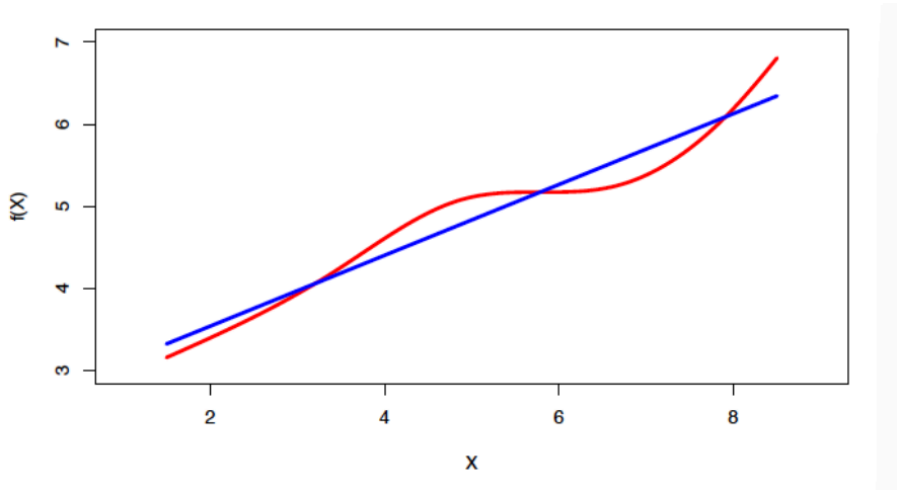
2 - Linear Regression A

La regressione lineare

> 💡 La regressione lineare è un approccio semplice per il Supervised Learning. E' il più semplice possibile, sfrutta sia le x sia le y.

Questo approccio parte assumendo che la dipendenza tra Y e X_1, X_2,... X_P è lineare, anche se le **vere** funzioni di regressione non sono **mai** lineari. Nel caso monodimensionale si parla di *simple linear regression*.



Anche se può sembrare troppo semplicistica la regressione lineare è estremamente utile sia concettualmente che praticamente.

Regressione lineare su dati pubblicitari

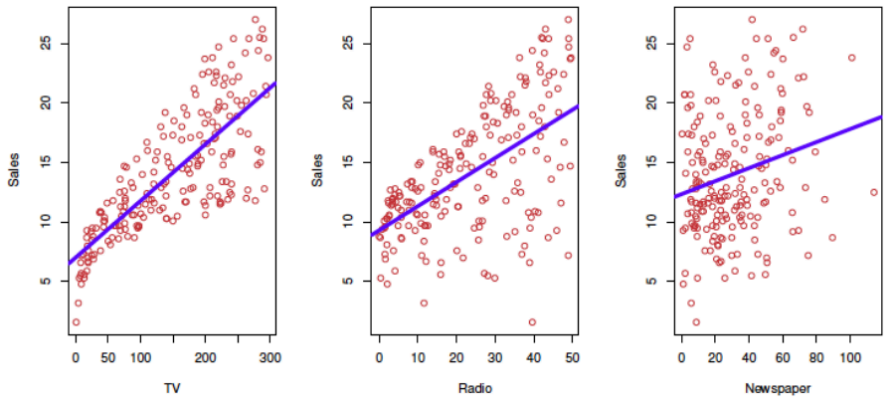
Il Data Set consiste dei dati di vendita di un particolare prodotto insieme ai budget pubblicitari per quel prodotto su TV, radio e giornali.

Sulla base di questi dati vogliamo formulare un piano di marketing per il prossimo anno che risulterà in molte vendite del prodotto.

Ci interessano diverse domande:

- C'è una **relazione tra vendite e pubblicità**?
- **Quanto è forte** la relazione?
- **Quale mezzo** pubblicitario contribuisce alle vendite?
- **Con quanta accuratezza** possiamo stimare l'effetto di ogni mezzo pubblicitario?

- Con quanta accuratezza possiamo **predire future** vendite?
- La relazione **è lineare**?
- C'è un effetto di **interazione tra i diversi mezzi** pubblicitari?



Regressione lineare semplice usando un solo predittore X

Assumiamo il modello:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Dove Y sono le vendite e X rappresenta il budget pubblicitario su TV, radio e giornali.

Nel modello in questione β_0 e β_1 sono due costanti ignote che rappresentano la intercept (intercetta) e la slope (pendenza), anche note come **coefficienti o parametri**; come al solito ε è il **termine di errore**.

Il nostro obiettivo è la stima di future Y a partire da future X, trattandosi di un modello parametrico ottenere la stima di Y cioè $\hat{y}$ diventa una questione di ottenere $\hat{\beta}_0$ e $\hat{\beta}_1$ cappello cioè le stime dei coefficienti.

Date delle stime dei coefficienti del modello $\hat{\beta}_0$ e $\hat{\beta}_1$, prediciamo le future vendite usando:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Dove $\hat{y}$ indica una predizione di Y sulla base di X = x, cioè lo stimatore di Y.

Il simbolo cappello indica una stima.

Stima dei parametri tramite least squares (minimi quadrati)

Ricordiamo di essere in possesso di un data set con n coppie osservate del tipo $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Sia $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la stima di Y basata sull'i-esimo valore di X, allora

$$e_i = y_i - \hat{y}_i$$

rappresenta l'i-esimo **residuo**.

9

But be aware that Sum of Squared Errors (SSE) and Residue Sum of Squares (RSS) sometimes are used interchangeably, thus confusing the readers. For instance, check this URL out.

Strictly speaking from statistic point of views, Errors and Residues are completely different concepts. Errors mainly refer to difference between actual observed sample values and your predicted values, and used mostly in the statistic metrics like Root Means Squared Errors (RMSE) and Mean Absolute Errors (MAE). In contrast, residues refer exclusively to the differences between dependent variables and estimations from linear regression.

Share Cite Improve this answer Follow

edited Nov 7, 2020 at 13:57 user5305519 103 5

answered Jun 16, 2019 at 17:04 Dr.CYY 87 1 1

Definiamo **residual sum of squares (RSS)** o somma residua dei quadrati il valore:

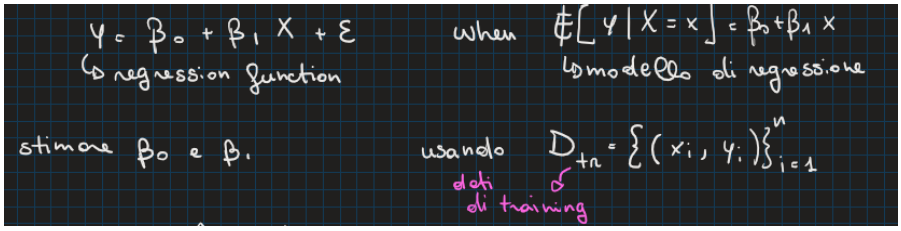
$$RSS = e_1^2 + e_2^2 + ... + e_n^2$$

o in maniera equivalente:

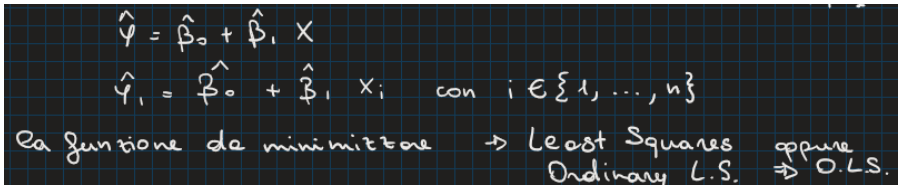
$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

L'approccio dei minimi quadrati (LS) sceglie $\hat{\beta}_0$ e $\hat{\beta}_1$ tali da minimizzare l'RSS.

Dimostrazione - importanza del RSS in relazione all'MSE



D_{tr} dato che siamo in contesto supervisionato sono x_i e y_i, n è la cardinalità delle osservazioni

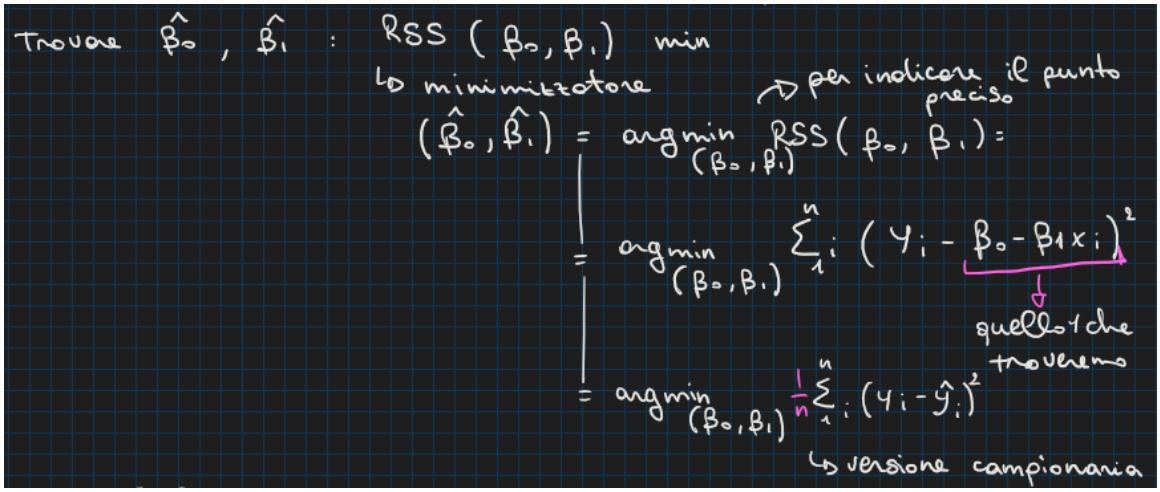


Riscriviamo la nostra funzione di regressione sostituendo $\hat{Y}$ con $\hat{y}_i$ e X con x_i.

La funzione da minimizzare è quella dei quadrati (RSS), secondo il criterio Least squares o talvolta Ordinary Least Squares (OLS).

Dobbiamo trovare $\hat{\beta}_0$ e $\hat{\beta}_1$ tali che minimizzano RSS (che è funzione di β_0 e β_1 nello spazio di tutti i possibili valori di β_0 e β_1). Il punto che stiamo cercando si chiama minimizzatore.

Quindi ($\hat{\beta}_0$, $\hat{\beta}_1$) sarà il **minimizzatore** dell'RSS, questo vuol dire che è l'argmin dell'RSS definito come sommatoria.



Quello sottolineato in rosa presenta una particolarità: β_0 e β_1 non hanno i cappelli, NON perché siano i valori veri ma sono perché il professore voleva evidenziare come questo processo serva a calcolare i valori $\hat{\beta}_0$, $\hat{\beta}_1$, ma comunque poi nella formula dell'RSS sono i valori cappello e infatti poi al rigo successivo li sostituisce anche con $\hat{y}$.

Per quanto riguarda 1/n, questo ci aiuta a capire perché è rilevante minimizzare l'RSS. Nell'approccio Bayesiano il nostro obiettivo era la minimizzazione del BMSE (che è l'MSE con una B per simboleggiare che stiamo lavorando con l'approccio Bayesiano).

