ΚΕΦΑΛΑΙΟ 7

ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ
(Non Parametric Regression)

Το κεφάλαιο αυτό συνδέεται άμεσα με το κεφάλαιο που αναφέρεται στην συσχέτιση τάξης μεγέθους με την έννοια υπό την οποία η κλασική παραμετρική παλινδρόμηση συνδέεται με τον συντελεστή συσχέτισης του Pearson που βασίζεται σε αυτές καθεαυτές τις τιμές των παρατηρήσεων του δείγματος.

Ας υποθέσουμε ότι έχουμε ένα τυχαίο δείγμα $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ n παρατηρήσεων πάνω στο τυχαίο διάνυσμα (X, Y). Οι μέθοδοι συσχέτισης δίνουν έμφαση στην εκτίμηση του βαθμού της εξάρτησης που υπάρχει μεταξύ X και Y. Οι μέθοδοι παλινδρόμησης δίνουν την δυνατότητα μιας βαθύτερης εξέτασης της μορφής της σχέσης που υφίσταται μεταξύ X και Y. Οπως είναι γνωστό, ένας σημαντικός στόχος των μεθόδων παλινδρόμησης είναι η πρόβλεψη μιας τιμής της μεταβλητής Y του ζεύγους (X, Y), όταν μόνο η τιμή της τυχαίας μεταβλητής X είναι γνωστή, με βάση πληροφορίες οι οποίες μπορούν να ληφθούν από προηγούμενες παρατηρήσεις $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$. Αν, για παράδειγμα, Y συμβολίζει τον βαθμό του πτυχίου ενός αποφοίτου Λυκείου και X τον αριθμό των μορίων τα οποία είχε πάρει στις Πανελλήνιες Εξετάσεις, παρατηρήσεις πάνω στις μεταβλητές αυτές για παλαιούς φοιτητές μπορούν να βοηθήσουν στο να προβλέψουμε με τι βαθμό θα αποφοιτήσει από το Πανεπιστήμιο ένας φοιτητής που πήρε $X=x_0$ μόρια στις Πανελλήνιες Εξετάσεις. Βέβαια, η μεταβλητή Y είναι μια τυχαία μεταβλητή, και επομένως δεν μπορούμε να περιμένουμε να προσδιορίσουμε την τιμή της βασιζόμενοι μόνο στην γνώση της

αντίστοιχης τιμής της μεταβλητής X. Αν όμως γνωρίζουμε την τιμή της X, μπορούμε ίσως να κάνουμε μια καλύτερη πρόβλεψη για την τιμή της Y. Μία εύλογη μέθοδος, που θα μπορούσε να χρησιμοποιήσει κάποιος για να προβλέψει την τιμή της μεταβλητής Y όταν η μεταβλητή X έχει την τιμή x, θα ήταν να χρησιμοποιήσει τον μέσο ή την διάμεσο ενός δείγματος αρκετών τιμών της Y που παρατηρήθηκαν στο παρελθόν όταν η X είχε την τιμή x. Μία εύλογη μέθοδος, δηλαδή, για την πρόβλεψη της τιμής της τυχαίας μεταβλητής Y όταν η X έχει την τιμή x βασίζεται στην εκτίμηση της E(Y|X=x) με βάση κάποια δεδομένα του παρελθόντος. Με τον τρόπο αυτό, μπορούμε να έχουμε σημειακές εκτιμήσεις αλλά και διαστήματα εμπιστοσύνης για την E(Y|X=x).

Η συνάρτηση E(Y|X=x) ονομάζεται *παλινδρόμηση της τυχαίας μεταβλητής* Y *πάνω στην τυχαία μεταβλητή* X. Η εξίσωση m(x) = E(Y|X=x) ονομάζεται *εξίσωση παλινδρόμησης*.

Η μορφή της εξίσωσης αυτής είναι κεντρικής σημασίας για το πρόβλημα της πρόβλεψης, γιατί αυτή καθορίζει την σχέση μεταξύ E(Y|X=x) και x. Συχνά, γίνεται η υπόθεση ότι η παλινδρόμηση είναι της μορφής E(Y|X=x) = α+βx, για κάποιες σταθερές α και β. Στην περίπτωση αυτή, λέμε ότι η *παλινδρόμηση της* Y *πάνω στην* X *είναι γραμμική*.

Στην συνέχεια, θα εξετάσουμε την περίπτωση της γραμμικής παλινδρόμησης, όπως επίσης και μια γενικότερη περίπτωση όπου η E(Y|X=x) είναι μία μονότονη (είτε αύξουσα είται φθίνουσα) συνάρτηση της τιμής x της X.

7.1 ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Συχνά, σε σχέση με πρακτικές εφαρμογές, οι σταθερές α και β είναι άγνωστες και απαιτείται να εκτιμηθούν από τα δεδομένα. Η κλασική μέθοδος για την προσαρμογή μιας ευθείας γραμμής σε διμεταβλητά δεδομένα είναι η *μέθοδος των ελαχίστων τετραγώνων (least squares method)* που, όπως είναι γνωστό (βλέπε π.χ. Ι. Πανάρετου *«Γραμμικά Μοντέλα με έμφαση στις Εφαρμογές»*, Αθήνα 1994), έχει βέλτιστες ιδιότητες κάτω από ορισμένες συνθήκες ανεξαρτησίας και κανονικότητας, οι οποίες επιτρέπουν την χρησιμοποίηση παραμετρικών τεχνικών εκτίμησης και ελέγχου υποθέσεων.

Παρά το ότι η μέθοδος των ελαχίστων τετραγώνων ενδέχεται να έχει σχετικά μικρή ευαισθησία σε ορισμένες αποκλίσεις από τις βασικές προϋποθέσεις, μπορεί να επηρεάζεται σημαντικά από άλλες. Για τον λόγο αυτό, τα τελευταία 30 χρόνια έχουν αναπτυχθεί πολλοί διαγνωστικοί έλεγχοι για τον εντοπισμό παρατηρήσεων που ενδέχεται να προκαλέσουν δυσκολίες στην εφαρμογή των παραμετρικών μεθόδων παλινδρόμησης ή παρατηρήσεων που ενδέχεται να επηρεάσουν τα αποτελέσματα, με την έννοια ότι σχετικά μικρές μεταβολές σ' αυτές μπορούν να έχουν εμφανή επίδραση στις εκτιμήσεις. Εχουν, επίσης, αναπτυχθεί *ευσταθείς (εύρωστες, robust)* μέθοδοι παλινδρόμησης, οι οποίες επηρεάζονται πολύ λίγο από *παρεκκλίνουσες* παρατηρήσεις, ενώ, ταυτόχρονα, συμπεριφέρονται σχεδόν το ίδιο καλά με την μέθοδο των ελαχίστων τετραγώνων. Έτσι, τις περισσότερες φορές, ο ερευνητής χρησιμοποιεί την μέθοδο των ελαχίστων τετραγώνων παράλληλα με μία ευσταθή μέθοδο. Αν οι δύο

μέθοδοι οδηγούν σε αποτελέσματα που αποκλίνουν, ο ερευνητής θα πρέπει να διερευνήσει τους λόγους.

Εύλογα, θα μπορούσε να θεωρήσει κανείς ότι είναι προτιμότερο να χρησιμοποιούμε πάντα ευσταθείς μεθόδους και απλώς να δεχόμαστε τα αποτελέσματά τους. Ο λόγος που κάτι τέτοιο δεν θα έπρεπε να γίνεται είναι ότι η προσαρμογή μιας ευθείας γραμμής είναι, ίσως, λανθασμένη επιλογή για τα δεδομένα, κάτι που ακόμη και μία ευσταθής μέθοδος ενδέχεται να μην αποκαλύψει χωρίς την εφαρμογή πρόσθετων διαγνωστικών ελέγχων. Οι μη παραμετρικές μέθοδοι ελέγχου υποθέσεων και εκτίμησης που εξετάζονται στην συνέχεια, ως βασιζόμενες στις τάξεις μεγέθους των παρατηρήσεων και όχι σε αυτές καθεαυτές τις τιμές τους, παρακάμπτουν προβλήματα που αναφύονται από την παραβίαση των βασικών προϋποθέσεων που απαιτούν τα παραμετρικά ανάλογά τους.

7.1.1 Η Μέθοδος των Ελαχίστων Τετραγώνων

Οπως ήδη αναφέρθηκε, μία ευρύτατα διαδεδομένη μέθοδος για την εκτίμηση των σταθερών α και β είναι η μέθοδος των ελαχίστων τετραγώνων. Η μέθοδος αυτή επιλέγει εκτιμήτριες $\hat{\alpha}$ και $\hat{\beta}$ των άγνωστων σταθερών α και β της συνάρτησης παλινδρόμησης $E(Y|X=x) = \alpha+\beta x$, οι οποίες οδηγούν σε εκτιμήσεις a και b που ελαχιστοποιούν το άθροισμα των τετραγωνικών αποκλίσεων των παρατηρούμενων τιμών y_i της Y από τις τιμές $a + bx_i$ που εκτιμάται ότι θα έπρεπε να παρατηρούνται για την Y (αφού, όπως έχουμε υποθέσει $E(Y|X=x) = \alpha+\beta x$). Συγκεκριμένα, η μέθοδος των ελαχίστων τετραγώνων οδηγεί στις εκτιμήτριες

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}}$$

$$\hat{\alpha} = \overline{Y} - \frac{S_{xY}}{S_{xx}}\overline{x},$$

όπου

$$S_{xY} = \sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(Y_i - \overline{Y}\right)$$

$$= \sum_{i=1}^{n} \left(x_i - \overline{x}\right) Y_i,$$

$$S_{xx} = \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2.$$

Επειδή (x_1, y_1), ..., (x_n, y_n) είναι μία πραγματοποίηση (realization) της ακολουθίας των τυχαίων διανυσμάτων (X_1, Y_1), ..., (X_n, Y_n) που παριστάνει το τυχαίο δείγμα από την κατανομή του διανύσματος (X, Y), τότε και οι εκτιμήσεις a και b των αγνώστων σταθερών α και β, αντίστοιχα, αποτελούν πραγματοποιήσεις των στατιστικών συναρτήσεων $\hat{\alpha}$ και $\hat{\beta}$. Δηλαδή, οι τιμές a και b είναι οι τιμές των εκτιμητριών $\hat{\alpha}$ και $\hat{\beta}$ των παραμέτρων α και β που αντιστοιχούν στο συγκεκριμένο δείγμα (x_1, y_1), ..., (x_n, y_n). Ως εκ τούτου,

$$b = \sum_{i=1}^{n}\left(x_i - \overline{x}\right)y_i \Big/ \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

και

$$a = \overline{y} - b\overline{x}.$$

Επομένως, αν με $\hat{Y}$ συμβολίσουμε την εκτιμήτρια της συνάρτησης παλινδρόμησης $\hat{E} = (Y|X) = \hat{\alpha} + \hat{\beta}X$, δηλαδή $\hat{Y}_i = \hat{E}(Y|X = x_i)$, i = 1, 2, …, n, τότε $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$, i = 1, 2, …, n.

Η εκτίμηση, επομένως, της ευθείας παλινδρόμησης, για το συγκεκριμένο δείγμα $(x_1, y_1), ..., (x_n, y_n)$, είναι η

$$\hat{y}_i = a + bx_i, \ i = 1, 2, .., n.$$

Αν η εκτιμηθείσα ευθεία ήταν η πραγματική άγνωστη ευθεία, πάνω στην οποία βρίσκεται η δεσμευμένη μέση τιμή της Y, τότε τα κατάλοιπα $e_i = y_i - \hat{y}_i = y_i - a - bx$ θα αντιστοιχούσαν στις άγνωστες αποκλίσεις $\varepsilon_i = y_i - \alpha - \beta x_i$ των παρατηρούμενων τιμών y_i της Y από τις τιμές $\alpha+\beta x_i$ (i = 1, 2, …, n), τις οποίες θα έπρεπε να παρατηρούμε για την Y. Οι αποκλίσεις αυτές ονομάζονται *σφάλματα*. Δηλαδή, τα κατάλοιπα $e_i = y_i - \hat{y}_i$ (οι πραγματοποιήσεις της μεταβλητής $e = Y - \hat{Y} = Y - \hat{\alpha} - \hat{\beta}X \equiv \hat{\varepsilon}$) μπορούν, με κάποια έννοια, να θεωρηθούν ως εκτιμήσεις των πραγματοποιήσεων ε_i, i = 1, 2, …, n μιας μη παρατηρήσιμης τυχαίας μεταβλητής ε $= Y - \alpha - \beta X \equiv Y - E(Y|X)$. (Η μεταβλητή ε που εκπροσωπεί ένα τυχαίο λάθος κατανέμεται ανεξάρτητα από την μεταβλητή X με $E(\varepsilon) = 0$).

Είναι φανερό ότι, αν αληθεύει η μορφή που υποθέσαμε για την παλινδρόμηση της Y πάνω στην X, δηλαδή, αν αληθεύει η υπόθεση ότι $E(Y|X=x) = \alpha+\beta x$, τότε η τυχαία μεταβλητή ε είναι ανεξάρτητη από την τυχαία μεταβλητή X. Το σημείο αυτό είναι κεντρικής σημασίας για οποιαδήποτε μορφή συμπερασματολογίας επιθυμούμε να κάνουμε. Πράγματι, ας υποθέσουμε ότι ενδιαφερόμαστε να

ελέγξουμε ένα από τα εξής τρία ζεύγη υποθέσεων για την κλίση β της ευθείας παλινδρόμησης:

Α. (Αμφίπλευρος έλεγχος)

H_0: $\beta=\beta_0$

H_1: $\beta\neq\beta_0$

Β. (Μονόπλευρος έλεγχος)

H_0: $\beta=\beta_0$

H_1: $\beta>\beta_0$

Γ. (Μονόπλευρος έλεγχος)

H_0: $\beta=\beta_0$

H_1: $\beta<\beta_0$

Για κάθε ζεύγος (X_i, Y_i), i = 1, 2, .., n, ας θεωρήσουμε την τυχαία μεταβλητή $U_i^* \equiv \varepsilon_i = Y_i - \alpha - \beta X_i$. Κάτω από την μηδενική υπόθεση, οι μεταβλητές X_i και $U_i^* = Y_i - \alpha - \beta_0 X_i$ είναι ανεξάρτητες. Επομένως, μπορούμε να χρησιμοποιήσουμε τον συντελεστή συσχέτισης τάξης μεγέθους ρ του Spearman για τα ζεύγη (X_i, U_i^*), i = 1, 2, ..., n.

Ας σημειωθεί ότι, επειδή η τάξη μεγέθους της μεταβλητής $U_i^* \equiv \varepsilon_i = Y_i - \alpha - \beta_0 X_i$ ταυτίζεται με την τάξη μεγέθους της μεταβλητής $U_i = Y_i - \beta_0 X_i$, (i = 1, 2, ..., n), μπορούμε να ελέγξουμε την υπόθεση H_0: $\beta=\beta_0$ χωρίς να γνωρίζουμε την σταθερά α και, επομένως, μπορούμε να χρησιμοποιήσουμε ως ελεγχοσυνάρτηση τον

συντελεστή συσχέτισης τάξης μεγέθους ρ του Spearman για τα ζεύγη (X_i, U_i), $i = 1, 2, ..., n$.

Παρατήρηση: Οπως ακριβώς ο συντελεστής ρ του Spearman αποτελεί το *μη παραμετρικό ανάλογο* του συντελεστή r του Pearson (στην πραγματικότητα, είναι ο συντελεστής r του Pearson υπολογιζόμενος στις τάξεις μεγέθους των παρατηρήσεων), έτσι και ο έλεγχος αυτός είναι το μη παραμετρικό ανάλογο του συνήθους παραμετρικού ελέγχου της ίδιας υπόθεσης που βασίζεται στον υπολογισμό του συντελεστή ρ του Pearson για τα ζεύγη (X_i, U_i), $i = 1, 2, ..., n$. Ο τελευταίος, βέβαια, γίνεται με την επιπλέον υπόθεση ότι το τυχαίο διάνυσμα (X, Y) ακολουθεί την διμεταβλητή κανονική κατανομή.

