ΕΙΣΑΓΩΓΗ

Ανάλυση παλινδρόμησης (regression analysis): εξετάζουμε τη σχέση μεταξύ 2 ή περισσοτέρων μεταβλητών

Σκοπός:Πρόβλεψη των τιμών της μιας, μέσω των τιμών της άλλης (ή των άλλων).

Δύο είδη μεταβλητών: **ανεξάρτητες ή ελεγχόμενες ή επεξηγηματικές (independent, predictor, casual, input, explanatory variables)** και **εξαρτημένες ή απόκρισης (dependent, response variables)**.

$$y = a_0 + a_1 X_1 + a_2 X_2 + \varepsilon \rightarrow \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2$$

$$\hat{y}$$

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).

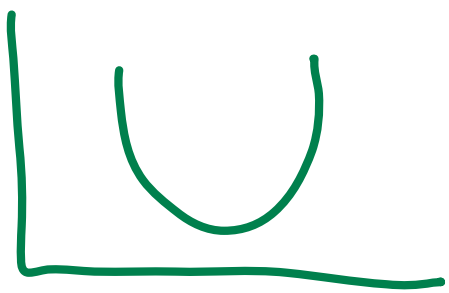
Εξαρτημένη μεταβλητή Y **είναι εκείνη στην οποία αντανακλάται το αποτέλεσμα των μεταβολών στις ανεξάρτητες** μεταβλητές (π.χ. η ζήτηση ενός προϊόντος, ο χρόνος αναμονής των πελατών ενός υποκαταστήματος τραπέζης).

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).

Αν οι μεταβλητές συνδέονται με μια σχέση της μορφής $Y = f(X)$ ωστε για κάθε τιμή της X να προβλέπουμε με απολυτη ακριβεια την Y (δηλ. χωρις σφάλμα) οι δύο μεταβλητές συνδέονται με **συναρτησιακή προσδιοριστική (deterministic) σχέση**.

Π.χ. το ρεύμα που καταναλώνει μια οικογένεια και το ποσό που πληρώνει ή το ποσό που καταθέτει κάποιος στην τράπεζα και ο τόκος που λαμβάνει.

Τοτε στο σκεδασμογράφημα όλα τα σημεία είναι ακριβώς πάνω στην καμπύλη $Y = f(X)$ **όσες φορές και αν επαναλάβουμε το πείραμα θέτοντας τις ιδιες τιμες στο X θα βρισκουμε πάντα το ιδιο Y.**



$y = ax^2 + \alpha = f(x)$

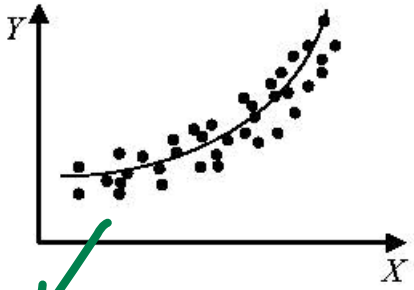
Οι μη προσδιοριστικές σχέσεις ονομάζονται **στοχαστικές – πιθανωτικές (stochastic, probabilistic) σχέσεις**.

Εδω αν επαναλάβουμε το πείραμα πολλές φορές **ΔΕΝ ΠΡΟΚΥΠΤΕΙ ΜΟΝΟ ΜΙΑ ΤΙΜΗ ΤΟΥ** Y για την ιδια τιμη του X. Π.χ. η τιμή ενός προϊόντος και η ζήτησή του.

Τωρα το **σκεδασμογράφημα (scatter plot) είναι ένα νέφος σημείων** που μπορεί να καθορίζει μια ιδεατή γραμμή (ή γενικότερα καμπύλη) η οποία δίνει μια πρώτη εικόνα της σχέσης που συνδέει τις δύο μεταβλητές.

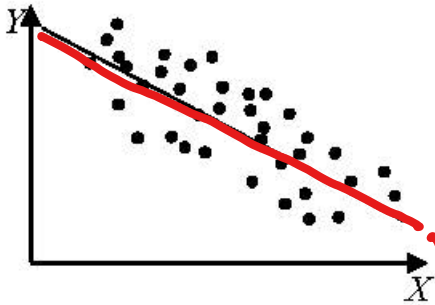
Για να περιγράψουμε τη στοχαστική εξάρτηση δύο μεταβλητών X και Υ προσπαθούμε να βρούμε, όπως και στην προσδιοριστική εξάρτηση, **μια σχέση μεταξύ των X και Y η οποία δεν θα δίνει ακριβή αλλά προσεγγιστική εικόνα της σχέσης** ενώ στο σκεδασμογράφημα τα ζεύγη τιμών (X, Y) δεν θα βρίσκονται πάνω, αλλά, γύρω από την καμπύλη.

Αν απο το σκεδασμογράφημα φαινεται η σχέση να προσεγγίζεται αρκετά από μια ευθεία τοτε εχουμε την πιο απλη περιπτωση , την περιπτωση της **απλής γραμμικής παλινδρόμησης**.



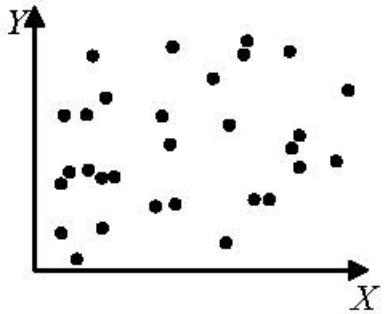
$y = \alpha + bx + \beta x^2 + \varepsilon$

