

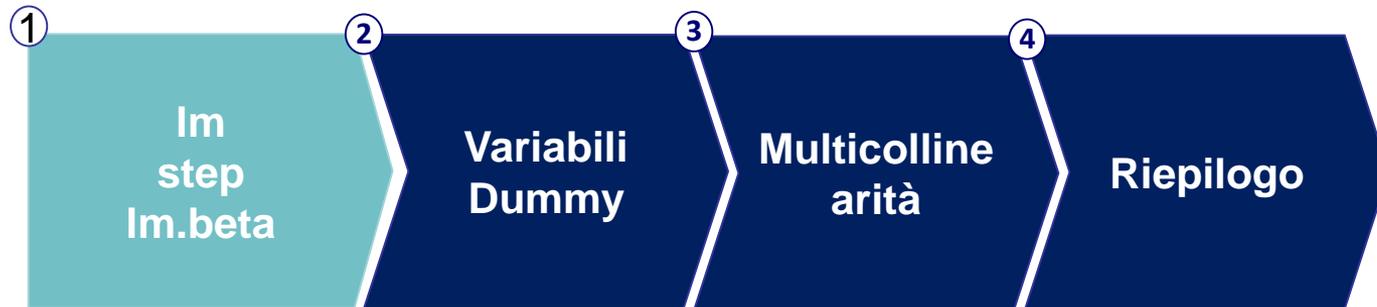
Regressione lineare

*Metodi Quantitativi per Economia,
Finanza e Management*

Esercitazione n°9

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:

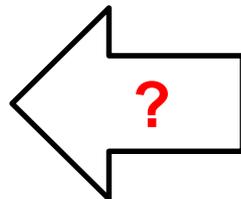


Modello di Regressione Lineare

L'analisi della regressione lineare è una metodologia asimmetrica che si basa sull'ipotesi dell'esistenza di una relazione di tipo **causa-effetto** tra una o più variabili indipendenti (o esplicative, X_i) e la variabile dipendente (Y).

Y

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



X_1, X_2, \dots, X_p

Variabili che si ritiene possano
influenzare Y

OBIETTIVO:

Individuare quali variabili tra X_1, \dots, X_p (variabili «indipendenti») influenzano la variabile Y (variabile «dipendente») e come la influenzano



Modello di Regressione Lineare

<u>Y</u>	<u>X₁</u>	<u>X₂</u>	<u>X₃</u>	<u>X_p</u>
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
- Y = variabile quantitativa continua oggetto dell'analisi
- p colonne corrispondenti alle variabili indipendenti (X₁, ..., 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

