Capitolo 13
Il modello di regressione lineare

La fase più operativa della statistica è diretta alla costruzione di modelli e cioè di rappresentazioni semplificate, analogiche e necessarie della realtà attraverso le quali provare a descrivere, prevedere, simulare e controllare fenomeni naturali, economici, sociali, ecc..

Per tali finalità, assume un ruolo centrale nell'analisi statistica la struttura logico-formale del modello di regressione mediante il quale si esplicita il **nesso funzionale** tra ciò che si intende spiegare (c.d. *variabile dipendente*) e ciò che può esserne la causa (solitamente, definito come *regressore* o *variabile indipendente*). Attraverso un modello di regressione, possiamo quindi:

1. riassumere il legame intercorrente tra le variabili osservate (due, nel caso della regressione lineare semplice e maggiori di due nel caso di regressione multipla) attraverso un'unica formula compatta;
2. effettuare e/o valutare previsioni;
3. verificare una legge scientifica descritta in termini di funzioni.

Il modello di regressione lineare è uno strumento di analisi particolarmente versatile e che, per questa ragione, può essere indicato come uno dei modelli maggiormente impiegato. La relazione funzionale può essere espressa come

$$Y = f(X) + e$$

Definiamo e come una variabile casuale che compendia l'insieme degli effetti e delle circostanze che impediscono alla relazione appena formula di essere un legame teorico di tipo matematico: la componente erratica sintetizza, quindi, la nostra ignoranza rispetto al fenomeno che intendiamo studiare. L'aggiunta della variabile casuale *e* qualifica il modello nel senso che, se f(X) è ancora una funzione matematica, allora sarà Y ad essere una variabile casuale risultante dalla somma di una componente deterministica e di una stocastica. L'aver posto $e = Y - f(X)$ come errore equivale a definire il legame funzionale $Y = f(X)$ come una relazione media tra X ed Y: se *e* vale in media zero (e cioè se $E(e) = 0$ [1] - cosa che può dirsi ragionevole trattandosi di uno *scarto*), avremo allora che

$$E(y) = E(f(X) + e) = f(X) + E(e) = f(X)$$

Per specificare il modello, occorre ora precisare la forma della funzione $f(\cdot)$ ed imporre delle relazioni sulla natura e sulle principali caratteristiche delle variabili casuali e_i. Prima di tutto, limitiamo la funzione f(X) assumendo la linearità della relazione, e cioè scrivendo

$$f(X) = \beta_0 + \beta_1 h(X)$$

[1] Nella teoria del calcolo delle probabilità, il valore atteso (aspettazione, attesa, media o speranza matematica) viene usualmente indicata con la notazione $E(\cdot)$. Il valore atteso di una variabile casuale è un numero che formalizza l'idea euristica di *valore medio* di un fenomeno aleatorio. Nel lanciodi una moneta, ad esempio, sappiamo di vincere 100 se scommettiamo su Testa ed tale evento si verifica e 0 nel caso inverso. Il valore atteso del gioco, quetso caso, è pari a 50, ovvero la media delle vincite e perdite pesata in base alle probabilità (50% per entrambi i casi): 100*0,5 + 0*0,5= 50, cioè il valore di "testa" per la sua probabilità e il valore di "croce" per la sua probabilità.

dove $h(X)$ è una funzione matematica *nota* della sola X non contente ulteriori parametri. La forma più semplice nello scegliere, $h(X)=X$ per cui il modello di regressione lineare diventa

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad i = 1,2,...,N$$

Essa esprime la variabile dipendente Y come una quota proporzionale alla variabile indipendente X (cioè la quantità $\beta_0 + \beta_1 X_i$) cui si aggiunge la componente casuale. In termini geometrici questo equivale a definire ogni punto come la risultante di una parte teorica e di una parte stocastica.

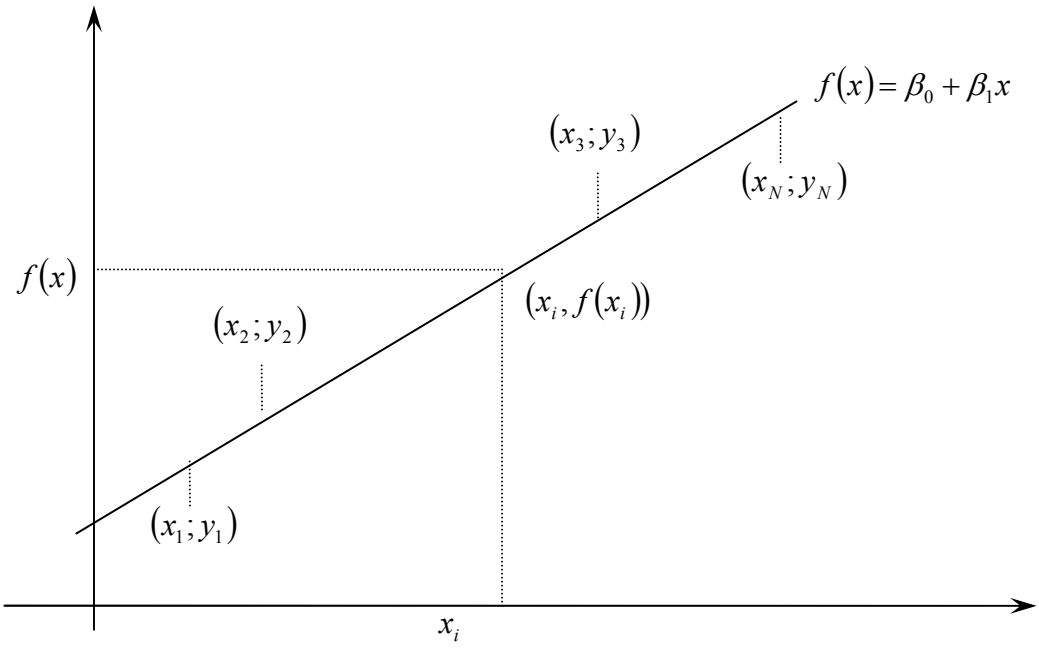


Figura 1 - Retta di regressione ed interpolazione dei dati.

Obiettivo dell'analisi di regressione è quello di individuare la retta "ottimale", intendendo per tale quella che è in grado di minimizzare la distanza tra punti teorici e punti osservati, distanza che viene interpretata come realizzazione $\hat{e}_i$ della variabile casuale e_i. Ogni variabile casuale Y_i possiede, così, una distribuzione di probabilità attorno al valore medio $E(Y_i) = \beta_0 + \beta_1 x$, che rappresenta il valore sulla retta corrispondente ad X=x, e la cui deviazione "casuale" è sintetizzata nelle variabili casuali e_i, per i=1, 2, 3, …, N.

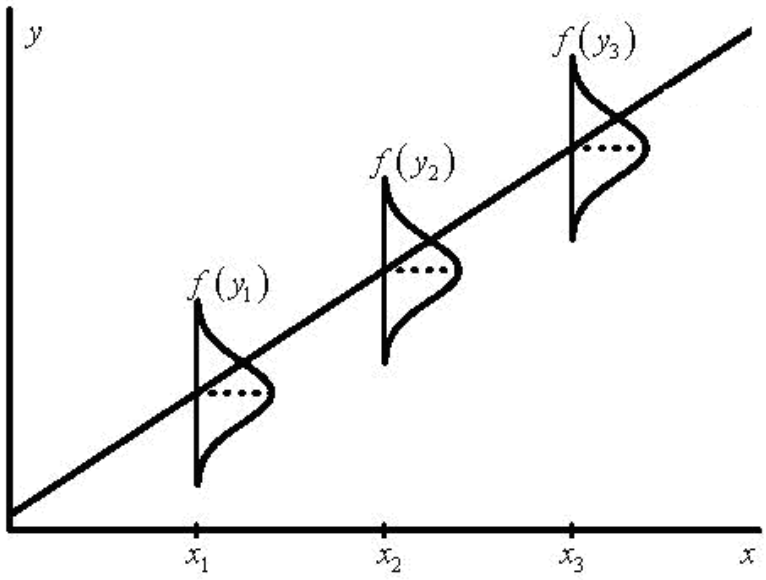


Figura 2 - Retta di regressione e funzioni di densità della Y in funzione di delta X: le distribuzioni delle Yi sono centrate sulla retta di regressione ed hanno tutte la stessa variabilità.

