν αποδίδει πάντα σωστά τη συσχέτιση, καθώς η συσχέτιση τους μπορεί να μήν είναι απλά γραμμική.

Στο Σχήμα 5.4 δίνονται τα διαγράμματα διασποράς για τρία δείγματα του ζεύγους (X, Y). Ο δειγματικός συντελεστής συσχέτισης για τα τρία δείγματα είναι ο ίδιος, $r = 0.84$. Για το πρώτο δείγμα στο Σχήμα 5.4α, το ζεύγος (X, Y) είναι από διμεταβλητή κανονική κατανομή και άρα το r αποδίδει σωστά το μέγεθος της συσχέτισης τους. Το δεύτερο και τρίτο δείγμα όμως δεν αντιστοιχεί σε γραμμικά συσχετισμένα X και Y. Είναι φανερό πως για το δεύτερο δείγμα στο Σχήμα 5.4β η σχέση των X και Y είναι απόλυτα γραμμική ($\rho = 1$) για όλα τα ζευγάρια εκτός από ένα απόμακρο σημείο (outlier), το οποίο μειώνει την εκτίμηση στο $r = 0.84$. Στο τρίτο δείγμα στο Σχήμα 5.4γ η συσχέτιση των X και Y είναι απόλυτη αλλά μη-γραμμική με αποτέλεσμα η συσχέτιση τους πάλι να υποεκτιμάται.

Από τα παραπάνω παραδείγματα είναι φανερό πως ο συντελεστής συσχέτισης Pearson δεν είναι πάντα κατάλληλο μέτρο συσχέτισης. Για την αντιμετώπιση απόμακρων σημείων άλλα μέτρα είναι πιο κατάλληλα, όπως ο εκτιμητής του συντελεστή συσχέτισης Spearman και Kendall. Όταν η συσχέτιση μπορεί να είναι μη-γραμμική άλλα μέτρα πρέπει να αναζητηθούν. Ένα τέτοιο μέτρο είναι η **αμοιβαία πληροφορία** (mutual information) που μετράει την πληροφορία που μπορούμε να έχουμε για τη μια τ.μ. γνωρίζοντας την άλλη.



Σχήμα 5.4: Διαγράμματα διασποράς δειγμάτων που δίνουν όλα τον ίδιο δειγματικό συντελεστή συσχέτισης $r = 0.84$. (α) Τα X και Y είναι από διμεταβλητή κανονική κατανομή. (β) $Y = X$ για όλα εκτός από ένα ζευγάρι παρατηρήσεων του δείγματος. (γ) $Y = 2 - 0.6(X - 0.585)^2$.

5.2 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 στη συνέχεια σχεδιάζουμε την εξάρτηση μιας τ.μ. Y, που την ονομάζουμε **εξαρτημένη μεταβλητή** (dependent variable), από άλλες τυχαίες μεταβλητές που τις ονομάζουμε **ανεξάρτητες μεταβλητές** (independent

variables). Γενικά θεωρούμε πως τις τιμές των ανεξάρτητων μεταβλητών τις ορίζουμε εμείς και δεν εμπεριέχουν σφάλματα, δεν είναι δηλαδή τυχαίες μεταβλητές. Σε πολλά πρακτικά προβλήματα όμως αυτό είναι περισσότερο μια παραδοχή για να εφαρμόσουμε τη μέθοδο των ελαχίστων τετραγώνων για να εκτιμήσουμε το μοντέλο παλινδρόμησης που δίνει τη μαθηματική έκφραση της εξάρτησης της εξαρτημένης από τις ανεξάρτητες μεταβλητές.

Στην αρχή θα θεωρήσουμε την εξάρτηση της Y από μία μόνο ανεξάρτητη μεταβλητή X και η περίπτωση αυτή αναφέρεται ως *απλή παλινδρόμηση*. Ενώ η συσχέτιση είναι συμμετρική ως προς τα X και Y, στην απλή παλινδρόμηση η εξαρτημένη μεταβλητή Y 'καθοδηγείται' από την ανεξάρτητη μεταβλητή X. Για αυτό και στην ανάλυση που κάνουμε παίζει ρόλο ποιόν από τους δύο παράγοντες που μετράμε ορίζουμε ως ανεξάρτητη μεταβλητή και ποιόν ως 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.

Στο πρώτο μέρος της μελέτης θα θεωρήσουμε πως η εξάρτηση είναι γραμμική, δηλαδή έχουμε την περίπτωση της *απλής γραμμικής παλινδρόμησης*.

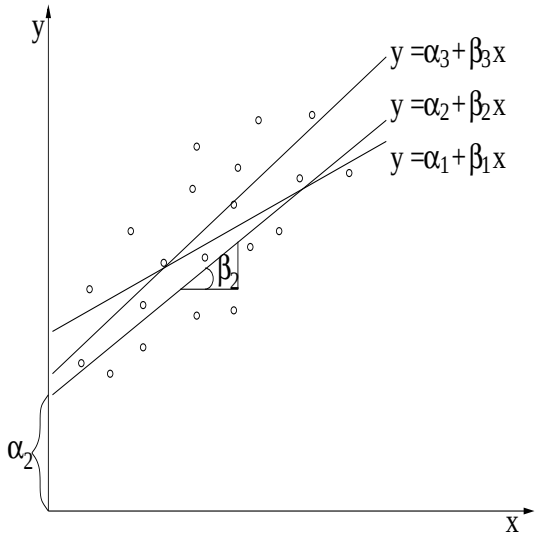
5.2.1 Το πρόβλημα της απλής γραμμικής παλινδρόμησης

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

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x \tag{5.6}$$

και η σχέση αυτή λέγεται **απλή γραμμική παλινδρόμηση της** Y **στη** X (linear regression). Το πρόβλημα της παλινδρόμησης είναι η εύρεση των παραμέτρων β_0 και β_1 που εκφράζουν καλύτερα τη γραμμική εξάρτηση της Y από τη X. Κάθε ζεύγος τιμών (β_0, β_1) καθορίζει μια διαφορετική γραμμική σχέση που εκφράζεται γεωμετρικά από ευθεία γραμμή. Η διαφοράς ύψους β_0 είναι η τιμή του y για $x = 0$ και λέγεται **διαφορά ύψους** (intercept) κι ο συντελεστής του x, β_1, 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 5.5.



Σχήμα 5.5: Ευθείες απλής γραμμικής παλινδρόμησης

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 ως 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

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{5.7}$$

όπου ϵ_i είναι και αυτή τ.μ., λέγεται **σφάλμα παλινδρόμησης** (regression error) και 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:

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

- Η σχέση (5.6) 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

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

δηλαδή η διασπορά της εξαρτημένης μεταβλητής Y είναι η ίδια για κάθε τιμή της X και μάλιστα είναι $\sigma_{Y|X}^2 = \sigma_\epsilon^2$. Η τελευταία σχέση προκύπτει από τη σχέση (5.7), αφού οι παράμετροι β_0 και β_1 είναι σταθερές και το x_i γνωστό. Η ιδιότητα αυτή λέγεται *ομοσκεδαστικότητα* 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 παραμετρικά διαστήματα εμπιστοσύνης για τις παραμέτρους ή να κάνουμε παραμετρικούς στατιστικούς ελέγχους 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

$$Y|X = x \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma_\epsilon^2).$$

Αν η κατανομή της Y δεν είναι κανονική, μπορούμε να χρησιμοποιήσουμε κάποιο μετασχηματισμό που την κάνει κανονική και για αυτό συχνά χρησιμοποιείται ο λογάριθμος.

5.2.2 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 β_0,

2. της κλίσης της ευθείας παλινδρόμησης β_1,

3. της διασποράς σφάλματος της παλινδρόμησης σ_ϵ^2.

Τα β_0 και β_1 προσδιορίζουν την ευθεία παλινδρόμησης και άρα καθορίζουν τη γραμμική σχέση εξάρτησης της τ.μ. Y από τη μεταβλητή X. Η παράμετρος σ_ϵ^2 προσδιορίζει το βαθμό μεταβλητότητας γύρω από την ευθεία παλινδρόμησης και εκφράζει την αβεβαιότητα της γραμμικής σχέσης.

Εκτίμηση των παραμέτρων της ευθείας παλινδρόμησης

Η εκτίμηση των παραμέτρων β_0 και β_1 γίνεται με τη μέθοδο των **ελαχίστων τετραγώνων** (method of least squares). Η μέθοδος λέγεται έτσι γιατί βρίσκει την ευθεία παλινδρόμησης με παραμέτρους b_0 και b_1 έτσι ώστε το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων από την ευθεία να είναι το ελάχιστο. Οι εκτιμήσεις των β_0 και β_1 δίνονται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \epsilon_i^2 \qquad \text{ή} \qquad \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{5.8}$$

Για να λύσουμε αυτό το πρόβλημα θέτουμε τις μερικές παραγώγους ως προς τα β_0 και β_1 ίσες με το μηδέν και καταλήγουμε στο σύστημα δύο εξισώσεων με δύο αγνώστους

$$\left.\begin{array}{l} \frac{\partial \sum (y_i-\beta_0-\beta_1 x_i)^2}{\partial \beta_0} = 0 \\ \frac{\partial \sum (y_i-\beta_0-\beta_1 x_i)^2}{\partial \beta_1} = 0 \end{array}\right\} \begin{array}{l} \sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i = \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 \end{array}$$

από το οποίο με αντικατάσταση του β_0 παίρνουμε την εκτίμηση για την κλίση

$$\hat{\beta}_1 \equiv b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}. \tag{5.9}$$

Το άθροισμα που συμβολίζεται στην παραπάνω σχέση ως S_{xy} διαιρώντας με $n-1$ δίνει τη δειγματική συνδιασπορά s_{XY} (δες σχέση (5.1)) και αντίστοιχα το S_{xx} διαιρώντας με $n-1$ δίνει τη δειγματική διασπορά της X s_{XY} (δες σχέση (3.3)). Με αντικατάσταση του b_1 στην πρώτη εξίσωση του παραπάνω συστήματος παίρνουμε την εκτίμηση του σταθερού όρου ως

$$\hat{\beta}_0 \equiv b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} \tag{5.10}$$

και άρα οι εκτιμήσεις των β_0 και β_1 δίνονται ως

$$b_1 = \frac{s_{XY}}{s_X^2}, \qquad b_0 = \bar{y} - b_1 \bar{x}. \tag{5.11}$$

Τα b_0 και b_1 ορίζουν την ευθεία

$$\hat{y} = b_0 + b_1 x,$$

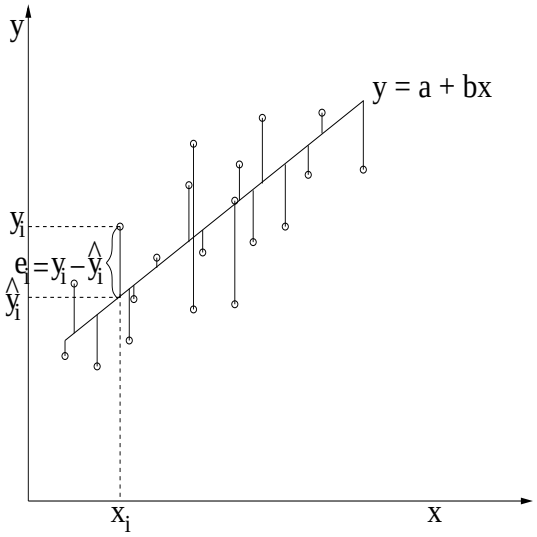
που λέγεται και **ευθεία ελαχίστων τετραγώνων**.

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

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$$

είναι η κατακόρυφη απόσταση της πραγματικής τιμής από την ευθεία ελαχίστων τετραγώνων και λέγεται σφάλμα ελαχίστων τετραγώνων ή απλά **υπόλοιπο** (residual). Στο Σχήμα 5.6 απεικονίζονται τα υπόλοιπα της παλινδρόμησης.



Σχήμα 5.6: Ευθεία ελαχίστων τετραγώνων και υπόλοιπα

Το υπόλοιπο e_i είναι η εκτίμηση του σφάλματος παλινδρόμησης ϵ_i αντικαθιστώντας απλά τις παραμέτρους παλινδρόμησης με τις εκτιμήσεις ελαχίστων τετραγώνων στον ορισμό του σφάλματος $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$. Άρα η εκτίμηση της διασποράς σ_ϵ^2 του σφάλματος (που είναι και η δεσμευμένη διασπορά της Y ως προς X) δίνεται από τη δειγματική διασπορά s_ϵ^2 των υπολοίπων e_i

$$s_\epsilon^2 \equiv \hat{\sigma}_\epsilon^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{5.12}$$

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_ϵ^2 μπορεί να εκφραστεί ως προς τις δειγματικές διασπορές των X και Y και της συνδιασποράς τους, αν αντικαταστήσουμε τις εκφράσεις των b_0 και b_1 από την (5.11) στην παραπάνω σχέση (όπου θέτουμε $\hat{y}_i = b_0 + b_1 x_i$)

$$s_\epsilon^2 = \frac{n-1}{n-2}\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right) = \frac{n-1}{n-2}(s_Y^2 - b_1^2 s_X^2) \tag{5.13}$$

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

   $$b_0 + b_1\bar{x} = \bar{y} - b_1\bar{x} + b_1\bar{x} = \bar{y}.$$

   Άρα η ευθεία ελαχίστων τετραγώνων μπορεί επίσης να οριστεί ως

   $$y_i - \bar{y} = b_1(x_i - \bar{x}).$$

2. Η σημειακή εκτίμηση των β_0 και β_1 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 b_0 και b_1 είναι οι εκτιμήτριες μεγίστης πιθανοφάνειας (και 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.

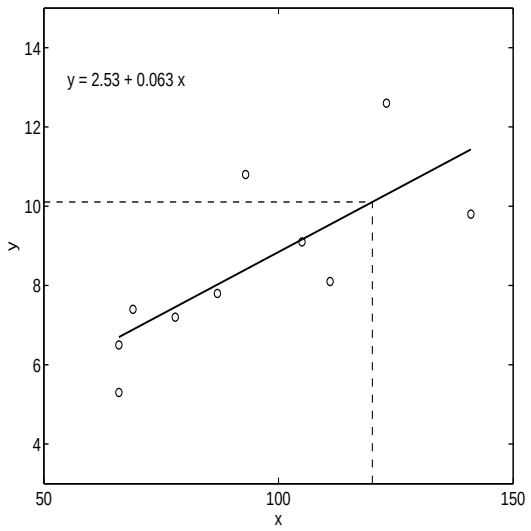
4. Για κάθε τιμή x της X μπορούμε να *προβλέψουμε* την αντίστοιχη τιμή y της Y από την ευθεία ελαχίστων τετραγώνων, $\hat{y} = b_0 + b_1 x$. Εδώ πρέπει να προσέξουμε ότι η τιμή x πρέπει να ανήκει στο εύρος τιμών της X που έχουμε από το δείγμα. Για τιμές έξω από αυτό το διάστημα η πρόβλεψη δεν είναι αξιόπιστη.

Παράδειγμα 5.2. Θέλουμε να μελετήσουμε σε ένα ολοκληρωμένο κύκλωμα την εξάρτηση της απολαβής ρεύματος κρυσταλλολυχνίας (τρανζίστορ) από την αντίσταση του στρώματος της κρυσταλλολυχνίας. Στον Πίνακα 5.2 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 5.2: 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 5.7 επιβεβαιώνει αυτήν την υπόθεση. Για να



Σχήμα 5.7: Διάγραμμα διασποράς για τα δεδομένα του Πίνακα 5.2 και ευθεία ελαχίστων τετραγώνων.

