Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Ανάλυση Παλινδρόμησης

Ανάλυση Παλινδρόμησης

Στα περισσότερα προβλήματα στατιστικής εξετάζουμε την συσχέτιση μεταξύ 2 ή περισσοτέρων μεταβλητών και τον τρόπο που η αλλαγή σε μια επηρεάζει την άλλη.

Για παράδειγμα η ηλικία και το ύψος ενός παιδιού εμφανίζουν θετική συσχέτιση μεταξύ τους καθώς όσο αυξάνεται η ηλικία ενός παιδιού αντίστοιχα αυξάνεται και το ύψος τους.

Ή η καθημερινή άσκηση έχει αρνητική εξάρτιση σε σχέση με την παχυσαρκία των ανθρώπων. Όσο περισσότερο ασκείται κάποιος τόσο λιγότερο ποσοστό λίπους εμφανίζεται στο σώμα του.

Ανάλυση Παλινδρόμησης

Στην απλή παλινδρόμηση έχουμε μόνο μια μεταβλητή X και μια μεταβλητή Y.

Η μεταβλητή Y υπολογίζεται συνάρτηση της μεταβλητής X.

Για παράδειγμα:

Y=10X+20

X:ανεξάρτητη μεταβλητή

Y:εξαρτημένη μεταβλητή

Ανάλυση Παλινδρόμησης

Η παλινδρόμηση στην οποία υπάρχει μόνο μια ανεξάρτητη μεταβλητή καλείται απλή παλινδρόμηση ενώ αν υπάρχουν περισσότερες από μια μεταβλητές τότε ονομάζεται πολλαπλή παλινδρόμηση.

Ανάλυση Παλινδρόμησης

- Για παράδειγμα η σχέση μεταξύ καθημερινής άσκησης και παχυσαρκίας είναι <u>απλή</u> παλινδρόμηση
  - Χ:Καθημερινή Άσκηση
  - Υ:Παχυσαρκία

Ανάλυση Παλινδρόμησης

- Αλλά η σχέση μεταξύ καθημερινής άσκησης και των διατροφικών συνήθειων σε σχέση με την παχυσαρκία είναι πολλαπλή παλινδρόμηση
  - X1:Καθημερινή Άσκηση
  - X2:Διατροφικές Συνήθειες
  - Y:Παχυσαρκία

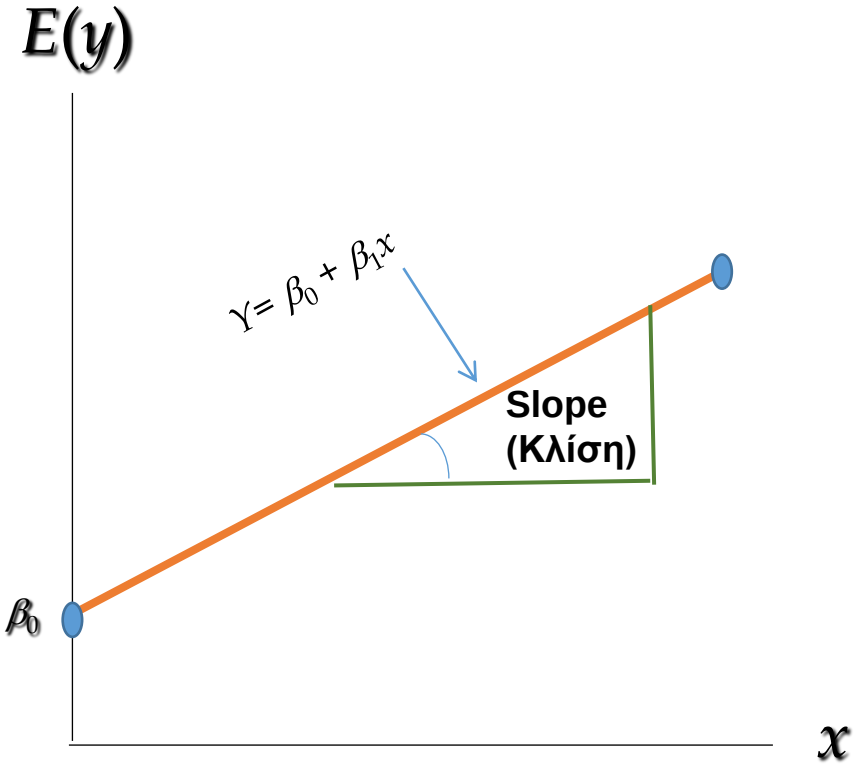
Ανάλυση Παλινδρόμησης

Η εξίσωση που περιγράφει την συσχέτιση του Y με το X είναι η παρακάτω.

