Universita' degli studi di Salerno

Statistica Applicata

Analisi di un dataset in R

Autori:
Mattia De Bartolomeis
Marco Cerino
Christian De Angelis
Mario Della Corte

Docente:
Fabio Postiglione

Febbraio,2023

INDICE

CAPITOLO 1

INTRODUZIONE

Il dataset utilizzato per l'analisi dei dati è composto da 100 campioni ricavati da questionari compilati da volontari sulle caratteristiche proprie delle immagini. In particolare :

1. **Variabile dipendente:** y_ImageQuality – Qualita' dell'immagine percepita.
2. **Variabili indipendenti (regressori):**
   x1_ISO – ISO (sensibilità del sensore)
   x2_FRatio – Rapporto Focale
   x3_Time – Tempo di Esposizione
   x4_MP – Megapixel del sensore
   x5_CROP – Fattore di Crop
   x6_FOCAL – Focale
   x7_PixDensity – Densita' di pixel

I valori assunti dai vari regressori sono standardizzati.

L'obiettivo di questo progetto è determinare un modello di regressione lineare ottimo sulla qualità dell'immagine attraverso le procedure analizzate durante il corso.

Ogni capitolo consiste di una premessa teorica sull'argomento seguita da una sezione relativa al commento del codice sorgente scritto in R.

CAPITOLO 2

STATISTICA DESCRITTIVA

2.1 Elementi teorici della statistica descrittiva

La statistica descrittiva è la branca della statistica che studia i criteri di rappresentazione dei dati sperimentali e la sintesi dell'informazione in essi contenuta.
L'insieme di elementi che caratterizza l'oggetto di studia è detto **popolazione**. Spesso non tutti i valori relativi alla popolazione in esame sono disponibili, a causa delle grandi dimensioni che quest'ultima può assumere. In questi casi la procedura consiste nell'esaminare solo una parte della popolazione detto "campione" e la procedura con la quale viene selezionato è detto **"campionamento"**. Esistono diversi tipi di campionamento tra cui il campionamento "casuale".
Nel caso di una popolazione fisica costituita da un numero finito di elementi, tale procedura si basa sul principio che "ogni elemento presente nella popolazione ha la stessa probabilità di essere selezionato".
Nel caso di una popolazione con un numero concettualmente infinito di elementi, si è in presenza di un campionamento casuale se la procedura di selezione è riconducibile allo schema degli esperimenti ripetuti indipendenti, cioè se il risultato di ogni singola selezione non dipende in alcun modo dai risultati precedenti.

2.1.1 Distribuzione di frequenza e frequenza cumulata

Uno strumento molto utile per sintetizzare le informazioni contenute in una popolazione o in un campione è la "distribuzione di frequenza cumulata o "distribuzione di frequenza". Il modo di procedere per la determinazione della distribuzione di frequenza della caratteristica di interesse dipende dal tipo di dati con cui si lavora; questi possono essere classificati come:

1. Dati di tipo qualitativo, cioè quei dati tra i quali è possibile stabilire solo una relazione di equivalenza ma non una relazione d'ordine.
2. Dati di tipo qualitativo-ordinato: è possibile stabilire sia una relazione d'ordine che di equivalenza.
3. Dati di tipo quantitativo discreto, che appartengono all'insieme dei numeri naturali.
4. Dati di tipo quantitativo continuo, che appartengono all'insieme dei numeri reali.

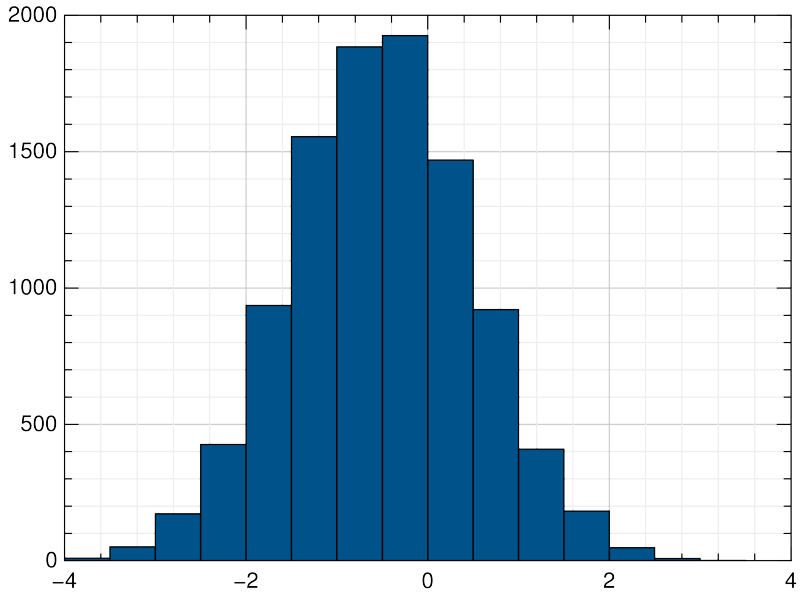
Nei primi tre casi,si procede contando il numero di elementi appartenenti a ciascuna classe. Tale numero prende il nome di **frequenza nella classe**. Se si divide la frequenza per il numero di elementi nella popolazione o nel campione si ottiene la **frequenza relativa**.
Nel caso di dati quantitativi continui, non possiamo applicare questa procedura. È necessario dividere l'intervallo di variazione osservato in un certo numero di sottointervalli di uguale ampiezza, detti **classi**, e valutare la frequenza asssociata a ciascuna classe, contando il numero di dati nel campione la cui misura appartiene alla classe. Solitamente il numero di classi è identificato da questa relazione:

$$k = 3.3 * log_{10}(N)$$

In questo modo è immediato calcolare la frequenza e la frequenza relativa e la corrispettiva distribuzione di frequenza. Un'altro parametro di sintesi dell'informazione è la **frequenza cumulata** calcolabile come la somma delle frequenze associate alle classi fino a quella il cui estremo superiore coincide con il livello prefissato.

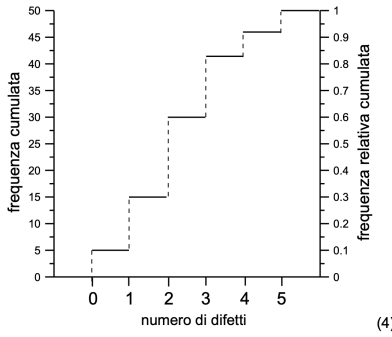
2.1.2 Istogrammi

La statistica descrittiva fornisce,oltre alla rappresentazione tabellare mediante le distribuzioni di frequenze, una rappresentazione grafica dell'informazione. Quest'ultima prende il nome di **istogramma**. L'istogramma è costituito da rettangoli adiacenti. Nel caso di dati quantitativi continui, ogni rettangolo ha base di lunghezza pari all'ampiezza della corrispondente classe; l'altezza invece è proporzionale alla frequenza o alla frequenza relativa associata a quella classe.

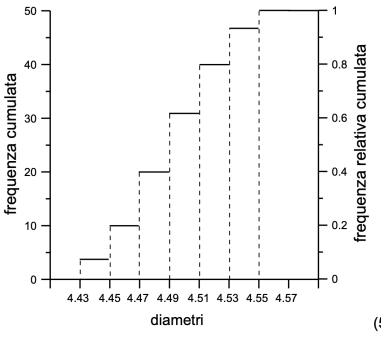


2.1.3 Istogrammi della distribuzione di frequenza cumulata

L'istogramma della distribuzione di frequenza cumulata è una rappresentazione grafica che assume un andamento a gradini,ottenuto riportando, in corrispondenza di ciascuna classe, il valore della frequenza cumulata fino a tale classe.



(4)



(5)

2.1.4 Indici di tendenza centrale

Gli indici di tendenza centrale sono dei parametri che forniscono un'idea su dove, numericamente, i dati sono concentrati.

Media aritmetica Si abbia un campione di n osservazioni sperimentali di tipo numerico (discreto o continuo) e sia xi (i =1,...,n) il generico valore osservato.La media aritmetica è definita come:

$$x = \sum_{i=1}^{n} x_i/n$$

Concettualmente essa rappresenta il centro di gravità delle osservazioni.
Vantaggi: è molto semplice da calcolare e di facile comprensione.
Svantaggi: dà eccessivo peso ai valori estremi anche se questi sono poco numerosi ed è valida solo per dati quantitativi.

Mediana La mediana di un campione (o di una popolazione) è quel valore della variabile rispetto al quale metà dei valori osservati, ordinati in senso crescente, risultano minori e l'altra metà maggiori.
Vantaggi: semplice da calcolare se n è piccolo, non è influenzato da valori estremi.
Svantaggi: non si presta a manipolazioni di tipo algebrico.

Moda La moda è definita come quel valore della variabile osservata in corrispondenza del quale la frequenza è massima. **Svantaggi:** In un campione possono apparire più mode e non sono sempre valori centrali.

2.1.5 Indici di dispersione

In statistica, un indice di dispersione è un indice che descrive sinteticamente la variabilità di un insieme di dati.

Escursione campionaria L'escursione campionaria (o range) è definita come differenza tra il più grande ed il più piccolo valore osservato della variabile di interesse nella popolazione o nel campione.

$$R = x_max - x_min$$

Varianza e deviazione standard Si definisce varianza di una popolazione il valore, generalmente indicato con il simbolo S^2, che risulta dalla somma dei quadrati delle differenze dei valori osservati, x_i, dalla media aritmetica, x, diviso per il numero delle osservazioni, n, ovvero:

$$s^2 = \sum_{i=1}^{n}(x_i - x)^2/n$$

Per motivi teorici (scomposizione dei gradi di libertà), però, quando ci si riferisce ad un campione, è preferibile definire la varianza mediante la seguente relazione:

$$s^2 = \sum_{i=1}^{n}(x_i - x)^2/n - 1$$

La varianza permette di identificare la dispersione dei valori della variabile X attorno al valor medio.
La deviazione standard S è definita come la radice quadrata della varianza.

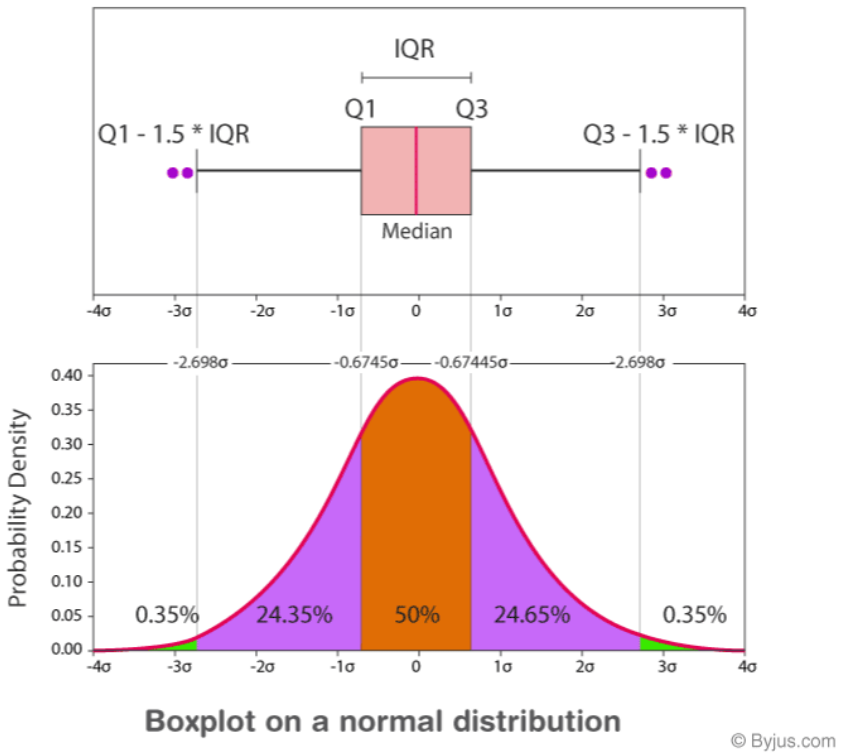
2.1.6 Box plot

Un box plot è un grafico utile che mostra la distribuzione dei dati di una variabile. Può essere letto in questo modo:

- La linea centrale nella scatola rappresenta la **mediana** dei dati. La metà dei dati si trova sopra questo valore, l'altra metà sotto. Se i dati sono simmetrici, la mediana è al centro della scatola.
- La parte inferiore e superiore della scatola mostrano il **primo** e il **terzo quartile**. La lunghezza della scatola è la differenza tra i due quartili e si chiama **range interquartile (IQR)**.

- Le linee che si estendono a partire dalla scatola sono chiamate **baffi**. I baffi rappresentano la variazione dei dati attesa e si estendono per 1,5 volte dall'IQR dalla parte superiore e inferiore della scatola. Se i dati non arrivano fino alla fine dei baffi, significa che i baffi si estendono fino ai valori di dati minimi e massimi. Se, invece, i dati ricadono sopra o sotto la fine dei baffi, sono rappresentati come punti, denominati spesso **outlier**.

NOTA: Un outlier è un punto più estremo della variazione attesa. Vale la pena esaminare questi punti di dati per determinare se sono errori di trascrizione o veri e propri outlier.



2.2 Commenti sul codice

Nella fase iniziale dell'analisi, andiamo a suddividere i nostri dati in due sottinsiemi riguardanti il **train set** e il **test set**. In particolare utilizzeremo il 70% dei dati per il training e l'30% per il testing

```
1    training=data[31:100,]
2    test=data[1:30,]
```

Successivamente andiamo ad analizzare le prime informazione sui dati del training.
In particolare mediante il comando summary, R ci calcola media, mediana, primo e terzo quartile, minimo e massimo di ogni variabile, trattandosi di dati quantitativi. Mediante il comando apply invece calcoliamo gli indici di dispersione: varianza e deviazione standard. Infine il comando diff andiamo a calcolare l'escursione campionaria.

```
summary(training)

diff(range(training$y_ImageQuality))
diff(range(training$x1_ISO))
diff(range(training$x2_FRatio))
diff(range(training$x3_TIME))
diff(range(training$x4_MP))
diff(range(training$x5_CROP))
diff(range(training$x6_FOCAL))
diff(range(training$x7_PixDensity))


apply(training,2,var)
apply(training, 2, sd)
```

I risultati di questa prima analisi sono i seguenti:

```
 y_ImageQuality        x1_ISO           x2_FRatio           x3_TIME            x4_MP
 Min.   : -8.682   Min.   :-1.69861   Min.   :-1.634843   Min.   :-1.65547   Min.   :-1.5531
 1st Qu.: 59.546   1st Qu.:-0.76927   1st Qu.:-0.837894   1st Qu.:-0.67861   1st Qu.:-0.5320
 Median : 78.774   Median :-0.07859   Median :-0.022622   Median : 0.09902   Median : 0.2285
 Mean   : 77.086   Mean   :-0.04455   Mean   :-0.004197   Mean   : 0.11106   Mean   : 0.1946
 3rd Qu.: 96.333   3rd Qu.: 0.58373   3rd Qu.: 0.844934   3rd Qu.: 0.97671   3rd Qu.: 0.8827
 Max.   :127.316   Max.   : 1.70996   Max.   : 1.706467   Max.   : 1.62455   Max.   : 1.7314
    x5_CROP           x6_FOCAL       x7_PixDensity
 Min.   :-1.7197   Min.   :-1.6237   Min.   :-1.9992
 1st Qu.:-1.1867   1st Qu.:-0.9927   1st Qu.:-0.5071
 Median :-0.3957   Median :-0.2432   Median : 0.1093
 Mean   :-0.1438   Mean   :-0.1869   Mean   : 0.1225
 3rd Qu.: 0.9075   3rd Qu.: 0.3819   3rd Qu.: 0.6420
 Max.   : 1.7261   Max.   : 1.7023   Max.   : 2.1444
> diff(range(training$y_ImageQuality))
[1] 135.9979
> diff(range(training$x1_ISO))
[1] 3.408564
> diff(range(training$x2_FRatio))
[1] 3.341309
> diff(range(training$x3_TIME))
[1] 3.280017
> diff(range(training$x4_MP))
[1] 3.284461
> diff(range(training$x5_CROP))
[1] 3.445884
> diff(range(training$x6_FOCAL))
[1] 3.325971
> diff(range(training$x7_PixDensity))
[1] 4.143585
> #indice di dispersione
> apply(training,2,var)
y_ImageQuality         x1_ISO      x2_FRatio        x3_TIME          x4_MP        x5_CROP       x6_FOCAL  x7_PixDensity
   724.3844029      0.8797329      0.9715301      0.9896845      0.8068841      1.2884222      0.8673521      0.8305836
> apply(training, 2, sd)
y_ImageQuality         x1_ISO      x2_FRatio        x3_TIME          x4_MP        x5_CROP       x6_FOCAL  x7_PixDensity
    26.9143903      0.9379408      0.9856623      0.9948289      0.8982673      1.1350869      0.9313174      0.9113636
```

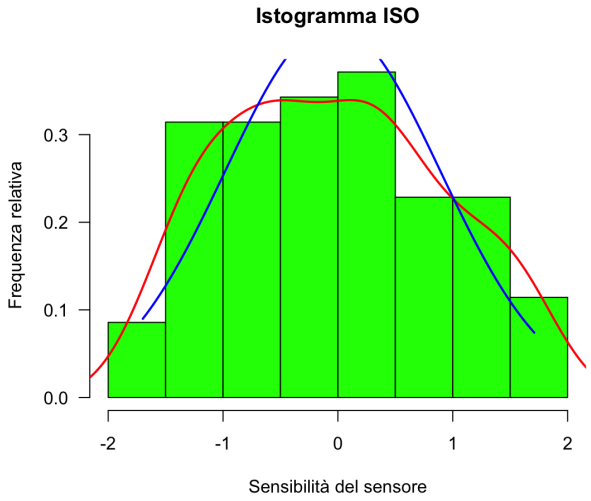
Osserviamo che i valori dei regressori hanno valori che oscillano intorno allo zero con escursione che rimane tra 3 e 4 (sono standardizzati), mentre la qualità dell'immagine va da circa -8 a 127, generando un'escursione pari a 136. Infine la varianza e deviazione standard della qualità dell'immagine sono molto alti mentre la standardizzazione permette ai regressori di avere una

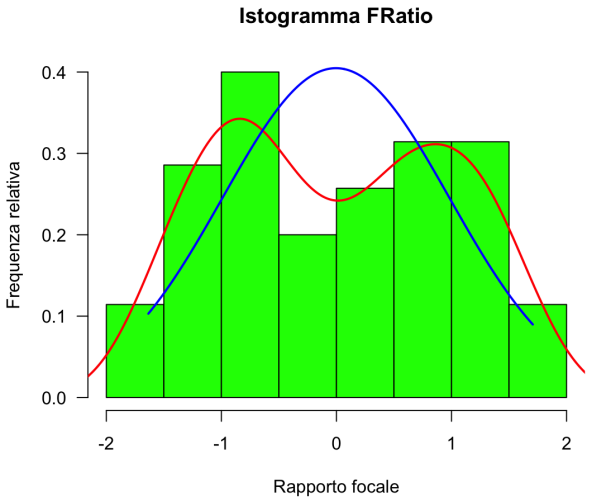
varianza e deviazione standard molto più contenuti.

