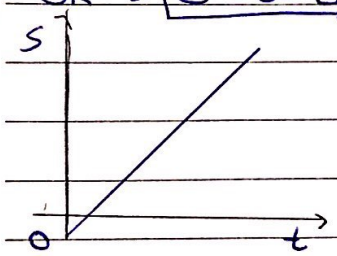
Δευτέρα 12/10/2020

~~Απλή γραμμική παλινδρόμηση~~

Παράδειγμα:

ΕΟΚ → $s = u \cdot t$ → Απόλυτο/αιτιοκρατικό/ντετερμινιστικό μοντέλο



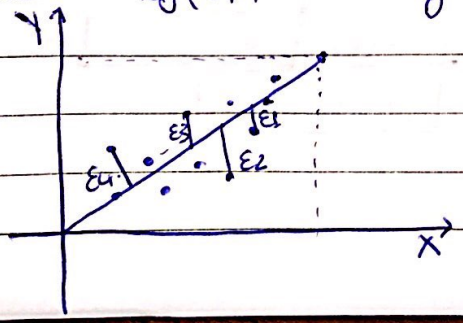
Όσα δεδομένα και να πάρω βρίσκονται απαραίτητα - απόλυτα πάνω στην γραμμή

Έστω ότι κάποιος σε μια μελέτη ασχολείται με το βάρος (=X) και το ύψος (=Y) των παιδιών. Υπάρχει κάποια σχέση μεταξύ τους. Υπάρχει κάποιο f (μοντέλο) ώστε $Y = f(X)$?

Διαλέγω n μέλη του πληθυσμού και παράγω πειραματικά δεδομένα στις X, Y

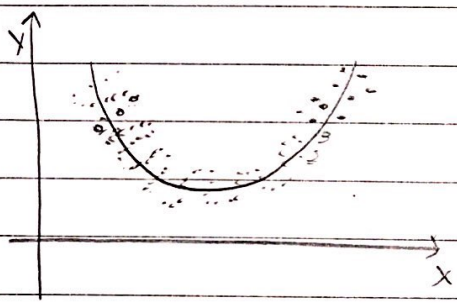
X	X_1	X_2	....	X_n
Y	y_1	y_2		y_n

Ερώτημα: Μπορούμε να κατασκευάσουμε ένα μαθηματικό μοντέλο στο οποίο να συνδέει τη X με την y μεταβλητή Δηλαδή έναν τύπο που να περιλαμβάνει τα σημεία (x_i, x_i) του ακόλουθου διαγράμματος για $i = 1, \dots n$.



Στο 1ο παράδειγμα είχαμε ένα αιτιοκρατικό πείραμα ενώ στο 2ο παράδειγμα έχουμε ένα τυχαίο πείραμα με τυχαίες μεταβλητές. Επίσης, δεν αναφερόμαστε στο 2ο παράδειγμα σε ένα μοντέλο απόλυτο αλλά έχουμε ένα μοντέλο τ.μ που διαφέρουν από άνθρωπο σε άνθρωπο λόγω της διαφορετικότητας των ανθρώπων. Για το σκοπό αυτό θεωρούμε ένα γραμμικό κατά παραμέτρους μοντέλο:

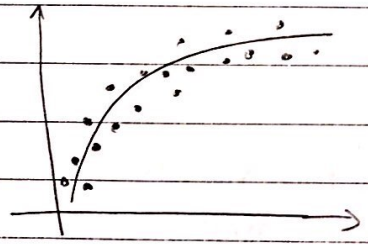
$Y = \beta_0 + \beta_1 X + \varepsilon$, με ε: οι αποκλίσεις των σημείων (x_i, y_i) $i = 1, \dots n$ από το γραμμικό μοντέλο



Ούτε εδώ υιοθετούμε το μοντέλο της γραμμής.

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

Η γραμμή είναι "φτωχή" περιγραφή για τη συγκεκριμένη κατανομή των σημείων



$Y = a \log x + \varepsilon$, όμοια με παραπάνω

Σκοπός του μαθήματος είναι η κατασκευή ενός γραμμικού μοντέλου που περιγράφει τη σχέση μεταξύ δύο ή περισσότερων τ.μ.

ΕΝΟΤΗΤΑ 1:

[redacted] Απλή Γραμμική Παλινδρόμηση (α.γ.π)

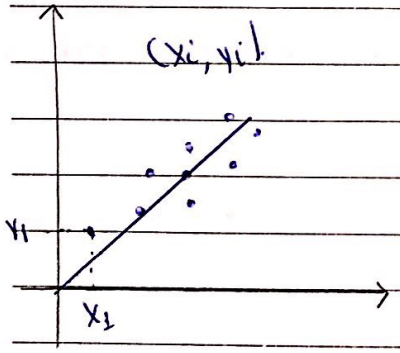
Έστω δυο τ.μ X και Y. Ειδικότερα η X είναι η μεταβλητή την οποία χειρίζεται ο στατιστικός και Y είναι η τ.μ που επιθυμούμε να δούμε πως σχετίζεται πως συμπεριφέρεται σε μεταβολές της X
Θεωρώ τυχ. δείγματα από τις X και Y μεγέθους n.

