Universidade do Minho
Escola de Ciências

Renato Fernandes dos Santos

Statistical Analysis of variables of tires' specifications and respective tests

outubro de 2019

Statistical Analysis of variables of tires' specifications and respective tests

Renato Santos

UMinho | 2019

Universidade do Minho
Escola de Ciências

Renato Fernandes dos Santos

Statistical Analysis of variables of tires' specifications and respective tests

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação do

Professor Doutor Luís Machado

outubro de 2019

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Agradecimentos

Ao Professor Doutor Luís Machado, pelo seu apoio e partilha de conhecimentos, em particular todas as sugestões e conselhos.

À empresa Continental Mabor, pela oportunidade enriquecedora e de crescimento pessoal e profissional.

Ao Francisco, Murilo e Rafael, por todo o apoio prestado durante o estágio, bem como aos restantes membros do departamento de R&D pela ajuda e prestabilidade.

À Engenheira Marta Malainho, pela orientação, conselhos e todo o apoio dados durante o decorrer do estágio.

Aos meus pais, Teresa e António, e à minha irmã Sabrina, pela ajuda, apoio e compreensão dados ao longo de todo o meu percurso e académico. Uma palavra de gratidão especial para a minha mãe, por todos os esforços que fez e continua a fazer para que tudo isto fosse possível, bem como por todo o amor e apoio incondicionais, e por sempre acreditar em mim e nas minhas capacidades, sem ela nada disto seria possível.

À Ana, pelo apoio constante, paciência, prestabilidade, e acima de tudo pelo seu coração e inabalável confiança que deposita em mim, que me faz ambicionar a cada vez mais.

Aos restantes professores e a todos os meus amigos e familiares, que de alguma forma contribuiram para a conclusão desta etapa.

A todos, o meu mais sincero obrigado.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.
I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Análise Estatística das variáveis das especificações de pneus e respectivos testes

O objectivo deste estudo passou por identificar quais as variáveis mais importantes tendo em conta os aspectos dimensionais de um pneu agricultural, bem como o desenvolvimento de uma ferramenta, usando vários métodos, que permitiriam prever a dimensão final do pneu, bem como entender o impacto no resultado final após realizar alterações nos valores das variáveis

Com esse objectivo em mente foram aplicadas duas diferentes metodologias, a regressão linear e logística.

De forma a realizar previsões sobre o resultado dimensional final do pneu foram desenvolvidos dois modelos distintos, usando regressão linear, um para cada uma de duas variáveis resposta (X_{36} e X_{37}). Este processo também permitiu a compreensão de quais as variáveis apresentam um maior impacto na dimensão final do pneu.

Os resultados foram posteriormente validados utilizando valores que não se encontravam na base de dados inicial, com o objectivo de entender se os modelos obtidos se ajustavam correctamente e se, consequentemente, as simulações obtidas eram aceitáveis. Depois de analisar os resultados foi possível concluir que o erro presente nas simulações estava dentro dos limites aceitáveis.

Por outro lado, a regressão logística foi utilizada de forma a entender, tendo em conta os valores apresentados pelas variáveis, qual seria a possibilidade de um pneu passar ou falhar no teste dimensional.

Por fim, de forma a facilitar o cálculo das simulações previamente mencionadas para novos pneus existiu a necessidade de desenvolver uma ferramenta que permitisse ao utilizador obter esse resultado apenas introduzindo os *inputs* necessários para cada uma das variáveis. Esta ferramenta foi desenvolvida utilizando o *R Shiny*.

Toda a modelação foi realizada utilizando o *software* R versão 3.5.2 (2018-12-20).

Palavras-chave: impacto das variáveis; pneu agrícola; previsões; regressão linear; regressão logística; resultados dimensionais; *R Shiny*

Statistical Analysis of variables of tires' specifications and respective tests

The objective of this study was to identify the most important predictor variables regarding the dimensions of an agricultural tire, as well as the development of a tool, using various methods, that would allow the user to predict the final dimensional results, as well as understand the impact that each variable's value change would have on that same result.

In order to do that linear regression and logistic regression were used.

To make the predictions about the final dimensional results, two different models based on linear regression were developed; one for each of the response variables (X_{36} and X_{37}) regarding the tire's dimensions. This also allowed the understanding of which variables were the most impactful ones on the final result for each of the dimensions.

The results were then validated using values that weren't present in the initial data, with the objective of understanding if the obtained models were correctly adjusted, and if the they were providing good simulations. After analyzing the results it was safe to conclude that the error of the simulations was within an acceptable range.

On the other hand, the logistic regression was applied to understand, given the values of the variables, what was the chance of a tire passing or failing the dimensional test.

Finally, in order to facilitate the calculation of the previously mentioned simulations for new tires, there was a need to develop a tool that would allow the user to obtain results in a friendly environment. This tool was developed in the R Shiny environment.

All the modeling was made using the software R, version 3.5.2. (2018-12-20).

Keywords: agricultural tire; dimensional results; linear regression; logistic regression; predictions; R Shiny; variable impact

Contents

List of Figures

List of Tables

List of Acronyms

SUV	Sport Utility Vehicle
SSR	Sum of Squared Residuals
SSE	Sum of Squared Estimate Errors
ANOVA	Analysis of Variance
SST	Total Sum of Squares
MSR	Mean Squared Regression
MSE	Mean Squared Error
MST	Mean Squared Total
LRT	Likelihood Ratio Test
AIC	Akaike's Information Criterion
VIF	Variance Inflation Factor
GLM	Generalized Linear Model
MLE	Maximum Likelihood Estimation
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
OLS	Ordinary Least Squares
ETO	Experimental Test Order

1 Introduction

With the constant advance of the tire industry, it has become normal practice to resort to mathematical and statistical approaches in order to aid on the development and advance of this industry.

With this in mind, the present dissertation was realized during an internship on the company Continental Mabor, more specifically on the Research & Development department. This specific department is responsible for the development of new sizes for agricultural tires, therefore those are the type of tires addressed in this manuscript.

In broad terms, the problem addressed on the following chapters consists on the study and understanding of the variables regarding the mold and specifications of a tire, and, upon having obtained that level of clarification, the problem evolved into finding correlations between them, as well as the creation of predictive models for two response variables, regarding the final size of the tires.

This problem arose after the dimensional results obtained for the tires weren't according to expectations, therefore there was a need to develop a deeper study in order to understand where the problems lied, and how they could be fixed.

Upon an extensive analysis of the data it was clear that the best approach, given its size and behaviour, would be to apply linear regression, and obtain two separate models, one for each of the response variables.

This dissertation is divided into four main chapters, to be presented next.

The first chapter consists on a brief presentation of the company where the internship was realized, Continental Mabor, as well as the type of tires under study, going into some detail on their technical features, construction and production processes. Regarding the construction there two main types, the radial and the bias/cross-ply. When it comes to the tyres addressed on this study they all used the radial construction. The components of said tires are then explained in detail on the chapter previously mentioned. When it comes to the production processes of a tire they can be divided into five stages: mixing, preparation, construction, vulcanization and final inspection.

The next chapter consists on a theoretical presentation of all the topics covered throughout the dissertation, being divided into two major parts, the linear regression and the generalized linear models. This chapter addresses all the different topics of the linear regression, namely, variable selection, multicollinearity, goodness-of-fit, parameter estimation, statistical inference and residual analysis, among others. For the generalized linear models the topics covered follow the same line of thought of the linear regression,

with some additions regarding the variable selection methods, and the tests to assess the quality of adjustment of the models obtained.

The ensuing chapter is divided into three major parts. The first part consists on the presentation of the data, as well as an exploratory analysis in order to understand its behaviour and characteristics. Some of the results obtained on this section are the correlation matrix, as well as location measures for the quantitative variables.

The second part refers to the analysis and modeling of both the response variables, being presented the models obtained, and analyzed its quality of adjustment, as well as the residual analysis for both of them, in order to understand if they were acceptable.

The third part consists on the presentation of the binary variable to be used to obtain the generalized linear regression model, that is nothing more than whether a tire passed or failed the dimensional tests. Once again, on this part of the chapter it's presented the variable under study, with some analysis and graphics, and afterwards it's obtained and presented the GLM model, where, once more, it's performed a residual analysis and it's tested the goodness-of-fit, through multiple tests and statistics, in order to obtain the best possible result.

Lastly, the last chapter refers to the application developed in order to allow the users to make predictions for the dimensional results of a tire, preventing the construction of unnecessary tires, for example. This application was developed on *R Shiny* environment, and, in broad terms, it allows the user to input the values for the mold and specification variables and predict the final dimensional results of the tire that is being developed, this by utilizing the linear regression models previously obtained.

This dissertation ends with a chapter dedicated to conclusions and future research.

2 Company

The present dissertation was developed during an internship on the company Continental Mabor – Indústria de Pneus S.A., whose facilities are located in Lousado, Portugal.

The following chapter regards a brief presentation of the company as well as a summary of the type of tires which were target of this study.

2.1 Continental AG

Continental AG was founded in Hannover (Germany) in October of 1871. On its prime days the manufacture consisted mainly of flexible rubber and solid tires for carriages and bicycles.

In 1898, initiated the production of flat tires (without tread drawing) for automobiles. Since then the company has been keeping up with the evolution of the automobile industry with the study and application of techniques, products and equipment in order to improve their tires. Its prestige goes beyond German borders and the Continental tires start to be used by winning cars of several racing competitions.

In 2007, the company acquires Siemens VDO Automotive AG and leaps forward to be one of the five major world suppliers of the automobile industry, and at the same time establishes its position in Europe, North America and Asia.

The Continental Group is specialist in the production of braking systems, dynamic controls for vehicles, potency transmission technologies, electronic systems and sensors. In addition to its operations connected to the automobile sector they also work towards machinery manufacturing, for the mining and printing industry.

Currently the company is located in 56 different countries, with 427 distinct locations.

2.2 Continental Mabor

Continental Mabor was created in December of 1989 as a company in the tire industry. Its name is the result of the junction of two major companies of rubber manufacture, Mabor, on a national level, and Continental AG, known worldwide.

Mabor – Manufactura Nacional de Borracha, S.A., was the first tire factory in Portugal. In July of 1990 began the restructuring program that turned the old Mabor facilities into the most modern ones of the Continental group, at the time. Averaging a daily production

of 5000 tires in 1990, the company quadrupled its daily production in a matter of six years, to 21 000 tires a day.

Presently the company's production is very varied when it comes to sizes, types and brands. Continental Mabor comprises in its portfolio tires destined to SUV's (Sport Utility Vehicles), high performance tires, ContiSeal tires and ContiSilent tires. Its range of manufacture includes tires of rim 14" to 22" and currently averages a daily production of 56 000 tires. Over 98% of the production is destined for exportation.

The currently called "replacement market" absorbs roughly 60% of the annual production of the company. The remaining share is distributed along the assembly lines of the most prestigious builders of the automobile industry.

2.3 Research & Development Department

The internship on which this manuscript was based was inserted in the Research & Development Department. This department is responsible for the "creation" of new agricultural tires according to certain requirements previously set.

2.3.1 Tire Construction

There are two main types of construction, the bias construction and the radial construction, both represented in the following figure.

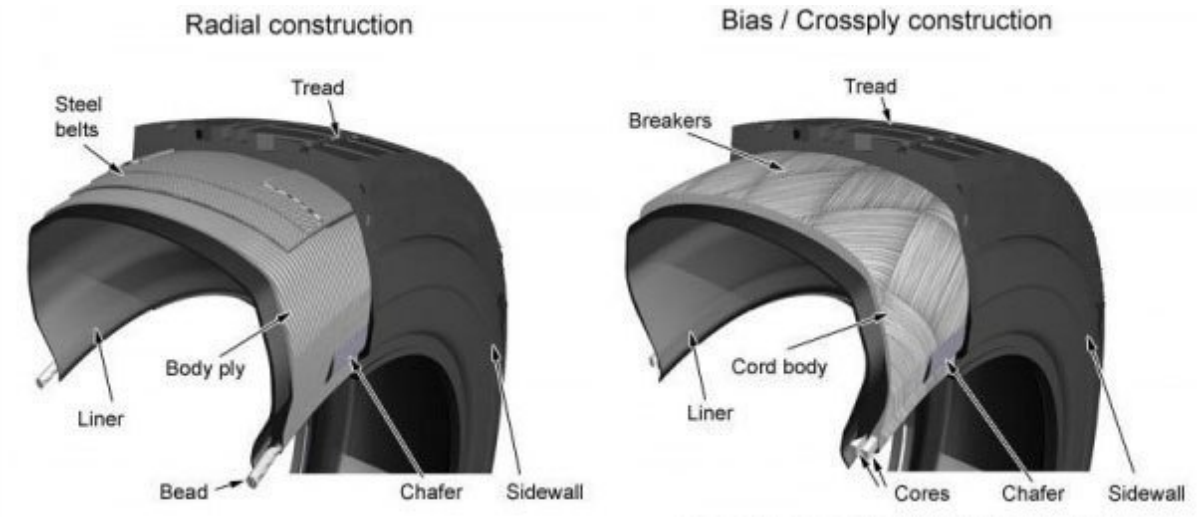


Figure 2.1: Radial construction vs bias construction.

The main differences between these two types of construction are that the radials' tires steel belts dissipate heat better, as well as the capacity of the tire to flex more, due to the fact of the radial tires having fewer layers of body cords on the sidewall.

The one mainly used and considered from now on is the radial construction, given that is the one that offers better overall results when it comes to the quality of the tire.

Tire composition

The composition of radial tires is displayed in the following figure, where the image on the left represents a transverse cut of the tire, and the one on the right a cross section cut in order to understand what are the components of a tire.

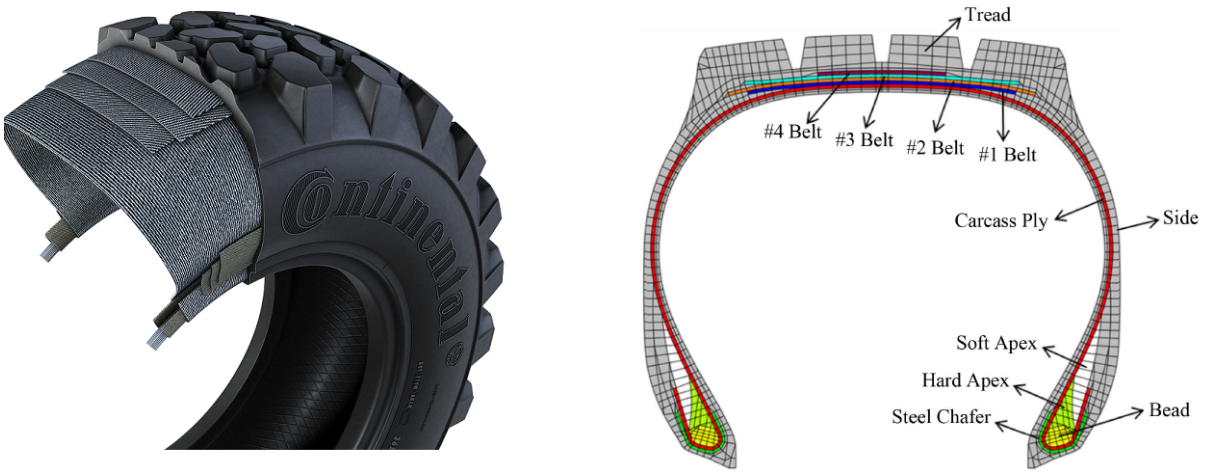


Figure 2.2: Left: transverse cut of a tire; Right: cross section cut.

As shown on the previous figure a tire is composed by several different components, each of them with its own importance, explained below.

- **Bead**

  The bead is where the tire adjusts to the rim. It's composed by a steel filament inextensible and with different shapes and proportions according to the size of the tire.

  Its main functions are:

  (i) Secure the tire to the rim;

  (ii) Provide the watertightness of the tire.

- **Carcass ply**

  The carcass is a flexible structure formed by textile or steel filaments, forming arches that wrap around the bead.

  Over the carcass are applied belts and layers of rubber that together will form the complete tire.

  Some of the most important functions of the carcass are:

  (i) Supporting the load and speed with the help of pressure;

  (ii) Plays a role in the comfort and stability;

  (iii) Plays a role in the energetic efficiency and yield of the tire.

- **Belts**

  The belts are made up by metallic filaments coated with rubber. They are placed above the carcass forming an area that guarantees the mechanical resistance of the tire to the velocity and the centrifugal force.

  The belts cross obliquely and are placed on top of each other. Its cross with the filaments of the carcass forms indeformable triangles, that guarantee the stiffness of the top.

  The role of these components is very complex:

  (i) In the circumferential sense of the tire they have to be rigid so that they don't extend under the effect of the forces when the tire is rotating, and to perfectly control the diameter of the tire, regardless of the conditions of utilization;

  (ii) Allow the tire to be flexible but not elastic.

- **Sidewall**

  Area that is situated between the tread/shoulder and the bead and that provides the tire with side stability. This can be considered as the height of the tire, and is also where the nomenclature of the tire can be found.

  Its main roles are:

  (i) Bear the load;

  (ii) Support constant mechanical flexion;

  (iii) Resistance to frictions and aggression;

  (iv) Plays a role in the stability and comfort.

Phases of the construction of a tire

The construction of a tire is divided in two different main phases: coming up with the appropriate mold and specifications for the tire, and the secondly, the physical construction of the tire (that includes several different phases itself).

The first phase, and perhaps the most crucial one, is the drawing of the mold. This, in conjunction with the specifications used for the tire is what's going to determine its dimensions, therefore it's extremely important to obtain the best possible mold, keeping in mind that after the mold is complete and ordered it's impossible to make any changes to its dimensions.

After that, the next step is to put together the best possible specification in order to obtain a good overall tire (dimension wise, resistance, etc), keeping in mind various factors like weight, material used, quantity and, evidently, costs. These specifications vary from the material used for the belts and plies, its quantity, the size of the bead, and many other factors.

When everything is agreed upon and the mold and specifications are ready, the next step is the physical construction of the tire. This process can also be divided into several others, explained below.

The process of construction of a tire (physical construction) is divided into five stages: mixing, preparation, construction, vulcanization and final inspection.

The first phase, the mixing, is where the natural and synthetic rubbers are combined with various chemical products (pigments, mineral oil, silica, among others) in order to create the necessary materials used in the fabrication of the tire.

The preparation stage, composed by the cold and hot preparation, is where the material needed for the specification of the tire is created. The hot preparation is responsible for the fabric of the sidewalls and treads, while the cold preparation is responsible for textile and metallic plies.

Next, on the construction stage, is where the components previously created are assembled together, giving origin to a tire denominated by “green tire”.

On the stage of the vulcanization the tires are placed in a diaphragm (that contains steam at 170ºC and at a pressure of 6 bar) where it’s applied a segmented mold. The tire is submitted to this process for some time, gaining the desired shape.

Lastly, on the last stage, the final inspection, the tire is submitted to an extensive inspection in order to guarantee its quality and safety, and to verify if it obliges to the set requirements.

After all these processes the tires are submitted to various tests (dimensional, resistance, etc) in order to conclude that the tire is ready to be produced and sold, according to certain parameters established by the law.

3 Linear Models

3.1 Linear Regression

Linear regression first appeared in the early 1800's, and its earliest form was the method of least squares, which was published by Legendre in 1805 [1], and later by Gauss in 1809 [2]. The term "regression" was coined by Francis Galton, in England, on a scientific article regarding the existence of a linear relation between the diameter of pea beans and the diameter of the descendant grains [3].

The main objective of linear regression is to study the relationship between a dependent variable, Y, that represents the output or outcome whose variation is being studied, and the remaining p independent variables, $X_1, X_2, ..., X_p$. Lastly, the main goal is to be able to explain the phenomenon which is being studied and predict its outcome, through the following mathematical formula,

$$Y = X\beta + \epsilon \tag{3.1}$$

where Y is a vector with dimension $n \times 1$, X represents the design matrix of the model, with dimension $n \times (p+1)$, associated to a vector β, of dimension $(p+1) \times 1$, that represents the regression parameters, and lastly, ϵ represents the random errors vector, with dimension $n \times 1$, and with a Gaussian distribution. With that being said, Y also follows a Gaussian distribution and, therefore, ϵ and Y are random variables that verify the following conditions:

- $E[Y] = X\beta$;
- $E[\epsilon_i, \epsilon_j] = 0$, if $i \neq j$;
- $Var[Y] = Var[\epsilon] = \sigma^2 I_n$, where I_n represents the identity matrix of order n;
- X is deterministic, measured without error.

Therefore, given a random sample, $(x_{11}, x_{12}, \ldots, x_{1p}, y_1), \ldots, (x_{n1}, x_{n2}, \ldots, x_{np}, y_n)$, with n independent observations, where x_{ij} and y_i represent, accordingly, the values of the variables X_j and Y for the individual i, we obtain the following:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, i = 1, ..., n \tag{3.2}$$

where

- y_i: i-th observation of the dependent variable;
- x_{ij}: i-th observation of the x_j covariable;
- β_0: expected value of Y when the independent variables are zero;
- β_j: average increase of Y with each unit of x_j, and maintaining all the other variables constant, in case this variable is quantitative. Alternatively, if the variable is qualitative, this represents the average increase of Y for each category of x_i, when compared with the reference category;
- ϵ_i: random error associated to the response of the subject i.

3.1.1 Assumptions of Linear Regression

Linear regression has a set of assumptions that need to be verified so that the model can be considered valid.

The first one states that the expected value of the errors is equal to zero, $E[\epsilon_i] = 0$.