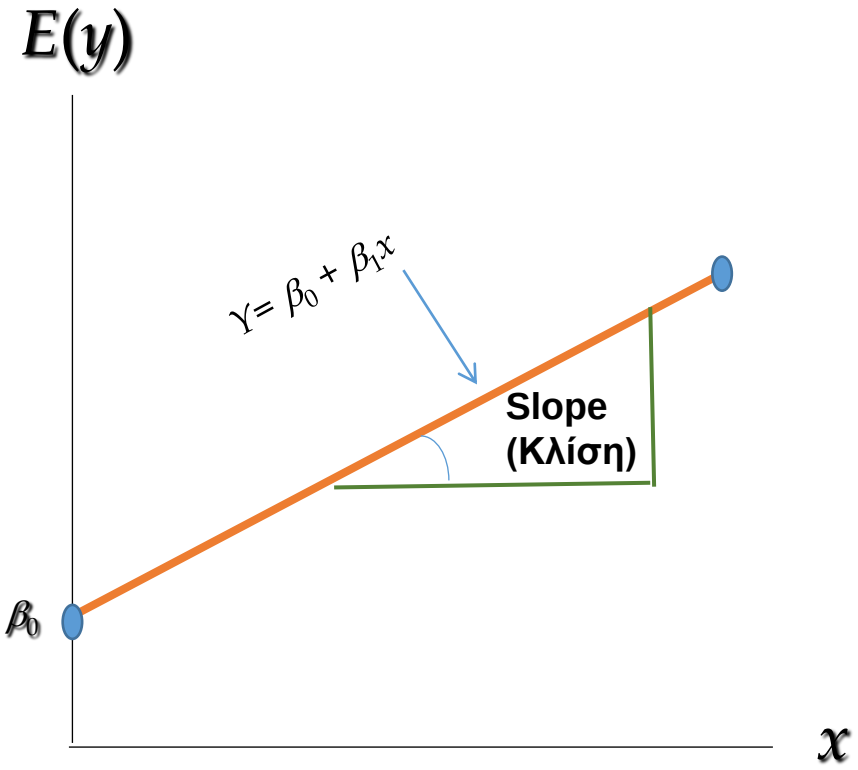
$$Y = \beta_0 + \beta_1 x$$

- Η γραφική αναπαράσταση αυτής της εξίσωσης είναι μία ευθεία γραμμή.
- Το β_0 είναι το σημείο από το οποίο ξεκινάει η γραμμική εξίσωση.
- Το β_1 είναι η κλίση της γραμμικής εξίσωσης.

