Statistica Computazionale B

Carlo Gaetan [1]
Dipartimento di Statistica
Università Ca' Foscari - Venezia
gaetan@unive.it

Anno Accademico 2008/2009

[1]Lucidi per il corso, preparati con `R version 2.9.2 (2009-08-24)`. Ringrazio Stefano Tonellato per aver permesso di utilizzare molta parte del suo materiale didattico.

Indice

A. Cercar casa, 3

B. Consumo di benzina, 55

C. Diete, 95

D. Cattedrali inglesi, 107

E. Risparmi, 117

F. Analisi dei residui, 121

Unità A

Cercar casa

- Diagramma di dispersione
- Modello di regressione lineare semplice
- Minimi quadrati
- Proprietà

I dati

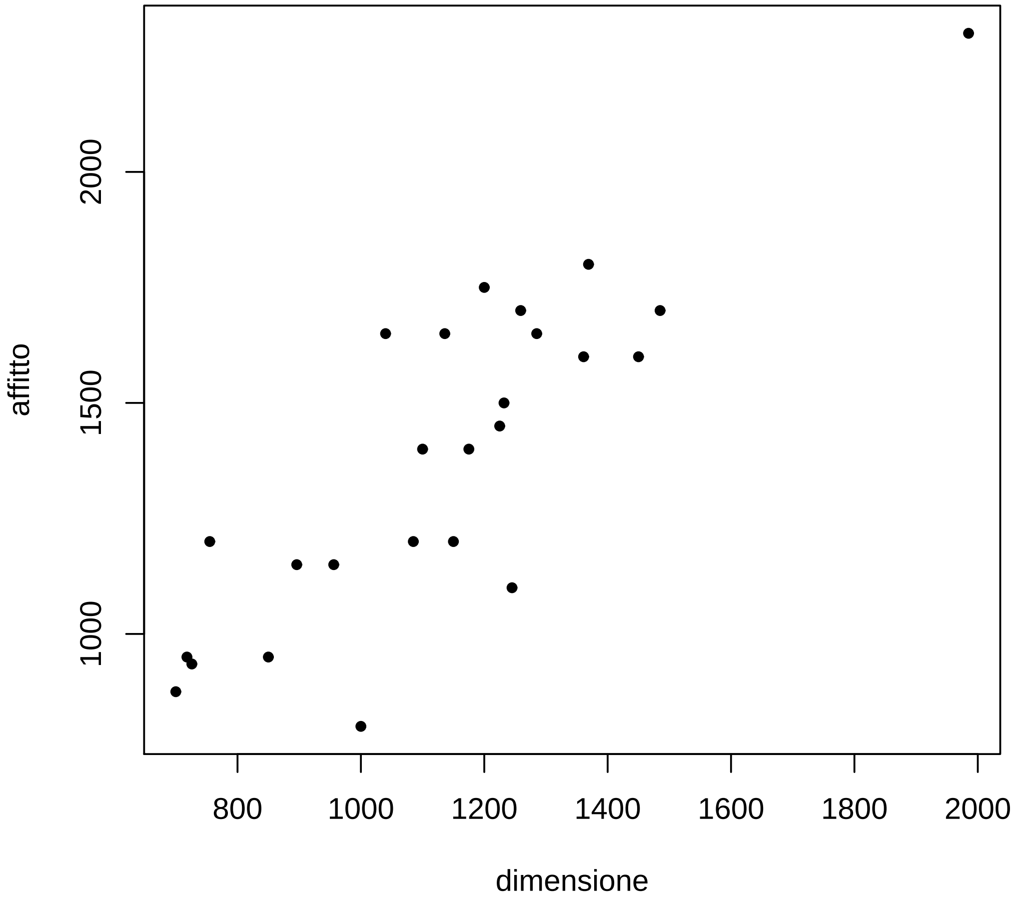
Un agente immobiliare [1] intende prevedere gli affitti mensili degli appartamenti sulla base della loro dimensione. Per questo conduce un'indagine e reperisce i dati su 25 appartamenti in una zona residenziale. La seguente tabella mostra i dati ottenuti per i 25 appartamenti, l'affitto è l'affitto mensile in dollari e la dimensione è espressa in piedi al quadrato.

	affitto	dimensione
1	950	850
2	1600	1450
3	1200	1085
4	1500	1232
5	950	718
6	1700	1485
7	1650	1136
8	935	726
9	875	700
10	1150	956
11	1400	1100
12	1650	1285
13	2300	1985
14	1800	1369
15	1400	1175
16	1450	1225
17	1100	1245
18	1700	1259
19	1200	1150
20	1150	896
21	1600	1361
22	1650	1040
23	1200	755
24	800	1000
25	1750	1200

Si vogliono utilizzare i dati per ottenere una equazione che permetta di prevedere l'affitto in base alla dimensione. Una simile equazione risulterà molto utile per fissare il prezzo d'affitto di un appartamento.

[1] Dati tratti da Levine, Krehbiel, Berenson (2002) Statistica, Apogeo.

Diagramma di dispersione



Abbiamo semplicemente disegnato i punti osservati sul piano. E' evidente una forte relazione, certamente crescente come ci si poteva attendere.

La covarianza campionaria come misura della direzione e della forza della relazione tra due variabili

Date n unità statistiche, supponiamo di osservare n coppie di valori (x_i, y_i), $i = 1, \ldots, n$ di due variabili numeriche x e y. La covarianza campionaria tra x e y è definita come

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \tag{A.1}$$

dove $\overline{x}$ e $\overline{y}$ indicano le due medie aritmetiche.

Si osservi che

1. nei caso in cui a valori crescenti di x corrispondano valori crescenti di y, ci aspettiamo che valori maggiori della media di x corrispondano a valori maggiori della media per y; in questo caso quindi la covarianza risulterà positiva.

2. Completamente simmetrico è quello che accade nel caso in cui al crescere della x la y tendenzialmente descresce. Quindi, in questo caso, ci aspettiamo una covarianza negativa.

3. Più è forte la relazione tra le due variabili più ci aspettiamo che la covarianza diventi grande in valore assoluto. Infatti, più è forte la relazione più il numero di addendi concordi nella (A.1) dovrebbe crescere ed inoltre un certo numero di

addendi sarà il prodotto di scarti dalle media grandi in valore assoluto.

4. In assenza di una qualche forma di relazione *monotona* tra le due variabili, viceversa, gli addendi della (A.1) saranno in parte positivi ed in parte negativi. Quindi in questi casi ci aspettiamo che la covarianza risulti nulla o comunque vicina allo zero.

5. **Osservazione importante**. Abbiamo già incontrato una definizione di covarianza: quella tra due **variabili casuali** X e Y definita come

$$\mathbb{C}ov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

   Non si confonda questa con la precedente.

Le considerazioni precedenti suggerisono l'uso della covarianza campionaria per *misurare* la direzione e la forza delle relazioni esistenti tra le variabili (quantomeno monotone, in realtà come vedremo essenzialmente lineari).

Calcolo della covarianza campionaria

Per il calcolo della covarianza è conveniente utilizzare la seguente relazione

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}.$$

ovvero

$$(\text{covarianza}) = \left(\begin{array}{c}\text{media dei}\\\text{prodotti}\end{array}\right) - \left(\begin{array}{c}\text{prodotto delle}\\\text{medie}\end{array}\right).$$

Infatti, abbiamo

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

Il secondo addendo è nullo poichè la somma degli scarti dalla media vale zero. Espandendo il primo addendo troviamo

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

Il coefficiente di correlazione (lineare) campionario

Per affermare se la covarianza campionaria è "piccola" o è "grande" dobbiamo confrontarla con il prodotto degli scarti quadratici medi.

Per semplificare il lavoro, è usuale presentare i risultati utilizzando non direttamente la covarianza ma una sua versione normalizzata nota come **coefficiente di correlazione (lineare) campionario**[2] e definito come

$$\mathsf{r}(x,y) = \frac{\mathsf{cov}(x,y)}{\mathsf{sqm}(x)\mathsf{sqm}(y)}.$$

dove $\mathsf{sqm}(z)$ indica lo scarto quadratico medio di z

$$\mathsf{sqm}(z) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \overline{z})^2}$$

Si può dimostrare che il coefficiente varia tra -1 e 1.

Nel nostro caso abbiamo $\mathsf{r}(x,y) = 0.85$.

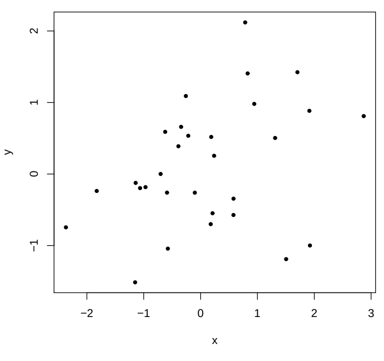
[2] Il "lineare" tra parantesi indica che a volte l'aggettivo lineare è omesso

La sua interpretazione è, a grande, linee la seguente.

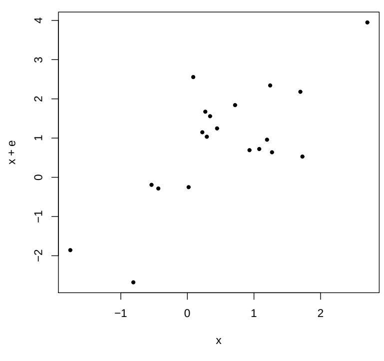
Se $\boxed{\mathsf{r}(x,y) < 0}$ allora i dati indicano una associazione negativa tra le due variabili (al crescere di una l'altra decresce). Questa associazione è man mano più forte più $\mathsf{r}(x,y)$ si avvicina a -1. Se $\boxed{\mathsf{r}(x,y) = -1}$ allora i dati sono perfettamente allineati su di una retta con coefficiente angolare negativo.

Se $\boxed{\mathsf{r}(x,y) = 0}$, ed in realtà da un punto di vista pratico, se $\mathsf{r}(x,y) \approx 0$, allora non esiste una relazione lineare (e più in generale una associazione monotona) tra le due variabili.

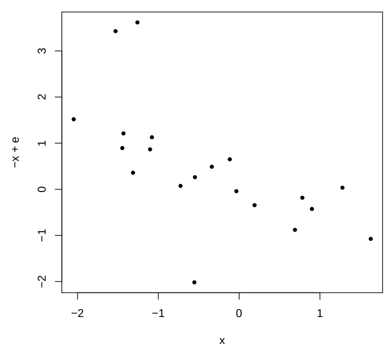
Se $\boxed{\mathsf{r}(x,y) > 0}$ l'interpretazione è simmetrica a quando detto per il caso "coefficiente di correlazione negativo". La relazione è crescente e se $\boxed{\mathsf{r}(x,y) = 1}$ i dati sono perfettamente allineati su di una retta con coefficiente angolare positivo.



0.35



0.75



-0.64

Due limiti di r(x, y) da tenere presente

- Dati posti perfettamente su di una curva *monotona*, pensiamola crescente, ma *non lineare* indicano una dipendenza perfetta tra le due variabili ma non risultano in r$(x, y) = 1$. Si pensi ad esempio, a dei dati posti sulla curva $y = \exp(x)$. In questo caso, i dati non risulteranno allineati. Quindi risulterà r$(x, y) < 1$. In definitiva, il coefficiente di correlazione campionario misura *accuratamente* la forza della relazione esistente solo se questa è lineare.

- r$(x, y) = 0$ non implica che non esiste nessuna relazione tra x e y. Ad esempio, lo studente verifichi che se le coppie di dati sulle due variabili sono $(x_1, y_1) = (-2, 4)$, $(x_2, y_2) = (-1, 1)$, $(x_3, y_3) = (0, 0)$, $(x_4, y_4) = (1, 1)$ e $(x_5, y_5) = (2, 4)$ allora r$(x, y) = 0$. Nonostante questo però i dati sono esattamente posti sulla parabola $y = x^2$. Il fatto è che il coefficiente di correlazione è, per costruzione, inutile nel valutare l'esistenza e la forza di relazioni non monotone.

Un possibile modello per il prezzo

Come si determina il prezzo? Il prezzo si determina per un coacervo di fattori alcuni noti altri no. Potremmo pensare che un fattore determinante sia dato dalla dimensione.

Adottiamo per il momento l'ipotesi di una relazione lineare.

Possiamo allora pensare ad un modello del tipo

$$(\text{affitto}) = \alpha + \beta(\text{dimensione}) + (\text{errore}) \tag{A.2}$$

dove l'ultima componente esprime la parte delle oscillazioni dell'affitto mensile non legate alla dimensione o, forse più precisamente, che una funzione lineare della dimensione non riesce a spiegare.

Modelli di regressione lineare semplice: caso generale e terminologia

Come possiamo intepretare la variabilità dei nostri dati a disposizione ?

$$\begin{array}{cccc} (y_1, x_1) & (y_2, x_2) & \dots & (y_{25}, x_{25}) \\ (950, 850) & (1600, 1450) & \dots & (1750, 1200) \end{array}$$

Ipotizziamo che i valori y_i, $i = 1, \dots, 25$ siano altrettante osservazioni da v.c. Y_i tra loro **incorrelate** ma proprio perché pensiamo che la diversa dimensione dell'appartamento influisca sul prezzo dell'affitto le v.c. **non sono identicamente distribuite**. Per esempio potremmo ipotizzare che Y_i sia una v.c. con valore atteso che dipende dalla dimensione dell'appartamento (che denotiamo x_i).

$$\mathbb{E}(Y_i) = \alpha + \beta x_i$$

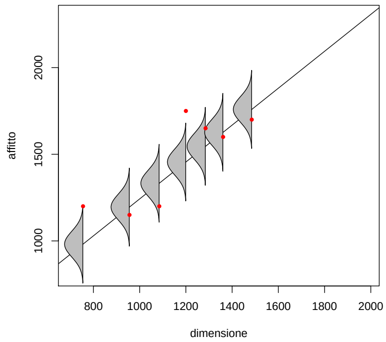
In questo modello ipotizziamo che il prezzo 'medio' dell'affitto sia determinato da dei fattori fissi (α) e dei fattori variabili (βx_i).

Una scrittura equivalente del modello che spiega la variabilità del prezzo d'affitto è la seguente

$$Y_i = \alpha + \beta x_i + \varepsilon_i. \tag{A.3}$$

dove ε_i è una v.c. detta **errore** tale che $\mathbb{E}(\varepsilon_i) = 0$.

Circa la variabilità dell'errore supponiamo che $\mathbb{V}ar(\varepsilon_i) = \sigma^2$ ovvero che $\mathbb{V}ar(Y_i) = \sigma^2$.



Un modello del tipo (A.3) viene usualmente chiamato **modello di regressione lineare semplice**. Nel caso generale, cerchiamo di **spiegare** una variabile, diciamo Y, utilizzando un'altra variabile, diciamo x.

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

Assunzioni del modello lineare

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1, \dots, n.$$

- Y viene usualmente indicata come la **variabile risposta** o la **variabile dipendente** e Y_i è la **variabile casuale** relativa alla $i-esima$ unità che fa parte del campione.

- x è il **regressore**, detto anche **variabile esplicativa** o **variabile indipendente**, e x_i è il valore **predeterminato** assunto dalla $i-esima$ unità che fa parte del campione. Si noti il diverso ruolo di x e Y.

- ε rappresenta il termine d'**errore**. È una **variabile casuale** con $\mathbb{E}(\varepsilon_i) = 0$ e $\mathbb{V}ar(\varepsilon_i) = \sigma^2$.

- le variabili Y_i sono supposte **incorrelate**

- α, β e σ^2 sono i **parametri** del modello e sono **ignoti**.

Minimi quadrati: idea

Il problema è come stimare α e β. Infatti, se riusciamo a calcolare un valore "ragionevole" per questi due parametri, diciamo $\widehat{\alpha}$ e $\widehat{\beta}$, possiamo poi pensare di "prevedere" il prezzo d'affitto

$$\widehat{\alpha} + \widehat{\beta}(\text{dimensione}). \tag{A.4}$$

Sembra ragionevole cercare di "calcolare" $\widehat{\alpha}$ e $\widehat{\beta}$ in modo tale che la (A.4) fornisca buone "previsioni" sull'insieme di dati osservato. Al proposito, indichiamo con n il numero delle osservazioni (in questo caso $n = 25$), e poniamo $y_i = (\text{affitto della casa } i\text{-sima})$ e $x_i = (\text{dimensione della casa } i\text{-sima})$. Quello che vorremmo è trovare dei valori per i parametri tali che

$$\begin{aligned} y_1 &\approx \widehat{\alpha} + \widehat{\beta}x_1 \\ y_2 &\approx \widehat{\alpha} + \widehat{\beta}x_2 \\ &\vdots \\ y_n &\approx \widehat{\alpha} + \widehat{\beta}x_n \end{aligned} \tag{A.5}$$

Per rendere "operativa" la (A.5), dobbiamo decidere (i) in che senso interpretiamo gli $\approx$ che abbiamo scritto e (ii) come combiniamo tra di loro le varie linee della (A.5) stessa. La soluzione più usata si concretizza nello scegliere i due parametri minimizzando

$$s^2(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{A.6}$$

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

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

per qualsivoglia $\alpha \in R$ e $\beta \in R$. In questo caso si dice che "i parametri sono stati calcolati utilizzando il **metodo dei minimi quadrati**".

Minimi quadrati: stima dei parametri α e β

Osserviamo, in primo luogo, che per ogni prefissato β, conosciamo già la soluzione del seguente problema

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

Infatti sappiamo che assegnati n numeri, diciamo $z_1, \ldots, z_n$, la media aritmetica delle z_i minimizza in a $\sum (z_i - a)^2$. Nel problema di minimizzazione precedente α gioca il ruolo di a e $(y_i - \beta x_i)$ quello di z_i. Quindi, per qualsivoglia β, la soluzione del problema la troviamo in corrispondenza di

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

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

Dalla definizione di $\alpha(\beta)$ segue che, per qualsivoglia α e β,

$$s^2(\alpha, \beta) \geq s^2(\alpha(\beta), \beta).$$

Quindi, $\widehat{\beta}$ può essere cercato risolvendo il problema di ottimizzazione

$$\inf_{\beta \in R} s^2(\alpha(\beta), \beta)$$

mentre

$$\widehat{\alpha} = \alpha(\widehat{\beta})$$

Ora,

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

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

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

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

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

Lasciamo allo studente il compito di vericare che questa soluzione corrisponde ad un punto di minimo (e non, ad esempio, ad un massimo).

La soluzione trovata può quindi essere scritta in modo compatto dividendo numeratore e denominatore per n come

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \tag{A.7}$$

$$\widehat{\beta} = \frac{\text{cov}(x, y)}{\text{var}(x)} \tag{A.8}$$

Le (A.7-A.8) forniscono la soluzione del problema che ci si era proposti solamente se var$(x) > 0$. Questo è molto ragionevole: β ci dice come varia la risposta al variare della esplicativa, ma se var$(x) = 0$ l'esplicativa non è variata affatto nei dati disponibili.

Quindi, per calcolare le stime di α e β è sufficiente conoscere le seguenti quantità:

$$\sum_{i=1}^{n} y_i \qquad \sum_{i=1}^{n} x_i \qquad \sum_{i=1}^{n} x_i^2 \qquad \sum_{i=1}^{n} x_i y_i \ .$$

Stima dei parametri nel caso degli affitti

In questo caso,

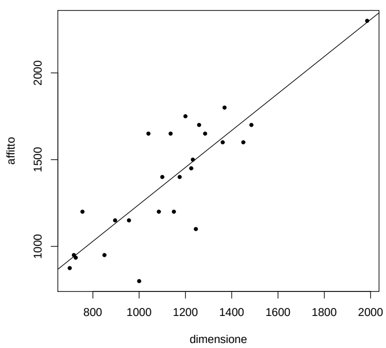
$$\begin{array}{ll} \sum y_i = 34660 & \sum x_i = 28383 \\ \sum x_i^2 = 34223535 & \sum x_i y_i = 41480210. \end{array}$$

Perciò

$$\begin{array}{l} \overline{y} = 34660/25 = 1386.4 \\ \overline{x} = 28283/25 = 1135.32 \\ \mathsf{var}(x) = (34223535/25) - 1135.32^2 = 79989.9 \\ \mathsf{cov}(x,y) = (41480210/25) - 1135.32 \times 1386.4 = 85200.75. \end{array}$$

Quindi

$$\begin{array}{l} \widehat{\beta} = 85200.75/79989.9 = 1.065144 \\ \widehat{\alpha} = 1386.4 - 1.065144 \times 1135.32 = 177.1207 \end{array}$$



Il grafico mostra i dati osservati con la retta di regressione stimata.

