Regressione lineare

Regressione semplice

- L'idea di modello lineare che abbiamo visto per l'anova può essere generalizzata e meglio articolata
- In particolare quando voglio rappresentare (spiegare) la parte lineare della relazione esistente tra due (o più) fenomeni posso costruire un *modello di regressione*
- Se ho una sola variabile dipendente Y ed una sola variabile indipendente X costruisco **il modello di regressione semplice**
- In questo modello assumo che fissati i valori osservati per la X la media di Y ($E(Y|X)$) sia descrivibile con una retta che dipende da X ovvero

$$E(Y|X) = \alpha + \beta X$$

Regressione semplice

Quindi assumendo che su ciascuna osservazione esista uno scostamento da questa media si ha

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

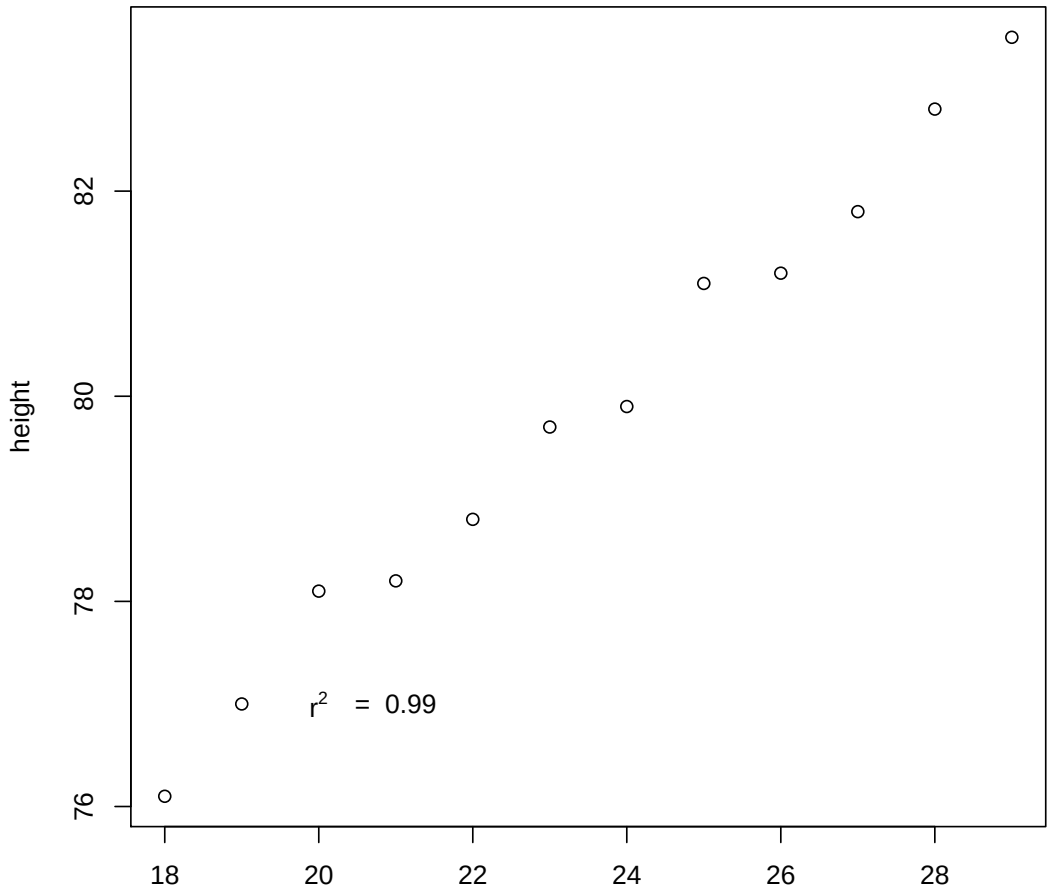
Le ipotesi alla base del modello sono le stesse che abbiamo già visto per l'analisi della varianza.
β è il coefficiente di regressione e misura l'intensità e la direzione (positivo o negativo) della relazione tra Y ed X

Regressione semplice: Esempio 1

- **Age and height of Egyptian children**
- **Data**: ageheight.txt.
- **Description**: Obviously the height of a child is not constant, but increases over time. On the other hand it is well-known that the growth pattern varies between children. In this dataset the focus is on determining the general growth pattern. One way to explore this is by using the average of several childrens heights, as presented in this dataset. The response variable is the average heights of a group of 161 children in Kalama, an Egyptian village: the site of a study of nutrition in developing countries. The data were obtained by measuring the heights of all 161 children in the village each month over several years. Time is the explanatory variable.
- **Number of observations**: 12
- **Variable Description**
- *age* Age in months
- *height* Average height in centimetres for children at this age

Regressione semplice: Esempio 1

Questi dati sono molto correlati e dal grafico a dispersione appare chiaro che la relazione lineare tra di essi è molto forte



