Statistica - 17ª lezione

1 giugno 2021

Regressione lineare multipla

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \longrightarrow \boxed{\begin{array}{c} \text{Legge fisica} \\ y = a + b_1 x_1 + \ldots + b_k x_k \\ = a + \mathbf{x}^t \mathbf{b} \\ (a \in \mathbb{R},\ \mathbf{b} \in \mathbb{R}^k) \end{array}} \xrightarrow{Y = a + \mathbf{x}^t\mathbf{b} + E}$$

input (vettore in $\mathbb{R}^k$) — output (variabile aleatoria)

Se facciamo n misure:

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \quad \longrightarrow \quad Y_1, Y_2, \ldots, Y_n \quad \text{con} \quad Y_i = a + \mathbf{x}_i^t \mathbf{b} + E_i$$

Ipotesi fondamentale del modello multilineare:

- $E_1, \ldots, E_n$ indipendenti $\quad \Leftrightarrow \quad Y_1, \ldots, Y_n$ indipendenti
- $E_i \sim N(0, \sigma^2)\ \forall i \quad \Leftrightarrow \quad Y_i \sim N(a + \mathbf{x}_i^t \mathbf{b}\,, \sigma^2)\ \forall i$

Parametri incogniti: $\quad a, \quad \mathbf{b}, \quad \sigma^2$

Stimatori, IC e test per i parametri

IPOTESI: $\begin{cases} Y_1, \ldots, Y_n & \text{indipendenti} \\ Y_i \sim N((1\ \mathbf{x}_i^t)\boldsymbol{\beta}\,,\, \sigma^2) & \forall i \end{cases}$ $\qquad \boldsymbol{\beta}$, σ^2 parametri incogniti

Stimatori, IC e test per i parametri

IPOTESI: $\begin{cases} Y_1, \dots, Y_n \quad \text{indipendenti} \\ Y_i \sim N((1\ \mathbf{x}_i^t)\boldsymbol{\beta}\,,\, \sigma^2) \quad \forall i \end{cases}$ $\qquad \boldsymbol{\beta}$, σ^2 parametri incogniti

CONSEGUENZE:

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\mathbf{y} \quad \xrightarrow[\text{vettore di v.a.}]{\text{diventa il}} \quad \hat{\boldsymbol{B}} = (\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\mathbf{Y} \quad \text{con} \quad \mathbf{Y} = (Y_1, \dots, Y_n)^t$$

IPOTESI: $\begin{cases} Y_1, \ldots, Y_n \quad \text{indipendenti} \\ Y_i \sim N((1\ \mathbf{x}_i^t)\boldsymbol{\beta}\,,\, \sigma^2) \quad \forall i \end{cases}$ $\boldsymbol{\beta}$, σ^2 parametri incogniti

CONSEGUENZE:

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\mathbf{y} \quad \underset{\text{vettore di v.a.}}{\overset{\text{diventa il}}{\longrightarrow}} \quad \hat{\boldsymbol{B}} = (\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\mathbf{Y} \quad \text{con} \quad \mathbf{Y} = (Y_1, \ldots, Y_n)^t$$

$$\hat{B}_r \sim N\left(\beta_r\,,\, \sigma^2\,[(\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}]_{rr}\right) \quad \Rightarrow \quad \hat{B}_r \ \text{stimatore corretto di } \beta_r$$

Stimatori, IC e test per i parametri

IPOTESI: $\begin{cases} Y_1, \ldots, Y_n & \text{indipendenti} \\ Y_i \sim N((1\ \mathbf{x}_i^t)\boldsymbol{\beta}\,,\, \sigma^2) & \forall i \end{cases}$ $\boldsymbol{\beta}$, σ^2 parametri incogniti

CONSEGUENZE:

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\mathbf{y} \quad \xrightarrow[\text{vettore di v.a.}]{\text{diventa il}} \quad \hat{\boldsymbol{B}} = (\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\mathbf{Y} \quad \text{con} \quad \mathbf{Y} = (Y_1, \ldots, Y_n)^t$$

$$\hat{B}_r \sim N\left(\beta_r\,,\, \sigma^2\,[(\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}]_{rr}\right) \quad \Rightarrow \quad \hat{B}_r \ \text{stimatore corretto di } \beta_r$$

$$\frac{SS_E}{\sigma^2} \sim \chi^2(n-k-1) \quad \text{indipendente da } \hat{B}_r$$

$$\Rightarrow \ \hat{\Sigma}^2 = \frac{SS_E}{n-k-1} \qquad \text{stimatore corretto di } \sigma^2$$

