Βιοστατιστική

Γραμμική παλινδρόμηση-Linear Regression

Χαράλαμπος Γναρδέλλης

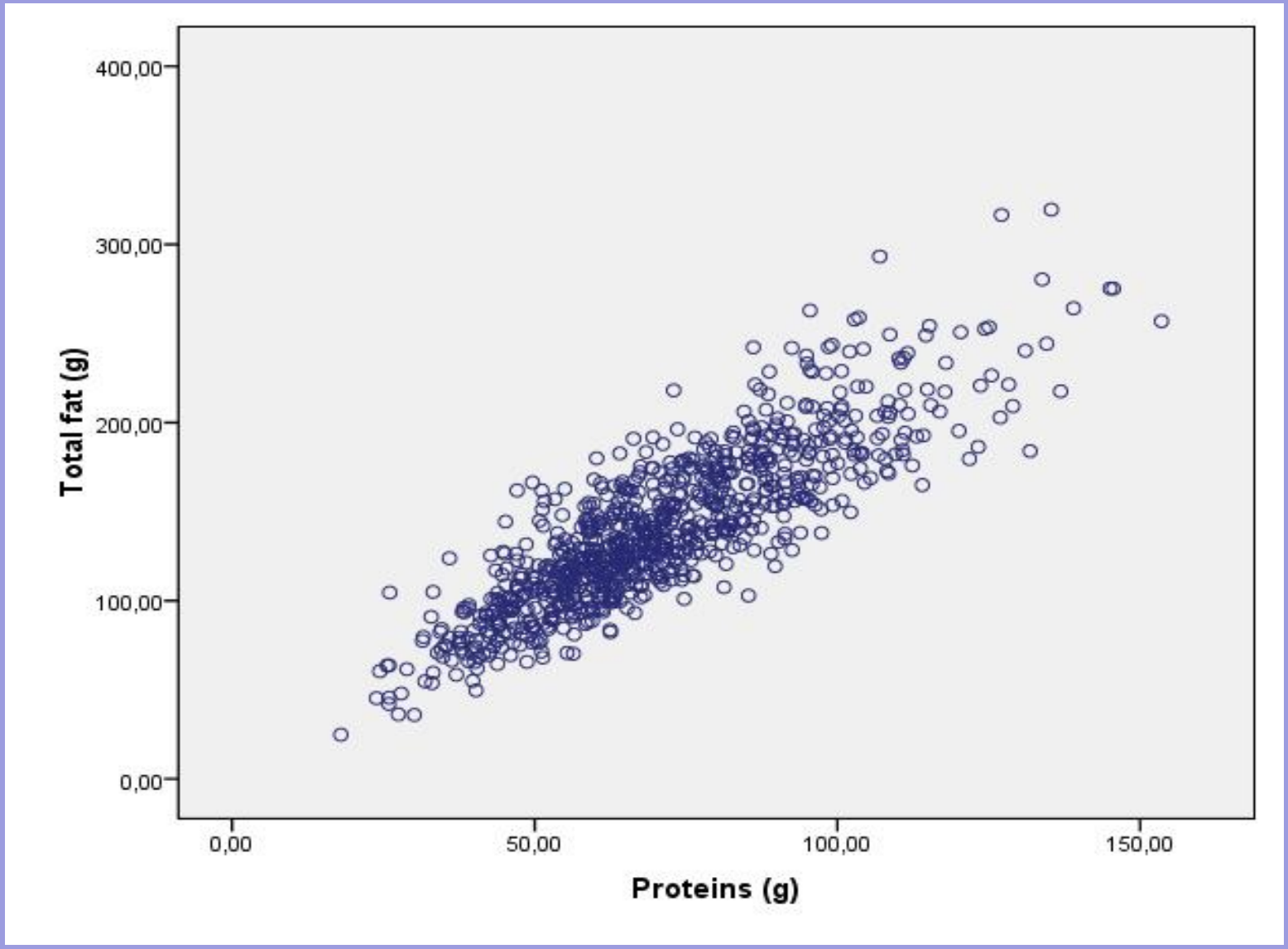
Τμήμα Ζωικής Παραγωγής Αλιείας και Υδατοκαλλιεργειών
Πανεπιστημίου Πατρών

