



Università degli Studi di Salerno

Dipartimento di Ingegneria dell'Informazione ed Elettrica e  
Matematica Applicata (DIEM)

## Relazione di progetto

Sviluppo di un modello di regressione lineare su dataset

Corso di Statistica Applicata - A.A. 2024/25

### **Studenti Gruppo 16:**

Corradomaria Giachetta

Matricola: 0612708054

Francesco Peluso

Matricola: 0612707469

Gerardo Selce

Matricola: 0612707692

Anuar Zouhri

Matricola: 0612707505

### **Docenti:**

Prof. Fabio Postiglione

Prof. Paolo Addesso



Last update: 22 giugno 2025

# Indice

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Descrizione del dataset fornito</b>           | <b>2</b>  |
| <b>2</b> | <b>Analisi delle caratteristiche del dataset</b> | <b>3</b>  |
| 2.1      | Boxplot dei dati . . . . .                       | 3         |
| 2.2      | Analisi di normalità . . . . .                   | 4         |
| <b>3</b> | <b>Analisi della dipendenza tra le variabili</b> | <b>6</b>  |
| 3.1      | Analisi di correlazione . . . . .                | 6         |
| 3.2      | Analisi di regressione . . . . .                 | 7         |
| <b>4</b> | <b>Analisi dei modelli</b>                       | <b>8</b>  |
| 4.1      | Modello 1 . . . . .                              | 8         |
| 4.2      | Modello 2 . . . . .                              | 8         |
| 4.3      | Modello 3 . . . . .                              | 9         |
| 4.4      | Modello 4 . . . . .                              | 10        |
| <b>5</b> | <b>Scelta del modello</b>                        | <b>12</b> |



# 1 Descrizione del dataset fornito

A completezza del progetto si riporta la descrizione del dataset da analizzare. Il dataset contiene  $n = 100$  osservazioni, costituite da:

## Variabile dipendente

**y\_VideoQuality** → Qualità percepita del video

Tale indice è immaginato come frutto di una opportuna trasformazione di un punteggio assegnato a un campione di immagini da volontari che compilano un questionario. Esso sarà funzione di diverse caratteristiche proprie dei video, tra cui:

- la presenza o meno di rumore;
- la presenza o meno di *motion blur*;
- la nitidezza;
- la profondità di campo;
- la risoluzione;
- le aberrazioni ottiche visibili;
- la gamma dinamica;
- la fedeltà cromatica.

## Variabili indipendenti (regressori)

Sono delle quantità di cui l'operatore ha il controllo (parziale o totale) selezionando:

- l'attrezzatura video da utilizzare;
- i parametri di ripresa.

Rappresentano indici standardizzati:

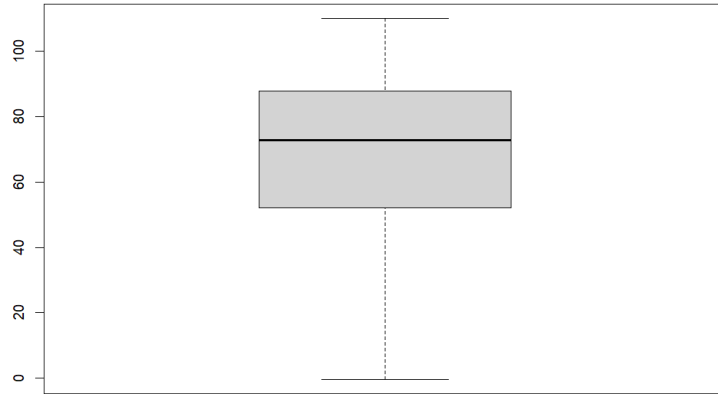
- **x1\_ISO** → ISO (sensibilità del sensore)
- **x2\_FRatio** → Rapporto Focale
- **x3\_Time** → Tempo di Esposizione (in relazione al frame rate utilizzato)
- **x4\_MP** → Megapixel del sensore
- **x5\_CROP** → Fattore di Crop
- **x6\_FOCAL** → Focale
- **x7\_PixDensity** → Densità di pixel

## 2 Analisi delle caratteristiche del dataset

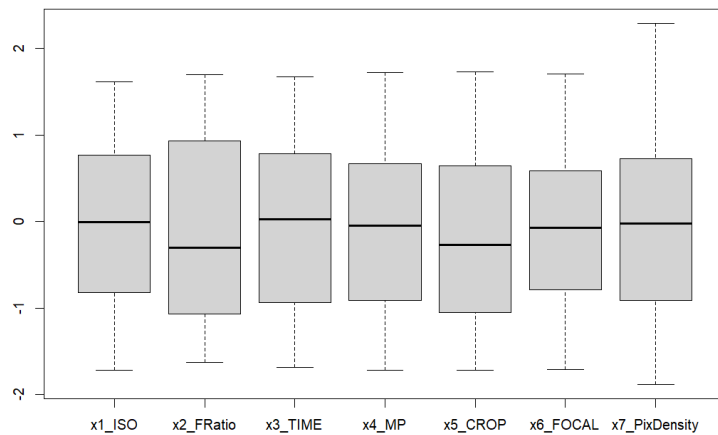
In questa fase preliminare si illustreranno le principali considerazioni fatte sul dataset fornito.

### 2.1 Boxplot dei dati

Si considerino i seguenti boxplot delle variabili del dataset.



