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

STATISTICA MULTIVARIATA
SSD MAT/06

LEZIONE 7- Regressione lineare multipla

Docente: Valentina MINI
valentina.mini@unife.it
RICEVIMENTO: su appuntamento 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: SRLS in R	2
21-ott-19	5	16-18	Questioni analitiche e interpretazione di SRLS	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

NOTA SULLA LETTURA DELL'OUTPUT DI REGRESSIONE LINEARE SEMPLICE

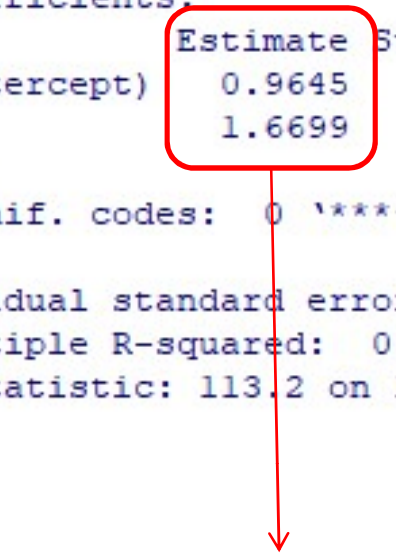
```
> summary(output.reg.lin)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.87406 -0.74834  0.08121  0.86255  1.15032

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9645     0.5262   1.833   0.0917 .
x             1.6699     0.1569  10.641 1.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9664 on 12 degrees of freedom
Multiple R-squared:  0.9042,    Adjusted R-squared:  0.8962
F-statistic: 113.2 on 1 and 12 DF,  p-value: 1.823e-07
```



$Y = 0.9645 + 1.6699 * X_i$

```
> help(anova)
starting httpd help server ... done
> anova(output.reg.lin)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 105.748 105.748  113.23 1.823e-07 ***
Residuals 12  11.207   0.934
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSR

SST = 105.748+11.207 =116.955

R^2 = SSR/SST=105.748/116.955 = 0.90417

```
> summary(output.reg.lin)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.87406 -0.74834  0.08121  0.86255  1.15032

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9645     0.5262   1.833   0.0917 .
x             1.6699     0.1569  10.641 1.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9664 on 12 degrees of freedom
Multiple R-squared:  0.9042,    Adjusted R-squared:  0.8962
F-statistic: 113.2 on 1 and 12 DF,  p-value: 1.823e-07
```

R^2 ←

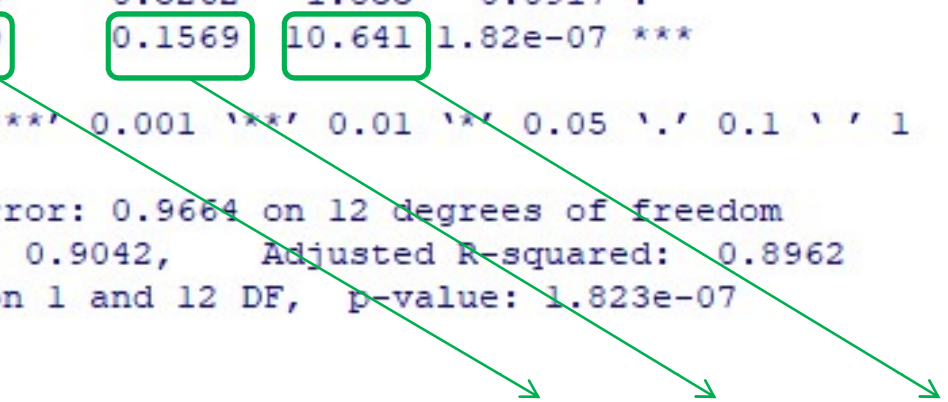
```
> summary(output.reg.lin)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.87406 -0.74834  0.08121  0.86255  1.15032

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9645     0.5262   1.833   0.0917 .
x             1.6699     0.1569  10.641 1.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9664 on 12 degrees of freedom
Multiple R-squared:  0.9042,    Adjusted R-squared:  0.8962
F-statistic: 113.2 on 1 and 12 DF,  p-value: 1.823e-07
```



T-stat = 1.6699/0.1569= 10.641

= b1 / Sb1 = T-stat

*La **regressione** formalizza e risolve il problema di una **relazione funzionale** tra variabili misurate sulla **base di dati campionari** estratti da **un'ipotetica popolazione infinita**.*

REGRESSIONE LINEARE MULTIPLA

E' la **generalizzazione del modello di regressione lineare semplice**:
per spiegare il fenomeno d'interesse Y
vengono introdotte p variabili esplicative (con p > 1).

