

DATA MINING PER IL MARKETING (63 ore)

Marco Riani

mriani@unipr.it

Sito web del corso

<http://www.riani.it/DMM>

REGRESSIONE INFERENZIALE

Introduzione agli elementi aleatori

	N. dipendenti (X)	Vendite in milioni di € (Y)		Prezzi in Euro (x)	Vendite (Y)
A	10	1,9	A	1.55	410
B	18	3,1	B	1.60	380
C	20	3,2	C	1.65	350
D	8	1,5	D	1.60	400
E	30	6,2	E	1.50	440
F	12	2,8	F	1.65	380
G	14	2,3	G	1.45	450
			H	1.50	420

Introduzione agli elementi aleatori

- Le vendite sono dovute in parte ai prezzi e in parte a fattori di natura aleatoria e perciò sono esse stesse delle v.c.
- Al contrario I dipendenti e/o i prezzi non sono v.c. poiché sono del tutto prevedibili dalla compagnia che li stabilisce

Introduzione agli elementi aleatori

- Una successione di valori fissi
- x_1, x_2, \dots, x_n
- a cui sono associate n v.c. indipendenti
- Y_1, Y_2, \dots, Y_n
- Il punto cruciale consiste nel descrivere in modo appropriato tali v.c.
- $E(Y_i)$? $\text{var}(Y_i)$? Distribuzione di Y_i ?

Assunzioni su Y_i

- Tutte le osservazioni sono caratterizzate dallo stesso grado di incertezza
- $\text{var}(Y_i) = \sigma^2$ $i=1, 2, \dots, n$
- σ^2 è un parametro incognito da stimare
- $\text{cov}(Y_i, Y_j)=0$ $i \neq j$

Assunzioni su Y_i

- $E(Y_i) = \mu_i$ $i=1, 2, \dots, n$
- i valori osservati della variabili dipendente provengono da n distribuzioni di probabilità con medie incognite
- Ip. le medie delle distribuzioni variano linearmente con la variabili indipendente
- $\mu_i = E(Y_i) = \alpha + \beta x_i$

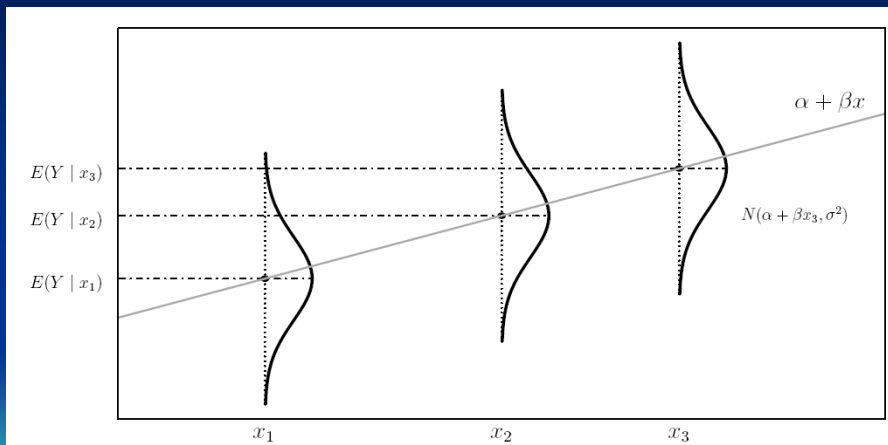
Assunzioni su Y_i (continua)

- Ip: $\mu_i = E(Y_i) = \alpha + \beta x_i$
- Questa ipotesi equivale ad affermare che i punti $(x_1, \mu_1), (x_2, \mu_2), \dots, (x_n, \mu_n)$ stiano tutti su una retta con parametri α, β
- Oss: questa assunzione non implica che tutti i punti (x_i, y_i) stiano sulla retta ma che i valori medi delle distribuzioni da cui i punti provengono verificano l'equazione della retta

Interpretazione di α e β

- I parametri α e β rappresentano l'intercetta ed il coeff. angolare della retta sulla quale giacciono le medie incognite delle distribuzioni Y_1, \dots, Y_n