La distribuzione degli stimatori di α e β

Consideriamo lo stimatore di β

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{j=1}^{n}(x_j - \overline{x})^2} = \sum_{i=1}^{n} w_i Y_i \\
&\text{con } w_i = \frac{x_i - \overline{x}}{\sum_{j=1}^{n}(x_j - \overline{x})^2}
\end{aligned}$$

e calcoliamo il valore atteso

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}) &= \mathbb{E}\left(\sum_{i=1}^{n} w_i Y_i\right) \\
&= \sum_{i=1}^{n} w_i \mathbb{E}(Y_i) \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(\alpha + \beta x_i)}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \\
&= \frac{\alpha \sum_{i=1}^{n}(x_i - \overline{x}) + \beta \sum_{i=1}^{n}(x_i - \overline{x})x_i}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \\
&= \frac{\beta \sum_{i=1}^{n}(x_i - \overline{x})x_i}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \\
&= \beta
\end{aligned}$$

da cui deduciamo che lo stimatore è non distorto.

$$
\begin{aligned}
\mathbb{V}ar(\widehat{\beta}) &= \mathbb{V}ar\left(\sum_{i=1}^{n} w_i Y_i\right) \\
&= \sum_{i=1}^{n} w_i^2 \mathbb{V}ar(Y_i) \\
&= \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{\left(\sum_{j=1}^{n}(x_j-\overline{x})^2\right)^2}\sigma^2 \\
&= \frac{\sigma^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}
\end{aligned}
$$

Per $\widehat{\alpha}$ valgono considerazioni del tutto analoghe a quanto appena visto:

$$
\begin{aligned}
\widehat{\alpha} &= \overline{Y} - \widehat{\beta}\overline{x} \\
&= \frac{\sum_{i=1}^{n} Y_i}{n} - \overline{x}\frac{\sum_{i=1}^{n}(x_i-\overline{x})Y_i}{\sum_{j=1}^{n}(x_j-\overline{x})^2} \\
&= \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{x}\frac{(x_i-\overline{x})}{\sum_{j=1}^{n}(x_j-\overline{x})^2}\right)Y_i \\
&= \sum_{i=1}^{n} \omega_i^* Y_i \\
&\quad \text{con} \quad \omega_i^* = \frac{1}{n} - \overline{x}\frac{(x_i-\overline{x})}{\sum_{j=1}^{n}(x_j-\overline{x})^2}\,.
\end{aligned}
$$

e cioé anche $\widehat{\alpha}$ è una combinazione lineare delle variabili Y_i.

Il valore atteso è

$$
\begin{aligned}
\mathbb{E}(\widehat{\alpha}) &= \sum_{i=1}^{n} \omega_i^* \mathbb{E}(Y_i) = \sum_{i=1}^{n} \omega_i^* (\alpha + \beta x_i) \\
&= \alpha \sum_{i=1}^{n} \left(\frac{1}{n} - \overline{x} \frac{(x_i - \overline{x})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) + \beta \sum_{i=1}^{n} \left(\frac{x_i}{n} - \overline{x} \frac{(x_i - \overline{x}) x_i}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) \\
&= \alpha \left(\frac{n}{n} - \overline{x} \frac{\sum_{i=1}^{n} (x_i - \overline{x})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) + \beta \left(\frac{\sum_{i=1}^{n} x_i}{n} - \overline{x} \frac{\sum_{i=1}^{n} (x_i - \overline{x}) x_i}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) \\
&= \alpha
\end{aligned}
$$

Quindi $\widehat{\alpha}$ è uno stimatore non distorto di α.

La varianza di $\widehat{\alpha}$ si ottiene con alcuni semplici passaggi algebrici:

$$
\begin{aligned}
\mathbb{V}ar(\widehat{\alpha}) &= \mathbb{V}ar\left(\sum_{i=1} \omega_i^* Y_i \right) = \sum_{i=1}^{n} (\omega_i^*)^2 Var(Y_i) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - \overline{x} \frac{(x_i - \overline{x})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right)^2 \sigma^2 \\
&= \sum_{i=1}^{n} \left(\frac{1}{n^2} + \overline{x}^2 \frac{(x_i - \overline{x})^2}{\left(\sum_{j=1}^{n} (x_j - \overline{x})^2 \right)^2} - \frac{2}{n} \overline{x} \frac{(x_i - \overline{x})}{\sum_{i=j}^{n} (x_j - \overline{x})^2} \right) \sigma^2 \\
&= \left(\frac{n}{n^2} + \overline{x}^2 \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\left(\sum_{j=1}^{n} (x_j - \overline{x})^2 \right)^2} - \frac{2}{n} \overline{x} \frac{\sum_{i=1}^{n} (x_i - \overline{x})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) \sigma^2 \\
&= \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=j}^{n} (x_j - \overline{x})^2} \right) \sigma^2
\end{aligned}
$$

Similmente si ottiene che la covarianza degli stimatori è

$$
\mathbb{C}ov(\widehat{\alpha}, \widehat{\beta}) = - \frac{\overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sigma^2
$$

Il teorema di Gauss-Markov

Gli stimatori $\widehat{\alpha}$ e $\widehat{\beta}$ ottenuti con il metodo dei minimi quadrati sono gli stimatori a varianza minima tra gli stimatori lineari corretti di α e β.

Dimostriamolo per $\widehat{\beta}$. Uno stimatore lineare è del tipo $b = \sum_{i=1}^n c_i Y_i$. Vogliamo determinare c_i $i = 1, \ldots, n$ in modo tale che sia non distorto $\mathbb{E}(b) = \beta$ e che la sua varianza, $\mathbb{V}ar(b)$ sia la più piccola possibile.

Si ha

$$b = \sum_{i=1}^n c_i Y_i = \alpha \sum_{i=1}^n c_i + \beta \sum_{i=1}^n c_i x_i + \sum_{i=1}^n c_i \varepsilon_i$$

e quindi

$$\mathbb{E}(b) = \alpha \sum_{i=1}^n c_i + \beta \sum_{i=1}^n c_i x_i.$$

Perché sia non distorto $\sum_{i=1}^n c_i = 0$ e $\sum_{i=1}^n c_i x_i = 1$.

Minimizziamo $\mathbb{V}ar(b) = \sigma^2 \sum_{i=1}^n c_i^2$ mediante il metodo dei **moltiplicatori di Lagrange**.

$$h = \sum_{i=1}^n c_i^2 - 2\lambda \sum_{i=1}^n c_i - 2\mu \left(\sum_{i=1}^n c_i x_i - 1 \right)$$

Calcolando la derivata parziale di h rispetto a c_i, λ e μ si ottiene

$$\begin{aligned} c_i &= \lambda + \mu x_i, \qquad i = 1, \ldots, n \\ \sum_{i=1}^n c_i &= 0 \\ \sum_{i=1}^n c_i x_i &= 1 \end{aligned}$$

Sommando per i vari valori di i si ottiene $\lambda = -\mu \overline{x}$. Questo valore sostituito nella i-esima equazione ci permette di ottenere $c_i = \mu(x_i - \overline{x})$.

Infine sostituendo questo nell'ultima equazione otteniamo $\mu \sum_{i=1}^n x_i(x_i - \overline{x}) = 1$ ovvero $\mu = 1/\sum_{i=1}^n x_i(x_i - \overline{x}) = 1/\sum_{i=1}^n (x_i - \overline{x})^2$ e quindi

$$c_i = \mu(x_i - \overline{x}) = \frac{(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

e quindi

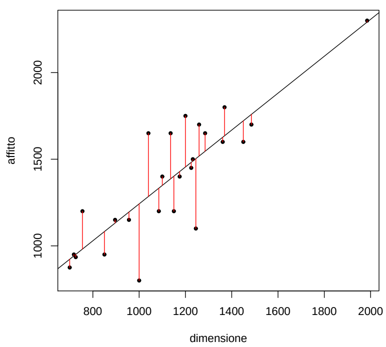
$$b = \sum_{i=1}^n \frac{(x_i - \overline{x}) Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = \widehat{\beta}.$$

I residui: stima di σ^2

Le differenze tra i valori osservati della risposta ed i valori "previsti" dal modello, ovvero,

$$r_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i \ \ (i = 1, \ldots, n)$$

sono usualmente chiamati **residui**.



E' facile verificare che la media dei residui è nulla. Infatti

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

La varianza dei residui, che, per quanto appena detto, coincide con la media dei quadrati dei residui, può essere utilizzata per avere una "idea numerica" della bontà di adattamento del modello ai dati. Infatti, più la varianza dei residui sarà piccola, più la retta di regressione "spiega" le variazioni della risposta.

Si osservi che la varianza dei residui è sempre non più grande della varianza della variabile y. Infatti

$$\begin{aligned} \mathsf{var}(y) &= \inf_{\alpha \in R} \sum (y_i - \alpha)^2/n \geq \\ \geq & \inf_{(\alpha,\beta) \in R^2} \sum (y_i - \alpha - \beta x_i)^2/n = \mathsf{var}(r_1, \ldots, r_n). \end{aligned}$$

$\mathsf{var}(r_1, \ldots, r_n)$ può agevolmente essere calcolata come

$$\mathsf{var}(r_1, \ldots, r_n) = \mathsf{var}(y) - \frac{\mathsf{cov}^2(x, y)}{\mathsf{var}(x)}. \tag{A.9}$$

Infatti

$$\begin{aligned} \mathsf{var}(r_1, \ldots, r_n) &= \frac{1}{n} \sum_{i=1}^{n} r_i^2 = \\ &= \frac{1}{n} \sum_{i=1}^{n} [(y_i - \overline{y}) - \widehat{\beta}(x_i - \overline{x})]^2 = \\ &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2 + \frac{\widehat{\beta}^2}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 - \\ & \quad - \frac{2\widehat{\beta}}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \\ &= \mathsf{var}(y) + \widehat{\beta}^2 \mathsf{var}(x) - 2\widehat{\beta}\mathsf{cov}(x, y) = \\ &= \mathsf{var}(y) + \mathsf{cov}^2(x, y)/\mathsf{var}(x) - 2\mathsf{cov}^2(x, y)/\mathsf{var}(x) = \\ &= \mathsf{var}(y) - \mathsf{cov}^2(x, y)/\mathsf{var}(x) \end{aligned}$$

Infine dalla (A.9) si ha

$$\begin{aligned}\mathsf{var}(r_1,\ldots,r_n) &= \mathsf{var}(y)\left(1-\frac{\mathsf{cov}^2(x,y)}{\mathsf{var}(x)\mathsf{var}(y)}\right)\\ &= \mathsf{var}(y)\left(1-\mathsf{r}^2(x,y)\right)\end{aligned}$$

relazione in cui appare il ruolo della correlazione lineare tra x e y.

Alla luce di quanto detto, la varianza dei residui può essere considerata una stima di σ^2 ovvero della varianza attorno alla retta $\alpha+\beta x$.

Per motivi legati alla distribuzione delle variabili casuali che entrano in gioco per la stima di σ^2 si preferisce utilizzare

$$\widehat{\sigma}^2=\frac{\sum_{i=1}^n r_i^2}{n-2}$$

perchè il relativo stimatore risulta essere non distorto.

Per il calcolo di $\widehat{\sigma}^2$, teniamo conto che:

$$\begin{aligned}\mathsf{var}(r_1,\ldots,r_n) &= \mathsf{var}(y)-\mathsf{cov}^2(x,y)/\mathsf{var}(x)\\ &= \mathsf{var}(y)-\widehat{\beta}^2\mathsf{var}(x)\\ &= \frac{1}{n}\sum_{i=1}^n y_i^2-\overline{y}^2-\widehat{\beta}^2\left(\frac{1}{n}\sum_{i=1}^n x_i^2-\overline{x}^2\right)\end{aligned}$$

e che $\widehat{\sigma}^2 = \frac{n}{n-2}\text{var}(r_1, \ldots, r_n)$; quindi

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n y_i^2 - n\overline{y}^2 - \widehat{\beta}^2 \left(\sum_{i=1}^n x_i^2 - n\overline{x}^2\right)}{n-2}$$

Notiamo che, oltre alle 4 quantità di pagina 20, ci basta calcolare $\sum_{i=1}^n y_i^2$.

Nel nostro caso otteniamo $\widehat{\sigma}^2 = 37867.37$.

Modello gaussiano e stime di massima verosimiglianza

Aggiungiamo ora alle ipotesi classiche del modello di regressione lineare semplice la seguente assunzione:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{A.10}$$

Le implicazioni di questa ipotesi sono le seguenti:

1. gli errori $\varepsilon_i,\ i = 1, \ldots, n$ sono stocasticamente indipendenti;
2. le variabili $Y_i,\ i = 1, \ldots, n$ sono stocasticamente indipendenti, e

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2). \tag{A.11}$$

Sulla base di questi risultati possiamo definire la funzione di densità congiunta delle osservazioni $y_i,\ i = 1, \ldots, n$:

$$f(y_1, \ldots, y_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}{\sigma^2}\right\}.$$

Siamo quindi in grado di definire la funzione di verosimiglianza ($y = (y_1, \ldots, y_n)'$):

$$L(\alpha, \beta, \sigma^2; y) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}{\sigma^2}\right\}$$

e la funzione di log-verosimiglianza:

$$l((\alpha, \beta, \sigma^2; y) = -\frac{n}{2}(\ln(2\pi)+\ln(\sigma^2))-\frac{1}{2}\frac{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}{\sigma^2}.$$

È immediato verificare che L e l dipendono da α, β attraverso la quantità $-\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$. Massimizzare la verosimiglianza rispetto a questi parametri equivale quindi a minimizzare la funzione $s^2(\alpha, \beta)$ introdotta in (A.6) . Gli stimatori di massima verosimiglianza di α e di β coincideranno quindi con gli stimatori dei minimi quadrati. Sostituendo questi stimatori nella log-verosimiglianza otteniamo la **log-verosimiglianza profilo**

$$l^*(\sigma^2; y) = -\frac{1}{2}\left(n\ln(\sigma^2) + \frac{\sum_{i=1}^{n} r_i^2}{\sigma^2}\right) + \text{costante.}$$

dove $r_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i$ sono i residui ottenuti con il metodo dei minimi quadrati.

Lo stimatore di massima verosimiglianza di σ^2 è ottenuto minimizzando l^* e si ottiene:

$$\widehat{\sigma}^2_{MV} = \frac{\sum_{i=1}^{n} r_i^2}{n} = \frac{(n-2)}{n}\widehat{\sigma}^2,$$

Proprietà degli stimatori di massima verosimiglianza

Essendo tali stimatori combinazioni lineari di variabili casuali normali e stocasticamente indipendenti, avremo

$$\widehat{\alpha} \sim \mathcal{N}\left(\alpha, \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\sigma^2\right)$$

$$\widehat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)$$

Infine si può dimostrare

1. lo stimatore $\widehat{\sigma}^2_{MV}$ è distorto

$$\mathbb{E}(\widehat{\sigma}^2_{MV}) = \frac{(n-2)}{n}\mathbb{E}(\widehat{\sigma}^2) = \frac{n-2}{n}\sigma^2.$$

2.

$$\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n R_i^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2 \sim \chi^2_{n-2}$$

e quindi

$$\widehat{\sigma}^2_{MV} \sim \frac{\sigma^2}{n}\chi^2_{n-2}$$

3. lo stimatore $\widehat{\sigma}^2$ è indipendente da $\widehat{\alpha}$ e $\widehat{\beta}$.

Riassumendo possiamo concludere che, nel caso in cui le Y_i abbiano distribuzione normale, si ha

$$\frac{\widehat{\alpha}-\alpha}{\sigma\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \sim \mathcal{N}(0,1)$$

$$\frac{\widehat{\beta}-\beta}{\sigma\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \sim \mathcal{N}(0,1)$$

Un problema di verifica d'ipotesi

L'agente immobiliare aveva sempre applicato una semplice valutazione del prezzo d'affitto: il prezzo era proporzionale alla dimensione dell'appartamento, e affittava gli appartamenti ad un dollaro per piede al quadrato. Vediamo ora se questa valutazione è errata in particolare sottoponiamo a verifica il sistema d'ipotesi

$$\begin{cases} \mathsf{H}_0 : \beta = \beta_0 \\ \mathsf{H}_1 : \beta \neq \beta_0 \end{cases}$$

con $\beta_0 = 1$. Accettare H_0, infatti, equivale a dire che,

$$\text{affitto medio} = \alpha + \text{dimensione}$$

- Per verificare il sistema d'ipotesi utilizziamo la statistica test

$$T = \frac{(\widehat{\beta} - \beta_0)}{\sigma \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}}.$$

Però noi non conosciamo σ. Quindi, con i dati a disposizione, non posso calcolare il valore osservato di T, e cioè t_{oss}. D'altra parte, poichè abbiamo a disposizione una stima di σ, $\widehat{\sigma}$, una statistica test analoga è data da

$$T = \frac{(\widehat{\beta} - \beta_0)}{\widehat{\sigma} \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}}.$$

Se H_0 è vera ci aspettiamo che t_{oss} assuma valori vicini zero. Invece, se H_1 è vera ci aspettiamo che t_{oss} cada lontano da zero.

Se sono valide le ipotesi del modello di regressione lineare semplice e gli errori sono v.c. $\mathcal{N}(0, \sigma^2)$

$$T \sim t_{n-2} \quad (t \text{ di Student con } n-2 \text{ gradi di libertà.})$$

Si noti qui che i gradi di libertà sono pari a $n - 2$.

Supponiamo di porre $\alpha = 0.05$. Allora

$$\begin{array}{c}
t_{n-2,1-\alpha/2} = t_{23,0.975} = 2.07 \\
\Big\downarrow \\
t_{oss} = \dfrac{1.065 - 1}{194.6\sqrt{\frac{1}{1999747}}} = 0.472 \\
\downarrow \\
-2.07 \leq 0.472 \leq 2.07 \text{ ?} \\
\Big\downarrow \text{si} \\
\text{accettiamo } H_0
\end{array}$$

E quindi concludiamo che la valutazione dell'agente era plausibile.

Intervalli di confidenza per β

Nelle ipotesi precedenti si può determinare un intervallo di confidenza per β. Infatti

$$\Pr(-t_{n-2,1-\alpha/2} \leq \frac{\widehat{\beta}-\beta}{\widehat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (x_i-\overline{x})^2}}} \leq t_{n-2,1-\alpha/2}) = 1-\alpha.$$

Ma allora, scrivendo le due disuguglianze in termini di β, si ha

$$\Pr\left(\widehat{\beta} - \sqrt{\widehat{\mathsf{var}}(\widehat{\beta})}\, t_{n-2,1-\alpha/2} \leq \beta \leq \widehat{\beta} + \sqrt{\widehat{\mathsf{var}}(\widehat{\beta})}\, t_{n-2,1-\alpha/2}\right)$$
$$= 1-\alpha$$

dove

$$\sqrt{\widehat{\mathsf{var}}(\widehat{\beta})} = \widehat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (x_i-\overline{x})^2}}$$

ovvero

$$\left[\widehat{\beta} - \widehat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (x_i-\overline{x})^2}}\, t_{n-2,1-\alpha/2}, \widehat{\beta} + \widehat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (x_i-\overline{x})^2}}\, t_{n-2,1-\alpha/2}\right]$$

è un intervallo di confidenza di livello $1-\alpha$ per β.

Supponiamo, ad esempio, di voler un intervallo di confidenza di livello 0.95. Allora, $t_{n-2,1-\alpha/2} = t_{23,0.975} = 2.07$. Ricordando che $\widehat{\beta} \approx 1.065$ e $\widehat{\sigma} \approx 194.6$ e quindi

$$\sqrt{\widehat{\mathsf{var}}(\widehat{\beta})} \approx 194.6\sqrt{1/1999747} \approx 0.1376,$$

la semi-ampiezza dell'intervallo richiesto è

$$0.285 \approx 0.1376 \times 2.07$$

mentre l'intervallo stesso è

$$[1.065 - 0.285 \ \ ; \ \ 1.065 + 0.285] = [0.780 \ \ ; \ \ 1.350]$$

Si osservi che l'intervallo include il valore $\beta = 1$. Questo era un risultato atteso. (Perché ?)

