UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche,
aziendali, matematiche e statistiche
“Bruno de Finetti”

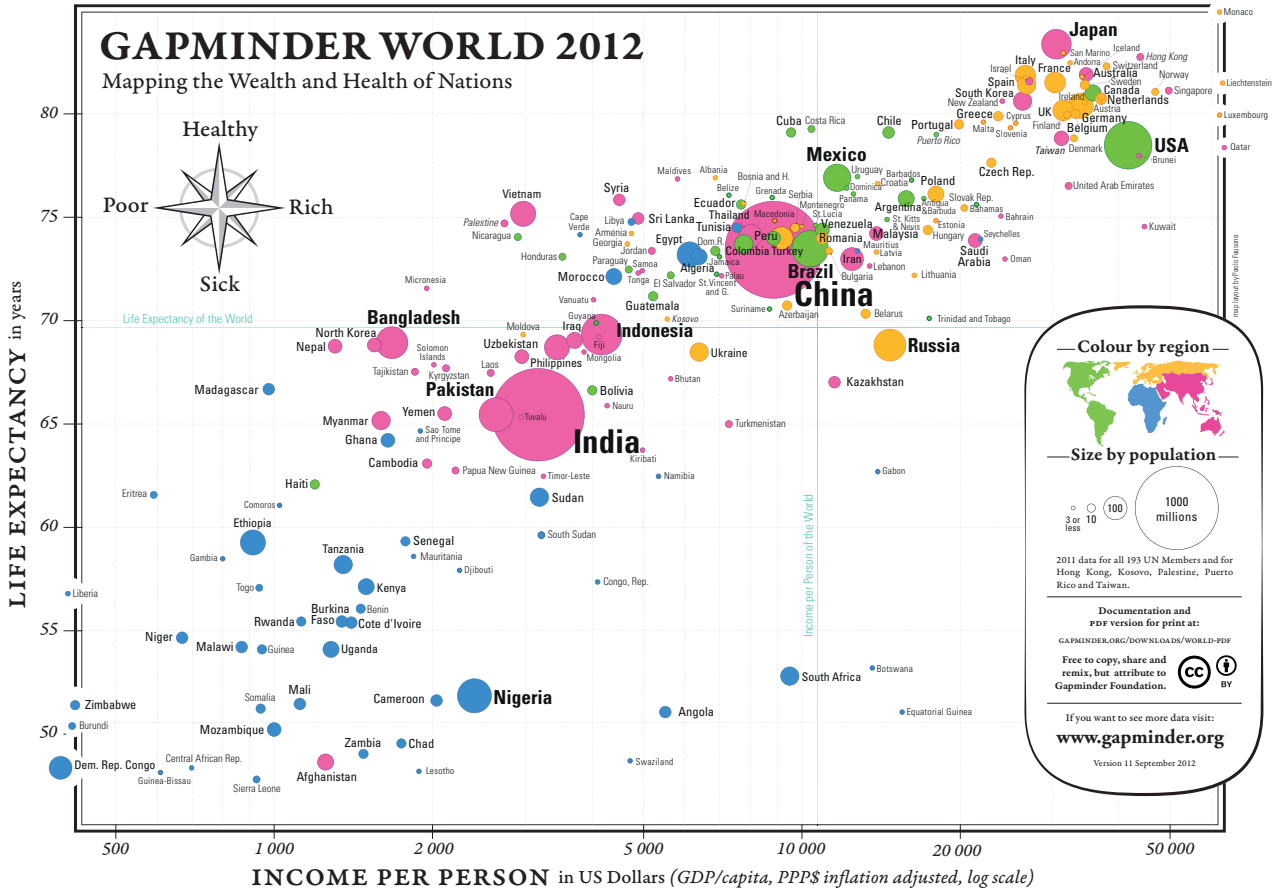
Statistica

Regressione

Francesco Pauli

A.A. 2016/2017

Ricchezza e salute



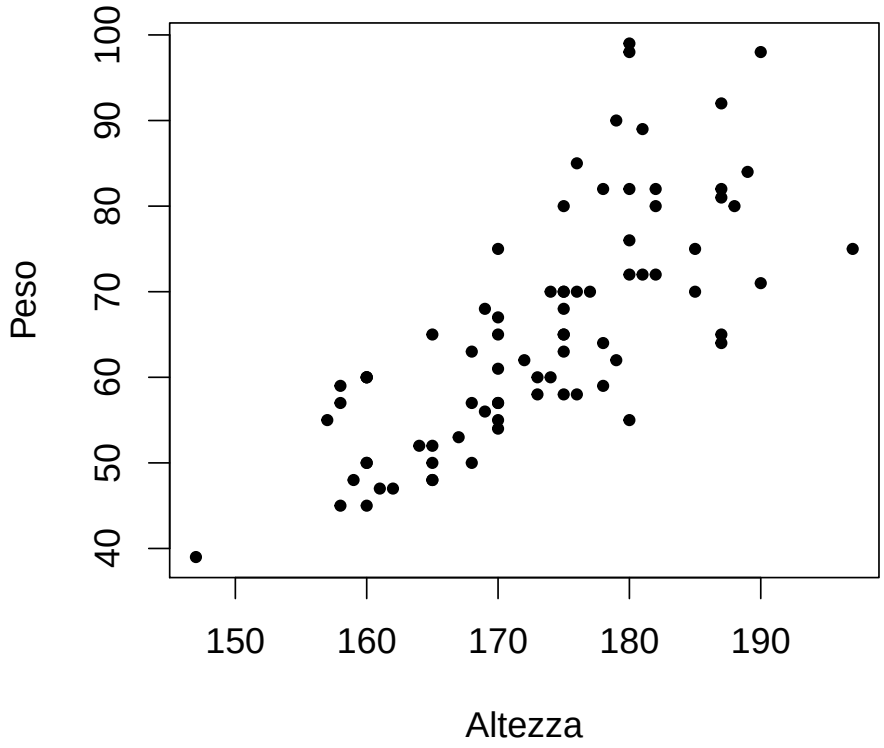
Un esempio più semplice: dati

	Altezza	Peso
1	160	60
2	174	60
3	173	58
4	168	63
5	175	68
6	170	61
7	179	90
8	165	65
9	160	60
10	158	45
11	176	70
12	170	75
13	158	59
14	180	55
15	197	75
16	181	89
17	190	98
18	157	55
19	180	98
20	170	57
21	187	92
22	182	72
23	160	60
24	181	72
25	165	52
26	164	52

	Altezza	Peso
27	187	82
28	174	70
29	158	57
30	180	82
31	178	59
32	180	76
33	169	68
34	168	57
35	185	75
36	147	39
37	161	47
38	190	71
39	170	65
40	160	50
41	187	81
42	167	53
43	185	70
44	182	80
45	173	60
46	180	72
47	175	58
48	188	80
49	165	48
50	189	84
51	187	64
52	187	65

	Altezza	Peso
53	170	54
54	170	55
55	180	99
56	175	63
57	175	70
58	175	65
59	165	50
60	162	47
61	178	64
62	165	48
63	159	48
64	160	45
65	175	70
66	178	82
67	170	57
68	182	82
69	169	56
70	168	50
71	172	62
72	175	65
73	176	85
74	177	70
75	176	58
76	179	62
77	160	50
78	170	67
79	175	80

Diagramma di dispersione: una relazione lineare



È evidente una relazione, che potrebbe essere adeguatamente descritta da una linea retta.

Un primo modello

Adottiamo l'ipotesi di una relazione lineare.

Possiamo allora pensare ad un modello del tipo

$$(\text{Peso}) = \alpha + \beta(\text{Altezza}) + (\text{errore})$$

dove l'errore esprime la parte delle oscillazioni del peso non legate all'altezza (o, meglio, che una funzione lineare dell'altezza non riesce a spiegare).

Un modello di questo tipo viene chiamato
modello di regressione lineare semplice.

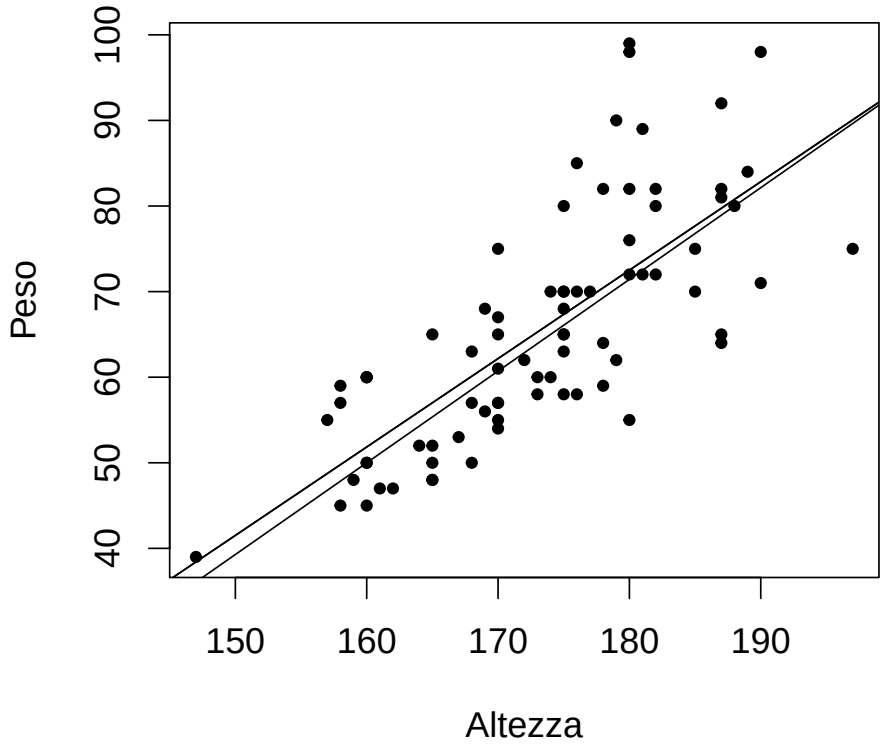
Modelli di regressione lineare semplice

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

- 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.

Rette "vicine" ai dati



Vogliamo una **linea retta** che si "avvicini" ai punti.

