Progetto statistica

Carlo Attanasio
Leo Filipović Grčić
Giorgia Bianchi

Dataset

Introduzione e obiettivo

Descrizione dei valori

Analisi del dataset

Introduzione e obiettivo

In questo progetto vogliamo studiare i fattori che influenzano il prezzo di vendita di un certo prodotto. Il dataset utilizzato si riferisce alle vendite di sedili per automobili, principalmente negli Stati Uniti e contiene 400 osservazioni.

Variabile	Descrizione	Tipo
Price	È la nostra variabile risposta, indica il prezzo di vendita del rivenditore al cliente	Numerica
CompPrice	Prezzo di vendita ad ogni rivenditore	Numerica
Income	Ricchezza media della zona in cui il prodotto è stato venduto	Numerica
Advertising	Budget destinato alla pubblicità per un particolare rivenditore	Numerica
Sales	Unità di prodotto vendute	Numerica
Population	Numero di abitanti nella zona del rivenditore	Numerica
Age	Età media della popolazione nella zona del rivenditore	Numerica
Education	Indice di educazione media della popolazione nella zona del rivenditore	Numerica
ShelveLoc	Bad, Good, Medium. Indica quanto il prodotto è esposto al cliente da ogni rivenditore	Categorica
Urban	Indica se il rivenditore si trova in una zona urbana	Categorica

Descrizione dei valori

```
> str(data)
```

```
$ Sales       : num [1:400] 9.5 11.22 10.06 7.4 4.15 ...
$ CompPrice   : num [1:400] 138 111 113 117 141 124 115 136 132 132 ...
$ Income      : num [1:400] 73 48 35 100 64 113 105 81 110 113 ...
$ Advertising: num [1:400] 11 16 10 4 3 13 0 15 0 0 ...
$ Population : num [1:400] 276 260 269 466 340 501 45 425 108 131 ...
$ Price       : num [1:400] 120 83 80 97 128 72 108 120 124 124 ...
$ ShelveLoc   : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
$ Age         : num [1:400] 42 65 59 55 38 78 71 67 76 76 ...
$ Education   : num [1:400] 17 10 12 14 13 16 15 10 10 17 ...
$ Urban       : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
```

Analisi del dataset

Alta correlazione

- Price, Sales
- Price, CompPrice
- Advertising, Sales
- Advertising, Population
- Age, Sales

Modello lineare

Goodness of fit

Analisi dei residui

Goodness of fit

```
> model <- lm(Price ~ . - Urban - ShelveLoc, data)
> summary(model)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-41.563  -9.691   0.202   8.707  42.210

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.939539   8.436043   4.853 1.76e-06 ***
Sales       -4.983530   0.272120 -18.314  < 2e-16 ***
CompPrice    0.953481   0.046936  20.314  < 2e-16 ***
Income       0.058395   0.025783   2.265   0.0241 *
Advertising  0.751801   0.115014   6.537 1.96e-10 ***
Population   0.001767   0.005047   0.350   0.7264
Age         -0.257487   0.045313  -5.682 2.59e-08 ***
Education   -0.193039   0.272699  -0.708   0.4794
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.16 on 392 degrees of freedom
Multiple R-squared:  0.6488,  Adjusted R-squared:  0.6425
F-statistic: 103.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

Primo modello senza variabili categoriche

- Residui simmetrici
- Molte covariate significative
- R^2 accettabile
- p-value molto basso

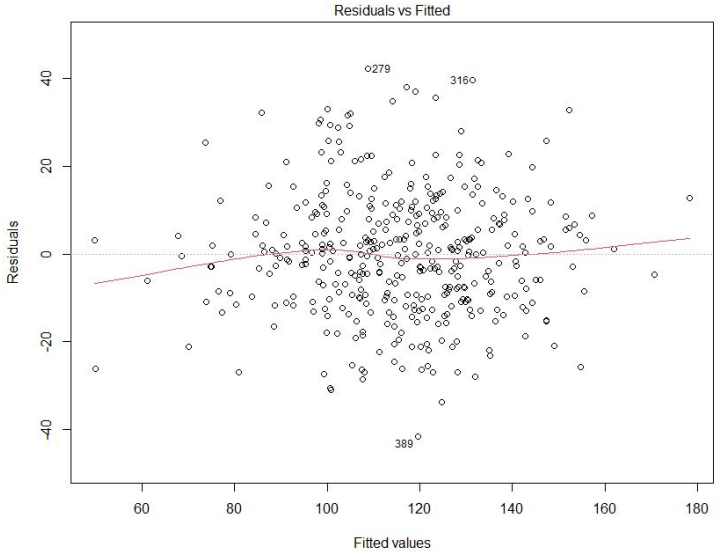
```
> AIC(model)
```

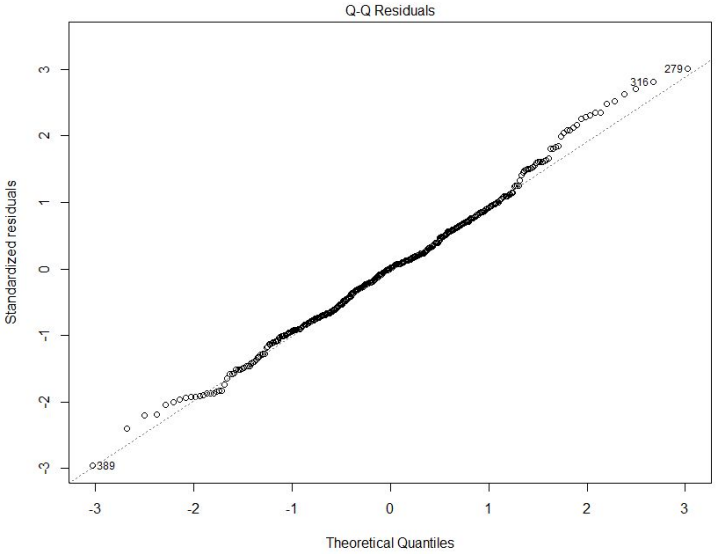
3265.224

Analisi dei residui