Che cosa significa questa immagine? Le variabili casuali nel modello di regressione lineare sono v.c. Normali doppie[2], funzioni di cinque parametri,

[2] Una variabile casuale si dice doppia quando si riferisce alla probabilità del contemporaneo verificarsi di due eventi (assunzione di zuccheri e glicemia, reddito e consumo, ecc.).

$\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho$, definite per $-\infty < x < +\infty$ e $-\infty < y < +\infty$ e la cui funzione di densità è

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x} \right) + -2\rho \left(\frac{x-\mu_x}{\sigma_x} \right) \left(\frac{y-\mu_x}{\sigma_y} \right) + \left(\frac{y-\mu_y}{\sigma_y} \right)^2 \right] \right\}$$

La Normale doppia possiede una serie di interessantissime proprietà tra le quali risulta importante, ai nostri fini, ricordare la seguente:

→ le v.c. condizionate di una Normale doppia sono le stesse Normali e precisamente

$$(X|Y=y) \text{ è } N\left[\mu_x + (\rho\sigma_x/\sigma_y)(y-\mu_y); \sigma_x^2(1-\rho^2)\right]$$

e, analogamente,

$$(Y|X=x) \text{ è } N\left[\mu_y + (\rho\sigma_y/\sigma_x)(x-\mu_x); \sigma_y^2(1-\rho^2)\right].$$

Avendo, quindi, posto Y come variabile dipendente del nostro modello di regressione, la retta disegnata nella figura 2 è appunto data da $N\left[\mu_y + (\rho\sigma_y/\sigma_x)(x-\mu_x); \sigma_y^2(1-\rho^2)\right]$: ogni punto stimato dal modello (e, quindi, giacente dalla retta) si discosta da quello osservato con una probabilità Normale.

Fatte queste indispensabili premesse, è ora necessario fare delle assunzioni e cioè delle ipotesi senza le quali non è possibile applicare il modello ai dati.

1. $Y_i = \beta_0 + \beta_1 X_i + e_i$: tale ipotesi enuncia il modello da stimare e la sua validità per ogni osservazione del campione;

2. $E(e_i) = 0$: è necessario che gli scarti positivi e negativi si compensino in media (la necessità matematica si giustifica anche in funzione della logica che vi è alla base e cioè del fatto che la componente erratica venga considerata accidentale nel modello). Da ciò discende che, in media, le variabili casuali e_i non esercitano alcuna influenza su Y_i.
3. $Var(e_i) = \sigma^2$: la varianza della componente erratica rimane costante al variare delle osservazioni.
4. $Cov(e_i; e_j) = 0$: le variabili casuali e_i sono tra di loro non correlate.
5. X è una variabile deterministica: la quinta ipotesi (detta anche ipotesi forte) considera i dati relativi alla variabile indipendente X come non soggetti a deviazioni di natura accidentale.

13.1. La stima dei parametri: minimi quadrati e massima verosimiglianza.

Dato il carattere prevalentemente applicativo del modello di regressione, tutto il capitolo verrà corredato di esempi pratici nella speranza che questi possano aiutare il lettore a comprendere la sostanza teorica e la potenza euristica di uno dei strumenti più versatili che la Statistica metta a disposizione della ricerca in moltissimi ambiti disciplinari.

Illustriamo, quindi, il modello facendo ricorso ad un esempio pratico: si studi la relazione intercorrente tra la variabile X (volume delle vendite di un punto vendita) e la variabile Y (prezzi praticati) a partire dai dati contenuti nella seguente tabella. In altri termini, si vuole studiare l'*influenza* che i prezzi (ossia, la **variabile indipendente**) esercitano sul volume delle vendite (c.d. **variabile dipendente**).

Supponiamo di avere i seguenti dati

STATO	VOLUME VENDITE	PREZZI
A	410	550
B	380	600
C	350	650
D	400	600
E	440	500
F	380	650
G	450	450
H	420	500

Tabella 1 - Analisi di marketing per il lancio di un nuovo prodotto.

Poiché lo scopo di questo lavoro è di introdurre il lettore nel vivo della ricerca operativa, ripercorriamo tutte le fasi dell'analisi che vengono fatte nelle realtà e che è bene imparare immediatamente. Prima di procedere alla stima dei parametri della retta di regressione, è buona norma, infatti, vedere graficamente la distribuzione dei dati, che nel nostro caso si presente nel seguente modo:

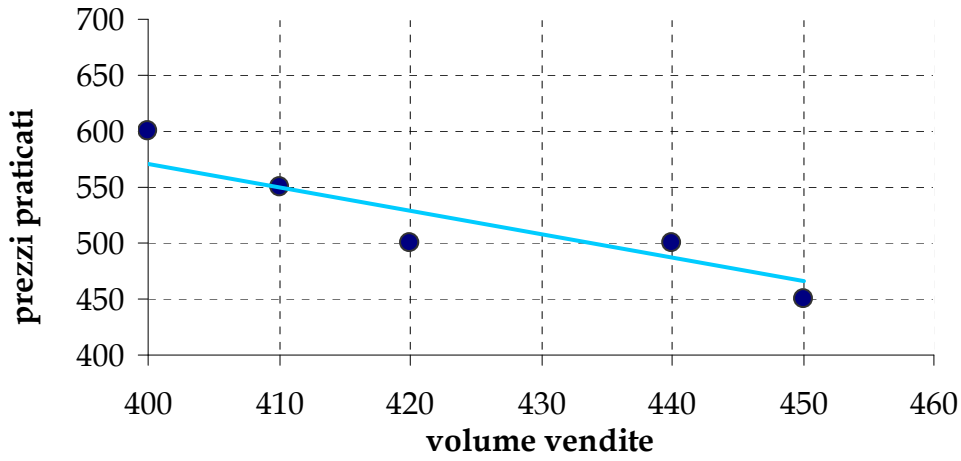


Figura 3 - Distribuzione dei dati contenuti nella tabella 13.1. "Analisi di marketing per il lancio di un nuovo prodotto in diversi punti vendita".

Uno degli assunti alla base del modello di regressione è la linearità del legame tra la variabile dipendente ed il regressore: attraverso l'analisi grafica della distribuzione dei dati possiamo immediatamente confermare o confutare tale ipotesi.

Quello proposto è chiaramente un caso molto semplice, per il quale si intuisce con un semplice "colpo d'occhio" la linearità dei dati, ma non sempre (anzi quali mai!) la realtà si presenta in modo altrettanto semplice. E' bene però partire da un caso meno complesso per comprendere la logica sottesa al modello.

Dopo l'analisi grafica dei dati, passiamo ora alla stima dei parametri del modello di regressione. I metodi a disposizione sono due:

- Minimi quadrati.
- Massima verosimiglianza.

