Applied Statistics

2024/25

A DATASET ANALYSIS

Design of a Linear Regression Model with RStudio

Group 4

Mario Pellegrino Ambrosone	0612707417
Emanuele Barbato	0612707998
Luca Celentano	0612707836
Francesco De Bonis	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

This report, to be understood as an appendix to the documentation of the code developed with the RStudio Framework, aims to present and explain the analysis of the dataset `"DataSet_gruppo4-RAW"`, which can be found in the directory `"Regression_Analysis/data/"`, along with the results produced as the output of the code, which can be found in the directory `"Regression_Analysis/results/"`, as well as a discussion of the theoretical foundations supporting the analysis itself.
The document is therefore structured into several sections, each responsible for expanding and deepening at least one of the stages of the dataset analysis development.
Before proceeding, it is important to briefly describe the dataset of interest, explain its role within the project, and discuss the importance of statistical analysis when dealing with data.
The provided data sample has indeed been interpreted as the training set for a multiple linear regression model, aimed at defining a functional relationship that could explain the trend of perceived video quality in a sample of images (`y_VideoQuality`) as the shooting/capture parameters and the equipment used vary.
No further assumptions were made on the potential predictors, i.e., the variables described as independent variables in `"Regression_Analysis/assignment/VariabiliProgStatAppl_DEF_24_25"`, beyond those provided in the aforementioned file path. This choice allows for evaluating the data while maintaining a level of simplicity sufficient to define a model that can explain the trend of the response variable, which will not be tested on a different dataset.
The analysis goal, then, is to carefully model the data currently available, revealing relationships and connections, without making prediction the central issue.

Contents

1 Data Handling and Preprocessing

In order to ensure a correct analysis and proper application of known statistical methods, an exploratory evaluation of the dataset is first necessary. Through this, it is possible to check the homogeneity and consistency of the information it represents, the absence of NA values (or at least a non-significant presence), the dimensions, and any other discrepancies.
The preprocessing goal is to clean the dataset, and in the case at hand, it is relatively simple (see `"./src/dataPreprocessing.R"`). The assigned dataset, henceforth referred to as `dataRaw`, indicating the set of data that has not yet been processed, contains all the variables declared in `"assignment/VariabiliProgStatAppl_DEF_24_25"`, including one response variable and seven potential regressors.

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

Figure 1: Output of `names(dataRaw)`

Each column in `dataRaw` is consistent, with no NA values, with mutually different and always meaningful names (as shown in Figure 1).
Closing the preprocessing phase, the function `str(dataRaw)` explicitly shows the domains of each variable, confirming that none of them are categorical (no binary variables), and that they are all of type `numeric`, making them suitable for linear regression modeling.
The provided data does not seem to require extensive cleaning procedures or further handling: the description of each variable follows.

2 Data Exploration

Descriptive statistics methods, during the preliminary phases of data exploration, are useful for characterizing the variables and defining their distribution. The file `"./src/descriptiveAnalysis.R"` specifically deals with the study of the data in order to extract what is necessary to observe the patterns linking them and to deduce properties from their distribution functions. This ensures a general understanding of the information they contain and provides the foundation for evaluating correlations and collinearity.

2.1 Data Structure and Outliers Evaluation

A first approach to this information is certainly the dataset summary, the output of which, besides being available in the results directory under the name `"./results/CharacterizedDataSet.txt"`, is reported below.

```
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
```

In addition to re-checking the dimensional correctness of `dataRaw`, reported as `Dimension`, it is possible to observe how the values of individual observations of the response variable have a significantly broader variability range compared to the predictors. The discrepancy between the `Median` and the `Mean` for this variable seems to indicate a marked skewness in the distribution, which could be a sign of the presence of potential outliers. Regarding the independent variables, the variability shown by `x7_PixDensity` and `x5_CROP` between their respective `Max.` values and the means could also suggest the presence of outliers. To rule out this hypothesis, we first observe the overall box-plot of the variables and then the individual box-plots for each of them, available in the directory "`./results/boxplots/`", and presented below.

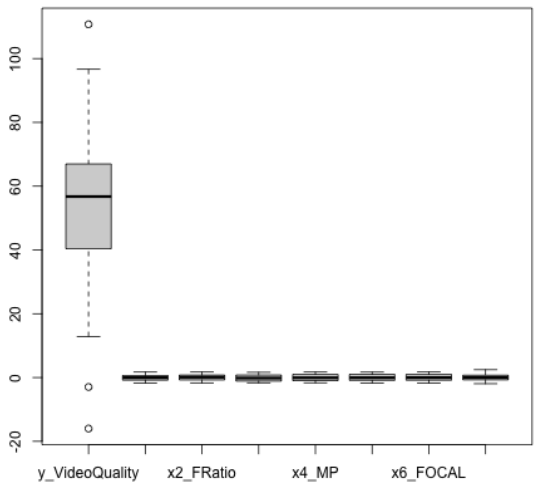


Figure 2: Box-plot of all variables. Outliers are visible on the response variable.

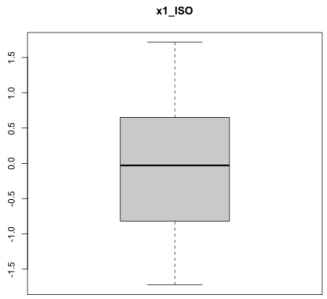


Figure 3: x1_ISO

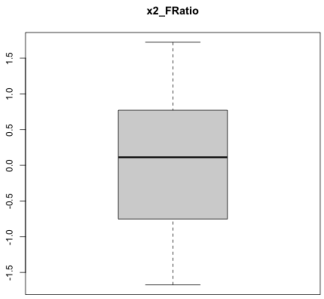


Figure 4: x2_FRatio

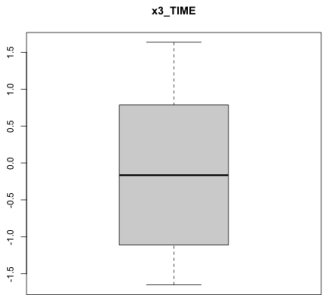


Figure 5: x3_TIME

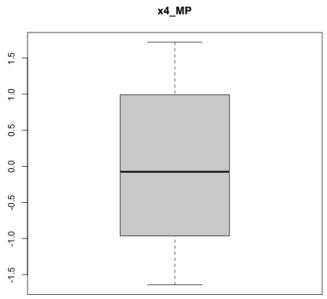


Figure 6: x4_MP

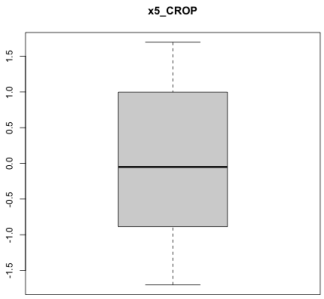


Figure 7: x5_CROP

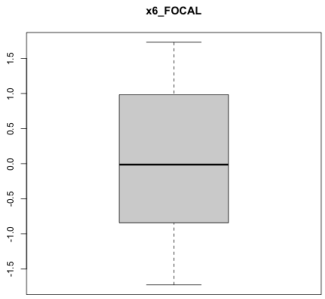


Figure 8: x6_FOCAL

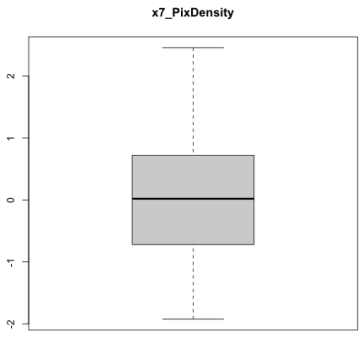


Figure 9: x7_PixDensity

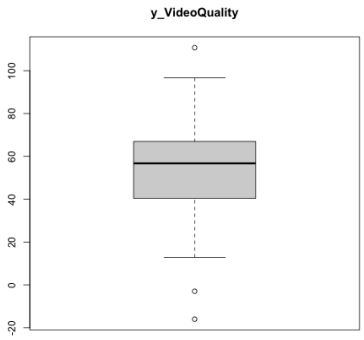


Figure 10: y_VideoQuality

A graphical analysis of the box-plots (from Figure 3 to Figure 9) allows us to observe the absence of outliers in the independent variables. The boxplot of the response variable (Figure 10) shows the outliers already identified in Figure 2, outside the whiskers of the boxplot, which are defined as 1.5 times the IQR (Interquartile Range). However, it is important to consider that these value's polarization, although may represent an issue, is not extreme. Therefore, it is hypothesized, for the time being and for simplification purposes, that the detected outliers do not have significant effects on the regression model coefficients, on the MSE, or on the normality of the residuals. This hypothesis will be further discussed during the model proposal phase (see Section Defining Regression Models), by evaluating the distribution of the residuals.

2.2 Variables Distribution

The characterization of the variables in `dataRaw` continues with the discussion of the distribution of each of them, clearly constructed through histograms, which are presented below. It is worth noting that all the distributions are accessible and contained in the directory `"./results/histograms/"`.