Interpretazione di α e β



Osservazione

- Dato il modello di regressione
- $Y_i = \alpha + \beta x_i + \varepsilon_i$
- L'ip: $\mu_i = E(Y_i) = \alpha + \beta x_i$
- equivale ad affermare che
- $E(\varepsilon_i) = 0$

Stima dei parametri

- I parametri da stimare sono
- $\alpha, \beta, \mu_1, \mu_2, \dots, \mu_n, \sigma^2$
- La conoscenza di α, β consente di ricostruire tutte le medie incognite $\mu_1, \mu_2, \dots, \mu_n$

Stime di α e β

- Pensando di ripetere più volte l'esperimento che ha generato le osservazioni y_1, \dots, y_n , per valori fissi di x_1, \dots, x_n si ottiene una distribuzione campionaria di valori

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})Y_i}{\sum (x_i - \bar{x})^2}$$

Stime di μ_j

Coeff. di regressione campionari e nella popolazione

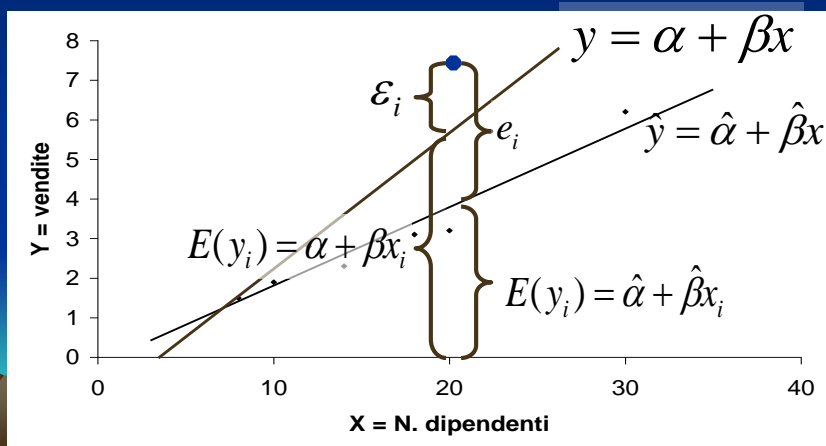
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$$

Coeff. di regressione campionari e nella popolazione

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$$



Stima di σ^2

- σ^2 = dispersione verticale attorno alla retta che unisce i valori medi delle popolazioni
- Dato che $\sigma^2 = E(\varepsilon_i^2)$
- Dato che e_i è una stima di ε_i sembra naturale utilizzare come stimatore di σ^2 la seguente espressione

$$s^2 = \frac{\sum (y_i - \hat{\mu}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

Stima di σ^2

- Utilizziamo gli scostamenti dalle medie delle popolazioni

$$s^2 = \frac{\sum (y_i - \hat{\mu}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

Stima di σ^2

- Excel definisce s come “errore standard nella stima di Y ” (se_y nel linguaggio di Excel)

$$s = \sqrt{\frac{\sum e_i^2}{n-2}}$$

- Si può ottenere direttamente tramite la funzione `ERR.STD.YX`.

Funzione regr.lin

- Ordine in cui vengono restituite le statistiche aggiuntive di regressione dalla funzione di Excel REGR.LIN

	A	B	C	D	E	F	G
1	b_k	b_{k-1}	...	b_2	b_1	b_0	a
2	s_k	s_{k-1}	...	s_2	s_1	s_0	s_a
3	R2	se_y					
4	F	d_f					
5	SS_{reg}	SS_{resid}					
6							

Ip. aggiuntiva

- Le distribuzioni Y_i sono normali
- y_1 è una realizzazione di $Y_1 \sim N(\mu_1, \sigma^2)$
- y_2 è una realizzazione di $Y_2 \sim N(\mu_2, \sigma^2)$
- ...
- y_n è una realizzazione di $Y_n \sim N(\mu_n, \sigma^2)$

- Y_1, Y_2, \dots, Y_n sono indipendenti

Richiami sulla v.c. normale

- se $Y \sim N(\mu, \sigma^2)$
- $Z = (Y - \mu) / \sigma \sim N(0, 1)$
- $\Pr(-1.96 < Z < 1.96) = 0.95$
- $aY + b \sim N(b + \mu, a^2\sigma^2)$