Iniziamo con il metodo dei minimi quadrati. Per comprendere la filosofia alla basse di tale metodo, consideriamo il caso in cui si abbia una popolazione X con funzione di densità pari a $f(x;\vartheta)$, con ϑ incognito (e, quindi, da stimare) e $E(X)=g(\vartheta)$. Se $X_1,...,X_n$ è un campione casuale estratto da X, risulta evidentemente che $E(X_i)=g(\vartheta)$, i=1,…,n: ciò significa che, in media le osservazioni campionarie individuano $g(\vartheta)$, mentre $e_i = X_i - g(\vartheta), \quad i=1,...,n$ rappresentano gli scarti casuali dalla media che si riscontrano nelle osservazioni. Alla luce di quanto sinora detto, possiamo, quindi, definire il metodo dei minimi quadrati nel seguente modo: lo stimatore $\hat{\vartheta}$ ottenuto sarà tale da minimizzare

$$S(\vartheta)=\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}[X_i - g(\vartheta)]^2$$

La formulazione dicotomica appena presentata può essere, facilmente, generalizzata al caso di più variabili supponendo di avere *n* distinte popolazioni (o variabili casuali) Y_i, ciascuna delle quali sia funzione di tutti o di alcuni fra gli H parametri noti $a_1,...,a_H$ e di tutti i *p* parametri incogniti da stimare $\vartheta_1,...,\vartheta_p$. Supponendo, inoltre, che la *i-ma* popolazione abbia valor medio $g(a_1,...,a_H;\vartheta_1,...,\vartheta_p)$ e che da ciascuna popolazione venga estratto un campione casuale di *una sola unità*, allora quello estratto dalla *i-ma* popolazione sarà

$$E(Y_i)=g_i(a_1,...,a_H;\vartheta_1,...,\vartheta_p)$$

da cui gli scarti dalla media risulteranno pari a

$$e_i = Y_i - g_i\left(a_1,...,a_H;\vartheta_1,...,\vartheta_p\right), \quad i = 1,...,n$$

Sulla scorta di quanto detto e generalizzando l'espressione $S(\vartheta)=\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left[X_i - g(\vartheta)\right]^2$, possiamo dire che lo stimatore determinato con il metodo dei minimi quadrati sarà dato da quella *p-pla* $\vartheta_1,...,\vartheta_p$ che minimizza la seguente espressione

$$S(\vartheta_1,...,\vartheta_n)=\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left[Y_i - g_i\left(a_1,...,a_H;\vartheta_1,...,\vartheta_p\right)\right]^2$$

I motivi per cui questo metodo viene così frequentemente impiegato sono di diversa natura. Anzitutto, non richiede alcun tipo di conoscenza della distribuzione della popolazione (o variabile casuale), cosa che rende questa tecnica applicabile a numerose situazioni di rilevanza pratica. Per poterla utilizzare è, infatti, sufficiente conoscere solo la forma funzionale delle $g_i\left(a_1,...,a_H;\vartheta_1,...,\vartheta_p\right)$ per i=1, 2, …, n. Con gli attuali calcolatori elettronici, inoltre, la determinazione degli stimatori OLS diviene ancora più semplice e veloce.

Ora, supponendo di avere una variabile casuale, X, con funzione di densità $f(x;\vartheta)$, $E(x)=\vartheta$ e supponendo inoltre di avere un campione estratto da X del tipo $X_1,...,X_n$ la stima ai minimi quadrati di ϑ si ottiene minimizzando

$$S(\vartheta)=\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left[X_i - g(\vartheta)\right]^2$$

e, cioè, derivando tale espressione rispetto al parametro incognito della popolazione e trovando quel $\hat{\vartheta}$ che la rende nulla e verificando che la derivata seconda in $\hat{\vartheta}$ sia positiva. In tal modo si avrà

$$\frac{dS(\vartheta)}{d\vartheta} = -2\sum\left(X_i - \vartheta\right) = 0$$

che si annulla per $\hat{\vartheta} = \sum X_i / n = \overline{X}$, mentre la derivata seconda è maggiore di zero. In altri termini, quindi, lo stimatore LS di ϑ è $\hat{\vartheta} = \overline{X}$.

Da un punto di vista strettamente matematico, perché $\hat{\vartheta}_1,...,\hat{\vartheta}_p$ è lo stimatore LS, esso deve necessariamente soddisfare il seguente sistema di p equazioni in p incognite

$$\sum\left[Y_i - g_i\left(a_1, ..., a_H; \vartheta_1, ..., \vartheta_p\right)\right]$$
$$\cdot\frac{\partial}{\partial\vartheta_j} g_i\left(a_1, ..., a_H; \vartheta_1, ..., \vartheta_p\right) = 0, \quad j = 1, 2, ..., p$$

dette equazioni normali.

Del metodo dei minimi quadrati esistono due casi particolari: il primo, si ha nel caso in cui, relativamente alla componente erratica, si supponga che

$$Var\left(e_i\right) = \sigma_e^2 \qquad i = 1, ..., n$$
$$Cov\left(e_i; e_j\right) = 0 \quad i \neq j$$

che oltre $E\left(e_i\right) = 0 \quad i = 1, 2, ..., n$.

Al ricorrere di tali caratteristiche, si parla di modello di regressione.

Si parlerà, invece, di modello di regressione *lineare* nel caso in cui si aggiunga alle precedenti condizioni anche quella in base alla quale le $g_i(a_1,...,a_H;\vartheta_1,...,\vartheta_p)$ possano essere tutte espresse sotto forma di combinazione lineare dei parametri incogniti e cioè

$$g_i(a_1,...,a_H;\vartheta_1,...,\vartheta_p)=\sum_{j=1}^{p} h_i(a_1,...,a_H)\vartheta_j$$

dove le h_i sono funzioni note.

Possiamo ora stimare i parametri della retta di regressione con il metodo dei minimi quadrati. Per individuare la retta che attraversa la nube delle osservazioni minimizzando la componente erratica, è infatti necessario stimare i parametri della retta di regressione e cioè l'intercetta, β_0, ed il coefficiente angolare, β_1, a partire da un campione di N osservazioni $(x_i;y_i)$ delle variabili $(X;Y)$ sotto le ipotesi del modello di regressione lineare.

Applichiamo, quindi, la $\sum_{i=1}^{n}\left[Y_i-g_i(a_1,...,a_H;\vartheta_1,...,\vartheta_p)\right]^2$ ed otteniamo

$$S=S(\beta_0,\beta_1)=\sum(y_i-\beta_0-\beta_1x_i)^2=$$
$$\text{minimo rispetto a } (\beta_0,\beta_1)$$

Le cui soluzioni si ricavano da

$$\begin{cases}\dfrac{\partial S}{\partial\beta_0}=-2\sum(y_i-\beta_0-\beta_1x_i)=0\\[2ex]\dfrac{\partial S}{\partial\beta_1}=-2\sum(y_i-\beta_0-\beta_1x_i)x_i=0\end{cases}$$

che, semplificando opportunamente, si riduce al sistema di due equazioni lineari nelle due incognite β_0 e β_1:

$$\begin{cases} \beta_0 N + \beta_1 \sum x_i = \sum y_i \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

Tali relazioni sono equazioni normali e, quindi, la loro soluzione è

$$\begin{cases} \hat{\beta}_0 = \dfrac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N \sum x_i^2 - \left(\sum x_i\right)^2} \\ \hat{\beta}_1 = \dfrac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - \left(\sum x_i\right)^2} \end{cases}$$

Applichiamo le formule ottenute ai dati presentati all'inizio del paragrafo. Per aiutare il lettore, lo svolgimento dell'esercizio verrà proposto in fasi, su ciascuna delle quali bene che il lettore rifletta attentamente.