Tale generalizzazione diventa molto più semplice utilizzando **l'algebra delle matrici.**

Il modello di regressione multipla genera però **nuovi problemi**:

- scelta delle **variabili**
- **test multipli**
- **multicollinearità**

Analisi di regressione lineare multipla

concetto: esaminiamo la relazione lineare tra
1 dipendente (Y) e 2 o più variabili indipendenti (X_i)

Modello di regressione multipla con k var. indipendenti:

Intercetta di Y

Inclinazione della popolazione

Errore casuale

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

dove:

- i varia tra le osservazioni, $i = 1, \ldots, n$;
- Y_i è il i-esimo valore della variabile dipendente
- $X_{1i}, X_{2i} + \cdots + X_{ki}$ sono le i-esime osservazioni di ciascuno dei k regressori;
- $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \cdots + \beta_k x_{ki}$ è la *retta di regressione*;
- β_0 è il valore atteso di Y quando tutte le X sono uguali a zero (cioè è l'intercetta);
- β_1 è il coefficiente angolare di X_1, β_2 è il coefficiente angolare di X_2, (tenendo costanti gli X_k non presi in considerazione), ecc.

Analisi di regressione lineare multipla: LA FORMA MATRICIALE

Siano:

- $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)'$ il vettore delle v.c. dipendenti, le cui realizzazioni campionarie saranno contenute nel vettore $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n)'$;
- $\mathbf{X}$ la matrice di dimensione $(n \times (p+1))$, contenente le osservazioni sulle variabili esplicative (regressori) e secondo la notazione usuale x_{ij} indica il valore assunto dalla variabile X_j, con $j = 1, 2, \ldots, p$, relativamente all' i-esima unità statistica, $i = 1, 2, \ldots, n$;
- $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ il vettore delle v.c. ϵ_i le cui realizzazioni (scarti) sono contenute nel vettore $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)'$;
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ il vettore dei $(p+1)$ parametri da stimare.

Analisi di regressione lineare multipla: LA FORMA MATRICIALE

Pertanto, avendo posto:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_i \\ \dots \\ Y_n \end{bmatrix} \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{i1} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_i \\ \dots \\ \beta_p \end{bmatrix} \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_i \\ \dots \\ \epsilon_n \end{bmatrix}$$

utilizzando la notazione matriciale, il modello di regressione multipla è dato da

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

ed esplicitando tale relazione per le singole unità statistiche equivale a

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i,\ i = 1, 2, \ldots, n.$$

Analisi di regressione lineare multipla: LA FORMA MATRICIALE

Sul campione osservato la relazione diventa

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

e, a livello delle singole unità statistiche, si specifica come segue

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i,\ i = 1, 2, \ldots, n.$$

Il vettore $\mathbf{e}$ contiene le realizzazioni del vettore di v.c. ϵ. Tali realizzazioni sono determinabili se conosciamo i parametri β, perchè:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$$

e, ovviamente, si esplicitano nel modo seguente:

$$e_i = y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) = y_i - y_i(\beta),\ i = 1, 2, \ldots, n.$$

Analisi di regressione lineare multipla: FORMALIZZAZIONE DEL MODELLO

I coefficienti del modello di regressione multipla sono stimati usando dati campionari

Equazione di regressione multipla con k variabili indipendenti:

Valori di Y stimati

Intercetta stimata

Coefficienti di inclinazione stimati

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

Analisi di regressione lineare multipla: LE IPOTESI DEL MODELLO

Le ipotesi del modello di regressione lineare multipla sono

1. $\mathbf{Y} = \mathbf{X}\beta + \epsilon$;
2. $\mathbb{E}(\epsilon) = \mathbf{0}$;
3. $\text{Var}(\epsilon) = \mathbb{E}(\epsilon\epsilon') = \sigma^2\mathbf{I}_n$;
4. $\mathbf{X}$ è una matrice (non stocastica) tale che $r(\mathbf{X}) = p + 1$.

Dopo aver ottenuto le stime $\hat{\beta}_j$ per i parametri β_j, il modello diventa

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip} + \hat{e}_i = \hat{y}_i + \hat{e}_i.$$

I residui $\hat{e}_i$ sono dati dalla differenza tra i valori osservati y_i e i valori stimati $\hat{y}_i$ calcolati secondo il modello di regressione.