Ανάλυση Παλινδρόμησης



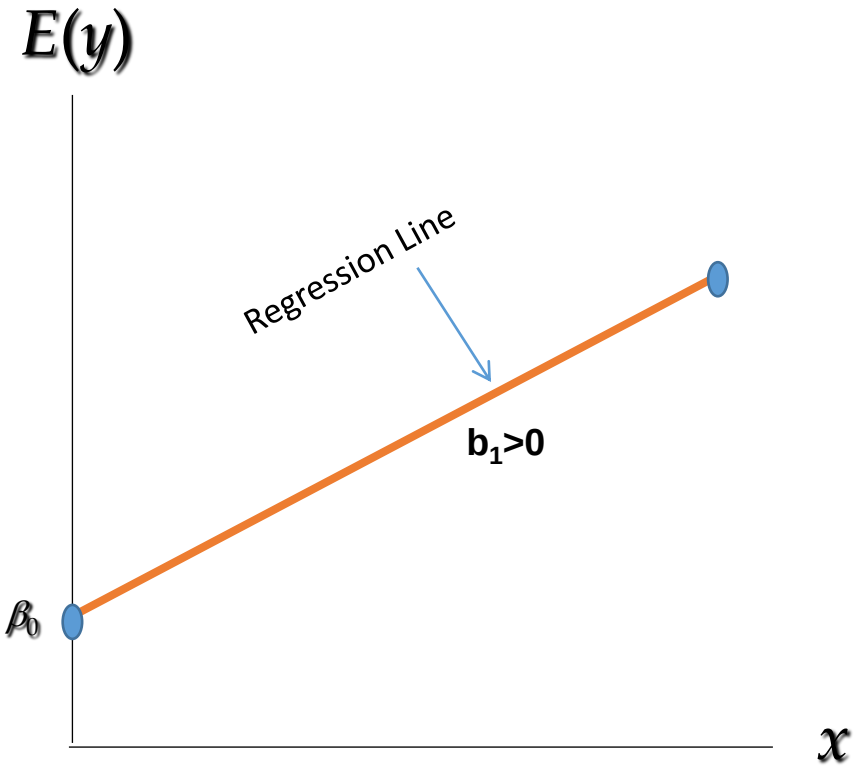
Ανάλυση Παλινδρόμησης



Linear Regression Function

$Y = \beta_0 + \beta_1 x$

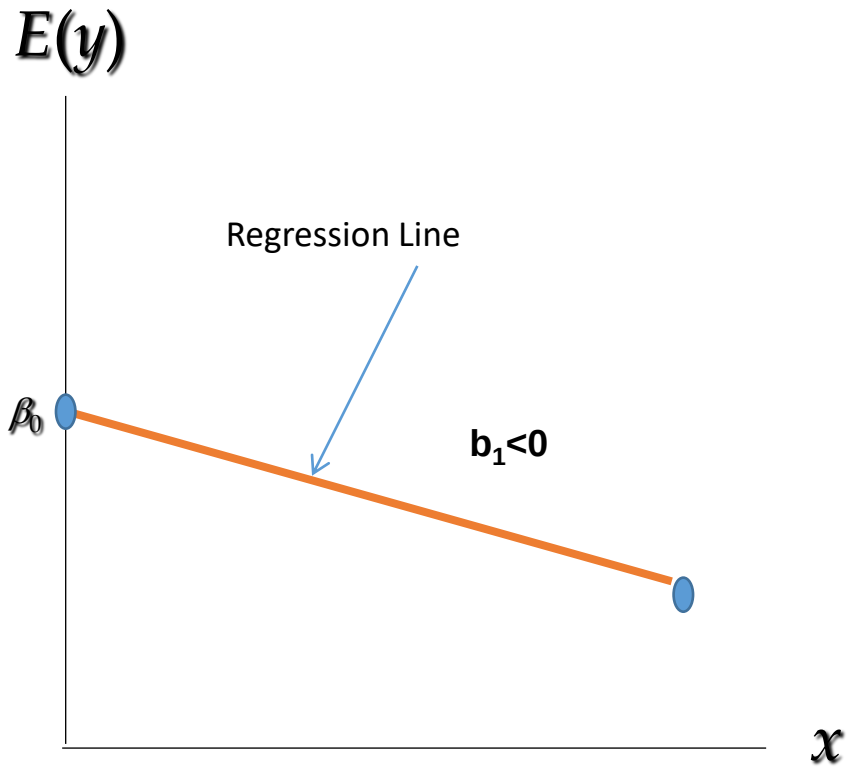
Ανάλυση Παλινδρόμησης



Positive Regression Line

Η κλίση είναι θετική

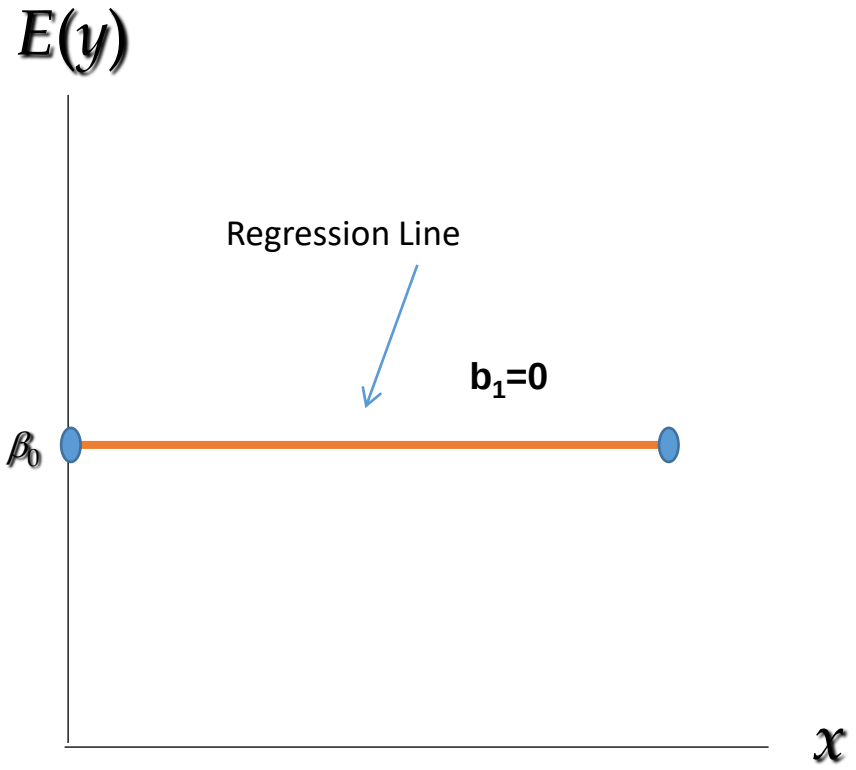
