

Statistica I

Unità L: gli alberi di ciliegio nero

Tommaso Rigon

Università Milano-Bicocca



Argomenti affrontati

- Modelli linearizzabili

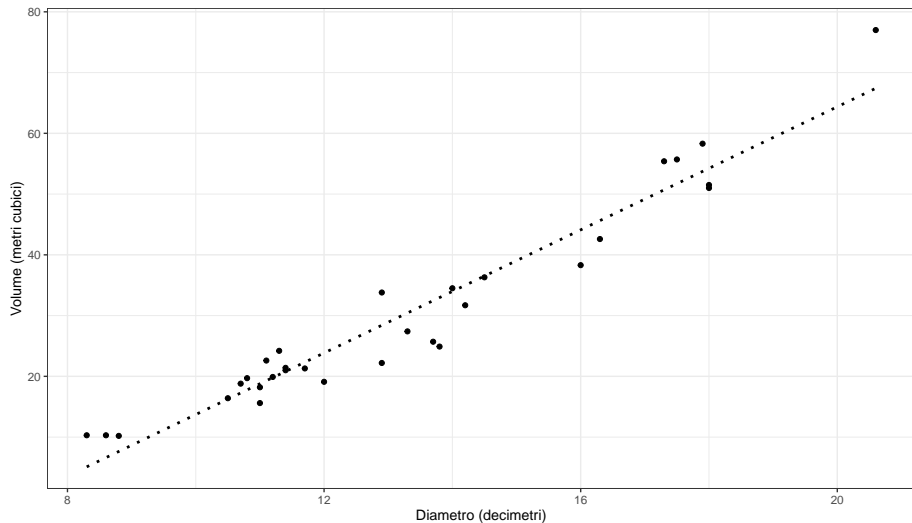
Riferimenti al libro di testo

- §22.9
- **Nota.** Alcuni paragrafi richiedono la conoscenza di nozioni di calcolo delle probabilità. Tali passaggi non sono materia d'esame.

Descrizione del problema

- Il modello costruito nell'unità K per gli alberi di ciliegio è una buona **approssimazione** della relazione tra diametro e volume.
- Ci sono tuttavia due caratteristiche non del tutto apprezzabili.
- In primo luogo, il modello pare non cogliere in maniera appropriata l'**andamento ai due estremi**: le osservazioni sembrano “curvare”, mentre il modello è una retta.
- Il comportamento per diametri molto piccoli è del tutto privo di senso. Si noti, infatti, che l'**intercetta stimata** $\hat{\alpha} = -36.88$ è **negativa**.
- Questo significa che valori prossimi a zero del diametro il **valore previsto** del volume è **negativo**!

Modello lineare



Alcune considerazioni geometriche

- Per tentare risolvere i problemi appena menzionati, possiamo ragionare sulla geometria degli alberi.
- Il volume del legno di un albero deriva dal tronco e dai rami. Si tratta sostanzialmente di un **solido** con dei “buchi”, ovvero gli spazi vuoti tra i rami.
- Il volume di un solido sufficientemente regolare è del tipo

$$(\text{volume}) = k \times h \times (\text{area della base}) \times (\text{altezza}),$$

dove k è una costante che dipende dalla **forma del solido** mentre h rappresenta la **frazione del solido** non costituita da spazi vuoti.

- Queste relazioni vanno sviluppate ulteriormente per poter diventare operative.

Alcune considerazioni geometriche

- L'**area della base** non è nota, ma è certamente legata al diametro del tronco. Sembra ragionevole supporre una relazione **non lineare** tra area e diametro.
- Si pensi ad esempio all'area del cerchio, pari a $(\text{area cerchio}) = \pi/4 \times (\text{diametro})^2$.
- Nel nostro problema possiamo quindi ipotizzare una relazione del seguente tipo:

$$(\text{area della base}) = \gamma_1(\text{diametro})^{\gamma_2},$$

dove γ_1 e γ_2 sono due costanti.

- Anche l'**altezza dell'albero** non è nota, ma possiamo tentare di descriverla in funzione del diametro del tronco. Ad esempio, possiamo supporre la semplice relazione

$$(\text{altezza}) = \delta(\text{diametro}),$$

per una qualche costante $\delta > 0$.