Stimatori, IC e test per i parametri

IPOTESI: $\begin{cases} Y_1, \ldots, Y_n \quad \text{indipendenti} \\ Y_i \sim N((1\ \mathbf{x}_i^t)\boldsymbol{\beta}\,,\ \sigma^2) \quad \forall i \end{cases}$ $\qquad \boldsymbol{\beta}$, σ^2 parametri incogniti

CONSEGUENZE:

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\mathbf{y} \quad \xrightarrow[\text{vettore di v.a.}]{\text{diventa il}} \quad \hat{\boldsymbol{B}} = (\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\mathbf{Y} \quad \text{con} \quad \mathbf{Y} = (Y_1, \ldots, Y_n)^t$$

$$\hat{B}_r \sim N\left(\beta_r\,,\ \sigma^2\,[(\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}]_{rr}\right) \quad \Rightarrow \quad \hat{B}_r \text{ stimatore corretto di } \beta_r$$

$$\frac{SS_E}{\sigma^2} \sim \chi^2(n-k-1) \quad \text{indipendente da } \hat{B}_r$$

$$\Rightarrow \ \hat{\Sigma}^2 = \frac{SS_E}{n-k-1} \qquad \text{stimatore corretto di } \sigma^2$$

$$\Rightarrow \ \mathrm{se}(\hat{B}_r) = \sqrt{\hat{\Sigma}^2[(\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}]_{rr}} \quad \text{stimatore approx. corretto di } \sqrt{\mathrm{Var}[\hat{B}_r]}$$

$$\frac{\hat{B}_r - \beta_r}{\mathrm{se}(\hat{B}_r)} \sim t(n-k-1)$$

Stimatori, IC e test per i parametri

$$\frac{\hat{B}_r - \beta_r}{\mathrm{se}(\hat{B}_r)} \sim t(n-k-1)$$

$$\swarrow$$

$$\left(\hat{\beta}_r \pm t_{\frac{1+\gamma}{2}}(n-k-1)\,\mathrm{se}(\hat{\beta}_r)\right)$$

è un $IC(\gamma)$ per β_r

Stimatori, IC e test per i parametri

$$\frac{\hat{B}_r - \beta_r}{\mathrm{se}(\hat{B}_r)} \sim t(n-k-1)$$

$$\swarrow \qquad\qquad\qquad \searrow$$

$$\left(\hat{\beta}_r \pm t_{\frac{1+\gamma}{2}}(n-k-1)\ \mathrm{se}(\hat{\beta}_r)\right)$$

è un $IC(\gamma)$ per β_r

"Rifiuto H_0 se

$$\left|\frac{\hat{B}_r - \beta_{r0}}{\mathrm{se}(\hat{B}_r)}\right| > t_{1-\frac{\alpha}{2}}(n-k-1)\text{"}$$

è un test di livello α per le ipotesi

$H_0 : \beta_r = \beta_{r0}$ vs. $H_1 : \beta_r \neq \beta_{r0}$

Eliminazione a ritroso dei predittori non significativi

$$k = k_0, \quad \text{predittori } x_1, \ldots, x_{k_0}$$

Eliminazione a ritroso dei predittori non significativi

$k = k_0, \quad$ predittori $x_1, \ldots, x_{k_0}$

$\downarrow$

Con R, ricavare l'output della regressione

Eliminazione a ritroso dei predittori non significativi

$k = k_0, \quad$ predittori $x_1, \dots, x_{k_0}$

$\downarrow$

Con R, ricavare l'output della regressione

$\downarrow$

Ordinare $x_1, \dots, x_k$ in modo che i k test

$$H_0 : \beta_r = 0 \quad \text{vs.} \quad H_1 : \beta_r \neq 0$$

abbiano p-value $p_1 < p_2 < \dots < p_k$

Eliminazione a ritroso dei predittori non significativi

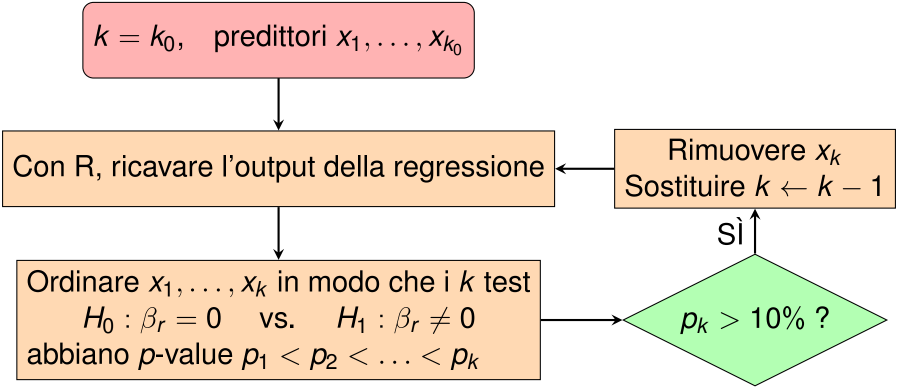
$k = k_0$, predittori $x_1, \ldots, x_{k_0}$