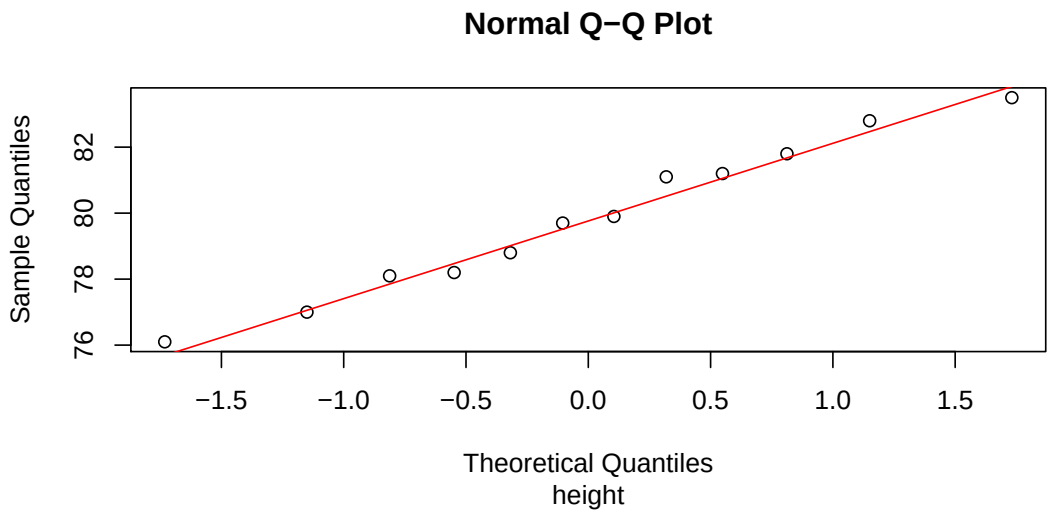
Regressione semplice: Esempio 1

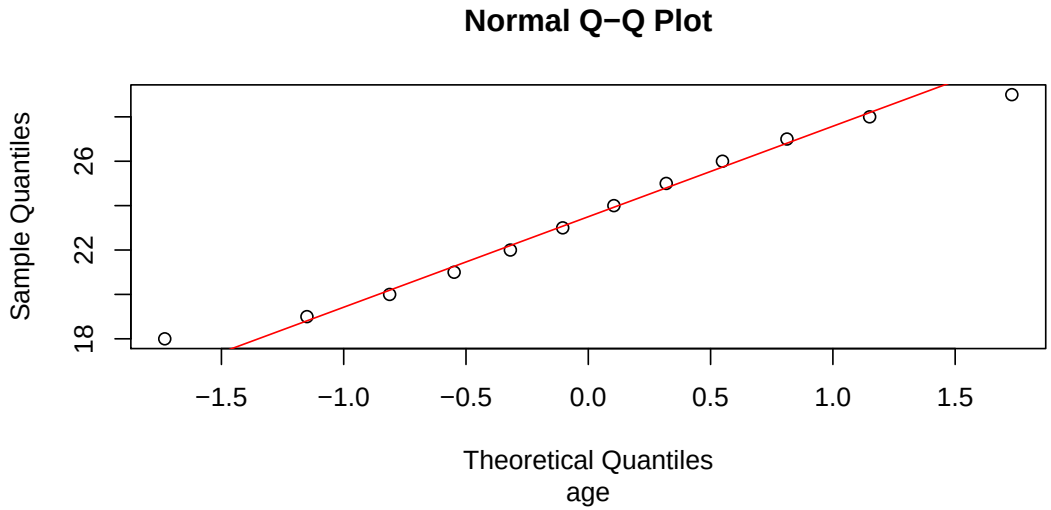
Statistiche di base

	age	height
Min.	18.00	76.10
1st Qu.	20.75	78.17
Median	23.50	79.80
Mean	23.50	79.85
3rd Qu.	26.25	81.35
Max.	29.00	83.50

Regressione semplice: Esempio 1

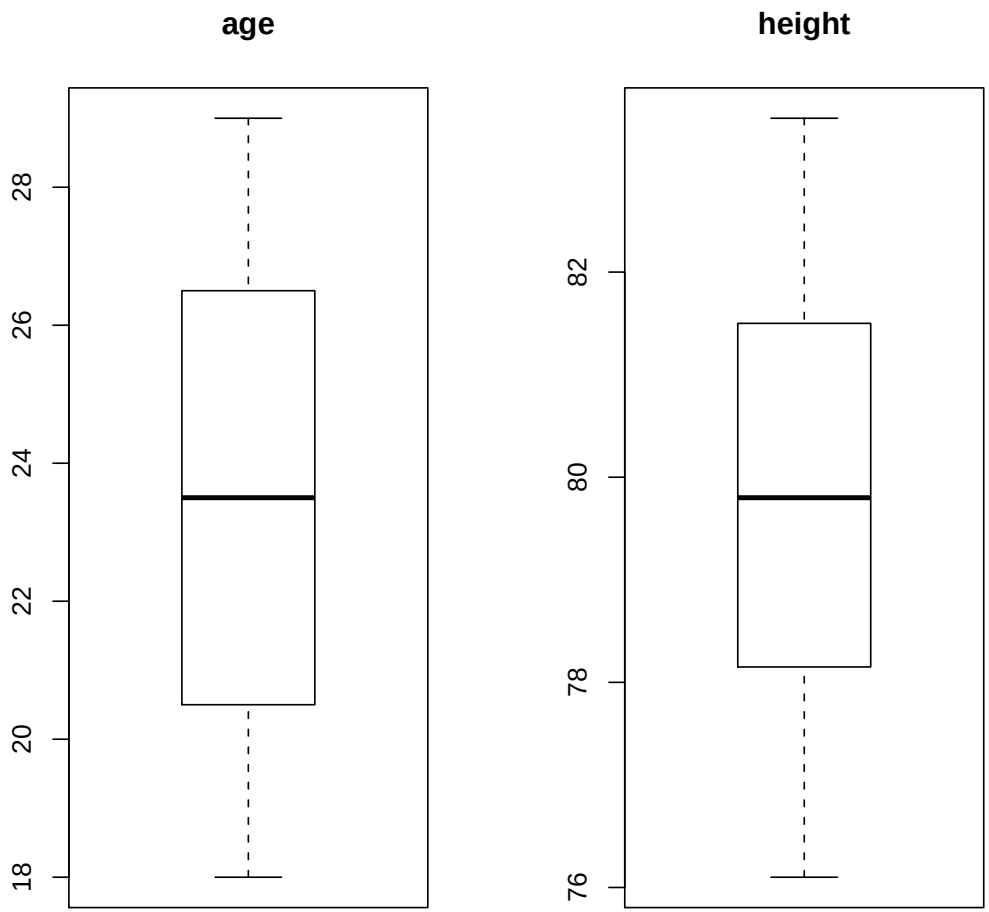
Verifichiamo le ipotesi di normalità dei dati:





Regressione semplice: Esempio 1

Verifichiamo l'eventuale presenza di dati anomali



Regressione semplice: Esempio 1

Cosa vediamo?

