Statistica I

Unità K: regressione lineare semplice

Tommaso Rigon

Università Milano-Bicocca

Unità K

Argomenti affrontati

- Modello di regressione lineare semplice
- Minimi quadrati
- Media e varianza residua, coefficiente di determinazione (R^2)

Riferimenti al libro di testo

- §22.1 — §22.4
- §22.8
- **Nota**. Alcuni paragrafi richiedono la conoscenza di nozioni di calcolo delle probabilità. Tali passaggi non sono materia d'esame.

Descrizione del problema

- Per $n = 31$ **alberi di ciliegio** nero sono disponibili le misure del diametro del tronco (misurato a circa 1m dal suolo) ed il volume ricavato dall'albero dopo l'abbattimento.

- Si vogliono utilizzare i dati per ottenere un'**equazione** che permetta di **prevedere** il volume, ottenibile solo dopo l'abbattimento dell'albero, avendo a disposizione il diametro, che è invece facilmente misurabile.

- In altri termini, stiamo cercando una qualche funzione $f(\cdot)$ tale che

$$(\text{volume}) \approx f(\text{diametro}).$$

- Una simile equazione ha differenti utilizzi.

- Ad esempio, può essere utilizzata per decidere quanti e quali alberi tagliare per ricavare un certo ammontare di legno, oppure per determinare il "prezzo" di un bosco.

I dati grezzi

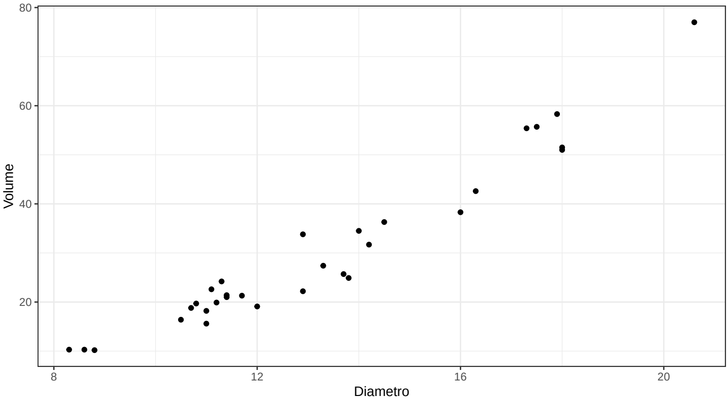
Diametro

```
 [1]  8.3  8.6  8.8 10.5 10.7 10.8 11.0 11.0 11.1 11.2 20.6 11.3
[13] 11.4 11.4 11.7 12.0 12.9 12.9 13.3 13.7 13.8 14.0 14.2 14.5
[25] 16.0 16.3 17.3 17.5 17.9 18.0 18.0
```

Volume

```
 [1] 10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 77.0 24.2
[13] 21.0 21.4 21.3 19.1 22.2 33.8 27.4 25.7 24.9 34.5 31.7 36.3
[25] 38.3 42.6 55.4 55.7 58.3 51.5 51.0
```

Diagramma di dispersione



- Possiamo quindi calcolare la correlazione:

$$\text{cor}(\texttt{diametro}, \texttt{volume}) = 0.967.$$

- È quindi evidente una **forte relazione** di tipo sostanzialmente lineare.

Un primo modello

- Adottiamo per il momento l'ipotesi di una relazione lineare.

- Possiamo allora definire un **modello lineare** del tipo

$$(\text{volume}) = \alpha + \beta\,(\text{diametro}) + (\text{errore}).$$

- L'ultima componente esprime la parte delle oscillazioni del volume non legate al diametro o che non è catturata dalla relazione lineare.

- Se $y_1, \dots, y_n$ rappresentano i volumi e $x_1, \dots, x_n$ rappresentano i diametri, allora scriveremo:

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

dove $\epsilon_1, \dots, \epsilon_n$ rappresentano invece gli errori.

Modello di regressione lineare: terminologia

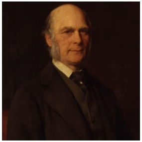
- Il modello che abbiamo appena descritto viene tipicamente chiamato **modello di regressione lineare semplice**.
- In generale, vogliamo spiegare una variabile y utilizzando un'altra variabile x, mediante un modello del tipo

$$y = \alpha + \beta x + \epsilon.$$

- La variabile y viene tipicamente chiamata **variabile risposta** o **variabile dipendente**.
- La variabile x viene chiamata **variabile esplicativa**, **regressore** oppure **variabile indipendente**.
- I valori $\alpha, \beta \in \mathbb{R}$ sono i **parametri** del modello.

Regressione lineare: cenni storici

- Il termine **regressione** deriva dalla famosa applicazione compiuta nel 1886 dal biologo e statistico **Francis Galton**.



