Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Estatística

SADRAQUE ENEAS DE FIGUEIREDO LUCENA

ESSAYS ON NONNORMAL REGRESSION MODELING

Recife
2017

SADRAQUE ENEAS DE FIGUEIREDO LUCENA

ESSAYS ON NONNORMAL REGRESSION MODELING

Doctoral thesis submitted to the Programa de Pós-Graduação em Estatística, Departamento de Estatística, Universidade Federal de Pernambuco as a partial requirement for obtaining a Ph.D. in Statistics.

Advisor: Prof. Ph.D. Francisco Cribari Neto

Recife
2017

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

L935e Lucena, Sadraque Eneas de Figueiredo
Essays on nonnormal regression modeling / Sadraque Eneas de Figueiredo Lucena. – 2017.
116 f.: il., fig., tab.

Orientador: Francisco Cribari Neto.
Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2017.
Inclui referências e apêndices.

1. Análise de regressão. 2. Regressão simplex. I. Cribari Neto, Francisco (orientador). II. Título.

519.536 CDD (23. ed.) UFPE- MEI 2017-57

SADRAQUE ENEAS DE FIGUEIREDO LUCENA

ESSAYS ON NONNORMAL REGRESSION MODELING

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

Aprovada em: 10 de fevereiro de 2017.

BANCA EXAMINADORA

Prof. Francisco Cribari Neto
UFPE

Prof. Francisco José de Azevêdo Cysneiros
UFPE

Prof. Raydonal Ospina Martínez
UFPE

Prof.ª Luz Milena Zea Fernández
UFRN

Prof. Aluísio de Souza Pinheiro
Unicamp

To my parents, Severino and Dorcas,

and to my sister, Stefanny.

Acknowledgements

First and foremost I have to thank my advisor, Francisco Cribari Neto, for his invaluable guidance of this research. Without his encouragement, support, patience and critical reviews this work would never be done. I am grateful not only for all his teachings, but for being an exemplary person in several aspects.

I would like thank all professors at UFPE for their teachings and incentive, especially Prof. Ph.D. Raydonal Ospina Martínez, Prof. Ph.D. Francisco José de Azevêdo Cysneiros, Prof. Ph.D. Gauss Moutinho Cordeiro, Prof. Ph.D. Klaus Leite Pinto Vasconcellos. I also thank the examiners for their contribution and suggestions to this work.

I thank for all the staff members, especially to Valéria, for her dedication and efficiency.

I also thank CAPES for financial support.

My gratitude to Ana Hermínia, for being with me during all this journey since undergraduate. For sharing all goods and bads and for being supportive whenever I needed.

I thank my UFPE classmates and my colleagues from UFPB, mainly my friends Marcelo, Juliana, Hemílio, Maria Lídia, Gilmara, Suely and José Carlos. You gave me an unlimited support. I would also like to thank my new colleagues at UFS, mostly to Eucymara, Rodrigo, Marcelo, Raphael and Vanessa for helping me in my adapting to Aracaju while writing this thesis.

Thanks also to all my friends, Joenio Oliveira, Leyde Klebia, Jobson Francisco Jr., Edjackson Ferreira, Márcia Albuquerque, Mônica Albuquerque, Joel Tavares, José Augusto, Nicole Gomes, Ligia Stansky, Lucio Ismael, Camila Spinelli, Carolina Pereira, Sabrinny Lima, Allyne Lacerda, Pedro Oliveira, Alexssandra Ribeiro, Thiago Almeida. You know how important you are to me.

I have to thank my parents (Severino and Dorcas), my sister (Stefanny) from the bottom of my heart for their unceasing support and love throughout my life.

Many people have contributed to my personal and professional growth while I was writing this thesis. I have to say my warmest thanks to all of them.

Phew! I have finished this thesis!

Abstract

In regression analysis a wide range of techniques can be used to investigate the relationship between the response and the regressors. In some situations, two or more competing models may fit the data equally well. When none of them can be obtained from the others by imposing parametric restrictions, we say the models are nonnested. In order to choose between competing nonnested linear regression models, one can use the J and MJ tests. In this PhD thesis we present an adaptation of such tests to nonnested models in the class of generalized additive models for location, scale and shape (GAMLSS). Monte Carlo evidence on the finite sample behaviour of the proposed tests and an application are reported. We also develop a frequentist approach to the augmented simplex regression model proposed by Bandyopadhyay, Galvis and Lachos [Bandyopadhyay, D., Galvis, D. M. & Lachos, V. H. (2014), 'Augmented mixed models for clustered proportion data', Statistical Methods in Medical Research (In Press)]. It can be used when the response assumes values in [0,1), (0,1] or [0,1] and we call it zero and/or one inflated simplex regression model. Inference, diagnostics measures and an application are also reported.

Keywords: Nonnested models. GAMLSS. J and MJ tests. Zero and/or one inflated simplex regression model. Diagnostic measures.

Resumo

Na modelagem de dados por meio de regressão, há uma ampla variedade modelos que podem ser ajustados para avaliar a relação entre a variável resposta e os regressores. Em algumas situações, a modelagem pode envolver dois ou mais modelos com ajustes semelhantes, embora com especificações distintas. Quando nenhum dos modelos ajustados pode ser obtido por meio de restrições paramétricas impostas aos outros modelos, dizemos que eles são não-encaixados. Dois possíveis métodos para selecionar o mais adequado entre modelos lineares não-encaixados são os testes J e MJ. Nesta tese é apresentada uma adaptação desses testes para a classe de modelos denominada *generalized additive models for location, scale and shape* (GAMLSS). Evidências obtidas a partir de simulações de Monte Carlo em pequenas amostras e uma aplicação são reportadas. Também é apresentada uma abordagem paramétrica para o modelo de regressão simplex aumentado. Este modelo pode ser ajustado nos casos em que a variável resposta assume valores nos intervalos [0,1), (0,1] ou [0,1]. Aqui o modelo é chamado de modelo de regressão simplex inflacionado em zero e/ou um. Inferência, medidas de diagnóstico e uma aplicação também são apresentados.

Palavras-chave: Modelos não-encaixados. GAMLSS. Testes J e MJ. Regressão simplex; Regressão simplex inflacionada em zero e/ou um. Medidas de diagnóstico.

List of Figures

List of Tables

Contents

CHAPTER 1

Nonnested hypothesis testing inference for GAMLSS models

Resumo

É comum em situações práticas a obtenção de dois ou mais modelos com ajustes semelhantes, mas com especificações distintas. Se qualquer um dos modelos não puder ser obtido a partir de restrições impostas sobre os parâmetros que os indexam, dizemos que estes são não-encaixados. Dois testes usados para avaliar qual dos modelos está corretamente especificado em modelos lineares de regressão são os testes J e MJ. Neste capítulo propomos variantes desses dois testes para a classe de modelos GAMLSS (*Generalized Additive Models for Location, Scale and Shape*). São reportadas evidências de Monte Carlo para avaliar o comportamento dos testes propostos em amostras finitas. De modo geral, a variante *bootstrap* do teste MJ apresentou melhor performance. Uma aplicação empírica da modelagem da renda líquida mensal na cidade de Munique é também apresentada e discutida.

1.1 Introduction

Regression theory received the first rigorous mathematical treatment with the development of the linear regression model by Pearson (1896). Since then several new models and inferential procedures were developed, such as the class of generalized linear models (GLMs) (Nelder & Wedderburn 1972), generalized additive models (GAMs) (Hastie & Tibshirani 1986, 1990) and beta regression models (Ferrari & Cribari-Neto 2004). Such classes of models were designed to cope with violations of some of the standard assumptions of the linear regression model (e.g., non-normal errors, heteroskedasticity and response variable restricted to a subinterval of the real line).

A class of models which has received attention in the past few years is the class of generalized additive models for location, scale and shape (GAMLSS). It was introduced by Rigby & Stasinopoulos (2001, 2005) and Akantziliotou et al. (2002) to overcome limitations of the GLMs and GAMs. In GAMLSS the response variable (Y) is assumed to belong to a distribution family which includes highly skewed and/or kurtotic continuous and discrete distributions.

In GAMLSS a general distribution family $\mathcal{D}$ is assumed for the response variable (Y). In general, the response distribution has four parameters and is denoted by $Y \sim \mathcal{D}(\mu, \sigma, \nu, \tau)$. The parameters μ and σ are usually the location and the scale parameters, whereas ν and τ are the shape parameters (e.g., skewness and kurtosis parameters). In GAMLSS, a submodel is estimated for each parameter related to a linear predictor through a link function. The submodel can be linear or nonlinear, parametric or semi-parametric (see Section 1.2). Hence, not only the location but other parameters of the distribution of Y can be modeled.

Once the model is estimated, practitioners usually carry out hypothesis testing inference to evaluate whether the data is well represented by the estimated model. However, it is not uncommon to have at disposal more than one model with different parametric (or semi-parametric) structures that fits the data equally well. In situations where no model can be obtained from the others by imposing parametric restrictions the models

are said to be nonnested. When that happens the usual asymptotic tests, such as the likelihood ratio (LR), Wald and score tests, cannot be applied to choose the best model (Cribari-Neto & Lucena 2015).

The literature on nonnested models has its origins in papers written by Sir David Cox (Cox 1961, 1962), which were revisited by Cox (2013). Three general approaches may be used for nonnested competing models (Pesaran & Weeks 2001): modified LR procedure or Cox test; comprehensive models, advocated by Atkinson (1970) and used by Davidson & MacKinnon (1981); and the encompassing procedure, developed by Deaton (1982) and Dastoor (1983) and extended by Gouriéroux et al. (1983) and Mizon & Richard (1986).

The most frequently applied nonnested hypothesis test in statistics and econometrics is the J test (Godfrey 2011, McAleer 1995), which was introduced by Davidson & MacKinnon (1981). Some extensions and modifications of the test were proposed since then. Wooldridge (1990) modified the test to make it robust under heterogeneity of unknown form. Michelis (1999) obtained the asymptotic null distribution of J test statistic for regression models with almost orthogonal nonnested regressors. Davidson & MacKinnon (2002) analyzed the finite-sample distribution of the J test statistic. Sapra (2008) proposed a modification on J test with superior finite sample performance when there are outliers in the data. Kelejian (2008), Kelejian & Piras (2014), Burridge & Fingleton (2010) and Piras & Lozano-Garcia (2012) studied the finite sample performance of J test using spatial data. Ghali et al. (2011) came up with a Bayesian variant of J test. Ramalho et al. (2011) evaluated the finite sample performance of nonnested hypothesis tests for choosing the link function of binary response models. Hagemann (2012) proposed a modification of the J test that avoids sequential testing and ambiguous outcomes: the MJ test. Cribari-Neto & Lucena (2015) adapted the J and MJ tests for beta regression models.

In GAMLSS, researchers may have at disposal two or more competing models with different regressors, link functions, or both of them. The models can also differ in the way nonparametric components are used. These differences may take place in one or more

submodels. Notice that the J and MJ tests as defined in literature are not suitable for these situations because they have been developed for linear regressions. The chief goal of this chapter is to present variants of J and MJ tests for GAMLSS models and evaluate their finite sample performance using Monte Carlo simulations. The tests were performed in the simulations for competing GAMLSS models that differ in the regressors and link functions. We also considered bootstrap-based testing inference.

This chapter is organized as follows. Section 1.2 introduces the GAMLSS. The J and MJ tests and their GAMLSS variants are presented in Section 1.3. Numerical results are presented and discussed in Section 1.4. In Section 1.5, we present and discuss an application using real (not simulated) data. Finally, Section 1.6 contains some concluding remarks.

1.2 The GAMLSS models

Let $y = (y_1, y_2, \ldots, y_n)^\top$ be a set of independent observations y_i, $i = 1, 2, \ldots, n$, each one with distribution $D(\boldsymbol{\theta}_i)$, where $\boldsymbol{\theta_i} = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i})^\top = (\mu_i, \sigma_i, \tau_i, \nu_i)^\top$ is a vector of four distribution parameters. D is a probability distribution that can be discrete, continuous or mixed. Each y_i has probability (density) function (p.d.f.) $f(y_i|\boldsymbol{\theta}_i)$. The parameters μ_i and σ_i are generally interpreted as the location and scale parameters whereas ν_i and τ_i are shape parameters (for example, skewness and kurtosis parameters). In a GAMLSS model each distribution parameter can be written as a function of regressors.

Let $g_k(\cdot)$, $k = 1, 2, 3, 4$, be known, strictly increasing and twice differentiable function. In the GAMLSS class of models introduced by Rigby & Stasinopoulos (2005), the distribution parameters are related to the regressors by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{t=1}^{T_k} \mathbf{Z}_{tk}\boldsymbol{\gamma}_{tk}, \tag{1.1}$$

that is,

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{t=1}^{T_1} \mathbf{Z}_{t1}\boldsymbol{\gamma}_{t1},$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{t=1}^{T_2} \mathbf{Z}_{t2}\boldsymbol{\gamma}_{t2},$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{t=1}^{T_3} \mathbf{Z}_{t3}\boldsymbol{\gamma}_{t3},$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{t=1}^{T_4} \mathbf{Z}_{t4}\boldsymbol{\gamma}_{t4},$$

where $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ and $\boldsymbol{\eta}_k$ are vectors of length n, $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{T_k k})$ is a vector of parameters of length T_k, $\mathbf{X}_k$ and $\mathbf{Z}_{tk}$ are known fixed matrices of regressors of order $n \times T_k$ and $n \times q_{tk}$, respectively. $\boldsymbol{\gamma}_{tk}$ is a q_{tk}-dimensional random variable with distribution $\gamma_{tk} \sim N_{q_{tk}}(\mathbf{0}, \mathbf{G}_{tk}^{-1})$, where $\mathbf{G}_{tk}^{-1}$ is a (generalized) inverse of a $q_{tk} \times q_{tk}$ symmetric matrix $\mathbf{G}_{tk} = \mathbf{G}_{tk}(\boldsymbol{\lambda}_{tk})$. $\mathbf{G}_{tk}$ may depend on a vector of hyperparameters $\boldsymbol{\lambda}_{tk}$. If $\mathbf{G}_{tk}$ is singular then $\boldsymbol{\gamma}_{tk}$ have improper prior density function proportional to $\exp(-\frac{1}{2}\boldsymbol{\gamma}_{tk}^{\top}\mathbf{G}_{tk}\boldsymbol{\gamma}_{tk})$, if $\mathbf{G}_{tk}$ is not singular then γ_{tk} follows a q_{tk}-dimensional multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{G}_{tk}^{-1}$. $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_{tk}$ are called, respectively, parametric vectors and random effects parameters.

The GAMLSS class contains important submodels. When $T_k = 0$ in Equation (1.1), the model is fully parametric, called *simple parametric linear* GAMLSS, and given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k. \tag{1.2}$$

Notice that $\boldsymbol{\theta}_k$, for $k = 1, 2, 3, 4$, contains the distribution parameter vectors $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$. If $\mathbf{Z}_{tk} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix, and $\boldsymbol{\gamma}_{tk} = \mathbf{h}_{tk} = h_{tk}(\mathbf{x}_{tk})$ for all combinations of t and k, model (1.1) becomes

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{t=1}^{T_k} h_{tk}(\mathbf{x}_{tk}), \tag{1.3}$$

where $\mathbf{x}_{tk}$ $(t = 1, 2, \ldots, T_k)$ is an n-vector. Here, h_{tk} is an unknown function of X_{tk} and $\mathbf{h}_{tk} = h_{tk}(\mathbf{x}_{tk})$ is the vector that contains h_{tk} evaluated at $\mathbf{x}_{tk}$. Model (1.3) is called *linear semi-parametric additive* GAMLSS.

Model (1.3) may be extended to include nonlinear parametric terms in each submodel. It is expressed by (Rigby & Stasinopoulos 2006)

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) + \sum_{t=1}^{T_k} h_{tk}(\mathbf{x}_{tk}), \tag{1.4}$$

where h_k, $k = 1, 2, 3, 4$, is a nonlinear function and $\mathbf{X}_k$ is a matrix of known regressors of order $n \times T_k$. Model (1.4) is called *nonlinear semi-parametric additive* GAMLSS. If $T_k = 0$ for all $k = 1, 2, 3, 4$, i.e., the submodels do not include additive terms, Model (1.4) reduces to a *nonlinear parametric* GAMLSS model, expressed by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k). \tag{1.5}$$

Further, if $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) = \mathbf{X}_k^\top \boldsymbol{\beta}_k$, for $k = 1, 2, 3, 4$ and $i = 1, 2, \ldots, n$, Equation (1.5) becomes the linear parametric model presented in (1.2). Any combination of Models (1.2) and (1.5) is called *parametric* GAMLSS model. The models above can be extended for p $(p > 4)$ parameters, that is., $k = 1, 2, \ldots, p$.

The R software (R Development Core Team 2011) has packages available to deal with GAMLSS models. The additive functions h_{tk} admitted in the GAMLSS packages are cubic splines, penalized splines, fractional polynomials, power polynomials, loess curves, varying coefficient terms, among others (Stasinopoulos & Rigby 2008).

It is noteworthy that not necessarily all the distribution parameters need to be modeled using regressors. Additionally, Rigby & Stasinopoulos (2005) emphasize that the GAMLSS is more general than GLM, GAM or GAMM, since the response variable is not restricted to the exponential family distribution and all parameters may be modeled in terms of fixed and random effects.

The form of the distribution which can be assumed for the response variable may be

very general. In the GAMLSS package for the R software there are several implemented continuous and discrete distributions. The implemented distributions are indexed by one parameter (such as the exponential and Poisson distributions), two parameters (such as the beta, Weibull and beta-binomial distributions), three parameters (e.g., generalized gamma, reverse Gumbel and Sichel distributions) or four parameters (e.g., the Box-Cox t (BCT) and generalized beta type 1 distributions) (Rigby et al. 2014).

1.2.1 Estimation

Estimation of $\boldsymbol{\beta}_k$, $k = 1, 2, 3, 4$, and of the random effects parameters $\boldsymbol{\gamma}_{tk}$, $t = 1, 2, \ldots, T_k$, is performed by maximizing the penalized likelihood function

$$\ell_{pen} = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{t=1}^{T_k}\lambda_{tk}\boldsymbol{\gamma}_{tk}^{\top}\mathbf{G}_{tk}\boldsymbol{\gamma}_{tk}, \tag{1.6}$$

for fixed values of the smoothing hyper-parameters λ_{tk}'s. In Equation (1.6), the term $\ell = \sum_{i=1}^{n}\log[f(y_i|\boldsymbol{\theta}_i)]$ is the log-likelihood function. More details on how the penalized log-likelihood ℓ_{pen} is maximized can be found in Stasinopoulos & Rigby (2008). Note that for parametric GAMLSS models, that is, models (1.2) and (1.5), ℓ_{pen} reduces to ℓ, and $\boldsymbol{\beta}_k$ is estimated by maximizing the log-likelihood function ℓ.

Two algorithms may be used in R in order to maximize (1.6): the CG and RS algorithms. The former is a generalization of the Cole & Green (1992) algorithm. It uses the first and the second (expected or approximated) order and cross derivatives of the likelihood function with respect to the distribution parameters $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)^{\top}$. For parametric GAMLSS models the CG algorithm is equivalent to Fisher's scoring method. If the parameters are orthogonal the cross derivatives of the log-likelihood function are 0. In this case, the RS algorithm is more suited. It is a generalization of the algorithm used by Rigby & Stasinopoulos (1996*a*, 1996*b*) for adjusting mean and dispersion additive models (MADAM), which does not uses cross derivatives. It is worth pointing out that the RS algorithm is not a particular case of the CG algorithm. Further details on both algorithms

can be found in Rigby & Stasinopoulos (2005).

1.2.2 Model selection

In order to compare different nested models, that is, when one model can be obtained from the others by imposing parametric restrictions, the global deviance or the generalized Akaike information criterion can be used. They are defined, respectively, as $GD = -2\hat{\ell}$ and GAIC $= -2\hat{\ell} + (\phi.df)$, where $\hat{\ell} = \sum_{i=1}^{n} \log[f(y_i|\hat{\boldsymbol{\theta}}_i)]$ is the maximized log-likelihood function, ϕ is a penalization term and df denotes the total (effective) degrees of freedom of the model. When $\phi = 2$, GAIC reduces to the usual Akaike information criterion (AIC) and when $\phi = \log(n)$ it equals the Schwartz Bayesian criterion (SBC). The SBC tends to favor more parsimonious models.

Consider two parametric GAMLSS models M_0 and M_1, where M_0 is a particular case of M_1. Both models can be compared using the generalized LR test statistic given by

$$\Lambda = GD_0 - GD_1, \tag{1.7}$$

where GD_0 and GD_1 are the global deviances of M_0 and M_1, respectively. GD_0 and GD_1 have df_0 and df_1 degrees of freedom, respectively. The Λ statistic has asymptotic χ^2 distribution under the null hypothesis that M_0 is the correct model, the distribution number of degrees of freedom being $d = df_0 - df_1$. When models M_0 and M_1 contain nonparametric additive terms the procedure proposed by Hastie & Tibshirani (1990) to compare nested GAMs can be used.

In order to evaluate the goodness-of-fit of GAMLSS models, the generalized pseudo R-squared of Nagelkerke (1991) can be used. It is defined as

$$R^2 = 1 - \left(\frac{L(0)}{L(\hat{\theta})}\right)^{2/n},$$

where $L(0)$ is the likelihood of the null model (only a constant is fitted to all parameters)

and $L(\hat{\theta})$ is the current fitted model. This definition is also referred to as the Cox and Snell R^2.

The GAIC and SBC cannot be used to select a model from a set of nonnested candidate models. In the next section we shall develop tests that can be used with nonnested models.

1.3 Nonnested hypothesis tests for GAMLSS models

Strategies have been proposed on literature for choosing a model from a set of regression nonnested linear models. The J test proposed by Davidson & MacKinnon (1981) is one of the most commonly used tests. Suppose we have M ($M > 2$) matrices of regressors that are associated to M competing nonnested regression models for the response variable $y = (y_1, y_2, \ldots, y_n)^\top$. The J test is performed by extending the model under evaluation using fitted values from the competing models. If the additional terms do not improve the model fit considerably, the model is not rejected. The artificial (augmented) model is

$$y = \left(1 - \sum_{\substack{l=1 \\ l \neq m}}^{M} \lambda_l\right) X_m \theta_m + \sum_{\substack{l=1 \\ l \neq m}}^{M} \lambda_l X_l \hat{\theta}_l + u, \tag{1.8}$$

where λ_l, $l = 1, \ldots, M$, is a scalar; X_m is the covariates matrix (of dimension $n \times k_m$) of the mth model; θ_m is a vector containing k_m unknown parameters; $\hat{\theta}_l$ is the the least squares estimator of the parameter vector of the lth model (i.e., $\hat{\theta}_l = (X_l^\top X_l)^{-1} X_l^\top y$) and u is a vector of random errors.

The J test consists in verify the validity of Model m against the $M - 1$ alternative models by testing $\lambda_l = 0$ ($l \neq m$) in Equation (1.8). Let

$$\omega_m = n^{-1/2} \left(y^\top P_l M_m y\right)_{l \in \mathcal{M} \setminus \{m\}}$$

and

$$\widehat{\Sigma}_m = n^{-1} \left(y^\top P_l M_m \widehat{\Omega}_m M_m P_{l'} y\right)_{l, l' \in \mathcal{M} \setminus \{m\}},$$

where $\mathcal{M} = \{1, \ldots, M\}$, $P_m = X_m(X_m^\top X_m)^{-1}X_m^\top$, $M_m = I_n - P_m$ and $\widehat{\Omega}_m = \text{diag}\{\hat{u}_{1,m}^2, \ldots, \hat{u}_{n,m}^2\}$ is a diagonal matrix that contains the square residuals ($\hat{u}_{i,m}^2 = y_i - x_{i,m}^\top \hat{\theta}_m$). The J statistic can be written as (Hagemann 2012)

$$J = \omega_m^\top \widehat{\Sigma}_m^{-1} \omega_m. \tag{1.9}$$

The test statistic in Equation (1.9) is asymptotically χ^2_{M-1}-distributed under the null hypothesis.

Model m is rejected if its fit is considerably improved by adding fitted values from the competing models as additional regressors. Otherwise, Model m is not rejected. The testing strategy consists of adding artificial regressors, that are obtained from the competing models, and then testing their exclusion. Notice that the test is sequentially applied for each candidate model. For example, when there are two competing nonnested models, each model should be tested against the other. The test may suggest that more than one model is not rejected.

In order to avoid sequential testing, Hagemann (2012) introduced the *MJ* test. The test is based on the idea that if the correct model (m^*) is in the set of candidate models, its associated J statistic has a well-defined asymptotic null distribution, whereas the other J statistics diverge. However, if m^* is not in the set of candidate models, all statistics diverge to ∞ as the sample size increases. Therefore, the model with the smallest J statistic is the natural candidate to be taken as the correct model. Moreover, all candidate models are safely rejected if the smallest J statistic is large. The *MJ* test is then useful to test the null hypothesis that the correct model is on of the candidate models. Notice that no sequential testing is needed. The test also provides a model selection procedure: if the null hypothesis is not rejected by the *MJ* test, the model with the smallest J statistic can be selected as the true model. Such a model selection strategy is asymptotically correct, that is, it selects the correct model asymptotic with probability equal to one. Readers are referred to (Hagemann 2012) for more details.

Given M nonnested models, the MJ test statistic is defined as

$$MJ = \min\{J_1, \ldots, J_M\}, \tag{1.10}$$

where J_m, $m = 1, 2, \ldots, M$, is given in Equation (1.9). Under certain regularity conditions (see Cox & Hinkley (1974), p. 281, for futher details) and under the null hypothesis , if the correct model m^* is among the candidate models, the asymptotic distribution of J_{m^*} is χ^2_{M-1} (Hagemann 2012). Furthermore, for every model $m \in \mathcal{M}\backslash\{m^*\}$ and for all $a \in \mathbb{R}$, $\lim_{n\to\infty} \mathbb{P}(J_m \geq a) = 1$. Therefore, the model that corresponds to the smallest J statistic is the only candidate for true model. At the α significance level, the MJ test is carried out as follows: (1) for each candidate $m \in \mathcal{M}$, estimate the augmented model given in Equation (1.8) and compute the corresponding J test statistic; (2) compute the MJ statistic using Equation (1.10); (3) reject the null hypothesis $\mathcal{H}_0 : m^* \in \mathcal{M}$ (the correct model is one of the candidate models) in favor of $\mathcal{H}_1 : m^* \notin \mathcal{M}$ (the correct model is none of the candidate models) if $MJ > \chi^2_{1-\alpha,M-1}$, where $\chi^2_{1-\alpha,M-1}$ is the $1-\alpha$ quantile of the χ^2_{M-1} distribution.

Notice that MJ statistic has the same asymptotic null distribution as J_{m^*}, the J statistic for the correct model m^*, whereas the $M-1$ test statistics diverge.

Let $\hat{m}$ denote the model associated to the smallest MJ statistic. If the null hypothesis is not rejected, that is, there is evidence that $m^* \in \mathcal{M}$, then $\hat{m}$ is the candidate for m^*. This model selection procedure is consistent when the correct model is one of the candidates.

1.3.1 J and MJ tests for GAMLSS models

Both J and MJ tests were developed for the linear regression model. Hereafter we consider their use in GAMLSS models. Nonnested GAMLSS models may differ in regressors and/or link functions or have different probability distributions for the response variable. Suppose the interest lies in testing M competing nonnested GAMLSS models.

That is, we want to choose a model from $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_M$, such that $\mathcal{H}_m$ is one of the models defined in Equations (1.1)-(1.5). For simplicity, we wish to test M nonnested simple parametric linear GAMLSS models, that is, we are interested in choosing a model from $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_m$, such that, for $m = 1, 2, \ldots, M$,

$$\begin{aligned} \mathcal{H}_m: \quad & g_{1m}(\boldsymbol{\mu}) = \boldsymbol{\eta}_{1m} = \mathbf{X}_{1m}\boldsymbol{\beta}_{1m}, \\ & g_{2m}(\boldsymbol{\sigma}) = \boldsymbol{\eta}_{2m} = \mathbf{X}_{2m}\boldsymbol{\beta}_{2m}, \\ & g_{3m}(\boldsymbol{\nu}) = \boldsymbol{\eta}_{3m} = \mathbf{X}_{3m}\boldsymbol{\beta}_{3m}, \\ & g_{4m}(\boldsymbol{\tau}) = \boldsymbol{\eta}_{4m} = \mathbf{X}_{4m}\boldsymbol{\beta}_{4m}. \end{aligned} \tag{1.11}$$

Let $\delta_{\mu(m)}, \delta_{\sigma(m)}, \delta_{\nu(m)}$ and $\delta_{\tau(m)}$ be the number of submodels for μ, σ, ν and τ that differ from the corresponding submodel in the mth model ($\mathcal{H}_m$). For testing Model $\mathcal{H}_m$ using the J test, one first estimates the parameters in the remaining models, and then includes the estimated predictors $\hat{\boldsymbol{\eta}}_{1r}$, $\hat{\boldsymbol{\eta}}_{2r}$, $\hat{\boldsymbol{\eta}}_{3r}$ and $\hat{\boldsymbol{\eta}}_{4r}$ ($r = 1, 2, \ldots, M$; $r \neq m$) that differ from $\mathcal{H}_m$ as additional regressors in its respective submodels. The LR J statistic for testing the joint exclusion of $\hat{\boldsymbol{\eta}}_{1r}$, $\hat{\boldsymbol{\eta}}_{2r}$, $\hat{\boldsymbol{\eta}}_{3r}$ and $\hat{\boldsymbol{\eta}}_{4r}$ ($r = 1, 2, \ldots, M$; $r \neq m$) is

$$J_m = 2\{\hat{\ell}_m - \tilde{\ell}_m\}, \tag{1.12}$$

where $\hat{\ell}_m$ and $\tilde{\ell}_m$ are the (penalized) log-likelihood functions, defined in Equation (1.6), evaluated at the maximum likelihood estimators for the augmented model and for the respective mth model ($\mathcal{H}_m$), respectively. Note that $\delta_{\mu(m)}, \delta_{\sigma(m)}, \delta_{\nu(m)}$ and $\delta_{\tau(m)}$ additional regressors are included in the submodels for μ, σ, ν and τ, respectively. Hence, the total number of regressors included in the augmented model is $\delta_m = \delta_{\mu(m)} + \delta_{\sigma(m)} + \delta_{\nu(m)} + \delta_{\tau(m)}$. The model specified in $\mathcal{H}_m$ is rejected at significance level α when $J_m > \chi^2_{1-\alpha,\delta_m}$.

