

## A DATASET ANALYSIS

Design of a Linear Regression Model with RStudio

### Group 4

Mario Pellegrino Ambrosone  
Emanuele Barbato  
Luca Celentano  
Francesco De Bonis

0612707417  
0612707998  
0612707836  
0612708153

# A Perceived Video Quality Model Based on Shooting Parameters

M. P. Ambrosone, E. Barbato, L. Celentano, F. De Bonis

June 22, 2025

## A Brief Introduction: Goals and Expectations

La presente trattazione, da intendersi come appendice di documentazione al codice sviluppato con il FrameWork RStudio, ha come obiettivo la presentazione e la spiegazione dell'analisi del dataset "DataSet\_gruppo4-RAW", da ritrovarsi nella directory "Regression\_Analysis/data/", e dei risultati prodotti come output del codice, da ritrovarsi nella directory "Regression\_Analysis/results/", oltre che la discussione dei fondamenti teorici a sostegno dell'analisi stessa.

Il documento è pertanto sviluppato su più sezioni, ognuna delle quali è responsabilizzata per l'espansione e l'approfondimento di almeno una delle fasi di sviluppo dell'analisi del dataset.

Prima di procedere, è bene dunque descrivere brevemente il dataset di interesse, spiegare il ruolo dello stesso all'interno del progetto, e discutere dell'importanza dell'analisi statistica quando ci si interfaccia con i dati.

Il campione di dati fornito è stato invero interpretato come il training set di un modello di regressione lineare multipla, atto alla definizione di una relazione funzionale che potesse spiegare l'andamento della qualità video percepita in un campione di immagini ( $y_{VideoQuality}$ ) al variare dei parametri di scatto/ripresa e delle attrezzature utilizzate. Sui potenziali predictors, cioè per le variabili descritte come variabili indipendenti in "Regression\_Analysis/assignment/VariabiliProgStatApp1\_DEF\_24\_25", non sono state effettuate assunzioni ulteriori rispetto a quelle fornite nel file di cui sopra il path.

Tale scelta permette di valutare i dati mantenendo un livello di semplicità sufficiente alla definizione di un modello che possa spiegare l'andamento della variabile di risposta, e che non verrà quindi testato su un dataset differente.

L'obiettivo d'analisi è allora quello di modellare attentamente i dati attualmente disponibili, rivelandone relazioni e collegamenti, senza farne della predizione un problema centrale.

# Contents

<b>1</b>	<b>Data Handling and Preprocessing</b>	<b>4</b>
<b>2</b>	<b>Data Exploration</b>	<b>4</b>
2.1	Data Structure and Outliers Evaluation . . . . .	4
2.2	Variables Distribution . . . . .	7
<b>3</b>	<b>Interaction Analysis: Investigating Correlations</b>	<b>10</b>
3.1	Visualization and Statistical Hypothesis Testing . . . . .	10
3.2	Polynomial Regression Analysis for Capturing Non-Linear Effects . . . . .	12
<b>4</b>	<b>Defining Regression Models</b>	<b>13</b>
4.1	Linear and Polynomial Approaches . . . . .	13
4.2	A Formal and Graphical Diagnostic . . . . .	15
4.2.1	Complete Model . . . . .	16
4.2.2	Alternative Model 1 . . . . .	17
4.2.3	Alternative Model 2 . . . . .	18
<b>5</b>	<b>Model Comparison and Validation</b>	<b>19</b>
5.1	Comparison Strategies and Evaluative Procedures . . . . .	19
5.2	Backward Selection Outputs . . . . .	21
5.3	Model Comparison and Final Choice . . . . .	23
<b>6</b>	<b>Domain Insights: Relating Statistical Findings to Video Quality</b>	<b>26</b>

# 1 Data Handling and Preprocessing

Al fine di garantire una corretta analisi ed un'opportuna applicazione dei metodi statistici noti, è anzitutto necessaria una valutazione esplorativa del dataset, attraverso la quale si possa controllare l'omogeneità e la consistenza delle informazioni che esso rappresenta, l'assenza di valori NA (o quantomeno una presenza non sensibilmente invalidante), le dimensioni ed altre eventuali discrepanze.

Il preprocessing ha come goal la pulizia del dataset e, nella fattispecie in esame, risulta poco complesso (vedi ". /src/dataPreprocessing.R").

Il dataset assegnato, da ora in poi referenziato mediante il nome dataRaw, ad indicare l'insieme di dati non ancora processati, ha tutte le variabili dichiarate in "assignment/VariabiliProgStatAppl\_DEF\_24\_25", di cui una di risposta e sei potenziali regressori.

```
[1] "y_VideoQuality" "x1_ISO"           "x2_FRatio"      "x3_TIME"       "x4_MP"  
[6] "x5_CROP"        "x6_FOCAL"        "x7_PixDensity"
```

Figure 1: Output di names(dataRaw)

Ognuna delle colonne di dataRaw è consistente, senza NA, con nomi vicendevolmente differenti e sempre significativi(come mostrato in Figure 1.).

A chiudere il preprocessing, la funzione str(dataRaw) esplicita i domini di ogni variabile, confermando non solo che nessuna di esse è categoriale (nessuna variabile binaria) ma che sono anche tutte di tipo numeric e quindi adatte al modeling mediante regressione lineare. I dati forniti non sembrano necessitare di procedimenti di cleaning estensivi o ulteriore handling: si procede alla descrizione delle singole variabili.

# 2 Data Exploration

Metodi di statistica descrittiva nelle fasi preliminari di esplorazione dei dati, sono utili per la caratterizzazione delle variabili e alla definizione della loro distribuzione.

Il file ". /src/descriptiveAnalysis.R" si occupa proprio dello studio dei dati al fine di ricavare quanto necessario per osservare i pattern che li legano e dedurre proprietà dalle loro funzioni di distribuzione, garantendo così una comprensione generale dell'informazione che esse contengono e fornendo le basi per la valutazione delle correlazioni e della collinearità.

## 2.1 Data Structure and Outliers Evaluation

Un primo approccio a tali informazioni è sicuramente il summary del dataset, il cui output, oltre che essere disponibile nella directory dei risultati sotto il nome ". /results/CharacterizedDataSet.txt", è di seguito riportato.

Summary of the non-processed Data Set:

y_VideoQuality	x1_ISO	x2_FRatio	x3_TIME	x4_MP	x5_CROP
Min.    :-16.00	Min.    :-1.72451	Min.    :-1.67486	Min.    :-1.6525	Min.    :-1.64276	Min.    :-1.70328
1st Qu.: 40.76	1st Qu.:-0.80196	1st Qu.:-0.73999	1st Qu.:-1.0902	1st Qu.:-0.95462	1st Qu.:-0.86830
Median : 56.75	Median :-0.03037	Median : 0.11254	Median :-0.1652	Median :-0.07534	Median :-0.05079
Mean    : 54.66	Mean    :-0.05141	Mean    : 0.04285	Mean    :-0.1147	Mean    : 0.01163	Mean    : 0.03188
3rd Qu.: 66.84	3rd Qu.: 0.65099	3rd Qu.: 0.76639	3rd Qu.: 0.7661	3rd Qu.: 0.96147	3rd Qu.: 0.99702
Max.    :110.74	Max.    : 1.71675	Max.    : 1.72301	Max.    : 1.6369	Max.    : 1.71967	Max.    : 1.69710
x6_FOCAL                         x7_PixDensity					
Min.    :-1.72865	Min.    :-1.92564				
1st Qu.:-0.83963	1st Qu.:-0.71872				
Median :-0.01376	Median : 0.01969				
Mean    : 0.04574	Mean    : 0.00000				
3rd Qu.: 0.97996	3rd Qu.: 0.71775				
Max.    : 1.73089	Max.    : 2.45975				

Dimension:

100

8

Oltre a verificare nuovamente la correttezza dimensionale di dataRaw, riportata come Dimension, è possibile vedere come i valori delle singole osservazioni della response variable abbiano un range di variabilità significativamente più ampio rispetto ai predittori. La discrepanza poi che tale variabile presenta tra il valore Median e il valore Mean sembra indicare una skewness marcata nella distribuzione e potrebbe essere sintomo della presenza di eventuali outliers. Per quanto riguarda le indipendenti, anche in questo caso la variabilità che x7\_PixDensity e x5\_CROP mostrano tra i propri Max. e le rispettive medie, potrebbe indicare la presenza di outliers. Al fine di scongiurare tale ipotesi si osserva dapprima il box-plot complessivo delle variabili e poi i box-plots di ognuna di esse, disponibili nella directory "./results/boxplots/", di seguito riportati.

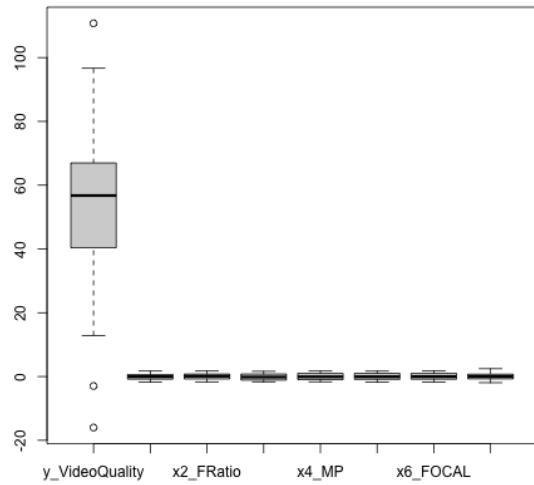


Figure 2: Box-plot di tutte le variabili. Si notano outliers sulla response variable.

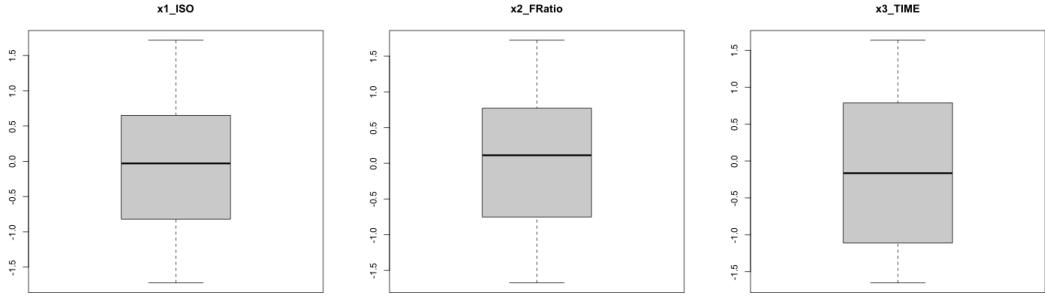


Figure 3: x1\_ISO

Figure 4: x2\_FRatio

Figure 5: x3\_TIME

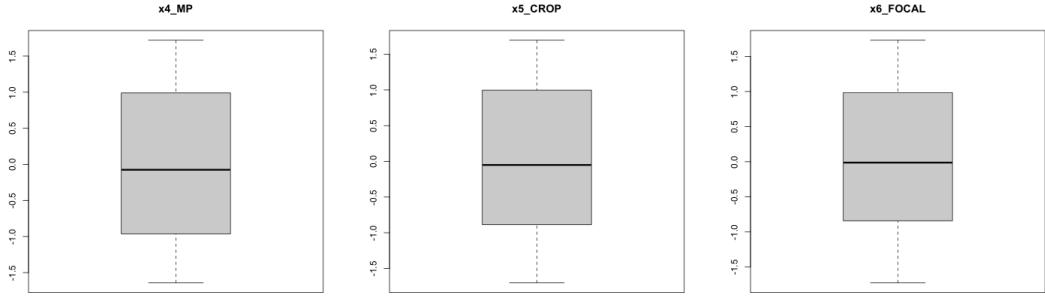


Figure 6: x4\_MP

Figure 7: x5\_CROP

Figure 8: x6\_FOCAL

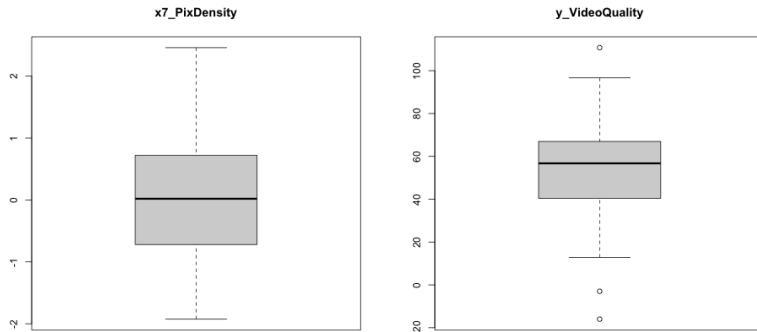


Figure 9: x7\_PixDensity

Figure 10: y\_VideoQuality

Un'analisi grafica dei box-plots (da Figure 3 a Figure 9) permette di osservare l'assenza di outliers nelle variabili indipendenti. Il boxplot della variabile di risposta (Figure 10) mostra gli outliers già identificati in Figure 2, al di fuori dei whiskers del boxplot, definiti come 1.5 volte l'IQR (Interquartile Range). È bene però considerare che tali valori, seppur possano rappresentare un inconveniente, la loro polarizzazione comunque non è estrema. Si ipotizza quindi, solo per il momento e per semplificazione di trattazione, che gli outliers rilevati non abbiano effetti sensibili sui coefficienti del modello di regressione, sull'MSE o sulla normalità dei residuals. L'ipotesi verrà poi discussa in fase di proposta dei modelli di regressione (vedi Sez. Model Comparison and Validation) a valutazione della distribuzione dei residui.

