Psicometria 1 (023-PS)

Michele Grassi
mgrassi@units.it

Università di Trieste

Lezione 21 22 23

Sommario

Sommario

Il materiale coperto è contenuto nel capitolo 9 del testo di Agresti e Finlay (2012)

- Relazioni tra variabili
- Diagrammi di dispersione
- Correlazione

Relazioni tra variabili

Relazioni tra variabili

I problemi più interessanti dell'analisi di dati riguardano le relazioni tra variabili.

- Vi sono differenze di genere nelle abilità cognitive?
- Le pene inflitte agli extra-comunitari sono più severe di quelle inflitte ai cittadini italiani?
- I tassi di mortalità infantile sono legati allo sviluppo economico nazionale?

Relazioni tra variabili

Quali sono i metodi che consentono di esaminare le relazioni tra variabili quantitative (come il livello di istruzione e il salario)?

- I diagrammi di dispersione sono dei grafici che visualizzano la relazione tra due variabili quantitative.
- Il coefficiente di correlazione misura il grado di associazione lineare tra due variabili quantitative.
- La regressione dei minimi quadrati è un metodo per trovare la retta che meglio riassume la relazione tra due variabili quantitative.

Variabili dipendenti e indipendenti

- Quando consideriamo la relazione tra due variabili, solitamente ipotizziamo che una variabile influenzi, almeno in parte, il comportamento dell'altra.
- Potremmo pensare, per esempio, che il livello di istruzione influenzi il reddito. In questo caso, diciamo che il reddito è la variabile dipendente (o variabile risposta) e il grado di istruzione è la variabile indipendente (detta anche esplicativa).

Variabili dipendenti e indipendenti

- Le variabili dipendenti e indipendenti sono facilmente distinguibili nella ricerca sperimentale, dove le varibili indipendenti sono direttamente manipolate dallo speriementatore.
- Per esempio, due gruppi di individui scelti a caso dalla popolazione vengono sottoposti a due diversi tipi di addestramento (*variabile indipendente*) e le prestazioni dei partecipanti vengono misurate in un certo compito (*variabile dipendente*).

Variabili dipendenti e indipendenti

- Nella maggior parte dei fenomeni psicologici, una variabile dipendente è influenzata da varie variabili esplicative.
  - Molti fattori oltre il livello di istruzione, per esempio, influenzano il reddito, come l'età, il genere, la razza, il luogo di residenza, eccetera.
- Nella ricerca non sperimentale è dunque importante riuscire a rendere conto degli effetti esercitati da molteplici variabili esplicative nei confronti di una variabile dipendente.

Variabili dipendenti e indipendenti

- Talvolta siamo interessati alla relazione tra due variabili senza distinguere tra variabile di risposta e variabile esplicativa (per esempio, componente verbale e componente matematica di un test d'abilità).
- Altre volte siamo interessati ad usare una variabile esplicativa per prevedere la variabile risposta, senza assumere un rapporto di *causa–effetto* tra le due.

Tipi di relazioni

- La relazione si dice positiva se all'aumentare o diminuire di x segue un aumento o decremento di y.
- La relazione si dice negativa se all'aumentare o diminuire di x segue rispettivamente un decremento o un incremento di y.
  - Es. Se all'aumento del livello di istruzione:
    - aumenta il reddito: *relazione positiva*;
    - diminuisce il livello di pregiudizio etnico: *relazione negativa*.

Tipi di relazioni

- La relazione si dice simmetrica se ad un cambiamento di x segue un corrispondente cambiamento in y e viceversa.
- La relazione si dice asimmetrica se ad un cambiamento di x segue un corrispondente cambiamento in y, ma non viceversa.

Diagrammi di dispersione

Diagrammi di dispersione

L'esame di un diagramma di dispersione rappresenta molto spesso il punto di partenza dell'analisi dei dati (in effetti dovremmo iniziare esaminando la distribuzione di ciascuna variabile separatamente prima di considerare la relazione tra variabili.)

- In un diagramma di dispersione i valori di una variabile (indipendente) sono rappresentati sull'ascissa (asse x) e i valori dell'altra variabile (dipendente) sono rappresentati sull'ordinata (asse y).
- Ciascuna osservazione è rappresentata nel grafico come un punto definito dai valori x_i, y_i delle due variabili.

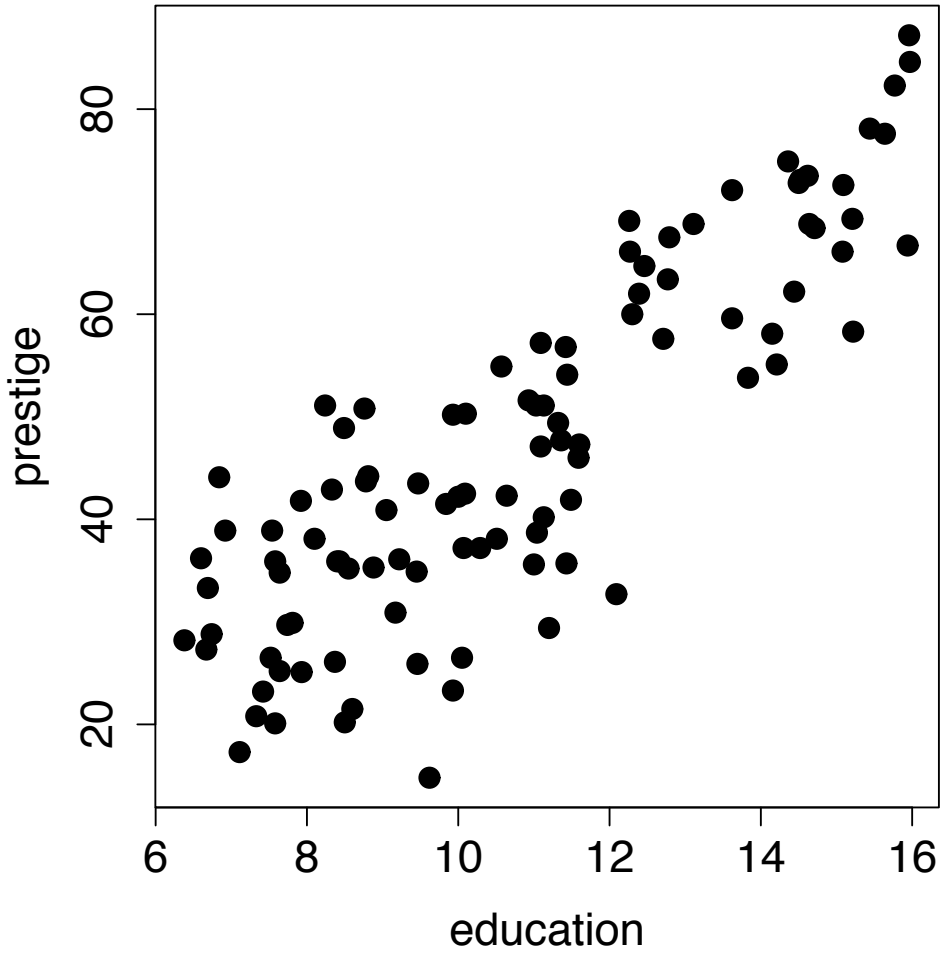
Diagrammi di dispersione

- Di seguito vengono riportati alcuni esempi di diagrammi di dispersione relativi ad un campione di dati in cui ciascuna osservazione rappresenta una professione.
- Sono rappresentate 102 professioni. A ciascuna osservazione sono associate 4 variabili.
- Le prime 3 derivano dal censimento canadese del 1971.

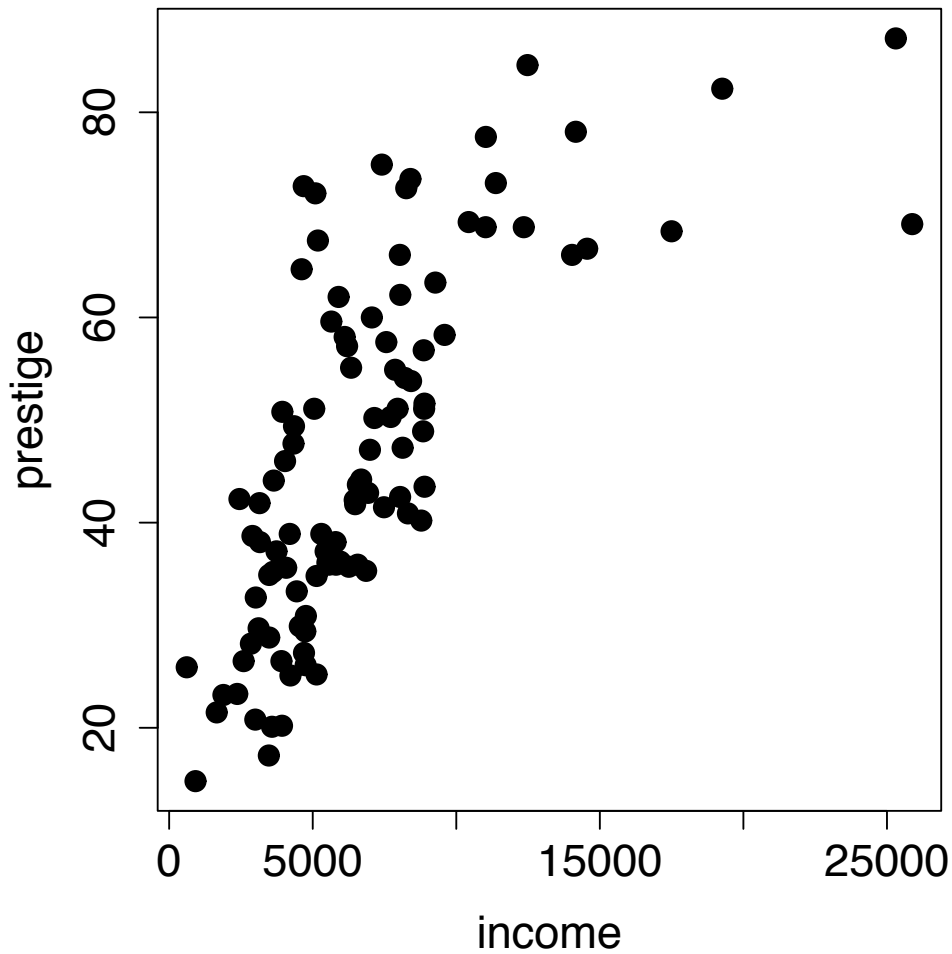
Diagrammi di dispersione

- *education* Average education of occupational incumbents, years, in 1971.
- *income* Average income of incumbents, dollars, in 1971.
- *women* Percentage of incumbents who are women.
- *prestige* Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
  - I dati sono contenuti nel *data frame* `Prestige` del *package* `car`.

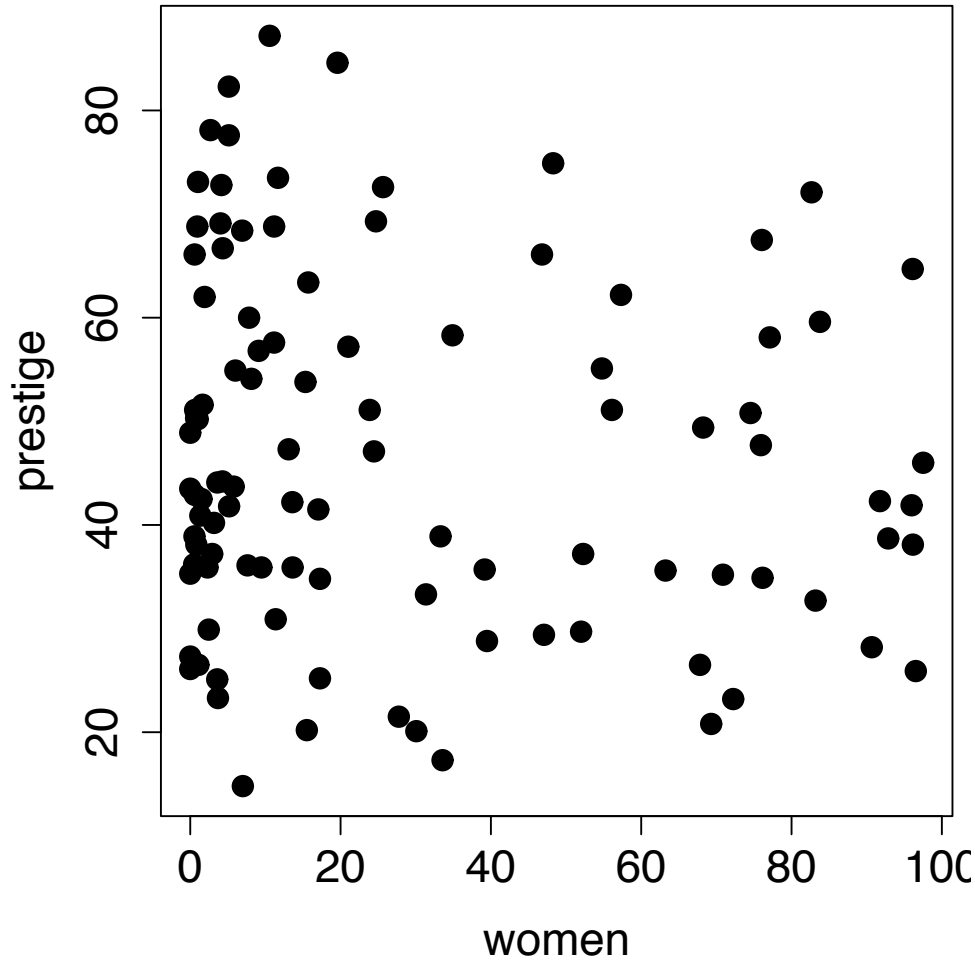
Diagrammi di dispersione



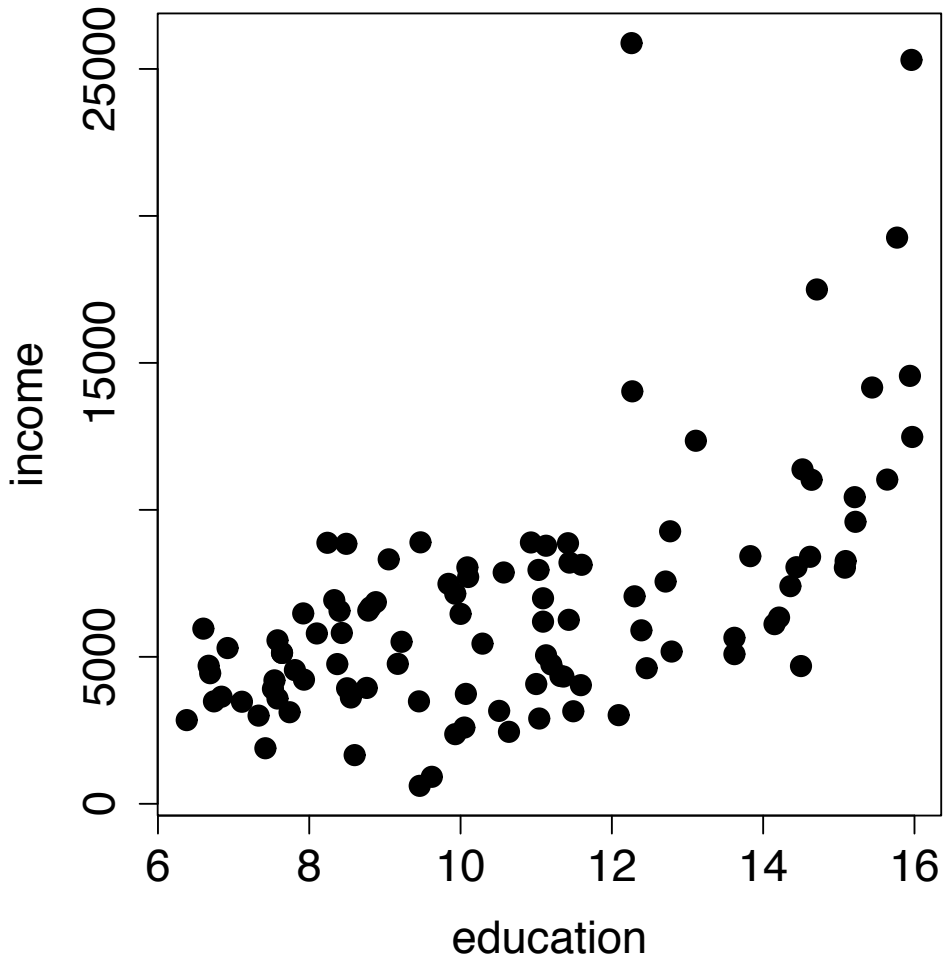
Diagrammi di dispersione



Diagrammi di dispersione



Diagrammi di dispersione



Diagrammi di dispersione

```
> library(car)
> data(Prestige)
> Prestige[1:10,]
                    education income women prestige census type
GOV.ADMINISTRATORS      13.11  12351 11.16     68.8   1113 prof
GENERAL.MANAGERS        12.26  25879  4.02     69.1   1130 prof
ACCOUNTANTS             12.77   9271 15.70     63.4   1171 prof
PURCHASING.OFFICERS     11.42   8865  9.11     56.8   1175 prof
CHEMISTS                14.62   8403 11.68     73.5   2111 prof
PHYSICISTS              15.64  11030  5.13     77.6   2113 prof
BIOLOGISTS              15.09   8258 25.65     72.6   2133 prof
ARCHITECTS              15.44  14163  2.69     78.1   2141 prof
CIVIL.ENGINEERS         14.52  11377  1.03     73.1   2143 prof
MINING.ENGINEERS        14.64  11023  0.94     68.8   2153 prof
```

Interpretazione

Per interpretare un diagramma di dispersione dobbiamo considerare l'andamento complessivo dei dati.

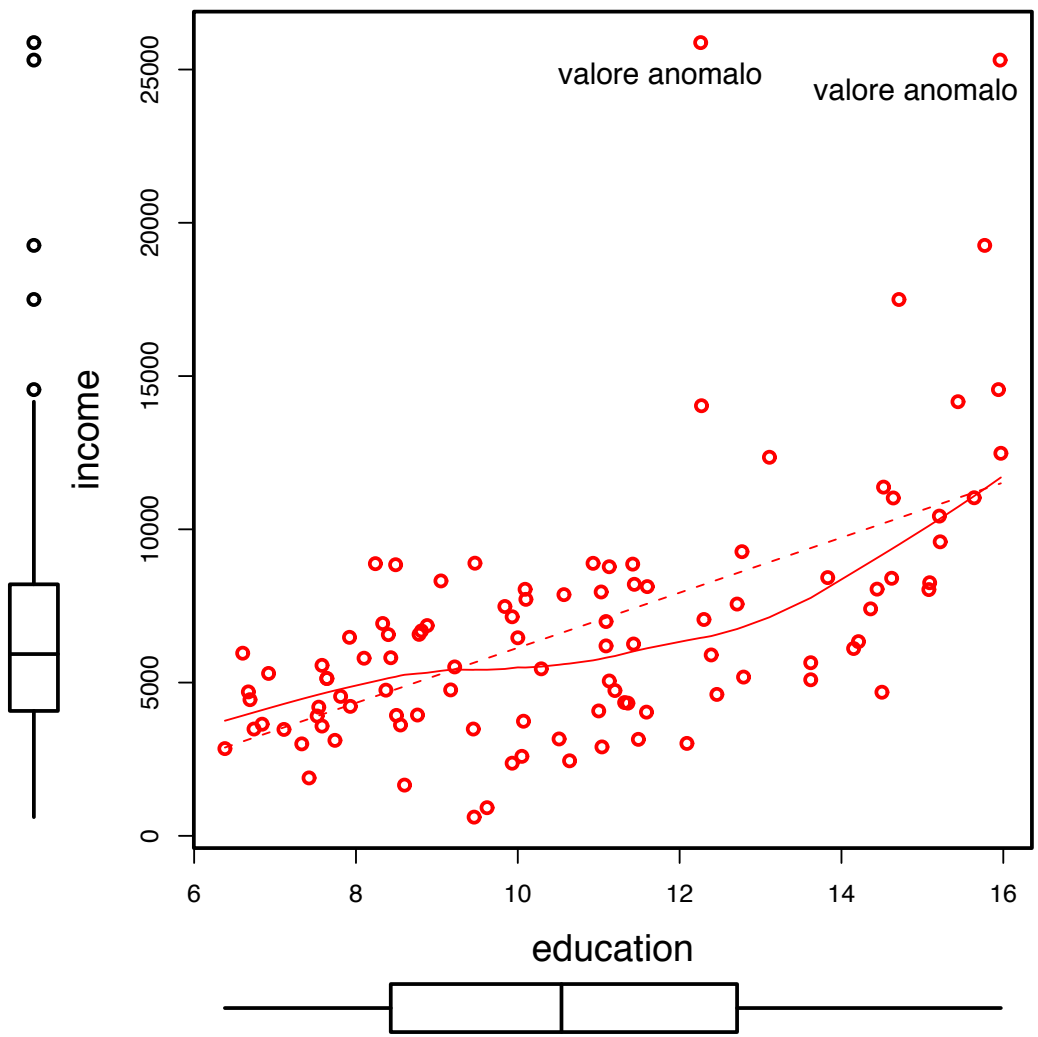
- Ci sono dei cluster di osservazioni nei dati?
- Ci sono dati anomali (outlier)? Un'osservazione anomala è un punto che si trova lontano dalla nuvola di punti che contiene la maggior parte dei dati.
- Il diagramma di dispersione rivela un'associazione tra le due variabili? Se sì, l'associazione ha una direzione?
- L'associazione può essere descritta da una retta oppure è chiaramente non lineare? Le relazioni lineari sono particolarmente semplici.