For instance, suppose there is interest in testing two nonnested GAMLSS models, $\mathcal{H}_1$ and $\mathcal{H}_2$, which differ in the μ and σ submodels. In order to test $\mathcal{H}_1$ using J test, the parameters of Model $\mathcal{H}_2$ are estimated and the estimated linear predictors ($\hat{\eta}_{12}$ and $\hat{\eta}_{22}$)

are included in the respective submodels of the model under test (Model $\mathcal{H}_1$). Note that $\delta_{\mu(m)} = 1$, $\delta_{\sigma(m)} = 1$ and $\delta_{\nu(m)} = \delta_{\tau(m)} = 0$, for $m = 1, 2$. The augmented model is then estimated and the J_1 statistic given in Equation (1.12) is computed. Model $\mathcal{H}_1$ is rejected at significance level α if $J_1 > \chi^2_{1-\alpha;2}$ (since $\delta_1 = 2$). For testing $\mathcal{H}_2$, one includes $\hat{\eta}_{11}$ and $\hat{\eta}_{21}$ as additional regressors in the Model $\mathcal{H}_2$ and test their joint exclusion.

If the competing models have a different number of submodels, only the submodels with regression structures in both models are considered. For example, $\mathcal{H}_1$ has three submodels (one for $\boldsymbol{\mu}$, one for $\boldsymbol{\sigma}$ and one for $\boldsymbol{\nu}$) and $\mathcal{H}_2$ has two submodels (one for $\boldsymbol{\mu}$ and one for $\boldsymbol{\sigma}$), only the submodels for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are to be considered in $\mathcal{H}_1$ and $\mathcal{H}_2$ when performing the J test.

For nonnested models with different distributions for the response variable, the respective linear predictors are included in the submodels with regression structure. Suppose, for instance, that two models were estimated. For Model H_{GU} it is assumed that the response variable has Gumbel(μ, σ) distribution (Crowder et al. 1991, p. 17) and for model H_{BCT} it is assumed the Box-Cox t distribution with parameters μ, σ, ν and τ, BCT(μ, σ, ν, τ) (Rigby & Stasinopoulos 2006). In order to perform the J test for H_{GU} and H_{BCT}, only the linear predictors of the submodels for μ and σ are included in the respective competing submodels and tested the exclusion. Here, the J_{GU} and J_{BCT} statistics have χ^2 distribution with 2 degrees of freedom.

Suppose there are M nonnested GAMLSS models under consideration. To perform the MJ test, one must compute the J test statistic for each model and then set $MJ = \min\{J_1, J_2, \ldots, J_M\}$, where J_m $(i = 1, 2, \ldots, M)$ is defined in Equation (1.12). The null hypothesis that the correct model is in the set of candidate models $(\mathcal{H}_1, \mathcal{H}_2, \ldots, H_M)$ is rejected if $MJ > \chi^2_{1-\alpha,s}$, where s is the number of degrees of freedom in the J test corresponding to the smallest J test statistic. Suppose the MJ test is to be performed for testing two nonnested GAMLSS models ($\mathcal{H}_1$ and $\mathcal{H}_2$) with differences in three submodels. After computing the J statistic for each model, the statistic $MJ = \min\{J_1, J_2\}$ is obtained. If $MJ > \chi^2_{1-\alpha,3}$ (notice that $\delta_{(m)} = 3$ for $m = 1, 2$), there is no evidence that one of the

model is the correct model. If the null hypothesis is not rejected, the model corresponding to the minimal J statistic is selected as the correct model.

The null distributions of J and MJ can be poorly approximated by their asymptotic counterpart (χ^2) in small sample sizes. Improved hypothesis testing inference can be achieved by using the bootstrap method (Fan & Li 1995, Godfrey 1998, Hagemann 2012).

Let $y = (y_1, y_2, \ldots, y_n)^\top$ be a vector of independent random variables with a certain distribution, $y_i \sim D(\mu_i, \sigma_i, \nu_i, \tau_i)$, $i = 1, 2, \ldots, n$. The bootstrap J test for testing the GAMLSS model $\mathcal{H}_1$ against $\mathcal{H}_2$, with differences in one submodel, say the submodel for $\boldsymbol{\theta}_i$, can be outlined as follows:

1. Estimate Model $\mathcal{H}_2$, obtain the $\hat{\boldsymbol{\eta}}_{i2}$ and include it as an additional regressors in model $\mathcal{H}_1$. Estimate the augmented model.

2. Compute the J test statistic.

3. Generate a bootstrap sample, y^*, with $y_i^* \overset{ind}{\sim} D(\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$, where $\hat{\mu}_i$, $\hat{\sigma}_i$, $\hat{\nu}_i$ and $\hat{\tau}_i$ are the (penalized) maximum likelihood estimates of the parameters that index Model $\mathcal{H}_1$.

4. Estimate the augmented model using y^* as response variable and compute J^*.

5. Execute steps (3) and (4) B times, where B is a large positive integer.

6. Compute $\varphi_{1-\alpha}$, the $1-\alpha$ quantile of all bootstrap test statistics $(J_1^*, J_2^*, \ldots, J_B^*)$.

7. Reject Model $\mathcal{H}_1$ at significance level α if $J > \varphi_{1-\alpha}$.

The decision rule can also be expressed using the bootstrap p-value, given by

$$p^* = \frac{\#J_b^* > J}{B}, \quad b = 1, 2, \ldots, B,$$

where # denotes the cardinality of a set and J_b^* is the J statistic computed using the bth bootstrap sample. Model $\mathcal{H}_1$ is rejected if p^* is smaller than the chosen nominal level. For testing $\mathcal{H}_2$ and for more than two nonnested models ($M > 2$), the procedure is similar.

The bootstrap MJ test is performed as follows.

1. Compute the MJ statistic as describe above.
2. Generate a bootstrap sample, y^*, with $y_i^* \overset{ind}{\sim} D(\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$, where $\hat{\mu}_i$, $\hat{\sigma}_i$, $\hat{\nu}_i$ and $\hat{\tau}_i$ are the (penalized) maximum likelihood estimates of the parameters that index Model $\mathcal{H}_1$.
3. Compute the bootstrap MJ statistic, MJ^*.
4. Execute steps (2) and (3) B times.
5. Compute $\varphi_{1-\alpha}$, the $1-\alpha$ quantile of all MJ bootstrap test statistics $(MJ_1^*, MJ_2^*, \ldots, MJ_B^*)$.
6. Reject the null hypothesis that the true model is in the set of candidates if $MJ > \varphi_{1-\alpha}$.

The tests described in this section for GAMLSS models are extensions of the J and MJ tests introduced by (Cribari-Neto & Lucena 2015) for varying dispersion beta regression models.

1.4 Numerical results

In order to evaluate the finite sample behavior of the J and MJ tests, several Monte Carlo Simulations were performed considering $M = 2$ GAMLSS models with different regressors, link functions and distributions. The finite sample performances of the tests and their bootstrap versions were evaluated. All tests are based on the LR statistic. The covariates values were chosen as random draws from the standard uniform distribution, $\mathcal{U}(0, 1)$, and were held constant all through the experiment. The number of Monte Carlo and bootstrap replications were 10,000 and 1000, respectively. The tests nominal levels were $\alpha = 1\%$, 5% and 10%. All the simulations were carried out using R software; `https://www.r-project.org/`. In order to speed up the execution time of the programs,

the function `enableJIT(3)` from the library `compiler` were used. The `GAMLSS` package were used.

The sample sizes considered were $n = 25$, 50, 75 and 100. For each regressors we generated 10 observations and then replicated the necessary number of times to achieve the sample size. This was done to guarantee that the degree of heterogeneity is held constant for all sample sizes.

We considered 10 different scenarios with two competing models with different regressors, link functions and distributions. The models were always defined as $\mathcal{H}_1$ against $\mathcal{H}_2$ and the true data-generating process corresponds to Model $\mathcal{H}_1$. In the first six scenarios the probability distribution assumed for the response variable in both competing models is Weibull. The parametrization used in the p.d.f. is such that μ is the mean of the distribution. For a random variable $Y \sim \text{Weibull}(\mu, \sigma)$, its p.d.f. is given by (Rigby et al. 2014)

$$f(y|\mu,\sigma) = \frac{\sigma}{\phi}\left(\frac{y}{\phi}\right)^{\sigma-1}\exp\left\{-\left(\frac{y}{\phi}\right)^{\sigma}\right\}, \tag{1.13}$$

$y > 0$, $\mu > 0$ and $\sigma > 0$ and $\phi = \mu/\Gamma(\frac{1}{\sigma}+1)$, where $\Gamma(\cdot)$ is the gamma function. The distribution mean is μ and its variance is $\mu^2\left\{\Gamma(\frac{2}{\sigma}+1)\left[\Gamma(\frac{1}{\sigma}+1)\right]^{-2}-1\right\}$.

Table 1.1 presents the tests null rejection rates (entries are percentages) for competing GAMLSS models with Weibull distributed response. We present numerical results for three scenarios, namely: competing models with different regressors in the μ submodel (SC1), in the σ submodel (SC2) and in both submodels (SC3). Both models are simple parametric GAMLSS models.

The competing models in the scenario SC1 are

$$\begin{aligned}\mathcal{H}_1:\ &\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \quad \text{and} \quad \mathcal{H}_2:\ \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3}\\ &\log(\sigma_i) = \gamma_0 + \gamma_1 x_{i1} \qquad\qquad\qquad\quad \log(\sigma_i) = \gamma_0 + \gamma_1 x_{i1},\end{aligned}$$

$i = 1, 2, \ldots, n$. Model $\mathcal{H}_1$ uses the regressors x_{i1} and x_{i2} in the μ submodel whereas in $\mathcal{H}_2$ the regressors are x_{i1} and x_{i3}. The parameter values are $\beta_0 = 1.0$, $\beta_1 = 2.7$, $\beta_2 = 1.6$,

$\gamma_0 = 1.4$ and $\gamma_1 = 2.0$.

Under scenario SC2, the simple parametric GAMLSS models under evaluation are

$$\begin{array}{lll} \mathcal{H}_1: & \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} & \text{and} \quad \mathcal{H}_2: \quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\ & \log(\sigma_i) = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i3} & \qquad\qquad\quad \log(\sigma_i) = \gamma_0 + \gamma_1 x_{i1} + \gamma_3 x_{i4}, \end{array}$$

$i = 1, 2, \ldots, n$. The models differ in the σ submodel ($\mathcal{H}_1$ has x_{i1} and x_{i3} as regressors whereas $\mathcal{H}_2$ uses x_{i1} and x_{i4}). The parameter values are $\beta_0 = 1.0$, $\beta_1 = 2.7$, $\beta_2 = 1.6$, $\gamma_0 = 1.0$ and $\gamma_1 = 2.0$ and $\gamma_2 = 1.3$.

In the scenario SC3 the competing models are

$$\begin{array}{lll} \mathcal{H}_1: & \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} & \text{and} \quad \mathcal{H}_2: \quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} \\ & \log(\sigma_i) = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i4} & \qquad\qquad\quad \log(\sigma_i) = \gamma_0 + \gamma_1 x_{i1} + \gamma_3 x_{i5}, \end{array}$$

$i = 1, 2, \ldots, n$. The models differ in both μ and σ submodels. The submodels have two regressors and differ in just one of them when compared to the corresponding submodel.

Since in scenario SC1 the models differ only in μ submodel, $\delta_{\mu(m)} = 1$ and $\delta_{\sigma(m)} = 0$, $m = 1, 2$. In SC2, we have $\delta_{\mu(m)} = 0$ and $\delta_{\sigma(m)} = 1$, because the difference lies in the σ submodel. Then, in both SC1 and SC2 the critical values were obtained from the χ^2 distribution with $\delta_m = \delta_{\mu(m)} + \delta_{\sigma(m)} = 1$ degree of freedom. For SC3, $\delta_{\mu(m)} = 1$ and $\delta_{\sigma(m)} = 1$, and the number of degrees of freedom is $\delta_{(m)} = 2$, $m = 1, 2$.

The results in Table 1.1 contain results relative to scenarios SC1–SC3. The bootstrap tests clearly outperform the corresponding tests based on asymptotic critical values. Large size distortions are observed when the sample size is small and the tests are based on χ^2 critical values. For example, when $n = 25$ and $\alpha = 0.05$ in SC3, the null rejection rates of the J and MJ tests are close to 13%. In contrast, the null rejection rates of the bootstrap tests are close to 5%. The tests based on asymptotic critical values are liberal, except for the MJ test in scenario SC2.

We computed the mean, the standard deviation and the coefficient of variation of the estimates. Table 1.2 contains such results for $n = 50$ for scenario SC1. Notice that the

mean estimates are close to the true parameters values. For the remaining models and other sample sizes the results are similar.

Table 1.1: Null rejection rates (%) for scenarios SC1, SC2 and SC3.

Scenario		$\alpha = 1\%$				$\alpha = 5\%$				$\alpha = 10\%$			
	n	25	50	75	100	25	50	75	100	25	50	75	100
SC1	J	2.3	1.8	1.5	2.2	9.1	6.5	5.5	6.8	15.3	12.3	10.7	13.0
	J_{boot}	1.5	1.4	1.3	1.5	4.8	5.3	4.7	5.2	9.8	9.2	9.6	10.2
	MJ	2.3	1.8	1.5	2.2	9.1	6.5	5.5	6.8	15.3	12.3	10.7	13.0
	MJ_{boot}	1.5	1.3	1.6	1.5	4.8	5.3	4.7	5.2	9.8	9.2	9.6	10.2
SC2	J	3.4	1.5	2.2	1.2	10.6	6.9	5.6	5.8	16.2	11.9	10.6	12.7
	J_{boot}	1.3	1.5	1.4	1.3	5.3	5.2	5.3	5.2	10.0	9.5	9.6	10.4
	MJ	0.7	0.5	1.1	0.4	3.8	3.1	3.4	3.8	6.3	7.1	7.1	10.2
	MJ_{boot}	0.9	1.3	1.3	1.1	5.4	4.9	4.9	5.2	9.7	9.3	9.5	10.3
SC3	J	4.7	1.9	1.2	1.3	13.2	7.8	6.4	6.1	21.4	13.7	12.8	12.4
	J_{boot}	1.8	1.5	1.0	1.1	5.2	5.3	5.1	4.8	10.4	10.1	10.3	10.3
	MJ	4.7	1.9	1.2	1.3	13.2	7.8	6.4	6.1	21.4	13.7	12.8	12.4
	MJ_{boot}	1.8	1.5	1.0	1.1	5.2	5.3	5.1	4.8	10.4	10.1	10.3	10.3

Table 1.2: Means, standard deviations and coefficients of variation, Weibull distributed response, $n = 50$.

Estimate	Parameter value	Mean	Std. deviation	Coef. of variation
$\hat{\beta}_0$	1.0	1.004	0.056	1.004
$\hat{\beta}_1$	2.7	2.703	0.053	1.001
$\hat{\beta}_2$	1.6	1.592	0.065	0.995
$\hat{\gamma}_0$	1.4	1.438	0.347	1.027
$\hat{\gamma}_1$	2.0	2.053	0.650	1.027

In the next three scenarios, different link functions were used in the submodels of the competing models. The models are

$$\mathcal{H}_1: \quad g_{11}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \quad \text{and} \quad \mathcal{H}_2: \quad g_{12}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i2}$$
$$g_{21}(\sigma_i) = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i3} \qquad\qquad g_{22}(\sigma_i) = \gamma_0 + \gamma_1 x_{i1} + \gamma_3 x_{i3},$$

$i = 1, 2, \ldots, n$; g_{12} and g_{22} being the logarithm link function. In scenario SC4, g_{11} is the inverse function and g_{21} is the logarithm function. Under scenario SC5, g_{11} is the logarithm function and g_{21} is the identity function. In scenario SC6, g_{11} and g_{21} are the

inverse and identity functions, respectively. Thus, the models in the scenario SC4, SC5 and SC6 differ in the μ submodels, in the σ submodel and in both submodel, respectively. Hence, $\delta_m = 1$ in SC4 and SC5, whereas $\delta_m = 2$ in SC6.

Table 1.3 contains the null rejection rates corresponding to SC4, SC5 and SC6. Once again, the figures show that J test is liberal in SC4 and SC5 when the submodels differ in the link functions. Overall, the bootstrap tests outperformed the corresponding tests based on asymptotic critical values.

Table 1.3: Null rejection rates (%), scenarios SC4, SC5 and SC6.

Scenario		$\alpha = 1\%$				$\alpha = 5\%$				$\alpha = 10\%$			
	n	25	50	75	100	25	50	75	100	25	50	75	100
SC4	J	3.1	2.1	1.6	1.4	9.6	6.7	7.4	6.8	17.6	11.5	12.9	11.8
	J_{boot}	1.1	1.4	1.5	1.4	5.3	4.9	6.0	5.9	10.0	9.8	10.8	11.1
	MJ	1.7	1.1	0.7	0.6	5.4	3.0	3.8	3.4	10.6	5.6	7.3	5.5
	MJ_{boot}	1.6	1.4	1.8	1.8	6.2	4.6	5.7	5.6	11.0	9.0	11.2	10.2
SC5	J	3.8	1.7	2.1	1.9	10.0	5.4	7.7	6.8	17.3	11.4	12.5	11.5
	J_{boot}	1.5	1.5	2.0	1.9	5.3	4.9	5.3	5.5	10.3	9.7	10.9	10.6
	MJ	2.8	1.5	1.8	1.7	9.0	4.9	7.1	6.2	15.6	10.0	11.6	10.7
	MJ_{boot}	1.7	1.3	2.0	2.0	5.0	4.9	5.6	5.6	10.3	9.6	11.6	10.5
SC6	J	5.2	1.6	1.6	1.4	12.0	7.0	7.1	7.1	20.0	13.7	13.1	13.0
	J_{boot}	1.5	0.7	1.3	1.3	5.4	4.9	5.5	5.2	10.3	10.1	11.0	11.2
	MJ	1.8	1.0	1.0	1.1	7.6	4.3	4.3	4.2	13.0	8.8	8.3	8.0
	MJ_{boot}	1.4	1.2	1.3	1.5	5.6	5.7	5.2	5.6	10.4	10.7	10.2	10.5

The behavior of J and MJ tests were also studied when different distributions are specified for the response variable. In scenario SC7 the true model ($\mathcal{H}_0$) uses the Weibull distribution for the response, with p.d.f. given by Equation (1.13), whereas the competing model ($\mathcal{H}_1$) uses the gamma distribution. The gamma p.d.f., denoted by GA(μ, σ) (Rigby et al. 2014), is given by

$$f(y|\mu,\sigma) = \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y^{\frac{1}{\sigma^2}-1} e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)},$$

for $y > 0$, $\mu > 0$ and $\sigma > 0$. Here, $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \sigma^2\mu^2$. The models in $\mathcal{H}_0$ and $\mathcal{H}_1$ were defined exactly as in Scenario SC3: the link function in both submodels is the logarithm function and the covariates in both submodels are different.

Competing GAMLSS models with three distributional paramaters are considered in scenario SC8. The type 2 power exponential distribution (PE2) is taken as true distribution and the alternative model assumes the t family distribution (TF). The PE2(μ, σ, ν) p.d.f. is given by (Rigby et al. 2014)

$$f(y|\mu, \sigma, \nu) = \frac{\nu \exp[-|z|^{\nu}]}{2\sigma\Gamma(\frac{1}{\nu})},$$

$-\infty < y, \mu < +\infty$, $\sigma, \nu > 0$, $z = (y - \mu)/\sigma$. Here, $E(y) = \mu$ and $\text{Var}(y) = \sigma^2 c^2$, where $c^2 = \Gamma(1/\nu)[\Gamma(3/\nu)]^{-1}$. The TF$(\mu, \sigma, \nu)$ p.d.f. is (Rigby et al. 2014)

$$f(y|\mu, \sigma, \nu) = \frac{1}{\sigma B(\frac{1}{2}, \frac{\nu}{2})\sqrt{\nu}} \left[1 + \frac{(y-\mu)^2}{\sigma^2 \nu}\right]^{-\frac{\nu+1}{2}},$$

$-\infty < y, \mu < +\infty$, $\sigma, \nu > 0$, where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ is the beta function. The mean is $E(y) = \mu$ and the variance is $\text{Var}(y) = \sigma^2\nu/(\nu - 2)$ for $\nu > 2$. The random variable $T = (y - \mu)/\sigma$ has standard t distribution with ν degrees of freedom. The regression structures defined for the competing models in scenario SC8 are

$$\begin{aligned}
\mathcal{H}_1: \quad & \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \qquad \text{and} \quad \mathcal{H}_2: \quad \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} \\
& \log(\sigma_i) = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i4} \qquad\qquad\qquad\quad \log(\sigma_i) = \gamma_0 + \gamma_1 x_{i1} + \gamma_3 x_{i5}, \\
& \log(\nu_i) = \varphi_0 + \varphi_1 x_{i2} + \varphi_2 x_{i6} \qquad\qquad\qquad\quad \log(\nu_i) = \varphi_0 + \varphi_1 x_{i2} + \varphi_3 x_{i7},
\end{aligned}$$

$i = 1, \ldots, n$, $\log(\cdot)$ denotes the logarithm function. Here, the models differ in one regressor in each submodel. Note that $\delta_{(m)} = \delta_{\mu(m)} + \delta_{\sigma(m)} + \delta_{\nu(m)} = 1 + 1 + 1 = 3$. The parameters values in model $\mathcal{H}_1$ (true model) are $\beta_0 = \gamma_0 = \varphi_0 = 1.0$, $\beta_1 = 1.6$, $\beta_2 = 2.7$, $\gamma_1 = 3.5$, $\gamma_2 = 2.2$, $\varphi_1 = 2.3$ and $\varphi_2 = 1.4$.

In scenario SC9, the nonnested models distributions are indexed by four parameters: the distribution used in model $\mathcal{H}_1$ is the Box-Cox t distribution (BCT) and the generalized beta distribution (GB2) is used in Model $\mathcal{H}_2$. The BCT(μ, σ, ν, τ) p.d.f. is given by (Rigby

et al. 2014)

$$f(y|\mu,\sigma,\nu,\tau)=\frac{y^{\nu-1}f_T(z)}{\mu\nu\sigma F_T\left(\frac{1}{\sigma|\nu|}\right)},$$

for $y,\mu,\sigma>0$ and $-\infty<\nu<+\infty$, where $f_T(t)$ and $F_T(t)$ is are, respectively, the p.d.f. and the cumulative distribution function of a random variable T having standard t distribution with $\tau>0$ degrees of freedom; z is defined as $\frac{1}{\sigma\nu}\left[\left(\frac{y}{\mu}\right)^{\nu}-1\right]$, if $\nu\neq 0$, and $\frac{1}{\sigma}\log\left(\frac{y}{\mu}\right)$, if $\nu=0$. The EGB(μ,σ,ν,τ) p.d.f. is (Rigby et al. 2014)

$$f(y|\mu,\sigma,\nu,\tau)=\mathrm{e}^{\nu z}\{|\sigma|B(\nu,\tau)[1+\mathbf{e}^{z}]^{\nu+\tau}\}^{-1},$$

$-\infty<y,\mu,\sigma<+\infty$, $\nu,\tau>0$, where $z=(y-\mu)/\sigma$. The competing models are

$$\begin{array}{lll}
\mathcal{H}_1: & \log(\mu_i)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2} & \text{and}\quad \mathcal{H}_2: \quad \log(\mu_i)=\beta_0+\beta_1x_{i1}+\beta_3x_{i3}\\
& \log(\sigma_i)=\gamma_0+\gamma_1x_{i1}+\gamma_2x_{i4} & \log(\sigma_i)=\gamma_0+\gamma_1x_{i1}+\gamma_3x_{i5},\\
& \log(\nu_i)=\varphi_0+\varphi_1x_{i2}+\varphi_2x_{i6} & \log(\nu_i)=\varphi_0+\varphi_1x_{i2}+\varphi_3x_{i7},\\
& \log(\tau_i)=\vartheta_0+\vartheta_1x_{i2}+\vartheta_2x_{i8} & \log(\tau_i)=\vartheta_0+\vartheta_1x_{i2}+\vartheta_3x_{i9},
\end{array}$$

$i=1,\ldots,n$. Note that $\delta_{(m)}=\delta_{\mu(m)}+\delta_{\sigma(m)}+\delta_{\nu(m)}+\delta_{\tau(m)}=1+1+1+1=4$. The parameter values in the true model ($\mathcal{H}_1$) are $\beta_0=1.0$, $\beta_1=1.6$, $\beta_2=2.7$, $\gamma_0=0.75$, $\gamma_1=2.1$, $\gamma_2=-2.7$, $\varphi_0=1.0$, $\varphi_1=2.3$, $\varphi_2=-2.4$, $\vartheta_0=2.0$, $\vartheta_1=4.1$ and $\vartheta_2=3.9$.

A tenth scenario was also considered. In SC10 we compare the behavior of J and MJ tests in the zero-inflated beta regression model (true model) against the beta regression model after transforming the response so that it assumes values in (0,1). The transformation used is $y'=[(n-1)y+0.5]/n$, where n is the sample size (Smithson & Verkuilen 2006). Inflated beta regressions were proposed by Ospina & Ferrari (2012). The model allows y to assume values in $[0,1)$, $(0,1]$ or $[0,1]$. We consider the case where $y\in[0,1)$.

The competing models in our simulations are

$$\begin{aligned}
\mathcal{H}_1: \quad & \log\left(\tfrac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \quad \text{and} \quad \mathcal{H}_2: \quad \log\left(\tfrac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\
& \sigma_i = \sigma \qquad\qquad\qquad\qquad\qquad\qquad\quad \sigma_i = \sigma, \\
& \log\left(\tfrac{\nu_i}{1-\nu_i}\right) = \varphi_0 + \varphi_1 x_{i3} + \varphi_2 x_{i4}
\end{aligned}$$

$i = 1, \ldots, n$; ν_t is a inflated beta regression parameter such that $\nu_t = P(y_t = 0)$. The parameters in the true model are $\beta_0 = -1.0$, $\beta_1 = 0.5$, $\beta_2 = 1.2$, $\sigma = 1.5$, $\varphi_0 = -1.5$, $\varphi_1 = 0.4$, $\varphi_2 = 1.1$.

Table 1.4 contains the J and MJ tests null rejection rates (%) for scenarios SC7, SC8, SC9 and SC10. The tests based on asymptotic critical values are considerably liberal, especially when the competing models employ distributions indexed by three or four parameters (scenarios SC8 and SC9, respectively). However, MJ tests based on asymptotical critical values in SC10 are conservative and a larger sample size is required for accurate inferences (in our simulations, something like $n = 700$). In general, larger samples are needed for the J and MJ tests based on asymptotic critical values display small size distortions. The corresponding bootstrap tests presented were considerably less distorted.

As explained in the previous section, the MJ statistic is the minimal J statistic. It can be used for model selection when the null hypothesis (that the correct model is one of the candidate models) is not rejected. We computed the percentages of correct model selections when the MJ statistic is used as a model selection criterion. We only considered the replications in which the null hypothesis was not rejected. The results for scenarios SC1 (difference in regressors of the μ submodel), SC2 (difference in regressors of the σ submodel), SC4 (difference in link functions) and SC7 for $\alpha = 5\%$ (different distributions assumed for the response variable) are presented in Table 1.5. The results for the other scenarios were quite similar, except for scenario SC10, which will be discussed later. When the competing models differ by a regressor of the μ submodel or when different distributions are assumed for the response variable, MJ model selection works well. When

the models differ in their link function(s) or when there is no difference in the regressors of the μ submodel, larger sample sizes are required for reliable model selection.

Table 1.4: Null rejection rates (%), scenarios SC7, SC8, SC9 and SC10.