- Galton esaminò le altezze dei figli (variabile risposta y) in funzione delle altezze dei genitori (variabile esplicativa x).
- Nella sua analisi, figli alti provenivano da genitori alti e viceversa figli bassi provenivano da genitori bassi.
- Galton notò inoltre una tendenza nelle altezze dei genitori a spostarsi verso l'altezza media nella generazione successiva.
- Galton chiamò questo fenomeno "**regression towards mediocrity**".

Metodo dei minimi quadrati: idea

- In pratica, è necessario **determinare il valore** dei parametri α e β.

- Se avessimo a disposizione un valore ragionevole dei parametri, diciamo $\hat{\alpha}$ e $\hat{\beta}$, potremmo prevedere il volume del legno usando

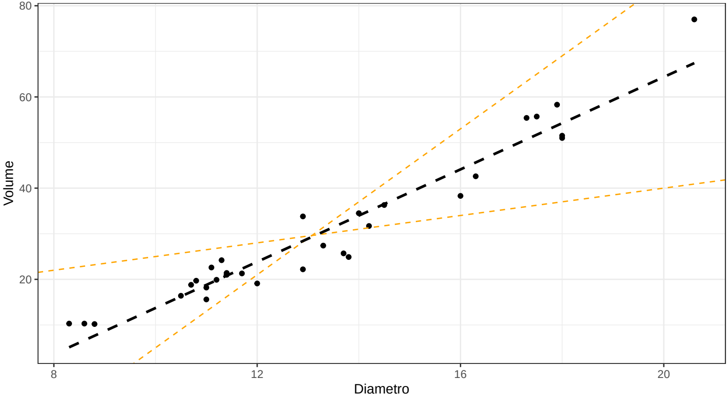
$$(\text{volume}) \approx \hat{\alpha} + \hat{\beta}(\text{diametro}).$$

- Sembra ragionevole cercare di determinare $\hat{\alpha}$ e $\hat{\beta}$ in modo tale da ottenere buone **previsioni** sull'insieme di dati osservato.

- Vogliamo quindi trovare dei valori per i parametri tali che

$$\begin{aligned} y_1 &\approx \hat{\alpha} + \hat{\beta} x_1, \\ y_2 &\approx \hat{\alpha} + \hat{\beta} x_2, \\ &\vdots \\ y_n &\approx \hat{\alpha} + \hat{\beta} x_n. \end{aligned}$$

Differenti scelte dei parametri



- Le linee **arancioni** rappresentano delle scelte **non ottimali** a fini previsivi.
- Viceversa, la linea nera attraversa la nuvola di punti e sembra una scelta appropriata.

Metodo dei minimi quadrati: la funzione di perdita

- Per rendere operativa la precedente intuizione, dobbiamo decidere cosa si intende precisamente per
$$y_i \approx \hat{\alpha} + \hat{\beta} x_i, \qquad i = 1, \dots, n.$$

- Una possibile soluzione, è scegliere i parametri che **minimizzano la funzione di perdita**
$$\ell(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^n \epsilon_i^2,$$
ovvero scegliendo $\hat{\alpha}$ e $\hat{\beta}$ tali che
$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha, \beta} \ell(\alpha, \beta).$$

- Questo criterio viene detto il **metodo dei minimi quadrati**, poiché minimizza la somma degli scarti al quadrato, ovvero la somma degli errori al quadrato.

Minimi quadrati: determinazione dei parametri

- Il criterio dei minimi quadrati è molto popolare perché la soluzione del problema di minimizzazione è semplice da calcolare.

Minimi quadrati

- L'unica soluzione al problema

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

è pari a

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{\text{cov}(x, y)}{\text{var}(x)}.$$

- La soluzione del problema è **ben definita** solamente se $\text{var}(x) > 0$.

- Questo è molto ragionevole: il parametro β indica quanto varia la risposta al variare della esplicativa, ma se $\text{var}(x) = 0$ allora l'esplicativa non varia affatto.

Dimostrazione I

- Per ogni prefissato β, conosciamo già la soluzione del seguente problema

$$\arg\min_{\alpha \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \arg\min_{\alpha \in \mathbb{R}} \sum_{i=1}^{n} (w_i - \alpha)^2,$$

avendo posto $w_i = y_i - \beta x_i$ per ogni $i = 1, \dots, n$. Infatti, dall'unità C sappiamo che il valore che minimizza tale funzione è la media aritmetica.