## 2.2 Variables Distribution

La caratterizzazione delle variabili di dataRaw procede con la discussione della distribuzione di ognuna di esse, costruita chiaramente mediante istogrammi, e di seguito riportata.

Si ricorda che tutte le distribuzioni sono accessibili e contenute nella directory

"./results/histograms/".

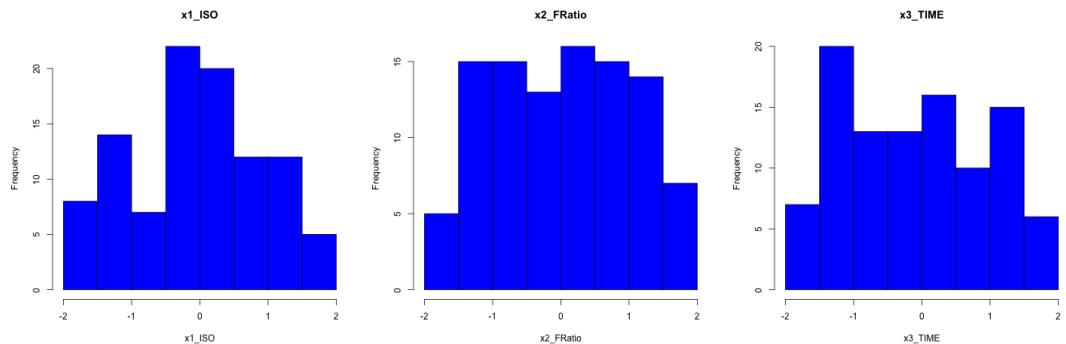


Figure 11: x1\_ISO

Figure 12: x2\_FRatio

Figure 13: x3\_TIME

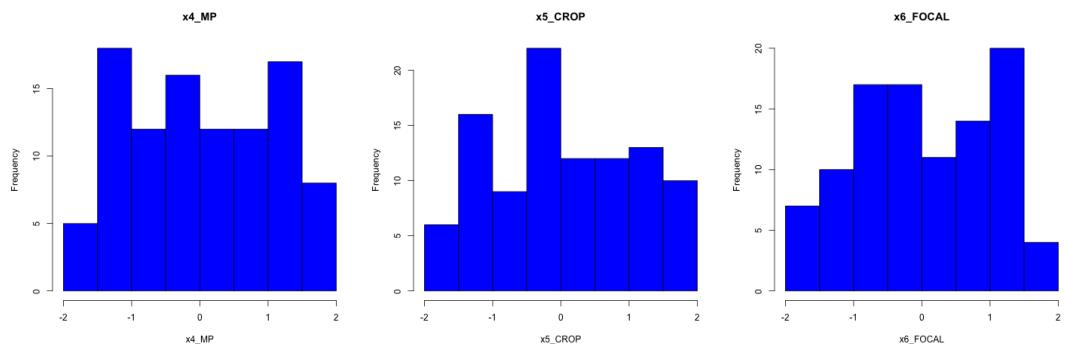


Figure 14: x4\_MP

Figure 15: x5\_CROP

Figure 16: x6\_FOCAL

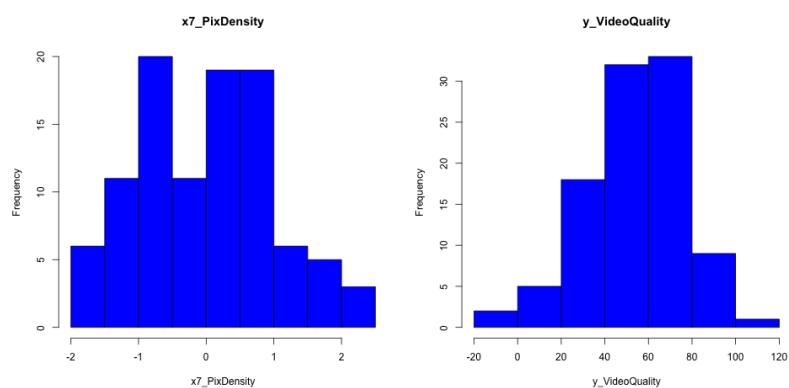


Figure 17: x7\_PixDensity

Figure 18: y\_VideoQuality

Seppur un'analisi grafica degli output suggerisca con forza che, a meno di quella di risposta, tutte le variabili del dataset non siano distribuite come una Normale, si effettua comunque il test di Shapiro-Wilk su ognuna. Ci si aspetta di rifiutare la null Hypothesis ( $H_0$  : I dati provengono da una distribuzione Normale), per tutte le variabili indipendenti, e di accettarla solo per y\_VideoQuality.

Seguono i risultati dei test, reperibili nel file "./results/Shapiro-Wilk.txt".

```
Shapiro-Wilk test for each dataRaw variable:

Shapiro-Wilk normality test
data: dataRaw$y_VideoQuality
W = 0.9849, p-value = 0.3124
p-value > alpha=0.05, y_VideoQuality is distributed as a Normal random variable

Shapiro-Wilk normality test
data: dataRaw$x1_ISO
W = 0.96069, p-value = 0.004508
p-value < alpha=0.05, x1_ISO is not distributed as a Normal random variable

Shapiro-Wilk normality test
data: dataRaw$x2_FRatio
W = 0.9575, p-value = 0.002678
p-value < alpha=0.05, x2_FRatio is not distributed as a Normal random variable

Shapiro-Wilk normality test
data: dataRaw$x3_TIME
W = 0.93515, p-value = 9.871e-05
p-value < alpha=0.05, x3_TIME is not distributed as a Normal random variable

Shapiro-Wilk normality test
data: dataRaw$x4_MP
W = 0.93457, p-value = 9.127e-05
p-value < alpha=0.05, x4_MP is not distributed as a Normal random variable

Shapiro-Wilk normality test
data: dataRaw$x5_CROP
W = 0.94712, p-value = 0.0005386
p-value < alpha=0.05, x5_CROP is not distributed as a Normal random variable

Shapiro-Wilk normality test
data: dataRaw$x6_FOCAL
W = 0.94806, p-value = 0.0006191
p-value < alpha=0.05, x6_FOCAL is not distributed as a Normal random variable

Shapiro-Wilk normality test
data: dataRaw$x7_PixDensity
W = 0.98484, p-value = 0.3091
p-value < alpha=0.05, x7_PixDensity is not distributed as a Normal random variable
```

Si nota che per la variabile dipendente y\_VideoQuality:

$$p\text{-value} = 0.3124 > \alpha = 0.05,$$

e pertanto si accetta l'ipotesi nulla  $H_0$ :

$H_0$  : I dati provengono da una distribuzione Normale.

Se la skewness emersa dal box-plot non sembra quindi avere grandi implicazioni sulla distribuzione della variabile, non è detto che si possa affermare lo stesso per gli outliers. Appurata la normalità in distribuzione, si valuta, per completezza, anche il Q-Q Plot (vedi Figure 19), anch'esso riportato nella directory ".*results*", al fine di caratterizzare ulteriormente il comportamento dei valori polarizzati osservati.

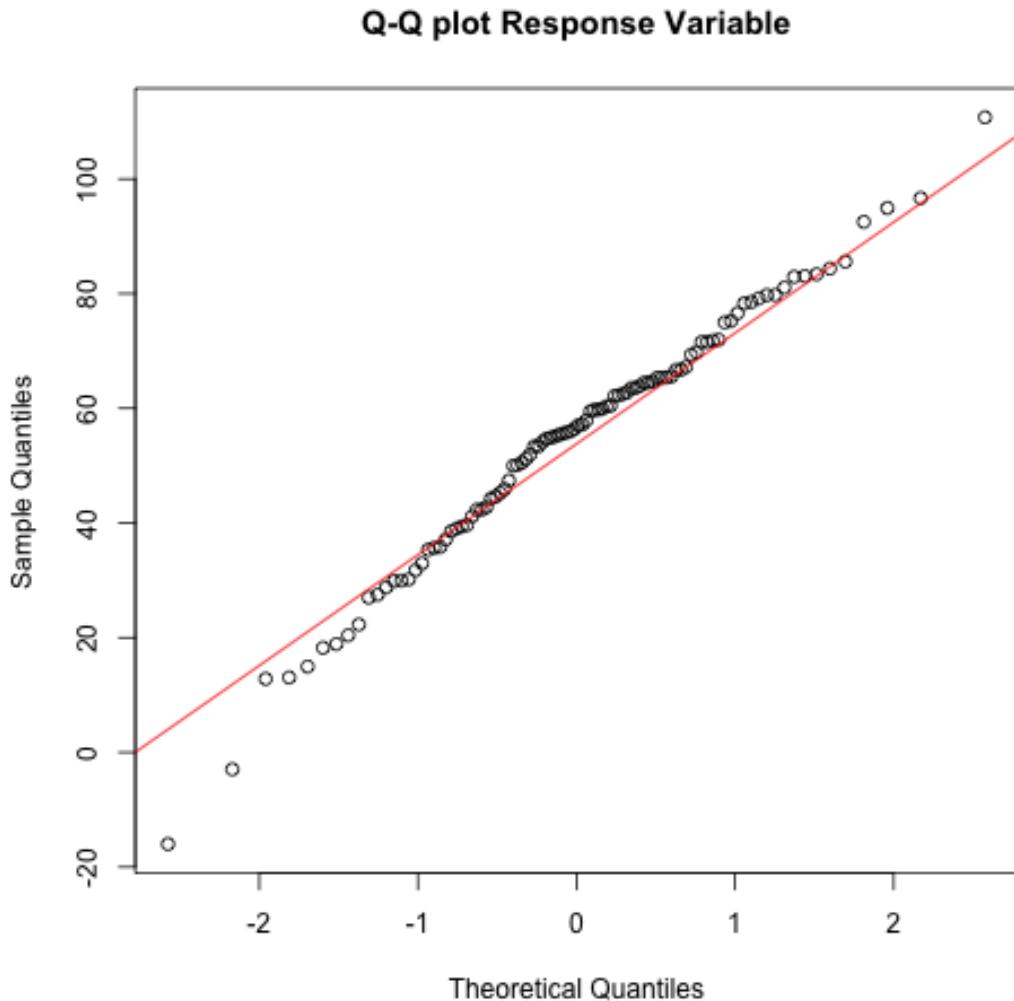


Figure 19: Q-Q plot per *y\_VideoQuality*. Sono sensibili gli effetti degli outliers.