Επομένως, ο κανόνας απόφασης έχει την μορφή:

Σε επίπεδο σημαντικότητας α, απορρίπτουμε την μηδενική υπόθεση H_0 στην περίπτωση Β, αν η τιμή του ρ είναι πολύ μεγάλη (δηλαδή, αν η τιμή του ρ υπερβαίνει το (1-α)-ποσοστιαίο σημείο της κατανομής του), θα απορρίπτουμε την μηδενική υπόθεση H_0 στην περίπτωση Γ, αν η τιμή του ρ είναι πολύ μικρή (δηλαδή, αν είναι μικρότερη από το α-ποσοστιαίο σημείο της κατανομής του) και θα απορρίπτουμε την μηδενική υπόθεση H_0 στην περίπτωση Α, αν η τιμή του ρ υπερβαίνει την τιμή του (1–α/2)-ποσοστιαίου σημείου της κατανομής του ή αν είναι μικρότερη από την τιμή του (α/2)-ποσοστιαίου σημείου της κατανομής του.

Παράδειγμα 7.1.1: Ας θεωρήσουμε τις μετρήσεις της επιθετικότητας των 12 ζευγαριών διδύμων που εξετάσαμε σε προηγούμενο παράδειγμα. Το μέτρο επιθετικότητας του πρωτότοκου συμβολίζεται

με X_i και του δευτερότοκου με Y_i. Οι 12 αυτές παρατηρήσεις ήταν (86,88), (71,77), (77,76), (68,64), (91,96), (72,72), (77,65), (91,90), (70,65), (71,80), (88,81) και (87,72).

Έχοντας ήδη απορρίψει την υπόθεση της ανεξαρτησίας μεταξύ της επιθετικότητας Χ του πρωτότοκου και την επιθετικότητας Υ του δευτερότοκου σε ένα ζεύγος διδύμων, έχει έννοια να προσπαθήσει κανείς να εκτιμήσει την γραμμική σχέση που υφίσταται μεταξύ των μεταβλητών Χ και Υ.

Από τα δεδομένα, προκύπτει ότι

$$\sum_{i=1}^{12} x_i = 949, \quad \bar{x} = 79.1, \quad \sum_{i=1}^{12} x_i^2 = 75919$$

$$\sum_{i=1}^{12} y_i = 926, \quad \bar{y} = 77.2, \quad \sum_{i=1}^{12} y_i^2 = 73976.$$

Επομένως, η ευθεία παλινδρόμησης ελαχίστων τετραγώνων είναι η

$$\hat{y} = 9.38 + 0.857x.$$

Μπορούμε να χρησιμοποιήσουμε την ευθεία παλινδρόμησης που εκτιμήσαμε ως μία περιγραφή της σχέσης μεταξύ Υ και Χ ή, ακριβέστερα, ως μία εκτίμηση της δεσμευμένης μέσης τιμής $E(Y|X=x)$ της Υ όταν δίνεται η τιμή x της Χ. Αν ο πρωτότοκος σε ένα επιπρόσθετο ζεύγος διδύμων έχει την βαθμολογία $x_0 = 80$, μπορούμε να προβλέψουμε την βαβμολογία του δευτερότοκου ως ίση περίπου με $9.38+(0.857)(80) = 77.9$.

Περιγράφοντας την σχέση που υπάρχει μεταξύ της επιθετικότητας των διδύμων, εύλογο είναι να υποθέσουμε ότι $\beta=1$,

αφού η εκτίμηση ελαχίστων τετραγώνων του συντελεστή β είναι b=0.857. Για να ελέγξουμε τις υποθέσεις

$$H_0: \beta=1$$

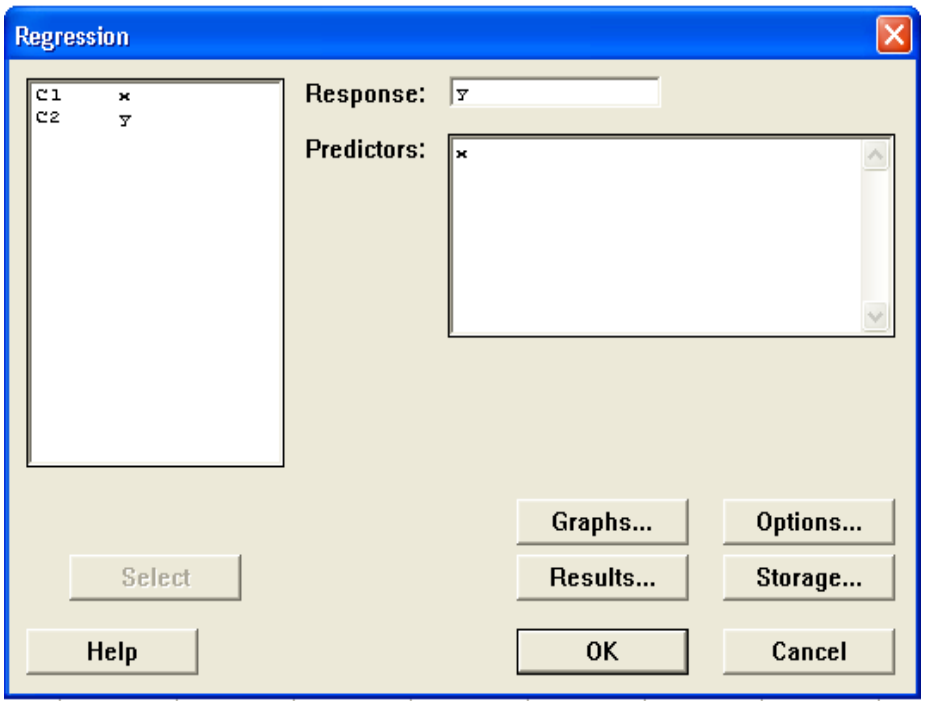
$$H_1: \beta\neq 1,$$

υπολογίζουμε τον συντελεστή συσχέτισης τάξης μεγέθους του Spearman μεταξύ των μεταβλητών X_i και $U_i = Y_i - (1)X_i = Y_i - X_i$, $i = 1, 2, \ldots, 12$ με βάση τον πίνακα που ακολουθεί.

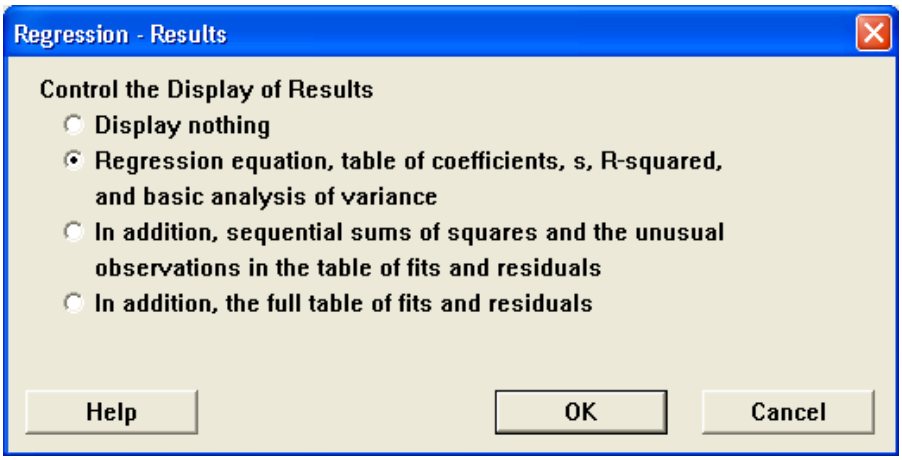
	Ζεύγος διδύμων											
	1	2	3	4	5	6	7	8	9	10	11	12
Πρωτότοκος X_I	86	71	77	68	91	72	77	91	70	71	88	87
ΔευτερότοκοςU_i	2	6	–1	–4	5	0	–12	–1	–5	9	–7	–15
$R(X_i)$	8	3.5	6.5	1	11.5	6.5	11.5	2	3.5	10	10	9
$R(U_i)$	9	11	6.5	5	10	2	6.5	4	12	3	3	1

Από τα στοιχεία του παραπάνω πίνακα, προκύπτει ότι η τιμή του συντελεστή ρ είναι –0.1232. Η τιμή αυτή δεν βρίσκεται στην κρίσιμη περιοχή μεγέθους α=0.05, η οποία, όπως προκύπτει από τον σχετικό πίνακα του παραρτήματος, ορίζεται από την ανισότητα $|\rho| > 0.5804$. Επομένως, η μηδενική υπόθεση δεν απορρίπτεται σε επίπεδο σημαντικότητας 0.05.

Λύση με το MINITAB: Για την εκτίμηση της ευθείας παλινδρόμησης της μεταβλητής Y πάνω στην μεταβλητή X, εργαζόμεθα ως εξής: Καταχωρίζουμε τα δύο δείγματα στις μεταβλητές **y** και **x** και επιλέγουμε **Stat, Regression** οδηγούμενοι στο εξής πλαίσιο διαλόγου:



Στο πεδίο **Response**, δηλώνουμε την εξαρτημένη μεταβλητή (**x**) και στο πεδίο **Predictors**, τις ανεξάρτητες μεταβλητές. (Εδώ είναι μία, δηλαδή η **x**). Κατόπιν, πιέζοντας **Results**, εμφανίζεται το εξής πλαίσιο διαλόγου:



Στο πλαίσιο αυτό, δηλώνουμε την δεύτερη επιλογή αφού μας ενδιαφέρουν μόνο οι τιμές των εκτιμητριών και τα τυπικά τους σφάλματα. Πιέζοντας **OK** και ξανά **OK** στο επανεμφανιζόμενο αρχικό πλαίσιο διαλόγου, οδηγούμεθα στον εξής πίνακα αποτελεσμάτων:

```
The regression equation is
y = 9,4 + 0,857 x

Predictor        Coef       StDev          T        P
Constant         9.38       19.92       0.47    0.648
x              0.8572      0.2505       3.42    0.007

S = 7.384       R-Sq = 53.9%     R-Sq(adj) = 49.3%
```

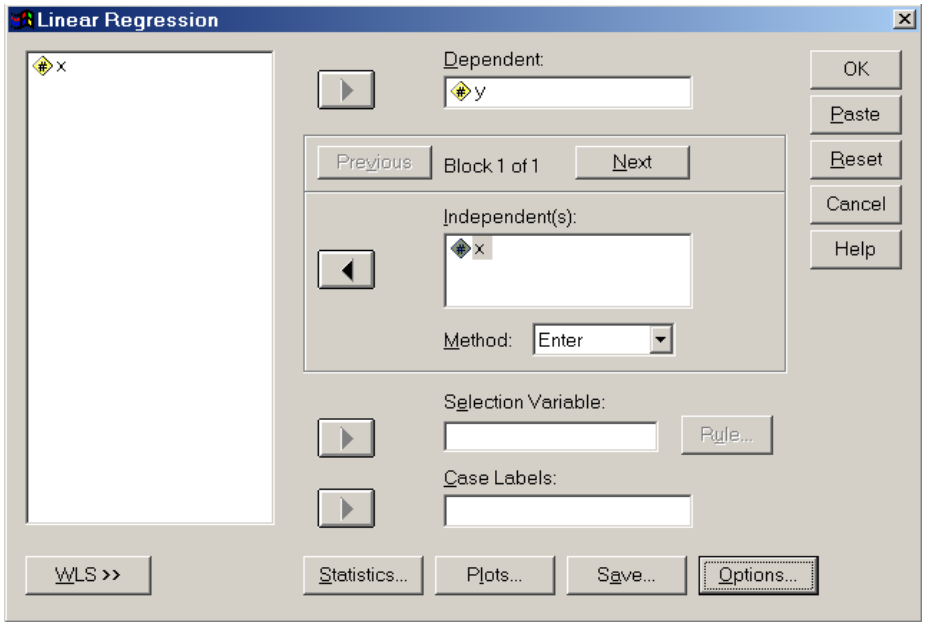
Για να ελέγξουμε την υπόθεση ο συντελεστής β ισούται με 1, δημιουργούμε την μεταβλητή **u=y-x** και, στην συνέχεια, υπολογίζουμε τον συντελεστή συσχέτισης κατά Spearman μεταξύ **u** και **x**. Επειδή το MINITAB δεν δίνει την δυνατότητα απ' ευθείας υπολογισμού, καταχωρίζουμε, σε δύο στήλες (έστω **rx** και **ru**), τις τάξεις μεγέθους των **x** και **u** αντίστοιχα και υπολογίζουμε τον συνήθη συντελεστή συσχέτισής τους (συντελεστή Pearson). Τα αποτελέσματα που προκύπτουν είναι:

```
Correlation of rx and ru = -0.123; P-Value = 0.703
```

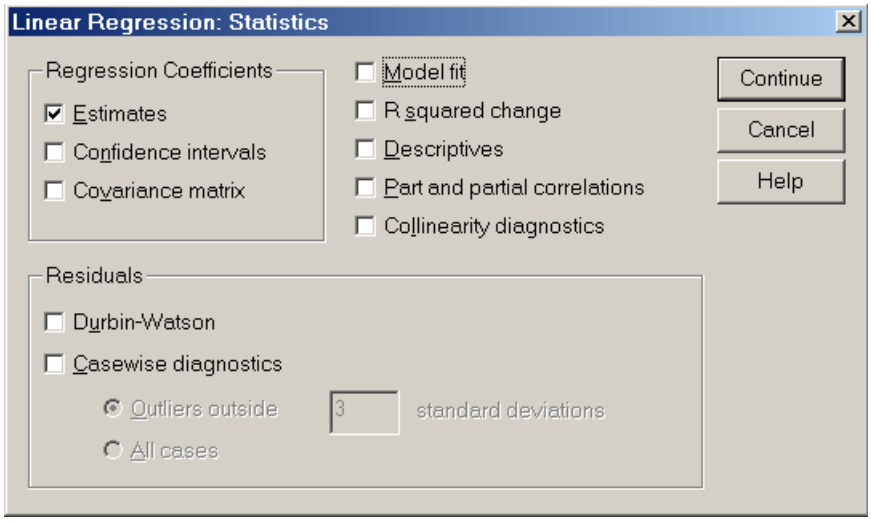
Παρατηρούμε ότι ο συντελεστής συσχέτισης κατά Spearman είναι -0.123. Ιδιαίτερη προσοχή χρειάζεται στο ότι το κρίσιμο επίπεδο που δίνει το πακέτο αναφέρεται σε συντελεστή συσχέτισης κατά Pearson. (Το MINITAB δεν γνωρίζει ότι υπολόγισε τον συντελεστή πάνω σε τάξεις μεγέθους και, επομένως, η οποιαδήποτε συμπερασματολογία θα πρέπει να βασίζεται σε υποθέσεις κανονικότητας των τάξεων μεγέθους, που φυσικά δεν ισχύουν). Για τον λόγο αυτό, η κρίση μας για το

εύλογο της H_0 θα πρέπει να βασισθεί σε σύγκριση της τιμής –0.123 με τις τιμές του πίνακα ποσοστιαίων σημείων της κατανομής του συντελεστή του Spearman. Οπως είδαμε στην αναλυτική λύση του παραδείγματος, η τιμή αυτή είναι εκτός της κρίσιμης περιοχής μεγέθους 0.05 (αυτή ορίσθηκε από την ανισότητα $|\rho|>0.5804$).

Λύση με το SPSS: Καταχωρίζουμε τα δύο δείγματα στις μεταβλητές **x** και **y** κατά τα ήδη γνωστά από προηγούμενα παραδείγματα, και επιλέγουμε **Analyze, Regression, Linear** προκειμένου να εκτιμήσουμε την ευθεία παλινδρόμησης της Y πάνω στην X. Αυτό μας οδηγεί στο εξής πλαίσιο διαλόγου:



Στο πεδίο **Dependent**, δηλώνουμε την εξαρτημένη μεταβλητή (**x**) και στο πεδίο **Independent(s)**, τις ανεξάρτητες μεταβλητές. (Εδώ είναι μία μόνο μεταβλητή, η **x**). Πιέζοντας το πλήκτρο **Statistics**, εμφανίζεται το εξής πλαίσιο:



Επειδή το μόνο που χρειαζόμαστε είναι η εκτίμηση της ευθείας παλινδρόμησης, επιλέγουμε μόνο **Estimates**, στο πεδίο **Regression Coefficients**. Τα αποτελέσματα της παλινδρόμησης (το τμήμα που μας ενδιαφέρει) είναι:

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	9.377	19.924		.471	.648
	X	.857	.250	.734	3.422	.007

a Dependent Variable: Y

Στην στήλη **B**, δίνονται οι τιμές που μας ενδιαφέρουν. Στην συνέχεια, για να ελέγξουμε την υπόθεση ότι η παράμετρος β είναι ίση με 1, δημιουργούμε την μεταβλητή U=Y-X και την καταχωρίζουμε στην μεταβλητή **u**. Τέλος, υπολογίζουμε τον συντελεστή συσχέτισης του Spearman μεταξύ των **x** και **u**. Τα αποτελέσματα περιέχονται στον εξής πίνακα:

Correlations

			X	U
Spearman's rho	X	Correlation Coefficient	1.000	-.123
		Sig. (2-tailed)	.	.703
		N	12	12
	U	Correlation Coefficient	-.123	1.000
		Sig. (2-tailed)	.703	.
		N	12	12

Παρατηρούμε ότι η τιμή του συντελεστή συσχέτισης είναι –0.123.