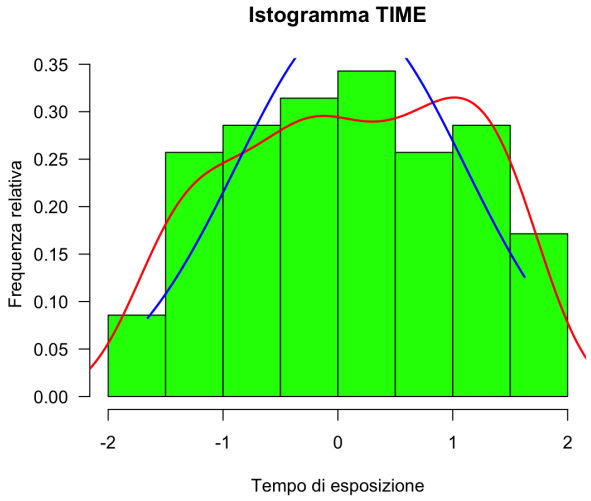
L'analisi dei dati prosegue attraverso la visualizzazzione degli istogrammi di densità (l'altezza indica la densità della classe, l'area la frequenza relativa) di tutte le variabili del nostro dataset. In aggiunta, con i comandi **lines** e **curve** inseriamo la PDF di ogni variabile e la PDF di una normale con i parametri della variabile stessa.

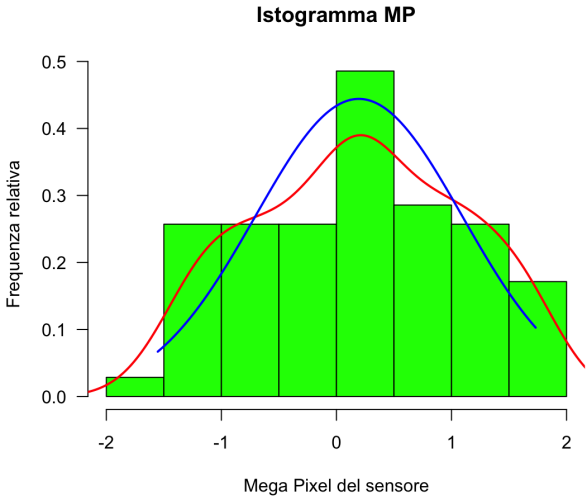
```
hist(training$x1_ISO,las=1,  freq=F,xlab="Sensibilita' del sensore",ylab
    ="Frequenza relativa", main="Istogramma ISO", col="green") #las=1
    mette i numeri in orizzontale
lines(density(training$x1_ISO),lwd=2,col="red")
curve(expr = dnorm(x = t,mean = mean(training$x1_ISO), sd = sd(training$
    x1_ISO)),from = min(training$x1_ISO),to = max(training$x1_ISO),n =
    500,add = TRUE,xname = "t",col = "blue", lwd=2)

hist(training$x2_FRatio,las=1,  freq=F,xlab="Rapporto focale",ylab="
    Frequenza relativa", main="Istogramma FRatio",col="green")
lines(density(training$x2_FRatio),lwd=2,col="red")
curve(expr = dnorm(x = t,mean = mean(training$x2_FRatio), sd = sd(
    training$x2_FRatio)),from = min(training$x2_FRatio),to = max(training
    $x2_FRatio),n = 500,add = TRUE,xname = "t",col = "blue", lwd=2)

.........
```

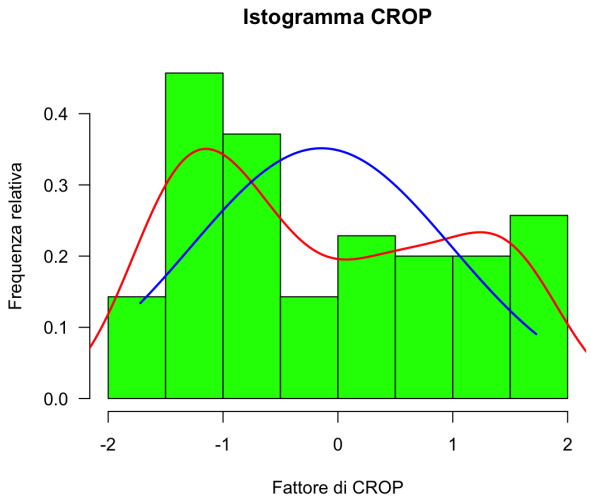
I risultati prodotti sono i seguenti:

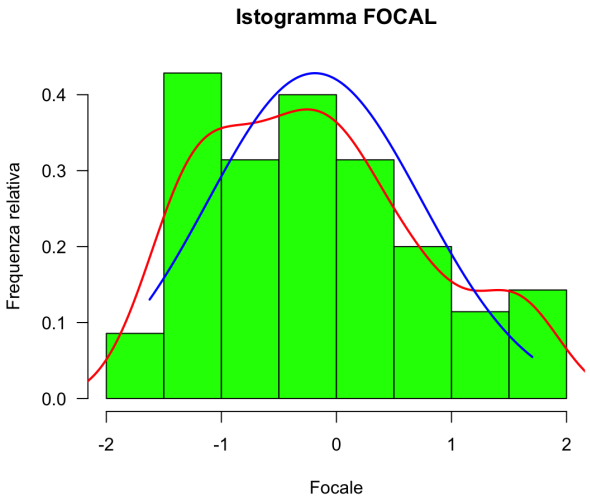


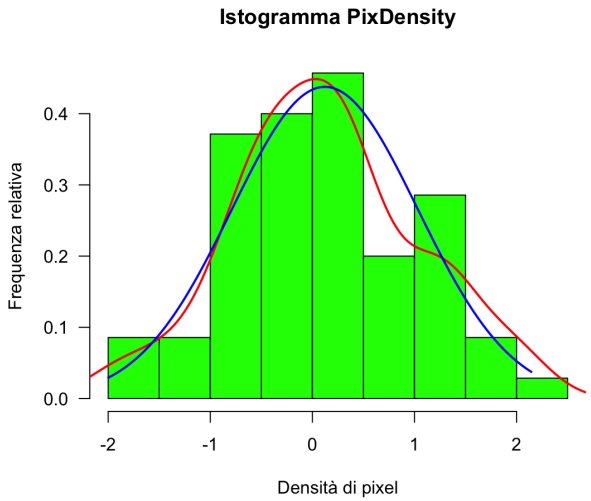


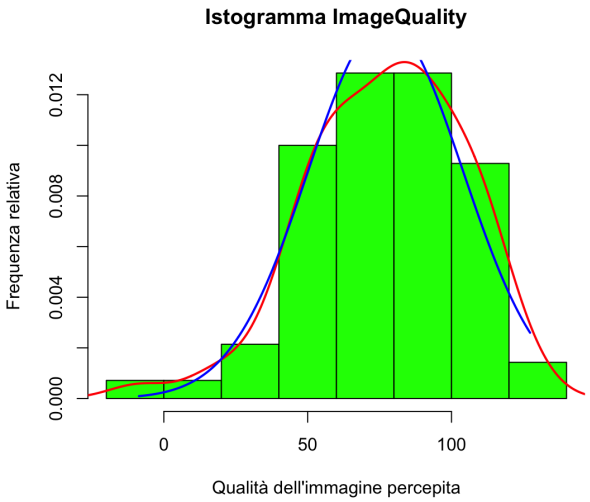












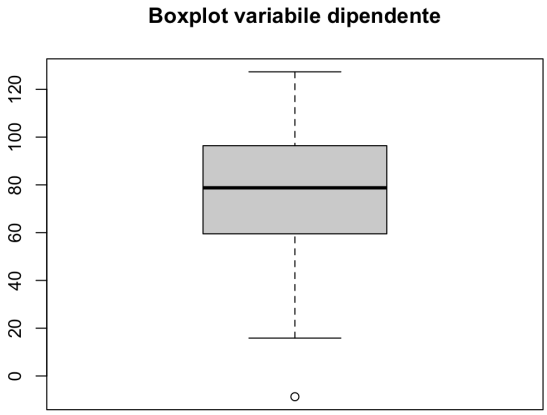
Come notiamo il numero delle classi formate è otto, essendo che di default R utilizza la regola del logaritmo sopra citata.
E' facile individuare la moda di ogni variabile, posizionata all'interno della classe caratterizzara dal rettangolo con maggiore altezza.

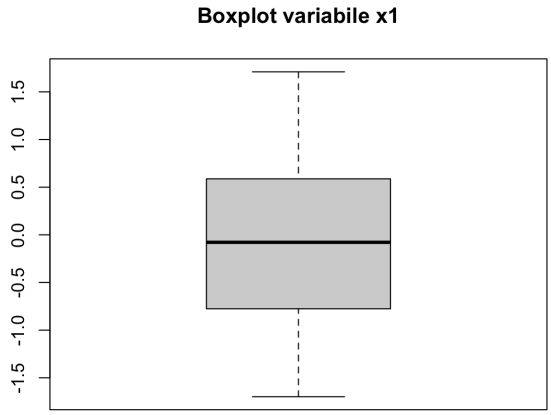
Notiamo inoltre, in particolar modo per X_2,X_3,X_5 e X_6 che la distribuzione dei dati si discosta molto da quella normale.

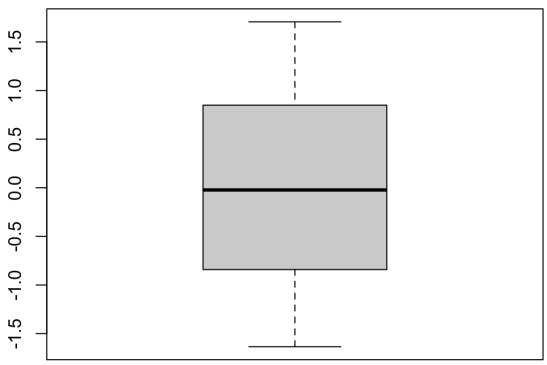
Proseguiamo attraverso dei **boxplot**

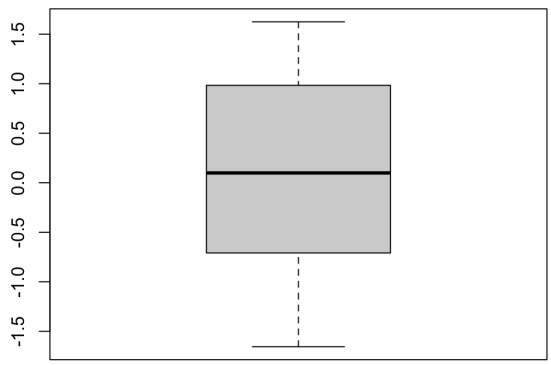
```
out0 = boxplot(training$y_ImageQuality,main="Boxplot variabile
    dipendente"); out0$out #1 outlier, -8.682078, riga 80
out1 = boxplot(training$x1_ISO,main="Boxplot variabile x1"); out1$out
out2 = boxplot(training$x2_FRatio,main="Boxplot variabile x2"); out2$out
.....
training=training[-50,] # rimuoviamo outlier
```

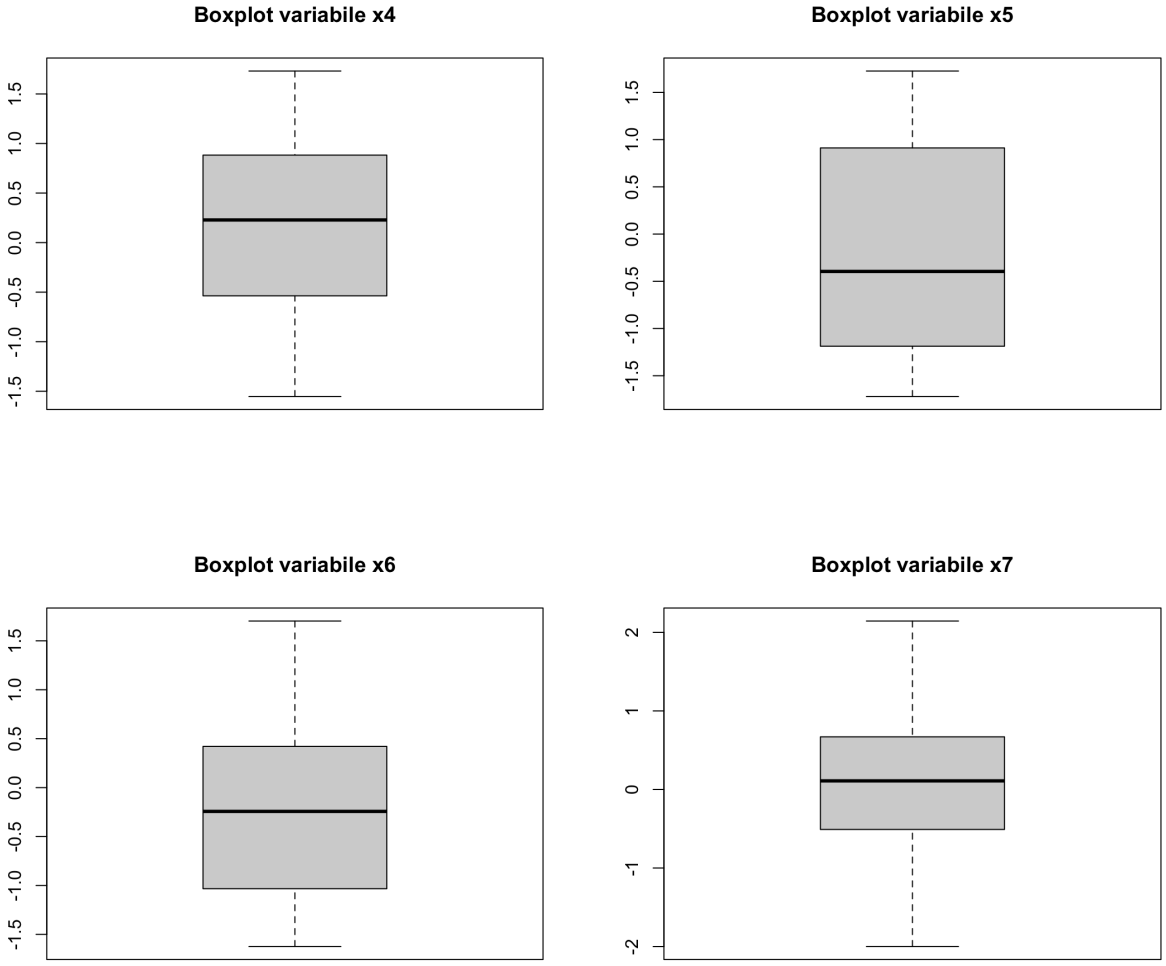
L'output prodotto è il seguente:









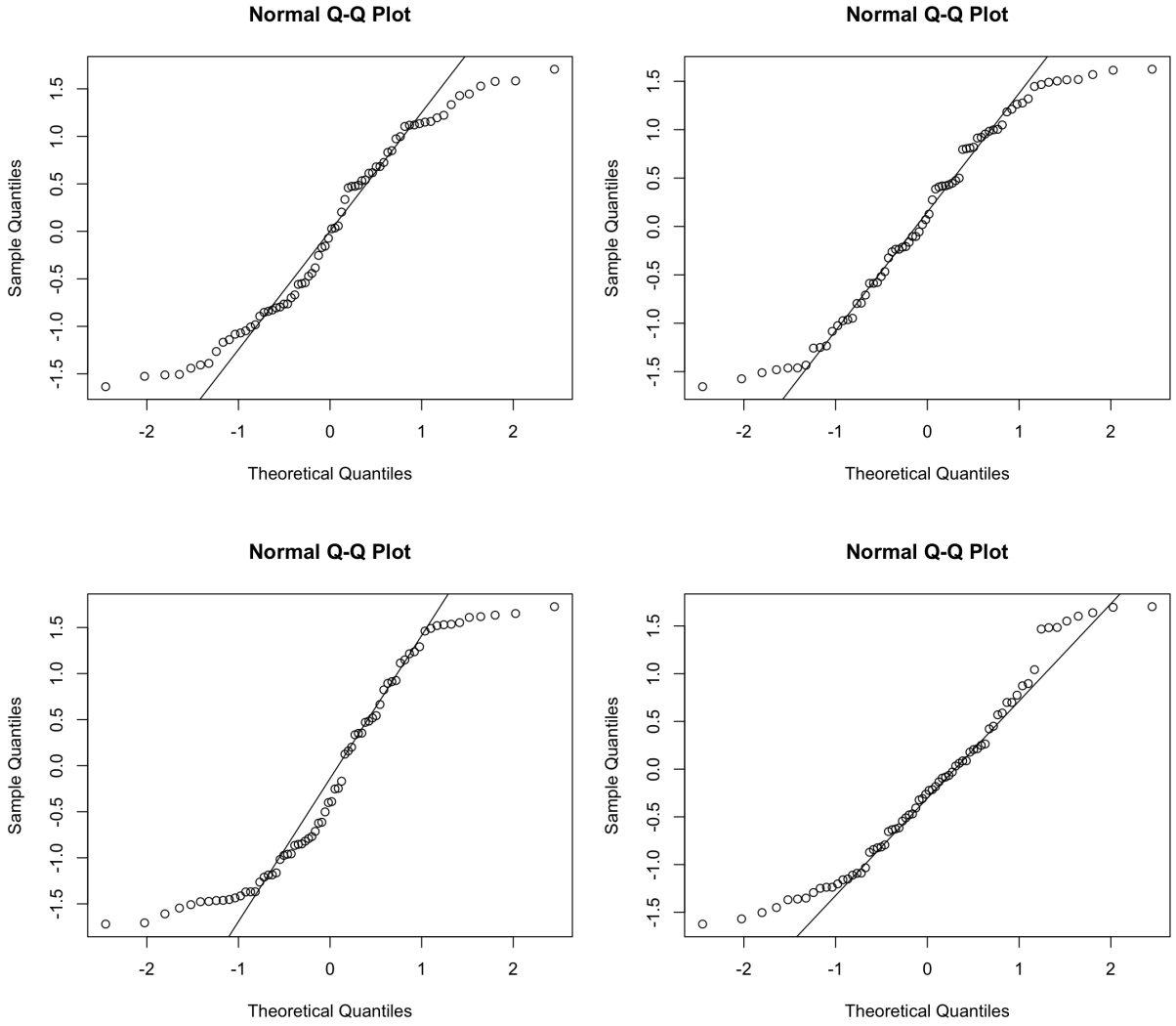


Dai grafici dei **boxplot** abbiamo conferma di alcuni dati ottenuti dal summary.Infatti facendo riferimento alla variabile dipendente y, notiamo una mediana circa pari a 80, un primo quartile circa pari a 60 e il terzo quartile circa pari a 100.
Una caratteristica importante che rileviamo soltanto nel boxplot relativo alla y, è la presenza di un outlier. Tale valore corrisponde precisamente, come evidenziato dal summary, a -8.68 e risulta essere anomalo rispetto al resto della distribuzione; abbiamo deciso di rimuoverlo dal training set.