Richiami sulla costruzione degli int. di confidenza

$$Z(\bar{X}_n) = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\text{var}(\bar{X}_n)}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$\Pr\left(-1.96 < \frac{\bar{X} - E(\bar{X})}{\sqrt{\text{var}(\bar{X})}} < 1.96\right) = 0.95$$

$$\Pr\left(-1.96 < \frac{\bar{X} - \mu}{\sqrt{\text{var}(\bar{X})}} < 1.96\right) = 0.95$$

$$\Pr\left(\bar{X} - 1.96\sqrt{\text{var}(\bar{X})} < \mu < \bar{X} + 1.96\sqrt{\text{var}(\bar{X})}\right) = 0.95$$

Obiettivo

Costruire intervalli di confidenza e
test di verifica d'ipotesi sul coeff.
angolare

$$\hat{\beta}$$

Studio della distribuzione di

$$\hat{\beta}$$

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})Y_i}{\sum (x_i - \bar{x})^2}$$

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$y_i = \hat{\alpha} + \hat{\beta}x_i + e_i$$

Studio della distribuzione di

$$\hat{\beta}$$

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})Y_i}{\sum (x_i - \bar{x})^2}$$

$$E(\hat{\beta}) = ?$$

$$E(\hat{\beta}) = \beta$$

$$\text{var}(\hat{\beta}) = ?$$

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Varianza di beta cappello

$$\text{var}(\hat{\beta}) = \text{var} \left(\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

$$\text{var}(\hat{\beta}) = \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \text{var} \left(\sum_{i=1}^n (x_i - \bar{x})Y_i \right)$$

$$\text{var}(\hat{\beta}) = \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \left(\sum_{i=1}^n \text{var}(x_i - \bar{x})Y_i \right)$$

Varianza di beta cappello

$$\text{var}(\hat{\beta}) = \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \left(\sum_{i=1}^n \text{var}(x_i - \bar{x}) Y_i \right)$$

$$\text{var}(\hat{\beta}) = \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \left(\sum_{i=1}^n (x_i - \bar{x})^2 \text{var} Y_i \right)$$

$$\text{var}(\hat{\beta}) = \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \left(\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 \right)$$

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Al posto di σ^2 sostituiamo il suo stimatore

$$\text{Stima}(\text{var}(\hat{\beta})) = s^2(\hat{\beta}) = \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- La radice quadrata della stima della varianza di uno stimatore è l'errore standard (standard error, SE) dello stimatore

$$s_{\hat{\beta}} = SE(\hat{\beta}) = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Interpretazione dello standard error di beta cappello

- Rappresenta l'errore quadratico medio che si commette quando si stima il coefficiente di regressione con le formule dei minimi quadrati

Funzione regr.lin

- Lo standard error di beta cappello è riportato nella zona di output di regr.lin all'incrocio della seconda riga e prima colonna)

	A	B	C	D	E	F	G
1	b_k	b_{k-1}	...	b_2	b_1	b_0	a
2	s_k	s_{k-1}	...	s_2	s_1	s_0	s_a
3	R ²	se_y					
4	F	d_f					
5	SS _{reg}	SS _{resid}					
6							

Studio della distribuzione di $\hat{\alpha}$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

$$E(\hat{\alpha}) = ?$$

$$E(\hat{\alpha}) = \alpha$$

$$\text{var}(\hat{\alpha}) = ?$$

$$\text{var}(\hat{\alpha}) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]$$

Esercizio: nell'esempio dei 7 supermercati calcolare lo standard error di beta cappello e alpha cappello

$$s_{\hat{\beta}} = SE(\hat{\beta}) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = 0.025$$

$$s_{\hat{\alpha}} = SE(\hat{\alpha}) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} = 0.44$$

Costruzione di intervalli di confidenza dei parametri

Punto di partenza: lo scostamento standardizzato di beta capello ha una distribuzione $N(0,1)$

$$\Pr\left(-Z_\gamma < \frac{\hat{\beta} - E(\hat{\beta})}{\sqrt{\text{var}(\hat{\beta})}} < Z_\gamma\right) = 1 - \gamma$$