Esercizio: Determinare l'intervallo di confidenza per α

Decomposizione della devianza e coefficiente di determinazione lineare

Consideriamo la seguente misura di variabilità

$$\mathrm{DEV}_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

dette devianza totale. Questa può essere decomposta in

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i - \overline{y})^2 & = & \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 & + & \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\ \mathrm{DEV}_{tot} & = & \mathrm{DEV}_{reg} & + & \mathrm{DEV}_{res} \end{array}$$

dove $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$ e DEV_{reg} rappresenta la parte di DEV_{tot} spiegata dalla regressione e si dice appunto *devianza della regressione* (si osservi che essa rappresenta la devianza dei termini $\widehat{y}$), mentre DEV_{res} rappresenta quella parte di variabilità di Y che il modello di regressione semplice non riesce a cogliere e si dice *devianza residua*.

si ha che

$$1 = \frac{\mathrm{DEV}_{reg}}{\mathrm{DEV}_{tot}} + \frac{\mathrm{DEV}_{res}}{\mathrm{DEV}_{tot}}.$$

Dal risultato precedente segue immediatamente che

$$R^2 = \frac{\mathrm{DEV}_{reg}}{\mathrm{DEV}_{tot}} = 1 - \frac{\mathrm{DEV}_{res}}{\mathrm{DEV}_{tot}} \tag{A.12}$$

sarà una quantità sempre compresa tra 0 e 1. Quanto più essa si avvicinerà a 1, tanto meglio i valori $\widehat{y}_i$ approssimeranno i valori y_i. Viceversa, quando R^2 si avvicinerà a 0 queste approssimazioni saranno poco soddisfacenti. L'indice R^2 quantifica quindi la bontà di adattamento del modello ai dati e si dice *coefficiente di determinazione lineare*.

Diamo ora una espressione alternativa per $\sum_{i=1}^n(\widehat{y}_i - \overline{y})^2$. Si ha

$$
\begin{aligned}
\sum_{i=1}^n(\widehat{y}_i - \overline{y})^2 &= \sum_{i=1}^n(\widehat{\alpha} + \widehat{\beta}x_i - \overline{y})^2 \\
&= \sum_{i=1}^n(\overline{y} - \widehat{\beta}\overline{x} + \widehat{\beta}x_i - \overline{y})^2 \\
&= \widehat{\beta}^2 \sum_{i=1}^n(x_i - \overline{x})^2
\end{aligned}
$$

Infine proviamo il seguente risultato

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2+\sum_{i=1}^{n}(y_i-\widehat{y_i})^2$$

Dimostrazione:

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \sum_{i=1}^{n}(y_i-\widehat{y_i}+\widehat{y_i}-\overline{y})^2\\ &= \sum_{i=1}^{n}(y_i-\widehat{y_i})^2+\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2\\ &\quad +2\sum_{i=1}^{n}(y_i-\widehat{y_i})(\widehat{y_i}-\overline{y})\end{aligned}$$

e

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\widehat{y_i})(\widehat{y_i}-\overline{y}) &= \sum_{i=1}^{n}(y_i-\widehat{y_i})\widehat{y_i}\\ &= \sum_{i=1}^{n}(y_i-\widehat{y_i})(\widehat{\alpha}+\widehat{\beta}x_i)\\ &= \widehat{\beta}\sum_{i=1}^{n}(y_i-\widehat{y_i})x_i\\ &= \widehat{\beta}\sum_{i=1}^{n}(y_i-\widehat{\alpha}-\widehat{\beta}x_i)x_i\\ &= \widehat{\beta}\sum_{i=1}^{n}[(y_i-\overline{y})-\widehat{\beta}(x_i-\overline{x})]x_i\\ &= n\widehat{\beta}[\mathsf{cov}(y,x)-\widehat{\beta}\mathsf{var}(x)]\\ &= n\widehat{\beta}\left[\mathsf{cov}(y,x)-\frac{\mathsf{cov}(y,x)}{\mathsf{var}(x)}\mathsf{var}(x)\right]=0\end{aligned}$$

Nota: La relazione

$$\sum_{i=1}^{n}(y_i - \widehat{y_i})x_i = \sum_{i=1}^{n} r_i x_i = 0$$

insieme alla

$$\sum_{i=1}^{n}(y_i - \widehat{y_i}) \cdot 1 = \sum_{i=1}^{n} r_i \cdot 1 = 0$$

che avevamo precedentemente dimostrato esprimono un'importante proprietà dei residui ovvero che il vettore dei residui $\mathbf{r} = (r_1, \ldots, r_n)'$ è ortogonale rispetto al vettore $\mathbf{x} = (x_1, \ldots, x_n)'$ e al vettore $\mathbf{1} = (1, \ldots, 1)'$ [3].

[3] Un vettore $\mathbf{a} = (a_1, \ldots, a_n)'$ dello spazio $\mathbb{R}^n$ si dice ortogonale ad un altro vettore di $\mathbb{R}^n$ $\mathbf{b} = (b_1, \ldots, b_n)'$ se $\sum_{i=1}^{n} a_i b_i = 0$.

L'analisi della varianza nella regressione

Sappiamo che

$$T = \frac{\frac{(\widehat{\beta}-\beta)}{\sigma\sqrt{\frac{1}{\sum_{i=1}^n (x_i-\overline{x})^2}}}}{\sqrt{\frac{\widehat{\sigma}^2}{\sigma^2}}} = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} \sim t_{n-2}$$

per cui

$$T^2 = \frac{\chi^2_1/1}{\chi^2_{n-2}/(n-2)} \sim F_{1,n-2}$$

Posto $\beta = 0$ il rapporto precedente diventa

$$\begin{aligned} T^2 &= \frac{\widehat{\beta}^2 \sum_{i=1}^n (x_i - \overline{x})^2}{\widehat{\sigma}^2} \\ &= \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2/1}{\sum_{i=1}^n r_i^2/(n-2)} \\ &= \frac{\mathrm{DEV}_{reg}/1}{\mathrm{DEV}_{res}/(n-2)} \sim F_{1,n-2} \end{aligned}$$

Per cui il sistema d'ipotesi $\begin{cases} \mathsf{H}_0 : \beta = 0 \\ \mathsf{H}_1 : \beta \neq 0 \end{cases}$ può essere verificato con livello di significatività α $(0 \leq \alpha \leq 0.5)$ calcolando il valore di

$$F_{oss} = \frac{\mathrm{DEV}_{reg}/1}{\mathrm{DEV}_{res}/(n-2)}$$

e rifiutando l'ipotesi H_0 se $F_{oss} > F_{1-\alpha,1,n-2}$.

Tabella dell'analisi della varianza

Causa	Somma dei quadrati	Gradi di libertà	Stima della varianza
x	$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$	1	$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$
Residuo	$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	n-2	$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2/(n-2)$
Totale	$\sum_{i=1}^{n}(y_i - \overline{y})^2$	n-1	$\sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)$

Nel nostro esempio

Causa	Somma dei quadrati	Gradi di libertà	Stima della varianza
x	2268776.55	1	2268776.55
Residuo	870949.45	23	37867.37
Totale	3139726	24	130821.92

Quindi $F_{oss} = 59.91$ e poiché $\alpha = 0.05$, $F_{0.95,1,23} = 4.28$, concludiamo rifiutando l'ipotesi nulla.

Un esempio di *output* di R

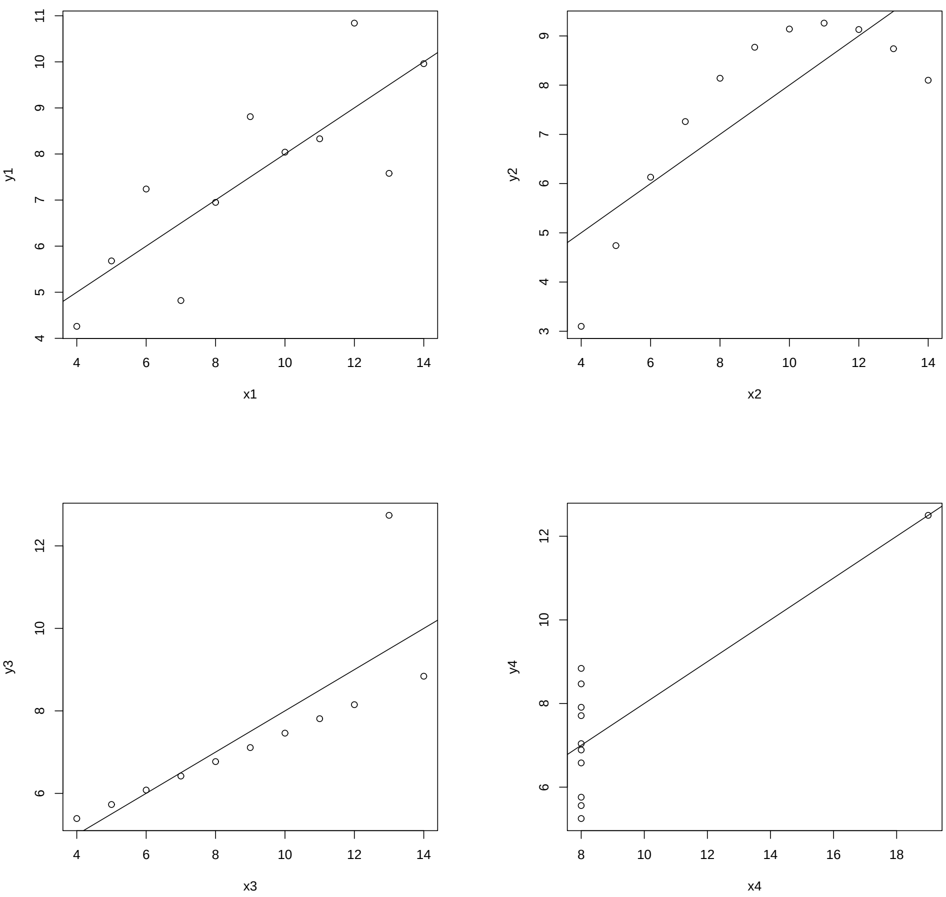
```
Call:
lm(formula = affitto ~ dimensione, data = rent)

Residuals:
    Min      1Q  Median      3Q     Max
-442.26  -58.86  -15.42  104.17  365.13

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 177.1208   161.0043    1.10    0.283
dimensione    1.0651     0.1376    7.74 7.52e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 194.6 on 23 degrees of freedom
Multiple R-squared: 0.7226,        Adjusted R-squared: 0.7105
F-statistic: 59.91 on 1 and 23 DF,  p-value: 7.518e-08
```

```
Analysis of Variance Table

Response: affitto
           Df  Sum Sq Mean Sq F value    Pr(>F)
dimensione  1 2268777 2268777  59.914 7.518e-08 ***
Residuals  23  870949   37867
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L'importanza dell'analisi grafica

Vogliamo illustrare con un esempio dovuto a Anscombe [4] l'importanza dei grafici nell'analisi della regressione.

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89



[4] F. Anscombe (1983). Graphs in statistical analysis, *American Statistician*, **27**, 17-21

```
Call:
lm(formula = y1 ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.92127 -0.45577 -0.04136  0.70941  1.83882 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   3.0001     1.1247   2.667  0.02573 * 
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6665,       Adjusted R-squared: 0.6295 
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.002170 



Call:
lm(formula = y2 ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9009 -0.7609  0.1291  0.9491  1.2691 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)    3.001      1.125   2.667  0.02576 * 
x2             0.500      0.118   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6662,       Adjusted R-squared: 0.6292 
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179 



Call:
lm(formula = y3 ~ x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1586 -0.6146 -0.2303  0.1540  3.2411 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   3.0025     1.1245   2.670  0.02562 * 
x3            0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6663,       Adjusted R-squared: 0.6292 
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176 



Call:
lm(formula = y4 ~ x4)

Residuals:
```

```
       Min         1Q     Median         3Q        Max
-1.751e+00 -8.310e-01  1.110e-16  8.090e-01  1.839e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6667,	Adjusted R-squared: 0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

Previsione

Il nostro agente immobiliare vuol conoscere il prezzo dell'affitto di un appartamento di 1300 piedi al quadrato. Se riconosce come valido il modello di regressione lineare semplice, l'agente può essere interessato a due quantità

$$\mathbb{E}(Y_0) = \alpha + \beta x_0$$

dove $x_0 = 1300$ oppure a

$$Y_0 = \alpha + \beta x_0 + \varepsilon_0.$$

Ambedue queste quantità sono non note ma si noti che la prima è un parametro (la media della distribuzione di Y_0 in corrispondenza di x_0) la seconda è una variabile casuale.

Consideriamo ora una stima di $\mathbb{E}(Y_0) = \alpha + \beta x_0$ utilizzando gli stimatori $\widehat{\alpha}$ e $\widehat{\beta}$ ottenuti con i minimi quadrati

$$\widehat{Y}_0 = \widehat{\alpha} + \widehat{\beta} x_0$$

. Tale stimatore è uno stimatore lineare e non distorto la cui varinza è data da

$$\begin{aligned}
\mathbb{V}ar(\widehat{Y}_0) &= \mathbb{V}ar(\widehat{\alpha}) + x_0^2\mathbb{V}ar(\widehat{\beta}) + 2x_0\mathbb{C}ov(\widehat{\alpha}, \widehat{\beta}) \\
&= \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=j}^n (x_j - \overline{x})^2}\right)\sigma^2 \\
&\quad + \frac{x_0^2}{\sum_{i=j}^n (x_j - \overline{x})^2}\sigma^2 \\
&\quad - \frac{2x_0\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}\sigma^2 \\
&= \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=j}^n (x_j - \overline{x})^2}\right)\sigma^2
\end{aligned}$$

Si può dimostrare nelle ipotesi del modello di regressione lineare semplice che $\widehat{Y}_0 = \widehat{\alpha} + \widehat{\beta}x_0$ è lo stimatore lineare a varianza minima tra gli stimatori lineari di $\mathbb{E}(Y_0)$.

Poichè $\widehat{Y}_0$ è una funzione lineare di $\widehat{\alpha}$ e $\widehat{\beta}$ quindi di $Y_i,\ i = 1, \ldots, n$, se il modello è gaussiano, allora

$$\widehat{Y}_0 \sim \mathcal{N}\left(\alpha + \beta x_0, \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=j}^n (x_j - \overline{x})^2}\right)\sigma^2\right). \quad \text{(A.13)}$$

$\widehat{\alpha}$ e $\widehat{\beta}$ sono indipendenti da $\widehat{\sigma}^2$ e quindi

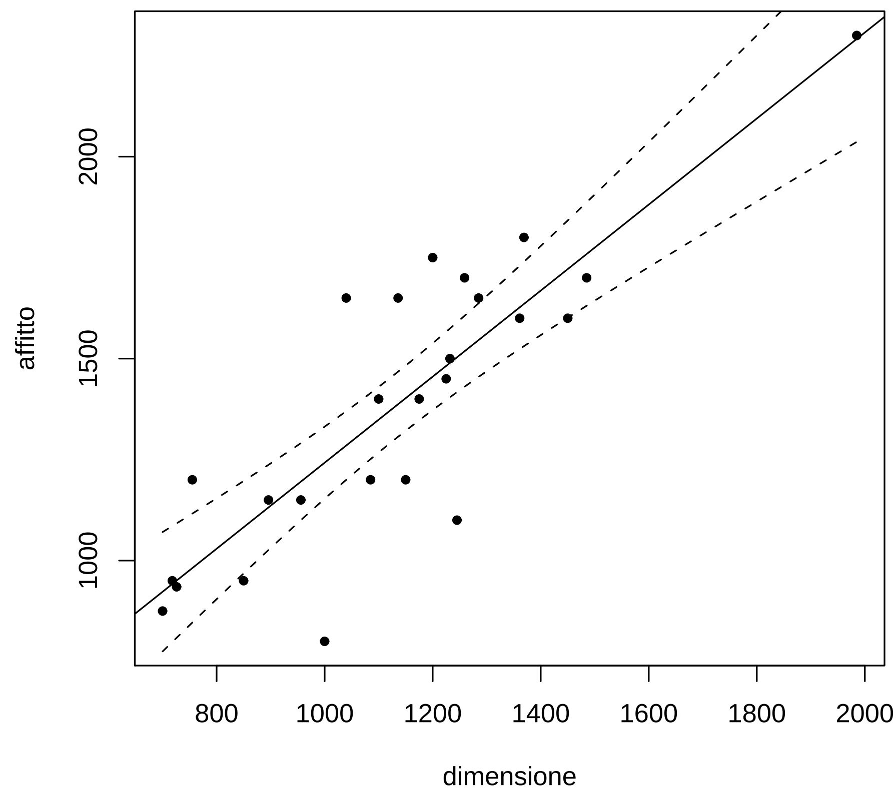
$$\frac{\widehat{Y}_0 - \alpha - \beta x_0,}{\widehat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=j}^n (x_j - \overline{x})^2}\right)}} \sim t_{n-2}.$$

Posto

$$s(x_0) = \widehat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=j}^{n}(x_j - \overline{x})^2}\right)}$$

e procedendo come abbiamo fatto con gli intervalli di confidenza per α e β, possiamo costruire un intervallo di confidenza al livello di confidenza $1 - \alpha$ ($0 \leq \alpha \leq 0.5$) anche per la funzione di regressione sul generico punto x_0:

$$[\widehat{Y}_0 - t_{n-2,1-\alpha/2}s(x_0), \widehat{Y}_0 + t_{n-2,1-\alpha/2}s(x_0)].$$



Vogliamo però spingerci oltre e fornire non tanto una stima di $E(Y_0)$, quanto piuttosto dare indicazioni circa il comportamento di

$$Y_0 = \alpha + \beta x_0 + \varepsilon_0$$

con $\mathbb{E}(\varepsilon_0) = 0$ $\mathbb{V}ar(\varepsilon_0) = \sigma^2$ e $\mathbb{E}(\varepsilon_0 \varepsilon_i) = 0$, $i = 1, \ldots, n$.

Ci muoveremo pertanto nel contesto della *previsione statistica*. In termini statistici la previsione non è necessariamente legata all'anticipazione di eventi futuri. Parleremo di previsione, in termini statistici, ogni qual volta ci porremo il problema di approssimare il comportamento di una variabile che per qualche ragione non si può osservare. Come previsore utilizziamo ancora $\widehat{Y}_0$ e definiamo

$$e_0 = Y_0 - \widehat{Y}_0$$

che è detto errore di previsione. Alcune sue caratteristiche

1. $\mathbb{E}(e_0) = \mathbb{E}(Y_0) - \mathbb{E}(\widehat{\alpha}) - \mathbb{E}(\widehat{\beta})x_0 = \alpha + \beta x_0 - \alpha - \beta x_0 = 0$

2.

$$\begin{aligned}
\mathbb{E}(e_0^2) &= \mathbb{E}\left[\varepsilon_0 - (\widehat{Y}_0 - \alpha - \beta x_0)\right]^2 \\
&= \mathbb{E}(\varepsilon_0^2) + \mathbb{V}ar(\widehat{Y}_0) \\
&= \sigma^2 + \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=j}^n (x_j - \overline{x})^2}\right)\sigma^2 \\
&= \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=j}^n (x_j - \overline{x})^2}\right)\sigma^2
\end{aligned}$$

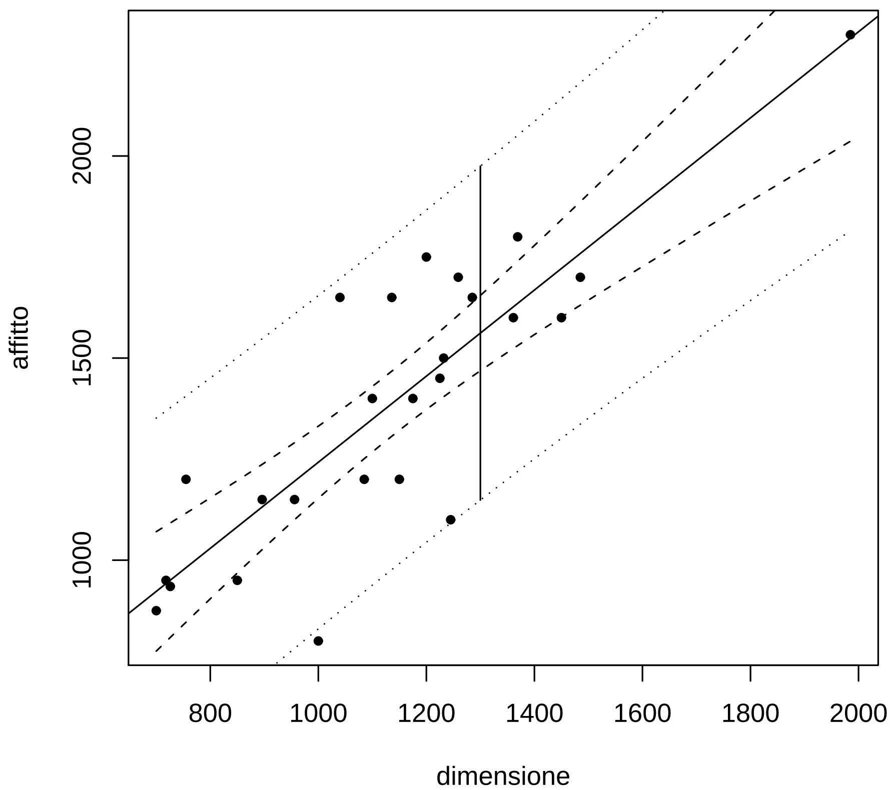
poiché ε_0 è incorrelato con gli altri errori ε_i e quindi con Y_i.

