Analisi dei dati

Regressione

Domenico De Stefano

a.a. 2021/2022

Indice

Che cos'è un modello statistico?

- Modello: Rappresentazione semplificata, analogica e necessaria della realtà
  - Semplificazione della realtà: il modello di un aereoplano, del flusso finanziario di un Paese ottenuti riproducendo gli aspetti "essenziali" e eliminando quelli ritenuti "superficiali"
  - Analogia della realtà: il modello è una riproduzione della realtà
  - Rappresentazione necessaria della realtà: anche se semplificato il modello è necessario per capire la realtà tramite lo studio di relazioni semplici e di maggiore intellegibilità

Che cos'è un modello statistico? (2)

- Modello statistico: modello di tipo matematico con
  - una componente deterministica
  - una componente casuale (o aleatoria o stocastica)
- La specificazione di un modello statistico (la sua forma "teorica") consiste nell'esplicitare un **legame** tra le variabili di interesse (che sia in linea con quanto osservato dai dati):

$$Y = f(X_1, X_2, \ldots, X_k)$$

Dove Y è la variabile da spiegare, mentre $X_1, X_2, \ldots, X_k$ sono le variabili scelte per spiegare Y tramite la funzione $f(.)$

- Inoltre non è quasi mai plausibile ipotizzare un legame deterministico quindi dobbiamo aggiungere un errore:

Che cos'è un modello statistico? (3)

$$Y = f(X_1, X_2, \ldots, X_k) + \epsilon$$

Dove ϵ è una variabile casuale e riassume la nostra ignoranza circa la vera relazione tra Y e X. Per questo motivo la chiameremo variabile errore.

Terminologia

$Y = f(X_1, X_2, \ldots, X_k) \rightarrow$ Modello statistico
$Y \rightarrow$ Variabile dipendente (o variabile di risposta)
$X_1, X_2, \ldots, X_k \rightarrow$ Variabile indipendenti (o variabile esplicative)
$\epsilon \rightarrow$ Variabile casuale errore

Nota: Il legame statistico implicato dal modello non è simmetrico. Sono le variabili esplicative a "determinare" la variabile dipendente e NON viceversa

Ad es.: se la X sono gli anni di formazione e la Y è il reddito di una persona è la Y che dipende dalla X e non viceversa!

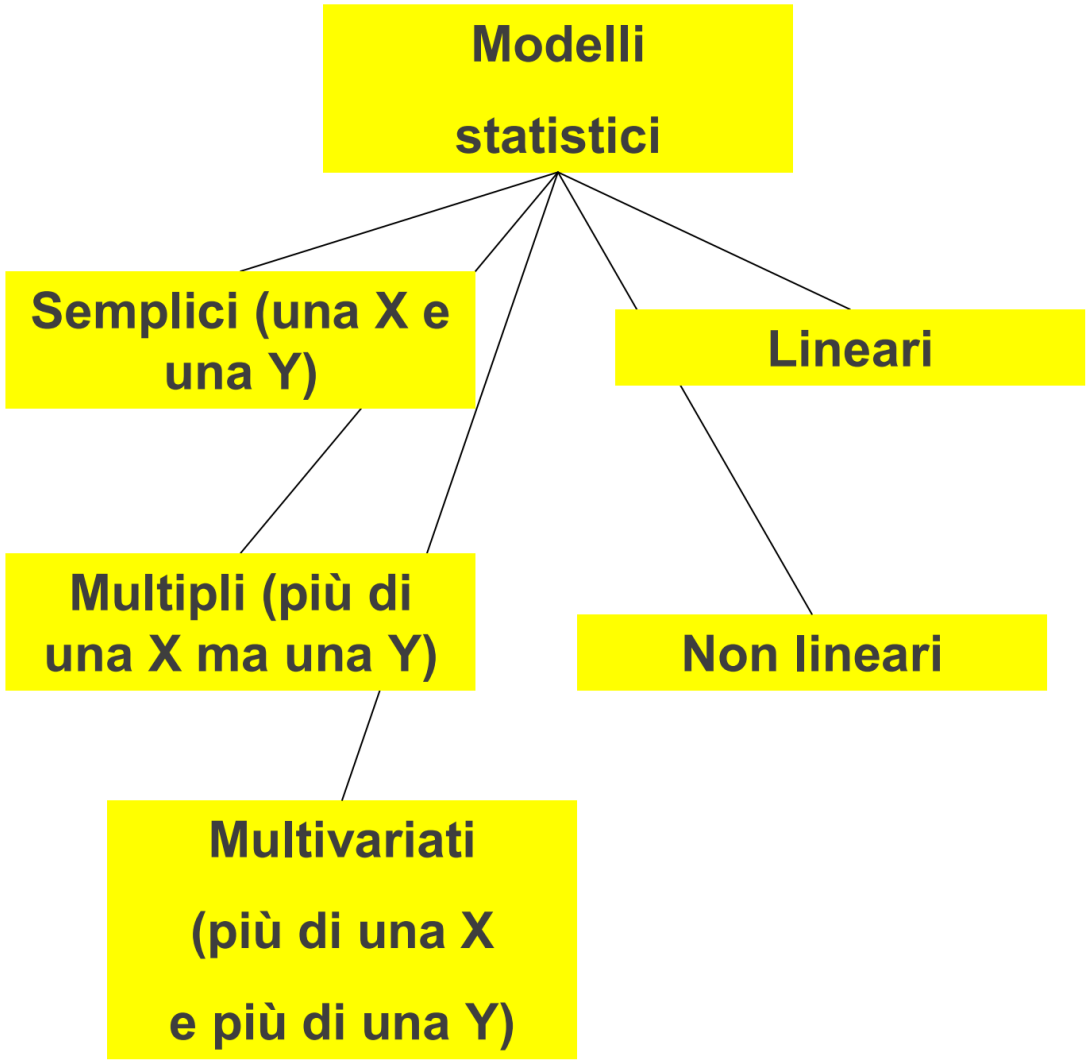
Semplici esempi di modelli statistici

In alcuni contesti la specificazione della relazione funzionale che lega la Y alle X risulta immediata dalla natura del problema:

1. e Y è il peso ed X è l'altezza di una persona adulta la prima relazione che viene in mente da specificare è quella proporzionale (maggiore il peso, maggiore l'altezza e viceversa) $Y = \beta X + \epsilon$
2. Se Y è il peso di una mattonella rettangolare per la quale X_1 e X_2 sono rispettivamente la lunghezza e la larghezza, allora una relazione funzionale può essere specificata mediante $Y = \beta X_1 X_2 + \epsilon$

Entrambe le specificazioni evidenziano un parametro β che deve essere determinato (stimato) per poter utilizzare il modello specificato

Possibili modelli statistici

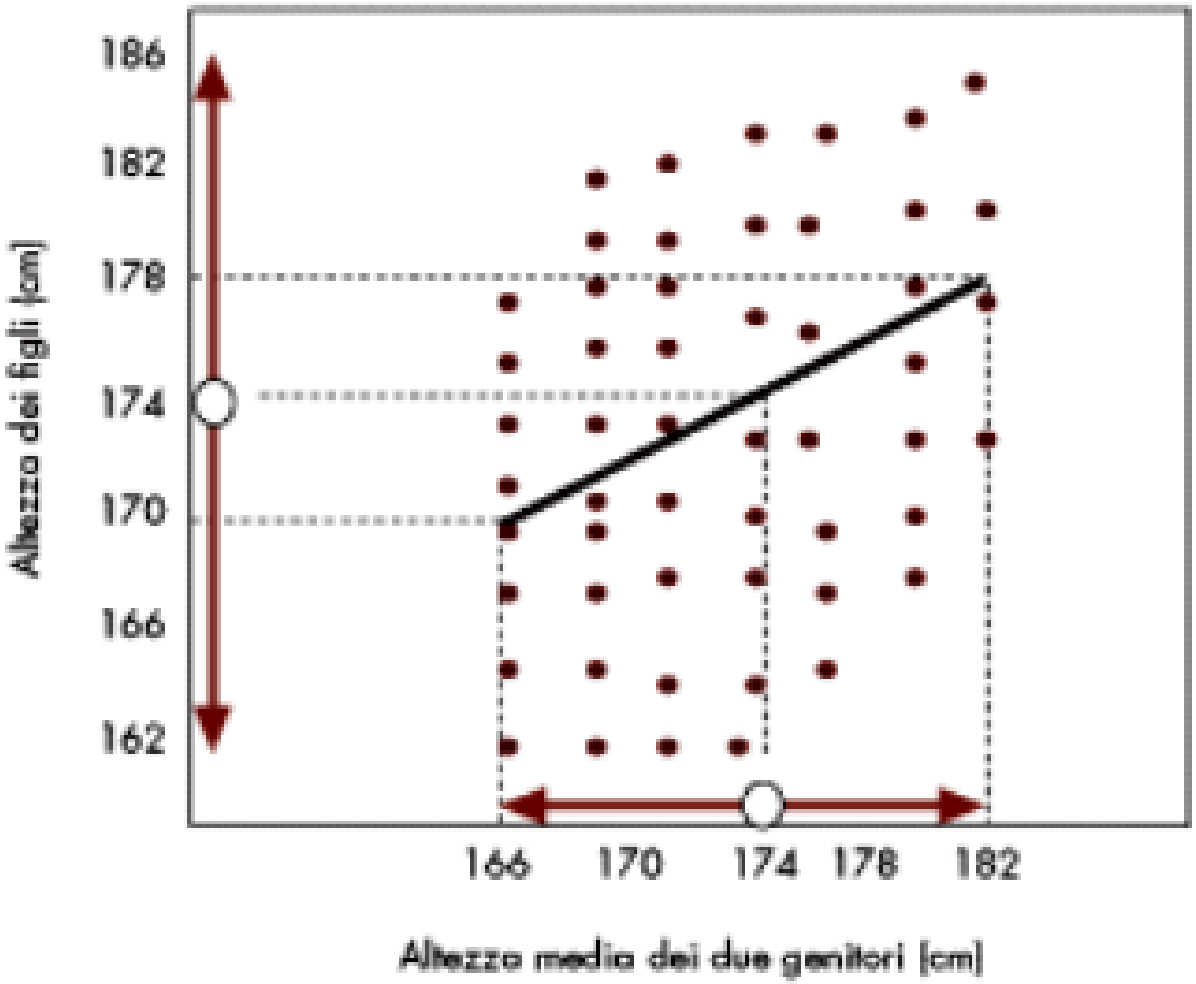


Indice

Il modello di regressione

- Il termine regressione deriva dall'applicazione svolta dal biologo Francis Galton che nel 1886 esaminò altezze dei figli (Y) in funzione delle altezze dei genitori (X) in Inghilterra
- notò una relazione funzionale tra le due variabili:v tendenzialmente a genitori più alti corrispondevano figli più alti e viceversa.
- Tuttavia ai genitori che si collocavano agli estremi (molto bassi o molto alti) non corrispondevano figli altrettanto estremi, ovvero Galton osservò che l'altezza dei figli **si spostava verso la media** e quindi concluse che questo costituiva una "regression towards mediocrity"
- la relazione funzionale fu chiamata "modello di regressione"

Il modello di regressione (2)



Il modello di regressione (3)