Dati anomali

Nel seguente diagramma di dispersione sono evidenziati due dati anomali.

```
> par(mar=c(4.5, 4.5, .5, .5), cex.lab=1.5, lwd=2)
> attach(Prestige)
> scatterplot(education, income)
> text(12, 25000, "valore anomalo")
> text(15, 24500, "valore anomalo")
> detach(Prestige)
```

Dati anomali

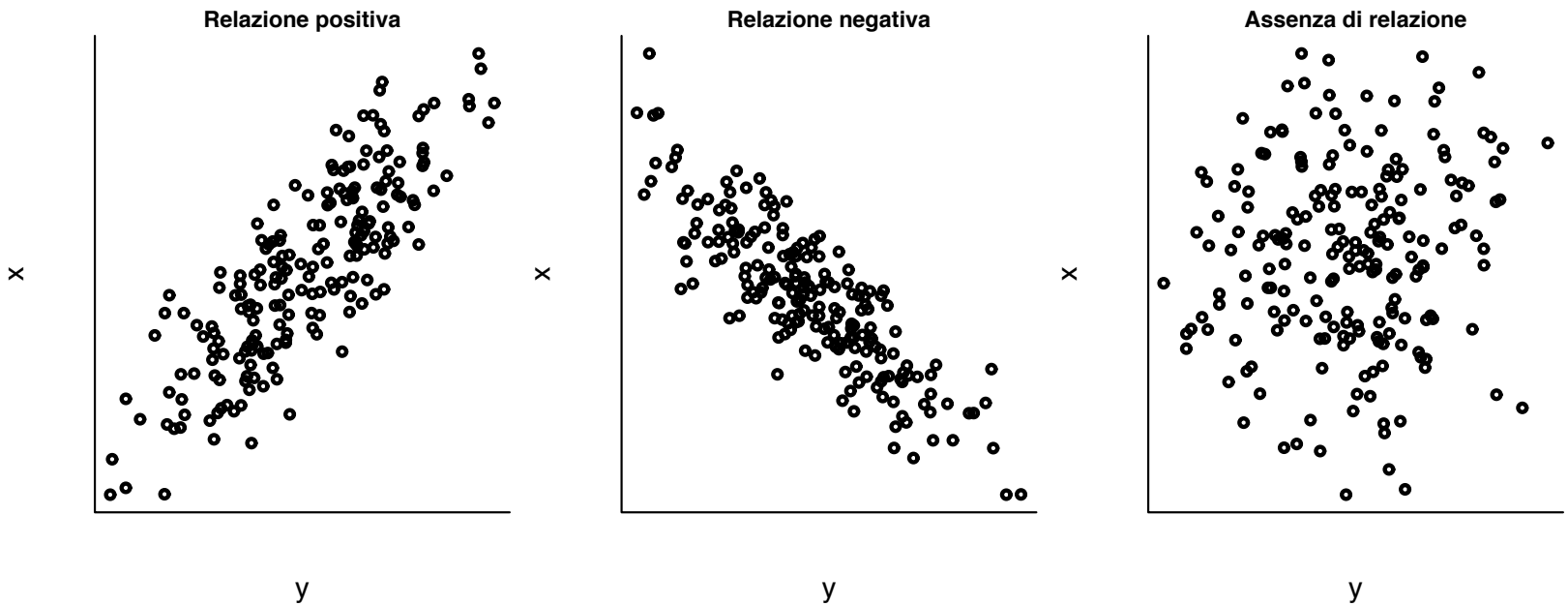


Dati anomali

- Si noti che nè il valore x nè il valore y di queste osservazioni, presi singolarmente, sono *anomali* rispetto a quelli delle altre osservazioni.
- In questo caso, è la *combinazione* dei valori x e y che fà in modo che questi due punti si distanzino dagli altri.

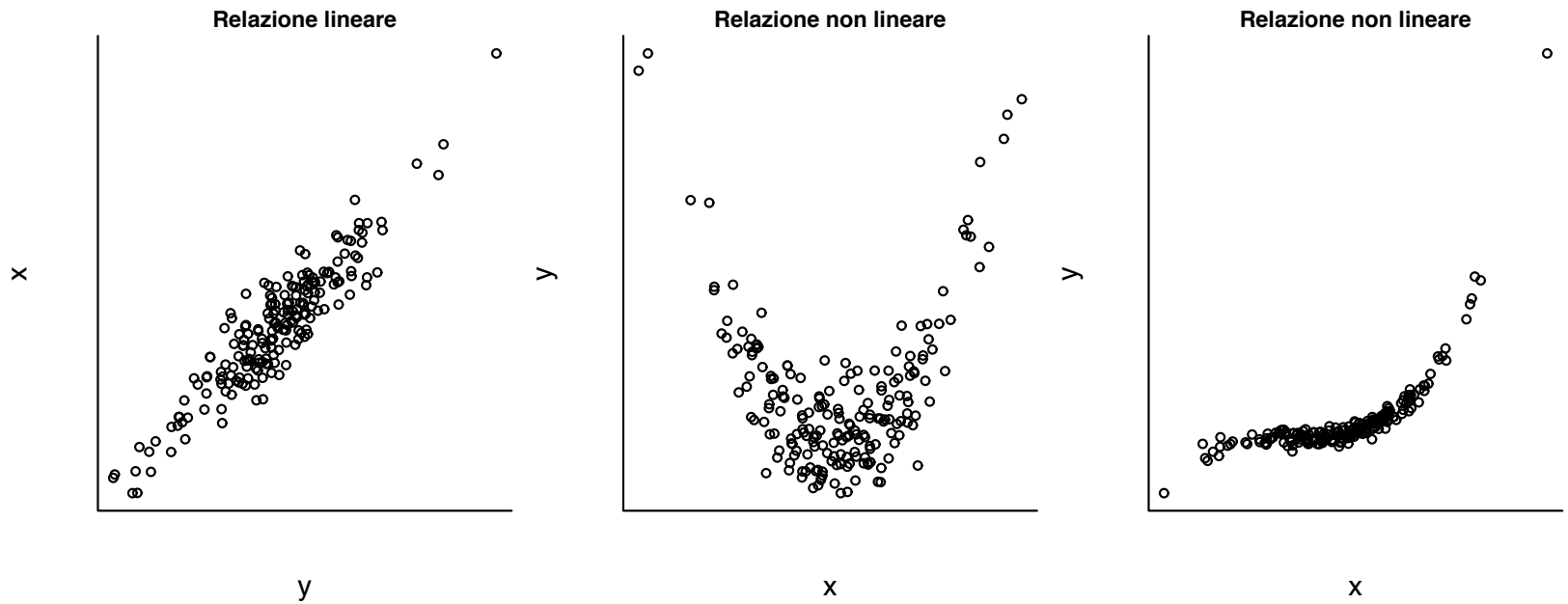
Associazione negativa e assenza di relazione

Vengono qui esemplificate le situazioni di associazione positiva, associazione negativa e assenza di relazione.



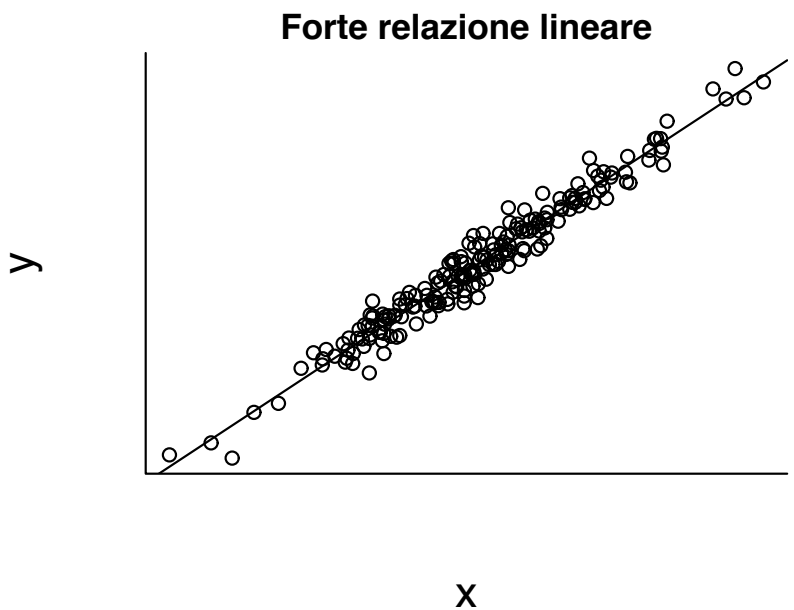
Relazione lineare e non lineare

Vengono qui esemplificate una relazione lineare e due relazioni non lineari.



Intensità della relazione lineare

Vengono qui esemplificate una forte e una debole relazione lineare.



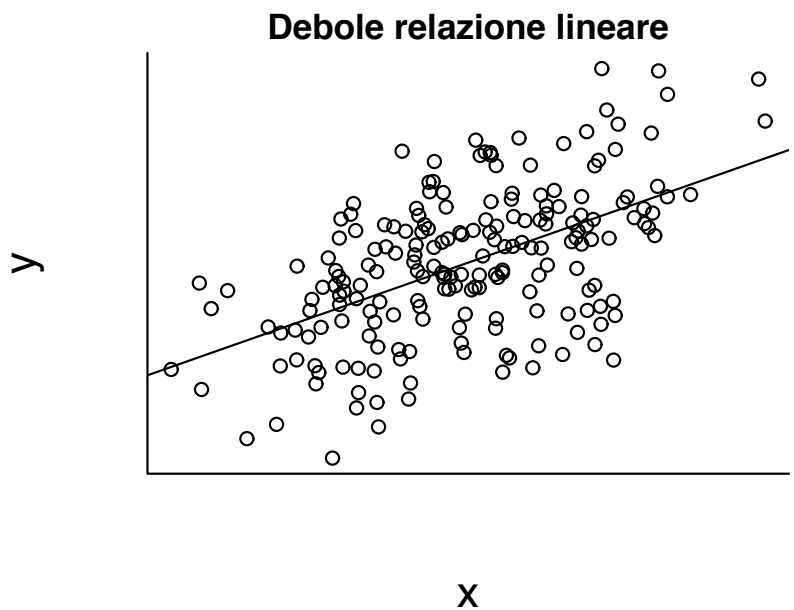
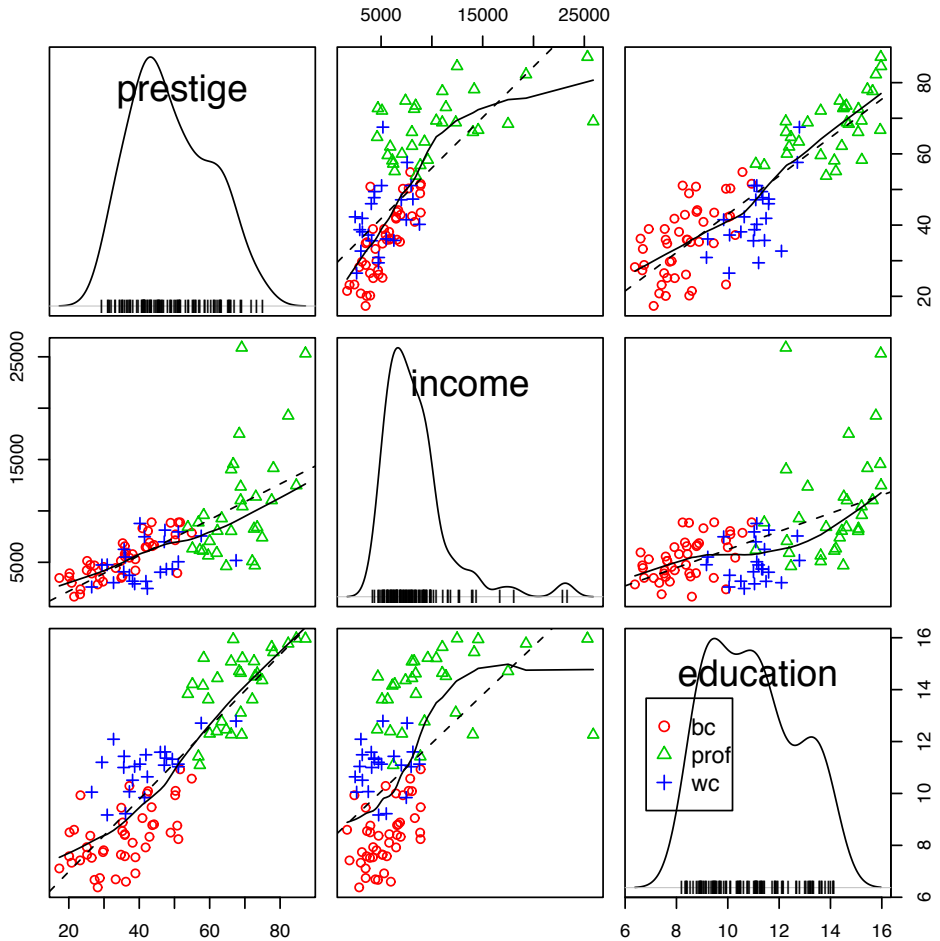


Diagramma di dispersione e fattori

- Usando simboli e colori diversi è possibile rappresentare un carattere esplicativo con modalità sconnesse (un fattore) all'interno di un diagramma di dispersione.
- Nel diagramma seguente, sono stati usati cerchi per la categoria *blue collars* (bc), croci per la categoria *whithe collars* (wc) e triangoli per *professionals* (prof).

Diagramma di dispersione e fattori



Correlazione

Correlazione

- Il coefficiente di correlazione r_{xy} misura la forza e la direzione di una relazione lineare tra due variabili x e y.
- Il coefficiente di correlazione r_{xy} è definito dalla seguente formula:

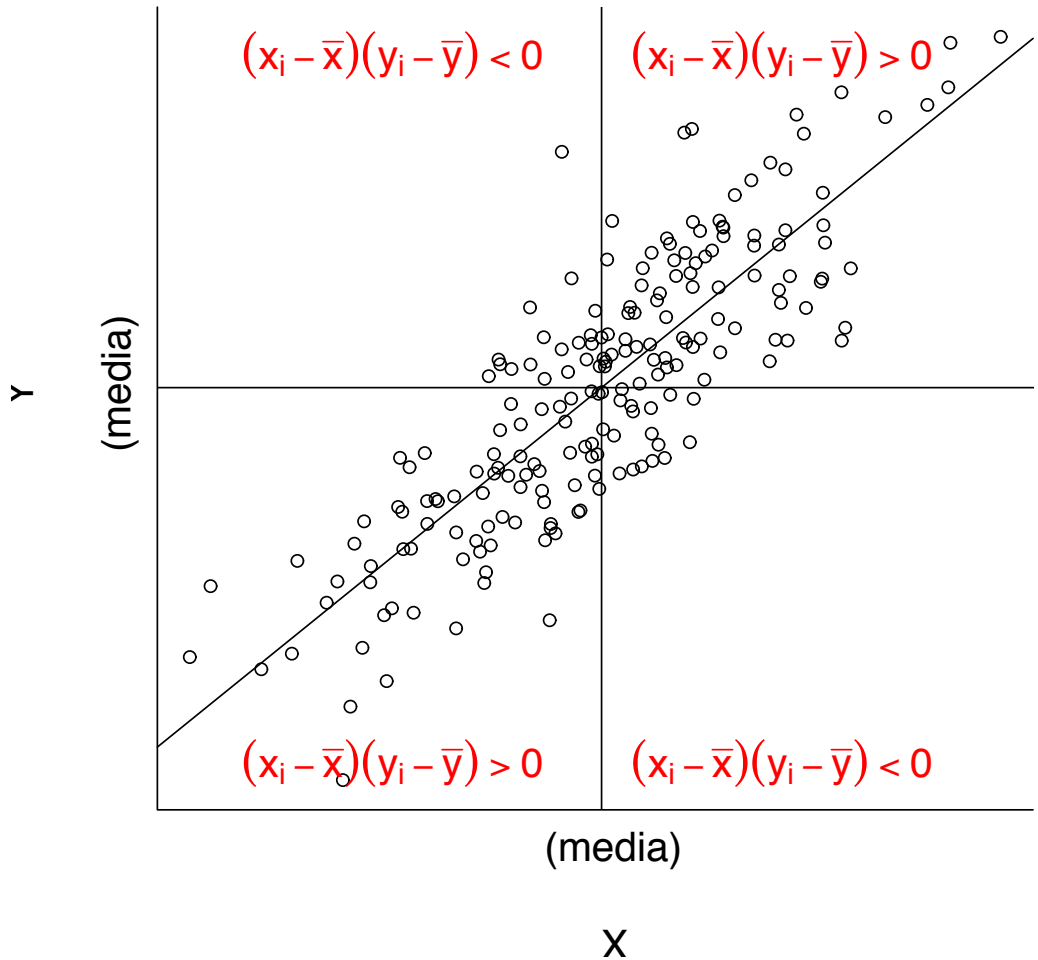
$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \frac{\sum z_x z_y}{n-1}$$

Correlazione

- s_x e s_y sono le deviazioni standard di x e y.
- $\left(\frac{x_i - \bar{x}}{s_x}\right)$ e $\left(\frac{y_i - \bar{y}}{s_y}\right)$ sono i punteggi standardizzati delle variabili x e y per l'osservazione i-esima. Si ricordi che questo non implica che x e y sono distribuiti normalmente.
- Il prodotto $\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ sarà positivo se sia x che y assumono valori superiori alle loro medie, o valori inferiori alle loro medie.
- Il prodotto $\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ sarà negativo se una variabile assume un valore superiore e l'altra un valore inferiore alla media.

Correlazione



Correlazione

- Il simbolo Σ significa che i prodotti $\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ sono sommati per tutte le osservazioni.
- Quando c'è una relazione positiva tra x e y la correlazione è positiva; quando c'è una relazione negativa tra x e y la correlazione è negativa.
- Si noti che la formula della correlazione non assegna un ruolo diverso alle variabili x e y. Il coefficiente di correlazione, dunque, non muta a seconda che la variabile x (o y) assuma il ruolo di variabile indipendente o di variabile dipendente.

Il coefficiente di correlazione r_{xy} varia tra -1 e $+1$.

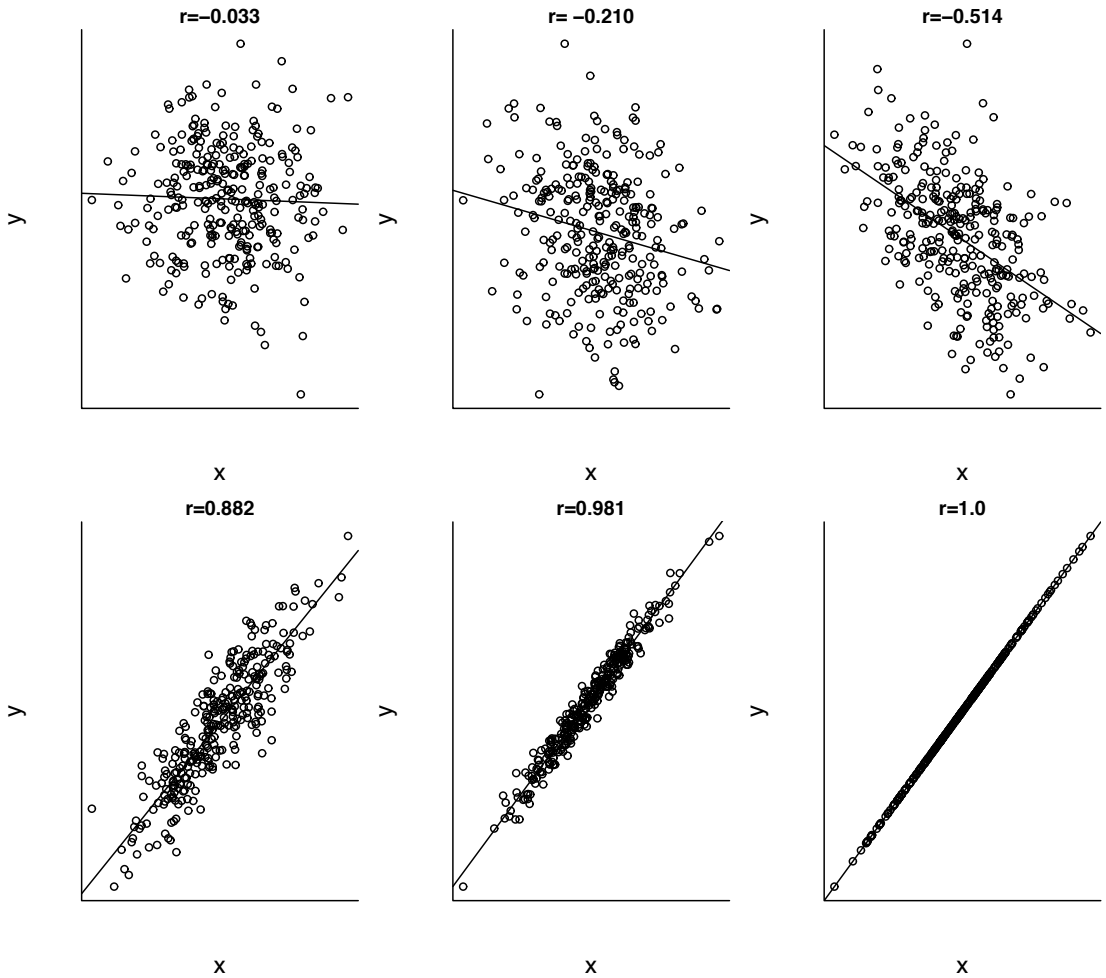
- $r_{xy} = -1$ perfetta relazione negativa: tutti i punti si trovano esattamente su una retta con pendenza negativa (dal quadrante in alto a sinistra al quadrante in basso a destra).
- $r_{xy} = +1$ perfetta relazione positiva: tutti i punti si trovano esattamente su una retta con pendenza positiva (dal quadrante in basso a sinistra al quadrante in alto a destra).
- $r_{xy} = 0$ assenza di relazione lineare tra x e y.

Correlazione

Valori $-1 < r_{xy} < +1$ indicano la presenza di una relazione lineare di intensità diversa.

- Dato che la correlazione è calcolata usando valori standardizzati, il suo valore non muta se vengono cambiate le unità di misura delle variabili x e y: l'indice di correlazione r_{xy} è privo di unità di misura.