Γραμμική παλινδρόμηση

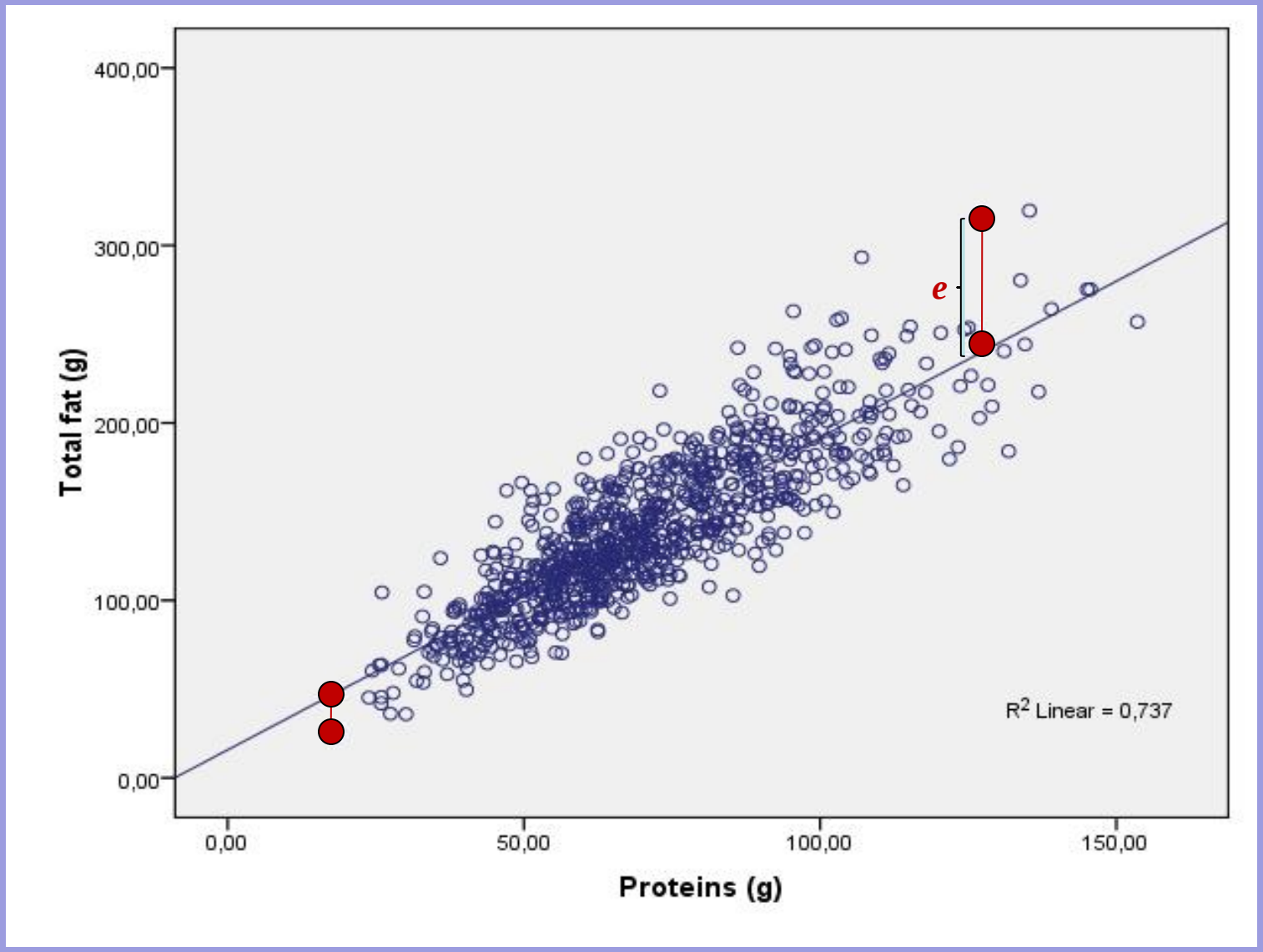
- **Οι συντελεστές συσχέτισης προσδιορίζουν το βαθμό (ή αλλιώς την ένταση) της σχέσης που υπάρχει μεταξύ δύο ποσοτικών μεταβλητών υπό την προϋπόθεση ότι η σχέση αυτή είναι γραμμική.**

