Università degli Studi di Napoli "Parthenope"

Corso di Laurea in Statistica e Informatica per l'Azienda, la Finanza e le Assicurazioni (SIAFA)

STATISTICA
II MODULO

Sergio LONGOBARDI
longobardi@uniparthenope.it

L'inferenza nel modello di regressione

Modello di regressione

Il modello che esprime la dipendenza lineare di Y da X nella popolazione

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- **Y è la variabile dipendente (o risposta)**
- **X è la variabile esplicativa (o indipendente)**
- **ε è il termine di errore**
- **β_0 e β_1 sono i parametri incogniti che descrivono la relazione di dipendenza nella popolazione**
- **β_0 e β_1 devono essere stimati sulla base del campione**

Specificazione del modello di regressione lineare

Per ogni osservazione i=1…n

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Assunzioni del modello:

1) ε_i sono v.c. indipendenti con valore atteso $E(\varepsilon_i)=0$

2) $Var(\varepsilon_i) = \sigma^2$

per ogni i=1,2…n (ipotesi di varianza costante o omoschedasticità)

3) i valori x_i della variabile esplicativa X sono noti senza errore

4) ε_i è una v.c. Normale, cioè $\varepsilon_i \sim N(0; \sigma^2)$

Valore atteso e varianza di Y

Dalle ipotesi fatte discende che Y è una variabile casuale Normale.

Il suo valore atteso condizionato al valore X=xi è dato da:

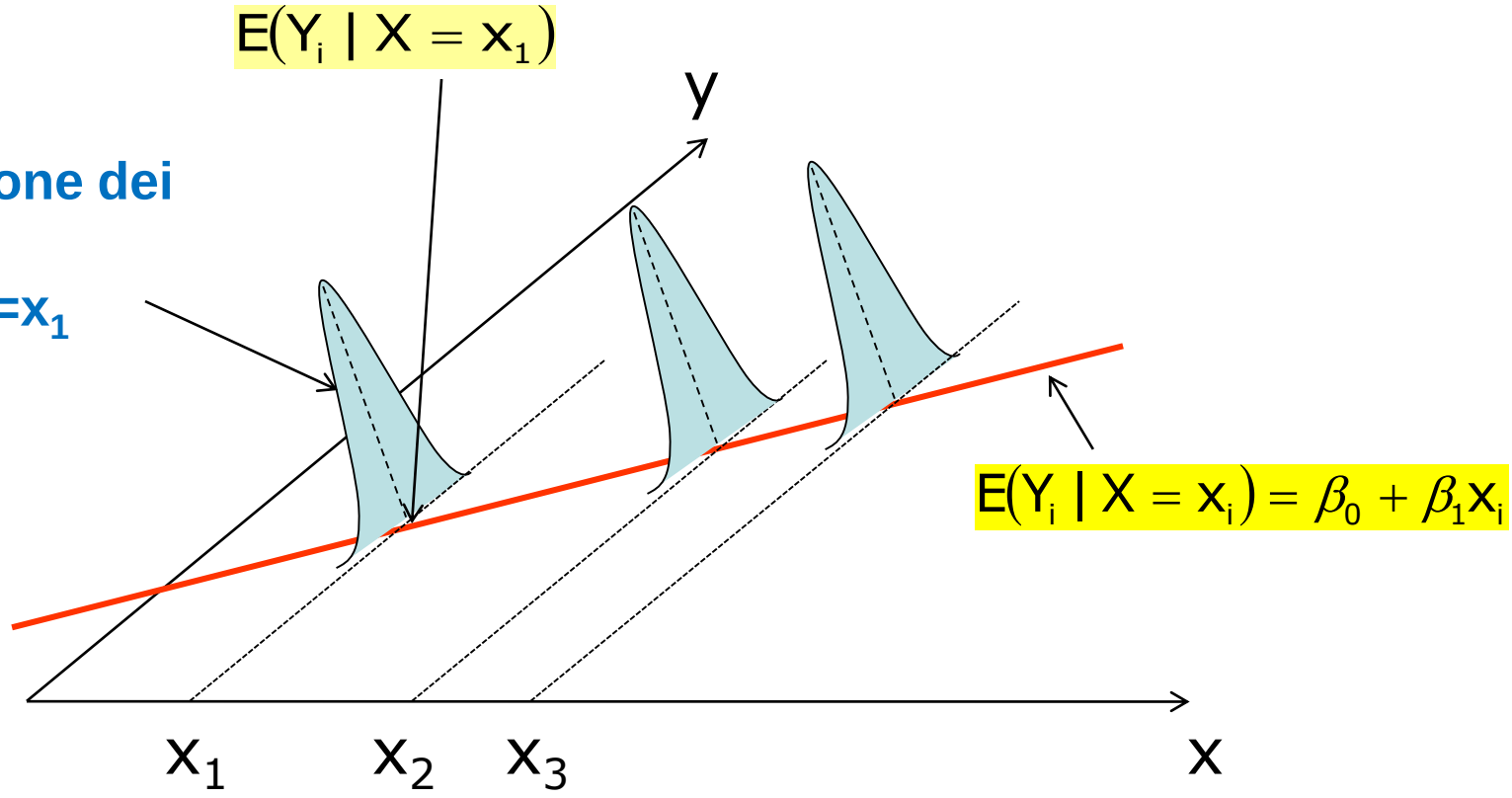
$$E(Y_i \mid X = x_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) =$$
$$= (\beta_0 + \beta_1 x_i + E(\varepsilon_i)) = \beta_0 + \beta_1 x_i$$