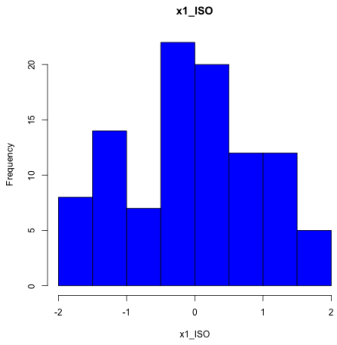


Figure 11: `x1_ISO`

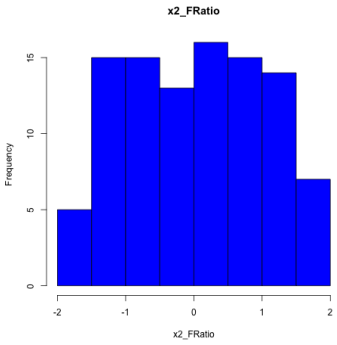


Figure 12: `x2_FRatio`

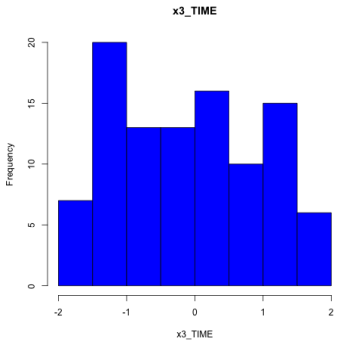


Figure 13: `x3_TIME`

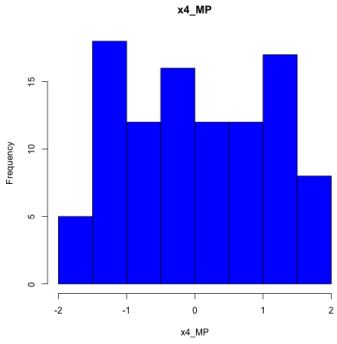


Figure 14: `x4_MP`

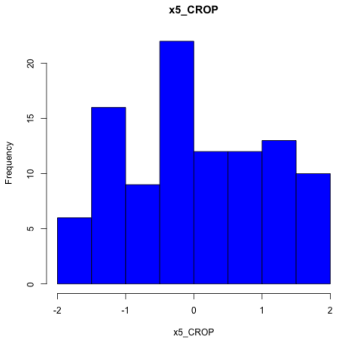


Figure 15: `x5_CROP`

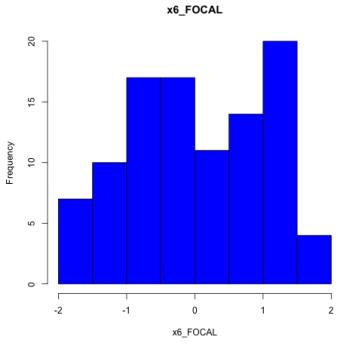


Figure 16: `x6_FOCAL`

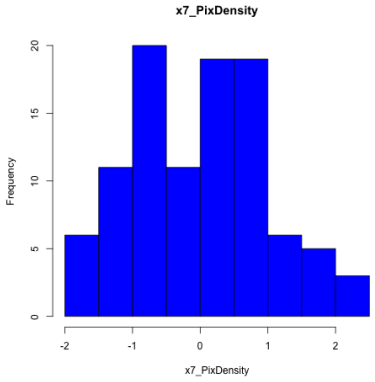


Figure 17: `x7_PixDensity`

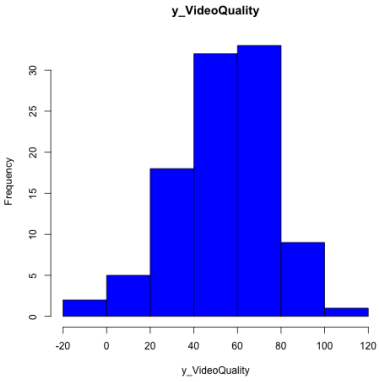


Figure 18: `y_VideoQuality`

Although a graphical analysis of the outputs strongly suggests that, except for the response variable, none of the variables in the dataset are normally distributed, the Shapiro-Wilk test is still performed on each of them. We expect to reject the null hypothesis (H_0: The data come from a normal distribution) for all the independent variables, and to accept it only for `y_VideoQuality`. The results of the tests, which can be found in the file `"./results/Shapiro-Wilk.txt"`, follow.

```
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
```

It is noted that for the dependent variable `y_VideoQuality` (first 4 rows of the output):

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

and therefore, the alternative hypothesis H_0 is rejected:

$$H_A : \text{The data do not come from a normal distribution.}$$

While the skewness observed in the box-plot does not seem to have major implications on the distribution of the variable, it cannot be stated the same for the outliers. After confirming normality in distribution, we also evaluate the Q-Q Plot (see Figure 19), which is also reported in the `"./results"` directory, in order to further characterize the behavior of the observed polarized values.

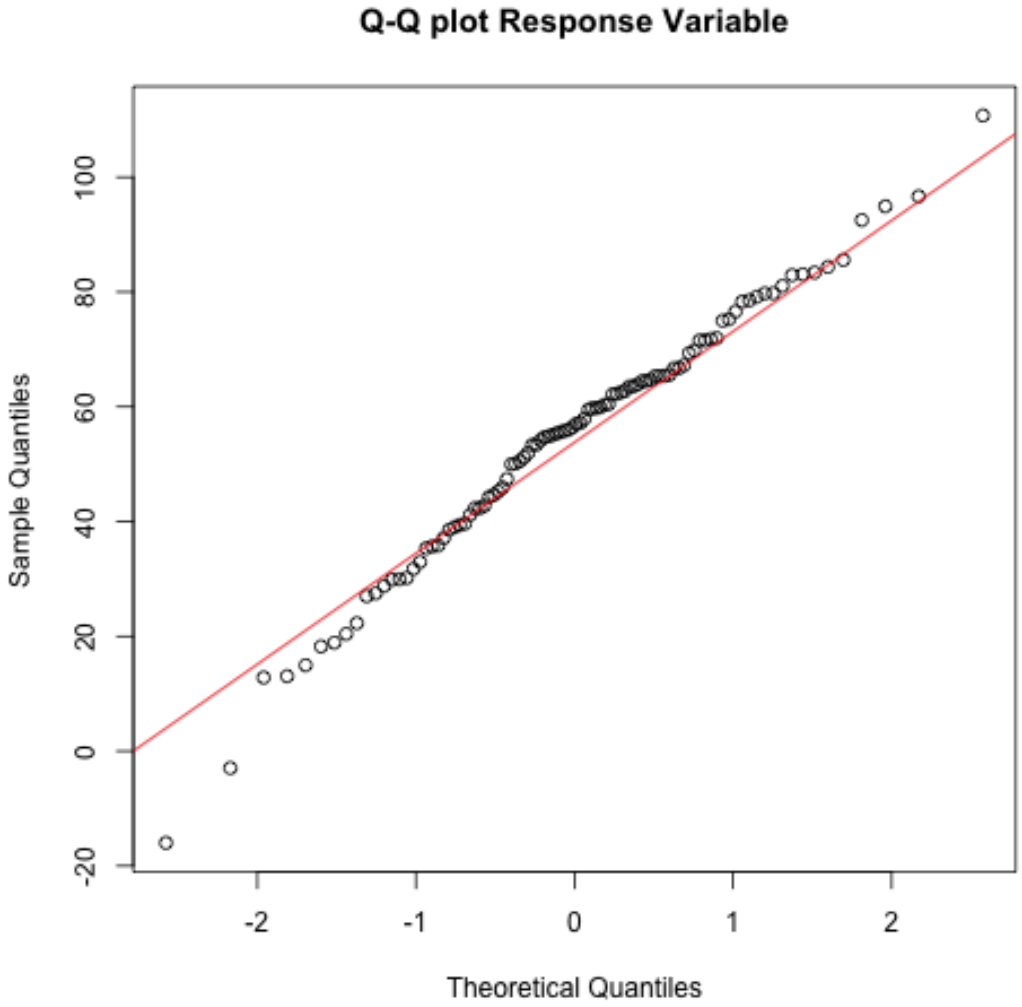


Figure 19: Q-Q plot for `y_VideoQuality`. The effects of outliers are noticeable.

The good alignment of the points along the diagonal confirms what emerged from the Shapiro–Wilk test, while the slight dispersion at the tails suggests the possible influence, potentially negligible, of the polarized values. The verification of the normality of `y_VideoQuality` represents an important prerequisite for the modeling phase, as it allows for the reasonable assumption of multivariate Gaussian error in the multiple linear regression.
The definitive conclusions on the impact of the discrepancies that have emerged, however, will be formulated in the model diagnostics section, with particular reference to the analysis of the residuals. As for all the other variables, the p-value associated with the test is always lower than the significance level, allowing us to accept the alternative hypothesis H_A.

3 Interaction Analysis: Investigating Correlations

In order to safeguard the development of the regression models, it is necessary to leverage the findings from the data exploration to study the functional relationships between the individual variables of `dataRaw`. Therefore, it is the responsibility of the Interaction Analysis phase to quantify the statistically significant correlation indices, as well as to produce a new training set should these indices invalidate the integrity of the models due to relevant collinearity.

3.1 Visualization and Statistical Hypothesis Testing

A first graphical evaluation of the correlation coefficients between the variables is provided by the heatmap, which can be found in the file `"./results/heatmaps/heatmap_raw_dataset.pdf"`, and is presented below.

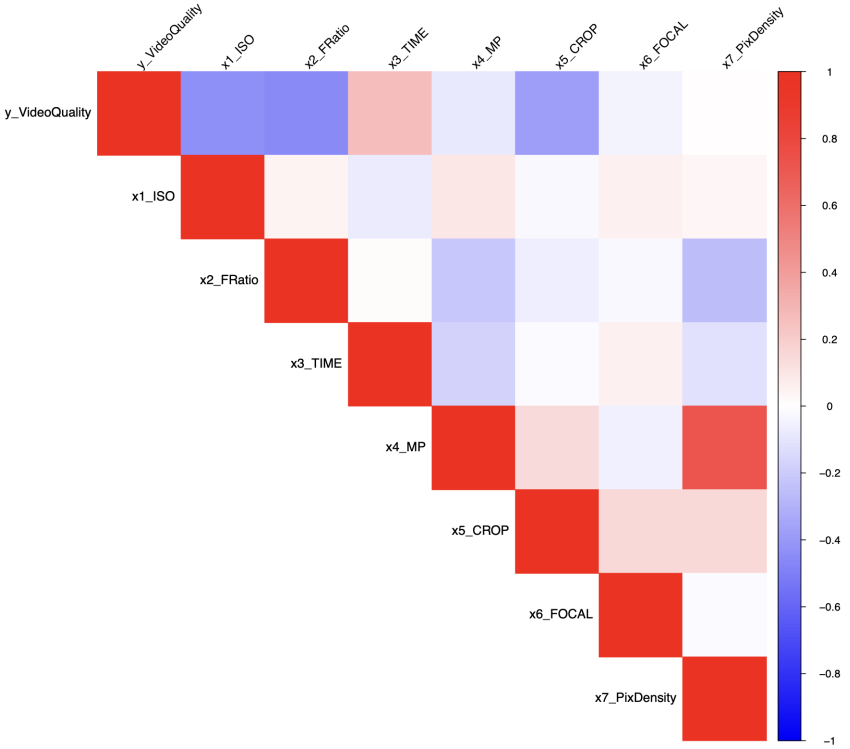


Figure 20: Heatmap for `dataRaw`.