```
> shapiro.test(model$res)
```

```
Shapiro-Wilk normality test
W = 0.99314, p-value = 0.06501
```





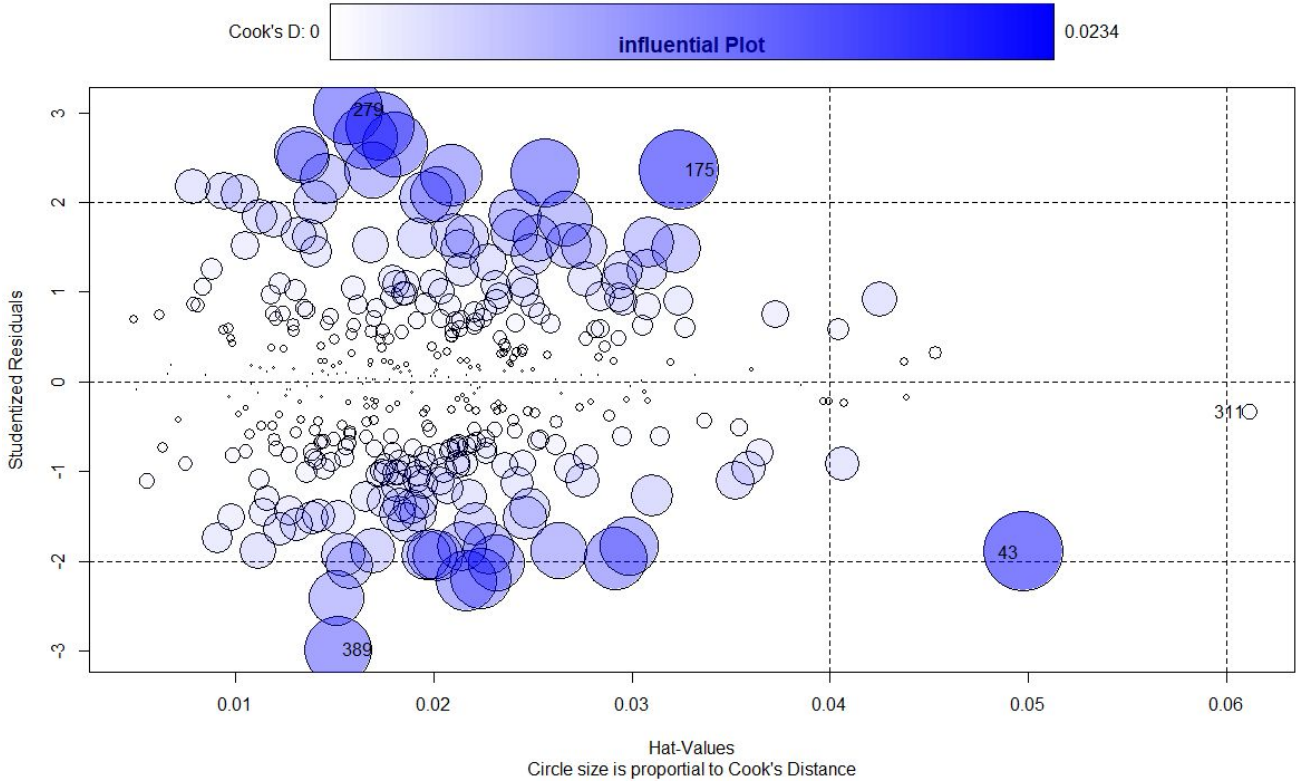
Punti influenti

Influence plot

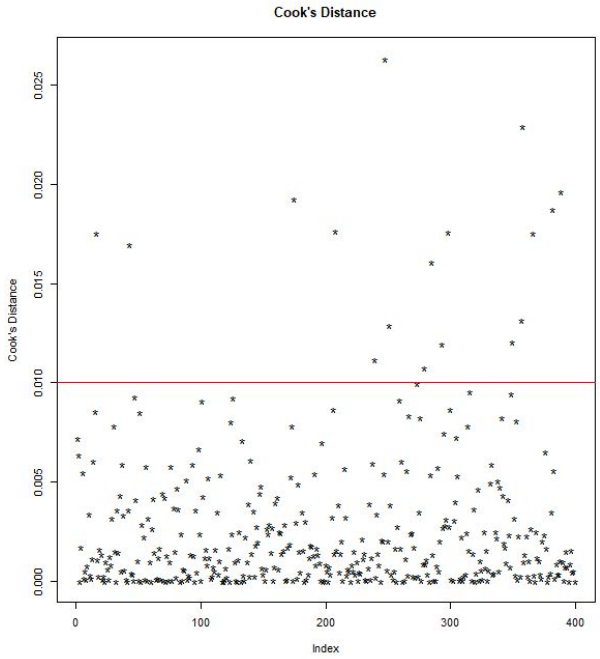
Cook's distance, leverage and residuals plot

