UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Comparative analysis of non-linear models with parametric link function based on the family of Czado functions and asymmetric functions of Aranda Ordaz

Maria Margarida Marques Campos de Azevedo

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Professora Doutora Lisete Sousa
Professor Doutor Carlos Geraldes

2021

Agradecimentos

Em primeiro lugar gostaria de agradecer aos meus orientadores, a Professora Lisete Sousa e o Professor Carlos Geraldes, por toda a disponibilidade e apoio ao longo de todo o processo de elaboração deste projecto. À minha família, mãe, pai, irmão e Selma o meu mais sincero obrigada pelo amor e apoio incondicional que sempre me deram. À Teresa, porque não podia pedir melhor companheira, amiga e apoio ao longo de todo o mestrado e durante todo o processo de elaboração deste projecto. À Cris e às Bisnasgas, porque não podia pedir amigas mais pacientes e que me transmitissem mais força. E por fim ao João, porque não existem palavras para descrever toda a paciência, companheirismo e apoio que demonstrou durante todo este percurso.

Resumo

Para a boa prática e gestão clínica é fundamental o desenvolvimento de modelos para a tomada de decisão médica. A admissão de doentes críticos nas Unidades de Cuidados Intensivos (UCIs) constitui um bom exemplo. Estes serviços têm como missão a prestação de cuidados de saúde a pacientes em situação crítica, o que constitui um desafio à gestão hospitalar tendo em conta os pesados orçamentos que são necessários para a manutenção da qualidade de resposta. Por isso, no dia a dia das UCIs, terão que ser tomadas decisões com base na eficácia do tratamento *versus* o seu custo. Para auxiliar essas decisões, utilizam-se métricas, normalmente obtidas a partir de modelos de regressão, sendo os mais utilizados o modelo linear generalizado (GLM), e o modelo aditivo generalizado (GAM). Estes são geralmente orientados para a quantificação do risco de mortalidade e caracterizam-se por um número reduzido de variáveis, a partir das quais se extrai uma pontuação que reflete o estado da gravidade do doente além de uma estimativa de mortalidade intra-hospitalar. De entre as componentes que se podem trabalhar no sentido de melhorar a qualidade dos modelos destacamos a função de ligação. Trabalhos recentes usando modelos com funções de ligação paramétricas flexíveis, nomeadamente com funções de ligação pertencentes à família de funções assimétricas de Aranda-Ordaz, revelaram uma melhoria no seu desempenho. Por outro lado, estudos que envolvam funções de ligação pertencentes à família de funções Czado são escassos. Neste último caso, a função depende de três parâmetros proporcionando maior flexibilidade do que a função de ligação Aranda. Assim, neste estudo pretende-se efetuar uma análise comparativa do desempenho dos modelos acima referidos (GLM, GAM), utilizando as funções de ligação Aranda e Czado, tendo como *baseline* a função de ligação Logística.

O tratamento estatístico de dados clínicos apresenta uma grande importância uma vez que, ao retirar conclusões da análise estatística, estas irão ser aplicadas em situações reais do quotidiano e influenciar diretamente o tratamento de doentes admitidos no hospital. Esta análise apresenta, assim, uma influência muito direta e determinante na tomada de decisões e deverá ser encarada com grande seriedade. A manutenção adequeada de tratamentos e respostas adequadas para doenças depende, em muitos casos, de um tratamento estatístico coerente e com resultados de fácil interpretabilidade que possam facilmente ser passados à comunidade científica. Os resultados estatísticos obtidos, influenciam assim, não só o tratamento direto de pacientes como, a eficiente gestão hospitalar e de verbas.

Os modelos lineares generalizados, são já bastantes comuns na análise de dados clínicos, e apresentam uma grande vantagem pela sua simplicidade de utilização e fácil interpretabilidade. No entanto, o facto de assumirem uma relação entre a combinação linear das variáveis explicativas e a variável de resposta pode apresentar um problema consoante os dados a serem

analisados. Caso a suposição acima referida esteja correta, a sua utilização não apresenta um problema, caso contrário, existe a necessidade de aplicar outro tipo de modelos para contornar esta suposição.

Os modelos aditivos generalizados encontram-se como uma boa alternativa para a limitação apresentada pelos modelos lineares generalizados, uma vez que não assumem a relação entre a combinação linear das variáveis independentes e a variável resposta. Para o efeito, esta classe de modelos utiliza funções suavizadores que permitem uma maior flexibilidade na relação entre variáveis. Apresentam ainda a vantagem de muita da interpretabilidade dos modelos lineares generalizados ser aplicável a estes modelos, facilitando assim a sua utilização num contexto real médico com consequências diretas na vida dos índividuos.

As funções de ligação utilizadas no contexto deste projecto, permitem uma maior flexibilidade relativamente às funções de ligação utilizadas comumente. O principal objetivo é tentar melhorar a adequabilidade da função de ligação aos dados utilizados, esperando assim obter um melhor resultado. Para isso duas famílias de funções foram utilizadas, Aranda-Ordaz e Czado.

A família de funções assimétricas de Aranda-Ordaz engloba as funções logística e log log, como casos especiais, e podem variar através de um único parâmetro, permitindo assim ter uma maior flexibilidade na função de ligação, apresentado já uma melhoria relativamente a utilizar somente, como é realizado frequentemente, a função de ligação logística.

A família de funções Czado, engloba igualmente a função logística, podendo, no entanto, ser adaptável através de três parâmetros independentes, permitindo uma grande flexibilidade e representado uma inovação em termos de funções de ligação, especialmente para o ramo médico.

Foram elaborados modelos diferentes utilizando a combinação de modelos lineares generalizados e modelos aditivos generalizados com as duas funções de ligação, Aranda-Ordaz e Czado. Os parâmetros de cada combinação de modelos foram variados de forma a obter os valores para cada parâmetro ideais dentro do problema apresentado, tendo-se, seguidamente, comparado os diferentes modelos obtidos de forma a poder selecionar qual o mais adequado aos dados. Para a comparação dos modelos foram utilizadas as medidas de qualidade AUC e o Brier *score*. Ambas as medidadas de qualidade foram utilizadas para comparar todos os modelos. Para garantir que a diferença ocorrida entre as AUCs de modelo para modelo era estatisticamente significativa foi utilizado o teste de DeLong para comparação de AUCs. Este teste permitiu perceber se o aumento do valor das AUCs de modelo para modelo, significava uma melhoria efetiva nos modelos, ou se a diferença nos valores não correspondia a uma melhoria estatisticamente significativa. Foram elaborados gráficos a representar a função de ligação de base, a logística, e as funções de ligação ótimas para cada um dos modelos, de forma a poder comparar visualmente as diferenças existentes enrte estas.

Nenhuma melhoria foi observada através da utilização dos modelos apresentados para a situação aplicada e para os dados utilizados. No entanto, o estudo utilizando funções de ligação que apresentam maior flexibilidade, e por isso, a obtenção de resultados mais precisos no contexto do problema, é sempre benéfica. O trabalho encontra-se desenvolvido e poderá ser

aplicado futuramente, noutro contexto, podendo, potencialmente, obter melhores resultados. Uma sugestão pricipal será de utilizar estas metodologias no contexto de outros problemas e outros dados na esperança de obtenção de resultados mais significativos e mais pertinentes para a solução do problema.

Este trabalho permitiu verificar que a utilização de funções de ligação mais comuns pode encontrar-se correta e não comprometer por isso os resultados, mas demonstrou igualmente como garantir que a escolha da função de ligação se encontra correta. A escolha de qual a função de ligação a ser utilizada deve passar sempre por um estudo mais aprofundado e baseado em evidência. Este trabalho, vem assim, apresentar também um método mais sistemático, eficiente e rigoroso na escolha da função de ligação adequada. Ao variar entre diferentes distribuições que apresentam um ou três parâmetros, permite de forma eficiente, simples e rápida garantir que a escolha da função de ligação é de facto a mais apropriada.

A escolha entre modelos lineares generalizados e modelos aditivos generalizados permite também a obtenção de resultados mais robustos, uma vez que não existe o pressuposto de linearidade entre a variável resposta e as varáveis independentes. Assim, a escolha do modelo é também feita de forma criteriosa e tendo em conta o problema em questão, de forma a obter os melhores resultados possíveis no contexto do problema. Este trabalho apresenta um avanço na forma como a estatística e a medicina se relacionam, na medida em que pretende fornecer, através de um trabalho estatístico coerente, os melhores resultados possíveis e e de fácil compreensão passíveis de serem aplicados em contexto médico.

Palavras Chave: GLM, GAM, Aranda-Ordaz, Czado.

Abstract

For a good clinical practice and management to be attained, statistical analysis can perform a major role. Statistical models can greatly aid in medical decision making, as for example in the admission of critically ill patients to the Intensive Care Units (ICUs). These services have the mission of providing health care to patients in critical situations, which constitutes a challenge to hospital management, considering the heavy budgets that are necessary to maintain the quality of response. Therefore, decisions are made daily bearing in mind the effectiveness of the treatment *versus* its cost. In order to help the decision-making process metrics can be obtained, usually via Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs). These are generally oriented towards the quantification of the risk of mortality and are characterized by a small number of variables, from which a score is extracted that reflects the patient's state of severity in addition to an in-hospital mortality estimate. Among the components of GLMs and GAMs which can be focused on in order to improve the quality of the models, the link function is highlighted. Recent work using models with flexible parametric link functions, namely with link functions belonging to the family of asymmetric functions of Aranda-Ordaz, showed an improvement in their performance. On the other hand, studies involving link functions belonging to the family of Czado functions are scarce. Using a Czado link function provides a greater flexibility, by it depending on three parameters, rather than the Aranda-Ordaz link function, which merely depends on one. Thus, a comparative analysis of the performance of both models referred (GLM, GAM), using the Aranda-Ordaz and Czado link functions, and considering the Logistic link function as a baseline was the primary line of work. The results presented themselves as inconclusive regarding the greater performance of either of the link functions, which can be related to the data used and not necessarily the actual performance of the models and link functions. Further studies should be carried using different data sets in order to truly access the performance of the models using both Czado and Aranda-Ordaz link functions.

Keywords: GLM, GAM, Aranda-Ordaz, Czado.

Abbreviations and Acronyms

GLM Generalized Linear Model

GAM Generalized Additive Model

ROC Receiver Operating Characteristics

AUC Area under the ROC curve

AIC Akaike Information Criterion

MLE Maximum Likelihood Estimation

RSS Residual Sum of Squares

MSE Mean Squared Error

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The first chapter of this work, aims to make an introduction both to the importance of developing it and the context in which is created. The problem at question is discussed during the chapter, making reference to already implemented solutions as well as presenting the alternative solution here proposed, the core of the project.

1.1 Clinical Data

Clinical data is often collected with the intent of conducting a study to obtain a better knowledge of a certain disease, group of patients or a diagnostic test. As its analysis is a primary step in decision making, a correct statistical analysis is of crucial importance [Geraldes, 2016].

Both planing and delivery of services depend greatly on data from clinical resources. Evidence-based practice, at the efficiency seen nowadays, is only possible through access to extensive research data, collated and presented in such a manner it can be easily understood by a clinician in order to make a diagnosis or in other decision making situations. It is only logical to conclude, the higher the quality of the data collected and the statistical analysis performed, the better will be the patients outcomes. The greater quality in decision-making implies a reduction in uncertainty and leads to more timely and accurate decision outcomes [Kerr et al., 2007].

In order to perform an analysis with useful, palpable results, data is collected based on two types of studies. One of which is called a Prognostic study, being longitudinal due to the continuous observation of a group of patients with the aim to observe a desired outcome. This type of study can be divided in two: a prospective design, where the outcome is awaited for in the future, and a retrospective design, where patients are followed continuously back in time, mainly through hospital records. The other type, the Diagnostic study, is often cross-sectional, and is characterized by having the study group defined by the presence of a symptom or the exposure to a certain factor, without the precise knowledge of the disease presence or not.

1.2 State of the Art

As stressed previously, statistical models take a major role in aiding clinicians making more accurate predictions based on data. A prognostic model can then be defined as a statistical tool that predicts a clinical outcome based on at least two points of patient data, being the patient information more often used rather than information regarding the disease or condition itself. Prognostic models can be divided into two categories: prognostic models at the patient population level, where the objective is to find a trend or discrepancy in groups of patients for a specific criterion, and prognostic models at the individual patient level [Vogenberg, 2009].

In longitudinal studies, referred to in the previous section, such as the one that will be presented in this work, the interest lies in the association between longitudinal response process and a binary outcome, and to predict a binary event in case of an existing association. The aim of modelling the longitudinal and binary data is to provide an estimated probability of the event of interest [Li et al.. 2015].

Both generalized linear models (GLMs) and generalized additive models (GAMs) have been widely used to elaborate prognostics models. GLMs are built on the basis of an existing linear relationship between a link function of the expected response variable and the explanatory variables. GAMs represent a step further, being an extension of GLMs by replacing the explanatory variables with smooth functions, which are often used to deal with nonlinear relationships between the response variable and explanatory variables [Yu et al., 2013]. These models are commonly used to quantify the risk of mortality, being characterized by the usage of a reduced number of variables, where a score is obtained in order to qualify the gravity of each patient. These models also allow for an estimate of mortality within a hospital [Geraldes, 2016].

Generalized linear models present a more common solution due to their transparency in terms of interpretability, which can present an advantage when dealing with clinical data. Models of this type are also able to deal with categorical predictors, common in the medical field, and allow for a clear understanding of how each predictor influences the outcome. These characteristics present as valuable, since they can connect the statistical results to the knowledge already acquired empirically in the field. However, it is necessary to assume the linear relationship between a link function of the expected response variable and the explanatory variables, which can be considered less than optimal, depending on the data itself [Amaral Turkman & Silva, 2000].

Generalized additive models present themselves as a more versatile solution compared to GLMs. It maintains most of the interpretability GLMs possess, adding the advantage of not establishing a linear relationship between a link function of the expected response variable and the explanatory variables. Instead a relation between the response variable and the predictors does not need to be assumed prior to the application of the model, as it is estimated [Geraldes, 2016].

Another aspect of building a prognostic model using both GLMs and GAMs lies in the importance of choosing a correct link function. A misspecification in the link function can carry terrible mistakes to a prognostic model by increasing the MSE of the estimated response probability which reflects in a considerable bias when estimating parameters [Li et al.. 2015].

This implies that the choice of a correct link function may have a definite importance to a correct characterization of a patient state. The most common link function for binary models is the logit link function, where the curve between the probability of an event and covariates is assumed to be symmetric. However, if there is an imbalance between the probability of the rate of a binary response approaching 0 and 1 may occur, the logit function no longer presents itself as satisfactory.