Il buon allineamento dei punti lungo la diagonale conferma quanto emerso dal test di Shapiro–Wilk sulla variabile di risposta, mentre la lieve dispersione in corrispondenza delle code suggerisce la possibile influenza, potenzialmente non trascurabile, dei valori polarizzati. La verifica della normalità di *y\_VideoQuality* rappresenta un prerequisito importante per la fase di modellazione, poiché consente di adottare, con ragionevole approssimazione, l'assunzione di residui multivariati gaussiani nella regressione lineare multipla.

Le conclusioni definitive sull'impatto delle discrepanze emerse, tuttavia, saranno formulate nella sezione di diagnostica del modello, con particolare riferimento all'analisi dei residui.

Per quanto riguarda poi tutte le altre variabili, il *p*-value associato al test risulta sempre minore del livello di rischio, permettendo quindi di rifiutare l'ipotesi  $H_0$ .

### 3 Interaction Analysis: Investigating Correlations

Al fine di salvaguardare lo sviluppo dei modelli di regressione è necessario strumentalizzare quanto ricavato dalla data exploration per studiare le relazioni funzionali che legano le singole variabili di `dataRaw`. È dunque responsabilità della fase di Variables Interaction Analysis la quantificazione degli indici di correlazione statisticamente significativi, oltre che la produzione di un nuovo training set qualora detti indici dovessero invalidare l'integrità dei modelli a causa di collinearità rilevanti.

#### 3.1 Visualization and Statistical Hypothesis Testing

Una prima valutazione grafica dei coefficienti di correlazione tra le variabili è fornita dall'heatmap, da ritrovarsi nel file `./results/heatmaps/heatmap_raw_dataset.pdf`, riportata di seguito.

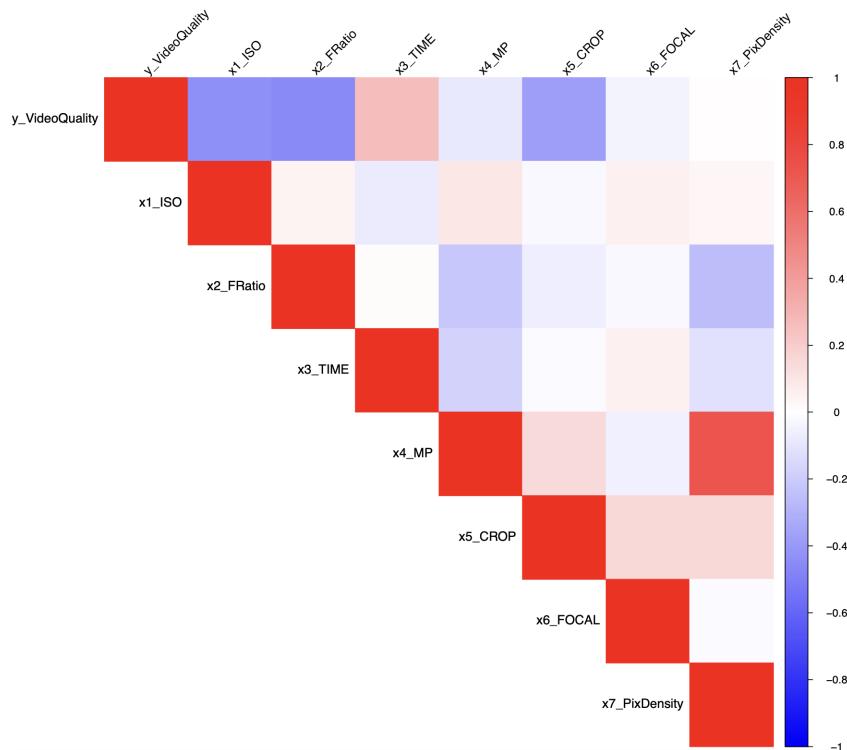


Figure 20: Heatmap per `dataRaw`.

È possibile constatare subito una forte correlazione positiva tra `x4_MP` e `x7_PixDensity` ed una moderata correlazione negativa tra la variabile di risposta e le variabili `x1_ISO`, `x2_FRatio` e `x5_CROP`. Per conferire una maggiore autorevolezza alle osservazioni è però necessario quantificare tali correlazioni mediante la computazione della matrice delle correlazioni, sui cui elementi verrà quindi condotto un test d'ipotesi volto all'apprezzamento della loro significatività statistica. In Figure 21 è dunque riportato l'output della funzione `GGally::ggpairs(dataRaw)` che, quando eseguita nel FrameWork, restituisce lo scatter plot delle variabili di `dataRaw` con gli associati coefficienti di correlazione e *p*-values. Si fa presente che il codice R implementato nel file sorgente `./src/regressionAnalysis.R` produce come output direttamente nel FrameWork la matrice di correlazione e salva nella directory `./results/` la matrice dei *p*-values; il file `./results/scatterplots/scatter-plot_All.png` propone invece una versione alternativa e snellita dello scatter di tutte le variabili.

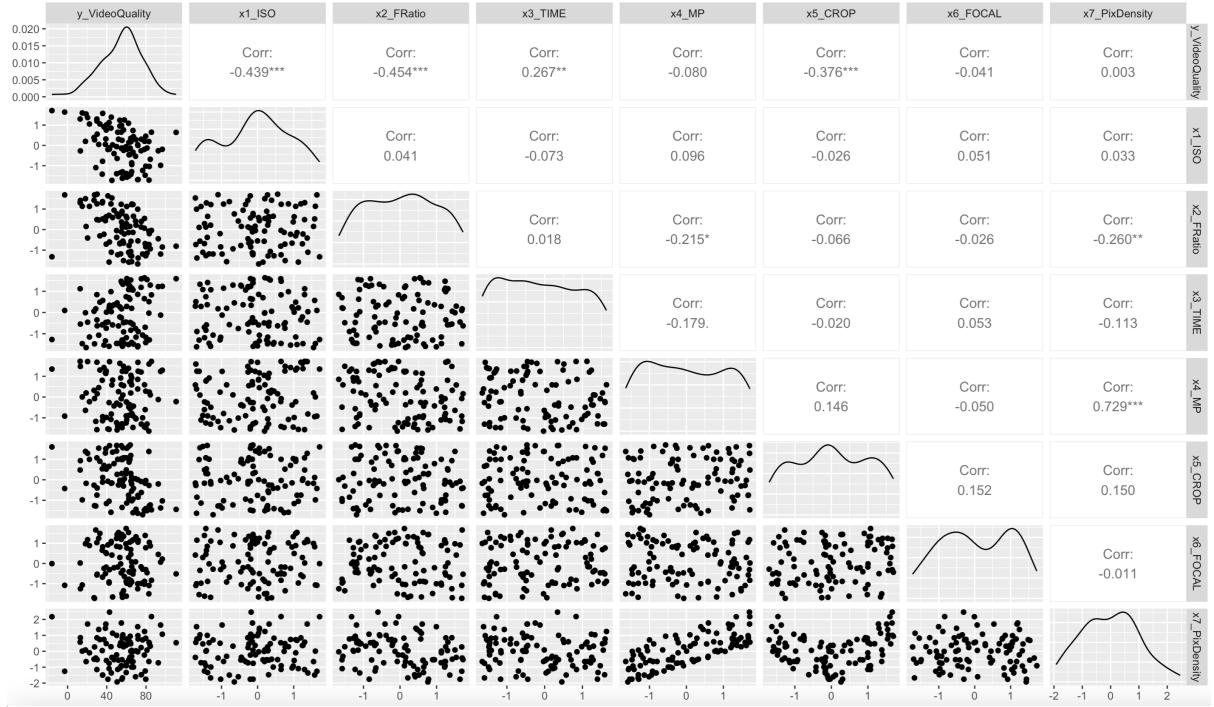


Figure 21: Scatter plot completo di coefficienti di correlazioni e distribuzione delle osservazioni.

Osservando i risultati mostrati in Figure 21, è possibile affermare che quanto dedotto dalla heatmap in Figure 20 è corretto: vi è un'evidente e preoccupante collinearità tra le variabili  $x4\_MP$  e  $x7\_PixDensity$ .

Su ogni coppia di variabili, sia condotto il test d'ipotesi tale che:

$$H_0 : R = 0 \quad H_A : R \neq 0,$$

allora, nella fattispecie di  $x4\_MP$  e  $x7\_PixDensity$ , risulta:

$$R = 0.729,$$

con un  $p$ -value:

$$p\text{-value} = 0.00 < 0.05 \Rightarrow H_0 \text{ rifiutata.}$$

Emergono poi come significative anche le seguenti correlazioni:

- $y\text{-VideoQuality}$  e  $x1\text{\_ISO}$  con un indice  $R = -0.439$  ed alta significatività statistica ( $p\text{-value} \approx 4.88 \cdot 10^{-6} \Rightarrow H_0$  rifiutata)
- $y\text{-VideoQuality}$  e  $x2\text{\_FRatio}$  con un indice  $R = -0.454$  ed alta significatività statistica ( $p\text{-value} \approx 2.13 \cdot 10^{-6} \Rightarrow H_0$  rifiutata)
- $y\text{-VideoQuality}$  e  $x3\text{\_TIME}$  con un indice  $R = 0.267$  ed alta significatività statistica ( $p\text{-value} \approx 0.0073 \Rightarrow H_0$  rifiutata)
- $y\text{-VideoQuality}$  e  $x5\text{\_CROP}$  con un indice  $R = -0.376$  ed alta significatività statistica ( $p\text{-value} \approx 0.0001 \Rightarrow H_0$  rifiutata)
- $x2\text{\_FRatio}$  e  $x7\text{\_PixDensity}$  con un indice  $R = -0.260$  ed alta significatività statistica ( $p\text{-value} \approx 0.0090 \Rightarrow H_0$  rifiutata)

- $x4\_MP$  e  $x2\_FRatio$  con un indice  $R = -0.215$  e moderata significatività statistica ( $p\text{-value} \approx 0.0031 \Rightarrow H_0$  rifiutata )

Si nota che il grado di correlazione tra ciascuna variabile indipendente e la variabile di risposta  $y\text{-VideoQuality}$  fornisce un'indicazione del peso relativo di quel preditore all'interno del modello di regressione, per cui valutare gli indici sarebbe al momento poco utile.

La relazione problematica è infatti quella tra  $x4\_MP$  e  $x7\_PixDensity$ , la quale, se ignorata, potrebbe impedire alla matrice di covarianza del modello di regressione di rispettare la proprietà di invertibilità, opponendosi quindi al processo di modeling. Per questa ragione si produce un nuovo dataset, chiamato "`./data/DataSet_gruppo4-PROCESSED.csv`", nuovo dataset di allenamento, che sia manchevole delle variabili ritenute inadatte alla modellazione tramite regressione.

Si elimina pertanto dal training set la variabile  $x7\_PixDensity$ . Per quanto riguarda invece le ultime due correlazioni elencate, il valore ridotto di  $R$  che esse presentano non crea de facto un problema di collinearità sensibile, ragion per cui non vi è necessità di eliminarne alcuna.

Per completezza è possibile consultare l'heatmap del nuovo dataset, contenuta nel file "`./results/heatmaps/heatmap_processed_dataset.pdf`".

### 3.2 Polynomial Regression Analysis for Capturing Non-Linear Effects

Al fine di esplorare la possibilità che alcune relazioni tra le variabili indipendenti e la risposta possano non essere di natura strettamente lineare, si è condotta un'analisi tramite regressione polinomiale. Tale approccio consente di modellare una dipendenza non lineare trasformandola in una relazione lineare nei parametri, mantenendo così la compatibilità con la struttura della regressione lineare classica.

È da porre particolare attenzione alla costruzione dei modelli polinomiali tra  $y\text{-VideoQuality}$  e i regressori  $x1\_ISO$ ,  $x2\_FRatio$ ,  $x3\_TIME$  e  $x5\_CROP$  così da caratterizzare ulteriormente le correlazioni emerse nella sezione precedente.

