MODELLO REGRESSIONE LINEARE SEMPLICE

Upload the file: Browse... MMPI.dat — Upload complete

Default max. file size is 5MB

Seleziona le variabili:

Seleziona variabile Dipendente Y: CA

Seleziona variabile Indipendente X: PAP

Ipotesi

$H_0 : \beta_1 = 0$

$H_1 : \beta_1$ > 0

Data Summary | Correlazione | Modello di Regressione | Residui | Inferenza | BIC e BF

Summary delle variabili selezionate

Comando in r

summary(var_x)

	Min	1sr Qu.	Median	Mean	3rd Qu.	Max
PAP	16	49	55.5	54.9117647058824	61	80
CA	26	47	53	53.7941176470588	62	82

Tabella Dati

Show 10 entries — Search:

	CA	PAP
1	65	51
2	52	60
3	48	40
4	59	40
5	46	53
6	56	62
7	67	56
8	57	59
9	54	63
10	45	54

Showing 1 to 10 of 102 entries

Previous 1 2 3 4 5 … 11 Next

MODELLO REGRESSIONE LINEARE SEMPLICE

Upload the file: Browse... MMPI.dat — Upload complete

Default max. file size is 5MB

Seleziona le variabili:

Seleziona variabile Dipendente Y: CA

Seleziona variabile Indipendente X: PAP

Ipotesi

$H_0 : \beta_1 = 0$

$H_1 : \beta_1$ > 0

Data Summary | Correlazione | Modello di Regressione | Residui | Inferenza | BIC e BF

Coefficiente di Correlazione

Il coefficiente di correlazione misura l'intensità della relazione lineare tra due variabili quantitative. Non permette di valutare, però, la relazione causa-effetto.
La formula del coefficiente di correlazione di Pearson è:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{Cov_{xy}}{S_x \cdot S_y}$$

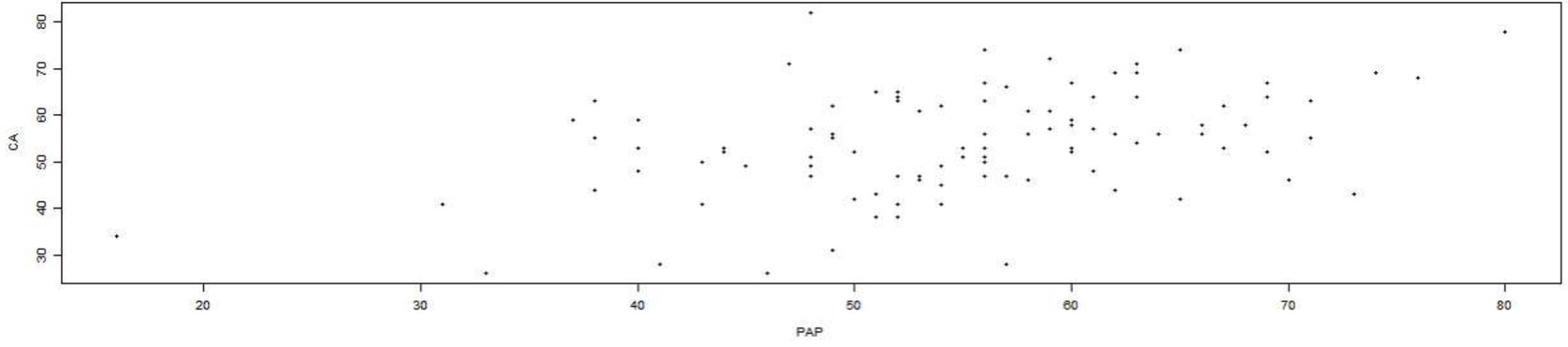
Il coefficiente va da $[-1; +1]$:

- $-1 \Rightarrow$ perfetta relazione lineare negativa, ovvero all'aumentare di una variabile l'altra diminuisce
- $+1 \Rightarrow$ perfetta relazione lineare positiva, ovvero all'aumentare di una variabile aumenta anche l'altra;
- $0 \Rightarrow$ non esiste una relazione lineare;

Rappresentazione grafica dei dati:

Comando in r

```
plot(var_x,var_y, pch = 20, cex = 1, col = 'black', xlab= var_x, ylab= var_y)
```