Secondly the variance of the errors needs to be constant (homoscedasticity), in other words, that the residuals are equal across the regression line, $Var[\epsilon_i] = \sigma^2$.

The third assumption falls upon the need for the residuals to be independent, meaning that there should be little or none autocorrelation in the data.

Another condition is the normality of residuals, meaning that the residuals should follow a normal distribution with mean zero and variance σ^2, $\epsilon_i \sim N(0, \sigma^2)$.

Lastly, there should be as little correlation between the independent variables as possible, avoiding the issue of multicollinearity that will be explained further along the dissertation.

The verification of all these conditions represents the ideal situation for the adequacy of the fitted model. However, in some cases, one or more assumptions can be disregarded, depending on the data that is subject of study.

3.1.2 Parameter Estimation

Through a sample of n observation of the variables Y and $\vec{X}$, it's possible to estimate, using the least squares method, the regression coefficients, $\beta_0, \beta_1, \ldots, \beta_p$.

The least squares method is based upon the determination of the values for all the $\vec{\beta}$ coefficients that minimize the sum of the square of the deviation between the observed values (Y_i) and the fitted values ($\hat{Y}_i$) of the regression function. In other words, this method aims to minimize the residual sum of squares (e_i), as given by,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_p X_{pi}))^2 \tag{3.3}$$

or in matrix notation,

$$e^T e = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X \hat{\beta} \tag{3.4}$$

where $e^T e$ is the residual sum of squares, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ represents the vector of the fitted regression coefficients and $X\hat{\beta} = (\hat{Y_1}, \hat{Y_2}, ..., \hat{Y_n})$ represents the vector of fitted values.

Consequently, deriving the expression to $\hat{\beta}$ and making it equal to zero, it's possible to obtain the estimates of the parameters, on its matrix form

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{3.5}$$

Therefore, the variance and covariance matrix of the estimates of the least squares is given by

$$Cov[\hat{\beta}] = \sigma^2 (X^T X)^{-1}, \tag{3.6}$$

where σ^2 represents the variance (constant) of the random errors, and $(X^T X)^{-1}$ is the inverse matrix of the crossed multiplications of X. Since σ^2 is usually unknown its estimation is given by the following expression,

$$\sigma^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p - 1} = \frac{\sum_{i=1}^{n} e_i^2}{n - p - 1} = \frac{SSE}{n - p - 1} \tag{3.7}$$

where SSE represents the residual sum of squares.

3.1.3 Statistical Inference

One example of statistical inference is, assuming the normality of the random errors, testing the statistical significance of a certain independent variable associated to a given parameter β_i, $i = 1, ..., p$, through an hypothesis test, such as, $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$. The previous hypothesis can be tested with the following formula,

$$T = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}, \tag{3.8}$$

that follows a *t*-Student distribution with, $n - p - 1$ degrees of freedom, and where $\hat{\sigma}^2_{\hat{\beta}_i}$ corresponds to the i-th element of the diagonal of the variance-covariance matrix of the estimators of the parameters, $\sigma^2 (X^T X)^{-1}$.

For a certain level of significance α, the rejection region for the null hypothesis is given by

$$\left] -\infty, -t_{1-\frac{\alpha}{2}; n-p-1} \right] \cup \left[t_{1-\frac{\alpha}{2}; n-p-1}, +\infty \right[\tag{3.9}$$

Therefore, if p-value $\leq \alpha$ the decision falls upon rejecting H_0, hence, one can conclude that β_i is statistically significant, or in other words, that the independent variable X_i contributes to the explanation of the dependent variable, when considering the remaining independent variables constant.

In order to determine the confidence interval for β_i with a confidence level of $(1-\alpha) \times 100\%$ we have,

$$\left(\hat{\beta}_i - t_{1-\frac{\alpha}{2}; n-p-1} \hat{\sigma}_{\hat{\beta}_i}, \hat{\beta}_i + t_{1-\frac{\alpha}{2}; n-p-1} \hat{\sigma}_{\hat{\beta}_i} \right) \tag{3.10}$$

Testing the global significance of the regression can be done through the following hypothesis test,

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \quad vs \quad H_1 : \exists j = 1, ..., p : \beta_j \neq 0 \tag{3.11}$$

which translates into the existence of at least one regression coefficient different than zero. The test statistic for this hypothesis test is given by,

$$F = \frac{(n - p - 1) \times SSR}{p \times SSE} \sim F_{p,n-p-1} \tag{3.12}$$

where SSR is the explained sum of squares, SSE is the residual sum of squares and $F_{p,n-p-1}$ is Fisher's distribution with p parameters and $n - p - 1$ degrees of freedom.

In order to implement the global significance test it's common practice to construct an analysis of variance (ANOVA) table, like Table 3.1.

Table 3.1: Analysis of variance table (ANOVA).

Variation Source	**Sum of Squares**	**Degrees of freedom**	**Mean of squares**	**F Value**
Regression	SSR=$\sum(\hat{Y}_i - \bar{Y})^2$	p	MSR=$\frac{SSR}{p}$	
Residual	SSE=$\sum(Y_i - \hat{Y}_i)^2$	$n - p - 1$	MSE=$\frac{SSE}{n-p-1}$	$\frac{MSR}{MSE}$
Total	SST=$\sum(Y_i - \bar{Y})^2$	$n - 1$	MST=$\frac{SST}{n-1}$	

3.1.4 Goodness-of-fit

The quality of regression for linear models can be ascertained in several ways. Usually, the first approach is a graphical analysis of the results in order to evaluate the quality obtained.

Another way to assess the quality of regression is through the coefficient of determination, designated by $R^2 (0 \leq R^2 \leq 1)$. This coefficient is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). In other words, it's the percentage of the variation of Y that is explained by the model.

One particularity of the coefficient of determination, R^2, is that it tends to increase every time the model is updated with more independent variables, regardless of their significance. With that being said, the best alternative is to analyze the adjusted coefficient of determination, R^2_a, where the previous situation only happens if the new variable is deemed significant for the model. It's possible to calculate the adjusted coefficient of determination with the following formula,

$$R^2_a = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} = 1 - \frac{n-1}{n-p-1}(1 - R^2) \tag{3.13}$$

with

$$R^2 = 1 - \frac{SSE}{SST} \tag{3.14}$$

and

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y_i})^2. \tag{3.15}$$

This coefficient allows an evaluation of the quality of a single model, however, if the goal were to compare two models among themselves there needs to be a different approach.

When comparing models we can have two cases, the models subject of comparisons are nested, this is, the independent variables of one of the models form a subset of the independent variables of the other, or the models are made up by different variables and can't be considered nested.

Regarding the first case, the easiest way to assess which of them can be considered to be more accurate is to test if the parameters corresponding to the independent variables that aren't present in both models are simultaneously zero. If that's the case the decision of what model is best falls upon the one with the least variables, in order to simplify the analysis.

One way to compare a more complex model with a simpler one, assuming they are hierarchically nested is through the likelihood ratio test (LRT). The likelihood scores can be calculated with the difference between the log-likelihoods, as follows,

$$LR = -2[l(\theta_0) - l(\hat{\theta})] \tag{3.16}$$

where θ is the given parameter and $LR \sim \chi^2_p$, with p being the number of parameters.

On the other hand, if the models are not nested the correct way to evaluate and compare the quality between both would be to use the adjusted coefficient of determination, or, alternatively, Akaike's information criterion (AIC). Akaike's information criterion is an estimator of the relative quality of statistical models for a given set of data. This criterion assesses the quality of each of the models, and orders them accordingly to the value obtained. The lower the value of AIC obtained the better the quality of regression, when comparing models. The AIC can be obtained with the following formula,

$$AIC = -2log(L) + 2p \tag{3.17}$$

where L represents the model's likelihood and p the number of parameters.

3.1.5 Variable Selection

The most common methods used to select which of the independent variables are significant for the model are the Backward, Forward and Stepwise methods.

On this manuscript were used the backward method, applying the likelihood ratio test and the stepwise method using Akaike's information criterion.

Both of the methods provide similar results, in fact, the stepwise method it's a combination of the backward and forward methods. This methodology begins with just one variable and adds more recursively. Every time a new variable is added to the model the resulting model is analyzed in order to make sure that all of the variables can be considered significant after that new input. This last process is nothing more than the backward

method, were the initial model is composed by all the variables possible and is simplified until all the remaining ones can be considered significant.

The decision on which of the models obtained (stepwise and backward) is the most accurate one is made by comparing them using Akaike's information criterion, previously mentioned.

3.1.6 Residual Analysis

The residual values in linear regression are nothing more than the difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$),

$$e = y - \hat{y}. \tag{3.18}$$

The main objective of the residual analysis is to comprehend if all the assumptions made upon the data are verified, and if that doesn't happen, understand how that impacts the results.

The assumption of the normality of residuals is easily tested using the Shapiro-Wilks test, or alternatively, the Kolmogorov-Smirnov test, under the null hypothesis that the residuals follow a normal distribution. This analysis can also be done graphically, through a QQ-plot, evaluating if the residuals are located mainly on top of the straight line, and if that's the case then there's graphical evidence that the residuals are normally distributed.

In order to verify the homogeneity of the variance, the independence of the errors and the null mean, one can also use a graphical approach, plotting the residual values versus the predicted values of the dependent variable. In order for the independence of the errors to be satisfied the points on the graph need to be randomly distributed around the residual with null value, forming a cloud of uniform width. If the dispersion of the residuals increases or diminishes along with the values of the dependent variable, the homoscedasticity assumption isn't verified.

This assumption can be tested using Bartlett's test for homogeneity of variances. The null hypothesis of this test states that all the variances are equal across all samples, against the alternative, that the variances are not equal for at least one pair, as follows,

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \quad vs \quad H_1 : \exists^1 i, j \in \mathbb{N}, (i \neq j), \quad \sigma_i^2 \neq \sigma_j^2. \tag{3.19}$$

3.1.7 Multicollinearity

Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present the statistical inferences made upon the latter may not be reliable. There are a few reasons why multicollinearity occurs:

- Can be caused by an inaccurate use of dummy variables;
- Can be caused by the inclusion of a variable which is computed from other variables in the data set;
- The repetition of the same kind of variable;

- Generally occurs when the variables are highly correlated to each other.

This particular issue can result in several problems, such as, the partial regression coefficient may not be estimated precisely, causing the standard errors to be high. Another issue with the coefficient estimated is the change of signs as well as its magnitude when using different samples. Given these two issues, multicollinearity makes it difficult to assess the relative importance of the independent variables in explaining the dependent variable. In the presence of high multicollinearity the confidence intervals of the coefficients tend to become very wide and the statistics tend to be very small. It becomes difficult to reject the null hypothesis of any study when multicollinearity is present in the data.

In order to detect the existence of multicollinearity, one can fall upon a few methods. One of those methods is the analysis of the correlation matrix of the independent variables, in order to determine if there are cases where the correlation values are too high. If one of those cases arises, one solution would be to eliminate one variable out of the pair that displays a high value of correlation, but only after analyzing the data and determining which one would make more sense to remove.

Another method to evaluate the existence of multicollinearity is the determination and analysis of the variance inflation factor (VIF), given by,

$$VIF = \frac{1}{1 - R_j^2} \tag{3.20}$$

where R_j^2 is the determination coefficient of the regression.

Ideally the value of the VIF is as close to 1 as possible, and if that's the case one can conclude that the variables are independent among each other. If the VIF is higher than 10 then we are looking at a case with the presence of multicollinearity. This cut-off point of 10 can vary according to the case study, being as low as 5 in some of them.

3.2 Generalized Linear Models

Advances in statistical theory and computer software allowed the use of methods analogous to those developed for linear models when considering the following more general situations [4]:

- Response variables have distributions other than the Normal distribution – they may even be categorical rather than continuous;
- Relationship between the response and explanatory variables need not be of the simple linear form such as the following,

$$E[Y_i] = \mu_i = x_i^T \beta; \qquad Y_i \sim N(\mu_i, \sigma^2). \tag{3.21}$$

One of these advances has been the recognition that many of the properties of the Normal distribution are shared by a wider class of distributions called the exponential family of distributions [4]. A second advance is the extension of the numerical methods to estimate the parameters β from the linear model described above to the situation

where there is some non-linear function relating $E[Y_i] = \mu_i$ to the linear component $x_i^T \beta$, that is

$$g(\mu_i) = x_i^T \beta \tag{3.22}$$

where the function g is called link function.

In the initial formulation of generalized linear models by Nelder and Wedderburn (1972) g is a simple mathematical function.

To summarize, generalized linear models (GLMs), are not only an extension of all models used to model non Gaussian data, but also of the classic linear model. These models have two major particularities, one of them being the fact that the distribution of the response variable is always a part of the exponential family of distributions, and the other is that the relation between the dependent and independent variables is given by any differential function, always maintaining the linear structure.

3.2.1 Exponential Family

The response variable Y, follows a distribution belonging to the exponential family if its probability density function can be written in the following way,

$$f(y \mid \theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{3.23}$$

where θ is the localization parameter, ϕ is the dispersion parameter and the functions $a(.)$, $b(.)$ and $c(.)$ are known functions for each distribution.

Regarding ϕ we have two possible cases. It can be a known parameter, then we're looking at a distribution of the exponential family with θ as the canonical parameter. Or, the other case where ϕ is unknown, which means that the distribution might not be part of the exponential family. We also admit that the function $b(.)$ is differentiable and that the support of the distribution doesn't rely on the parameters.

3.2.2 Components of a Generalized Linear Model

A generalized linear model is composed by the following three distinct components:

1. The Random Component which identifies the dependent variable that is case of modeling, being a random variable with n independent observations and with a distribution belonging to the exponential family.

2. A linear predictor – that is a linear function of regressors,

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} \tag{3.24}$$

3. A link function – that is smooth and invertible, in this case $g(.)$, which transforms the expectation of the response variable, $\mu_i = E[Y_i]$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}. \tag{3.25}$$

There's special interest in the cases where the linear predictor and the localization parameter match, namely, $\eta_i = \theta_i$. When this happens the link function is named canonical link function. The canonical link functions are mostly used in the context of generalized linear models, since these guarantee the concavity of the likelihood function, and, therefore, a vast quantity of asymptotic results are obtained in an easier way.

3.2.3 Link Functions

On the present dissertation two distinct functions were used as link functions, the *logit* and the *probit*.

The *logit* link function is a fairly simple transformation of the prediction curve, and also provides odds ratios, both features that make it popular among researchers. This link function takes the natural log of the ratio of the probability that Y is equal to 1 compared to the probability that it is not equal to one. Hence, the logistic equation can be written as,

$$ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X \tag{3.26}$$

where π is the probability that $Y = 1$ and $1 - \pi$ the probability that $Y = 0$.

The left hand of the equation represents the *logit* transformation. This transformation can be written in terms of the mean rather than the probability,

$$ln\left(\frac{\mu}{1-\mu}\right) = \alpha + \beta X. \tag{3.27}$$

The transformation of the mean represents a link to the central tendency of the distribution, one of the important defining aspects of any given probability distribution.

Another possibility is the *probit* regression, that, as the names suggests, uses the *probit* link function. This regression utilizes a (inverse) normal distribution link for a binary variable, instead of the *logit* link where $Y^* = \phi^{-1}(\pi)$, and is given by,

$$\phi^{-1}(\pi) = \frac{1}{\sqrt{2\pi^*}} \int_{-\infty}^{\pi} e^{\frac{-z^2}{2}} dz. \tag{3.28}$$

3.2.4 Parameter Estimation

The estimation of parameters of a generalized linear model is usually made resorting to the maximum likelihood estimation (MLE), therefore, the estimators obtained are consistent, asymptotically efficient and with asymptotically normal distribution.

The MLE's objective is to maximize the likelihood function, which is equivalent to maximizing the log-likelihood of β, with a known ϕ, given by

$$log(L(\beta)) = l(\beta) = \sum_{i=1}^{n} \left\{ \frac{w_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi, w_i) \right\} = \sum_{i=1}^{n} l_i(\beta) \tag{3.29}$$

where $l_i(\beta) = \frac{w_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi, w_i)$ is the contribution of each observation of the dependent variable to the likelihood [5].

Assuming that certain conditions of regularity are met, the maximum likelihood estimators for β can be obtained as a solution of the likelihood equations

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l(\beta)}{\partial \beta_j} = 0, \qquad j = 1, ..., p \tag{3.30}$$

equations given by

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \qquad j = 1, ..., p. \tag{3.31}$$

One way to solve these equations is through Fisher's scoring, considered to be the simplest way to do it. This method utilizes Fisher's information matrix as the covariance matrix, $I(\beta)$,

$$I(\beta) = E\left[-\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} \right] = E\left[-\frac{\partial S(\beta)}{\partial \beta} \right] \tag{3.32}$$

where $S(\beta) = \sum_{i=1}^{n} S_i(\beta) = \sum_{i=1}^{n} \frac{\partial l_i(\beta}{\partial \beta_j}, j = 1, ..., p$ is the score function.

On its matrix form, Fisher's information matrix is given by

$$I(\beta) = Z^T W Z \tag{3.33}$$

where W is the diagonal matrix and where its i-th element is given by

$$W_i = \frac{(\frac{\partial \mu_i}{\partial \eta_i})^2}{var(Y_i)}. \tag{3.34}$$

The dispersion parameter can also be estimated using the MLE, however, the simplest way to do it is utilizing a method based on the sampling distribution, for large values of n, of Pearson's goodness-of-fit statistic [5]. Hence, $\hat{\phi}$ is given by,

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i (y_i - \hat{\mu_i})^2}{V(\mu_i)} \quad . \tag{3.35}$$

3.2.5 Hosmer and Lemeshow method

Just like some of the methods previously mentioned, the Hosmer and Lemeshow method is a statistical approach with the objective of minimizing the number of variables used in the modeling of a variable, maintaining only the variables considered relevant. This method of selection of variables can be divided into four different steps.

The first step consists of the creation of simple regression models for each of the variables present on the data, obtaining the corresponding p-values through Wald's test. In general, any variable with a p-value higher than 0.25 is considered in the model. This value is higher than the usual value considered (0.05) so that some variables that may seem not important in the simple regression can be selected, because those variables can reveal themselves as important in the multiple regression.

The next step consists on adjusting the multiple regression model with the variables previously selected, and, once again, the objective is to evaluate the p-values obtained

for each variable and select the ones that have are less than 0.25, excluding the ones that don't verify that condition.

Thirdly, using the variables obtained on the second step, this step consists on adjusting a new multiple regression model. The selection process continues until all the important variables are included in the model, making sure that all the excluded ones are not statistically significant. Lastly, still on this third step, all the variables excluded in the first step should be included on this new model, one by one, by decreasing order of the p-value obtained, with the objective of identifying variables that when analyzed by themselves are not relevant but become such when in the presence of other variables.

The fourth and final step consists on introducing in the model all the interactions that may be relevant. Once all the relevant interactions are identified the previous variable selection process is repeated and all the interaction with a p-value of more than 0.05 are removed from the model.

3.2.6 Statistical Inference

Most of the problems of statistical inference related with hypothesis tests upon the parameter vector β can be formulated such as:

$$H_0 : C\beta = \xi \quad vs \quad H_1 : C\beta \neq \xi, \tag{3.36}$$

where C is a $p \times q$ matrix, with $q \leq p$, of complete characteristic q and ξ is a vector of dimension q, previously specified [5].

In general, there are three statistics to test the hypothesis mentioned above, that are deduced from the asymptotic distributions of the maximum likelihood estimators and from suitable functions of those estimators.

- **Wald's Statistic**, based on the asymptotic normality of the maximum likelihood estimator $\hat{\beta}$, and usually used to test null hypothesis made upon individual components;
- **Wilk's Statistic** or **Maximum likelihood statistic**, based on the asymptotic distribution of the ratio of the maximum likelihoods under the hypothesis H_0 and $H_0 \cup H_1$, used to compare nested models;
- **Rao's Statistic** or **Score Statistic**, based on the asymptotic properties of the score function.

3.2.7 Goodness-of-Fit

Deviance

The deviance for logistic regression plays the same role as the residual sum of squares plays in linear regression. In fact, the deviance, is given by the following function, [5]

$$D^*(y, \hat{\theta}) = -2 \times (l_M(\hat{\beta}_M) - l_S(\hat{\beta}_S)) \tag{3.37}$$

where M is the model being studied and S represents the saturated model.

Considering two nested models, M_1 and M_2, with p_1 and p_2 parameters, accordingly, it's possible to obtain the following function (under the null hypothesis that the less complex model, M_2 is a better fit),

$$\frac{D(M_2) - D(M_1)}{\phi} \sim \chi^2_{p_1 - p_2} \tag{3.38}$$

where $D(M_1)$ and $D(M_2)$ represent the deviance for the models M_1 and M_2, accordingly.

Hosmer-Lemeshow Test

The Hosmer-Lemeshow goodness of fit test is based on dividing the sample up according to their predicted probabilities, or risks. The observations in the sample are split into g groups, according to their predicted probabilities. In other words, supposing that $g = 10$, then the first group consists of the observations with the lowest 10% predicted probabilities. The second group consists of the 10% of the sample whose predicted probabilities are next smallest, and so on.

A formula defining the calculation of this statistic is given by,

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \hat{\pi}_k)} \tag{3.39}$$

where n'_k is the total number of subjects in the k^{th} group, c^k denotes the number of covariate patterns in the k^{th} decile,

$$o_k = \sum_{j=1}^{c_i} y_j \tag{3.40}$$

is the number of responses among the c_k covariate patterns, and

$$\bar{\pi_k} = \sum_{j=1}^{c_i} \frac{m_j \hat{\pi}_j}{n'_k} \tag{3.41}$$

is the average estimated probability.

