Analyzing Survival Factors of the Titanic Disaster: A Logistic Regression Approach

Maria Nascimento Cunha
Instituto Superior de Educação e Ciências (ISEC Lisboa -Portugal).
Member of the Scientific Council of CIAC – Centro de Investigação de Artes e Comunicação

Jorge Figueiredo
Universidade Lusíada, Famalicão, Portugal

Isabel Oliveira
Universidade Lusíada, Porto, Portugal

Manuel Maçães
Universidade Lusíada, Porto, Portugal

Silvia Costa Pinto
Universidade Fernando Pessoa

Abstract: Since that fateful and chilly dawn of April 15, 1912, the world has witnessed the construction of larger ships, some already dismantled or lying, solitary, in the darkness of the bottom of the oceans and others still in circulation. However, no other ship has become as famous and significant for popular naval and imaginary history as the Royal Mail Ship Titanic. The RMS Titanic joined the imaginary of the navy, literature and cinema. It fed the dreams and nightmares of generations, from the one from 1912 who was perplexed to receive the news of the disaster to the our present generation that has it in the ambivalence of an engineering feat of its time, as well as a fruit of the arrogance of its creators.

Its history is known to all and its data used in many studies. It should be mentioned that these data are composed of records of various variables and of various natures. In addition, they are easily generalizable to several other situations.

In this study the researcher will make use of the regression models. This model are one of the most important statistical tools in data analysis when the objective is to study relationships between variables, or more particularly, to analyze the influence that one or more variables (explanatory variables) may have on a variable of interest (response variable).

The purpose of this study is to describe in detail the construction of this type of model using a dataset on the Titanic tragedy.

Keywords: Titanic; Statistics; regression models.

Introduction

The RMS Titanic was a British passenger ship operated by the White Star Line and built by the Harland and Wolff shipyards in Belfast. The second vessel of the Olympic Class of ocean liners, after the RMS Olympic and followed by the HMHS Britannic, was designed by naval engineers Alexander Carlisle and Thomas Andrews. Its construction began in March 1909 and its launch into the sea took place in May 1911. The Titanic was thought to be the most luxurious and safest ship of its time, spawning legends that it was supposedly "unsinkable" (Pestana e Gageiro, 2014; Siena, 2019).

Its history is known to all and its data used in many studies. It should be mentioned that these data are composed of records of various variables and of various natures. In addition, they are easily generalizable to several other situations.

In this study the researcher will make use of the regression models. This model are one of the most important statistical tools in data analysis when the objective is to study relationships between variables, or more particularly, to analyze the influence that one or more variables (explanatory variables) may have on a variable of interest (response variable). The purpose of this study is to describe in detail the construction of this type of model using a dataset on the Titanic tragedy.

Regression models are one of the most important statistical tools in data analysis when the objective is to study relationships between variables, or more particularly, to analyze the influence that one or more variables (explanatory variables) may have on a variable of interest (response variable).

For this, several types of models can be considered, and the type of response variable involved has special interest in the choice of this model. Data analysis with binary response, which admits only two outcomes, is one such situation. In this case, it is intended to study the occurrence of a "success" taking into account the other variables of interest.

Thus, to describe a Logistic Regression model, in section 2, the methodology of an MLG will be introduced succinctly, introduced in the literature by Nelder and Wedderburn (1972) and Turkman and Silva (2000). Section 3 shows the particular case of the Logistic Regression Model. Section 4 describes in detail the construction of this type of model using a dataset on the Titanic tragedy. Its history is known by all and its data used in many studies, because it is a good set because it contains records of various variables and of various natures, in addition to that, it is easily generalizable to several other situations. There are different versions of this dataset available for free online, and the train set has been used. CVS available at the link "https://www. Kaggle. com/c/titanic/data". This contains data on 891 passengers and among other variables, contains one called " Survived " which takes the value 1 if the passenger survived or 0 otherwise. If we consider as an objective, to predict the survival of the individual taking into account, for example, the class where he traveled, sex, age, etc., we are faced with the intention of modeling a dichotomous response variable taking into account a set of explanatory variables, so that they are suitable for the adjustment of a logistic regression model (Afonso & Nunes. 2019).

The choice of this data set is due to the fact that it contains records of several variables and of various natures, in addition to which, it is easily generalizable to several other situations.

R statistical package will be used to support the resolution of the example. This is followed by the conclusion in section 5 and the bibliographical references in section 6.

Regression models

As already mentioned in this case, it is intended to study the occurrence of a "success" taking into account the other variables of interest. There are several practical situations in which this type of response appears (Pestana e Gageiro, 2014; Negas. 2021). Examples are:

- the result of the diagnosis of a laboratory test − positive or negative;
- the result of the inspection of a newly manufactured part − defective or non-defective;
- a voter's opinion of voting for a particular candidate – favorable or unfavorable;
- the result of a promotion of a chain of stores by sending each customer a coupon or discount code − used or unused;
- Credit granting models, where from information offered by the applicant, the financial institution decides whether or not to release the credit.
- the result of a knowledge test with a minimum grade to be approved − approved or not approved;

This last example also illustrates situations in which only two possibilities are considered of interest for a continuous variable, values less than a reference value and values greater than or equal to that value. In these cases, a new binary variable is considered for these two possibilities. In this way, binary variables can be existing variables in a study or can be created if there is interest.

