REGRESSIONE LINEARE

Note Title 04/12/2017

$$D_n := \sup_{t\in\mathbb{R}} \left| \frac{\#\{i : F_0(X_i) \le F_0(t)\}}{n} - F_0(t) \right|$$

$$\mathbb{P}(D_n \ge d) \overset{\uparrow}{=} \mathbb{P}\left(\sup_{y\in(0,1)} \left| \frac{\#\{i : U_i \le y\}}{n} - y \right| \ge d \right)$$

dove $U_1 \dots U_n$ sono v.a. i.i.d con $\mathbb{P}_{U_i} = U(0,1)$

purché $F_0 : \mathbb{R} \to [0,1]$ sia continua.

X v.a. con legge F $\Rightarrow$ $U = F_0 X$ ha distribuzione $U((0,1))$

$\mathbb{P}_X = f(x)\,dx$ $\quad F(t) = \int_{-\infty}^{t} f(x)\,dx$

$\psi : \mathbb{R} \to \mathbb{R}$ di Borel non negativa

$$\int_{\mathbb{R}} \psi(u)\, \mathbb{P}_U(du) = \int_{\mathbb{R}} \psi(u)\, \mathbb{P}_{F_0 X}(du) = \int_{\mathbb{R}} \psi(F(x))\, \mathbb{P}_X(dx)$$

$$= \int_{\mathbb{R}} \psi(F(x))\, f(x)\, dx$$

rispetto alla misura di Lebesgue

$u = F(x)$ $\quad du = F'(x)\,dx$ $\Rightarrow$ per p.o. $x \in \mathbb{R}$

$F'(x) = f(x)$ $\Rightarrow$ $du = f(x)\,dx$

$x \to -\infty$ $\quad u \to 0$

$x \to +\infty$ $\quad u \to 1$

$$\Rightarrow \int_{\mathbb{R}} \psi(u)\, \mathbb{P}_U(du) = \int_0^1 \psi(u)\, du = \int_{\mathbb{R}} \psi(u)\, \mathbb{1}_{(0,1)}(u)\, du$$

$$\lim_{n\to\infty} \mathbb{P}(D_n\sqrt{n} \le t) = \begin{cases} 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2t^2) & t > 0 \\ 0 & t \le 0 \end{cases}$$

Rifiuto H_0 se $d_n > \varepsilon$

$$\alpha = \mathbb{P}(D_n > \varepsilon) = \mathbb{P}(D_n\sqrt{n} > \varepsilon\sqrt{n}) = 1 - \mathbb{P}(D_n\sqrt{n} \le \varepsilon\sqrt{n})$$

$$= 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp\left(-2j^2\varepsilon^2 n\right) =$$

$$= 2\exp\left(-2\varepsilon^2 n\right) + 2\sum_{j=2}^{\infty}(-1)^{j-1}\exp\left(-2j^2\varepsilon^2 n\right)$$

Serie a segni alterni

$$\alpha = 2\exp\left(-2\varepsilon^2 n\right)$$

$$\exp\left(-2\varepsilon^2 n\right) = \frac{\alpha}{2} \qquad -2\varepsilon^2 n = \log\frac{\alpha}{2} \qquad \alpha < \frac{1}{2}$$

$$\varepsilon^2 = \frac{-1}{2n}\log\left(\frac{\alpha}{2}\right) = \frac{1}{2n}\log\left(\frac{2}{\alpha}\right)$$

$$\varepsilon = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$$

——— ○ ———

RETTA DI REGRESSIONE

Suppongo di fare un esperimento in cui si può controllare direttamente un dato in ingresso che indico con x

x_1 ——— x_n

Ad ogni prova ho un risultato y_1 — y_n

(y_i la risposta in corrispondenza dell'input x_i)