- Se $1 - \gamma = 0.95$

$$\Pr(-1.96 < \frac{\hat{\beta} - E(\hat{\beta})}{\sqrt{\text{var}(\hat{\beta})}} < 1.96) = 0.95$$

$$\Pr(-1.96 < \frac{\hat{\beta} - E(\hat{\beta})}{\sqrt{\text{var}(\hat{\beta})}} < 1.96) = 0.95$$

$$\Pr\left(-1.96 < \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}} < 1.96\right) = 0.95$$

Problema: σ^2 è ignoto (occorre sostituire il suo stimatore s^2)

Studio della distribuzione di s^2

- Si può dimostrare che $E(S^2) = \sigma^2$ e che

$$\frac{(n-2) s^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

Sostituendo al posto di σ^2 il suo stimatore

$$\Pr \left(-1.96 < \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}}} < 1.96 \right) = 0.95$$

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}}} = \frac{(\hat{\beta} - \beta) / \sqrt{\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{N(0, 1)}{\sqrt{\frac{\chi_{n-2}^2}{n-2}}}$$

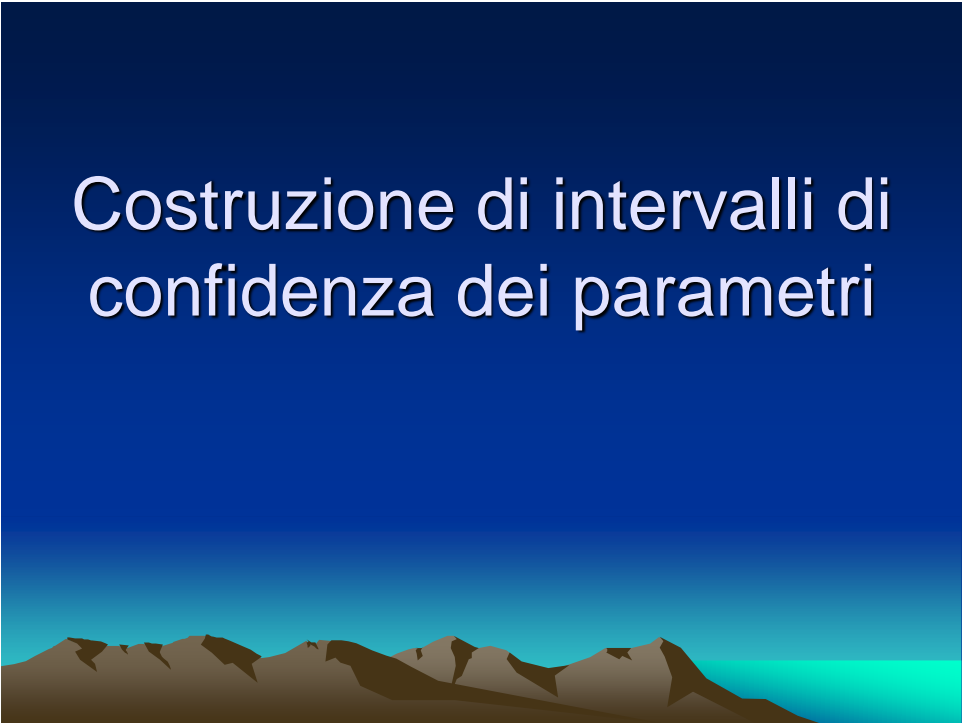
Costruzione di un intervallo di confidenza per il coeff. angolare

$$\Pr \left(\hat{\beta} - t_\gamma s_{\hat{\beta}} \leq \beta \leq \hat{\beta} + t_\gamma s_{\hat{\beta}} \right) = 1 - \gamma$$