Thus, when the objective of the study is to explain a binary response variable, linear regression, whose variables are continuous in nature, does not represent the most appropriate model. One of the particular cases of Generalized Linear Models (MLG) are models where the response variable presents only two categories or that has somehow been dichotomized, and the Logistic Regression model is the most popular of these models because it takes into account the fact that the response variable is categorical and the explanatory variables can be continuous or categorical (Pestana e Gageiro, 2014; Negas. 2021).

Brief considerations on a generalized linear model

Generalized linear models (MLG) constitute a class of statistical models and generalize classical linear models allowing the inclusion of many other models considered useful in statistical analysis. These have as main objective to study the relationship between variables, more particularly, to analyze the influence that one or more explanatory variables, independent or covariate, measured in individuals or objects, have on a variable of interest called response or dependent variable.

Consider that there are n experimental units and that the measurements refer to p explanatory variables that are believed to explain part of the variability inherent in Y, the response variable.

Description of the Generalized Linear Model

Generalized linear models, as already mentioned, are an extension of the classical linear model.

$$Y = X\beta + \varepsilon$$

where X is the n-× $(p+1)$ matrix of model specification n, associated with a vector $\beta = (\beta 0, \dots , \beta p)T$ of parameters and ε is a random error vector with distribution that is assumed Nn (0; σ2I).

These hypotheses imply that $E(Y|X) = \mu$ with $\mu = X\beta$, that is, the expected value of the response variable is a linear function of the covariates.

To specify a generalized linear model, three components are needed: the *random component of the model* (identifies the distribution of the response variable and can have any distribution belonging to the exponential family), the *systematic component* (consisting of a linear predictor that is a linear combination of the explanatory variables) and a *linkage function* (which combines the two previous components, establishing a relationship between the parameters of the distribution and the explanatory variables). Like this:

Componente Aleatória

Given the vectors of covariates $x's$, the variables Yi are (conditionally) independent and have distribution belonging to exponential family with mean value $E(Yi|xi) = \mu i$.

Note that the response variable to be studied is binary thus, let Y be the response variable with Bernoulli distribution, and a sample $y1, \dots , yn$ of that distribution that can take only two values, assigning $yi = 1$ to the event of interest and $yi = 0$ to the complementary event, called "success" and "failure" respectively and whose probability function is given by

$$f(y_i|p_i) = p_i{}^{y^i} (1 - p_i)^{(1-y^i)}, \ y_i = 0{,}1;\ i = 1, \dots , n \qquad (3.1)$$

where pi is the unknown parameter, which means the probability of success, i.e. $P(Yi = 1) = pi$. Thus, the probability of failure is given by $P(Yi = 0) = 1 - pi.$

It is intended to formulate a model for the probability of an object or individual characterized by a vector of explanatory variables (x) take the value 1, that is, to formulate a model for the mean value of the response variable Yi , which corresponds to $P(Yi = 1|xi)$.

Systematic or Structural Component

Assume that $xi1$, $xi2$, . . . , xip represent the values of p explanatory variables referring to the i-th individual. The systematic component allows the elaboration of a linear model in the explanatory variables, called a linear predictor and represented by the vector of parameters $\eta = (\eta 1, \eta 2, \dots \eta n)T$ *where*:

$$\eta_i = \sum_{j=0}^{p} \beta_j x_{ij} = \boldsymbol{x}^T\boldsymbol{\beta}, \quad i \qquad i = 1, \dots, n$$

$\boldsymbol{xi} = (xi0, xi1, xi2, \dots , xip)$ tendo-se $xi0 = 1$. The linear predictor can be matrix-represented by:

$$\eta = X\beta$$

The systematic component is said to be continuous if the explanatory variables are continuous and categorized if the explanatory variables are discrete. When the systematic part of the model consists of both continuous and discrete covariates, the systematic component is said to be mixed.

Connection Function

The third component of an MLG, the binding function, is the link link $\eta_i = g(\mu_i)$ which describes the functional relationship between the systematic component and the expected value of the random component, where μ_i represents the mean of the response variable and g to a differentiable monotone function. It is this function that allows to make the relation linear, that is:

$$g(\mu_i) = \sum_{j=0}^{p} \beta_j x_{ij} = \boldsymbol{x}^T\boldsymbol{\beta}, \qquad i = 1, \dots, n. \qquad (2.1)$$

In this case, although there are other possible connection functions, the $função\ logit$, given by

$$\eta_i = \ln \left\{ \frac{\mu_i}{1-\mu_i} \right\}$$

Methodology of Generalized Linear Models

When trying to model data through an MLG, three essential steps must be followed: the formulation of the models, the adjustment of the models, and the selection and validation of the models (Afonso & Nunes. 2019). In a first phase, three factors have to be taken into account:

- the choice of the distribution for the response variable, and for this it is necessary to carefully examine the data, where a preliminary analysis of them is fundamental to make an adequate choice of the family of distributions to be considered;
- the choice of covariates and appropriate formulation of the specification matrix, taking into account the specific problem under study and appropriate coding of the variables;
- the choice of a linking function resulting from prior considerations of the problem at hand, intensive study of the data, ease of interpretation of the model, etc.

The adjustment phase of the model or models goes through the estimation of the parameters, that is, the estimation of the coefficients *β's* associated with the covariates, and of the dispersion parameter if it is present. At this point, it is important to estimate parameters that represent measures of adequacy of the estimated values, obtain confidence intervals and perform adjustment tests.