La varianza è data da:

$$V(Y_i) = V(\beta_0 + \beta_1 x_i + \varepsilon_i) = V(\varepsilon_i) = \sigma^2$$

$$\Rightarrow \quad Y_i \sim N(\beta_0 + \beta_1 x_i ; \sigma^2)$$

Illustrazione delle assunzioni del modello



Stime puntuali

coefficiente angolare della retta (pendenza)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

intercetta all'origine

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

la risposta media

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Inferenza sulla regressione

- Le stime dei coefficienti del modello β_0 e β_1 e la stima della risposta media dipendono dal campione osservato.
- Al variare dei campioni si generano le v.c. stimatori dei coefficienti di regressione e della risposta media che indichiamo con B_0, B_1 e $\hat{Y}_i$

Stimatori dei coefficienti di regressione

Proprietà 1: B_0 e B_1 sono **corretti**:

$$\mathbf{E(B_0)= \beta_0 \quad E(B_1)= \beta_1}$$

Proprietà 2: nella classe degli stimatori corretti che sono funzioni lineari di Y_i gli stimatori dei minimi quadrati B_0 e B_1 sono i più efficienti (Teorema Gauss-Markov).

Proprietà 3:

$$\text{var}(B_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right); \quad \text{var}(B_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Stime e stimatori nella regressione

Parametro	Stima	Stimatore		
		Stima	Media	Varianza
β_0	$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$	B_0	β_0	$\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$
β_1	$\widehat{\beta_1} = \sigma_{xy}/\sigma_x^2$	B_1	β_1	$\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Distribuzione campionaria di B_0 e B_1

$$\mathbf{B_1} \sim N\left(\beta_1; \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \Rightarrow Z = \frac{\mathbf{B_1} - \beta_1}{\sqrt{\sigma^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim N(0;1)$$

$$\mathbf{B_0} \sim N\left(\beta_0; \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right) \Rightarrow Z = \frac{\mathbf{B_0} - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}} \sim N(0;1)$$

Stima della varianza σ^2

$$\sigma^2 = V(\varepsilon_i)$$

σ^2 non è nota. Lo stimatore corretto di σ^2 è dato da:

$$s^2 = \frac{\sum_{i=1}^{n}(\hat{e}_i - E(\hat{e}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}(\hat{e}_i)^2}{n-2}$$

La radice quadrata di s^2 si definisce come *errore standard di regressione* e misura la dispersione dei punti osservati intorno alla retta di regressione

$$s = \sqrt{s^2}$$

Distribuzione campionaria di B0 e B1

Quando al posto di σ^2 utilizziamo la stima corretta s^2

$$\frac{B_1 - \beta_1}{\sqrt{s^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim t_{n-2} \Rightarrow \frac{B_1 - \beta_1}{s(B_1)} \sim t_{n-2}$$

Errore standard di B_1

$$\frac{B_0 - \beta_0}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}} \sim t_{n-2} \Rightarrow \frac{B_0 - \beta_0}{s(B_0)} \sim t_{n-2}$$

Errore standard di B_0

Distribuzione campionaria di B_0 e B_1

$$s(B_0) = \sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}$$

Errore standard di B_0

$$s(B_1) = \sqrt{s^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Errore standard di B_1

Intervalli di confidenza dei coefficienti di regressione

Intervallo di confidenza per β_1 al livello di confidenza 1-α:

$$P\left(B_1 - t_{\alpha/2;n-2} \cdot s(B_1) < \beta_1 < B_1 + t_{\alpha/2;n-2} \cdot s(B_1)\right) = 1 - \alpha$$

