Analisi dei Dati Multivariata

Nicola Torelli

2024

1 Introduzione

L'analisi multivariata riguarda l'analisi di un insieme di variabili $x_1, x_2, \ldots, x_p$ (con $p \geq 3$) misurate su n unità. Tale analisi permette di cogliere relazioni complesse presenti nei dati, che potrebbero non emergere dall'analisi di coppie di variabili.

Le variabili possono essere quantitative o categoriali. Nel caso di dati misti, è importante distinguere le variabili categoriali, definendole come fattori in R.

2 Strumenti di Analisi

2.1 Regressione Multipla

L'estensione dell'analisi di regressione semplice a quella multipla coinvolge una variabile risposta quantitativa e più variabili esplicative. Questa tecnica, fondamentale nell'analisi statistica, è ripresa in altri corsi, come il machine learning e i modelli statistici.

2.2 Cluster Analysis

L'analisi di raggruppamento o cluster analysis cerca pattern nei dati multivariati, identificando unità simili. Questo approccio, tipico dell'apprendimento non supervisionato, non prevede una variabile risposta.

3 Analisi di Variabili Categoriali

3.1 Associazione Marginale e Condizionale

L'associazione marginale tra due variabili categoriali può differire dall'associazione condizionale, come illustrato dal paradosso di Simpson. Un esempio sono i dati delle ammissioni ai dipartimenti dell'università di Berkeley nel 1973.

4 Analisi di Variabili Quantitative

4.1 Matrice di Varianza-Covarianza e di Correlazione

Per più variabili quantitative, si calcolano le covarianze o i coefficienti di correlazione lineare, organizzati in matrici simmetriche. Esempi includono i dati *iris* e *Cars93*, visualizzati tramite la funzione `ggcorrplot`.

4.2 Scatterplot Matrix

La funzione `pairs()` rappresenta tutti gli scatterplot delle coppie di variabili. Per dati misti, la funzione `ggpairs` del pacchetto GGally di ggplot adotta la rappresentazione grafica più appropriata.

5 Regressione Lineare Multipla

Nei modelli di regressione lineare multipla, una variabile risposta quantitativa Y è modellata come combinazione lineare di più variabili esplicative $X_1, X_2, \ldots, X_p$. La funzione di regressione ha la forma:

$$M(Y|x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$