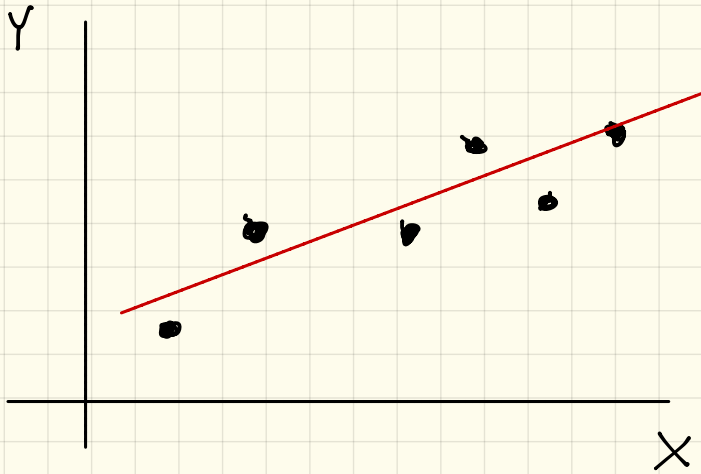
Lezione n. 34

(martedì, 1 marzo 2016)

	Task Description	Due Date	Status
☐	Richiami sulla regressione e differenza nel caso di indagini campionarie		
☐	Le ipotesi alla base del modello classico di regressione lineare		
☐	Teorema di Gauss - Markov		
☐	Le ipotesi alla base del modello classico normale di regressione		
☐	Stima della varianza dell'errore		
☐	Stima intervallare e verifica di ipotesi per i parametri di regressione		
☐	ARGOMENTO AGGIUNTIVO R^2 e formule inverse per calcolare Dev(R) e Dev(E)		
☐			
☐			
☐			

RICHIAMI REGRESSIONE (1)



→ obiettivo:
"migliore" retta che interpola il sistema di punti

criterio minimi quadrati

$$\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$\hat{y}_i$ → $b_0 + b_1 x_i$

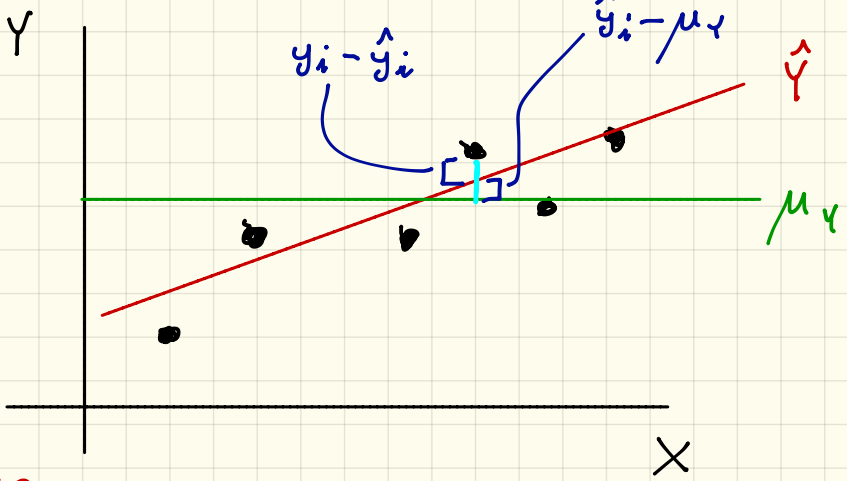
soluzioni dei minimi quadrati

$$\begin{cases} b_1 = \dfrac{\text{coder}(X,Y)}{\text{Der}(X)} = \dfrac{\text{cor}(X,Y)}{\text{var}(X)} \\ b_0 = \mu_Y - b_1 \mu_X \end{cases}$$

RICHIAMI REGRESSIONE (2)

