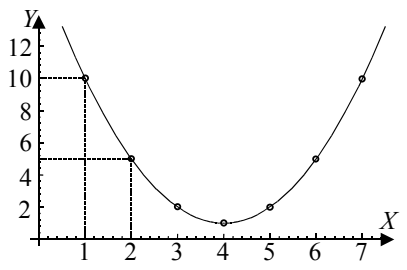
Ανάλυση Παλινδρόμησης

Με την *ανάλυση παλινδρόμησης (regression analysis)* εξετάζουμε τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών με σκοπό την πρόβλεψη των τιμών της μιας, μέσω των τιμών της άλλης (ή των άλλων). Σε κάθε πρόβλημα παλινδρόμησης διακρίνουμε δύο είδη μεταβλητών: τις ***ανεξάρτητες ή ελεγχόμενες ή επεξηγηματικές (independent, predictor, casual, input, explanatory variables)*** και τις ***εξαρτημένες ή απόκρισης (dependent, response variables)***. Σε **πειραματικές έρευνες**, *ανεξάρτητη μεταβλητή* X είναι εκείνη την οποία μπορούμε να ελέγξουμε, δηλαδή, να καθορίσουμε τις τιμές της *(π.χ. το ύψος της διαφημιστικής δαπάνης ενός προϊόντος, ο αριθμός των λειτουργούντων ταμείων σε ένα υποκατάστημα τραπέζης, η ποσότητα λιπάσματος, η θερμοκρασία επεξεργασίας ενός προϊόντος). Εξαρτημένη μεταβλητή* Y είναι εκείνη στην οποία αντανακλάται το αποτέλεσμα των μεταβολών στις *ανεξάρτητες μεταβλητές (π.χ. η ζήτηση ενός προϊόντος, ο χρόνος αναμονής των πελατών ενός υποκαταστήματος τραπέζης, η απόδοση μιας καλλιέργειας, η αντοχή ενός υλικού).* Σε **μη πειραματικές έρευνες (δειγματοληψίες)** η διάκριση μεταξύ *ανεξάρτητων* και *εξαρτημένων μεταβλητών* δεν είναι πάντοτε σαφής γιατί καμία μεταβλητή δεν είναι ελεγχόμενη αλλά όλες είναι τυχαίες (π.χ. *το ύψος και το βάρος των φοιτητών*, *οι ώρες μελέτης των φοιτητών ενός πανεπιστημιακού τμήματος και η απόδοση τους σε ένα τεστ, οι εβδομάδες εμπειρίας ενός εργάτη σε μια επιχείρηση* και *ο αριθμός των ελαττωματικών προϊόντων που παράγει, η κατάταξη δέκα προϊόντων από έναν κριτή και η κατάταξη των ιδίων προϊόντων από έναν άλλο κριτή, ο αριθμός των πωλήσεων μουσικών CD σε μια περιοχή και ο αριθμός των νέων στην ίδια περιοχή*).

Ας θεωρήσουμε δύο μεταβλητές X, Y. Αν οι μεταβλητές αυτές συνδέονται με μια σχέση της μορφής $Y = f(X)$ μέσω της οποίας για κάθε τιμή της X μπορούμε να προβλέψουμε ακριβώς την τιμή της Y, δηλαδή, αν οι τιμές της Y δεν υπόκεινται σε σφάλματα, τότε λέμε ότι οι δύο μεταβλητές συνδέονται με τη *συναρτησιακή-προσδιοριστική (deterministic) σχέση* $Y = f(X)$. Για παράδειγμα, το ρεύμα που καταναλώνει μια οικογένεια σε ένα δίμηνο και το ποσό που πληρώνει για την κατανάλωση αυτή συνδέονται με συναρτησιακή-προσδιοριστική σχέση[6]. Επίσης, το ποσό που καταθέτει κάποιος στο Ταμιευτήριο και ο τόκος που παίρνει για το ποσό αυτό, συνδέονται με συναρτησιακή-προσδιοριστική σχέση. Σε αυτές τις περιπτώσεις τα σημεία του διαγράμματος διασποράς βρίσκονται όλα πάνω στην καμπύλη που έχει εξίσωση $Y = f(X)$ και όσες φορές και αν επαναλάβουμε το πείραμα θέτοντας το X στο ίδιο επίπεδο $X = x_i$, θα παίρνουμε πάντα την ίδια τιμή για το Y. Για παράδειγμα, η εξίσωση $Y = (X-4)^2 + 1$ (που παριστάνει μια παραβολή) περιγράφει προσδιοριστικά τη σχέση μεταξύ των X και Y του παρακάτω πίνακα:

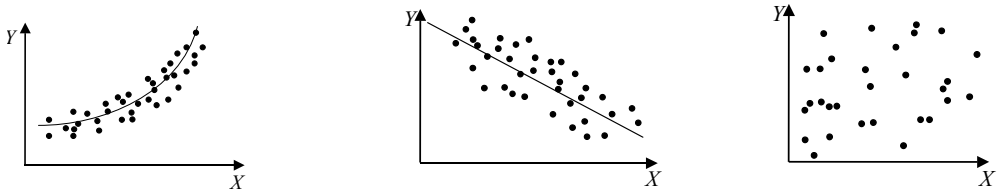
x_i	y_i
1	10
2	5
3	2
4	1
5	2
6	5
7	10



[6] *Για την ελληνική πραγματικότητα, αυτό το παράδειγμα προσδιοριστικής σχέσης, μάλλον είναι άστοχο.*

Οι μη προσδιοριστικές σχέσεις μεταξύ μεταβλητών ονομάζονται *στοχαστικές – στατιστικές (stochastic, probabilistic)* σχέσεις. Στην περίπτωση αυτή, αν επαναλάβουμε το πείραμα πολλές φορές θέτοντας το X στο ίδιο επίπεδο $X = x_i$ τότε στην τιμή x_i της X δεν αντιστοιχεί μια μόνο τιμή y_i της Y αλλά, γενικά, αντιστοιχεί ένα πλήθος διαφορετικών τιμών της Y. Για παράδειγμα, αν X είναι η τιμή ενός προϊόντος και Y είναι η ζήτησή του, η Y βρίσκεται σε στοχαστική σχέση-εξάρτηση από τη X, γιατί η ζήτηση ενός προϊόντος επηρεάζεται και από άλλους παράγοντες όπως είναι το ύψος του εισοδήματος των καταναλωτών, οι τιμές ομοειδών προϊόντων, οι καταναλωτικές συνήθειες, κ.ά.

Σε μια στοχαστική σχέση το διάγραμμα διασποράς είναι, γενικά, ένα *νέφος σημείων* το οποίο πολλές φορές καθορίζει μια ιδεατή γραμμή η οποία δίνει μια πρώτη εικόνα της σχέσης που συνδέει τις δύο μεταβλητές. Η σχέση μάλιστα μεταξύ των δύο μεταβλητών είναι τόσο περισσότερο ισχυρή όσο πιο κοντά στην ιδεατή γραμμή βρίσκονται τα σημεία του διαγράμματος διασποράς. Στο πρώτο από τα παρακάτω σχήματα έχουμε το διάγραμμα διασποράς μιας ισχυρής σχέσης στην οποία όταν αυξάνουν οι τιμές της X αυξάνουν γενικά και οι τιμές της Y, ενώ στο δεύτερο σχήμα έχουμε μια λιγότερο ισχυρή σχέση στην οποία όταν αυξάνουν οι τιμές της X ελαττώνονται γενικά και οι τιμές της Y. Τέλος, στην περίπτωση του τρίτου σχήματος δε φαίνεται να υπάρχει κάποια σχέση μεταξύ των X και Y.



Γενικά, δύο μεταβλητές που συνδέονται είτε με συναρτησιακή-προσδιοριστική σχέση είτε με στοχαστική σχέση λέγονται *«εξαρτημένες»*. Αν υπάρχει εξάρτηση μεταξύ δύο μεταβλητών, τότε μπορούμε τη μια από αυτές να τη χαρακτηρίσουμε ως «αιτία» και την άλλη ως «αποτέλεσμα». Αυτό όμως, μόνο στην περίπτωση που η εξάρτηση οφείλεται σε σχέση αιτιότητας των δύο μεταβλητών και όχι σε μια απλή συμμεταβολή η οποία μπορεί να οφείλεται σε εξάρτηση των δύο μεταβλητών από μια τρίτη μεταβλητή. Αν, για παράδειγμα, X είναι το ετήσιο εισόδημα μιας οικογένειας και Y, Z είναι τα ποσά που ξοδεύει η οικογένεια αυτή σε ένα έτος για κρέας και για αγορά λογοτεχνικών βιβλίων, τότε: αν διαπιστώσουμε σε ένα σύνολο οικογενειών σχέση μεταξύ των X και Y (ή μεταξύ των X και Z) δεχόμαστε ότι υπάρχει εξάρτηση μεταξύ των δύο μεταβλητών και τότε μπορούμε να χαρακτηρίσουμε τη X ως «αιτία» και την Y (ή τη Z) ως «αποτέλεσμα». Αν όμως διαπιστωθεί σχέση μεταξύ των Y και Z (που είναι πολύ πιθανό, αφού και οι δύο μεταβάλλονται με το ετήσιο εισόδημα X) ασφαλώς θα πρόκειται για «νόθα» εξάρτηση.

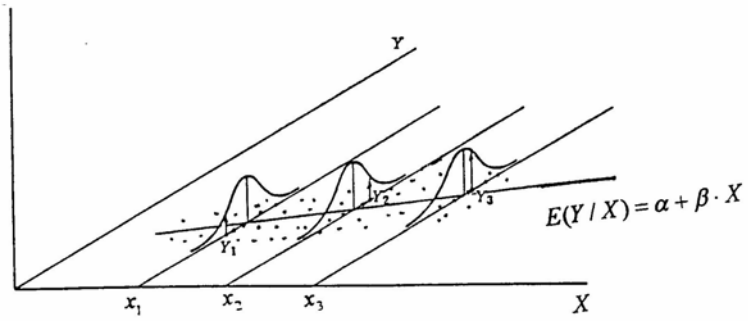
Για να περιγράψουμε τη *στοχαστική εξάρτηση* δύο μεταβλητών X και Y προσπαθούμε να βρούμε, όπως και στην *προσδιοριστική εξάρτηση*, μια σχέση μεταξύ των X και Y η οποία όμως τώρα δε θα δίνει ακριβή αλλά προσεγγιστική μόνο εικόνα της εξάρτησης των X και Y και τα σημεία του διαγράμματος διασποράς των X και Y δε θα βρίσκονται πάνω, αλλά, γύρω από μια καμπύλη. Μια μέθοδος που χρησιμοποιείται για την περιγραφή της στοχαστικής εξάρτησης δύο μεταβλητών είναι η **μέθοδος των**

ελαχίστων τετραγώνων και αυτή θα εφαρμόσουμε στη συνέχεια για να μελετήσουμε την πιο απλή μορφή στοχαστικής εξάρτησης, τη **γραμμική**.

Απλή Γραμμική Παλινδρόμηση

Αν το διάγραμμα διασποράς δύο μεταβλητών X και Y έχει μορφή *επιμήκους κεκλιμένης έλλειψης* ή *πλατυσμένου* ***J***, η σχέση των X και Y είναι κατά προσέγγιση γραμμική. Στην περίπτωση αυτή έχουμε την απλούστερη μορφή παλινδρόμησης, την **απλή γραμμική παλινδρόμηση** όπου υπάρχει μόνο μια ανεξάρτητη μεταβλητή X και η εξαρτημένη μεταβλητή Y μπορεί να προσεγγισθεί ικανοποιητικά από μια γραμμική συνάρτηση του X.

Η γραμμική σχέση $Y = \alpha + \beta \cdot X$ δε μπορεί, ασφαλώς, να περιγράψει τη γραμμική στοχαστική εξάρτηση των μεταβλητών X και Y αφού αν, για παράδειγμα, X είναι η τιμή ενός προϊόντος και Y είναι η ζήτηση του προϊόντος αυτού, και διατηρήσουμε τη X στο ίδιο επίπεδο $X = x_1$ τότε οι αντίστοιχες τιμές του Y θα είναι φυσικά διαφορετικές στις διάφορες επαναλήψεις. Επίσης, αν X είναι η ποσότητα λιπάσματος και Y είναι η απόδοση μιας καλλιέργειας, και διατηρήσουμε τη X στο ίδιο επίπεδο $X = x_1$ τότε οι αντίστοιχες τιμές του Y θα είναι φυσικά διαφορετικές στις διάφορες επαναλήψεις αφού παράγοντες όπως, η θερμοκρασία, οι βροχοπτώσεις, η ποιότητα του εδάφους, θα επηρεάζουν, επίσης, την παραγωγή. Επιπλέον, συμβαίνει να παρατηρούνται και σφάλματα μέτρησης των τιμών της Y (λόγω οργάνων ή ελλιπούς πληροφόρησης). Έτσι, για $X = x_1$ το αντίστοιχο Y είναι μια τυχαία μεταβλητή Y_1 που ακολουθεί κάποια κατανομή. Ομοίως, για $X = x_2$ θα έχουμε κάποια άλλη κατανομή Y_2 κ.ό.κ..



Επομένως, στην εξίσωση $Y = \alpha + \beta \cdot X$, πρέπει να προσθέσουμε έναν ακόμη όρο ε ο οποίος, για δεδομένη τιμή της X, να περιγράφει τη διαφορά της παρατηρούμενης από τη θεωρητική $(\alpha + \beta \cdot X)$ τιμή της Y. Δηλαδή, $\varepsilon = Y - (\alpha + \beta \cdot X)$. Προκύπτει, επομένως, το στοχαστικό μοντέλο

$$Y = \alpha + \beta \cdot X + \varepsilon .$$

Για λόγους απλούστευσης των υπολογισμών και εφικτότητας λύσης του προβλήματος, κάνουμε κάποιες υποθέσεις, όπως $E(\varepsilon) = 0$ και $E(Y/X) = \alpha + \beta \cdot X$. Δηλαδή, υποθέτουμε ότι τα σφάλματα έχουν μέση τιμή μηδέν και ότι για τις διάφορες τιμές της X, οι αντίστοιχες μέσες τιμές της Y βρίσκονται πάνω σε μια ευθεία (βλ. και *Παράρτημα Α). Η ευθεία αυτή* $(E(Y/X) = \alpha + \beta \cdot X)$, *ονομάζεται* ***πληθυσμιακή ευθεία παλινδρόμησης***.

Με τη μέθοδο των ελαχίστων τετραγώνων θα προσδιορίσουμε στη συνέχεια μια εκτίμηση $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ της ευθείας $E(Y/X) = \alpha + \beta \cdot X$ όπου $\hat{\alpha}$ και $\hat{\beta}$ εκτιμήτριες των α και β αντίστοιχα.

Η εκτίμηση $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ της *πληθυσμιακής ευθείας παλινδρόμησης* $E(Y/X) = \alpha + \beta \cdot X$, ονομάζεται ***ευθεία ελαχίστων τετραγώνων*** από τη μέθοδο υπολογισμού των παραμέτρων της.