nejvhodnější



$y = \alpha + \beta x + \varepsilon$

přesně



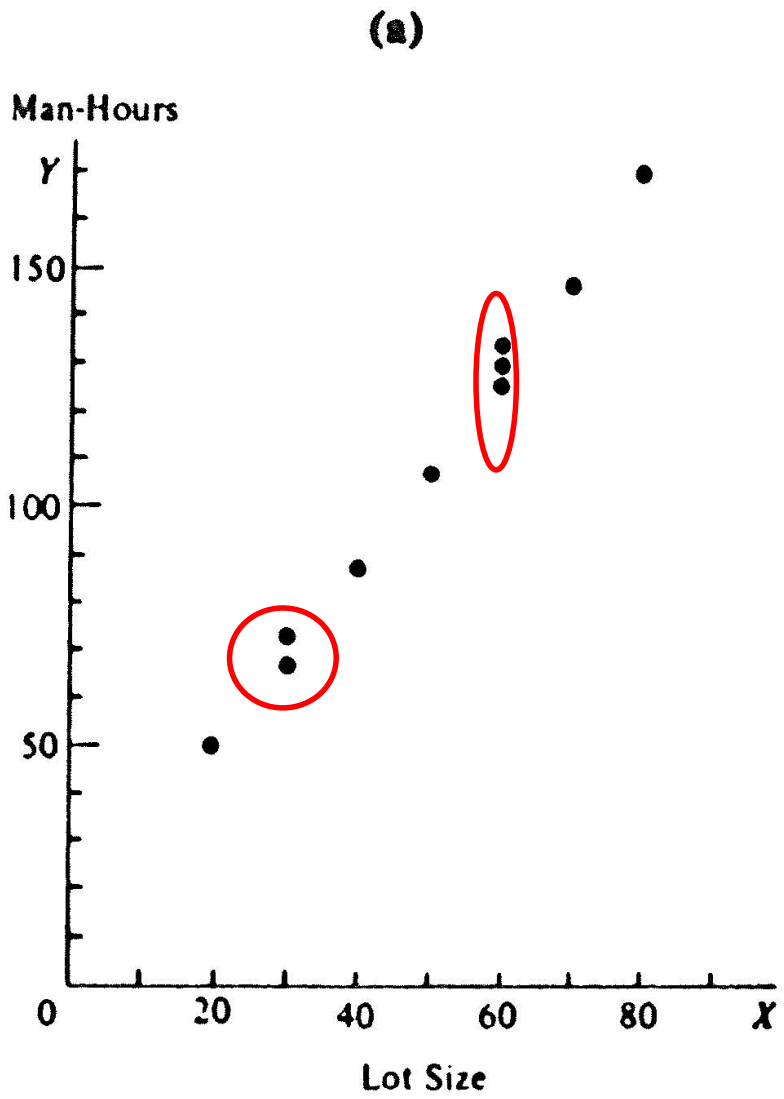
$\hat{y} = \hat{\alpha} + \hat{b}x$

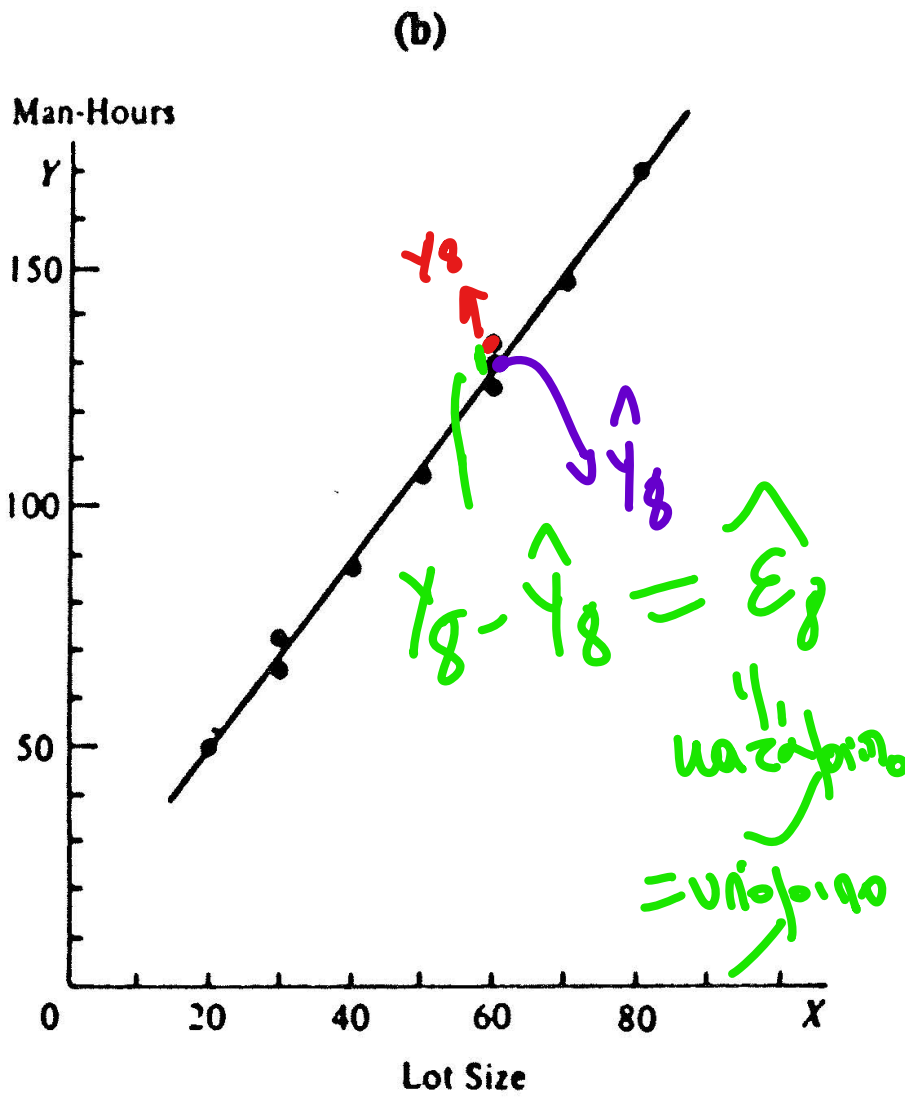
Η ευθεία

$$Y = \alpha + \beta X$$

με α και β σταθερες, **δεν μπορει να περιγραψει ιδανικα τη σχεση** των X και Y αφου οπως ειπώθηκε πριν, αν στο X δωσουμε π.χ. δυο φορες μια συγκεκριμένη τιμη x το αποτέλεσμα δηλ. το Y δεν θα βγει ιδιο.

Οι διαφορετικες τιμες του Y μπορει να οφειλονται απλά, **σε (προφανώς ΤΥΧΑΙΑ) σφάλματα μέτρησης**. Τα σφάλματα αυτα ως τυχαία θα πρέπει να έχουν μια κατανομή. Αυτό συνεπάγεται η Y να μεταβάλεται (τυχαία) ανάλογα με τις τιμές της X και άρα να αποτελεί μια **τυχαια μεταβλητη** και ως τέτοια θα πρέπει να έχει μια κατανομή.


(a)
Man-Hours
Y
150
100
50
0
20
40
60
80
X
Lot Size


(b)
Man-Hours
Y
150
100
50
0
20
40
60
80
X
Lot Size
Y_8
$\hat{Y}_8$
$Y_8 - \hat{Y}_8 = \hat{\varepsilon}_8$

Έτσι στην αρχική εξίσωση προσθέτουμε έναν ακόμα όρο ο οποίος, για δεδομένη τιμή της X, να περιγράφει τη διαφορά της παρατηρούμενης από τη θεωρητική (δηλ. απο αυτη που θα έπρεπε αν ΔΕΝ υπήρχε το σφάλμα, δηλ. $\alpha + \beta X$) τιμή της Y.

Η ποσότητα αυτή ονομάζεται **σφάλμα** και ισχύει $\epsilon = Y - (\alpha + \beta X)$. Ετσι προκύπτει το **στοχαστικό μοντέλο**

$$Y = \alpha + \beta X + \epsilon$$

$\varepsilon \sim N(0, \sigma^2)$

Η πιο συνήθης υπόθεση για τα σφάλματα είναι ότι έχουν μέση τιμή $E(\epsilon) = 0$ και διασπορά $E(\epsilon^2) = Var(\epsilon) = \sigma^2$.

Αργότερα θα υποθέσουμε επιπλέον οτι η κατανομή είναι η **Κανονική (Normal, Gauss)**.