Minimi quadrati: idea

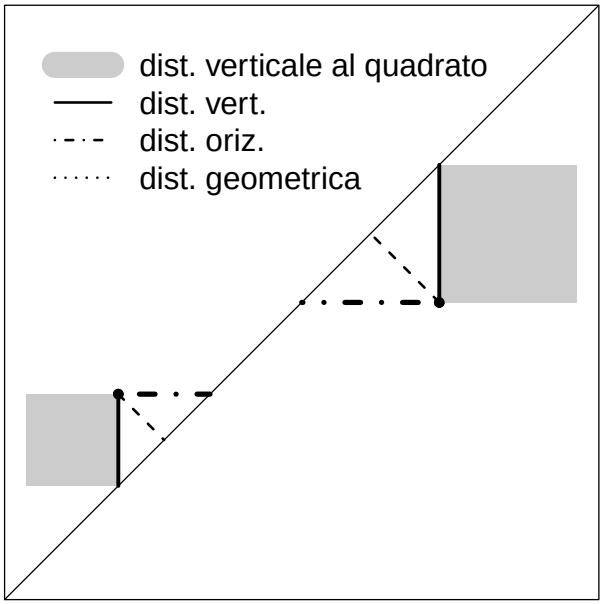
Sembra ragionevole scegliere per i parametri due valori, $\hat{\alpha}$ e $\hat{\beta}$, in modo tale che la retta di regressione "riproduca" bene i nostri dati, ovvero in modo tale che

$$\begin{aligned} y_1 &\approx \hat{\alpha} + \hat{\beta} x_1 \\ y_2 &\approx \hat{\alpha} + \hat{\beta} x_2 \\ &\vdots \\ y_N &\approx \hat{\alpha} + \hat{\beta} x_N \end{aligned}$$

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.

Distanza tra i punti e la retta



Diverse opzioni sono ragionevoli

Appare logico guardare alle distanze verticali, visto che vediamo la retta come approssimazione del valore di y.

È matematicamene conveniente usare i quadrati delle distanze.

Retta dei minimi quadrati

Scegliamo 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**.

Proprietà dei minimi quadrati

La media aritmetica minimizza la somma (e quindi la media) dei quadrati degli scarti delle singole osservazioni da una costante a.
Sia a un numero qualsiasi. Allora

$$\sum_{i=1}^{N}(y_i - a)^2 = \sum_{i=1}^{N}(y_i - \overline{y})^2 + N(\overline{y} - a)^2 \tag{1}$$

Dimostrazione.
NB: tutte le sommatorie vanno da 1 a N

$$\begin{aligned}
\sum(y_i - a)^2 &= \sum(y_i - a + \overline{y} - \overline{y})^2 = \\
&= \sum[(y_i - \overline{y}) + (\overline{y} - a)]^2 = \\
&= \sum\left[(y_i - \overline{y})^2 + (\overline{y} - a)^2 + 2(\overline{y} - a)(y_i - \overline{y})\right] = \\
&= \sum(y_i - \overline{y})^2 + \sum(\overline{y} - a)^2 + 2(\overline{y} - a)\sum(y_i - \overline{y}) = \\
&= \sum(y_i - \overline{y})^2 + N(\overline{y} - a)^2 + 2(\overline{y} - a) \times 0.
\end{aligned}$$

Quindi, quando $a = \overline{y}$, la quantità $\sum(y_i - a)^2$ è minima.

Minimi quadrati: determinazione dei parametri

(1) Fissato β ad un qualunque valore, il problema diventa

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

con $z_i = y_i - \beta x_i$.

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 (continua)

(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 (continua)

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

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

Esercizio: verificare che questa soluzione corrisponde ad un punto di minimo.

Si noti che

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

Minimi quadrati: determinazione dei parametri (continua)

Quindi

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

$$\hat{\beta} = \frac{\sigma_{XY}}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}$$

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.

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.

Calcolo delle stime

Sostituendo i valori del campione si ha $n = 79$ e

$$\frac{1}{N}\sum_{i=1}^{n} x_i = \frac{1}{79}1.3688 \times 10^4 = 173.27, \qquad \frac{1}{N}\sum_{i=1}^{n} y_i = \frac{1}{79}5178 = 65.544,$$

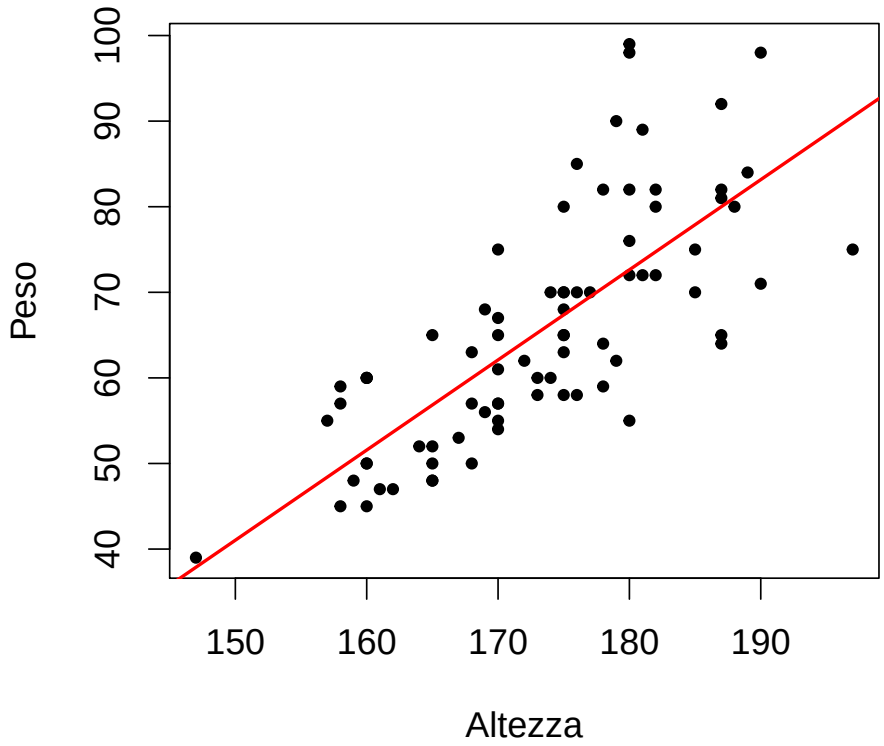
$$\frac{1}{N}\sum_{i=1}^{n} x_i^2 = \frac{1}{79}2.3791 \times 10^6 = 3.0115 \times 10^4, \qquad \frac{1}{N}\sum_{i=1}^{n} x_i y_i = \frac{1}{79}9.049 \times 10^5 = 1.1454 \times 10^4,$$

Si ha allora

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i/n - \bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2/n - \bar{x}^2} = \frac{9.049 \times 10^5/79 - 173.27 \times 65.544}{2.3791 \times 10^6/79 - 173.27^2} = 1.0531$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 65.544 - 1.0531 \times 173.27 = -116.93,$$

Diagramma di dispersione con retta di regressione



La capacità di descrivere le variazione del peso sembra discreta, naturalmente c'è una certa variabilità intorno alla retta stessa.

Valori osservati, previsti (o teorici), residui

Le seguenti quantità sono di interesse.

- y_i valore **osservato** per Y sulla i-sima unità statistica
- $\hat{y}_i$ valore **previsto** (o teorico) 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: $r_i = y_i - \hat{\alpha} - \hat{\beta} x_i$.

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

Modelli stimato, valori teorici, residui

Il modello stimato è

$$y_i = -116.93 + 1.0531 x_i + (\text{errore}).$$

Valori teorici:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

ad esempio, $x_5 = 168$ e quindi

$$\hat{y}_5 = -116.93 + 1.0531 \times (168) = 60.11.$$

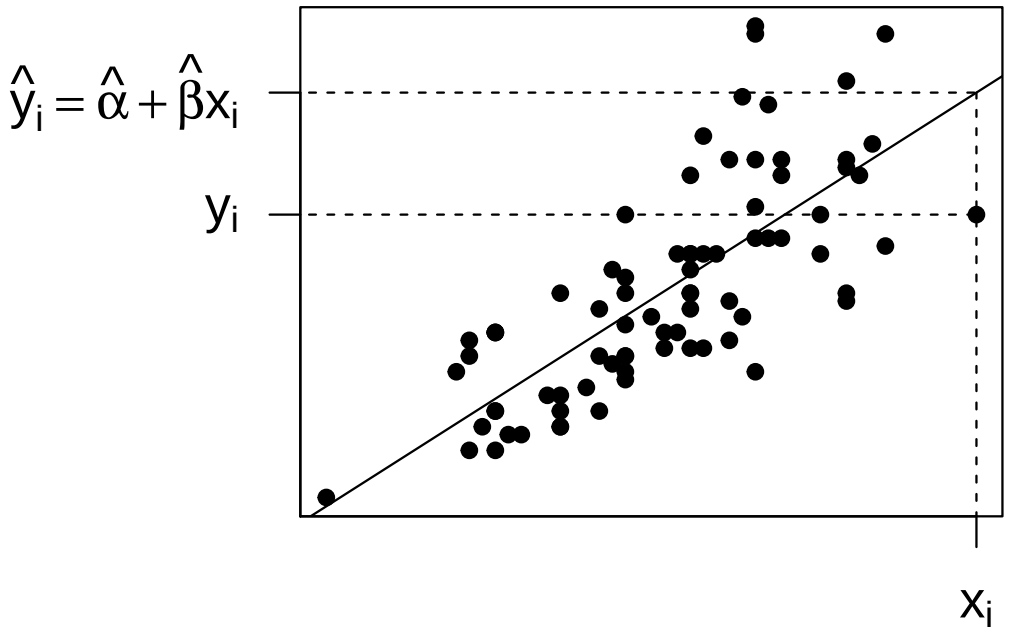
Residui

$$r_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

ad esempio

$$r_5 = 57 - 60.11 = -3.11.$$

Graficamente



Proprietà dei residui: media

È facile verificare che la media dei residui è nulla.

$$
\begin{aligned}
\sum_{i=1}^{N} r_i &= \sum_{i=1}^{n} (y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) \\
&= \sum_{i=1}^{N} y_i - N\hat{\alpha} - \hat{\beta} \sum_{i=1}^{N} x_i \\
&= N\overline{y} - N(\overline{y} - \hat{\beta}\overline{x}) - N\hat{\beta}\overline{x} \\
&= 0
\end{aligned}
$$

Varianza dei residui

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

Scomposizione della varianza

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2
\end{aligned}$$

poiché

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)(\hat{\alpha}+\hat{\beta}x_i-\bar{y}) \\
&= \sum_{i=1}^{n}(y_i-\bar{y}+\hat{\beta}\bar{x}-\hat{\beta}x_i)(\bar{y}-\hat{\beta}\bar{x}+\hat{\beta}x_i-\bar{y}) \\
&= \sum_{i=1}^{n}(y_i-\bar{y}-\hat{\beta}(x_i-\bar{x}))\hat{\beta}(x_i-\bar{x})=0 \quad \text{cfr Passo 2 sopra}
\end{aligned}$$

