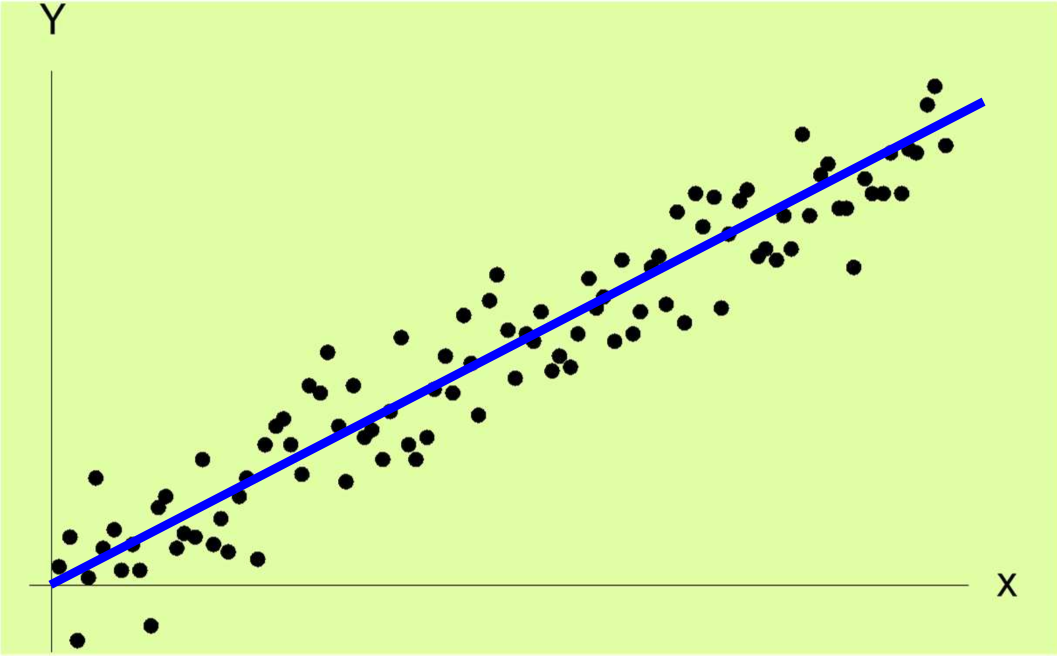
STATISTICA

Regressione-4
Il modello lineare generale

Inferenza



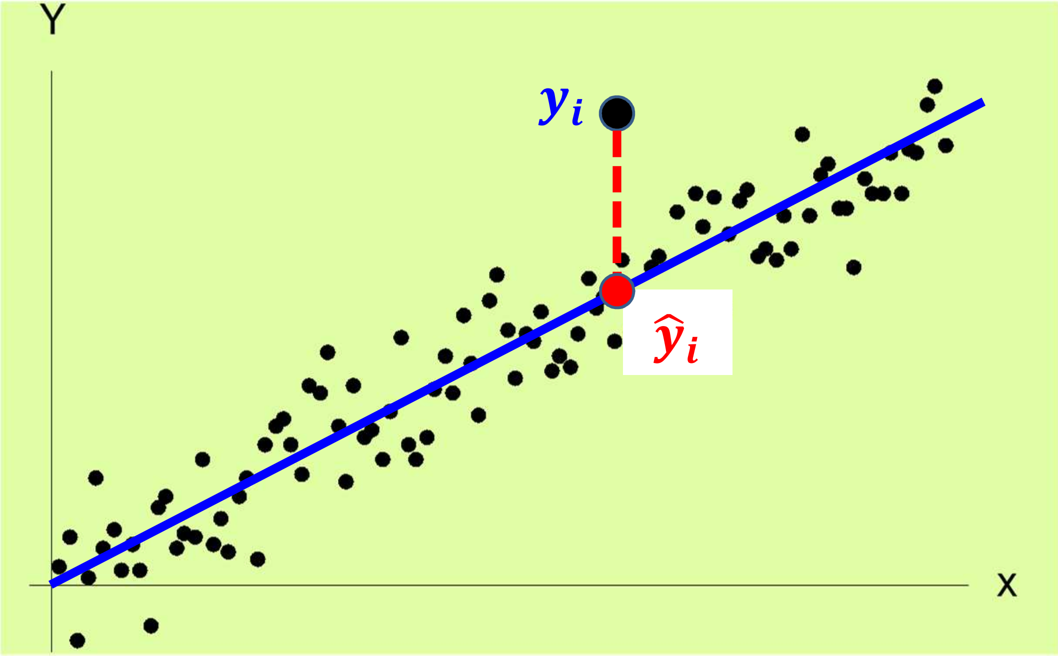
Il modello della **regressione lineare semplice**:

$$f(x) = a + bx$$
$$\approx \Longleftrightarrow \varepsilon_i \sim N(0, \sigma^2)$$
$$\varepsilon_i \text{ indipendenti}$$

$$Y_i = a + bx_i + \varepsilon_i$$

$$Y_i \sim N(a + bx_i, \sigma^2)$$

Inferenza



$$f(x) = a + bx$$

$$\approx \Longleftrightarrow \quad \varepsilon_i \sim N(0, \sigma^2)$$

ε_i **indipendenti**

$$Y_i = a + bx_i + \varepsilon_i$$

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

$$e_i = y_i - \hat{y}_i$$

$$\sum_{i=1}^{n} e_i = 0$$

$$\hat{b} = \frac{\sigma_{xy}}{\sigma^2{}_x} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

Inferenza

dalle stime agli stimatori:

$$B_n = \frac{\sum (Y_i - \bar{Y}_n)(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$A_n = \bar{Y}_n - B_n \bar{x}$$

$\varepsilon_i \sim N(0, \sigma^2)$

$Y_i \sim N(a + bx_i, \sigma^2)$

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} {e_i}^2$$

A_n e B_n v.c. gaussiane

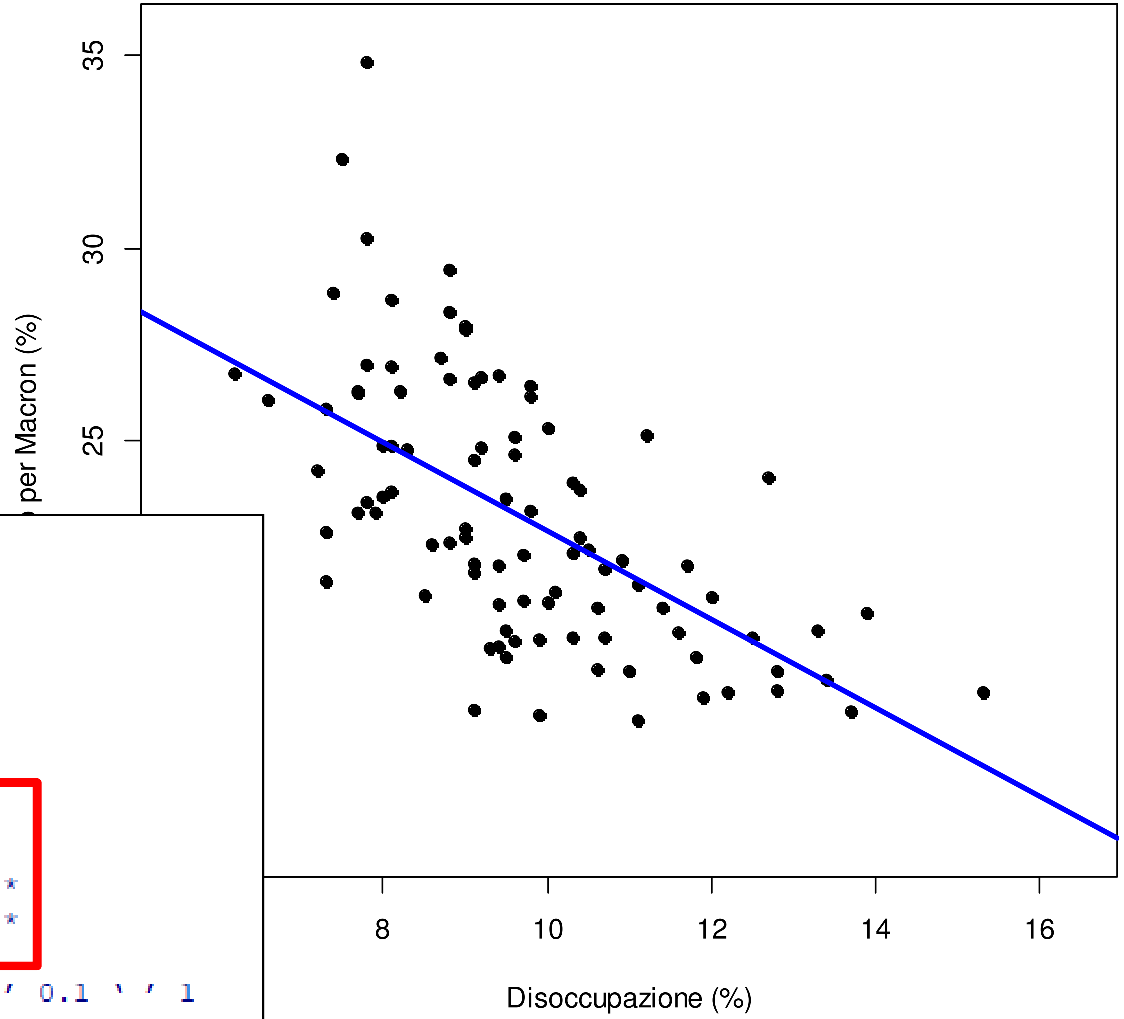
$H_0 : b = 0$ $\quad$ $H_1 : b \neq 0$

rifiutiamo H_0 se:

(rifiutiamo la casualità di una pendenza $\neq 0$)

$$\frac{|\hat{b}|}{\sqrt{\dfrac{s^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}} > t(\boldsymbol{n-2})_{\frac{\alpha}{2}}$$

La regressione con R




```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6817 -1.9000 -0.2081  1.6560  9.6499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.1702     1.5616   21.88  < 2e-16 ***
X            -1.1526     0.1592   -7.24 1.21e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.779 on 94 degrees of freedom