Analisi di regressione lineare multipla: LA STIMA DEI COEFFICIENTI

Per stimare i parametri del modello di regressione multipla, senza fare ulteriori assunzioni circa la forma distributiva degli errori, si utilizza il metodo dei minimi quadrati (LS). Tale metodo consente di trovare il vettore β che minimizza la somma degli scarti al quadrato, ovvero la funzione $G(\beta)$ data da

$$G(\beta) = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta).$$

Sviluppando si ha che

$$G(\beta) = \mathbf{y}'\mathbf{y} + \beta'(\mathbf{X}'\mathbf{X})\beta - 2\beta'\mathbf{X}'\mathbf{y}$$

ed uguagliando a 0 la derivata prima di $G(\beta)$ rispetto a β si ottiene

$$0 = G'(\beta) = -2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X})\beta \Longrightarrow \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Analisi di regressione lineare multipla: LA STIMA DEI COEFFICIENTI

- Le equazioni normali dei minimi quadrati sono:

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1}y_i \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots & \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik}y_i
\end{aligned}
$$

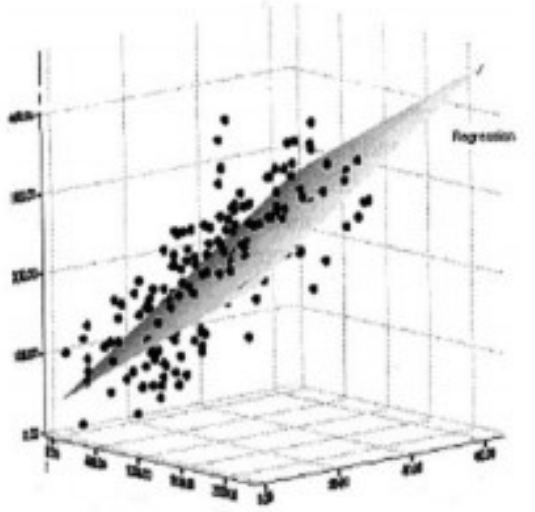
- La soluzione delle equazioni normali sono gli stimatori dei minimi quadrati dei coefficienti di regressione.

Analisi di regressione lineare multipla: RAPPRESENTAZIONE GRAFICA

Geometricamente l'equazione

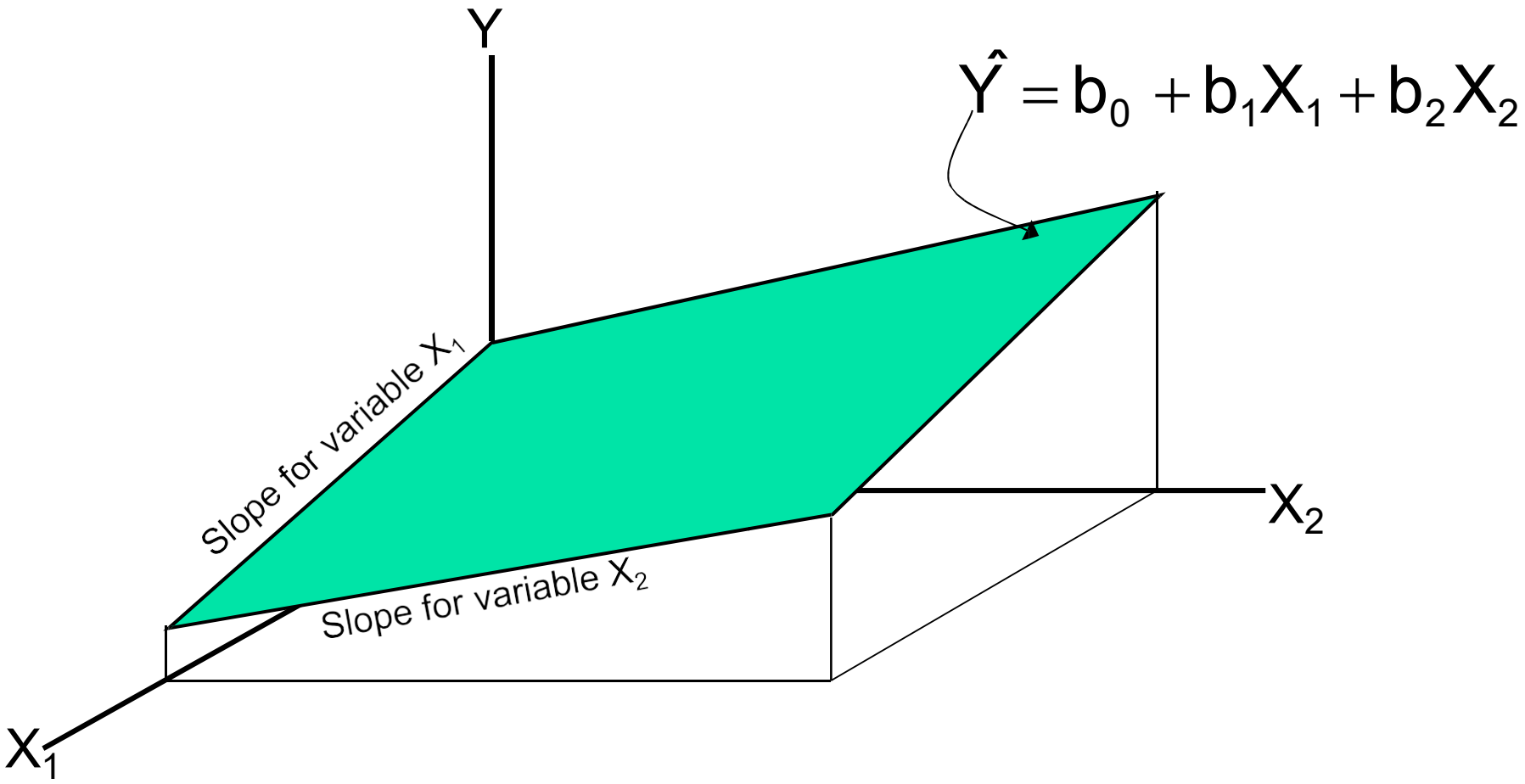
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip},\ i = 1, 2, \ldots, n,$$

definisce un iperpiano nello spazio a $p + 1$ dimensioni. Per avere un'idea del procedimento di stima dei minimi quadrati, il piano rappresentato in figura è, tra gli infiniti piani, quello che rende minima la somma dei quadrati delle lunghezze dei segmenti congiungenti i punti osservati al piano stesso.



Analisi di regressione lineare multipla: RAPPRESENTAZIONE GRAFICA

Modello a due variabili



Analisi di regressione lineare multipla: TEOREMA DI GAUSS-MARKOV

Teorema di Gauss-Markov