εκτιμήσουμε τις παραμέτρους b_0 και b_1 της ευθείας ελαχίστων τετραγώνων

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 (5.1) 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_1 και b_0 είναι

$$\begin{aligned} b_1 &= \frac{41.81}{662.1} = 0.063 \\ b_0 &= 8.46 - 0.063 \cdot 93.9 = 2.53. \end{aligned}$$

Από τη σχέση (5.13) υπολογίζουμε την εκτίμηση διασποράς των σφαλμάτων παλινδρόμησης

$$s_\epsilon^2 = \frac{9}{8}(4.66 - 0.063^2 \cdot 41.81) = 2.271.$$

Τα αποτελέσματα αυτά ερμηνεύονται ως εξής:

1. b_1: Για αύξηση της αντίστασης στρώματος κατά μία μονάδα μέτρησης (1 ohm/cm) η απολαβή του ρεύματος της κρυσταλλολυχνίας αυξάνεται κατά 0.063.

2. b_0: 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_ϵ^2: Η εκτίμηση της διασποράς γύρω από την ευθεία παλινδρόμησης για κάθε τιμή της X (στο διάστημα τιμών του πειράματος) είναι 2.271, ή αλλιώς το τυπικό σφάλμα της εκτίμησης της παλινδρόμησης είναι 1.507 μονάδες, που είναι σχετικά 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 5.7 απεικονίζεται η

πρόβλεψη της απολαβής ρεύματος για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$ και είναι

$$\hat{y} = 2.53 + 0.063 \cdot 120 = 10.11.$$

Στο παραπάνω παράδειγμα η πρόβλεψη της απολαβής ρεύματος μπορεί να διαφέρει σημαντικά αν αλλάξουν οι σημειακές εκτιμήσεις των παραμέτρων παλινδρόμησης. Στη συνέχεια θα μελετήσουμε την αβεβαιότητα στην εκτίμηση των παραμέτρων παλινδρόμησης και της πρόβλεψης της εξαρτημένης μεταβλητής.

5.2.3 Σχέση του συντελεστή συσχέτισης και παλινδρόμησης

Η παλινδρόμηση ορίζεται θεωρώντας την ανεξάρτητη μεταβλητή X ελεγχόμενη 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 b_1 (της εκτίμησης του συντελεστή της παλινδρόμησης β_1 με τη μέθοδο των ελαχίστων τετραγώνων) δίνεται ως εξής (συνδυάζοντας τις σχέσεις $r = \frac{s_{XY}}{s_X s_Y}$ και $b_1 = \frac{s_{XY}}{s_X^2}$)

$$r = b_1 \frac{s_X}{s_Y} \quad \text{ή} \quad b_1 = r \frac{s_Y}{s_X}. \tag{5.14}$$

Και τα δύο μεγέθη, r και b_1, εκφράζουν ποσοτικά τη γραμμική συσχέτιση των μεταβλητών X και Y, αλλά το b_1 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_1 περιγράφεται ως εξής:

- Αν η συσχέτιση είναι θετική ($r > 0$) τότε η κλίση της ευθείας παλινδρόμησης b_1 είναι επίσης θετική.

- Αν η συσχέτιση είναι αρνητική ($r < 0$) τότε η κλίση της ευθείας παλινδρόμησης b_1 είναι επίσης αρνητική.

- Αν οι μεταβλητές X και Y δε συσχετίζονται ($r = 0$) τότε η ευθεία παλινδρόμησης είναι οριζόντια ($b_1 = 0$).

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

$$s_\epsilon^2 = \frac{n-1}{n-2} s_Y^2 (1 - r^2) \quad \text{ή} \quad r^2 = 1 - \frac{n-2}{n-1} \frac{s_\epsilon^2}{s_Y^2}. \tag{5.15}$$

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 5.3. Στο παραπάνω παράδειγμα 5.2, ο συντελεστής συσχέτισης της απολαβής 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 και από τις σχέσεις (5.14) ή (5.15). Ο συντελεστής συσχέτισης δηλώνει την ασθενή θετική συσχέτιση της απολαβής ρεύματος και της αντίστασης στρώματος, που δε μπορούμε όμως να συμπεράνουμε από την τιμή του συντελεστή παλινδρόμησης $b_1 = 0.063$ ή την τιμή της διασποράς των σφαλμάτων $s_\epsilon^2 = 2.249$ γιατί εξαρτιούνται από τις μονάδες μέτρησης των δύο μεταβλητών.

5.2.4 Διάστημα εμπιστοσύνης των παραμέτρων της απλής γραμμικής παλινδρόμησης

Όπως η δειγματική μέση τιμή $\bar{x}$ και η τυπική απόκλιση s μπορούν να διαφέρουν από δείγμα σε δείγμα παρατηρήσεων μιας τ.μ. X, έτσι και η εκτιμώμενη κλίση b_1 και διαφορά ύψους b_0 μπορούν να διαφέρουν επίσης από δείγμα σε δείγμα ζευγαρωτών παρατηρήσεων των (X, Y). Σε αντιστοιχία με την προσέγγιση για τη μέση τιμή και τυπική απόκλιση, για να υπολογίσουμε διαστήματα εμπιστοσύνης (ή να κάνουμε στατιστικό έλεγχο όπως θα δούμε αμέσως μετά) για τις παραμέτρους β_1 και β_0 θα μελετήσουμε την κατανομή των b_1 και b_0 αντίστοιχα. Σημειώνεται ότι μόνο η Y είναι τυχαία μεταβλητή, καθώς θεωρήσαμε πως η X είναι ελεγχόμενη (ανεξάρτητη) μεταβλητή.

Μπορεί να δειχθεί με βάση της σχέση (5.9) πως ο εκτιμητής b_1 της κλίσης δίνεται ως γραμμικός συνδυασμός των $y_1, \ldots, y_n$, που κάθε ένα από αυτά είναι τ.μ. όπως το Y. Επιπλέον το b_1 έχει μέση τιμή

$$\mu_{b_1} \equiv \mathrm{E}(b_1) = \beta_1$$

και διασπορά

$$\sigma_{b_1}^2 \equiv \mathrm{Var}(b_1) = \frac{\sigma_\epsilon^2}{S_{xx}}$$

ή αντίστοιχα τυπική απόκλιση $\sigma_{b_1} = \sigma_\epsilon / \sqrt{S_{xx}}$. Καθώς η διασπορά των σφαλμάτων εκτιμάται από την ευθεία ελαχίστων τετραγώνων ως s_ϵ^2, η εκτίμηση της τυπικής απόκλισης του b_1 είναι

$$s_{b_1} = \frac{s_\epsilon}{\sqrt{S_{xx}}}. \tag{5.16}$$

Θεωρώντας πως η Y ακολουθεί κανονική κατανομή, ο εκτιμητής κλίσης b_1 ακολουθεί επίσης κανονική κατανομή και άρα το $(1-a)\%$ διάστημα εμπιστοσύνης της κλίσης β_1 δίνεται όπως και για τη μέση τιμή ως

$$b_1 \pm t_{n-2,1-a/2} s_{b_1} \quad \text{ή} \quad b_1 \pm t_{n-2,1-a/2} \frac{s_\epsilon}{\sqrt{S_{xx}}}. \tag{5.17}$$

Η κατανομή Student που δίνει την κρίσιμη τιμή έχει $n-2$ βαθμούς ελευθερίας όσους θεωρούμε και στην εκτίμηση του s_ϵ (δες σχέση (5.12)).

Με αντίστοιχη προσέγγιση μπορεί να δειχθεί πως η τυπική απόκλιση της εκτίμησης της διαφοράς ύψους b_0 είναι

$$\sigma_{b_0} = s_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \tag{5.18}$$

και το $(1-a)\%$ διάστημα εμπιστοσύνης του σταθερού όρου β_0 είναι

$$b_0 \pm t_{n-2,1-a/2} s_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}. \tag{5.19}$$

5.2.5 Έλεγχος υπόθεσης για τις παραμέτρους της απλής γραμμικής παλινδρόμησης

Κάτω από την υπόθεση της κανονικότητας μπορούν να γίνουν αντίστοιχοι παραμετρικοί έλεγχοι υπόθεσης για την κλίση και τη διαφορά ύψους.

Για τη μηδενική υπόθεση H_0: $\beta_1 = \beta_1^0$, δηλαδή ότι η κλίση μπορεί να είναι β_1^0, ο παραμετρικός έλεγχος γίνεται με το στατιστικό

$$t = \frac{b_1 - \beta_1^0}{s_b} = \frac{(b_1 - \beta_1^0)\sqrt{S_{xx}}}{s_\epsilon}, \tag{5.20}$$

όπου $t \sim t_{n-2}$, δηλαδή το στατιστικό κάτω από την H_0 ακολουθεί κατανομή Student με $n-2$ βαθμούς ελευθερίας. Ιδιαίτερο ενδιαφέρον έχει η μηδενική υπόθεση H_0: $\beta_1 = 0$, δηλαδή ότι η γραμμή παλινδρόμησης είναι οριζόντια και άρα η τ.μ. Y δεν εξαρτάται από την X.

Αντίστοιχα ο έλεγχος υπόθεσης για τη διαφορά ύψους, H_0: $\beta_0 = \beta_0^0$, δίνεται από το στατιστικό

$$t = \frac{b_0 - \beta_0^0}{s_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \tag{5.21}$$

όπου και πάλι $t \sim t_{n-2}$.

5.2.6 Διαστήματα πρόβλεψης

Έχοντας μελετήσει την κατανομή των παραμέτρων της ευθείας ελαχίστων τετραγώνων b_0 και b_1 μπορούμε να μελετήσουμε την κατανομή της εκτίμησης της μέσης τιμής του Y για κάποιο x, $E(Y|X = x) = \beta_0 + \beta_1 x$, από την ευθεία ελαχίστων τετραγώνων $\hat{y}$ που δίνεται ως $\hat{y} = b_0 + b_1 x$.

Μπορεί να δειχθεί και πάλι πως ο εκτιμητής $\hat{y}$ δίνεται ως γραμμικός συνδυασμός των τ.μ. $y_1, \ldots, y_n$ και των σταθερών $x_1, \ldots, x_n$ και x. Το $\hat{y}$ είναι αμερόληπτος εκτιμητής του $E(Y|X = x)$, δηλαδή

$$\mu_{\hat{y}} \equiv E(\hat{y}) = E(Y|X = x) = \beta_0 + \beta_1 x.$$

Η διασπορά του $\hat{y}$ μπορεί να δειχθεί ότι είναι

$$\sigma^2_{\hat{y}} \equiv \mathrm{Var}(\hat{y}) = \sigma^2_\epsilon \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)$$

Αντικαθιστώντας τη διασπορά των σφαλμάτων από την ευθεία ελαχίστων τετραγώνων ως s^2_ϵ, η εκτίμηση της τυπικής απόκλισης του $\hat{y}$ είναι

$$s_{\hat{y}} = s_\epsilon \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}. \tag{5.22}$$

Όπως και για τις παραμέτρους b_0 και b_1, ο εκτιμητής της ευθείας ελαχίστων τετραγώνων για κάποιο x ακολουθεί επίσης κανονική κατανομή. Άρα το $(1 - a)$**% διάστημα εμπιστοσύνης της μέσης τιμής του** Y **για κάποιο** x είναι

$$\hat{y} \pm t_{n-2,1-a/2} s_{\hat{y}} \quad \text{ή} \quad (b_0 + b_1 x) \pm t_{n-2,1-a/2} s_\epsilon \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}. \tag{5.23}$$

Το διάστημα εμπιστοσύνης για μέσης τιμής του Y για κάποιο x εξαρτάται από την απόσταση του x από το μέσο όρο των τιμών της μεταβλητής Q του δείγματος. Υπάρχει λοιόν μεγαλύτερη βεβαιότητα για το Y όταν το η τιμή x είναι κοντά στο κέντρο των δεδομένων της Q που ήδη γνωρίζουμε (βάσει των οποίων έγινε η εκτίμηση της ευθείας ελαχίστων τετραγώνων).

Το παραπάνω διάστημα εμπιστοσύνης της μέσης τιμής του Y για κάποιο x, $\mathrm{E}(Y|X=x)$, είναι το **διάστημα της μέσης πρόβλεψης**, δηλαδή δίνει σε επίπεδο εμπιστοσύνης $(1-a)\%$ τα όρια της πρόβλεψης για τη μέση (αναμενόμενη) τιμή της Y όταν δίνεται μια συγκεκριμένη τιμή x για τη μεταβλητή X. Συχνά μας ενδιαφέρει να γνωρίζουμε και τα όρια της πρόβλεψης για μια (μελλοντική) τιμή y της Y που θα πάρουμε για κάποια τιμή x της X. Το ζητούμενο διάστημα εμπιστοσύνης δεν αναφέρεται στην παράμετρο της δεσμευμένης μέσης τιμής $\mathrm{E}(Y|X=x)$ αλλά κάποιας τιμής y της Y για το ίδιο x και άρα περιμένουμε τα όρια να είναι μεγαλύτερα. Η διαφορά είναι ίδια με την ακρίβεια της μέσης τιμής κάποιας τ.μ. X και μιας παρατήρησης της X. Πράγματι το $(1-a)$**% διάστημα πρόβλεψης για μια παρατήρηση** y **της** Y **για κάποιο** x είναι

$$\hat{y} \pm t_{n-2,1-a/2}\sqrt{s_\varepsilon^2 + s_{\hat{y}}^2} \quad \text{ή} \quad (b_0+b_1x) \pm t_{n-2,1-a/2}s_\varepsilon\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}. \quad (5.24)$$

Παράδειγμα 5.4. Στο παραπάνω παράδειγμα 5.2, βρήκαμε τις εκτιμήσεις των παραμέτρων της ευθείας παλινδρόμησης $b_1 = 0.063$, $b_0 = 2.53$ και την εκτίμηση της τυπικής απόκλισης των σφαλμάτων παλινδρόμησης $s_\varepsilon = 1.507$. Θα εξετάσουμε στη συνέχεια την ακρίβεια και σημαντικότητα των παραμέτρων της ευθείας παλινδρόμησης και θα δώσουμε διαστήματα πρόβλεψης για την απολαβή ρεύματος όταν η αντίσταση στρώματος είναι $x = 120\,\mathrm{ohm/cm}$.

Η εκτίμηση της τυπικής απόκλιση της κλίσης της ευθείας ελαχίστων τετραγώνων b_1 είναι από τη σχέση (5.16) $s_{b_1} = 0.0195$. Σύμφωνα με την σχέση (5.17) το 95% διάστημα εμπιστοσύνης για την κλίση είναι

$$0.063 \pm 2.306 \cdot 0.0195 \quad \Rightarrow \quad [0.018, 0.108].$$

Το διάστημα περιέχει μόνο θετικές τιμές που δείχνει πως η κλίση β_1 της ευθείας παλινδρόμησης είναι σημαντικά διάφορη του μηδενός, τουλάχιστον σε επίπεδο σημαντικότητας $a = 0.05$.

