Metodi Statistici per la Ricerca Sociale: Formulario

Attenzione. Il Formulario contiene una selezione delle formule.

Le formule non riportate sono supposte note

Associazione tra variabili categoriche

Misure di associazione in tabelle 2×2

Inferenza per l'odds-ratio

Intervallo di confidenza al livello di confidenza $1-\alpha$ per l'odds ratio

$$IC_{1-\alpha}(OR) = \left[\exp\left\{ \log(\widehat{OR}) - z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \right\}; \right.$$
$$\left. \exp\left\{ \log(\widehat{OR}) + z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \right\} \right]$$

Correlazione e Regressione lineare semplice

Inferenza per il coefficiente di correlazione

- Statistica test

$$T = \frac{r_{XY}}{\sqrt{(1 - r_{XY}^2)/(n-2)}}$$

Il modello di regressione lineare semplice

Inferenza nel modello di regressione lineare semplice

- **Stimatori degli errori standard degli stimatori dei minimi quadrati**

$$\widehat{e.s.}\left(\widehat{B}_1\right) = \sqrt{\widehat{Var}\left(\widehat{B}_1\right)} = \sqrt{\frac{s^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$$
$$\widehat{e.s.}\left(\widehat{B}_0\right) = \sqrt{\widehat{Var}\left(\widehat{B}_0\right)} = \sqrt{s^2 \cdot \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}$$

- **Intervalli di confidenza per i coefficienti di regressione**

$$IC_{1-\alpha}(\beta_0) = \widehat{\beta}_0 \pm t_{n-2}(\alpha/2) \cdot \widehat{e.s.}(\widehat{B}_0)$$
$$IC_{1-\alpha}(\beta_1) = \widehat{\beta}_1 \pm t_{n-2}(\alpha/2) \cdot \widehat{e.s.}(\widehat{B}_1)$$

- **Verifica di ipotesi per β_1**
  - Statistica Test

$$T = \frac{\widehat{B}_1 - 0}{\widehat{e.s.}(\widehat{B}_1)}$$

Analogo il procedimento per sottoporre a test ipotesi sul coefficiente β_0.

- **Inferenza per la risposta media**
  - Stimatore dell'errore standard dello stimatore di $\widehat{Y}_{x^*}$

$$\widehat{e.s.}\left(\widehat{Y}_{x^*}\right) = \sqrt{s^2 \cdot \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]}$$

  - Intervallo di confidenza al livello di confidenza del $(1-\alpha)$

$$IC_{1-\alpha}(Y_{x^*}) = \widehat{Y}_{x^*} \pm t_{n-2,\alpha/2}\widehat{e.s.}\left(\widehat{Y}_{x^*}\right)$$

- **Previsione**
  - Stimatore dell'errore standard dello stimatore della previsione $\widehat{Y}^P_{x^*}$

$$\widehat{e.s.}\left(Y^P_{x^*} - \widehat{Y}^P_{x^*}\right) = \sqrt{s^2 \cdot \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]}$$

  - Intervallo di confidenza al livello di confidenza del $(1-\alpha)$

$$IC_{1-\alpha}(Y^P_{x^*}) = \widehat{Y}^P_{x^*} \pm t_{n-2,\alpha/2}\widehat{e.s.}\left(Y^P_{x^*} - \widehat{Y}^P_{x^*}\right)$$

Il modello di regressione lineare multipla

Indice di determinazione multipla corretto

$$R^2 = 1 - \frac{SSE/(n-k-1)}{TSS/(n-1)}$$

Inferenza nel modello di regressione lineare multipla

$$Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_k \cdot x_{ik} + \epsilon_i$$

- **Intervalli di confidenza per i coefficienti di regressione**

$$IC_{1-\alpha}(\beta_j) = \left[\widehat{\beta}_j - t_{n-k-1}(\alpha/2) \cdot \widehat{e.s.}(\widehat{B}_j); \widehat{\beta}_j + t_{n-k-1}(\alpha/2) \cdot \widehat{e.s.}(\widehat{B}_j)\right]$$

- **Test delle ipotesi**
  - Statistica test per test su un solo coefficiente β_j

$$T = \frac{\widehat{B}_j - 0}{\widehat{e.s.}(\widehat{B}_j)}$$

- **Il test F**

- Statistica test (per il confronto del modello con il modello nullo)

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{RSS/k}{SSE/(n-k-1)}$$

- **Modelli di regressione a confronto**
  - Modello esteso e modello ridotto

$$\text{Modello esteso : } Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_h \cdot x_{ih} + \cdots + \beta_k \cdot x_{ik} + \epsilon_i$$
$$\text{versus}$$
$$\text{Modell ridotto : } Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_h \cdot x_{ih} + \epsilon_i$$

  - Statistica Test

$$F = \frac{(SSE_r - SSE_e)/(k-h)}{SSE_e/(n-k-1)} = \frac{(R_e^2 - R_r^2)/(k-h)}{(1-R_e^2)/(n-k-1)}$$