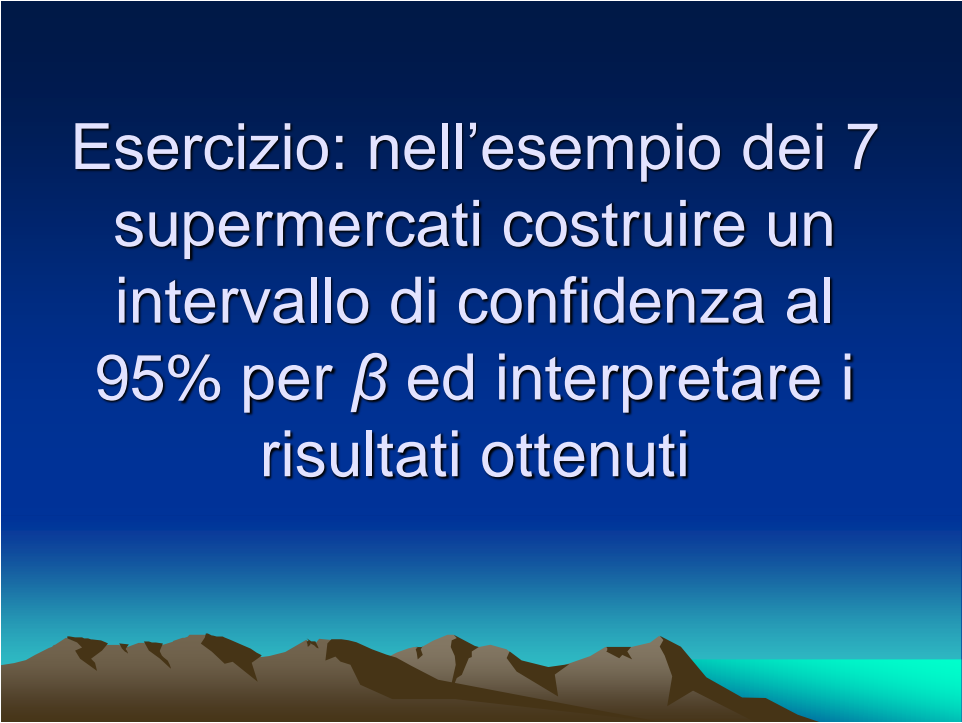
- Dove t_γ è il quantile (percentile) associato alla distribuzione T di Student con $(n-2)$ gradi di libertà tale che (v. p. 44)

dove t_γ è il percentile di una variabile T di Student con $(n-2)$ gradi di libertà tale per cui $\Pr(T \leq -t_\gamma) = \Pr(T \geq t_\gamma) = \gamma/2$.

Costruzione di intervalli di confidenza dei parametri



Esercizio: nell'esempio dei 7
supermercati costruire un
intervallo di confidenza al
95% per β ed interpretare i
risultati ottenuti



Costruzione di un intervallo di confidenza al 95% per il coeff. angolare

$$\Pr(\hat{\beta} - t_{\gamma} s_{\hat{\beta}} \leq \beta \leq \hat{\beta} + t_{\gamma} s_{\hat{\beta}}) = 1 - \gamma$$

- $t_{0.05}(5) = +2.5706$ (=INV.T(0.05;5))
- (Oss: $\Pr.(T(5) > 2.5706) = 0.025$)
- $\Pr(0.198 - 2.5706 \times 0.0253 < \beta < 0.198 - 2.5706 \times 0.0253) = 0.95$
- $\Pr(0.133 < \beta < 0.263) = 0.95$

Interpretazione

- L'intervallo di confidenza del coefficiente di regressione, con probabilità uguale a 0.95, va da 0,133 a 0,263.
- Questo significa che nell'universo di riferimento, all'aumento di un dipendente può corrispondere un aumento delle vendite compreso tra 133 mila Euro e 263 mila Euro circa (con probabilità del 95%).
- Oss: l'intervallo è piuttosto ampio e questo dipende dalla ridotta numerosità campionaria (solo 7 supermercati).