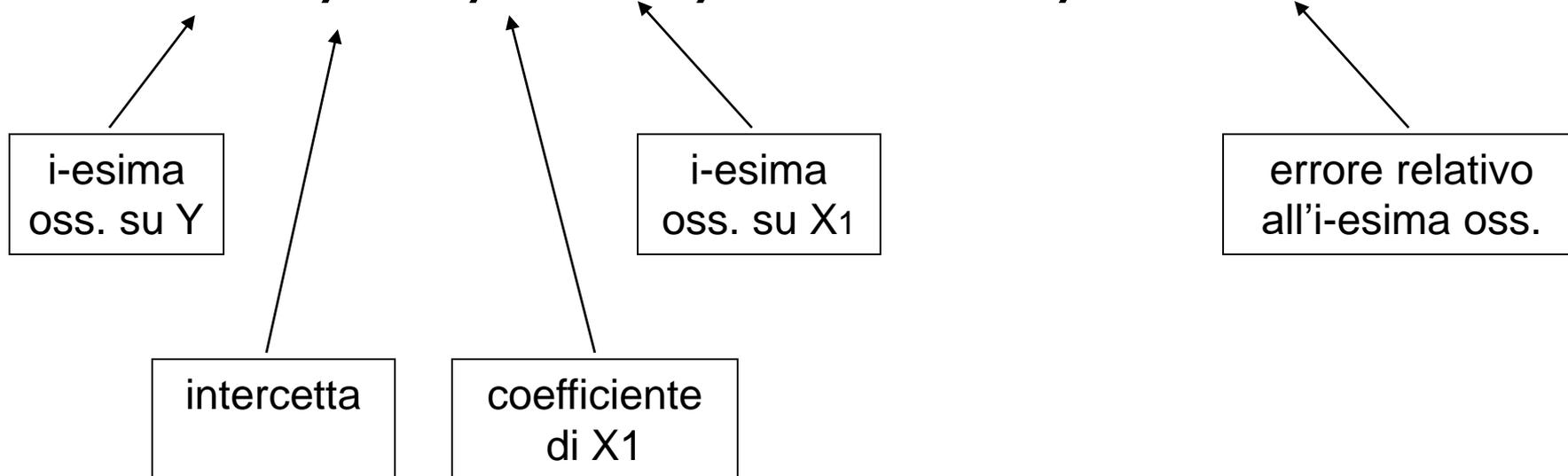


Modello di Regressione Lineare

Vogliamo descrivere la relazione esistente tra la variabile dipendente Y e le variabili indipendenti (X_1, \dots, X_p) tramite una funzione lineare.

Equazione di regressione lineare multipla

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



Im – 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 rivedere 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



lm – Sintassi

La funzione che in R calcola il modello di regressione lineare è la ***lm (linear model)***.

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

```
Nome_oggetto = lm (variabile_dipendente  
~ variabile_indipendente,  
data=dataset_input)
```



Im – Esempio

Modello di regressione lineare:

Variabile dipendente= SODDISFAZIONE_GLOBALE,

Regressori= 9 variabili di soddisfazione (livello di soddisfazione relativo a tariffe, promozioni, ecc.)

VARIABILE DIPENDENTE

Soddisfazione = Im (soddisfazione_globale
~ CambioTariffa_2 + ComodatoUsa_2 +
AltriOperatori_2 + assistenza_2 +
ChiamateTuoOperatore_2 + Promozioni_2 +
Autoricarica_2 + CostoMMS_2 +
vsPochiNumeri_2, *data=telefonia*)

DATASET DI INPUT DEI DATI

R
E
G
R
E
S
S
O
R
I



Valutazione modello

Valutazione della bontà del modello (**output della Im**)

- **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). **R-quadro adjusted**, come R^2 ma indipendente dal numero di regressori
- **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)



Im – Output

Summary (soddisfazione)

```
> summary(soddisfazione)
```

Call:

```
lm(formula = soddisfazione_globale ~ CambioTariffa_2 + ComodatoUso_2 +  
  AltriOperatori_2 + assistenza_2 + ChiamateTuoOperatore_2 +  
  Promozioni_2 + Autoricarica_2 + CostoMMS_2 + vsPochiNumeri_2,  
  data = telefonia)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.05388	-0.54353	0.02425	0.55546	2.86406

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.655292	0.299960	5.518	9.38e-08	***
CambioTariffa_2	0.118383	0.031784	3.725	0.000247	***
ComodatoUso_2	0.074904	0.027020	2.772	0.006035	**
AltriOperatori_2	0.089570	0.032851	2.727	0.006904	**
assistenza_2	0.104721	0.035066	2.986	0.003136	**
ChiamateTuoOperatore_2	0.209688	0.035710	5.872	1.53e-08	***
Promozioni_2	0.174532	0.039617	4.405	1.63e-05	***
Autoricarica_2	-0.001675	0.026596	-0.063	0.949838	
CostoMMS_2	0.009812	0.027645	0.355	0.722991	
vsPochiNumeri_2	0.015711	0.030118	0.522	0.602437	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8868 on 225 degrees of freedom
(1 observation deleted due to missingness)

Multiple R-squared: 0.5949, Adjusted R-squared: 0.5787

F-statistic: 36.71 on 9 and 225 DF, p-value: < 2.2e-16

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

FORMULA

STIMA DEI
COEFFICIENTI
DEL MODELLO

TEST
STATISTICI

Im – Output – TEST STATISTICI

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