Alcune considerazioni geometriche

- Componendo tutte le assunzioni, otteniamo quindi un **modello** che lega la variabile volume alla variabile diametro, ovvero

$$(\text{volume}) = h \times k \times \delta \times \gamma_1 \times (\text{diametro})^{1+\gamma_2}.$$

- In forma più compatta, scriveremo quindi che

$$(\text{volume}) = \eta (\text{diametro})^\lambda,$$

per due costanti η e λ .

- In maniera analoga a quanto fatto nell'unità K, potremmo determinare i valori appropriati per η e λ utilizzando i **minimi quadrati**, ovvero considerando

$$(\hat{\eta}, \hat{\lambda}) = \arg \min_{\eta, \lambda} \sum_{i=1}^n (y_i - \eta x_i^\lambda)^2.$$

- Purtroppo non esiste una **soluzione in forma chiusa** a questo problema, che infatti necessita dell'utilizzo di **tecniche numeriche**. Esistono fortunatamente delle strategie alternative.

Linearizzazione del modello

- Il problema precedente può essere risolto con un **cambio di prospettiva**, tramite una procedura chiamata **linearizzazione** del modello.
- Supponendo che la relazione sia del tipo

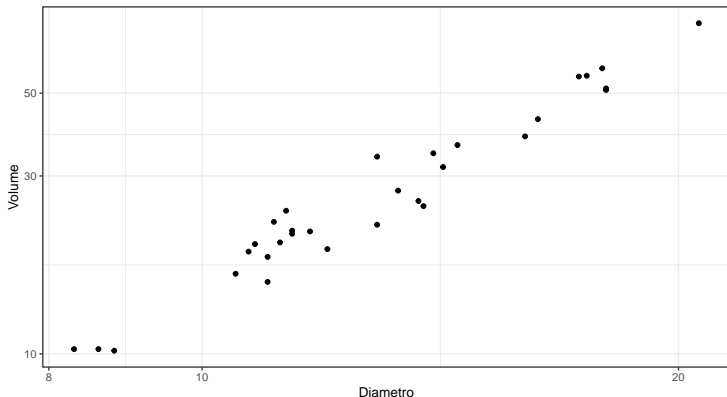
$$(\text{volume}) = \eta (\text{diametro})^\lambda,$$

allora applicando la funzione log ambo i lati, si ottiene

$$\log(\text{volume}) = \log \eta + \lambda \log(\text{diametro}).$$

- Quindi, la **relazione non lineare** che abbiamo supposto tra diametro e volume corrisponde ad una **relazione lineare** tra i logaritmi delle due variabili.

Diagramma a dispersione, scala logaritmica



- Le due variabili sono disegnate in **scala logaritmica**, anche se sono riportati i valori originali.
- In questa scala, la relazione sembra lineare, soprattutto agli estremi.

Il modello linearizzato

- La relazione in scala logaritmica descrive un modello **linearizzato**.
- Si tratta di un modello di regressione lineare semplice in cui

$$z_i = \log y_i, \quad w_i = \log x_i, \quad i = 1, \dots, n.$$

- Introducendo esplicitamente il termine di errore, avremo quindi che

$$z_i = \alpha + \beta w_i + \epsilon_i$$

in cui $\alpha = \log \eta$ e $\beta = \lambda$.

- A questo punto, possiamo determinare i **parametri trasformati** ottimali $\hat{\alpha}$ e $\hat{\beta}$ come descritto nell'unità K, ovvero utilizzando il criterio dei minimi quadrati.
- Quindi, possiamo ottenere le seguenti stime per i **parametri originali**

$$\hat{\eta} = \exp\{\hat{\alpha}\}, \quad \hat{\lambda} = \hat{\beta}.$$

Calcolo dei parametri, scala trasformata

- Passando alla scala logaritmica, otteniamo quindi che

$$\begin{aligned}\sum_{i=1}^n z_i &= 101.455, & \sum_{i=1}^n w_i &= 79.277, \\ \sum_{i=1}^n z_i^2 &= 340.343, & \sum_{i=1}^n w_i^2 &= 204.376, & \sum_{i=1}^n w_i z_i &= 263.056.\end{aligned}$$

- Perciò possiamo calcolare medie, varianze e covarianza

$$\begin{aligned}\bar{z} &= \frac{101.455}{31} = 3.273, & \bar{w} &= \frac{79.28}{31} = 2.557, \\ \text{var}(z) &= 0.266, & \text{var}(w) &= 0.055, & \text{cov}(w, z) &= 0.117.\end{aligned}$$