Analisi della varianza

Analisi della varianza a un fattore

- **Devianze**
  - Devianza totale (Somma dei quadrati totale)

$$D_T = \sum_{\ell=1}^{g}\sum_{i=1}^{n_\ell}\left(Y_{\ell i} - \overline{Y}\right)^2 = \sum_{i=1}^{n_1}\left(Y_{1i} - \overline{Y}\right)^2 + \cdots + \sum_{i=1}^{n_\ell}\left(Y_{gi} - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

  - Devianza entro gruppi (Somma dei quadrati entro gruppi): Somma delle devianze entro gruppi

$$D_W = \sum_{\ell=1}^{g}\sum_{i=1}^{n_\ell}\left(Y_{\ell i} - \overline{Y}_\ell\right)^2 = \sum_{\ell=1}^{g} D_\ell$$

    dove $D_\ell = \sum_{i=1}^{n_\ell}\left(Y_{\ell i} - \overline{Y}_\ell\right)^2 = s_\ell^2 \cdot (n_\ell - 1)$

  - Devianza tra gruppi (Somma dei quadrati tra gruppi): Devianza delle medie entro gruppi

$$D_B = \sum_{\ell=1}^{g}\left(\overline{Y}_\ell - \overline{Y}\right)^2 \cdot n_\ell = \left(\overline{Y}_1 - \overline{Y}\right)^2 \cdot n_1 + \cdots + \left(\overline{Y}_g - \overline{Y}\right)^2 \cdot n_g$$

- **Medie dei quadrati (Varianze)**

$$\text{Varianza tra gruppi:} \quad s_B^2 = \frac{D_B}{g-1}$$
$$\text{Varianza entro i gruppi:} \quad s^2 = \frac{D_W}{n-g} \quad = \frac{s_1^2 \cdot (n_1-1) + s_2^2 \cdot (n_2-1) + \cdots + s_g^2 \cdot (n_g-1)}{(n_1-1) + (n_2-1) + \cdots + (n_g-1)}$$

- **Verifica di ipotesi per l'uguaglianza tra le medie**
  - Statistica test
$$F = \frac{D_B/(g-1)}{D_W/(n-g)}$$
- **Il modello di analisi della varianza (con vincolo baseline)**
$$Y_i = \beta_0 + \beta_2 \cdot A_{i2} + \cdots + \beta_g \cdot A_{ig} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \text{ indipendenti}$$
$$A_{i\ell} = \begin{cases} 1 & \text{se l'unità } i \text{ appartiene al gruppo } \ell \\ 0 & \text{Altrimenti} \end{cases} \qquad \ell = 1, \ldots, g$$
  - Stime dei parametri
$$\widehat{\beta}_0 = \overline{Y}_1 \qquad \widehat{\beta}_1 = 0 \qquad \widehat{\beta}_\ell = \overline{Y}_\ell - \overline{Y}_1 \qquad \ell = 2, \ldots, g$$
  - Il test F
$$F = \frac{D_B/(g-1)}{D_W/(n-g)} = \frac{RSS/(g-1)}{SSE/(n-g)}$$

Analisi della varianza con due fattori

- **Variabili indicatrici**
  - Variabili indicatrici per il fattore A
$$A_{ih} = \begin{cases} 1 & \text{se l'unità } i \text{ appartiene al gruppo } h \\ 0 & \text{Altrimenti} \end{cases} \qquad h = 1, \ldots, H$$
  - Variabili indicatrici per il fattore B
$$B_{ik} = \begin{cases} 1 & \text{se l'unità } i \text{ appartiene al gruppo } k \\ 0 & \text{Altrimenti} \end{cases} \qquad k = 1, \ldots, K$$
- **Modello di analisi della varianza con due fattori senza interazione (Vincolo Baseline)**
$$Y_i = \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} + \epsilon_i$$
  con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti
- **Il test F per il confronto del modello con il modello nullo**
  - Statistica test
$$F = \frac{RSS/(H-1+K-1)}{SSE/(n-H-K+1)}$$
- **Modelli a confronto**

– Modello esteso e modello ridotto

Modello esteso : $Y_i = \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} + \epsilon_i$

versus

Modell ridotto : $Y_i = \beta_0 + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iH} + \epsilon_i$

