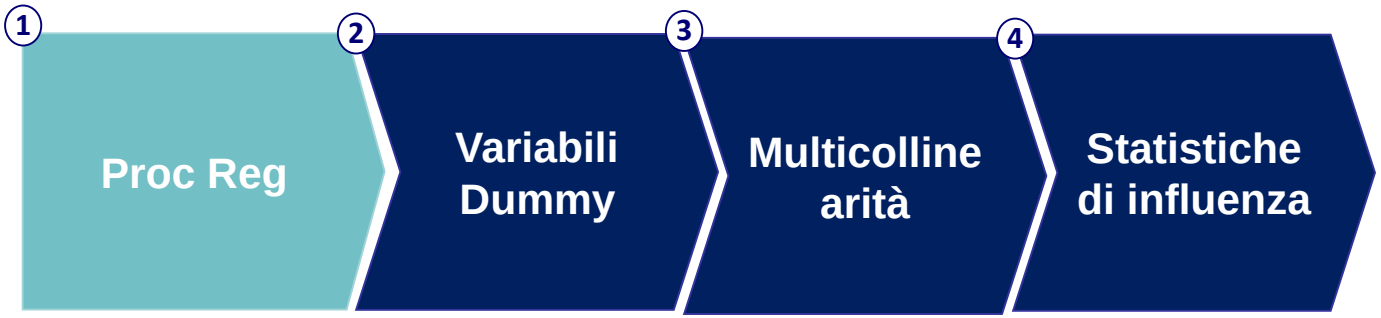
Regressione lineare

Metodi Quantitativi per Economia, Finanza e Management

Esercitazione n°9

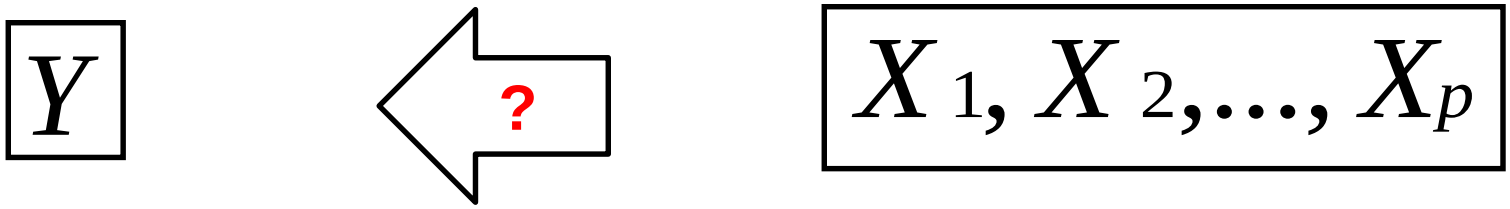
Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Modello di Regressione Lineare

I **modelli di Regressione Lineare** costituiscono una classe di modelli utili alla rappresentazione di relazioni di dipendenza non simmetriche tra variabili.



Variabile «target»: rappresenta un fenomeno di interesse (variabile quantitativa continua)

Variabili che si ritiene possano influenzare Y

OBIETTIVO:

Individuare quali variabili tra X1,…,Xp (variabili «indipendenti») influenzano la variabile Y (variabile «dipendente») e come la influenzano

Modello di Regressione Lineare

Y	**X_1**	**X_2**	**X_3**	…	…	…	**X_p**
y_1	x_{11}	x_{12}	x_{13}	…	…	…	x_{1p}
y_2	x_{21}	x_{22}	x_{23}	…	…	…	x_{2p}
y_3	x_{31}	x_{32}	x_{33}	…	…	…	x_{3p}
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
y_n	x_{n1}	x_{n2}	x_{n3}	…	…	…	x_{np}
(nx1)				(nxp)			

- n righe → n unità statistiche
- una colonna di n misurazioni sulla variabile dipendente Y (quantitativa continua)
- p colonne corrispondenti alle variabili indipendenti (X_1,…,X_p) (consideriamo variabili di natura quantitativa)
- in corrispondenza di ogni riga abbiamo (p+1) misurazioni: (y_i,x_{i1},x_{i2},x_{i3},…,x_{ip}) i=1,…,n

Modello di Regressione Lineare

Vogliamo descrivere la relazione esistente tra la variabile dipendente Y e le variabili indipendenti (X1,…Xp) tramite una funzione lineare.

Equazione di regressione lineare multipla

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

- Y_i: i-esima oss. su Y
- β_0: intercetta
- β_1: coefficiente di X1
- X_{i1}: i-esima oss. su X1
- ε_i: errore relativo all'i-esima oss.

PROC REG – Esempio

Variabile dipendente (soddisfazione globale) e 9 regressori (variabili indipendenti)

Nome variabile	Descrizione variabile
AltriOperatori_2	Livello di soddisfazione relativo ai costi verso altri operatori
assistenza_2	Livello di soddisfazione relativo al servizio di assistenza
Autoricarica_2	Livello di soddisfazione relativo alla possibilità di autoricarica
CambioTariffa_2	Livello di soddisfazione relativo alla facilità di cambiamento della tariffa
ChiamateTuoOperatore_2	Livello di soddisfazione relativo alla possibilità di effettuare chiamate a costi inferiori verso numeri dello stesso operatore
ComodatoUso_2	Livello di soddisfazione relativo alla possibilità di rivecere un cellulare in comodato d'uso
CostoMMS_2	Livello di soddisfazione relativo al costo degli MMS
Promozioni_2	Livello di soddisfazione relativo alla possibilità di attivare promozioni sulle tariffe
vsPochiNumeri_2	Livello di soddisfazione relativo alle agevolazioni verso uno o più numeri di telefono
soddisfazione_globale	Livello di soddisfazione globale relativo al telefono cellulare

