Introduzione alla Regressione Lineare

Corso di Informatica Avanzato

Liceo Peano

Febbraio 2018

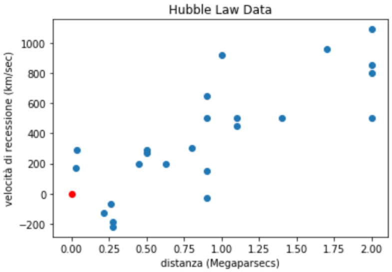
Motivazioni

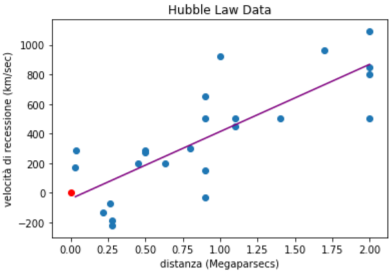
Come dedurre una legge generale partendo da dati sperimentali?

Motivazioni

Come dedurre una legge generale partendo da dati sperimentali?

Ad esempio, come predire la velocitá di oggetti extragalattici basandosi sulla loro distanza dalla Terra: la Legge di Hubble





$$v = H_0 d$$

I dati

I dati sono raccolti in tabelle (file *.csv) come questa

distance (MegaParsec)	recession velocity (km/sec)
.032	170
.034	290
.214	-130
.263	-70
.275	-185
.275	-220
.45	200
.5	290
.5	270
.63	200
.8	300
.9	-30
.9	650
.9	150
.9	500
1.0	920

Modello lineare

L'idea è cercare di rappresentare i dati tramite un modello lineare, ovvero... una retta!

$$\hat{y}_i = a + bx_i$$

Perchè una retta?

- è la funzione più semplice (il rasoio di Occam è sempre da prendere in considerazione)
- si è visto essere in accordo coi dati sperimentali

La funzione di Costo

Definiamo la Funzione di Costo

$$C = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Dove si ha:

- $n \rightarrow$ numero di punti disponibili
- $y_i \rightarrow$ output
- $\hat{y}_i \rightarrow$ output predetto

La funzione di Costo

Definiamo la Funzione di Costo

$$C = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Dove si ha:

- $n \rightarrow$ numero di punti disponibili
- $y_i \rightarrow$ output
- $\hat{y}_i \rightarrow$ output predetto

Siccome stiamo usando un modello lineare l'output predetto diventa

$$\hat{y}_i = a + bx_i$$

La funzione di Costo

Definiamo la Funzione di Costo

$$C = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Dove si ha:

- $n \rightarrow$ numero di punti disponibili
- $y_i \rightarrow$ output
- $\hat{y}_i \rightarrow$ output predetto

Siccome stiamo usando un modello lineare l'output predetto diventa

$$\hat{y}_i = a + bx_i$$

E quindi la funzione di Costo si puó riscrivere come

$$C = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Derivazione

$$C = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Derivazione

$$C = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- L'idea è ricavare i parametri a e b tali da minimizzare la funzione di costo C.

Derivazione

$$C = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- L'idea è ricavare i parametri a e b tali da minimizzare la funzione di costo C.
- Utilizziamo quindi le derivate!

$$\frac{\partial C}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - b) \qquad \frac{\partial C}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i)\, x_i$$

Derivazione

$$C = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- L'idea è ricavare i parametri a e b tali da minimizzare la funzione di costo C.
- Utilizziamo quindi le derivate!

$$\frac{\partial C}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - b) \qquad \frac{\partial C}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i)\, x_i$$

- Imponiamo la condizione di minimizzazione ponendo le derivate a zero:

$$\frac{\partial C}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \qquad \frac{\partial C}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i)\, x_i = 0$$

Richiami su sommatorie

Prima di dedicarci ai calcoli, ricaviamoci alcune formule utilizzando le sommatorie

- $\sum_{i=1}^{n} 1 = n\,; \qquad \sum_{i=1}^{n} c = c \sum_{i=1}^{n} 1 = c\,n\,;$

Richiami su sommatorie

Prima di dedicarci ai calcoli, ricaviamoci alcune formule utilizzando le sommatorie

- $\sum_{i=1}^{n} 1 = n\,; \qquad \sum_{i=1}^{n} c = c \sum_{i=1}^{n} 1 = c\, n\,;$
- Definiamo la media $\rightarrow \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Richiami su sommatorie