Ora attraverso lo **Shapiro Test** verifichiamo la normalità della variabile dipenente. Infatti con l'assunzione di normalità possiamo calcolare intervalli di confidenza e test di ipotesi sulla retta di regressione.
Successivamente facciamo lo stesso sui regressori. Lì dove lo Shapiro test decide di rifiutare H_0 analizziamo anche i **QQ-Plot**

```
shapiro.test(training$y_ImageQuality)

shapiro.test(training$x1_ISO)

shapiro.test(training$x2_FRatio)
qqnorm(training$x2_FRatio)
qqline(training$x2_FRatio)

shapiro.test(training$x3_TIME)
qqnorm(training$x3_TIME)
qqline(training$x3_TIME)
```

Per i regressori X_2,X_3,X_5 e X_6 il P_{value} è minore di α e ciò mi porta a rifiutare H_0.Analizziamo i relativi **QQ-Plot**:



Il **QQ-Plot** confronta la distribuzione cumulata della variabile osservata con la distribuzione cumulata della normale.È chiaro dai grafici che la distribuzione dei regressori analizzati si discosta da quella normale, in particolar modo nel grafico 3.
Infatti questo prova il valore molto basso del p-Value ottenuto con lo Shapiro-Test effettuato su tale variabile (p-value = 5.81e-05).

CAPITOLO 3

ANALISI DELLA CORRELAZIONE

3.1 Elementi di teoria di analisi della correlazione

La relazione tra due (o più) variabili può essere utilizzata sia al fine di agire su variabili di input così da condizionare la variabile di output, sia per utilizzare l' osservazione effettuata su di una determinata variabile per stimare o predire il valore dell' altra.
Tali due problematiche sono trattati da ampi capitoli dell'Analisi Statistica dei dati e prendono il nome di Analisi di Correlazione e Analisi di Regressione.

Analisi di correlazione

Un metodo preliminare per poter analizzare se vi è **legame lineare** tra due variabili quantitative X e Y, che viene generalmente seguito dalla costruzione del **diagramma di correlazione**.

Per poter costruire tale diagramma bisogna seguire tali passi:

- **Rilevazione dei dati:** Definire il valore che le due variabili assumono in corrispondenza di un determinato tempo o luogo. Tale rilevazione va ripetuta in un numero N di condizione diverse (circa 20-30 coppie di valori).
- **Costruzione del grafico di correlazione:** Rappresentare due assi ortogonali e tracciare le opportune scale per la rappresentazione delle due variabili. Nel caso in cui si voglia studiare la dipendenza della variabile Y dalla variabile X in genere si utilizza l'asse verticale per le Y e quello orizzontale per le X.

Possono verificarsi diverse situazioni tipiche:

1. **Forte correlazione positiva:** situazione dove ad una crescita di X corrisponde quasi sicuramente una crescita di Y. In tal caso la nube dei punti tende ad assumere la forma di un ‘ ellissi schiacciata;
2. **Debole correlazione positiva:** situazione dove vi è una tendenza (non marcata) a crescere della Y quando cresce la X.
3. **Assenza di correlazione:** in tal caso non esiste legame tra le variabili, ovvero la nube di punti tende ad assumere una forma circolare.

4. **Debole correlazione negativa:** vi è una debole correlazione tra X e Y con la differenza che stavolta la dipendenza di una diminuzione di Y corrisponde ad una crescita di X.
5. **Forte correlazione negativa:** ad una diminuzione di Y corrisponde quasi sicuramente una crescita di X.

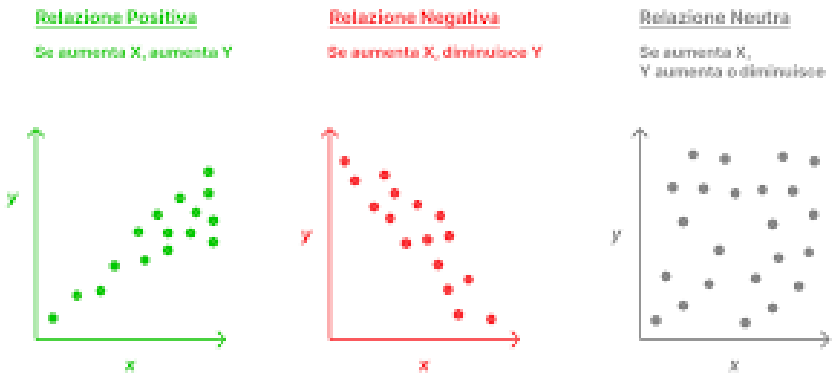
Possiamo affermare che tra le due variabili X e Y vi sono diversi gradi di associazione. Si va dall' assenza di correlazione fino al caso di perfetta correlazione quando tutti i punti si allineano su di una retta. Per misurare il grado di associazione tra due variabili è utilizzato il coefficiente di correlazione R:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Il coefficiente R è compreso tra i valori di -1 e 1. Quanto più è prossimo ad 1 in valore assoluto, tanto più le due variabili sono correlate, mentre quanto più esso è vicino allo 0 tanto più il grado di associazione è basso.
Valori positivi di R indicano una correlazione positiva, al contrario valori negativi indicano correlazione negativa. Nel caso ideale di perfetta correlazione R=±1, mentre nel caso di R=0 le due variabili non sono correlate linearmente.

Scatter Plot

Un grafico di notevole importanza per lo studio di tale legame è lo scatter plot: ogni punto su tale grafico rappresenta una coppia di valori di due variabili. Lo scatter plot è molto utile anche per visualizzare la distribuzione dei dati.



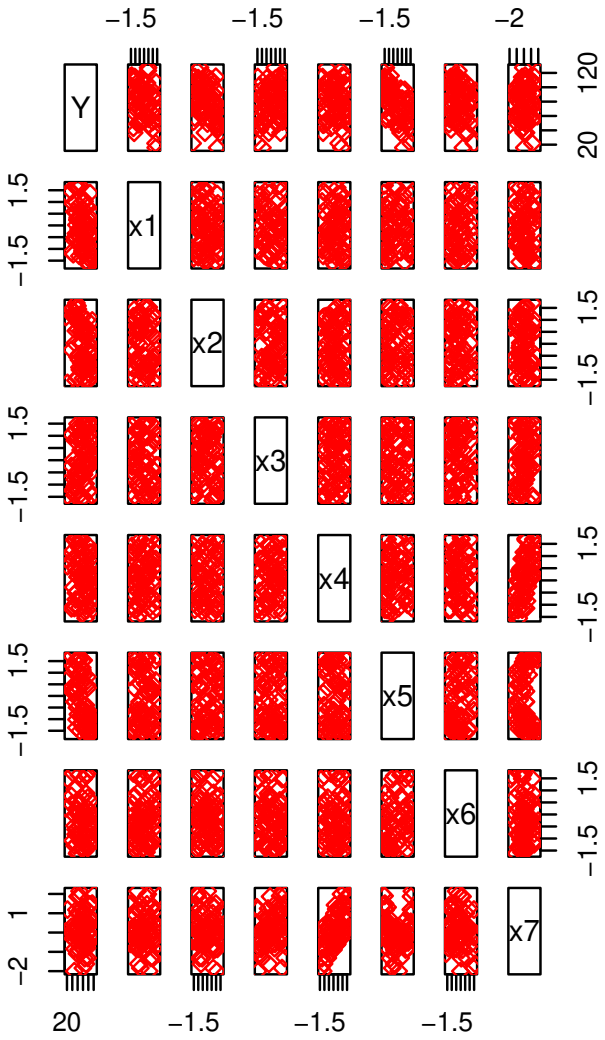
Precauzioni nell' interpretazione di un grafico di correlazione:
Esiste una differenza tra correlazione ed il concetto di causa-effetto. Un valore elevato del coefficiente di correlazione non implica che una determinata variabile è la causa della variazione dell'altra. Spesso vi è una causa comune non individuata che causa la variazione di entrambe le variabili. Inoltre il coefficiente di correlazione misura il grado di associazione lineare, per cui se pari a 0, è possibile che tra le due variabili esista una forte dipendenza non lineare.

3.2 Commenti sul codice

Mediante il comando:

```
    pairs(training[, c(1,2,3,4,5,6,7,8)],
        col = red,
        pch = 5,
        labels = c("Y","x1","x2","x3","x4","x5","x6","x7"),
)
```

vengono generati gli scatter plot per analizzare la correlazione tra la coppia di variabili dipendenti ed indipendenti del nostro dataset.



Per una maggiore comprensione, è possibile usufruire del comando:

```
    cor1 = round(cor(training_new), digits = 2); cor1
```

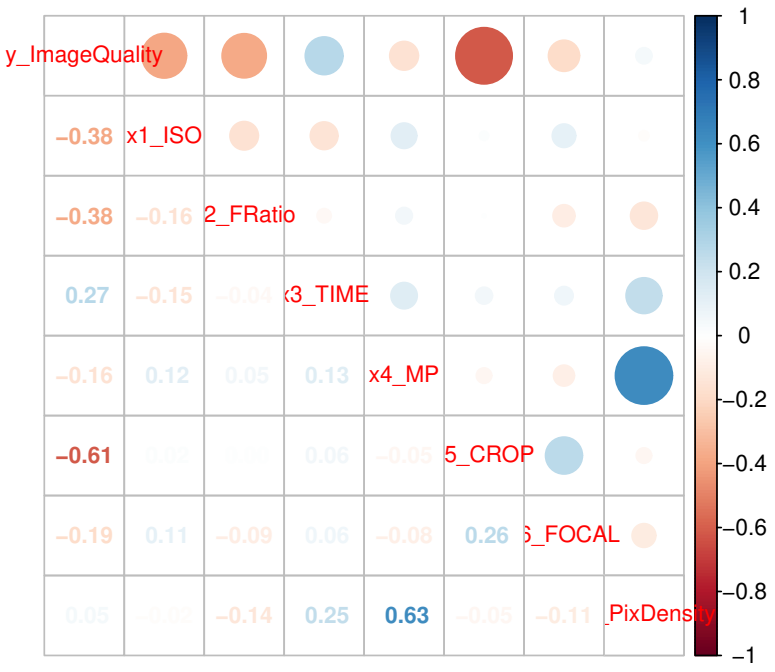
con il quale è possibile calcolare e far stampare la matrice di correlazione.

	y_ImageQuality	x1_ISO	x2_FRatio	x3_TIME	x4_MP	x5_CROP	x6_FOCAL	x7_PixDensity
y_ImageQuality	1.00	-0.38	-0.38	0.27	-0.16	-0.61	-0.19	0.05
x1_ISO	-0.38	1.00	-0.16	-0.15	0.12	0.02	0.11	-0.02
x2_FRatio	-0.38	-0.16	1.00	-0.04	0.05	0.00	-0.09	-0.14
x3_TIME	0.27	-0.15	-0.04	1.00	0.13	0.06	0.06	0.25
x4_MP	-0.16	0.12	0.05	0.13	1.00	-0.05	-0.08	0.63
x5_CROP	-0.61	0.02	0.00	0.06	-0.05	1.00	0.26	-0.05
x6_FOCAL	-0.19	0.11	-0.09	0.06	-0.08	0.26	1.00	-0.11
x7_PixDensity	0.05	-0.02	-0.14	0.25	0.63	-0.05	-0.11	1.00

Inoltre, servendosi del comando:

```
1  corrplot.mixed(cor(training_new),number.cex=0.8,tl.cex=0.8)
```

È possibile ottenere una rappresentazione grafica della matrice di correlazione precedente.



Analizzando tale grafico possiamo dedurre che vi è una forte correlazione lineare (la quale verrà trattata più nello specifico nelle figure successive) tra la variabile dipendente y e x_5, e tra x_7 e x_4. Notiamo anche una debole correlazione lineare tra y e x_1 e tra y e x_2.

CAPITOLO 4

ANALISI DELLA REGRESSIONE

4.1 Elementi di teoria di analisi della regressione

L' Analisi di Regressione è una tecnica statistica con l' obiettivo di determinare la superficie di risposta, ovvero stimare il valore atteso della grandezza di interesse in funzione di una qualsiasi combinazione dei livelli delle variabili misurate che possono spiegarne l' andamento, nel momento in cui tali variabili, che prendono anche il nome di fattori, assumono valori in un insieme continuo (tale analisi si può estendere facilmente anche al caso di variabili discrete e categoriche).

Per quanto riguarda le convenzioni, i fattori prendono il nome di variabili indipendenti o anche regressori (indicati in genere con la x), i livelli delle variabili indipendenti prendono il nome di valori, mentre la grandezza di interesse viene definita variabile dipendente(indicata in genere con la y).
Nell' ambito della regressione, stimare il valore atteso di una variabile dipendente Y in corrispondenza di un valore fissato della variabile indipendente X, significa disporre di un determinato modello matematico che riesca a legare le due variabili caratterizzato da un livello di accuratezza quantificabile.
Tra i modelli più importanti vi sono i **modelli lineari**. Un esempio molto semplice di quest' ultimo è il seguente:

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{4.1}$$

Dove:

- β_0 rappresenta il termine costante, chiamato intercetta, qualunque sia il valore assunto dalla variabile indipendente;
- $\beta_1 X$ è il contributo dovuto dalla regressione(l' effetto su Y di X al livello di x) che rappresenta un' aliquota fissa per ogni valore fissato di X;
- ϵ è una quantità aleatoria che definisce la variabilità della quantità osservata Y rispetto alla quantità fissa $\beta_0 + \beta_1 X$ (rappresenta dunque l' errore).

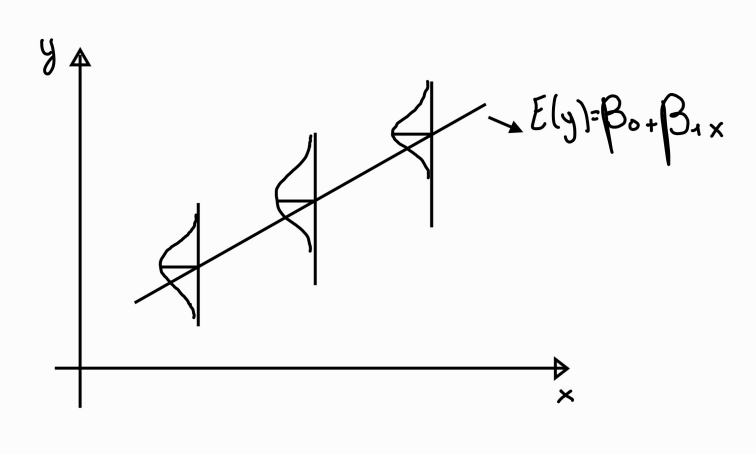
Dunque si assume che tale variabile aleatoria (errore) abbia una distribuzione Normale con :

$$E(\epsilon) = 0 \text{ e } Var(\epsilon) = \sigma^2$$

ed inoltre che le osservazioni sperimentali siano tra loro stocasticamente indipendenti. La distribuzione gaussiana dell'errore è opzionale e serve per la stima dei minimi quadrati delle varie costanti, così da ottenere gli stessi stimatori calcolati dalla stima di Massima Verosimiglianza. Sulla base di tali ipotesi possiamo definire che:

$$E(Y) = \beta_0 + \beta_1 X$$

Dunque il modello adottato afferma che la relazione che lega il valore atteso della variabile aleatoria Y alla variabile X è di tipo lineare e gli scostamenti da tale relazione sono imputabili ai valori assunti da una variabile aleatoria Normale con valore atteso pari a zero e varianza σ^2 indipendentemente dal valore attuale della variabile X.