Intervallo di confidenza per l'intercetta

Costruzione di un intervallo di confidenza al 95% per l'intercetta

$$\Pr(\hat{\alpha} - t_{\gamma} s_{\hat{\alpha}} \leq \alpha \leq \hat{\alpha} + t_{\gamma} s_{\hat{\alpha}}) = 1 - \gamma$$

- $t_{0.05}(5) = +2.5706$ (=INV.T(0.05;5))
- (Oss: $\Pr.(T(5) > 2.5706) = 0.025$)
- $\Pr(-1.31 < \alpha < 0.96) = 0.95$

Costruzione di un intervallo di confidenza al 95% per σ^2

Punto di partenza

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

$$\Pr\left(\chi^2_{0.025}(n-2) \leq \frac{(n-2)s^2}{\sigma^2} \leq \chi^2_{0.975}(n-2)\right) = 0.95$$

$$\Pr\left(\frac{(n-2)s^2}{\chi^2_{0.975}(n-2)} \leq \sigma^2 \leq \frac{(n-2)s^2}{\chi^2_{0.025}(n-2)}\right) = 0.95$$

- per trovare $\chi^2_{0.975}$ utilizzo
=INV.CHI(0.025;5)=0.83
- per trovare $\chi^2_{0.025}$ utilizzo
=INV.CHI(0.975;5)=12.83

$$\Pr\left(\frac{(n-2)s^2}{\chi^2_{0.975}(n-2)} \leq \sigma^2 \leq \frac{(n-2)s^2}{\chi^2_{0.025}(n-2)}\right) = 0.95$$

- $\Pr(0.08 < \sigma^2 < 1.30) = 0.95$

Costruzione di test di ipotesi per

$$\alpha \quad \beta \quad \sigma^2$$

Dato che

$$\frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \sim t_{n-2}$$

Sotto $H_0: \beta = 0$

$$\frac{\hat{\beta}}{SE(\hat{\beta})} \sim t_{n-2}$$

Funzione regr.lin

- Ordine in cui vengono restituite le statistiche aggiuntive di regressione dalla funzione di Excel REGR.LIN

	A	B	C	D	E	F	G
1	b_k	b_{k-1}	...	b_2	b_1	b_0	a
2	s_k	s_{k-1}	...	s_2	s_1	s_0	s_a
3	R2	se_y					
4	F	d_f					
5	SS_{reg}	SS_{resid}					
6							

Calcolo delle statistiche t con Excel e del relativo p-value

- p value → Funzione distrib.T

Esercizio: nell'esempio dei 7 supermercati testare $H_0:\beta=0$, trovare il relativo p-value ed interpretare il risultato del test

$$t_{\beta}=7.82 \quad \text{p-value} = 0.000548$$

Interpretazione : rifiuto decisamente l'ipotesi nulla

Esercizio: nell'esempio dei 7 supermercati testare $H_0:\alpha=0$, trovare il relativo p-value ed interpretare il risultato del test

$$t_{\alpha}=0.39 \quad \text{p-value} = 0.714$$

Interpretazione : non posso rifiutare l'ipotesi nulla

Esercizio

- Calcolare

$$\text{Cov} [\hat{\alpha}, \hat{\beta}]$$

Intervallo di confidenza delle
previsioni con il metodo dei
minimi quadrati

Strumenti necessari propedeutici

$$\begin{aligned} \text{var}(aX_1 + bX_2) &= E((aX_1 + bX_2) - (a\mu_1 + b\mu_2))^2 \\ &= E(a(X_1 - \mu_1) + b(X_2 - \mu_2))^2 \\ &= E(a^2(X_1 - \mu_1)^2 + b^2(X_2 - \mu_2)^2 + 2ab(X_1 - \mu_1)(X_2 - \mu_2)) \\ &= a^2 \text{var}(X_1) + b^2 \text{var}(X_2) + 2ab \text{cov}(X_1, X_2). \end{aligned}$$

Calcolo della var. dell'errore di previsione

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$$

$$\begin{aligned} e_0 &= y_0 - \hat{y}_0 \\ &= \alpha + \beta x_0 + \varepsilon_0 - \hat{\alpha} - \hat{\beta}x_0 \\ &= (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})x_0 + \varepsilon_0 \end{aligned}$$

$$\text{Var}[e_0] = \text{Var}[\hat{\alpha}] + (x_0)^2 \text{Var}[\hat{\beta}] + 2x_0 \text{Cov}[\hat{\alpha}, \hat{\beta}] + \text{Var}[\varepsilon_0]$$