Multiple R-squared:  0.358,     Adjusted R-squared:  0.3512
F-statistic: 52.42 on 1 and 94 DF,  p-value: 1.212e-10
```

Analisi della Varianza

Fonte di variabilità	Gradi di libertà (gl)	SS (Sum of Squares)	Mean Square (SS/gl)
Retta di regressione	1	$\sum_i (\hat{y}_i - \bar{y})^2$	
Attorno alla retta	$n - 2$	$\sum_i (y_i - \hat{y}_i)^2$	$\frac{1}{n-2}\sum_{i=1}^{n} {e_i}^2$
Totale	$n - 1$	$\sum_i (y_i - \bar{y})^2$	

varianza ***spiegata***

varianza ***totale***

num. di parametri stimati (a e b)

Il caso generale

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i \quad , \ i = 1, \dots, n$$

$\varepsilon_i \sim N(0, \sigma^2)$
ε_i indipendenti

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{p,1} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1,n} & \dots & x_{p,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \dots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

Fonte di variabilità	Gradi di libertà (gl)	SS (Sum of Squares)	Mean Square (SS/gl)
Retta di regressione	p	$\sum_i (\hat{y}_i - \bar{y})^2$	
Attorno alla retta	$n - (p + 1)$	$\sum_i (y_i - \hat{y}_i)^2$	
Totale	$n - 1$	$\sum_i (y_i - \bar{y})^2$	

Il caso generale

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i \quad , \; i = 1, \dots, n$$

$\varepsilon_i \sim N(0, \sigma^2)$
ε_i indipendenti

$$\begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \cdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