Sotto le ipotesi del modello di regressione lineare, gli stimatori LS $\mathbf{B}$ per i parametri β, sono lineari, non distorti,ed i più efficienti nella classe degli stimatori lineari e non distorti (BLUE).

Per applicare il metodo ML, occorre aggiungere l'ipotesi che il vettore $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. gli stimatori ML
coincidono con quelli LS prima ricavati, che sono lineari, non distorti, sufficienti ed efficienti nella classe di tutti gli stimatori non distorti.

Analisi di regressione lineare multipla: INFERENZA E TEST DA APPLICARE **su un singolo parametro**

Per ottenere la regione critica di un test o un intervallo di confidenza per i parametri del modello di regressione, è necessario ipotizzare, per n finito, che le v.c. errori siano normali e indipendenti, utilizzando quindi gli stimatori ML. Per verificare $H_0 : \hat{\beta}_i = 0$ contro l'alternativa $H_1 : \hat{\beta}_i \neq 0$ basta calcolare il rapporto

$$T = \frac{\hat{\beta}_i - 0}{s\sqrt{v^{j+1,j+1}}}, \; j = 0, 1, 2, \ldots, p.$$

Infatti la stima della varianza dello stimatore B_j per il parametro β_j è data da $es^2(B_j) = s^2 v^{j+1,j+1}$ dove $v^{j+1,j+1}$ è l'elemento di posto $(j+1, j+1)$ sulla diagonale principale della matrice $(\mathbf{X}'\mathbf{X})$. Tale rapporto, sotto H_0, si distribuisce come una v.c. t di Student con $n - p - 1$ gradi di libertà.

Analisi di regressione lineare multipla: INFERENZA E TEST DA APPLICARE

ANOVA su tutti i parametri

Consiste in un test globale su tutti i parametri del modello (eccetto β_0) e in particolare nel confronto tra la devianza del modello saturo (ovvero con tutte le variabili in considerazione) e del modello ridotto. Le ipotesi saranno:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta p = 0 \quad \text{vs.} \quad H_1 : \text{almeno un } \beta_j \neq 0$$

Se indichiamo con

- Q_1 la devianza della regressione
- Q_2 la devianza dei residui

siamo interessati a valutare la statistica

$$F = \frac{Q_1/p}{Q_2/n-p-1}$$

che sotto H_0 ha distribuzione $F(p, n-p-1)$.