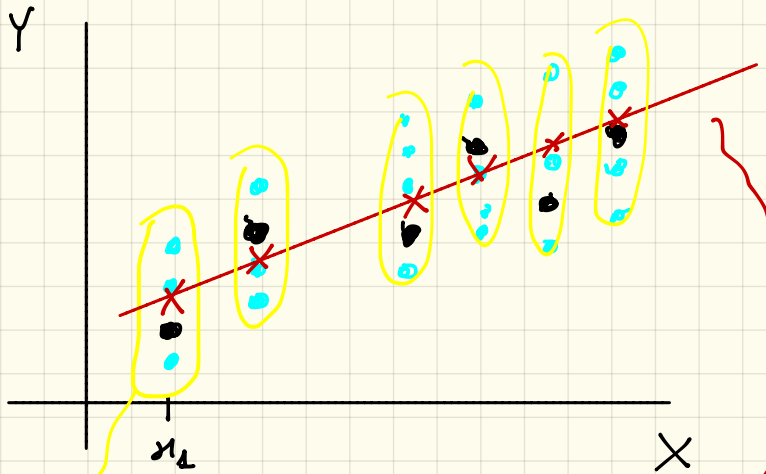
$$R^2 = \frac{\text{Dev. spiegata}}{\text{Dev. totale}} = \frac{\text{Dev}(\hat{Y})}{\text{Dev}(Y)} = 1 - \frac{\text{Dev}(E)}{\text{Dev}(Y)}$$

$$= \rho^2_{XY} = \hat{\beta}_1^2 \frac{\text{Dev}(X)}{\text{Dev}(Y)}$$



MEMO:

$$\sum_i (y_i - \mu_Y)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \mu_Y)^2$$



valori osservati (campione)

$\mu_{Y|X=x_i}$ media condizionata

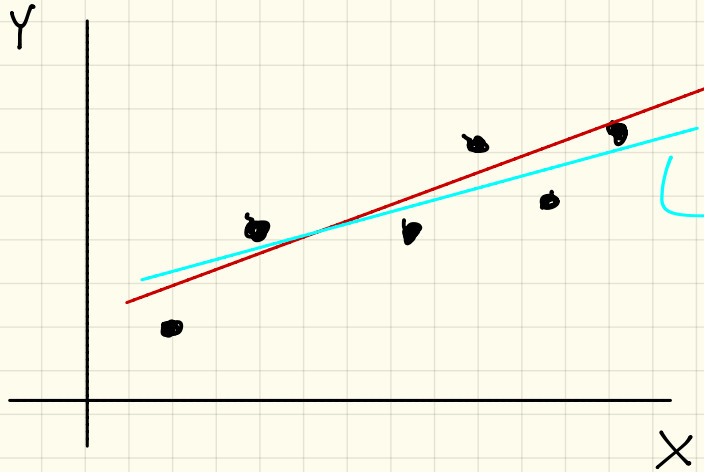
$Y|X=x_1$

Retta di regressione: luogo geometrico su cui giacciono le medie condizionate

RETTA REGRESSIONE POPOLAZIONE

le medie condizionate $Y|X=x_i$ cadono su una retta

↳ legame lineare tra X e Y



retta popolazione → β_0 e β_1 coefficiente retta popolaz.

→ stima retta usando i dati campionari

$\hat{\beta}_0$ (o b_0) e $\hat{\beta}_1$ (o b_1) sono gli stimatori di β_0 e β_1, rispettivamente, usando il criterio dei minimi quadrati