- Oggi il termine regressione è divenuto significato di "relazione funzionale tra variabili ottenuta con metodi statistici"
- la frase "regredire Y su $(X_1, X_2, \ldots, X_k)$" significa ricercare una relazione statistica del tipo:

$$Y = f(X_1, X_2, \ldots, X_k) + \epsilon$$

Il modello di regressione semplice (semplice perché include una sola variabile esplicativa) è specificato dalla relazione::

$$y_i = f(x_i; \beta) + \epsilon_i$$

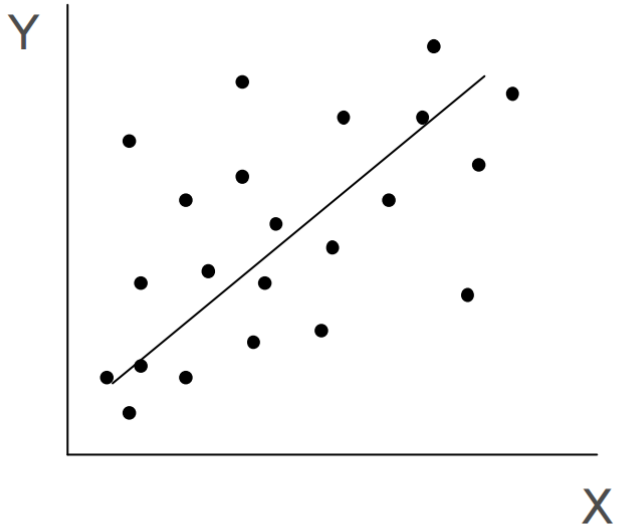
Il modello di regressione (4)

La funzione $f(.)$ può essere di primo grado ossia avremo che la specificazione del modello è la seguente:

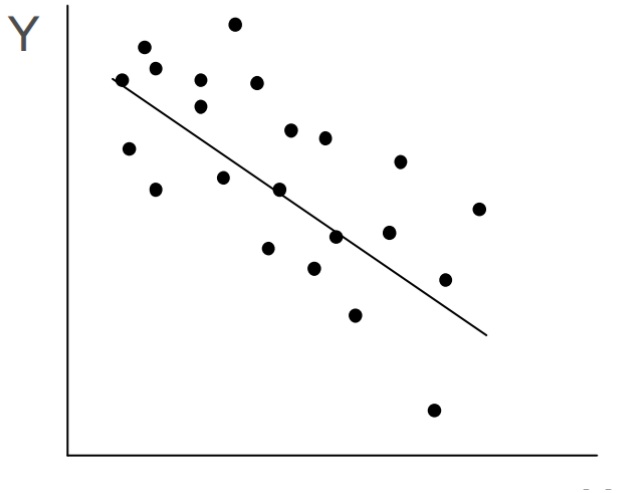
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

che identifica una retta, cioè la retta di regressione dove:
β_0 (o α) $\rightarrow$ intercetta, il valore di Y_i quando $x_i = 0$
$\beta_1 \rightarrow$ pendenza, di quanto cambia Y_i quando x_i incrementa di un'unità
$\epsilon_i \rightarrow$ l'errore che si commette nella spiegazione della variabile y_i tramite una funzione lineare di x (quanto ci discostiamo da una retta?)

Che relazione c'è tra X e Y?

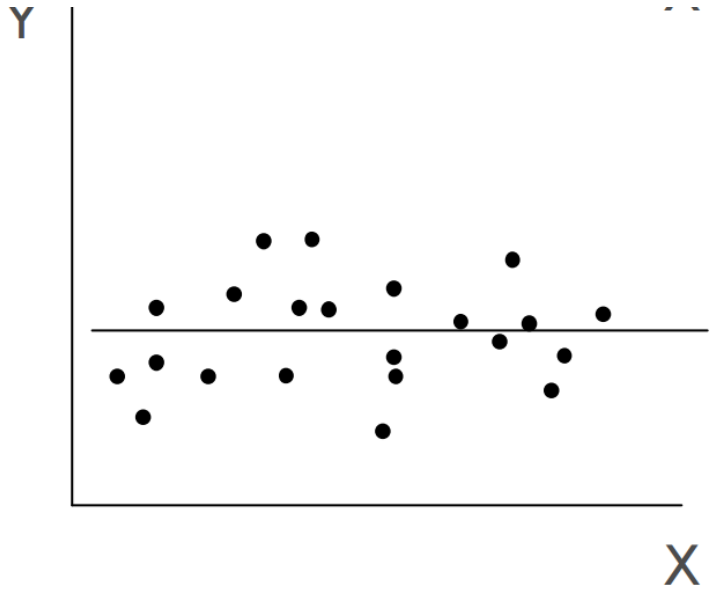


Covariano positivamente



Covariano negativamente

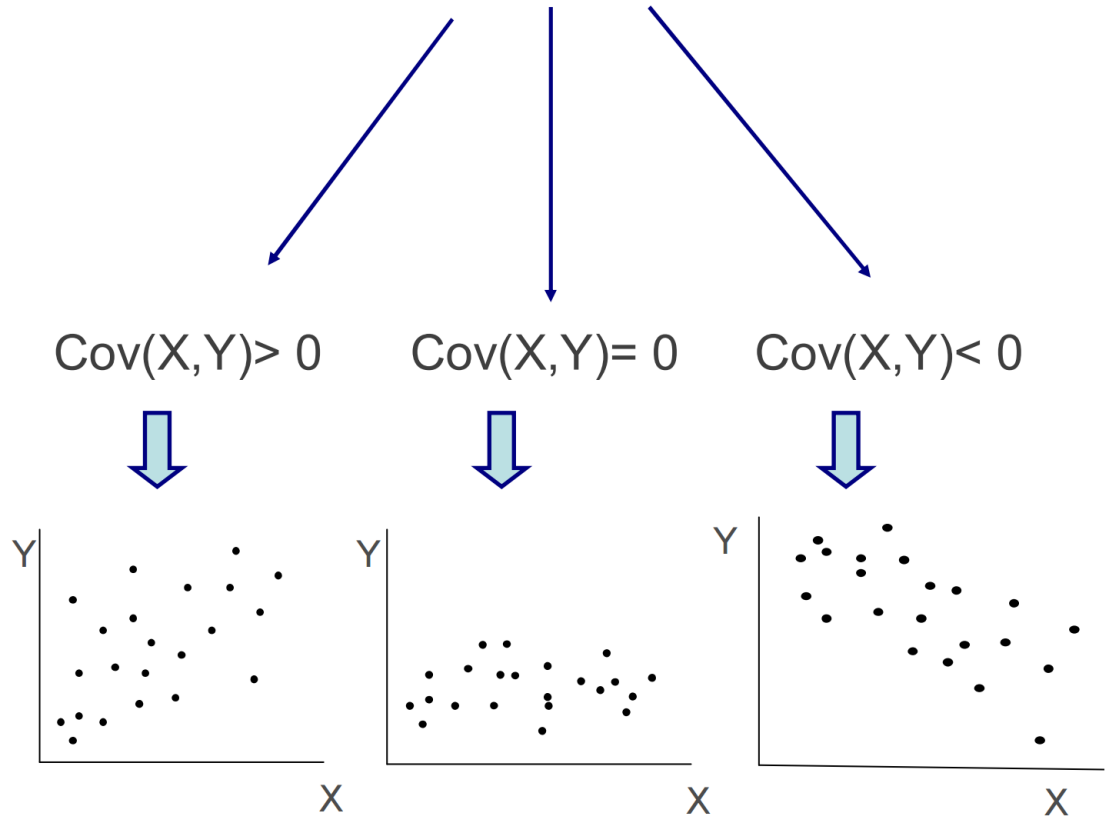
Che relazione c'è tra X e Y? (2)



Non covariano

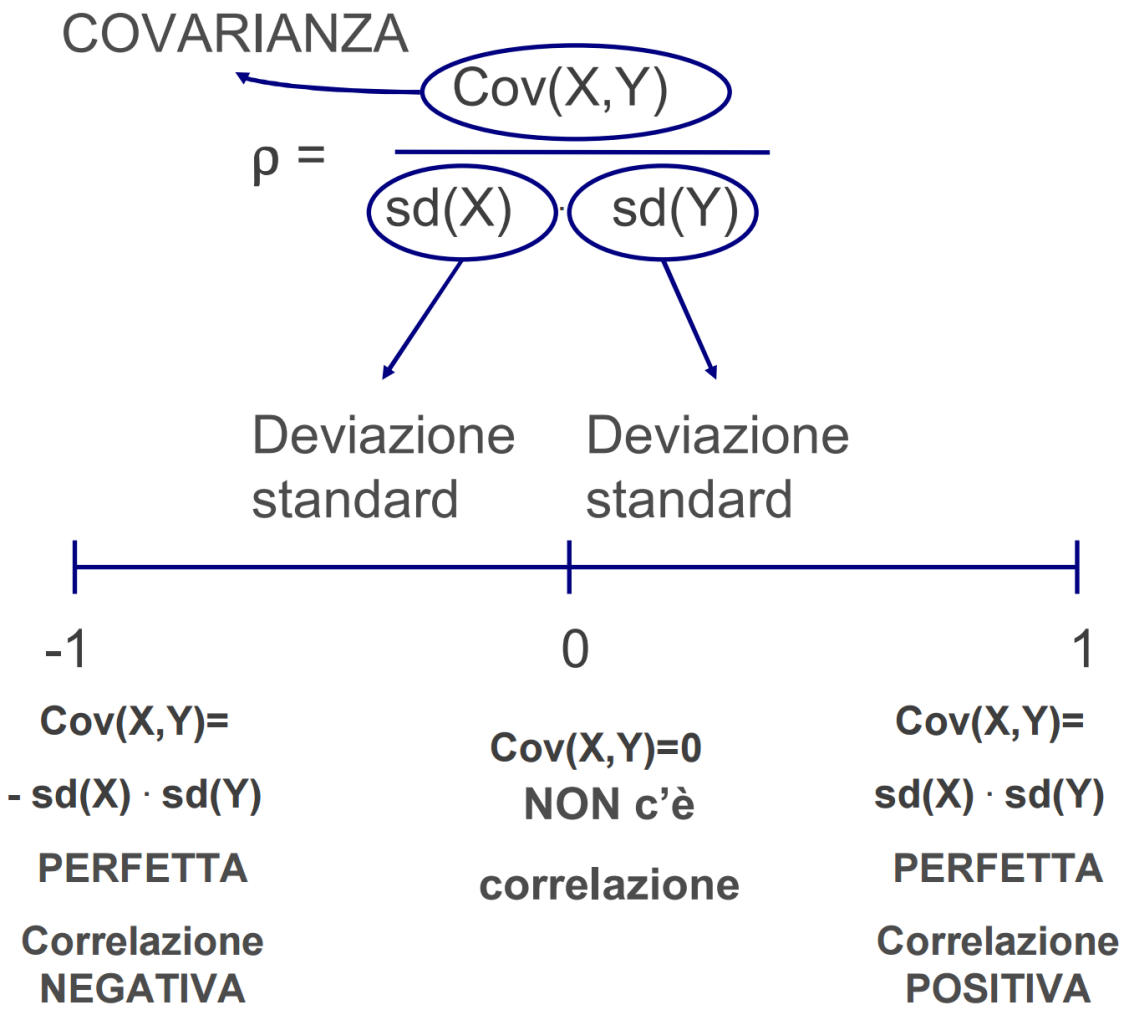
La covarianza misura l'attitudine a covariare di due variabili quantitative

$$Cov(X,Y) = \frac{\Sigma_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{n-1}$$