FASE 1: PREDISPORRE UN FOGLIO DI CALCOLO O COSTRUIRE UNA TABELLA NEL SEGUENTE MODO

i	**vendite (yi)**	**prezzi (xi)**	**y^2**	**x^2**	**x*y**
A	410	550	168100	302500	225500
B	380	600	144400	360000	228000
C	350	650	122500	422500	227500
D	400	600	160000	360000	240000
E	440	500	193600	250000	220000
F	380	650	144400	422500	247000
G	450	450	202500	202500	202500
H	420	500	176400	250000	210000
TOT	**3230**	**4500**	**1311900**	**2570000**	**1800500**

FASE 2: CALCOLO DEI VALORI DA INSERIRE NELLE FORMULE PER LA STIMA DEI PARAMETRI CON IL METODO DEI MINIMI QUADRATI.

I valori da calcolare sono:

MEDIA	403,75	562,5	163987,5	321250	225062,5
QUADR.	163014,063	316406,25			

$\sum_{i=1}^{n} x_i = 3230$ grazie alla quale è possibile calcolare $\bar{y} = 403{,}75$ [3]

$\sum_{i=1}^{n} y_i = 4500$ grazie alla quale è possibile calcolare $\bar{y} = 562{,}5$ [4]

Calcoliamo ora la **varianza di x**, ricordando che essa può essere calcolata tramite la differenza tra la media delle x_i al quadrato meno il quadrato della media di x_i

$$\sum_{i=1}^{n} x^2 = 2570000 \rightarrow \text{media di } x^2 = 321250 \rightarrow S_x^2 = 4843{,}75$$

la **varianza della y**

$$\sum_{i=1}^{n} y^2 = 1311900 \rightarrow \text{media di } y^2 = 163987{,}5 \rightarrow S_y^2 = 973{,}44$$

[3] La formula della media aritmetica è $\frac{\sum_{i=1}^{n} x_i}{N}$.

[4] Per la formula della media aritmetica si guardi la nota precedente.

e la **covarianza di x ed y**, che può essere calcolata come la differenza tra la media dei prodotti di x per y meno il prodotto delle medie di x e di y

$$\sum_{i=1}^{n} xy = 1800500 \rightarrow \text{media} = 225062{,}5$$

il prodotto delle medie è 227109,38

$$\rightarrow S_{xy} = -2046{,}875$$

Inseriamo i valori ottenuti nelle formule

$$\begin{cases} \hat{\beta}_0 = \dfrac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \\ \hat{\beta}_1 = \dfrac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \end{cases}$$

ed otteniamo che

$$\begin{cases} \hat{\beta}_1 = \dfrac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} = \dfrac{S_{xy}}{S_x^2} = \dfrac{-2046{,}875}{4843{,}75} = -0{,}422 \\ \hat{\beta}_0 = \dfrac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} = \bar{y} - \hat{\beta}_1 \bar{x} = (403{,}75) - [(-0{,}422)(562{,}5)] = 641{,}125 \end{cases}$$

FASE 3: DETERMINAZIONE DELLA RETTA DI REGRESSIONE.

Avendo stimato i parametri, ed essendo $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ l'equazione della retta di regressione, possiamo quindi ottenere

$$\hat{Y} = 641{,}125 - 0{,}422X$$

13.2. Proprietà delle stime ottenute con il metodo dei minimi quadrati.

Abbiamo già anticipato alcune delle ragioni che spiegano la diffusione del metodo dei minimi quadrati e che ci aiutano a comprenderne la diffusione. A quanto precedentemente detto, va qui aggiunto che la retta stimata con il metodo dei minimi quadrati gode di alcune importantissime proprietà:

1. innanzitutto, è l'unica retta che rende minima la somma dei quadrati degli scarti $S(\beta_0;\beta_1)$: non è, quindi, possibile trovare nessun'altra retta che possieda la medesima somma dei quadrati, minima per ipotesi;
1. passa sempre per il punto di coordinate $(\bar{x};\bar{y})$ e il punto $X=\bar{x}\ e\ Y=\bar{y}$ soddisfa sempre l'equazione $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X$;
2. infine, è tale che $\sum \hat{e}_i = 0$, posto che $e_i = y_i - \hat{y}_i$, la qual cosa implica che la media campionaria delle Y osservate coincide con la media delle $\hat{Y}$ calcolate mediante la retta dei minimi quadrati.

Il metodo dei minimi quadrati, ci consente inoltre di stimare i parametri $(\hat{\beta}_0;\hat{\beta}_1)$ in modo tale che essi godano di talune specifiche proprietà.

Prima di presentarle, è bene fare una breve leggenda che contenga le notazioni relative alle quantità di nostro interesse. Indichiamo con:

- $(\beta_0;\beta_1)$ i valori veri dei parametri della relazione lineare ipotizzata;
- $(\hat{\beta}_0;\hat{\beta}_1)$ le stime dei minimi quadrati per un campione specifico;

- $(B_0; B_1)$ le variabili casuali stimatori dei minimi quadrati.

La bontà del modello dei minimi quadrati non può essere giudicata sulla base dei valori numerici $(\hat{\beta}_0; \hat{\beta}_1)$ ottenuti da un solo campione, ma bisognerà esaminare le proprietà della variabile casuale doppia $(B_0; B_1)$ generata dalla stima dei minimi quadrati. Ciò significa che per comprendere le proprietà delle stime fatte con i minimi quadrati, sarà necessario guardare alle caratteristiche di una variabile casuale doppia che sono la media e la varianza delle variabili casuali componenti la loro covarianza oppure il coefficiente di correlazione. Si può dimostrare che[5]:

1. $E(B_0) = \beta_0 \qquad E(B_1) = \beta_1$
2. $Var(B_0) = \sigma^2 \dfrac{\sum x_i^2}{N\sum (x_i - \bar{x})^2} = \dfrac{\sigma^2}{N}\left(1 + \dfrac{\bar{x}^2}{S_x^2}\right)$ e

   $Var(B_1) = \dfrac{\sigma^2}{\sum (x_i - \bar{x}^2)} = \dfrac{\sigma^2}{NS_x^2}$
3. $Cov(B_0; B_1) = \sigma^2 \dfrac{-\bar{x}}{\sum (x_i - \bar{x})^2} = \dfrac{-\sigma^2 \bar{x}}{NS_x^2}$
4. $Corr(B_0; B_1) = \dfrac{-\bar{x}}{\sqrt{\sum x_i^2 / N}}$

Soffermiamo la nostra attenzione sulla dimostrazione della prima espressione. Sapendo che X è una variabile deterministica, che $\hat{\beta}_1$ è un valore osservato per la variabile casuale B_1 mentre y_i è un valore osservato per la variabile casuale Y_i, gli stimatori $(B_0; B_1)$ si possono scrivere come

[5] Vd. Johnston e Goldberger.

$$\begin{cases} B_1 = \dfrac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} \\ B_0 = \bar{Y} - B_1 \bar{x} \end{cases}$$

In base a $E(B_0) = \beta_0$ *e* $E(B_1) = \beta_1$ e a quanto sinora detto, otteniamo

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + e_i) = \beta_0 + \beta_1 x_i + E(e_i) = \beta_0 + \beta_1 x_i$$
$$E(\bar{Y}) = E\left(\frac{1}{N}\sum Y_i\right) = \frac{1}{N}\sum E(Y_i) = \frac{1}{N}\sum(\beta_0 + \beta_1 x) = \frac{1}{N}\left[N\beta_0 + \beta_1 \sum x_i\right] = \beta_0 + \beta_1 \bar{x}$$
$$E(Y_i - \bar{Y}) = E(Y_i) - E(\bar{Y}) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x) = \beta_1 x_i - \beta_1 \bar{x} = \beta_1 (x_i - \bar{x})$$

Tenendo conto della proprietà del valor medio, possiamo ottenere

$$E(B_1) = \frac{\sum (x_i - \bar{x}) E(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) \beta_1 (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \beta_1 \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} = \beta_1$$
$$E(B_0) = E(\bar{Y} - B_1 \bar{x}) = E(\bar{Y}) - E(B_1)\bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$$

Abbiamo in questo modo dimostrato che gli stimatori OLS sono non distorti poiché, in media, essi equivalgono ai parametri che cercano di stimare.