Σημείωση: Παρά το ότι το SPSS μας δίνει συντελεστή συσχέτισης κατά Spearman, το κρίσιμο επίπεδο έχει υπολογισθεί με βάση την κατανομή του δειγματικού συντελεστή συσχέτισης κατά Pearson κάτω από την υπόθεση μη συσχέτισης και συνεπώς είναι λάθος. Γι' αυτό εμείς από τα αποτελέσματα χρησιμοποιούμε μόνο την τιμή του συντελεστή ρ= –0.123, ενώ η κρίση μας για το εύλογο της H_0 θα πρέπει να βασισθεί σε σύγκριση της τιμής –0.123 με τις τιμές του πίνακα ποσοστιαίων σημείων της κατανομής του συντελεστή του Spearman.

Λύση με το SAS: Οι εντολές που χρησιμοποιούνται για την εκτίμηση της ευθείας παλινδρόμησης του παραδείγματος είναι οι εξής:

```
data twins;
input x y @@;
u=y-x;
cards;
86 88 71 77 77 76 68 64 91 96 72 72 77 65 91 90 70 65 71 80 88 81 87 72
;
run;
proc reg;
model y=x;
run;

proc corr spearman;
var x u;
run;
```

Το αποτέλεσμα περιέχεται στον πίνακα που ακολουθεί.

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	638.46939	638.46939	11.711	0.0065
Error	10	545.19728	54.51973		
C Total	11	1183.66667			

Root MSE	7.38375	R-square	0.5394
Dep Mean	77.16667	Adj R-sq	0.4933
C.V.	9.56857		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	9.376618	19.92381000	0.471	0.6480
X	1	0.857198	0.25048849	3.422	0.0065

The SAS System
Correlation Analysis

2 'VAR' Variables: X U

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
X	12	79.083333	8.887768	77.000000	68.000000	91.000000
U	12	-1.916667	7.153617	-1.000000	-15.000000	9.000000

Spearman Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 12

	X	U
X	1.00000 0.0	-0.12324 0.7028
U	-0.12324 0.7028	1.00000 0.0

Επομένως, όπως προκύπτει από τον πίνακα με τίτλο **Parameter Estimates**, η ευθεία παλινδρόμησης ελαχίστων τετραγώνων είναι η

$$\hat{y} = 9.376618 + 0.857198x.$$

Επίσης, ο συντελεστής συσχέτισης τάξης μεγέθους του Spearman μεταξύ των μεταβλητών X_i και $U_i = Y_i - (1)X_i = Y_i - X_i$, $i = 1, 2, \ldots, 12$ υπολογίσθηκε ίσος με –0.12324. Οπως αναφέρθηκε και στην

επίλυση του παραδείγματος με το SPSS, το κρίσιμο επίπεδο έχει υπολογισθεί με βάση την κατανομή του δειγματικού συντελεστή συσχέτισης κατά Pearson κάτω από την υπόθεση μη συσχέτισης. Ετσι, η κρίση μας για το εύλογο της H_0 θα πρέπει να βασισθεί σε σύγκριση της τιμής –0.123 με τις τιμές του πίνακα ποσοστιαίων σημείων της κατανομής του συντελεστή του Spearman.

7.1.2 Η Μέθοδος Παλινδρόμησης του Theil

Το 1950, ο Theil πρότεινε την εκτίμηση της κλίσης μιας ευθείας παλινδρόμησης χρησιμοποιώντας την διάμεσο των κλίσεων όλων των ευθυγράμμων τμημάτων που ενώνουν ζεύγη σημείων με διαφορετικές Χ τιμές.

Για το ζεύγος σημείων (X_i, Y_i) και (X_j, Y_j), $X_i \neq X_j$, $i < j$, η *σχετική κλίση* ορίζεται ως

$$S_{ij} = \frac{Y_J - Y_i}{X_j - X_i}.$$

Ας υποθέσουμε ότι οι τιμές X_i, i = 1, 2, ..., n είναι διακεκριμένες. Για ευκολία, διατάσσουμε τις παρατηρήσεις κατά αύξουσα σειρά μεγέθους. Προφανώς, $S_{ij} = S_{ji}$, για όλα τα i και j, με αποτέλεσμα να υπάρχουν $\binom{n}{2} = \frac{n(n-1)}{2}$ διακεκριμένες τιμές S_{ij}. Επομένως, μπορούμε να παρουσιάσουμε τις τιμές αυτές με την μορφή ενός άνω τριγωνικού πίνακα:

$$\begin{array}{ccccc} S_{12} & S_{13} & S_{14} & \cdots & S_{1n} \\ & S_{23} & S_{24} & \cdots & S_{2n} \\ & & S_{34} & \cdots & S_{3n} \\ & & & . & S_{n-1,n} \end{array}$$

Η διάμεσος των S_{ij} μπορεί, εύκολα, να προσδιορισθεί από τον πίνακα αυτό. Αν, λοιπόν, συμβολίσουμε με S την διάμεσο των τιμών S_{ij}, τότε, σύμφωνα με τον Theil, η εκτιμήτρια της κλίσης β της ευθείας παλινδρόμησης είναι η

$$\tilde{\beta} = \tilde{S} .$$

Για την σταθερά α, ο Theil πρότεινε ως εκτιμήτρια $\tilde{\alpha}$ την διάμεσο των τιμών $Y_i - \tilde{\beta} X_i$. Δηλαδή,

$$\tilde{\alpha} = \text{διάμεσος } \{Y_i - \tilde{\beta} X_i, i = 1, 2, \ldots, n\}$$

Παρατήρηση: Μια παραλλαγή της μεθόδου του Theil χρησιμοποιεί μία εναλλακτική εκτιμήτρια για την σταθερά α. Συγκεκριμένα, στην θέση της κατά Theil εκτιμήτριας $\tilde{\alpha}$ του α χρησιμοποιείται η εκτιμήτρια $\tilde{\alpha}' = \tilde{Y} - \tilde{\beta}\tilde{X}$, όπου $\tilde{X}$ και $\tilde{Y}$ συμβολίζουν αντίστοιχα τις διαμέσους των X_i και Y_i, i = 1, 2, …, n. Στην περίπτωση αυτή, η εκτιμώμενη ευθεία διέρχεται από την διάμεσο των παρατηρήσεων, ενώ η ευθεία ελαχίστων τετραγώνων διέρχεται από τον μέσο των παρατηρήσεων.

Παράδειγμα 7.1.2: Ας υποθέσουμε ότι έχουμε τα εξής δεδομένα:

	i						
	1	2	3	4	5	6	7
x_i	0	1	2	3	4	5	6
y_i	2.5	3.1	3.4	4.0	4.6	5.1	11.1

Ξεκινώντας από το ζεύγος τιμών $(x_1, y_1) = (0, 2.5)$ και $(x_2, y_2) = (1, 3.1)$, έχουμε

$$S_{12} = (y_2 - y_1)/(x_2 - x_1) = (3.1 - 2.5)/(1 - 0) = 0.6.$$

Προχωρώντας με τον ίδιο τρόπο και γράφοντας τα αποτελέσματα με την μορφή άνω τριγωνικού πίνακα, έχουμε τις εξής αλγεβρικά διακεκριμένες κλίσεις:

0.600	0.450	0.300	0.525	0.520	1.433
	0.300	0.450	0.500	0.500	1.600
		0.600	0.600	0.567	1.900
			0.600	0.550	2.367
				0.500	3.250
					6.000

Η διάμεσος των παραπάνω τιμών είναι $\tilde{S} = 0.567$. Η εκτίμηση της σταθεράς α θα είναι η διάμεσος των τιμών $y_i - \tilde{S}\, x_i$, $i = 1, 2, ..., 7$, δηλαδή, των τιμών

2.5, 2.533, 2.266, 2.299, 2.332, 2.265, 7.698,

που είναι, προφανώς, ίση με 2.332. Επομένως,

$$\tilde{\beta} = 0.567 \text{ και } \tilde{\alpha} = 2.332,$$

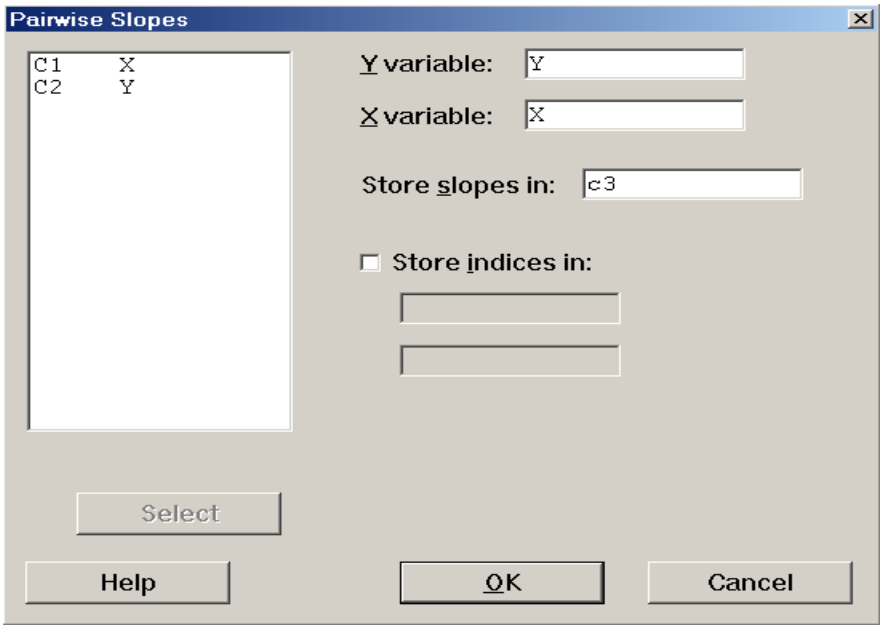
και, κατά συνέπεια, η κατά Theil ευθεία παλινδρόμησης είναι η

$$\tilde{y} = 2.332 + 0.567x.$$

Σημείωση: Αν εκτιμήσουμε την σταθερά α χρησιμοποιώντας την εκτιμήτρια $\tilde{\alpha}' = \tilde{Y} - \tilde{\beta}\tilde{X}$, καταλήγουμε σε εκτίμηση της σταθεράς α ίση με 4 – (0.567) (3) = 2.299. (Αυτό είναι άμεση συνέπεια του ότι η παρατηρηθείσα τιμή της διαμέσου των Y_i είναι 4 και αυτή της διαμέσου των X_i είναι 3).

Λύση με το MINITAB: Το MINITAB δεν δίνει απ' ευθείας εκτιμήσεις της ευθείας παλινδρόμησης με την μέθοδο Theil. Μπορεί, όμως να χρησιμοποιηθεί για τον υπολογισμό των συντελεστών της. Καταχωρίζουμε, στις στήλες **C1** και **C2** (με ονόματα **X** και **Y**), τις

μεταβλητές X και Y αντίστοιχα. Στην συνέχεια, επιλέγουμε **Stat, Nonparametrics, Pairwise Slopes** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



Στα πεδία **Y variable** και **X variable**, δηλώνουμε την εξαρτημένη (**Y**) και την ανεξάρτητη μεταβλητή (**X**) αντίστοιχα. Στο πεδίο **Store slopes in**, δηλώνουμε την στήλη όπου θα καταχωρισθούν οι κλίσεις όλων των ευθυγράμμων τμημάτων που ενώνουν ζεύγη σημείων με διαφορετικές τιμές X. Η διάμεσος της στήλης αυτής είναι η εκτίμηση της τιμής της εκτιμήτριας $\tilde{\beta}$ του συντελεστή β της ευθείας παλινδρόμησης. Η διάμεσος αυτή είναι ίση με 0.567 όπως προκύπτει από τα παρακάτω αποτελέσματα:

```
Variable            N       Mean     Median     TrMean
StDev    SE Mean
C3                 21      1.159      0.567      0.949
1.349      0.294
```

Variable	Minimum	Maximum	Q1	Q3
C3	0.300	6.000	0.500	1.517

Για να εκτιμήσουμε την σταθερά α της ευθείας παλινδρόμησης, καταχωρίζουμε στην στήλη **C4** τις διαφορές **Y-0.567*X** και υπολογίζουμε την διάμεσό τους. Η τιμή της διαμέσου είναι η ζητούμενη εκτίμηση $\tilde{\alpha}$ της σταθεράς α. Προκύπτει ότι $\tilde{\alpha}$ =2.332.

Λύση με το SPSS: Η παλινδρόμηση Theil δεν είναι διαθέσιμη στο SPSS. Μπορούμε όμως να χρησιμοποιήσουμε το πακέτο γαι τον υπολογισμό των εκτιμητριών των συντελεστών της παλινδρόμησης. Ο τρόπος που εργαζόμεθα προσφέρεται μόνο για μικρά δείγματα: Καταχωρίζουμε τα δείγματα τιμών των X και Y σε δύο μεταβλητές (έστω **x** και **y**) και διατάσσουμε τις τιμές της μεταβλητής **x** κατ' αύξουσα τάξη μεγέθους. Αυτό συνεπάγεται ταυτόχρονα αναδιάταξη των τιμών της μεταβλητής **y**. Στην συνέχεια, σε δύο ζεύγη στηλών (έστω **xx1**, **yy1** και **xx2**, **yy2**), καταχωρίζουμε (με επαναλήψεις της διαδικασίας **copy**, **paste** και πολλή προσοχή) όλα τα δυνατά ζεύγη σημείων (X_i, Y_i), (X_j, Y_j) με $X_i \neq X_j$, i<j. Στο τέλος της διαδικασίας αυτής, το φύλλο δεδομένων (το άνω μέρος του) δείχνει ως εξής:

x	y	xx1	yy1	xx2	yy2	var
0	2.5	0	2.5	1	3.1	
1	3.1	0	2.5	2	3.4	
2	3.4	0	2.5	3	4.0	
3	4.0	0	2.5	4	4.6	
4	4.6	0	2.5	5	5.1	
5	5.1	0	2.5	6	11.1	
6	11.1	1	3.1	2	3.4	
.	.	1	3.1	3	4.0	
.	.	1	3.1	4	4.6	
.	.	1	3.1	5	5.1	
.	.	1	3.1	6	11.1	
.	.	2	3.4	3	4.0	
.	.	2	3.4	4	4.6	
.	.	2	3.4	5	5.1	

Στην συνέχεια, δημιουργούμε με **Transform**, **Compute** και πληκτρολογώντας **(yy2-yy1)/(xx2-xx1)** στο πεδίο **Numeric Expression** μία μεταβλητή (έστω **s**) που θα περιέχει τις κλίσεις των ευθυγράμμων τμημάτων μεταξύ των παραπάνω δυνατών ζευγών σημείων. Εύκολα υπολογίζεται ότι η διάμεσος της στήλης αυτής, που αποτελεί την τιμή της εκτιμήτριας $\tilde{\beta}$ του συντελεστή β, είναι ίση με 0.5667. Τέλος, δημιουργούμε την μεταβλητή **a** με **Transform**, **Compute** και πληκτρολογώντας **y-0.5667*x** στο πεδίο **Numeric Expression**. Η διάμεσος αυτής είναι η τιμή της εκτιμήτριας $\tilde{\alpha}$ της σταθεράς $\tilde{\alpha}$ της ευθείας παλινδρόμησης. Προκύπτει ότι η διάμεσος αυτής της μεταβλητής είναι 2.3332. Επομένως, η ευθεία παλινδρόμησης εκτιμάται ως **Y=2.3332+0.5667*X**.

Εναλλακτικά, η σταθερά α της ευθείας παλινδρόμησης μπορεί να εκτιμηθεί από την $\tilde{\alpha} = \tilde{Y} - \tilde{\beta}\tilde{X}$. Αυτό επιτυγχάνεται με υπολογισμό των διαμέσων $\tilde{X}$ και $\tilde{Y}$ (με **Analyze**, **Descriptive Statistics**, **Frequencies**). Οι διάμεσοι προκύπτουν ίσες με 3 για την X και 4 για την Y. Άρα, η σταθερά είναι **4–0.5667*3=2.2999**. Με αυτόν τον τρόπο δηλαδή η ευθεία παλινδρόμησης προκύπτει ίση με **Y=2.299+0.5667*X**.