Sia definito un modello target polinomiale del tipo:

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

allora, si costruiscono i seguenti modelli aventi grado ottimale individuato empiricamente.

$$\begin{aligned} x1\_ISO : \quad y\text{-VideoQuality}_x &= \hat{\beta}_0 + \hat{\beta}_1 x1\_ISO + \hat{\beta}_2 x1\_ISO^2, \\ x2\_FRatio : \quad y\text{-VideoQuality}_x &= \hat{\beta}_0 + \hat{\beta}_1 x2\_FRatio + \hat{\beta}_2 x2\_FRatio^2, \\ x3\_TIME : \quad y\text{-VideoQuality}_x &= \hat{\beta}_0 + \hat{\beta}_1 x3\_TIME + \hat{\beta}_2 x3\_TIME^2, \\ x5\_CROP : \quad y\text{-VideoQuality}_x &= \hat{\beta}_0 + \hat{\beta}_1 x5\_CROP. \end{aligned}$$

Ulteriori approfondimenti e dettagli in merito allo sviluppo dei modelli di regressione polinomiale sulle coppie di predittori, oltre che dettagliate descrizioni in merito ai funzionali sopra discussi, sono da ritrovarsi nel file "`./results/Polynomial_Regression.txt`"; al suo interno è infatti riportato il sommario di ognuno di essi, completo di stime sui parametri,  $p\text{-value}$  dei  $t$ -Test su ogni parametro, sommario sui residui, deviazione standard dei residui, coefficiente di determinazione  $R^2$  e statistica  $F$  per la quantificazione della variabilità spiegata dal modello. Alla luce di quanto emerso, è allora possibile confermare con autorevolezza la bontà dei predittori  $x1\_ISO$ ,  $x2\_FRatio$ ,  $x3\_TIME$  e  $x5\_CROP$  nel contribuire alla spiegazione della qualità video percepita. Queste evidenze polinomiali confermano l'importanza di includere termini non lineari in alcuni predittori, ragion per cui si terrà in considerazione l'influenza dei funzionali sopra ricavati nella formulazione dei modelli di regressione, presentati nella sezione successiva.

## 4 Defining Regression Models

In seguito alle evidenze emerse in data exploration e interaction analysis, è bene concentrare l'attenzione sulla costruzione dei modelli di regressione veri e propri.

Lo script `./src/regressionModel.R` formalizza a tal proposito diverse ipotesi modellistiche - dal modello lineare multiplo completo a configurazioni multiple e polinomiali. Questa fase permette quindi di evidenziare eventuali criticità – ad esempio anomalie nelle dipendenze collineari – che potrebbero dover essere oggetto di un'analisi diagnostica più approfondita. Il goal è quindi la selezione del miglior modello che possa meglio spiegare la variabilità delle osservazioni di `y_VideoQuality`, concentrandosi sulla goodness of fit, nella speranza di avere accuratezza predittiva, e nel rispetto del rasoio di Occam (parsimony principle).

### 4.1 Linear and Polynomial Approaches

Si consideri il dataset processato di cui precedentemente discusso, d'ora in poi referenziato come `processedData`, allora è possibile costruire un modello di regressione multipla nella forma:

$$Y = f(X_1, \dots, X_p) + \varepsilon,$$

che, per  $f(X_1, \dots, X_p)$  lineare,  $Y = y_{\text{VideoQuality}}$  e  $p = 6$ , è tale che:

$$\begin{aligned} y_{\text{VideoQuality}} = & \beta_0 + \beta_1 x_{1,\text{ISO}} + \beta_2 x_{2,\text{FRatio}} + \\ & + \beta_3 x_{3,\text{TIME}} + \beta_4 x_{4,\text{MP}} + \beta_5 x_{5,\text{CROP}} + \beta_6 x_{6,\text{FOCAL}} + \varepsilon. \end{aligned}$$

Questi è in effetti il modello principale in analisi (d'ora in poi chiamato *Complete Model* poiché rappresenta il modello massimo), ma non l'unico. Vi sono infatti due modelli alternativi, proposti in funzione della sensibile influenza che alcuni regressori hanno dimostrato sulla variabile dipendente ma anche e soprattutto in virtù degli andamenti polinomiali ricavati e caratterizzati precedentemente.

Siano considerate le variabili `x1_ISO`, `x2_FRatio`, `x3_TIME` e `x5_CROP`, i cui scatter plot, oltre che disponibili nella directory `./results/scatterplots`, sono riportati in Figure 22.

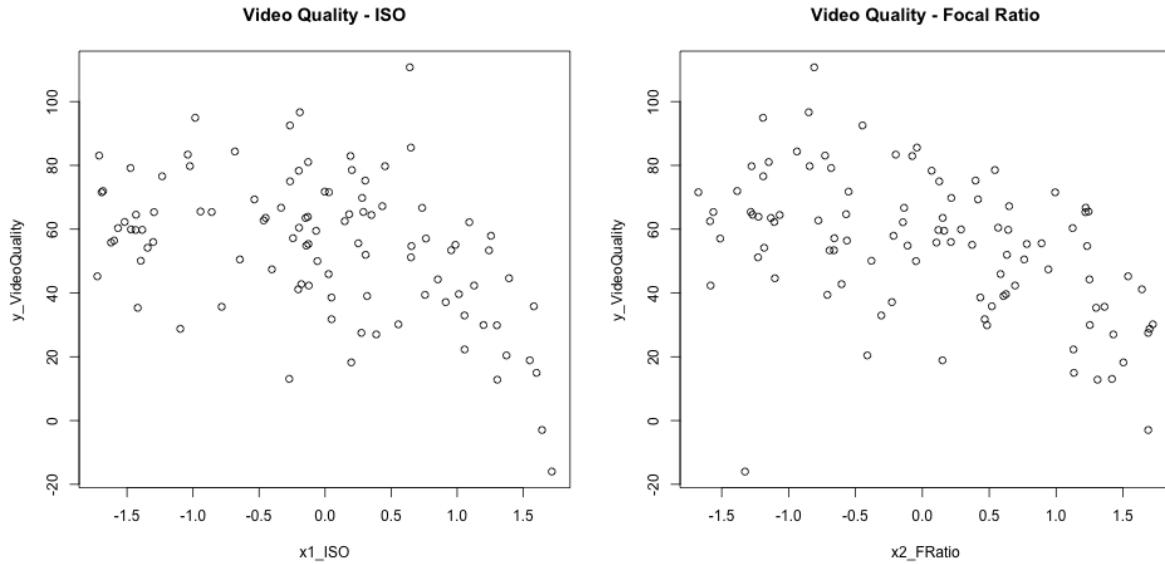
Allora è possibile definire il primo modello alternativo (detto *Alternative Model 1*) come un modello multiplo ridotto, tale che:

$$y_{\text{VideoQuality}} = \beta_0 + \beta_1 x_{1,\text{ISO}} + \beta_2 x_{2,\text{FRatio}} + \beta_3 x_{3,\text{TIME}} + \beta_4 x_{5,\text{CROP}} + \varepsilon.$$

Vi è comunque la possibilità che tale modello alternativo non sia in grado di spiegare in maniera sufficientemente precisa la variabilità su `y_VideoQuality` poiché non tiene conto dei termini quadratici che invece caratterizzano a fondo le interazioni proprio tra questi regressori e la variabile di risposta.

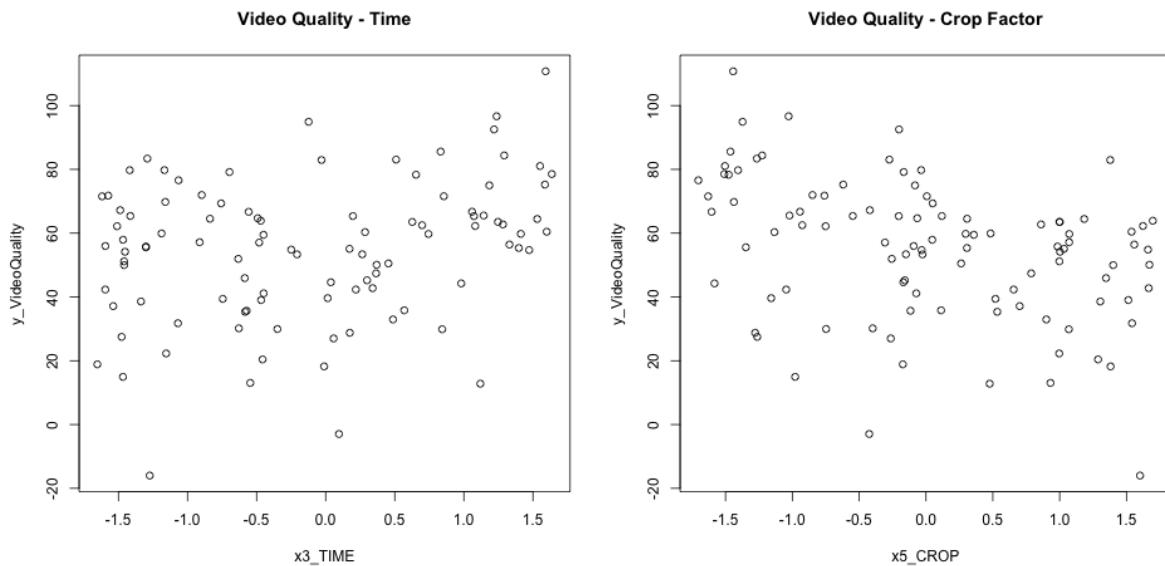
Si formula pertanto un secondo modello alternativo (detto *Alternative Model 2*) che metta in evidenza le regressioni polinomiali, discusse in sezione 3.2, tale che:

$$\begin{aligned} y_{\text{VideoQuality}} = & \beta_0 + \beta_1 x_{1,\text{ISO}} + \beta_2 (x_{1,\text{ISO}})^2 + \\ & + \beta_3 x_{2,\text{FRatio}} + \beta_4 (x_{2,\text{FRatio}})^2 + \beta_5 x_{3,\text{TIME}} + \beta_6 (x_{3,\text{TIME}})^2 + \beta_7 x_{5,\text{CROP}} + \varepsilon. \end{aligned}$$



Andamento quadratico concavo-downward:  
la qualità cresce fino a un ISO  
moderato e poi decresce rapidamente.

Relazione quadratica marcata, concava-downward: i valori estremi di F-Ratio associano diminuzioni nette di qualità.



Andamento quadratico concavo-upward:  
la qualità migliora per tempi  
molto brevi o molto lunghi.

Chiara correlazione lineare decrescente:  
la qualità diminuisce  
all'aumentare del crop factor.

Figure 22: Scatter Plot dei regressori critici.

### About Quadratic Terms

La formulazione di modelli alternativi ridotti, cioè quei modelli di regressione multipla aventi meno variabili di quelle previste dal dataset, come ad esempio *Alternative Model 1*, non porta

con sé particolari problemi, a meno di una valutazione delle variazioni in precisione e della significatività statistica dei parametri associati. Questione diversa è la formulazione di modelli aventi termini polinomiali, come *Alternative Model 2*, nei quali non si conosce nulla in merito ad eventuali collinearità generate dai regressori al quadrato. A tal proposito, una volta generato un dataset temporaneo comprensivo delle componenti quadratiche, chiamato `dataModel12`, si costruisce la rinnovata matrice di correlazione con associati *p*-value e si analizzano le correlazioni tramite scatterplot.