... e il coefficiente di correlazione

è una misura NORMALIZZATA della covarianza ed esprime quanto X e Y covariano

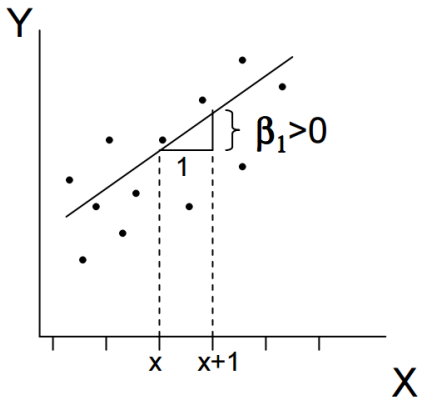


Correlazione e regressione

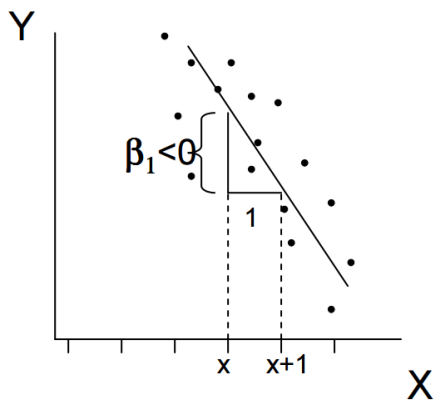
Quindi la correlazione esprime "quanto possiamo approssimare" con una **retta** la relazione tra due variabili quantitative (ossia quanto 'adatto' è un modello di regressione per i nostri dati)

Correlazione e regressione (2)

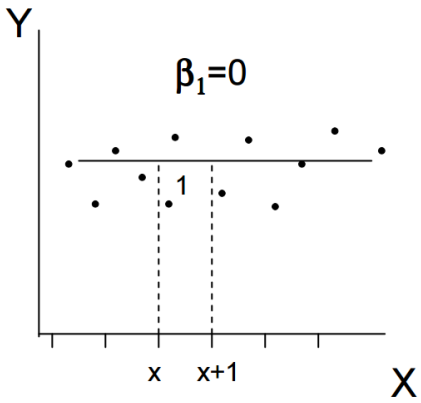
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



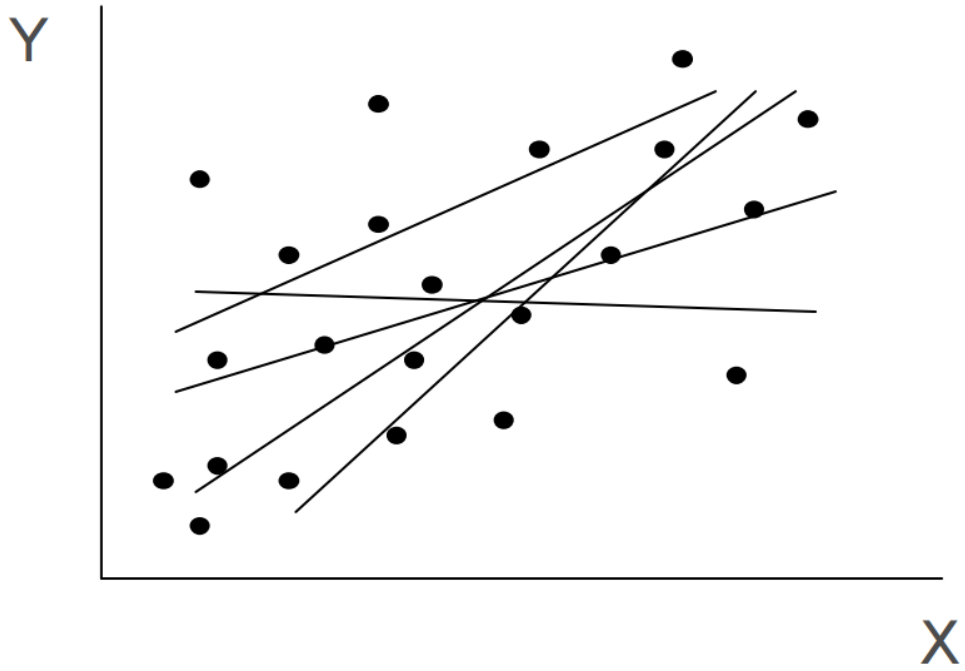
$$y_i = \beta_0 - \beta_1 x_i + \varepsilon_i$$



$$y_i = \beta_0 + \varepsilon_i$$

Correlazione e regressione (3)

ma ricordiamo che per un insieme di punti possono passare infinite rette! (ossia infiniti valori per β_0 e β_1)



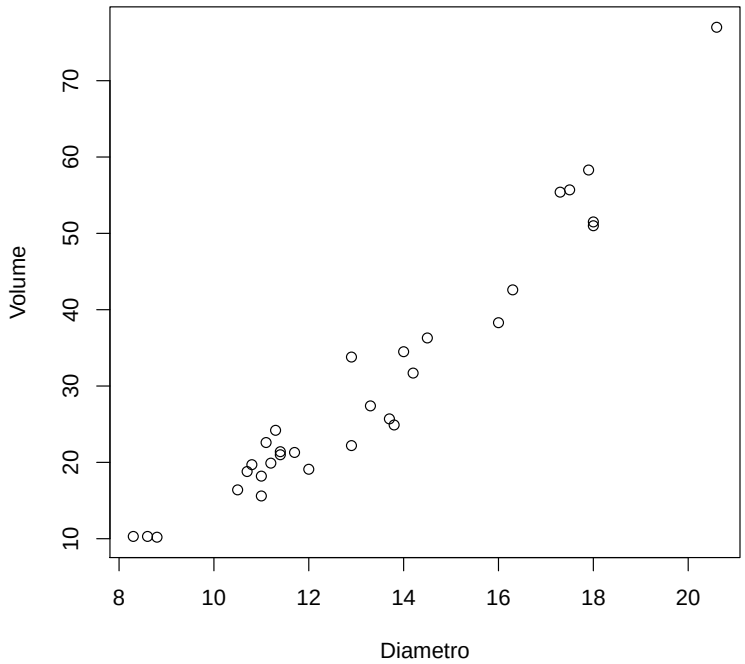
Correlazione e regressione (4)

Metodo di stima dei coefficienti del modello: Metodo dei minimi quadrati per individuare la retta più adatta (stima di β_0 e β_1)
Non è l'unico ma è quello standard, detto anche ols (ordinary least square)

Un esempio: ciliegi neri (trees)

diametro (in pollici)	altezza (in piedi)	volume del legno (in piedi^3)
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21.0
11.4	76	21.4
11.7	69	21.3
12.0	75	19.1
12.9	74	22.2
12.9	85	33.8
13.3	86	27.4
13.7	71	25.7
13.8	64	24.9
14.0	78	34.5
14.2	80	31.7
14.5	74	36.3
16.0	72	38.3
16.3	77	42.6
17.3	81	55.4
17.5	82	55.7
17.9	80	58.3
18.0	80	51.5
18.0	80	51.0
20.6	87	77.0

Un esempio: ciliegi neri (`trees`)



È evidente una forte relazione. Sostanzialmente lineare con forse un po' di problemi agli estremi.

Un primo modello

Adottiamo l'ipotesi di una relazione lineare.

Possiamo allora pensare ad un modello del tipo

$$(\text{volume}) = \alpha + \beta(\text{diametro}) + (\text{errore})$$

dove l'errore esprime la parte delle oscillazioni del volume non legate al diametro (o, meglio, che una funzione lineare del diametro non riesce a spiegare).

Un modello di questo tipo viene chiamato modello di regressione lineare semplice.
Per quanto riguarda il nome, regressione viene dalla storia, lineare perchè è lineare, semplice perchè si tenta di "spiegare" la risposta utilizzando una sola variabile esplicativa.

Modelli di regressione lineare semplice

$$y = \alpha + \beta x + \text{errore}.$$

- y variabile risposta o dipendente mentre
- x variabile esplicativa o indipendente o regressore.
- α intercetta
- β coefficiente angolare

α e β sono i parametri del modello. Il problema è ora come "determinare" α e β.

Minimi quadrati: idea

Sembra ragionevole scegliere per i parametri due valori, $\hat{\alpha}$ e $\hat{\beta}$, in modo tale che la retta di regressione "riproduca" bene i nostri dati, ovvero in modo tale che

$$\begin{aligned} y_1 &\approx \hat{\alpha} + \hat{\beta} x_i \\ y_2 &\approx \hat{\alpha} + \hat{\beta} x_2 \\ &\vdots \\ y_N &\approx \hat{\alpha} + \hat{\beta} x_N \end{aligned}$$

Per rendere "operativa" l'idea, dobbiamo decidere

- in che senso interpretiamo gli $\approx$ che abbiamo scritto e
- come combiniamo tra di loro le varie approssimazioni.

Minimi quadrati: idea (2)

La soluzione più usata si concretizza nello scegliere i due parametri minimizzando

$$s^2(\alpha, \beta) = \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2$$

ovvero scegliendo $\hat{\alpha}$ e $\hat{\beta}$ in maniera tale che

$$s^2(\hat{\alpha}, \hat{\beta}) \leq s^2(\alpha, \beta)$$

per qualsivoglia $\alpha \in R$ e $\beta \in R$.

In questo caso si dice che i parametri sono stati calcolati utilizzando il metodo dei minimi quadrati.

Minimi quadrati: determinazione dei parametri

Step 1 Fissato β ad un qualunque valore, il problema diventa

$$\inf_{\alpha \in R} \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2 = \inf_{\alpha \in R} \sum_{i=1}^{N} (z_i - \alpha)^2$$

con $z_i = y_i - \beta x_i$.

Sappiamo la costante che minimizza la media dei quadrati degli scarti da un valore è la media aritmetica delle z_i.

Quindi

$$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} z_i = \frac{1}{n} \sum_{i=1}^{N} (y_i - \beta x_i) = \overline{y} - \beta \overline{x}$$

dove $\overline{y}$ e $\overline{x}$ indicano rispettivamente la media delle y_i e quella delle x_i.

Minimi quadrati: determinazione dei parametri (2)

Step 2 La quantità da minimizzare diventa quindi

$$s^2(\hat{\alpha}, \beta) = \sum_{i=1}^{N} [y_i - \overline{y} - \beta(x_i - \overline{x})]^2.$$

Derivando rispetto a β e mettendo a zero la derivata si ottiene l'equazione (per β)

$$-2 \sum_{i=1}^{N} (x_i - \overline{x})[(y_i - \overline{y}) - \beta(x_i - \overline{x})] = 0,$$

che possiamo riscrivere come

$$\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}) = \beta \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

Minimi quadrati: determinazione dei parametri (3)

Se $\sum_{i=1}^{N}(x_i - \overline{x})^2 > 0$, l'equazione precedente ammette l'unica soluzione

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Quindi

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

$$\hat{\beta} = \frac{CODEV(X,Y)}{DEV(X)} = \frac{\sigma_{XY}}{\sigma_X^2}$$

dove $\overline{y}$, $\overline{x}$, σ_X^2 e σ_{XY} sono rispettivamente la media della variabile risposta, la media e la varianza della variabile esplicativa e la covarianza tra risposta e esplicativa.

Minimi quadrati: determinazione dei parametri (4)

Deve valere $\sigma_X^2 > 0$. Questo è molto ragionevole: β ci dice come varia la risposta al variare della esplicativa, ma se $\sigma_X^2 = 0$ l'esplicativa non è variata affatto nei dati disponibili.