Scenario		$\alpha = 1\%$				$\alpha = 5\%$				$\alpha = 10\%$			
	n	25	50	75	100	25	50	75	100	25	50	75	100
SC7	J	5.0	2.3	1.7	1.3	13.9	8.8	7.4	6.1	23.3	14.9	12.8	12.4
	J_{boot}	1.9	1.6	1.3	1.1	5.7	5.6	5.7	4.8	11.4	11.0	10.4	10.3
	MJ	5.0	2.3	1.7	1.3	13.9	8.8	7.4	6.1	23.3	14.9	12.8	12.4
	MJ_{boot}	1.9	1.6	1.3	1.1	5.7	5.6	5.7	4.8	11.4	11.0	10.4	10.3
SC8	J	11.6	7.9	5.3	4.7	30.4	21.4	16.9	16.3	44.0	23.8	18.2	17.5
	J_{boot}	1.8	1.6	1.4	1.3	5.8	5.5	5.5	5.3	11.7	11.2	10.9	10.4
	MJ	8.0	5.2	4.7	4.6	26.0	19.6	16.8	16.0	40.5	23.2	18.3	17.2
	MJ_{boot}	1.8	1.6	1.5	1.1	5.9	5.6	5.5	5.2	10.9	10.6	10.2	10.1
SC9	J	14.2	9.9	8.4	8.1	27.0	23.2	18.0	16.7	34.1	30.0	26.5	18.3
	J_{boot}	1.8	1.5	1.2	1.2	6.2	5.4	5.6	5.3	11.8	10.7	10.1	10.3
	MJ	12.6	9.5	8.4	6.4	23.1	17.7	15.4	15.7	31.8	29.6	26.3	17.5
	MJ_{boot}	1.9	1.4	1.3	1.2	5.5	5.5	5.4	5.4	11.4	10.6	10.7	10.3
SC10	J	2.5	3.2	1.5	1.1	9.3	7.6	6.7	7.3	15.6	12.4	11.5	13.1
	J_{boot}	1.1	2.3	1.7	1.3	6.1	6.0	5.5	6.6	10.6	10.7	10.9	13.2
	MJ	0.0	0.0	0.0	0.0	0.4	0.5	0.5	0.3	1.6	1.8	1.2	1.2
	MJ_{boot}	1.1	1.3	0.9	0.7	4.3	4.8	3.8	4.4	8.8	9.3	8.1	8.5

Table 1.5: Frequencies (%) of correct model selection using the MJ statistic (when the null hypothesis is not rejected, $\alpha = 5\%$).

Scenario	Criterion	$n = 25$	$n = 50$	$n = 75$	$n = 100$
SC1	MJ	100,0%	100,0%	100,0%	100,0%
	MJ_{boot}	100,0%	100,0%	100,0%	100,0%
SC2	MJ	68.2%	82.0%	89.9%	93.4%
	MJ_{boot}	68.6%	82.2%	90.0%	93.4%
SC4	MJ	62.3%	71.2%	77.8%	78.5%
	MJ_{boot}	62.9%	71.3%	78.0%	78.6%
SC7	MJ	100.0%	100.0%	100.0%	100.0%
	MJ_{boot}	100.0%	100.0%	100.0%	100.0%

In order to compare MJ model selection to other approaches, we performed Monte Carlo simulations under scenarios SC2 and SC10. Two alternative model selection procedures were used: those based on the GAIC criterion and on the SBC criterion (Rigby & Stasinopoulos 2005). Model selection was performed regardless of MJ testing and also

only when the MJ test indicated that one of the competing models is the true model. Table 1.6 presents the results for $n = 50$ and $\alpha = 5\%$ in SC2. The three procedures behave similarly. However, MJ model selection has a noteworthy advantage: it is coupled with a test that indicates whether the true model is in the set of competing models.

Table 1.6: Frequencies (%) of correct model selection in scenario SC2 using different criteria ($n = 50$).

	MJ	GAIC	SBC
Regardless the rejection of $\mathcal{H}_0$	–	81.8	81.8
$\mathcal{H}_0$ is not rejected by MJ (at the 5% sig. level)	82.0	82.0	82.0
$\mathcal{H}_0$ is not rejected by MJ boot. (at the 5% sig. level)	82.2	82.2	82.2

In scenario SC10, model selection based on MJ test statistic was performed for $n = 100$, 500 and 700 and $\alpha = 5\%$. The frequencies of correct model selection (zero-inflated beta regression) based on asymptotic and bootstrap critical values of MJ test were 47.2% ($n = 100$), 66.0% ($n = 500$) and 81.1% ($n = 700$). In contrast, GAIC and SBC always selected the wrong model (beta regression with transformed) for all sample sizes considered. Here, MJ model selection had was clearer superior.

1.5 Application

We shall now present an empirical application. The interest lies in modeling the monthly net rent (in Euros) per month in Munich. The covariates are living area in square meters (x_1); year of construction (x_2); a factor with three levels, indicating the quality of location according to an expert assessment ($x_{3(2)}$: good location, $x_{3(3)}$: top location); a factor with two levels indicating whether the bath facilities are standard or premium (x_4); a factor with two levels indicating the quality of the kitchen (x_5); a factor indicating the presence of central heating (x_6); and district in Munich (x_7). More details about the data can be found in Fahrmeir et al. (2013). Four different GAMLSS nonnested models were considered with different covariates and probability distributions for the response.

Model H_{BCCG} uses Box-Cox Cole-Green distribution for the response, model H_{BCPE} uses the Box-Cox power exponential distribution, model H_{WEI} uses the Weibull distribution and model H_{GG} uses the generalized gamma distribution. The models are

$$\begin{aligned}
H_{BCCG}: \quad \log(\mu) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3(2)} + \beta_4 x_{3(3)} + \beta_5 x_4 + \beta_6 x_5 + \beta_6 x_7 + \beta_8 x_7 \\
&\quad + \beta_9 x_1 x_4 + \beta_{10} x_1 x_5 + \beta_{11} x_2 x_5 + \beta_{12} x_2 x_7 + \beta_{13} x_{3(2)} x_4 + \beta_{14} x_{3(3)} x_4 \\
\log(\sigma) &= \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_{3(2)} + \gamma_4 x_{3(3)} + \gamma_5 x_6 + \gamma_6 x_1 x_2 + \gamma_7 x_1 x_{3(2)} \\
&\quad + \gamma_8 x_1 x_{3(3)} + \gamma_9 x_2 x_6 \\
\log(\nu) &= \varphi_0 + \gamma_1 x_2 + \gamma_2 x_5
\end{aligned}$$

$$\begin{aligned}
H_{BCPE}: \quad \log(\mu) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3(2)} + \beta_4 x_{3(3)} + \beta_5 x_4 + \beta_6 x_6 + \beta_7 x_7 \\
&\quad + \beta_8 x_1 x_4 + \beta_9 x_2 x_7 \\
\log(\sigma) &= \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_{3(2)} + \gamma_4 x_{3(3)} + \gamma_5 x_5 + \gamma_6 x_6 + \gamma_7 x_1 x_2 \\
&\quad + \gamma_8 x_1 x_{3(2)} + \gamma_9 x_1 x_{3(3)} + \gamma_{10} x_2 x_5 + \gamma_{11} x_2 x_6 \\
\log(\nu) &= \varphi_0 + \gamma_1 x_2 + \gamma_2 x_5 + \gamma_3 x_2 x_5 \\
\tau &= \vartheta_0 + \vartheta_1 x_6
\end{aligned}$$

$$\begin{aligned}
H_{WEI}: \quad \log(\mu) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3(2)} + \beta_4 x_{3(3)} + \beta_5 x_4 + \beta_6 x_5 + \beta_7 x_6 + \beta_8 x_1 x_4 \\
&\quad + \beta_9 x_1 x_6 + \beta_{10} x_2 x_5 + \beta_{11} x_{3(2)} x_4 + \beta_{12} x_{3(2)} x_4 \\
\log(\sigma) &= \gamma_0 x_1 + \gamma_1 x_2 + \gamma_2 x_6 + \gamma_3 x_2 x_6
\end{aligned}$$

$$\begin{aligned}
H_{GG}: \quad \log(\mu) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3(2)} + \beta_4 x_{3(3)} + \beta_5 x_4 + \beta_6 x_5 + \beta_7 x_6 + \beta_8 x_7 \\
&\quad + \beta_9 x_1 x_4 + \beta_{10} x_1 x_6 + \beta_{11} x_2 x_5 + \beta_{11} x_2 x_7 \\
\log(\sigma) &= \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_{3(2)} + \gamma_4 x_{3(3)} + \gamma_5 x_5 + \gamma_6 x_6 + \gamma_7 x_2 x_6 \\
\log(\nu) &= \varphi_0 + \varphi_1 x_1 + \varphi_2 x_2 + \varphi_3 x_{3(2)} + \varphi_4 x_{3(3)} + \varphi_5 x_5 + \varphi_6 x_1 x_2.
\end{aligned}$$

The parameter estimates are presented in Table 1.8. The generalized pseudo R^2 for GAMLSS models Nagelkerke (1991), GAIC and SBC of the estimated models were also computed; see Table 1.7. Model H_{WEI} has the highest pseudo-R^2. On the other hand, the GAIC favors Model H_{BCCG} and the SBC favors Model H_{BCPE}.

Table 1.7: Generalized R^2, GAIC and SBC of models H_{BCCG}, H_{BCPE}, H_{WEI} and H_{GG}.

Model	Generalized pseudo R^2	GAIC	SBC
H_{BCCG}	0.543	38,343.86	38,512.79
H_{BCPE}	0.541	38,338.88	38,507.81
H_{WEI}	0.565	38,493.37	38,601.97
H_{GG}	0.542	38,353.68	38,522.61

The J and MJ tests and their bootstrap versions were also performed. The number of bootstrap replications was 1,000. First, one should note that Models H_{BCCG}, H_{BCPE} and H_{GG} have three regression structures (submodels) whereas Model H_{WEI} has two regression structures. In order to carry out the tests, we consider the minimum number of regression structures (submodels) in all models. In this case, Model H_{WEI} has the minimum number of submodels (one for μ and another one for σ). Then, when J and MJ tests, only the submodels for μ and σ are augmented in each competing model. For example, to test Model H_{BCCG} against the other models using J test, one should include the respective estimated linear predictors of μ and σ submodels from the competing models and test their exclusion using the likelihood ratio test. We performed the tests at 5% significance level. The J test p-values for each model in the presence of the other three competing models are: 0.245 ($J = 7.902$) for H_{BCCBG}, 0.289 ($J = 7.353$) for H_{BCPE}, 7.296×10^{-11} ($J = 58.967$) for H_{WEI} and 0.576 ($J = 4.754$) for H_{GG}. We note that Model H_{WEI} is the only rejected model. The MJ test statistic is the smallest J statistic observed. It corresponds to Model H_{GG}; the p-value (0.576) indicates that the correct model is among the candidate models. Since the smallest J statistic corresponds to Model H_{GG}, such a model is selected using the MJ approach. Hence, we conclude that Model H_{GG} as the appropriate model.

Table 1.8: Parameter estimates for models H_{BCCG}, H_{BCPE}, H_{WEI} and H_{GG}.

Model H_{BCCG}		Model H_{BCPE}		Model H_{WEI}		Model H_{GG}	
$\hat{\beta}_0$	-3.110	$\hat{\beta}_0$	3.053	$\hat{\beta}_0$	-5.186	$\hat{\beta}_0$	-2.432
$\hat{\beta}_1$	1.106×10^{-2}	$\hat{\beta}_1$	-1.087×10^{-1}	$\hat{\beta}_1$	0.010	$\hat{\beta}_1$	6.824×10^{-3}
$\hat{\beta}_2$	4.182×10^{-3}	$\hat{\beta}_2$	-2.071×10^{-3}	$\hat{\beta}_2$	0.005	$\hat{\beta}_2$	4.019×10^{-3}
$\hat{\beta}_3$	7.195×10^{-2}	$\hat{\beta}_3$	-6.431×10^{-3}	$\hat{\beta}_3$	0.098	$\hat{\beta}_3$	8.677×10^{-2}
$\hat{\beta}_4$	1.944×10^{-1}	$\hat{\beta}_4$	8.885×10^{-1}	$\hat{\beta}_4$	0.192	$\hat{\beta}_4$	3.181×10^{-1}
$\hat{\beta}_5$	1.878×10^{-1}	$\hat{\beta}_5$	1.997×10	$\hat{\beta}_5$	0.173	$\hat{\beta}_5$	2.358×10^{-1}
$\hat{\beta}_6$	5.627	$\hat{\beta}_6$	1.691×10	$\hat{\beta}_6$	4.967	$\hat{\beta}_6$	7.328
$\hat{\beta}_7$	2.772×10^{-1}	$\hat{\beta}_7$	5.631×10^{-5}	$\hat{\beta}_7$	0.120	$\hat{\beta}_7$	-7.096×10^{-3}
$\hat{\beta}_8$	-3.266×10^{-3}	$\hat{\beta}_8$	1.312×10^{-3}	$\hat{\beta}_8$	-0.002	$\hat{\beta}_8$	-3.069×10^{-3}
$\hat{\beta}_9$	-2.016×10^{-3}	$\hat{\beta}_9$	-9.878×10^{-3}	$\hat{\beta}_9$	0.002	$\hat{\beta}_9$	-2.097×10^{-3}
$\hat{\beta}_{10}$	1.157×10^{-3}	$\hat{\beta}_{10}$	-9.713×10^{-3}	$\hat{\beta}_{10}$	-0.002	$\hat{\beta}_{10}$	3.630×10^{-3}
$\hat{\beta}_{11}$	-2.815×10^{-3}	$\hat{\beta}_{11}$	-8.855×10^{-3}	$\hat{\beta}_{11}$	0.084	$\hat{\beta}_{11}$	-3.586×10^{-3}
$\hat{\beta}_{12}$	1.649×10^{-6}	$\hat{\gamma}_0$	-13.905	$\hat{\beta}_{12}$	0.082	$\hat{\beta}_{12}$	1.549×10^{-6}
$\hat{\beta}_{13}$	8.678×10^{-2}	$\hat{\gamma}_1$	0.007	$\hat{\gamma}_0$	5246	$\hat{\gamma}_0$	-4.043
$\hat{\beta}_{14}$	1.065×10^{-1}	$\hat{\gamma}_2$	-13.029	$\hat{\gamma}_1$	-1.842×10^{-3}	$\hat{\gamma}_1$	0.003
$\hat{\gamma}_0$	3.156	$\hat{\gamma}_3$	0.008	$\hat{\gamma}_2$	-2.139×10^{-3}	$\hat{\gamma}_2$	0.001
$\hat{\gamma}_1$	-1.011×10^{-1}	$\hat{\varphi}_0$	0.372	$\hat{\gamma}_3$	-1.780×10	$\hat{\gamma}_3$	0.058
$\hat{\gamma}_2$	-2.128×10^{-3}	$\hat{\varphi}_1$	0.350	$\hat{\gamma}_4$	9.322×10^{-3}	$\hat{\gamma}_4$	-0.024
$\hat{\gamma}_3$	-1.950×10^{-2}					$\hat{\gamma}_5$	-0.401
$\hat{\gamma}_4$	9.116×10^{-1}					$\hat{\gamma}_6$	15.376
$\hat{\gamma}_5$	1.635×10					$\hat{\gamma}_7$	-0.008
$\hat{\gamma}_6$	5.239×10^{-5}					$\hat{\varphi}_0$	16.702
$\hat{\gamma}_7$	1.426×10^{-3}					$\hat{\varphi}_1$	-0.554
$\hat{\gamma}_8$	-9.900×10^{-3}					$\hat{\varphi}_2$	-0.008
$\hat{\gamma}_9$	-8.559×10^{-3}					$\hat{\varphi}_3$	0.069
$\hat{\varphi}_0$	3.156					$\hat{\varphi}_4$	0.962
$\hat{\varphi}_0$	-2.128×10^{-3}					$\hat{\varphi}_5$	1.575
$\hat{\varphi}_0$	-1.950×10^{-2}					$\hat{\varphi}_6$	0.003×10^{-1}
$\hat{\varphi}_0$	9.116×10^{-1}						
$\hat{\varphi}_0$	1.635×10						
$\hat{\varphi}_0$	5.239×10^{-5}						
$\hat{\varphi}_0$	1.426×10^{-3}						
$\hat{\varphi}_0$	-9.900×10^{-3}						
$\hat{\varphi}_0$	-8.559×10^{-3}						

1.6 Concluding remarks

In this chapter, we addressed the problem of nonnested hypothesis testing in GAMLSS models. We proposed variants of the J and MJ tests for such a class of models. The finite sample performances of the two tests and of their bootstrap counterparts were evaluated via Monte Carlo simulations. In GAMLSS models with three or more submodels, the tests based on asymptotic critical values tend to be liberal (oversized) in small samples. However, the size distortions become quite small when bootstrap critical values are used. The MJ test statistic may also be used as a model selection criterion. It works well when the competing models differ in distributions and regressors of the mean submodel. In other situations, a larger sample is needed for the procedure to work well. An empirical application was presented and discussed.

CHAPTER 2

The Inflated Simplex Regression Model

Resumo

Neste capítulo é proposta uma abordagem frequentista para o modelo de regressão simplex inflacionado em zero e/ou um. Esse modelo é adequado aos casos em que a variável resposta está restrita ao intervalo [0,1), (0,1] ou [0,1], tais como taxas e proporções. A distribuição simplex inflacionada é descrita como uma combinação entre a distribuição simplex e uma distribuição degenerada em 0 e/ou 1. O modelo de regressão é composto por submodelos que possuem uma estrutura de regressão associada a cada parâmetro da distribuição. O processo de estimação dos parâmetros do modelo, intervalos de confiança e testes de hipóteses são apresentados. Aplicações com dados simulados avaliando a estimação dos parâmetros do modelo de regressão também são reportados.

2.1 Introduction

Statistical modeling is oftentimes based on distributional assumptions. When such assumptions do not hold statistical inferences can be inaccurate and even invalid. A strategy to cope with violations of some assumptions of the model is to transform the response variable. However, this procedure has some limitations because it affects the variance and the parameters are no longer interpretable in terms of the response (Atkinson 1985, Ch. 7). An alternative strategy is to use regression models which assume an adequate probability distribution for the response. For example, in order to analyze data observed in the unit interval, Cox (1996) discuss some practical aspects of fitting and interpreting nonlinear quasi-likelihood regression models. Another approach was introduced by Song & Tan (2000) and Song et al. (2004) who proposed the simplex regression model to such data. Furthermore, Ferrari & Cribari-Neto (2004) and Simas et al. (2010) proposed to model data using the beta regression model.

The presence of excess of ones in a dataset precludes the possibility of using a distribution defined in the open unity interval. In this situation, the distribution may be augmented in order to allow the data to assume values in (0,1]. In this case the model is inflated in one. Many inflated models can be found in literature. Lambert (1992) introduced the zero-inflated Poisson (ZIP) regression model. Greene (1994) presented an alternative to ZIP model by extending the negative binomial model for excess of zeros count data, i.e., by developing the zero-inflated negative binomial (ZINB) model. Hall (2000) and Vieira et al. (2000)described the zero-inflated binomial (ZIB) regression model. Paul et al. (2004) proposed a zero-inflated beta-binomial (ZIBB) model. Famoye & Singh (2006) proposes the zero-inflated generalized Poisson (ZIGP) regression and modeled domestic violence data with too many zeros. Tong et al. (2013) proposed the zero-adjusted gamma (ZAGA) regression and empirically validates the model. Ospina & Ferrari (2010) presented mixed continuous-discrete distributions called inflated beta distributions. Bandyopadhyay et al. (2014) proposed, the zero-one augmented simplex regression (ZOAS-RE).

The regression model proposed by Bandyopadhyay et al. (2014) is an alternative to

the model proposed by Ospina & Ferrari (2010). Both may be used to model variates that assume values in [0,1), (0,1] or [0,1]. This is the case of modeling rates and proportions, which are typically restricted to the standard unit interval (0, 1), but it is not uncommon for the data to contain zeros and/or ones. The zero-one augmented simplex regression is a mixed continuous-discrete distribution with probability mass at zero and/or one. In the model the continuous part and the parameter(s) associated to the degenerated part of the model are modeled through a regression structure. We shall explore a frequentist approach to the model, since the authors of the ZOAS-RE model only developed a Bayesian inference for that model. Once we perform an approach similar to Ospina & Ferrari (2010), in what follows the model will be called inflated simplex regression (IS-RE). In what follows, we shall use a notation similar to that of Ospina & Ferrari (2010).

The inflated simplex model is obtained by combining the simplex and Bernoulli distributions. It is suitable for variables with support in [0,1]. For responses that assume values in [0,1) or (0,1], the new distribution is obtained by combining the simplex distribution with a distribution degenerated at 0 or 1. The simplex density is quite flexible for data modeling since it can assume many different forms. It can be, e.g., highly skewed, flat or bimodal. This makes the IS-RE a good choice for modeling proportion data that include zeros and/or ones since their continuous part is usually skewed and multimodal.

This chapter unfolds as follows. Section 2.2 describes the simplex distribution. Section 2.3 presents the inflated simplex distribution. In Section 2.4 the inflated simplex regression model is presented and inference strategies are developed and discussed.

2.2 Dispersion Models and The Simplex Distribution

The simplex distribution was introduced by Barndorff-Nielsen & Jørgensen (1991) to model data restricted to the standard unit interval (0,1). The distribution is a special case of the class of probability models known as *dispersion models*, introduced by Jørgensen (1997), which extends the class of generalized linear models, GLM (Nelder & Wedder-

burn 1972). A dispersion model, denoted by DM(μ, σ^2), with location parameter μ and dispersion parameter σ^2, is a family of distributions whose probability density function takes the form

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\}, \quad y \in C, \tag{2.1}$$

where $\mu \in \Omega \subseteq C \subseteq \mathbb{R}$, $\sigma^2 > 0$, $a(\cdot) > 0$ is a suitable function independent of μ. Here, $d(y;\mu)$ is the *unit deviance*.

The unity deviance satisfies two properties: (i) it is zero when μ equals the observed y, i.e., $d(y;y) = 0$, $\forall y \in \Omega$; (ii) it is positive when μ and the observed y are different, i.e., $d(y;\mu) > 0$, $\forall y \neq \mu$. Moreover, we say the unit deviance is said to be *regular* if the function $d(y;\mu)$ is twice continuously differentiable with respect to (y,μ) and satisfies

$$\frac{\partial^2 d(y;y)}{\partial\mu^2} = \left.\frac{\partial^2 d(y;\mu)}{\partial\mu^2}\right|_{y=\mu} > 0, \quad \forall y \in \Omega.$$

The variance function $V : \Omega \to (0,\infty)$ of a regular unity deviance is defined as

$$V(\mu) = \frac{2}{\left.\frac{\partial^2 d(y;\mu)}{\partial\mu^2}\right|_{y=\mu}}, \quad \mu \in \Omega. \tag{2.2}$$

2.2.1 Simplex Distribution

Let y be a random variable following the simplex distribution $S^-(\mu,\sigma^2)$ with parameters $\mu \in (0,1)$ and $\sigma^2 > 0$. Then, its probability density function is given by

$$f(y;\mu,\sigma^2) = \{2\pi\sigma^2[y(1-y)]^3\}^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\}, \tag{2.3}$$

for $y \in (0,1)$ and unit deviance is given by

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}. \tag{2.4}$$

Note that $a(y;\sigma^2) = \{2\pi\sigma^2[y(1-y)]^3\}^{-1/2}$ in (2.1). Using Equation (2.2), it follows that the variance function is $V(\mu) = \mu^3(1-\mu)^3$. The expectation is $E_{S^-}(y) = \mu$ and the variance is given by (Jørgensen 1997)

$$\mathrm{Var}_{S^-}(y) = \mu(1-\mu) - \frac{1}{\sqrt{2\sigma^2}} \exp\left(\frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right) \Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right), \qquad (2.5)$$

where $\Gamma(a,b)$ is the incomplete gamma function, i.e., $\Gamma(a,b) = \int_b^\infty t^{a-1}\mathrm{e}^{-t}dt$. Properties of the simplex distribution can be found in Barndorff-Nielsen & Jørgensen (1991) and Song (2007).

Different simplex densities are presented in Figure 2.1 for some values of the parameters (μ, σ^2). The distribution can be unimodal or bimodal. It can also be bell-shaped, U-shaped, J-shaped and reverse-J-shaped (i.e., L-shaped). The density function is symmetric when $\mu = 1/2$ and asymmetric when $\mu \neq 1/2$.

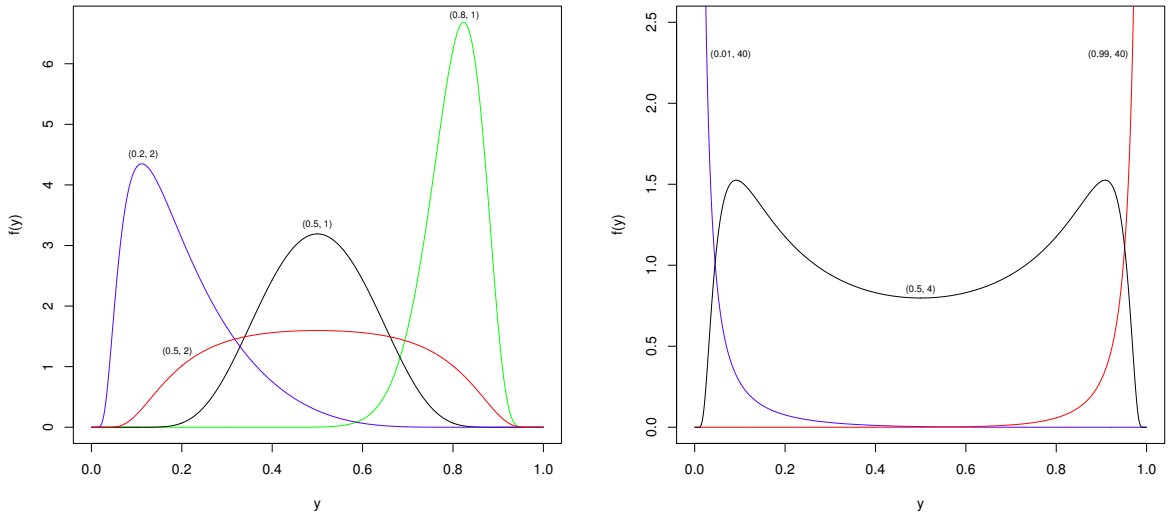


Figure 2.1: Simplex densities for different values of (μ, σ^2).

2.3 The Inflated Simplex Distribution

Rates and proportions may contain zeros and/or ones and in this case the simplex distribution is not suitable. In what follows, extensions of the simplex distribution are

presented. These extensions are divided into two groups: one suitable for modeling data in $[0, 1)$ and $(0, 1]$, and another for data in $[0, 1]$.

2.3.1 The Zero or One Inflated Simplex Distribution

If the data contain zeros or ones (but not both at the same time), the simplex distribution may be combined with a distribution degenerated at a known point c (where $c = 0$ or $c = 1$) in order to model the data. The support of the mixed distribution is $[0, 1)$ or $(0, 1]$ according to the value attributed to c. The probability density function of the mixture is defined as

$$\text{is}_c(y; \alpha, \mu, \sigma^2) = \begin{cases} \alpha, & y = c, \\ (1-\alpha) f(y; \mu, \sigma^2), & y \in (0, 1), \end{cases} \tag{2.6}$$

or, alternatively, as

$$\text{is}_c(y; \alpha, \mu, \sigma^2) = \left\{\alpha^{\mathbb{1}_{\{c\}}(y)}(1-\alpha)^{1-\mathbb{1}_{\{c\}}(y)}\right\} \left\{f(y; \mu, \sigma^2)^{1-\mathbb{1}_{\{c\}}(y)}\right\},$$

$\alpha \in (0, 1)$, $\mu \in (0, 1)$, $\sigma^2 > 0$. Here, $f(y; \mu, \sigma^2)$ is the probability density function of the simplex distribution $S^-(\mu, \sigma^2)$, which is given in (2.3), and $\mathbb{1}_{\{c\}}(y)$ is an indicator function that equals 1 if $y = c$ and 0 if $y \neq c$. The probability mass at c is α, i.e., the probability of observing $c = 0$ or $c = 1$ equals α is called the mixture parameter. Note that density (2.6) is the product of two terms: the first depends only on α and the second depends solely on (μ, σ^2).

Let y be a random variable following the inflated simplex distribution with probability density function (2.6). Depending on where inflation takes place (value of c), two nomenclatures can be defined for the distribution:

1. If $c = 0$, the distribution is called *zero-inflated simplex distribution* (ZIS), here denoted by $y \sim \text{ZIS}(\alpha, \mu, \sigma^2)$.

2. If $c = 1$, the distribution is called *one-inflated simplex distribution* (OIS), here denoted by $y \sim \text{OIS}(\alpha, \mu, \sigma^2)$.

Note that if $y \sim \text{ZIS}(\alpha, \mu, \sigma^2)$, then $\alpha = \Pr(y = 0)$ and if $y \sim \text{OIS}(\alpha, \mu, \sigma^2)$, then $\alpha = \Pr(y = 1)$.

