Lezione 8

La regressione multipla:
Modello statistico e assunzioni

Argomenti della lezione:

- **Il modello teorico**
- **Il calcolo dei parametri**
- **Assunzioni e verifica**

Nella regressione multipla abbiamo una variabile dipendente che regredisce su almeno due variabili indipendenti

Equazione di regressione:

$$Y' = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

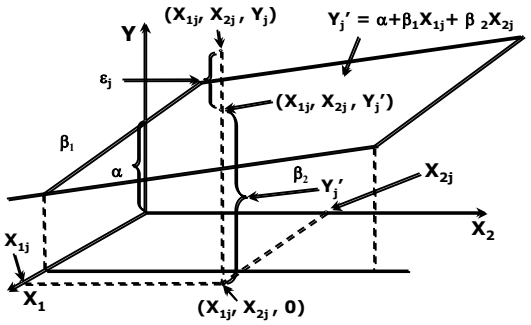
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon_i$$

Piano di regressione (per due VI) o iperpiano (per più di 2 VI)

Equazione del piano di regressione:

$$Y' = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Il piano di regressione



Coefficienti di regressione multipla: coefficienti "parziali"

Per ogni VI rappresentano:

→ **l'influenza della VI sulla VD, al netto delle altre VI**

Ovunque

Nell'equazione di regressione multipla $Y' = \alpha + \beta_1 X_1 + \beta_2 X_2$:

- β_1 rappresenta l'inclinazione della retta di regressione di Y su X_1 quando si mantiene costante X_2
- β_2 rappresenta l'inclinazione della retta di regressione di Y su X_2 quando si mantiene costante X_1

Stime dei coefficienti: minimi quadrati multipli

$\Sigma[Y - (\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)]^2 = \min$

Individuare un iperpiano di dimensioni k che si adatti meglio ai punti dispersi in uno spazio di dimensioni k+1 (k VI e 1 VD)

Espressioni matriciali:

equazione di regressione

$y = bX + e$

coefficienti di regressione

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

residui

$e = Y - (Xb + a)$

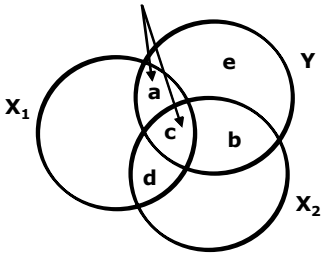
Misure di associazione tra VI e VD

Consideriamo:

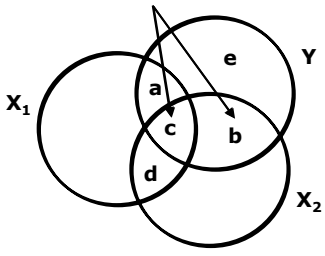
2 VI: X_1 e X_2

1 VD: Y

a+c = varianza in comune tra X_1 e Y

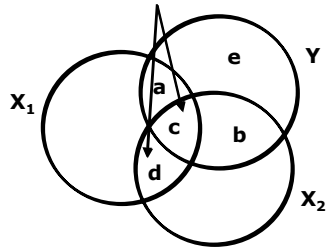


c+b = varianza in comune tra X_2 e Y

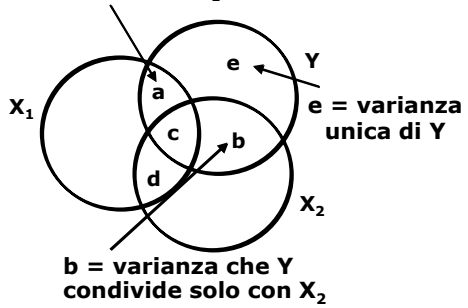


Ovunque

d+c = varianza in comune tra X_1 e X_2



a = varianza che Y condivide solo con X_1



e = varianza unica di Y

b = varianza che Y condivide solo con X_2

Correlazione Semi - parziale:

$$sr_{y1.2} = \frac{r_{y1} - r_{y2}\, r_{12}}{\sqrt{1 - r_{12}^2}}$$

Correlazione tra X_1 e Y, quando X_2 viene parzializzata solo da X_1

$$sr^2_{y1.2} = a/(a+c+b+e)$$

proporzione della varianza totale di Y spiegata unicamente da X_1 al netto di X_2