It is immediately evident that there is a strong positive correlation between `x4_MP` and `x7_PixDensity`, and a moderate negative correlation between the response variable and the variables `x1_ISO`, `x2_FRatio`, and `x5_CROP`. However, in order to give greater authority to these observations, it is necessary to quantify these correlations by computing the correlation matrix. A hypothesis test will then be performed on its elements to assess their statistical significance. In Figure 21, the output of the `GGally::ggpairs(dataRaw)` function is shown, which, when executed in the framework, returns the scatter plot of the variables in `dataRaw` with the associated correlation coefficients and p-values. It is worth noting that the R code implemented in the source file `"./src/regressionAnalysis.R"` produces as output the correlation matrix directly in the framework and saves the p-value matrix in the `./results/` directory. The file `./results/scatterplots/scatter-plot_All.png` offers an alternative and streamlined version of the scatter plot for all variables.

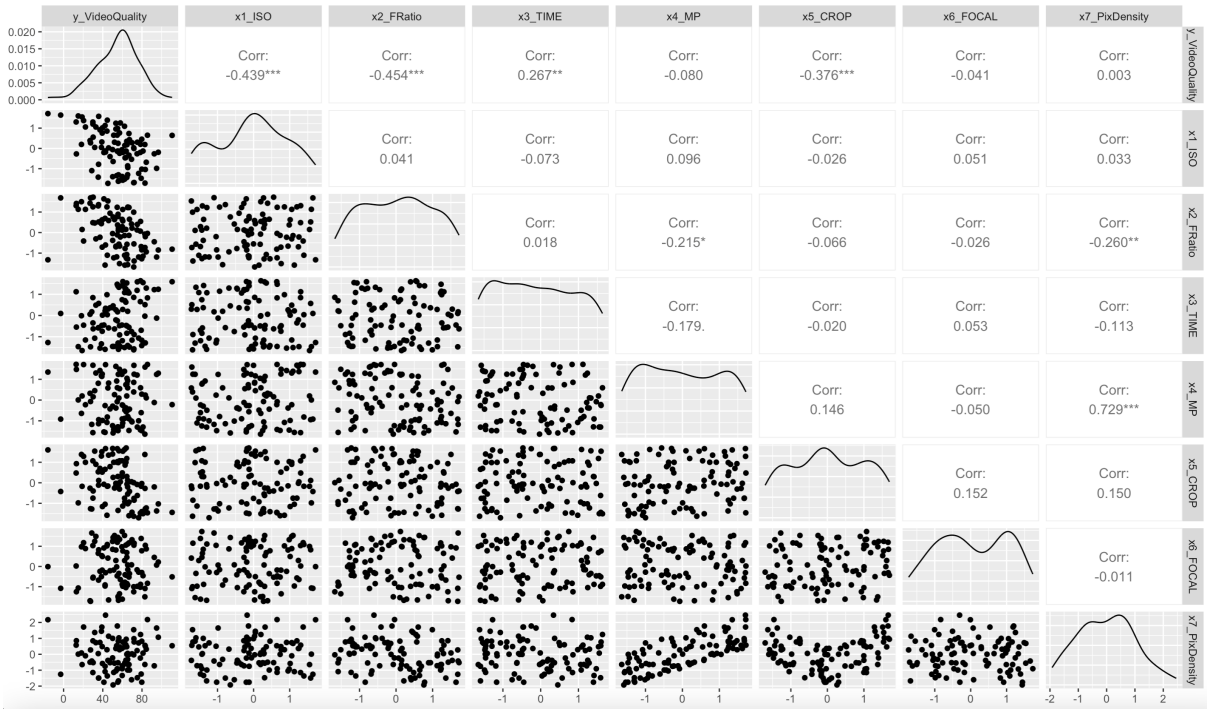


Figure 21: Complete scatter plot with correlation coefficients and distribution of observations for `dataRaw`.

Upon observing the results shown in Figure 21, it is possible to affirm that the conclusion drawn from the heatmap in Figure 20 is correct: there is a clear and concerning collinearity between the variables `x4_MP` and `x7_PixDensity`.
For each pair of variables, a hypothesis test is conducted such that:

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

In the case of `x4_MP` and `x7_PixDensity`, we have:

$$R = 0.729,$$

with a p-value:

$$p\text{-value} = 0.00 < 0.05 \Rightarrow H_A \text{ accepted.}$$

The following correlations also emerge as significant:

- `y_VideoQuality` and `x1_ISO` with a correlation coefficient of $R = -0.439$ and high statistical significance (p-value $\approx 4.88 \cdot 10^{-6} \Rightarrow H_0$ rejected)
- `y_VideoQuality` and `x2_FRatio` with a correlation coefficient of $R = -0.454$ and high statistical significance (p-value $\approx 2.13 \cdot 10^{-6} \Rightarrow H_0$ rejected)
- `y_VideoQuality` and `x3_TIME` with a correlation coefficient of $R = 0.267$ and high statistical significance (p-value $\approx 0.0073 \Rightarrow H_0$ rejected)
- `y_VideoQuality` and `x5_CROP` with a correlation coefficient of $R = -0.376$ and high statistical significance (p-value $\approx 0.0001 \Rightarrow H_0$ rejected)
- `x2_FRatio` and `x7_PixDensity` with a correlation coefficient of $R = -0.260$ and high statistical significance (p-value $\approx 0.0090 \Rightarrow H_0$ rejected)

– x4_MP and x2_FRatio with a correlation coefficient of $R = -0.215$ and moderate statistical significance (p-value $\approx 0.0031 \Rightarrow H_0$ rejected)

It is noted that the degree of correlation between each independent variable and the response variable y_VideoQuality provides an indication of the relative weight of that predictor within the regression model. Therefore, at this point, evaluating the correlation coefficients would be of little use.
The problematic relationship is, in fact, just the one between x4_MP and x7_PixDensity, which, if ignored, could prevent the covariance matrix of the regression model from respecting the invertibility property, thus hindering the modeling process. For this reason, a dataset, called "./data/DataSet_gruppo4-PROCESSED.csv", is produced as a new training set, which lacks the variables deemed unsuitable for regression modeling.
Thus, the variable x7_PixDensity is removed from the training set. Regarding the last two correlations listed, the reduced value of R they exhibit does not, in fact, create a significant collinearity problem. Therefore, there is no need to remove either of them.
For completeness, the heatmap of the new dataset can be consulted, contained in the file "./results/heatmaps/heatmap_processed_dataset.pdf".

3.2 Polynomial Regression Analysis for Capturing Non-Linear Effects

In order to explore the possibility that some relationships between the independent variables and the response may not be strictly linear, a polynomial regression analysis was conducted. This approach allows modeling a nonlinear dependency by transforming it into a linear relationship in the parameters, thus maintaining compatibility with the structure of classical linear regression.
Particular attention should be given to the construction of polynomial models between y_VideoQuality and the regressors x1_ISO, x2_FRatio, x3_TIME, and x5_CROP in order to further characterize the correlations identified in the previous section.
Let a target polynomial model be defined as:

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

Then, the following models are constructed with the optimal degree empirically identified.

$$\begin{aligned}
\texttt{x1_ISO}: &\quad \hat{y}_\texttt{VideoQuality}_x = \hat{\beta}_0 + \hat{\beta}_1 x1_\texttt{ISO} + \hat{\beta}_2 x1_\texttt{ISO}^2,\\
\texttt{x2_FRatio}: &\quad \hat{y}_\texttt{VideoQuality}_x = \hat{\beta}_0 + \hat{\beta}_1 x2_\texttt{FRatio} + \hat{\beta}_2 x2_\texttt{ISO}^2,\\
\texttt{x3_TIME}: &\quad \hat{y}_\texttt{VideoQuality}_x = \hat{\beta}_0 + \hat{\beta}_1 x3_\texttt{TIME} + \hat{\beta}_2 x3_\texttt{TIME}^2,\\
\texttt{x5_CROP}: &\quad \hat{y}_\texttt{VideoQuality}_x = \hat{\beta}_0 + \hat{\beta}_1 x5_\texttt{CROP}.
\end{aligned}$$

Further insights and details regarding the development of polynomial regression models on the predictor pairs, as well as detailed descriptions of the functionals discussed above, can be found in the file "./results/Polynomial_Regression.txt"; it contains a summary of each model, including parameter estimates, p-values of the t-Tests on each parameter, residual summaries, standard deviation of the residuals, the coefficient of determination R^2, and the F-statistic for quantifying the variability explained by the model.
In light of the findings of the cited file, it is then possible to confidently confirm the goodness of the predictors x1_ISO, x2_FRatio, x3_TIME, and x5_CROP in contributing to the explanation of perceived video quality. These polynomial findings confirm the importance of including nonlinear terms in some predictors, which is why the influence of the functionals derived above will be considered in the formulation of the regression models, presented in the next section.