PROC REG – Sintassi

Modello di regressione lineare – a partire da *p* regressori (variabili indipendenti)

```
proc reg data=dataset;
   model variabile_dipendente=
         regressore_1 ... regressore_p
    /option(s);
run;
quit;
```

PROC REG – Esempio

Modello di regressione lineare→ variabile dipendente= SODDISFAZIONE_GLOBALE, regressori= 9 variabili di soddisfazione (livello di soddisfazione relativo a tariffe, promozioni, ecc.)

```
proc reg data= corso.telefonia ;

model soddisfazione_globale  =

CambioTariffa_2  ComodatoUso_2
AltriOperatori_2 assistenza_2
ChiamateTuoOperatore_2  Promozioni_2
Autoricarica_2  CostoMMS_2  vsPochiNumeri_2

/ stb ;
run;
quit;
```

VARIABILE DIPENDENTE

REGRESSORI

opzione per ottenere i coefficienti standardizzati

Valutazione modello

Valutazione della bontà del modello (output della PROC REG)

- Coefficiente di determinazione **R-quadro** per valutare la capacità esplicativa del modello → capacità di rappresentare la relazione tra la variabile dipendente e i regressori
  (varia tra 0 e 1, quanto più si avvicina ad 1 tanto migliore è il modello)

- Test F per valutare la significatività congiunta dei coefficienti (se il p-value del test è inferiore al livello di significatività fissato, rifiuto l'ipotesi che i coefficienti siano tutti nulli →il modello ha capacità esplicativa)

- Test **t** per valutare la significatività dei *singoli* coefficienti
  (se il p-value del test è inferiore al livello di significatività fissato, rifiuto l'ipotesi di coefficiente nullo → il regressore corrispondente è rilevante per la spiegazione della variabile dipendente)

PROC REG – Output

attenzione!! → se la variabile dipendente o almeno uno dei regressori contiene un valore mancante, SAS scarta l'intero record nella stima del modello

Number of Observations Read	236
Number of Observations Used	235
Number of Observations with Missing Values	1

Root MSE	0.88676	R-Square	0.5949
Dependent Mean	6.49362	Adj R-Sq	0.5787
Coeff Var	13.65594		

Il modello è abbastanza buono, spiega il 60% della variabilità della variabile dipendente.
Quanto più R-Square si avvicina ad 1 tanto migliore è il modello.

PROC REG – Output

Test F per valutare la significatività congiunta dei coefficienti

$$H_0 : \beta_1 = \ldots = \beta_p = 0$$

$$H_1 : \text{almeno un } \beta_j \neq 0$$

Fissato un livello di significatività pari a 0.05, il p-value associato al test F è < 0.05 → Rifiuto l'ipotesi H0 → il modello ha capacità esplicativa

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	259.81139	28.86793	36.71	<.0001
Error	225	176.92903	0.78635		
Corrected Total	234	436.74043			

PROC REG – Output

Test t per valutare la significatività dei singoli coefficienti

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

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	**t Value**	**Pr > |t|**	Standardized Estimate
Intercept	Intercept	1	1.65529	0.29996	**5.52**	**<.0001**	0
CambioTariffa_2	CambioTariffa_2	1	0.11838	0.03178	**3.72**	**0.0002**	0.19265
ComodatoUso_2	ComodatoUso_2	1	0.07490	0.02702	**2.77**	**0.0060**	0.12760
AltriOperatori_2	AltriOperatori_2	1	0.08957	0.03285	**2.73**	**0.0069**	0.13297
assistenza_2	assistenza_2	1	0.10472	0.03507	**2.99**	**0.0031**	0.14126
ChiamateTuoOperatore_2	ChiamateTuoOperatore_2	1	0.20969	0.03571	**5.87**	**<.0001**	0.29775
Promozioni_2	Promozioni_2	1	0.17453	0.03962	**4.41**	**<.0001**	0.25256
Autoricarica_2	Autoricarica_2	1	-0.00168	0.02660	**-0.06**	**0.9498**	-0.00300
CostoMMS_2	CostoMMS_2	1	0.00981	0.02765	**0.35**	**0.7230**	0.01612
vsPochiNumeri_2	vsPochiNumeri_2	1	0.01571	0.03012	**0.52**	**0.6024**	0.02457

PROC REG – Output

Fissato un livello di significatività pari a 0.05, il p-value associato al test t è < 0.05 → Rifiuto l'ipotesi H0 di coefficiente nullo → il regressore corrispondente è rilevante per la spiegazione della variabile dipendente

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	**t Value**	**Pr > |t|**	Standardized Estimate
Intercept	Intercept	1	1.65529	0.29996	**5.52**	**<.0001**	0
CambioTariffa_2	CambioTariffa_2	1	0.11838	0.03178	**3.72**	**0.0002**	0.19265
ComodatoUso_2	ComodatoUso_2	1	0.07490	0.02702	**2.77**	**0.0060**	0.12760
AltriOperatori_2	AltriOperatori_2	1	0.08957	0.03285	**2.73**	**0.0069**	0.13297
assistenza_2	assistenza_2	1	0.10472	0.03507	**2.99**	**0.0031**	0.14126
ChiamateTuoOperatore_2	ChiamateTuoOperatore_2	1	0.20969	0.03571	**5.87**	**<.0001**	0.29775
Promozioni_2	Promozioni_2	1	0.17453	0.03962	**4.41**	**<.0001**	0.25256
Autoricarica_2	Autoricarica_2	1	-0.00168	0.02660	**-0.06**	**0.9498**	-0.00300
CostoMMS_2	CostoMMS_2	1	0.00981	0.02765	**0.35**	**0.7230**	0.01612
vsPochiNumeri_2	vsPochiNumeri_2	1	0.01571	0.03012	**0.52**	**0.6024**	0.02457