Fonte di variabilità	Gradi di libertà (gl)	SS (Sum of Squares)	Mean Square (SS/gl)
Retta di regressione	p	$\sum_i (\hat{y}_i - \bar{y})^2$	
Attorno alla retta	$n - (p + 1)$	$\sum_i (y_i - \hat{y}_i)^2$	
Totale	$n - 1$	$\sum_i (y_i - \bar{y})^2$	

***F*-test di linearità**

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

$$\frac{\sum_i (\hat{y}_i - \bar{y})^2 / p}{\frac{1}{n-p-1} \sum_{i=1}^{n} e_i^2} \sim F(p, n-p-1)$$

Il caso particolare!

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i \ \ , \ i = 1, \dots, n$$

$\varepsilon_i \sim N(0, \sigma^2)$
ε_i indipendenti

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} \\ \dots & \dots \\ 1 & x_{1,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \dots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

Fonte di variabilità	Gradi di libertà (gl)	SS (Sum of Squares)	Mean Square (SS/gl)
Retta di regressione	$p = 1$	$\sum_i (\hat{y}_i - \bar{y})^2$	
Attorno alla retta	$n - (1 + 1)$	$\sum_i (y_i - \hat{y}_i)^2$	
Totale	$n - 1$	$\sum_i (y_i - \bar{y})^2$	

F-test di linearità

$$H_0 : \beta_1 = 0$$

$$\frac{\sum_i (\hat{y}_i - \bar{y})^2 / 1}{\frac{1}{n-2} \sum_{i=1}^{n} {e_i}^2} \sim F(1, n-1-1)$$

$$= T^2 \ , T \sim t(n-2)$$

Il caso particolare!

$Y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i \;,\; i = 1, \dots, n$

$\varepsilon_i \sim N(0, \sigma^2)$
ε_i indipendenti

$+ \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix}$

Macron e la disoccupazione

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6817 -1.9000 -0.2081  1.6560  9.6499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.1702                21.88  < 2e-16 ***
X            -1.1526                -7.24 1.21e-10 ***
---
Signif. codes:         0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.779 on 94 degrees of freedom
Multiple R-squared:  0.358,     Adjusted R-squared:  0.3512
F-statistic: 52.42 on 1 and 94 DF,  p-value: 1.212e-10
```

$(-7.24)^2 = 52.42$

			...ean ...uare ...S/gl)
...retta		$\sum_i$	
Totale	$n-1$	$\sum_i (y_i - \bar{y})^2$	

***F*-test di linearità**

$H_0 : \beta_1 = 0$

$\frac{\sum_i (\hat{y}_i - \bar{y})^2 / 1}{\frac{1}{n-2} \sum_{i=1}^{n} {e_i}^2} \sim F(1, n-1-1)$

$= T^2 \,, T \sim t(n-2)$

Facciamo un salto in R

dataset "attitude", Chatterjee, S. and Price, B. (1977) *Regression Analysis by Example*. New York: Wiley. (Section 3.7, p.68ff of 2nd ed.(1991); Section 3.3, p. 52ff of 3rd ed. (2000))

From a survey of the clerical employees of a large financial organization, the data are aggregated from the questionnaires of the approximately 35 employees for each of 30 (randomly selected) departments. The numbers give the percent proportion of favorable responses to seven questions in each department.

There was a question designed to measure the overall performance of a supervisor, as well as questions that were related to specific activities involving interaction between supervisor and employee.

X_1, X_2, X_5 related to direct interpersonal relationships between superv. and empl., whereas X_3 and X_4 related to the job as a whole. X_6 (rate of advancing to better jobs) served as a general measure of how the empl. perceives his/her own progress in the company.

The response to any item ranged 1-5 (1=very satisfactory, 5=very insatisfactory). A dichotomus index was created to each item: {1,2} = favorable response. Data have been aggregated for departments.

Facciamo un salto in R

dataset "attitude", Chatterjee, S. and Price, B. (1977) *Regression Analysis by Example*. New York: Wiley. (Section 3.7, p.68ff of 2nd ed.(1991); Section 3.3, p. 52ff of 3rd ed. (2000))

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_6 x_{6,i} + \varepsilon_i \quad , \; i = 1, \ldots, 30$$