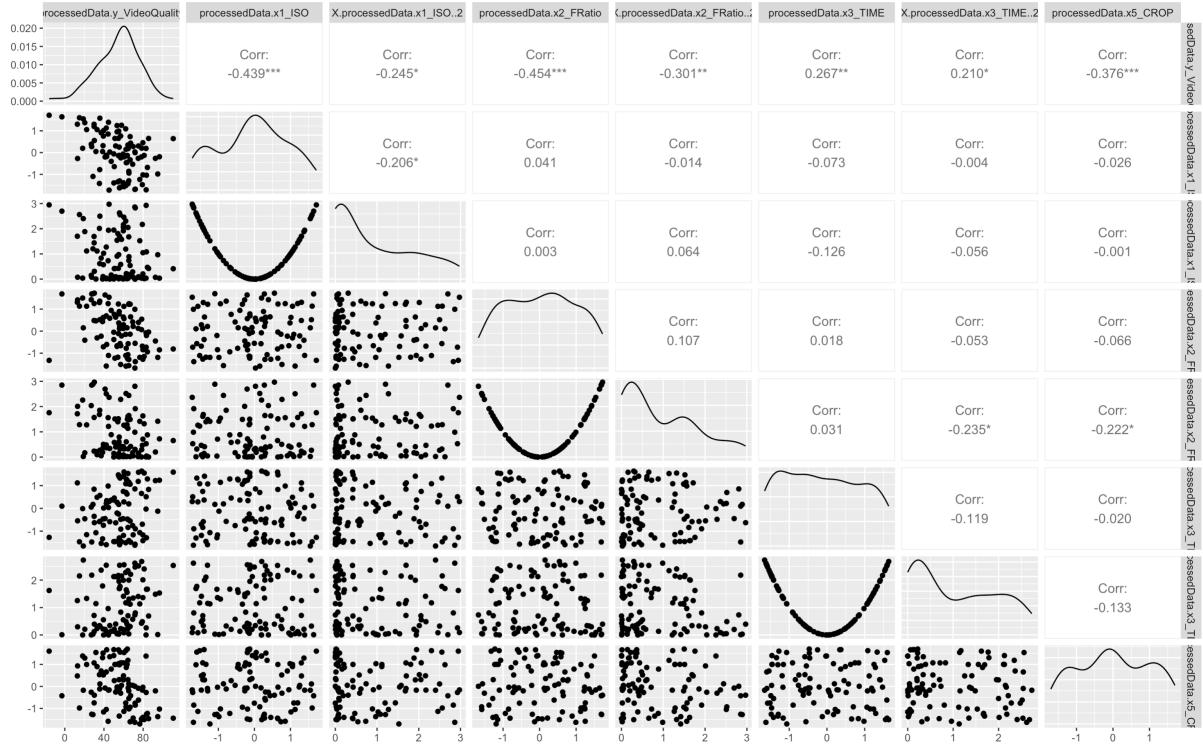


Figure 23: Scatter plot di valutazione delle collinearità/multicollinearità tra i regressori per *Alternative Model 2*.

A meno degli ovvi andamenti quadratici che si possono osservare sulle colonne relative ai regressori ricavati come square di quelli originali, lo scatter non evidenzia alcuna collinearità particolarmente influente o preoccupante. Quanto emerge invece sulla colonna delle variabili  $(x1\_ISO)^2, (x2\_FRatio)^2$  e  $(x3\_TIME)^2$ , ovvero la concentrazione di osservazioni sullo zero, è questione legata ai campioni, già centrati attorno allo zero. L'analisi può pertanto procedere con la caratterizzazione dei modelli candidati.

## 4.2 A Formal and Graphical Diagnostic

L'analisi dei modelli, implementata nei file `./src/regressionModel.R` e `./src/decision.R`, si basa sulla valutazione dei *p*-value ottenuti dai test t sui parametri stimati, sulla misurazione dei guadagni in precisione di stima, sulla quantificazione della variabilità spiegata e sull'analisi della distribuzione dei residui. È infatti necessario, per via degli outliers che la variabile di risposta ha mostrato in data exploration, giudicare il comportamento e quotare la deviazione standard (MSE) dei residuals, al fine di verificare la ridotta influenza dei valori polarizzati e confermare la scelta di mantenere inalterato il campione originale.

Si ricorda che, quanto segue (Figure 24-26) è riportato direttamente nei files ".*/results/models/NomeModello.pdf*" e nel file ".*/results/Multiple\_Regression.txt*", per i summary di ogni model.

#### 4.2.1 Complete Model

Il modello principale, come riportato nel file testuale sopra referenziato, e come si può evincere osservando il primo dei due plot in Figure 24, offre una moderata qualità nella stima della media della variabile indipendente. Con un coefficiente di determinazione  $R^2$ :

$$R^2 = \frac{SQTOT - SQE_6}{SQTOT} \approx 0.614,$$

al sesto regressore, il modello è in grado di spiegare un'aliquota pari a circa il 60% della variabilità totale. Il vero problema del modello è però da ritrovarsi nella significatività dei parametri associati alle variabili *x4\_MP* e *x6\_FOCAL*. Le due variabili, già precedentemente identificate come non sensibilmente correlate ad *y\_VideoQuality*, hanno parametri stimati il cui *p*-value al test della *t* di Student è rispettivamente pari a 0.5418 e 0.8096, indicando quindi una innecessaria complessità del modello.

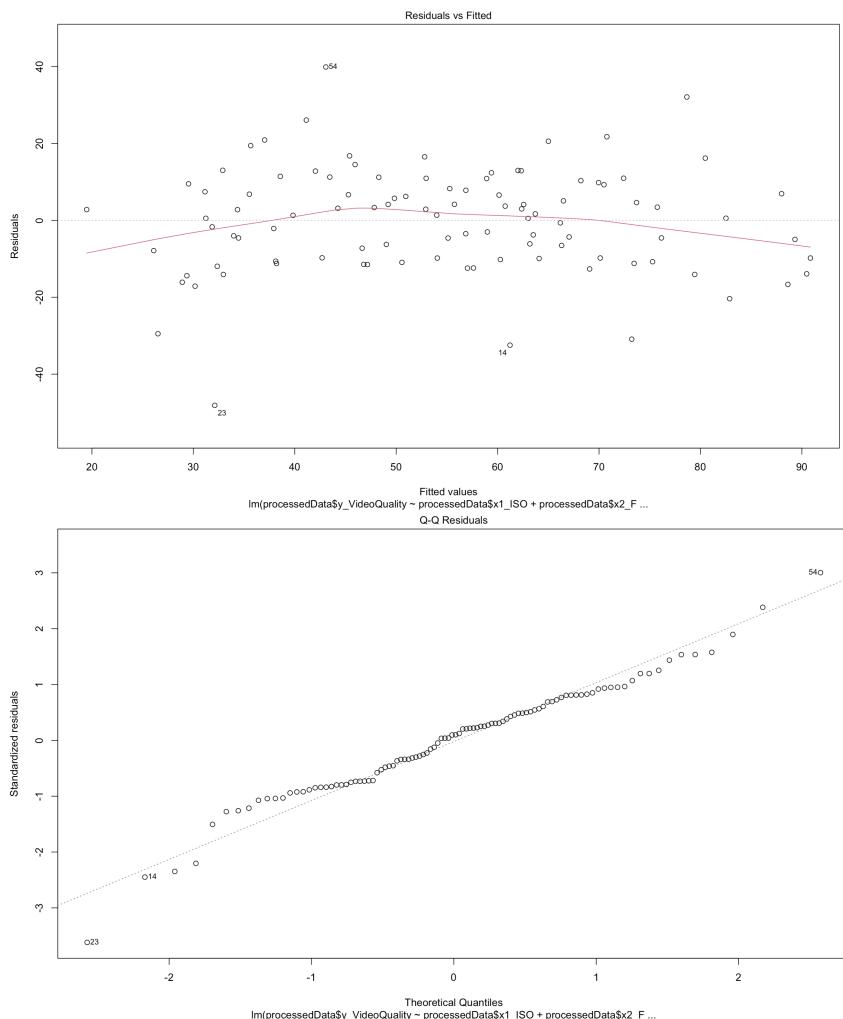


Figure 24: Residual vs Fitted e Q-Q Plot dei residui di *Complete Model*.

Il Residual Standard Error poi risulta pari a 14.02 - che vedremo essere il più alto tra tutti i

candidati - indicando pertanto un apprezzabile discostamento tra il modello e i dati osservati. Nonostante, secondo quanto suggerisce il Q-Q Plot, gli outliers della v.a. indipendente non sembrano avere grandi impatti sui modelli - si osservino le code -, probabilmente in virtù della loro bassa frequenza (circa l'1% dei dati sono outliers).

#### 4.2.2 Alternative Model 1

Nell'*Alternative Model 1* la semplificazione sui predittori, che secondo quanto emerso da una prima analisi del *Complete Model* è corretta, comporta una riduzione dell'errore standard :

$$RSE = 13.91 \Rightarrow \text{Riduzione dello } 0,79\%,$$

e un mantenimento sostanziale della capacità esplicativa del modello ( $R^2_{\%} \approx 61\%$ ).

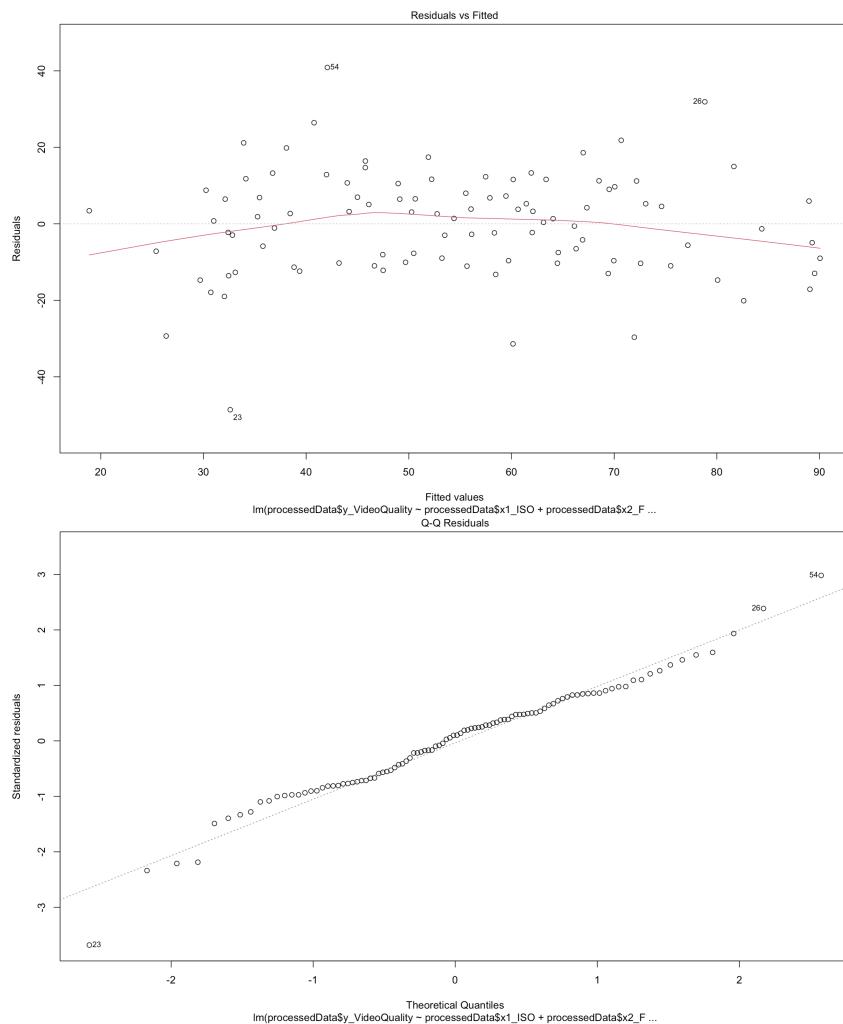


Figure 25: Residual vs Fitted e Q-Q Plot dei residui di *Alternative Model 1*.

Tutti i coefficienti risultano poi altamente significativi, lasciando circa invariato l'ordine di grandezza dei  $p$ -value per ogni parametro ( $p \ll 0.001$ ), a conferma della solidità del sottoinsieme selezionato. Per quanto concerne la distribuzione dei residui, anche per *Alternative Model 1* non appaiono dal Q-Q plot (Figure 25) particolari discostamenti dai quantili della Normale. Chiaramente la maggiore compattezza del modello, a fronte di prestazioni pressoché

invariate, rappresenta un chiaro vantaggio in termini di parsimonia e interpretabilità, eppure la riduzione minima osservata sull'RSE suggerisce, nonostante si abbiano meno parametri, come il modello tenda a performare meglio. A tal proposito ed in questa misura si hanno ottime speranze nei riguardi del secondo modello alternativo.

#### 4.2.3 Alternative Model 2

Si conclude l'esame dei candidati con l'esplorazione di *Alternative Model 2*, i cui grafici diagnostici sono riportati di seguito.