When it comes to choosing the number of groups to use for the test there isn't a clear criteria. Hosmer and Lemeshow's conclusions for simulations were based on using $g > p+1$, where p represents the number of parameters, suggesting that if the model has 10 covariates, for example, the number of groups would be at least 11. Intuitively, using a small value of g ought to give less opportunity to detect miss-specification. However, if g is too large, the numbers in each group may be so small that it would be difficult to determine whether differences between observed and expected are due to chance or indicative or model mis-specification.

The advantage of a summary goodness of fit statistic like $\hat{C}$ is that it provides a single, easily interpretable value that can be used to assess fit. The great disadvantage is that in the process of grouping there's a chance one might miss an important deviation from fit due to a small number of individual data points [5]. Hence, bibliography advocates that, before finally accepting that a model fits, an analysis of the individual residuals and relevant diagnostic statistics be performed. These methods are presented further along the dissertation.

When it comes to performing this test on the statistical software *R* it can be done through the library *ResourceSelection* and utilizing the command *hoslem.test*.

Prediction Error

The prediction error, as the name suggests, evaluates the error associated to each prediction obtained through the fitted model. Counterintuitively, we want this value to be as close to 1 as possible, given that it translates into the proportion of correct classifications given by the model. It is calculated establishing first a cut-off point that ranges from 0 to 1. Usually the cut-off value used is around 0.5.

This error can be divided in two components: sensitivity and specificity.

Sensitivity (also called the true positive rate) measures the proportion of actual positives which are correctly identified as such (in this specific case, the percentage of failed tires that are correctly identified as such), and is complementary to the false negative rate, and can be calculated with the following formula,

$$Sensitivity = \frac{TruePositives}{(TruePositives + FalseNegatives)} \quad . \tag{3.42}$$

On the other hand, specificity (true negative rate) measures the proportion of negatives which are correctly identified as such (the percentage of approved tires that are identified as such). The formula for the specificity is given by,

$$Specificity = \frac{TrueNegatives}{(TrueNegatives + FalsePositives)} \quad . \tag{3.43}$$

In general, the higher the value of both the sensitivity and specificity, the better, meaning that the fitted model is predicting accurately the classification of each subject.

Area Under the ROC Curve

Area under the ROC (Receiver Operating Characteristic) curve (AUC) is a performance measurement for a classification problem at various threshold settings.

Sensitivity and specificity rely on a single cut point to classify a test result as positive. A more complete description of classification accuracy is given by the area under the ROC curve. This curve, originating from signal detection theory, shows how the receiver operates the existence of signal in the presence of noise. It plots the probability of detecting true signal (sensitivity) and false signal (1-specificity) for an entire range of possible cut-off points [6]. In other words, ROC is a probability curve and AUC represents degree or measure of separability.

The area under the ROC curve, which ranges from zero to one, provides a measure of the model's ability to discriminate between those subjects who experience the outcome of interest versus those who do not.

It's possible to interpret the value obtained for the ROC curve according to the following table,

Table 3.2: AUC value classification.

Value Range	Classification
$AUC \leq 0.5$	This suggests no discrimination
$0.7 < AUC < 0.8$	This is considered acceptable discrimination
$0.8 < AUC < 0.9$	This is considered excellent discrimination
$AUC > 0.9$	This is considered outstanding discrimination

When it comes to calculating the value of the AUC and plotting the curve on *R* it can be done through multiple libraries. One of them is the library *ROCR*, where utilizing the command *plot.roc* one can plot the corresponding ROC curve for a given model. This library also allows the determination of the value of the AUC for each model obtained.

Brier Score

Brier's Score is a measure of the accuracy of a set of probability assessments, it allows the evaluation of the calibration of probabilistic predictions of binary events. Proposed by Brier (1950), it's the average deviation between predicted probabilities for a set of events and their outcomes, so a lower score represents higher accuracy [7]. The score value can be obtained using the following [15],

$$BS = \frac{1}{n}\sum_{t=1}^{n}(f_t - o_t)^2, \tag{3.44}$$

where f_t represents the probability of success estimated for the subject t, o_t the true classification of the subject (0 or 1) and n the size of the sample.

Pseudo R^2 (McFadden & Cox-Snell)

Another way to evaluate the goodness-of-fit of a model is through the adjustment coefficient. In this case there are two specific coefficients of interest, called pseudo-R^2, one by McFadden and the other by Cox-Snell.[14]

The first one, the McFadden coefficient, or ρ^2 is nothing more than a transformation of the log-likelihood function into an index analogous to the multiple correlation coefficient by defining [9]

$$\rho^2 = 1 - \frac{L(\hat{\theta})}{L(\bar{\theta})} \tag{3.45}$$

where $\hat{\theta}$ is the maximum likelihood estimator and $\bar{\theta}$ is zero or is zero except for coefficients of alternative dummies.

The ρ^2 and R^2 indices both vary in the unit interval (except when some coefficients in $\bar{\theta}$ are excluded from θ, in which case a poor fit may yield ρ^2 or R^2 negative). [10]

The values obtained for the McFadden ρ^2, are considerably lower than the usual R^2 obtained in *OLS* (Ordinary Least Squares), therefore, when interpreting this coefficient, values between 0.2 and 0.4 are usually considered to be excellent, and indicate a very good fit for the model.

On the other hand, Cox-Snell's adjustment coefficient, is, similarly to McFadden's, a transformation of the log-likelihood function, and is given by,

$$R^2 = 1 - \left[\frac{L(\bar{\theta})}{L(\hat{\theta})}\right]^{\frac{2}{N}} \tag{3.46}$$

where $L(\bar{\theta})$ and $\hat{\theta}$ are the same as previously defined, and N is the size of the sample. [16]

3.2.8 Residual Analysis

The residual analysis for the logistic regression models follows roughly the same assumptions of a regular analysis for linear regression.

With that being said, in order to assess the goodness-of-fit of the chosen model all the assumptions previously addressed on the section 3.1.6 need to be verified once more.

3.2.9 Parameter Interpretation

Unlike the linear regression models, when it comes to the logistic regression the interpretation of the coefficients is not as linear as evaluating the value obtained and draw conclusions. In order to interpret the coefficients obtained through the logistic regression there's the necessity to introduce a measure of association termed the odds ratio.

Assuming the utilization of the *logit* link function, the odds ratio are then obtained as follows,

$$\hat{OR} = exp(\beta_k), \quad k \in \mathbb{N}. \tag{3.47}$$

When it comes to the interpretation of the odds ratio it depends on the value obtained, as well as the type of variable under study. Regarding the case where the variable is quantitative, for an odds ratio of 3, for example, for each increase of 1 unit the estimated odds of the event increases by a factor of 3. In the same way, for the same kind of variable, if the odds ratio is $\frac{1}{3}$, for each increase in 1 unit the estimated odds of the event decreases by a factor of 3.

However, regarding qualitative variables, if the value is between 0 and 2, taking for example 1.1, this means that regarding the reference category, the variable under study presents a chance of verifying the event 10% higher. In the same way, if the value obtained is, for example, 0.9, this means that the chance of the event being verified is 10% less, when compared to the reference category.

In case the value obtained is between 2 and 3 (once again a random value just for explanatory purposes), this means that the event is approximately two times more likely to occur.

4 Study of tests related to the production of agricultural tires

4.1 Presentation of the data set

The data used on this manuscript refers to dimensional tests to which the tires where subjected in order to evaluate if the final results are within the limits established by the law, and, therefore, if the tires can be cleared for production.

This data can be divided into four main components - nominal variables with identification purposes, variables corresponding to the mold used for each tire, the specifications of each tire, and finally, the two dimensional response variables -, representing tires "created" since the beginning of the department in 2016.

The first component consists of four nominal variables, with identification purposes, as previously mentioned. These four variables are the size of the tire, the experimental test order (ETO) number, the article of the tire, and the number of the mold used. The size of the tire is given by, for example, 280/85 R 24, where the first number, 280, corresponds to the tire's article nominal section width, the second number the article's cross section (height of the tire) and the third number the article's diameter (rim size).

The second component, the mold variables, corresponds to the values used for certain parameters of the mold used for each tire.

Thirdly, the specification variables, refer to all the components utilized on the tire, and values given to certain parameters.

Lastly, the two response variables refer to the result of the final dimensional measures performed upon the tires.

Outside of these four components there's an extra last variable that corresponds to the final evaluation of the tire, that consists on comparing the values obtained on the dimensional tests with the tabled maximum values permitted by law.

Due to confidentiality reasons all the variable names have been coded, hence, from this point on all the conclusions and inferences made upon the data will be mentioned using the coded value assigned to each variable.

4.2 Exploratory Analysis

To understand how the data behaves it's important to submit it to simpler statistical analysis, calculating basic descriptive statistics and interpreting them, in order to achieve

a better understanding of the values present in each variable, the existing variations, possible computation errors, and an overall understanding of all the data.

The data is composed by 38 variables, denominated by X_1 through X_{38}, where nine of them are qualitative, and the remaining 29 are quantitative. Each variable is classified as shown on the following table,

Table 4.1: Variable classification.

Classification	**Variables**
Qualitative	X1 (nominal)
	X2 (nominal)
	X3 (nominal)
	X4 (nominal)
	X21 (ordinal)
	X22 (ordinal)
	X24 (ordinal)
	X34 (ordinal)
	X38 (ordinal)
Quantitative ($\in \mathbb{R}$)	X5 to X20
	X23
	X25 to X33
	X35 to X37

To the present date there were 146 ETOs that were in conditions to be added to the database, therefore that's the number of observations used during the study.

Like previously mentioned, the database has nine qualitative variables, where four of those are just for identification purposes, and the remaining five are composed by different levels of interest. All those variables, their levels and respective absolute and relative frequencies are presented on the Table 4.2.

Table 4.2: Absolute and relative frequencies of the qualitative variables.

Variables	Levels	Absolute Freq.	Relative Freq. ($\times 100\%$)
X_{21}	1	57	39 %
	2	89	61 %
X_{22}	1	6	4 %
	2	70	48 %
	3	49	34 %
	4	15	10 %
	5	6	4 %
X_{24}	4	121	83 %
	5	9	6 %
	6	16	11 %
X_{34}	1	117	80 %
	2	21	15 %
	3	8	5 %
X_{38}	0	29	20 %
	1	117	80 %

Analyzing the values presented on the previous table it's possible to see that regarding the variable X_{22} the values are divided between five levels, and that most of the tires are distributed between levels 2 and 3.

When it comes to variable X_{24}, roughly 83% of the tires assume the level 4 of the variable, which is roughly eight times more than level 6, with 16 tires assuming that level.

Similarly, the variable X_{34} has one level with considerably more observation than the remaining two, the first one, with 115 observations, corresponding to approximately 80% of the sample.

Lastly, the variable X_{38} corresponds to the result of the tire, or in other words, if the tire passed the tests, assuming value 1 if that happened, and 0 otherwise. Hence, it's possible to observe that 115 tires passed the tests, around 80% of the entire sample, while the remaining 29 tires failed, due to either width or diameter issues.

Regarding the remaining variables in the database, they are quantitative, and 29 in total. In Table 4.3 are presented some of the measures of location for these variables.

Table 4.3: Summary statistics for quantitative variables.

Variables	Min.	1^{st} Quad.	Median	3^{rd} Quad.	Max.	Mean	Std. Deviation
X_5	1092.0	1328.0	1554.0	1760.0	2060.0	1561.0	252.4
X_6	266.0	399.0	455.0	515.0	700.0	469.3	102.0
X_7	286.0	442.2	486.2	550.0	763.0	505.0	111.3
X_8	279.4	419.1	457.2	539.9	762.0	485.7	109.6
X_9	565.0	820.0	915.0	1190.0	1760.0	1007.0	290.5
X_{10}	39.0	46.0	50.0	54.0	63.0	50.2	4.9
X_{11}	632.0	738.0	788.0	990.0	1192.0	829.5	150.4
X_{12}	75.6	120.7	133.4	162.9	235.0	142.4	32.7
X_{13}	280.0	420.0	460.0	540.0	800.0	492.5	118.2
X_{14}	60.0	70.0	85.0	85.0	85.0	77.3	8.9
X_{15}	24.0	28.0	30.0	38.0	46.0	31.61	6.0
X_{16}	7.7	8.5	9.5	13.4	16.7	10.5	2.7
X_{17}	11.4	12.1	13.4	15.0	24.3	13.8	2.8
X_{18}	480.0	770.0	875.0	1022.5	1350.0	892.6	196.7
X_{19}	47.7	85.2	116.7	173.9	310.6	135.3	64.0
X_{20}	2.5	2.5	2.5	2.5	4.0	2.6	0.3
X_{23}	250.0	380.0	410.0	490.0	720.0	445.0	109.2
X_{25}	107.0	179.0	205.0	240.0	362.0	213.1	50.7
X_{26}	0.0	8.0	10.0	11.0	25.0	9.8	4.5
X_{27}	17.0	22.0	25.0	27.0	35.0	25.2	3.8
X_{28}	19.0	23.0	25.0	28.0	36.0	26.1	4.2
X_{29}	20.0	26.0	28.0	32.0	39.0	28.6	4.5
X_{30}	135.0	195.0	220.0	260.0	370.0	234.1	56.1
X_{31}	30.0	50.0	50.0	50.0	130.0	51.9	21.8
X_{32}	7.0	8.0	8.0	9.0	14.0	8.3	1.3
X_{33}	7.0	8.0	8.0	10.0	14.0	8.8	1.4
X_{35}	10.0	13.0	15.0	16.0	27.0	15.8	3.9
X_{36}	292.0	436.9	439.4	545.2	818.9	510.1	116.7
X_{37}	1088.0	1323.0	1569.0	1774.0	2101.0	1570.0	264.4

4.2.1 Correlation

Determining the correlation between all the variables provides an important insight in order to understand how the data and the variables behave.

The following matrix represents the correlations between all variables,

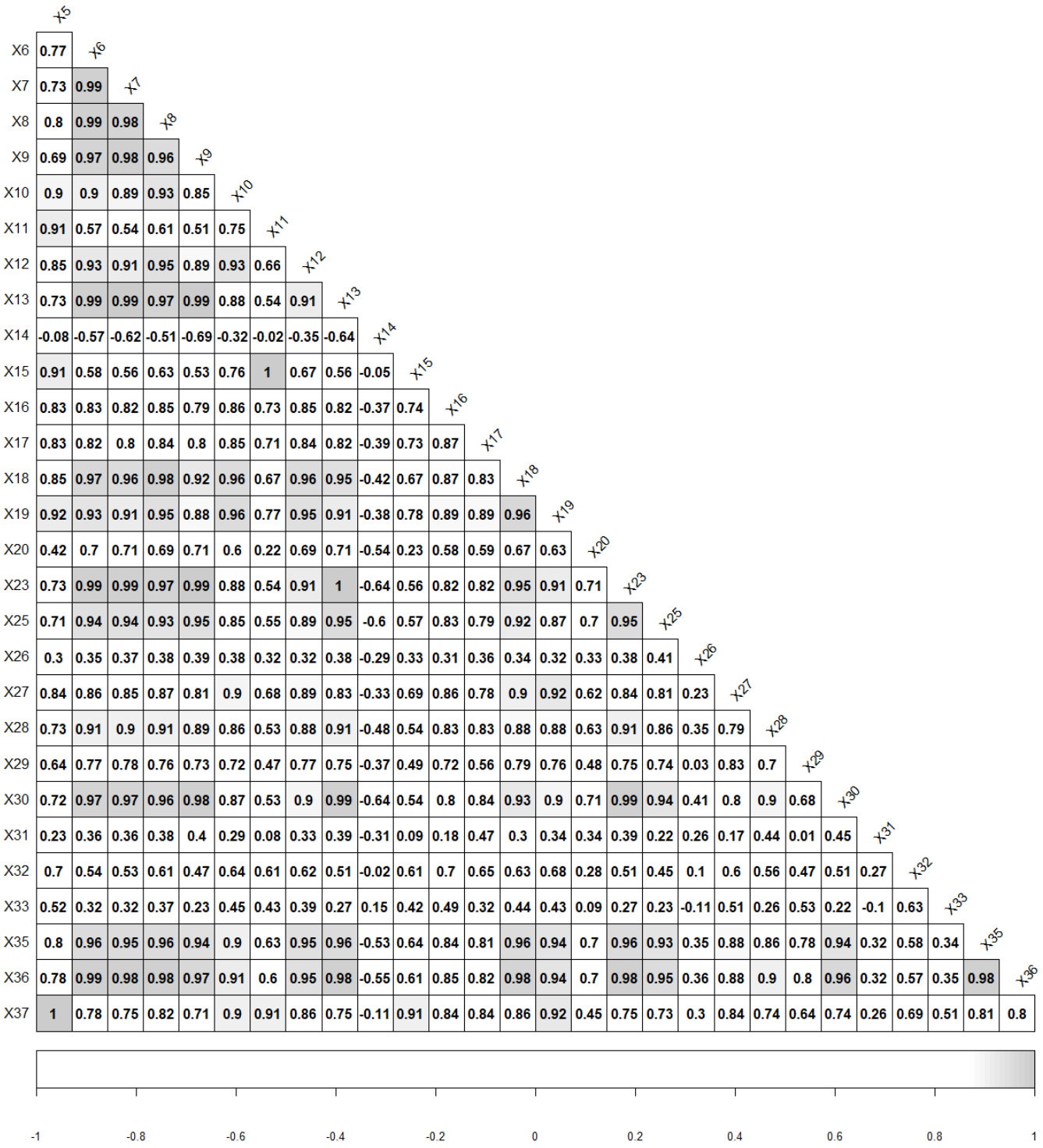


Figure 4.1: Spearman correlation matrix for all variables.

Given the context of the data on which this manuscript is based upon, it was expectable from the beginning to obtain such high values of correlation between all the variables. Therefore, in order to analyze which ones are most correlated amongst each other, the rest of the analysis is made assuming high correlation when the value is higher than 0.9.

As it's possible to see there are several variables whose correlation is extremely high, close to 1. Like previously mentioned, this happens because some of the variables are somewhat obtained through each other, or, in some other cases, because a change in that specific variable implies a change on another one, in order to compensate the alteration made so that the tire doesn't present construction problems.

Analyzing all the correlations, and in particular the ones involving the response variables we can see that when it comes to the dependent variable X_{36} there are several variables suggested to have an high correlation with this one.

On the other hand, the other response variable, X_{37}, presents fewer variables with correlation value above 0.9 when compared with the other dependent variable. It's important to point out that out of the five variables with correlation value superior to 0.9, the correlation between the response variable and the variable X_5 is close to 1, suggesting that this variable offers an almost complete explanation of the dependent variable.

It's also important to mention that the values of correlation obtained imply, and justify, the existence of multicollinearity on the data, as it was expected given the nature of said data.

4.2.2 Response variable X_{36}

X_{36} is one of the response variables evaluated in this manuscript and, as previously mentioned, it's a quantitative variable regarding one of the dimensional results.

The boxplots represented on the Figure 4.2 enable the possibility of establishing a connection between X_{36} and the qualitative variables.

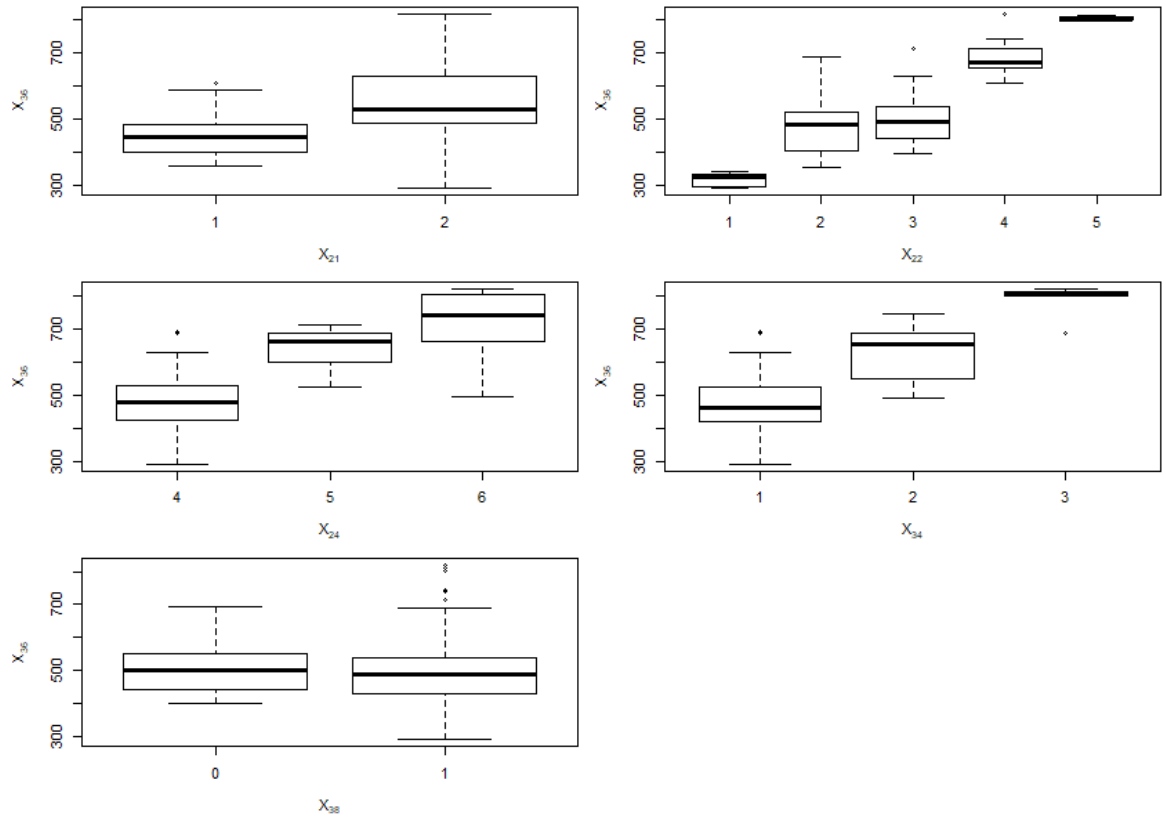


Figure 4.2: Boxplots for the relation between the qualitative variables and X_{36}.