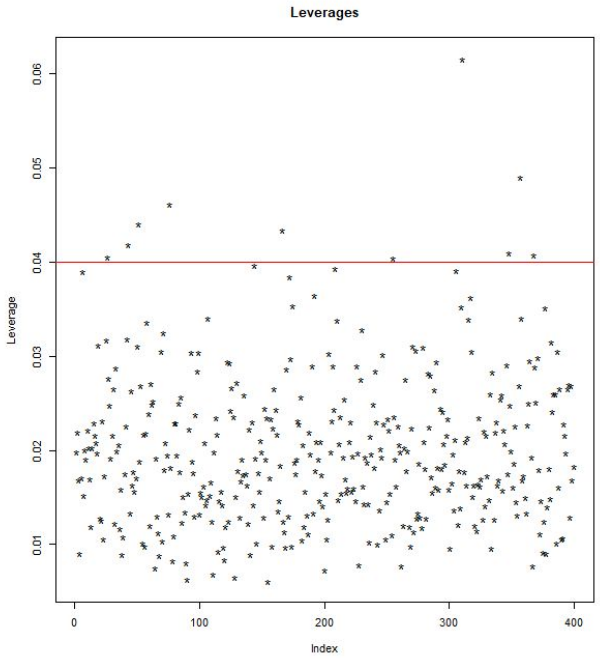
Analisi dei punti influenti

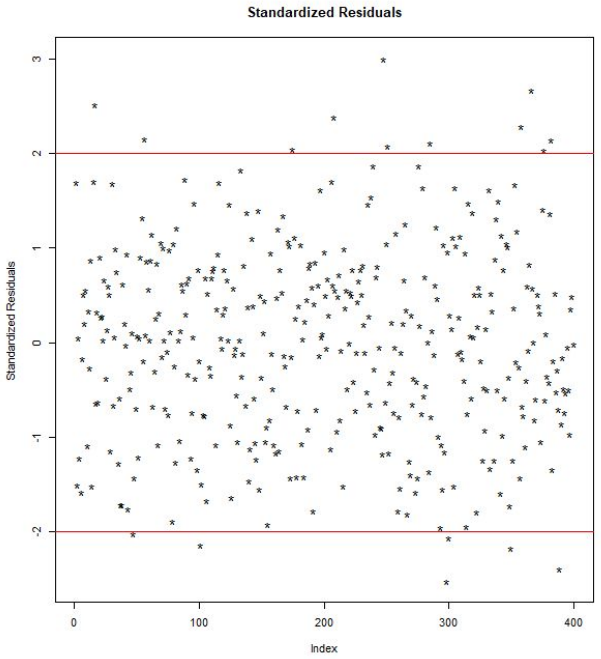
Influence plot



Cook's distance, leverages, standardized residuals







Analisi dei punti influenti

Esempi di punti influenti