Con R, ricavare l'output della regressione

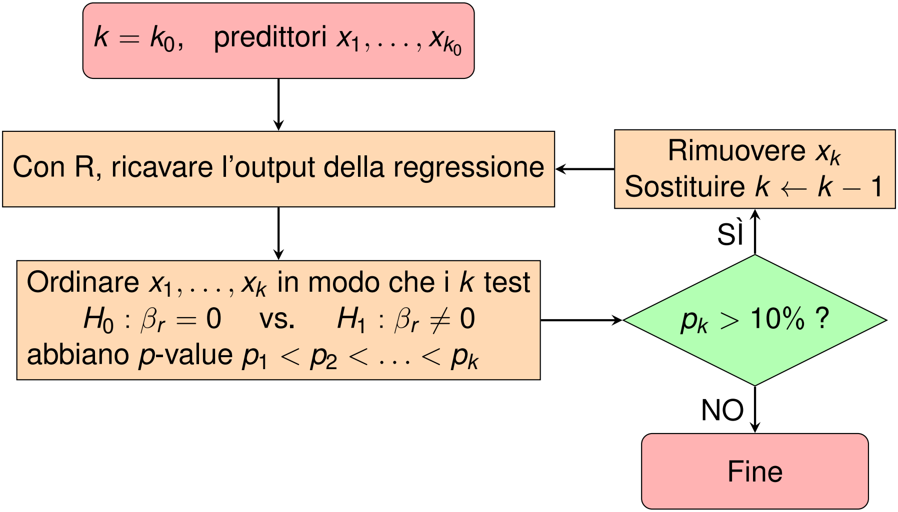
Ordinare $x_1, \ldots, x_k$ in modo che i k test
$H_0 : \beta_r = 0$ vs. $H_1 : \beta_r \neq 0$
abbiano p-value $p_1 < p_2 < \ldots < p_k$

$p_k > 10\%$?

Eliminazione a ritroso dei predittori non significativi



Eliminazione a ritroso dei predittori non significativi



Test per la bontà del modello multilineare

Ipotesi del modello multilineare:

(1) $Y_1, \ldots, Y_n$ indipendenti

(2) $Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \, , \, \sigma^2)$

Test per la bontà del modello multilineare

Ipotesi del modello multilineare:

(1) $Y_1, \ldots, Y_n$ indipendenti

(2) $Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\,,\, \sigma^2)$

Supponendo vera (1), testiamo (2):

$$H_0 : \ \exists\, \boldsymbol{\beta}, \sigma \quad \text{t. c.} \quad Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\,,\, \sigma^2) \quad \forall i = 1, \ldots, n$$

$$H_1 : \ H_0 \text{ è falsa}$$

Test per la bontà del modello multilineare

Ipotesi del modello multilineare:

(1) $Y_1, \ldots, Y_n$ indipendenti

(2) $Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\,,\, \sigma^2)$

Supponendo vera (1), testiamo (2):

$$H_0 : \exists\, \boldsymbol{\beta}, \sigma \quad \text{t. c.} \quad Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\,,\, \sigma^2) \quad \forall i = 1, \ldots, n$$

$$H_1 : H_0 \text{ è falsa}$$

Se (1) - (2) sono vere, allora i *residui studentizzati*

$$R_i' := \frac{Y_i - \hat{Y}_i}{\hat{\Sigma}\sqrt{1 - h_{ii}}} = \frac{R_i}{\sqrt{1 - h_{ii}}} \quad \text{con} \quad h_{ii} = \left[\tilde{\mathbf{x}}(\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\right]_{ii}$$

- sono (approssimativamente) i.i.d.
- $R_i' \approx N(0, 1)$

Test per la bontà del modello multilineare

Ipotesi del modello multilineare:

(1) $Y_1, \ldots, Y_n$ indipendenti

(2) $Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\,,\, \sigma^2)$

Supponendo vera (1), testiamo (2):

$$H_0 : \exists\, \boldsymbol{\beta}, \sigma \quad \text{t. c.} \quad Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\,,\, \sigma^2) \quad \forall i = 1, \ldots, n$$

$$H_1 : H_0 \text{ è falsa}$$

Se (1) - (2) sono vere, allora i *residui studentizzati*

$$R'_i := \frac{Y_i - \hat{Y}_i}{\hat{\Sigma}\sqrt{1 - h_{ii}}} = \frac{R_i}{\sqrt{1 - h_{ii}}} \quad \text{con} \quad h_{ii} = \left[\tilde{\mathbf{x}}(\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}^t\right]_{ii}$$