<u>*Μέθοδος ελαχίστων τετραγώνων*</u>
Θεωρούμε ν ζεύγη παρατηρήσεων (x_i, y_i), $i = 1,2,3,...,\nu$. Αναζητούμε προσέγγιση της μορφής:

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

όπου τα ε_i παριστάνουν τις αποκλίσεις της πραγματικής τιμής y_i από την προσαρμοσμένη (θεωρητική) $\alpha + \beta \cdot x_i$. Δηλαδή, $\varepsilon_i = y_i - (\alpha + \beta \cdot x_i)$.

Είναι φανερό, ότι η εκλογή (εκτίμηση) των α και β θα πρέπει να γίνει έτσι ώστε να ελαχιστοποιηθούν οι ποσότητες ε_i. Για το σκοπό αυτό, θα αναζητήσουμε τις τιμές των α και β για τις οποίες ελαχιστοποιείται το άθροισμα των τετραγώνων των ε_i. Δηλαδή, η ποσότητα

$$\sum_{i=1}^{\nu} \varepsilon_i^2 = \sum_{i=1}^{\nu} (y_i - \alpha - \beta \cdot x_i)^2 \qquad (1)$$

(Η ελαχιστοποίηση του αθροίσματος $\sum \varepsilon_i$ δεν αποτελεί ασφαλές κριτήριο επιλογής διότι κάποια αρνητικά ε_i θα αναιρούν αντίστοιχες θετικές ποσότητες του αθροίσματος).

Παραγωγίζοντας την (1) ως προς α και β και εξισώνοντας με μηδέν παίρνουμε τις ακόλουθες δύο εξισώσεις που ονομάζονται **κανονικές εξισώσεις**:

$$\sum_{i=1}^{\nu} y_i = \nu \cdot \alpha + \beta \cdot \sum_{i=1}^{\nu} x_i$$

$$\sum_{i=1}^{\nu} x_i y_i = \alpha \cdot \sum_{i=1}^{\nu} x_i + \beta \cdot \sum_{i=1}^{\nu} x_i^2$$

Λύνοντας το σύστημα των *κανονικών εξισώσεων*, παίρνουμε:

$$\hat{\beta} = \frac{\nu \cdot \sum_{i=1}^{\nu} x_i y_i - \left(\sum_{i=1}^{\nu} x_i\right) \cdot \left(\sum_{i=1}^{\nu} y_i\right)}{\nu \cdot \sum_{i=1}^{\nu} x_i^2 - \left(\sum_{i=1}^{\nu} x_i\right)^2} = \frac{\sum_{i=1}^{\nu} x_i y_i - \nu \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{\nu} x_i^2 - \nu \cdot \bar{x}^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

ή

$$\boxed{\hat{\beta} = \frac{s_{xy}}{s_x^2} \ \ και \ \ \hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}}$$

Η **εκτίμηση ελαχίστων τετράγωνων** $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ της **ευθείας παλινδρόμησης** από το δείγμα των ν ζευγών παρατηρήσεων είναι, επομένως, η

$$\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X = \bar{y} - \hat{\beta} \cdot \bar{x} + \hat{\beta} \cdot X = \bar{y} + \hat{\beta} \cdot (X - \bar{x})$$

ή

$$\boxed{\hat{Y} = \bar{y} + \frac{s_{xy}}{s_x^2} \cdot (X - \bar{x})}$$

Προφανώς, η *ευθεία ελαχίστων τετραγώνων*, διέρχεται από το σημείο $(\bar{x}, \bar{y})$.

Επισημαίνουμε ότι πρέπει να γίνεται διάκριση μεταξύ της παρατηρούμενης τιμής του Y και της $\hat{Y}$ που εκτιμάμε. Η **παρατηρούμενη** τιμή y_i είναι η πραγματική τιμή της Y, ενώ η τιμή $\hat{y}_i$ της $\hat{Y}$, είναι εκτίμηση της μέσης τιμής $E(Y / X = x_i)$.

Πόσο «καλή» είναι η ευθεία ελαχίστων τετραγώνων $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ **ως εκτίμηση της ευθείας παλινδρόμησης** $E(Y / X) = \alpha + \beta \cdot X$ **;**

Από την προφανή σχέση $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$, μπορεί εύκολα να δειχθεί (αλγεβρικά) ότι

$$\sum_{i=1}^{\nu} (y_i - \bar{y})^2 = \sum_{i=1}^{\nu} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{\nu} (\hat{y}_i - \bar{y})^2 \quad (2)$$

Το άθροισμα

$$SSTO = \sum_{i=1}^{\nu} (y_i - \bar{y})^2$$

λέγεται **ολικό άθροισμα τετραγώνων (total sum of squares)** ή **ολική μεταβλητότητα (total variation)** των y_i και όπως φαίνεται από τη (2) αναλύεται σε δύο συνιστώσες: στο **άθροισμα τετραγώνων παλινδρόμησης (regression sum of squares)**

$$SSR = \sum_{i=1}^{\nu} (\hat{y}_i - \bar{y})^2$$

και στο **άθροισμα τετραγώνων των σφαλμάτων (error sum of squares)** ή **υπόλοιπο μεταβλητότητας (residual variation)**

$$SSE = \sum_{i=1}^{\nu} (y_i - \hat{y}_i)^2$$

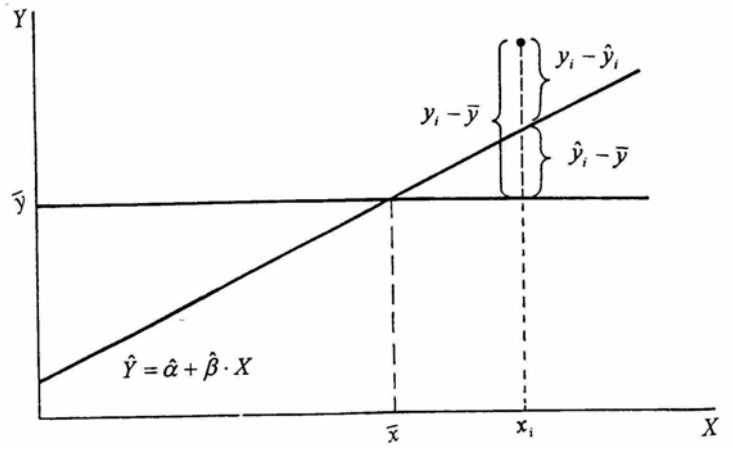
Δηλαδή,

$$SSTO = SSR + SSE$$

Το *SSTO* μετράει τη συνολική μεταβλητότητα των παρατηρήσεων y_i δηλαδή εκφράζει την αβεβαιότητα στην πρόβλεψη του Y όταν δε χρησιμοποιείται το X. Το *SSR* εκφράζει το μέρος της μεταβλητότητας που μπορεί να οφείλεται στο X και το *SSE=SSTO-SSR* εκφράζει την υπόλοιπη μεταβλητότητα που δεν εξηγείται από την παλινδρόμηση ενώ ο λόγος

$$r^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^{\nu}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{\nu}(y_i - \bar{y})^2} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum_{i=1}^{\nu}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{\nu}(y_i - \bar{y})^2}$$

εκφράζει το ποσοστό της συνολικής μεταβλητότητας των y_i που εξηγείται (απορροφάται) από την παλινδρόμηση. Το r^2 λέγεται ***συντελεστής προσδιορισμού (coefficient of determination)*** και παίρνει τιμές στο κλειστό διάστημα **[0, 1]**. Όταν όλα τα σημεία $M_1(x_1, y_1), M_2(x_2, y_2), ..., M_\nu(x_\nu, y_\nu)$ βρίσκονται πάνω στην *ευθεία ελαχίστων τετραγώνων* θα έχουμε $y_i = \hat{y}_i$ και άρα $SSE = \sum_{i=1}^{\nu}(y_i - \hat{y}_i)^2 = 0$ οπότε, $r^2 = 1$ ενώ όταν η κλίση της ευθείας ελαχίστων τετραγώνων είναι μηδέν δηλαδή $\hat{\beta} = 0$ θα είναι $r = 0$. Στις διάφορες πρακτικές εφαρμογές η τιμή του r^2 βρίσκεται μεταξύ 0 και 1 και όσο πλησιέστερα βρίσκεται προς το 1 τόσο καλύτερη είναι η ευθεία ελαχίστων τετραγώνων ως εκτίμηση της ευθείας παλινδρόμησης.



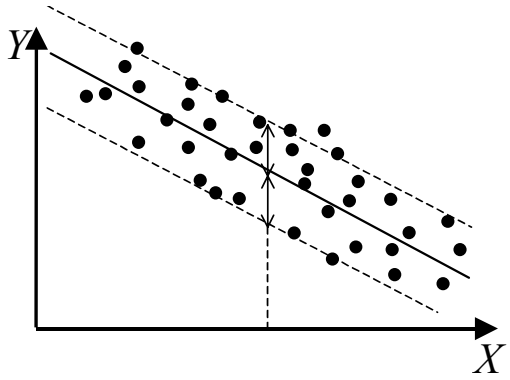
Η μέση απόκλιση μεταξύ της πραγματικής και της εκτιμούμενης τιμής της μεταβλητής ονομάζεται ***τυπικό σφάλμα της εκτίμησης (standard error of the estimate)***, συμβολίζεται με s και δίνεται από τον τύπο

$$s = \sqrt{\frac{1}{\nu - 2} \cdot \sum_{i=1}^{\nu}(y_i - \hat{y}_i)^2} = \sqrt{\frac{SSE}{\nu - 2}}$$

Εάν το *τυπικό σφάλμα* της εκτίμησης είναι μικρό τότε οι παρατηρούμενες και οι εκτιμούμενες τιμές δε διαφέρουν πολύ και η ευθεία παλινδρόμησης μας δίνει μια καλή περιγραφή της σχέσης μεταξύ των *X* και *Y*. Αν το *τυπικό σφάλμα* της εκτίμησης είναι μεγάλο τότε δε μπορούμε να ισχυρισθούμε ότι έχουμε μια καλή περιγραφή της σχέσης.

Είναι φανερό, ότι το *τυπικό σφάλμα* της εκτίμησης, είναι ένα μέτρο της *διασποράς* των (x_i, y_i) γύρω από την *ευθεία ελαχίστων τετραγώνων* $\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$ (το s^2 είναι μια εκτίμηση της διασποράς των σφαλμάτων). Έχει, επομένως, ιδιότητες ανάλογες με αυτές της *τυπικής απόκλισης*. Έτσι, αν φέρουμε δύο ευθείες παράλληλες προς την ευθεία ελαχίστων τετραγώνων και σε κατακόρυφες προς αυτήν αποστάσεις s, $2s$, $3s$ τότε, για μεγάλα ν (μεγαλύτερα του 30), μεταξύ των δύο αυτών ευθειών θα βρίσκεται

περίπου το 68%, το 95% και το 99,7% των σημείων του διαγράμματος διασποράς αντίστοιχα.



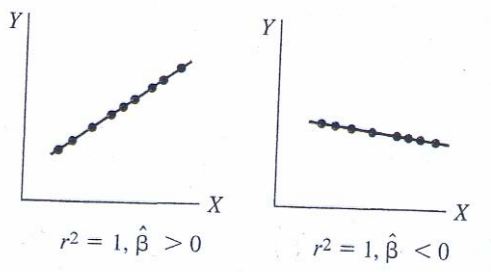
Σημείωση:
Στο σχήμα, οι παράλληλες έχουν σχεδιασθεί σε κατακόρυφη απόσταση από την ευθεία ελαχίστων τετραγώνων ίση με $2 \cdot s$.

Εύκολα μπορεί να αποδειχθεί ότι,

$$s = \sqrt{\frac{1}{\nu - 2} \cdot \sum_{i=1}^{\nu} (y_i - \hat{y}_i)^2} = \sqrt{\frac{\nu - 1}{\nu - 2} \cdot (s_y^2 - \hat{\beta}^2 \cdot s_x^2)} = \sqrt{\frac{\nu - 1}{\nu - 2} \cdot s_y^2 (1 - r^2)} .$$

Παρατηρήσεις για την ευθεία ελαχίστων τετραγώνων

1. Είναι φανερό ότι το $\hat{\beta}$ της *ευθείας ελαχίστων τετραγώνων* $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ εκφράζει την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής *Y* (σε μονάδες μέτρησης της *Y*) όταν η ανεξάρτητη μεταβλητή *X* αυξηθεί κατά μια μονάδα (μέτρησής της). Πράγματι αν $X = x_1$ έχουμε $\hat{y}_1 = \hat{\alpha} + \hat{\beta} \cdot x_1$ και αν $X = x_1 + 1$ έχουμε $\hat{y}_2 = \hat{\alpha} + \hat{\beta} \cdot (x_1 + 1) = \hat{\alpha} + \hat{\beta} \cdot x_1 + \hat{\beta} = \hat{y}_1 + \hat{\beta}$. Έτσι όταν το x_i αυξηθεί κατά μια μονάδα το $\hat{y}_i$ αυξάνεται κατά $\hat{\beta}$ μονάδες αν $\hat{\beta} > 0$ ή ελαττώνεται κατά $\hat{\beta}$ μονάδες αν $\hat{\beta} < 0$.
2. Το $\hat{\alpha}$ της ευθείας ελαχίστων τετραγώνων $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ εκφράζει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής *Y* όταν η ανεξάρτητη μεταβλητή *X* πάρει την τιμή 0.
3. Η ποσότητα $1 - r^2$ εκφράζει το ποσοστό της συνολικής μεταβλητότητας που οφείλεται στο τυχαίο σφάλμα.
4. Το r^2 **δεν μετρά** πόσο μεγάλη είναι η κλίση $\hat{\beta}$ της ευθείας παλινδρόμησης!



5. Όταν έχουμε πειραματικά δεδομένα όπου ο ερευνητής ελέγχει-καθορίζει τις τιμές της μιας μεταβλητής θεωρούμε τη μεταβλητή αυτή ανεξάρτητη (*X*) και την άλλη

εξαρτημένη (*Y*). Σε αυτή την περίπτωση εκτιμάμε την ευθεία παλινδρόμησης **της *Y* πάνω στη *X***, $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$. Όταν έχουμε μη πειραματικά δεδομένα όπου ο ερευνητής επιλέγει ένα τυχαίο δείγμα ατόμων και σε κάθε ένα από αυτά μετρά τις τιμές των μεταβλητών, τότε μπορούμε να θεωρήσουμε ως ανεξάρτητη μεταβλητή οποιαδήποτε από τις δύο και να μελετήσουμε είτε την παλινδρόμηση **της *Y* πάνω στη *X*** είτε την παλινδρόμηση **της *X* πάνω στη *Y***. Στην περίπτωση αυτή, και οι δύο μεταβλητές είναι τυχαίες, και ως μέτρο της γραμμικής συσχέτισης χρησιμοποιούμε το *συντελεστή γραμμικής συσχέτισης* $r = \frac{s_{xy}}{s_x \cdot s_y}$ και επειδή

$\hat{\beta} = \frac{s_{xy}}{s_x^2}$ θα είναι, $r = \hat{\beta} \cdot \frac{s_x}{s_y}$ (I)

Έτσι, αν το *r* πλησιάζει το 1 τότε τα σημεία του διαγράμματος διασποράς τείνουν να βρίσκονται σε μια ευθεία με συντελεστή διεύθυνσης $\hat{\beta} > 0$ ενώ, αν το r πλησιάζει το -1 τότε τα σημεία του διαγράμματος διασποράς τείνουν να βρίσκονται σε μια ευθεία με συντελεστή διεύθυνσης $\hat{\beta} < 0$. Αν $r \approx 0$ τότε $\hat{\beta} \approx 0$ και δεν υπάρχει γραμμική σχέση των μεταβλητών. Ο συντελεστής γραμμικής συσχέτισης έχει επομένως το ίδιο πρόσημο με το $\hat{\beta}$.