Correlazione tra PAP e CA

☑ Comando in r

cor.test(var_y,var_x)

```
        Pearson's product-moment correlation

data:  var$y and var$x
t = 4.532, df = 100, p-value = 1.622e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2373693 0.5621370
sample estimates:
      cor
0.4127889
```

MODELLO REGRESSIONE LINEARE SEMPLICE

Upload the file: Browse... MMPI.dat — Upload complete

Default max. file size is 5MB

Seleziona le variabili:

Seleziona variabile Dipendente Y: CA

Seleziona variabile Indipendente X: PAP

Ipotesi

$H_0 : \beta_1 = 0$

$H_1 : \beta_1$ > 0

Data Summary | Correlazione | Modello di Regressione | Residui | Inferenza | BIC e BF

Modello Regressione Lineare Semplice

La regressione mira a stabilire se tra due variabili vi sia una dipendenza funzionale, in particolare a quantificare la misura in cui una variabile (chiamata predittore) agisca su un'altra (chiamata dipendente).

La regressione lineare consiste nel determinare il legame tra le due variabili attraverso una funzione lineare del tipo:

$$Y = \beta_0 + \beta_1 X$$

dove

- Y è la variabie dipendente,
- X è il predittore,
- β_0 rappresenta l'intercetta,
- β_1 rappresenta il coefficiente angolare

Retta di Regressione dei dati

La retta di regressione riguardo i dati inseriti e':

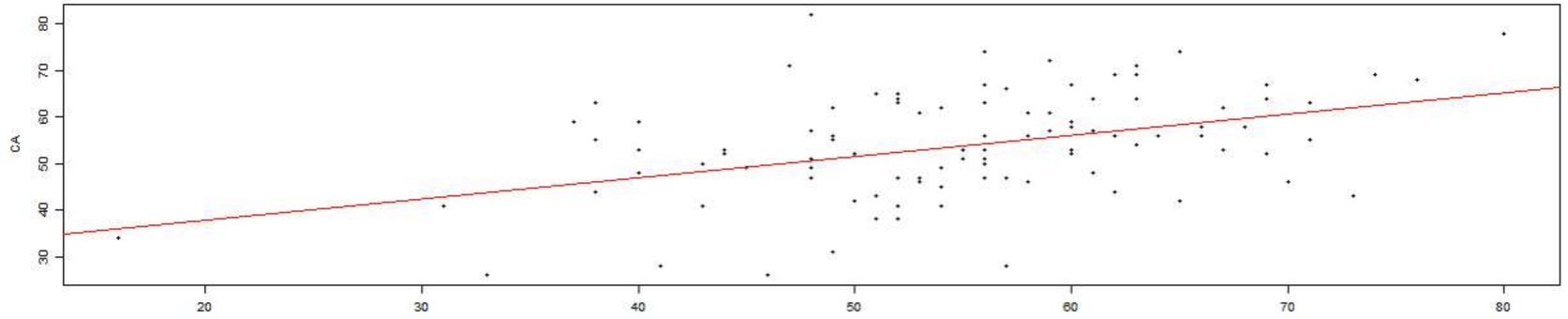
$$Y = 28.815 + 0.455X$$

☑ Comando modello in r

```
lm(var_y ~ var_x, data = 'nome_dataframe')
```

☑ Comando grafico in r

```
plot( var_x, var_y, pch = 20, cex = 1, col = 'black', xlab = var_x, ylab = var_y)
abline( lm(var_y ~ var_x), col = 'red')
```



MODELLO REGRESSIONE LINEARE SEMPLICE

Upload the file: Browse... MMPI.dat — Upload complete

Default max. file size is 5MB

Seleziona le variabili:

Seleziona variabile Dipendente Y: CA

Seleziona variabile Indipendente X: PAP

Ipotesi

$H_0 : \beta_1 = 0$

$H_1 : \beta_1$ > 0

Data Summary | Correlazione | Modello di Regressione | Residui | Inferenza | BIC e BF

Assunti Modello di Regressione Lineare

È importante controllare che gli assunti del modello di regressione siano rispettati prima di fare inferenze sui dati.
Gli assunti sono:

- **Linearità:** Il valore attesso dell'errore per un dato valore di X è zero:

$$E(\epsilon_i) = E(\epsilon|x_i) = 0$$

  In pratica significa che il valore atteso della variabile dipendente, E (Y), è una funzione lineare del predittore

