METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 12: ULTIME NOTE SULLA REGRESSIONE LINEARE

A.A. 2024-2025

Prof. Martina Vettoretti

IL PROBLEMA DEI MINIMI QUADRATI GENERALIZZATI

- Residui non a varianza omogenea e/o a campioni correlati → violate le assunzioni su cui si basa il metodo dei minimi quadrati lineari, ovvero:

$$Var(\varepsilon_i) = \sigma^2 \quad \forall\, i = 1, \dots, n$$
$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall\, i \neq j$$
$$\Rightarrow \quad Cov(\boldsymbol{\varepsilon}) = \sigma^2 \cdot \boldsymbol{I}$$
$$\boldsymbol{\varepsilon} = [\varepsilon_1\ \varepsilon_2\ \cdots \varepsilon_n]^{\boldsymbol{T}}$$

- Se tali assunzioni risultano violate, possiamo comunque stimare i parametri del modello di regressione lineare, se ipotizziamo di conoscere la matrice di covarianza dei termini ε_i a meno di un fattore incognito σ^2:

$$Cov(\boldsymbol{\varepsilon}) = \sigma^2 \cdot \boldsymbol{V}$$

  - σ^2 incognito
  - $\boldsymbol{V}$ nota

- Problema dei **minimi quadrati lineari generalizzati:**

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left((\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta})\right)$$

I quadrati dei residui vengono pesati per $\boldsymbol{V}^{-1}$ → osservazioni per cui la varianza dell'errore è minore vengono pesati di più

LO STIMATORE DEI MINIMI QUADRATI GENERALIZZATI

- Soluzione del problema dei minimi quadrati generalizzati:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T V^{-1} X})^{-1} \boldsymbol{X^T V^{-1} Y}$$

- Stima a posteriori di σ^2:

$$\hat{\sigma}^2 = \frac{SSE}{n - m - 1}$$

- Proprietà dello stimatore $\widehat{\boldsymbol{\beta}}$:
  - $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
  - $cov(\widehat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\boldsymbol{X^T V^{-1} X})^{-1}$

REGRESSIONE LINEARE E CORRELAZIONE

- Due strumenti statistici per valutare la presenza di una relazione lineare tra due variabili:
  - **Coefficiente di correlazione lineare di Pearson:**

$$Corr(X,Y) := \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

$$r_{X,Y} := \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1)s_x s_y}$$ ← Stimatore campionario

s_X, s_Y: deviazioni standard campionarie dei campioni osservati per X ed Y rispettivamente ($x_i, i = 1, \dots, n;\ y_i, i = 1, \dots, n$).

  - **Regressione lineare semplice:**

$$Y = \beta_0 + \beta \cdot X + \varepsilon$$

- Che relazione sussiste tra i due?

LEGAME TRA CORRELAZIONE E REGRESSIONE LINEARE SEMPLICE

- Si può dimostrare che il coefficiente di determinazione R^2 di una regressione lineare semplice tra X e Y è equivalente al quadrato del coefficiente di correlazione lineare di Pearson tra X e Y, $r_{X,Y}$.

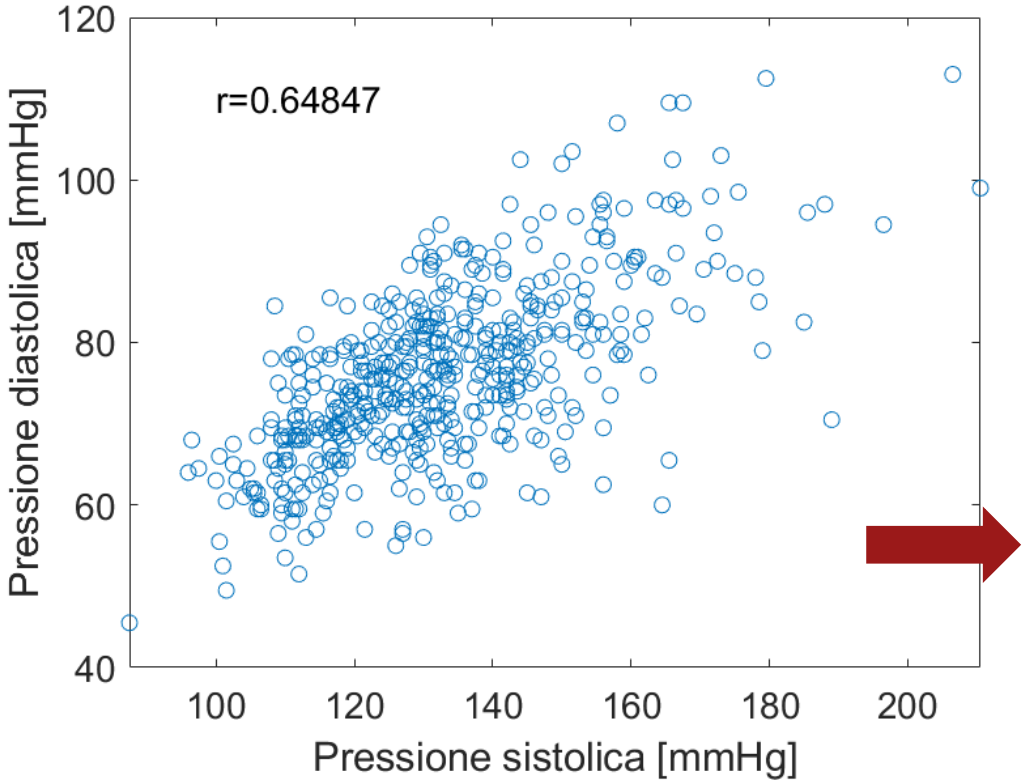
$$R^2 = r_{X,Y}^2$$

- Inoltre si ha:

$$r_{X,Y} = \hat{\beta} \cdot \frac{s_X}{s_Y}$$

ESEMPIO

- Analizziamo la relazione tra pressione arteriosa sistolica (X) e diastolica (Y) in un campione di 600 individui. Esercizio svolto in Matlab.