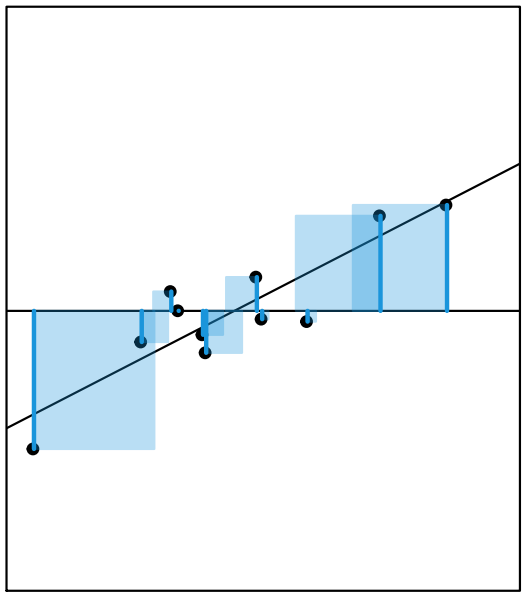
Scomposizione della varianza

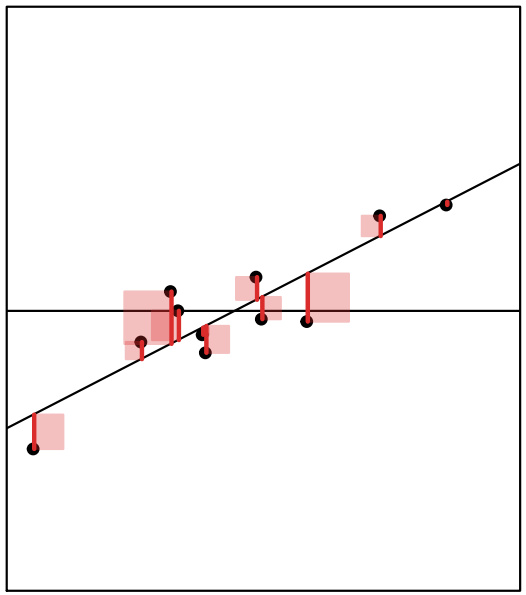
$$\frac{1}{N}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad = \quad \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad + \quad \frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

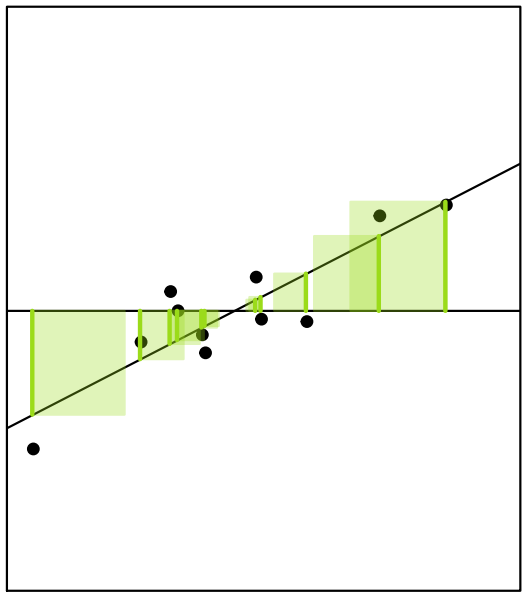
varianza totale

varianza residui

varianza valori teorici (spiegata)







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.
- 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: peso e altezza

$$\begin{aligned}\overline{y} &= 65.544\\ \sigma_X^2 &= 92.507\\ \sigma_{XY} &= 97.191.\end{aligned}$$

Inoltre

$$\sum_{i=1}^{n} y_i^2 = 3.5377 \times 10^5$$

Quindi

$$\sigma_Y^2 = 3.5377 \times 10^5/79 - 65.544^2 = 182.08$$

e perciò

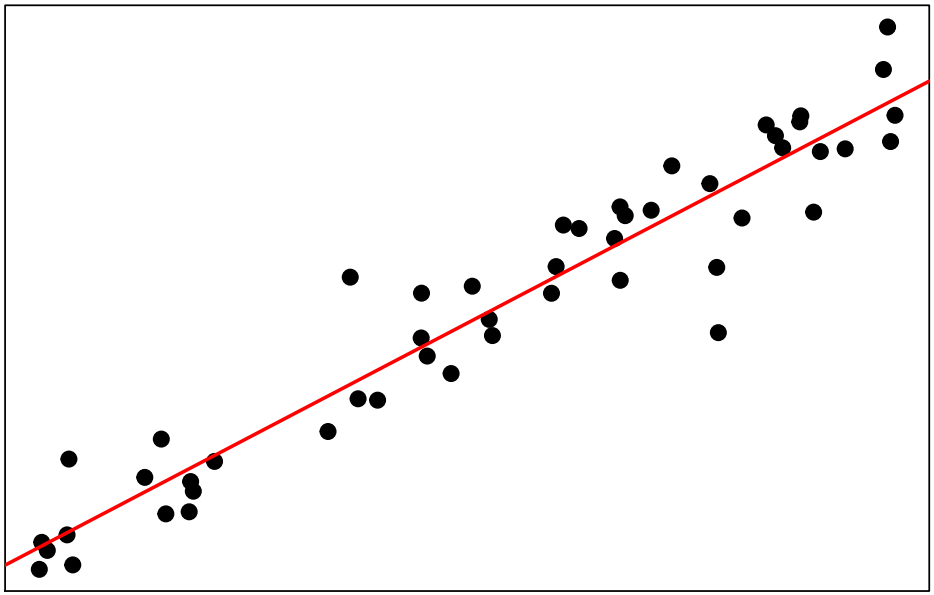
$$\sigma_R^2 = 182.08 - 97.191^2/92.507 = 79.9678303$$

Il coefficiente di determinazione vale

$$R^2 = 1 - 79.9678303/182.08 = 0.5608094,$$

ovvero il modello spiega il 56.1% della varianza della risposta.