Αν $\hat{X} = \hat{\gamma} + \hat{\delta} \cdot Y$ είναι η εκτίμηση ελαχίστων τετραγώνων της ευθείας παλινδρόμησης **της *X* πάνω στην *Y*** θα ισχύει: $\hat{\delta} = \frac{s_{xy}}{s_y^2}$ και $\hat{\gamma} = \bar{x} - \hat{\delta} \cdot \bar{y}$.

Συνεπώς, $r = \hat{\delta} \cdot \frac{s_y}{s_x}$ (II). Από τις (I) και (II) προκύπτει, επίσης, ότι $r^2 = \hat{\beta} \cdot \hat{\delta}$.

6. Οι προβλέψεις που μπορούμε να κάνουμε για την εξαρτημένη μεταβλητή *Y* από τις τιμές της ανεξάρτητης μεταβλητής *X* μέσω της *ευθείας ελαχίστων τετραγώνων* $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$ πρέπει να γίνονται μόνο για τις τιμές της ανεξάρτητης μεταβλητής, οι οποίες βρίσκονται στο διάστημα που έχει γίνει η μελέτη ή πολύ κοντά στα άκρα του διαστήματος αυτού.
7. Η εξίσωση της *ευθείας ελαχίστων τετραγώνων* $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$, **δε μας επιτρέπει** να κάνουμε προβλέψεις για τις τιμές της *X*, όταν δίνονται οι τιμές της *Y*. Για να είναι αυτό δυνατόν, πρέπει να προσδιορίσουμε εξαρχής την *ευθεία ελαχίστων τετραγώνων* **της *X* πάνω στην *Y*,** $\hat{X} = \hat{\gamma} + \hat{\delta} \cdot Y$, η οποία γενικά είναι διαφορετική από την $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot X$. Και στις δύο όμως περιπτώσεις οι ευθείες διέρχονται από το σημείο $(\bar{x}, \bar{y})$.
8. Επισημαίνουμε ότι για δοσμένη τιμή x_i της *X*, η εκτίμηση $\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$ αφορά **τη μέση τιμή** $E(Y / X = x_i)$ της *Y* και **όχι την πραγματική τιμή** του *Y*.
9. Αξίζει να σημειωθεί ότι πάντα ισχύει $\sum_{i=1}^{\nu} \hat{\varepsilon}_i = 0$ αφού

$$\sum_{i=1}^{\nu} \hat{\varepsilon}_i = \sum_{i=1}^{\nu} (y_i - \hat{y}_i) = \sum_{i=1}^{\nu} (y_i - \hat{\alpha} - \hat{\beta} \cdot x_i) = \sum_{i=1}^{\nu} y_i - \nu \cdot \hat{\alpha} - \hat{\beta} \sum_{i=1}^{\nu} x_i = \nu \cdot (\bar{y} - \hat{\alpha} - \hat{\beta} \cdot \bar{x}) = 0$$

Παράδειγμα-1:
Ο πίνακας που ακολουθεί δίνει τη *ζήτηση ενός προϊόντος (Y)*, για διάφορα επίπεδα *διαφημιστικής δαπάνης (X).*

y_i *(σε χιλιάδες τεμάχια)*	x_i *(σε χιλιάδες €)*	$x_i \cdot y_i$	x_i^2
12	2	24	4
13	2	26	4
13	3	39	9
14	3	42	9
15	4	60	16
15	4	60	16
14	5	70	25
16	5	80	25
17	6	102	36
18	6	108	36
$\sum y_i = 147$	$\sum x_i = 40$	$\sum x_i \cdot y_i = 611$	$\sum x_i^2 = 180$

Είναι:

$$\bar{x} = \frac{40}{10} = 4$$

$$\bar{y} = \frac{147}{10} = 14{,}7$$

$$\hat{\beta} = \frac{\sum_{i=1}^{\nu} x_i \cdot y_i - \nu \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{\nu} x_i^2 - \nu \cdot \bar{x}^2} = \frac{611 - 10 \cdot 4 \cdot 14{,}7}{180 - 10 \cdot 4^2} = 1{,}15$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} = 14{,}7 - 1{,}15 \cdot 4 = 10{,}1$$

και άρα η εξίσωση της *ευθείας ελαχίστων τετραγώνων* είναι η
$\hat{Y} = 10{,}1 + 1{,}15 \cdot X$.

Ερμηνεία του $\hat{\beta}$

Επειδή $\hat{\beta} = 1{,}15 > 0$, αύξηση της διαφημιστικής δαπάνης συνεπάγεται αύξηση της ζήτησης του προϊόντος. Αν η διαφημιστική δαπάνη αυξηθεί κατά 1000 €, η μέση ζήτηση του προϊόντος εκτιμάται ότι θα αυξηθεί κατά 1,15 χιλιάδες τεμάχια.

Ερμηνεία του $\hat{\alpha}$
Για μηδενική διαφημιστική δαπάνη, η μέση ζήτηση του προϊόντος εκτιμάται ότι θα είναι 10,1 χιλιάδες τεμάχια. Επειδή η τιμή 0 είναι μακριά από το διάστημα μελέτης, η ερμηνεία του $\hat{\alpha}$ δεν έχει πρακτική αξία (δες και *Παρατήρηση-5*).

Θα υπολογίσουμε το συντελεστή προσδιορισμού r^2

x_i	y_i	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
2	12	12,4	-2,3	5,29	-2,7	7,29
2	13	12,4	-2,3	5,29	-1,7	2,89
3	13	13,55	-1,15	1,32	-1,7	2,89
3	14	13,55	-1,15	1,32	-0,7	0,49
4	15	14,7	0	0	0,3	0,09
4	15	14,7	0	0	0,3	0,09
5	14	15,85	1,15	1,32	-0,7	0,49
5	16	15,85	1,15	1,32	1,3	1,69
6	17	17	2,3	5,29	2,3	5,29
6	18	17	2,3	5,29	3,3	10,89
$\sum x_i = 40$	$\sum y_i = 147$			*SSR=26,44*		*SSTO=32,1*

$$r^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^{\nu}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{\nu}(y_i - \bar{y})^2} = \frac{26,44}{32,1} = 0,82$$

Ερμηνεία του r^2
Οι μεταβολές του ύψους της διαφημιστικής δαπάνης ερμηνεύουν το *82%* της μεταβλητότητας της ζήτησης του προϊόντος.

Ερμηνεία του $1 - r^2$
Το *18%* της μεταβλητότητας της ζήτησης του προϊόντος, οφείλεται σε τυχαία σφάλματα.

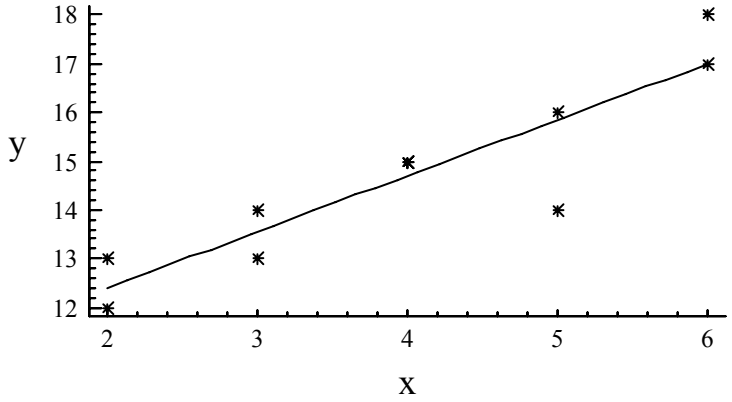
Το *τυπικό σφάλμα της εκτίμησης* s είναι

$$s = \sqrt{\frac{1}{\nu - 2} \cdot \sum_{i=1}^{\nu}(y_i - \hat{y}_i)^2} = \sqrt{\frac{SSE}{\nu - 2}} = \sqrt{\frac{32,1 - 26,44}{8}} = \sqrt{0,7} = 0,84$$

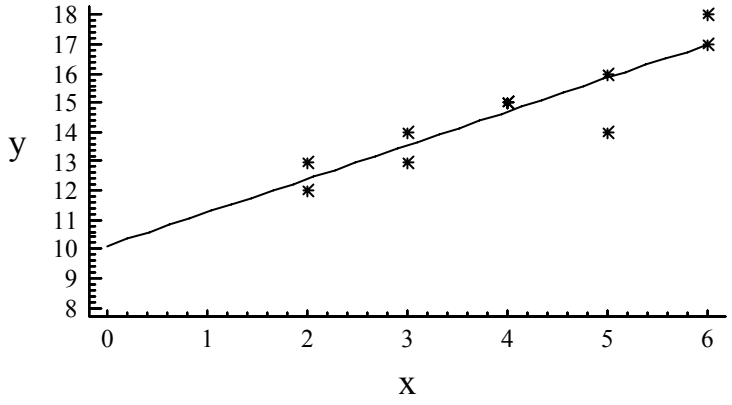
Πρόβλεψη με το μοντέλο $\hat{Y} = 10,1 + 1,15 \cdot X$ *που εκτιμήσαμε*
Αν το ύψος της διαφημιστικής δαπάνης είναι π.χ. 3,5 *χιλιάδες €*, η μέση ζήτηση του προϊόντος, εκτιμάται ότι θα είναι $10,1 + 1,15 \cdot 3,5 = 14,125$ χιλιάδες τεμάχια.

Αξιολόγηση του μοντέλου
Το μοντέλο $\hat{Y} = 10,1 + 1,15 \cdot X$ ερμηνεύει το *82%* της μεταβλητότητας της ζήτησης του προϊόντος.

Παρατήρηση:
Για συγκεκριμένη ζήτηση του προϊόντος, δε μπορούμε από το μοντέλο αυτό, να προβλέψουμε το απαιτούμενο ύψος διαφημιστικής δαπάνης.



Η *ευθεία ελαχίστων τετραγώνων* $\hat{Y} = 10{,}1 + 1{,}15 \cdot X$ έχει σχεδιασθεί με κατάλληλο πρόγραμμα υπολογιστή. Παρατηρείστε ότι η νοητή προέκταση της, δεν τέμνει τον άξονα των Y στο 10,1. Υπάρχει λάθος; Τι εξήγηση δίνετε; *(Δείτε και το επόμενο σχήμα σε συνδυασμό και με την Παρατήρηση-5)*



<u>*Παράδειγμα-2:*</u>
Μετρήσαμε το βάρος (σε gr) 32 νεογέννητων παιδιών και την αύξηση του βάρους τους τρεις μήνες μετά τη γέννησή τους. Η αύξηση του βάρους τους, εκφράζεται ως ποσοστό (%) του αρχικού τους βάρους. Έστω X το βάρος και Y η αύξηση του βάρους. Από τις τιμές (x_i, y_i) $i = 1,2,...,32$ πήραμε:

$\sum y_i = 2.279$, $\sum x_i = 107.280$, $\sum x_i \cdot y_i = 7.380.960$, $\sum x_i^2 = 368.892.000$, $\sum y_i^2 = 179.761$

Οι μεταβλητές X και Y είναι και οι δύο τυχαίες και επομένως μπορούμε ως μέτρο συσχέτισης να χρησιμοποιήσουμε το *συντελεστή γραμμικής συσχέτισης του Pearson r.*

Είναι:
$\bar{x} = 3.352{,}5$
$\bar{y} = 71{,}22$

$$s_{xy} = \frac{\sum x_i y_i - \nu \cdot \bar{x} \cdot \bar{y}}{\nu - 1} = \frac{7.380.960 - 32 \cdot 3.352,5 \cdot 71,22}{31} = -8.371,7$$

$$s_x{}^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{\nu} \chi_i^2 - \nu \cdot \bar{\chi}^2 \right) = \frac{368.892.000 - 32 \cdot 3.352,5^2}{31} = 29.7929,3 \text{ άρα}$$

$$s_x = \sqrt{29.7929,3} = 545,3$$

$$s_y{}^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{\nu} y_i^2 - \nu \cdot \bar{y}^2 \right) = \frac{179.761 - 32 \cdot 71,22^2}{31} = 562,83 \text{ άρα } s_y = \sqrt{562,83} = 23,72$$

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{-8.371,7}{5455,3 \cdot 23,72} = -0,65$$

Το αρνητικό πρόσημο του r δείχνει ότι αύξηση του βάρους των νεογέννητων συνεπάγεται ελάττωση του ποσοστού αύξησης του βάρους στο πρώτο τρίμηνο μετά τη γέννηση.

Θα εκτιμήσουμε την παλινδρόμηση της Y πάνω στη X.

$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{-8.371,7}{29.7929,3} = -0,0281 \text{ και } \hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} = 71,22 + 0.0281 \cdot 3.352,5 = 165,42.$$

Άρα η εξίσωση ελαχίστων τετραγώνων της Y πάνω στη X είναι: $\hat{Y} = 165,42 - 0,0281 \cdot X$.

Ερμηνεία του $\hat{\beta}$

Αύξηση του βάρους γέννησης κατά ένα gr εκτιμάται ότι θα προκαλέσει μείωση του μέσου ποσοστού αύξησης τους βάρους το πρώτο τρίμηνο μετά τη γέννηση κατά 0,0281%.

Ερμηνεία του $\hat{\alpha}$

Για μηδενικό βάρος γέννησης (!!!) το μέσο ποσοστό αύξησης του βάρους το πρώτο τρίμηνο μετά τη γέννηση εκτιμάται ότι θα είναι 165,42 %. Επειδή η τιμή 0 είναι μακριά από το διάστημα μελέτης (και … όχι μόνο) η ερμηνεία του $\hat{\alpha}$ δεν έχει πρακτική αξία (δες και *Παρατήρηση-5*).

Αξιολόγηση του μοντέλου.

Ο συντελεστής προσδιορισμού είναι: $r^2 = (-0,65)^2 = 0,42$. Δηλαδή, οι μεταβολές στο βάρος κατά τη γέννηση ερμηνεύουν το 42% της μεταβλητότητας του ποσοστού αύξησης του βάρους στο πρώτο τρίμηνο μετά τη γέννηση.