Ανάλυση Παλινδρόμησης



Negative Regression Line

Η κλίση είναι αρνητική

Ανάλυση Παλινδρόμησης



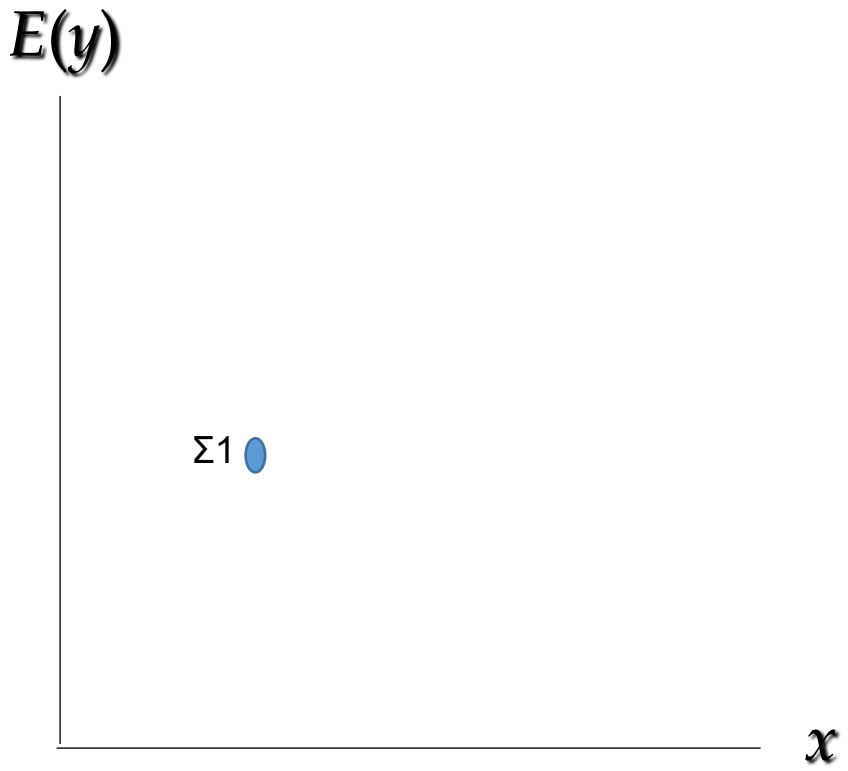
<u>Χωρίς Συσχέτιση</u>

Η κλίση είναι μηδέν

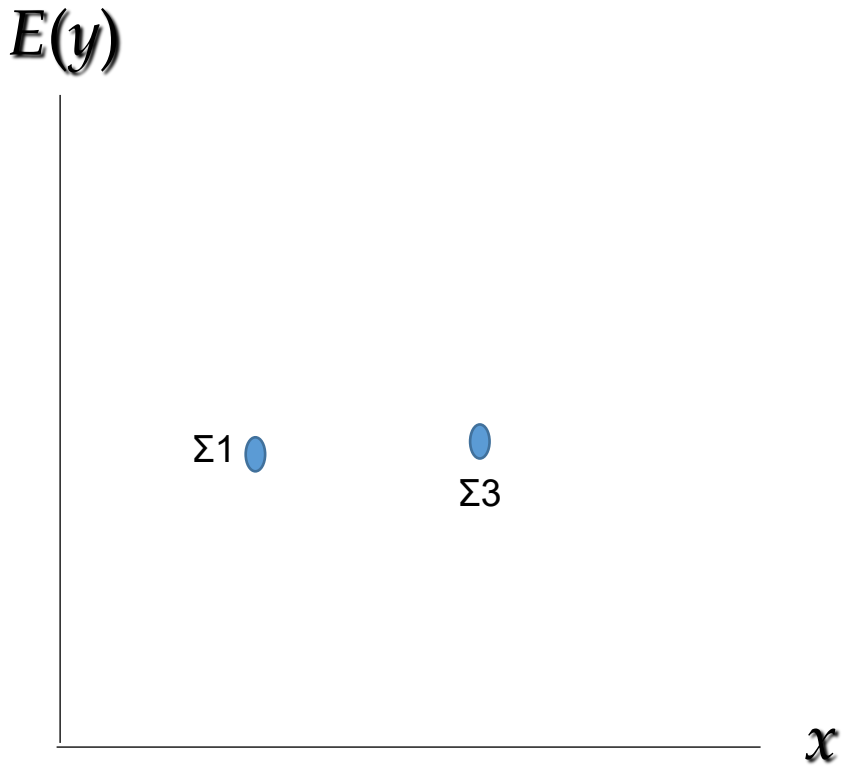
Ανάλυση Παλινδρόμησης

Τι γίνεται στην περίπτωση που έχουμε 3 σημεία;