Stima puntuale

valore t

$S(B_1)$

Intervallo di confidenza per β_0 al livello di confidenza 1-α

$$P\left(B_0 - t_{\alpha/2;n-2} \cdot s(B_0) < \beta_0 < B_0 + t_{\alpha/2;n-2} \cdot s(B_0)\right) = 1 - \alpha$$

Esempio consumo-reddito

X	Y	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
600	310	-89	-116	10324	13456
650	320	-79	-66	5214	4356
670	340	-59	-46	2714	2116
690	380	-19	-26	494	676
700	400	1	-16	-16	256
720	420	21	4	84	16
760	430	31	44	1364	1936
780	440	41	64	2624	4096
790	470	71	74	5254	5476
800	480	81	84	6804	7056
716	399			34860	39440

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

$$\hat{\beta}_0 = 399 - 0{,}884 * 716 = -233{,}8$$

$$\hat{\beta}_1 = \frac{34860}{39440} = 0{,}884$$

Stima per intervallo

$\hat{e}_i = y_i - \hat{y}_i$	$\hat{e}_i^2$
13,53	183,04
-20,66	427,01
-18,34	336,42
3,98	15,85
15,14	229,28
17,46	305,01
-7,89	62,26
-15,57	242,36
5,59	31,29
6,75	45,62
	1878,14

$$\sum_{i=1}^{n} \hat{e}_i^2$$

$$s^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2} = \frac{1878,14}{8} = 234,77$$

Errore standard della regressione

$$s = \sqrt{234,77} = 15,32$$

Errore standard di B_1

$$s(B_1) = \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{234,77}{39440}} = 0,078$$

Errore standard di B_0

$$s(B_0) = \sqrt{s^2\left(\frac{1}{n} + \frac{\overline{x^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} =$$

$$= \sqrt{234,77\left(\frac{1}{10} + \frac{716^2}{39440}\right)} = 55,45$$

Stime per intervallo. Dati consumo-reddito

Al livello di confidenza: 1-a=95% $t_{0,025;\,8} = \pm 2{,}31$

I.C. β_1

$$P(0{,}88 - 2{,}31 \times 0{,}078 < \beta_1 < 0{,}88 + 2{,}31 \times 0{,}078) = 0{,}95$$

Stima puntuale **valore t** **$S(B_1)$**

$$P(0{,}71 < \beta_1 < 1{,}06) = 0{,}95$$

I.C. β_0

$$P(-233{,}85 - 2{,}31 \times 55{,}45 < \beta_0 < -233{,}85 + 2{,}31 \times 55{,}45) = 0{,}95$$

$$P(-361{,}73 < \beta_0 < -105{,}26) = 0{,}95$$

Verifica di ipotesi sui coefficienti di regressione

Possiamo essere interessati a verificare:

$$H_0 : \beta_1 = b_1 \qquad H_1 : \beta_1 \neq b_1$$

$$H_0 : \beta_1 = b_1 \qquad H_1 : \beta_1 > b_1$$

$$H_0 : \beta_1 = b_1 \qquad H_1 : \beta_1 < b_1$$

Gli stessi sistemi di ipotesi si possono specificare per il parametro intercetta β_0

Verifica di ipotesi sui coefficienti di regressione

Generalmente si verifica che:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Se accetto H_0 vuol dire che non c'è nella popolazione una significativa relazione di dipendenza lineare di Y da X

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

Se accetto H_0 vuol dire che la relazione nella popolazione può essere rappresentata mediante una retta passante per l'origine

Statistica test

per $H_0:\beta_1=b_1$

$$\frac{B_1 - b_1}{s(B_1)} \sim t_{n-2}$$

per $H_0:\beta_1=0$

$$\frac{B_1}{s(B_1)} \sim t_{n-2}$$

$H_0:\beta_1=0$
$H_1:\beta_1\neq 0$

Al livello di significatività α, accetto H_0 se il valore della statistica test calcolato sul campione cade nell'area di accettazione dell'ipotesi nulla, cioè se

$$-t_{\alpha/2;n-2} < \frac{\hat{\beta}_1}{s(B_1)} < t_{\alpha/2;n-2}$$

ESERCIZIO

In un ipermercato di Napoli è stata svolta un'indagine per rilevare il prezzo del pane negli ultimi cinque mesi (in euro al Kg) e le quantità consumate in media in un giorno (in Kg)

prezzo	quantità
1,65	210
1,67	198
1,68	176
1,69	175
1,7	174

Stimare la retta di regressione che mette in relazione la quantità in funzione del prezzo