- Sono **evidenziati** i valori ottenuti con calcoli ad alta precisione numerica. Quindi:

$$\begin{aligned}\hat{\beta} &= \frac{0.117}{0.055} = \mathbf{2.20}, & \hat{\alpha} &= 3.273 - 2.20 \times 2.557 = \mathbf{-2.353}. \\ \text{var}(r) &= 0.266 - \frac{0.117^2}{0.055} = \mathbf{0.012}, & R^2 &= 1 - \frac{0.012}{0.27} = \mathbf{0.95}.\end{aligned}$$

Il ritorno alla scala originale

- Ritrasformando i parametri nella scala originale, otteniamo quindi che

$$\hat{\eta} = \exp\{\hat{\alpha}\} = 0.10, \quad \hat{\lambda} = \hat{\beta} = 2.20.$$

- **Nota importante.** Non ha senso confrontare il valore di R^2 nella scala trasformata rispetto al valore R^2 ottenuto nella scala originale.

- Il valore di R^2 dipende infatti dalla scala stessa: gli indici non sono confrontabili.

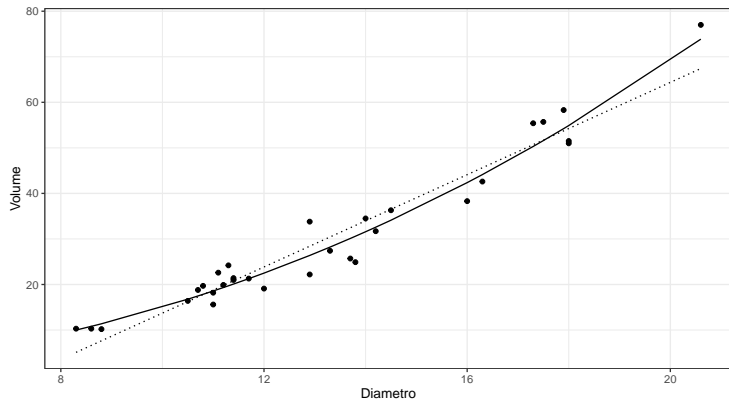
- Un approccio più corretto consiste invece nel calcolare la **varianza dei residui della scala originale**, ovvero

$$\frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{\eta} x_i^{\hat{\lambda}} \right)^2,$$

che in questo caso è pari a 10.3.

- Questa quantità è confrontabile con la varianza residuale ottenuta nell'unità K, pari a **16.9** (valore ad alta precisione numerica, pari a 17.35 nelle slides), evidenziando quindi un netto miglioramento.

Analisi grafica



- **Linea tratteggiata** è la retta di regressione. **Linea continua**: modello linearizzato.
- Anche da un punto di vista grafico il nuovo modello sembra migliore, poiché coglie la curvatura agli estremi.

- I metodi basati sui minimi quadrati descritti nell'unità K possono essere applicati **non solo** a modelli del tipo

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

ma anche a **modelli più generali**, ad esempio

$$g(y_i) = \alpha + \beta h(x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

dove $g(\cdot)$ e $h(\cdot)$ sono due funzioni appropriate.

- L'importante è che **il modello sia lineare nei parametri e non nelle variabili**.
- Ad esempio, risulta facilmente trattabile il seguente modello

$$y_i = \alpha + \beta \sin^{31}(x_i) + \epsilon_i, \quad i = 1, \dots, n.$$

Commenti all'analisi

- I modelli lineari nelle variabili possono essere visti spesso come **approssimazioni** di relazioni non lineari. Si pensi ad esempio alla formula di Taylor.
- In queste situazioni, ottenere **estrapolazioni** dal modello è pericoloso e può dare luogo a risultati insensati. Nel caso considerato, previsioni negative per il volume.
- Non bisognerebbe ignorare le informazioni a disposizione.
- Ad esempio, in questo caso, poche conoscenze di geometria hanno condotto ad un modello che pare **adattarsi meglio ai dati** e soprattutto che è più ragionevole da un punto di vista fisico.
- In generale, lo statistico ha il dovere di recuperare le conoscenze sul fenomeno che sta analizzando.
- Inoltre, è spesso utile che lo statistico “vada sul campo” (in laboratorio, nello stabilimento di produzione, etc) per vedere in prima persona come i dati sono raccolti.