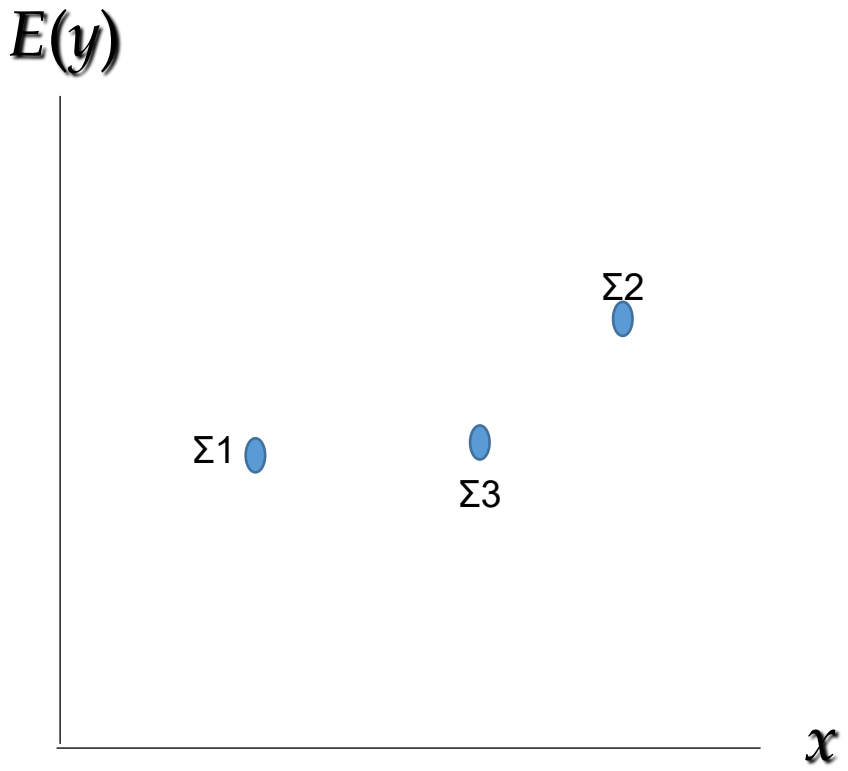
Ανάλυση Παλινδρόμησης



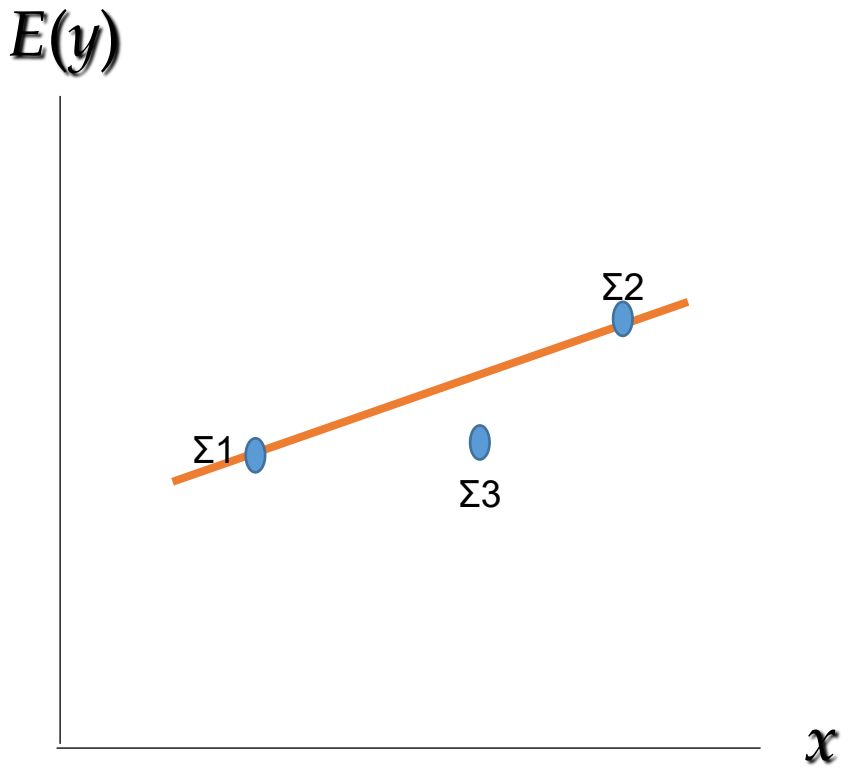
Ανάλυση Παλινδρόμησης



Ανάλυση Παλινδρόμησης

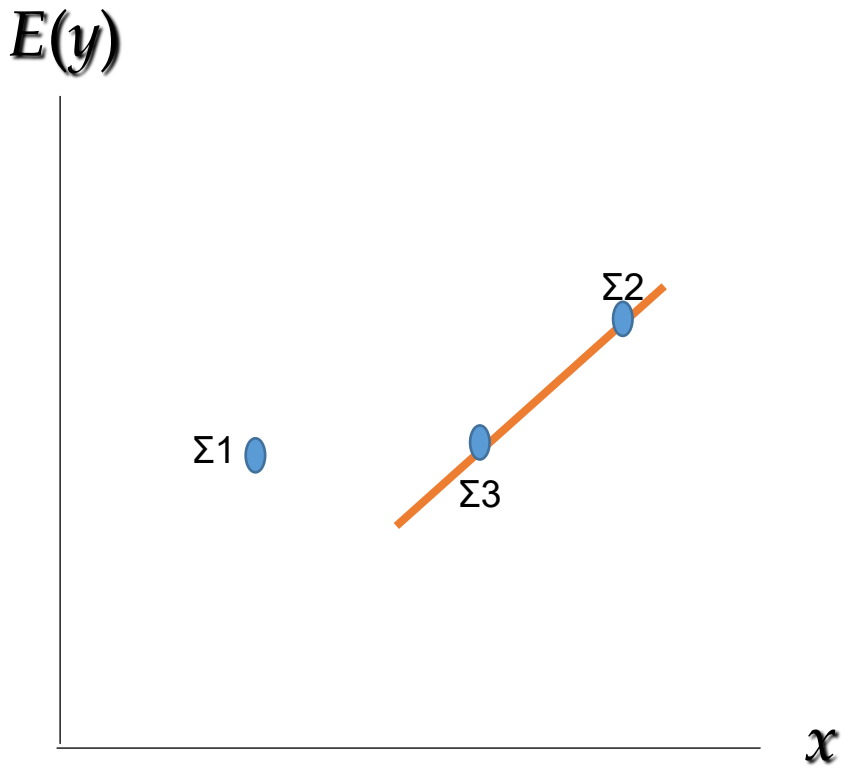


Ανάλυση Παλινδρόμησης



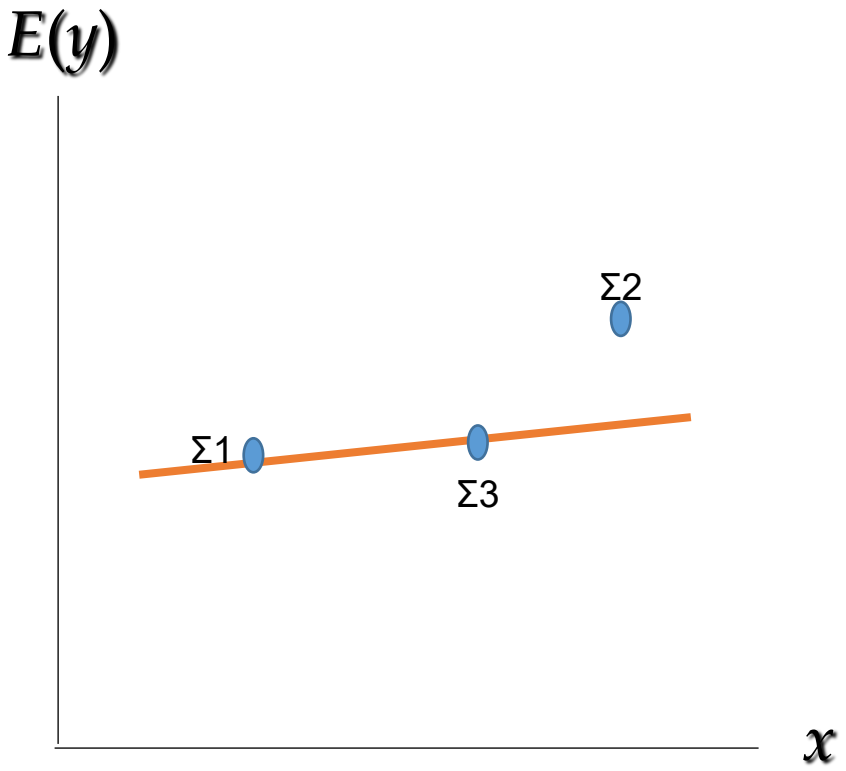
Υπάρχει ευθεία που να διέρχεται από όλα τα σημεία;