4 Defining Regression Models

Following the findings from the Data Exploration and Interaction Analysis, attention should be focused on the construction of the actual regression models.
The script `./src/regressionModel.R` formalizes several modeling hypotheses, ranging from the complete multiple linear model to multiple and polynomial configurations. This phase allows for the identification of potential issues, such as anomalies in collinear dependencies, which may require a more in-depth diagnostic analysis. The goal is to formulate the model that can best explain the variability in the observations of `y_VideoQuality`, focusing on goodness of fit, with the hope of achieving predictive accuracy, while adhering to Occam's razor (the parsimony principle).

4.1 Linear and Polynomial Approaches

Consider the processed dataset, discussed earlier, hereafter referred to as `processedData`, and it is then possible to construct a multiple regression model in the form:

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

For a linear function $f(X_1, \cdots, X_p)$, where $Y = \texttt{y\_VideoQuality}$ and $p = 6$, it is such that:

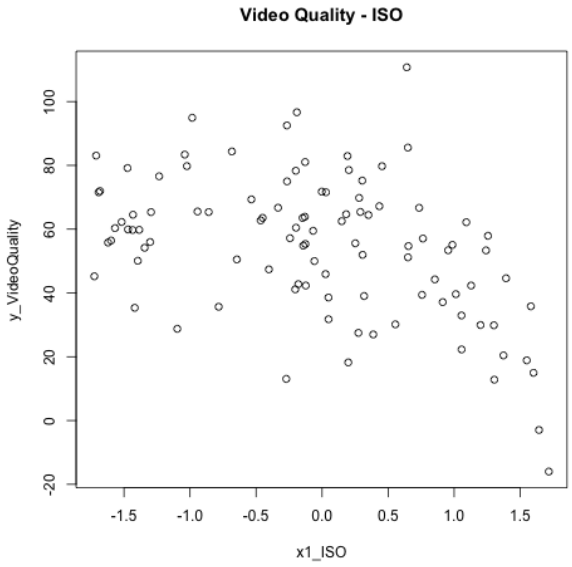
$$\begin{aligned} \texttt{y\_VideoQuality} &= \beta_0 + \beta_1 x1\texttt{\_ISO} + \beta_2 x2\texttt{\_FRatio} + \\ &+ \beta_3 x3\texttt{\_TIME} + \beta_4 x4\texttt{\_MP} + \beta_5 x5\texttt{\_CROP} + \beta_6 x6\texttt{\_FOCAL} + \varepsilon. \end{aligned}$$

This is indeed the main model under analysis (hereafter referred to as the *Complete Model* since it represents the maximum model), but not the only one. There are two alternative models, proposed based on the significant influence that some regressors have shown on the dependent variable, but also and especially due to the polynomial trends previously derived and characterized.Consider the variables `x1_ISO`, `x2_FRatio`, `x3_TIME`, and `x5_CROP`, whose scatter plots, in addition to being available in the `"./results/scatterplots"` directory, are shown in Figure 22.
It is then possible to define the first alternative model (referred to as *Alternative Model 1*) as a reduced multiple model, such that:

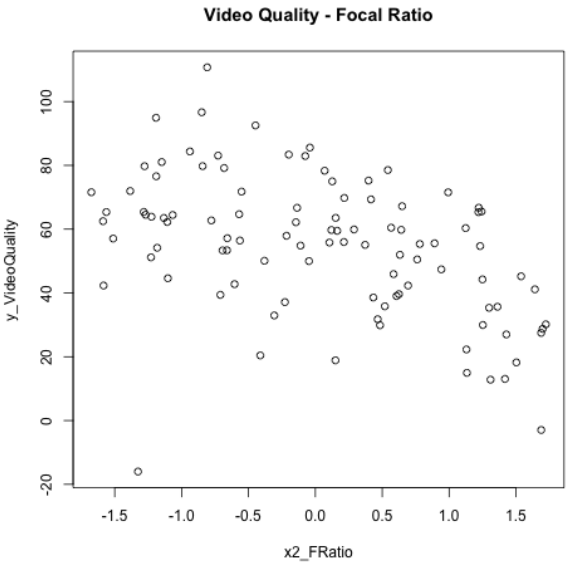
$$\texttt{y\_VideoQuality} = \beta_0 + \beta_1 x1\texttt{\_ISO} + \beta_2 x2\texttt{\_FRatio} + \beta_3 x3\texttt{\_TIME} + \beta_4 x5\texttt{\_CROP} + \varepsilon.$$

There is still the possibility that this alternative model may not be able to explain the variability in `y_VideoQuality` sufficiently accurately, as it does not account for the quadratic terms that fundamentally characterize the interactions between these predictors and the response variable. Therefore, a second alternative model (referred to as *Alternative Model 2*) is formulated, which highlights the polynomial regressions discussed in section 3.2, such that:

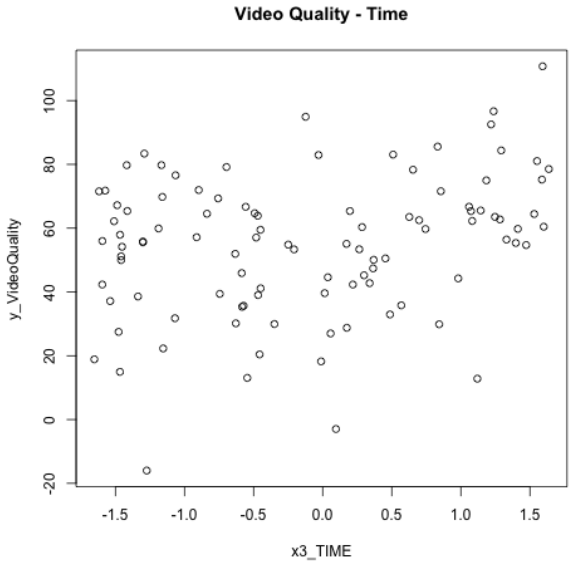
$$\begin{aligned} \texttt{y\_VideoQuality} &= \beta_0 + \beta_1 x1\texttt{\_ISO} + \beta_2 (x1\texttt{\_ISO})^2 + \\ &+ \beta_3 x2\texttt{\_FRatio} + \beta_4 (x2\texttt{\_FRatio})^2 + \beta_5 x3\texttt{\_TIME} + \beta_6 (x3\texttt{\_TIME})^2 + \beta_7 x5\texttt{\_CROP} + \varepsilon. \end{aligned}$$



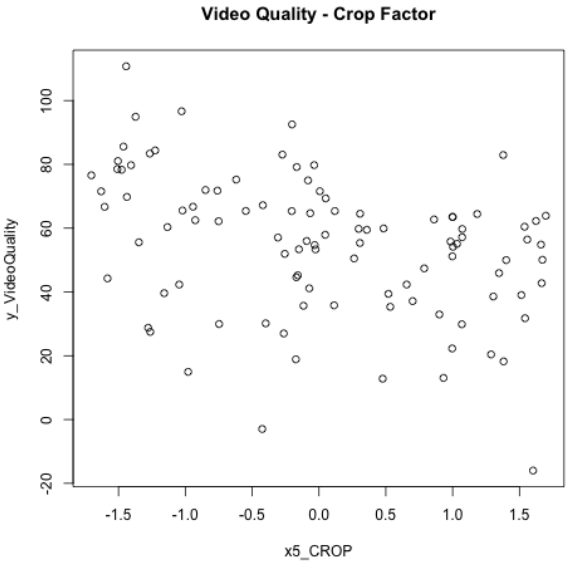
Concave-down quadratic trend:
The quality increases up to a moderate ISO and then decreases rapidly.



Marked quadratic relationship, concave-downward: extreme values of F-Ratio are associated with sharp decreases in quality.



Concave-upward quadratic trend:
Quality improves for very short
or very long times



Clear decreasing linear correlation:
Quality decreases,
as the crop factor increases.

Figure 22: Scatter Plot dei regressori critici.

About Quadratic Terms

The formulation of reduced alternative models, i.e., those multiple regression models with fewer variables than those provided by the dataset, such as *Alternative Model 1*, does not pose

particular issues, unless one evaluates the variations in accuracy and the statistical significance of the associated parameters. A different matter is the formulation of models with polynomial terms, such as *Alternative Model 2*, where nothing is known about potential collinearity generated by the squared predictors. In this regard, once a temporary dataset including the quadratic components, called `dataModel2`, is generated, the renewed correlation matrix is constructed with associated p-values, and the correlations are analyzed through scatterplots.

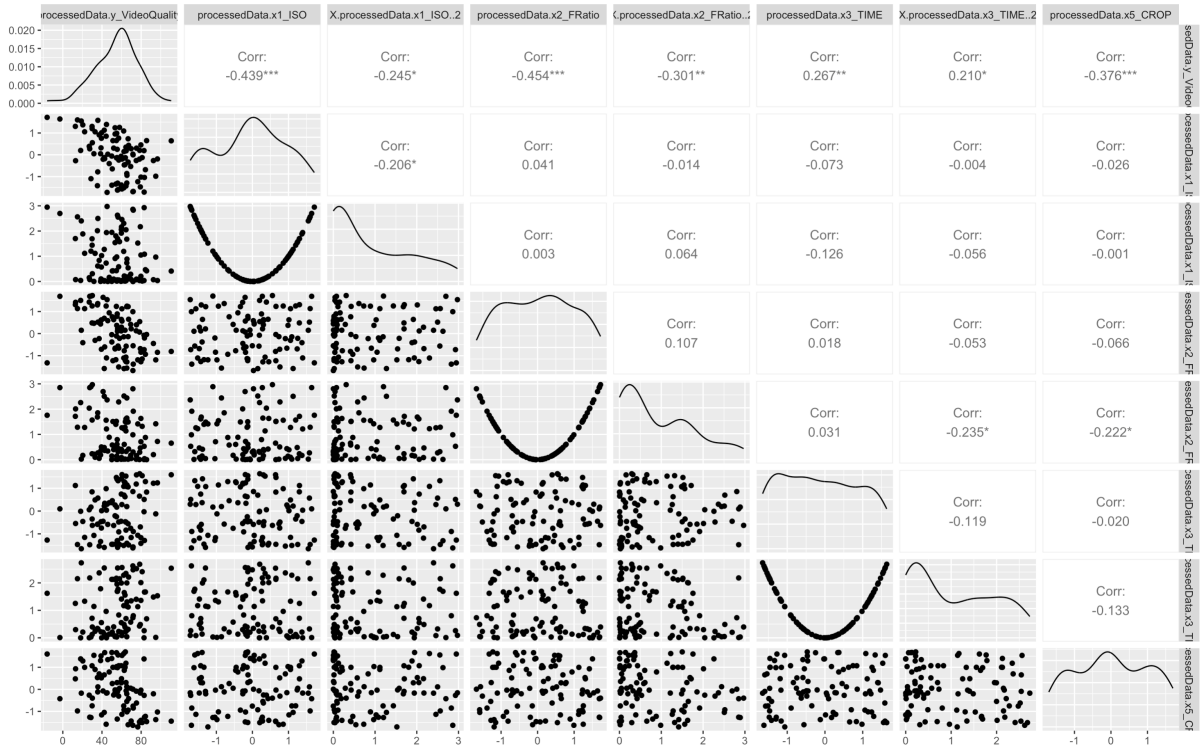


Figure 23: Scatter plot for evaluating collinearity/multicollinearity among the predictors for *Alternative Model 2*.

