Psicometria 2
1. regressione lineare (iv)

Michele Grassi
Dipartimento di Psicologia, Università di Trieste
Modulo 2, 4 C.F.U.
Marzo – Aprile 2012

Regressione lineare bivariata

1. **Dati questi vettori x e y, utilizzando solo l'algebra lineare calcola i parametri *a* e *b* della retta di regressione:**

$$\mathbf{x} = \begin{bmatrix} 2 \\ 3 \\ 1 \\ 4 \\ 5 \\ 8 \end{bmatrix}; \quad \mathbf{y} = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 7 \\ 6 \\ 7 \end{bmatrix}.$$

2. **Dai dati precedenti, calcola SQ_{TOT}, SQ_{REG} e SQ_{ERR} .**

3. **Calcola inoltre l'indice R^2**

Regressione - soluzione

1. **Dati i vettori x e y, utilizzando solo l'algebra lineare calcola i parametri *a* e *b* della retta di regressione:**

Occorre applicare questa formula: $\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum y_i x_i \end{bmatrix};$

a. Costruiamo la matrice $\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix},$

dove $n = 6;\ \sum x_i = \mathbf{1}' \cdot \mathbf{x} = 23;\ \sum x_i^2 = \mathbf{x}' \cdot \mathbf{x} = 119,$

sostituendo otteniamo $\begin{bmatrix} 6 & 23 \\ 23 & 119 \end{bmatrix}$

b. Troviamo l'inversa: $\begin{bmatrix} 0.643 & -0.124 \\ -0.124 & 0.032 \end{bmatrix}$

c. Troviamo $\begin{bmatrix} \sum y_i = \mathbf{1}' \cdot \mathbf{y} = 27 \\ \sum y_i x_i = \mathbf{y}' \cdot \mathbf{x} = 128 \end{bmatrix};$

Regressione - soluzione

d. Infine: $\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0.643 & -0.124 \\ -0.124 & 0.032 \end{bmatrix} \begin{bmatrix} 27 \\ 128 \end{bmatrix} = \begin{bmatrix} 1.454 \\ 0.795 \end{bmatrix}.$

La retta di regressione è dunque: $\hat{y} = 1.454 + 0.795x.$

2. **Dai dati precedenti, calcola SQ_{TOT}, SQ_{REG} e SQ_{ERR} .**

a. Sappiamo che

$$SQ_{TOT} = SQ_{REG} + SQ_{ERR}$$
$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Li calcoleremo tutti e tre, per controllare l'esattezza dei calcoli

Regressione - soluzione

b. Calcoliamo SQ_{TOT}:

$$\sum(y_i-\bar{y})^2=\left(\mathbf{y}-n^{-1}\sum y_i\cdot\mathbf{1}\right)'\cdot\left(\mathbf{y}-n^{-1}\sum y_i\cdot\mathbf{1}\right)=\left(\begin{bmatrix}3\\2\\2\\7\\6\\7\end{bmatrix}-\begin{bmatrix}4.5\\4.5\\4.5\\4.5\\4.5\\4.5\end{bmatrix}\right)'\bullet\left(\begin{bmatrix}3\\2\\2\\7\\6\\7\end{bmatrix}-\begin{bmatrix}4.5\\4.5\\4.5\\4.5\\4.5\\4.5\end{bmatrix}\right)=29.5$$

c. Calcoliamo SQ_{REG}:

$$\sum(\hat{y}_i-\bar{y})^2=(\hat{\mathbf{y}}-\bar{\mathbf{Y}})'\cdot(\hat{\mathbf{y}}-\bar{\mathbf{Y}})=\left\{(a\cdot\mathbf{1}+b\cdot\mathbf{x})-\bar{\mathbf{Y}}\right\}'\cdot(\hat{\mathbf{y}}-\bar{\mathbf{Y}})=$$

$$=\left\{\left(\begin{bmatrix}1.454\\1.454\\1.454\\1.454\\1.454\\1.454\end{bmatrix}+\begin{bmatrix}1.589\\2.383\\0.795\\3.178\\3.973\\6.357\end{bmatrix}\right)-\begin{bmatrix}4.5\\4.5\\4.5\\4.5\\4.5\\4.5\end{bmatrix}\right\}'\bullet\begin{bmatrix}-1.457\\-0.662\\-2.251\\0.132\\0.927\\3.311\end{bmatrix}=19.468$$