Parametric links are commonly chosen based on the fact that they include the canonical link, their flexibility of different shapes, their mathematical simplicity and their comparison of maximum likelihood fits in data sets. Their usage presents an improvement in terms of fit in maximum likelihood. However the cost in terms of increasing the variances of the estimated regression coefficients and mean response predictions when the link is estimated should also be taken into consideration [Czado, 1992].

It is possible to find already some work develop towards finding more flexible and better fitting link functions for such cases, both using parametric and non-parametric functions.

The work of Li, Xang & Seongho (2015) [Li et al.. 2015] presented two families of flexible link functions used in joint models of longitudinal measurements and a binary outcome. One of the families was the generalized extreme link, which allows for a more flexible skewness controlled by a shape parameter. This is a particularly beneficial model when the binary outcome possesses an imbalance between observed ones (1) and zeros (0). The other was the power link function proposed by Jiang et al. (2013) [Jiang et al., 2013], based on the cumulative distribution function corresponding to a symmetric baseline link function and its mirror reflection. The introduction of a power parameter allows for flexibility in skewness both in positive and negative directions, and allows to maintain the symmetric baseline link as a special case.

Another example of application of less common, yet more flexible link functions, is the work of Geraldes (2016) [Geraldes, 2016], where, for a generalized additive neural network, both a parametric link function and non-parametric were applied. The parametric link function utilized was the Aranda-Ordaz function and for the non-parametric a multi-layer perceptron was used to estimate the link function.

With the problematic of heterogeneous sets of binary data in mind Aranda-Ordaz proposed in 1981 [Aranda-Ordaz, 1981] a family of power transformations for probabilities in order to provide a representation of alternative scales for analysing binary response data. Such family of transformations presents itself as a proper alternative for a more flexible link function already seen in Geraldes (2016).

The Aranda-Ordaz asymmetric transformation is an extended model from the frequently used logistic distribution, which includes the distribution mentioned, as well as others, as special cases. Such is possible by the variation of a single parameter, which in a range from 0 to 1, allows for a more flexible link function [Aranda-Ordaz, 1981]. This variation is the primary factor here studied, as the greater the flexibility of the link function and the model, hopefully the better the results according to the data.

Czado (1992) [Czado, 1992], proposed parametric link families, which, due to parameter orthogonality and standardization, are able to overcome the problem of estimating the link parameter, reducing the variance inflation and thus increasing numerical stability while maintaining the likelihood fit.

As flexibility is so important, the Czado family of transformations takes a step forward. The family, which is a unified method for choosing parametric link functions, can use up to three parameters for doing so [Czado, 1992]. The variation of these three parameters allows for a far better adaptability, and thus for a finer selection and fit to the data.

Both lastly mentioned propositions of families can be considered as potential beneficial link functions, as they try to overcome both the problematic of flexibility and cost implied in the estimation of the link parameter. The combination between the more adaptable link functions and greater versatility of models, by obtaining a final advantageous model, presents as an attempted solution to the optimal prediction of outcomes in the medical field, and thus representing an advance in formulation of a correct prognosis and a better health assistance.

1.3 Objectives

The objective of this report is to compare the performance of different statistical methods in predicting a correct outcome, in a medical context. As mentioned previously, a correct statistical prediction is of crucial importance, taking into consideration that a clinical prognosis can derive from it. The aim is to compare the performance of two link parametric functions, Aranda-Ordaz and Czado, using two different classes of models, GLMs and GAMs, in order to assess which method has the most accurate results, and thus more reliable when applying to a real life situation.

In order to elaborate this work, the following steps were defined:

- Understand the clinical data used, by doing an exploratory analysis of each variable individually and choosing which variables should be incorporated in each model at test.
- Build different models using both GLMs and GAMs, applying both link functions Aranda-ordaz and Czado family of functions.
- Analyse each model obtained through measurements such as AUC and Brier score. Apply DeLong test to validate if differences between AUC values are considered statistically significant.
- Compare the performance for the best models obtained, taking into special consideration the potential difference in performance of models using the Aranda-Ordaz and Czado families as link functions.

The data, which is comprised of the measurements made to several indicators on patients on arrival day at São Jos*é* Hospital, Lisbon, is merely used to fulfill the statistical purpose of the work, as no clinical conclusions are going to derive from its use. Nevertheless, the statistical conclusions itself can, hopefully, be transposed into a real life situation and be applied in order to aim in a medical context. All analysis were performed using software R.

All things considered, the main goal here proposed is to obtain better fitting statistical models by having a greater flexibility to adapt to different types of sets of data, and thus producing more accurate and usable results in a real life context.

1.4 Work Structure

The work here presented is divided into 4 chapters, namely, Introduction, Methods, Results and Conclusions and Discussion.

In the first chapter a brief introduction is made to the overall theme of this report, explaining the potential importance of the work here developed, presenting some previous related works and the objectives to be attained.

Chapter two consists of explaining in somewhat detail the methods used for the development of this report. GLMs, GAMs, the family of functions Aranda-Ordaz and Czado are discussed. Also, evaluation methods used to compare models, variable selection and every necessary methodology from data preparation to the conclusion to which is the best and final model are included.

Chapter three contains the results obtained throughout the elaboration of this work. It is the result of applying the methodology discussed in chapter two, to the data collected. Hence forth, chapter three is the summary of all statistical analysis performed and their results. Exploratory analysis, variable selection, estimation of each model and comparison of model performance are the main sections included.

Lastly, chapter four is a critical discussion of the results obtained accompanied by suggestions of possible future developments in the same line of work.

Chapter 2

Methods

The chapter contains an explanation regarding all methods used to developed this project. The detailed description of each method can be found in each section. Information on both types of models used, generalized linear models (GLMs) and generalized additive models (GAMs), as both link functions, Aranda-Ordaz and Czado, as well as the evaluation methods used, can be found in the next pages of this work.

2.1 Data Collection and Variables

In order to conduct the work here developed, a set of clinical data, previously collected, was used. The data used was obtained by measuring several indicators on the arrival day, on patients admitted to São Josė Hospital, Lisbon, and observing if they were deceased by the third day of hospitalization. The characteristics of the data collection are consistent with the ones of a Prognostic study with a prospective design. A total of eight indicators were collected, accounting for eight variables which can be used for the statistical analysis, plus an outcome variable observed three days later from admission.

The indicators measured on the arrival day, the independent variables $\mathbf{x} = (x_1, x_2, \ldots, x_8)$, consist of:

- Blood pressure of the patient at the admission moment (BPre);
- Serum sodium level of the patient at the admission moment (SSLev);
- Urinary output of the patient at the admission moment (UOut);
- Age of the patient at the admission moment (Age);
- Serum urea Level of the patient at the admission moment (SULev);
- Bilirubin Level of the patient at the admission moment (BLev);
- Serum bicarbonate Level of the patient at the admission moment (SBLev);
- If the patient was ventilated at time of admission in the hospital (Ventilated).

The outcome variable y, consists of:

- State of the patient, deceased or alive, after three days of the admission date (Death).

2.2 Exploratory analysis and Variable Selection

The first step for any statistical analysis is to understand the data being analysed. The exploratory analysis intends to do so. For each variable, basic metrics such as the mean and quantiles, measurements of location, standard deviation, a measurement of variation, should be calculated, as well as a graphic representation of binary variables. It should be taken into consideration the symmetry, or not of the distribution of each variable.

For testing the symmetry of the distribution of each variable, the Cabilio-Masaro test of symmetry about an unknown unique median of a distribution, θ, of a probability distribution with density function f and distribution function F, can be used. The test, considering a random sample $X_1, \ldots, X_n$ identically drawn from a probability distribution, has the hypothesis [Cabilio & Masaro, 1996]:

$$H_0 : f(\theta - x) = f(\theta + x)$$

The test statistic under H_0 is:

$$S_K = \frac{\sqrt{n}(\bar{X} - m)}{S} \sim \mathcal{N}(0, \sigma_0^2(F)), \tag{2.1}$$

where $\bar{X}$ is the sample mean, m is the sample median, S^2 is the sample variance and

$$\sigma_0^2(F) = 1 + \frac{1}{e_{m,\bar{X}}(F)} - \frac{2}{e_{m,\bar{X}}(F)} E\left|\frac{X - \mu}{\sigma}\right|. \tag{2.2}$$

The test hypothesis H_0, for a sample of size n and a significance level of α, is rejected if $|S_K| \geq P_{1-\alpha/2}$, where $P_{1-\alpha/2}$ is the quantile of order $1 - \alpha/2$ of distribution $\mathcal{N}(0, \sigma_0^2(F))$. For values of $P_{1-\alpha/2}$ and further developments on equation (2.2) please refer to Cabilio and Masaro (1996).

The Pearson's correlation between continuous variables should also be included. The correlation, which varies between -1 and 1, allows for understanding the existence of a linear relation between the variables and therefore understand how variables behave together and influence each other in a linear fashion. The correlation equation is as follows [Benesty et al., 2009]:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{2.3}$$

Moreover, boxplots for each continuous variable, represent as well a good visualization technique for understanding, initially, how the remaining variables relate to the outcome variable and consequently if they should be included in the model or not.

In order to obtain an optimal model, a rather important component is the variable selection. The main objective in doing so, is to be able to find the most parsimonious model while still being able to correctly explain the data [Hosmer & Lemeshow, 2000]. Models which contain a lower number of variables tend to be numerically stable, making it easier for their generalization. Moreover, a higher number of variables may mean the model is more dependent on the data. On the other hand, an argument in favor of including a higher number of variables in the model is that in doing so, it allows for a complete control of confounding, given that variables may only display confounding when incorporated together.

For each variable a univariable logistic regression model, a regression model using a logistic link function, discussed ahead, should be fitted. For decision making purposes, the estimated coefficient, the univariable Wald statistic and its correspondent p-value should be obtained. Each variable is incorporated in the multivariable model if its p-value is below 0.25. The p-value is higher than the more conservative, and more commonly used, value of 0.05, in order to identify variables known as important. Nevertheless, a critical look at the variables should always be taken before incorporating them in the model.

Another measurement important for model selection is the AIC (Akaike Information Criterion). The information criterion, which presents itself as a review of the maximum likelihood estimation procedure, can be defined as [Akaike, 1974]:

$$\mathrm{AIC} = -2\log(L) + 2k. \tag{2.4}$$

As understood by the formula, where L is the likelihood function and k is the number of selected parameters for the given model, the lower the value of the AIC, the better the model. The AIC is particularly useful for nested models, being much more precise in this case rather than for non-nested ones.

2.3 Generalized Linear Models

A linear model presents as follows :

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \tag{2.5}$$

where Y is the response variable, $\beta_0, \beta_1, ..., \beta_p$ are the regression coefficients, $x_1, ..., x_p$ are the explanatory variables, and ε accounts for the random component. The model attempts to explain the relationship between one, or more predictor variables and one response variable, where the linearity of the model is assumed. The linearity is only applicable to the regression coefficients and not to the variables x, using the least squares theory as the analytical technique. For the random component, the error, a normal distribution is assumed [Rencher & Schaalje, 2008].

A generalized linear model (GLM) can be obtained from the linear model using techniques for non-normal data. The model uses other functions than the identity function as a linear predictor, extending the linear model to fit data in which the response variable probability function can be different from the normal distribution, though still belonging to the exponential family

[Geraldes, 2016]. The exponential family includes distributions such as normal, Bernoulli, binomial, Poisson, exponential and gamma distributions. Distributions belonging to the exponential family present as follows:

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{2.6}$$

where θ is the canonical form of the location parameter and ϕ is the scale parameter, supposedly known. Functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are known. The following expressions represent the mean and variance of functions which belong to the exponential family [Amaral Turkman & Silva, 2000]:

$$\begin{aligned} E[Y] &= b'(\theta) \\ var(Y) &= a(\phi)b''(\theta) \end{aligned} \tag{2.7}$$

The class of generalized linear models was created to unify the procedure for fitting models regarding the distributions previously mentioned, through the usage of maximum likelihood estimation, by Nelder and Wedderburn (1972) [Nelder & Wedderburn, 1972].

GLMs are then comprised of two components:

- A random component, which, given a vector of covariates $x_i = (x_{i1}, \dots, x_{ip})$, the components Y_i have independent normal distributions with $E(Y_i|x_i) = \mu_i = b'(\theta_i), i = 1, ..., n$;
- A systematic component, a linear predictor defined by

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{2.8}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

is a specification matrix, function of the vectors of covariates $x_i, i = 1, ..., n$ and $\boldsymbol{\beta}^T = (\beta_0, \dots, \beta_p)$ is a vector of parameters of dimension $p+1$.

A relationship between the linear predictor $\eta_i = \mathbf{z}_i^T\boldsymbol{\beta}^*$ and the mean value can be established:

$$\mu_i = h(\eta_i) = h(\mathbf{z}_i^T\boldsymbol{\beta}^*), \quad \eta_i = g(\mu_i), \quad i = 1, \dots, n \tag{2.9}$$

where h is a monotonous, differential function, $g = h^{-1}$ is the link function and, as previously mentioned, can take other forms than the identity, $\boldsymbol{\beta}^*$ is a parameter vector of dimension p and $\mathbf{z}_i$ is a specification vector of dimension p, function of the covariate vector $\mathbf{x}_i$.

Generally, $\mathbf{z}_i = (1, x_{i1}, \dots, x_{ik})^T$ where $k = p - 1$. However, for qualitative variables a codification as to be made recurring to *dummy* variables.

Regarding the link function, its choice should depend on the type of response and the particular study under consideration. For when the linear predictor coincides with the canonical parameter, $\theta_i = \eta_i$, which implies $\theta_i = \mathbf{z}_i^T \boldsymbol{\beta}^*$, the corresponding link function is then called *canonical link function* [Amaral Turkman & Silva, 2000].

Finally, in order to be considered adequate, a GLM should follow the subsequent assumptions [Montgomery et al., 2012]:

- The error term ε has zero mean;
- The error term ε has constant variance σ^2;
- The errors are uncorrelated;
- The errors are normally distributed.

2.4 Generalized Additive Models

Generalized additive models (GAMs) represent a step forward when compared to GLMs, when it comes to the relationship between the independent variables and the dependent variable. As mentioned, GLMs assume a linear relationship between a link function of the expected response variable and the explanatory variables, being GAMs created to overcome this limitation allowing for more accurate model adjustment [Geraldes, 2016].

GAMs are characterized by replacing in the GLM definition, the linear predictor $\eta = \sum_{j=1}^{p} \beta_j X_j$ by an additive predictor $\eta = \sum_{j=1}^{p} s_j(X_j)$. The local scoring technique is used for estimating the $s_j(\cdot)$, where, in order to allow the generalization of Fisher scoring procedure necessary for the calculation of maximum likelihood estimates, scatter-plot smoothers are used [Hastie & Tibshirani, 1986].