- Esiste una forte relazione lineare.
- Dalle statistiche descrittive desumiamo una certa regolarità delle osservazioni. Media e mediana coincidenti sono una prima verifica dell'ipotesi di normalità.
- Dal QQ-plot abbiamo un?ulteriore conferma in tal senso.
- L'assenza di valori anomali ci rassicura ulteriormente sulla possibilità di ottenere una buona stima.

Regressione semplice: Esempio 1

L'implementazione in R è molto semplice e si basa sulla funzione `lm`, il risultato in forma tabellare si ottiene tramite la funzione `summary` ovvero `yy=lm(height~age, data=ageheight)` seguito da `summary(yy)`:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	64.9283	0.5084	127.71	$< 2e-16$ ***
age	0.6350	0.0214	29.66	$4.43e-11$ ***

Nella tabella oltre alle stime dei coefficienti del modello, sono riportati i relativi errori standard, utili per costruire intervalli di confidenza. Inoltre su ogni coefficiente viene svolto un test d'ipotesi in cui l'ipotesi nulla è H_0 : *il coefficiente è uguale a zero* e l'ipotesi alternativa H_1: *il coefficiente è diverso da zero*, vengono quindi aggiunti i relativi p-value.

La stima è ottenuta con il metodo dei minimi quadrati

Regressione semplice: Minimi quadrati

Il metodo dei minimi quadrati produce una retta *dei minimi quadrati* di equazione

$$y = a + bx$$

dove $a = \bar{y} - b\bar{x}$ e $b = r\frac{s_y}{s_x}$
r coefficiente di correlazione ed s_y, s_x deviazioni standard delle due variabili.
a e b sono gli stimatori non distorti di α e β Questo significa che $E(a) = \alpha$ e $E(b) = \beta$.

Regressione semplice: Minimi quadrati

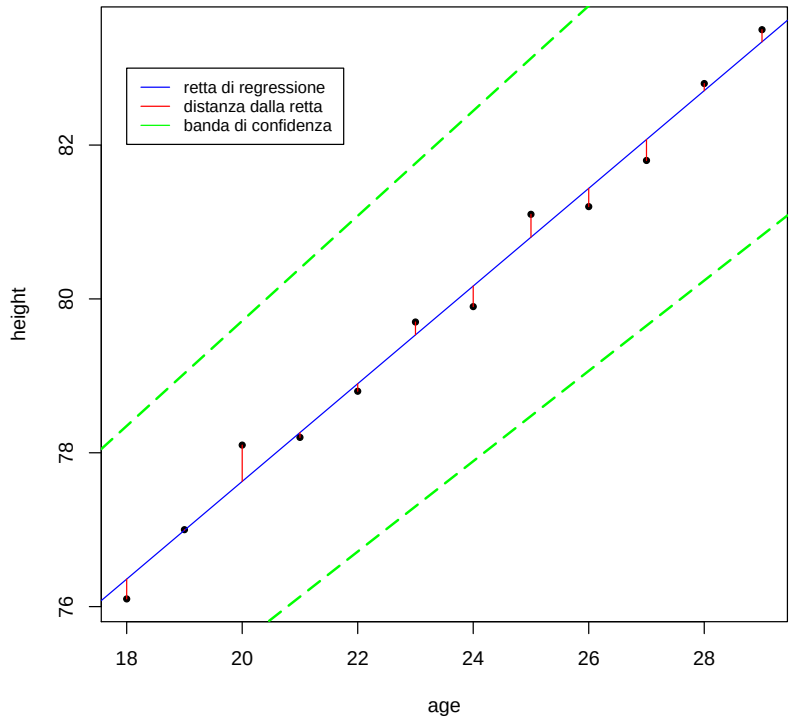
Minimi Quadrati

Con questa tecnica cerco quella retta che passando tra i dati si avvicina di più ad essi.

Formalmente:

$$\min_{a,b} \sum_i (y_i - a + bx_i)^2$$

Regressione semplice: Minimi quadrati



Retta di regressione, evidenziata la distanza tra retta e punti osservati. Inoltre sono state aggiunte due rette che delimitano una fascia di *rette possibili*. Queste due rette hanno come coefficienti i valori ottenuti dai limiti inferiore e superiore degli intervalli di confidenza costruiti per le stime ovvero $a \in [63.796, 66.061]$, $b \in [0.587, 0.683]$

Regressione semplice: Esempio 1

Nel nostro esempio:

- Il modello stimato è molto aderente ai dati
- La relazione tra le due variabili è tale per cui al crescere di un'unità della variabile età si ha una crescita di 0.64 unità dell'altezza (cfr valore del coefficiente angolare)
- Mentre quando l'età è pari a zero (neonato) l'altezza vale 64.93 cm che rappresenta l'altezza media alla nascita.

Come posso supportare rigorosamente la prima affermazione?

Regressione semplice: Valutazione del modello

- Un modo per valutare la *bontà* del modello con un singolo numero si basa sul **coefficiente di determinazione o percentuale della varianza spiegata** (in R: *multiple R squared*)
- In sostanza conduciamo un'analisi della varianza sui risultati del modello. Più precisamente scomponiamo la devianza totale nella parte spiegata dal modello ed una parte residua, poniamo $f_i = a + bx_i$, $\bar{f} = \frac{1}{n}\sum_i f_i$:

$$\begin{array}{ccccc} \sum_i \sum_i (y_i - \bar{y})^2 & = & \sum_i (f_i - \bar{f})^2 & + & \sum_i (Y_i - f_i)^2 \\ SS(y) & = & SS(reg) & + & SS(e) \end{array}$$

- il coefficiente di determinazione è definito come $R^2 = \frac{SS(reg)}{SS(y)} = 1 - \frac{SS(e)}{SS(y)}$ tanto più è vicino ad 1 tanto migliore è l'adattamento del modello ai dati.