(a) Boxplot della variabile dipendente  $y\_VideoQuality$



(b) Boxplot delle variabili indipendenti  $x_i$

Figura 1: Boxplot delle variabili considerate

Si osservi innanzitutto che i valori per ciascuna variabile sono tutti contenuti all'interno dell'intervallo interquartile e che quindi non sono presenti outliers. Per quel che riguarda la variabile dipendente  $y\_VideoQuality$  si è osservato che il valore della media e della mediana sono simili, infatti valgono rispettivamente  $media = 72.8135$ ,  $mediana = 68.6081$ . Si è osservato inoltre che i valori assunti dalla variabile  $x7\_PixDensity$  coprono un intervallo maggiore rispetto alle altre variabili indipendenti.

## 2.2 Analisi di normalità

Anche se non strettamente necessario ai fini del metodo di regressione, si è comunque deciso di verificare se qualcuna delle variabili indipendenti avesse una distribuzione normale. Tra i diversi qq-plot, si osserva che la variabile `x6_Focal` sembrerebbe avere una

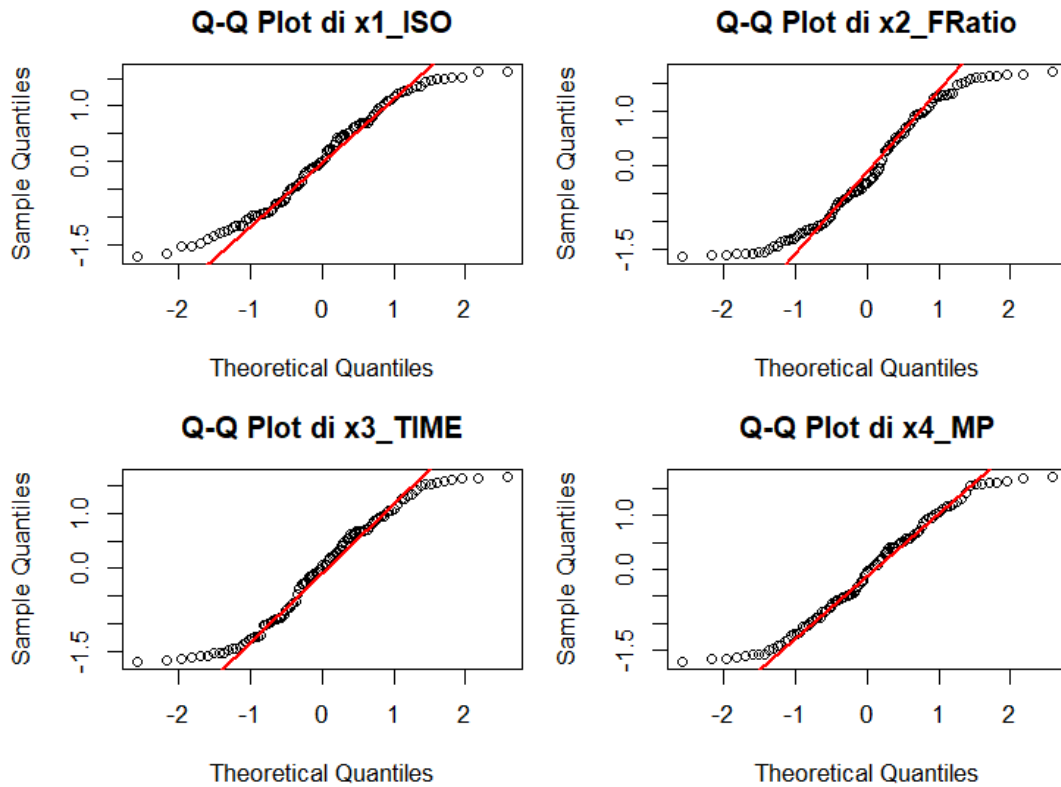
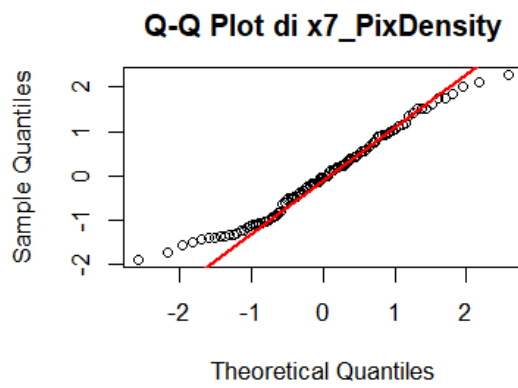
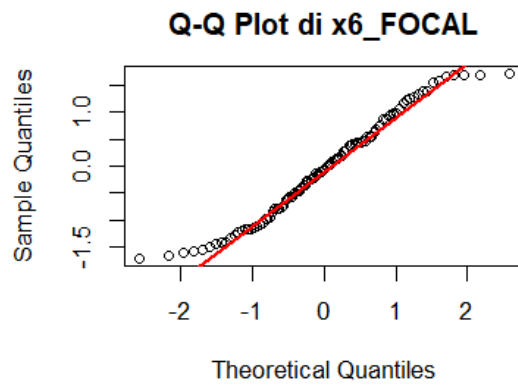
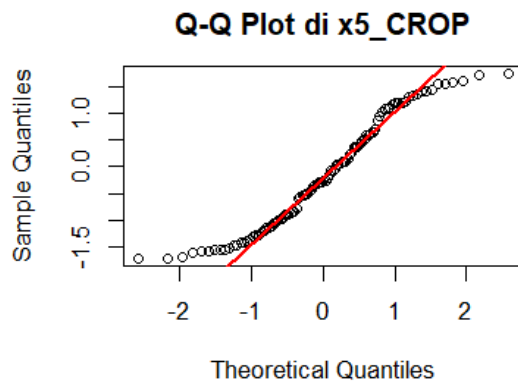