– Statistica Test

$$F = \frac{(SSE_r - SSE_e)/(H-1)}{SSE_e/(n-H-K+1)}$$

- **Modello di analisi della varianza con due fattori con interazione (Vincolo Baseline)**

$$\begin{aligned} Y_i &= \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} \\ &+ \beta_{22}^{AB} A_{i2} \cdot B_{i2} + \ldots + \beta_{2K}^{AB} A_{i2} \cdot B_{iK} + \ldots \\ &+ \beta_{H2}^{AB} A_{iH} \cdot B_{i2} + \ldots + \beta_{HK}^{AB} A_{iH} \cdot B_{iK} + \epsilon_i \end{aligned}$$

con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

– **Valutare la significatività dell'interazione**

– Modello esteso:

$$\begin{aligned} Y_i &= \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} \\ &+ \beta_{22}^{AB} A_{i2} \cdot B_{i2} + \ldots + \beta_{2K}^{AB} A_{i2} \cdot B_{iK} + \ldots \\ &+ \beta_{H2}^{AB} A_{iH} \cdot B_{i2} + \ldots + \beta_{HK}^{AB} A_{iH} \cdot B_{iK} + \epsilon_i \end{aligned}$$

versus

Modello ridotto:

$$Y_i = \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} + \epsilon_i$$

– Statistica Test

$$F = \frac{(SSE_r - SSE_e)/(H \cdot K - H - K + 1)}{SSE_e/(n - H \cdot K)}$$

Analisi della covarianza

Modello additivo

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \cdots + \beta_g^A \cdot A_{ig} + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

- **Stime dei coefficienti**

$$\begin{aligned} \widehat{\beta}_0 &= \overline{y}_1 - \widehat{\beta}_1 \cdot \overline{x}_1 \\ \widehat{\beta}_\ell^A &= (\overline{y}_\ell - \overline{y}_1) - \widehat{\beta}_1 \cdot (\overline{x}_\ell - \overline{x}_1) \quad \ell = 2, \ldots, g \\ \widehat{\beta}_1 &= \frac{\sum_{\ell=1}^{g} \sum_{i=1}^{n_\ell} (y_{i\ell} - \overline{y}_\ell) \cdot (x_{i\ell} - \overline{x}_\ell)}{\sum_{\ell=1}^{g} \sum_{i=1}^{n_\ell} (x_{i\ell} - \overline{x}_\ell)^2} \end{aligned}$$

Modello con interazione

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \cdots \beta_g^A \cdot A_{ig} + \beta_2^{AX} \cdot A_{i2} \cdot x_i + \cdots + \beta_g^{AX} \cdot A_{ig} \cdot x_i + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

- **Valutare la significatività dell'interazione**
  - Modello esteso:

    $$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \cdots + \beta_g^A \cdot A_{ig} + \beta_2^{AX} \cdot A_{i2} \cdot x_i + \cdots + \beta_g^{AX} \cdot A_{ig} \cdot x_i + \epsilon_i$$

    versus

    Modello ridotto:

    $$Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2^A \cdot A_{i2} + \cdots \beta_g^A \cdot A_{ig} + \epsilon_i$$

    con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti
  - Statistica Test

    $$F = \frac{(SSE_r - SSE_e)/(g-1)}{SSE_e/(n - 2 \cdot g)} = \frac{(R_e^2 - R_r^2)/(g-1)}{(1 - R_e^2)/(n - 2 \cdot g)}$$

Modelli per variabili risposta categoriche

Modello di regressione logistica

$$\begin{aligned} logit(\pi_i) &= logit\left(\frac{Pr(Y_i = 1 | X_{i1} = x_{i1}, \ldots, X_{ik} = x_{ik})}{1 - Pr(Y_i = 1 | X_{i1} = x_{i1}, \ldots, X_{ik} = x_{ik})}\right) \\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \end{aligned}$$

$$\Longleftrightarrow$$

$$\pi_i = Pr(Y_i = 1 | X_{i1} = x_{i1}, \ldots, X_{ik} = x_{ik}) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}$$

Inferenza su un parametro β_j nel modello logit

- Statistica test

  $$Z = \frac{\widehat{B}_j}{\widehat{e.s.}(\widehat{B}_j)}$$

- Intervallo di confidenza per $\exp(\beta_j)$ al livello di confidenza $1 - \alpha$

  $$IC_{1-\alpha}(\exp(\beta_j)) = \left[\exp\left\{\widehat{\beta}_j - z_{\alpha/2} \cdot \widehat{e.s.}(\widehat{B}_j)\right\}; \exp\left\{\widehat{\beta}_j + z_{\alpha/2} \cdot \widehat{e.s.}(\widehat{B}_j)\right\}\right]$$

Confronto tra modelli

- Modello esteso

$$logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_h x_{ih} + \cdots + \beta_k x_{ik}$$

- Modello ridotto

$$logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_h x_{ih}$$

- Statistica test del rapporto di verosimiglianza

$$-2\log\left(\frac{\mathcal{L}(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_h)}{\mathcal{L}(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_h, \ldots, \widehat{\beta}_k)}\right) = -2\log\left(\frac{\mathcal{L}_r}{\mathcal{L}_e}\right) = -2\left(\log\mathcal{L}_r - \log\mathcal{L}_e\right)$$

Information Criteria

$$AIC = -2\log\mathcal{L} + 2\cdot(k+1) \qquad BIC = -2\log\mathcal{L} + (k+1)\cdot\log(n)$$

dove $k+1$ è il numero di parametri del modello (inclusa l'intercetta)