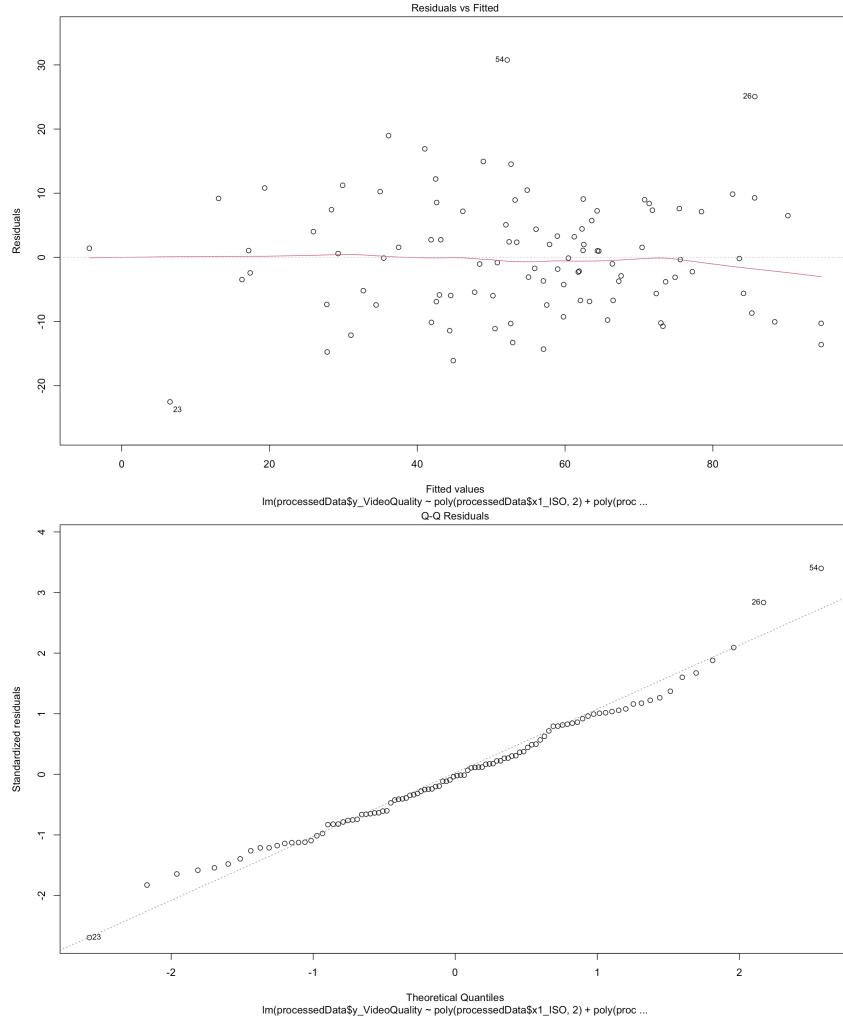


Figure 26: Residual vs Fitted e Q-Q Plot dei residui di *Alternative Model 2*.

Detto modello, arricchito dell'esplicitazione delle relazioni polinomiali caratteristiche dei campioni proposti, per i predittori principali, si dimostra quello in grado di catturare con maggior precisione la complessità intrinseca del rapporto tra variabili di ripresa e qualità percepita. L' $R^2_{adj}$  raggiunge 0.818 e l' $R^2$  cresce fino a 0.831, evidenziando un guadagno di oltre il 35% nella quota di variabilità spiegata rispetto al modello massimo.

I grafici (Figure 26) avvalorano con forza la scelta: il diagramma Residuals vs Fitted mostra nuvole di punti distribuite in modo omogeneo attorno allo zero, segno di una varianza costante sugli scarti, mentre il Q-Q plot rimanda un perfetto allineamento dei residui rispetto alla diagonale teorica, confermando la plausibilità dell'ipotesi di normalità. L'errore standard dei

residui poi si riduce del 32.96% rispetto ad Alternativ Model 1, raggiungendo un valore  $RSE = 9.33$ , con una conseguente riduzione del gap medio tra valori osservati e predetti. Infine, il  $p$ -value al Test  $t$  conferma l'ingresso dei parametri stimati nel modello, con un'accettazione dell'ipotesi alternativa su tutti, tranne che sul termine di secondo grado di  $x3\_TIME$ . La stima del parametro  $\beta_6$  è infatti pari a:

$$\hat{\beta}_6 = 10.3298,$$

con  $p$ -value:

$$p = 0.296 > \alpha = 0.05.$$

Ciò obbliga l'esclusione del parametro dal modello, suggerendo come l'incremento di complessità con termine quadratico sulla variabile  $x3\_TIME$  non apporta beneficio. L'esclusione del parametro discusso è riservata alla fase di applicazione dei criteri di selezione dei parametri e confronto dei modelli.

Resta comunque evidente che la trasformazione non lineare migliora la capacità esplicativa complessiva del modello, pur richiedendo una valutazione critica sull'effettiva necessità di ciascun termine.

## 5 Model Comparison and Validation

È essenziale, dopo aver esplorato e messo alla prova ciascun modello con le prime analisi diagnostiche, passare alla fase decisiva: il confronto diretto tra i modelli candidati.

L'obiettivo della fase di model comparison and validation è infatti, sotto la guida dello script ".src/decision.R", l'applicazione dei criteri di scelta, discussi nella sezione seguente, l'individuazione del modello che meglio coniughi capacità predittiva, semplicità interpretativa e robustezza statistica.

### 5.1 Comparison Strategies and Evaluative Procedures

Si immagini di studiare un fenomeno  $y$  su un insieme di  $p$  variabili predittive  $\mathbf{X} = (X_1, X_2, \dots, X_p)$  con l'intento di definire un modello matematico di tipo lineare, espresso nella forma:

$$Y = \beta_0 + \sum_{i=1}^p \beta_i X_i + \varepsilon,$$

In tale contesto, l'individuazione del modello ottimale può essere effettuata attraverso appositi criteri di selezione applicati nell'ambito della regressione stepwise, effettuando test d'ipotesi in merito alla media degli stimatori dei coefficienti  $\beta_i$ .

In merito alle scelte del criterio di analisi Stepwise, la teoria statistica contempla principalmente tre approcci: *Forward*, *Backward* e *Hybrid* (o bidirezionale). Tali metodi si applicano al classico modello lineare multiplo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad \forall i \in (1, \dots, n)$$

che, in forma matriciale, è:

$$\underline{y} = \underline{X} \underline{\beta} + \underline{\varepsilon},$$

dove:

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ è il vettore delle osservazioni } (n \times 1),$$

$$\underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \text{ è il vettore dei parametri } ((p+1) \times 1),$$

$$\underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ è il vettore degli errori aleatori } (n \times 1),$$

$$\underline{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \text{ è la design matrix, di dimensione } n \times (p+1).$$

È comunque da tenere in considerazione che, per modelli come *Alternative Model 2*, le trasformazioni sui predittori  $x_{ik} = x_{i2}^2$  possono essere incorporate nella matrice di design per modellare relazioni non lineari in modo parametrico.

Tenendo quindi a mente i tre modelli proposti, è stata adottata la procedura di regressione Stepwise nella sua formulazione backward, così che, partendo dai completi di tutti i predittori potenzialmente rilevanti, e procedendo mediante l'eliminazione iterativa delle variabili che forniscono un contributo predittivo trascurabile, si è in grado di costruire i minimi modelli ottimali.

Si intende con ciò dire che una regressione a ritroso permette di individuare quelle interazioni tra predittori che potrebbero non emergere in approcci più restrittivi, oppure identificabili ma con approcci a complessità computazionale troppo elevata come il BSS.

Se quindi la backward selection permette di valutare i singoli modelli, allora la comparazione tra essi, una volta ridotti, è posta in capo a indici qualitativi e quantificanti diversi da quelli già incontrati. È di conseguenza opportuno considerare anche criteri informativi, oltre che i criteri di vaglio delle significatività statistiche e degli incrementi in precisione di stima (i.e.  $R^2$ ), che possano valutare globalmente la qualità e la parsimonia del modello.

Per un generico modello di regressione, sia  $k$  il numero di parametri stimati e  $\hat{L}$  la funzione di maximum-likelihood del modello, allora si dice AIC(Akaike Information Criterion) la quantità:

$$AIC = 2k - 2\ln(\hat{L}),$$

e BIC(Bayesian Information Criterion) la quantità:

$$BIC = \ln(n) \cdot k - 2\ln(\hat{L}).$$

I due appena descritti sono i principali criteri informazionali di misurazione della qualità di un modello statistico, distinti, di fatto, solo dalla severità con cui essi trattano il numero di parametri e quindi la complessità del modello. Entrambi comunque penalizzano i modelli poco

generali, al fine di evitare selezioni di modelli con problemi di overfitting sul training set, e all'applicazione di entrambi si seleziona il modello cui è associato il valore minore.

Chiaramente, l'andamento logaritmico del BIC al variare del numero di parametri tende a sfavorire con più forza modelli che invece, con AIC, avrebbero livelli di qualità migliori. Ciò però rende i due information criteria adatti a contesti differenti e: se da un lato l'Akaike è più indicato per modelli che mirano alla predizione, dall'altro il Bayesian è più idoneo quando si vogliono solo modellare i dati, ragion per cui quest'ultimo sarà scelto come principale discriminante tra i candidati ridotti.

Infine, sempre in riferimento agli indici qualitativi per la selezione del miglior modello, si considera il coefficiente di determinazione aggiustato  $R_{\text{adj}}^2$ :

$$R_{\text{adj}}^2 = 1 - \left( \frac{(1 - R^2)(n - 1)}{n - p - 1} \right).$$

Questi penalizza l'inserimento di variabili esplicative irrilevanti, favorendo modelli più robusti e parsimoniosi. A differenza del  $R^2$  classico, che tende a crescere con l'aggiunta di nuovi predittori,  $R_{\text{adj}}^2$  cresce solo se la nuova variabile migliora effettivamente l'adattamento del modello. La selezione del miglior modello avverrà quindi confrontando le distribuzioni dei residui mediante istogrammi e test di Shapiro-Wilk, gli indici qualitativi descritti e, dove lo si ritiene opportuno, l'applicazione del Test  $F$  all'ultimo regressore.

## 5.2 Backward Selection Outputs

**Complete Model** Come già evinto in fase di formulazione dei modelli di regressione candidati, *Complete Model* ha ampi margini di miglioramento in termini di parsimonia, poiché popolato da regressori effettivamente poco utili alla stima del valore atteso della variabile indipendente. Ci si prospetta pertanto, avvalendosi della backward selection, implementata mediante il comando R `step(NomeModello, direction = "backward")`, di ricavare un modello almeno equipollente ad *Alternative Model 1*.

In effetti, siano  $H_0$  e  $H_A$  le ipotesi del Test  $t$  tali che:

$$H_0 : \beta_i = 0 \quad \text{e} \quad H_A : \beta_i \neq 0,$$

allora si elimina in sequenza ogni regressore il cui coefficiente non risulta significativamente diverso da zero fino ad ottenere:

$$\hat{y}_x = 55.493 - 9.461(x1\_IS0) - 10.572(x2\_FRatio) + 5.042(x3\_TIME) - 8.893(x5\_CROP),$$

con il rifiuto dell'ipotesi alternativa per i regressori `x6_FOCAL` e `x4_MP` e l'accettazione per i restanti elencati.

Consultando quindi il file `./results/Multiple_Regression.txt` si può facilmente verificare la totale sovrapponibilità del funzionale lineare appena descritto e quello associato ad *Alternative Model 1*. Quanto descritto ha quindi una duplice interpretazione: i risultati ottenuti, anche in riferimento ai valori  $R^2$ ,  $RSE$  e  $F$ -Statistic, di cui si discuterà, per ogni modello, nella sezione 5.3, mostrano non solo un enhancement della parsimonia e delle prestazioni in termini di miglioramento nella precisione di stima, ma anche e soprattutto come il primo alternativo sia il minimo modello non polinomiale in grado di spiegare con significatività statistica l'andamento della qualità video.