- sono (approssimativamente) i.i.d.
- $R'_i \approx N(0, 1)$

$\Rightarrow$ test di normalità per $R'_1, \ldots, R'_n$

Stima e IP per una nuova osservazione

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ \vdots \\ x_k^* \end{pmatrix}$$

nuovo input

Legge fisica

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$
$$= (1\ \mathbf{x}^t)\,\boldsymbol{\beta}$$

$$Y^* = (1\ \mathbf{x}^{*\,t})\,\boldsymbol{\beta} + E^*$$

nuovo output

Stima e IP per una nuova osservazione

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ \vdots \\ x_k^* \end{pmatrix} \longrightarrow$$

nuovo input

Legge fisica

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$
$$= (1\ \mathbf{x}^t)\,\boldsymbol{\beta}$$

$$\longrightarrow Y^* = (1\ \mathbf{x}^{*\,t})\,\boldsymbol{\beta} + E^*$$

nuovo output

IPOTESI: $E^* \sim N(0, \sigma^2)$ e indipendente da $E_1, \ldots, E_n$

$\Rightarrow$ $Y^* \sim N(\beta_0 + \beta_1 x_1^* + \ldots + \beta_k x_k^*\,,\ \sigma^2)$ e indipendente da $Y_1, \ldots, Y_n$

Stima e IP per una nuova osservazione

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ \vdots \\ x_k^* \end{pmatrix}$$

nuovo input

Legge fisica

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$
$$= (1\ \mathbf{x}^t)\,\boldsymbol{\beta}$$

$$Y^* = (1\ \mathbf{x}^{*\,t})\,\boldsymbol{\beta} + E^*$$

nuovo output

IPOTESI: $E^* \sim N(0, \sigma^2)$ e indipendente da $E_1, \ldots, E_n$

$\Rightarrow\ Y^* \sim N(\beta_0 + \beta_1 x_1^* + \ldots + \beta_k x_k^*\,,\ \sigma^2)$ e indipendente da $Y_1, \ldots, Y_n$

$\hat{Y}^* := \hat{B}_0 + \hat{B}_1 x_1^* + \ldots + \hat{B}_k x_k^*$ stimatore non distorto del parametro

$$\mathbb{E}\,[Y^*] = \beta_0 + \beta_1 x_1^* + \ldots + \beta_k x_k^*$$

Stima e IP per una nuova osservazione

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ \vdots \\ x_k^* \end{pmatrix}$$

nuovo input

Legge fisica

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$
$$= (1\ \mathbf{x}^t)\,\boldsymbol{\beta}$$

$$Y^* = (1\ \mathbf{x}^{*\,t})\,\boldsymbol{\beta} + E^*$$

nuovo output

IPOTESI: $E^* \sim N(0, \sigma^2)$ e indipendente da $E_1, \ldots, E_n$

$\Rightarrow\ Y^* \sim N(\beta_0 + \beta_1 x_1^* + \ldots + \beta_k x_k^*\,,\ \sigma^2)$ e indipendente da $Y_1, \ldots, Y_n$

$\hat{Y}^* := \hat{B}_0 + \hat{B}_1 x_1^* + \ldots + \hat{B}_k x_k^*$ stimatore non distorto del parametro

$$\mathbb{E}\,[Y^*] = \beta_0 + \beta_1 x_1^* + \ldots + \beta_k x_k^*$$

$IC_{\mathbb{E}[Y^*]}$ e IP_{Y^*} $\longleftarrow$ comando `predict()` di R

Test per la significatività di un gruppo di predittori

Per vedere se il gruppo dei primi r predittori è significativo, si testano

$$H_0 : \underbrace{\beta_1 = \beta_2 = \ldots = \beta_r = 0}_{\text{tutte contemporaneamente}} \qquad H_1 : \underbrace{\beta_s \neq 0 \text{ per qualche } s = 1, \ldots, r}_{\text{almeno una } \neq 0}$$

Test per la significatività di un gruppo di predittori

Per vedere se il gruppo dei primi r predittori è significativo, si testano

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_r = 0 \qquad H_1 : \beta_s \neq 0 \text{ per qualche } s = 1, \ldots, r$$

Un test di livello α è dato dalla regola

$$\text{“rifiuto } H_0 \text{ se } \quad \frac{[SS_E(\mathrm{ridotto}) - SS_E(\mathrm{completo})]/r}{SS_E(\mathrm{completo})/(n-k-1)} > f_{1-\alpha}(r, n-k-1)\text{”}$$

dove

$SS_E(\mathrm{completo}) :=$ varianza residua del modello con tutti i k predittori

