UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Lucas Araújo da Silva

Influential diagnostics for location parameter within GAMLSS

Recife

2021

Lucas Araújo da Silva

Influential diagnostics for location parameter within GAMLSS

Dissertação apresentada ao Programa de pós-graduação em Estatística do Departamento de Estatística da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador (a): Fernanda de Bastiani

Recife

2021

Catalogação na fonte
Bibliotecária Fernanda Bernardo Ferreira, CRB4-2165

S586i Silva, Lucas Araújo da
Influential diagnostics for location parameter within GAMLSS / Lucas Araújo da Silva. – 2021.
72 f.: il., fig., tab.

Orientadora: Fernanda de Bastiani.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2021.
Inclui referências.

1. Estatística Aplicada. 2. Bootstrap. 3. Distância de Cook. I. Bastiani, Fernanda de (orientadora). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2021 - 79

LUCAS ARAÚJO DA SILVA

INFLUENTIAL DIAGNOSTICS FOR LOCATION PARAMETER WITHIN GAMLSS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 18 de fevereiro de 2021.

BANCA EXAMINADORA

Prof.ª Fernanda De Bastiani
DE/UFPE

Prof.ª Aline Tsuyuguchi
DE/UFPE

Prof.º Jorge Luis Bazán Guzmán
ICMC-USP

Dedicated to all Brazilian researchers.

ACKNOWLEDGEMENTS

Obtain a master title it's not a easy task, so i would like to thank some people. I am sincerely grateful to my supervisor Professor Fernanda De Bastiani for her guidance, attention and patience during this research. Also, to the professors of department of statistics at UFPE especially to Gauss Cordeiro, Francisco Cribari, Maria do Carmo and Audrey Cysneiros for theirs professionalism.

To my colleagues Vinícius, Yuri, Jordan and Luis, for his patience and for having provided great moments sharing housing with me during my first year in Recife. To my classmates, especially Tatiane Fontana and Alexsandra Gomes for the hours we studied together.

The professors who accompanied me during graduation were not forgotten, they were important for me to get where i got. My special thanks to Professor Juvêncio Nobre, Professor Luis Gustavo, Professor Julio Barros, Professor Rafael Farias and Professor Ricardo Coelho.

To my parents, and my sister Livia for their support.

And finally, to the chaos in the universe.

"I've seen things you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I watched C-beams glitter in the dark near the Tannhäuser Gate. All those moments will be lost in time, like tears in rain." (SCOTT et al., 1982).

ABSTRACT

Modelling the functional relationship between a variable response and a set of explanatory variables is at the core of the regression problems in statistics. Several studies have proposed different models. More recently, generalized additive models for scale and shape location (GAMLSS) have gained attention for generalizing other already popular models such as the linear model, the generalized linear models, semiparametric models and the generalized additive models, and allowing any parametric distribution to model the response variable. In addition, all distribution parameters can be modeled with linear, non-linear or smoothing functions for explanatory variables. Various tools of influence diagnostics have been proposed in the literature, and this work shows some of these tools and proposes techniques to detect possible influential observations in the GAMLSS model class. This work considers several measures of influence such as: the generalized Cook distance, the likelihood distance, the adjusted Peña measure, differences in the generalized Akaike information criterion and the Kim measure for simulated data and applications. It is also proposed algorithms to obtain the reference values of these measures using bootstrap, adapting for the other measures the procedure suggested by (KIM; PARK; KIM, 2002). The study is still limited to situations where we model the location parameter (in general the mean) of the response variable, whether or not we have smoothing additives, in this case univariate penalized splines were used as a smoother, since the Peña and Kim measures need to calculate the matrix of smoothing that varies according to the smoothed covariate and the smoother in question. For the simulation studies, several scenarios were considered with some relevant distributions and several sample sizes, taking into account continuous and discrete distributions as well. Analysis of real data illustrates the approached methodology.

Keywords: Bootstrap. Cook's Distance. Peña's Measure. P-splines. Semiparametric Model.

RESUMO

Modelar a relação funcional entre uma váriável resposta e um conjunto de variáveis explicativas é o cerne dos problemas de regressão em estatística. Diversos estudos tem propostos diferentes modelos. Mais recentemente os modelos aditivos generalizados para locação escala e forma (GAMLSS) tem ganhado atenção por generalizar outros modelos já populares como o modelo linear, os modelos lineares generalizados, modelos semiparamétricos e os modelos aditivos generalizados, e permitir qualquer distribuição paramétrica para modelar a variável resposta. Além disso, todos os parâmetros da distribuição podem ser modelados com funções lineares, não lineares ou funções de suavização das variáveis explicativas. Várias ferramentas de diagnósticos de influência tem sido propostas na literatura, e este trabalho mostra algumas dessas ferramentas e propõe técnicas para detectar possíveis observações influentes na classe de modelos GAMLSS. Este trabalho considera diversas medidas de influência como: a distância de Cook generalizada, o afastamento de verossimilhanças, a medida de Peña ajustada, diferenças do critério de informação de Akaike generalizada e a medida de Kim para dados simulados e aplicações. É proposto ainda algoritmos para obter os valores de referência destas medidas utilizando bootstrap, adaptando para as outras medidas o procedimento sugerido por Kim et al. (2002). O estudo ainda limita-se a situações que se é modelado o parâmetro de locação (em geral a média) da variável resposta, incluindo ou não termos aditivos de suavização, neste caso utilizou-se splines penalizados univariados como suavizador, já que a medida de Peña e de Kim necessitam do cálculo da matriz de suavização que varia de acordo com a covariável suavizada e o suavizador em questão. Para os estudos de simulação, foram considerados diversos cenários com algumas distribuições relevantes e diversos tamanhos amostrais, considerando distribuições tanto de natureza contínua quanto discretas. Análise de dados reais ilustram a metodologia abordada.

Palavras-chaves: Bootstrap. Distância de Cook. Medida de Peña. Modelo Semiparamétrico. P-Splines.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 PRELIMINARIES

In many practical situations it is desirable to model the functional relation of one or more populational aspects between a particular response variable and one or more explanatory variables. For this goal, a popular approach is to fit a regression model.

In general, we desire a model that represents the reality of the studied phenomenon, but in some cases the model can be excessively complex, and the principle of parsimony suggest than a good model need to capture the essential of the data behavior with the best possible simplicity.

Regression analysis is attributed to Francis Galton in the 1870s for the regression to the mean as presented by (SENN, 2011). Moreover, currently several fields of science have used regression techniques to predict and understanding a phenomena. In this chapter, a review about some important classes of regression models is presented.

1.1 INTRODUCTION AND STRUCTURE

There are several classes of regression models in the literature. The first chapter of this work present some important GAMLSS submodels and theirs main respectives features and parameter estimation as well.

The chapter two explores some diagnostic tools for the GAMLSS models, we focus on the leave-one-out measures and present the algorithms to compute the reference values. The main goal of this work is provide some global influence measures to GAMLSS models, also a criteria to identify when a observation is influential.

The chapter three provide simulations for the main GAMLSS submodels by artificially including influential cases, also different distributions are simulated to the response variable and fitted models.

The chapter four perform two applications with real data, and the measures are computed, finally the chapter five present the concluding remarks and the chapter six suggest some future woks to complement this one.

1.2 THE LINEAR REGRESSION MODEL (LM)

One of the simplest and more popular regression model, is the linear regression model. This section introduces this model, and shows some advantages, assumptions, and disadvantages. (MONTGOMERY; PECK; VINING, 2012) provide several details about this class of models.

Initially, consider a data set with n observations, from random variable Y_i. Let $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ be a vector of observed response variable and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r)^\top$ be a matrix with fixed values and r co-variate columns, with expectation, $\mu_i \equiv \mathbb{E}(Y_i|\boldsymbol{X})$. Denote $\boldsymbol{Y}$ as the response variable and $\boldsymbol{X}$ as the design matrix. A possible suitable model for the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the regression linear model, having the following functional form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_r x_{ir} + \epsilon_i, \tag{1.1}$$

where, ϵ_i for $i = 1, \ldots, n$ are independently normal distributed that is, $\epsilon_i \overset{ind}{\sim} N(\mu, \sigma^2)$, the normal distribution is given by the following probability density function:

$$f(x_i|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}.$$

We can rewrite the model (1.1) in the follow equivalent specification:

$$Y_i \sim N(\mu_i, \sigma^2), \tag{1.2}$$

where $\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_r x_{ir}$, for $i = 1, \ldots, n$.

Using matrix notation, the model (1.2) specification can be write as

$$\boldsymbol{Y} \overset{ind}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2),$$

where, $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{y} = (Y_1, \ldots, Y_n)^\top$ is the response vector, $\boldsymbol{X}$ is the design matrix with dimensions $n \times p\,(p = r + 1)$, if the constant is required the first column is ones, and plus r covariate columns, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_r)^\top$ is the coefficient vector, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ is the mean vector, and $\boldsymbol{\sigma}^2 = (\sigma^2, \ldots, \sigma^2)^\top$ is the vector of constant variance.

A common way, to estimate $\boldsymbol{\beta}$ is using the least squares estimator, the idea is minimizing the sum of squared differences between the observations Y_i and the means μ_i, with respect to the betas coefficients. We can write this as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

the solution is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \tag{1.3}$$

it can also shown that least squares estimator in (1.3) is equivalent to the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$. The fitted values of the linear model are $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and the residuals (fitted errors) are $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$. An unbiased estimator for σ^2 is

$$\hat{s}^2 = \frac{\hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}}}{n-p},$$

since we have a established mean and variance of $\hat{\beta}$, we have that

$$\hat{\beta} \sim N\left(\beta, \frac{\sum_{i=1}^{n} x_i^2}{\sigma^2}\right).$$

Note, the model premise that $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$, often this can be not true. Moreover, situations like not constant variance, correlated errors, nonlinear trend of the data or the response variable are taken from others distributions. Also, count data, or even rates and proportions can be not well fitted by the linear model. Some transformations in the response variable were proposed to stabilize the variance of errors as the Box-Cox transformation (BOX; COX, 1964) and the Yeo-Johnson transformation (YEO; JOHNSON, 2000), but this approaches compromises the interpretation of the fitted parameters. Knowing this, several models were developed to get around this situations.

1.3 THE GENERALIZED LINEAR MODEL (GLM)

The Generalized Linear Model (GLM) was first introduced by (NELDER; WEDDERBURN, 1972), we can highlight three innovations in their approach: (i) the exponential family (denoted as EF($\cdot$)) replaces the normal distribution for modeling the response variable, thus are useful to practical modeling, allowing count or binary data for example; (ii) a monotonic link function $g(\cdot)$ is used in modeling the relationship between $\mathbb{E}(Y)$ and the explanatory variables (iii) in order to find the MLE for the parameters β it uses an interactively re-weighted least squares algorithm.

The GLM can be written as:

$$Y_i \overset{ind}{\sim} \text{EF}(\mu_i, \phi), \text{ where } g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_r x_{ir}, \text{ for } i = 1, 2, \ldots, n,$$

and ϕ is the dispersion parameter. Using matrix notation, we rewrite this as

$$\boldsymbol{Y} \overset{ind}{\sim} \text{EF}(\boldsymbol{\mu}, \boldsymbol{\phi}),$$

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta},$$

where η is called the linear predictor and $\boldsymbol{\phi} = (\phi, \ldots, \phi)^\top$ is a vector of constant ϕ.

The exponential family distribution $\text{EF}(\mu, \phi)$ is defined by the probability (density) function $f(y|\mu, \phi)$ having the form:

$$f(y|\mu, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right\}, \tag{1.4}$$

where $\mathbb{E} = \mu = b'(\theta)$ and $\text{var}(Y) = \phi V(\mu)$, where $V(\mu) = b''[\theta(\mu)]$ and $V(\mu)$ is called the variance function. The exponential family in 1.4 includes many important distributions as the normal distribution, binomial, gamma, Poisson, inverse Gaussian, negative binomial, and others.

(WOOD, 2017) showed that GLMs can be estimated by the *iteratively re-weighted least square (IRLS) algorithm*, the IRLS are described as follows:

Algoritmo 1: Iteratively Re-Weighted Least Square Algorithm.

(1) Initialize $\hat{\mu} = y_i + \delta_i$ and $\hat{\eta}_i = g(\hat{\mu}_i)$, where δ_i is usually zero, but may be a small constant ensuring that $\hat{\eta}_i$ is finite. Iterate the following two steps to convergence;

(2) Compute pseudodata $z_i = \dfrac{g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)}{\alpha(\hat{\mu}_i)} + \hat{\eta}_i$, and iterative weights $w_i = \dfrac{\alpha(\hat{\mu}_i)}{g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)}$;

(3) Find $\hat{\beta}$ witch minimise of the weighted least squares objective function

$$\sum_{i=1}^{n} w_i (z_i - \boldsymbol{X}_i\boldsymbol{\beta})^2,$$

then update $\hat{\eta} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

To more details about the GLMs see also (DOBSON; BARNETT, 2018).

1.4 THE GENERALIZED ADDITIVE MODEL (GAM)

The generalized additive model (GAM) introduced by (HASTIE; TIBSHIRANI, 1990) is a generalized linear model with a sum of smooth functions of covariates in the linear predictor. (WOOD, 2017) has contributed extensively to GAM theory and popularity by allowing, in his implementation of GAM in R (package **mgcv**).

The GAM can be written as:

$$\boldsymbol{Y} \overset{ind}{\sim} \text{EF}(\boldsymbol{\mu}, \boldsymbol{\phi})$$

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + s_1(\boldsymbol{x}_1) + \ldots + s_J(\boldsymbol{x}_J), \tag{1.5}$$

where s_j is a nonparametric smoothing function applied to covariate $\boldsymbol{x}_j$, for $j = 1, \ldots, J$, $\text{EF}(\boldsymbol{\mu}, \boldsymbol{\phi})$ denotes an exponential family distribution with vector of mean $\boldsymbol{\mu}$ and vector of scale $\boldsymbol{\phi}$. Y_i is the response variable, s_j is not limited to a univariate case, that is the smoothing terms $s(\cdot)$ can smooth two or more covariables in a single term, but in this work we focus on models with just one univariate smoother term with penalized splines.

The main idea of this formulation is allow a better flexibility to the model fit, that means leaving the data determine the relationship between the linear predictor $\eta = g(\mu)$ and the explanatory variables. the sections 1.7 and 2.2.4 provide more information about the smoothing parameter estimation. To read more details about the additive smoothing terms see the chapter 9 of (STASINOPOULOS et al., 2017).

1.5 MEAN AND DISPERSION ADDITIVE MODEL (MADAM)

In some occasions, in which the assumption of a constant scale parameter is not appropriate. To deal with this situations and modelling σ, in the 1970s new approaches were proposed. (HARVEY, 1976) and (AITKIN, 1987) was the first to modeling the variance of the normal distribution. As a solution to the problem of heterocedasticity.

(NELDER; WEDDERBURN, 1972), (SMYTH, 1989) and (VERBYLA, 1993) are introduced approaches for modeling the dispersion parameter within the GLM framework. A model to modeling both μ and σ was introduced by (RIGBY; STASINOPOULOS, 1996), (RIGBY; STASINOPOULOS, 1996), which they called the mean and dispersion additive model (MADAM). The MADAM using the pseudo-likelihood method to estimate the parameters, the same used in GAMLSS which allows any two-parameters distribution for the response variable, but in the original formulation the response distribution had to be in the exponential family.

The MADAM formulation has the form:

$$Y \overset{ind}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma})$$

$$\begin{aligned}\boldsymbol{\eta}_1 &= g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + s_{11}(x_{11}) + \ldots + s_{1J_1}(x_{1J_1}) \\ \boldsymbol{\eta}_2 &= g_2(\boldsymbol{\mu}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 + s_{21}(x_{21}) + \ldots + s_{2J_2}(x_{2J_2})\end{aligned} \tag{1.6}$$

where $D(\boldsymbol{\mu}, \boldsymbol{\sigma})$ is any two-parameters distribution and both μ and σ are linear and/or smooth functions of the explanatory variables.

1.6 THE GENERALIZED ADDITIVE MODEL FOR LOCATION, SCALE AND SHAPE (GAMLSS)

Use a two-parameter distribution bounded us to the fact that the skewness and kurtosis of the distribution are fixed for fixed μ and σ. In some situations, there may be interest to model the skewness and/or kurtosis. The GAMLSS has been used in several fields like: actuarial science, biology, bio-sciences, energy economic, genomics, finance, fisheries, food consumption, growth curves estimation, marine research, medicine, meteorology, rainfalls, vaccines, and others. Huge institutes are using GAMLSS in their analysis, like the World Health Organisation (WHO) (PAIVA; FREIRE; CECATTI, 2008), the International Monerary Fund (IMF) (MONETARY; DEPARTMENT, 2015), and the European Bank (GIRAUD; KOCKEROLS, 2015).

Therefore, the model 1.6 can be extended as follows:

$$\boldsymbol{Y} \overset{ind}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$$

$$\begin{aligned}\boldsymbol{\eta}_1 &= g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + s_{11}(\boldsymbol{x}_{11}) + \ldots + s_{1J_1}(\boldsymbol{x}_{1J_1}) \\ \boldsymbol{\eta}_2 &= g_1(\boldsymbol{\sigma}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 + s_{21}(\boldsymbol{x}_{21}) + \ldots + s_{2J_2}(\boldsymbol{x}_{2J_2}) \\ \boldsymbol{\eta}_3 &= g_1(\boldsymbol{\nu}) = \boldsymbol{X}_3\boldsymbol{\beta}_3 + s_{31}(\boldsymbol{x}_{31}) + \ldots + s_{3J_3}(\boldsymbol{x}_{3J_3}) \\ \boldsymbol{\eta}_4 &= g_4(\boldsymbol{\tau}) = \boldsymbol{X}_4\boldsymbol{\beta}_4 + s_{41}(\boldsymbol{x}_{41}) + \ldots + s_{4J_4}(\boldsymbol{x}_{4J_4})\end{aligned} \tag{1.7}$$

where $D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ is any four-parameter distribution and where μ are a location parameter, σ are a scale parameter and ν and τ are shape parameters which are often related to the skewness and kurtosis of the distribution. The model 1.7 defines de formulation of the generalized additive model for location, scale and shape (GAMLSS) and was first introduced by (RIGBY; STASINOPOULOS, 2005). GAMLSS dispose of a computational package in the R language, which one enable to fit a model, implement new distributions (in addition to the more than one hundred already existing), more than ten kinds of additive terms, and several others functionalities pre implemented.

The GAM model provides a more flexible approach, in terms of the specification of the dependence of the response on the covariates. This models can be represented using basis expansions for each smooth term. In this particular work, we focus in models with one single univariate smooth component.

The GAMLSS lets choose between a wide range of options of regression models, (STASINOPOULOS et al., 2017) underscore some basic properties of the GAMLSS, which follows:

- GAMLSS is a very flexible unifying framework for univariate regression models.
- It allows any distribution for the response variable. All the parameters of the distribution can be modelled as functions of explanatory variables.
- It allows a variety of additive terms in the models for the distribution parameters.
- The fitted algorithm is modular, where different components can be added easily.
- It extends basic statistical models allowing flexible modeling of overdispersion, excess of zeros, skewness and kurtosis in the data.

Generally, the smooth functions used in the GAMLSS models can be written as $s(\boldsymbol{x}) = \boldsymbol{Z}\boldsymbol{\gamma}$ where $\boldsymbol{Z}$ is the basis matrix which depends on the explanatory variable $\boldsymbol{x}$. $\boldsymbol{\gamma}$ is a parameter vector to be estimated, subject to a quadratic penalty of the form $\lambda\boldsymbol{\gamma}^\top\boldsymbol{G}\boldsymbol{\gamma}$, for a known matrix $\boldsymbol{G} = \boldsymbol{D}^\top\boldsymbol{D}$ and where the hyperparameter λ regulates the amount of smoothing needed for the fit.