Except for the obvious quadratic trends that can be observed in the columns related to the regressors derived as the square of the original ones, the scatter plot does not reveal any particularly influential or concerning collinearity. What emerges instead in the column of variables $(\texttt{x1\_ISO})^2$, $(\texttt{x2\_FRatio})^2$, and $(\texttt{x3\_TIME})^2$, i.e., the concentration of observations at zero, is related to the samples, already centered around zero. The analysis can therefore proceed with the characterization of the candidate models.

4.2 A Formal and Graphical Diagnostic

The analysis of the models, implemented in the files `"./src/regressionModel.R"` and `"./src/decision.R"`, is based on the evaluation of the p-values obtained from the t-tests on the estimated parameters, the measurement of gains in estimation accuracy, the quantification of explained variability, and the analysis of the residual distribution. Indeed, due to the outliers shown by the response variable in data exploration, it is necessary to assess the behavior and calculate the standard deviation (MSE) of the residuals, in order to verify the reduced influence of the polarized values and confirm the decision to keep the original sample unchanged.
It is reminded that what follows (Figures 24-26) is directly reported in the files `"./results/models/`**ModelName**`.pdf"` and in the file `"./results/Multiple_Regression.txt"` for the `summary` of each model.

4.2.1 Complete Model

The main model, as reported in the above-referenced text file, and as can be seen by observing the first of the two plots in Figure 24, provides a moderate quality in estimating the mean of the independent variable. With a coefficient of determination R^2:

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

for the sixth regressor, the model is able to explain about 60% of the total variability. However, the true issue with the model lies in the significance of the parameters associated with the variables `x4_MP` and `x6_FOCAL`. These two variables, already identified as not significantly correlated with `y_VideoQuality`, have estimated parameters whose p-values in the Student's t-test are 0.5418 and 0.8096, respectively, indicating an unnecessary complexity in the model.

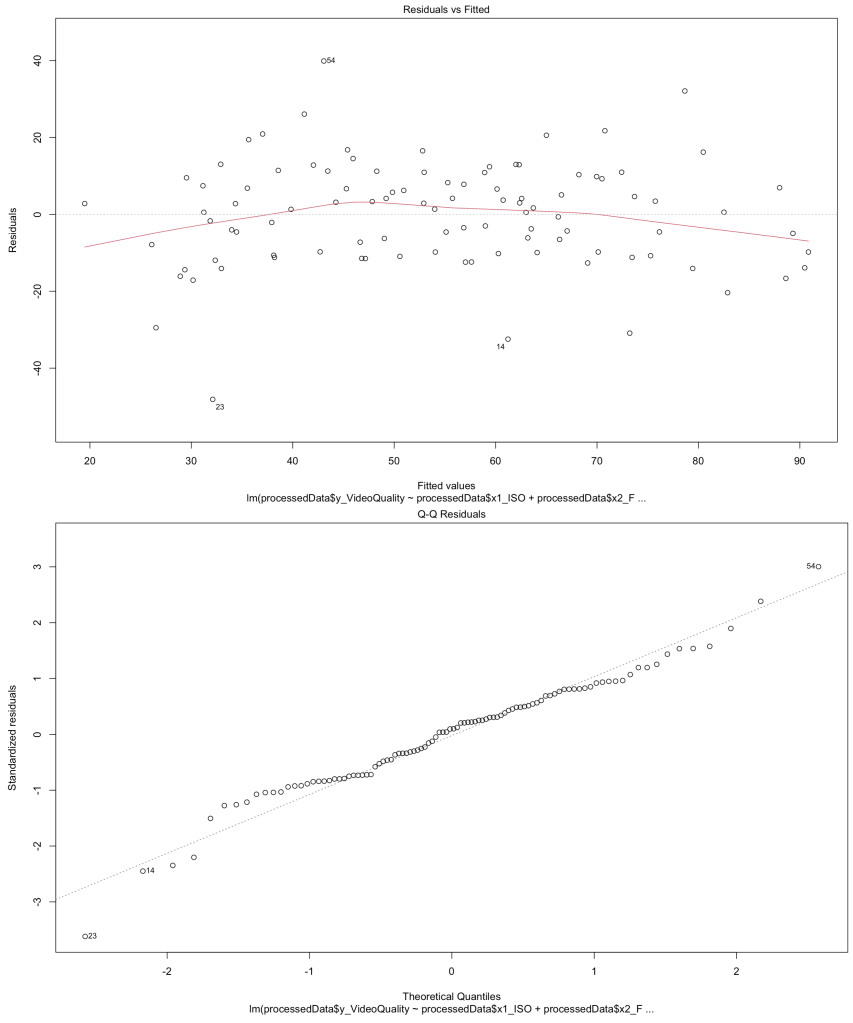


Figure 24: Residual vs Fitted and residuals's Q-Q Plot of *Complete Model*.

The Residual Standard Error is found to be 14.02, which we will see is the highest among all the candidates, indicating a noticeable deviation between the model and the observed data. Finally, as suggested by the Q-Q Plot, the outliers of the dependent variable do not seem to have a significant impact on the models — note the tails — probably due to their low frequency (about 1% of the data are outliers).

4.2.2 Alternative Model 1

In *Alternative Model 1*, the simplification of the predictors, which, according to the initial analysis of the *Complete Model*, is correct, leads to a reduction in the standard error:

$$RSE = 13.91 \Rightarrow \text{Reduction of } 0.79\%,$$

and a substantial maintenance of the model's explanatory power $\left(R^2_{\%} \approx 61\%\right)$.

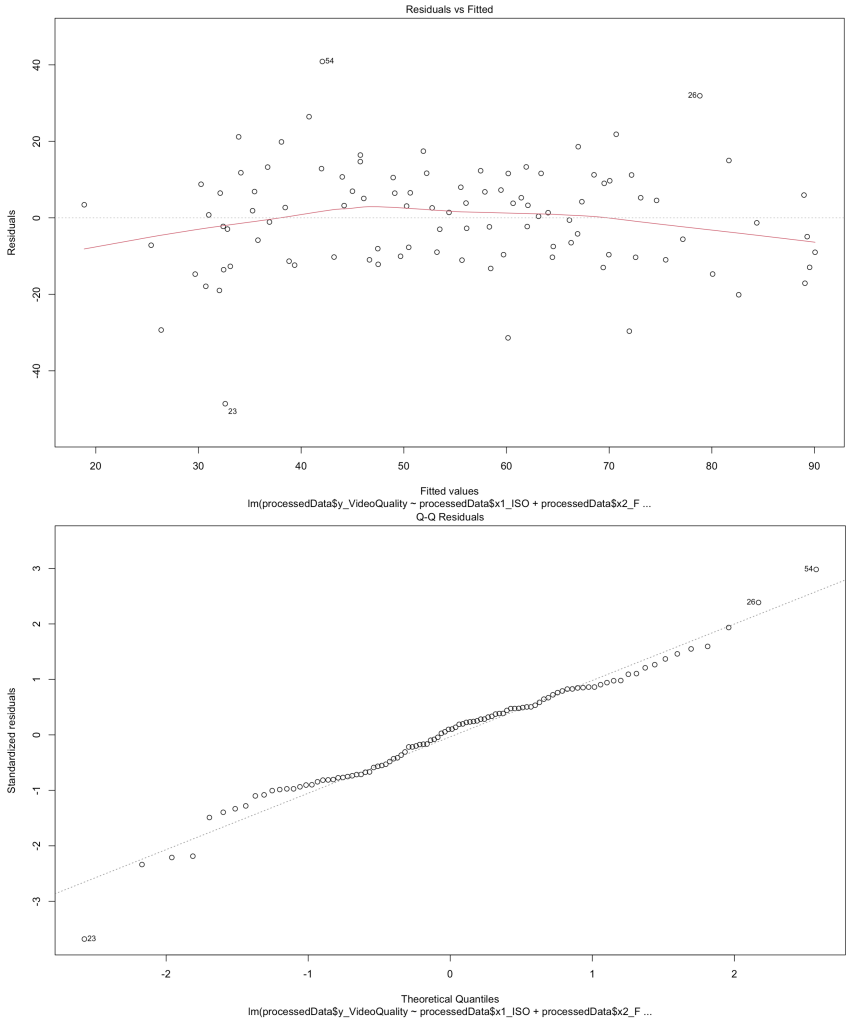


Figure 25: Residual vs Fitted and residuals's Q-Q Plot of *Alternative Model 1*.