Il ragionamento è che il minimizzatore (o massimizzatore) di una funzione non cambia se moltiplichiamo per una costante, quindi nulla vieta di aggiungere quella 1/n, ricordando che l'MSE è la media della differenza al quadrato tra il valore vero e quello stimato, osserviamo che con l'aggiunta di 1/n abbiamo ottenuto la media campionaria della differenza al quadrato tra il valore vero e quello stimato.

In altre parole usando l'approccio OLS (Ordinary Least Squares) non facciamo altro che minimizzare NON l'MSE teorico ma invece quello campionario.

—fine dimostrazione—

Si può dimostrare che il valore minimizzante è:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

stime dei coefficienti per i minimi quadrati (per la regressione lineare semplice)

dove

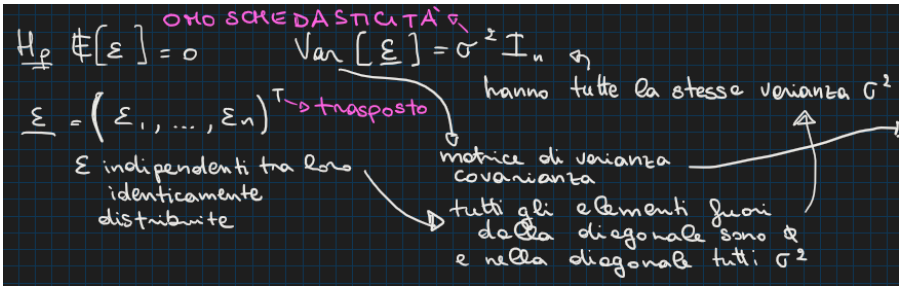
$$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

sono le medie campionarie.

La dimostrazione sarà presentata in seguito.

Ipotesi del procedimento in corso

Quali sono le ipotesi per il procedimento di regressione lineare che stiamo seguendo?



Ipotizziamo che la nostra ε ha valore medio uguale a 0.

Assumiamo inoltre che tutte le nostre ε (che possiamo mettere in un vettore) hanno la varianza che è uguale per tutte e uguale a σ^2.

Assumiamo anche le nostre ε sono tutte i.i.d.

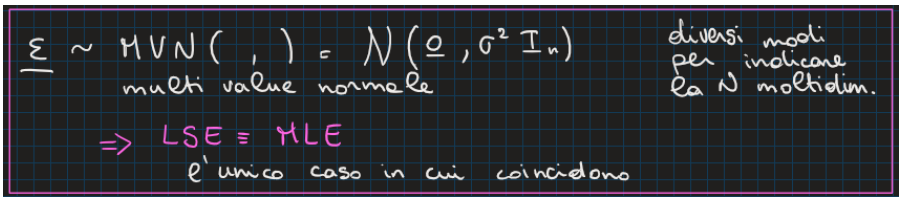
La varianza non esiste per un vettore, **var di un vettore è la matrice di varianza covarianza**, avrà sulla diagonale le varianze e gli altri valori covarianze.

Indipendenza implica covarianza 0 quindi tutti i valori della matrice sono 0 e la varianza è per tutti uguali, quindi possiamo scrivere la matrice come σ^2 per la matrice identità di dimensioni n.

> 💡 Varianza costante si chiama **omoschedasticità** (homoscedasticity), diverse varianze si chiama eteroschedasticità.

💡 Il miglior stimatore lineare possibile si può ottenere con OLS anche **senza il modello delle ε**.

E se avessimo anche il modello delle ε?



multi variate normal

Se vogliamo considerare il modello delle ε (per caso semplice Gaussiana) questo sarà una normale a più dimensioni, MVN (multi varied normal).

💡 Le stime LS corrispondono alle stime a massima verosimiglianza $LSE = MLE$ nel caso di normalità (Gaussiana). (è l'unico caso in cui succede questo).

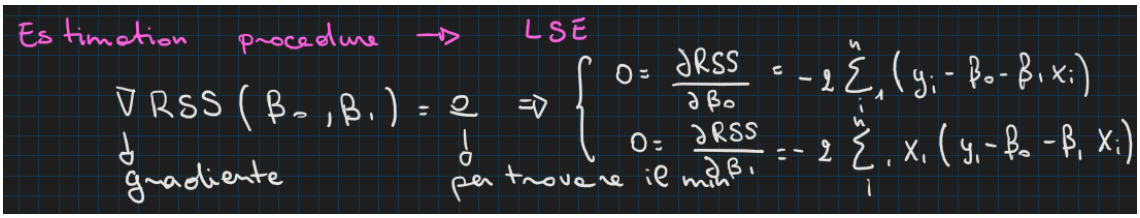
L'utilità di questa affermazione è che il MLE è il migliore, quindi ipotizzando la normalità possiamo usarlo, poi però bisogna tornare indietro a fare la diagnostica e verificare se le ipotesi fatte erano davvero attendibili.

Faremo la diagnostica del modello più avanti.

Dimostrazione formule per il calcolo di $\hat{\beta}_0$ e $\hat{\beta}_1$

Per ottenere le stime LSE dobbiamo trovare i valori che minimizzano l'RSS.

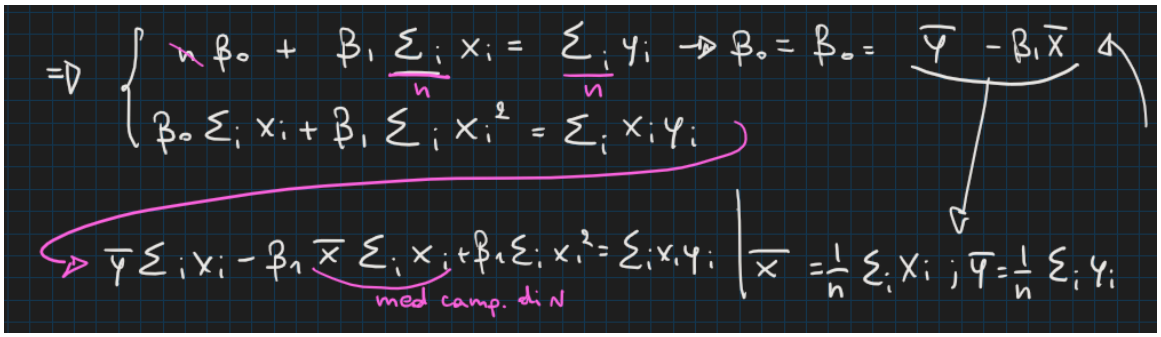
Il minimo di una funzione di due variabili si ottiene ponendo il gradiente uguale a 0.



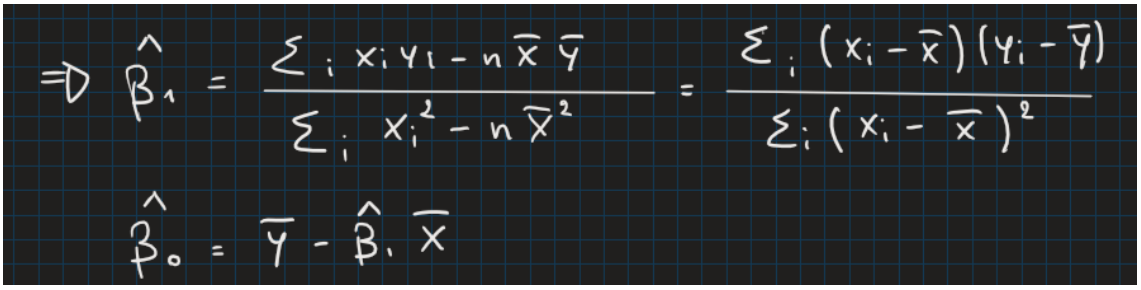
Il meno 2 viene dalla derivata del quadrato.

Il meno 2 si può togliere perché dall'altro lato ci sta 0.

Beta 0 è costante quindi si può far uscire dalla sommatoria e poi dividiamo tutto per n.



Le sommatorie sono da i = 1 a n anche se non c'è scritto.



A questo punto $\hat{\beta}_0$ sarà uguale alla differenza tra media campionaria di y meno media campionaria di x per $\hat{\beta}_1$.

▼ Passaggi per capire l'ultimo pezzo di $\hat{\beta}_1$

Espandiamo $(x_i - \bar{x})(y_i - \bar{y})$:

Partiamo da questa espressione:

$$(x_i - \bar{x})(y_i - \bar{y}).$$

Moltiplicando:

$$(x_i - \bar{x})(y_i - \bar{y}) = x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}.$$

Sommando per tutti gli i:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - \sum_i x_i \bar{y} - \sum_i \bar{x} y_i + \sum_i \bar{x}\bar{y}.$$

Analizziamo ciascun termine:

1. **Primo termine:** $\sum_i x_i y_i$ rimane invariato.
2. **Secondo termine:** $\sum_i x_i \bar{y}$:
   - $\bar{y}$ è costante rispetto alla somma, quindi:

$$\sum_i x_i \bar{y} = \bar{y} \sum_i x_i.$$

3. **Terzo termine:** $\sum_i \bar{x} y_i$:
   - $\bar{x}$ è costante, quindi:

$$\sum_i \bar{x} y_i = \bar{x} \sum_i y_i.$$

4. **Quarto termine:** $\sum_i \bar{x}\bar{y}$:
   - Sia $\bar{x}$ che $\bar{y}$ sono costanti, quindi:

$$\sum_i \bar{x}\bar{y} = n\bar{x}\bar{y}.$$

Uniamo i risultati:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - \bar{y} \sum_i x_i - \bar{x} \sum_i y_i + n\bar{x}\bar{y}.$$

Ricorda che:

- $\sum_i x_i = n\bar{x}$,
- $\sum_i y_i = n\bar{y}$.

Sostituendo questi risultati:

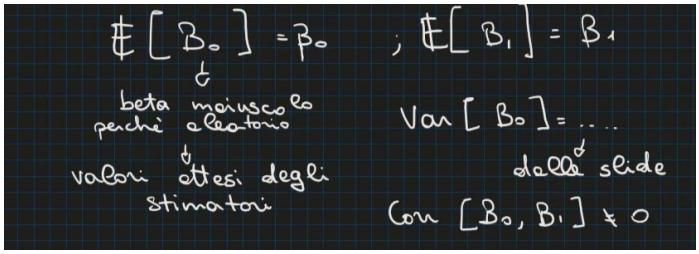
$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - \bar{y}(n\bar{x}) - \bar{x}(n\bar{y}) + n\bar{x}\bar{y}.$$

Osserva che i termini $-\bar{y}n\bar{x}$, $-\bar{x}n\bar{y}$ e $+n\bar{x}\bar{y}$ si combinano esattamente per formare $-n\bar{x}\bar{y}$.

Quindi:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y}.$$

Teorema di Gauss-Markov



Il valore atteso di B_0 (che si può calcolare perché in questo momento lo vediamo come stima, non come stimatore, quindi invece dei parametri "dentro" ci sono proprio le variabili aleatorie che lo rendono a sua volta una variabile aleatoria, per questo per essere precisi lo abbiamo scritto in maiuscolo) è proprio precisamente il vero β_0 cioè quello della retta di regressione della popolazione (population regression line).

Per essere precisi ricordiamo che β_0 vero è quello che noi sappiamo ci serve visto che vogliamo rappresentare la relazione tra Y ed X come lineare e che dobbiamo stimare per mezzo dei dati. Un oggetto del quale non conosco il valore ma devo stimarlo per mezzo dei dati prende il nome di "valore vero" o "stato di natura".