Παρατήρηση:

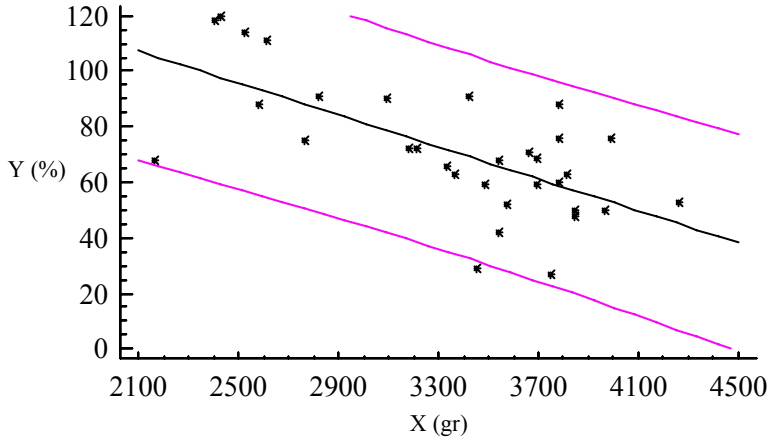
Επειδή είναι δυσνόητη η ειδική εννοιολογική ερμηνεία της τετραγωνικής ρίζας ενός ποσοστού (όπως ο συντελεστής προσδιορισμού r^2), για σκοπούς ερμηνείας προτιμάται η χρησιμοποίηση του συντελεστή προσδιορισμού r^2 παρά του συντελεστή γραμμικής συσχέτισης r. Το r^2 ως μη αρνητικός αριθμός μικρότερος ή ίσος του 1 έχει τετραγωνική ρίζα αριθμό μεγαλύτερό του (εκτός από την περίπτωση που είναι 0 ή 1) και συνεπώς, αν για την αξιολόγηση του μοντέλου προτιμηθεί η τετραγωνική του ρίζα (δηλ. το r) υπάρχει

κίνδυνος υπερεκτίμησής του. Για παράδειγμα αν $r^2 = 0{,}49$ *το r θα είναι 0,7. Δηλαδή, ενώ το μοντέλο ερμηνεύει τη μεταβλητότητα σε ποσοστό μικρότερο του 50% ο συντελεστής συσχέτισης δείχνει ισχυρή γραμμική συσχέτιση.*

Το τυπικό σφάλμα της εκτίμησης είναι,

$$s = \sqrt{\frac{\nu-1}{\nu-2}\cdot s_y^2(1-r^2)} = \sqrt{\frac{31}{30}\cdot 562{,}83\cdot(1-0{,}42)} = 18{,}36\,.$$

Επειδή $\nu = 32 > 30$, αν φέρουμε δύο ευθείες παράλληλες προς την ευθεία ελαχίστων τετραγώνων και σε κατακόρυφες προς αυτήν αποστάσεις $18{,}36$, $2\cdot 18{,}36$, $3\cdot 18{,}36$ τότε, μεταξύ των δύο αυτών ευθειών θα βρίσκεται περίπου το 68%, το 95% και το 99,7% των σημείων του διαγράμματος διασποράς αντίστοιχα.



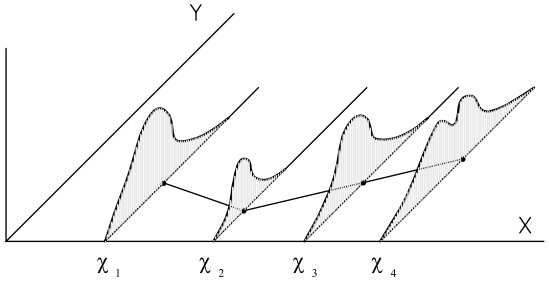
Μεταξύ των δύο παράλληλων βρίσκεται περίπου το 95% των σημείων του διαγράμματος διασποράς (η κάθε ευθεία έχει κατακόρυφη απόσταση από την ευθεία ελαχίστων τετραγώνων $2\cdot 18{,}36 = 36{,}72$).

ΠΑΡΑΡΤΗΜΑ Α

Προϋποθέσεις-παραδοχές για την εφαρμογή του Απλού Γραμμικού Μοντέλου

$$\boldsymbol{Y = \alpha + \beta \cdot X + \varepsilon}$$

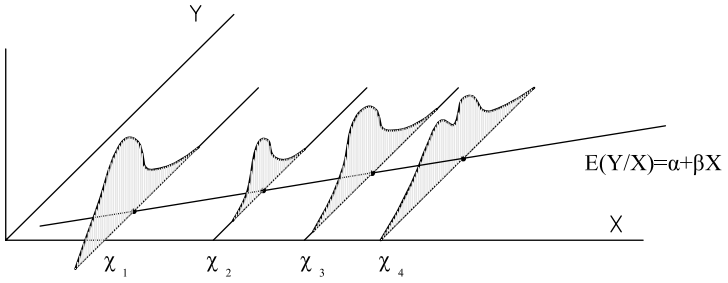
Η γενική υπόθεση-παραδοχή που κάνουμε για ένα μοντέλο παλινδρόμησης (γραμμικό ή όχι), είναι ότι η μεταβλητή *X* μετράται χωρίς σφάλμα και ότι η *Y,* για κάθε επίπεδο x_i της *X*, είναι τυχαία μεταβλητή με πεπερασμένη μέση τιμή και διασπορά.



Για το ***απλό γραμμικό μοντέλο*** κάνουμε επιπλέον τις ακόλουθες υποθέσεις-παραδοχές:

<u>Υπόθεση 1: Γραμμικότητα (Linearity)</u>

Η κατανομή της *Y* έχει, για τα διάφορα επίπεδα x_i $i = 1,2,...,\nu$ της *X,* μέση τιμή $E(Y / X = x_i) = \alpha + \beta \cdot x_i$ ή $E(Y / X) = \alpha + \beta \cdot X$, όπου, α και β παράμετροι που εκτιμώνται από το δείγμα (x_i, y_i) $i = 1,2,...,\nu$. Δηλαδή, υποθέτουμε ότι οι μέσες τιμές της *Y,* για τα διάφορα επίπεδα της *X,* είναι γραμμικές συναρτήσεις της *X* (ότι βρίσκονται δηλαδή σε ευθεία γραμμή). Σημειώνουμε ότι στο μοντέλο $Y = \alpha + \beta \cdot X + \varepsilon$, τυχαίες μεταβλητές είναι μόνο οι *Y* και *ε*.

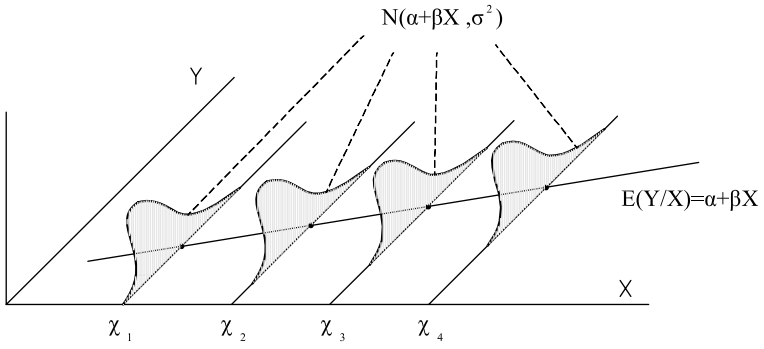


<u>Υπόθεση 2: Ομοσκεδαστικότητα-Σταθερότητα Διασποράς (Homoscedasticity - Variance Stability)</u>

Οι κατανομές της *Y* έχουν ίδια διασπορά για όλα τα επίπεδα της *X,* δηλαδή, $Var(Y / X = x_i) = \sigma^2$. Ένα παράδειγμα παραβίασης της υπόθεσης αυτής *(heteroscedasticity)* φαίνεται στο προηγούμενο σχήμα (η διασπορά της *Y*, π.χ. στο επίπεδο x_1, είναι μεγαλύτερη από τη διασπορά της *Y* στο επίπεδο x_2).

Υπόθεση 3: Ανεξαρτησία (Independence)
Οι τιμές της Y που αντιστοιχούν στα διάφορα επίπεδα της X είναι ανεξάρτητες μεταξύ τους.

Υπόθεση 4: Κανονικότητα (Normality)
Η κατανομή της Y για όλα τα επίπεδα της X είναι κανονική.

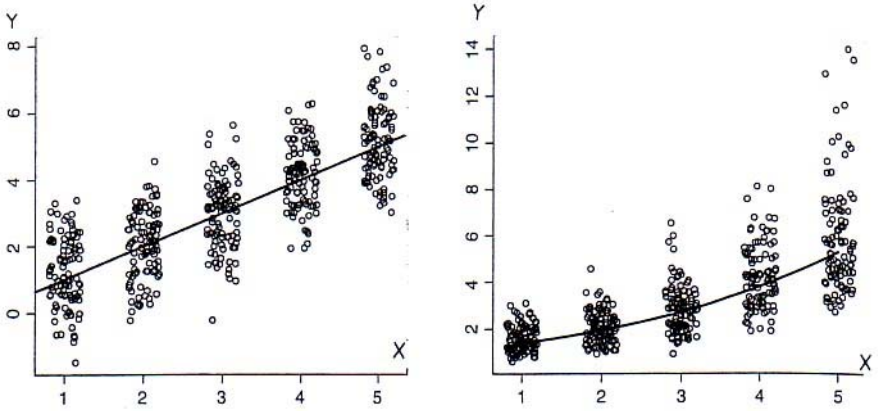


Με βάση τις παραπάνω υποθέσεις για την τυχαία μεταβλητή Y, για την τυχαία μεταβλητή $\varepsilon = Y - (\alpha + \beta \cdot X)$ (δηλαδή για τα σφάλματα-residuals) δεχόμαστε ότι:

1. $\varepsilon \sim N(0, \sigma^2)$
2. Οι τιμές της ε που αντιστοιχούν στα διάφορα επίπεδα της X είναι μεταξύ τους ανεξάρτητες.

Στη συνέχεια, παρουσιάζουμε ορισμένες μεθόδους (γραφικές κυρίως) για τον έλεγχο των παραπάνω προϋποθέσεων-παραδοχών προσαρμογής του απλού γραμμικού μοντέλου. Οι παραδοχές αυτές αποτελούν την αναγκαία μαθηματική (πιθανοθεωρητική) βάση για την εφαρμογή μεθόδων της στατιστικής συμπερασματολογίας (π.χ. έλεγχοι υποθέσεων, διαστήματα εμπιστοσύνης). Ο έλεγχος επομένως αυτών των παραδοχών είναι αναγκαίος προκειμένου να αποφεύγουμε λανθασμένες διαδικασίες εξαγωγής συμπερασμάτων για τον πληθυσμό.

Ένας πρώτος, άμεσος, έλεγχος μπορεί να γίνει με προσεκτική παρατήρηση του ***διάγραμματος διασποράς*** του δείγματος. Ας δούμε δύο παραδείγματα:



Στο πρώτο *διάγραμμα διασποράς* (αριστερά) φαίνεται ότι για όλα τα επίπεδα της X,

- οι κατανομές της Y είναι συμμετρικές και έχουν σταθερή διασπορά

- οι αναμενόμενες μέσες τιμές της Y βρίσκονται σε ευθεία γραμμή.

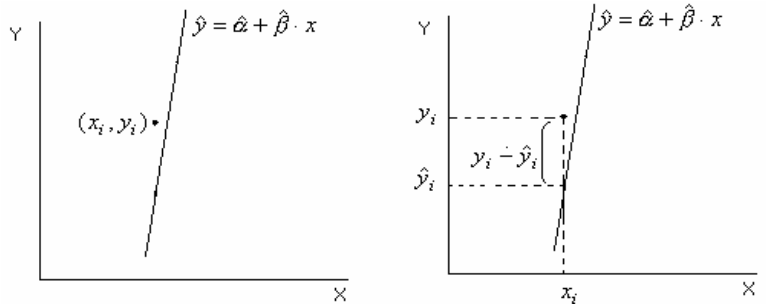
Στο δεύτερο *διάγραμμα διασποράς* (δεξιά) φαίνεται ότι,

- οι κατανομές της Y για τα διάφορα επίπεδα της X δεν είναι συμμετρικές και ούτε έχουν σταθερή διασπορά. Μάλιστα, φαίνεται ότι αυξανομένου του X αυξάνεται η διασπορά καθώς και η ασσυμετρία (θετική) της κατανομής του Y
- οι αναμενόμενες μέσες τιμές της Y για τα διάφορα επίπεδα της X δεν βρίσκονται σε ευθεία γραμμή αλλά σε καμπύλη.

Ας δούμε πιο αναλυτικά, ανά υπόθεση, πώς μπορούμε να διαπιστώσουμε και να αντιμετωπίσουμε πιθανές παραβιάσεις.

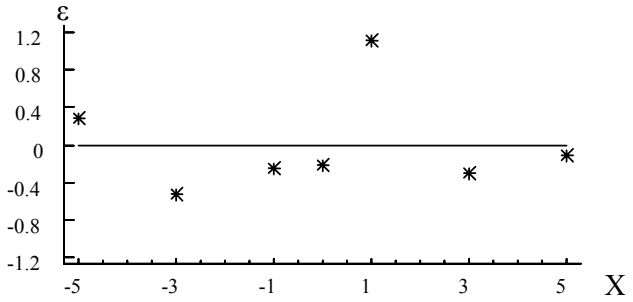
1. Γραμμικότητα (Linearity)

Ένας πρώτος έλεγχος της γραμμικότητας μπορεί να γίνει γραφικά με το *διάγραμμα διασποράς*. Είναι όμως δυνατόν, ιδίως όταν η κλίση της ευθείας παλινδρόμησης που προσεγγίζει τα δεδομένα είναι μεγάλη, να μας δίνεται η εντύπωση ότι τα σημεία (x_i, y_i) είναι κοντά στην ευθεία παλινδρόμησης ενώ στην πραγματικότητα δεν είναι! (Δείτε τα παρακάτω σχήματα και, επίσης, θυμηθείτε με βάση ποιο κριτήριο εκτιμώνται οι παράμετροι α και β της πληθυσμιακής ευθείας παλινδρόμησης $E(Y/X) = \alpha + \beta \cdot X$.)

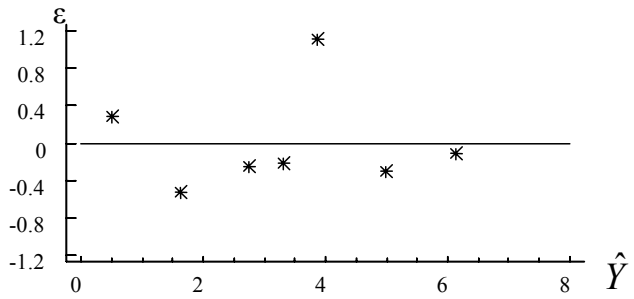


Για το λόγο αυτό, συνήθως, χρησιμοποιούμε τα ***διαγράμματα υπολοίπων (residual plots)*** όπου, αντί των (x_i, y_i) αναπαρίστανται γραφικά τα $(x_i, \hat{\varepsilon}_i)$ ή τα $(\hat{y}_i, \hat{\varepsilon}_i)$ (όπου $\hat{\varepsilon}_i = y_i - \hat{y}_i$ τα υπόλοιπα-σφάλματα).

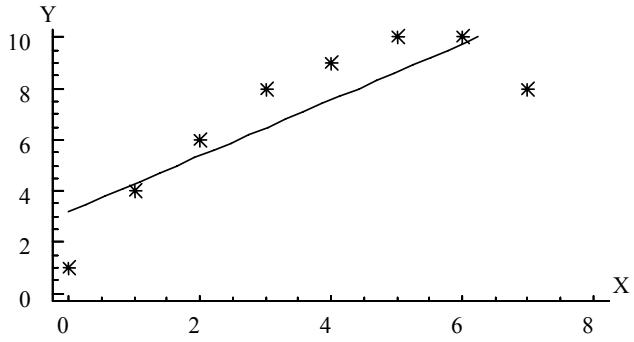
Αν στο διάγραμμα υπολοίπων, τα σημεία $(x_i, \hat{\varepsilon}_i)$ (ή τα $(\hat{y}_i, \hat{\varepsilon}_i)$) δεν ακολουθούν κάποιο πρότυπο (κάποια συστηματική τάση) αλλά είναι τυχαία διεσπαρμένα σε μια οριζόντια ζώνη γύρω από την ευθεία $\varepsilon = 0$, τότε η επιλογή γραμμικού μοντέλου δικαιολογείται.