```
> data[16, ]
$ Sales       : 8.71
$ CompPrice   : 149
$ Income      : 95
$ Advertising: 5
$ Population  : 400
$ Price       : 144
$ ShelveLoc   : Medium
$ Age         : 76
$ Education   : 18
$ Urban       : No
$ US          : No
```

Reddito e popolazione elevati

Vendite basse

```
> data[43, ]
$ Sales       : 10.4
$ CompPrice   : 77
$ Income      : 69
$ Advertising: 0
$ Population  : 25
$ Price       : 24
$ ShelveLoc   : Medium
$ Age         : 50
$ Education   : 18
$ Urban       : Yes
$ US          : No
```

Vendite alte

Pubblicità e popolazione bassa

Manteniamo i punti influenti

- **Rappresentatività del campione**: Escludere i punti influenti potrebbe ridurre la rappresentatività del dataset, specialmente se questi punti riflettono condizioni rare ma reali.
- **Robustezza del modello**: Mantenendo i punti influenti, possiamo valutare se il modello è robusto e in grado di gestire variabilità reale nei dati.

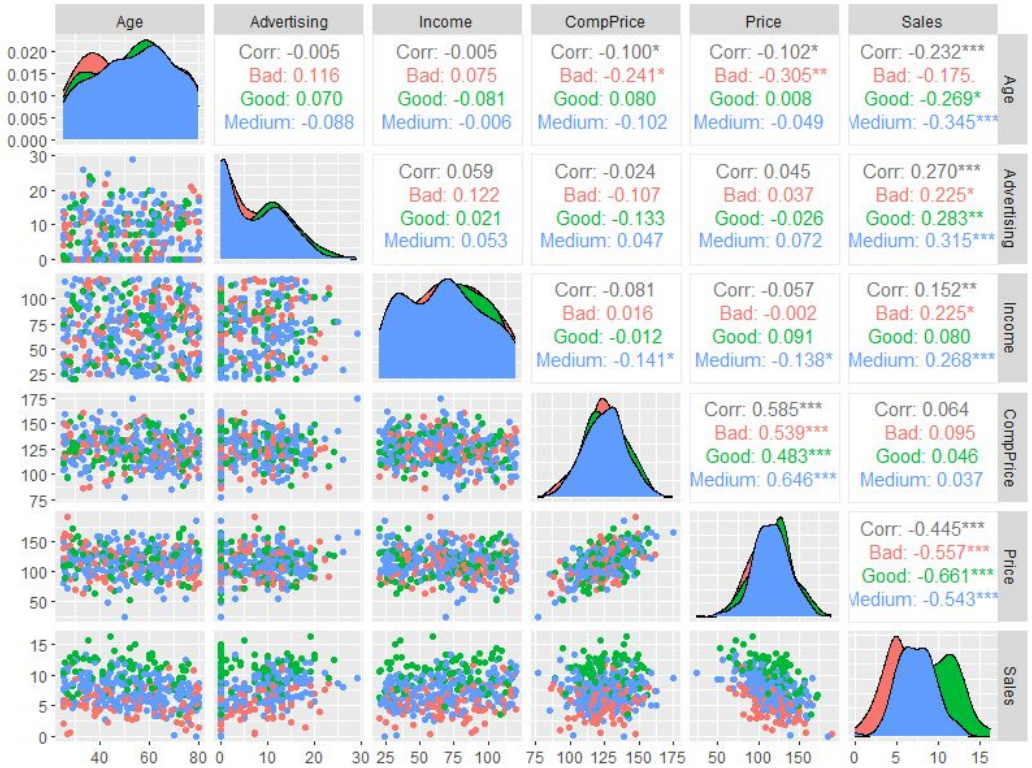
Analisi della varianza (ANOVA)

Variabili categoriche

- *ShelveLoc*
- *Urban*

Update del modello lineare

Variabili categoriche - *ShelveLoc*



Forte dipendenza tra la variabile categorica *ShelveLoc* e la variabile *Sales*.

Dal momento che *Sales* è fortemente correlata con *Price*, ci aspettiamo che sia influente nel modello.

ANOVA - *ShelveLoc*

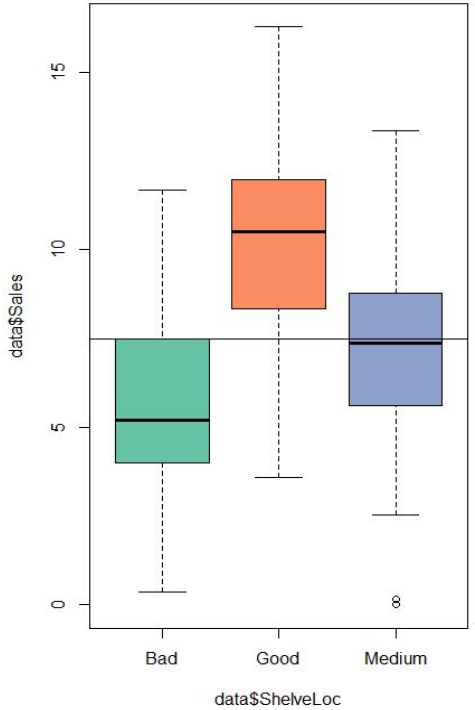
```
> anova_model <- aov(Sales ~ ShelveLoc, data)
> summary(anova_model)
```

Test ANOVA $\begin{cases} \mathcal{H}_0: & \mu_0 = \mu_1 = \cdots = \mu_k \\ \mathcal{H}_1: & \exists i, j : \mu_i \neq \mu_j \end{cases}$

Aggiungendo la variabile *ShelveLoc* otteniamo la seguente tabella ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
ShelveLoc	2	1010	504.8	92.23	<2e-16	***
Residuals	397	2173	5.5			

Il p-value del test ANOVA è basso, quindi possiamo rifiutare l'ipotesi nulla. Includiamo *ShelveLoc* nel modello.



Ipotesi per il test ANOVA - *ShelveLoc*

Controllo delle ipotesi per il test ANOVA

- Normalità dei gruppi