Prima di dedicarci ai calcoli, ricaviamoci alcune formule utilizzando le sommatorie

- $\sum_{i=1}^{n} 1 = n\,; \qquad \sum_{i=1}^{n} c = c\sum_{i=1}^{n} 1 = c\,n\,;$
- Definiamo la media $\rightarrow \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Dimostriamo ora che $\sum_{i=1}^{n}\left(\overline{x}^2 - x_i\overline{x}_i\right) = 0$

Richiami su sommatorie

Prima di dedicarci ai calcoli, ricaviamoci alcune formule utilizzando le sommatorie

- $\sum_{i=1}^{n} 1 = n\,; \qquad \sum_{i=1}^{n} c = c \sum_{i=1}^{n} 1 = c\,n\,;$
- Definiamo la media $\rightarrow \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
- Dimostriamo ora che $\sum_{i=1}^{n} \left(\overline{x}^2 - x_i \overline{x}_i \right) = 0$

$$\sum_{i=1}^{n} \left(\overline{x}^2 - x_i \overline{x}_i \right) = \sum_{i=1}^{n} \overline{x}^2 - \sum_{i=1}^{n} \overline{x} x_i$$

Richiami su sommatorie

Prima di dedicarci ai calcoli, ricaviamoci alcune formule utilizzando le sommatorie

- $\displaystyle\sum_{i=1}^{n} 1 = n\,; \qquad \sum_{i=1}^{n} c = c\sum_{i=1}^{n} 1 = c\,n\,;$
- Definiamo la media $\rightarrow \overline{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$
- Dimostriamo ora che $\displaystyle\sum_{i=1}^{n}\left(\overline{x}^2 - x_i\overline{x}_i\right) = 0$

$$\sum_{i=1}^{n}\left(\overline{x}^2 - x_i\overline{x}_i\right) = \sum_{i=1}^{n}\overline{x}^2 - \sum_{i=1}^{n}\overline{x}x_i \quad = \overline{x}^2\sum_{i=1}^{n} 1 - \overline{x}\sum_{i=1}^{n} x_i$$

Richiami su sommatorie

Prima di dedicarci ai calcoli, ricaviamoci alcune formule utilizzando le sommatorie

- $\sum_{i=1}^{n} 1 = n\,; \qquad \sum_{i=1}^{n} c = c \sum_{i=1}^{n} 1 = c\,n\,;$
- Definiamo la media $\rightarrow \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
- Dimostriamo ora che $\sum_{i=1}^{n} \left(\overline{x}^2 - x_i \overline{x}_i\right) = 0$

$$\sum_{i=1}^{n} \left(\overline{x}^2 - x_i \overline{x}_i\right) = \sum_{i=1}^{n} \overline{x}^2 - \sum_{i=1}^{n} \overline{x} x_i \quad = \overline{x}^2 \sum_{i=1}^{n} 1 - \overline{x} \sum_{i=1}^{n} x_i = \overline{x}^2 n - \overline{x} n \overline{x} = 0$$

Il primo parametro

Consideriamo la prima derivata

$$\frac{\partial C}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - b) = 0$$

Il primo parametro

Consideriamo la prima derivata

$$\frac{\partial C}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - b) = 0$$

Facciamo un po' di calcoli, ricordando che a e b sono costanti

$$0 = \sum_{i=1}^{n}(y_i - a - bx_i) = \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} 1 - b\sum_{i=1}^{n} x_i$$

Il primo parametro

Consideriamo la prima derivata

$$\frac{\partial C}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - b) = 0$$

Facciamo un po' di calcoli, ricordando che a e b sono costanti

$$0 = \sum_{i=1}^{n}(y_i - a - bx_i) = \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} 1 - b\sum_{i=1}^{n} x_i = n\overline{y} - na - bn\overline{x}$$

Il primo parametro

Consideriamo la prima derivata

$$\frac{\partial C}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - b) = 0$$

Facciamo un po' di calcoli, ricordando che a e b sono costanti

$$0 = \sum_{i=1}^{n} (y_i - a - bx_i) = \sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} 1 - b \sum_{i=1}^{n} x_i = n\overline{y} - na - bn\overline{x}$$

Il primo parametro

Consideriamo la prima derivata

$$\frac{\partial C}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - b) = 0$$