Esempio: forte relazione lineare



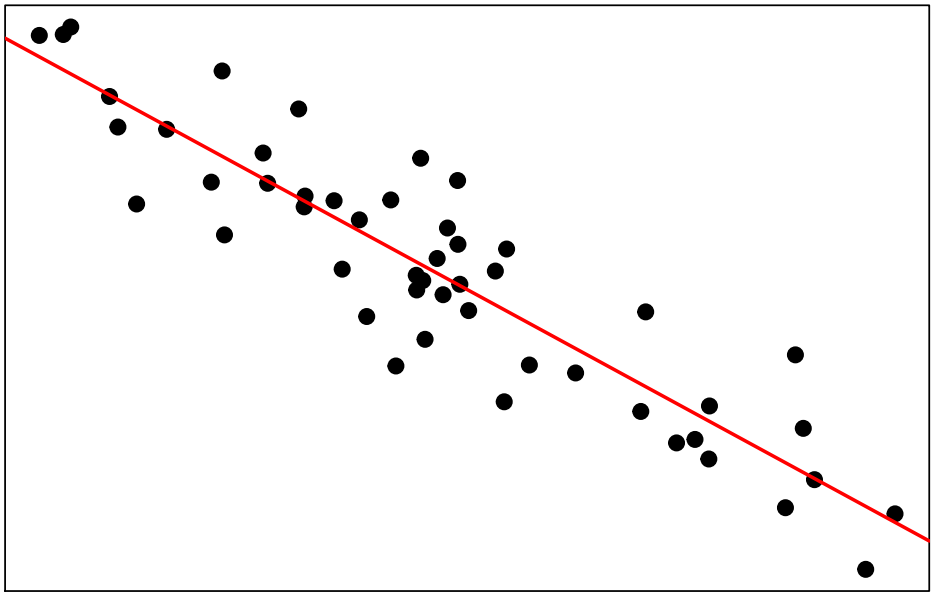
$$\rho = 0.9581$$

$$R^2 = 0.9179$$

$$\hat{\alpha} = -0.015458$$

$$\hat{\beta} = 1.0478$$

Esempio: forte relazione lineare bis



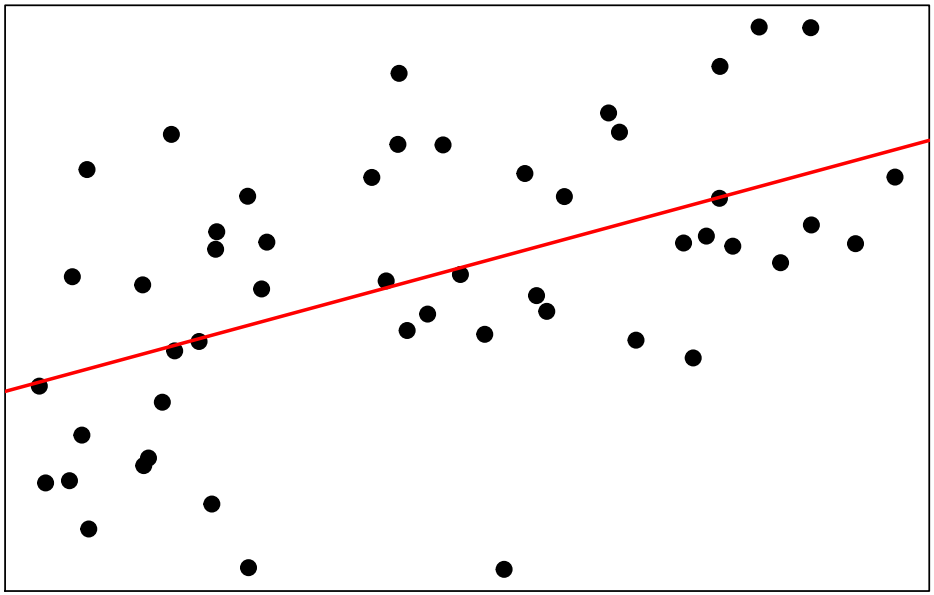
$$\rho = -0.9135$$

$$R^2 = 0.8344$$

$$\hat{\alpha} = -0.0097761$$

$$\hat{\beta} = -0.9788$$

Esempio: relazione lineare



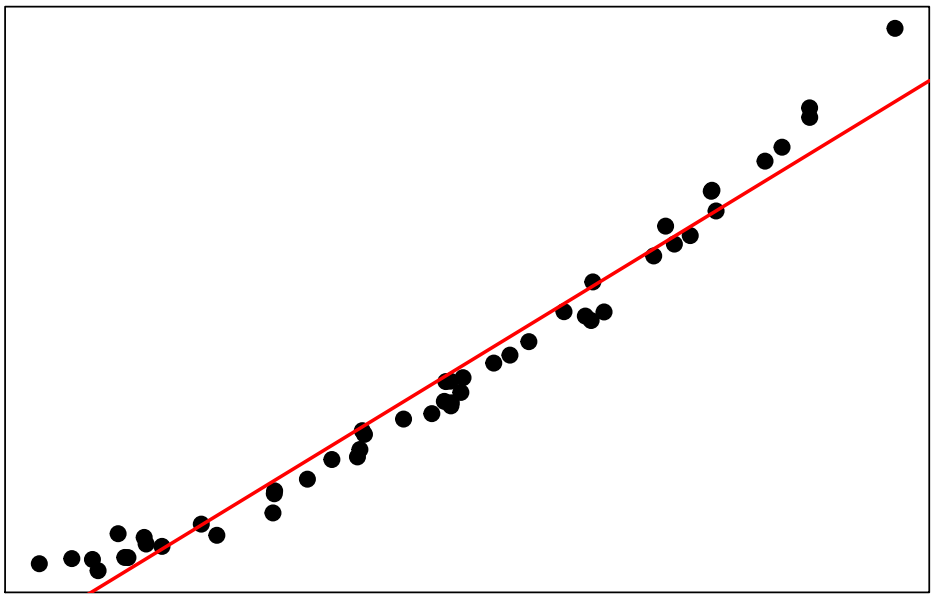
$$\rho = 0.5132$$

$$R^2 = 0.2634$$

$$\hat{\alpha} = 0.014617$$

$$\hat{\beta} = 0.91287$$

Esempio: relazione forte, ben approssimata da una retta, ma non lineare



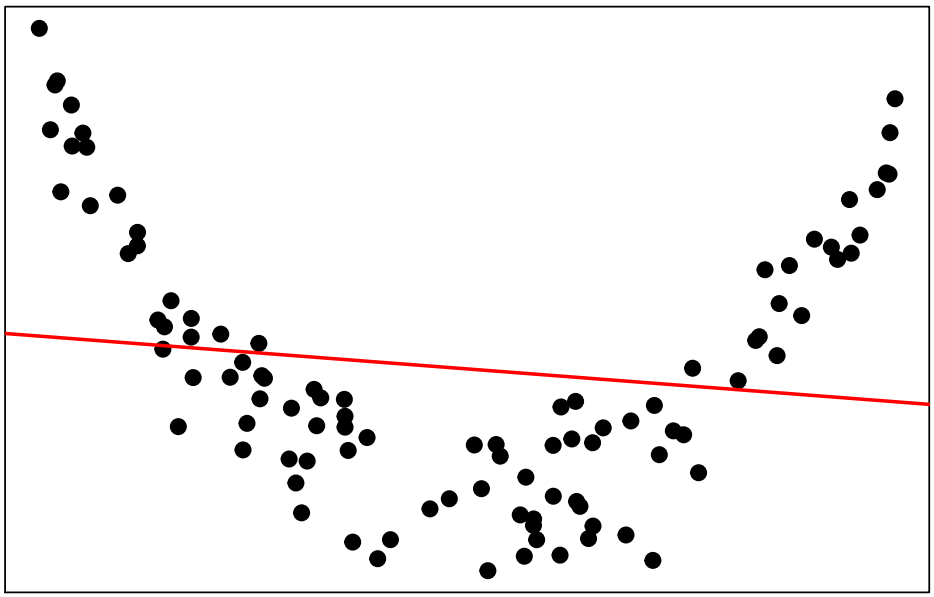
$$\rho = 0.9833$$

$$R^2 = 0.9669$$

$$\hat{\alpha} = -7.1706$$

$$\hat{\beta} = 5.6811$$

Esempio: relazione forte, non lineare al punto che non è rilevata



$$\rho = -0.146$$

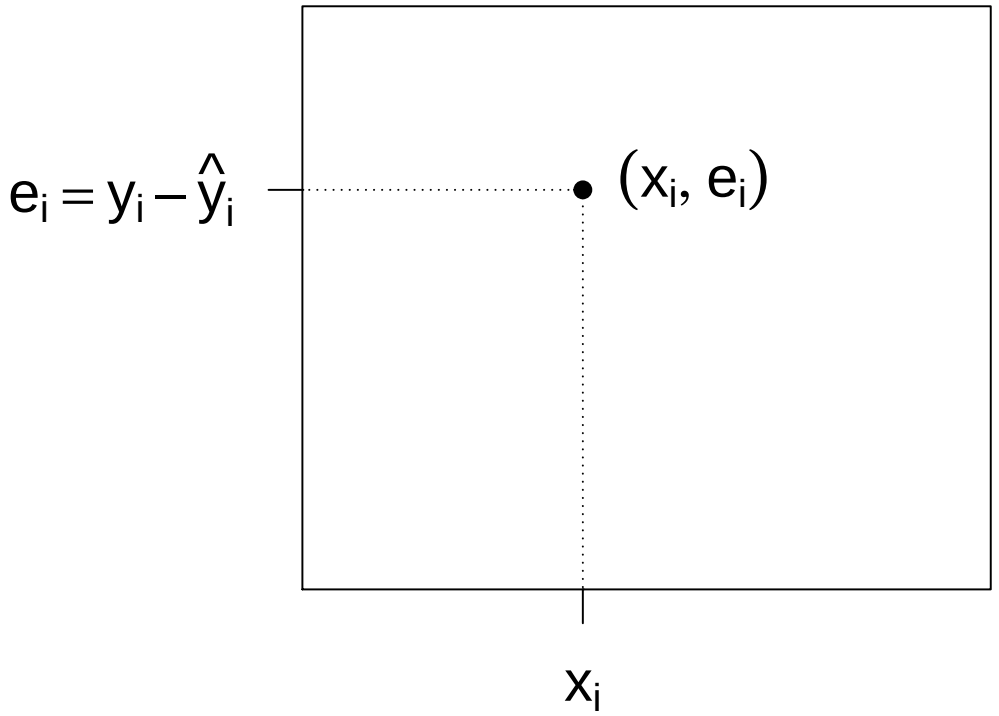
$$R^2 = 0.02133$$

$$\hat{\alpha} = 0.31443$$

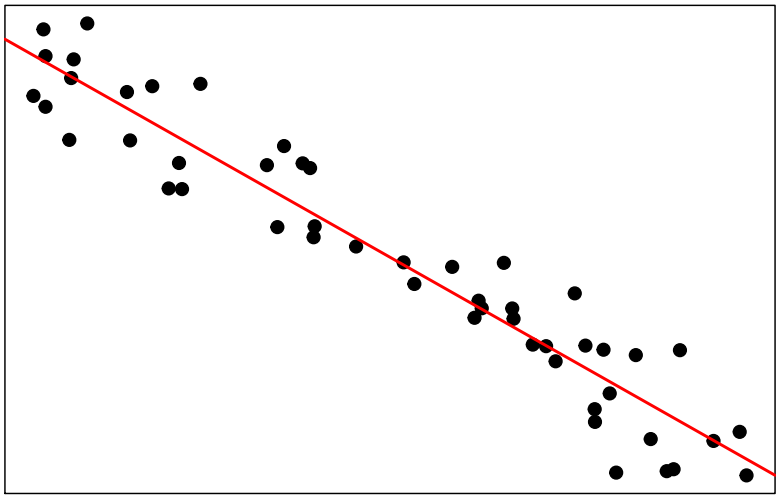
$$\hat{\beta} = -0.08578$$

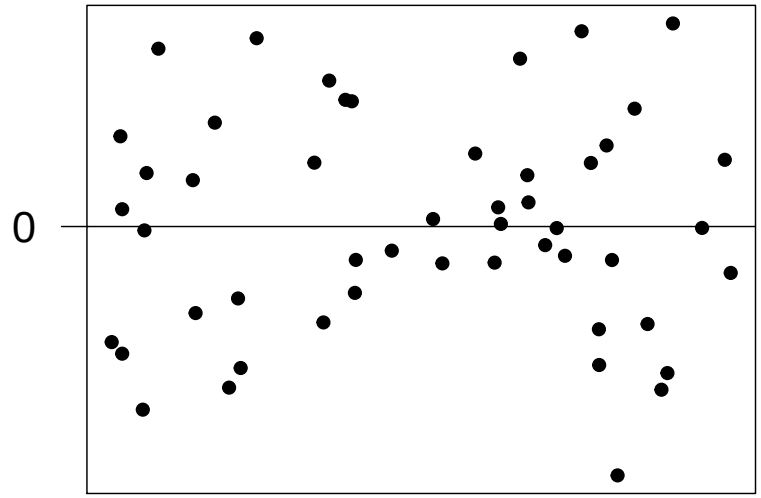
Rilevare le non linearità

L'analisi grafica negli esempi sopra è più che sufficiente, tuttavia è spesso utile fare un grafico dei residui contro i valori delle esplicative.

