DATA MINING PER IL MARKETING (63 ore)

Marco Riani

mriani@unipr.it

Sito web del corso

http://www.riani.it/DMM

Ripasso: vincoli del sistema di equazioni normali nella regressione semplice

$$\sum_{i=1}^{n} e_i = 0$$

$$\sum_{i=1}^{n} x_i e_i = 0$$

vincoli del sistema di equazioni normali nella regressione multipla

- L'equazione

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

- implica X'e=0

Interpretazione dei parametri nella regressione lineare multipla

Modello di regressione multiplo (y investimenti)

STATISTICHE	
	Parametri $\hat{\beta}_j$
Intercetta (X_0)	$-441,272$
PIL (X_1)	$0,625$
Trend (X_2)	$-12,522$

Interpretazione dei parametri nella regressione lineare multipla

si parla, infatti, di *coefficienti di regressione parziale* o di coefficienti *netti* dato che una volta fissato il valore delle variabili $\mathbf{x}_0, ..., \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, ..., \mathbf{x}_{k-1}$, il parametro β_j rappresenta il coefficiente angolare della retta

$$E(\mathbf{y}) = K + \beta_j \mathbf{x}_j$$

dove K è il valore comprensivo di tutte le variabili in cui sono stati fissati i valori. Si può quindi dire che i coefficienti di regressione parziale misurano la relazione tra la variabile dipendente e la corrispondente variabile esplicativa, mantenendo costante il livello delle altre.

Interpretazione di $\hat{\beta}_j$

- Valore previsto per l'unità i

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,0} + \ldots + \hat{\beta}_{j-1} x_{i,j-1} + \hat{\beta}_j x_{i,j} + \hat{\beta}_{j+1} x_{i,j+1} + \ldots + \hat{\beta}_{k-1} x_{i,k-1} \quad (1)$$

- Valore previsto per l'unità i quando la variabile j viene aumentata di un'unità e tutto il resto viene mantenuto fisso

$$\hat{\beta}_0 + \hat{\beta}_1 x_{i,0} + \ldots + \hat{\beta}_{j-1} x_{i,j-1} + \hat{\beta}_j (x_{i,j} + 1) + \hat{\beta}_{j+1} x_{i,j+1} + \ldots + \hat{\beta}_{k-1} x_{i,k-1} \quad (2)$$

- La differenza tra (2) e (1) è $\hat{\beta}_j$

Criterio alternativo per trovare I coefficienti di regr. lineare multipla (p. 70)

- Es. trovare il coeff. di regressione parziale del PIL
- 1) Regressione Investimenti su tutte le variabili tranne il PIL
- 2) Regressione del PIL su tutte le altre variabili esplicative
- 3) Regressione tra i residui di 1) e i residui di 2)

File excel

- File regr-mult0.xlsx

In generale

- Dato

$y = 1\,\beta_0 + X_1\,\beta_1 + X_2\,\beta_2 + \ldots + X_i\,\beta_i + \ldots + X_{k-1}\,\beta_{k-1} + \varepsilon$

- Per trovare beta cappello_i
  - Si regredisce y su tutte le variabili tranne X_i e si prendono i residui (di INPUT)
  - Si regredisce X_i su tutte le altre variabili esplicative e si prendono i residui (di OUTPUT)
  - Il coefficiente di regressione semplice calcolato sulle due serie dei residui produce beta cappello_i

Modello di regressione nell-universo e nel campione

$$y = X\beta + \varepsilon$$

$$y = X\hat{\beta} + e$$

- Qual è la relazione tra *e* ed ε?