The model (1.7) can be extended including random effects in the following form:

$$\boldsymbol{Y}|\boldsymbol{\gamma} \overset{ind}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$$

$$\begin{aligned}
\boldsymbol{\eta}_1 &= g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + Z_{11}(\boldsymbol{\gamma}_{11}) + \ldots + Z_{1J_1}(\boldsymbol{\gamma}_{1J_1}) \\
\boldsymbol{\eta}_2 &= g_1(\boldsymbol{\sigma}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 + Z_{21}(\boldsymbol{\gamma}_{21}) + \ldots + Z_{2J_2}(\boldsymbol{\gamma}_{2J_2}) \\
\boldsymbol{\eta}_3 &= g_1(\boldsymbol{\nu}) = \boldsymbol{X}_3\boldsymbol{\beta}_3 + Z_{31}(\boldsymbol{\gamma}_{31}) + \ldots + Z_{3J_3}(\boldsymbol{\gamma}_{3J_3}) \\
\boldsymbol{\eta}_4 &= g_4(\boldsymbol{\tau}) = \boldsymbol{X}_4\boldsymbol{\beta}_4 + Z_{41}(\boldsymbol{\gamma}_{41}) + \ldots + Z_{4J_4}(\boldsymbol{\gamma}_{4J_4})
\end{aligned} \tag{1.8}$$

where the β's are the fixed effects parameters:

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top)^\top$$

and the γ's are the random effects parameters:

$$\boldsymbol{\gamma} = (\gamma_{11}^\top, \ldots, \gamma_{1J_1}^\top, \gamma_{21}^\top, \ldots, \gamma_{4J_4}^\top)^\top$$

Also, assume for the model (1.8) that the γ's are independent of each other, each with prior distribution

$$\gamma_{kj} \overset{ind}{\sim} N(\mathbf{0}, [\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})]^{-1}) \tag{1.9}$$

where $[\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})]^{-1}]$ is the (generalized) inverse of a $q_{kj} \times q_{kj}$ symmetic matrix $\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})$ which may depend on a vector of hyperparameters $\boldsymbol{\lambda}_{kj}$. Moreover, we can simplify the model (1.8), if there are no random effects, and write as:

$$\boldsymbol{Y} \overset{ind}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$$

$$\begin{aligned}
\boldsymbol{\eta}_1 &= g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 \\
\boldsymbol{\eta}_2 &= g_1(\boldsymbol{\sigma}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 \\
\boldsymbol{\eta}_3 &= g_1(\boldsymbol{\nu}) = \boldsymbol{X}_3\boldsymbol{\beta}_3 \\
\boldsymbol{\eta}_4 &= g_4(\boldsymbol{\tau}) = \boldsymbol{X}_4\boldsymbol{\beta}_4
\end{aligned} \tag{1.10}$$

The model (1.10) is called as the parametric GAMLSS model, and the model (1.8) as the random effects GAMLSS model.

1.6.1 Parameter Estimation

The parametric GAMLSS model only requires estimates the β's. The random effects GAMLSS model requires estimates for the β's, $\boldsymbol{\gamma}$'s and also the $\boldsymbol{\lambda}$. The **gamlss** package in R fitted the parametric model by maximum likelihood estimation with respect to $\boldsymbol{\beta}$. In this section we summarise the model parameter estimation for the GAMLSS models, for more details the reader can see (STASINOPOULOS; RIGBY et al., 2007). Also as introduced by (RIGBY; STASINOPOULOS, 2005), the random effects model is fitted using maximum penalized likelihood estimation, or in an equivalent way by maximum a posterior estimation (MAP).

Let $f(y_i|\boldsymbol{\beta}, \boldsymbol{\gamma})$ be the conditional probability function of Y_i given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, which can be any distribution. Assume that the observations Y_i, for $i = 1, 2, \ldots, n$, are conditionally independent given $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Assume that he $\boldsymbol{\gamma}_{ks}'$s have prior independent normal distributions

given in (1.9), for $k = 1, 2, 3, 4$ over the distribution parameters and $j = 1, 2, \ldots, J_k$ over the different smoothers. A constant improper prior distribution for $\boldsymbol{\beta}$ is also assumed for $\boldsymbol{\lambda}$ fixed. Then the posterior distribution for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ given $\boldsymbol{y}$ and $\boldsymbol{\lambda}$ is given by:

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{\lambda}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\boldsymbol{\lambda}) \propto L(\boldsymbol{\beta}, \boldsymbol{\gamma}) \prod_k \prod_j f(\gamma_{kj}|\lambda_{kj}), \tag{1.11}$$

note that $f(\boldsymbol{\gamma}|\boldsymbol{\lambda}) = \prod_k \prod_j f(\gamma_{kj}|\lambda_{kj})$ is the prior probability density distribution for $\boldsymbol{\gamma}$. Also, $f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = L(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \prod_i f(y_i|\boldsymbol{\beta}, \boldsymbol{\gamma})$ is the likelihood function.

Furthermore, we can also work with the log-likelihood for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, in this case we have:

$$\begin{aligned}
\log f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{\lambda}) &= l(\beta, \gamma) + \sum \sum \log f(\gamma_{kj}|\lambda_{kj}|\lambda_{kj}) + c(\boldsymbol{y}, \boldsymbol{y}) && (1.12)\\
&= l_h(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\lambda}) + c(\boldsymbol{y}, \boldsymbol{\lambda}) && (1.13)\\
&= l(\boldsymbol{\beta}, \boldsymbol{\lambda}) - \frac{1}{2} \sum_k \sum_j \gamma_{kj}^\top \boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj}) \gamma_{kj} + c_1(\boldsymbol{y}, \boldsymbol{\lambda}) && (1.14)\\
&= l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\lambda}) + c_1(\boldsymbol{y}, \boldsymbol{\lambda}). && (1.15)
\end{aligned}$$

Fist, in the equation (1.12) the log-likelihood of the data for a given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, that is, $l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \log L(\boldsymbol{\beta}, \boldsymbol{\gamma})$, together with the logarithm of the assumed probability density function for $\boldsymbol{\gamma}$ given $\boldsymbol{\lambda}$. The $c(\boldsymbol{y}, \boldsymbol{\lambda}) = -\log f(\boldsymbol{y}, |\boldsymbol{\lambda})$ is a constant for the proportionality in (1.11). The hierarchical likelihood is obtained by combining the first two terms of (1.13), the definition of the hierarchical likelihood is provided by (LEE; NELDER; PAWITAN, 2018). In the equation (1.14) substitute the general form $f(\gamma_{kj}|\lambda_{kj})$ with the assumed normal distribution, so is resulting the quadratic penalty $\gamma_{kj}^\top \boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})$. The others elements of the logarithm of the normal probability density function is are absorbed by the constant, this changes c to c_1. Finally, the log-posterior for the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is equal to the penalized likelihood of the equation (1.16).

To estimate the parameters using maximum likelihood, we need to integrate $\boldsymbol{\gamma}$ out of the joint likelihood of $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ given in equation

The log-likelihood function for the parametric model (1.10) is given by

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i),$$

and the penalized log-likelihood function for the random effects GAMLSS model (1.7) is given by

$$l_p = l - \frac{1}{2} \sum_{k=1}^{4} \sum_{j=1}^{J_k} \boldsymbol{\gamma}_{kj}^\top \boldsymbol{G}_{kj}(\lambda_{kj}) \boldsymbol{\gamma}_{kj}. \tag{1.16}$$

There are two basic algorithms for fitting the parametric model with respect to $\boldsymbol{\beta}$, and the nonparametric model with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for fixed $\boldsymbol{\lambda}$: the RS algorithm (a generalization

of the same algorithm used by (RIGBY; STASINOPOULOS, 1996) and (RIGBY; STASINOPOULOS, 1996) for fitting the MADAM models). The RS algorithm using three nested procedures: the *outer iteration*, the *inner iteration* and a *modified backfitting*.

The *outer iteration* starts by maximise the log-likelihood over $\boldsymbol{\mu}$ to fit a model for $\boldsymbol{\mu}$, latest estimates $\hat{\boldsymbol{\sigma}}$, $\hat{\boldsymbol{\nu}}$ and $\hat{\boldsymbol{\tau}}$, then fit a model for $\boldsymbol{\sigma}$ given the latest estimates $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\nu}}$ and $\hat{\boldsymbol{\tau}}$, then fi a model for $\boldsymbol{\nu}$ given the latest estimates $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\nu}}$ and $\hat{\boldsymbol{\tau}}$, and finally fit a model for $\boldsymbol{\tau}$ given the latest estimates $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\nu}}$ and $\hat{\boldsymbol{\tau}}$.

The *inner iteration* is a local scoring algorithm resembling to the used in the GLMs, for practice we defined $\boldsymbol{\theta}_1 = \boldsymbol{\mu}$, $\boldsymbol{\theta}_2 = \boldsymbol{\sigma}$, $\boldsymbol{\theta}_3 = \boldsymbol{\nu}$ and $\boldsymbol{\theta}_4 = \boldsymbol{\tau}$. Let define a modified response variable (also called *working variable*) for fitting the parameter θ_k is given by

$$\boldsymbol{z}_k = \boldsymbol{\eta}_k + \boldsymbol{w}_k^{-1} \circ u_k \tag{1.17}$$

where, $\boldsymbol{z}_k = (z_{k1}, \ldots, z_{kn})^\top$, $\boldsymbol{\eta}_k = (\eta_{k1}, \ldots, \eta_{kn})^\top$ and $\boldsymbol{u}_k = (u_{k1}, \ldots, u_{kn})^\top$, $w_k^{-1} \circ u_k = (w_{k1}^{-1} u_{k1}, \ldots, w_{kn}^{-1} u_{kn})$, the operator $\circ$ is the Hadamard element by element product, $\boldsymbol{\eta}_1 = g_k(\boldsymbol{\theta}_k)$ is the predictor vector of the k-th parameter vectors for $k = 1, 2, 3, 4$. The first derivative of the log-likelihood (score function) with respect to the predictor, are given by

$$\boldsymbol{u}_k = \frac{\partial l}{\partial \boldsymbol{\eta}_k} = \left(\frac{\partial l}{\partial \boldsymbol{\theta}_k}\right) \circ \left(\frac{d\boldsymbol{\theta}_k}{d\boldsymbol{\eta}_k}\right)$$

where $\dfrac{\partial l}{\partial \boldsymbol{\eta}_k} = \left(\dfrac{\partial l_1}{\partial \eta_{k1}}, \ldots, \dfrac{\partial l_n}{\partial \eta_{kn}}\right)^\top$, $\dfrac{\partial l}{\partial \boldsymbol{\theta}_k} = \left(\dfrac{\partial l_1}{\partial \theta_{k1}}, \ldots, \dfrac{\partial l_n}{\partial \theta_{kn}}\right)^\top$, for $k = 1, 2, 3, 4$, $\dfrac{d\boldsymbol{\theta}_k}{d\boldsymbol{\eta}_k} = \left(\dfrac{d\theta_{k1}}{d\eta_{k1}}, \ldots, \dfrac{d\theta_{kn}}{d\eta_{kn}}\right)^\top$, for $k = 1, 2, 3, 4$.

We define the *iterative weights* $\boldsymbol{w}_k$ as

$$\boldsymbol{w}_k = -\boldsymbol{f}_k \circ \left(\frac{d\boldsymbol{\theta}_k}{d\boldsymbol{\eta}_k}\right) \circ \left(\frac{d\boldsymbol{\theta}_k}{d\boldsymbol{\eta}_k}\right)$$

where the method to compute $\boldsymbol{f}_k$ depends on the information available for the specific distribution. If the expectation $\boldsymbol{f}_k = \mathbb{E}\left(\dfrac{\partial^2 l}{\partial \theta_k^2}\right)$ exists, we using a Fisher's scoring algorithm, if $\boldsymbol{f} = \dfrac{\partial l}{\partial \boldsymbol{\theta}_k^2}$ we using the usual Newton-Raphson algorithm, if $\boldsymbol{f}_k = -\left(\dfrac{\partial l}{\partial \boldsymbol{\theta}_k}\right) \circ \left(\dfrac{\partial l}{\partial \boldsymbol{\theta}_k}\right)$, where $\dfrac{\partial^2 l}{\partial \theta_k^2} = \left(\dfrac{\partial^2 l_1}{\partial \theta_{k1}^2}, \ldots, \dfrac{\partial^2 l_n}{\partial \theta_{kn}^2}\right)^\top$, we using a quasi-newton scoring algorithm. This procedure for the GLMs is the IRLS algorithm, described in the algorithm 1.

The modified backfitting is a version of the Gauss-Seidel algorithm (HASTIE; TIBSHIRANI, 1990) responsible for the estimation of the beta and gamma parameters. The backfitting algorithm need a least squares algorithm and a penalized weighted least squares algorithm.

We desire to fit linear explanatory variables and smoothers to $\boldsymbol{z}_k$ with working weights $\boldsymbol{w}_k$ using backfitting and the inner iteration for updating the estimate of distribution parameter $\boldsymbol{\theta}_k$. For given iterative weights $\boldsymbol{w}_k$, working response variable $\boldsymbol{z}_k$ and previosly inialized or estimated values for the coefficients of the two smoothers $\hat{\gamma}_{k1}$ and $\hat{\gamma}_{k2}$, calculate the partial residuals $\boldsymbol{\epsilon}$ for the beta parameters $\boldsymbol{\beta}_k$ (equivalently offsetting for $\hat{\gamma}_{k1}$ and $\hat{\gamma}_{k2}$) and fit a weighted least squares algorithm to the residuals to obtain the partial residual with respect to the second smoother and use the penalized least squares algorithm to obtain a new estimate of $\hat{\gamma}_{k1}$. Then obtain the partial residual with respect to the second smoother and use the penalized least squares to obtain a new estimate of $\hat{\gamma}_{k2}$. Then, repeat the process until the $\hat{\beta}_k$, $\hat{\gamma}_{k1}$ and $\hat{\gamma}_{k2}$ reach convergence.

On the other hand, the CG algorithm, is a generalisation of the algorithm introduced by (COLE; GREEN, 1992). The RS algorithm does not use cross derivatives of the log-likelihood and the CG algorithm requires first and second cross derivatives of the log-likelihood function with respect to the distribution parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$. The CG algorithm maximizes the penalized log-likelihood (1.16) with respect to the betas and gammas for fixed $\boldsymbol{\lambda}$.

In the outer iteration of the CG algorithm, the working variable and the iterative weights for the parameter vectors $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ are updated by:

$$\boldsymbol{z} = \boldsymbol{\eta} + \boldsymbol{w}_{kk}^{-1} \circ \boldsymbol{u}_k$$

equivalent to $\boldsymbol{z}_k$ defined in equation (1.17), for $k = 1, 2, 3, 4$ and the $\boldsymbol{w}_{ks}$ vectors contain the elements of the iterative weights, for $k = 1, 2, 3, 4$, $s = 1, 2, 3, 4$ and $s \leq k$ defined by

$$\boldsymbol{w}_{ks} = \boldsymbol{f}_{ks} \circ \left(\frac{\partial \boldsymbol{\theta}_k}{\partial \boldsymbol{\eta}_k}\right) \circ \left(\frac{\partial \boldsymbol{\theta}_k}{\partial \boldsymbol{\eta}_k}\right)$$

where $\boldsymbol{f}_{ks}$ is computed depending on the information available for the specific distribution. If the expectation $\boldsymbol{f}_{ks} = \mathbb{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_s}\right)$ exists we use a Fisher's scoring algorithm, if $\boldsymbol{f}_{ks} = \frac{\partial^2 l}{\partial \boldsymbol{\theta}}$ we use a Newton-Raphson scoring algorithm, if $\boldsymbol{f}_{ks} = -\left(\frac{\partial l}{\partial \boldsymbol{\theta}_k}\right) \circ \left(\frac{\partial l}{\partial \boldsymbol{\theta}_s}\right)$ we use a quasi Newton scoring algorithm. Where $\frac{\partial^2 l}{\partial \boldsymbol{\theta}_{\boldsymbol{k}}^{\boldsymbol{2}}} = \left(\frac{\partial^2 l_1}{\partial \boldsymbol{\theta}_{k1}^2}, \ldots, \frac{\partial^2 l_1}{\partial \boldsymbol{\theta}_{k1}^2}\right)^\top$.

Thereby, in the inner iteration takes the current working variable $\boldsymbol{z}_k$, current weights $\boldsymbol{w}_{ks}$, and current predictors denoted by $\boldsymbol{\eta}_l^\circ$ for $k = 1, 2, 3, 4$ and $s = 1, 2, 3, 4$ are fixed. Next, for $k = 1, 2, 3, 4$, updates the new adjusted working variable as $\boldsymbol{z}_k^\top = \boldsymbol{z}_k + \boldsymbol{z}_k^a$, where $\boldsymbol{z}_k^a$ is a combination of the 'cross derivatives' multiplied by the difference in the relevant predictors,

defined for the four parameters as:

$$\mu : \boldsymbol{z}_1^a = -\boldsymbol{w}_{11}^{-1} \circ [\boldsymbol{w}_{12} \circ (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_2^\circ) + \boldsymbol{w}_{13} \circ (\boldsymbol{\eta}_3 - \boldsymbol{\eta}_3^\circ) + \boldsymbol{w}_{14} \circ (\boldsymbol{\eta}_4 - \boldsymbol{\eta}_4^\circ)]$$

$$\sigma : \boldsymbol{z}_2^a = -\boldsymbol{w}_{22}^{-1} \circ [\boldsymbol{w}_{12} \circ (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_1^\circ) + \boldsymbol{w}_{23} \circ (\boldsymbol{\eta}_3 - \boldsymbol{\eta}_3^\circ) + \boldsymbol{w}_{24} \circ (\boldsymbol{\eta}_4 - \boldsymbol{\eta}_4^\circ)]$$

$$\nu : \boldsymbol{z}_3^a = -\boldsymbol{w}_{33}^{-1} \circ [\boldsymbol{w}_{13} \circ (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_1^\circ) + \boldsymbol{w}_{23} \circ (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_2^\circ) + \boldsymbol{w}_{34} \circ (\boldsymbol{\eta}_4 - \boldsymbol{\eta}_4^\circ)]$$

$$\tau : \boldsymbol{z}_4^a = -\boldsymbol{w}_{44}^{-1} \circ [\boldsymbol{w}_{14} \circ (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_1^\circ) + \boldsymbol{w}_{23} \circ (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_2^\circ) + \boldsymbol{w}_{34} \circ (\boldsymbol{\eta}_3 - \boldsymbol{\eta}_3^\circ)].$$

Then, repeat using the modified backfitting algorithm until the convergence of the inner global deviance. The algorithm returns to the outer iteration which recalculates the quantities $\boldsymbol{z}_k, \boldsymbol{w}_{ks}$ and $\boldsymbol{\eta}_k^\circ$, for $k = 1, 2, 3, 4$ and $s = 1, 2, 3, 4$ and starts the inner iteration again.

Additionally, for estimating the hyperparameters $\boldsymbol{\lambda}$ there are tree main approach's: Likelihood-methods (RIGBY; STASINOPOULOS, 2005), Generalized Akaike information criteria (RIGBY; STASINOPOULOS, 2004) and Generalized cross validation (WOOD, 2017).

1.7 UNIVARIATE PENALIZED SMOOTHERS

Consider a usual regression problem where we have n observations of a random, variable Y, for a single explanatory variable with real values. In this particular case nonparametric regression methods are also called *scatterplot smoothers*. Therefore, the goal is estimating a function f such as:

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{x}) = f(\boldsymbol{x}) \tag{1.18}$$

To estimate f, using the usual methods as the IRLS algorithm, f must be represented in such a way that 1.18 becomes a linear model. For this purpose, we choose a linear basis function to represent f. If $b_j(x)$ is the j^{th} basis function, then f is assumed to have a representation

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j.$$

The linear smoother, such the fitted values $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^\top$ can be written in the form $\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$. The $\boldsymbol{S}$ matrix, is called *smoother matrix* and (BUJA; HASTIE; TIBSHIRANI, 1989) defines a $\boldsymbol{S}$ matrix, called *smoothing matrix*. (STASINOPOULOS et al., 2017) underscored the importance of penalized smoothers in the GAMLSS family of smoothers because of their flexibility and the fact they can be applied in a variety of different situations.

Let $\boldsymbol{Z}$ be an $n \times p$ basis matrix, which is defined in (2.5), γ a $p \times 1$ vector of parameters, $\boldsymbol{W}$ an $n \times n$ diagonal matrix of weights, $\boldsymbol{G}$ a $p \times p$ penalty matrix, λ a smoothing parameter

and y the variable of interest. Penalized smoothers are the solution to minimizing the quantity Q with respect to γ, more details are provided in (EILERS; MARX, 1996):

$$Q = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})^\top \boldsymbol{W}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^\top \boldsymbol{G}\boldsymbol{\gamma}. \tag{1.19}$$

The solution to the equation 1.19 is given by:

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{Z} + \lambda \boldsymbol{G})^{-1} \boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{y},$$

moreover, the fitted values for y are given in (STASINOPOULOS et al., 2017), by:

$$\hat{\boldsymbol{y}} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{Z} + \lambda \boldsymbol{G})^{-1} \boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{y} = \boldsymbol{S}\boldsymbol{y}. \tag{1.20}$$

So, the smoothing matrix plays a similar role to the hat matrix $\boldsymbol{H}$ in least squares estimation. The penalty matrix $\boldsymbol{G}$ is defined as $\boldsymbol{G} = \boldsymbol{D}_k^\top \boldsymbol{D}_k$, where the matrix $\boldsymbol{D}_k$ is a $(p-k) \times p$ difference matrix of order k. For example $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ matrices of order 1 and 2, respectively, look like:

$$\boldsymbol{D}_1 = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -1 & \cdots \end{bmatrix} \text{ and } \boldsymbol{D}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}$$

Usually, the matrix $\boldsymbol{D}_2$ is used to compute the $\boldsymbol{G}$ matrix used to estimate the *smoothing matrix*.