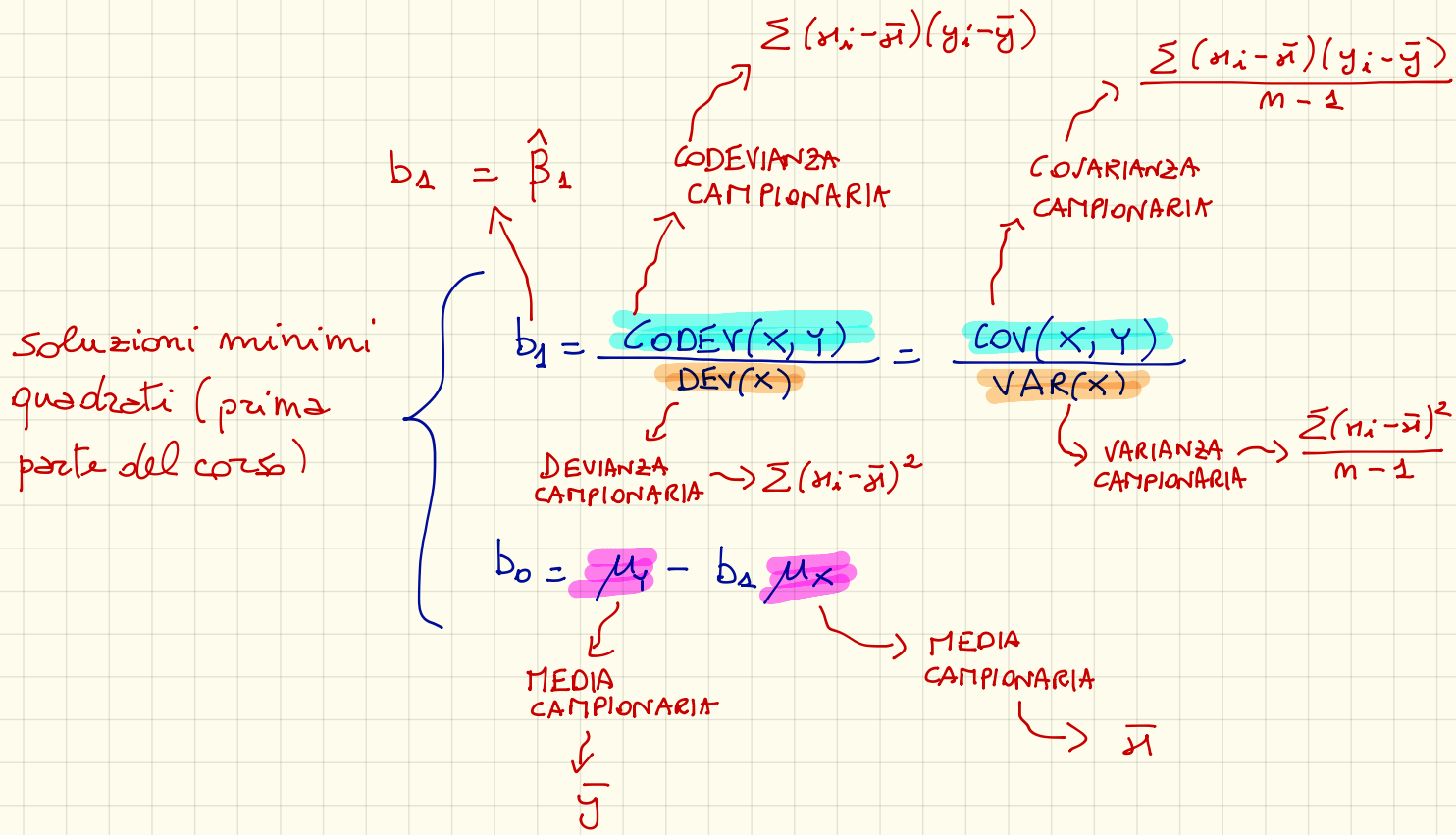
POPOLAZIONE : due parametri

β_0 → $b_0 \equiv \hat{\beta}_0$

β_1 → $b_1 \equiv \hat{\beta}_1$

stimatori usando il criterio dei minimi quadrati

Come cambiano le formule nel passaggio ai dati campionari:



IPOTESI (ASSUNZIONI) MODELLO CLASSICO DI REGRESSIONE

(0) Esiste un legame lineare tra X e Y

↳ le medie condizionate della distribuzione $Y|X=x_i$ giacciono su una retta

↳ $E(Y|X=x_i)$ = funzione lineare = $\beta_0 + \beta_1 x$

$E(Y|X=x_i)$: funzione di regressione

β_0: intercetta o termine noto

β_1: pendenza o coeffic. angolare

se estendo il punto di vista alle intere distribuzioni condizionate $Y|X=x_i$

$$Y = \underbrace{\beta_0 + \beta_1 x}_{\text{deterministica}} + \varepsilon$$

$\beta_0 + \beta_1 x$: parte lineare

ε: parte casuale → v.c. errore

↳ sintesi di tutto quello che spiega Y ma che non ho inserito nel modello

Su ε (v.c. errore) si fanno una serie di ipotesi

① $E(\varepsilon) = 0$

↳ dovrei scrivere $E(\varepsilon \mid X = x_i)$

non ci sono fattori sistematici che non ho considerato nel modello

↳ tecnicamente $Y = \beta_0 + \beta_1 X + \varepsilon$

$$E(Y \mid X = x) = E(\overbrace{\beta_0 + \beta_1 X}^{\text{deterministica}} + \underbrace{\varepsilon}_{\text{casuale}} \mid X = x) =$$

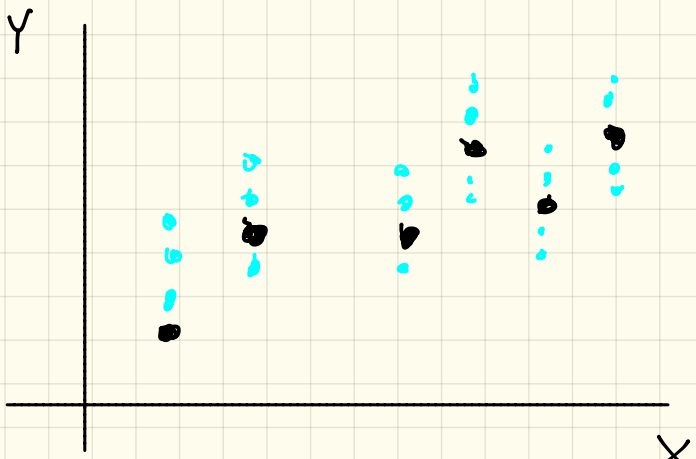
$$= \underline{E}(\beta_0 + \beta_1 X \mid X = x) + \underbrace{\underline{E}(\varepsilon \mid X = x)}_{=0 \text{ sotto l'ipotesi ①}} =$$

$$= \beta_0 + \beta_1 X$$

(2) $Var(\varepsilon) = \sigma_\varepsilon^2 = \sigma^2$ → costante

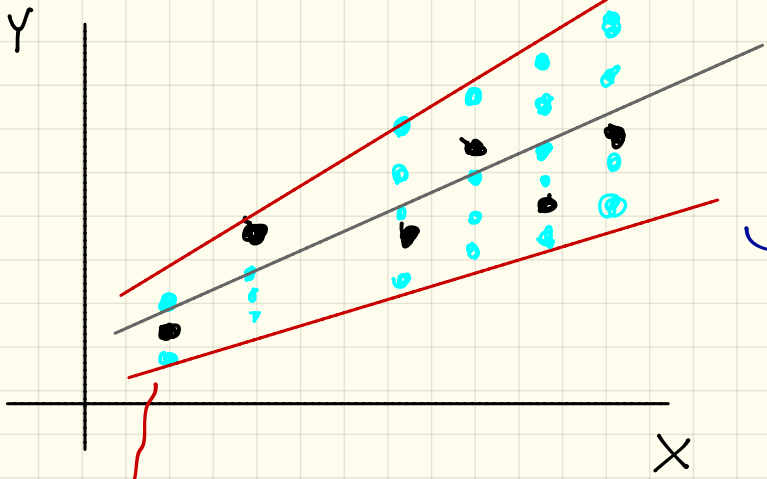
↳ $Var(\varepsilon | X = x) = \sigma^2$ non dipende da X

ipotesi di omoschedasticità



→ le distribuz. condizionate $Y|X=x_i$ sono tutte caratterizzate dalla stessa variabilità (σ_ε^2)

da un punto di vista pratico questo richiede di stimare un unico parametro (σ^2) per "governare" la v.c. ε