Calcolo della var. dell'errore di previsione

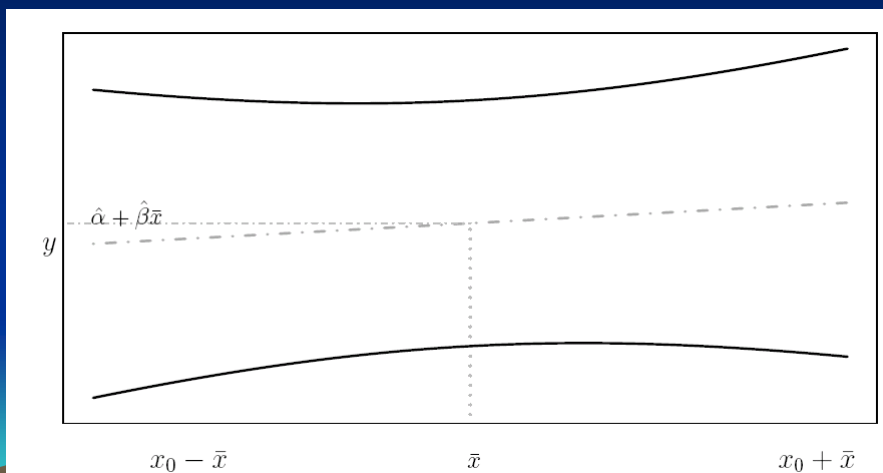
$$Var [e_0] = Var [\hat{\alpha}] + (x_0)^2 Var [\hat{\beta}] + 2x_0 Cov [\hat{\alpha}, \hat{\beta}] + Var [\varepsilon_0]$$

$$Var [e_0] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} + \frac{(x_0)^2}{\sum_i (x_i - \bar{x})^2} - 2x_0 \frac{\bar{x}}{\sum_i (x_i - \bar{x})^2} + 1 \right]$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right]$$

Osservando attentamente quest'ultima relazione si può notare che la varianza dell'errore di previsione è minima quando $x_0 = \bar{x}$ e cresce in modo non lineare all'allontanarsi di x_0 da \bar{x} .

Bande di confidenza dell'errore di previsione (p. 55)



Costruzione di un intervallo di confidenza per y_0

- Tenendo presente che

$$\frac{e_0 - E(e_0)}{\sqrt{\text{var}(e_0)}} \sim N(0,1)$$

$$\frac{e_0 - E(e_0)}{\sqrt{\hat{\text{var}}(e_0)}} \sim T(n-2)$$

$$\frac{e_0}{\sqrt{\hat{\text{var}}(e_0)}} \sim T(n-2)$$

$$\frac{y_0 - \hat{y}_0}{\sqrt{\hat{\text{var}}(e_0)}} \sim T(n-2)$$

Costruzione di un intervallo di confidenza per y_0

$$\frac{y_0 - \hat{y}_0}{\sqrt{\hat{\text{var}}(e_0)}} \sim T(n-2)$$

$$\Pr\left(-t_\gamma < \frac{y_0 - \hat{y}_0}{\sqrt{\hat{\text{var}}(e_0)}} < t_\gamma\right) = 1 - \gamma$$

$$\Pr\left[\hat{y}_0 - t_\gamma \sqrt{\hat{\text{var}}[e_0]} \leq y_0 \leq \hat{y}_0 + t_\gamma \sqrt{\hat{\text{var}}[e_0]}\right] = 1 - \gamma$$

vedi p. 167

Esercizio: per un numero di dipendenti pari a 16 costruire un intervallo di previsione delle vendite al 95%

$$\Pr \left[\hat{y}_0 - t_{\gamma/2} \sqrt{\widehat{Var} [e_0]} \leq y_0 \leq \hat{y}_0 + t_{\gamma/2} \sqrt{\widehat{Var} [e_0]} \right] = 1 - \gamma$$

$$Var [e_0] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right]$$

$$\Pr(3 - 2.57 \times 0.4966 < y_0 < 3 + 2.57 \times 0.4966) = 0.95$$

$$\Pr(1.72 < y_0 < 4.28) = 0.95$$