Visto che il valore atteso è proprio il valore vero allora lo stimatore $\hat{\beta}_0$ è unbiased.

Stesso discorso vale per lo stimatore di β_1, cioè la slope del modello di regressione lineare semplice.

Questa cosa di scrivere $E[B_0]$ invece di $E[\hat{\beta}_0]$ viene dai libri tradizionali di statistica ma è la stessa cosa.

Visto che B_0 e B_1 sono variabili aleatorie avranno una varianza, quella che nelle slide abbiamo chiamato standard error al quadrato.

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$\text{where } \sigma^2 = \mathrm{Var}(\epsilon)$$

richiamo dalle slide

Una cosa importante da ricordare è che la Correlazione tra B_0 e B_1 è diversa da 0, non sono incorrelati.

> 💡 **L'enunciato del Teorema di Gauss-Markov**
>
> Gli stimatori B_0 e B_1 così trovati rappresentano i migliori stimatori lineari tra quelli non polarizzati (unbiased).

Per il teorema di Gauss-Markov sono fondamentali tutte le ipotesi che abbiamo introdotto.

Qualunque altro stimatore lineare unbiased avrà varianza maggiore dello stimatore ottenuto in questo modo.

Lo stimatore unbiased della varianza

La qualità dello stimatore sicuramente è influenzata dalla varianza di ε, per una varianza di ε più piccola si hanno i valori che "ballano" di meno.

Ma noi stiamo stimando tutto, non sappiamo nulla di questo modello, quindi perché dovremmo sapere la varianza di ε_i?

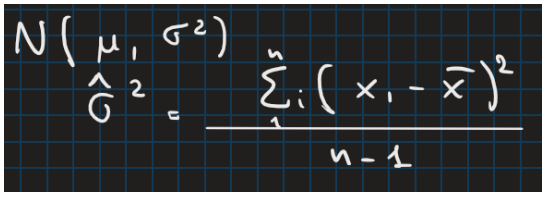
Potrebbe capitare in un caso raro che la conosciamo ma generalmente non è così, la varianza di ε dobbiamo ricavarla, anche questa va stimata a partire dai dati.

A σ^2, approccio plug-in, sostituisco una sua stima a partire dai dati, $\hat{\sigma}^2$.

Come la stimo σ^2?

Si usa un approccio che poi useremo anche per stimare le varianze delle Gaussiane in generale.

Vedremo che gli stimatori della varianza sono del tipo:



La dimostrazione è lunga ma n - 1 al denominatore invece di n perché bisogna sottrarre i gradi di libertà. Perché?

Essenzialmente perché quando calcolo lo stimatore di una varianza o in generale di qualcosa di quadratico, nelle forme quadratiche che sono al denominatore devo considerare quelle che sono indipendenti.

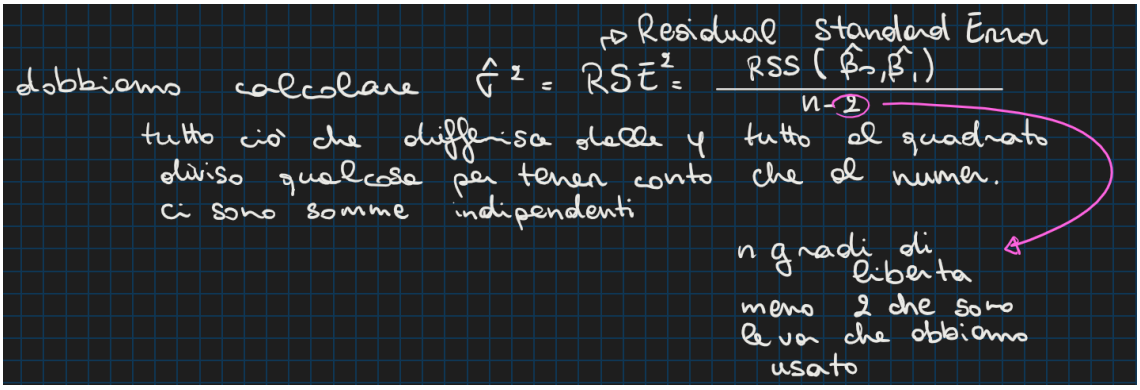
Quando si sommano insieme queste forme quadratiche bisogna sempre tenere a mente quante sono indipendenti, in questo caso n - 1 perché un grado di libertà è stato perso per la stima della media.

Più cose si stimano più forme quadratiche tolgo al denominatore.

Questo è lo stimatore unbiased della varianza.

Per calcolare la sigma quadro ci serve una misura di variabilità, e qual è nel nostro problema? Tutto ciò che differisce nelle y vere dalle $\hat{y}$ che ho appena calcolato e poi tutto al quadrato... questo è proprio l'RSS, solo che poi bisogna fare diviso un numero che tenga conto dei gradi di libertà lasciati da quello che abbiamo fatto.

La σ^2 è



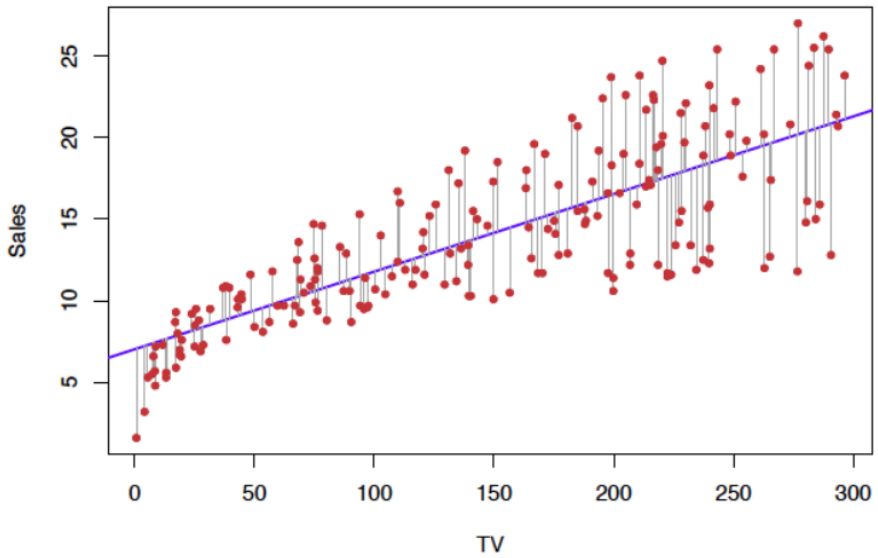
n-2 perché nel problema LS abbiamo fatto due stime, $\hat{\beta}_0$ e $\hat{\beta}_1$.

> 💡 Il **Residual Standard Error** è uguale alle somme dei quadrati, cioè l'RSS **calcolato nei punti di minimo** (un'altra ragione per la quale ci servono), diviso il numero di gradi di libertà che è n - 2.
>
> Questa varianza sarà tutta la variabilità che non siamo riusciti a spiegare con la retta di regressione.

Esempio

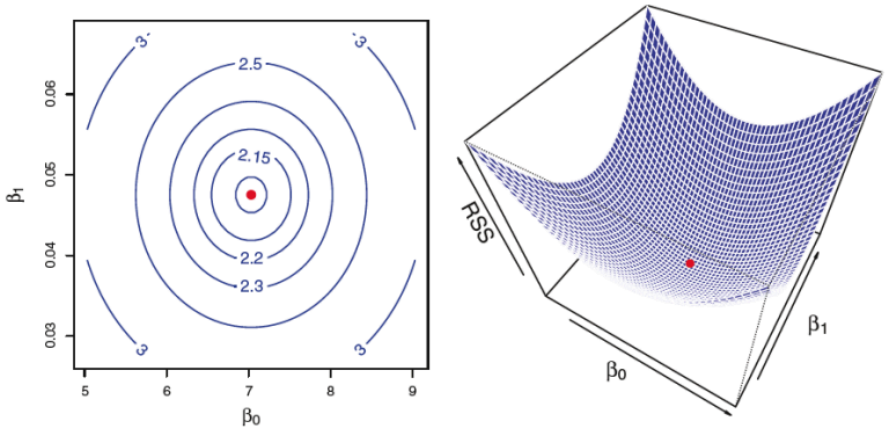
$\hat{\beta}_0 = 7.03$ e $\hat{\beta}_1 = 0.0475$

Un aumento di spesa di 1000 dollari per la pubblicità in TV sono associati con la vendita di circa 47.5 unità aggiuntive del prodotto.



Il fit least squares per la regressione di vendite su TV.

In questo caso un fit lineare cattura l'essenza della relazione, anche se in un modo che è leggermente carente nella parte sinistra del plot.



Contour Plot e Plot tridimensionale dell'RSS sui dati pubblicitari usando sales come response e TV come predictor. I punti rossi corrispondono alle stime per minimi quadrati $\hat{\beta}_0$ e $\hat{\beta}_1$.

Valori che minimizzano l'RSS: $\hat{\beta}_0 = 7.03$ e $\hat{\beta}_1 = 0.0475$

Verificare l'accuratezza delle stime dei coefficienti

Ricordiamo che la vera relazione tra X e Y è quella a sinistra, dove f è una funzione ignota e ε è un termine di errore casuale a media zero.

L'approssimazione lineare della vera relazione tra X e Y è quella a destra, dove β_0 è il termine intercetta, cioè il valore di Y quando X=0, mentre β_1 è la pendenza, cioè l'aumento medio di Y associato ad un aumento di una unità di X.

Il termine di errore è, tipicamente, ritenuto indipendente da X ed è fondamentale in quanto serve a considerare tutto ciò che non consideriamo con questo semplice modello: la relazione di natura è probabilmente non lineare, potrebbero esserci altre variabili che influenzano Y e inoltre c'è sempre errore di misurazione.

$$Y = f(X) + \epsilon \quad \Rightarrow \quad Y = \beta_0 + \beta_1 X + \epsilon.$$

$f(\,)$ linear function

> 💡 Il modello lineare che stiamo considerando definisce la **population regression line**, che è la migliore approssimazione lineare della vera relazione tra X e Y.

La coppia ottima $\hat{\beta}_0$ e $\hat{\beta}_1$ di stime di coefficienti per la regressione a minimi quadrati caratterizza la cosiddetta **least squares line (LS regression line)**.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$\widehat{\beta_0}$ and $\widehat{\beta_1}$ (B_0 and B_1 in some textbooks) are *unbiased* estimator of β_0 and β_1:
$E[\widehat{\beta_0}] = \beta_0$ and $E[\widehat{\beta_1}] = \beta_1$

9

La media degli stimatori di β_0 e β_1, calcolati su dataset diversi (o sotto-dataset) permette di ottenere i veri β_0 e β_1, perché gli stimatori least squares sono unbiased.

L'errore standard di uno stimatore riflette quanto cambia se sottoposto a ripetuti sampling (campionature).