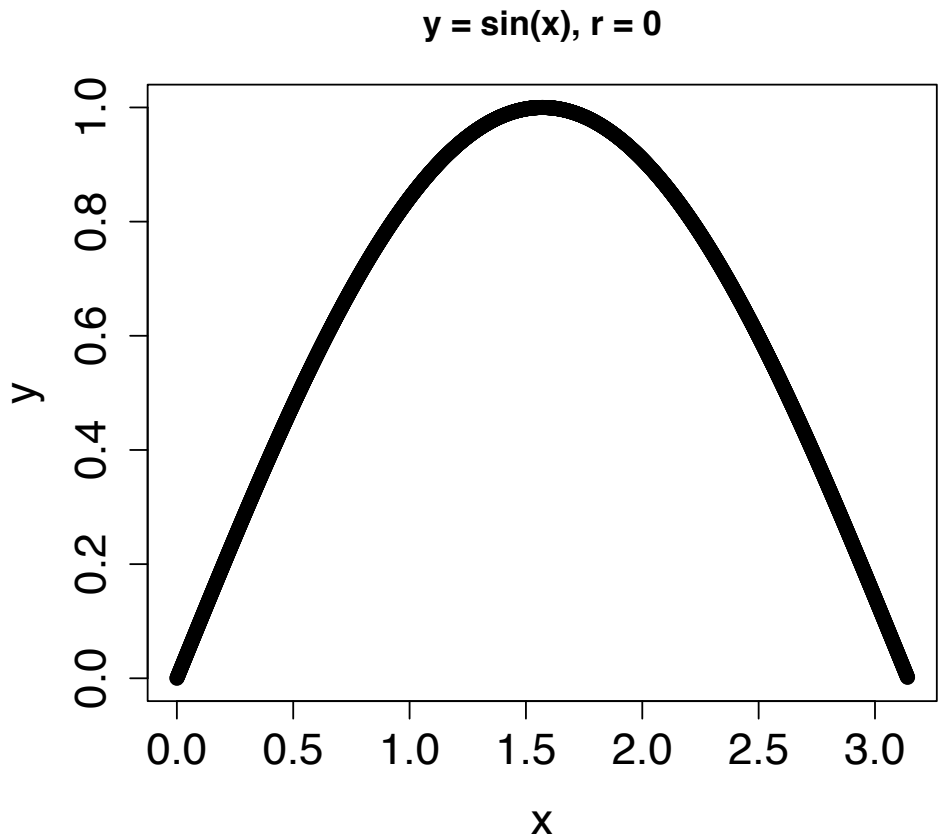
Correlazione



Correlazione

- Relazioni non lineari, anche di forte intensità, possono produrre una correlazione $r_{xy} = 0$.
- Anche se $r_{xy} \neq 0$, il coefficiente di correlazione non è appropriato per rappresentare il grado di associazione tra variabili nel caso di una relazione non lineare.

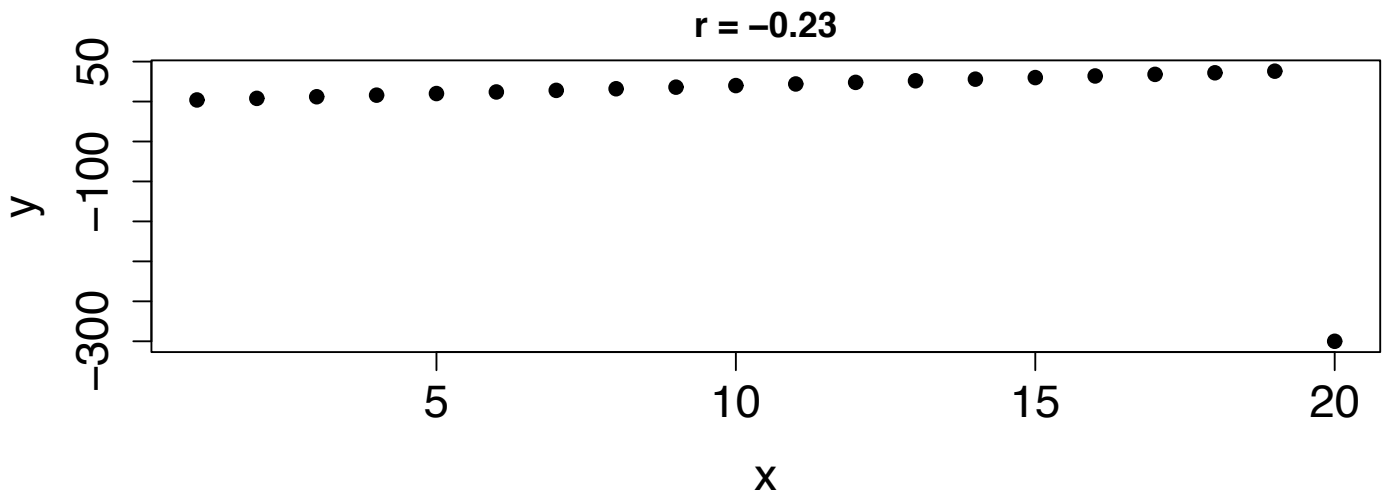
Correlazione



Correlazione

- Il coefficiente di correlazione r_{xy} è appropriato solo per variabili quantitative.
- Non ha senso calcolare la correlazione tra una variabile quantitativa e una variabile qualitativa (così come non avrebbe senso calcolare la media di una variabile qualitativa).

Come per la media e la varianza, anche il coefficiente di correlazione può essere fortemente influenzato da un unico dato anomalo (o da un piccolo numero di dati anomali).



Conclusioni

Conclusioni

Come si stabilisce l'esistenza di un'associazione tra due variabili?

- Il diagramma di dispersione consente di visualizzare la relazione tra due variabili e di stabilire se la relazione è approssimativamente lineare.
- Consente inoltre di individuare gli eventuali dati anomali.
- Il coefficiente di correlazione r di Pearson descrive l'intensità e la direzione della relazione lineare tra due variabili.

Regressione lineare semplice

Sommario

Sommario

Il materiale coperto è contenuto nel capitolo 9 del testo di Agresti e Finlay (2012)

- Regressione lineare semplice
- Medoto dei minimi quadrati
- Interpretazione della retta di regressione
- Regressione e correlazione
- Indice di determinazione
- Inferenza sul modello di regressione lineare
- Problemi del modello di regressione lineare

Regressione lineare semplice

Regressione lineare semplice

- Se la relazione tra una variabile dipendente (y) e una variabile indipendente (x) è lineare, allora è ragionevole rappresentare l'andamento di tale relazione con una retta.
- Ci occuperemo ora della regressione lineare semplice, ovvero del metodo comunemente usato per trovare l'orientamento della retta che meglio si adatta alla nuvola di punti rappresentata mediante un diagramma di dispersione.

Notazione Nella trattazione che segue, userò indifferentemente x o X per indicare la variabile indipendente, e y o Y per indicare la variabile dipendente.

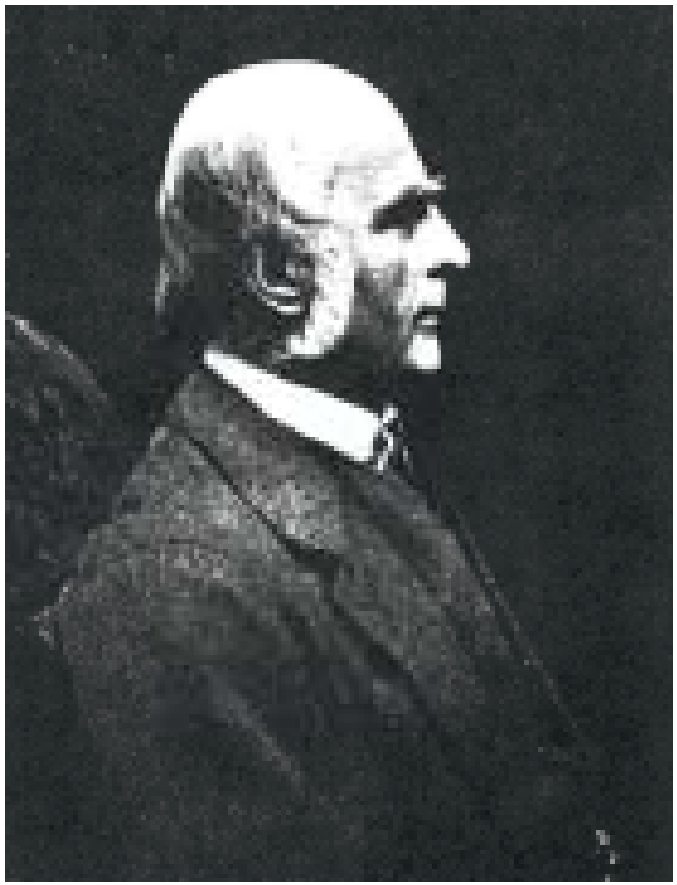
Regressione verso la media

- Il termine regressione fu introdotto da Sir Francis Galton (1822-1911) che, nei suoi studi di eugenica, voleva verificare se la statura dei figli potesse essere prevista sulla base di quella dei genitori.
- Analizzando i dati di circa 200 coppie in cui aveva potuto misurare l'altezza del padre e quella di un figlio maschio di circa 20 anni, Galton scoprì che i padri più alti della media avevano figli con un'altezza media minore di quella dei padri, e padri meno alti della media avevano figli con un'altezza media maggiore di quella dei padri.

Regressione verso la media

Nel suo articolo del 1886 "*Regression towards mediocrity in hereditary stature*" questo fenomeno fu visto come negativo in quanto ostacola la selezione di una popolazione "migliore".

Sir Francis Galton (1822-1911)



Regressione verso la media

- Fortemente colpito dalla lettura de *L'origine delle specie* scritto da suo cugino Charles Darwin, Galton era convinto che anche le "facoltà mentali", quali l'intelligenza, e i "valori morali", quali l'onestà, si trasmettessero per via ereditaria.
- Secondo Galton, il progresso della civiltà aveva ostacolato il libero corso della selezione naturale, permettendo la protezione e una riproduzione indefinita a esistenze "mediocri" e inducendo così un forte rischio di degenerazione.
- Fu quindi tra i fondatori dell'eugenetica, finalizzata a migliorare le "razze e le classi inferiori" attraverso misure tese ad evitare il diffondersi dei caratteri ereditari indesiderati.

Excursus

- La nozione di "razza" viene solitamente definita nei termini del colore della pelle degli individui.
- Tale caratteristica definiente è accompagnata dalla presupposizione che le cause genetiche e biologiche responsabili di tale carattere consentono di dividere la specie umana in razze separate.
- Tuttavia, è stato dimostrato che non esistono variazioni genetiche di tale entità da giustificare la divisione della specie umana in "razze".
- La nozione di "razza" è un concetto sociologico, non un costrutto biologico.

Regressione verso la media

- In seguito, dal suo significato originario di "ritornare indietro" verso la media e verso "la mediocrità", il termine regressione assunse solo il significato neutro di funzione che esprime matematicamente la relazione tra
  - la variabile attesa o predetta o teorica, indicata con $\hat{Y}$
  - la variabile empirica o esplicativa, indicata con X .
- Mentre Galton aveva descritto la regressione in termini geometrici, Karl Pearson, a sua volta maestro di Fisher e di Gosset, arrivò alla formulazione dell'indice di correlazione prodotto-momento che utilizziamo a tutt'oggi.

Karl Pearson (1857-1936)

Regressione verso la media

In realtà, il fenomeno di regressione verso la media è un fenomeno statistico e può essere descritto confrontando le prestazioni ottenute in un pre-test e un post-test.

- Supponiamo di scegliere il 10% degli studenti che in un pre-test hanno ottenuto i punteggi più bassi.
- Nel post-test è probabile che molti di questi studenti continuino ad essere nel decimo percentile inferiore, ma anche se solo alcuni ottengono un voto più alto (per caso), allora nel post-test la media di questo gruppo sarà più vicina alla media della popolazione che nel pre-test.

Regressione verso la media

- La regressione verso la media si verifica perché abbiamo scelto un campione non-casuale dalla popolazione (ovvero, un campione selezionato nella coda inferiore della distribuzione).
- Se avessimo scelto un campione casuale dalla popolazione, i punteggi medi al pre-test e al post-test sarebbero stati simili.
- In questo caso, dato che al pre-test il punteggio del campione è già molto simile a quello della popolazione, al post-test non può regredire verso la media della popolazione.

Regressione verso la media

- La regressione verso la media è un fenomeno relativo: ci sarà un "miglioramento" per gli individui con i punteggi più bassi nel pre-test e, corrispondentemente, un "peggioramento" per gli individui con i punteggi più alti nel pre-test.
- L'entità della regressione verso la media dipende da quanto sono estremi i quantili che scegliamo: se i campioni selezionati hanno una media molto diversa dalla media della popolazione, la variazione dal pre-test al post-test sarà maggiore.

Regressione verso la media

- L'altro fattore che influenza l'entità della regressione verso la media è la correlazione tra le due variabili.
- Se le due variabili sono perfettamente correlate, allora non ci sarà regressione verso la media.
- Soltanto se vi è errore casuale, allora si verificherà il fenomeno della regressione verso la media come conseguenza del campionamento dai quantili più estremi.

Illustrazione

```
> pre.test <- 50 + 10*rnorm(200)
> summary(pre.test)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.31   40.99   49.72   48.58   55.92   73.48
>
> post.test <- 50 + 10*rnorm(200)
> summary(post.test)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  31.67   43.88   50.21   50.42   56.96   77.32
>
> dati <- data.frame(pre.test, post.test)
> dati[1:3, ]
  pre.test post.test
1 45.42602  52.28430
2 69.64085  47.12002
3 60.57446  42.99103
```

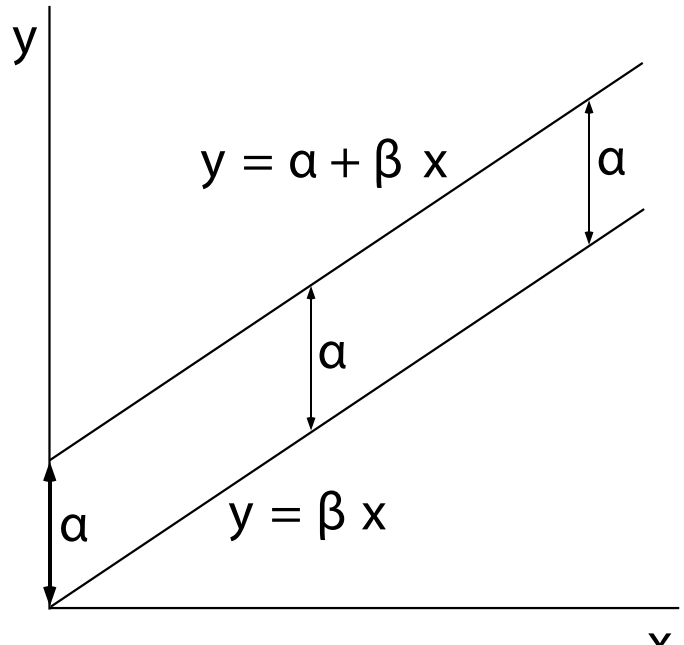
Illustrazione

```
> quantile.inf <- subset(dati, pre.test < 40)
> # nel post-test i punteggi sono migliorati
> mean(quantile.inf)
 pre.test post.test
 34.85926  52.98482
> dim(quantile.inf)
[1] 43  2
>
> quantile.sup <- subset(dati, pre.test > 60)
> # nel post-test i punteggi sono peggiorati
> mean(quantile.sup)
 pre.test post.test
 63.35422  47.92948
> dim(quantile.sup)
[1] 29  2
```

Funzione lineare

Funzione lineare

- La funzione $y = \beta x$ rappresenta una relazione di proporzionalità diretta tra la variabile x e la variabile y.
- La funzione $y = \alpha + \beta x$ somma una costante α a ciascuno dei valori y definiti dalla funzione precedente.



Funzione lineare

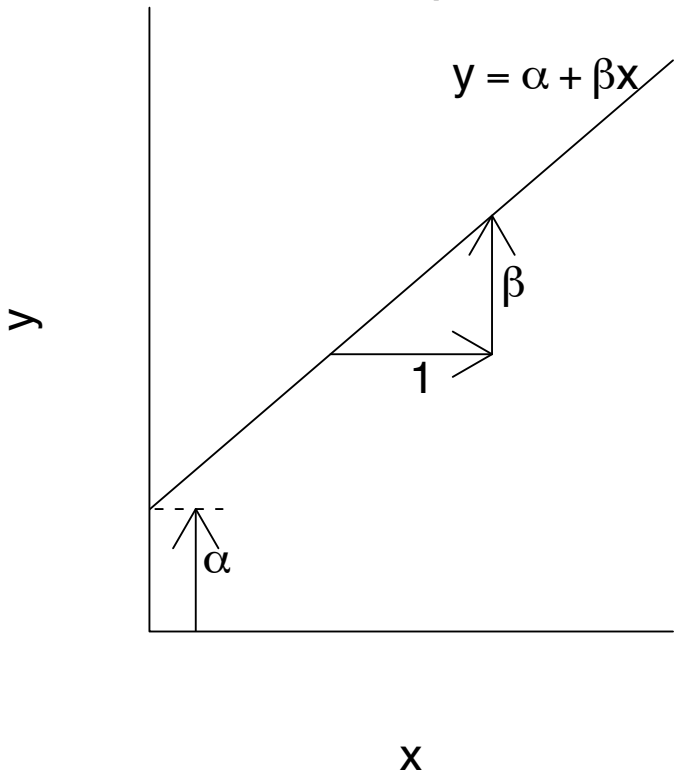
Una retta è definita dall'equazione

$$y = \alpha + \beta x$$

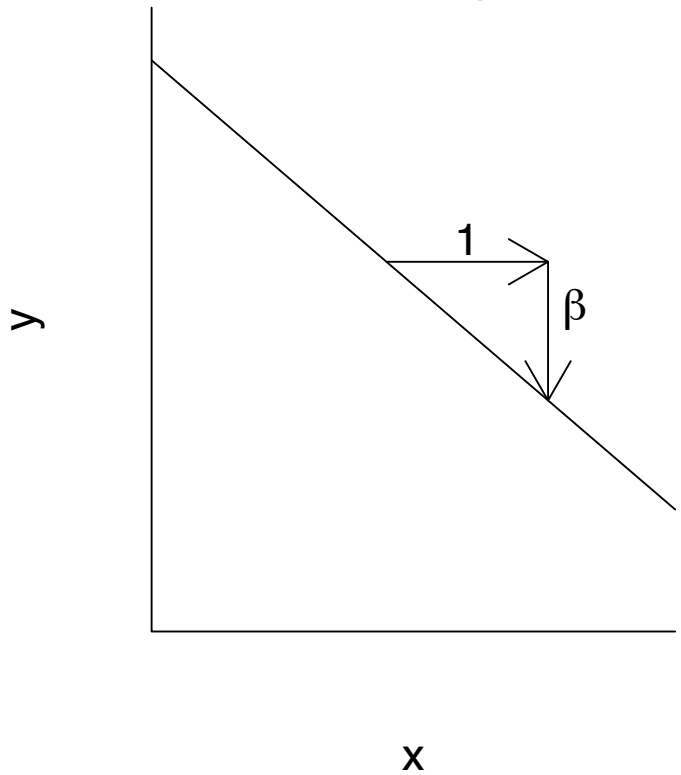
- Il parametro α è chiamato intercetta e indica il valore y corrispondente a $x = 0$.
- Il parametro β è chiamato pendenza o coefficiente angolare ed esprime la variazione di y per una variazione unitaria della variabile x.
  - Se β è positivo, allora il valore y aumenta al crescere di x.
  - Se β è negativo, allora il valore y diminuisce al crescere di x.
  - Se $\beta = 0$, allora la retta è orizzontale – il valore y non cambia al variare di x.

Funzione lineare

Pendenza positiva



Pendenza negativa



Consideriamo più attentamente la pendenza della retta.

- Le differenze

$$\Delta x = x - x_0$$

$$\Delta y = y - y_0$$

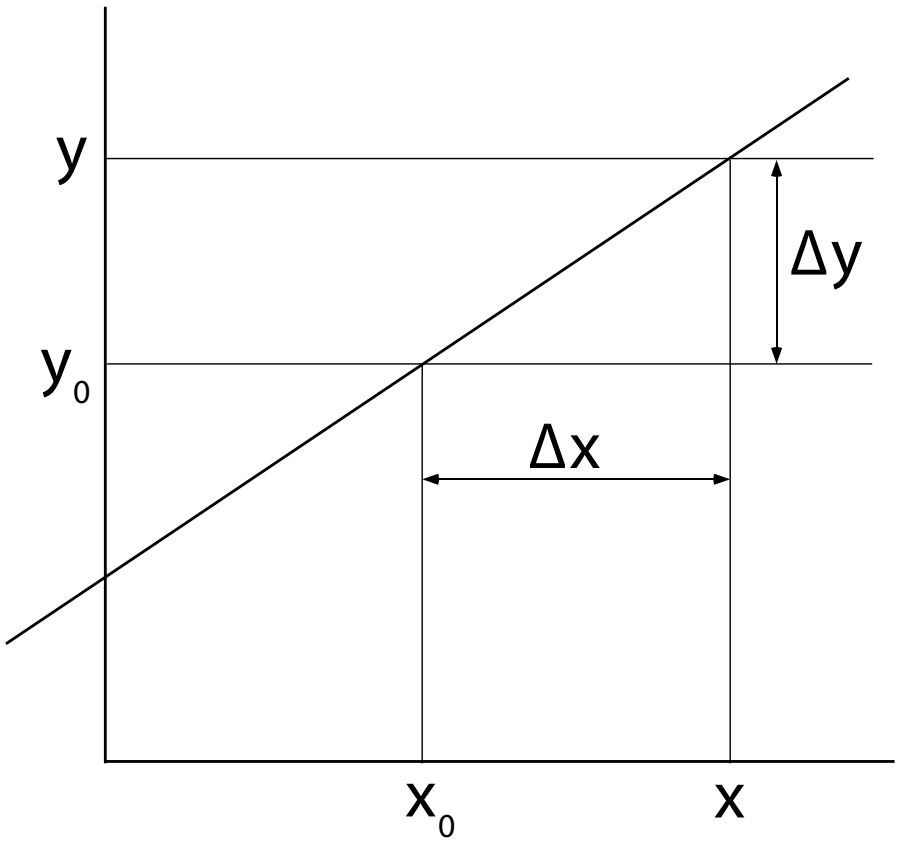
sono detti incrementi di x e y.