- Pertanto per qualsiasi valore di β, otteniamo che

$$\hat{\alpha}(\beta) = \frac{1}{n} \sum_{i=1}^{n} w_i = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta x_i) = \bar{y} - \bar{x}\beta.$$

- Dalla definizione di $\hat{\alpha}(\beta)$ segue che per ogni α, β

$$\ell(\alpha, \beta) \geq \ell(\hat{\alpha}(\beta), \beta).$$

Dimostrazione II

- Abbiamo quindi ridotto il problema iniziale al seguente sotto-problema

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}} \ell(\hat{\alpha}(\beta), \beta) = \arg\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} [(y_i - \bar{y}) - \beta(x_i - \bar{x})]^2$$

e ovviamente porremo $\hat{\alpha} = \hat{\alpha}(\hat{\beta}) = \bar{y} - \hat{\beta}\bar{x}$.

- Prendendo la derivata rispetto a β e ponendola pari a 0, si ottiene che

$$-2\sum_{i=1}^{n}(x_i - \bar{x})[(y_i - \bar{y}) - \beta(x_i - \bar{x})] = 0,$$

che possiamo riscrivere come

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \beta \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Dimostrazione III

- Quindi, se $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$ la soluzione al problema è pari a

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{\text{var}(x)},$$

  dove l'ultimo passaggio si ottiene moltiplicando numeratore e denominatore per n.

- **Nota matematica**. Per concludere la dimostrazione bisogna infine verificare che la soluzione trovata è un punto di minimo e non, ad esempio, un massimo.

- **Esercizio**. Si verifichi che la soluzione è effettivamente un punto di minimo, ad esempio valutando il segno della derivata seconda di $\ell(\hat{\alpha}(\beta), \beta)$.

Calcolo dei parametri: gli alberi di ciliegio

- In questo caso abbiamo che

$$\sum_{i=1}^{n} y_i = 935.3, \qquad \sum_{i=1}^{n} x_i = 410.7,$$

$$\sum_{i=1}^{n} x_i^2 = 5736.55, \qquad \sum_{i=1}^{n} x_i y_i = 13887.86.$$

- Perciò possiamo calcolare medie, varianza e covarianza

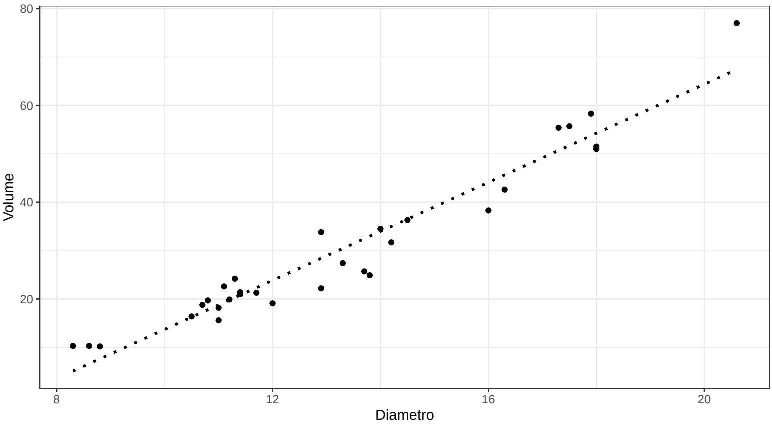
$$\bar{y} = \frac{935.5}{31} = 30.17, \qquad \bar{x} = \frac{410.7}{31} = 13.25,$$

$$\mathrm{var}(x) = \frac{5736.55}{31} - 13.25^2 = 9.53, \qquad \mathrm{cov}(x, y) = \frac{13887.86}{31} - 13.25 \times 30.17 = 48.24.$$

- Possiamo quindi determinare i parametri

$$\hat{\beta} = \frac{48.24}{9.53} = 5.06, \qquad \hat{\alpha} = 30.17 - 5.06 \times 13.25 = -36.88.$$

Diagramma di dispersione con retta di regressione



- La capacità di descrivere le variazione del volume sembra buona, con l'eccezione forse delle osservazioni più esterne.

I residui: media e varianza

- Le differenze tra i valori osservati della variabile risposta ed i valori previsti dal modello, ovvero
$$r_i = y_i - (\hat{\alpha} + \hat{\beta} x_i), \qquad i = 1, \dots, n,$$
vengono spesso chiamati **residui**.

- **Proprietà**. La media dei residui è nulla, infatti:
$$\sum_{i=1}^n r_i = \sum_{i=1}^n y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^n x_i = n\bar{y} - n(\bar{y} - \hat{\beta}\bar{x}) - n\hat{\beta}\bar{x} = 0.$$

- La varianza dei residui essere utilizzata per valutare la **bontà di adattamento** del modello ai dati.

- Infatti, quanto più la varianza dei residui è piccola, tanto più la retta di regressione è vicina alle osservazioni.