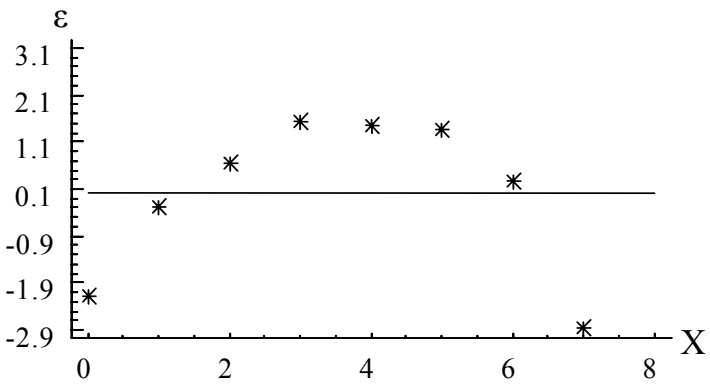
Τα διαγράμματα υπολοίπων συνήθως παρουσιάζουν την ίδια εικόνα και όταν τα υπόλοιπα $\hat{\varepsilon}_i$ παρασταθούν γραφικά συναρτήσει των προσαρμοσμένων τιμών $\hat{y}_i$.



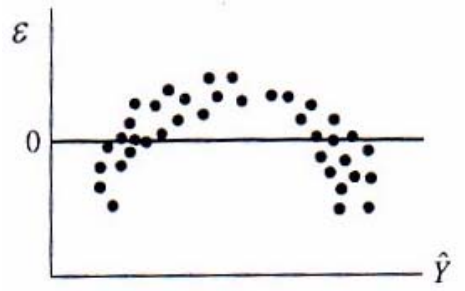
Στο ακόλουθο παράδειγμα, η προσαρμογή της ευθείας $\hat{Y} = 3.17 - 1.09 \cdot X$



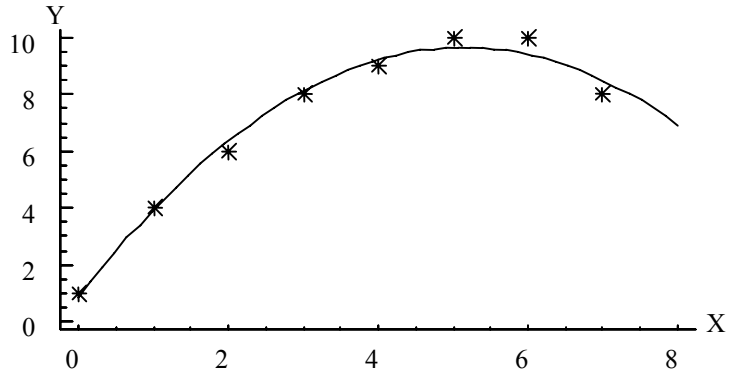
δίνει το ακόλουθο διάγραμμα υπολοίπων:



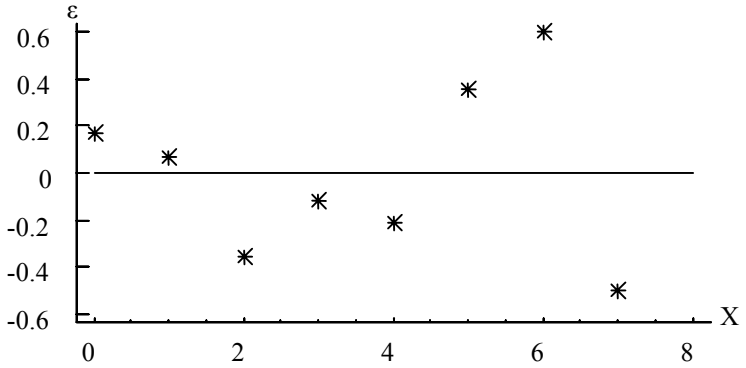
Παρατηρούμε ότι αυξανόμενου του X τα υπόλοιπα δεν συγκεντρώνονται τυχαία γύρω από την ευθεία $\varepsilon = 0$, αλλά ακολουθούν ένα κυκλικό πρότυπο (αρνητικές-θετικές-αρνητικές τιμές). Αυτή η κυκλική συμπεριφορά (βλ. και επόμενο σχήμα) φανερώνει παλινδρόμηση δεύτερου βαθμού ως προς X $(Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \varepsilon)$.



Έτσι, αν στα ίδια δεδομένα προσαρμοσθεί η παραβολή $\hat{Y} = 0.83 + 3.42 \cdot X - 0.33 \cdot X^2$



τα υπόλοιπα συγκεντρώνονται τυχαία σε μια οριζόντια ζώνη γύρω από την ευθεία *ε = 0.*



Η καταλληλότητα ή όχι του γραμμικού μοντέλου ελέγχεται και με το ποσοστό της μεταβλητότητας του *Y* που εξηγείται από την παλινδρόμηση, δηλαδή, με το συντελεστή προσδιορισμού r^2. Στο προηγούμενο παράδειγμα, το μοντέλο $\hat{Y} = 3.17 - 1.09 \cdot X$ δίνει $r^2 = 72\%$ ενώ το μοντέλο $\hat{Y} = 0.83 + 3.42 \cdot X - 0.33 \cdot X^2$ δίνει $r^2 = 98.6\%$. Μπορεί επίσης να ελεγχθεί με το Lack-of-Fit test.

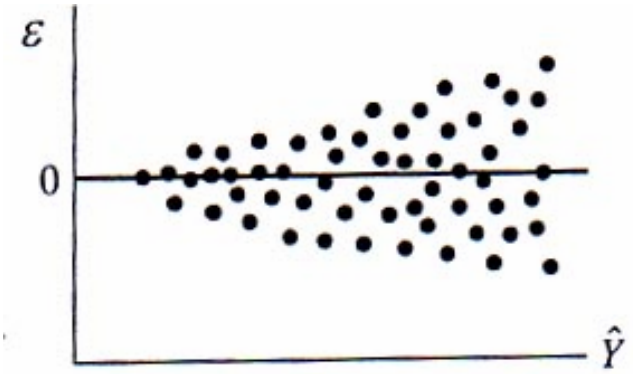
Όταν διαπιστώνεται ότι η σχέση μεταξύ *X* και *Y* είναι μη γραμμική, σε αρκετές περιπτώσεις είναι δυνατόν, με κατάλληλους μετασχηματισμούς στα *X* ή/και στα *Y* να προκύψει γραμμική σχέση. Έχουμε έτσι τη δυνατότητα να αξιοποιήσουμε τη στατιστική θεωρία του γραμμικού μοντέλου και σε μη γραμμικά μοντέλα (αφού, αντιστρέφοντας στη συνέχεια τις μετασχηματισμένες μεταβλητές, μπορούμε να πάρουμε τα ζητούμενα συμπεράσματα για τις αρχικές). Στο Παράρτημα Β΄ δίνουμε παραδείγματα τέτοιων μετασχηματισμών. Γενικά, η στατιστική μελέτη μη γραμμικών μοντέλων, με εξαίρεση τα πολυωνυμικά, παραμένει δύσκολο και ανοικτό πρόβλημα.

2. <u>Ομοσκεδαστικότητα ή Σταθερότητα Διασποράς (Homoscedasticity-Variance Stability)</u>

Ένας πρώτος έλεγχος της σταθερότητας ή μη της διασποράς της *Y* (ή της *ε*) για τα διάφορα επίπεδα της *X* μπορεί να γίνει με το *διάγραμμα διασποράς* και τα *διαγράμματα υπολοίπων*. Αν για παράδειγμα, το διάγραμμα υπολοίπων έχει μορφή τραπεζίου (ανοιχτής βεντάλιας), όπως το παρακάτω, η πιο πιθανή αιτία αυτής της διαταραχής[7] είναι η μη σταθερότητα της διασποράς των τυχαίων σφαλμάτων *ε*. Σε πολλές

[7] *Της απόκλισης από την τυχαία συγκέντρωση των σημείων γύρω από την ευθεία ε = 0*

οικονομικές και εμπορικές εφαρμογές η μεταβολή της διασποράς σ^2 με το X ή με το $\hat{Y}$ δίνει διαγράμματα υπολοίπων μορφής τραπεζίου (αυξανομένου του X ή του $\hat{Y}$, αυξάνει το σ^2 ή αντιστρόφως). Αυτό συμβαίνει διότι τέτοιες εφαρμογές ακολουθούν πολλαπλασιαστικά μοντέλα όπου $\sigma_Y^2 = [E(Y)]^2 \cdot \sigma^2$ και σ^2 η διασπορά των σφαλμάτων ε (γιατί;)[8]. Επίσης, ανάλογα διαγράμματα υπολοίπων δίνουν μεταβλητές που μετρούν αριθμό συμβάντων στη μονάδα χρόνου, χώρου, μήκους, κ.τλ. δηλαδή μεταβλητές που ακολουθούν κατανομή Poisson (γιατί ;)[9].



Αν από τα *διαγράμματα υπολοίπων* δημιουργούνται υπόνοιες ότι δεν έχουμε σταθερές διασπορές, μπορούμε να ελέγξουμε στατιστικά αν υπάρχει σημαντική διαφορά στις διασπορές ή όχι εφόσον για τα διάφορα επίπεδα της X έχουμε περισσότερες της μιας παρατηρήσεις. Μπορούμε, επίσης, να ταξινομήσουμε τις παρατηρήσεις σε αύξουσα σειρά των X, να τις χωρίσουμε σε δύο ή περισσότερες ομάδες και να ελέγξουμε στατιστικά αν οι ομάδες έχουν σημαντική διαφορά στις διασπορές ή όχι.

Όταν διαπιστώνεται μη σταθερότητα διασπορών μπορούμε, σε αρκετές περιπτώσεις, να αντιμετωπίσουμε το πρόβλημα με κατάλληλους μετασχηματισμούς στις μεταβλητές. Στο Παράρτημα Β΄ δίνουμε παραδείγματα τέτοιων μετασχηματισμών.

3. Ανεξαρτησία (Independence)

Εξαρτημένα Y εμφανίζονται συνήθως όταν παίρνουμε παρατηρήσεις από την ίδια πειραματική μονάδα σε διαφορετικές χρονικές στιγμές (π.χ. μετράμε την πίεση ή το βάρος του ιδίου ατόμου ανά εβδομάδα). Επίσης, σε περιπτώσεις όπου χρησιμοποιούνται μηχανές (όργανα μέτρησης, κ.τλ) που αλλάζει η απόδοσή τους με τη χρήση ή ο χειριστής βελτιώνεται (ή χειροτερεύει) με την πάροδο του χρόνου. Είναι επομένως χρήσιμο, όταν έχουμε πειραματικά δεδομένα που παίρνονται με χρονική σειρά, να κάνουμε ένα διάγραμμα υπολοίπων ως προς το χρόνο έστω και αν ο χρόνος δεν χρησιμοποιείται ως μεταβλητή στο μοντέλο. Αν το διάγραμμα υπολοίπων έχει τη μορφή του παρακάτω σχήματος τότε είναι πιθανόν να υπάρχει στοχαστική εξάρτηση μεταξύ των σφαλμάτων. Στη συνέχεια, πρέπει να ελέγξουμε στατιστικά την υπόνοια αυτή με το Durbin-Watson test. Αν διαπιστωθεί εξάρτηση των τιμών της Y τότε για την προσαρμογή κατάλληλου μοντέλου και την εξαγωγή στατιστικών συμπερασμάτων πρέπει να χρησιμοποιηθούν ειδικές μέθοδοι.

[8] *Στο πολλαπλασιαστικό μοντέλο έχουμε* $Y = E(Y) \cdot \varepsilon$ *ενώ στο προσθετικό έχουμε* $Y = E(Y) + \varepsilon$

[9] *Θυμηθείτε ότι αν η Y ακολουθεί κατανομή Poisson τότε* $\sigma_Y^2 = E(Y)$

4. Κανονικότητα (Normality)

Η κανονικότητα μπορεί να ελεγχθεί με διάφορους τρόπους όπως:
Με *ιστόγραμμα*
Με *φυλλογράφημα (steam and leaf plot)*
Με *θηκόγραμμα (box plot)*
Με *διάγραμμα πιθανοτήτων (normal probability plot)*
Με στατιστικούς ελέγχους καλής προσαρμογής (*goodness–of-fit test*) όπως *Kolmogorov-Smirnov test* ή X^2 *test.*

Όταν διαπιστώνεται παραβίαση της κανονικότητας μπορούμε, σε αρκετές περιπτώσεις, να αντιμετωπίσουμε το πρόβλημα με κατάλληλους μετασχηματισμούς στις μεταβλητές. Στο Παράρτημα Β΄ δίνουμε παραδείγματα τέτοιων μετασχηματισμών.

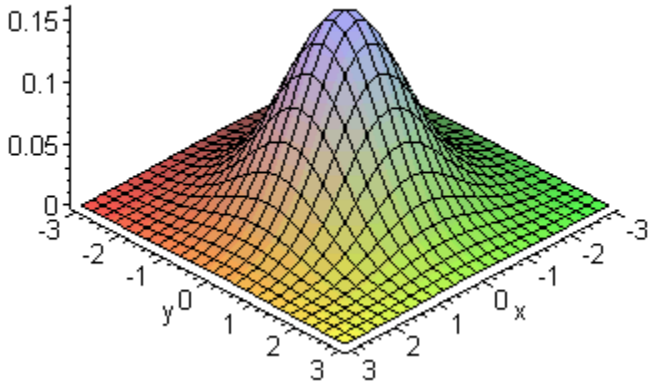
Πέραν των παραπάνω υποθέσεων-παραδοχών, είναι χρήσιμο να ελέγχουμε την ύπαρξη ή μη *ακραίων παρατηρήσεων* (*outliers*). Οι *ακραίες παρατηρήσεις* μπορούν να ανιχνευθούν αποτελεσματικά με το *θηκόγραμμα* των παρατηρήσεων ή και με το *διάγραμμα υπολοίπων*. Αν διαπιστωθεί ακραία παρατήρηση, πρέπει πρώτα να ερευνηθεί αν οφείλεται σε λανθασμένη παρατήρηση ή πιθανόν σε απότομη στιγμιαία διαταραχή του συστήματος που παρατηρούμε. Αν αυτό συμβαίνει, πρέπει να παραληφθεί από το δείγμα. Αν όμως η ακραία παρατήρηση ανήκει στον πληθυσμό είναι λάθος να παραληφθεί από το δείγμα. Η γενική αρχή που πρέπει να τηρούμε είναι ότι ποτέ δεν απορρίπτουμε μια ακραία παρατήρηση αν δεν είμαστε βέβαιοι ότι πρόκειται για λάθος ή απότομη στιγμιαία διαταραχή. Έγκυρες ακραίες παρατηρήσεις μπορεί να αποδειχθούν οι πλέον ενδιαφέρουσες!