Λύση με το SAS: Η παλινδρόμηση κατά Theil δεν είναι διαθέσιμη στο SAS. Για τον λόγο αυτό, θα εργασθούμε όπως και στο SPSS. Ετσι, πληκτρολογούμε τις παρακάτω εντολές στο παράθυρο εντολών:

```
data theil;
input x1 y1 x2 y2 ;
s=(y2-y1)/(x2-x1);
cards;
0  2.5  1  3.1
0  2.5  2  3.4
0  2.5  3  4
0  2.5  4  4.6
0  2.5  5  5.1
0  2.5  6  11.1
1  3.1  2  3.4
1  3.1  3  4
1  3.1  4  4.6
1  3.1  5  5.1
1  3.1  6  11.1
2  3.4  3  4
2  3.4  4  4.6
2  3.4  5  5.1
2  3.4  6  11.1
3  4    4  4.6
3  4    5  5.1
3  4    6  11.1
4  4.6  5  5.1
4  4.6  6  11.1
5  5.1  6  11.1
;
run;
proc print;
run;
proc univariate noprint;
var s;
output out=med median=med_s;
run;
proc print;
run;

data theil2;
input x y;
a=y-(0.56667*x);
cards;
```

```
0      2.5
1      3.1
2      3.4
3      4
4      4.6
5      5.1
6      11.1
;
run;
proc univariate noprint;
var a;
output out=med median=med_a;
run;
proc print;
run;
```

Τα αποτελέσματα που δίνει το πακέτο εμφανίζονται στους πίνακες που ακολουθούν.

The SAS System

OBS	X1	Y1	X2	Y2	S
1	0	2.5	1	3.1	0.60000
2	0	2.5	2	3.4	0.45000
3	0	2.5	3	4.0	0.50000
4	0	2.5	4	4.6	0.52500
5	0	2.5	5	5.1	0.52000
6	0	2.5	6	11.1	1.43333
7	1	3.1	2	3.4	0.30000
8	1	3.1	3	4.0	0.45000
9	1	3.1	4	4.6	0.50000
10	1	3.1	5	5.1	0.50000
11	1	3.1	6	11.1	1.60000
12	2	3.4	3	4.0	0.60000
13	2	3.4	4	4.6	0.60000
14	2	3.4	5	5.1	0.56667
15	2	3.4	6	11.1	1.92500
16	3	4.0	4	4.6	0.60000
17	3	4.0	5	5.1	0.55000
18	3	4.0	6	11.1	2.36667
19	4	4.6	5	5.1	0.50000
20	4	4.6	6	11.1	3.25000
21	5	5.1	6	11.1	6.00000

The SAS System

OBS	MEDIANS
1	0.56667

The SAS System

OBS	MED_A
1	2.33332

Παρατηρούμε ότι ο η εκτίμηση του συντελεστή β της ζητούμενης εξίσωσης παλινδρόμησης είναι ίσος με 0.56667, ενώ η εκτίμηση της σταθεράς α είναι ίση με 2.33332.

Ας θεωρήσουμε τώρα μια εκτιμήτρια β^* της σταθεράς β και ας συμβολίσουμε με α^* την αντίστοιχη εκτιμήτρια της σταθεράς α. (Η εκτιμήτρια αυτή είναι συνάρτηση της β^*, δηλαδή $\alpha^* = \alpha^*(\beta^*)$). Είναι σαφές ότι, αν η β^* είναι μία *καλή εκτιμήτρια* της παραμέτρου β, τότε τα κατάλοιπα $e_i^* = y_i - a^* - b^* x_i$, που αντιστοιχούν σε κάθε μία παρατήρηση, θα πρέπει να είναι θετικά ή αρνητικά με την ίδια πιθανότητα, όπου b^* και a^* είναι οι εκτιμήσεις που δίνουν οι εκτιμήτριες β^* και α^*, αντίστοιχα. Αυτό σημαίνει ότι η τυχαία μεταβλητή $e^* = Y - \alpha^* - \beta^* X$, της οποίας πραγματοποιήσεις είναι οι τιμές των καταλοίπων e_i^*, πρέπει να κατανέμεται με τυχαίο τρόπο ανεξάρτητα από την τυχαία μεταβλητή X και με μέση ή διάμεση τιμή ίση με το 0. (Το τελευταίο αυτό συμπέρασμα βρίσκεται σε πλήρη αρμονία με το γεγονός της ανεξαρτησίας των μεταβλητών $\varepsilon = Y - \alpha - \beta X$ και X που ισχύει από τον ορισμό της θεωρηθείσας μορφής παλινδρόμησης).

Παρατηρούμε ότι

$$S_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$$

$$= \frac{a + \beta X_j + \varepsilon_j - a - \beta X_i - \varepsilon_i}{X_j - X_i}$$

$$= \beta + \frac{\varepsilon_j - \varepsilon_i}{X_j - X_i}, \quad i < j, \; X_i \neq X_j$$

(όπου $\varepsilon_k = Y_k - \beta X_k - \alpha$ είναι η απόκλιση της Y_k από την υποτεθείσα ευθεία παλινδρόμησης (σφάλμα)).

Είναι προφανές, από την παραπάνω σχέση, ότι η κλίση S_{ij} είναι μεγαλύτερη (αντίστοιχα, μικρότερη) από την σταθερά β οποτεδήποτε

$\frac{\varepsilon_j - \varepsilon_i}{X_j - X_i} > 0$ (αντίστοιχα, $\frac{\varepsilon_j - \varepsilon_i}{X_j - X_i} < 0$). Δηλαδή, η κλίση S_{ij} είναι μεγαλύτερη ή μικρότερη από την σταθερά β, ανάλογα με το εάν τα ζεύγη των τυχαίων μεταβλητών $(X_i, \varepsilon_i) \equiv (X_i, Y_i - \alpha - \beta X_i)$ και $(X_j, \varepsilon_j) \equiv (X_j, Y_j - \alpha - \beta X_j)$ είναι εναρμονισμένα ή μη εναρμονισμένα, με την έννοια που οι όροι αυτοί χρησιμοποιήθηκαν στην περίπτωση του συντελεστή συσχέτισης τ του Kendall. Ισοδύναμα, η κλίση S_{ij} είναι μεγαλύτερη ή μικρότερη της σταθεράς β, ανάλογα με το εάν τα ζεύγη των τυχαίων μεταβλητών $(X_i, U_i) \equiv (X_i, Y_i - \beta X_i)$ και $(X_j, U_j) \equiv (X_j, Y_j - \beta X_j)$ είναι εναρμονισμένα ή μη εναρμονισμένα, με την έννοια του εναρμονισμένου ζεύγους τυχαίων μεταβλητών, όπως αυτό ορίσθηκε στην περίπτωση του συντελεστή τ του Kendall. Αυτό, στην ουσία, σημαίνει ότι μία τιμή β είναι αποδεκτή ως τιμή της κλίσης της ευθείας παλινδρόμησης, αν για την τιμή αυτή δεν απορρίπτεται η υπόθεση ανεξαρτησίας των μεταβλητών X και $U = Y - \beta X$ με βάση τον έλεγχο τ του Kendall (αφού η ελεγχοσυνάρτηση τ εκφράζεται μέσω των εναρμονισμένων και μη εναρμονισμένων ζευγών (X_i, U_i) και (X_j, U_j)). Με άλλα λόγια, δεχόμαστε οποιαδήποτε τιμή β, η οποία δεν οδηγεί σε έναν αριθμό εναρμονισμένων ή μη εναρμονισμένων ζευγών που να υποδηλώνει ότι ο συντελεστής τ του Kendall διαφέρει από το μηδέν. Συγκεκριμένα, δεν επιθυμούμε ο αριθμός N_d των μη εναρμονισμένων (αντίστοιχα, ο αριθμός N_c των εναρμονισμένων) ζευγών να είναι πολύ μικρός ή πολύ μεγάλος. Επομένως, αν, για τον έλεγχο της υπόθεσης H_0: $\beta=\beta_0$, χρησιμοποιήσουμε ως ελεγχοσυνάρτηση T τον αριθμό των κλίσεων που είναι μικρότερες από την τιμή β, δηλαδή, την στατιστική συνάρτηση T = αριθμός των S_{ij} που είναι μικρότερες από την τιμή

β_0, δεν απορρίπτουμε την τιμή β_0, αν ο αριθμός N_d των μη εναρμονισμένων ζευγών $(X_i, Y_i - \beta_0 X_i)$ και $(X_j, Y_j - \beta_0 X_j)$ δεν είναι πολύ μικρός ή πολύ μεγάλος. Γνωρίζουμε, όμως, ότι ο αριθμός N_d συνδέεται με τον αριθμό N_c με την σχέση $N_c + N_d = N$, όπου N είναι ο συνολικός αριθμός των ζευγών παρατηρήσεων. Τότε, χρησιμοποιώντας τον πίνακα ποσοστιαίων σημείων της στατιστικής συνάρτησης $N_c - N_d$ (πίνακας 12 του παραρτήματος), θα θεωρούμε ότι η τιμή N_d είναι πολύ μικρή, αν η τιμή της στατιστικής συνάρτησης $N_c - N_d$ υπερβαίνει το (1–p)- ποσοστιαίο σημείο της κατανομής της, όπου $p=\alpha$ ή $\alpha/2$ σύμφωνα με το εάν πρόκειται για μονόπλευρη ή αμφίπλευρη εναλλακτική. Δηλαδή, η τιμή N_d θεωρείται πολύ μικρή αν $N_c - N_d > w_{1-p}$. Η τελευταία ανισότητα είναι ισοδύναμη με την ανισότητα $N - 2N_d > w_{1-p}$, δηλαδή

$$N_d < \frac{1}{2}\left(N - w_{1-p}\right).$$

Επομένως, η τιμή N_d είναι πολύ μικρή, αν $N_d < r$, όπου $r = \frac{1}{2}\left(N - w_{1-p}\right)$. Κατά συνέπεια, για τον έλεγχο των υποθέσεων της μορφής $B\ (H_1 : \beta > \beta_0)$ σε επίπεδο σημαντικότητας α, η κρίσιμη περιοχή είναι της μορφής $N_d < r$, όπου $r = \frac{1}{2}\left(N - w_{1-a}\right)$. Με άλλα λόγια, η τιμή β είναι *αποδεκτή*, αν $N_d \geq r$, όπου r ορίζεται όπως παραπάνω. Αυτό σημαίνει ότι η τιμή β_0 είναι αποδεκτή αν υπερβαίνει τουλάχιστον r κλίσεις S_{ij}, δηλαδή, αν $\beta_0 > S^{(r)}$, όπου $S^{(r)}$ είναι η r κλίση στο διατεταγμένο δείγμα των κλίσεων.

Από τα παραπάνω, συνάγεται ότι ο κανόνας απόφασης στην περίπτωση των υποθέσεων της μορφής Β έχει την μορφή:

Η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν $\beta_0 \le S^{(r)}$, $r = \frac{1}{2}\left(N - w_{1-a}\right)$.

Κατ' αναλογία, στην περίπτωση των υποθέσεων της μορφής Γ ο κανόνας απόφασης είναι:

Η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν $\beta_0 \ge S^{(t)}$, $t = \frac{1}{2}\left(N + w_{1-a}\right) + 1$.

Τέλος, για τον αμφίπλευρο έλεγχο της περίπτωσης Α, ο κανόνας απόφασης είναι:

Η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν $\beta_0 \le S^{(r)}$ ή $\beta_0 \ge S^{(t)}$, όπου $r = \frac{1}{2}\left(N - w_{1-a/2}\right)$, $t = \frac{1}{2}\left(N + w_{1-a/2}\right) + 1$.

(Υπενθυμίζεται ότι N είναι ο αριθμός των κλίσεων).

Η παραπάνω επιχειρηματολογία, μπορεί να οδηγήσει στην κατασκευή ενός διαστήματος εμπιστοσύνης για την κλίση β της ευθείας παλινδρόμησης. Πράγματι, το (1–α)-διάστημα εμπιστοσύνης για την κλίση β της ευθείας παλινδρόμησης είναι

$$(S^{(r)}, S^{(t)}),\ r = \frac{1}{2}\left(N - w_{1-a/2}\right),\ t = \frac{1}{2}\left(N + w_{1-a/2}\right) + 1.$$

Δηλαδή, αν r και t ορίζονται όπως παραπάνω, τότε ο συντελεστης εμπιστοσύνης είναι τουλάχιστον ίσος με 1–α. Συμβολικά,

$$P(S^{(r)} < \beta < S^{(t)}) \ge 1 - \alpha.$$

(Η τιμή $w_{1-a/2}$ προσδιορίζεται από τον πίνακα 12 του παραρτήματος. Επίσης, η τιμή του r στρογγυλοποιείται προς τα κάτω, ενώ η τιμή του t στρογγυλοποιείται προς τα πάνω αν δεν είναι ήδη ακέραιες).

Παράδειγμα 7.1.3: Ας υποθέσουμε ότι ενδιαφερόμαστε να κατασκευάσουμε ένα 95% - διάστημα εμπιστοσύνης για την κλίση β της ευθείας παλινδρόμησης για τα δεδομένα του αμέσως προηγούμενου παραδείγματος. Από τον σχετικό πίνακα του παραρτήματος, έχουμε , για n=7, ότι $w_{0.975}$ = 13. Σύμφωνα με την θεωρία που αναπτύχθηκε, προκύπτει, τότε, ότι r=4 και t=18. Επομένως, $S^{(4)}$ = 0.450, $S^{(18)}$ = 1.900 και το 95%–διάστημα εμπιστοσύνης για την κλίση β της ευθείας παλινδρόμησης είναι το διάστημα (0.450, 1.900).

Λύση με το MINITAB: Για να λύσουμε το παράδειγμα με τη βοήθεια του MINITAB, αρκεί να διατάξουμε την στήλη των κλίσεων κατά αύξουσα τάξη μεγέθους κατά τα ήδη γνωστά και, αφού υπολογίσουμε τα r και t από τον πίνακα 12, να χρησιμοποιήσουμε τις τιμές των γραμμών r και t της στήλης των κλίσεων ως άκρα του ζητούμενου διαστήματος εμπιστοσύνης.

Λύση με το SPSS: Για να υπολογίσουμε το 95% διάστημα εμπιστοσύνης του συντελεστή β, χρειαζόμαστε την τέταρτη και την δέκατη όγδοη τιμή της διατεταγμένης κατ' αύξουσα τάξη μεγέθους μεταβλητής **s** που δημιουργήσαμε στο προηγούμενο παράδειγμα. Η διάταξη της γίνεται κατά τα ήδη γνωστά με **Data**, **Sort Cases**.

Λύση με το SAS: Το 95% διάστημα εμπιστοσύνης του συντελεστή β, υπολογίζεται με κάτω όριο την τέταρτη τιμή και με άνω όριο την δέκατη όγδοη τιμή της διατεταγμένης κατ' αύξουσα τάξη μεγέθους μεταβλητής **s**. Για να διατάξουμε κατά αύξουσα σειρά μεγέθους τις τιμές της μεταβλητής s, χρησιμοποιούμε τις παρακάτω εντολές.

```
data theil;
input x1 y1 x2 y2 ;
s=(y2-y1)/(x2-x1);
cards;
0  2.5  1  3.1
0  2.5  2  3.4
0  2.5  3  4
0  2.5  4  4.6
0  2.5  5  5.1
0  2.5  6  11.1
1  3.1  2  3.4
1  3.1  3  4
1  3.1  4  4.6
1  3.1  5  5.1
1  3.1  6  11.1
2  3.4  3  4
2  3.4  4  4.6
2  3.4  5  5.1
2  3.4  6  11.1
3  4    4  4.6
3  4    5  5.1
3  4    6  11.1
4  4.6  5  5.1
4  4.6  6  11.1
5  5.1  6  11.1
;
run;
proc sort; by s;
run;
proc print;
run;
```

Παράδειγμα 7.1.4: Έστω ότι για τα δεδομένα του προβλήματος των διδύμων, η ευθεία ελαχίστων τετραγώνων θεωρείται μη ικανοποιητική. Εφαρμόζοντας την παραλλαγή της μεθόδου του Theil, τα N=63 ζεύγη παρατηρήσεων (X_i, Y_i) και (X_j, Y_j) με $i<j$ και $X_i \neq Y_j$ χρησιμοποιούνται για τον προσδιορισμό των κλίσεων S_{ij}. Εύκολα μπορεί να διαπιστωθεί ότι $\tilde{S}=0.89$, $\tilde{Y}=76.5$, $\tilde{X}=77$, όπου $\tilde{Y}$ (αντίστοιχα, $\tilde{X}$) παριστάνει την διάμεσο τιμή του δείγματος παρατηρήσεων $Y_1, Y_2, ..., Y_n$ (αντίστοιχα, $X_1, X_2, ..., X_n$). Επομένως, οι εκτιμήσεις των συντελεστών της ευθείας παλινδρόμησης είναι:

$$\tilde{b} = \tilde{s} = 0.89, \quad \tilde{a}' = \tilde{y} - \tilde{s}\tilde{x} = 7.79.$$

Κατά συνέπεια, η εκτίμηση της ευθείας παλινδρόμησης με την παραλλαγμένη μέθοδο του Theil είναι η ευθεία

$$\tilde{y} = 7.79 + 0.89x.$$

Για ένα 95%-διάστημα εμπιστοσύνης για την κλίση β της ευθείας παλινδρόμησης, οι N=63 κλίσεις S_{ij} διατάσσονται κατά αύξουσα σειρά μεγέθους. Από τον σχετικό πίνακα του παραρτήματος, προκύπτει, για n=12, ότι $w_{0.975} = 28$. Επομένως, r=17 και t=47. Κατά συνέπεια, $S^{(r)} = S^{(17)} = 0.24$ και $S^{(t)} = S^{(47)} = 1.48$ και το 95%–διάστημα εμπιστοσύνης για την κλίση β της ευθείας παλινδρόμησης, είναι το διάστημα (0.24, 1.48).