PROC REG – Output

→ se il p-value associato al test t è > 0.05 (livello di significatività fissato a priori) si accetta l'ipotesi H0 di coefficiente nullo → il regressore corrispondente **NON** è rilevante per la spiegazione della variabile dipendente.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	1.65529	0.29996	**5.52**	**<.0001**	0
CambioTariffa_2	CambioTariffa_2	1	0.11838	0.03178	**3.72**	**0.0002**	0.19265
ComodatoUso_2	ComodatoUso_2	1	0.07490	0.02702	**2.77**	**0.0060**	0.12760
AltriOperatori_2	AltriOperatori_2	1	0.08957	0.03285	**2.73**	**0.0069**	0.13297
assistenza_2	assistenza_2	1	0.10472	0.03507	**2.99**	**0.0031**	0.14126
ChiamateTuoOperatore_2	ChiamateTuoOperatore_2	1	0.20969	0.03571	**5.87**	**<.0001**	0.29775
Promozioni_2	Promozioni_2	1	0.17453	0.03962	**4.41**	**<.0001**	0.25256
Autoricarica_2	Autoricarica_2	1	-0.00168	0.02660	**-0.06**	**0.9498**	-0.00300
CostoMMS_2	CostoMMS_2	1	0.00981	0.02765	**0.35**	**0.7230**	0.01612
vsPochiNumeri_2	vsPochiNumeri_2	1	0.01571	0.03012	**0.52**	**0.6024**	0.02457

Selezione regressori

- ✓ Nella scelta dei regressori bisogna cercare di mediare tra due esigenze:
  1) maggior numero di variabili per migliorare il fit
  2) parsimonia per rendere il modello più robusto e interpretabile

- ✓ Scelta dei regressori che entrano nel modello

metodi di selezione automatica

Selezione regressori

E’ possibile ricorrere a procedure di calcolo automatico per selezionare il sottoinsieme di regressori ottimale tra quelli possibili

• **forward selection** → inserisce nel modello una variabile per volta, scegliendo ad ogni passo il regressore che contribuisce maggiormente alla spiegazione della variabilità di Y

• **backward selection** → parte da un modello che considera tutti i regressori; rimuove dal modello una variabile per volta, scegliendo ad ogni passo il regressore che comporta la minor perdita di capacità esplicativa della variabilità di Y

• **stepwise selection** (forward+backward selection) → ogni variabile può entrare/uscire dal modello

Selezione Stepwise

Procedura sequenziale che valuta l'ingresso/uscita dal modello dei singoli regressori:

- test statistico (test «F parziale») che valuta la significatività del contributo del regressore alla spiegazione della variabilità di Y;
- vengono fissati a priori due livelli di significatività (ingresso/uscita)

- **Step 0**→si considerano tutti i potenziali regressori

- **Step 1**→entra il primo regressore. Ossia, viene stimato un modello contenente un unico regressore tra quelli proposti (viene scelto il regressore che dà il contributo maggiore alla spiegazione della variabilità, purché sia significativo)

- **Step 2**→si valutano tutti i possibili modelli contenenti il regressore individuato allo step 1 e uno dei rimanenti regressori, e si tiene il modello con il fit migliore (ossia entra il regressore che dà il contributo maggiore alla spiegazione della variabilità, purché sia significativo)

Selezione Stepwise

- **Step 3 e seguenti** →si valuta l'uscita di ognuno dei regressori presenti (in base alla minor perdita di capacità esplicativa del modello) e l'ingresso di un nuovo regressore (in base al maggior incremento nella capacità esplicativa del modello).

- **NB**: un regressore incluso ai passi precedenti può essere rimosso a seguito dell'inserimento di altri regressori che rendono non più significativo il suo contributo originale alla spiegazione della variabilità di Y

- **Criterio di arresto** → la procedura si arresta quando nessun regressore rimanente può essere inserito in base al livello di significatività scelto (slentry) e nessun regressore incluso può essere eliminato in base al livello di significatività scelto (slstay). In pratica quando non si riesce in alcun modo ad aumentare la capacità esplicativa del modello

Esercizio

Variabile dipendente (soddisfazione globale) e 21 regressori (variabili di soddisfazione)

