Regressione lineare

Y, X

Y variabile risposta (variabile dipendente)
X variabile esplicativa o predittore (variabile indipendente)

$$Y = \beta_0 + \beta X + E$$

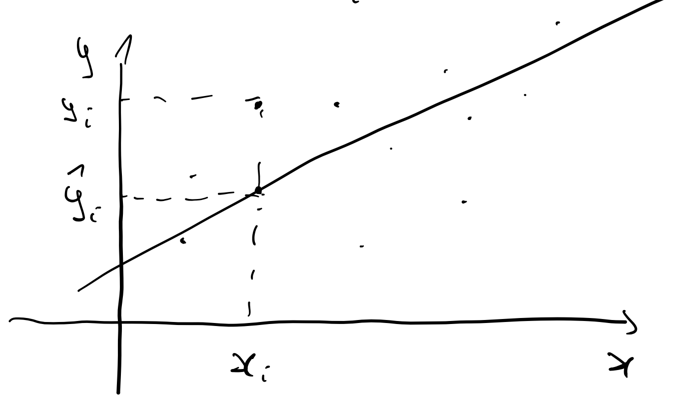
Vengono raccolte n osservazioni (x_i, y_i).

- Stimare i coefficienti β_0 e β i cui stimatori verranno indicati con $\hat{\beta}_0$, $\hat{\beta}$

- Metodo dei minimi quadrati

$$\boxed{y = \hat{\beta}_0 + \hat{\beta} x}$$

(x_i, y_i) $\quad y_i \longrightarrow \hat{y}_i$

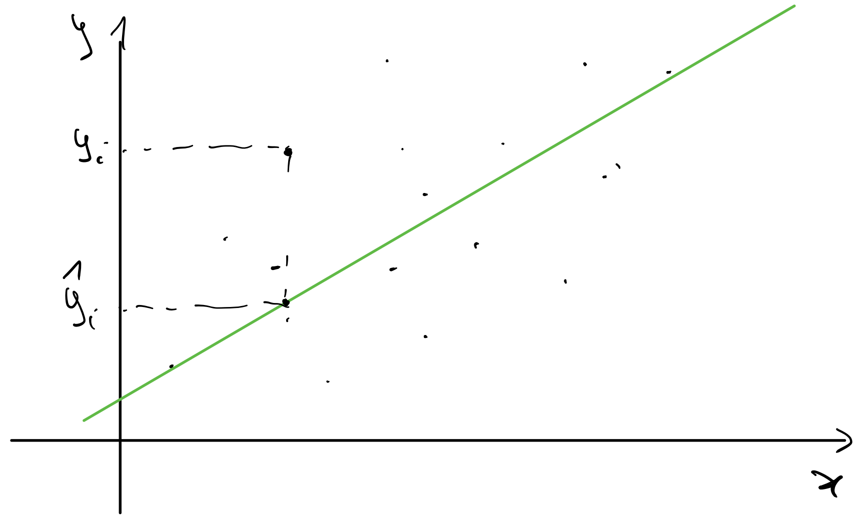


$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$y_i - \hat{y}_i = y_i - (\beta_0 + \beta x_i)$$

Si scelgono come stimatori quei particolari valori che minimizzano la somma di quadrati delle differenze

$$\sum_j \left(y_j - (\beta_0 + \beta x_j) \right)^2$$



$$f(\beta_0, \beta) = \sum_j \left(y_j - (\beta_0 + \beta x_j) \right)^2$$

Proposizione

La coppia di valori $(\hat{\beta}_0, \hat{\beta})$ che minimizza la funzione $f(\beta_0, \beta)$ è data da

$$\hat{\beta} = \frac{\sum_{i,j} (x_j - \bar{x})(y_j - \bar{y})}{\sum_i (x_j - \bar{x})^2}$$

$\bar{x}$

$\bar{y}$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}$$

<u>Dim.</u>

$$f(\beta_0, \beta) = \sum_j \left(y_j - (\beta_0 + \beta x_j)\right)^2$$

① $$\frac{\partial f}{\partial \beta_0} = -2 \sum_j \left(y_j - (\beta_0 + \beta x_j)\right) = 0$$

② $$\frac{\partial f}{\partial \beta} = -2 \sum_j x_j \left(y_j - (\beta_0 + \beta x_j)\right) = 0$$