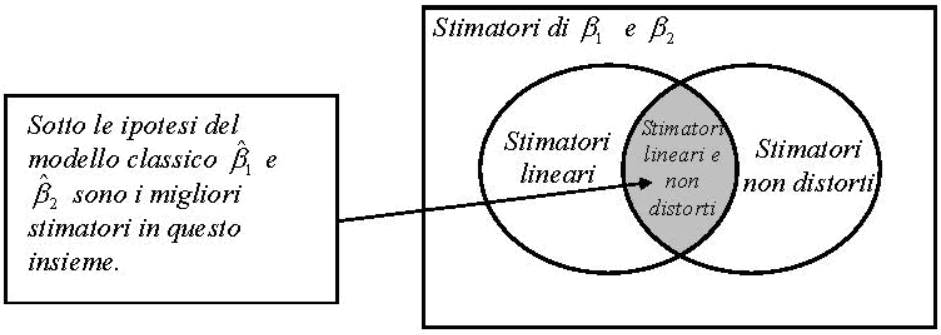
La varianza di tali stimatori tende a zero al crescere di N: poiché non sono distorti, tale proprietà implica che essi siano anche consistenti. L'importanza di questa osservazione è di immediata evidenza: se la non distorsione assicura che tali stimatori coincidono in media con i parametri veri della popolazione, il tendere a zero della loro varianza implica che la precisione delle stime aumenta con N. A ciò si aggiunga, inoltre, che le stime OLS sono lineari perché possono essere scritte come una combinazione lineare delle osservazioni Y_i. Grazie al teorema di Gauss – Markov, è inoltre possibile dimostrare che, tra tutte le stime lineari non distorte, gli stimatori dei minimi

quadrati sono anche quelli più efficienti, ossia con varianza minima. La covarianza dipende, infatti, dal segno di x: se è positivo la covarianza è negativ (da cui discende che, se si sovrastima la pendenza, si sottostima l'intercetta e viceversa), se invece x è negativo la covarianza è positiva e pertanto tendenzialmente gli errori nella stima di β_1 e β_2 hanno lo stesso segno.

Il risultato più importante sulle proprietà degli stimatori dei minimi quadrati è costituito dal teorema di Gauss Markov.

Teorema di Gauss Markov*: sotto le ipotesi del modello classico, gli stimatori dei minimi quadrati sono i più efficienti nella classe degli stimatori lineari e non distorti.*

Non è quindi possibile che esista un'altra coppia di stimatori per β_1 e β_2 che siano lineari e non distorti e abbiano varianza minore degli stimatori dei minimi quadrati.



Riguardo alle proprietà asintotiche è possibile affermare che gli stimatori dei minimi quadrati sono consistenti in media quadratica. Infatti sono non distorti e la loro varianza tende a zero,

$$\lim_{n\to\infty} Var\left(\hat{\beta}_1\right) = \sigma^2 \lim_{n\to\infty}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)$$

$$\lim_{n\to\infty} Var\left(\hat{\beta}_2\right) = \lim_{n\to\infty} \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = 0$$

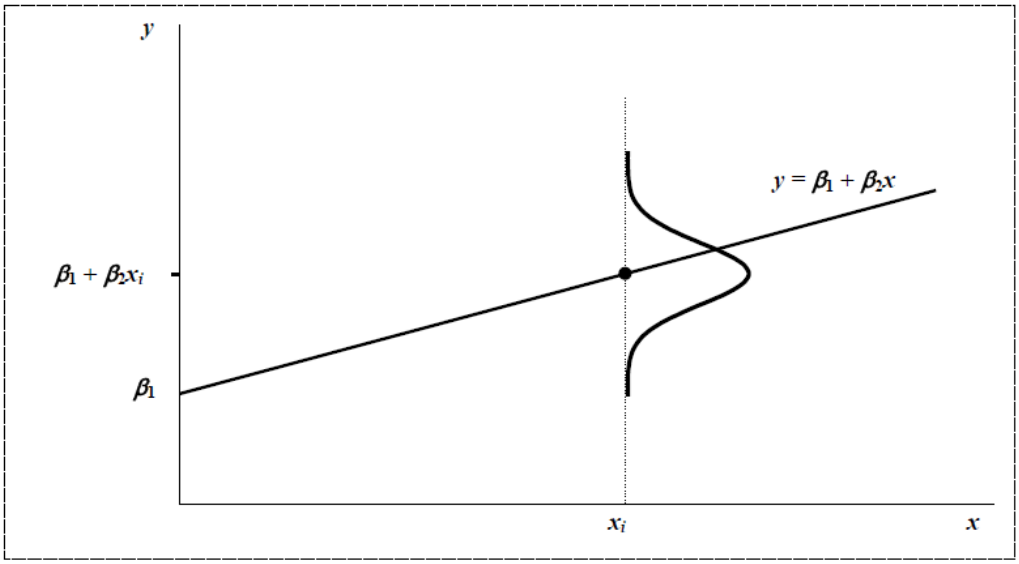
Occorre inoltre aggiungere che, asintoticamente, gli stimatori dei minimi quadrati sono normali.

13.3. Il metodo della massima verosimiglianza.

Un altro metodo per stimare i parametri della retta di regressione è qullo c.d. della massima verosimiglianza. Il metodo consiste nel massimizzare la funzione di verosimiglianza, definita in base alla probabilità di osservare una data realizzazione campionaria, condizionatamente ai valori assunti dai parametri oggetto di stima.

Nel modello di regressione lineare semplice, una delle ipotesi fondamentali è che $E(\varepsilon \mid x) = 0$, da cui $y = E(y \mid x) + \varepsilon$, ossia la retta di regressione rappresenta la media condizionale della variabile y. Se a questa aggiungiamo l'ipotesi di normalità degli errori, $\varepsilon|x \sim N(0,\sigma 2)$, otteniamo $y \mid x \sim N(\beta_1 + \beta_2, x, \sigma^2)$.

La distribuzione condizionale di *y*, quindi, è nota a meno di un numero limitato di parametri, (β1, β2, σ) e la stima dei parametri del modello di regressione equivale alla stima della media e della varianza di tale distribuzione.



Per trovare lo stimatore di massima verosimiglianza del vettore di parametri (β_1, β_2, σ) occorre costruire e massimizzare la funzione di verosimiglianza.

$$f\left(y_i|x_i,\beta_0,\beta_1,\sigma\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y_i-\beta_0-\beta_1x_i}{\sigma}\right)^2}$$

$$f\left(y_1,...,y_n|x_1,...,x_n;\beta_0,\beta_1,\sigma\right)=\prod_{i=1}^{n}f\left(y_i|x_i,\beta_0,\beta_1,\sigma\right)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y_i-\beta_0-\beta_1x_i}{\sigma}\right)^2}$$

$$L\left(\beta_0,\beta_1,\sigma|y_1,...,y_n;x_1,...,x_n\right)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y_i-\beta_0-\beta_1x_i}{\sigma}\right)^2}$$

$$\ln L\left(\beta_0,\beta_1,\sigma|y_1,...,y_n;x_1,...,x_n\right)=n\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_2x_i\right)^2\right)$$

formula precedente, ossia della sommatoria tra parentesi. Ma ciò equivale esattamente alla minimizzazione della somma dei residui al quadrato, da cui si ottiene lo stimatore dei minimi quadrati ordinari. Ne consegue, quindi, che i due stimatori sono equivalenti $\left(\hat{\beta}_{MLE}=\hat{\beta}_{OLS}\right)$.

