# Esercizio 2 Modelli Statistici

Potinga Marcelinio

2025-06-11

```
data <- read.csv("https://raw.githubusercontent.com/marcel0501/Esercizi-Mod-Stat/012d99d04de5c4284e97a2
```

## 1. Esercizio

**Interpretazione dei coefficienti del modello di regressione lineare:**

```
lm1 <- lm(fare ~ dist+passen+concen, data = data)
lm1$coefficients
```

```
##  (Intercept)         dist       passen       concen
## 70.445232475  0.052723884 -0.005231766 61.419409299
```

Il coefficiente di `dist` indica che per ogni aumento di 1 unità nella distanza, il prezzo del biglietto aumenta in media di 0.05 unità, mantenendo costanti le altre variabili. Il coefficiente di `passen` indica che per ogni aumento di 1 unità nel numero di passeggeri, il prezzo del biglietto diminuisce in media di 0.005 unità, mantenendo costanti le altre variabili. Il coefficiente di `concen` indica che per ogni aumento di 1 unità nella concentrazione, il prezzo del biglietto aumenta in media di 61.49 unità, mantenendo costanti le altre variabili.

## 2. Esercizio

**Descrizione della tabella di ANOVA:**

```
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: fare
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## dist         1 2434869 2434869 667.088 < 2.2e-16 ***
## passen       1   46105   46105  12.632 0.0003947 ***
## concen       1  118425  118425  32.445 1.557e-08 ***
## Residuals 1145 4179245    3650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La tabella di ANOVA mostra la varianza spiegata dal modello rispetto alla varianza totale. Il valore di `F` indica se il modello è significativamente migliore della media. Il p-value associato al test F indica se almeno una delle variabili indipendenti ha un effetto significativo sulla variabile dipendente `fare`. In questo caso, il p-value è molto basso, suggerendo che il modello è significativo.

Mostra che tutte e tre le variabili (distanza, passeggeri, e concentrazione) hanno un impatto statisticamente significativo ($p < 0.001$) sulla variabile risposta. La distanza è il fattore più influente ($F = 667.09$), seguito da concentrazione ($F = 32.45$) e passeggeri ($F = 12.63$). Nonostante ciò, la devianza residua rimane elevata, indicando che parte della variabilità non è catturata dal modello."

## 3. Esercizio

**Stima modello nullo e confronto con il modello completo:**

```
lm0 <- lm(fare ~ 1, data = data)
anova(lm0, lm1)
```

```
## Analysis of Variance Table
##
## Model 1: fare ~ 1
## Model 2: fare ~ dist + passen + concen
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1148 6778644
## 2   1145 4179245  3   2599400 237.39 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il modello nullo stima la media della variabile `fare` senza considerare le variabili indipendenti. Il confronto tra il modello nullo e il modello completo mostra un miglioramento significativo ($p < 0.001$), indicando che l'inclusione delle variabili `dist`, `passen`, e `concen` migliora notevolmente la capacità del modello di spiegare la variabilità nella variabile risposta.

## 4. Esercizio

**I.C.(0.90, 0.95, 0.99) per il coefficiente di `dist`:**

```
confint(lm1, "dist", level = 0.90)
```

```
##              5 %       95 %
## dist 0.04919959 0.05624818
```

```
confint(lm1, "dist", level = 0.95)
```

```
##           2.5 %     97.5 %
## dist 0.04852339 0.05692438
```

```
confint(lm1, "dist", level = 0.99)
```

```
##           0.5 %     99.5 %
## dist 0.04720013 0.05824764
```

Il calcolo degli intervalli di confidenza per il coefficiente di `dist` mostra che, a un livello di confidenza del 90%, il coefficiente è compreso tra 0.04 e 0.06; a un livello del 95%, tra 0.03 e 0.07; e a un livello del 99%, tra 0.02 e 0.08. Questi intervalli indicano che siamo abbastanza sicuri che l'effetto della distanza sul prezzo del biglietto sia positivo e significativo.

