

# Κεφάλαιο 10

Γραμμική συσχέτιση – Γραμμική  
παλινδρόμηση

# Συνδιασπορά

- Ορίζουμε ως δειγματική Συνδιασπορά ή Συνδιακύμανση των ποσοτικών μεταβλητών  $X$  και  $Y$  το στατιστικό
- $$s_{XY} = \frac{1}{n-1} \left( \sum_1^n x_i y_i - \frac{(\sum_1^n x_i) \cdot (\sum_1^n y_i)}{n} \right)$$
- Μοιάζει με τον υπολογιστικά εύκολο τύπο της διασποράς μιας μεταβλητής  $X$ , αλλάζοντας όμως το  $\sum_1^n x_i^2$  με  $\sum_1^n x_i y_i$  και το  $(\sum_1^n x_i)^2$  με  $(\sum_1^n x_i) \cdot (\sum_1^n y_i)$ .
- Η συνδιασπορά μπορεί να έχει και αρνητική τιμή. Εκφράζει τη διάχυση του νέφους των σημείων αλλά και την κλίση του.
- Η τιμή της συνδιασποράς μπορεί να είναι μεγάλη ή μικρή, ή μηδενική, αλλά σε κάθε περίπτωση δεν είναι εύκολο να αξιολογήσουμε το μέγεθός της.

# Συντελεστής γραμμικής συσχέτισης

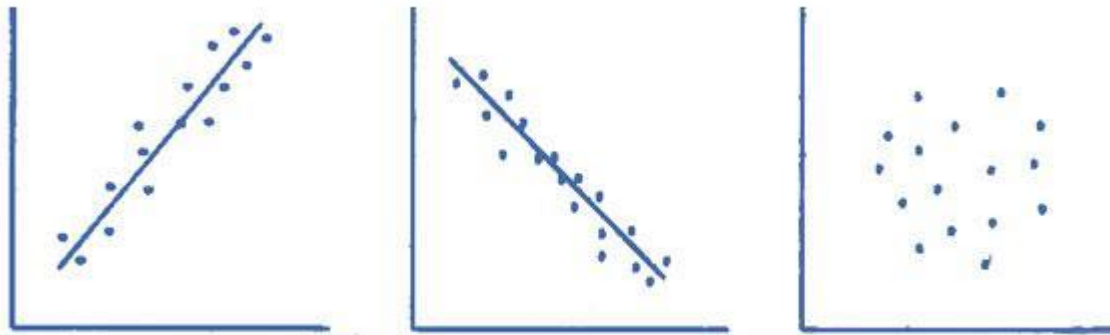
- Αν υπολογίσουμε και τις διασπορές των  $X$  και  $Y$
- 
- $s_X^2 = \frac{1}{n-1} \left( \sum_1^n x_i^2 - \frac{(\sum_1^n x_i)^2}{n} \right)$
- 
- $s_Y^2 = \frac{1}{n-1} \left( \sum_1^n y_i^2 - \frac{(\sum_1^n y_i)^2}{n} \right)$
- 
- μπορούμε τώρα να συνδυάσουμε τα δύο παραπάνω με τη συνδιασπορά. Ορίζουμε το *συντελεστή γραμμικής συσχέτισης του Pearson*,  $r$  ως
- 
- $r = \frac{s_{XY}}{s_X s_Y}$

# Συντελεστής γραμμικής συσχέτισης

- Γενικά, ο συντελεστής συσχέτισης παίρνει τιμές από  $-1$  έως και  $1$ , αποτελεί μια σταθμισμένη εκδοχή της συνδιασποράς, και ανάλογα με το πρόσημο έχουμε αρνητική ή θετική συσχέτιση, ή μηδενική συσχέτιση αν ισούται με μηδέν.
- Όταν πλησιάζει στο  $-1$  ή στο  $1$  έχουμε γραμμικό σχήμα στο νέφος των σημείων του διαγράμματος διασποράς και αυτό σημαίνει ότι η αύξηση της μίας μεταβλητής συνοδεύεται από μείωση ή από αύξηση της άλλης, αντίστοιχα.
- Ανάλογα με το μέγεθος και το πρόσημο του συντελεστή μπορούμε να έχουμε μια περιγραφή του διαγράμματος διασποράς των  $X$  και  $Y$  (προσέξτε, μόνο περιγραφή έχουμε). Για παράδειγμα, ένας συντελεστής  $-0,86$  δηλώνει αρνητική κλίση του νέφους και σχεδόν γραμμικό σχήμα. Το πρώτο δηλώνεται από το πρόσημο και το δεύτερο από τον αριθμό που πλησιάζει στη μονάδα (γραμμικότητα). Ένας συντελεστής συσχέτισης με τιμή  $0,86$  σχολιάζεται ως υψηλή θετική συσχέτιση, ενώ αν έχει τιμή  $-0,86$  λέμε ότι έχουμε υψηλή αρνητική συσχέτιση.
- Ο συντελεστής συσχέτισης θα πρέπει να είναι μεγαλύτερος από  $0,5$  ή μικρότερος από  $-0,5$  για να έχουμε υψηλή (θετική ή αντίστοιχα αρνητική) συσχέτιση. Το όριο συχνά ανεβαίνει στο  $0,6$  ή  $0,7$  (αντίστοιχα κατεβαίνει στο  $-0,6$  και  $-0,7$ ) ανάλογα με τις εφαρμογές και τις χρήσεις των συντελεστών.

# Συντελεστής γραμμικής συσχέτισης

Διαγράμματα διασποράς με παραδείγματα θετικής, αρνητικής και μηδενικής συσχέτισης (αντίστοιχα από αριστερά προς τα δεξιά). Οι ευθείες γραμμές έχουν προστεθεί για να καταδείξουν τη γραμμικότητα.



# Συντελεστής γραμμικής συσχέτισης

- Ο συντελεστής συσχέτισης πρέπει να υπολογίζεται και να ερμηνεύεται με σύνεση. Πρέπει να υπολογίζεται σε παρατηρήσεις που προέρχονται από τυχαία δείγματα. Αν δεν υπολογίζεται σε τέτοιου είδους παρατηρήσεις μπορεί να καταγράφει συστηματικά λάθη ή δομικές διαφοροποιήσεις (patterns) των δεδομένων σας, όπως διαφοροποιήσεις ανάμεσα σε φύλα, κλπ.
- Επίσης ο συντελεστής γραμμικής συσχέτισης είναι μόνον ένα περιγραφικό στατιστικό και δεν τεκμηριώνει κάποια σχέση. Μια υψηλή θετική ή αρνητική συσχέτιση δεν εκφράζει κατανάγκη κάποια σχέση ανάμεσα στις δύο μεταβλητές  $X$  και  $Y$ , με την έννοια ότι η  $X$  προκαλεί την  $Y$ , αλλά απλά μια συμμεταβολή.

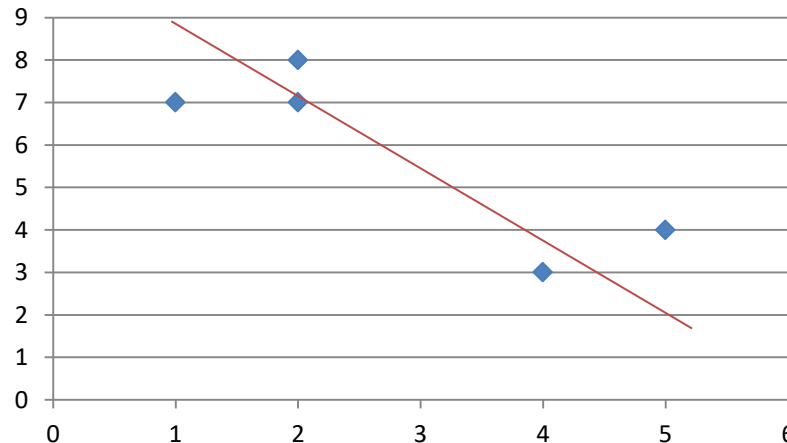
# Παράδειγμα 10.1

- Δίνονται δύο ποσοτικές μεταβλητές  $X$  και  $Y$  με 5 παρατηρήσεις η κάθε μία. Οι παρατηρήσεις συνήθως είναι - και πρέπει να είναι - πολύ περισσότερες, αλλά επιλέγουμε αυτό το παράδειγμα για ευκολία. Στον πίνακα που ακολουθεί περιγράφουμε τις παρατηρήσεις.
- Να υπολογιστεί η συνδιασπορά των  $X$  και  $Y$  και ο συντελεστής γραμμικής συσχέτισης του Pearson.

| $X$ | $Y$ |
|-----|-----|
| 1   | 7   |
| 2   | 7   |
| 2   | 8   |
| 4   | 3   |
| 5   | 4   |

# Παράδειγμα 10.1

- Διάγραμμα διασποράς των  $X$  και  $Y$  (η ευθεία δεν αποτελεί μέρος του διαγράμματος αλλά τοποθετήθηκε για να καταδείξει τη γραμμικότητα).



# Παράδειγμα 10.1

| x             | y             | $x^2$           | $y^2$            | xy              |
|---------------|---------------|-----------------|------------------|-----------------|
| 1             | 7             | 1               | 49               | 7               |
| 2             | 7             | 4               | 49               | 14              |
| 2             | 8             | 4               | 64               | 16              |
| 4             | 3             | 16              | 9                | 12              |
| 5             | 4             | 25              | 16               | 20              |
| $\Sigma x=14$ | $\Sigma y=29$ | $\Sigma x^2=50$ | $\Sigma y^2=187$ | $\Sigma xy =69$ |

# Παράδειγμα 10.1

- $$s_{XY} = \frac{1}{n-1} \left( \sum_1^n x_i y_i - \frac{(\sum_1^n x_i) \cdot (\sum_1^n y_i)}{n} \right) =$$
$$\frac{1}{5-1} \left( 69 - \frac{14 \cdot 29}{5} \right) = -3,05$$
- Η συνδιασπορά είναι αρνητική. Αυτό που γενικά δεν μπορούμε να κάνουμε βλέποντας την τιμή της συνδιασποράς είναι να την αξιολογήσουμε ως προς το μέγεθός της, αν είναι μεγάλη ή μικρή.

# Παράδειγμα 10.1

- $s_{XY} = \frac{1}{n-1} \left( \sum_1^n x_i y_i - \frac{(\sum_1^n x_i) \cdot (\sum_1^n y_i)}{n} \right) = \frac{1}{5-1} \left( 69 - \frac{14 \cdot 29}{5} \right) = -3,05$
- Η συνδιασπορά είναι αρνητική. Αυτό που γενικά δεν μπορούμε να κάνουμε βλέποντας την τιμή της συνδιασποράς είναι να την αξιολογήσουμε ως προς το μέγεθός της, αν είναι μεγάλη ή μικρή.
- Υπολογίζουμε τώρα και τις διασπορές των X και Y
- $s_X^2 = \frac{1}{n-1} \left( \sum_1^n x_i^2 - \frac{(\sum_1^n x_i)^2}{n} \right) = \frac{1}{5-1} \left( 50 - \frac{14^2}{5} \right) = 2,7$
- 
- $s_Y^2 = \frac{1}{n-1} \left( \sum_1^n y_i^2 - \frac{(\sum_1^n y_i)^2}{n} \right) = \frac{1}{5-1} \left( 187 - \frac{29^2}{5} \right) = 4,7$

# Παράδειγμα 10.1

- Οπότε
- $$r = \frac{s_{XY}}{s_X s_Y} = \frac{-3,05}{\sqrt{2,7}\sqrt{4,7}} = -0,86$$
- Ο συντελεστής γραμμικής συσχέτισης του Pearson εκφράζει μια υψηλή αρνητική συσχέτιση.

# Απλή παλινδρόμηση

- Ένα μοντέλο απλής παλινδρόμησης, επιδιώκει να εκφράσει μια εξαρτημένη ποσοτική μεταβλητή  $Y$  ως γραμμική συνάρτηση μιας ανεξάρτητης ποσοτικής μεταβλητής  $X$ . Κάθε μοντέλο παλινδρόμησης είναι μοντέλο αιτίου-αποτελέσματος, δηλαδή θεωρούμε ότι η  $X$  επηρεάζει την  $Y$  και προσπαθούμε να προσδιορίσουμε ένα γραμμικό μοντέλο που να περιγράφει πώς την επηρεάζει.
- Αν είχαμε περισσότερες ανεξάρτητες μεταβλητές δεν θα μιλούσαμε πια για απλή, αλλά για πολλαπλή παλινδρόμηση.