$SS_E(\mathrm{ridotto}) :=$ varianza residua del modello senza i primi r predittori

Test per la significatività di un gruppo di predittori

Per vedere se il gruppo dei primi r predittori è significativo, si testano

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_r = 0 \qquad H_1 : \beta_s \neq 0 \text{ per qualche } s = 1, \ldots, r$$

Un test di livello α è dato dalla regola

$$\text{“ rifiuto } H_0 \text{ se} \quad \frac{[SS_E(\mathrm{ridotto}) - SS_E(\mathrm{completo})]/r}{SS_E(\mathrm{completo})/(n-k-1)} > f_{1-\alpha}(r, n-k-1) \text{ ”}$$

dove

$SS_E(\mathrm{completo}) :=$ varianza residua del modello con tutti i k predittori

$SS_E(\mathrm{ridotto}) :=$ varianza residua del modello senza i primi r predittori

Se $k = r$, otteniamo un test per la significatività dell'intero modello:

$$\text{“ rifiuto } H_0 \text{ se} \quad \frac{[SS_R(\mathrm{completo})]/k}{SS_E(\mathrm{completo})/(n-k-1)} > f_{1-\alpha}(k, n-k-1) \text{ ”}$$

perchè $SS_E(\mathrm{ridotto}) = SS_T(\mathrm{completo})$ in questo caso

Output della regressione in R

```
Call:
lm(formula = Distance ~ Temperature + Fuel + I(Temperature^2),
    data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-32.078  -8.482  -0.377   9.223  34.466

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.743e+02  7.902e+01  -2.206   0.0308 *
Temperature       2.321e+01  5.066e+00   4.581 2.03e-05 ***
Fuel             -2.981e-03  7.689e-02  -0.039   0.9692
I(Temperature^2) -4.356e-01  8.139e-02  -5.352 1.10e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.27 on 68 degrees of freedom
Multiple R-squared:  0.631,     Adjusted R-squared:  0.6147
F-statistic: 38.76 on 3 and 68 DF,  p-value: 1.012e-14
```

k

$n - k - 1$

r_A^2

$$\frac{[SS_R\,(\text{completo})]/k}{SS_E\,(\text{completo})/(n-k-1)}$$

p-value del test

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

$H_1 : \beta_s \neq 0$ per qualche $s = 1, \ldots, k$

Generalizzazioni

Caso non lineare	Sostituzione	Linearizzazione
$y = a + bf(x)$	$x' = f(x)$	$y = a + bx'$

Generalizzazioni

Caso non lineare	Sostituzione	Linearizzazione
$y = a + bf(x)$	$x' = f(x)$	$y = a + bx'$
$y = a + b_1 f(x_1) + b_2 g(x_2)$	$x'_1 = f(x_1)$ $x'_2 = g(x_2)$	$y = a + b_1 x'_1 + b_2 x'_2$

Generalizzazioni

Caso non lineare	Sostituzione	Linearizzazione
$y = a + bf(x)$	$x' = f(x)$	$y = a + bx'$
$y = a + b_1 f(x_1) + b_2 g(x_2)$	$x_1' = f(x_1)$ $x_2' = g(x_2)$	$y = a + b_1 x_1' + b_2 x_2'$
$y = a + b_1 x_1^2 + b_2 x_2$	$x_1' = x_1^2$	$y = a + b_1 x_1' + b_2 x_2$

Generalizzazioni

Caso non lineare	Sostituzione	Linearizzazione
$y = a + bf(x)$	$x' = f(x)$	$y = a + bx'$
$y = a + b_1 f(x_1) + b_2 g(x_2)$	$x'_1 = f(x_1)$ $x'_2 = g(x_2)$	$y = a + b_1 x'_1 + b_2 x'_2$
$y = a + b_1 x_1^2 + b_2 x_2$	$x'_1 = x_1^2$	$y = a + b_1 x'_1 + b_2 x_2$
$y = a + b_1 x_1 + b_2 x_1 x_2$	$x'_2 = x_1 x_2$	$y = a + b_1 x_1 + b_2 x'_2$

$x'_2 = x_1 x_2$ *termine di interazione*

Generalizzazioni

Caso non lineare	Sostituzione	Linearizzazione
$y = a + bf(x)$	$x' = f(x)$	$y = a + bx'$
$y = a + b_1 f(x_1) + b_2 g(x_2)$	$x_1' = f(x_1)$ $x_2' = g(x_2)$	$y = a + b_1 x_1' + b_2 x_2'$
$y = a + b_1 x_1^2 + b_2 x_2$	$x_1' = x_1^2$	$y = a + b_1 x_1' + b_2 x_2$
$y = a + b_1 x_1 + b_2 x_1 x_2$	$x_2' = x_1 x_2$	$y = a + b_1 x_1 + b_2 x_2'$

A noi interessa solo fare inferenza su a, b, b_1, b_2

$\Rightarrow$ Ci basta che $\tilde{\mathbf{x}}^t \tilde{\mathbf{x}}$ sia invertibile!