$y_i = ax_i + b$ $\qquad$ $y_i = ax_i + b + \varepsilon_i$

Vedo la risposta dell'esperimento come una v.a. Y_i di valore atteso $ax_i + b$

$$\forall i = 1 \text{ — } n \qquad \sum_{i=1}^{n}\left|y_i - (ax_i + b)\right|^2 \longrightarrow \min$$

$$\sum_{i=1}^{n}\left(Y_i - (Ax_i + B)\right)^2 \longrightarrow \min$$

$$A = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = {}^{2} \quad (Y_1 - Y_n, x_n - x_n)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$B = \bar{Y} - A\bar{x} = b(Y_n - Y_n, x_n - x_n)$$

$$Y - \bar{Y} = A(x - \bar{x})$$ si dice STIMA DELLA REGRESSIONE

$$A = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sum_{i=1}^{n} \left(x_i Y_i - \bar{x} Y_i - x_i \bar{Y} + \bar{x} \bar{Y} \right) =$$

$$= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \left(\sum_{i=1}^{n} x_i Y_i - \bar{x} \sum_{i=1}^{n} Y_i - n\bar{x}\bar{Y} + n\bar{x}\bar{Y} \right)$$

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

$$B = \bar{Y} - A\bar{x} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{1}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$$

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right) Y_i$$

$\mathbb{E}[Y_i] = ax_i + b$ $\quad Y_1 \dots Y_n$ indipendenti

In generale supponiamo $\mathbb{P}_{Y_i} = N(ax_i + b, \sigma^2)$

$$\mathbb{P}_{Y_i - ax_i + b} = N(0, \sigma^2)$$

$$A = \frac{1}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i \qquad B = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum (x_j - \bar{x})^2} \right) Y_i$$

Poiché $Y_1 \dots Y_n$ sono indipendenti e gaussiane lo sono anche A e B sono anche gaussiane.

$$E[A] = \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) E[Y_i] =$$

$$= \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) \underbrace{(ax_i + b - a\bar{x} + a\bar{x})}_{a(x_i - \bar{x}) + a\bar{x} + b}$$

$$= \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \sum_{i=1}^{n} \left(a(x_i - \bar{x})^2 + (a\bar{x} + b)(x_i - \bar{x}) \right) =$$

$$= \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \left(a \sum_{i=1}^{n} (x_i - \bar{x})^2 + (a\bar{x} + b) \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{= 0} \right)$$

$$\sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x}$$

$$= a$$

$$Var[A] = Var\left[\frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \sum_{i=1}^{n} \underbrace{(x_i - \bar{x}) Y_i}_{} \right] =$$

$$= \frac{1}{\left(\sum_{j=1}^{n}(x_j - \bar{x})^2 \right)^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \underbrace{Var[Y_i]}_{= \sigma^2 \ \forall i = 1 \dots n}$$

$$= \frac{1}{\left(\sum_{j=1}^{n}(x_j - \bar{x})^2 \right)^2} \sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$B = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right) Y_i \qquad E[Y_i] = ax_i + b$$

$$E[B] = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right) \underbrace{E[Y_i]}_{(ax_i + b)} =$$

$$= \sum_{i=1}^{n} \frac{1}{n}(ax_i + b) - \frac{\bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) \left(a(x_i - \bar{x}) + a\bar{x} + b \right)$$

$$= 2\bar{x} + b - \frac{\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\left(2\sum_{i=1}^{n}(x_i-\bar{x})^2 + (2\bar{x}+b)\sum_{i=1}^{n}(x_i-\bar{x})\right)$$

$n\bar{x} - n\bar{x} = 0$

$$= 2\bar{x} + b - 2\bar{x} = b$$

$$Var[B] = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{\sum_j (x_j-\bar{x})^2}\right)^2 \underbrace{Var[Y_i]}_{=\sigma^2 \ \forall i=1\dots n}$$

$$= \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n^2} + \frac{\bar{x}^2(x_i-\bar{x})^2}{\left(\sum_{j=1}^{n}(x_j-\bar{x})^2\right)^2} - \frac{2\bar{x}(x_i-\bar{x})}{n\sum_{j=1}^{n}(x_j-\bar{x})^2}\right)$$