Nome variabile	Descrizione variabile
soddisfazione_globale	Livello di soddisfazione globale relativo al telefono cellulare
AccessoWeb_2	Livello di soddisfazione relativo al costo di accesso a internet
AltriOperatori_2	Livello di soddisfazione relativo ai costi verso altri operatori
assistenza_2	Livello di soddisfazione relativo al servizio di assistenza
Autoricarica_2	Livello di soddisfazione relativo alla possibilità di autoricarica
CambioTariffa_2	Livello di soddisfazione relativo alla facilità di cambiamento della tariffa
ChiamateTuoOperatore_2	Livello di soddisfazione relativo alla possibilità di effettuare chiamate a costi inferiori verso numeri dello stesso operatore
ChiarezzaTariffe_2	Livello di soddisfazione relativo alla chiarezza espositiva delle tariffe
ComodatoUso_2	Livello di soddisfazione relativo alla possibilità di rivecere un cellulare in comodato d'uso
copertura_2	Livello di soddisfazione relativo alla copertura della rete
CostoMMS_2	Livello di soddisfazione relativo al costo degli MMS
CostoSMS_2	Livello di soddisfazione relativo al costo degli SMS
diffusione_2	Livello di soddisfazione relativo alla diffusione
DurataMinContratto_2	Livello di soddisfazione relativo alla presenza di una durata minima del contratto
immagine_2	Livello di soddisfazione relativo all'immagine
MMSTuoOperatore_2	Livello di soddisfazione relativo alla possibilità inviare MMS a costi inferiori verso numeri dello stesso operatore
NavigazioneWeb_2	Livello di soddisfazione relativo al costo di navigazione in internet
NoScattoRisp_2	Livello di soddisfazione relativo all'assenza di scatto alla risposta
NumeriFissi_2	Livello di soddisfazione relativo alle agevolazioni verso numeri fissi
Promozioni_2	Livello di soddisfazione relativo alla possibilità di attivare promozioni sulle tariffe
SMSTuoOperatore_2	Livello di soddisfazione relativo alla possibilità inviare SMS a costi inferiori verso numeri dello stesso operatore
vsPochiNumeri_2	Livello di soddisfazione relativo alle agevolazioni verso uno o più numeri di telefono

PROC REG – Sintassi

Modello di regressione lineare

```
proc reg data=dataset;
   model variabile_dipendente=
         regressore_1 ... regressore_p
    /option(s);
run;
```

OPTIONS:

- STB calcola i coefficienti standardizzati
- selection=stepwise applica la procedura stepwise per la selezione dei regressori
- slentry=… livello di significatività richiesto per il test F parziale affinché il singolo regressore possa entrare nel modello
- slstay=… livello di significatività richiesto per il test F parziale affinché il singolo regressore non sia rimosso dal modello

PROC REG – Esempio

Modello di regressione lineare→ variabile dipendente= SODDISFAZIONE_GLOBALE, regressori= 21 variabili di soddisfazione (livello di soddisfazione relativo a tariffe, promozioni, ecc.)

```
proc reg data= corso.telefonia;
 model soddisfazione_globale=
 CambioTariffa_2 ChiarezzaTariffe_2 …/stb

 selection=stepwise

slentry=0.05 slstay=0.05;
run;
quit;
```

VARIABILE DIPENDENTE = REGRESSORI

opzione per ottenere i coefficienti standardizzati

criterio di selezione automatica dei regressori

soglia di significatività per testare l'entrata di un regressore nel modello (valore di default=0.15)

soglia di significatività per testare l'uscita di un regressore dal modello (valore di default=0.15)

PROC REG – Output

Il metodo Stepwise seleziona 6 regressori tra le 21 variabili di soddisfazione

Fissato un livello di significatività pari a 0.05, il p-value associato al test t è < 0.05 → i regressori selezionati sono rilevanti per la spiegazione della variabile dipendente

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**
Intercept	**1**	1.70973	0.28331	6.03	<.0001	0
CambioTariffa_2	**1**	0.11876	0.03154	3.77	0.0002	0.19327
ComodatoUso_2	**1**	0.07698	0.02577	2.99	0.0031	0.13114
AltriOperatori_2	**1**	0.09132	0.03212	2.84	0.0049	0.13557
assistenza_2	**1**	0.10482	0.03476	3.02	0.0029	0.14139
ChiamateTuoOperatore_2	**1**	0.21579	0.0343	6.29	<.0001	0.30641
Promozioni_2	**1**	0.17767	0.03695	4.81	<.0001	0.2571

Interpretazione coefficienti

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- Il coefficiente esprime la variazione che subisce la variabile dipendente Y in seguito a una variazione unitaria del regressore, posto che il valore degli altri regressori rimanga costante

- ATTENZIONE!!→ i valori dei coefficienti dipendono dall'unità di misura della variabile a cui sono associati, quindi non sono direttamente confrontabili ed utilizzabili per stabilire un ordine di importanza tra i regressori rispetto all'impatto sulla variabile Y.

- in genere si considerano i coefficienti standardizzati (opzione STB della PROC REG) che non sono influenzati dall'unità di misura delle variabili

PROC REG – Output

se la variabile CambioTariffa_2 aumenta di una unità allora la soddisfazione globale aumenta del 19%

se la variabile CambioTariffa_2 diminuisce di una unità allora la soddisfazione globale diminuisce del 19%

N.B.:attenzione al segno del coefficiente!!

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**
Intercept	**1**	1.70973	0.28331	6.03	<.0001	0
CambioTariffa_2	**1**	0.11876	0.03154	3.77	0.0002	0.19327
ComodatoUso_2	**1**	0.07698	0.02577	2.99	0.0031	0.13114
AltriOperatori_2	**1**	0.09132	0.03212	2.84	0.0049	0.13557
assistenza_2	**1**	0.10482	0.03476	3.02	0.0029	0.14139
ChiamateTuoOperatore_2	**1**	0.21579	0.0343	6.29	<.0001	0.30641
Promozioni_2	**1**	0.17767	0.03695	4.81	<.0001	0.2571

PROC REG – Output

se il regressore3 aumenta di una unità allora la variabile dipendente diminuisce del 31%

se il regressore3 diminuisce di una unità allora la variabile dipendente aumenta del 31%

N.B.:attenzione al segno del coefficiente!!