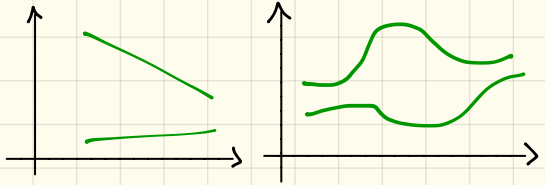


retta regressione

→ al crescere di X aumenta la varianza delle distribuzioni condizionate, ovvero la varianza di $\varepsilon | X=x$

per questa distribuzione condizionata c'è minore probabilità di sbagliare usando il modello di regressione

esempio ETEROSCHEDASTICITÀ



NOTA: $Var(\varepsilon | X=x) = \sigma^2 \Longrightarrow Var(Y | X=x) = \sigma^2$

(3) le distribuzioni condizionate $Y \mid X = x$ sono incorrelate

↓

ovvero le ε sono incorrelate

$$Cov(\varepsilon_i, \varepsilon_j) = 0$$

↳ più precisamente

$$Cov(\varepsilon \mid X = x_i, \varepsilon \mid X = x_j) = 0$$

Sotto le tre ipotesi del modello classico vale il teorema di Gauss-Markov:

$\hat{\beta}_0(b_0)$ e $\hat{\beta}_1(b_1)$ sono BLUE stimatori

- Linear
- Unbiased (non distorti)
- Best (efficienti in senso assoluto)

Se si è interessati solo alle stime puntuali di β_0 e β_1, il teorema di Gauss-Markov, se valgono le ipotesi:

$$E(\varepsilon|X)=0 \implies E(Y|X)=\beta_0+\beta_1 x$$

$$Var(\varepsilon|X)=\sigma^2 \implies Var(Y|X)=\sigma^2$$

$$Cov(\varepsilon_i,\varepsilon_j)=0$$

assicura che:

gli stimatori sono non distorti:
- $E(\hat{\beta}_0)=\beta_0$
- $E(\hat{\beta}_1)=\beta_1$

gli stimatori sono efficienti in senso assoluto:
- $Var(\hat{\beta}_0)=$ minima
- $Var(\hat{\beta}_1)=$ minima

Il teorema di Gauss-Markov non fornisce però alcuna informazione circa la distribuzione di $\hat{\beta}_0$ e $\hat{\beta}_1$:

$$\hat{\beta}_0 \sim (?) \qquad \hat{\beta}_1 \sim (?)$$

Senza informazioni su queste distribuzioni campionarie, non è possibile costruire stime intervallari o effettuare test di ipotesi

Per procedere, è necessario modificare le ipotesi di base

↳ modello classico normale

↓

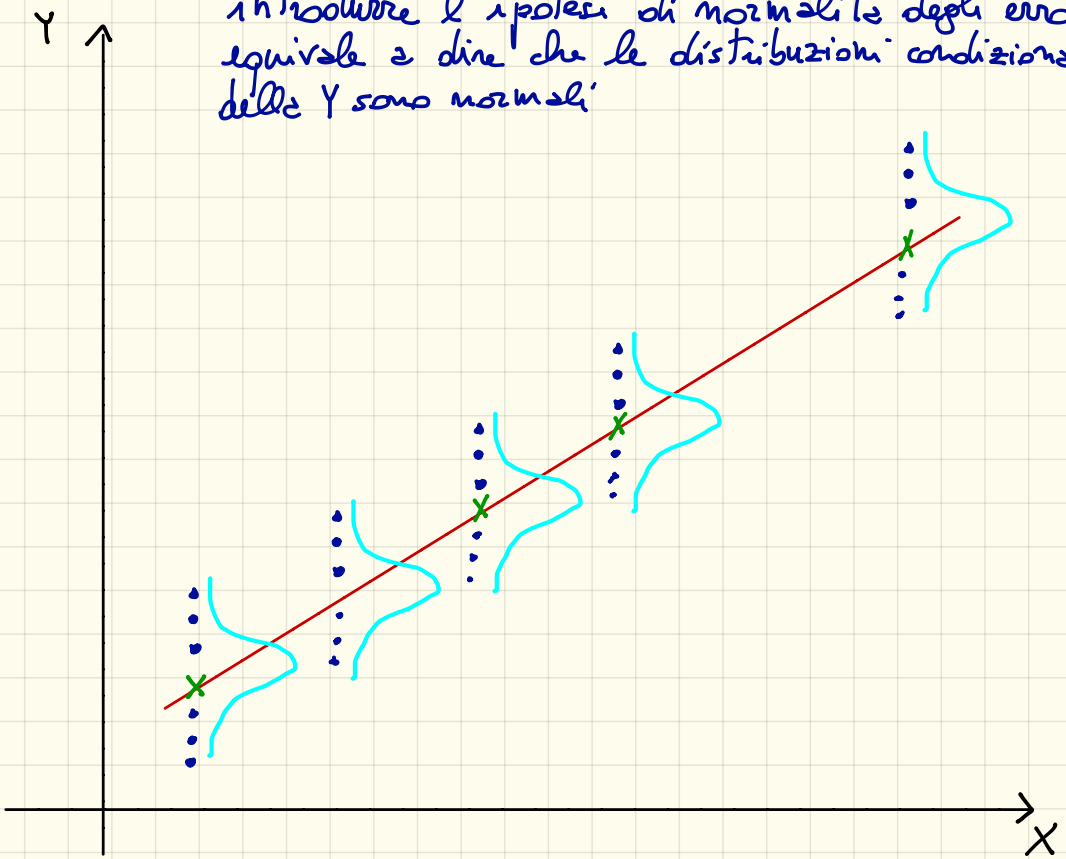
- si introduce l'ipotesi di normalità degli errori
- l'ipotesi di incorrelazione viene rafforzata in ipotesi di indipendenza

