Statistica I

Esercitazione 5: covarianza, correlazione, regressione lineare semplice

Tommaso Rigon

Università Milano-Bicocca

Descrizione del problema

- **Francis Galton**, scienziato vittoriano e cugino di Charles Darwin, era appassionato dello studio dei **fenomeni ereditari**. Si interessò al problema seguente:

- Come prevedere la statura dei figli adulti conoscendo quella dei genitori?

I dati grezzi

- Nel 1886 Galton raccolse le **stature** di un gruppo consistente di genitori e figli adulti ($n = 465$). È probabile che i soggetti fossero ben nutriti e di classe agiata.

- Alcuni padri compaiono più volte perchè hanno vari figli.

Altezza padre, in cm (x)

```
 [1] 199.390 191.770 191.770 190.500 190.500 190.500 190.500 190.500
 [9] 190.500 187.960 187.960 187.960 187.960 187.960 187.960 185.420
[17] 185.420 185.420 185.420 185.420 185.420 185.420 185.420 185.420
[...]
```

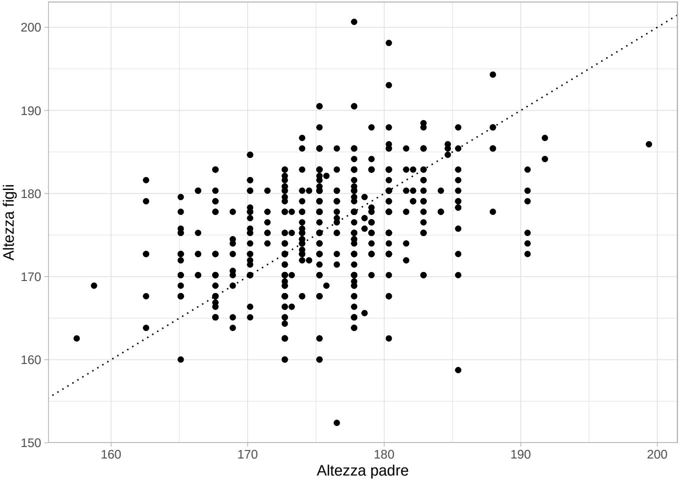
Altezza figlio, in cm (y)

```
 [1] 185.928 186.690 184.150 180.340 179.070 173.990 182.880 175.260
 [9] 172.720 194.310 187.960 185.420 185.420 187.960 177.800 172.720
[17] 170.180 180.340 179.070 182.880 179.070 178.308 178.308 175.768
[...]
```

Domande

- Si rappresentino i dati tramite un diagramma a dispersione.
- Si ottengano medie e varianze di x ed y.
- Si ottengano le **stime ai minimi quadrati** e si disegni la retta di regressione stimata nel diagramma a dispersione.
- Si ottengano i **valori previsti** ed i **residui** della regressione. Si ottenga quindi la varianza residuale.
- Si calcoli R^2 e la correlazione ρ.
- Sulla base di tutte queste informazioni, si risponda alla domanda posta da Galton e si fornisca un'**interpretazione** dei risultati ottenuti.

Diagramma a dispersione



- La linea tratteggiata rappresenta la **bisettrice**.

Alcune statistiche descrittive

- Vista la mole di dati, vengono riportate nel seguito alcune statistiche descrittive. **Nota** in altri esercizi il calcolo di queste quantità è lasciata allo studente.

- Si ricordi anzitutto che la **numerosità campionaria** è $n = 465$. Inoltre

$$\sum_{i=1}^{n} x_i y_i = 14372354, \qquad \sum_{i=1}^{n} x_i = 81694.53, \qquad \sum_{i=1}^{n} y_i = 81766.16.$$

- Inoltre sono rese note le seguenti quantità

$$\sum_{i=1}^{n} x_i^2 = 14368514, \qquad \sum_{i=1}^{n} y_i^2 = 14398590.$$

- Grazie a queste poche statistiche, possiamo procedere con lo svolgimento dell'intero esercizio. Verranno **evidenziati** tra parentesi i valori ad altissima precisione numerica.

Medie e varianze

- Sulla base delle statistiche indicate, otteniamo le **medie**:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{81694.53}{465} = 175.68716, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{81766.16}{465} = 175.8412.$$

- Inoltre, i **momenti secondi** sono pari a

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{14368514}{465} = 30900.03, \qquad \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \frac{14398590}{465} = 30964.71,$$

- Di conseguenza, si ottengono le seguenti **varianze**:

$$\text{var}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = 30900.03 - 175.68716^2 \approx 34.05 \quad (\mathbf{34.05347}),$$

e

$$\text{var}(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = 30964.71 - 175.8412^2 \approx 44.58 \quad (\mathbf{44.58311}).$$

Stima ai minimi quadrati

- Otteniamo inoltre che

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{14372354}{465} = 30908.2882,$$

da cui si può calcolare la **covarianza**:

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = 30908.2882 - 175.68716 \times 175.8412 \approx 15.247.$$

- Utilizzando le quantità appena calcolate, possiamo quindi ottenere la **stima ai minimi quadrati** $(\hat{\alpha}, \hat{\beta})$ di un modello di regressione lineare semplice:

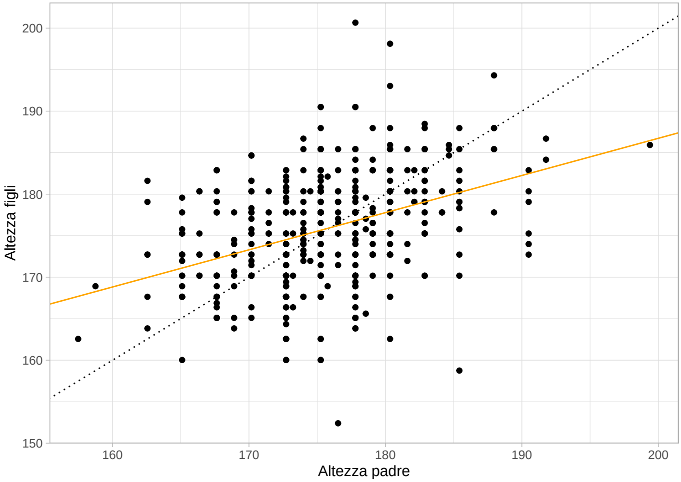
$$\hat{\beta} = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} = \frac{15.247}{34.05} \approx 0.4477, \quad (\mathbf{0.44775}),$$

ed inoltre

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 175.8412 - 0.4477 \times 175.68716 \approx 97.18, \quad (\mathbf{97.17764}).$$

- Di conseguenza, la retta di regressione stimata è $y = 97.18 + 0.4477x$.

Retta di regressione



- La linea tratteggiata rappresenta la **bisettrice**, la linea arancione è invece la **retta ai minimi quadrati**.

Valori previsti e residui

- Si ricordi che i **valori previsti** si ottengono come $\hat{y}_i = 97.18 + 0.4477 x_i$, mentre i **residui** sono pari a $y_i - \hat{y}_i$.

- Per ovvie ragioni di spazio, riportiamo qui solamente i primi 10 valori di 465.

x_i (`altezza padre`)	y_i (`altezza figlio`)	$\hat{y}_i$ (**valori previsti**)	r_i (**residui**)
199.39	185.93	186.45	-0.53
191.77	186.69	183.04	3.65
191.77	184.15	183.04	1.11
190.50	180.34	182.47	-2.13
190.50	179.07	182.47	-3.40
190.50	173.99	182.47	-8.48
190.50	182.88	182.47	0.41
190.50	175.26	182.47	-7.21
190.50	172.72	182.47	-9.75
187.96	194.31	181.34	12.97
$\vdots$	$\vdots$	$\vdots$	$\vdots$

La varianza residuale

- Ci sono due modi per calcolare la **varianza residuale**. Il primo metodo è "diretto" ma dispedioso in termini di tempo, ovvero si considera:

$$\text{var}(r) = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{1}{465}\left[(-0.53)^2 + 3.65^2 + 1.11^2 + \cdots\right] = 37.75613.$$

  Si ricordi infatti che la **media dei residui** è sempre nulla $\bar{r} = 0$.

- In questo caso necessariamente il calcolo è svolto da un **computer**. Tuttavia, usando le formule presentate nell'unità K si può ottenere "a mano" lo stesso risultato:

$$\text{var}(r) = \text{var}(y) - \frac{\text{cov}(x, y)^2}{\text{var}(x)} = 44.58 - \frac{15.247^2}{34.05} = 37.75266 \quad (\mathbf{37.75613}).$$

- Come previsto, la varianza residuale è più piccola della variabilità di y.

Indice R^2 e correlazione

- Anche indice R^2 si può calcolare in **due modi diversi**, entrambi semplici. Il primo è tramite la definizione:

$$R^2 = 1 - \frac{\text{var}(r)}{\text{var}(y)} = 1 - \frac{37.75613}{44.58} \approx 0.153 \quad (\mathbf{0.1531}).$$

- In alternativa, si ricordi che vale la seguente relazione $R^2 = \rho^2$. Pertanto si ottiene in primo luogo:

$$\rho = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} = \frac{15.247}{\sqrt{44.58 \times 34.05}} \approx 0.3913, \quad (\mathbf{0.3913174}),$$

da cui si ottiene

$$R^2 = 0.3913^2 = 0.1531.$$

- Questo valore di R^2 è "grande" o "piccolo"?

- Sebbene vi sia una certa variabilità negli errori di previsione, è comunque interessante notare che una relazione genetica sia ben presente!

Commenti ai risultati

- In primo luogo: si, è possibile prevedere quantomeno in parte l'altezza dei figli sulla base di quella dei padri.

- È interessante confrontare la previsione data dalla **bisettrice** ($\hat{\alpha} = 0$ e $\hat{\beta} = 1$) con la previsione ai **minimi quadrati** ($\hat{\alpha} = 97.18$ e $\hat{\beta} = 0.4477$).

- La bisettrice prevede che l'altezza dei padri sia uguale a quella dei figli. I dati però suggeriscono una situazione diversa.

- Il coefficiente di regressione stimato $\hat{\beta} = 0.4477 < 1$ implica che, in media, a padri bassi avranno figli un po' più alti, mentre padri alti avranno figli un po' più bassi.

- **Francis Galton** diede il nome "**regressione verso la mediocrità**" a questo fenomeno.

- Questo spiega l'orgine del termine **regressione**, che ora viene usato semplicemente per indicare ogni processo di adattamento delle rette ai dati.