Ο έλεγχος για τη μηδενική υπόθεση H_0: $\beta_1 = 0$ επιβεβαιώνει τη σημαντικότητα της κλίσης. Το στατιστικό από το δείγμα από τη σχέση (5.20) έχει την τιμή

$$t = \frac{0.063}{0.0195} = 3.235$$

που είναι αρκετά μεγαλύτερη από την κρίσιμη τιμή $t_{0.975,8} = 2.306$ που δίνει το όριο της t για να δεχθούμε την H_0. Η τιμή $t = 3.235$ από το δείγμα αντιστοιχεί στην πιθανότητα (p-τιμή) $p = 0.012$. Η p-τιμή δηλώνει πως δε θα μπορούσαμε να απορρίψουμε την H_0 ότι η κλίση είναι μηδενική σε επίπεδο σημαντικότητας $a = 0.01$, δηλαδή το μοντέλο παλινδρόμησης δεν έχει μεγάλη σημαντικότητα. Αυτό οφείλεται σε μεγάλο βαθμό στο μικρό αριθμό των

παρατηρήσεων σε συνδυασμό με την όχι τόσο ισχυρή εξάρτηση της απολαβής ρεύματος από την αντίσταση στρώματος ($r = 0.753$).

Αντίστοιχα η εκτίμηση της τυπικής απόκλισης της διαφοράς ύψους β_0 από τη σχέση (5.18) είναι $s_{b_0} = 1.894$ και το 95% διάστημα εμπιστοσύνης είναι σύμφωνα με τη σχέση (5.19)

$$2.53 \pm 2.306 \cdot 1.894 \quad \Rightarrow \quad [-1.837, 6.898]$$

Το διάστημα περιέχει το μηδέν και άρα η διαφορά ύψους δεν είναι σημαντική σε επίπεδο σημαντικότητας $a = 0.05$. Αυτό επιβεβαιώνεται από τον έλεγχο της μηδενικής υπόθεσης H_0: $\beta_0 = 0$. Το στατιστικό από το δείγμα από τη σχέση (5.21) είναι

$$t = \frac{2.53}{1.894} = 1.336$$

που είναι μικρότερο της κρίσιμης τιμής $t_{0.975,8} = 2.306$ και δίνει p-τιμή $p = 0.218$. Το αποτέλεσμα αυτό δηλώνει πως η εξάρτηση της απολαβής ρεύματος από την αντίσταση στρώματος μπορεί να περιγραφτεί με μοντέλο γραμμικής παλινδρόμησης χωρίς σημαντική διαφορά ύψους.

Στη συνέχεια ας εξετάσουμε τα όρια πρόβλεψης για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$. Από τη σχέση (5.22) υπολογίζουμε την τυπική απόκλιση της μέσης απολαβής ρεύματος για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$ ως

$$s_{\hat{y}} = 1.507 \sqrt{\frac{1}{10} + \frac{(120 - 93.9)^2}{9 \cdot 662.1}} = 0.698.$$

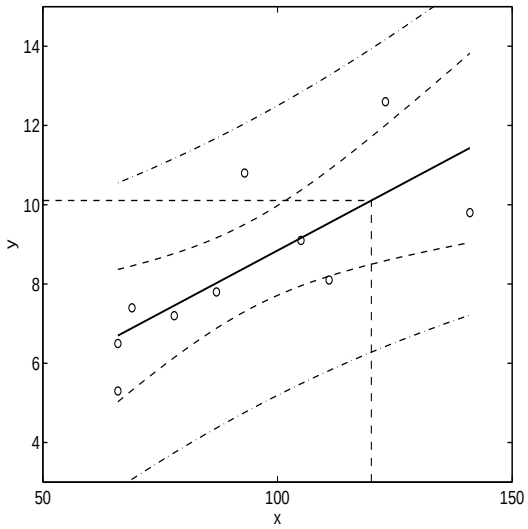
Το 95% διάστημα πρόβλεψης της μέσης απολαβής ρεύματος για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$ δίνεται από τη σχέση (5.23) ως

$$10.108 \pm 2.306 \cdot 0.698 \quad \Rightarrow \quad [8.499, 11.717]$$

Το αντίστοιχο 95% διάστημα πρόβλεψης για μια (μελλοντική) παρατήρηση y της απολαβής ρεύματος για αντίσταση στρώματος $x = 120\,\text{ohm/cm}$ είναι από τη σχέση (5.24)

$$10.108 \pm 2.306 \cdot 1.507 \sqrt{1 + \frac{1}{10} + \frac{(120 - 93.9)^2}{9 \cdot 662.1}} \quad \Rightarrow \quad [6.279, 13.937]$$

Τα διαστήματα πρόβλεψης της μέσης απολαβής ρεύματος και μελλοντικής απολαβής για τιμές αντίστασης στρώματος στο διάστημα που ορίζεται από το δείγμα δίνονται στο Σχήμα 5.8. Παρατηρούμε πως και τα δύο διαστήματα πρόβλεψης είναι πιο μικρά για τιμές αντίστασης στρώματος κοντά στη δειγματική μέση τιμή της αντίστασης στρώματος. Σε αυτό το παράδειγμα των 10 ζευγαρωτών παρατηρήσεων το 95% διάστημα πρόβλεψης των τιμών απολαβής ρεύματος περιέχει και τις 10 τιμές.



Σχήμα 5.8: Διάγραμμα διασποράς για τα δεδομένα του Πίνακα 5.2, ευθεία ελαχίστων τετραγώνων και διαστήματα πρόβλεψης μέσης και μελλοντικής απολαβής ρεύματος.

5.2.7 Επάρκεια μοντέλου απλής γραμμικής παλινδρόμησης

Γενικά όταν έχουμε ένα πραγματικό πρόβλημα, όπως εδώ το πρόβλημα της απλής γραμμικής παλινδρόμησης, δε γνωρίζουμε από πριν ποιο είναι το κατάλληλο μοντέλο. Θα πρέπει λοιπόν να διερευνήσουμε αν το μοντέλο είναι επαρκές, δηλαδή αν εξηγεί ικανοποιητικά τις σχέσεις των μεταβλητών στο πρόβλημα μας. Για την περίπτωση της απλής γραμμικής παλινδρόμησης που μελετάμε εδώ θα πρέπει να εξετάσουμε αν η προσαρμογή ευθείας ερμηνεύει πλήρως τη σχέση της εξαρτημένης μεταβλητής Y από την ανεξάρτητη μεταβλητή X, ή μπορεί μια άλλη σχέση (πιο πολύπλοκη από την απλή ευθεία) να εξηγεί καλύτερα. Κάποιος μπορεί να προσπαθήσει να απαντήσει σε αυτό το ερώτημα μελετώντας το διάγραμμα διασποράς της Y προς τη X, δηλαδή αν φαίνεται κάποια άλλη καμπύλη να προσαρμόζεται καλύτερα στα σημεία (X, Y), αλλά την απάντηση καλύτερα μας τη δίνουν τα υπόλοιπα $e_i = y_i - \hat{y}_i$ της προσαρμογής της ευθείας των ελαχίστων τετραγώνων.

Η διερεύνηση της καταλληλότητας του μοντέλου γίνεται στα τυποποιημένα σφάλματα προσαρμογής, $e_i^* = e_i/s_e$, όπου s_e είναι η εκτιμώμενη τυπική απόκλιση του σφάλματος e_i που δίνεται από τη ρίζα της διασποράς στη σχέση (5.12) ή (5.13) και η μέση τιμή του είναι 0. Θεωρώντας κανονική κατανομή για το σφάλμα e_i υπάρχει πιο ακριβής έκφραση για τη διασπορά του αλλά θα περιοριστούμε εδώ στο s_e^2 όπως το ορίσαμε γενικά.

Κατάλληλα γραφήματα των σφαλμάτων μπορούν να διαγνώσουν την καταλληλότητα και επάρκεια του μοντέλου. Το πιο σημαντικό **διαγνωστικό γράφημα** (diagnostic plot) είναι το διάγραμμα διασποράς των τυποποιημένων σφαλμάτων ως προς την εκτιμώμενη εξαρτημένη μεταβλητή $\hat{y}_i$. Σε αυτό συνήθως σχηματίζονται και δύο οριζόντιες γραμμές στο επίπεδο 2 και -2 (για την ακρίβεια ± 1.96) που αντιστοιχούν στα 95% όρια των τιμών του e_i^* αν αυτό ακολουθεί τυπική κανονική κατανομή. Κάποια από τα προβλήματα στην καλή προσαρμογή του μοντέλου απλής γραμμικής παλινδρόμησης δίνονται στο Σχήμα 5.9 και εξηγούνται παρακάτω.

- Το ποσοστό των σημείων στο γράφημα e_i^* προς $\hat{y}_i$ έξω από τα όρια υπερβαίνει το 5% (περίπου). Τότε η κανονικότητα των σφαλμάτων αμφισβητείται και άρα τα παραμετρικά διαστήματα και ο έλεγχος για τις παραμέτρους, καθώς και τα διαστήματα πρόβλεψης, δεν είναι ακριβή (δες Σχήμα 5.9α).

- Η διασπορά των σφαλμάτων δεν είναι σταθερή αλλά αλλάζει με το $\hat{y}_i$. Η υπόθεση της σταθερής διασποράς των σφαλμάτων δεν ισχύει και άρα και πάλι τα παραμετρικά διαστήματα δεν ισχύουν (δες Σχήμα 5.9β).

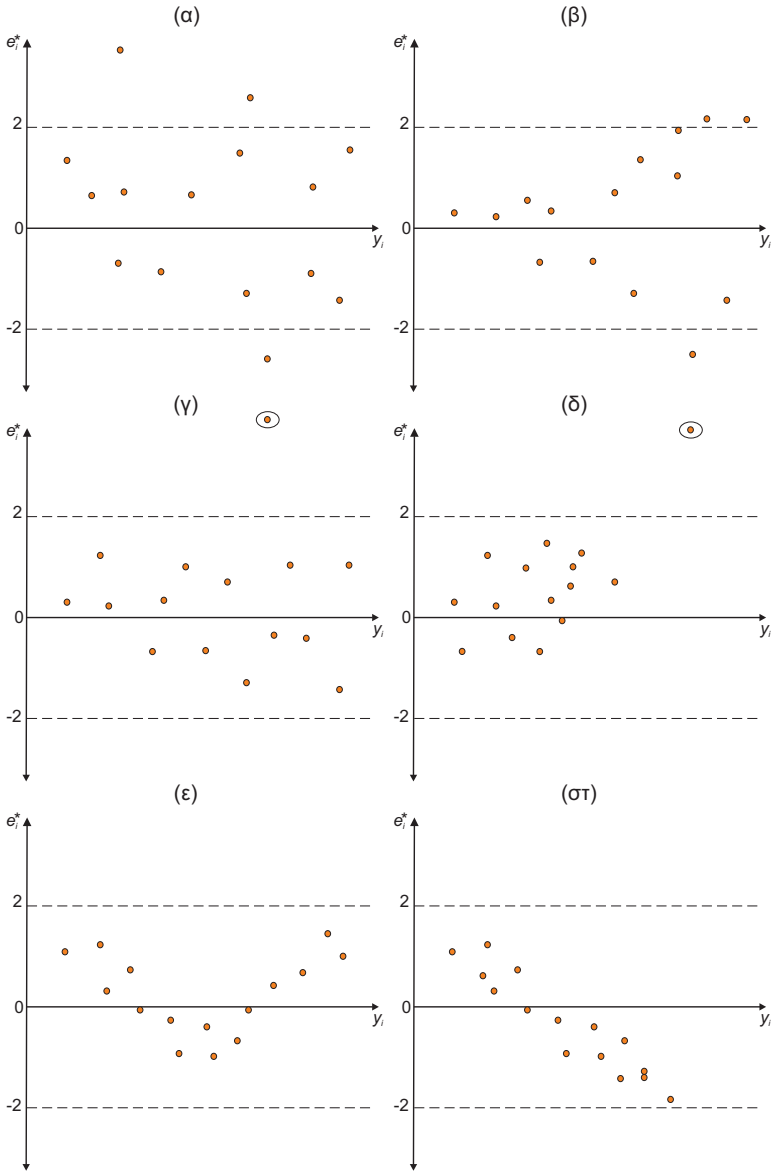
- Κάποιο σφάλμα είναι πολύ μεγαλύτερο από όλα τα άλλα. Αυτό έχει επηρεάσει σημαντικά την εκτίμηση των παραμέτρων του μοντέλου (δηλαδή τους συντελεστές της ευθείας ελαχίστων τετραγώνω). Άρα αν παραλείψουμε το ζεύγος τιμών των (X, Y) που αντιστοιχεί στο μεγάλο σφάλμα, το εκτιμούμενο μοντέλο στο νέο δείγμα μπορεί να είναι σημαντικά διαφορετικό (δες Σχήμα 5.9γ).

- Υπάρχει απόμακρο σημείο (outlier). Αυτό και πάλι επηρεάζει σημαντικά την εκτίμηση των παραμέτρων του μοντέλου και αν το απαλείψουμε μπορεί το μοντέλο της απλής γραμμικής παλινδρόμησης να μη φαίνεται πλέον κατάλληλο (δες Σχήμα 5.9δ).

- Τα σφάλματα διαμορφώνουν κάποιο σχήμα καμπύλης. Η υπόθεση της γραμμικής εξάρτησης δεν ισχύει και η εξάρτηση μπορεί να είναι μηγραμμική (δες Σχήμα 5.9ε).

- Τα σφάλματα φαίνεται να συγκεντρώνονται γύρω από μια γραμμή. Η εξαρτημένη μεταβλητή μπορεί να εξαρτιέται γραμμικά και από κάποια δεύτερη μεταβλητή που πρέπει να συμπεριληφθεί στο μοντέλο γραμμικής παλινδρόμησης (δες Σχήμα 5.9στ).

Ειδικότερα οι δύο τελευταίες δυσκολίες που παρουσιάστηκαν στην προσαρμογή του μοντέλου απλής γραμμικής παλινδρόμησης στα Σχήματα 5.9ε



Σχήμα 5.9: Διαγράμματα διασποράς των τυποποιημένων σφαλμάτων προς τις εκτιμήσεις της εξαρτημένης μεταβλητής, που δείχνουν ανεπάρκεια του μοντέλου. (α) Μη-κανονική κατανομή (πολλά σημεία εκτός των ορίων του 95% της κανονικής κατανομής. (β) Η διασπορά των σφαλμάτων δεν είναι σταθερή. (γ) Ακραία τιμή σφάλματος. (δ) Απομακρυσμένη τιμή σφάλματος και ανεξάρτητης μεταβλητής. (ε) Ύπαρξη μη-γραμμικής εξάρτησης. (στ) Ύπαρξη επιπλέον ανεξάρτητης μεταβλητής.

και 5.9στ, συνιστούν τη χρήση άλλου τύπου μοντέλου, απλής μη-γραμμικής παλινδρόμησης και πολλαπλής γραμμικής παλινδρόμησης, αντίστοιχα, που θα δούμε παρακάτω.

5.3 Μη-Γραμμική Παλινδρόμηση

Σε κάποια προβλήματα μπορεί να έχουμε θεωρητικές ενδείξεις ότι η εξάρτηση μιας εξαρτημένης τ.μ. Y από μια ανεξάρτητη μεταβλητή X είναι κάποιας συγκεκριμένης μη-γραμμικής μορφής. Σε κάποιες περιπτώσεις, μπορεί τη μη-γραμμική μορφή να μας την υποδείξει το διαγνωστικό γράφημα για ένα μοντέλο απλής γραμμικής παλινδόμησης, όπως στο Σχήμα 5.9ε. Σε κάθε περίπτωση έχουμε να εκτιμήσουμε τις παραμέτρους μιας μη γραμμικής συνάρτησης.

5.3.1 Εγγενής γραμμική συνάρτηση παλινδρόμησης

Κάποιες μη-γραμμικές συναρτήσεις μπορεί με κατάλληλο μετασχηματισμό να γίνουν γραμμικές, και τότε μια τέτοια συνάρτηση λέγεται **εγγενής γραμμική** (intrinsically linear). Στον Πίνακα 5.3 δίνονται τέσσερις από τις πιο γνωστές εγγενείς γραμμικές συναρτήσεις και για κάθε μια δίνεται ο κατάλληλος μετασχηματισμός και η γραμμική μορφή που δίνει ο μετασχηματισμός.