```
Residual standard error: 0.8868 on 225 degrees of freedom  
(1 observation deleted due to missingness)  
Multiple R-squared: 0.5949, Adjusted R-squared: 0.5787  
F-statistic: 36.71 on 9 and 225 DF, p-value: < 2.2e-16
```

R² : 0.5949, Il modello è abbastanza buono, spiega il 60% della variabilità della variabile dipendente. Quanto più R-Squared si avvicina ad 1 tanto migliore è il modello.

R² corretto: il 57.87% della variabilità della soddisfazione può essere spiegato dal modello proposto, tenuto conto del numero di regressori e dell'ampiezza campionaria



Im – Output – TEST STATISTICI

Test F per valutare la significatività congiunta dei coefficienti

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

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

Residual standard error: 0.8868 on 225 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.5949, Adjusted R-squared: 0.5787
F-statistic: 36.71 on 9 and 225 DF, p-value: < 2.2e-16

Test F: 36.71 e rispettivo p-value < 0.05.

Fissato un livello di significatività α pari a 0.05, il p-value associato al test F è < 0.05, quindi Rifiuto l'ipotesi H_0 .

Il modello ha capacità esplicativa



Im – Output – STIMA DEI COEFFICIENTI DEL MODELLO

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Test t per valutare la significatività dei singoli coefficienti

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.655292	0.299960	5.518	9.38e-08	***
CambioTariffa_2	0.118383	0.031784	3.725	0.000247	***
ComodatoUsa_2	0.074904	0.027020	2.772	0.006035	**
AltriOperatori_2	0.089570	0.032851	2.727	0.006904	**
assistenza_2	0.104721	0.035066	2.986	0.003136	**
ChiamateTuoOperatore_2	0.209688	0.035710	5.872	1.53e-08	***
Promozioni_2	0.174532	0.039617	4.405	1.63e-05	***
Autoricarica_2	-0.001675	0.026596	-0.063	0.949838	
CostoMMS_2	0.009812	0.027645	0.355	0.722991	
vsPochiNumeri_2	0.015711	0.030118	0.522	0.602437	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R identifica con gli * il livello di significatività del p-value associato al test T:

- se il p-value è <0.05, 1 asterisco
- se p-value<0.01, 2 asterischi
- se p-value<0.001, 3 asterischi



Im – Output – STIMA DEI COEFFICIENTI DEL MODELLO

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	1.655292	0.299960	5.518	9.38e-08	***						
CambioTariffa_2	0.118383	0.031784	3.725	0.000247	***						
ComodatoUsa_2	0.074904	0.027020	2.772	0.006035	**						
AltriOperatori_2	0.089570	0.032851	2.727	0.006904	**						
assistenza_2	0.104721	0.035066	2.986	0.003136	**						
ChiamateTuoOperatore_2	0.209688	0.035710	5.872	1.53e-08	***						
Promozioni_2	0.174532	0.039617	4.405	1.63e-05	***						
Autoricarica_2	-0.001675	0.026596	-0.063	0.949838							
CostoMMS_2	0.009812	0.027645	0.355	0.722991							
vsPochiNumeri_2	0.015711	0.030118	0.522	0.602437							

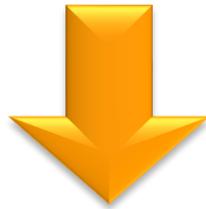
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

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.



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 (sl_{entry}) e nessun regressore incluso può essere eliminato in base al livello di significatività scelto (sl_{stay}). 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 rivedere 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



step – Sintassi

1. **Stimo il modello di regressione lineare**
2. **Uso la funzione step per stimare il modello con il metodo stepwise**

```
Nome_modello_lm = step(nome_dataset_lm,  
direction='both')
```

applica la procedura stepwise per la
selezione dei regressori



step – Esempio

Modello di regressione lineare:

Variabile dipendente= SODDISFAZIONE_GLOBALE,

Regressori= 21 variabili di soddisfazione (livello di soddisfazione relativo a tariffe, promozioni, ecc.)

1.

```
soddisfazione2 = lm ( soddisfazione_globale ~ AccessoWeb_2 +  
AltriOperatori_2 + assistenza_2 + Autoricarica_2 +  
CambioTariffa_2 + ChiamateTuoOperatore_2 +  
ChiarezzaTariffe_2 + ComodatoUso_2 + copertura_2 + CostoMMS_2  
+ CostoSMS_2 + diffusione_2 + DurataMinContratto_2 +  
immagine_2 + MMSTuoOperatore_2 + NavigazioneWeb_2 +  
NoScattoRisp_2 + NumeriFissi_2 + Promozioni_2 +  
SMSTuoOperatore_2 + vsPochiNumeri_2, data=telefonia)
```

2. `p<-step(soddisfazione2, direction='both')`

`summary(p)`

**criterio di selezione
automatica dei regressori**



step – Output

Il primo Output di R mostra tutti i vari passaggi della stepwise.
L'output della summary, invece, mostra il modello ottimale scelto dalla procedura.