The expected value and the variance of the zero-or-one inflated simplex distribution are obtained from the relationships

$$\text{E}(y) = \text{E}\left[\text{E}(y|\mathbb{1}_{\{c\}}(y))\right] \quad \text{and} \quad \text{Var}(y) = \text{E}\left[\text{Var}(y|\mathbb{1}_{\{c\}}(y))\right] + \text{Var}\left[\text{E}(y|\mathbb{1}_{\{c\}}(y))\right].$$

It can be shown that

$$\text{E}[y|\mathbb{1}_{\{c\}}(y)] = \begin{cases} c, & \text{with probability } \alpha, \\ \mu, & \text{with probability } 1-\alpha, \end{cases}$$

$$\text{Var}[y|\mathbb{1}_{\{c\}}(y)] = \begin{cases} 0, & \text{with probability } \alpha, \\ \text{Var}_{S^-}(y), & \text{with probability } 1-\alpha. \end{cases}$$

$\text{Var}_{S^-}(y)$ is the variance of the distribution $S^-(\mu, \sigma^2)$ given in (2.5). Therefore, the mean and variance of the ZIS distribution are

$$\text{E}(y) = (1-\alpha)\mu \quad \text{and} \quad \text{Var}(y) = (1-\alpha)\text{Var}_{S^-}(y) + \alpha(1-\alpha)\mu^2.$$

The mean and the variance of the OIS distribution are

$$\text{E}(y) = \alpha + (1-\alpha)\mu \quad \text{and} \quad \text{Var}(y) = (1-\alpha)\text{Var}_{S^-}(y) + \alpha(1-\alpha)(1-\mu)^2.$$

2.3.2 The Zero and One Inflated Simplex Distribution

The distribution presented in Subsection 2.3.1 is not suitable for modeling variates that assume values in $[0, 1]$. For such variables, the *zero and one inflated simplex dis-*

tribution (ZOIS) is suitable (Bandyopadhyay et al. 2014). The notation used is $y \sim \text{ZOIS}(\delta_0, \delta_1, \mu, \sigma^2)$. The probability density function of the ZOIS distribution is

$$\text{zois}(y; \delta_0, \delta_1, \mu, \sigma^2) = \begin{cases} \delta_0, & y = 0, \\ (1 - \delta_0 - \delta_1) f(y; \mu, \sigma^2), & y \in (0, 1), \\ \delta_1, & y = 1, \end{cases} \tag{2.7}$$

$0 < \delta_0 + \delta_1 < 1$, $\mu \in (0, 1)$ and $\sigma^2 > 0$. Here, $f(y; \mu, \sigma^2)$ is the simplex density function given in (2.3). Note that $\Pr(y = 0) = \delta_0$ and $\Pr(y = 1) = \delta_1$. When $\delta_0 + \delta_1 \to 1$, the ZOIS distribution tends to concentrate its probability mass in the extreme values of the interval [0,1]. When $\delta_0 + \delta_1 \to 0$, the ZOIS distribution tends to the simplex distribution.

The expected value and variance of the zero-and-one inflated simplex distribution are obtained from the relationship

$$\text{E}(y) = \text{E}\left[\text{E}(y|\mathbb{1}_{\{0,1\}}(y))\right] \quad \text{and} \quad \text{Var}(y) = \text{E}\left[\text{Var}(y|\mathbb{1}_{\{0,1\}}(y))\right] + \text{Var}\left[\text{E}(y|\mathbb{1}_{\{0,1\}}(y))\right].$$

It follows that

$$\text{E}[y|\mathbb{1}_{\{0,1\}}(y)] = \begin{cases} \frac{\delta_1}{\delta_0 + \delta_1}, & \text{with probability } \delta_0 + \delta_1, \\ \mu, & \text{with probability } 1 - \delta_0 - \delta_1, \end{cases}$$

$$\text{Var}[y|\mathbb{1}_{\{0,1\}}(y)] = \begin{cases} \frac{\delta_0 \delta_1}{(\delta_0 + \delta_1)^2}, & \text{with probability } \delta_0 + \delta_1, \\ \text{Var}_{S^-}(y), & \text{with probability } 1 - \delta_0 - \delta_1. \end{cases}$$

Here, $\text{Var}_{S^-}(y)$ is the variance of the simplex distribution $S^-(\mu, \sigma^2)$ given in (2.5). It can be shown, after some algebra, that the mean and variance of the ZOIS distribution are

$$\text{E}(y) = \delta_1 + (1 - \delta_0 - \delta_1)\mu$$

and

$$\mathrm{Var}(y) = \delta_1(1-\delta_1) + (1-\delta_0-\delta_1)\left[\mathrm{Var}_{S^-}(y) - 2\delta_1\mu + (\delta_0+\delta_1)\mu^2\right].$$

2.4 The Zero or One Inflated Simplex Regression Model

Let $y_1, \ldots, y_n$ be independent random variables, each following an inflated simplex distribution at point c ($c = 0$ or $c = 1$). Then, for $t = 1, \ldots, n$, y_t has p.d.f. given in (2.6). The simplex regression model inflated at point c ($c = 0$ or $c = 1$), denoted by IS-RE$_c$, is defined by the relationships

$$\begin{aligned} h(\alpha_t) &= \sum_{i=1}^{M} z_{ti}\varphi_i = z_t^\top \varphi = \zeta_t, \\ g(\mu_t) &= \sum_{i=1}^{m} x_{ti}\beta_i = x_t^\top \beta = \eta_t, \end{aligned} \tag{2.8}$$

where $\varphi = (\varphi_1, \ldots, \varphi_M)^\top$ and $\beta = (\beta_1, \ldots, \beta_m)^\top$ are unknown parameter vectors such that $\varphi \in \mathbb{R}^M$ and $\beta \in \mathbb{R}^m$; $z_t = (z_{t1}, \ldots, z_{tM})^\top$ and $x_t = (x_{t1}, \ldots, x_{tm})^\top$ are known observations on covariates; $M + m < n$. The z's and x's can coincide partially or completely. The link functions $h : (0,1) \to \mathbb{R}$ and $g : (0,1) \to \mathbb{R}$ are strictly monotone and twice differentiable. Commonly used link functions are logit, probit, log-log, Cauchy and complementary log-log, among others.

Note that c is fixed for all observations and $\alpha_t = \Pr(y_t = c)$. The parameters μ_t and σ^2 are the conditional mean and the dispersion parameter of y_t, for $y_t \in (0,1)$. The parameter σ^2 (as well as the other parameters) may be considered constant or be regressed onto some covariates. In what follows, we shall assume that σ^2 is constant. The zero inflated simplex regression model ($c = 0$) will be called ZIS-RE and the one inflated simplex regression model ($c = 1$) will be denoted by OIS-RE.

2.4.1 Likelihood Inference

Consider the parameter vector $\theta = (\varphi^\top, \beta^\top, \sigma^2)^\top$. The log-likelihood function of the simplex regression model inflated at point c is

$$L(\theta) = \prod_{t=1}^{n} \text{is}_c(y_t; \alpha_t, \mu_t, \sigma^2) = L_1(\varphi) L_2(\beta, \sigma^2), \tag{2.9}$$

where $\text{is}_c(\cdot;\cdot,\cdot,\cdot)$ is the probability density function defined in Equation (2.6) and

$$L_1(\varphi) = \prod_{t=1}^{n} \alpha_t^{\mathbb{1}_{\{c\}}(y_t)} (1-\alpha_t)^{1-\mathbb{1}_{\{c\}}(y_t)},$$
$$L_2(\beta, \sigma^2) = \prod_{t:\, y_t \in (0,1)} f(y_t; \mu_t, \sigma^2).$$

The parameters α_t and μ_t are defined as functions of φ and β, respectively, in Equation (2.8), i.e., $\alpha_t = h^{-1}(\zeta_t)$ and $\mu_t = g^{-1}(\eta_t)$. Note that the likelihood function factors into two terms: one depending only on φ and another one depending only on $(\beta^\top, \sigma^2)^\top$. Therefore, the parameter vectors are separable (Pace & Salvan 1997, p.128) and maximum likelihood inference on $(\beta^\top, \sigma^2)^\top$ can be performed as if φ were known and vice-versa. Note that the component $L_1(\varphi)$ only involves the parameters used to model the probability of observing zero or observing one. On the other hand, $L_2(\beta, \sigma^2)$ only involves the parameters used to model the conditional distribution of the response variable on the interval (0,1).

The logarithm of the likelihood function for $\theta = (\varphi^\top, \beta^\top, \sigma^2)^\top$ is

$$\ell(\theta) = \log[L(\theta)] = \ell_1(\varphi) + \ell_2(\beta, \sigma^2), \tag{2.10}$$

where

$$\ell_1(\varphi) = \sum_{t=1}^{n} \ell_t(\alpha_t),$$
$$\ell_2(\beta, \sigma^2) = \sum_{t:y_t \in (0,1)} \ell_t(\mu_t, \sigma^2).$$

Here,

$$\ell_t(\alpha_t) = \mathbb{1}_{\{c\}}(y_t)\log(\alpha_t) + (1 - \mathbb{1}_{\{c\}}(y_t))\log(1-\alpha_t),$$
$$\ell_t(\mu_t, \sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{3}{2}\log[y_t(1-y_t)] - \frac{1}{2\sigma^2}d(y_t;\mu_t),$$

where $c = 0$ or $c = 1$, depending on the case; $d(y_t;\mu_t)$ is as defined in (2.4). For $t = 1,\ldots,n$, the random variable $\mathbb{1}_{\{c\}}(y_t)$ is Bernoulli distributed with parameter α_t. Note that $\alpha_t = P\left(\mathbb{1}_{\{c\}}(y_t) = 1\right)$. The parameter α_t is associated to the linear predictor ζ_t (which includes regressors and parameters) through a link function $h(\cdot)$, as defined in (2.8). Therefore, $\ell_1(\varphi)$ is the log-likelihood function of a generalized linear model with a binary response. Further details can be found in McCullagh & Nelder (1989). Additionally, $\ell_2(\beta,\sigma^2)$ is the log-likelihood function of a simplex regression model with response restricted to the interval (0,1).

The score function is obtained from differentiation of the log-likelihood function with respect to each unknown parameter. The score function for φ_R can be written as

$$U_{\varphi_R} = Z^\top PG(y^c - \alpha^*), \tag{2.11}$$

where Z is a matrix of dimension $n \times M$, whose tth line is $z_t^\top$ and the other elements are $P = \text{diag}\{1/[\alpha_1(1-\alpha_1)],\ldots,1/[\alpha_n(1-\alpha_n)]\}$, $G = \text{diag}\{1/h'(\alpha_1),\ldots,1/h'(\alpha_n)\}$, $y^c = (\mathbb{1}_{\{c\}}(y_1),\cdots,\mathbb{1}_{\{c\}}(y_1))^\top$ and $\alpha^* = (\alpha_1,\ldots,\alpha_n)^\top$.

The score function for β_r can be expressed as

$$U_{\beta_r} = \sigma^{-2}X^\top THu, \tag{2.12}$$

where X is an $n \times m$ matrix, whose tth line is $x_t^\top$, $T = \text{diag}\{1/g'(\mu_1),\ldots,1/g'(\mu_n)\}$, $H = \text{diag}\{1-\mathbb{1}_{\{c\}}(y_1),\cdots,1-\mathbb{1}_{\{c\}}(y_n)\}$, $u^\top = (u_1,\ldots,u_n)^\top$ and

$$u_t = \frac{(y_t-\mu_t)(y_t - 2\mu_t y_t + \mu_t^2)}{y_t(1-y_t)\mu_t^3(1-\mu)^3},$$

The score function U_{σ^2} is given by

$$U_{\sigma^2} = \mathrm{tr}(H\, D^*), \tag{2.13}$$

where $D^* = \mathrm{diag}\{d_1^*, \ldots, d_n^*\}$, with $d_t^* = -1/(2\sigma^2) + [1/(2\sigma^4)]d(y_t; \mu_t)$, and $\mathrm{tr}(\cdot)$ is the trace of a square matrix.

Details on how the first order derivatives of the logarithm of the likelihood function (2.10) were obtained can be found in Appendix A (section A.1).

In order to obtain Fisher's information matrix, similarly to the work of Ospina & Ferrari (2012), we calculate the moments and the cumulants of the second order derivatives of the log-likelihood function (2.10) (see Appendix A.2 for details). Fisher's information matrix for the simplex regression model inflated at point c ($c = 0$ or $c = 1$) has the form

$$K(\theta) = \begin{pmatrix} K_{\varphi\varphi} & 0 & 0 \\ 0 & K_{\beta\beta} & 0 \\ 0 & 0 & K_{\sigma^2\sigma^2} \end{pmatrix}, \tag{2.14}$$

where $K_{\varphi\varphi} = Z^\top Q Z$, $K_{\beta\beta} = -\sigma^{-2} X^\top \Delta A X$ and $K_{\sigma^2\sigma^2} = \mathrm{tr}(\Delta D)$, $Q = \mathrm{diag}\{q_1, \ldots, q_n\}$, $A = \mathrm{diag}\{a_1, \ldots, a_n\}$, $\Delta = \mathrm{diag}\{\delta_1, \ldots, \delta_n\}$, $D = \mathrm{diag}\{d_1, \ldots, d_n\}$. For $t = 1, \ldots, n$, $q_t = -p_t[1/h'(\alpha_t)]^2$, $p_t = 1/[\alpha_t(1-\alpha_t)]$, $\delta_t = 1 - \alpha_t$, $d_t = -1/(2\sigma^4)$ and

$$a_t = \left(\frac{3\sigma^2}{\mu_t(1-\mu_t)} + \frac{1}{\mu_t 3(1-\mu_t)^3}\right)\left(\frac{1}{g'(\mu_t)}\right)^2.$$

It is noteworthy that the matrix in (2.14) does not depend on the inflation point c.

The inverse of Fisher's information matrix is

$$K(\theta)^{-1} = \begin{pmatrix} K^{\varphi\varphi} & 0 & 0 \\ 0 & K^{\beta\beta} & 0 \\ 0 & 0 & K^{\sigma^2\sigma^2} \end{pmatrix}, \tag{2.15}$$

where $K^{\varphi\varphi} = (Z^\top QZ)^{-1}$, $K^{\beta\beta} = -\sigma^2(X^\top \Delta AX)^{-1}$ and $K^{\sigma^2\sigma^2} = [\mathrm{tr}(\Delta D)]^{-1}$.

Because of the separability of φ and $(\beta^\top, \sigma^2)^\top$, the maximum likelihood estimators (MLE) are obtained separately. The MLE of φ is obtained as solution of the nonlinear system $U_\varphi = 0$ whereas the MLE of $(\beta^\top, \sigma^2)^\top$ is the solution of the nonlinear system $(U_\beta^\top, U_{\sigma^2})^\top = 0$. Only the estimator for σ^2 can be expressed in closed form. Therefore, the log-likelihood function (2.10) need to be numerically maximized when obtaining estimatives of φ and β. Nonlinear optimization can be carried out using the Newton-Raphson method, Fisher's method of scoring or the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm; see Press et al. (1992) for further details.

2.4.2 Estimation Process

We shall now present an estimation procedure of the parameters that index IS-RE$_c$ model. The estimation process using Fisher's method of scoring depends on the score function and on Fisher's information matrix. The MLE of φ is obtained using the following iterative mechanism:

$$\begin{aligned} \varphi^{(m+1)} &= \varphi^{(m)} + (Z^\top Q^{(m)} Z)^{-1} Z^\top P^{(m)} G^{(m)} (y^c - \alpha^{*(m)}) \\ &= (Z^\top Q^{(m)} Z)^{-1} Z^\top Q^{(m)} \tau_1^{(m)}, \end{aligned} \tag{2.16}$$

where $\tau_1^{(m)} = Z\varphi^{(m)} + (Q^{(m)})^{-1} P^{(m)} G^{(m)} (y^c - \alpha^{*(m)})$ and $m = 0, 1, 2, \ldots$.

The scoring iterative scheme used to obtain the MLE of β is expressed by

$$\begin{aligned} \beta^{(m+1)} &= \beta^{(m)} + (X^\top \Delta^{(m)} A^{(m)} X)^{-1} X^\top T^{(m)} H^{(m)} u^{(m)} \\ &= (X^\top \Delta^{(m)} A^{(m)} X)^{-1} X^\top \Delta^{(m)} A^{(m)} \tau_2^{(m)}, \end{aligned} \tag{2.17}$$

where $\tau_2^{(m)} = X\beta^{(m)} + (\Delta^{(m)} A^{(m)})^{-1} T^{(m)} H^{(m)} u^{(m)}$.

The iterations in (2.16) and (2.17) are carried on until the distances between $\varphi^{(m+1)}$ and $\varphi^{(m)}$ and between $\beta^{(m+1)}$ and $\beta^{(m)}$ are smaller than a specified tolerance. The distance used may be the Euclidean distance.

The MLE of σ^2 has closed-form. From $U_{\sigma^2} = 0$, we obtain

$$\widehat{\sigma^2} = \frac{\sum_{t=1}^{n} \left(1 - \mathbb{1}_{\{c\}}(y_t)\right) d(y_t; \mu_t)}{n - \sum_{t=1}^{n} \mathbb{1}_{\{c\}}(y_t)}.$$

2.4.3 Confidence Interval and Hypothesis Tests

Under mild regularity conditions (Cox & Hinkley 1974, Sen & Singer 1993), $\widehat{\theta}$ and $K(\widehat{\theta})$ are consistent for θ and $K(\theta)$; here, $K(\widehat{\theta})$ denotes Fisher's information matrix evaluated at $\widehat{\theta}$. Let $J(\theta) = \lim_{n\to\infty} K(\theta)/n$ be nonsingular. Then

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}_{M+m+1}(0, J(\theta)^{-1}),$$

where $\widehat{\theta} = (\widehat{\varphi}^\top, \widehat{\beta}^\top, \widehat{\sigma^2})^\top$ is the maximum likelihood estimator of $\theta = (\varphi^\top, \beta^\top, \sigma^2)^\top$, N_{M+m+1} is the multivariate normal distribution of dimension $M+m+1$ and $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. Using the asymptotic normality of the MLE $\widehat{\theta}$, asymptotic confidence intervals can be easily obtained for the parameters that index the simplex regression model inflated at c.

For a confidence level of $100(1-\nu)\%$, $\nu \in (0, 0.5)$, the asymptotic confidence intervals for φ_R, β_r and σ^2 are, respectively,

$$\left(\widehat{\varphi}_R - z_{1-\nu/2}\left(\widehat{K}^{\varphi\varphi}_{RR}\right)^{1/2},\ \ \widehat{\varphi}_R + z_{1-\nu/2}\left(\widehat{K}^{\varphi\varphi}_{RR}\right)^{1/2}\right),$$

for $R = 1, \ldots, M$,

$$\left(\widehat{\beta}_r - z_{1-\nu/2}\left(\widehat{K}^{\beta\beta}_{rr}\right)^{1/2},\ \ \widehat{\beta}_r + z_{1-\nu/2}\left(\widehat{K}^{\beta\beta}_{rr}\right)^{1/2}\right),$$

for $r = 1, \ldots, m$, and, finally,

$$\left(\widehat{\sigma^2} - z_{1-\nu/2}\left(\widehat{K}^{\sigma^2\sigma^2}\right)^{1/2},\ \ \widehat{\sigma^2} + z_{1-\nu/2}\left(\widehat{K}^{\sigma^2\sigma^2}\right)^{1/2}\right).$$

The estimated asymptotic variances of $\widehat{\varphi}_R$, $\widehat{\beta}_r$ and $\widehat{\sigma}^2$ are $\widehat{K}^{\varphi\varphi}_{RR}$, $\widehat{K}^{\beta\beta}_{rr}$ and $\widehat{K}^{\sigma^2\sigma^2}$, respectively. Here, $\widehat{K}^{\varphi\varphi}_{RR}$ is the element (R, R) of the matrix $K^{\varphi\varphi}$ evaluated at $\widehat{\varphi}$; $\widehat{K}^{\beta\beta}_{rr}$ is the element (r, r) of the matrix $K^{\beta\beta}$ evaluated at $\widehat{\beta}$ and $\widehat{K}^{\sigma^2\sigma^2}$ is the element $K^{\sigma^2\sigma^2}$ evaluated at $\widehat{\sigma}^2$. In the above confidence intervals, $z_{1-\nu/2}$ denotes the $1-\nu/2$ standard normal quantile.

Using the multivariate delta method (Lehmann & Casella 2002), the asymptotic confidence interval for the mean response $\mu_t^\circ = \mathrm{E}(y_t)$, $t = 1, \ldots, n$, of the model IS-RE$_c$ with $100(1-\nu)\%$ of confidence can be shown to be

$$\left(\widehat{\mu}_t^\circ - z_{1-\nu/2}\ \mathrm{se}(\widehat{\mu}_t^\circ),\ \ \widehat{\mu}_t^\circ + z_{1-\nu/2}\ \mathrm{se}(\widehat{\mu}_t^\circ)\right),$$

where

$$\widehat{\mu}_t^\circ = c\,\widehat{\alpha}_t + (1-\widehat{\alpha}_t)\widehat{\mu}_t = c\,h^{-1}(\widehat{\zeta}_t) + (1-h^{-1}(\widehat{\zeta}_t))g^{-1}(\widehat{\eta}_t)$$

and

$$\mathrm{se}(\widehat{\mu}_t^\circ) = \sqrt{\left(\frac{1-\widehat{\mu}_t}{h(\widehat{\zeta}_t)}\right)^2 z_t^\top \widehat{K}^{\varphi\varphi}\, z_t + \left(\frac{1-\widehat{\alpha}_t}{g(\widehat{\eta}_t)}\right)^2 x_t^\top \widehat{K}^{\beta\beta}\, x_t},$$

where $\widehat{K}^{\varphi\varphi}$ and $\widehat{K}^{\beta\beta}$ are the elements $K^{\varphi\varphi}$ and $K^{\beta\beta}$ of the inverse of Fisher's information matrix (2.15) evaluated at $\widehat{\varphi}$ and $\widehat{\beta}$.

Hypothesis testing inference can also be easily performed. Suppose the interest lies in testing a subset of the parameter vectors φ and β. We partition φ and β as $\varphi = (\varphi_1^\top, \varphi_2^\top)^\top$ and $\beta = (\beta_1^\top, \beta_2^\top)^\top$, where $\varphi_1 = (\varphi_1, \ldots, \varphi_{M_1})^\top$, $\varphi_2 = (\varphi_{M+1}, \ldots, \varphi_M)^\top$, $\beta = (\beta, \ldots, \beta_{m_1})^\top$ and $\beta_2 = (\beta_{m_1+1}, \ldots, \beta_m)^\top$. Here, $0 < M_1 \leq M$ and $0 < m_1 \leq m$. Suppose we wish to test $\mathcal{H}_0 : \varphi_1 = \varphi_1^{(0)}; \beta_1 = \beta_1^{(0)}$ against $\mathcal{H}_1$: violation of at least one equality. Here, $\varphi_1^{(0)}$ and $\beta_1^{(0)}$ are vectors of dimension M_1 and m_1, respectively.

The hypothesis can be tested using the log-likelihood ratio test statistic, which is given by

$$\Lambda = 2\{\ell(\widehat{\varphi}, \widehat{\beta}, \widehat{\sigma}^2) - \ell(\widetilde{\varphi}, \widetilde{\beta}, \widetilde{\sigma}^2)\},$$

where $\ell(\varphi, \beta, \sigma^2)$ is the log-likelihood function given in (2.10) and $(\widehat{\varphi}, \widehat{\beta}, \widehat{\sigma^2})$ and $(\widetilde{\varphi}, \widetilde{\beta}, \widetilde{\sigma^2})$

are the unrestricted and restricted (under hypothesis $\mathcal{H}_0$) MLEs of $(\varphi, \beta, \sigma^2)^\top$, respectively. Under some regularity conditions, Λ is asymptotically distributed as χ^2 with $M_1 + m_1$ degrees of freedom under the null hypothesis $\mathcal{H}_0$, i.e., $\Lambda \xrightarrow{\mathcal{D}} \chi^2_{M_1+m_1}$. The test can be performed using critical values from $\chi^2_{M_1+m_1}$.

An alternative to the likelihood ratio test is the score test. Let Z_1, Z_2, X_1 and X_2 be full rank matrices of dimension $n \times M_1$, $n \times (M - M_1)$, $n \times m_1$ and $n \times (m - m_1)$, respectively. From the null hypothesis $\mathcal{H}_0 : \varphi_1 = \varphi_1^{(0)}; \beta_1 = \beta_1^{(0)}$, we can define the matrices of regressors Z and X as partitioned matrices $Z = [Z_1, Z_2]$ and $X = [X_1, X_2]$. Here, if $M_1 = M$ we define $Z_1 = Z$ and if $m_1 = m$ we have $X_1 = X$. Let $U_{1\varphi}$ be the vector of dimension M_1 containing the first M_1 elements of score vector U_φ and let $K_{11}^{\varphi\varphi}$ be the matrix containing the first M_1 lines and the first M_1 columns of the matrix $K^{\varphi\varphi}$, defined in (2.15). In similar fashion, let $U_{1\beta}$ be the vector of dimension m_1 containing the first m_1 elements of the score vector U_β and $K_{11}^{\beta\beta}$ be the matrix of dimension $m_1 \times m_1$ formed using the first m_1 lines and the first m_1 columns of $K^{\beta\beta}$, which is defined in (2.15). Therefore, $U_{1\varphi} = Z_1^\top PG(y^c - \alpha^*)$ and $U_{1\beta} = \sigma^{-2} X_1^\top THu$. The score statistic ξ can be written as

$$\xi = \widetilde{U}_{1\varphi}^\top \widetilde{K}_{11}^{\varphi\varphi} \widetilde{U}_{1\varphi} + \widetilde{U}_{1\beta}^\top \widetilde{K}_{11}^{\beta\beta} \widetilde{U}_{1\beta},$$

where "$\widetilde{\ }$" indicates that the quantities are evaluated at the restricted MLEs. Under some the regularity conditions and the null hypothesis, the score statistic is asymptotically distributed as $\chi^2_{M_1+m_1}$.

The null hypothesis $\mathcal{H}_0$ can also be tested using the Wald test. The Wald statistic is given by

$$\varpi = \left(\widehat{\varphi}_1 - \widehat{\varphi}_1^{(0)}\right)^\top \left(\widehat{K}_{11}^{\varphi\varphi}\right)^{-1} \left(\widehat{\varphi}_1 - \widehat{\varphi}_1^{(0)}\right) + \left(\widehat{\beta}_1 - \widehat{\beta}_1^{(0)}\right)^\top \left(\widehat{K}_{11}^{\beta\beta}\right)^{-1} \left(\widehat{\beta}_1 - \widehat{\beta}_1^{(0)}\right),$$

where "$\widehat{\ }$" indicates the quantities evaluated at the unrestricted MLEs. Under the null hypothesis and under some regularity conditions, $\varpi \xrightarrow{\mathcal{D}} \chi^2_{M_1+m_1}$. The test can thus be performed using critical values obtained from the χ^2 distribution with $M_1 + m_1$ degrees

of freedom. In particular, the significance of the Rth parameter φ_R, $R = 1, \ldots, M$, can be tested using the square root of the Wald statistic, i.e., $\widehat{\varphi}_R/\mathrm{se}(\widehat{\varphi}_R)$, where $\mathrm{se}(\widehat{\varphi}_R)$ is the standard error of $\widehat{\varphi}_R$. Note that $\mathrm{se}(\widehat{\varphi}_R)$ is the square root ot the (R, R) element of $K^{\varphi\varphi}$ evaluated at the MLE. The asymptotic distribution of $\widehat{\varphi}_R/\mathrm{se}(\widehat{\varphi}_R)$ under the null hypothesis is standard normal, $\mathcal{N}(0, 1)$. Hypothesis testing inferences on β_r, $r = 1, \ldots, m$, can be performed analogously.

2.4.4 Application to simulated data

We fitted the zero-inflated simplex regression (ZIS-RE) model to a simulated dataset. To the end, we used the `gamlss` and `simplexreg` `R` (R Development Core Team 2011) packages. The source code is included in Appendix A.6. We consider the model with the following structure:

$$g(\mu_t) = \beta_0 + \beta_1 x_t,$$
$$h(\alpha_t) = \varphi_0 + \varphi_1 z_t,$$

$t = 1, \ldots, n$, where g and h are the logit link functions. The values of x_t and z_t were obtained as standard uniform random draws. The true values of the parameters are $\beta_0 = -1.5$, $\beta_1 = 1.5$, $\varphi_0 = -1$, $\varphi_1 = 0.5$ and $\sigma^2 = 1$.

In our example, 32.6% of the values of the response variable are equal to zero. Dispersion diagrams of the response variable y_t against x_t and z_t are displayed in Figure 2.2. Maximum likelihood estimates and their respective standard errors can be found in Table 2.1. Notice that all regressors are significant at the useful significance levels. We point out that the MLEs of the parameters that index the submodel for α_t are biased in small samples. Preliminary numerical experiments suggest that such estimators only become nearly unbiased when $n \geq 400$. Further details on the simulated data can be found in Appendix A.6.

Table 2.1: Maximum likelihood estimatives and standard errors for the simulated data in ZIS-RE.

Parameter	β_0	β_1	φ_0	φ_1	σ^2
Estimate	-1.5135	1.4989	-1.0066	0.5522	0.9840
Std. error	0.0407	0.0748	0.1959	0.3317	0.0379

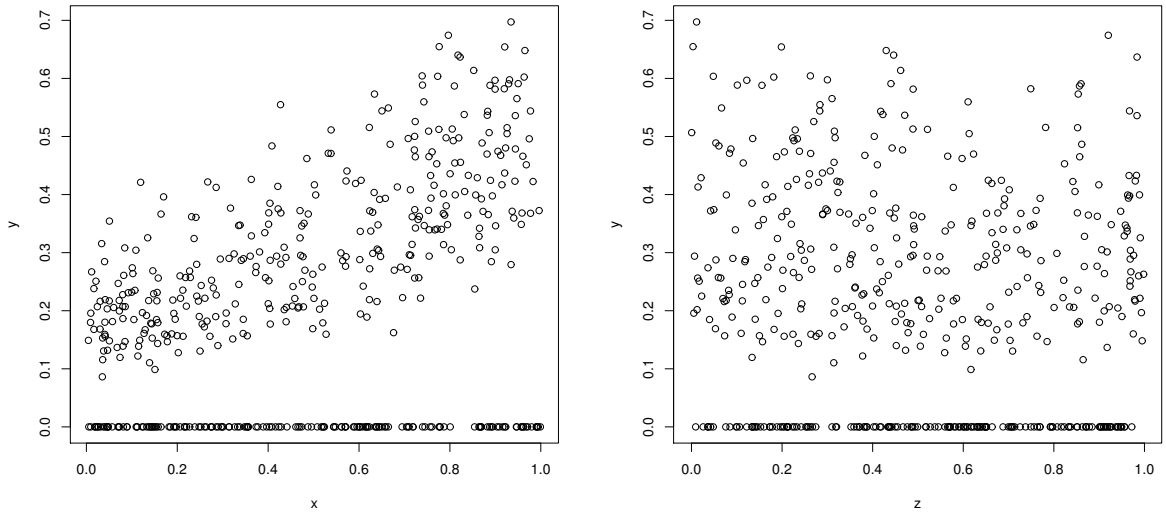


Figure 2.2: Dispersion diagrams of the response y_t against x_t (left) and z_t (right).