Considering a structure similar to the one presented in section 2.3, with a random component composed of a response variable Y and a vector of covariates $X_1, X_2, \ldots, X_p$, a linear regression model can also be defined as:

$$E(Y|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{2.10}$$

Based on the previous expression, a GAM can be easily defined by:

$$E(Y|X_1, X_2, \ldots, X_p) = s_0 + \sum_{j=1}^{p} s_j(X_j), \tag{2.11}$$

where $s_j(\cdot)$ are smooth standardized functions so the equality $E(s_j(X_j)) = 0$ can be true.

For a model as referenced in section 2.3, with equation (2.9) being rewritten as $\eta(\mathbf{X}) = g(\mu)$, where η is a function of p variables, it is now possible to write the expression for a multiple covariate additive model [Hastie & Tibshirani, 1986]:

$$\eta(\mathbf{X}) = s_0 + \sum_{j=1}^{q} s_j(X_j) + \sum_{j=q+1}^{p} \beta_j X_j, \tag{2.12}$$

where each function is estimated trough smoothing on only a coordinate at a time.

Estimation of the model is done by means of a back-fitting algorithm, an iterative process. Considering

$$E(Y|\mathbf{X}) = s_0 + \sum_{j=1}^{p} s_j(X_j), \tag{2.13}$$

where for every j the condition $E(s_j(X_j)) = 0$ is true, and the partial residual, defined as

$$R_j = Y - s_0 - \sum_{k \neq j} s_k(X_k), \tag{2.14}$$

then $E(R_j|X_j) = s_j(X_j)$ which minimizes $E(Y - s_0 - \sum_{k\neq j} s_k(X_k))^2$, allows for the estimation of each $\hat{s}_j(\cdot)$, for $j = 1, \ldots, p$.

The backfitting algorithm, where $s_j^m(\cdot)$ is the estimate at the m-th iteration of $s_j(\cdot)$ runs as follows:

Initialization: $s_0 = E(Y), s_1^1(\cdot) \equiv s_2^1(\cdot) \equiv \cdots \equiv s_p^1(\cdot) \equiv 0, \quad m = 0.$

Iterate $m = m + 1$ for $j = 1$ to p do:

$$R_j = Y - s_0 - \sum_{k=1}^{j-1} s_k^m(X_k) - \sum_{k=j+1}^{p} s_k^{n-1}(X_k)$$

$$s_j^m(X_j) = E(R_j|X_j).$$

Until: $RSS = E(Y - s_0 - \sum_{j=1}^{p} s_j^m(X_j))^2$ fails to decrease.

For a function of sample size n, $E(\hat{s}_j^m(X) - s_j^m(X))^2 \to 0$ as $n \to \infty$, when m is fixed.

2.5 Logistic Function

The logistic function is presented as the most common solution as a GLMs link function used for binary response data. It was invented in the 19th century for the description of the growth of populations and the course of *chain reactions* [Cramer, 2003].

If considered n independent response variables $Y_i \sim Binomial(1, \pi_i)$ then:

$$f(y_i|\pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \quad y_i = 0, 1 \tag{2.15}$$

and that each individual i, is associated with a specification vector $\mathbf{z}_i$, resulting of the covariate vector $x_i, i = 1, \ldots, n$. Since $\mu_i = E(Y_i) = \pi_i$ and since $\theta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$, by doing $\theta_i = \eta_i = \mathbf{z}_i^T \boldsymbol{\beta}$, it is possible to conclude that the logistic function is the canonical link function.

The probability of success, $P(Y_i = 1) = \pi_i$ is related to vector $\mathbf{z}_i$:

$$\pi_i = \frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{z}_i^T \boldsymbol{\beta})}. \tag{2.16}$$

The distribution function $F : \mathbb{R} \to [0, 1]$ can then be defined by:

$$F(x) = \frac{\exp(x)}{1 + \exp(x)}, \tag{2.17}$$

The visual representation of the corresponding link function of the logistic function can be seen in figure 2.1.

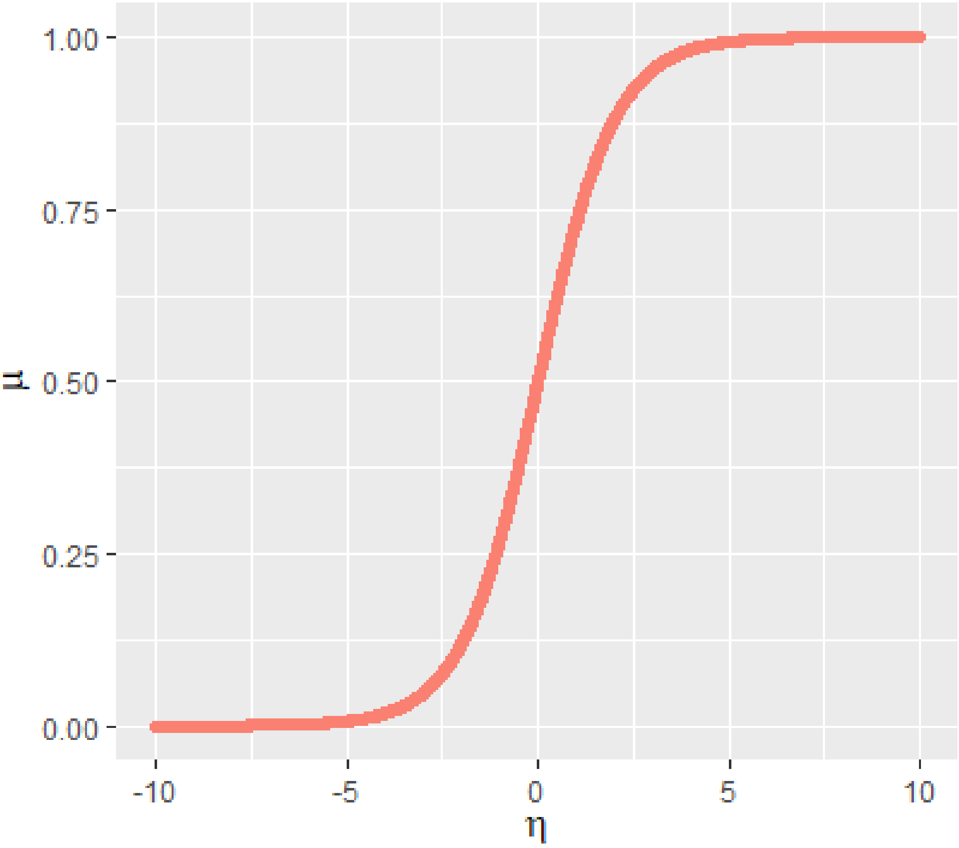

Figure 2.1: Graphic representation of the Logistic distribution function.

2.6 Aranda-Ordaz Asymmetric Family of Functions

As mentioned previusly, using a logistic function as a link function is the most common alternative used in GLMs for binary response data. Nevertheless, it may not be the most correct alternative as a link function, according to the data used. In order to overcome the limitation of using a logistic function, Aranda-Ordaz proposed in 1981 [Aranda-Ordaz, 1981], two new families of transformations for binary response data. These are extended models, which not only include the logistic distribution but also others, as special cases.

The asymmetric family, here presented, is most beneficial when talking about extreme value problems, for example. Considering $0 < \theta < 1$ denotes the probability of success and λ, $0 \leqslant \lambda \leqslant 1$, denotes the transformation parameter, a family designed to respond appropriately is:

$$W(\theta) = \frac{(1-\theta)^{-\lambda} - 1}{\lambda}. \tag{2.18}$$

Assuming:

$$\log W(\theta) = \tau, \tag{2.19}$$

where τ is real.

As mentioned, the logistic function represents a special case of the family of transformations here presented, for $\lambda = 1$, whereas the complementary log log model represents a special case for $\lambda = 0$. It can then be easily concluded that the models can be compared through a single parameter, here represented as λ.

The inverse of (2.19) is as follows:

$$\theta(\tau) = \begin{cases} 1 - (1 + \lambda e^{\tau})^{-\frac{1}{\lambda}} & \text{if} \quad \lambda e^{\tau} > -1, \\ 1 & \text{otherwise.} \end{cases} \tag{2.20}$$

The same structure mentioned in section 2.3, for GLMs, can be defined for the example here discussed. The same two components can be identified for m sets of independent observations, where for each set the probability of success is the same:

- The random component, where components Y_i have binomial distributions, $B(n_i, \theta_i)$, where n_i is the number of trials and θ_i is the probability of success in the *ith* set ($i = 1, \ldots, m$);
- The systematic component, a linear predictor defined by $\eta = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of unknown parameters and $\mathbf{X}$ is a specification matrix, as previously defined in section 2.3.

The link function can be defined by equation (2.20), and the moment parameter by $\mu_i = n_i\theta_i, i = 1, \ldots, m$, making the association between the former and the latter dependent on the family chosen.

Figure 2.2 allows for a visual representation of the flexibility characteristics of the Aranda-Ordaz family of asymmetric functions as a link function:

The code corresponding to the Aranda-Ordaz link function is available in Appendix A.

2.7 Czado Family of Functions

The usage of parametric link families, even though providing an improvement in terms of maximum likelihood fit compared to a more commonly used GLM, means an increase in the

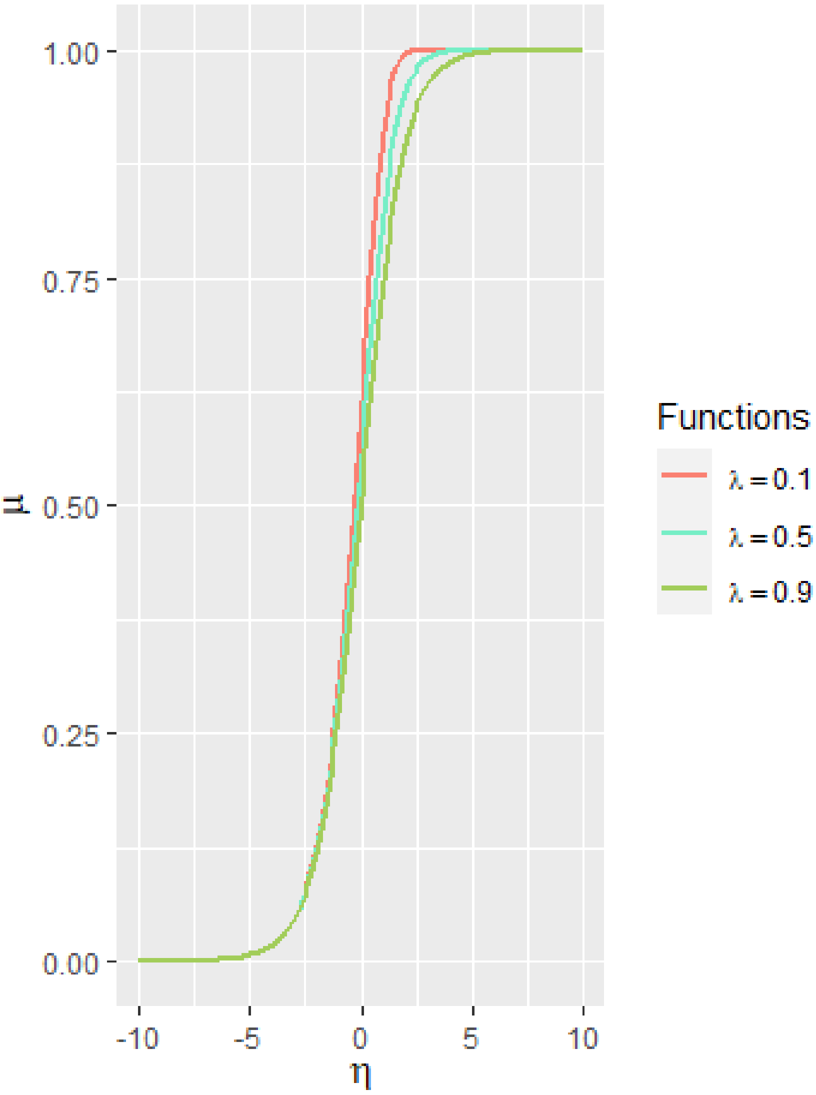

Figure 2.2: Graphic representation of Aranda-Ordaz distribution function.

variances of estimated regression coefficients and mean response predictions, leading to numeric instability and consequently to a more difficult interpretation of the model. However the Czado family of link functions, due to parameter orthogonality and standardization, allows for a reduction of variance inflation while maintaining the advantage of a better maximum likelihood fit.

The Czado family of link functions also adds to the flexibility allowed for the link function. Depending on three parameters, instead of just one similar to Aranda-Ordaz link function, allows for a greater fit to the data and problem at hand.

As in previous sections, the definition of a GLM should be kept in mind. As it is defined in section 2.3, it is comprised of a random component, Y_i, a systematic component, $\eta_i = \beta_0 + \mathbf{X_i^T}\boldsymbol{\beta}$ and a parametric link. The parametric link can be defined as $\mu_i = F(\eta_i, \boldsymbol{\psi})$ for some $F(\cdot, \boldsymbol{\psi})$ in $\{F(\cdot, \boldsymbol{\psi}) : \boldsymbol{\psi} \in \boldsymbol{\Psi}\}$, where $\mu_i = E(Y_i)$ and $\boldsymbol{\psi} = (\psi_1, \psi_2)$ [Czado, 1992].