```
> summary(p)
```

Call:

```
lm(formula = soddisfazione_globale ~ AltriOperatori_2 + CambioTariffa_2 +  
  ChiamateTuoOperatore_2 + ChiarezzaTariffe_2 + ComodatoUso_2 +  
  copertura_2 + diffusione_2 + NumeriFissi_2 + Promozioni_2,  
  data = telefonia)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.74320	-0.57538	-0.00447	0.48309	2.50958

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.95386	0.38247	2.494	0.013353	*
AltriOperatori_2	0.06421	0.03355	1.914	0.056899	.
CambioTariffa_2	0.12878	0.03143	4.097	5.84e-05	***
ChiamateTuoOperatore_2	0.17471	0.03693	4.730	3.96e-06	***
ChiarezzaTariffe_2	0.07886	0.03200	2.464	0.014488	*
ComodatoUso_2	0.06857	0.02572	2.667	0.008221	**
copertura_2	0.10223	0.04026	2.539	0.011784	*
diffusione_2	0.07766	0.04274	1.817	0.070574	.
NumeriFissi_2	0.04926	0.03380	1.457	0.146395	
Promozioni_2	0.14375	0.03935	3.653	0.000322	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8686 on 225 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.6113, Adjusted R-squared: 0.5958

F-statistic: 39.32 on 9 and 225 DF, p-value: < 2.2e-16

Il metodo Stepwise
seleziona 9 regressori tra
le 21 variabili di
soddisfazione, di cui 6
sono significative



step – Output

```
> summary(p)
```

```
Call:
lm(formula = soddisfazione_globale ~ AltriOperatori_2 + CambioTariffa_2 +
    ChiamateTuoOperatore_2 + ChiarezzaTariffe_2 + ComodatoUso_2 +
    copertura_2 + diffusione_2 + NumeriFissi_2 + Promozioni_2,
    data = telefonia)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-2.74320 -0.57538 -0.00447  0.48309  2.50958
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.95386	0.38247	2.494	0.013353	*
AltriOperatori_2	0.06421	0.03355	1.914	0.056899	.
CambioTariffa_2	0.12878	0.03143	4.097	5.84e-05	***
ChiamateTuoOperatore_2	0.17471	0.03693	4.730	3.96e-06	***
ChiarezzaTariffe_2	0.07886	0.03200	2.464	0.014488	*
ComodatoUso_2	0.06857	0.02572	2.667	0.008221	**
copertura_2	0.10223	0.04026	2.539	0.011784	*
diffusione_2	0.07766	0.04274	1.817	0.070574	.
NumeriFissi_2	0.04926	0.03380	1.457	0.146395	
Promozioni_2	0.14375	0.03935	3.653	0.000322	***

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.8686 on 225 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared:  0.6113,    Adjusted R-squared:  0.5958
F-statistic: 39.32 on 9 and 225 DF,  p-value: < 2.2e-16
```

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



Interpretazione coefficienti

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

- ❑ 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 (**Im.beta in R**) che non sono influenzati dall'unità di misura delle variabili



Im.beta – Sintassi

Per calcolare le stime standardizzate dei coefficienti, è necessario scaricare un pacchetto: **QuantPsyc** e richiamarlo.

Successivamente si potrà usare la funzione **Im.beta**

```
library(QuantPsyc)
```

```
Im.beta(nome_modello_lm)
```



lm.beta – Interpretazione output

Interpretiamo solo i coefficienti delle variabili che nell'output della regressione lineare erano significativi (p-value < 0.05).

```
> lm.beta(p)
  AltriOperatori_2      0.09532746
  ComodatoUso_2       0.11681706
  Promozioni_2        0.20801836
  CambioTariffa_2     0.20958047
  copertura_2         0.11527455
  ChiamateTuoOperatore_2 0.24807429
  diffusione_2        0.08362539
  ChiarezzaTariffe_2  0.12966234
  NumeriFissi_2       0.07125797
```

Se la variabile CambioTariffa_2 aumenta di una unità allora la soddisfazione globale aumenta di 0.20 punti.

Se la variabile CambioTariffa_2 diminuisce di una unità allora la soddisfazione globale diminuisce di 0.20 punti.

