

Introduzione all'analisi di regressione

Dottorato in Metodologia della ricerca

a.a. 2008/09

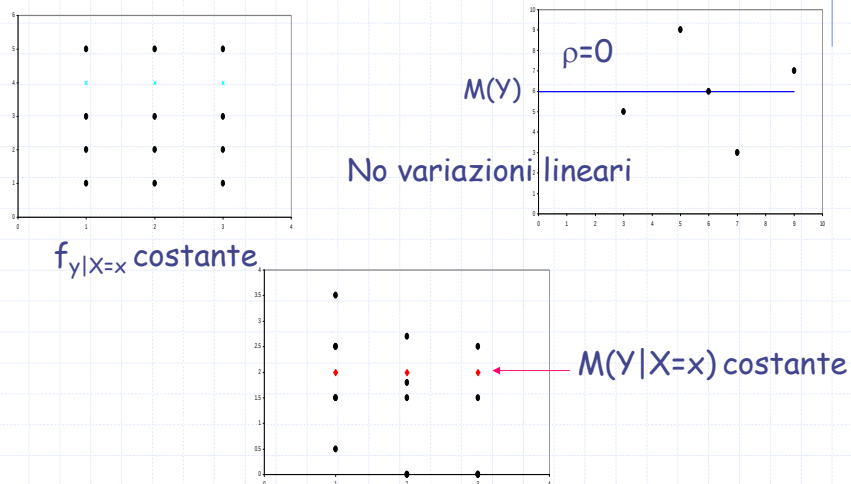
Carla Rampichini
rampichini@ds.unifi.it

Testi di riferimento

- ◆ Cook R. D. e Weisberg S. (1999) *Applied Regression Including Computing and Graphics*, John Wiley and Sons.
- ◆ Greene W. (2003), *Econometric Analysis*, Upper Saddle River (N.J.), Prentice Hall.
- ◆ Montgomery D.C., Peck E.A., Vining G.G. (2001), *Introduction to linear regression analysis*, 3. Ed., New York: John Wiley & Sons.
- ◆ Moore D. S. e McCabe G., 1999, *Introduction to the Practice of Statistics*, 3rd edition, New York: Freeman.
- ◆ Piccolo D., 1998, *Statistica*, Bologna: Il Mulino.

Relazione tra caratteri

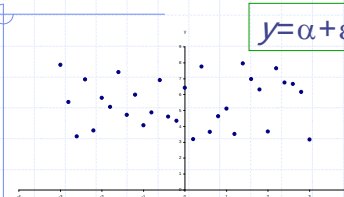
Al variare di X in S_x , si osserva una variazione sistematica della distribuzione di Y ?