Supponendo che i disturbi del modello siano gaussiani, avremo

$$Y_0 - \widehat{Y}_0 \sim \mathcal{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}\right)\right)$$

Esercizio. Costruire un intervallo di previsione al livello di fiducia $1-\alpha$ per l'ignoto valore di Y_0 in corrispondenza di x_0.

Il grafico qui sotto mostra per l'appunto un insieme di questi intervalli di previsione che vengono confrontati con gli intervalli di confidenza per la media.



Infine ecco la risposta al problema dell'agente immobiliare. La sua previsione per il prezzo d'affitto è

$$\widehat{Y}_0 = 177.12 + 1.07 * 1300 = 1561.81$$

con un intervallo di previsione di livello 0.95 pari a

$$[1561.81 - 413.19, 1561.81 - 413.19] = [1148.62, 1975].$$

Unità B

Consumo di benzina

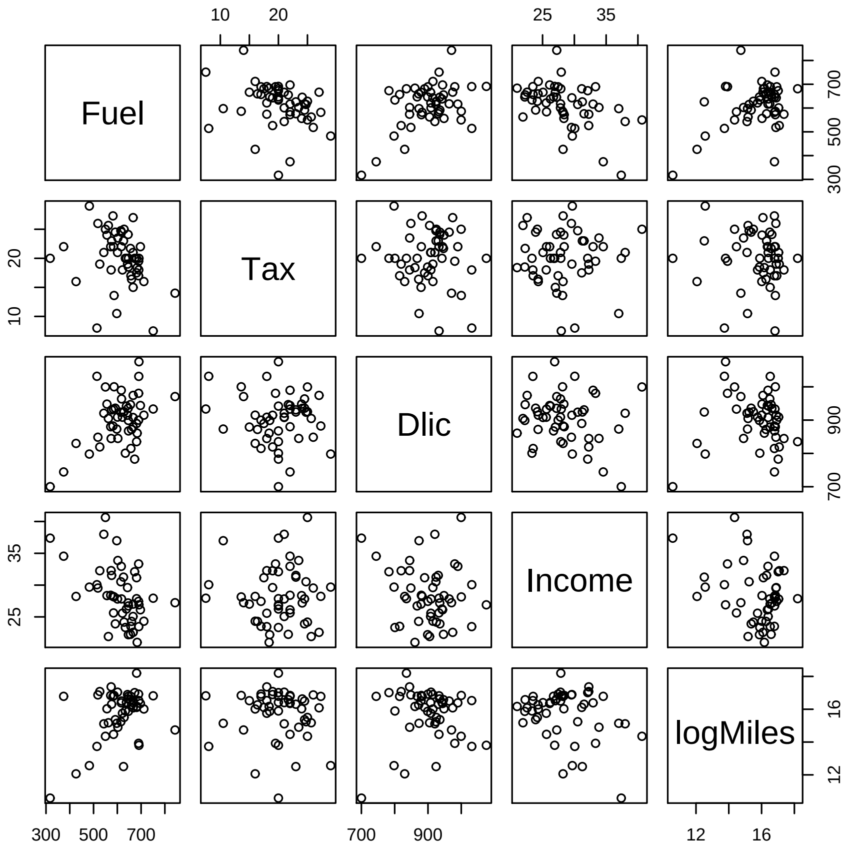
I dati

L'obbiettivo di questo esempio[1] è quello di comprendere quali sono le determinanti del consumo di benzina in 50 stati degli Stati Uniti e nel distretto della Columbia. Nello specifico si è interessati a comprendere quali sono gli effetti delle tasse locali sulla benzina sul consumo di questa. Per ogni stato vengono osservate queste variabili:

- il numero di patenti (`Drivers`);
- quantità di benzina venduta (in migliaia di galloni) (`FuelC`);
- reddito personale medio per l'anno 2000 (in migliaia di dollari) (`Income`);
- miglia di autostrade pubbliche (`Miles`);
- popolazione d'età maggiore o uguale a 16 anni (nel 2001) (`Pop`);
- imposte locali sulla benzina (centesimi per gallone) (`Tax`).

Poiché sia (`Drivers`) e (`FuelC`) sono dei totali e dipendono dalla numerosità della popolazione di uno stato e il reddito è calcolato per persona di derivano due nuove variabili `Dlic = Drivers/Pop` e `Fuel = FuelC/Pop`. Inoltre anche la variabile `Miles` viene trasformata e si considera $\log_2$(`Miles`).

[1] Tratto da Weisberg, S. (2005). *Applied Linear Regression (thrd ed.)*, Wiley, New York.


Fuel
Tax
Dlic
Income
logMiles
10
20
25
35
300
500
700
900
12
16

Sino ad ora abbiamo supposto che una variabile di interesse, la variabile risposta, venisse posta in relazione con una sola variabile indipendente. Definiremo ora una classe di modelli più generale, di cui il modello di regressione lineare semplice costituisce un caso particolare. Supponiamo ora che la variabile risposta, Y, venga posta in relazione con più variabili $x_1, x_2, \ldots, x_p$, precisamente

$$Y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon,$$

con $\mathbb{E}(\varepsilon) = 0$ e $\mathbb{V}ar(\varepsilon) = \sigma^2$. Per questo avremo

$$\begin{aligned} \mathbb{E}(Y) &= \beta_1 x_1 \cdots + \beta_p x_p \\ &= \mathbf{x}'\boldsymbol{\beta} \end{aligned}$$

con $\mathbf{x} = [x_1, \ldots, x_p]'$ e $[\beta_1, \ldots, \beta_p]'$ appartenenti a $\mathbb{R}^p$.

Consideriamo ora un campione di dimensione n dal precedente modello le variabili

$$\begin{aligned} Y_i &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \end{aligned}$$

con $\mathbf{x}_i' = [x_{i,1}, \ldots, x_{i,p}]$.

Nel nostro caso abbiamo osservato y_i e i valori $x_{i,1}, \ldots, x_{i,p}$ sono stati assegnati. Mostriamo i dati relativi ai primi 10 stati.

	Fuel	Dlic	Income	logMiles	Tax
1	690.2644	1031.3801	23.471	16.52711	18.0
2	514.2792	1031.6411	30.064	13.73429	8.0
3	621.4751	908.5972	25.578	15.75356	18.0
4	655.2927	946.5706	22.257	16.58244	21.7
5	573.9129	844.7033	32.275	17.36471	18.0
6	616.6115	989.6062	32.949	16.38960	22.0
7	549.9926	999.5934	40.640	14.35191	25.0
8	626.0239	924.3448	31.255	12.50532	23.0
9	317.4924	700.1953	37.383	10.58308	20.0
10	586.3461	1000.1242	28.145	16.83983	13.6

Notazione matriciale

Definendo

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_i \\ \vdots \\ \mathbf{x}'_n \end{bmatrix}, \qquad \text{e} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

possiamo scrivere in forma matriciale

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

dove $\mathbf{Y}$ ed $\boldsymbol{\varepsilon}$ sono vettori di dimensione $n \times 1$, $\mathbf{X}$ è una matrice di dimensione $n \times p$ e $\boldsymbol{\beta}$ è un vettore di dimensione $p \times 1$.

Due rappresentazioni di $\mathbf{X}$

1. $\mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_i \\ \vdots \\ \mathbf{x}'_n \end{bmatrix}$ in termini degli n vettori riga $\mathbf{x}'_i$ $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,p}]'$ (è il vettore costituito dai p valori sulle variabili $x_1, \ldots, x_p$ per il campione i)

2. $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_p]$ in termini dei p vettori colonna $\mathbf{X}_j = [x_{1,j}, \ldots, x_{n,j}]'$ (è il vettore costituito dalle n osservazioni sulla variabile x_j)

Vettori aleatori

Sia $\mathbf{W}' = [W_1, \ldots, W_i, \ldots, W_k] \in \mathrm{R}^{\mathrm{k}},\ \mathrm{k} > 1$, un vettore i cui elementi sono variabili casuali univariate. Diremo allora che $\mathbf{W}$ è un *vettore aleatorio, o variabile casuale multivariata*. La media di questo vettore aleatorio sarà data da

$$\mathbb{E}(\mathbf{W}) = \begin{bmatrix} \mathbb{E}(W_1) \\ \vdots \\ \mathbb{E}(W_i) \\ \vdots \\ \mathbb{E}(W_k) \end{bmatrix}.$$

La matrice di varianza e covarianza di $\mathbf{W}$, $\mathbf{\Sigma}_W$ è definita come

$$\mathbf{\Sigma}_W = E[(\mathbf{W} - \mathbb{E}(\mathbf{W}))(\mathbf{W} - \mathbb{E}(\mathbf{W}))']$$

ed è una matrice di dimensione $k \times k$. Da questa definizione segue che l'elemento di $\mathbf{\Sigma}_W$ sulla i-esima riga e sulla j-esima colonna sarà:

$$\sigma_{i,j} = \begin{cases} \mathbb{V}ar(W_i) & \text{se } i = j \\ \mathbb{C}ov(W_i, Wj) & \text{se } i \neq j. \end{cases}$$

Evidentemente $\mathbf{\Sigma}_W$ è una matrice simmetrica. Generalmente $\mathbf{\Sigma}_W$ è definita positiva. Sarà semidefinita positiva quando un elemento di $\mathbf{W}$ può essere è combinazione lineare degli altri elementi (cioè delle altre variabili casuali) che costituiscono il vettore aleatorio.

È possibile dimostrare che se $\mathbf{Z} = \mathbf{a} + \mathbf{A}\mathbf{W}$, con $\mathbf{Z}$ e $\mathbf{a}$ vettori di dimensione m, $m \geq 1$, e $\mathbf{A}$ matrice di dimensione $m \times k$, allora

$$\mathbb{E}(\mathbf{Z}) = \mathbf{a} + \mathbf{A}\mathbb{E}(\mathbf{W}) \quad \text{e} \quad \mathbf{\Sigma}_Z = \mathbf{A}\mathbf{\Sigma}_{\mathbf{W}}\mathbf{A}'.$$

Ipotesi classiche del modello di regressione lineare

Il modello di regressione lineare multipla si fonda sulle seguenti assunzioni:

1. Y_i è una variabile univariata e i vettori $\mathbf{x}_i'$ hanno dimensione p, con $p \leq n$;
2. $Y_i = \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i, \quad i = 1, \ldots, n$;
3. $\mathbb{E}(\varepsilon_i) = 0, \quad i = 1, \ldots, n$; i disturbi aleatori hanno media nulla;
4. $\mathbb{V}ar(\varepsilon_i) = \sigma^2 < \infty, \quad i = 1, 2, \ldots, n$; i disturbi aleatori sono omoschedastici;
5. $\mathbb{C}ov(\varepsilon_i, \varepsilon_j) = \mathbb{E}(\varepsilon_i \varepsilon_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \ldots, n$; i disturbi aleatori sono incorrelati;
6. la matrice $\mathbf{X}$ ha rango p, ovvero i p regressori $x_1, \ldots, x_p$ sono linearmente indipendenti.

Nota 1 Può apparire strano che nel modello di regressione multipla non ci sia un termine che non dipenda da x_i. In realtà il modello può includere un'intercetta (valore costante) se si impone che $\mathbf{x}_j = (1, \ldots, 1)'$. Solitamente si pone $j = 1$ e pertanto β_1 ha il ruolo di intercetta. È inoltre evidente che il modello di regressione lineare semplice costituisce un caso particolare del modello di regressione lineare multipla con $p = 2$.

Nota 2 Perché è importante che rango$(\mathbf{X}) = p$?

La condizione di rango implica che i p vettori $\mathbf{X}_j$ siano linearmente indipendenti. Dalla definizione di indipendenza lineare sappiamo che se i p vettori $\mathbf{X}_j$, sono linearmente indipendenti, allora

$$\sum_{j=1}^{p} c_j \mathbf{X}_j = \mathbf{0}, \text{se e solo se } c_j = 0\ \forall j$$

(i coefficienti c_j sono scalari). Supponiamo ora che rango$(\mathbf{X}) = p - 1$ e che quindi solo $p - 1$ colonne di $\mathbf{X}$ siano linearmente indipendenti. Questo significa che esisteranno p coefficienti, non tutti nulli, tali che

$$\sum_{j=1}^{p} c_j \mathbf{X}_j = 0.$$

Supponiamo, in particolare, che $c_p \neq 0$. Dall'uguaglianza precedente segue immediatamente che

$$\mathbf{X}_p = -\sum_{j=1}^{p-1} \frac{c_j}{c_p} \mathbf{X}_j.$$

Questo significa che il $p-$esimo regressore è una combinazione lineare degli altri $p-1$ regressori e che quindi esso è ridondante nel modello, potendo la variabile risposta essere messa in relazione con soli $p-1$ regressori mediante un'equazione del tipo:

$$Y_i = \sum_{j=1}^{p-1} \gamma_j x_{i,j} + \varepsilon_i.$$

Si può dimostrare che, se rango$(\mathbf{X}) = k < p$, allora $p-k$ regressori possono essere considerati ridondanti. Chiaramente ciò significa che si è costruito un cattivo modello e che solo k delle p variabili esplicative devono essere incluse nel modello!

Stima dei parametri con il metodo dei minimi quadrati

Applicando il metodo dei minimi quadrati, dovremo minimizzare rispetto a $\boldsymbol{\beta} = [\beta_1, \ldots \beta_p]'$ la seguente funzione:

$$\begin{aligned} s(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{B.1}$$

Come nel caso del modello di regressione lineare semplice, la stima di $\boldsymbol{\beta}$ sarà quel valore $\widehat{\boldsymbol{\beta}}$ per il quale è minima la distanza tra il vettore delle osservazioni $\mathbf{y}$ e il vettore delle loro approssimazioni fornito dal modello di regressione,

$$\widehat{\mathbf{y}} = [\widehat{y}_1, \ldots, \widehat{y}_n]' = \mathbf{X}\widehat{\boldsymbol{\beta}},$$

essendo la (B.1) equivalente a:

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

Sviluppando il prodotto della (B.1) si ottiene:

$$s(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \tag{B.2}$$

Il gradiente di tale funzione è dato da [2]:

[2] Ricordando che

$$\frac{\partial}{\partial \boldsymbol{\beta}} s(\boldsymbol{\beta}) = 2(\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{y}) \tag{B.3}$$

e la matrice hessiana sarà

$$\mathbf{H}(\boldsymbol{\beta}) = 2\mathbf{X}'\mathbf{X}.$$

Si osservi che, se $\mathbf{X}$ ha rango p, allora $\mathbf{H}(\boldsymbol{\beta})$ è una matrice definita positiva per qualsiasi valore di $\boldsymbol{\beta}$. Ne consegue che il valore $\boldsymbol{\beta}$ che annulla (B.3), quel valore cioè che soddisfa alle condizioni necessarie per la minimizzazione di $s(\boldsymbol{\beta})$, costituirà un punto di minimo assoluto della medesima funzione. Dovrà quindi verificarsi che:

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0, \tag{B.4}$$

ovvero, risolvendo l'equazione rispetto a $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{B.5}$$

Questa equazione definisce la stima dei minimi quadrati di $\boldsymbol{\beta}$ nel modello di regressione lineare multipla. È importante

---

1. $$\frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{a}'\boldsymbol{\alpha} = \mathbf{a}, \qquad \mathbf{a} \in \mathbb{R}^d$$

2. $$\frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{a}'\mathbf{A}\boldsymbol{\alpha} = 2\mathbf{A}\boldsymbol{\alpha}, \qquad \mathbf{A} \text{ matrice quadrata}$$

3. $$\frac{\partial}{\partial \boldsymbol{\alpha}'\partial \boldsymbol{\alpha}} \mathbf{a}'\mathbf{A}\boldsymbol{\alpha} = 2\mathbf{A}$$

sottolineare che se il rango di $\mathbf{X}$ fosse inferiore a p, la matrice $(\mathbf{X}'\mathbf{X})^{-1}$ non sarebbe invertibile e quindi non si potrebbe definire $\widehat{\boldsymbol{\beta}}$!

Si osservi che la (B.4) può essere riscritta come

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{X}'(\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{X}'\mathbf{r} = \mathbf{0}, \tag{B.6}$$

dove l'$i-$esimo elemento di $\mathbf{r}$ è $r_i = y_i - \widehat{y}_i$.

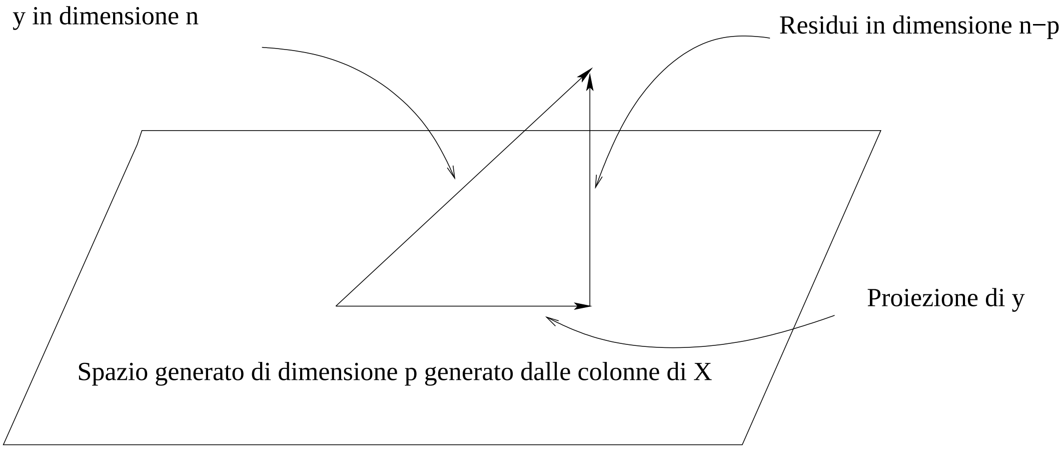
Dalla (B.6) si evince che, se nel modello compare l'intercetta, cioè se la prima colonna di $\mathbf{X}$ coincide con il vettore $n-$dimensionale $\mathbf{1_n}$ con elementi identicamente uguali a 1, allora

$$\mathbf{1}_n'\mathbf{r} = \sum_{i=1}^{n} r_i = 0.$$

Ne consegue immediatamente che, come nel caso del modello di regressione lineare semplice, se nel modello compare l'intercetta,

1. la media aritmetica dei residui è nulla;
2. i residui sono incorrelati con <u>tutti</u> i regressori;
3. i residui sono incorrelati con i valori stimati $\widehat{y}_i$.

Interpretazione geometrica



- $\|\mathbf{a}-\mathbf{b}\| = \sqrt{(\mathbf{a}-\mathbf{b})'(\mathbf{a}-\mathbf{b})} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$ è detta distanza tra i vettori $\mathbf{a}$ e $\mathbf{b}$ di $\mathbb{R}^n$

- $s(\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2$ quindi il metodo dei minimi quadrati minimizza la distanza tra $\mathbf{y}$ e $\mathbf{X}\boldsymbol{\beta}$.