$EY = E(a+bx+\varepsilon) = a+bx+E(\epsilon) = a+bx$

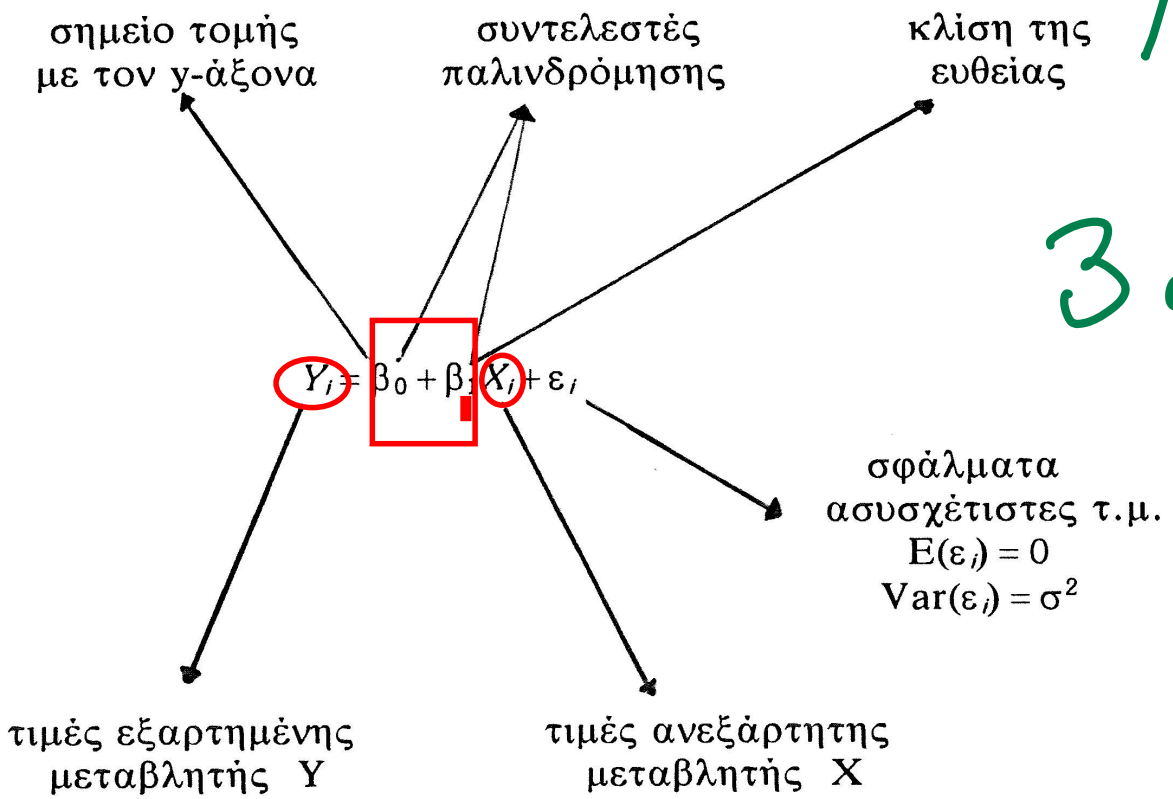
ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

ΜΟΝΤΕΛΟ ΠΡΩΤΗΣ ΤΑΞΗΣ

$Var(Y) = Var(\alpha + \beta x + \varepsilon)$

$\equiv Var(\varepsilon) = \sigma^2$

$Y \sim N(\alpha + \beta x, \sigma^2)$



3 αγνωστοι:

b_0, b_1, σ^2

Ως συνέπεια της υπόθεσης

$$E(Y) = \alpha + \beta X \quad \& \quad Var(Y) = \sigma^2$$

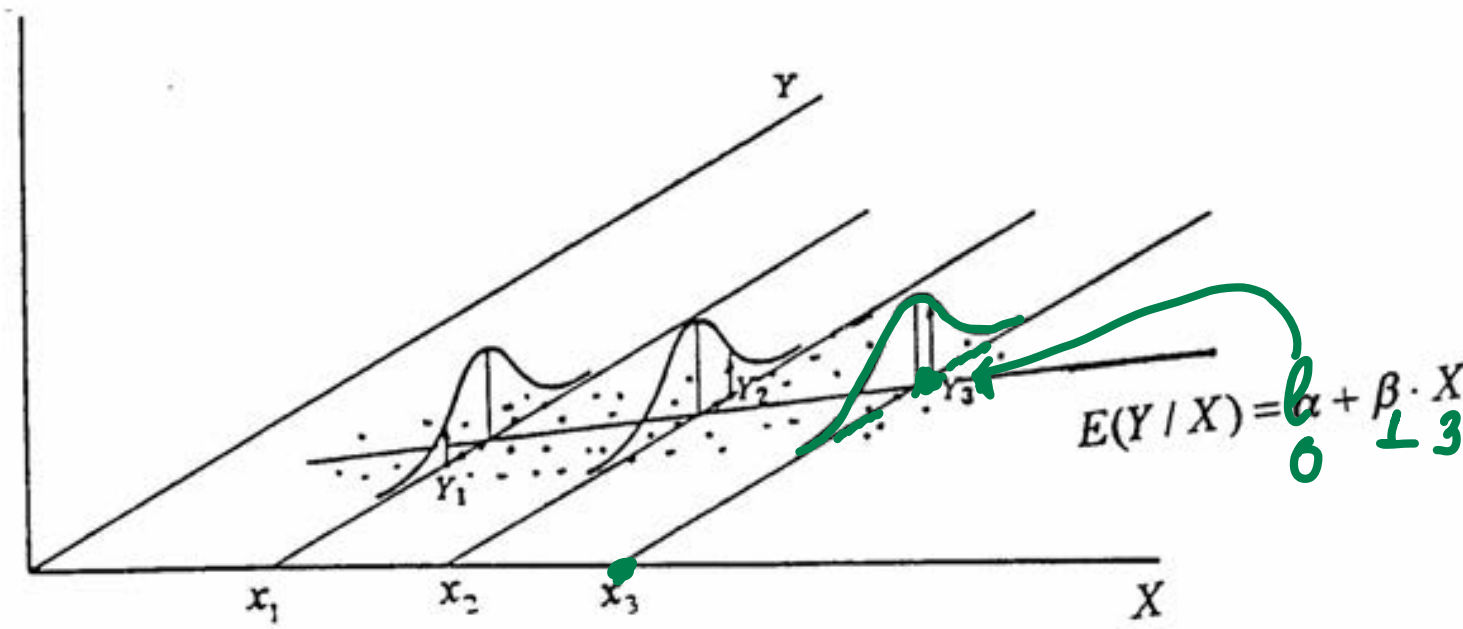
ή καλύτερα

$$E(Y|X) = \alpha + \beta X \quad \& \quad Var(Y|X) = \sigma^2$$

Φυσικά κάτω από την πρόσθετη υπόθεση θα έχουμε

$$Y \equiv Y|X \sim N(\alpha + \beta X, \sigma^2)$$

αφου η Y ειναι μια απλή συνάρτηση της τ.μ. ε.


Y
Y_1
Y_2
Y_3
$E(Y/X) = \alpha + \beta \cdot X$
x_1
x_2
x_3
X

Στόχος είναι ο **εντοπισμός εκτιμητών** έστω $\hat{\alpha}$ και $\hat{\beta}$ των παραμέτρων α, β καθως και $\hat{\sigma}^2$ του σ^2 ωστε να μπορουμε να προχωρήσουμε σε **προβλέψεις**. Η εκτίμηση θα γίνει με τη βοήθεια διαθέσιμων παρατηρήσεων $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Εχοντας τα $\hat{\alpha}$ και $\hat{\beta}$ μπορούμε για οποιαδήποτε τιμή της X να προσεγγίσουμε (εκτιμήσουμε - προβλέψουμε) την τιμή της Y με την τιμή $\hat{Y}$ απο τη σχέση

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X$$

η οποία ονομάζεται **εκτιμήτρια ευθεία παλινδρόμησης** και **αποτελεί εκτίμηση/προσέγγιση της καλύτερης ευθείας που περιγράφει με τον ιδανικότερο τρόπο τα δεδομένα**.

Το σ^2 θα χρησιμοποιηθεί για να διαπιστωθεί **ποσο καλή (ακριβής) είναι η πρόβλεψη** $\hat{Y}$ (αλλά και οι εκτιμήτριες των παραμέτρων).

ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Το στοχαστικο μοντέλο στην **Απλή Γραμμική Παλινδρόμηση (simple linear regression)** δίνεται από τη σχέση

$$Y = \beta_0 + \beta_1 X + \epsilon$$

NORMAL

όπου υποθέτουμε ότι τα σφάλματα είναι ανεξάρτητα με μέση τιμή $E(\epsilon) = 0$ και διασπορά $E(\epsilon^2) = Var(\epsilon) = \sigma^2$.