Finally, the phase of selection and validation of models that aims to find submodels with a moderate number of parameters that is still appropriate to the data, detect relevant differences between the data and the predicted values, ascertain the existence of outliers, etc. In any case, the balance between suitability, parsimony and interpretation is important in selecting the best model (Pestana e Gageiro, 2014; Negas.2021).

Logistic Regression Model

Any regression is based on the calculation of the expected value of the response variable conditioned to the values of the explanatory variables, usually represented by *E(Y|x)*, where x represents the vector of the p covariates (Afonso & Nunes. 2019). When looking for a linear regression model it is because the data obey a relationship of this type, that is, linear, and it can be written that:

$$E(Y|\mathbf{x}) = \beta 1 + \beta 2xi2 + \cdots + \beta pxip$$

This quantity can take on any value in the set of real numbers. However, in the case under study, the response variable assumes only two values, not being a continuous variable so that this expected value can only vary between 0 and 1. To solve this problem, logistic regression rewrites the linear model so that the value of the response variable varies between 0 and 1, through the following equation

$$P(Y_i = 1|\mathbf{x}_i) = \frac{e^{\beta 1+\beta 2xi2+\cdots+\beta pxip}}{1+e^{\beta 1+\beta 2xi2+\cdots+\beta pxip}} \qquad (3.1)$$

It is usual to represent this quantity by *π(xi)*, where the parameters designated by *βj, j = 1,2, ... ,* p represent the effect of the explanatory variables on the response variable Y.In this case, the link link of the MLG (link function) that allows to restrict the values that vary between−∞ e +∞ at half-time [0,1] will be the transformation ***logit***, which is the logarithm of the ratio between the probability of success and the probability of failure, given by

$$logit\ [P(Y = 1|\mathbf{x}\ )] = \ln \left[\frac{P(Yi = 1|\mathbf{x}i)]. }{\text{1-P(Yi=1|xi)}} \right. \qquad (3.2)$$

Its goal is to linearize the model by applying the logarithm. Note that by substituting in the *logit* expression 3.1 is obtained:

$$logit\ [P(Yi\ = 1|\mathbf{x}i)] = \beta 1 + \beta 2xi2 + \cdots + \beta pxip.$$

Parameter estimation β

In order to be able to apply the methodology of generalized linear models to a set of data, it is necessary, after the formulation of the model that is considered appropriate, to estimate the parameters involved in it and to make inferences about this model.

Being $(yi, \boldsymbol{xi})$, $i = 1,2, . . , n$ the i-th observation where *yi* represents the value of the response and xi the vector of covariates. It is known that the response variable can only take two values, 0 and 1. At this stage, the objective is to determine the estimators of the unknown parameters *β = (β0, ... , βp).* It should be noted that, unlike a simple linear regression model, the logistic regression model does not allow us to obtain directly, through the method of least squares, estimates for these parameters. Thus the *maximum likelihood method* (Paula, 2013; Pestana e Gageiro, 2014; Negas, 2021) is usually used.

Quality of adjustment

After obtaining the estimates of the regression coefficients, it is necessary to evaluate the quality of the adjusted model. The first step of this evaluation is to verify whether the estimated coefficients are significant, that is, whether there is a statistically significant association between the explanatory variables and the response variable. For this, the Wald test and the likelihood ratio test are used.

_Wald *Test*

The Wald test is used to test the null hypothesis that the parameter βj, $j = 1, \dots , p$ estimated is equal to zero. The test statistic and its distribution, under the validity of $H0$ are:

$$W_{j} = \frac{\beta j}{\text{Se}(\beta)} \quad \text{N}(0.1)$$

Likelihood Ratio Test

The likelihood ratio test is used to compare the quality of fit of two nested models, that is, models in which one has the subset of variables of the other model. It can also be said that this test evaluates the significance of the coefficients estimated simultaneously, that is, it verifies whether the estimated model is globally significant (Pestana e Gageiro, 2014; Negas, 2021).

Given two nested models, *Mp* and *Mq*, with a number of variables *p* and *q* respectively, such that p < q, to compare the quality of fit of two models can be applied the likelihood ratio test, under the hypothesis that the $q - p$ variables in the model do not present a significant increase in the quality of the model.

Where ln $(L_{Mp}\ (\boldsymbol{\beta}))$ e ln $(L_{Mq}\ (\boldsymbol{\beta}))$ são respetivamente a função verosimilhança do modelo *Mp* e do modelo *Mq*.

However, there are still other tests that prove to be very useful. This is the case of the Hosmer and Lemeshow test to evaluate in a general way the quality of the fit of a model, that is, it tests whether the model fits well to the data (Paula, 2013).

Predictive capability of the mode

When the adjustment objective of the Logistic regression model is prediction, it is necessary that the model has great power of discrimination, because the misclassification error has its consequences. The analysis of the power of discrimination is done through some performance measures such as sensitivity, specificity and the total percentage of correct answers. McCullagh, and Nelder (1989) suggest two methods: ROC Curve and Contingency Tables.

Roc Curves

Be $\hat{Y} = 1$ if an individual selected in the study population is classified as an event of interest, and $\hat{Y} = 0$, otherwise. For this classification, it is necessary to establish a cutoff point that determines the probability of a given individual being classified in one of these classes. The most commonly used cutoff point is 0.5, which means that for an estimated value greater than or equal to 0.5 The individual will be classified in class 1, otherwise they will be classified in class 0. Through an ROC curve it is possible to choose a cut-off point that simultaneously maximizes sensitivity and specificity. This is represented by means of a graph allowing to study the variation of sensitivity and specificity for all possible cutoff points between 0 and 1. Generally, the best cutoff point is based on a *combination of sensitivity* and 1− *specificity that most closely matches* the upper-left corner of the graph.

