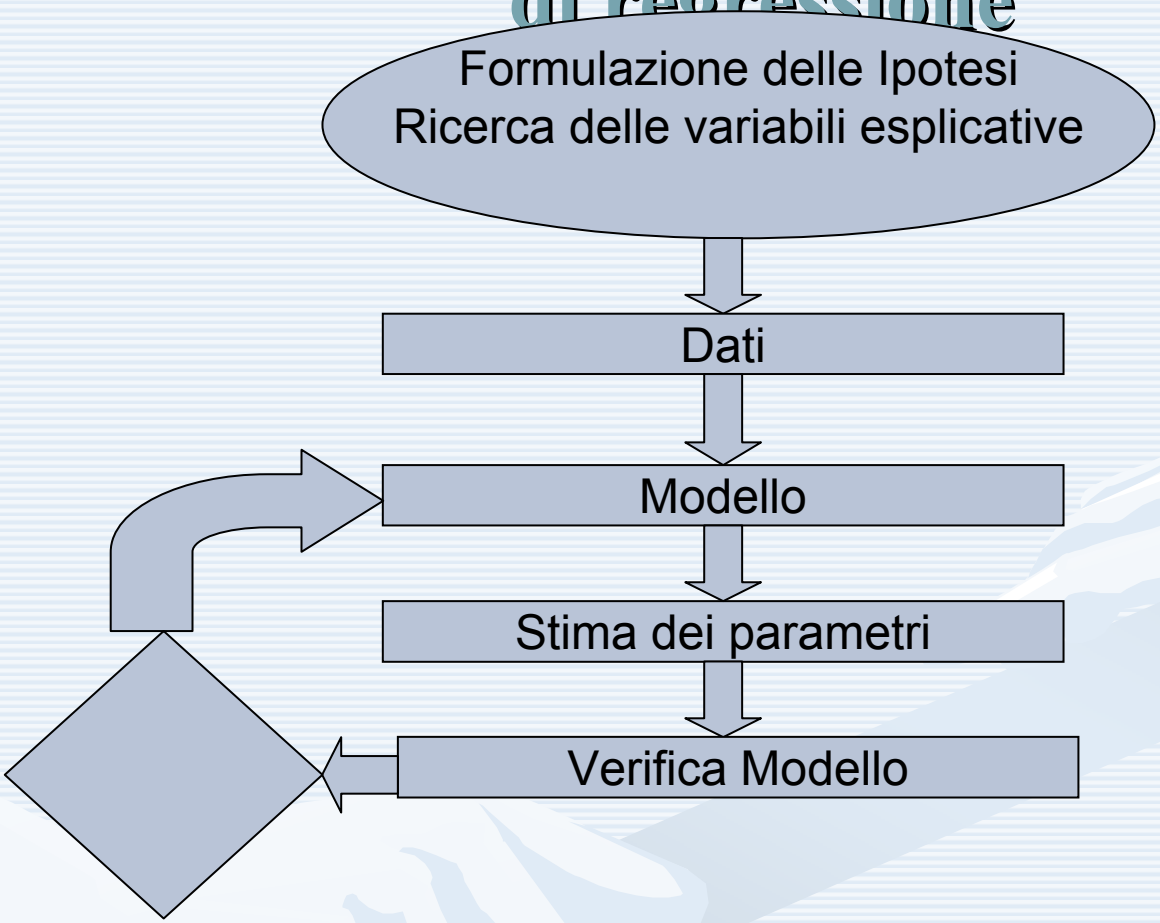
Descrizione per la costruzione del modello di regressione



IL MODELLO DI REGRESSIONE LINEARE MULTIPLA

L'analisi della regressione multipla è una tecnica statistica che può essere impiegata per analizzare la relazione tra una **variabile dipendente** e diverse **variabili indipendenti** (**predittori**).

L'**OBIETTIVO** dell'analisi è prevedere i valori assunti da una variabile dipendente a partire dalla conoscenza di quelli osservati su più variabili indipendenti.

Se il problema coinvolge una sola variabile indipendente, la tecnica statistica viene definita **regressione semplice**.

Quando invece il problema coinvolge due o più variabili indipendenti, è detta, appunto, **regressione multipla**.

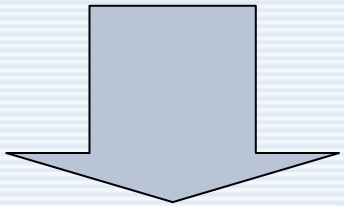
La relazione tra le variabili esplicative e la variabile dipendente può essere scritta come:

$$Y = f(X_1, X_2, ..., X_m) + \varepsilon = f(\mathbf{X}) + \varepsilon$$

Se si esplicita una **relazione di tipo lineare** si ottiene l'equazione:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + \varepsilon = \mathbf{X\beta} + \varepsilon$$

nella quale dovranno essere stimati i **parametri** β_i



Metodo dei minimi quadrati

A tal scopo è necessario osservare le variabili esplicative e la variabile dipendente su un campione di ***n*** osservazioni

Regressione lineare semplice (1 dip, 1 indip)

$$Y_i = a + bX_i + \varepsilon_i$$

intercetta

pendenza

variabile indipendente

errore

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i+} \varepsilon_i$$

Regressione lineare multipla (2 indip, 1 dip)

RAPPRESENTAZIONE MATRICIALE

Dato il modello

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

la rappresentazione dei dati campionari potrà allora essere la seguente:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ . & . & & & . \\ . & . & & & . \\ . & . & & & . \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_m \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i+} \varepsilon_i$$

y	X_1	X2	
3	2	1	$3=1\beta_0+2\beta_1+1\beta_2+e_1$
2	3	5	$2=1\beta_0+3\beta_1+5\beta_2+e_2$
4	5	3	$4=1\beta_0+5\beta_1+3\beta_2+e_3$
5	7	6	$5=1\beta_0+7\beta_1+6\beta_2+e_4$
8	8	7	$8=1\beta_0+8\beta_1+7\beta_2+e_5$

$3=1\beta_0+2\beta_1+1\beta_2+e_1$
$2=1\beta_0+3\beta_1+5\beta_2+e_2$
$4=1\beta_0+5\beta_1+3\beta_2+e_3$
$5=1\beta_0+7\beta_1+6\beta_2+e_4$
$8=1\beta_0+8\beta_1+7\beta_2+e_5$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 5 \\ 1 & 5 & 3 \\ 1 & 7 & 6 \\ 1 & 8 & 7 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

IPOTESI DEL MODELLO DI REGRESSIONE MULTIPLA

Corretta specificazione del modello

$$E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \Leftrightarrow \quad E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

$$VAR(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n \quad \Leftrightarrow \quad VAR(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$$

Normalità distributiva della variabile d'errore e, da cui segue la normalità distributiva della variabile dipendente

Matrice di osservazioni **X** non stocastica, e rango(**X**) = *m*+1

Quando *m*=1 queste ipotesi coincidono con quelle del modello di regressione semplice.