Η εκτιμήτρια ευθεία παλινδρόμησης δίνεται από τη σχέση

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

όπου $\hat{\beta}_0$ και $\hat{\beta}_1$ οι εκτιμητες των παραμέτρων β_0, β_1, ενώ με $\hat{\sigma}^2$ συμβολίζεται η εκτιμήτρια της διασποράς σ^2 των σφαλμάτων.

Ενδιαφέρει για $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Έστω n ζεύγη τιμών $(X_1, Y_1), \ldots, (X_n, Y_n)$. Με βαση τα προηγουμενα οι αποκλίσεις των παρατηρηθεισών τιμών από αυτές που θα έπρεπε να παρατηρηθούν αν δεν υπήρχαν σφάλματα είναι

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

όπου τα β_0 και β_1 ειναι οι προς εκτίμηση παράμετροι.

$$\epsilon_i^2 = [Y_i - (b_0 + b_1 X_i)]^2 \qquad i = 1, 2, \ldots, n$$

$$\sum \epsilon_i^2 = \sum [Y_i - (b_0 + b_1 X_i)]^2$$

↑ βρίσκω b_0, b_1 ωστε $\sum \epsilon_i^2 \equiv$ ελάχιστο

- $Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$
- Ελαχιστοποιούμε την Q ως προς β_0, β_1 και έχουμε τους εκτιμητές ελαχίστων τετραγώνων:

LEAST SQUARES METHOD

- $$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
- $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum (x_i - \bar{x})}{\sum (\ )^2} \cdot Y_i + 0$$

$= k_i$

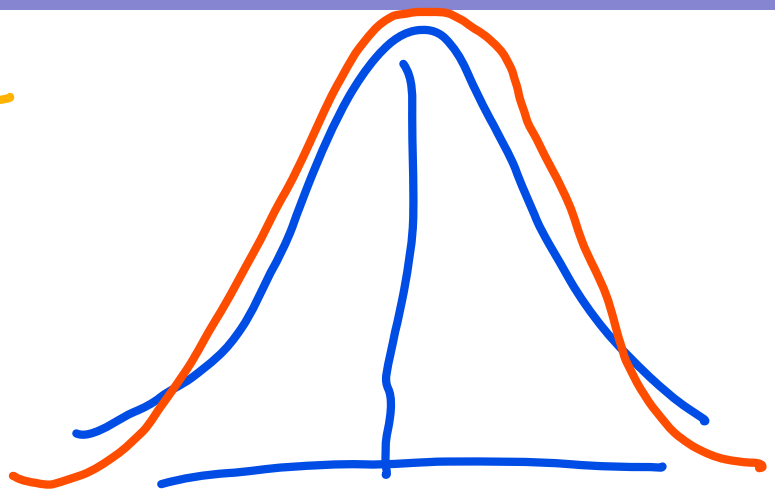
$\frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \cdot \bar{y}$

ex1:

R: lm($Y \sim X_1 + X_2 + X_3$)

summary(ex1)

- $\hat{\beta}_1 = \sum_{i=1}^{n} \kappa_i \cdot Y_i$
- $\hat{\beta}_0 = \sum_{i=1}^{n} (\frac{1}{n} - \kappa_i \bar{X}) Y_i$



όπου

$$\kappa_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\sum_{i=1}^{n} \kappa_i = 0, \quad \sum_{i=1}^{n} \kappa_i X_i = 1, \quad \sum_{i=1}^{n} \kappa_i^2 = \left\{ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right\}^{-1}$$

- $\hat{\epsilon}_i = \hat{Y}_i - Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i - Y_i$

έχοντας,

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0, \quad \sum_{i=1}^{n} \hat{\epsilon}_i X_i = 0$$

$$\sum_{i=1}^{n} \hat{\epsilon}_i \hat{Y}_i = 0, \quad \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \text{minimum}.$$

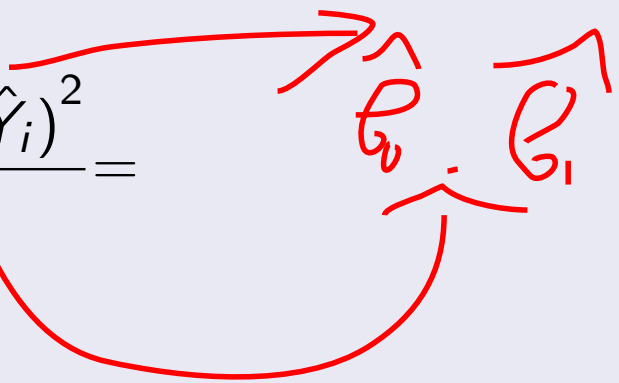
Εκτιμήτρια Διασποράς

- $\hat{\sigma}^2 = s^2 = MSE$

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

$$= \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-2}$$

$$= \frac{SSE}{n-2}$$



- $E(\hat{\beta}_1) = \beta_1, \quad E(\hat{\beta}_0) = \beta_0$
- $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- $Var(\hat{\beta}_0) = \sigma^2 \left\{ \dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right\}$

Κανονικότητα-Ομοσκεδαστικότητα

$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

- $\epsilon_i \sim N(0, \sigma^2)$, με
- $Likelihood(\beta_0, \beta_1, \sigma^2) =$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 \right\}, \text{ και έχουμε:}$$

$$MLE \equiv LSE$$

Συμπερασματολογία

- $\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1))$
- $\hat{\beta}_0 \sim N(\beta_0, Var(\hat{\beta}_0))$ έχοντας,

$$t = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\widehat{Var}(\hat{\beta}_i)}} \sim t_{n-2}, \qquad i = 0, 1, \ldots$$

Διαστήματα Εμπιστοσύνης

- $\hat{\beta}_i \pm t_{n-2;\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\beta}_i)}, \quad i = 0, 1, \ldots$

  Έλεγχοι: $H_0 : \beta_i = \beta_{i0}$

- Με το Στατιστικό τέστ να είναι:

$$t = \frac{\hat{\beta}_i - \beta_{i0}}{\sqrt{\widehat{Var}(\hat{\beta}_i)}}, \quad i = 0, 1, \ldots$$

F-test

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

- Με το Στατιστικό τέστ να είναι:

$$F = \frac{MSR}{MSE} \overset{H_0}{\sim} F_{1,n-2} \equiv t^2_{n-2}$$

R: anova(ex1)

Αναλυση της SSTO

	Sum of Squares	d.f.	Mean Square	F
Regression	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	p-1	$MSR = \frac{SSR}{(p-1)}$	$\frac{MSR}{MSE}$
Error	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	n-p	$MSE = \frac{SSE}{(n-p)}$	
Total	$SSTo = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	n-1		

$r^2 = SSR/SSTO \approx 1$ τότε καλή

όπου n = # παρατηρήσεων και p = # επεξηγηματικών μεταβλητών +1 = # παραμέτρων β

n-25 ββ1

df

1
23
24

p=2 για την απλη γραμμικη παλινδρομηση

Ερμηνεία ANOVA

SSTo = Ολικό Άθροισμα Τετραγώνων: Εκφράζει τη συνολική μεταβλητότητα (διακύμανση) των παρατηρήσεων y_i, δηλ. την αβεβαιότητα στην πρόβλεψη της Y διχως τη χρήση της X.