In the chapter 4 of (WOOD, 2017) the reader can find a detailed description of univariate smoothers, in particular how to represent the piecewise linear basis. Also, the chapter 8 and 9 of (STASINOPOULOS et al., 2017) provides a description for B-basis and linear additive terms for the GAMLSS context.

2 GLOBAL INFLUENCE FOR THE MODELS WITH LOCATION PARAMETER

It is important to ensure for the regression model that assumptions are not violated, otherwise the model may have undesirable characteristics like: bad predictions, misinterpretation of the fitted data behaviour, incoherent inferences, etc. Furthermore, occasionally, only a few observations can affect significantly the model and consequently the inferences as a whole. Firstly, (COOK, 1977) and (COOK, 1979) introduced the influence analysis for linear regression models. Basically, an observation is flagged as influential if after removing it from the data set, affects significantly the parameters of the fitted model.

Influential observations may occur for for a variety of reasons, including a typo, a system error, a measure error, experimental error, even just by causality or some other unknown cause on the experiment.

Several studies has been extended methods and measures to detect influential observations to other statistical models. (PREGIBON et al., 1981) has presented in diagnostics for the logistic regression model. (KIM; STORER, 1996) suggested a way to compute reference values for Cook's distance in the linear model via Monte Carlo simulation. (KIM; PARK; KIM, 2002) are provided some influence diagnostics for the semiparametric regression models and a way to compute references values for the Cook's distance in the semiparametric regression models using bootstrap.

Therefore, (PEÑA, 2005) proposed a new way to measure the influence, the Peña's measure is defined as the squared norm of the vector of changes of the forecast of one observation when each of the sample points are deleted one by one. More recently, (TÜRKAN; TOKTAMIS, 2013) compared the Cook's distance and Peña's measure for the semiparametric regression model using real and simulated data.

The main goal of this chapter is to study how to detect influential observations in the GAMLSS models. For this purpose, we focused on the most well known perturbation schemes: case-deletion (COOK; WEISBERG, 1982) using two measures: the generalised Cook's distance and the likelihood distance.

Moreover, we have suggested a approach using the Generalised Akaike Information Criteria (GAIC) and also a technique to compute a reference value for each measure based on bootstrap resamples.

2.1 DIAGNOSTIC TOOLS BASED ON RESIDUALS

The checking of model assumptions via normality of residuals is widely used the statistical literature for the classic regression linear model. (DUNN; SMYTH, 1996) proposed the normalised (randomised) quantile residuals, the main advantage of use this residuals is for any distribution of the response variables, the distribution of the residuals follow a standard normal distribution when the adopt model is correct. For the GAMLSS given the distribution of $f(y|\boldsymbol{\theta})$ and fit the observations $y_i, i = 1, \ldots, n$. The normalised (randomised) quantile residuals are given by

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i),$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function (cdf) of the normal distribution. The $\hat{u}_i's$ are the quantile residuals have distinct definitions for continuous and discrete responses.

If y is an observation of a continuous random variable so, let $u = F(y|\boldsymbol{\theta})$ and $\hat{u} = F(y|\hat{\boldsymbol{\theta}})$ is the model and the cdf's respectively. If the model is correctly specified, u have uniform distribution between zero and one, that is, $u \sim U(0, 1)$. This is called probability integral transform. The u is transformed in the true residual r(z-score) for $r = \Phi^{-1}(u)$ which have normal distribution if the model is correct. In a similar way, $\hat{u}$ is transformed in the adjusted residual $\hat{r}$ by $\hat{r} = \Phi(\hat{u}) = \Phi^{-1}[F(y|\hat{\theta})]$, and $\hat{r}$ have a approximate standard normal distribution.

If y is an observation of a discrete random variable, so $F(y|\boldsymbol{\theta})$ is a step function with jumps at the integers. The distribution of $u = F(y|\boldsymbol{\theta})$ has range zero to one, but is discrete with positive probability at the points $F(y|\boldsymbol{\theta})$. To deal with the discrete response variable, u is defined with a random value of a uniform distribution in the interval $[u_1, u_2] = [F(y - 1|\boldsymbol{\theta}), F(y|\boldsymbol{\theta})]$ and similarly $\hat{u}$ is a random variable of a uniform distribution in $[\hat{u}_1, \hat{u}_2] = [F(y - 1|\hat{\theta}), F(y|\hat{\theta})]$.

Using the fitted cdf, y is transformed for $\hat{u}$, randomly select of $(\hat{u}_1, \hat{u}_2)$, so the transformed residual for the adjusted residual $\hat{r} = \Phi^{-1}(\hat{u}))$ and $\hat{r}$ has approximate standard normal distribution.

The normalised residuals are useful to obtain important plots in the GAMLSS as:

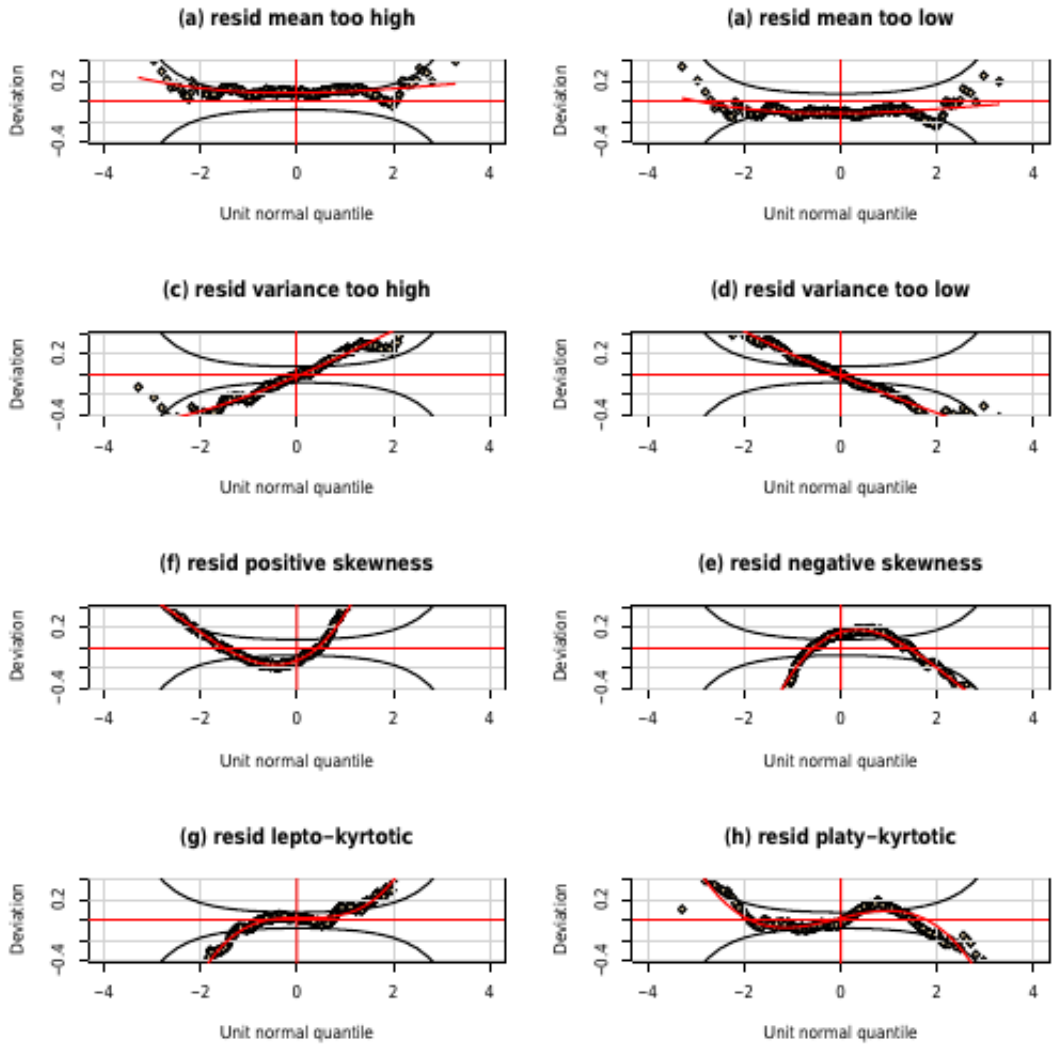
- Residuals against the fitted values of the μ parameter;
- residuals against the index;
- kernel density estimate of the residuals;

- Quantile-Quantile plot of the residuals;
- worm plots.

The worm plots are introduced by (BUUREN; FREDRIKS, 2001), and are used in the GAMLSS to identify possible model violations. The Figure 1, taken from (STASINOPOULOS et al., 2017), illustrate the possible failures in the model which can be identify by the worm plot, including the failures associated to the fit in the location, scale, skewness and kurtosis.

Originally, the worm plots are used as a diagnostic device for modelling growth reference curves, for the GAMLSS the normalised quantile residuals are used for generate the worm plots. Also, for the mean, if the worm passes above the origin, the fitted mean is too small and if the worm passes below the origin, the fitted mean is too large. For the variance, if the worm has a positive slope the fitted variance is too small and if the worm has a negative slope the fitted variance is too large. For the skewness, if the worm has a U-shape the fitted distribution is too skew to the left, and if the worm has an inverse U-shape the fitted distribution is too skew to the rigth. Finally for the kurtosis, if the worm has an S-shape on the left bent down the tails of the fitted distribution are too light and if if the worm has an S-shape on the left bent up the tails of the fitted distribution are too heavy.

Figure 1 – Worm plots indicating different types of model failures.



Source: (STASINOPOULOS et al., 2017).

2.2 LEAVE-ONE-OUT MEASURES BASED

A possible approach to perform a measure of global influence for the GAMLSS model (1.7) is dropping the i-th observation on the data set, and study the effect of this in the model. This methodology is known as case-deletion or leave-one-out. Initially, in particular we are interested to study the impact of the i-th observation on the estimates of μ, σ, ν and τ, respectively.

There are some measures to reveal the impact of the ith observation on the estimates. Moreover, let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator of $\boldsymbol{\theta}$, we have special desire to study the difference between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{(i)}$, where $\hat{\boldsymbol{\theta}}_{(i)}$ is the MLE of $\hat{\boldsymbol{\theta}}_{(i)}$ removing the ith observation. The basic idea behind this approach is to compare $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{(i)}$ and verify if this make a seriously

changing in the model, in other words if the inference is considerable affected by removing this particular case. In this work we focus on this methodology to detect influential observations. The R source code for generate the index plots and all measures of this work is available in <https://rpubs.com/LucasSilva/743391>.

2.2.1 Likelihood Distance

The basic idea of the likelihood distance (COOK, 1977), (COOK; WEISBERG, 1982), (COOK, 1986), is assess the influence of the i-th observation on the maximum likelihood distance, so we have to compare the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$. defined by:

$$LD_i(\boldsymbol{\theta}) = 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})].$$

for the GAMLSS models $l(\cdot)$ are computed as the pseudo-likelihood, note then $LD_i(\boldsymbol{\theta})$ can be assume negative values. Similarly, we can assess the specifics measures in each parameter for the model by $LD_i(\boldsymbol{\mu})$, $LD_i(\boldsymbol{\sigma})$, $LD_i(\boldsymbol{\nu})$, and $LD_i(\boldsymbol{\tau})$. In this work we focus in situations tha modelling the parameter μ, so we focus in computing $LD_i(\boldsymbol{\mu})$, the measures $LD_i(\boldsymbol{\sigma})$, $LD_i(\boldsymbol{\nu})$, and $LD_i(\boldsymbol{\tau})$ may be not comparable with the others measures used like Peña's measure or Kim's measure.

The LD_i can take positive or negative values, the interpretation of is that positive values indicate poorer fit associated with removing case i as the log-likelihood of the full sample solution decreased, also negative values indicate that removing case i to improve the model fit compared with the original sample, more details about the likelihood distance can be find in (PEK; MACCALLUM, 2011).

2.2.2 Generalised Cook's Distance

For the classic linear model 1.2 is well established the Cook's distance, which provides the measure between $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\boldsymbol{\beta}}$, defined by (COOK, 1977), (COOK, 1986):

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps^2}, \text{ for } i = 1, \ldots, n.$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the fitted coefficients when the observation i is deleted.

For the GAMLSS another alternative is using, the standardised norm of $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$, this measure is known as the generalised Cook distance (RAMIRES et al., 2018), defined by

$$\text{CD}_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^\top [-\boldsymbol{I}(\hat{\boldsymbol{\theta}})](\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}), \tag{2.1}$$

where, $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ is the observed information matrix, defined in (2.2)

For the parametric GAMLSS model (1.10), the asymptotic distribution of $\boldsymbol{\theta}_T$ is

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_T, \boldsymbol{i}(\boldsymbol{\theta})^{-1}),$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_T$ is the assumed true vale. The Fisher expected information matrix evaluated at $\hat{\boldsymbol{\theta}}_T$ given by

$$\boldsymbol{i}(\boldsymbol{\theta}_T) = \mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}}^\top\right]_{\boldsymbol{\theta}_T},$$

in some situations it is not possible to compute the expected information $\boldsymbol{i}(\boldsymbol{\theta}_T)$ analytically and therefore given by (2.2) is used

$$I(\boldsymbol{\theta}_T) = -\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_T}, \tag{2.2}$$

in others words, $I(\boldsymbol{\theta})$ is equal to the negative of the Hessian matrix of the log-likelihood function at $\boldsymbol{\theta}$. In general, for the parametric GAMLSS models the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_T, \boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}).$$

If the model is a semiparametric GAMLSS model, the pseudo-likelihood is used instead and the Hessian matrix is numerically computed. Moreover, if the model specification is incorrect then, under regularity conditions, the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ may be approximated by

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_c, \boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1} K(\hat{\boldsymbol{\theta}}) \boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}),$$

where $\boldsymbol{\theta}_c$ is the closest value of $\boldsymbol{\theta}$ to the true model measured by a weighted Kullback-Leibler distance, and $K(\hat{\boldsymbol{\theta}})$ is an estimate of the variance-covariance matrix of the fist derivative of log-likelihood function with respect to the parameters. If the model is incorrect then $\hat{\boldsymbol{\theta}}$ is not a consistent estimator of $\boldsymbol{\theta}_T$.

In a similar way, it's possible to calculate the value $\text{CD}_i(\boldsymbol{\mu})$, $\text{CD}_i(\boldsymbol{\sigma})$, $\text{CD}_i(\boldsymbol{\nu})$ and $\text{CD}_i(\boldsymbol{\tau})$. In this work we focus in situations that modelling the parameter μ, so we focus in computing $\text{CD}_i(\boldsymbol{\mu})$. The measures $\text{CD}_i(\boldsymbol{\sigma})$ and $\text{CD}_i(\boldsymbol{\nu})$, $\text{CD}_i(\boldsymbol{\tau})$ may be not comparable with the others measures used like Peña's measure or Kim's measure.

2.2.3 Leave-One-Out GAIC

The basic idea behind the Akaike Information Criteria (AIC), introduced by (AKAIKE, 1974), is select a model with a reduced number of parameters, that is a more parsimonious model. The AIC for the usual linear regression model is defined by

$$\text{AIC} = -L(\hat{\beta}) + p = n\text{log}\left(\frac{D(y|\hat{\mu})}{n}\right) + 2p,$$

where $D(y|\hat{\mu}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$.

An alternative, for the usual linear regression model, as well, is the Schwarz Bayesian Criteria (BIC) introduced by (SCHWARZ et al., 1978), which is defined by

$$\text{BIC} = p\ \text{log}(n) - 2\text{log}(\hat{L}).$$

where, $\hat{L}$ is the maximised value of the likelihood function of the model.

A extension for the GAMLSS is the Generalised AIC (GAIC), is given by

$$\text{GAIC}(\kappa) = -2l(\hat{\beta}, \hat{\gamma}) + \kappa \cdot \text{df},$$

where κ is a constant of penalty and df is the effective degrees of freedom. The GAIC(κ) is maximised globally over λ. The AIC can be obtained from the GAIC with $\kappa = 2$ and the SBC with $\kappa = \text{log}\ n$.

In this work we compare the difference between a GAIC of the model fitted with the full sample (GAIC(κ)) and the model fitted dropping the i-th observation (GAIC$_{(i)}(\kappa)$). Thus, we call this method Leave-One-Out GAIC, with is defined as

$$\text{GAIC}(i) = \text{GAIC}(\kappa) - \text{GAIC}_{(i)}(\kappa). \tag{2.3}$$

Similarly to the likelihood distance the values in the equation (2.3) can be negative values. In the section 2.3 we provide the method for compute the reference values for the Leave-One-Out GAIC.

2.2.4 Kim's Measures for Semiparametric Models

Consider a semiparametric regression model, which is a GAMLSS submodel like (1.7), but with a single univariate additive term $s_1(\cdot)$, and modelling the $\boldsymbol{\mu}$, with any response distribution. Hence, consider the model

$$\boldsymbol{Y} \overset{ind}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}),$$

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + s(\boldsymbol{t}). \tag{2.4}$$

Following the approach of (KIM; PARK; KIM, 2002), we consider a measure for three components in this GAMLSS submodel, the fitted $\hat{\boldsymbol{\beta}}$'s, the fitted $\hat{s}$'s that is to say, the influence corresponding to the smoothing additive terms, and the mean response.

In this work we focus in a important class of smothers, so we use P-Splines as smoother for the model (2.4). The P-Splines are based in penalised B-Splines, where each basis function is only non-zero over the intervals between $m+3$ adjacent knots, where $m+1$ is the order of the basis (hence, $m=2$ is used for the cubic splines). We use $k+m+2$ knots for the basis, where $x_1 < x_2 < \ldots < x_{k+m+2}$, by default in the GAMLSS models we use $m=2$ and $k=10$ if $n<99$ and $k=20$ if $n \geq 100$. An $(m+1)^{\text{th}}$ order spline can be represented as

$$f(x) = \sum_{i=1}^{k} Z_i^m(x)\beta_i$$

where the B-spline function are defined recursively as:

$$Z_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} Z_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m+2} - x_{i+1}} Z_{i+1}^{m-1}(x), i = 1, \ldots, k. \tag{2.5}$$

and

$$Z_i^{-1} = \begin{cases} 1 & x_i \leq x \leq x_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

More details about the linear basis for the P-Splines can be find in (WOOD, 2017).

Let $\boldsymbol{S}$ be the linear smoother matrix, (BUJA; HASTIE; TIBSHIRANI, 1989), shows that $\boldsymbol{S}$ depends on covariates as well the particular smoother, but not on $\boldsymbol{y}$. And given a linear smoothing algorithm, we can compute the corresponding smoother matrix $\boldsymbol{S}$.

Let $\boldsymbol{X}$ be the design matrix with $\boldsymbol{x}_i^\top$ as its ith row, and $\boldsymbol{y}$ be a response vector. Now, we consider the follow hat matrix $\tilde{\boldsymbol{H}} = (\boldsymbol{I} - \boldsymbol{S})^{-1}\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^\top(\boldsymbol{I} - \boldsymbol{S})$. Then, we can write

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{H}}\boldsymbol{y} \text{ and } \hat{\boldsymbol{m}}(\boldsymbol{t}) = \boldsymbol{S}(\boldsymbol{I} - \tilde{\boldsymbol{H}})\boldsymbol{y},$$

respectively. Also, the vector of fitted responses equals $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ where

$$\boldsymbol{H} = \boldsymbol{S} + (\boldsymbol{I} - \boldsymbol{S})\tilde{\boldsymbol{H}}, \tag{2.6}$$

is the hat matrix. The residual vector is then given by

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \tilde{\boldsymbol{H}})(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{y}.$$

Note that $\boldsymbol{H} = \tilde{\boldsymbol{H}} + \boldsymbol{H}^*$, where $\boldsymbol{H}^* = \boldsymbol{S}(\boldsymbol{I} - \tilde{\boldsymbol{H}})$, which will be used in defining and interpreting Cook's distances for measuring the influence over the mentioned components of the GAMLSS model.

An influence measure for the ith observation on $\hat{\boldsymbol{\beta}}$ may be defined as a type of Cook's distance by:

$$\tilde{C}_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^\top \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{\sigma^2 \text{tr}(\tilde{\boldsymbol{H}})}.$$

We can express it as a function of the ith residual and leverage, i.e.,

$$\tilde{C}_i = \frac{1}{p\sigma^2} \frac{\tilde{e}_i^2 \tilde{h}_{ii}}{(1 - \tilde{h}_{ii})^2}, \tag{2.7}$$

where $\tilde{e}_i$ is the component of residual vector $\boldsymbol{e} = (\boldsymbol{I} - \tilde{\boldsymbol{H}})\tilde{\boldsymbol{y}}$ and $\tilde{h}_{ii}$ is the ith diagonal component of $\tilde{\boldsymbol{H}}$. Here we take the P-splines smoother and define the corresponding cook's distance for $\hat{\boldsymbol{s}}$ suggested by (KIM; PARK; KIM, 2002). We may define a type of Cook's distance for the ith observation by

$$C_i^* = \frac{\{\hat{m}(t_i) - \hat{m}_{(i)}(t_i)\}^2}{\sigma^2 \text{tr}(\boldsymbol{H}^*)},$$

where h_{ii}^* is the ith diagonal element of $\boldsymbol{H}^*$ and e_i^* is the ith component of residual vector $\boldsymbol{e}^* = (\boldsymbol{I} - \boldsymbol{H}^*)\boldsymbol{y}$.