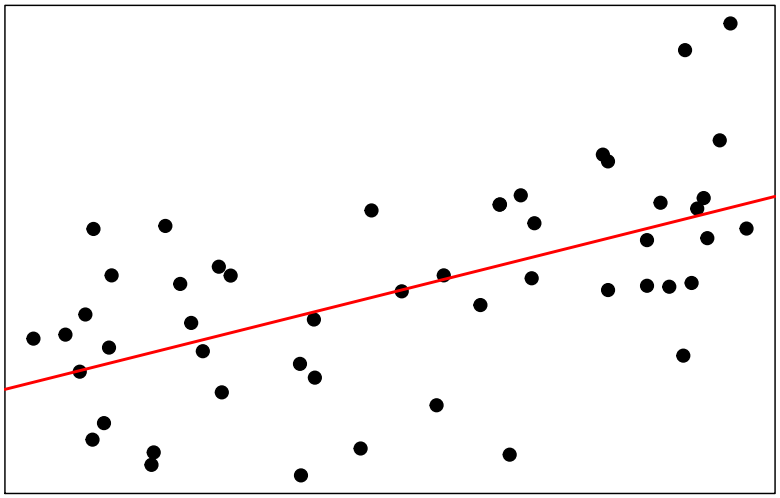


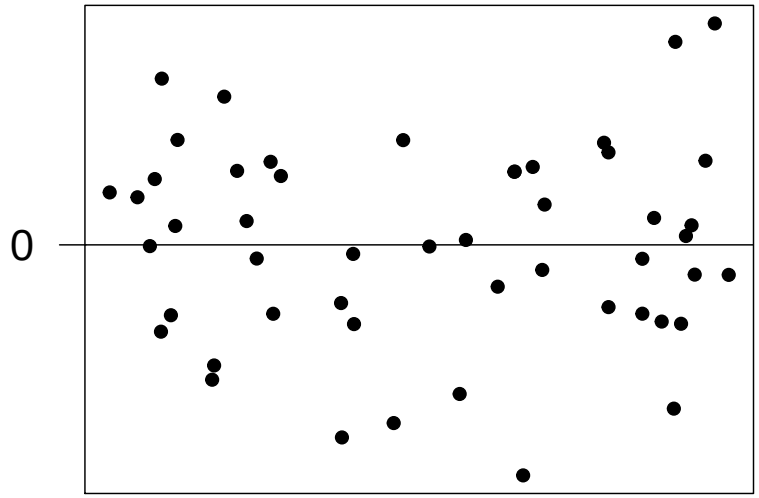
Esempi



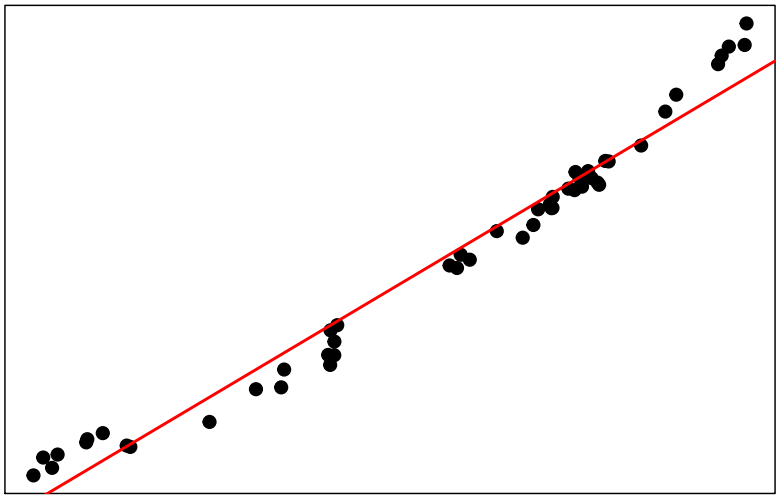


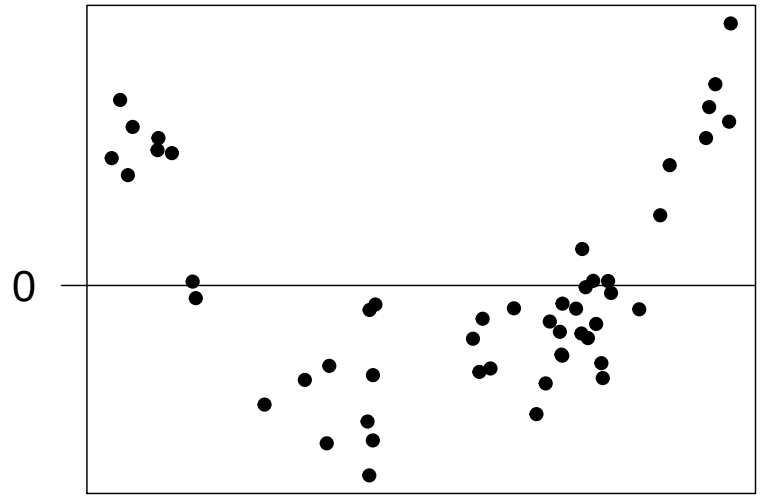
Esempio 1



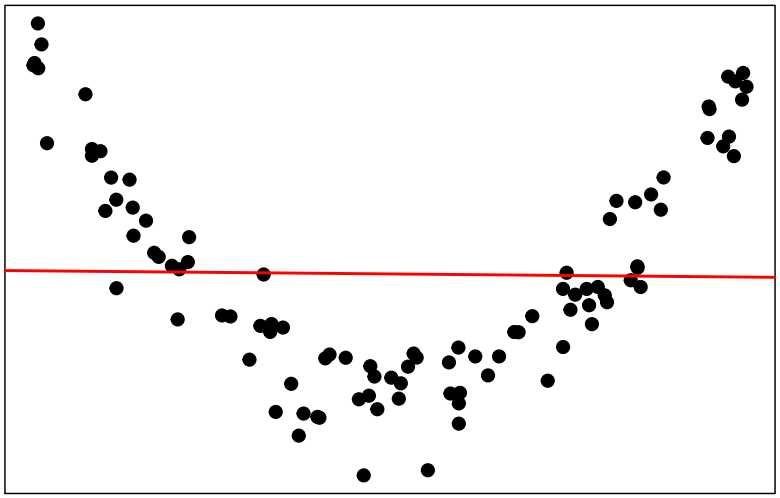


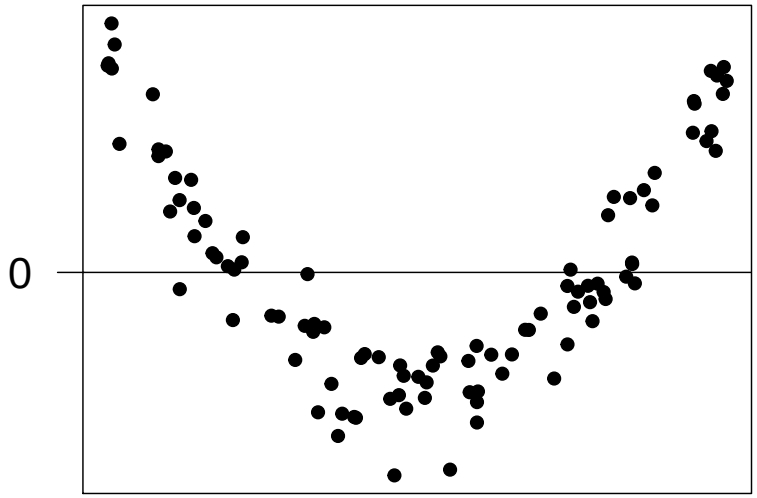
Esempio 2





Esempio 3



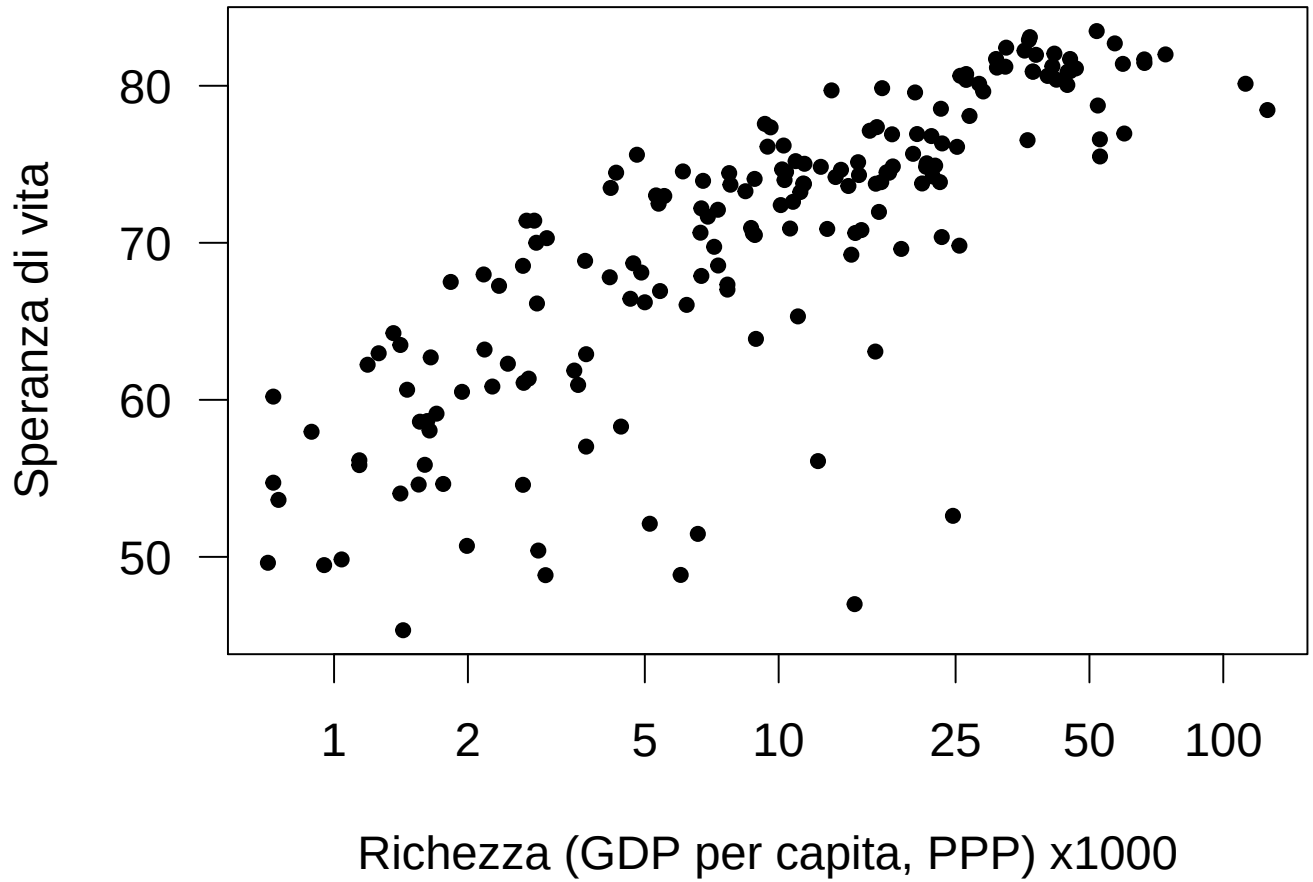


Ricchezza e salute

Semplifichiamo, e consideriamo solo le due grandezze

ricchezza GDP per capita Purchasing Power Parity

salute Speranza di vita



Retta di regressione

Stimiamo la relazione

$$(\text{salute}) = \alpha + \beta \log_{10}(\text{ricchezza}) + (\text{errore})$$