Contingency Tables

The contingency table is, in this case, a 2 x 2 table for the chosen cut-off point, i.e.

Table 1: Contingency table

Classification	Observed Values	Total

	Class (0)	Class (1)	
Class (0)	n11	n12	n1
Class (1)	n21	n22	n2
Total	n1	n2	

Own Source

Where the performance measures of the model are given by:

- sensitivity: $P(\hat{Y} = 1|Y = 1) = \frac{n22}{n.2}$, which represents the probability of correct classification of the event of interest;
- specificity: $P(\hat{Y} = 0|Y = 0) = \frac{n11}{n.1}$, which represents the probability of correct classification of the event of interest does not occur;
- total percentage of correct answers: $\frac{n11+n22}{n} \times 100$.

The Roc curve plot not only provides the best cutoff point, but the area under the curve ranging from 0 to 1 is a measure of the model's ability to discriminate the values of the response variable, $Y = 1$, from the values of $Y = 0$.

In Hosmer and Lemeshow (2013). is considered a general rule for evaluating the result of the area under the ROC curve:

- If the área is equal to 0.5 there is no discrimination;
- if this area is between 0.7 e 0.8 Discrimination is acceptable;
- if the area is between 0.8 e 0.9 discrimination is good;
- higher than 0.9 is good.

Model Selection and Validation

Faced with several MLGs candidates for a data set, it becomes necessary to determine the most appropriate model. The determination of the model is based on the selection and validation of models and comprises two important questions: "the suitability of the model?" and whether "among the adequate, which is better?"

Model Selection

The selection or comparison of models is the statistical procedure that determines which one should be chosen (Afonso & Nunes. 2019). This template should incorporate all the essential information, excluding the less relevant features, so that the important aspects are highlighted. That is, in selecting the appropriate model the balance between fit (the model should describe the data set as best as possible) and parsimony (the model should allow good predictions without containing unnecessary parameters) is very important.

At this stage, the validation of the model must be carried out, that is, it must be ascertained whether the selected model is suitable.

In practice there are usually a high number of variables that may be potentially important to explain the variability of the response variable. This implies the existence of several models with different combinations of the explanatory variables to explain the phenomenon in question, which makes the selection process more difficult and more time-consuming. To facilitate the selection process, the selection method is widely used stepwise.

The *stepwise model* is an automatic procedure of selecting the variables in backward, forward or both direction. The forward direction starts from a null model and adds the variables, one at a time, that can be significant to explain the variability of the response variable. The null model is a simple model, with no covariates, with only one parameter representing the same mean value for all observations *yi*.

The *backward* steering case, unlike forward steering, starts from a complete model and checks at each step whether or not a variable can be eliminated from the model. The complete or saturated model is the largest model we have the possibility to consider. Given a sample with n observations, the maximum number of parameters for this model is equal to n, that is, one parameter for each observation.

The method *both stepwise* is a combination of two methods (*forward* e *backward*).

The phase of including or excluding the variable from the model is the phase of assessing the significance of the variables or comparing the models. For this, appropriate statistical measures are used for its evaluation.

Model Validation

The deviance is a statistical measure that evaluates the significance of the estimated coefficients and is based on the likelihood ratio test (Pestana e Gageiro, 2014; Afonso & Nunes. 2019).

Considering two models, the first with the variable present and the second without this variable, the likelihood ratio test, already described, allows us to affirm that, under the hypothesis of the model with the variable present being the true model, the deviance is given by

$$D = -2ln\left[\frac{L(modelo\ com\ uma\ variável)}{modelo\ saturado}\quad 2\right] \sim \chi n^{-q}$$

Da mesma forma, se o modelo sem essa variável for o modelo verdadeiro, a *deviance* é dada por:

$$D = -2ln\left[\frac{L(modelo\ com\ uma\ variável)}{modelo\ saturado}\quad 2\right] \sim \chi n^{-q}$$

Thus, the *D* value represents the deviation of the adjusted model from the saturated model. The closer the fitted model, $\hat{\mu}$, is to the observed data, y, the lower the value of *D*.

To assess the significance of an explanatory variable in the model, the difference between the deviance value of the model without the variable and the deviance value of the model with the variable is calculated. The value of this difference coincides with the likelihood ratio statistic, and this value is compared with the quantile of the Chi-Square distribution with $q - p$ degrees of freedom. For a given level of significance, the hypothesis that $q - p$ explanatory variables included in the model are not significant is rejected if the value of the likelihood ratio test statistic is greater than the probability quantile $(1 - \alpha)$ of the Chi-Square distribution (Afonso & Nunes. 2019).

Another measure used to evaluate the model is the Akaike Information Criterion, developed by Hirotugu Akaike and proposed in 1974 (Hosmer and Lemeshow. 2013). This measure is not a hypothesis test, it is a statistic that is based on the logarithm of likelihood and penalizes the model with many variables. The AIC measure is given by

$$AIC = -2[\log(L) - k]$$

where *k* is the number of parameters of the model, and *L* is the likelihood value for the estimated model.

AIC is a relative measure of the information lost by fitting a given model. Unlike the deviance measure, which only compares nested models, it allows you to compare nested or non-nested models. The lower this value, the lower the information lost and, therefore, the better the adjustment of the model (Hosmer and Lemeshow. 2013).

