Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2019/2020 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

LEZIONE 5 - Questioni di analisi e applicazione della regressione lineare
Pratica in R con creazione del database e caricamento esterno

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: lunedì 14-16 previa mail

giorno	lezione	ora	argomento	h
07-ott-19	1	16-18	Introduzione al corso, alla materia e all'ambiente R	2
09-ott-19	2	14-16	Vettori e matrici: overview e laboratorio in R	2
14-ott-19	3	16-18	Modello di regressione lineare semplice	2
16-ott-19	4	14-16	Applicazione pratica: MRLS in R	2
21-ott-19	5	16-18	Questioni analitiche e interpretazione di MRLS	2
23-ott-19	6	14-16	Modello di regressione lineare multivariata	2
28-ott-19	7	16-18	Applicazione pratica: MRLM in R	2
30-ott-19	8	14-16	Analisi per componenti principali (PCA)	2
04-nov-19	9	16-18	PCA: Applicazione pratica in R	2
06-nov-19	10	14-16	Analisi fattoriale	2
11-nov-19	11	16-18	AF: applicazione pratica in R	2
13-nov-19	12	14-16	Approfondimento: distinzione tra analisi fattoriale confermativa ed esplorativa	2
18-nov-19	13	16-18	Analisi per gruppi (CA)	2
20-nov-19	14	14-16	Cluster gerarchici e applicazione in R	2
25-nov-19	15	16-18	Cluster non gerarchici e applicazione in R	2
27-nov-19	16	14-16	Cluster gerarchici e non gerarchici: laboratorio in R	2
02-dic-19	17	16-18	Test di permutazione	2
04-dic-19	18	14-16	Analisi di dipendenza e interdipenenza: overview	2
09-dic-19	19	16-18	ESERCITAZIONI (RLS-RLM)	2
11-dic-19	20	14-16	ESERCITAZIONI (FA-PCA)	2
16-dic-19	21	16-18	ESERCITAZIONI (CA)	2

…punto della situazione…

Regressione

→ Metodo dei minimi quadrati per determinare b_0 e b_1

→ Calcolo coefficiente di determinazione (R^2)

→ Calcolo errore standard (Syx)

→ Analisi dei residui per verificare le assunzioni base (relazione tra e_i - x_i)

→ **SE** le assunzioni **sono rispettate** → ci chiediamo

LA RELAZIONE TRA X E Y
NELLA POPOLAZIONE
E' SIGNIFICATIVA?

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE

Partendo da un campione per fare inferenza abbiamo calcolato le stime.

Per verificare che

tali stime stiano “significativamente rispecchiando la realtà”,

dobbiamo esaminare **SE** nella realtà (nella popolazione)

la relazione lineare **è significativa**

IMPOSTIAMO LE IPOTESI:

- H_0 = ipotesi nulla

-H_1= ipotesi alternativa

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE

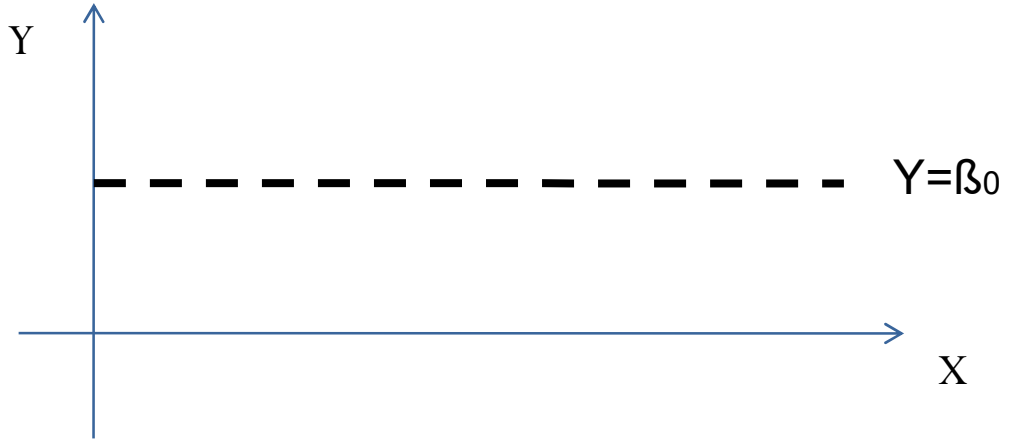
Per verificare che la relazione tra X e Y nella popolazione sia significativa, dobbiamo verificare l'ipotesi che **$ß_1=0$**