Facciamo un po' di calcoli, ricordando che a e b sono costanti

$$0 = \sum_{i=1}^{n}(y_i - a - bx_i) = \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} 1 - b\sum_{i=1}^{n} x_i = n\overline{y} - na - bn\overline{x}$$

Ricavando a otteniamo quindi il primo parametro

$$a = \overline{y} - b\overline{x}$$

Il secondo parametro

Consideriamo ora la seconda derivata

$$\frac{\partial C}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - b x_i)\, x_i = 0$$

Il secondo parametro

Consideriamo ora la seconda derivata

$$\frac{\partial C}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0$$

Nei calcoli, utilizziamo $a = \overline{y} - b\overline{x}$

$$0 = \sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = \sum_{i=1}^{n}(y_i - (\overline{y} - b\overline{x}) - bx_i)\,x_i$$

Il secondo parametro

Consideriamo ora la seconda derivata

$$\frac{\partial C}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0$$

Nei calcoli, utilizziamo $a = \overline{y} - b\overline{x}$

$$0 = \sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = \sum_{i=1}^{n}(y_i - (\overline{y} - b\overline{x}) - bx_i)\,x_i$$
$$= \sum_{i=1}^{n}(y_i x_i - \overline{y}x_i) - b\sum_{i=1}^{n}\left(x_i^2 - \overline{x}x_i\right)$$

Il secondo parametro

Consideriamo ora la seconda derivata

$$\frac{\partial C}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0$$

Nei calcoli, utilizziamo $a = \overline{y} - b\overline{x}$

$$\begin{aligned} 0 = \sum_{i=1}^{n}(y_i - a - bx_i)\,x_i &= \sum_{i=1}^{n}(y_i - (\overline{y} - b\overline{x}) - bx_i)\,x_i \\ &= \sum_{i=1}^{n}(y_i x_i - \overline{y}x_i) - b\sum_{i=1}^{n}\left(x_i^2 - \overline{x}x_i\right) \end{aligned}$$

Il secondo parametro

Consideriamo ora la seconda derivata

$$\frac{\partial C}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i)\, x_i = 0$$

Nei calcoli, utilizziamo $a = \overline{y} - b\overline{x}$

$$\begin{aligned} 0 = \sum_{i=1}^{n} (y_i - a - bx_i)\, x_i &= \sum_{i=1}^{n} (y_i - (\overline{y} - b\overline{x}) - bx_i)\, x_i \\ &= \sum_{i=1}^{n} (y_i x_i - \overline{y} x_i) - b \sum_{i=1}^{n} \left(x_i^2 - \overline{x} x_i \right) \end{aligned}$$

Ricavando b si ottiene

$$b = \frac{\sum_{i=1}^{n} (y_i x_i - \overline{y} x_i)}{\sum_{i=1}^{n} \left(x_i^2 - \overline{x} x_i \right)}$$

Cosmesi

$$b = \frac{\sum_{i=1}^{n} \left(y_i x_i - \overline{y} x_i\right)}{\sum_{i=1}^{n} \left(x_i^2 - \overline{x} x_i\right)}$$

Cosmesi

$$b = \frac{\sum_{i=1}^{n}\left(y_i x_i - \overline{y} x_i\right)}{\sum_{i=1}^{n}\left(x_i^2 - \overline{x} x_i\right)}$$

La formula ottenuta puó essere semplificata ricordando che

$$\sum_{i=1}^{n}\left(\overline{x}^2 - \overline{x} x_i\right) = 0 \; ; \qquad \sum_{i=1}^{n}\left(\overline{x}\,\overline{y} - \overline{x} y_i\right) = 0$$

Cosmesi

$$b = \frac{\sum_{i=1}^{n}(y_i x_i - \overline{y} x_i)}{\sum_{i=1}^{n}(x_i^2 - \overline{x} x_i)}$$

La formula ottenuta puó essere semplificata ricordando che

$$\sum_{i=1}^{n}\left(\overline{x}^2 - \overline{x} x_i\right) = 0 \; ; \qquad \sum_{i=1}^{n}(\overline{x}\,\overline{y} - \overline{x} y_i) = 0$$

Sommando la prima al denominatore e la seconda al numeratore si ottiene

Cosmesi

$$b = \frac{\sum_{i=1}^{n} \left(y_i x_i - \overline{y} x_i\right)}{\sum_{i=1}^{n} \left(x_i^2 - \overline{x} x_i\right)}$$