The following expressions represent then the η_0 standardize parametric link function for GLMs proposed by Czado (1992). In the second expression the modification of the right tail is given by the first branch, and the left tail by the second branch [Geraldes, 2016].

$$h(\eta, \boldsymbol{\psi}) = \frac{e^{F(\eta,\boldsymbol{\psi})}}{1 + e^{F(\eta,\boldsymbol{\psi})}} \tag{2.21}$$

$$f(\eta, \boldsymbol{\psi}) = \begin{cases} \eta_0 + \frac{(\eta-\eta_0+1)^{\psi_1}-1}{\psi_1} & if \quad \eta \geq \eta_0, \\ \eta_0 - \frac{(-\eta+\eta_0+1)^{\psi_2}-1}{\psi_2} & if \quad \eta < \eta_0. \end{cases} \tag{2.22}$$

Figure 2.3 allows for a visual representation of the flexibility characteristics, varying the three different parameters, of the Czado family of functions:

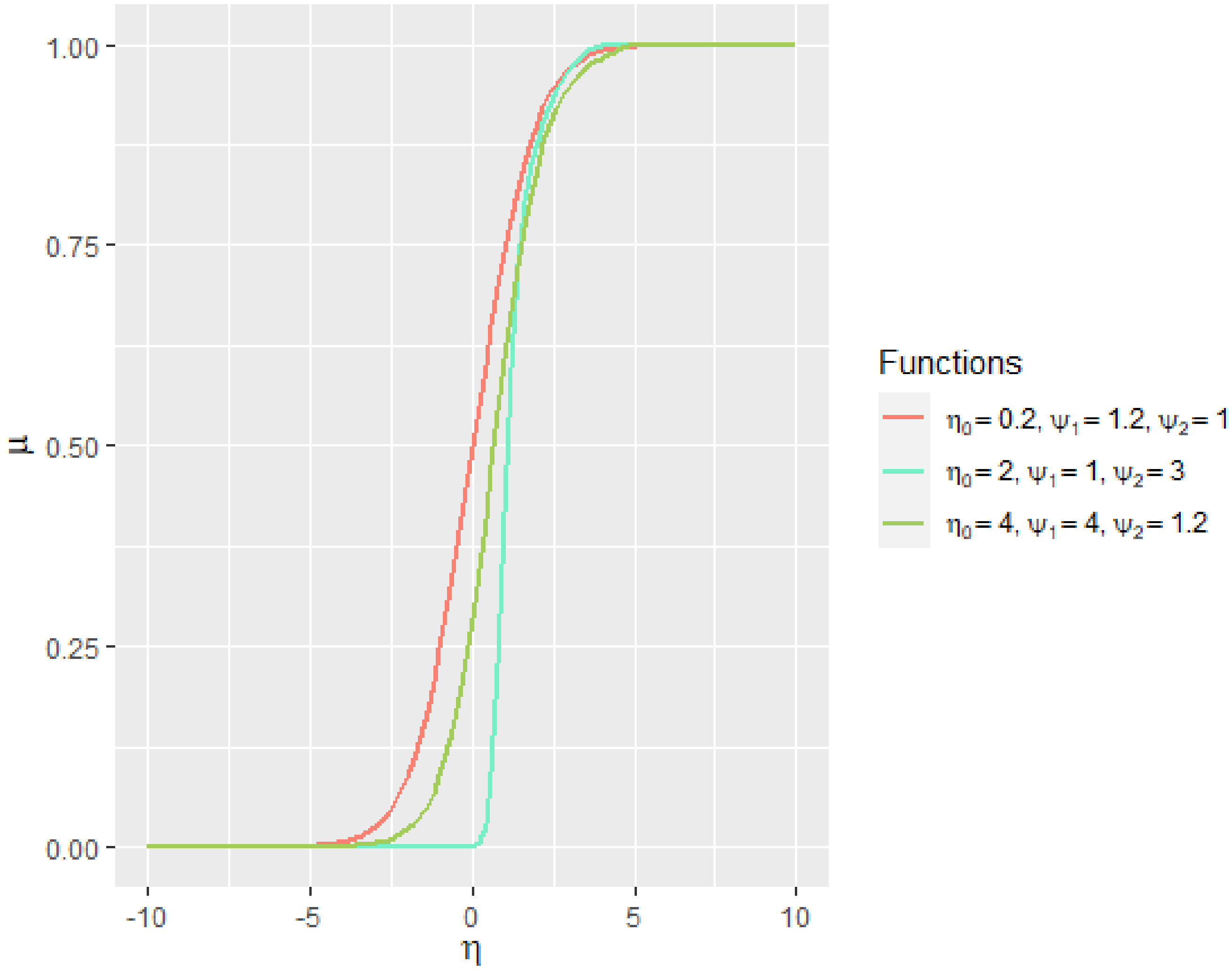

Figure 2.3: Graphic representation of Czado family of functions.

2.7.1 Parameter Orthogonality

As mentioned in the beginning of the chapter, being Czado a parametric link family can represent an improvement in terms of maximum likelihood fit, needing, however, parameter orthogonality for a reduction of variance inflation.

As a general example of what parameter orthogonality is, for a vector of length n of Y random variables, with density function $f_Y(y;\eta)$ depending on a p-dimensional vector of unknown parameters $\boldsymbol{\theta}$. Partitioning vector $\boldsymbol{\theta}$ into two, $\boldsymbol{\theta_1}$ of length p_1 and $\boldsymbol{\theta_2}$ of length p_2 where $p_1 + p_2 = p$, then $\boldsymbol{\theta_1}$ is orthogonal to $\boldsymbol{\theta_2}$ if the elements of the information matrix satisfy the following condition:

$$I_{\theta_s\theta_t} = \frac{1}{n}E\left(\frac{\partial l}{\partial\theta_s}\frac{\partial l}{\partial\theta_t};\boldsymbol{\theta}\right) = \frac{1}{n}E\left(-\frac{\partial^2 l}{\partial\theta_s\partial\theta_t};\boldsymbol{\theta}\right) = 0, \tag{2.23}$$

where I is the information for each observation, $s = 1,\ldots,p_1$ and $t = p_1+1,\ldots,p_1+p_2$, and l is the log-likelihood. Local orthogonality can always be achieved, occurring when the previous equation is only valid for a single parameter value, θ^0, yet global orthogonality can only be achieved in certain cases [Cox & Reid, 1987].

Consequently, a sufficient condition for the previous expression, even though only a local condition as $F(\cdot,\psi)$ would have to be independent of ψ, is given by:

$$\frac{\partial}{\partial\psi}F(\eta_i,\psi) = 0 \quad \text{for every} \quad 1 \le i \le n. \tag{2.24}$$

2.7.2 Parameter Standardization

For GLMs with parametric link belonging to the exponential family, both β_0 and $\boldsymbol{\beta}$ can be seen as parameters for finding the most suitable location and scale of the covariates, making the link family $F(\cdot,\psi_0)$ somewhat invariant in terms of location and scale.

Approaches can be defined in order for a link family to be called location and scale invariant, such as:

- η_0-location invariant if $\exists$ a value η_0 such that $F(\eta_0,\psi) = \alpha_0$ for all $\psi \in \Psi$.
- (η_1,η_2)-scale invariant if $\exists$ η_1 and η_2 such that $F(\eta_1,\psi) = \alpha_1$ and $F(\eta_2,\psi) = \alpha_2$ are independent of ψ.
- η_0-location invariant if $\exists$ a value η_0 such that $F(\eta_0,\psi) = \alpha_0$ and $\frac{\partial}{\partial\eta}F(\eta,\psi)|_{\eta=\eta_0}$ is independent of ψ for all $\psi \in \Psi$.

In terms of choosing a value for η_0, a reference given by Czado (1992) is, if in binary regression the observed proportions in a data set are approximately symmetric around 0.5, then $\eta_0 = 0$ is considered a good choice.

The R code corresponding to the Czado link function is available in Appendix A.

2.8 Evaluation Methods

As the main objective of this work is to evaluate the performance of different models and different link functions, having an efficient mechanism of comparison between them is of great importance. These measurements will be used for the selection of the best fit model in each combination possible, between GLMs, GAMs and the two types of link functions, Aranda-Ordaz and Czado, but also to determine the final model.

2.8.1 Brier Score

Brier Score was initially proposed by Brier (1950), in which he designed a verification scheme for forecasts expressed probabilistically [Brier, 1950]. Brier defined a verification score Sc, varying from 0 to 1, where 0 represents a perfect prediction, meaning the event is correctly predicted with a probability of 1, and 1 the least accurate predicted possible, in which a probability different than 0 is given for an event which did not occur.

Given an event occurring on n occasions, with r possible classes or categories, on occasion i, the forecast probabilities are $f_{t1}, f_{t2}, ..., f_{tr}$ that the event will occur in classes 1, 2, ..., r, respectively. The r possible classes have to be chosen in a manner the following condition must apply:

$$\sum_{j=1}^{r} f_{ij} = 1, i = 1, 2, 3, ..., n \tag{2.25}$$

The verification score Sc is defined as:

$$Sc = \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n} (f_{ij} - I_{ij})^2 \tag{2.26}$$

where I_{ij} according to the event happening in class j or not, assumes the value 1 or 0, respectively.

A variation of this score for assessing the accuracy of binary prediction was created, and its usage in terms of clinical data is increasing. The score addresses calibration, statistical consistency between the predicted probability and the observations and sharpness, defined as the concentration of the predictive distribution [Rufibach, 2010].

The Brier score, which in this case equals the mean square error of prediction, is the following:

$$\begin{gathered} \mathrm{B}(p, x) = n^{-1} \sum_{i=1}^{n} (x_i - p_i)^2 = \\ = n^{-1} \sum_{i=1}^{n} (x_i - pi)(1 - 2p_i) + n^{-1} \sum_{i=1}^{n} p_i(1 - p_i), \end{gathered} \tag{2.27}$$

where $p = (p_1, ..., p_n)$ refers to the predictive probabilities, with $0 \leq p_i \leq 1$, and n realizations $\mathbf{x} = (x_1, ..., x_n)$ of Bernoulli random variables $X_i \sim$ Bernoulli(π_i) with $0 \leq \pi_i \leq 1, \pi = (\pi_1, ..., \pi_n)$ and $x_i \in \{0, 1\}$, i=1, . . . , n.

2.8.2 Area under a ROC curve

The area under the ROC curve (AUC) represents the probability that a positive example, chosen randomly, is correctly rated with greater suspicion when compared to a negative example chosen at random [Bradley, 1997]. However, in order to understand what the AUC measurement is, it is necessary to understand firstly what a Receiver Operating Characteristics (ROC) curve is, and to do so, understand what Sensitivity and Specificity are.

In the case of a binary outcome, the classification can be exemplified in a contingency table, where the the concepts of true positives and true negatives can be visualized:

Table 2.1: Contingency table of predicted values *versus* actual values.

	Actual Values	
Predicted Values	Positive	Negative
Positive	True Positive (T_p)	False Positive
Negative	False Negative	True Negative (T_n)
	P	N

Sensitivity can be defined as the proportion of real positive cases that are correctly predicted positive [Powers, 2008]. Given T_p represents the number of true positives and P the number of real positive cases, then Sensitivity presents as follows:

$$Sensitivity = \frac{T_p}{P}. \tag{2.28}$$

Specificity is the proportion of real negative cases that are correctly identified as negative. Representing T_n as the number of correctly identified negative cases and N as the total number of negative cases, Specificity can be defined [Powers, 2008]:

$$Specificity = \frac{T_n}{N}. \tag{2.29}$$

However, Specificity in not directly used in this particular case, but the False Positive Rate which can be obtained as $1 - Specificity$.

Receiver Operating Characteristics (ROC) curves, are two-dimensional graphs were Sensitivity is displayed in Y axis and the False Positive Rate is plotted on the X axis, accounting for a visual display of the trade-offs between benefits, the true positives and costs, the true negatives [Fawcett, 2006]. This type of curve can be interpreted having the (0,0) point has a reference, which represents the strategy of never issuing a positive classification, meaning, even though there is no possibility of obtaining false positive cases, there are also no true positive cases. The point (1,1) represents the opposite strategy of issuing only positive cases. A perfect classification would then be obtained at point (0,1). The following graphic represents an example of a ROC curve.

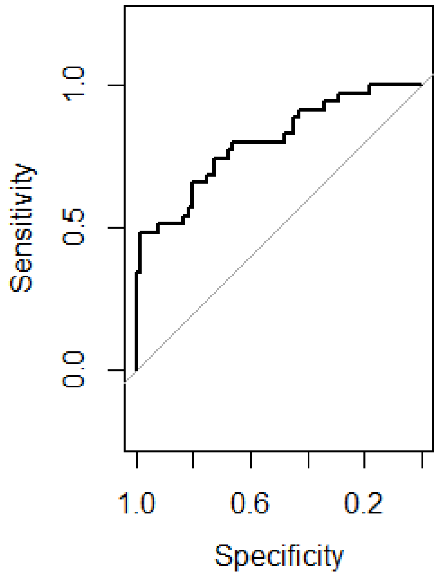

Figure 2.4: Example of a ROC curve.

In order to reduce ROC performance interpretability to a single scalar value, so that classifiers can easily be compared, the area under the ROC curve is usually calculated. Given it is an area under a unit square, AUC values vary between a minimum of 0.5 and a maximum of 1. One major characteristic of the AUC is that its value is equivalent to the probability that the classifier will rank a randomly chosen positive case higher than a randomly chosen negative case [Fawcett, 2006].

2.8.3 DeLong test for comparing AUCs

The DeLong test is a non-parametric approach to the analysis of two or more areas under correlated ROC curves, by means of the generalized U-statistics. The area under the points comprising an empirical ROC curve calculated by the trapezoidal rule is equal to the Mann-Whitney U-statistic, a statistic applied to two samples, $\{X_i\}$ and $\{Y_i\}$. The result is an estimated covariance matrix [DeLong et al., 1988].

Considering a sample of N individuals, where m of which undergo the event of interest, denominated C_1, and n individuals who did not undergo any occurrence of the event of interest, denominated C_2, and considering the definitions of sensitivity and specificity previously presented, the probability, θ, to randomly select an observation from the population represented by C_2 be less than or equal to randomly select an observation from the population represented by C_1, is presented below, as an average over a kernel, ψ:

$$\hat{\theta} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \psi(X_i, Y_j), \tag{2.30}$$

where

$$\psi(X, Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X. \\ 0 & Y > X \end{cases} \tag{2.31}$$

Generalizing the previous equation to k binary classifiers, where for observation i in C_1, X_i^k denotes classifier k estimated probability that it belongs to class 1. Likewise, Y_j^k can be defined for observations in C_2. The definition of k-th empirical AUC is:

$$\hat{\theta}^k = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \psi(X_i^k, \psi_j^k). \tag{2.32}$$

Considering $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)^T \in \mathbb{R}^K$ is the vector of K empirical AUCs, $\theta = (\theta_1, \ldots, \theta_k)$ is the vector of true AUCs, and $\mathbf{L}$ is a row vector of coefficients, in order to compare two AUCs, the null hypothesis is then given:

$$H_0 : \theta_1 = \theta_2, \quad i.e. \quad \mathbf{L}^T \boldsymbol{\theta} = 0.$$

and then the test statistic is given by:

$$\frac{\mathbf{L}\hat{\boldsymbol{\theta}}^T - \mathbf{L}\boldsymbol{\theta}^T}{\left[\mathbf{L}\left(\frac{1}{m}\mathbf{S}_{10} + \frac{1}{n}\boldsymbol{S}_{01}\right)\mathbf{L}^T\right]^{\frac{1}{2}}} \sim \mathcal{N}(0,1) \quad \text{under} \quad H_0. \tag{2.33}$$

The elements (r,s)th of a matrices $\boldsymbol{S}_{10}$ and $\mathbf{S}_{01}$ of size $K \times K$ are defined as:

$$s_{10}^{r,s} = \frac{1}{m-1} \sum_{i=1}^{m} [V_{10}^r(X_i) - \hat{\theta}^r][V_{10}^s(X_i) - \hat{\theta}^s], \tag{2.34}$$

and

$$s_{01}^{r,s} = \frac{1}{n-1} \sum_{j=1}^{n} [V_{01}^r(Y_j) - \hat{\theta}^r][V_{01}^s(Y_j) - \hat{\theta}^s], \tag{2.35}$$

where

$$V_{10}^r(X_i) = \frac{1}{n} \sum_{j=1}^{n} \psi(X_i^r, Y_j^r) \quad (i = 1, 2, \ldots, m), \tag{2.36}$$

and

$$V_{01}^r(Y_j) = \frac{1}{m} \sum_{i=1}^{m} \psi(X_i^r, Y_j^r) \quad (j = 1, 2, \ldots, n). \tag{2.37}$$

If the p-value associated with the test statistic (2.33) is inferior to the significance level considered, usually of 0.05, it is assumed there is statistical evidence to reject the null hypothesis.