Generalizzazioni

$x_k \in \{c_1, c_2, \ldots, c_l\}$ predittore categorico

(p.es., $x_k \in \{\text{rosso}, \text{verde}, \text{blu}\}$)

Generalizzazioni

$$x_k \in \{c_1, c_2, \ldots, c_l\} \quad \text{predittore categorico}$$

Definiamo $l - 1$ nuovi predittori $z_k,\ z_{k+1}, \ldots,\ z_{k+l-2}$ con $z_r \in \{0, 1\}$:

z_k	z_{k+1}	z_{k+2}	...	x_k
0	0	0	...	c_1
1	0	0	...	c_2
0	1	0	...	c_3
0	0	1	...	c_4
...	...	...	...	...

e facciamo la regressione per i $k + l - 2$ predittori numerici

$$x_1, \ldots,\ x_{k-1},\ z_k, \ldots,\ z_{k+l-2}$$

Generalizzazioni

$$x_k \in \{c_1, c_2, \ldots, c_l\} \quad \text{predittore categorico}$$

Definiamo $l-1$ nuovi predittori $z_k,\ z_{k+1}, \ldots,\ z_{k+l-2}$ con $z_r \in \{0,1\}$:

z_k	z_{k+1}	z_{k+2}	...	x_k
0	0	0	...	c_1
1	0	0	...	c_2
0	1	0	...	c_3
0	0	1	...	c_4
...	...	...	...	...

e facciamo la regressione per i $k+l-2$ predittori numerici

$$x_1, \ldots,\ x_{k-1},\ z_k, \ldots,\ z_{k+l-2}$$

Facciamo così perché vogliamo che $\tilde{\mathbf{x}}^t\tilde{\mathbf{x}}$ sia invertibile.

Generalizzazioni

ESEMPIO: Due modelli diversi:

$$y = \begin{cases} a + bx & \text{se la caratteristica } z = 0 \\ c + dx & \text{se la caratteristica } z = 1 \end{cases}$$

Generalizzazioni

ESEMPIO: Due modelli diversi:

$$y = \begin{cases} a + bx & \text{se la caratteristica } z = 0 \\ c + dx & \text{se la caratteristica } z = 1 \end{cases}$$

$$\Rightarrow \quad y = a + b \quad x + (c - a) \quad z + (d - b) \quad xz$$

Generalizzazioni

ESEMPIO: Due modelli diversi:

$$y = \begin{cases} a + bx & \text{se la caratteristica } z = 0 \\ c + dx & \text{se la caratteristica } z = 1 \end{cases}$$

$$\Rightarrow \quad y = \underbrace{a}_{\beta_0} + \underbrace{b}_{\beta_1}\underbrace{x}_{x_1} + \underbrace{(c-a)}_{\beta_2}\underbrace{z}_{x_2} + \underbrace{(d-b)}_{\beta_3}\underbrace{xz}_{x_3}$$

$$\Rightarrow \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Un esempio

Il famoso esploratore e statistico Jack Bettazzi, recatosi in Groenlandia per condurre la sua ricerca sulle specie locali di foche, dopo mesi di rilievi e osservazioni riesce a misurare i seguenti caratteri di $n = 30$ esemplari: il numero di `anni` di vita dell'esemplare; il suo `peso` in Kg; e infine la sua `specie`, che può essere bianca, marrone o nera.
Decide di impostare questo modello linare per spiegare la variabile `peso` tramite l'età e la specie:

$$\begin{aligned}\texttt{peso}_i &= \beta_0 + \beta_1 \texttt{anni}_i + \beta_2 \texttt{speciemarrone}_i + \beta_3 \texttt{specienera}_i \\ &\quad + \beta_4 \texttt{anni}_i \cdot \texttt{speciemarrone}_i + \beta_5 \texttt{anni}_i \cdot \texttt{specienera}_i + \varepsilon_i\end{aligned}$$

con $\varepsilon_1, \dots, \varepsilon_{30}$ i.i.d. e $\varepsilon_i \sim N(0, \sigma^2)$. Nel modello, il regressore categorico `specie` è codificato come segue:

$$(\texttt{speciemarrone}, \texttt{specienera}) = \begin{cases} (0,\ 0) & \text{se la specie è bianca} \\ (1,\ 0) & \text{se la specie è marrone} \\ (0,\ 1) & \text{se la specie è nera} \end{cases}$$