Analisi di regressione lineare multipla: BONTA' DEL MODELLO (R^2)

Ricordando che $SQT = SQE + SQR$, il modello si adatterà tanto più ai dati quanto più modesta sarà la variabilità dell'errore rispetto alla variabilità totale. Si introduce pertanto l'indice di determinazione multipla R^2 dato da

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

L'indice R^2 varia in $[0, 1]$ e più si avvicina a 1 migliore è l'adattamento del modello ai dati. Tuttavia è opportuno sottolineare che il valore R^2 aumenta con l'aumentare del numero di regressori, per cui è conveniente considerare la versione corretta dell'indice R^2, data da

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}.$$

Analisi di regressione lineare multipla: SCELTA DELLE VARIABILI DA UN DATABASE AMPIO

Oltre all'indice R^2, vi è l'indice proposto da Mallows (1973)

$$C_p = \frac{(1 - R_p^2)(n - T)}{1 - R_T^2} - [n - 2(p + 1)]$$

Quando le variabili esplicative sono molte si ricorre a procedure di tipo stepwise, nelle varianti *per inclusione* e *per eliminazione*. In particolare, partendo da un modello parziale si procede per passi e di volta in volta si aggiunge una variabile che contribuisce in maniera significativa al miglioramento del modello o si elimina una variabile il cui coefficiente non significativo. Altro approccio è il best-subset, in cui si valutano tutti i possibili modelli di regressione ricavabili da un certo insieme di variabili esplicative e si individuano i sottinsiemi migliori secondo uno dei criteri sopra riportati (R^2 e C_p).

Analisi di regressione lineare multipla: ANALISI DELLA MULTICOLLINEARITA'

Si verifica quando il rango della matrice $\mathbf{X}$ non è massimo e si traduce nella presenza di un'elevata correlazione tra le variabili esplicative. Le variabili collineari non forniscono informazioni aggiuntive e risulta difficile individuare l'effetto che ciascuna di esse ha sulla variabile risposta. Una misura della multicollinearità è data dall'indice VIF (Variance Inflationary Factor). In particolare, per la j-esima variabile si ha

$$VIF_j = \frac{1}{1 - R_j^2},$$

dove R_j^2 è il coefficiente di determinazione che caratterizza il modello in cui la variabile dipendente è X_j e tutte le altre variabili esplicative sono incluse nel modello.

Analisi di regressione lineare multipla: ALTRE FORME DI REGRESSIONE MULTIPLA QUADRATICA (E POLINOMIALE)

Supponiamo ora che tra $\mathbf{Y}$ e $\mathbf{X}$ non vi sia una relazione di tipo lineare. Tra le relazioni non lineari più comuni vi è quella quadratica. Il modello di regressione quadratica è simile ad un modello di regressione multipla con due variabili esplicative in cui la seconda variabile esplicativa è il quadrato della prima. In particolare:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \epsilon_i,\ i = 1, 2, \ldots, n.$$

dove

- β_0 è l'intercetta,
- β_1 è il coefficiente che esprime l'effetto lineare su Y,
- β_2 è il coefficiente che esprime l'effetto quadratico su Y,
- ϵ_i è l'errore casuale.

Tale modello è generalizzabile ad un modello polinomiale.