Υπόθεση για την εφαρμογή του απλού γραμμικού μοντέλου παλινδρόμησης σε μη πειραματικά δεδομένα

Για την ανάπτυξη της στατιστικής θεωρίας του απλού γραμμικού μοντέλου $Y = \alpha + \beta \cdot X + \varepsilon$, υποθέσαμε ότι η μεταβλητή X **δεν είναι τυχαία** (μετράται χωρίς σφάλμα) και ότι τυχαίες μεταβλητές είναι μόνο οι Y και ε. Αυτή η υπόθεση ικανοποιείται στις **πειραματικές έρευνες** όπου ο ερευνητής ελέγχει (καθορίζει) τις τιμές της X και παρατηρεί πώς οι μεταβολές στις τιμές της X αντανακλώνται στην Y.

Σε **μη πειραματικές έρευνες (δειγματοληψίες)**, όπου ο ερευνητής επιλέγει ένα τυχαίο δείγμα (x_i, y_i) $i = 1,2,...\nu$, δηλαδή, όταν όχι μόνο η Y αλλά και η X είναι τυχαία μεταβλητή, τότε με την υπόθεση ότι η από κοινού κατανομή των X και Y είναι **διδιάστατη κανονική κατανομή,** μπορούμε και πάλι να εφαρμόσουμε τη θεωρία του απλού γραμμικού μοντέλου και να υπολογίσουμε την ευθεία ελαχίστων τετραγώνων της Y πάνω στην X ή της X πάνω στην Y διότι από τη θεωρία πιθανοτήτων είναι γνωστό ότι οι δεσμευμένες κατανομές της Y δεδομένης της X και της X δεδομένης της Y είναι κανονικές με

$\mu_{Y/X} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)$ και $\mu_{X/Y} = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y)$ αντίστοιχα.



ΠΑΡΑΡΤΗΜΑ Β
Μετασχηματισμοί
Σταθεροποίησης Διασπορών – Κανονικοποίησης - Γραμμικοποίησης

Κατά τη διερεύνηση της σχέσης μεταξύ δύο μεταβλητών X και Y για την εφαρμογή του γραμμικού μοντέλου παλινδρόμησης, πολλές φορές, διαπιστώνεται παραβίαση μιας ή και περισσότερων εκ των προϋποθέσεων-παραδοχών εφαρμογής της αντίστοιχης στατιστικής θεωρίας. Σε αρκετές περιπτώσεις, μπορούμε να αντιμετωπίσουμε αυτά τα προβλήματα με κατάλληλους μετασχηματισμούς των μεταβλητών.

Πιο συγκεκριμένα, υπάρχουν τρεις βασικοί λόγοι για την αναζήτηση κατάλληλων μετασχηματισμών των μεταβλητών:

1. Για τη **σταθεροποίηση των διασπορών**, όταν παραβιάζεται η παραδοχή της ομοσκεδαστικότητας. Δηλαδή, όταν οι διασπορές της εξαρτημένης μεταβλητής Y δεν είναι ίσες για τα διάφορα επίπεδα της X.
2. Για την **κανονικοποίηση**, όταν οι κατανομές της εξαρτημένης μεταβλητής Y για τα διάφορα επίπεδα της X δεν είναι κανονικές.
3. Για την **γραμμικοποίηση**, όταν τα αρχικά δεδομένα υποδεικνύουν όχι γραμμικό αλλά μη γραμμικό μοντέλο (είτε ως προς τις παραμέτρους παλινδρόμησης είτε ως προς τις μεταβλητές).

Παρότι, για τους ενδεικνυόμενους κατά περίπτωση μετασχηματισμούς, υπάρχει πλούσια βιβλιογραφία, εντούτοις, η αναζήτηση κατάλληλων μετασχηματισμών, για το συγκεκριμένο κάθε φορά πρόβλημα, απαιτεί αρκετή σχετική εμπειρία. Απαιτεί επίσης καλή γνώση της φύσης του υπό μελέτη προβλήματος, ιδιαίτερα όταν τα δεδομένα παραβιάζουν (δεν υποστηρίζουν) περισσότερες από μία προϋποθέσεις-παραδοχές. Γιατί σε αυτή την περίπτωση, είναι δυνατόν, μετασχηματισμοί που προσφέρονται για την άρση μιας παραβίασης να μην προσφέρονται για την άρση των άλλων ή και να δημιουργούν νέες.

Στη συνέχεια, σταχυολογούμε από τη βιβλιογραφία κάποιες χαρακτηριστικές περιπτώσεις ενδεικνυόμενων μετασχηματισμών.

1. Λογαριθμικοί μετασχηματισμοί

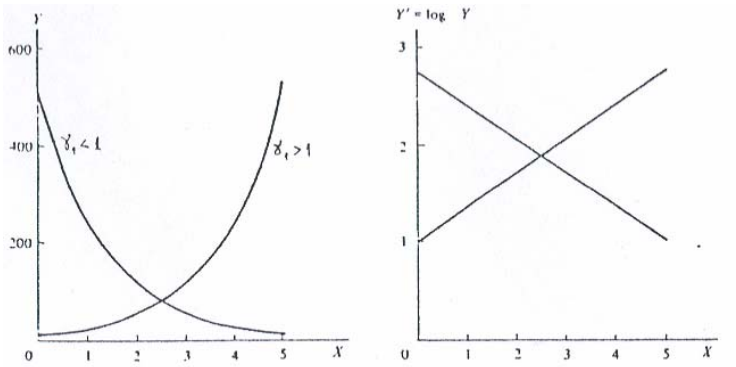
Ο λογαριθμικός μετασχηματισμός $\ln(Y) = Y'$ ενδείκνυται:

α) για <u>σταθεροποίηση της διασποράς της *Y*</u>, όταν αυξάνεται με το *Y*.

β) για <u>κανονικοποίηση της *Y*</u>, όταν η κατανομή των υπολοίπων παρουσιάζει θετική ασυμμετρία.

γ) για <u>γραμμικοποίηση του μοντέλου</u> όταν τα αρχικά δεδομένα υποδεικνύουν το πολλαπλασιαστικό μοντέλο:

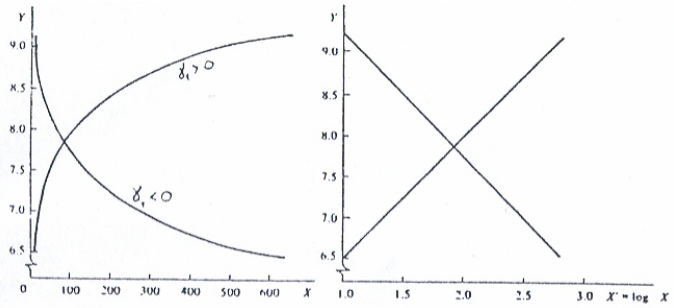
$$Y = \gamma_0 \cdot \gamma_1^X \cdot \varepsilon .$$



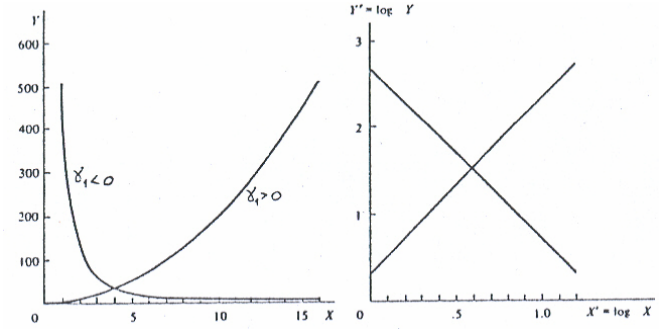
Στην περίπτωση αυτή, το αρχικό μοντέλο (αριστερά) μετασχηματίζεται στο γραμμικό (δεξιά): $Y' = \alpha + \beta \cdot X + \varepsilon'$, όπου $Y' = \ln(Y)$, $\alpha = \ln(\gamma_0)$, $\beta = \ln(\gamma_1)$, $\varepsilon' = \ln(\varepsilon)$

Με λογαριθμικούς μετασχηματισμούς γίνεται, επίσης, γραμμικοποίηση των πολλαπλασιαστικών μοντέλων:

$$e^Y = \gamma_0 \cdot X^{\gamma 1} \cdot \varepsilon \quad \text{(με το μετασχηματισμό } \ln(X) = X' \text{)}$$



και $Y = \gamma_0 \cdot X^{\gamma 1} \cdot \varepsilon$ (με το μετασχηματισμό $\ln(Y) = Y'$ και $\ln(X) = X'$)



2. Αντίστροφοι μετασχηματισμοί

Ο αντίστροφος μετασχηματισμός $\frac{1}{Y} = Y'$ ενδείκνυται:

α) για <u>σταθεροποίηση της διασποράς της *Y*,</u> όταν έχουμε μεγάλη αύξηση της διασποράς πάνω από κάποια τιμή του *Y*.

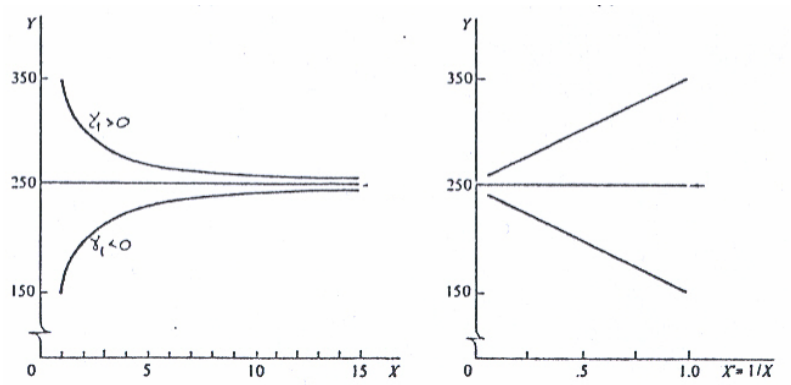
β) για <u>γραμμικοποίηση του μοντέλου</u> όταν τα αρχικά δεδομένα υποδεικνύουν το αντίστροφο μοντέλο: $Y = \frac{1}{\gamma_0 + \gamma_1 \cdot X + \varepsilon}$.

Στην περίπτωση αυτή, το αρχικό μοντέλο μετασχηματίζεται στο γραμμικό:

$$Y' = \alpha + \beta \cdot X + \varepsilon', \text{ όπου } Y' = \frac{1}{Y},\ \alpha = \gamma_0,\ \beta = \gamma_1,\ \varepsilon' = \varepsilon$$

Με τον αντίστροφο μετασχηματισμό $\frac{1}{X} = X'$ γίνεται, γραμμικοποίηση του αντίστροφου μοντέλου:

$$Y = \gamma_0 + \gamma_1 \cdot \frac{1}{X} + \varepsilon$$



3. Μετασχηματισμοί τετραγωνικής ρίζας

Ο μετασχηματισμός $\sqrt{Y} = Y'$ ενδείκνυται:

α) για <u>σταθεροποίηση της διασποράς της *Y*,</u> όταν η διασπορά είναι ανάλογη της μέσης τιμής της *Y*.

β) για <u>γραμμικοποίηση του μοντέλου</u> όταν τα αρχικά δεδομένα υποδεικνύουν το μοντέλο: $Y = (\gamma_0 + \gamma_1 \cdot X + \varepsilon)^2$.

Στην περίπτωση αυτή, το αρχικό μοντέλο μετασχηματίζεται στο γραμμικό:

$$Y' = \alpha + \beta \cdot X + \varepsilon', \text{ όπου } Y' = \sqrt{Y},\ \alpha = \gamma_0,\ \beta = \gamma_1,\ \varepsilon' = \varepsilon$$

Με τον μετασχηματισμό $\sqrt{X} = X'$ γίνεται γραμμικοποίηση του μοντέλου: $Y = \gamma_0 + \gamma_1 \cdot \sqrt{X} + \varepsilon$.

4. Μετασχηματισμός $Y^2 = Y'$

Ο μετασχηματισμός αυτός ενδείκνυται:

α) για σταθεροποίηση της διασποράς της *Y*, όταν ελαττώνεται με τη μέση τιμή της *Y*.

β) για κανονικοποίηση της *Y*, όταν η κατανομή των υπολοίπων παρουσιάζει αρνητική ασυμμετρία.

γ) για γραμμικοποίηση του μοντέλου όταν τα αρχικά δεδομένα υποδεικνύουν καμπυλόγραμμο μοντέλο π.χ. $Y = \sqrt{\gamma_0 + \gamma_1 \cdot X + \varepsilon}$. Στην περίπτωση αυτή, το αρχικό μοντέλο μετασχηματίζεται στο γραμμικό: $Y' = \alpha + \beta \cdot X + \varepsilon'$, όπου $Y' = Y^2$, $\alpha = \gamma_0$, $\beta = \gamma_1$, $\varepsilon' = \varepsilon$.

Οι παραπάνω μετασχηματισμοί μπορούν φυσικά να συνδυασθούν για την αντιμετώπιση πιο πολύπλοκων περιπτώσεων. Για παράδειγμα το μη γραμμικό μοντέλο

$$Y = \frac{1}{1 + e^{\gamma_0 + \gamma_1 \cdot X + \varepsilon}}$$

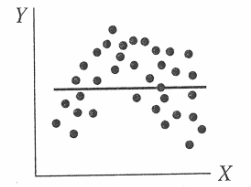
εύκολα μετασχηματίζεται σε γραμμικό με το μετασχηματισμό $Y' = \ln(\frac{1}{Y} - 1)$ που είναι ένας αντίστροφος και ένας λογαριθμικός μετασχηματισμός (διαδοχικά).

ΠΑΡΑΡΤΗΜΑ Γ
Επισημάνσεις - Σχόλια - Διευκρινήσεις

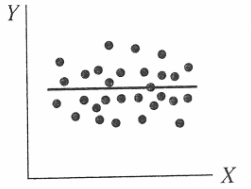
1. Ερμηνεία του ελέγχου της υπόθεσης $H_0 : \beta = 0$ έναντι της $H_1 : \beta \neq 0$ για την κλίση της ευθείας παλινδρόμησης $Y = \alpha + \beta \cdot X + \varepsilon$.

α) Όταν **δεν απορρίπτεται** η μηδενική υπόθεση, τότε[10] συμβαίνει ένα από τα παρακάτω:

- Η σχέση μεταξύ X και Y ***δεν είναι γραμμική***



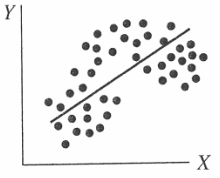
- Πρόκειται για το μοντέλο $E(Y/X) = E(Y) = \alpha$. Δηλαδή, για την περίπτωση όπου η X δεν συνεισφέρει στην πρόβλεψη της $E(Y/X)$. Έτσι, η εκτίμηση $\hat{Y} = \bar{y} + \hat{\beta} \cdot (X - \bar{x})$ προβλέπει τη μέση τιμή της Y όσο και η $\hat{Y} = \bar{y}$.



β) Όταν **απορρίπτεται** η μηδενική υπόθεση, τότε[11] συμβαίνει ένα από τα παρακάτω:

- Η X, μέσω του γραμμικού μοντέλου, συνεισφέρει στην πρόβλεψη της $E(Y/X)$. Δηλαδή, η εκτίμηση $\hat{Y} = \bar{y} + \hat{\beta} \cdot (X - \bar{x})$ είναι καλύτερη (στατιστικά πιο σημαντική) από την $\hat{Y} = \bar{y}$.