- **Αν η γραμμική σχέση των δύο μεταβλητών οριστεί με όρους εξάρτησης της μίας από την άλλη, δηλαδή, αν η μεταβολή των τιμών της μιας μεταβλητής θεωρηθεί ότι προκύπτει με γραμμικό τρόπο από τη μεταβολή των τιμών της άλλης, τότε η ανάλυση της σχέσης των δύο μεταβλητών πραγματοποιείται με τη βοήθεια ενός υποδείγματος *απλής γραμμικής παλινδρόμησης*.**

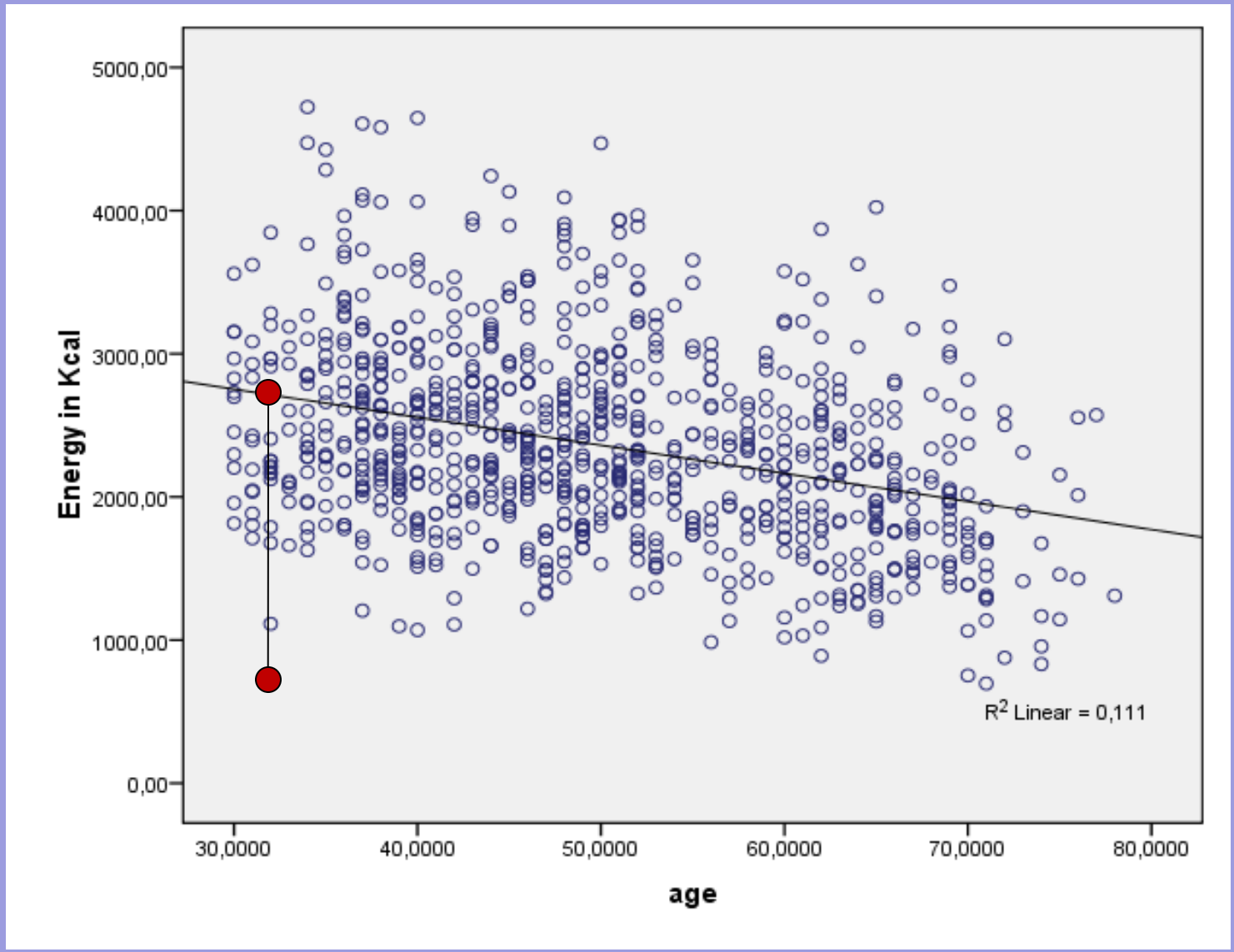
Διάγραμμα διασποράς δύο γραμμικά συσχετιζόμενων ποσοτικών μεταβλητών