Ανάλυση Παλινδρόμησης



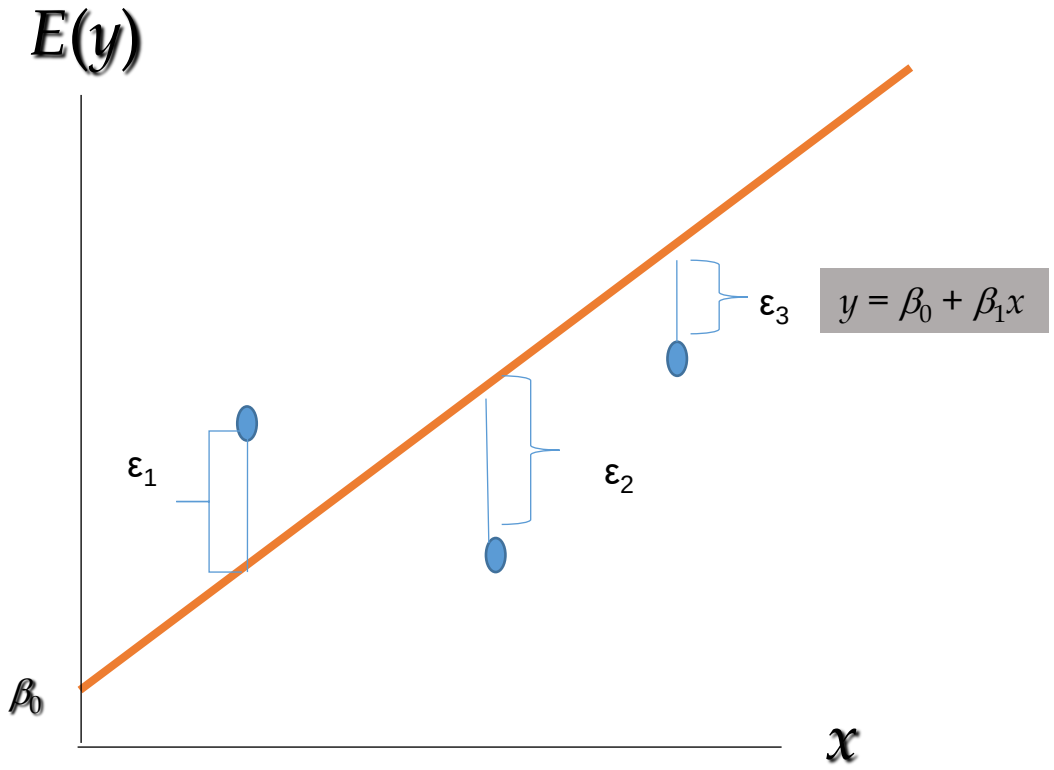
Υπάρχει ευθεία που να διέρχεται από όλα τα σημεία;

Ανάλυση Παλινδρόμησης



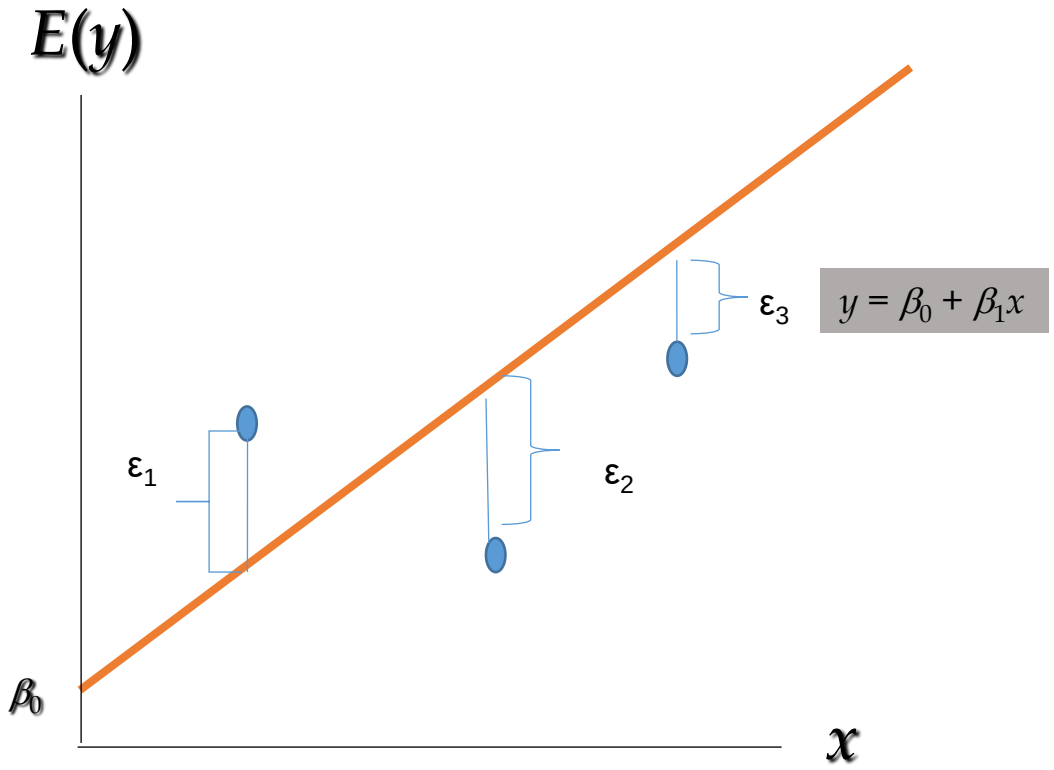
Υπάρχει ευθεία που να διέρχεται από όλα τα σημεία;

Ανάλυση Παλινδρόμησης



Όπως παρατηρούμε η ιδανική ευθεία δεν περνάει από όλα τα σημεία.