Figura 2:

distribuzione normale. Applicando il test di shapiro a questa variabile si ottiene

$$W = 0.97, p\text{-value} = 0.02.$$

Il valore di p-value ottenuto non si discosta molto da 0.05 e si potrebbe perciò supporre che la variabile sia distribuita come una normale.



### 3 Analisi della dipendenza tra le variabili

#### 3.1 Analisi di correlazione

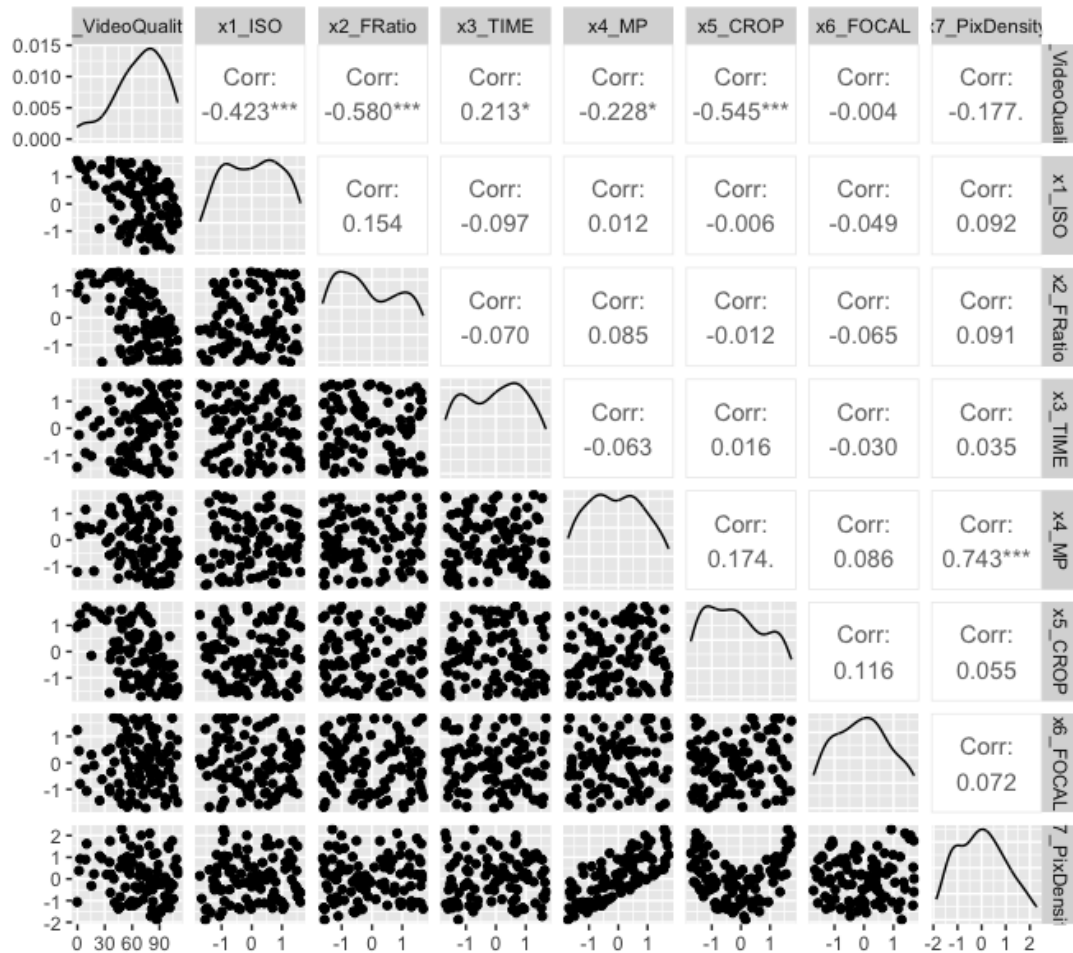


Figura 3: Scatter plot delle variabili presenti nel dataset.

Dalla Figura (3) notiamo, anche dal coefficiente di correlazione, una dipendenza lineare tra le variabili:

- **x4\_MP** e **x7\_PixDensity**

Invece notiamo la presenza di dipendenze non lineari che non vengono descritte dal coefficiente di correlazione. In particolare la notiamo tra le variabili:

- **y\_VideoQuality** e **x1\_ISO**
- **y\_VideoQuality** e **x2\_FRatio**
- **y\_VideoQuality** e **x3\_Time**
- **y\_VideoQuality** e **x5\_CROP**
- **x5\_CROP** e **x7\_PixDensity**

### 3.2 Analisi di regressione

Le dipendenze tra la variabile  $y\_VideoQuality$  e le diverse variabili indipendenti sono state analizzate attraverso una regressione semplice sulle singole variabili indipendenti.

| Variabile indipendente | p-value      |
|------------------------|--------------|
| x1_ISO                 | $1.17e - 05$ |
| x2_FRatio              | $2.63e - 10$ |
| x3_TIME                | $0.0331e$    |
| x4_MP                  | $0.0227$     |
| x5_CROP                | $4.39e - 09$ |
| x6_FOCAL               | $0.97$       |
| x7_PixDensity          | $0.0775$     |

Tabella 1: Sono rappresentati i p-value relativi alle regressioni delle singole variabili indipendenti al primo grado.