- La pendenza della retta, β, è uguale a:

$$\beta = \frac{\Delta y}{\Delta x} = \frac{(y - y_0)}{(x - x_0)}$$

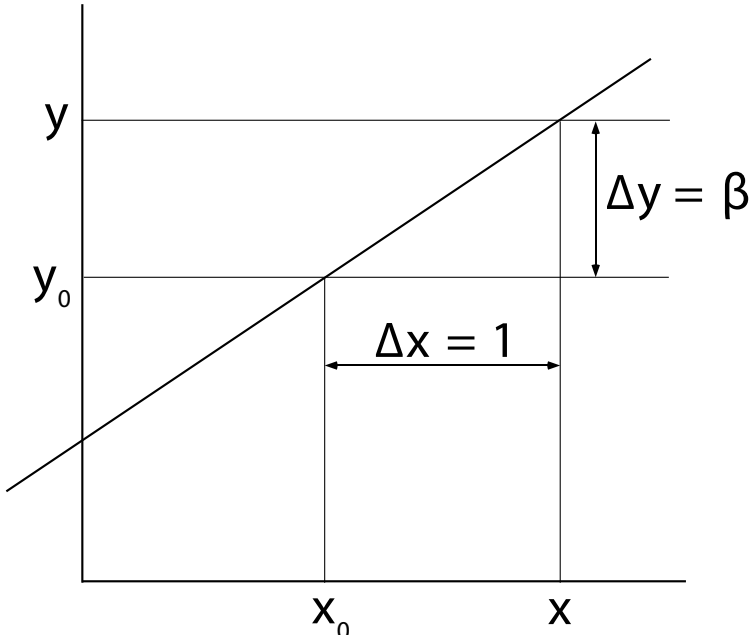
indipendentemente dalla grandezza degli incrementi x e y.

Funzione lineare



Se poniamo $\Delta x = 1$, la nozione di coefficiente angolare della retta assume l'interpretazione più semplice:

la pendenza della retta, β, è uguale all'incremento Δy associato all'incremento unitario $\Delta x = 1$.



Metodo dei minimi quadrati

Metodo dei minimi quadrati

A meno che la relazione tra X e Y non sia perfetta (il che non è mai il caso con dati reali), non è possibile trovare un'unica retta che coincida con tutti i punti del diagramma di dispersione.

- Quando vi è una forte relazione lineare tra X e Y è facile stabilire visivamente qual è la retta che meglio approssima la distribuzione dei punti.
- Questo non succede, invece, quando la relazione tra X e Y è debole, come solitamente avviene con dati psicologici. Abbiamo dunque bisogno di un metodo per trovare la retta che meglio si adatta ai dati e che non dipenda dal nostro giudizio soggettivo.

Metodo dei minimi quadrati

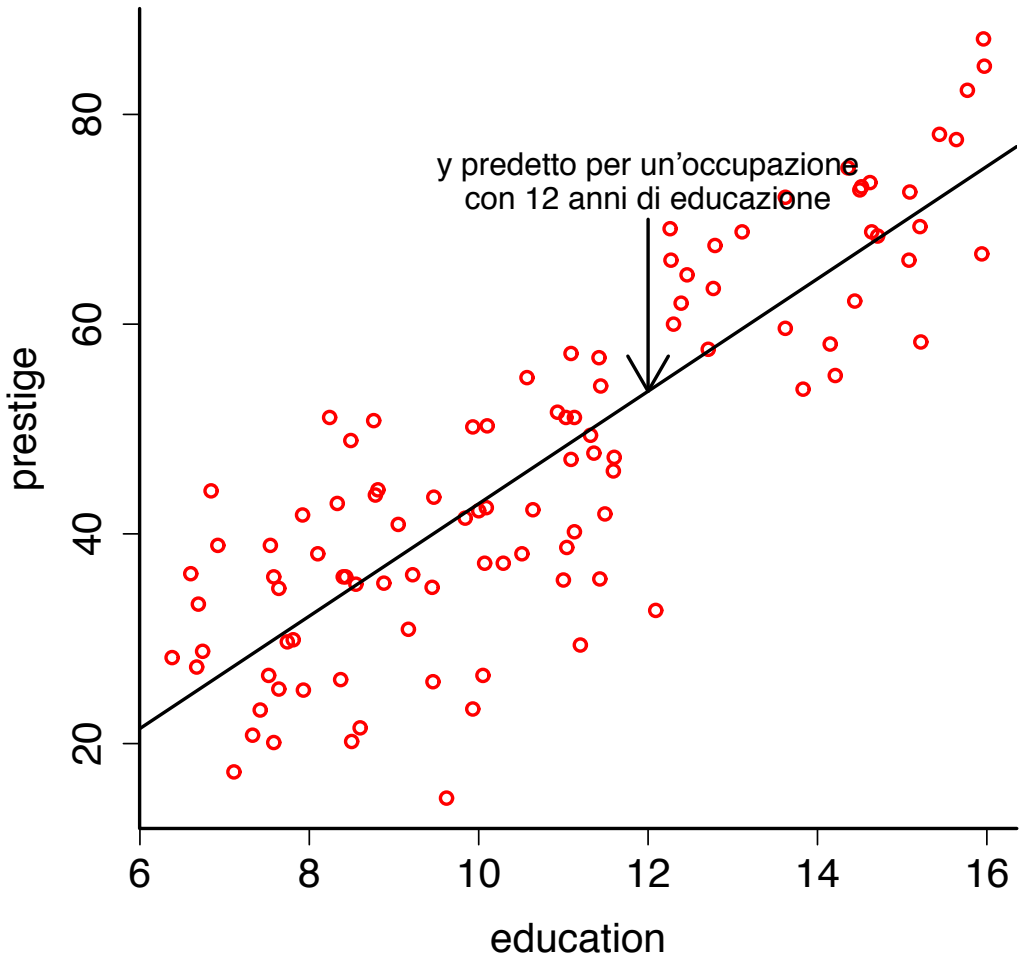
Illustrazione. Consideriamo, per esempio, la relazione desunta dai dati del censimento canadese del 1971 tra il prestigio delle professioni e il livello di istruzione.

- Vogliamo trovare la retta che che interpola al meglio i punti del diagramma di dispersione.
- Tale retta può essere impiegata per la previsione del valore della Y (*prestigio*) per ciascuno specifico valore di X (*istruzione*).

Illustrazione

- Per esempio, il valore Y teorico, denotato da $\hat{Y}$, per una professione che richiede un livello di istruzione pari a 12 anni, corrisponderà al punto della retta di regressione in corrispondenza del valore $X = 12$, in questo caso, $\hat{Y} = 53.6$.
- Si noti che il valore teorico $\hat{Y}$ è solitamente diverso dal valore osservato Y.
- In questo campione, adirittura, non c'è nessun valore osservato in corrispondenza di $X = 12$.

Illustrazione



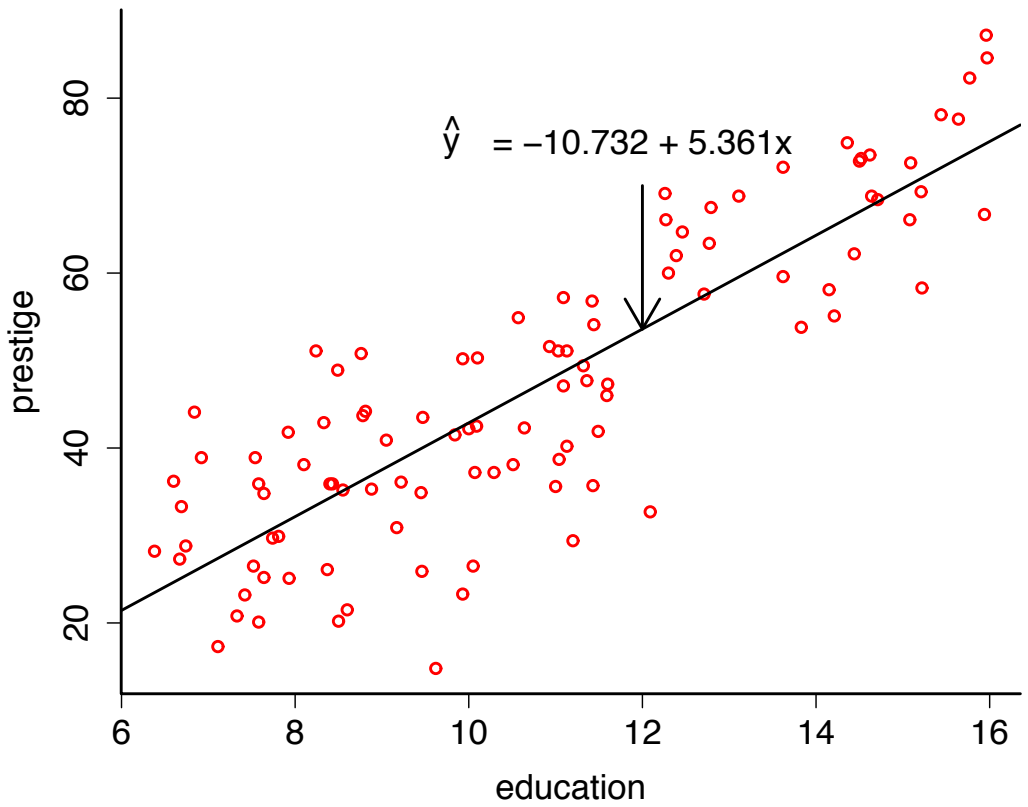
Illustrazione

La retta di regressione stimata è

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X = -10.732 + 5.361X$$

dove $\hat{\alpha} = -10.732$ e $\hat{\beta} = 5.361$ sono le stime dei minimi quadrati dell'intercetta e della pendenza della retta di regressione che interpola al meglio i dati.

Illustrazione



Calcolo delle stime dei minimi quadrati

- Come vengono calcolate le stime dei parametri della retta di regressione?
- Per stimare α e β con il metodo dei minimi quadrati si considerano le distanze dei punti (X_i, Y_i) dalla retta di regressione

$$Y_i - (\alpha + \beta X_i)$$

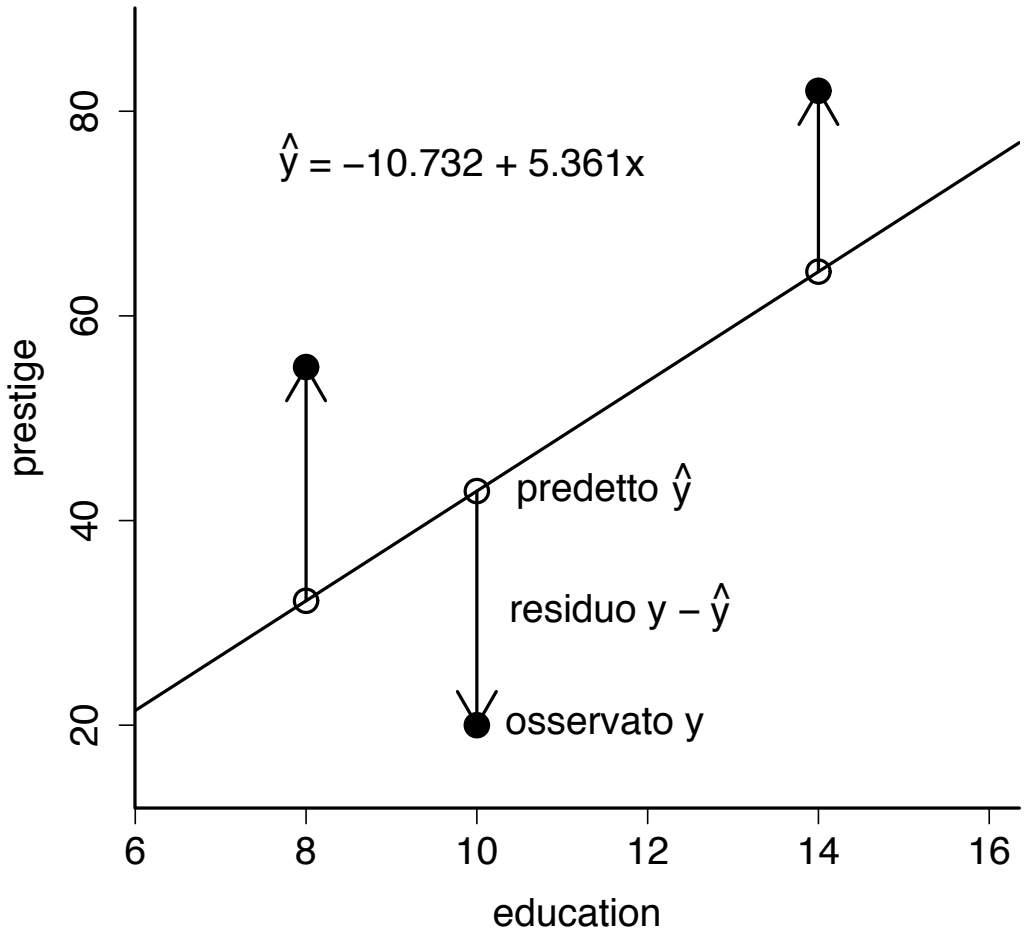
- Le stime di α e β sono scelte in modo tale da minimizzare le distanze dei punti (X_i, Y_i) dalla retta di regressione stimata.

Calcolo delle stime dei minimi quadrati

Poiché alcune distanze sono positive e altre negative, si considera la somma delle distanze al quadrato

$$\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2$$

Calcolo delle stime dei minimi quadrati



Calcolo delle stime dei minimi quadrati

Le stime dei minimi quadrati sono quei valori $\hat{\alpha}$ e $\hat{\beta}$ che minimizzano la somma dei quadrati delle distanze $Y_i - \hat{Y}_i$:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$$

dove

- r è il coefficiente di correlazione;
- s_{xy} è la covarianza;
- s_y e s_y sono le deviazioni standard di Y e X ;
- $\bar{Y}$ e $\bar{X}$ sono le medie di Y e X .

Interpretazione della retta di regressione

Interpretazione della retta di regressione

La retta di regressione stimata $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ definisce un legame in media fra la variabile dipendente e quella esplicativa.

L'intercetta α della retta di regressione è il valore atteso della variabile dipendente quando $X = 0$.

La pendenza β esprime la variazione del valore atteso della variabile dipendente per una variazione unitaria della variabile esplicativa.

Interpretazione della retta di regressione

Illustrazione. Consideriamo l'interpretazione della retta di regressione stimata

$$\hat{Y} = -10.733 + 5.361X$$

per l'esempio precedente.

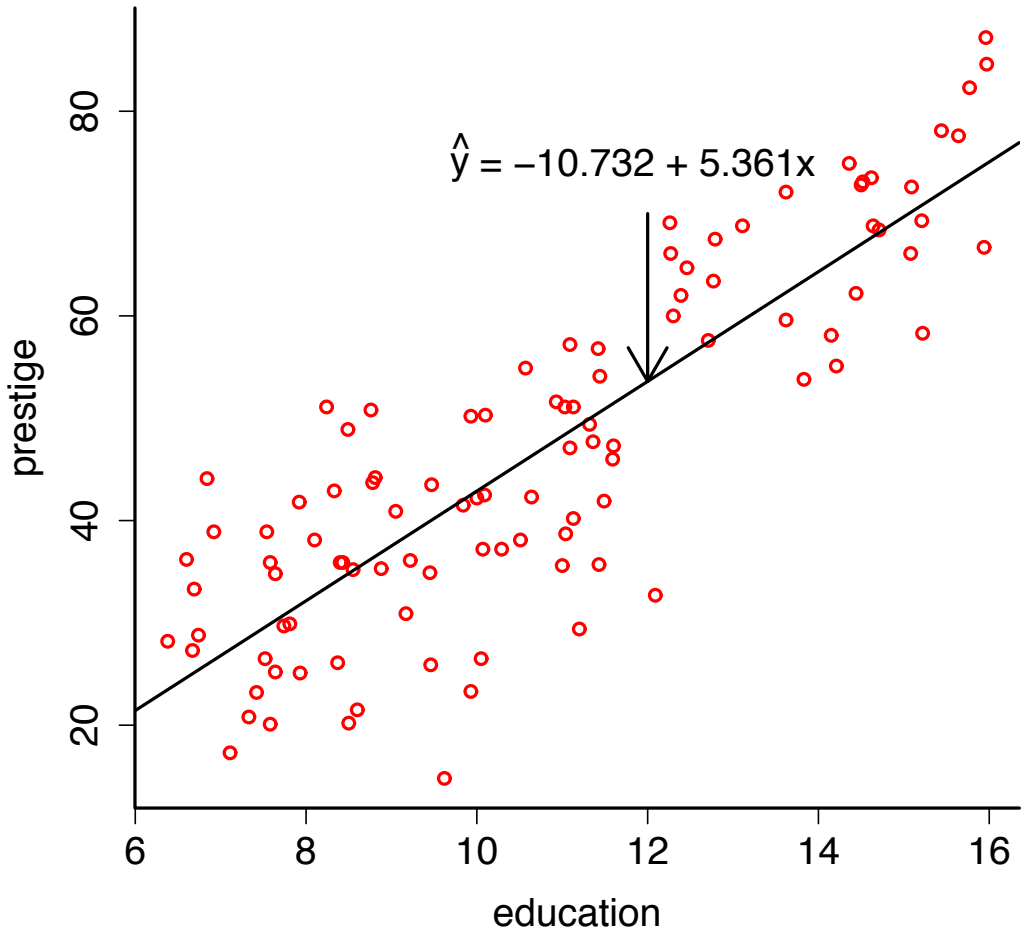
$\hat{\alpha} = -10.733$ è il valore teorico della variabile "prestigio" per una professione con 0 anni di istruzione media.

$\hat{\beta} = 5.361$: un incremento unitario (un anno) nel livello di istruzione provoca una variazione positiva di 5.361 punti nella media della variabile dipendente (scala di prestigio Pineo-Porter).

Interpretazione della retta di regressione

- Dato che gli assi del diagramma di dispersione non hanno l'origine $(0, 0)$, l'intercetta non compare nel grafico.
- Nel caso presente, non si deve interpretare il valore $\hat{\alpha}$ alla lettera in quanto
  - per tutte le 102 professioni, l'istruzione media è di almeno 6 anni,
  - i punteggi della scala Pineo-Porter non possono essere negativi.
- In generale, anche se la retta di regressione è adeguata per riassumere la relazione tra X e Y, è sempre pericoloso estrapolare tale relazione al di là della gamma di valori che è stata effettivamente osservata.

Interpretazione della retta di regressione



Rappresentazione grafica

Noti i valori dell'intercetta $\hat{\alpha}$ e del coefficiente angolare $\hat{\beta}$, è possibile procedere alla rappresentazione grafica della retta.

- Per la regressione del prestigio sull'istruzione, per esempio, troviamo i valori $\hat{Y}$ corrispondenti a $X = 6$ e $X = 16$:

$$X = 6 \quad \hat{Y} = -10.733 + 5.361 \times 6 = 21.433$$

$$X = 16 \quad \hat{Y} = -10.733 + 5.361 \times 16 = 75.043$$

- Colleghiamo poi i punti $(6, 21.433)$ e $(16, 75.043)$.

Rappresentazione grafica

- E’ importante ricordare che la retta passa sempre dal baricentro del diagramma di dispersione, individuato dal punto d’incontro delle due medie $\bar{X}$ e $\bar{Y}$.

Regressione lineare con R

Regressione lineare con R

- L'analisi della regressione lineare si esegue in **R** mediate la funzione `lm()`:

```
> fit <- lm( y ~ x )
```

- Notate la sintassi di `lm( y ~ x)`. A sinistra di ~ vi è il regressore a destra la variabile esplicativa.
- La costante α è automaticamente inclusa, ovvero il modello utilizzato da **R** è:

$$Y = \alpha + \beta X + \varepsilon$$

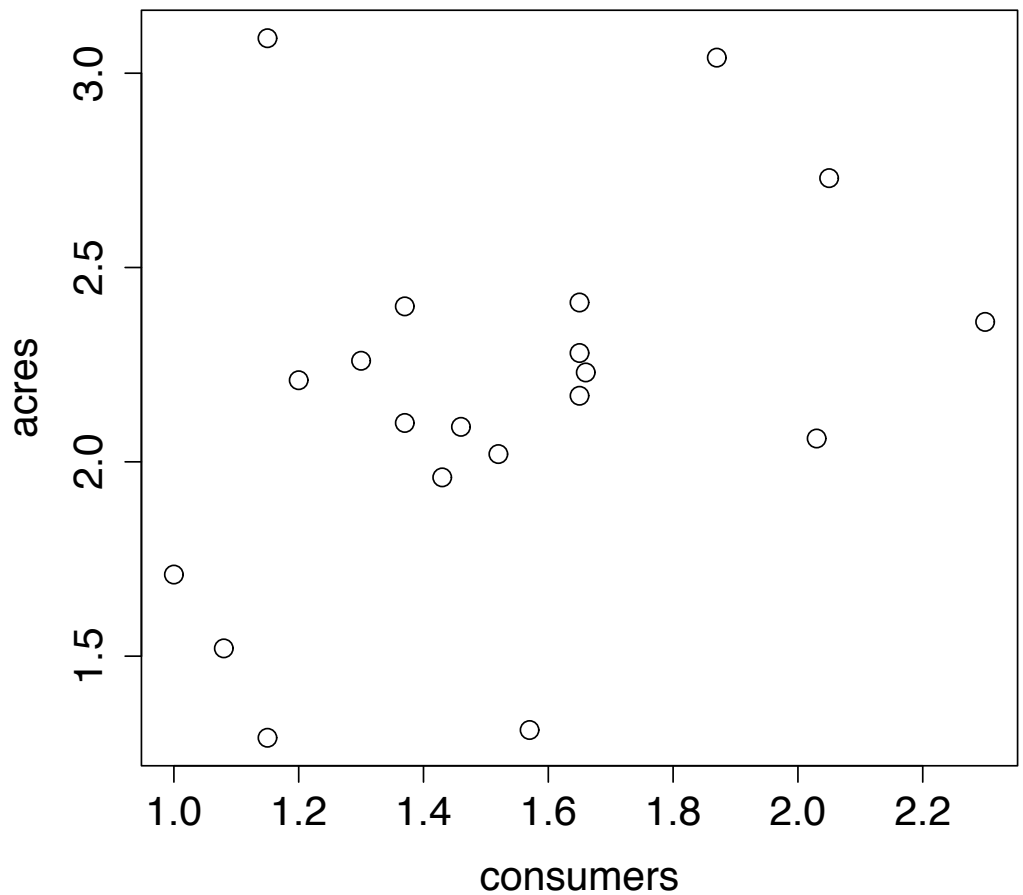
Regressione lineare con R

- Mediante la funzione di attribuzione `<-` abbiamo creato un oggetto, che abbiamo chiamato `fit`.
- Un oggetto è qualcosa di più complicato di un vettore o di una matrice. E' una lista di elementi su cui si può applicare una serie di funzioni.
- Ad esempio con il comando `coef( fit )` vediamo le stime dei minimi quadrati.