# Απλή παλινδρόμηση

- $Y = b_0 + b_1X$
- Το  $b_0$  ονομάζεται σταθερός όρος ή σταθερά του μοντέλου και το  $b_1$  συντελεστής της  $X$ . Να θυμάστε ότι ο  $b_1$  είναι αυτός που ενδιαφέρει ιδιαίτερα γιατί εκφράζει πόσο η  $X$  επηρεάζει την  $Y$ .
- Οι συντελεστές του μοντέλου υπολογίζονται ως εξής
- $\hat{b}_1 = \frac{s_{XY}}{s_X^2}$
- $\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}$

# Υπολογισμός των εκτιμώμενων τιμών της Y

- έχουμε τη δυνατότητα να υπολογίσουμε τις εκτιμήσεις της Y που παράγονται αν στο παρακάτω μοντέλο αντικαταστήσουμε τις τιμές της X:
- 
- $\hat{y} = \hat{b}_0 + \hat{b}_1 x$
- 
- Μπορούμε να αντικαταστήσουμε διαδοχικά τις πέντε τιμές της X του Παραδείγματος 10.1 στο μοντέλο, και να πάρουμε τα εκτιμώμενα y από το μοντέλο. Αυτά συμβολίζονται με  $\hat{y}$  και δεν είναι κατ' ανάγκη ίσα με τα πραγματικά y, αλλά όσο πιο κοντά τους είναι, τόσο καλύτερη είναι η γραμμική παλινδρόμηση. Γενικά υπολογίζονται ως εξής:
- 
- $\hat{y}_1 = \hat{b}_0 + \hat{b}_1 x_1$
- $\hat{y}_2 = \hat{b}_0 + \hat{b}_1 x_2$
- $\hat{y}_3 = \hat{b}_0 + \hat{b}_1 x_3$
- .
- .
- .
- $\hat{y}_{n-1} = \hat{b}_0 + \hat{b}_1 x_{n-1}$
- $\hat{y}_n = \hat{b}_0 + \hat{b}_1 x_n$

# Σφάλματα της παλινδρόμησης

- Τα σφάλματα  $e$  της παλινδρόμησης είναι οι διαφορές των πραγματικών  $y$  μείον τις τιμές της ευθείας παλινδρόμησης. Στην πράξη, υπολογίζονται ως διαφορές των  $y$  μείον τις εκτιμήσεις  $\hat{y}$  (δηλαδή τις τιμές που υπολογίσαμε με βάση τα  $x$  και χρησιμοποιώντας το μοντέλο) και τότε τα σφάλματα, όταν υπολογίζονται εμπειρικά, λέγονται και υπόλοιπα ή κατάλοιπα.
- $e_i = y_i - \hat{y}_i$
- Το άθροισμά τους θεωρητικά ισούται με μηδέν. Στη συνέχεια υπολογίζουμε το άθροισμα των τετραγώνων τους
- $\sum_{i=1}^n e_i^2$
- Αυτό είναι το μικρότερο δυνατόν επειδή επιλέξαμε την ευθεία ελαχίστων τετραγώνων, που ελαχιστοποιεί το άθροισμα τετραγώνων των αποστάσεων των  $y$  από τις εκτιμήσεις τους. Αυτή είναι η τιμή των ελαχίστων τετραγώνων (σαν το νούμερο  $2^2+2^2=8$  στο παράδειγμα με τον καθηγητή και τους δύο φοιτητές).

# Συντελεστής προσδιορισμού

- $$R^2 = \frac{SS_R}{SS_T} = \frac{\sum_1^n (\hat{y}_i - \bar{y})^2}{\sum_1^n (y_i - \bar{y})^2} = \frac{\sum \hat{y}_i^2 - \frac{(\sum \hat{y}_i)^2}{n}}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}$$
- *Ειδικά και μόνο για μοντέλα απλής παλινδρόμησης (με μία ανεξάρτητη μεταβλητή), ο συντελεστής προσδιορισμού ισούται με τον συντελεστή συσχέτισης των X, Y στο τετράγωνο*
- $R^2 = r^2$

# Τυπικό σφάλμα εκτίμησης

- Υπολογίζουμε το τυπικό σφάλμα εκτίμησης ως

- $$s = \sqrt{\frac{\sum_1^n (y_i - \hat{y})^2}{n-2}} = \sqrt{\frac{\sum_1^n e_i^2}{n-2}}$$

- Αποτελεί την τυπική απόκλιση των σφαλμάτων. Οι βαθμοί ελευθερίας στην περίπτωση της απλής παλινδρόμησης είναι  $n - 2$  ενώ γενικά όταν έχουμε  $k$  ανεξάρτητες μεταβλητές είναι  $n - k - 1$ .

# Έλεγχος σημαντικότητας των συντελεστών των ανεξάρτητων μεταβλητών

- Η διασπορά του συντελεστή  $b_1$  της  $X$  είναι
- $$s_{b_1}^2 = \frac{s^2}{(n-1)s_X^2}$$
- Υπολογίζουμε λοιπόν το στατιστικό
- $$t_{b_1} = \frac{\hat{b}_1}{s_{b_1}}$$
- και αν η απόλυτη τιμή αυτής της ποσότητας είναι μεγαλύτερη από την κρίσιμη τιμή της κατανομής Student  $t_{n-2;0,025}$  συμπεραίνουμε ότι ο συντελεστής της  $X$  είναι στατιστικά σημαντικός σε επίπεδο σημαντικότητας  $\alpha=0,05=5\%$ .

# Παράδειγμα 10.2

- Ας συνεχίσουμε τους υπολογισμούς για το μοντέλο παλινδρόμησης χρησιμοποιώντας τα δεδομένα του προηγούμενου παραδείγματος.

| x     | y     | x <sup>2</sup>      | y <sup>2</sup>       | xy      |
|-------|-------|---------------------|----------------------|---------|
| 1     | 7     | 1                   | 49                   | 7       |
| 2     | 7     | 4                   | 49                   | 14      |
| 2     | 8     | 4                   | 64                   | 16      |
| 4     | 3     | 16                  | 9                    | 12      |
| 5     | 4     | 25                  | 16                   | 20      |
| Σx=14 | Σy=29 | Σx <sup>2</sup> =50 | Σy <sup>2</sup> =187 | Σxy =69 |

## Παράδειγμα 10.2

- $$s_{XY} = \frac{1}{n-1} \left( \sum_1^n x_i y_i - \frac{(\sum_1^n x_i) \cdot (\sum_1^n y_i)}{n} \right) =$$
$$\frac{1}{5-1} \left( 69 - \frac{14 \cdot 29}{5} \right) = -3,05$$
- $$s_X^2 = \frac{1}{n-1} \left( \sum_1^n x_i^2 - \frac{(\sum_1^n x_i)^2}{n} \right) = \frac{1}{5-1} \left( 50 - \frac{14^2}{5} \right) = 2,7$$
- $$\hat{b}_1 = \frac{s_{XY}}{s_X^2} = \frac{-3,05}{2,7} = -1,13$$

# Παράδειγμα 10.2

- $\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} = \frac{29}{5} - (-1,13) \frac{14}{5} = 8,964$
- Άρα το μοντέλο είναι
- $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 8,964 - 1,13X$
- Το  $\hat{Y}$  στο παραπάνω μοντέλο δηλώνει ότι βάσει αυτού, υπολογίζουμε εκτιμήσεις των τιμών της  $Y$  και όχι τις πραγματικές-αρχικές τιμές της. Στη συνέχεια υπολογίζουμε την εκτιμώμενες τιμές της  $Y$  ή τις εκτιμήσεις της  $Y$ , δηλαδή τα  $\hat{y}$  θέτοντας στο μοντέλο τις τιμές τις  $X$ .
- $\hat{y}_1 = \hat{b}_0 + \hat{b}_1 x_1 = 8,964 - 1,13 \cdot 1 = 7,834$
- $\hat{y}_2 = \hat{b}_0 + \hat{b}_1 x_2 = 8,964 - 1,13 \cdot 2 = 6,704$
- $\hat{y}_3 = \hat{b}_0 + \hat{b}_1 x_3 = 8,964 - 1,13 \cdot 2 = 6,704$
- $\hat{y}_4 = \hat{b}_0 + \hat{b}_1 x_4 = 8,964 - 1,13 \cdot 4 = 4,444$
- $\hat{y}_5 = \hat{b}_0 + \hat{b}_1 x_5 = 8,964 - 1,13 \cdot 5 = 3,314$

# Παράδειγμα 10.2

|        | x  | y  | x <sup>2</sup> | xy | $\hat{y}$ | e      | e <sup>2</sup> |
|--------|----|----|----------------|----|-----------|--------|----------------|
|        | 1  | 7  | 1              | 7  | 7,834     | -0,834 | 0,695556       |
|        | 2  | 7  | 4              | 14 | 6,704     | 0,296  | 0,087616       |
|        | 2  | 8  | 4              | 16 | 6,704     | 1,296  | 1,679616       |
|        | 4  | 3  | 16             | 12 | 4,444     | -1,444 | 2,085136       |
|        | 5  | 4  | 25             | 20 | 3,314     | 0,686  | 0,470596       |
| Σύνολα | 14 | 29 | 50             | 69 |           |        | 5,01852        |

## Παράδειγμα 10.2

- $\sum_1^n e_i^2 = 5,01852$
- $R^2 = r^2 = (-0,86)^2 = 0,74 = 74\%$

# Παράδειγμα 10.2

- Στη συνέχεια υπολογίζουμε το τυπικό σφάλμα εκτίμησης

- $$s = \sqrt{\frac{\sum_1^n (y_i - \hat{y})^2}{n-2}} = \sqrt{\frac{5,01852}{5-2}} = 1,29$$

- 

- Και στη συνέχεια υπολογίζουμε τη διασπορά του συντελεστή της X

- 

- $$s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2} = \frac{1,29^2}{(5-1)2,7} = 0,15$$

# Παράδειγμα 10.2

- Για το T-Test για το  $b_1$  υπολογίζουμε
- $t_{b_1} = \frac{\hat{b}_1}{s_{b_1}} = \frac{-1,13}{\sqrt{0,15}} = -2,92$
- Η απόλυτη τιμή αυτού συγκρίνεται με την τιμή
- $t_{n-2;0,025} = t_{5-2;0,025} = t_{3;0,025} = 3,18$
- Επειδή  $|t_{b_1}| = 2,92$  δεν είναι μεγαλύτερο από  $3,18 = t_{n-2;0,025}$  ο συντελεστής  $b_1$  ΔΕΝ είναι στατιστικά σημαντικός. Δηλαδή με βάση τα δεδομένα

# Πολλαπλή παλινδρόμηση

- Θεωρούμε δύο μεταβλητές  $X$  και  $Y$  με παρατηρήσεις που περιγράφονται στον παρακάτω πίνακα. Θα επιχειρήσουμε ένα μεταβατικό στάδιο στο οποίο με τα ίδια δεδομένα της απλής παλινδρόμησης θα εφαρμόσουμε μια νέα τεχνική υπολογισμού των συντελεστών του μοντέλου που κάνει χρήση πινάκων.
- Σημειώστε ότι η τεχνική αυτή εκτιμά μόνο τους συντελεστές του μοντέλου. Μόλις υπολογιστεί το μοντέλο, όλοι οι υπόλοιποι υπολογισμοί (συντελεστής προσδιορισμού, έλεγχος σημαντικότητας των συντελεστών, κλπ), γίνονται κανονικά όπως είδαμε πριν.

# Πολλαπλή παλινδρόμηση

| Χ | Υ |
|---|---|
| 1 | 7 |
| 2 | 7 |
| 2 | 8 |
| 4 | 3 |
| 5 | 4 |

# Πολλαπλή παλινδρόμηση

| x             | y             | $x^2$           | xy              |
|---------------|---------------|-----------------|-----------------|
| 1             | 7             | 1               | 7               |
| 2             | 7             | 4               | 14              |
| 2             | 8             | 4               | 16              |
| 4             | 3             | 16              | 12              |
| 5             | 4             | 25              | 20              |
| $\Sigma x=14$ | $\Sigma y=29$ | $\Sigma x^2=50$ | $\Sigma xy =69$ |

# Πολλαπλή παλινδρόμηση

- Για τα δεδομένα του παραδείγματός μας ορίζουμε τους πίνακες  $X$ ,  $Y$  και  $\hat{B}$  ως εξής:

- $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$

- $Y = \begin{bmatrix} 7 \\ 7 \\ 8 \\ 3 \\ 4 \end{bmatrix}$

- $\hat{B} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix} = (X'X)^{-1}X'Y$

- Δηλαδή δημιουργούμε έναν πίνακα  $X$  με μονάδες στην πρώτη στήλη και τη μεταβλητή  $X$  στην άλλη και έναν πίνακα  $Y$  με τις τιμές της μεταβλητής  $Y$ . Ο πίνακας  $\hat{B}$  περιλαμβάνει τους εκτιμημένους συντελεστές του μοντέλου και το  $(X'X)^{-1}X'Y$  περιγράφει τις πράξεις που πρέπει να κάνουμε με χρήση των πινάκων για να υπολογίζουμε τους συντελεστές του μοντέλου.