↓

sotto questo quadro aggiornato si ha che:

- $Y \sim N$
- $\hat{\beta}_0$ e $\hat{\beta}_1$, essendo combinazioni lineari delle Y_i, sono a loro volta normali

introdurre l'ipotesi di normalità degli errori equivale a dire che le distribuzioni condizionate delle Y sono normali



POPOLAZIONE ⇝ $Y = \beta_0 + \beta_1 X + \varepsilon$

componente deterministica

parametri
- β_0
- β_1

stimatori
- $\hat{\beta}_0$ (oppure b_0)
- $\hat{\beta}_1$ (oppure b_1)

$E(Y|X) = \beta_0 + \beta_1 X$

$Var(Y|X) = \sigma^2$

modello classico
- $E(\varepsilon|X) = 0$
- $Var(\varepsilon|X) = \sigma^2 = \sigma^2_\varepsilon$
- ε_i e ε_j sono incorrel. $i \neq j$

TEOREMA DI GAUSS MARKOV
gli stimatori dei minimi quadrati sono BLUE

non permette di ottenere gli intervalli di confidenza e i test per β_0 e β_1

modello classico normale
- $E(\varepsilon|X) = 0$
- $Var(\varepsilon|X) = \sigma^2 = \sigma^2_\varepsilon$
- incorrelazione viene rafforzata (ε_i e ε_j sono indipendenti)

NUOVA ⇝ $\varepsilon \sim N$

⇝ $\varepsilon_i \sim N(0, \sigma^2)$

con ε_i ed ε_j indipend. quando $i \neq j$

Sotto le ipotesi del modello classico normale si ha che $Y \sim N$

↓

è possibile mostrare che $\hat{\beta}_0$ e $\hat{\beta}_1$ sono combinazioni lineari delle Y_i

↳ essendo le $Y_i \sim N$ allora $\hat{\beta}_0$ e $\hat{\beta}_1$ saranno distribuite normalmente

Se consideriamo $\hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} = \dfrac{\sum (x_i - \bar{x})\, Y_i}{\sum (x_i - \bar{x})^2}$

$$\sum (x_i - \bar{x}) Y_i - \sum (x_i - \bar{x}) \bar{Y} = \sum (x_i - \bar{x}) Y_i - \bar{Y} \underbrace{\sum (x_i - \bar{x})}_{=0}$$

essendo le x_i deterministiche nel modello classico si vede come $\hat{\beta}_1$ è una combinazione lineare delle Y_i

Lo stesso avviene per $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y}_m - \hat{\beta}_1 \bar{x}_m = \sum \left[1 - \frac{(x_i - \bar{x}) \bar{x}}{\sum (x_i - \bar{x})^2} \right] y_i$$