I residui: media e varianza

- **Proprietà**. La varianza dei residui è sempre minore di quella della variabile risposta. Infatti:

$$\text{var}(y) = \min_{\alpha \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^n (y_i - \alpha)^2 \geq \min_{(\alpha,\beta)\in\mathbb{R}^2} \frac{1}{n}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = \text{var}(r).$$

- **Proprietà**. La varianza dei residui è pari a

$$\text{var}(r) = \text{var}(y) - \frac{\text{cov}(x,y)^2}{\text{var}(x)}.$$

Infatti, usando le proprietà della varianza, otteniamo che

$$\begin{aligned}
\text{var}(r) &= \frac{1}{n}\sum_{i=1}^n [(y_i - \hat{\beta}x_i) - (\bar{y} - \hat{\beta}\bar{x})]^2 = \text{var}(y - \hat{\beta}x) \\
&= \text{var}(y) + \hat{\beta}^2\text{var}(x) - 2\hat{\beta}\text{cov}(x,y) \\
&= \text{var}(y) + \frac{\text{cov}(x,y)^2}{\text{var}(x)} - 2\frac{\text{cov}(x,y)^2}{\text{var}(x)} = \text{var}(y) - \frac{\text{cov}(x,y)^2}{\text{var}(x)}.
\end{aligned}$$

Coefficiente di determinazione R^2

- La varianza dei residui dipende dalla scala del fenomeno osservato. Pertanto per valutare la bontà di adattamento si utilizza spesso l'indice R^2.

- **Coefficiente di determinazione R^2**. Il coefficiente R^2 per un modello di regressione lineare semplice è definito come:

$$R^2 = 1 - \frac{\text{var}(r)}{\text{var}(y)}.$$

- L'indice R^2 misura la frazione di varianza della variabile risposta (**varianza totale**) spiegata dal modello. Si ha pertanto che $0 \leq R^2 \leq 1$.

- Si ha che $R^2 = 0$ se $\text{var}(r) = \text{var}(y)$, ovvero quando il modello non "spiega" la risposta.

- Viceversa, si ha che $R^2 = 1$ quando $\text{var}(r) = 0$, ovvero quando il modello "spiega" perfettamente la risposta.

Coefficiente di determinazione: gli alberi di ciliegio

- Abbiamo calcolato in precedenza le seguenti quantità:

$$\bar{y} = 30.17, \qquad \bar{x} = 13.25,$$
$$\text{var}(x) = 9.53, \qquad \text{cov}(x, y) = 48.24.$$

È inoltre noto che $\sum_{i=1}^n y_i^2 = 36324.99$.

- Pertanto possiamo ottenere

$$\text{var}(y) = \frac{36324.99}{31} - 30.17^2 = 261.54, \qquad \text{var}(r) = 261.54 - \frac{48.24^2}{9.53} = 17.35.$$

- Pertanto, il **coefficiente di determinazione** vale circa

$$R^2 = 1 - \frac{17.35}{261.54} = 0.934,$$

ovvero il modello spiega poco meno del 95% della varianza spiegata.

Correlazione e coefficiente di determinazione

- **Proprietà**. Il coefficiente di determinazione è pari al coefficiente di correlazione al quadrato, infatti:

$$R^2 = 1 - \frac{\text{var}(r)}{\text{var}(y)} = \frac{\text{cov}(x, y)^2}{\text{var}(x)\text{var}(y)} = \text{cor}(x, y)^2.$$

- Questa equivalenza chiarisce che il coefficiente di correlazione (e quindi la covarianza) misura una relazione di tipo lineare.

- Infatti, il coefficiente R^2 e quindi $\text{cor}(x, y)$ catturano la vicinanza dei dati ad una retta.

- **Nota**. Nel caso dei ciliegi, abbiamo ottenuto $R^2 = 0.934$ e $\text{cor}(x, y) = 0.967^2 = 0.935$. Questa leggera discrepanza è dovuta alle varie approssimazioni numeriche effettuate.

- Se avessimo tenuto traccia di un maggior numero di decimali, avremmo ottenuto

$$\text{cor}(x, y) = 0.9671194, \qquad R^2 = 0.9353199.$$

Regressione e correlazione

- Le analogie con l'unità J, dove abbiamo introdotto la covarianza e la correlazione, sono molte.

- Il problema di base è lo stesso (studio delle relazioni tra variabili) e gli "ingredienti" che abbiamo maneggiato pure (medie, varianze e covarianze).

- Nonostante ciò, si noti che esiste una **importante differenza**.

- In questa unità abbiamo considerato l'effetto di una variabile esplicativa su una variabile risposta. Le variabili erano poste in maniera **asimmetrica**, poichè eravamo interessati ad una relazione del tipo `diametro` $\rightarrow$ `volume`.

- Viceversa nell'unità J ci siamo posti in maniera **simmetrica** rispetto alle variabili. Non abbiamo cercato di spiegarne una sulla base di un altra ma abbiamo semplicemente valutato le relazioni intercorrenti.