```
Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708   11.58926   0.931 0.361634
complaints   0.61319    0.16098   3.809 0.000903 ***
privileges  -0.07305    0.13572  -0.538 0.595594
learning     0.32033    0.16852   1.901 0.069925 .
raises       0.08173    0.22148   0.369 0.715480
critical     0.03838    0.14700   0.261 0.796334
advance     -0.21706    0.17821  -1.218 0.235577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3 Residual standard error: 7.068 on 23 degrees of freedom
  Multiple R-squared:  0.7326,    Adjusted R-squared:  0.6628   2
1 F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

$$s^2 = \frac{1}{n - p - 1} \sum_{i=1}^{n} {e_i}^2$$

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Facciamo un salto in R

dataset "attitude", Chatterjee, S. and Price, B. (1977) *Regression Analysis by Example*. New York: Wiley. (Section 3.7, p.68ff of 2nd ed.(1991); Section 3.3, p. 52ff of 3rd ed. (2000))

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_6 x_{6,i} + \varepsilon_i \quad , \; i = 1, \ldots, 30$$

```
Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t val   Pr(>|t|)
(Intercept) 10.78708   11.58926    0.9  0.361634
complaints   0.61319    0.16098    3.8  0.000903 ***
privileges  -0.07305    0.13572   -0.5  0.595594
learning     0.32033    0.16852    1.9  0.069925 .
raises       0.08173    0.22148    0.3  0.715480
critical     0.03838    0.14700    0.2  0.796334
advance     -0.21706    0.17821   -1.2  0.235577
---
Signif. codes:  0 '***' 0.001 '**' 0.   '*' 0.05 '.'    1 ' ' 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,    Adjusted R-squared:  0.6628
F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

$$\frac{1}{n-p-1}\sum_{i=1}^{n} {e_i}^2$$

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Il caso generale

$H_0 : \beta_1 = 0$

$Y_i = \beta_0 + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i \quad , \; i = 1, \ldots, n$

modello *ridotto*

Fonte di variabilità	Gradi di libertà (gl)	SS (Sum of Squares)	Mean Square (SS/gl)
Retta di regressione		$\sum_i (\hat{y}_i - \bar{y})^2$=SSR	
Attorno alla retta		$\sum_i (y_i - \hat{y}_i)^2$=SSE	
Totale	$n-1$	$\sum_i (y_i - \bar{y})^2$	

rifiuto H_0 se supero il quantile della $F(1, n-p-1)$

$$\frac{(\text{SSR}-\text{SSR}_\text{R})/1}{\text{SSE}/(n-p-1)} \sim F(1, n-p-1)$$

$$= T^2 , T \sim t(n-p-1)$$

Facciamo un salto in R

dataset "attitude", Chatterjee, S. and Price, B. (1977) *Regression Analysis by Example*. New York: Wiley. (Section 3.7, p.68ff of 2nd ed.(1991); Section 3.3, p. 52ff of 3rd ed. (2000))

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_6 x_{6,i} + \varepsilon_i \quad , \; i = 1, \ldots, 30$$

```
Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708   11.58926   0.931 0.361634
complaints   0.61319    0.16098   3.809 0.000903 ***
privileges  -0.07305    0.13572  -0.538 0.595594
learning     0.32033    0.16852   1.901 0.069925 .
raises       0.08173    0.22148   0.369 0.715480
critical     0.03838    0.14700   0.261 0.796334
advance     -0.21706    0.17821  -1.218 0.235577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,    Adjusted R-squared:  0.6628
F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

Quante variabili è più economico tenere?