① $\Rightarrow$ $$\sum_j y_j - \sum_j \beta_0 - \beta \sum_j x_j = 0$$

$$\bar{y} = \frac{1}{n} \sum_j y_j \qquad \bar{x} = \frac{1}{n} \sum_j x_j$$

$$n\bar{y} - n\beta_0 - n\beta\bar{x} = 0$$

$$\hat{\beta}_0 = \bar{y} - \beta\bar{x}$$

② $\Rightarrow$ $$\sum_j x_j y_j - \beta_0 \sum_j x_j - \beta \sum_j x_j^2 = 0$$

$$\sum_j x_j y_j - (\bar{y} - \beta\bar{x}) \sum_j x_j - \beta \sum_j x_j^2 = 0$$

$$\sum_j x_j y_j - \bar{y} \sum_j x_j + \beta \bar{x} \sum_j x_j - \beta \sum_j x_j^2 = 0$$

$$\sum_j (y_j - \bar{y}) x_j - \beta \sum_j (x_j - \bar{x}) x_j = 0$$

$$\beta = \frac{\sum_j (y_j - \bar{y}) x_j}{\sum_j (x_j - \bar{x}) x_j}$$

Oss. che $\sum_j \bar{x} (x_j - \bar{x}) = 0$

$$\sum_j \bar{x} x_j - \sum_j \bar{x}^2 = \bar{x} \sum_j x_j - n \bar{x}^2 = \bar{x} n \bar{x} - n \bar{x}^2 = 0$$

$$\sum_j \bar{x} (y_j - \bar{y}) \leftarrow$$

$$\beta = \frac{\sum_j (y_j - \bar{y}) x_j - \sum_j (y_j - \bar{y}) \bar{x}}{\sum_j (x_j - \bar{x}) x_j - \sum_j (x_j - \bar{x}) \bar{x}} =$$

$$= \frac{\sum_j (y_j - \bar{y})(x_j - \bar{x})}{\sum_j (x_j - \bar{x})^2} = \hat{\beta}$$

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta} x_j$$

$$r_j = y_j - \hat{y}_j = y_j - \left(\hat{\beta}_0 + \hat{\beta} x_j \right)$$

$$\hat{V} = \frac{1}{n-2} \sum_j \left(Y_j - \hat{Y}_j \right)^2$$

$$\hat{V} = \frac{1}{n-2} \sum_j \left(y_j - \hat{y}_j \right)^2 = \frac{1}{n-2} \sum_j r_j^2$$

Notazione

$$S_{xx} = \sum_j \left(x_j - \bar{x} \right)^2 = \sum_j x_j^2 - n\bar{x}$$

$$S_{yy} = \sum_j \left(y_j - \bar{y} \right)^2$$

$$S_{xy} = \sum_j \left(x_j - \bar{x} \right) \left(y_j - \bar{y} \right) = \sum_j x_j y_j - n\bar{x}\bar{y}$$

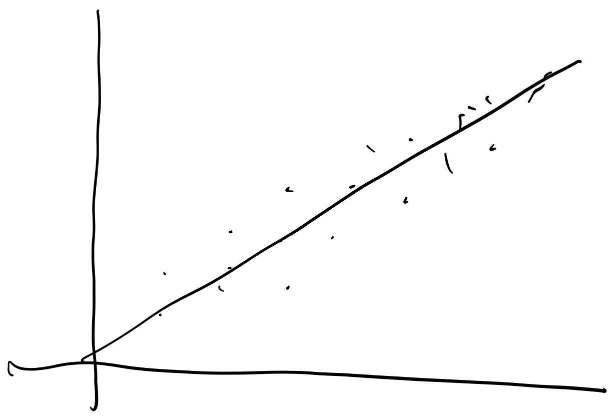
Definizioni

$$SS_{Reg} = \sum_j \left(\left(\hat{\beta}_0 + \beta x_j \right) - \bar{Y} \right)^2 \quad \text{varianza spiegata}$$

$$SS_{Res} = \sum_j \left(Y_j - \left(\hat{\beta}_0 + \hat{\beta} x_j \right) \right)^2 \quad \text{varianza residua}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{Y} - \beta \bar{X} \qquad \hat{V} = \frac{SS_{Res}}{n-2}$$

$$\rightarrow \underline{SS_{Reg} = \frac{(S_{xy})^2}{S_{xx}}} \qquad S_{xx} = \frac{(S_{xy})^2}{S_{xx}} \; ?$$



Oss.

Le medie campionarie per y_j e $\hat{y}_j$ coincidono

$$\sum_j \hat{y}_j = \sum_j \left(\hat{\beta}_0 + \hat{\beta} x_j\right) = \sum_j \left(\bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_j\right) =$$

$$= \sum_j \bar{y} = n\bar{y} \qquad \bar{y} = \frac{1}{n}\sum_j y_j$$

Varianza per y

$$S_y^2 = \frac{1}{n-1}\sum_j (y_j - \bar{y}) = \frac{1}{n-1} S_{yy}$$

$$S_{\hat{y}}^2 = \frac{1}{n-1}\sum_j \left((\hat{\beta}_0 + \hat{\beta} x_j) - \bar{y} \right) = \frac{1}{n-1} SS_{Reg}$$