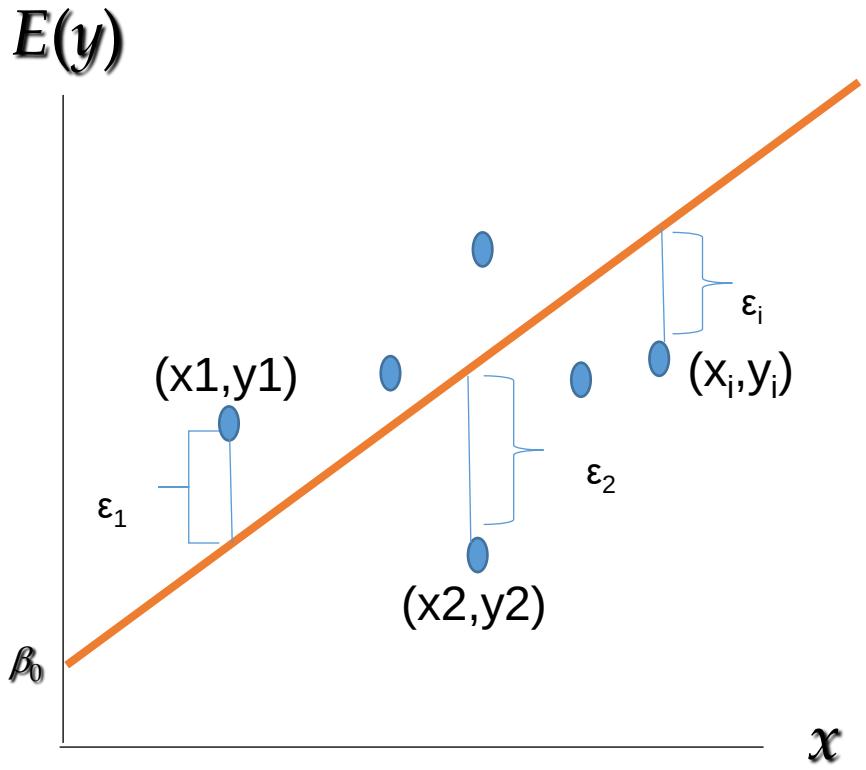
Ανάλυση Παλινδρόμησης



Όπως παρατηρούμε η ιδανική ευθεία δεν περνάει από όλα τα σημεία.

Για όλα τα σημεία υπάρχει μια απόσταση (ε) μεταξύ των σημείων και της ευθείας

Ανάλυση Παλινδρόμησης



Οπότε η εξίσωση της γραμμικής παλινδρόμησης γίνεται

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Ανάλυση Παλινδρόμησης

Η εξίσωση που περιγράφει την συσχέτιση του Y με το X μαζί με τον συντελεστή **ε** ονομάζεται γραμμικό μοντέλο.

$$E(y) = \beta_0 + \beta_1 x + \varepsilon$$

όπου:

- β_0 και β_1 είναι οι παράμετροι του μοντέλου,
- ε είναι ο συντελεστής σφάλματος

Ανάλυση Παλινδρόμησης

Για να υπολογίσουμε την απόσταση κάθε σημείου i και της ευθείας η εξίσωση γίνεται

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Ανάλυση Παλινδρόμησης

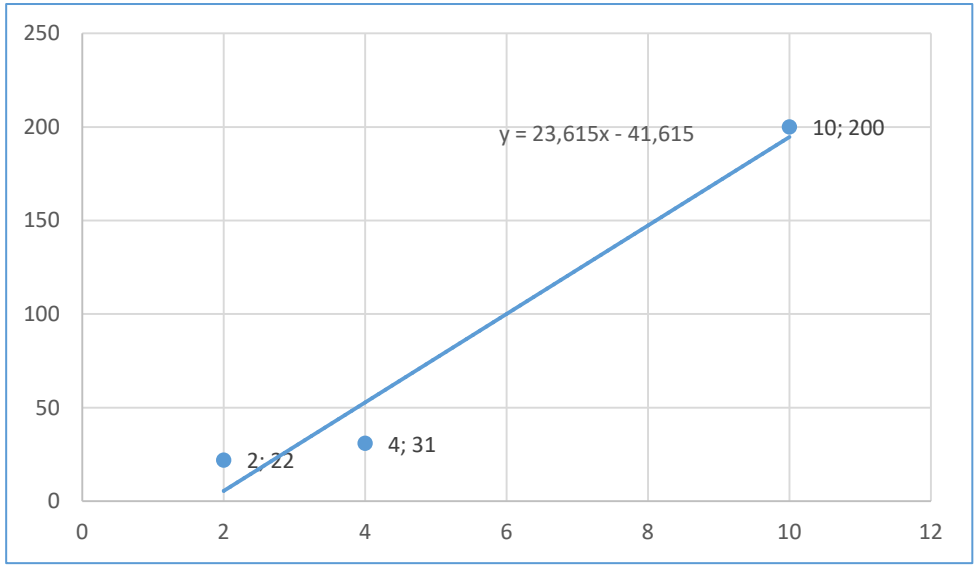
Για να υπολογίσουμε την απόσταση κάθε σημείου i και της ευθείας η εξίσωση γίνεται

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Στην συνέχεια προκειμένου να υπολογίσουμε την ελάχιστη δυνατή απόσταση μεταξύ των σημείων θα χρησιμοποιούμε την μέθοδο των ελαχίστων τετραγώνων μέσω της οποίας ψάχνουμε την ευθεία που ελαχιστοποιεί το άθροισμά των τετραγώνων των ε_i.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Μέθοδος ελαχίστων τετραγώνων



Θα πρέπει να βρω τα β_0, β_1 για τα οποία το άθροισμα των $\varepsilon_1+\varepsilon_2....+\varepsilon_v$ γίνεται ελάχιστο

Μέθοδος ελαχίστων τετραγώνων

Έχει βρεθεί ότι οι τιμές των β_0, β_1 υπολογίζονται από τις σχέσεις

$$\hat{\beta}_1 = \frac{\nu \sum_{i=1}^{v} \boldsymbol{x_i y_i} - (\sum_{i=1}^{v} x_i)(\sum_{i=1}^{v} y_i)}{\nu \sum_{i=1}^{v} \boldsymbol{x_i}^2 - (\sum_{i=1}^{v} \boldsymbol{x_i})^2}, \hat{\beta}_0 = \frac{1}{\nu}\sum_{i=1}^{v} \boldsymbol{y_i} - \hat{\beta}_1 \cdot \frac{1}{\nu}\sum_{i=1}^{v} \boldsymbol{x_i}$$

Οπότε η ευθεία $y=\hat{\beta}_0+\hat{\beta}_1 x$ ονομάζεται ευθεία ελαχίστων τετραγώνων ή ευθεία παλινδρόμησης της Υ σε σχέση με το Χ.