$$s^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} {e_i}^2$$

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Il caso generale

$H_0 : \beta_1 = \cdots = \beta_q = 0$

$Y_i = \beta_0 + \beta_{q+1} x_{q+1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i \quad , \ i = 1, \dots, n$

modello *ridotto*

Fonte di variabilità	Gradi di libertà (gl)	SS (Sum of Squares)	Mean Square (SS/gl)
Retta di regressione		$\sum_i (\hat{y}_i - \bar{y})^2$=SSR	
Attorno alla retta		$\sum_i (y_i - \hat{y}_i)^2$=SSE	
Totale	$n-1$	$\sum_i (y_i - \bar{y})^2$	

rifiuto H_0 se supero il quantile della $F(q, n-p-1)$

$\frac{(\mathrm{SSR} - \mathrm{SSR_R})/q}{\mathrm{SSE}/(n-p-1)} \sim F(q, n-p-1)$

Il caso generale

$H_0 : \beta_1 = \cdots = \beta_q = 0$

$Y_i = \beta_0 + \beta_{q+1} x_{q+1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i \quad , \quad i = 1, \ldots, n$

modello *ridotto*

Se voglio risparmiare sul modello
NON DEVO RIFIUTARE
quindi preferisco i
p-valori > 0.05

Fonte di		
Totale	$n - 1$	$\sum_i (y_i - \bar{y})^2$

rifiuto H_0 se supero il quantile della $F(q, n-p-1)$

$$\frac{(\mathrm{SSR} - \mathrm{SSR_R})/q}{\mathrm{SSE}/(n-p-1)} \sim F(q, n-p-1)$$

Il caso generale

$H_0 : \beta_1 = \cdots = \beta_q = 0$

$$\frac{(R^2 - {R_R}^2)/q}{(1 - R^2)/(n - p - 1)}$$

$Y_i = \beta_0 + \beta_{q+1} x_{q+1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i \quad , \; i = 1, \ldots, n$

modello *ridotto*

Fonte di variabilità	Gradi di libertà (gl)	SS (Sum of Squares)	Mean Square (SS/gl)
Retta di regressione		$\sum_i (\hat{y}_i - \bar{y})^2$=SSR	
Attorno alla retta		$\sum_i (y_i - \hat{y}_i)^2$=SSE	
Totale	$n - 1$	$\sum_i (y_i - \bar{y})^2$	

rifiuto H_0 se supero il quantile della $F(q, n - p - 1)$

$$\frac{(\mathrm{SSR} - \mathrm{SSR_R})/q}{\mathrm{SSE}/(n-p-1)} \sim F(q, n - p - 1)$$

$$\frac{(R^2 - {R_R}^2)/q}{(1 - R^2)/(n - p - 1)} = \frac{(0.7326 - 0.708)/4}{(1 - 0.7326)/23} = 0.621$$

dataset "attitude", Chatterj

Example. New York: Wiley. (Section

3rd ed. (2000))

```
Call:
lm(formula = rating ~ complaints + learning)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.5568  -5.7331   0.6701   6.5341  10.3610 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.8709     7.0612   1.398    0.174    
complaints    0.6435     0.1185   5.432 9.57e-06 ***
learning      0.2112     0.1344   1.571    0.128    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 6.817 on 27 degrees of freedom
Multiple R-squared:  0.708,	Adjusted R-squared:  0.6864 
F-statistic: 32.74 on 2 and 27 DF,  p-value: 6.058e-08
```

```
Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
     Min       1Q   Median       3Q      M
-10.9418  -4.3555   0.3158   5.5425  11.59

Coefficients:
            Estimate Std. Error t value Pr
(Intercept) 10.78708   11.58926   0.931 0.361634    
complaints   0.61319    0.16098   3.809 0.000903 ***
privileges  -0.07305    0.13572  -0.538 0.595594    
learning     0.32033    0.16852   1.901 0.069925 .  
raises       0.08173    0.22148   0.369 0.715480    
critical     0.03838    0.14700   0.261 0.796334    
advance     -0.21706    0.17821  -1.218 0.235577    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,	Adjusted R-squared:  0.6628 
F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```