- la matrice $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ è detta matrice di proiezione essa genera il vettore $\widehat{\mathbf{y}} = \mathbf{P}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\beta}}$

- il vettore $\mathbf{r} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ è ortogonale ai vettori $\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_p$ che costituiscono la matrice $\mathbf{X}$

Proprietà dello stimatore dei minimi quadrati

Da quanto visto a proposito dei vettori aleatori, risulta immediatamente che $\widehat{\boldsymbol{\beta}}$ è uno stimatore corretto di $\boldsymbol{\beta}$:

$$\begin{aligned}\mathbb{E}(\widehat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}\end{aligned}$$

e che la sua matrice varianza è data da

$$\begin{aligned}\mathbb{V}ar(\widehat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{V}ar(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I_n}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\mathbf{-1}} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

È possibile dimostrare che il teorema di Gauss-Markov vale anche nel modello di regressione lineare multipla: lo stimatore dei minimi quadrati, $\widehat{\boldsymbol{\beta}}$, è il più efficiente tra tutti gli stimatori di $\boldsymbol{\beta}$ corretti e lineari rispetto a $\mathbf{Y}$, nel senso che, se $\widehat{\boldsymbol{\beta}}^* = \mathbf{AY}$, con $\mathbf{A}$ matrice arbitraria di dimensione $p \times n$, e $\mathbb{E}(\widehat{\boldsymbol{\beta}}^*) = \boldsymbol{\beta}$, allora la matrice $\mathbb{V}ar(\widehat{\boldsymbol{\beta}}^*) - \mathbb{V}ar(\widehat{\boldsymbol{\beta}})$ è semidefinita positiva.

Stima della varianza degli errori

Intuitivamente si potrebbe essere indotti a stimare la varianza σ^2 utilizzando la varianza campionaria dei residui $R_i = Y_i - \widehat{Y}_i$ di regressione e definendo quindi lo stimatore:

$$S^{*2} = \sum_{i=1}^{n} R_i^2/n$$

Si può dimostrare che tale stimatore è distorto, con $\mathbb{E}(S^{*2}) = (n-p)\sigma^2/n$. Si preferisce pertanto utilizzare lo stimatore corretto

$$S^2 = \sum_{i=1}^{n} R_i^2/(n-p) \qquad \text{(B.7)}$$

Esempio in R

```
> XX <- t(X) %*% X
> Xy <- t(X) %*% y
> invXX <- solve(XX)
> beta.hat <- invXX %*% Xy
> beta.hat
```

```
                   [,1]
(Intercept) 154.1928446
Tax          -4.2279832
Dlic          0.4718712
Income       -6.1353310
logMiles     18.5452745
```

```
> y.hat <- X %*% beta.hat
> r <- y - y.hat
> sigma2.hat <- as.numeric((t(r) %*% r)/(length(y) - dim(X)[2]))
> varbeta.hat <- sigma2.hat * invXX
> varbeta.hat
```

```
            (Intercept)           Tax         Dlic        Income      logMiles
(Intercept) 37988.41145 -120.09607926 -17.18034682 -251.85813715 -813.33358142
Tax          -120.09608    4.12139164   0.02357820    0.17952544    0.67471601
Dlic          -17.18035    0.02357820   0.01651570    0.05006761    0.02274627
Income       -251.85814    0.17952544   0.05006761    4.81202826    4.21173557
logMiles     -813.33358    0.67471601   0.02274627    4.21173557   41.88904244
```

Modello gaussiano e stimatore di massima verosimiglianza

Se, oltre alle ipotesi classiche del modello di regressione lineare multipla, introduciamo l'assunto che $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ le variabili casuali $Y_i,\ i = 1, \ldots, n$, saranno stocasticamente indipendenti e $Y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$.

La funzione di densità congiunta delle osservazioni $y_i,\ i = 1, \ldots, n$:

$$
\begin{aligned}
f(y_1, \ldots, y_n) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{i,j})^2}{\sigma^2}\right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right\}
\end{aligned}
$$

Come per il caso univariato risulta evidente che il valore di $\boldsymbol{\beta}$ che massimizza la funzione di verosimiglianza,

$$
L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\frac{s(\boldsymbol{\beta})}{\sigma^2}\right\}
$$

dove

$$
s(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2
$$

indipendentemente dal valore assunto da σ^2, coinciderà con il valore che minimizza la funzione $s(\boldsymbol{\beta})$. La stima di massima verosimiglianza di $\boldsymbol{\beta}$ sarà quindi:

$$\widehat{\boldsymbol{\beta}}_{MV} = \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Si può dimostrare che la stima di massima verosimiglianza di σ^2 è data da:

$$\begin{aligned}\widehat{\sigma}^2_{MV} &= \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})^2/n. \\ &= \frac{n-p}{n}s^2\end{aligned}$$

A questo punto è facile derivare le espressioni degli stimatori di massima verosimiglianza di $\boldsymbol{\beta}$ e di σ^2 che saranno:

$$\widehat{\boldsymbol{\beta}}_{MV} = \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

e

$$\begin{aligned}\widehat{\sigma}^2_{MV} &= \sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})^2/n \\ &= \frac{1}{n}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \frac{n-p}{n}S^2\end{aligned}$$

dove S^2 è lo stimatore introdotto nella (B.7). Si osservi che $\widehat{\sigma}^2_{MV}$ è uno stimatore distorto di σ^2, essendo $\mathbb{E}(\widehat{\sigma}^2_{MV}) = (n - p)\sigma^2/n$.

Poiché, nel caso del modello gaussiano, $\widehat{\boldsymbol{\beta}}_{MV}$ è una trasformazione lineare di n variabili casuali normali e indipendenti (le Y_i), allora si verifica immediatamente che:

$$\widehat{\boldsymbol{\beta}}_{MV} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Si può inoltre dimostrare che

$$\frac{n\widehat{\sigma}^2_{MV}}{\sigma^2} \sim \chi^2_{(n-p)}.$$

È importante ricordare che nel modello gaussiano gli stimatori $\widehat{\boldsymbol{\beta}}$ e S^2 (e quindi $\widehat{\sigma}^2_{MV}$) rimangono ancora stocasticamente indipendenti. Utilizzeremo questa proprietà quando affronteremo il problema della stima intervallare per i coefficienti di regressione.

Verifica di ipotesi su un singolo coefficiente di regressione

Supponiamo di voler saggiare il sistema di ipotesi:

$$\begin{array}{ll} H_0: & \beta_j = \beta_{j0} \\ H_1: & \beta_j \neq \beta_{j0} \end{array}$$

per $j = 1, \ldots, p$.

Sappiamo che nel modello gaussiano $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ e quindi, per le proprietà della distribuzione normale multivariata, la distribuzione dello stimatore $\widehat{\beta}_j$ di β_j sarà:

$$\widehat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 k_j),$$

dove k_j rappresenta il j-esimo elemento sulla diagonale di $(\mathbf{X}'\mathbf{X})^{-1}$. Se fosse vera H_0 avremmo che

$$\frac{\widehat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 k_j}} \sim \mathcal{N}(0, 1).$$

Non essendo σ^2 noto, nell'espressione precedente si sostituisce ad esso il suo stimatore S^2 introdotto nella (B.7), ottenendo

$$T_j = \frac{\widehat{\beta}_j - \beta_{j0}}{\sqrt{S^2 k_j}}$$

(si osservi che il denominatore di questa espressione è lo stimatore dello standard error di $\widehat{\beta}_j$). Poiché $\widehat{\beta}_j$ e S^2 sono stocasticamente indipendenti e $(n-p)S^2/\sigma^2 \sim \chi^2_{(n-p)}$, sotto H_0 avremo che

$$T_j \sim t_{n-p}.$$

Ad un livello di significatività α, quindi, rifiuteremo l'ipotesi nulla quando $|t_j^{oss}| \geq t_{(n-p),1-\alpha/2}$, dove

$$t_j^{oss} = \frac{\widehat{\beta}_j - \beta_{j0}}{\sqrt{s^2 k_j}}$$

ovvero quando il livello di significatività osservato

$$\alpha^{oss} = \Pr(|T_j| \geq |t_j^{oss}|; H_0) \leq \alpha.$$

Esempio

```
Call:
lm(formula = Fuel ~ Tax + Dlic + Income + logMiles, data = fuel2001)

Residuals:
     Min       1Q   Median       3Q      Max
-163.145  -33.039    5.895   31.989  183.499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.1928   194.9062   0.791 0.432938
Tax          -4.2280     2.0301  -2.083 0.042873 *
Dlic          0.4719     0.1285   3.672 0.000626 ***
Income       -6.1353     2.1936  -2.797 0.007508 **
logMiles     18.5453     6.4722   2.865 0.006259 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64.89 on 46 degrees of freedom
Multiple R-Squared: 0.5105,       Adjusted R-squared: 0.4679
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.33e-07
```

Decomposizione della devianza

Come nel caso del modello di regressione lineare semplice, possiamo ancora definire la devianza totale come

$$
\begin{aligned}
DEV_{tot} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \\
&= \mathbf{y}'\mathbf{y} - n\bar{y}^2. \qquad \text{(B.8)}
\end{aligned}
$$

Ricordando che il vettore $\mathbf{y}$ può essere scritto come

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= \widehat{\mathbf{y}} + \mathbf{r}
\end{aligned}
$$

allora

$$
\begin{aligned}
\mathbf{y}'\mathbf{y} &= (\widehat{\mathbf{y}} + \mathbf{r})'(\widehat{\mathbf{y}} + \mathbf{r}) \\
&= \widehat{\mathbf{y}}'\widehat{\mathbf{y}} + \mathbf{r}'\mathbf{r} + 2\widehat{\mathbf{y}}'\mathbf{r}.
\end{aligned}
$$

Ma poiché sappiamo che $\mathbf{X}'\mathbf{r} = \mathbf{0}$, allora $2\widehat{\mathbf{y}}'\mathbf{r} = 2\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{r} = 0$. Quindi avremo che

$$
\mathbf{y}'\mathbf{y} = \widehat{\mathbf{y}}'\widehat{\mathbf{y}} + \mathbf{r}'\mathbf{r}. \qquad \text{(B.9)}
$$

Quindi, potremo scrivere:

$$\begin{aligned} DEV_{tot} &= \mathbf{y}'\mathbf{y} - n\bar{y}^2 \\ &= \widehat{\mathbf{y}}'\widehat{\mathbf{y}} - n\bar{y}^2 + \mathbf{r}'\mathbf{r}. \end{aligned}$$

Ricordando che, quando nel modello compare l'intercetta, la somma dei residui campionari è nulla, possiamo definire le quantità:

$$\begin{aligned} DEV_{reg} &= \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 \\ &= \widehat{\mathbf{y}}'\widehat{\mathbf{y}} - n\bar{y}^2 \end{aligned}$$

e

$$DEV_{res} = \sum_{i=1}^{n} r_i^2 = \mathbf{r}'\mathbf{r}.$$

Abbiamo quindi dimostrato che, come nel modello di regressione lineare semplice,

$$DEV_{tot} = DEV_{reg} + DEV_{res}. \qquad \text{(B.10)}$$

Inoltre,

$$\widehat{\mathbf{y}}'\widehat{\mathbf{y}} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

Questa somma di quadrati avrà p gradi di libertà, perché essa dipende dai dati attraverso le p funzioni che definiscono $\widehat{\boldsymbol{\beta}}$. Nel calcolo di DEV_{reg} si introduce un vincolo attraverso la stima della media di Y, quindi DEV_{reg} avrà $p-1$ gradi di libertà.

Abbiamo visto che DEV_{res} rappresenta la devianza dei residui. Ricordando che i residui devono soddisfare i p vincoli dervanti dalle p equazioni della (B.4), i gradi di libertà di DEV_{res} saranno $n-p$. Possiamo quindi giungere alla definizione della *tabella dell'analisi della varianza*:

Tabella dell'analisi della varianza

Causa	Somma dei quadrati	Gradi di libertà	Stima della varianza
$x_1, \ldots, x_p$	$\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2$	p-1	$\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2/(p-1)$
Residuo	$\sum_{i=1}^n (y_i - \widehat{y}_i)^2$	n-p	$\sum_{i=1}^n (y_i - \widehat{y}_i)^2/(n-p)$
Totale	$\sum_{i=1}^n (y_i - \overline{y})^2$	n-1	$\sum_{i=1}^n (y_i - \overline{y})^2/(n-1)$

$$\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} - n\bar{Y}^2 + \mathbf{R}'\mathbf{R}.$$

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

Si può dimostrare che

- $DEV_{tot}/\sigma^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/\sigma^2 \sim \chi^2_{n-1}$
- $DEV_{reg}/\sigma^2 = \sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2/\sigma^2 \sim \chi^2_{p-1}$
- $DEV_{res}/\sigma^2 = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2/\sigma^2 \sim \chi^2_{n-p}$
- DEV_{reg} e DEV_{res} sono indipendenti.

Infine ricordiamo che anche in questo caso possiamo definire il *coefficiente di determinazione lineare*

$$R^2 = \frac{\mathrm{DEV}_{reg}}{\mathrm{DEV}_{tot}} = 1 - \frac{\mathrm{DEV}_{res}}{\mathrm{DEV}_{tot}}. \tag{B.11}$$

Comandi in R

```
> fit <- lm(Fuel ~ Tax + Dlic + Income + logMiles, data = fuel2001)
> fit0 <- lm(Fuel ~ 1, data = fuel2001)
> anova(fit0, fit)
```

```
Analysis of Variance Table

Model 1: Fuel ~ 1
Model 2: Fuel ~ Tax + Dlic + Income + logMiles
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     50 395694
2     46 193700  4    201994 11.992 9.33e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Verifica di ipotesi su più coefficienti di regressione.

Supponiamo voler adattare un modello di regressione con p regressori ai dati rilevati su un campione di dimensione n. Nel seguito assumeremo che sussistano le ipotesi classiche del modello di regressione lineare multipla e che il modello sia gaussiano. La forma generale del modello sarà quindi

$$M_1: \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n), \tag{B.12}$$

con $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_p]$ matrice $n \times p$ e $\boldsymbol{\beta} \in \mathbb{R}^p$.

Nelle applicazioni a dati reali è importante saper decidere se sia o meno opportuno escludere dal modello alcuni dei regressori che compaiono in (B.12). In particolare, fissando $q < p$, potremmo ritenere che un modello adeguato contempli solo i q regressori $X_1, \ldots, X_q$. Un modello così definito è:

$$M_0: \quad \mathbf{Y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*,\ \boldsymbol{\varepsilon}^* \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n), \tag{B.13}$$

con $\mathbf{X}^* = [\mathbf{X}_1, \cdots, \mathbf{X}_q]$ matrice $n \times q$ e $\boldsymbol{\beta}^* \in \mathbb{R}^q$.

È evidente che confrontare i due modelli equivale a saggiare il seguente sistema di ipotesi sul modello (B.12):

$$\begin{aligned} H_0: &\quad \beta_{q+1} = \cdots = \beta_p = 0, \\ H_1: &\quad \textit{almeno uno di questi coefficienti è diverso da 0.} \end{aligned}$$

In effetti, il modello M_0 equivale al modello M_1 con l'introduzione del vincolo $\beta_{q+j} = 0, \quad 1 \leq j \leq p - q$. Supponiamo ora di adattare entrambi i modelli ad un fissato campione, ottenendo la stima $\widehat{\boldsymbol{\beta}}$ per il modello M_1 e la stima $\widehat{\boldsymbol{\beta}}^*$ per il modello M_0. Nel seguito, con $\mathbf{r}$ e $\mathbf{r}^*$, $\widehat{\mathbf{y}}$ e $\widehat{\mathbf{y}}^*$ indicheremo i residui e i valori previsti rispettivamente da M_1 e di M_0.

La stima secondo il metodo di MV per il modello M_0 porta a considerare

$$\min_{\boldsymbol{\beta}^*} s^*(\boldsymbol{\beta}^*) = \min_{\boldsymbol{\beta}^*} (\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^*)'(\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^*)$$

La determinazione di questo minimo è equivalente a

$$\min_{\boldsymbol{\beta}} s(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

soggetto a

$$\beta_{q+1} = \cdots = \beta_p = 0.$$

E' allora chiaro che

$$s(\widehat{\boldsymbol{\beta}^*}) \geq s(\widehat{\boldsymbol{\beta}})$$

ovvero

$$\begin{aligned} \mathbf{r}^{*\prime}\mathbf{r}^* \; &\geq \; \mathbf{r}'\mathbf{r} \\ DEV_{res}(M_0) \; &\geq \; DEV_{res}(M_1) \\ DEV_{reg}(M_0) \; &\leq \; DEV_{reg}(M_1). \end{aligned}$$

Intuitivamente saremo propensi a preferire il modello M_0, ovvero ad accettare H_0 qualora $DEV_{res}(M_0)$ fosse non troppo grande rispetto a $DEV_{res}(M_1)$, ovvero qualora

$$\mathbf{r}^{*\prime}\mathbf{r}^{*} - \mathbf{r}'\mathbf{r} \geq 0.$$

fosse prossima a 0, saremmo indotti ad accettare H_0. Relativizziamo questa quantità

$$\tilde{f} = \frac{\mathbf{r}^{*\prime}\mathbf{r}^{*} - \mathbf{r}'\mathbf{r}}{\mathbf{r}'\mathbf{r}}.$$

Evidentemente, la corrispondente v.c. $\tilde{F}$ potrebbe essere utilizzata per saggiare il sistema di ipotesi che ci interessa. Essendo propensi a rifiutare l'ipotesi nulla per valori alti di $\tilde{f}$, se fossimo in grado di determinare la distribuzione di tale statistica sotto l'ipotesi nulla potremmo determinare la regione di rifiuto al livello di significatività α:

$$\mathcal{R} = \{\tilde{f} \geq f_{1-\alpha}\},$$

dove $f_{1-\alpha}$ rappresenta il quantile di ordine $1-\alpha$ della distribuzione di $\tilde{F}$ sotto H_0.

In pratica si utilizza la seguente statistica:

$$F = \frac{n-p}{p-q}\tilde{F} = \frac{(\mathbf{R}^{*\prime}\mathbf{R}^{*} - \mathbf{R}'\mathbf{R})/(p-q)}{\mathbf{R}'\mathbf{R}/(n-p)}.$$

Si può dimostrare che, sotto H_0,

- $(\mathbf{R}^{*\prime}\mathbf{R}^{*} - \mathbf{R}'\mathbf{R})/\sigma^2 \sim \chi^2_{p-q}$ $\qquad (n-q)-(n-p)=p$;
- $\mathbf{R}'\mathbf{R}/\sigma^2 \sim \chi^2_{n-p}$;
- le due v.c. sono indipendenti;

e quindi $F \sim \mathcal{F}_{p-q,n-p}$. Ad un livello di significatività α la regione di rifiuto sarà:

$$\mathcal{R} = \{f \geq f_{p-q,n-p;1-\alpha}\},$$

dove $f_{p-q,n-p;1-\alpha}$ rappresenta il quantile di ordine $1-\alpha$ della $\mathcal{F}_{p-q,n-p}$.

Sia f_{oss} il valore osservato sul campione della statistica F. Definiremo allora il *p-value di F* come

$$\alpha_{oss} = \Pr(F \geq f_{oss}; H_0)$$

e, come di consueto, rifiuteremo H_0 quando $\alpha_{oss} \leq \alpha$.

Un caso speciale è dato da (supponendo che la variabile x_1 sia una costante)

$$\begin{aligned} H_0: &\quad \beta_2 = \cdots = \beta_p = 0, \\ H_1: &\quad \textit{almeno uno di questi coefficienti è diverso da 0.} \end{aligned}$$

Il modello M_0 è $Y_i = \beta_1^* + \varepsilon_i^*$ e abbiamo

$$F = \frac{(\mathbf{R}^{*\prime}\mathbf{R}^* - \mathbf{R}'\mathbf{R})/(p-1)}{\mathbf{R}'\mathbf{R}/(n-p)}.$$

e quindi

$$r_i^* = y_i - \widehat{y}_i^* = y_i - \widehat{\beta^*}_1 = y_i - \overline{y}$$

da cui

$$\begin{aligned} F &= \frac{(DEV_{tot} - DEV_{res})/(p-1)}{DEV_{res}/(n-p)} \\ &= \frac{DEV_{reg}/(p-1)}{DEV_{res}/(n-p)} \\ &= \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \end{aligned}$$

Esempio

Supponiamo di voler verificare che uno qualsiasi dei regressori `Tax, Dlic, Income, logMiles` contribuisce significativamente a spiegare la variabilità di `Fuel` ovvero specifichiamo $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ contro l'alternativa $H_1 : \overline{H}_0$
Ecco un esempio di codice in R