It's possible to observe that for the variable X_{21} the two existing levels present different median levels, and don't indicate the existence of any outliers.

For the variable X_{22}, the tires with the value 2 and 3 have similar median results. It's also clear the existence of two outliers, on levels 3 and 4, accordingly.

Doing the same analysis for the variable X_{24} it's clear that, unlike the previous case, all three levels have distinct medians. This variable, similarly to the latter, also suggests the presence of two outliers, but this time on the same level.

Looking at the variable X_{34} it's clear the existence of two major levels, the first two, and a third one with considerably less observations, as seen in Table 4.2. The last level, besides having an inferior number of observations also has the particularity of suggesting the existence of an outlier.

Lastly, the boxplot for the variable X_{38} suggests very close levels for the medians of the two levels, and, once again, one of the levels shows signs of the existence of outliers, but on a larger quantity this time.

4.2.3 Response variable X_{37}

The other response variable is X_{37}, and, similarly to the other one, an analysis of the boxplots for the qualitative data is an important approach to understand the behaviour of those variables.

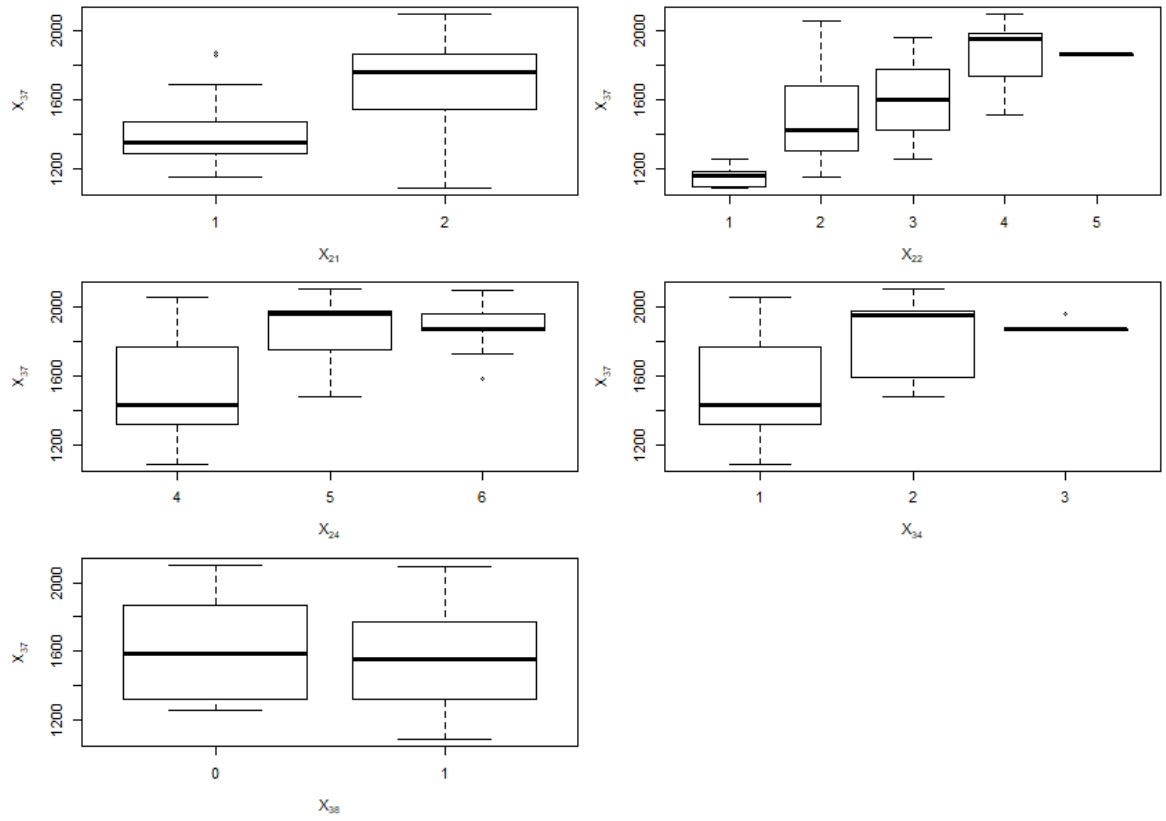


Figure 4.3: Boxplots for the relation between the qualitative variables and X_{37}.

Unlike the other response variable, for X_{37} only the variable X_{38} shows signs of having equal medians for both levels, while the other variables don't suggest that scenario for any of their levels.

Another aspect that it's important to point out is that, on this case, the total number of outliers suggested is just two, being distributed each on the third level of the variables X_{24} and X_{34}, accordingly.

4.3 Modeling of the variable X_{36}

In order to be able to make predictions for the outcomes of both the response variables the data had to be fitted to two models (one for each variable) obtained using linear regression.

In this case, the results presented in this sub-chapter are related to the variable X_{36}.

The expected outcome of the regression would be a model with the following appearance,

$$X_{36} = \beta_0 + \beta_1 X_5 + ... + \beta_{36} X_{35} \tag{4.1}$$

where only the first four variables of the database are omitted given the fact that they don't provide any important information when it comes to explaining the outcome of the dependent variable.

The Table 4.6 represents the estimated coefficients of β for the complete model,

Table 4.4: Coefficients for the complete model of the variable X_{36}.

Variables	**Level**	**Coefficient ($\hat{\beta}_i$)**	**p-value**	
Intercept	—	-22.57	0.14	
X_5	—	0.05	0.04	*
X_6	—	0.01	0.99	
X_7	—	0.30	≈ 0.00	*
X_8	—	-0.07	≈ 0.00	*
X_9	—	0.04	0.02	*
X_{10}	—	0.09	0.74	
X_{11}	—	0.27	0.14	
X_{12}	—	0.09	0.40	
X_{13}	—	0.04	0.66	
X_{14}	—	-0.21	0.13	
X_{15}	—	-8.13	0.09	
X_{16}	—	-1.45	≈ 0.00	*
X_{17}	—	0.12	0.72	
X_{18}	—	0.16	≈ 0.00	*
X_{19}	—	-0.06	0.06	
X_{20}	—	4.17	0.01	*
X_{21}	2	-3.03	≈ 0.00	*
X_{22}	2	-4.06	0.01	*
	3	-7.98	≈ 0.00	*
	4	-14.33	≈ 0.00	*
	5	-21.09	≈ 0.00	*
X_{23}	—	0.01	0.88	
X_{24}	5	1.96	0.14	
	6	8.84	≈ 0.00	*
X_{25}	—	0.01	0.99	
X_{26}	—	0.29	≈ 0.00	*
X_{27}	—	-0.01	0.96	
X_{28}	—	0.43	0.01	*
X_{29}	—	0.29	0.01	*
X_{30}	—	0.03	0.57	
X_{31}	—	-0.02	0.48	
X_{32}	—	1.42	0.01	*
X_{33}	—	-0.75	0.03	*
X_{34}	2	-0.10	0.93	
	3	9.73	0.13	
X_{35}	—	9.38	≈ 0.00	*
$^* < 0.05$		$R^2 \approx 0.99$	$R^2_a \approx 0.99$	

In order to obtain the best possible model, and therefore, reduce its error and number of variables included maintaining only the important ones several methods were applied, with the final objective of choosing the one that provides the best outcome.

With this goal in mind, in order to minimize the number of variables used in the model, the method applied was the backward method, with two variants, one of them using AIC, and the other one using the likelihood ratio's test. After that the choice fell upon the model with the best combination for the value of AIC and the number of parameters. The results are presented in the Table 4.5,

Table 4.5: Descriptive table of the models obtained through different methods for X_{36}.

Method	**Parameters**	**AIC**
Complete	36	652.29
AIC	29	639.18
LRT	21	640.77

Analyzing the results presented in the previous table it's possible to observe that both models present an almost equal value of AIC, however, the model obtained through the likelihood ratio test displayed eight less variables when compared to the other model.

Upon analyzing all the data and both the models obtained the next step would be to decide which one of the two models would be used for further analysis and prediction purposes. The decision ended up falling upon the model obtained through the AIC, given that it's the one that has a lower value of AIC, and, even though it presents more variables (less simple), those extra variables could provide important information and would give the opportunity to observe how its variation affects the outcome of the response variable on the predictions.

The model's equation is then given by,

$$\begin{aligned} X_{36} =& \beta_0 + \beta_1 \cdot X_5 + \beta_2 \cdot X_7 + \beta_3 \cdot X_8 + \beta_4 \cdot X_9 + \beta_5 \cdot X_{10} \\ &+ \beta_6 \cdot X_{11} + \beta_7 \cdot X_{12} + \beta_8 \cdot X_{13} + \beta_9 \cdot X_{14} + \beta_{10} \cdot X_{15} \\ &+ \beta_{11} \cdot X_{16} + \beta_{12} \cdot X_{18} + \beta_{13} \cdot X_{19} + \beta_{14} \cdot X_{20} + \beta_{15} \cdot If(X_{21} = 2) \\ &+ \beta_{16} \cdot If(X_{22} = 2) + \beta_{17} \cdot If(X_{22} = 3) + \beta_{18} \cdot If(X_{22} = 4) \\ &+ \beta_{19} \cdot If(X_{22} = 5) + \beta_{20} \cdot If(X_{24} = 5) + \beta_{21} \cdot If(X_{24} = 6) \\ &+ \beta_{22} \cdot X_{26} + \beta_{23} \cdot X_{28} + \beta_{24} \cdot X_{29} + \beta_{25} \cdot X_{31} + \beta_{26} \cdot X_{32} \\ &+ \beta_{27} \cdot X_{33} + \beta_{28} \cdot If(X_{34} = 2) + \beta_{29} \cdot If(X_{34} = 3) + \beta_{30} \cdot X_{35} \end{aligned} \tag{4.2}$$

where the estimates for the β_i coefficients are presented 4.6.

Table 4.6: Coefficients for the adjusted model of the variable X_{36}.

Variables	**Level**	**Coefficient ($\hat{\beta}_i$)**	**p-value**	
Intercept	–	-17.39	0.19	
X_5	–	0.05	0.01	*
X_7	–	0.30	≈ 0.00	*
X_8	–	-0.07	≈ 0.00	*
X_9	–	0.04	0.02	*
X_{11}	–	0.28	0.10	
X_{12}	–	0.12	0.17	
X_{13}	–	0.08	0.09	
X_{14}	–	-0.24	0.06	
X_{15}	–	-8.41	0.05	
X_{16}	–	-1.40	≈ 0.00	*
X_{18}	–	0.15	≈ 0.00	*
X_{19}	–	-0.06	0.04	*
X_{20}	–	3.98	0.01	*
X_{21}	2	-3.04	≈ 0.00	*
X_{22}	2	-4.00	≈ 0.00	*
	3	-7.98	≈ 0.00	*
	4	-14.51	≈ 0.00	*
	5	-21.14	≈ 0.00	*
X_{24}	5	1.98	0.10	
	6	8.69	≈ 0.00	*
X_{26}	–	0.30	≈ 0.00	*
X_{28}	–	0.41	≈ 0.00	*
X_{29}	–	0.25	0.01	*
X_{31}	–	-0.02	0.17	
X_{32}	–	1.54	≈ 0.00	*
X_{33}	–	-0.76	0.02	*
X_{34}	2	-0.02	0.98	
	3	8.77	0.04	*
X_{35}	–	9.35	≈ 0.00	*
$^* < 0.05$		$R^2 \approx 0.99$	$R^2_a \approx 0.99$	

Observing the values obtained it was possible to cross reference those with the existent knowledge, and understand if they made sense.

Some of the variables that were considered significant by the model and that corresponded to the already known facts were, X_{18}, X_{22}, X_{24} and X_{35}.

However, some other variables that were considered significant came as a surprise, since they were being overlooked. The two most relevant ones among this group are X_7 and X_{16}.

When it comes to X_7 it came as a surprise that this variable was responsible for determining a substantial part of the final dimension of the tire represented by the response variable X_{36}.

Just like X_7, X_{16} was also a surprising result, so much so that there was a need to test

it. For that, two sets of tires were built, with the same size, mold and overall specification, where the only difference between them was the value of the variable X_{16}. The results obtained confirmed that, in fact, X_{16} affected the final dimension of the tire, however, it was also possible to observe that the coefficient obtained for this variable was slightly higher than what the real result came to be.

4.4 Modeling of the variable X_{37}

Equally to the variable X_{36}, the same methodologies were applied for the modeling of the variable X_{37}.

Once again were applied the same two methods in order to obtain the best possible model, with the results presented in Table 4.7.

Table 4.7: Descriptive table of the models obtained through different methods for X_{37}.

Method	Parameters	AIC
Complete	36	767.85
AIC	21	748.54
LRT	15	750.44

Just like for X_{36}, the choice for the best adjustment for this variable falls upon the model obtained through AIC, even though it has more parameters than the model obtained by applying the likelihood ratio test.

The equation for the adjusted model is then given by,

$$\begin{aligned} X_{37} = & -15.86 + 0.87 \cdot X_5 + 1.03 \cdot X_{10} + 1.22 \cdot X_{11} + 0.11 \cdot X_{12} \\ & - 0.22 \cdot X_{14} - 26.85 \cdot X_{15} + 1.92 \cdot X_{16} - 0.88 \cdot X_{17} \\ & + 0.10 \cdot X_{18} + 0.19 \cdot X_{19} - 0.16 \cdot X_{23} - 9.35 \cdot If(X_{24} = 5) \\ & - 19.25 \cdot If(X_{24} = 6) + 0.08 \cdot X_{25} - 0.49 \cdot X_{26} - 1.21 \cdot X_{29} \\ & + 0.07 \cdot X_{31} - 1.93 \cdot X_{32} + 0.68 \cdot X_{33} + 23.06 \cdot If(X_{34} = 2) \\ & + 54.92 \cdot If(X_{34} = 3) \end{aligned} \tag{4.3}$$

Upon analyzing the variables considered significant on the model, it was possible to understand that the variables X_5 and X_{34} were the ones with an increased importance in the determination of the final dimension of the tire represented by this response variable.

Furthermore, a consequent study upon the relation/correlation of these two variables lead to a quite important conclusion. The value assumed by the variable X_5 is determined through a relation with another parameter, which varies for different sizes of tires. This was thought to be the only variation parameter for that value, however, further analysis determined that, the value presented by the variable X_{34} also influences the final value for the variable X_5, meaning that, taking into consideration the the factor of X_{34}, X_5 can be different for tires of the same size. This relation was previously unknown, and actually explained quite a few of the unexpected results obtained until then.

4.5 Residual Analysis

In order to evaluate the quality of adjustment of the model obtained there's the need to perform a residual analysis. This analysis aims at identifying the difference between the observed values and the fitted values obtained, where disparity between this difference can result in a bad adjustment of the observations. Another goal is to ascertain about the existence of isolated deviations of the model, in other words, the existence of one or more observations that don't follow the same pattern as the remaining. These observations can be: outliers, leverage points and influence points.

The following six graphics allow the evaluation of the normality of the residuals, the influential observations and how the residuals behave according to the fitted values for both the response variables.

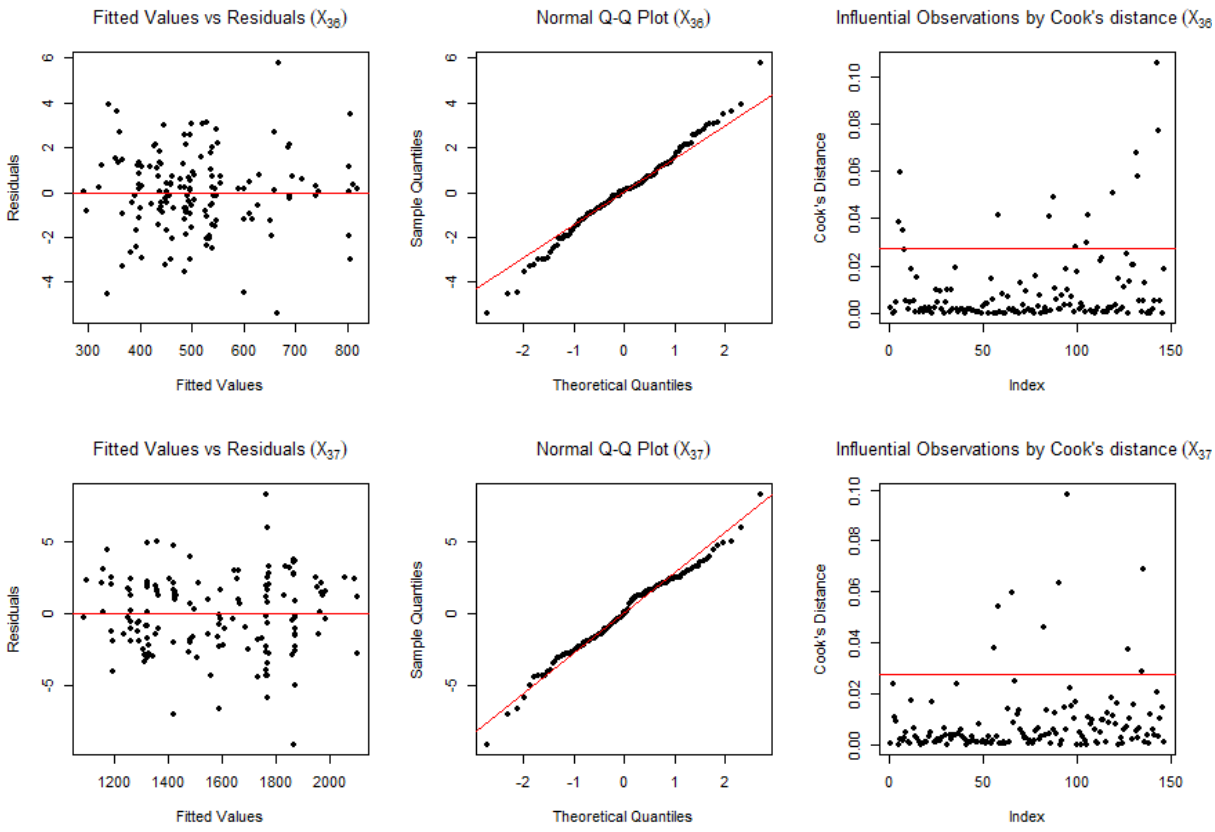


Figure 4.4: Left: Fitted values vs residuals; Center: Normality QQ-Plot; Right: Influential observations by Cook's distance.

Analyzing the left graphic for both of the variables it's possible to see that most of the residual values are within a range of 4 mm for the variable X_{36}, and 5 mm for X_{37}, with some exceptions, called outliers. These outliers correspond to the observations 6, 119, 131 and 132 for the first variable, and 65 and 95 for the second one. Upon analysis of these observations it doesn't seem to be a computation or human error, therefore can't be considered "wrong". Further along will be studied the case where these observations are removed from the data.

The center graphic is a normal QQ-plot, whose purpose is to indicate whether the residuals follow a Gaussian distribution, in other words, if the assumption of normality

is verified. Upon analysis of the graphic it seems to suggest that the residuals do in fact verify the assumption. As previously done for the variables, utilizing Lilliefor's test for normality it was possible to obtain a p-value for the hypothesis test of normality of residuals (H_0: the errors follow a normal distribution) obtaining a value of 0.35 for X_{36} and 0.14 for X_{37}, hence, not rejecting the null hypothesis for either of them and confirming the normality of residuals suggested by the graphical analysis.