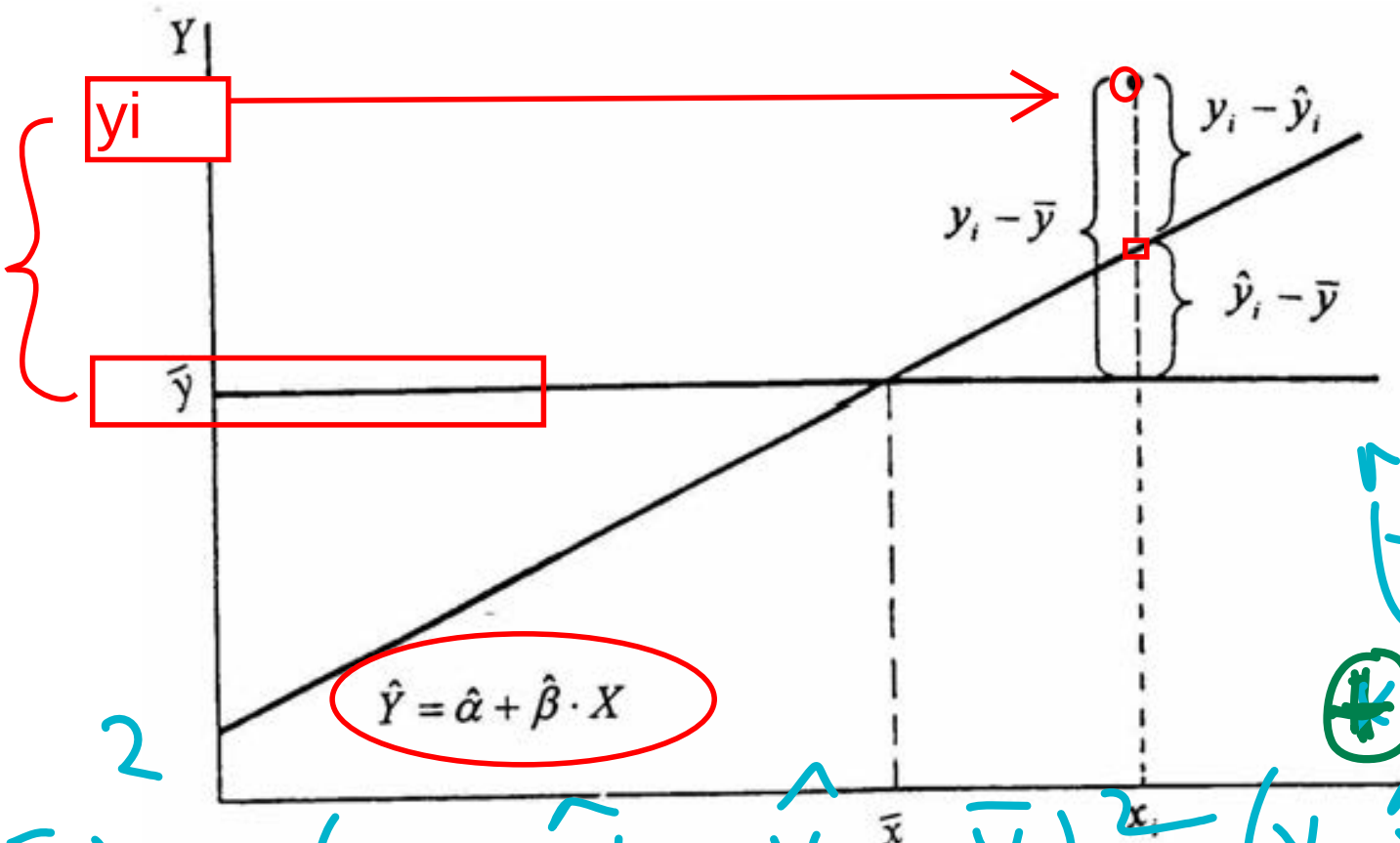
SSR = Άθροισμα Τετραγώνων Παλινδρόμησης: Εκφράζει το μέρος της μεταβλητότητας που μπορεί να οφείλεται στον ανεξάρτητο παράγοντα X, δηλ. στη μεταβλητότητα που επεξηγείται (περιγράφεται, απορροφάται) από το μοντέλο παλινδρόμησης
Εναλλακτικά η πηγή μεταβλητότητας R-Regression μπορεί να διατυπωθεί, μεταξύ άλλων, ως εξής: μεταξύ των δειγμάτων (Between Groups), Επεμβάσεις (Treatmens (Tr)), Παράγοντας (factor) και **Επεξηγηματική (explained)**.

Ερμηνεία ANOVA

SSE = Άθροισμα Τετραγώνων Σφάλματος: Εκφράζει το μέρος της μεταβλητότητας που δεν μπορεί να περιγραφτεί από την παλινδρόμηση και συνεπώς αποτελεί το τμήμα που ΔΕΝ εξηγήθηκε από το μοντέλο (σφάλμα).
Εναλλακτικά η πηγή μεταβλητότητας E-Error μπορεί να διατυπωθεί, μεταξύ άλλων, ως εξής: εντός των δειγμάτων (Within Groups) και **Κατάλοιπα (Residuals)**
MSR και MSE είναι αντίστοιχα το Μέσο Άθροισμα Τετραγώνων της Παλινδρόμησης και του Σφάλματος με το MSE να αποτελεί την εκτιμήτρια της διασποράς σ^2

Ισοδυναμίες Αθροισμάτων Τετραγώνων

$$\text{SSTo=SSR+SSE} \implies 1 = \frac{SSR}{SSTo} + \frac{SSE}{SSTo}$$



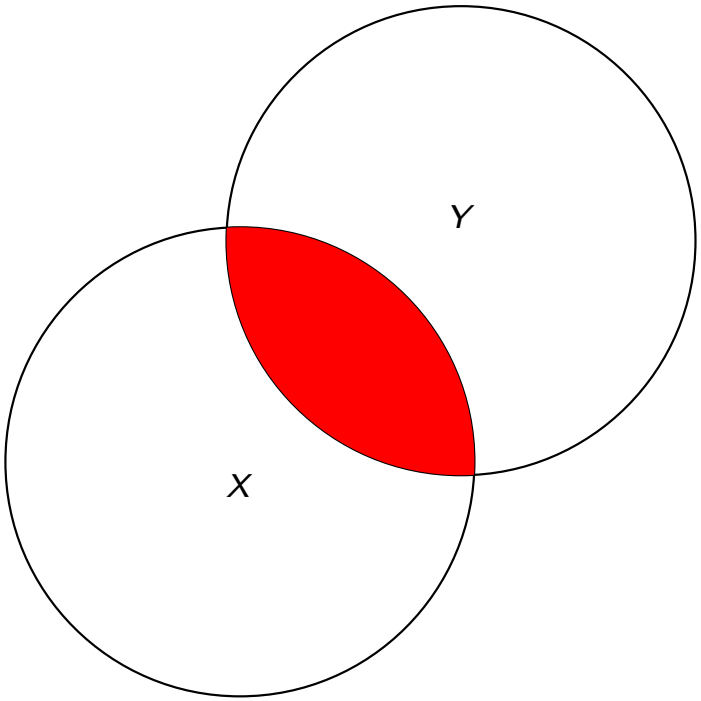
$$(Y_i - \bar{Y})^2 = (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \equiv (Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2$$

Συντελεστής Προσδιορισμού (Coefficient of Determination)

$$R^2 = \frac{SSR}{SST_o} = 1 - \frac{SSE}{SST_o} = \rho^2$$

$$\Longleftrightarrow$$

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



Απλή Γραμμική = Συντελεστής Συσχέτισης Pearson (γραμμική συσχέτιση)

Το R^2 εκφράζει **το ποσοστό της συνολικής μεταβλητότητας** των y_i που **εξηγείται (απορροφάται, περιγράφεται, ερμηνεύεται)** από την παλινδρόμηση.

Το R^2 ονομάζεται **συντελεστής προσδιορισμού (coefficient of determination)** και παίρνει τιμές στο κλειστό διάστημα $[0, 1]$.