Esempio: dataset `trees`

$$\sum y_i = 935{,}3 \quad \sum x_i = 410{,}7$$
$$\sum x_i^2 = 5736{,}5 \quad \sum x_i y_i = 13887{,}86.$$

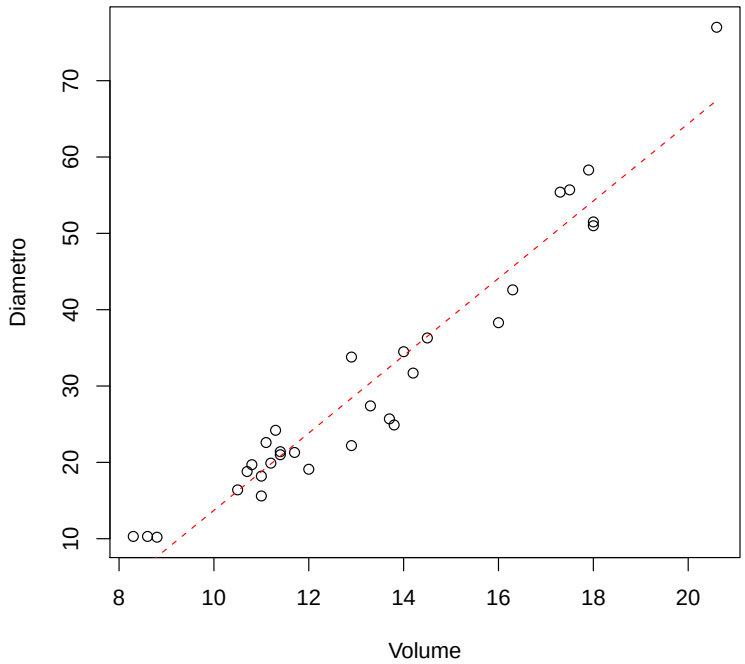
Perciò

$$\overline{y} = 935{,}3/31 = 30{,}2$$
$$\overline{x} = 410{,}7/31 = 13{,}2$$
$$\sigma_X^2 = (5736{,}5/31) - 13{,}2^2 = 9{,}5$$
$$\sigma_{XY} = (13887{,}86/31) - 13{,}2 \times 30{,}2 = 48{,}3.$$

Quindi

$$\hat{\beta} = 48{,}3/9{,}5 = 5{,}1$$
$$\hat{\alpha} = 30{,}2 - 5{,}1 \times 13{,}2 = -37{,}1.$$

Diagramma di dispersione con retta di regressione



La capacità di descrivere le variazione del volume sembra discreta con l'eccezione forse delle osservazioni più "esterne".

Indice

Valori osservati, previsti, residui

Per valutare la bontà del modello (in inglese GOF cioè goodness-of-fit), ossia l'adattamento del modello ai dati e seguenti quantità sono di interesse.

- y_i valore osservato per Y sulla i-sima unità statistica
- $\hat{y}_i$ valore previsto per Y sulla i-sima unità statistica: $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$
- r_i residuo per la i-sima unità statistica (realizzazioni della variabille casuale ϵ): $r_i = y_i - \hat{\alpha} - \hat{\beta} x_i$.

Il valore previsto sta sulla retta di regressione stimata; i residui misurano la distanza tra il valore osservato e la retta di regressione.

Media dei residui

È facile verificare che la media dei residui è nulla.

$$\sum_{i=1}^{N} r_i = \sum_{i=1}^{N} y_i - N\hat{\alpha} - \hat{\beta} \sum_{i=1}^{N} x_i =$$
$$= N\overline{y} - N(\overline{y} - \hat{\beta}\overline{x}) - N\hat{\beta}\overline{x} = 0$$

Varianza dei residui

$$
\begin{aligned}
\text{var}(r_1, \ldots, r_N) &= \sigma_R^2 = \frac{1}{N}\sum_{i=1}^{N} r_i^2 = \\
&= \frac{1}{N}\sum_{i=1}^{N}[(y_i - \overline{y}) - \hat{\beta}(x_i - \overline{x})]^2 = \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 + \frac{\hat{\beta}^2}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 - \\
&\quad -\frac{2\hat{\beta}}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \\
&= \sigma_Y^2 + \hat{\beta}^2\sigma_X^2 - 2\hat{\beta}\sigma_{XY} = \\
&= \sigma_Y^2 + \sigma_{XY}^2/\sigma_X^2 - 2\sigma_{XY}^2/\sigma_X^2 = \\
&= \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2
\end{aligned}
$$

Varianza dei residui (cont)

- La varianza dei residui, che coincide con la media dei quadrati dei residui, è sempre non più grande della varianza della risposta.
- la quantità detta $SSE = \sum_{i=1}^{N} r_i^2$ è il numeratore della varianza dei residui

- Può essere utilizzata per avere una "idea numerica" della bontà di adattamento del modello ai dati.

- Più σ_R^2 è piccola, più la retta di regressione "spiega" le variazioni della risposta. Quando $\sigma_R^2 = 0$, tutti le osservazioni giacciono sulla retta di regressione.

- Quando $\sigma_{XY} = 0$, cioè in assenza di una relazione lineare, $\sigma_R^2 = \sigma_Y^2$.

Coefficiente di determinazione

La frazione della varianza della risposta (Y) spiegata dal modello di regressione lineare semplice è data da

$$R^2 = 1 - \frac{\sigma_R^2}{\sigma_Y^2}$$

$$0 \leq R^2 \leq 1$$

$R^2 = 1 \longrightarrow \sigma_R^2 = 0$: il modello spiega perfettamente la risposta.
$R^2 = 0 \longrightarrow \sigma_R^2 = \sigma_Y^2$: il modello non spiega per niente.

Esempio: dataset `trees`

$$\bar{y} = 935{,}3/31 = 30{,}2$$
$$\sigma_X^2 = (5736{,}5/31) - 13{,}2^2 = 9{,}5$$
$$\sigma_{XY} = (13887{,}86/31) - 13{,}2 \times 30{,}2 = 48{,}3.$$

Inoltre

$$\sum y_i^2 = 36324{,}99$$

Quindi

$$\sigma_Y^2 = 36324{,}99/31 - 30{,}2^2 = 261{,}5$$

e perciò

$$\sigma_R^2 = 261{,}5 - 48{,}3^2/9{,}5 = 15{,}9.$$

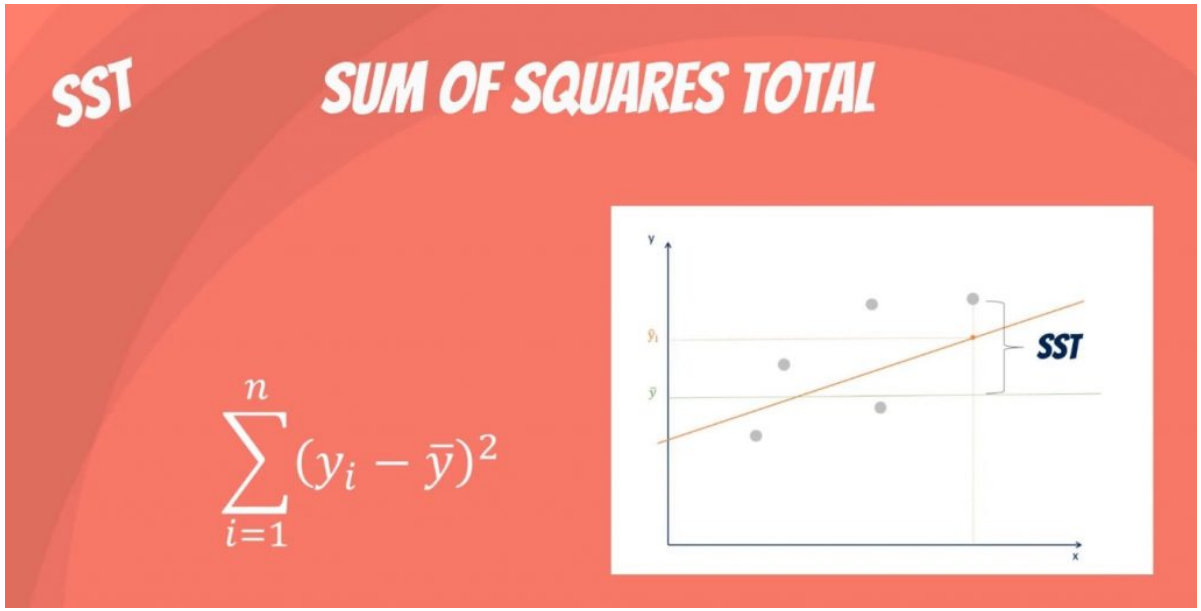
Il coefficiente di determinazione vale

$$R^2 = 1 - 15{,}9/261{,}5 = 0{,}94,$$

ovvero il modello spiega il 94% della varianza della variabile di risposta.

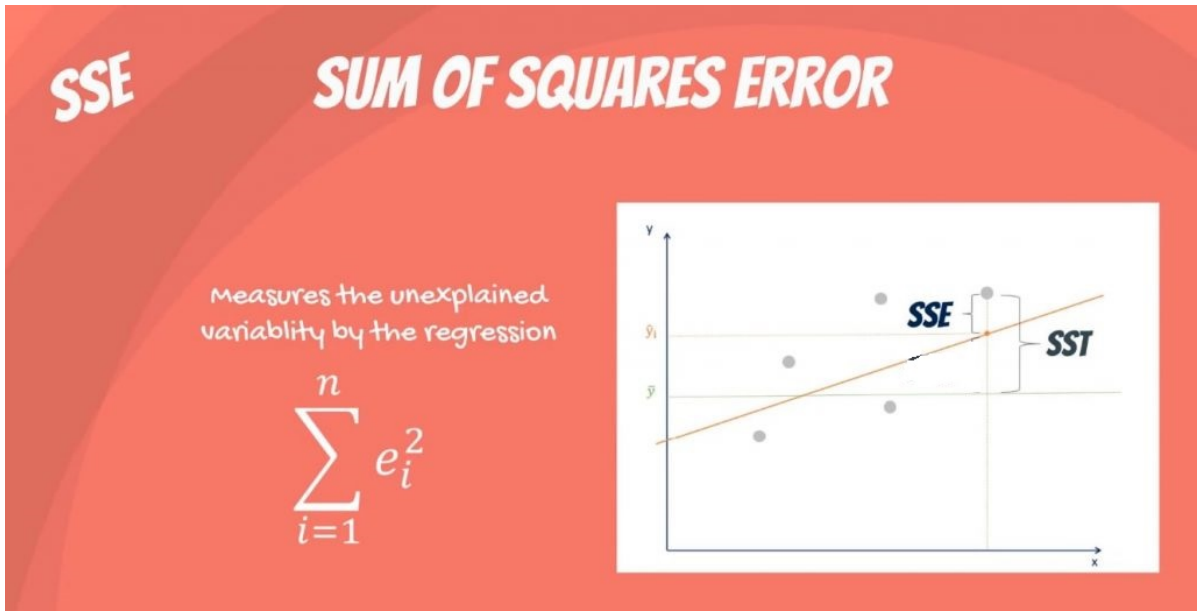
Analisi della varianza

- se non vi fosse associazione tra la Y e la X $\beta = 0$ e il miglior predittore per ciascuna osservazione y_i di Y sarebbe la sua media $\alpha = \overline{y}$, cioè il modello sarebbe $Y = \overline{y}$
- in tal caso la deviazione dal modello è data dalla "devianza totale" nei dati SST (total sum of squares) cioè la distanza al quadrato tra le osservazioni di Y e la retta $\overline{y}$