Analisi dei valori previsti

$$\hat{y} = X\hat{\beta}$$

$$\hat{y} = X(X'X)^{-1}X'y$$

$$\hat{y} = Hy$$

$$H = X(X'X)^{-1}X'$$

Analisi della matrice H

- Simmetrica e idempotente

$$\begin{pmatrix}
0.25 & 0.24 & 0.21 & 0.18 & 0.15 & 0.11 & 0.06 & 0.02 & -0.01 & -0.02 & -0.02 & 0.00 & -0.03 & -0.07 & -0.08 \\
0.24 & 0.25 & 0.22 & 0.19 & 0.15 & 0.10 & 0.02 & -0.03 & -0.06 & -0.06 & -0.03 & 0.05 & 0.02 & -0.04 & -0.02 \\
0.21 & 0.22 & 0.20 & 0.17 & 0.13 & 0.09 & 0.02 & -0.02 & -0.05 & -0.04 & -0.02 & 0.06 & 0.03 & -0.02 & 0.00 \\
0.18 & 0.19 & 0.17 & 0.14 & 0.12 & 0.09 & 0.04 & 0.01 & -0.01 & -0.01 & 0.00 & 0.05 & 0.03 & 0.00 & 0.01 \\
0.15 & 0.15 & 0.13 & 0.12 & 0.10 & 0.08 & 0.06 & 0.04 & 0.03 & 0.02 & 0.03 & 0.04 & 0.03 & 0.01 & 0.01 \\
0.11 & 0.10 & 0.09 & 0.09 & 0.08 & 0.08 & 0.09 & 0.09 & 0.09 & 0.07 & 0.05 & 0.02 & 0.02 & 0.02 & 0.00 \\
0.06 & 0.02 & 0.02 & 0.04 & 0.06 & 0.09 & 0.16 & 0.19 & 0.20 & 0.16 & 0.10 & -0.04 & -0.03 & 0.01 & -0.04 \\
0.02 & -0.03 & -0.02 & 0.01 & 0.04 & 0.09 & 0.19 & 0.24 & 0.26 & 0.21 & 0.13 & -0.06 & -0.05 & 0.02 & -0.05 \\
-0.01 & -0.06 & -0.05 & -0.01 & 0.03 & 0.09 & 0.20 & 0.26 & 0.27 & 0.23 & 0.14 & -0.06 & -0.04 & 0.04 & -0.03 \\
-0.02 & -0.06 & -0.04 & -0.01 & 0.02 & 0.07 & 0.16 & 0.21 & 0.23 & 0.20 & 0.14 & -0.01 & 0.01 & 0.07 & 0.02 \\
-0.02 & -0.03 & -0.02 & 0.00 & 0.03 & 0.05 & 0.10 & 0.13 & 0.14 & 0.14 & 0.12 & 0.07 & 0.08 & 0.12 & 0.10 \\
0.00 & 0.05 & 0.06 & 0.05 & 0.04 & 0.02 & -0.04 & -0.06 & -0.06 & -0.01 & 0.07 & 0.22 & 0.22 & 0.18 & 0.25 \\
-0.03 & 0.02 & 0.03 & 0.03 & 0.03 & 0.02 & -0.03 & -0.05 & -0.04 & 0.01 & 0.08 & 0.22 & 0.23 & 0.20 & 0.27 \\
-0.07 & -0.04 & -0.02 & 0.00 & 0.01 & 0.02 & 0.01 & 0.02 & 0.04 & 0.07 & 0.12 & 0.18 & 0.20 & 0.21 & 0.25 \\
-0.08 & -0.02 & 0.00 & 0.01 & 0.01 & 0.00 & -0.04 & -0.05 & -0.03 & 0.02 & 0.10 & 0.25 & 0.27 & 0.25 & 0.32
\end{pmatrix}$$

Analisi degli elementi sulla diagonale principale della matrice H

$$h_i = \sum_{i=1}^{n} h_{ij}^2 = h_i^2 + \sum_{j \neq i}^{n} h_{ij}^2$$

- Gli elementi sulla diagonale principale sono compresi tra 0 e 1

Nel modello di regressione semplice (p. 187-188)

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}$$

- Di conseguenza h_{ii} è elevato se x_i è distante dalla nuvola dei punti

- In letteratura le osservazioni a cui corrisponde

$$\text{un } h_{ii} > 2\bar{h} = 2\sum_i^n h_{ii}/n = 2k/n$$

- vengono detti punti di leverage

Esercizio

- Rappresentare graficamente gli elementi che sono sulla diagonale principale della matrice H e trovare eventuali punti di leverage