- Το γραμμικό μοντέλο είναι μόνο μια καλή γραμμική προσέγγιση, μιας μη γραμμικής, στην πραγματικότητα, σχέσης.



Συνοψίζοντας: Είτε απορρίπτεται η μηδενική υπόθεση είτε όχι, το γραμμικό μοντέλο μπορεί να μην είναι κατάλληλο. Κάποιο άλλο μοντέλο (μη γραμμικό), μπορεί να περιγράφει τη σχέση μεταξύ *X* και *Y* καλύτερα.

[10] *με το αντίστοιχο σφάλμα λανθασμένης μη απόρριψης*
[11] *με το αντίστοιχο σφάλμα λανθασμένης απόρριψης*

2. Ο Διορθωμένος Συντελεστής Προσδιορισμού (adjusted r^2)

Ο ***συντελεστής προσδιορισμού*** r^2 εκφράζει το ποσοστό της μεταβλητότητας των τιμών της εξαρτημένης μεταβλητής που εξηγείται, μέσω του μοντέλου, από τις ανεξάρτητες μεταβλητές. Αν στο γραμμικό μοντέλο εισαχθεί μια επιπλέον μεταβλητή, δεν είναι δυνατόν η τιμή του *συντελεστή προσδιορισμού* να μειωθεί. Ο **διορθωμένος *συντελεστής προσδιορισμού*** συμπεριφέρεται διαφορετικά. Αν εισαχθεί μια επιπλέον μεταβλητή, η τιμή του μπορεί είτε να αυξηθεί είτε να ελαττωθεί. Αν η νέα μεταβλητή που εισάγεται δεν συνεισφέρει σημαντικά στην ερμηνεία της μεταβλητότητας των τιμών της εξαρτημένης μεταβλητής, η τιμή του δ*ιορθωμένου* r^2 ελαττώνεται!

Ο δ*ιορθωμένος συντελεστής προσδιορισμού* υπολογίζεται από τον τύπο:

$r^2_{adj} = 1 - (\frac{n-1}{n-k}) \cdot (1 - r^2)$ όπου k, ο αριθμός των παραμέτρων του μοντέλου.

Ο δ*ιορθωμένος συντελεστής προσδιορισμού* είναι πιο κατάλληλος από το *συντελεστή προσδιορισμού* στις εξής περιπτώσεις:
α) όταν ο αριθμός των παραμέτρων του μοντέλου είναι κοντά στο μέγεθος του δείγματος
β) όταν συγκρίνουμε μοντέλα που περιλαμβάνουν διαφορετικό αριθμό ανεξάρτητων μεταβλητών (γιατί;)[12].

[12] *Παρατηρείστε τον τύπο υπολογισμού του και σκεφθείτε τι «τιμωρείται».*

Προβλήματα

1. Στον πίνακα που ακολουθεί φαίνονται τα ποσοστά χαλκού σε 30 δείγματα μεταλλεύματος που ελήφθησαν σε διαφορετικές αποστάσεις κατά μήκος μιας στοάς ενός μεταλλείου. Φαίνονται, επίσης, οι αποστάσεις των σημείων δειγματοληψίας από την είσοδο της στοάς.

Απόσταση (m)	**Ποσοστό Χαλκού (%)**	**Απόσταση (m)**	**Ποσοστό Χαλκού (%)**
0.0	0.67667	17.8	0.6
0.8	0.64333	19.7	0.63667
1.0	0.70333	20.2	0.88667
1.3	0.59333	21.1	0.77667
2.7	0.64333	24.7	1.2
4.5	0.54	25.2	0.66
5.1	0.43	26.3	1.13667
6	0.5833	28.1	1.12333
7.2	0.41	30.5	1.19333
8.7	0.49667	32	1.23333
10.2	0.16667	35.3	1.12667
11.8	0.62	36.2	1.55667
13.1	0.51	37.7	2.10667
15.2	0.57	38.4	1.6
16.3	0.9	41	2.03333

(α) Προσαρμόστε στα δεδομένα το απλό γραμμικό μοντέλο.

α_i) Μέσω του μοντέλου που προσαρμόσατε, τι ποσοστό της μεταβλητότητας της περιεκτικότητας χαλκού εξηγείται από την απόσταση;

α_{ii}) Ελέγξτε αν το μοντέλο που προσαρμόσατε είναι στατιστικά σημαντικό.

α_{iii}) Για την κλίση της ευθείας παλινδρόμησης β, ελέγξτε την υπόθεση $H_0 : \beta = 0$ έναντι της $H_1 : \beta \neq 0$. Ερμηνεύστε το αποτέλεσμα του ελέγχου αυτού.

α_{iv}) Επιβεβαιώνονται από τα δεδομένα οι υποθέσεις-παραδοχές της στατιστικής θεωρίας του απλού γραμμικού μοντέλου;

(β) Προέκυψαν ενδείξεις ότι πρέπει να αναζητηθεί άλλο μοντέλο; Αν ναι, διερευνήστε.

2. Στο πλαίσιο μιας περιβαλλοντικής μελέτης, μετρήθηκαν σε έξι διαφορετικούς χρόνους T, οι συγκεντρώσεις Y μιας χημικής ουσίας σε 18 διαφορετικά διαλύματα (έγιναν τρεις μετρήσεις σε καθέναν από τους έξι διαφορετικούς χρόνους). Στον πίνακα που ακολουθεί φαίνονται τα αποτελέσματα των μετρήσεων αυτών.

Αριθμός Διαλύματος	**Χρόνος** (t_i) ***σε ώρες***	**Συγκέντωση** (y_i) ***σε mg/ml***
1	6	0.029
2	6	0.032
3	6	0.027
4	8	0.079
5	8	0.072
6	8	0.088
7	10	0.181
8	10	0.165
9	10	0.201
10	12	0.425
11	12	0.384
12	12	0.472
13	14	1.13
14	14	1.02
15	14	1.249
16	16	2.812
17	16	2.465
18	16	3.099

(α) Προσαρμόστε στα δεδομένα το απλό γραμμικό μοντέλο $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot T$.

α_i) Μέσω του μοντέλου που προσαρμόσατε, τι ποσοστό της μεταβλητότητας της συγκέντρωσης της χημικής ουσίας *Y* εξηγείται από τη μεταβλητότητα του χρόνου *T*;

α_{ii}) Ελέγξτε αν το μοντέλο που προσαρμόσατε είναι στατιστικά σημαντικό. Για την κλίση της ευθείας παλινδρόμησης β, ελέγξτε την υπόθεση $H_0 : \beta = 0$ έναντι της $H_1 : \beta \neq 0$. Ερμηνεύστε το αποτέλεσμα του ελέγχου αυτού. Τέλος, ελέγξτε (στατιστικά) αν πρέπει να εξετάσετε προσαρμογή κάποιου άλλου μοντέλου.

α_{iii}) Σχολιάστε συνολικά τις επιμέρους απαντήσεις στο ερώτημα *(α_{ii})*. Είναι κάποιες αντιφατικές; Είναι κάποιες ταυτόσημες; (εξηγείστε.)

α_{iv}) Επιβεβαιώνονται από τα πειραματικά δεδομένα οι υποθέσεις-παραδοχές της στατιστικής θεωρίας του απλού γραμμικού μοντέλου;

(β) Προσθέστε έναν ακόμη όρο ($\beta_2 \cdot T^2$) στο μοντέλο.

β_i) Επαναλάβατε τα ερωτήματα α_i – α_{iv} για το νέο μοντέλο: $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 \cdot T + \hat{\beta}_2 \cdot T^2$.

β_{ii}) Ελέγξτε αν ο όρος $\beta_2 \cdot T^2$ είναι στατιστικά σημαντικός. Ερμηνεύστε το αποτέλεσμα του ελέγχου αυτού.

β_{iii}) Βελτιώθηκε το ποσοστό της μεταβλητότητας του *Y* που εξηγείται από την παλινδρόμηση;

(γ) Προσαρμόστε το μοντέλο $\ln(\hat{Y}) = \hat{\alpha} + \hat{\beta} \cdot T$.

γ_i) Επαναλάβετε τα ερωτήματα α_i – α_{iv} για το νέο μοντέλο.

γ_{ii}) *Τι* ποσοστό της μεταβλητότητας του ln(*Y*) εξηγείται από την παλινδρόμηση;

(δ) Ποιο από τα τρία μοντέλα παλινδρόμησης είναι το καταλληλότερο για προσαρμογή στα δεδομένα του πειράματος; Εξηγείστε γιατί. Εξηγείστε επίσης πώς οδηγηθήκαμε να εξετάσουμε αυτά τα μοντέλα.
(ε) Έστω ότι οι 18 αναλύσεις δεν είχαν γίνει σε 18 διαφορετικά διαλύματα αλλά σε 3. Δηλαδή, έστω ότι είχαν γίνει 6 αναλύσεις σε καθένα από 3 διαφορετικά διαλύματα (μια σε καθέναν από τους 6 διαφορετικούς χρόνους). Στην περίπτωση αυτή, θα υπήρχαν προβλήματα στην εφαρμογή της στατιστικής θεωρίας της παλινδρόμησης;

3. Σε κοιλάδες τρίτης τάξης μετρήθηκαν: α) ο αριθμός των ρυακιών πρώτης τάξης *(Y)* β) η πυκνότητα αποστράγγισης[13] *(X_1)* γ) το εμβαδόν κάθε κοιλάδας *(X_2,* δ) η υψομετρική διαφορά του υψηλότερου και του χαμηλότερου σημείου της λεκάνης κάθε κοιλάδας *(X_3)* και ε) το σχήμα[14] κάθε κοιλάδας *(X_4)*. Τα αποτελέσματα των μετρήσεων φαίνονται στον πίνακα που ακολουθεί.

Κοιλάδα	**Αριθμός ρυακιών** Υ	**Πυκνότητα αποστράγγισης** X_1 **(Km/Km^2)**	**Εμβαδόν** X_2 **(Km^2)**	**Υψομετρική διαφορά** X_3 **(m)**	**Σχήμα** X_4
1	25	7.16	0.968	998	0.42
2	7	8.28	0.198	562	0.53
3	12	11.73	0.254	542	0.33
4	59	11.47	1.018	817	0.25
5	5	14.62	0.117	635	0.17
6	12	10.53	0.339	332	0.41
7	6	14.76	0.126	275	0.65
8	23	10.57	0.564	786	0.73
9	6	11.62	0.154	695	0.47
10	7	11.28	0.218	885	0.45
11	5	7.32	0.254	690	0.71
12	10	9.43	0.332	592	0.36
13	9	7.76	0.595	735	0.66
14	6	7.06	0.306	548	0.42
15	5	12.14	0.098	576	0.38
16	9	11.76	0.272	713	0.25
17	11	12.52	0.440	805	0.31
18	7	12.44	0.156	384	0.39
19	17	8.46	0.766	910	0.32
20	5	9.55	0.179	507	0.42

(α) Να εξετάσετε αν μεταξύ του αριθμού των ρυακιών Y και κάθε μιας εκ των μεταβλητών X_1, X_2, X_3, X_4 υπάρχει γραμμική ή άλλη εξάρτηση.
(β) Εκτιμείστε κατάλληλο στοχαστικό μοντέλο το οποίο θα σας επιτρέψει να απαντήσετε στο ερώτημα: Σε επίπεδο σημαντικότητας 5%, μπορούμε με αυτά τα δεδομένα να ισχυριστούμε ότι μια κοιλάδα που έχει 35 ρυάκια και εμβαδόν 0.6 Km^2 ανήκει στον πληθυσμό των κοιλάδων που μελετάμε;
(γ) Εκτιμείστε το μοντέλο: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$.
(δ) Ποιες εκ των μεταβλητών X_1, X_2, X_3, X_4 θα επιλέξετε για να συμπεριλάβετε στο μοντέλο; Όρους αλληλεπίδρασης θα συμπεριλάβετε; Εξηγείστε.

[13] *Η **πυκνότητα αποστράγγισης** της κοιλάδας ορίζεται ως το πηλίκο του συνολικού μήκους όλων των ρυακιών της κοιλάδας προς το εμβαδόν της κοιλάδας.*
[14] *Ως **σχήμα** της κοιλάδας ορίζεται το πηλίκο του πλάτους προς το μήκος της κοιλάδας.*

4. Στον πίνακα που ακολουθεί φαίνονται οι μετρήσεις του βάρους *Y* και του μήκους *X* είκοσι βρεφών τα οποία κατά τη γέννηση είχαν βάρος μικρότερο των 1.500 gr (λιπόβαρα).

Μήκος (x_i) ***σε cm***	**Βάρος** (y_i) ***σε gr***
41	1360
40	1490
38	1490
38	1180
38	1200
32	680
33	620
38	1060
30	1320
34	830
32	880
39	1130
38	1140
39	1350
37	950
39	1220
38	980
42	1480
39	1250
38	1250

(α) Με βάση τα παραπάνω δεδομένα να εκτιμήσετε κατάλληλο στοχαστικό μοντέλο μέσω του οποίου να μπορεί να εκτιμηθεί το μέσο βάρος βρεφών συγκεκριμένου μήκους.
(β) Αξιολογείστε το μοντέλο που εκτιμήσατε και τεκμηριώστε την καταλληλότητά του *(επιβεβαίωση των υποθέσεων-παραδοχών της στατιστικής θεωρίας του μοντέλου, τυπικό σφάλμα της εκτίμησης, ζώνη εμπιστοσύνης, τυπικά σφάλματα των εκτιμήσεων των παραμέτρων και αντίστοιχα διαστήματα εμπιστοσύνης, έλεγχοι υποθέσεων για τις παραμέτρους, συντελεστής προσδιορισμού, Lack-of-Fit test, διερεύνηση πιθανών ακραίων τιμών, σύγκριση με άλλα επίσης κατάλληλα μοντέλα)*.
(γ) Εκτιμείστε το μέσο βάρος των λιπόβαρων κατά τη γέννηση βρεφών μήκους 36 cm. Τι αξία έχει αυτή η εκτίμηση; *(δώστε ένα διάστημα εμπιστοσύνης για το μέσο βάρος του πληθυσμού των λιπόβαρων βρεφών μήκους 36 cm και ερμηνεύστε)*.
(δ) Από τον εξεταζόμενο πληθυσμό των λιπόβαρων κατά τη γέννηση βρεφών, επιλέγετε ένα βρέφος και βρίσκετε ότι έχει μήκος 36 cm. Τι βάρος προβλέπετε να έχει αυτό το βρέφος; Τι αξία έχει αυτή η πρόβλεψη; *(δώστε ένα διάστημα εμπιστοσύνης για το βάρος αυτού του βρέφους (διάστημα πρόβλεψης) και ερμηνεύστε)*.
(ε) Το μοντέλο που εκτιμήσατε μπορεί να δώσει «αξιόπιστη» εκτίμηση του μέσου βάρους λιπόβαρων βρεφών μήκους 46 cm; (εξηγείστε.)
(στ) Για ποιο μήκος (μεταξύ των δεδομένων) προκύπτει το καλύτερο διάστημα εμπιστοσύνης για το μέσο βάρος του πληθυσμού των βρεφών;