Ευθεία παλινδρόμησης δύο γραμμικά συσχετιζόμενων μεταβλητών



Ευθεία παλινδρόμησης δύο γραμμικά συσχετιζόμενων μεταβλητών



- **Η απλή γραμμική παλινδρόμηση ορίζει τη σχέση δύο ποσοτικών μεταβλητών *X* και *Y*, μέσω ενός υποδείγματος της μορφής**

$$y = b_0 + b_1 x + e,$$

όπου *x* και *y* είναι οι τιμές των δύο μεταβλητών για μια οποιαδήποτε παρατήρηση των δειγματικών δεδομένων.

- **Το πρώτο μέρος του υποδείγματος συνοψίζει τη γραμμική σχέση των δύο μεταβλητών, ενώ η ποσότητα *e*, η οποία ονομάζεται *υπόλοιπο* ή *σφάλμα*, αφορά την απόκλιση από τη γραμμικότητα, δηλαδή την απόκλιση των τιμών της *Y* από την ευθεία**

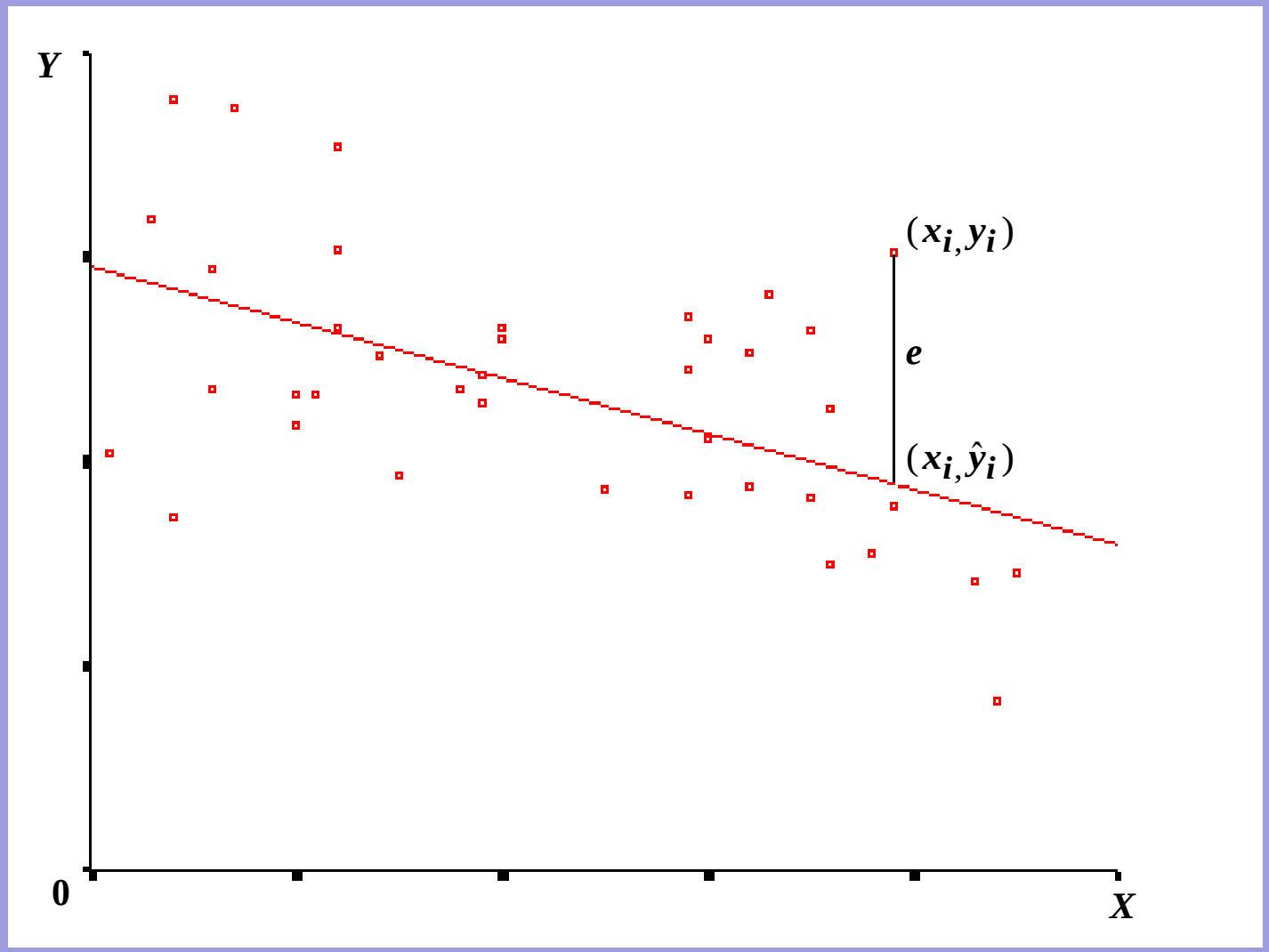
$$b_0 + b_1 x$$

- **Η ευθεία $b_0 + b_1 x$ ονομάζεται ευθεία της παλινδρόμησης**

- **Μέσω του υποδείγματος της απλής γραμμικής παλινδρόμησης οι τιμές της *Y* εκτιμώνται (ή προβλέπονται) από τις τιμές της *X*. Για το λόγο αυτό η *Y* ονομάζεται *εξαρτημένη μεταβλητή* του υποδείγματος *(dependent variable ή response variable)* ενώ η μεταβλητή *X* ονομάζεται *ανεξάρτητη* *(independent ή predictor variable)*.**

- **Το υπόδειγμα της απλής γραμμικής παλινδρόμησης είναι η απλούστερη περίπτωση ενός γενικότερου υποδείγματος που χρησιμοποιείται στη στατιστική συμπερασματολογία, σύμφωνα με το οποίο *η μεταβολή των τιμών μιας ποσοτικής μεταβλητής ερμηνεύεται γραμμικά από τη μεταβολή των τιμών ενός συνόλου* k *άλλων ποσοτικών μεταβλητών* .**