```
> Ps = c(shapiro.test(data$Sales[data$ShelveLoc == "Bad"])$p,
         shapiro.test(data$Sales[data$ShelveLoc == "Medium"])$p,
         shapiro.test(data$Sales[data$ShelveLoc == "Good"])$p)
Ps = 0.6065787, 0.3139415, 0.8793911
```

- Omoschedasticità tra i gruppi

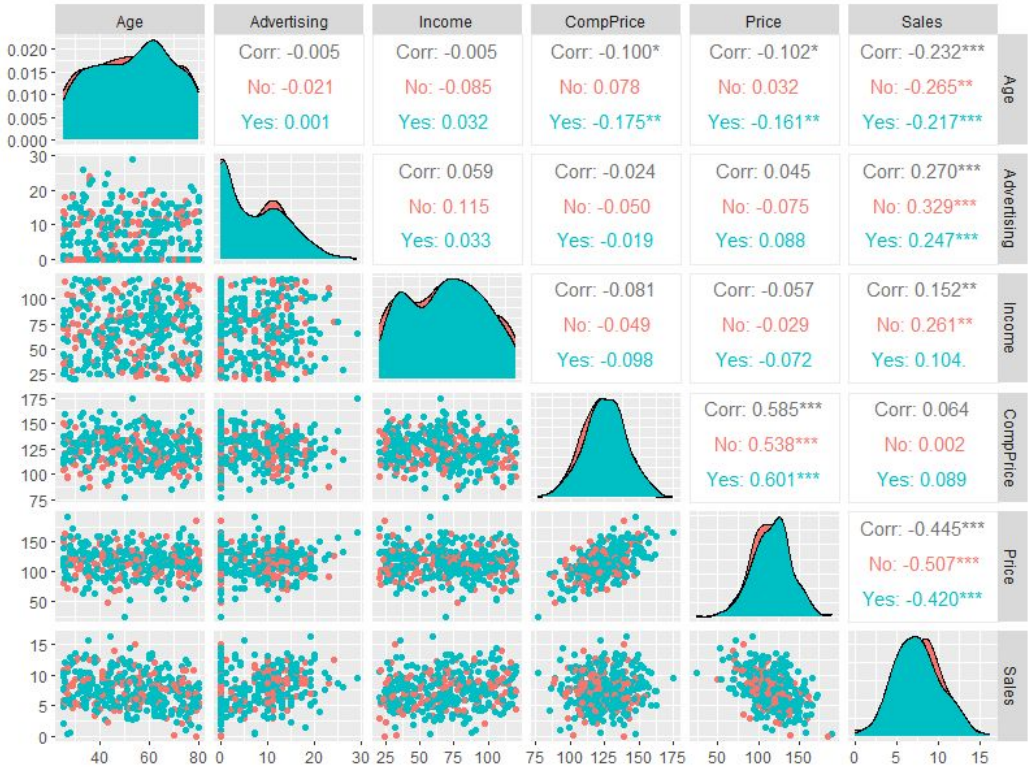
```
> bartlett.test(data$Sales, data$ShelveLoc)
        Bartlett test of homogeneity of variances

data:  data$Sales and data$ShelveLoc
Bartlett's K-squared = 1.2208, df = 2, p-value = 0.5431

