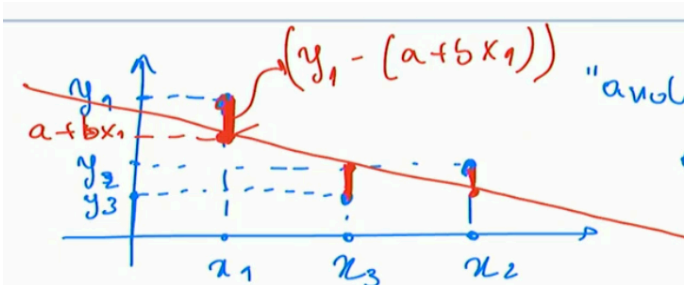
Regressione lineare

Tuesday, 6 June 2023 14:26

- Si lavora con dati accoppiati
  Coefficente correlazione lineare
  $$r_{x,y} = \frac{1}{n-1} * \frac{\sum x_i y_i - n\bar{x}\bar{y}}{S_x S_y}$$
  Questo va a misurare
  - Covarianza campionaria
    $$\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$$
    Questo ci dice se c'è un andamento crescente/decrescente
    Però non ci dice se c'è un collegamento lineare con x e y

  In generale se:
  - $y_i = a + bx_i \Rightarrow |r_{xy}| = 1$
    Quindi solamente quando c'è un legame lineare
  - $r_{xy} = 0 \Rightarrow no\ legame$

  Noi dobbiamo trovare una retta di equazione che meglio approssima i dati del diagramma di dispersione.
  - Per ogni valore si prende lo scarto quadratico e poi si sommano

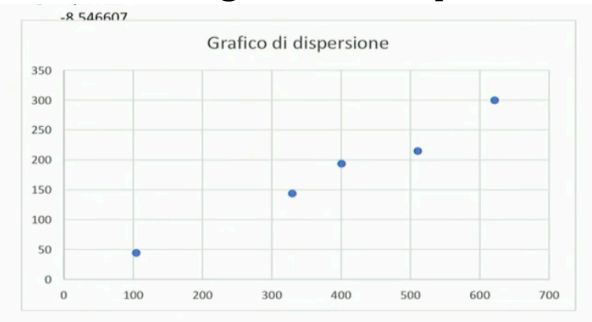
    $$\sum (y_i - a - bx_i)^2$$

    Io ci ho provato a seguirla, vado dritto agli esercizi
- Esercizio
  Abbiamo i seguenti dati:

X	Y
105	44
511	214
401	193

622	299
330	143

a. Tracciare un grafico di dispersione dei dati



b. Calcolare il coef. Correlazione

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x},\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}} \simeq 0.99$$

c. Scrivere l'equazione retta regressione

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = 0.4752$$

$$a = \bar{y} - b\bar{x} \simeq -8{,}5466$$

Formula finale:

$$f(x) = y = bx + a = 0.4752x - 8.5466$$

d. Calcolare il tempo previsto per processare 200, 300, 400, 500 dati
Qui sostituendo $f(x)$ esce questo

- Modello di regressione lineare semplice
  - Andiamo a studiare la dipendenza tra 2 variabili
    - La x è detta ingresso/predittore/input
    - La y è detta uscita/risposta/output
    - Si assuma che il legame sia $Y_i = \alpha + \beta x_i + e_i$
      e_i = errore, supponiamo un errore indipendenti identicamente distribuite

$e_i \sim N(0, \sigma^2)$ → Supposizione
$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ → Calcolato
3 incognite:

- α, β che derivano dalla retta di regressione
- σ^2

Creeremo gli stimatori:

$$\begin{cases} A = \bar{Y} - B\bar{x} \\ B = \dfrac{(x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} \end{cases}$$

Per semplificare la formula noi diciamo che

$$S_{xy} = \sum x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = (n-1)s_n^2$$

$$S_{yy} = \sum Y_i^2 - n\bar{Y}^2$$

$$\begin{cases} B = \dfrac{Sx\,Y}{S_{xx}} \\ A = \bar{Y} - B\bar{x} \end{cases}$$

Quindi ora abbiamo gli stimatori

$$\begin{cases} \hat{\beta} = \dfrac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \\ \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \end{cases}$$

E le leggi sono:

$$B \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

$$A \sim N\left(\alpha, \frac{\sigma^2 \sum x_i^2}{xS_{xx}}\right)$$

Quindi con questo
$Y_i \sim N(0, 1)$
Quante belle formule inutili
Skippo all'esercizio

- Esercizio
  Un automobilista osserva i consumi quando va veloce

Velocità	45	50	55	60	65	70	75
Gallone	24.2	25	23.3	22	21.5	20.6	19.8

Possiamo supporre che valga un modello di regressione semplice che lega miglie percorse con velocità
Vogliamo un intervallo di confidenza per β livello 95%
$\gamma = 100\% - 95\% = 0.05$

$n = 7$

$S_{xx} = \sum x_i^2 - n\bar{x}\bar{y} = 700$

$S_{yy} = \sum Y_i^2 - n\bar{Y}^2 \sim 21.757$

$S_{xy} = -11.9$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -0.17$

$SS_r = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}} \sim 1.527$

Ora calcoliamo intervallo confidenza

$IC = \hat{\beta} \pm t_{n-2,\frac{\alpha}{2}} * \sqrt{\frac{SS_r}{(n-2)S_{xx}}} = -0.17 \pm t_{5,\frac{0.05}{2}} \sqrt{\frac{1527}{3500}} = -0.17 \pm 0.054$

$= (-0.224, -0.116)$

Quindi sappiamo che all'aumentare della velocità percorro meno strada con la stessa quantità di carburante

- Io sto soffrendo nel cercare di ascoltare sta prof, sto facendo veramente tanta fatica. Ultima lezione, ce la posso fare, ce la posso fare.