# Πολλαπλή παλινδρόμηση

$$\begin{aligned}
 \bullet \quad X'X &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 4 & 5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \\
 &\begin{bmatrix} 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 & 1 \cdot 1 + 1 \cdot 2 + 1 \cdot 2 + 1 \cdot 4 + 1 \cdot 5 \\ 1 \cdot 1 + 1 \cdot 2 + 1 \cdot 2 + 1 \cdot 4 + 1 \cdot 5 & 1 \cdot 1 + 2 \cdot 2 + 2 \cdot 2 + 4 \cdot 4 + 5 \cdot 5 \end{bmatrix} = \\
 &\begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix} = \begin{bmatrix} 5 & 14 \\ 14 & 50 \end{bmatrix}
 \end{aligned}$$

# Πολλαπλή παλινδρόμηση

- Υπολογίζουμε την ορίζουσα του  $X'X$
- $|X'X| = \begin{vmatrix} 5 & 14 \\ 14 & 50 \end{vmatrix} = 5 \cdot 50 - 14 \cdot 14 = 54 \neq 0$
- Παρατηρούμε ότι η ορίζουσα δεν είναι μηδέν. Έτσι μπορούμε να συνεχίσουμε την εφαρμογή. Αν η ορίζουσα ήταν ίση με μηδέν, δεν θα μπορούσαμε να συνεχίσουμε έχοντας μόνον τις εισαγωγικές γνώσεις που παρέχονται εδώ.

# Πολλαπλή παλινδρόμηση

- Υπολογίζουμε τα αλγεβρικά συμπληρώματα του πίνακα  $A = X'X$ . Για κάθε στοιχείο του πίνακα που βρίσκεται στη θέση  $(i, j)$  το αλγεβρικό συμπλήρωμα υπολογίζεται πολλαπλασιάζοντας το  $(-1)^{i+j}$  επί την ελάσσονα που παράγεται αν διαγράψουμε τη στήλη και τη γραμμή που βρίσκεται το συγκεκριμένο στοιχείο.
- $A_{11} = (-1)^{1+1} \cdot 50 = 50$
- $A_{12} = (-1)^{1+2} \cdot 14 = -14$
- $A_{21} = (-1)^{2+1} \cdot 14 = -14$
- $A_{22} = (-1)^{2+2} \cdot 5 = 5$
- Για τον αντίστροφο κατασκευάζουμε έναν πίνακα με τα αλγεβρικά συμπληρώματα σε ανάστροφη τοποθέτηση και διαιρούμε με την ορίζουσα. Να θυμάστε ότι τόσο ο  $X'X$  όσο και ο αντίστροφός του είναι συμμετρικοί.
- $$(X'X)^{-1} = \frac{1}{|X'X|} \cdot \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix} = \frac{1}{54} \cdot \begin{bmatrix} 50 & -14 \\ -14 & 5 \end{bmatrix}$$

# Πολλαπλή παλινδρόμηση

- Υπολογίζουμε τώρα τον  $X'Y$

$$\begin{aligned} X'Y &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 4 & 5 \end{bmatrix} \cdot \begin{bmatrix} 7 \\ 7 \\ 8 \\ 3 \\ 4 \end{bmatrix} = \\ &= \begin{bmatrix} 1 \cdot 7 + 1 \cdot 7 + 1 \cdot 8 + 1 \cdot 3 + 1 \cdot 4 \\ 1 \cdot 7 + 2 \cdot 7 + 2 \cdot 8 + 4 \cdot 3 + 5 \cdot 4 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} = \begin{bmatrix} 29 \\ 69 \end{bmatrix} \end{aligned}$$

# Πολλαπλή παλινδρόμηση

- $$\hat{B} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix} = (X'X)^{-1}X'Y = \frac{1}{54} \cdot \begin{bmatrix} 50 & -14 \\ -14 & 5 \end{bmatrix} \cdot \begin{bmatrix} 29 \\ 69 \end{bmatrix} =$$
$$\frac{1}{54} \cdot \begin{bmatrix} 50 \cdot 29 - 14 \cdot 69 \\ -14 \cdot 29 + 5 \cdot 69 \end{bmatrix} = \begin{bmatrix} \frac{50 \cdot 29 - 14 \cdot 69}{54} \\ \frac{-14 \cdot 29 + 5 \cdot 69}{54} \end{bmatrix} = \begin{bmatrix} 8,96 \\ -1,13 \end{bmatrix}$$

- Δηλαδή το μοντέλο παλινδρόμησης είναι

- $\hat{Y} = 8,96 - 1,13X$

# Πολλαπλή παλινδρόμηση

- Στην πολλαπλή παλινδρόμηση δεν έχουμε μια ευθεία γραμμή, αλλά χρησιμοποιούμε μια γραμμική εξίσωση επιπέδου ή υπερεπιπέδου ανάλογα με το πόσες ανεξάρτητες μεταβλητές έχουμε. Αν έχουμε  $k$  ανεξάρτητες μεταβλητές τότε το μοντέλο πολλαπλής παλινδρόμησης είναι
- $$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k$$

# Παράδειγμα 10.4

- Ας υπολογίσουμε τώρα ένα μοντέλο πολλαπλής παλινδρόμησης κάνοντας χρήση της μεθόδου των πινάκων. Έστω το παρακάτω παράδειγμα με δύο ανεξάρτητες μεταβλητές  $X_1$  και  $X_2$  και την εξαρτημένη μεταβλητή  $Y$ .

| $X_1$ | $X_2$ | $Y$ |
|-------|-------|-----|
| 3     | 3     | 7   |
| 6     | 7     | 8   |
| 4     | 5     | 3   |
| 6     | 5     | 8   |
| 2     | 2     | 3   |

# Παράδειγμα 10.4

$$Y = b_0 + b_1X_1 + b_2X_2$$

| x1      | x2      | y       | (x1) <sup>2</sup>       | (x2) <sup>2</sup>       | x1·x2      | x1·y       | x2·y       |
|---------|---------|---------|-------------------------|-------------------------|------------|------------|------------|
| 3       | 3       | 7       | 9                       | 9                       | 9          | 21         | 21         |
| 6       | 7       | 8       | 36                      | 49                      | 42         | 48         | 56         |
| 4       | 5       | 3       | 16                      | 25                      | 20         | 12         | 15         |
| 6       | 5       | 8       | 36                      | 25                      | 30         | 48         | 40         |
| 2       | 2       | 3       | 4                       | 4                       | 4          | 6          | 6          |
| Σ x1=21 | Σ x2=22 | Σ y =29 | Σ(x1) <sup>2</sup> =101 | Σ(x2) <sup>2</sup> =112 | Σ x1x2=105 | Σ x1y =135 | Σ x2y =138 |

# Παράδειγμα 10.4

- $X = \begin{bmatrix} 1 & 3 & 3 \\ 1 & 6 & 7 \\ 1 & 4 & 5 \\ 1 & 6 & 5 \\ 1 & 2 & 2 \end{bmatrix}$

- $Y = \begin{bmatrix} 7 \\ 8 \\ 3 \\ 8 \\ 3 \end{bmatrix}$

- $\hat{B} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = (X'X)^{-1}X'Y$

# Παράδειγμα 10.4

- $X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 4 & 6 & 2 \\ 3 & 7 & 5 & 5 & 2 \end{bmatrix}$

- $X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 4 & 6 & 2 \\ 3 & 7 & 5 & 5 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 3 & 3 \\ 1 & 6 & 7 \\ 1 & 4 & 5 \\ 1 & 6 & 5 \\ 1 & 2 & 2 \end{bmatrix} =$   
 $\begin{bmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum (x_1)^2 & \sum (x_1)(x_2) \\ \sum x_2 & \sum (x_1)(x_2) & \sum (x_2)^2 \end{bmatrix} = \begin{bmatrix} 5 & 21 & 22 \\ 21 & 101 & 105 \\ 22 & 105 & 112 \end{bmatrix}$

# Παράδειγμα 10.4

- 
- Υπολογίζουμε την ορίζουσα του  $X'X$
- 

$$\begin{aligned} |X'X| &= \begin{vmatrix} 5 & 21 & 22 \\ 21 & 101 & 105 \\ 22 & 105 & 112 \end{vmatrix} = 5 \cdot \begin{vmatrix} 101 & 105 \\ 105 & 112 \end{vmatrix} - 21 \cdot \\ &\begin{vmatrix} 21 & 105 \\ 22 & 112 \end{vmatrix} + 22 \cdot \begin{vmatrix} 21 & 101 \\ 22 & 105 \end{vmatrix} = \\ &5 \cdot (101 \cdot 112 - 105 \cdot 105) - 21 \cdot (21 \cdot 112 - 22 \cdot \\ &105) + 22 \cdot (21 \cdot 105 - 22 \cdot 101) = 179 \neq 0 \end{aligned}$$

# Υπολογίζουμε τα αλγεβρικά συμπληρώματα του πίνακα $A = X'X$

- $A_{11} = (-1)^{1+1} \cdot \begin{vmatrix} 101 & 105 \\ 105 & 112 \end{vmatrix} = 101 \cdot 112 - 105 \cdot 105 = 287$
- $A_{12} = (-1)^{1+2} \cdot \begin{vmatrix} 21 & 105 \\ 22 & 112 \end{vmatrix} = -(21 \cdot 112 - 22 \cdot 105) = -42$
- $A_{13} = (-1)^{1+3} \cdot \begin{vmatrix} 21 & 101 \\ 22 & 105 \end{vmatrix} = 21 \cdot 105 - 22 \cdot 101 = -17$
- $A_{21} = (-1)^{2+1} \cdot \begin{vmatrix} 21 & 22 \\ 105 & 112 \end{vmatrix} = -(21 \cdot 112 - 22 \cdot 105) = -42$
- $A_{22} = (-1)^{2+2} \cdot \begin{vmatrix} 5 & 22 \\ 22 & 112 \end{vmatrix} = 5 \cdot 112 - 22 \cdot 22 = 76$
- $A_{23} = (-1)^{2+3} \cdot \begin{vmatrix} 5 & 21 \\ 22 & 105 \end{vmatrix} = -(5 \cdot 105 - 22 \cdot 21) = -63$
- $A_{31} = (-1)^{3+1} \cdot \begin{vmatrix} 21 & 22 \\ 101 & 105 \end{vmatrix} = 21 \cdot 105 - 22 \cdot 101 = -17$
- $A_{32} = (-1)^{3+2} \cdot \begin{vmatrix} 5 & 22 \\ 21 & 105 \end{vmatrix} = -(5 \cdot 105 - 22 \cdot 21) = -63$
- $A_{33} = (-1)^{3+3} \cdot \begin{vmatrix} 5 & 21 \\ 21 & 101 \end{vmatrix} = 5 \cdot 101 - 21 \cdot 21 = 64$

# Παράδειγμα 10.4

- $(X'X)^{-1} = \frac{1}{|X'X|} \cdot \begin{bmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{bmatrix} = \frac{1}{179} \cdot \begin{bmatrix} 287 & -42 & -17 \\ -42 & 76 & -63 \\ -17 & -63 & 64 \end{bmatrix}$

- Υπολογίζουμε τον πίνακα

- $X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 4 & 6 & 2 \\ 3 & 7 & 5 & 5 & 2 \end{bmatrix} \cdot \begin{bmatrix} 7 \\ 8 \\ 3 \\ 8 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \cdot 7 + 1 \cdot 8 + 1 \cdot 3 + 1 \cdot 8 + 1 \cdot 3 \\ 3 \cdot 7 + 6 \cdot 8 + 4 \cdot 3 + 6 \cdot 8 + 2 \cdot 3 \\ 3 \cdot 7 + 7 \cdot 8 + 5 \cdot 3 + 5 \cdot 8 + 2 \cdot 3 \end{bmatrix} =$

$$\begin{bmatrix} \sum y \\ \sum x_1 \cdot y \\ \sum x_2 \cdot y \end{bmatrix} = \begin{bmatrix} 29 \\ 135 \\ 138 \end{bmatrix}$$

# Παράδειγμα 10.4

$$\begin{aligned} \bullet \quad \hat{B} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} &= (X'X)^{-1}X'Y = \frac{1}{179} \cdot \begin{bmatrix} 287 & -42 & -17 \\ -42 & 76 & -63 \\ -17 & -63 & 64 \end{bmatrix} \cdot \begin{bmatrix} 29 \\ 135 \\ 138 \end{bmatrix} = \\ &= \begin{bmatrix} \frac{287 \cdot 29 - 42 \cdot 135 - 17 \cdot 138}{179} \\ \frac{-42 \cdot 29 + 76 \cdot 135 - 63 \cdot 138}{179} \\ \frac{-17 \cdot 29 - 63 \cdot 135 + 64 \cdot 138}{179} \end{bmatrix} = \begin{bmatrix} 1,715 \\ 1,944 \\ -0,927 \end{bmatrix} \end{aligned}$$