X	x_1	x_2 . .	x_n
Y	y_1	y_2	y_n

Ερώτημα:
Υπάρχει κάποια σχέση (γραμμική) που συνδέει τα X και Y ?
Πόσο αξιόπιστη είναι αυτή?

Βήμα 1ο: [Διάγραμμα Διασποράς]

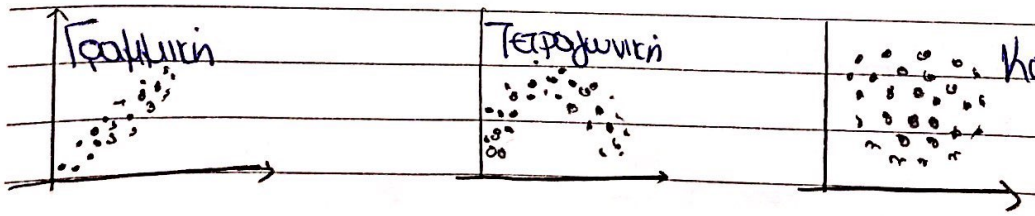
Απεικονίζουμε τα δεδομένα σε ένα σύστημα αξόνων



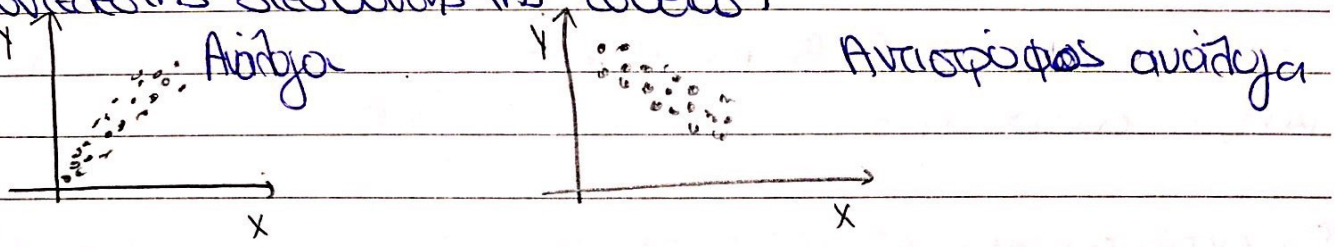
Βλέπουμε ότι δεν βρίσκονται όλα τα σημεία των πειραματικών δεδομένων πάνω στην ευθεία. Δηλαδή υπάρχει μια ευθεία $Y = b_0 + b_1 X$ που τείνει κατά προσέγγιση να συμπεριλάβει τα (x_i, y_i). Τα σημεία (x_i, y_i) $i = 1, 2, .. n$ παλινδρομούν γύρω από το γραμμικό μοντέλο

Δηλαδή υπάρχει η σχέση $Y = b_0 + b_1 X + \varepsilon$, με ε: η απόκλιση των σημείων από την πορισθείσα ευθεία.

Το διάγραμμα διασποράς μας δείχνει:
α) Τη μορφή της σχέσης μεταξύ των X και Y.



β) Την κατεύθυνση της σχέσης (αν αυτή είναι γραμμική)
δηλαδή αν το Υ μεγαλώνει όταν μεγαλώνει το Χ ή αντίστοιχα
αν μικραίνει. (αύξηση ή αντίστροφος ανάλογα, παίζει ρόλο ο
συντελεστής διεύθυνσης της ευθείας)



γ) Βαθμίδα της σχέσης, δηλαδή πόσο κοντά είναι τα σημεία του διαγράμματος διασποράς στο μοντέλο που το περιγράφει.

Η αδυναμία του διαγράμματος διασποράς είναι ότι αναφέρεται σε δυο μόνο μεταβλητές.

Αν το διάγραμμα διασποράς υποδεικνύει γραμμικότητα των Υ και Χ
επιλέγουμε το μοντέλο της Απλής Γραμμικής Παλινδρόμησης.