OSSERVAZIONI

La terza ipotesi include sia la omoschedasticità $VAR(\varepsilon_i) = \sigma^2$

che l'incorrelazione delle variabili casuali errori $COVAR(\varepsilon_i, \varepsilon_j) = 0$ per ogni *i* e *j*

L'assunzione riguardante il rango della matrice **X** impone in pratica che il numero di informazioni campionarie non ridondanti sia almeno pari al numero dei parametri da stimare.

1) linearità

2) $E(\boldsymbol{\varepsilon}) = 0$ …con $\boldsymbol{\varepsilon}$ vettore, $\Rightarrow$

$$E(Y_i \mid X_{i1}, X_{i2}, \ldots, X_{ip}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

3) omoschedasticità degli errori (delle Y_i) e assenza di autocorrelazione tra errori:

$$E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right) = \begin{bmatrix} E(\varepsilon_1\varepsilon_1) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ \vdots & \vdots & \cdots & \vdots \\ E(\varepsilon_n\varepsilon_1) & & \cdots & E(\varepsilon_n\varepsilon_n) \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ \vdots & \sigma^2 & \cdots & \vdots \\ 0 & & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

4) **X** contiene variabili non stocastiche

5) normalità degli errori (e quindi di Y_i)

6) indipendenza lineare delle variabili esplicative (rango della matrice **X** = p+1), cioè nessuna X può essere espressa come combinazione lineare delle altre var. espl. -- Tale assunzione implica che n≥ p+1

STIMA DEI PARAMETRI: METODO DEI MINIMI QUADRATI

L'obiettivo è determinare, sulla base dei dati campionari, il vettore **b** delle stime che minimizza:

$$\Phi(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) =$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} =$$

$$= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Derivando rispetto a **b** e uguagliando a zero di ottiene:

$$\frac{\partial \Phi(\beta)}{\partial \beta} = -2\mathrm{X}'\mathrm{y} + 2(\mathrm{X}'\mathrm{X})\hat{\beta} = 0$$

da cui si ricava il vettore **b** delle stime dell'intercetta e dei coefficienti di regressione:

$$\mathrm{b} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

N $\quad \sum x_1, \sum x_2 \quad \sum x_1^2, \sum x_2^2 \quad \sum x_1 x_2$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \\ 1 & 5 & 3 & 6 & 7 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 5 \\ 1 & 5 & 3 \\ 1 & 7 & 6 \\ 1 & 8 & 7 \end{bmatrix} = \begin{bmatrix} 5 & 25 & 22 \\ 25 & 151 & 130 \\ 22 & 130 & 120 \end{bmatrix}$$

X′ **X** **X′X**

Calcolare l'inversa

La diamo per scontata

$$\begin{bmatrix} 5 & 25 & 22 \\ 25 & 151 & 130 \\ 22 & 130 & 120 \end{bmatrix}^{-1} = \begin{bmatrix} 1{,}201 & -0.138 & -0{,}071 \\ -1.138 & 0{,}114 & -0{,}098 \\ -0{,}071 & -0{,}098 & 0{,}128 \end{bmatrix}$$

X'X **inversa**

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \\ 1 & 5 & 3 & 6 & 7 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 22 \\ 131 \\ 111 \end{bmatrix}$$

$\sum y$

$\sum x_1 y$

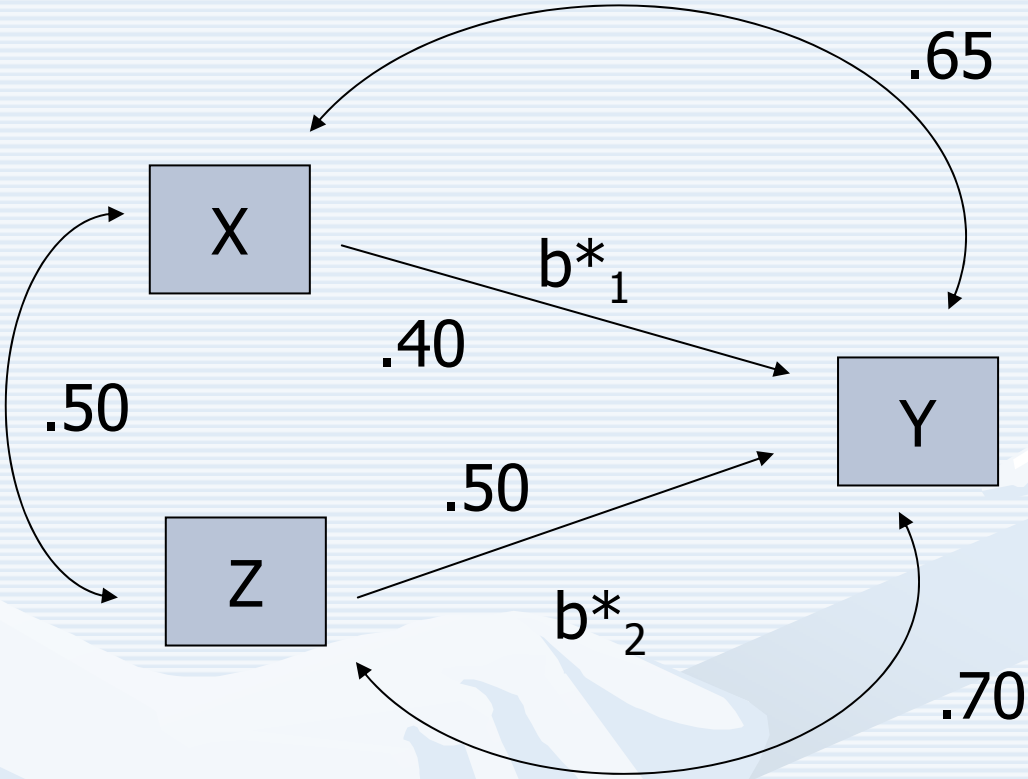
$\sum x_2 y$

X′ **y** **X′y**

$$\begin{bmatrix} 1{,}201 & -0.138 & -0{,}071 \\ -1.138 & 0{,}114 & -0{,}098 \\ -0{,}071 & -0{,}098 & 0{,}128 \end{bmatrix} \begin{bmatrix} 22 \\ 131 \\ 111 \end{bmatrix} = \begin{bmatrix} 0.50 \\ 1 \\ -0.25 \end{bmatrix}$$

$$\hat{Y}_i = .50 + 1X_{1i} + (-.25)X_{2i}$$

La correlazione fra 2 variabili è la somma delle influenze dirette e indirette delle due variabili