$n\bar{x} - n\bar{x} = 0$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\left(\sum(x_j-\bar{x})^2\right)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 - \frac{2\bar{x}\sum_{i=1}^{n}(x_i-\bar{x})}{n\sum_{j=1}^{n}(x_j-\bar{x})^2}\right)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right) =$$

$$= \sigma^2 \frac{n\bar{x}^2 + \sum_{j=1}^{n}\left(x_j^2 - 2\bar{x}x_j + \bar{x}^2\right)}{n\sum_{j=1}^{n}(x_j-\bar{x})^2} =$$

$$= \frac{\sigma^2\left(n\bar{x}^2 + \sum_{j=1}^{n}x_j^2 - 2\bar{x}\,n\bar{x} + n\bar{x}^2\right)}{n\sum_{j=1}^{n}(x_j-\bar{x})^2} = \frac{\sigma^2\sum_{j=1}^{n}x_j^2}{n\sum_{j=1}^{n}(x_j-\bar{x})^2}$$

$$\sum_{i=1}^{n}\left|Y_i - Ax_i + B\right|^2 = S_R \quad \text{residuo}$$

$\frac{S_R}{\sigma^2}$ ha distribuzione χ^2_{n-2}

A, B e S_R sono v.a. indipendenti

$$E\left[\frac{S_R}{n-2}\right] = E\left[\frac{S_R}{\sigma^2}\frac{\sigma^2}{n-2}\right] = \frac{\sigma^2}{n-2}E\left[\frac{S_R}{\sigma^2}\right] =$$

$$= \frac{\sigma^2}{n-2}(n-2) = \sigma^2$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = S_{xx}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = S_{xY}$$

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

$$A = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xY}}{S_{xx}}$$

$$B = \bar{Y} - A\bar{x}$$

$$S_R = \sum_{i=1}^{n} (Y_i - Ax_i - B)^2 = \sum_{i=1}^{n} (Y_i - Ax_i - \bar{Y} + A\bar{x})^2$$

$$= \sum_{i=1}^{n} \left((Y_i - \bar{Y}) - A(x_i - \bar{x}) \right)^2 = \sum_{i=1}^{n} \left(Y_i - \bar{Y} - \frac{S_{xY}}{S_{xx}} (x_i - \bar{x}) \right)^2$$

$$S_R = \sum_{i=1}^{n} \left((Y_i - \bar{Y})^2 + \frac{S_{xY}^2}{S_{xx}^2} (x_i - \bar{x})^2 - \frac{2 S_{xY}}{S_{xx}} (x_i - \bar{x})(Y_i - \bar{Y}) \right)$$

$$= S_{YY} + \frac{S_{xY}^2}{S_{xx}^2} S_{xx} - \frac{2 S_{xY}}{S_{xx}} S_{xY} =$$

$$= S_{YY} - \frac{S_{xY}^2}{S_{xx}} = \frac{S_{xx} S_{YY} - S_{xY}^2}{S_{xx}}$$

— o —

$H_0 : \quad a = \bar{a}$ $\qquad$ $H_A : \quad a \neq \bar{a}$

$$P_A = N\left(a, \frac{\sigma^2}{S_{xx}}\right) \qquad P \frac{S_R}{\sigma^2} = \chi^2_{n-2}$$

$$Z = \frac{A - a}{\sigma / \sqrt{S_{xx}}} = \frac{(A - a)\sqrt{S_{xx}}}{\sigma} \qquad P_Z = N(0,1)$$

$$T = \frac{Z \sqrt{n-2}}{\sqrt{S_R / \sigma^2}} = \frac{Z \sqrt{n-2}}{\sqrt{S_R}} \sigma =$$

$$T = \frac{(A-\alpha)\sqrt{S_{xx}}}{\sigma} \cdot \frac{\sqrt{n-2}\ \sigma}{\sqrt{S_R}} \qquad P_T = t(n-2)$$