Chapter 3

Results

The following chapter is a compilation of all the results obtained throughout the development of this work. It is divided into five sections, namely, Exploratory Data Analysis, Variable Selection, Aranda-Ordaz GLM and GAM Model Estimation, Czado GLM and GAM Model Estimation, and Model Comparison. Each section results from the application of the previously discussed methods to the data collected from patients admitted to *São José* Hospital.

3.1 Exploratory Data Analysis

As stated in section 2.2, the first step in a thorough statistical analysis is the exploratory analysis. The data is comprised of eight independent variables, one of which, Ventilated is binary, being the remaining continuous. The outcome variable, Death, is also binary.

Firstly, all continuous covariates were analysed to determine whether the population from which each sample was drawn, is symmetric or asymmetric, in order to most accurately present summary statistics. The Cabilio and Masaro symmetry test was used, referenced in section 2.2. For variable blood pressure (BPre) the p-value obtained was close to 0, which, considering the commonly used significance level of 0.05, indicates there is evidence towards rejecting the null hypothesis, being the variable distribution considered asymmetric. For serum sodium level (SSLev) the p-value obtained was 0.17 which means there is no evidence for rejecting the null hypothesis and the population is considered symmetric. Variable urinary output (UOut) had a p-value close to 0, which means the population from which the sample was taken from is considered asymmetric. Covariate Age had a p-value close to 0 as well , which according to the significance level of 0.05 considered, accounts for a variable with an asymmetric distribution. For serum urea level (SULev) the p-value obtained was equally close to 0, meaning the distribution from which the variable is drawn is considered asymmetric. Variable bilirubin level (BLev) had a p-value close to 0 which leads to the immediate conclusion that the population from which the sample was drawn is asymmetric. Finally, for serum bicarbonate level (SBLev) the p-value obtained was close to 0, similarly to the majority of variables here discussed, being the variable distribution considered asymmetric as well.

The following tables summarize the location and variability statistics for all the continuous independent variables considered further for variable selection. The variables are divided according to the symmetry of their distributions, or lack there of, and the statistics presented are accordingly to such distinction, already presented. Variables with symmetric distribution should be represented by their mean, which in fact should coincide with the median, and standard deviation, while variables with an asymmetric distribution should be represented by their median, first and third quantiles.

Table 3.1. presents the only variable with a symmetric distribution, showing statistics as the minimum (Min.), maximum (Max.), mean and standard deviation (Std. Dev.). It is possible to see SSLev varies from 119 to 164 with a mean of 138.556 and a median of 139.

Table 3.1: Summary statistics for the variable with a symmetric distribution, SSLev.

	Min.	Mean	Max.	Std. Dev.	Median
SSLev	119.000	138.556	164.000	7.722	139.000

On the other hand, table 3.2. shows variables with an asymmetric distribution, with statistics such as the minimun, maximum, median, the first (1st Qu.) and the third quantile (3rd Qu.). For BPre the median is 97, while for UOut is 2.2. Variable Age has a median of 63 years, and SULev has a median of 57. BLev has a median of 0.89 and the median of the variable SBLev is 20.1.

Table 3.2: Summary statistics for continuous variables with an asymmetric distribution.

	Min.	1st Qu.	Median	3rd Qu.	Max.
BPre	30.000	78.000	97.000	147.000	268.000
UOut	0.000	1.500	2.200	3.090	7.775
Age	14.000	46.500	62.000	73.000	100.000
SULev	6.000	31.000	57.000	95.500	384.000
BLev	0.100	0.530	0.890	1.600	39.900
SBLev	2.000	16.100	20.100	27.000	59.900

Table 3.3 presents the correlation matrix between covariates. It is possible to see there is a light tendency for BPre to increase as UOut also increases and as well that SULev increases as Age increases. On the other hand there is a tendency for SULev to decrease as UOut increases and vice versa. However, none of the values presented in table is high, meaning covariates are very little correlated between themselves, having a very weak linear relationship. In terms of modeling, this lack of correlation between variables presents itself as favourable when performing a regression analysis, since it indicates there is no sign of existence of multicollinearity and therefore the statistical significance of an independent variable is not undermined due to it being correlated to another independent variable.

Table 3.3: Pearson correlation matrix.

	BPre	SSLev	UOut	Age	SULev	BLev	SBLev
BPre	1.00	-0.02	0.26	-0.09	-0.16	-0.08	0.13
SSLev	-0.02	1.00	0.00	0.07	0.01	0.02	0.00
UOut	0.26	0.00	1.00	-0.17	-0.20	-0.06	0.13
Age	-0.09	0.07	-0.17	1.00	0.25	-0.06	0.10
SULev	-0.16	0.01	-0.20	0.25	1.00	0.18	-0.20
BLev	-0.08	0.02	-0.06	-0.06	0.18	1.00	-0.17
SBLev	0.13	0.00	0.13	0.10	-0.20	-0.17	1.00

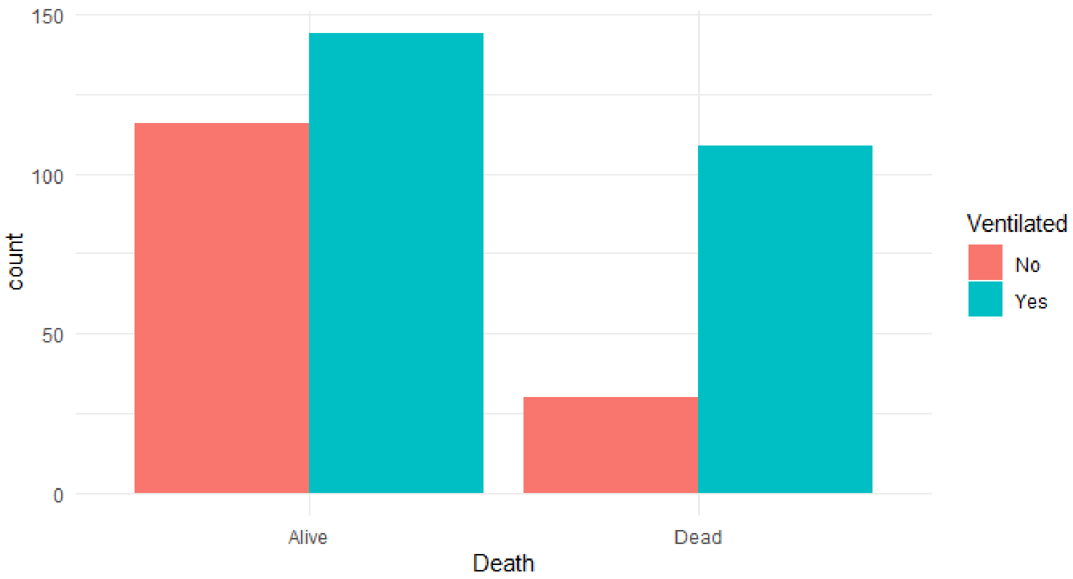

Figure 3.1: Count of ventilated patients at time of admission according to observed outcome variable, Death.

Regarding the only categorical variable, Ventilated, figure 3.1 was built. The graph allows for a visual representation of the outcome observed, Death, according to the patient being ventilated, or not at the moment of arrival at the hospital. Among patients who survived, 116 were not ventilated, whereas 144 were. For patients who died, only 30 were not ventilated while 109 were indeed ventilated. A distinction can be made in the relationship between the necessity of ventilating a patient and its final outcome. This distinction can be an indicator that variable Ventilated does indeed explain at least to some extent the outcome variable and should then be included in the final model.

The following set of graphics display each continuous variable through boxplots against each level of the outcome variable, Death.

Variable Age appears to be fairly equally distributed either the event of interest, Death, occurred or not. The median of ages for patients who did not die is 60, while for patients who died is a little higher, being 66. No outliers are observed in figure 3.2. The similarity between medians and overall distribution of both categories of Death, may be an indication the variable does not explain very clearly the outcome variable and may be a potential candidate to be excluded from the final model.

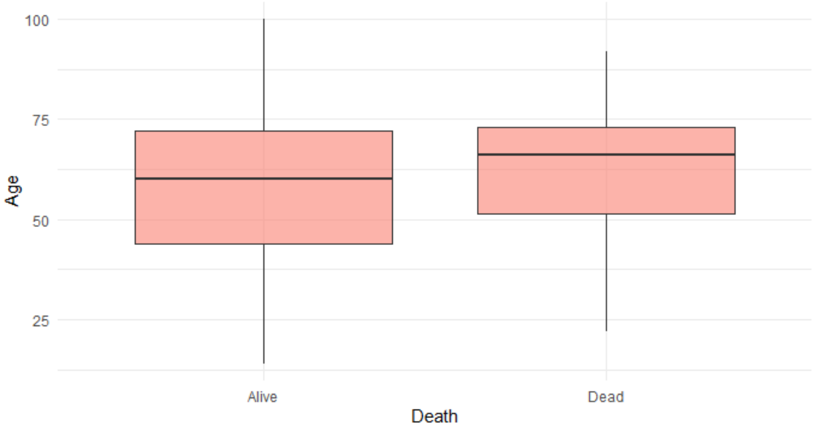

Figure 3.2: Boxplot of Age according to the outcome variable, Death.

For variable BLev the case is different. The majority of values are low, and it is possible to see the presence of a few outliers in figure 3.3. The median for individuals who did not die is of 0.86 and for individuals who died is 1. The maximum value of BLev for deceased individuals is 39.9, while for individuals who did not die is 21.3. In this particular case, the medians do not seem too different between both categories of variable Death, however the clear distinction of extreme values given by the outliers seems to be an indication of how the variable can contribute to explain the outcome variable, and therefore be a good candidate to be included in the model.

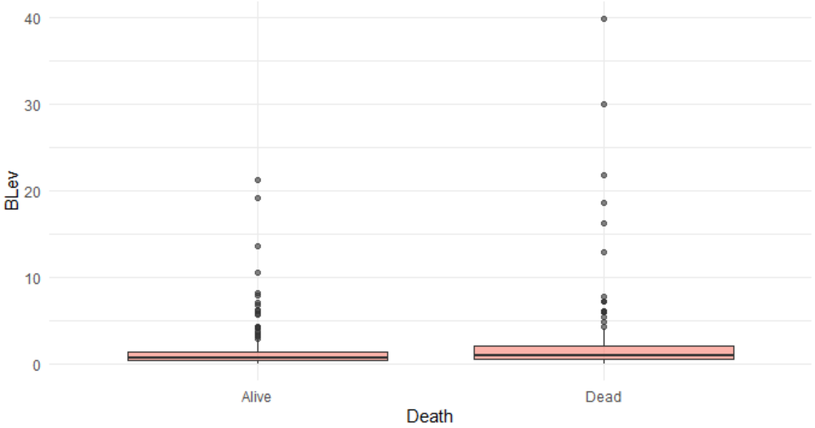

Figure 3.3: Boxplot of BLev according to the outcome variable, Death.

Boxplots corresponding to variable BPre are shown in figure 3.4. The boxplots are quite distinct according to the levels of the outcome variable. The median for individuals who did not present the event of interest is 103 while the median for the individuals who died is 81. For the first group mentioned no outliers are observed, while for the second group it is possible to observe their presence, being the highest value 214. Given there is such a clear difference in the distribution of values of BPre between categories of the variable Death, BPre is a strong candidate to be included in the final regression model.

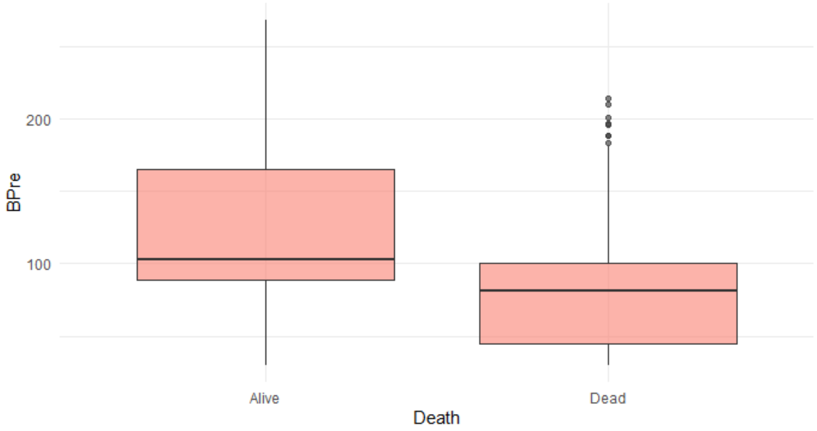

Figure 3.4: Boxplot of BPre according to the outcome variable, Death.

Variable SBLev (figure 3.5) has, for both categories of the outcome variable, a similar median. For patients who did not suffer the event of interest the median is 20.9, and for patients who were deceased by the third day of hospitalization the median is 18.7. Both levels have outliers associated to them, where for category 0 of variable Death the highest outlier corresponds to 59.9, whereas for category 1 of variable Death the solely outlier has a value of 45.1. SBLev presents some differences across both categories fo variable Death, being therefore a potential candidate to be included in the regression model.

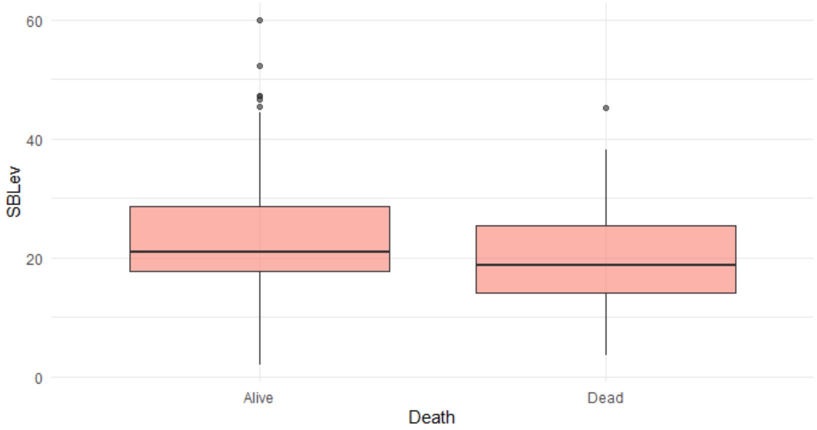

Figure 3.5: Boxplot of SBLev according to the outcome variable, Death.

The next variable to be graphically analysed (figure 3.6), SSLev possesses the same median for both levels of variable Death, 139, which can be verified visually through the boxplots. The difference between both groups lies in the distribution of the number of individuals throughout the values of SSLEV. From the analysis of the boxplot corresponding to the group of patients who died, there are no visible outliers, and values range from 119 to 161. As for the remaining group, values vary from 122 to 164, corresponding the latter to an outlier. As the distribution of SSLev between both categories of the outcome variable is so similar, it seems it does not contribute greatly for the explanation of Death and seems to be likely its exclusion from the final regression model.