Όταν όλα τα σημεία βρίσκονται πάνω στην εκτιμήτρια ευθεία παλινδρόμησης τότε $R^2 \equiv 1$ αφού

$$\hat{y}_i \equiv y_i \quad \& \quad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \equiv 0.$$

Αντίθετα αν δεν υπάρχει ουδεμία (γραμμική) σχέση δηλ. $\beta_1 \equiv 0$ τότε $\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = 0$ και άρα $R^2 \equiv 0$.

Συντελεστής Συσχέτισης και R^2

- $Cov(X, Y) = E(X - \mu_x)(Y - \mu_y)$
- Συντελεστής Συσχέτισης Pearson

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot (Var(Y))}}, \quad |\rho| \leq 1$$

Εκτιμητές

- $\widehat{Cov(X, Y)} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$
- $\hat{\rho}^2 = \left(\frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \right)^2 = r^2 = R^2$

ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Το στοχαστικό μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης είναι:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, 2, ..., n$$

όπου τα σφάλματα είναι *iid* $N(0, \sigma^2)$.

Όλα ισχύουν όπως και για 2 συντελεστές παλινδρόμησης ($p = 2$)

Η εκτιμήτρια πολλαπλής παλινδρόμησης:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + .. + \hat{\beta}_{p-1} X_{p-1}$$

p παράμετροι

όπου $\hat{\beta}_i$ ο **ΕΕΤ** του β_i και $\hat{\sigma}^2$ η εκτιμήτρια του σ^2.

vs. H_1: ΟΧΙ H_0

Γραφική γενικά: $H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_{p-1} = 0$

ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Έστω n ζεύγη τιμών

$$(X_{11}, X_{12}, ..., X_{1,p-1}, Y_1), \ldots, (X_{n1}, X_{n2}, ..., X_{n,p-1}, Y_n)$$

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i - ... - \beta_{p=1} X_{i,p-1})^2$$

Ελαχιστοποιούμε την Q ως προς $\beta_0, \beta_1, ..., \beta_{p-1}$ και έχουμε τους εκτιμητές ελαχίστων τετραγώνων (ΕΕΤ).

Οπως και στην απλή παλινδρόμηση έχουμε ότι **οι εκτιμητές είναι γραμμικός συνδυασμός των** Y_i:

$$\hat{\beta}_i = \sum_{j=1}^{n} \kappa_j \cdot Y_j$$

για κάποιες σταθερές κ.

Προφανώς **τα κατάλοιπα** είναι

$$\hat{\epsilon}_i = \hat{Y}_i - Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_{p-1} X_{i,p-1} - Y_i$$

Εκτιμήτρια Διασποράς

- $\hat{\sigma}^2 = s^2 = MSE$

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

$$= \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-p} = \frac{SSE}{n-p}$$

$p=2$ για απλή γραμμική παλινδρόμηση

$y = b_0 + b_1 x + \varepsilon$

Κανονικότητα

- $\epsilon_i \sim N(0, \sigma^2)$, με

- $Likelihood(\beta_0, \beta_1, \sigma^2) = =$

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1})^2 \right\},$$

και έχουμε:

$$MLE \equiv LSE$$

ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ

- $\hat{\beta}_i \sim N(\beta_i, Var(\hat{\beta}_i))$
- $\hat{\beta}_i \pm t_{n-p;\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\beta}_i)}, \quad i = 0, 1, \ldots$

Έλεγχοι: $H_0 : \beta_i = \beta_{i0}$

- Με το Στατιστικό τέστ να είναι:

$$t = \frac{\beta_i - \beta_{i0}}{\sqrt{\widehat{Var}(\hat{\beta}_i)}}, \quad i = 0, 1, \ldots, p-1$$

- $SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad (d.f = n-1)$
- $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2, \quad (d.f = p-1)$
- $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \quad (d.f = n-p)$. Επίσης έχουμε:

- $MSR = \dfrac{SSR}{(p-1)}$ και
- $MSE = \dfrac{SSE}{(n-p)}$

όπου $p-1 = \#$ ανεξάρτητων μεταβλητών .

Το SSTO μετράει τη συνολική μεταβλητότητα (διακύμανση) των παρατηρήσεων y_i δηλαδή εκφράζει την αβεβαιότητα στην πρόβλεψη του Y όταν δεν χρησιμοποιούνται οι μεταβλητές X.

Το SSR εκφράζει το μέρος της μεταβλητότητας που μπορεί να οφείλεται στις μεταβλητές X δηλ. που εξηγειται (περιγράφεται, απορροφάται) από το μοντέλο (δηλ. την πολλαπλή παλινδρόμηση) και

Το SSE=SSTO-SSR εκφράζει την υπόλοιπη μεταβλητότητα που δεν εξηγείται από την παλινδρόμηση και άρα αποτελεί το τμήμα που ΔΕΝ εξηγήθηκε από το μοντέλο (σφάλμα).

Πίνακας ANOVA

Παλινδρόμηση	SSR	$p-1$	MSR	$F = MSR/MSE$
Σφάλματα	SSE	$n-p$	MSE	
Ολική Διακύμανση	$SSTO$	$n-1$		

Ο Έλεγχος F αφορά:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0 \quad vs \quad H_1 : \text{όχι η } H_0$$

Απορρίπτω H_0 αν $F > F_{p-1,n-p,\alpha/2}$

Συντελεστής Συσχετίσεως

Coefficient of Determination

$$r^2 = \frac{SSR}{SSTO}$$

Το r^2 εκφράζει **το ποσοστό της συνολικής μεταβλητότητας** των y_i που **εξηγείται (απορροφάται, περιγράφεται, ερμηνεύεται)** από την παλινδρόμηση.

Το r^2 ονομάζεται **συντελεστής προσδιορισμού (coefficient of determination)** και παίρνει τιμές στο κλειστό διάστημα $[0, 1]$.

ΠΡΟΫΠΟΘΕΣΕΙΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ- ΑΝΑΛΥΣΗ ΚΑΤΑΛΟΙΠΩΝ

Η πιο συνήθης υπόθεση για τα σφάλματα είναι ότι έχουν **σταθερή μέση τιμή** $E(\epsilon) = 0$ και **σταθερή διασπορά** $E(\epsilon^2) = Var(\epsilon) = \sigma^2$.

ομοσκεδαστικότητα

Ειναι επισης αυτονόητο ότι δεν είναι δυνατόν το σφάλμα μιας μέτρησης να επηρεάζει την επόμενη ή τις επόμενες μετρήσεις. Ως εκ τούτου **τα σφάλματα υποθέτουμε ότι είναι ασυσχέτιστα**, δηλ. $Cov(\epsilon_i, \epsilon_j) = 0$.

→ ανεξαρτησία

Τέλος υποθέσουμε ότι η κατανομή των σφαλμάτων είναι η **Κανονική** **(Normal, Gauss)**.

Και φυσικά θα πρέπει να υφίσταται **ισχυρή γραμμική σχέση** μεταξύ των Υ και X.

ΔΙΕΡΕΥΝΗΣΗ ΚΑΝΟΝΙΚΟΤΗΤΑΣ

Γραφικές Μέθοδοι: Ιστόγραμμα, BoxPlot, Σκεδασμογράφημα, P-P Plot & Q-Q Plot

Διαγνωστικοί Έλεγχοι: Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilks, Lilliefors

Οι έλεγχοι εφαρμόζονται είτε στα κατάλοιπα είτε στα ΗμιΤυποποιημένα Κατάλοιπα.