All coefficients are highly significant, with the order of magnitude of the p-values for each parameter remaining approximately unchanged ($p \ll 0.001$), confirming the robustness of the selected subset. Regarding the residuals' distribution, even for *Alternative Model 1*, no significant deviations from the normal quantiles are observed in the Q-Q plot (Figure 25). Clearly, the greater compactness of the model, with virtually unchanged performance, represents a clear advantage in terms of parsimony and interpretability. However, the minimal reduction observed in the RSE suggests, despite having fewer parameters, that the model tends to perform better. In this regard, and to this extent, there are high expectations for the second alternative model.

4.2.3 Alternative Model 2

The examination of the candidates concludes with the exploration of *Alternative Model 2*, whose diagnostic plots are shown below.

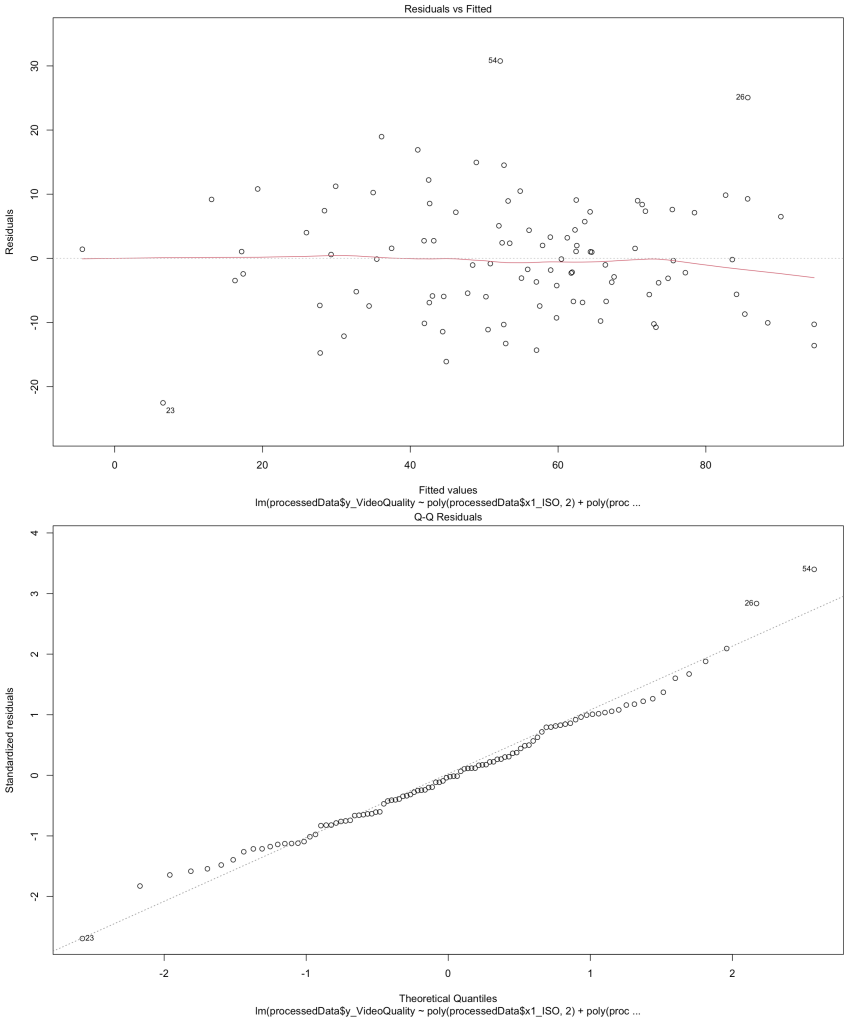


Figure 26: Residual vs Fitted e and residuals's Q-Q Plot of *Alternative Model 2*.

The model, enriched with the explicit representation of the polynomial relationships characteristic of the proposed samples for the main predictors, proves to be the one that most accurately captures the intrinsic complexity of the relationship between shooting variables and perceived quality.
The R^2_{adj} reaches 0.818 and R^2 increases to 0.831, highlighting a gain of over 35% in the proportion of explained variability compared to the maximum model.
The plots (Figure 26) strongly support this choice: the Residuals vs Fitted plot shows clouds of points uniformly distributed around zero, indicating constant variance in the residuals, while the Q-Q plot shows a perfect alignment of the residuals with the theoretical diagonal, confirming the plausibility of the normality assumption. The residual standard error is then reduced by 32.96% compared to `Alternative Model 1`, reaching a value of $RSE = 9.33$, with a resulting reduction in the mean gap between observed and predicted values.
Finally, the p-value from the t-test confirms the inclusion of the estimated parameters in the model, with the acceptance of the alternative hypothesis for all, except for the second-degree term of `x3_TIME`.

The estimate of the parameter β_6 is in fact equal to:

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

with a p-value:

$$p = P[|t| > |t_n| \,|H_0] = 0.296 > \alpha = 0.05.$$

This forces the exclusion of the parameter from the model, suggesting that the increase in complexity with the quadratic term on the variable `x3_TIME` does not provide any benefit. The exclusion of the discussed parameter is reserved for the phase of applying parameter selection criteria and model comparison.
It remains clear that the nonlinear transformation improves the overall explanatory power of the model, while still requiring a critical evaluation of the actual necessity of each term.

5 Model Comparison and Validation

It is essential, after exploring and testing each model with the initial diagnostic analyses, to move on to the decisive phase: the direct comparison between the candidate models.
The goal of the model comparison and validation phase is, under the guidance of the script `".src/decision.R"`, to apply the selection criteria discussed in the following section, and to identify the model that best combines predictive ability, interpretative simplicity, and statistical robustness.

5.1 Comparison Strategies and Evaluative Procedures

Consider studying a phenomenon y based on a set of p predictive variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ with the goal of defining a linear mathematical model, expressed in the form:

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

In this context, the identification of the optimal model can be carried out through appropriate selection criteria applied within the stepwise regression framework, performing hypothesis tests on the mean of the coefficient estimators β_i.
Regarding the choices for the Stepwise analysis criterion, statistical theory primarily includes three approaches: *Forward*, *Backward*, and *Hybrid* (or bidirectional). These methods apply to the classic multiple linear model:

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

which, in matrix form, is:

$$\underset{\_}{y} = \underset{\_}{\underset{\_}{X}}\,\underset{\_}{\beta} + \underset{\_}{\varepsilon},$$

where:

$\underset{\_}{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ is the vector of observations $(n \times 1)$,

$$\underset{-}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \text{ is the vector of parameters } ((p+1)\times 1),$$

$$\underset{-}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ is the vector of random errors } (n\times 1),$$

$$\underset{-}{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \text{ is the design matrix, of dimension } n\times(p+1).$$

It is important to consider that, for models like *Alternative Model 2*, transformations on the predictors $x_{ik} = x_{i2}^2$ can be incorporated into the design matrix to model nonlinear relationships parametrically.

Keeping in mind the three proposed models, the Stepwise regression procedure was adopted in its backward formulation. Starting from the full set of all potentially relevant predictors, and proceeding through the iterative elimination of variables that provide a negligible predictive contribution, it is possible to construct the minimal optimal models.
This means that a backward regression allows for the identification of interactions between predictors that may not emerge in more restrictive approaches, or may be identifiable but with computationally expensive methods such as BSS.
While backward selection allows for the evaluation of individual models, the comparison between them, once reduced, is based on qualitative and quantitative indices different from those previously encountered. It is therefore appropriate to consider information criteria, in addition to the significance criteria and the improvements in estimation accuracy (i.e. R^2), that can globally evaluate the quality and parsimony of the model.
For a generic regression model, let k be the number of estimated parameters and $\hat{L}$ the maximum-likelihood function of the model. The Akaike Information Criterion (AIC) is defined as:

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

and the Bayesian Information Criterion (BIC) is defined as:

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

The two criteria described above are the main informational criteria for measuring the quality of a statistical model, differing primarily in the severity with which they penalize the number of parameters and, consequently, the complexity of the model. Both criteria penalize overly specific models to avoid selecting models with overfitting issues on the training set, and in both cases, the model with the lower value is selected.
Clearly, the logarithmic behavior of the BIC as the number of parameters increases tends to more strongly penalize models that, with the AIC, would have better quality levels. However, this makes the two information criteria suitable for different contexts: while Akaike is more appropriate for models aimed at prediction, the Bayesian criterion is better suited when the goal is simply to model the data. For this reason, the latter will be chosen as the main criterion

for discriminating between the reduced candidates.

Finally, also with respect to the qualitative indices for selecting the best model, the adjusted coefficient of determination R^2_{adj} is considered:

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

This penalizes the inclusion of irrelevant explanatory variables, favoring more robust and parsimonious models. Unlike the classic R^2, which tends to increase with the addition of new predictors, R^2_{adj} only increases if the new variable genuinely improves the model's fit. The selection of the best model will therefore be carried out by comparing the residual distributions through histograms and Shapiro-Wilk tests, the described qualitative indices, and, where appropriate, the application of the F-test to the last regressor.

5.2 Backward Selection Outputs

Complete Model As already observed during the formulation of the candidate regression models, the *Complete Model* has ample room for improvement in terms of parsimony, as it is populated by regressors that are actually not very useful for estimating the expected value of the independent variable. Therefore, using backward selection, implemented with the R command `step(ModelName, direction = "backward")`, we aim to derive a model that is at least equivalent to *Alternative Model 1*.
In fact, let H_0 and H_A be the hypotheses of the t-test such that:

$$H_0 : \beta_i = 0 \qquad \text{and} \qquad H_A : \beta_i \neq 0,$$

then each regressor whose coefficient is not significantly different from zero is sequentially eliminated, resulting in:

$$\hat{y_x} = 55.493 - 9.461\,(x1\_\texttt{ISO}) - 10.572\,(x2\_\texttt{FRatio}) + 5.042\,(x3\_\texttt{TIME}) - 8.893\,(x5\_\texttt{CROP}),$$

with the rejection of the alternative hypothesis for the regressors `x6_FOCAL` and `x4_MP`, and acceptance for the remaining ones.
By consulting the file `"./results/Multiple_Regression.txt"`, it is easy to verify the complete overlap between the linear functional just described and the one associated with *Alternative Model 1*. What has been described thus has a dual interpretation: the results obtained, including the values of R^2, *RSE*, and F-Statistic, which will be discussed for each model in Section 5.3, show not only an enhancement in parsimony and performance in terms of improved estimation accuracy, but also, and most importantly, how the first alternative is the minimal non-polynomial model capable of explaining the video quality trend with statistical significance.
For this reason, and to avoid excessive redundancy, the presentation of the results from the backward selection on *Alternative Model 1* is omitted.