> leveneTest(data$Sales, data$ShelveLoc)
Levene's Test for Homogeneity of Variance (center = median)
         Df    F value    Pr(>F)
group     2     0.8088    0.4461
        397
```

Le ipotesi sono soddisfatte, quindi posso aggiungere la variabile categorica al modello.

Variabili categoriche - *Urban*



Non sembrerebbe esserci nessuna differenza sostanziale tra le due categorie.

***Urban* non risulta essere influente nel modello.**

Update del modello lineare

```
> model <- update(model, Price ~ . + ShelveLoc, data)
> summary(model)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-23.8457  -6.7042   0.4552   6.2350  27.9166

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      46.267734   5.605451   8.254 2.39e-15 ***
Sales            -8.026320   0.224890 -35.690  < 2e-16 ***
CompPrice         0.956841   0.031018  30.848  < 2e-16 ***
Income            0.123978   0.017311   7.162 4.00e-12 ***
Advertising       0.957920   0.076552  12.513  < 2e-16 ***
Population        0.003176   0.003338   0.952    0.342
Age              -0.383607   0.030510 -12.573  < 2e-16 ***
Education        -0.153790   0.180213  -0.853    0.394
ShelveLocGood    39.247704   1.748426  22.447  < 2e-16 ***
ShelveLocMedium  15.812270   1.229196  12.864  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.355 on 390 degrees of freedom
Multiple R-squared:  0.8474,  Adjusted R-squared:  0.8439
F-statistic: 240.6 on 9 and 390 DF,  p-value: < 2.2e-16
```

Modello con aggiunta delle variabili categoriche

- Residui circa simmetrici
- Molte covariate significative
- R^2 buono
- p-value molto basso

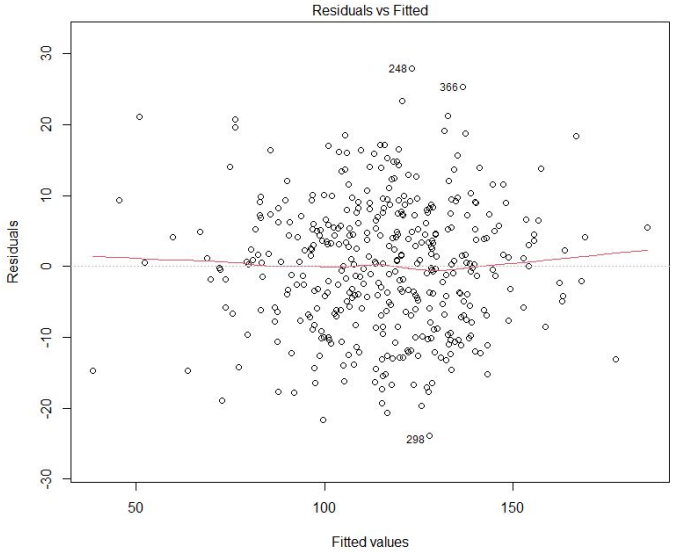
```
> AIC(model)
```

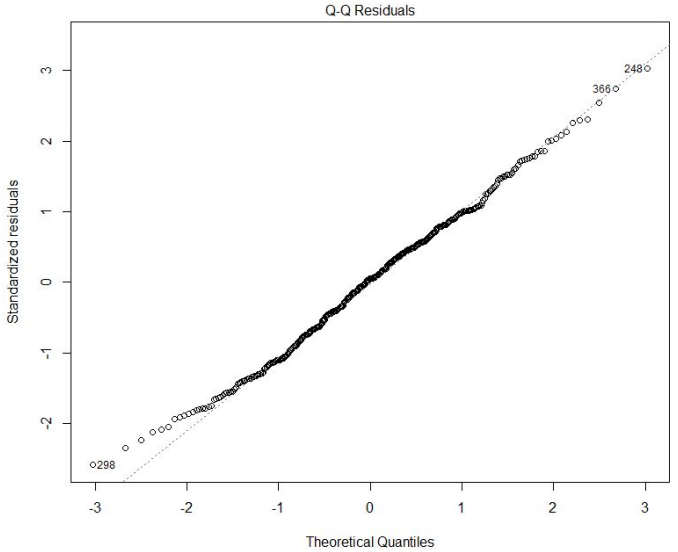
2935.755

Update del modello lineare - Analisi dei residui