Προσδιορισμός της ευθείας της παλινδρόμησης για τις μεταβλητές X και Y



Αν δηλαδή η σχέση των μεταβλητών *X* και *Y* είναι γραμμική τότε αυτή μπορεί να συνοψιστεί με τη βοήθεια μιας ευθείας γραμμής η οποία προσεγγίζει με βέλτιστο τρόπο τα σημεία του διαγράμματος. Κάθε σημείο επομένως του διαγράμματος με συντεταγμένες (x_i, y_i), μπορεί να θεωρηθεί ότι βρίσκεται επάνω σε μια ευθεία γραμμή, η οποία ονομάζεται ευθεία της παλινδρόμησης, και η οποία προσεγγίζει με τον καλύτερο δυνατό τρόπο το σύνολο των σημείων. Αυτή η παραδοχή απλοποιεί τη μορφή του νέφους των σημείων, υποκαθιστώντας το με τα αντίστοιχα σημεία $(x_i, \hat{y}_i)$ που προκύπτουν από τις κατακόρυφες προβολές των δειγματικών τιμών της *Y* στην ευθεία της παλινδρόμησης.

- **Οι κατακόρυφες προβολές των δειγματικών τιμών της Y, ορίζουν τα υπόλοιπα** $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$.

- **Αν όλα τα υπόλοιπα είναι ίσα με 0, θα έχουμε την πλήρη προσαρμογή της ευθείας επί των σημείων του διαγράμματος. Η πλήρης προσαρμογή της ευθείας είναι απίθανο να προκύψει (εκτός και αν οι δύο μεταβλητές είναι απολύτως γραμμικά εξαρτημένες), μπορούν όμως να ελαχιστοποιηθούν τα υπόλοιπα. Η ελαχιστοποίηση των υπολοίπων ισοδυναμεί με την ελαχιστοποίηση της ποσότητας**

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

η οποία ονομάζεται *άθροισμα τετραγώνων των υπολοίπων* (*residual sum of squares*) ή *άθροισμα τετραγώνων των σφαλμάτων* (*error sum of squares*).