2.5 The Zero and One Inflated Simplex Regression Model

Let $y_1, \ldots, y_n$ be independent random variables with probability density function given by (2.7), i.e., $y_t \sim \text{ZOIS}(\delta_0, \delta_1, \mu_t, \sigma^2)$. The zero and one inflated simplex regression model (ZOIS-RE) is defined by

$$\begin{aligned} g(\mu_t) &= \sum_{i=1}^{m} x_{ti}\beta_i = x_t^\top \beta = \eta_t, \\ H(\delta_{0t}, \delta_{1t}) &= (h_0(\delta_{0t}, \delta_{1t}), h_1(\delta_{0t}, \delta_{1t})) = \left(\nu_t^\top \rho,\, z_t^\top \varphi\right) = (\zeta_{0t}, \zeta_{1t}), \end{aligned} \tag{2.18}$$

where $\mu_t = \text{E}(y_t | y_t \in (0,1))$, $\delta_{0t} = P(y_t = 0)$, $\delta_{1t} = P(y_t = 1)$ and $1 - \delta_{0t} - \delta_{1t} = P(y_t \in (0,1))$; η_t, ζ_{0t} and ζ_{1t} are linear predictors; $\beta = (\beta_1, \ldots, \beta_k)^\top$, $\rho = (\rho, \ldots, \rho_{k_0})^\top$

and $\varphi = (\varphi_1, \ldots, \varphi_{k_1})^\top$ are vectors of unknown parameters, where $\beta \in \mathbb{R}^k$, $\rho \in \mathbb{R}^{k_0}$ and $\varphi \in \mathbb{R}^{k_1}$. Here, $x_t = (x_{t1}, \ldots, x_{tk})^\top$, $\nu_t = (\nu_{t1}, \ldots, \nu_{tk_0})^\top$ and $z_t = (z_{t1}, \ldots, z_{tk_1})^\top$ are vectors of regressors of dimension k, k_0 and k_1, respectively.

The link function $g : (0, 1) \to \mathbb{R}$ is strictly monotone and twice differentiable. The function H is a bijective from the set $\mathcal{C} = \{(\delta_{0t}, \delta_{1t}) : 0 < \delta_{0t} < 1,\, 0 < \delta_{1t} < 1 - \delta_{0t}\}$ to $\mathbb{R}$. H is also twice differentiable. The partial derivatives of $\delta_{0t} = h_0^*(\zeta_{0t}, \zeta_{1t})$ and $\delta_{1t} = h_1^*(\zeta_{0t}, \zeta_{1t})$ are continuous in $\mathbb{R}^2$ and δ_{0t} and δ_{1t} can be uniquely expressed in terms of ζ_{0t} and ζ_{1t} (Rudin 1976).

2.5.1 Likelihood Inference

Let $\theta = (\rho^\top, \varphi^\top, \beta^\top, \sigma^2)^\top$ be the vector of unknown parameters. The likelihood function for the zero-and-one inflated simplex distribution is given by

$$L(\theta) = \prod_{t=1}^{n} \text{zois}(y; \delta_0, \delta_1, \mu, \sigma^2) = L_1(\rho, \varphi) L_2(\beta, \sigma^2), \tag{2.19}$$

where $\text{zois}(\cdot; \cdot, \cdot, \cdot, \cdot)$ is the probability density function of the zero and one inflated simplex distribution defined in (2.7) and

$$L_1(\rho, \varphi) = \prod_{t=1}^{n} \delta_{0t}^{\mathbb{1}_{\{0\}}(y_t)} \delta_{1t}^{\mathbb{1}_{\{1\}}(y_t)} (1 - \delta_{0t} - \delta_{1t})^{1 - \mathbb{1}_{\{0\}}(y_t) - \mathbb{1}_{\{1\}}(y_t)},$$

$$L_2(\beta, \sigma^2) = \prod_{t : y_t \in (0,1)} f(y_t; \mu_t, \sigma^2),$$

where δ_{0t}, δ_{1t} and μ_t are functions of the parameters ρ, φ and β, respectively, through (2.18). The function $f(y_t \cdot, \cdot)$ is the simplex density defined in (2.3).

The function $L(\theta)$ factors into two terms: one depending solely on $(\rho^\top, \varphi^\top)^\top$ and another one depending just on $(\beta^\top, \sigma^2)^\top$. The parameters are thus separable (Pace & Salvan 1997). It follows that, maximum likelihood inference on $(\rho^\top, \varphi^\top)^\top$ can be performed as if $(\beta^\top, \sigma^2)^\top$ were known and vice-versa. Note that $L_1(\rho, \varphi)$ involves only the parameters used to model the probability of occurrence of zeros and ones (discrete component). The

likelihood function $L_2(\beta^\top, \sigma^2)$ only involves the parameters used to model the distribution of the response in the interval $(0, 1)$ – the continuous component.

Using (2.19), the log-likelihood function for $\theta = (\rho^\top, \varphi^\top, \beta^\top, \sigma^2)^\top$ is given by

$$\ell(\theta) = \log[L(\theta)] = \ell_1(\rho, \varphi) + \ell_2(\beta, \sigma^2), \tag{2.20}$$

where

$$\ell_1(\rho, \varphi) = \sum_{t=1}^{n} \ell_t(\delta_{0t}, \delta_{1t}),$$
$$\ell_2(\beta, \sigma^2) = \sum_{t:y_t \in (0,1)} \ell_t(\mu_t, \sigma^2).$$

Here,

$$\begin{aligned}\ell_t(\delta_{0t}, \delta_{1t}) &= \mathbb{1}_{\{0\}}(y_t)\log(\delta_{0t}) + \mathbb{1}_{\{1\}}(y_t)\log(\delta_{1t}) \\ &\quad + [1 - \mathbb{1}_{\{0\}}(y_t) - \mathbb{1}_{\{1\}}(y_t)]\log(1 - \delta_{0t} - \delta_{1t}), \\ \ell_t(\mu_t, \sigma^2) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{3}{2}\log[y_t(1 - y_t)] - \frac{1}{2\sigma^2}d(y_t; \mu_t).\end{aligned}$$

Note that the function $\ell_2(\beta, \sigma^2)$ in (2.20) is the same as the one given in (2.10). It occurs because the simplex distribution is used to model the continuous component of the response, i.e., the part of the response that lies in the interval $(0, 1)$.

By differentiating the log-likelihood function in (2.20) with respect to each unknown parameter we obtain the score function. The score vector for $(\rho^\top, \varphi^\top)^\top$ is

$$U(\rho^\top, \varphi^\top)^\top = \left(U_\rho(\rho, \varphi)^\top, U_\varphi(\rho, \varphi)^\top\right)^\top,$$

where

$$\begin{aligned}U_\rho(\rho, \varphi) &= V^\top T_0 \left(\Delta_0 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)}\right) + V^\top T_{10} \left(\Delta_1 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)}\right), \\ U_\varphi(\rho, \varphi) &= Z^\top T_{01} \left(\Delta_0 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)}\right) + Z^\top T_1 \left(\Delta_1 y_{\{1\}} - \Delta_{(0,1)} y_{(0,1)}\right),\end{aligned} \tag{2.21}$$

where V is a $n \times k_0$ matrix of regressors whose tth line is ν_t and Z is a $n \times k_1$ matrix of regressors whose tth line is z_t. Furthermore,

$$
\begin{aligned}
T_0 &= \mathrm{diag}\{\partial\delta_{01}/\partial\zeta_{01}, \ldots, \partial\delta_{0n}/\partial\zeta_{0n}\},\\
T_1 &= \mathrm{diag}\{\partial\delta_{11}/\partial\zeta_{11}, \ldots, \partial\delta_{1n}/\partial\zeta_{1n}\},\\
T_{01} &= \mathrm{diag}\{\partial\delta_{01}/\partial\zeta_{11}, \ldots, \partial\delta_{0n}/\partial\zeta_{1n}\},\\
T_{10} &= \mathrm{diag}\{\partial\delta_{11}/\partial\zeta_{01}, \ldots, \partial\delta_{1n}/\partial\zeta_{0n}\},\\
\Delta_0 &= \mathrm{diag}\{1/\delta_{01}, \ldots, 1/\delta_{0n}\},\\
\Delta_1 &= \mathrm{diag}\{1/\delta_{11}, \ldots, 1/\delta_{1n}\},\\
\Delta_{(0,1)} &= \mathrm{diag}\{1/(1-\delta_{01}-\delta_{11}), \ldots, 1/(1-\delta_{0n}-\delta_{1n})\},\\
y_{\{0\}} &= (\mathbb{1}_{\{0\}}(y_1), \ldots, \mathbb{1}_{\{0\}}(y_n))^\top,\\
y_{\{1\}} &= (\mathbb{1}_{\{1\}}(y_1), \ldots, \mathbb{1}_{\{1\}}(y_n))^\top,\\
y_{(0,1)} &= (\mathbb{1}_{(0,1)}(y_1), \ldots, \mathbb{1}_{(0,1)}(y_n))^\top.
\end{aligned}
$$

Note that T_0, T_1, T_{01}, T_{10}, δ_0, δ_1 and $\delta_{(0,1)}$ are diagonal matrices of dimension n. The vectors $y_{\{0\}}$, $y_{\{1\}}$ and $y_{(0,1)}$ are of dimension n.

We define the augmented matrices $\widetilde{T}$ and $\widetilde{Z}$, of dimension $2n \times 2n$ and $2n \times (k_0 + k_1)$, respectively, as

$$
\widetilde{T} = \begin{pmatrix} T_0 & T_{10} \\ T_{01} & T_1 \end{pmatrix}, \qquad \widetilde{Z} = \begin{pmatrix} V & 0 \\ 0 & Z \end{pmatrix}
$$

and the vector $y_\Delta^\top = ((\Delta_0 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)})^\top, (\Delta_1 y_{\{1\}} - \Delta_{(0,1)} y_{(0,1)})^\top)^\top$, the score vector for $(\rho^\top, \varphi^\top)^\top$ can now be written as

$$
U(\rho^\top, \varphi^\top)^\top = \widetilde{Z}^\top \widetilde{T}\, y_\Delta. \tag{2.22}
$$

The score vector for β is

$$
U_\beta(\beta, \sigma^2) = \sigma^{-2} X^\top \mathrm{diagm}(y_{(0,1)}) T\, u, \tag{2.23}
$$

where the the operator diagm$(\cdot)$ transforms a vector into a diagonal matrix; T and u are defined in (2.12). The score function of σ^2 is

$$U_{\sigma^2}(\beta, \sigma^2) = \text{tr}(\text{diagm}(y_{(0,1)})\, D^*), \tag{2.24}$$

where D^* is defined in Equation (2.13).

Details on how the first order derivatives of the log-likelihood function in (2.20) were obtained can be found in Appendix A (subsection A.3).

Fisher's information matrix for $\theta = (\rho^\top, \varphi^\top, \beta^\top, \sigma^2)^\top$, $K(\theta)$, for the zero and one inflated simplex regression model is

$$K(\theta) = \begin{pmatrix} K_{\rho\rho} & K_{\rho\varphi} & 0 & 0 \\ K_{\varphi\rho} & K_{\varphi\varphi} & 0 & 0 \\ 0 & 0 & K_{\beta\beta} & 0 \\ 0 & 0 & 0 & K_{\sigma^2\sigma^2} \end{pmatrix}, \tag{2.25}$$

where

$$\begin{aligned}
K_{\rho\rho} &= V^\top \left\{T_0^2\Delta_0 - \Delta_{(0,1)}(T_0 + T_{10})^2 + \Delta_1 T_{10}^2\right\} V, \\
K_{\varphi\varphi} &= Z^\top \left\{T_1^2\Delta_1 - \Delta_{(0,1)}(T_1 + T_{01})^2 + \Delta_0 T_{01}^2\right\} Z, \\
K_{\rho\varphi} &= K_{\varphi\rho}^\top = Z^\top \left\{T_0\Delta_0 T_{01} - (T_0 + T_{10})\Delta_{(0,1)}(T_1 + T_{01}) + T_1\Delta_1 T_{10}\right\} V, \\
K_{\beta\beta} &= -\sigma^2 X^\top \Delta_{(0,1)}^{-1} A X, \\
K_{\sigma^2\sigma^2} &= \text{tr}(\Delta_{(0,1)}^{-1} D).
\end{aligned}$$

The expressions for A and D are identical to those given (2.14). The vector of parameters $(\rho^\top, \varphi\top)^\top$ is orthogonal to $(\beta^\top, \sigma^2)^\top$.

We define $\widetilde{Q}$ as a $2n \times 2n$ matrix given by

$$\widetilde{Q} = \begin{pmatrix} Q_0 & Q_{01} \\ Q_{01}^\top & Q_1 \end{pmatrix}, \tag{2.26}$$

where

$$Q_0 = T_0^2\Delta_0 - \Delta_{(0,1)}(T_0 + T_{10})^2 + \Delta_1 T_{10}^2,$$
$$Q_{01} = Q_{01}^\top = T_0\Delta_0 T_{01} - (T_0 + T_{10})\Delta_{(0,1)}(T_1 + T_{01}) + T_1\Delta_1 T_{10},$$
$$Q_1 = T_1^2\Delta_1 - \Delta_{(0,1)}(T_1 + T_{01})^2 + \Delta_0 T_{01}^2.$$

The second order leading principal submatrix of $K(\theta)$ (i.e., the first two lines and columns of $K(\theta)$) can be written as

$$K_\Upsilon(\Upsilon) = \begin{pmatrix} K_{\rho\rho} & K_{\rho\varphi} \\ K_{\varphi\rho} & K_{\varphi\varphi} \end{pmatrix} = \widetilde{Z}^\top \widetilde{Q} \widetilde{Z},$$

where $\Upsilon = (\rho^\top, \varphi^\top)^\top$.

The inverse of Fisher's information matrix for the ZOIS-RE model is given by

$$K(\theta)^{-1} = \begin{pmatrix} K^{\rho\rho} & K^{\rho\varphi} & 0 & 0 \\ K^{\varphi\rho} & K^{\varphi\varphi} & 0 & 0 \\ 0 & 0 & K^{\beta\beta} & 0 \\ 0 & 0 & 0 & K^{\sigma\sigma^2} \end{pmatrix}, \tag{2.27}$$

where

$$K^{\rho\rho} = K_{\rho\rho}^{-1}[I_{k_0} + K_{\rho\varphi}(K_{\varphi\varphi} - K_{\rho\varphi}K_{\rho\rho}^{-1}K_{\rho\varphi})^{-1}K_{\rho\varphi}K_{\rho\rho}^{-1}],$$
$$K^{\rho\varphi} = K^{\varphi\rho\top} = -K_{\rho\rho}^{-1}K_{\rho\varphi}(K_{\varphi\varphi} - K_{\rho\varphi}K_{\rho\rho}^{-1}K_{\rho\varphi})^{-1},$$
$$K^{\varphi\varphi} = (K_{\varphi\varphi} - K_{\rho\varphi}K_{\rho\rho}^{-1}K_{\rho\varphi})^{-1},$$

and I_{k_0} is the identity matrix of dimension $k_0 \times k_0$. The components $K^{\beta\beta}$, $K^{\sigma\sigma}$ are the same as those given in (2.15).

2.5.2 Estimation Process

The estimation process of the parameters that index the ZOIS-RE model uses Fisher's scoring method. The MLE of the vector of parameters $\Upsilon = (\rho^\top, \varphi^\top)^\top$ is obtained using the following iterative mechanism:

$$\begin{aligned}\Upsilon^{(m+1)} &= \Upsilon^{(m)} + (\widetilde{Z}^\top \widetilde{Q}^{(m)} \widetilde{Z})^{-1} \widetilde{Z}^\top \widetilde{T}^{(m)} y_\Delta^{(m)} \\ &= (\widetilde{Z}^\top \widetilde{Q}^{(m)} \widetilde{Z})^{-1} \widetilde{Z}^\top \widetilde{Q}^{(m)} \widetilde{y}^{(m)},\end{aligned}$$

$m = 0, 1, 2, \ldots$, where $\widetilde{Q}$ is as defined in (2.26) and the matrices $\widetilde{Z}$, $\widetilde{Q}$ and the vector y_Δ are given in (2.22). The vector $\widetilde{y}^{(m)} = \widetilde{Z}\Upsilon(m) + (\widetilde{Q}^{(m)})^{-1}\widetilde{T}^{(m)} y_\Delta^{(m)}$ is a modified response variable. The estimation process of $(\beta^\top, \sigma^2)^\top$ is identical to that of the IS-RE$_c$ model.

2.5.3 Confidence Intervals and Hypothesis Tests

Under standard regularity conditions, $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_{k_0+k_1+k+1})^\top = (\widehat{\rho}_1, \ldots, \widehat{\rho}_{k_0}, \widehat{\varphi}_1, \ldots, \widehat{\varphi}_{k_1}, \widehat{\beta}_1, \ldots, \widehat{\beta}_k, \widehat{\sigma}^2)^\top$ and $K(\widehat{\theta})$ are consistent estimators for θ and $K(\theta)$, respectively, $K(\widehat{\theta})$ being Fisher's information matrix (2.25) evaluated at $\widehat{\theta}$. Suppose that $J(\theta) = \lim_{n\to\infty} K(\theta)/n$ exists and is not singular. It follows that

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}_{k_0+k_1+k+1}(0, J(\theta)^{-1}),$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution and $\mathcal{N}_{k_0+k_1+k+1}$ is the multivariate normal distribution of dimension k_0+k_1+k+1. Asymptotic confidence intervals for the parameters that index the simplex regression can be obtained using the asymptotic normality of the MLE $\widehat{\theta}$. For $r = 1, \ldots, k_0 + k_1 + k + 1$, the $100(1-\nu)\alpha\%$ level asymptotic confidence interval for θ_r is

$$\left(\widehat{\theta}_r - z_{1-\frac{\nu}{2}}(K(\widehat{\theta})^{rr})^{1/2}, \widehat{\theta}_r + z_{1-\frac{\nu}{2}}(K(\widehat{\theta})^{rr})^{1/2}\right).$$

Here, $K(\widehat{\theta})^{rr}$ is the (r,r)th element of the inverse of Fisher's information matrix $K(\theta)$ evaluated at $\widehat{\theta}$.

Suppose we wish to test restrictions on the parameters of ZOIS-RE model. Let $\rho = (\rho_1^\top, \rho_2^\top)^\top$, $\varphi = (\varphi_1^\top, \varphi_2^\top)^\top$ and $\beta = (\beta_1^\top, \beta_2^\top)^\top$ be partitions of the parameter vectors. Here, $\rho_1 = (\rho_1, \ldots, \rho_{k_{0_1}})^\top$, $\rho_2 = (\rho_{k_{0_1}+1}, \ldots, \rho_{k_0})^\top$, $\varphi_1 = (\varphi_1, \ldots, \varphi_{k'})^\top$, $\varphi_2 = (\varphi_{k'+1}, \ldots, \varphi_{k_1})^\top$, $\beta_1 = (\beta_1, \ldots, \beta_{k''})^\top$ and $\beta_2 = (\beta_{k''+1}, \ldots, \beta_k)^\top$. The interest lies in testing $\mathcal{H}_0 : \rho_1 = \rho_1^{(0)}; \varphi_1 = \varphi_1^{(0)}; \beta_1 = \beta_1^{(0)}$ against $\mathcal{H}_1$: at least one equality is violated. The vectors $\rho_1^{(0)}$, $\varphi_1^{(0)}$ and $\beta_1^{(0)}$ are given and have dimensions k_{0_1}, k' and k'', respectively. We assume that $0 \leq k_{0_1} \leq k_0$, $0 \leq k' \leq k_1$ and $0 \leq k'' \leq k$. The trivial case $k_{0_1} = k' = k'' = 0$ is excluded.

The log-likelihood ratio statistic is

$$\Lambda = 2\{\ell(\widehat{\rho}, \widehat{\varphi}, \widehat{\beta}, \widehat{\sigma}^2) - \ell(\widetilde{\rho}, \widetilde{\varphi}, \widetilde{\beta}, \widetilde{\sigma}^2)\},$$

where $\ell(\rho, \varphi, \beta, \sigma^2)$ is the log-likelihood function (2.20); $(\widehat{\rho}, \widehat{\varphi}, \widehat{\beta}, \widehat{\sigma}^2)$ and $(\widetilde{\rho}, \widetilde{\varphi}, \widetilde{\beta}, \widetilde{\sigma}^2)$ are the unrestricted and restricted MLEs, the latter is obtained by imposing the null hypothesis. If the restricted model is not on the boundary of the parametric space, then, under same regularity conditions and under $\mathcal{H}_0$, $\Lambda \xrightarrow{\mathcal{D}} \chi^2_{k_{0_1}+k'+k''}$. Hence, the likelihood ratio test can be performed using asymptotic critical values from the χ^2 distribution with $k_{0_1} + k' + k''$ degrees of freedom.

The score test is an alternative to the log-likelihood ratio test. Let $V = [V_1, V_2]$, $Z = [Z_1, Z_2]$ and $X = [X_1, X_2]$ be matrices of regressors partitioned according to the null hypothesis. V_1, V_2, Z_1, Z_2, X_1 and X_2 are full rank matrices of dimension $n \times k_{0_1}$, $n \times (k_0 - k_{0_1})$, $n \times k'$, $n \times (k_1 - k')$, $n \times k''$ and $n \times (k - k'')$, respectively. Note that if $k_{0_1} = k_0$ we have $V_1 = V$, when $k' = k_1$ $Z_1 = Z$ and for $k'' = k$ we define $X_1 = X$. Let $U_{1\rho}$, $U_{1\varphi}$ and $U_{1\beta}$ be vectors of dimensions k_{0_1}, k' and k'' containing the first k_{0_1}, k' and k'' elements of the respective score vectors $U_\rho(\rho, \varphi)$, $U_\varphi(\rho, \varphi)$ and $U_\beta(\beta, \sigma^2)$. Let $K_{11}^{\rho\rho}$, $K_{11}^{\varphi\varphi}$ and $K_{11}^{\beta\beta}$ be matrices of dimension $k_{0_1} \times k_{0_1}$, $k' \times k'$ and $k'' \times k''$ having the first k_{0_1}, k' and k'' lines and columns of matrices $K^{\rho\rho}$, $K^{\varphi\varphi}$ and $K^{\beta\beta}$ defined in (2.27), respectively.

The partition defined according to the hypothesis $\mathcal{H}_0$ leads to

$$U_{1\rho} = V_1^\top T_0(\Delta_0 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)}) + V_1^\top T_{10}(\Delta_1 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)}),$$
$$U_{1\varphi} = Z_1^\top T_{01}(\Delta_0 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)}) + Z_1^\top T_1(\Delta_1 y_{\{1\}} - \Delta_{(0,1)} y_{(0,1)}),$$
$$U_{1\beta} = \sigma^{-2} X_1^\top \text{diagm}(y_{(0,1)}) T u.$$

The score statistic ξ can the be written as the sum of three quadratic forms:

$$\xi = \widetilde{U}_{1\rho}^\top K_{11}^{\rho\rho} \widetilde{U}_{1\rho} + \widetilde{U}_{1\varphi}^\top K_{11}^{\varphi\varphi} \widetilde{U}_{1\varphi} + \widetilde{U}_{1\beta}^\top K_{11}^{\beta\beta} \widetilde{U}_{1\beta}.$$

where "$\widetilde{\ }$" indicates that the quantities are evaluated imposing the null hypothesis. Under some regularity conditions and the null hypothesis, the score statistic is asymptotically distributed as $\chi^2_{k_{0_1}+k'+k''}$.

The Wald statistic can also be used to test the null hypothesis $\mathcal{H}_0$. It is given by

$$\varpi = (\widehat{\rho}_1 - \rho_1^{(0)})^\top (\widehat{K}_{11}^{\rho\rho})^{-1} (\widehat{\rho}_1 - \rho_1^{(0)}) + (\widehat{\varphi}_1 - \varphi_1^{(0)})^\top (\widehat{K}_{11}^{\varphi\varphi})^{-1} (\widehat{\varphi}_1 - \varphi_1^{(0)})$$
$$+ (\widehat{\beta}_1 - \beta_1^{(0)})^\top (\widehat{K}_{11}^{\beta\beta})^{-1} (\widehat{\beta}_1 - \beta_1^{(0)}),$$

where "$\widehat{\ }$" denotes that quantities are evaluated under the unrestricted estimators. Under the null hypothesis and under the regularity conditions, $\varpi \xrightarrow{\mathcal{D}} \chi^2_{k_{0_1}+k'+k''}$. In particular, if we want to test the null hypothesis $\mathcal{H}_0 : \beta_i = 0$ we can use the square root of the Wald statistic, i.e., $\widehat{\beta}_i/\text{s.e.}(\widehat{\beta}_i)$, where s.e.$(\widehat{\beta}_i)$ is the asymptotic standard error of the MLE $\widehat{\beta}_i$; it is obtained as the (i,i)th element of he matrix $K^{\beta\beta}$ evaluated at the maximum likelihood estimator. Under the null hypothesis, this statistic is asymptotically distributed as standard normal.

2.5.4 Application to simulated data

We should now report maximum likelihood estimation of the parameters of the zero-and-one inflated simplex regression (ZOIS-RE) model using `R` (R Development Core Team

2011). We consider the model with the following structure:

$$g(\mu_t) = \beta_0 + \beta_1 x_t,$$
$$h_1(\delta_{0t}) = \rho_0 + \rho_1 \nu_t,$$
$$h_2(\delta_{1t}) = \varphi_0 + \varphi_1 z_t,$$

$t = 1, \ldots, n$, where g is the logit function and h_1 and h_2 are the logarithmic functions. The values of x_t, ν_t and z_t are obtained as standard uniform distribution draws. The true values of the parameters are $\beta_0 = -1.5$, $\beta_1 = 1.5$, $\rho_0 = -2$, $\rho_1 = 0.5$, $\varphi_0 = -2$, $\varphi_1 = 0.7$ and $\sigma^2 = 1$. The sample size is $n = 500$.

Here, 16.2% of the values of the response variable are equal to zero and 15.0% are equal to one. Dispersion plots of the response variable y_t against x_t, ν_t and z_t are displayed in Figure 2.3. The maximum likelihood estimates and their respective standard errors can be found in Table 2.2. The figures indicate that the parameter estimators are noticeably biased even when $n = 500$. Further analyses on the bias of the MLE are called for. Our implementation uses the `R` software (R Development Core Team 2011) and can be found in Appendix A.7.

Table 2.2: Maximum likelihood estimate and standard errors for the simulated data in ZOIS-RE.

Parameter	β_0	β_1	ρ_0	ρ_1	φ_0	φ_1	σ^2
Estimate	-1.6687	1.5536	-1.7792	0.2637	-1.9661	0.4577	0.7550
Std. error	0.0290	0.0552	0.2432	0.4077	0.2531	0.4234	0.0288

2.6 Concluding remarks

In this chapter we developed a frequentist approach to the zero-one inflated simplex regression model proposed by Bandyopadhyay et al. (2014). The model allows the response variable to assume values in [0,1), (0,1] or [0,1]. Each model parameter is related to explanatory variables through a regression structure. We presented the maximum likelihood

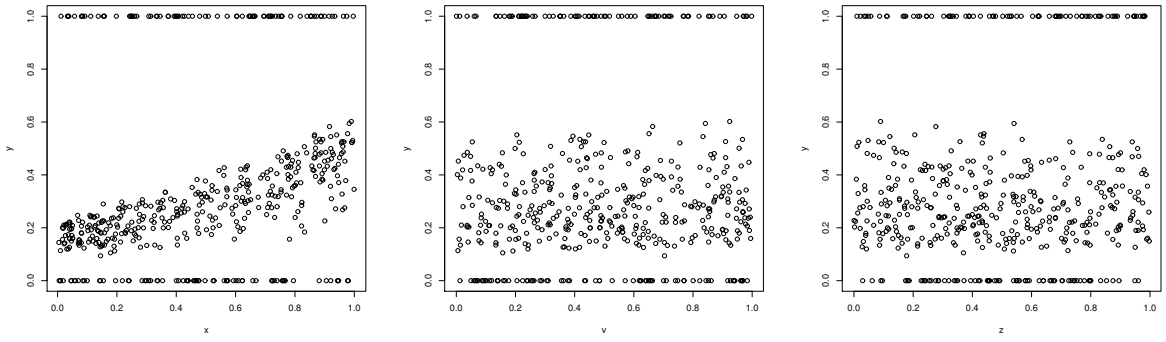


Figure 2.3: Dispersion diagrams of the response y_t against x_t (left), ν_t (center) and z_t (right).

estimation process for each model. We obtained matrix expressions for the score vector and for the Fisher's information matrix. Asymptotic confidence interval and hypothesis tests were also presented. Simulated data were analyzed.