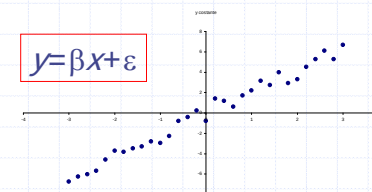
Modello di regressione

La relazione tra x e $M(Y|x)$ è sintetizzata attraverso un modello parametrico

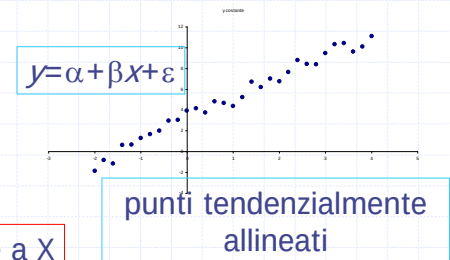
$$y = g(x, \theta) + \varepsilon$$



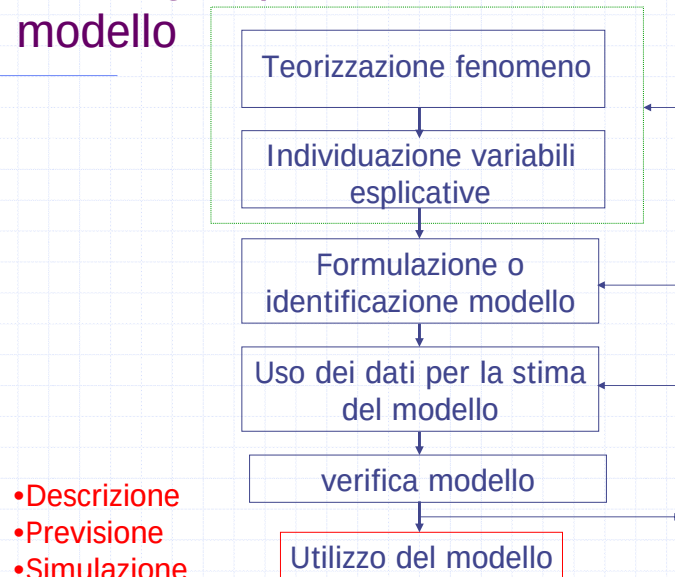
Y tendenzialmente costante



Y tendenzialmente proporzionale a X

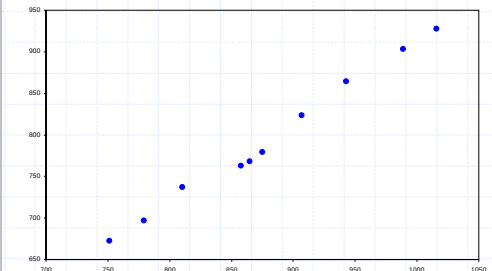


Schema logico per la costruzione di un modello



Esempio: reddito disponibile e consumi (Greene, 2001)

anno	reddito	consumo
1970	751.6	672.1
1971	779.2	696.8
1972	810.3	737.1
1973	864.7	767.9
1974	857.5	762.8
1975	874.9	779.4
1976	906.8	823.1
1977	942.9	864.3
1978	988.8	903.2
1979	1015.7	927.6
media	879.2	793.4
varianza	6719.2	6497.2
ds	82.0	80.6
covarianza		6579.9
rho		0.9959



X: reddito in bilioni di \$ (valuta '72)
Y: consumi in bilioni di \$ (valuta '72)

Modello di regressione lineare semplice

- ◆ Forma funzionale $y_i = \alpha + \beta x_i + \varepsilon_i \quad i=1, \dots, n$
- ◆ $E(\varepsilon_i) = 0$
- ◆ Omoschedasticità: $\text{Var}(\varepsilon_i) = \sigma^2$, costante
- ◆ Residui incorrelati: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$
- ◆ Regressori e errori incorrelati
 $\text{cov}(x_i, \varepsilon_i) = 0$
- ◆ $\varepsilon_i \sim N(0, \sigma^2)$

Segue esempio consumo/reddito

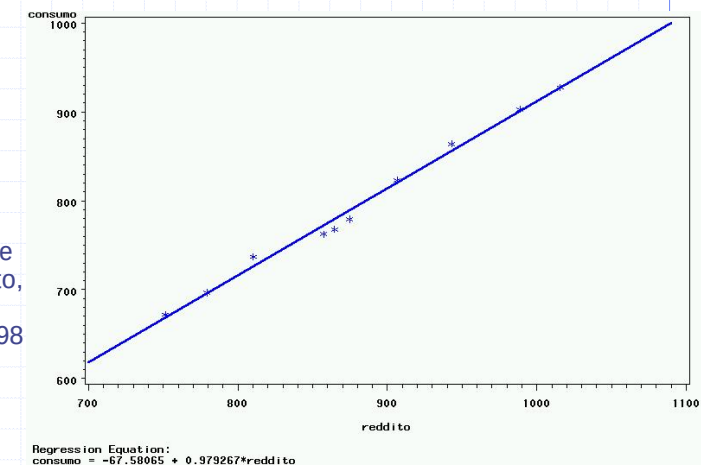
Stima OLS

$a = -67.58$

consumi per
reddito=0

$b = 0.979$

per ogni bilione
in più di reddito,
i consumi
crescono di 0.98
bilioni



consumi_green.sas

Scomposizione della varianza

◆ $SST = SSR + SSE$

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

◆ $R^2 = SSR/SST$ indice di determinazione

◆ $R^2 = \rho^2$

◆ $R^2 = \text{corr}(\hat{y}_i, y_i) = \frac{\sum_{i=1}^n [(\hat{y}_i - \bar{y})(y_i - \bar{y})]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$

Bontà di adattamento

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	64435	64435	959.92	<.0001
Error	8	537	67.13		
Corrected Total	9	64972			