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**
Intercept	**1**	1.71	0.283	6.03	<.0001	**0**
regressore 1	**1**	0.12	0.032	3.77	<.0001	**0.19**
regressore 2	**1**	0.08	0.026	2.99	<.0001	**0.13**
regressore 3	**1**	-0.22	0.034	6.29	<.0001	**-0.31**
regressore 4	**1**	0.18	0.037	4.81	<.0001	**0.26**

Importanza dei regressori

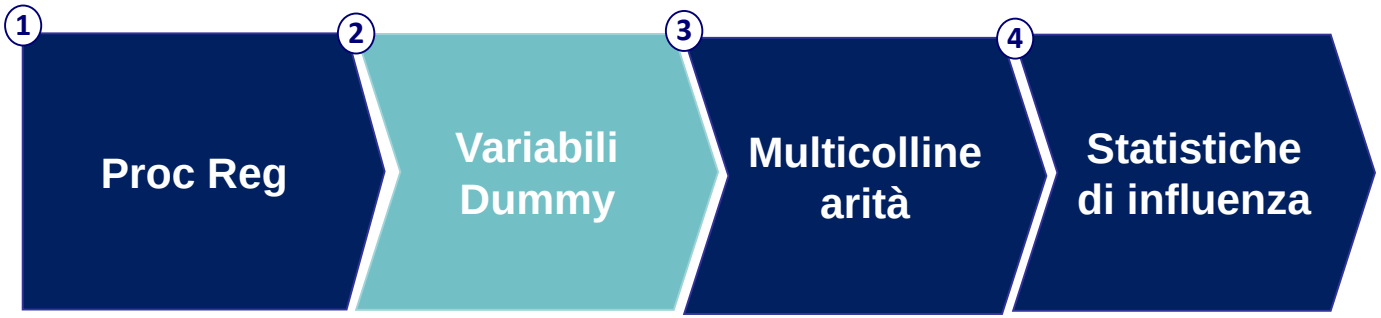
Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**
Intercept	**1**	1.71	0.283	6.03	<.0001	**0**
regressore 1	**1**	0.12	0.032	3.77	<.0001	**0.19**
regressore 2	**1**	0.08	0.026	2.99	<.0001	**0.13**
regressore 3	**1**	-0.22	0.034	6.29	<.0001	**-0.31**
regressore 4	**1**	0.18	0.037	4.81	<.0001	**0.26**

- I coefficienti standardizzati sono utili per valutare l'importanza relativa dei regressori. Possiamo ordinare i regressori in base all'importanza che hanno nello spiegare la variabile dipendente. Il regressore con valore assoluto del coefficiente standardizzato più alto è il più importante.

- Nell'esempio il regressore 3 è il più importante, poi il regressore 4, l'1 e infine il 2.

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Regressione lineare – Variabili qualitative

Considerazioni da fare <u>prima di stimare il modello</u>

- Non si possono inserire variabili qualitative tra i regressori
- Per considerare questo tipo di variabili all'interno del modello bisogna costruire delle variabili dummy (dicotomiche (0-1)) che identificano le modalità della variabile originaria.

Variabile qualitativa con k modalità → costruire (k-1) dummy

- Le variabili dummy saranno utilizzate come regressori.

Costruzione variabili dummy - esempio

Es. Si vuole considerare tra i regressori la variabile qualitativa nominale "Area" che identifica l'area di residenza degli intervistati

N° questionario	AREA
1	nord
2	nord
3	sud
4	nord
5	centro
6	nord
7	centro
8	sud
9	nord
10	centro

La variabile "Area" assume tre modalità (nord-centro-sud) → si costruiscono due variabili dummy

Costruzione variabili dummy - esempio

Le variabili dummy da costruire sono due (la terza sarebbe ridondante→può essere ottenuta come combinazione delle altre due)

- Area_nord→vale 1 se l'intervistato è residente al nord e 0 in tutti gli altri casi
- Area_centro→vale 1 se l'intervistato è residente al centro e 0 in tutti gli altri casi

Costruzione variabili dummy - esempio

N° questionario	AREA	AREA_NORD	AREA_CENTRO
1	nord	1	0
2	nord	1	0
3	sud	0	0
4	nord	1	0
5	centro	0	1
6	nord	1	0
7	centro	0	1
8	sud	0	0
9	nord	1	0
10	centro	0	1

VARIABILE ORIGINARIA (non entra nel modello)

VARIABILI DUMMY (entrano nel modello)

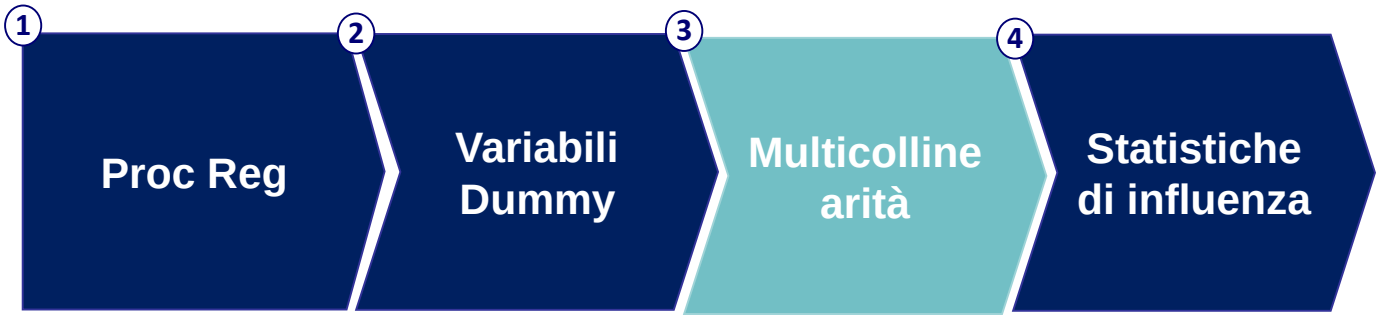
Costruzione variabili dummy - esempio

Nella PROC REG si inseriscono le due variabili dummy (ma non la variabile originaria!) nella lista dei regressori→i relativi coefficienti rappresentano l'effetto della singola modalità (nord/centro) della variabile "Area".

```
proc reg data= … ;
 model Y= X1 X2 … area_nord area_centro
 /stb;
run;
quit;
```

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Multicollinearità

Quando un regressore è combinazione lineare di altri regressori nel modello, le stime sono instabili e hanno standard error elevato. Questo problema è chiamato multicollinearità.