Teorema

La varianza totale è la somma della varianza spiegata più la varianza residua

$$S_{yy} = SS_{Reg} + SS_{Res}$$

Dim. ... — .

$$SS_{Res} = S_{yy} - SS_{Reg} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = \frac{S_{xx}S_{yy} - (S_{xy})^2}{S_{xx}}$$

Coefficiente di determinazione

$$R^2 = \frac{SS_{Reg}}{SS_{yy}} \simeq 1$$

$$R^2 = \frac{S_{yy} - SS_{Res}}{S_{yy}} = 1 - \frac{SS_{Res}}{S_{yy}} \simeq 0$$

Si dimostra che

$$R^2 = \left(\frac{S_{xx}}{\sqrt{S_{xx} S_{yy}}} \right)^2$$

$$r = \hat{\beta} \sqrt{\frac{S_{xx}}{S_{yy}}}$$

Coefficiente di correlazione corretto

$$R^2_{corr} = 1 - \frac{SS_{Res}/(n-2)}{S_{yy}/(n-1)}$$

Test relativi ai parametri

Ipotesi di normalità $\qquad Y = \beta_0 + \beta X + E$

$$E \sim Nor\left(\mu = 0,\ \sigma^2 = v\right)$$

Proprietà

$$\hat{\beta} \sim Nor\left(\mu = \beta,\ \sigma^2 = \frac{v}{S_{xx}}\right)$$

$$\hat{\beta}_0 \sim Nor\left(\mu = \beta_0 \,,\ \sigma^2 = v\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

Si ha

$$\frac{(n-2)}{v}\hat{v} = \frac{1}{v}\sum_j \left(Y_j - \hat{Y}_j\right) \sim \chi^2(\nu = n-2)$$

Test su β

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{v}{S_{xx}}}} \sim Nor\left(\mu = 0 \,,\ \sigma^2 = 1\right)$$

v è incognita e consideriamo la stima $\hat{v}$

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{\hat{v}}{S_{xx}}}} \sim T(\nu = n-2)$$

Test d'ipotesi

$$\text{Tipo I} : \begin{cases} H_0 : & \beta = \beta^* \\ H_1 : & \beta \neq \beta^* \end{cases}$$

tipo II : $\begin{cases} H_0 : \beta = \beta^* & \text{opp. } \beta \leq \beta^* \\ H_1 : \beta > \beta^* \end{cases}$

tipo III : $\begin{cases} H_0 : \beta = \beta^* & \text{opp. } \beta \geq \beta^* \\ H_1 : \beta < \beta^* \end{cases}$

Statistica test

$$U = \frac{\hat{\beta} - \beta^*}{\sqrt{\frac{\hat{V}}{S_{xx}}}} \overset{H_0}{\sim} T \, (\nu = n-2)$$

$$\frac{\hat{\tilde{\beta}}}{\sqrt{\frac{\hat{V}}{S_{xx}}}} \qquad \hat{\tilde{\beta}} = \hat{\beta} - \beta^*$$

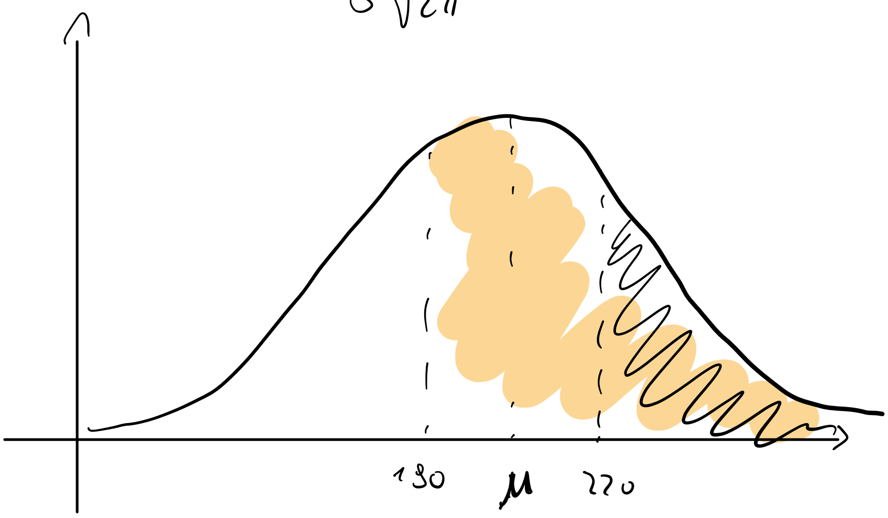
Problema 1

Dati:

$\mu = 200\,g$

$\sigma^2 = 10\,g$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



1) $\displaystyle\int_{220}^{+\infty} f(x)\,dx$ $\qquad$ $f = @\ \text{normpdf}(x, \mu, \sigma)$

2) $\displaystyle\int_{190}^{+\infty} f(x)\,dx$

3)