Λύση με το MINITAB: Εργαζόμενοι όπως στο προηγούμενο παράδειγμα, οι εκτιμήσεις των συντελεστών α και β, που το MINITAB δίνει με την παραλλαγή της μεθόδου Theil, προκύπτουν ίσες με $\tilde{\beta}$ =0.889, $\tilde{\alpha}'$=7.77. Διατάσσουμε κατά τα ήδη γνωστά τις τιμές της $\tilde{\beta}$ κατά αύξουσα τάξη μεγέθους και προσδιορίζουμε τα r και s από τον σχετικό πίνακα του παραρτήματος. Προκύπτει ότι τα άκρα του ζητούμενου διαστήματος εμπιστοσύνης είναι οι τιμές που βρίσκονται στις γραμμές 17 και 47 της στήλης που περιέχει τις τιμές της κλίσης β διατεταγμένες κατ' αύξουσα τάξη μεγέθους. Το ζητούμενο 95% διάστημα εμπιστοσύνης προκύπτει ίσο με (0.2353,1.4762).

Λύση με το SPSS: Και αυτό το παράδειγμα λύνεται εύκολα, όπως και τα δύο προηγούμενα, αλλά με χρονοβόρο τρόπο. Αφού καταχωρίσουμε τα δύο δείγματα στις στήλες **x** και **y** και ταξινομήσουμε κατά αύξουσα σειρά μεγέθους ώς προς τις τιμές της **x**, δημιουργούμε τις στήλες των δυνατών ζευγών σημείων (x_i, y_i) και (x_j, y_j) με $x_i \neq x_j$ και $i<j$. (Δεν πρέπει να εξετάζουμε ζεύγη σημείων που έχουν την ίδια x συντεταγμένη). Στην συνέχεια, δημιουργούμε την στήλη των αντίστοιχων κλίσεων. Η διάμεσός της, που είναι η εκτίμηση του β,

προκύπτει ίση με 0.8889. Κατόπιν, προκύπτει ότι η εκτίμηση του α (με βάση την εκτιμήτρια $\tilde{\alpha}$) είναι 7.777. Άρα η εκτιμώμενη ευθεία παλινδρόμησης είναι **Y=7.777+0.8889*X**.

Σύμφωνα με τον σχετικό πίνακα του παραρτήματος, το 95% διάστημα εμπιστοσύνης για το β θα έχει ως όρια την δέκατη έβδομη και την τεσσαρακοστή έβδομη κλίση, όταν αυτές έχουν διαταχθεί κατά αύξουσα τάξη μεγέθους. Εύκολα προκύπτει ότι αυτές είναι ίσες με 0.24 και 1.48, αντίστοιχα. Άρα το 95% διάστημα εμπιστοσύνης του β είναι (0.24, 1.48).

Λύση με το SAS: Η διαδικασία που μπορεί να ακολουθηθεί είναι η ίδια όπως στα προηγούμενα παραδείγματα με τις ίδιες ακριβώς εντολές. Παρόλα αυτά, η εισαγωγή των δεδομένων που αποτελούν τα δυνατά ζεύγη σημείων (x_i, y_i) και (x_j, y_j) με $x_i \neq y_i$ και $i<j$ είναι ιδιαίτερα χρονοβόρα και δεν ενδείκνυται η χρήση του SAS, επειδή η εισαγωγή των δεδομένων γίνεται σε απλό επεξεργαστή κειμένου και δεν βοηθάει η χρησιμοποίηση των διαδικασιών copy και paste.

7.2 ΜΕΘΟΔΟΙ ΜΟΝΟΤΟΝΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

(Monotone Regression)

Συχνά, παρουσιάζονται περιπτώσεις όπου η παλινδρόμηση μίας μεταβλητής Y πάνω σε μία άλλη μεταβλητή X δεν είναι εύλογο να υποτεθεί ότι είναι γραμμική, αλλά ενδέχεται να είναι εύλογο να υποτεθεί ότι η E(Y|X) αυξάνει (τουλάχιστον δεν ελαττώνει) καθώς η X αυξάνει. Σε μία τέτοια περίπτωση, θα λέμε ότι η παλινδρόμηση είναι *μονότονα αύξουσα* (*monotonically increasing*). Αντίστοιχα, αν η E(Y|X) ελαττώνει (τουλάχιστον δεν αυξάνει) καθώς η X αυξάνει, θα λέμε ότι η παλινδρόμηση είναι *μονότονα φθίνουσα* (*monotonically decreasing*). Οι περιπτώσεις αυτές, όπου μία ευθεία παλινδρόμησης δεν είναι η καλύτερη μορφή παλινδρόμησης για την έκφραση των δεδομένων ενός προβλήματος, είναι συνήθως εκείνες στις οποίες τα κατάλοιπα δεν φαίνονται να κατανέμονται με τυχαίο τρόπο (δεν έχουν ένα τυχαίο *άπλωμα*). Είναι, δηλαδή, οι περιπτώσεις εκείνες στις οποίες τα κατάλοιπα είναι *σειριακά συσχετισμένα* (*serially correlated*). Η σειριακή συσχέτιση που ενδέχεται να υπάρχει μεταξύ των καταλοίπων μιας εκτιμηθείσας ευθείας παλινδρόμησης μετριέται συνήθως με τον λεγόμενο *σειριακό συντελεστή συσχέτισης* (*serial correlation coefficient*) που ορίζεται ως εξής:

Εστω $e_1, e_2, ..., e_n$, όπου $e_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$, $i = 1, 2, ..., n$ η ακολουθία των καταλοίπων μιας εκτιμηθείσας ευθείας παλινδρόμησης. Ορίζουμε, ως *συντελεστή σειριακής συσχέτισης* (*serial correlation coefficient*) της ακολουθίας των τυχαίων μεταβλητών $e_1, e_2, ..., e_n$ και συμβολίζουμε με R, τον κατά Pearson συντελεστή συσχέτισης των μεταβλητών e_i και e_{i+1} . Δηλαδή, $R = r_{e_i, e_{i+1}}$, ή, ισοδύναμα,

$$R = \frac{\sum_{i=1}^{n-1} e_i e_{i+1} - \left(\sum_{i=1}^{n-1} e_i\right)\left(\sum_{i=2}^{n} e_i\right) \Big/ n}{\left[\sum_{i=1}^{n-1} e_i^2 - \left(\sum_{i=1}^{n-1} e_i\right)^2\right]^{1/2} \left[\sum_{i=2}^{n} e_i^2 - \left(\sum_{i=2}^{n} e_i\right)^2\right]^{1/2}}$$

Μεγάλες τιμές του συντελεστή R αποτελούν ένδειξη ότι το εκτιμηθέν μοντέλο δεν είναι, ενδεχομένως, το καλύτερο που θα μπορούσαμε να χρησιμοποιήσουμε. Στην περίπτωση αυτή, μπορούμε να χρησιμοποιήσουμε την τεχνική της *μονότονης παλινδρόμησης*.

Η μέθοδος αυτή χρησιμοποιεί τεχνικές γραμμικής παλινδρόμησης, τις οποίες εφαρμόζει στις τάξεις μεγέθους των παρατηρήσεων και όχι στις παρατηρήσεις αυτές καθεαυτές. Η λογική της τεχνικής της μονότονης παλινδρόμησης βασίζεται στο γεγονός ότι, αν δύο μεταβλητές έχουν μία μονότονη σχέση, τότε οι τάξεις μεγέθους τους θα έχουν μία γραμμική σχέση. Οι τάξεις μεγέθους, δηλαδή, των μεταβλητών αποτελούν το αποτέλεσμα ενός μετασχηματισμού των αρχικών μεταβλητών, ο οποίος επιδιώκει να μετατρέψει την μονότονη συνάρτηση παλινδρόμησης σε μία γραμμική συνάρτηση παλινδρόμησης.

Ας υποθέσουμε ότι έχουμε ένα τυχαίο δείγμα $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ από κάποια διμεταβλητή κατανομή. Για να προβλέψουμε μία τιμή της μεταβλητής Y όταν $X=x_0$ με βάση την μονότονη σχέση των μεταβλητών X και Y, ακολουθούμε την εξής διαδικασία:

1. Προσδιορίζουμε τις τάξεις μεγέθους $R(X_i)$ των τιμών X_i, $i = 1, 2, ..., n$ και $R(Y_i)$ των τιμών Y_i, $i = 1, 2, ..., n$. (Χρησιμοποιούμε τους μέσους των τάξεων μεγέθους των παρατηρήσεων που οι τιμές

τους ταυτίζονται στην θέση των τάξεων μεγέθους που αυτές θα είχαν αν οι τιμές τους ήταν διαφορετικές).

2. Προσαρμόζουμε μία ευθεία γραμμικής παλινδρόμησης των μεταβλητών $R(Y_i)$ πάνω στις μεταβλητές $R(X_i)$, i = 1, 2, ..., n:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n(n+1)^2/4}{\sum_{i=1}^{n} [R(X_i)]^2 - n(n+1)^2/4}$$

$$\hat{\alpha} = (1-\hat{\beta})(n+1)/2 .$$

(Οι παραπάνω τύποι είναι οι απλοποιημένες μορφές των σχετικών τύπων για τους συντελεστές ελαχίστων τετραγώνων, όταν χρησιμοποιούνται οι τάξεις μεγέθους των παρατηρήσεων στην θέση των παρατηρήσεων). Επομένως, η εξίσωση γραμμικής παλινδρόμησης μεταξύ των τάξεων μεγέθους των μεταβλητών Y_i και των τάξεων μεγέθους των μεταβλητών X_i , i = 1, 2, ..., n είναι η

$$\hat{R} = (Y_i) = \hat{\alpha} + \hat{\beta}R(X_i) .$$

3. Η τάξη μεγέθους της τιμής x προσδιορίζεται ως εξής:

$$R(x_0)=\begin{cases} R(X^{(1)})\equiv 1, & \text{αν } x_0 \le X^{(1)} \\ R(X_i), & \text{αν } x_0 = X_i, \text{ για κάποιο i} \\ R(X_i)+\dfrac{x_0-X_i}{X_j-X_i}\left[R(X_j)-R(X_i)\right], & \text{αν } X_i < x_0 < X_j \text{ για κάποια i, j} \\ R(X^{(n)})\equiv n, & \text{αν } x_0 \ge X^{(n)}. \end{cases}$$

Δηλαδή, ως τάξη μεγέθους της τιμής x_0, θεωρούμε την τιμή 1, δηλαδή, την τάξη μεγέθους της μικρότερης παρατήρησης, αν η τιμή x είναι μικρότερη ή ίση από αυτήν. Αν η x_0 ταυτίζεται με κάποια από τις παρατηρήσεις του δείγματος των X, τότε ως τάξη μεγέθους της τιμής x_0 θεωρούμε την τάξη μεγέθους της συγκεκριμένης παρατήρησης. Αν η x_0 βρίσκεται μεταξύ δύο διαδοχικών τιμών του δείγματος των X, η τάξη μεγέθους της υπολογίζεται με την μέθοδο της γραμμικής παρεμβολής. Τέλος, αν η τιμή x_0 είναι ίση ή μεγαλύτερη από την μεγαλύτερη τιμή του δείγματος των παρατηρήσεων πάνω στη μεταβλητή X, τότε ως τάξη μεγέθους της x_0, ορίζεται η τιμή n, δηλαδή, η τάξη της μέγιστης παρατήρησης.

4. Αντικαθιστούμε την τιμή $R(x_0)$ στην παραπάνω εκτιμηθείσα ευθεία παλινδρόμησης μεταξύ των τάξεων μεγέθους των παρατηρήσεων και παίρνουμε μία εκτίμηση της τάξης μεγέθους $R(y_0)$ της τιμής $E(Y|X=x_0)$:

$$R(y_0) = a + bR(x_0),$$

όπου a και b συμβολίζουν τις εκτιμήσεις στις οποίες οδηγούν οι εκτιμήτριες $\hat{\alpha}$ και $\hat{\beta}$ αντίστοιχα για το συγκεκριμένο δείγμα.

5. Μετατρέπουμε την τάξη μεγέθους $R(y_0)$ σε μία εκτίμηση $\hat{y}_0 \equiv \hat{E}(Y|X=x_0)$ της τιμής $E(Y|X=x_0)$ ως εξής:

$$\hat{E}(Y \mid X = x_0) = \begin{cases} Y^{(1)}, & \text{αν } R(y_0) \le \min\{R(Y_i), i = 1,2,...,n\} \\ Y_i, & \text{αν } R(y_0) = R(Y_i), \text{για κάποιο } i \\ Y_i + \dfrac{R(y_0) - R(Y_i)}{R(y_j) - R(Y_i)}(Y_j - Y_i), & \text{αν } R(y_i) < R(y_0) < R(y_j) \text{ για κάποια } i, j \\ Y^{(n)}, & \text{αν } R(y_0) \ge \max\{R(Y_i), i = 1,2,...,n\} \end{cases}$$

Δηλαδή, αν η τάξη μεγέθους $R(y_0)$ είναι μικρότερη ή ίση από την ελάχιστη παρατηρηθείσα τάξη μεγέθους των Y τιμών, τότε η $\hat{E}(Y|X=x_0)$ ορίζεται ίση με την μικρότερη παρατηρηθείσα Y τιμή. Αντίστοιχα, αν η τάξη μεγέθους $R(y_0)$ είναι μεγαλύτερη ή ίση από την μέγιστη παρατηρηθείσα τάξη μεγέθους των Y τιμών, τότε η $\hat{E}(Y|X=x_0)$ ορίζεται ίση με την μέγιστη παρατηρηθείσα Y τιμή. Αν η τάξη μεγέθους $R(y_0)$ ισούται με την τάξη μεγέθους μιας από τις παρατηρήσεις Y_i, i = 1, 2, ..., n, τότε η εκτίμηση $\hat{E}(Y|X=x_0)$ ορίζεται ίση με την τιμή της παρατήρησης αυτής. Τέλος, αν η τάξη μεγέθους $R(y_0)$ βρίσκεται μεταξύ των τάξεων μεγέθους δύο γειτονικών τιμών της μεταβλητής Y, τότε η εκτίμηση $\hat{E}(Y|X=x_0)$ υπολογίζεται με την μέθοδο της γραμμικής παρεμβολής.

Παράδειγμα 7.2.1: Δεκαπέντε άτομα εκπαιδεύθηκαν για διαφορετικές χρονικές περιόδους (X) σε ώρες πάνω στον χειρισμό μιας μηχανής. Μετά την συμπλήρωση της περιόδου εκπαίδευσης, τα άτομα αυτά εξετάσθηκαν και αξιολογήθηκαν με βάση τον χρόνο (Y) σε λεπτά που χρειάσθηκαν για να ολοκληρώσουν την ίδια εργασία με τον χειρισμό της μηχανής. Τα αποτελέσματα συνοψίζονται στον πίνακα που ακολουθεί.