SSR / SST = 0.9917

Root MSE = 8.19303

Dependent Mean = 793.43000

Coeff Var = 1.03261

R-Square = 0.9917

Adj R-Sq = 0.9907

TEST F

$H_0: \beta = 0, H_a: \beta \neq 0$

$$F_{[1, (N-2)]} = \frac{SSR/1}{SSE/(N-2)}$$

Proprietà degli stimatori OLS

Teorema di Gauss-Markov

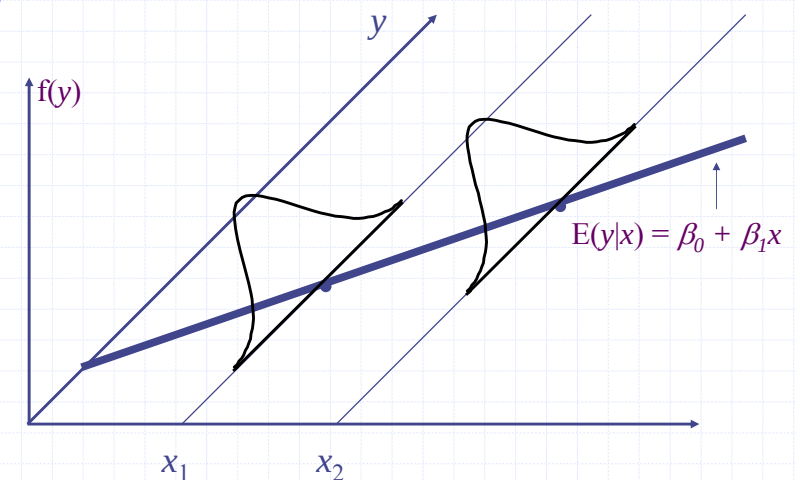
Lo stimatore OLS $\mathbf{b}$ è lo stimatore lineare e corretto a varianza minima di β

• Se $\varepsilon \sim N(0, \sigma^2) \rightarrow \mathbf{b} | \mathbf{X} \sim N(\beta, \text{Var}(\mathbf{b}))$

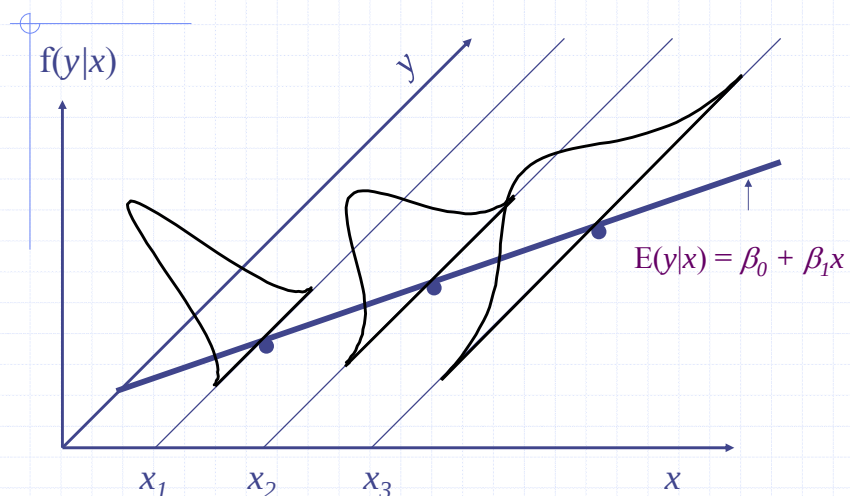
Proprietà utile per definire test e intervalli di confidenza

$$\text{Var}(\mathbf{b}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$E(y|x)$ è una funzione lineare di x , dove per ogni x la distribuzione di y è centrata intorno a $E(y|x)$



In caso di eteroschedasticità...



Test e intervalli di confidenza

◆ Stima di σ^2

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

◆ Statistica per il test $H_0: \beta = 0$

$$t = \frac{b}{SE_b} \sim t_{(n-2)}$$

◆ Intervallo di confidenza al livello c per β

$$b \pm t^* SE_b$$

t^* valore di t di Student con $(n-2)$ g.l.: $P(-t^* < t < t^*) = c$

segue esempio consumo/reddito

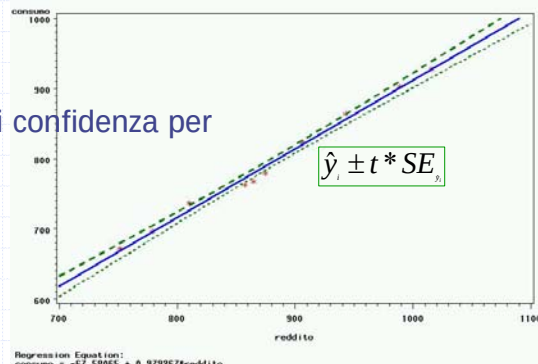
Parameter Estimates

$$Var(b) = \frac{s^2}{\sum (x_i - \bar{x})^2}$$

$$t = \frac{b}{SE_b} \sim t_{(n-2)}$$

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-67.58065	27.91071	-2.42	0.0418
reddito	reddito	1	0.97927	0.03161	30.98	<.0001