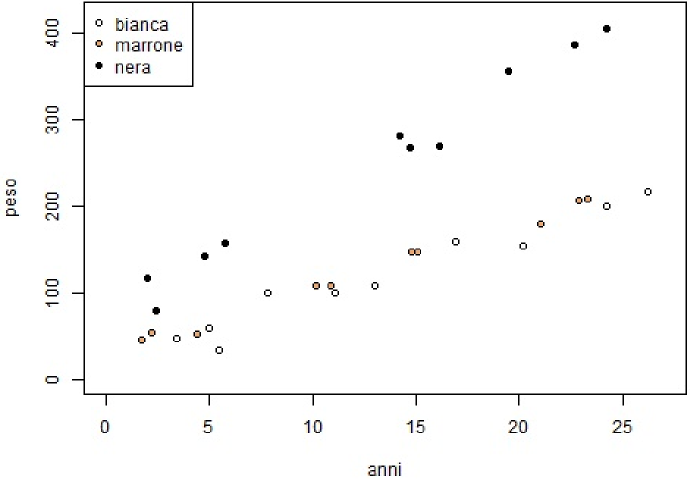
Un esempio

Questi sono i dati che ha trovato:

```
> anni
 [1]  7.8 24.2  2.4 16.1  2.0 10.2 19.5 14.7  2.2  5.5  3.4 10.9  1.7 11.1
[15] 14.2  5.0 22.7  5.8 15.1 20.2 13.0 24.2 22.9 14.8 16.9 23.3  4.4  4.8
[29] 21.0 26.2
> specie
 [1] "bianca"  "nera"    "nera"    "nera"    "nera"    "marrone" "nera"
 [8] "nera"    "marrone" "bianca"  "bianca"  "marrone" "marrone" "bianca"
[15] "nera"    "bianca"  "nera"    "nera"    "marrone" "bianca"  "bianca"
[22] "bianca"  "marrone" "marrone" "bianca"  "marrone" "marrone" "nera"
[29] "marrone" "bianca"
> peso
 [1] 100.1 406.2  80.5 270.0 117.6 108.0 356.4 268.1  55.1  34.4  47.1
[12] 108.8  45.7 100.6 282.5  60.3 386.7 158.4 148.3 155.1 109.0 200.4
[23] 206.3 148.2 160.3 209.6  52.7 143.0 180.5 216.7
```

Un esempio

E questo è il loro scatterplot:



Un esempio

```
Call:
lm(formula = peso ~ anni + specie + anni:specie)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.447  -6.275   2.071   5.379  23.053 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        18.71817    8.20343   2.282   0.0317 *  
anni                7.47801    0.53113  14.080 4.28e-13 ***
speciemarrone      12.05817   11.39650   1.058   0.3006    
specienera         53.24377   11.32394   4.702 8.86e-05 ***
anni:speciemarrone  0.07485    0.75225   0.100   0.9216    
anni:specienera     6.36519    0.74566   8.536 9.82e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.1 on 24 degrees of freedom
Multiple R-squared:  0.9861,    Adjusted R-squared:  0.9832 
F-statistic: 340.6 on 5 and 24 DF,  p-value: < 2.2e-16
```

R usa i due punti per il termine d'interazione

$$\texttt{peso}_i = \beta_0 + \beta_1 \texttt{anni}_i + \beta_2 \texttt{speciemarrone}_i + \beta_3 \texttt{specienera}_i$$

$$+ \beta_4 \texttt{anni}_i \cdot \texttt{speciemarrone}_i + \beta_5 \texttt{anni}_i \cdot \texttt{specienera}_i + \varepsilon_i$$

Un esempio

```
Call:
lm(formula = peso ~ anni + specie + anni:specie)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.447  -6.275   2.071   5.379  23.053 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          18.71817    8.20343   2.282   0.0317 *  
anni                  7.47801    0.53113  14.080 4.28e-13 ***
speciemarrone        12.05817   11.39650   1.058   0.3006    
specienera           53.24377   11.32394   4.702 8.86e-05 ***
anni:speciemarrone    0.07485    0.75225   0.100   0.9216    
anni:specienera       6.36519    0.74566   8.536 9.82e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.1 on 24 degrees of freedom
Multiple R-squared:  0.9861,    Adjusted R-squared:  0.9832 
F-statistic: 340.6 on 5 and 24 DF,  p-value: < 2.2e-16
```