Sostituendo, però, i valori così ottenuti nella condizione del primo ordine relativa a σ, si ottiene:

$$\hat{\sigma}^2{}_{MLE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)^2=\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

che è, invece, diverso dallo stimatore corretto della varianza calcolato con il metodo dei minimi quadrati.

13.4. Proprietà degli stimatori di massima verosimiglianza.

Nella sezione precedente abbiamo visto che lo stimatore di massima verosimiglianza coincide con quello dei minimi quadrati. Rispetto ad esso, però, in piccoli campioni, il stimatore MLE non ha talune prorietà, quali ad esempio la non distorsione, di cui gode invece lo stimatore determinato con il metodo dei minimi quadrati.

Vanno comunque riconosciute a $\hat{\beta}_{MLE}$ alcune fondamentali proprietà asintotiche. Dato un vettore di parametri, è infatti possibile dimostrare che $\hat{\beta}_{MLE}$ è consistente, asintoticamente efficiente[6] ed asintoticamente distribuito secondo una normale[7].

Affinché lo stimatore di massima verosimiglianza goda delle suddette proprietà è necessario che la funzione di verosimiglianza sia correttamente specificata e, quindi, che l'ipotesi sulla distribuzione degli errori sia corretta. Ciò non è evidente nel caso del modello lineare con ipotesi di normalità qui trattato perché lo stimatore di massima verosimiglianza coincide con quello dei minimi quadrati (almeno per quanto riguarda i coefficienti delle variabili esplicative) e quindi ne assume le proprietà asintotiche di consistenza e normalità, anche se la distribuzione degli errori non è normale. Tale ipotesi risulta dunque essere cruciale.

---

[6] tra tutti gli stimatori consistenti, lo stimatore di massima verosimiglianza è quello con varianza più "piccola". Per questo motivo, sotto l'ipotesi di normalità, lo stimatore OLS è BUE (*best unbiased estimator*) e non semplicemente BLUE, poiché coincide con lo stimatore MLE.

[7] Abbiamo, infatti, che

$$\sqrt{N}(\hat{\theta}\text{-}\theta) \rightarrow N(0,V) \text{ dove } V = \{I(\theta)\}^{-1} \text{ e } I(\theta) = -E\left\{\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right\} \text{ (matrice di informazione)}$$

13.5. Intervalli di confidenza per i coefficienti di regressione.

Assumendo che la variabile dipendente Y si distribuisce come una normale, la distribuzione degli stimatori dei minimi quadrati tende ad approssimarsi ad una variabile casuale normale. Inoltre, dal momento che B1 è una combinazione lineare delle variabili casuali Y, anche B1 si distribuirà normalmente con media β_1 e σ^2 / S_{xx}. Anche B0, al variare delle osservazioni campionarie, tenderà a distribuirsi secondo una normale con β_0 e varianza $\sigma^2 \sum_{i=1}^{n} x_i^2 / nS_{xx}$.

Data questa premessa, anche le quantità $Z = \frac{(B_1 - \beta_1)}{\sigma / \sqrt{S_{xx}}}$ e $Z = \frac{(B_0 - \beta_0)}{\sqrt{\frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}}$ tenderanno a distribuirsi come leggi normali standardizzate.

Poiché le formule impiegate contengono varianza incognita, questa andrà stimata utilizzando MSE*=SSE*/(n-2). Sostituendo nelle precedenti espressioni il valore incognito con quello stimato, otteniamo $t = (B_1 - \beta_1) / \sqrt{MSE^* / S_{xx}}$; $t = (B_0 - \beta_0) / \sqrt{MSE^* \left(\sum_{i=1}^{n} x_i^2 / nS_{xx} \right)}$[8], e cioè due t di Student che si distribuiscono con n-2 gradi di libertà.

Sotto l'ipotesi di normalità, $(MSE^* / \sigma^2)(n-2)$ ha una distribuzione χ^2 con n-2 gradi di libertà e si può inoltre dimostrare che $(MSE^* / \sigma^2)(n-2)$ e B1 sono indipendenti. Sotto queste condizioni,

[8] Nella ricerca empirica, accade spesso che la varianza, σ^2, sia incognita. Si rende, quindi, necessario stimare la varianza corretta, che qui indichiamo con la sigla MSE.

$$t = \frac{\frac{(B_1 - \beta_1)}{\sigma / \sqrt{S_{xx}}}}{\sqrt{MSE^*(n-2)/\sigma^2(n-2)}} = \frac{B_1 - \beta_1}{\sqrt{MSE^*}}\sqrt{S_{xx}} = \frac{(B_1 - \beta_1)\sqrt{n-2}\sqrt{S_{xx}}}{\sqrt{SSE^*}}$$

è una v.c. pivot con una distribuzione t di Student con n-2 gradi di libertà.

Un intervallo di confidenza per β_1 con un livello di confidenza pari a $(1-\alpha)$ è dato da

$$P\left(B_1 - t_{\alpha/2,n-2}\sqrt{MSE^*/S_{xx}}\right) \le \beta_1 \le B_1 + t_{\alpha/2,n-2}\sqrt{MSE^*/S_{xx}} = 1-\alpha$$

mentre per β_0 è dato da

$$P\left(B_0 - t_{\alpha/2,n-2}\sqrt{MSE^*\left(\sum_{i=1}^{n} x_i^2/nS_{xx}\right)}\right) \le \beta_1 \le \left(B_0 - t_{\alpha/2,n-2}\sqrt{MSE^*\left(\sum_{i=1}^{n} x_i^2/nS_{xx}\right)}\right) = 1-\alpha$$

Molto spesso, può esser utile determinare un intervallo di confidenza intorno alla vera retta di regressione. Per poterlo fare, al 100(1-α%), si ricorre alle seguenti quantità:

$$L_1 = \hat{Y} - t_{\alpha/2,n-2}\sqrt{MSE^*\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - x)^2}\right)}$$

$$L_2 = \hat{Y} + t_{\alpha/2,n-2}\sqrt{MSE^*\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - x)^2}\right)}$$

Tali valori possono essere rappresentati su di un piano in funzione di x_0, insieme alla retta di regressione.

Nel seguito, proponiamo un quadro di sintesi relativamente a tutti gli intervalli di confidenza nel modello di regressione lineare.

Parametro	Intervallo di confidenza
β_0	$b_0 \pm t_{\alpha/2,n-2}\sqrt{MSE*\left(\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}\right)}$
β_1	$b_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{MSE*}{S_{xx}}}$
$\mu_{y/x_0} = E(Y_0)$	$\hat{Y}_0 \pm t_{\alpha/2,n-2}\sqrt{MSE*\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$
Y_0	$\hat{Y}_0 \pm t_{\alpha/2,n-2}\sqrt{MSE*\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$
$\frac{1}{2}\log\frac{1+\rho}{1-\rho}$	$\frac{1}{2}\log\frac{1+r*}{1-r*} \pm z_{\alpha/2}\frac{1}{\sqrt{n-3}}$

con

r* = stimatore di r (coefficiente di correlazione)

MSE* = stimatore della varianza.