È inoltre utile precisare che la variabilità definita dalla variabile aleatoria ϵ è originata dalle differenze tra le unità sperimentali, ovvero da tutti quei fattori sui quali chi effettua l' esperimento non ha controllo.
C'è da dire che l' errore di previsione dipende anche dall' accuratezza del modello di regressione, infatti se quest' ultimo si presta bene a descrivere il modello di popolazione in esame , l' errore sarà minore rispetto ad un modello più semplice.
Per migliorare il nostro modello è utile servirsi anche di modelli polinomiali del tipo:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

o anche modelli di tipo non polinomiale del tipo:

$$y = \beta_0 + \beta_1 f(\mathbf{x}) + \epsilon$$

Tali modelli sono tutti di regressione lineare, in quanto la linearità si riferisce alla dipendenza della variabile aleatoria dipendente Y e dai parametri dei regressori.

4.2 Commento sul codice

Per poter comprendere quali regressori , dal punto di vista lineare, posso considerare nei miei modelli di regressione polinomiale analizzo la loro correlazione lineare con la variabile dipendente y. Per eseguire tale operazione utilizzo il comando ggpairs:

```
ggpairs(training[,c(8,1)])
ggpairs(training[,c(7,1)])
ggpairs(training[,c(6,1)])
ggpairs(training[,c(5,1)])
ggpairs(training[,c(4,1)])
ggpairs(training[,c(3,1)])
ggpairs(training[,c(2,1)])
```

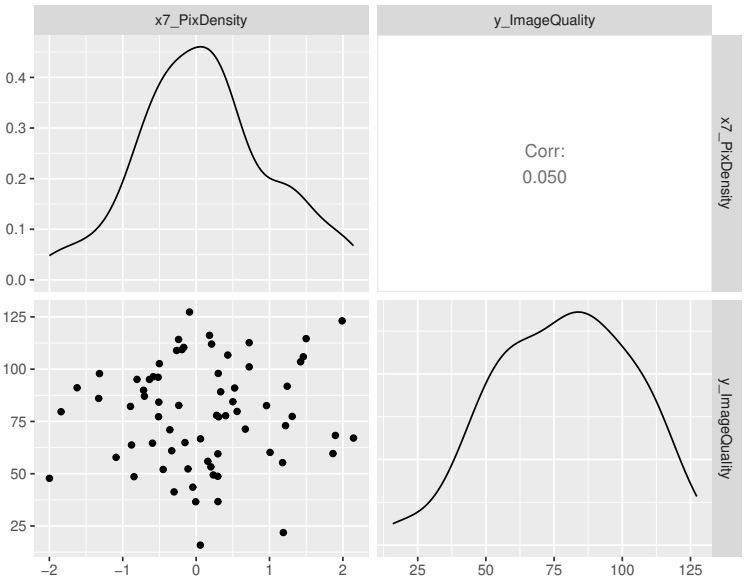


Figura 4.1: Notiamo che vi è una debole correlazione positiva tra le due variabili.

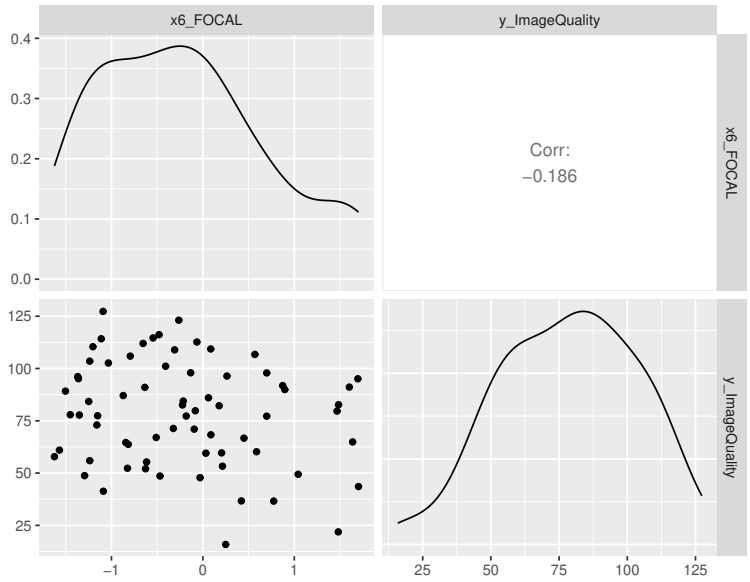


Figura 4.2: Notiamo che tali variabili sono incorrelate dal punto di vista lineare.

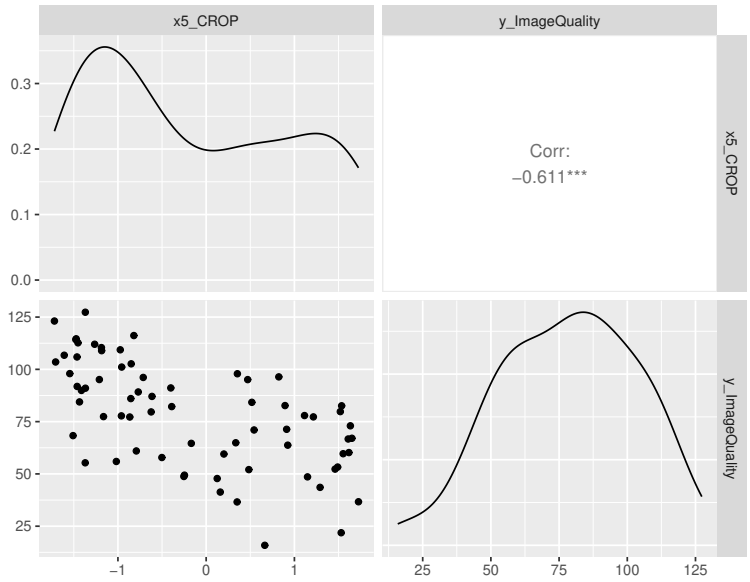


Figura 4.3: Notiamo che tra le due variabili vi è una forte correlazione negativa, ci aspettiamo dunque che il modello di regressione polinomiale abbia tale termine.

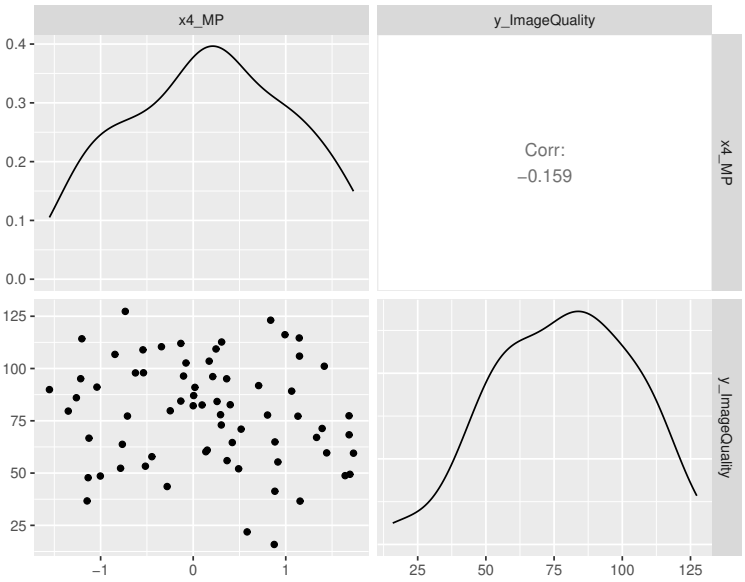


Figura 4.4: Notiamo che tali variabili sono incorrelate dal punto di vista lineare.

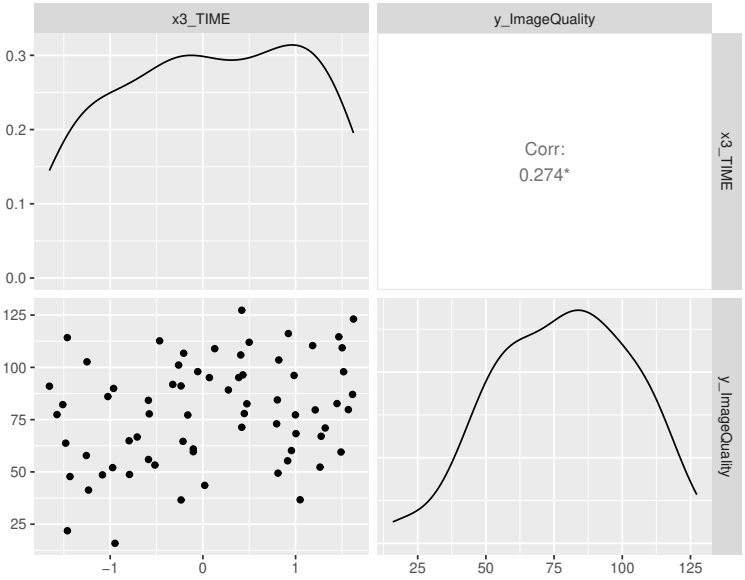


Figura 4.5: Notiamo che vi è una debole correlazione positiva tra le due variabili.

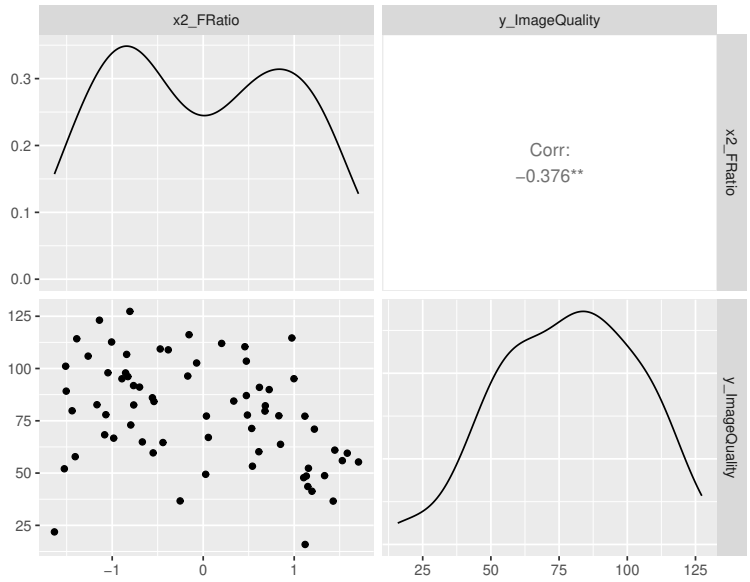


Figura 4.6: Notiamo che tra le due variabili vi è una debole correlazione negativa.

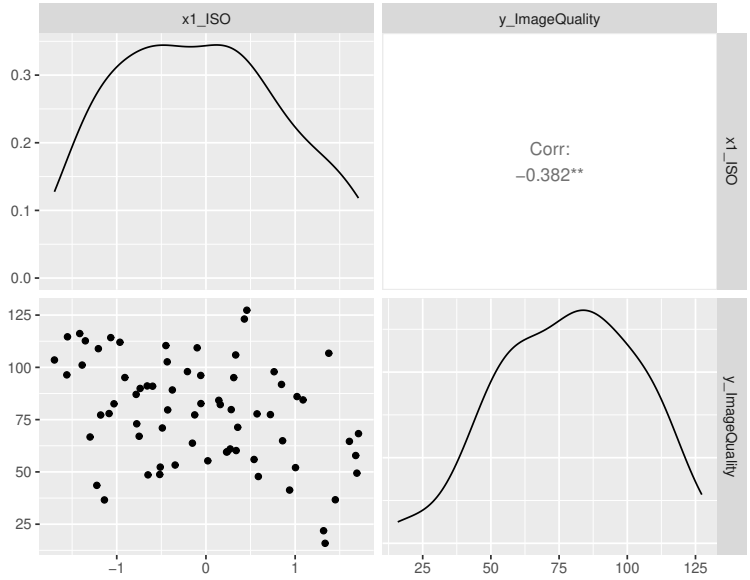


Figura 4.7: Notiamo che tra le due variabili vi è una debole correlazione negativa.

Adesso, sulla base della conoscenza acquisita riguardo la correlazione lineare di ogni singolo regressore e la variabile dipendente possiamo creare dei modelli di regressione polinomiali . La tecnica utilizzata è lo stepwise forward con la tecnica dell' AIC(vedere paragrafi successivi per maggiori dettagli). Per la creazione dei diversi modelli viene utilizzato il comando step,dove trace=1 indica che R dovrà mostrare i passaggi per arrivare a quel modello.In particolar modo:

```
1   step_1= step(lm(y_ImageQuality ~ 1 , data=training),scope = ~ x1_ISO
       + I(x1_ISO^2) + I(x1_ISO^3),direction="forward",trace=1)
```

Il risultato finale è il seguente:

```
Step:  AIC=433.8
y_ImageQuality ~ I(x1_ISO^3)
```

ovvero come modello ottimo di regressione polinomiale si ha:

$$Y = \beta_0 + \beta_1 x_1^3$$

Tale risultato c' era da aspettarselo in quanto tra la y e x1 vi è una debole correlazione negativa lineare, dunque un termine quadratico o cubico potrebbe essere più adatto a descrivere il modello.

```
step_2= step(lm(y_ImageQuality ~ 1, data=training),scope = ~ x2_FRatio
    + I(x2_FRatio^2) + I(x2_FRatio^3),direction="forward",trace=1)
```

abbiamo il seguente risultato:

```
Step:  AIC=429.15
y_ImageQuality ~ x2_FRatio + I(x2_FRatio^2)
```

ovvero come modello ottimo di regressione polinomiale si ha:

$$Y = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2$$

Potevamo aspettarcelo in quanto la correlazione che vi è tra la v.a. dipendente e quella indipendente è debole , dunque per spiegare al meglio il modello, oltre che di un regressore lineare viene utilizzato anche un altro di grado superiore(in tal caso quadratico).

```
step_3=step(lm(y_ImageQuality ~ 1 , data=training),scope = ~ x3_TIME +
    I(x3_TIME^2) + I(x3_TIME^3),direction="forward",trace=1)
```

abbiamo il seguente risultato:

```
Step:  AIC=441.83
y_ImageQuality ~ x3_TIME
```

ovvero come modello ottimo di regressione polinomiale si ha:

$$Y = \beta_0 + \beta_1 x_3$$

anche se effettivamente la correlazione tra la y e x3 è alquanto bassa.

```
step_4=step(lm(y_ImageQuality ~ 1 , data=training),scope = ~ x4_MP +
    I(x4_MP^2) + I(x4_MP^3),direction="forward",trace=1)
```

abbiamo il seguente risultato:

```
Step:  AIC=444.9
y_ImageQuality ~ I(x4_MP^2)
```

ovvero come modello ottimo di regressione polinomiale si ha:

$$Y = \beta_0 + \beta_1 x_4^2$$

Tale modello è giustificato dal fatto che tra le due variabili non vi è nessuna correlazione lineare.

```
step_5=step(lm(y_ImageQuality ~ 1 , data=training),scope = ~ x5_CROP +
    I(x5_CROP^2) + I(x5_CROP^3),direction="forward",trace=1)
```

abbiamo il seguente risultato:

```
Step:  AIC=412.37
y_ImageQuality ~ x5_CROP + I(x5_CROP^2)
```

ovvero come modello ottimo di regressione polinomiale si ha:

$$Y = \beta_0 + \beta_1 x_5 + \beta_2 x5^2$$

Il termine di primo grado è giustificato dal fatto che tra le due variabili sussiste una forte correlazione.

```
step_6=step(lm(y_ImageQuality ~ 1 , data=training),scope = ~ x6_FOCAL
   + I(x6_FOCAL^2) + I(x6_FOCAL^3),direction="forward",trace=1)
```

abbiamo il seguente risultato:

```
Step:  AIC=444.79
y_ImageQuality ~ x6_FOCAL
```

ovvero come modello ottimo di regressione polinomiale si ha:

$$Y = \beta_0 + \beta_1 x_6$$

Questo significa che anche se vi è assenza di correlazione lineare tra le due variabili, il termine quadratico e cubico di x6 si adattano addirittura peggio al modello rispetto al termine lineare.

```
step_7=step(lm(y_ImageQuality ~ 1 , data=training),scope = ~ x7_
   PixDensity + I(x7_PixDensity^2) + I(x7_PixDensity^3),direction="
   forward",trace=1)
```