Analisi dei residui

$$y = X\beta + \varepsilon$$

$$y = X\hat{\beta} + e$$

$$e = M\varepsilon$$

$$M = I - X(X'X)^{-1}X' = I - H$$

Analisi dei residui

$$var(e) = var(M\epsilon) = ME(\epsilon\epsilon')M' = \sigma^2 M$$

$$var(e_i) = \sigma^2(1 - h_{ii})$$

$$cov(e_i, e_j) = -\sigma^2 h_{ij}$$

Caratteristiche della matrice M

- Simmetrica
- Idempotente

- La somma dei quadrati dei residui si può scrivere come

$$\sum_{i=}^{n} e_i^2 = e'e = \varepsilon' M \varepsilon$$

- Forma quadratica idempotente

Ulteriore interpretazione dei punti di leverage

$$var(e_i) = \sigma^2(1 - h_{ii})$$

- I punti in cui h_{ii} è grande sono i valori influenti nella regressione, ossia quelli che attirano a sé l'iperpiano di regressione

Studio della distribuzione di $\hat{\beta}$

$$E(\hat{\beta}) = \beta$$

$$\text{var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

$$\hat{\beta} \sim N\left[\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right]$$

Nella regressione semplice avevamo ($\hat{\alpha}$ e $\hat{\beta}$ scalari)

$$\text{var}(\hat{\alpha}) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]$$

$$\text{cov}(\hat{\alpha}, \hat{\beta}) = -\bar{x} \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

- In generale ($\hat{\beta}$ vettore di k elementi)

$$\text{var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

Nella regressione semplice

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

$$(X'X)^{-1} = \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

$$\text{var}(\hat{\alpha}) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]$$

$$\text{cov}(\hat{\alpha}, \hat{\beta}) = -\bar{x} \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Stima di σ^2

- E(s^2)?

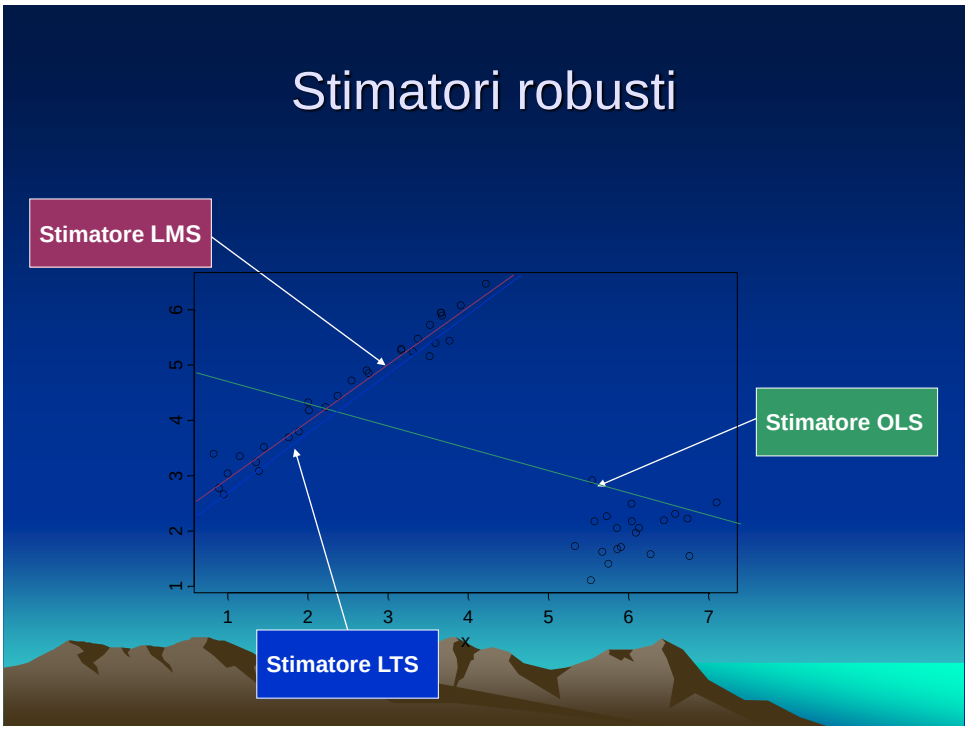
$$E(e'e) = \sigma^2(n-k)$$

- Qual è la distribuzione di s^2 (somma dei quadrati dei residui diviso i gradi di libertà)

Statistica robusta

- Obiettivo: trovare una funzione interpolante che descriva la maggior parte delle osservazioni e non sia influenzata dalla presenza di valori atipici

Stimatori robusti



Analisi di diversi stimatori

- Min. somma dei quadrati dei residui (OLS)
- Min. somma dei valori assoluti dei residui (MAD)
- Min. mediana dei quadrati dei residui (LMS)
- Min. la somma del 50% dei residui al quadrato più piccoli (LTS)