$A \sim N\left(\alpha, \frac{\sigma^2 \sum x_i^2}{nS_{xx}}\right), \qquad B \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$

$\frac{SS_r}{\sigma^2} \sim X^2(n-2)$

Ipotesi di β

  - β=0

    Vuol dire che la risposta non dipende dal predittore

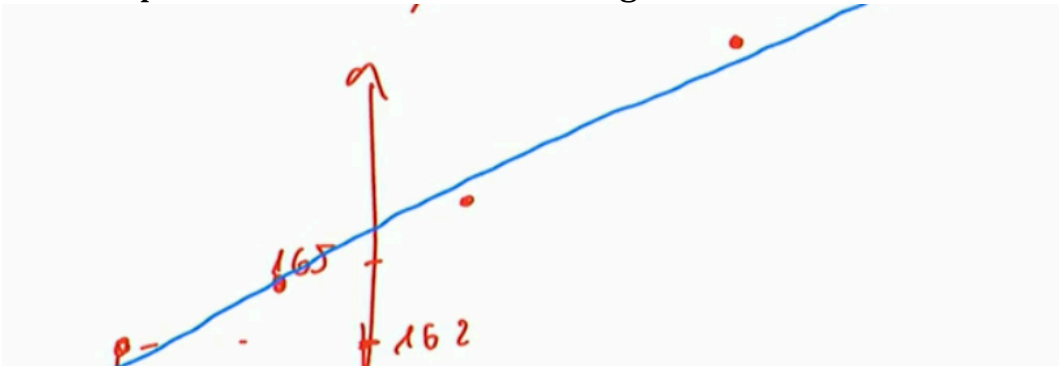
    Aka il modello di regressione lineare semplice non è un modello buono

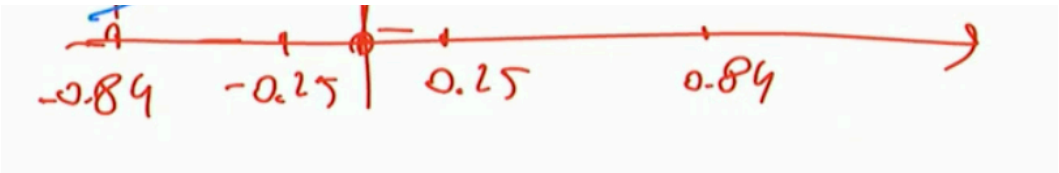
- Verificare che n=4 dati 165, 162, 171, 167 sono in buona approssimazione normale provenienti da una popolazione normale.

Prima li ordino

162, 165, 167, 171

Per comprenderlo dobbiamo fare un grafico:





A seconda del grafico comprendiamo se è una normale.
Si va ad intuito lol

- Abbiamo i seguenti dati

X	Y
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165

  - Fare grafico dispersione
    Iniziamo espandendo la tabella

X	Y	X^2	y^2	X*Y
42	130	1764	16900	5460
46	115	2116	13225	5290
42	148	1764	21904	6216
71	100	5041	10000	7100
80	156	6400	24336	12480
74	162	5476	26244	11988
70	151	4900	22801	10570
80	156	6400	24336	12480
85	162	7225	26244	13770
72	158	5184	24964	11376
64	155	4096	24025	9920
81	160	6561	25600	12960
41	125	1681	15625	5125
61	150	3721	22500	9150
75	165	5625	27225	12375

$\bar{x} = 65.6$
$\bar{y} = 146.2$

$$S_{xx} = \sum x_i^2 - n\bar{x}^2 = 3303.6$$

$$S_{yy} = \sum y_i^2 - n\bar{y}^2 = 5312.4$$

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 2399.2$$

Ora con questi dati possiamo calcolare:

$$\beta = \frac{S_{xy}}{S_{xx}} = 0.7049$$

$$\alpha = \bar{y} - \hat{\beta}\bar{x} = 99.95$$

$$SS_r = \frac{S_{xx}S_{yy} - S_x^2 Y}{S_{xx}} = 3621.202$$

$$f(x) = y = bx + a = 0.7049x + 99.95$$

- Retta regressione e rappresentarla
- Grafico dei residui
- Costruire IC al 95 e 99% β

$$\left(\hat{\beta} \mp \sqrt{\frac{SS_r}{(n-2)S_{xx}}}\, t_{n-2,\frac{\alpha}{2}}\right)$$

$$= 0.7049 \pm \sqrt{\frac{5312.4}{13 * 3403.6}}\, t_{13,\frac{1-0.95}{2}} = 0.7049 \pm 0.6190$$

E notiamo che $0 \notin IC$

Ora calcoliamo 99%

$$0.7049 \pm \cdots * t_{13,\frac{1-0.9}{2}} = (-0.1569, 1.5666)$$

E notiamo che $0 \in IC$

- Test l'ipotesi che non ci sia legame tra età e pressione arteriosa

$$R^2 = 1 - \frac{SS_r}{S_{YY}} = 0.32$$

- Calcolare il coefficente di determinazione
- Calcolare coefficente di correlazione campionaria

$$r = \pm\sqrt{032} \rightarrow r = 0.567$$