Diversamente da quanto ottenuto nell'analisi di correlazione, dalla Tabella (1) risultano rilevanti i regressori x1, x2, x3, x5. La stessa analisi è stata poi effettuata considerando anche i regressori al secondo ordine.

| Variabile indipendente | p-value      |
|------------------------|--------------|
| x1_ISO                 | $2.46e - 03$ |
| x2_FRatio              | $1.28e - 3$  |
| x3_TIME                | $0.3094$     |
| x4_MP                  | $0.2899$     |
| x5_CROP                | $0.368$      |
| x6_FOCAL               | $0.770$      |
| x7_PixDensity          | $0.8038$     |

Tabella 2: Sono rappresentati i p-value relativi alle regressioni delle singole variabili indipendenti al secondo grado.

Dalla Tabella (2) risulta evidente una dipendenza quadratica della variabile dipendente dai regressori x1, x2.



## 4 Analisi dei modelli

In questa sezione si analizzeranno differenti modelli e successivamente li si confronteranno verificando quale dei modelli meglio soddisfa l'ipotesi di normalità dei residui tramite dei grafici e test diagnostici. Inoltre, dato il numero non elevato di campioni si confronteranno i valori di AIC e di adjusted- $R^2$ .

### 4.1 Modello 1

Il primo modello analizzato è quello che include i regressori (di primo grado) più significativi (in base al valore di  $p\_value$  misurato precedentemente). Ovvero:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5.$$

La stima dei parametri ottenuti per questo modello è

| Parametro | Stima  | Dev. Std. |
|-----------|--------|-----------|
| $\beta_0$ | 65.62  | 1.30      |
| $\beta_1$ | -9.37  | 1.38      |
| $\beta_2$ | -13.33 | 1.24      |
| $\beta_3$ | 4.01   | 1.26      |
| $\beta_5$ | -14.52 | 1.26      |

Tabella 3: Stime dei coefficienti e deviazioni standard del modello

Gli intervalli di confidenza al 5%, ottenuti tramite il metodo `confint()` di R, sono:

| Parametro | Lower bound | Upper bound |
|-----------|-------------|-------------|
| $\beta_0$ | 63.04       | 68.20       |
| $\beta_1$ | -12.11      | -6.62       |
| $\beta_2$ | -15.79      | -10.87      |
| $\beta_3$ | 1.51        | 6.51        |
| $\beta_5$ | -17.01      | -12.03      |

Tabella 4: Intervalli di confidenza al 95% per i coefficienti del modello

I valori dell'adjusted  $R^2$  e AIC ottenuti sono:

$$R^2 = 0.77, \quad AIC = 514.69.$$

### 4.2 Modello 2

Il prossimo modello analizzato è quello ottenuto aggiungendo tutti i regressori più significativi con l'aggiunta di alcuni regressori al quadrato.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_3 + \beta_6 x_5$$

La stima dei parametri ottenuti per questo modello è

| Parametro | Stima  | Dev. Std. |
|-----------|--------|-----------|
| $\beta_0$ | 79.93  | 1.95      |
| $\beta_1$ | -8.66  | 1.05      |
| $\beta_2$ | -8.03  | 1.23      |
| $\beta_3$ | -13.49 | 0.94      |
| $\beta_4$ | -6.38  | 1.09      |
| $\beta_5$ | 3.94   | 0.95      |
| $\beta_6$ | -13.23 | 0.96      |

Tabella 5: Stime dei coefficienti e errori standard del modello

Gli intervalli di confidenza al 5%, ottenuti tramite il metodo `confint()` di R, sono:

| Parametro | Lower bound | Upper bound |
|-----------|-------------|-------------|
| $\beta_0$ | 76.06       | 83.80       |
| $\beta_1$ | -10.75      | -6.58       |
| $\beta_2$ | -10.48      | -5.58       |
| $\beta_3$ | -15.36      | -11.63      |
| $\beta_4$ | -8.55       | -4.22       |
| $\beta_5$ | 2.05        | 5.84        |
| $\beta_6$ | -15.14      | -11.32      |

Tabella 6: Intervalli di confidenza al 95% per i coefficienti del modello

I valori dell'adjusted  $R^2$  e AIC ottenuti sono:

$$R^2 = 0.87, \quad AIC = 460.76.$$

### 4.3 Modello 3

Questo modello è ottenuto tramite la funzione `step()` di R eliminando i regressori non significativi, a partire dal modello contenente tutti i regressori di primo grado. Il modello ottenuto è:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_5 + \beta_5 x_7$$

La stima dei parametri ottenuti per questo modello è:

| Parametro | Stima  | Dev. Std. |
|-----------|--------|-----------|
| $\beta_0$ | 65.65  | 1.29      |
| $\beta_1$ | -9.18  | 1.38      |
| $\beta_2$ | -13.17 | 1.23      |
| $\beta_3$ | 4.11   | 1.25      |
| $\beta_4$ | -14.41 | 1.25      |
| $\beta_5$ | -2.02  | 1.29      |