$r_{xz}=.5$
$r_{xy}=.65$
$r_{zy}=.70$

$r_{xy}=b^*_1+b^*_2 r_{zx}$
$r_{zy}=b^*_2+b^*_1 r_{zx}$

$b^*_1=r_{xy}-r_{xz}b^*_2=.65-.50b^*_2$
$b^*_2=r_{zy}-r_{xz}b^*_1=.70-.50b^*_1$

Sviluppando…

poniamo $X=X_1$, $Z=X_2$

$r_{xy}=b_1+b_2r_{zx}$
$r_{zy}=b_2+b_1r_{zx}$

$r_{y1}=b_1r_{11}+b_2r_{12}= b_1r_{11}+b_2r_{12}$
$r_{y2}=b_2r_{22}+b_1r_{12}= b_1r_{12}+b_2r_{22}$

$$\begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{21} \\ r_{12} & r_{22} \end{bmatrix} \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix}$$

$$r_{yx} = R_{xx} b_{yx}^*$$

$$r_{yx} = R_{xx} b^*_{yx}$$

$$b^* = R^{-1} r$$

$$r_{y1} = b^*_{y1.23} r_{11} + b^*_{y2.13} r_{12} + b^*_{y3.12} r_{13}$$

$$r_{y2} = b^*_{y1.23} r_{21} + b^*_{y2.13} r_{22} + b^*_{y3.12} r_{23}$$

$$r_{y3} = b^*_{y1.23} r_{31} + b^*_{y2.13} r_{32} + b^*_{y3.12} r_{33}$$

con $r_{ij} = r_{ji}$

$$b^*_{yx} = \begin{bmatrix} b^*_{y1.23} \\ b^*_{y2.13} \\ b^*_{y3.12} \end{bmatrix} \quad R_{XX} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \quad r_{yx} = \begin{bmatrix} r_{y1} \\ r_{y2} \\ r_{y3} \end{bmatrix}$$

Regressione matriciale

formule alternative:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\boldsymbol{\beta} = \mathbf{C}_{xx}^{-1}\mathbf{c}_{yx}$$

$$\boldsymbol{\beta}^{*} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{yx}$$

$\mathbf{C}_{xx}$ è la matrice varianza/covarianza fra le X

$\mathbf{c}_{yx}$ è il vettore delle covarianze fra le x e la y

$\mathbf{R}_{xx}$ è la matrice di correlazione fra le X

$\mathbf{r}_{yx}$ è il vettore delle correlazioni fra le x e la y

Esempio di $b=C^{-1}c$

$$C_{XX} = \begin{bmatrix} 6.5 & 5.0 \\ 5.0 & 5.8 \end{bmatrix} \qquad c_{xy} = \begin{bmatrix} 5.25 \\ 3.55 \end{bmatrix}$$

varianza e covarianza calcolate con N-1

$$\frac{1}{12.7}\begin{bmatrix} 5.8 & -5.0 \\ -5.0 & 6.5 \end{bmatrix}\begin{bmatrix} 5.25 \\ 3.55 \end{bmatrix} = \begin{bmatrix} 1.00 \\ -0.25 \end{bmatrix} \begin{matrix} \leftarrow b_1 \\ \leftarrow b_2 \end{matrix}$$

$$b_0 = \bar{Y} - \sum (b_i \bar{X}_i) = 4.4 - 1(5) - (-.25)4.4 = 0.5$$

Beta standardizzati

$$b^*_{yx_i} = b_{yx_i}\frac{s_{x_i}}{s_y}$$

$$b_{yx_i} = b^*_{yx_i}\frac{s_y}{s_{x_i}}$$

Con i dati dell'esempio precedente:

$$b^*_{yx_1} = \frac{2.54}{2.3}\times 1 = 1.109$$

$$b^*_{yx2} = \frac{2.408}{2.30}(-.25) = -0.262$$

Esempio con $b^*=R^{-1}r$

$$R_{XX} = \begin{bmatrix} 1 & .814 \\ .814 & 1 \end{bmatrix} \qquad r_{xy} = \begin{bmatrix} .894 \\ .640 \end{bmatrix}$$

$$\frac{1}{0.337}\begin{bmatrix} 1 & -.814 \\ -.814 & 1 \end{bmatrix}\begin{bmatrix} .894 \\ .640 \end{bmatrix} = \begin{bmatrix} 1.107 \\ -0.261 \end{bmatrix}$$

$b_0 = 0$

Propor. di varianza spiegata

$$r^2 = r^2_{y\hat{y}} = \frac{spiegata}{totale} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

$$= \frac{\sum(Y - \bar{Y})^2 - \sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$$

$$= \underbrace{r_{y1}b^*_{y1.2} + r_{y2}b^*_{y2.1}}_{\text{con 2 X}} = \underbrace{\sum r_{yi}b^*_i}_{\text{generico}}$$

Stimatore dei Minimi Quadrati: le proprietà

$$B = (X'X)^{-1}X'Y$$

$$E(B) = \beta \qquad Stimatore \quad Corretto$$

$$Var(B) = (X'X)^{-1}\sigma^2$$

Cosa fare se σ è incognito?

Stimare σ

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-m-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-m-1}$$

$$Var(B_i) = MSE * c_{ii} \qquad con \quad i = 1,2,...,m$$

Dove c_{ii} rappresenta l'i-esimo elemento sulla diagonale della matrice $(\mathbf{X}'\mathbf{X})^{-1}$

MISURE DI BONTA' DEL MODELLO: INDICE DI DETERMINAZIONE LINEARE

$$R^2 = \frac{\text{Devianza di regressione}}{\text{Devianza totale}}$$

Nel modello di regressione multipla l'indice di determinazione lineare può presentare alcuni problemi calcolatori e di interpretazione. Ad esempio, in caso di assenza di relazione lineare non è pari a zero.

R^2 tende ad aumentare al numero delle X

E' bene ricorrere perciò all'indice **R^2 corretto**: che varia sempre tra zero e uno.

m=numero di variabili indipendenti (X)

$$\overline{R}^2 = \left(R^2 - \frac{m}{n-1} \right) \frac{n-1}{n-m-1}$$

$$adjR^2 = 1 - (1 - R^2) \frac{n-1}{n-m-1}$$

CONTROLLO D'IPOTESI SUL MODELLO:
esiste un legame effettivo tra la variabile dipendente e i regressori?

Si tratta di saggiare l'ipotesi nulla

$$H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0$$

Tale ipotesi si controlla con il **test F di Fisher**.
La statistica test si ottiene dal rapporto tra la varianza di regressione e la varianza di dispersione del modello:

$$F = \frac{\frac{Dev(Y)_{regr}}{m}}{\frac{Dev(Y)_{residua}}{n-m-1}} = \frac{Var(Y)_{regr}}{Var(Y)_{residua}} = \frac{SSR/m}{MSE}$$

L'ipotesi nulla viene rigettata se, a un prefissato livello di significatività α, la *F* così calcolata sui dati campionari è maggiore del valore della F di Fisher tabulato in corrispondenza di m e (n-m-1) gradi di libertà: $F_{\alpha,m,n-m-1}$

un test globale: che include tutte le variabili

Confronto fra: $Y = b_0 + \varepsilon$ df_r=N-1 (ristretto)

$Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon$ df_f=N-3 (completo)

$H_0 : b_1 = b_2 = 0$

Usiamo la statistica F di Fisher

Se è significativa, c'è una relazione consistente fra le x e la y; la regressione ha senso. N.B.: In genere è significativa

$$F = \frac{(R_f^2 - R_r^2)/(d_r - d_f)}{(1 - R_f^2)/d_f}$$

f=full (completo)
r=ristretto [$R^2=0$]

$$= \frac{\sum (Y - \bar{Y})^2 - \sum (Y - \hat{Y})^2 / (d_r - d_f)}{\sum (Y - \hat{Y})^2 / d_f}$$

$$= \frac{R_f^2 / m}{(1 - R_f^2)/(N - m - 1)}$$

Se il modello globale è significativo, si può fare:

un test per ciascuna var. indip. (X)

Anche se il modello globale è significativo, questo non significa che tutte le X siano significativamente associate a Y

La maggior parte dei programmi utilizza un semplice t-test. Se il test è significativo, la X_n può stare nel modello, altrimenti si dovrebbe togliere.