13.6. Verifica delle ipotesi dei parametri del modello di regressione.

Dopo aver stimato i parametri della retta di regressione (fasi 1, 2 e 3), continuiamo lo svolgimento dell'esercizio proposto all'inizio del capitolo, dedicandoci ora alla verifica delle ipotesi.

Attraverso la verifica delle ipotesi si vuole verificare l'esistenza di un legame lineare tra la variabile dipendente e quella indipendente.

Le ipotesi sono

$H_0 : \beta_1 = 0$ non esiste alcuna relazione lineare tra variabile dipendente e regressore

$H_1 : \beta_1 \neq 0$ esiste una relazione lineare tra le due variabili

Per stabilire se la relazione lineare verificata a livello campionario mediante il valore **b** (b = coefficiente di regressione campionario) possa essere considerata valida anche per la popolazione**,** occorre fare inferenza sui parametri della retta di regressione: facendo inferenza sui parametri della popolazione si vuole, infatti, verificare se la retta di regressione campionaria (retta **CAM)** possa essere ritenuta una buona espressione della vera relazione lineare esistente nella popolazione (retta **POP**).

Si noti anzitutto che il ricercatore è chiamato a verificare l'ipotesi nulla (H_0) e non l'ipotesi alternativa: quest'ultima, quindi, viene accettata o rifiutata solo come conseguenza di ciò che verrà fatto della prima ipotesi → ***Rifiutare H_0 significa ammettere che nella popolazione vi è dipendenza lineare.***

La relativa statistica test è

$$t_{n-2} = \frac{b_1}{S_{b_1}}$$

con

$$S_{b_1} = \frac{S_{yx}}{\sqrt{\sum_{i=1}^{n} x_i - \frac{\sum_{i=1}^{n} x_i}{n}}}.$$

La regola di decisione da adottare è $|t| = t_{\alpha/2,n-2}$ si rifiuta l'ipotesi nulla e si conclude che la variabile X, il regressore, ha un significato statistico nello spiegare le variazioni della variabile dipendente.

FASE 4: PER VERIFICARE L'ESISTENZA DI UNA RELAZIONE LINEARE TRA VARIABILE DIPENDENTE (Y) E VARIABILE INDIPENDENTE (X)

FASE 4.1. PREDISPORRE UN ADEGUATO FOGLIO DI CALCOLO O PREPARARE UNA TABELLA NEL SEGUENTE MODO:

i	vendite (yi)	prezzi (xi)	vendite stimate con il modello	residui	quadrati
A	410	550	409,025	0,975	0,950625
B	380	600	387,925	-7,925	62,805625
C	350	650	366,825	-16,825	283,08063
D	400	600	387,925	12,075	145,80563
E	440	500	430,125	9,875	97,515625
F	380	650	366,825	13,175	173,58063
G	450	450	451,225	-1,225	1,500625
H	420	500	430,125	-10,125	102,51563
TOT	**3230**	**4500**	**3230**	**0**	**867,755**

Il valore delle vendite stimate è stato calcolato inserendo nella $\hat{Y} = 641{,}125 - 0{,}422X$ i diversi livelli di prezzo. Si può immediatamente notare

che la somma dei valori osservati e di quelli stimati dal modello è la stessa. In questo caso, calcolare tutti i valori teorici è stato possibile poiché abbiamo un n=8. In molti casi, invece, questa operazione risulta proibitiva perché la numerosità campionaria è molto più consistente. In tal caso, possiamo ricorrere alle seguenti formule

$$\sum_{i=1}^{n}\left(y_i - y_i^*\right) = \sum_{i=1}^{n} y_i^{\,2} - b_0\sum_{i=1}^{n}\gamma_i - b_1\sum_{i=1}^{n} x_i\gamma_i \,,$$

da cui si ottiene

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^{n}\gamma_i^{\,2} - b_0\sum_{i=1}^{n}\gamma_i - b_1\sum_{i=1}^{n} x_i\gamma_i}{n-2}}\,.$$

FASE 4.2. CALCOLARE I RESIDUI.

Dopo aver calcolato i valori teorici, possiamo calcolare il residuo in base alla formula $\hat{e} = y_i - \hat{y}_i$ (si noti che la somma dei residui, coerentemente con le ipotesi assunte alla base del modello di regressione lineare, è nulla).

FASE 4.3. DETERMINAZIONE DEL TEST EMPIRICO.

Dopo aver calcolato i residui, eleviamo al quadrato tutti valori ottenuti in modo da poter determinare il valore del test critico applicando la seguente formula:

$$t_{n-2} = \frac{b_1}{S_{b_1}}$$

con b_1 lo stimatore ai minimi (intercetta della retta di regressione) e con

$$S_{b_1} = \frac{S_{yx}}{\sqrt{\sum_{i=1}^{n} x_i - \frac{\sum_{i=1}^{n} x_i}{n}}}$$

e

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - y_i^{*} \right)}{n-2}}$$

Svolgiamo i calcoli partendo dall'ultima formula, che nel nostro caso appare nella seguente forma

$$S_{yx} = \sqrt{\frac{867,755}{8-2}} = 12,03$$

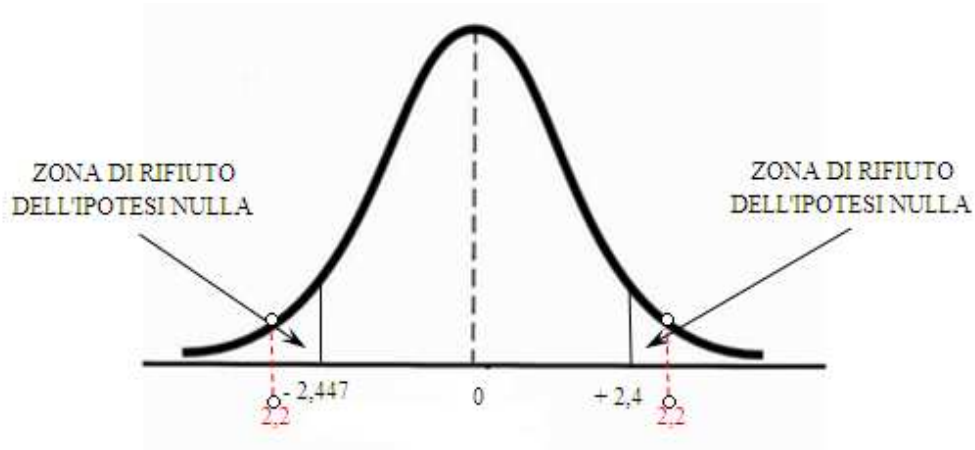
Otteniamo, così

$$S_{b_1} = \frac{12,03}{\sqrt{4500 - \frac{4500}{8}}} = 0,19$$

Il valore del test empirico è, quindi, pari a

$$t_{n-2} = \frac{(-0,422)}{0,19} = -2,22$$

Per accettare o rifiutare l'ipotesi nulla, la regola di decisione è: se $|t| > t_{\alpha/2,n-2}$ allora si rifiuta l'ipotesi nulla e si conclude che la variabile X ha significato statistico nello spiegare le variazione della variabile dipendente. Per calcolare il valore critico del test per α/2 ed (n-2) g.l., ricorriamo alle tavole della t di Student. Supponendo di aver fissato α/2=0,025, il valore del test è 2,447. Essendo $|2{,}447| > 1{,}943$ rifiutiamo l'ipotesi nulla e verifichiamo, quindi, che il prezzo praticato incide sul volume delle vendite.