↳ anche $\hat{\beta}_0$ è esprimibile come combinazione lineare delle Y

Per definire completamente $\hat{\beta}_0$ e $\hat{\beta}_1$ ci servono le corrispondenti varianze:

$$\sigma^2_{\hat{\beta}_1} = Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

σ^2 → varianza v.c. errore

$\sum (x_i - \bar{x})^2 = (n-1)\, s_x^2$

$$\sigma^2_{\hat{\beta}_0} = Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\, s_x^2} \right]$$

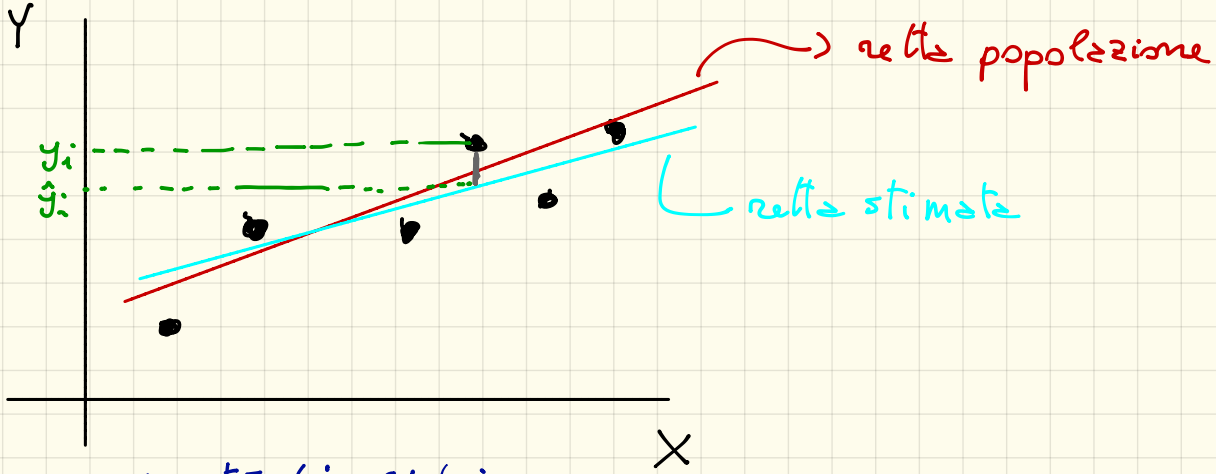
$$\hat{\beta}_0 \sim N\left(\beta_0 , \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right] \right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1 , \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \right)$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]}} \sim N(0,1)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}} \sim N(0,1)$$

Siccome σ^2 non è noto lo devo stimare a partire dai dati campionari: si utilizzano i residui ovvero $y_i - \hat{y}_i$ per stimare σ^2



NOTA: proprietà dei residui

Gli $e_i = y_i - \hat{y}_i$ sono tali che $\sum e_i = \sum (y_i - \hat{y}_i) = 0$

Per stimare σ^2 considero la distanza al quadrato:

$$\sum (y_i - \hat{y}_i)^2 = \text{Dev}(E)$$

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{Dev(E)}{n-2}$$

$\hat{\sigma}^2$ → stimatore per σ^2

$n-2$ → si perde un grado di libertà per ciascun parametro

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}} \sim t_{n-2}$$

ottenute le statistiche pivotali e le statistiche test

Invertendo le statistiche pivotali per β_0 e β_1 si hanno:

$$IC_{1-\alpha}(\beta_0) = \left[\hat{\beta}_0 \pm t_{n-2,\frac{\alpha}{2}} \underbrace{\sqrt{Var(\hat{\beta}_0)}}_{\text{errore standard}} \right] = \left[\hat{\beta}_0 \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]} \right]$$

$$IC_{1-\alpha}(\beta_1) = \left[\hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}} \sqrt{Var(\hat{\beta}_1)} \right] = \left[\hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}} \right]$$

Le due statistiche test sono:

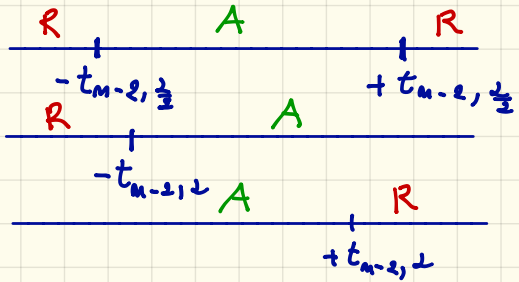
$H_0: \beta_0 = \text{valore}$

$H_1: \beta_0 \neq \text{valore}$

$<$

$>$

$$T_n = \frac{\hat{\beta}_0 - \beta_0 \text{ sotto } H_0}{\sqrt{Var(\hat{\beta}_0)}}$$



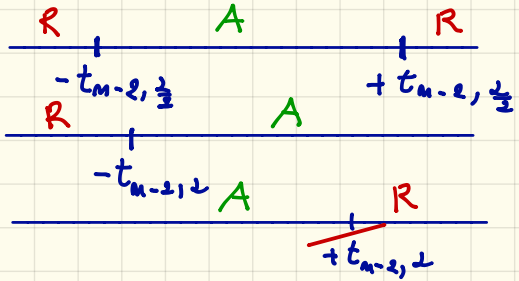
$H_0: \beta_1 = \text{valore}$

$H_1: \beta_1 \neq \text{valore}$

$<$

$>$

$$T_n = \frac{\hat{\beta}_1 - \beta_1 \text{ sotto } H_0}{\sqrt{Var(\hat{\beta}_1)}}$$



Di solito si utilizza il valore 0 come valore del parametro sotto H_0:

$$\left.\begin{array}{l} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \\ \qquad \lessgtr \end{array}\right] \quad T_m = \frac{\hat{\beta}_0 - 0}{\sqrt{Var(\hat{\beta}_0)}} = \frac{\hat{\beta}_0}{\sqrt{Var(\hat{\beta}_0)}}$$

↳ test significatività dell'intercetta

$$\left.\begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \\ \qquad \lessgtr \end{array}\right] \quad T_m = \frac{\hat{\beta}_1}{\sqrt{Var(\hat{\beta}_1)}}$$

↳ test significatività intercetta

↳ nel modello semplice equivale a testare la significatività del modello

FORMULE ALTERNATIVE CALCOLO R^2

$$R^2 = \frac{Dev(R)}{Dev(Y)} = 1 - \frac{Dev(E)}{Dev(Y)} = r^2_{xy}$$

r^2_{xy} → coefficiente di correlazione campionario

Si ha anche che:

$$R^2 = r^2_{xy} = \left[\frac{CODEV(X,Y)}{\sqrt{Dev(X)}\sqrt{Dev(Y)}}\right]^2 =$$

$$= \frac{codev(X,Y)^2}{Dev(X)\,Dev(Y)} \frac{Dev(X)}{Dev(X)} = \left[\frac{CODEV(X,Y)}{Dev(X)}\right]^2 \frac{Dev(X)}{Dev(Y)} =$$

$$= \hat{\beta}_1^2 \frac{Dev(X)}{Dev(Y)}$$

Dalle relazioni sopra esposte si ha quindi:

$$R^2 = \frac{Dev(R)}{Dev(Y)} = \hat{\rho}_{XY}^2 \implies Dev(R) = \rho_{XY}^2 \; Dev(Y)$$

$$R^2 = \frac{Dev(R)}{Dev(Y)} = \hat{\beta}_1^2 \frac{Dev(X)}{Dev(Y)} \implies Dev(R) = \hat{\beta}_1^2 \; Dev(X)$$

E' possibile cioè calcolare la $Dev(R)$, e quindi anche la $Dev(E) = Dev(Y) - Dev(R)$, a partire da $\hat{\rho}_{XY}$ e $Dev(Y)$ oppure a partire da $\hat{\beta}_1$ e $Dev(Y)$

↳ questo permette di evitare il calcolo di $\sum (y_i - \hat{y}_i)^2$ e $\sum_i (\hat{y}_i - \mu_Y)^2$ che può essere laborioso