$$Y = \beta_0 + \beta \cdot X + \varepsilon$$

$$\hat{\beta}_0 = 25.52$$

$$\hat{\beta} = 0.38$$

$$r^2 = 0.4205 \quad \Leftrightarrow \quad R^2 = 0.4205$$

$$\hat{\beta} \cdot \frac{s_X}{s_Y} = 0.6485$$

LEGAME TRA CORRELAZIONE E REGRESSIONE LINEARE MULTIPLA

- Per la regressione lineare multipla, si può dimostrare che il coefficiente di determinazione R^2 è equivalente al quadrato del coefficiente di correlazione lineare di Pearson tra l'outcome reale e quella predetta dal modello.

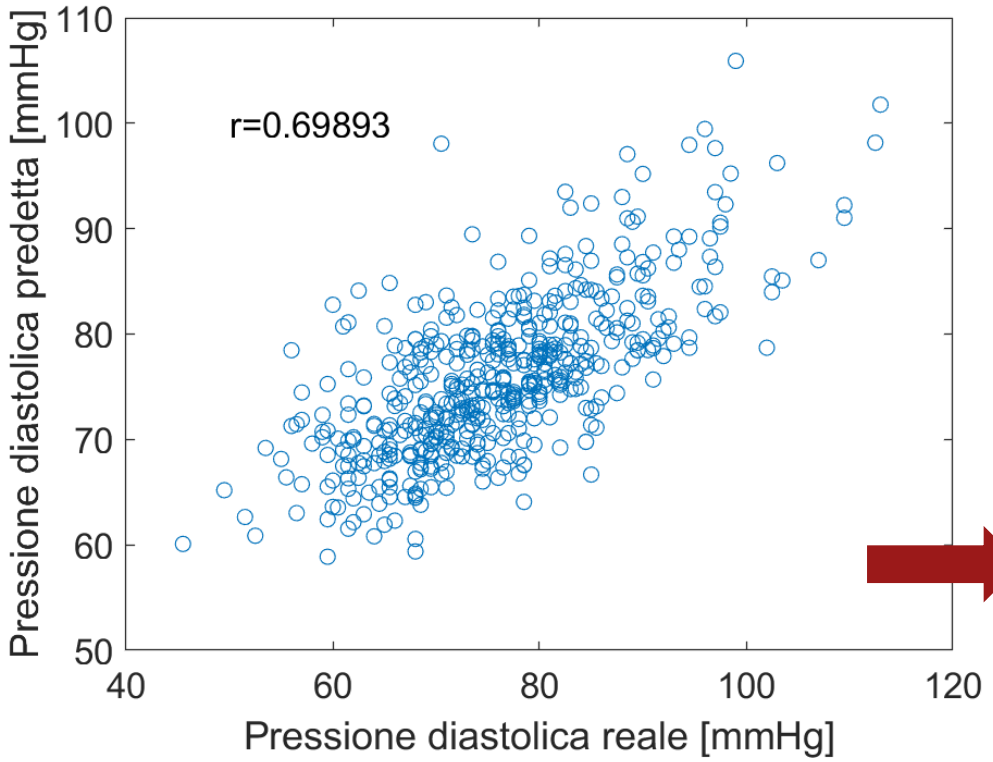
$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X} \cdot \widehat{\boldsymbol{\beta}}$$

$$R^2 = r^2_{Y,\hat{Y}}$$

ESEMPIO

- Analizziamo la relazione tra pressione arteriosa diastolica (Y) e pressione arteriosa sistolica (X_1) ed età (X_2) in un campione di 600 individui. Esercizio svolto in Matlab.



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$\hat{\beta}_0 = 45.68$$
$$\hat{\beta}_1 = 0.42$$
$$\hat{\beta}_2 = -0.41$$

$$r^2 = 0.4885 \quad \leftrightarrow \quad R^2 = 0.4885$$

ESEMPI DI DOMANDE D'ESAME SULLA REGRESSIONE LINEARE

QUIZ 1

- Nel modello di regressione lineare semplice:

$$Y = \beta_0 + \beta \cdot X + \varepsilon$$

se il coefficiente β è negativo, e significativamente diverso da 0, significa che:

a. Al crescere di X, il valor medio di Y aumenta
b. Al crescere di X, il valor medio di Y diminuisce
c. Al decrescere di X, il valor medio di Y diminuisce
d. Nessuna delle precedenti

QUIZ 2

- In un modello di regressione lineare che predice la pressione sanguigna diastolica [mmHg], la somma dei residui al quadrato (SSE) risulta pari a 25 [$mmHg^2$] e la devianza della pressione sanguigna diastolica (SST) risulta pari a 100 [$mmHg^2$]. Che percentuale della devianza della pressione sanguigna diastolica è spiegata dal modello di regressione lineare?
  a. 40%
  b. 25%
  c. 75%
  d. 60%

QUIZ 3

- Quale dei seguenti metodi di regolarizzazione non consente di ridurre la complessità di un modello di regressione lineare multipla?
  a. La regolarizzazione L2 o Ridge
  b. La regolarizzazione L1 o LASSO
  c. La regolarizzazione elastic net
  d. Nessuna delle precedenti

DOMANDE APERTE

1. Descrivi il modello di regressione lineare multipla. Come è possibile stimarne i parametri a partire da un campione di n osservazioni per le variabili coinvolte?

2. Cosa si intende per problema della multicollinearità nella regressione lineare multipla? Descrivi un approccio per mitigarlo.