- **Normalità:** Gli errori sono distribuiti normalmente intorno lo zero:

$$\epsilon_i \; N(0, \sigma^2)$$

- **Omogeneità delle varianze:** La varianza degli errori è costante per qualunque valori di X :

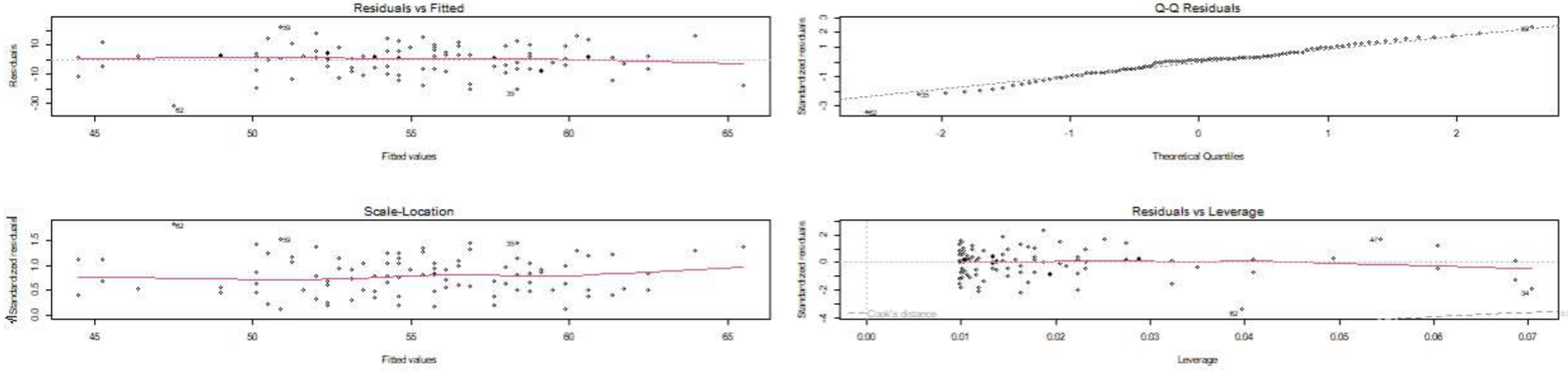
$$V(\epsilon|x_i) = \sigma^2$$

- **Indipendenza:** Tutte le coppie di errori ϵ_i ed ϵ_j sono tra loro indipendenti per ogni $i \neq j$

Grafici Residui

Comando in r

```
par(mfrow = c(2,2))
plot(lm(var_x ~ var_y))
```



Guida per interpretare i grafici:

1. **Residual vs Fitted:** Rappresenta i residui ϵ_i in funzione dei valori attesi $\hat{y}^2$
   Questo grafico permette di valutare l'assunto di indipendenza. Nel caso ideale i residui si distribuiscono normalmente intorno all media zero e la linea continua (che rappresenta il trend stimato delle medie dei residui) risulta sostanzialmente piatta. In presenza di una violazione dell'assunto di indipendenza dei residui, la linea può assumere ad esempio una forma crescente/decrescente.

2. **Normal Q-Q:** Rappresenta i residui standardizzati in funzione dei quantili della normale e serve per valutare l'assunto di normalità.
   Se la distribuzione dei residui fosse perfettamente normale, i punti sarebbero tutti allineati lungo la retta grigia tratteggiata. In presenza di una violazione dell'assunto di normalità, i punti non si ditribuirebbero lungo una retta.

3. **Scale-Location:** Rappresenta la radice quadrata dei residui standardizzati in funzione dei valori attesi $\hat{y}^2$.
   Questo grafico permette di valutare l'assunto di omogeneità delle varianze. La linea continua indica il trend stimato e non dovrebbe esprimere alcun tipo di trend, indicando così l'omogeneità delle varianze dei residui.

4. **Residual vs Leverage:** Rappresenta i residui in funzone del valore di leverage.
   Con questo grafico si possono individuare casi anomali e casi influenti. Il valore di leverage misura la potenziale influenza di un dato sulle stime dei parametri del modello o quanto possa variare l'inclinazione della retta.
   Ad esempio, un alto valore di leverage unito ad un alto residuo implica che il dato ha un forte valore sulle stime.
   Una buona misura che riassume l'influenza di ciascun dato è la distanza di Cook. Nel grafico viene rappresentata utilizzando le delle linee trattegiate. I punti che cadono nelle aree esterne delimitate da tali linee rappresentano distanze grandi e quindi risultano essere valori influenti.