Analisi di regressione lineare multipla: COME CONSIDERARE VARIABILI QUALITATIVE CREAZIONE DI DUMMIES

Nel caso di variabili esplicative discrete è opportuno ricorrere ad un modello che includa variabili indicatrici (dummy) per poter valutare l'effetto di un fenomeno che presenta modalità qualitative su una risposta. Sia E un evento che si suppone abbia un effetto nel modificare Y_i. Sia

$$D_i = \begin{cases} 1 & \text{se per l'unità } i\text{-esima } E \text{ è presente} \\ 0 & \text{altrimenti} \end{cases}$$

la variabile indicatrice (dummy). Se consideriamo il modello

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 D_i + \epsilon_i,\ i = 1, 2, \ldots, n$$

si avrà che

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{i1} + \epsilon_i, & \text{per le unità dove } E \text{ è assente} \\ (\beta_0 + \beta_2) + \beta_1 x_{i1} + \epsilon_i, & \text{per le unità dove } E \text{ è presente} \end{cases}$$

Analisi di regressione lineare multipla:
ESEMPIO PRATICO

- Un distributore di torte congelate vuole valutare i fattori che influenzano la domanda di mercato
  - Variabile dipendente: Vendite torte(unità per settimana)
  - Variabili indipendenti:
    - Prezzo (in $)
    - Pubblicità (in $100)
- Dati raccolti per 15 settimane

Analisi di regressione lineare multipla: ESEMPIO
IL DATABASE DI RIFERIMENTO

		Y	X1	X2
Settimana	Unità vendute	**Prezzo unitario ($)**	**Pubblicità ($100s)**
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Equazione di regressione multipla che stiamo cercando è:

$$\widehat{\text{Vendite}} = b_0 + b_1\,(\text{Prezzo}) + b_2\,(\text{Pubblicità})$$

Analisi di regressione lineare multipla: ESEMPIO

Regression Statistics	
Multiple R	**0.72213**
R Square	**0.52148**
Adjusted R Square	**0.44172**
Standard Error	**47.46341**
Observations	**15**

b_1 = -24.975: le vendite Diminuiscono in media di 24.975 torte per settimana per ogni aumento di prezzo di $1, al netto di cambiamenti nella pubblicità

b_2 = 74.131: le vendite aumentano in media di 74.131 torte a settimana per ogni aumento di $100 di pubblicità, al netto di effetto di modifiche dovute al Prezzo

$$\text{Vendite} = 306.526 - 24.975(\text{Prezzo}) + 74.131(\text{Pubblicità})$$

ANOVA	***df***	***SS***	***MS***	***F***	***Significance F***
Regression	**2**	**29460.027**	**14730.013**	**6.53861**	**0.01201**
Residual	**12**	**27033.306**	**2252.776**		
Total	**14**	**56493.333**			

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***
Intercept	**306.52619**	**114.25389**	**2.68285**	**0.01993**	**57.58835**	**555.46404**
Prezzo	**-24.97509**	**10.83213**	**-2.30565**	**0.03979**	**-48.57626**	**-1.37392**
Pubblicità	**74.13096**	**25.96732**	**2.85478**	**0.01449**	**17.55303**	**130.70888**

Analisi di regressione lineare multipla: ESEMPIO PREDIRE Y SULLA BASE DI VALORI DATI DI Xi

Prediciamo le vendite settimanali con un prezzo di vendita $5.50 e pubblicità $350:

$$\widehat{\text{Vendite}} = 306.526 - 24.975(\text{Prezzo}) + 74.131(\text{Pubblicità})$$

$$= 306.526 - 24.975\,(5.50) + 74.131(3.5)$$

$$= 428.62$$

Vendite stimate 428.62 torte

Pubblicità è in $100's, quindi $350 significa che $X_2 = 3.5$

Analisi di regressione lineare multipla: ESEMPIO
LA BONTA' DELLA REGRESSIONE

- riportiamo la proporzione della variazione totale in Y spiegata da tutte le variabili X insieme

$$R^2 = \frac{SSR}{SST} = \frac{\text{somma ^2 scarti regr.}}{\text{somma ^2 scarti tot.}}$$

Analisi di regressione lineare multipla: ESEMPIO

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$R^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% della variazione nelle vendite di torte è spiegata dalla variazione in P e pubblicità

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Analisi di regressione lineare multipla: ESEMPIO
R^2 AGGIUSTATO

- Mostra la proporzione di variazione in Y spiegata da tutte le variabili X aggiustata per il **numero di variabili** X usate e per la **dimensione del campione**

$$R^2_{adj} = 1 - \left[(1 - r^2) \left(\frac{n-1}{n-k-1} \right) \right]$$

(dove n = dim. campione, k = numero var. indipendenti)

- Si penalizza l'utilizzo di variabili indipendenti non necessarie
- E' più piccolo di R^2
- Utile per paragonare modelli diversi

Analisi di regressione lineare multipla: ESEMPIO LA SIGNIFICATIVITA' GENERALE DEL MODELLO

- F Test per la significatività generale del modello
- Mostra se c'è una relazione lineare tra tutte le variabili X considerate insieme e la Y
- Si utilizza il test statistico F
- Si impostano le ipotesi seguenti:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (no relazione lineare)