X_i:	0	0	1	1.6	3	4	4	5	6.5	8	8	10	12	12.6	14
Y_i:	18.4	17.4	16.2	16.4	14.4	10.5	11.2	10.8	9.0	8.4	7.0	7.2	6.6	5.0	5.6

Να εκτιμήσετε τον μέσο χρόνο που θα απαιτηθεί για τον χειρισμό της μηχανής από ένα άτομο που εκπαιδεύθηκε για ένα διάστημα 11 ωρών.

Λύση: Από την μορφή των δεδομένων, μπορούμε να θεωρήσουμε ότι οι μεταβλητές X και Y έχουν μία μόνοτονη σχέση, χωρίς απαραίτητα να θεωρήσουμε ότι η σχέση αυτή είναι γραμμική. Τα δεδομένα είναι ήδη διατεταγμένα κατά αύξουσα σειρά μεγέθους των τιμών της μεταβλητής X.

Στον πίνακα που ακολουθεί δίνονται οι τάξεις μεγέθους των ζευγών (X_i, Y_i), $i = 1, 2, ..., 15$.

X_i	Y_i	$R(X_i)$	$R(Y_i)$
0	18.4	1.5	15
0	17.4	1.5	14
1	16.2	3	12
1.6	16.4	4	13
3	14.4	5	11
4	11.2	6.5	10
4	10.5	6.5	8
5	10.8	8	9
6.5	9.0	9	7
8	8.4	10.5	6
8	7.0	10.5	4
10	7.2	12	5
12	6.6	13	3
12.6	5.0	14	1
14	5.6	15	2

Με βάση τα στοιχεία του πίνακα αυτού, υπολογίζουμε τις τιμές a και b των εκτιμητριών $\hat{\alpha}$ και $\hat{\beta}$:

$$a = 15.856,\ b = -0.982.$$

Επομένως, η ευθεία ελαχίστων τετραγώνων που εκτιμά την γραμμική σχέση μεταξύ των τάξεων μεγέθους των μεταβλητών X και Y είναι η

$$\hat{R}(y_i) = 15.856 - 0.982R(x_i).$$

Για την εκτίμηση της τιμής $y_0 = E(Y|X=11)$, ακολουθούμε τα βήματα που περιγράψαμε παραπάνω θέτοντας $x_0 = 11$:

Προφανώς,

$$X_{12} \equiv 10 < x_0 \equiv 11 < 12 \equiv X_{13}.$$

Επομένως, $R(x_0)=12.5$ και, από την εκτιμηθείσα εξίσωση παλινδρόμησης, έχουμε ότι $R(y_0) = 15.856 - (0.982)(12.5) = 3.581$. Δηλαδή, $R(6.6) \equiv 3 < R(y_0) \equiv 3.581 < 4 \equiv R(7)$. Επομένως, η εκτίμηση της τιμής $y_0 \equiv E(Y|X=11)$ είναι η

$$\begin{aligned}\hat{y}_0 &\equiv \hat{E}(Y|X=11)\\ &= 6.6 + \frac{3.581-3}{4-3}(7-6.6)\\ &= 6.8324.\end{aligned}$$

Λύση με το MINITAB: Η χρήση του MINITAB για την εφαρμογή της μεθόδου της μονότονης παλινδρόμησης είναι πολύ απλή. Καταχωρίζουμε τις δύο μεταβλητές σε δύο στήλες (έστω **C1** και **C2**). Σε δύο άλλες στήλες, καταχωρίζουμε τις τάξεις μεγέθους των τιμών των στηλών αυτών. Στην συνέχεια, εργαζόμενοι με τον ήδη γνωστό τρόπο από προηγούμενα παραδείγματα, προχωρούμε σε εκτίμηση της ευθείας παλινδρόμησης των τάξεων μεγέθους της Y πάνω στις τάξεις

μεγέθους της X. Το αποτέλεσμα που προκύπτει δίνεται στον εξής πίνακα αποτελεσμάτων:

```
The regression equation is
RY = 15.9 - 0.982 RX

Predictor        Coef       StDev          T        P
Constant      15.8564      0.5101      31.08    0.000
RX           -0.98205     0.05614     -17.49    0.000

S = 0.9369      R-Sq = 95.9%     R-Sq(adj) = 95.6%
```

Η στήλη **Coef** δίνει τις εκτιμήσεις a και b των συντελεστών α και β της εκτιμηθείσας ευθείας γραμμικής παλινδρόμησης των τάξεων μεγέθους της μεταβλητής Y πάνω στις τάξεις μεγέθους της μεταβλητής **SX**. Παρατηρούμε ότι a=15.856 και b= –0.982.
Το υπόλοιπο της λύσης συνίσταται σε απλούς υπολογισμούς που γίνονται ευκολότερα με τον υπολογιστή (calculator) των Windows.

Λύση με το SPSS: Για την λύση του παραδείγματος με την εφαρμογή της μεθόδου της μονότονης παλινδρόμησης, εργαζόμαστε ως εξής: Στο **syntax window**, καταχωρίζουμε τις τιμές των δύο μεταβλητών σε δύο μεταβλητές του SPSS (έστω **x** και **y**). Στην συνέχεια, εκτελούμε την εντολή

```
rank variables=x y.
```

Με την εκτέλεσή της, δημιουργούνται αυτόματα οι μεταβλητές **rx** και **ry** που περιέχουν τις τάξεις μεγέθους των παρατηρήσεων των **x** και **y**, αντίστοιχα. Η εφαρμογή (κατά τον ήδη γνωστό τρόπο) της μεθόδου της γραμμικής παλινδρόμησης των τιμών της μεταβλητής **ry** πάνω στις τιμές της μεταβλητής **rx**, οδηγεί στα εξής αποτελέσματα:

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	15.856	.510		31.084	.000
	RANK of X	-.982	.056	-.979	-17.493	.000

a Dependent Variable: RANK of Y

Η στήλη **Unstandardized Coefficients** δίνει τις εκτιμήσεις a και b των συντελεστών α και β της ευθείας γραμμικής παλινδρόμησης των τάξεων μεγέθους της μεταβλητής Y πάνω στις τάξεις μεγέθους της μεταβλητής **SX**. Παρατηρούμε ότι a=15.856 και b= –0.982.

Οι υπόλοιποι υπολογισμοί που αφορούν την εκτίμηση της τιμής y_0=E(Y/X=11) γίνονται πιο εύκολα με τον υπολογιστή (calculator) των Windows.

Λύση με το SAS: Στο παράθυρο εντολών, πληκτρολογούμε τις εξής εντολές:

```
data regres;

input x y @@;

cards;
0 18.4 0 17.4 1 16.2 1.6 16.4 3 14.4 4 10.5 4 11.2 5 10.8 6.5 9 8 8.4 8
7
10 7.2 12 6.6 12.6 5 14 5.6
;
run;
PROC RANK TIES= MEAN;
          VAR x y;
      RANKS rx ry;
run;

proc reg;
model ry=rx;
run;
```

Η εντολή **proc rank** δίνει τις τάξεις μεγέθους των μεταβλητών **x** και **y** (όπως δηλώνονται με την υποεντολή **VAR x y;**) και τις αποθηκεύει σε νέες μεταβλητές με ονόματα **rx** και **ry** (υποεντολή **RANKS rx ry;**).

Με την εντολή **proc reg; model ry=rx;** παίρνουμε την ζητούμενη εκτίμηση της ευθείας παλινδρόμησης.
Ο πίνακας που ακολουθεί περιέχει τα αποτελέσματα.

```
Model: MODEL1
Dependent Variable: RY         RANK FOR VARIABLE Y

                                    Parameter Estimates

                                Parameter      Standard    T for H0:
             Variable  DF      Estimate         Error   Parameter=0
Prob > |T|

             INTERCEP   1     15.856373    0.51011179        31.084
0.0001
             RX         1     -0.982047    0.05613878       -17.493
0.0001

                          Variable
             Variable  DF     Label

             INTERCEP   1  Intercept
             RX         1  RANK FOR VARIABLE X
```

Η εκτίμηση της σταθεράς α δίνεται στο κελλί που αντιστοιχεί στην στήλη **Parameter estimate** και στην γραμμή **INTERCEP** του πίνακα και ισούται με 15.85637. Η εκτίμηση της κλίσης β της ευθείας παλινδρόμησης δίνεται στο κελλί που αντιστοιχεί στην στήλη **Parameter estimate** και στην γραμμή **RX** και ισούται με –0.982047.

7.2.1 Εκτίμηση της Καμπύλης Παλινδρόμησης της Μεταβλητής Y πάνω στην Μεταβλητή X

Για να προσδιορισθεί η καμπύλη παλινδρόμησης, η οποία αποτελείται από όλα τα ζεύγη σημείων τα οποία μπορούν να προσδιορισθούν με τον τρόπο που περιγράψαμε παραπάνω, συνήθως ακολουθείται η εξής διαδικασία:

1. Προσδιορίζουμε τα άκρα της καμπύλης παλινδρόμησης χρησιμοποιώντας τις τιμές $X^{(1)}$ και $X^{(n)}$ στην παραπάνω διαδικασία για να υπολογίσουμε τις τιμές $\hat{E}\left(Y|X = X^{(1)}\right)$ και $\hat{E}\left(Y|X = X^{(n)}\right)$.

2. Από την εκτιμηθείσα εξίσωση παλινδρόμησης $R(y_0) = \hat{\alpha} + \hat{\beta} R(x_0)$, προσδιορίζουμε, για κάθε τάξη μεγέθους $R(Y_i)$ των μεταβλητών Y_i, i = 1, 2, ..., n, μία εκτίμηση $\hat{R}\left(X_i\right)$ της τάξης μεγέθους της μεταβλητής X_i, i = 1, 2, ..., n:

$$\hat{R}\left(X_i\right) = \left[R\left(Y_i\right) - \hat{\alpha}\right]/\hat{\beta}, \quad i = 1, 2, \ldots, n.$$

3. Μετατρέπουμε κάθε τιμή $R(X_i)$ σε μία εκτίμηση $\hat{X}_i$ με τον τρόπο που περιγράφεται στο βήμα 5 της προηγούμενης διαδικασίας. Συγκεκριμένα:

(α) Αν $\hat{R}\left(X_i\right) = R\left(X_j\right)$ για κάποιο j, τότε θέτουμε $\hat{X}_i = X_j$.

(β) Αν $R\left(X_j\right) < \hat{R}\left(X_i\right) < R\left(X_k\right)$, για κάποια γειτονικά j, k με j< k, τότε χρησιμοποιώντας την μεθοδο της γραμμικής παρεμβολής, έχουμε

$$\hat{X}_i = X_j + \frac{\hat{R}\left(X_i\right) - R\left(X_j\right)}{R\left(X_k\right) - R\left(X_j\right)}\left(X_k - X_j\right).$$

(γ) Αν $\hat{R}\left(X_i\right) < \min\left\{R(X_j),\ j = 1, 2, ..., n\right\}$ ή $\hat{R}\left(X_i\right) > \max\left\{R(X_j),\ j = 1, 2, ..., n\right\}$, τότε δεν επιχειρούμε να προσδιορίσουμε την τιμή της $\hat{X}_i$.

4. Σε ένα σύστημα ορθογωνίων αξόνων, με τις τιμές των X_i να μετριούνται στον οριζόντιο άξονα και τις τιμές των Y_i να

μετριούνται στον κατακόρυφο άξονα, σημειώνουμε τα σημεία $(\hat{X}_i, Y_i)$, i = 1, 2, ..., n, όπως, επίσης και τα σημεία $(X^{(1)}, \hat{E}(Y|X = X^{(1)}))$ και $(X^{(n)}, \hat{E}(Y|X = X^{(n)}))$.

5. Συνδέουμε ανά δύο τα γειτονικά σημεία του βήματος 4 με ευθύγραμμα τμήματα. Η σειρά αυτή των συνδεδεμένων ευθυγράμμων τμημάτων αποτελεί την εκτίμηση της παλινδρόμησης της μεταβλητής Y πάνω στη μεταβλητή X. Η παλινδρόμηση αυτή πρέπει να είναι μονότονη, αύξουσα αν η $\hat{\beta}$ οδηγεί σε εκτίμηση b της κλίσης που είναι θετική και φθίνουσα αν η $\hat{\beta}$ οδηγεί σε εκτίμηση b της κλίσης που είναι αρνητική.

Παράδειγμα 7.2.2: Δεκαεπτά δοχεία με χυμό σταφυλιού εξετάσθηκαν, με σκοπό να μελετηθεί ο χρόνος που απαιτείται για την μετατροπή του χυμού σε κρασί ως συνάρτηση της ποσότητας ζάχαρης που προστέθηκε στον χυμό. Διαφορετικές ποσότητες ζάχαρης, από 0 έως 10 χιλιόγραμμα, προστέθηκαν στα δοχεία, τα οποία ελέγχονταν καθημερινά για να διαπιστωθεί αν η μετατροπή σε κρασί είχε ολοκληρωθεί. Μετά από 30 μέρες, το πείραμα διεκόπη με αποτέλεσμα να μην ολοκληρωθεί η ζύμωση σε τρία δοχεία. Ας υποθέσουμε ότι επιθυμούμε να προσδιορίσουμε μία εκτίμηση της καμπύλης παλινδρόμησης του αριθμού Y των ημερών μέχρι την ολοκλήρωση της ζύμωσης πάνω στην ποσότητα ζάχαρης X.

Οι παρατηρήσεις (X_i, Y_i), οι τάξεις μεγέθους $R(X_i)$ και $R(Y_i)$ και οι τιμές $\hat{R}(X_i)$ και $\hat{X}_i$, i = 1, 2, ..., 17 υπολογίζονται από τα βήματα 2 και 3 της προηγούμενης διαδικασίας και δίνονται στον πίνακα που ακολουθεί. Για τον προσδιορισμό των τιμών $\hat{R}(X_i)$ και

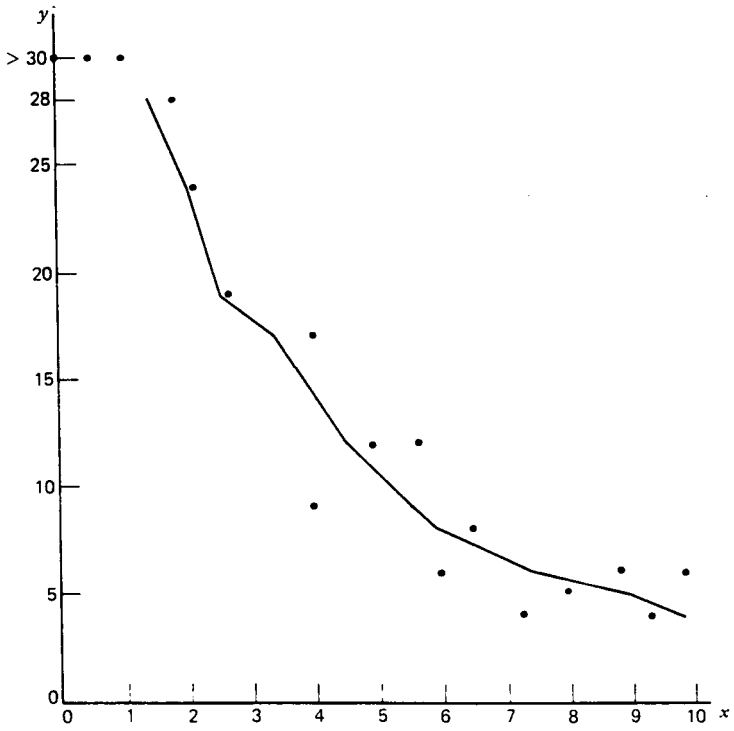
$\hat{X}_i$, i = 1, 2, ..., 17, χρησιμοποιήθηκε η εκτίμηση ελαχίστων τετραγώνων της ευθείας παλινδρόμησης των τάξεων μεγέθους $R(Y_i)$ πάνω στις τάξεις μεγέθους $R(X_i)$:

$$\hat{R}(y_i) = 17.4 - 0.934R(x_i).$$

X_i	Y_i	$R(X_i)$	$R(Y_i)$	$\hat{R}(X_i)$	$\hat{X}_i$
0	> 30	1	16	1.50	0.25
0.5	>30	2	16	1.50	0.25
1.0	> 30	3	16	1.50	0.25
1.8	28	4	14	3.64	1.51
2.2	24	5	13	4.71	2.08
2.7	19	6	12	5.78	2.59
4.0	17	7.5	11	6.85	3.44
4.0	9	7.5	8	10.06	5.62
4.9	12	9	9.5	8.46	4.58
5.6	12	10	9.5	8.46	4.58
6.0	6	11	5	13.28	7.50
6.5	8	12	7	11.13	6.07
7.3	4	13	1.5	17.02	9.80
8.0	5	14	3	15.42	9.01
8.8	6	15	5	13.28	7.50
9.3	4	16	1.5	17.02	9.80
9.8	6	17	5	13.28	7.50

Τα ακραία σημεία προσδιορίζονται με αντικατάσταση των τάξεων μεγέθους $R(X^{(1)}) = 1$ και $R(X^{(n)}) = 17$ στην παραπάνω εξίσωση παλινδρόμησης, η οποία οδηγεί στις τάξεις μεγέθους 16.47 και 1.52 για τις τιμές $\hat{E}\left(Y|X = X^{(1)}\right)$ και $\hat{E}\left(Y|X = X^{(n)}\right)$. Χρησιμοποιώντας την μέθοδο της γραμμικής παρεμβολής μεταξύ διαδοχικών παρατηρήσεων πάνω στην μεταβλητή Y, η τάξη μεγέθους 1.52 μετατρέπεται σε $\hat{E}\left(Y|X = 9.8\right) = 4.01$. Η άλλη τιμή, δηλαδή η τιμή E(Y|X=0), θεωρείται ότι είναι ">30" γιατί η εκτιμηθείσα τάξη μεγέθους 16.47 υπερβαίνει την μέγιστη παρατηρηθείσα τάξη μεγέθους των Y τιμών. Στο γράφημα των σημείων $\left(\hat{X}_i, Y_i\right)$, δεν σημειώνονται τα ακραία σημεία, γιατί η τάξη μεγέθους $\hat{R}\left(X_i\right) = 1.52$ διαφέρει πολύ λίγο από την τιμή 1.50 που είναι η πιο μικρή από τις υπολογισθείσες τάξεις μεγέθους του βήματος 2, με αποτέλεσμα, το σημείο αυτό να βρίσκεται πάνω στο ευθύγραμμο τμήμα που κατασκευάσθηκε. Από την άλλη μεριά, η παρατήρηση ">30" δεν μπορεί να χρησιμοποιηθεί ως άκρο ευθύγραμμου τμήματος.