La PROC REG fornisce nell'output un indicatore per ogni regressore per investigare questo problema:

Variance Inflation Factors
(→opzione VIF nell'istruzione MODEL).

Multicollinearità

R2	VIF
0.1	1.11
0.2	1.25
0.3	1.43
0.4	1.67
0.5	2.00
0.6	2.50
0.7	3.33
0.8	5.00
0.9	10.00
0.95	20.00
0.98	50.00
0.99	100.00

Per verificare la presenza di multicollinearità:

- regressione lineare di Xj sui rimanenti p-1 regressori
  - Rj^2 misura la quota di varianza di Xj spiegata dai rimanenti p-1 regressori →
    valori > 0.2 / 0.3 → presenza di multicollinearità

  - $VIFj = 1 / (1 – Rj^2)$ misura il grado di relazione lineare tra Xj e i rimanenti p-1 regressori →
    valori > 1.2 / 1.3 → presenza di multicollinearità

PROC REG – Sintassi

Verifica presenza multicollinearità

```
proc reg data=dataset;
   model variabile_dipendente=
         regressore_1 ... regressore_p /VIF;
run;
```

per verificare presenza
di multicollinearietà

Esempio

Variabile dipendente (SODDISFAZIONE_GLOBALE) e 21 regressori (variabili di soddisfazione)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	1.05063	0.40052	2.62	0.0093	0	0
CambioTariffa_2	1	0.12034	0.0331	3.64	0.0003	0.19584	1.63324
MMSTuoOperatore_2	1	-0.00139	0.01586	-0.09	0.9303	-0.00425	1.32504
copertura_2	1	0.06579	0.04557	1.44	0.1503	0.07419	1.48622
NoScattoRisp_2	1	-0.02286	0.02539	-0.9	0.3689	-0.04257	1.25835
Autoricarica_2	1	-0.00947	0.02736	-0.35	0.7295	-0.01698	1.35409
CostoMMS_2	1	0.00949	0.03211	0.3	0.768	0.01558	1.56654
NumeriFissi_2	1	0.0584	0.03599	1.62	0.1062	0.08448	1.52619
DurataMinContratto_2	1	0.03014	0.03124	0.96	0.3358	0.04964	1.49001
vsPochiNumeri_2	1	-0.01002	0.03212	-0.31	0.7555	-0.01566	1.42001
diffusione_2	1	0.05247	0.05206	1.01	0.3147	0.0565	1.76896
ComodatoUso_2	1	0.06531	0.02891	2.26	0.0249	0.11126	1.36501
ChiarezzaTariffe_2	1	0.06117	0.03412	1.79	0.0744	0.10058	1.77144
AccessoWeb_2	1	0.02487	0.05946	0.42	0.6762	0.04024	5.21015
AltriOperatori_2	1	0.06777	0.03564	1.9	0.0586	0.1006	1.57539
SMSTuoOperatore_2	1	0.01755	0.03696	0.47	0.6354	0.02923	2.13318
assistenza_2	1	0.0504	0.04082	1.23	0.2183	0.06798	1.70656
immagine_2	1	0.01288	0.04614	0.28	0.7803	0.01638	1.9376
ChiamateTuoOperatore_2	1	0.15362	0.04424	3.47	0.0006	0.21813	2.22145
Promozioni_2	1	0.14316	0.0426	3.36	0.0009	0.20717	2.13881
CostoSMS_2	1	0.02739	0.04167	0.66	0.5117	0.04154	2.24805
NavigazioneWeb_2	1	-0.04249	0.06017	-0.71	0.4809	-0.06822	5.25292

Alcuni dei VIFj presentano valori alti

Multicollinearità

Esempio

Possibile risoluzione: utilizzo dell'analisi fattoriale

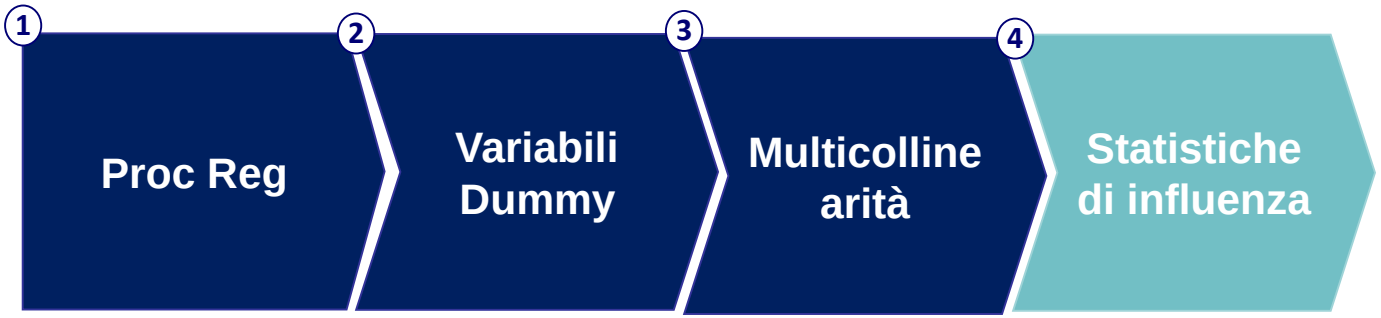
Variabile dipendente (SODDISFAZIONE_GLOBALE) e 6 fattori creati con un'analisi fattoriale sulle 21 variabili di soddisfazione