H_1: at least one $\beta_i \neq 0$ (almeno una variabile indipendente influenzaY)

Analisi di regressione lineare multipla: ESEMPIO

- Test statistico:

$$F_{STAT} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

Dove:

-il numeratore ha g.l. = k

- il denominatore ha g.l. = (n – k - 1)

Analisi di regressione lineare multipla: ESEMPIO

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

Con 2 e 12 gradi di libertà

P-value del F Test

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Analisi di regressione lineare multipla: ESEMPIO

Errori (residui) del modello di regressione:

$$e_i = (Y_i - \hat{Y}_i)$$

Assunzioni:

- Indipendenza degli errori
  - Valori di errore statisticamente indipendenti
- Distribuzione normale degli errori
  - I valori di errore sono normalmente distribuiti rispetto ai valori di X
- Varianza costante (omoschedasticità)
  - La distribuzione di variabilità degli errori ha varianza costante

Analisi di regressione lineare multipla: ESEMPIO ANALISI GRAFICA DEGLI ERRORI

- Grafici a dispersione usati in regressione multipla
  - Residui vs. $\hat{Y}_i$
  - Residui vs. X_{1i}
  - Residui vs. X_{2i}
  - Residuals vs. tempo (se dati temporali)

Usiamo il grafico sui residui per controllare l'eventuale violazione delle assunzioni di regressione

Analisi di regressione lineare multipla: ESEMPIO TESTARE LA RELAZIONE TRA OGNI X E LA Y

- Usiamo il test t per l'inclinazione delle variabili individuali
- Mostra se c'è relazione lineare tra la variabile X_j e Y mantenendo costanti gli effetti delle altre variabili X
- Ipotesi:
  - H_0: $\beta_j = 0$ (no relazione lineare)
  - H_1: $\beta_j \neq 0$ (relazione lineare tra X_j eY)

Analisi di regressione lineare multipla: ESEMPIO TESTARE LA RELAZIONE TRA OGNI X E LA Y

(continua)

H_0: $\beta_j = 0$ (no relazione lineare)

H_1: $\beta_j \neq 0$ (relazione lineare tra X_j eY)

Test statistico:

$$t_{STAT} = \frac{b_j - 0}{S_{b_j}}$$

(g.l. = n – k – 1)

Analisi di regressione lineare multipla: ESEMPIO TESTARE LA RELAZIONE TRA OGNI X E LA Y

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

t Stat for Price is t_{STAT} = -2.306, with p-value .0398

t Stat for Advertising is t_{STAT} = 2.855, with p-value .0145

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Analisi di regressione lineare multipla: ESEMPIO

Multicollinearità (o collinearità) avviene quando due o più variabili esplicative del modello di regressione multipla sono altamente correlate (xi e xj)

in presenza di collinearità le stime dei coefficienti possono cambiare con alta variabilità in conseguenza di piccoli cambiamenti nei dati (low efficiency o bassa efficienza).

Perfetta multicollinearità⇒ ***matrice X*** *è singola e non può essere invertita* ⇒ le stime dei minimi quadrati non possono essere calcolate → si esclude la REG.LIN.MULT.

Analisi di regressione lineare multipla : ESEMPIO

Un indicatore di multicollinearità spesso utilizzato nella pratica è il *variance inflation factor* (**fattore di inflazione della varianza**), o VIF.
Il VIF è calcolato per ciascuna variabile del modello (spesso automaticamente da diversi software statistici), in base all'espressione:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2},$$

R_j^2: coefficiente di determinazione di regressione di X_j su tutte le altre variabili esplicative

Un VIF maggiore o uguale a 5 indica un problema di multicollinearità

In presenza di multicollinearità una o più variabili esplicative devono essere **rimosse dal modello**

Esempio: VIF(Prezzo)=VIF(Pubblicità)= $1/(1-R_1^2)= 1/(1-0.0009264^2) \cong 1$
pezzo e pubblicità sono almeno non-correlate⇒ assenza di collinearità

Analisi di regressione lineare multipla: I PASSAGGI PRATICI

Procedura di analisi di regressione

- Specificazione del modello di regressione multipla
- Test di significatività del modello
- Test di significatività dei coefficienti
- Discutere R^2 aggiustato
- Utilizzare il grafico a dispersione degli errori per verificare le assunzioni del modello

LBORATORIO R

Esercizi in R

Problema 1 – Dtabase Torta

- Effettuare una analisi di regressione per predire le vendite in funzione del prezzo e della spesa in pubblicità

- Predire il valore di VENDITE
  quando prezzo =5.5 pubblicità =4.2

- Predire il valore di VENDITE quando
  il prezzo =5.2 e pubblicità=4.7

INTRODUZIONE AI DATABASE

tastes and habits linked to the consumption of wine

Passito

A marketing survey on the demand of the wine «Passito» has been performed.

A sample of n=386 people has been interviewed. The questionnaire includes several questions about their preferences and behaviors related to drinking wine.