```
> fit <- lm(Fuel ~ Tax + Dlic + Income + logMiles, data = fuel2001)
```

calcoliamo ora la statistica test

```
> DEV.tot <- sum((fuel2001$Fuel - mean(fuel2001$Fuel))^2)
> DEV.res <- sum(fit$res^2)
> DEV.reg <- DEV.tot - DEV.res
> test.F <- (DEV.reg/(length(fit$coeff) - 1))/(DEV.res/fit$df.residual)
> test.F
```

```
[1] 11.99242
```

```
> p.value <- 1 - pf(test.F, length(fit$coeff) - 1, fit$df.residual)
> p.value
```

```
[1] 9.33078e-07
```

Sapreste riconoscere questi valori in

```
Call:
lm(formula = Fuel ~ Tax + Dlic + Income + logMiles, data = fuel2001)

Residuals:
     Min       1Q   Median       3Q      Max
-163.145  -33.039    5.895   31.989  183.499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.1928   194.9062   0.791 0.432938
Tax          -4.2280     2.0301  -2.083 0.042873 *
Dlic          0.4719     0.1285   3.672 0.000626 ***
Income       -6.1353     2.1936  -2.797 0.007508 **
```

```
logMiles     18.5453     6.4722   2.865 0.006259 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64.89 on 46 degrees of freedom
Multiple R-Squared: 0.5105,        Adjusted R-squared: 0.4679
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.33e-07
```

Altra possibilità

```
> fit <- lm(Fuel ~ Tax + Dlic + Income + logMiles, data = fuel2001)
> fit0 <- lm(Fuel ~ 1, data = fuel2001)
> anova(fit0, fit)
```

```
Analysis of Variance Table

Model 1: Fuel ~ 1
Model 2: Fuel ~ Tax + Dlic + Income + logMiles
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     50 395694
2     46 193700  4    201994 11.992 9.33e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$f_{oss} = \frac{201994/4}{193700/46} = 11.99$$

Scelta dei regressori

Supponiamo di dover porre in relazione una variabile di interesse, Y, con k potenziali regressori e di disporre di un campione di dimensione $n > k$. Come si decide quali siano i regressori da inserire nel modello lineare? Non esiste un unico criterio di scelta, ma noi qui useremo quello che viene definito il metodo di *backward selection*. Secondo questo criterio di scelta, si adatta in una prima fase il modello più generale che possiamo considerare, cioè quello con k regressori. Supponendo che il modello sia gaussiano (ed accertando questa assunzione attraverso una analisi empirica dei residui) possiamo applicare il test basato sulla t di Student a ciascun coefficiente di regressione, saggiando il sistema di ipotesi

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned}$$

per $j = 1, \ldots, k$. Se esistessero dei regressori i cui coefficienti non risultassero significativamente diversi da 0 ad un fissato livello α, si escluderebbe dal modello la variabile esplicativa per il coefficiente della quale il valore osservato della statistica T determinasse il più grande dei k *p-value calcolati*. Si ristimerebbe il modello con $k-1$ regressori e si ripeterebbe la procedura sino ad ottenere un modello con p variabili esplicative i cui coefficienti risultassero tutti diversi da 0, secondo il test basato sulla t di Student, al livello di significatività α.

Ulteriori aspetti del modello lineare

Sino ad ora abbiamo sempre considerato modelli del tipo

$$Y = \beta_0 + \beta_1 x_1 + \dots \beta_{p-1} x_{p-1} + e,$$

sotto le assunzioni classiche.

Non si deve però pensare che il modello debba essere necessariamete lineare nei regressori. Esso infatti deve essere *lineare solo rispetto ai coefficienti di regressione*. Se, ad esempio, Y fosse una variabile che assume valori positivi e fosse legata ad altre p variabili, pure positive, dalla seguente relazione:

$$Y = \alpha x_1^{\beta_1} x_2^{\beta_2} \dots x_{p-1}^{\beta_{p-1}} u,$$

potremmo sempre considerare la rappresentazione

$$\log(Y) = \beta_0 + \beta_1 \log(x_1) + \dots \beta_{p-1} \log(x_{p-1}) + e,$$

ponendo $\beta_0 = \log(\alpha)$ (con $\alpha > 0$) e $e = \log(u)$, purché in questa formulazione sussistano le ipotesi classiche del modello lineare. In questo caso la variabile risposta sarebbe $Y^* = \log(Y)$ e i regressori sarebero $x_j^* = \log(x_j)$, $1 \leq j \leq p-1$ (oltre al regressore x_0 identicamente uguale a 1). Questo tipo di rappresentazione è utile per alcune forme di eteroschedasticità (cioè in alcuni di quei casi in cui $Var(Y)$ non è costante su tutte le unità statistiche); essa può inoltre essere utilizzata allo scopo di ottenere una variabile risposta approssimativamente gaussiana. Naturalmente non esiste una regola universale che possa garantire l'adeguatezza di una trasformazione logaritmica, la cui adeguatezza deve essere valutata con cautela.

Un altro tipo di modello di notevole diffusione è quello della regressione polinomiale:

$$Y = \beta_0 + \sum_{j=1}^{p-1} \beta_j x^j + e$$

in cui il $j-$esimo regressore non è altro che la potenza $j-esima$ di una singola variabile X.

Naturalmente con questa trattazione non possiamo né intendiamo esaurire la gamma infinita delle trasformazioni che si possono considerare.

Intervalli di confidenza per i coefficienti di regressione

Come sappiamo, nel modello di regressione lineare gaussiano con p regressori, lo stimatore di massima verosimiglianza $\widehat{\boldsymbol{\beta}}$ si distribuisce come una $\mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. Lo stimatore del generico coefficiente di regressione β_i, $\widehat{\beta}_i$, si distribuisce quindi come una v.c. $\mathcal{N}(\beta_i, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ii})$. Lo stimatore di σ^2, S^2, è stocasticamente indipendente da $\widehat{\boldsymbol{\beta}}$ e

$$\frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Da queste condizioni segue quindi che

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} \sim t_{n-p},$$

dove t_{n-p} indica la distribuzione di una variabile casuale t di student con $n-p$ gradi di libertà. Un intervallo di confidenza di livello $1-\alpha,\ 0 \leq \alpha \leq 1$, può essere costruito e i suoi estremi saranno dati da

$$[\beta_i - \sqrt{s^2(\mathbf{X}'\mathbf{X})_{ii}}t_{1-\alpha/2,n-p},\ \beta_i + \sqrt{s^2(\mathbf{X}'\mathbf{X})_{ii}}t_{1-\alpha/2,n-p}]$$

dove $t_{1-\alpha/2,n-p}$ rappresenta il quantile di ordine $1-\alpha/2$ di una v.c. t di Student con $n-p$ gradi di libertà.

Ecco un esempio di codice in R per il calcolo degli intervalli di confidenza di livello $1 - \alpha = 0.95$

```
> XX <- t(X) %*% X
> Xy <- t(X) %*% y
> invXX <- solve(XX)
> beta.hat <- invXX %*% Xy
> y.hat <- X %*% beta.hat
> r <- y - y.hat
> sigma2.hat <- as.numeric((t(r) %*% r)/(length(y) - dim(X)[2]))
> varbeta.hat <- sigma2.hat * invXX
> alpha <- 0.05
> sebeta.hat <- sqrt(diag(varbeta.hat))
> ll <- beta.hat - qt(1 - alpha/2, dim(X)[1] - dim(X)[2]) * sebeta.hat
> ul <- beta.hat + qt(1 - alpha/2, dim(X)[1] - dim(X)[2]) * sebeta.hat
> cbind(ll, ul)
```

```
                    [,1]        [,2]
(Intercept) -238.1329083 546.5185975
Tax           -8.3144050  -0.1415614
Dlic           0.2131871   0.7305553
Income       -10.5508863  -1.7197756
logMiles       5.5174630  31.5730860
```

Sapreste ricavare questi valori a partire da

```
Call:
lm(formula = Fuel ~ Tax + Dlic + Income + logMiles, data = fuel2001,
    x = TRUE, y = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-163.145  -33.039    5.895   31.989  183.499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.1928   194.9062   0.791 0.432938
Tax          -4.2280     2.0301  -2.083 0.042873 *
Dlic          0.4719     0.1285   3.672 0.000626 ***
Income       -6.1353     2.1936  -2.797 0.007508 **
logMiles     18.5453     6.4722   2.865 0.006259 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64.89 on 46 degrees of freedom
Multiple R-Squared: 0.5105,       Adjusted R-squared: 0.4679
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.33e-07
```

Unità C

Diete

Il problema

Supponiamo di osservare le realizzazione della variabile Y su un campione causale di dimensione n e che le osservazioni campionarie siano suddivise in p gruppi, ciascuno di numerosità $n_j,\ 0 \leq n_j \leq n,\ 1 \leq j \leq p$, in modo tale che $\sum_{j=1}^{p} n_j = n$.

Supponiamo inoltre che le n variabili casuali campionarie siano stocasticamente indipendenti e che le osservazioni appartenenti al $j-$esimo gruppo provengano da una $N(\beta_j, \sigma^2)$, con σ^2 costante sui p gruppi. Ciascun gruppo, quindi, si distingue dagli altri per il valore assunto dalla media della variabile aleatoria normale da cui è stato generato.

In altri termini, possiamo pensare che ciascuna unità statistica sia stata sottoposta ad uno ed un solo trattamento e che il numero dei trattamenti effettuati sia uguale a p. In modo analogo, possiamo anche assumere che tutte le unità statistiche siano state sottoposte ad un solo trattamento con p modalità diverse, dette *livelli*, e che ciascuna unità statistica sia stata trattata secondo una ed una sola modalità del trattamento.

Ciò che interessa, da un punto di vista statistico, è accertare se i p livelli del trattamento influenzino effettivamente il valore atteso di Y in altre parole se i valori β_j sono in realtà uguali tra loro.

Analisi della varianza ad un criterio e a più livelli

Specifichiamo ora un modello di regressione multipla utile allo scopo. Per questo definiamo p variabili indicatrici, $x_1, \ldots, x_j, \ldots, x_p$, definite come

$$x_{i,j} = \begin{cases} 1 & \text{se l'unità } i \text{ appartiene al gruppo } j \\ 0 & \text{altrimenti} \end{cases}$$

(si osservi che queste variabili sono qualitative, nel senso che si limitano ad indicare l'appartenenza o meno ad un fissato gruppo: non si tratta di misurazioni di un carattere quantitativo!).

$$Y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i,$$

con $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ tra loro indipendenti. In termini matriciali si ha

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Quali saranno le stime di β_j? Queste coincideranno con le medie di $\overline{y}_j$ di tutte le osservazioni all' interno del gruppo $j \Rightarrow \widehat{\beta}_j = \overline{y}_j$

Se non ci fosse effetto dovremmo avere

$$\beta_j = \mu, \qquad j = 1, \ldots, p$$

e il sistema d'ipotesi è del tipo

$$H_0 \; : \; \beta_j = \mu, \qquad j = 1, \ldots, p \tag{C.1}$$

$$H_1 \; : \; \text{esiste almeno un } j \text{ tale che } \beta_j \neq \mu. \tag{C.2}$$

L'ipotesi nulla, H_0, rappresenta un'ipotesi di inefficacia del trattamento, in base alla quale tutte le n variabili casuali campionarie hanno media uguale a μ, a prescindere dal gruppo di appartenenza delle unità statistiche. L'ipotesi alternativa, H_1, equivale quindi ad un'ipotesi di efficacia del trattamento, essendo la media di Y diversa da μ in almeno un gruppo!

Una diversa formulazione del modello

Scegliamo ora una diversa parametrizzazione del modello ovvero modifichiamo la struttura, conservando però il significato originario. Poniamo

$$\beta_j = \mu + \alpha_j.$$

Il sistema d'ipotesi precedente diventa

$$\begin{aligned} H_0 &: \alpha_j = 0, \qquad j = 1, \ldots, p \\ H_1 &: \text{esiste almeno un } j \text{ tale che } \alpha_j \neq 0. \end{aligned}$$

Attenzione: siamo passati da p parametri a $p+1$ parametri e questo non è senza conseguenze. Infatti

$$Y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i,$$

$$Y_i = \mu(x_{i,1} + \cdots + x_{i,p}) + \alpha_1 x_{i,1} + \cdots + \cdots + \alpha_p x_{i,p} + \varepsilon_i.$$

Poiché ogni unità statistica appartiene ad uno e ad un solo gruppo si ha

$$x_{i,1} + \cdots + x_{i,p} = 1$$

e quindi

$$Y_i = \mu + \alpha_1 x_{i,1} + \cdots + \cdots + \alpha_p x_{i,p} + \varepsilon_i$$

Questo modello può essere scritto in forma matriciale come

$$\mathbf{Y} = \mathbf{X}^* \boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{C.3}$$

dove $\mathbf{X}^* = [\mathbf{1}_n, \mathbf{X}_1, \cdots, \mathbf{X}_p] = [\mathbf{1}_n, \mathbf{X}]$ è una matrice di dimensioni $n \times (p+1)$, $\boldsymbol{\gamma} = [\mu, \alpha_1, \ldots, \alpha_p]' \in \mathbb{R}^{p+1}$ e il simbolo $\mathbf{1}_n$ rappresenta il vettore in $\mathbb{R}^n$ i cui elementi sono identicamente uguali a 1.

La relazione $x_{i,1} + \cdots + x_{i,p} = 1$ mostra che si ha l'uguaglianza $\mathbf{1}_n - \sum_{j=0}^{p-1} \mathbf{X}_j = \mathbf{0}$. ne consegue che la matrice $\mathbf{X}^*$ non potrà mai avere rango pieno ($rango(\mathbf{X}^*) < p+1$).

Il modello di regressione costruito nella (C.3), nel quale compaiono $p+1$ regressori, non rispetta le ipotesi classiche del modello di regressione lineare multipla: $rango(\mathbf{X}^*) < p+1$.

Una formulazione del modello stimabile

Poniamo $\beta_j = \mu + \alpha_j$ ma con $\alpha_1 = 0$.

In questa formulazione β_1 rappresenta il valore atteso di Y nel gruppo 1; il termine α_j rappresenta la differenza tra $\mathbb{E}(Y)$ nel gruppo j e $\mathbb{E}(Y)$ nel gruppo 1. Il gruppo 1 si dice *gruppo di controllo*, essendo il termine di paragone rispetto al quale si confrontano i valori attesi di Y negli altri gruppi.

Il modello

$$Y_i = \mu + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} \cdots + \cdots + \alpha_p x_{i,p} + \varepsilon_i$$

può essere scritto come:

$$Y_i = \mu + \alpha_2 x_{i,2} + \cdots + \alpha_p x_{i,p} + \varepsilon_i,$$

e in forma matriciale come:

$$\mathbf{Y} = \mathbf{X}^{**}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{C.4}$$

dove $\mathbf{X}^{**} = [\mathbf{1}_n, \mathbf{X}_2, \cdots, \mathbf{X}_p]$ e $\boldsymbol{\delta} \in \mathbb{R}^p$. È facile verificare che $\mathbf{X}^{**}$ ha rango p e sono quindi rispettate tutte le ipotesi del modello di regressione lineare multipla.

Confrontiamo ora i due modelli sotto H_0 e H_1

$$\begin{aligned} M_0 : Y_i &= \mu + \varepsilon_i^* \\ M_1 : Y_i &= \mu + \alpha_2 x_{i,2} \cdots + \cdots + \alpha_p x_{i,p} + \varepsilon_i \end{aligned}$$

A questo punto, per saggiare il sistema di ipotesi

$$\begin{aligned} H_0 &: \alpha_j = 0, \qquad j = 2, \ldots, p \\ H_1 &: \text{esiste almeno un } j \text{ tale che } \alpha_j \neq 0. \end{aligned}$$

ad un fissato livello α, possiamo utilizzare il test basato sulla F di Snedecor visto nell'unità precedente utilizzando la statistica test:

$$F = \frac{(\mathbf{R}^{*\prime}\mathbf{R}^* - \mathbf{R}'\mathbf{R})/(p-1)}{\mathbf{R}'\mathbf{R}/(n-p)}$$

che sotto H_0 si distribuirà come una $F_{p-1,n-p}$.

Si osservi che la stima dei minimi quadrati di μ nel modello M_0 è dato da $\overline{y}$. Il modello M_1

$$Y_i = \mu + \alpha_2 x_{i,2} \cdots + \cdots + \alpha_p x_{i,p} + \varepsilon_i$$

è equivalente a

$$Y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i,$$

con $\beta_j = \mu + \alpha_j$ e $\alpha_1 = 0$. Per cui

$$\overline{y}_1 = \widehat{\beta}_1 = \widehat{\mu}$$

$$\overline{y}_j = \widehat{\beta}_j = \widehat{\mu} + \widehat{\alpha}_j \Rightarrow \widehat{\alpha}_j = \overline{y}_j - \overline{y}_1,\ j = 2, \ldots, p$$

Esempio

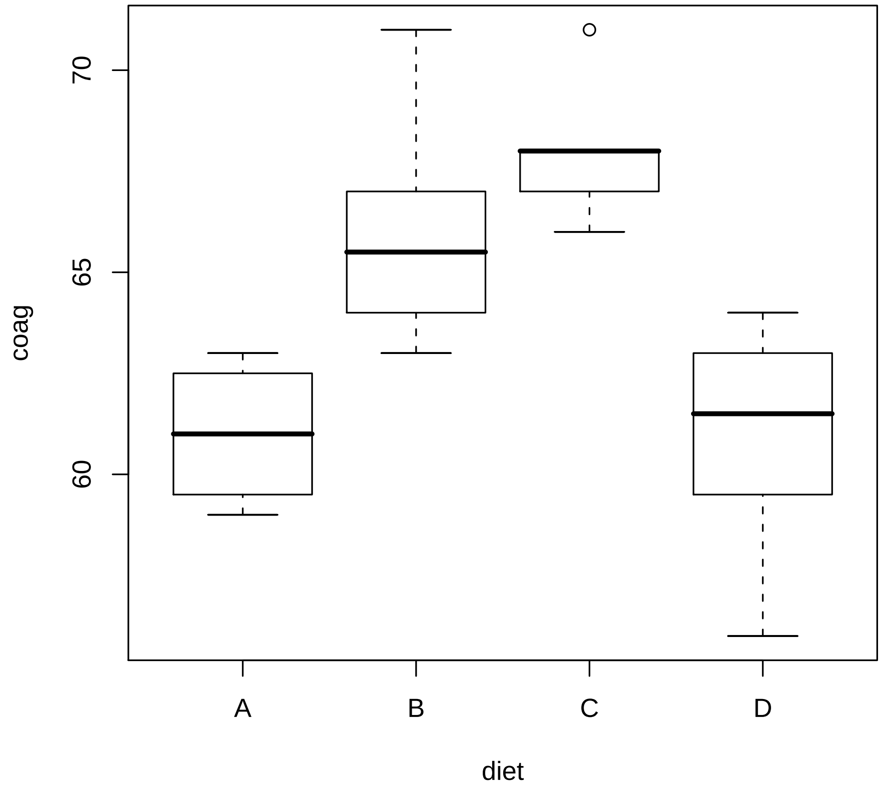
Un insieme di 24 animali viene casualmente diviso in 4 gruppi ed a ogni gruppo viene somministrata una particolare dieta (A,B, C, D). Quindi ad ogni animale viene misurato il tempo di coagulazione del sangue [1].

```
> library(faraway)
> data(coagulation)
> coagulation[1:13, ]

   coag diet
1    62    A
2    60    A
3    63    A
4    59    A
5    63    B
6    67    B
7    71    B
8    64    B
9    65    B
10   66    B
11   68    C
12   66    C
13   71    C
```

Ecco i diagrammi a scatola per ogni gruppo

[1] Dati tratti dalla libreria `faraway` Version: 1.0.3

```
> plot(coag ~ diet, data = coagulation)
```



che rivelano come la dieta C abbia un comportamento particolare.