So che $E[A] = \alpha = \bar{\alpha}$ SSE $A - \bar{\alpha}$ ha valore atteso nullo SSE $E[T] = 0$

A e S_R sono indipendenti $= 0$

$$E[T] = E\left[\frac{(A-\alpha)\sqrt{S_{xx}}\sqrt{n-2}}{\sqrt{S_R}}\right] =$$

$$\sqrt{S_{xx}}\sqrt{n-2}\ E\left[(A-\alpha)\, S_R^{-1/2}\right] =$$

$$= \sqrt{S_{xx}}\sqrt{n-2}\ E[A-\alpha]\, E[S_R^{-1/2}] = 0 \quad \text{SSE} \quad E[A-\alpha] = 0$$

$E[S_R^{-1/2}] > 0$

$E[A] = \alpha$

Quello Ho SSE

$$\left| \frac{A(x_1 - x_n, y_1 - y_n) - \bar{\alpha}\sqrt{S_{xx}}\sqrt{n-2}}{\sqrt{S_R(x_1 - x_n, y_1 - y_n)}} \right| < \varepsilon$$

$$\alpha = P\left(\frac{|A-\bar{\alpha}|\sqrt{S_{xx}}\sqrt{n-2}}{\sqrt{S_R}} > \varepsilon \;\middle|\; E[A] = \bar{\alpha} \right) = P(|T_{n-2}| > \varepsilon)$$

$$= P(T_{n-2} > \varepsilon) + P(T_{n-2} < -\varepsilon) =$$

$$= 1 - F_{T_{n-2}}(\varepsilon) + F_{T_{n-2}}(-\varepsilon) = 2\left(1 - F_{T_{n-2}}(\varepsilon)\right)$$

$$F_{T_{n-2}}(\varepsilon) = 1 - \frac{\alpha}{2} \qquad \varepsilon = t_{n-2,\,1-\frac{\alpha}{2}}$$

INTERVALLO DI CONFIDENZA PER IL PARAMETRO α

$\frac{(A-\alpha)\sqrt{n-2}\sqrt{S_{xx}}}{\sqrt{S_R}}$ ha distribuzione $t(n-2)$

$$1-\alpha = P\left(\frac{|A-\alpha|\sqrt{n-2}\sqrt{S_{xx}}}{\sqrt{S_R}} < t_{n-2,1-\frac{\alpha}{2}} \right) =$$

$$= P\left(|A-\alpha| < \frac{\sqrt{S_R}\, t_{n-2,1-\frac{\alpha}{2}}}{\sqrt{n-2}\sqrt{S_{xx}}} \right) =$$

$$= P\left(A - \frac{\sqrt{S_R}\, t_{n-2,1-\frac{\alpha}{2}}}{\sqrt{(n-2)S_{xx}}} < \alpha < A + \frac{\sqrt{S_R}\, t_{n-2,1-\frac{\alpha}{2}}}{\sqrt{(n-2)S_{xx}}} \right)$$

— o —

H_0 $b = E[B] = \bar{b}$ $\quad H_A : b = E[B] \neq \bar{b}$

$$P_B = N\left(b, \frac{\sigma^2 \sum x_i^2}{n S_{xx}} \right)$$

$$Z = \frac{(B-b)\sqrt{S_{xx}\, n}}{\sigma\sqrt{\sum x_i^2}}$$ ha distribuzione $N(0,1)$

$\frac{S_R}{\sigma^2}$ ha distribuzione χ^2_{n-2}

$$T := \frac{Z\sqrt{n-2}}{\sqrt{S_R/\sigma^2}} = \frac{(B-b)\sqrt{S_{xx}\, n(n-2)}}{\sqrt{\sum x_i^2}\sqrt{S_R}}$$ ha distribuzione $t(n-2)$

$$E\left[\frac{(B-\bar{b})\sqrt{S_{xx}\, n(n-2)}}{\sqrt{\sum x_i^2}\sqrt{S_R}} \right] = 0 \quad \text{sse} \quad \bar{b} = E[B]$$