abbiamo il seguente risultato:

```
Start:  AIC=445.23
y_ImageQuality ~ 1
```

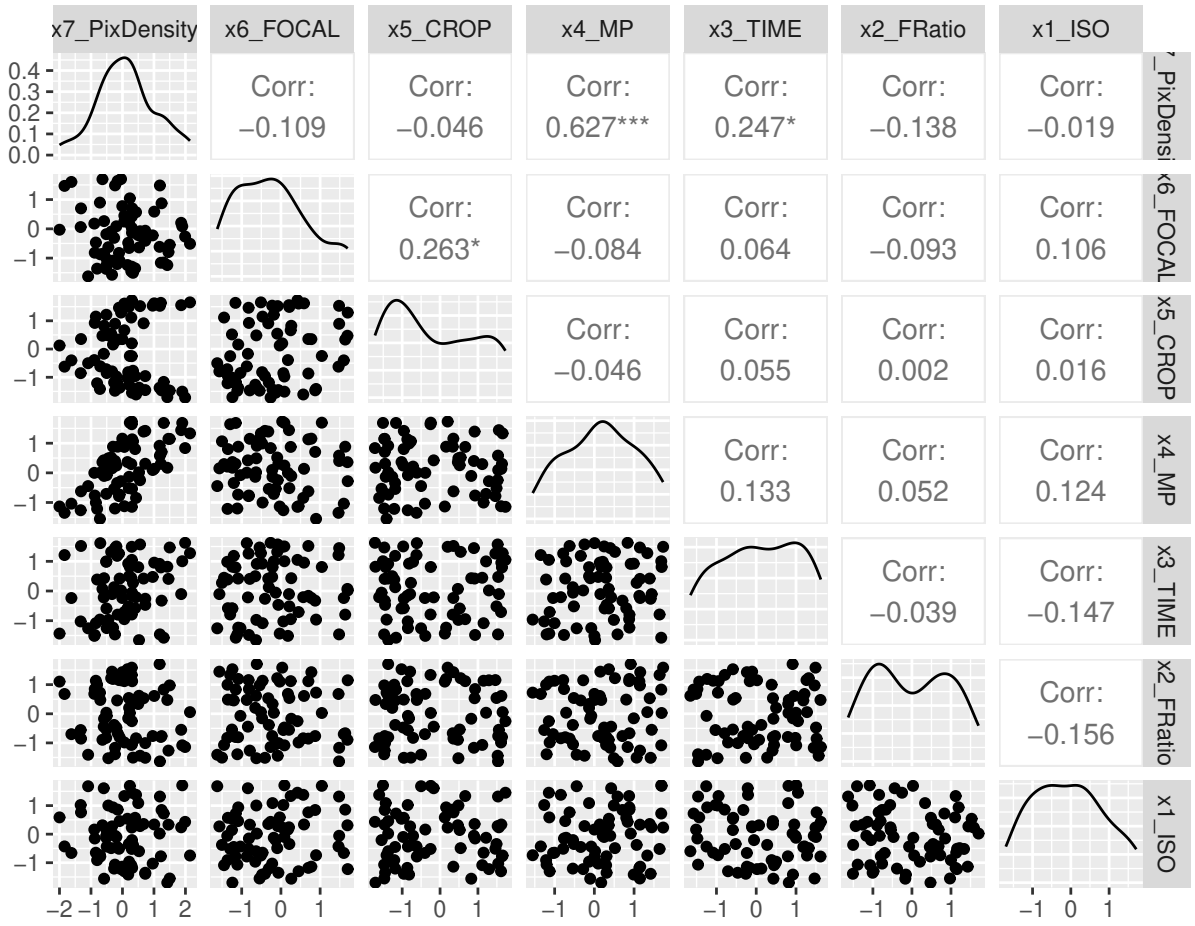
ovvero come modello ottimo di regressione polinomiale si ha:

$$Y = \beta_0$$

Questo è giustificato dall' assenza di correlazione lineare tra le due variabili.
Per quanto riguarda la regressione multipla(verrà trattata nel dettaglio nei capitoli seguenti) è di fondamentale importanza capire se i regressori sono linearmente indipendenti tra loro poiché, in caso contrario, il determinante della matrice $(\underline{X}^T\underline{X})^{-1}$ si avvicina allo 0 , così da avere stimatori troppo grandi .
Per verificare se due regressori sono tra loro dipendenti utilizziamo il comando:

```
ggpairs(training[,c(8,7,6,5,4,3,2)])
```

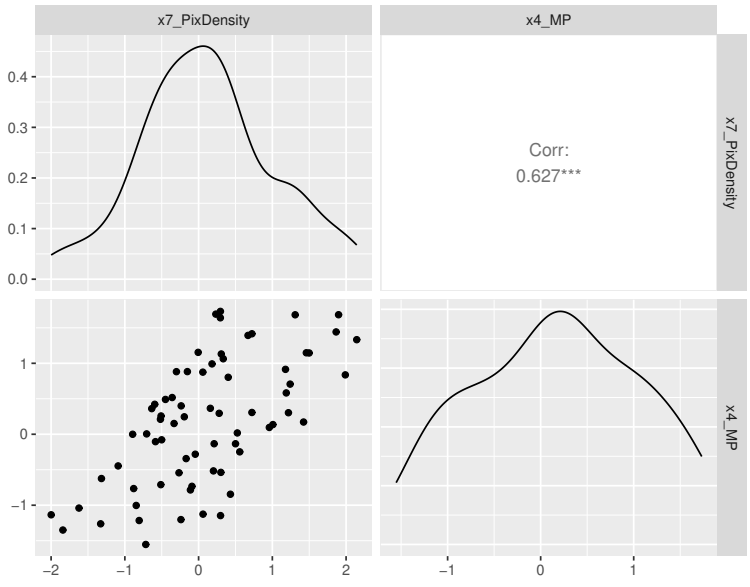
il quale produce la seguente figura(scatter plot, pdf della variabile aleatoria e calcolo del coefficiente di correlazione):



Dunque per un'analisi completa utilizziamo il comando:

```
ggpairs(training[,c(8,5)])
```

con la quale si ottengono tali grafici:


x7_PixDensity
x4_MP
Corr:
0.627***
x7_PixDensity
x4_MP
0.4
0.3
0.2
0.1
0.0
1
0
−1
−2
−1
0
1
2
−1
0
1

CAPITOLO 5

INTERVALLI DI CONFIDENZA E METODO DEI MINIMI QUADRATI

5.1 Elementi di teoria sugli Intervalli di confidenza

Stimare un parametro di una popolazione vuol dire mettere in atto una procedura che, partendo dal campionamento, porti ad avere delle informazioni sul valore del parametro incognito.
Ci sono due tipi di procedure :

1. Stima puntuale
2. Stima per intervallo

Seppur la stima puntuale sia un metodo valido per la stima di un parametro essa non fornisce, in maniera esplicita, informazioni sul livello di incertezza del valore vero del parametro incognito. Di conseguenza si preferisce utilizzare gli intervalli di confidenza.

La procedura di stima per intervallo si basa sul ricavare, sulla base dei dati campionari, un intervallo in cui il valore del parametro incognito è presumibilmente contenuto in quanto, essendo tale intervallo ricavato sulla base di osservazioni campionarie, potrebbe anche non contenere il valore vero del parametro.

Tuttavia si può definire una regola in maniera tale che, ripetendo la stima intervallare, la percentuale di volte che gli intervalli generati contengano il valore vero del parametro può essere fissata a priori.

Possiamo quindi confidare sul fatto che la procedura che abbiamo utilizzato abbia un'elevata probabilità di produrre un intervallo che contiene il valore vero del parametro. Quindi anche se nel singolo esperimento non si potrà mai sapere se l'intervallo generato contiene o meno il valore del parametro incognito si può confidare nel fatto che la regola con cui ho generato l'intervallo, in un gran numero di casi, dà origine ad intervalli che contengono il valore del parametro da stimare.

Il livello di confidenza quindi è la probabilità con cui nella ripetizione dell' esperimento la regola dà luogo ad intervalli che contengono il valore vero ed è uguale ad $1 - \alpha$.

Il livello di rischio α è la probabilità di errore in cui su incorre nel ritenere che l'intervallo ottenuto contenga effettivamente il valore vero del parametro.
Per ottenere rischio 0, l'intervallo dovrebbe avere come estremi -∞ +∞

Quindi un intervallo di confidenza al livello 1-α su un parametro incognito è un intervallo ottenuto mediante una procedura che ha nella ripetizione dell'estrazione degli intervalli una probabilità pari ad 1-α di generare un intervallo che contiene il valore vero del parametro.

Per stimare un intervallo dopo aver estratto il campione e fissato la regola si sceglie una statistica Pivot da cui si determinano i quantili per poi stimare gli estremi inferiore e superiore dell'intervallo.

Il parametro da stimare è deterministico, di conseguenza la probabilità ricade negli estremi che sono variabili aleatorie.
Minore è l'incertezza maggiore sarà la probabilità di errore.
Si calcolano spesso gli intervallo di confidenza sul parametro media di una popolazione Normale o sul parametro varianza di una popolazione Normale.
Le distribuzioni più utilizzate sono la Normale Standard, la chi-quadrato e la T di Student.

Ad esempio

$$\Pr\left(-z_{1-\frac{\alpha}{2}} \leq \ Z \ \leq \ z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

Può essere usato per stimare l'intervallo di confidenza della media di una popolazione Normale conoscendo la varianza.

$$\Pr\left(-t_{1-\frac{\alpha}{2};v} \leq \ T \ \leq \ t_{1-\frac{\alpha}{2};v}\right) = 1 - \alpha$$

Può essere usato per stimare l'intervallo di confidenza della media di una popolazione Normale non conoscendo la varianza.

$$\Pr\left(-\chi^2_{\frac{\alpha}{2};n-1} \leq \chi^{\ 2} \leq \ \chi^2_{1-\frac{\alpha}{2};n-1}\right) = 1 - \alpha$$

Può essere usato per stimare l'intervallo di confidenza della varianza di una popolazione Normale.

5.1.1 Metodo dei minimi quadrati

Forma di un modello lineare con k+1 parametri incogniti

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + B_k x_k + \varepsilon$$

Voglio stimare i parametri sulla base di $n > k+1$ osservazioni sperimentali.
Il metodo dei minimi quadrati fornisce stimatori non distorti e a minima varianza ovvero i migliori stimatori lineari.

Inoltre, sotto l'ipotesi di normalità della variabile aleatoria ϵ, le stime ai minimi quadrati coincidono con le stime che si ottengono applicando il metodo della Massima Verosimiglianza. (Tuttavia la normalità non è richiesta necessariamente).

In particolare, il metodo dei Minimi Quadrati consente di determinare la forma funzionale degli stimatori $(\hat{\beta_0} - \hat{\beta_1})$, imponendo che la somma dei quadrati delle differenze tra i valori osservati Y e i valori stimati $\hat{Y_i}$ sia minima.

$$\hat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{Valore atteso stimato}$$

$$SQE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \min_{\hat{\beta}_0, \hat{B}_1} \left\{ \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2 \right\}$$

Bisogna risolvere il sistema di equazioni che si ottiene uguagliando a zero le sue derivate parziali prime.

$$\frac{\partial SQE}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial SQE}{\partial \hat{\beta}_1} = 0$$

Gli stimatori non distorti dei parametri quindi sono :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \frac{y_i}{n} - \left(\sum_{i=1}^{n} \frac{x_i}{n} \right) \hat{\beta}_1 = \overline{y} - \overline{x}\hat{\beta}_1$$

Essendo combinazioni lineari delle osservazioni yi.

Si può notare che $\hat{\beta}_0$ e $\hat{\beta}_1$ non sono indipendenti tra loro, c'è un termine di covarianza.

Per definire completamente il modello lineare posso stimare anche la varianza.
L'SQE diviso per un numero di gradi di libertà, in questo caso n-(k+1), è uno stimatore della varianza.

$$Var\left(\hat{\beta}_0\right) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}} \right]$$

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}}$$

$$Var\left(\hat{Y_x}\right) = \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}} \right]$$

È possibile ora calcolare gli intervalli di confidenza sui parametri
$E(Y|X = x)$, β_0, β_1.

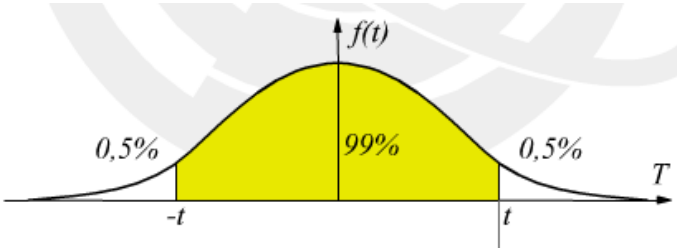
Siccome lo stimatore $\hat{Y}_x$ è una variabile aleatoria Normale, l'intervallo di confidenza potrebbe essere calcolato usando la Normale Standard conoscendo la varianza.

Tuttavia si preferisce assumere il caso più generale ed utilizzare una T di Student con un livello di confidenza 1-α , con v= n-2 gradi di libertà.

$$T = \frac{\hat{Y}_x - \mathrm{E}(Y|X=x)}{S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}}$$

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}}$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{S\sqrt{\frac{1}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}}$$



Ottenendo:

$$\Pr\left\{\hat{Y}_x - t_{1-\alpha/2;v} S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}} < \mathrm{E}(Y \mid X = x) \leq \hat{Y}_x + t_{1-\alpha/2;v} S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}\right\} = 1-\alpha$$

$$\Pr\left\{\hat{\beta}_0 - t_{1-\alpha/2;v} S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}} < \beta_0 \leq \hat{\beta}_0 + t_{1-\alpha/2;v} S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}\right\} = 1-\alpha$$

$$\Pr\left\{\hat{\beta}_1 - t_{1-\frac{\alpha}{2};v} S\sqrt{\frac{1}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}} < \beta_1 \leq \hat{\beta}_1 + t_{1-\frac{\alpha}{2};v} S\sqrt{\frac{1}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}\right\} = 1-\alpha$$

5.1.2 P-value

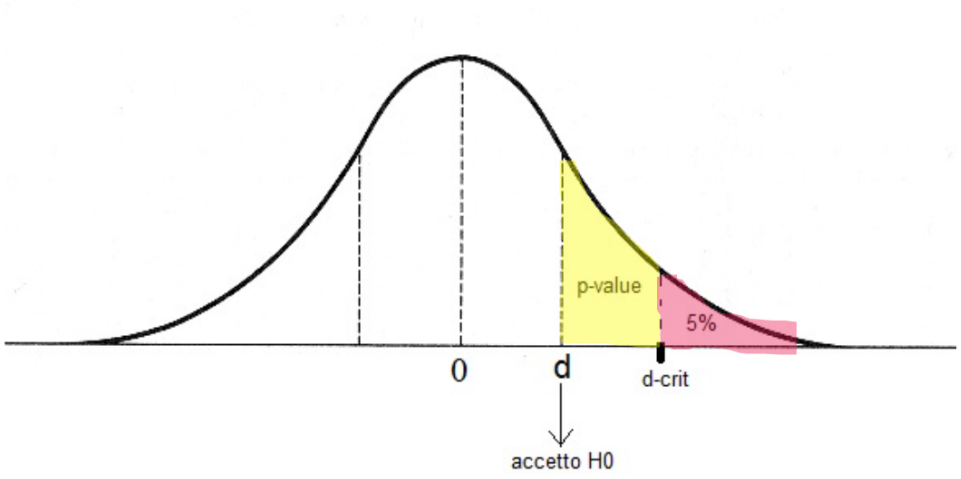
Il p-value, nell'ambito dei test di ipotesi, è definito come la probabilità, sotto ipotesi nulla, che il valore ottenuto dalla statistica pivot sia più estremale della stima generata dai dati; in altre parole il p-value quantifica la **forza del rifiuto dell'ipotesi nulla**, in quanto un p-value basso indica che l'ipotesi nulla non spiega adeguatamente i dati osservati, cioè è poco credibile che il valore osservato sia stato effettivamente estratto dalla statistica pivot sotto ipotesi H_0. Indicando con T_n la statistica e t_n la stima, il p-value p si definisce come:

1. $p = Pr\{|T_n| > |t_n||H_0\}$, in caso di test bilaterale;
2. $p = Pr\{T_n > t_n|H_0\}$, in caso di test unilaterale con $H_0 : \mu \leq \mu_0$;
3. $p = Pr\{T_n < t_n|H_0\}$, in caso di test unilaterale con $H_0 : \mu \geq \mu_0$.

Il p-value viene utilizzato confrontandolo con il rischio di prima specie α:

1. se $p < \alpha$, si rifiuta H_0
2. se $p > \alpha$, non si rifiuta H_0

Il vantaggio di usare il p-value invece di verificare se la stima è inclusa nella regione di accettazione è che il p-value non dipende dal valore di α fissato al contrario della regione di accettazione, quindi α può essere fissato anche a posteriori, al momento della decisione.



Caso di test unilaterale con $H_0 : \mu \leq \mu_0$ (p è evidenziato in giallo mentre la regione critica con $\alpha = 0.05$ è evidenziata in rosso); $p > \alpha$ quindi si accetta H_0.

5.1.3 Intervalli di predizione su un futuro valore della variabile aleatoria Y

Oltre agli intervalli di confidenza esiste anche un altro problema inferenziale che riguarda **la misura dell'incertezza sul valore che verrà assunto dalla variabile dipendente Y in un futuro esperimento in corrispondenza di un assegnato valore x.**

La logica è la stessa degli intervalli di confidenza ovvero determinare un intervallo che nella ripetizione dell'esperimento ha una probabilità 1-α di contenere il futuro valore osservato della variabile aleatoria Y.

Siccome sia Y che $\hat{Y}_x$ sono variabili aleatorie normali con lo stesso valore atteso $E(Y|X = x)$ ma diversa varianza, la variabile aleatoria data dalla differenza delle variabili aleatorie sarà anch'essa una normale a media 0 e varianza:

$$Var\left(\hat{Y}_x - Y_x\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}}\right]$$

In quanto le variabili sono indipendenti tra loro.

L'intervallo di predizione può essere stimato con una Normale Standard come statistica pivot nel caso la varianza sia nota. Tuttavia in via più generale si usa una T-di Student con $S = \sqrt{MSQE}$

$$T = \frac{\hat{Y}_x - Y}{S\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x^2}}}}$$

E quindi l'intervallo di predizione isolando Y :

$$\Pr\left\{\hat{Y}_x - t_{1-\frac{\alpha}{2};v} S\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}}\right\} < \ Y \ \leq \hat{Y}_x + t_{1-\frac{\alpha}{2};v} S\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}} = 1-\alpha \tag{5.1}$$

Si nota che gli intervalli di predizione sono più ampi rispetto a quello di confidenza in quanto, dovendo generare Y c'è un σ^2 dato dalla varianza di Y che si aggiunge.

5.1.4 Intervalli di confidenza dei parametri nella regressione multipla

Il metodo da applicare è quello dei minimi quadrati tuttavia deve essere considerato il vettore dei parametri anziché il parametro singolo

$$\underline{\hat{\beta}} = \arg\min \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij}\right)^2$$

Per trovare il minimo la matrice deve essere definita positiva, ovvero deve avere tutti gli autovalori positivi.