Regressione - soluzione

d. Calcoliamo SQ_{ERR}:

$$\sum(y_i-\hat{y})^2=(\mathbf{y}-\hat{\mathbf{y}})'\cdot(\mathbf{y}-\hat{\mathbf{y}})=\left(\begin{bmatrix}3\\2\\2\\7\\6\\7\end{bmatrix}-\begin{bmatrix}3.043\\3.838\\2.248\\4.632\\5.427\\7.810\end{bmatrix}\right)'\bullet\left(\begin{bmatrix}3\\2\\2\\7\\6\\7\end{bmatrix}-\begin{bmatrix}3.043\\3.838\\2.248\\4.632\\5.427\\7.810\end{bmatrix}\right)=10.032$$

e. Controlliamo:

$$SQ_{TOT}=SQ_{REG}+SQ_{ERR}$$
$$29.5\cong 19.468+10.032$$

3. Calcoliamo il coefficiente di determinazione R^2

$$R^2=\frac{SQ_{REG}}{SQ_{TOT}}=\frac{19.468}{29.5}=0.66$$

Regressione lineare multipla

4. I dati del vettore y precedente dipendono non solo dalla variabile indipendente x (d'ora in avanti, x_1), ma anche dalle variabili indipendenti x_2 e x_3:

$$\mathbf{x}_2=\begin{bmatrix}2\\2\\1\\5\\9\\2\end{bmatrix};\ \mathbf{x}_3=\begin{bmatrix}3\\7\\2\\6\\8\\9\end{bmatrix}.$$

Utilizzando solo l'algebra lineare, calcolare i coefficienti parziali di regressione, calcolarne la significatività con il test-*t* e calcolare la significatività del modello complessivo con la statistica *F*.

5. Confrontare il *modello ridotto* $\hat{y}=a+b_1x_1$ con il *modello completo* $\hat{y}=a+b_1x_1+b_2x_2+b_3x_3$, utilizzando gli indici R^2 e la statistica *F*.

Regressione multipla - soluzione

Noi sappiamo che $\underset{(n\times1)}{\mathbf{Y}}=\underset{(n\times k+1)}{\mathbf{X}}\underset{(k+1\times1)}{\boldsymbol{\beta}}+\underset{(n\times1)}{\boldsymbol{\varepsilon}}$

e che $\underset{(k+1\times1)}{\mathbf{B}}=\left(\underset{(k+1\times n)}{\mathbf{X}'}\underset{(n\times k+1)}{\mathbf{X}}\right)^{-1}\underset{(k+1\times n)}{\mathbf{X}'}\underset{(n\times1)}{\mathbf{Y}}$

$$\mathbf{y}=\begin{bmatrix}3\\2\\2\\7\\6\\7\end{bmatrix};\mathbf{X}=\begin{bmatrix}1&2&2&3\\1&3&2&7\\1&1&1&2\\1&4&5&6\\1&5&9&8\\1&8&2&9\end{bmatrix}.$$

Regressione multipla - soluzione

$$\left(\mathbf{X'X}\right)^{-1}=\begin{bmatrix} 1.1434 & 0.1243 & -0.0021 & -0.2479 \\ & 0.1693 & 0.0189 & -0.1439 \\ & & 0.0301 & -0.0301 \\ & & & 0.1551 \end{bmatrix};\ \mathbf{X'Y}=\begin{bmatrix} 27 \\ 128 \\ 115 \\ 180 \end{bmatrix};\ \mathbf{B}=\begin{bmatrix} 1.93 \\ 1.29 \\ 0.40 \\ -0.65 \end{bmatrix}.$$

4(a). L'equazione di regressione multipla è quindi

$$\boxed{\hat{y}=1.93+1.29x_1+0.40x_2-0.65x_3}$$

il vettore degli errori (o *residui*) è dato da

$$\mathbf{e}=\mathbf{y}-\hat{\mathbf{y}}=\underbrace{\begin{bmatrix} 3 \\ 2 \\ 2 \\ 7 \\ 6 \\ 7 \end{bmatrix}}_{\mathbf{y}}-\underbrace{\begin{bmatrix} 1 & 2 & 2 & 3 \\ 1 & 3 & 2 & 7 \\ 1 & 1 & 1 & 2 \\ 1 & 4 & 5 & 6 \\ 1 & 5 & 9 & 8 \\ 1 & 8 & 2 & 9 \end{bmatrix}\begin{bmatrix} 1.93 \\ 1.29 \\ 0.40 \\ -0.65 \end{bmatrix}}_{\mathbf{XB}}=\underbrace{\begin{bmatrix} 3 \\ 2 \\ 2 \\ 7 \\ 6 \\ 7 \end{bmatrix}-\begin{bmatrix} 3.36 \\ 2.03 \\ 2.31 \\ 5.20 \\ 6.81 \\ 7.21 \end{bmatrix}}_{\mathbf{e}}$$