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



Im.beta – Esempio Output

Se il regressore3 aumenta di una unità allora la variabile dipendente diminuisce di 0,31

Se il regressore3 diminuisce di una unità allora la variabile dipendente aumenta di 0,31

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

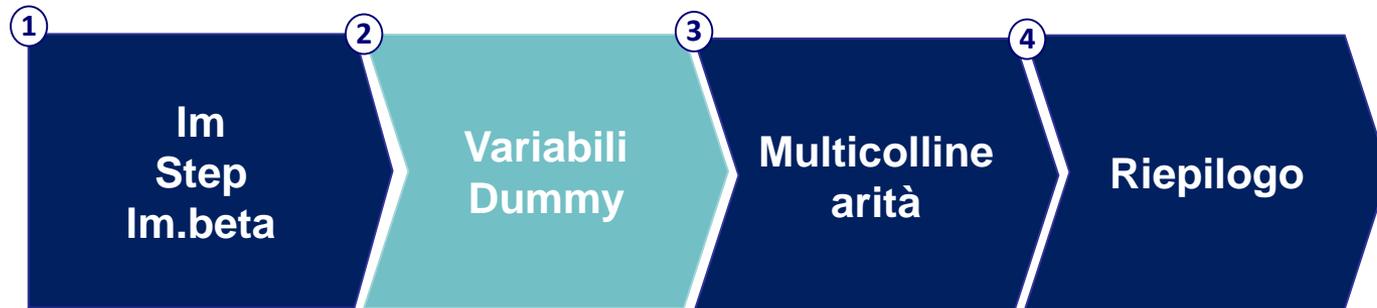
Variable	DF	Parameter Estimates				Standardized Estimate
		Parameter Estimate	Standard Error	t Value	Pr > t	
Intercept	1	1.71	0.283	6.03	<.0001	0
regressore 1	1	0.12	0.032	3.77	<.0001	0.19
regressore 2	1	0.08	0.026	2.99	<.0001	0.13
regressore 3	1	-0.22	0.034	6.29	<.0001	-0.31
regressore 4	1	0.18	0.037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

- 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 lm 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", rispetto alla modalità che non è stata inserita nel modello (sud).

```
area= lm ( y ~ x1 x2 ... area_nord area_centro,  
data=dataset_input)
```

```
summary(area)
```



Interpretazione variabili dummy

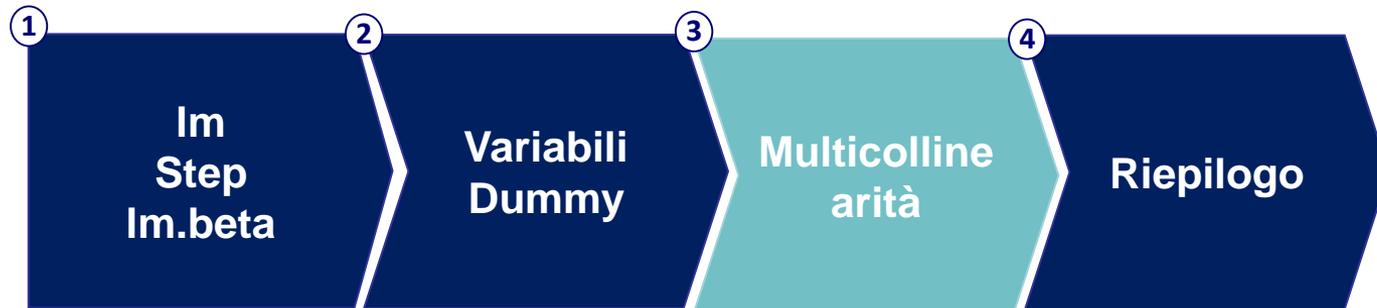
$Soddisfazione_globale = b_0 + area_nord * b_1 + area_centro * b_2 + chiamate_estero * b_3 + err$

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	1	0.7	0.283	6.03	<.0001	0
Area_nord	1	1.8	0.032	3.77	<.0001	0.30
Area_centro	1	-0.8	0.026	2.99	<.0001	-0.19
Chiamate_estero	1	-0.3	0.034	6.29	<.0001	-0.22

- ❑ A **parità di altre condizioni**, chi abita al nord ha un incremento della soddisfazione globale di 0.30 punti rispetto a chi abita al sud
- ❑ A **parità di altre condizioni**, chi abita al centro ha un decremento della soddisfazione globale di 0.19 punti rispetto a chi abita al sud

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.

VIF: indicatore che serve per individuare la presenza di multicollinearità ed è calcolato per ciascuna variabile del modello.

Variance Inflation Factors

$VIF > 1.2$ o $1.3 =$ multicollinearità (nella pratica $VIF > 2$)



Multicollinearità

R ²	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 X_j sui rimanenti $p-1$ regressori

- R_j^2 misura la quota di varianza di X_j spiegata dai rimanenti $p-1$ regressori →

valori > 0.2 / 0.3 → presenza di multicollinearità

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



vif – Sintassi

Verifica presenza multicollinearità

Per calcolare l' indicatore VIF, è necessario scaricare un pacchetto: **usdm** e richiamarlo.