MODELLO REGRESSIONE LINEARE SEMPLICE

Upload the file: Browse... MMPI.dat — Upload complete

Default max. file size is 5MB

Seleziona le variabili:

Seleziona variabile Dipendente Y: CA

Seleziona variabile Indipendente X: PAP

Ipotesi

$H_0 : \beta_1 = 0$

$H_1 : \beta_1$ > 0

Data Summary | Correlazione | Modello di Regressione | Residui | Inferenza | BIC e BF

Inferenza

Ci si pone ora il problema di stabilire se 'PAP' rappresenti un predittore statisticamente significativo di 'CA'.
L'intensità della relazione tra le due varibili è legata al parametro β_1 chiamato anche coefficiente di regressione.

Informazioni sul modello:

Comando in r

```
fit <- lm(var_y ~ var_x)
summary(fit)
```

```
Call:
lm(formula = var$y ~ var$x)

Residuals:
     Min       1Q   Median       3Q      Max
-26.7441  -6.2480  -0.5597   6.7295  31.3501

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.8146     5.6079   5.138 1.38e-06 ***
var$x         0.4549     0.1004   4.532 1.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.44 on 100 degrees of freedom
Multiple R-squared:  0.1704,    Adjusted R-squared:  0.1621
F-statistic: 20.54 on 1 and 100 DF,  p-value: 1.622e-05
```

Info Residual

Nella prima parte dell'output sono riportate le statistiche generali sui residui (Residuals).
L'analisi dei residui permette una prima valutazione sulla bontà dell'adattamento ai dati della retta stimata. In generale dovremmo attenderci dei residui non troppo elevati con valori sia positivi che negativi. La mediana dovrebbe essere vicino allo zero e minimo e massimo approssimativamente uguali.

Info Coefficients

Nella seconda parte dell'output (Coefficients) sono riportati i risultati relativi ai test sull'intercetta (Intercept)) e sul coefficiente di regressione (PAP)
Nell'ordine si possono leggere i valori stimati dei parametri (Estimate) con relativo errore standard (Std. Error), la statistica test utilizzata (t value) con relativa probabilità associata (Pr(> | t |))

Interpretazione t value

Il valore del t value ci permette di verificare se le nostre ipotesi sono congruenti ai dati.
Guardare il valore del t value e confrontarlo con H_1. Se si vuole fare un'ipotesi monodirezionale:

- Se il valore è $negativo$ allora H_1 deve ipotizzare $\beta_1 < 0$
- Se il valore è $positivo$ allora H_1 deve ipotizzare $\beta_1 > 0$

Interpretazione p value

Il valore di p value ci permette di determinare se il test è statisticamente significativo e quindi rigettare H_0 oppure no. Tre sono gli aspetti essenziali:

- Il p value indica la probabilità di un risultato uguale o più estremo di quello osservato, posto che sia vera H_0. Non rappresenta in alcun modo la probabilità che sia vera H_0 e pertanto non può essere considerato come un misuratore del grado di falsità della stessa ipotesi
- Il p value risente, in generale, della numerosità campionaria. Aumentando la dimensione del campione il valore di p tende a diminuire
- Il p value non può essere considerato una misura dell'evidenza statistica, va usato come criterio decisionale per rigettare o meno H_0

In psicologia α (livello di significatività) è fissato al 5%. Pertanto se il p-value risulta:

- superiore al 0.05 (5%) il test non è statisticamente significativo, di conseguenza non si può rigettare l'ipotesi H_0;
- inferiore al 0.05 (5%) il test è statisticamente significativo, di conseguenza si può rigettare l'ipotesi H_0.

Per quanto riguarda questo modello, il test risulta statisticamente significativo e quindi si può rigettare H_0 perchè p value = 1e-05 ed è minore di 0.05

Ultime info

Nell'ultima parte dell'output son riportate alcune misure che permettono di valutare l'intensità della relazione tra le variabili PAP e CA.
Il Residual standard error chiamato anche *errore standard della regressione* , indica la variabilità media dei punti intorno alla retta di regressione. Pertanto, quanto maggiore sarà questo valore tanto minore sarà il potere predittivo della retta di regressione.

