R - Esercitazione 6

Lorenzo Di Biagio
dibiagio@mat.uniroma3.it

Università Roma Tre

Lunedì 16 Dicembre 2013

Il modello di regressione lineare semplice (I)

Esempi tratti da:
Stock, Watson — Introduzione all'econometria — Pearson
Pieraccini — Fondamenti di inferenza statistica — Giappichelli

Un distretto scolastico riduce la dimensione delle classi delle scuole elementari: qual è l'effetto sul punteggio dei suoi studenti in un test standardizzato?

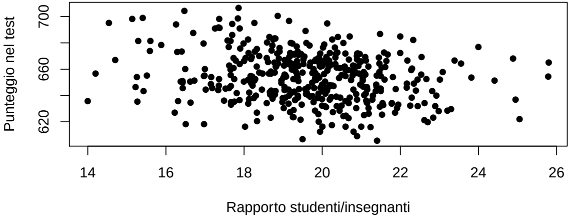
All'interno dell'azienda ACME vi è una relazione tra il numero di ore lavorate e il salario percepito? O il numero di ore lavorate è ininfluente a livello di salario?

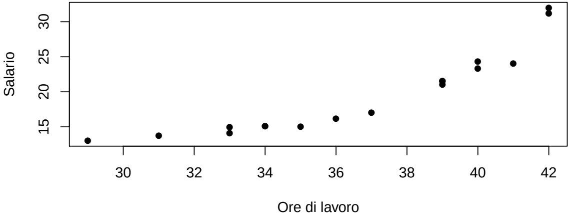
Il modello di regressione lineare semplice (II)

L'analisi di regressione è una tecnica statistica che permette di stimare la relazione tra due variabili X e Y.

Il modello di regressione lineare *postula* una relazione lineare tra X e Y. Il termine noto e il coeff. angolare della retta che mette in relazione X e Y sono una caratteristica ignota della distribuzione congiunta di X e Y: l'analisi di regressione permette di stimare tali parametri a partire da un campione (stima puntuale di parametri, intervalli di confidenza, test di ipotesi...)

Dati campionari:





Il modello di regressione lineare semplice (III)

Assunzioni del modello (caso classico):

i dati osservati sono n coppie (x_i, y_i); consideriamo i valori della variabile X come assegnati (anche se scelti casualmente).

I valori $y_1, \ldots, y_n$ sono valori osservati delle variabili $Y_1, \ldots, Y_n$ con $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ e Y_i tra loro indipendenti.

Per ogni i, $Y_i = \alpha + \beta x_i + U_i$, con U_i v.c. i.i.d. $\sim N(0, \sigma^2)$.

X si dice variabile indipendente, o regressore (regressor)
Y si dice variabile dipendente
$\alpha + \beta X$ è la retta di regressione della popolazione (regression line)
α si chiama intercetta (intercept)
β si chiama pendenza (slope)
$u_i = y_i - \alpha - \beta x_i$ si chiama errore (error)

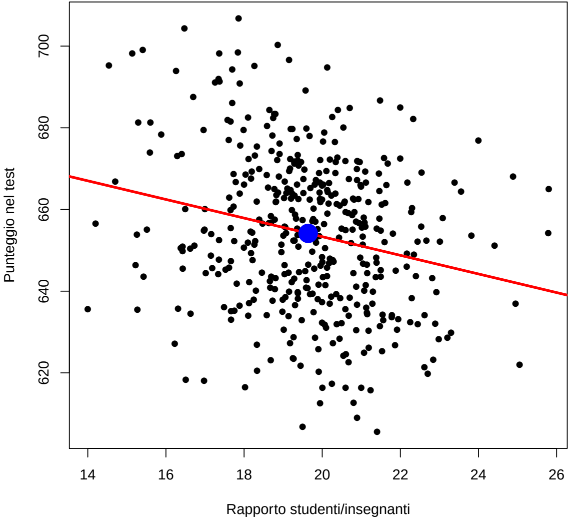
Il modello di regressione lineare semplice (IV)

Dal punto di vista strettamente geometrico: vogliamo costruire una retta $y = \alpha + \beta x$ che meglio approssima la "nuvola" di punti. Come possibile criterio scegliamo i parametri $\hat{\alpha}$ e $\hat{\beta}$ che minimizzano la somma dei quadrati degli scarti tra i valori osservati e i valori "predetti".