CONTROLLO D'IPOTESI SUL MODELLO:
esiste un legame lineare tra la variabile dipendente e il singolo regressore Xi?

Si tratta di saggiare l'ipotesi nulla

$$H_0 : \beta_i = 0 \qquad \forall i = 1,\ldots,m$$

ale ipotesi si controlla con il **test t di Student**. La statistica test si ottiene:

$$t = \frac{B_i}{\sqrt{\mathrm{var}(B_i)}} = \frac{B_i}{\sqrt{MSE \cdot c_{ii}}}$$

Errore Standard dell'i-esimo coefficiente di regressione

Dove c_{ii} rappresenta l'i-esimo elemento sulla diagonale della matrice $(\mathbf{X}'\mathbf{X})^{-1}$

L'ipotesi nulla viene rigettata se, a un prefissato livello di significatività α, la *t* così calcolata sui dati campionari è maggiore del valore della t di Student tabulato in corrispondenza di (n-m-1) gradi di libertà: $t_{\alpha,n-m-1}$

Intervallo di confidenza per la previsione di un singolo valore

$\mathbf{x}_I' = [1 \quad x_{I1} \quad \cdots \quad x_{Ip}]$

$\hat{Y}_I = \mathbf{x}_I'\hat{\mathbf{b}}$ = previsione per il singolo valore

Varianza della previsione per il singolo valore:

$$Var(\hat{Y}_I) = \sigma^2(1 + \mathbf{x}_I'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_I)$$

Fissato il livello di confidenza $(1-\alpha)$, l'intervallo di confidenza per la previsione sarà:

$$\hat{Y}_I \pm t_{\alpha/2,n-p-1} s\sqrt{(1 + \mathbf{x}_I'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_I)}$$

in cui t_{n-p-1} è il valore soglia teorico della distribuzione t-Student con n-p-1 g.d.l.

ANALISI DEI RESIDUI

L'analisi grafica dei residui consente di valutare, a posteriori, se il modello ipotizzato è corretto.
In tal caso, infatti, gli errori dovrebbero distribuirsi in modo normale.

Ancora, la rappresentazione grafica dei residui rispetto ai valori stimati della variabile dipendente consente di valutare la sussistenza delle ipotesi del modello:

$$E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{e} \quad VAR(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$$

Nel caso in cui si disponga di dati temporali, si può valutare l'esistenza di auto-correlazione tra i residui con il test di Durbin-Watson, che saggia l'ipotesi nulla di ASSENZA DI AUTOCORRELAZIONE tra i residui. La statistica test è:

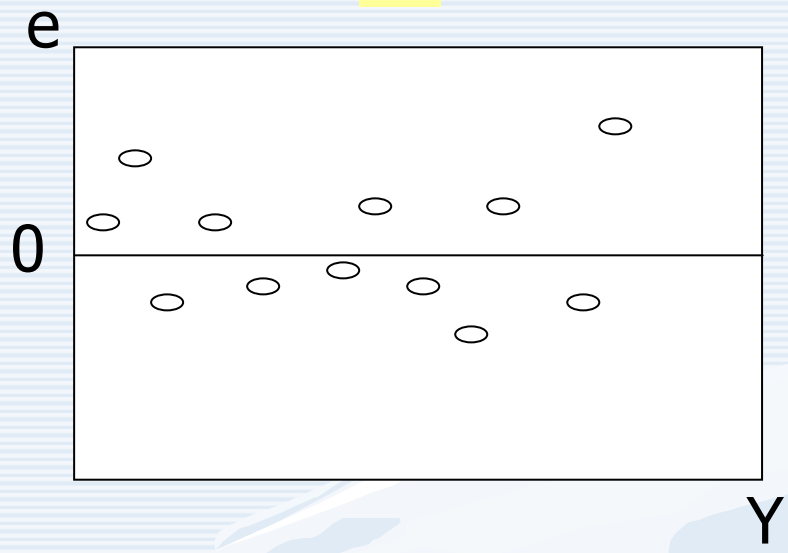
$$d = \frac{\sum_{i=1}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} (e_i)^2}$$

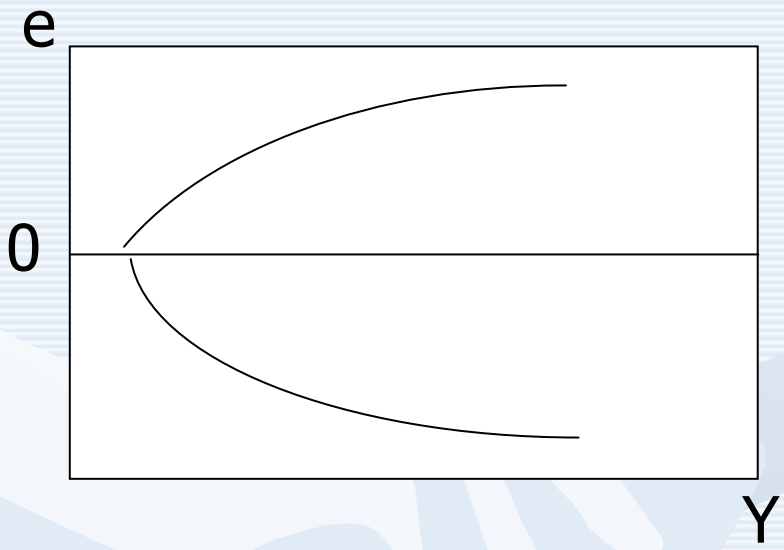
Un valore tra 1,3 e 1,4 indica autocorrelazione tra i residui