Lo standard error degli stimatori di β_0 e β_1 è il seguente:

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$

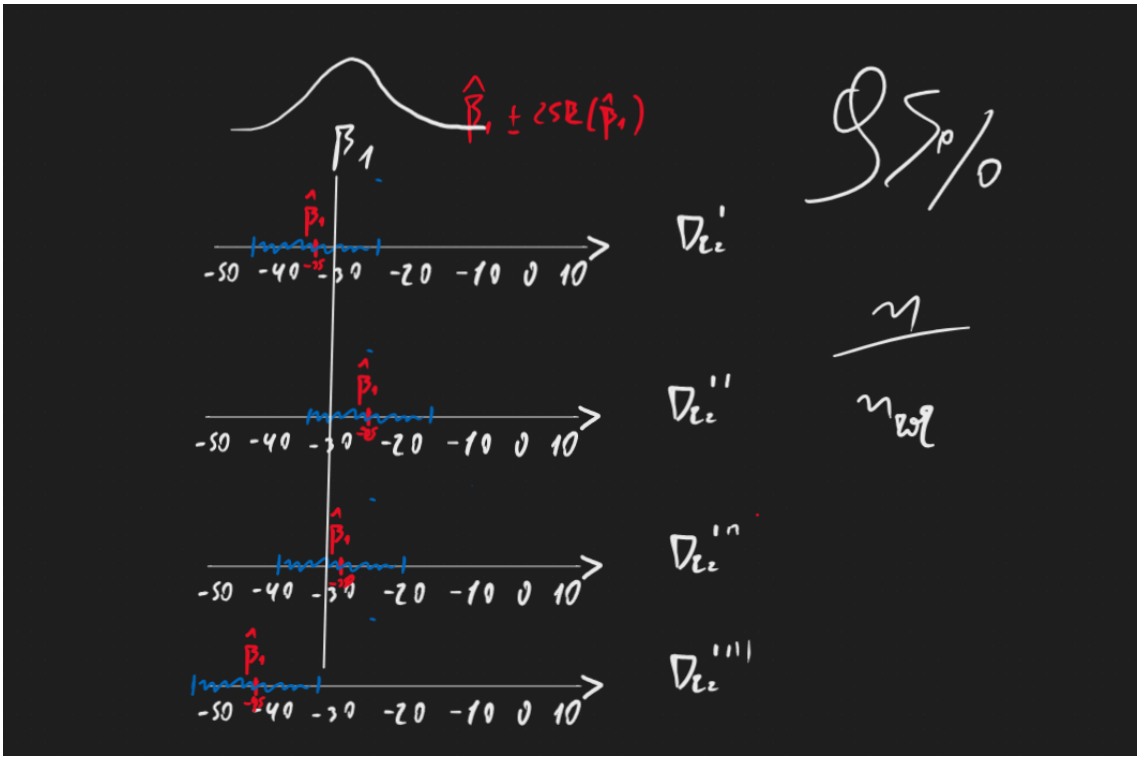
Lo $SE(\hat{\beta}_1)$ è più piccolo quando le x_i sono più sparpagliate.

La deviazione standard σ è stimata dal residual standard error (RSE) , errore standard residuo:

$$RSE = \sqrt{RSS/(n-2)}$$

Intervallo di Confidenza

Gli errori standard appena visti possono essere usati per calcolare gli intervalli di confidenza.



💡 Un intervallo di confidenza al 95% è definito come un range di valori tali per i quali quel range contiene il vero, e ignoto, valore del parametro con probabilità del 95%.

L'intervallo di confidenza ha questa forma:

$$\hat{\beta}_1 \pm \underbrace{2}_{\cong t_{1-\frac{\alpha}{2},\nu} \text{ for } \alpha = 0.05 \text{ and } \nu = n-2} \cdot \mathrm{SE}(\hat{\beta}_1).$$

Cioè c'è approssimativamente (per diverse motivazioni non approfondite) una probabilità del 95% che l'intervallo

$$[\hat{\beta}_1 - 2 * SE(\hat{\beta}_1), \hat{\beta}_1 + 2 * SE(\hat{\beta}_1)]$$

conterrà il vero valore di β_1 (nello scenario nel quale ci siano molteplici campioni ripetuti come nel caso presente).

Nell'esempio dei dati pubblicitari, l'intervallo di confidenza del 95% per $\hat{\beta}_0$ è [6.130, 7.935] mentre l'intervallo di confidenza del 95% per $\hat{\beta}_1$ è [0.042, 0.053].

✂️ **Precisazioni sull'intervallo per $\hat{\beta}_1$.**
L'intervallo calcolato per
$\hat{\beta}_1$ si basa sull'assunzione che gli errori siano Gaussiani. Inoltre, il fattore 2 che moltiplica $SE(\hat{\beta}_1)$ varierà leggermente sulla base del numero di osservazioni n nella regressione lineare. Ad essere precisi, piuttosto che il numero 2, l'equazione dovrebbe contenere il Quantile al 97.5% di una t-distribution con n-2 gradi di libertà.

Intervallo di confidenza - note

Alla base dell'idea di intervallo di confidenza c'è il concetto che più di conoscere un valore (in questo caso β_0 e β_1) è per noi di interesse sapere, con una certa probabilità, entro che intervallo quel numero si muoverà e inoltre è importante se sono vicini a 0 o significativamente diversi da 0.

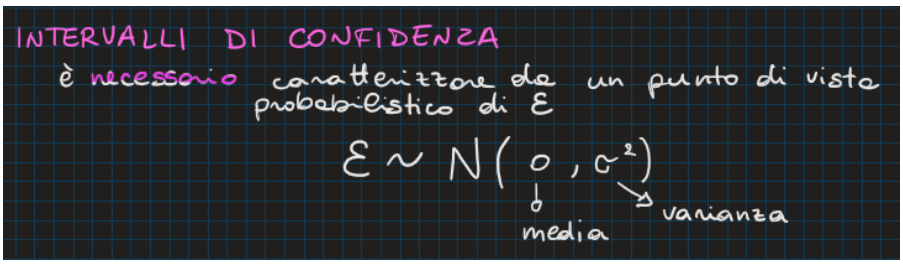
Prima che farlo con la regressione (più complicato) introduciamo il concetto con la stima della media di una variabile aleatoria con distribuzione normale per mezzo di raccolta di dati.

Ricordiamo che per la stima dei parametri della regressione lineare secondo least squares non era necessaria la conoscenza del modello probabilistico di ε, anche se ricordiamo che conoscendolo se era normale allora OLS = MLE.

Per il calcolo dell'Intervallo di Confidenza è necessario caratterizzare da un punto di vista probabilistico ε e quindi conseguentemente le X. Questo perché gli intervalli si vedono proprio sulla distribuzione di probabilità.

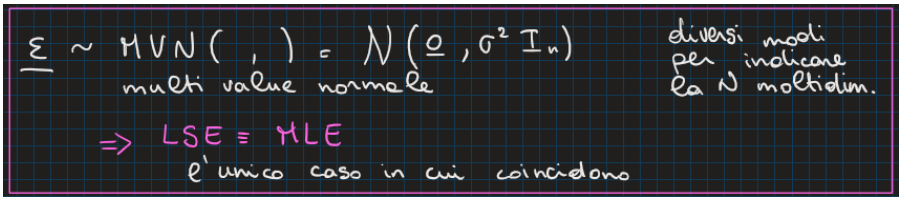
Detto questo assumiamo come modello probabilistico per ε la normale ma poi si può estendere ad altri casi.

La ε per noi sarà una normale a media 0 varianza σ^2.



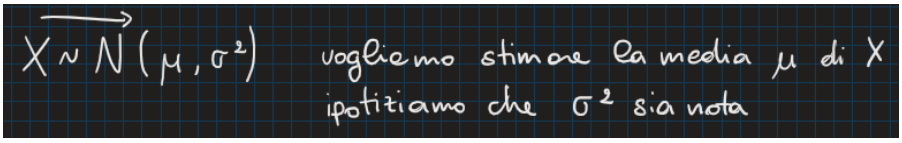
Questa è la ε della formula Y = f(X) + ε dove avevamo assunto f(X) lineare.

Ovviamente visto che abbiamo caratterizzato ε il vettore delle ϵ sarà MVN come visto recentemente.



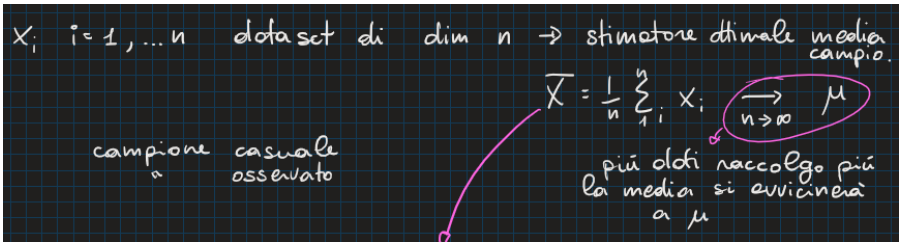
Calcolo intervalli di confidenza relativi alla stima della media una gaussiana

Iniziamo lavorando su una variabile aleatoria X (per mimare ciò che ci servirà per la regressione diciamo che anche X sarà una normale però a media μ e varianza sigma quadro (diverso dall'altro sigma quadro)).



Dobbiamo stimare **μ** e assumiamo σ^2 sia nota, iniziamo così che non è realistico ma poi ci torneremo per riuscire a stimare anche sigma quadro.

Vogliamo stimare la media **μ**, per farlo per prima caso raccogliamo campioni, un dataset di x_i di dimensione n.

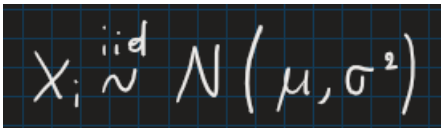


Ricordiamo che il miglior stimatore (è quello intuitivo e anche MLE) che si può costruire per la media è la media campionaria.

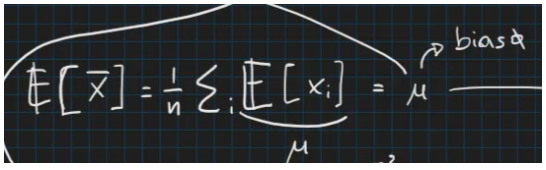
La media campionaria convergerà a **μ** per n che va ad infinito (la convergenza può essere o forte o debole, infatti ci sono due versioni della legge dei grandi numeri sulla base del tipo di convergenza, poi lo vedremo).

x_i è il campione osservato, X non osservato che è quindi una variabile aleatoria.

Il modello della X si chiama **modello di popolazione**, tutte le x_i seguiranno il modello della popolazione cioè X.



Le x_i saranno tutte uguali a quelle della popolazione e saranno tutte iid, del resto stiamo facendo estrazione indipendente quindi non c'è correlazione.