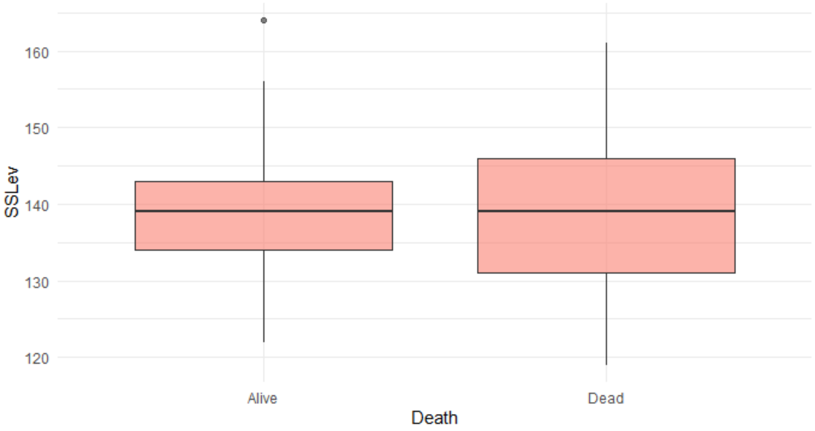

Figure 3.6: Boxplot of SSLev according to the outcome variable, Death.

Variable SULev has quite disparate medians, as seen in figure 3.7. For patients who died, the median is 87, while for the other group the median is 45. Both groups possess quite a few outliers, where for the group where patients died the outlier furthest from the median has a value of 384, and for remaining group the highest outlier has a value of 297. Variable SULev seems a good candidate to be included in the final model, since the disparity between medians implies that the variable ads a significant contribution to the explanation of the outcome variable, Death.

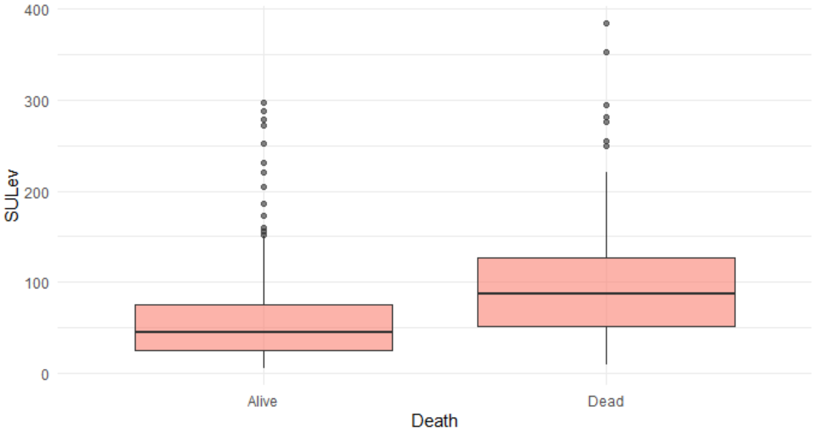

Figure 3.7: Boxplot of SULev according to the outcome variable, Death.

Finally, boxplots in figure 3.8 correspond to the graphical representation of variable UOut according to the two levels of the outcome variable Death. For UOut the median is higher for the group of patients who did not die, with a value of 2.467, while for the group were patients who did, the median is 1.7. As in the previous analysed variable, both groups present outliers. The furthest from the median in the group with level 0 for the outcome variable is 7.775, while for the other group the highest value corresponds to 7.4, much closer together in value when compared to the median values. Similarly to variable SULev, UOut distribution differs between both categories of variable Death, and therefore is a good candidate to be included in the final regression model.

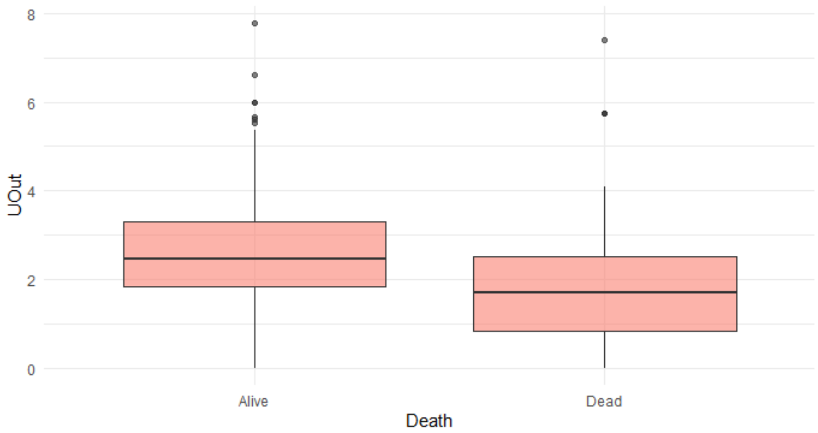

Figure 3.8: Boxplot of UOut according to the outcome variable, Death.

3.2 Variable Selection

As explained in section 2.2., the variable selection takes great importance in order to find the best fitting model to the data. Table 3.4 represents the result of the function *summary()* in R, for each univariable model.

For variable BPre, the outcome of the univariable model with Death as the dependent variable, puts BPre as a significant variable. The p-value of the variable in the model is close to zero, being lower than 0.25, the criteria for variable selection previously explained. According to such criteria the variable will be considered for the final model. Variable SSLev is considered not significant, as represented below, as its p-value in the model is largely over the threshold of 0.25. As for variable UOut, the outcome below, shows a p-value of approximately 0 for the univariable model. The p-value is close to zero, making the the variable important to include in the final model. The p-value for SULev is similar to the one for UOut, and so logically, SULev is also of importance to include in the final model. Variable Age, which output can be seen below, has a p-value of 0.016, when its univariable model is built. Since the p-value is below 0.25, this variable should also be taken into account for the final model. Variable BLev has, as shown below, a p-value of 0.0112 as a covariate in its univariable model. Taking into account the rule in which every p-value below 0.25 should make the variable it is linked to, of importance to include in the final model, BLev is then selected for such purpose. SBLev, shows a p-value of approximately 0, considerable below 0.25, and given so, SBLev should be considered for the final model. Finally, variable Ventilated, has a p-value close to zero, and for that, should be considered for the final model.

The analysis above, allows for a model in which only the variable SSLev should be excluded from it, between the variables taken into consideration. The generalized linear model is as follows:

$$y = \beta_0 + \beta_1(BPre) + \beta_2(UOut) + \beta_3(Age) + \beta_4(SULev) + \beta_5(BLev) + \beta_6(SBLev) + \beta_7(Ventilated) + \varepsilon$$

Table 3.4: Results for the univariable logistic regression model for each variable.

	Estimate	Std. Error	z value	Pr(> |z|)
BPre	-0.0199	0.0030	-6.731	< 0.0001 ***
SSLev	-0.0110	0.0137	-0.808	0.4190
UOut	-0.6116	0.1031	-5.934	< 0.0001 ***
SULev	0.0124	0.0021	-5.934	< 0.0001 ***
Age	0.0153	0.0063	2.419	0.0155 *
BLev	0.0930	0.0367	2.536	0.0112 *
SBLev	-0.0478	0.0133	-3.591	0.0003 ***
Ventilated	1.0739	0.2410	4.456	< 0.0001 ***

[*] Statistically significant at 0.05 significance level.
[**] Statistically significant at 0.01 significance level.
[***] Statistically significant at 0.001 significance level.

Further analysis can be done, by exploring the significance of each variable when the resulting model is compiled. The following table presents the coefficients corresponding to the model mentioned above, in which variable Age seems to no longer have a statistical importance, using 0.25 as the threshold, once the other variables are taken into account for the regression model.

Table 3.5: Coefficients for model: $Death \sim BPre + UOut + Age + SULev + BLev + SBLev + Ventilated.$

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.6529	0.6923	0.943	0.345583
BPre	-0.0132	0.0031	-4.303	< 0.0001 ***
UOut	-0.3824	0.1068	-3.579	0.0003 ***
Age	0.0037	0.0077	0.477	0.6330
SULev	0.0086	0.0023	3.712	0.0002 ***
BLev	0.0620	0.0404	1.533	0.1252
SBLev	-0.0231	0.0148	-1.558	0.1193
Ventilated1	0.6914	0.2824	2.448	0.0144 *

[*] Statistically significant at 0.05 significance level.
[**] Statistically significant at 0.01 significance level.
[***] Statistically significant at 0.001 significance level.

Considering the AIC, explained in section 2.2., as the final criteria for variable selection, a backward elimination can be performed, The model $Death \sim BPre + UOut + Age + SULev + BLev + SBLev + Ventilated$ has an AIC of 413.4. Removing variable Age, still considering the 0.25 significance level, makes the AIC drop, as desirable, to 411.63. If a more conservative p-value is used, the ever so common 0.05, variables BLev and SBLev should also be removed from the model. However, if BLev is removed the AIC goes up to 411.99, if SBLev is removed AIC goes up to 411.9, and if both are removed the final AIC is of 412.93. Although AIC differences are minor, it goes up if BLev and SBLev are removed, making for the conclusion only Age should be taken from the final model. The final model is then presented:

$$y = \beta_0 + \beta_1(BPre) + \beta_2(UOut) + \beta_3(SULev) + \beta_4(BLev) + \beta_5(SBLev) + \beta_6(Ventilated) + \varepsilon$$

3.3 Logistic Model Estimation

After variable selection, it is possible to start analysing different models using both families of link functions Aranda-Ordaz and Czado, and both GLMs and GAMs. A train dataset containing seventy-five percent of the total dataset was used for model training and a test dataset, containing the remaining twenty-five percent of the data was used for testing each model predictive capability. The baseline model used for comparison was a GLM using the logistic link function (figure 2.1), since the main goal is to try to find improvements to using such a conservative link function.

Each value of AUC and the Brier score, used for evaluating each model predictive capability, was stored accordingly to each value used for the functions' parameters, either λ for Aranda-Ordaz function or ψ_1, ψ_2, η_0 for Czado function. The DeLong test was also included in the analysis, were a p-value below the significance level of 0.05 entails there is evidence to refuse the null hypothesis, which states the difference between the AUC for the baseline logistic model and the AUC for the model at test is not significant.

Table 3.6 presents the results obtained for the GLM using the logistic function as its link function. The results of the logistic function will be reproduced in each analysis, under the corresponding parameters' values within each function for a better visualization of significant differences and better comparison.

Table 3.6: Results of AUC, p-value from the DeLong test and Brier score for model GLM using the logistic function as the link function (baseline model).

Link function	AUC	p-value	Brier
Logistic	0.8013	1.0000	0.1657

3.4 Aranda-Ordaz GLM and GAM Model Estimation

The first model to be analyzed is a GLM with Aranda-Ordaz link function. Only the variables selected previously were used when building this model. The single parameter used in the Aranda-Ordaz asymmetric function was varied between the values of 0 and 1, by steps of 0.01.

Table 3.7 presents itself as a summary of the highest values for the measures used, values of AUC and Brier score. The first line of the table also presents the results obtained for the baseline model using a logistic link function. The following two lines represent the results corresponding to the highest AUC value and Brier score, respectively. The third column corresponds to the p-value associated with the DeLong test.

The values of AUC varied from 0.8008791 when λ is 0.98, to 0.8061538 when λ is 0 (figure 3.9). On the other hand, the Brier score varied from 0.1650408 when λ is 0.34 to 0.1657095 when λ is 1. As the variation of Brier score is so little, making it impossible to fully perceive the difference in prediction capability of the models, taking conclusions considering solely the AUC values is advisable. The best model obtained was for λ equal to 0, which corresponds to the log log link function.

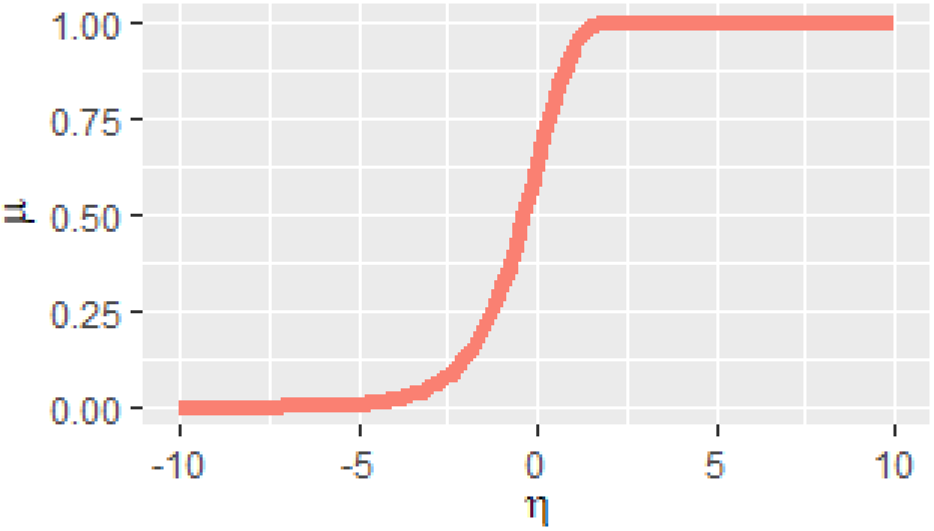

Figure 3.9: Complementary log log link function, corresponding to $\lambda = 0$.

Table 3.7: Results of AUC, p-value from the DeLong test comparing AUCs to the logistic link function and Brier score for model GLM with Aranda-Ordaz as link functions.

Link function	λ	AUC	p-value	Brier
Logistic	1	0.8013	1.0000	0.1657
Aranda-Ordaz	0	0.8062	0.6614	0.1656
	0.34	0.8031	0.7815	0.1650

Nevertheless, the DeLong test shows, through a p-value of 0.6614 that there is no statistical evidence to reject the null hypothesis, at the significance level of 0.05, meaning the difference between values of AUC for both models, logistic and log log, is not sufficient to be considered statistically significant.

Figure 3.10 represents the variation of AUC values according to the values of λ. It is possible to see how the highest value is clearly associated with the smallest value for the parameter, 0. The variation of the value of the AUC is not constant, dropping initially until it reaches the value of 0.8013187 when λ is 0.45, augmenting to 0.8043956 when λ is between 0.74 and 0.78, and dropping to its lowest value, 0.8008791 when the parameter is equal to 0.98.