Regressione semplice: Valutazione del modello

Nell'output di R, oltre alle informazioni già viste abbiamo altre indicazioni:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
age           0.6350     0.0214   29.66 4.43e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.256 on 10 degrees of freedom
Multiple R-squared: 0.9888, Adjusted R-squared: 0.9876
F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

La F statistics che viene indicata deriva dall'analisi della varianza condotta sul modello.

Regressione semplice: Valutazione del modello

Un altro modo per valutare la bontà del modello si basa sul condurre un'analisi della varianza usando la scomposizione della devianza appena vista. H_0: *il modello non spiega i dati* (ovvero $SS(reg)/gdl < SS(e)/gdl$)
H_1: *il modello spiega i dati* (ovvero $SS(reg)/gdl >> SS(e)/gdl$)
I gradi di libertà sono: $gdl_{tot} = n - 1$ se p è il numero di variabili indipendenti, $gdl_{reg} = p$ (uno per ogni variabile indipendente) e $gdl_e = n - p - 1$.
La statistica test è:

$$F = \frac{SS(reg)/p}{SS(e)/(n-p-1)} \sim F_{p,(n-p-1)}$$

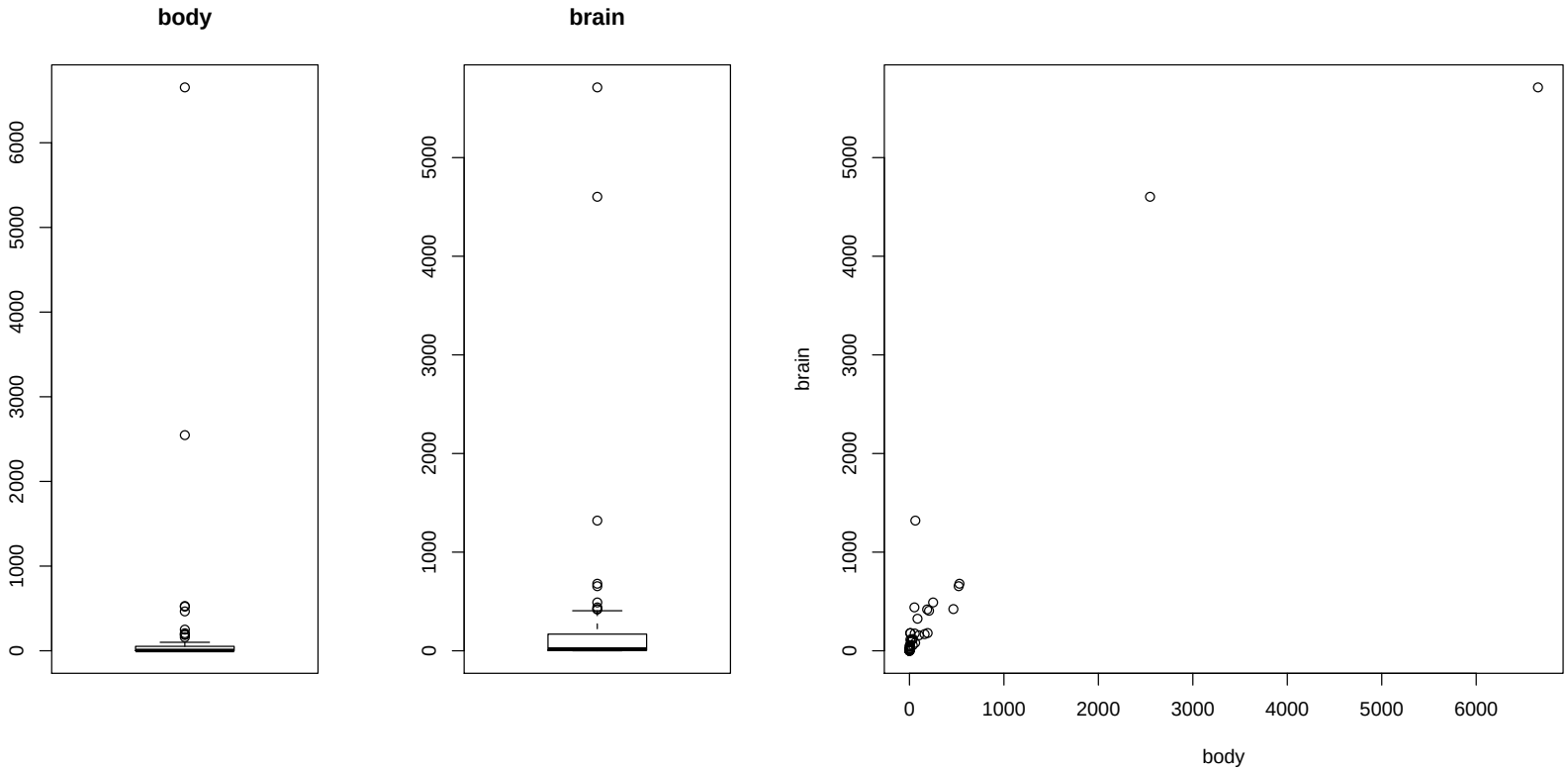
Regressione semplice: Esempio II

Abbiamo raccolto dati su 62 specie animali misurando il peso corporeo e il peso del cervello. Le specie considerate vanno dall'elefante africano al pipistrello bruno all'essere umano. Ci possiamo quindi aspettare una elevata variabilità nei dati.
Dal summary vediamo:

Animal frequency	Statistics	body	brain
africal giant pouched rat: 1	Min.	0.005	0.14
african elephant : 1	1st Qu.	0.600	4.25
arctic fox : 1	Median	3.342	17.25
arctic ground squirrel : 1	Mean	198.791	283.13
asian elephant : 1	3rd Qu.	48.203	166.00
baboon : 1	Max.	6654.000	5712.00
(Other) :56			