$\not\exists\ \varepsilon \rightarrow$ μόνο αυτά βρίσκω $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots \hat{\varepsilon}_n$

$\forall$ διαφόρων παρατηρήσεων λέω σε ε_i

ΓΡΑΦΙΚΕΣ ΜΕΘΟΔΟΙ

Τα **P-P Plot (Probability-Probability Plot) και Q-Q Plot (Quantile-Quantile Plot)** είναι γραφικές μέθοδοι για να διερευνηθεί αν τα δεδομένα προέρχονται από συγκεκριμένη κατανομή (π.χ. κανονική).

Τα γραφήματα αυτά βασίζονται στην ακόλουθη παρατήρηση:

Αν $X_1, \ldots, X_n$ τ.δ. από συνεχή κατανομή τότε $Y_1 = F(X_1), \ldots, Y_n = F(X_n)$ είναι ανεξάρτητες και $\sim U(0,1)$.

Επίσης αποδεικνύεται ότι οι διατεταγμένες $Y_{(i)} \equiv F(X_{(i)}) \sim Beta(i, n-i+1)$ με μέσο $E(Y_{(i)}) = i/(n+1)$.

Έτσι

$$Y_{(i)} \equiv F(X_{(i)}) \approx i/(n+1) \quad or \quad X_{(i)} \approx F^{-1}(i/(n+1))$$

Αρα αν $X_i \sim F_0$ τότε τα σημεία

$$\left(F_0(X_{(i)}), i/(n+1)\right), \; i = 1, \ldots, n$$ P-P

ή ισοδυναμα τα σημεία

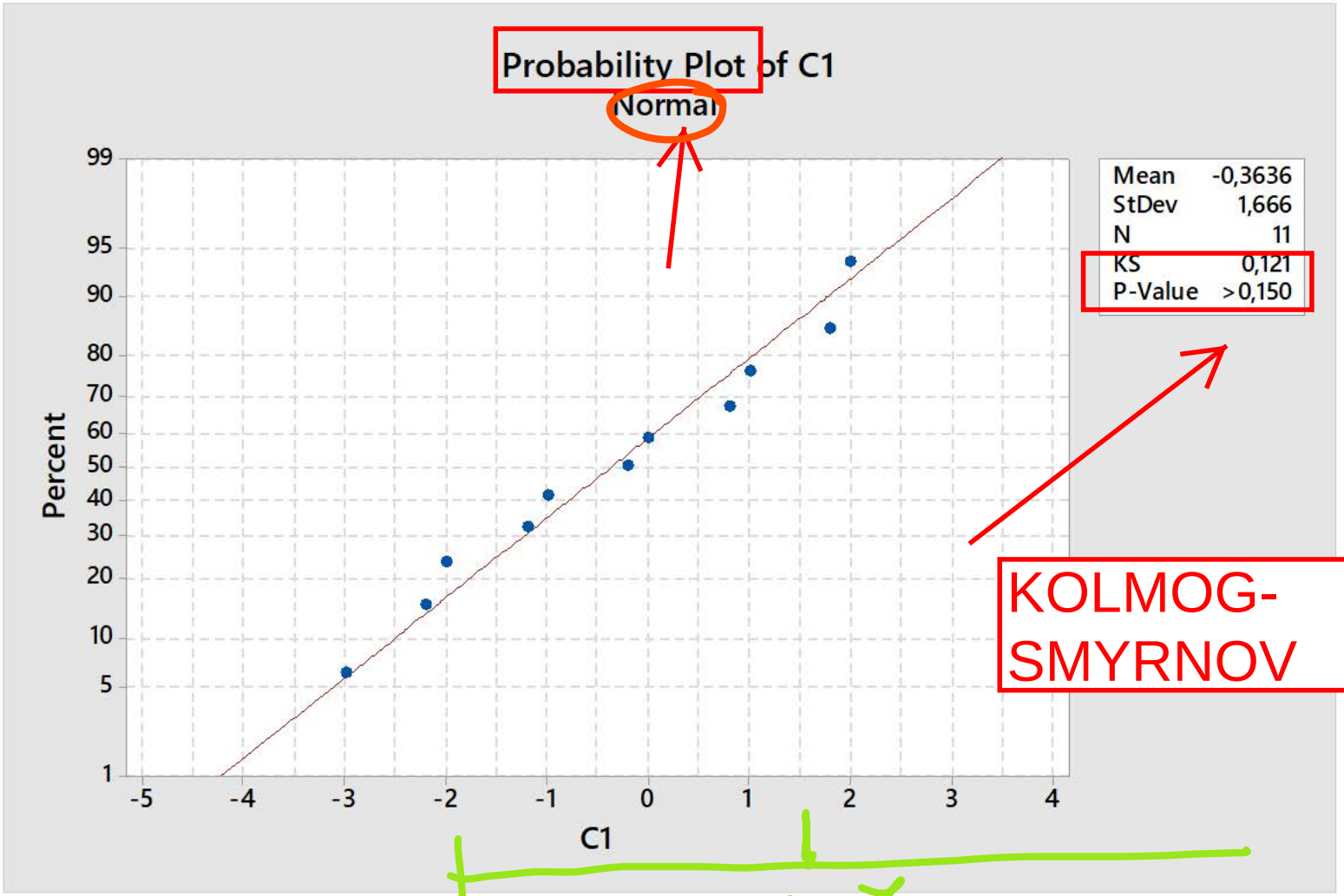
$$\left(X_{(i)}, F_0^{-1}(i/(n+1))\right), \; i = 1, \ldots, n$$ Q-Q

θα βρίσκονται κοντά στη διαγώνιο που περνά από την αρχή των αξόνων.

Το P-P plot είναι το γράφημα των πρώτων σημείων και το Q-Q plot των δεύτερων.

Σημεία κοντά στη διαγώνιο (και τυχαία γύρω απο αυτήν) συνηγορούν υπέρ της μηδενικής F_0.

Αγνωστες πράμετροι της F_0 εκτιμώνται απο τα δεδομένα.


Probability Plot of C1
Normal
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-5
-4
-3
-2
-1
0
1
2
3
4
C1
Mean -0,3636
StDev 1,666
N 11
KS 0,121
P-Value >0,150
KOLMOG-SMYRNOV
0.07

ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ/ΕΛΕΓΧΟΙ

Kolmogorov Test

$$X_1, X_2, ..., X_n \sim F(x)$$

1. $H_0 : F(x) = F^*(x)$
2. $H_0 : F(x) \leq F^*(x)$
3. $H_0 : F(x) \geq F^*(x)$

Test Statistic

1. $T^1 = sup|F_n(x) - F^*(x)|$
2. $T^2 = sup[F_n(x) - F^*(x)]$
3. $T^3 = sup[F^*(x) - F_n(x)]$

Χ. Α. H_0: T^1 ή T^2 ή $T^3 > w_n(a)$ (ειδικοί πίνακες) όπου **η εμπειρική κατανομή (EDF)** F_n δίδεται από τον τύπο:

$$F_n(x') = \frac{1}{n}\{\#x_i \leq x'\}$$

Έλεγχος Lilliefors - Κανονικότητα

$H_0 : X_i \sim N(\mu, \sigma^2) \quad \Leftrightarrow \quad H_0 : Z_i = \frac{X_i - \bar{X}}{S} \sim N(0, 1)$