Illustrazione. Prendiamo ora in esame il campione di dati analizzato dall'antropologo Sahlins (1972) e riguardante la produzione agricola in una comunità primitiva della valle dello Gwemba nell'Africa centrale.

- La variabile X (*consumers*) è il rapporto tra consumatori e produttori in ciascun nucleo familiare; la variabile Y è il numero di acri coltivati da ciascuna famiglia (*acres*).
- Sahlins ipotizza che l'area coltivata da ciascuna famiglia aumenta all'aumentare del rapporto tra produttori e consumatori.

Regressione lineare con R



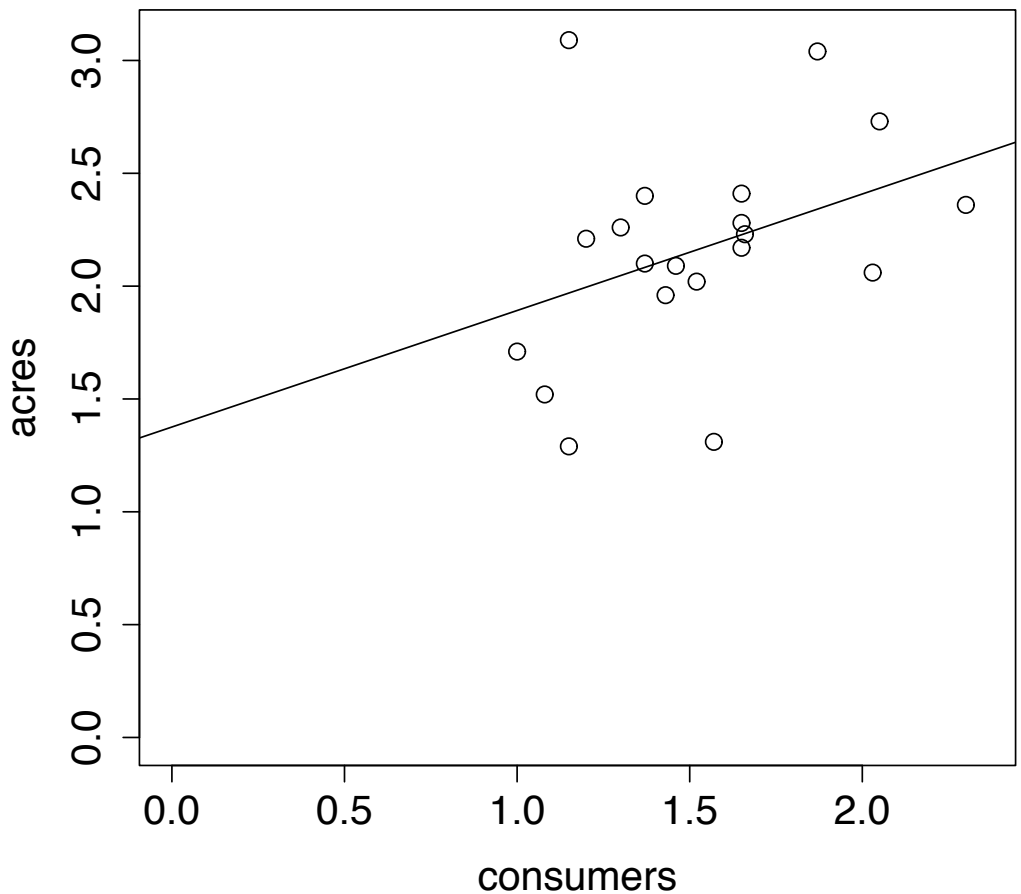
Regressione lineare con R

- L'ispezione visiva del diagramma di dispersione suggerisce la presenza di un'associazione positiva tra le due variabili, confermata da una correlazione $r = 0.3757$.
- La retta di regressione per i dati di Sahlins è riportata nella figura seguente.
- La funzione di regressione stimata è

$$\hat{Y} = 1.3756 + 0.5163X$$

- I comandi **R** utilizzati per trovare le stime dei minimi quadrati sono riportati di seguito.

Regressione lineare con R



Regressione lineare con R


```
> library(car)
> data(Sahlins)
> attach(Sahlins)
> fit <- lm(acres ~ consumers)
```

L'oggetto `fit` contiene più quantità

```
> names(fit)
 [1] "coefficients"  "residuals"     "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"
[10] "call"          "terms"         "model"
```

Regressione lineare con R

I coefficienti dei minimi quadrati si trovano mediante la funzione `coefficients()`:

```
> coef(fit)
(Intercept)   consumers
  1.3756445   0.5163201
```

Interpretazione.

- $\hat{\beta} = 0.52$: se il rapporto tra produttori e consumatori aumenta di un'unità, l'area coltivata da una famiglia aumenta, in media, di 0.52 acri.
- Solitamente interpretiamo l'intercetta $\hat{\alpha}$ come il valore atteso di Y condizionato a $X = 0$, ma in questo caso è impossibile un rapporto pari a 0 tra consumatori e produttori.
- L'intercetta $\hat{\alpha}$ è raramente informativa, dato che il valore $\hat{Y}$ in corrispondenza di $X = 0$ di solito non è importante.

Scomposizione di Y

Scomposizione di Y

Il modello di regressione lineare semplice assume che la variabile dipendente Y_i sia costituita dalla somma di due componenti:

1. una componente deterministica $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ predicibile dal modello lineare a partire dalla variabile X ;
2. una componente casuale E_i, detta residuo, non predicibile dal modello lineare.

$$Y_i = \hat{Y}_i + E_i$$
$$= \hat{\alpha} + \hat{\beta} X_i + E_i$$

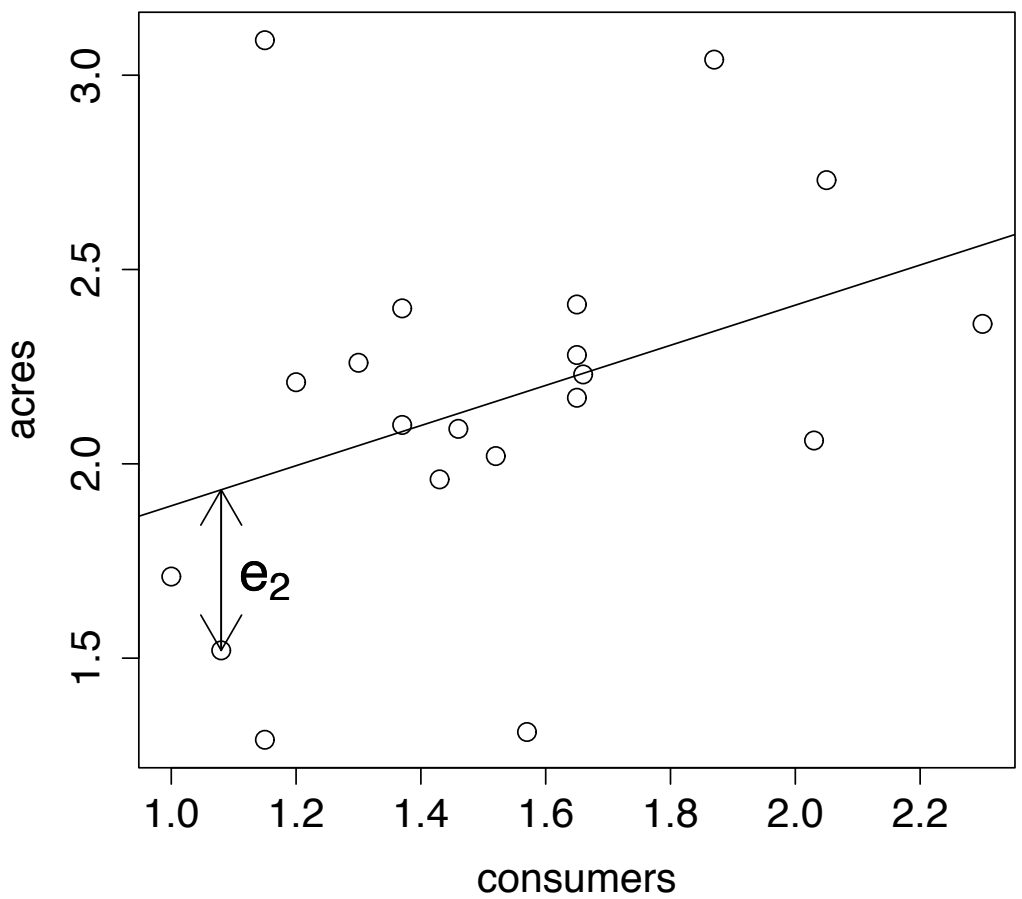
Illustrazione. Per la seconda delle osservazioni di Sahlins, per esempio, avremo

$$
\begin{aligned}
Y_2 &= \hat{Y}_2 + E_2 \\
&= \hat{\alpha} + \hat{\beta} X_2 + E_2 \\
1.52 &= 1.3756 + 0.5163 \times 1.08 + (-0.4133) \\
&= 1.9333 + (-0.4133)
\end{aligned}
$$

Il residuo E_2 rappresenta dunque la distanza verticale tra la seconda osservazione e la retta di regressione.

Illustrazione

Scomposizione di Y



Illustrazione

Esaminiamo i valori Y_i, $\hat{Y}_i$ e $E_i = Y_i - \hat{Y}_i$ per alcune osservazioni del campione raccolto da Sahlins.

Y_i	$\hat{Y}_i$	$E_i = Y_i - \hat{Y}_i$
1.7100	1.8920	-0.1820
1.5200	1.9333	-0.4133
1.2900	1.9694	-0.6794
3.0900	1.9694	1.1206
2.2100	1.9952	0.2148
2.2600	2.0469	0.2131
...	...	...
2.3600	2.5632	-0.2032

Calcoliamo con **R** i residui e i valori previsti dal modello.

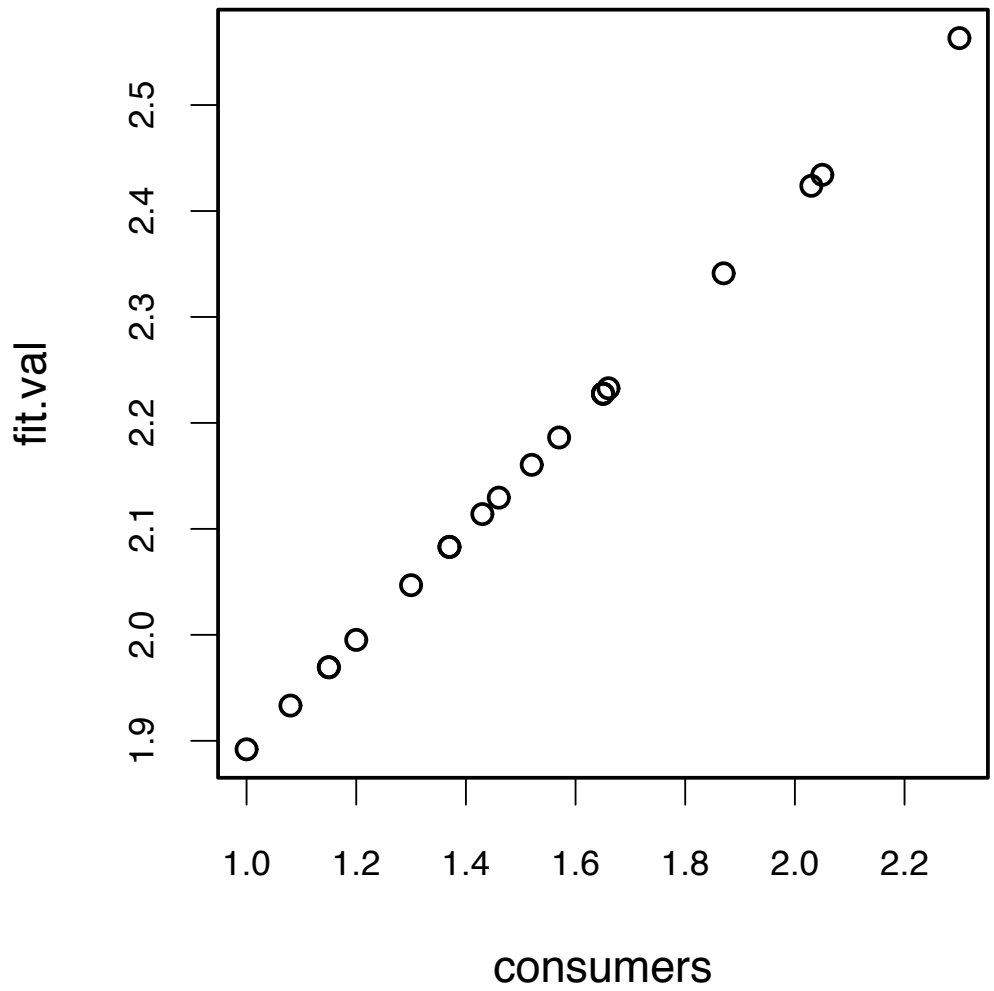
```
> res <- resid(fit)
> fit.val <- fitted(fit)
```

Rappresentiamo ora i valori predetti `fit.val` in funzione della variabile esplicativa `consumers`:

```
> plot(consumers, fit.val)
```

Illustrazione

Scomposizione di Y



Proprietà degli stimatori dei minimi quadrati

Proprietà degli stimatori dei minimi quadrati

Discuteremo ora due proprietà degli stimatori dei minimi quadrati:

1. La somma dei residui della retta di regressione calcolata con il metodo dei minimi quadrati è uguale a zero.
2. La soluzione dei minimi quadrati è fortemente influenzata dalla presenza di dati anomali

Media dei residui

- In precedenza abbiamo detto che il metodo dei minimi quadrati consente di trovare:

  i parametri $\hat{\alpha}$ e $\hat{\beta}$ che minimizzano $\sum \text{residui}_i^2$

- Perchè non trovare semplicemente i parametri $\hat{\alpha}$ e $\hat{\beta}$ che minimizzano la somma dei residui (non innalzati al quadrato)?
- Può essere dimostrato che la somma dei residui è uguale a zero per qualsiasi retta passante per il punto di coordinate $(\bar{X}, \bar{Y})$. Ci sono dunque infinite rette che rendono uguale a zero la somma dei residui.

Media dei residui

- Dato che lo stimatore dei minimi quadrati dell'intercetta è uguale a

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

  ne segue che la retta trovata con il metodo dei minimi quadrati passa per il punto $\bar{X}$, $\bar{Y}$.
- Di conseguenza, la somma dei residui calcolati con il metodo dei minimi quadrati sarà necessariamente uguale a zero.
  - Per i dati di Sahlins, infatti,

```
> sum(res)
[1] -8.283305e-17
```

Esaminiamo ora l'influenza dei dati anomali sulla soluzione dei minimi quadrati.

Illustrazione. Si considerino i dati della tabella 9.1 e si calcoli la retta di regressione per il tasso di omicidi (*murder rate*) in funzione del tasso di povertà (*poverty rate*).

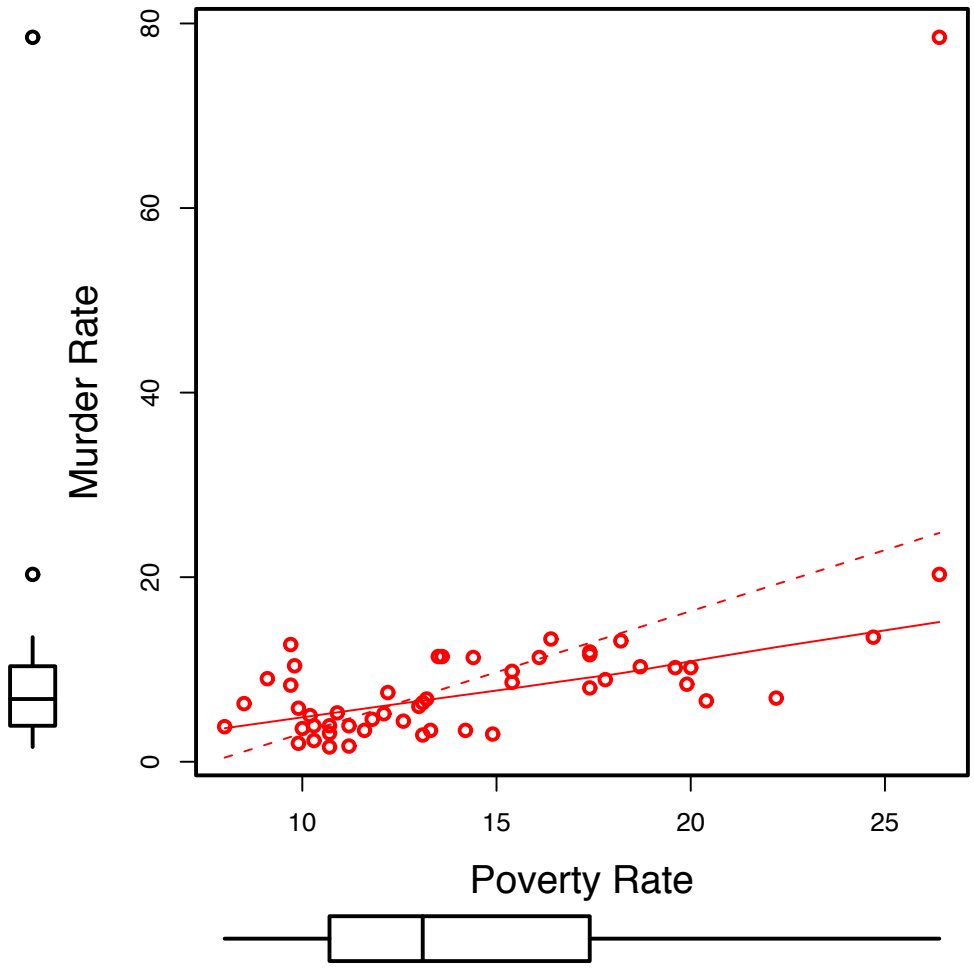
- Il diagramma di dispersione della figura 9.6 riprodotto nella figura seguente rivela che la retta dei minimi quadrati (retta tratteggiata) è fortemente influenzata da un'osservazione anomala (Washington D.C.).

Dati anomali

Si chiama *osservazione influente* un dato che, se venisse rimosso, produrrebbe un grande cambiamento nella pendenza della retta di regressione (come in questo caso).

```
> scatterplot(MR ~ P, data=murder, span=1,
+     xlab=c("Poverty Rate"), ylab=c("Murder Rate"))
```

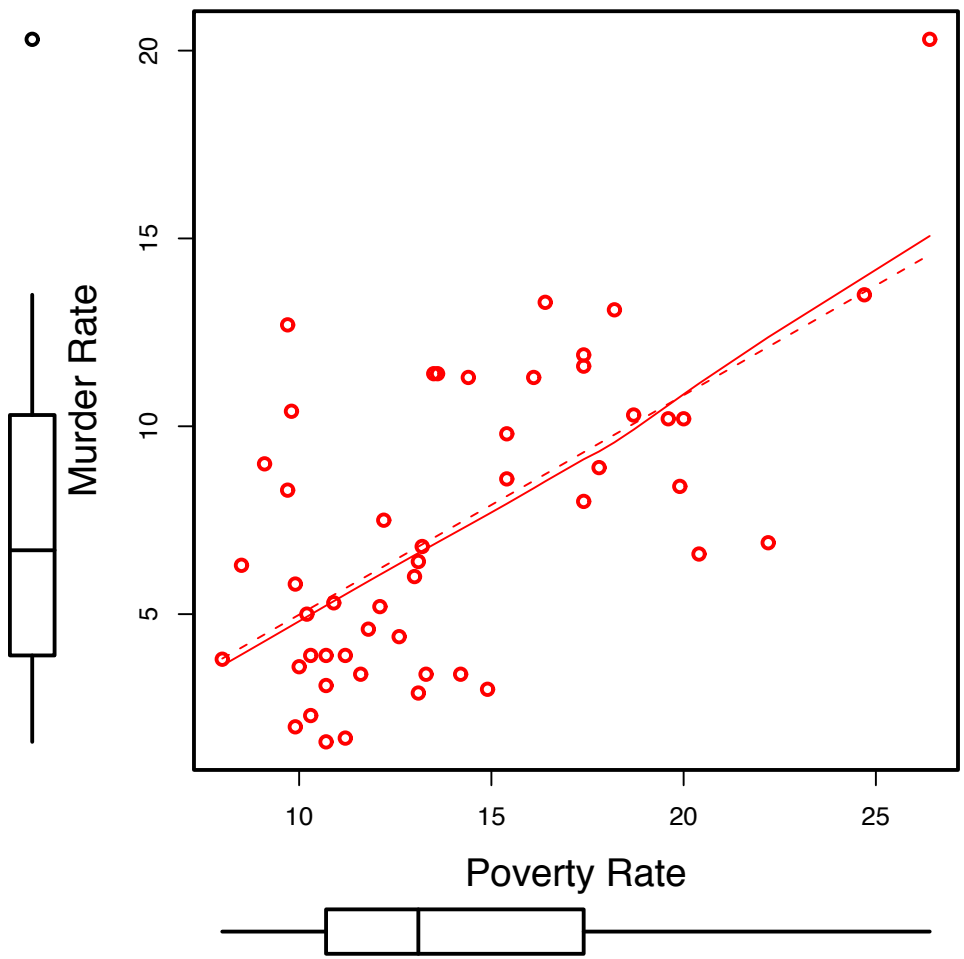
Dati anomali



Si noti che, se l'osservazione relativa a Washington D.C. viene rimossa, allora la pendenza della retta di regressione si dimezza.

```
> fm1 <- lm(MR ~ P, data=murder)
>
> murder2 <- murder[murder$State != "DC", ]
> fm2 <- lm(MR ~ P, data=murder2)
>
> fm1$coef
(Intercept)           P
 -10.136396    1.322960
>
> fm2$coef
(Intercept)           P
 -0.8567156   0.5842406
```

Dati anomali



La pendenza B e l'indice r

La pendenza B e l'indice r

La pendenza $\hat{\beta}$ della retta di regressione e il coefficiente di correlazione r sono legati dalla seguente equazione:

$$\hat{\beta}_{yx} = r \frac{s_y}{s_x}$$

Il coefficiente di correlazione r e la pendenza $\hat{\beta}$ sono simili per alcuni aspetti e diversi per altri.