E decide lui come codificare le categorie bianca, marrone e nera (qui ha scelto la codifica di prima)

$$\texttt{peso}_i = \beta_0 + \beta_1 \texttt{anni}_i + \beta_2 \texttt{specie}\textcolor{red}{\texttt{marrone}}_i + \beta_3 \texttt{specie}\textcolor{red}{\texttt{nera}}_i$$

$$+ \beta_4 \texttt{anni}_i \cdot \texttt{specie}\textcolor{red}{\texttt{marrone}}_i + \beta_5 \texttt{anni}_i \cdot \texttt{specie}\textcolor{red}{\texttt{nera}}_i + \varepsilon_i$$

Un esempio

```
Call:
lm(formula = peso ~ anni + specie + anni:specie)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.447  -6.275   2.071   5.379  23.053 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         18.71817    8.20343   2.282   0.0317 *  
anni                 7.47801    0.53113  14.080 4.28e-13 ***
speciemarrone       12.05817   11.39650   1.058   0.3006    
specienera          53.24377   11.32394   4.702 8.86e-05 ***
anni:speciemarrone   0.07485    0.75225   0.100   0.9216    
anni:specienera      6.36519    0.74566   8.536 9.82e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.1 on 24 degrees of freedom
Multiple R-squared:  0.9861,	Adjusted R-squared:  0.9832 
F-statistic: 340.6 on 5 and 24 DF,  p-value: < 2.2e-16
```

p-value bassissimo!
Forte evidenza che la specie nera ha un'intercetta diversa dalle altre

$$\texttt{peso}_i = \beta_0 + \beta_1 \texttt{anni}_i + \beta_2 \texttt{speciemarrone}_i + \color{red}{\beta_3} \texttt{specienera}_i$$

$$+ \beta_4 \texttt{anni}_i \cdot \texttt{speciemarrone}_i + \beta_5 \texttt{anni}_i \cdot \texttt{specienera}_i + \varepsilon_i$$

Un esempio

```
Call:
lm(formula = peso ~ anni + specie + anni:specie)

Residuals:
    Min      1Q  Median      3Q     Max
-25.447  -6.275   2.071   5.379  23.053

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         18.71817    8.20343   2.282   0.0317 *
anni                 7.47801    0.53113  14.080 4.28e-13 ***
speciemarrone       12.05817   11.39650   1.058   0.3006
specienera          53.24377   11.32394   4.702 8.86e-05 ***
anni:speciemarrone   0.07485    0.75225   0.100   0.9216
anni:specienera      6.36519    0.74566   8.536 9.82e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.1 on 24 degrees of freedom
Multiple R-squared:  0.9861,    Adjusted R-squared:  0.9832
F-statistic: 340.6 on 5 and 24 DF,  p-value: < 2.2e-16
```

Ed evidenza ancora più forte che la specie nera ha pure il coefficiente angolare diverso dalle altre

$$\texttt{peso}_i = \beta_0 + \beta_1 \texttt{anni}_i + \beta_2 \texttt{speciemarrone}_i + \beta_3 \texttt{specienera}_i$$

$$+ \beta_4 \texttt{anni}_i \cdot \texttt{speciemarrone}_i + {\color{red}\beta_5} \texttt{anni}_i \cdot \texttt{specienera}_i + \varepsilon_i$$

Un esempio

```
Call:
lm(formula = peso ~ anni + specie + anni:specie)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.447  -6.275   2.071   5.379  23.053 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        18.71817    8.20343   2.282   0.0317 *  
anni                7.47801    0.53113  14.080 4.28e-13 ***
speciemarrone      12.05817   11.39650   1.058   0.3006    
specienera         53.24377   11.32394   4.702 8.86e-05 ***
anni:speciemarrone  0.07485    0.75225   0.100   0.9216    
anni:specienera     6.36519    0.74566   8.536 9.82e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.1 on 24 degrees of freedom
Multiple R-squared:  0.9861,	Adjusted R-squared:  0.9832 
F-statistic: 340.6 on 5 and 24 DF,  p-value: < 2.2e-16
```

Invece, non sembra esserci nessuna evidenza che l'intercetta e il coefficiente angolare della specie bianca e di quella marrone siano diversi.
Però bisognerebbe eliminarli uno alla volta!

$$\texttt{peso}_i = \beta_0 + \beta_1 \texttt{anni}_i + \cancel{\beta_2 \texttt{speciemarrone}_i} + \beta_3 \texttt{specienera}_i$$

$$+ \cancel{\beta_4 \texttt{anni}_i \cdot \texttt{speciemarrone}_i} + \beta_5 \texttt{anni}_i \cdot \texttt{specienera}_i + \varepsilon_i$$