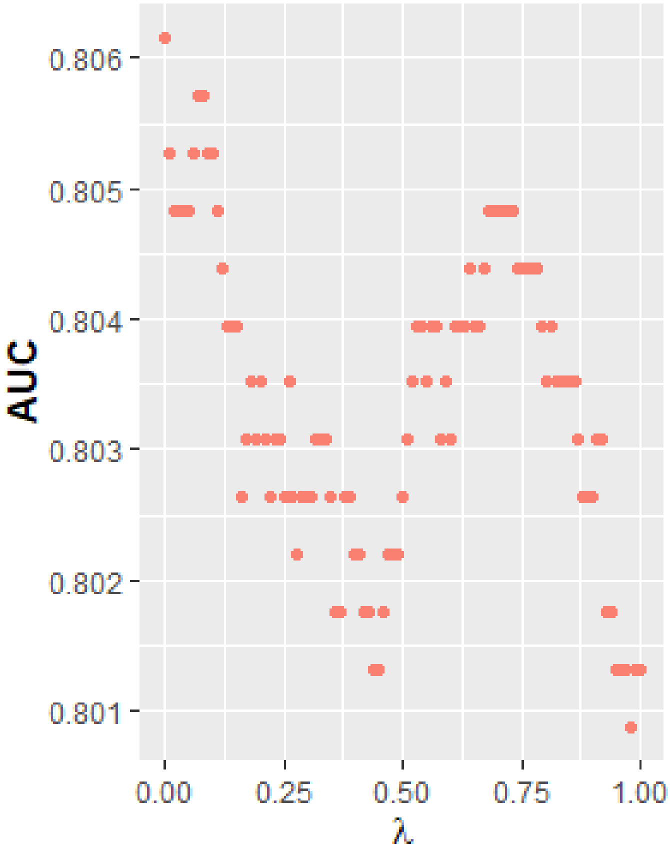

Figure 3.10: AUC values according to different Aranda-Ordaz distribution parameter, for GLM.

The values for Brier score can be seen in figure 3.11, presenting a variation according to the values of the parameter, represented as λ. It starts with a value of 0.1655828 when λ is 0, dropping to its lowest value of 0.1650408 when λ is 0.34. Its highest value is when the parameter equals 1, with a value of 0.1657095. Between the highest and the lowest value for the coefficient the variation is only of 0.0007, which can be considered too small a difference to make any correct assumptions based on it, as mentioned.

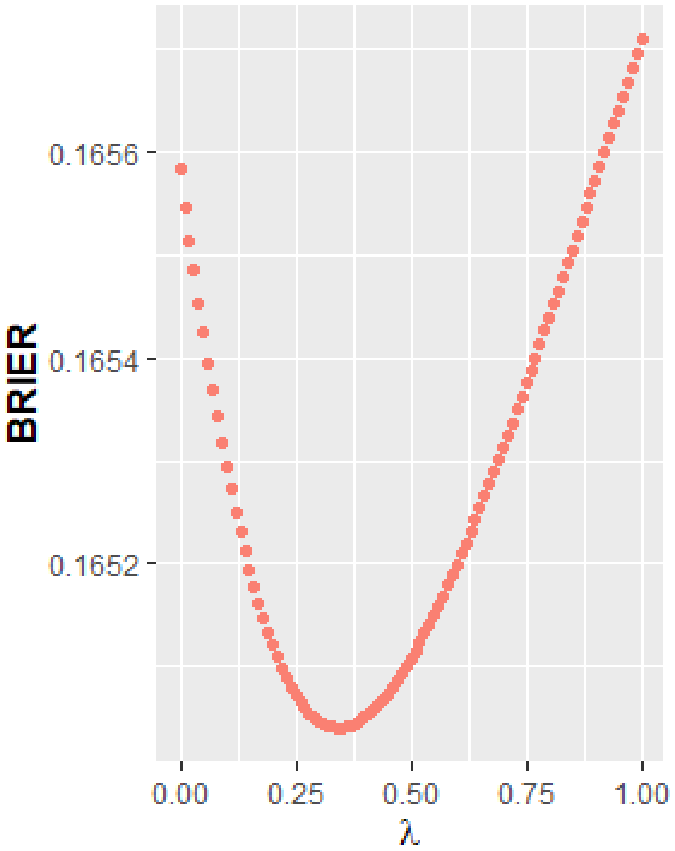

Figure 3.11: Brier coefficient values according to different Aranda-Ordaz distribution parameter, for GLM.

When the exact same analysis was performed using a GAM instead of a GLM the same results were obtained. The analysis was performed using the variation of parameter λ between 0 and 1, with steps of 0.01. The variables included were the same previously selected ones. The variations of AUC values and Brier score were the same, with exactly the same highest and lowest values. The Brier score variation was then too small to take any proper conclusions and the difference in AUC values between the baseline model and the model at test were not considered significant. The unexisting difference between the usage of either GLM or GAM indicates that the assumption of a linear relation between the outcome variable and covariates is correct.

3.5 Czado GLM and GAM Model Estimation

In order to continue the pursue of a better fitting link function, the next analysis was performed using the Czado family of functions as the link function. As mentioned, the inverse of equations (2.21) and (2.22) were used. However, the inverse function has some limitations to its usage, where the following condition $\psi_1 \times (\mu - \eta_0) + 1 > 0$, where μ is the expected value and η corresponds to the input data, must be met in order for the function to be able to retrieve a value.

Table 3.8: Results of AUC, DeLong test comparing AUCs to the logistic link function and Brier score for model GLM using Czado as the link function.

Link function	η_0	ψ_1	ψ_2	AUC	p-value	Brier
Logistic	1.00	1.00	0.00	0.8013	1.0000	0.1657
Czado	2.00	1.00	0.50	0.8018	0.9580	0.1666

The function was tested with different combinations of values for the three parameters at test. For ψ_1 and ψ_2 between 1 and 5 by steps of 1, and η_0 between 0 and 5 by steps of 0.5. If the combination of any of these values did not follow the condition previously shown, the function could not be applied, and the respective values were excluded from the analysis.

The same method as for Aranda-Ordaz was applied. Both GLM and GAM were used, and values of AUC and Brier score were obtained. Table 3.8 summarizes these results, along with the DeLong test for a GLM, having only the highest values for each statistic.

The AUC values varied from 0.2692308, when $\psi_1 = 1, \quad \psi_2 = 4, \quad \eta_0 = 0.0$ to 0.8017582, when $\psi_1 = 2, \quad \psi_2 = 1, \quad \eta_0 = 0.5$, and the Brier score from 0.1657095, when $\psi_1 = 1, \quad \psi_2 = 1, \quad \eta_0 = 0.0$, to 0.7224080, when $\psi_1 = 1, \quad \psi_2 = 5, \quad \eta_0 = 1.5$. Taking into consideration the Brier score, it indicates, still, the best solution is the logistic function. However, AUC values are best when parameters are $\psi_1 = 2, \quad \psi_2 = 1, \quad \eta_0 = 0.5$. The DeLong test, with a p-value of 0.9579531 indicates the difference from the AUC value of the baseline logistic model and the Czado link function model is not statistically significant at a 0.05 significance level.

The link function corresponding to the best model obtained can be seen in figure 3.12. The slightest difference in curvature in the right branch of the function can be seen especially when compared to figure 2.1.

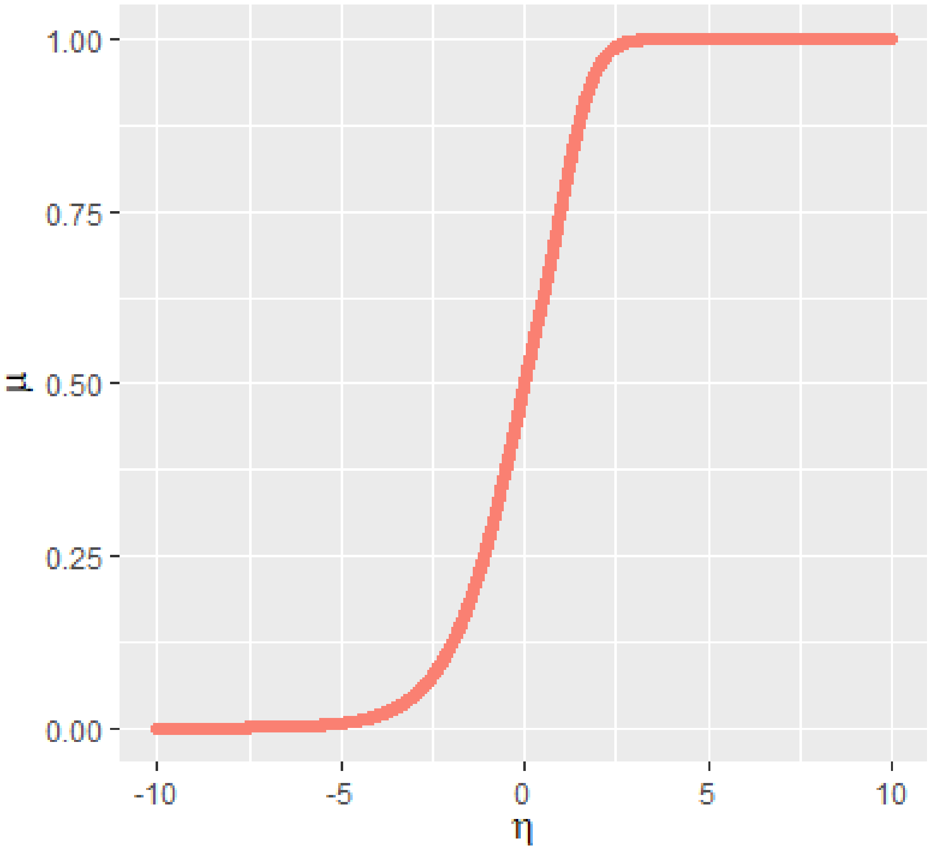

Figure 3.12: Czado function for parameters $\psi_1 = 2, \quad \psi_2 = 1, \quad \eta_0 = 0.5$.

When using a GAM, the results, similarly to what happened for Aranda-Ordaz, are the same as for the GLM. However more values were excluded from the analysis due to the incapability of building a model when the necessary condition of the inverse Czado function cannot be met. The conclusions are, then, similar to the ones presented for the GLM, adding the indication, once more, that assuming a linear relation between covariates and the outcome variable in this particular case seem to be correct.

3.6 Model Comparison

Considering solely the Aranda-Ordaz link function, no major improvements were observed regarding both the model used and the link function. Using a GLM and a GAM did not make a difference, indicating that having a prior assumption of a linear relationship between a link function of the expected response variable and the explanatory variables may be correct in this context. The link function Aranda-Ordaz did not produce any improvement in the specif conditions of this work, making the usage of another link function important in order to better understand the impact of using a parametric link function.

Regarding the Czado link function, however its greater flexibility, the results obtained were not considered statistically significant when compared to the logistic regression model. Considering the Brier score, the best result was obtained using the logistic link function, even though it was not the case considering the AUC as a measure of capability of the regression to correctly predict the classification of each individual in each category of the outcome variable. However, as mentioned, the difference between the values of the AUC corresponding to the best fitted model and the logistic regression model were not statistically significant according to the DeLong test. As occurred fo the Aranda-link function there was no distinction in results between the GLM and GAM obtained. This conclusion only strengthens the belief a linear relation between the outcome variable and the model covariates can be assumed.

By soling comparing AUC values, given the Brier score was discarded as informative for the Aranda-Ordaz link function model and considering none of the models obtained had statistically significant differences to the logistic function in their AUC values, it is possible to determine which model produced better estimations for the outcome variable. Using the Aranda-Ordaz link function, the best AUC value obtained was 0.8061538, for a parameter value of 0, corresponding to the log log model. The best model obtained with Czado GLM had an AUC value of 0.8017582, when $\psi_1 = 2$, $\psi_2 = 1$, $\eta_0 = 0.5$. Comparing AUC values for both models, the log log model presents a higher value of AUC, indicating it is the best data fitting model, for the data at study.

Chapter 4

Conclusion & Discussion

The aim of this work is to present a more accurate alternative when it comes to building prognostic models in a medical context. In order to try to achieve such a challenging goal, the strategy utilized was to test different models, GLMs and GAMs, with never before used link functions, which allow for more flexibility than commonly used link functions.

The constant search for better statistical methods to be applied in the medical field, reflects a necessity for statistical models which can approximate themselves to reality. Several times response variables are considered to have a normal distribution, when in reality the normal distribution is indiscriminately applied to data which should be handled otherwise, specially when considering count data [Lindsey & Jones, 1988].

GLMs have still been quite unexplored in a clinical context, except for the use of normal and logistic models [Lindsey & Jones, 1988]. With this in mind, the work here developed enables a solution for a still limited statistical approach in the medical field. It explores not only GLMs but also GAMs. GLMs are well known for analyzing dependencies between a possibly non normal outcome variable and a number of covariates. However, GAMs allow for the incorporation of non-parametric covariate effects [Czado et al., 2010].

Is is possible to find some work already related to the use of more versatile link functions [Geraldes, 2016, Li et al.. 2015]. Even though Aranda-Ordaz was already used for a similar study [Geraldes, 2016], it was not applied to a GAM, and, as the versatility of each link function is important to try to better explain the relation between the covariates and the outcome variable, different sets of data should be explored in order to fully understand how well the link function can adapt to the data in question. Regarding the Czado family of functions, no previous work was found to use it as a link function, despite its favorable adaptability properties and robustness.

The Aranda-Ordaz link function can be applied to binary response data and has a transformation parameter, which already represents added flexibility to the commonly used link functions. The Czado link function relies on three distinct parameters, allowing for a growing increase on the adaptability of the link function. Both link functions mentioned have the logistic function as a special case, becoming this the baseline function for comparison throughout this work.

Firstly, an exploratory analysis was performed in order to understand the data used. Even though the aim of the study was not to draw any clinical conclusions, understanding the behaviour of each variable before building any model is essential. The exploratory analysis, through a boxplot, gave the indication the SSLev variable, since it had the exact same median for both levels of variable Death, and possibly the Age variable for its similarity in distribution for both categories of the outcome variable, added no relevant information about variable Death, and that they might have been good candidates to be excluded from the model.

By performing a variable selection analysis, such suspicions about variables SSLev and Age were confirmed, and both variables were excluded from the final regression model, according to the criteria of excluding variables with a p-value for the Wald test, associated with a univariable regression model, greater than 0.25.

When estimating both GLM and GAM models using the Aranda-Ordaz link function, the best values obtained corresponded to the log log function. This is not the ideal result expected once it corresponds to one of the extreme values the single parameter of the function can assume. When the function parameter is equal to 1, the logistic link function is assumed and when the parameter is equal to 0 the complementary log log link function is assumed. Both of these functions are commonly used, and so the greatest advantage of the flexibility of the function was to be able to fit the model to any parameter placed in between both parameters corresponding to already used functions. However, two major aspects should be taken into consideration.

Firstly the data used. The asymmetric Aranda-Ordaz distribution is particular beneficial when there is a disparity in the count of both levels of the outcome variable. Since the data used presented a similar value of patients who died (139) and patients who did not die (260) an advisable improvement regarding this analysis is to use data where the imbalance between patients in which the event of interest occurred and not, is greater. With such improvement in terms of data, the advantages of using Aranda-Ordaz as the link function can be greatly seen.