IPOTESI NULLA

il coefficiente angolare $ß_1=0$ →

H_0= al variare di x non vi è alcuna variazione di Y

Graficamente: $Yi=ß_0+0$ → retta costante parallela all'asse delle ascisse



TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE

IPOTESI ALTERNATIVA = H1:

Se $ß_1 \neq 0$ → vi è una variazione in Y al variare di X

(OVVERO ESISTE UNA RELAZIONE TRA Y E X)

PROCEDIAMO CON IL TEST DI IPOTESI

Se si rifiuta H0 vi è evidenza empirica

dell'esistenza della relazione (lineare) tra X e Y nella popolazione

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE

H0: **l'ipotesi nulla** contiene sempre un segno di eguale relativo al valore specificato del parametro della popolazione

H1: **l'ipotesi alternativa** non contiene mai un segno di eguale relativo al valore specificato del parametro della popolazione

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE

Ambito statistico di riferimento:

VERIFICA DI IPOTESI BASATA SU

TEST AD **UN CAMPIONE**

A **DUE CODE** (= e $\neq$)

UTILIZZANDO IL **METODO DEL VALORE CRITICO** (v.c.)

SCOPO:
La logica sottostante alla verifica di ipotesi è quella di stabilire la plausibilità dell'ipotesi nulla alla luce delle informazioni campionarie

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE: verifica di ipotesi

La verifica di ipotesi è una **procedura inferenziale** che ha come scopo quello di considerare l'informazione empirica (ottenuta da una statistica campionaria) e di stabilire se questa è favorevole ad una asserzione di interesse sui parametri della popolazione

La verifica di ipotesi ha inizio proprio con **una considerazione di una teoria o proposizione riguardante un particolare parametro della popolazione** e l'ipotesi che il valore del parametro della popolazione sia uguale ad un dato valore prende il nome di ipotesi nulla

L'ipotesi nulla in genere coincide con lo stato delle cose e viene indicata con il simbolo H0, quindi nell'esempio della regressione **H0: ß1= 0**

Sebbene le informazioni siamo tratte a partire dal campione, l'ipotesi è espressa con **riferimento a un parametro della popolazione**.

Se i risultati campionari non fossero favorevoli all'ipotesi nulla si dovrebbe concludere che l'ipotesi nulla sia falsa e chiaramente ci deve essere un'altra ipotesi che risulti vera. L'ipotesi **alternativa** H1 è l'asserzione opposta all'ipotesi nulla, e nell'esempio in questione **H1: ß1 ≠ 0**

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE: verifica di ipotesi e decisione finale

L'ipotesi alternativa rappresenta la conclusione a cui si giunge quando si **rifiuta l'ipotesi nulla (decisione forte),** cioè quando il campione osservato fornisce sufficiente evidenza del fatto che l'ipotesi nulla sia falsa

D'altro canto **il manca rifiuto dell'ipotesi nulla non prova che essa è vera**. Quello che si può concludere è che non vi è sufficiente evidenza empirica contraria ad essa (**decisione debole**)