5. Σε δείγματα μελιού έγιναν επεμβάσεις (treatments) με malathion και fluvalinate σε συνθήκες incubator και storage. Για να μελετηθεί ο ρυθμός αποδόμησης των ουσιών αυτών, έγιναν μετρήσεις της συγκέντρωσης Y κάθε ουσίας σε διάφορους χρόνους T μετά την αντίστοιχη επέμβαση. Τα αποτελέσματα των μετρήσεων αυτών φαίνονται στους παρακάτω πίνακες[15]:

Malathion

Χρόνος μετά την αγωγή (t_i) *σε εβδ.*	Συγκέντρωση y_i *σε ppd*	
	Incubator	Storage
0	98.7	99.3
1	96.0	97.5
3	83.7	81.9
4	77.7	81.2
5	40.3	52.4
6	29.7	37.7
7	17.0	25.4
8	10.3	18.8
9	5.7	15.9
10	3.3	10.2

Fluvalinate

Χρόνος μετά την επέμβαση (t_i) *σε εβδ.*	Συγκέντρωση y_i *σε ppd*	
	Incubator	Storage
0	193.5	202.5
4	179.9	203.7
8	160.2	142.6
12	79.3	98.1
16	37.4	95.6
20	18.3	76.1
24	6.9	62.3

Οι ερευνητές, μεταξύ άλλων, προσάρμοσαν στις πειραματικές μετρήσεις και για κάθε περίπτωση ξεχωριστά (malathion σε incubator, malathion σε Storage, fluvalinate σε incubator, fluvalinate σε Storage), το απλό γραμμικό μοντέλο παλινδρόμησης.
(α) Να βρείτε (εκτιμήσετε) αυτά τα μοντέλα γραμμικής παλινδρόμησης και να ερμηνεύσετε τις τιμές των παραμέτρων τους.
(γ) Να ελέγξετε αν επιβεβαιώνονται από τα πειραματικά δεδομένα οι υποθέσεις-παραδοχές της στατιστικής θεωρίας του απλού γραμμικού μοντέλου.
(δ) Να δώσετε το τυπικό σφάλμα και ένα 95% διάστημα εμπιστοσύνης για κάθε μια από τις παραμέτρους των μοντέλων. Για κάθε μοντέλο, να ερμηνεύσετε (με όρους του προβλήματος) τις τιμές των άκρων του διαστήματος εμπιστοσύνης κάθε παραμέτρου.
(ε) Για κάθε περίπτωση, να εκτιμήσετε τη μέση συγκέντρωση της ουσίας δύο εβδομάδες μετά την αντίστοιχη επέμβαση. Τι αξία έχουν αυτές οι εκτιμήσεις;
(στ) Να ελέγξετε αν υπάρχει στατιστικώς σημαντική διαφορά μεταξύ των ρυθμών αποδόμησης i) της ουσίας malathion σε συνθήκες incubator και σε συνθήκες Storage ii) της ουσίας fluvalinate σε συνθήκες incubator και σε συνθήκες σε Storage iii) της ουσίας malathion σε συνθήκες incubator και της ουσίας fluvalinate σε συνθήκες incubator.
(ζ) Να αξιολογήσετε τα μοντέλα.

[15] *P. G. Balayannis; L. A. Santas, Journal of Apicultural Research, 31(2): 70-76 (1992)*

6. Στον πίνακα που ακολουθεί φαίνονται οι τιμές του ρυθμού θανατηφόρων γεωργικών ατυχημάτων *Y*, που αντιστοιχούν σε 12 έτη μετά τη λήψη μέτρων ασφάλειας (νομοθετικών, συμβουλευτικής, εκπαίδευσης κ.τλ). Τα έτη έχουν κωδικοποιηθεί.

Έτος x_i	**Ρυθμός θανατηφόρων ατυχημάτων ανά 100 γεωργούς** y_i
1	0.2419
2	0.1732
3	0.1361
4	0.1108
5	0.0996
6	0.0952
7	0.0904
8	0.0792
9	0.0701
10	0.0891
11	0.0799
12	0.1084

(α) Για να μοντελοποιήσετε την τάση των ρυθμών θανατηφόρων ατυχημάτων, επιλέξτε να προσαρμόσετε στα δεδομένα με τη μέθοδο των ελαχίστων τετραγώνων το καταλληλότερο από τα παρακάτω μοντέλα:

i) $Y = \alpha + \beta \cdot X + \varepsilon$

ii) $e^{Y} = \alpha \cdot X^{\beta} \cdot \varepsilon$

iii) $Y = \alpha + \beta \cdot \frac{1}{X} + \varepsilon$

Τεκμηριώστε την επιλογή σας.

(β) Οι τιμές y_i είναι τιμές χρονολογικής σειράς. Τι συνεπάγεται η διαπίστωση αυτή για τη «στατιστική αξία» του μοντέλου που προσαρμόσατε;

7. Στον πίνακα που ακολουθεί φαίνονται οι πειραματικές τιμές της πίεσης *P* και του αντίστοιχου όγκου *V* μιας μάζας αέρα.

Όγκος	54.3	61.8	72.4	88.7	118.6	194.0
Πίεση	61.2	49.5	37.6	28.4	19.2	10.1

Σύμφωνα με τη θερμοδυναμική θεωρία, για τα *P* και *V* ισχύει η μη γραμμική σχέση: $P \cdot V^{\gamma} = C$ όπου γ και *C* σταθερές.

α) Το διάγραμμα διασποράς των πειραματικών δεδομένων επιβεβαιώνει-υποδεικνύει τη σχέση της θερμοδυναμικής θεωρίας;

β) Αφού μετασχηματίσετε κατάλληλα τις πειραματικές τιμές της πίεσης *P* ή/και του όγκου *V,* προσαρμόστε το απλό γραμμικό μοντέλο για την εκτίμηση της πίεσης από τον όγκο. Τι εκτιμήσεις για τις παραμέτρους γ και *C* δίνει το μοντέλο αυτό; Αξιολογείστε τις εκτιμήσεις αυτές. Εκτιμείστε την τιμή της πίεσης *P* για *V = 100*.

γ) Εκτιμείστε τις παραμέτρους γ και C με προσαρμογή απευθείας στις πειραματικές τιμές του μη γραμμικού μοντέλου της θερμοδυναμικής. Αξιολογείστε τις εκτιμήσεις αυτές. Εκτιμείστε την τιμή της πίεσης P για $V = 100$.
δ) Σχολιάστε τις δύο μεθόδους (προσαρμογή γραμμικού μοντέλου στις κατάλληλα μετασχηματισμένες πειραματικές τιμές - προσαρμογή κατάλληλου μη γραμμικού μοντέλου απευθείας στις πειραματικές τιμές).

8. Στον πίνακα που ακολουθεί φαίνονται οι μετρήσεις της συστολικής πίεσης του αίματος *SBP*, του δείκτη $QUET = {}^{\beta\acute{\alpha}\rho o\varsigma}\!/_{\acute{\upsilon}\psi o\varsigma^2} \cdot 100$, της ηλικίας *AGE* και του ιστορικού σχετικά με το κάπνισμα *SMK* (*SMK = 0* για μη καπνιστές, *SMK = 1* για καπνιστές ή πρώην καπνιστές) 32 ανδρών ηλικίας άνω των 40 ετών από μια συγκεκριμένη περιοχή.

Άτομο	SBP	QUET	AGE	SMK
1	135	2.876	45	0
2	122	3.251	41	0
3	130	3.1	49	0
4	148	3.768	52	0
5	146	2.979	54	1
6	129	2.79	47	1
7	162	3.668	60	1
8	160	3.612	48	1
9	144	2.368	44	1
10	180	4.637	64	1
11	166	3.877	59	1
12	138	4.032	51	1
13	152	4.116	64	0
14	138	3.673	56	0
15	140	3.562	54	1
16	134	2.998	50	1
17	145	3.36	49	1
18	142	3.024	46	1
19	135	3.171	57	0
20	142	3.401	56	0
21	150	3.628	56	1
22	144	3.751	58	0
23	137	3.296	53	0
24	132	3.21	50	0
25	149	3.301	54	1
26	132	3.017	48	1
27	120	2.789	43	0
28	126	2.956	43	1
29	161	3.8	63	0
30	170	4.132	63	1
31	152	3.962	62	0
32	164	4.01	65	0

(α) Προσαρμόστε το απλό γραμμικό μοντέλο παλινδρόμησης για την εκτίμηση της μέσης συστολικής πίεσης του αίματος των ανδρών (ηλικίας άνω των 40 ετών της συγκεκριμένης περιοχής) μέσω της τιμής του δείκτη *QUET*.

$\alpha_{i})$ Μέσω του μοντέλου που προσαρμόσατε, τι ποσοστό της μεταβλητότητας της συστολικής πίεσης του αίματος εξηγείται από τη μεταβλητότητα του δείκτη *QUET*;

$\alpha_{ii})$ Ελέγξτε αν το μοντέλο που προσαρμόσατε είναι στατιστικά σημαντικό. Για την κλίση της ευθείας παλινδρόμησης β, ελέγξτε την υπόθεση $H_0 : \beta = 0$ έναντι της $H_1 : \beta \neq 0$. Ερμηνεύστε το αποτέλεσμα του ελέγχου αυτού. Τέλος, ελέγξτε (στατιστικά) αν πρέπει να εξετάσετε προσαρμογή κάποιου άλλου μοντέλου.

$\alpha_{iii})$ Σχολιάστε συνολικά τις επιμέρους απαντήσεις στο ερώτημα (α_{ii}). Είναι κάποιες αντιφατικές; Είναι κάποιες ταυτόσημες; (εξηγείστε.)

$\alpha_{iv})$ Επιβεβαιώνονται από τα πειραματικά δεδομένα οι υποθέσεις-παραδοχές της στατιστικής θεωρίας του απλού γραμμικού μοντέλου;

$\alpha_{v})$ Δώστε μια εκτίμηση και ένα 95% διάστημα εμπιστοσύνης για τη μέση συστολική πίεση του αίματος των ανδρών που έχουν δείκτη *QUET* ίσο με 3.5. Στη συνέχεια, προσαρμόστε κατάλληλο γραμμικό μοντέλο παλινδρόμησης για να μπορέσετε απαντήσετε στα παρακάτω ερωτήματα:

$\alpha_{vi})$ Δώστε μια εκτίμηση και ένα 95% διάστημα εμπιστοσύνης για τη μέση συστολική πίεση του αίματος των ανδρών που έχουν δείκτη *QUET* ίσο με 3.5 και είναι καπνιστές και αντίστοιχα των ανδρών ίδιου δείκτη *QUET* (= 3.5) που δεν είναι καπνιστές.

$\alpha_{vii})$ Εκτιμείστε πόσο θα μεταβληθεί η συστολική πίεση του αίματος των ανδρών που είναι καπνιστές αν ο δείκτης *QUET* αυξηθεί κατά μια μονάδα και αντίστοιχα πόσο των μη καπνιστών.

$\alpha_{vii})$ Ελέγξτε αν υπάρχει διαφορετική γραμμική σχέση μεταξύ της συστολικής πίεσης του αίματος και του δείκτη *QUET* για τους καπνιστές σε σύγκριση με τους μη καπνιστές άνδρες

(β) Προσαρμόστε το απλό γραμμικό μοντέλο παλινδρόμησης για την εκτίμηση της μέσης συστολικής πίεσης του αίματος των ανδρών (ηλικίας άνω των 40 ετών της συγκεκριμένης περιοχής) μέσω της ηλικίας *AGE* και επαναλάβατε τα ερωτήματα α_i *έως* α_{iv}.

(γ) Προσαρμόστε το απλό γραμμικό μοντέλο παλινδρόμησης για την εκτίμηση του δείκτη *QUET* των ανδρών (ηλικίας άνω των 40 ετών της συγκεκριμένης περιοχής) μέσω της ηλικίας *AGE* και επαναλάβατε τα ερωτήματα α_i *έως* α_{iv}.

(δ) Προσαρμόστε το απλό γραμμικό μοντέλο παλινδρόμησης για την εκτίμηση της μέσης συστολικής πίεσης του αίματος των ανδρών (ηλικίας άνω των 40 ετών της συγκεκριμένης περιοχής) μέσω του ιστορικού για το κάπνισμα *SMK*.

$\delta_{i})$ Να συγκρίνετε την εκτίμηση $\hat{\alpha}$ της παραμέτρου α του μοντέλου, με τη μέση συστολική πίεση αίματος των μη καπνιστών. Επίσης, να συγκρίνετε την τιμή του αθροίσματος $\hat{\alpha} + \hat{\beta}$ με τη μέση συστολική πίεση αίματος των καπνιστών (ερμηνεύστε).

$\delta_{ii})$ Για την κλίση της ευθείας παλινδρόμησης β, ελέγξτε την υπόθεση $H_0 : \beta = 0$ έναντι της $H_1 : \beta \neq 0$. Ερμηνεύστε το αποτέλεσμα του ελέγχου αυτού.

$\delta_{iii})$ Είναι ο έλεγχος στο ερώτημα δ_{ii}, ισοδύναμος με το *t-test* για την ισότητα δύο πληθυσμιακών μέσων (με ίσες αλλά άγνωστες διασπορές);

(ε) Προσαρμόστε κατάλληλο γραμμικό μοντέλο παλινδρόμησης για να μπορέσετε απαντήσετε στα παρακάτω ερωτήματα:

$\varepsilon_{i})$ Πόση εκτιμάτε ότι είναι η μέση συστολική πίεση των ανδρών ηλικίας 50 ετών, που έχουν δείκτη *QUET* ίσο με 3.5 και είναι καπνιστές;

ε_{ii}) Πόση εκτιμάτε ότι είναι η μέση συστολική πίεση των ανδρών ηλικίας 50 ετών, που έχουν δείκτη *QUET* ίσο με 3.5 και είναι μη καπνιστές;

ε_{iii}) Εκτιμείστε πόσο θα μεταβληθεί η μέση συστολική πίεση των μη καπνιστών ανδρών ηλικίας 50 ετών αν ο δείκτης *QUET* αυξηθεί από 3.5 σε 4.5

ε_{iv}) Εκτιμείστε πόσο θα μεταβληθεί η μέση συστολική πίεση των μη καπνιστών ανδρών ηλικίας 42 ετών αν ο δείκτης *QUET* αυξηθεί από 3.5 σε 4.5

ε_{v}) Εκτιμείστε πόσο θα μεταβληθεί η μέση συστολική πίεση των μη καπνιστών ανδρών ηλικίας 51 ετών αν ο δείκτης *QUET* αυξηθεί από 3 σε 4.

(στ) Προσθέστε στο μοντέλο έναν όρο ο οποίος να εκφράζει την αλληλεπίδραση μεταξύ ηλικίας και δείκτη *QUET*. Έχει η ηλικία διαφορετική επίδραση στη συστολική πίεση του αίματος εξαρτώμενη από την τιμή του δείκτη *QUET;*

(ζ) Ποιες από τις μεταβλητές ηλικία, δείκτης *QUET* και αλληλεπίδραση μεταξύ ηλικίας και δείκτη *QUET* θα επιλέξετε για να συμπεριλάβετε στο μοντέλο; Εξηγείστε.