Correlazione Parziale:

$$pr_{y1.2} = \frac{r_{y1} - r_{y2}\, r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}}$$

Correlazione tra X_1 e Y, quando X_2 viene parzializzata da Y e da X_1

$$pr^2_{y1.2} = a/(a+e)$$

proporzione della varianza di Y non spiegata da X_2, spiegata unicamente da X_1 al netto di X_2

Coefficiente di regressione:

$$b_{y1.2} = \frac{b_{y1} - b_{y2} b_{12}}{1 - r_{12}^2}$$

$$\beta_{y1.2}\hat{} = b_{y1.2} \frac{s_1}{s_y} = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2}$$

Cambiamento atteso in Y in seguito ad un cambiamento di una unità (b) o di una deviazione standard (β^) in X_1 al netto di X_2

Varianza spiegata

- **Coefficiente di correlazione multiplo (R o RM): associazione tra una VD e un insieme di VI**
- **Coefficiente di determinazione multiplo (R^2): proporzione di varianza della VD spiegata dalle VI prese nel loro complesso**

Ovunque

$$R^2_{y.12\ldots k} = \sum r_{yi}\ \beta_{yi}\ \hat{}$$

nel caso di due VI la formula è:

$$R^2_{y.12} = r_{y1}\ \beta_{y1}\ \hat{} + r_{y2}\ \beta_{y2}\ \hat{}$$

Somma dei prodotti delle correlazioni semplici e dei coefficienti β^ tra la VD e ogni VI

Il coefficiente di correlazione multiplo si ottiene da R^2:

$$R_{y.12\ldots k} = \sqrt{R^2_{y.12\ldots k}}$$

Coefficiente di determinazione multiplo corretto (Adjusted)

$$AR^2 = R^2 - (1-R^2)*(k/(N-k-1))$$

Significatività statistica del coefficiente di determinazione (R^2)

Ipotesi statistiche:

H_0: $\rho = 0$;
(equivale a H_0: $\beta_1=\beta_2= \ldots =\beta_k=0$)

H_1: $\rho > 0$

Varianza	Somme dei Quadrati	GDL	Stime della Varianza
Totale	Σy^2	N-1	
Spiegata	$R^2\Sigma y2$	K	$R^2\Sigma y^2/k$
Non Sp.	$(1-R^2)\Sigma y^2$	N-k-1	$(1-R^2)\Sigma y^2/(N-k-1)$

$$F = R^2(N-k-1)/(1- R^2)k$$

Significatività statistica della differenza dei singoli b da 0

H_0: $\beta = 0$; H_1: $\beta \neq 0$ (<, >)

Test appropriato su dati campionari:

$$t = (b - 0)/S_b$$

con N-k-1 gradi di libertà

S_{bi} è la stima campionaria dell'errore standard di β definita come:

$$S_{bi} = \frac{S_y}{S_i}\sqrt{\frac{1-R^2_y}{N-k-1}}\sqrt{\frac{1}{1-R^2_i}}$$

Ovunque

Errore standard della stima ed errore standard dei parametri

$$S_e = \sqrt{\frac{\sum(Y-Y')^2}{N - k - 1}}$$

$$S_{bi} = \sqrt{\frac{S_e^2}{\sum(X - \bar{X})^2 (1 - R_i^2)}}$$

Assunzioni alla base della regressione multipla

→ **Assenza errore di specificazione**

- **Relazione tra le X_i e Y: lineare**
- **Non sono state omesse variabili indipendenti rilevanti**
- **Non sono state incluse variabili indipendenti irrilevanti**

→ **Assenza di errore di misurazione: X_i e Y misurate senza errore**

→ **Le VI sono quantitative o dicotomiche, la VD è quantitativa**

→ **Le varianze sono > 0**

→ **Il campionamento è casuale**

→ **Assenza di multicollinearità**

→ **Assunzioni sui residui**

- **Media uguale a zero: $E(e_i)=0$**
- **Omoschedasticità, $VAR(e_i) = \sigma 2$**
- **Assenza di autocorrelazione: $Cov(e_i, e_j) = 0$**
- **Non correlazione tra VI e residui: $Cov(e_i, X_i) = 0$**
- **Normalità**