Tabella 7: Stime dei coefficienti e deviazioni standard del modello

Gli intervalli di confidenza al 5%, ottenuti tramite il metodo `confint()` di R, sono:

| Parametro | Lower bound | Upper bound |
|-----------|-------------|-------------|
| $\beta_0$ | 63.08       | 68.21       |
| $\beta_1$ | -11.92      | -6.45       |
| $\beta_2$ | -15.62      | -10.72      |
| $\beta_3$ | 1.62        | 6.59        |
| $\beta_4$ | -16.89      | -11.93      |
| $\beta_5$ | -4.58       | 0.54        |

Tabella 8: Intervalli di confidenza al 95% per i coefficienti del modello

I valori dell'adjusted  $R^2$  e AIC ottenuti sono:

$$R^2 = 0.77, \quad AIC = 514.11.$$

#### 4.4 Modello 4

L'ultimo modello analizzato, è ottenuto tramite la seguente istruzione R, adottando la funzione step():

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_1^2 + \beta_9 x_2^2 + \beta_{10} x_6^2 + \beta_{11} x_7^2 + \beta_{12} x_1 x_6 + \beta_{13} x_2 x_4 + \beta_{14} x_3 x_4 + \beta_{15} x_3 x_5 + \beta_{16} x_3 x_7 + \beta_{17} x_4 x_7.$$

In particolare il modello di partenza da cui si è partiti:

```
model_step_interactions <- lm(y_VideoQuality ~ (.)^2 + I(x1_IS0^2) +
I(x2_FRatio^2) + I(x3_TIME^2) + I(x4_MP^2) + I(x5_CROP^2) + I(x6_FOCAL^2) +
I(x7_PixDensity^2), data = data)
```

La stima dei parametri ottenuti per questo modello è:

| Parametro    | Stima  | Dev. Std. |
|--------------|--------|-----------|
| $\beta_0$    | 81.64  | 2.18      |
| $\beta_1$    | -8.77  | 1.00      |
| $\beta_2$    | -13.56 | 0.90      |
| $\beta_3$    | 4.31   | 1.03      |
| $\beta_4$    | -0.25  | 1.46      |
| $\beta_5$    | -13.37 | 0.92      |
| $\beta_6$    | 0.62   | 0.99      |
| $\beta_7$    | -2.96  | 1.60      |
| $\beta_8$    | -8.85  | 1.16      |
| $\beta_9$    | -6.57  | 1.01      |
| $\beta_{10}$ | -1.89  | 1.07      |
| $\beta_{11}$ | 2.91   | 1.86      |
| $\beta_{12}$ | -1.71  | 1.18      |
| $\beta_{13}$ | 1.66   | 0.99      |
| $\beta_{14}$ | -2.81  | 1.42      |
| $\beta_{15}$ | 2.83   | 0.99      |
| $\beta_{16}$ | 3.24   | 1.54      |
| $\beta_{17}$ | -3.55  | 2.25      |

Tabella 9: Stime dei coefficienti e deviazioni standard del modello

Gli intervalli di confidenza al 5%, ottenuti tramite il metodo `confint()` di R, sono:

| Parametro | L.B.   | U.B.   |
|-----------|--------|--------|
| $\beta_0$ | 77.29  | 85.99  |
| $\beta_1$ | -10.76 | -6.78  |
| $\beta_2$ | -15.34 | -11.77 |
| $\beta_3$ | 2.26   | 6.37   |
| $\beta_4$ | -3.16  | 2.65   |
| $\beta_5$ | -15.20 | -11.53 |
| $\beta_6$ | -1.34  | 2.59   |
| $\beta_7$ | -6.14  | 0.22   |
| $\beta_8$ | -11.16 | -6.55  |

| Parametro    | L.B.  | U.B.  |
|--------------|-------|-------|
| $\beta_9$    | -8.58 | -4.57 |
| $\beta_{10}$ | -4.01 | 0.23  |
| $\beta_{11}$ | -0.78 | 6.61  |
| $\beta_{12}$ | -4.05 | 0.64  |
| $\beta_{13}$ | -0.31 | 3.62  |
| $\beta_{14}$ | -5.65 | 0.02  |
| $\beta_{15}$ | 0.86  | 4.81  |
| $\beta_{16}$ | 0.19  | 6.30  |
| $\beta_{17}$ | -8.02 | 0.93  |

Tabella 10: Intervalli di confidenza al 95% per i coefficienti del modello (suddivisi in due tabelle)