Ponendo il gradiente uguale a 0 :

$$\nabla RSS\left(\underline{\beta}\right) = 0$$

si ottiene :

$$\underline{\hat{\beta}} = \left(\underline{\underline{x}}^T \, \underline{\underline{x}}\right)^{-1} \underline{\underline{x}} \, \underline{y}$$

Che ha varianza:

$$VAR \ = \ \left(\underline{\underline{x}}^T \, \underline{\underline{x}}\right)^{-1} \ \sigma^2$$

Considerando come statistica pivot una T di student posso anche calcolare l'intervallo di confidenza di B_j con incertezza 1-α

$$\Pr\left\{ \ \hat{B}_j \ - \ t_{1-\frac{\alpha}{2};n-p-1} \ SE\left(\ \hat{B}_j \ \right) \ \leq \ B_j \ \leq \ \hat{B}_j \ + \ t_{1-\frac{\alpha}{2};n-p-1} \, SE \ \left(\ \hat{B}_j \ \right) \ \right\} = 1-\alpha$$

5.2 Commento sul codice

Dopo aver trovato diversi modelli di regressione lineare abbiamo scelto il modello step_6 per calcolare gli intervalli di confidenza dei parametri attraverso la stima ai minimi quadrati in modo manuale ovvero applicando le formule illustrate precedentemente con un livello di confidenza 95%.

Ricavo i parametri per costruire gli intervalli.

```
b0 = step_6$coefficients[1];

b1 = step_6$coefficients[2];

y.fitted=step_6$fitted.values

y=training$y_ImageQuality

x=training$x6_FOCAL

n=length(y)

sqe=sum((y - y.fitted)^2);

msqe=sqe/(n-2); msqe

S=sqrt(msqe);

t.val = qt(0.975, n - 2)
```

Il modello, dopo aver ricavato i parametri è:

$$y = 77.25 \ - \ 5.10x_6$$

I dati del training set quindi sono stati usati per calcolare i valori stimati di y,l'SQE e l'MSQE insieme ai quantili della statistica test della T di Student ad 1-$\frac{\alpha}{2}$ per costruire gli estremi inferiore e superiore dell'intercetta e di β_1.

```
> conf_int_b0=t.val*S*sqrt(1/n+( (mean(x))^2 / (sum(x^2)-n*(mean(x))^2 ) ) )
> L_b0=b0-conf_int_b0;L_b0
(Intercept)
   71.13973
> U_b0=b0+conf_int_b0;U_b0
(Intercept)
   83.35523
```

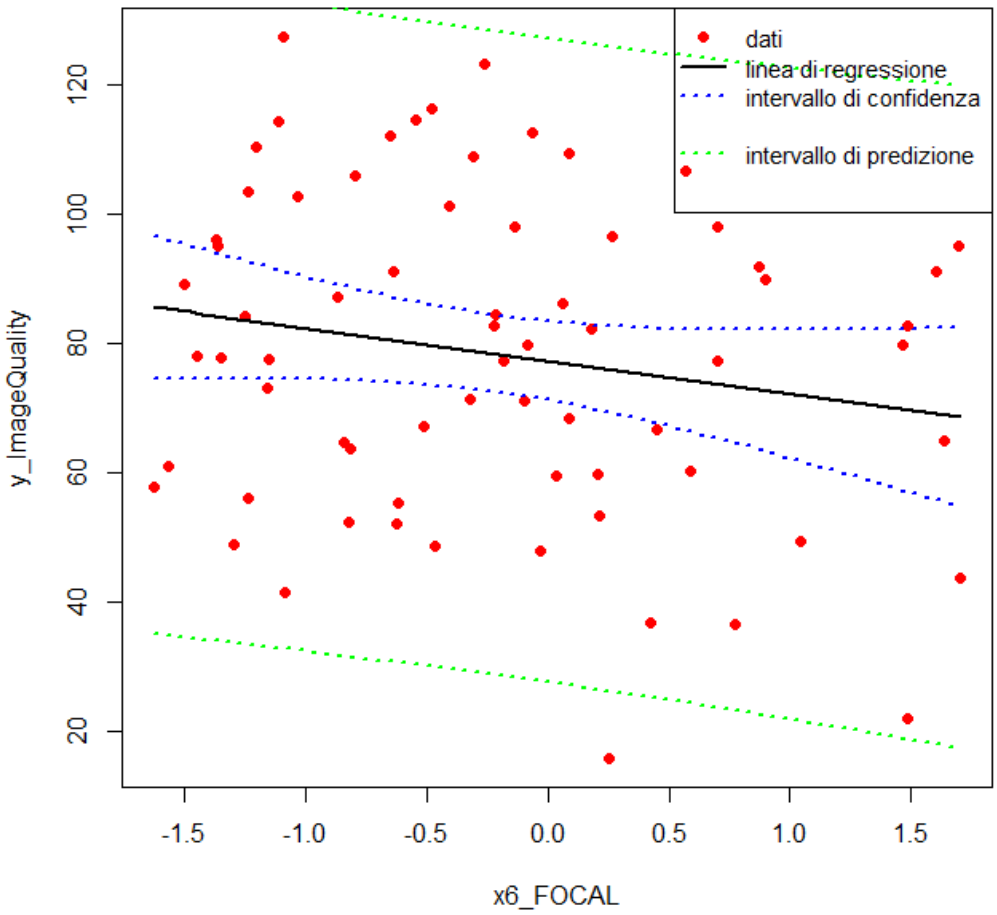
```
> conf_int_b1=t.val*S*sqrt( (1 / (sum(x^2)-n*(mean(x))^2 ) ) )
> L_b1=b1-conf_int_b1;L_b1
 x6_FOCAL
-11.65512
> U_b1=b1+conf_int_b1;U_b1
x6_FOCAL
1.456842
```

Si ricava che l'estremo inferiore per l'intercetta è 71.14 mentre quello superiore è 83.36.

L'estremo inferiore per B1 è -11.66 mentre quello superiore è 1.46.

I risultati calcolati manualmente sono stati poi confrontati con quelli calcolati dal comando confint(step_6) e risultano coincidere.

Per lo stesso modello sono stati calcolati anche gli intervalli di confidenza e predizione della Y e poi rappresentati in un grafico riassuntivo del modello.



Si può notare che quasi tutti i valori rientrano negli intervalli, inoltre il grafico è schiacciato nella parte centrale in quanto la maggioranza dei valori si concentra nel valore medio ciò significa che in mezzo il modello reale e quello osservato si sovrappongono.

Sono stati calcolati anche gli intervalli di confidenza per il miglior modello di regressione multipla trovato considerando il T-test che dopo il calcolo dei parametri risulta essere :

$$y = 83.70 - 5.74x_1 - 10.87x_2 - 8.52x_2^2 + 5.57x_3 - 3.08x_4 - 13.46x_5$$

Per ogni parametro sono stati stimati gli intervalli di confidenza con un livello di confidenza del 95%.

	2.5 %	97.5 %
(Intercept)	80.067427	87.3387800
I(x1_ISO^3)	-7.056481	-4.4348651
x2_FRatio	-13.259525	-8.4915520
I(x2_FRatio^2)	-11.584097	-5.4653303
x3_TIME	3.174998	7.9730280
x4_MP	-5.772219	-0.3801051
x5_CROP	-15.490466	-11.4335347

Una volta scelto il modello finale sono stati stimati i coefficienti, di conseguenza sostituendo risulta :

$$y = 88.48 - 5.13x_1^2 - 5.56x_1^3 - 11,71x_2 - 9,16x_2^2 + 5.27x_3 - 2.39x_4 - 13.89x_5$$

E i loro relativi intervalli di confidenza con incertezza pari al 95%.

	Estremo Inferiore	Estremo Superiore
Intercetta β_0	84,55	92,41
β_1	-7,55	-2,71
β_2	-6,72	-4,40
β_3	-13,86	-9,56
β_4	-11,89	-6,43
β_5	3,14	7,40
β_6	-4,80	0,02
β_7	-15,71	-12,08

	2.5 %	97.5 %
(Intercept)	84.546887	92.41014601
x5_CROP	-15.706084	-12.08697681
I(x1_ISO^3)	-6.722719	-4.39230863
x2_FRatio	-13.859605	-9.56083450
I(x2_FRatio^2)	-11.886788	-6.43060736
x3_TIME	3.141403	7.40358707
I(x1_ISO^2)	-7.550397	-2.71053592
x4_MP	-4.800249	0.02278943

CAPITOLO 6

REGRESSIONE MULTIPLA

6.1 Elementi di teoria sulla regressione multipla

Con regressione multipla si intende un modello in cui la variabile dipendente è condizionata da più di un regressore. Il caso più generale è il seguente (p è il numero di regressori)

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

> Un caso particolare di regressione multipla è la regressione multipla lineare, in cui si ha linearità rispetto ai coefficienti (e non necessariamente rispetto ai regressori)
> $$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = \beta_0 + \sum_1^p \beta_j X_j + \epsilon$$

Per poter fare inferenza sull'efficacia del modello e la stima dei parametri, assumiamo l'errore ϵ sia distribuito secondo una Normale a media nulla e varianza finita e costante indipendentemente dal valore assunto dai regressori (condizione di omoschedasticità).

Per la stima dei coefficienti di regressione e per l'inferenza sull'errore si ricorre all'analisi matriciale; è quindi utile scrivere l'insieme dei valori assunti dalla variabile indipendente ($\underline{Y}$), dai coefficienti ($\underline{\beta}$) e dall'errore ($\underline{\epsilon}$) sottoforma di vettori colonna di dimensione n x 1, (p+1) x 1 e n x 1 rispettivamente.
I valori assunti dai regressori per ogni i-esimo campione vengono riuniti nella matrice dei dati ($\underline{\underline{X}}$) che il valore 1 per l'intera prima colonna per includere i coefficienti β_0 e ha dimensione n x (p+1).
Da questa impostazione abbiamo che l'equazione di regressione è definita in questo modo

$$\underline{Y} = \underline{\underline{X}}\underline{\beta} + \underline{\epsilon}$$

e l'errore ϵ è distribuito secondo una Multivariata Gaussiana con media $\underline{0}$ e matrice di varianza-covarianza $\sigma^2 I_n$ (dove I_n è la matrice identità di ordine n).
Per quanto riguarda la stima ai minimi quadrati dei coefficienti essa si ottiene, nel caso $n > p+1$, come il vettore di coefficienti che riduce al minimo la somma dei qudrati dei residui (RSS).

$$RSS(\underline{\beta}) = (\underline{Y} - \underline{\underline{X}}\underline{\beta})^T (\underline{Y} - \underline{\underline{X}}\underline{\beta})$$

Si pongono le derivate prime dell'RSS a zero (e l'Hessiano deve essere definito positivo per ricavare il minimo) e si ottiene

$$\underline{\hat{\beta}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}$$

$$\underline{\hat{Y}} = \underline{\underline{X}} \underline{\hat{\beta}} = \underline{\underline{X}} (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y} = \underline{\underline{H}} \underline{Y}$$

$\underline{\underline{H}}$ è la **matrice di predizione**, che fa proiettare $\underline{Y}$ sullo spazio vettoriale generato dai regressori.

Per poter fare inferenza sulla stima dei coefficienti (intervalli di confidenza e test di ipotesi) bisogna conoscere la varianza dello stimatore di $\underline{\hat{\beta}}$. Essa si calcola in relazione alla varianza dell'errore ϵ ottenendo la **matrice di varianza-covarianza** $(\underline{X}^T \underline{X})^{-1}\sigma^2$.
Nella maggioranza dei casi la varianza σ^2 non è nota e va stimata a partire dai dati. Sappiamo che dividendo l'RSS per un opportuno numero di gradi di libertà otteniamo uno stimatore non distorto di σ^2. In questo caso i gradi di libertà a disposizione per stimare la varianza sono $n-(p+1)$

$$\hat{\sigma}^2 = \frac{RSS(\underline{\hat{\beta}})}{n-(p+1)} = \frac{\sum_{i=1}^{p}(Y_i - \hat{Y}_i)^2}{n-(p+1)}$$

In definitiva lo stimatore di $\underline{\hat{\beta}}$ è distribuito secondo una Multivariata Gaussiana con media $\underline{\beta}$ e matrice varianza covarianza $(\underline{X}^T \underline{X})^{-1}\sigma^2$.

6.2 Coefficiente di determinazione

Il coefficiente di determinazione R^2 rappresenta l'adeguatezza del modello di regressione ad interpretare il contesto sperimentale d'interesse.
E' quindi una misura dell'importanza relativa che l'intero insieme di variabili indipendenti X_i ha nell'interpretare il comportamento della variabile dipendente Y.

$$R^2 = \frac{SQR}{SQTOT}$$

SQTOT rappresenta la variabilità associata al modello quindi il coefficiente di determinazione misura

$$\frac{variabilitàregressione}{variabilitàtotale}$$

Essendo associato a somme di quadrati

$$0 < R^2 \leq 1$$

$R^2 = 1$ spiega la variabilità totale, implica SQE=0 quindi i valori stimati coincidono con i valori osservati.

$R^2 = 0$ implica SQR=0 ovvero il modello di regressione non fornisce alcun contributo alla interpretazione del fenomeno osservato.

6.2.1 Relazione tra coefficiente di correlazione, β_1 e R^2

Siccome è possibile scrivere β_1 come

$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{r_{xy} S_y}{S_x}$

Dove S_xy è la covarianza campionaria,S_x e S_y sono le varianze campionare e r è il coefficiente di correlazione

Essendo $SQR = \beta_1^2 \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$

$R^2 = \frac{SQR}{SQTOT} = \frac{\beta_1^2 \cdot S_x^2}{S_y^2}$ in questo modo è possibile notare che è legato al coefficiente β_1 Sostituendo proprio β_1 si ottiene :

$$R^2 = (r_{xy})^2$$

È possibile dire che il coefficiente di determinazione e quello di correlazione sono direttamente legati tra loro, in particolare uno è il quadrato dell'altro.

6.3 Commenti sul codice

Abbiamo deciso di confrontare quattro modelli multipli lineari.
Il primo contiene tutti i regressori lineari. Eseguendo i seguenti comandi

```
lin_mult1 = lm(y_ImageQuality ~ x1_ISO + x2_FRatio + x3_TIME + x4_MP +
    x5_CROP + x6_FOCAL + x7_PixDensity, data = training)
summary(lin_mult1)
```

otteniamo un modello lineare dalle seguenti caratteristiche:

lin_mult1

$$y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-22.2852  -8.6184   0.2112   8.5184  23.0077

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    74.9708     1.5675  47.828  < 2e-16 ***
x1_ISO        -10.1720     1.7018  -5.977 1.28e-07 ***
x2_FRatio     -10.9134     1.6166  -6.751 6.20e-09 ***
x3_TIME         6.7200     1.5932   4.218 8.31e-05 ***
x4_MP          -3.9372     2.2502  -1.750   0.0852 .
x5_CROP       -13.5114     1.3740  -9.833 3.38e-14 ***
x6_FOCAL       -1.5647     1.7315  -0.904   0.3697
x7_PixDensity  -0.7766     2.2526  -0.345   0.7315
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.34 on 61 degrees of freedom
Multiple R-squared:  0.7815,    Adjusted R-squared:  0.7564
F-statistic: 31.16 on 7 and 61 DF,  p-value: < 2.2e-16
```

L'indice di determinazione è $R^2 = 0.7815$
Ponendo come rischio di prima specie $\alpha = 0.05$ risulta che i regressori x_4, x_6 e x_7 non hanno un effetto statisticamente significativo sulla quantità Y.

Abbiamo deciso di rimuovere i suddetti regressori ottenendo con i seguenti comandi il secondo modello lineare multiplo

```
lin_mult2 = lm(y_ImageQuality ~ x1_ISO + x2_FRatio + x3_TIME + x5_CROP,
    data = training)
summary(lin_mult2)
```

con le seguenti caratteristiche:

lin_mult2

$$y \sim x_1 + x_2 + x_3 + x_5$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-28.494  -7.728   1.187   7.795  27.796

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.539      1.559  47.810  < 2e-16 ***
x1_ISO       -10.935      1.707  -6.405 2.03e-08 ***
x2_FRatio    -11.011      1.607  -6.850 3.41e-09 ***
x3_TIME        5.877      1.586   3.705 0.000442 ***
x5_CROP      -13.624      1.367  -9.965 1.21e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.73 on 64 degrees of freedom
Multiple R-squared:  0.7562,    Adjusted R-squared:  0.741
F-statistic: 49.64 on 4 and 64 DF,  p-value: < 2.2e-16
```

L'indice di determinazione è $R^2 = 0.7562$; è diminuito di poco perchè il modello è diventato più semplice.
Dai T-Test osserviamo che adesso tutte le variabili indipendenti sono statisticamente significative.