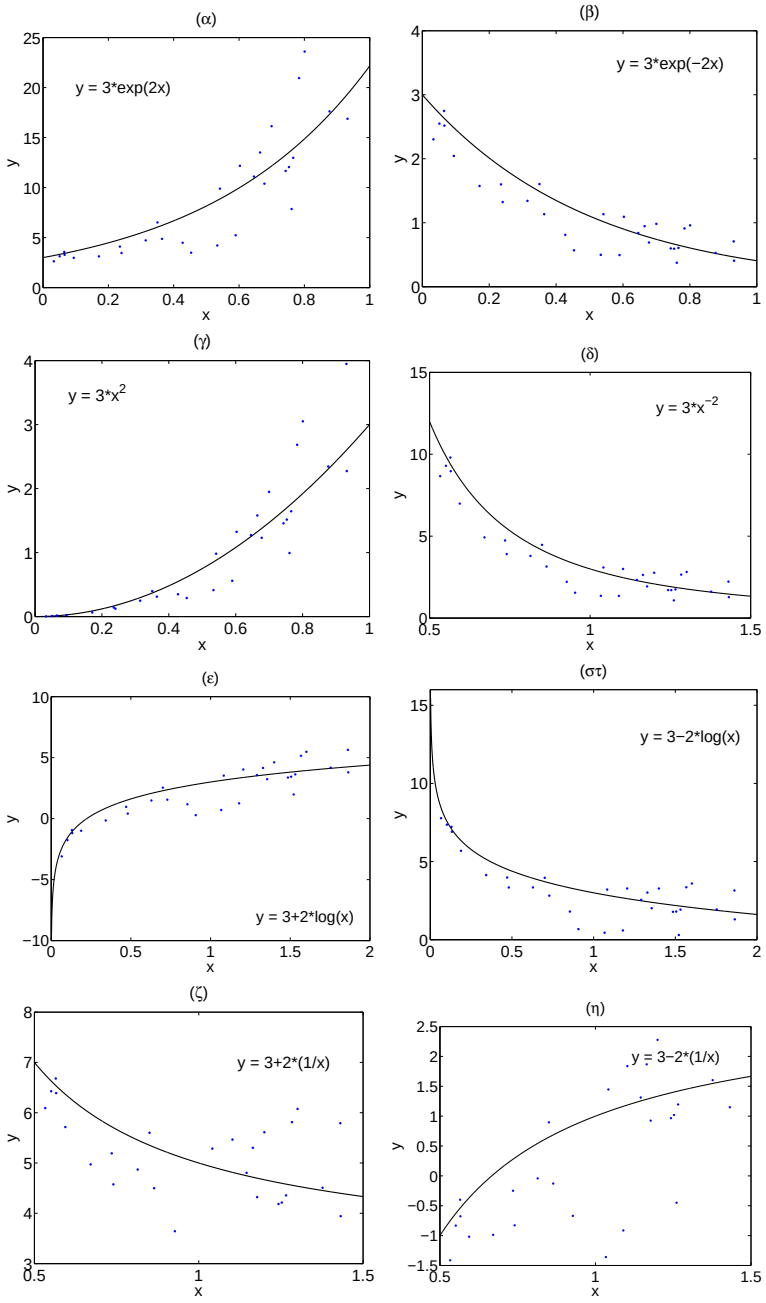
Εγγενής συνάρτηση	*Μετασχηματισμός*	*Γραμμική συνάρτηση*
1. Εκθετική: $y = ae^{\beta x}$	$y' = \ln(y)$	$y' = \ln(a) + \beta x$
2. Δύναμης: $y = ax^{\beta}$	$y' = \log(y)$, $x' = \log(x)$	$y' = \log(a) + \beta x'$
3. $y = a + \beta \log(x)$	$x' = \log(x)$	$y = a + \beta x'$
4. Αντίστροφη: $y = a + \beta \frac{1}{x}$	$x' = \frac{1}{x}$	$y = a + \beta x'$

Πίνακας 5.3: Εγγενείς γραμμικές συναρτήσεις, οι κατάλληλοι μετασχηματισμοί και οι γραμμικές συναρτήσεις που προκύπτουν από τους μετασχηματισμούς. Όπου δίνεται ο δεκαδικός λογάριθμος μπορεί ισοδύναμα να χρησιμοποιηθεί ο νεπέρειος λογάριθμος.

Το βασικό πλεονέκτημα που έχουμε όταν γνωρίζουμε πως η μορφή της συνάρτησης παλινδρόμησης δεν είναι μια οποιαδήποτε μη-γραμμική συνάρτηση αλλά είναι εγγενής γραμμική είναι πως μπορούμε να εκτιμήσουμε τις παραμέτρους της συνάρτησης με τη μέθοδο των ελαχίστων τετραγώνων το ίδιο

εύκολα όπως στη γραμμική παλινδρόμηση. Αυτό συμβαίνει γιατί η συνάρτηση του αθροίσματος των τετραγώνων των σφαλμάτων παραμένει γραμμική ως προς τις παραμέτρους.

Για κάθε εγγενή γραμμική συνάρτηση, η αντίστοιχη στοχαστική συνάρτηση που σχηματίζεται προσθέτοντας θόρυβο ϵ δεν είναι πάντα εγγενής γραμμική. Για παράδειγμα το εκθετικό στοχαστικό μοντέλο $y = ae^{\beta x} + \epsilon$ ή το στοχαστικό μοντέλο δύναμης $y = ax^{\beta} + \epsilon$ (και τα δύο θεωρώντας προσθετικό θόρυβο ϵ) δεν είναι εγγενείς στοχαστικές συναρτήσεις γιατί ο μετασχηματισμός της λογαρίθμησης εφαρμόζεται σε άθροισμα με έναν όρο την ανεξάρτητη μεταβλητή x και δεύτερο όρο το θόρυβο ϵ και άρα δε μπορούν αυτά να διαχωριστούν. Αν όμως θεωρήσουμε πολλαπλασιαστικό θόρυβο ϵ, δηλαδή $y = ae^{\beta x} \cdot \epsilon$ ή $y = ax^{\beta} \cdot \epsilon$, τότε ο διαχωρισμός είναι δυνατός. Μάλιστα αν ο θόρυβος ϵ έχει λογαριθμική κανονική κατανομή (lognormal distribution) τότε ο μετασχηματισμός δίνει θόρυβο $\epsilon' = \ln \epsilon$ με κανονική κατανομή. Για τις δύο άλλες εγγενείς γραμμικές συναρτήσεις (τις δύο τελευταίες στον Πίνακα 5.3) ο θόρυβος πρέπει να είναι προσθετικός και τότε οι στοχαστικές συναρτήσεις $y = a + \beta \log(x) + \epsilon$ και $y = a + \beta \frac{1}{x} + \epsilon$ είναι εγγενείς γραμμικές, δηλαδή ο μετασχηματισμός δίνει ισοδύναμο γραμμικό μοντέλο παλινδρόμησης. Στο Σχήμα 5.10 παρουσιάζονται οι εγγενείς συνάρτησεις και παρατηρήσεις που προκύπτουν από αυτές προσθέτοντας κατάλληλο θόρυβο (από λογαριθμική κατανομή στις δύο πρώτες εγγενείς συναρτήσεις και από κανονική κατανομή για τις άλλες δύο).



Σχήμα 5.10: Οι 4 εγγενείς γραμμικές συναρτήσεις του Πίνακα 5.3 για θετικό και αρνητικό συντελεστή β, καθώς και παρατηρήσεις από αυτές με θόρυβο: εγγενής συνάρτηση 1 με $\beta > 0$ και $\beta < 0$ στο (α) και (β), αντίστοιχα συνάρτηση 2 στο (γ) και (δ), συνάρτηση 3 στο (ε) και (στ), και συνάρτηση 4 στο (ζ) και (η).

Παράδειγμα 5.5. Για ένα ιδανικό αέριο ισχύει $pV^\gamma = C$ για κάποια σταθερά C, όπου p είναι η απόλυτη πίεση του αερίου, V ο όγκος του και γ είναι ένας εκθέτης χαρακτηριστικός για το ιδανικό αέριο (λέγεται και λόγος των ειδικών θερμοτήτων (ratio of the specific heats)). Θέλουμε να εκτιμήσουμε τον εκθέτη γ καθώς και τη σταθερά C από τις παρακάτω μετρήσεις απόλυτης πίεσης και όγκου του ιδανικού αερίου στον Πίνακα 5.4. Επίσης θέλουμε να προβλέψουμε την απόλυτη πίεση για όγκο ιδανικού αερίου 25 in.^3.

A/A	p [psi]	V [in.^3]
1	16.6	50
2	39.7	30
3	78.5	20
4	115.5	15
5	195.3	10
6	546.1	5

Πίνακας 5.4: Μετρήσεις απόλυτη πίεσης και όγκου για ένα ιδανικό αέριο.

Πράγματι στο Σχήμα 5.11α οι μετρήσεις υποδεικνύουν μια μη-γραμμική σχέση της απόλυτης πίεσης και του όγκου του ιδανικού αερίου. Η σχέση του προβλήματος δηλώνει εγγενή γραμμική συνάρτηση της μορφής δύναμης (συνάρτηση 2 του Πίνακα 5.3)

$$pV^\gamma = C \quad \Leftrightarrow \quad y = ax^\beta,$$

όπου η εξαρτημένη μεταβλητή είναι $y = p$, η ανεξάρτητη μεταβλητή είναι $x = V$, και οι παράμετροι είναι $a = C$ και $\beta = -\gamma$. Εφαρμόζοντας λογάριθμο και στα δύο μέρη της ισότητας και θέτοντας $y' = \ln(y) = \ln(p)$ και $x' = \ln(x) = \ln(V)$ παίρνουμε το γραμμικό μοντέλο

$$y' = \ln(a) + \beta x' \quad \Leftrightarrow \quad \ln(p) = \ln(C) - \gamma \ln(V).$$

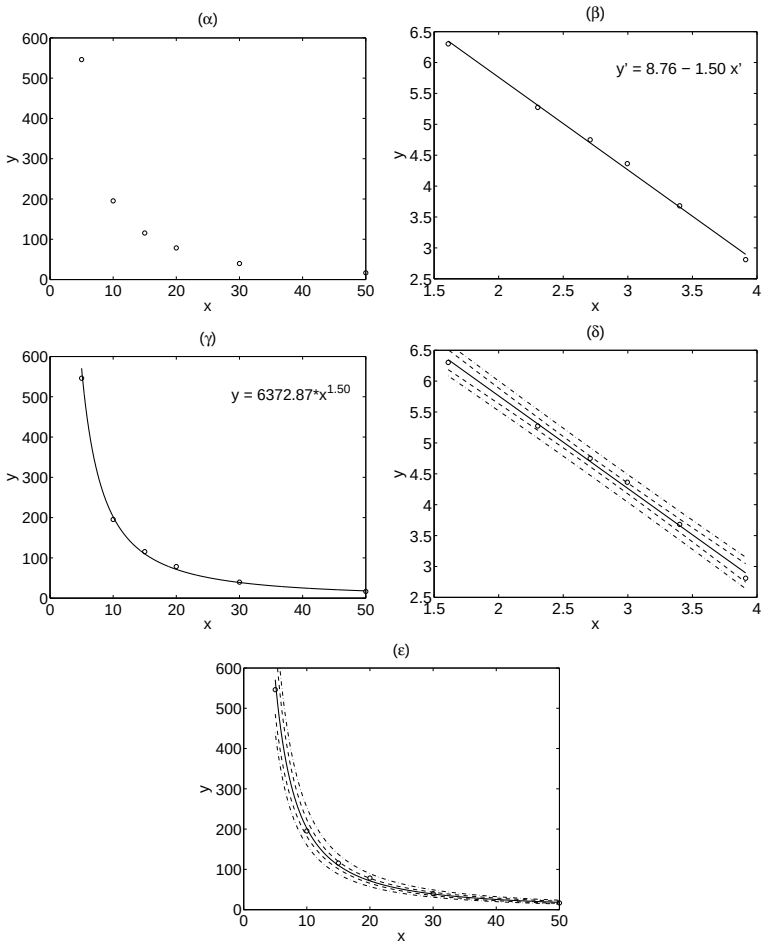
Θεωρώντας ότι ο θόρυβος στις παρατηρήσεις του Πίνακα 5.4 εισέρχεται πολλαπλασιαστικά στο αρχικό μοντέλο, δηλαδή

$$pV^\gamma = C \cdot \epsilon \quad \Leftrightarrow \quad y = ax^\beta \cdot \epsilon,$$

και θέτοντας $\epsilon' = \ln(\epsilon)$ έχουμε το πρόβλημα της απλής γραμμικής παλινδρόμησης του y' προς τη x' στη μορφή

$$y' = \ln(a) + \beta x' + \epsilon' \quad \Leftrightarrow \quad \ln(p) = \ln(C) - \gamma \ln(V) + \ln(\epsilon).$$

Τα μετασχηματισμένα δεδομένα για x' και y' δίνονται στον Πίνακα 5.5. Το διάγραμμα διασποράς τους στο Σχήμα 5.11β υποδηλώνει τη γραμμική



Σχήμα 5.11: (α) Διάγραμμα διασποράς των 6 παρατηρήσεων απόλυτης πίεσης p και όγκου V για ένα ιδανικό αέριο. (β) Διάγραμμα διασποράς των μετασχηματισμένων p και V και προσαρμογή ευθείας ελαχίστων τετραγώνων. (γ) Διάγραμμα διασποράς όπως στο (α) και προσαρμογή της συνάρτησης που προκύπτει από τον αντίστροφο μετασχηματισμό της ευθείας ελαχίστων τετραγώνων. (δ) Όπως στο (β) αλλά με τα 95% διαστήματα πρόβλεψης για τη μέση τιμή και για μια τιμή της y' για κάθε x'. (ε) Όπως στο (γ) αλλά με τα 95% διαστήματα πρόβλεψης για τη μέση τιμή και για μια τιμή της απόλυτης πίεσης για κάθε τιμή του όγκου του ιδανικού αερίου.

σχέση του λογαρίθμου της απόλυτης πίεσης y' και του λογαρίθμου του όγκου x' του ιδανικού αερίου.

Εκτιμούμε τις παραμέτρους b_0 και b_1 της ευθείας ελαχίστων τετραγώνων. Έχουμε $\bar{x} = 2.82$, $\bar{y} = 4.53$, $s_X^2 = 0.6614$ και $s_{XY} = -0.9915$, οπότε εκτιμή-

A/A	x'	y'
1	1.609	6.303
2	2.303	5.275
3	2.708	4.749
4	2.996	4.363
5	3.401	3.681
6	3.912	2.809

Πίνακας 5.5: Τιμές λογαρίθμων των μετρήσεων απόλυτη πίεσης και όγκου για ένα ιδανικό αέριο.

σεις των συντελεστών της ευθείας είναι

$$\begin{aligned} b_1 &= \frac{-0.9915}{0.6614} = -1.4991 \\ b_0 &= 4.53 + 1.4991 \cdot 2.82 = 8.7598. \end{aligned}$$

Από τη σχέση (5.13) και έχοντας $s_Y^2 = 1.4908$, υπολογίζουμε την εκτίμηση διασποράς των σφαλμάτων παλινδρόμησης

$$s_\epsilon^2 = \frac{5}{4}(1.4908 - 1.4991^2 \cdot 0.6614) = 0.00554,$$

και αντίστοιχα οι τυπικές αποκλίσεις των σφαλμάτων είναι $s_\epsilon = 0.07442$.