Alternative Model 2 By acting on the second alternative model as was done with the maximum model, due to the previously discussed p-value for the estimate of the parameter β_6 on the second-degree term associated with the variable `x3_TIME`, one would expect to obtain a new reduced model. In reality, due to the implementation of the method `lm(Alternative Model 2, direction = "backward", ...)`, the output in the RStudio framework that this procedure provides is as follows:

```
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
```

Clearly, this is a bug, the cause of which lies in the algorithm provided by the RStudio software, and it can be resolved by explicitly constructing the correct model from the backward selection. Referred to as *Alternative Model 3*, the new reduced model presents the coefficients $\hat{\beta}_i$:

```
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 ***
---
```

The reduced form of `Alternative Model 2`, i.e., the `Alternative Model 3`, returns the estimate of the mean $E[Y|X = x]$, $\hat{y_x}$:

$$\hat{y_x} = 54.6649 - 92.4341\,(x1\_\texttt{ISO}) - 63.0276\,(x1\_\texttt{ISO})^2 - 102.1467\,(x2\_\texttt{FRatio})$$
$$-77.817\,(x2\_\texttt{FRatio})^2 + 44.1998\,(x3\_\texttt{TIME}) - 107.5113\,(x5\_\texttt{CROP})\,.$$

In this case, the backward procedure has proven effective in producing a statistically solid and parsimonious model, reducing the risk of overfitting. Its progressive application – first on a complete linear model, then on a polynomial specification – has allowed for a gradual refinement of the predictive structure. Final evaluations regarding the performance of the models are deferred to the next section, although, based on what has already been presented regarding the added polynomial terms, the model just presented seems to be the most promising.

5.3 Model Comparison and Final Choice

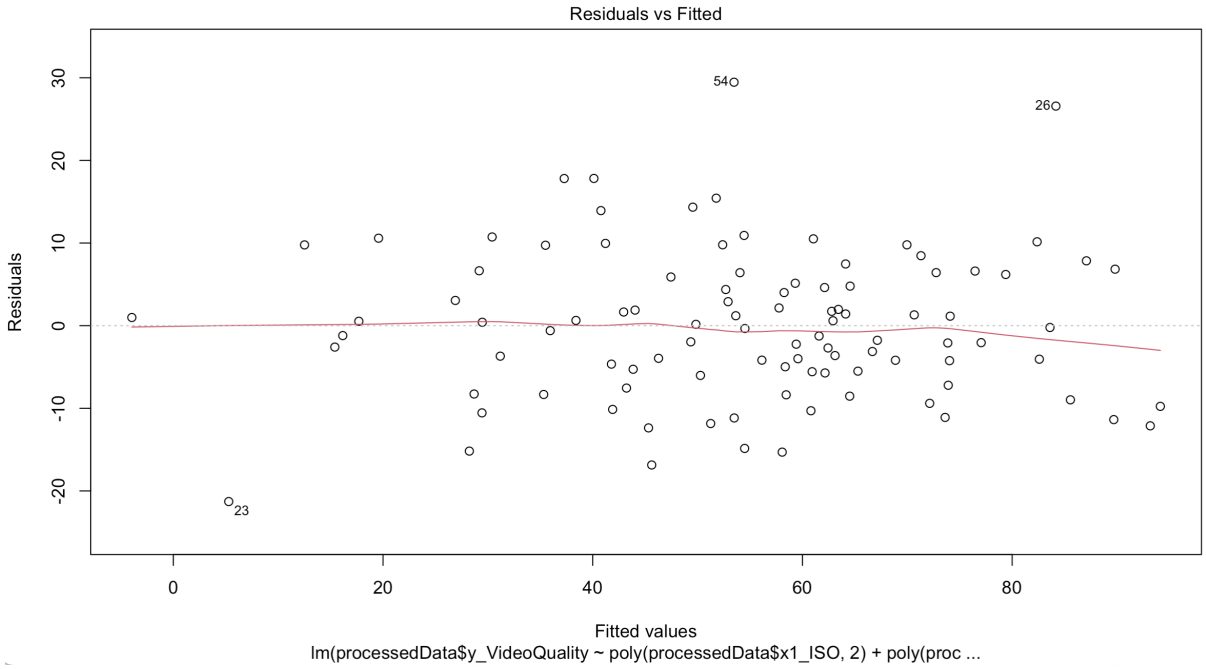
We now proceed with the formal comparison; to identify the optimal model among those obtained, evaluated according to the criteria in Section 5.1, the following summary table is presented:

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

It is evident that the models obtained as reductions – namely, `Alternative Model 1` and `Alternative Model 3` – are not only more economical but also more performant in terms of the proportion of explained variability. Furthermore, as expected, the third alternative model proves to be the best in terms of the performance/cost ratio: with one less quadratic term, it is possible to maintain virtually unchanged the standard error of the residuals and R^2.
By analyzing Figure 27 (on the following page), one can observe a good cloud of residuals $e_i = y_i - \hat{y_i}$, indicating a rather constant variance $\forall\ i$. However, it is important to evaluate how the residuals are distributed not only graphically, in order to verify their normality, due to the inconsistencies in the tails shown by the histogram. By performing a Shapiro-Wilk test, a p-value of approximately 0.216 is obtained, thereby rejecting the alternative hypothesis and dispelling any doubts about the confirmed normality of the residuals.
The final comparison between the models, in order to make the final choice, involves the evaluation of the BIC, whose respective values are reported in the table below.

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

The script `"./src/decision.R"` also includes, for completeness, the evaluation of the AIC associated with the models, from which it emerges that *Alternative Model 3* is indeed the best among the candidates, not only for the fitting to the training set but also for prediction on a potential test set, as it has the lowest values for both the Akaike and Bayesian Information Criteria.



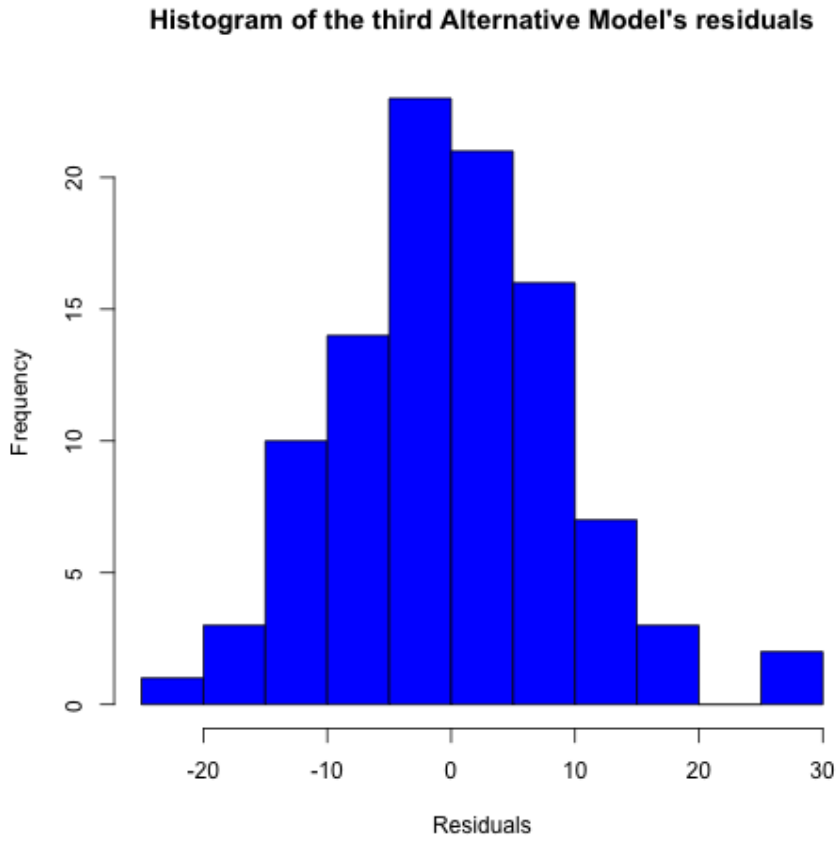


Figure 27: Diagnostica grafica del terzo modello alternativo.

With confidence intervals on the parameter estimates at the 2.5th and 97.5th percentiles such that:

```
Confidence Intervals:
                                            2.5 %    97.5 %
(Intercept)                              52.81050  56.51939
poly(processedData$x1_ISO, 2)1         -111.05703 -73.81123
poly(processedData$x1_ISO, 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
```

and, with SQE and $MSQE$ such that:

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

$$MSQE = \frac{SQE}{\nu} \approx 0.8721, \qquad \text{with } \nu \text{ df},$$

The selection of the second reduced alternative, namely *Alternative Model 3*, is decreed as the multiple linear regression model that estimates the mean of the video quality perceived by users as the sensor sensitivity, focal ratio, exposure time, and crop factor vary.
Further insights and key-values, such as F-Statistic, SD of residuals, and a summary of the residuals' distribution, can be found in the file `"./results/models/chosen/Chosen.txt"`.