- Δηλαδή το μοντέλο παλινδρόμησης είναι
- $\hat{Y} = 1,715 + 1,944 \cdot X_1 - 0,927 \cdot X_2$

# εκτιμήσεις των $Y$

- $\hat{y}_1 = 1,715 + 1,944x1_1 - 0,927x2_1$
- $\hat{y}_2 = 1,715 + 1,944x1_2 - 0,927x2_2$
- $\hat{y}_3 = 1,715 + 1,944x1_3 - 0,927x2_3$
- $\hat{y}_4 = 1,715 + 1,944x1_4 - 0,927x2_4$
- $\hat{y}_5 = 1,715 + 1,944x1_5 - 0,927x2_5$
- Δηλαδή
- $4,766 = 1,715 + 1,944 \cdot 3 - 0,927 \cdot 3$
- $6,89 = 1,715 + 1,944 \cdot 6 - 0,927 \cdot 7$
- $4,856 = 1,715 + 1,944 \cdot 4 - 0,927 \cdot 5$
- $8,744 = 1,715 + 1,944 \cdot 6 - 0,927 \cdot 5$
- $3,749 = 1,715 + 1,944 \cdot 2 - 0,927 \cdot 2$

# Παράδειγμα 10.4

| x1             | x2             | y              | $\hat{y}$ | $e = y - \hat{y}$ | $e^2$                 | $y^2$            | $\hat{y}^2$               |
|----------------|----------------|----------------|-----------|-------------------|-----------------------|------------------|---------------------------|
| 3              | 3              | 7              | 4,766     | 2,234             | 4,990756              | 49               | 22,71476                  |
| 6              | 7              | 8              | 6,89      | 1,11              | 1,2321                | 64               | 47,4721                   |
| 4              | 5              | 3              | 4,856     | -1,856            | 3,444736              | 9                | 23,58074                  |
| 6              | 5              | 8              | 8,744     | -0,744            | 0,553536              | 64               | 76,45754                  |
| 2              | 2              | 3              | 3,749     | -0,749            | 0,561001              | 9                | 14,055                    |
| $\Sigma x1=21$ | $\Sigma x2=22$ | $\Sigma y =29$ |           |                   | $\Sigma e^2=10,78213$ | $\Sigma y^2=195$ | $\Sigma \hat{y}^2=184,28$ |

# Παράδειγμα 10.4

- $$R^2 = \frac{\sum_1^n (\hat{y}_i - \bar{y})^2}{\sum_1^n (y_i - \bar{y})^2} = \frac{\sum \hat{y}_i^2 - \frac{(\sum \hat{y}_i)^2}{n}}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}} = \frac{184,28 - \frac{29^2}{5}}{195 - \frac{29^2}{5}} = 0,6 = 60\%$$
- Ο συντελεστής  $R^2$  ή συντελεστής προσδιορισμού, παίρνει τιμές από το 0 έως το 1 και εκφράζει (όταν αναφέρεται ως ποσοστό) το ποσοστό της ερμηνείας της εξαρτημένης μεταβλητής από το σύνολο των ανεξάρτητων μεταβλητών. Στην περίπτωσή μας ισούται με 0,6 ή 60%. Και οι δύο ανεξάρτητες μεταβλητές ερμηνεύουν το 60% της εξαρτημένης, ή πιο αυστηρά η διασπορά της Y ερμηνεύεται κατά 60% από τη διασπορά της παλινδρόμησης (τη διασπορά των εκτιμήσεων της Y).

# Παράδειγμα 10.4

- Στη συνέχεια υπολογίζουμε το τυπικό σφάλμα εκτίμησης (προσέξτε ότι διαιρούμε με n-3 δηλαδή με n-k-1 όπου k=2 ο αριθμός των ανεξάρτητων μεταβλητών)
- $\sum_1^n e_i^2 = 10,78213$
- $s = \sqrt{\frac{\sum_1^n (y_i - \hat{y})^2}{n-3}} = \sqrt{\frac{10,78213}{5-3}} = 2,32$
- Στη συνέχεια υπολογίζουμε τις διασπορές  $s_{b_1}^2, s_{b_2}^2$
- $s_{b_1}^2 = v_{11}s^2$
- $s_{b_2}^2 = v_{22}s^2$
- όπου  $v_{ii}$  είναι τα διαγώνια στοιχεία του πίνακα
- $(X'X)^{-1} = \frac{1}{179} \cdot \begin{bmatrix} 287 & -42 & -17 \\ -42 & 76 & -63 \\ -17 & -63 & 64 \end{bmatrix}$

# Παράδειγμα 10.4

- $s_{b_1}^2 = 0,42 \cdot 2,32^2 = 2,26$
- Για το T-Test για το  $b_1$  υπολογίζουμε
- $t_{b_1} = \frac{\hat{b}_1}{s_{b_1}} = \frac{1,944}{\sqrt{2,26}} = 1,29$
- Η απόλυτη τιμή του συγκρίνεται με την τιμή
- $t_{n-3;0,025} = t_{5-3;0,025} = t_{2;0,025} = 4,303$
- Επειδή  $|t_{b_1}| = 1,29 < 4,303 = t_{n-3;0,025}$  ο συντελεστής ΔΕΝ είναι στατιστικά σημαντικός.
- Για το T-Test για το  $b_2$  υπολογίζουμε
- $s_{b_2}^2 = 0,36 \cdot 2,32^2 = 1,94$
- $t_{b_2} = \frac{\hat{b}_2}{s_{b_2}} = \frac{-0,927}{\sqrt{1,94}} = -0,67$
- Η απόλυτη τιμή του συγκρίνεται με την τιμή
- $t_{n-3;0,025} = t_{5-3;0,025} = t_{2;0,025} = 4,303$
- Επειδή  $|t_{b_2}| = 0,67 < 4,303 = t_{n-3;0,025}$  ο συντελεστής είναι ΔΕΝ είναι στατιστικά σημαντικός.  
Η έλλειψη στατιστικής σημαντικότητας οφείλεται σε μεγάλο βαθμό στο μικρό μέγεθος των δειγμάτων που λόγω ευκολίας χρησιμοποιούμε. Για μεγαλύτερα δείγματα μπορεί να έχουμε στατιστικά σημαντικούς συντελεστές.

# Ταξινόμηση των ανεξάρτητων μεταβλητών ως προς την ερμηνευτική τους ικανότητα

- Ο τρόπος που εργαζόμαστε έχει να κάνει με τον υπολογισμό των τυπικών ή τυποποιημένων τιμών των μεταβλητών (z-τιμών). Αυτές τις τιμές τις είδαμε όταν χρησιμοποιούσαμε την Κανονική Κατανομή στο Κεφάλαιο 6. Για να υπολογιστούν, αφαιρούμε από κάθε τιμή μιας μεταβλητής το μέσο όρο της μεταβλητής και στη συνέχεια διαιρούμε με την τυπική της απόκλιση. Το αποτέλεσμα είναι να έχουμε ισάριθμες παρατηρήσεις που δεν έχουν κάποια φυσική υπόσταση είναι δηλαδή καθαροί αριθμοί χωρίς μονάδες μέτρησης, με μέσο όρο μηδέν και τυπική απόκλιση ένα. Ο τύπος που χρησιμοποιούμε είναι
- $$Z = \frac{X - \bar{X}}{s}$$
- όπου  $\bar{X}$  είναι ο μέσος όρος της μεταβλητής  $X$  και  $s$  η τυπική της απόκλιση.

# Ταξινόμηση των ανεξάρτητων μεταβλητών ως προς την ερμηνευτική τους ικανότητα

- Για τα δεδομένα του παραδείγματός μας θα εφαρμόζουμε τον υπολογισμό των z-τιμών στις δύο ανεξάρτητες μεταβλητές μας και στην εξαρτημένη.
- $$ZX1 = \frac{X1 - \bar{x1}}{s_{X1}}$$
- $$ZX2 = \frac{X2 - \bar{x2}}{s_{X2}}$$
- $$ZY = \frac{Y - \bar{y}}{s_Y}$$
- Ισχύει ότι
- Μέσος όρος της  $X1=4,2$  τυπική απόκλιση=1,789
- Μέσος όρος της  $X2=4,4$  τυπική απόκλιση=1,949
- Μέσος όρος της  $Y=5,8$  τυπική απόκλιση=2,588.

# Ταξινόμηση των ανεξάρτητων μεταβλητών ως προς την ερμηνευτική τους ικανότητα

- $ZX1 = \frac{X1-4,2}{1,789}$

- $ZX2 = \frac{X2-4,4}{1,949}$

- $ZY = \frac{Y-5,8}{2,588}$

# Ταξινόμηση των ανεξάρτητων μεταβλητών ως προς την ερμηνευτική τους ικανότητα

| x1 | x2 | y | zx1      | zx2      | zy       | zx1 <sup>2</sup> | zx2 <sup>2</sup> | zx1·zx2  | zx1·zy       | zx2·zy   |
|----|----|---|----------|----------|----------|------------------|------------------|----------|--------------|----------|
| 3  | 3  | 7 | -0,67082 | -0,71818 | 0,4636   | 0,449999         | 0,515783         | 0,48177  | -0,310992152 | -0,33295 |
| 6  | 7  | 8 | 1,00623  | 1,33377  | 0,84993  | 1,012499         | 1,778942         | 1,342079 | 0,855225064  | 1,133611 |
| 4  | 5  | 3 | -0,1118  | 0,30779  | -1,08173 | 0,012499         | 0,094735         | -0,03441 | 0,120937414  | -0,33295 |
| 6  | 5  | 8 | 1,00623  | 0,30779  | 0,84993  | 1,012499         | 0,094735         | 0,309708 | 0,855225064  | 0,2616   |
| 2  | 2  | 3 | -1,22984 | -1,23117 | -1,08173 | 1,512506         | 1,51578          | 1,514142 | 1,330354823  | 1,331794 |
|    |    |   |          |          | σύνολα   | 4,000003         | 3,999974         | 3,613288 | 2,850750213  | 2,061111 |

## Ταξινόμηση των ανεξάρτητων μεταβλητών ως προς την ερμηνευτική τους ικανότητα

- $Z_X = \begin{bmatrix} -0,67082 & -0,71818 \\ 1,00623 & 1,33377 \\ -0,1118 & 0,30779 \\ 1,00623 & 0,30779 \\ -1,22984 & -1,23117 \end{bmatrix}$

- $Z_Y = \begin{bmatrix} 0,4636 \\ 0,84993 \\ -1,08173 \\ 0,84993 \\ -1,08173 \end{bmatrix}$

- $\hat{B} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (Z_X' Z_X)^{-1} Z_X' Z_Y$

# Ταξινόμηση των ανεξάρτητων μεταβλητών ως προς την ερμηνευτική τους ικανότητα

- $Z_X' = \begin{bmatrix} -0,67082 & 1,00623 & -0,1118 & 1,00623 & -1,22984 \\ -0,71818 & 1,33377 & 0,30779 & 0,30779 & -1,23117 \end{bmatrix}$
- $Z_X'Z_X = \begin{bmatrix} -0,67082 & 1,00623 & -0,1118 & 1,00623 & -1,22984 \\ -0,71818 & 1,33377 & 0,30779 & 0,30779 & -1,23117 \end{bmatrix}$
- $\begin{bmatrix} -0,67082 & -0,71818 \\ 1,00623 & 1,33377 \\ -0,1118 & 0,30779 \\ 1,00623 & 0,30779 \\ -1,22984 & -1,23117 \end{bmatrix} = \begin{bmatrix} \sum z_{X1}^2 & \sum z_{X1}z_{X2} \\ \sum z_{X1}z_{X2} & \sum z_{X2}^2 \end{bmatrix} = \begin{bmatrix} 4,000003 & 3,613288 \\ 3,613288 & 3,999974 \end{bmatrix}$
- Υπολογίζουμε την ορίζουσα του  $Z_X'Z_X$
- $|Z_X'Z_X| = \begin{vmatrix} 4,000003 & 3,613288 \\ 3,613288 & 3,999974 \end{vmatrix} = 4,000003 \cdot 3,999974 - 3,613288 \cdot 3,613288 = 2,944 \neq 0$

## Ταξινόμηση των ανεξάρτητων μεταβλητών ως προς την ερμηνευτική τους ικανότητα

- $A_{11} = (-1)^{1+1} \cdot 3,999974 = 3,999974$
- $A_{12} = (-1)^{1+2} \cdot 3,613288 = -3,613288$
- $A_{21} = (-1)^{2+1} \cdot 3,613288 = -3,613288$
- $A_{22} = (-1)^{2+2} \cdot 4,000003 = 4,000003$
- Για τον αντίστροφο κατασκευάζουμε έναν πίνακα με τα αλγεβρικά συμπληρώματα σε ανάστροφη τοποθέτηση και διαιρούμε με την ορίζουσα. Ο  $Z_X'Z_X$  όσο και ο αντίστροφός του είναι συμμετρικοί.
- $$(Z_X'Z_X)^{-1} = \frac{1}{|Z_X'Z_X|} \cdot \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix} =$$
$$\frac{1}{2,944} \cdot \begin{bmatrix} 3,999974 & -3,613288 \\ -3,613288 & 4,000003 \end{bmatrix}$$