La formula ottenuta puó essere semplificata ricordando che

$$\sum_{i=1}^{n} \left(\overline{x}^2 - \overline{x} x_i\right) = 0 \; ; \qquad \sum_{i=1}^{n} \left(\overline{x}\,\overline{y} - \overline{x} y_i\right) = 0$$

Sommando la prima al denominatore e la seconda al numeratore si ottiene

$$b = \frac{\sum_{i=1}^{n} \left(y_i x_i - \overline{y} x_i - y_i \overline{x} + \overline{x}\,\overline{y}\right)}{\sum_{i=1}^{n} \left(x_i^2 - \overline{x} x_i - x_i \overline{x} + \overline{x}^2\right)}$$

Cosmesi

$$\boxed{b = \frac{\sum_{i=1}^{n} \left(y_i x_i - \overline{y} x_i\right)}{\sum_{i=1}^{n} \left(x_i^2 - \overline{x} x_i\right)}}$$

La formula ottenuta puó essere semplificata ricordando che

$$\sum_{i=1}^{n} \left(\overline{x}^2 - \overline{x} x_i\right) = 0 \; ; \qquad \sum_{i=1}^{n} \left(\overline{x}\,\overline{y} - \overline{x} y_i\right) = 0$$

Sommando la prima al denominatore e la seconda al numeratore si ottiene

$$b = \frac{\sum_{i=1}^{n} \left(y_i x_i - \overline{y} x_i - y_i \overline{x} + \overline{x}\,\overline{y}\right)}{\sum_{i=1}^{n} \left(x_i^2 - \overline{x} x_i - x_i \overline{x} + \overline{x}^2\right)} = \frac{\sum_{i=1}^{n} \left[y_i (x_i - \overline{x}) - \overline{y}(x_i - \overline{x})\right]}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}$$

Cosmesi

$$b = \frac{\sum_{i=1}^{n} (y_i x_i - \overline{y} x_i)}{\sum_{i=1}^{n} (x_i^2 - \overline{x} x_i)}$$

La formula ottenuta puó essere semplificata ricordando che

$$\sum_{i=1}^{n} \left(\overline{x}^2 - \overline{x} x_i\right) = 0 \;; \qquad \sum_{i=1}^{n} (\overline{x}\,\overline{y} - \overline{x} y_i) = 0$$

Sommando la prima al denominatore e la seconda al numeratore si ottiene

$$b = \frac{\sum_{i=1}^{n} (y_i x_i - \overline{y} x_i - y_i \overline{x} + \overline{x}\,\overline{y})}{\sum_{i=1}^{n} \left(x_i^2 - \overline{x} x_i - x_i \overline{x} + \overline{x}^2\right)} = \frac{\sum_{i=1}^{n} [y_i (x_i - \overline{x}) - \overline{y}(x_i - \overline{x})]}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Cosmesi

$$b = \frac{\sum_{i=1}^{n} (y_i x_i - \overline{y} x_i)}{\sum_{i=1}^{n} (x_i^2 - \overline{x} x_i)}$$

La formula ottenuta puó essere semplificata ricordando che

$$\sum_{i=1}^{n} \left(\overline{x}^2 - \overline{x} x_i \right) = 0 \; ; \qquad \sum_{i=1}^{n} (\overline{x}\,\overline{y} - \overline{x} y_i) = 0$$

Sommando la prima al denominatore e la seconda al numeratore si ottiene

$$b = \frac{\sum_{i=1}^{n} (y_i x_i - \overline{y} x_i - y_i \overline{x} + \overline{x}\,\overline{y})}{\sum_{i=1}^{n} \left(x_i^2 - \overline{x} x_i - x_i \overline{x} + \overline{x}^2 \right)} = \frac{\sum_{i=1}^{n} [y_i (x_i - \overline{x}) - \overline{y} (x_i - \overline{x})]}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

da cui otteniamo

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Conclusioni

Abbiamo quindi ricavato i tre ingredienti per studiare la Legge di Hubble:

- l'intercetta a e il coefficiente angolare b del modello lineare $\hat{y}_i = a + bx_i$

$$a = \overline{y} - b\overline{x} \; ; \qquad b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- La stima della bontà della regressione linare: la funzione di Costo

$$C = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

E ora..

Possiamo ora implementare tutto su python!