Stima dei coefficienti di regressione

	Coefficienti
Intercetta	1529
Variabile X 1	-800

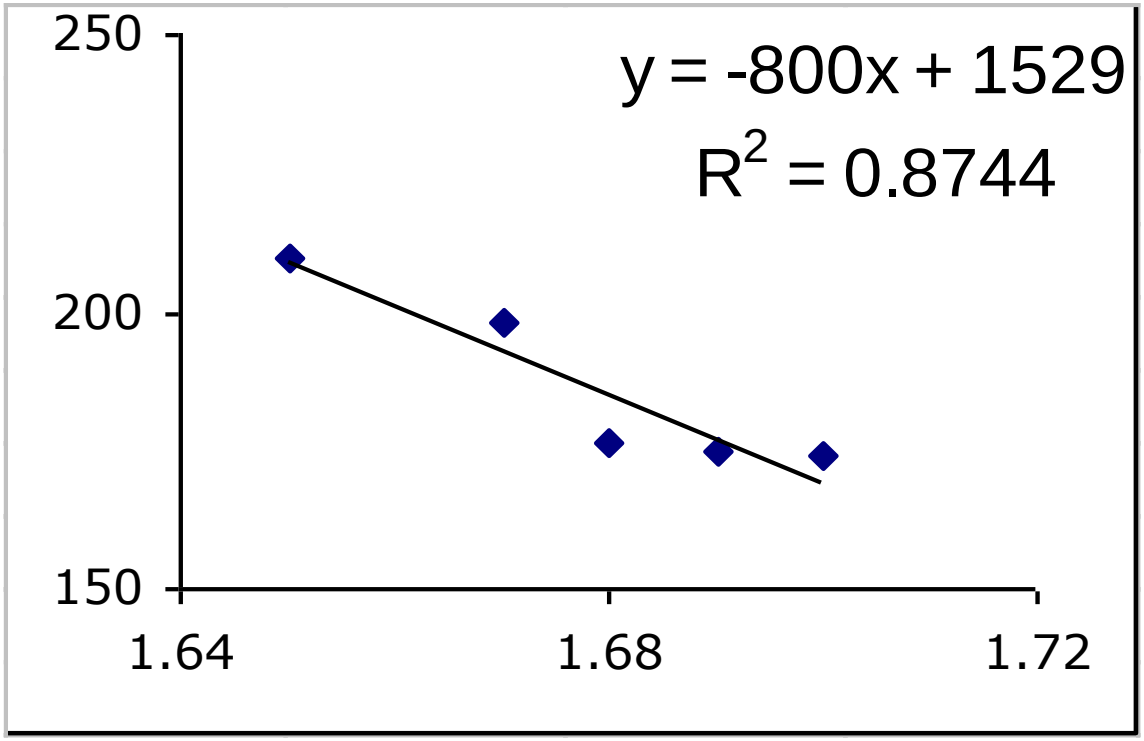
La retta stimata è Y=1529-800X
Un aumento di 1€ del prezzo al Kg del pane fa diminuire la quantità media di pane consumato giornalmente di 800 kg

Statistica della regressione	
R al quadrato	0,87
Errore standard	6,73

La dipendenza lineare è forte (R^2=0,87)

s=6,73

Rappresentazione punti osservati e retta di regressione



Test t

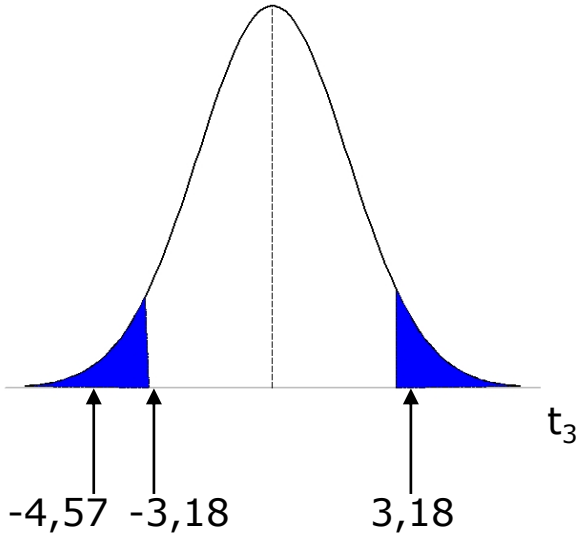
	Coefficienti	*Errore standard*	*Stat t*	*p-value*
Intercetta	1529	293,69	5,21	0,01
Variabile X 1	-800	175,02	-4,57	0,02

Per verificare $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{s(B_1)} = \frac{-800}{175{,}02} = -4{,}57$$

Al livello α=0,05 $t_{0,025;3} = \pm 3{,}18$



-4,57<-3,18

Si rifiuta H_0

C'è evidenza sufficiente per concludere che la quantità consumata di pane dipende linearmente dal prezzo