Residui

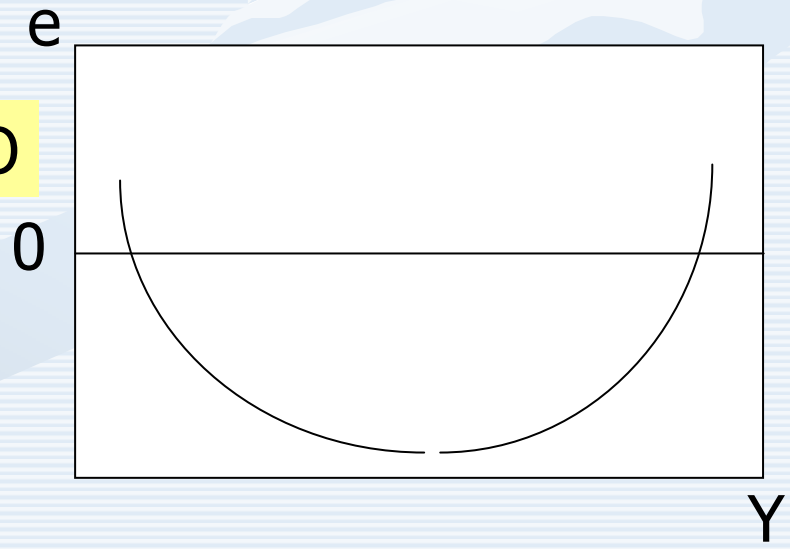
I residui (e=Y-Y') dovrebbero essere dispersi casualmente attorno a Y

si





NO



Se **non** sono dispersi casualmente, esiste un'altra variabile X che può spiegarne una parte, oppure la relazione non è lineare

MULTICOLLINEARITA'

Con il termine **multicollinearità** ci si riferisce alla correlazione fra le variabili indipendenti di un modello di regressione.
Il suo effetto consiste nel ridurre la capacità previsiva di ogni singola variabile indipendente in modo proporzionale alla forza della sua associazione con le altre variabili indipendenti.
L'effetto della multicollinearità può interessare sia la capacità di *spiegazione* del modello (capacità della procedura di regressione e del ricercatore di rappresentare e capire l'influenza di ciascuna variabile indipendente) sia la sua *stima* (la sua presenza rende problematica la determinazione dei contributi individuali delle variabili indipendenti, perché i loro effetti vengono "mescolati" o confusi).
Va pertanto valutata e individuata. Due strumenti a disposizione sono la Tolleranza (**Tolerance**) e il Fattori di Accrescimento della Varianza (**Variance Inflaction Factor**).

Tolerance = $1 - R_{i0}^2$

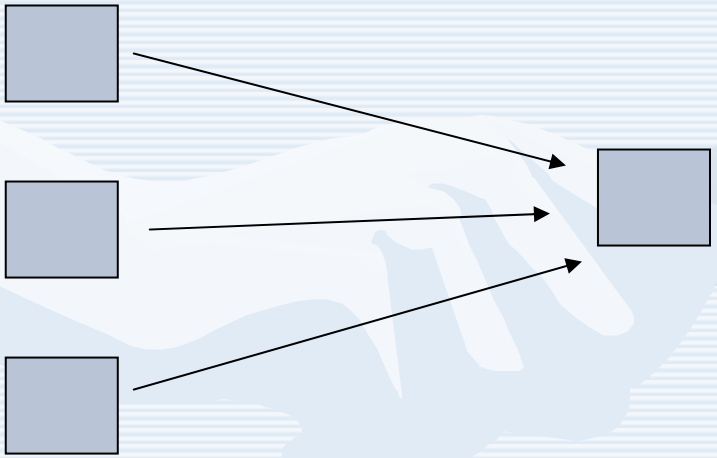
$$VIF_i = \frac{1}{1 - R_{i0}^2}$$

dove R_{i0}^2 rappresenta il quadrato del coefficiente che misura la correlazione fra la i-esima variabile esplicativa e tutte le altre.
In generale un VIF>5 è indice di alta multicollinearità.

Multicollinearità 1

La situazione ideale per una regressione multipla dovrebbe essere: ogni X è altamente correlata con Y, ma le X non sono correlate fra loro

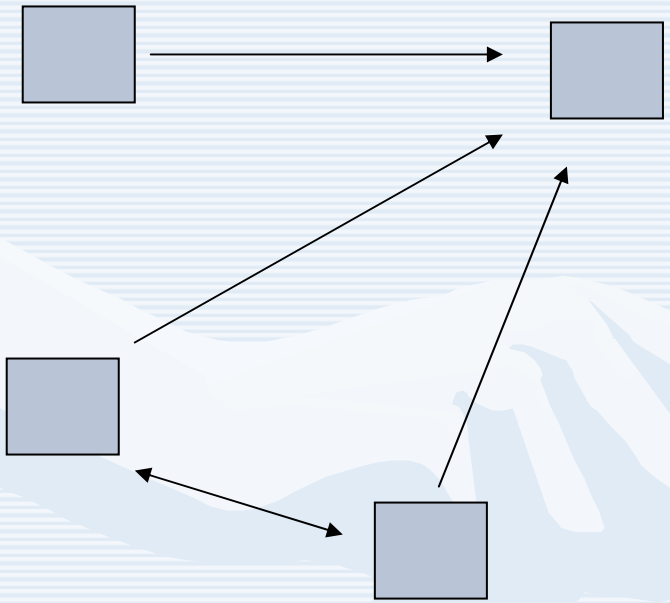
	X_1	X_2	X_3
Y	.60	.50	.70
X_1		.20	.30
X_2			.20



Idealmente, le correlazioni tra le X, dovrebbero essere 0; in questo modo beta dovrebbe coincidere con r e non con r parzializzato

Multicollinearità 2

Spesso però, due o più X sono correlate fra loro



	X_1	X_2	X_3
Y	.60	.50	.70
X_1		**.70**	.30
X_2			.20

Quando due variabili X o più, sono tra loro correlate (moderatamente o più), parliamo di "**multicollinearità**".

Problemi della multicollinearità

- fa diminuire la R multipla
- l'effetto dei predittori si confonde
- aumenta la varianza e l'instabilità dell'equazione

Diminuire la multicollinearità

- combinare fra loro i predittori altamente correlati (ad esempio sommandoli)
- se ci sono molti predittori altamente correlati, usare un'analisi delle componenti principali per ridurre il numero delle X

Scegliere i predittori

- Usare la teoria (ricerca bibliografica)
- metodi semi-automatici sequenziali
  - forward
  - stepwise
  - backward

Regressione standard

- Tutte le variabili X vengono considerate assieme e tutti i coefficienti di regressione (B o beta) stimati contemporaneamente

Forward

- Le variabili X vengono inserite una alla volta (in genere la X con la correlazione XY più alta) e vengono poi calcolate le correlazioni parziali e i test di significatività di tutte le altre.
- Una nuova variabile viene inserita se risulta statisticamente associata al modello
- Ci si ferma quando non ci sono variabili signficative

Backword

- Le X vengono inserite tutte assieme e poi pian piano tolte se non risultano significative al t-test
- Ci si ferma quando tutte le non significative sono state tolte

Stepwise

- Si parte con “alcune” variabili X e poi
- Le altre X vengono inserite e / o tolte a seconda della loro importanza e significatività
- Il modello finale identificato “dovrebbe” essere il migliore

Esercizio sulla regressione Multipla: 1 variabile indipendente (Y) e 3 variabili dipenden
(X).