Per tale ragione, e a salvaguardia di un contenuto non eccessivamente ridondante, si evita la presentazione dei risultati della backward selection su *Alternative Model 1*.

**Alternative Model 2** Agendo sul secondo modello alternativo come si è fatto con il modello massimo, in virtù del già discusso  $p$ -value alla stima del parametro  $\beta_6$  sul termine di secondo grado associato alla variabile `x3_TIME`, ci si aspetterebbe di ottenere un nuovo modello ridotto. In realtà, per questioni di implementazione del metodo `lm(Alternative Model 2, direction = "backward, ...")` l'output nel FrameWork RStudio che tale procedura fornisce è il seguente:

```

Call:
lm(formula = processedData$y_VideoQuality ~ poly(processedData$x1_ISO,
2) + poly(processedData$x2_FRatio, 2) + poly(processedData$x3_TIME,
2) + poly(processedData$x5_CROP, 1))

Residuals:
    Min      1Q  Median      3Q     Max 
-22.5355 -6.1691 -0.2754  6.6517 30.7669 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 54.6649   0.9333  58.570 < 2e-16 ***
poly(processedData$x1_ISO, 2)1 -92.2181   9.3749 -9.837 5.04e-16 ***
poly(processedData$x1_ISO, 2)2 -62.3988   9.4738 -6.586 2.75e-09 ***
poly(processedData$x2_FRatio, 2)1 -101.4934  9.3857 -10.814 < 2e-16 ***
poly(processedData$x2_FRatio, 2)2 -75.0058   9.9444 -7.543 3.18e-11 ***
poly(processedData$x3_TIME, 2)1   44.2548   9.4681  4.674 1.01e-05 ***
poly(processedData$x3_TIME, 2)2   10.3298   9.8214  1.052   0.296  
poly(processedData$x5_CROP, 1)   -105.4467  9.7863 -10.775 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.333 on 92 degrees of freedom
Multiple R-squared:  0.831,    Adjusted R-squared:  0.8182 
F-statistic: 64.63 on 7 and 92 DF,  p-value: < 2.2e-16

```

Chiaramente questo è un bug, la cui causa è da ritrovare nell'algoritmo che il software RStudio mette a disposizione, risolvibile con la costruzione esplicita del modello corretto dalla backward. Detto *Alternative Model 3*, il nuovo modello ridotto presenta i coefficienti  $\hat{\beta}_i$ :

