Metodi Statistici per le decisioni

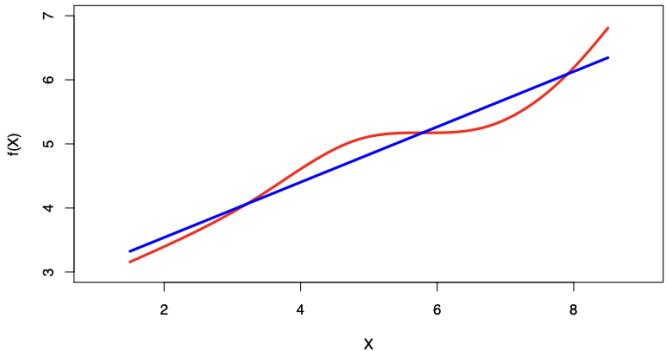
2024-2025

Vincenzo Nardelli

vincenzo.nardelli@unicatt.it

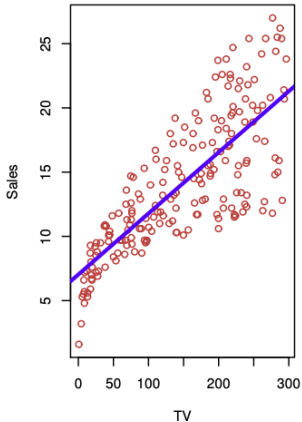
Regressione Lineare

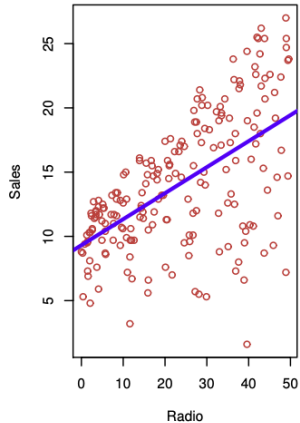
- La regressione lineare è un approccio semplice per l'apprendimento supervisionato. Assume che la dipendenza di Y da $X_1, X_2, \ldots, X_p$ sia lineare.

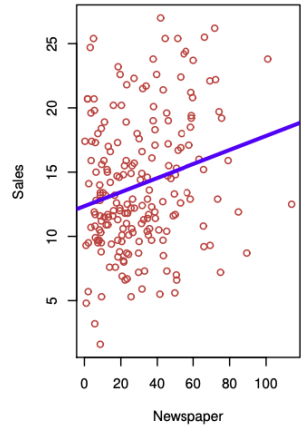


- Anche se può sembrare troppo semplicistico, la regressione lineare è estremamente utile sia concettualmente che praticamente.

Dati pubblicitari







Regressione lineare per i dati pubblicitari

Domande che potremmo porci:

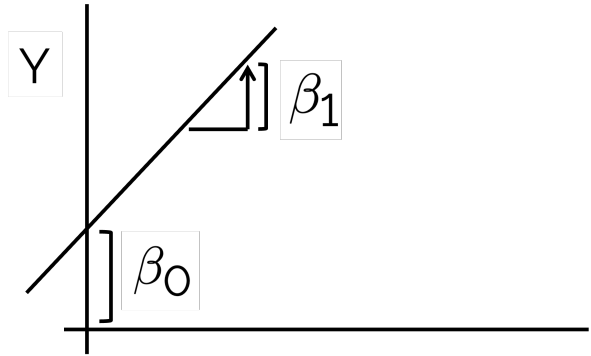
- Esiste una relazione tra budget pubblicitario e vendite?
- Quanto è forte la relazione tra budget pubblicitario e vendite?
- Quali mezzi contribuiscono alle vendite?
- Con quanta accuratezza possiamo prevedere le vendite future?
- La relazione è lineare?
- Esiste sinergia tra i mezzi pubblicitari?