Residue Analysis is useful for evaluating the fit quality of a model with respect to choice of distribution, binding function, and linear predictor terms, as well as identifying observations that are poorly adjusted by the model

The techniques used for residue analysis in generalized linear models are similar to those of the classical regression model. Thus, for the $i - th$ observation, the residue is defined as the difference between the observed value yi and the value estimated by the model.

One can calculate the *Pearson Residue*, the *Deviance Residue*. It should also be noted that for a proper analysis it is necessary to standardize them by their standard deviation. For more information see (Hosmer and Lemeshow. 2013; Pestana e Gageiro, 2014; Afonso & Nunes. 2019).

Evaluating the existence of *influential Observations* consists of verifying the dependence of the statistical model on the various observations that have been collected and adjusted. The Outlier is an observation far removed from the others in terms of the explanatory variables, and may or may not be influential. An influential observation is one whose elimination from the dataset results in substantial changes in certain aspects of the model (Paula. 2013; Pestana e Gageiro, 2014; Sharpe et al. 2018).

Interpretation of regression coefficients

Assuming the assumption that the model fits the data well and that the estimated coefficients are significant, it is necessary to interpret the values associated with the model coefficients. The interpretation of the coefficients of the regression model depends on the nature of the explanatory variables that can be categorical or continuous. In the case of the categorical explanatory variable it is necessary to create auxiliary variables, as already mentioned in the previous section, the so-called variables *dummy*.

Dichotomous independent variable

An explanatory variable is categorical dichotomous if it can assume two possible values (Sharpe et al. 2018). Considering that it takes the values 0 and 1, we can construct a contingency table, that is,

Table 2: Contigency Table

	X=1	x= 0
Y=1	*p1*	*p0*
Y=0	1- p1	1-p0

Own source

with the probabilities that are intended to be estimated, namely the probability in which the response variable can assume the value 1 taking into account the two values that the covariate can assume, i.e. $p_1 = P(Y = 1|x = 1)$ e $p_0 = P(Y = 0|x = 0)$. The expression of the calculation of these probabilities may, taking into account the expression 3.1, be given by

$$p1 = \frac{e\beta 1+\beta 2}{1 + e^{\beta}1+\beta 2} \quad \text{e} \quad p0 = \frac{e\beta 1}{1 + e^{\beta}1}$$

the possibility, or *odds*, may be defined as follows,

$$e\beta 1+\beta 2 = \frac{p1}{1-p1} \quad \text{e} \quad e^{\beta}1 = \frac{p0}{1-p0}$$

Whe

- The reason $\frac{p1}{1-p1}$ represents the possibility of the response variable taking on the value 1 in
- relation to the value 0 when the explanatory variable is equal to 1;
- The reazon $\frac{p0}{1-p0}$ means the possibility of the response variable assuming value 1 in relation to to the value 0 when the explanatory variable is equal to 0;

Applying the logit function, comes

$$logit\ [P(Y = 1|\mathbf{x} = 1)\ ] = \ln\left(\frac{p1}{1-p1}\right) = \beta 1+ \beta 2,$$

In the same way,

$$logit\ [P(Y = 0|\mathbf{x} = 0)\ ] = \ln\left(\frac{p0}{1-p0}\right) = \beta 1,$$

Thus the ratio between the possibilities is called the odds ratio (OR) and its expression is given by

Which represents the risk of the response variable taking value 1 when the explanatory variable is also 1, in relation to the value 0.

Polychotomous independent variable

When the explanatory variable is categorical with more than 2 categories, it is called k categories, then it is necessary to create k − 1 dummy variables. These new variables can take only the values 0 or 1. For convenience the k categories are numbered from 0 to k − 1, with category 0 being the reference class. The possible values 0 or 1 of the variables dummy means that if the characteristic of an object or individual belongs to class i (with *i = 1, ... , k − 1),* then to all dummy variables will correspond the value 0, except for the i- th class which will take the value 1. In the case of the reference class, if the characteristic of an object belongs to this class, then all k − 1 *dummy* variables will correspond to the value *0.*

Thus, for each category of the explanatory variable, the probability of the response variable assuming the value 1 in relation to the value 0 can be estimated.

For example, in the data that will be used there is a variable called *embarked*, which concerns the Boarding Gate, this can be transformed into a *dummy* variable, as follows:

Table 3: Dummy variable

Embarked	Embarked Q	Embarked S
C = Cherbourg	0	0
Q = Queenstown	0	1
S = Southampton	1	0

Own source

The calculation and interpretation of the Odds Ratio value is similar to the case of the dichotomous variable.

Variável independente contínua

When a model contains a continuous independent variable, the interpretation of the corresponding estimated coefficient will be made based on the assumption of linearity between the response variable and the independent variable (Sharpe et al. 2018). It has already been mentioned that to establish this linear relationship the *logit* 3.2 function is used, in which case,

Thus, the interpretation of the estimated coefficient is similar to that of the classical regression model. The coefficient *β2* represents the change in the logarithm of the possibility by a unit of change in the value of the independent variable, x.

By increasing one unit in the value of the variable, x, there will be a difference β2 in the logarithm of the possibility and if we increase k units, there will be a difference of kβ2 units. One can estimate the value of the odds ratio through the exponential of *β2* or *kβ2.*

An interval with (1 − α) 100% confidence for the eβ2 estimate is given by:

Practical application

As already mentioned will be used the set train.cvs available in the link " *https://www. Kaggle. com/c/titanic/data*". This contains data on 891 passengers and among other variables, contains one called "Survived" that takes the value 1 if the passenger survived or 0 otherwise.