```
> shapiro.test(model$res)
```

```
Shapiro-Wilk normality test
W = 0.99581, p-value = 0.3665
```





Selezione delle covariate

Collinearità e VIF

Selezione delle variabili più significative

Analisi dei residui del nuovo modello

Calcolo del Variance Inflation Factor

```
> vif(model)
```

	GVIF	Df	GVIF^(1/(2*Df))
Sales	1.837543	1	1.355560
CompPrice	1.021632	1	1.010758
Income	1.066169	1	1.032555
Advertising	1.102532	1	1.050015
Age	1.110891	1	1.053988
ShelveLoc	1.588323	2	1.122625

Tutte le covariate hanno GVIF sufficientemente vicino a 1, quindi possiamo concludere che non ci sono problemi di collinearità nel nostro modello.

Update del modello lineare

```
> summary(model)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-23.8457  -6.7042   0.4552   6.2350  27.9166

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      46.267734   5.605451   8.254 2.39e-15 ***
Sales            -8.026320   0.224890 -35.690  < 2e-16 ***
CompPrice         0.956841   0.031018  30.848  < 2e-16 ***
Income            0.123978   0.017311   7.162 4.00e-12 ***
Advertising       0.957920   0.076552  12.513  < 2e-16 ***
Population        0.003176   0.003338   0.952    0.342
Age              -0.383607   0.030510 -12.573  < 2e-16 ***
Education        -0.153790   0.180213  -0.853    0.394
ShelveLocGood    39.247704   1.748426  22.447  < 2e-16 ***
ShelveLocMedium  15.812270   1.229196  12.864  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.355 on 390 degrees of freedom
Multiple R-squared:  0.8474,  Adjusted R-squared:  0.8439
F-statistic: 240.6 on 9 and 390 DF,  p-value: < 2.2e-16
```

Notiamo che ci sono alcune covariate non particolarmente significative:

- *Population*
- *Education*

Selezione delle covariate

```
> model <- step(model, direction = "both")
> summary(model)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-23.352  -6.626   0.285   6.311  27.765

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      45.35149    4.75990   9.528  < 2e-16 ***
Sales            -8.02319    0.22475 -35.699  < 2e-16 ***
CompPrice         0.95336    0.03086  30.890  < 2e-16 ***
Income            0.12413    0.01728   7.185 3.41e-12 ***
Advertising       0.97808    0.07393  13.230  < 2e-16 ***
Age              -0.38509    0.03046 -12.641  < 2e-16 ***
ShelveLocGood    39.22310    1.74755  22.445  < 2e-16 ***
ShelveLocMedium  15.76152    1.22781  12.837  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.353 on 392 degrees of freedom
Multiple R-squared:  0.8467,  Adjusted R-squared:  0.844
F-statistic: 309.3 on 7 and 392 DF,  p-value: < 2.2e-16
```

Modello dopo aver selezionato le covariate più significative

- Residui simmetrici
- Tutte le covariate significative
- R^2 buono
- p-value molto basso

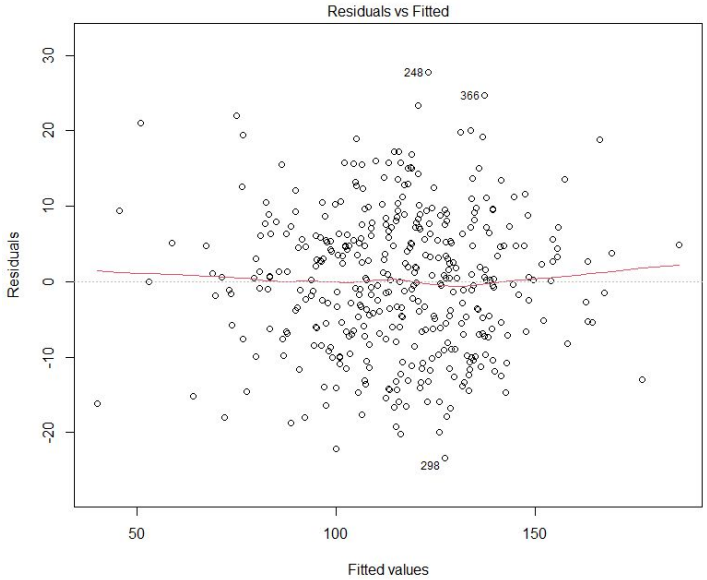
```
> AIC(model)
```

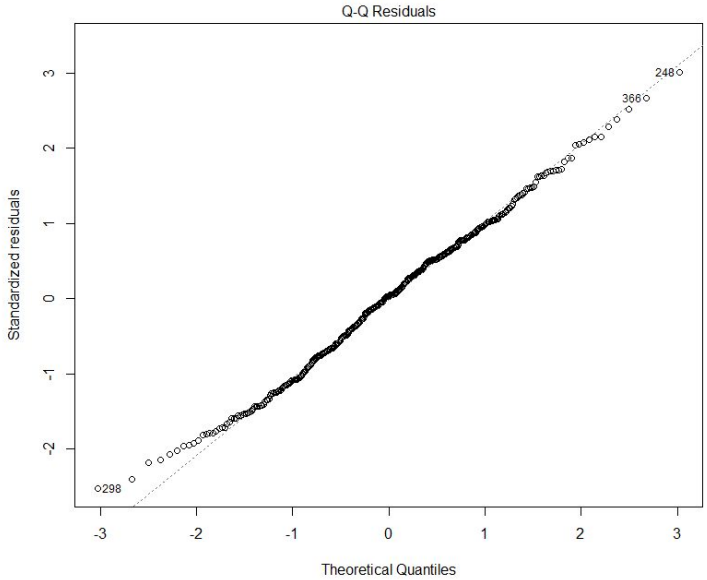
2933.616

Selezione delle covariate – Analisi dei residui