CHAPTER 3

Residual Analysis in Inflated Simplex Regressions

Resumo

Medidas de diagnósticos são fundamentais para avaliar a adequação de modelos de regressão. A partir do gráfico de resíduos é possível identificar a presença de observações atípicas e verificar se as suposições do modelo são satisfeitas. Neste capítulo, são obtidos os resíduos padronizados, resíduos padronizados ponderados e os resíduos quantis aleatorizados para o modelo de regressão simplex inflacionado. Algumas medidas do tipo pseudo-R^2 são apresentadas para avaliar a qualidade do ajuste do modelo. A seleção de modelos via AIC, SBC e GAIC e a utilização de envelopes simulados são descritas. Uma aplicação que envolve a modelagem da proporção de uso de bebidas alcoólicas entre estudantes de escolas públicas por meio do modelo de regressão simplex inflacionado em zero é apresentada e discutida.

3.1 Introduction

Diagnostic analysis plays an important role in regression modeling. The model residuals contain important information for analyzing the regression fit. Residuals can be used to identify atipical data points and to determine whether the relevant of distributional assumptions hold.

A residual measures the discrepancy between the observed data and fitted values. Residual analysis can be based on ordinary residuals, on its standardized versions, on residuals defined based on components of the deviance function (McCullagh & Nelder 1989) or on generalized residuals (Cox & Snell 1968). Belsley et al. (1980) and Cook & Weisberg (1982) discussed the use of standardized residuals in normal linear regressions. Pregibon (1981) introduced the deviance residual in GLMs and defined a standardization using approximations proposed by Cox & Snell (1968). Williams (1984, 1987) found, using simulation, evidence of agreement between the empirical distribution of the deviance and the standard normal distribution for different GLMs. McCullagh (1987) presented an alternative standardization for the deviance aiming eliminating asymmetry and kurtosis. Atkinson (1981) showed how to construct confidence bands residuals in linear regression models using simulation. Williams (1987) discussed the computation of envelopes for residuals in GLMs. Farhrmeir & Tutz (1994) extended McCullagh's (1987) result for models that do not belong to the exponential family of distributions.

Residual analysis for different models have also been considered in literature. Ferrari & Cribari-Neto (2004) introduced some diagnostic measures for the beta regression model. Some other residuals for the same class of models were also proposed by Espinheira et al. (2008). Ferrari et al. (2011) and Rocha & Simas (2011) presented diagnostic tools in beta regression models with varying dispersion. Ospina & Ferrari (2012) introduced the zero and/or one inflated beta regression model and presented some diagnostic measures and model selection tools for that model. Tang et al. (2000) developed diagnostic tools for nonlinear dispersion models. Jørgensen (1997) presented Pearson and Wald residuals, score and dual score residuals, deviance and modified deviance residuals for exponential

dispersion models. Zhang et al. (2016) provided approximate Pearson residuals for simplex regression. Miyashiro (2008) presented the standardized weighted residual 2 (*resíduos ponderados padronizados 2*, in Portuguese) for simplex regression models. Dunn & Smith (1996) defined the (randomized) quantile residuals to check model adequacy, largely used in GAMLSS models. In this chapter we define some residuals for the zero and/or one simplex regression model using a similar approach to that of Ospina & Ferrari (2012) combined with the approach in Miyashiro (2008).

Another helpful and well known tool in regression models is the global goodness-of-fit measure known as coefficient of determination or R^2. Its original definition for linear models, however, cannot be used in zero and/or one simplex regression. Some pseudo-R^2 type goodness-of-fit summary statistics should then be used. They are also described for this model in this chapter. Some model selection procedures like the Akaike information criterion (AIC) (Akaike 1973, 1974), the Schwarz Bayesian criterion (SBC) (Schwarz 1978) and their generalized version (GAIC) (Lv & Liu 2014) are also described in this chapter for choosing between competing models. Simulated envelopes are also an interesting tools for evaluating the model at hand. An empirical application based on the zero inflated simplex regression is also presented in the following sections.

3.2 Residuals

Residuals measure disagreements between the fitted model and the data. They can be defined as a function $r(y_t, \widehat{E(y_t)})$ which measures the distance between the observed value and the estimated mean response (Snell 1968). Hereafter, residuals for the inflated simplex regression model are presented.

3.2.1 Residuals for the zero or one inflated simplex regression

Standardized Residuals

We considered the convergence of the Fisher's scoring method for φ and β in inflated simplex regression models in Equations (B.5) and (B.4) in Appendix B.1. The standardized residuals for the zero or one inflated simplex regression model can then be defined as

$$r_t^{(1)} = \frac{\mathbb{1}_{\{c\}}(y_t) - \widehat{\alpha}_t}{\sqrt{\widehat{\alpha}_t(1-\widehat{\alpha}_t)(1-\widehat{h}^*_{1_{tt}})}} \tag{3.1}$$

and

$$r_t^{(2)} = \frac{\widehat{u}_t}{\sqrt{b_t(1-\widehat{\alpha}_t)(1-\widehat{h}^*_{2_{tt}})}}, \tag{3.2}$$

where the terms $\widehat{u}_t$ and b_t are given by

$$\widehat{u}_t = (y_t - \widehat{\mu}_t)\frac{y_t - 2\widehat{\mu}_t y_t + \widehat{\mu}_t^2}{y_t(1-y_t)\widehat{\mu}_t^3(1-\widehat{\mu}_t)^3} \tag{3.3}$$

and

$$b_t = \widehat{\mathrm{Var}}(u_t) = \widehat{\sigma}^2\left[\frac{3\widehat{\sigma}^2}{\widehat{\mu}_t(1-\widehat{\mu}_t)} + \frac{1}{\widehat{\mu}_t^3(1-\widehat{\mu}_t)^3}\right], \tag{3.4}$$

$\widehat{\mu}_t = g^{-1}(x_t^\top\widehat{\beta})$, $t = 1, \ldots, n$. The components $\widehat{h}^*_{1_{tt}}$ and $\widehat{h}^*_{2_{tt}}$ are the tth diagonal elements of the projection matrices

$$\widehat{H}_1^* = \widehat{Q}^{1/2}Z(Z^\top\widehat{Q}Z)^{-1}Z^\top\widehat{Q}^{1/2}$$

and

$$\widehat{H}_2^* = (\widehat{\Delta}\widehat{A})^{1/2}X(X^\top\widehat{\Delta}\widehat{A}X)^{-1}X^\top(\widehat{\Delta}\widehat{A})^{1/2},$$

respectively. Here, Z $(n \times M)$ and X $(n \times m)$ are matrices of known fixed covarites values whose tth lines are $z_t^\top = (z_{t1}, \ldots, z_{tM})$ and $x_t^\top = (x_{t1}, \ldots, x_{tm})$, respectively. Matrices Q, Δ and A are defined in (B.1) and (B.2).

Note that $r_t^{(1)}$ and $r_t^{(2)}$ are, respectively, the standardized residuals for the discrete

and the continuous components of the zero or one inflated simplex regression model. These residuals tipically follow empirical distributions which renders the use of diagnostic measures more dificult. However, plots of $r_t^{(1)}$ and $r_t^{(2)}$ against $\widehat{\alpha}_t$ and $\widehat{\mu}_t$, respectively, may reveal outliers corresponding to each submodel.

From Equations (3.1) and (3.2), the standardized residuals of the IS-RE$_c$ model can be defined as

$$r_t = \begin{cases} r_t^{(1)}, & if\ y_t = c, \\ r_t^{(2)}, & if\ y_t \in (0,1). \end{cases} \tag{3.5}$$

Note that the distribution of the residuals is asymmetric due to the presence of probability mass in $y = c$.

We also introduce the weighted standardized residuals

$$r_t^* = \widehat{\alpha} r_t^{(1)} + (1 - \widehat{\alpha}) r_t^{(2)}, \tag{3.6}$$

where $c = 0$ or $c = 1$. Plots of r_t and r_t^* against adjusted values $\widehat{\mathrm{E}(y_t)}$ allow us to identify the outliers and influential points.

Randomized Quantile Residuals

For a continuous response (y), residuals can be defined using the cumulative distribution function $F(y)$. Let $U = F(y, \theta)$, which is uniformly distributed in the unity interval, and let $\Phi(\cdot)$ the cumulative standard normal distribution function. Then we have that

$$V = \Phi^{-1}(F(y, \theta)) = \Phi^{-1}(U)$$

is standard normal distributed. When θ is known we obtain the Cox and Snell residual (Snell 1968). In practice, θ is unknown and it should be replaced by its maximum likelihood estimate. Fortunately, $\widehat{\theta}$ is a consistent estimator for θ and $V \xrightarrow{D} N(0,1)$, i.e., V is asymptotically distributed as standard normal. Then, $E(V) \approx 0$ and $\mathrm{Var}(V) \approx 1$.

The zero or one inflated simplex distribution is not absolutely continuous and it is necessary to provide a more general definition of the residual V. Based on Dunn & Smith (1996), the randomized quantile residual for the zero or one inflated simplex regression model is

$$r_t^q = \Phi^{-1}(v_t),$$

$t = 1, \ldots, n$, where v_t is a random uniform variable in $(a_t, b_t]$. Here, $a_t = \lim_{y \uparrow y_t} IS_c(y, \widehat{\alpha}, \widehat{\mu}, \widehat{\sigma}^2)$ and $b_t = IS_c(y, \widehat{\alpha}, \widehat{\mu}, \widehat{\sigma}^2)$. In the zero inflated simplex regression, for $y = 0$, $a_t = \lim_{y \uparrow 0} ZIS(y, \widehat{\alpha}, \widehat{\mu}, \widehat{\sigma}^2) = (1 - \widehat{\alpha}) \lim_{y \uparrow 0} F(y; \widehat{\mu}, \widehat{\phi}) = 0$ and $b_t = ZIS(y, \widehat{\alpha}, \widehat{\mu}, \widehat{\sigma}^2) = \widehat{\alpha}$. Then, v_t is a uniform random variable in $(0, \widehat{\alpha}]$. On the other hand, for $y_t \in (0, 1)$, we have $v_t = ZIS(y, \widehat{\alpha}, \widehat{\mu}, \widehat{\sigma}^2)$. In the one inflated simplex regression, for $y_t = 1$ we have $a_t = \lim_{y \uparrow 1} OIS(y, \widehat{\alpha}, \widehat{\mu}, \widehat{\sigma}^2) = (1 - \widehat{\alpha}) \lim_{y \uparrow 1} F(y; \widehat{\mu}, \widehat{\phi}) = (1 - \widehat{\alpha})$ and $b_t = OIS(1; \widehat{\alpha}, \widehat{\mu}, \widehat{\sigma}^2) = 1$. Hence, v_t is a random uniform variable in $(1 - \widehat{\alpha}, 1]$. For $y \in (0, 1)$, $v_t = ZIS(y_t, \widehat{\alpha}, \widehat{\mu}, \widehat{\sigma}^2)$.

Randomization of residuals is used to produce continuous normal residuals. We point out that different values for the randomized quantile residuals should be observed in each realization of this procedure. We then advise readers to compute these residuals at least four times in order to detect eventual patterns.

3.2.2 Residuals for the zero and one inflated simplex regression

We shall now obtain residuals for the zero and one inflated simplex regression. Using Fisher's scoring method we obtained the MLEs for ρ, φ and β in Equations (B.10) and (B.5) (see Appendix B.1 and B.2). The standardized residuals for the zero and one inflated simplex regression model (r_t) can be split into three terms as

$$r_t = \begin{cases} r_t^{\{0\}}, & \text{if } y_t = 0, \\ r_t^{\{1\}}, & \text{if } y_t = 1, \\ r_t^{(c)}, & \text{if } y_t \in (0, 1). \end{cases}$$

As one may note, each residual is associated to the submodel that models the probability of occurrence of zeros, the probability of occurrence of ones and the occurrences of values in (0,1). The first two terms can be expressed as

$$r_t^{\{0\}} = \frac{\mathbb{1}_{\{0\}}(y_t) - \delta_{0t}}{\sqrt{\widehat{q}_{1_{tt}}(1 - \widehat{h}_{tt}^{\{0\}})}} \tag{3.7}$$

and

$$r_t^{\{1\}} = \frac{\mathbb{1}_{\{1\}}(y_t) - \delta_{1t}}{\sqrt{\widehat{q}_{2_{tt}}(1 - \widehat{h}_{tt}^{\{1\}})}}, \tag{3.8}$$

where $q_{1_{tt}}$ and $q_{2_{tt}}$ are, respectively, the tth diagonal elements of the matrices Q_1 and Q_2 defined in Appendix B.2 and $h_{tt}^{\{0\}}$ and $h_{tt}^{\{1\}}$ are the tth diagonal elements of the projection matrices $H^{\{0\}} = \Psi_0^\top H_d^* \Psi_0$ and $H^{\{1\}} = \Psi_1^\top H_d^* \Psi_1$, respectively. Here, $\Psi_0 = (I_n, 0_n)^\top$ and $\Psi_1 = (0_n, I_n)^\top$, I_n is the $n \times n$ identity matrix, 0_n is an $n \times n$ matrix of zeros and $H_d^* = Q^{1/2}\widetilde{Z}(\widetilde{Z}^\top Q \widetilde{Z})^{-1}\widetilde{Z}^\top Q^{1/2}$ is a projection matrix. The quantities are evaluated in the maximum likelihood estimator. Residual plots can be used to identify outliers in each submodel. We suggest plotting $r_t^{\{0\}}$ against $\widehat{\delta}_{0t}$ and $r_t^{\{1\}}$ against $\widehat{\delta}_{1t}$, for $t = 1, \ldots, n$, separately.

For the continuous component, the same standardized weighted residual defined for the zero or one inflated simplex regression model can be used. In this case,

$$r_t^{(c)} = \frac{\widehat{u}_t}{\sqrt{q_t(1 - \widehat{\delta}_{0t} - \widehat{\delta}_{1t})(1 - \widehat{h}_{c_{tt}}^*)}}, \tag{3.9}$$

$t = 1, \ldots, n$, where u_t and q_t are defined in Equations (3.3) and (3.4), respectively; $\widehat{h}_{c_t t}^*$ is the tth element of the main diagonal of the projection matrix $\widehat{H}_c^* = (\widehat{\Delta}\widehat{A})^{1/2} X (X^\top \widehat{\Delta}\widehat{A} X)^{-1}$ $X^\top (\widehat{\Delta}\widehat{A})^{1/2}$. Here, X is an $n \times m$ matrix of known fixed values, where $x_t^\top = (x_{t1}, \ldots, x_{tm})$ and Δ and A are defined in Equation (B.2). Plots of $r_t^{(c)}$ against $\widehat{\mu}_t$ may reveal outliers in the submodel of the zero and one inflated simplex regression that models observations in $(0, 1)$.

We can also define the weighted standardized residual. It is given by

$$r_t^* = \widehat{\delta}_{0t} r_t^{\{0\}} + \widehat{\delta}_{1t} r_t^{\{1\}} + (1 - \widehat{\delta}_{0t} - \widehat{\delta}_{1t}) r_t^{(c)}.$$

Note that r_t^* is a weighted sum of $r_t^{\{0\}}$, $r_t^{\{1\}}$ and $r_t^{(c)}$. Plots of r_t^* against adjusted values can help identifying atypical observations.

Randomized Quantile Residuals

In the zero and one inflated simplex regression model we can define the randomized quantile residual as

$$r_t^q = \Phi^{-1}(u_t), \quad t = 1, \ldots, n,$$

where u_t is a uniform random variable on the interval $(a_t, b_t]$, in which $a_t = \lim_{y \uparrow y_t} \text{ZOIS}(y; \delta_0, \delta_1, \mu, \sigma^2)$ and $b_t = \text{ZOIS}(y; \delta_0, \delta_1, \mu, \sigma^2)$, respectively. Here, $\text{ZOIS}(y; \delta_0, \delta_1, \mu, \sigma^2)$ is defined in Equation (2.7). A plot of r_t^q against the indices of the observations may reveal atypical data points. A detectable trend in the plot of residuals against estimated predictors may indicate link function misspecification. Normal probability plots with simulated envelopes are also a helpful diagnostic tool (Atkinson 1985).

3.3 Global Goodness-of-fit Measure

The goodness-of-fit of a zero and/or one inflated simplex regression model can be measured using a pseudo-R^2. A simple pseudo-R^2, say R_p^2, is given by the square of the correlation coefficient between the response, $y_1, \ldots, y_n$, and the respective predicted values, $\breve{\mu}_1, \ldots, \breve{\mu}_n$, where $\breve{\mu}_t = \widehat{\text{E}(y_t)} = c\,\widehat{\alpha}_t + (1 - \widehat{\alpha}_t)\widehat{\mu}_t$. Perfect disagreement between y's and $\breve{\mu}$'s yields $R_p^2 = 0$, whereas perfect agreement leads to $R_p^2 = 1$. Two alternative pseudo-R^2 measures can be defined as $R_p^{2*} = 1 - \log \widehat{L}_0 / \log \widehat{L}$ (McFadden 1974) and $R_{LR}^2 = 1 - (\widehat{L}_0/\widehat{L})^{2/n}$ (Cox. & Snell 1989, p.208-209). Here, $\widehat{L}_0$ and $\widehat{L}$ are, respectively, the maximized likelihood functions of the null model and the fitted model. Note that R_p^{2*}

is valid only for positive $\widehat{L}_0$ and $\widehat{L}$.

3.4 Model Selection

Likelihood ratio tests, which were defined in Section 2.5.3 of Chapter 2, can be used to compare nested zero and/or one inflated simplex regression models. Another useful approach to select the most parsimonious model, i.e., a well adjusted model with a can help identifying atypical observations number of parameter is the generalized Akaike information criterion (GAIC). This selection procedure can also be used in competing nonnested models. It is defined as GAIC $= -2\widehat{\ell} + (\phi.df)$, where $-2\widehat{\ell}$ is the fitted deviance (Rigby & Stasinopoulos 2005), $\widehat{\ell}$ is the maximized log-likelihood, ϕ is a penalization term and df denotes the degrees of freedom of the model. The first term of GAIC can be interpreted as a measure of lack of fit. Tthe model that corresponds to the smallest GAIC is selected. Special cases of the GAIC are the Akaike information criterion AIC (Akaike 1974) for $\phi = 2$, the Schwarz Bayesian criterion SBC (Schwarz 1978) for $\phi = \log(n)$ and the consistent Akaike information criterion (CAIC) for $\phi = \log(n) + 1$. For nonnested zero and/or one inflated simplex regression models we recommend the use of J and MJ, as defined in Section 1.3.1 of Chapter 1.

3.5 Simulated Envelopes

Normal probability plots with simulated envelopes are a helpful diagnostic tool to evaluate a fitted model. It is based on standardized residuals. Further details can be found in Neter et al. (1996). The simulated envelope is determined by the confidence bands. Points that lie outside the confidence bands indicate that the model is not appropriate. We recommend the use of the weighted standardized residuals defined in Section 3.2.1 in simulated envelope plots.

3.6 Application

We shall now present an empirical application of the inflated simplex regression. The interest lies in modeling the proportion of alcohol use by public school students in the past 30 days in California in years 2008 to 2010 (`Percentage`). The data consists of 1340 observations. There are five covariates: a factor with 56 levels/clusters/counties (`County`); a grade level indicating 7th, 9th or 11th grade (`Grade`); a factor with levels [1,2], [3,9], [10,19] and [20,30] (`Days`); the med point of each of the intervals defined in `Days` (`MedDays`); a factor with levels Female and Male (`Gender`). The data can be found at `http://www.kidsdata.org`.

The response contains observed values between 0 and 0.3330, of which 3.88% are zeros. The mean value is 0.0656, the median is 0.0380 and the standard deviation is 0.0615. Figure 3.1 shows the histogram and the boxplot of the response. Notice that the response is asymmetric, unimodal and there is a concentration in the smaller values. Given the presence of zeros in the data, the zero inflated simplex regression model may be adjusted.

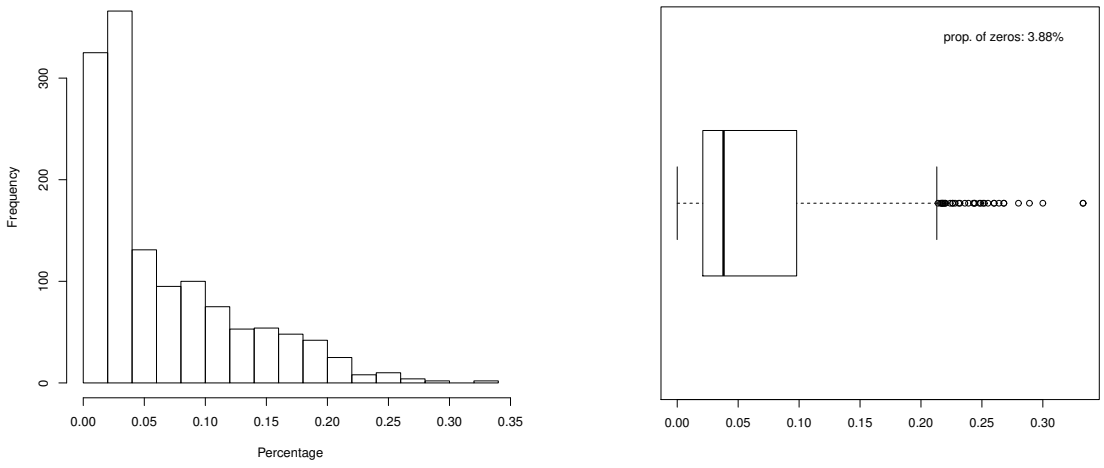


Figure 3.1: Histogram (left) and boxplot (right) for the proportion of public school students in 4 buckets of days in which they drank alcohol in the past 30 days in California in years 2008 to 2010.

We consider the zero inflated simplex regression model where $y_t \sim ZIS(\alpha, \mu_t, \sigma^2)$ and such that

$$\begin{aligned} \text{logit}(\alpha) &= Z\varphi, \\ \text{logit}(\alpha) &= X\beta, \end{aligned} \tag{3.10}$$

where .Z is a matrix of regressors containing the factors of *County*, *Days* and *Gender*; X is a matrix of regressors containing Grade and the factor of $\widehat{\text{Days}}$; φ and β are the respective parameter vectors to be estimated in each submodel. Here, we considered $\sigma^2 = \exp(\nu)$. The model parameters were estimated by maximum likelihood using RS algorithm (Rigby & Stasinopoulos 2005). Table 3.1 shows the maximum likelihood estimates and their respective standard errors.

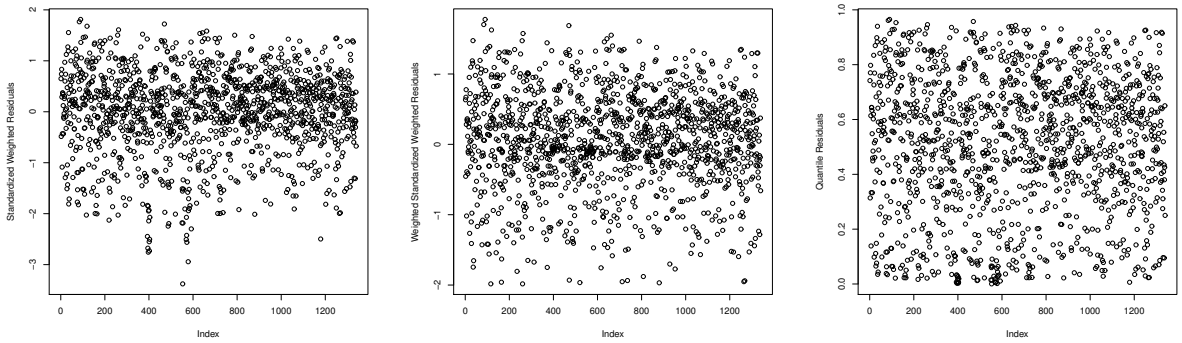
The AIC and the BIC were computed to compare the saturated model (containing all possible covariates in each submodel) and the model given in Equation (3.10). For the saturated model we observed AIC $= -5839.126$ and SBC $= -5475.096$. For the model in Equation (3.10) were obtained AIC $= -5860.403$ and SBC $= -5501.573$. The likelihood ratio test was also performed to compare both models. The test statistics equals $\Lambda = 0.7538$ and the corresponding p-value $= 0.3853$. According to the AIC, the SBC and also the likelihood ratio criteria, the model in Equation (3.10) may be used with the data at hand. The pseudo-R^2 (defined by the square of the correlation coefficient between the response and the respective predicted values) for the fitted model equals 0.7621, thus indicating a good fit.

In order to identify possible deviations from the model assumptions, we use plots of r_t, r_t^* and r_t^q against the observations indices. Figure 3.2 shows standardized residuals (Figure 3.2(a)), weighted standardized residuals (Figure 3.2(b)) and randomized quantile residuals (Figure 3.2(c)). Such plots do not indicate the presence of outliers and there is no clear systematic pattern.

In Figure 3.3 we plot the standardized residuals $r_t^{(1)}$ (for the discrete component) and $r_t^{(1)}$ (for the continuous component) against $\widehat{\alpha}_t$ and $\widehat{\mu}_t$, respectively. Such plots do not

Table 3.1: Maximum likelihood estimates of ZIS-RE model for the proportion of alcohol use by public school students in the past 30 days in California in years 2008 to 2010.

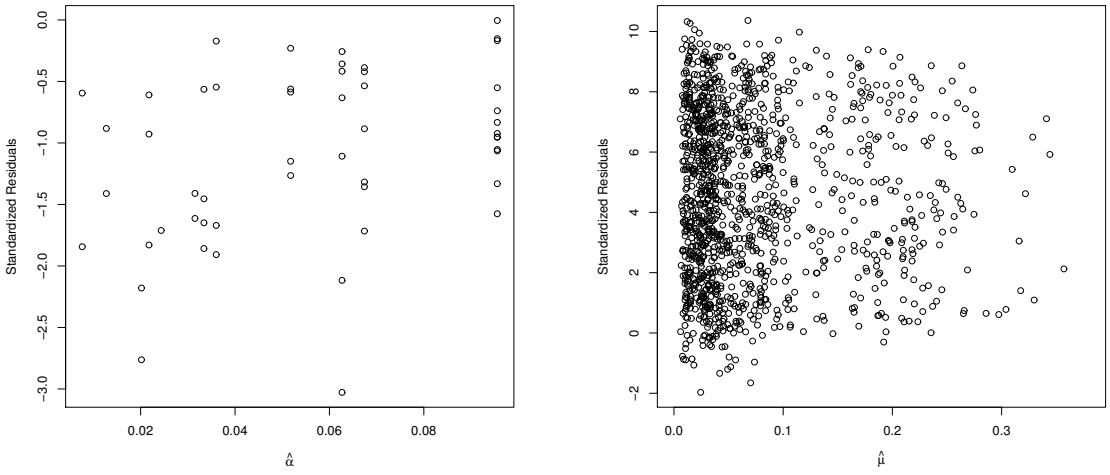
Submodel for α					
Covariate	Estimate	Std. Error	Covariate	Estimate	Std. Error
Intercept	−3.6906	0.4699	Days[3,9]	0.9855	0.5333
Grade9	−0.6601	0.3285	Days[10,19]	1.4430	0.5075
Grade11	−1.1779	0.3903	Days[20,30]	1.0634	0.5282
Submodel for μ					
Covariate	Estimate	Std. Error	Covariate	Estimate	Std. Error
Intercept	−2.7609	0.0847	CountyRiverside	0.3001	0.1118
CountyAmador	0.4205	0.1190	CountySacramento	0.0113	0.1041
CountyButte	0.2509	0.1103	CountySanBenito	0.2577	0.1105
CountyCalaveras	0.1369	0.1156	CountySanBernardino	0.4073	0.1148
CountyColusa	0.5131	0.1177	CountySanDiego	0.3158	0.1120
CountyContraCosta	0.2289	0.1096	CountySanFrancisco	−0.2682	0.0977
CountyDelNorte	0.7177	0.1258	CountySanJoaquin	0.2475	0.1102
CountyElDorado	0.1174	0.1065	CountySanLuisObispo	0.3130	0.1119
CountyFresno	0.2999	0.1114	CountySanMateo	0.0412	0.1047
CountyGlenn	0.3566	0.1225	CountySantaBarbara	0.2959	0.1119
CountyHumboldt	0.2337	0.1095	CountySantaClara	−0.1273	0.1008
CountyImperial	0.4567	0.1165	CountySantaCruz	0.6065	0.1208
CountyInyo	0.6327	0.1247	CountyShasta	0.4421	0.1153
CountyKern	0.3159	0.1121	CountySiskiyou	0.1051	0.1118
CountyKings	0.4556	0.1160	CountySolano	0.3405	0.1126
CountyLake	0.8682	0.1279	CountySonoma	0.3087	0.1121
CountyLassen	0.5063	0.1602	CountyStanislaus	0.6107	0.1208
CountyLosAngeles	0.1436	0.1074	CountySutter	0.3563	0.1130
CountyMadera	0.4381	0.1156	CountyTehama	0.4844	0.1218
CountyMarin	0.2882	0.1110	CountyTrinity	0.9188	0.1317
CountyMariposa	1.2907	0.1430	CountyTulare	0.4834	0.1172
CountyMendocino	0.6012	0.1206	CountyTuolumne	0.8053	0.1271
CountyMerced	0.5098	0.1179	CountyVentura	0.3130	0.1119
CountyModoc	0.4697	0.1756	CountyYolo	0.4013	0.1145
CountyMono	0.5303	0.1565	CountyYuba	0.6139	0.1204
CountyMonterey	0.3955	0.1144	Grade9	0.8255	0.0290
CountyNapa	0.1026	0.1067	Grade11	1.1270	0.0325
CountyNevada	0.0841	0.1057	Days[10.19]	−2.1198	0.0463
CountyOrange	0.0886	0.1060	Days[20.30]	−1.7842	0.0475
CountyPlacer	0.0387	0.1045	Days[3.9]	−1.1255	0.0512
CountyPlumas	0.8321	0.1479	GenderMale	0.0564	0.0244
$\widehat{\sigma}^2$	2.3740	0.0197			



(a) (b) (c)

Figure 3.2: Residual plots.

indicate the presence of outliers.