Accetto H_0 se $\frac{|B-\bar{b}|\sqrt{S_{xx}\, n(n-2)}}{\sqrt{\sum x_i^2}\sqrt{S_R}} < \varepsilon$

$$\alpha = \mathbb{P}\left(\frac{|B-\bar{b}|\sqrt{S_{xx}\,n(n-2)}}{\sqrt{\Sigma x_i^2}\sqrt{S_R}} > \varepsilon \;\middle|\; \bar{b} = \mathbb{E}[B] \right)$$

$$= \mathbb{P}\left(|T_{n-2}| > \varepsilon \right) = 2\left(1 - F_{|T_{n-2}|}(\varepsilon)\right)$$

$\Rightarrow$ otengo livello di significatività α scegliendo $\varepsilon = t_{n-2,\,1-\frac{\alpha}{2}}$

$$\left(B - \frac{\sqrt{\Sigma x_i^2}\sqrt{S_R}}{\sqrt{S_{xx}\,n(n-2)}}\, t_{n-2,\,1-\frac{\alpha}{2}} \;,\; B + \frac{\sqrt{\Sigma x_i^2}\sqrt{S_R}}{\sqrt{S_{xx}\,n(n-2)}}\, t_{n-2,\,1-\frac{\alpha}{2}} \right)$$

— o —

x_1 —— x_n

x_0 $\qquad \mathbb{E}[Ax_0+B] = x_0\,\mathbb{E}[A] + \mathbb{E}[B] = ax_0 + b$

$$\hat{A}x_0 + B = Ax_0 + \bar{Y} - A\bar{x} = \bar{Y} + A(x_0 - \bar{x}) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i + (x_0 - \bar{x})\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i$$

$$= \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_0-\bar{x})(x_i-\bar{x})}{S_{xx}} \right) Y_i$$

$$\mathrm{Var}[Ax_0+B] = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_0-\bar{x})(x_i-\bar{x})}{S_{xx}} \right)^2 \underbrace{\mathrm{Var}[Y_i]}_{=\sigma^2} =$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \frac{(x_0-\bar{x})^2(x_i-\bar{x})^2}{S_{xx}^2} + \frac{2(x_0-\bar{x})}{n\,S_{xx}}(x_i-\bar{x}) \right) =$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}^2} S_{xx} + \frac{2(x_0-\bar{x})}{n\,S_{xx}} \underbrace{\sum_{i=1}^{n}(x_i-\bar{x})}_{n\bar{x} - n\bar{x} = 0} \right)$$

$$\mathrm{Var}[Ax_0+B] = \sigma^2 \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}} \right)$$

$$Z = \frac{(Ax_0+B) - (\alpha x_0+b)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \qquad P_Z = N(0,1)$$

$\frac{S_R}{\sigma^2}$ ha distribuzione χ^2_{n-2}

A, B e S_R sono indipendenti $\Rightarrow$ Z e $\frac{S_R}{\sigma^2}$ sono indipendenti.

$\Rightarrow$ $\frac{Z\sqrt{n-2}}{\sqrt{S_R/\sigma^2}}$ ha distribuzione $t(n-2)$ cioè

$$\frac{(Ax_0+B) - (\alpha x_0+b)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \cdot \frac{\sqrt{n-2}\,\sigma}{\sqrt{S_R}}$$ ha distribuzione $t(n-2)$

$$P\left(\frac{|(Ax_0+B) - (\alpha x_0+b)|}{\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \cdot \frac{\sqrt{n-2}}{\sqrt{S_R}} < t_{n-2,1-\frac{\alpha}{2}} \right) = 1-\alpha$$

$$Ax_0+B \pm t_{n-2,1-\frac{\alpha}{2}} \frac{\sqrt{S_R}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}}{\sqrt{n-2}}$$