I valori dell'adjusted  $R^2$  e AIC ottenuti sono:

$$R^2 = 0.89, \quad AIC = 448.27.$$



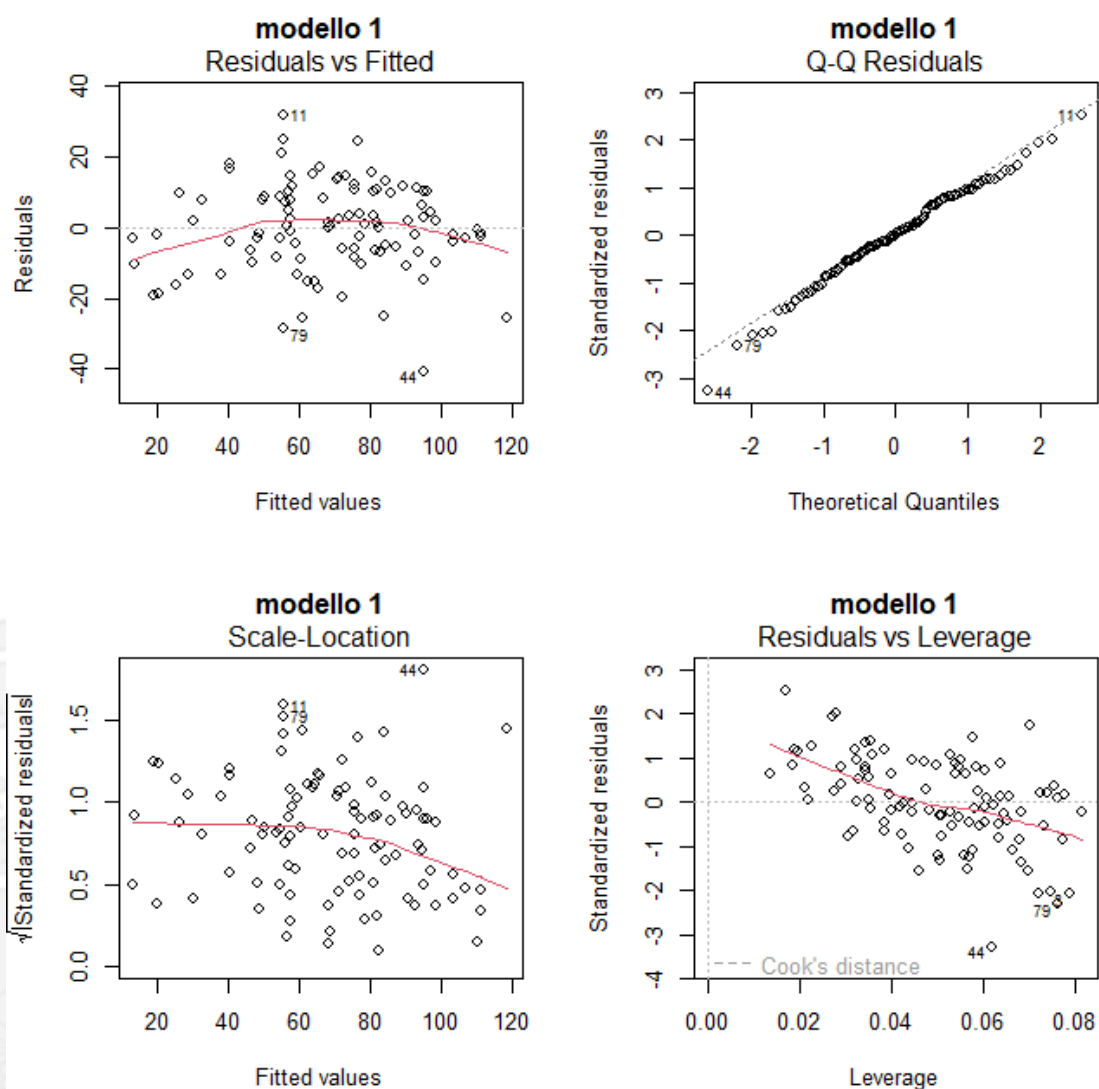
## 5 Scelta del modello

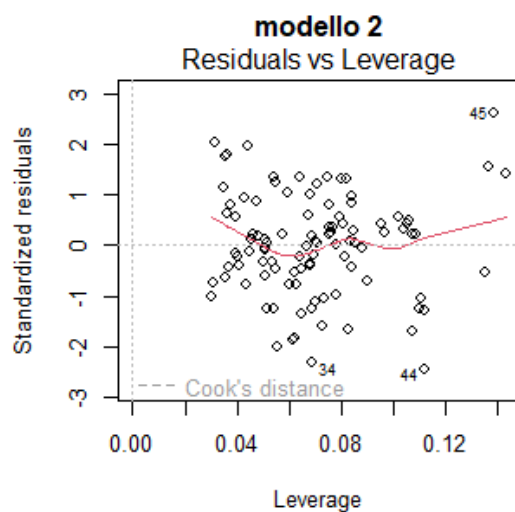
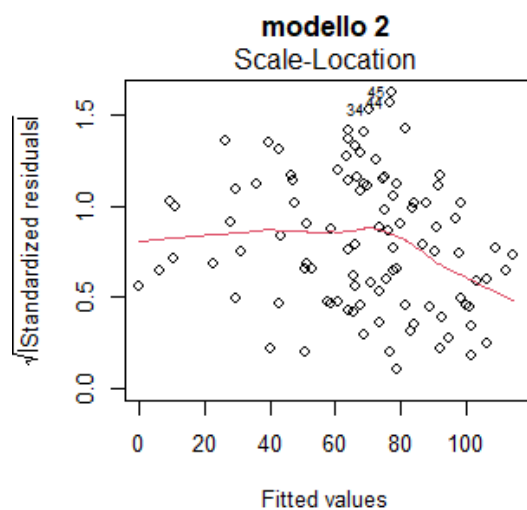
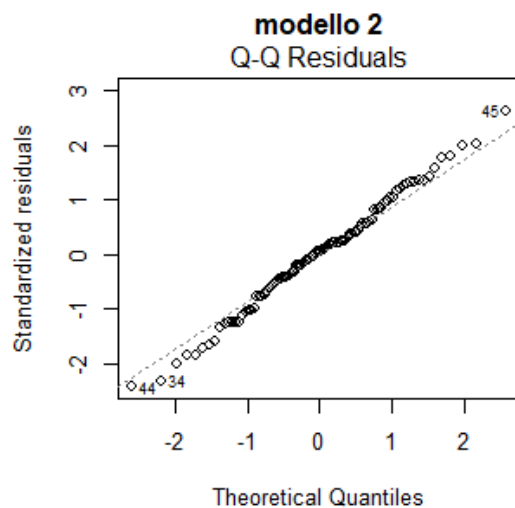
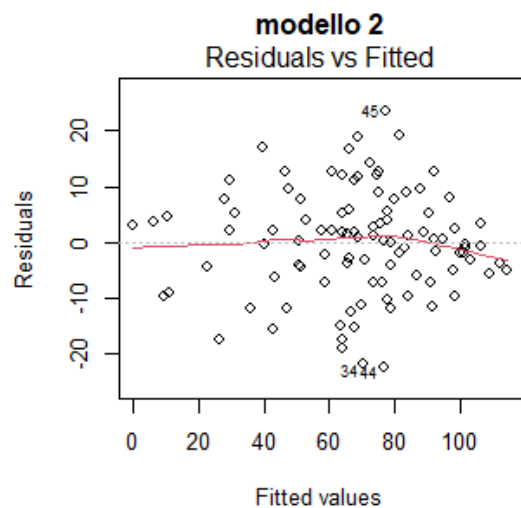
Si riportano i valori di  $R^2$  e AIC dei quattro modelli.