Dataset variables:

Label	Description	Coding
ID	Personal ID of the interviewed	Increasing integer number
AgeClass	Age of the person	Age (years)
AGE_CLASS	Age class of the person	1-6
SEX	Sex of the person	M or F
PROV	Province where the interviewed lives	Province code
LIKE_WINE	How much do you like drinking wine?	Integrer number from 1 to 7
FREQ_HOME	How often do you drink wine at home with meals?	Integrer number from 1 to 5
FREQ_BAR	How often do you drink wine in bars/pubs?	Integrer number from 1 to 5
FREQ_REST	How often do you drink wine at restaurants with meals?	Integrer number from 1 to 5
KNOW_PAS	Do you know the wine Passito?	Integrer number from 1 to 7
FREQ_PAS	How often do you drink Passito?	Integrer number from 1 to 5
FREQ_P_HOL	How often do you drink Passito on holidays and celebrations?	Integrer number from 1 to 5
FREQ_P_ALO	How often do you drink Passito when you are alone?	Integrer number from 1 to 5
FREQ_P_MEA	How often do you drink Passito at the end of meals?	Integrer number from 1 to 5
FREQ_P_OFF	How often do you drink Passito offered by someone?	Integrer number from 1 to 5
HOW_MUCH	How much wine do you drink in one year?	Integrer number from 1 to 4
LIKE_PAS	How much do you like drinking Passito?	Integrer number from 1 to 7
LIKE_AROMA	How much do you like aroma and smell of Passito?	Integrer number from 1 to 7
LIKE_SWEET	How much do you like the sweetness of Passito?	Integrer number from 1 to 7
LIKE_ALCOHOL	How much do you like the alcohol content of Passito?	Integrer number from 1 to 7
LIKE_TASTE	How much do you like the intensity of taste of Passito?	Integrer number from 1 to 7
PRICE	How much could you pay for one bottle of Passito? (0.5 litre)	Integrer number from 1 to 5

Esercizi in R

Problema 2 – Dtabase Passito

- Effettuare una analisi di regressione per predire LIKE_PAS come funzione di LIKE_AROMA, LIKE_SWEET, LIKE_ALCOHOL e LIKE_TASTE
- Predire il valore di LIKE_PAS

  quando

  LIKE_AROMA=LIKE_ALCOHOL=5

  LIKE_TASTE=LIKE_SWEET=6

INTRODUZIONE AI DATABASE

Hotel

A customer satisfaction survey where four hotels have been evaluated by 40 customers (10 for each hotel) with respect to k=3 variables: cleanliness, courtesy and price.

The data consist of rates from 0 (minimum satisfaction) to 100 (maximum satisfaction).

Dataset variables:

Name	***Type***
Hotel	Categorical
Cleanliness	Numeric
Courtesy	Numeric
Price	Numeric

Esercizi in R

Problema 3 – Database Hotel

- Effettuare un'analisi di regressione multipla per predire *Price* come funzione di *Cleanliness e Courtesy*
- Predire il valore di *Price* quando *Cleanliness*=80 e *Courtesy*=40

INTRODUZIONE AI DATABASE

Mall

A customer satisfaction survey about a a recently opened shopping center.

A sample of n=29 customers was asked to evaluate k=5 different aspects of the shopping center, such as the environmental temperature, the brightness, the presence of sales assistents, the range of products, the background music volume.

Evaluations are expressed on a scale from -100 («too little») to +100 («too much»), where 0 corresponds to «just right».

Dataset variables:

Temp_Level
Brightness
Salesman
Product_assortmant
Music_volume

Esercizi in R

Problema 4 – Database Mall

- Effettuare un'analisi di regressione multipla per predire *Product_assortment in* funzione di *Temp_Level*, *Brightness*, *Salesman* e *Music_volume*
- Predire il valore di *Product_assortment* quando *Temp_Level*=-50, *Brightness=20*, *Salesman=30 e Music_volume=-70*

INTRODUZIONE AI DATABASE

Students

Let us consider an example of teaching evaluation of k=3 university programs (undergraduate degree in Economics) evaluated by n=20 students with a rate from 0 to 100.

Dataset variables:

Statistics
Mathematics
Econometrics

Esercizio in R

Problema 5 – database studenti

- Effettuare un'analisi di regressione multipla per prevedere *Econometrics* come funzione di *Statistics* e *Mathematics*
- Predire il valore di *Econometrics* quando *Statistics=8* e *Mathematics=7*