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	6.49839	0.05783	112.38	<.0001	0	0
Factor1	1	0.51102	0.05838	8.75	<.0001	0.37142	1.00102
Factor2	1	0.437	0.05822	7.51	<.0001	0.31847	1.00080
Factor3	1	0.06409	0.05821	1.1	0.272	0.04672	1.00079
Factor4	1	0.69395	0.05813	11.94	<.0001	0.50651	1.00064
Factor5	1	0.24529	0.05833	4.2	<.0001	0.17843	1.00096
Factor6	1	0.32203	0.05782	5.57	<.0001	0.23622	1.00000

L'analisi fattoriale ci permette di trasformare i regressori in componenti non correlate e risolvere il problema della multicollinearità. Tutti i Variance Inflation Factors sono prossimi a 1, cioè l'Rj^2 della regressione lineare di Xj sui rimanenti p-1 regressori è prossimo a zero.

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Osservazioni influenti

- Se un valore y_j è particolarmente distante rispetto a tutti gli altri allora la stima del modello di regressione può essere notevolmente influenzata da tale osservazione.
- Per valutare la presenza di osservazioni influenti si elimina una osservazione per volta e si stima nuovamente il modello.
- Osservazioni la cui esclusione produce variazioni rilevanti nelle stime dei coefficienti sono dette ***influenti***

Statistiche di influenza

Misure di influenza:

- **Distanza di Cook** : misura la distanza tra la stima dei coefficienti senza l'*i-esima* osservazione e con l'*i-esima* osservazione.

→ Le unità per cui $\mathbf{D_i > 1}$ sono potenzialmente osservazioni influenti

- **Leverage H** :

→ Le unità per cui $\mathbf{H_i > 2*(p+1)/n}$ sono potenzialmente osservazioni influenti (dove p è il numero di regressori e n il numero di osservazioni)

Statistiche di influenza

Sintassi

La PROC REG fornisce nell'output i valori della distanza di Cook e del levarage H per ogni osservazione del dataset:

```
proc reg data=dataset noprint;
  model variabile_dipendente=
        regressore_1 ... regressore_p
   / influence;
output out=dataset_output cookd=cook H=leverage;
run;
```

OPTIONS:

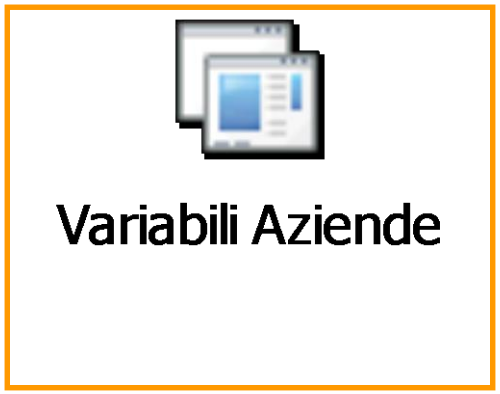
- Influence fornisce una serie di indicatori di influenza tra cui D e H
- Cookd= crea nel dataset di output una variabile con i valori della Distanza di Cook per ogni osservazione
- H= crea nel dataset di output una variabile con i valori del Leverage per ogni osservazione
- Noprint = utile soprattutto per dataset con molte informazioni, permette di non stampare l'output

Esempio

- Il data set AZIENDE contiene informazioni relative ai comportamenti di 500 clienti del segmento Aziende di una banca.

- L'obiettivo è stimare il margine totale del cliente

- Viene stimato un modello di regressione lineare con 66 variabili dipendenti → tramite selezione automatica delle variabili, vengono selezionati 12 regressori rilevanti



Esempio

Output analisi influenza :

	NDG cliente	Cook's D Influence Statistic	Leverage	flag_cook	flag_leverage
101	297752	0.0000206016	0.0091951557	0	0
102	988113	3.362712E-10	0.0032931669	0	0
103	764449	5.9803595E-7	0.0029707706	0	0
104	416050	3.2010873E-7	0.0032221258	0	0
105	466496	0.0000787361	0.0057903792	0	0
106	495415	0.0005268767	0.0124103938	0	0
107	192174	9.466682E-7	0.0042966574	0	0
108	718322	0.0006957082	0.0082709555	0	0
109	402419	0.0000142729	0.0040465626	0	0
110	691241	1.2718634378	0.8093134316	1	1
111	441493	0.0006199363	0.0096788994	0	0
112	23857	9.0029983E-6	0.0029266512	0	0
113	931377	1.0003113E-6	0.0034783268	0	0

Distanza di Cook>1 e Leverage >0.052 =2*(12+1)/500 ⟶ oss influente

```
proc reg data=corso.aziende noprint;
 model tot_margine= lista 12 regressori /stb influence ;
output out=corso.aziende_out cookd=cook H=leverage;
run;
```

Eliminazione osservazioni influenti

Come si individuano e eliminano le osservazioni influenti (quelle con Distanza di Cook>1 e Leverage >0.052)?

```
data CORSO.AZIENDE_NEW;

   set corso.aziende_out;

  where cook<=1 or leverage<=0.052;
run;
```