# Ταξινόμηση των ανεξάρτητων μεταβλητών ως προς την ερμηνευτική τους ικανότητα

- $$Z_X'Z_Y = \begin{bmatrix} -0,67082 & 1,00623 & -0,1118 & 1,00623 & -1,22984 \\ -0,71818 & 1,33377 & 0,30779 & 0,30779 & -1,23117 \end{bmatrix}.$$

$$\begin{bmatrix} 0,4636 \\ 0,84993 \\ -1,08173 \\ 0,84993 \\ -1,08173 \end{bmatrix} = \begin{bmatrix} \sum z_{X1}z_Y \\ \sum z_{X2}z_Y \end{bmatrix} = \begin{bmatrix} 2,850750213 \\ 2,061111 \end{bmatrix}$$
- Οπότε
- $$\hat{B} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (Z_X'Z_X)^{-1}Z_X'Z_Y = \frac{1}{2,944} \cdot \begin{bmatrix} 3,999974 & -3,613288 \\ -3,613288 & 4,000003 \end{bmatrix}.$$

$$\begin{bmatrix} 2,850750213 \\ 2,061111 \end{bmatrix} = \begin{bmatrix} 1,3435 \\ -0,6984 \end{bmatrix}$$

# Πίνακας ανάλυσης της διασποράς της παλινδρόμησης

- ισχύει η εξής ταυτότητα
- $y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$
- επειδή προσθέτουμε και αφαιρούμε τον όρο  $\hat{y}_i$ .
- Αποδεικνύεται ότι ισχύει και
- $(y_i - \bar{y})^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2$
- Και αν αυτό ισχύει για κάθε  $i$  τότε θα ισχύει και για το σύνολο των παρατηρήσεων-σημείων. Δηλαδή
- $$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e^2$$

# ανάλυση της διασποράς της παλινδρόμησης

- Δηλαδή το άθροισμα των τετραγώνων των  $y_i$  ισούται με το άθροισμα δύο αθροισμάτων τετραγώνων, των  $\hat{y}_i$  και των καταλοίπων. Αυτό γράφεται και ως
- $SS_T = SS_R + SS_e$
- Το  $SS_R$  θα παίξει το ρόλο του  $SS_B$  της γνωστής ως τώρα ανάλυσης διακύμανσης, και το  $SS_e$  θα παίξει το ρόλο της  $SS_W$ .

# ανάλυση της διασποράς της παλινδρόμησης

| Άθροισμα τετραγώνων | Βαθμοί ελευθερίας |
|---------------------|-------------------|
| $SS_T$              | $n-1$             |
| $SS_R$              | $k$               |
| $SS_e$              | $n-k-1$           |

# ανάλυση της διασποράς της παλινδρόμησης

- Υπολογίζουμε τρεις λόγους:
- $\frac{SS_R}{k}$
- 
- που είναι η διασπορά της παλινδρόμησης (των εκτιμήσεων των  $\gamma$ ),
- $\frac{SS_e}{n-k-1}$
- που είναι η διασπορά των σφαλμάτων (είναι το τετράγωνο του τυπικού σφάλματος εκτίμησης), και τέλος, το λόγο
- $F = \frac{\frac{SS_R}{k}}{\frac{SS_e}{n-k-1}}$
- που είναι ο λόγος των δύο διασπορών και κατανέμεται σύμφωνα με την F κατανομή με  $k$  και  $n-k-1$  βαθμούς ελευθερίας. Όσο πιο μεγάλος είναι αυτός ο λόγος, τόσο μπορούμε να είμαστε πιο σίγουροι ότι η παλινδρόμηση είναι «καλή» δηλαδή ερμηνεύει σε σημαντικό βαθμό τη διασπορά των  $\gamma$ . *Ο πίνακας ανάλυσης*

# ανάλυση της διασποράς της παλινδρόμησης

| Αθροίσματα τετραγώνων                      | Βαθμοί<br>ελευθερίας | Μέσα τετράγωνα<br>(διασπορές) | F                                                   |
|--------------------------------------------|----------------------|-------------------------------|-----------------------------------------------------|
| Εξαιτίας της παλινδρόμησης $SS_R$          | k                    | $\frac{SS_R}{k}$              | $F = \frac{\frac{SS_R}{k}}{\frac{SS_e}{n - k - 1}}$ |
| Εξαιτίας των σφαλμάτων (καταλοίπων) $SS_e$ | n-k-1                | $\frac{SS_e}{n - k - 1}$      |                                                     |
| Συνολικά $SS_T$                            | n-1                  |                               |                                                     |

# ανάλυση της διασποράς της παλινδρόμησης

- Οι ποσότητες που μας χρειάζονται υπολογίζονται και με χρήση των παρακάτω υπολογιστικά ευκολότερων τύπων
- $SS_R = \sum \hat{y}_i^2 - \frac{(\sum \hat{y}_i)^2}{n}$
- $SS_T = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$
- $SS_e = SS_T - SS_R = \sum e^2$
- Αλλά πόσο μεγάλο θα πρέπει να είναι το F ώστε να μπορούμε να βγάλουμε το παραπάνω συμπέρασμα; Επειδή το F ακολουθεί την κατανομή  $F_{k,n-k-1}$  αν η τιμή του υπερβεί την κρίσιμη τιμή για κάποιο επίπεδο σημαντικότητας  $\alpha$ , τότε μπορούμε να βγάλουμε ένα τέτοιο συμπέρασμα, γιατί τότε το F θεωρείται μεγάλο. Οπότε υπολογίζουμε το  $F_{k,n-k-1;\alpha}$  (κρίσιμη τιμή) και αν το F που υπολογίσαμε παραπάνω το υπερβαίνει τότε το F είναι μεγάλο και άρα η διασπορά της παλινδρόμησης είναι στατιστικά σημαντικά μεγαλύτερη από τη διασπορά των σφαλμάτων, ή η παλινδρόμηση είναι στατιστικά σημαντική, ή το μοντέλο «προσαρμόζεται καλά» στα δεδομένα. Βέβαια το τελευταίο καλό είναι να συνοδεύεται και από ένα υψηλό συντελεστή προσδιορισμού του μοντέλου,  $R^2$ . Μια στατιστικά σημαντική παλινδρόμηση που συνοδεύεται από υψηλό  $R^2$ , προσαρμόζεται καλά στα δεδομένα.
- Συνήθως το  $\alpha$  είναι ίσο με 0,05=5%. Υπενθυμίζουμε ότι μεγάλο F σημαίνει ότι επειδή ξεπερνά την κρίσιμη τιμή  $F_{k,n-k-1;0,05}$  βάσει της F κατανομής, το συγκεκριμένο F βρίσκεται στο 5% των μεγαλύτερων τιμών.

# Παράδειγμα 10.5

- Ας δούμε τώρα πώς εφαρμόζεται η ανάλυση διακύμανσης στα δεδομένα του Παραδείγματος 10.4 της πολλαπλής

| x1             | x2             | y              | $\hat{y}$ | $e = y - \hat{y}$ | $e^2$             | $y^2$            | $\hat{y}^2$               |
|----------------|----------------|----------------|-----------|-------------------|-------------------|------------------|---------------------------|
| 3              | 3              | 7              | 4,766     | 2,234             | 4,990756          | 49               | 22,71476                  |
| 6              | 7              | 8              | 6,89      | 1,11              | 1,2321            | 64               | 47,4721                   |
| 4              | 5              | 3              | 4,856     | -1,856            | 3,444736          | 9                | 23,58074                  |
| 6              | 5              | 8              | 8,744     | -0,744            | 0,553536          | 64               | 76,45754                  |
| 2              | 2              | 3              | 3,749     | -0,749            | 0,561001          | 9                | 14,055                    |
| $\Sigma x1=21$ | $\Sigma x2=22$ | $\Sigma y =29$ |           |                   | $\Sigma e^2=10,7$ | $\Sigma y^2=195$ | $\Sigma \hat{y}^2=184,28$ |

# Παράδειγμα 10.5

- $$R^2 = \frac{\sum_1^n (\hat{y}_i - \bar{y})^2}{\sum_1^n (y_i - \bar{y})^2} = \frac{\sum \hat{y}_i^2 - \frac{(\sum \hat{y}_i)^2}{n}}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}} = \frac{184,28 - \frac{29^2}{5}}{195 - \frac{29^2}{5}} = 0,6 = 60\%$$

- Έχουμε τις ποσότητες που μας ενδιαφέρουν

- $$SS_R = \sum \hat{y}_i^2 - \frac{(\sum \hat{y}_i)^2}{n} = 184,28 - \frac{29^2}{5} = 16,08$$

- $$SS_T = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 195 - \frac{29^2}{5} = 26,8$$

$$SS_e = SS_T - SS_R = 26,8 - 16,08 = 10,7$$

# Παράδειγμα 10.5

| Αθροίσματα τετραγώνων | Βαθμοί ελευθερίας |
|-----------------------|-------------------|
| $SS_T = 26,8$         | $n-1=5-1=4$       |
| $SS_R = 16,08$        | $k=2$             |
| $SS_e = 10,7$         | $n-k-1=5-2-1=2$   |

# Παράδειγμα 10.5

- Υπολογίζουμε τρεις λόγους:
- $\frac{SS_R}{k} = \frac{16,08}{2} = 8,04$
- που είναι η διασπορά της παλινδρόμησης (των εκτιμήσεων των  $\gamma$ ),
- $\frac{SS_e}{n-k-1} = \frac{10,7}{5-2-1} = \frac{10,7}{2} = 5,35$
- που είναι η διασπορά των σφαλμάτων της παλινδρόμησης, και τέλος τον λόγο
- $F = \frac{\frac{SS_R}{k}}{\frac{SS_e}{n-k-1}} = \frac{8,04}{5,35} = 1,5$
- Ο λόγος αυτός των δύο διασπορών κατανέμεται σύμφωνα με την F κατανομή με  $k=2$  και  $n-k-1=2$  βαθμούς ελευθερίας.

# Παράδειγμα 10.5

| Αθροίσματα τετραγώνων                             | Βαθμοί ελευθερίας | Μέσα τετράγωνα (διασπορές)  | F                                                     |
|---------------------------------------------------|-------------------|-----------------------------|-------------------------------------------------------|
| Εξαιτίας της παλινδρόμησης $SS_R = 16,08$         | $k=2$             | $\frac{SS_R}{k} = 8,04$     | $F = \frac{\frac{SS_R}{k}}{\frac{SS_e}{n-k-1}} = 1,5$ |
| Εξαιτίας των σφαλμάτων (καταλοίπων) $SS_e = 10,7$ | $n-k-1=2$         | $\frac{SS_e}{n-k-1} = 5,35$ |                                                       |
| Συνολικά $SS_T = 26,8$                            | $n-1=4$           |                             |                                                       |

# Παράδειγμα 10.5

- Έχουμε λοιπόν ότι  $F(2,2)=1,5$ . Υπολογίζουμε τώρα την κρίσιμη τιμή  $F_{k,n-k-1;\alpha} = F_{2,2;0,05}$  αν υποθέσουμε ότι  $\alpha=5\%=0,05$ .
- $F_{2,2;0,05} = 19$
- Επειδή το  $F$  που υπολογίσαμε είναι 1,5 και δεν είναι μεγαλύτερο από το  $F_{2,2;0,05} = 19$  η παλινδρόμηση ΔΕΝ είναι στατιστικά σημαντική. Αυτό βέβαια μπορεί να αποδοθεί σε μεγάλο βαθμό στο μικρό μέγεθος του δείγματος που χρησιμοποιήσαμε. Με ένα μεγαλύτερο δείγμα (όπως είναι και το σωστό να έχουμε) μπορεί η παλινδρόμηση να ήταν στατιστικά σημαντική. Εδώ χρησιμοποιούμε μικρά δείγματα για να έχουμε απλά και εύληπτα παραδείγματα.