Possiamo aumentare il coefficiente di determinazione includendo termini polinomiali di grado maggiore al primo. Abbiamo deciso di creare il terzo modello lineare includendo termini polinomiali suggeriti dai modelli polinomiali a singolo fattore (il settimo regressore è completamente escluso dato che il suo modello polinomiale singolo è del tipo Y 1). Quindi, eseguendo i seguenti comandi

```
lin_mult3 = lm(y_ImageQuality ~
                  x1_ISO + I(x1_ISO^3) +
                  x2_FRatio + I(x2_FRatio^2) +
                  x3_TIME +
                  x4_MP + I(x4_MP^2) +
                  x5_CROP + I(x5_CROP^2) +
                  x6_FOCAL,
                  data = training)
summary(lin_mult3)
```

otteniamo un modello lineare dalle seguenti caratteristiche:

lin_mult3

$$y \sim x_1 + x_1^3 + x_2 + x_2^2 + x_3 + x_4 + x_4^2 + x_5 + x_5^2 + x_6$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-28.0529  -6.0489  -0.1763   6.2300  18.2179

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     85.448977   2.671848  31.981  < 2e-16 ***
x1_ISO          -0.004073   2.950373  -0.001 0.998903
I(x1_ISO^3)     -5.758357   1.533484  -3.755 0.000404 ***
x2_FRatio      -11.366322   1.255307  -9.055 1.08e-12 ***
I(x2_FRatio^2)  -8.819661   1.531854  -5.758 3.41e-07 ***
x3_TIME          5.965392   1.235038   4.830 1.04e-05 ***
x4_MP           -3.456249   1.484262  -2.329 0.023387 *
I(x4_MP^2)      -0.043136   1.537931  -0.028 0.977720
x5_CROP        -12.947371   1.061653 -12.195  < 2e-16 ***
I(x5_CROP^2)    -1.451466   1.312450  -1.106 0.273327
x6_FOCAL        -2.519739   1.353665  -1.861 0.067755 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.361 on 58 degrees of freedom
Multiple R-squared:  0.8805,    Adjusted R-squared:  0.8599
F-statistic: 42.72 on 10 and 58 DF,  p-value: < 2.2e-16
```

L'indice di determinazione è $R^2 = 0.8805$; è decisamente aumentato rispetto ai precedenti modelli data l'aumentata complessità e un miglior "fitting".
Dai T-Test osserviamo che x_1, x_4^2, x_5^2 e x_6 non hanno un effetto statisticamente significativo sul valore di Y.

Abbiamo deciso di rimuovere i suddetti regressori ottenendo con i seguenti comandi il quarto e ultimo modello lineare multiplo

```
lin_mult4 = lm(y_ImageQuality ~
                  I(x1_ISO^3) +
                  x2_FRatio + I(x2_FRatio^2) +
                  x3_TIME +
                  x4_MP +
                  x5_CROP,
                  data = training)
summary(lin_mult4)
```

con le seguenti caratteristiche:

lin_mult4

$$y \sim x_1^3 + x_2 + x_2^2 + x_3 + x_4 + x_5$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-28.2689  -6.1076   0.3892   6.8877  17.9072

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     83.7031     1.8188  46.022  < 2e-16 ***
I(x1_ISO^3)     -5.7457     0.6557  -8.762 1.90e-12 ***
x2_FRatio      -10.8755     1.1926  -9.119 4.65e-13 ***
I(x2_FRatio^2)  -8.5247     1.5305  -5.570 5.85e-07 ***
x3_TIME          5.5740     1.2001   4.645 1.82e-05 ***
x4_MP           -3.0762     1.3487  -2.281    0.026 *
x5_CROP        -13.4620     1.0148 -13.266  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.415 on 62 degrees of freedom
Multiple R-squared:  0.8707,    Adjusted R-squared:  0.8582
F-statistic: 69.62 on 6 and 62 DF,  p-value: < 2.2e-16
```

L’indice di determinazione è $R^2 = 0.8707$; è diminuito di poco perchè il modello è diventato più semplice.
Dai T-Test osserviamo che tutti i regressori sono statisticamente significativi.

CAPITOLO 7

MODEL SELECTION

7.1 Elementi di teoria sulla Model Selection

7.1.1 Criteri di scelta del modello

Nell'ambito della regressione polinomiale e multipla risulta evidente che è possibile descrivere l'evoluzione della variabile dipendente utilizzando diversi set di regressori ottenendo modelli diversi. Risulta necessario avere a disposizione uno o più criteri per selezionare il modello migliore per la nostra applicazione.

Abbiamo visto come l'**indice di determinazione** R^2 permette di identificare la percentuale della variabilità che viene spiegata dal modello rispetto alla variabilità totale e quindi quantifica la precisione del modello. Nonostante ciò l'indice di determinazione non è un buon criterio da utilizzare senza l'ausilio di altri perchè tende ad aumentare sempre all'aggiunta di regressori, ma non ci indica se quest'aggiunta sia statisticamente significativa nel migliorare effettivamente il nostro modello.

Nel caso in cui vogliamo confrontare due modelli che differiscono per l'aggiunta o rimozione di un singolo regressore, possiamo ricorrere a questi quattro criteri o metodi.

1. **T-TEST** test di ipotesi che valuta se il regressore X_i è statisticamente significativo nel modello di regressione. Ha come ipotesi nulla $H_0 : \beta_i = 0$ e come ipotesi alternativa $H_A : \beta_i \neq 0$ e si basa sulla seguente statistica pivot (Sd indica la deviazione standard)

$$\frac{\hat{\beta}_i - \beta_i}{Sd(\hat{\beta}_i)}$$

   che sotto H_0 è distribuita secondo una **T di Student** con $n - (p + 1)$ gradi di libertà ν. Ne deriva la seguente regione di accettazione I_{ac}

$$I_{ac} : -t_{1-\alpha/2;\nu} Sd(\hat{\beta}_i) < \hat{\beta}_i < t_{1-\alpha/2;\nu} Sd(\hat{\beta}_i)$$

   Alternativamente si rifiuta H_0 se il p-value è minore di α.

2. **F-TEST (ANOVA):** possiamo eseguire lo stesso test di ipotesi considerando che la variabilità totale dell'esperimento $SQTOT$ può essere divisa in variabilità dovuta alla regressione SQR e variabilità non controllata SQE. Possiamo affermare che il modello 2

(che aggiunge il regressore X_i) è più preciso del modello 1 se $SQE2 < SQE1$, quindi la differenza $SQE2 - SQE1$ rappresenta il contributo che il modello 2 introduce rispetto al modello 1 nel ridurre la variabilità non controllata. Risulta che $SQE1 - SQE2$ e $SQE2$, divisi per un opportuno numero di gradi di libertà, sono entrambi stimatori non distorti del parametro σ^2 (sotto ipotesi H_0).

A questo punto possiamo eseguire il test basandoci sulla seguente statistica pivot

$$F = \frac{(SQE1 - SQE2)/\nu_1}{SQE2/\nu_2}$$

distribuita secondo una **Fisher-Snedecor** con $\nu_1 = 1$ e $\nu_2 = n-(p+1)$, dove p è il numero di regressori nel modello 2. Ne deriva la seguente regione di accettazione

$$I_{ac} : F < F_{1-\alpha;\nu_1,\nu_2}$$

Anche in questo caso alternativamente si rifiuta H_0 se il p-value è minore di α.

3. **AIC (Akaike Information Criterion):** è un valore che dipende dalla funzione di verosimiglianza del modello $L(\underline{\hat{\beta}})$, che quantifica l'adattamento del modello ai dati, e dalla dimensione del modello d, che è pari al numero di parametri da stimare nel modello (p+2). Quantifica la perdita di informazioni che si ha usando il modello per descrivere il processo che ha generato i dati, quindi il modello con AIC minore è migliore nell'adattarsi ai dati (minore underfitting) e non è troppo complesso e quindi non dipendende eccessivamente dai dati su cui è stato allenato (minore overfitting). Si definisce con la seguente formula

$$AIC = 2d - 2logL(\underline{\hat{\beta}})$$

4. **BIC (Bayesian Information Criterion):** è analogo all'AIC, ma penalizza maggiormente l'introduzione di più regressori

$$BIC = log(n)d - 2logL(\underline{\hat{\beta}})$$

7.1.2 Stepwise Selection

Teoricamente per scegliere quale sia il modello migliore bisognerebbe confrontare tutti i modelli che si possono ottenere con i p regressori a disposizione, ossia confrontare tra di loro 2^p modelli usando uno dei criteri esposti. Quando il numero di regressori è troppo grande (genericamente oltre i 40) questo metodo diventa impraticabile.

La **stepwise selection** consiste nel partire da un modello, che può essere quello senza regressori, quello completo oppure uno intermedio, e aggiungere o rimuovere un regressore alla volta basando la decisione su uno dei criteri. Si prosegue fin quando si ritiene opportuno (si decide una regola di stop) oppure fin quando la decisione migliore secondo il criterio selezionato è nè aggiungere nè rimuovere uno dei regressori. In questo modo il numero di confronti (step) viene drasticamente ridotto e viene semplificato il problema della selezione del modello. Si può procedere principalmente in tre modi o direzioni:

1. **Forward**, ossia partendo dal modello senza regressori ($Y = \beta_0+\epsilon$) e aggiungendo regressori fin quando non si raggiunge la regola di stop;
2. **Backward**, ossia partendo dal modello completo e rimuovendo regressori fin quando non si raggiunge la regola di stop;
3. **Hybrid**, ossia partendo da un modello completo o intermedio e aggiungendo o rimuovendo regressori.

7.2 Commenti sul codice

Abbiamo deciso di utilizzare la Stepwise Selection usando i criteri AIC e BIC e impostando come direzione Forward (partendo dal modello $Y \sim 1$) e Hybrid (partendo dall'ultimo modello lineare multiplo, ossia lin_mult4), ottenendo quattro modelli in totale. Lo scope utilizzato comprende tutti i regressori con grado uno, due e tre.

Il primo modello si ottiene con criterio AIC e direzione Forward eseguendo i seguenti comandi

```
best1 = step(lm(y_ImageQuality ~ 1 , data=training),scope = ~ x1_ISO + I
   (x1_ISO^2) + I(x1_ISO^3) +
               x2_FRatio + I(x2_FRatio^2) + I(x2_FRatio^3) +
               x3_TIME + I(x3_TIME^2) + I(x3_TIME^3) +
               x4_MP + I(x4_MP^2) + I(x4_MP^3) +
               x5_CROP + I(x5_CROP^2) + I(x5_CROP^3) +
               x6_FOCAL + I(x6_FOCAL^2) + I(x6_FOCAL^3) +
               x7_PixDensity + I(x7_PixDensity^2) + I(x7_PixDensity^3)
             ,direction="forward",trace=1)
summary(best1)
```

e ha le seguenti caratteristiche:

best1

$$y \sim x_1^2 + x_1^3 + x_2 + x_2^2 + x_3 + x_3^2 + x_4 + x_4^3 + x_5 + x_6 + x_7^2$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-17.2240  -4.9758  -0.7519   4.9154  16.8064

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         87.1673     2.2683  38.428  < 2e-16 ***
x5_CROP            -13.7290     0.8940 -15.356  < 2e-16 ***
I(x1_ISO^3)         -5.8616     0.5764 -10.169 2.03e-14 ***
x2_FRatio          -12.3955     1.0393 -11.927  < 2e-16 ***
I(x2_FRatio^2)      -9.5550     1.3100  -7.294 1.03e-09 ***
x3_TIME              5.1788     1.0219   5.068 4.54e-06 ***
I(x1_ISO^2)         -4.8688     1.1732  -4.150 0.000112 ***
x4_MP               -5.9819     2.4538  -2.438 0.017918 *
I(x4_MP^3)           2.3404     1.3153   1.779 0.080506 .
x6_FOCAL            -1.4603     1.1239  -1.299 0.199069
I(x3_TIME^2)         2.1496     1.1935   1.801 0.076983 .
I(x7_PixDensity^2)  -1.4657     0.8937  -1.640 0.106488
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.892 on 57 degrees of freedom
Multiple R-squared:  0.9165,    Adjusted R-squared:  0.9004
F-statistic: 56.88 on 11 and 57 DF,  p-value: < 2.2e-16
```

Dal primo comando otteniamo $AIC = 295.90$, mentre dal secondo $R^2 = 0.9165$, un valore maggiore dei precedenti modelli lineari multipli in quanto in modello e più complesso.

Il secondo modello si ottiene con criterio AIC e direzione Hybrid eseguendo i seguenti comandi:

```
best2 = step(lm(y_ImageQuality ~
                  I(x1_ISO^3) +
                  x2_FRatio + I(x2_FRatio^2) +
                  x3_TIME +
                  x4_MP +
                  x5_CROP, data=training),
             scope = ~ x1_ISO + I(x1_ISO^2) + I(x1_ISO^3) +
               x2_FRatio + I(x2_FRatio^2) + I(x2_FRatio^3) +
               x3_TIME + I(x3_TIME^2) + I(x3_TIME^3) +
               x4_MP + I(x4_MP^2) + I(x4_MP^3) +
               x5_CROP + I(x5_CROP^2) + I(x5_CROP^3) +
               x6_FOCAL + I(x6_FOCAL^2) + I(x6_FOCAL^3) +
               x7_PixDensity + I(x7_PixDensity^2) + I(x7_PixDensity^3)
             ,direction="both",trace=1)
summary(best2)
```

e ha le seguenti caratteristiche:

best2

$$y \sim x_1^2 + x_1^3 + x_2 + x_2^2 + x_3 + x_3^2 + x_4 + x_4^3 + x_5 + x_6 + x_7^2$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-17.2240  -4.9758  -0.7519   4.9154  16.8064

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         87.1673     2.2683  38.428  < 2e-16 ***
I(x1_ISO^3)         -5.8616     0.5764 -10.169 2.03e-14 ***
x2_FRatio          -12.3955     1.0393 -11.927  < 2e-16 ***
I(x2_FRatio^2)      -9.5550     1.3100  -7.294 1.03e-09 ***
x3_TIME              5.1788     1.0219   5.068 4.54e-06 ***
x4_MP               -5.9819     2.4538  -2.438 0.017918 *
x5_CROP            -13.7290     0.8940 -15.356  < 2e-16 ***
I(x1_ISO^2)         -4.8688     1.1732  -4.150 0.000112 ***
I(x4_MP^3)           2.3404     1.3153   1.779 0.080506 .
x6_FOCAL            -1.4603     1.1239  -1.299 0.199069
I(x3_TIME^2)         2.1496     1.1935   1.801 0.076983 .
I(x7_PixDensity^2)  -1.4657     0.8937  -1.640 0.106488
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.892 on 57 degrees of freedom
Multiple R-squared:  0.9165,    Adjusted R-squared:  0.9004
F-statistic: 56.88 on 11 and 57 DF,  p-value: < 2.2e-16
```

Osserviamo che cambiando l'ordine il modello ottenuto è esattamente lo stesso del precedente.

Il terzo modello si ottiene con criterio BIC e direzione Forward eseguendo i seguenti comandi (il parametro k=log(n) ci permette di basarci su BIC e non su AIC)

```
best3 = step(lm(y_ImageQuality ~ 1 , data=training),scope = ~ x1_ISO + I
   (x1_ISO^2) + I(x1_ISO^3) +
               x2_FRatio + I(x2_FRatio^2) + I(x2_FRatio^3) +
               x3_TIME + I(x3_TIME^2) + I(x3_TIME^3) +
               x4_MP + I(x4_MP^2) + I(x4_MP^3) +
               x5_CROP + I(x5_CROP^2) + I(x5_CROP^3) +
               x6_FOCAL + I(x6_FOCAL^2) + I(x6_FOCAL^3) +
               x7_PixDensity + I(x7_PixDensity^2) + I(x7_PixDensity^3)
             ,trace=1, direction = "forward", k=log(n))
summary(best3)
```

e ha le seguenti caratteristiche:

best3

$$y \sim x_1^2 + x_1^3 + x_2 + x_2^2 + x_3 + x_4 + x_5$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-17.3942  -6.5545  -0.7788   6.2467  15.8563

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      88.4785     1.9662  45.000  < 2e-16 ***
x5_CROP         -13.8965     0.9049 -15.356  < 2e-16 ***
I(x1_ISO^3)      -5.5575     0.5827  -9.537 1.06e-13 ***
x2_FRatio       -11.7102     1.0749 -10.894 6.08e-16 ***
I(x2_FRatio^2)   -9.1587     1.3643  -6.713 7.19e-09 ***
x3_TIME           5.2725     1.0657   4.947 6.23e-06 ***
I(x1_ISO^2)      -5.1305     1.2102  -4.239 7.72e-05 ***
x4_MP            -2.3887     1.2060  -1.981   0.0521 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.342 on 61 degrees of freedom
Multiple R-squared:  0.9002,    Adjusted R-squared:  0.8887
F-statistic: 78.57 on 7 and 61 DF,  p-value: < 2.2e-16
```

Dal primo comando otteniamo $BIC = 318.11$, mentre dal secondo $R^2 = 0.9002$, un valore di poco minore rispetto al modello basato su AIC.

Infine il quarto modello si ottiene con criterio BIC e direzione Hybrid eseguendo i seguenti comandi

```
best4 = step(lm(y_ImageQuality ~
                  I(x1_ISO^2) + I(x1_ISO^3) +
                  x2_FRatio + I(x2_FRatio^2) +
                  x3_TIME +
                  x5_CROP +
                  I(x7_PixDensity^2), data=training),
             scope = ~ x1_ISO + I(x1_ISO^2) + I(x1_ISO^3) +
               x2_FRatio + I(x2_FRatio^2) + I(x2_FRatio^3) +
               x3_TIME + I(x3_TIME^2) + I(x3_TIME^3) +
               x4_MP + I(x4_MP^2) + I(x4_MP^3) +
               x5_CROP + I(x5_CROP^2) + I(x5_CROP^3) +
               x6_FOCAL + I(x6_FOCAL^2) + I(x6_FOCAL^3) +
               x7_PixDensity + I(x7_PixDensity^2) + I(x7_PixDensity^3)
             ,direction="both",trace=1, k=log(n))
summary(best4)
```

e ha le seguenti caratteristiche:

best4

$$y \sim x_1^2 + x_1^3 + x_2 + x_2^2 + x_3 + x_4 + x_5$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-17.3942  -6.5545  -0.7788   6.2467  15.8563

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     88.4785     1.9662  45.000  < 2e-16 ***
I(x1_ISO^2)     -5.1305     1.2102  -4.239 7.72e-05 ***
I(x1_ISO^3)     -5.5575     0.5827  -9.537 1.06e-13 ***
x2_FRatio      -11.7102     1.0749 -10.894 6.08e-16 ***
I(x2_FRatio^2)  -9.1587     1.3643  -6.713 7.19e-09 ***
x3_TIME          5.2725     1.0657   4.947 6.23e-06 ***
x5_CROP        -13.8965     0.9049 -15.356  < 2e-16 ***
x4_MP           -2.3887     1.2060  -1.981   0.0521 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.342 on 61 degrees of freedom
Multiple R-squared:  0.9002,    Adjusted R-squared:  0.8887
F-statistic: 78.57 on 7 and 61 DF,  p-value: < 2.2e-16
```

Osserviamo che anche in questo caso cambiando l'ordine il modello non cambia.