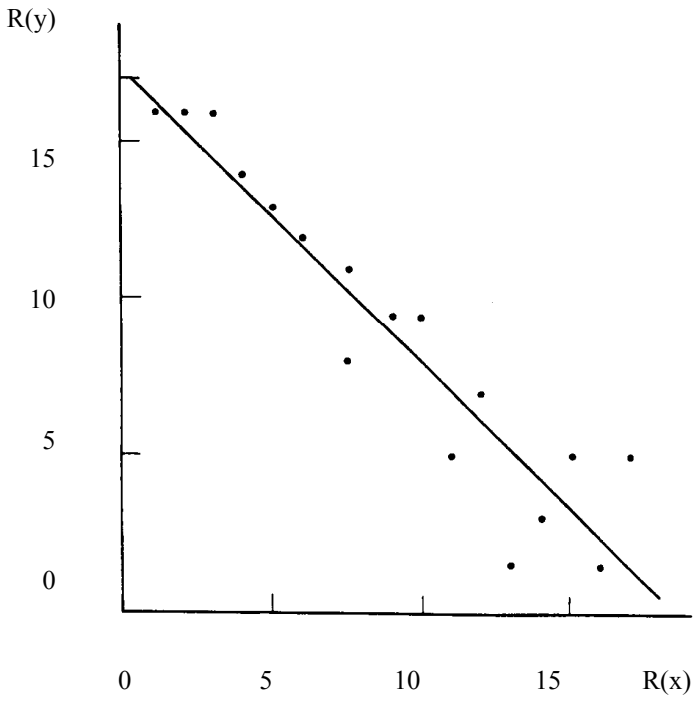
Τα σημεία $\left(\hat{X}_i, Y_i\right)$ σημειώνονται σε ένα σύστημα ορθογωνίων αξόνων, με τις X τιμές να μετριούνται στον οριζόντιο άξονα και τις Y τιμές να μετριούνται στον κατακόρυφο άξονα. Τα γειτονικά διαδοχικά σημεία ενώνονται, στην συνέχεια, με ευθύγραμμα τμήματα οδηγώντας, έτσι, στην εκτίμηση της καμπύλης παλινδρόμησης της τυχαίας μεταβλητής Y πάνω στην τυχαία μεταβλητή X, όπως φαίνεται στο σχήμα που ακολουθεί.



Σχήμα 7.2.1

Εκτιμηθείσα μονότονη καμπύλη παλινδρόμησης του χρόνου μέχρι την ολοκλήρωση της ζύμωσης πάνω στην χρησιμοποιηθείσα ποσότητα ζάχαρης

Είναι ενδιαφέρον να παρατηρηθεί, στο σημείο αυτό, πώς ένα σύνολο παρατηρήσεων με καμπύλη παλινδρόμησης, η οποία είναι προφανώς μη γραμμική, όπως φαίνεται από το σχήμα 7.3.1, μετατρέπεται σε τάξεις μεγέθους, των οποίων η καμπύλη παλινδρόμησης μοιάζει να είναι γραμμική, όπως φαίνεται από το σχήμα 7.3.2.



Σχήμα 7.2.2

Εκτιμηθείσα ευθεία παλινδρόμησης των τάξεων μεγέθους $R(Y_i)$ πάνω στις τάξεις μεγέθους $R(X_i)$, i = 1, 2, …, n

Λύση με το MINITAB: Σε δύο μεταβλητές (έστω **x** και **y**), καταχωρίζουμε τα δύο δείγματα και, σε δύο άλλες (έστω **rx** και **ry**), καταχωρίζουμε τις τάξεις μεγέθους τους. Η εκτίμηση της ευθείας παλινδρόμησης των τιμών της **ry** πάνω στις τιμές της **rx** με το πακέτο δίνει:

```
The regression equation is
ry = 17.4 - 0.934 rx
```

Στην συνέχεια, καταχωρίζουμε τις τιμές της $\hat{R}(X)$ στην μεταβλητή **rxx**. Αυτό γίνεται με την επιλογή **Calc**, **Calculator** και πληκτρολογώντας **(ry–17.4)/(–0.934)** στο πεδίο **Numeric Expression**.

Με το MINITAB, έχουμε δυσκολία να υπολογίσουμε, στην συνέχεια τις τιμές της $\hat{X}$. Αυτό συμβαίνει γιατί, σε αντίθεση με το SPSS, το MINITAB δεν παρέχει στον χρήστη την ευχέρεια να ορίζει με διαφορετικό τρόπο τις τιμές μιας στήλης ανάλογα με τους εκάστοτε περιορισμούς. Για τον λόγο αυτό, οι υπολογισμοί που απομένει να γίνουν για την ολοκλήρωση της λύσης είναι προτιμότερο να γίνουν «με το χέρι».

Λύση με το SPSS: Καταχωρίζουμε στις μεταβλητές **x** και **y**, τα δύο δείγματα και, στις μεταβλητές **rx** και **ry**, τις τάξεις μεγέθους τους. Στην συνέχεια, προχωρούμε στην εκτίμηση της ευθείας παλινδρόμησης των τιμών της μεταβλητής **ry** πάνω στις τιμές της μεταβλητής **rx** με τα εξής αποτελέσματα:

Coefficients[a]

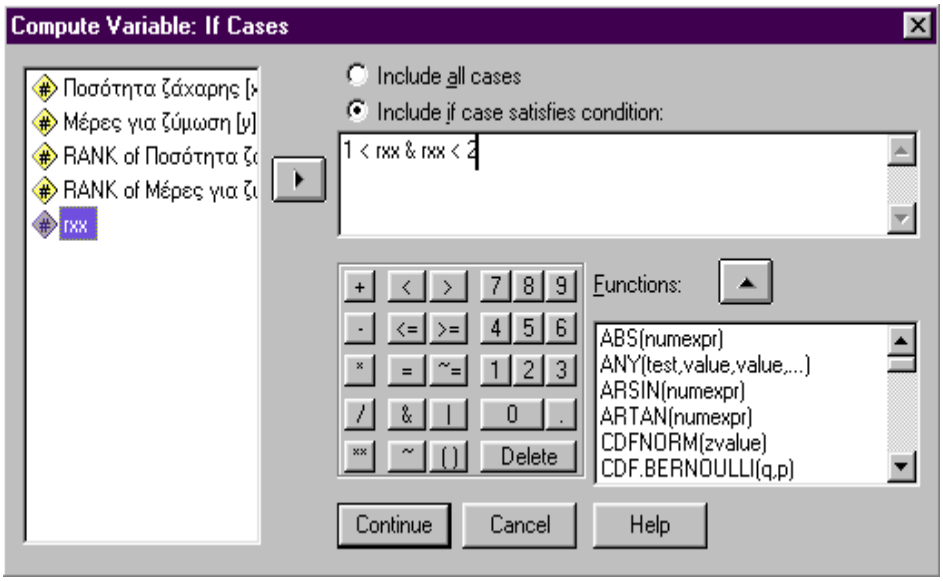
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	17.404	.905		19.227	.000
	RANK of Ποσότητα ζάχαρης	-.934	.088	-.939	-10.569	.000

a Dependent Variable: RANK of Μέρες για ζύμωση

Το επόμενο βήμα είναι να καταχωρίσουμε, στην μεταβλητή **rxx**, τις τιμές της $\hat{R}(X)$. Αυτό γίνεται με την επιλογή **Transform**, **Compute** και πληκτρολογώντας **(ry–17.404)/(–0.934)** στο πεδίο **Numeric Expression**.

Το τελευταίο βήμα είναι να υπολογίσουμε τις τιμές της $\hat{X}$. Αυτό γίνεται με επιλογή **Transform**, **Compute** και χρήση του πλήκτρου **If** ώστε να δηλώνουμε το κατάλληλο **expression** ανάλογα με το που βρίσκεται η τιμή της $\hat{R}(X)$. Για παράδειγμα, ξέρουμε ότι αν

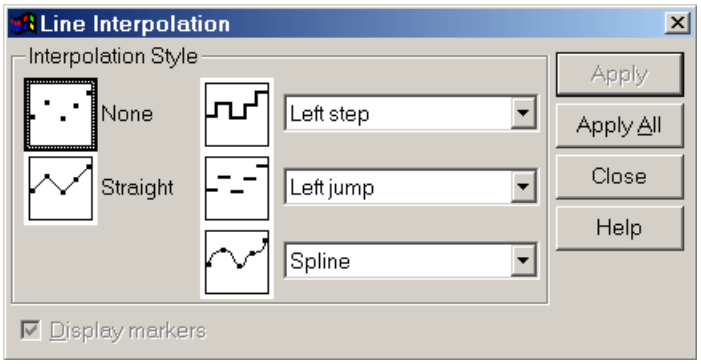
$R(X_1) < \hat{R} < R(X_2)$ τότε $\hat{X} = X_1 + \frac{\hat{R} - R(X_1)}{R(X_2) - R(X_1)}(X_2 - X_1)$. Αυτό επιτυγχάνεται με την επιλογή **Transform**, **Compute**. Δηλώνουμε το όνομα της μεταβλητής στόχου και ως **expression** δίνουμε **0+(0.5–0)*(rxx–1)/(2–1)**. Πιέζοντας **If**, προκύπτει το εξής πλαίσιο διαλόγου:



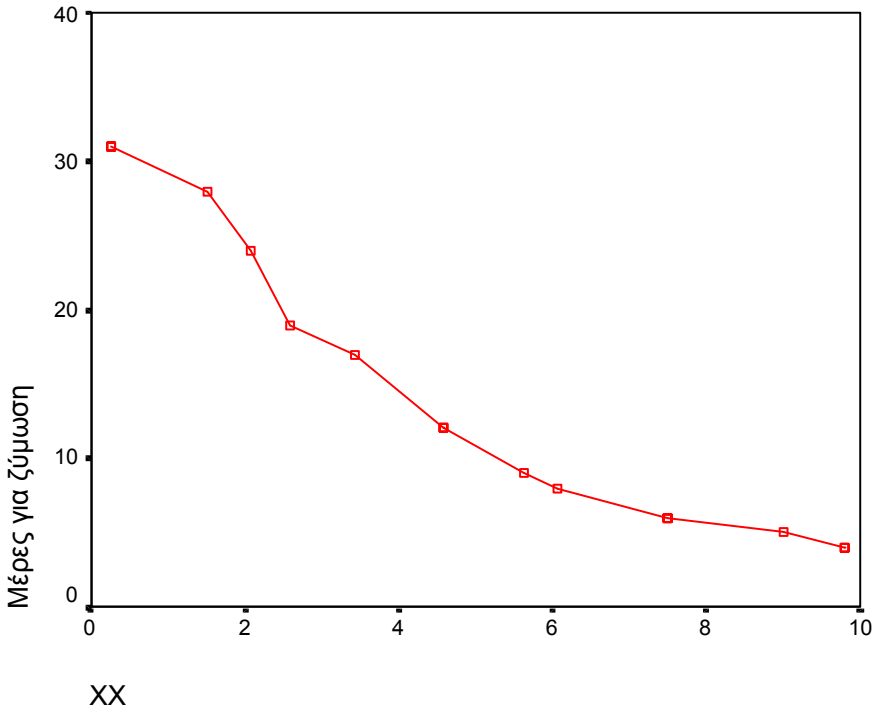
Στο πλαίσιο **Include if case satisfies condition**, δηλώνουμε για ποιες τιμές της **rxx** θα γίνεται ο υπολογισμός που δηλώσαμε στο πεδίο **expression**. Η έκφραση που δώσαμε δηλώνει την συνθήκη $R(X_1) < \hat{R} < R(X_2)$. Εργαζόμενοι με τον ίδιο τρόπο για τους υπόλοιπους υπολογισμούς, δημιουργείται η μεταβλητή **xx** που περιέχει τις τιμές της $\hat{X}$.

Απομένει να κατασκευασθεί το γράφημα της Y έναντι της $\hat{X}$. Ταξινομούμε τις τιμές της μεταβλητής **xx** κατ' αύξουσα τάξη μεγέθους και, στην συνέχεια, κατασκευάζουμε ένα διάγραμμα διασποράς των τιμών της **y** έναντι των τιμών της **xx**. Με διπλό κλικ στο εμφανιζόμενο

γράφημα, οδηγούμεθα στον **Chart editor**. Επιλέγοντας **Format**, **Interpolation** εμφανίζεται το εξής πλαίσιο διαλόγου:



Επιλέγουμε **Straight** και πιέζοντας **Apply All**, **Close** προκύπτει το ζητούμενο γράφημα.

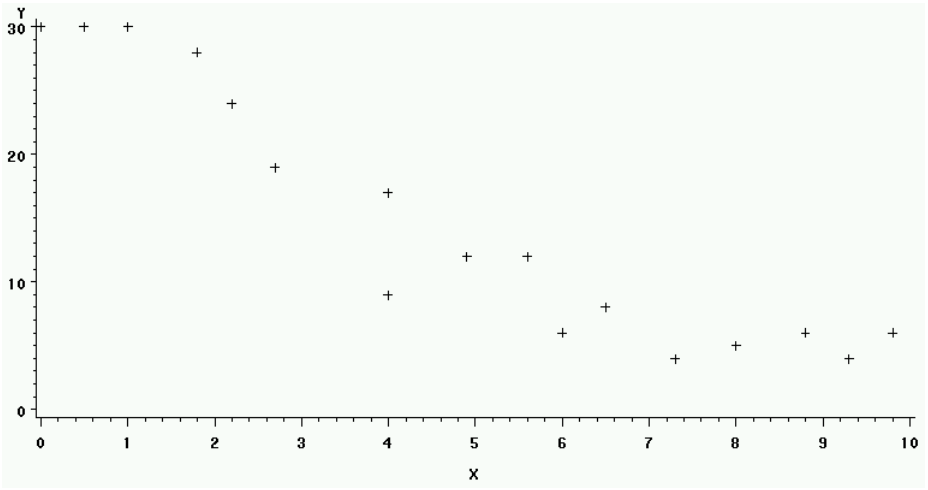


Αυτό ολοκληρώνει την διαδικασία.

Λύση με το SAS: Ένα διάγραμμα διασποράς των δύο μεταβλητών από το οποίο μπορεί να διαφανεί η μορφή της σχέσης που τις συνδέει, μπορεί να κατασκευασθεί με τις εξής εντολές:

```
data regres;

input x y @@;

cards;
0 30 0.5 30 1 30 1.8 28 2.2 24 2.7 19 4 17 4 9 4.9 12 5.6 12
6 6 6.5 8 7.3 4 8 5 8.8 6 9.3 4 9.8 6
;
run;

proc gplot;
 plot y*x;
run;
```

Το προκύπτον διάγραμμα δίνεται στο παρακάτω γράφημα.