# Χρήση βοηθητικών - δίτιμων μεταβλητών

- Βοηθητικές μεταβλητές (Dummy variables) ονομάζουμε τις διχοτομικές - δίτιμες μεταβλητές με τιμές δηλαδή 0 ή 1 που χρησιμοποιούνται συχνά σε διάφορες στατιστικές επεξεργασίες και στην γραμμική παλινδρόμηση. Ας δούμε μερικά απλά παραδείγματα που θα μας βοηθήσουν να καταλάβουμε πώς εργαζόμαστε με βοηθητικές μεταβλητές. Συχνά οι δίτιμες μεταβλητές εκφράζουν κάποιο ατομικό χαρακτηριστικό με δύο τιμές, ή μια ιδιότητα που κάθε παρατηρούμενο υποκείμενο μπορεί να την έχει ή να μην την έχει. Αν αναφερόμαστε σε χώρες για παράδειγμα θα μπορούσε μια δίτιμη βοηθητική μεταβλητή να εκφράζει αν η χώρα ανήκει ή όχι στην Ευρωπαϊκή Ένωση, στον ΟΟΣΑ, στο ΝΑΤΟ, σε ατομικό επίπεδο μπορεί μια δίτιμη μεταβλητή να είναι το φύλο με τιμές άντρας, γυναίκα, κλπ.

# Παράδειγμα 10.6

- Το παράδειγμα χρησιμοποιεί τα δεδομένα βαθμών φοιτητών και του φύλου τους (Α/Γ).

| x1 | x2 | φύλο | Υ |
|----|----|------|---|
| 3  | 3  | Α    | 7 |
| 6  | 7  | Γ    | 8 |
| 4  | 5  | Γ    | 3 |
| 6  | 5  | Α    | 8 |
| 2  | 2  | Α    | 3 |
| 4  | 4  | Α    | 4 |
| 5  | 4  | Α    | 4 |
| 6  | 4  | Α    | 4 |
| 6  | 7  | Γ    | 6 |
| 7  | 7  | Γ    | 6 |
| 6  | 6  | Γ    | 7 |
| 7  | 6  | Α    | 6 |
| 6  | 5  | Α    | 7 |

# Παράδειγμα 10.6

| Χ1 | Χ2 | Χ3 (φύλο) | Υ |
|----|----|-----------|---|
| 3  | 3  | 0         | 7 |
| 6  | 7  | 1         | 8 |
| 4  | 5  | 1         | 3 |
| 6  | 5  | 0         | 8 |
| 2  | 2  | 0         | 3 |
| 4  | 4  | 0         | 4 |
| 5  | 4  | 0         | 4 |
| 6  | 4  | 0         | 4 |
| 6  | 7  | 1         | 6 |
| 7  | 7  | 1         | 6 |
| 6  | 6  | 1         | 7 |
| 7  | 6  | 0         | 6 |
| 6  | 5  | 0         | 7 |

# Παράδειγμα 10.6

- Το  $R^2$  του μοντέλου είναι 0,385 και το μοντέλο είναι
- $\hat{Y} = 1,274 - 0,412 \cdot X1 + 1,482 \cdot X2 - 2,364 \cdot X3$
- Θυμίζουμε ότι οι συντελεστές του μοντέλου εκφράζουν πόσο μεταβάλλεται η  $Y$  όταν η  $X$  αυξηθεί κατά μία μονάδα (όταν οι άλλες ανεξάρτητες παραμένουν αμετάβλητες). Οπότε, επειδή ο συντελεστής της  $X3$  (του φύλου) είναι -2,364 αυτό σημαίνει ότι όταν αυξήσουμε την  $X3$  κατά μία μονάδα, η  $Y$  μειώνεται περισσότερο από 2. Αλλά τότε αυξάνεται κατά μία μονάδα η  $X3$ ; Όταν αντί για Άντρα με τιμή 0, αυξάνουμε την τιμή κατά 1, δηλαδή στο 1 που δηλώνει Γυναίκα. Άρα με αμετάβλητους τους βαθμούς στα δύο μαθήματα, που παριστάνονται με τις μεταβλητές  $X1$  και  $X2$ , οι γυναίκες επιτυγχάνουν χειρότερη επίδοση από τους άντρες. Πάντα το συμπέρασμα βγαίνει σε σχέση με την κατηγορία που έχει την τιμή μηδέν, εδώ τους άντρες. Η κατηγορία αυτή ονομάζεται κατηγορία αναφοράς.

# Παράδειγμα 10.7

- Στο παράδειγμα αυτό αντί για το φύλο χρησιμοποιούμε μια ποιοτική μεταβλητή με τρεις τώρα κατηγορίες. Αυτή η μεταβλητή μπορεί να είναι για παράδειγμα το αν ένας φοιτητής είναι παλιών εξαμήνων, ή νέος φοιτητής ή φοιτητής στο πρόγραμμα ανταλλαγών Erasmus. Πώς θα συμπεριλάβουμε αυτήν τη μεταβλητή στο μοντέλο πολλαπλής παλινδρόμησης με βοηθητικές μεταβλητές;

# Παράδειγμα 10.7

| Χ1 | Χ2 | Παλαιότητα φοιτητή/φοιτήτριας | Υ |
|----|----|-------------------------------|---|
| 3  | 3  | Παλιός/ά                      | 7 |
| 6  | 7  | Νέος/α                        | 8 |
| 4  | 5  | Νέος/α                        | 3 |
| 6  | 5  | Παλιός/ά                      | 8 |
| 2  | 2  | Παλιός/ά                      | 3 |
| 4  | 4  | Παλιός/ά                      | 4 |
| 5  | 4  | Νέος/α                        | 4 |
| 6  | 4  | Erasmus                       | 4 |
| 6  | 7  | Παλιός/ά                      | 6 |
| 7  | 7  | Erasmus                       | 6 |
| 6  | 6  | Erasmus                       | 7 |
| 7  | 6  | Erasmus                       | 6 |
| 6  | 5  | Erasmus                       | 7 |

# Παράδειγμα 10.7

- Εδώ έχουμε μια τρίτιμη μεταβλητή. Υπάρχουν διάφοροι τρόποι να την εισάγουμε σε ένα μοντέλο με χρήση βοηθητικών μεταβλητών (dummy variables). Αν  $k$  είναι το πλήθος των κατηγοριών της μεταβλητής, ένας απλός τρόπος που χρησιμοποιείται συχνά είναι να χρησιμοποιήσουμε  $k-1$  δίτιμες βοηθητικές μεταβλητές. Πώς γίνεται αυτό; Ας το δούμε στην περίπτωση μας που έχουμε  $k=3$ . Θα χρησιμοποιήσουμε δηλαδή δύο δίτιμες βοηθητικές μεταβλητές.
- Πρώτα ορίζουμε ότι η μία από τις κατηγορίες θα είναι η *κατηγορία αναφοράς*. Στις ποιοτικές μεταβλητές οι οποίες με χρήση βοηθητικών (dummy) μεταβλητών εισάγονται σε μοντέλα, δεν μπορούμε να βγάζουμε συνολικά συμπεράσματα, αλλά το συχνότερο είναι να συγκρίνονται οι κατηγορίες της μεταβλητής με μία συγκεκριμένη κατηγορία της, την οποία και ονομάζουμε κατηγορία αναφοράς. Εμείς αποφασίζουμε ποια θα είναι η κατηγορία αναφοράς. Μπορεί να είναι μια πολύ κοινή κατηγορία (με τη μεγαλύτερη συχνότητα), ή μια κατηγορία για την οποία έχουμε προηγούμενη γνώση ή αυτή που τελικά θέλουμε να εκφράζει τη βάση των συγκρίσεών μας.

# Παράδειγμα 10.7

- Στο παράδειγμά μας επιλέγουμε ως κατηγορία αναφοράς την κατηγορία «Φοιτητής/φοιτήτρια Erasmus». Στη συνέχεια δημιουργούμε τις δύο  $(k-1)$  βοηθητικές μεταβλητές ως εξής:
- Χπαλιός: 0 αν δεν είναι παλιός φοιτητής, 1 αν είναι παλιός φοιτητής (φοιτητές κανονικής φοίτησης).
- Χνέος: 0 αν δεν είναι νέος φοιτητής, 1 αν είναι νέος φοιτητής (φοιτητές κανονικής φοίτησης).

# Παράδειγμα 10.7

|                                | Χπαλιός | Χνέος |
|--------------------------------|---------|-------|
| Παλιός/ά<br>φοιτητής/φοιτήτρια | 1       | 0     |
| Νέος/α<br>φοιτητής/φοιτήτρια   | 0       | 1     |
| Φοιτητής/φοιτήτρια<br>Erasmus  | 0       | 0     |

# τα δεδομένα μας

| Χ1 | Χ2 | Χπαλιός | Χνέος | Υ |
|----|----|---------|-------|---|
| 3  | 3  | 1       | 0     | 7 |
| 6  | 7  | 0       | 1     | 8 |
| 4  | 5  | 0       | 1     | 3 |
| 6  | 5  | 1       | 0     | 8 |
| 2  | 2  | 1       | 0     | 3 |
| 4  | 4  | 1       | 0     | 4 |
| 5  | 4  | 0       | 1     | 4 |
| 6  | 4  | 0       | 0     | 4 |
| 6  | 7  | 1       | 0     | 6 |
| 7  | 7  | 0       | 0     | 6 |
| 6  | 6  | 0       | 0     | 7 |
| 7  | 6  | 0       | 0     | 6 |
| 6  | 5  | 0       | 0     | 7 |

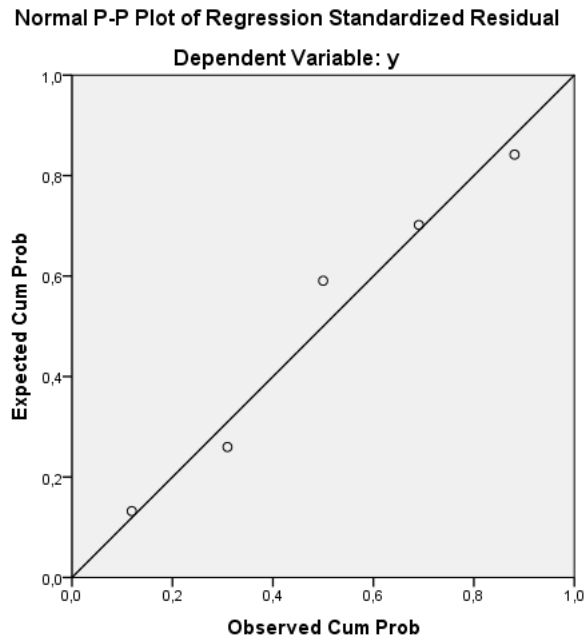
# Παράδειγμα 10.7

- Κάνουμε τώρα την πολλαπλή παλινδρόμηση με τέσσερις ανεξάρτητες μεταβλητές, οι δύο από αυτές είναι βοηθητικές.
- Το  $R^2$  του μοντέλου είναι 0,4.
- Και το μοντέλο είναι
- $\hat{Y} = 0,584 + 0,557 \cdot X1 + 0,330 \cdot X2 + 1,288 \cdot X_{\text{παλιός}} - 0,132 \cdot X_{\text{νέος}}$
- Τι εκφράζει τώρα αυτό το μοντέλο; Μας λέει ότι με τις υπόλοιπες μεταβλητές να μη μεταβάλλονται, οι παλιοί φοιτητές επιτυγχάνουν λίγο καλύτερα από τους φοιτητές Erasmus (την κατηγορία αναφοράς) επειδή έχουν συντελεστή 1,288. Επίσης οι νέοι φοιτητές έχουν μικρότερη συγκριτικά επίδοση από τους φοιτητές Erasmus (την κατηγορία αναφοράς) επειδή έχουν συντελεστή - 0,132 με τις υπόλοιπες μεταβλητές να μη μεταβάλλονται. Η σύγκριση πάντα γίνεται με την κατηγορία αναφοράς.