(a) (b)

Figure 3.3: Residual plots for the discrete and the continuous component.

Normal probability plots with simulated envelopes are shown in Figure 3.4. Notice that residuals lie inside the confidence bands. There is no evidence agaisnt the fitted zero inflated simplex regression model. The distribution of residuals are left skewed because of the probability mass observed in zero point.

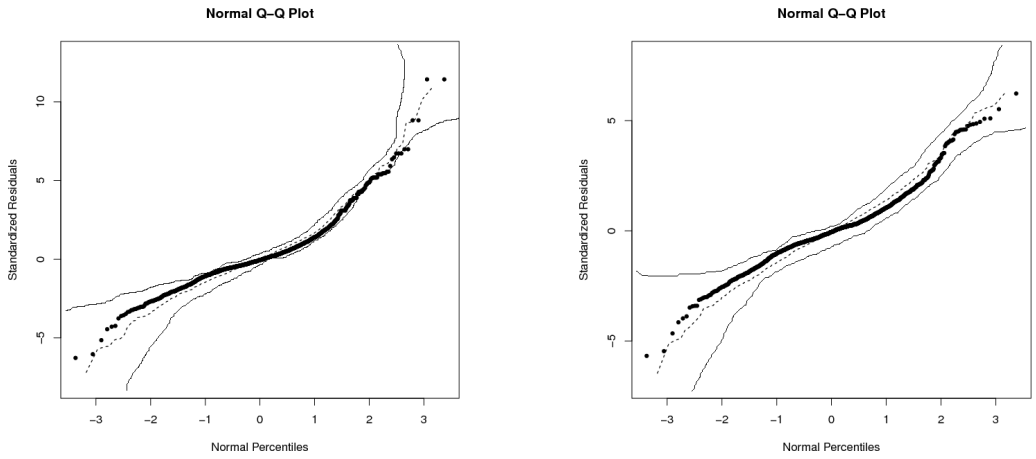


Figure 3.4: Normal probability plots with simulated envelopes.

3.7 Concluding remarks

In this chapter we obtained the standardized weighted residuals for the zero and/or one inflated simplex regression model. These residuals are divided into two parts: residuals for the continuous components of the model (i.e., when the response lies in the open unit interval) and residuals for the discrete components of the model (when the response assumes 0 and/or 1). Such residuals are asymmetrically distributed which may render diagnostic analysis difficult to perform. To overcome that shortcoming we defined the weighted standardized residuals. Randomized quantile residuals were also defined in order to produce continuous residuals that are approximately normally ditributed.

Pseudo-R^2 measures were also provided. A model selection procedure based on the generalized Aikaike information criterion was described. We constructed normal probability plots with simulated envelopes as a diagnostic tool. We also presented and discussed an application with real (nor simulated) data.

References

Akaike, H. (1973), *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, chapter Information theory and an extension of the maximum likelihood principle, pp. 267–281.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE. Transactions on Automatic Control* **19**, 716–723.

Akantziliotou, K., Rigby, R. A. & Stasinopoulos, D. M. (2002), The R implementation of generalized additive models for location, scale and shape, *in* M. Stasinopoulos & G. Touloumi, eds, 'Statistical modelling in Society: Proceedings of the 17th International Workshop on Statistical Modelling', Chania, Greece, pp. 75–83.

Atkinson, A. (1970), 'A method for discriminating between models (with discussion)', *Journal of the Royal Statistical Society B* **32**(3), 323–353.

Atkinson, A. C. (1981), 'Two graphical display for outlying and influential observations in regression', *Biometrika* **68**, 13–20.

Atkinson, A. C. (1985), *Plots, Transformations and Regression: An Introduction to*

Graphical Methods of Diagnostic Regression Analysis, Oxford University Press, New York.

Bandyopadhyay, D., Galvis, D. M. & Lachos, V. H. (2014), 'Augmented mixed models for clustered proportion data', *Statistical Methods in Medical Research* (forthcoming).

Barndorff-Nielsen, O. & Jørgensen, B. (1991), 'Some parametric models on the simplex', *Journal of Multivariate Analysis* **39**(1), 106–116.

Belsley, D. A., Kuh, E. & Eelsch, R. E. (1980), *Regression Diagnostics*, Wiley, New York.

Burridge, P. & Fingleton, B. (2010), 'Bootstrap inference in spatial econometrics: the J-test', *Spatial Economic Analysis* **5**(1), 93–119.

Cole, T. J. & Green, P. J. (1992), 'Smoothing reference centile curves: the LMS method and penalized likelihood', *Statistics in Medicine* **11**, 1305–1319.

Cook, R. D. & Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, New York.

Cox, C. (1996), 'Nonlinear quasi-likelihood models: applications to continuous proportions', *Computational Statistics & Data Analysis* **21**(4), 449–461.

Cox, D. (1961), 'Tests of separated families of hypothesis', *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 105–123.

Cox, D. (1962), 'Further results on tests of separete families of hypotheses', *Journal of the Royal Statistical Society B* **24**(2), 406–424.

Cox, D. R. (2013), 'A return to an old paper: 'tests of separate families of hypotheses'', *Journal of the Royal Statistical Society B* **75**(2), 207–215.

Cox., D. R. D. R. & Snell, J. (1989), *Analysis of Binary Data*, Chapman and Hall, London.

Cox, D. R. & Hinkley, D. V. (1974), *Theoretical Statistics*, Chapman and Hall, London.

Cox, D. & Snell, E. (1968), 'A general definition of residuals', *Journal of the Royal Statistical Society B* **30**, 248–275.

Cribari-Neto, F. & Lucena, S. E. F. (2015), 'Nonnested hypothesis testing in the class of varying dispersion beta regressions', *Journal of Applied Statistics* **42**(5), 967–985.

Crowder, M. J., Kimber, A. C., Smith, R. L. & Sweeting, T. J. (1991), *Statistical Analysis of Reliability Data*, Chapman and Hall, London.

Dastoor, N. (1983), 'Some aspects of testing nonnested hypotheses', *Journal of Econometrics* **21**(2), 213–218.

Davidson, R. & MacKinnon, J. (1981), 'Several tests for model specification in the presence of alternative hypotheses', *Econometrica* **49**(3), 781–793.

Davidson, R. & MacKinnon, J. G. (2002), 'Bootstrap J tests of nonnested linear regression models', *Journal of Econometrics* **109**(1), 167–193.

Deaton, A. S. (1982), *Evaluating the Reliability of Macroeconomic Models*, Wiley, New York, chapter Model selection procedures, or, does the consumption function exist?, pp. 43–65.

Dunn, P. K. & Smith, G. K. (1996), 'Randomized quantile residuals', *Journal of Computational and Graphical Statistics* **5**, 236–244.

Espinheira, P. L., Ferrari, S. L. P. & Cribari-Neto, F. (2008), 'On beta regression residuals', *Journal of Applied Statistics* **35**, 407–419.

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013), *Regression: models, methods and applications*, Springer.

Famoye, F. & Singh, K. P. (2006), 'Zero-inflated generalized Poisson regression model with an application to domestic violence data', *Journal of Data Science* **4**(1), 117–130.

Fan, Y. & Li, Q. (1995), 'Bootstrapping J-type tests for non-nested regression models', *Economics Letters* **48**(2), 107–112.

Farhrmeir, L. & Tutz, G. (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer, New York.

Ferrari, S. L., Espinheira, P. L. & Cribari Neto, F. (2011), 'Diagnostic tools in beta regression with varying dispersion', *Statistica Neerlandica* **65**(3), 337–351.

Ferrari, S. L. P. & Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* **31**(7), 799–815.

Ghali, M., Krieg, J. M. & Rao, K. S. (2011), 'A Bayesian extension of the J-test for non-nested hypotheses', *Journal of Quantitative Economics* **9**(1), 53–72.

Godfrey, L. (1998), 'Tests of non-nested regression models: some results on small sample behaviour and the bootstrap', *Journal of Econometrics* **84**(1), 59–74.

Godfrey, L. G. (2011), 'Robust non-nested testing for ordinary least squares regression when some of the regressors are lagged dependent variables', *Oxford Bulletin of Economics and Statistics* **73**(5), 651–668.

Gouriéroux, C., Monfort, A. & Trognon, A. (1983), 'Testing nested or nonnested hypotheses', *Journal of Econometrics* **21**(1), 83–115.

Greene, W. (1994), Accounting for excess zeros and sample selection in Poisson and negative binomial regression models, Working papers, New York University, Leonard N. Stern School of Business, Department of Economics.

Hagemann, A. (2012), 'A simple test for regression specification with non-nested alternatives', *Journal of Econometrics* **166**(2), 247–254.

Hall, D. B. (2000), 'Zero-inflated poisson and binomial regression with random effects: A case study', *Biometrics* **56**(4), 1030–1039.

Hastie, T. J. & Tibshirani, R. J. (1986), 'Generalized additive models', *Statistical Science* **1**(3), 297–318.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Jørgensen, B. (1997), *The Theory of Dispersion Models*, Chapman and Hall, London.

Kelejian, H. H. (2008), 'A spatial J-test for model specification against a single or a set of non-nested alternatives', *Letters in Spatial and Resource Sciences* **1**(1), 3–11.

Kelejian, H. H. & Piras, G. (2014), 'An extension of the J-test to a spatial panel data framework', *Journal of Applied Econometrics* **31**(2), 387–402.

Lambert, D. (1992), 'Zero-inflated poisson regression, with an application to defects in manufacturing', *Technometrics* **34**(1), 1–14.

Lawley, D. (1956), 'A general method for approximating to the distribution of likelihood ratio criteria', *Biometrika* **43**, 295–303.

Lehmann, E. L. & Casella, G. (2002), *Theory of Point Estimation*, 2nd edn, Springer, New York.

Lv, J. & Liu, J. S. (2014), 'Model selection principles in misspecified models', *Journal of the Royal Statistical Society B* **76**(1), 141–167.

Marínez, R. O. (2008), Modelos de regressão beta inflacionados, PhD thesis, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

McAleer, M. (1995), 'The significance of testing empirical non-nested models', *Journal of Econometrics* **67**(1), 149–171.

McCullagh, P. (1987), *Tensor Methods in Statistics*, Chapman and Hall, London.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

McFadden, D. (1974), *Frontiers in Econometrics*, Academic Press, New York, chapter Conditional logit analysis of qualitative choice behavior, pp. 105–142.

Michelis, L. (1999), 'The distribution of the J and Cox non-nested tests in regression models with weakly correlated regressors', *Journal of Econometrics* **93**(2), 369–401.

Miyashiro, E. S. (2008), Modelos de regressão beta e simplex para análise de proporções, Master's thesis, USP, São Paulo.

Mizon, G. E. & Richard, J. F. (1986), 'The encompassing principle and its applications to non-nested hypotheses', *Econometrica* **54**(3), 657–678.

Nagelkerke, N. J. D. (1991), 'A note on a general definition of the coefficient of determination', *Biometrika* **78**(3), 691–692.

Nelder, J. M. K. & Wedderburn, R. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society A* **135**(3), 370–384.

Neter, J., Kutner, M. H., Naschtheim, C. J. & Wasserman, W. (1996), *Appplied Linear Statistical Models*, 4 edn, McGraw Hill, Chicago.

Ospina, R. & Ferrari, S. L. P. (2010), 'Inflated beta distributions', *Statistical Papers* **51**, 111–216.

Ospina, R. & Ferrari, S. L. P. (2012), 'A general class of zero-or-one inflated regression models', *Computational Statistics & Data Analysis* **56**(6), 1609–1623.

Pace, L. & Salvan, A. (1997), *Principles of Statistical Inference from a Neo-Fisherian Perspective*, Vol. 4, World Scientific Publishing Co., Singapore.

Paul, S. R., Jiang, X., Rai, S. N. & Balasooriya, U. (2004), 'Test of treatment effect in pre-drug and post-drug count data with zero-inflation', *Statistics in Medicine* **23**(10), 1541–1554.

Pearson, K. (1896), 'Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia', *Philosophical Transactions of the Royal Society of London* **187**, 253–318.

Pesaran, M. H. & Weeks, M. (2001), *A Companion to Theoretical Econometrics*, Blackwell, Malden, chapter Nonnested hypothesis testing: an overview, pp. 279–309.

Piras, G. & Lozano-Garcia, N. (2012), 'Spatial J-test: some Monte Carlo evidence', *Statistics and Computing* **22**(1), 169–183.

Pregibon, D. (1981), 'Logistic regression diagnostics', *Annals of Statistics* **9**, 705–724.

Press, W. H., Teulosky, S. A., Vetterling, W. T. & Flannery, B. P. (1992), *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edn, Cambridge.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramalho, E. A., Ramalho, J. J. S. & Murteira, J. M. R. (2011), 'Alternative estimating and testing empirical strategies for fractional regression models', *Journal of Economic Surveys* **25**(1), 19–68.

Rigby, B., Stasinopoulos, M., Heller, G. & Voudouris, V. (2014), *The distribution toolbox of GAMLSS*, The GAMLSS Team, London.

Rigby, R. A. & Stasinopoulos, D. M. (2001), The GAMLSS project: a flexible approach to statistical modelling, *in* B. Klein & L. Korsholm, eds, 'New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling', Odense, Denmark, pp. 337–345.

Rigby, R. A. & Stasinopoulos, D. M. (2005), 'Generalized additive models for location, scale and shape (with discussion)', *Applied Statistics* **54**(3), 507–554.

Rigby, R. A. & Stasinopoulos, D. M. (2006), 'Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis', *Statistical Modelling* **6**(3), 209–229.

Rigby, R. & Stasinopoulos, D. (1996*a*), 'A semi-parametric additive model for variance heterogeneity', *Statistics and Computing* **6**(1), 57–65.

Rigby, R. & Stasinopoulos, M. (1996*b*), Mean and dispersion additive models, *in* W. Härdle & M. Schimek, eds, 'Statistical Theory and Computational Aspects of Smoothing', Contributions to Statistics, Physica-Verlag HD, London, pp. 215–230.

Rocha, A. V. & Simas, A. B. (2011), 'Influence diagnostics in a general class of beta regression models', *Test* **20**, 95–119.

Rudin, W. (1976), *Principles of Mathematical Analysis*, 3rd edn, McGraw-Hill.

Sapra, S. K. (2008), 'Robust nonnested hypothesis testing', *Applied Economics Letters* **15**(1), 1–4.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.

Sen, P. & Singer, J. M. (1993), *Large Sample Methods in Statistics: An Introduction With Applications*, Chapman and Hall, New York.

Simas, A. B., Barreto-Souza, W. & Rocha, A. V. (2010), 'Improved estimators for a general class of beta regression models', *Computational Statistics and Data Analysis* **54**(2), 348–366.

Smithson, M. & Verkuilen, J. (2006), 'A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables', *Psychological Methods* **11**(1), 54–71.

Snell, D. C. C. (1968), 'A general definition of residuals', *Journal of the Royal Statistical Society B* **30**, 248–275.

Song, P. X.-K. & Tan, M. (2000), 'Marginal models for longitudinal continuous proportional data', *Biometrics* **56**, 496–502.

Song, P. X., Qiu, Z. & Tan, M. (2004), 'Modelling heterogeneous dispersion in marginal models for longitudinal proportional data', *Biometrical Journal* **46**(5), 540–553.

Song, X. K. (2007), *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer, New York.

Stasinopoulos, D. M. & Rigby, R. A. (2008), 'Generalized additive models for location scale and shape (GAMLSS) in R', *Journal of Statistical Software* **23**(7), 1–46.

Tang, N.-S., Wei, B.-C. & Wang, X.-R. (2000), 'Influence diagnostics in nonlinear reproductive dispersion models', *Statistics & Probability Letters* **46**(1), 59–68.

Tong, E. N., Mues, C. & Thomas, L. (2013), 'A zero-adjusted gamma model for mortgage loan loss given default', *International Journal of Forecasting* **29**(4), 548 – 562.

Vieira, A. M. C., Hinde, J. P. & Demetrio, C. G. B. (2000), 'Zero-inflated proportion data models applied to a biological control assay', *Journal of Applied Statistics* **27**(3), 373–389.

Williams, D. A. (1984), Residuals in generalized linear models, *in* 'Proceedings of the 12th International Biometrics Conference', Tokyo, pp. 59–68.

Williams, D. A. (1987), 'Generalized linear model diagnostic using the deviance and single case deletion', *Applied Statistics* **36**, 181–191.

Wooldridge, J. M. (1990), 'A unified approach to robust, regression-based specification tests', *Econometric Theory* **6**(1), 17–43.

Zhang, P., Qiu, Z. & Shi, C. (2016), 'simplexreg: An R package for regression analysis of proportional data using the simplex distribution', *Journal of Statistical Software* **71**(11), 1–21.

APPENDIX A

Appendices of Chapter 2

A.1 First order derivatives of the log-likelihood function for the simplex regression inflated at $c = 0$ or $c = 1$

From the separability of the vectors of parameters φ and $(\beta^\top, \sigma^2)^\top$, the score function for φ can be expressed independently from the score function for $(\beta^\top, \sigma^2)^\top$. From the log-likelihood function in (2.10), the score function for φ_R, for $R = 1, \ldots, M$, is

$$U_{\varphi_R} = \frac{\partial \ell_1(\varphi)}{\partial \varphi_R} = \sum_{t=1}^{n} \frac{\partial \ell_t(\alpha_t)}{\partial \alpha_t} \frac{d\alpha_t}{d\zeta_t} \frac{\partial \zeta_t}{\partial \varphi_R},$$

where

$$\frac{\partial \ell_t(\alpha_t)}{\partial \alpha_t} = \frac{\mathbb{1}_{\{c\}}(y_t) - \alpha_t}{\alpha_t(1-\alpha_t)}, \qquad \frac{d\alpha_t}{d\zeta_t} = \frac{dh^{-1}(\zeta_t)}{d\zeta_t} = \frac{1}{h'(\alpha_t)} \qquad \text{and} \qquad \frac{\partial \zeta_t}{\partial \varphi_R} = \frac{\partial h(\alpha_t)}{\partial \varphi_R} = z_{tR}.$$

Here, h' is the first order derivative of the link function h. Therefore, the score function for φ_R can be expressed as

$$U_{\varphi_R} = \sum_{t=1}^{n} \frac{\mathbb{1}_{\{c\}}(y_t) - \alpha_t}{\alpha_t(1-\alpha_t)} \frac{1}{h'(\alpha_t)} z_{tR}. \tag{A.1}$$

The score function for β_r, $r = 1, \ldots, m$, is

$$U_{\beta_r} = \frac{\partial \ell_2(\beta, \sigma^2)}{\partial \beta_r} = \sum_{t:y\in(0,1)} \frac{\partial \ell_t(\mu_t, \sigma^2)}{\partial \mu_t} \frac{d\mu_t}{d\eta_t} \frac{\partial \eta_t}{\partial \beta_r},$$

where

$$\frac{\partial \ell_t(\mu_t, \sigma^2)}{\partial \mu_t} = -\frac{1}{2\sigma^2} d'(y_t, \mu_t), \quad \text{with} \;\; d'(y_t; \mu_t) = \frac{\partial d(y_t; \mu_t)}{\partial \mu_t}.$$

Let

$$u_t = -\frac{1}{2} d'(y_t; \mu_t) = \frac{(y_t - \mu_t)(y_t - 2\mu_t y_t + \mu_t^2)}{y_t(1-y_t)\mu_t^3(1-\mu)^3}. \tag{A.2}$$

Then

$$\frac{\partial \ell_t(\mu_t, \sigma^2)}{\partial \mu_t} = \sigma^{-2} u_t.$$

Furthermore,

$$\frac{d\mu_t}{\eta_t} = \frac{dg^{-1}(\eta_t)}{d\eta_t} = \frac{1}{g'(\mu_t)} \quad \text{and} \quad \frac{\partial \eta_t}{\partial \beta_r} = \frac{\partial g(\mu_t)}{\beta_r} = x_{tr},$$

where g' is the first derivative of the link function g. The score function U_{β_r} is then given by

$$U_{\beta_r} = \sum_{t=1}^{n} \left(1 - \mathbb{1}_{\{c\}}(y_t)\right) \sigma^{-2} u_t \frac{1}{g'(\mu_t)} x_{tr}. \tag{A.3}$$

The score function for σ^2 is

$$U_{\sigma^2} = \frac{\partial \ell_2(\beta, \sigma^2)}{\partial \sigma^2} = \sum_{t:y\in(0,1)} \frac{\partial \ell_t(\mu_t, \sigma^2)}{\partial \sigma^2},$$

where

$$\frac{\partial \ell_t(\mu_t, \sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} d(y_t; \mu_t),$$

where $d(y_t;\mu_t)$ is given by (2.4). The score function U_{σ^2} is then given by

$$U_{\sigma^2} = \sum_{t=1}^{n}(1-\mathbb{1}_{\{c\}}(y_t))\left[-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}d(y_t;\mu_t)\right]. \tag{A.4}$$

The score functions in (A.1), (A.3) and (A.4) are presented in matrix form in section 2.4.1.

A.2 Second order derivatives and cumulants of the log-likelihood function for the simplex regression inflated at $c = 0$ or $c = 1$

Fisher's information matrix is obtained from the moments of the second order log-likelihood derivatives. We use the notation of Lawley (1956), in which $-\kappa_{rs} = \kappa_{r,s}$ denotes the (r, s) element of Fisher's information matrix $K(\theta)$.

The second order derivative of the log-likelihood function in (2.10) with respect to φ_R, $R = 1, \ldots, M$, and φ_S is

$$\begin{aligned}
U_{\varphi_R\varphi_S} &= \frac{\partial^2\ell_1(\varphi)}{\partial\varphi_R\varphi_S} = \sum_{t=1}^{n}\frac{\partial}{\partial\alpha_t}\left(\frac{\partial\ell_t(\alpha_t)}{\partial\alpha_t}\frac{d\alpha_t}{d\zeta_t}\frac{\partial\zeta_t}{\partial\varphi_R}\right)\frac{d\alpha_t}{d\zeta_t}\frac{\partial\zeta_t}{\partial\varphi_S} \\
&= \sum_{t=1}^{n}\left\{\frac{\partial^2\ell_t(\alpha_t)}{\partial\alpha_t^2}\left(\frac{d\alpha_t}{d\zeta_t}\right)^2+\frac{\partial\ell_t(\alpha_t)}{\partial\alpha_t}\left(\frac{\partial}{\partial\alpha_t}\frac{d\alpha_t}{d\zeta_t}\right)\frac{d\alpha_t}{d\zeta_t}\right\}z_{tS}z_{tR} \\
&= \sum_{t=1}^{n}\left\{\left(\frac{-\mathbb{1}_{(0,1)}(y_t)}{(1-\alpha_t)^2}-\frac{\mathbb{1}_{\{c\}}(y_t)}{\alpha_t^2}\right)\left(\frac{d\alpha_t}{d\zeta_t}\right)^2\right. \\
&\quad \left.+\left(\frac{\mathbb{1}_{\{c\}}(y_t)}{\alpha_t}-\frac{\mathbb{1}_{(0,1)}(y_t)}{1-\alpha_t}\right)\left(\frac{\partial}{\partial\alpha_t}\frac{d\alpha_t}{d\zeta_t}\right)\frac{d\alpha_t}{d\zeta_t}\right\}z_{tS}z_{tR}.
\end{aligned}$$

Note that $\mathbb{1}_{(0,1)}(y_t) = 1-\mathbb{1}_{\{c\}}(y_t)$.

Under the regularity conditions, $\mathrm{E}(\partial\ell_t(\alpha_t))/\partial\alpha_t = 0$. Thus,

$$\begin{aligned}
\kappa_{\varphi_R\varphi_S} = \mathrm{E}(U_{\varphi_R\varphi_S}) &= \mathrm{E}\left[\sum_{t=1}^{n}\left\{\frac{\partial^2\ell_t(\alpha_t)}{\partial\alpha_t^2}\left(\frac{d\alpha_t}{d\zeta_t}\right)^2 + \frac{\partial\ell_t(\alpha_t)}{\partial\alpha_t}\left(\frac{\partial}{\partial\alpha_t}\frac{d\alpha_t}{d\zeta_t}\right)\frac{d\alpha_t}{d\zeta_t}\right\} z_{tS}z_{tR}\right] \\
&= \sum_{t=1}^{n}\left\{\mathrm{E}\left[\frac{\partial^2\ell_t(\alpha_t)}{\partial\alpha_t^2}\left(\frac{d\alpha_t}{d\zeta_t}\right)^2 z_{tS}z_{tR}\right] + \mathrm{E}\left[\frac{\partial\ell_t(\alpha_t)}{\partial\alpha_t}\left(\frac{\partial}{\partial\alpha_t}\frac{d\alpha_t}{d\zeta_t}\right)\frac{d\alpha_t}{d\zeta_t} z_{tS}z_{tR}\right]\right\} \\
&= \sum_{t=1}^{n}\mathrm{E}\left[\frac{\partial^2\ell_t(\alpha_t)}{\partial\alpha_t^2}\left(\frac{d\alpha_t}{d\zeta_t}\right)^2 z_{tS}z_{tR}\right] \\
&= \sum_{t=1}^{n}\mathrm{E}\left[\left(\frac{-\mathbb{1}_{(0,1)}(y_t)}{(1-\alpha_t)^2} - \frac{\mathbb{1}_{\{c\}}(y_t)}{\alpha_t^2}\right)\left(\frac{d\alpha_t}{d\zeta_t}\right)^2 z_{tS}z_{tR}\right].
\end{aligned}$$

Note that $\mathrm{E}(\mathbb{1}_{\{c\}}(y_t)) = \alpha_t$ and $\mathrm{E}(1 - \mathbb{1}_{\{c\}}(y_t)) = 1 - \alpha_t$. Then, the expression above reduces to

$$\kappa_{\varphi_R\varphi_S} = -\sum_{t=1}^{n}\frac{1}{\alpha_t(1-\alpha_t)}\left(\frac{1}{h'(\alpha_t)}\right)^2 z_{tS}z_{tR}.$$

The cross second order derivative of the log-likelihood function in (2.10) with respect to β_R, $r = 1, \ldots, m$, and β_s is

$$\begin{aligned}
U_{\beta_r\beta_s} = \frac{\partial^2\ell_2(\beta,\sigma^2)}{\partial\beta_r\beta_s} &= \sum_{t:y_t\in(0,1)}\frac{\partial}{\partial\mu_t}\left(\frac{\partial\ell_t(\mu_t,\sigma^2)}{\partial\mu_t}\frac{d\mu_t}{d\eta_t}\frac{\partial\eta_t}{\partial\beta_r}\right)\frac{d\mu_t}{d\eta_t}\frac{\partial\eta_t}{\partial\beta_s} \\
&= \sum_{t:y_t\in(0,1)}\left\{\frac{\partial^2\ell_t(\mu_t,\sigma^2)}{\partial\mu_t^2}\left(\frac{d\mu_t}{d\eta_t}\right)^2 + \frac{\partial\ell_t(\mu_t,\sigma^2)}{\partial\mu_t}\left(\frac{\partial}{\partial\mu_t}\frac{d\mu_t}{d\eta_t}\right)\frac{d\mu_t}{d\eta_t}\right\} x_{ts}x_{tr},
\end{aligned} \tag{A.5}$$

where

$$\frac{\partial^2\ell_t(\mu_t,\sigma^2)}{\partial\mu_t^2} = -\frac{1}{2\sigma^2}d''(y_t;\mu_t),$$

with

$$\begin{aligned}
\frac{1}{2}d''(y_t;\mu_t) &= \frac{1}{2}\frac{\partial^2 d(y_t;\mu_t)}{\partial\mu_t^2} \\
&= \frac{1}{\mu_t(1-\mu_t)} + \frac{(1-2\mu_t)}{\mu_t^2(1-\mu_t)^2}(y_t-\mu_t)d(y_t;\mu_t) \\
&+ \frac{1}{\mu_t^3(1-\mu_t)^3} + \frac{1-2\mu_t}{\mu_t^4(1-\mu_t)^4}(y_t-\mu_t) \\
&- \frac{1}{\mu_t(1-\mu_t)}(y_t-\mu_t)d'(y_t;\mu_t) - \frac{2(2\mu_t-1)}{\mu_t^4(1-\mu_t)^4}(y_t-\mu_t),
\end{aligned} \tag{A.6}$$

where $d(y_t;\mu_t)$ is given in (2.4).

In order to obtain the cumulants involving β and σ^2, the following proposition shall be proved. It can be found in PhD thesis of Raydonal Ospina Martínez (Marínez 2008, p. 108, in portuguese).