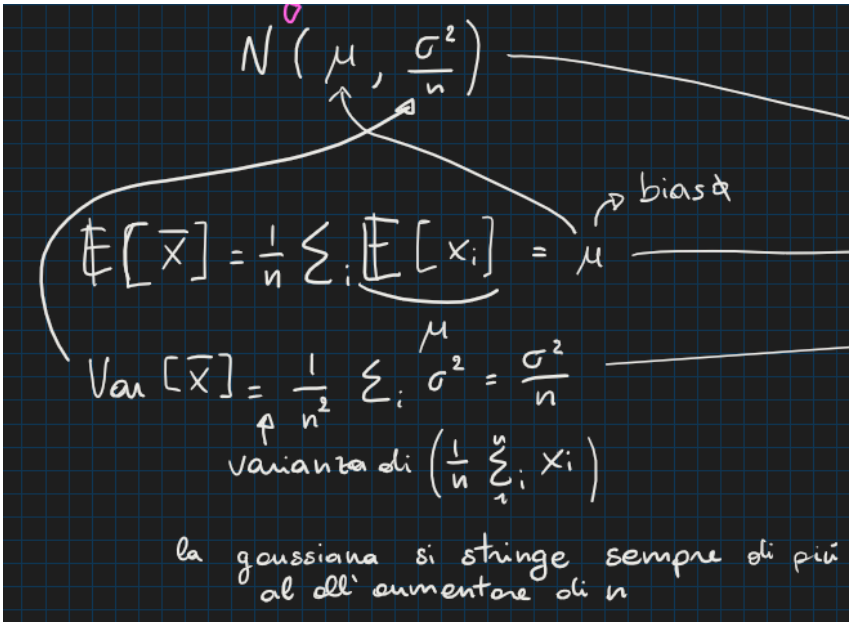


La media la calcoliamo e ne esce μ. Con ciò dimostriamo che il mio stimatore ha bias 0, lo sapevamo.

Ho un ottimo stimatore della media, ma qual è l'intervallo in cui si muove?

La combinazione lineare di x_i è combinazione lineare di Gaussiane che è una Gaussiana, che va da meno infinito a infinito, quindi se voglio un intervallo di confidenza che mi da certezza (probabilità 1) dovrei andare da meno infinito a più infinito, che è un risultato irrilevante, quindi **limito**, si parla spesso di intervalli di confidenza al 90% o al 95% ecc.

> 💡 Caratterizzo probabilisticamente la media campionaria ($\overline{x}$), sarà una Gaussiana come abbiamo detto, quindi ci servono media (già mostrato) e varianza di questa **nuova** gaussiana.



▼ Dimostrazione calcolo varianza

1. **Substitute the definition of $\bar{X}$**:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

2. **Apply the scaling property**:

$$\mathrm{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)$$

3. **Apply the sum of independent random variables property** (assuming X_i are independent and identically distributed):

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\mathrm{Var}(X)$$

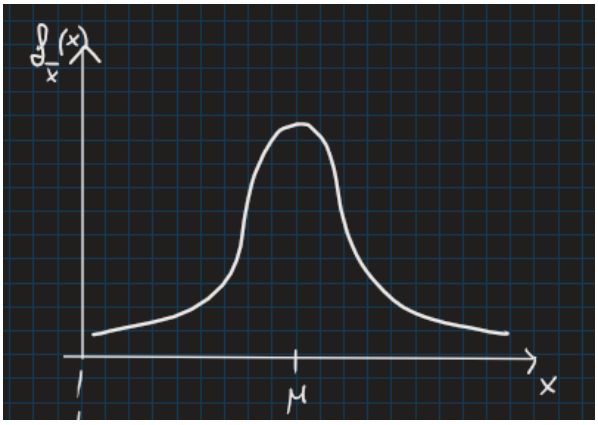
4. **Combine these results**:

$$\mathrm{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 (n\mathrm{Var}(X)) = \frac{\mathrm{Var}(X)}{n}$$

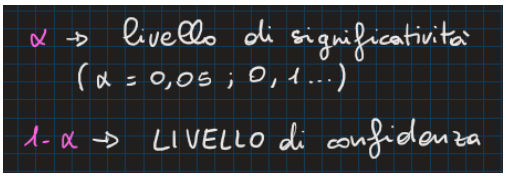
Ora la varianza, Var[$\overline{x}$].
Potremmo fare il percorso classico di calcoli o potremmo, meglio,
sfruttare le espressioni che già abbiamo, quindi scriviamo varianza di 1/n per la sommatoria delle x, che è la definizione della $\overline{x}$.

Le covarianze sono 0 quindi la varianza della somma sarà la somma delle varianze.

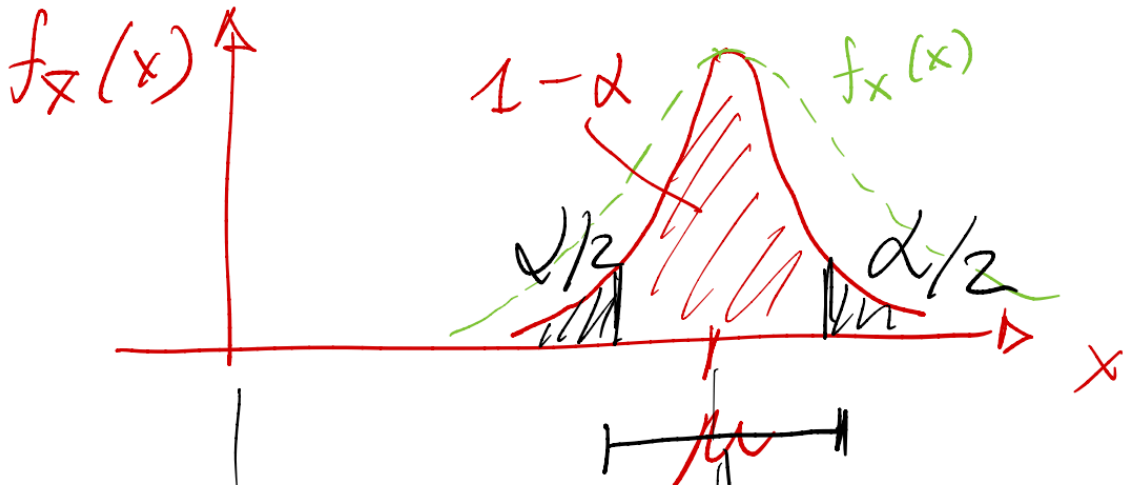
La varianza tende a 0 per n infinito, dove diventa una delta di dirac.



Adesso abbiamo la pdf, ora voglio trovare l'intervallo di confidenza, lo stabilisco a livello di 1 - α (90 per cento, 95 per cento ecc).

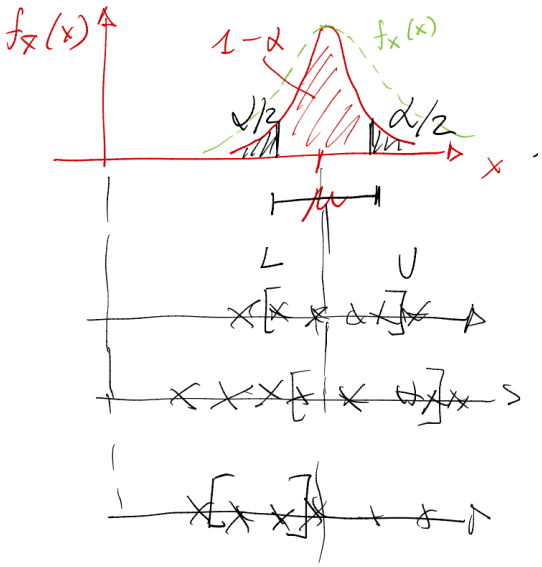


Quindi chiamo α il livello di significatività (significance level) e chiamo 1-α livello di confidenza (confidence level).



Fissiamo α e poi ci interessa 1-α, perché sarà il valore che corrisponde ai risultati corretti.

Dobbiamo calcolare 1-α a partire dalla distribuzione, cosa può succedere però?



Riportiamo i campioni come crocette sull'asse x.

Mi calcolo l'intervallo che contiene la media.

Poi faccio una seconda generazione che avrà punti diversi ma ci interessa sempre l'intervallo che contiene la media.

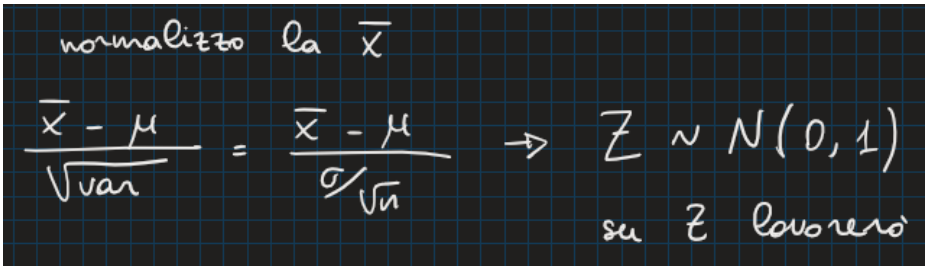
Per mera fluttuazione a partire dalla variabile aleatoria x succede al nostro terzo tentativo che non c'è la media nell'intervallo.

Abbiamo stabilito l'intervallo di confidenza intorno ai campioni che abbiamo scelto.

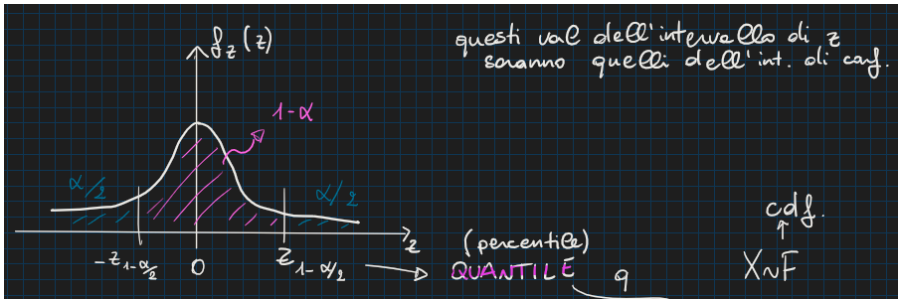
💡 Per tutto quello che abbiamo detto 1-α volte il nostro intervallo contiene il valore medio vero ma α volte non sarà così, errore, non si può avere una cosa infallibile perché per averla ci vorrebbe l'intervallo meno infinito infinito che è irrilevante.

Abbiamo detto che siamo nel caso di σ^2 noto, detto questo come faccio a calcolare l'intervallo di confidenza?

Per la Gaussiana i valori della CDF sono tabellari, non è risolta in maniera analitica, quindi come faccio a trovare questi valori di probabilità (i Quantili)? Standardizzo la $\overline{x}$ che ho trovato.



Su questa Z mi interessa trovare i livelli relativi al mio intervallo di interesse.



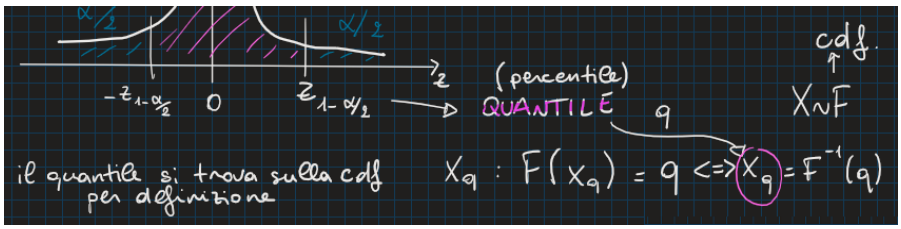
I livelli sono scelti in modo che nell'1- α per cento dei casi l'intervallo scelto contiene la media.

Le punte saranno tali che la somma delle due sarà pari ad α, il resto 1- α, del resto l'area sotto la pdf deve essere 1.