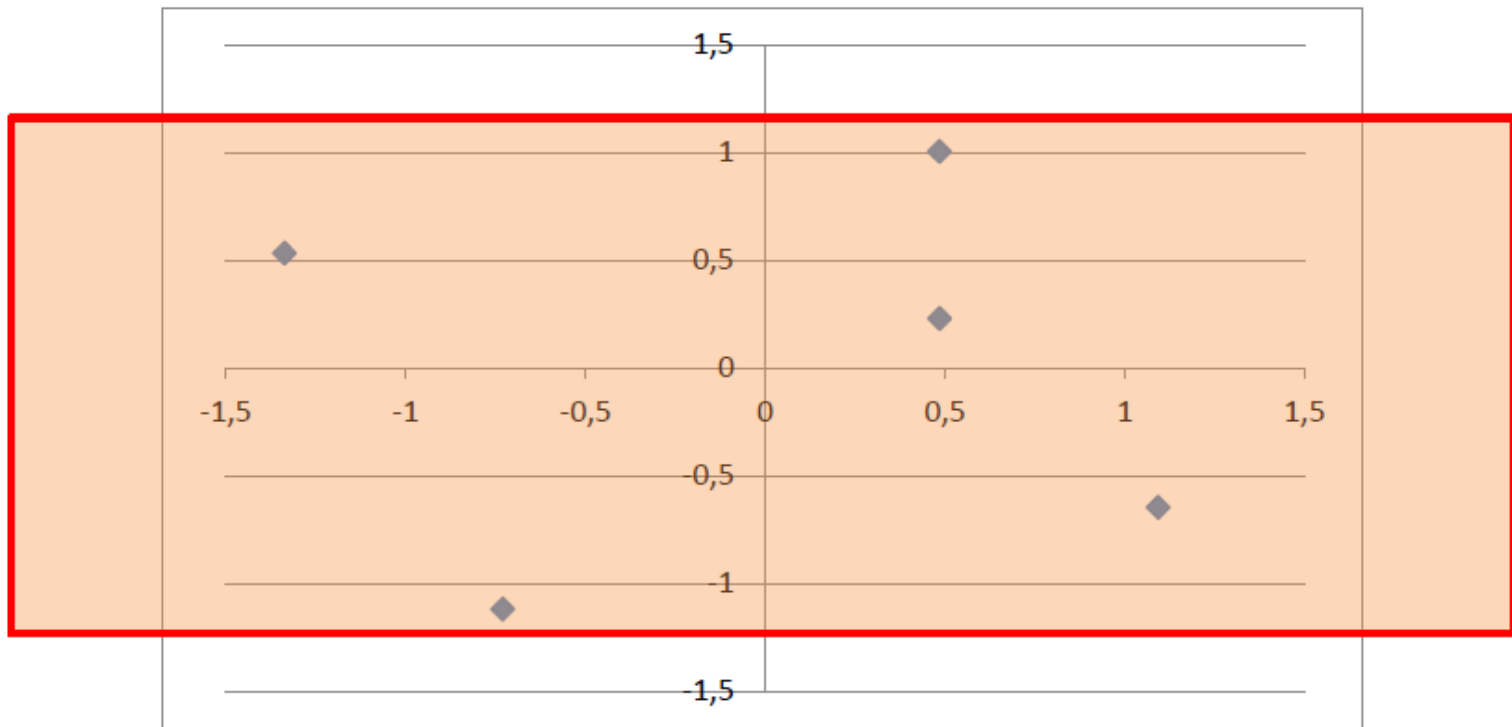
# Προϋποθέσεις της παλινδρόμησης

- Υπάρχει γραμμικότητα. Δηλαδή υποθέτουμε ότι οι μέσες τιμές της  $Y$  για τα διάφορα επίπεδα της  $X$  βρίσκονται πάνω σε ευθεία γραμμή.
- Οι τιμές της  $Y$  που αντιστοιχούν στα διάφορα επίπεδα της  $X$  είναι ανεξάρτητες μεταξύ τους.
- Η κατανομή της  $Y$  για κάθε επίπεδο της  $X$  είναι κανονική.
- Θα πρέπει να έχουμε *ομοσκεδαστικότητα* και όχι *ετεροσκεδαστικότητα*. Δηλαδή η διασπορά των καταλοίπων για κάθε τιμή της  $X$  θα πρέπει να είναι σταθερή.

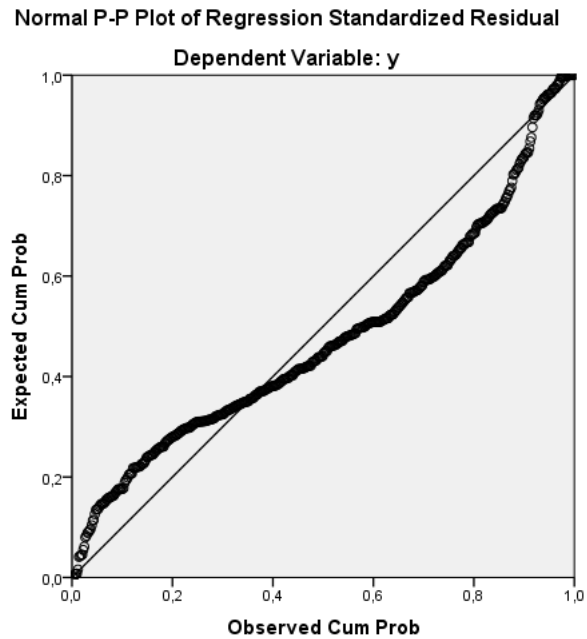
# Παραδείγματα (κανονικότητα)



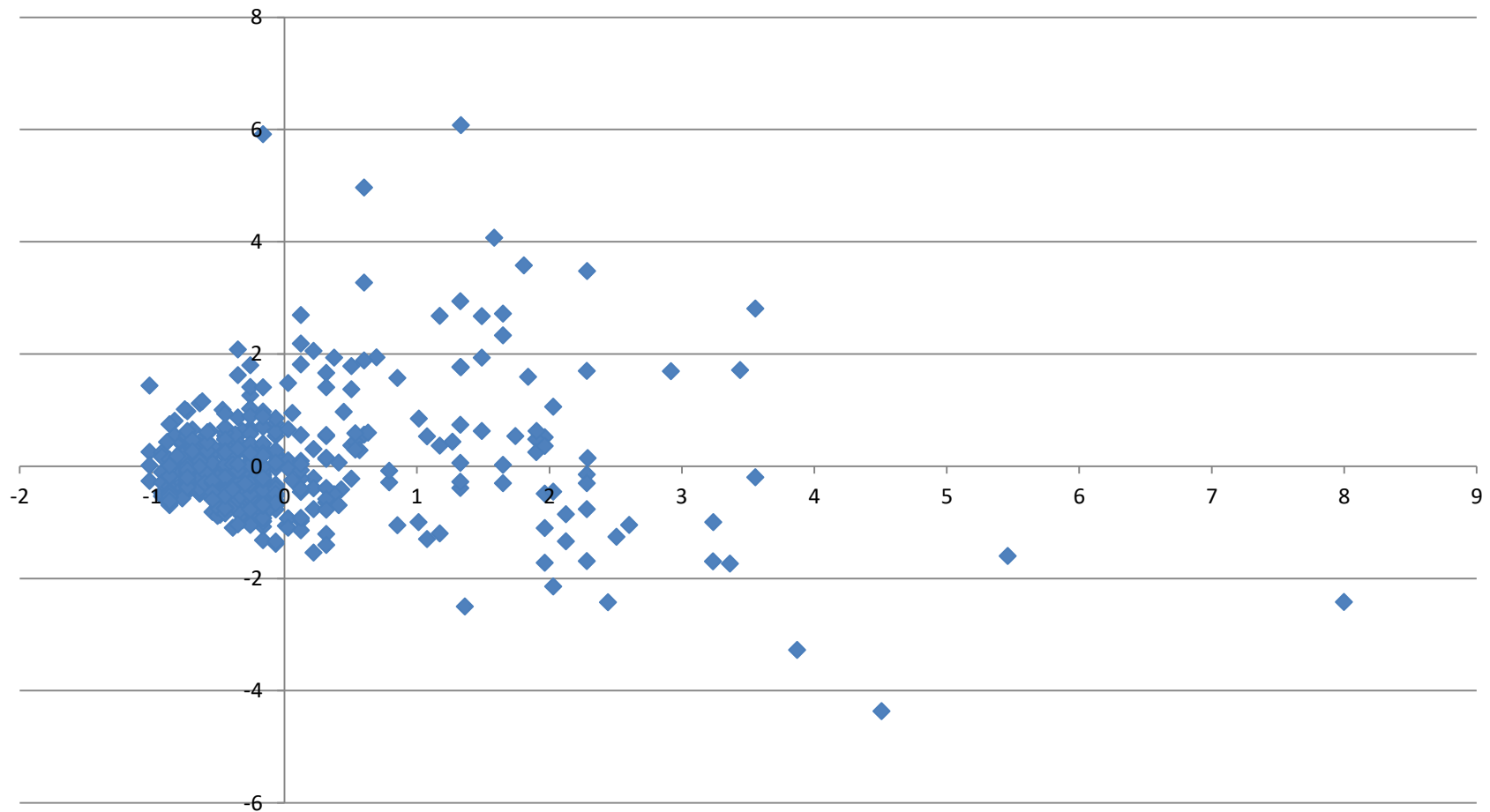
# Ομοσκεδαστικότητα



# Χωρίς κανονικότητα



# Ετεροσκεδαστικότητα



# Θεραπεία μη κανονικότητας και ετεροσκεδαστικότητας

- Συνήθως χρησιμοποιούμε νέες μεταβλητές που είναι οι αντίστροφες ποσότητες των αρχικών μεταβλητών (με την προϋπόθεση ότι οι αρχικές παρατηρήσεις είναι μη μηδενικές).
- $X^* = \frac{1}{X}$
- $Y^* = \frac{1}{Y}$
- άλλος συχνός μετασχηματισμός είναι ο λογαριθμικός, εφόσον οι παρατηρήσεις των  $X$  και  $Y$  είναι μεγαλύτερες από το μηδέν.
- $X^* = \ln(X)$
- $Y^* = \ln(Y)$

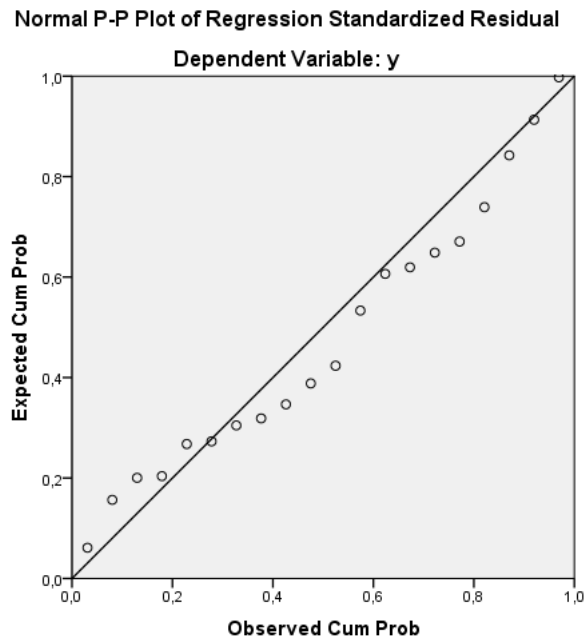
# Παράδειγμα 10.10

| Χ    | Υ     |
|------|-------|
| 2,7  | 5,7   |
| 1,88 | 4,02  |
| 1,2  | 2,15  |
| 1,32 | 2,19  |
| 2,1  | 4,5   |
| 1,35 | 3,21  |
| 1,88 | 3,6   |
| 0,98 | 2,19  |
| 1,28 | 2,79  |
| 1,35 | 2,4   |
| 1,65 | 3,03  |
| 1,2  | 2,84  |
| 1,43 | 2,78  |
| 1,68 | 3,51  |
| 1,35 | 2,73  |
| 1,5  | 4,08  |
| 1,43 | 4,6   |
| 2,75 | 10,38 |
| 1,43 | 4,23  |
| 1,16 | 2,63  |

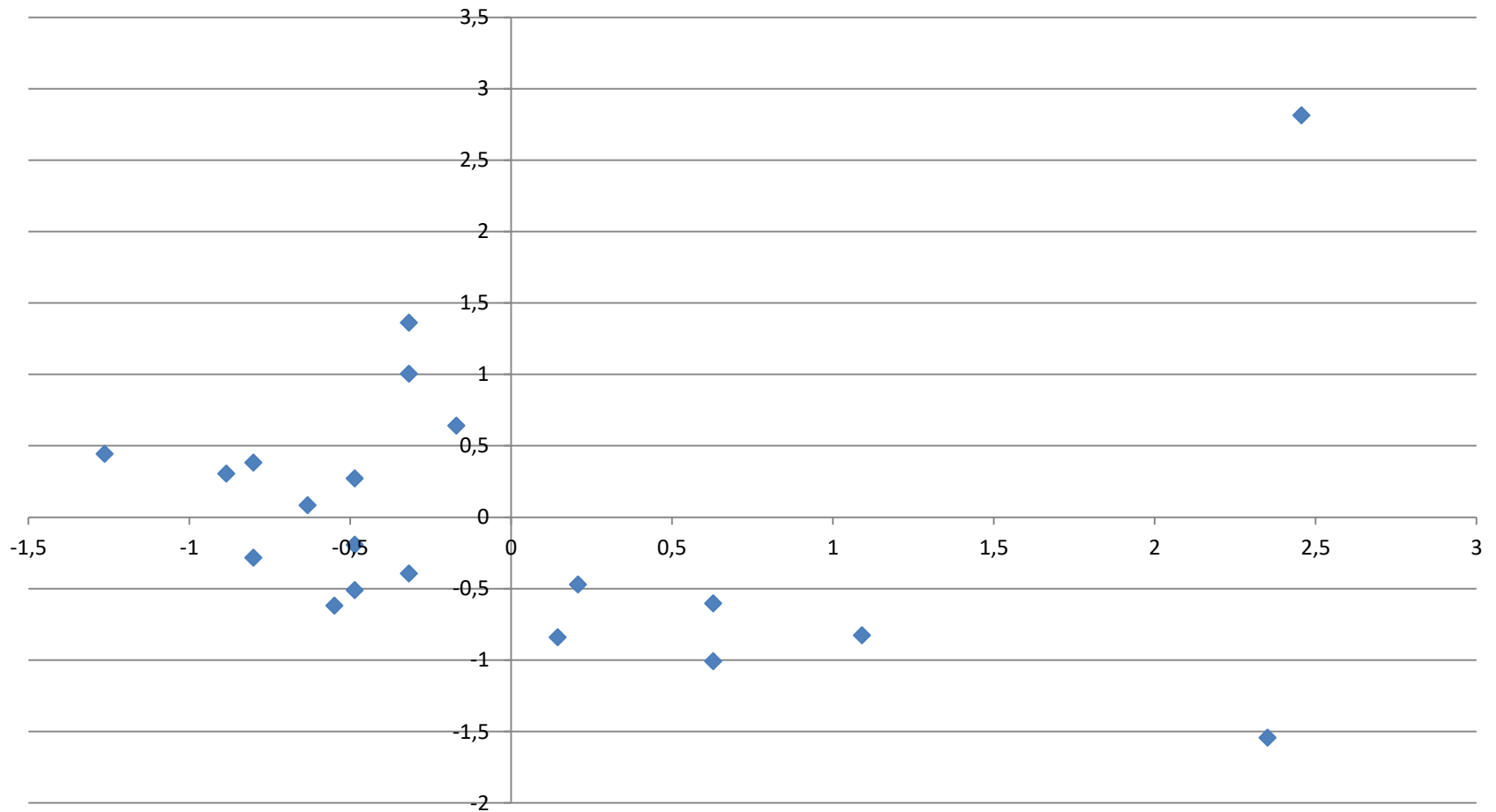
# Παράδειγμα 10.10

- Δίνουμε κατευθείαν το εκτιμημένο μοντέλο που είναι
- $\hat{Y} = -1,441 - 3,238 \cdot X$