Proviamo ora stimare un modello per l'analisi della varianza

```
> g <- lm(coag ~ diet, data = coagulation)
> summary(g)
```

```
Call:
lm(formula = coag ~ diet, data = coagulation)

Residuals:
       Min         1Q     Median         3Q        Max
-5.000e+00 -1.250e+00 -2.498e-16  1.250e+00  5.000e+00

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 6.100e+01  1.183e+00   51.554  < 2e-16 ***
dietB       5.000e+00  1.528e+00    3.273 0.003803 **
dietC       7.000e+00  1.528e+00    4.583 0.000181 ***
dietD       2.991e-15  1.449e+00 2.06e-15 1.000000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-Squared: 0.6706,       Adjusted R-squared: 0.6212
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

Poiché la statistica F presenta un livello di significatività quasi pari a zero, concludiamo che vi è differenza nell'effetto delle diete. I coefficienti stimati si interpretano come effetti addizionali rispetto alla dieta A.

Per ottenere le stime delle medie di gruppo

```
> g <- lm(coag ~ diet - 1, data = coagulation)
> summary(g)
```

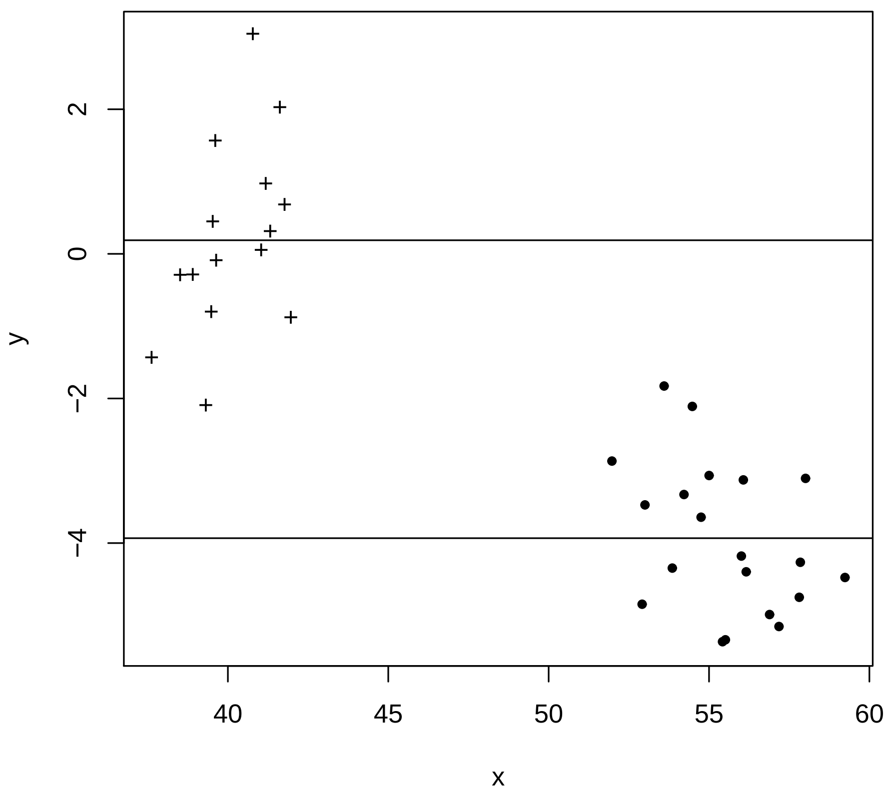
```
Call:
lm(formula = coag ~ diet - 1, data = coagulation)

Residuals:
       Min         1Q     Median         3Q        Max
-5.000e+00 -1.250e+00  1.110e-16  1.250e+00  5.000e+00

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
dietA  61.0000     1.1832   51.55   <2e-16 ***
dietB  66.0000     0.9661   68.32   <2e-16 ***
dietC  68.0000     0.9661   70.39   <2e-16 ***
dietD  61.0000     0.8367   72.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-Squared: 0.9989,       Adjusted R-squared: 0.9986
F-statistic:  4399 on 4 and 20 DF,  p-value: < 2.2e-16
```

Unità D

Cattedrali inglesi

Esempio simulato

Si considerino 35 pazienti di diversa età, a 20 (•) è stato somministrato un farmaco, a 15 è stato somministrato un placebo (+). Per ogni paziente è stata rilevata la differenza di pressione tra prima e dopo la somministrazione, e l'età.



Le medie della differenza di pressione per i due gruppi (0.19, -3.93) sembrano indicare una sensibile differenza tra i due gruppi. Tuttavia questa differenza può essere imputata al'età.

Analisi della covarianza

Supponiamo di considerare un modello di regressione lineare semplice:

$$M_0: \qquad Y_i = \mu + \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \tag{D.1}$$

Complichiamo ora la situazione introducendo l'ulteriore assunzione che le n unità statistiche siano suddivise in k gruppi, ciascuno dei quali caratterizzato da uno specifico livello di un fissato trattamento. Come nel caso dell'analisi della varianza, assumeremo che ciascuna unità statistica possa essere sottoposta al trattamento secondo uno ed un solo livello. Ci interroghiamo circa l'efficacia del trattamento, e proprio in questa domanda stanno le complicazioni! Infatti il trattamento potrebbe determinare semplicemente una variazione dell'intercetta della retta di regressione implicitamente definita dalla (D.1), cioè potrebbe dar luogo al modello con $k+1$ regressori:

$$M_1: \qquad Y_i = \mu + \sum_{j=1}^{k-1} \alpha_j z_{i,j} + \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \tag{D.2}$$

In questa equazione le quantità $z_{i,j}$ sono valori di variabili indicatrici X_j definite come nell'analisi della varianza e il gruppo di controllo sarà costituito dal gruppo 1.

Per cui se l'unità i appartiene ad un gruppo 1

$$Y_i = \mu + \beta x_i + \varepsilon_i$$

oppure se l'unità i appartiene ad un gruppo $j \neq 1$

$$Y_i = \mu + \alpha_j + \beta x_i + \varepsilon_i$$

Ciò equivale a dire che sul gruppo di controllo il modello per le variabili casuali Y_i è M_0, mentre negli altri gruppi, se il trattamento fosse efficace, il modello è M_1. In altri termini i coefficienti α_j, $j = 1, \ldots, k-1$, rappresentano le variazioni dell'intercetta della retta di regressione nei $k-1$ gruppi diversi da quello di controllo.

Saggiare l'efficacia del trattamento ad un livello di significatività α equivale quindi a confrontare il modello M_1 con il modello M_0 attraverso il sistema di ipotesi:

$$\begin{aligned} H_0 &: \ \alpha_j = 0, j = 1, \ldots, k-1 \\ H_1 &: \ \text{esiste } j : \alpha_j \neq 0. \end{aligned}$$

L'efficacia del trattamento potrebbe però estrinsecarsi anche in variazioni del coefficiente angolare della retta (β). In tal caso dovremmo definire un modello più generale:

$$M_1 : \qquad Y_i = \mu + \beta x_i + \sum_{j=1}^{k-1} \beta_j z_{i,j} x_i + \varepsilon_i$$

Per cui se l'unità i appartiene ad un gruppo 1

$$M_0 : Y_i = \mu + \beta x_i + \varepsilon_i$$

oppure se l'unità i appartiene ad un gruppo $j \neq 1$

$$M_1 : Y_i = \mu + (\beta + \beta_j) x_i + \varepsilon_i$$

ovvero i coefficienti β_j, $j = 1, \ldots, k-1$, rappresentano le variazioni della pendenza della retta di regressione nei $k-1$ gruppi diversi da quello di controllo.

Infine un modello generale contempla variazioni sia nell'intercetta che nel coefficiente angolare

$$M_1 : Y_i = \mu + \sum_{j=1}^{k-1} \alpha_j z_{i,j} + \beta x_i + \sum_{j=1}^{k-1} \beta_j z_{i,j} x_i + \varepsilon_i$$

Per cui se l'unità i appartiene ad un gruppo 1

$$M_0 : Y_i = \mu + \beta x_i + \varepsilon_i$$

oppure se l'unità i appartiene ad un gruppo $j \neq 1$

$$M_1 : Y_i = \mu + \alpha_j + (\beta + \beta_j) x_i + \varepsilon_i$$

Stimando i modelli M_0 e M_1, si otterrà ancora una statistica F definita come:

$$F = \frac{(\mathbf{R}^{*\prime}\mathbf{R}^{*} - \mathbf{R}^{\prime}\mathbf{R})/(p-q)}{\mathbf{R}^{\prime}\mathbf{R}/(n-p)}.$$

dove q e p indicano rispettivamente il numero di parametri di M_0 e M_1, e $\mathbf{R}^{*\prime}\mathbf{R}^{*}$ e $\mathbf{R}^{\prime}\mathbf{R}$ rappresentano la somma dei quadrati dei residui di M_0 e M_1. Sotto H_0 ancora una volta $F \sim F_{p-q,n-p}$.

Esempio

I dati che esaminiamo si riferiscono a 25 cattedrali inglesi construite nel Medioevo. Per ognuna di esse viene rilevata l'altezza della navata (`x`) e la lunghezza totale (`y`) misurate in piedi. Alcune sono in stile romanico (`style='r'`) le altre sono in stile gotico (`style='g'`) [1].

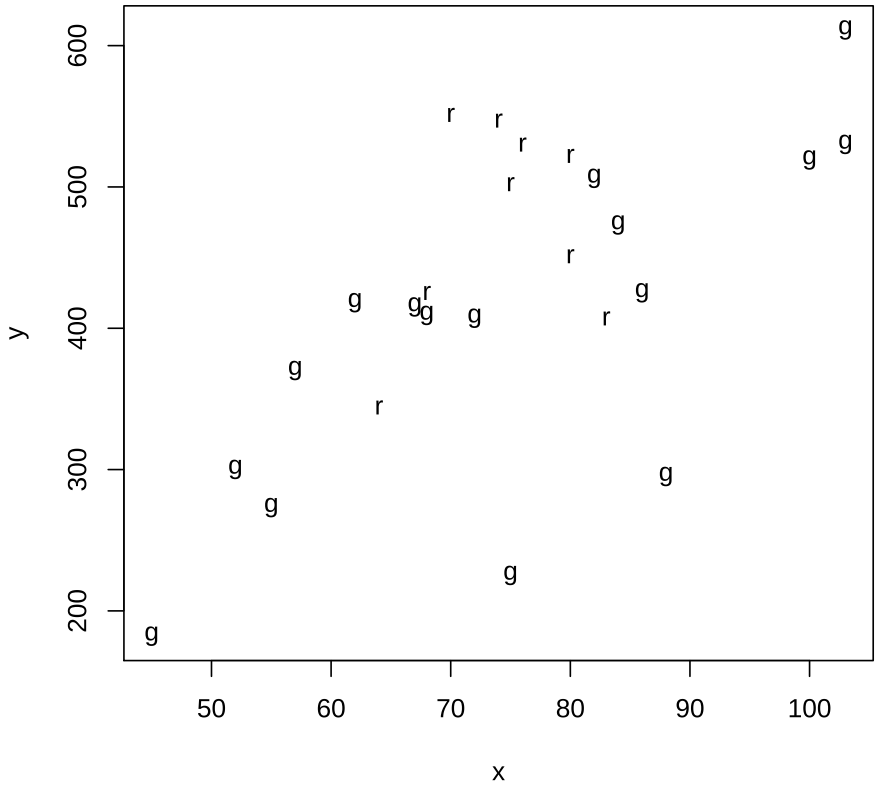
```
> library(faraway)
> data(cathedral)
> cathedral
```

```
              style   x   y
Durham            r  75 502
Canterbury        r  80 522
Gloucester        r  68 425
Hereford          r  64 344
Norwich           r  83 407
Peterborough      r  80 451
St.Albans         r  70 551
Winchester        r  76 530
Ely               r  74 547
York              g 100 519
Bath              g  75 225
Bristol           g  52 300
Chichester        g  62 418
Exeter            g  68 409
GloucesterG       g  86 425
Lichfield         g  57 370
Lincoln           g  82 506
NorwichG          g  72 407
Ripon             g  88 295
Southwark         g  55 273
Wells             g  67 415
St.Asaph          g  45 182
WinchesterG       g 103 530
Old.St.Paul       g 103 611
Salisbury         g  84 473
```

Ecco i diagrammi di dispersione per ogni stile

```
> attach(cathedral)
> plot(y ~ x, data = cathedral, type = "n")
> points(x[style == "g"], y[style == "g"], pch = "g")
> points(x[style == "r"], y[style == "r"], pch = "r")
```

[1] Dati tratti dalla libreria `faraway` Version: 1.0.3



Proviamo ora stimare un modello di regressione semplice per ciascun stile

```
> fit <- lm(y ~ x * style, data = cathedral)
> summary(fit)
```

```
Call:
lm(formula = y ~ x * style, data = cathedral)

Residuals:
    Min      1Q  Median      3Q     Max
-172.68  -30.22   23.75   55.78   89.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   37.111     85.675   0.433 0.669317
x              4.808      1.112   4.322 0.000301 ***
styler       204.722    347.207   0.590 0.561733
x:styler      -1.669      4.641  -0.360 0.722657
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.11 on 21 degrees of freedom
Multiple R-Squared: 0.5412,       Adjusted R-squared: 0.4757
F-statistic: 8.257 on 3 and 21 DF,  p-value: 0.0008072
```

Come codifica R la matrice $\mathbf{X}$?

```
> model.matrix(fit)
```

```
             (Intercept)   x styler x:styler
Durham                 1  75      1       75
Canterbury             1  80      1       80
Gloucester             1  68      1       68
Hereford               1  64      1       64
Norwich                1  83      1       83
Peterborough           1  80      1       80
St.Albans              1  70      1       70
Winchester             1  76      1       76
Ely                    1  74      1       74
York                   1 100      0        0
Bath                   1  75      0        0
Bristol                1  52      0        0
Chichester             1  62      0        0
Exeter                 1  68      0        0
GloucesterG            1  86      0        0
Lichfield              1  57      0        0
Lincoln                1  82      0        0
NorwichG               1  72      0        0
Ripon                  1  88      0        0
Southwark              1  55      0        0
Wells                  1  67      0        0
St.Asaph               1  45      0        0
WinchesterG            1 103      0        0
Old.St.Paul            1 103      0        0
Salisbury              1  84      0        0
attr(,"assign")
[1] 0 1 2 3
attr(,"contrasts")
attr(,"contrasts")$style
[1] "contr.treatment"
```

Osserviamo come il termine di interazione (`x:styler`) non è significativamente differente da zero

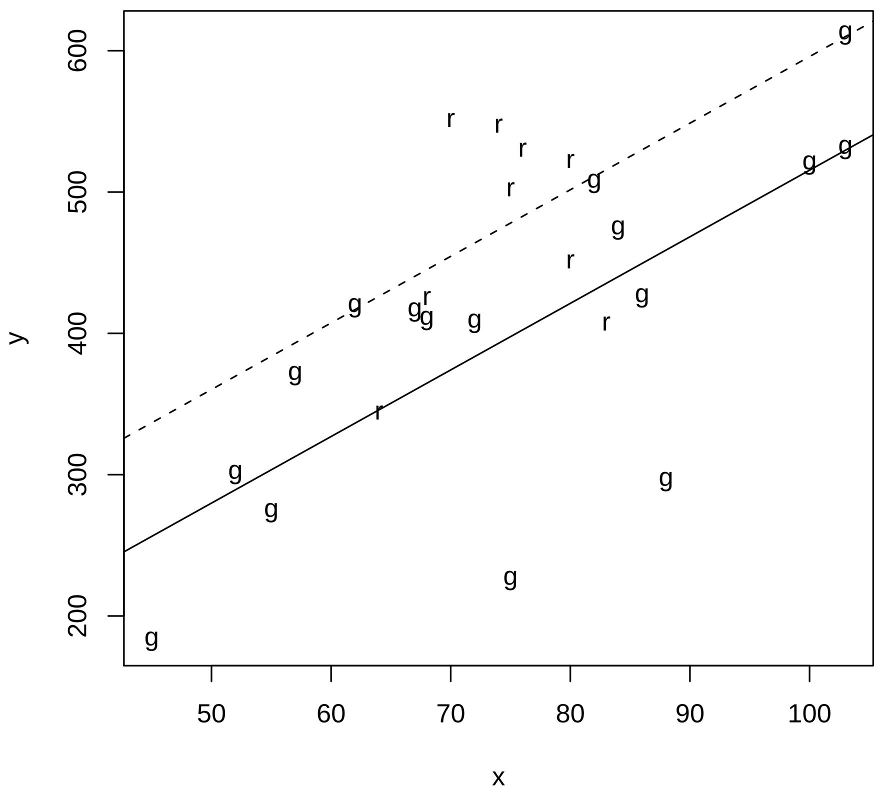
```
> fit2 <- lm(y ~ x + style, data = cathedral)
> summary(fit2)
```

```
Call:
lm(formula = y ~ x + style, data = cathedral)

Residuals:
    Min      1Q  Median      3Q     Max
-172.67  -30.44   20.38   55.02   96.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.298     81.648   0.543   0.5929
x              4.712      1.058   4.452   0.0002 ***
```

```
styler        80.393     32.306   2.488   0.0209 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.53 on 22 degrees of freedom
Multiple R-Squared: 0.5384,        Adjusted R-squared: 0.4964
F-statistic: 12.83 on 2 and 22 DF,  p-value: 0.0002028
```

Il valore 80.39 rappresenta la differenza tra le due intercette. Disegnamo ora il grafico delle due rette stimate.

```
> plot(y ~ x, data = cathedral, type = "n")
> points(x[style == "g"], y[style == "g"], pch = "g")
> points(x[style == "r"], y[style == "r"], pch = "r")
> abline(fit2$coeff[1], fit2$coeff[2])
> abline(fit2$coeff[1] + fit2$coeff[3], fit2$coeff[2], lty = 2)
```



Unità E

Risparmi

I dati

I dati [1] qui sotto riportati si riferiscono a 50 paesi. Le variabili considerate, intese come valori medi nel periodo 1960-1970, sono:

- `dpi` il reddito disponibile pro-capite in dollari;
- `ddpi` il tasso di variazione percentuale del reddito disponibile pro capite:
- `sr` il risparmio disponibile diviso per il reddito disponibile;
- `pop15` e `pop75` sono rispettivamente la percentuale della popolazione al di sotto dei 15 anni e al di sopra dei 75.

