Statistical Models

Kevin Michael Frick

3 gennaio 2020

1 Formulario

1.1 Stima di valori attesi e intervalli di predizione

$$E[\hat{y}_h] = X^T\beta \tag{1}$$

$$Var[\hat{y}_h] = \sigma^2 x_h^T (X^T X)^{-1} x_h \tag{2}$$

$$s^2[\hat{y}_h] = MSE \cdot x_h^T (X^T X)^{-1} x_h \tag{3}$$

$$\epsilon_i \sim N \implies \hat{y}_h \sim N \tag{4}$$

$$\epsilon_i \sim N \implies \frac{\hat{y}_h - x_h^T\beta}{\sqrt{s^2[\hat{y}_h]}} \sim t(n-p) \tag{5}$$

$$E[y_{h(new)} - \hat{y}_h] = 0 \tag{6}$$

$$Var[y_{h(new)} - \hat{y}_h] = \sigma^2(1 + x_h^T (X^T X)^{-1} x_h) \tag{7}$$

$$s^2[y_{h(new)} - \hat{y}_h] = MSE \cdot (1 + x_h^T (X^T X)^{-1} x_h) \tag{8}$$

$$\frac{y_{h(new)} - \hat{y}_h}{\sqrt{s^2[y_{h(new)} - \hat{y}_h]}} \sim t(n-p) \tag{9}$$

1.2 Variabili standardizzate

$$y_{is} = \frac{y_i - \bar{y}}{\sqrt{s_y^2}} \tag{10}$$

$$x_{ks} = \frac{x_k - \bar{x}_k}{\sqrt{s_k^2}} \tag{11}$$

$$X_{RS} = X_{RC} D_R^{-1} \tag{12}$$

$$D_R = \text{diag}_k\{\sqrt{s_k^2}\} \tag{13}$$

$$b_{0s} = 0 \tag{14}$$

$$R_{xx} = D_R^{-1} S_{xx} D_R^{-1} \tag{15}$$

$$y_s = \frac{M_0 y}{\sqrt{s_y^2}} \tag{16}$$

$$r_{yx} = \frac{D_R^{-1} s_{yx}}{\sqrt{s_y^2}} \tag{17}$$

$$b_{Rs} = \frac{D_R b_R}{\sqrt{s_y^2}} = R_{xx}^{-1} r_{yx} \tag{18}$$

$$b_{kR} = \sqrt{\frac{s_y^2}{s_k^2}} b_{ks} \tag{19}$$

1.3 Diagnostiche di regressione

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \tag{20}$$

$$e \sim MVN(0, \sigma^2 M) \tag{21}$$

$$r_i = \frac{e_i}{\sqrt{MSE \cdot (1 - h_{ii})}} \tag{22}$$

$$d_i = y_i - \hat{y}_{i(i)} \tag{23}$$

$$t_i = \frac{d_i}{\sqrt{s^2[d_i]}} = e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}} \tag{24}$$

$$y_i \text{ non outlier} \implies t_i \sim t(n-1-p) \tag{25}$$

$$\Pr[\text{Err. t. 1 su famiglia}] \approx (1-\alpha)^n \tag{26}$$

$$\Pr[\text{Err. t. 1 su famiglia (corr. Bonf.)}] = 1-(1-\frac{\alpha}{n})^2 \approx \alpha \tag{27}$$

$$\sum_i h_{ii} = p \tag{28}$$

$$\bar{h} = \frac{p}{n} \tag{29}$$

$$n \gg 3p \implies h_{ii} \geq 2p/n \text{ (o } 3p/n) \text{ valore grande} \tag{30}$$

$$D_i = \frac{\sum_{j-1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot MSE} \tag{31}$$

$$D_i = \frac{r_i^2 h_{ii}}{p(1-h_{ii})} \tag{32}$$

$$D_i > F(1/2; n, p) \implies y_i \text{ influente} \tag{33}$$

1.4 Regressori categorici

$$b_0 = \bar{y}_{reg} \tag{34}$$

$$b_k = \bar{y}_{k.} - \bar{y}_{reg} \tag{35}$$

1.5 Selezione del modello

$$R_a^2 = 1 - (\frac{SSE}{SSTOT} \cdot \frac{n-1}{n-p}) = 1 - \frac{MSE}{s_y^2} \tag{36}$$

$$AIC = n\log(\frac{SSE}{n}) + 2p \tag{37}$$

$$SBC = n\log(\frac{SSE}{n}) + p\log(n) \tag{38}$$

1.6 Regressione logistica semplice

$$L(\beta_0, \beta_1) = \prod_i (\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}})^{y_i} (1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}})^{1-y_i} \tag{39}$$

$$l(\beta_0, \beta_1) = \log L(\beta_0, \beta_1) \tag{40}$$

$$\hat{\pi}_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \tag{41}$$

1.7 Regressione logistica multipla

$$E[b] \longrightarrow \beta \tag{42}$$

$$Var[b] \longrightarrow -E^{-1}[\frac{\partial^2}{\partial\beta\partial\beta^T}\log L(\beta)] \tag{43}$$

$$b \sim MVN_p \tag{44}$$

$$s[b] = (-G)^{-1} \tag{45}$$

$$G = \frac{\partial^2}{\partial\beta\partial\beta^T}\log L(\beta)|_{\beta=b} \tag{46}$$

$$b_k \in_\alpha \beta_k \pm Z(1-\alpha/2)\sqrt{s^2[b_k]} \tag{47}$$

$$\frac{b_k - \beta_k}{\sqrt{s^2[b_k]}} \longrightarrow N(0,1) \tag{48}$$

$$G^2 = 2\log\frac{L(F)}{L(R)} \geq 0 \approx \chi^2(p-q)|H_0 \tag{49}$$

$$G^2 \approx N^2(0,1) = \frac{b_k^2}{s^2[b_k]} \tag{50}$$

2 Ipotesi

2.1 Approccio di regressione all'analisi della varianza

1. $y_{ij} = \beta_0 + \sum_k \beta_k x_{ijk} + \epsilon_{ij}$;
2. x_{ijk} fissi $\implies n$ fissi;
3. $\epsilon_{ij} \sim N(0,\sigma)$ iid.

2.2 Regressione logistica semplice

1. $y_i \sim Be(\pi_i)$ iid;
2. $\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}; \pi_i' = \log(\frac{\pi_i}{1-\pi_i})$;
3. x_i fissi.

2.3 Regressione logistica multipla

1. $y_i \sim Be(\pi_i)$ iid;
2. $\pi_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}; \pi_i' = \log(\frac{\pi_i}{1-\pi_i})$;
3. x_i fissi.

Disclaimer: Questo documento pu contenere errori e imprecisioni che potrebbero danneggiare sistemi informatici, terminare relazioni e rapporti di lavoro, liberare le vesciche dei gatti sulla moquette e causare un conflitto termonucleare globale. Procedere con cautela.