Για να προβλέψουμε την απόλυτη πίεση για όγκο ιδανικού αερίου 25 in.^{3}, χρησιμοποιούμε το λογάριθμο της τιμής του όγκου $x' = \ln(25) = 3.219$ και η πρόβλεψη του λογαρίθμου της απόλυτης πίεσης είναι

$$\ln(p) = y' = 8.7598 - 1.4991 \cdot 3.219 = 3.9344.$$

Άρα η πρόβλεψη της απόλυτης πίεσης για όγκο ιδανικού αερίου 25 in.^{3} είναι $p = \exp(3.9344) = 51.13$ psi. Στο Σχήμα 5.11γ δίνεται η καμπύλη που δίνει την απόλυτη πίεση p ως συνάρτηση του όγκου V.

Μπορούμε επίσης να υπολογίσουμε διαστήματα πρόβλεψης για τη μέση απόλυτη πίεση και για μια πρόβλεψη της απόλυτης πίεσης για κάποιο δεδομένο όγκο του ιδανικού αερίου. Υπολογίζουμε πρώτα τα διαστήματα πρόβλεψης για το μετασχηματισμένο γραμμικό μοντέλο και παίρνουμε τον αντίστροφο μετασχηματισμό στα όρια του διαστήματος πρόβλεψης.

Για $x' = \ln(25) = 3.219$ από τη σχέση (5.23) υπολογίζουμε το 95% διάστημα πρόβλεψης της μέσης y' ως $[3.839, 4.030]$ και εφαρμόζοντας την εκθετική συνάρτηση στα άκρα προκύπτει το 95% διάστημα πρόβλεψης της μέσης απόλυτης πίεσης για όγκο ιδανικού αερίου 25 in.^{3} ως $[46.465, 56.264]$. Αντίστοιχα τα όρια του 95% διαστήματος πρόβλεψης για μια τιμή του y' όταν

$x' = \ln(25) = 3.219$ είναι $[3.707, 4.162]$ και τα αντίστοιχα όρια για μια τιμή απόλυτης πίεσης για όγκο ιδανικού αερίου 25 in.^3 είναι $[40.718, 64.205]$. Τα 95% διαστήματα πρόβλεψης (μέσης και μιας τιμής) για το γραμμικοποιημένο και το αρχικό μοντέλο παλινδρόμησης δίνονται στο Σχήμα 5.11δ και 5.11ε, αντίστοιχα. Παρατηρούμε πως η πρόβλεψη της απόλυτης πίεσης είναι λιγότερη ακριβής για μικρούς όγκους ιδανικού αερίου.

Θα πρέπει να σημειωθεί πως στην περίπτωση εγγενούς γραμμικής συνάρτησης παλινδρόμησης, οι εκτιμήσεις των παραμέτρων στο μετασχηματισμένο γραμμικό μοντέλο παλινδρόμησης είναι οι καλύτερες. Από αυτές μπορούμε να εκτιμήσουμε τις παραμέτρους του αρχικού μη-γραμμικού μοντέλου παλινδρόμησης, αλλά αυτές οι εκτιμήσεις δεν είναι οι καλύτερες, με την έννοια της ελαχιστοποίησης των σφαλμάτων. Για να το πετύχουμε αυτό θα πρέπει να εφαρμόσουμε τη μέθοδο ελαχίστων τετραγώνων απευθείας στο αρχικό μη-γραμμικό σύστημα. Η μέθοδος αυτή απαιτεί τη λύση ενός μη-γραμμικού συστήματος εξισώσεων ως προς τις παραμέτρους, που ανάλογα με τη μορφή της μη-γραμμικής συνάρτησης παλινδρόμησης μπορεί να είναι αρκετά σύνθετη.

Στη συνέχεια θα δούμε έναν τύπο μη-γραμμικής παλινδρόμησης, την πολυωνυμική παλινδρόμηση, που δεν αντιμετωπίζεται με μετασχηματισμό σε γραμμική συνάρτηση αλλά εκτιμάται απευθείας με χρήση της μεθόδου ελαχίστων τετραγώνων.

5.3.2 Πολυωνυμική παλινδρόμηση

Τα μη-γραμμικά μοντέλα παλινδρόμησης που δίνονται από εγγενείς γραμμικές συναρτήσεις της εξαρτημένης μεταβλητής y προς την ανεξάρτητη μεταβλητή x έχουν ως κοινό χαρακτηριστικό ότι οι συναρτήσεις είναι μονότονες, αύξουσες ή φθίνουσες (δες Σχήμα 5.10). Σε πολλά προβλήματα η θεωρητική προσέγγιση ή το διάγραμμα διασποράς συνιστά ότι η συνάρτηση έχει ένα ή περισσότερα σημεία καμπής. Σε τέτοιες περιπτώσεις η πολυωνυμική συνάρτηση κάποιου βαθμού k μπορεί να αποτελεί ικανοποιητική προσέγγιση της πραγματικής συνάρτησης παλινδρόμησης.

Το **μοντέλο πολυωνυμικής γραμμικής παλινδρόμησης βαθμού** k (k-th degree polynomial regression model) δίνεται από την εξίσωση

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon. \tag{5.25}$$

Όπως και στη γραμμική παλινδρόμηση υποθέτουμε πως τα σφάλματα παλινδρόμησης ακολουθούν κανονική κατανομή με μέση τιμή 0 και διασπορά σ_e^2. Η υπόθεση αυτή μας επιτρέπει να εκτιμήσουμε διαστήματα εμπιστοσύνης και

να κάνουμε ελέγχους για τις παραμέτρους του μοντέλου, όπως και για να εκτιμήσουμε διαστήματα πρόβλεψης. Η μέθοδος ελαχίστων τετραγώνων όμως δεν προϋποθέτει κανονικότητα των σφαλμάτων για να δώσει τις καλύτερες (σημειακές) εκτιμήσεις των παραμέτρων.

Η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο ελαχίστων τετραγώνων όπως και για το γραμμικό μοντέλο παλινδρόμησης γιατί ενώ η πολυωνυμική συνάρτηση παλινδρόμησης είναι μη-γραμμική ως προς την ανεξάρτητη μεταβλητή x, είναι γραμμική ως προς τους συντελεστές $\beta_0, \beta_1, \ldots, \beta_k$. Το άθροισμα των τετραγώνων των σφαλμάτων για κάποιο διμεταβλητό δείγμα μεγέθους n των (X, Y), είναι

$$f(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k) \right)^2 .$$

Οι μερικοί παράγωγοι της συνάρτησης αυτής ως προς κάθε παράμετρο β_0, $beta_1, \ldots, \beta_k$, δίνει ένα σύστημα k εξισώσεων, γνωστό και ως σύστημα **κανονικών εξισώσεων** (normal equations) από το οποίο βρίσκουμε τις εκτιμήσεις $b_0, b_1, \ldots, b_k$

$$\begin{aligned}
b_0 n + b_1 \sum x_i + b_2 \sum x_i^2 + \cdots b_k \sum x_i^k &= \sum y_i \\
b_0 \sum x_i + b_1 \sum x_i^2 + b_2 \sum x_i^3 + \cdots b_k \sum x_i^{k+1} &= \sum x_i y_i \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad &\vdots \\
b_0 \sum x_i^k + b_1 \sum x_i^{k+1} + b_2 \sum x_i^{k+2} + \cdots b_k \sum x_i^{2k} &= \sum x_i^k y_i
\end{aligned} \tag{5.26}$$

Τα σφάλματα του μοντέλου πολυωνυμικής παλινδρόμησης που εκτιμήθηκε με τη μέθοδο ελαχίστων τετραγώνων είναι $e_i = y_i - \hat{y}_i$, όπου

$$\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_k x_i^k .$$

Η εκτίμηση της διασποράς των σφαλμάτων e_i ορίζεται ως

$$s_e^2 = \frac{1}{n - (k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{5.27}$$

Όπως και για την απλή γραμμική παλινδρόμηση και συσχέτιση, ορίζουμε τον **συντελεστή του πολλαπλού προσδιορισμού** (coefficient of multiple determination) R^2 ως

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5.28}$$

που δηλώνει την αναλογία της μεταβλητότητας που εξηγείται από το μοντέλο.

Είναι φυσικό πως η πρόσθεση περισσότερων μη-γραμμικών όρων (δυνάμεων) της ανεξάρτητης μεταβλητής x στο πολυωνυμικό μοντέλο παλινδρόμησης θα βελτιώσει την προσαρμογή του στις n ζευγαρωτές παρατηρήσεις, χωρίς αυτό να σημαίνει ότι ένας μεγάλος βαθμός k είναι πάντα ο πιο κατάλληλος. Γι αυτό χρησιμοποιούμε τον **προσαρμοσμένο συντελεστή του πολλαπλού προσδιορισμού** (adjusted coefficient of multiple determination)

$$\text{adj}R^2 = 1 - \frac{n-1}{n-(k+1)} \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.29}$$

που δίνει μικρότερες τιμές από το R^2 της σχέσης (5.28) που το ποσό μείωσης προσαρμόζεται στο πλήθος των μη-γραμμικών όρων k.

Για τις παραμέτρους $\beta_0, \beta_1, \ldots, \beta_k$ μπορούμε να εκτιμήσουμε διαστήματα εμπιστοσύνης και να κάνουμε στατιστικούς ελέγχους όπως και για τις παραμέτρους του γραμμικού μοντέλου παλινδρόμησης, αφού υποθέσουμε πως τα σφάλματα ακολουθούν κανονική κατανομή με σταθερή διασπορά για κάθε τιμή του x. Για την πολυωνυμική παλινδρόμηση η διασπορά της εκτίμησης για κάθε μια από τις $\beta_0, \beta_1, \ldots, \beta_k$ δίνεται από σχετικά πολύπλοκους τύπους που εξαρτώνται και από το βαθμό του πολυώνυμου. Το ίδιο ισχύει και για τα διαστήματα πρόβλεψης.

Παράδειγμα 5.6. Στον Πίνακα 5.6 δίνονται τα δεδομένα για την ημέρα της συγκομιδής (αριθμός ημερών αφού ανθίσει) και του μεγέθους σοδειάς (σε kg/ha) ενός είδους Ινδικού ρυζιού που λέγεται paddy.

Το διάγραμμα διασποράς δίνεται στο Σχήμα 5.12α. Η καμπύλη που φαίνεται να προσαρμόζεται καλύτερα είναι παραβολή, δηλαδή πολυώνυμο δευτέρου βαθμού. Προσαρμόζουμε πολυώνυμα από πρώτο ως τέταρτο βαθμό. Οι προσαρμογές δίνονται στο Σχήμα 5.12. Οι εκτιμήσεις των παραμέτρων δίνονται ως λύση των κανονικών εξισώσεων στη σχέση (5.31). Υπολογίζουμε τα σφάλματα προσαρμογής του κάθε μοντέλου και αντίστοιχα τον συντελεστή προσδιορισμού και τον προσαρμοσμένο συντελεστή προσδιορισμού, τα οποία και εμφανίζοντα με κάθε προσαρμογή μοντέλου στο Σχήμα 5.12.

Είναι φανερό πως το γραμμικό μοντέλο παλινδρόμησης δεν είναι κατάλληλο. Πράγματι τα σφάλματα παλινδρόμησης δεν είναι τυχαία αλλά θα σχηματίζουν καμπύλη, όπως στο Σχήμα 5.9ε. Επίσης ο συντελεστής προσδιορισμού είναι πολύ κοντά στο 0, που σημαίνει ότι το μοντέλο αδυνατεί πλήρως να ερμηνεύσει τις παρατηρήσεις. Η πρόσθεση του όρου του τετραγώνου των ημερών για τη συγκομιδή (ανεξάρτητη μεταβλητή) δίνει την παραβολή που προσαρμόζεται πολύ καλά στα δεδομένα. Ο συντελεστής προσδιορισμού R^2, καθώς και ο προσαρμοσμένος συντελεστής προσδιορισμού $\text{adj}R^2$, εκτοξεύτηκαν στο επίπεδο του 0.8, που σημαίνει ότι με το πολυωνυμικό μοντέλο

Α/Α	*Ημέρες*	*Σοδειά*
1	16	2508
2	18	2518
3	20	3304
4	22	3423
5	24	3057
6	26	3190
7	28	3590
8	30	3883
9	32	3823
10	34	3646
11	36	3708
12	38	3333
13	40	3517
14	42	3241
15	44	3103
16	46	2776

Πίνακας 5.6: Τιμές ημερών για τη συγκομιδή και μεγέθους σοδειάς για το Ινδικό ρύζι paddy.

δευτέρου βαθμού μπορούμε να εκτιμήσουμε τη σοδειά του paddy όταν δίνεται ο αριθμός ημερών για τη συγκομιδή. Η παραπέρα αύξηση του βαθμού του πολυωνύμου δε φαίνεται να βελτιώνει την παλινδρόμηση του μεγέθους σοδειάς του paddy προς τον αριθμό ημερών για τη συγκομιδή. Πράγματι, το R^2 παραμένει το ίδιο ενώ το adjR^2 μειώνεται.

Το πολυωνυμικό μοντέλο δευτέρου βαθμού εκτιμήθηκε να είναι

$$y = -1.1242 + 0.2979x - 0.0046x^2$$

και μπορούμε να το χρησιμοποιήσουμε για να κάνουμε προβλέψεις του μεγέθους της σοδειάς για κάθε δεδομένη χρονική περίοδο ως τη συγκομιδή.

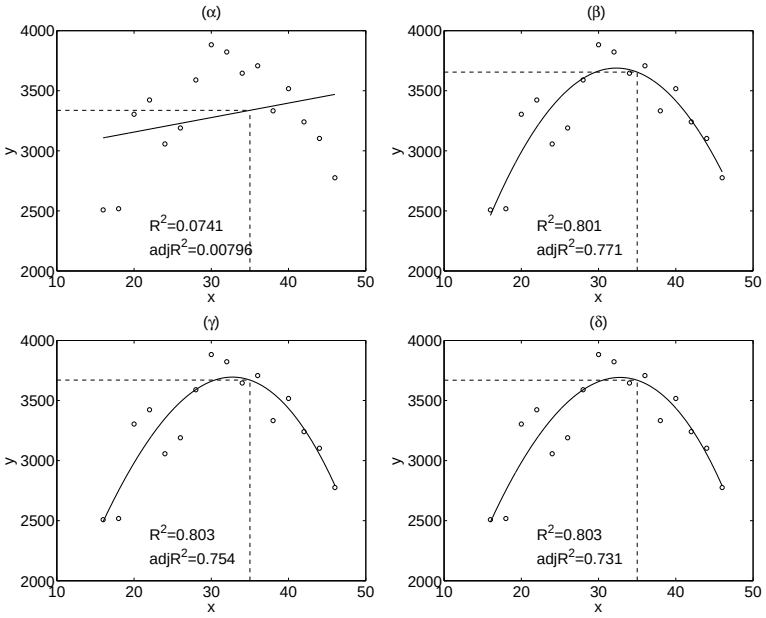
5.4 Πολλαπλή Παλινδρόμηση

Στο μοντέλο πολυωνυμικής παλινδρόμησης που δίνεται στη σχέση (5.25), αν θέσουμε

$$x_1 = x, \quad x_2 = x^2, \quad \ldots \quad x_k = x^k,$$

το μοντέλο γίνεται

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \tag{5.30}$$



Σχήμα 5.12: (α) Διάγραμμα διασποράς των 16 παρατηρήσεων του μεγέθους σοδειάς για το Ινδικό ρύζι paddy για δεδομένες ημέρες για τη συγκομιδή. Σχηματίζεται η ευθεία ελαχίστων τετραγώνων που προσαρμόζεται στα σημεία. Ο συντελεστής προσδιορισμού και ο προσαρμοσμένος συντελεστής προσδιορισμού δίνονται μέσα στο σχήμα. (β) Όπως το (α) αλλά για μοντέλο παλινδρόμησης από πολυώνυμο δευτέρου βαθμού. (γ) Όπως το (α) αλλά για πολυώνυμο τρίτου βαθμού. (δ) Όπως το (α) αλλά για πολυώνυμο τετάρτου βαθμού.