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE: processo decisionale

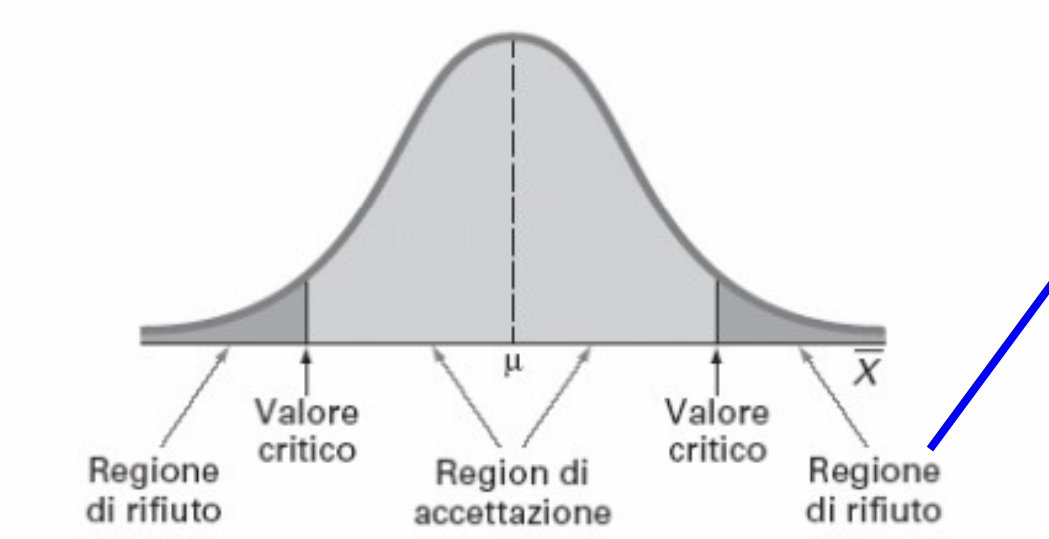
Il processo decisionale è sostenuto dal punto di vista quantitativo, **valutando la probabilità di ottenere un dato risultato campionario**, se l'ipotesi nulla fosse vera

Tale **probabilità** si ottiene determinando prima la **distribuzione campionaria della statistica di interesse** e poi **calcolando la probabilità che la statistica test assuma il valore osservato in corrispondenza del campione estratto**

La distribuzione campionaria della statistica test spesso è una distribuzione statistica nota, **come la normale o la t,** e quindi possiamo ricorrere a queste distribuzioni per decidere se rifiutare o meno a un'ipotesi nulla

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE: processo decisionale

La distribuzione campionaria della statistica test è divisa in due regioni: una regione di rifiuto (chiamata anche regione critica) e una regione di accettazione



La regione di rifiuto è data da tutti i valori della statistica test che non è probabile si verifichino quando l'ipotesi nulla è vera, mentre è probabile che questi valori si verifichino quando l'ipotesi nulla è falsa

A: se la statistica test cade nella regione di **accettazione**, l'ipotesi nulla **non può essere rifiutata**

B: se la statistica test cade nella **regione di rifiuto,** l'ipotesi nulla **deve essere rifiutata**

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE: processo decisionale

Per prendere una decisione sull'ipotesi nulla, dobbiamo in primo luogo definire le **regioni di rifiuto e di accettazione** e questo viene fatto determinando il cosiddetto **valore critico della statistica test**

Si definisce a priori il **livello α di significatività** (ovvero siamo certi ad un livello di confidenza pari a (1- α)*100 delle conclusioni tratte). Generalmente i valori assegnati ad α sono 0.01, 0.05 o 0.1.

Il coefficiente di confidenza (1- α) rappresenta la probabilità che l'ipotesi nulla non sia rifiutata quando è vera (quindi non dovrebbe essere rifiutata).

Una volta specificato il valore di α, si ottiene anche la regione di rifiuto perché è la probabilità che la **statistica test** cada nella regione di rifiuto quando l'ipotesi nulla è vera.
Il valore critico che separa la regione di accettazione da quella di rifiuto viene determinato di conseguenza

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE: processo decisionale

Elementi da utilizzare:

t-stat e Valore Critico (V.C.)

SE t-stat **>** valore critico o [–t-stat]< [-vc] → SI RIFIUTA H_0

(esiste una relazione lineare tra X e Y)

T-stat= (b1-ß1)/Sb1

Dove :
b1= coefficiente angolare campionario
ß1=valore ipotizzato nella popolazione
(=0 per H0)
Sb1= errore standard del coefficiente angolare campionario

VALORE CRITICO $t_{a/2}$

-Con α fissato a priori
-Con gradi di libertà n-2

→ Tavola E3 *di t-student*

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE: processo decisionale

T STATISTICO

$$t_{stat} = \frac{b_1 - \beta_1}{S_{b1}}$$

→ Considerato pari a 0 in conformità all'ipotesi nulla

$$S_{b1} = \frac{S_{yx} = \sqrt{\frac{\sum_{1=i}^{n}(y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{SSX} = \sqrt{\sum_{1=i}^{n}(x_i - \bar{x})^2}}$$