```
> shapiro.test(model$res)
```

```
Shapiro-Wilk normality test
W = 0.99575, p-value = 0.3545
```





Prediction & Cross-validation

Predizione su metà del dataset

K-fold cross-validation

Prediction

Separo il dataset in due gruppi e effettuo la prediction

```
> train = sample(nrow(data), floor(nrow(data)/2))
> data_train = data[train,]
> data_test = data[-train,]

> predict_model = update(model, . ~ ., data = data_train)
> mean((data_test$Price - predict(predict_model, data_test))^2)
```

MSE = 84.73227

```
> mean((model$residuals)^2)
```

MSE = 85.72649

L'errore del modello creato su *data_train* non è molto diverso dall'errore del modello con dataset completo.
Questa stima di errore di predizione non è molto affidabile.

Cross-validation

Trasformo il mio modello lineare in un GLM per poter applicare la funzione `cv.glm()` *per la cross-validation*

```
> glm.fit = glm(Price ~ Sales + CompPrice + Income + Advertising +
                        ShelveLoc + Age, data = data)
> cv.error = cv.glm(data,glm.fit, K = 5 )$delta[1]
```

```
cv.error = 89.35949
```

```
> MSE = mean((data$Price-predict(model, data))^2)
```

```
MSE = 85.72649
```

L'errore della cross-validation è abbastanza simile al MSE del nostro modello, quindi:

- **Non c'è overfitting**
- **Generalizza bene**
- **Performance consistente**

Conclusione

```
> summary(model)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-23.352  -6.626   0.285   6.311  27.765

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      45.35149    4.75990   9.528  < 2e-16 ***
Sales            -8.02319    0.22475 -35.699  < 2e-16 ***
CompPrice         0.95336    0.03086  30.890  < 2e-16 ***
Income            0.12413    0.01728   7.185 3.41e-12 ***
Advertising       0.97808    0.07393  13.230  < 2e-16 ***
Age              -0.38509    0.03046 -12.641  < 2e-16 ***
ShelveLocGood    39.22310    1.74755  22.445  < 2e-16 ***
ShelveLocMedium  15.76152    1.22781  12.837  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.353 on 392 degrees of freedom
Multiple R-squared:  0.8467,  Adjusted R-squared:  0.844
F-statistic: 309.3 on 7 and 392 DF,  p-value: < 2.2e-16
```

Il modello finale indica che il prezzo di vendita del prodotto cresce se la zona di vendita è ricca, giovane e viene investito molto in pubblicità.

I prodotti bene esposti vengono venduti ad un prezzo più alto.

Il numero di unità vendute aumenta se il prezzo scende.

Non c’è differenza significativa tra zone rurali e urbane.