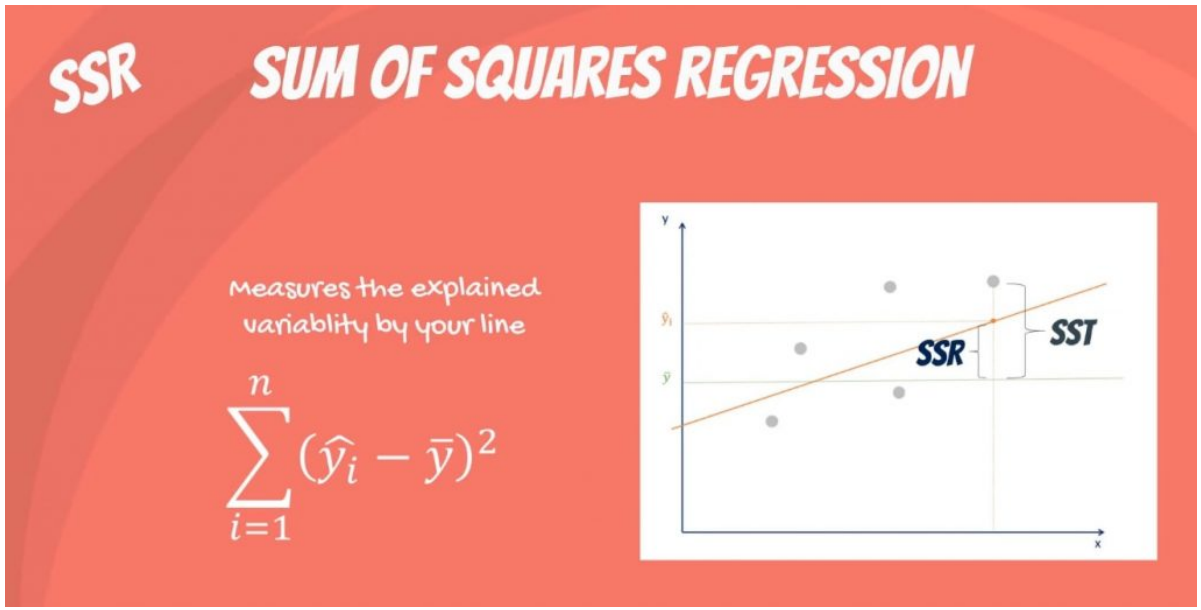
Analisi della varianza (2)

- Quando vi è associazione tra Y e X ($\beta \neq 0$), il miglior predittore di ciascuna osservazione y_i di Y sarebbe $\hat{y}_i = \hat{\alpha} + \beta x_i$
- In tal caso, la deviazione dal modello da considerare è proprio $SSE = (y_i - \hat{y}_i)^2 = (y_i - \hat{\alpha} - \beta x_i)^2$ ossia il numeratore della varianza dei residui che indica la distanza tra le osservazioni della Y e la retta di regressione



Analisi della varianza (3)

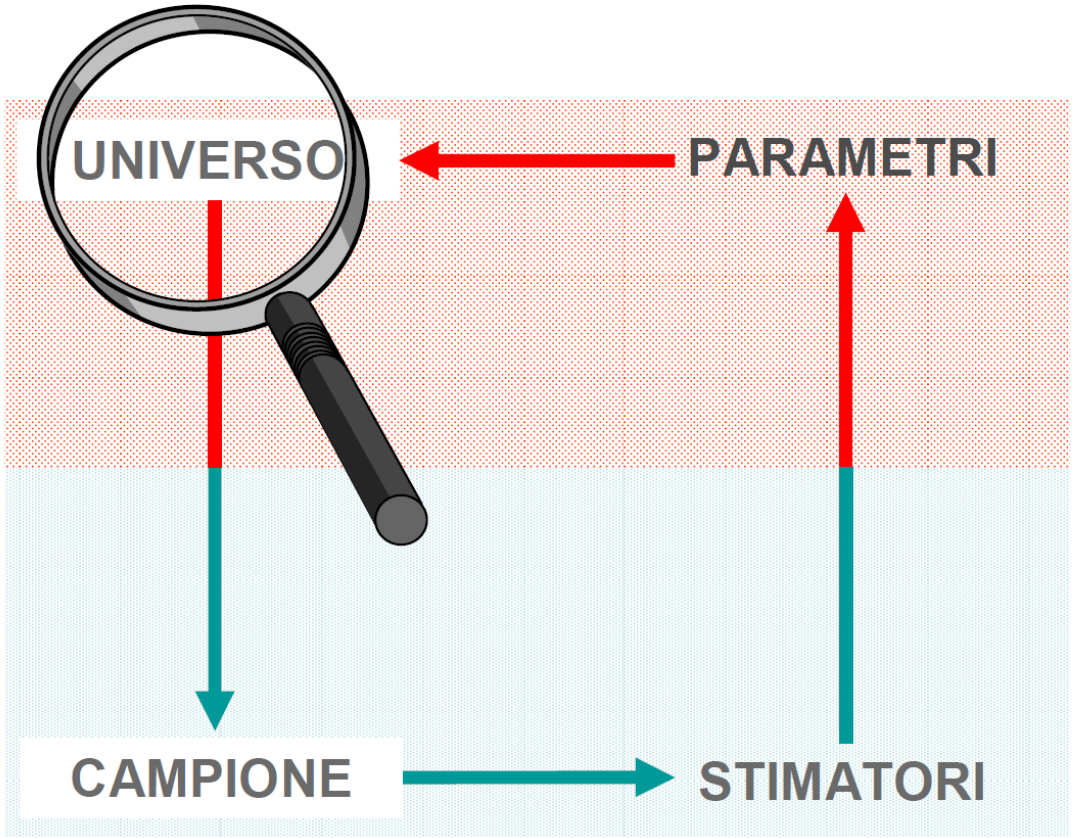
- La differenza tra la SST e SSE è la varianza "spiegata" di Y dal modello di regressione usando la X.
- Rappresenta la distanza al quadrato tra i valori predetti dal modello (sulla retta) e la media di Y: $SSR = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$ detta Regression Sum of Squares



Indice

Test nullità dei coefficienti

Siamo interessati a valutare l'esistenza di una relazione tra X e Y nella popolazione tramite un modello di regressione.



Test nullità dei coefficienti (2)

Prima di poter "usare" un modello di regressione stimato occorre testare se il coefficiente angolare (la pendenza) è statisticamente diverso da 0... cioè possiamo supporre che il valore stimato della pendenza è diverso da 0 nella popolazione?

Per valutare questo si ricorre ad un test di verifica delle ipotesi:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

La statistica test espressa in maniera semplice è:

$$T_{oss} = \frac{\hat{\beta}_1 - 0}{E.S._{\hat{\beta}_1}}$$

Test nullità dei coefficienti (3)

Nei test di verifica delle ipotesi l'evidenza a favore o meno dell'ipotesi nulla si valuta mediante il valore-p (p-value)...

Se valore-p < 0.05 si può concludere che vi è sufficiente evidenza contro H_0!

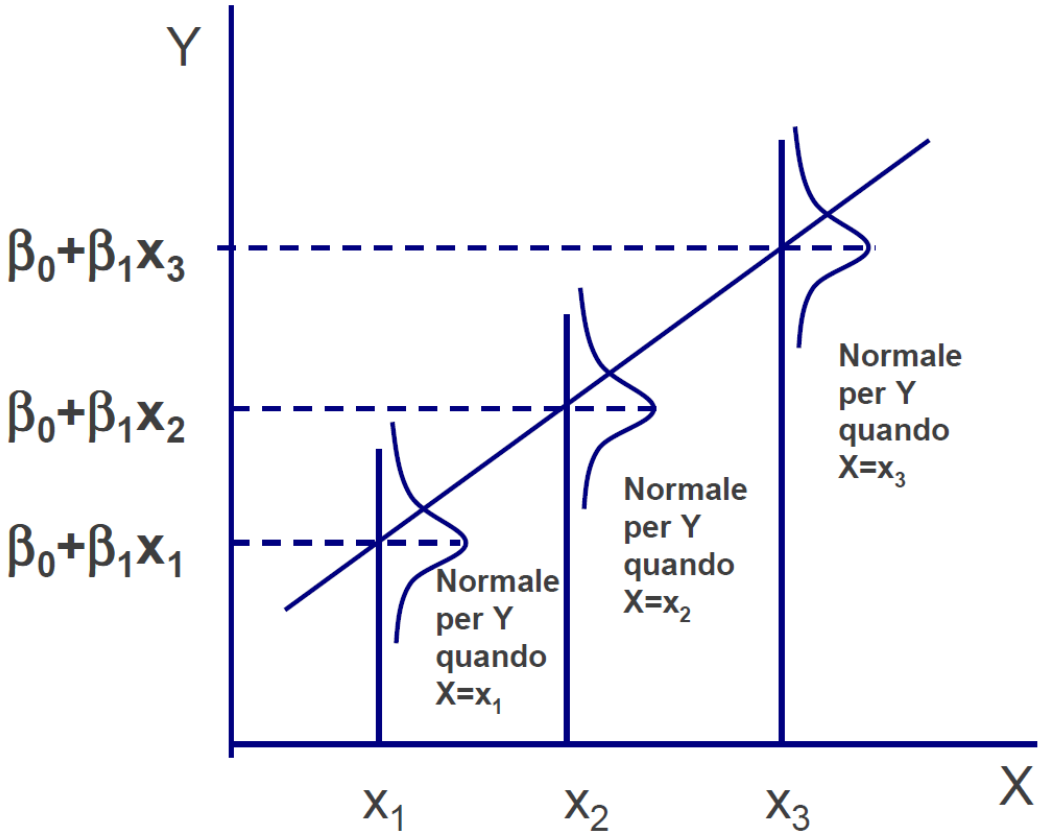
Cioè possiamo usare il modello posto che si adatti bene ai dati e che non si violino le assunzioni su cui si basa.