όπου και πάλι υποθέτουμε πως τα σφάλματα παλινδρόμησης ϵ έχουν μέση τιμή 0 και διασπορά σ_ϵ^2 (για να εκτιμήσουμε παραμετρικά διαστήματα εμπιστοσύνης ή να κάνουμε παραμετρικό έλεγχο θα χρειαστεί να υποθέσουμε πως τα σφάλματα ακολουθούν κανονική κατανομή). Το μοντέλο της σχέσης (5.30) είναι το μοντέλο πολλαπλής παλινδρόμησης για την εξαρτημένη μεταβλητή y προς τις ανεξάρτητες μεταβλητές $x_1, x_2, \ldots, x_k$. Βέβαια σε αυτήν την περίπτωση οι μεταβλητές $x_1, x_2, \ldots, x_k$ δεν είναι ανεξάρτητες μεταξύ τους αφού όλες είναι δυνάμεις της ίδιας μεταβλητής x. Η εκτίμηση του μοντέλου στη σχέση (5.30) παραμένει ίδια σα να ήταν ανεξάρτητες. Γενικά οι $x_1, x_2, \ldots, x_k$ αντιπροσωπεύουν διαφορετικά μεγέθη που πιθανόν να επηρεάζουν την y, που και πάλι δε ξέρουμε αν πράγματι είναι μεταξύ τους ανεξάρτητα. Το μοντέλο της σχέσης (5.30) λέγεται **μοντέλο γραμμικής πολλαπλής παλινδρόμησης** (liner multiple regression model).

Ένα μοντέλο πολλαπλής παλινδρόμησης μπορεί να περιέχει και μη-γραμμικούς

όρους, όπως δυνάμεις των (υποθετικά) ανεξάρτητων μεταβλητών $x_1, x_2, \ldots, x_k$ και όρους αλληλεπίδρασης τους. Στη γενική του μορφή ένα τέτοιο μοντέλο λέγεται **μοντέλο προσθετικής πολλαπλής παλινδρόμησης** (additive multiple regression model), όπου ό όρος 'προσθετικής' τονίζει ότι όλοι οι όροι του μοντέλου συμπεριλαμβάνονται αθροιστικά στο μοντέλο. Ας δούμε ένα απλό παράδειγμα με δύο ανεξάρτητες μεταβλητές x_1 και x_2. Τα δυνατά μοντέλα προσθετικής πολλαπλής παλινδρόμησης είναι:

1. Το μοντέλο πρώτου πολυωνυμικού βαθμού (γραμμικής πολλαπλής παλινδρόμησης)
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

2. Το μοντέλο δεύτερου πολυωνυμικού βαθμού χωρίς αλληλεπίδραση
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \epsilon.$$

3. Το μοντέλο πρώτου πολυωνυμικού βαθμού με αλληλεπίδραση
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

4. Το πλήρες μοντέλο δευτέρου πολυωνυμικού βαθμού (με αλληλεπίδραση)
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon.$$

Υπάρχουν διαφορές μεταξύ των τεσσάρων αυτών μοντέλων ως προς τις μεταβλητές x_1 και x_2 αλλά όλα είναι γραμμικά ως προς τις παραμέτρους και η εκτίμηση των παραμέτρων μπορεί να γίνει με την κλασσική μέθοδο των ελαχίστων τετραγώνων. Το ίδιο βέβαια ισχύει και όταν οι μεταβλητές είναι περισσότερες από δύο.

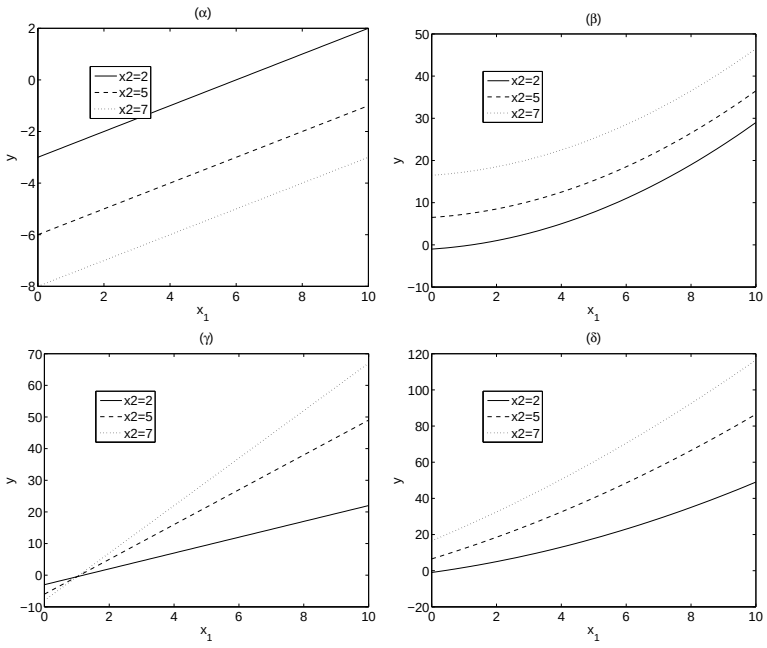
Σε προβλήματα πολλαπλής παλινδρόμησης όπου εμπλέκονται περισσότερες από μια ανεξάρτητες μεταβλητές δε μπορούμε να αποφασίσουμε έυκολα για τη μορφή του μοντέλου, όπως μπορούμε να κάνουμε για την απλή παλινδρόμηση της y ως προς τη x, σχηματίζοντας το διάγραμμα διασποράς του ζεύγους (x, y). Όταν οι ανεξάρτητες μεταβλητές είναι δύο, γραφική εκτίμηση του κατάλληλου μοντέλου μπορεί να γίνει από το γράφημα της (x_1, y) για διαφορετικές τιμές της x_2 (ή ισοδύναμα αντιστρέφοντας τις θέσεις των x_1 και x_2). Για τα παραπάνω 4 μοντέλα δύο ανεξάρτητων μεταβλητών θα περιμέναμε τα εξής για τα γραφήματα του αιτιοκρατικού μέρους του κάθε μοντέλου (αγνοώντας την ύπαρξη του θορύβου ϵ):

1. Το γράφημα των σημείων (x_1, y) είναι σε ευθείες παράλληλες για κάθε τιμή του x_2 γιατί η μεταβολή της y για κάποια μεταβολή της x_1 (π.χ. κατά μια μονάδα) είναι ανεξάρτητη της x_2 (δες Σχήμα 5.13α).

2. Το γράφημα των σημείων (x_1, y) είναι σε καμπύλες παραβολής παράλληλες για κάθε τιμή του x_2 λόγω της παρουσίας των τετραγωνικών όρων (δες Σχήμα 5.13β).

3. Το γράφημα των σημείων (x_1, y) είναι σε ευθείες για κάθε τιμή του x_2 που τέμνονται γιατί η μεταβολή της y ως προς τη x_1 δεν είναι τώρα ανεξάρτητη της x_2 λόγω της παρουσίας του όρου αλληλεπίδρασης (δες Σχήμα 5.13γ).

4. Το γράφημα των σημείων (x_1, y) είναι σε καμπύλες παραβολής για κάθε τιμή του x_2 που δεν είναι παράλληλες λόγω της παρουσίας του όρου αλληλεπίδρασης (δες Σχήμα 5.13γ).



Σχήμα 5.13: Προβολή του γραφήματος της συνάρτησης παλινδρόμησης δύο ανεξάρτητων μεταβλητών (x_1, x_2) στο επίπεδο των (y, x_1) για τρεις τιμές του x_2. Η έκφραση της συνάρτησης είναι: (α) $y = -1 + 0.5x_1 - x_2$, (β) $y = -1 + 0.5x_1 + 25x_1^2 - x_2 + 0.5x_2^2$, (γ) $y = -1 + 0.5x_1 - x_2 + x_1x_2$, (δ) $y = -1 + 0.5x_1 + 25x_1^2 - x_2 + 0.5x_2^2 + x_1x_2$.

Όταν η παλινδρόμηση αφορά περισσότερες από δύο μεταβλητές τότε δεν έχουμε γραφικά εργαλεία για να καθορίσουμε τη μορφή του μοντέλου προσθετικής πολλαπλής παλινδρόμησης και πρέπει να δοκιμάσουμε διαφορετικά

μοντέλα και να επιλέξουμε από αυτά που βρέθηκαν να είναι κατάλληλα (μετά από διαγνωστικό έλεγχο) και προσαρμόζονται καλά στα δεδομένα το πιο απλό.

5.4.1 Εκτίμηση μοντέλου πολλαπλής γραμμικής παλινδρόμησης

Είδαμε παραπάνω ότι σε κάθε περίπτωση ένα μοντέλο προσθετικής πολλαπλής παλινδρόμησης μπορεί να γραφτεί στη μορφή της σχέσης (5.30), όπου οι μεταβλητές $x_1, x_2, \ldots, x_k$, μπορεί να αντιπροσωπεύουν διαφορετικά μεγέθη ή δυνάμεις και αλληλεπιδράσεις τους.

Ας υποθέσουμε πως το πολυ-μεταβλητό δείγμα μεγέθους n είναι $\{x_{1i}, x_{2i}, \ldots, x_{ki}, y_i\}_{i=1}^n$. Η εκτίμηση των παραμέτρων του μοντέλου στη σχέση (5.30) γίνεται με τη μέθοδο ελαχίστων τετραγώνων. Το άθροισμα των τετραγώνων των σφαλμάτων είναι

$$f(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}))^2 .$$

Το σύστημα κανονικών εξισώσεων που προκύπτει από τις μερικές παραγώγους της συνάρτησης αυτής ως προς κάθε παράμετρο $\beta_0, \beta_1, \ldots, \beta_k$, είναι

$$\begin{aligned}
b_0 n + b_1 \sum x_{1i} + b_2 \sum x_{2i} + \cdots b_k \sum x_{ki} &= \sum y_i \\
b_0 \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i} x_{2i} + \cdots b_k \sum x_{1i} x_{ki} &= \sum x_{1i} y_i \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots & \qquad \vdots \\
b_0 \sum x_{ki} + b_1 \sum x_{1i} x_{ki} + b_2 \sum x_{2i} x_{ki} + \cdots b_k \sum x_{ki}^2 &= \sum x_{ki} y_i
\end{aligned} \tag{5.31}$$

από το οποίο βρίσκουμε τις εκτιμήσεις $b_0, b_1, \ldots, b_k$.

Η εκτίμηση της εξαρτημένης μεταβλητής με το μοντέλο πολλαπλής παλινδρόμησης που εκτιμήθηκε με τη μέθοδο ελαχίστων τετραγώνων είναι

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$$

και τα σφάλματα του μοντέλου είναι $e_i = y_i - \hat{y}_i$. Η εκτίμηση της διασποράς των σφαλμάτων s_e^2 ορίζεται όπως και για την πολυωνυμική παλινδρόμηση από τη σχέση (5.27) και αντίστοιχα ορίζεται ο συντελεστής του πολλαπλού προσδιορισμού R^2 από τη σχέση (5.28) και ο προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού $\text{adj}R^2$ από τη σχέση (5.29).

Είναι δυνατόν να υπολογίσουμε παραμετρικά διαστήματα εμπιστοσύνης για τις παραμέτρους $\beta_0, \beta_1, \ldots, \beta_k$ για τις οποίες όμως η εκτίμηση της διασποράς είναι σύνθετη. Γενικά αν η εκτίμηση της διασποράς είναι $s^2_{b_j}$ για $j = 0, 1, \ldots, k$ τότε το $(1-a)\%$ διάστημα εμπιστοσύνης για το συντελεστή β_j είναι

$$b_j \pm t_{n-(k+1),1-a/2} s_{b_j}.$$

Ο έλεγχος για μια τιμή β_j^0 της β_j, H_0: $\beta_j = \beta_j^0$ γίνεται με το στατιστικό

$$t = \frac{\beta_j - \beta_j^0}{s_{b_j}} \sim t_{n-(k+1)}.$$

Το $(1-a)\%$ διάστημα εμπιστοσύνης για τη μέση τιμή της y όταν δίνονται τα $x_1, \ldots, x_k$ είναι

$$\hat{y} \pm t_{n-(k+1),1-a/2} s_{\hat{y}}$$

όπου η διασπορά της εκτίμησης $\hat{y}$, $s^2_{\hat{y}}$, δίνεται από επίσης σύνθετη έκφραση. Αντίστοιχα το $(1-a)\%$ διάστημα πρόβλεψης μιας (μελλοντικής) τιμής της y είναι

$$\hat{y} \pm t_{n-(k+1),1-a/2} \sqrt{s_e^2 + s^2_{\hat{y}}}.$$

Παράδειγμα 5.7. Σε μελέτη της επίδρασης γεωργικών χημικών στην προσρόφηση ιζημάτων και εδάφους, δίνονται στον Πίνακα 5.7 13 δεδομένα για το δείκτη προσρόφησης φωσφορικού άλατος (y), για το εξαγώγιμο σίδηρο (x_1) και το εξαγώγιμο αργίλλιο (x_2).

Το μοντέλο που θα εκτιμήσουμε είναι

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

Οι εκτιμήσεις των συντελεστών του μοντέλου με τη μέθοδο των ελαχίστων τετραγώνων καθώς και η εκτίμηση της τυπικής τους απόκλισης δίνονται στον Πίνακα 5.8.

Το 95% διάστημα εμπιστοσύνης για το συντελεστή του εξαγώγιμου σιδήρου β_1 είναι ($t_{10,0.975} = 2.228$)

$$0.11273 \pm 2.228 \cdot 0.02969 = [0.0466, 0.1789]$$

και αντίστοιχα για το συντελεστή του εξαγώγιμου αργιλλίου β_2 είναι

$$0.34900 \pm 2.228 \cdot 0.07131 = [0.1901, 0.5079].$$

Παρατηρούμε πως και τα δύο διαστήματα εμπιστοσύνης δεν περιέχουν το 0, αλλά μπορούμε με βεβαιότητα σε επίπεδο 95% να συμπεράνουμε πως το εξαγώγιμο σίδηρο και αργίλλιο επηρεάζουν σημαντικά το δείκτη προσρόφησης

Α/Α	Εξαγώγιμο σίδηρο	Εξαγώγιμο αργίλλιο	Δείκτης προσρόφησης
1	61	13	4
2	175	21	18
3	111	24	14
4	124	23	18
5	130	64	26
6	173	38	26
7	169	33	21
8	169	61	30
9	160	39	28
10	244	71	36
11	257	112	65
12	333	88	62
13	199	54	40

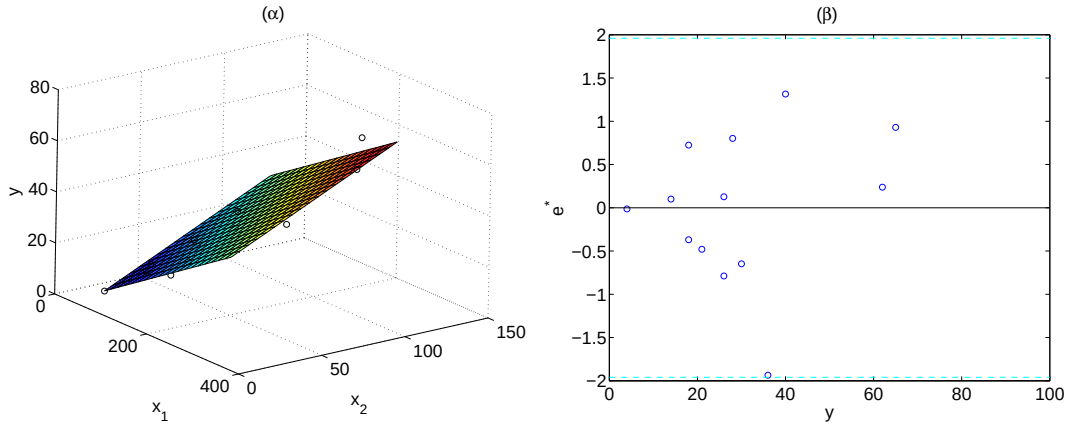
Πίνακας 5.7: Τιμές του δείκτη προσρόφησης, του εξαγώγιμου σιδήρου και εξαγώγιμου αργιλλίου για τη μελέτη της επίδρασης γεωργικών χημικών στο έδαφος.