- Se $r = 0$ (assenza di relazione lineare tra X e Y) allora anche $\hat{\beta} = 0$.
- Se le variabili X e Y sono standardizzate (in modo tale che $s_x = s_y = 1$), allora $\hat{\beta} = r$.

La pendenza B e l'indice r

- Il coefficiente di correlazione r non dipende dal fatto che ad una variabile sia stato assegnato il ruolo di variabile dipendente e all'altra il ruolo di variabile esplicativa.
- Questo non è vero per $\hat{\beta}$. Se X (piuttosto che Y) viene considerata la variabile dipendente allora

$$\hat{\beta}_{xy} = r \frac{s_x}{s_y}$$

che, in generale, è diverso da $\hat{\beta}_{yx}$.

$$r = \sqrt{r \frac{s_y}{s_x} r \frac{s_x}{s_y}} = \sqrt{\hat{\beta}_{yx} \hat{\beta}_{xy}} = \text{media geometrica di } \hat{\beta}_{yx} \text{ e } \hat{\beta}_{xy}$$

La pendenza B e l'indice r

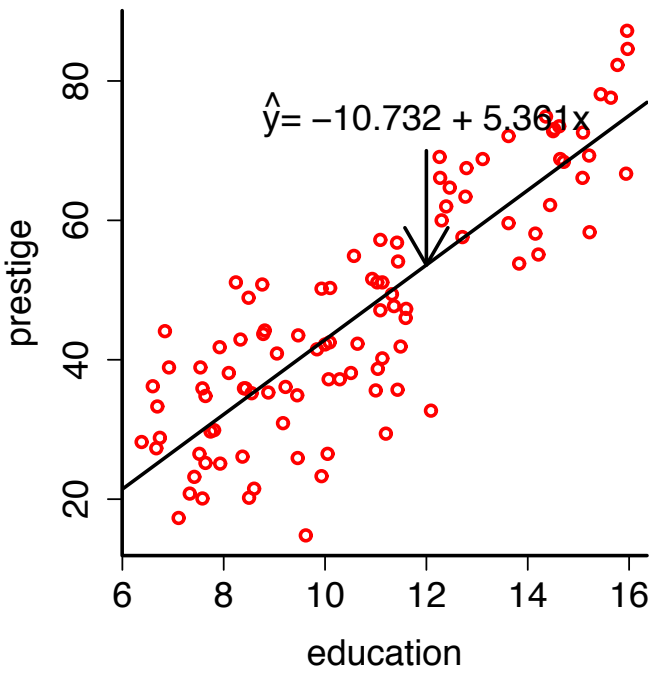
- Ci sono due rette di regressione
  - una per la regressione di Y su X e
  - l'altra per la regressione di X su Y .
- Queste due rette di regressione saranno diverse, a meno che r = 1.

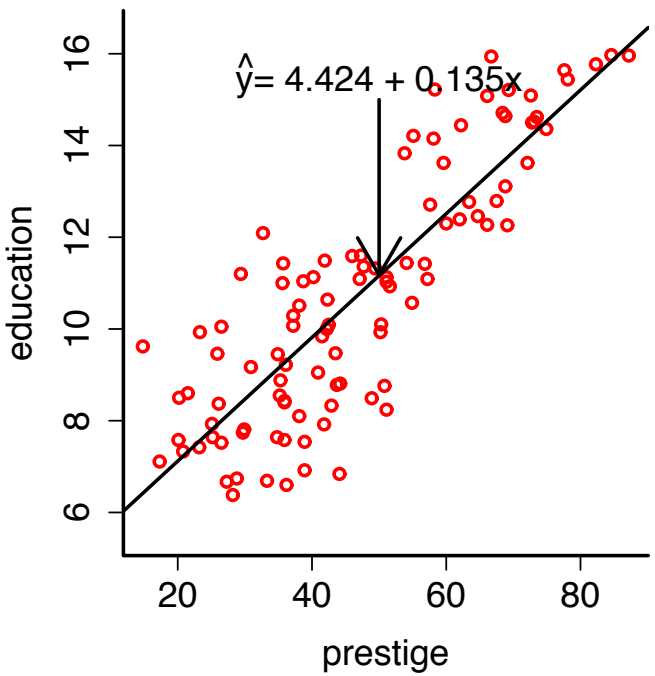
Precisamente, dato che $r = \sqrt{\hat{\beta}_{yx}\hat{\beta}_{xy}}$, allora

$$\frac{r^2}{\hat{\beta}_{yx}} = \hat{\beta}_{xy};$$

con $r = 1$, i due coefficienti angolari sono esattamente l'uno il reciproco dell'altro

Illustrazione





Regressione e correlazione

Una distinzione schematica potrebbe essere formulata nel modo seguente.

- L'*analisi della regressione* ricava dai dati campionari un modello statistico che consente di
  - prevedere i valori di una variabile, detta dipendente e (talvolta) individuata come effetto, a partire dai valori dell'altra variabile, detta indipendente o esplicativa, (talvolta) individuata come causa.
- L'*analisi della correlazione*
  - misura l'intensità dell'associazione tra due variabili quantitative, senza assumere che tra esse esista un nesso causale.

Associazione e causalità

E’ sempre importante distinguere tra associazione e causalità.

- L’associazione statistica tra due variabili indica che esse variano congiuntamente, non che tra esse esista una relazione diretta di causa-effetto. Associazione non significa causalità.
- L’associazione tra due variabili, misurata con test statistici quali la correlazione e la regressione, può essere prodotta da una terza variabile confondente a cui entrambe sono legate.

Indice di determinazione

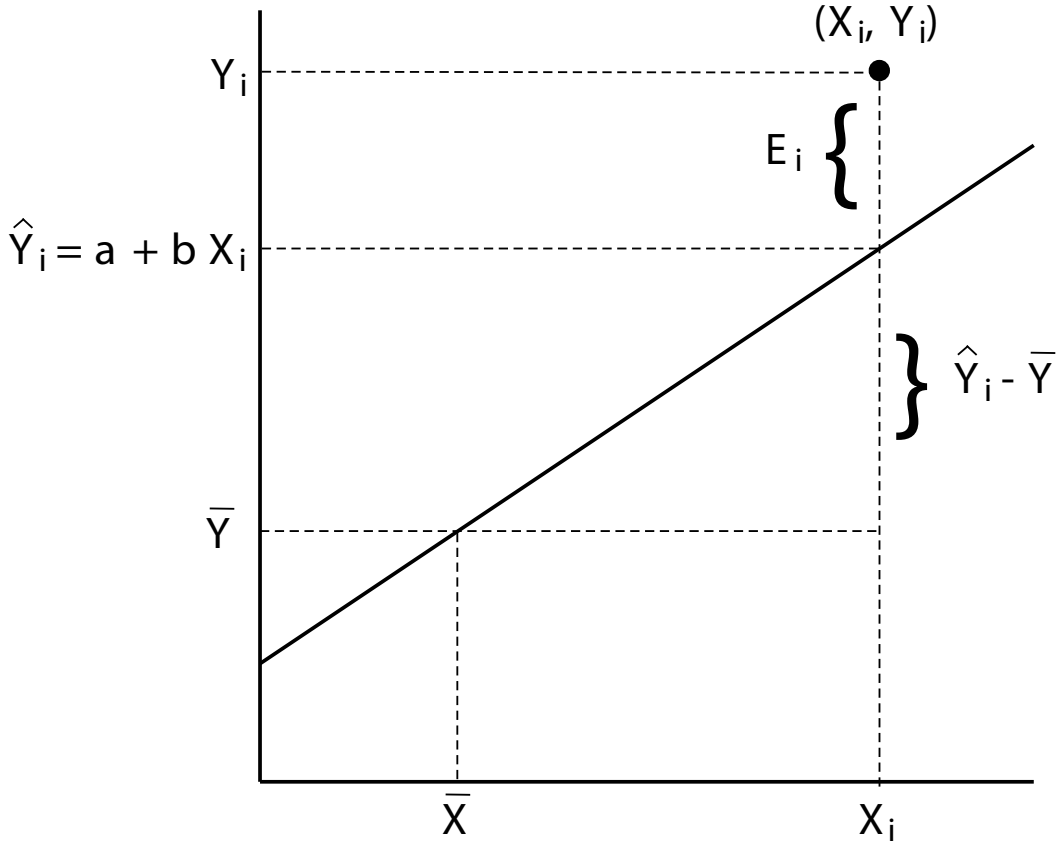
Indice di determinazione

- La variazione di Y_i rispetto alla media $\bar{Y}$ si può scomporre nella somma del residuo E_i e dello scarto di $\hat{Y}_i$ rispetto alla media $\bar{Y}$:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}) + (\hat{Y} - \bar{Y})$$

- La quantità $\hat{Y} - \bar{Y}$ è la frazione di variazione di Y rispetto alla media che viene spiegata dalla funzione di regressione.
- Il residuo costituisce invece la frazione di variazione di Y che non è possibile spiegare dalla funzione di regressione.

Indice di determinazione



Teorema di scomposizione della devianza

Il teorema della scomposizione della devianza dimostra che la devianza totale $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ si scompone nella somma di due componenti:

- la devianza spiegata $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- la devianza residua $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

ovvero

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Teorema della scomposizione della devianza

La differenza tra la devianza totale e la devianza residua rappresenta la quota della variazione di Y rispetto alla media $\bar{Y}$ che viene "spiegata" dalla regressione di Y su X :

$$\begin{array}{lcll} \text{devianza} & = & \text{devianza} & - \quad \text{devianza} \\ \text{spiegata} & & \text{totale} & \quad \text{residua} \end{array}$$

$$= \sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y}_i)^2$$

$$= \sum (\hat{Y}_i - \bar{Y})^2$$

Indice di determinazione

Il quadrato del coefficiente di correlazione, detto indice di determinazione r^2, misura la frazione di devianza totale spiegata dal modello di regressione:

$$r^2 = \frac{\text{devianza spiegata}}{\text{devianza totale}}$$

$$= \frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

$$= 1 - \frac{\sum(Y_i - \hat{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

Indice di determinazione

- Quando c'è una perfetta relazione lineare tra X e Y i residui sono tutti uguali a zero; la somma dei residui al quadrato sarà 0 e $r^2 = 1$.
- Quando non c'è relazione lineare tra X e Y la variazione spiegata è uguale a zero e $r^2 = 0$.
  - Per la regressione del prestigio sull'istruzione $r = .8502$, e quindi la regressione spiega circa il 72% della variazione dei punteggi della variabile "prestigio" ($r^2 = 0.8502^2 = 0.7228$).

Riduzione proporzionale dell'errore

- Parliamo di previsione, in termini statistici, ogni qual volta ci poniamo il problema di approssimare il comportamento di una variabile che per qualche ragione non si può osservare.
- Una approssimazione puntuale del comportamento di Y è data dal predittore $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$.

Riduzione proporzionale dell'errore

L'indice r^2 viene presentato da Agresti e Finlay (1999) come una misura di riduzione proporzionale dell'errore di previsione (PRE):

$$PRE = \frac{E_1 - E_2}{E_1}$$

dove

- E_1 è l'errore che si commette volendo prevedere Y senza utilizzare la conoscenza di X ;
- E_2 è l'errore che si commette volendo prevedere Y come funzione lineare di X .

Riduzione proporzionale dell'errore

Senza usare la conoscenza di X , la stima migliore di Y è data da $\bar{Y}$.

- Quale errore si commette in tali circostanze? Per una singola osservazione, l'errore di previsione è

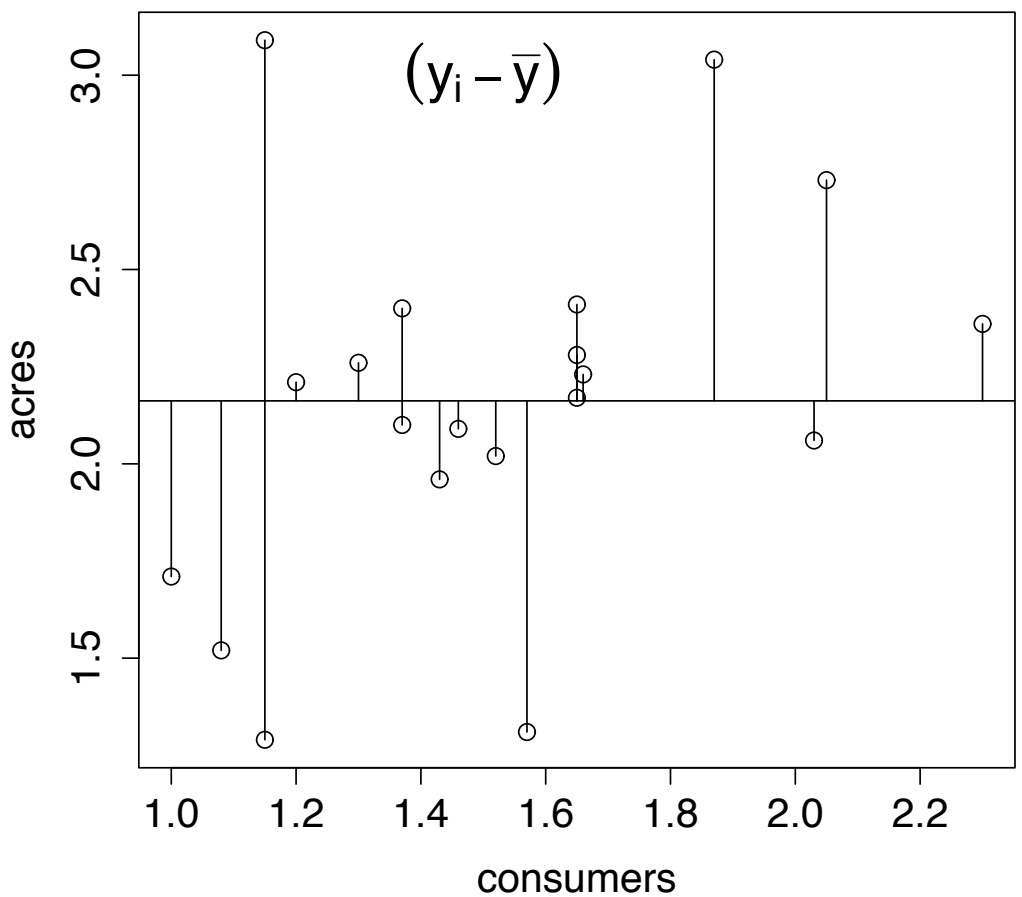
$$Y_i - \bar{Y}$$

- Per tutto il campione, sommiamo gli n errori elevati al quadrato:

$$E_1 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Tale somma è uguale alla devianza totale.

Riduzione proporzionale dell'errore



Riduzione proporzionale dell'errore

La previsione migliore di Y *che fa uso della conoscenza di* X è $\hat{Y}$.

- Quale errore si commette volendo predire Y con la retta di regressione? Per una singola osservazione, l'errore di previsione è

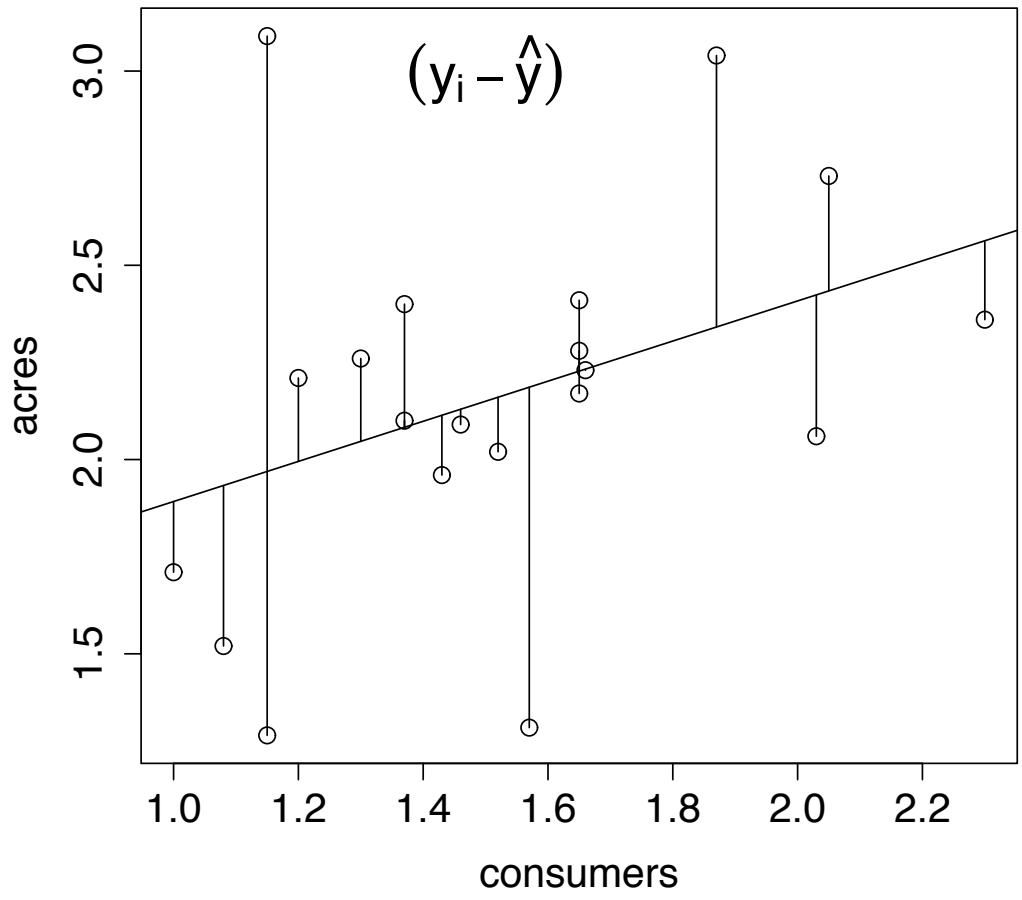
$$Y_i - \hat{Y}$$

- Per tutto il campione, sommiamo gli n errori elevati al quadrato:

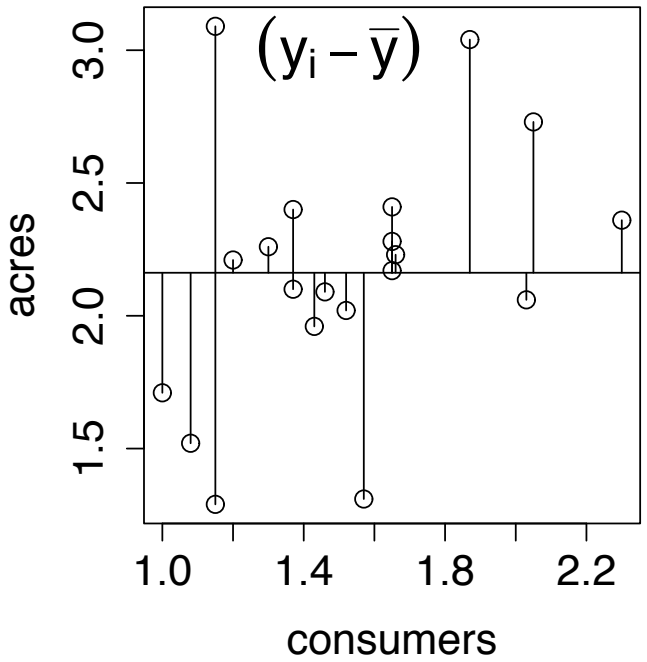
$$E_2 = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

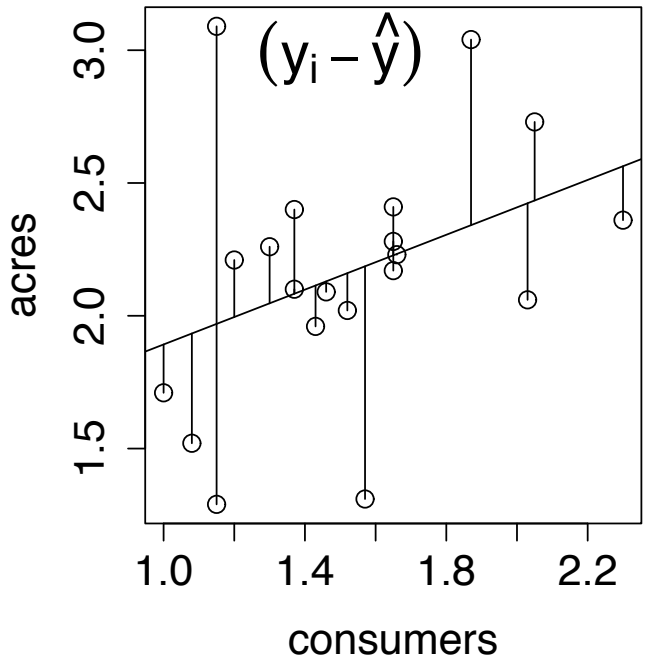
- Tale somma è detta devianza residua.