Nuovo Dataset

Dataset calcolato nella proc reg con opzioni: *influence, Cookd, H e output out*

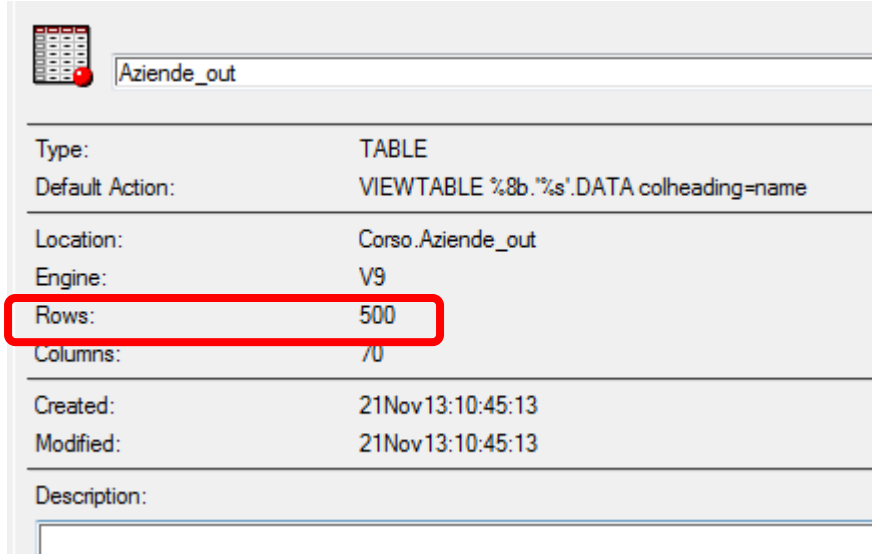
Soglia per la statistica Cook (fissa)

Vogliamo tenere tutte le osservazioni che soddisfano la statistica di Cook ***OPPURE*** la statistica di Leverage

Soglia per la statistica Leverage (variabile)

Eliminazione osservazioni influenti

Quante osservazioni influenti sono state eliminate?



Leggere il LOG e confrontare la numerosità del data set CORSO.AZIENDE_NEW rispetto alla numerosità dataset corso.aziende_out (proprietà della tabella).

```
21    data CORSO.AZIENDE_NEW;
22      set corso.aziende_out;
23      where cook<=1 or leverage<=0.52;
24    run;

NOTE: There were 499 observations read from the data set CORSO.AZIENDE_OUT.
      WHERE (cook<=1) or (leverage<=0.52);
NOTE: The data set CORSO.AZIENDE_NEW has 499 observations and 70 variables.
NOTE: DATA statement used (Total process time):
      real time           0.24 seconds
      cpu time            0.03 seconds
```

Esempio

Output ristima coefficienti di regressione al netto della osservazione influente :

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	**1**	13.02789	53.52084	0.24	0.8078	0
PROFT_T1	Reddività totale al T1	**1**	1.72412	0.02464	69.98	<.0001	0.89474
_cs_np12	Possesso Risparmio Gestito	**1**	811.93397	184.46258	4.4	<.0001	0.03174
racc_dir_t1	Raccolta diretta al T1	**1**	-0.00282	0.00067655	-4.17	<.0001	-0.05604
scanumt2	Scost. Ass N. Prod. Cross Selling	**1**	-158.54071	21.75354	-7.29	<.0001	-0.0508
sccnmov1	C.Correnti - Numero movimenti	**1**	7.30311	1.21192	6.03	<.0001	0.05551
sccvsma1	C.Correnti - Saldo Contabile Medio Avere	**1**	0.00729	0.00121	6.04	<.0001	0.08317
sccvsmd1	C.Correnti - Saldo Contabile Medio Dare	**1**	0.00457	0.00081505	5.61	<.0001	0.05708
sesinae1	Estero: Anticipi in Euro Import	**1**	0.04441	0.01022	4.35	<.0001	0.03441
sesoptot1	Estero: Operatività Totale	**1**	-0.00662	0.00206	-3.21	0.0014	-0.02624

```
proc reg data=aziende_new;
 model tot_margine= lista 66 regressori
/stb selection= stepwise slentry=0.01 slstay=0.01;
run;
```

N.B.: aziende_new è lo stesso dataset iniziale SENZA l'osservazione influente

PROC REG – Riepilogo

1. Individuazione variabili dipendente e regressori
2. Trasformazione di eventuali variabili qualitative in dummy
3. Stimare un modello di regressione lineare utilizzando la procedura automatica di selezione delle variabili (stepwise)
4. Valutare la bontà del modello (R-square, Test F, Test t)
5. Analisi di influenza con i soli regressori scelti nella stepwise.
   - ✓ Se si è in presenza di osservazioni influenti: eliminarle e ripetere i punti 3 e 4
   - ✓ In assenza di osservazioni influenti: passare al punto 6

PROC REG – Riepilogo

6. Verificare la presenza di multicollinearità (se i regressori del modello sono i fattori di un'analisi fattoriale non è necessario perchè risultano non correlati per costruzione → tutti i VIFj =1)
   - ✓ Se si è in presenza di multicollinearità: azioni per eliminarla e ripetere i punti 3, 4, 5
   - ✓ In assenza di multicollinearità: passare al punto 7
7. Verificare l'impatto dei regressori nella spiegazione del fenomeno (ordinarli usando il valore assoluto dei coefficienti standardizzati e controllare il segno dei coefficienti)
8. Interpretazione del coefficienti standardizzati