Intervallo di confidenza per
 $y_i = a + bx_i$



Regressione multipla: matrice di dati

ID	y	x_1	x_2	...	x_j	...	x_p
1	y_1	x_{11}	x_{21}	...	x_{j1}	...	x_{p1}
2	y_2	x_{12}	x_{22}	...	x_{j2}	...	x_{p2}
...	...	...	...	...	...	...	...
i	y_i	x_{1i}	x_{2i}	...	x_{ji}	...	x_{pi}
...	...	...	...	...	...	...	...
n	y_n	x_{1n}	x_{2n}	...	x_{jn}	...	x_{pn}

Regressione lineare multipla

Modello:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

$i=1, \dots, n$

β_j coefficienti di regressione **parziali**
 incremento di Y per una variazione unitaria di x_j , quando le altre X sono costanti

Ipotesi modello di regressione lineare multipla

- Modello: $Y = X\beta + \varepsilon$
- $E(\varepsilon) = \mathbf{0} \rightarrow E(Y) = X\beta$
- $E(\varepsilon \varepsilon') = \sigma^2 I$
omoschedasticità + incorrelazione
- $E(\varepsilon | X) = \mathbf{0} \rightarrow \text{cov}(\varepsilon, X) = \mathbf{0}$
- X Matrice $n \times p$ a pieno rango
- $\varepsilon \sim MN(\mathbf{0}, \sigma^2 I)$

Metodo di stima

Minimi quadrati

$$\min_{\beta_0, \dots, \beta_p} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_p x_{pi})^2$$

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Coefficienti di regressione

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \text{regr. multipla}$$

$$y_i = b_0 + b_1 x_{1i} + v_i$$

regr. semplice

$$y_i = c_0 + b_2 x_{2i} + \gamma_i$$

In generale $b_1 \neq \beta_1$ e $b_2 \neq \beta_2$, se $r_{x_1, x_2} = 0 \rightarrow \beta_1 = b_1$

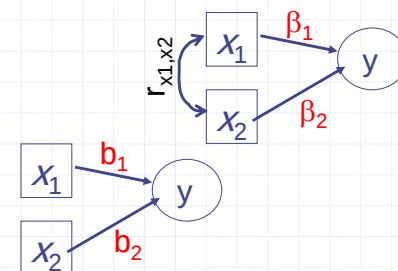
coefficienti

netti

$$\beta_1 = \frac{\sigma_y}{\sigma_{x_1}} \frac{r_{y, x_1} - r_{y, x_2} r_{x_1, x_2}}{1 - r_{x_1, x_2}^2}$$

lordi

$$b_1 = \frac{\text{cov}(y, x_1)}{\sigma_{x_1}^2} = \frac{\sigma_y}{\sigma_{x_1}} r_{y, x_1}$$



Coefficiente di correlazione multipla

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad 0 \leq R^2 \leq 1$$

Indica la % di variabilità totale spiegata dalla regressione

$R^2=1 \rightarrow$ adattamento perfetto $y_i = \hat{y}_i$

$R^2=0 \rightarrow$ non c'è relazione lineare $y_i = \bar{y}$

in tal caso $\bar{y}$ minimizza $\sum_{i=1}^n (y_i - \hat{y}_i)^2$

Esempio: regressione multipla

giorno	x1	x2	y
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	5.8
5	50	2	4.3
6	80	1	6.8
7	75	3	6.6
8	80	2	5.9
9	90	3	7.6
10	90	2	6.1

X_1 miglia percorse
 X_2 numero di consegne
 Y tempo di percorrenza

Come varia il tempo giornaliero di consegna?