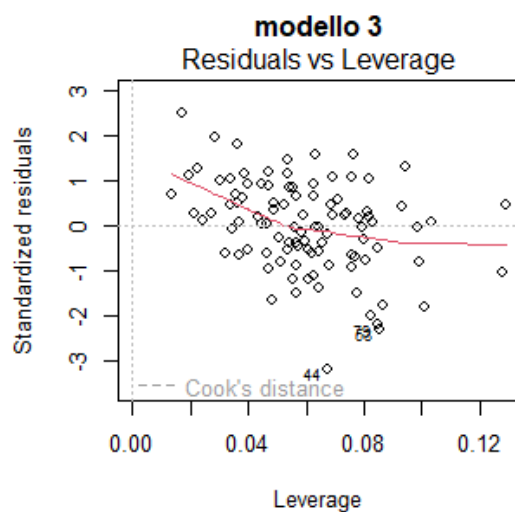
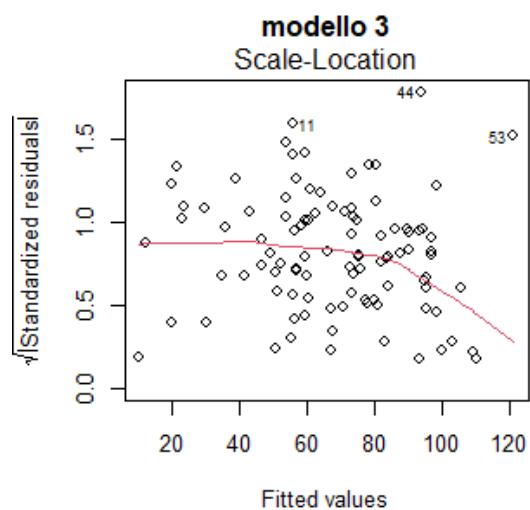
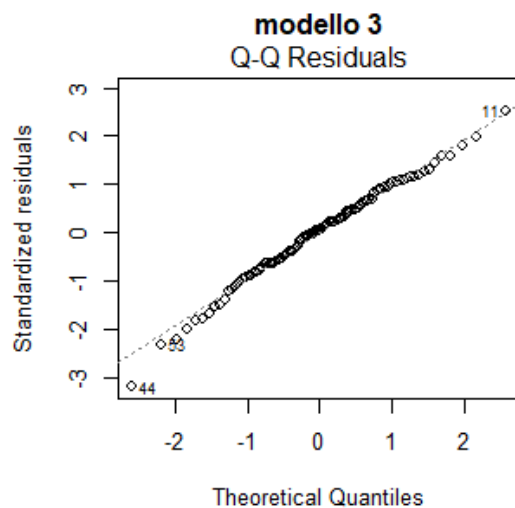
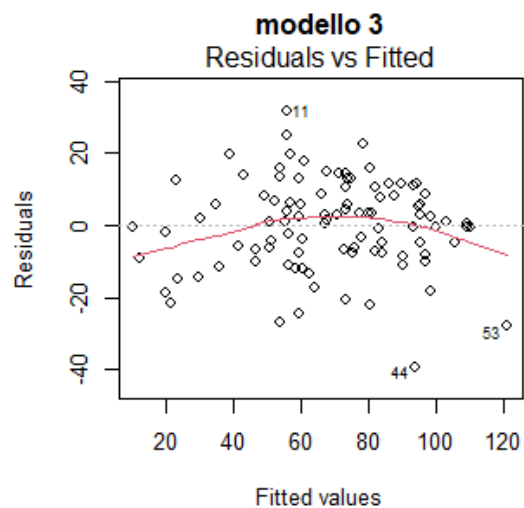
| Modello | adjusted $R^2$ | AIC    |
|---------|----------------|--------|
| 1       | 0.77           | 514.69 |
| 2       | 0.87           | 460.76 |
| 3       | 0.77           | 514.11 |
| 4       | 0.89           | 448.27 |

Tabella 11: Valori di  $R^2$  e AIC per i quattro modelli

Di seguito vengono mostrati i grafici diagnostici ottenuti sui quattro modelli.