Rifiutiamo l'ipotesi nulla perché il valore del test empirico cade oltre la soglia critica, e cioè nella zona di rifiuto dell'ipotesi nulla. Possiamo, quindi, verificare che il prezzo praticato incide sul volume delle vendite.

13.7. Misura della bontà di adattamento

Ogni modello statistico è un'approssimazione della realtà e, come tale, quindi, assolutamente imprecisa: tutti i modelli sono sbagliati (nel senso che non riproducono esattamente il fenomeno oggetto di studio) e deve, quindi, essere obiettivo del ricercatore quello di individuare il modello che meglio si accosta rappresenti la realtà.

Per riuscire nell'intento, vi sono diversi strumenti. Tra questi, ricordiamo:

1. indice di determinazione;
2. coefficiente di correlazione;
3. analisi dei residui.

L'indice di determinazione è un indice attraverso il quale poter verificare il livello di accostamento tra i valori teorici (giacenti sulla retta di regressione) e quelli osservati. La formula attraverso cui determinarlo è

$$r^2 = SSR / S_{yy} = \frac{\text{devianza di regressione}}{\text{devianza totale}}$$

Per verificare quanto i valori teorici stimati mediante la retta di regressione siano vicini ai valori osservati, e cioè per avere una misura della bontà di accostamento, occorre ricordare che

$$\text{dev. totale} = \text{dev. residua meno dev. di regressione}$$

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

$$S_{yy} = SSE - SSR$$

Il coefficiente di determinazione è, quindi, interpretabile come la porzione della devianza totale piegata dal modello di regressione prescelto. Il calcolo dell'indice di determinazione è molto semplice ed è

$$R^2 = 1 - \frac{Dev(\hat{e})}{Dev(\hat{y})} = 1 - \frac{N\left(S_y^2 - S_{xy}^2 / S_x^2\right)}{NS_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2} = r_{xy}^2 \ .$$

Questo risultato dimostra che l'indice di determinazione non è altro che il quadrato della stima del coefficiente di correlazione lineare $\rho(X,Y)$, anche detto coefficiente di correlazione lineare di Pearson. Esso varia tra -1 e +1 ed assume le seguenti sfumature di significato:

> 0	le variabili *x* e *y* si dicono *direttamente correlate*, oppure *correlate positivamente* (ad una variazione della variabile indipendente corrisponde una variazione dello stesso segno della variabile dipendente).
= 0	le variabili *x* e *y* si dicono *indipendenti* (le variazioni della variabili dipendente non influenzano "reazioni" della variabile dipendente).
< 0	le variabili *x* e *y* si dicono *inversamente correlate*, oppure *correlate negativamente* (ad una variazione della variabile indipendente corrisponde una variazione di segno opposto della variabile dipendente).

Nel caso di indipendenza lineare il coefficiente assume valore zero, mentre non vale la conclusione opposta, ovvero dal coefficiente nullo non si può desumere l'indipendenza lineare: tale condizione può, quindi, dirsi necessaria ma non sufficiente per l'indipendenza delle due variabili.

La verifica delle ipotesi può essere condotta anche sul coefficiente di determinazione e sul coefficiente di correlazione. Illustriamo entrambi nel seguente specchietto riassuntivo, nel quale verrà ricordata anche la verifica delle ipotesi fatta sulla ipotesi di relazione lineare tra variabile dipendente e variabile indipendente.

Vediamo un'applicazione facendo ricorso ai nostri dati.

FASE 5: MISURA DELLA BONTÀ DI ADATTAMENTO DEL MODELLO AI DATI

FASE 5.1. PER PRIMA COSA, OCCORRE PREDISPORRE NUOVAMENTE LA TABELLA DI CALCOLO NEL SEGUENTE MODO.

i	**vendite (yi)**	**prezzi (xi)**	**y^2**	**x^2**	**x*y**
A	410	550	168100	302500	225500
B	380	600	144400	360000	228000
C	350	650	122500	422500	227500
D	400	600	160000	360000	240000
E	440	500	193600	250000	220000
F	380	650	144400	422500	247000
G	450	450	202500	202500	202500
H	420	500	176400	250000	210000
TOT	3230	4500	1311900	2570000	1800500
MEDIA	403,75	562,5	163987,5	321250	225062,5
QUADR.	163014,06	316406,25			

Per misura la bontà di adattamento del modello ai dati, useremo l'indice di determinazione lineare e la sua radice quadrato, il coefficiente di correlazione lineare.

L'indice di determinazione lineare è calcolato tramite la formula

$$R^2 = b_{xy} - b_{yx}$$

con

$$b_{xy} = \frac{\sum_{i=1}^{n}(xy) - n(\bar{x}\,\bar{y})}{\sum_{i=1}^{n} x^2 - n(\bar{x})^2} = \frac{cod(x;y)}{dev(x)}$$

e

$$b_{yx} = \frac{\sum_{i=1}^{n}(xy) - n(\bar{x}\,\bar{y})}{\sum_{i=1}^{n} y^2 - n(\bar{y})^2} = \frac{cod(x;y)}{dev(y)}$$

Il coefficiente di correlazione è, quindi, dato da

$$r = \pm\sqrt{b_0 - b_1}\ .$$

Inserendo i dati nelle formule, otteniamo

cod (x,y) =	-16375
dev (x) =	38750
dev (y) =	7787,5
bxy=	-0,4225806
byx=	-2,1027287

E, quindi:

R^2=	0,8885725
r=	0,9426412

FASE 5.1. COSTRUZIONE DEL TEST SU ρ.

Le ipotesi da formulare sono le seguenti

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

Il test empirico per verificare le ipotesi formulate viene determinato ricorrendo alla seguente formula

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-R^2}}$$

Che nel nostro caso risulta pari a 6,92. Quale test utilizziamo per la verifica delle ipotesi? Utilizziamo una t di Student con (n-2)g.l. [9], il cui valore, in questo caso, è pari a

$$t_{0,025;8-2\,g.l.} = 2{,}447$$

La regola di decisione è la stessa vista in precedenza: se il test empirico è maggiore del test critico allora si rifiuta l'ipotesi nulla.

[9] Quando n (=numerosità campionaria) è grande, lo stimatore r* approssima una Normale N[0;1/(n-1)].

Nel seguito, proponiamo un quadro di sintesi relativamente alla verifca delle ipotesi nel modello di regressione lineare:

H_0	Stimatore	Statistica test	Distribuzione sotto H_0
$\beta_0 = 0$	B_0	$\frac{b_0}{\sqrt{MSE\left[\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}\right]}}$	t_{n-2} t di Student
$\beta_1 = 0$	B_1	$\frac{b_1}{\sqrt{\frac{MSE}{S_{xx}}}}$	t_{n-2} t di Student
$\rho^2 = 0$	r^{*2}	$\frac{MSR*}{MSE*}$	$F*_{1,n-2}$ F di Fischer - Snedecor
$\rho = 0$	r^*	$\frac{r^*\sqrt{n-2}}{\sqrt{1-r^{*2}}}$	t_{n-2} t di Student
$\rho = \rho_0$	r^*	$\frac{r^*\left[\frac{1}{2}\log_e\frac{1+r^*}{1-r*}-\left(\frac{1}{2}\log_e\frac{1+\rho_0}{1-\rho_0}+\frac{\rho_0}{2(n-1)}\right)\right]}{1/\sqrt{n-3}}$	Z Normale

con

B_1= stimatore del coefficiente angolare della retta di regressione

r^{*2}=stimatore del coefficiente di determinazione

MSR*=SSR*/1=stimatore della varianza di regressione