Indice

Assunzioni del modello di regressione

1. Le unità statistiche sono indipendenti tra loro (campionamento casuale)
2. Y distribuita normalmente
3. Fissato un valore di X abbiamo una popolazione di valori di Y distribuiti normalmente con media situata sulla retta di regressione (in altri termini la distribuzione condizionata di Y dato X deve essere distribuita come una normale)

Assunzioni del modello di regressione (2)



$$Y|X=x_1 \sim N(\beta_0+\beta_1 x_1, \sigma^2)$$
$$Y|X=x_2 \sim N(\beta_0+\beta_1 x_2, \sigma^2)$$
$$Y|X=x_3 \sim N(\beta_0+\beta_1 x_3, \sigma^2)$$

Assunzioni del modello di regressione (3)

4. La varianza della Y rimane la stessa indipendentemente da X
   Omoschedasticità: $Var(y_i) = \sigma^2$

Analisi dei residui (Parte pratica)

- I residui dovrebbero essere distribuiti come una variabile casuale normale con varianza costante

Indice

Il modello di regressione multipla

- Quando si vuole stimare un modello di regressione con più di una variabile esplicativa (ad esempio p variabili esplicative) si parla di modello di regressione multipla

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Il modello di regressione multipla (2)

Cioè abbiamo i seguenti dati (matrice dati):

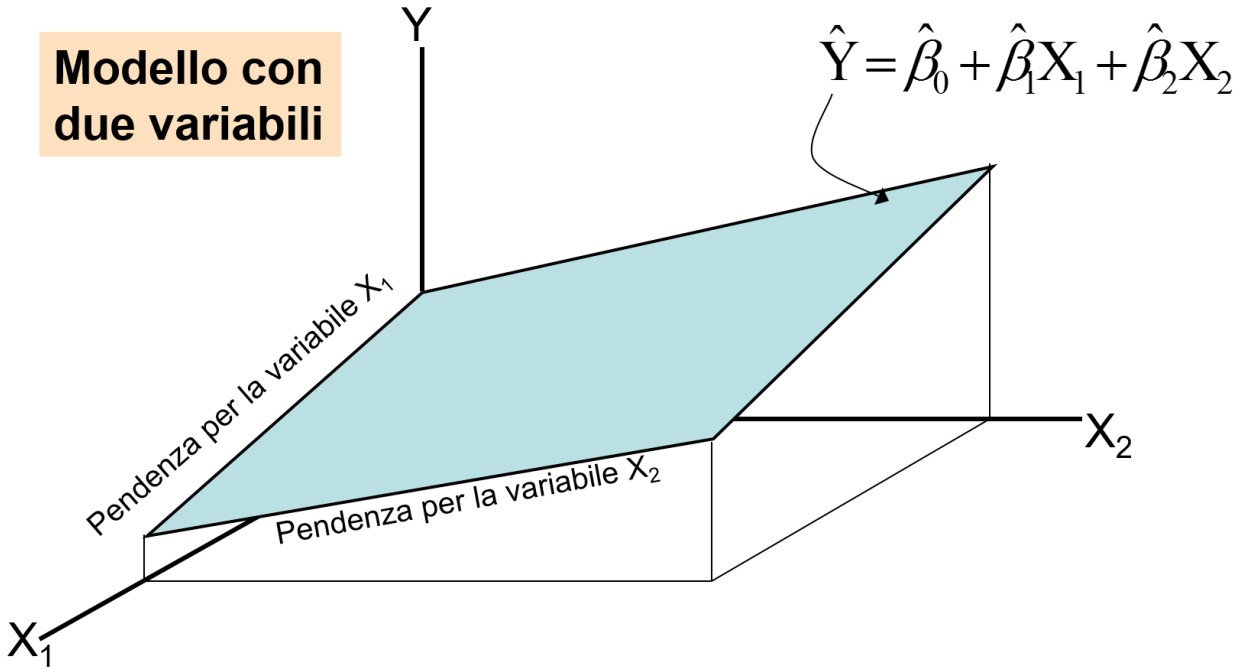
$\underline{Y}$	$\underline{X_1}$	$\underline{X_2}$	$\underline{X_3}$	...	...	...	$\underline{X_p}$
y_1	X_{11}	X_{12}	X_{13}	...	...	...	X_{1p}
y_2	X_{21}	X_{22}	X_{23}	...	...	...	X_{2p}
y_3	X_{31}	X_{32}	X_{33}	...	...	...	X_{3p}
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
y_n	X_{n1}	X_{n2}	X_{n3}	...	...	...	X_{np}
(nx1)				(nxp)			

Così come una sola variabile X dava luogo ad una retta di regressione (funzione lineare in due dimensioni, piano X–Y) quando si considerano due variabili esplicative X_1 e X_2 il modello specifica un piano di regressione (funzione lineare in tre dimensioni, spazio tridimensionale X_1-X_2-Y) e così via...

Il piano di regressione

Nel caso di due sole variabili esplicative X_1 e X_2 si ha il piano di regressione

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$



Esempio: estensione di un modello

Con riferimento ad un campione casuale di 20 famiglie si cerca di spiegare il consumo alimentare (Y) utilizzando come variabile esplicativa il reddito (X_1).

Esempio: estensione di un modello (2)

famiglia	Spesa annua per l'alimentazione (000*Euro)	Reddito annuo (000*Euro)	Dimensione della famiglia (numero di componenti)
	SPESA	REDDITO	NC
1	5.2	28	3
2	5.1	26	3
3	5.6	32	2
4	4.6	24	1
5	11.3	54	4
6	8.1	59	2
7	7.8	44	3
8	5.8	30	2
9	5.1	40	1
10	18	82	6
11	4.9	42	3
12	11.8	58	4
13	5.2	28	1
14	4.8	20	5
15	7.9	42	3
16	6.4	47	1
17	20	112	6
18	13.7	85	5
19	5.1	31	2
20	2.9	26	2

Esempio: estensione di un modello (3)

Il modello stimato è il seguente:

$$\hat{y}_i = -0.412 + 0.184 x_{1i}, (i = 1, 2, \ldots, 20)$$

Ora estendiamo il modello per considerare anche la dimensione della famiglia (X_2), misurata in termini di numero di componenti (NC) della famiglia. Il modello diventa:

$$Y = \beta_0 + \beta_1 REDDITO + \beta_2 NC + \epsilon$$

- Dovremmo aspettarci che i segni di β_1 e di β_2 siano entrambi positivi, cioè che sia il reddito sia la dimensione della famiglia abbiano effetti positivi sulla spesa alimentare della famiglia. Ciò vale nel caso di singole regressioni lineari semplici

Esempio: estensione di un modello (4)

- Invece β_1 misura l'effetto parziale del reddito sulla spesa alimentare, **tenendo costante** la dimensione della famiglia
- Analogamente β_2 misura l'effetto parziale della dimensione della famiglia sulla spesa, tenendo costante il reddito

Esempio: estensione di un modello (5)

Lo strumento principale per valutare le relazioni tra le variabili in gioco è la matrice di correlazione.
Infatti mentre con una sola X ci dobbiamo preoccupare solo della correlazione tra X e Y ora con più variabili esplicative dobbiamo considerare le correlazioni tra le X e la Y e tra le X stesse. Attraverso la matrice di correlazione possiamo valutare tali quantità:

	SPESA	*REDDITO*	*NC*
SPESA	1		
REDDITO	0.95	1	
NC	0.79	0.68	1

Esempio: estensione di un modello (6)

in linea di principio le condizioni desiderate sono:

1) alta correlazione (in valore assoluto) tra le X e la Y
2) bassa correlazione tra le $X \rightarrow$ Alta corr tra le X (collinearità) crea problemi di stima dei coefficienti del modello

Esempio: estensione di un modello (7)

Risultati

$$\widehat{SPESA} = -1{,}118 + 0{,}148\,(\text{Reddito}) + 0{,}793(\text{NC})$$

OUTPUT RIEPILOGO						
Statistica della regressione						
R multiplo	0.967					
R al quadrato	0.935					
R al quadrato corretto	0.927					
Errore standard	1.261					
Osservazioni	20					
ANALISI VARIANZA						
	gdl	*SQ*	*MQ*	*F*	*Significatività F*	
Regressione	2	386.3129	193.15643	121.4702	8.558E-11	
Residuo	17	27.03264	1.5901551			
Totale	19	413.3455				
	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività (p-value)*	*Inferiore 95%*	*Superiore 95%*
Intercetta	-1.118	0.655	-1.708	0.106	-2.500	0.263
REDDITO	0.148	0.016	9.049	0.000	0.114	0.183
NC	0.793	0.244	3.245	0.005	0.277	1.309

Esempio: estensione di un modello (8)

$$\hat{SPESA} = -1{,}118 + 0{,}148\,(Reddito) + 0{,}793(NC)$$

Dove
- SPESA è in Euro*1000
- REDDITO è in Euro*1000
- NC è in numero di componenti.

b1 = 0,148: la SPESA alimentare aumenta, in media, di **148** Euro all'anno all'aumentare di **1000 Euro del REDDITO**, al netto (fermo restando) degli effetti dovuti alle variazioni di **NC**

b_2 = 0,793: la SPESA alimentare aumenta, in media, di **793** Euro all'anno all'aumentare di **1 di NC** , al netto (fermo restando) degli effetti dovuti alle variazioni del **REDDITO**

Esempio: estensione di un modello (9)

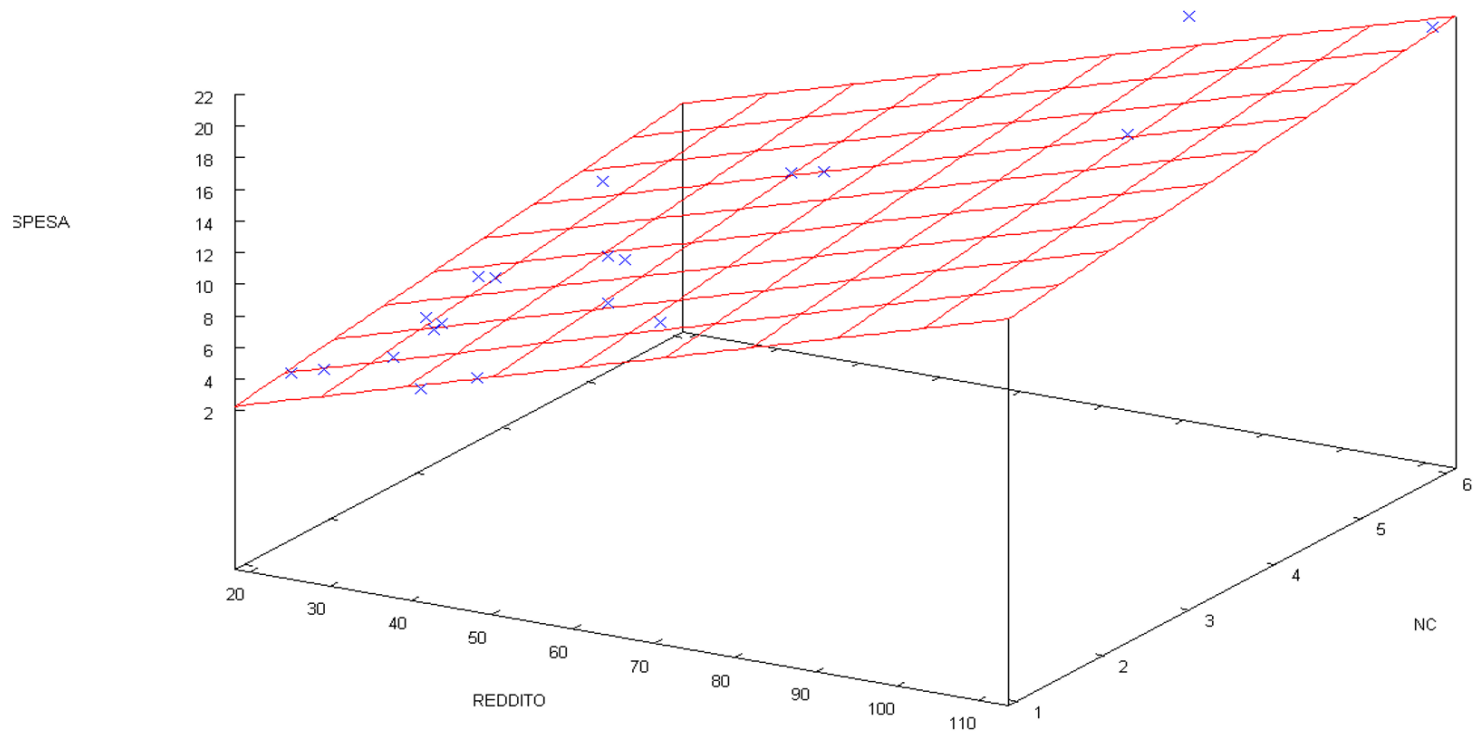
Commento e significato dei parametri (coefficienti)

- $\hat{\beta}_0 = -1.118$ nessun significato interpretabile perché il livello medio della spesa alimentare della famiglia non può essere negativo anche quando nessun componente ha una occupazione remunerata. Inoltre, non è realistico pensare all'esistenza di una famiglia che pur non avendo reddito e zero componenti presenta una spesa per alimentazione. Ciò nonostante, questo valore non dovrebbe essere scartato; svolge un ruolo importante quando si utilizza la equazione di regressione stimata per la previsione.
- $\hat{\beta}_1 = 0.148$ Rappresenta l'effetto parziale del reddito annuale della famiglia sulla spesa per alimentazione, tenendo costante la dimensione. Il segno positivo stimato implica che tale effetto è positivo mentre il valore assoluto implica che il consumo alimentare aumenta di 148 euro per ogni 1000 euro di aumento nel reddito (ricordiamo che la X_1 è espressa in migliaia di euro).