Regressione lineare semplice con un singolo predittore X

Assumiamo un modello

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

dove β_0 e β_1 sono due costanti sconosciute che rappresentano l'*intercetta* e la *pendenza*, noti anche come *coefficienti* o *parametri*, e ϵ è il termine di errore.



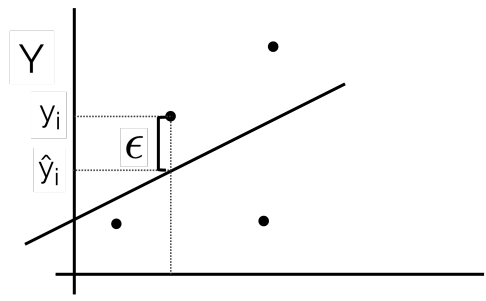
Regressione lineare semplice con un singolo predittore X

Dati alcuni valori stimati $\hat{\beta}_0$ e $\hat{\beta}_1$ per i coefficienti del modello, prevediamo le vendite future usando

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

dove $\hat{y}$ indica una previsione di Y sulla base di $X = x$. Il simbolo con il cappello denota un valore stimato.

Sia $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la previsione per Y basata sul valore i-esimo di X. Allora $e_i = y_i - \hat{y}_i$ rappresenta il residuo i-esimo.



Stima dei parametri tramite minimi quadrati

- Definiamo la *somma dei quadrati dei residui* (RSS) come

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

o equivalentemente come

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

Stima dei parametri tramite minimi quadrati

- Definiamo la *somma dei quadrati dei residui* (RSS) come

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

o equivalentemente come

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

- L'approccio dei minimi quadrati sceglie $\hat{\beta}_0$ e $\hat{\beta}_1$ per minimizzare l'RSS. I valori che minimizzano possono essere mostrati come

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

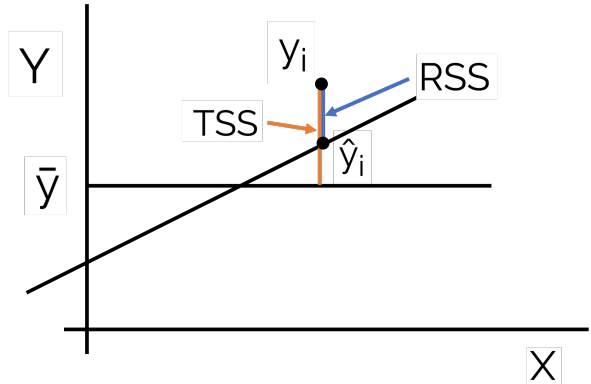
dove $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ e $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ sono le medie campionarie.

Valutazione dell'Accuratezza Complessiva del Modello

- Il *R-quadrato* o frazione della varianza spiegata è

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

dove $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ è la *somma totale dei quadrati*.



LAB Consumo di Sigarette

Un ente sanitario sta studiando come il prezzo e la tassazione influenzino il consumo di sigarette a livello statale. Analizzando i dati storici, l'obiettivo è comprendere l'efficacia delle politiche di aumento dei prezzi e delle tasse sul controllo del consumo di tabacco. Questa analisi fornirà basi solide per formulare raccomandazioni a favore della salute pubblica, con l'intento di ridurre il consumo di sigarette e i rischi correlati alla salute.

Attraverso lo studio di queste relazioni, l'ente sanitario intende valutare l'effetto delle variazioni di prezzo e tassazione per pianificare strategie che possano incentivare un decremento nel consumo di sigarette.

Descrizione delle Variabili - Dataset Cigarette (Pacchetto Ecdat)

Il dataset contiene informazioni raccolte a livello statale negli Stati Uniti e include variabili chiave per l'analisi dell'effetto dei prezzi e della tassazione sul consumo di sigarette. Le principali variabili sono:

- **state**: Stato in cui sono stati raccolti i dati.
- **year**: Anno della raccolta dei dati.
- **avgprs**: Prezzo medio di un pacchetto di sigarette (in dollari).
- **packpc**: Consumo di sigarette (pacchetti pro capite).
- **taxs**: Totale delle tasse su un pacchetto di sigarette (in dollari).