Another possibility for the graphical analysis of normality is through the histogram for both the variables and the respective density lines, with the objective of evaluating its shape. Hence it was possible to obtain the following histograms,

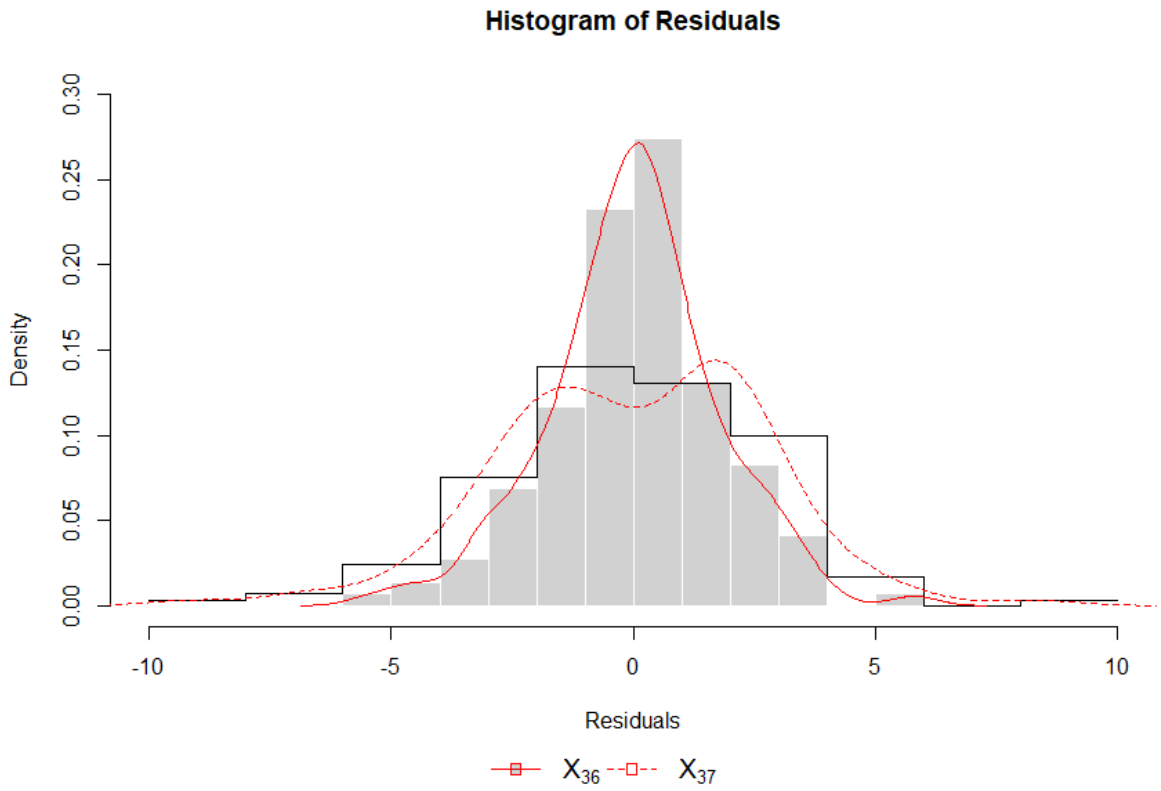


Figure 4.5: Histogram of residuals for both response variables.

Lastly, the third graphic of Figure 4.4 allows the identification of the influential observations calculated with the Cook's distance. As it's possible to see there seem to be quite a few observations classified as such (points above the red line), especially for the variable X_{36}. The next two plots show in detail those influential observations, as well as their index.

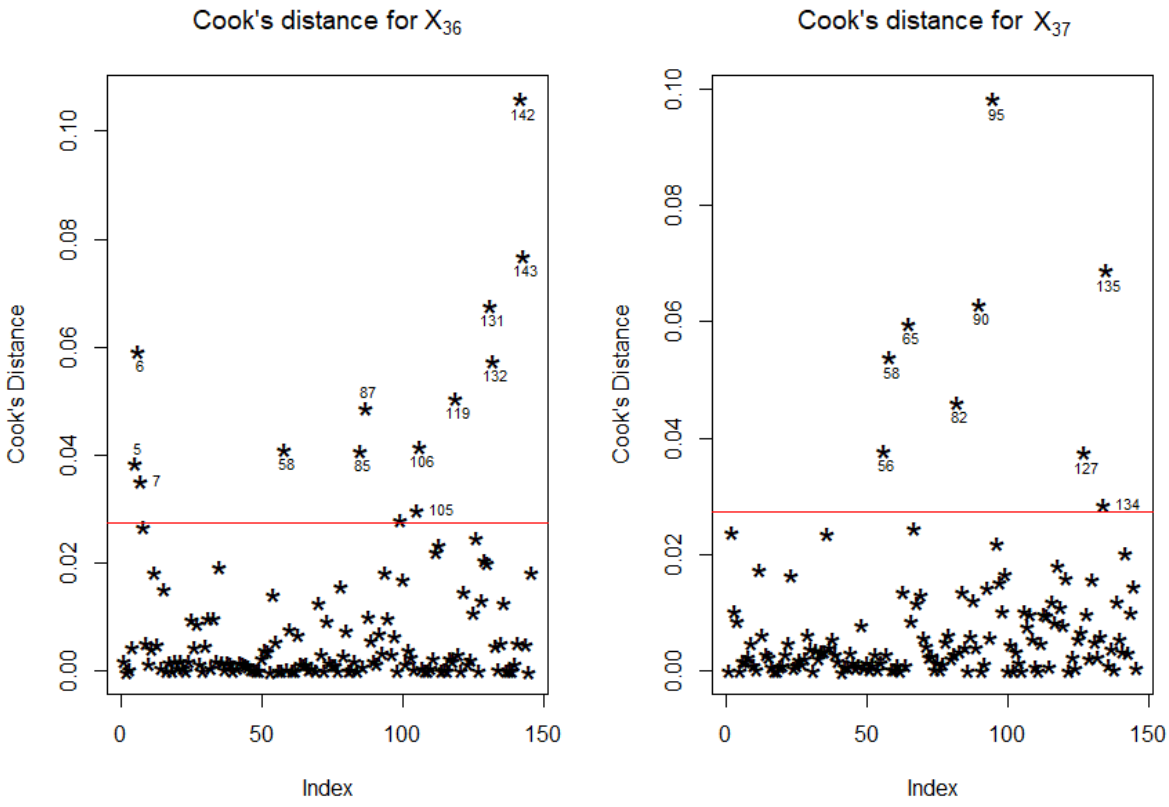


Figure 4.6: Cook's distance plot for both response variables.

Both the model's leverage points are given by the next two plots,

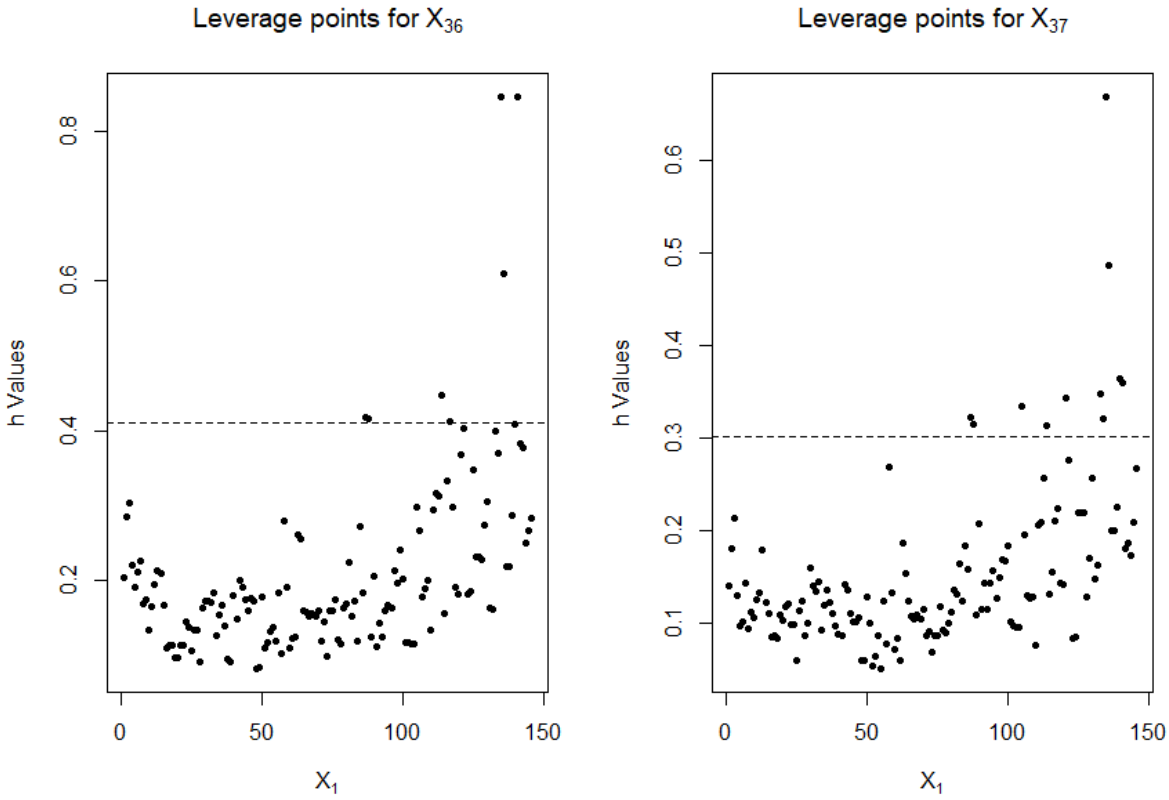


Figure 4.7: Leverage points for both the response variables.

Pregibon [13] suggests the utilization of elements of the diagonal of the H matrix and the analysis of the points that stand out in order to identify the leverage points. Some authors [5] suggest a practical rule in order to identify such observations: values that verify $h_{ii} \geq \frac{2n}{p}$. On Figure 4.7 the leverage points reveal the influence of the observation on its predicted value. Through the graphical inspection it's possible to verify that X_{36} displays only seven points above the dotted line, that can be considered influential observations. On the other hand, X_{37} displays quite a few more observations in that condition, 11 of them to be more precise.

In order to analyze the deviation of the residuals were obtained the plots shown on Figure 4.8.

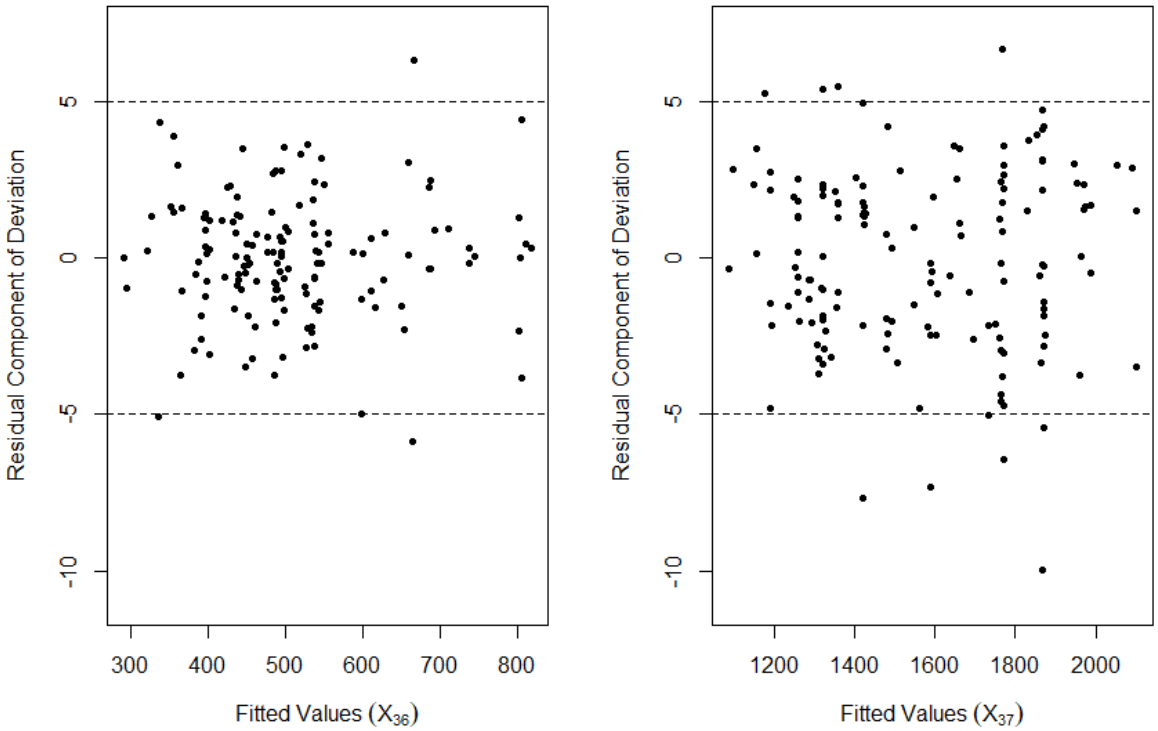


Figure 4.8: Residual deviation for both the response variables.

It's possible to observe that regarding the variable X_{36} only two observations have a deviation bigger than 5 mm (with a couple more marginal observations), which can be considered quite acceptable.

On the other hand, the graphic for the variable X_{37} shows evidence of considerably more observations with a deviation superior to 5 mm (absolute value), around nine observations with a couple more marginal ones.

When it comes to the outliers present in the residual values they are represented in the following two boxplots, for each response variable.

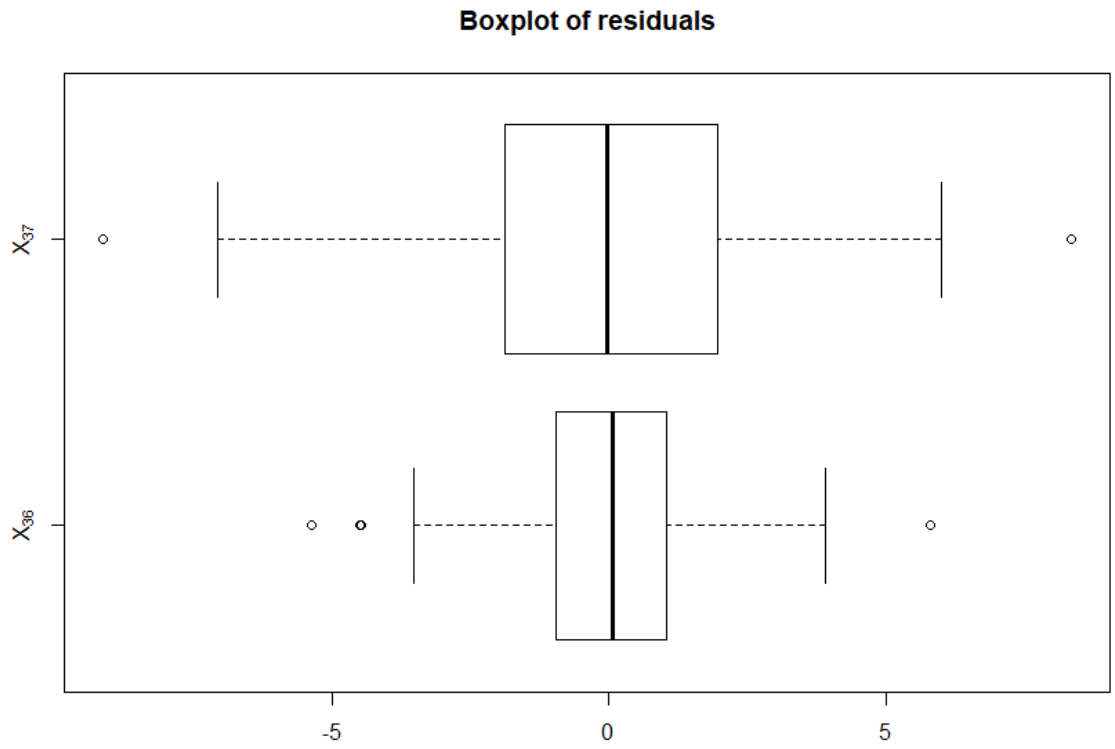


Figure 4.9: Residual outliers for both the response variables.

As is observable on the previous boxplot, the variable X_{36} presents four outliers, corresponding to the observations 6, 119, 131 and 132. On the other hand, the variable X_{37} shows evidence of the existence of only two outliers, corresponding to the observations 65 and 95. Even though the second variable has less outliers, their distance to the average residual value is great than any of the outliers of the variable X_{36}.

4.6 Models' Validation

One important aspect to test after obtaining any kind of modeling is testing it with data not included in the database. In this particular case, the important aspect to test would be the quality of the predictions obtained with the model for new values not included in the data. Given the lack of data, this was only possible to do for five different data entries presented 4.8,

Table 4.8: Prediction for entries not included in the data.

Entry	Real Value (X_{36})	Real Value (X_{37})	Prediction (X_{36})	Prediction (X_{37})
A	495.00	1953.80	496.10	1952.10
B	485.90	1769.20	486.60	1770.30
C	358.90	1191.40	355.30	1189.10
D	353.70	1194.00	351.40	1190.20
E	488.80	1421.30	489.60	1421.20

Analyzing the values obtained it's possible to see that the biggest difference between the real value and the value obtained through the prediction using the model is on the entry C for the variable X_{36}, with approximately 3.6 mm of difference. When it comes to the variable X_{37} the entry with the biggest difference between values is D, with around 3.8 mm of difference, almost the same as the maximum error obtained for X_{36}.

Thus, it's clear that the outputs given by the adjusted model are within an acceptable error range from the real values, providing very important information and facilitating the comprehension of how certain changes on the variables affect the final expected result.

The ideal approach, in order to test the quality of the model, would be to split the data into two different sets: test and train data (30% and 70%, for example) and perform all the analysis on the train data and further along, in order to validate the results obtained, test them on the test data, and draw conclusions from the values obtained. However, given the reduced size of data used for this study, and also some particularities present on it, this approach was not viable in this situation.

4.7 Modeling without outliers

It's advisable not to remove outliers unless there's certainty that they are input errors or measurement errors, which is not the case in the data treated in this manuscript. However, in order to understand the influence of these observations on the results, they were removed and all the analysis were remade and compared to the results previously obtained.

The observations removed in order to realize this analysis were the ones mentioned as outliers on the section 4.5 (6, 65, 95, 119, 131 and 132).

In order to obtain the best possible model for the data, this time without the outliers, were used the same approaches as before (stepwise and backward using the LRT).

Table 4.9: Descriptive table of the models obtained through different methods for X_{36} and X_{37} (without outliers).

Variable	Method	Parameters	AIC
X_{36}	Complete	36	597.03
	AIC	27	585.70
	LRT	25	588.45
X_{37}	Complete	36	725.85
	AIC	28	715.23
	LRT	26	715.03

By analyzing the values obtained it's possible to see that, regarding the variable X_{36}, the choice of the model falls upon the one obtained through the AIC method, even though the one resulting from the LRT method presents less variables. This choice was made given the fact that, just as before, even though the second model has less variables, they can be important to explain the behaviour of the response variable, adding to the fact that the AIC value for the first model is lower.

Regarding the variable X_{37} the choice is the same, and the reasons are similar. Even though the values of AIC for both the models are almost the same, the two extra variables present in the model obtained through the AIC method can be important, like previously mentioned.

With the choice being made the resulting models for X_{36} and X_{37} are then presented in Table 4.10 and Table 4.11,

Table 4.10: Coefficients for the adjusted model of the variable X_{36} (without outliers).

Variables	**Level**	**Coefficient ($\hat{\beta}_i$)**	**p-value**	
Intercept	—	-28.31	≈ 0.00	*
X_5	—	0.06	≈ 0.00	*
X_7	—	0.28	≈ 0.00	*
X_8	—	-0.07	≈ 0.00	*
X_9	—	0.06	≈ 0.00	*
X_{11}	—	0.24	0.11	
X_{14}	—	-0.13	0.06	
X_{15}	—	-7.34	0.05	
X_{16}	—	-1.45	≈ 0.00	*
X_{18}	—	0.17	≈ 0.00	*
X_{19}	—	-0.07	≈ 0.00	*
X_{20}	—	4.60	≈ 0.00	*
X_{21}	2	-3.37	≈ 0.00	*
X_{22}	2	-5.80	≈ 0.00	*
	3	-9.74	≈ 0.00	*
	4	-16.56	≈ 0.00	*
	5	-23.03	≈ 0.00	*
X_{23}	—	0.06	0.05	
X_{24}	5	1.23	0.26	
	6	8.58	≈ 0.00	*
X_{26}	—	0.27	≈ 0.00	*
X_{28}	—	0.42	≈ 0.00	*
X_{29}	—	0.27	≈ 0.00	*
X_{32}	—	1.49	≈ 0.00	*
X_{33}	—	-0.67	0.01	*
X_{34}	2	0.19	0.85	
	3	7.47	0.01	*
X_{35}	—	9.23	≈ 0.00	*
$^* < 0.05$		$R^2 \approx 0.99$	$R^2_a \approx 0.99$	

Table 4.11: Coefficients for the adjusted model of the variable X_{37} (without outliers).

Variables	**Level**	**Coefficient ($\hat{\beta}_i$)**	**p-value**	
Intercept	–	-10.64	0.49	
X_5	–	0.88	≈ 0.00	*
X_6	–	0.10	≈ 0.03	*
X_{10}	–	0.73	0.01	*
X_{11}	–	1.25	≈ 0.00	*
X_{12}	–	0.18	0.01	*
X_{14}	–	-0.38	0.01	*
X_{15}	–	-28.00	≈ 0.00	*
X_{16}	–	1.65	≈ 0.00	*
X_{18}	–	0.09	≈ 0.00	*
X_{19}	–	0.17	≈ 0.00	*
X_{21}	2	-1.33	0.07	
X_{22}	2	-3.18	0.07	
	3	-4.94	0.01	*
	4	-4.59	0.12	
	5	-3.90	0.36	
X_{23}	–	-0.19	≈ 0.00	*
X_{24}	5	-9.68	≈ 0.00	*
	6	-18.97	≈ 0.00	*
X_{25}	–	0.07	0.06	
X_{26}	–	-0.45	≈ 0.00	*
X_{28}	–	0.27	0.14	
X_{29}	–	-1.34	≈ 0.00	*
X_{30}	–	-0.15	0.01	*
X_{31}	–	0.07	0.05	
X_{32}	–	-1.43	0.02	*
X_{33}	–	0.59	0.14	*
X_{34}	2	23.39	≈ 0.00	*
	3	48.40	≈ 0.00	*
$^* < 0.05$		$R^2 \approx 0.99$	$R^2_a \approx 0.99$	

4.7.1 Residual Analysis

Just like for the previously obtained models the new ones ask for an analysis in order to evaluate the overall results and quality of adjustment. So, just like before, the next six graphics are an important asset for the understanding of the results obtained for both the response variables, this time without the outliers.