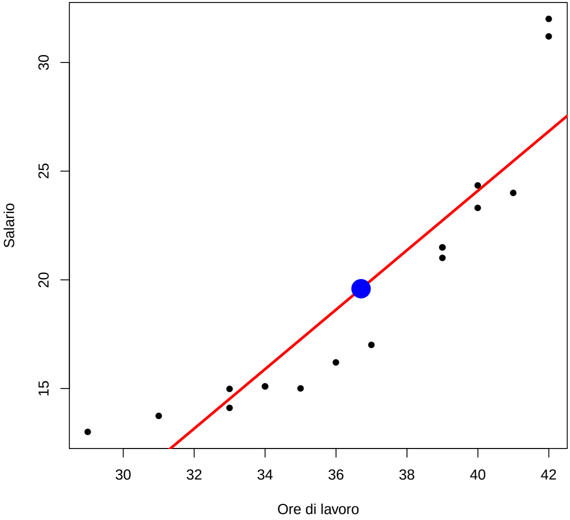
$$f(a, b) := \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$(\hat{\alpha}, \hat{\beta})$ minimizza la funzione f.

In rosso la retta di regressione stimata. Il punto blu ha coordinate $(\overline{x}, \overline{Y})$ (calcolate sul campione).



In rosso la retta di regressione stimata. Il punto blu ha coordinate $(\overline{x}, \overline{Y})$ (calcolate sul campione).



Il modello di regressione lineare semplice (V)

Date le assunzioni del modello di regressione lineare, le stime di α e β basate sul metodo dei minimi quadrati coincidono con le stime di massima verosimiglianza per α e β.

Stimatore per β: $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$.
Quindi $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)$

Dato che conosciamo la distribuzione di $\hat{\beta}$ possiamo fare inferenza su β.
Sia $\hat{U}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$.
Sia $S^2 = \frac{1}{n-2}\sum \hat{U}_i^2$ (stimatore corretto di σ).
Sia $D_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$.
L'inferenza su β è basata sul fatto che

$$\frac{\hat{\beta} - \beta}{S/\sqrt{D_x}} \sim T, \text{con } T \text{ t di Student con } n-2 \text{ gradi di libertà}$$

Il modello di regressione lineare semplice (VI)

Calcolando $\hat{\beta}$ sul campione $y_1, \ldots, y_n$ otteniamo una stima puntuale di β.

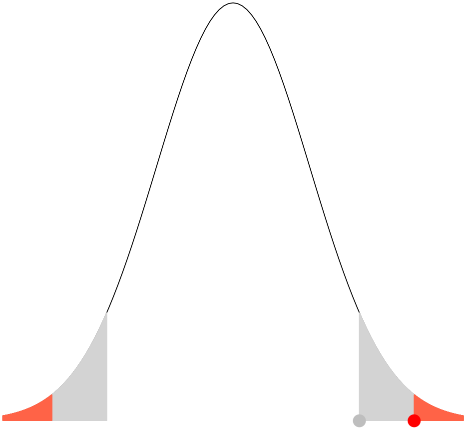
Sia $\gamma = 1 - \alpha$, con $0 < \alpha < 1$. Gli stimatori degli estremi dell'intervallo di confidenza per β al 100γ percento sono: $\hat{\beta} - S/\sqrt{D_x} \cdot t_{1-\frac{\alpha}{2}}; \hat{\beta} + S/\sqrt{D_x} \cdot t_{1-\frac{\alpha}{2}}$.

L'ipotesi nulla $H_0 : \beta = \beta_0$ viene rifiutata, rispetto all'ipotesi $H_1 : \beta \neq \beta_0$, ad un livello di significatività α, se il valore $\tilde{\tau}$ della statistica $\tau = \frac{\hat{\beta}-\beta_0}{S/\sqrt{D_x}}$ calcolata sul campione è maggiore di $t_{1-\frac{\alpha}{2}}$ (se positivo) o minore di $-t_{1-\frac{\alpha}{2}}$ (se negativo). Ovvero se $P(|\tau| > |\tilde{\tau}|) < \alpha$.