R-Squared chiamato anche *coefficiente di determinazione* (R^2) rappresenta la porzione di devianza spiegata dalla relazione lineare. Ajusted R-Squared è lo stesso coefficiente di determinazione corretto per i gradi di libertà.

Il valore di r^2 = 0.1704

ANOVA

L'analisi della varianza (ANOVA) permette di avere informazioni riguardo ai livelli di variabilità all'interno del modello di regressione e forma le basi per i test di significatività.
Il concetto base della linea di regressione è dato da:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Dove:

- $y_i - \bar{y}$: Variazione totale di y
- $\hat{y}_i - \bar{y}$: Variazione della risposta media
- $y_i - \hat{y}_i$: Valore residuo.

Mettendo a quadrato ciascuno di questi termini e sommando tutte le n osservazioni si ottiene l'equazione:

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

.
che può anche essere scritta come $SST = SSM + SSE$, dove SST è la devianza totale, SSM è la devianza di regressione, SSE è la devianza residua.

Il quadrato della correlazione del campione è uguale al rapporto tra la somma del modello dei quadrati e la somma totale dei quadrati: r^2 = SSM / SST.
Ciò formalizza l'interpretazione di r^2 come spiegazione della frazione di variabilità nei dati spiegati dal modello di regressione.

La varianza del campione ${s_y}^2$ è uguale a $(y_i - \bar{y})^2/(n-1) = SST/DFT$, la somma totale dei quadrati divisa per i gradi di libertà totali DFT.

Per la regressione lineare semplice, MSM (mean square model):

$$MSM = \sum(\hat{y}_i - \bar{y})^2/(1) = SSM/DFM$$

L' MSE corrispondente (errore quadrato medio) $= \sum(y_i - \hat{y}_i)^2/(n-2) = SSE/DFE$, la stima della varianza attorno alla linea di regressione della popolazione σ^2.

Anova del modello

☑ Comando grafico in r

anova(lm(var_y ~ var_x))

Df	Sum Sq	Mean Sq	F value	Pr(>F)
1	2239.10	2239.10	20.54	0.00
100	10901.58	109.02	NA	NA

MODELLO REGRESSIONE LINEARE SEMPLICE

Upload the file: Browse... MMPI.dat — Upload complete

Default max. file size is 5MB

Seleziona le variabili:

Seleziona variabile Dipendente Y: CA

Seleziona variabile Indipendente X: PAP

Ipotesi

$H_0 : \beta_1 = 0$

$H_1 : \beta_1$ > 0

Data Summary | Correlazione | Modello di Regressione | Residui | Inferenza | BIC e BF

Bayesian Information Criterion (BIC)

Il Bayesian Information Criterion (BIC; Schwarz, 1978) è un criterio utile per la selezione di modelli.
Dato un modello M, viene definito con

$$BIC(M) = k \cdot ln(n) - 2ln(L)$$

in cui L è il valore di massima verosimiglianza del modello, k il numero di parametri e n il numero di osservazioni.
In generale, migliore è il modello più basso risulta essere il valore di BIC, pertanto possiamo utilizzare tale statistica per scegliere il modello migliore.

Il BIC di questo modello è: 779.851

Comando in r

```
BIC(lm(var_y ~ var_x))
```

Bayesian Factor (BF)

Una misura di quantificazione dell'evidenza di un'ipotesi rispetto ad un'altra è il Bayes Factor.
Nel caso classico con H_0 e H_1 contrapposte:

$$BF_10 = f(x|H_1) \cdot f(x|H_0)$$

Quando il valore supera 1 allora è più evidente H1. In generale BF consente di valutare quanto un'ipotesi (o un modello) sia più evidente di un'altra/o.

BIC_1 = 780

BIC_2 =

1

BF = 1.33440934269402e+169

Comando in r

```
library(BayesFactor)
exp( (BIC_1 - BIC_2) / 2)
```

Created and Developed by
Valeria Pamato

Bibliography
Pastore, M. (2015) . “Analisi dei dati in psicologia con applicazioni in R”. Bologna: Il Mulino