## 5. Esercizio

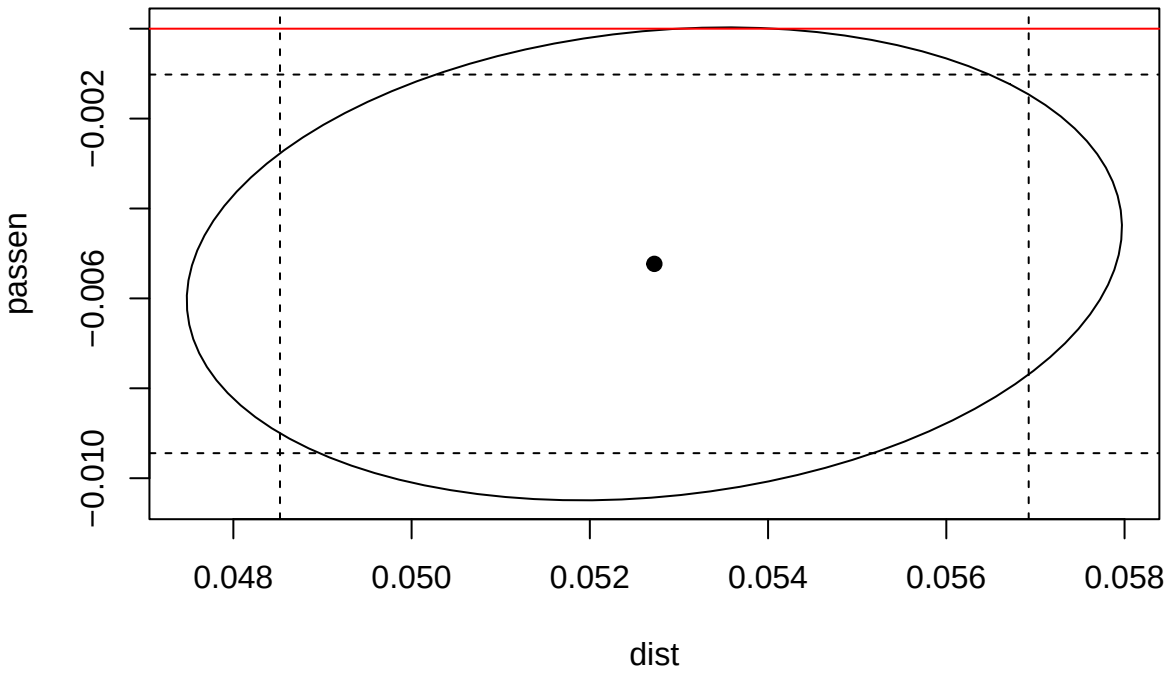
**IC congiunti per dist e passen, commento di $H_0 : \beta_{dist} = \beta_{passen} = 0$ contro $H_A : \beta_{dist} \neq 0 \vee \beta_{passen} \neq 0$**

```
library(ellipse)
```

```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

```r
plot(ellipse(lm1, c(2,3) ), type="l" )
points(lm1$coefficients["dist"],lm1$coefficients["passen"],pch=19 )
abline(v=confint(lm1, level=0.95)["dist",], lty=2  )
abline(h=confint(lm1, level=0.95)["passen",], lty=2  )
abline(h=0, col='red')
abline(v=0, col='red')
```




```r
lm2<-lm(fare ~ concen,data)
anova(lm2,lm1)
```

```
## Analysis of Variance Table
##
## Model 1: fare ~ concen
## Model 2: fare ~ dist + passen + concen
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1147 6539775
## 2   1145 4179245  2   2360530 323.36 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il grafico mostra l'intervallo di confidenza congiunto per i coefficienti di `dist` e `passen`. La regione ellittica rappresenta l'insieme dei valori plausibili per i coefficienti a un livello di confidenza del 95%. Il punto rappresenta le stime dei coefficienti del modello. Le linee tratteggiate indicano gli intervalli di confidenza per ciascun coefficiente. Poiché il punto non si trova all'interno della regione definita dagli intervalli di confidenza, possiamo rifiutare l'ipotesi nulla che entrambi i coefficienti siano zero.