Following the work of (KIM; PARK; KIM, 2002), we may express it as

$$C_i^* = \frac{(h_{ii}^* e_i^*)^2}{(1 - h_{ii}^*)^2 \sigma^2 \text{tr}(\boldsymbol{H}^*)}. \tag{2.8}$$

An influence measure for the ith observation on the vector of fitted values can be similarly defined by

$$C_i = \frac{(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(i)})^\top (\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(i)})}{\sigma^2 \text{tr}(\boldsymbol{H})},$$

where $\boldsymbol{H} = \boldsymbol{S} + (\boldsymbol{I} - \boldsymbol{S})\tilde{\boldsymbol{H}}$, as is given in (2.6). Also, we can express it as a function of the corresponding residual and leverage. Let h_{ii} be the ith diagonal element of $\boldsymbol{H}$ and e_i be the ith component of $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$. Then

$$C_i = \frac{1}{\sigma^2 \text{tr}(\boldsymbol{H})} \frac{e_i^2 h_{ii}}{(1 - h_{ii})^2}. \tag{2.9}$$

Note that this has the same form as the original Cook's distance in the classical linear model.

This particular in approach has the advantage of provide the influence of an observation on $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{y}}$ separately by (2.7), (2.8), (2.9) respectively. This can be handy for know how to raise the model to improve the fit, in other words, identify what component in the model can be change.

2.2.5 Peña's Measure

In some situations outliers can be undetected by Cook's distance in linear models, but can be detected by the Peña's measure. This fact has been demonstrated and in (PEÑA, 2005) when the Peña's measure was proposed.

The Peña's measure is defined as the squared norm of the vector of changes of the forecast of one observation when each of the sample points are deleted one by one. This measure is very effective in detection of high leverage outliers, that can be not deteced by the Cook's distance in large datasets.

Consider the model (1.1), let us denote $\hat{\boldsymbol{\beta}}_{(i)}$ as the vector of coefficients when the i-th case are deleted, and $\hat{\boldsymbol{y}}_{(i)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(i)}$ be the corresponding vector of forecast. Essentially, this approach measure how each observation is being influenced by the rest of the data. For that purpose, this can be done computing the follows vectors:

$$\boldsymbol{s}_i = (\hat{y}_i - \hat{y}_{i(1)}, \ldots, \hat{y}_i - \hat{y}_{i(n)})^\top,$$

that is to say, we see at how sensitive the forecast of the i-th observation is to the deletion of each observation in the sample.

Finally, the Peña's measure for the i-th observation, S_i, as the squared norm of the standardized vector $\boldsymbol{s}_i$, that is,

$$P_i = \frac{\boldsymbol{s}_i^\top \boldsymbol{s}_i}{p\hat{\text{var}}(\hat{y}_i)}, \tag{2.10}$$

where $\hat{\text{var}}(\hat{y}) = s^2 h_{ii}$, and h_{ii} is the i-th diagonal element of the hat matrix and $s^2 = \dfrac{\boldsymbol{e}^\top \boldsymbol{e}}{n-p}$.

One computational advantage is its more ease to obtain reference values, (PEÑA, 2005) suggested a criteria to determine if the observation are influential (in other words, if P_i are large enough), if P_i exceeds the median value of $P_i + 4.5\text{MAD}(P_i)$, where $\text{MAD}(P_i) = \text{median}\left\{\dfrac{|P_i - \text{median}(P_i)|}{0.6745}\right\}$.

The equation (2.10), we consider the original measure for linear models, but when additive terms are included, the computation of hat matrix changes. Let consider the GAMLSS submodel (1.5), using a single smother additive term s, that is:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + s(\boldsymbol{t}),$$

the fitted values vector can be written as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{S}\boldsymbol{y},$$

where $\boldsymbol{S}$ are the smoothing matrix. For this class of models, the diagonal elements of the hat matrix h_{ii} are taken from the modified version for the semiparametric models in (2.6) proposed by (KIM; PARK; KIM, 2002).

The adjusted Penã's measure for the semiparametric regression are derived in (TÜRKAN; TOKTAMIS, 2013), for this work we estimating the smoother matrix for the P-Splines (1.20) and not for the local polynomial smoother as the original work.

2.3 REFERENCE VALUES

For the semi-parametric models (KIM; PARK; KIM, 2002) proposed a approach to calculate reference values for Cook distance using bootstrap. For the GAMLSS we following the same idea. But in that case, we run a non-parametric bootstrap for approximate a confidence interval and use it as references values to detect influential cases.

The procedure is based in compute a confidence interval based on bootstrap percentiles for a more detailed reading about this see (EFRON; TIBSHIRANI, 1994) chapter 13. This approach consists in re-sample with replacement the observations of the original data set D_o, and generate B bootstrap samples $D_b, b = 1, \ldots, B$, with a large number of samples with same sample size N of the original data for each bootstrap replica.

The next step, is fit a model with the same specification for each one of this B replicas, and compute the likelihood distance for each case in his respective replica. Let $\mathsf{LD}_b^*(\hat{\boldsymbol{\theta}})$ be the of the likelihood distances for a b-th fitted model.

Let $\hat{G}$ be the empirical cumulative distribution function of $\mathsf{LD}_i^*(\hat{\boldsymbol{\theta}})$, the $(1-2\alpha)$ percentile interval is defined by the α and $1-\alpha$ percentiles of $\hat{G}$

$$[\mathsf{LD}(\hat{\boldsymbol{\theta}})_{\%,\mathsf{lower}}; \mathsf{LD}(\hat{\boldsymbol{\theta}})_{\%,\mathsf{upper}}] = [\hat{G}^{-1}(\alpha); \hat{G}^{-1}(1-\alpha)],$$

in this case we use $\alpha = 0.05$ as default. Since, by definition $\hat{G}^{-1}(\alpha) = LD_b^*(\hat{\boldsymbol{\theta}})$, the $100 \cdot \alpha$th percentile of the bootstrap distribution, we can also write the percentile interval as

$$[\mathsf{LD}(\hat{\boldsymbol{\theta}})_{\%,\mathsf{lower}}; \mathsf{LD}(\hat{\boldsymbol{\theta}})_{\%,\mathsf{upper}}] = [\mathbf{LD}^{*(\alpha)}(\hat{\boldsymbol{\theta}})_{\%,\mathsf{lower}}; \mathbf{LD}^{*(1-\alpha)}(\hat{\boldsymbol{\theta}})_{\%,\mathsf{upper}}], \tag{2.11}$$

we consider a i-th observation as influential if the value of $\mathsf{LD}_i(\hat{\boldsymbol{\theta}})$ it's out of the interval (2.11). The Algorithm 2, provides the detailed steps of this procedure using pseudo-code.

A similar approach was conducted to obtain reference values for generalized cook distance. However, the expression (2.1) only provides positive values, so we compute the $(1-\alpha) = 95\%$ percentile instead in this case.

Algoritmo 2: Computing reference values for likelihood distance.

Input: A data set;
Output: A confidence interval $[\mathbf{LD}^{*(\alpha)}(\hat{\boldsymbol{\theta}})_{\%,\text{lower}}; \mathbf{LD}^{*(1-\alpha)}(\hat{\boldsymbol{\theta}})_{\%,\text{upper}}]$;
Declare: LD[n] a empty vector with size n (to the likelihood distances of the original sample);
LD*[B] a empty vector with size B (to the likelihood distances of the bootstrap samples);
LD$_F$[B] a empty vector with size nB (to all computed likelihood distance of the boostraps samples);
(1) fit the observed data using a GAMLSS model;
(2) **for** i := 1 **to** n **do** ;
LD[i] $\leftarrow 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})]$;
(3) generate B bootstrap samples (B large e.g 1000) ;
(4) fit a GAMLSS model with the same specification for each model fitted in step (3);
(5) **for** n = 0 **to** B **do**;
for j = 0 **to** n **do**;
LD$^*_j[b] \leftarrow 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})]$;
(6) **for** j = 0 **to** nB **do**;
LD$_F[j] \leftarrow$ LD$^*[b]$;
(7) **Return**: the 95% bootstrap percentile confidence interval of $LD_F(\hat{\boldsymbol{\theta}})$:;
$[\mathbf{LD}^{*(\alpha)}(\hat{\boldsymbol{\theta}})_{\%,\text{lower}}; \mathbf{LD}^{*(1-\alpha)}(\hat{\boldsymbol{\theta}})_{\%,\text{upper}}]$.

The idea is re-sampling the original data to generate B bootstrap samples and fit a GAMLSS model with the same specification for each bootstrap sample, and compute the generalized Cook's distance for each sample. Then finally, compute the $(1-\alpha)$ bootstrap percentile of all generalized Cook's distance computed. The Algorithm 3 give us a guideline for this procedure.

Similary with the Algorithm 3, to obtain the references values for the Kim's measure we adjust the original procedure of (KIM; PARK; KIM, 2002). In this case for each bootstrap sample we compute C_i^*, $\tilde{C}_i$ and C_i. Next, we compute the $(1-\alpha)$ bootstrap percentile for all each measure computed on the bootstraps samples (tree bootstraps generated on the reference process, in this case we use the bootstraps percentiles, one for each measure), usually $\alpha = 0.05$. In this case we obtain tree reference values, one for each measure provided by the Kim's measure. The algorithm 4 provides the details of each step.

In the sections 3 and 4, we use this procedure to obtain reference values for simulated data and the original application of (KIM; PARK; KIM, 2002).

Similarly to Likelihood distance, the reference values for the Leave-One-Out GAIC for a given κ is described in the algorithm 5.

Algoritmo 3: Computing reference values for generalized Cook's distance.

Input: A data set and the GAMLSS model;
Output: R (a numeric reference value);
Declare: CD[n] (a empty vector with size n);
CD*[B] (a empty vector with size B);
CD$_F$[B] (a empty vector with size nB);
(1) fit the observed data using a GAMLSS model;
(2) **for** i := 0 **to** n **do** ;
CD$_i(\boldsymbol{\theta}) \leftarrow (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^\top [-\boldsymbol{I}(\hat{\boldsymbol{\theta}})](\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})$;
(3) generate B bootstrap samples (B large e.g 1000) ;
(4) fit a GAMLSS model with the same specification for each model fitted in the step (3);
(5) **for** b = 0 **to** B **do**;
for j = 0 **to** n **do**;
CD$^*_j[b] \leftarrow (\hat{\boldsymbol{\theta}}^*_{(i)} - \hat{\boldsymbol{\theta}}^*)^\top [-\boldsymbol{I}(\hat{\boldsymbol{\theta}}^*)](\hat{\boldsymbol{\theta}}^*_{(i)} - \hat{\boldsymbol{\theta}}^*)$;
(6) **for** j = 0 **to** nB **do**;
CD$_F[j] \leftarrow$ CD$^*[b]$;
(7) **Return**: the $(1-\alpha)$ bootstrap percentile of $CD_F(\hat{\boldsymbol{\theta}})$.

Algoritmo 4: Computing reference values for kim's Measure.

Input: A data set and the GAMLSS model;
Output: R (a numeric reference value);
Declare: $C^*_i[n]$ (a empty vector with size n);
$C^*_i[n]$ (a empty vector with size n);
$\tilde{C}_i[n]$ (a empty vector with size n);
C_i (a empty vector with size n);
$C^*_i[B]$ (a empty vector with size B);
$\tilde{C}_i[B]$ (a empty vector with size B);
C_i (a empty vector with size B);
(1) fit the observed data using a GAMLSS model;
(2) **for** i := 0 **to** n **do** ;
compute: C^*_i, $\tilde{C}_i$ and C_i;
(3) generate B bootstrap samples (B large e.g 1000) ;
(4) fit a GAMLSS model with the same specification for each model fitted in the step (3);
(5) **for** b :=0 **to** B **do**;
compute: $C^*_i[B]$, $\tilde{C}_i[B]$ and C_i;
(7) **Return**: the $(1-\alpha)$ bootstrap percentile of $CD_i(\hat{\boldsymbol{\theta}})$.

Algoritmo 5: Computing reference values for Leave-One-Out GAIC.

Input: A data set;

Output: $[\text{GAIC}^{*(\alpha)}(\hat{\boldsymbol{\theta}})_{\%,\text{lower}}; \text{GAIC}^{*(1-\alpha)}(\hat{\boldsymbol{\theta}})_{\%,\text{upper}}]$ (a confidence interval);

Declare: GAIC[n] (a empty vector with size n);
GAIC*[B] (a empty vector with size B);

(1) fit the observed data using a GAMLSS model;

(2) **for** i $= 1$ **to** n **do** ;
GAIC$(i) =$ GAIC$(\kappa) -$ GAIC$_{(i)}(\kappa)$;

(3) generate B bootstrap samples (B large e.g 1000) ;

(4) fit a GAMLSS model with the same specification for each model fitted in step (3);

(5) **for** n $= 0$ **to** B **do**;
for j $= 0$ **to** n **do**;
GAIC$^*_j[b] \leftarrow$ GAIC$(\kappa) -$ GAIC$_{(i)}(\kappa)$;

(6) **for** j $= 0$ **to** nB **do**;
GAIC$_F[j] \leftarrow$ GAIC$^*[b]$;

(7) **Return**: the 95% bootstrap percentile confidence interval of $LD_i(\hat{\boldsymbol{\theta}})$.

3 INFLUENCE MEASURES FOR ARTIFICIAL DATA

The main goal of this chapter is simulate several different scenarios for univariate splines and semiparametrical models with different responses distributions: normal, poisson, gumbel and the skew power exponential type 3. Then, we compute the influential measures aforementioned, and compute the correspondents references values.

3.1 SIMULATED DATA FOR THE SIMPLE LINEAR REGRESSION MODEL

For the simple linear regression model the data was generated from

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n, \tag{3.1}$$

where $\beta_0 = 4$, $\beta_1 = 2$ and $\epsilon_i \overset{ind}{\sim} N(0, 4)$, $\boldsymbol{x}$ is the fixed vector of co-variables, in this case $\boldsymbol{x}$ is the sequence $1, 2, \ldots, n$, because $\boldsymbol{x}$ can be any fixed vector. We consider four different scenarios with $n_1 = 90, n_2 = 150, n_3 = 300$ and $n_4 = 500$. For this model the low variance are used to obtain a good fit and a evident influential observation. To artificially include a influential point we generate an outlier observation in the response variable based on the Tukey's method for detect outlier in univariate data, (TUKEY, 1977) proposes a approach based on the interquartile range, a observation of a univariate data are flagged as outlier if there are out of range:

$$[LB, UB] = [Q_1(Y) - k(Q_3(Y) - Q_1(Y)), Q_3(Y) + k(Q_3(Y) - Q_1(Y))], \tag{3.2}$$

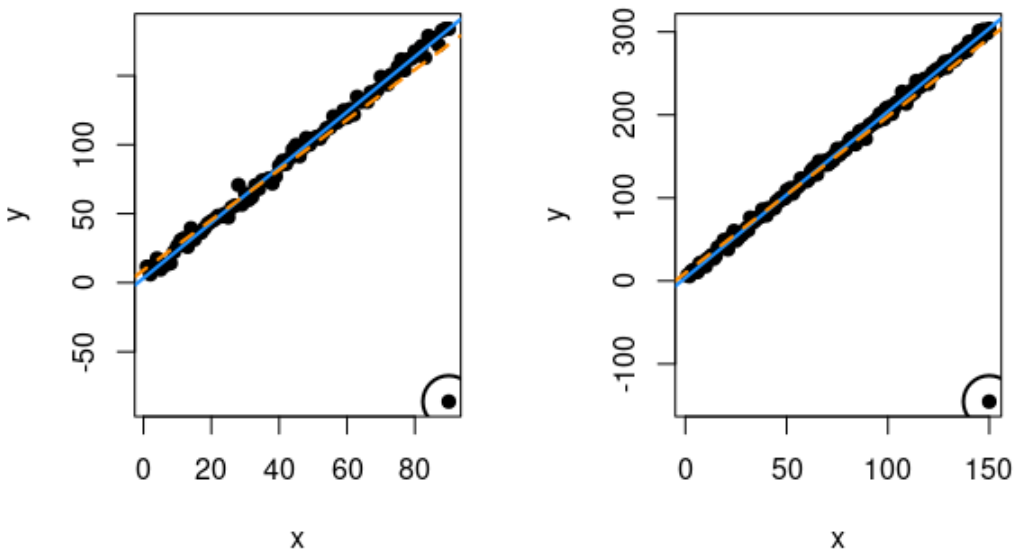
where $Q_1(Y)$ are the lower quartile (25th percentile) and $Q_3(Y)$ are the uper quartile (75th percentile), k is a positive constant, usually $k = 1.5$ are used, and $k = 3$ to outliers "far out", that is problematic cases. We use $k = 1.5$ to generate a lower outlier for the response variable and the maximum value of the explanatory variable x_i, $i = 1, \ldots, n + 1$.

To include a specific new outlier we take the maximum value of the covariables and use the lower-bound computed by the tukey criteria, or similarly use the upper bound of the tukey criteria with the minimum value of the covariables.

The Figures 2 and 3 shown the dispersion and the fitted linear models with and without the influential case (blue curve and yellow respectively).

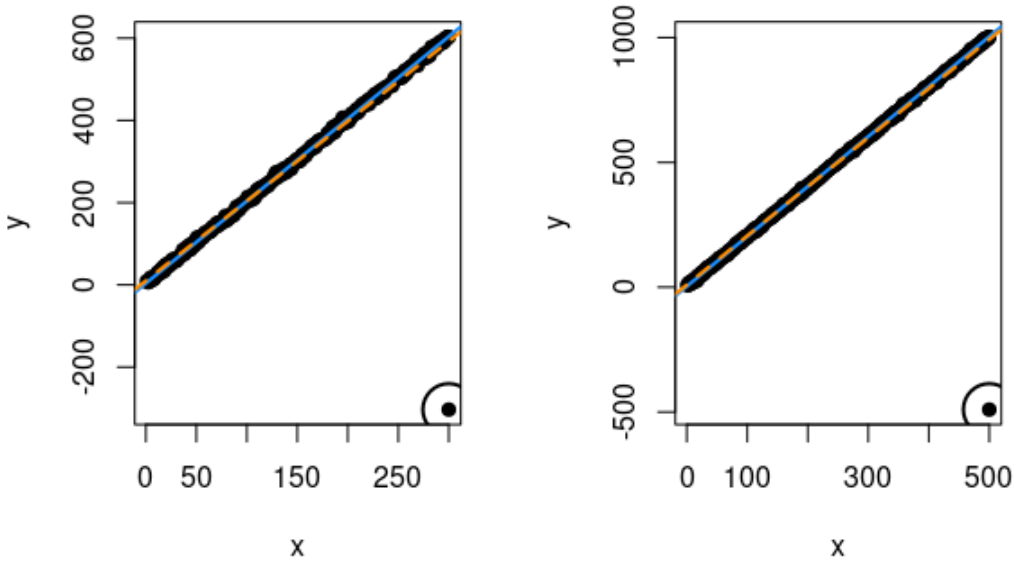
The index of the influential observation is $n+1$, so when we generate a sample with $n_1 = 90$ the observation 91 is artificially included as influential. Also the Figure 4 shown the index

Figure 2 – Dispersion and fitted linear model with and without the influential case with sample size $n_1 = 90$, $n_2 = 150$ respectively.



Source: Author's own.