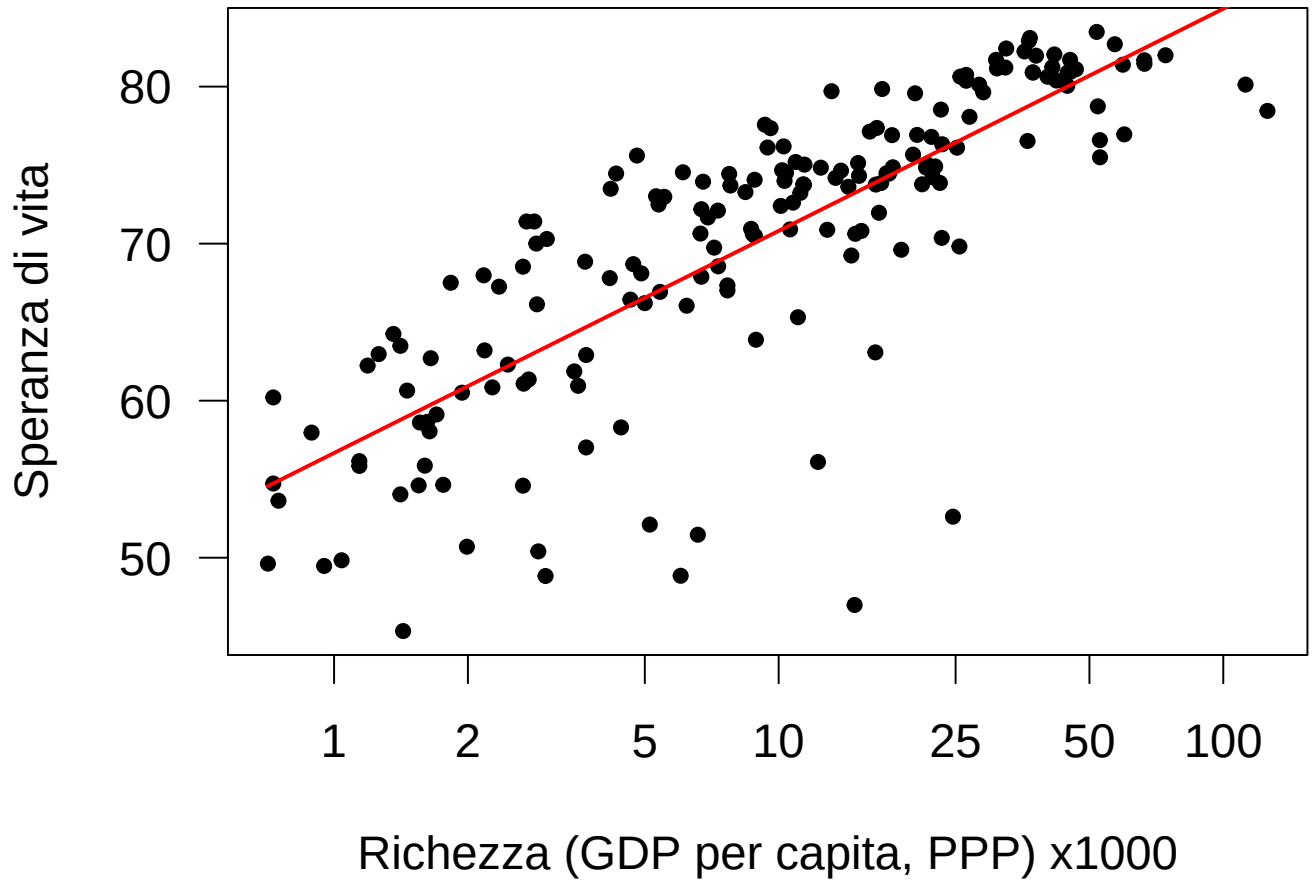
ne risulta

$$\hat{\alpha} = 14.235, \quad \hat{\beta} = 14.144$$

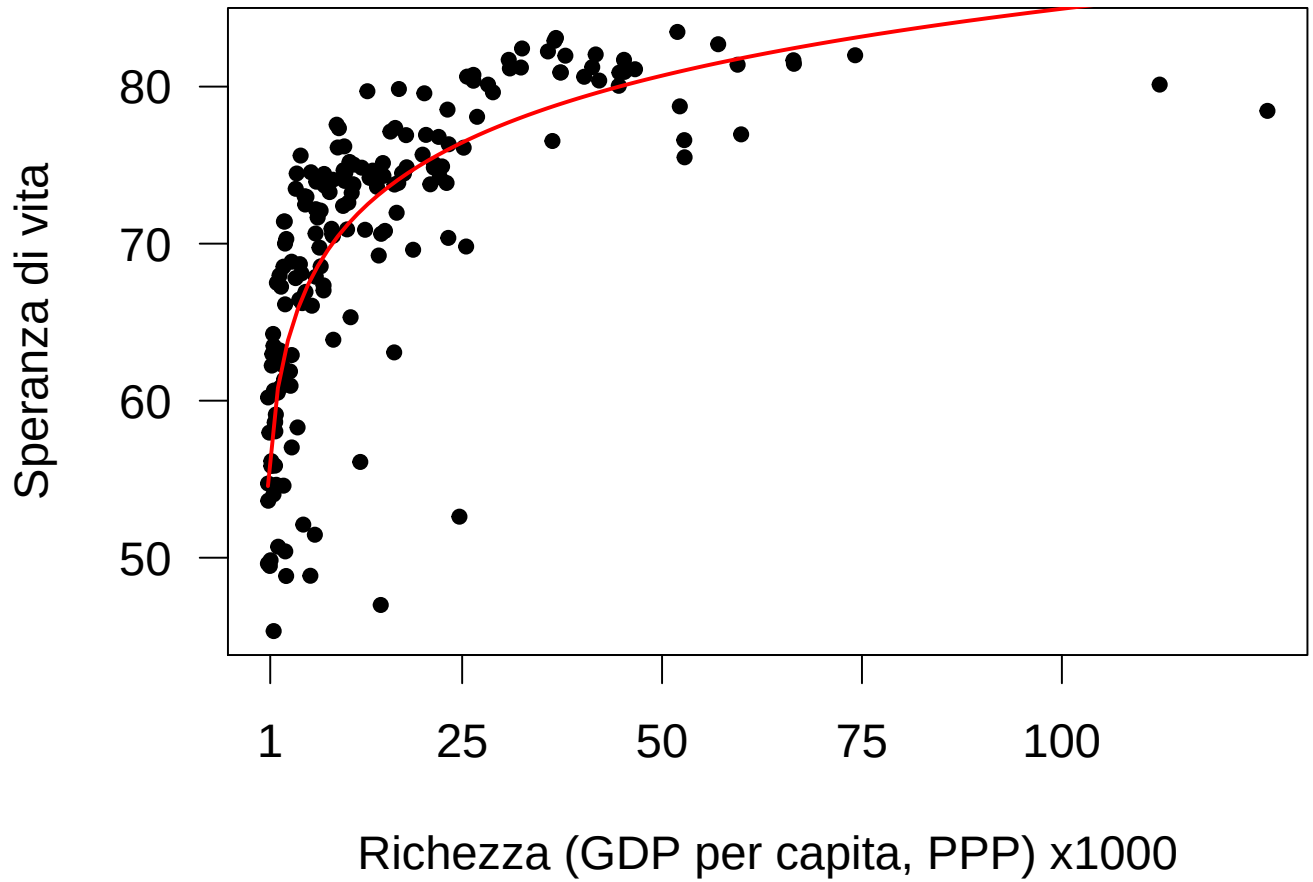
con

$$R^2 = 0.6326$$

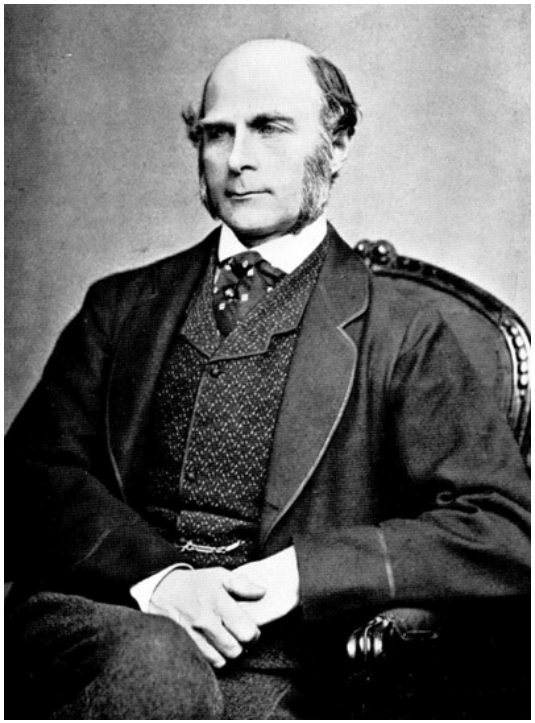
Retta di regressione



... su scala orginale

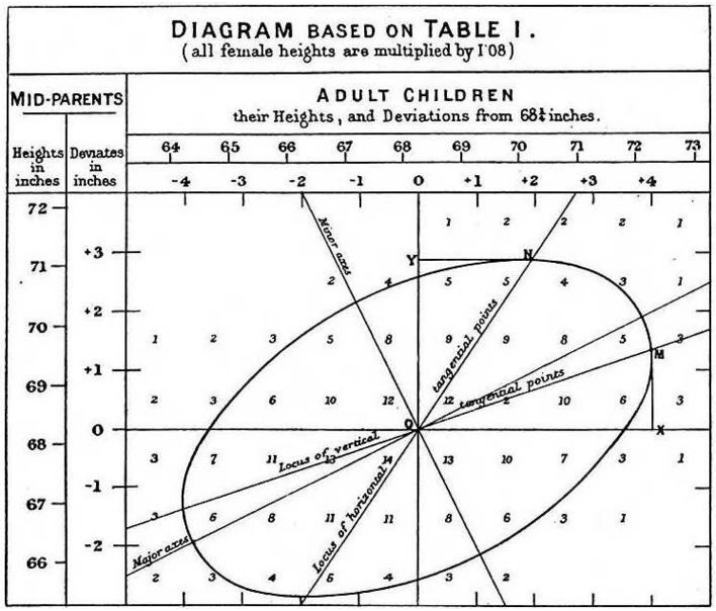


Nascita della regressione (e il perché del nome)



Il metodo è stato sviluppato intorno al 1888 da Sir Francis Galton (1822-1911) sociologo, psicologo, antropologo, esploratore, geografo, ecc. e anche statistico.

Nascita della regressione (e il perché del nome)



Il metodo è stato sviluppato intorno al 1888 da Sir Francis Galton (1822-1911) sociologo, psicologo, antropologo, esploratore, geografo, ecc. e anche statistico.

Tra le altre cose lavora sulla relazione tra le altezze dei genitori e dei figli, in figura una sua rappresentazione (non usuale oggi) dei dati.

Galton nota che

- genitori alti tendono ad avere figli alti,
- ma non così alti come loro.

chiama questo fenomeno 'regressione verso la media'.