Matrice di correlazione

	x1	x2	y
x1	1.00000	0.28046	0.77330
x2	0.28046	1.00000	0.63511
y	0.77330	0.63511	1.00000

Indici univariati

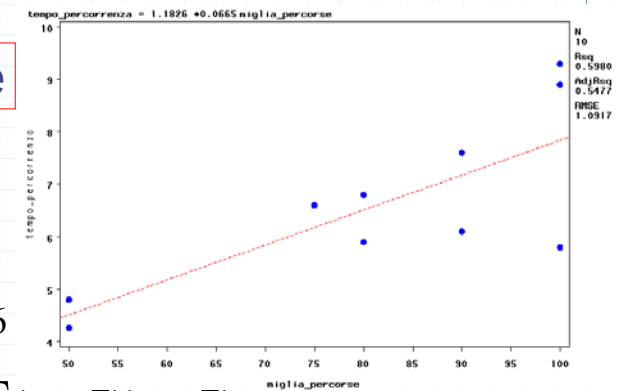
Variabile	Media	Varianza	Std Dev
x1	81.50	320.25	17.90
x2	2.60	0.84	0.92
y	6.61	2.37	1.54

Modello di regressione lineare semplice

$$Y = a + bx_1 + e$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 1.1826$$

$$\hat{b} = \frac{\text{cov}(X, Y)}{\text{var}(X)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0.0665$$



Indice di determinazione lineare

$$R^2 = 0.598$$

Può essere calcolato come:

- $1 - \text{EQM} / \text{Var}(Y)$: $1 - 0.9533 / 2.37$
- Proporzione di varianza spiegata: $1.418 / 2.37$
- ρ^2 : $(0.7733)^2$

$$\text{Varianza spiegata} = \sum (y^* - M(y))^2 = 14.18 / 10 = 1.418$$

Errore quadratico medio

$$\text{EQM} = \sum (y_i - y_i^*)^2 / N = \sum (y_i - a - bx_i)^2 / N = 9.53386 / 10 = 0.9533$$

obs	Y	Y*	e	$e_i = Y_i - Y_i^*$
1	9.3000	7.8371	1.4629	
2	4.8000	4.5098	0.2902	
3	8.9000	7.8371	1.0629	
4	5.8000	7.8371	-2.0371	
5	4.2600	4.5098	-0.2498	
6	6.8000	6.5062	0.2938	
7	6.6000	6.1735	0.4265	
8	5.9000	6.5062	-0.6062	
9	7.6000	7.1716	0.4284	
10	6.1000	7.1716	-1.0716	

$$Y^* = 1.1826 + 0.0665 X_1 \quad \text{Valori stimati dal modello}$$

Modello di regressione lineare multipla

$$Y = \alpha + \beta x_1 + \beta x_2 + \varepsilon$$

$$R^2 = 0.7878$$

$$\alpha = 0.09239$$

$$\beta_1 = 0.05559$$

$$\beta_2 = 0.76273$$

Rispetto al modello lineare:

R^2 è più alto (0.7878 contro 0.5980)

EQM più piccolo (0.5032 contro 0.9533)

EQM da regressione multipla

$$\text{EQM} = 1/N \sum (y_i - y_i^*)^2 = \sum (y_i - a - bx_i)^2 / N = 5.03154 / 10 = 0.5032$$

Obs	y	y*	e	$e = Y_i - Y_i^*$
1	9.3000	8.7022	0.597	
2	4.8000	5.1600	-0.3600	
3	8.9000	8.7022	0.1978	
4	5.8000	7.1768	-1.3768	
5	4.2600	4.3973	-0.1373	
6	6.8000	5.3023	1.4977	
7	6.6000	6.5498	0.0502	
8	5.9000	6.0650	-0.1650	
9	7.6000	7.3836	0.2164	
10	6.1000	6.6209	-0.5209	

$$Y^* = 0.09239 + 0.0559 X_1 + 0.76273 X_2 \quad \text{Valori stimati}$$

Indice di determinazione lineare

$$R^2=0.7878$$

Può essere calcolato come:

- $1 - \text{EQM} / \text{Var}(Y)$: $1 - 0.5032 / 2.37$
- Proporzione di varianza spiegata: $1.8684 / 2.37$

$$[R^2]^{0.5} = (0.598)^{0.5} \quad \text{coeff. correlazione parziale}$$

$$\text{Varianza spiegata} = \sum (y^* - M(y))^2 = 18.684 / 10 = 1.8684$$

Coefficienti di regressione totali e parziali

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \text{regr. multipla}$$

$$y_i = b_0 + b x_{1i} + v_i \quad \text{regr. semplice}$$

$$\text{In generale } b \neq \beta_1: 0.0665 \neq 0.05559$$

coefficienti

netti

$$\beta_1 = \frac{\sigma_y}{\sigma_{x_1}} \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2}$$

lordi

$$b_1 = \frac{\text{cov}(y, x_1)}{\sigma_{x_1}^2} = \frac{\sigma_y}{\sigma_{x_1}} r_{y,x_1}$$

$$\begin{aligned} r_{x_1,x_2} &= 0.28046 \\ r_{x_1,x_2}^2 &= 0.0787 \\ r_{y,x_2} &= 0.6351 \end{aligned}$$

$$r_{y,x_1} = 0.7733$$

Troppe o poche variabili?

- ◆ Cosa succede se includiamo nel modello variabili che non servono?
→ Non c'è effetto sulle stime dei parametri: stime OLS corrette
- ◆ Cosa accade se escludiamo delle variabili rilevanti?
→ Le stime OLS saranno distorte!

Distorsione per omissione di variabili

- ◆ Il modello vero sia
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$
- ◆ Ma noi stimiamo
$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \varepsilon$$
- ◆ allora si ha:
$$\tilde{\beta}_1 = \frac{\sum (x_{1i} - \bar{x}_1) y_i}{\sum (x_{1i} - \bar{x}_1)^2}$$

Distorsione per omissione di variabili (2)

Il vero modello è

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

quindi il numeratore diventa:

$$\begin{aligned} \sum (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i) = \\ \beta_1 \sum (x_{i1} - \bar{x}_1)^2 + \beta_2 \sum (x_{i1} - \bar{x}_1)x_{i2} + \sum (x_{i1} - \bar{x}_1)\varepsilon_i \end{aligned}$$

Distorsione per omissione di variabili (3)

$$\tilde{\beta} = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1)x_{i2}}{\sum ((x_{i1} - \bar{x}_1)^2)} + \frac{\sum (x_{i1} - \bar{x}_1)\varepsilon_i}{\sum ((x_{i1} - \bar{x}_1)^2)}$$

poichè $E(\varepsilon_i) = 0$, calcolando il valore atteso si ha

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1)x_{i2}}{\sum ((x_{i1} - \bar{x}_1)^2)}$$

Distorsione per omissione di variabili (4)

◆ Si consideri la regressione di x_1 su x_2

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1 \Rightarrow \tilde{\delta}_1 = \frac{\sum (x_{i1} - \bar{x}_1)x_{i2}}{\sum ((x_{i1} - \bar{x}_1)^2)}$$

$$\Rightarrow E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1$$

Direzione della distorsione (bias)

	$\rho(x_1, x_2) > 0$	$\rho(x_1, x_2) < 0$
$\beta_2 > 0$	Positiva	Negativa
$\beta_2 < 0$	Negativa	Positiva

Distorsione per omissione di variabili: in sintesi...

- ◆ Ci sono due casi nei quali la distorsione è nulla:
 - $\beta_2 = 0$, ossia x_2 non è una variabile rilevante
 - x_1 e x_2 sono incorrelate nel campione
- ◆ Se la correlazione tra x_2 , x_1 e x_2 , y hanno lo stesso segno, la distorsione sarà positiva
- ◆ Se la correlazione tra x_2 , x_1 e x_2 , y hanno segno opposto, la distorsione sarà negativa

Il test t

- ◆ $H_0: \beta_j = 0$
- ◆ H_1 può essere unidirezionale o bidirezionale
- ◆ La statistica test segue una distribuzione t di Student con $(n-k-1)$ g.l.