```
> library(faraway)
> data(savings)
> savings[1:10, ]
```

```
              sr pop15 pop75     dpi ddpi
Australia  11.43 29.35  2.87 2329.68 2.87
Austria    12.07 23.32  4.41 1507.99 3.93
Belgium    13.17 23.80  4.43 2108.47 3.82
Bolivia     5.75 41.89  1.67  189.13 0.22
Brazil     12.88 42.19  0.83  728.47 4.56
Canada      8.79 31.72  2.85 2982.88 2.43
Chile       0.60 39.74  1.34  662.86 2.67
China      11.90 44.75  0.67  289.52 6.51
Colombia    4.98 46.64  1.06  276.65 3.08
Costa Rica 10.78 47.64  1.14  471.24 2.80
```

[1] Vedi Belsley, D. A., E. Kuh, and R. E. Welsch (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley e la libreria `faraway` Version: 1.0.3

Consideriamo un primo modello con tutti i regressori

```
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
> summary(g)
```

```
Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-Squared: 0.3385,       Adjusted R-squared: 0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Alcune regressori potrebbero non essere considerati ad esempio `dpi`

```
> g2 <- lm(sr ~ pop15 + pop75 + ddpi, data = savings)
> summary(g2)
```

```
Call:
lm(formula = sr ~ pop15 + pop75 + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2539 -2.6159 -0.3913  2.3344  9.7070

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.1247     7.1838   3.915 0.000297 ***
pop15        -0.4518     0.1409  -3.206 0.002452 **
pop75        -1.8354     0.9984  -1.838 0.072473 .
ddpi          0.4278     0.1879   2.277 0.027478 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.767 on 46 degrees of freedom
Multiple R-Squared: 0.3365,       Adjusted R-squared: 0.2933
F-statistic: 7.778 on 3 and 46 DF,  p-value: 0.0002646
```

```
> test <- (sum(g2$res^2) - sum(g$res^2))/(sum(g$res^2)/g$df.residual)
> 1 - pf(test, 1, g$df.residual)
```

```
[1] 0.7191732
```

Se troviamo due variabili X_j e X_k per cui il livello di significatività significatività osservat non è elevato, questo significa che ambedue X_j e X_k possono essere eliminate dal modello? Non necessariamente in quanto l'eliminazione di una variabile ad esempio X_j porta ad una ristima del modello per la quale la variabile X_k potrebbe divenire significativa. Se si vuole realmente verificare la significatività di X_j e X_k, si dovrebbe stimare un modello M_1 con le due variabili e un modello M_0 senza di essere e utilizzare un test F

```
> g3 <- lm(sr ~ pop15 + ddpi, data = savings)
> summary(g3)
```

```
Call:
lm(formula = sr ~ pop15 + ddpi, data = savings)

Residuals:
     Min       1Q   Median       3Q      Max
-7.58314 -2.86323  0.04535  2.22734 10.47530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.59958    2.33439   6.682 2.48e-08 ***
pop15       -0.21638    0.06033  -3.586 0.000796 ***
ddpi         0.44283    0.19240   2.302 0.025837 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.861 on 47 degrees of freedom
Multiple R-Squared: 0.2878,       Adjusted R-squared: 0.2575
F-statistic: 9.496 on 2 and 47 DF,  p-value: 0.0003438
```

```
> test <- ((sum(g3$res^2) - sum(g$res^2))/2)/(sum(g$res^2)/g$df.residual)
> 1 - pf(test, 2, g$df.residual)
```

```
[1] 0.1900451
```

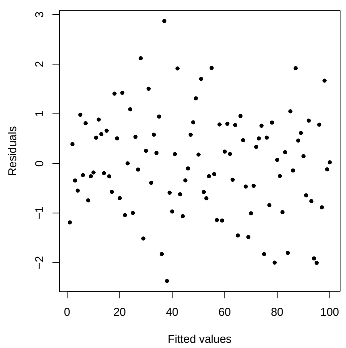
Quale conclusione si può trarre ?

Unità F

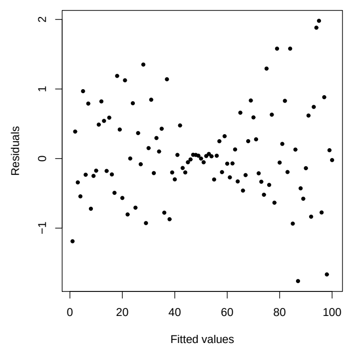
Analisi dei residui

Diagrammi di dispersione

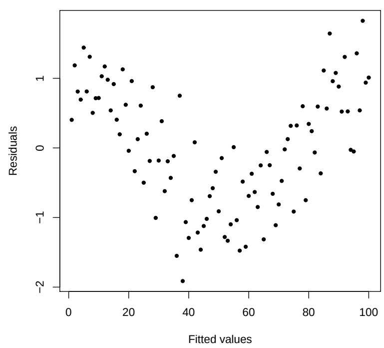
Consideriamo i residui $r_i = y_i - \widehat{y}_i$. Un diagramma di dispersione dei residui rispetto ai valori $\widehat{y}_i$ può rilevare che il modello non è stato correttamente specificato. Le tre situazioni qui sotto riportate si riferiscono ai residui nei casi in cui (a) un modello correttamente specificato, (b) la varianza degli errori era in realtà non costante, (c) esiste qualche forma di non linearità nei regressori.



(a)



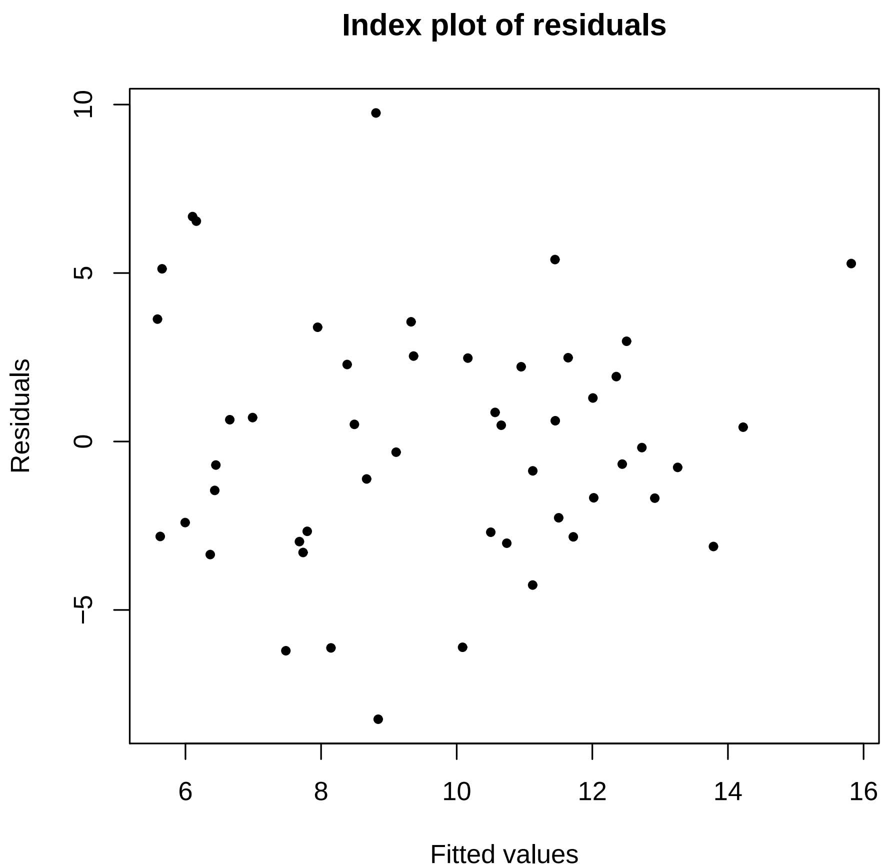
(b)



(c)

Nel caso dei dati sul risparmio abbiamo

```
> plot(g$fitted, g$res, xlab = "Fitted values", ylab = "Residuals",
+     main = "Index plot of residuals", pch = 20)
```



Punti leva

Abbiamo definito i residui come

$$\mathbf{R} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}$$

Posto $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, otteniamo [1]

$$\mathbf{R} = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}$$

e quindi

$$\mathbb{E}(\mathbf{R}) = \mathbf{0}, \text{ e } \mathbb{V}ar(\mathbf{R}) = \sigma^2(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H})' = \sigma^2(\mathbf{I}_n - \mathbf{H}),$$

poichè

$$(\mathbf{I}_n - \mathbf{H})\mathbf{X} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{X} - \mathbf{X}$$

$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$.

Da ciò si deduce che mentre gli errori $\boldsymbol{\varepsilon}$ sono incorrelati e omoschedastici i residui non lo sono. Condideriamo gli elementi sulla diagonale di $\mathbf{H}$, $h_i = (\mathbf{H})_{ii}$, e si ha

$$\mathbb{V}ar(R_i) = \sigma^2(1 - h_i)$$

[1] La matrice $\mathbf{H}$ è pari a $\mathbf{P}$ la matrice di proiezione del vettore $\mathbf{y}$ sullo spazio generato dai vettori colonna che compongono la matrice $\mathbf{X}$.

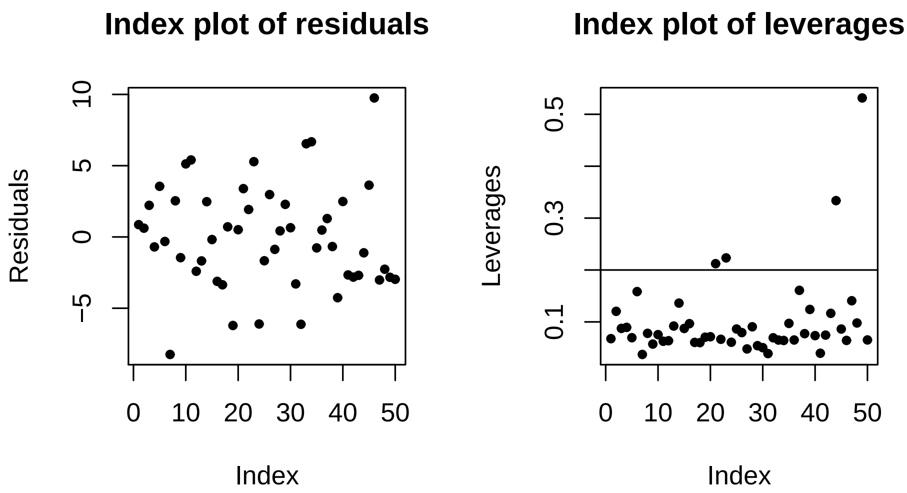
La quantità h_i è detta *leverage*. Tanto più h_i è grande tanto più $\mathbb{V}ar(R_i)$ è piccola. Si può dimostrare che

$$\sum_{i=1}^{n} h_i = \text{tr}(\mathbf{H}) = p, \qquad h_i \geq 1/n.$$

La media aritmentica dei *leverage* vale p/n, per cui una regola pratica porta a considerare attentamente quelle unità statistiche per cui $2p/n$. Valori grandi di h_i corrispondono a valori grandi in $\mathbf{X}$. Identifichiamo ora i punti leverage

```
> library(faraway)
> par(mfrow = c(1, 2), pty = "s")
> data(savings)
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
> plot(g$res, ylab = "Residuals", main = "Index plot of residuals",
+     pch = 20)
> x <- model.matrix(g)
> lev <- hat(x)
> plot(lev, ylab = "Leverages", main = "Index plot of leverages",
+     pch = 20)
> abline(h = 2 * 5/50)
> countries <- row.names(savings)
> names(lev) <- countries
> lev[lev > (2 * 5/50)]
```

```
      Ireland         Japan United States         Libya
    0.2122363     0.2233099     0.3336880     0.5314568
```

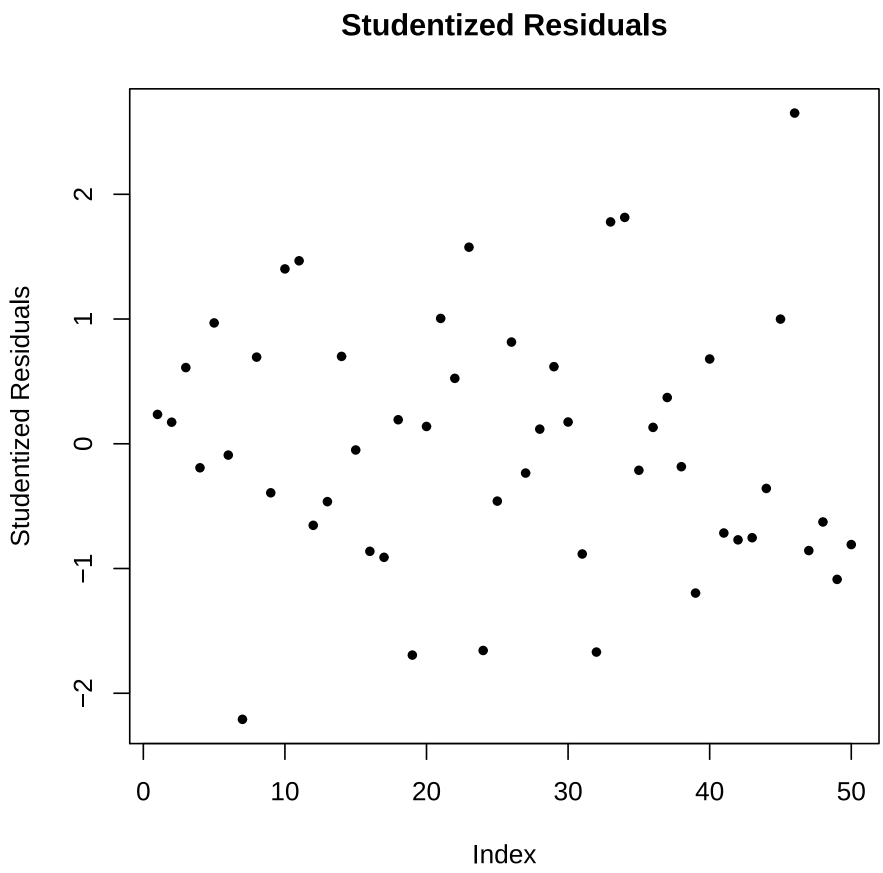


Infine possiamo definire dei residui detti *studentized* come

$$\tilde{r}_i = \frac{r_i}{\widehat{\sigma}\sqrt{1 - h_i}}$$

se il modello è correttamente specificato allora $\mathbb{V}ar(\tilde{R}_i) \simeq 1$. Consideriamo il calcolo di questi

```
> sigma.hat <- sqrt(sum(g$res^2)/g$df.residual)
> stud <- g$res/(sigma.hat * sqrt(1 - lev))
> plot(stud, ylab = "Studentized Residuals", main = "Studentized Residuals",
+     pch = 20)
```



e osserviamo che non esiste nessuna particolare anomalia.

Le distanze di Cook

Sin dall'inizio del corso ci siamo posti il problema di valutare se talune unità statistiche possano essere eccessivamente influenti sulle procedure inferenziali che adottiamo. Se ciò accadesse per una unità statistica sulla quale, ad esempio, fossero commessi degli errori di misurazione, le conseguenze potrebbero essere disastrose. È quindi molto importante valutare se nel campione esistano osservazioni anomale e se esse esercitino una forte influenza sulle stime dei coefficienti di regressione. Un utile strumento a tale proposito è costituito dalle distanze di Cook.

Consideriamo la stima dei coefficienti di regressione che otteniamo applicando il criterio dei minimi quadrati $\widehat{\boldsymbol{\beta}}$ il vettore dei valori previsti essa associato è definito come $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.

Supponiamo ora di escludere la i-esima unità statistica dal campione e definiamo

$\mathbf{y}(i) =$ vettore delle osservazioni su y, esclusa la $i-$esima;

$\mathbf{X}(i) =$ matrice di regressione uguale a $\mathbf{X}$ con l'esclusione però della $i-$esima riga;

$\widehat{\boldsymbol{\beta}}(i) =$ stima dei minimi quadrati ottenuta dal campione di $n - 1$ osservazioni

$\widehat{\mathbf{y}}(i) = \mathbf{X}\widehat{\boldsymbol{\beta}}(i)$ vettore dei valori previsti calcolati su <u>tutte</u> le unità statistiche.

Se la $i-$esima unità fosse molto influente sulla stima di $\boldsymbol{\beta}$, i valori $\widehat{\mathbf{y}}$ ottenuti con l'intero campione sarebbero lontani da quelli calcolati sul campione da cui fosse estromessa la $i-$esima unità statistica. Viceversa, se quell'unità non fosse influente, i due vettori sarebbero molto vicini, nel senso della distanza euclidea, $\|\cdot\|$, tra di loro. Ha senso quindi considerare il quadrato della distanza euclidea tra $\widehat{\mathbf{y}}$ e $\widehat{\mathbf{y}}(i)$ come un indicatore dell'influenza della $i-$esima unità del campione:

$$\tilde{D}_i = \|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}(i)\|^2 = (\widehat{\mathbf{y}} - \widehat{\mathbf{y}}(i))'(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}(i)) = \sum_{j=1}^{n}(\widehat{y}_j - \widehat{y}_j(i))^2.$$

Cook ha proposto una misura dell'influenza di una generica unità statistica strettamente legata a questa:

$$D_i = (\widehat{\mathbf{y}} - \widehat{\mathbf{y}}(i))'(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}(i))/(p\widehat{\sigma}^2), i = 1, \ldots, n$$

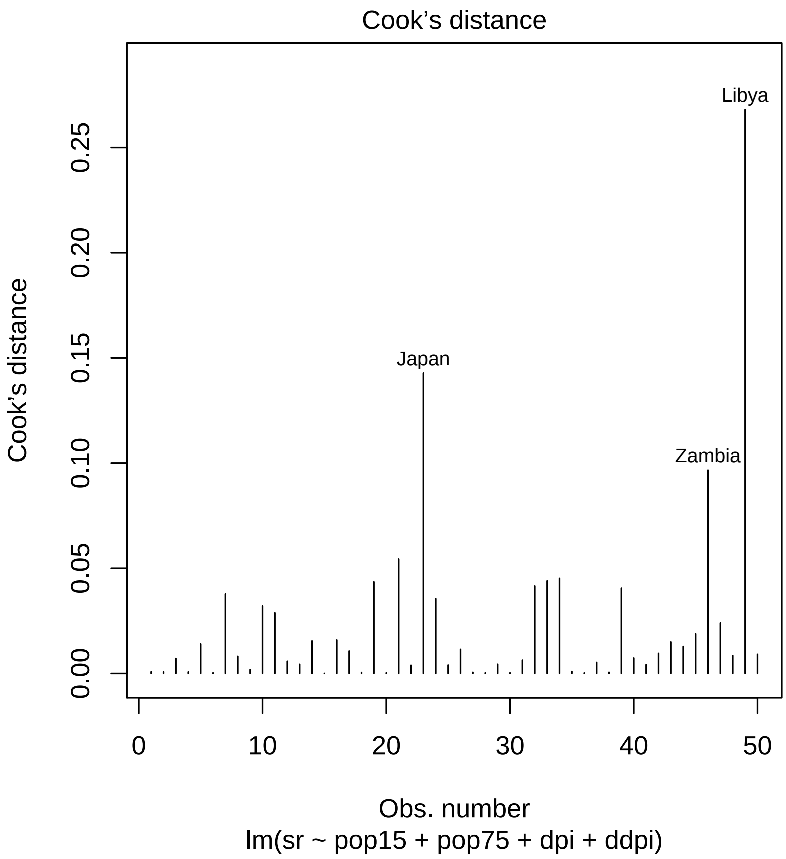
che non è altro che $\tilde{D}_i$ riscalata dal prodotto tra il numero dei regressori e la stima di σ^2 ottenuta dal campione completo. Chiaramente questa misura può essere calcolata per ogni unità statistica del campione. Le unità per le quali D_i risultasse particolarmente elevato devono considerarsi con estrema cautela, in quanto potrebbero distorcere notevolmente le procedure inferenziali.

Due possibili soluzioni:

1. esclusione di quelle unità

2. l'elaborazione di tecniche inferenziali più **robuste** a queste anomalie, ma questo è un tema che esula dagli obiettivi del corso

```
> par(pty = "s")
> plot(g, which = 4)
```



Stimando un modello lineare dopo aver rimosso l'osservazione più influente, otteniamo una variazione del coefficiente di `ddpi` del 50%

```
> cook <- cooks.distance(g)
> gl <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings, subset = (cook <
+     max(cook)))
> summary(gl)
```

```
Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings,
    subset = (cook < max(cook)))

Residuals:
    Min      1Q  Median      3Q     Max
-8.0699 -2.5408 -0.1584  2.0934  9.3732
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.5240460  8.2240263   2.982  0.00465 **
pop15       -0.3914401  0.1579095  -2.479  0.01708 *
pop75       -1.2808669  1.1451821  -1.118  0.26943
dpi         -0.0003189  0.0009293  -0.343  0.73312
ddpi         0.6102790  0.2687784   2.271  0.02812 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.795 on 44 degrees of freedom
Multiple R-Squared: 0.3554,        Adjusted R-squared: 0.2968
F-statistic: 6.065 on 4 and 44 DF,  p-value: 0.0005617
```

Verifica sulla distribuzione normale

Test ed intervalli di confidenza sono basati sull'assunto che $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Se il modello è correttamente specificato allora i residui $\tilde{r}_i$ dovrebbero avere una distribuzione simile a quella di una v.c. normale. Per questo si utilizza il grafico quantile-quantile così costruito

1. ordina i residui $\tilde{r}_{(1)} \leq \tilde{r}_{(2)} \leq \cdots \leq \tilde{r}_{(n)}$;
2. calcola i quantili $u_i = \Phi^{-1}(i/(n+1))$
3. disegna il diagramma di dispersione di $\tilde{r}_{(i)}$ rispetto a u_i

Se i residui sono in accordo con la distribuzione normale allora i punti si dovrebbero disporre lungno una linea retta.

```
> par(pty = "s")
> plot(g, which = 2, pch = 20)
```