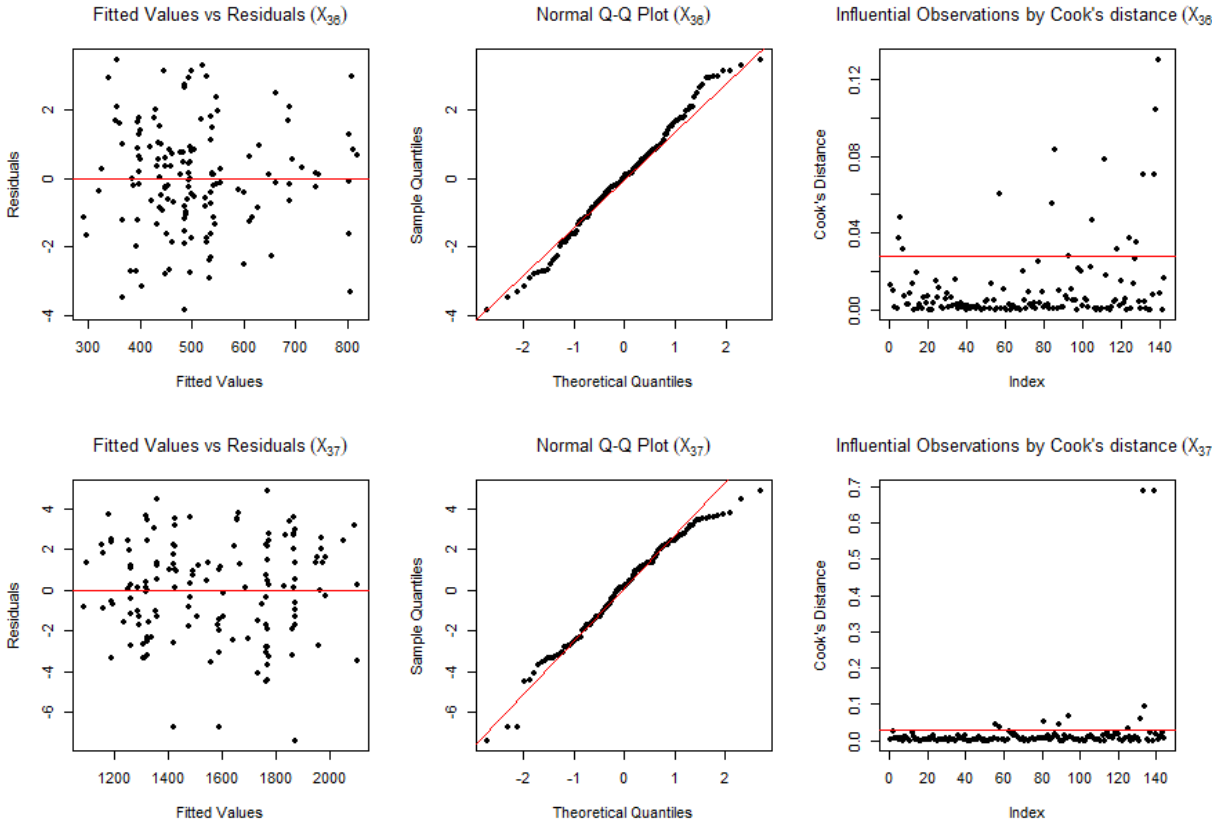


Figure 4.10: Left: Fitted values vs residuals; Center: Normality QQ-Plot; Right: Influential observations by Cook's distance (for models without outliers).

Analyzing both the left most graphics it's possible to observe that this time around all the observations are within a range of 4 mm when it comes to the variable X_{36}, unlike before. However, regarding the variable X_{37} it's observable that some of the fitted values present a distance superior to 6 mm from the observed values, even superior than the ones previously obtained on the models with the outliers.

The plots on the center represent the normal QQ-plots to assert the hypothesis of normality of the residuals of both the response variables. Observing the graphic corresponding to the response variable X_{36} it's possible to conclude that the residuals seem to follow a Gaussian distribution, confirmed by Lilliefor's test for normality, with a value of 0.91, not rejecting the null hypothesis of normality of residuals. Similarly, for the variable X_{37}, the graphical analysis also suggests some kind of normality of the residuals, that is once again confirmed by the Lilliefor's test, with a value of 0.1. Once again, the graphical analysis of normality can also be done through the following histograms for both variables.

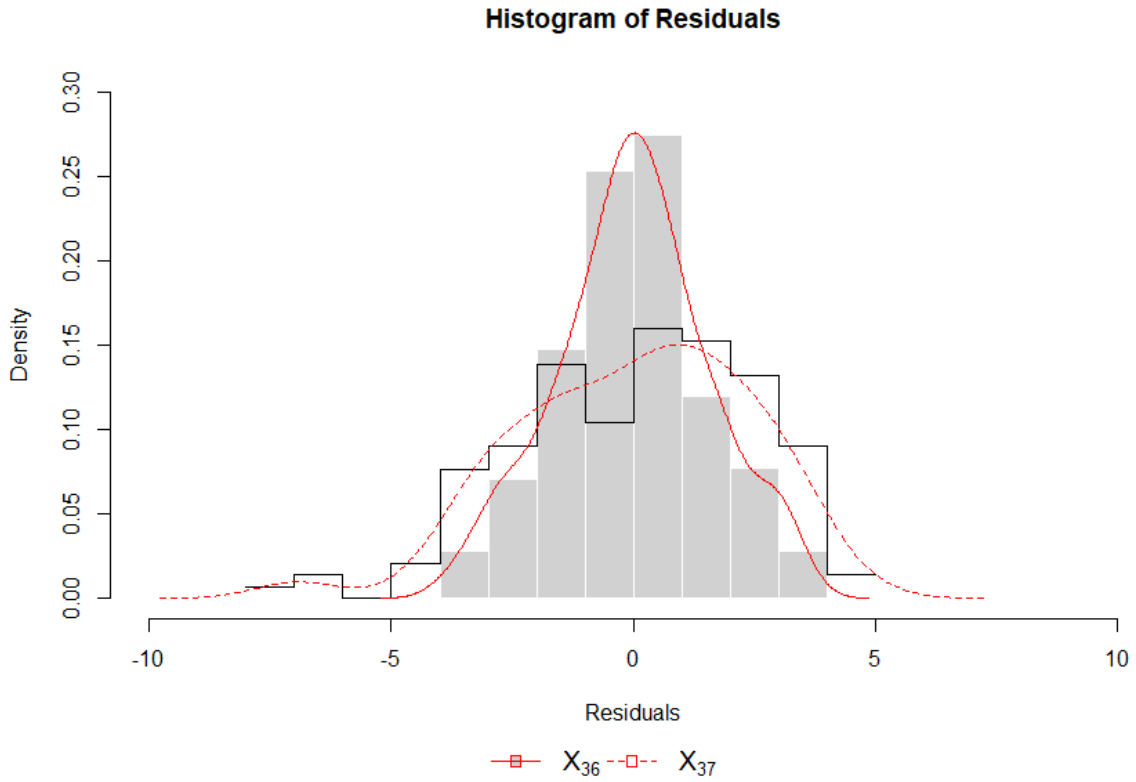


Figure 4.11: Histogram of residuals for both response variables (without outliers).

Lastly, the right most plots refer to the influential observations obtained through Cook's distance, and it's possible to observe that there are still quite a few of these observations on both response variables, where the variable X_{36} is clearly the one that evidences the presence of these observations in a larger quantity.

In conclusion, the models obtained by removing the outlier observations exhibit, overall, better results, with an inferior error compared to the previous models, even though the maximum value for the residuals is higher when compared to the previous one.

4.8 Modeling of the variable X_{38}

The variable X_{38} is qualitative with two levels, 0 and 1, and, as previously mentioned, it classifies if a tire passed or failed the test to which it was subject. If the tire passed the test, this variable assumes the value 1, and 0 otherwise. As shown 4.2, out of the 144 tires studied in this database, 115 of them passed the test, and the remaining 29 failed.

Given the nature of this variable, the most advisable methodology to apply in this case is the logistic regression.

The variables used for the modeling of this variable are all the previously used, removing the ones used as response variables for the linear regression X_{36} and X_{37}. Therefore, some of the exploratory analysis of these variables can be consulted on the section 4.2.

The quantitative variables were subject to the two-sample Wilcoxon-Mann-Whitney test, in order to test if the median points are equal on both samples ($H_0 : \eta_i = \eta_j$). In order to evaluate this hypothesis for all the quantitative variables were obtained the following set of plots.

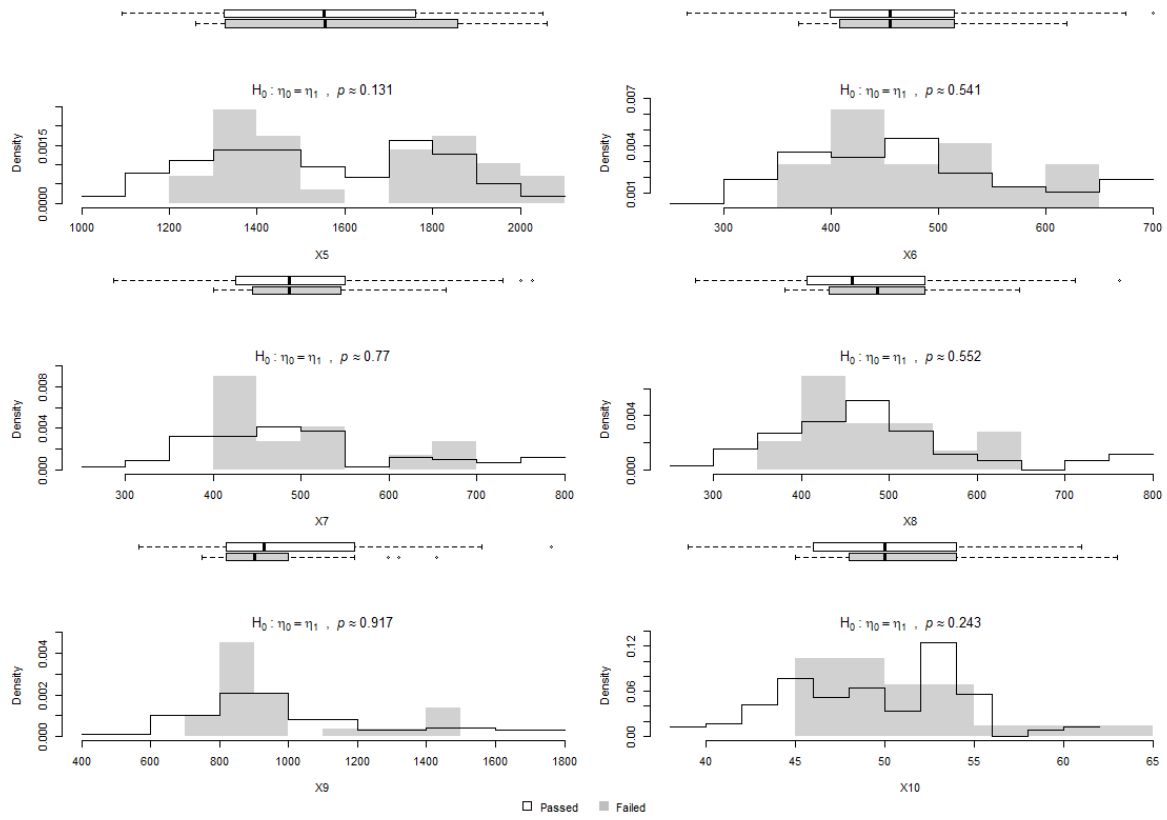
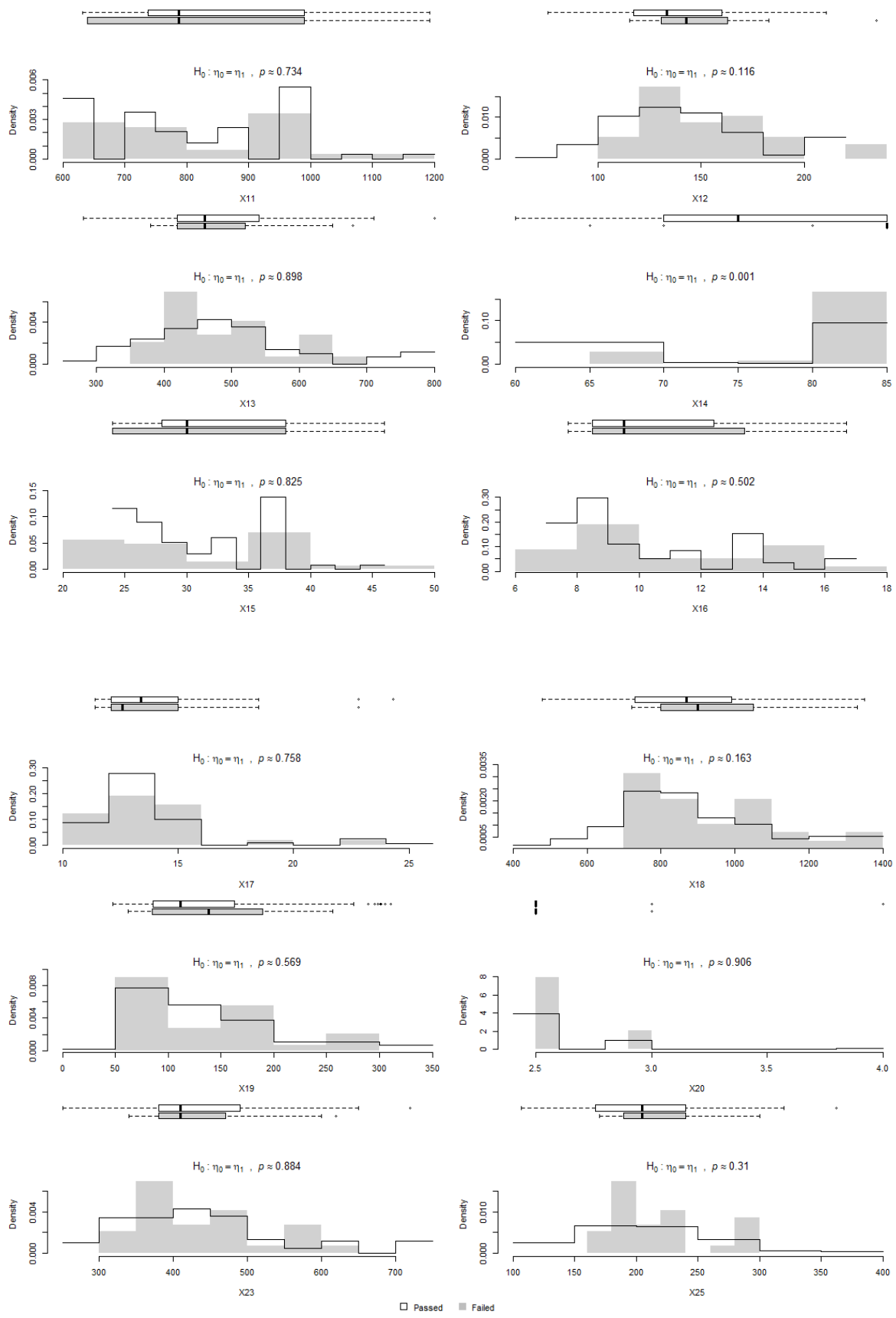


Figure 4.12: Histograms and boxplots for the quantitative variables and p-values for the Wilcoxon-Mann-Whitney two-sample test.


H0 : η0 = η1 , p ≈ 0.734
Density
X11
H0 : η0 = η1 , p ≈ 0.116
X12
H0 : η0 = η1 , p ≈ 0.898
X13
H0 : η0 = η1 , p ≈ 0.001
X14
H0 : η0 = η1 , p ≈ 0.825
X15
H0 : η0 = η1 , p ≈ 0.502
X16
H0 : η0 = η1 , p ≈ 0.758
X17
H0 : η0 = η1 , p ≈ 0.163
X18
H0 : η0 = η1 , p ≈ 0.569
X19
H0 : η0 = η1 , p ≈ 0.906
X20
H0 : η0 = η1 , p ≈ 0.884
X23
H0 : η0 = η1 , p ≈ 0.31
X25
Passed
Failed

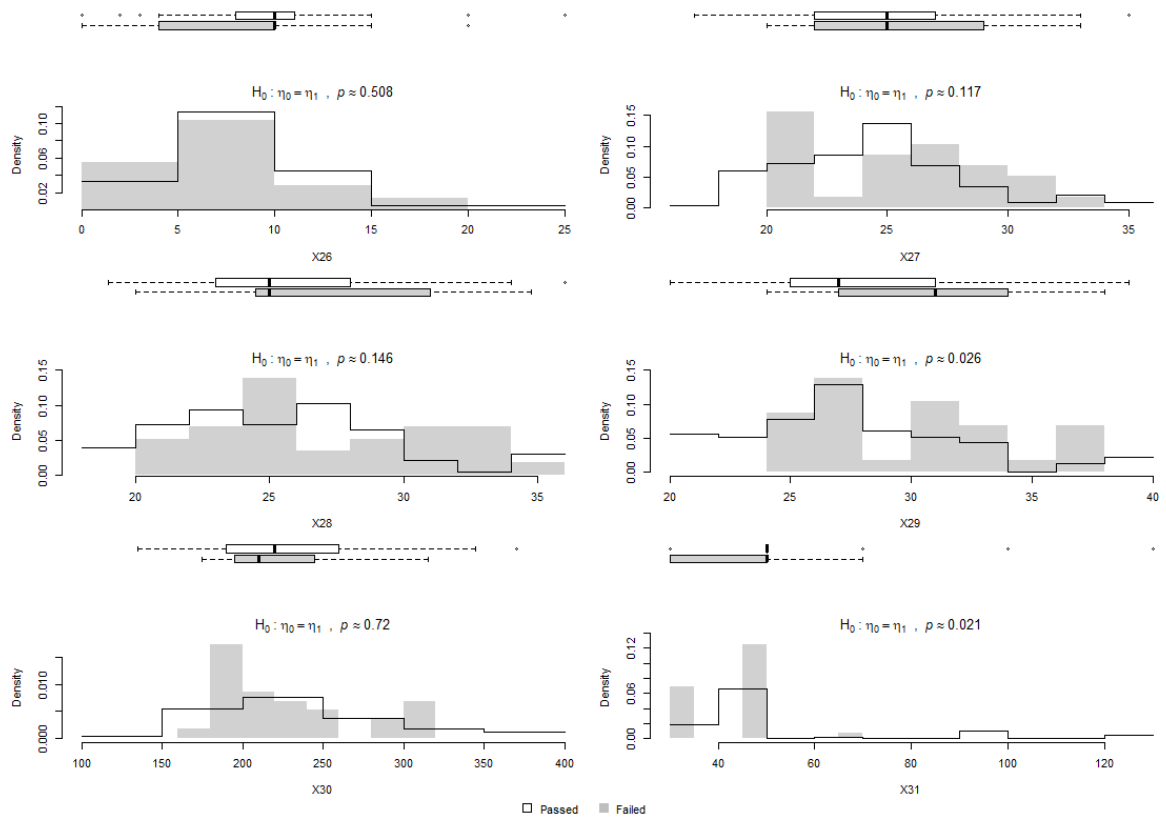


Figure 4.13: Histograms and boxplots for the quantitative variables and p-values for the Wilcoxon-Mann-Whitney two-sample test.

The results of the Wilcoxon-Mann-Whitney two-sample test referenced in the previous plots are presented in the following table in a more descriptive manner [23].

Table 4.12: Results of Wilcoxon-Mann-Whitney two-sample test.

Variables	p-value	Decision	
X_5	0.13	H_0	
X_6	0.54	H_0	
X_7	0.77	H_0	
X_8	0.55	H_0	
X_9	0.92	H_0	
X_{10}	0.24	H_0	
X_{11}	0.73	H_0	
X_{12}	0.12	H_0	
X_{13}	0.90	H_0	
X_{14}	0.01	H_1	*
X_{15}	0.83	H_0	
X_{16}	0.50	H_0	
X_{17}	0.76	H_0	
X_{18}	0.16	H_0	
X_{19}	0.57	H_0	
X_{20}	0.91	H_0	
X_{23}	0.88	H_0	
X_{25}	0.31	H_0	
X_{26}	0.51	H_0	
X_{27}	0.12	H_0	
X_{28}	0.15	H_0	
X_{29}	0.03	H_1	*
X_{30}	0.72	H_0	
X_{31}	0.02	H_1	*
X_{32}	0.94	H_0	
X_{33}	0.46	H_0	
X_{35}	0.58	H_0	

$^* < 0.05$

Upon analyzing the p-values obtained for all the variables, the null hypothesis was rejected on three different occasions, or more precisely, for three different variables, being X_{14}, X_{29} and X_{31}. Therefore, it's possible to conclude that there's statistical evidence that for those three variables the median points on both samples are not the same. The same conclusion can be obtained by performing a graphical analysis upon the plots presented on the Figure 4.13.

4.8.1 Variable selection

Just like previously studied, in order to obtain the best possible model for the response variable several methods were used to minimize the number of variables, maintaining only those considered to be important in explaining the dependent variable.

In this section three methods were used, two of which were already mentioned and used on the chapter regarding the linear regression, being the stepwise and backward.

The third one is a new method not used so far in this manuscript, the Hosmer and Lemeshow method for variable selection.

In order to find the best possible fitted model for the response variable besides using these three different methods for variable selection were also used two different link functions, the *logit* and the *probit*.

Firstly, all the models were calculated for both of the link functions using the three different methods previously mentioned, and then all of them were compared in order to determine which one was the right choice, ending with one model for each link function.