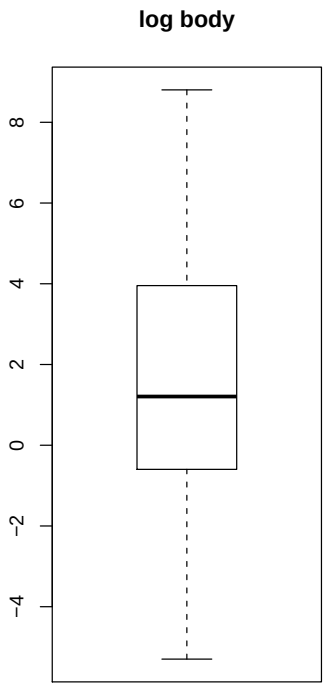
Regressione semplice: Esempio II

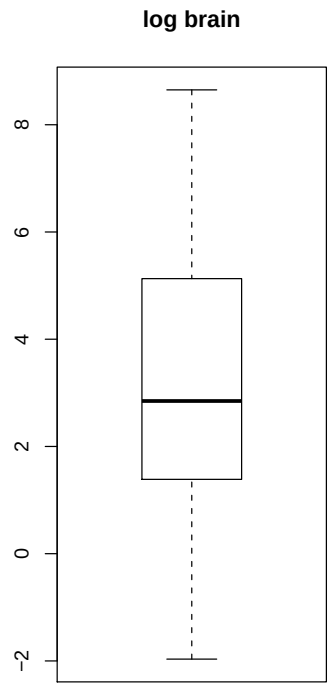
Dai boxplot abbiamo conferma della enorme variabilità dei dati a nostra disposizione e dal diagramma a dispersione vediamo come questa renda poco comprensibile l'eventuale presenza di una relazione lineare tra le due variabili peso del corpo e del cervello.

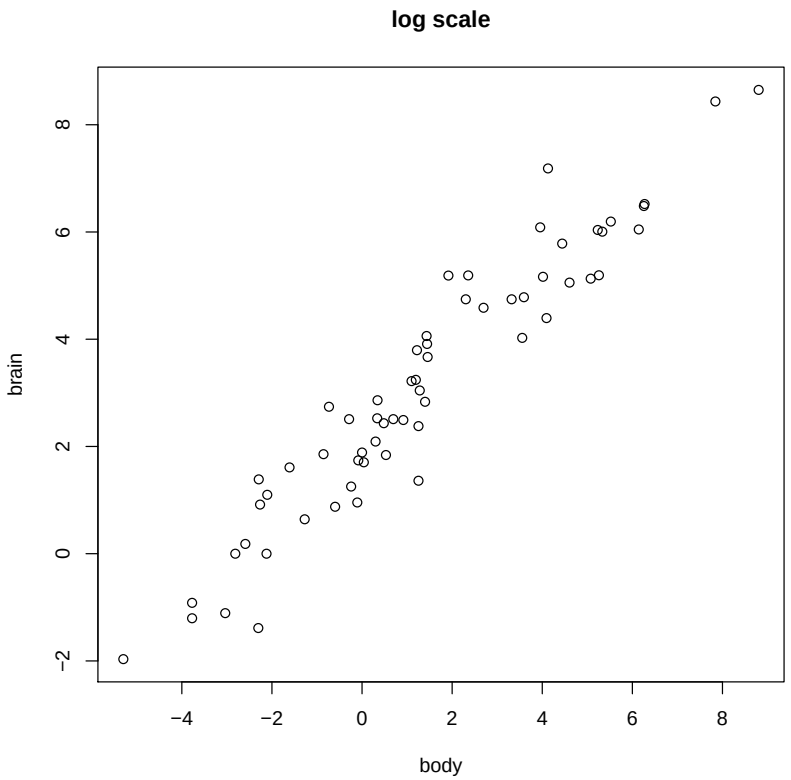


Regressione semplice: Esempio II

Passiamo alla scala log per entrambe le variabili e vediamo come cambiano i grafici

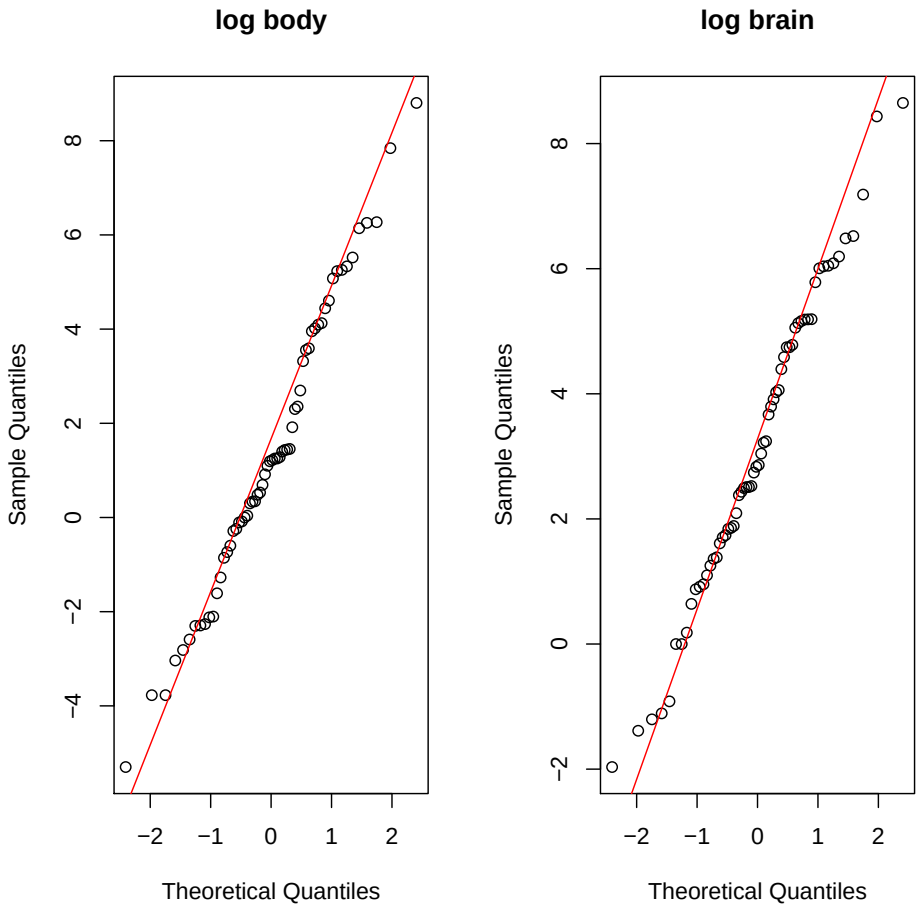






Regressione semplice: Esempio II

Su scala log log la relazione lineare è molto evidente, verifichiamo anche l'aderenza all'ipotesi di normalità tramite i qq-plot