- **Η ευθεία δηλαδή της παλινδρόμησης κατασκευάζεται με την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων, γι' αυτό το λόγο ονομάζεται και *ευθεία των ελαχίστων τετραγώνων*.**

- **Από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων προκύπτουν οι τιμές των συντελεστών του υποδείγματος , b_0 και b_1.**

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

Ο συντελεστής b_1 *(κλίση)* της ευθείας της παλινδρόμησης, ορίζεται ως η εκτιμώμενη μεταβολή της εξαρτημένης μεταβλητής (της *Y* δηλαδή) για κάθε μία μονάδα αύξησης της ανεξάρτητης (της *X*). Ο ορισμός αυτός προκύπτει από την παρακάτω σχέση

$$y_{x+1} - y_x = \left[b_0 + b_1(x+1)\right] - \left(b_0 + b_1 x\right) = b_1$$

Αν το πρόσημο του συντελεστή b_1 είναι θετικό, τότε για κάθε μονάδα αύξησης της *X* η *Y* αυξάνεται κατά b_1, ενώ αν το πρόσημο του συντελεστή είναι αρνητικό για κάθε μονάδα αύξησης της *X* η *Y* ελαττώνεται κατά b_1.

Επιπλέον, ο σταθερός όρος b_0 του υποδείγματος ορίζεται ως η εκτιμώμενη τιμή της εξαρτημένης μεταβλητής Y για την τιμή 0 της μεταβλητής X,

$$b_0 = b_0 + b_1 \cdot 0$$

Παράδειγμα

Σε ένα τυπικό πρόβλημα γραμμικής παλινδρόμησης, το ενδιαφέρον εστιάζεται στον προσδιορισμό της ευθείας της παλινδρόμησης, δηλαδή της ευθείας που περιγράφει την πραγματική σχέση που υπάρχει μεταξύ των μεταβλητών X και Y. Ο προσδιορισμός αυτής της ευθείας ισοδυναμεί με την εκτίμηση των συντελεστών της παλινδρόμησης b_0 και b_1.