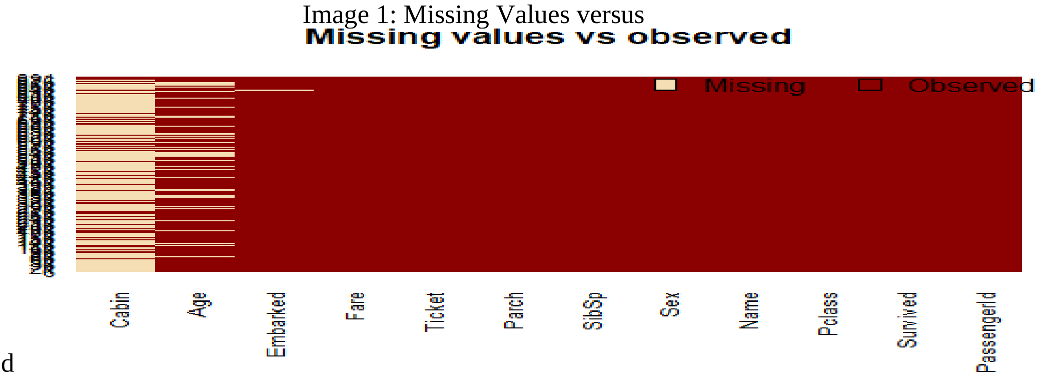
Using the R package, the glm function will be applied. There are several packages in R, with the "glm" function. For this application, the glm2 package was used.

Data Preparation

Before proceeding with the model adjustment, it is very important to prepare the data set for analysis. This step proves to be in many situations and, especially in real data, crucial to fit a good model and with good predictability. Thus, analyzing the available variables, the variable "*PassengerId*" for being only an index and the variable "*Ticket*" will not be considered.

On the other hand, it is necessary to check for missing data. Through the function sapply, it is verified which variables have *missing values*. The Amelia package has a function that, through a graph, highlights the "missing values" called *missmap*. It was found that the variable "*Cabin*" has many missing values so it will not be considered either. Thus, another set of predictor variables was found to be considered: Survived, PcClasse, Sex, Age, SibSp, Parch, Fare, Embarked.

Image 1: Missing Values versus



Observed

Own source

In this new group of data there were still missing values in the variables "*Age*" and "*Embarked*". In the first we chose to replace these missing values with the mean of the ages and in the second, as there were only 2 unknown elements, the corresponding lines were removed. Thus, from 891 observations, 889 were considered. It remains to be taken some precaution with the categorical variables.

So we have a new set of date. Given the number of records available, it became possible to divide them into two groups, the training group, consisting of 800 observations, and the test group with the others. Note that the test group has about 10% of the available data. The training group will be used to adjust the model that will be tested with the test group.

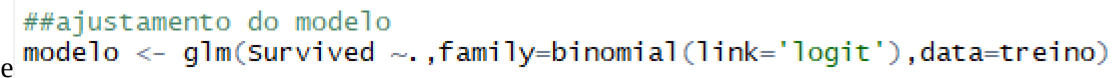
Model Estimation, Selection, and Validation

To estimate the logistic regression model, R was used and the glm function was used. Note that the type of response variable, the binding function (logit) and the data used (in this case, the new data set called training) were indicated(Pestana e Gageiro, 2014; Sharpe et al. 2018). Like this:

Image 2: Model adjustment

Own

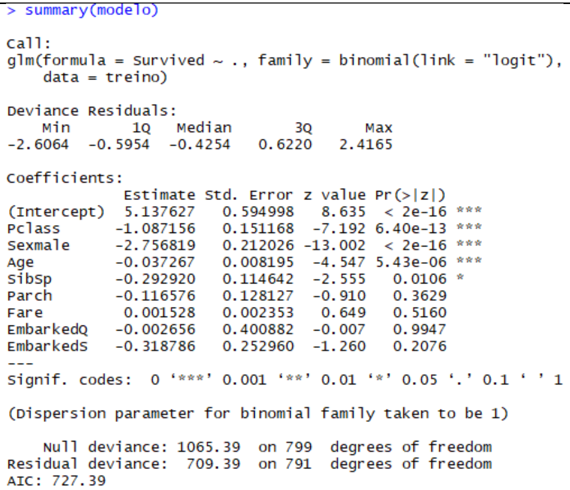
source



```
##ajustamento do modelo
modelo <- glm(Survived ~.,family=binomial(link='logit'),data=treino)
```

To view the result of this model, just use the function *summary*

Image 3: Summary



```
> summary(modelo)

Call:
glm(formula = Survived ~ ., family = binomial(link = "logit"),
    data = treino)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6064  -0.5954  -0.4254   0.6220   2.4165

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.137627   0.594998   8.635  < 2e-16 ***
Pclass      -1.087156   0.151168  -7.192 6.40e-13 ***
Sexmale     -2.756819   0.212026 -13.002  < 2e-16 ***
Age         -0.037267   0.008195  -4.547 5.43e-06 ***
SibSp       -0.292920   0.114642  -2.555   0.0106 *
Parch       -0.116576   0.128127  -0.910   0.3629
Fare         0.001528   0.002353   0.649   0.5160
EmbarkedQ   -0.002656   0.400882  -0.007   0.9947
EmbarkedS   -0.318786   0.252960  -1.260   0.2076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1065.39  on 799  degrees of freedom
Residual deviance:  709.39  on 791  degrees of freedom
AIC: 727.39
```