Esempio: estensione di un modello (10)

- $\hat{\beta}_2 = 0.793$ rappresenta l'effetto parziale della dimensione della famiglia sulla spesa per alimentazione, tenendo costante il reddito della famiglia. Il segno positivo stimato implica che tale effetto sia positivo mentre il valore assoluto implica che la spesa alimentare aumenta di 793 euro per ogni componente della famiglia in più (per matrimonio, nascita, adozione, ecc.)

Esempio: estensione di un modello (11)



Indice

Variabili indipendenti qualitative

- Di solito le variabili nella regressione sono variabili continue
- In molte applicazioni si rende necessario l'introduzione di una variabile qualitative le cui modalità saranno due o più
  - Ad esempio possibili variabili esplicative possono essere: genere (modalità maschio/femmina), titolo di studio (nessuno/elementare/medio-superiore/laurea), anno di iscrizione (primo/secondo/terzo/fuori corso)
  - in generale le variabili qualitative determinano sottogruppi nel campione (sottocampioni nel gergo di XLstat)

Variabili indipendenti qualitative (2)

- nel modello di regressione le variabili qualitative non vengono utilizzate così come sono ma occorre utilizzare la codifica in variabili dummy
  - Le dummy sono variabili che assumono in genere solo i valori (0, 1) a seconda che la variabile di interesse abbia assunto una delle sue modalità (siamo liberi di scegliere quale, ad esempio per la variabile "genere": Maschio=0 e Femmina=1).
  - Se la variabile ha solo due modalità **basta una sola variabile dummy per rappresentarla completamente**
  - Altrimenti se la variabile di interesse ha k modalità **servono** $k-1$ **variabili dummy per rappresentarla**

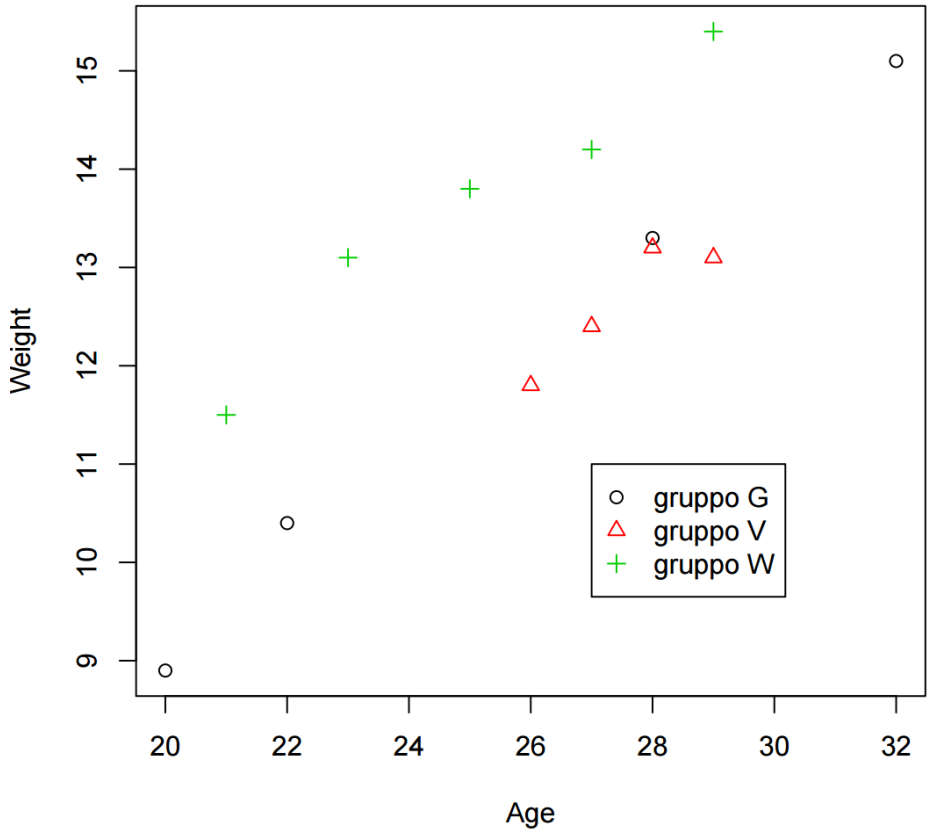
Variabili indipendenti qualitative: Esempio

Il peso (weight) e l'età (age) di 13 tacchini consumati durante il Giorno del Ringraziamento sono riportati nella seguente tabella, con la regione di provenienza (county).

	age	weight	county
1	28.00	13.30	G
2	20.00	8.90	G
3	32.00	15.10	G
4	22.00	10.40	G
5	29.00	13.10	V
6	27.00	12.40	V
7	28.00	13.20	V
8	26.00	11.80	V
9	21.00	11.50	W
10	27.00	14.20	W
11	29.00	15.40	W
12	23.00	13.10	W
13	25.00	13.80	W

Variabili indipendenti qualitative: Esempio (2)

Il grafico seguente mostra il diagramma di dispersione dei tacchini considerando tutte e tre le variabili:



Variabili indipendenti qualitative: Esempio (3)

Stimiamo i parametri della retta di regressione $y = \beta_0 + \beta_1 x + \epsilon$ dove con y indichiamo il peso e con x l'età.
La bontà di adattamento è $R^2 = 0.66$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9833	2.3327	0.85	0.4133
age	0.4167	0.0892	4.67	0.0007

il modello è quindi accettabile. Ma per migliorare l'adattamento potremmo provare a verificare se la regione di provenienza abbia qualche effetto sul peso.

Variabili indipendenti qualitative: Esempio (4)

- Poiche la variabile assume tre modalità costruiamo due variabili z_1 e z_2 ciascuna delle quali assume solo i valori (0, 1)

	x	z_1	z_2
1	28	0	0
1	20	0	0
1	32	0	0
1	22	0	0
1	29	1	0
1	27	1	0
1	28	1	0
1	26	1	0
1	21	0	1
1	27	0	1
1	29	0	1
1	23	0	1
1	25	0	1

Variabili indipendenti qualitative: Esempio (5)

- In pratica quando entrambe le dummy valgono zero **indichiamo** la modalità "Regione G" (questa si chiama modalità di riferimento o reference category)
- Quando $z_1 = 1$ E $z_2 = 0$ indichiamo la "regione V"
- Quando $z_1 = 0$ E $z_2 = 1$ indichiamo la "regione W"

In questo senso per una variabile con k modalità ci vogliono $k - 1$ dummy

- **IMPORTANTE**: per includere qualunque variabile qualitativa in un modello di regressione dobbiamo SEMPRE usare la codifica dummy di quella variabile

Variabili indipendenti qualitative: Esempio (6)

- Pertanto il nostro modello con inclusa la variabile regione (county) sarà

$$y = \beta_0 + \beta_1 x + \alpha_1 z_1 + \alpha_2 z_2 + \epsilon$$

- In pratica quando entrambe le dummy valgono zero si ha il modello di riferimento (quello relativo alla reference category), quando una delle due vale 1 si hanno gli altri due modelli da confrontare con quello di riferimento:

$$y = \beta_0 + \beta_1 x + \epsilon$$
$$y = (\beta_0 + \alpha_1) + \beta_1 x + \epsilon$$
$$y = (\beta_0 + \alpha_2) + \beta_1 x + \epsilon$$

Variabili indipendenti qualitative: Esempio (7)

I risultati della stima sono riassunti nella seguente tabella:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−0.4875	0.6734	−0.72	0.4875
age	0.4868	0.0257	18.91	0.0000
countyV	−0.2735	0.2184	−1.25	0.2421
countyW	1.9184	0.2018	9.51	0.0000

I valori delle stime nelle righe denominate countyV e countyW sono rispettivamente i valori di α_1 e α_2. Il modello stimato risulta quindi:

$$y = -0.4875 + 0.4868x - 0.2735z_1 + 19184z_2 + \epsilon$$

Ossia:

$$\widehat{y} = -0.4875 + 0.4868x \quad \text{per G}$$
$$\widehat{y} = -0.761 + 0.4868x \quad \text{per V}$$
$$\widehat{y} = 1.4309 + 0.4868x \quad \text{per W}$$

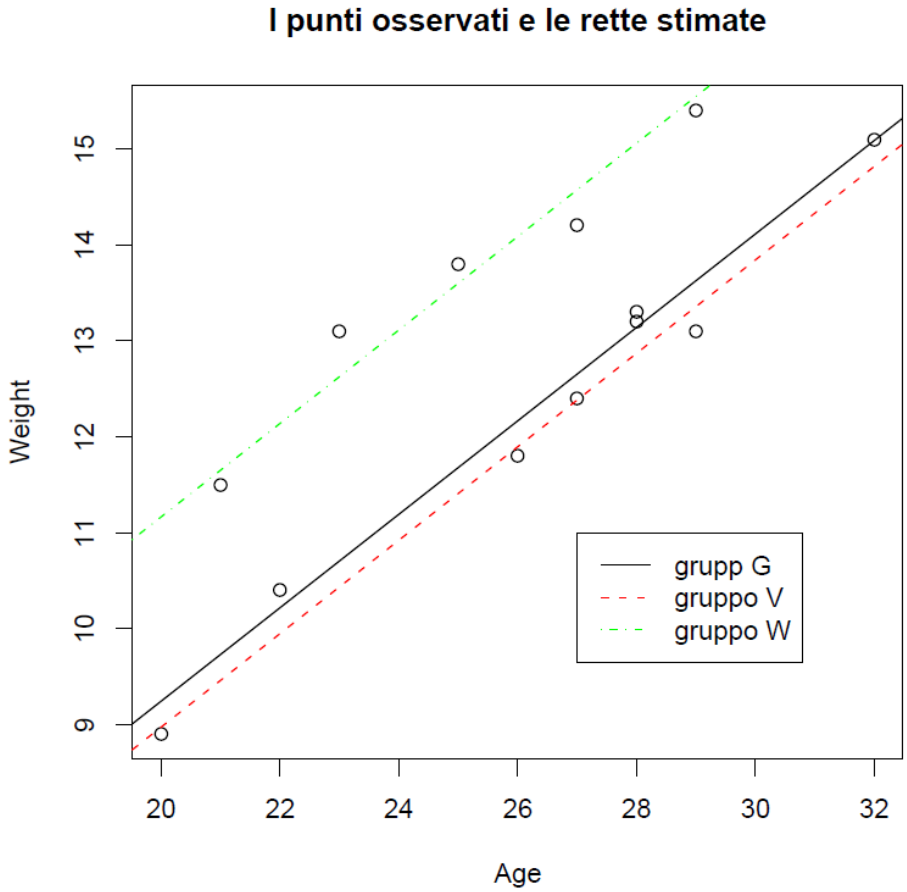
Variabili indipendenti qualitative: Esempio (8)