The Brier score, even though it was included initially in the analysis, was not used to draw any conclusions here presented, as stated previously, since its variation was too little to be possible to make any accurate decisions based on it.

Regarding the Czado function the results could be considered a little more promising, since the best model obtained, considering AUC values, correspond to different values of parameters when compared to the baseline logistic function. The parameter ψ_2 was 1, similar to the one in the logistic function, however both remaining parameters, η_0 and ψ_1, differed. Nevertheless, when AUC values where compared through the DeLong test, it showed the difference between the performance of both models was not significant. When Brier score was concerned the best model obtained corresponded to the logistic link function. Once again, the results presented unsatisfactory when compared to the expectation of improvement in terms of flexibility of the link function. However, as previously brought up, the flexibly of the link function allowed for a more thorough research of the best link function and the choice of either the logistic link function, when considering the Brier score, or the function with parameters $\psi_1 = 2$, $\psi_2 = 1$, $\eta_0 = 0.5$.

Independently of the link function used, the results obtained were equal either using GLMs or GAMs. This indicates that in this specific case, for the data used the prior assumption of a linear relation between the outcome variable and the independent variables seems to be correct. Nevertheless, the analysis taking into consideration both types of models validates this prior assumption of linearity, rather than just performing the analyses assuming it with no statistical evidence to so. May further studies be performed the same analysis using a different set of data, both models should be taken into consideration as the results may differ significantly.

As results were not as resounding as initially expected, the work developed allows for a easily reproducible and systematic analysis regarding the best fitting link function to each problem. It can be applied in multiple contexts, and specially in the medical field, as already demonstrated throughout the work, as statistical results have such an impact in treatment outcome and hospital management. For future developments it is suggested to use different sets of data, especially data in which the relation between the outcome variable and the independent variables is not linear given it will be more suitable to highlight the potential a flexible link function has. Another suggestion for future developments on the theme is to perform a simulation analysis where, with a controlled set of data, it is possible to better analyse the behaviour, according to distinct situations and types of data, of the both considered link functions Aranda-Ordaz and Czado.

Appendix A

Aranda-Oradz and Czado Link Functions

This appendix contains both `R` functions used for the Aranda-Ordaz link function and the Czado link function.

A.1 Aranda-Ordaz Link Funcion

The code here available corresponds to the code which allows for the GLM Aranda-Ordaz link function to be implemented. The function can be easily replaced bya a GAM link function by replacing the class `"family"` for `"link-gamlss"`.

Listing A.1: Aranda-Ordaz GLM link function

```
aranda_glm <- function(lambda = 1) {

  if(lambda == 0) {
    binomial(link = cloglog)
  } else {
    care.exp <- function(x, thresh = about36) {
      about36 <- - log(.Machine$double.eps)
      thresh <- min(thresh, about36)
      x[x > thresh] <- thresh
      x[x < ( - thresh)] <- - thresh
      exp(x)
    }
    linkfun <- function(mu) { log(((1-mu)^(-lambda)-1)/lambda) }
    linkinv <- function(eta) {1-(lambda*care.exp(eta)+1)^(-1/lambda) }
    mu.eta <- function(eta) { care.exp(eta)*(lambda*
    care.exp(eta)+1)^(-1/lambda-1)}
    valideta <- function(eta) TRUE
    variance <- function(mu) mu * (1 - mu)
    validmu <- function(mu) all(mu > 0) && all(mu < 1)
    dev.resids <- function(y, mu, wt) {
      devy<-y
  nz<-y!=0
  devy[nz]<-y[nz]*log(y[nz])
  nz<-(1-y)!=0
  devy[nz]<-devy[nz]+(1-y[nz])*log(1-y[nz])
  devmu<-y*log(mu)+(1-y)*log(1-mu)
  if(any(small <- mu*(1-mu) < .Machine$double.eps)) {
    warning("fitted values close to 0 or 1")
    smu<-mu[small]
    sy<-y[small]
    smu<-ifelse(smu < .Machine$double.eps, .Machine$double.eps,smu)
    onemsmu<-ifelse((1-smu) < .Machine$double.eps, .Machine$double.
       eps,1-smu)
    devmu[small]<-sy*log(smu)+(1-sy)*log(onemsmu)
  }
  devi<-2*(devy-devmu)
  wt*devi
}
aic <- function(y, n, mu, wt, dev) {
  m <- if (any(n > 1))
    n
  else wt
  -2 * sum(ifelse(m > 0, (wt/m), 0) * dbinom(round(m * y), round(m),
      mu, log = TRUE))
}
initialize = expression( {
  if (NCOL(y) == 1) {
    if (is.factor(y)) y <- y != levels(y)[1]
    n <- rep(1, nobs)
    if (any(y < 0 | y > 1)) stop("y values must be 0 <= y <= 1")
    mustart <- (weights * y + 0.5)/(weights + 1)
    m <- weights * y
    if (any(abs(m - round(m)) > 0.001)) warning("non-integer #
       successes in a binomial glm!")
  } else if (NCOL(y) == 2) {
    if (any(abs(y - round(y)) > 0.001)) warning("non-integer counts
       in a binomial glm!")
    n <- y[, 1]+ y[, 2]
    y <- ifelse(n == 0, 0, y[, 1]/n)
    weights <- weights * n
    mustart <- (n * y + 0.5)/(n + 1)
  } else stop(paste("For the binomial family, y must be",
                    "a vector of 0 and 1's or a 2 column", "matrix
                       where col 1 is no. successes",
                    "and col 2 is no. failures"))
} )
structure(list(family = "Aranda", link = lambda, linkfun = linkfun,
   linkinv = linkinv,
               variance = variance, dev.resids = dev.resids, aic =
                  aic, mu.eta = mu.eta,
               initialize = initialize, validmu = validmu, valideta
                  = valideta), class = "family")
  }

}
```

A.2 Czado Link Function

The code here available corresponds to the code necessary for the GLM Czado link function. In order for the code to work for a GAM the only difference lies in replacing the class "family" for "link-gamlss".

Listing A.2: Czado GLM link function

```
czado_glm <- function(psi1 = 1, psi2 = 1, eta0 = 0) {

    care.exp <- function(x, thresh = about36) {
      about36 <- -log(.Machine$double.eps)
      thresh <- min(thresh, about36)
      x[x > thresh] <- thresh
      x[x < (-thresh)] <- -thresh
      exp(x)
    }

    linkfun <- function(mu) {
      f <- log(mu/(1-mu))

      for (i in 1:length(f)){
        h<-f
        if(((alpha1*(mu-eta0)+1)^(1/psi1)-1+eta0)[i] >= eta0){

          h[i] <- (psi1*(f[i]-eta0) + 1)^(1/psi1) + eta0 - 1
        } else {
          h[i] <- eta0 + 1 - (psi2*(eta0-f[i]) + 1)^(1/psi2)
        }
      }

      return(h)

    }

    linkinv <- function(eta) {

      f<-eta
      for (i in 1:length(eta)){

        if(eta[i] >= eta0){
          f[i] <- eta0 + ((eta[i] - eta0 + 1)^psi1 - 1)/psi1
        } else {
          f[i] <- eta0 - ((-eta[i] + eta0 + 1)^psi2 - 1)/psi2
    }}
  out = care.exp(f)/(1+care.exp(f))

  return(out)

}

mu.eta <- function(eta){

  f<-eta
  g<-eta
  logistica.f <- eta
  a <- eta

  for (i in 1:length(eta)){
    if(eta[i] >= eta0){
      g[i] <- (eta[i] - eta0 + 1)^(psi1-1)
      f[i] <- eta0 + ((eta[i] - eta0 + 1)^psi1 - 1)/psi1
      logistica.f[i] <- 1/(1+care.exp(-f[i]))
      a[i] <- (logistica.f[i]*(1-logistica.f[i]))*g[i]

    }else{

      g[i] <- (-eta[i] + eta0 + 1)^(alpha2-1)
      f[i] <- eta0 - ((-eta[i] + eta0 + 1)^alpha2 - 1)/alpha2
      logistica.f[i] <- 1/(1+care.exp(-f[i]))
      a[i] <- (logistica.f[i]*(1-logistica.f[i]))*g[i]

    }
  }
  return(a)
}

valideta <- function(eta) TRUE
variance <- function(mu) mu * (1 - mu)
validmu <- function(mu) all(mu > 0) && all(mu < 1)
dev.resids <- function(y, mu, wt) {
  devy<-y
  nz<-y!=0
  devy[nz]<-y[nz]*log(y[nz])
  nz<-(1-y)!=0
  devy[nz]<-devy[nz]+(1-y[nz])*log(1-y[nz])
  devmu<-y*log(mu)+(1-y)*log(1-mu)
  if(any(small <- mu*(1-mu) < .Machine$double.eps)) {
    warning("fitted values close to 0 or 1")
    smu<-mu[small]
    sy<-y[small]
    smu<-ifelse(smu < .Machine$double.eps, .Machine$double.eps,smu)
    onemsmu<-ifelse((1-smu) < .Machine$double.eps, .Machine$double.
       eps,1-smu)
    devmu[small]<-sy*log(smu)+(1-sy)*log(onemsmu)
    }
    devi<-2*(devy-devmu)
    wt*devi
  }
  aic <- function(y, n, mu, wt, dev) {
    m <- if (any(n > 1))
      n
    else wt
    -2 * sum(ifelse(m > 0, (wt/m), 0) * dbinom(round(m * y), round(m),
        mu, log = TRUE))
  }

  initialize = expression( {
    if (NCOL(y) == 1) {
      if (is.factor(y)) y <- y != levels(y)[1]
      n <- rep(1, nobs)
      if (any(y < 0 | y > 1)) stop("y values must be 0 <= y <= 1")
      mustart <- (weights * y + 0.5)/(weights + 1)
      m <- weights * y
      if (any(abs(m - round(m)) > 0.001)) warning("non-integer #
         successes in a binomial glm!")
    } else if (NCOL(y) == 2) {
      if (any(abs(y - round(y)) > 0.001)) warning("non-integer counts
         in a binomial glm!")
      n <- y[, 1] + y[, 2]
      y <- ifelse(n == 0, 0, y[, 1]/n)
      weights <- weights * n
      mustart <- (n * y + 0.5)/(n + 1)
    } else stop(paste("For the binomial family, y must be",
                      "a vector of 0 and 1's or a 2 column", "matrix
                         where col 1 is no. successes",
                      "and col 2 is no. failures"))
  } )

  structure(list(family = "Czado", link = c(psi1, psi2), linkfun =
     linkfun, linkinv = linkinv,
                 variance = variance, dev.resids = dev.resids, aic =
                    aic, mu.eta = mu.eta,
                 initialize = initialize, validmu = validmu, valideta
                    = valideta), class = "family")
}
```

Bibliography

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *Springer Serires in Statistics Selected Papers of Hirotugu Akaike,* 215-222.

[Amaral Turkman & Silva, 2000] Amaral Turkman, M. A., Silva, G. (2000). *Modelos Lineares Generalizados - da Teoria à Prática.* Edições SPE, Lisboa.

[Aranda-Ordaz, 1981] Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika,* 68(2), 357-363.

[Benesty et al., 2009] Benesty J., Chen J., Huang Y., Cohen I. (2009) Pearson Correlation Coefficient. In: *Noise Reduction in Speech Processing. Springer Topics in Signal Processing,* vol 2. Springer, Berlin, Heidelberg.

[Bradley, 1997] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition,* 30(7), 1145-1159.

[Brier, 1950] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review,* 78, 1-3.

[Cabilio & Masaro, 1996] Cabilio, P., & Masaro, J. (1996). A simple test of symmetry about an unknown median. *Canadian Journal of Statistics,* 24(3), 349-361.

[Cox & Reid, 1987] Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate condition. *Journal of the Royal Statistical Society: Series B (Methodological),* 49(1), 1-18.

[Cramer, 2003] Cramer, J. S. (2003). The origins and development of the logit model. *Logit Models from Economics and Other Fields,* 149-157.

[Czado, 1992] Czado, C. (1992). On link selection in generalized linear models. *Advances in GLIM and Statistical Modelling Lectures Notes Statistics,* 60-65.

[Czado et al., 2010] Czado, C., Pfettner, J., Gschlößl, S., Schiller, F. (2010) Nonnested model comparison of GLM and GAM count regression models for life insurance data.

[DeLong et al., 1988] DeLong, E. R., DeLong D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics,* 44(3), 837.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 27(8), 861-874

[Geraldes, 2016] Geraldes, C. J. B. (2016). *Aplicação das Redes Neuronais Aditivas Generalizadas à Medicina*, Doctoral dissertation.

[Hastie & Tibshirani, 1986] Hastie, T. & Tibshirani, R. (1986). Generalized additive models. *Statistical Science,* 1(3), 297-318

[Hosmer & Lemeshow, 2000] Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression.* New York: Wiley

[Jiang et al., 2013] Jiang, X., Dey, DK., Prunier, R., Wilson, AM., & Holsinger, KE. (2013). A new class of flexible link functions with application to species co-occurrence in Cape Floristic region. *The Annals of Applied Statistics,* 7, 2180-2204.

[Kerr et al., 2007] Kerr, K., Norris, T., & Stockdale, R. (2007). Data quality information and decision making: a healthcare case study. 18^th *Australasian Conference on Information Systems,* 5-7

[Li et al.. 2015] Li, D., Wang, X., S., Zhang, N., & Dey, D. K. (2015). Flexible link functions in a joint model of binary and longitudinal data. *Stat,* 4(1), 320-330.

[Lindsey & Jones, 1988] Lindsey, J. K., & Jones, B. (1998). Choosing among generalized linear models applied to medical data. *Statistics in Medicine,* 17(1), 59-68.

[Montgomery et al., 2012] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

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

[Powers, 2008] Powers, D. (2008). Evaluation: from precision, recall and f-factor to ROC, informedness, markedness & correlation. *Mach. Learn. Technol.,* 2.

[Rencher & Schaalje, 2008] Rencher, A. C. & Schaalje, G. B. (2008). *Linear Models in Statistics.* Hoboken: Wiley & Sons. Retrived September 3, 2020, from http://www.utstat.toronto.edu/ brunner/books/LinearModelsInStatistics.pdf

[Rufibach, 2010] Rufibach, K. (2010). Use of Brier score to assess binary predictions. *Journal of Clinical Epidemiology,* 63(8), pp.938-939.

[Vogenberg, 2009] Vogenberg, F. R. (2009). Predictive and prognostic models: implications for healthcare decision-making in a modern recession. *AHDB,* 2(6): 218-222.

[Yu et al., 2013] Yu, H., Jiao, Y., & Carstensen, L. W. (2013). Performance comparison between spatial interpolation GLM/GAM in estimating relative abundance indices through a simulation study. *Fisheries Research,* 147, 186-195.