Si vuole analizzare la relazione tra il numero di Carte di Credito di una famiglia in
relazione a tre possibili variabili di influenza

Numero Carte di Credito (Y)	Ampiezza della Famiglia (X1)	Reddito della Famiglia (in migliaia di €) (X2)	Numero di auto della famiglia (X3)
4	2	14	1
6	2	16	2
6	4	14	2
7	4	17	1
8	5	18	3
7	5	21	2
8	6	17	1
10	6	25	2

Fasi dell'analisi:

1) Stima dei parametri di regressione
2) Inferenza sui parametri di Regressione Multipla (Test di Ipotesi ,Intervalli di confidenza)
3) Diagnostica di Regressione: Plot dei Residui
4) Previsioni

Stima dei Parametri di Regressione (utilizzo di Excel o di software Statistici)

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*
Intercetta	0,286	1,606	0,178	0,867
Ampiezza della Famiglia	0,635	0,271	2,341	0,0792
Reddito della Famiglia (in migliaia di €)	0,200	0,119	1,671	0,170
Numero di auto della famiglia	0,272	0,470	0,578	0,594

$$Y = 0{,}286 + 0{,}635X_1 + 0{,}2X_2 + 0{,}272X_3$$

Interpretazione dei Coefficienti : Attenzione

La Bontà dell'adattamento del Modello Lineare

R al quadrato	0,872
R al quadrato corretto	0,776

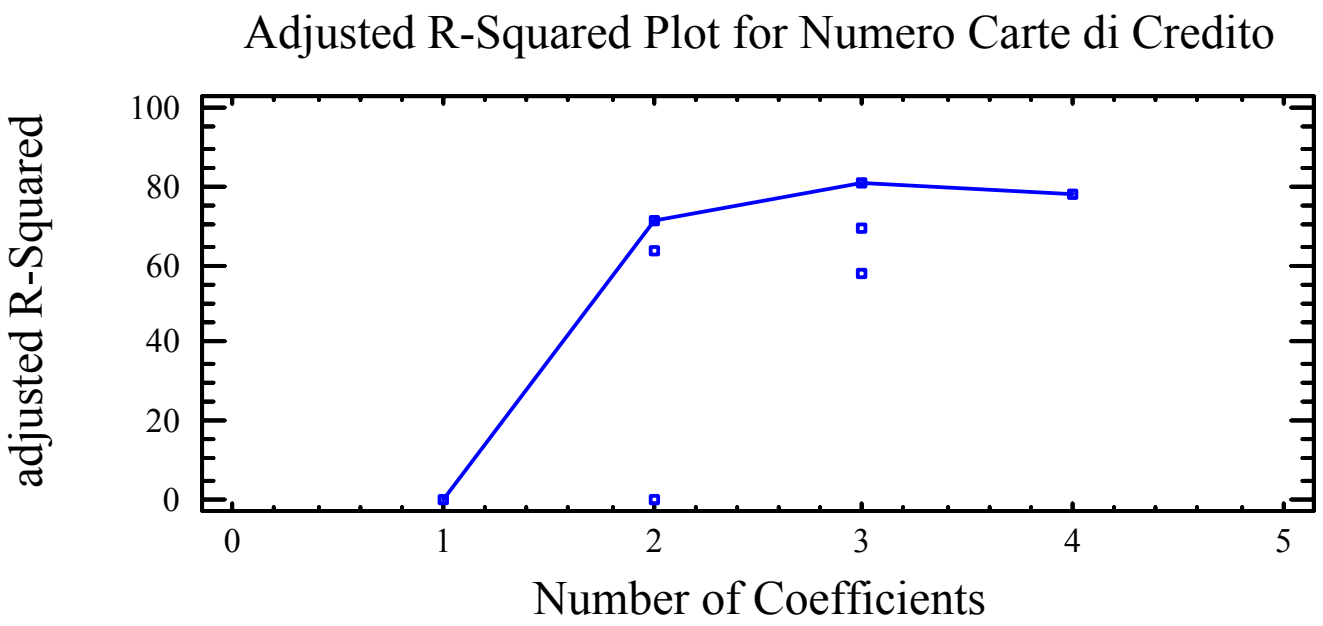
Inferenza sui Coefficienti : La significatività dei coefficienti e la Selezione delle Variabili Esplicative

Regression Model Selection

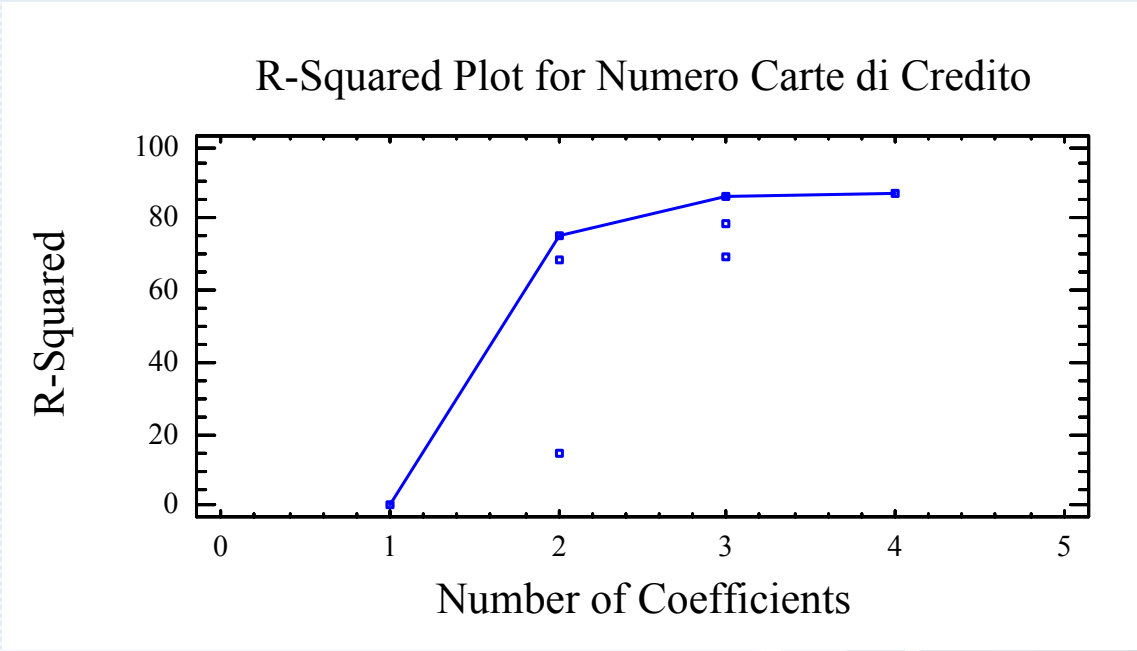
Dependent variable: Numero Carte di Credito
Independent variables: A=Ampiezza della Famiglia B= Numero di Auto C=Reddito