Regressione semplice: Esempio II

Possiamo quindi procedere tranquillamente a stimare il modello di regressione, ottenendo:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	2.09180	0.10179	20.55	$< 2e-16 * **$
body	0.76256	0.03051	24.99	$< 2e-16 * **$

Residual standard error: 0.7303 on 60 degrees of freedom
Multiple R-squared: 0.9124, Adjusted R-squared: 0.9109
F-statistic: 624.6 on 1 and 60 DF, p-value:$< 2.2e-16$

Regressione semplice: Valutazione del modello

I Residui

- Dall'analisi dei residui di un modello possiamo capire molto di come quest'ultimo rappresenta i dati.
- Per ipotesi i residui devono avere media nulla, quindi riportandoli su di un grafico a dispersione insieme ai valori predetti dal modello mi aspetto di ottenere una *nuvola di punti* disposta attorno allo zero
- Se ciò non accade il modello non sta *spiegando* correttamente (o completamente) l'informazione contenuta nei dati.

Regressione semplice: Valutazione del modello

- Posso individuare osservazioni *influenti* ovvero dati anomali (outliers) che hanno un effetto di *trascinamento* sul modello stimato. Spesso di usano anche i residui standardizzati

$$\begin{aligned} \tilde{e}_i &= \frac{e_i - m_e}{s_e} \\ m_e &= \text{media residui} \\ s_e &= \text{sd. residui} \end{aligned}$$

- Si utilizzano anche diverse misure di influenza come *la distanza di Cook* e i *leverage*

Regressione semplice: Valutazione del modello

Distanza di Cook

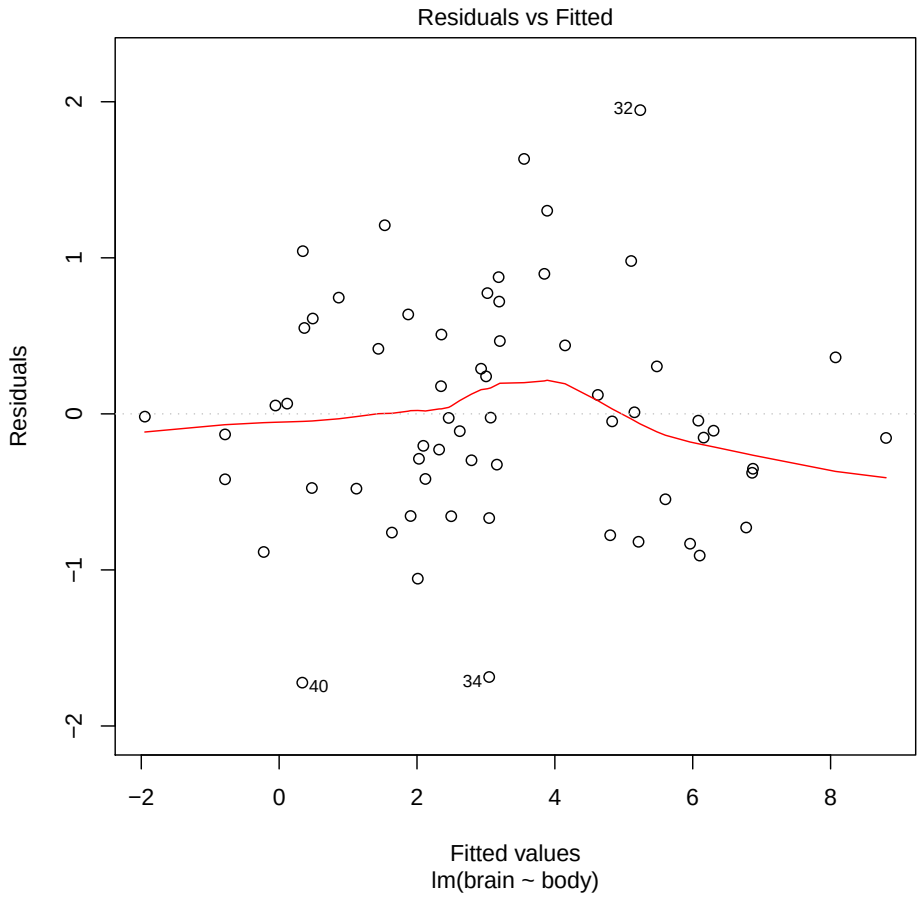
$$C_i = \frac{\sum_{j=1}^{n} (\hat{y}_j^{(i)} - \hat{y}_i)^2}{(p+1)s^2}$$

Dove:

- $\hat{y}_j^{(i)}$ valore previsto dal modello quando l'i-esima osservazione è esclusa dalla stima
- $\hat{y}_j$ i-esimo valore predetto dalla regressione;
- n numero di osservazioni
- p numero di variabili indipendenti
- s^2 varianza stimata dei residui