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE: processo decisionale

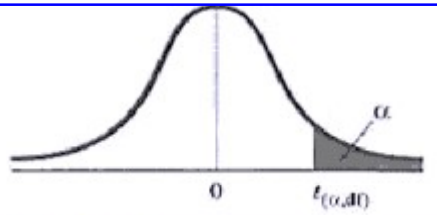
VALORE CRITICO (V.C.)

n-2 gradi di libertà

α/2

Incrocio i valori sulla Tavola T Student per trovare il V.C.

Tavola della distribuzione T di Student



Gradi di libertà	Area nella coda di destra					
	0,25	0,1	0,05	0,025	0,01	0,005
1	1,0000	3,0777	6,3138	12,7062	31,8205	63,6567
2	0,8165	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,7649	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,7407	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,7267	1,4759	2,0150	2,5706	3,3649	4,0321
6	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,7111	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,6974	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,6955	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,6938	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,6924	1,3450	1,7613	2,1448	2,6245	2,9768

EX 1- esempio

Verifichiamo che vi sia evidenza empirica per la relazione lineare tra *Mq di dimensione* e *volume delle vendite* dei punti vendita.

A priori fissiamo un livello di significatività a=0.05 (95%)

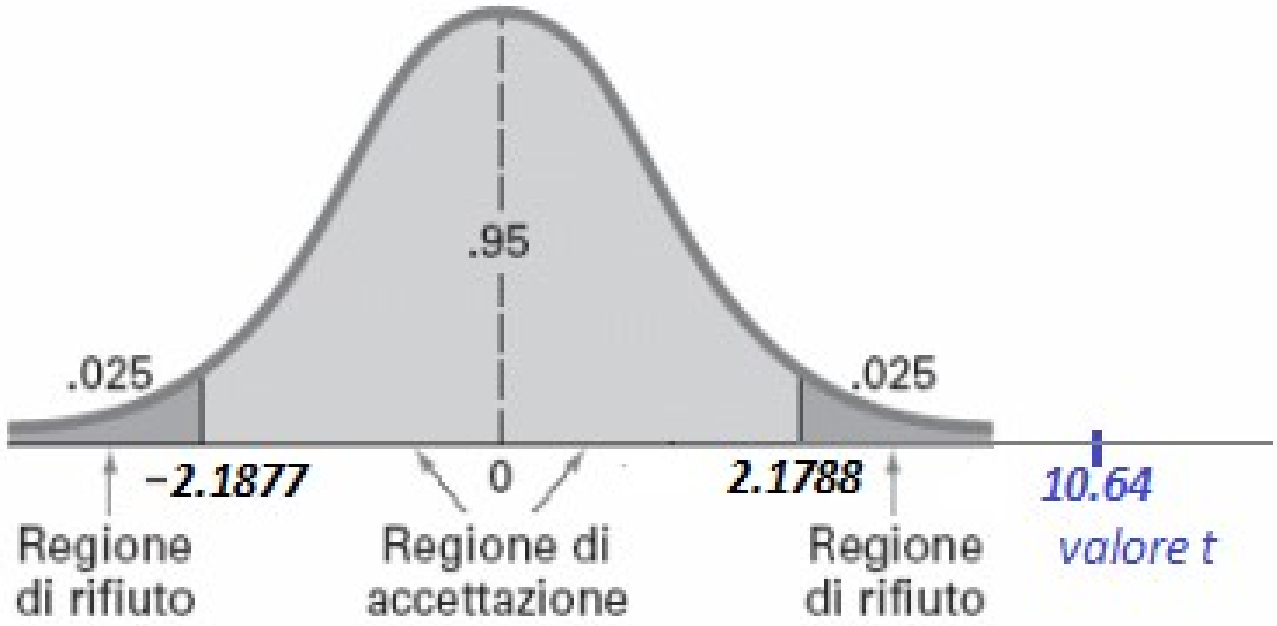
A) Valore critico:
   A) n-2= 12 gradi di libertà
   B) a=0.05 → a/2 = 0.025
   C) TAVOLA E3 → vc = 2.1788