Model Results

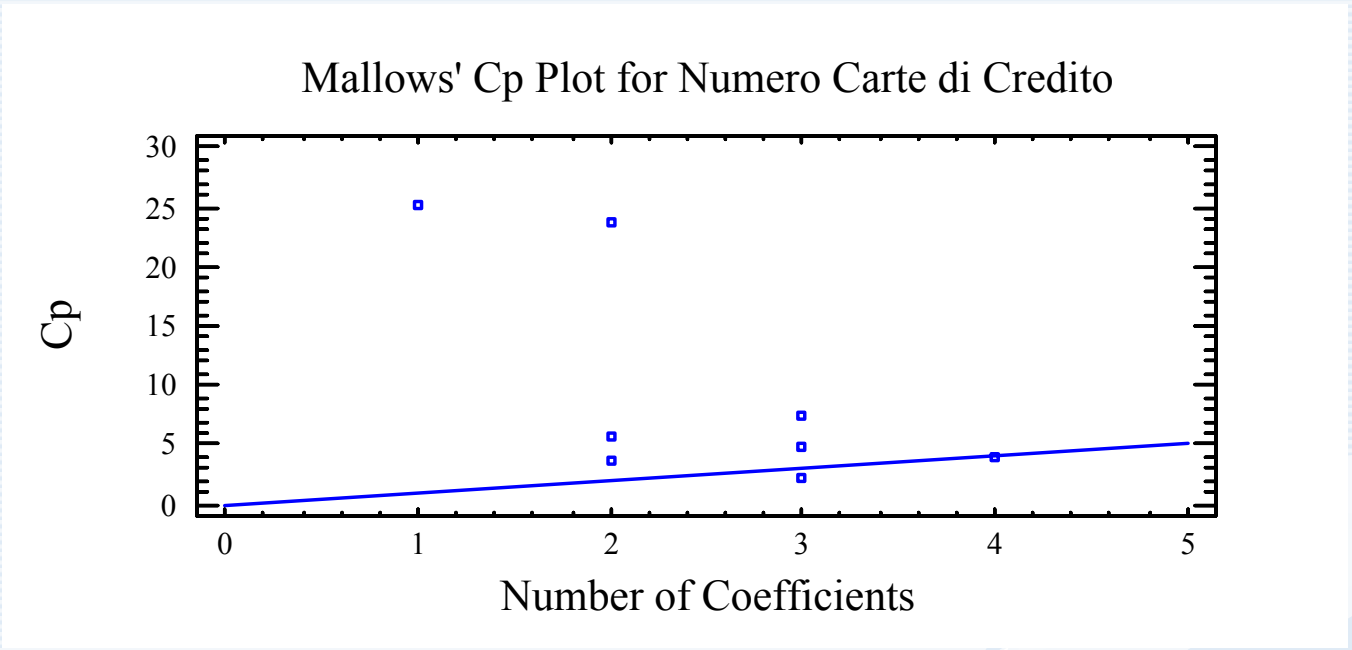
MSE	R-Squared	Adjusted R-Squared	Cp	Included Variables
3,14	0,0	0,0	25,2622	
0,91	75,0649	70,91	3,79524	A
3,14	14,2857	0,0	23,6081	B
1,15	68,7292	63,52	5,77594	C
0,96	78,2743	69,58	4,79194	AB
0,61	86,1376	80,59	2,33369	AC
1,33	69,6696	57,54	7,48195	BC
0,70	87,205	77,61	4,0	ABC



$$\overline{R}^2 = \left(R^2 - \frac{m}{n-1} \right) \frac{n-1}{n-m-1}$$



$$R^2 = \frac{\text{Devianza di regressione}}{\text{Devianza totale}}$$



Cp is a measure of the bias in the model based on a comparison of total Mean Squared Error to the true error variance.

Unbiased models have an expected Cp value of approximately p, where p is the number of coefficients in the fitted model. Cp is based on the assumption that the model that contains all the candidate variables is unbiased; therefore, the full model will always have Cp = p. Look for models that have Cp values close to p.

CONTROLLO D'IPOTESI SUL MODELLO:
esiste un legame effettivo tra la variabile dipendente e i regressori?

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$$

$$F = \frac{\frac{Dev(Y)_{regr}}{m}}{\frac{Dev(Y)_{residua}}{n-m-1}} = \frac{Var(Y)_{regr}}{Var(Y)_{residua}} = \frac{SSR/m}{MSE} = 9{,}09$$

Da confrontare con il valore tabulato

$$F_{\frac{0{,}05}{2};3;4} = 9{,}98$$

$$F_{\frac{0{,}10}{2};3;4} = 6{,}59$$

Dev (Y) = 22

Dev (Y) *regressione* = *19,185*

Dev (Y) residua = *2,815*

Stima Intervallare dei Coefficienti di Regressione

	Inferiore 95%	*Superiore 95%*	*Inferiore 90,0%*	*Superiore 90,0%*
Intercetta	-4,17	4,74	-3,14	3,71
Ampiezza della Famiglia	-0,12	1,39	0,06	1,21
Reddito della Famiglia (in migliaia di €)	-0,13	0,53	-0,06	0,45
Numero di auto della famiglia	-1,03	1,58	-0,73	1,27

$$\left[\left[B_i - t_{(\frac{\alpha}{2};n-m-1)} * \sqrt{\mathrm{var}(B_i)} \;;\; B_i + t_{(\frac{\alpha}{2};n-m-1)} * \sqrt{\mathrm{var}(B_i)} \right]\right]$$

Esempio di Calcolo per il coefficiente della Variabile Ampiezza della Famiglia

	Coefficiente	**t di Student**	**Standard Error**	**Limite Inferiore**	**Limite Superiore**
90%	0,635	2,132	0,271	0,057	1,212
95%	0,635	2,776	0,271	-0,118	1,387

Diagnostica di regressione

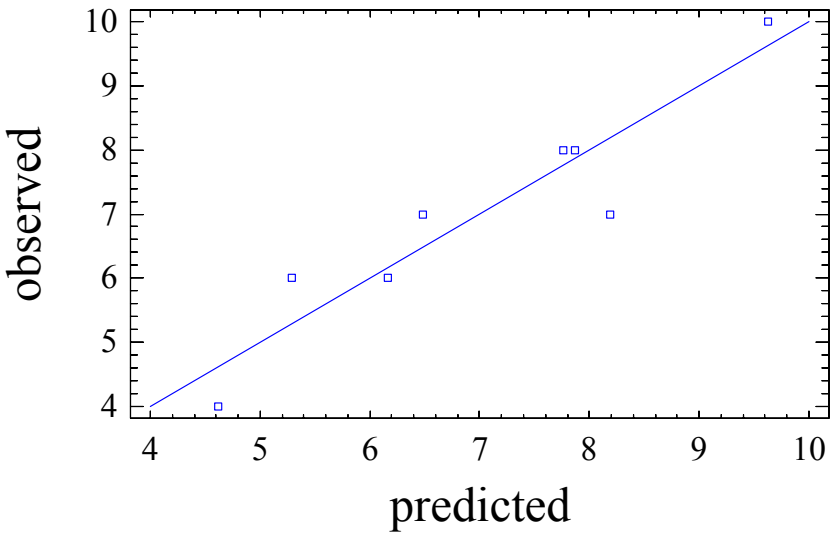
Regression Results for Numero Carte di Credito