Violazione delle assunzioni

Esame della distribuzione dei residui e = (Y-Y') rispetto ai punteggi teorici Y'

Utile per rilevare:

→ **La non linearità**
→ **La non omogeneità della varianza**
→ **La non normalità dei residui**

Ovunque

Violazione delle assunzioni

La Multicollinearità (correlazione elevata tra i predittori) può essere rilevata:

- → **dalle correlazioni tra le VI (>.8)**
- → **da R^2 elevati e β bassi**
- → **da errori standard elevati**
- → **dagli indici di tolleranza e VIF**

Indice di tolleranza: quantità di varianza di una variabile indipendente non spiegata dalle altre variabili indipendenti

$$T_i = (1 - R_i^2)$$

dove R_i^2 è il coefficiente di determinazione nella regressione della variabile indipendente i sulle altre variabili indipendenti

Il Variance Inflaction Factor (VIF)

costituisce il reciproco dell'indice di tolleranza

$$VIF_i = 1/T_i = 1/(1 - R_i^2)$$

Valori bassi ⇒ bassa collinearità
valori alti ⇒ elevata collinearità

Non indipendenza degli errori (Autocorrelazione), rilevata tramite:

- → **l'esame dei residui (Y-Y') rispetto all'ordine di acquisizione dei dati**
- → **il Test di Durbin-Watson**

Test di Durbin-Watson:
ha un valore compreso tra 0 e 4

Se n>100 e le VI sono almeno 2:

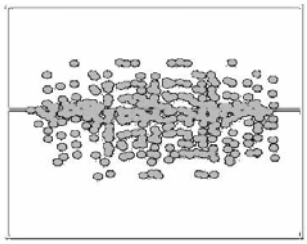
Valori tra 1.5 e 2.2 ⇒ assenza di autocorrelazione

Valori < 1.5 ⇒ autocorr. positiva
Valori > 2.2 ⇒ autocorr. negativa

Esame della distribuzione dei residui:

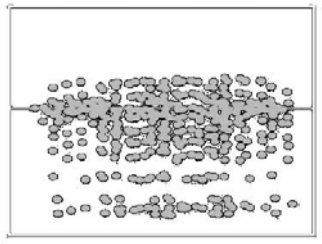
Alcuni esempi

Ovunque

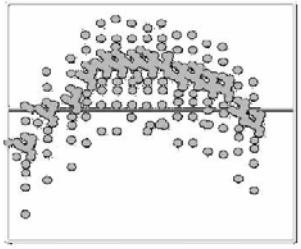


Assunzioni rispettate

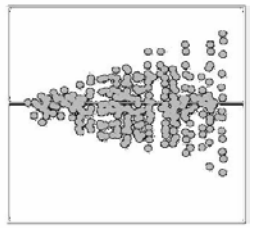
Punteggi predetti Y': in ascisse
Residui (Y-Y'): in ordinate



Punteggi predetti Y': in ascisse
Residui (Y-Y'): in ordinate

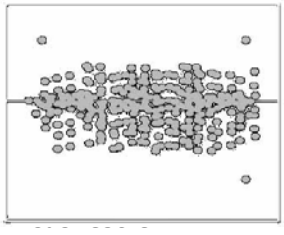


Punteggi predetti Y': in ascisse
Residui (Y-Y'): in ordinate

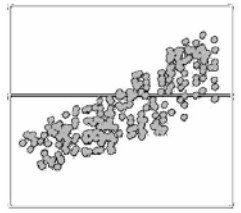


Eteroschedasticità

Punteggi predetti Y': in ascisse
Residui (Y-Y'): in ordinate



Punteggi predetti Y': in ascisse
Residui (Y-Y'): in ordinate



Autocorrelazione

Tempo o ordine di acquisizione: in ascisse
Residui (Y-Y'): in ordinate

Ovunque

CONCLUSIONE

- Minimi quadrati multipli
- Coefficienti parziali
- Verifica di ipotesi
- Assunzioni e loro verifica