Own Source

In this, first of all, it can be seen that the variables $Parch$, $Fare$ e $EmbarkedQ$ e $EmbarkedS$ are not considered statistically significant. Of the statistically significant variables, *Sexmale* has the lowest *p*-value, which suggests a strong association of the passenger's sex with the likelihood of having survived. The fact that the coefficient associated with this variable is negative suggests that if all other variables have the same value, the male passenger (*Sexmale*) is less likely to have survived. Not forgetting that in the *logit* model the response variable is given by the ln(odds), comes that the saturated model would be given by:

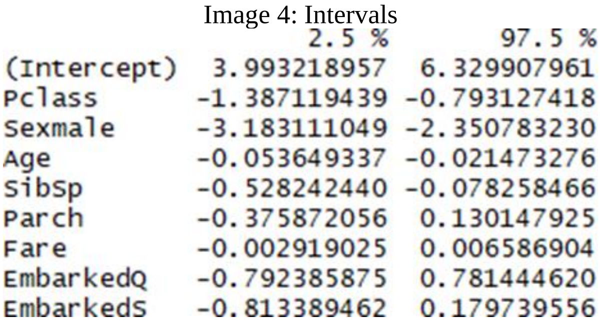
$$ln(odds) = 5.137 - 1.087Pclass - 2.757Sexmale - 0.037Age - 0.293SibsSp - 0.117Parch \\ + 0.02Fare - 0.003EmbarkedQ - 0.319EmbarkedS$$

The analyses of the parameters would be done as described in section 3, for example, as the variable *Sexmale* is a *dummy* variable, the fact that the passenger is male reduces the response variable (not forgetting that as it is the logistic regression reduces the logarithm of the response variable) by *2.75*, while a unit incremented in age reduces this value by *0.037.*

The *confint* function allows you to calculate confidence intervals for the estimates of the coefficients. These are calculated at 95%.

Getting:

Image 4: Intervals



```
                  2.5 %       97.5 %
(Intercept)  3.993218957  6.329907961
Pclass      -1.387119439 -0.793127418
Sexmale     -3.183111049 -2.350783230
Age         -0.053649337 -0.021473276
SibSp       -0.528242440 -0.078258466
Parch       -0.375872056  0.130147925
Fare        -0.002919025  0.006586904
EmbarkedQ   -0.792385875  0.781444620
EmbarkedS   -0.813389462  0.179739556
```

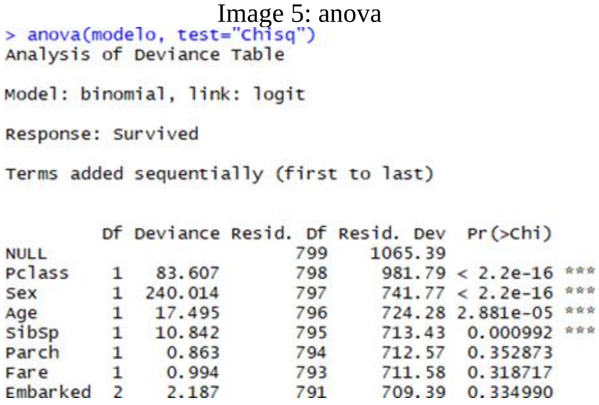
Own Source

However, these estimates are not easy to interpret. A simpler way is to convert the estimates through the ratio of possibility or OR, i.e., by doing,

One can now interpret the results more directly: for example, for every woman who was saved on the Titanic, 0.06 men were saved (or, more intuitively, for every 100 women who survived 6 men were saved), if all other variables have the same value.

Doing: **>anova(modelo, test=”chisq”)** one can see, through the difference between the null deviation and the residual deviation, how the adjusted model improves the null model (a model without covariates). The bigger this difference the better (Pestana e Gageiro, 2014; Sharpe et al. 2018; Afonso & Nunes. 2019). Analyzing the output,

Image 5: anova



```
> anova(modelo, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Survived

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                       799    1065.39              
Pclass    1   83.607       798     981.79 < 2.2e-16 ***
Sex       1  240.014       797     741.77 < 2.2e-16 ***
Age       1   17.495       796     724.28 2.881e-05 ***
SibSp     1   10.842       795     713.43  0.000992 ***
Parch     1    0.863       794     712.57  0.352873    
Fare      1    0.994       793     711.58  0.318717    
Embarked  2    2.187       791     709.39  0.334990    
```

Own Source

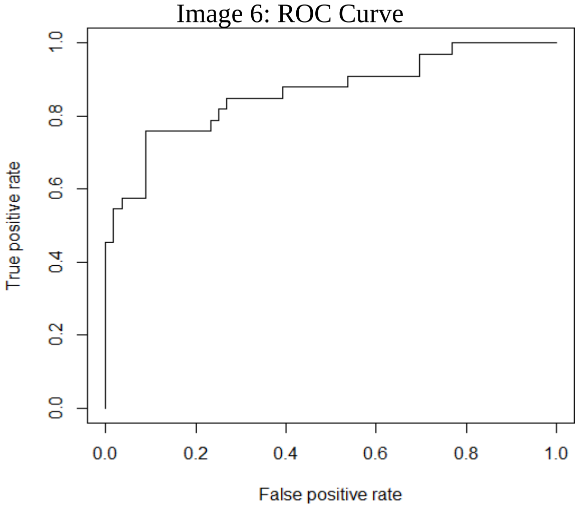
it is also verified that the deviation decreases as the variables are added, one by one, to the model. It is also verified that adding the variables *Pclass*, *Sex* and *Age* significantly reduces the residual deviation. The other variables cause a smaller effect in the reduction of this deviation, even the variable *SibSp*, which although has a small p-value and can be considered significant. In this situation, a high p-value indicates that the inclusion of this variable in the model does not significantly increase the explained value of the response variable. On the other hand, it is intended that the inclusion of a variable causes a significant drop in the value of the deviation and the AIC. Thus, the model with only the variables *Pclass, Sex, Age* and *SibSp* seems to show good results and with fewer variables.