# Παράδειγμα 10.10



# Παράδειγμα 10.10



# Παράδειγμα 10.10

| X    | Υ     | $X^* = \frac{1}{X}$ | $Y^* = \frac{1}{Y}$ |
|------|-------|---------------------|---------------------|
| 2,7  | 5,7   | 0,37                | 0,18                |
| 1,88 | 4,02  | 0,53                | 0,25                |
| 1,2  | 2,15  | 0,83                | 0,47                |
| 1,32 | 2,19  | 0,76                | 0,46                |
| 2,1  | 4,5   | 0,48                | 0,22                |
| 1,35 | 3,21  | 0,74                | 0,31                |
| 1,88 | 3,6   | 0,53                | 0,28                |
| 0,98 | 2,19  | 1,02                | 0,46                |
| 1,28 | 2,79  | 0,78                | 0,36                |
| 1,35 | 2,4   | 0,74                | 0,42                |
| 1,65 | 3,03  | 0,61                | 0,33                |
| 1,2  | 2,84  | 0,83                | 0,35                |
| 1,43 | 2,78  | 0,7                 | 0,36                |
| 1,68 | 3,51  | 0,6                 | 0,28                |
| 1,35 | 2,73  | 0,74                | 0,37                |
| 1,5  | 4,08  | 0,67                | 0,25                |
| 1,43 | 4,6   | 0,7                 | 0,22                |
| 2,75 | 10,38 | 0,36                | 0,1                 |
| 1,43 | 4,23  | 0,7                 | 0,24                |
| 1,16 | 2,63  | 0,86                | 0,38                |

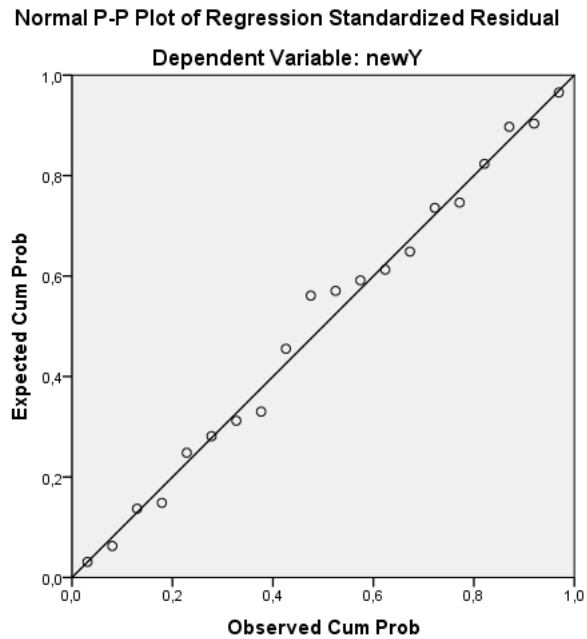
# Παράδειγμα 10.10

- Κάνουμε τώρα παλινδρόμηση με τις δύο νέες μεταβλητές και το μοντέλο είναι
- $\hat{Y}^* = -0,026 + 0,501 \cdot X^*$
- Υπολογίζουμε τις εκτιμήσεις  $\hat{Y}^*$  και τα κατάλοιπα του μοντέλου που μόλις γράψαμε και υπολογίζουμε και τις τυποποιημένες τους τιμές.

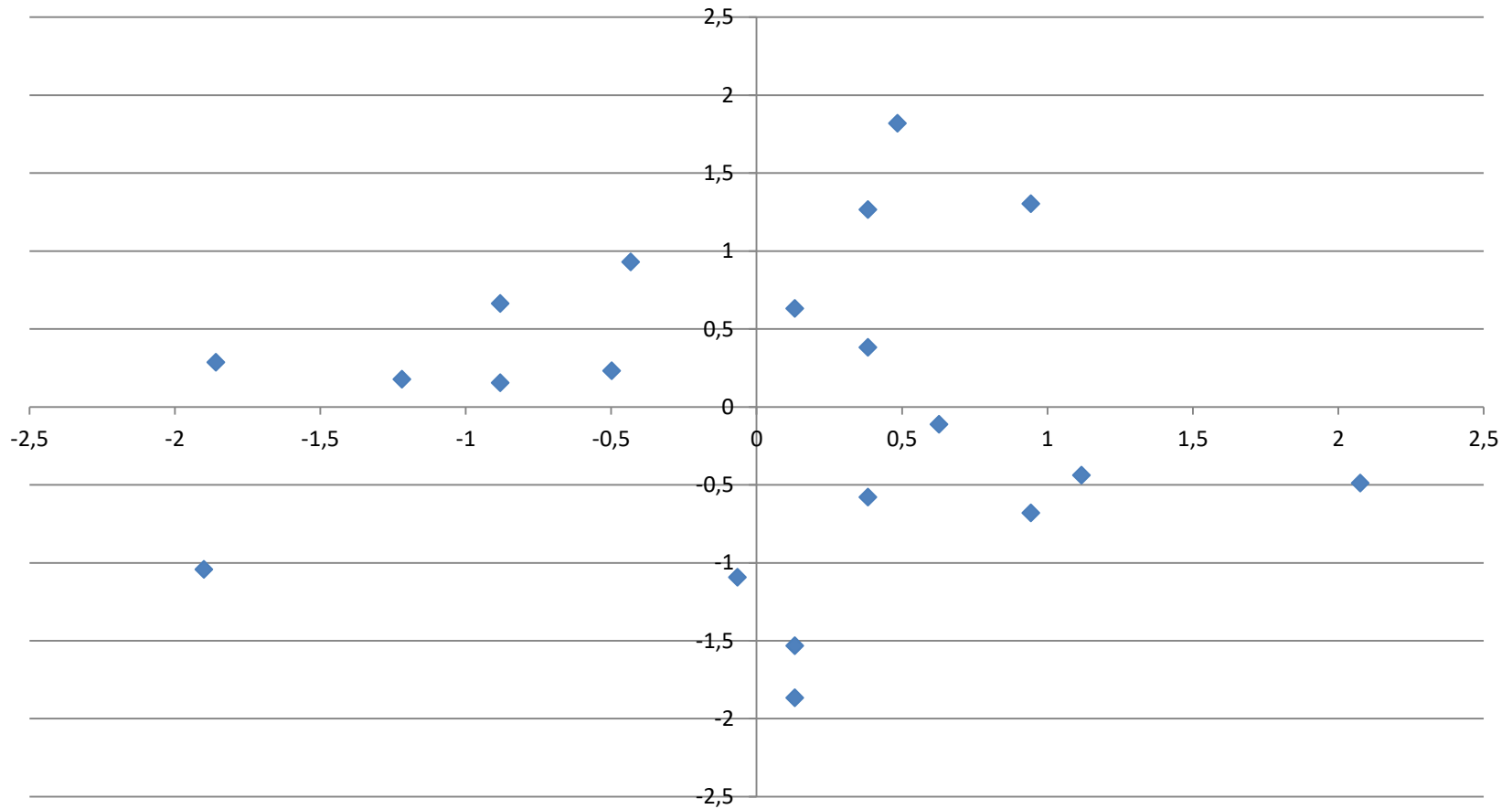
# Παράδειγμα 10.10

| $x$  | $y$   | $x^*$ | $y^*$ | $\widehat{y^*}$ | $e^*$    | Τυποποιημένες τιμές<br>$z_{\widehat{y^*}}$ | Τυποποιημένες τιμές<br>$z_{e^*}$ |
|------|-------|-------|-------|-----------------|----------|--------------------------------------------|----------------------------------|
| 2,7  | 5,7   | 0,37  | 0,18  | 0,15913         | 0,01631  | -1,85911                                   | 0,28618                          |
| 1,88 | 4,02  | 0,53  | 0,25  | 0,24            | 0,00876  | -0,88126                                   | 0,15363                          |
| 1,2  | 2,15  | 0,83  | 0,47  | 0,3909          | 0,07422  | 0,94325                                    | 1,30207                          |
| 1,32 | 2,19  | 0,76  | 0,46  | 0,35297         | 0,10365  | 0,48468                                    | 1,81839                          |
| 2,1  | 4,5   | 0,48  | 0,22  | 0,2121          | 0,01012  | -1,21857                                   | 0,17754                          |
| 1,35 | 3,21  | 0,74  | 0,31  | 0,34454         | -0,03302 | 0,38278                                    | -0,57921                         |
| 1,88 | 3,6   | 0,53  | 0,28  | 0,24            | 0,03778  | -0,88126                                   | 0,66277                          |
| 0,98 | 2,19  | 1,02  | 0,46  | 0,48455         | -0,02793 | 2,07562                                    | -0,48997                         |
| 1,28 | 2,79  | 0,78  | 0,36  | 0,36482         | -0,0064  | 0,62798                                    | -0,11226                         |
| 1,35 | 2,4   | 0,74  | 0,42  | 0,34454         | 0,07212  | 0,38278                                    | 1,26531                          |
| 1,65 | 3,03  | 0,61  | 0,33  | 0,27712         | 0,05291  | -0,43245                                   | 0,92831                          |
| 1,2  | 2,84  | 0,83  | 0,35  | 0,3909          | -0,03878 | 0,94325                                    | -0,68039                         |
| 1,43 | 2,78  | 0,7   | 0,36  | 0,3238          | 0,03592  | 0,13194                                    | 0,63009                          |
| 1,68 | 3,51  | 0,6   | 0,28  | 0,2717          | 0,0132   | -0,49796                                   | 0,23158                          |
| 1,35 | 2,73  | 0,74  | 0,37  | 0,34454         | 0,02176  | 0,38278                                    | 0,38171                          |
| 1,5  | 4,08  | 0,67  | 0,25  | 0,30746         | -0,06236 | -0,0656                                    | -1,09402                         |
| 1,43 | 4,6   | 0,7   | 0,22  | 0,3238          | -0,10641 | 0,13194                                    | -1,86671                         |
| 2,75 | 10,38 | 0,36  | 0,1   | 0,15575         | -0,05942 | -1,89987                                   | -1,04235                         |
| 1,43 | 4,23  | 0,7   | 0,24  | 0,3238          | -0,08739 | 0,13194                                    | -1,53311                         |
| 1,16 | 2,63  | 0,86  | 0,38  | 0,40528         | -0,02505 | 1,11718                                    | -0,43953                         |

# Παράδειγμα 10.10



# Παράδειγμα 10.10



# Παράδειγμα 10.10

- Όπως είπαμε το μοντέλο είναι το παρακάτω
- $Y^* = -0,026 + 0,501 \cdot X^*$
- και είναι αποδεκτό. Άρα, αν στο τέλος θέλουμε να πάρουμε ένα μοντέλο που να εκτιμά την  $Y$  από την  $X$ , τότε έχουμε
- $\frac{1}{Y} = -0,026 + 0,501 \cdot \frac{1}{X}$
- Λύνοντας ως προς  $Y$  έχουμε
- $Y = \frac{1}{-0,026 + 0,501 \cdot \frac{1}{X}}$