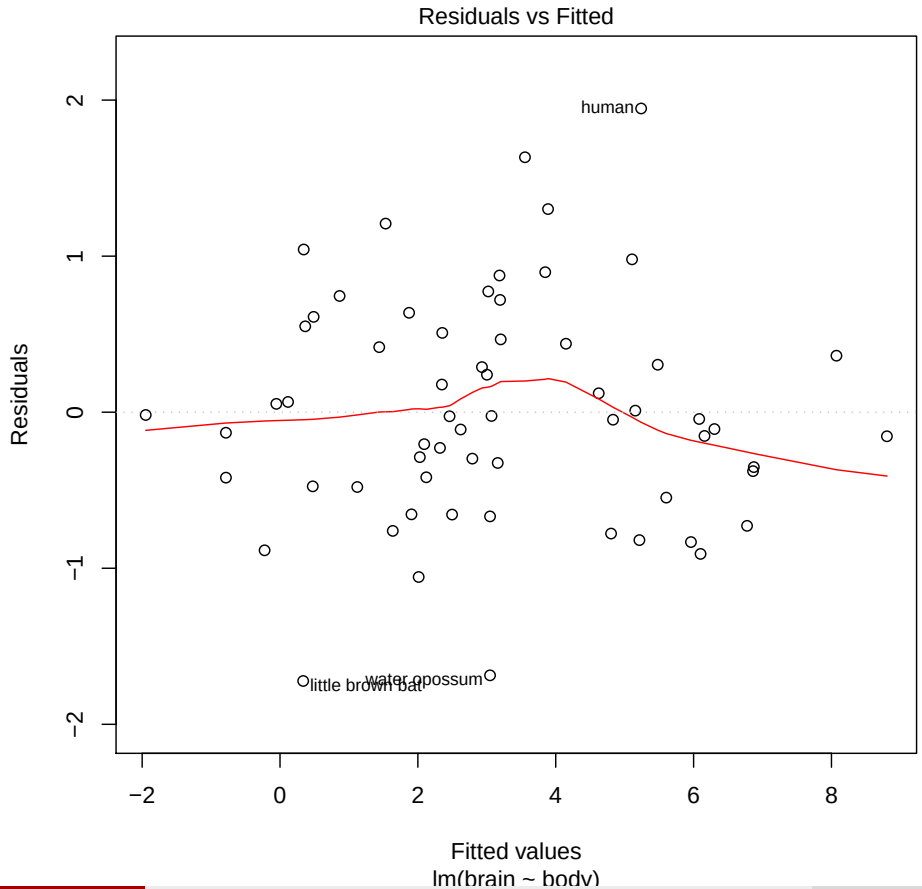
Regressione semplice: Residui Esempio II

Sull'asse verticale riportiamo $e_i = \hat{y}_i - y_i$ dove $\hat{y}_i$ sono i valori stimati tramite il modello e che a loro volta sono riportati sull'asse orizzontale