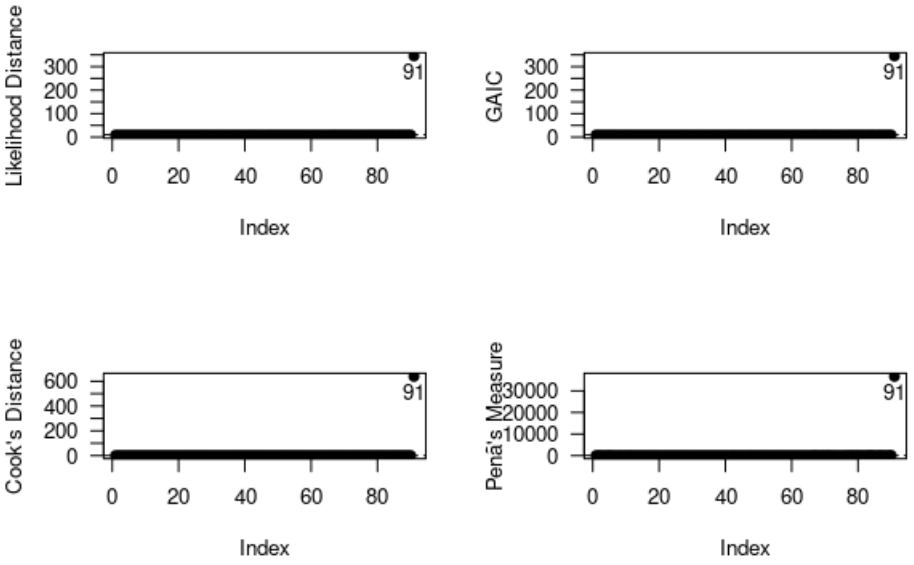
Figure 3 – Dispersion and fitted linear model with and without the influential case with sample size $n_3 = 300$, $n_4 = 500$ respectively.



Source: Author's own.

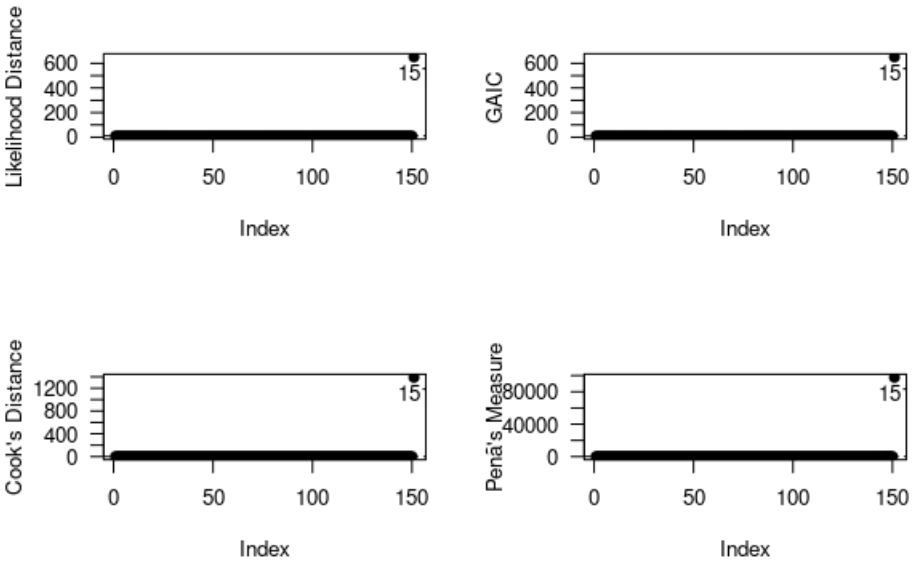
plot for the Likelihood and Cook's Distance. For example, the bootstrap confidence interval obtained for the likelihood distance using 1000 bootstrap resamples are $[1.631801; 7.491019]$, thus the observation 91 for the scenario with $n = 91$ flagged as the observation 91 as influential. The Figures 4, 5, 6 and 7. Shows the results for the Likelihood distance, Leave-One-Out GAIC, Cook's distance and Peña's measure for the scenarios with $n_1 = 90, n_2 = 150, n_3 = 300$ and $n_4 = 500$ respectively. For this cases, the four scenarios has similar results for any measure used.

Figure 4 – Index Plots for the simulated data from the model (3.1), with Likelihood Distance, GAIC Distance, Cook's Distance, and Peña's Measure, respectively. For the scenario with sample size $n_1 = 90$.



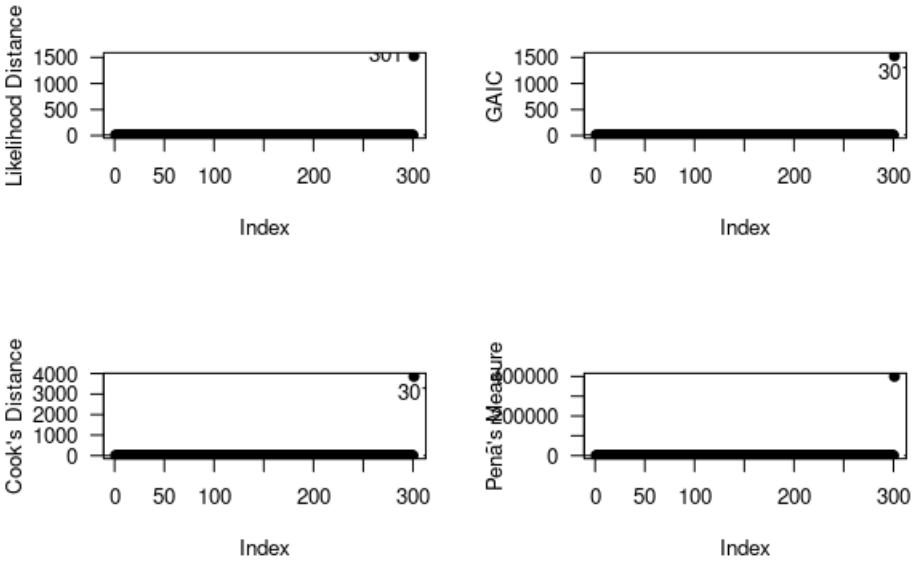
Source: Author's own.

Figure 5 – Index Plots for the simulated data from the model (3.1), with Likelihood Distance, GAIC Distance, Cook's Distance, and Peña's Measure, respectively. For the scenario with sample size $n_2 = 150$.



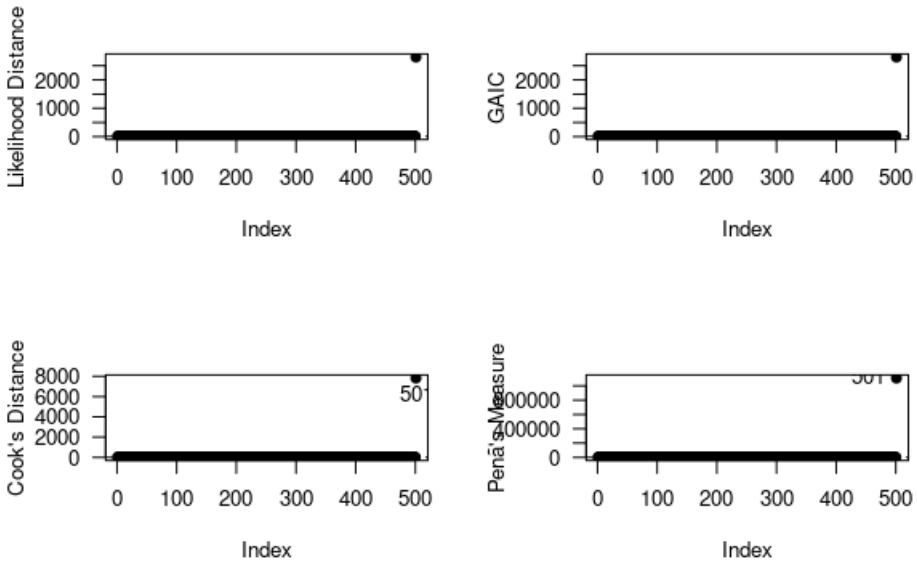
Source: Author's own.

Figure 6 – Index Plots for the simulated data from the model (3.1), with Likelihood Distance, GAIC Distance, Cook's Distance, and Peña's Measure, respectively. For the scenario with sample size $n_3 = 300$.



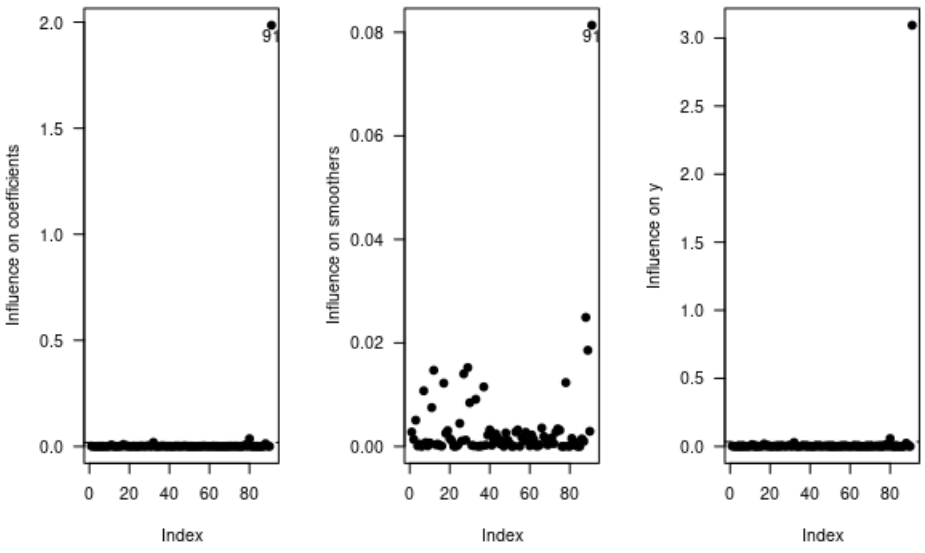
Source: Author's own.

Figure 7 – Index Plots for the simulated data from the model (3.1), with Likelihood Distance, GAIC Distance, Cook's Distance, and Peña's Measure, respectively. For the scenario with sample size $n_4 = 500$.



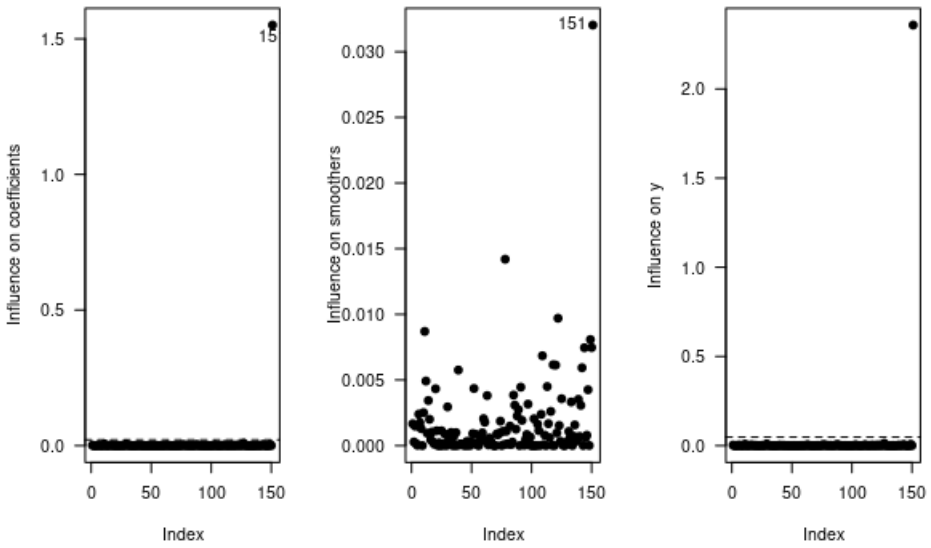
Source: Author's own.

Figure 8 – Index Plots of the Kim's measure for the simulated data from the model (3.1), for the scenario with sample size $n_1 = 90$.



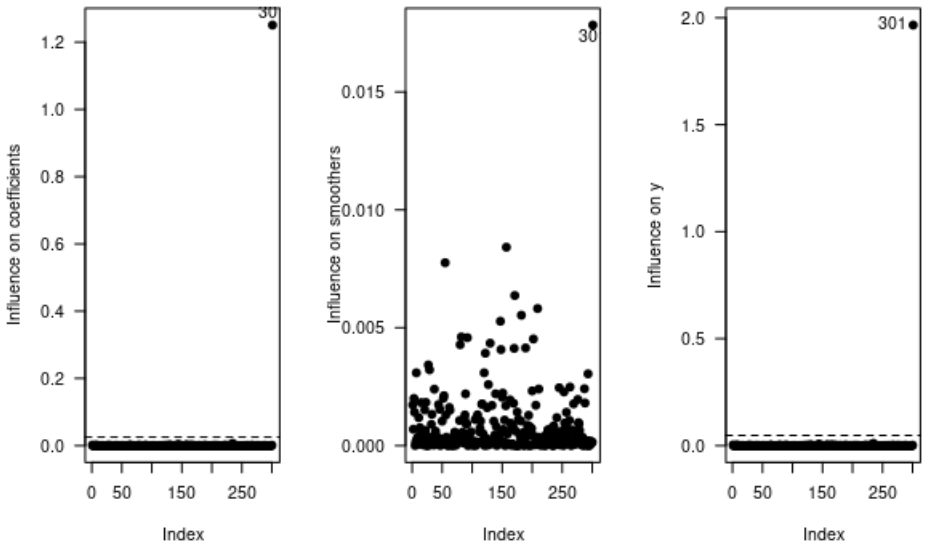
Source: Author's own.

Figure 9 – Index Plots of the Kim's measure for the simulated data from the model (3.1), for the scenario with sample size $n_2 = 150$.



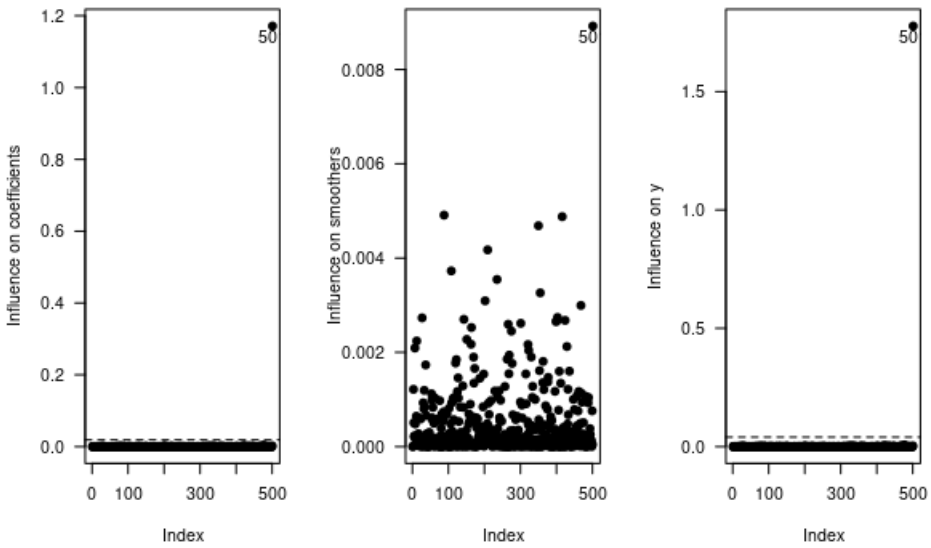
Source: Author's own.

Figure 10 – Index Plots of the Kim's measure for the simulated data from the model (3.1), for the scenario with sample size $n_3 = 300$.



Source: Author's own.

Figure 11 – Index Plots of the Kim's measure for the simulated data from the model (3.1), for the scenario with sample size $n_4 = 500$.



Source: Author's own.

3.2 SIMULATED DATA FOR UNIVARIATE PENALIZED SMOOTHERS

Now, we simulate a dataset to modeling a response variable with a p-spline with the model 1.18, for four different scenarios $n_1 = 90, n_2 = 150, n_3 = 300$ and n_4500 observations, using the follow functional relationship:

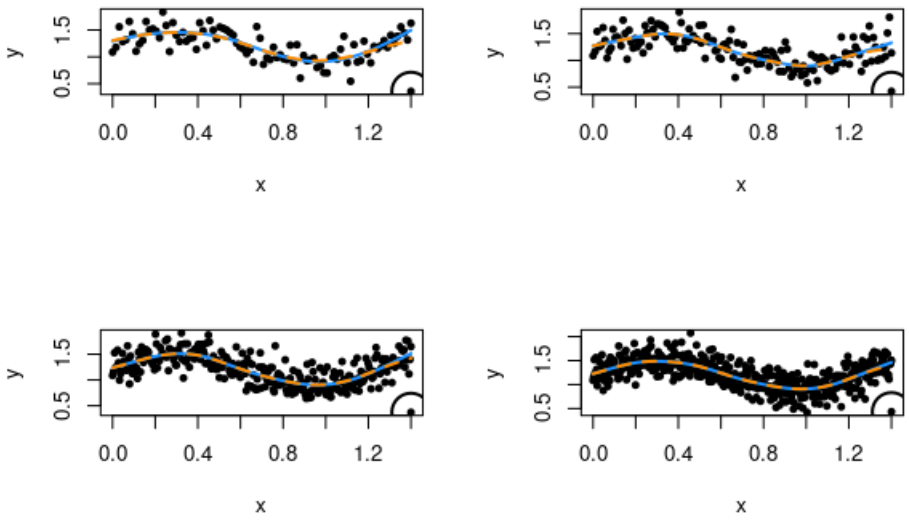
$$Y_i = 1 + 2\text{sin}(5\pi x_i) + \varepsilon_i, \tag{3.3}$$

where $\varepsilon_i \sim N(0, 1)$. The Figure 12 show the dispersion and the fitted curves for each scenario, the circle highlight the influential observation. $\boldsymbol{x}$ is the vector of observed covariates. For each scenario we artificially included a influential point by the Tukey's method criteria, where the index of the influential observation is $n+1$, that is, for the simulation with $n_1 = 90$ for example the observation 91 was included as influential. The response was included by the lower bound of the interval (3.2) and the covariate x are the maximum value in $\boldsymbol{x}$.

The model (3.3) is non-parametric, so in this case we use the likelihood distance, Leave-One-Out GAIC and generalized Cook's distance, the peñas measure and kim's measures need the $\boldsymbol{H}$ matrix and this model uses the smoother matrix instead. The Figures 13, 14, 15 and 16 show the results of the index plots for the scenarios with $n_1 = 90$, $n_2 = 150$, $n_3 = 300$, and $n_4 = 500$, respectively.

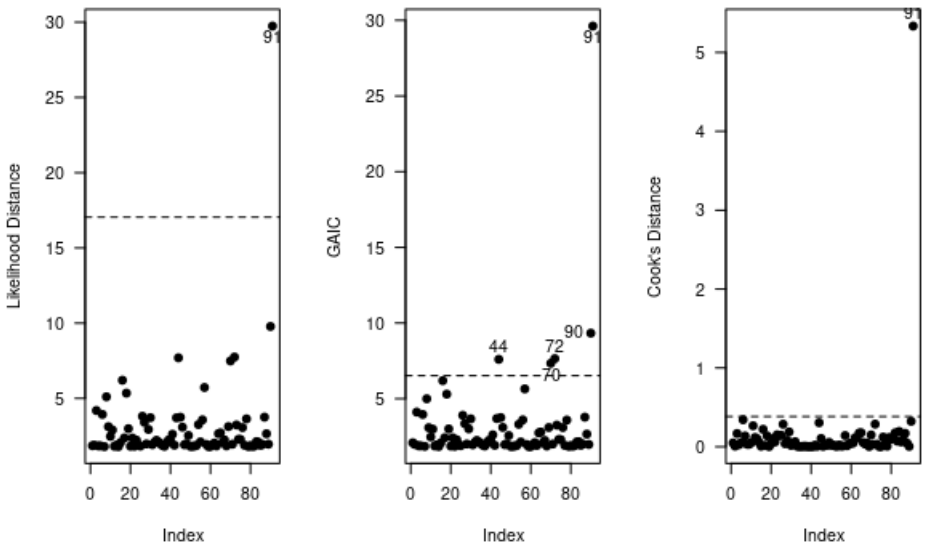
When n is large and a single influential case is included more points can be miss identified as influential, one possible solution can be choice a more conservative percentile value on the reference value algorithm when n is large.

Figure 12 – Scatter plot and fitted curves with and without the influential observation for the simulated data based on the functional form in (3.3), with sample size $n_1 = 90$ and $n_2 = 150$, $n_2 = 90$ and $n_3 = 150$, respectively.



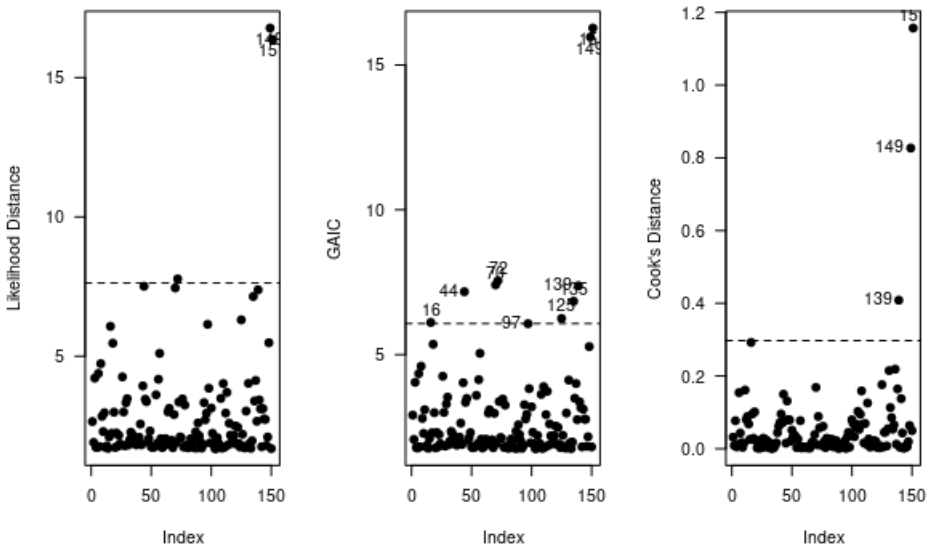
Source: Author's own.