Non è detto che va scelto simmetrico l'intervallo visto che talvolta sbagliare in una direzione costa più che in un'altra, se questo non è il caso va bene che le due punte sono α/2 ciascuna.

Trovati questi intervalli mi interessano i valori della Z sotto a questi intervalli, cioè i valori della Z corrispondenti alla sua pdf.

Questi valori ci daranno gli intervalli "limite inferiore" e "limite superiore".



I punti che delimitano i margini tra l'area giusto e l'area di errore sono chiamati Quantile di 1-(α/2).

Il Quantile

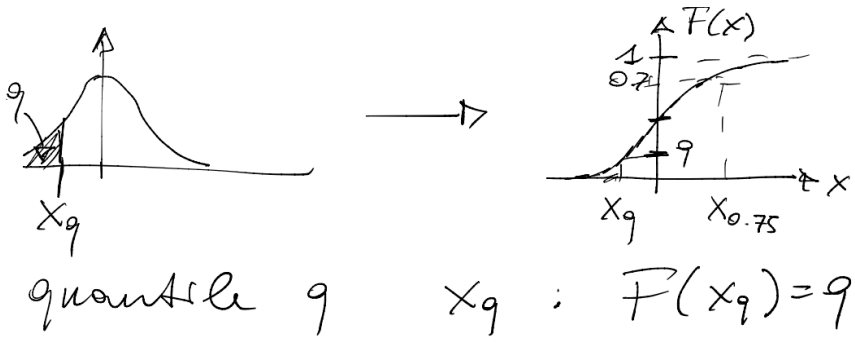
> In termini di **CDF** (Funzione di Distribuzione Cumulativa), un **quantile** corrisponde al valore x_q per cui la CDF assume un determinato valore q. In altre parole, il quantile x_q è il valore tale che:
>
> $$$$
>
> dove $F(x_q) = q$ è la CDF e q è la probabilità associata al quantile (espressa come numero tra 0 e 1). Ad esempio, il 25° percentile è il valore x per cui F(x)=0.25, ossia la probabilità che un valore casuale preso dalla distribuzione sia minore o uguale a x è 0.25.

Percentile e **quantile** sono concetti strettamente correlati, ma non esattamente la stessa cosa:

- Un **quantile** è un termine più generico che si riferisce a qualsiasi valore che divide una distribuzione in intervalli uguali. Ad esempio, i quartili dividono la distribuzione in quattro parti uguali, i decili in dieci parti, e così via.
- Un **percentile** è un tipo specifico di quantile che divide la distribuzione in 100 parti uguali. Il **p-esimo percentile** è il valore al di sotto del quale cade il p% dei dati.

Quindi, ogni percentile è un quantile, ma non ogni quantile è un percentile.

A noi interessa il Quantile 1 - (α/2) perché a noi interessa arrivare da meno infinito a 1 - (α/2), perché α/2 è il residuo (la coda a destra).



Leggere il quantile sulla CDF è come vedere l'area della pdf.

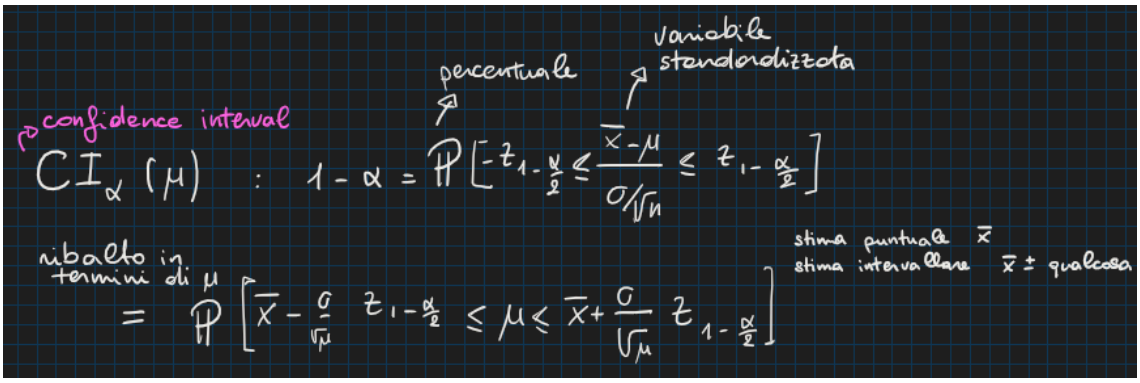
Il Quantile è il numero che corrisponde ad un certo livello di probabilità.

Il valore che nella pdf lascia fuori dall'integrale solo una certa quantità.

Ricerca del Confidence Interval

Il problema iniziale era la ricerca del Confidence Interval fissato α (livello di significatività) sullo stimatore della μ.

Lo ricaviamo a partire dalla Z, chiedendoci per prima cosa cosa sarà 1 - α.



è probabilità non percentuale

1-α è la Probabilità che la nostra distribuzione si trovi all'interno dell'intervallo corretto.

Ribaltando trovo l'intervallo sulla μ.

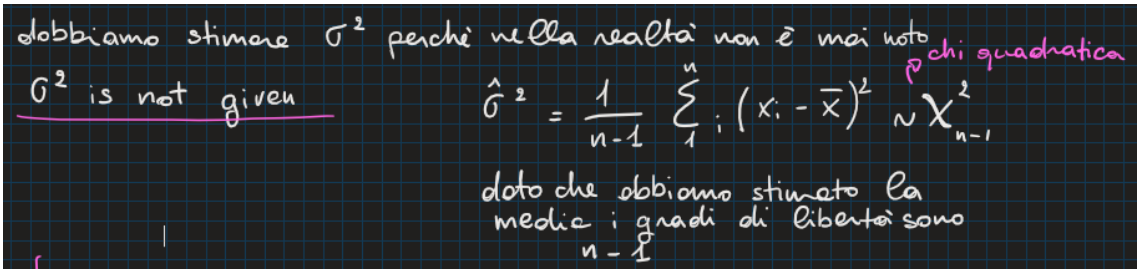
$$CI_{\alpha}(\mu) : [L, U]$$

Il termine additivo che rende la stima intervallare per n che va a infinito va a 0 e si capisce perché lo stimatore asintoticamente tende a μ.

α per cento delle volte il campione sarà fuori da quello che ci aspettiamo ma va bene perché altrimenti potevamo solo dire meno infinito infinito come intervallo.

Stimare σ^2

Inizialmente abbiamo detto che σ^2 era noto, visto che questo è improbabile ora vediamo come stimarlo.



Questo è **lo stimatore non polarizzato per la varianza**. Lo otteniamo a partire dai dati.

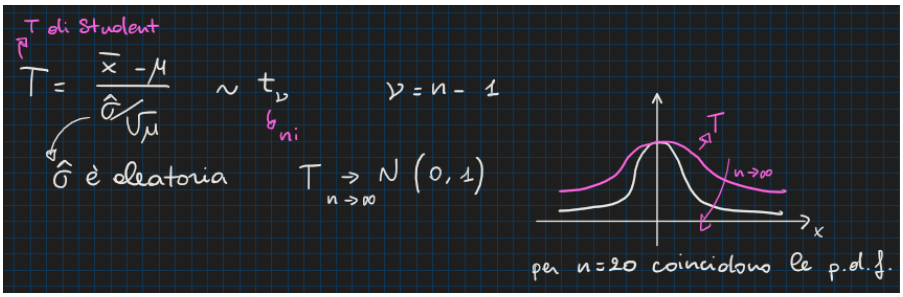
Si può dimostrare che questo stimatore è distribuito su una particolare distribuzione, la distribuzione che è i quadrati di n gaussiane e che prende il nome di "chi squared" χ^2, dove l'unica informazione da aggiungere è il numero di quadrati indipendenti. (Chi è una lettera greca, assomiglia alla x.) Ovviamente affinché questo valga è necessario che in partenza sia tutto Gaussiano.

A questo punto abbiamo ottenuto $\hat{\sigma}^2$.

T di student

Prima l'avevamo usata per la normalizzazione della Gaussiana, ora lì ci andrà lo stimatore.

Avere una variabile aleatoria a denominatore **cambia la distribuzione**. Viene quindi generato un nuovo modello di variabile aleatoria.



A numeratore abbiamo una normale e a denominatore la radice quadrata di una χ^2 ("chi" squared).

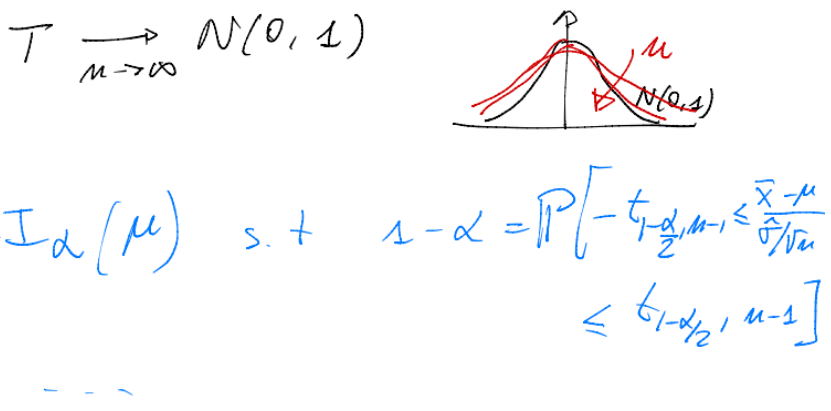
Questo rapporto produce una speciale distribuzione chiamata **t di student**.

La T di student è identificata con t con un unico parametro "ni" (v) che è il numero di gradi di libertà, in questo caso n-1.

T per n che tende all'infinito tende ad una normale standard.

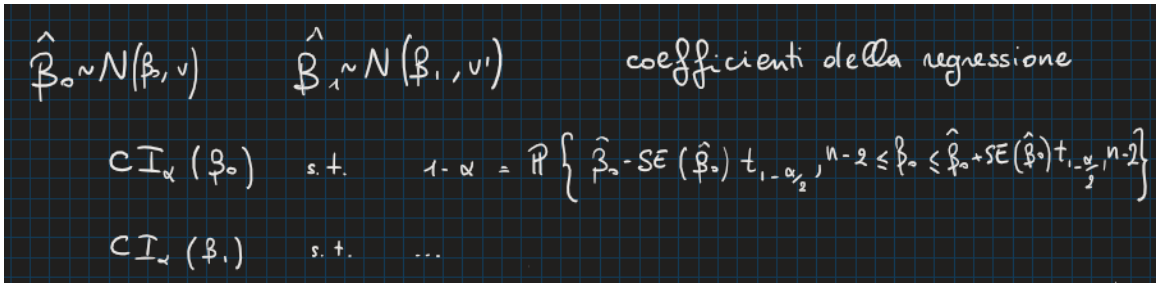
In realtà già per n uguale a 20 o 30 è uguale ad una gaussiana, i quantili di t di student e normale standard diventano uguali fino alla seconda, terza cifra decimale.

Riguardo i nostri intervalli di confidenza concludiamo:



Abbiamo quindi trovato gli intervalli di confidenza, ora trasferiamo tutte le conclusioni raggiunte a quanto riguarda $\hat{\beta}_0$ e $\hat{\beta}_1$, i coefficienti della nostra regressione.