παράμετροι

$$Y = \beta_0 + \beta_1 X + \varepsilon \iff Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, \dots n$$

σφάλματα του μοντέλου

εξαρτημένη ή απόκριση

ανεξάρτητη ή επεξηγηματική

Το ε περιλαμβάνει αποσιωπήσεις μεταβλητές μέχρι κυρίως την ύπαρξη άλλων παραγόντων που επηρεάζουν τις Χ και Υ και τους οποίους δεν λαμβάνουμε υπόψη στο μοντέλο

2ο Βήμα: Κατασκευή του μοντέλου α.ε.π.
Σκοπός είναι να προσδιοριστούν τα β_0 και β_1.
Εκτιμητές ελαχίστων τετραγώνων (Pearson ~ 1900

Ιδέα εκτίμησης β_0 β_1 } Τα καλύτερα β_0, β_1 είναι εκείνα για τα οποία ελαχιστοποιούνται οι αποστάσεις (ε_i) των σημείων του διαγρ. διασποράς από το μοντέλο της α.γ.π.

Οπότε θα βρω τα β_0, β_1 που ελαχιστοποιούν συνολικά τα ε_i ή κάποια συνάρτηση των ε_i ή την $S^2 = \sum_{i=1}^{n} \varepsilon_i^2$

Οι εκτιμητές των ελάχιστων τετραγώνων των β_0, β_1 προκύπτουν από ελαχιστοποίηση ως προς β_0, β_1 της.

$$S^2 = S^2(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 \overset{\text{α.γ.π}}{=} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Για την εύρεση των β_0, β_1 που ελαχιστοποιούν την S^2.

$$\left.\begin{array}{l} \dfrac{\partial S^2}{\partial \beta_0} = 0 \Rightarrow n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ \dfrac{\partial S^2}{\partial \beta_1} = 0 \Rightarrow \beta_1 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \end{array}\right\} \text{Κανονικές Εξισώσεις}$$

Καταλήγουμε μετά από πράξεις.

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Η ευθεία $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ είναι η πιο "καλή" ευθεία που περνάει με τέτοιο τρόπο ώστε όλα τα σημεία να πολινδρομούν όσο το δυνατόν πιο κοντά στο γραμμικό μοντέλο $Y = \beta_0 + \beta_1 X + \varepsilon$
Αυτό το μοντέλο λέγεται εκτιμώμενο μοντέλο α.γ.π
Χρησιμοποιείται για την πρόβλεψη της Y για συγκεκριμένες τιμές της X.

Scanned with CamScanner

Για το βήμα 3:

Υπόλοιπα: Η απόκλιση του μοντέλου από την πραγματικότητα που "έλπιζει" το μοντέλο να περιγράψει

Συμβολίζουμε $e_i = Y_i - \hat{Y}_i$, $i=1,\dots n$.

Ιδιότητα $\sum_{i=1}^{n} e_i = 0$

Απόδ:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n}(Y_i - \hat{\beta_0} - \hat{\beta_1} x_i) =$$

$$= \sum_{i=1}^{n}(Y_i - (\bar{Y} - \hat{\beta_1}\bar{x}) - \hat{\beta_1} x_i) = \sum_{i=1}^{n}(Y_i - \bar{Y}) + \hat{\beta_1}\sum_{i=1}^{n}(\bar{x_i} - x_i) = 0$$

αφού $\sum_{i=1}^{n}(Y_i - \bar{Y}) = \sum Y_i - \sum \bar{Y} = n\bar{Y} - n\bar{Y} = 0$

Ανάλυση της ολικής μεταβλητότητας στο μοντέλο α.γ.π.

Ολική μεταβλητότητα ονομάζουμε και σαν την δειγματική διακύμανση των $Y_1, \dots Y_n$ δηλαδή την $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \propto \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Θέλουμε να μελετήσουμε πως η ολική μεταβλητότητα διέπεται από το εκτιμώμενο μοντέλο.

→ SStot.

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(Y_i+\hat{Y}_i-\hat{Y}_i-\bar{Y})^2 =$$

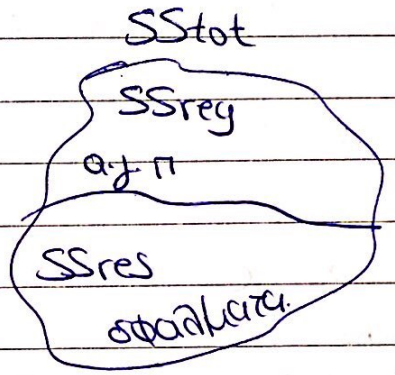
$$= \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \rightarrow$$ μεταβλητότητα που οφείλεται στα υπόλοιπα (SSres).

→ μεταβλητότητα που οφείλεται στο μοντέλο της. α.β.π (SSreg)

Άρα $$SStot = \sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

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

$$SSres = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$



- Αν SSreg >> SSres το μοντέλο εξηγεί, εξηγεί μεγάλο μέρος της ολικής μεταβλητότητας SStot. Άρα είναι ικανοποιητικό μοντέλο α.β.π.

- Αν SSreg << SSres. Τότε το μοντέλο της α.β.π φαίνεται ότι δεν είναι ικανοποιητικό