Figure 13 – Likelihood Distance, Leave-One-Out GAIC and generalized Cook's distance for the simulated data based on the functional form in (3.3), with sample size $n_1 = 90$.



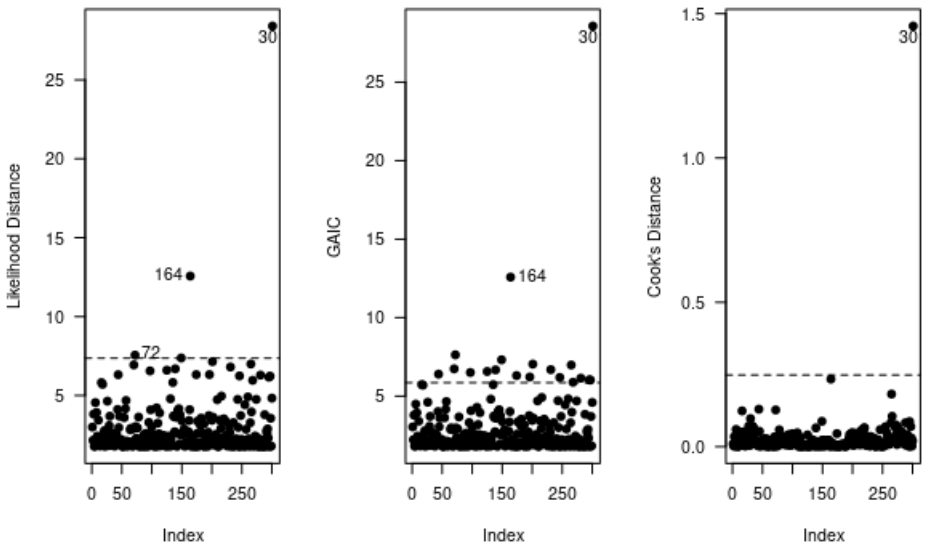
Source: Author's own.

Figure 14 – Likelihood Distance, Leave-One-Out GAIC and generalized Cook's distance for the simulated data based on the functional form in (3.3), with sample size $n_2 = 150$.



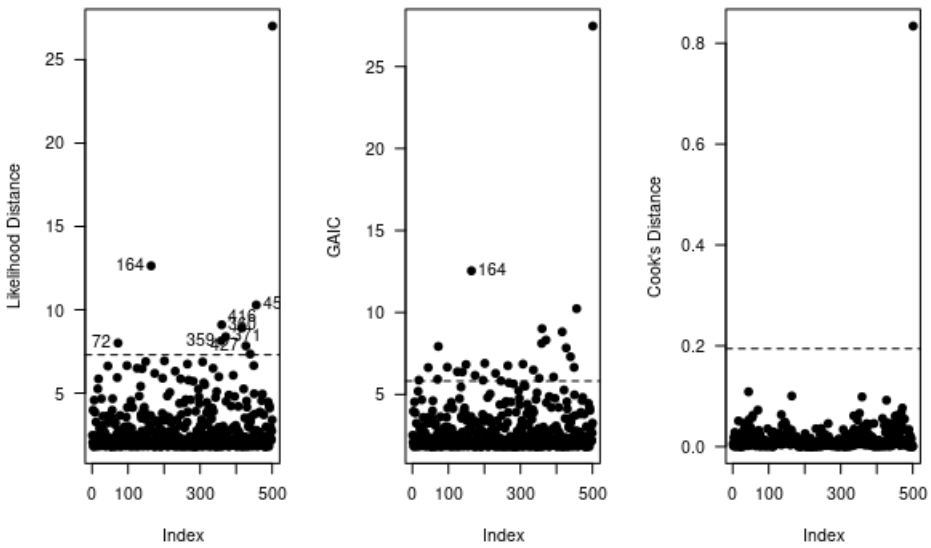
Source: Author's own.

Figure 15 – Likelihood Distance, Leave-One-Out GAIC and generalized Cook's distance for the simulated data based on the functional form in (3.3), with sample size $n_3 = 300$.



Source: Author's own.

Figure 16 – Likelihood Distance, Leave-One-Out GAIC and generalized Cook's distance for the simulated data based on the functional form in (3.3), with sample size $n_4 = 500$.



Source: Author's own.

3.3 SIMULATED DATA FOR A SEMIPARAMETRIC MODEL WITH POISSON RESPONSE

The Poisson distribution is widely used to model count data. If the response variable Y represents the number of occurrences of some event, the probability distribution can be written as:

$$f(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}, y = 0, 1, 2, \ldots,$$

where μ is the average number of occurrences. One important feature of this distribution is: the mean and the variance are equal, that is $\mathbb{E}(Y) = \text{Var}(Y) = \mu$.

For the Poisson regression model, consider $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, independent random variables with Y_i denoting the number of events observed in the exhibition number n_i. The model (3.4), is the GAMLSS submodel for the Poisson response for modeling the μ_i parameter, including the p-spline $s(u)$.

$$Y_i \overset{iid}{\sim} \text{PO}(\mu_i); \mathbb{E}(Y_i) = \mu_i = n_i e^{\boldsymbol{x_i}^\top \boldsymbol{\beta} + s(u)}, i = 1, \ldots, n; \tag{3.4}$$

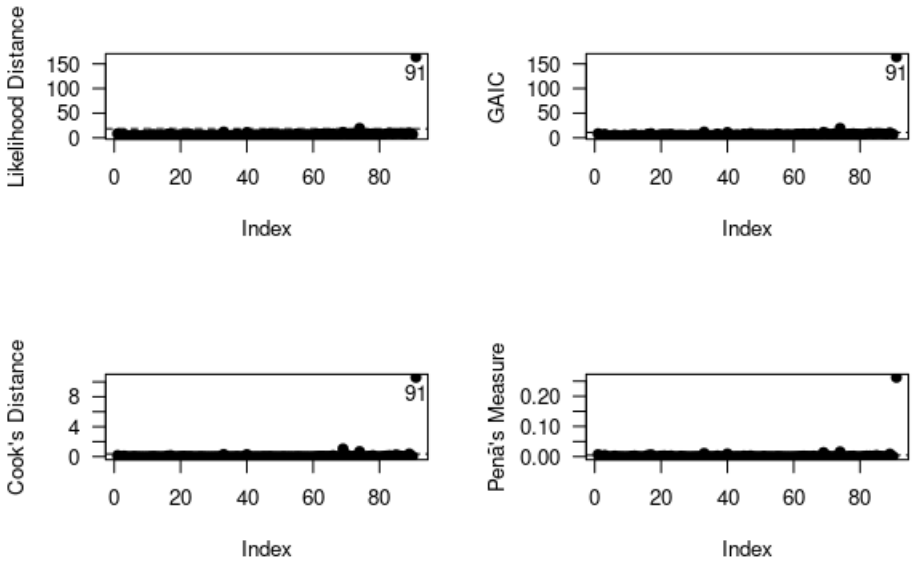
the natural link function is the logarithmic function, that is:

$$log(\mu_i) = log(n_i) + \boldsymbol{x}_i^\top \boldsymbol{\beta} + s(u),$$

in this case, for the simulation process, we take $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2)$ and, $\beta_0 = 2$, $\beta_1 = 3$ and $\beta_3 = -2$. Also, $\boldsymbol{x}$ is the fixed vector of co-variables, in this case $\boldsymbol{x}^\top = (\boldsymbol{x}_1, \boldsymbol{x}_2)^\top$ both are a sequence generated from a uniform fixing the seed in the generation process. We consider four different sample sizes $n_1 = 90, n_2 = 150, n_3 = 300$ and $n_4 = 500$. To artificially include a influential point we generate a outlier observation in the response variable based on the Tukey's method presented in (3.2).

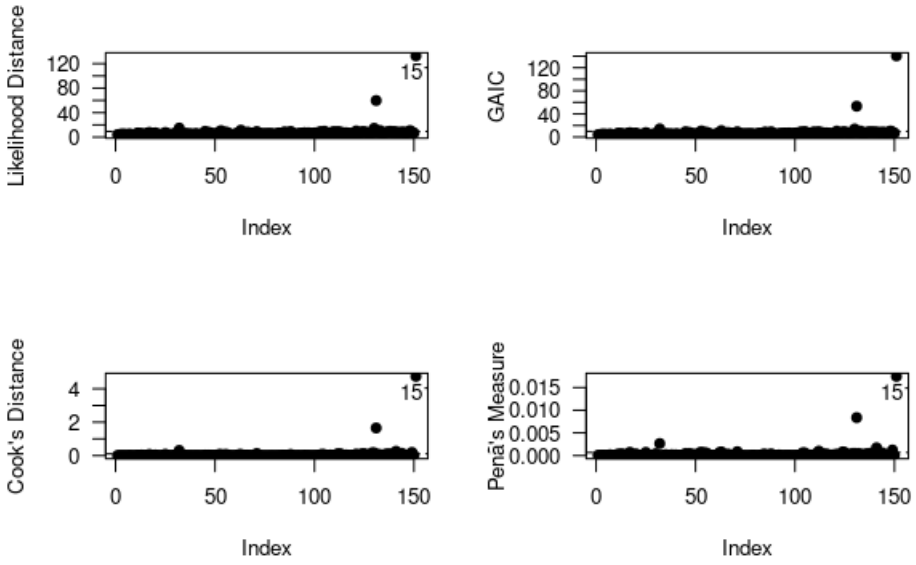
For all scenarios the influential observation was detected, some others observations was possible miss flagged as influential the results are show in the Figures 17, 18 and 19. We don't compare different measures itself, but the results of this different measures, in this particular case the figures 17, 18 and 19 provide similar conclusions for all scenarios.

Figure 17 – Likelihood distance, Leave-One-Out-GAIC, generalized Cook's distance and Peña's measure respectively, for the simulated data based on the functional form in (3.4), with sample size $n_1 = 90$.



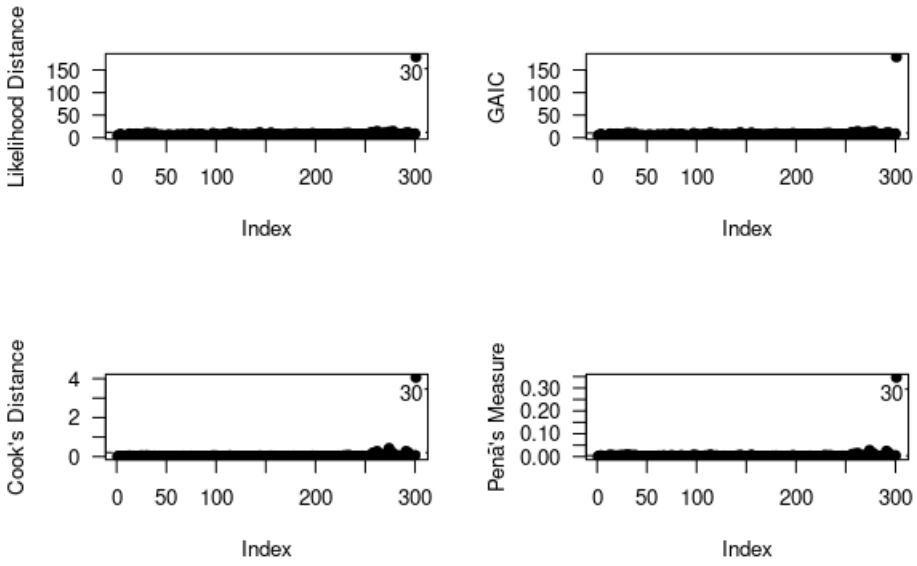
Source: Author's own.

Figure 18 – Likelihood distance, Leave-One-Out-GAIC, generalized Cook's distance and Peña's measure respectively, for the simulated data based on the functional form in (3.4), with sample size $n_2 = 150$.



Source: Author's own.

Figure 19 – Likelihood distance, Leave-One-Out-GAIC, generalized Cook's distance and Peña's measure respectively, for the simulated data based on the functional form in (3.4), with sample size $n_3 = 300$.



Source: Author's own.

3.4 SIMULATED DATA FOR A SEMIPARAMETRIC MODEL WITH GUMBEL RESPONSE

The Gumbel distribution (GUMBEL, 1948), is a continuous normally used to modeling rare events or extreme values situations. Let consider Y a random variable with Gumbel distribution, with location parameter μ and scale parameter σ, that is $Y \sim GU(\mu, \sigma)$. Thus, the probability density function is given by:

$$f(y; \mu, \sigma) = \frac{1}{\sigma}\exp\{-(z + \exp(-z)\},$$

where, $z = \dfrac{x - \mu}{\sigma}$. More details about the Gumbel distribution for the GAMLSS framework can be find in (RIGBY et al., 2019).

For the Gumbel regression model, consider $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, a vector of independent random variables, with Y_i. The canonical link function for the Gumbel response in the GAMLSS framework are the identity function. The Gumbel response semiparametrical model has the follow functional form

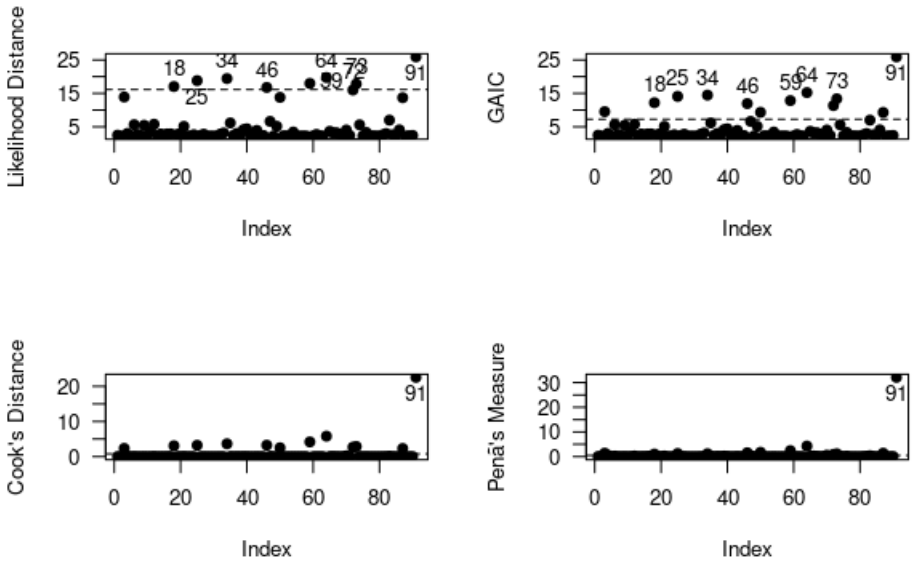
$$\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + s(u), \tag{3.5}$$

where, $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2)$ and, $\beta_0 = 2$, $\beta_1 = 3$ and $\beta_3 = -2$. Also, $\boldsymbol{x}$ is the fixed vector of co-variables, in this case $\boldsymbol{x}^\top = (\boldsymbol{x}_1, \boldsymbol{x}_2)^\top$ both are a sequence generated from a uniform fixing the seed in the generation process.

Simulating the four scenarios with $n_1 = 90, n_2 = 150, n_3 = 300$ and $n_4 = 500$. All measures detected the influential observation, but for the Gumbel, others observations are detected as influential as well. This can be associated to the fact of then Gumbel distribution generate extreme values, so in this cases this result is already expected, in this case we can use a bigger value of κ in the tukey criteria to force a higher extreme value. For all measures, the influential observation was the most influential on the sample.

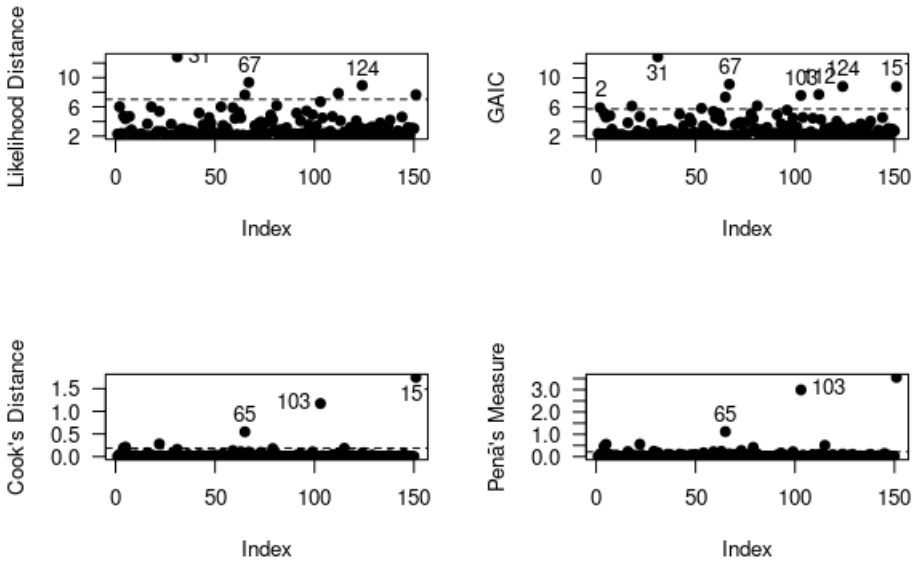
In this cases when n is large less observations are miss identified as influential, this indicate a better model fitting but the single influential case are so influential as the others.

Figure 20 – Likelihood distance, Leave-One-Out-GAIC, generalized Cook's distance and Peña's measure for the simulated data based on the functional form in (3.5), with sample size $n_1 = 90$.



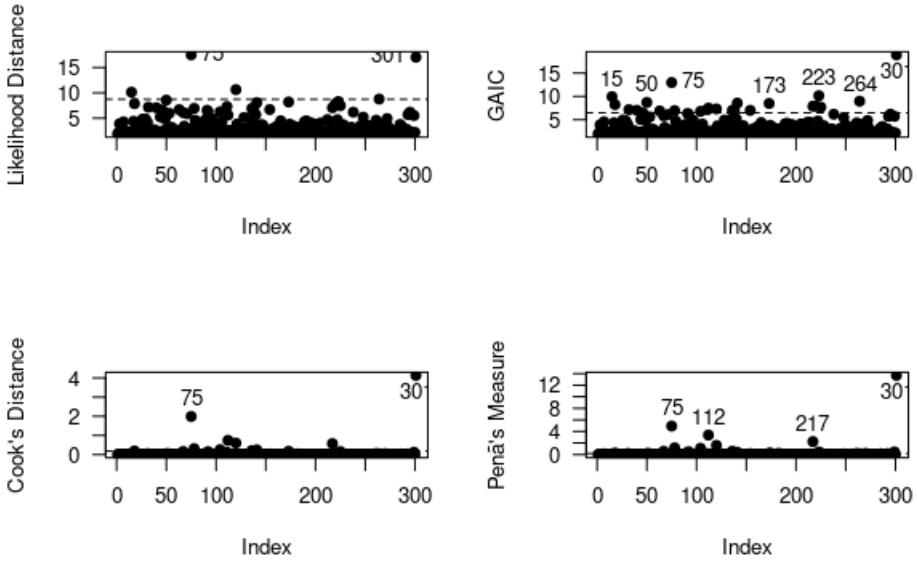
Source: Author's own.

Figure 21 – Likelihood distance, Leave-One-Out-GAIC, generalized Cook's distance and Peña's measure for the simulated data based on the functional form in (3.5), with sample size $n_2 = 150$.



Source: Author's own.

Figure 22 – Likelihood distance, Leave-One-Out-GAIC, generalized Cook's distance and Peña's measure for the simulated data based on the functional form in (3.5), with sample size $n_2 = 150$.



Source: Author's own.

4 APPLICATIONS

In this section we present some applications. Initially, we modeling a Diabetes data with a semiparametric model with gamma response, and use the Cook's distance, Likelihood Distance, Leave-One-Out GAIC, Adjusted Peña's measure, and Kim's measure to detect possible influential observations. Also, we fit a semiparametric model wih Negative Binomial Type-II for the Tidal data, the response is the counts of a mollusk find in New Zealand coast and compute the same measures.

4.1 DIABETES DATA

To illustrate the method described in the section 2.3, we consider the Diabetes data used in (KIM; PARK; KIM, 2002) taken from a study by (SOCHETT et al., 1987). The response variable is the logarithm of C-peptide blood concentration and the explanatory variables are the age $\boldsymbol{X}_1 = (x_{11}, \ldots, x_{1n})^\top$ and the base deficit $\boldsymbol{X}_2 = (x_{21}, \ldots, x_{2n})^\top$. For this example we consider $n = 41$ observations, the original data has $n = 43$, but (KIM; PARK; KIM, 2002) fitted a semiparametric model after removing the fifth (i=5) and the tenth (i=10) observation, and we follow the same line to have comparable results.