Regressione verso la media

Partiamo dalle osservazioni (x_i, y_i) e standardizziamole, consideriamo cioè

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sigma_x}, \quad \tilde{y}_i = \frac{y_i - \bar{y}}{\sigma_y}$$

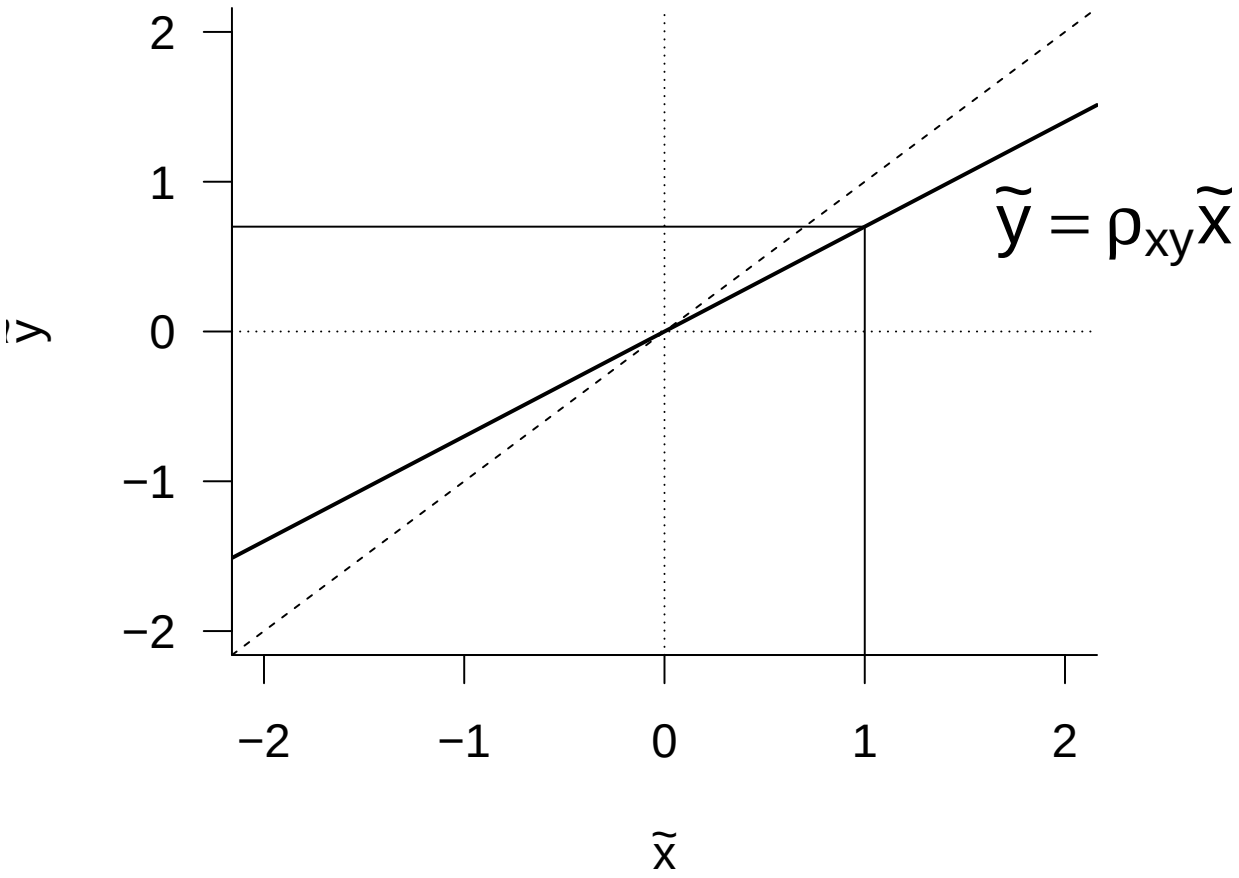
x^* e y^* misurano, in unità di deviazione standard, lo scarto di un'osservazione dalla media.

La retta di regressione tra x^* e y^* è

$$\tilde{y} = \rho_{xy}\tilde{x}$$

dove, si noti, $\rho_{xy} = \rho_{\tilde{x}\tilde{y}}$.

Regressione verso la media

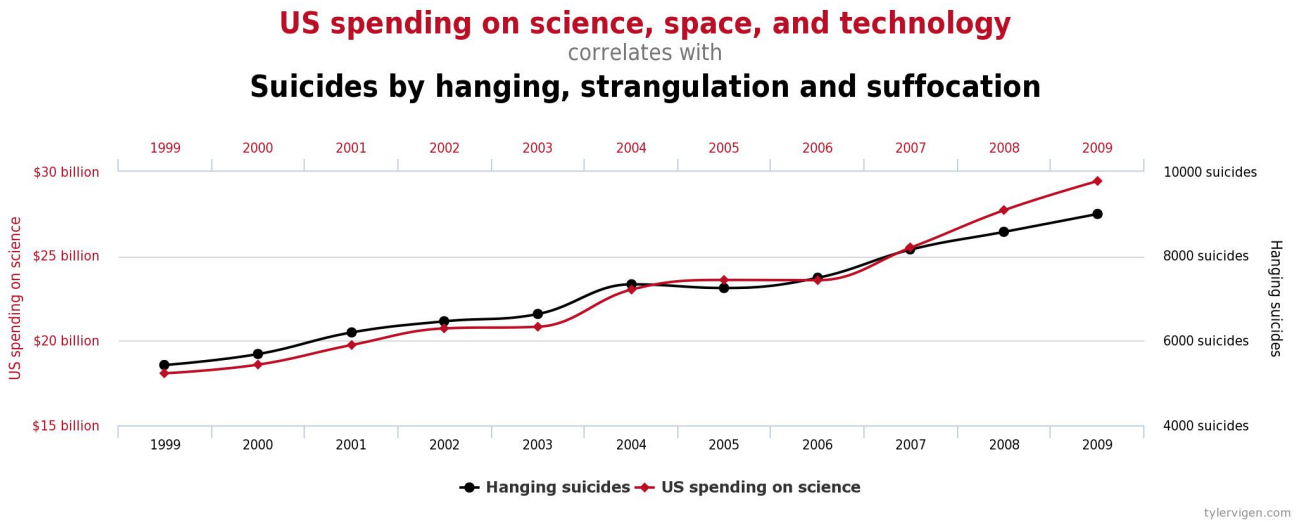


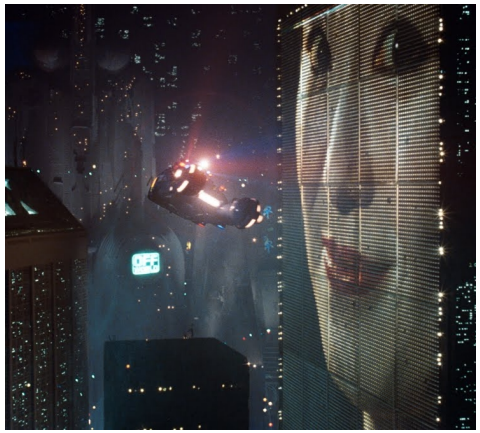
Relazione $\not\Rightarrow$ causa ed effetto

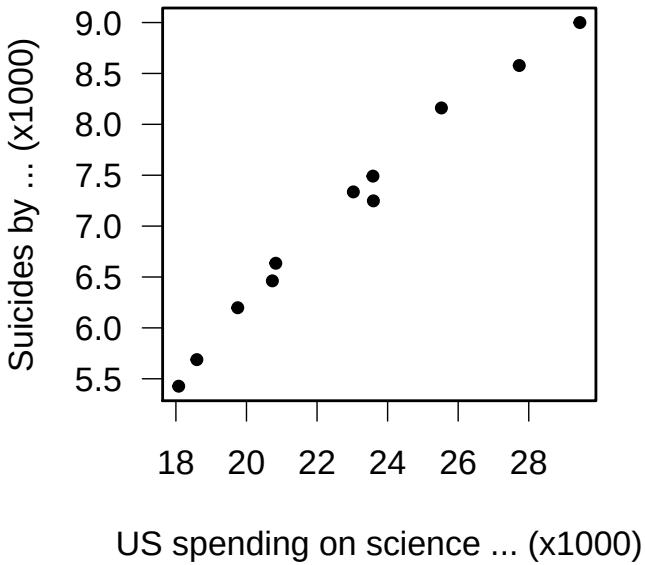
Attenzione all'interpretazione!

- Quando mettiamo in relazione due variabili e troviamo una forte associazione è forte la tentazione di interpretarlo come se x "causasse" y o viceversa.
- Una relazione statistica, anche stretta, tra y e x **non implica** una relazione causa effetto.
- Per esempio, entrambe potrebbero essere legate a una terza variabile, che le "causa" entrambe.
- Ci sono metodi statistici per l'inferenza su relazioni causa effetto ma richiedono una maggiore sofisticazione oppure un campione costruito in un certo modo.
- I prossimi esempi vengono dal sito www.tylervigen.com