Παράμετρος	Εκτιμητής b_i	Εκτίμηση SD s_{b_i}
β_0	-7.351	3.485
β_1	0.11273	0.02969
β_2	0.34900	0.07131

Πίνακας 5.8: Εκτίμηση παραμέτρων και τυπική απόκλισης τους.

φωσφορικού άλατος και καλώς συμπεριλαμβάνονται στο μοντέλο. Πράγματι το γράφημα του μοντέλου στο χώρο που ορίζεται από τα σημεία (x_1, x_2, y) είναι ένα επίπεδο που δεν είναι παράλληλο ούτε προς τον άξονα x_1 (εξαγώγιμο σίδηρο) ούτε και προς τον άξονα x_2 (εξαγώγιμο αργίλλιο), όπως φαίνεται στο Σχήμα 5.14α. Το επίπεδο φαίνεται να προσαρμόζεται ικανοποιητικά στα 13 σημεία του δείγματος. Πράγματι η τυπική απόκλιση των σφαλμάτων είναι $s_e = 4.616$, ο συντελεστής του πολλαπλού προσδιορισμού είναι $R^2 = 0.948$ και ο προσαρμοσμένος συντελεστής είναι $adjR^2 = 0.931$. Επίσης, όπως φαίνεται στο διαγνωστικό διάγραμμα διασποράς των τυποποιημένων σφαλμάτων του μοντέλου προς τις τιμές του y στο Σχήμα 5.14β, τα σφάλματα είναι μέσα στα 95% όρια της κανονικής κατανομής και κατανέμονται τυχαία ως προς y.

Έστω ότι θέλουμε να προβλέψουμε το δείκτη προσρόφησης y όταν ο εξαγώγιμος σίδηρος είναι $x_1 = 160$ και ο εξαγώγιμος αργίλλιος είναι $x_2 = 39$. Η



Σχήμα 5.14: (α) Διάγραμμα διασποράς των 13 παρατηρήσεων του δείκτη προσρόφησης (y) για δοθείσες τιμές του εξαγώγιμου σιδήρου (x_1) και του εξαγώγιμου αργιλλίου (x_2). Δίνεται το γράφημα του μοντέλου γραμμικής πολλαπλής παλινδρόμησης που εκτιμήθηκε με τη μέθοδο ελαχίστων τετραγώνων. (β) Διαγνωστικό διάγραμμα διασποράς των τυποποιημένων σφαλμάτων του μοντέλου στο (α) προς τις αντίστοιχες τιμές του y.

πρόβλεψη είναι

$$\hat{y} = -7.351 + 0.11273 \cdot 160 + 0.34900 \cdot 39 = 24.30.$$

Η εκτίμηση της τυπικής απόκλισης για αυτήν την πρόβλεψη $\hat{y}$ βρέθηκε να είναι $s_{\hat{y}} = 1.30$. Το 95% διάστημα εμπιστοσύνης για το μέσο δείκτη προσρόφησης y όταν ο εξαγώγιμος σίδηρος είναι $x_1 = 160$ και ο εξαγώγιμος αργίλλιος είναι $x_2 = 39$ βρίσκεται ως

$$24.30 \pm 2.228 \cdot 1.30 = [21.40, 27.20]$$

και το αντίστοιχο 95% διάστημα πρόβλεψης για μια μελλοντική τιμή του y (για $x_1 = 160$ και $x_2 = 39$) είναι

$$24.30 \pm 2.228 \cdot \sqrt{(4.616)^2 + (1.30)^2} = [13.62, 34.98]$$

5.4.2 Επιλογή μεταβλητών

Σε κάποια προβλήματα παλινδρόμησης έχουμε στη διάθεση μας δεδομένα από πολλούς παράγοντες που μπορεί να επηρεάζουν την εξαρτημένη μεταβλητή που μας ενδιαφέρει να καθορίσουμε ή να προβλέψουμε. Θα θέλαμε λοιπόν να επιλέξουμε το μικρότερο δυνατόν υποσύνολο ανεξάρτητων

(επεξηγηματικών) μεταβλητών που εξηγεί σχεδόν το ίδιο καλά την εξαρτημένη μεταβλητή όπως και μεγαλύτερα υποσύνολα ανεξάρτητων μεταβλητών ή ακόμα και ολόκληρο το σύνολο ανεξάρτητων μεταβλητών.

Σε μια πρώτη προσέγγιση η λύση είναι να προσαρμόσει κάποιος όλα τα δυνατά μοντέλα, δηλαδή για όλους του συνδυασμούς υποσυνόλων των ανεξάρτητων μεταβλητών, και να βρει αυτό που προσαρμόζεται καλύτερα. Το κριτήριο προσαρμογής θα πρέπει να δίνει κάποια μορφή ποινής σε πιο πολύπλοκα μοντέλα (με περισσότερες ανεξάρτητες μεταβλητές) όπως ο προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού adjR^2. Αυτή η απλοϊκή προσέγγιση μπορεί να είναι αρκετά επίπονη όταν το πλήθος των ανεξάρτητων μεταβλητών είναι μεγάλο και συνήθως δεν ακολουθείται.

Έχουν προταθεί διάφορες μέθοδοι επιλογής των ανεξάρτητων μεταβλητών που υπολογίζουν το βέλτιστο μοντέλο πολλαπλής παλινδρόμησης βηματικά (stepwise regression). Όλες αυτές οι μέθοδοι εφαρμόζουν διαδοχικούς ελέγχους για το αν κάποια ανεξάρτητη μεταβλητή x_j είναι σημαντική, δηλαδή H_0: $\beta_j = 0$. Για παράδειγμα η **μέθοδος απαλοιφής προς τα πίσω** (backward elimination) αρχίζει με το μοντέλο να περιέχει όλες τις ανεξάρτητες μεταβλητές $x_1, x_2, \ldots, x_k$ και με διαδοχικούς ελέγχους απαλοίφει κάθε φορά μια ανεξάρτητη μεταβλητή όταν φαίνεται να 'περισσεύει', δηλαδή να μην εξηγεί την y όταν παρουσιάζεται με άλλες ανεξάρτητες μεταβλητές στο μοντέλο. Οι στατιστικοί έλεγχοι μπορούν επίσης να γίνουν με αντίστροφη πορεία, ξεκινώντας από το απλό μοντέλο σταθερού όρου και προσθέτοντας κάθε φορά μια ανεξάρτητη μεταβλητή στο μοντέλο που φαίνεται να είναι η πιο σημαντική για να εξηγήσει τη y όταν ήδη υπάρχουν κάποιες άλλες στο μοντέλο. Αυτή είναι η μέθοδος της **επιλογής προς τα μπρος** (forward selection).

Ένα πρόβλημα που είναι συνδεδεμένο με την επιλογή μεταβλητών είναι η **πολλαπλή συγγραμικότητα** (multicollinearity), δηλαδή δύο ή περισσότερες από τις k ανεξάρτητες μεταβλητές του μοντέλου, να είναι κατά όνομα ανεξάρτητες αλλά να είναι ισχυρά αλληλοεξαρτημένες. Αυτό μπορεί κάποιος να το διαπιστώσει προσαρμόζοντας ένα μοντέλο παλινδρόμησης σε κάθε μια από τις ύποπτες για αλληλοεξάρτηση μεταβλητές x_j ως προς όλες τις υπόλοιπες. Αν η x_j μπορεί να προβλεφθεί καλά από τις υπόλοιπες $k-1$ ανεξάρτητες μεταβλητές (μεγάλο R^2 ή adjR^2), τότε τα δεδομένα έχουν το πρόβλημα της πολλαπλής συγγραμικότητας. Πολλές φορές το πρόβλημα της συγγραμικότητας παραβλέπεται σε προβλήματα παλινδρόμησης. Γενικά δεν υπάρχει συγκεκριμένη μέθοδος αντιμετώπισης της πολλαπλής συγγραμικότητας.

5.4.3 Άλλα μη-γραμμικά μοντέλα

Τα μοντέλα προσθετικής (πολλαπλής) παλινδρόμησης που εξετάσαμε έχουν ως κοινό σημείο ότι είναι γραμμικά ως προς τις παραμέτρους τους.

Σε κάποιο πρόβλημα όπου οι παράμετροι φαίνονται να εμπλέκονται μη-γραμμικά, είναι ενδεχομένως δυνατόν με κατάλληλους μετασχηματισμούς να φέρουμε το μοντέλο σε προσθετική μορφή. Τέτοια μοντέλα αποτελούν τη γενίκευση της εγγενής γραμμική συνάρτησης παλινδρόμησης σε περισσότερες ανεξάρτητες μεταβλητές. Σε άλλες περιπτώσεις η μορφή του μοντέλου δεν απλοποιείται και τότε πρέπει να υπολογιστούν οι παράμετροι με κάποια μέθοδο μη-γραμμικής βελτιστοποίησης.

Υπάρχουν και άλλες κλάσεις μοντέλων που δεν έχουν κάποια γνωστή αναλυτική μορφή αλλά δίνονται ως άθροισμα διαφορετικών βασικών συναρτήσεων, όπως τα **νευρωνικά δίκτυα** (neural networks). Τέλος υπάρχουν και μη παραμετρικά μοντέλα που κάνουν εκτίμηση ή πρόβλεψη για τις δεδομένες τιμές των ανεξάρτητων μεταβλητών χρησιμοποιώντας από τα υπάρχοντα δεδομένα αυτά που είναι 'γειτονικά'. Τέτοια μοντέλα είναι τα μοντέλα **πυρήνων** (kernels).

Ασκήσεις Κεφαλαίου 5

1. Δημιουργείστε $M = 1000$ δείγματα μεγέθους $n = 20$ ζευγαρωτών παρατηρήσεων των (X, Y) από διμεταβλητή κανονική κατανομή με μέσες τιμές $\mu_X = 0$, $\mu_Y = 0$, τυπικές αποκλίσεις $\sigma_X = 1$, $\sigma_Y = 1$ και για δύο τιμές του συντελεστή συσχέτισης, $\rho = 0$ και $\rho = 0.5$.

   (α΄) Υπολογίστε το παραμετρικό 95% διάστημα εμπιστοσύνης κάνοντας χρήση του μετασχηματισμού Fisher για κάθε ένα από τα M δείγματα. Κάνετε το ιστόγραμμα για κάθε ένα από τα δύο άκρα του διαστήματος εμπιστοσύνης. Σε τι ποσοστό το διάστημα εμπιστοσύνης περιέχει τη πραγματική τιμή του ρ; Κάνετε τους υπολογισμούς ξεχωριστά για $\rho = 0$ και $\rho = 0.5$.

   (β΄) Κάνετε έλεγχο της υπόθεσης για μηδενική συσχέτιση των X και Y χρησιμοποιώντας το στατιστικό της κατανομής Student t της σχέσης (5.5) για κάθε ένα από τα M δείγματα. Σε τι ποσοστό απορρίπτεται η μηδενική υπόθεση; Κάνετε τους υπολογισμούς ξεχωριστά για $\rho = 0$ και $\rho = 0.5$.

   (γ΄) Επαναλάβεται τους παραπάνω υπολογισμούς για δείγματα μεγέθους $n = 200$. Υπάρχει διαφορά στα αποτελέσματα του διαστήματος εμπιστοσύνης και στατιστικής υπόθεσης;

   (δ΄) Επαναλάβεται τους παραπάνω υπολογισμούς για δείγματα μεγέθους $n = 20$ και $n = 200$ αλλά παίρνοντας τα τετράγωνα των παρατηρήσεων, δηλαδή θεωρείστε τις τ.μ. X^2 και Y^2. Υπάρχει διαφορά στα αποτελέσματα του διαστήματος εμπιστοσύνης και στατιστικής υπόθεσης από τα αντίστοιχα για τις τ.μ. X και Y;

2. Μελετήσαμε τον παραμετρικό έλεγχο για ανεξαρτησία (ή καλύτερα μηδενική συσχέτιση) δύο τ.μ. X και Y κάνοντας χρήση του στατιστικού t της σχέσης (5.5) και θεωρώντας ότι ακολουθεί κατανομή Student. Μπορούμε να κάνουμε τον έλεγχο χωρίς να θεωρήσουμε γνωστή κατανομή του στατιστικού κάτω από τη μηδενική υπόθεση και αυτός λέγεται μη-παραμετρικός έλεγχος. Θα χρησιμοποιήσουμε έναν τέτοιο έλεγχο που λέγεται έλεγχος τυχαιοποίησης και θα δημιουργήσουμε L τυχαιοποιημένα δείγματα από το αρχικό μας διμεταβλητό δείγμα των X και Y σύμφωνα με την μηδενική υπόθεση. Για αυτό θα αλλάξουμε τυχαία τη σειρά όλων των παρατηρήσεων της μιας από τις δύο τ.μ. στο δείγμα, και θα το κάνουμε αυτό L φορές. Στη συνέχεια θα υπολογίσουμε το στατιστικό t της σχέσης (5.5) στο αρχικό δείγμα, έστω t_0, αλλά και στα L τυχαιοποιημένα δείγματα, έστω $t_1, \ldots, t_L$. Η μηδενική υπόθεση απορρίπτεται αν η τιμή t_0 δεν περιέχεται στην κατανομή των $t_1, \ldots, t_L$,

δηλαδή στην εμπειρική (μη-παραμετρική) κατανομή του t κάτω από τη μηδενική υπόθεση της ανεξαρτησίας των X και Y. Συγκεκριμένα για επίπεδο σημαντικότητας a θα εξετάσουμε αν το t_0 είναι μεταξύ των $a/2\%$ και $(1-a/2)\%$ ποσοστιαίων σημείων, δηλαδή μεταξύ των σημείων με σειρά $La/2$ και $L(1-a/2)$ (προσεγγιστικά στον πλησιέστερο ακέραιο), όταν βάζουμε τα $t_1, \ldots, t_L$ σε αύξουσα σειρά.