6 Domain Insights: Relating Statistical Findings to Video Quality

The final selected model, *Alternative Model 3*, in addition to representing the best solution in terms of statistical performance, also consistently reflects the technical logic that links the independent variables to the perceived video quality (`y_VideoQuality`). This latter is, as a reminder, an index derived from human judgment on visual aspects such as sharpness, noise, dynamic range, color fidelity, depth of field, resolution, and presence of artifacts.
However, before describing the model's properties in the application domain, it is useful to provide an overview highlighting the information carried by the model's regressors.

The predictor `x1_ISO` represents the sensor's sensitivity to light. It is known that medium-low ISO values ensure sharp, noise-free images, while high values, in unsuitable contexts, increase digital noise, compromising sharpness and also impacting color fidelity and dynamic range, as shown in Figure 28. The proposed model proves to be quite accurate in this regard, displaying negative coefficients for both the linear and quadratic terms, thus generating a concave decreasing curve that harshly penalizes observations with either very high or excessively low values.

Figure 28: From left to right: shots with the same lighting conditions, but different ISO values.

The variable `x2_FRatio`, on the other hand, describes the aperture of the diaphragm; if the diaphragm is either too open or too closed, it causes visual problems, such as reduced depth of field or optical aberrations. Therefore, it is not surprising to find negative correlations between this variable and video quality. Here too, `y_VideoQuality` is optimal at intermediate aperture values due to the quadratic term.

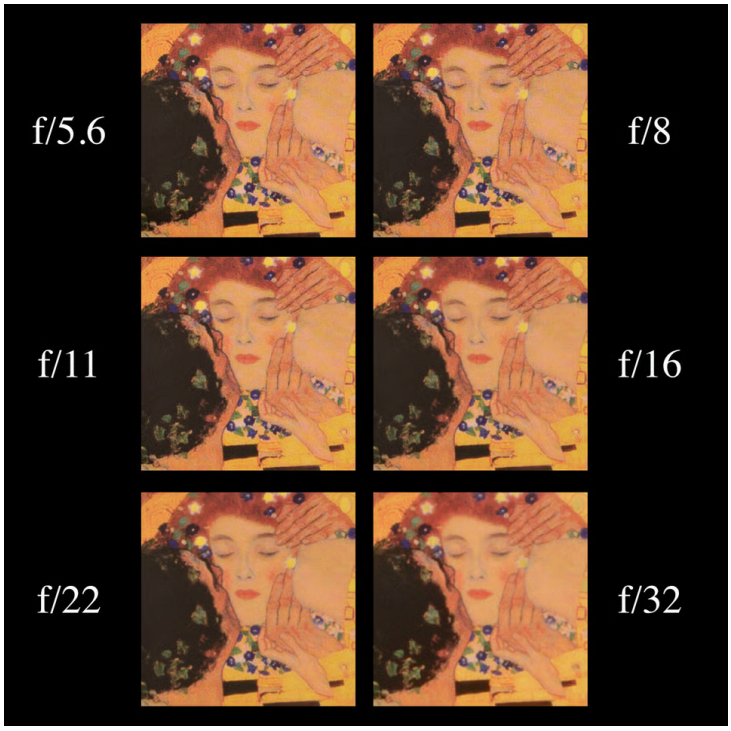


Figure 29: From left to right and top to bottom: shots with different focal ratios, with the respective values shown next to them.

The exposure time, represented by `x3_TIME`, has a positive linear relationship with the dependent variable. A longer exposure allows more light to be captured per frame, reducing underexposure and improving the light rendering. This results in greater sharpness and fewer visual artifacts, contributing to better quality. However, it should be noted that, since we are talking about quality perceived by users, the linear relationship might also be indicative of the impact that longer exposure times (i.e., motion blur) have on the human eye.

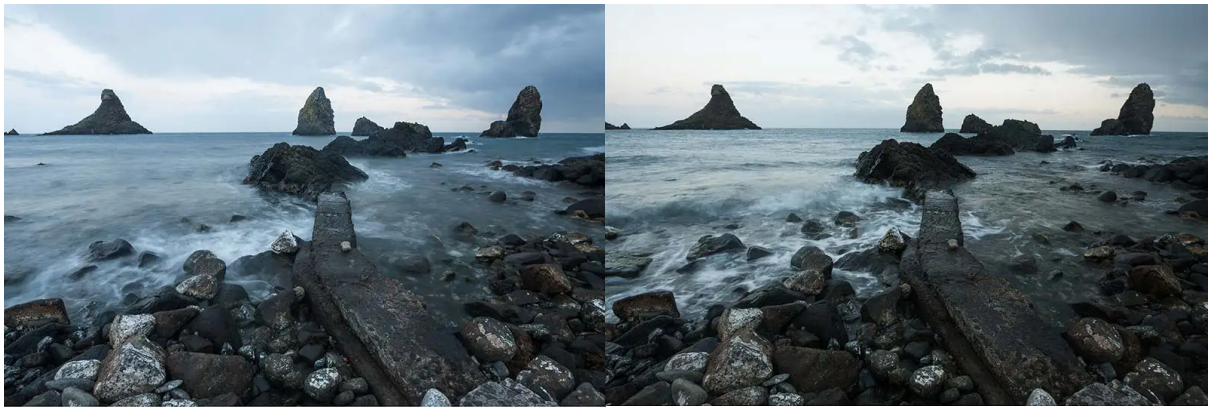


Figure 30: From left to right: shots with different exposure times, respectively equal to $2s$ and $1/2s$.

Finally, the sensor's crop factor (`x5_CROP`) is negatively associated with quality. Higher crop values indicate a smaller sensor or a larger portion of the image being cropped, resulting in a loss of depth of field, field of view, and overall optical performance, accentuating the effects caused by motion. The effect, well represented in the model by a marked negative coefficient, reflects the strong unfavorable impact of the crop on visual perception. It can also be interpreted as a symptom of the perception an observer has when viewing a frame with reduced depth.

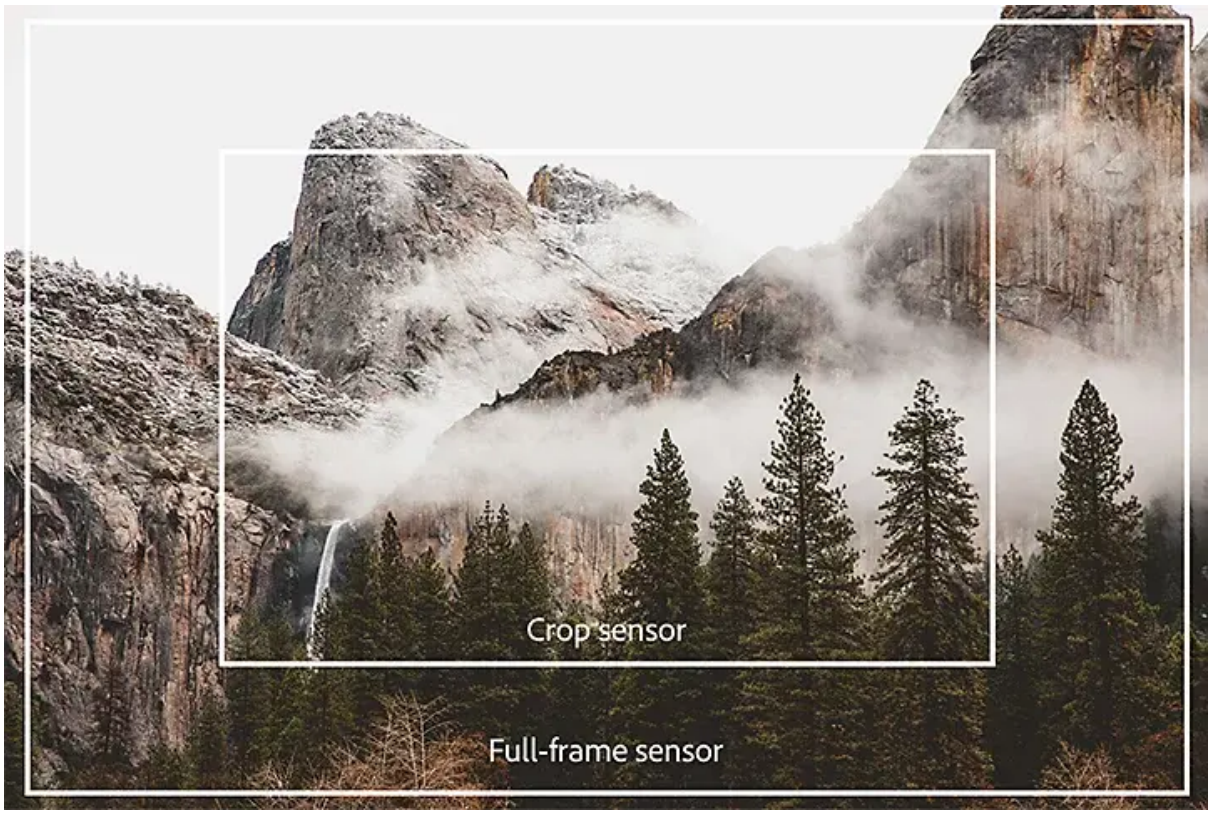


Figure 31: Difference between Full-frame and Crop sensors

Each of the relationships observed between the selected predictors and the response variable seems to have a concrete and logical justification in the physical and perceptual functioning of video capture and photo shooting, confirming the interpretative reliability of the final model. However, it should be kept in mind that the analysis conducted, due to the cardinality of the provided samples and the number of variables, only takes into account those shooting conditions over which an operator might have control, which clearly are not the only factors influencing the final results.
In this regard, a larger dataset, possibly divided into train and test sets, and a greater number of parameters, such as regressors providing information about the shooting time and natural lighting conditions, would have guided the analysis in a different direction, leading to the development of potentially predictive models even more relevant to the application domain.
Therefore, the results obtained, as a whole, provide a solid foundation for future developments, both in terms of extending the model to broader contexts and in deepening interactive effects or factors not observed in the present dataset.