N° Oss.	Y	Y predetta	Errore di regressione
1	4,0	4,62019	-0,620192
2	6,0	5,29087	0,709135
3	6,0	6,16106	-0,161058
4	7,0	6,48798	0,512019
5	8,0	7,86538	0,134615
6	7,0	8,19231	-1,19231
7	8,0	7,75721	0,242788
8	10,0	9,625	0,375

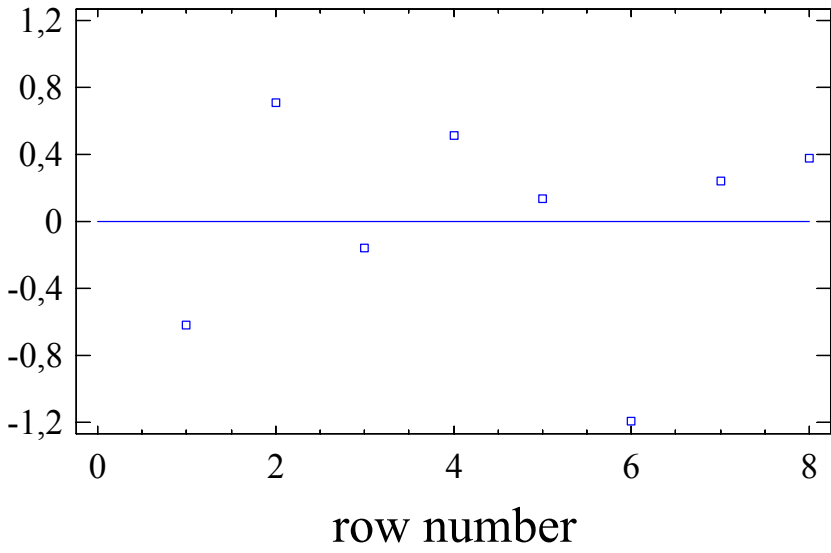
$$d = \frac{\sum_{i=1}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} (e_i)^2}$$

DW= 2,47 (Assenza di correlazione tra i residui)

Plot of Numero Carte di Credito



Residual Plot



Un'azienda propone un nuovo prodotto, una barretta energetica. Tale barretta viene commercializzata in 34 negozi ad un certo prezzo e sulla base di una certa attività promozionale (sia il prezzo che la promozione sono diversi per negozio).

Sulla base delle osservazioni relative a

- numero di barrette vendute in un mese (Y)
- prezzo in centesimi di una barretta (X_1)
- spesa mensile per le attività promozionali, in euro (X_2)

l'azienda vuole stabilire l'effetto che il prezzo e le promozioni all'interno dei negozi possono avere sulle vendite.

Matrice di covarianza S

	Y	X1	X2
Y	1578596	-14600,4	106272,7
X1		265,24	-249,14
X2			26524,06

Matrice di correlazione R

	Y	X1	X2
	1	-0,735	0,535
		1	-0,097
			1

Il modello stimato da luogo ai seguenti risultati:

$\hat{\beta}_0 = 5837{,}52$: numero medio di barrette che ci si aspetterebbe di vendere ogni mese se il prezzo e l'ammontare delle spese in promozione fossero entrambi pari a zero (?!?)...

$\hat{\beta}_1 = -53{,}22$: per un dato ammontare della spesa in promozioni pubblicitarie, il numero di barrette vendute in un mese diminuisce di 53,22 per ogni centesimo di aumento del prezzo.

$\hat{\beta}_2 = 3{,}61$: per un dato prezzo, il numero di barrette vendute aumenta di 3,61 per ogni euro speso in più in attività promozionali

modello stimato:

$\hat{Y}_i = 5837{,}52 - 53{,}2173X_1 + 3{,}6131X_2$

Stima di σ^2: $\hat{\sigma}^2 = 638{,}07$

$R^2 = 0{,}758$ $R^2_{adj} = 0{,}742$

Il 74,21% della variabilità delle vendite può essere spiegato dal modello proposto.

Errore standard del vettore dei coefficienti

E' una matrice (p+1)x(p+1) data dall'espressione

$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Tuttavia σ^2 non è nota. Al fine di ottenere una stima della varianza dello stimatore, sostituiamo σ^2 con $\hat{\sigma}^2$:

$S^2_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$

$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 5837{,}52 \\ -53{,}22 \\ 3{,}61 \end{bmatrix}$ $S^2_{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} 628{,}15 \\ 6{,}85 \\ 0{,}68 \end{bmatrix}$

Verifica dell'ipotesi di significatività del modello

$F_{obs} = 48{,}48 \qquad F_{0{,}05;2,31} \approx 3{,}32$

Poiché $F_{obs} > F_{\alpha;p,n-p}$ rifiutiamo l'ipotesi nulla di assenza di relazione tra il numero di barrette vendute e il prezzo e/o la spesa promozionale.

ANOVA

Modello	Somma dei quadrati	df	Media dei quadrati	F	Sig.
Regressione	39472730,773	2	19736365,387	48,477	,000
Residuo	12620946,668	31	407127,312		
Totale	52093677,441	33			

Verifica dell'ipotesi di significatività di singoli coefficienti

$$t_{obs} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = -7{,}77 < -2{,}04 = t_{31;0.025} \quad \rightarrow \quad \text{rif. } H_0$$

$$t_{obs} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = 5{,}27 > 2{,}04 = t_{31;0.025} \quad \rightarrow \quad \text{rif. } H_0$$

- Vi è una correlazione significativa tra il prezzo (X1) e le vendite, dato l'ammontare della spesa promozionale.
- Vi è una correlazione significativa tra la spesa promozionale (X2) e le vendite, dato il prezzo.

Intervalli di confidenza per i parametri

$$-67{,}19 \leq \beta_1 \leq -39{,}24$$

$$2{,}22 \leq \beta_2 \leq 5{,}01$$

- Dato l'effetto della spesa promozionale, l'aumento di un centesimo del prezzo delle barrette determina una riduzione del numero di barrette vendute compresa tra 67 e 39 (nel 95% dei casi…)

- Dato l'effetto del prezzo, per ciascun euro di aumento delle spese promozioni il numero di barrette vendute aumenta di un ammontare compreso tra 2,2 e 5 barrette

Se vogliamo **prevedere il numero di barrette vendute** in un negozio in cui il prezzo praticato è stato di 79 centesimi e la spesa in promozioni è stata pari a 400€ in un mese, possiamo utilizzare il modello stimato:

$$\hat{Y}_i = 5837{,}52 - 53{,}21(79) + 3{,}61(400) = 3077{,}93$$

intervallo di confidenza: [1758,4399]