Successivamente si potrà usare la funzione **vif**

```
library(usdm)
```

```
vif(nome_subset_input)
```



vif – Sintassi

Verifica presenza multicollinearità

Creiamo prima il subset delle sole variabili su cui vogliamo verificare la presenza di multicollinearità.

```
tel = telefonia[,c("AccessoWeb_2", "AltriOperatori_2", "assistenza_2",  
"Autoricarica_2", "CambioTariffa_2", "ChiamateTuoOperatore_2",  
"ChiarezzaTariffe_2", "ComodatoUso_2", "copertura_2", "CostoMMS_2",  
"CostoSMS_2", "diffusione_2", "DurataMinContratto_2", "immagine_2",  
"MMSTuoOperatore_2", "NavigazioneWeb_2", "NoScattoRisp_2",  
"NumeriFissi_2", "Promozioni_2", "SMSTuoOperatore_2",  
"vsPochiNumeri_2")]
```



vif – Sintassi

Verifica presenza multicollinearità

`vif(tel)`

```
> vif(tel)
```

	Variables	VIF
1	AccessoWeb_2	5.237226
2	AltriOperatori_2	1.575417
3	assistenza_2	1.719988
4	Autoricarica_2	1.363115
5	CambioTariffa_2	1.642613
6	ChiamateTuoOperatore_2	2.152962
7	ChiarezzaTariffe_2	1.780802
8	ComodatoUso_2	1.356863
9	copertura_2	1.485851
10	CostoMMS_2	1.575195
11	CostoSMS_2	2.246523
12	diffusione_2	1.776925
13	DurataMinContratto_2	1.496311
14	immagine_2	1.949202
15	MMSTuoOperatore_2	1.328668
16	NavigazioneWeb_2	5.257093
17	NoScattoRisp_2	1.258780
18	NumeriFissi_2	1.526655
19	Promozioni_2	2.153628
20	SMSTuoOperatore_2	2.113280
21	vsPochiNumeri_2	1.418556

Alcuni dei VIF_j
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

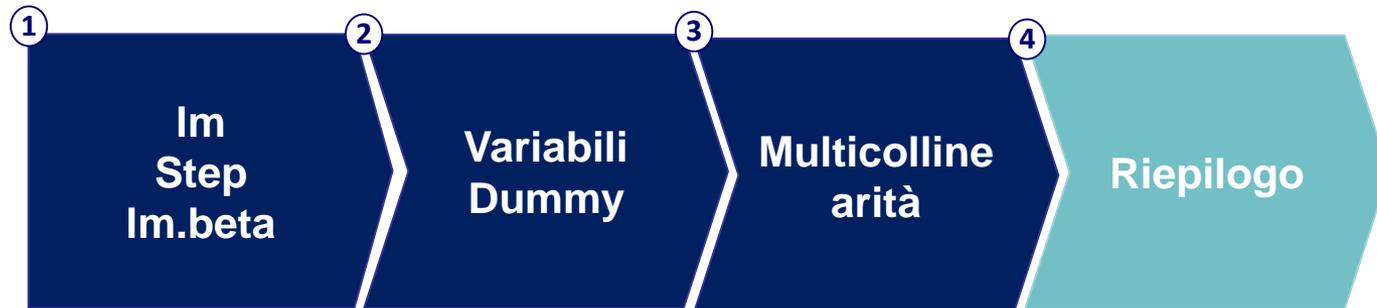


L'analisi fattoriale ci permette di trasformare i regressori in componenti non correlate e risolvere il problema della multicollinearità. Tutti i Variance Inflation Factors sono prossimi a 1, cioè l' R_j^2 della regressione lineare di X_j sui rimanenti $p-1$ regressori è prossimo a zero.



Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Regressione lineare – 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. Se la procedura stepwise non ha prodotto tutte stime significative, provare a stimare un modello di regressione lineare con i soli parametri le cui stime sono significative. Tornare al punto 4, poi al punto 6.



Regressione lineare – 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 $VIF_j = 1$)
 - ✓ Se si è in presenza di multicollinearità: azioni per eliminarla e ripetere i punti 3, 4
 - ✓ 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 dei coefficienti standardizzati