LAB Consumo di Sigarette: Domande di Analisi

- Esiste una relazione tra il prezzo medio di un pacchetto di sigarette (`avgprs`) e il consumo di sigarette (`packpc`)? La relazione è positiva o negativa?
- Qual è la correlazione tra prezzo medio e consumo di sigarette? Come possiamo interpretare questo valore per valutare l'influenza del prezzo sul consumo?
- In che modo la tassazione totale (`taxs`) influenza il consumo di sigarette? Analizza questa relazione utilizzando uno scatterplot e calcola la correlazione.
- Tramite un modello di regressione lineare, quale sarebbe l'impatto sui consumi di sigarette se la tassazione in uno stato passasse da 50$ a 100$?

LAB Real Estate

Un'agenzia immobiliare è interessata a comprendere come il livello socioeconomico di una zona influenzi il valore medio delle case di Boston. L'obiettivo è valutare l'impatto del livello di povertà sul valore medio delle abitazioni, così da poter offrire raccomandazioni più precise agli investitori e ai pianificatori urbani. L'analisi di questa

relazione è fondamentale per definire strategie di investimento mirate e per prevedere l'andamento del mercato immobiliare in funzione delle variabili socioeconomiche, contribuendo a migliorare l'efficacia delle decisioni di business.

Descrizione delle Variabili - Dataset Boston (Pacchetto ISLR2)

Il dataset contiene informazioni su vari quartieri, includendo variabili chiave per l'analisi socioeconomica del valore delle abitazioni. Di seguito alcune delle principali variabili utilizzate:

- **lstat**: Percentuale di popolazione con basso livello socioeconomico, un indicatore del livello di povertà nella zona.
- **medv**: Valore medio delle abitazioni in migliaia di dollari, rappresenta il target di interesse per il mercato immobiliare.
- **rm**: Numero medio di stanze per abitazione, una misura della dimensione abitativa media in ciascun quartiere.
- **age**: Percentuale di abitazioni costruite prima del 1940, indica la vetustà del patrimonio immobiliare.
- **dis**: Distanza media dai centri di lavoro di Boston, importante per valutare l'accessibilità ai servizi urbani.

LAB Real Estate: Domande di Analisi

- Qual è la relazione tra il livello di povertà (`lstat`) e il valore medio delle case (`medv`)? La relazione è positiva o negativa?
- Quali sono le frequenze relative e cumulative della variabile `lstat` nei vari quartieri? Cosa possiamo dedurre?
- Qual è la covarianza tra `lstat` e `medv`? Che cosa ci indica questo valore?
- Qual è la correlazione tra `lstat` e `medv`? Come si interpreta questa correlazione in termini di influenza del livello di povertà sui prezzi immobiliari?
- Possiamo prevedere il valore delle case in funzione del livello di povertà? Quali sono le previsioni del valore medio delle case per quartieri con livelli di povertà pari al 5%, 10%, e 15%?

LAB Real Estate: Visualizzazione dei Dati

- Qual è la distribuzione del livello di povertà (`lstat`) tra i quartieri? Visualizza e descrivi la distribuzione.
- Qual è la distribuzione del valore medio delle case (`medv`) tra i quartieri?
- Esiste una relazione lineare tra livello di povertà e valore medio delle case? Mostra il grafico e discuti i risultati.

LAB Real Estate: Analisi del Modello

- Quali sono i coefficienti di regressione tra livello di povertà e valore medio delle case? Come si interpretano l'intercetta e la pendenza?
- Qual è il livello di bontà di adattamento del modello? È una relazione forte o debole?
- In che modo possiamo utilizzare i risultati del modello per stimare il valore delle case in quartieri con livelli diversi di povertà?
- Prova a stimare altri modelli con le rimanenti variabili. Come si interpretano i risultati?