Μέθοδος ελαχίστων τετραγώνων – Σφάλμα Εκτίμησης

Η μέση απόκλιση μεταξύ της πραγματικής και της εκτιμώμενης τιμής μιας μεταβλητής ονομάζεται τυπικό σφάλμα (standard error of the estimate),

Έστω για παράδειγμα η συνάρτηση

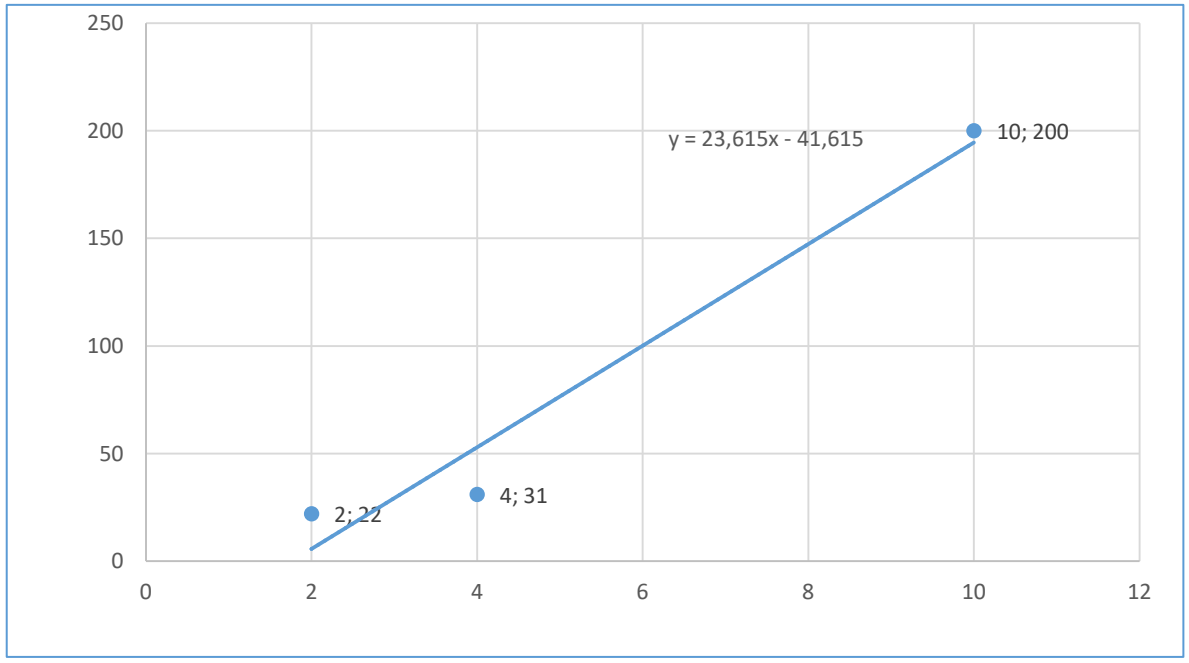
Y=23,615*X-41,615 (X=10, Y=236,1).

Το Y είναι η εκτιμώμενη τιμή και την συμβολίζου με $\hat{y}$.

Η πραγματική τιμή του Y όμως είναι 200.

Πραγματική-Εκτιμώμενη=$y - \hat{y}$=200-236,1=-36,1

Μέθοδος ελαχίστων τετραγώνων – Σφάλμα Εκτίμησης



Τυπικό σφάλμα της εκτίμησης

Το τυπικό σφάλμα της εκτίμησης (standard error of the estimate), συμβολίζεται με s και δίνεται από τον τύπο

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

Τυπικό σφάλμα της εκτίμησης

Το τυπικό σφάλμα της εκτίμησης (standard error of the estimate), συμβολίζεται με s και δίνεται από τον τύπο

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

Εάν το τυπικό σφάλμα της εκτίμησης είναι μικρό τότε η ευθεία παλινδρόμησης μας δίνει μια καλή περιγραφή της σχέσης μεταξύ των Χ και Υ. Αν το τυπικό σφάλμα της εκτίμησης είναι μεγάλο τότε η ευθεία δεν περιγράφει καλά την σχέση μεταξύ των πραγματικών και των εκτιμούμενων τιμών.

Πόσο «καλή» είναι η ευθεία ελαχίστων τετραγώνων;

Προκειμένου να ελέγξουμε κατά πόσο είναι καλή ή όχι η ευθεία της εξίσωσης που δημιουργήσαμε θα χρησιμοποιήσουμε τον συντελεστή προσδιορισμού (coefficient of determination) R^2 ,που παίρνει τιμές στο κλειστό διάστημα [0, 1].

Coefficient of determination - R^2

SST = SSR + SSE

- SST = άθροισμα τετραγώνων (total sum of squares)

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

Coefficient of determination - R^2

SST = SSR + SSE

- SST = άθροισμα τετραγώνων (total sum of squares)

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

- SSR = άθροισμα τετραγώνων παλινδρόμησης (regression sum of squares)

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

Coefficient of determination - R^2

SST = SSR + SSE

- SST = άθροισμα τετραγώνων (total sum of squares)

$$SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

- SSR = άθροισμα τετραγώνων παλινδρόμησης (regression sum of squares)

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

- SSE = άθροισμα τετραγώνων των σφαλμάτων (error sum of squares)

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

Coefficient of determination - R^2

Ο τύπος του R^2 είναι :

$$R^2 = SSR/SST$$

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.

ΤΕΛΟΣ