Alternative unidirezionali

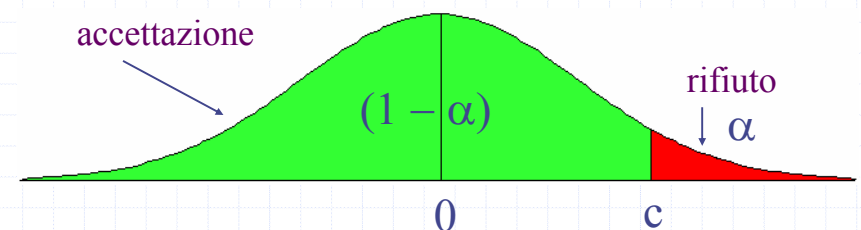
- ◆ posto il livello di significatività, α , cerchiamo il percentile $(1 - \alpha)$ della distribuzione t con $(n - k - 1)$ g.l. e lo indichiamo con c , valore critico
- ◆ Si rifiuta l'ipotesi nulla se la statistica t è maggiore del valore critico: $t > c$
- ◆ Se $t < c$ non rifiutiamo l'ipotesi nulla

Alternative unidirezionali (2)

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i$$

$$H_0: \beta_j = 0$$

$$H_1: \beta_j > 0$$

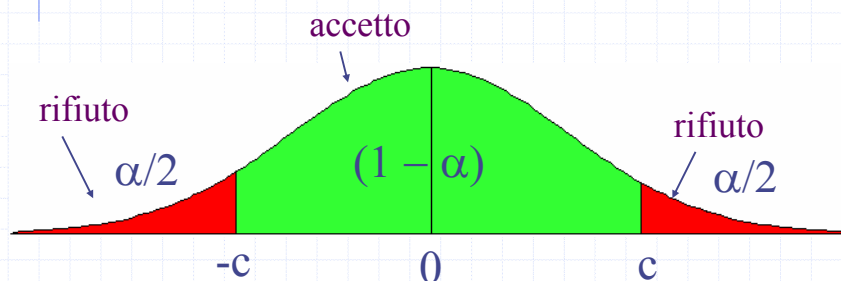


Alternative bidirezionali

$$y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i$$

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$



Test per l'esclusione di 2 o più variabili esplicative

- ◆ Ipotesi nulla $H_0: \beta_{k-q+1} = 0, \dots, \beta_k = 0$
- ◆ Ipotesi alternativa $H_1: H_0$ falsa
- ◆ Non possiamo sottoporre a verifica questa ipotesi con test t separati, perchè vogliamo sapere se i q parametri sono congiuntamente significativi

Test per l'esclusione di più variabili esplicative (2)

- ◆ Per fare questo test dobbiamo stimare il modello "vincolato" senza $x_{k-q+1}, \dots, x_k$, e il modello "non vincolato" con tutte le x
- ◆ Intuitivamente, vogliamo sapere se il cambiamento di SSR è abbastanza grande per includere $x_{k-q+1}, \dots, x_k$

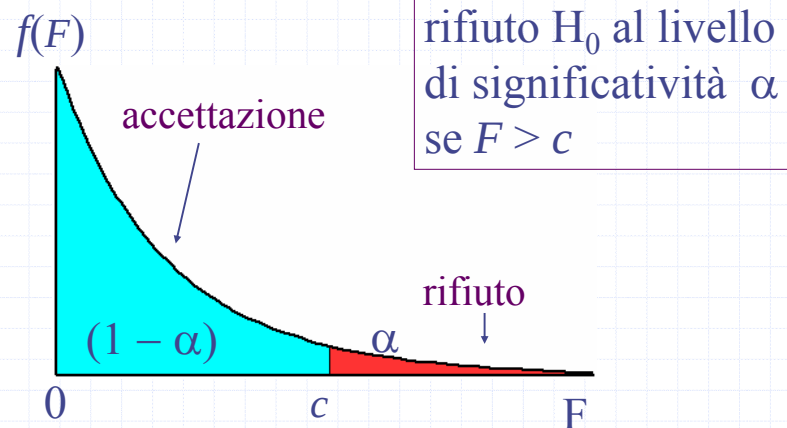
$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)}$$

dove r indica il modello vincolato e ur il modello non vincolato

La statistica F

- ◆ La statistica F è sempre positiva, poichè SSR dal modello vincolato non può mai essere più piccola di SSR dal modello non vincolato
- ◆ la statistica F misura l'incremento relativo di SSR quando si passa dal modello non vincolato a quello vincolato
- ◆ q = numero di vincoli, o $gl_r - gl_{ur}$
- ◆ $n - k - 1 = gl_{ur}$

La statistica F (2)



La statistica F in termini di R^2

◆ Ricordando che $SSR = SST(1 - R^2)$ si ha:

$$F \equiv \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)}$$

Confronto con modello nullo

- ◆ Caso speciale $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
- ◆ Poichè R^2 del modello nullo (solo intercetta) è zero, la statistica F è pari a

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$