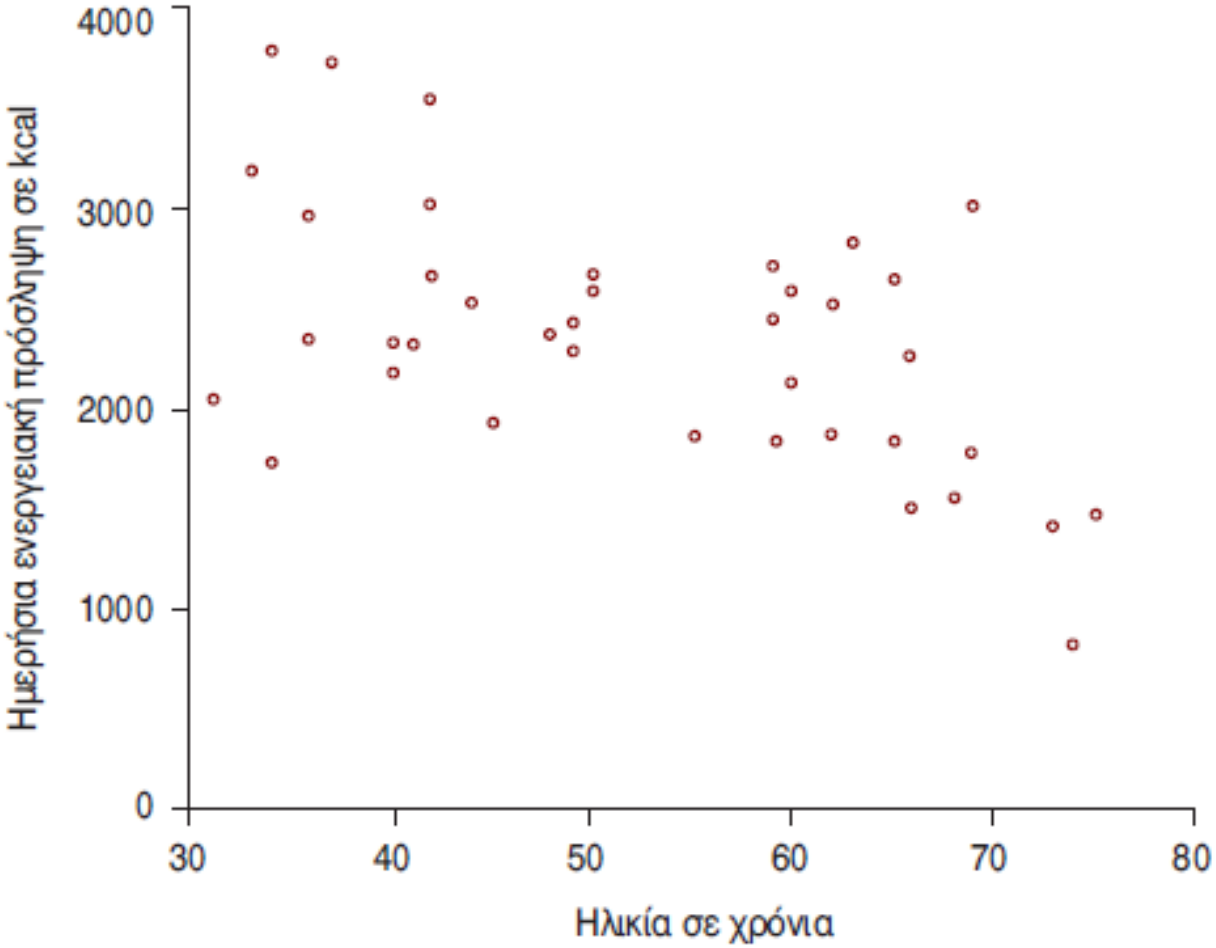
Πριν όμως προσδιοριστεί η δειγματική ευθεία της παλινδρόμησης, είναι απαραίτητο να επιβεβαιωθεί η γραμμική σχέση που υπάρχει μεταξύ των δύο μεταβλητών στα δειγματικά δεδομένα. Η διαδικασία αυτή μπορεί να γίνει με τη βοήθεια ενός διαγράμματος διασποράς.

Έστω οι τιμές της ημερήσιας ενεργειακής πρόσληψης (σε Kcal) 40 ενήλικων ατόμων μαζί με την ηλικία τους. Μεταξύ της ηλικίας και της ημερήσιας ενεργειακής πρόσληψης υπάρχει γραμμική σχέση, σύμφωνα με την οποία αυξανομένης της ηλικίας η ενεργειακή πρόσληψη ελαττώνεται. Η ύπαρξη της γραμμικής σχέσης μεταξύ των δύο μεταβλητών επιβεβαιώνεται από τη μορφή του διαγράμματος διασποράς που απεικονίζεται στη συνέχεια.