McFadden's R2 value allows quantifying the fit of the model. Using the *pscl* package one can calculate this value, being in this case *0.334*.

It is now necessary to test the predictive capabilities of the response variable of the model found in a new data set, the test set. In *R*, defining the parameter type as "response" with probability of y being 1 given the covariate vector *X*, i.e. $P\ (y\ =\ 1|X)$. *If* $P\ (y\ =\ 1|X)\ >\ 0.5$, then the answer is y = 1, otherwise y = 0. In other different situations, different decision thresholds may prove to be better options. An "Accuracy" value of approximately *0.84* was obtained for the test data for the variables, which is a good result.

Finally, the ROC curve was plotted and the AUC (area under the curve) was calculated, which are the typical performance measurements when the response variable is binary. This curve is obtained by tracing the sensitivity and specificity at each cutoff point in pairs. This shows the relationship between the sensitivity and specificity of a test and can be used in deciding the best cutoff point. As a general rule, a model with good ability to discriminate values according to classes, should have an AUC closer to 1 (1 is ideal) than for *0.5* and a curve approaching the upper left corner of the chart. In this case, we obtained an $AUC = 0.86$ and an ROC curve near the upper left corner, that is, we can affirm that the model has a good discrimination capacity. In the following figure, the ROC curve for this data set is represented.

Image 6: ROC Curve



Own source

It would be a case to say to anyone who was about to travel on the Titanic: Tell me your age and your gender, who you travel with and what class you are going to, and I will tell you your destination!

Conclusion

Logistic regression is the most widely used method to model the binary response of data. Modeling a binary response variable using normal linear regression introduces bias in the estimation of parameters, and does not fulfill, for example, the assumption that a Standard linear model contains the response variable with normal distribution. This is because the binary response model is derived from the Bernoulli distribution. We have seen that the probability function of a Bernoulli is part of the exponential family, this family of distributions, which allows for easier estimates to determine. We have also seen that when it comes to estimating models based on the exponential family, the Generalized Linear Models algorithm is the best option.

Thus, using this algorithm, the Logistic Regression model was adjusted to a group of data known to everyone (the data on the Titanic tragedy), and tried to make predictions to determine whether a given passenger would survive or not, depending on other variables, such as gender, the class where he traveled, etc. It was concluded that there was a model, which in relation to the saturated model (where all variables were included), one could consider only the variables classes where he traveled, sex, age and number of siblings/spouses who accompanied this individual, without the variability of the response variable, suffering major changes.

We also tested a group of data not used in the estimation, which demonstrated good predictions and with the ability to discriminate in the intended response.

It should also be noted that the characteristics of the data set used would easily be found in several areas of study, so the methodology used is generalized to many other situations in which the response variable is binary.

References

[1]. Afonso, A. & Nunes, C. (2019). Probabilidades e estatística - Versão revista e aumentada Editora: Universidade de Évora.

[2]. Agresti, Alan (2002). *Categorical Data Analysis*. Wiley.

[3]. Alvarenga, A.M.T (2015).*Modelos lineares generalizados: aplicação a dados de acidentes rodoviários*, Tese de Mestrado, Dept. Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa.

[4]. Faraway, J.(2006), *Extending the Linear Model with R*, Chapman and Hall Ltd, London

[5]. Hosmer, D. W., and Lemeshow, S. (2013). Applied Logistic Regression. Wiley.

[6]. McCullagh, P.; Nelder, J. A. (1989). *Generalized Linear models - second edition*. Chapman and Hall Ltd, London

[7]. Murteira, B. J. F. (1988). *Estatística: Inferência e decisão*. Imprensa Nacional - Casa da Moeda.

[8]. Negas, Elsa. (2021). Estatística Descritiva: Explicação teórica, casos de aplicações e exercícios resolvidos (2ª Edição). Edições Sílabo

[9]. Nelder, J.A. and Wedderburn, R.W.M. (1972). *Generalized linear models*. Journal of the Royal Statistical Society, A 135, 370-384.

[10]. Paula, G.A. (2013). *Modelos de Regressão com Apoio Computacional*, São Paulo; IME - Universidade de São Paulo.

[11]. Pestana, M. H. e Gageiro, J. N. (2014). Análise de dados para ciências sociais. A complementaridade do SPSS.

[12]. 6ª Edição. Edições Sílabo

[13]. Sharpe, Norean, Velleman, Paul F. e Veaux, Richard D. De, (2018). Estatística Aplicada Administração, Economia e Negócios. Bookman

[14]. Sienna, Kleber. (2019) Titanic, o legado do grande navio. Permanências e alterações no imaginário das catástrofes. Dissertação de Mestrado apresentada à Universidade Federal da Uberlândia- Brasil

[15]. Turkman, M. A., & Silva, G. L. (2000). Modelos Lineares Generalizados - da teoria prática. Lisboa: Edições SPE.