Test Statistic: $T = sup[F_n(z) - F(z)]$

$F_n(z)$ η εμπειρική κατανομή των z_i και

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

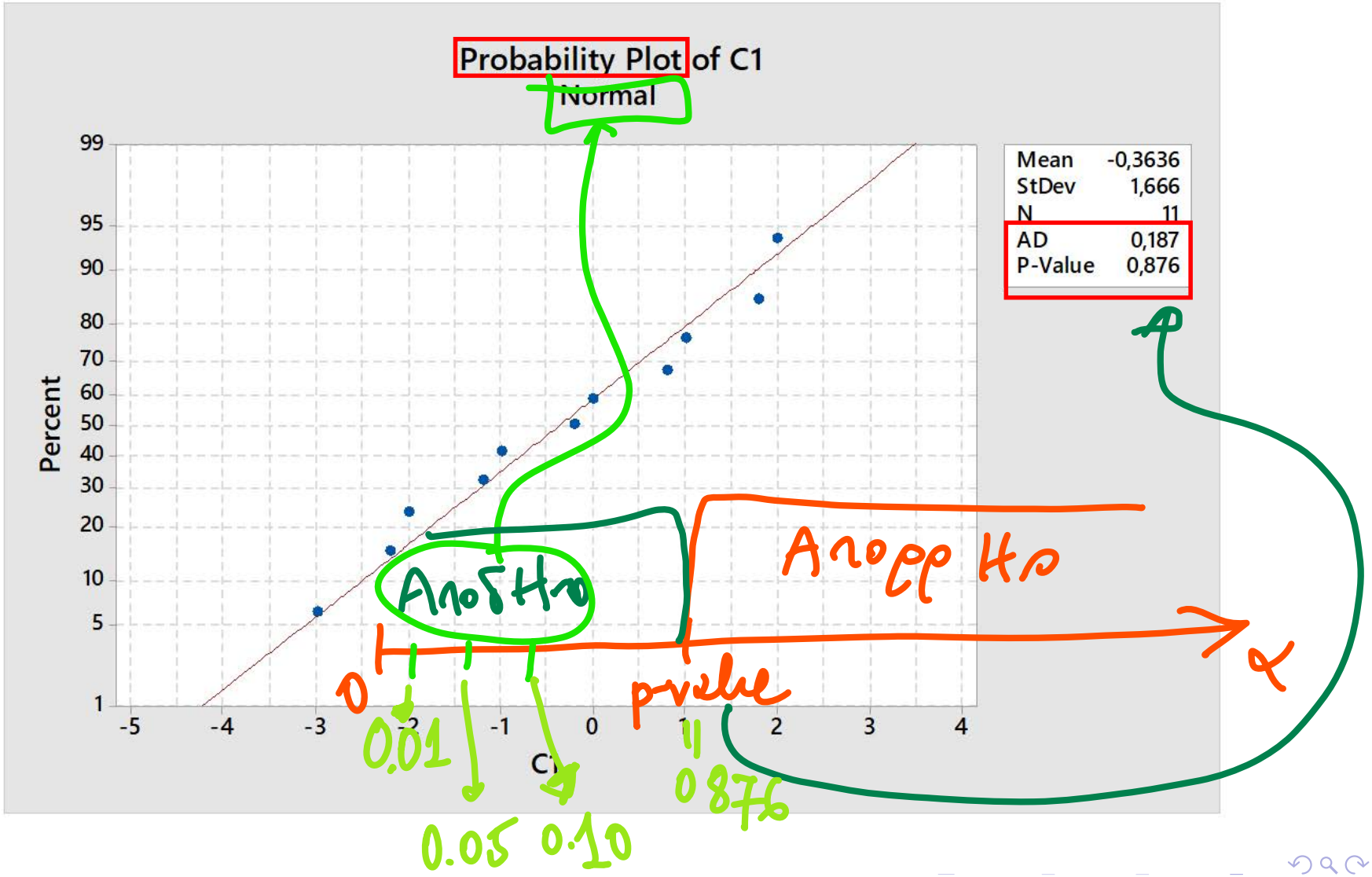
Εκτός απο τον Kolmogorov έλεγχο εχουμε αρκετούς ακομα ελέγχους που βασίζονται στην εμπειρική συνάρτηση κατανομής. Οι έλεγχοι αυτοί αναφέρονται ώς *EDF* **έλεγχοι και ειδικότερα ως quadratic EDF έλεγχοι** επειδή σε αντίθεση με τον έλεγχο Kolmogorov, βασίζοται στην τετραγωνική απόκλιση και όχι στην απόλυτη απόκλιση. Οι γνωστοτεροι απο τους ελέγχους αυτούς είναι

Anderson-Darling (AD)

Cramer-von Mises (CvM)

Watson (παραλλαγή του προηγούμενου)

ενώ υπάρχει και ο έλεγχος **Shapiro-Wilk (SW)** ο οποιος βασίζεται στο διατεταγμένο δείγμα ενώ κάνει χρήση ειδικών συντελεστών.


Probability Plot of C1
Normal
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-5
-4
-3
-2
-1
0
1
2
3
4
C1
Mean -0,3636
StDev 1,666
N 11
AD 0,187
P-Value 0,876
Απορρ Η0
Απορ Η0
0
0.01
0.05
0.10
pvalue
0876
α

ΔΙΕΡΕΥΝΗΣΗ ΤΥΧΑΙΟΤΗΤΑΣ/ΑΝΕΞΑΡΤΗΣΙΑΣ

Συχνά σε οικονομετρικές μελέτες εξετάζεται **η γραφική παράσταση των καταλοίπων με τον χρόνο** t (ακομα και αν δεν περιλαμβάεται στις υπο εξέταση μεταβλητές).

Αν τα σημεία βρίσκονται τυχαία κατανεμημένα γύρω από το 0 **η υπόθεση της τυχαιότητας γίνεται δεκτή.**

Αν υπάρχει μια αυξητική/ανοδική τάση τότε υπάρχει εξάρτηση σε σχέση με το χρόνο και συστήνεται να περιληφθεί στο μοντέλο έξτρα πολυωνυμικός όρος του χρόνου, π.χ.

$$\beta_{00} t^2$$

ΔΙΕΡΕΥΝΗΣΗ ΤΥΧΑΙΟΤΗΤΑΣ/ΑΝΕΞΑΡΤΗΣΙΑΣ -RUNS TEST

Στο Runs Test κάθε κατάλοιπο αντικαθίσταται απο το πρόσημο του και δημιουργείται μια ακολουθία προσήμων ¨+' και ¨-' της μορφής (τα 0 δεν λαμβάνονται υπόψη) π.χ.

$$+, +, +, +, -, -, +, -, -, +, +, +$$

Αν τα προσημα εναλλάσονται με εναν τυχαίο τρόπο (τυχαία διάταξη) και χωρις κάποιο μοτιβο (pattern) τότε γίνεται δεκτή η τυχαιότητα.

ΔΙΕΡΕΥΝΗΣΗ ΤΥΧΑΙΟΤΗΤΑΣ/ΑΝΕΞΑΡΤΗΣΙΑΣ -RUNS TEST

Ροή μήκους k ονομάζεται μια σειρά από k διαδοχικά ¨+' (ή ¨-') που έπονται από ένα τουλάχιστον ¨-' (ή ¨+'). Το παράδειγμα έχει 5 ροές.

Αν u **ο αριθμός των ροών** σε ένα τ.δ. τότε

$$E(u) = \mu = \frac{2n_1n_2}{n_1 + n_2} + 1$$

και

$$Var(u) = \sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

όπου $n_1 = \#" + "$ και $n_2 = \#" - "$.

ΔΙΕΡΕΥΝΗΣΗ ΤΥΧΑΙΟΤΗΤΑΣ/ΑΝΕΞΑΡΤΗΣΙΑΣ -RUNS TEST

Αποδεικνύεται ότι

$$Z = \frac{u - \mu}{\sigma} \sim N(0, 1)$$

οπότε **η τυχαιότητα/ανεξαρτησία απορρίπτεται** αν

$$|Z| \geq z_{\alpha/2}.$$