La scienza rende il mondo arido e triste



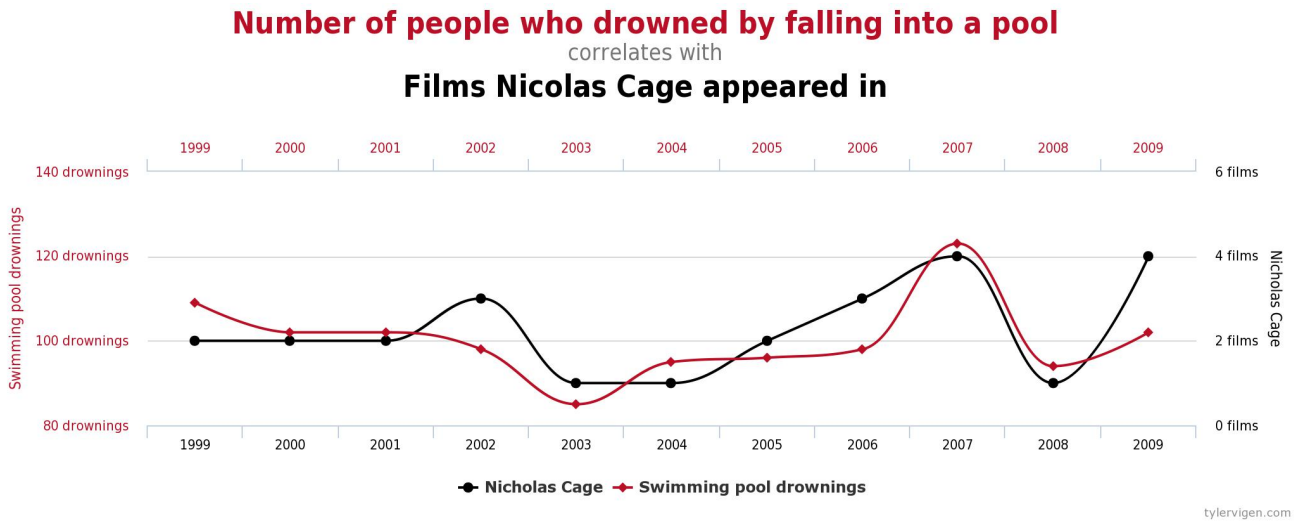


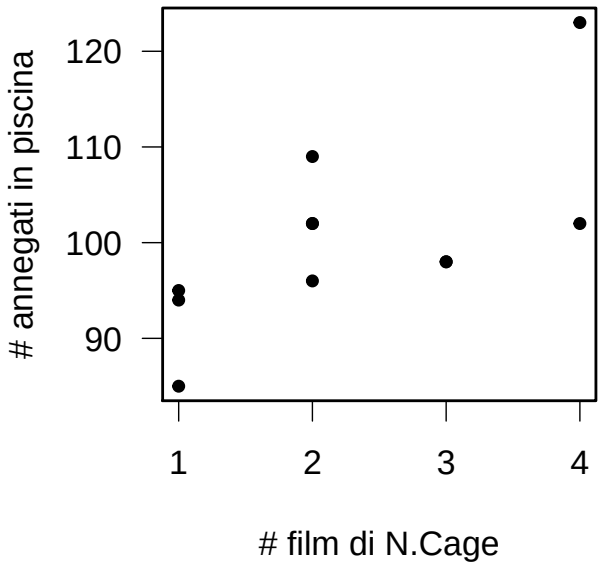


La correlazione è 0.992, ma diminuire la spesa per scienza e tecnologia non è una strategia per diminuire i suicidi.

Quale potrebbe essere una spiegazione per questa correlazione?

Nicolas Cage un pericolo per i nuotatori (in piscina)





La correlazione è 0.666 ma ...

Notare come il grafico sopra suggerisca una relazione molto stretta, ridimensionata dal diagramma di dispersione: quello sopra **non è un buon modo di rappresentare due serie storiche.**

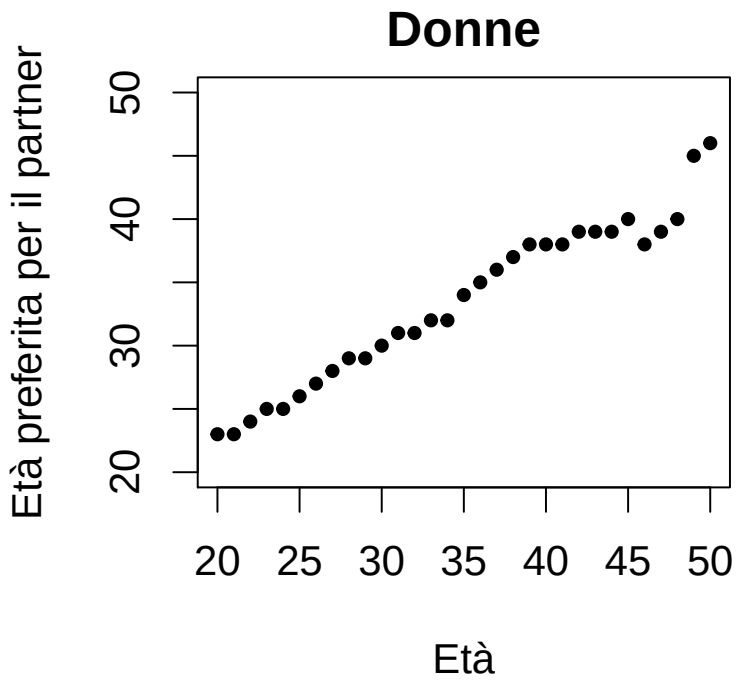
da Repubblica - Donna

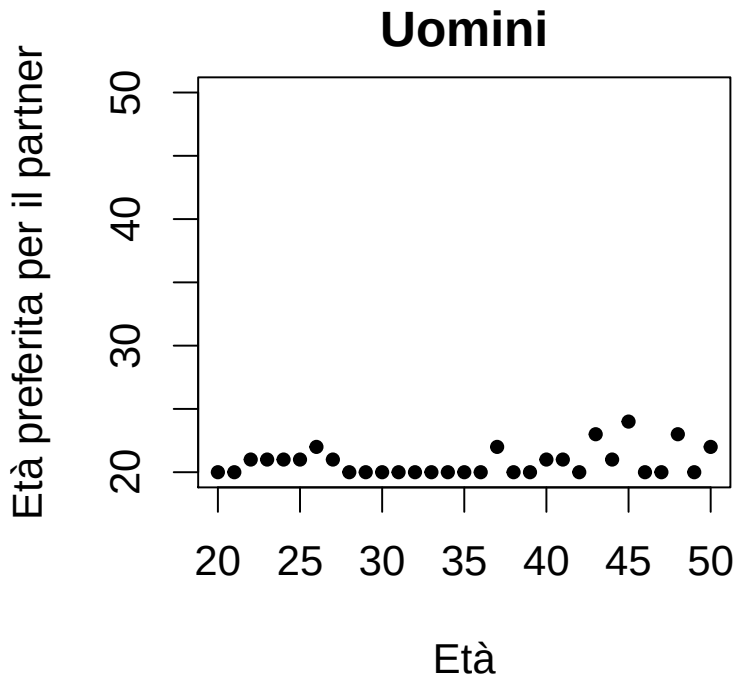
L'algoritmo di Cupido

Gli uomini preferiscono le ventenni. I dubbi già erano pochi ma ora la conferma arriva da OKCupid, che raccoglie statistiche e dati legati ai profili di uomini e donne che cercano un partner. Ne è nato così una sorta di algoritmo di Cupido ovvero un grafico che mostra l'andamento tra l'età dell'uomo e quella della donna che si sceglie. Secondo questo schema l'età media delle donne desiderate dai maschi si attesta intorno ai 22 anni e non supera i 24. Ma se dai 20 ai 30 anni d'età appare normale che un uomo cerchi una compagna di quell'età quando il grafico sale ai 50 anni la situazione cambia. Allo stesso modo un altro grafico mostra come le donne invece cerchino partner più vicini alla loro età con uno scarto in più o in meno di 2/5 anni. Il presidente di OKCupid, Christian Rudder, autore di quattro libri sulla psicologia legata alle coppie, è rimasto sorpreso da questi risultati. "Le donne però non devono prendere questi risultati alla lettera perché rispecchiano solo i desideri degli uomini e non quello che poi realmente mettono in pratica. Quando guardiamo le statistiche delle coppie notiamo che uomini di 40 anni di solito si legano a donne di 35. Però certo si nota come l'uomo medio 42enne accetti facilmente una compagna di 15 anni più giovane ma difficilmente una che superi i 45 anni". E anche i vip non sfuggono alla regola...

L'analisi statistica

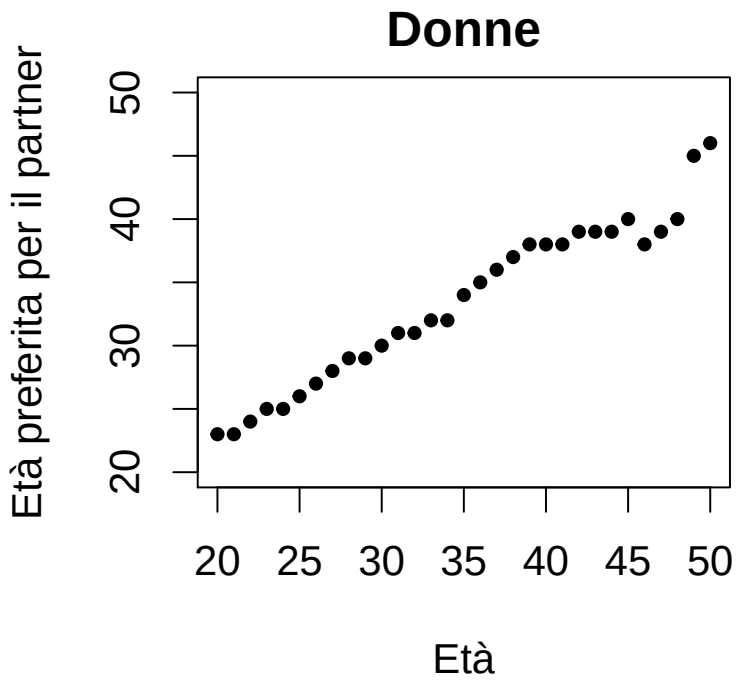
Età preferita per il partner rispetto alla propria, distinguendo uomini e donne.



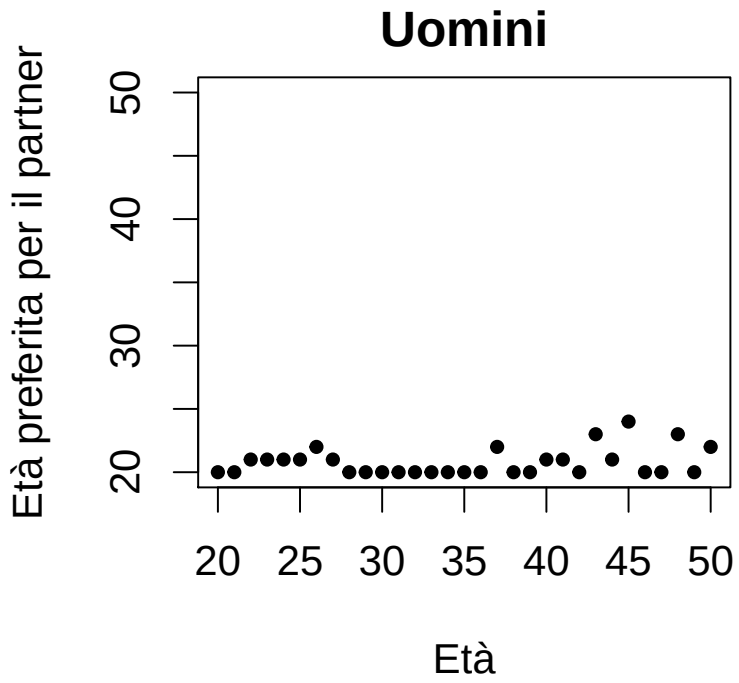


L'analisi statistica

Età preferita per il partner rispetto alla propria, distinguendo uomini e donne.



$r = 0.982$



$r = 0.287$