B) t-stat:
   A) b1= 1.6699 ß1=0 (imposto per H0)
   B) Syx= 0.9664 (errore standard di regressione)
   C) Calcolo di SSX e pongo sotto radice quadrata
   D) Otteniamo Sb1= 0.9664/RQ(SSX) = 0.1569

   → t-stat = (1.6699-0)/0.1569 = 10.64

EX 1- esempio

Verifichiamo che vi sia evidenza empirica per la relazione lineare tra *Mq di dimensione* e *volume delle vendite* dei punti vendita.
A priori fissiamo un livello di significatività a=0.05 (95%)



EX 1- esempio

t-stat (10.64)> vc (2.1788)

→RIFIUTO H0

→Esiste evidenza empirica che

a livello di significatività 0.05 (confidenza al 95%)

esiste una relazione lineare

tra X e Y nella popolazione

TESTARE LA RELAZIONE TRA X E Y NELLA POPOLAZIONE:

approccio del P-value

Il p-value rappresenta la **probabilità di osservare un valore della statistica** test **uguale o più estremo del valore critico** che si calcola a partire dal campione, quando l'ipotesi H0 è vera.

Un p-value basso porta a rifiutare l'ipotesi nulla H0.

Il p-value è anche chiamato livello di significatività osservato, in quanto coincide con il più piccolo livello di significatività in corrispondenza del quale H0 è rifiutata.

In base all'approccio del p-value, la regola decisionale per rifiutare H0 è la seguente:

Se il p-value è ≥ α, l'ipotesi nulla non è rifiutata.

Se il p-value è < α, l'ipotesi nulla è rifiutata.

INFERENZA 2:
IL COEFFICIENTE DI CORRELAZIONE

La forza relativa di un legame tra due variabili quantitative è data dal

COEFFICIENTE DI CORRELAZIONE CAMPIONARIO: r

In relazione al coefficiente di correlazione nella popolazione indicata con ρ

Si utilizza per testare la correlazione statisticamente significativa nella popolazione

$$r = \frac{Cov_{xy}}{S_x S_y} \qquad -1 \leq r \leq +1$$

INFERENZA 2:
IL COEFFICIENTE DI CORRELAZIONE

Graficamente

ρ = -1 ρ =0 ρ =+1

INFERENZA 2:
IL COEFFICIENTE DI CORRELAZIONE

Per verificare che ci sia una relazione lineare statisticamente significativa tra X e Y nella popolazione dobbiamo testare le ipotesi seguenti:

$$H0 : \rho = 0$$

$$H1 : \rho \neq 0$$

Si procede con il calcolo di v.c e t-stat → confronto

Se I t-stat I> I v.c I → rifiuto H0

→ Se rifiuto H0 esiste una significativa correlazione tra x e y nella popolazione

INFERENZA 2:
IL COEFFICIENTE DI CORRELAZIONE

$$Vc: \frac{a}{2} \quad (n-2)$$

$$t - stat = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$dove\ r = +\sqrt{R^2}\ se\ b1 > 0; \quad -\sqrt{R^2}\ se\ b1 < 0$$

ESERCIZIO SU CUI GLI STUDENTI FANNO PRATICA:
APPLICAZIONE ALL'ESEMPIO "MINI MARKET"

Letture consigliate

Levin, Krenbiel, Berenson (2010) Statistica. *Quinta edizione. Cap 12*

ESERCITAZIONI
(tipologia esame)

Q. La differenza tra analisi di regressione lineare multipla e regressione lineare semplice dipende da…:

a. Il numero di variabili dipendenti
b. Il numero di variabili esplicative
c. Il numero di equazioni di regressione

Q. Quale delle seguenti non è una assunzione tipica sugli errori del modello di regressione lineare classico:

a. Indipendenza
b. Normalità
c. Eteroschedasticità

Q. I residui del modello di regressione sono:

a. Le differenze tra valori osservati e stimati della variabile dipendente
b. Le differenze tra valori osservati e stimati delle variabili esplicative
c. Le differenze tra valori osservati e media campionaria della variabile dipendente

Laboratorio in R