Riduzione proporzionale dell'errore



Riduzione proporzionale dell'errore

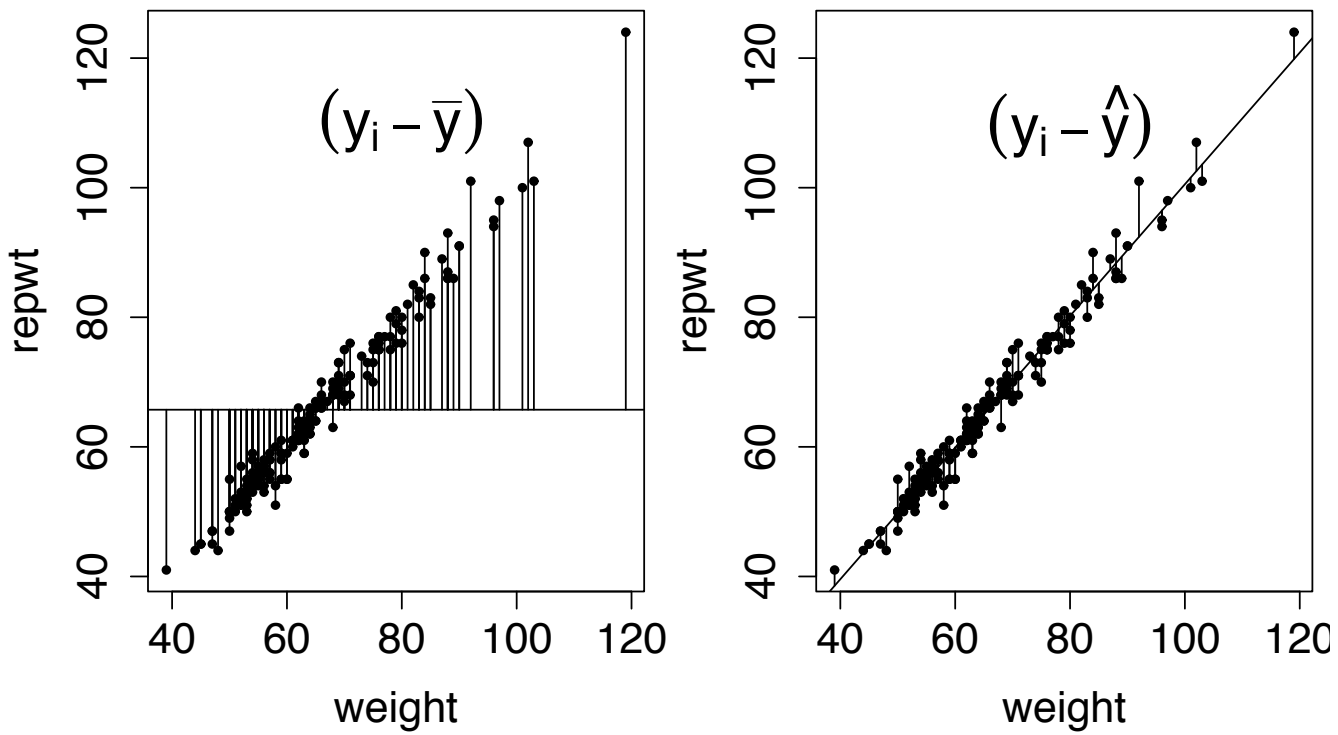




Riduzione proporzionale dell'errore

- Per i dati di Sahlins la riduzione proporzionale dell'errore non è molto evidente.
- Nella figura seguente è riportato un esempio più chiaro, relativo ad un campione di dati che verrà utilizzato in seguito.
  - La variabile dipendente è il peso effettivo (*weight*) e la variabile indipendente è il peso riferito (*repwt*) in un campione di 112 donne raccolto da Caroline Davis.

Riduzione proporzionale dell'errore



Riduzione proporzionale dell'errore

La differenza tra la devianza totale e la devianza residua indica la diminuzione dell'errore che si commette prevedendo Y con $\hat{Y}_i$ anziché con $\bar{Y}$:

$$PRE = \frac{E_1 - E_2}{E_1}$$

$$= \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

$$= r^2$$

Inferenza sul modello di regressione

Inferenza sul modello di regressione

Finora abbiamo considerato la regressione lineare quale strumento di descrizione di un campione di dati.

- Il problema che ci eravamo posti era quello di trovare la retta che interpola al meglio i dati.
- E' però importante interrogarsi sulla relazione che intercorre tra $\hat{\alpha}$ e $\hat{\beta}$ calcolati sulla base delle informazioni fornite da un particolare campione e i parametri α e β della retta di regressione nella popolazione.
- Questo è il problema dell'inferenza statistica sul modello di regressione lineare.

Inferenza sul modello di regressione

- In precedenza, il problema dell'inferenza statistica è stato affrontato considerando la relazione che intercorre tra una statistica del campione e il corrispondente parametro della popolazione.
- Nel caso di un campione casuale di dimensioni n estratto da una popolazione con media μ e varianza σ^2, per esempio, la media del campione veniva considerata una variabile aleatoria avente valore atteso μ, varianza σ^2/n e distribuzione tendente alla normale al crescere delle dimensioni del campione.

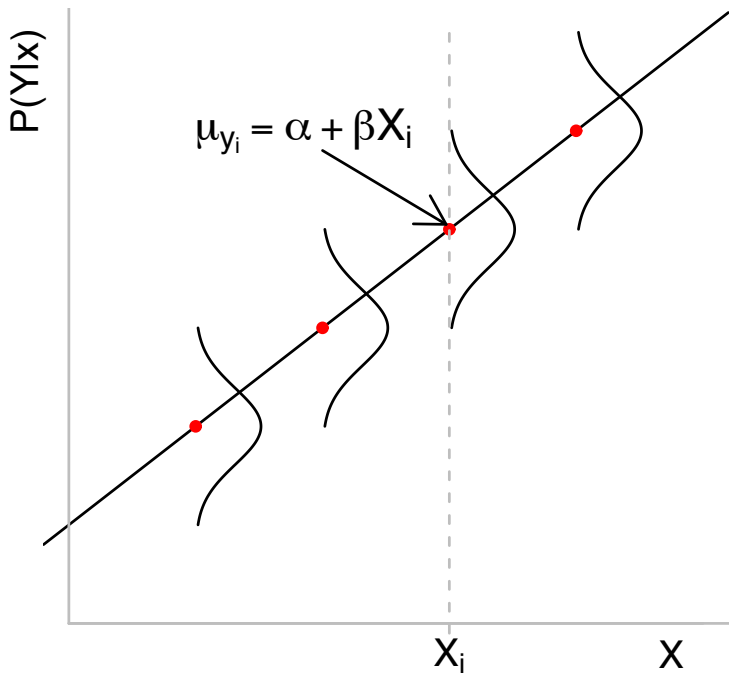
Inferenza sul modello di regressione

Per il modello di regressione lineare, il valore atteso della variabile aleatoria Y non è costante, ma bensì è funzione di una variabile (non aleatoria) X .

- Si noti la precisazione: la variabile esplicativa X non è considerata una variabile aleatoria, ma bensì una costante conosciuta.
- Ciò significa che la variabile X continuerebbe ad avere gli stessi valori anche se un altro campione venisse osservato.
- In altre parole, i valori della variabile X sono sotto il controllo dello sperimentatore.

Inferenza sul modello di regressione

- Il punto importante è che, a ciascun valore X , corrisponde una popolazione di valori Y .
- Uno solo di questi valori viene osservato in un particolare campione.



Inferenza sul modello di regressione

- Si genera un campione casuale estraendo a caso un'osservazione da ciascuna popolazione dei valori di Y dato X .
- Sulla base dei dati di un particolare campione si stimano i parametri della retta di regressione $\hat{\alpha}$ e $\hat{\beta}$.
- Come si può eseguire un test statistico sui parametri del modello di regressione?

Test sui parametri del modello di regressione

- Se nella popolazione le variabili X e Y sono indipendenti, allora il parametro β sarà uguale a zero.
- Le ipotesi da sottoporre a test sono

$$H_0 : \beta = 0 \qquad \text{verso} \qquad H_a : \beta \neq 0$$

Test sui parametri del modello di regressione

La statistica test è

$$t = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

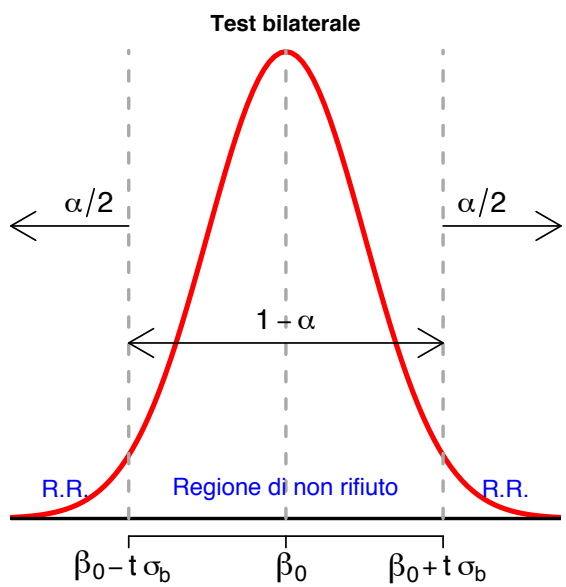
Il rapporto $\hat{\beta}/\hat{\sigma}_{\hat{\beta}}$ è chiamato rapporto-t e si distribuisce come una t di Student con $n - 2$ gradi di libertà.

Regione di accettazione e regione critica

La regione di accettazione e la regione critica risultano

$$\text{R.A.}: |t_0| \leq t_{n-2,1-\alpha/2}$$

$$\text{R.C.}: |t_0| > t_{n-2,1-\alpha/2}$$



Ipotesi

Si può utilizzare il rapporto-t può per i test sui parametri del modello se le seguenti ipotesi risultano soddisfatte.

- La media di ciascuna distribuzione Y è associata a X in base ad un'equazione lineare (linearità):

$$\mathbb{E}(y) = \alpha + \beta(x)$$

- Tutte le distribuzioni Y hanno la stessa varianza (omoschedasticità).
- Ciascuna distribuzione Y è normale (normalità).
- Le osservazioni costituiscono un campione casuale indipendente (indipendenza).

Deviazione standard di $\hat{\beta}$

Resta da stabilire come si calcola la deviazione standard di $\hat{\beta}$.

Una stima della deviazione standard di $\hat{\beta}$ è data da

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}_{\varepsilon}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

dove una stima della deviazione standard dei residui nella popolazione, $\hat{\sigma}_{\varepsilon}$, (chiamata *root mean square error*) si ottiene con

$$\hat{\sigma}_{\varepsilon} = \sqrt{\frac{\sum E_i^2}{n-2}}$$

Illustrazione

```
> fm <- lm(acres ~ consumers)
> summary(fm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3756     0.4684   2.937  0.00881 **
consumers     0.5163     0.3002   1.720  0.10263
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4543 on 18 degrees of freedom
Multiple R-Squared: 0.1411,     Adjusted R-squared: 0.0934
F-statistic: 2.957 on 1 and 18 DF,  p-value: 0.1026
```

Stima della deviazione standard dei residui:

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{\sum E_i^2}{n-2}}$$

```
> e <- residuals(fm)
> se <- sqrt( sum(e^2) / 18 )
> se
[1] 0.4543179
```

Stima della deviazione standard di $\hat{\beta}$:

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}_{\varepsilon}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

```
> sb <- se / sqrt( sum( (consumers - mean(consumers))^2 ) )
> sb
[1] 0.3002335
```

Illustrazione

- La statistica $t = \hat{\beta}/\hat{\sigma}_{\hat{\beta}}$ diventa quindi:

```
> t <- 0.5163201 / sb
> t
[1] 1.719728
```

- Dato che Sahlins ha ipotizzato che la variabile "acres" aumenta all'aumentare di "consumers", l'ipotesi alternativa è $H_a : \beta > 0$.
- Il p-valore della statistica test corrisponde dunque alla sola coda destra della distribuzione t di Student con 18 gradi di libertà.

```
> pt(t, 18, lower.tail = FALSE)
[1] 0.05131462
```

Illustrazione

Si noti che questo p-valore è la metà di quello riportato da **R**, dato che **R** riporta di default il p-valore di un test bilaterale.

Intervallo di confidenza

L'intervallo di confidenza per β è

$$\beta \pm t_{n-2,1-\alpha/2}\hat{\sigma}_{\hat{\beta}}$$

dove $t_{n-2,1-\alpha/2}$ è il percentile della distribuzione t di Student con $n-2$ gradi di libertà tale che $|t_{n-2,1-\alpha/2}| = 1-\alpha/2$.

Intervallo di confidenza

Illustrazione. Per i dati di Sahlins, si calcoli l'intervallo di confidenza al 95% per il coefficiente angolare $\hat{\beta}$.

```
> tc <- qt(.975, 18)
> tc
[1] 2.100922
>
> 0.5163201 + tc*sb
[1] 1.147087
> 0.5163201 - tc*sb
[1] -0.1144471
```

$$P(-0.11 \leq \beta \leq 0.52) = 0.95$$

Condizioni di validità della regressione

Condizioni di validità della regressione

La retta di regressione fornisce un sommario adeguato della relazione tra X e Y se

- le due variabili sono effettivamente associate in maniera lineare,
- non ci sono dati anomali (*outliers*), e
- non è stato commesso un errore di specificazione.

Condizioni di validità della regressione

Come la media, la deviazione standard e la correlazione, anche la pendenza della retta dei minimi quadrati può essere fortemente influenzata dalla presenza di uno o di pochi valori anomali.

Un dato anomalo è un punto che, in un diagramma di dispersione, si distanzia in maniera marcata dalla nuvola che contiene la maggioranza dei dati.

Condizioni di validità della regressione

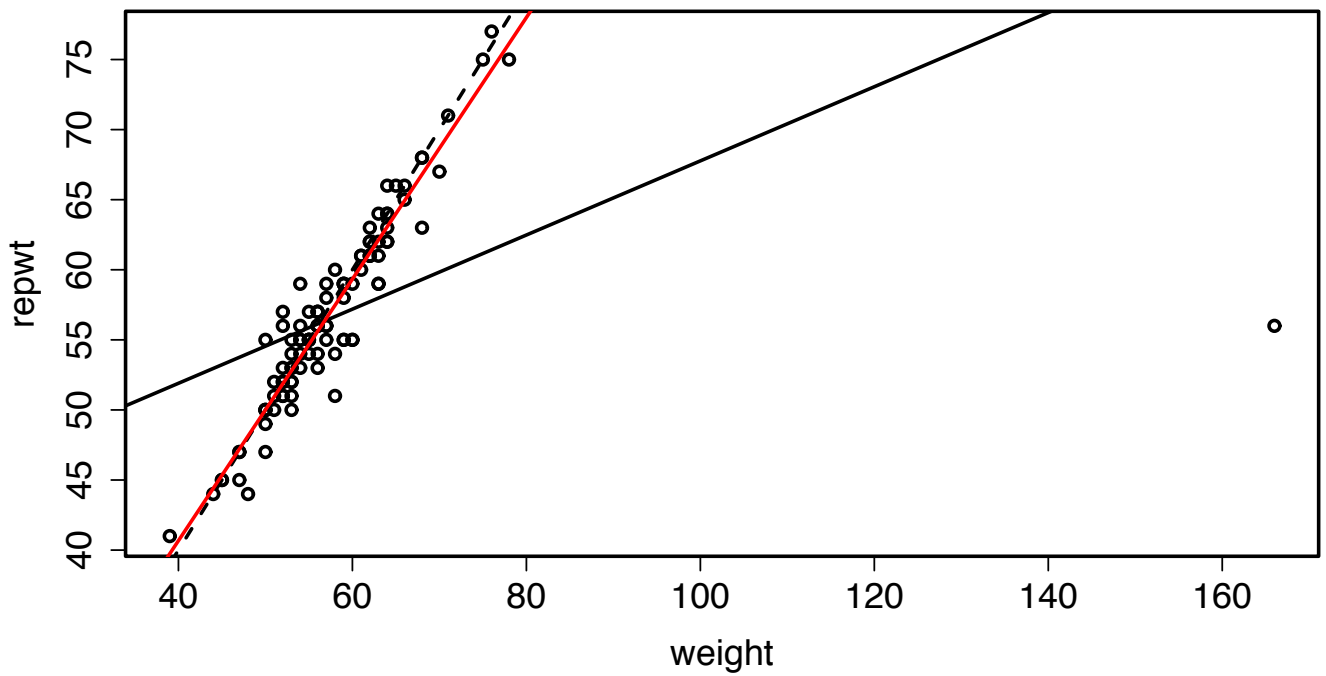
Illustrazione. I dati della figura seguente, riportati da Caroline Davis, una psicologa che si occupa di disordini alimentari, rappresentano il peso effettivo (*weight*) e il peso riferito (*repwt*) da 112 donne.

Se le donne riportano il loro peso senza distorsioni, allora la retta di regressione dovrebbe essere $\hat{Y} = X$ (ovvero, intercetta $\alpha = 0$ e pendenza $\beta = 1$ – linea tratteggiata).

Illustrazione

- Un solo valore anomalo fa sì che il metodo dei minimi produca una stima errata dell'orientamento della retta di regressione (linea nera).
- Se il dato anomalo viene omesso, i valori stimati della retta di regressione sono simili ai valori attesi $\hat{Y} = X$ (linea rossa).
- In questo caso, il dato anomalo è prodotto da un'errore nella tabulazione dei dati.

Illustrazione



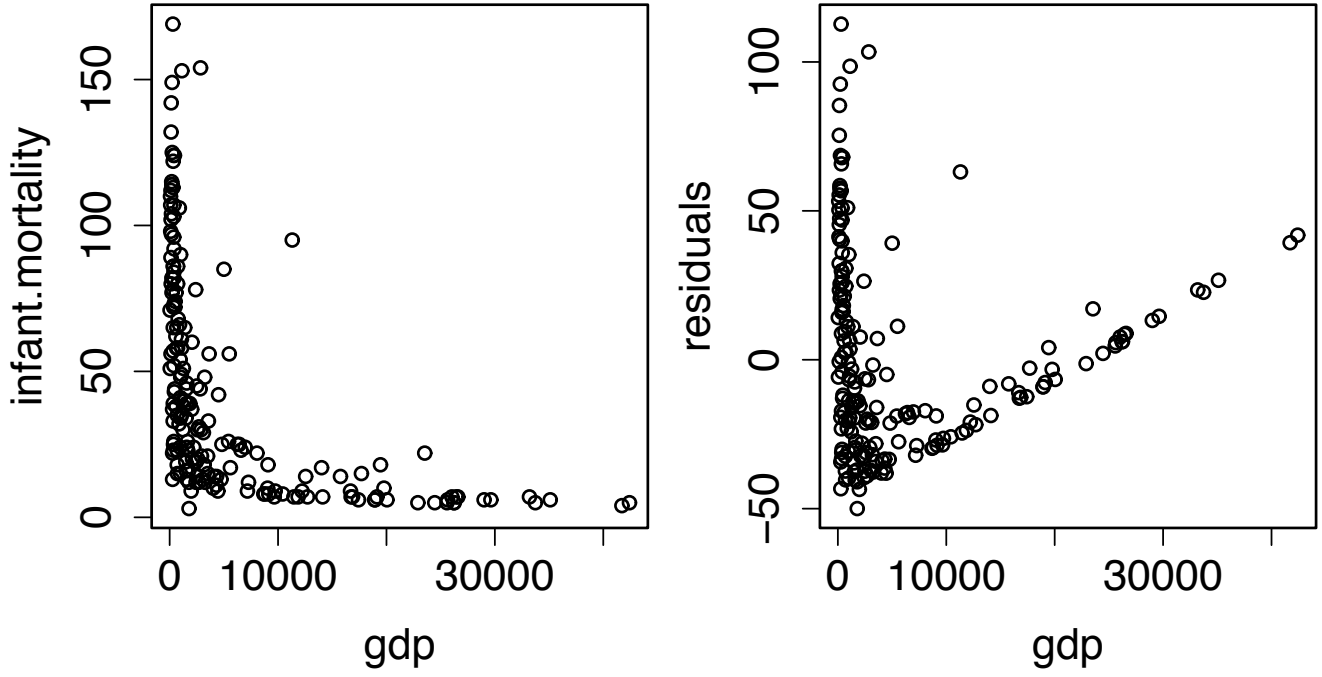
Illustrazione

Spesso i problemi del modello di regressione possono essere rivelati più chiaramente da un'analisi esplorativa realizzata collocando i residui (E_i) in un diagramma cartesiano in cui

- l'ordinata riporta gli scarti rispetto alla retta,
- l'ascissa indica il valore corrispondente della variabile indipendente X .

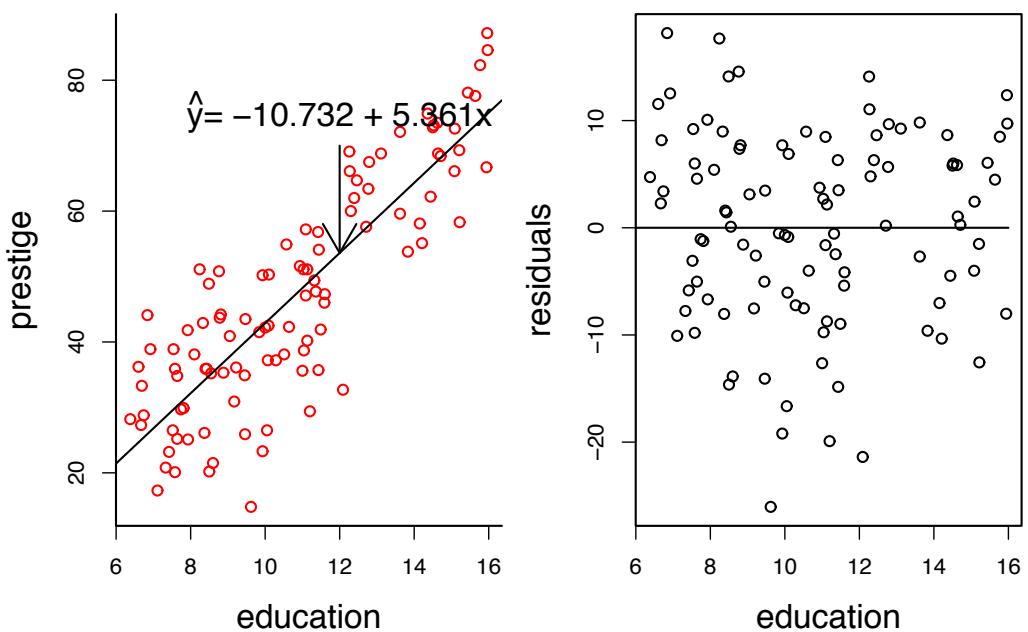
Illustrazione

La figura di sinistra rivela una relazione non lineare tra Y e X e la figura di destra riporta i residui della regressione in funzione di X .