Regressione multipla - soluzione

La varianza dei parametri di regressione è:

$$S_E^2=\frac{\mathbf{e'e}}{n-k-1}=\frac{4.156}{6-3-1}=2.078$$

$$Var(a)=S_E^2\left[\left(\mathbf{X'X}\right)^{-1}\right]_{11}=2.078\times 1.1434=2.376$$

$$Var(b_1)=S_E^2\left[\left(\mathbf{X'X}\right)^{-1}\right]_{22}=2.078\times 0.1693=0.352$$

$$Var(b_2)=S_E^2\left[\left(\mathbf{X'X}\right)^{-1}\right]_{33}=2.078\times 0.0301=0.063$$

$$Var(b_3)=S_E^2\left[\left(\mathbf{X'X}\right)^{-1}\right]_{44}=2.078\times 0.1551=0.322$$

La statistica t (2 g.d.l) relativa al primo parametro (intercetta) risulta essere:

$$t=\frac{\mathbf{B}_1-0}{\sqrt{S_E^2\left[\left(\mathbf{X'X}\right)^{-1}\right]_{11}}}=\frac{1.93}{\sqrt{2.376}}$$

4(b). Il vettore di parametri di regressione trasformati in punti t si ottiene come:

$$\boldsymbol{t}=\begin{bmatrix} \sqrt{1/2.376} & 0 & 0 & 0 \\ 0 & \sqrt{1/0.352} & 0 & 0 \\ 0 & 0 & \sqrt{1/0.063} & 0 \\ 0 & 0 & 0 & \sqrt{1/0.322} \end{bmatrix}\begin{bmatrix} 1.93 \\ 1.29 \\ 0.40 \\ -0.65 \end{bmatrix}=\begin{bmatrix} 1.93/\sqrt{2.376} \\ 1.29/\sqrt{0.352} \\ 0.40/\sqrt{0.063} \\ -0.65/\sqrt{0.322} \end{bmatrix};$$

…tutti con $p > 0.05$

Regressione multipla - soluzione

4(c). Calcoliamo ora la significatività del modello completo con:

$$F_{(k,n-k-1)}=\frac{SQ_{REG_C}/k}{SQ_{ERR_C}/(n-k-1)}=\frac{SQ_{REG_C}}{SQ_{ERR_C}}\times\frac{(n-k-1)}{k}$$

Noi sappiamo che

$$n=6;\ \ k=3;$$

$$SQ_{REG_C}=\sum\left(\hat{y}_i-\bar{Y}\right)^2=;\left(\hat{\mathbf{y}}-\bar{\mathbf{Y}}\right)'\cdot\left(\hat{\mathbf{y}}-\bar{\mathbf{Y}}\right)=$$

$$=\left(\left(b_0\cdot\mathbf{1}+b_1\cdot\mathbf{x}_1+b_2\cdot\mathbf{x}_2+b_3\cdot\mathbf{x}_3\right)-\bar{\mathbf{Y}}\right)'\cdot\left(\hat{\mathbf{y}}-\bar{\mathbf{Y}}\right)=25.344$$

$$SQ_{ERR_C}=\sum\left(y_i-\hat{y}_i\right)^2=\left(\mathbf{y}-\hat{\mathbf{y}}\right)'\cdot\left(\mathbf{y}-\hat{\mathbf{y}}\right)=4.156;$$

$$\boxed{F_{(3,2)}=\frac{25.344}{4.156}\times\frac{2}{3}=4.065;\ \ p>0.05.}$$

Regressione multipla - soluzione

5. Confrontiamo il *modello ridotto* (senza $\mathbf{x}_2$ e $\mathbf{x}_3$; $q = 2$) con il *modello completo* utilizzando gli indici R^2 e la statistica F:

$$F_{(q,n-k-1)}=\frac{\left(R_C^2-R_R^2\right)/q}{\left(1-R_C^2\right)/(n-k-1)}=\frac{R_C^2-R_R^2}{1-R_C^2}\times\frac{(n-k-1)}{q}$$

Noi sappiamo che

$$n=6;\ \ k=3;\ q=2$$

$$R_R^2=\frac{SQ_{REG_R}}{SQ_{TOT}}=\frac{19.468}{29.5}=0.66$$

$$R_C^2=\frac{SQ_{REG_C}}{SQ_{TOT}}=\frac{25.344}{29.5}=0.859$$

$$\boxed{F_{(2,2)}=\frac{0.859-0.66}{1-0.859}\times\frac{2}{2}=1.414;\ \ p>0.05.}$$