The data are show in the table 1, the response variable take positive real values, and we select the Gamma distribution for the GAMLSS model in this case with the GAIC automatic selection criteria in the **gamlss** R package. For each observation, Y_i is Gamma distributed with mean μ and $\phi^{-1/2}$, that is $Y_i \overset{ind}{\sim} GA(\mu_i, \phi)$, with probability density function given by:

$$f(y;\mu,\phi) = \frac{1}{\Gamma(\phi)}\left(\frac{\phi y}{\mu}\right)^{\phi} \exp\left(-\frac{\phi y}{\mu}\right) \mathrm{d}(\log y) =$$

$$= \exp[\phi\{-\left(\frac{y}{\mu} + \log\,\mu\right)\} - \log\Gamma(\phi) + \phi\log(\phi y) - \log\, y],$$

where, $y > 0$, $\phi > 0$, $\mu > 0$, $\Gamma(\phi) = \int_0^\infty t^{\phi-1}e^{-t}\mathrm{dt}$ is the gamma function. The canonical link function for the Gamma regression model is the logarithmic function:

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) = \beta_0 + \log(\boldsymbol{\mu}) = \boldsymbol{X}_1\beta_1 + s(\boldsymbol{X}_2), \tag{4.1}$$

in this case $s(\cdot)$ is the smoothing P-Spline. The fitted parameters are $\beta_0 = 1.50971$ and $\beta_1 = 0.01173$.

The Table 1 shown the data, and the respective i-th likelihood distance, generalized cook distance, Peña's measure, and Kim's measure (C_i, C_i^* and $\tilde{C}$) for each observation on sample. Also the Figure 27 present the histograms of the bootstrap values generated for compute the likelihood distance, Leave-One-Out GAIC and Cook's distances respectively.

The Figure 26 show off the index plot for the likelihood and Cook's distances, the horizontal line is the upper limit reference value calculated with 1000 bootstrap samples. For the likelihood distance, we computed the confidence interval $[-0.626935; 7.773088]$. That is, by the likelihood distance approach, the observation 20 are possible influential. Based on the Leave-One-Out GAIC, the computed reference value is 5.364664, so the observations 20 and 34 are possible influential by this approach.

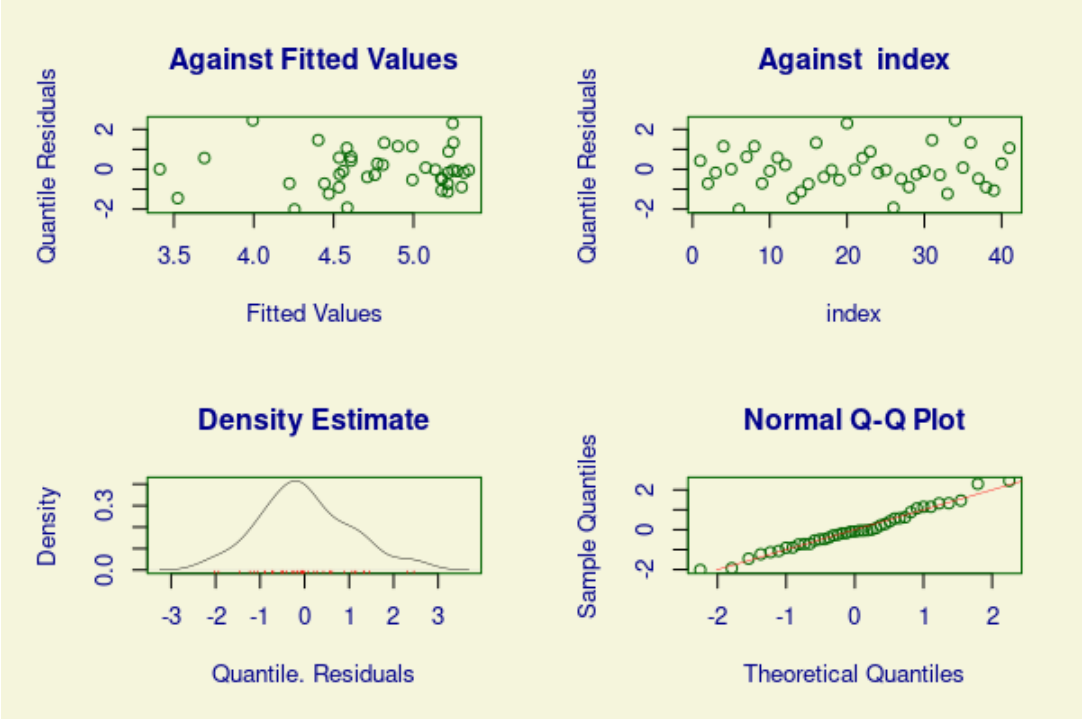
Based on the Cook's distance the computed reference value is 1.181677, so the observations 13 and 34 are possible influential, on the other hand. For the adjusted Peña's measure the obtained reference value is 0.0103735 and also flagged the observations 6, 13, 20, 26 and 34, as possible influential.

The Figure, 27 show the distribution of the bootstrap of likelihood distance, Leave-One-Out GAIC and Cook's distance used to obtain the references values.

Somehow, the results agree one with each other, the observations 6, 13, 20 and 34 was detected as influential in almost all measures so this four observations can be potentially influential. For a semiparameric model but using a local polynomial smoother (KIM; PARK; KIM, 2002), in a similar way detect the observations 6, 22 and 34 as possible influential.

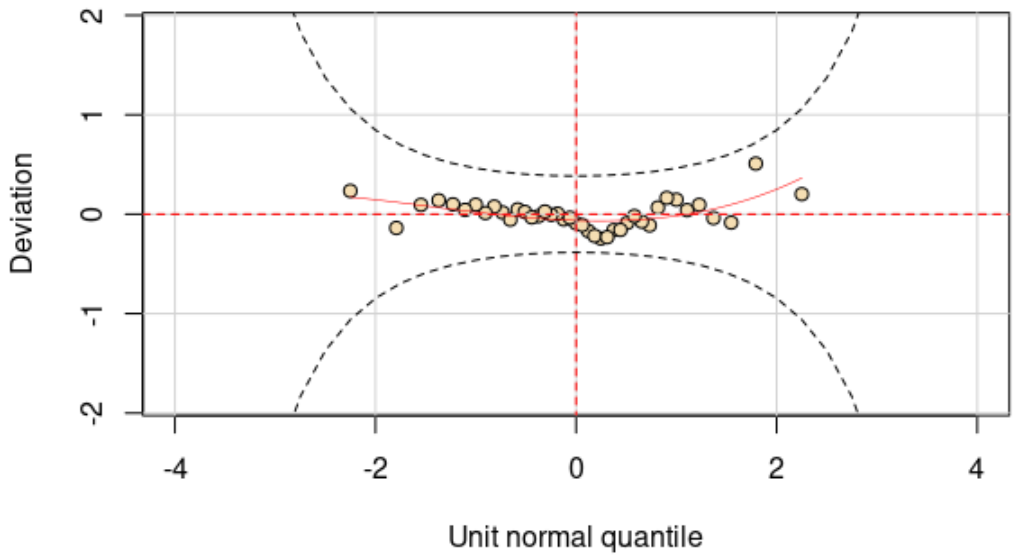
The results of this work can be comparable with the results of (KIM; PARK; KIM, 2002), but the influential observations can be different because the model is not the same and the others measures (likelihood distance, Leave-One-Out GAIC, Cook's distance and Peña's measure) can indicate different results for the influence.

Figure 23 – The residuals against fitted values, residuals against the index, kernel density estimate for the normalised residuals and Q-Q plot of the normalised residuals respectively for the model (4.1).



Source: Author's own.

Figure 24 – Worm-plot for the model (4.1).

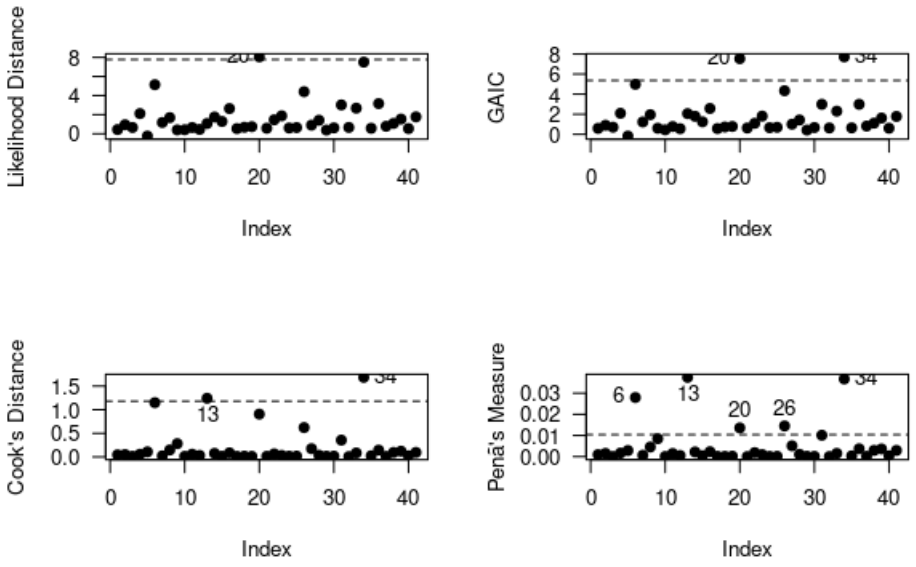


Source: Author's own.

Table 1 – Likelihood distances, generalized Cook's Distance, and Kim's measure for the diabetes data fitted with the model 4.1.

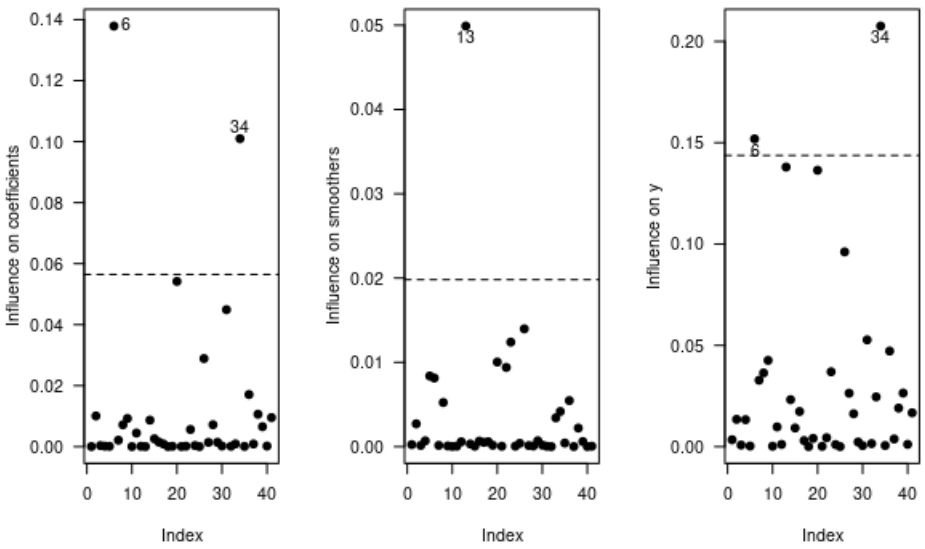
	$\boldsymbol{Y}$	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\mathsf{LD}_i(\boldsymbol{\mu})$	$\mathsf{CD}_i(\boldsymbol{\mu})$	P_i	C_i	C_i^*	$\tilde{C}_i$
1	4.80	5.20	−8.10	0.436248	0.045151	0.000930	0.000017	0.000238	0.003398
2	4.10	8.80	−16.10	0.940130	0.050683	0.001495	0.010046	0.002702	0.013452
3	5.20	10.50	−0.90	0.653117	0.014057	0.000050	0.000331	0.000140	0.000778
4	5.50	10.60	−7.80	2.109778	0.056517	0.001634	0.000140	0.000684	0.013304
5	3.40	1.80	−19.20	−0.247866	0.108730	0.002835	0.000087	0.008375	0.000288
6	3.40	12.70	−18.90	5.135500	1.150154	0.027619	0.137887	0.008133	0.151889
7	4.90	15.60	−10.60	1.177152	0.022439	0.000619	0.002092	0.000146	0.032787
8	5.60	5.80	−2.80	1.685129	0.148490	0.004482	0.007153	0.005231	0.036372
9	3.90	2.20	−3.10	0.398059	0.281553	0.008372	0.009214	0.000078	0.042580
10	4.50	4.80	−7.80	0.415537	0.012915	0.000053	0.000000	0.000022	0.000196
11	4.80	7.90	−13.90	0.647521	0.057503	0.001439	0.004453	0.000048	0.009728
12	4.90	5.20	−4.50	0.452536	0.031446	0.000506	0.000099	0.000571	0.001098
13	3.00	0.90	−11.60	1.069903	1.242706	0.037044	0.000011	0.049891	0.137972
14	4.60	11.80	−2.10	1.749776	0.075746	0.002179	0.008676	0.000313	0.023183
15	4.80	7.90	−2.00	1.283771	0.014627	0.000425	0.002558	0.000045	0.009206
16	5.50	11.50	−9.00	2.635045	0.086497	0.002294	0.001468	0.000651	0.017336
17	4.50	10.60	−11.20	0.535396	0.012748	0.000129	0.000856	0.000500	0.002988
18	5.30	8.50	−0.20	0.674441	0.016793	0.000118	0.000011	0.000576	0.000030
19	4.70	11.10	−6.10	0.738299	0.012737	0.000193	0.000093	0.000158	0.004083
20	6.60	12.80	−1.00	8.051237	0.907519	0.013416	0.054164	0.010030	0.136401
21	5.10	11.30	−3.60	0.575160	0.013472	0.000002	0.000023	0.000043	0.000148
22	3.90	1.00	−8.20	1.459061	0.065160	0.001855	0.000124	0.009394	0.004403
23	5.70	14.50	−0.50	1.873735	0.031933	0.000935	0.005613	0.012394	0.036941
24	5.10	11.90	−2.00	0.600435	0.016199	0.000102	0.000346	0.000049	0.001032
25	5.20	8.10	−1.60	0.643374	0.017317	0.000139	0.000003	0.000386	0.000020
26	3.70	13.80	−11.90	4.401796	0.622400	0.014324	0.028869	0.013970	0.096119
27	4.90	15.50	−0.70	0.898874	0.175165	0.005039	0.001375	0.000123	0.026367
28	4.80	9.80	−1.20	1.411375	0.035053	0.001041	0.007169	0.000072	0.016214
29	4.40	11.00	−14.30	0.380166	0.015802	0.000147	0.001403	0.000695	0.002280
30	5.20	12.40	−0.80	0.612800	0.017225	0.000112	0.000190	0.000177	0.000497
31	5.10	11.10	−16.80	3.015145	0.356155	0.009975	0.044883	0.000040	0.052697
32	4.60	5.10	−5.10	0.657765	0.007654	0.000022	0.000113	0.000000	0.001497
33	3.90	4.80	−9.50	2.676957	0.084021	0.001496	0.000878	0.003411	0.024548
34	5.10	4.20	−17.00	7.501751	1.686115	0.036300	0.100974	0.004164	0.207562
35	5.10	6.90	−3.30	0.571309	0.024787	0.000336	0.000060	0.000421	0.000561
36	6.00	13.20	−0.70	3.155435	0.144020	0.003653	0.017070	0.005462	0.047183
37	4.90	9.90	−3.30	0.780475	0.011351	0.000130	0.000857	0.000009	0.003777
38	4.10	12.50	−13.60	1.079980	0.095771	0.002904	0.010610	0.002198	0.019027
39	4.60	13.20	−1.90	1.532794	0.121837	0.003577	0.006525	0.000597	0.026441
40	4.90	8.90	−10.00	0.538494	0.024470	0.000362	0.000172	0.000005	0.001133
41	5.10	10.80	−13.50	1.765195	0.096247	0.002889	0.009515	0.000042	0.016703

Figure 25 – Index plot for the Cook's distance, Likelihood distance and Peña's measure for the model used for the diabetes data and fitted with the model (4.1), the horizontal line is the reference value computed with $B = 1000$ bootstraps resamples.



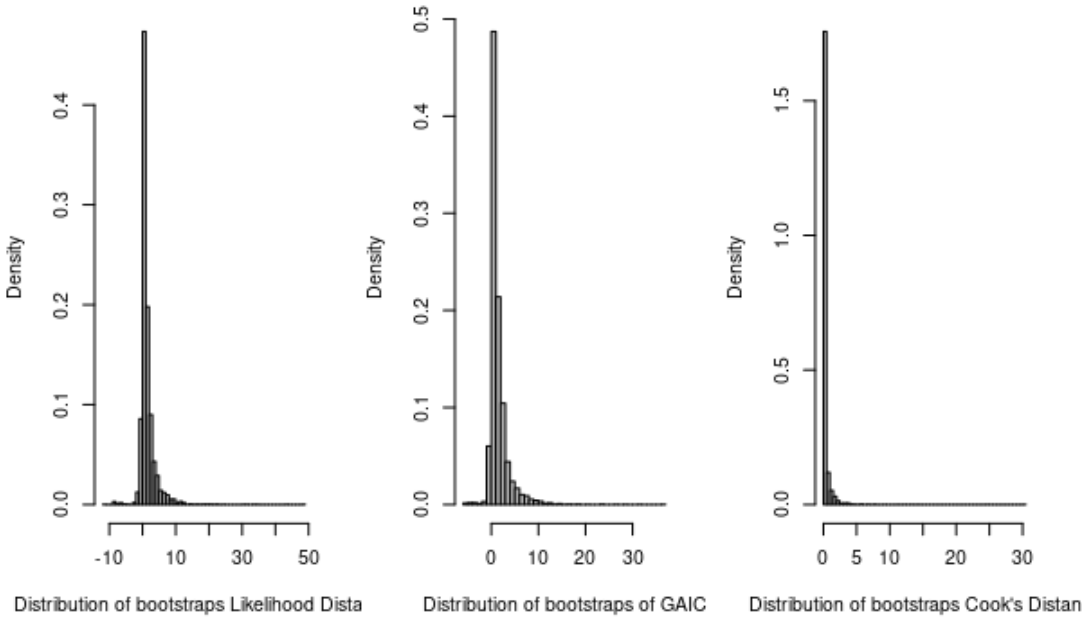
Source: Author's own.

Figure 26 – Index plot for the Kim's measure for the model used for the diabetes data and fitted with the model (4.1), the horizontal line is the reference value computed using the algorithm 4.



Source: Author's own.

Figure 27 – Histograms for the bootstrap likelihood and cook's distances for the diabetes data fitted with the model (4.1).



Source: Author's own.

4.2 TIDAL DATA

The organism *intertidal bivalve A. Stutchburyi* with common name *New Zealand cockle*, is a marine bivalve mollusk frequently often found in the New Zealand. (MCARDLE; ANDERSON, 2004) provided the data about the count in thee different coastal areas in the Bay of Planty, New Zealand. They also modelling the data with a particular transformation, in this work we use a Negative Binomial type II distribution.