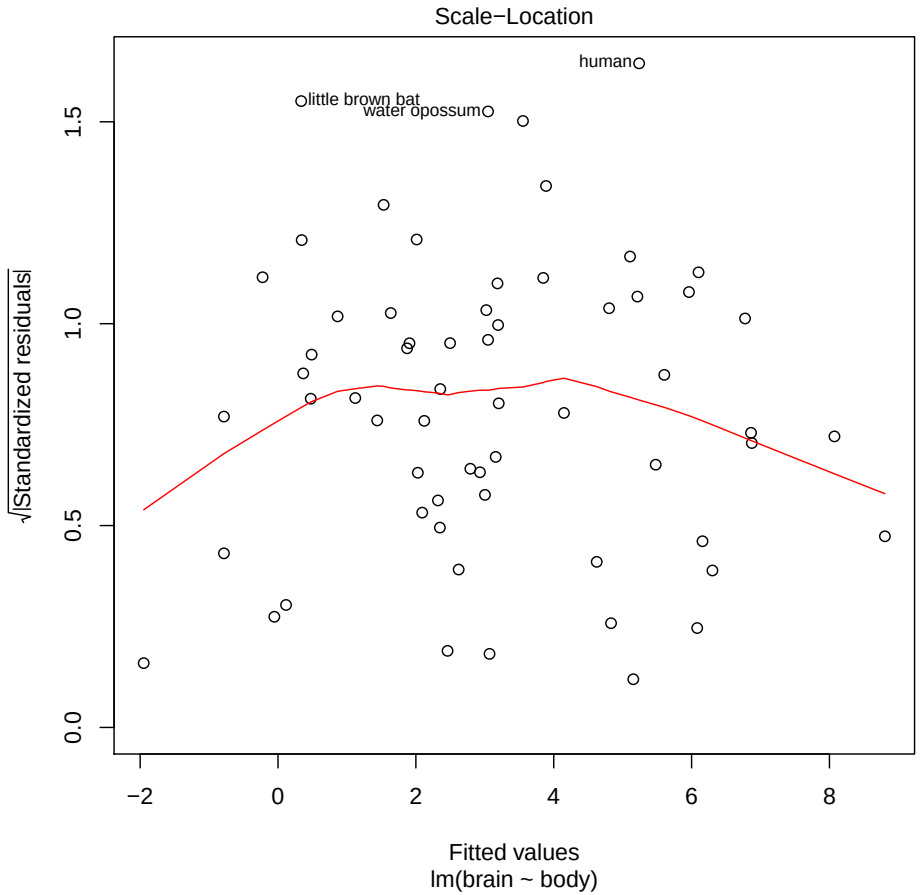


Regressione semplice: Residui Esempio II

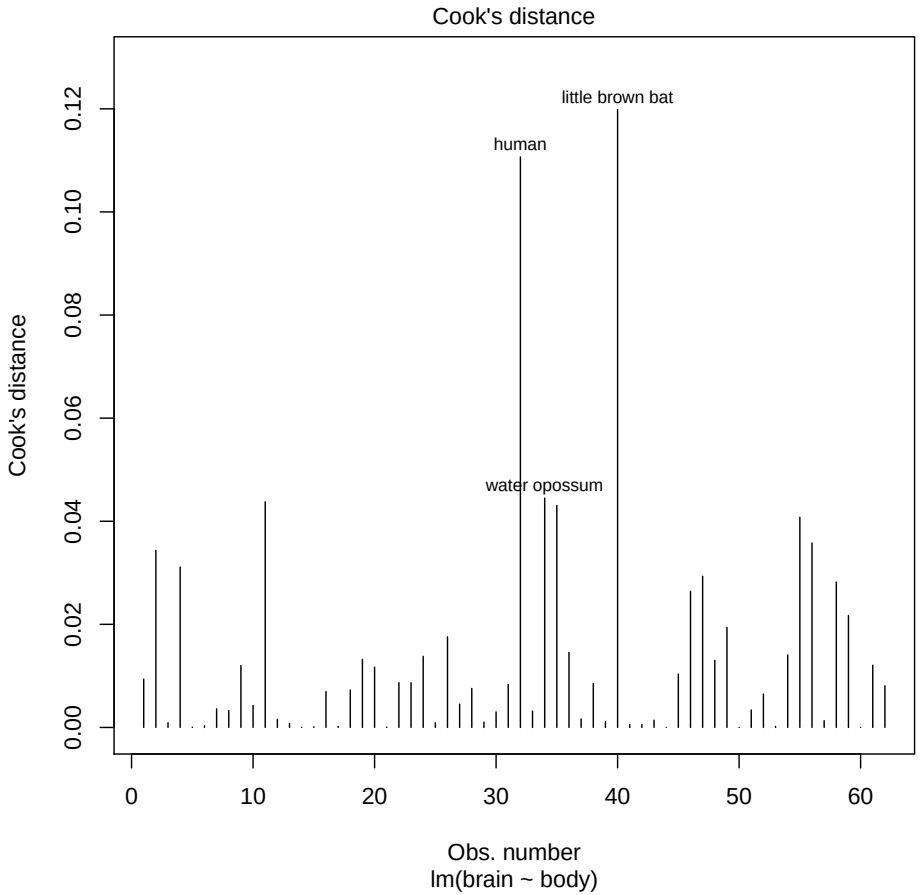
Alcune osservazioni sono evidenziate con il numero della riga corrispondente e sono quelle con residui più elevati in valore assoluto. Questo indica che quelle osservazioni sono mal descritte dal modello



Regressione semplice: Residui standardizzati Esempio II



Regressione semplice: Distanza di Cook Esempio II



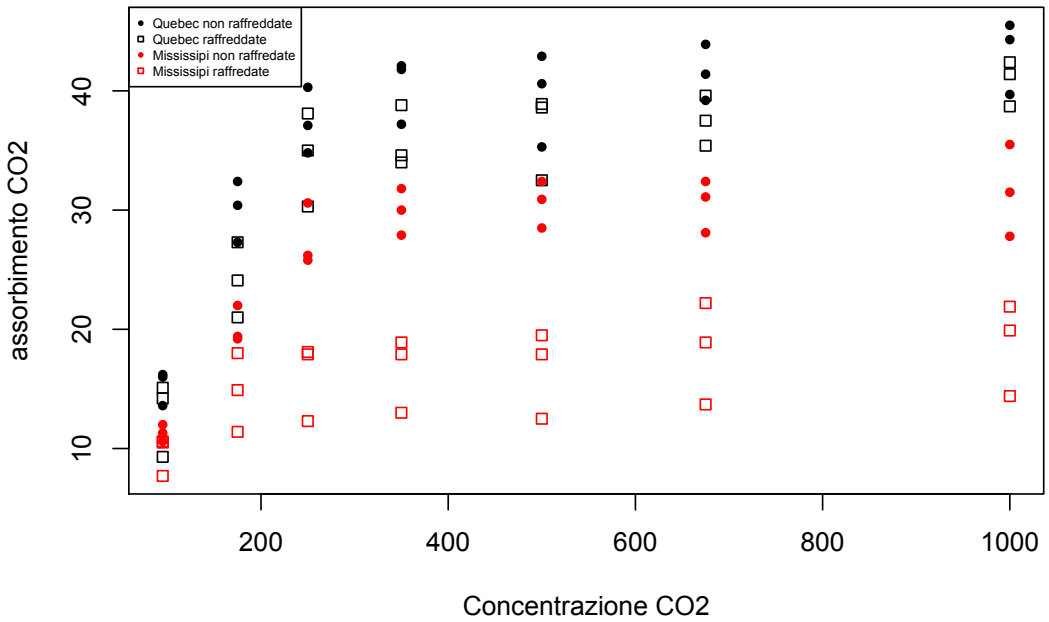
Regressione semplice: Esempio II

In conclusione

- Il valore di R^2 è molto alto indicando un ottimo adattamento del modello su scala log-log
- Dall'analisi dei grafici dei residui vediamo che ci sono sempre 3 osservazioni che risultano relativamente mal rappresentate dal modello
- Queste sono la 32,34 e 40 che corrispondono a due animali piccoli (opossum d'acqua e pipistrello marrone 34 e 40) e all'uomo che ha il rapporto peso del cervello-peso corporeo più sbilanciato.

Regressione semplice con variabili qualitative

Consideriamo la seguente situazione: In un esperimento si misura l'assorbimento (uptake) di CO2 di alcune piante, stessa specie, provenienti da due aree diverse, Quebec e Mississip (Type). Le piante sono esposte a concentrazioni diverse di CO2 (conc) e una parte di esse viene raffreddata prima di esser esposta alla CO2 (Treatment). Vogliamo studiare la relazione tra assorbimento e concentrazione.



Regressione semplice con variabili qualitative

- Dal grafico sembra probabile che sia necessario includere nella stima di un modello di regressione la distinzione relativa al tipo di pianta e forse anche al trattamento.
- Limitiamoci al solo fattore Type.
- Possiamo procedere stimando due modelli di regressione separati per ogni gruppo oppure stimare un solo modello inserendo delle variabili 0/1 che descrivono quando un'osservazione è di un tipo o dell'altro.

Analizziamo l'esempio in pratica.

Regressione semplice con variabili qualitative