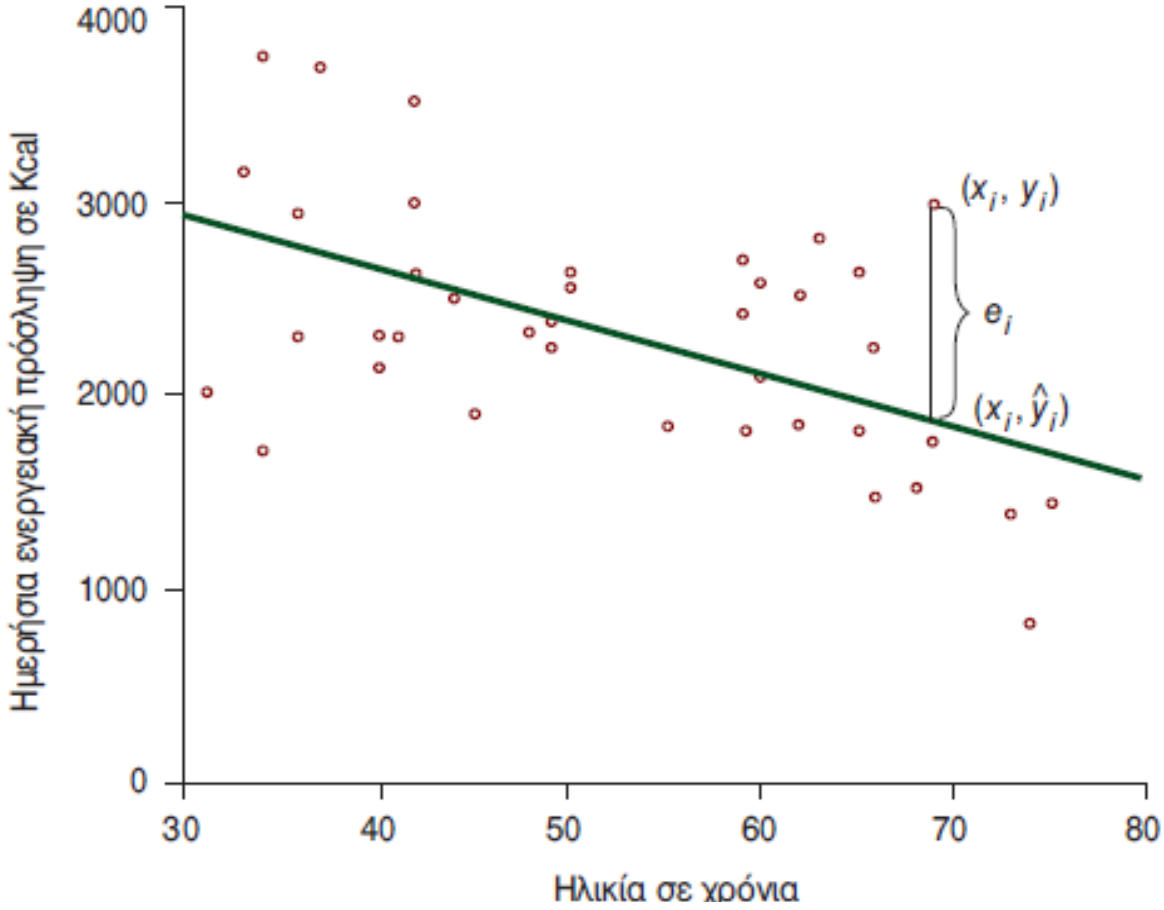
Τιμές ηλικίας και ενεργειακής πρόσληψης

Αριθμός ατόμου	Ηλικία	Ενέργεια σε Kcal
1	63	2822
2	49	2419
3	44	2518
4	69	3015
5	55	1857
6	62	1875
7	41	2322
8	42	3536
9	36	2943
10	50	2658
11	60	2596
12	50	2593
13	42	2655
14	37	3728
15	34	3766
16	40	2327
17	65	2637
18	62	2524
19	45	1928
20	42	3027
21	36	2352
22	59	2454
23	40	2175
24	48	2358
25	66	1495
26	59	2705
27	73	1411
28	34	1725
29	49	2291
30	31	2047
31	69	1776
32	66	2265
33	65	1821
34	66	1501
35	75	1458
36	60	2111
37	68	1546
38	33	3189
39	59	1837
40	74	832

Διάγραμμα διασποράς της ηλικίας και της ημερήσιας ενεργειακής πρόσληψης 40 ενηλίκων



Προσδιορισμός της ευθείας των ελαχίστων τετραγώνων για τη σχέση της ενεργειακής πρόσληψης με την ηλικία



Η διαδικασία προσδιορισμού της ευθείας των ελαχίστων τετραγώνων, η οποία συμβολικά ορίζεται από την εξίσωση

$$y = b_0 + b_1 x$$

απαιτεί τον προσδιορισμό των ποσοτήτων b_0, b_1

Υπολογίζουμε τους συντελεστές του υποδείγματος από τις σχέσεις

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

Μορφή του υποδείγματος μετά τον υπολογισμό των

$$b_0 \quad \text{και} \quad b_1$$

$$b_0 = 3756{,}6 \qquad b_1 = -27$$

$$\hat{y} = 3756{,}6 - 27x$$