Statistica τ sotto l'ipotesi H_0

punto grigio: $t_{1-\frac{\alpha}{2}}$

punto rosso: $\tilde{\tau}$.



Esercizio 1

Aprire il dataset "test.csv", che contiene - per un certo numero di distretti scolastici - il valore del rapporto studenti/insegnanti (stins) e il corrispondente valore del punteggio medio ottenuto dagli studenti in un test standardizzato (punteggio).

1. Calcolare i coefficienti α e β della retta di regressione di "punteggio" su "stins".
2. Porre i dati campionari in un grafico di dispersione e aggiungere al grafico la retta di regressione.
3. Calcolare l'intervallo di confidenza al 97.5% per β.
4. Dire se l'ipotesi $H_0 : \beta = 0$ si può rifiutare ad un livello di significatività dello 0.1%.

Il modello di regressione lineare semplice (VII)

Una volta stabilito se vi è una relazione statisticamente significativa tra le variabili X e Y (con il test su β) è opportuno verificare la bontà dell'adattamento della retta ai punti osservati, chiedendosi quanta parte della variabilità di Y è dovuta alla componente sistematica (riassunta dalla retta) e quanta parte è invece dovuta alla componente accidentale.

In qualche modo ci chiediamo quanto il modello di regressione lineare può spiegare circa la relazione tra X e Y.

Il modello di regressione lineare semplice (VIII)

La devianza campionaria delle Y_i si decompone nella devianza dei valori interpolati e nella devianza della componente accidentale:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}\hat{u}_i^2.$$

Si porrà:

$$r^2 := \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

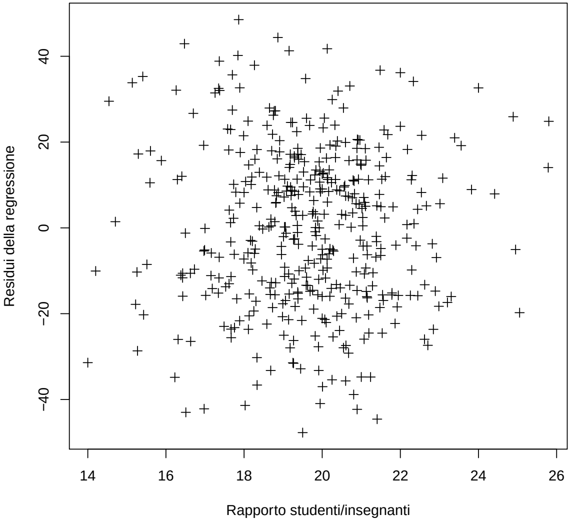
Esercizio 2

In relazione all'Esercizio 1, calcolare l' r^2 della regressione di "punteggio" su "stins".

Il modello di regressione lineare semplice (IX)

Alcune informazioni sulla validità dell'assunto della dipendenza lineare di Y da X si possono ottenere con un grafico dei residui dell'interpolazione (i.e., $y_i - \hat{y}_i$).

Se i residui si dispongono intorno all'asse delle ascisse senza mostrare andamenti particolari allora la relazione tra X e Y può essere ben rappresentata da una forma lineare. Se invece i residui mostrano un andamento curvilineo allora si può essere in presenza di un effetto non lineare tra X e Y.


Residui della regressione
40
20
0
−20
−40
14
16
18
20
22
24
26
Rapporto studenti/insegnanti

Esercizio 3

È stato estratto un campione di 17 lavoratori dell'azienda ACME. Per ognuno di essi sono riportate le ore lavorate (medie settimanali) e il corrispondente salario percepito (netto annuale, in migliaia di euro).

Ore:
$29, 31, 33, 33, 35, 34, 34, 36, 37, 39, 39, 39, 40, 41, 40, 42, 42$
Salario:
$13, 13.75, 14.1, 14.98, 15, 15.1, 15.1, 16.2, 17, 21, 21.5, 21.5, 23.3, 24, 24.35, 31.2, 32$

Verificare che l'ipotesi di indipendenza lineare tra numero di ore lavorate e salario percepito si può rifiutare ad un livello di significatività dell'1%. Generare un grafico dei residui dell'interpolazione e commentarlo.