Table 4.13: Comparison of models obtained through the methods: Backward, Forward, Stepwise and Hosmer-Lemeshow.

Link Function	**Method**	$\mathbf{p^1}$	**p-value**
Logit	Stepwise	27	≈ 0.00
	Backward	**16**	
	Hosmer-Lemeshow	7	≈ 0.01
	Backward	**16**	
Probit	Stepwise	25	≈ 0.00
	Backward	**11**	
	Hosmer-Lemeshow	7	≈ 0.00
	Backward	**11**	

p^1 - number of parameters;

Looking at the results presented on the table, obtained through the likelihood ratio test, using a Chi-Squared distribution, it's possible to see that for both link functions, the choice fell upon the model obtained through the backward selection method. This choice was made because, when compared to the stepwise models, this one displays considerably less variables. When comparing it to the models obtained through the Hosmer-Lemeshow method it's possible to see that the latter presents less variables, however, the variables that were removed are considered to be important to the study, therefore the choice fell once again upon the model obtained through the backward selection method.

The coefficients obtained for each model are then presented in Table **??**.

Table 4.14: Coefficients for the chosen adjusted models using the *logit* and *probit* link functions.

Variables	**Level**	**Model 1 (*Logit*)**	**Model 2 (*Probit*)**
Slope	–	28.07	35.86
X_5	–	0.12 *	–
X_6	–	–	−0.11 *
X_8	–	–	0.08 *
X_9	–	0.08 *	–
X_{11}	–	−0.13 *	–
X_{12}	–	−0.33 *	–
X_{14}	–	–	−0.31 *
X_{18}	–	−0.14 *	–
X_{19}	–	0.19 *	0.05 *
X_{22}	2	−12.39	–
	3	−9.99	–
	4	−16.72	–
	5	−2.94	–
X_{25}	–	–	−0.20 *
X_{28}	–	−0.66 *	–
X_{29}	–	−0.60 *	–
X_{30}	–	–	−0.19 *
X_{32}	–	–	−0.82 *
X_{33}	–	1.12 *	1.02 *
X_{34}	2	−8.30 *	−2.56 *
	3	−27.52	−1.21

* < 0.05

The next step is to naturally define which of the two could be considered the “best” model, using once again Akaike’s information criterion to compare both of them.

Table 4.15: AIC value for the models using *logit* and *probit* functions.

Model	**AIC**
Model 1 (*Logit*)	91.90
Model 2 (*Probit*)	89.78

Analyzing the values in Table 4.15 the model using the *probit* link functions displays a slightly inferior value of AIC when compared to the one resultant from using the *logit* function. However, since the values are not so different, the choice of “best” model was made not only considering the AIC value but also the interpretability of the coefficients obtained. Therefore, with that in mind, the model chosen and to be considered from this point forward is the one obtained using the *logit* link function.

4.8.2 Goodness-of-fit

Similarly to the linear regression the model obtained calls for an analysis in order to ascertain its quality of adjustment and validity. In order to do that various metrics were

calculated, and will be discussed next.

Even though the model chosen was the one obtained through the utilization of the *logit* link function, the next metrics were calculated for both the models, in order to understand their behaviour and differences, and confirm the choice made.

ROC Curve

With the objective of studying the quality of adjustment of the models obtained one of the metrics oftenly used is the area under the ROC curve. The goal is to obtain a value of AUC as close to 1 as possible, translating into an outstanding discriminating power by the model.

The plots shown in Figure 4.14 represent the ROC curves obtained for both models.

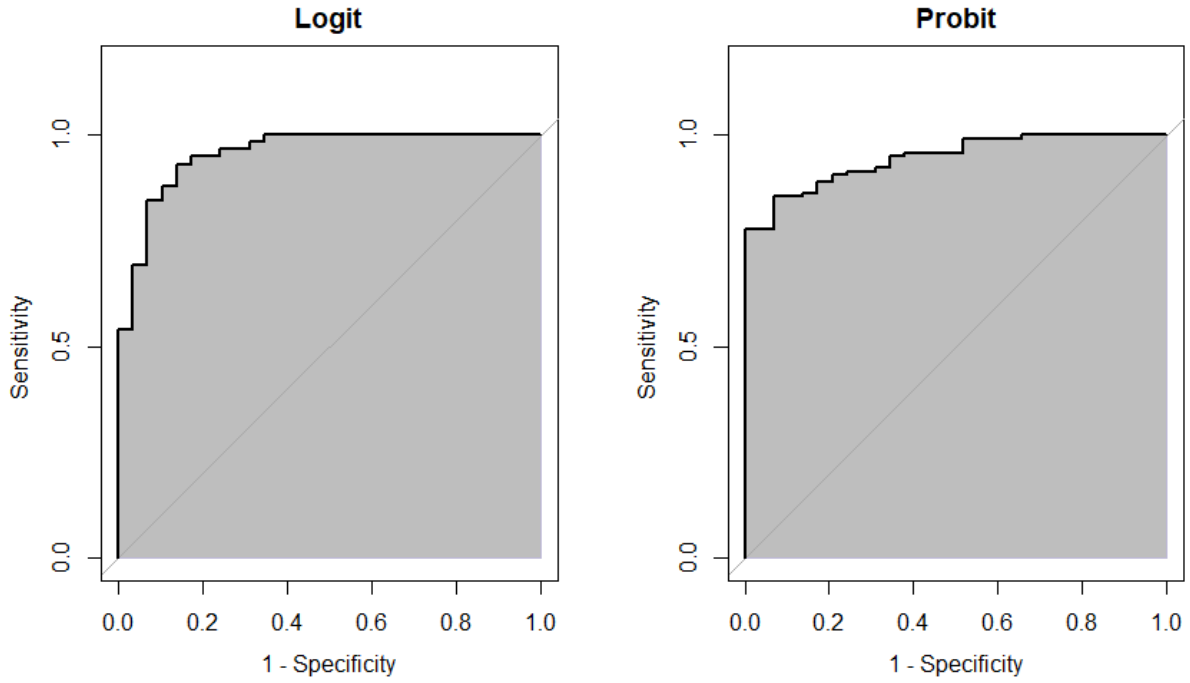


Figure 4.14: Area under the ROC curve for both models.

Just by analyzing the shape of the curves obtained for each model it's possible to observe that the curve corresponding to the *logit* model appears to be slightly higher than the one corresponding to the *probit* model, meaning that the value of the area under the curve will be consequently higher, and, therefore, the discriminative power of the respective model will be better than the other option.

Analyzing the curves value-wise, the area under the curve for the *logit* model is of 0.96, while the one for the *probit* model is 0.95, which means that while the values are extremely close, the *logit* model presents a slightly better outcome, confirming the conclusion drawn by the graphical analysis.

Prediction Error

Another way to evaluate the goodness-of-fit of the models is to analyze its prediction error. The determination of this prediction error is usually made utilizing a cut-off point

of 0.5, however, for this analysis, with the objective of obtaining a broader spectrum of results, the prediction error was calculated with various cut-off points, ranging from 0 to 1, in steps of 0.2 units.

The results obtained are presented in the Figure 4.15.

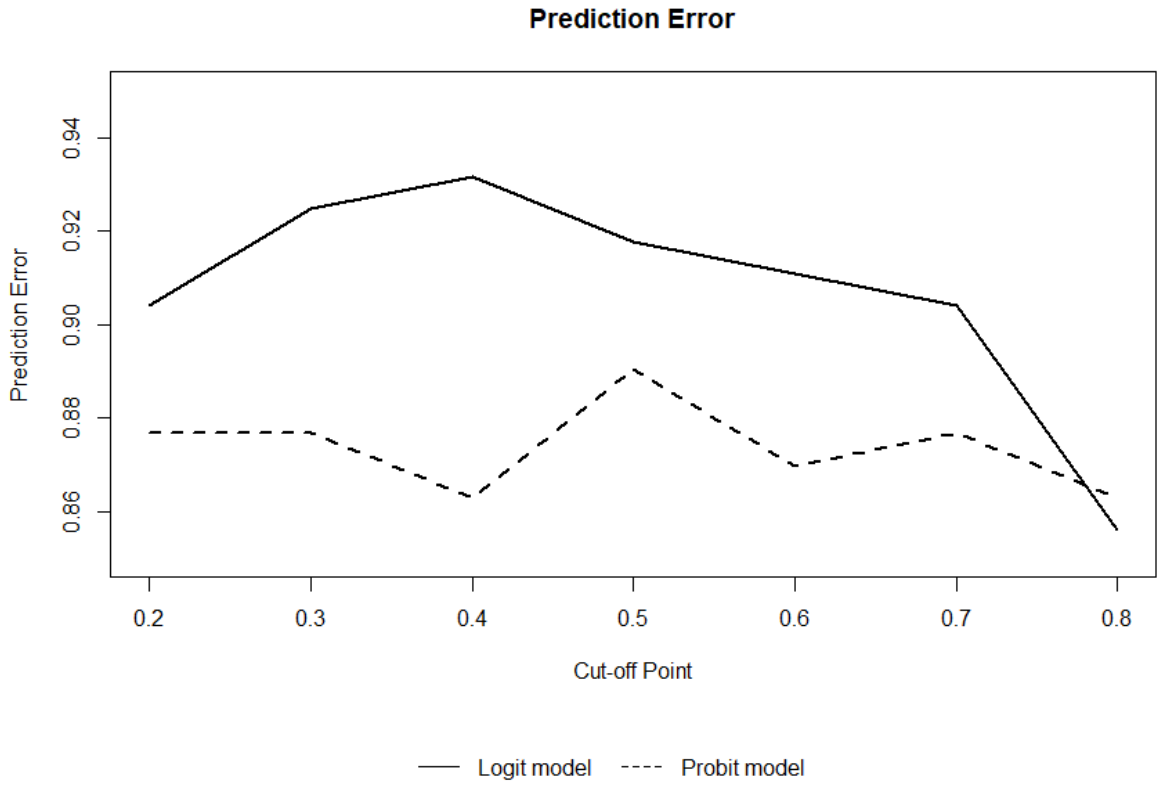


Figure 4.15: Prediction error with various cut-off points for both models.

Analyzing the plot it's clear that the model corresponding to the *logit* link function presents a clearly superior prediction error at most cut-off points, being lower than the prediction error of the *probit* model only when the cut-off point is superior to approximately 0.78. Therefore, looking just at these results, the *logit* model was also the right choice between the two options.

The side by side comparison of both models for these two metrics is then given by the Figure 4.16.

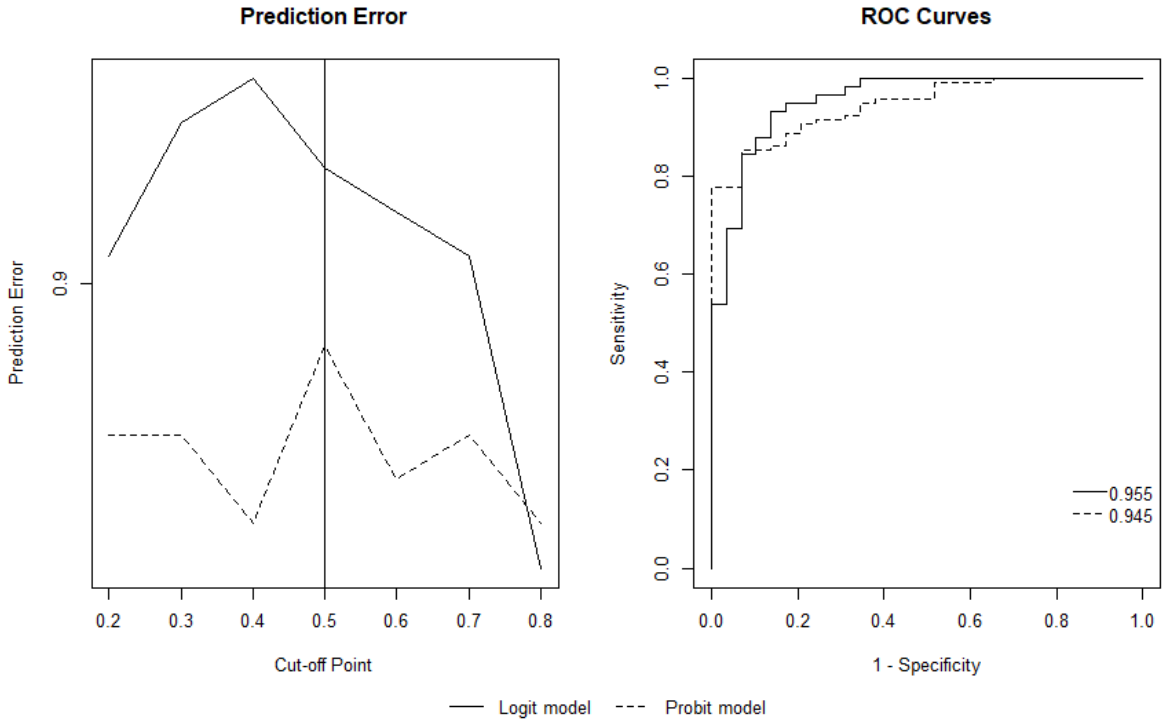


Figure 4.16: Side by side comparison of prediction error and area under the ROC curve for both models.

Hence, just by analyzing these two metrics it's possible to conclude that the choice of selecting the *logit* model as the "best" one was the correct one, being confirmed by both the area under ROC curve values and the prediction error values.

In order to calculate the prediction error was also necessary to calculate the specificity and sensitivity of the models, or in other words, the correct identification of true negatives and true positives, accordingly.

Table 4.16: Specificity, sensitivity and prediction error for both models.

	Model 1 (*Logit*)	**Model 2 (*Probit*)**
Specificity	0.16	0.15
Sensitivity	0.97	0.95
Prediction Error	0.92	0.90

Once again, analyzing the values presented in the Table 4.18 it's clear that the *logit* model can be considered as the model that offers a better goodness-of-fit, when compared to its alternative.

Brier's score

A third valid option to evaluate the quality of adjustment of a model is through the Brier's score. As mentioned before, the lowest the value of the score the better is the accuracy of the model, therefore, the model with the lowest score is the one that represents a better goodness-of-fit.

The values of Brier's score obtained for the *logit* and *probit* models are, accordingly, 1.06 and 1.08. Hence, once again, the *logit* model is the one that can be considered to have a best fit to the data.

The set of plots shown in Figure 4.17 are a visual representation of the estimated probabilities for the observations that were correctly and wrongfully classified by the model.

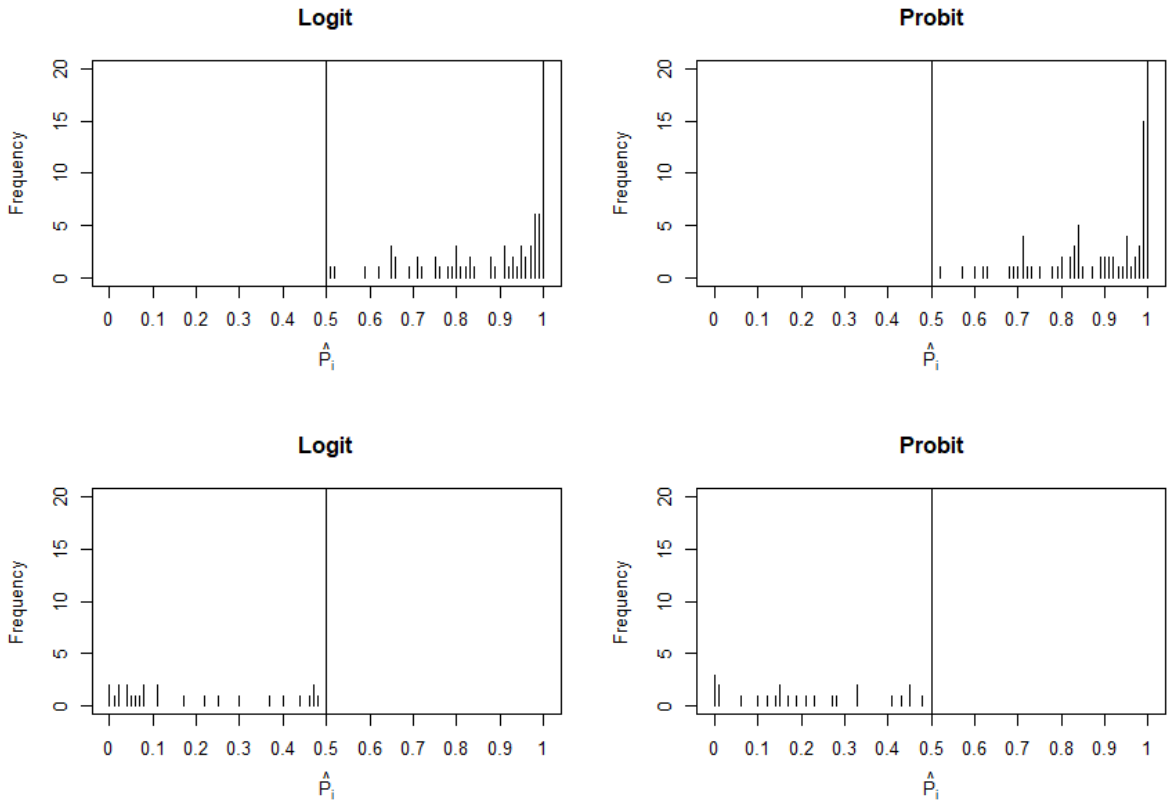


Figure 4.17: Top: Estimated probabilities for the observations classified correctly; Bottom: Estimated probabilities for the observations wrongfully classified.

With the top graphics the objective is to have as many observations as possible close to the value of probability 1 or 0 (the extremes), meaning that the degree of certainty that the observation was correctly classified was high, and the exact same for the opposite. It's possible to see that this happens in both graphics, all the values are mostly concentrated between 0.8 and 1, which indicates a good result.

On the other hand, for the observations that were wrongly classified the objective is to have the observations concentrated near the middle of the graphic, on the cut-off point 0.5, so that the result indicates uncertainty. However, on both the graphics this is not the case, given that although some of the values are near 0.5 some others are close to 0, which doesn't represent a good result.

Still, upon analyzing each pair of graphics for both the models, it's clear once again that the best choice is the one using the *logit* link function, given the reasons previously mentioned.

Hosmer-Lemeshow test

Given the fact that the models have 16 (*logit*) and 11 (*probit*) parameters, the number of groups used to calculate the Hosmer-Lemeshow statistic was 17, following the criteria proposed by the developers of the test, and previously mentioned. However, in order to understand what values this statistic would present, it was calculated for some other number of groups, as shown in Table **??**.

Table 4.17: Results of the Hosmer-Lemeshow test for different number of groups for both models.

Link Function	Groups	p-value
Logit	10	2.63×10^{-1}
	11	9.31×10^{-6} *
	12	4.36×10^{-1}
	13	3.23×10^{-3} *
	14	5.03×10^{-1}
	15	1.14×10^{-2} *
	16	4.53×10^{-7} *
	17	9.31×10^{-2}
	18	4.13×10^{-6} *
	19	7.65×10^{-2}
	20	2.63×10^{-6} *
Probit	10	5.97×10^{-1}
	11	9.80×10^{-1}
	12	9.98×10^{-1}
	13	9.99×10^{-1}
	14	9.98×10^{-1}
	15	9.20×10^{-1}
	16	7.23×10^{-1}
	17	9.65×10^{-1}
	18	9.97×10^{-1}
	19	9.99×10^{-1}
	20	9.58×10^{-1}

$^* < 0.05$

Looking at the values on the table, it's clear that, when it comes to the *probit* model, there isn't statistical evidence to reject the null hypothesis that the model is well adjusted to the data, considering any number of groups.

On the other hand, when it comes to the *logit* model, there are some number of groups that reject that hypothesis, namely, 11, 13, 15, 16, 18 and 20. However, for the number of groups considered (17) given the number of variables, the null hypothesis wasn't rejected, therefore there's statistical evidence that the model is well adjusted to the data.

Pseudo R^2 (McFadden & Cox-Snell)

Two other metrics to evaluate the quality of adjustment of the models are the pseudo R^2 coefficients, particularly the McFadden and the Cox-Snell coefficients. The values for

these two coefficients, as mentioned before, are considerably lower than the usual R^2 obtained in OLS, therefore a lower value is to be expected.

Table 4.18: Pseudo R^2 coefficients for both models.

	Model 1 (*Logit*)	**Model 2 (*Probit*)**
McFadden	0.59	0.53
Cox-Snell	0.44	0.41

As it was expected the values obtained are considerably lower when compared to the usual R^2, being 0.59, 0.53 and 0.44, 0.41, for McFadden and Cox-Snell, accordingly.

Although the values obtained are lower than the usual R^2, in theory, in order to indicate a good fit the values should be slightly lower, between 0.2 and 0.4. However, the result is acceptable and it's possible to infer that the quality of adjustment is acceptable.

Overall

After analyzing all the quality of adjustment tests and metrics and their results it's possible to conclude that, overall, the quality of adjustment of both the models is acceptable, having obtained similar values in most of the tests and statistics.

Hence, given that information, it's possible to conclude that choosing the model obtained through the *logit* link function was justified and correct.

4.8.3 Residual Analysis

Just like in the linear regression, the logistic one also calls for a residual analysis in order to understand the results obtained and the "behaviour" of the observations.