In conclusione abbiamo scelto come modello migliore quello basato sul **criterio BIC** (best3), in quanto il criterio BIC penalizza maggiormente la complessità del modello e quindi evita maggiormente l'overfitting, fattore problematico quando si va a valutare la capacità di predizione del modello usando il test-set.

CAPITOLO 8

DIAGNOSTICA DEL MODELLO

8.1 Elementi di teoria della diagnostica del modello

Nell' analisi di regressione multipla è buona norma eseguire l'analisi dei residui, ovvero la fase di diagnostica del modello. Anche in questo caso è di fondamentale importanza lo scatter plot.

Analisi dei residui Per comprendere se il mio modello di regressione multipla si adatta al modello della popolazione, utilizziamo lo scatter plot dove, sull' asse orizzontale vi sono i valori fittati e sull' asse verticale i residui ($e_i = y_i – \hat{y}_i$).
Dunque se il residuo è a media nulla, allora possiamo concludere che il modello di regressione trovato approssima bene i dati. Nel caso in cui la variabilità è fissa si verifica la proprietà di omoschedasticità, ovvero quando una collezione di variabili aleatorie hanno variabilità fissa e finita. Nel caso in cui la variabilità non è fissa vi è la proprietà di eteroschedasticità. In tal caso il nostro modello non approssima bene i dati ed una soluzione è effettuare la trasformazione dei dati, ovvero tentare di trasformare i dati in modo che la varianza dell'errore sia più uniforme. Ad esempio, si può utilizzare il logaritmo.

QQ-Plot Inoltre un'altra parte fondamentale della diagnostica è la verifica della normalità utilizzando il Q-Q plot (il cui aspetto teorico è stato già discusso nei paragrafi precedenti). Sull' asse orizzontale va posizionato il quantile teorico(normale) e sull' asse orizzontale il quantile campionario: nel caso in cui i residui sono allineati alla bisettrice del primo e terzo quadrante allora si può approssimare la curva ad una normale, in caso contrario non è possibile fare tale approssimazione.

Outlier e leverage Infine un altro aspetto critico è il rilevamento degli outlier. Sempre servendoci di uno scatter plot disponiamo sull' asse orizzontale le i e sull' asse verticale τ_i, ovvero i residui standardizzati. Nel caso in cui il $|\tau_i|$ supera il valore 3, allora è un possibile outlier. Gli outlier sono possibili errori di trascrizione dei dati , dunque vanno eliminati in quanto potrebbero compromettere il nostro modello di regressione. Punti particolati sono i leverage point analizzati sullo scatter plot dove sull' asse verticale vi sono i residui e sull' asse orizzontale vi sono i valori fittati (i valori previsti dal modello). Tali punti possono influenzare

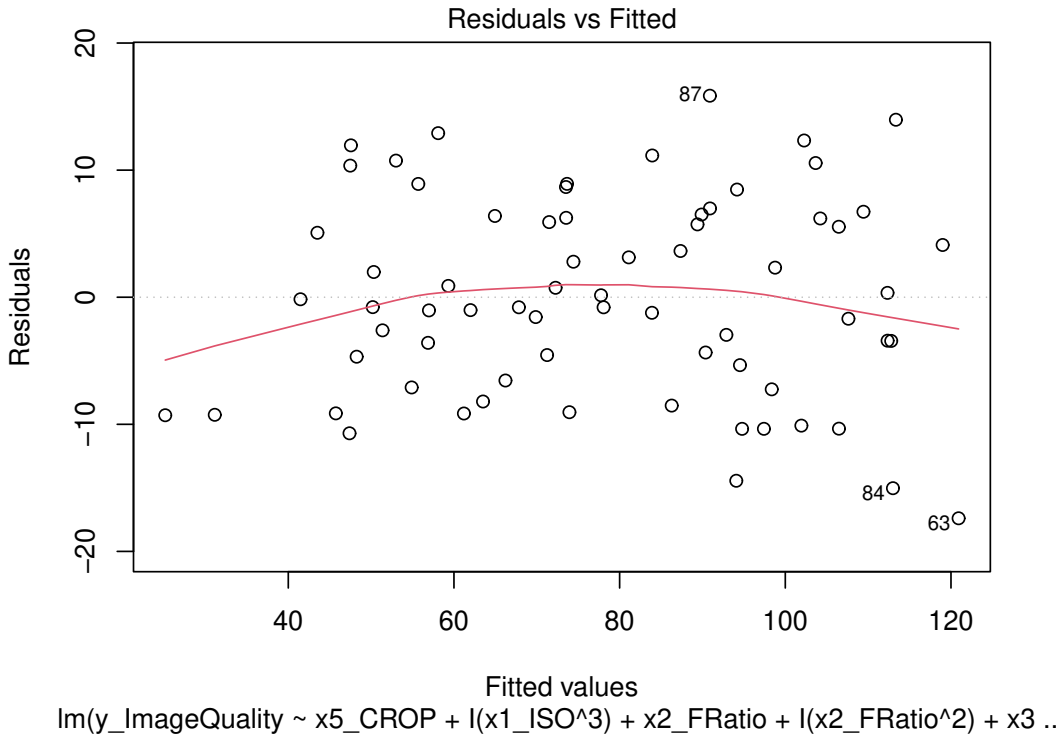
notevolmente la forma della regressione e avere un effetto significativo sulla stima dei coefficienti del modello. È importante individuare e gestire tali punti affinché il modello restituisca risultati affidabili e robusti. La distanza di Cook è una misura che indica l'effetto di un punto di leverage sul modello di regressione. La distanza di Cook è calcolata come la differenza tra la regressione originale e la regressione che esclude il punto di leverage, in relazione alla devianza totale del modello. Più alto è il valore della distanza di Cook per un punto di leverage, maggiore è l'effetto sul modello. In genere i punti di leverage con una distanza di Cook superiore ad una certa soglia sono considerati influenti e dunque potrebbero essere rimossi o presi in considerazione nell' interpretazione dei risultati.

8.2 Commento sul codice

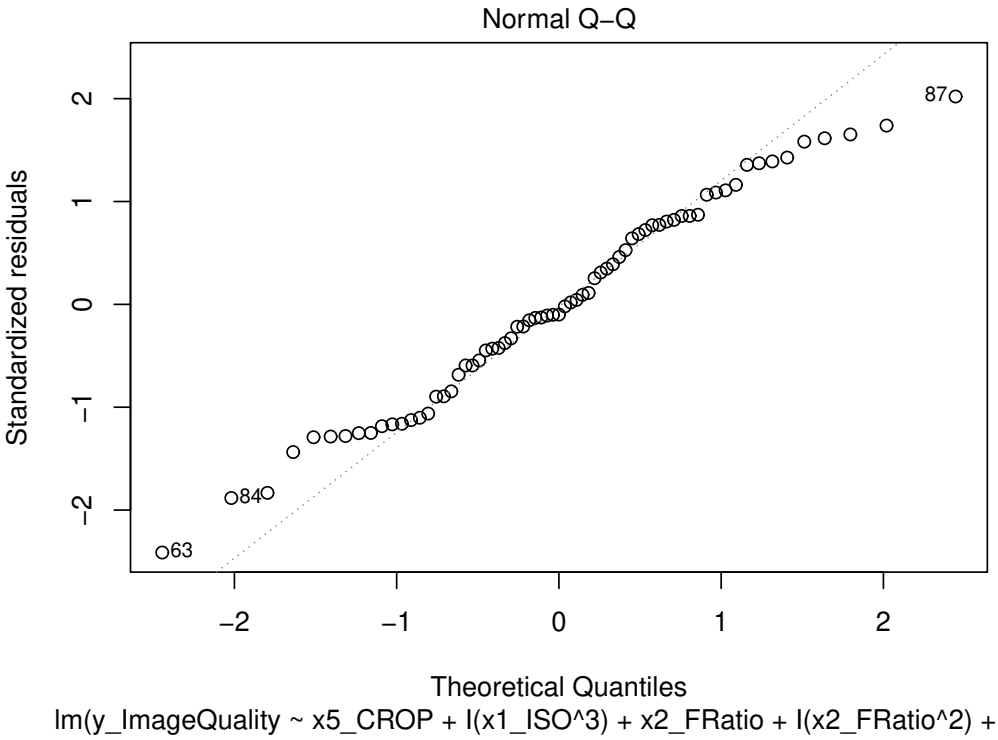
In questo caso per analizzare i valori fittati vs i residui, è stato utilizzato il comando:

```
1 plot(best3)
```

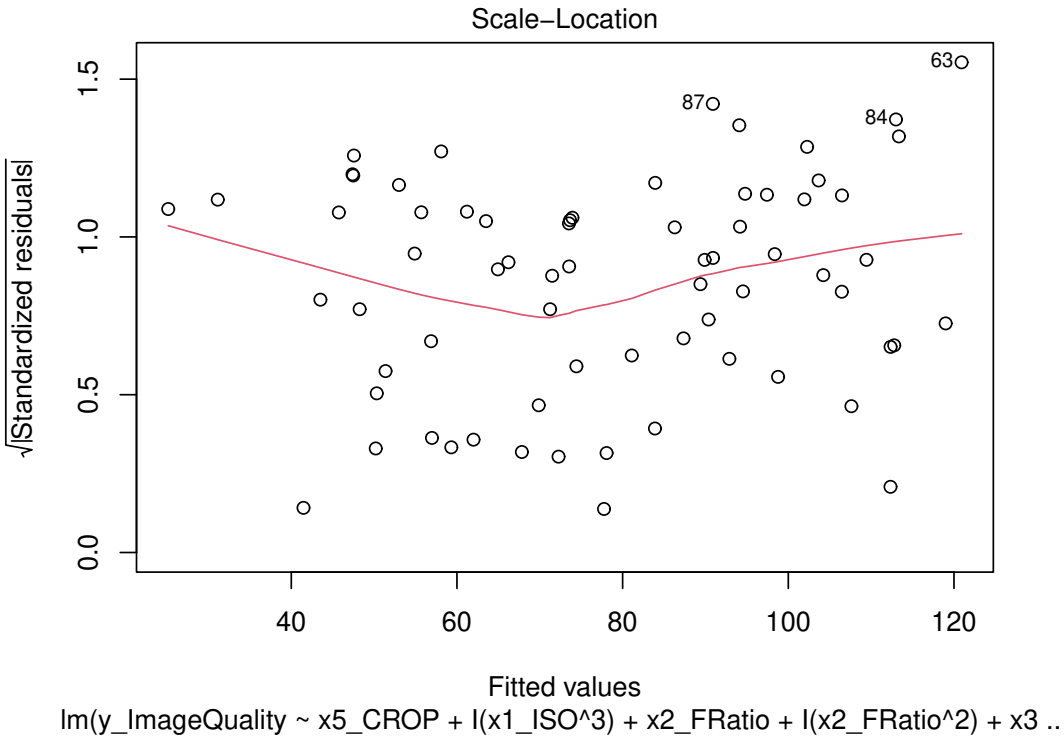
ottenendo tali grafici:



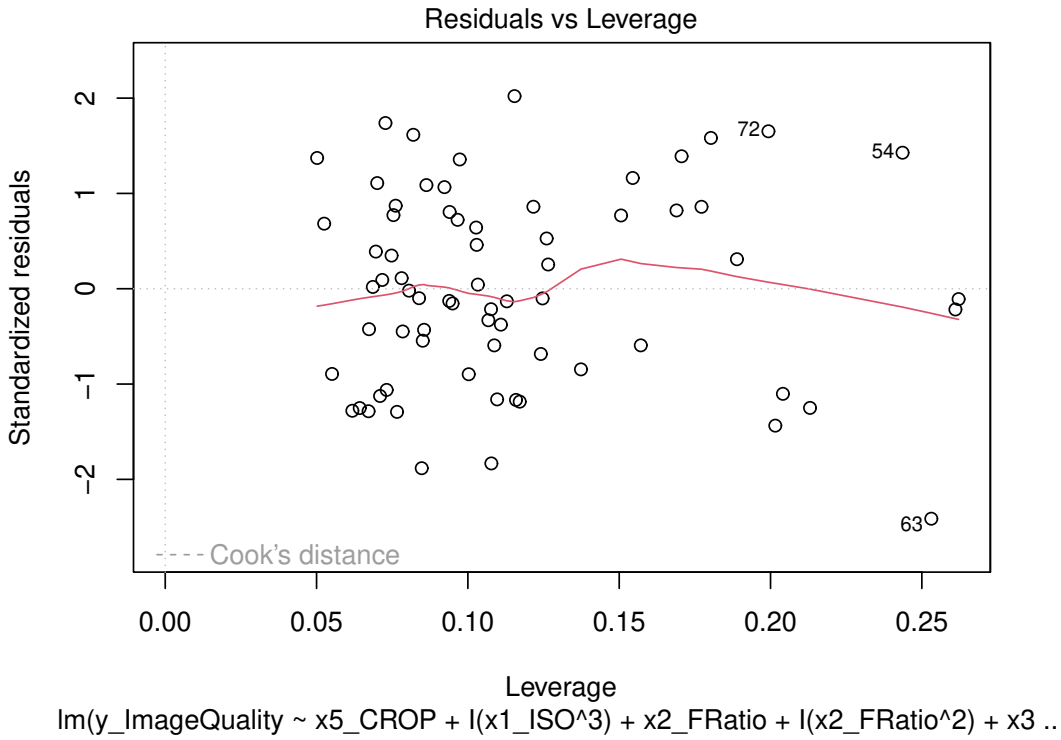
Si deduce che i residui si distribuiscono intorno all' asse delle ascisse con varianza abbastanza omogenea; dunque, è un buon modello e possiamo assumere l'omoschedasticità.



In questo caso notiamo come i residui si distribuiscono sulla bisettrice primo e terzo quadrante, dunque, si può assumere un andamento Gaussiano.



In tale grafico si apprezza come non vi siano outlier, in quanto non viene superato il valore massimo (valore assoluto maggiore di 3) dove però in questo caso è utilizzata la radice quadrata.



Infine da tale rappresentazione notiamo come non vi siano punti di Leverage rilevanti in quanto non viene superata la distanza di Cook.
Per la verifica della omoschedasticità è utilizzato il comando:

```
bptest(best3)
```

il quale indica come p-value $= 0.2226 > 0.05$ dunque si può affermare che l'ipotesi
H_0**: Presenza di omoschedasticità**
non può essere rifiutata.
Per avere dati analitici sulla normalità dei residui si utilizza il seguente comando:

```
shapiro.test(best3$residuals)
```

il quale produce p-value $= 0.3069$, ovvero è possibile approssimare la distribuzione dei residui ad una Gaussiana.

CAPITOLO 9

TRAINING SET E TEST SET

9.1 Training Set

Il training set è un insieme di dati che viene utilizzato per addestrare un modello di machine learning.

Il modello utilizza questi dati per ”imparare” come associare input e output desiderati. Il training set rappresenta un sottoinsieme dei dati totali a disposizione, che viene utilizzato per la formazione del modello. Durante il processo di addestramento, il modello utilizza i dati del training set per adattare i suoi parametri in modo da ottenere la migliore performance possibile.

9.2 Test Set

Il test set è l'insieme di dati (diverso da quello del training set) con cui si valuta la capacità di predizione di un modello.

In questa fase viene eseguito il testing error, calcolato come la media della differenza al quadrato tra le previsioni del modello e i valori osservati. Un errore di test basso indica che il modello ha una buona capacità di fare previsioni accurate su dati mai visti prima, mentre un errore di test alto indica una scarsa capacità di generalizzazione. Si deduce che il corrispondente del testing error in esame è comparabile con l' MSE calcolato sul training set.
Generalmente l' MSE del training set è più basso di quello del test set, in quanto i vari parametri del modello sono stati stimati sui dati proprio del training, però si può osservare che una differenza minima ci può far dire che non siamo in overfitting, ovvero che non abbiamo iper allenato il nostro modello sui dati iniziali. In caso contrario , ovvero iper addestrando il nostro modello, andiamo a ridurre il bias ma conseguentemente la varianza aumenterà. Ciò comporta che il nostro modello sarà ottimo per i dati utilizzati nel training ma pessimo per altri dati.
Per verificare il testing error è stato utilizzato i due comandi:

```
predictions = predict(best3,newdata=test);
predictions = predict(best3,newdata=test);
testing_error = mean((predictions - test$y_ImageQuality)^2);
```

producendo un errore di **104.8651**, il quale non si discosta molto dall' MSE calcolato sul training set mediante il comando:

```
mse_func=function(actual,predicted)
{
  mean( (actual-predicted)^2 )
}
mse_best3 = mse_func(training$y_ImageQuality, best3$fitted.values);
```

ottenendo il risultato di **61.52153**. Infatti analizzando l' RMSE(root mean square error) , il primo risulta circa **10,2** , mentre il secondo pari a **7.8**. RMSE può essere considerato come una sorta di distanza (normalizzata) tra il vettore dei valori previsti e il vettore dei valori osservati. Possiamo dedurre che il nostro modello potrebbe non essere in overfitting, dunque risulta essere non troppo specifico per i dati del training set (la variabilità non è eccessivamente grande).