Il valore di $R^2 = 0.9794$ è notevolmente cresciuto, quindi la regione di provenienza ha un notevole effetto sul peso y. Il significato dei parametri è il seguente:

- α_1 stima la differenza della risposta y tra la regione G di riferimento e la regione V. Il test $H0 : \alpha_1 = 0$ ci porta a concludere che questa ipotesi è plausibile, cioè tra i due gruppi la differenza non è significativa
- α_2 stima la differenza della risposta tra la regione G di riferimento e la regione W. Il test $H0 : \alpha_2 = 0$ ci porta a concludere che questa ipotesi va rigettata, cioè tra i due gruppi la differenza è significativa.
  - In particolare l'effetto sul peso dei tacchini della provenienza dalla regione W è 1.9184 ossia i tacchini provenienti da quella regione pesano in media $+1.9184$ kg rispetto alla regione G
- La differenza tra i gruppi V e W è data da:
  $\alpha_1 - \alpha_2 = -0.2735 - 1.9184 = -2.1919$

Variabili indipendenti qualitative: Esempio (9)

Il grafico mostra le tre rette stimate (per le tre modalità della variabile County)



Indice

1) R^2 ed R^2 corretto

Rispetto al modello di regressione semplice in quello multiplo occorre valutare anche la complessità del modello (cioè il numero di variabili esplicative utilizzate)

- Verifica della bontà del modello si può ancora usare l'indice R^2
- ma se si vuole tenere conto anche della complessità del modello si usa R^2 corretto (adjusted R^2) che tiene conto della numerosità del campione e del numero di variabili esplicative
- questi due indici ci dicono se le variabili esplicative sono idonee a prevedere (o "spiegare") i valori della variabile dipendente mediante la proporzione di varianza della Y spiegata dal modello di regressione

1) R^2 ed R^2 corretto (2)

L'R^2 corretto...

- Penalizza l'impiego eccessivo di variabili indipendenti poco importanti
- Ha un valore inferiore di R^2
- Utile per capire se una variabile esplicativa addizionale ha un contributo importante nella modellizzazione della Y

1) R^2 ed R^2 corretto (3)

$$R^2_{adj} = 0{,}927$$

Il 92,7% della variabilità nella spesa alimentare è spiegato tramite la variazione nel reddito e nella dimensione della famiglia, tenendo conto della dimensione del campione e del numero di variabili indipendenti

OUTPUT RIEPILOGO						
Statistica della regressione						
R multiplo	0.967					
R al quadrato	0.935					
R al quadrato corretto	0.927					
Errore standard	1.261					
Osservazioni	20					
ANALISI VARIANZA						
	gdl	*SQ*	*MQ*	*F*	*Significatività F*	
Regressione	2	386.3129	193.1564	121.4702	8.56E-11	
Residuo	17	27.03264	1.590155			
Totale	19	413.3455				
	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività(p-value)*	*Inferiore 95%*	*Superiore 95%*
Intercetta	-1.1183	0.6549	-1.7077	0.1059	-2.4999	0.2633
REDDITO	0.1482	0.0164	9.0491	0.0000	0.1137	0.1828
NC	0.7931	0.2444	3.2446	0.0048	0.2774	1.3088

1) R^2 ed R^2 corretto (4)

L'R^2 e R^2 corretto dicono se quanto adatta è la relazione lineare tra le X e la Y

NON dicono se

- Una variabile inclusa sia statisticamente significativa
- Le variabili esplicative sono la vera causa della variabilità della variabile dipendente
- Il modello sia ben specificato

2) La multicollinearità

- Multicollinearità significa elevata correlazione fra le variabili esplicative X
- Se vi è multicollinearità, le variabili non forniscono informazioni aggiuntive ed è difficile valutare l'effetto di ciascuna di esse
- Le stime dei coefficienti presentano elevata variabilità (elevato errore standard o standard error SE)

2) La multicollinearità (2)

Come si rivela la presenza di multicollinearità?

- Esame della matrice di correlazione
  - La correlazione fra coppie di variabili X è più elevata di quella con la variabile Y
- Poche soluzioni
  - Utilizzare nuovi dati (di solito impraticabile)
  - **Eliminare una delle variabili X altamente correlate** (soluzione più usata)

2) La multicollinearità (3)

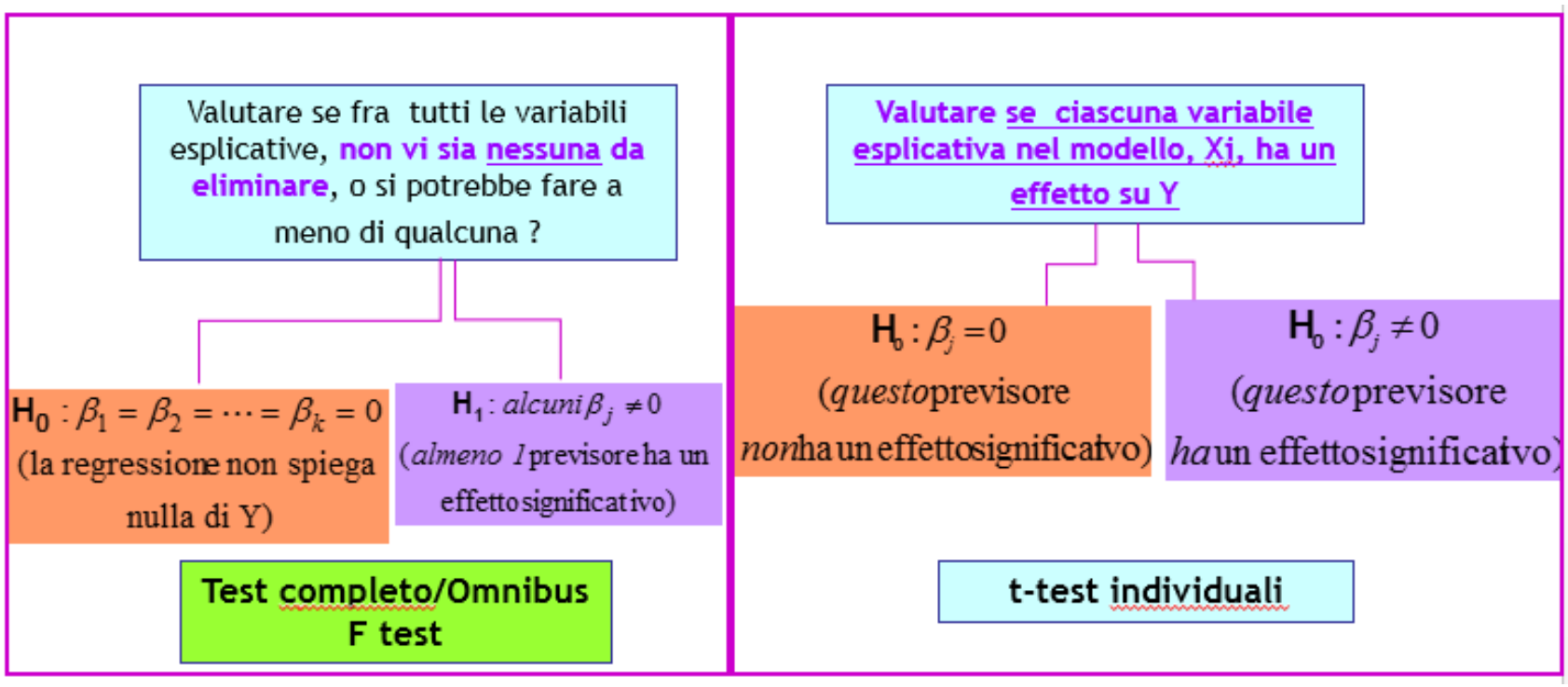
	SPESA	*REDDITO*	*NC*
SPESA	1		
REDDITO	0.9456	1	
NC	0.7871	0.6755	1

r_{YX1} r_{YX2} r_{X1X2}

Secondo voi vi è possibile collinearità tra X_1 e X_2?

3) Test sulla nullità dei coefficienti

Nella regressione multipla possono essere utilizzati due tipi di test di verifica delle ipotesi per la nullità dei coefficienti



3) Test sulla nullità dei coefficienti (2)

Con una sola variabile esplicativa (cioè nella regressione lineare semplice), questi due test **sono identici**. Nella regressione multipla, questi due test sono **decisamente differenti**!

3) Test sulla nullità dei coefficienti (3)

Test F per la significatività globale del modello

- Ipotesi (modello con k variabili esplicative):
  - $H_0 : \beta_1 = \beta_2 = \ldots = \beta_j = \ldots = \beta_k = 0$ (nessuna relazione lineare)
  - H_1 : Almeno un $\beta_j \neq 0$ (almeno una variabile esplicativa nel modello influenza la Y)
- Test F: Rapporto tra devianza della Y spiegata dal modello di regressione (detta SSR) e varianza della Y residua (non spiegata ossia devianza dei residui o SSE); se divido per i gradi di libertà ottengo MQR o (Mean square of regression o anche MSR) e MQE (Mean Square of Errors o anche MSE)
- Se tale rapporto è statisticamente maggiore di 1 allora rifiutiamo H_0
- come al solito valutiamo mediante l'approccio del valore p (p-value)

3) Test sulla nullità dei coefficienti (4)

OUTPUT RIEPILOGO						
Statistica della regressione						
R multiplo	0.967					
R al quadrato	0.935					
R al quadrato corretto	0.927					
Errore standard	1.261					
Osservazioni	20					
ANALISI VARIANZA						
	gdl	*SQ*	*MQ*	*F*	*Significatività F*	
Regressione	2	386.3129	193.1564	121.4702	8.56E-11	
Residuo	17	27.03264	1.590155			
Totale	19	413.3455				
	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività(p-value)*	*Inferiore 95%*	*Superiore 95%*
Intercetta	-1.1183	0.6549	-1.7077	0.1059	-2.4999	0.2633
REDDITO	0.1482	0.0164	9.0491	0.0000	0.1137	0.1828
NC	0.7931	0.2444	3.2446	0.0048	0.2774	1.3088

$$F_{2,17} = \frac{MQR}{MQE} = 121{,}4702$$

P-value per il test F

Dal valore p cosa decidereste rispetto all'ipotesi nulla? Cosa concludete quindi sul modello in questione?

3) Test sulla nullità dei coefficienti (5)

Vi è sufficiente evidenza per rifiutare H_0. Ciò significa che il modello di regressione multipla proposto non è una mera costruzione teorica, ma effettivamente esiste ed è statisticamente significativo. Infatti, vi è evidenza che almeno una variabile indipendente influenzi significativamente la Y

3) Test sulla nullità dei coefficienti (6)

Per quanto riguarda i t-test individuali sugli effetti delle singole variabili esplicative sulla Y l'approccio è analogo a quello della regressione semplice (valutazione della nullità dei singoli coefficienti di regressione)