Torniamo agli stimatori $\hat{\beta}_0$ e $\hat{\beta}_1$



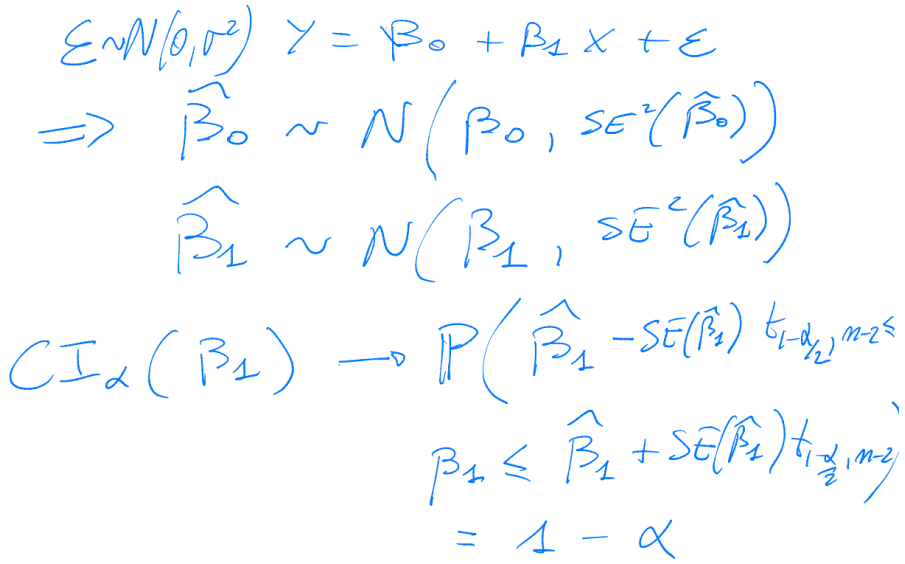
> 💡 Con ipotesi di normalità di ε (errore) anche gli stimatori sono normali.

Gli stimatori Gaussiani per Gauss-Markov sappiamo che sono non polarizzati.

σ^2 non è noto ma la cosa non ci preoccupa, invece di avere Z avremo la T di student, che per n 20 o più comunque è quasi normale.

Se stimiamo σ^2 dobbiamo prendere il quantile della T che è vicinissimo a quello della Z (che sarebbe quello da considerare se avessimo il vero sigma quadro) quindi va bene comunque.

I gradi di libertà per la stima di σ sono gli stessi che dobbiamo mettere nella T.



Hypothesis testing

Lo standard error può essere utilizzato per performare l'hypothesis testing sui coefficienti. Il tipo più comune di hypothesis test riguarda testare

l'ipotesi nulla:

- H_0 → non c'è relazione tra X ed Y;

e l'ipotesi alternativa:

- H_A → c'è una qualche relazione tra X ed Y.

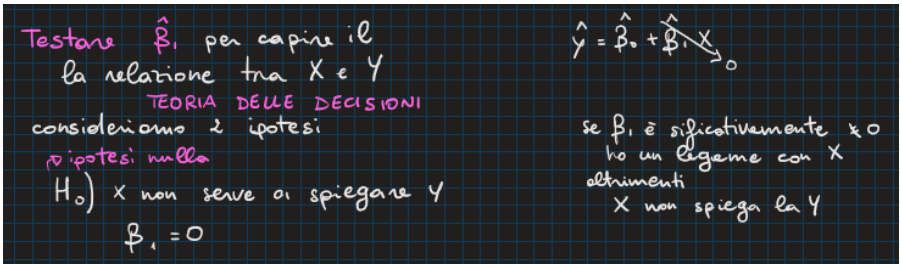
Matematicamente questo corrisponde a testare

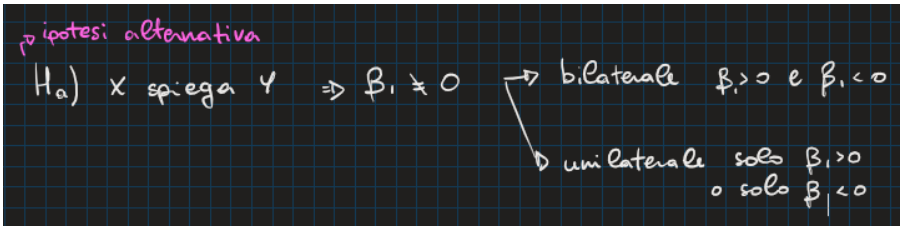
$$H_0 : \beta_1 = 0$$

contro

$$H_1 : \beta_1 \neq 0$$

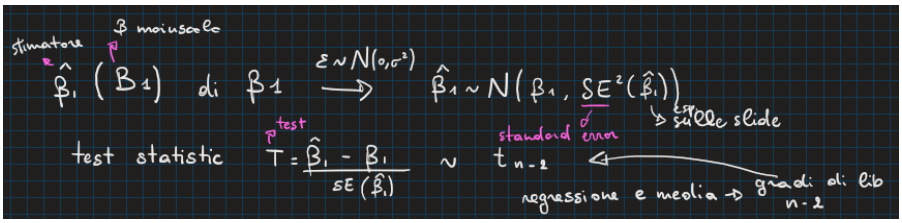
visto che se $\beta_1 = 0$ allora il modello si riduce a $Y = \beta_0 + \varepsilon$ e X non è associata ad Y.





▼ Possibile spiegazione di unilaterale e bilaterale

In un test unilaterale la zona di rifiuto dell'ipotesi nulla è solamente in una coda della distribuzione; in un test bilaterale essa è equamente divisa nelle due code della distribuzione.



La statistica di decisione (test statistic) ci permette di valutare la nostra ipotesi.

Voglio usare $\hat{\beta}_1$ come mia statistica di decisione, ma non è normale standard quindi la standardizzo, in questo modo ottengo approssimativamente una t di student con n-2 gradi di libertà perché ho stimato β_0 e β_1.

▼ Possibile spiegazione della test statistic

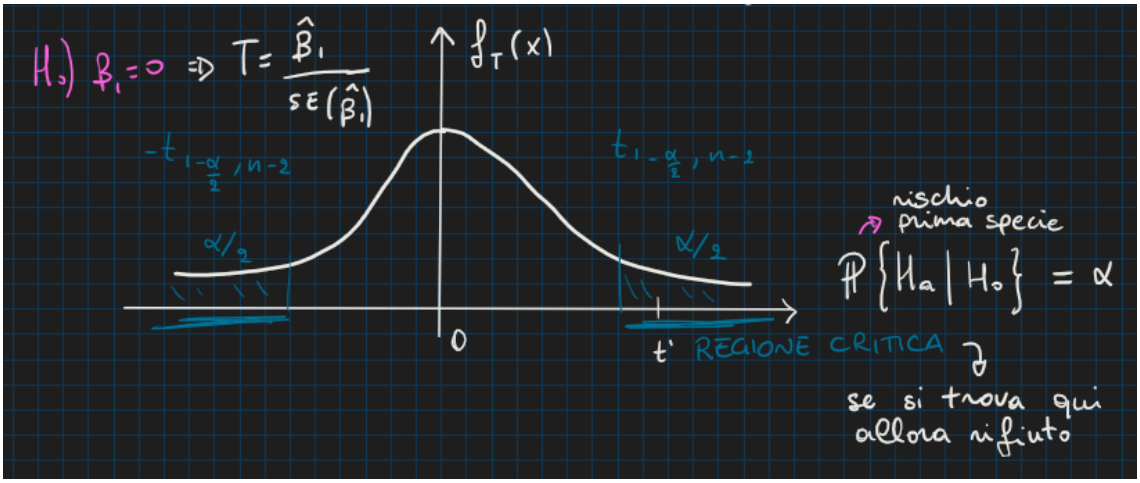
La **test statistic** (statistica del test) è una misura calcolata a partire dai dati osservati che viene utilizzata per prendere una decisione su un'ipotesi statistica. Nel contesto della **regressione lineare**, la statistica del test è utilizzata per verificare se i coefficienti stimati della regressione siano significativamente diversi da zero, ovvero se esista una relazione tra la variabile indipendente e quella dipendente.

Nella regressione lineare, dopo aver stimato i coefficienti del modello tramite il metodo dei minimi quadrati ordinari (OLS), è importante capire se questi coefficienti sono statisticamente significativi. Per ogni coefficiente, possiamo calcolare una statistica del test basata sulla **distribuzione t di Student**.

Supponiamo vera H_0

Nella T $\hat{\beta}_1 - \beta_1$ è vicina allo zero, ma noi supponiamo di essere in una ipotesi specifica, supponiamo che $\beta_1 = 0$, cioè supponiamo vera l'ipotesi nulla.

La curva a campana con $\beta_1 = 0$ rende molto probabile che $\hat{\beta}_1 = 0$ ma si deve comunque ammettere un errore perché per semplice fluttuazione del valore di $\hat{\beta}_1$ questo talvolta non sarà 0.



Ci interessa fissare quello che definiamo **Rischio di Prima Specie**, cioè la probabilità che scelgo H_a ma è vero H_0. Questa probabilità la chiamo α (il Rischio di Seconda Specie sarà β).

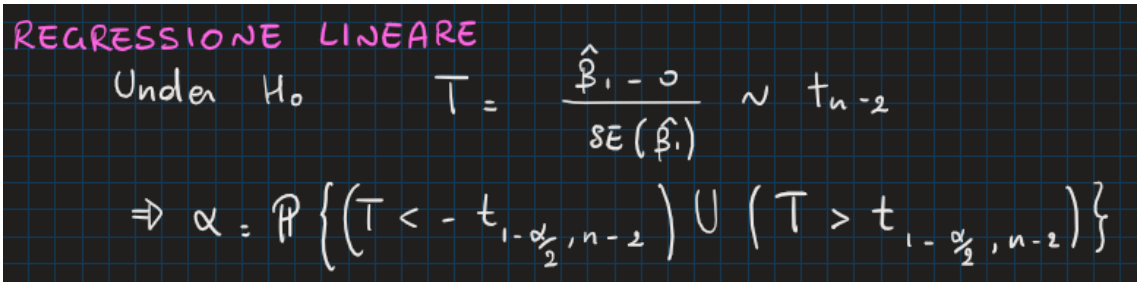
β_1 vale 0, noi abbiamo una variabile aleatoria che se estratta da un valore molto vicino a β_1 cioè 0, ma talvolta per mera fluttuazione non sarà così. Il Rischio di prima specie è la probabilità che una estrazione da un valore molto lontano da 0 nonostante β_1 = 0.

La Regione Critica è divisa in due code che rappresentano probabilità α/2.

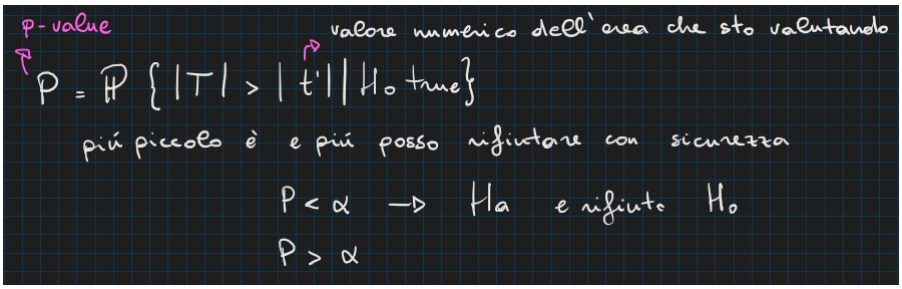
Per trovare le regioni si usano i quantili $1 - \alpha/2$ della t di Student.