Οπως και στο προηγούμενο παράδειγμα, οι εντολές που μπορούν να χρησιμοποιηθούν προκειμένου να εκτιμήσουμε την ευθεία παλινδρόμησης, είναι οι εξής:

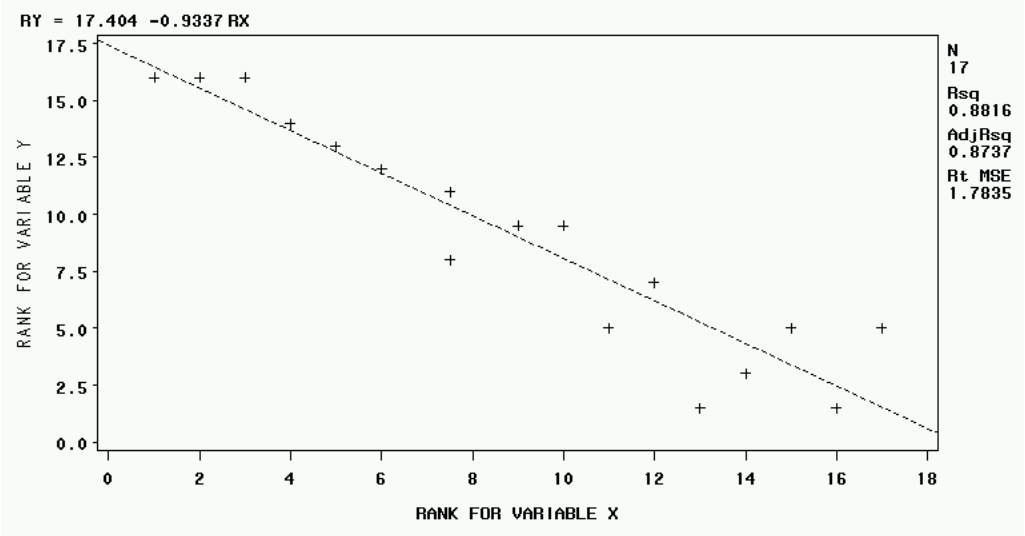
```
PROC RANK TIES= MEAN;
          VAR x y;
      RANKS rx ry;
run;

proc reg graphics;
model ry=rx;
plot ry*rx='*';
run;
```

Τα αποτελέσματα περιέχονται στον πίνακα που ακολουθεί.

```
Model: MODEL1
```

```
Dependent Variable: RY         RANK FOR VARIABLE Y

                    Parameter Estimates
                 Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

INTERCEP   1     17.403681    0.90517768        19.227        0.0001
RX         1     -0.933742    0.08834867       -10.569        0.0001

                      Variable
          Variable  DF     Label

          INTERCEP   1  Intercept
          RX         1  RANK FOR VARIABLE X
```

Οι παραπάνω εντολές, παρέχουν ταυτόχρονα το παρακάτω διάγραμμα στο οποίο εικονίζεται το διάγραμμα διασποράς των τάξεων μεγέθους και το γράφημα της εκτιμηθείσας ευθείας ελαχίστων τετραγώνων $\hat{R}(y_i) = 17.4037 - 0.9337R(x_i)$.



ΑΣΚΗΣΕΙΣ

1. Ο πίνακας που ακολουθεί περιέχει στοιχεία για την ποσοστιαία αύξηση των δαπανών διαφήμισης, X, δέκα εταιρειών και για την ποσοστιαία αύξηση των πωλήσεων, Y, για την περυσινή χρονιά, σε σύγκριση με την χρονιά που προηγήθηκε.

	Εταιρεία									
	1	2	3	4	5	6	7	8	9	10
Χ (διαφήμηση)	4	62	31	-11	47	88	16	-1	74	21
Υ (πωλήσεις)	10	33	39	-14	37	39	18	-8	45	33

α) Από το γράφημα των παραπάνω σημείων, θα μπορούσατε να πείτε ότι η αναμενόμενη τιμή της ποσοστιαίας αύξησης των πωλήσεων μοιάζει να είναι γραμμική συνάρτηση της ποσοστιαίας αύξησης των δαπανών διαφήμισης; Μία μονότονη συνάρτηση;

β) Να εκτιμηθεί η αναμενόμενη ποσοστιαία αύξηση των πωλήσεων για μία αύξηση 25% στις δαπάνες διαφήμισης.

γ) Να εκτιμηθεί η παλινδρόμηση της τυχαίας μεταβλητής Υ πάνω στην τυχαία μεταβλητή Χ. Να κατασκευασθεί το γράφημα της καμπύλης που εκτιμήσατε στην ερώτηση (α) και να σχολιάσετε την εικόνα που δημιουργείται.

2. Στην βιολογική έρευνα και στην έρευνα στην περιοχή της φαρμακευτικής βιομηχανίας, παρουσιάζουν μεγάλο ενδιαφέρον καμπύλες που εκφράζουν την σχέση μεταξύ της δοσολογίας ενός φαρμάκου και της αντίδρασης σε αυτό. Ας υποθέσουμε ότι ένα συγκεκριμένο φάρμακο δίνεται σε δόση Χ (σε χιλιοστόλιτρα) σε κάποια πειραματόζωα, για να εξετασθεί η μορφή της αντίδρασής τους σ' αυτό (καρκίνος, διαβήτης κ.λ.π.). Δέκα διαφορετικές δόσεις του φαρμάκου δίνονται σε πέντε πειραματόζωα για ένα διάστημα, στο τέλος του οποίου κατεγράφη το ποσοστό των ζώων, Υ, που παρουσίασαν αντίδραση.

Χ (δόση) :	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Υ (% αντιδράσεων):	0	0	20	0	40	60	40	80	100	100

α) Από το γράφημα των παραπάνω σημείων, θα μπορούσατε να πείτε ότι το αναμενόμενο ποσοστό αντιδράσεων φαίνεται να συνδέεται με την δόση μέσω κάποιας γραμμικής συνάρτησης; Μιάς μονότονης συνάρτησης;

β) Να εκτιμηθεί το αναμενόμενο ποσοστό αντιδράσεων όταν η δόση είναι 3 χιλιοστόλιτρα.

γ) Να επαναληφθεί το ερώτημα (β) όταν η δόση είναι 3.3 χιλιοστόλιτρα.

δ) Να εκτιμηθεί η παλινδρόμηση της τυχαίας μεταβλητής Y πάνω στην τυχαία μεταβλητή X.

3. Μία εταιρεία καλλυντικών ενδιαφέρεται να μελετήσει την μεταβλητότητα που παρατηρείται στις ετήσιες πωλήσεις ενός από τα καλλυντικά που παράγει. Θεωρώντας ότι οι πωλήσεις σε μία πόλη εξαρτώνται κυρίως από το εισόδημα που έχουν οι κάτοικοι της πόλης αυτής, η διεύθυνση της εταιρείας συγκέντρωσε τα ακόλουθα στοιχεία, με βάση ένα τυχαίο δείγμα 10 πόλεων (σε εκατοντάδες χιλιάδες δραχμές).

Πόλη (i)	1	2	3	4	5	6	7	8	9	10
Ετήσιο "κατά κεφαλή" εισόδημα (X)	15.5	17.0	18.5	20.0	21.0	21.0	215	22.0	23.0	24.0
Ετήσιες "κατά κεφαλή" πωλήσεις (Y)	24.1	23.1	26.1	28.0	26.2	26.1	26.1	31.6	28.3	31.2

Ο στατιστικός αναλυτής της εταιρείας χρησιμοποιώντας την μέθοδο των ελαχίστων τετραγώνων εκτίμησε την εξής ευθεία παλινδρόμησης για τα δεδομένα αυτά: $y = 10.207 + 0.829\ x$

Πώς ερμηνεύετε το αποτέλεσμα αυτό;

(Οικ. Παν/μιο Αθηνών – Εξετ. Φεβρ. 1998)

4. Τα δεδομένα του παρακάτω πίνακα είναι οι ετήσιοι μισθοί και τα χρόνια υπηρεσίας 15 υπαλλήλων μιας εταιρίας.

Χρόνια Υπηρεσίας X	Μισθός (σε δολάρια) Y
7	26075
28	79370
23	65726
18	41983
19	62308
15	41154
24	53610
13	33697
2	22444
8	32562
20	43076
21	56000
18	58667
7	22210
2	20521

Να εκτιμηθεί ο αναμενόμενος μέσος μισθός ενός υπαλλήλου που δουλεύει στην εταιρία αυτή 16 χρόνια.

(Οικ. Παν/μιο Αθηνών – Εξετ. Σεπτ. 2000)

5. Τα παρακάτω ζεύγη τιμών αναφέρονται στους αριθμούς μορίων στις πανελλαδικές εξετάσεις και στους αντίστοιχους βαθμούς πτυχίου ενός τυχαίου δείγματος 10 φοιτητών του τμήματος Στατιστικής.

Βαθμολογία στις Πανελλαδικές	Βαθμός Πτυχίου
5212	6.15
5287	5.9
5310	6.47
5310	6.7
5366	6.9
5400	6.95
5490	7.67
5559	8.83
5800	8.29
5974	9.57

Ενας νεοεισαχθείς φοιτητής του τμήματος Στατιστικής συγκέντρωσε στις πανελλαδικές εξετάσεις 5824 μόρια και ισχυρίζεται ότι όταν

αποφοιτήσει ο χαρακτηρισμός του πτυχίου του θα είναι *άριστα*. Με βάση τους βαθμούς των φοιτητών των παρελθόντων ετών πιστεύετε ότι ο ισχυρισμός του είναι βάσιμος;

(Οικ. Παν/μιο Αθηνών – Εξετ. Φεβρ. 1999)

6. Σε μία οικολογική έρευνα που έκαναν οι J.R. Gat και A. Nissenbaum (National Geographic Research Reports – 1976 Projects, σελ. 413-18) για την συγκέντρωση αμμωνίας σε διαφορετικά βάθη στη Νεκρή Θάλασσα, προέκυψαν τα στοιχεία που συνοψίζονται στον πίνακα που ακολουθεί.

Βάθος (m)	25	50	100	150	155	
Αμμωνία (mg/l)	6.13	5.51	6.18	6.70	7.22	
Βάθος (m)	187	200	237	287	290	300
Αμμωνία (mg/l)	7.28	7.22	7.48	7.38	7.38	7.64

Να εκτιμηθεί μια ευθεία παλινδρόμησης της συγκέντρωσης αμμωνίας πάνω στην μεταβλητή βάθος, χρησιμοποιώντας την μέθοδο ελαχίστων τετραγώνων.

α) Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης για την κλίση της πραγματικής ευθείας παλινδρόμησης.

β) Να ελεγχθεί η υπόθεση H_0: β=1.

(Οικ. Παν/μιο Αθηνών – Εξετ. Ιανουαρίου 1994)

7. Ένας οδηγός κατέγραφε τον αριθμό των χιλιομέτρων που ταξίδευε και την ποσότητα βενζίνης (σε λίτρα) που έβαζε κάθε φορά που αγόραζε βενζίνη.

Χιλιόμετρα (Y)	Λίτρα (X)	Χιλιόμετρα (Y)	Λίτρα (X)
142	11.1	157	12.5

116	5.7	255	17.9
194	14.2	159	8.8
250	15.8	43	3.4
88	7.5	208	15.2

(α) Να κατασκευάσετε ένα διάγραμμα των σημείων (X_i, Y_j), i = 1, 2, ..., 10, με την κατανάλωση βενζίνης να μετριέται στον οριζόντιο άξονα.

(β) Να εκτιμηθεί η ευθεία παλινδρόμησης ελαχίστων τετραγώνων της διανυόμενης απόστασης (Y) πάνω στην κατανάλωση βενζίνης (X).

(γ) Χρησιμοποιώντας το διάγραμμα της ερώτησης (α), να κατασκευάσετε το γράφημα της ευθείας που εκτιμήσατε και να σχολιάσετε την εικόνα που προκύπτει.

(δ) Ενα περιοδικό για αυτοκίνητα ανέφερε ότι, στα σχετικά τεστ που έκανε για τον συγκεκριμένο τύπο αυτοκινήτου, βρήκε ότι η μέση κατανάλωση βενζίνης, εκφραζόμενη ως απόσταση που διανύεται ανά ένα λίτρο βενζίνης για αυτοκίνητα του ιδίου έτους κατασκευής με το συγκεκριμένο, είναι 18 χιλιόμετρα ανά λίτρο. Να ελεγχθεί η μηδενική υπόθεση ότι η τιμή αυτή ισχύει και για το συγκεκριμένο αυτοκίνητο με τον συγκεκριμένο οδηγό. (Να χρησιμοποιηθεί έλεγχος για την κλίση της ευθείας παλινδρόμησης).

(ε) Να κατασκευασθεί ένα 95%-διάστημα εμπιστοσύνης για την κατανάλωση (ως διανυόμενη απόσταση ανά λίτρο βενζίνης) του συγκεκριμένου αυτοκίνητου με τον συγκεκριμένο οδηγό.

8. Ο πίνακας που ακολουθεί δίνει στοιχεία για τον αριθμό των σάπιων φρούτων, Y, που βρέθηκαν σε κάθε ένα από 10 τυχαία επιλεγμένα κιβώτια ενός μεγάλου φορτίου μετά που αυτό

παρέμεινε στην αποθήκη για ένα αριθμό X ημερών. Να χρησιμοποιηθεί η μέθοδος του Theil για να εκτιμηθεί η ευθεία παλινδρόμησης της μεταβλητής Y πάνω στην μεταβλητή X.

Διάρκεια αποθήκευσης X	3	5	8	11	15	18	20	25	27	30
Αριθμός σάπιων φρούτων Y	2	4	7	10	17	23	29	45	59	73

Να παρασταθεί γραφικά η εκτιμηθείσα ευθεία και τα σημεία του παραπάνω πίνακα. Θα μπορούσατε να συμπεράνετε ότι η προσαρμογή μίας ευθείας γραμμής είναι εύλογη; (Να ελέγξετε τα συμπεράσματά σας ελέγχοντας κατάλληλες υποθέσεις για την κλίση της ευθείας παλινδρόμησης).

9. Σε μία οικολογική έρευνα που έκαναν οι J. R Gat και A. Nissenbaum (National Geographic Research Reports - 1976 Projects, σελ. 413-18) για την συγκέντρωση αμμωνίας σε διαφορετικά βάθη στη Νεκρή Θάλασσα, προέκυψαν τα στοιχεία που συνοψίζονται στον πίνακα που ακολουθεί.

Βάθος (m)	25	50	100	150	155	187	200	237
	287	290	300					
Αμμωνία	6.13	5.51	6.18	6.70	7.22	7.28	7.22	7.48
(mg/1)	7.38	7.38	7.64					

Να εκτιμηθεί μία ευθεία παλινδρόμησης της συγκέντρωσης αμμωνίας πάνω στην μεταβλητή βάθος, χρησιμοποιώντας (α) την μέθοδο ελαχίστων τετραγώνων και (β) την μέθοδο του Theil. Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης για την κλίση της πραγματικής ευθείας παλινδρόμησης.

10. Τα στοιχεία που ακολουθούν αναφέρθηκαν από τον S. K. Katti (Biometrics, 1965, τ. 21, σελ. 957-974) σε μία μελέτη για την συσχέτιση του βάρους της τροφής (X) και της αύξησης βάρους (Y) δέκα γουρουνιών στα οποία χορηγήθηκε ένας τύπος τροφής και δέκα άλλων γουρουνιών στα οποία χορηγήθηκε ένας δεύτερος τύπος τροφής.

Τύπος	x:	575	585	628	632	637	638	661	674	694	713
τροφής Α	y:	130	146	156	164	158	151	159	165	167	170

Τύπος	x:	625	646	651	678	710	722	728	754	763	831
τροφής Β	y:	142	164	149	160	184	173	193	189	200	201

Να χρησιμοποιηθεί η μέθοδος του Theil για την εκτίμηση μιας ευθείας παλινδρόμησης της αύξησης του βάρους πάνω στο βάρος της τροφής, για κάθε τύπο τροφής. Να κατασκευασθούν 95% και 99%-διαστήματα εμπιστοσύνης για τις κλίσεις των δύο πραγματικών ευθειών παλινδρόμησης.