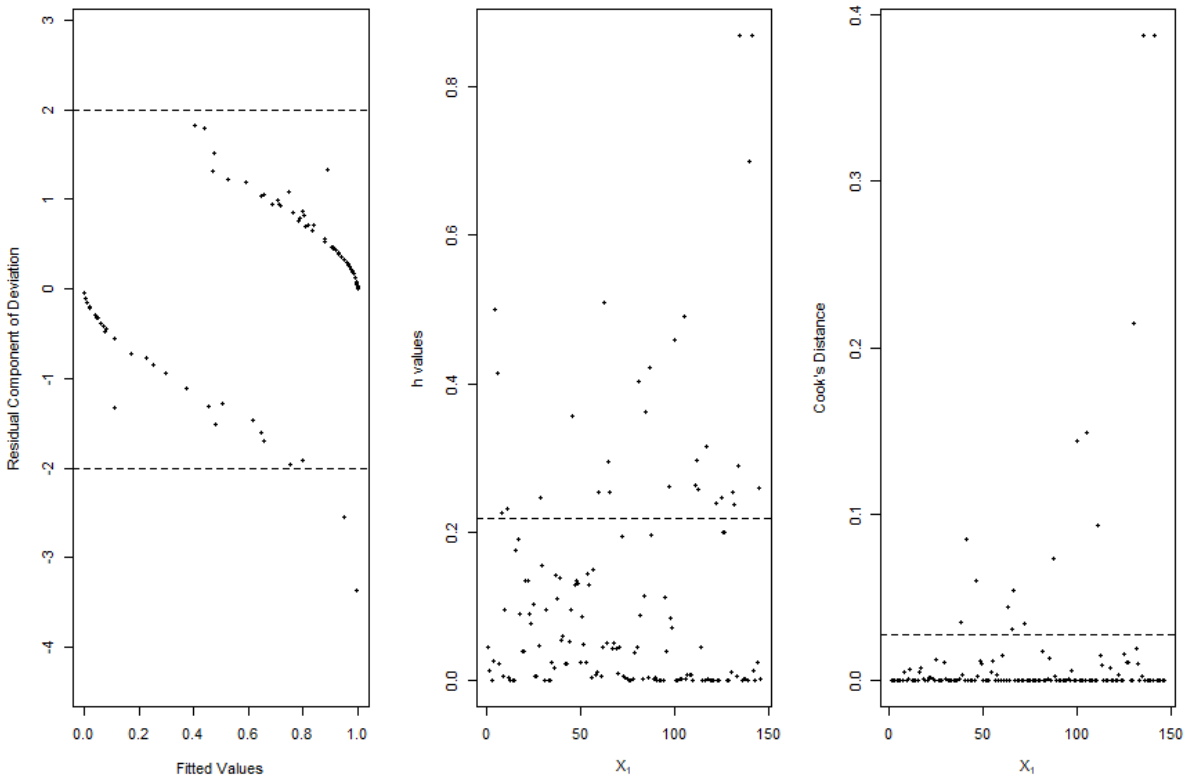


Figure 4.18: Left: Fitted values vs deviation (outliers); Center: Leverage points; Right: Influential observations.

Analyzing the left plot of the Figure 4.18, it's possible to observe the presence of two outliers, corresponding to the observations 41 and 130, while the remaining values are within the expected ranges.

When it comes to the middle plot, it's observable the existence of quite a few leverage points, utilizing the same criteria previously used on the residual analysis for the linear regression models.

Lastly, the third plot of the Figure 4.18 represents the influential observations, and it's observable that there are 14 observations that can be classified as such.

When it comes to the normality of the residuals, utilizing the Lilliefor's normality test, the p-value indicates that the normality is not verified ($p - value < 2.2 \times 10^{-16}$).

4.8.4 Interpretation of the coefficients

The estimated model allows for the interpretation the chance of a tire passing the dimensional tests when compared to another tire with different features.

The estimated odds ratio associated to the different tires, according to each variable, are presented in the Table 4.19.

Table 4.19: Estimated odds ratio for each variable.

Variables	Level		Odds Ratio
Slope	—		1.55×10^{12}
X_5	—		1.13
X_9	—		1.08
X_{11}	—		0.88
X_{12}	—		0.72
X_{18}	—		0.87
X_{19}	—		1.21
X_{22}	2	$\approx$	0.00
	3	$\approx$	0.00
	4	$\approx$	0.00
	5		0.05
X_{28}	—		0.52
X_{29}	—		0.55
X_{33}	—		3.07
X_{34}	2	$\approx$	0.00
	3	$\approx$	0.00

Next it's done the interpretation of the average impact of the isolated variation of the features that describe a tire when it comes to its chance of passing the tests. Therefore, from now on, it's made the assumption that when comparing two tires they possess the same technical features regarding the variables contemplated in the model, except for the variable being analyzed.

Analyzing just the variable X_5 it's possible to conclude that an unitary increase in its value increases the average chance of a tire passing by approximately 13%. Similarly, the thought process can be applied to the variable X_9, where a positive unitary change will increase the average chance of a tire passing by approximately 8%.

On the other hand, upon looking at the variables X_{11}, X_{12} and X_{18} it's possible to see that the same unitary increase will have the opposite effect on the average chance of a tire passing, meaning that, said chance will decrease approximately 12%, 28% and 13%, accordingly.

The variable X_{19}, similarly to the first two previously mentioned, presents an average chance roughly 21% higher for every unitary increase.

Regarding the variable X_{22}, a tire with value within the level 5 has an average chance of passing the tests approximately 95% lower when compared to tires with the other 4 levels.

Similarly to the variables X_{11}, X_{12} and X_{18}, the variables X_{28} and X_{29} have an average chance approximately 48 and 45% lower for each unitary increase on the variable.

Lastly, with the variable X_{33}, an unitary increase will make the chances of the tire passing approximately 3 times higher.

5 Shiny Application

Shiny is an *R* package that makes it easy to build interactive web apps straight from *R*. This method combines the computational power of *R* with the interactivity of the modern web.

The Shiny package comes with eleven built-in examples that demonstrate how Shiny works, so that even beginners can pick it up easily and understand the concept behind the package.

To put it simply, a Shiny application is simply a directory containing an *R* script called *app.R*, which is made up of a user interface object and a server function. This folder can also contain any additional data, scripts, or other resources required to support the application.

As previously mentioned, the application is divided into two different components, the interface object and the server function.

The interface object is the bit of code that is going to define how the application is going to look like, as well as what inputs it's going to require from the user in order to accomplish a certain task.

On the other hand, the server function is responsible for the "calculations" that the application is going to do. In other words, it's the segment of code that decides how the inputs introduced by the user are going to be utilized, and also how the result is going to be presented as an output (as a table, plot, etc.), keeping in mind that the aesthetics of the outputs are controlled by the interface object, just like the overall look of the application.

After the application is completed there are several ways to make it possible for outside users to utilize the application.

One of them is to host the application on the *R* public servers, for example, making it available to use right away.

Further along this manuscript will be mentioned some of the features available on this kind of application, and which ones were implemented.

Interface

The overall interface of the application is as shown on Figure 5.1.

On the top side of the application there are seven tabs, each with its own purpose. The first two represent two possible simulation results, each with its own tab. The next three, named database 1 through 3, like the name hints, are three distinct databases regarding test results for all the ETOs available up until the date, this for three different tests. Finally,

the last two tabs represent two databases created to store the input data introduced for a certain simulation and its corresponding result.

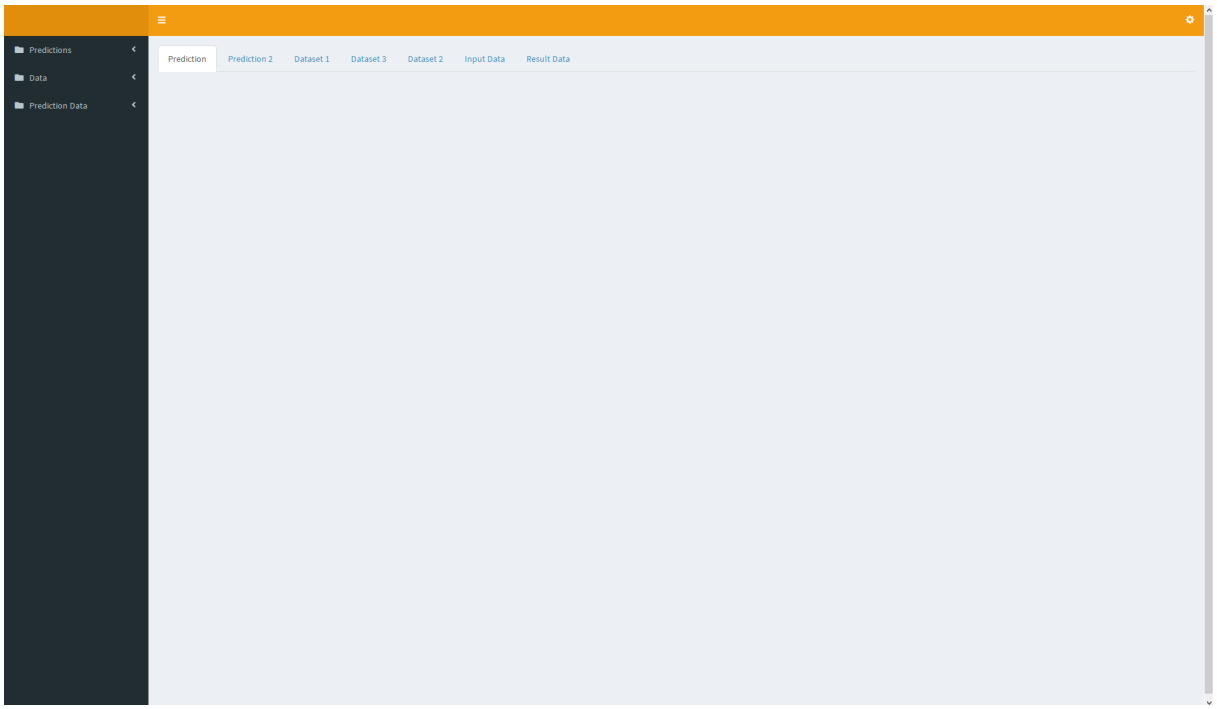


Figure 5.1: Main interface of the application.

Inputs

The application requires 35 inputs, two of them just for identification purposes and the remaining 33 in order to calculate the prediction. After filling in all the values required there are three options, as shown on Figure 5.2. The first option, “Test”, calculates the value of the prediction. This action requires all the empty camps to be filled.

After having decided upon all the values of inputs, the second option allows to save them and the prediction results in a new database, consisting in one of the tabs previously mentioned.

Lastly, the third option allows the user to save the prediction result as a PDF file.

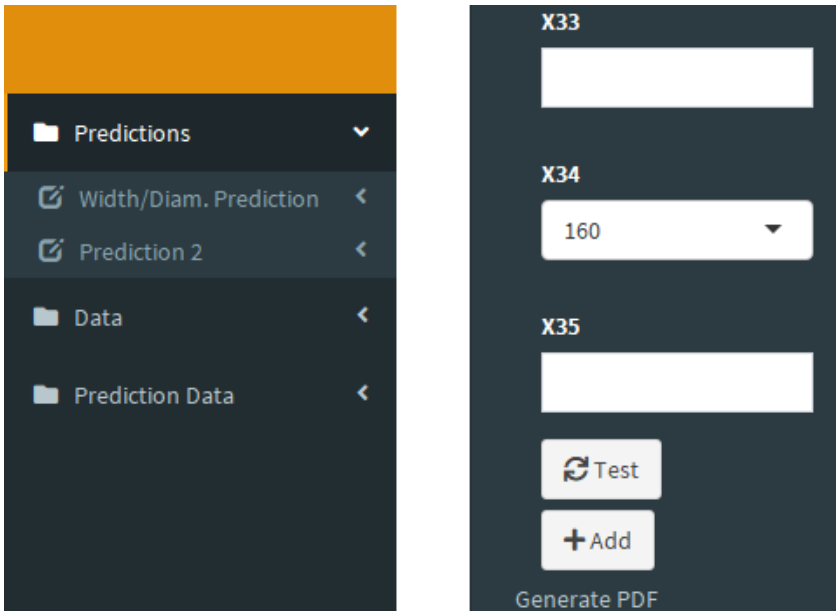


Figure 5.2: Left: sidebar options; Right: input camps examples and action options.

Prediction

After filling all the input camps and testing for the result the display is then given as shown on the Figure 5.3.

Figure 5.3: Left: sidebar options; Right: input camps examples and action options.

The first three values presented as “NA” are used for labelling, and therefore aren’t correctly filled in this example.

The next value is the result for the simulation of one of the response variables, followed by the maximum allowed for that specific tire. The same happens for next two values, but now regarding the other response variable.

The next two columns represent the evaluation of the results obtained. Here is made a comparison between the prediction result and the maximum allowed, where the value

on this collumn is displayed as “PASSED” if the result is lower than the maximum and “FAILED” otherwise.

The last collumn is an overall evaluation, displaying “PASSED” if both the predictions are lower than the maximums presented, and “FAILED” otherwise.

Data storage

Other two options available on the application are to store the input data and the results obtained, as well as access to databases from the existing ETOs, as shown on Figure 5.4.

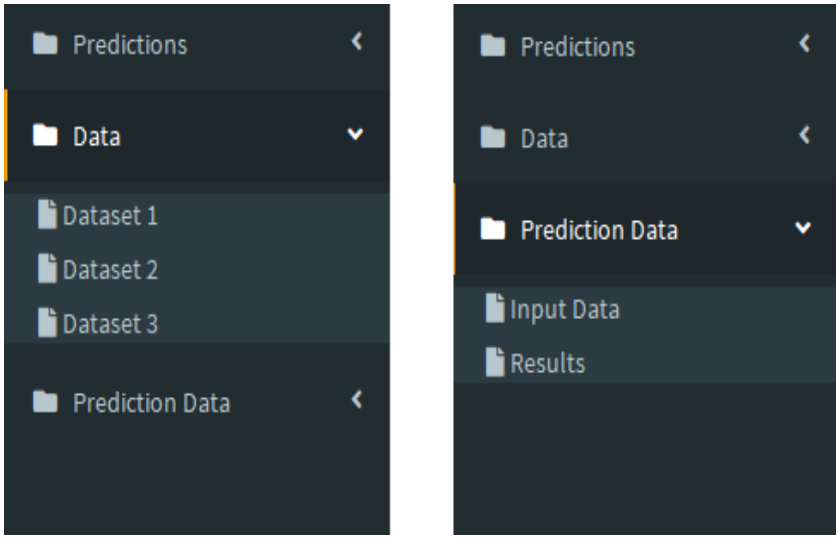


Figure 5.4: Left: Databases tab; Right: input and results storage tab.

The databases are representative of the different tests realized by the tires so far, while the “Prediction Data” tab, like the name suggests, corresponds to the storage of all the data regarding the inputs and the corresponding results.

6 Conclusions and Future Research

Conclusions

The major aim of this dissertation was to understand the impact of all the variables regarding the tire's mold and specifications in the final dimensional results. Another major objective was to develop a tool capable of predicting said final result for each tire.

With this in mind, the second part of this dissertation (Chapter 3) consists on the introduction of some theoretical concepts regarding Linear Regression and Logistic Regression, while the first part consists on a brief introduction of the company were the internship was realized, as well as the explanation of the various types of construction in existence, the components of a tire, and the several steps since the beginning to the end of the construction.

The fourth chapter portrays the development of the regression models in order to fulfill the objective previously presented, as well as the evaluation of said models to understand their validity and quality of adjustment, with the ultimate goal of understanding what were the most important variables and to also understand if the models had the requirements to be used for future simulations and predictions. The main conclusions to be drawn out of this chapter, regarding the linear regression, are the two models obtained, one for each of the response variables, as well as the coefficients obtained for each of the significant variables, which made it possible to understand the importance of each of them on the final dimension of the tire.

Looking into the coefficients obtained for all the variables it was possible to understand if the values made sense, and if they were in agreement with the existing knowledge. In fact, the results obtained confirmed some of the existing suspicions. For example, the coefficients obtained for the variables X_{18}, X_{22} and X_{24}, mainly, could be considered correct according to existing results. The variable X_{34} is another example of that, where tests were realized and the value of the coefficients for each factor of the variable was confirmed by said tests (this regarding the response variable X_{37}).

On the other hand, some of the variables that were considered significant by the models came as a surprise, such as X_7, X_{16} and X_{17}. The first one, upon further analysis, came to be considered one of the most important variables when it comes to the determination of one of the dimensional results of the tire, and was overlooked up until that point. As for the last two, they also came as an unexpected result, and therefore, in order to confirm the results, two sets of tires were made, where the only difference would be the value of those variables, with the purpose of directly comparing and understanding

its importance. The results obtained confirmed that X_{16} and X_{17} were in fact affecting the dimensions of the tire, however, the coefficient obtained initially for both of them was slightly higher than the true value.

Another result obtained was the correlation between the value of the variable X_5 and the variable X_{34}. Upon analyzing all the values present in the variable X_5 and understanding how they varied taking into consideration the tire size, it was possible to establish a correlation between that and the values of the variable X_{34}. With this study it was possible to conclude that, according to the value assumed by the variable X_{34}, the variable X_5 needed to take an appropriate value, otherwise the tire would fail on one of the dimensional criteria. This conclusion came to be as one of the most relevant ones, explaining various unexpected results.

The chapter previously mentioned also contains the logistic analysis performed in order to understand what's the chance of a tire passing or failing the dimensional tests.

Lastly, the fifth chapter is an overall look of the tool developed, evidencing and explaining the multiple features available, as well as a demonstration of how the outputs are displayed.

It's also important to note that, all the knowledge obtained through the entirety of the internship, such as the models obtained, and the tool developed, were carefully documented, and guidelines were written, so that everything can continue to be used and updated by the employees, as need be.

Future Research

As future work, it would be interesting to analyze different variables, such as the ones existent in the production, and understand what's their impact on the dimensions of the tire.

Taking into account that the tires are submitted to various other tests besides the dimensional ones, it would also be relevant to realize a similar study, and obtain models, for those studies as well. Obtaining said models it would be possible to simulate if a given tire, with a certain mold and specification, would pass all of the tests needed to be certified and cleared for the selling market. Such analysis wasn't possible given the reduced size of data for these specific tests, which was approximately a third of the data used on this dissertation.

Another possibility of future work would be to add more features to the application, regarding the possible models previously mentioned, for example, and possibly an overall result taking into consideration all of the simulations done and with a final decision based on the results obtained for all of said tests.

Bibliography

[1] Legendre, A.M., *Nouvelles méthodes pour la détermination des orbites des comètes*, Paris, Firmin Didot, 1805.

[2] Gauss., C.F., *Theoria Motus Corporum Coelestium* in *Sectionibus Conicis Solem Ambientum.*, 1809.

[3] Hanley, J.A., *"Transmuting" Women into Men: Galton's Family Data on Human Stature.* The American Statistician, 58 (3), 2004, pp. 237-243.

[4] Dobson, A. J., *An Introduction to Generalized Linear Models.* Chapman e Hall, 3ed, 2008.

[5] Turkman, A., Silva, G., *Modelos Lineares Generalizados - da teoria à prática.* Lisboa: Sociedade Portuguesa de Estatística, 2000.

[6] PSY 510/610 Categorical Data Analysis, Fall 2016.

[7] Hosmer, D. W., and Lemeshow, S., *Applied Logistic Regression.* Wiley, 2013.

[8] Armstrong, J., *Principles of Forecasting: A Handbook for Researchers and Practitioners, International Series in Operations Research & Management Science*, Springer, 2001.

[9] Burnham, K. P., Anderson, D. R. *Model Selection and Multimodel Inference: A practical information-theoretic approach* (2nd ed.), Springer-Verlag, 2002.

[10] McFadden, D., *"Conditional logit analysis of qualitative choice behavior."* Pp. 105-142 in P. Zarembka (ed.), Frontiers in Econometrics. Academic Press, 1974.

[11] Domencich, T. and McFadden, D.L., *Urban Travel Demand: A Behavioral Analysis*, North-Holland Publishing Co., 1975. Reprinted 1996.

[12] Cox, D. R., and E. J. Snell., *The Analysis of Binary Data*, 2nd ed. London: Chapman and Hall, 1989.

[13] Pregibon, D. *et al.*, *Logistic regression diagnostics*, The Annals of Statistics 9 (1981), pp. 705-724.

[14] Allison, P.D., *Measures of fit for logistic regression*, in *Proceedings of the SAS Global Forum 2014 Conference.* 2014, pp. 1-13.

[15] Armstrong, J., *Principles of Forecasting: A Handbook for Researchers and Practicioners*, International Series in Operations Research & Management Science, Springer, 2001.

[16] Allison, P., *What's the Best R-Squared for Logistic Regression?*, webpage: https://statisticalhorizons.com/r2logistic. Accessed on July 2019.

[17] Columbus, A., *Introduction to R Shiny*, webpage: https://medium.com/@ODSC/introduction-to-r-shiny-b6acdf17c963. Accessed on May 2019.

[18] Borchani, H., *A survey on multi-output regression*, Machine Intelligence Group, Department of Computer Science, Spain.

[19] Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., *Applied Linear Statistical Models*, 5th ed. McGraw-Hill/Irwin, 2004.

[20] Turner H., *Introduction to Generalized Linear Models*, ESRC National Centre for Research Methods, UK and Department of Statistics University of Warwick, UK, 2008.

[21] Smith, T. J., McKenna, C. M., *A Comparison of Logistic Regression Pseudo R^2 Indices*, Northern Illinois University.

[22] R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[23] Happ, M., Bathke, A. C., Brunner, E., *Optimal Sample Size Planning for the Wilcoxon-Mann-Whitney-Test*, Department of Mathematics, University of Salzburg, Austria, 2018.