Θεωρείστε $L = 1000$ τυχαιοποιημένα δείγματα και επαναλάβετε τον έλεγχο, αλλά τώρα τυχαιοποιημένο αντί για παραμετρικό, για την άσκηση 1, για την περίπτωση $n = 20$, και για τα ζευγάρια (X, Y) και (X^2, Y^2). Υπάρχουν διαφορές στα αποτελέσματα;

Βοήθεια (`matlab`*):* Για τη δημιουργία ενός τυχαιοποιημένου δείγματος αρκεί να αλλάξετε τυχαία τη σειρά των παρατηρήσεων της μιας τ.μ.. Μπορείτε να δημιουργείστε μια τυχαιοποιημένη σειρά δεικτών $1, \ldots, n$ με την συνάρτηση `randperm` και όρισμα n. Έτσι αν x είναι το διάνυσμα των n παρατηρήσεων της X και y το αντίστοιχο της Y, τότε αν r = randperm(n) είναι το διάνυσμα τυχαίων δεικτών, θέτοντας xr = x(r) δημιουργούμε το τυχαιοποιημένο δείγμα των xr και y.

3. Δίνονται οι μέσες μηνιαίες τιμές θερμοκρασίας και βροχόπτωσης στη Θεσσαλονίκη για την περίοδο 1959 - 1997. Ελέγξετε αν υπάρχει συσχέτιση μεταξύ της θερμοκρασίας και της βροχόπτωσης για κάθε μήνα ξεχωριστά. Χρησιμοποιείστε το στατιστικό t της σχέσης (5.5) και κάνετε παραμετρικό και έλεγχο τυχαιοποίησης (σύμφωνα με την άσκηση 2). Τα δεδομένα δίνονται στην ιστοσελίδα του μαθήματος σε δύο αρχεία πινάκων, ένας για τη θερμοκρασία και ένας για τη βροχόπτωση, που έχουν 39 γραμμές για τα 39 έτη και 12 στήλες για τους 12 μήνες κάθε έτους, από Ιανουάριο ως Δεκέμβριο.

4. Στο αρχείο `lightair.dat` στην ιστοσελίδα του μαθήματος δίνονται 100 μετρήσεις της πυκνότητας του αέρα (σε $\mathrm{kg/m^3}$) σε διαφορετικές συνθήκες θερμοκρασίας και πίεσης (στην πρώτη στήλη) και οι αντίστοιχες μετρήσεις της ταχύτητας φωτός (-299000 km/sec) στη δεύτερη στήλη.

   (α′) Σχεδιάστε το κατάλληλο διάγραμμα διασποράς και υπολογίστε τον αντίστοιχο συντελεστή συσχέτισης.

   (β′) Εκτιμείστε το μοντέλο γραμμικής παλινδρόμησης με τη μέθοδο ελαχίστων τετραγώνων για τη γραμμική εξάρτηση της ταχύτητας φωτός από την πυκνότητα του αέρα. Υπολογίστε παραμετρικό διάστημα εμπιστοσύνης σε επίπεδο 95% για τους δύο συντελεστές της ευθείας ελαχίστων τετραγώνων (διαφορά ύψους β_0 και κλίση β_1).

(γ′) Σχηματίστε στο διάγραμμα διασποράς την ευθεία ελαχίστων τετραγώνων, τα όρια πρόβλεψης σε επίπεδο 95% για τη μέση ταχύτητα φωτός καθώς και για μια τιμή της ταχύτητας φωτός. Επίσης κάνετε πρόβλεψη για πυκνότητα αέρα $1.29\ \mathrm{kg/m^3}$ δίνοντας και τα όρια μέσης τιμής και μιας παρατήρησης της ταχύτητας φωτός.

(δ′) Η πραγματική σχέση της ταχύτητας φωτός στον αέρα με την πυκνότητα του αέρα είναι:

$$c_{air} = c\left(1 - 0.00029\frac{d}{d_0}\right),$$

όπου οι δύο σταθερές είναι:

- $c = 299792.458\ \mathrm{km/sec}$, η ταχύτητα φωτός στο κενό, και
- $d_0 = 1.29\ \mathrm{kg/m^3}$, η πυκνότητα του αέρα σε θερμοκρασία και πίεση δωματίου.

Από την παραπάνω σχέση υπολογίστε την εξίσωση της πραγματικής ευθείας παλινδρόμησης της ταχύτητας φωτός στον αέρα ως προς την πυκνότητα του αέρα. Στη συνέχεια κάνετε έλεγχο ξεχωριστά για κάθε συντελεστή της πραγματικής ευθείας παλινδρόμησης αν τον δεχόμαστε με βάση το δείγμα των 100 ζευγαρωτών παρατηρήσεων (σύμφωνα με τις εκτιμήσεις τους στο 4β′). Είναι οι πραγματικές μέσες τιμές της ταχύτητας φωτός μέσα στα όρια μέσης πρόβλεψης για κάθε τιμή πυκνότητας αέρα στο δείγμα; (σύμφωνα με το διάστημα μέσης πρόβλεψης που υπολογίσατε και σχηματίσατε στο 4γ′).

5. Θα υπολογίσουμε μη παραμετρικό διάστημα εμπιστοσύνης για τους δύο συντελεστές της ευθείας ελαχίστων τετραγώνων (διαφορά ύψους β_0 και κλίση β_1) και θα το εφαρμόσουμε στα δεδομένα της προηγούμενης άσκησης (έξάρτηση της ταχύτητας φωτός από την πυκνότητα του αέρα). Η μέθοδος που θα χρησιμοποιήσουμε λέγεται bootstrap και για τον υπολογισμό των διαστημάτων εμπιστοσύνης των β_0 και β_1 ορίζεται ως εξής:

   - Πάρε ένα νέο δείγμα 100 τυχαίων ζευγαρωτών παρατηρήσεων από το δείγμα των 100 ζευγαρωτών παρατηρήσεων (πυκνότητας αέρα και ταχύτητας φωτός). Το κάθε ζευγάρι επιλέγεται τυχαία με επανάθεση, δηλαδή το ίδιο ζευγάρι μπορεί να εμφανιστεί πολλές φορές στο νέο δείγμα (και άλλο ζευγάρι να μην εμφανιστεί καθόλου).

- Υπολόγισε τις εκτιμήσεις b_0 και b_1 των συντελεστών της ευθείας ελαχίστων τετραγώνων για αυτό το νέο δείγμα.
- Επανέλαβε τα παραπάνω δύο βήματα $M = 1000$ φορές.
- Από τις M τιμές για το b_0 υπολόγισε τα όρια του $(1 - a)\%$ (εδώ $a = 0.05$) διαστήματος εμπιστοσύνης για το β_0 από τα ποσοστιαία σημεία $a/2\%$ και $(1-a/2)\%$. Για αυτό βάλε τις $M = 1000$ τιμές σε αύξουσα σειρά και βρες τις τιμές για τάξη $Ma/2\%$ και $M(1-a/2)\%$. Κάνε το ίδιο για το b_1.

Σύγκρινε τα bootstrap διαστήματα εμπιστοσύνης των β_0 και β_1 με τα παραμετρικά που βρήκες στην προηγούμενη άσκηση.

Βοήθεια (`matlab`*):* Για τη δημιουργία n τυχαίων αριθμών από 1 ως N με επανάθεση, χρησιμοποίησε τη συνάρτηση `unidrnd` με κατάλληλα ορίσματα ως `unidrnd(`N`,`n`,1)`. Θα δώσει το διάνυσμα δεικτών για το νέο δείγμα από τα n στοιχεία του αρχικού δείγματος (εδώ τα στοιχεία είναι οι ζευγαρωτές παρατηρήσεις).

6. Στον παρακάτω πίνακα δίνονται τα δεδομένα για το ποσοστό υψηλής επίδοσης που ακόμα έχουν ελαστικά (με ακτινωτή ενίσχυση) ενώ έχουν ήδη χρησιμοποιηθεί για τα αντίστοιχα χιλιόμετρα.

Α/Α	*Απόσταση* σε χιλιάδες km	*ποσοστό δυνατότητας* χρήσης
1	2	98.2
2	3	91.7
3	8	81.3
4	16	64.0
5	32	36.4
6	48	32.6
7	64	17.1
8	80	11.3

(α′) Κάνε το διάγραμμα διασποράς και προσάρμοσε το κατάλληλο μοντέλο για τον προσδιορισμό του ποσοστού δυνατότητας χρήσης (υψηλής επίδοσης) προς τα αντίστοιχα χιλιόμετρα χρήσης του ελαστικού. Για να ελέγξεις την καταλληλότητα του μοντέλου, κάνε διαγνωστικό διάγραμμα διασποράς των τυποποιημένων σφαλμάτων προσαρμογής του επιλεγμένου μοντέλου προς την εξαρτημένη μεταβλητή (ποσοστό δυνατότητας χρήσης).

(β′) Πρόβλεψε το ποσοστό δυνατότητας χρήσης για ελαστικό που χρησιμοποιήθηκε για 25000km.

7. Ο θερμοστάτης είναι αντιστάτης με αντίσταση που εξαρτιέται από τη θερμοκρασία. Είναι φτιαγμένος (συνήθως) από ημι-αγωγό υλικό με ενεργειακό διάκενο E_g. Η αντίσταση R του θερμοστάτη αλλάζει σύμφωνα με τη σχέση

$$R \propto R_0 e^{E_g/2kT},$$

όπου T είναι η θερμοκρασία (σε ^{o}K) και R_0, k είναι σταθερές. Για κατάλληλες παραμέτρους β_0 και β_1 ($\beta_1 = 2k/E_g$) η παραπάνω εξίσωση μπορεί να απλοποιηθεί στη γραμμική μορφή

$$\frac{1}{T} = \beta_0 + \beta_1 \ln(R).$$

Το ενεργειακό διάκενο E_g έχει κάποια μικρή εξάρτηση από τη θερμοκρασία έτσι ώστε η παραπάνω έκφραση να μην είναι ακριβής. Διορθώσεις μπορούν να γίνουν προσθέτοντας πολυωνυμικούς όρους του $\ln(R)$.

Στον παρακάτω πίνακα δίνονται 32 μετρήσεις της αντίστασης R (σε Ωμ) και της θερμοκρασίας σε ^{o}C (θα πρέπει να μετατραπούν σε ^{o}K, δηλαδή να προστεθεί σε κάθε τιμή 273.15).

(α′) Βρείτε το κατάλληλο πολυωνυμικό μοντέλο της παλινδρόμησης του $1/T$ ως προς $\ln(R)$, κάνοντας διαγνωστικό έλεγχο με το διάγραμμα διασποράς των τυποποιημένων υπολοιπών προς $1/T$ για κάθε μοντέλο που δοκιμάζετε (πρώτου βαθμού, δευτέρου βαθμού κτλ).

(β′) Συγκρίνετε την προσαρμογή και κατλληλότητα του μοντέλου που καταλήξατε με το μοντέλο του Steinhart-Hart που δίνεται από την εξίσωση

$$\frac{1}{T} = \beta_0 + \beta_1 \ln(R) + \beta_3 (\ln(R))^3.$$

Α/Α	*Αντίσταση*	*θερμοκρασία (σε ^oC)*
1	0.76	110
2	0.86	105
3	0.97	100
4	1.11	95
5	1.45	85
6	1.67	80
7	1.92	75
8	2.23	70
9	2.59	65
10	3.02	60
11	3.54	55
12	4.16	50
13	4.91	45
14	5.83	40
15	6.94	35
16	8.31	30
17	10.00	25
18	12.09	20
19	14.68	15
20	17.96	10
21	22.05	5
22	27.28	0
23	33.89	-5
24	42.45	-10
25	53.39	-15
26	67.74	-20
27	86.39	-25
28	111.30	-30
29	144.00	-35
30	188.40	-40
31	247.50	-45
32	329.20	-50

8. Μετρήθηκε το βάρος και 10 δείκτες σώματος σε 22 άνδρες νεαρής ηλικίας. Τα δεδομένα δίνονται στην ιστοσελίδα του μαθήματος στο αρχείο `physical.txt`. Η πρώτη γραμμή του πίνακα του αρχείου έχει τα ονόματα των δεικτών σε κάθε στήλη δεδομένων και δίνονται στον παρακάτω πίνακα.

Α/Α	*Όνομα*	*Περιγραφή*
1	Mass	Βάρος σε κιλά
2	Fore	μέγιστη περιφέρεια του πήχυ χεριού
3	Bicep	μέγιστη περιφέρεια του δικέφαλου μυ
4	Chest	περιμετρική απόσταση στήθους (στο ύψος κάτω από τις μασχάλες)
5	Neck	περιμετρική απόσταση λαιμού (στο μέσο ύψος λαιμού)
6	Shoulder	περιμετρική απόσταση ώμου
7	Waist	περιμετρική απόσταση μέσης (οσφίου)
8	Height	ύψος από την κορυφή στα δάχτυλα ποδιού
9	Calf	μέγιστη περιφέρεια κνήμης
10	Thigh	περιμετρική απόσταση γοφού
11	Head	περιμετρική απόσταση κεφαλιού

Διερευνείστε το κατάλληλο μοντέλο γραμμικής παλινδρόμησης για το βάρος. Δοκιμάστε το μοντέλο με τις 10 ανεξάρτητες μεταβλητές και συγκρίνετε το με το μοντέλο που δίνει κάποια μέθοδος βηματικής παλινδρόμησης. Υπολογίστε για το κάθε μοντέλο τις εκτιμήσεις των παραμέτρων, τη διασπορά των σφαλμάτων και το συντελεστή προσδιορισμού (καθώς και τον προσαρμοσμένο συντελεστή προσδιορισμού).

Βοήθεια (`matlab`*):* Για να εφαρμόσετε βηματική παλινδρόμηση, το `matlab` παρέχει γραφικό περιβάλλον με την εντολή `stepwise` και κάνει τους ίδιους υπολογισμούς στη συνάρτηση `stepwisefit`. Για να φορτώσετε τα δεδομένα του αρχείου με την εντολή `load` θα πρέπει πρώτα να διαγράψετε την πρώτη σειρά με τα ονόματα των μεταβλητών.

9. Μετρήθηκαν σε 12 νοσοκομεία των ΗΠΑ οι μηνιαίες ανθρωπο-ώρες που σχετίζονται με την υπηρεσία αναισθησιολογίας, καθώς και άλλοι δείκτες που ενδεχομένως επηρεάζουν την απασχόληση προσωπικού στην υπηρεσία αναισθησιολογίας. Τα δεδομένα δίνονται στην ιστοσελίδα του μαθήματος στο αρχείο `hospital.txt`. Η πρώτη γραμμή του πίνακα του αρχείου έχει τα ονόματα των δεικτών σε κάθε στήλη δεδομένων και δίνονται στον παρακάτω πίνακα.

Α/Α	*Όνομα*	*Περιγραφή*
1	ManHours	οι ανθρωπο-ώρες στην υπηρεσία αναισθησιολογίας
2	Cases	τα περιστατικά χειρουργίου μηνιαία
3	Eligible	ο πληθυσμός που εξυπηρετείται ανά χιλιάδες
4	OpRooms	οι αίθουσες χειρουργείου

Διερευνείστε το κατάλληλο μοντέλο γραμμικής παλινδρόμησης για τις ανθρωπο-ώρες. Δοκιμάστε το μοντέλο με τις 3 ανεξάρτητες μεταβλητές και συγκρίνετε το με το μοντέλο που δίνει κάποια μέθοδος βηματικής παλινδρόμησης. Υπολογίστε για το κάθε μοντέλο τις εκτιμήσεις των παραμέτρων, τη διασπορά των σφαλμάτων και το συντελεστή προσδιορισμού (καθώς και τον προσαρμοσμένο συντελεστή προσδιορισμού). Επίσης διερευνείστε το φαινόμενο πολλαπλής συγγραμικότητας για τους 3 δείκτες.