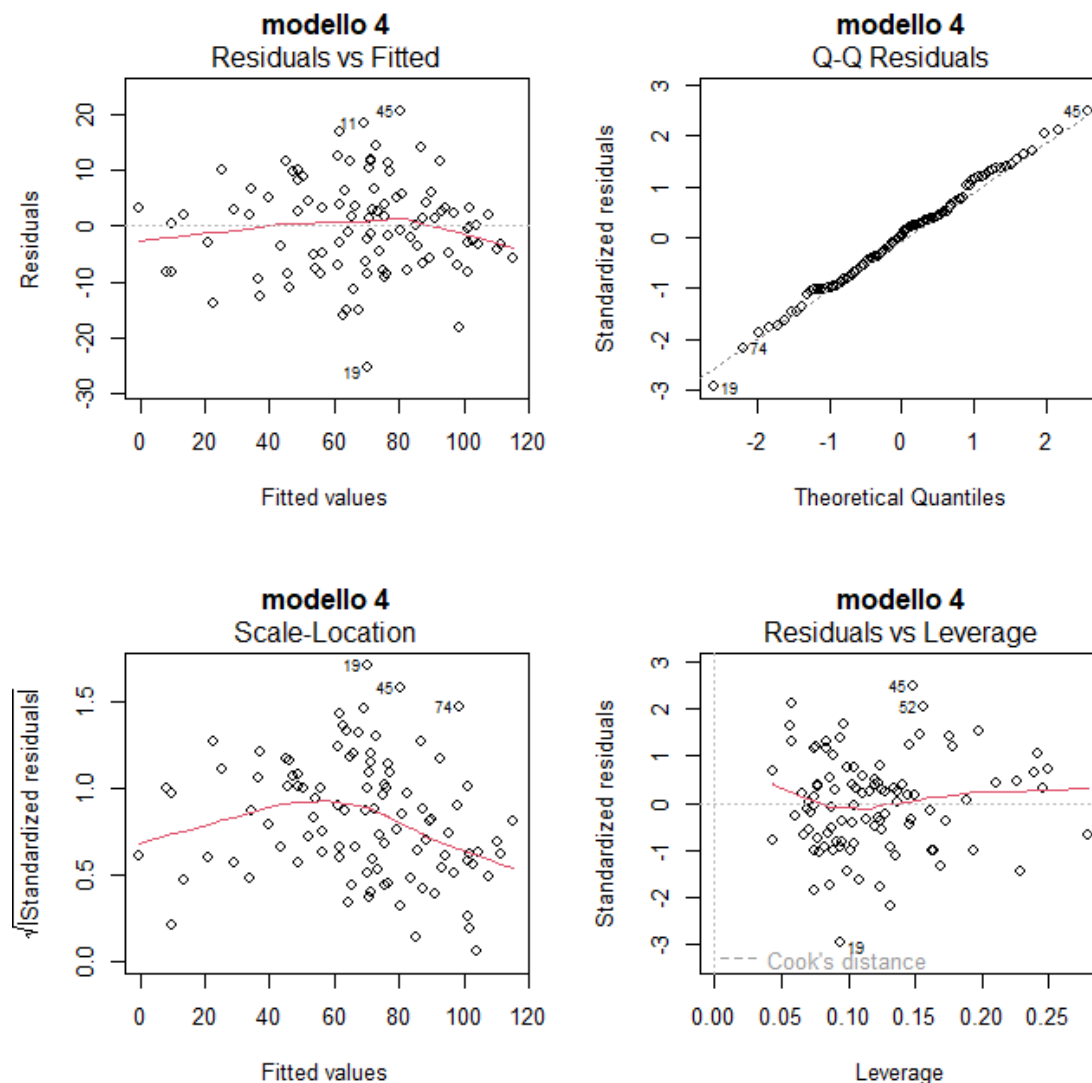


Figura 4: Residuals vs Fitted e Q-Q Residuals dei quattro modelli.

Osservando i grafici 'Residuals vs Fitted' si nota che solo nei modelli 2 e 4, la linea rossa non presenta alcun pattern soddisfacendo in buona maniera l'ipotesi di linearità. Inoltre, sempre i modelli 2 e 4 nei grafici 'Q-Q Residuals' l'ipotesi di normalità sembra essere soddisfatta.

Si osservi (dal grafico 'Scale-Location') che però su nessuno dei modelli considerati si può supporre che la varianza sia costante.

Infine comparando i valori di adjusted  $R^2$  e AIC, il modello 4 sarebbe da preferire. Infatti, usando l'AIC, si sceglie il modello che ha valore minore; un valore maggiore di  $R^2$  implica che il modello è in grado di interpretare meglio il fenomeno osservato.

A fronte dei dati ricavati si è stimato che il modello che meglio rappresenta il dataset fornito è il modello 4.