The probability mass function of the Binomial Type II, is given by the

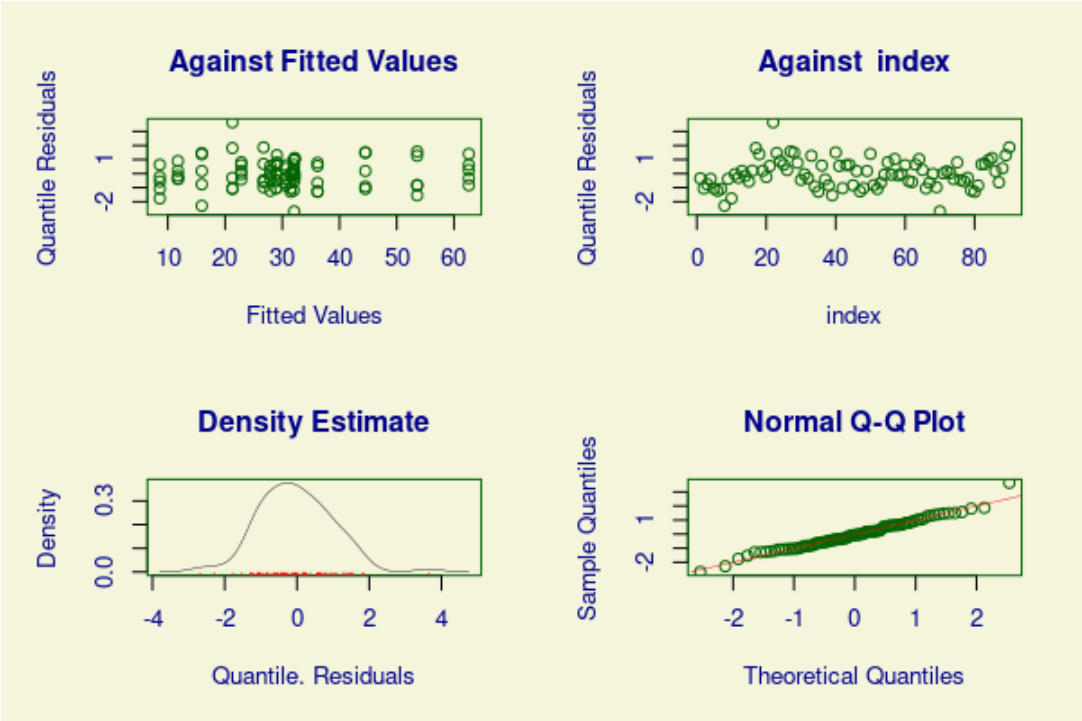
$$P(Y = y|\mu, \sigma) = \frac{\Gamma\left(y + \frac{\mu}{\sigma}\right)}{\Gamma\left(\frac{\mu}{\sigma}\right)\Gamma(y+1)}\sigma^y(1+\sigma)^{y+\frac{\mu}{\sigma}}. \tag{4.2}$$

This parameterization was used by (EVANS, 1953) and also by (JOHNSON; KEMP; KOTZ, 2005). For the Binomial Type II semiparametric regression model, the natural link function is the logarithmic function. The model in the GALMSS framework is:

$$Y_i \overset{iid}{\sim} \text{NBII}(\mu, \sigma)$$

$$\boldsymbol{\eta} = log(\boldsymbol{\mu}) = \beta_0 + \beta_1 \boldsymbol{x}_1 + s(\boldsymbol{u}), \tag{4.3}$$

Figure 28 – The residuals against fitted values, residuals against the index, kernel density estimate for the normalised residuals and Q-Q plot of the normalised residuals respectively for the model (4.3).



Source: Author's own.

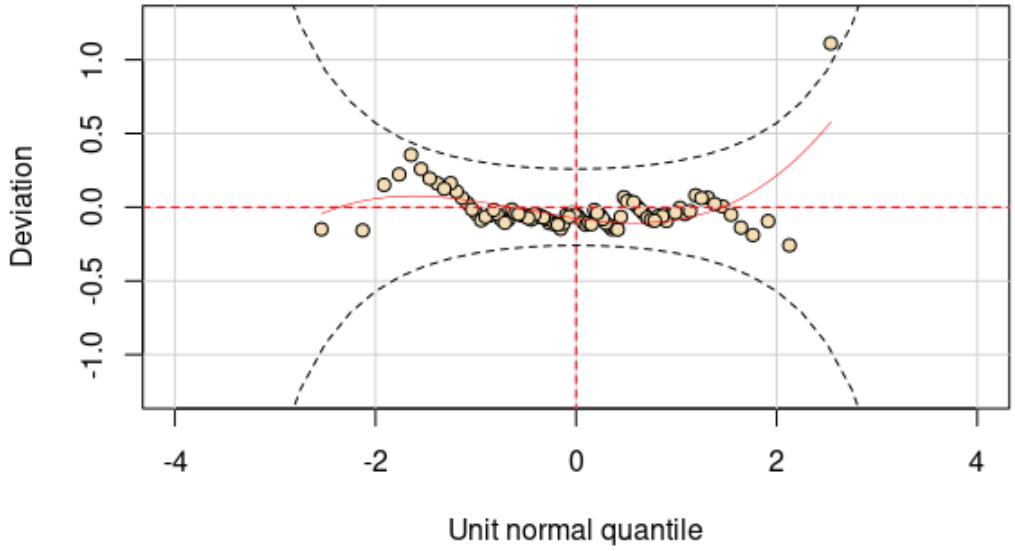
where the response variable is the count of *New Zealand cockles*, the covariable $\boldsymbol{u}$ is the vertical tidal height, s$(\cdot)$ is the p-spline smoother function, $\boldsymbol{x}_1$ is the factor indicating one of the tree tidal areas based on the tidal vertical height: lower (< 0.33m), middle ($0.33 - 0.66$m) and upper (> 0.66m). The ecologists are interested in the effect of tidal height (either raw or classified) on the number of organisms.

The estimated parameters for the model are $\beta_0 = 0.8585$ and $\beta_1 = -0.4541\mathbb{I}_{x_1=1}(x)$ and $\beta_2 = -1.4244\mathbb{I}_{x_1=2}(x)$, where $\mathbb{I}$ is the indicator function, defined as

$$\mathbb{I}_A : X \to \{0, 1\},$$

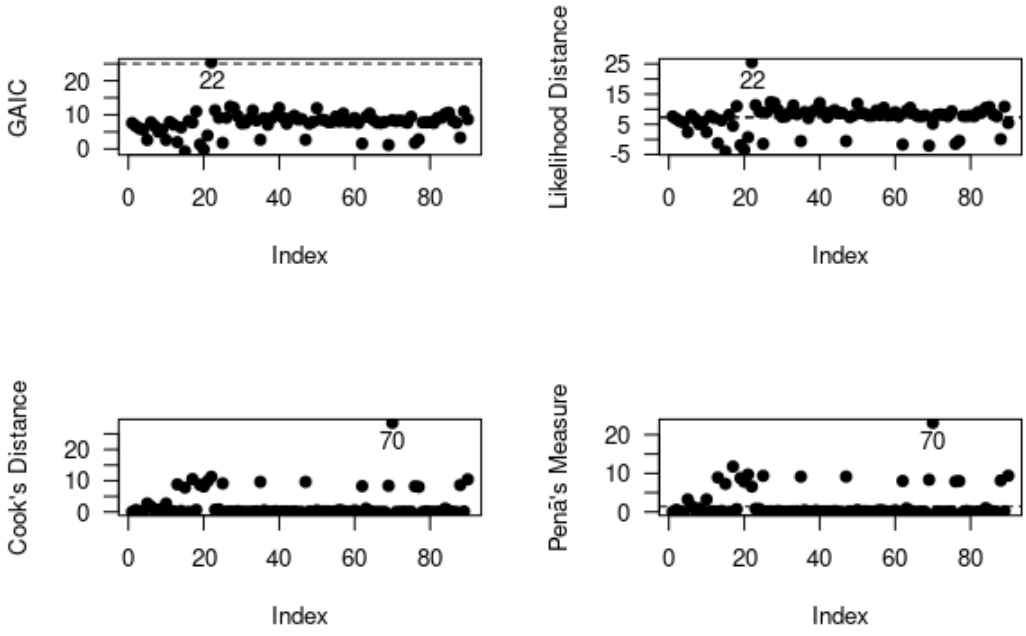
$$\mathbb{I}_A = \begin{cases} 1, \text{ if } x \in A, \\ 0, \text{ if } x \notin A \end{cases}$$

Figure 29 – Worm-plot for the model (4.3).



Source: Author's own.

Figure 30 – Index plot for the Cook's distance, Likelihood distance and Peña's measure for the model used for the diabetes data and fitted with the model (4.3), the horizontal line is the reference value computed using $B = 1000$ bootstraps resamples.

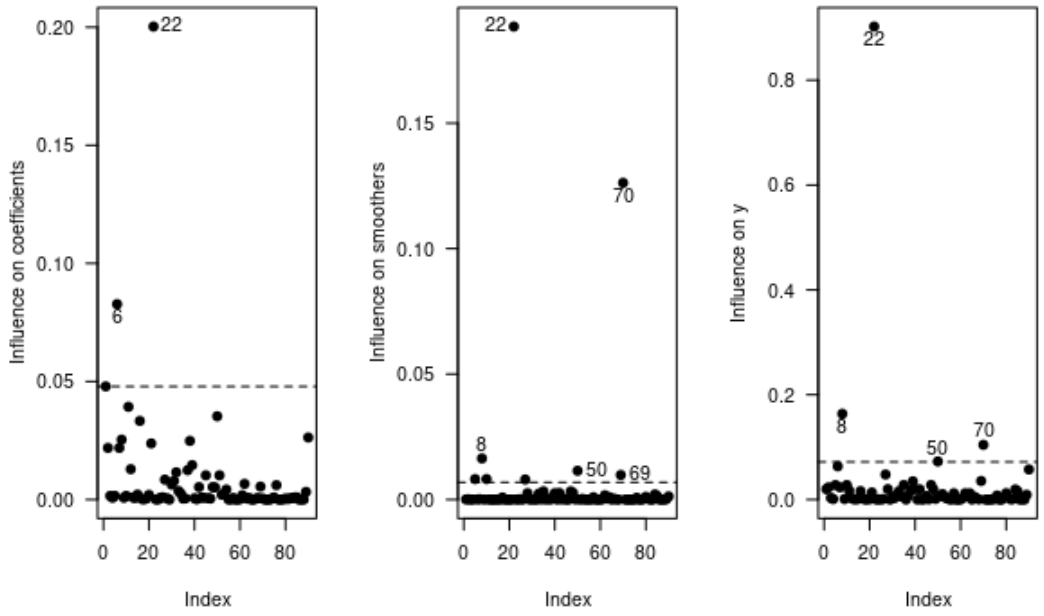


Source: Author's own.

Table 2 – The tidal data-set, where $\boldsymbol{y}$ is the number of organisms, $\boldsymbol{u}$ is the vertical tidal height in meters and $\boldsymbol{x}_1$ is the tidal area 1-lower, 2-middle and 3-upper.

i	$\boldsymbol{y}$	$\boldsymbol{u}$	$\boldsymbol{x}_1$	i	$\boldsymbol{y}$	$\boldsymbol{u}$	$\boldsymbol{x}_1$	i	$\boldsymbol{y}$	$\boldsymbol{u}$	$\boldsymbol{x}_1$
1	15	0.32	1	31	16	0.37	2	61	13	0.72	3
2	3	0.27	1	32	45	0.62	2	62	10	0.67	3
3	3	0.22	1	33	98	0.57	2	63	43	0.97	3
4	3	0.17	1	34	18	0.52	2	64	57	0.92	3
5	0	0.12	1	35	8	0.47	2	65	32	0.87	3
6	5	0.32	1	36	37	0.42	2	66	8	0.82	3
7	3	0.27	1	37	10	0.37	2	67	12	0.77	3
8	0	0.22	1	38	32	0.62	2	68	22	0.72	3
9	3	0.17	1	39	13	0.57	2	69	6	0.67	3
10	0	0.12	1	40	96	0.52	2	70	0	0.97	3
11	21	0.32	1	41	45	0.47	2	71	26	0.92	3
12	10	0.27	1	42	7	0.42	2	72	29	0.87	3
13	13	0.22	1	43	30	0.37	2	73	26	0.82	3
14	5	0.17	1	44	80	0.62	2	74	16	0.77	3
15	1	0.12	1	45	26	0.57	2	75	44	0.72	3
16	26	0.32	1	46	43	0.52	2	76	12	0.67	3
17	71	0.27	1	47	8	0.47	2	77	15	0.97	3
18	43	0.22	1	48	7	0.42	2	78	19	0.92	3
19	8	0.17	1	49	20	0.37	2	79	7	0.87	3
20	2	0.12	1	50	118	0.62	2	80	6	0.82	3
21	35	0.32	1	51	26	0.57	2	81	10	0.77	3
22	187	0.27	1	52	15	0.52	2	82	40	0.72	3
23	46	0.22	1	53	16	0.47	2	83	38	0.67	3
24	21	0.17	1	54	21	0.42	2	84	54	0.97	3
25	10	0.12	1	55	36	0.37	2	85	60	0.92	3
26	65	0.62	2	56	24	0.97	3	86	29	0.87	3
27	114	0.57	2	57	57	0.92	3	87	13	0.82	3
28	96	0.52	2	58	22	0.87	3	88	33	0.77	3
29	52	0.47	2	59	24	0.82	3	89	63	0.72	3
30	14	0.42	2	60	40	0.77	3	90	84	0.67	3

Figure 31 – Index plot for the Kim's measure for the model used for the diabetes data and fitted with the model (4.1), the horizontal line is the reference value computed using the algorithm 4.



Source: Author's own.

5 CONCLUDING REMARKS

In this work we using five measures to detect influential observations for the GAMLSS model: the Likelihood distance, Leave-One-Out GAIC, generalized Cook's distance, Peñas measure and Kim's measure. The Leave-One-Out GAIC is an new approach introduced in this work, usually the GAIC has the purpose to select the model, here we adjust the GAIC with the Leave-One-Out method for obtain a approach similar to the likelihood distance to detect influential observations.

The smoother matrix is not easy to estimate for a general smoother, and is necessary to compute the Kim's measure and Peña's measure. The smoother matrix estimation method changes when the smoother additive terms changes, that is for the penalized univariate splines we have a method to estimate the smoother matrix and for the local polynomial smoother we have other method to estimate the matrix. This hamper the computational method to calculate this measure.

One considerable contribution of this work is the methods to obtain the reference values, the bootstrap approach is used since the real distribution of the measures is unknown. However, this approach has a considerable computational disadvantage, we need to fit $n \times B$ models, in some situations some of this models can not reach convergence. Also, sometimes when n and B are huge, compute and store this models on the memory can be a problem, that is compute this references values sometimes need a considerable computational effort.

Using multiple measures to detect influential observation has a advantage to observe if the results agree one with other, sometimes the results can be very similar and sometimes the measures can detected different observations as influential.

6 FUTURE WORKS

Now we mention some future works that we have interest to extend the results obtained in this work.

There are two type of miss identification of influential observations, when the influential observation is not detected as influential and when a not influential observation is detected as possible influential. A future work can simulate several scenarios and compute a confusion matrix, that is understand this measures and reference values as a problem of classification. For this purpose its necessary a huge computational time to generate the samples and fit the models.

Furthermore, we can study in particular the Leave-One-Out GAIC to observe how the κ changes the detection of influence.

Other future work that we have interest is adjust this measures for models with parameters of scale and shape, that is models with σ, ν and τ. And perform simulations and applications as we done in this work.

At last, we have interest in apply this measures to detect influential observations with different smoother, that is by fixing the parametric terms on the model, observe what changes on the measures when the smothers terms changes, by using local polynomial, fractional polynomial, cubic splines, or neural networks.

REFERENCES

AITKIN, M. Modelling variance heterogeneity in normal regression using glim. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 36, n. 3, p. 332–339, 1987.

AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974.

BOX, G. E.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 26, n. 2, p. 211–243, 1964.

BUJA, A.; HASTIE, T.; TIBSHIRANI, R. Linear smoothers and additive models. *The Annals of Statistics*, JSTOR, p. 453–510, 1989.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine*, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992.

COOK, R. D. Detection of influential observation in linear regression. *Technometrics*, Taylor & Francis Group, v. 19, n. 1, p. 15–18, 1977.

COOK, R. D. Influential observations in linear regression. *Journal of the American Statistical Association*, Taylor & Francis, v. 74, n. 365, p. 169–174, 1979.

COOK, R. D. Assessment of local influence. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 48, n. 2, p. 133–155, 1986.

COOK, R. D.; WEISBERG, S. *Residuals and influence in regression*. [S.l.]: New York: Chapman and Hall, 1982.

DOBSON, A. J.; BARNETT, A. G. *An introduction to generalized linear models*. [S.l.]: CRC press, 2018.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996.

EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. [S.l.]: CRC press, 1994.

EILERS, P. H.; MARX, B. D. Flexible smoothing with b-splines and penalties. *Statistical science*, JSTOR, p. 89–102, 1996.

EVANS, D. Experimental evidence concerning contagious distributions in ecology. *Biometrika*, JSTOR, v. 40, n. 1/2, p. 186–211, 1953.

GIRAUD, G.; KOCKEROLS, T. Making the european banking union macro-economically resilient: Cost of non-europe report. *Report to the European Parliament*, 2015.

GUMBEL, E. J. *Statistical theory of extreme values and some practical applications: a series of lectures*. [S.l.]: US Government Printing Office, 1948. v. 33.

HARVEY, A. C. Estimating regression models with multiplicative heteroscedasticity. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 461–465, 1976.

HASTIE, T. J.; TIBSHIRANI, R. J. *Generalized additive models*. [S.l.]: CRC press, 1990. v. 43.

JOHNSON, N. L.; KEMP, A. W.; KOTZ, S. *Univariate discrete distributions*. [S.l.]: John Wiley & Sons, 2005. v. 444.

KIM, C.; PARK, B. U.; KIM, W. Influence diagnostics in semiparametric regression models. *Statistics & probability letters*, Elsevier, v. 60, n. 1, p. 49–58, 2002.

KIM, C.; STORER, B. E. Reference values for cook's distance. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 25, n. 3, p. 691–708, 1996.

LEE, Y.; NELDER, J. A.; PAWITAN, Y. *Generalized linear models with random effects: unified analysis via H-likelihood*. [S.l.]: CRC Press, 2018. v. 153.

MCARDLE, B. H.; ANDERSON, M. J. Variance heterogeneity, transformations, and models of species abundance: a cautionary tale. *Canadian Journal of Fisheries and Aquatic Sciences*, NRC Research Press, v. 61, n. 7, p. 1294–1302, 2004.

MONETARY, I. M. F.; DEPARTMENT, C. M. *United States: Financial Sector Assessment Program-Detailed Assessment of Observance on the Basel Core Principles for Effective Banking Supervision*. [S.l.]: International Monetary Fund, 2015.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2012. v. 821.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

PAIVA, C. S. M.; FREIRE, D. M. C.; CECATTI, J. G. Modelos aditivos generalizados para posição, escala e forma (gamlss) na modelagem de curvas de referência. *Rev. bras. ciênc. saúde*, v. 12, n. 3, p. 289–310, 2008.

PEK, J.; MACCALLUM, R. C. Sensitivity analysis in structural equation models: Cases and their influence. *Multivariate Behavioral Research*, Taylor & Francis, v. 46, n. 2, p. 202–228, 2011.

PEÑA, D. A new statistic for influence in linear regression. *Technometrics*, Taylor & Francis, v. 47, n. 1, p. 1–12, 2005.

PREGIBON, D. et al. Logistic regression diagnostics. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 9, n. 4, p. 705–724, 1981.

RAMIRES, T. G.; ORTEGA, E. M.; CORDEIRO, G. M.; PAULA, G. A.; HENS, N. New regression model with four regression structures and computational aspects. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 47, n. 7, p. 1940–1962, 2018.

RIGBY, R. A.; STASINOPOULOS, D. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, Springer, v. 6, n. 1, p. 57–65, 1996.

RIGBY, R. A.; STASINOPOULOS, D. M. Smooth centile curves for skew and kurtotic data modelled using the box–cox power exponential distribution. *Statistics in medicine*, Wiley Online Library, v. 23, n. 19, p. 3053–3076, 2004.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005.

RIGBY, R. A.; STASINOPOULOS, M. D. Mean and dispersion additive models. In: *Statistical theory and computational aspects of smoothing*. [S.l.]: Springer, 1996. p. 215–230.

RIGBY, R. A.; STASINOPOULOS, M. D.; HELLER, G. Z.; BASTIANI, F. D. *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. [S.l.]: CRC press, 2019.

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978.

SCOTT, R.; FORD, H.; HAUER, R.; YOUNG, S.; FANCHER, H.; VANGELIS. *Blade runner*. [S.l.]: Warner Home Video Los Angeles, 1982.

SENN, S. Francis galton and regression to the mean. *Significance*, Wiley Online Library, v. 8, n. 3, p. 124–126, 2011.

SMYTH, G. K. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 51, n. 1, p. 47–60, 1989.

SOCHETT, E.; DANEMAN, D.; CLARSON, C.; EHRLICH, R. Factors affecting and patterns of residual insulin secretion during the first year of type 1 (insulin-dependent) diabetes mellitus in children. *Diabetologia*, Springer, v. 30, n. 7, p. 453–459, 1987.

STASINOPOULOS, D. M.; RIGBY, R. A. et al. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, n. 7, p. 1–46, 2007.

STASINOPOULOS, M. D.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; BASTIANI, F. D. *Flexible regression and smoothing: using GAMLSS in R*. [S.l.]: CRC Press, 2017.

TUKEY, J. W. *Exploratory data analysis*. [S.l.]: Reading, MA, 1977. v. 2.

TÜRKAN, S.; TOKTAMIS, Ö. Detection of influential observations in semiparametric regression model. *Revista Colombiana de Estadística*, Universidad Nacional de Colombia., v. 36, n. 2, p. 271–284, 2013.

VERBYLA, A. P. Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 55, n. 2, p. 493–508, 1993.

WOOD, S. N. *Generalized additive models: an introduction with R*. [S.l.]: CRC press, 2017.

YEO, I.-K.; JOHNSON, R. A. A new family of power transformations to improve normality or symmetry. *Biometrika*, Oxford University Press, v. 87, n. 4, p. 954–959, 2000.