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 54.6649   0.9339  58.537 < 2e-16 ***
poly(processedData$x1_ISO, 2)1 -92.4341   9.3780 -9.856 4.12e-16 ***
poly(processedData$x1_ISO, 2)2 -63.0276   9.4603 -6.662 1.87e-09 ***
poly(processedData$x2_FRatio, 2)1 -102.1467  9.3705 -10.901 < 2e-16 ***
poly(processedData$x2_FRatio, 2)2 -77.7817  9.5933 -8.108 2.02e-12 ***
poly(processedData$x3_TIME, 1)    44.1998   9.4734  4.666 1.03e-05 ***
poly(processedData$x5_CROP, 1)   -107.5113  9.5928 -11.207 < 2e-16 ***
---

```

La forma ridotta di Alternative Model 2, cioè il modello Alternative Model 3, restituisce la stima della media  $E[Y|X = x]$ ,  $\hat{y}_x$ :

$$\begin{aligned}\hat{y}_x = & 54.6649 - 92.4341(x_{1\_ISO}) - 63.0276(x_{1\_ISO})^2 - 102.1467(x_{2\_FRatio}) \\ & - 77.817(x_{2\_FRatio})^2 + 44.1998(x_{3\_TIME}) - 107.5113(x_{5\_CROP}).\end{aligned}$$

Anche in questo caso la procedura backward si è dunque rivelata efficace nel produrre un modello statisticamente solido e parco, riducendo il rischio di overfitting. La sua applicazione progressiva – dapprima su un modello lineare completo, poi su una specificazione polinomiale – ha consentito un affinamento graduale della struttura predittiva. Le valutazioni finali in merito alle performance dei modelli sono rimandate alla sezione successiva, anche se, per via di quanto già presentato rispetto ai termini polinomiali aggiunti, quello appena presentato sembra essere il più promettente.

### 5.3 Model Comparison and Final Choice

Si procede ora con la comparazione formale; per individuare il modello ottimale tra quelli ottenuti, valutati secondo i criteri di Sez. 5.1, si presenta la seguente tabella riassuntiva:

Modello	$R^2$	$R^2_{adj}$	RSE
Complete Model	0.6143	0.5894	14.02
Alternative Model 1	0.6124	0.5961	13.91
Alternative Model 2	0.831	0.8182	9.333
Alternative Model 3	0.829	0.8179	9.339

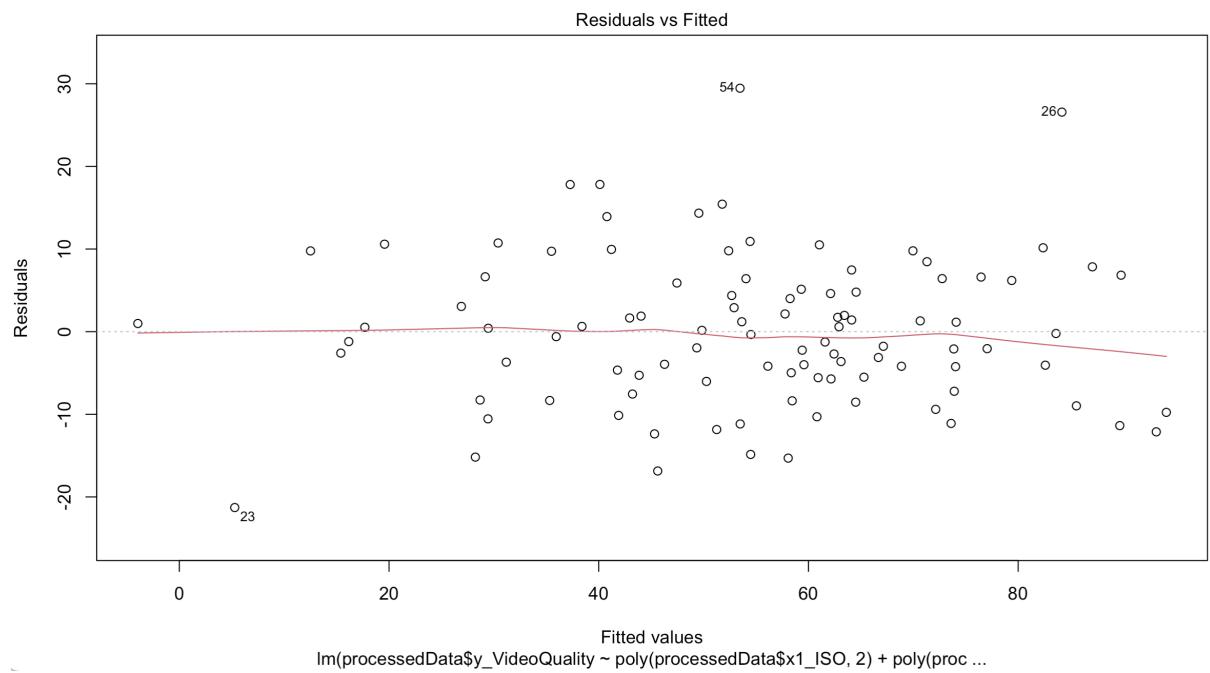
È evidente che i modelli ottenuti come riduzioni - ovvero Alternative Model 1 e Alternative Model 3 - non solo sono più economici ma anche più prestanti in termini di aliquota di variabilità spiegata. Inoltre, come previsto, il terzo modello alternativo risulta essere il migliore in termini di rapporto prestazione/costo: con un termine quadratico in meno si riesce a mantenere sostanzialmente invariato l'errore standard sui residui e  $R^2$ .

Analizzando infatti Figure 27 (a pagina successiva) si può osservare una buona nuvola dei residui  $e_i = y_i - \hat{y}_i$ , ad indicare una varianza piuttosto costante  $\forall i$ . È bene però valutare come i residui si distribuiscono non solo in maniera grafica, al fine di verificarne la normalità, a causa delle inconsistenze sulle code che l'istogramma mostra. Conducendo un test di Shapiro-Wilk si ottiene un  $p$ -value  $\approx 0.216$ , rifiutando così l'ipotesi alternativa e sciogliendo ogni dubbio sulla confermata normalità dei residu.

Il confronto conclusivo tra i modelli, così da decretarne la scelta, coinvolge la valutazione del BIC, i cui rispettivi valori sono riportati nella tabella di seguito.

Modello	BIC
Complete Model	841.5301
Alternative Model 1	832.8119
Alternative Model 2	763.6115
Alternative Model 3	760.20

Lo script "`./src/decision.R`" contiene, per completezza, anche la valutazione dell'AIC associato ai modelli, da cui emerge come *Alternative Model 3* sia in effetti anche il migliore tra i candidati non solo per il fitting al training set ma anche per la predizione su un eventuale test set, avendo difatti i valori più bassi sia per l'Akaike che per il Bayesian Information Criterion.



**Histogram of the third Alternative Model's residuals**

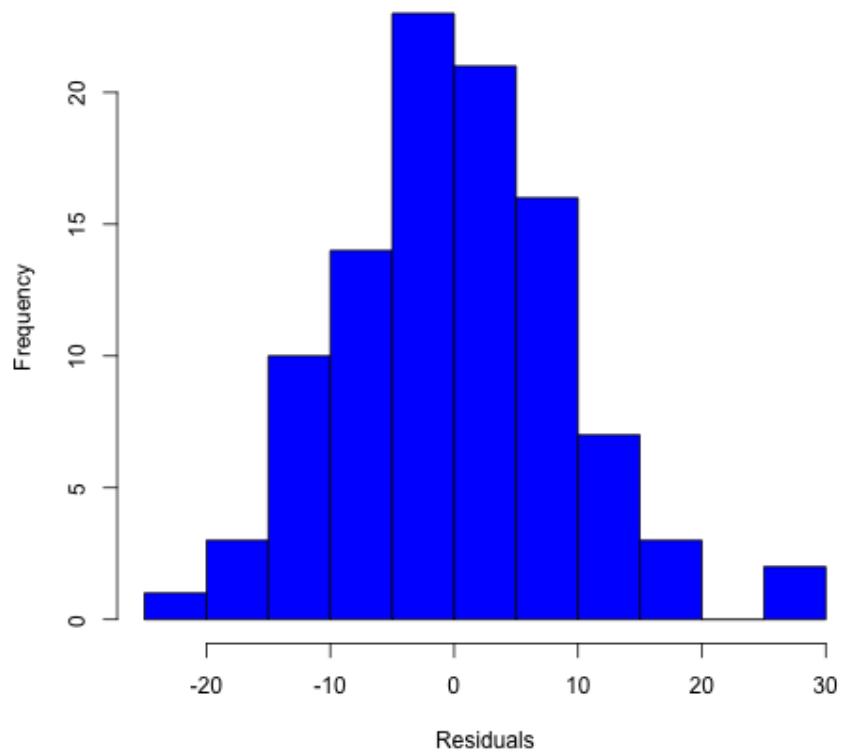


Figure 27: Diagnostica grafica del terzo modello alternativo.

Con intervalli di confidenza sulle stime dei parametri al 2.5-esimo e 97.5-esimo percentile tali che:

**Confidence Intervals:**

	2.5 %	97.5 %
(Intercept)	52.81050	56.51939
poly(processedData\$x1_IS0, 2)1	-111.05703	-73.81123
poly(processedData\$x1_IS0, 2)2	-81.81388	-44.24133
poly(processedData\$x2_FRatio, 2)1	-120.75452	-83.53878
poly(processedData\$x2_FRatio, 2)2	-96.83199	-58.73138
poly(processedData\$x3_TIME, 1)	25.38759	63.01207
poly(processedData\$x5_CROP, 1)	-126.56083	-88.46186

e con  $SQE$  e  $MSQE$  tali che:

$$SQE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \approx 8110.34,$$

$$MSQE = \frac{SQE}{v} \approx 0.8721, \quad \text{con } v \text{ gradi di libertà,}$$

si decreta la selezione del secondo alternativo ridotto, vale a dire *Alternative Model 3*, come il modello di regressione lineare multipla che stima la media della qualità video percepita dagli utenti al variare della sensibilità del sensore, del rapporto focale, del tempo di esposizione e del fattore di crop.

Ulteriori approfondimenti e key-values, come  $F$ -Statistic, SD residuals e summary della distribuzione dei residuals possono essere consultati al file "../results/models/chosen/Chosen.txt".

## 6 Domain Insights: Relating Statistical Findings to Video Quality

Il modello finale selezionato, *Alternative Model 3*, oltre a rappresentare la migliore soluzione in termini di prestazioni statistiche, riflette coerentemente anche la logica tecnica che lega le variabili indipendenti alla qualità video percepita (*y\_VideoQuality*). Quest'ultima, si ricorda essere un indice derivato dal giudizio umano su aspetti visivi quali nitidezza, rumore, gamma dinamica, fedeltà cromatica, profondità di campo, risoluzione e presenza di artefatti.

Prima però di descrivere le proprietà del modello nel dominio applicativo, è bene proporre una panoramica che metta in evidenza l'informazione che i regressori del modello trasportano.

Il predittore *x1\_ISO*, rappresenta la sensibilità del sensore alla luce. È noto che valori mediobassi di ISO garantiscono lo scatto di immagini nitide e prive di rumore, mentre valori elevati, in contesti inadatti, incrementano il rumore digitale, compromettendo tale nitidezza e mostrando impatti anche sulla fedeltà cromatica e la gamma dinamica, come mostrato in Figure 28. Il modello proposto si dimostra essere in ciò piuttosto accurato, esplicitando coefficienti negativi sia su termine lineare che quadratico, generando quindi una curva concava decrescente che penalizza con durezza osservazioni a valori alti o eccessivamente bassi.

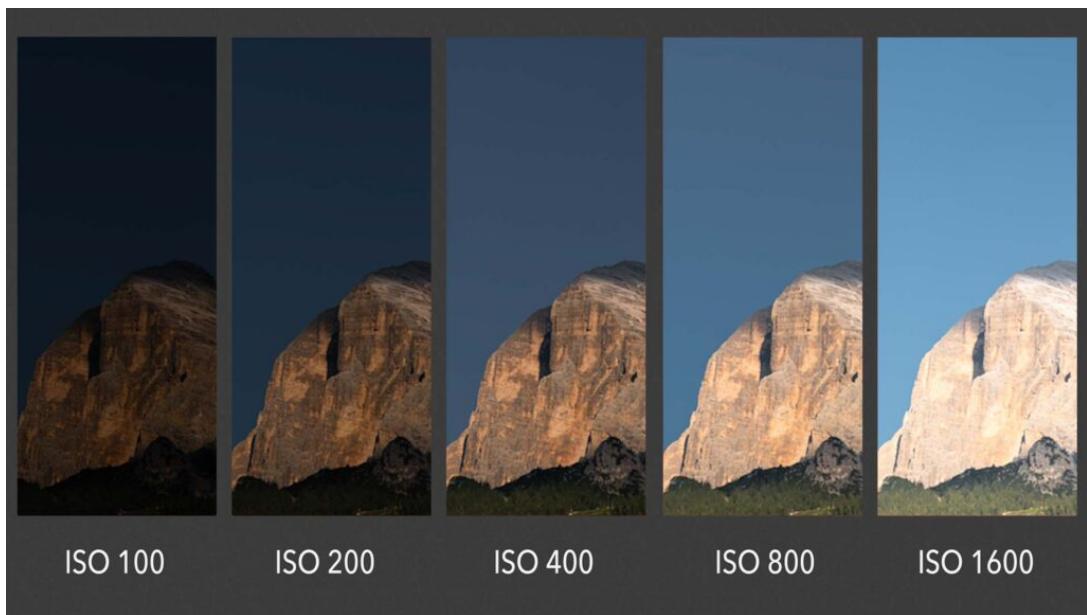


Figure 28: Da sinistra a destra: scatti con medesime condizioni di luce, ma differenti ISO.

La variabile *x2\_FRatio*, invece, descrive l'apertura del diaframma; se un diaframma troppo aperto o troppo chiuso causa problemi visivi, come ad esempio una ridotta profondità di campo o aberrazioni ottiche, allora non stupisce ritrovare correlazioni negative della variabile rispetto alla qualità video. Anche qui infatti, la *y\_VideoQuality* risulta ottimale a valori intermedi di apertura per via del termine quadratico.

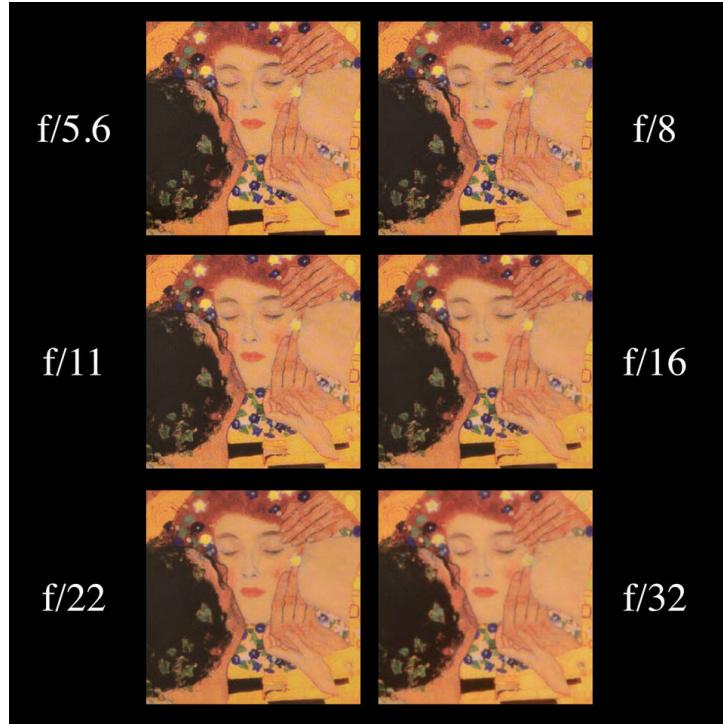


Figure 29: Da sinistra a destra e dall’alto al basso: scatti con rapporto focale differente, con rispettivi valori di fianco riportati.

Il tempo di esposizione poi, espresso tramite `x3_TIME`, ha una relazione lineare positiva con la variabile dipendente. Un’esposizione più lunga consente una maggiore quantità di luce catturata per ogni fotogramma, riducendo la sottoesposizione e migliorando la resa luminosa. Questo comporta una maggiore nitidezza e una minore incidenza di artefatti visivi, contribuendo ad una qualità migliore. È però da notare che, poiché si sta parlando di qualità percepita da utenti, la relazione lineare potrebbe anche essere sintomo dell’impatto che ha, all’occhio umano, uno scatto a tempi di esposizione più lunghi (i.e. motion blur).



Figure 30: Da sinistra a destra: scatti con tempi di esposizione differenti, rispettivamente pari a  $2s$  e  $1/2s$ .

Infine, il fattore di Crop del sensore (`x5_CROP`) è associato negativamente alla qualità. Valori più elevati di crop indicano un sensore più piccolo o una maggiore porzione di immagine

ritagliata, con conseguente perdita di profondità di campo, ampiezza visiva e resa ottica complessiva, accentuando gli effetti causati dal movimento. L'effetto, ben rappresentato nel modello da un coefficiente negativo marcato, a testimoniare il forte impatto sfavorevole del crop sulla percezione visiva, può essere anche interpretato come una sintomatologia della percezione che un utente ha alla visione di un frame meno profondo.

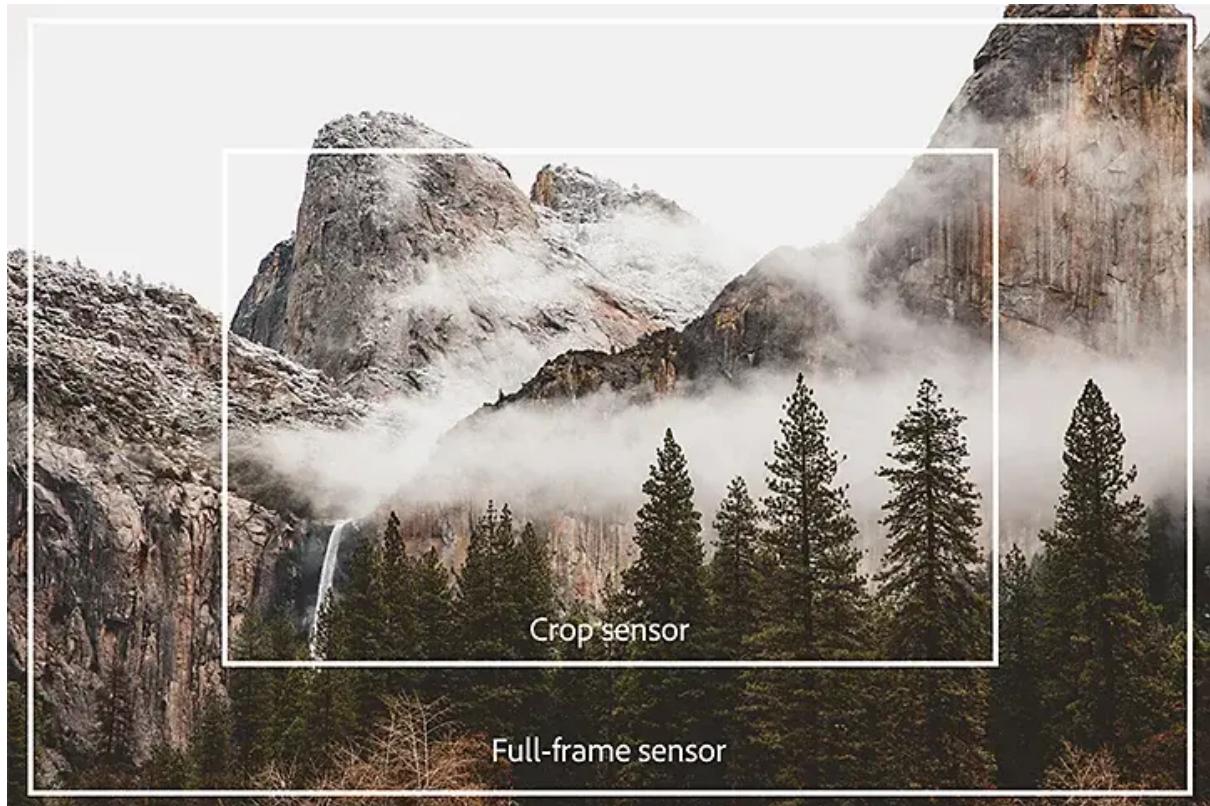


Figure 31: Differenza tra sensori a Full-frame e crop.

Ciascuna delle relazioni osservate tra i predittori selezionati e la variabile di risposta sembra trovare una motivazione concreta e logica nel funzionamento fisico e percettivo della ripresa video, e nello scatto di foto, a conferma dell'affidabilità interpretativa del modello finale. È però da tenere a mente che l'analisi condotta, per via della cardinalità dei campioni forniti e del numero di variabili, tiene in considerazione solo quelle condizioni di ripresa sulle quali un operatore macchina potrebbe avere il controllo, le quali chiaramente non sono le uniche ad influenzare i risultati finali.

In tal senso infatti, un data set più vasto, eventualmente divisibile in train e test set, e un numero maggiore di parametri, ad esempio con regressori che forniscono informazioni in merito all'orario di scatto e alle condizioni di luce naturale, avrebbero guidato l'analisi in una direzione differente, portando allo sviluppo di modelli potenzialmente predittivi e ancor più attinenti al dominio applicativo.

I risultati ottenuti quindi, nel loro insieme, offrono una solida base per eventuali sviluppi futuri, sia in termini di estensione del modello a contesti più ampi, sia nell'approfondimento di effetti interattivi o non osservabili nel presente dataset.