Illustrazione

- Se non vi sono problemi, i residui del modello di regressione lineare occupano un'area omogenea centrata sullo 0 in tutta la gamma di X .
- Per la regressione della variabile "prestigio" sulla variabile "istruzione" tale condizione è soddisfatta.



Illustrazione

- I seguenti quattro campioni di dati proposti da Anscombe sono stati generati in maniera tale da avere esattamente la stessa retta di regressione e lo stesso coefficiente di correlazione:

$$\hat{Y} = 3.0 + 0.5X$$

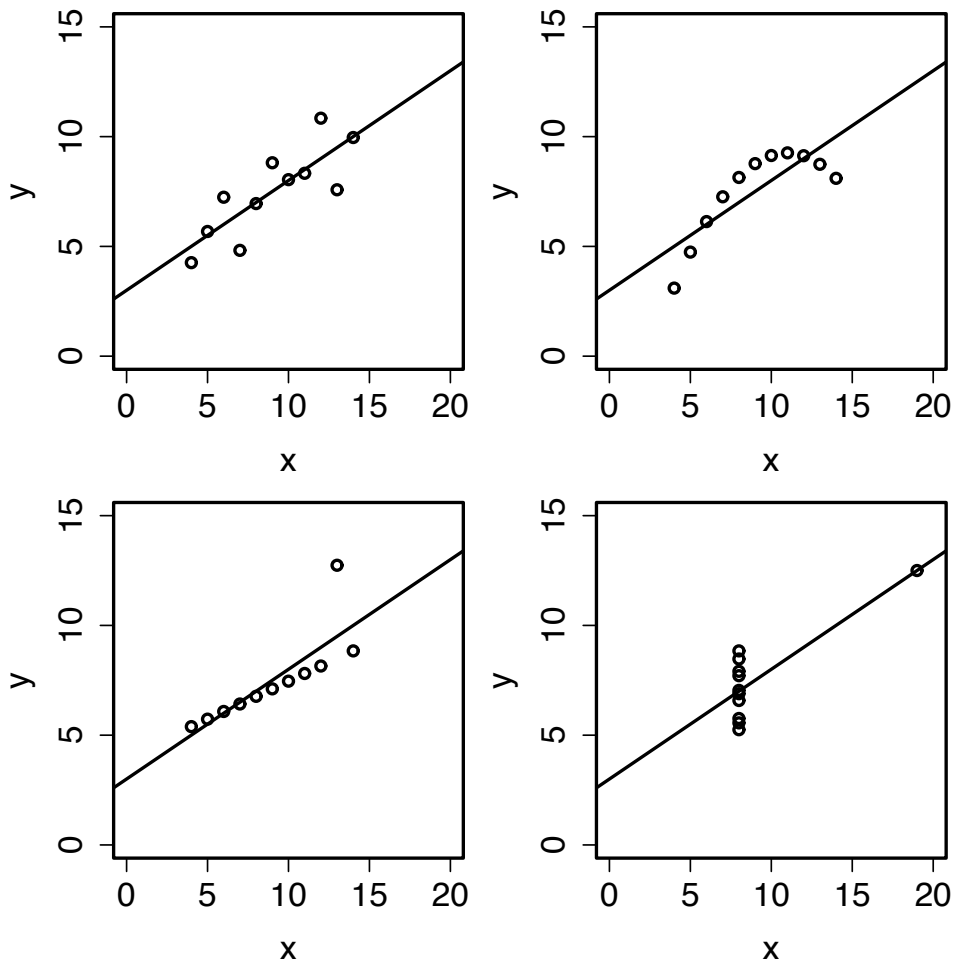
$$r = 0.82$$

- Inoltre, $\bar{X}$, $\bar{Y}$, s_x e s_y sono gli stessi in ciascun campione di dati.
- Ciò nonostante, la retta di regressione stimata con il metodo dei minimi quadrati rappresenta in maniera adeguata la relazione tra X e Y solo nel caso del primo campione di dati.

Illustrazione

	Campione 1		Campione 2		Campione 3		Campione 4	
	X	Y	X	Y	X	Y	X	Y
1	10.00	8.00	13.00	9.00	11.00	14.00	6.00	4.00
2	12.00	7.00	5.00	8.04	6.95	7.58	8.81	8.33
3	9.96	7.24	4.26	10.84	4.82	5.68	10.00	8.00
4	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00
5	5.00	9.14	8.14	8.74	8.77	9.26	8.10	6.13
6	3.10	9.13	7.26	4.74	10.00	8.00	13.00	9.00
7	11.00	14.00	6.00	4.00	12.00	7.00	5.00	7.46
8	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15
9	6.42	5.73	8.00	8.00	8.00	8.00	8.00	8.00
10	8.00	19.00	8.00	8.00	8.00	6.58	5.76	7.71
11	8.84	8.47	7.04	5.25	12.50	5.56	7.91	6.89

Illustrazione



Illustrazione

- Nel secondo campione di dati la relazione non è lineare.
- Il terzo campione contiene un dato anomalo.
- Nel quarto campione non c'è relazione lineare tra le variabili e l'orientamento della retta di regressione dipende da un'unico un valore anomalo.

- Nessuno di questi problemi, però, risulta evidente se consideriamo soltanto l'equazione dei minimi quadrati e il coefficiente di correlazione.
- E' dunque importante procedere ad un'ispezione visiva dei dati che vengono sottoposti all'analisi della regressione.

Errore di specificazione

Errore di specificazione

Una variabile confondente è una variabile indipendente che è stata omessa dall'analisi e che ha un importante effetto sulla relazione tra X e Y.

Un modello che esclude una o più variabili importanti si dice affetto da un errore di specificazione se tale omissione altera in maniera sostanziale le stime del modello.

Errore di specificazione

Illustrazione. Consideriamo uno studio di psicologia sociale condotto Moore e Krupat (1971). Ai partecipanti viene chiesto di fornire un giudizio percettivo a proposito di uno stimolo intrinsecamente ambiguo.

Prima di fornire la risposta, i soggetti vengono esposti al giudizio di un altro individuo ("partner") che, in realtà, è un collaboratore dello sperimentatore.

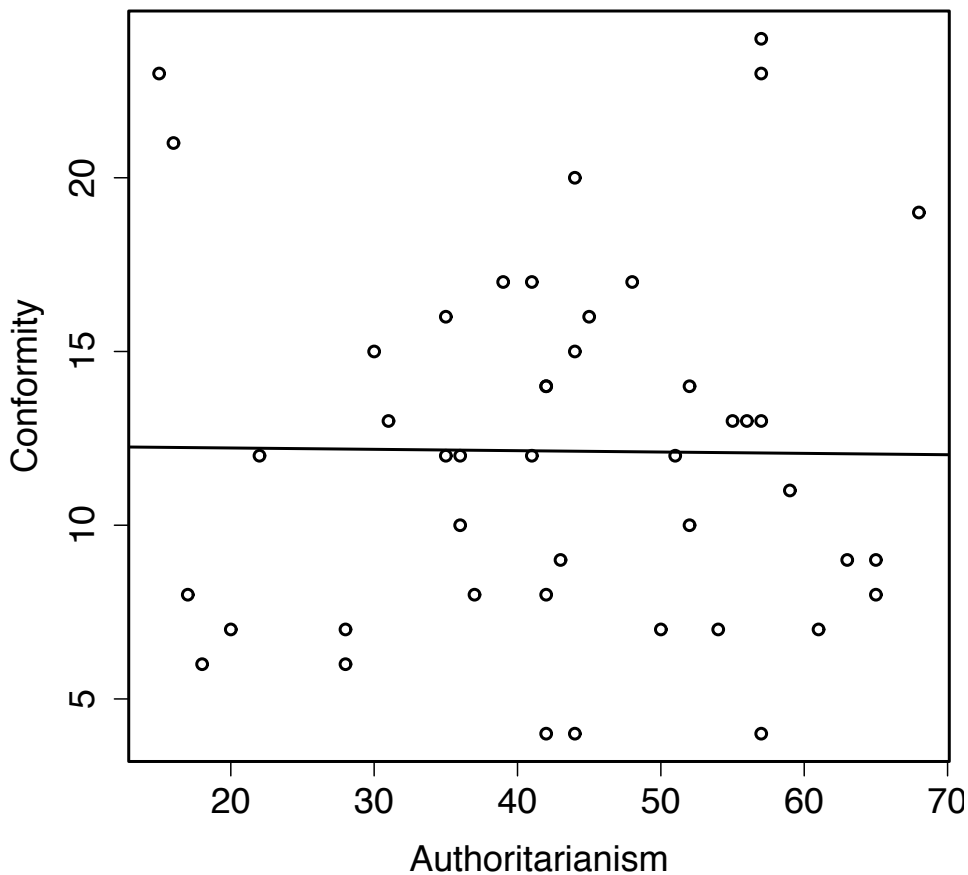
Si vuole stabilire se esiste una relazione tra la conformità sociale (misurata dal numero di volte in cui il soggetto si adegua al giudizio del partner) e l'"autoritarismo".

Illustrazione

- La figura seguente riporta i punteggi delle due variabili.
- Si noti che il diagramma di dispersione non rivela alcuna relazione tra *conformità* e *autoritarismo.*

Illustrazione

Errore di specificazione

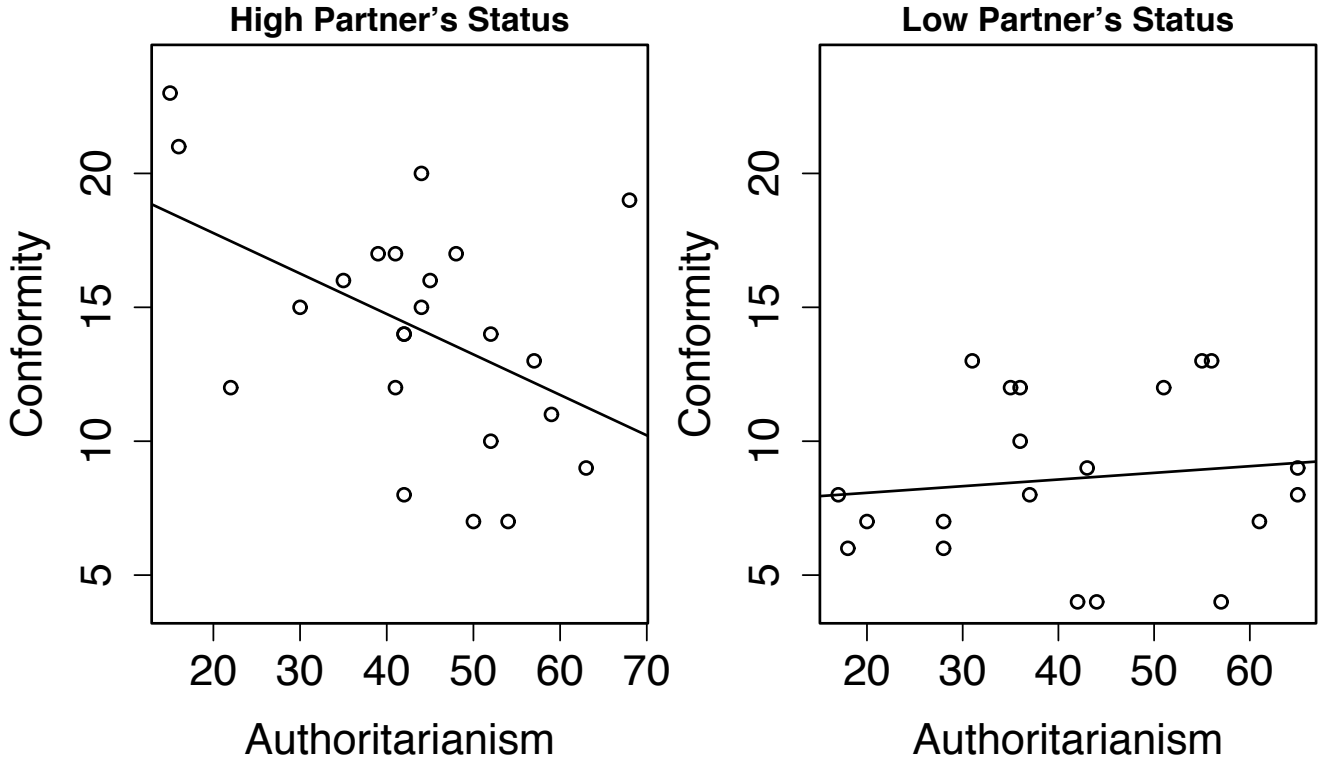


Illustrazione

- Stimiamo ora l'orientamento della retta di regressione distinguendo tra alto (medico) e basso (impiegato) status sociale del partner.
- Dato che la scala di autoritarismo contiene una componente di "rigidità concettuale", Moore e Krupat (1971) ipotizzano che gli individui con *punteggi bassi* di autoritarismo reagiscono in maniera diversa allo status sociale del partner degli individui con punteggi alti su tale scala.
- Questo infatti è quello che succede, come indicato dalla figura seguente.

Illustrazione

Errore di specificazione



Illustrazione

- Nel caso presente, dunque, lo status sociale del partner è una variabile confondente.
- La mancata considerazione di questa variabile produce infatti una stima distorta della relazione tra Y (*conformità*) e X (*autoritarismo).*

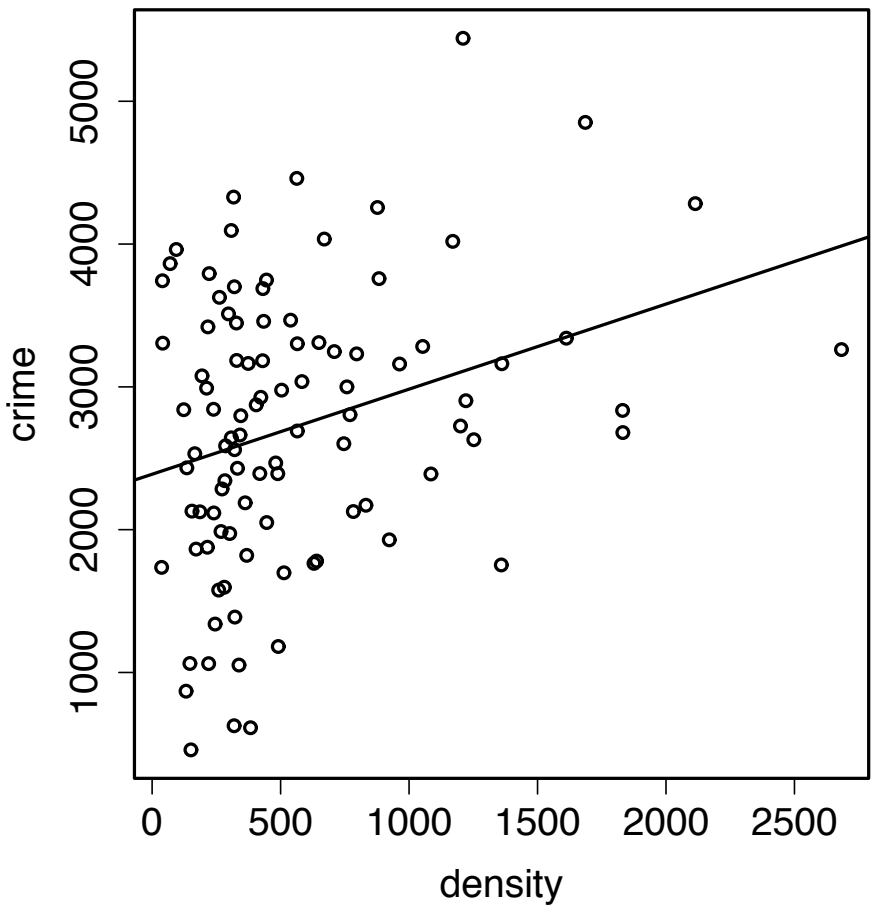
Può anche darsi l'effetto opposto – una relazione apparente tra due variabili X e Y può essere indotta dall'omissione di un'importante variabile Z.

Illustrazione. Consideriamo lo studio di Freedman (1975) sulla relazione tra densità della popolazione urbana e tasso di criminalità. Il data frame *Freedman* riporta i dati relativi a 110 aree metropolitane degli Stati Uniti.

La figura seguente rivela una associazione tra il tasso di criminalità (*"crime"*) e la densità della popolazione urbana (*"density"*).

Illustrazione

Errore di specificazione

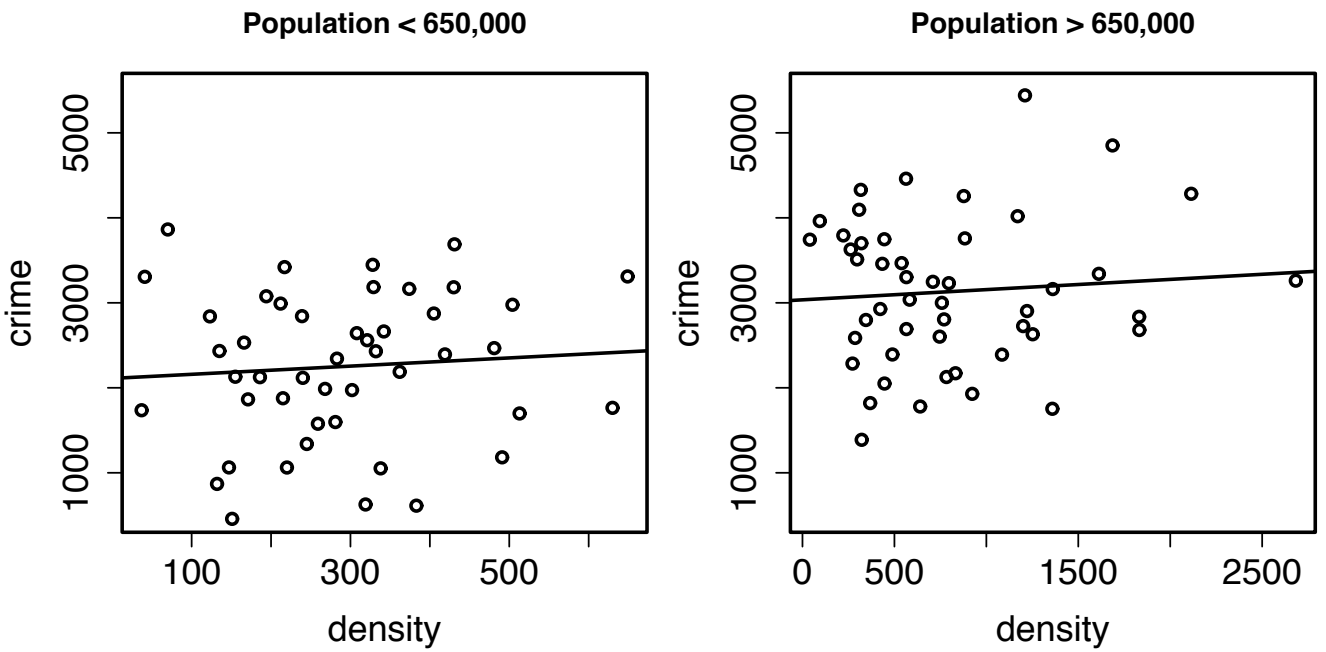


Illustrazione

- Questa associazione, però, è dovuta ad altri fattori che sono legati sia alla densità della popolazione che al tasso di criminalità.
- Per esempio, città con un grande numero di abitanti tendono ad avere sia un'alta densità della popolazione che alti tassi di criminalità.
- Se esaminiamo separatamente le città aventi approssimativamente lo stesso numero di abitanti, l'associazione tra *"density"* e *"crime"* scompare.

Illustrazione

Errore di specificazione



Conclusioni

Conclusioni

- Il modello di regressione semplice

$$\mathbb{E}(y) = \alpha + \beta(x)$$

  descrive la relazione che intercorre tra una variabile esplicativa X e la media $\mathbb{E}(y)$ della variabile dipendente.

- Il modello di regressione è appropriato quando c'è effettivamente una relazione lineare tra X e Y .

Conclusioni

- L'ispezione di un grafico a dispersione consente di stabilire se la relazione tra X e Y è approssimativamente lineare.
- Se questo è il caso, mediante il metodo dei minimi quadrati vengono calcolati i coefficienti $\hat{\alpha}$ e $\hat{\beta}$ della retta

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

  che interpola al meglio i dati.
- Il coefficiente di correlazione $-1 \leq r_{yx} \leq 1$ rappresenta la pendenza della retta di regressione per le variabili standardizzate z_x e z_y.

Conclusioni

- Il quadrato del coefficiente di correlazione r^2_{yx} è detto indice di determinazione e descrive l'intensità dell'associazione lineare tra le variabili X e Y.
- r^2_{yx} si interpreta come la riduzione proporzionale della variazione di Y attorno ai valori adattati $\hat{Y}_i$ rispetto alla variazione di Y attorno alla media $\bar{Y}$.

Conclusioni

- L'inferenza statistica sul modello di regressione verifica l'ipotesi nulla di indipendenza, ovvero l'ipotesi che, nella popolazione, la pendenza della retta di regressione sia uguale a zero.
- Più informativo del test statistico $H_0 : \beta = 0$ è l'intervallo di confidenza per il parametro β.

Conclusioni

Ipotesi nulla: H_0: indipendenza

	Scala di misura delle variabili		
	Nominale	Ordinale	Intervalli
Statistica test	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$	$z = \frac{\hat{\gamma}}{\sigma_{\hat{\gamma}}}$	$t_{n-2} = \frac{b}{\hat{\sigma}_b}$
Misura di associazione	odds ratio	$\hat{\gamma} = \frac{C - D}{C + D}$	$r = b\frac{s_x}{s_y}$
			$r^2 = \frac{SQ_{reg}}{SQ_{tot}}$