Il **quantile** rappresenta un punto specifico sull'asse delle variabili (nel nostro caso, sull'asse x), non una probabilità. Il quantile al 95% indica il valore sull'asse x sotto il quale si trova il 95% della probabilità cumulativa (l'area sotto la curva fino a quel punto).

Tornando alla Regressione Lineare, se è vera l'ipotesi nulla abbiamo concluso che la statistica di test sarà una t di Student con $n - 2$ gradi di libertà e possiamo calcolare la probabilità di α.



Quando facciamo una ipotesi è importante capire con quanta sicurezza possiamo rigettare una certa ipotesi, si chiama forza di rifiuto.



Chiamiamo p-value la probabilità di osservare un qualsiasi numero uguale a |t'| (dove t' è la test statistic) o maggiore in valore assoluto, quando $\beta_1 = 0$. Possiamo interpretare un p-value piccolo come il fatto che è improbabile osservare una associazione sostanziale tra predittore e risposta per puro caso, in assenza di una reale associazione tra predittore e risposta. In altre parole se il p-value è piccolo rigettiamo la null hypothesis perché riconosciamo che una relazione tra X ed Y esiste.

t' è il valore della statistica calcolata sulle code.

Chiameremo P-value la probabilità dell'evento che ci troviamo nelle code, cioè che la mia statistica è maggiore di quella appena calcolata.

Testiamo la null hypothesis

Per testare la null hypothesis calcoliamo una t-statistic, data da:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Questa avrà una t-distribution con $n - 2$ gradi di libertà assumendo $\beta_1 = 0$.

Usando software statistici è facile comparare la probabilità di osservare un qualsiasi valore uguale a |t| o più grande. Questo è chiamato **p-value**.

> 💡 Più specificatamente, la t-distribution ha una forma a campana e per valori di n più grandi di approssimatamente 30 è molto simile alla distribuzione normale.

Risulta abbastanza semplice calcolare il p-value, cioè la probabilità di osservare un qualsiasi numero uguale a |t| o maggiore in valore assoluto, assumendo $\beta_1 = 0$.

Se osserviamo un p-value piccolo, possiamo dedurre che c'è una associazione tra il predittore e la risposta.

Nello specifico il p-value è interpretato nel seguente modo:

> 💡 Un p-value piccolo indica che è improbabile osservare tale associazione tra il predittore e la risposta per caso, se non c'è una vera associazione.

Rigettiamo l'ipotesi nulla (dichiariamo quindi che **una relazione tra X ed Y esiste**) se il p-value è più piccolo di un valore cutoff (di separazione) α = Pr{reject H_0 | H_0 true}, questo valore è anche detto livello di significanza (significance level).

Tipici valori per α sono 5% o 1%. Quando n = 30 questi valori corrispondono rispettivamente a t-statistics di circa 2 e 2.75, in test bilaterali.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

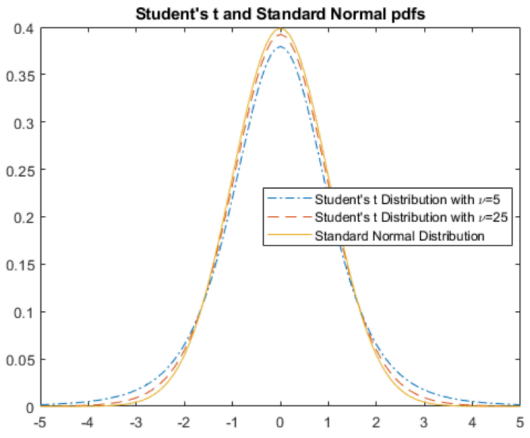
I valori stimati dei coefficienti β_0 e β_1 sono molto grandi rispetto ai loro Standard Error, quindi possiamo concludere che anche le t-statistics sono grandi: $t > t_{1-\frac{\alpha}{2},v}$, con $v = n - 2$, che sono i gradi di libertà.

Detto questo, la probabilità di osservare questi valori se H_0 è vero sono virtualmente 0.

Possiamo quindi concludere che β_0 e β_1 sono diversi da 0.

> 💡 Un piccolo p-value per la intercetta (β_0) indica che possiamo rigettare l'ipotesi nulla che $\beta_0 = 0$ ed un piccolo p-value per TV indica che dobbiamo rigettare la null hypothesis che $\beta_1 = 0$.
>
> Queste due conclusioni hanno implicazioni diverse.
>
> - Rigettare la null hypothesis che $\beta_0 = 0$ implica che in assenza di spese pubblicitarie in TV le vendite NON sono 0.
> - Rigettare la nulla hypothesis per TV implica che c'è una relazione tra TV e vendite.

La t di student e l'hypothesis testing



La t di student fu introdotta da W.S. Gosset sotto lo pseudonimo Student nei primi anni del ventesimo secolo.

$$y = f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}}$$

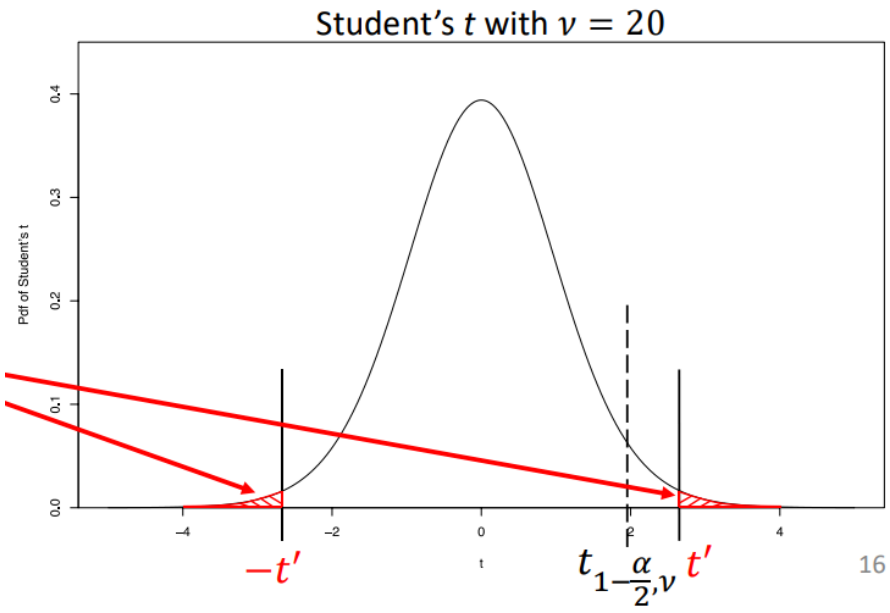
dove $\Gamma(\cdot)$ è la funzione gamma di Eulero (completa):

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t$$

Il p-value si calcola:

$$p = \{|t| > |t'| \mid H_0\}$$

Graficamente il p-value è la somma delle due aree indicate nel seguente grafico.



16

Ultima precisazione

> 💡 Equivalentemente, la stessa conclusione raggiunta poteva anche essere derivata calcolando gli intervalli di confidenza $(1-\alpha)$ per β_0 e β_1 e notando che entrambi NON includono 0.

Ad esempio i (approssimati) 95% confidence intervals sono

(con $t_{1-\frac{0.05}{2},198} \approx z_{1-\frac{0.05}{2}} \approx 1.96$, assumendo che n sia grande abbastanza):

- per β_0: [7.035 - 1.96*0.4578, 7.035 + 1.96*0.4578] → [6.117, 7.930];
- per β_1: [0.0475 - 1.96*0.0027, 0.0475 + 1.96*0.0027] → [0.0422, 0.0528];
- nessuno dei due intervalli include 0.

Confermiamo quindi che sia β_0 che β_1 sono significativamente maggiori di 0.

Stabilire l'accuratezza del modello

Una volta rigettata la null hypothesis in favore dell'ipotesi alternativa, è tipicamente desiderabile quantificare come il modello si adatta (fit) ai dati.

La qualità del fit della regressione lineare è tipicamente stabilito usando due quantità tra loro legate: il **Residual Standard Error** (RSE, errore standard residuo) e la R^2 **statistic**.

Ricordando che dal modello associata ad ogni osservazione vi è un termine di errore ε, **a** causa della presenza di termini di errore anche se conoscessimo la vera population regression line (β_0 e β_1 noti) non potremmo predire perfettamente Y a partire da X.

Residual Standard Error

Il Residual Standard Error si calcola:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

dove la residual sum-of squares (somma residua dei quadrati) è:

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

> 💡 L'RSE è una stima della deviazione standard di ε.

💡 L'RSE è considerato una misura della mancanza di fit del modello ai dati.

- Se le previsioni ottenute usando il modello sono molto vicine ai veri valori degli outcomes, l'RSE è piccolo e possiamo concludere che il modello si adatta (fit) molto bene ai dati.
- D'altro canto, se le previsioni del modello sono molto lontane dai veri valori per una o più osservazioni, allora l'RSE potrebbe risultare grande, indicando che il modello non si adatta (fit) bene ai dati.

R^2 statistic

L'RSE fornisce una misura assoluta della mancanza di fit del modello rispetto ai dati, ma visto che è misurato in unita di Y non è sempre chiaro cosa costituisca un buon RSE o meno.

💡 La R^2 statistic fornisce una misura alternativa del fit, questa rappresenta la proporzione di varianza spiegata (tra 0 ed 1) ed è indipendente dalla scala di Y.

La $R-squared$ o frazione di varianza spiegata è

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

dove TSS è la somma totale dei quadrati e si calcola:

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

- Il TSS misura la varianza totale nella risposta Y e rappresenta la quantità di variabilità intrinseca nella risposta prima che sia effettuata la regressione.
- In contrasto l'RSS misura la quantità di variabilità che rimane inspiegata dopo aver performato la regressione.

💡 Di conseguenza TSS - RSS misura la quantità di variabilità nella response che è “spiegata” (o rimossa) effettuando la regressione, ed R^2 misura la proporzione di variabilità in Y che può essere cambiata usando X.

- una R^2 statistic che è vicina ad 1 indica che una grande proporzione della variabilità nella risposta è stata spiegata dalla regressione;
- una R^2 statistic che è vicina a 0 indica che la regressione non ha spiegato molto della variabilità nella risposta; questo può succedere perché il modello lineare è sbagliato o perché l’errore σ^2 intrinseco è alto o per entrambi.

Correlazione

Può essere dimostrato che nel nostro caso di Simple Linear Regression vale che $R^2 = r^2$, dove r è la correlazione tra X ed Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

dove $r = \hat{Cor}(X, Y)$ è una misura della **relazione lineare** tra X ed Y.

La statistica R^2 estende il concetto di correlazione tra multipli predittori e la riposta, come nei problemi di Multiple Linear Regression (dove sono usati molti predittori).

La Correlazione quantifica l’associazione tra una singola coppia di variabili invece che tra un numero più grande di variabili.

ESEMPIO

Nell’esempio dei dati pubblicitari:

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1