Proposition 1. *Let $(y_1,\ldots,y_n)^\top$ be a vector of n independent random variables where y_t follows the inflated simplex distribution with p.d.f. given in (2.6), i.e., $y_t \sim IS_c(\alpha_t,\mu_t,\sigma^2)$, $t=1,\ldots,n$. Let $\mathcal{I}:(0,1)\to\mathbb{R}$ be a continuous function. Thus,*

$$E\left(\sum_{t:y_t\in(0,1)}\mathcal{I}(y_t)\right)=\sum_{t=1}^{n}(1-\alpha_t)E\left(\mathcal{I}(y_t)|\mathbb{1}_{\{c\}}(y_t)=0\right).$$

Proof. The support of the $IS_c(\alpha_t,\mu_t,\sigma^2)$ is set $(0,1)\cup\{c\}$, where $c=0$ or $c=1$. Let

$$\mathcal{I}^*(y_t)=\begin{cases}0, & \text{if } y_t=c,\\ \mathcal{I}(y_t), & \text{if } y_t\in(0,1).\end{cases}$$

It then follows that

$$\mathrm{E}\left(\sum_{t:y_t\in(0,1)}\mathcal{I}(y_t)\right)=\mathrm{E}\left(\sum_{t=1}^{n}\mathcal{I}^*(y_t)\right)=\sum_{t=1}^{n}\mathrm{E}(\mathcal{I}^*(y_t)).$$

Thus,

$$\begin{aligned}\mathrm{E}(\mathcal{I}^*(y_t)) &= \mathrm{E}\left(\mathcal{I}^*(y_t)|\mathbb{1}_{\{c\}}(y_t)=0\right)\Pr\left(\mathbb{1}_{\{c\}}(y_t)=0\right)\\ &\quad+\mathrm{E}\left(\mathcal{I}^*(y_t)|\mathbb{1}_{\{c\}}(y_t)=1\right)\Pr\left(\mathbb{1}_{\{c\}}(y_t)=1\right)\\ &=\mathrm{E}\left(\mathcal{I}^*(y_t)|\mathbb{1}_{\{c\}}(y_t)=0\right)\Pr\left(\mathbb{1}_{\{c\}}(y_t)=0\right)\\ &=(1-\alpha_t)\mathrm{E}\left(\mathcal{I}(y_t)|\mathbb{1}_{\{c\}}(y_t)=0\right).\end{aligned}$$

The result follows. □

Under the standard regularity conditions, $\mathrm{E}(\partial\ell_t(\mu_t,\sigma^2)/\partial\mu_t)=0$. Using the result in

Proposition 1, we obtain

$$\begin{aligned}
\kappa_{\beta_r\beta_s} &= \mathrm{E}(U_{\beta_r\beta_s}) \\
&= \mathrm{E}\left[\sum_{t:y_t\in(0,1)}\left\{\frac{\partial^2\ell_t(\mu_t,\sigma^2)}{\partial\mu_t^2}\left(\frac{d\mu_t}{d\eta_t}\right)^2+\frac{\partial\ell_t(\mu_t,\sigma^2)}{\partial\mu_t}\left(\frac{\partial}{\partial\mu_t}\frac{d\mu_t}{d\eta_t}\right)\frac{d\mu_t}{d\eta_t}\right\}x_{ts}x_{tr}\right] \\
&= \sum_{t=1}^{n}(1-\alpha_t)\mathrm{E}\left[\frac{\partial^2\ell_t(\mu_t,\sigma^2)}{\partial\mu_t^2}\right]\left(\frac{d\mu_t}{d\eta_t}\right)^2 x_{ts}x_{tr} \\
&= -\sum_{t=1}^{n}(1-\alpha_t)\frac{1}{2\sigma^2}\mathrm{E}\left[d''(y_t;\mu_t)\right]\left(\frac{1}{g'(\mu_t)}\right)^2 x_{ts}x_{tr}.
\end{aligned}$$

From Equation (A.6),

$$\begin{aligned}
\frac{1}{2}\mathrm{E}\left[d''(y_t;\mu_t)\right] &= \frac{1}{\mu_t(1-\mu_t)}\left\{\mathrm{E}[d(y_t;\mu_t)]-\mathrm{E}[(y_t-\mu_t)d'(y_t;\mu_t)]\right\} \\
&\quad+\frac{1-2\mu_t}{\mu_t^2(1-\mu_t)^2}\mathrm{E}[(y_t-\mu_t)d(y_t;\mu_t)]+\frac{1}{\mu_t^3(1-\mu_t)^3} \\
&= \frac{3\sigma^2}{\mu_t(1-\mu_t)}+\frac{1}{\mu_t^3(1-\mu_t)^3},
\end{aligned}$$

because $\mathrm{E}[d(y_t;\mu_t)]=\sigma^2$, $\mathrm{E}[(y_t-\mu_t)d'(y_t;\mu_t)]=-2\sigma^2$ and $\mathrm{E}[(y_t-\mu_t)d(y_t;\mu_t)]=0$. Then,

$$\kappa_{\beta_r\beta_s} = -\frac{1}{\sigma^2}\sum_{t=1}^{n}(1-\alpha_t)a_t x_{ts}x_{tr},$$

where

$$a_t = \left(\frac{3\sigma^2}{\mu_t(1-\mu_t)}+\frac{1}{\mu_t 3(1-\mu_t)^3}\right)\left(\frac{1}{g'(\mu_t)}\right)^2.$$

The second derivative with respect to σ^2 of the log-likelihood function in (2.10) is

$$\begin{aligned}
U_{\sigma^2\sigma^2} = \frac{\partial^2\ell_2(\beta,\sigma^2)}{\partial\sigma^4} &= \sum_{t:y_t\in(0,1)}\frac{\partial^2\ell_t(\mu_t,\sigma^2)}{\partial\sigma^4} \\
&= \sum_{t:y_t\in(0,1)}\frac{\partial}{\partial\sigma^2}\left(-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}d(y_t;\mu_t)\right) \\
&= \sum_{t:y_t\in(0,1)}\left(\frac{1}{2\sigma^4}-\frac{1}{\sigma^6}d(y_t;\mu_t)\right).
\end{aligned} \tag{A.7}$$

Since $E[d(y_t;\mu_t)] = \sigma^2$ and using the result in the Proposition 1,

$$\kappa_{\sigma^2\sigma^2} = \mathrm{E}\left[U_{\sigma^2\sigma^2}\right] = -\sum_{t=1}^{n}(1-\alpha_t)\frac{1}{2\sigma^4}.$$

By differentiating (2.10) with respect to β_r and σ^2, we obtain

$$U_{\beta_r\sigma^2} = \frac{\partial^2\ell_2(\beta,\sigma^2)}{\partial\beta_r\partial\sigma^2} = \sum_{t:y_t\in(0,1)}\frac{\partial^2\ell_t(\mu_t,\sigma^2)}{\partial\mu_t\partial\sigma^2}\frac{d\mu_t}{d\eta_t}\frac{\partial\eta_t}{\partial\beta_r} = -\frac{1}{\sigma^4}\sum_{t:y_t\in(0,1)}u_t\frac{d\mu_t}{d\eta_t}x_{tr}, \tag{A.8}$$

where u_t is given in Equation (A.2). Then, using the result in the Proposition 1,

$$\kappa_{\beta_r\sigma^2} = \mathrm{E}(U_{\beta_r\sigma^2}) = -\frac{1}{\sigma^4}\sum_{t=1}^{n}(1-\alpha_t)\mathrm{E}(u_t)\frac{d\mu_t}{d\eta_t}x_{tr}. \tag{A.9}$$

We have

$$\mathrm{E}(u_t) = \mathrm{E}\left[\frac{y_t\mu_t}{\mu_t(1-\mu_t)}\left\{d(y_t;\mu_t) + \frac{1}{\mu^2(1-\mu)^2}\right\}\right].$$

Using the fact that $\mathrm{E}[d(y_t;\mu_t)] = \sigma^2$ and $\mathrm{E}[(y_t-\mu_t)d(y_t;\mu_t)] = 0$, it follows that

$$\mathrm{E}(u_t) = \frac{1}{\mu_t(1-\mu_t)}\left[\mathrm{E}[(y_t-\mu_t)d(y_t;\mu_t)] + \frac{1}{\mu^2(1-\mu)^2}\mathrm{E}(y_t-\mu_t)\right] = 0. \tag{A.10}$$

Then, using Equations (A.9) and (A.10), we obtain

$$\kappa_{\beta_r\sigma^2} = 0.$$

From the separability of φ and $(\beta^\top,\sigma^2)^\top$ we have that $U_{\varphi_R\sigma^2} = U_{\beta_r\varphi_R} = 0$ and hence $\kappa_{\varphi_R\sigma^2} = \kappa_{\beta_r\varphi_R} = 0$.

A.3 First order derivatives of the log-likelihood function for the zero and one inflated simplex regression model

Consider the vector of parameters $(\rho^\top, \varphi^\top)^\top$. The first order derivatives of $\ell_1(\rho, \varphi)$ which is in (2.20), for $r' = 1, \ldots, k_0$ and $r'' = 1, \ldots, k_1$, are

$$
\begin{aligned}
U_{\rho_{r'}} = \frac{\partial \ell_1(\rho,\varphi)}{\partial \rho_{r'}} &= \sum_{t=1}^{n} \left\{ \frac{\partial \ell_t(\delta_{0t},\delta_{1t})}{\partial \delta_{0t}} \frac{\partial \delta_{0t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} + \frac{\partial \ell_t(\delta_{0t},\delta_{1t})}{\partial \delta_{1t}} \frac{\partial \delta_{1t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} \right\} \\
&= \sum_{t=1}^{n} \left\{ \frac{\mathbb{1}_{\{0\}}(y_t)}{\delta_{0t}} - \frac{\mathbb{1}_{(0,1)}(y_t)}{1-\delta_{0t}-\delta_{1t}} \right\} \frac{\partial \delta_{0t}}{\partial \zeta_{0t}} \nu_{tr'} + \sum_{t=1}^{n} \left\{ \frac{\mathbb{1}_{\{1\}}(y_t)}{\delta_{1t}} - \frac{\mathbb{1}_{(0,1)}(y_t)}{1-\delta_{0t}-\delta_{1t}} \right\} \frac{\partial \delta_{1t}}{\partial \zeta_{1t}} \nu_{tr'}
\end{aligned}
$$

and

$$
\begin{aligned}
U_{\varphi_{r''}} = \frac{\partial \ell_1(\rho,\varphi)}{\partial \varphi_{r''}} &= \sum_{t=1}^{n} \left\{ \frac{\partial \ell_t(\delta_{0t},\delta_{1t})}{\partial \delta_{0t}} \frac{\partial \delta_{0t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} + \frac{\partial \ell_t(\delta_{0t},\delta_{1t})}{\partial \delta_{1t}} \frac{\partial \delta_{1t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} \right\} \\
&= \sum_{t=1}^{n} \left\{ \frac{\mathbb{1}_{\{0\}}(y_t)}{\delta_{0t}} - \frac{\mathbb{1}_{(0,1)}(y_t)}{1-\delta_{0t}-\delta_{1t}} \right\} \frac{\partial \delta_{0t}}{\partial \zeta_{1t}} z_{tr''} + \sum_{t=1}^{n} \left\{ \frac{\mathbb{1}_{\{1\}}(y_t)}{\delta_{1t}} - \frac{\mathbb{1}_{(0,1)}(y_t)}{1-\delta_{0t}-\delta_{1t}} \right\} \frac{\partial \delta_{1t}}{\partial \zeta_{1t}} z_{tr''}.
\end{aligned}
$$

From Equation (2.20), the first order derivatives of $\ell_2(\beta, \sigma^2)$ are

$$
U_{\beta_r} = \sigma^{-2} \sum_{t=1}^{n} \mathbb{1}_{(0,1)}(y_t) u_t \frac{1}{g'(\mu_t)} x_{tr} \tag{A.11}
$$

and

$$
U_{\sigma^2} = \sum_{t=1}^{n} \mathbb{1}_{(0,1)}(y_t) \left[-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} d(y_t;\mu_t) \right]. \tag{A.12}
$$

Note that (A.11) and (A.12) are the same quantities given in Equations (A.3) and (A.4), respectively.

Expressions for $U_{\rho_{r'}}$, $U_{\varphi_{r''}}$, U_{β_r} and U_{σ^2} in matrix form are presented in Section 2.5.1.

A.4 Second order derivatives of the log-likelihood function for the zero and one inflated simplex regression model

The second order derivatives of the function $\ell_1(\rho, \varphi)$ in (2.20) are given by

$$
\begin{aligned}
U_{\rho_{r'}\rho_{s'}} = \frac{\partial^2 \ell_1(\rho,\varphi)}{\partial \rho_{r'} \partial \rho_{s'}} = \sum_{t=1}^{n} \Bigg\{ & \left[\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t}^2} \frac{\partial \delta_{0t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{s'}} + \frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t} \partial \delta_{1t}} \frac{\partial \delta_{1t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{s'}} \right] \frac{\partial \delta_{0t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} \\
& + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t}} \frac{\partial^2 \delta_{0t}}{\partial \zeta_{0t}^2} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} \frac{\partial \zeta_{0t}}{\partial \rho_{s'}} + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}} \frac{\partial \delta_{0t}}{\partial \zeta_{0t}} \frac{\partial^2 \zeta_{0t}}{\partial \rho_{s'} \partial \rho_{r'}} \\
& + \left[\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t} \partial \delta_{0t}} \frac{\partial \delta_{0t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{s'}} + \frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}^2} \frac{\partial \delta_{1t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{s'}} \right] \frac{\partial \delta_{1t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} \\
& + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}} \frac{\partial^2 \delta_{1t}}{\partial \zeta_{0t}^2} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} \frac{\partial \zeta_{0t}}{\partial \rho_{s'}} + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}} \frac{\partial \delta_{1t}}{\partial \delta_{0t}} \frac{\partial^2 \zeta_{0t}}{\partial \rho_{s'} \partial \rho_{r'}} \Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
U_{r''s''} = \frac{\partial^2 (\rho,\varphi)}{\partial \varphi_{r''} \partial \varphi_{s''}} = \sum_{t=1}^{n} \Bigg\{ & \left[\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t}^2} \frac{\partial \delta_{0t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{s''}} + \frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t} \partial \delta_{1t}} \frac{\partial \delta_{1t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{s''}} \right] \frac{\partial \delta_{0t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} \\
& + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t}} \frac{\partial^2 \delta_{0t}}{\partial \zeta_{1t}^2} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} \frac{\partial \zeta_{1t}}{\partial \varphi_{s''}} + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t}} \frac{\partial \delta_{0t}}{\partial \zeta_{1t}} \frac{\partial^2 \zeta_{1t}}{\partial \varphi_{s''} \partial \varphi_{r''}} \\
& + \left[\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t} \partial \delta_{0t}} \frac{\partial \delta_{0t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{s''}} + \frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}^2} \frac{\partial \delta_{1t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{s''}} \right] \frac{\partial \delta_{1t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} \\
& + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}} \frac{\partial^2 \delta_{1t}}{\partial \zeta_{1t}^2} \frac{\partial \zeta_{0t}}{\partial \varphi_{r''}} \frac{\partial \zeta_{1t}}{\partial \varphi_{s''}} + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}} \frac{\partial \delta_{1t}}{\partial \delta_{1t}} \frac{\partial^2 \zeta_{1t}}{\partial \varphi_{s''} \partial \varphi_{r''}} \Bigg\}
\end{aligned}
$$

and

$$
\begin{aligned}
U_{\rho_{r'}\varphi_{r''}} = \frac{\partial^2 \ell_1(\rho,\varphi)}{\partial \rho_{r'} \partial \varphi_{r''}} = \sum_{t=1}^{n} \Bigg\{ & \left[\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t}^2} \frac{\partial \delta_{0t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} + \frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t} \partial \delta_{1t}} \frac{\partial \delta_{1t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} \right] \frac{\partial \delta_{0t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} \\
& + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t}} \frac{\partial^2 \delta_{0t}}{\partial \zeta_{1t} \partial \zeta_{0t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} + \frac{\partial \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}} \frac{\partial^2 \delta_{1t}}{\partial \zeta_{0t} \partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} \\
& + \left[\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t} \partial \delta_{0t}} \frac{\partial \delta_{0t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} + \frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}^2} \frac{\partial \delta_{1t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \varphi_{r''}} \right] \frac{\partial \delta_{1t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_{r'}} \Bigg\},
\end{aligned}
$$

where

$$\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t}^2} = -\frac{\mathbb{1}_{\{0\}}(y_t)}{\delta_{0t}^2} + \frac{\mathbb{1}_{(0,1)}(y_t)}{(1-\delta_{0t}-\delta_{1t})^2},$$
$$\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}^2} = -\frac{\mathbb{1}_{\{1\}}(y_t)}{\delta_{0t}^2} + \frac{\mathbb{1}_{(0,1)}(y_t)}{(1-\delta_{0t}-\delta_{1t})^2},$$
$$\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}\delta_{0t}} = \frac{\mathbb{1}_{(0,1)}(y_t)}{(1-\delta_{0t}-\delta_{1t})^2}.$$

The second order derivatives of the function $\ell_2(\beta, \sigma^2)$ in (2.20), i.e., $U_{\beta_r\beta_s}$, $U_{\sigma^2\sigma^2}$ and $U_{\beta_r\sigma^2}$, are the same as those given in (A.5), (A.7) and (A.8), respectively. From the separability of the parameters $(\rho^\top, \varphi^\top)^\top$ and $(\beta^\top, \sigma^2)^\top$ we have that

$$U_{\beta_r\rho_{r'}} = \frac{\partial \ell_1(\rho, \varphi)}{\partial \beta_r \partial \rho_{r'}} = 0,$$
$$U_{\beta_r\varphi_{r''}} = \frac{\partial \ell_1(\rho, \varphi)}{\partial \beta_r \partial \varphi_{r''}} = 0,$$
$$U_{\sigma^2\rho_{r'}} = \frac{\partial \ell_1(\rho, \varphi)}{\partial \sigma^2 \partial \rho_{r'}} = 0,$$
$$U_{\sigma^2\varphi_{r''}} = \frac{\partial \ell_1(\rho, \varphi)}{\partial \sigma^2 \partial \varphi_{r''}} = 0.$$

A.5 Cumulants of the log-likelihood function for the zero and one inflated simplex regression model

In order to obtain the cumulants, the following results are useful. Under some regularity conditions (Lehmann & Casella 2002), we have

$$E\left(\frac{\partial \ell_t(\mu_t, \sigma^2)}{\partial \delta_{0t}}\right) = 0,$$
$$E\left(\frac{\partial \ell_t(\mu_t, \sigma^2)}{\partial \delta_{1t}}\right) = 0,$$
$$E\left(\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{0t}^2}\right) = -\frac{1}{\delta_{0t}} + \frac{1}{1-\delta_{0t}-\delta_{1t}},$$
$$E\left(\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}^2}\right) = -\frac{1}{\delta_{1t}} + \frac{1}{1-\delta_{0t}-\delta_{1t}},$$
$$E\left(\frac{\partial^2 \ell_t(\delta_{0t}, \delta_{1t})}{\partial \delta_{1t}\partial \delta_{0t}}\right) = \frac{1}{1-\delta_{0t}-\delta_{1t}},$$

Another useful result is that stated in Proposition 2. It is useful to obtain the cumulants that involve $(\beta^\top, \sigma^2)^\top$. It can be found in PhD thesis of Raydonal Ospina Martínez (Marínez 2008, p. 115, in portuguese).

Proposition 2. *Let $(y_1, \ldots, y_n)^\top$ be a vector of n independent random variables, where y_t follows the zero and one inflated simplex distribution with p.d.f. given in (2.7), i.e., $y_t \sim ZOIS(\delta_{0t}, \delta_{1t}, \mu_t, \sigma^2)$, $t = 1, \ldots, n$. Let $\mathcal{I} : (0,1) \to \mathbb{R}$ is a continuous function. Thus,*

$$E\left(\sum_{t:y_t\in(0,1)} \mathcal{I}(y_t)\right) = \sum_{t=1}^{n}(1-\delta_{0t}-\delta_{1t})E\left(\mathcal{I}(y_t)|\mathbb{1}_{\{c\}}(y_t)=0\right).$$

Proof. Let

$$\mathcal{I}^*(y_t) = \begin{cases} 0, & \text{if } y_t \in \{0,1\}, \\ \mathcal{I}(y_t), & \text{if } y_t \in (0,1). \end{cases}$$

Then

$$\mathrm{E}\left(\sum_{t:y_t\in(0,1)} \mathcal{I}(y_t)\right) = \mathrm{E}\left(\sum_{t=1}^{n}\mathcal{I}^*(y_t)\right) = \sum_{t=1}^{n}\mathrm{E}(\mathcal{I}^*(y_t)).$$

Additionally,

$$\begin{aligned}
\mathrm{E}(\mathcal{I}^*(y_t)) &= \mathrm{E}\left(\mathcal{I}^*(y_t)|\mathbb{1}_{(0,1)}(y_t)=0\right)\Pr\left(\mathbb{1}_{(0,1)}(y_t)=0\right) \\
&\quad + \mathrm{E}\left(\mathcal{I}^*(y_t)|\mathbb{1}_{(0,1)}(y_t)=1\right)\Pr\left(\mathbb{1}_{(0,1)}(y_t)=1\right) \\
&= \mathrm{E}\left(\mathcal{I}^*(y_t)|\mathbb{1}_{(0,1)}(y_t)=1\right)\Pr\left(\mathbb{1}_{(0,1)}(y_t)=1\right) \\
&= (1-\delta_{0t}-\delta_{1t})\mathrm{E}\left(\mathcal{I}(y_t)|\mathbb{1}_{(0,1)}(y_t)=1\right).
\end{aligned}$$

The result follows. □

We now obtain some log-likelihood cumulants:

$$
\begin{aligned}
\kappa_{\rho_{r'}\rho_{s'}} &= \mathrm{E}(U_{\rho_{r'}\rho_{s'}}) \\
&= \sum_{t=1}^{n}\left\{\left[\left(-\frac{1}{\delta_{0t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{0t}}{\partial\zeta_{0t}}\nu_{ts'}+\left(\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{1t}}{\partial\zeta_{0t}}\nu_{ts'}\right]\frac{\partial\delta_{0t}}{\partial\zeta_{0t}}\nu_{tr'}\right. \\
&\quad \left.+\left[\left(\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{0t}}{\partial\zeta_{0t}}\nu_{ts'}+\left(-\frac{1}{\delta_{1t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{1t}}{\partial\zeta_{0t}}\nu_{ts'}\right]\frac{\partial\delta_{1t}}{\partial\zeta_{0t}}\nu_{tr'}\right\} \\
&= \sum_{t=1}^{n}\left(-\frac{1}{\delta_{0t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\left(\frac{\partial\delta_{0t}}{\partial\zeta_{0t}}\right)^2\nu_{tr'}\nu_{ts'} \\
&\quad +2\sum_{t=1}^{n}\left(\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{1t}}{\partial\zeta_{0t}}\frac{\partial\delta_{0t}}{\partial\zeta_{0t}}\nu_{tr'}\nu_{ts'} \\
&\quad +\sum_{t=1}^{n}\left(-\frac{1}{\delta_{1t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\left(\frac{\partial\delta_{1t}}{\partial\zeta_{0t}}\right)^2\nu_{tr'}\nu_{ts'},
\end{aligned}
$$

$$
\begin{aligned}
\kappa_{\varphi_{r''}\varphi_{s''}} &= \mathrm{E}(U_{\varphi_{r''}\varphi_{s''}}) \\
&= \sum_{t=1}^{n}\left\{\left[\left(-\frac{1}{\delta_{0t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{0t}}{\partial\zeta_{0t}}z_{ts''}+\left(\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{1t}}{\partial\zeta_{1t}}z_{ts''}\right]\frac{\partial\delta_{0t}}{\partial\zeta_{1t}}z_{tr''}\right. \\
&\quad \left.+\left[\left(\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{0t}}{\partial\zeta_{1t}}z_{ts''}+\left(-\frac{1}{\delta_{1t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{1t}}{\partial\zeta_{1t}}z_{ts''}\right]\frac{\partial\delta_{1t}}{\partial\zeta_{1t}}z_{tr''}\right\} \\
&= \sum_{t=1}^{n}\left(-\frac{1}{\delta_{0t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\left(\frac{\partial\delta_{0t}}{\partial\zeta_{1t}}\right)^2 z_{tr''}z_{ts''} \\
&\quad +2\sum_{t=1}^{n}\left(\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{0t}}{\partial\zeta_{1t}}\frac{\partial\delta_{1t}}{\partial\zeta_{1t}}z_{tr''}z_{ts''} \\
&\quad +\sum_{t=1}^{n}\left(-\frac{1}{\delta_{1t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\left(\frac{\partial\delta_{1t}}{\partial\zeta_{1t}}\right)^2 z_{tr''}z_{ts''}
\end{aligned}
$$

and

$$
\begin{aligned}
\kappa_{\rho_{r'}\varphi_{s''}} &= \mathrm{E}(U_{\rho_{r'}\varphi_{s''}}) \\
&= \sum_{t=1}^{n}\left(-\frac{1}{\delta_{0t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{0t}}{\partial\zeta_{0t}}\frac{\partial\delta_{0t}}{\partial\zeta_{0t}}z_{tr''}\nu_{r'}+\sum_{t=1}^{n}\left(\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{1t}}{\partial\zeta_{1t}}\frac{\partial\delta_{0t}}{\partial\zeta_{0t}}z_{tr''}\nu_{r'} \\
&\quad +\sum_{t=1}^{n}\left(\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{0t}}{\partial\zeta_{1t}}\frac{\partial\delta_{1t}}{\partial\zeta_{0t}}z_{r''}\nu_{r'}+\sum_{t=1}^{n}\left(-\frac{1}{\delta_{1t}}+\frac{1}{1-\delta_{0t}-\delta_{1t}}\right)\frac{\partial\delta_{1t}}{\partial\zeta_{1t}}\frac{\partial\delta_{1t}}{\partial\zeta_{0t}}z_{r''}\nu_{r'}.
\end{aligned}
$$

Using Proposition 2, the $\ell_2(\beta, \sigma^2)$ cumulants are

$$\kappa_{\beta_r\beta_s} = \mathrm{E}(U_{\beta_r\beta_s}) = \sum_{t=1}^{n} \mathrm{E}\left(\mathbb{1}_{(0,1)}(y_t)\right)\left\{-\sigma^{-2}a_t\right\} x_{ts}x_{tr} = -\sigma^{-2}\sum_{t=1}^{n}(1-\delta_{0t}-\delta_{1t})a_t x_{tr}x_{ts},$$

where

$$a_t = \left(\frac{3\sigma^2}{\mu_t(1-\mu_t)} + \frac{1}{\mu_t 3(1-\mu_t)^3}\right)\left(\frac{1}{g'(\mu_t)}\right)^2,$$

$$\kappa_{\sigma^2\sigma^2} = \mathrm{E}(U_{\sigma^2\sigma^2}) = \sum_{t=1}^{n} -\mathrm{E}\left(\mathbb{1}_{(0,1)}(y_t)\right)\frac{1}{2\sigma^4} = -\sum_{t=1}^{n}(1-\delta_{0t}-\delta_{1t})\frac{1}{2\sigma^4}$$

and

$$\kappa_{\beta_r\sigma^2} = \mathrm{E}(U_{\beta_r\sigma^2}) = -\frac{1}{\sigma^4}\sum_{t=1}^{n}\mathrm{E}(\mathbb{1}_{(0,1)}(y_t))\mathrm{E}(u_t)\frac{d\mu_t}{d\eta_t}x_{tr} = 0.$$

we note that u_t is given in Equation (A.2) and, using (A.10), $\mathrm{E}(u_t) = 0$.

It follows from the saparability of the parameters $(\rho^\top, \varphi^\top)^\top$ and $(\beta^\top, \sigma^2)^\top$ that

$$\kappa_{\beta_r\rho_{r'}} = \mathrm{E}(U_{\beta_r\rho_{r'}}) = 0,$$

$$\kappa_{\beta_r\varphi_{r''}} = \mathrm{E}(U_{\beta_r\varphi_{r''}}) = 0,$$

$$\kappa_{\sigma^2\rho_{r'}} = \mathrm{E}(U_{\sigma^2\rho_{r'}}) = 0,$$

$$\kappa_{\sigma^2\varphi_{r''}} = \mathrm{E}(U_{\sigma^2\varphi_{r''}}) = 0.$$

A.6 Maximum likelihood estimation of Zero Inflated Simplex regression model (ZIS-RE)

```
# Packages
require(simplexreg)
require(gamlss)

# Adapted distribution function of simplex
psim2 = function (q, mu, sig)
{
   ll <- length(q)
   pp <- rep(0, ll)
   for (i in 1:ll) {
      dsimp <- function(x) {
      1/sqrt(2 * pi * sig[i]^2 * (x * (1 - x))^3) * exp(-1/2/sig[i]^2 *
      (x - mu[i])^2/(x * (1 - x) * mu[i]^2 * (1 - mu[i])^2))
      }
      if (sig[i] < 0.001 | (1 - mu[i]) * sig[i] < 0.01) {
         pp[i] <- psim.norm(q[i], mu[i], sig[i])
      }
```

```
      else {
         tem <- integrate(Vectorize(dsimp), lower = 10^{-100}, upper = q[i])
         pp[i] <- tem$value
      }
   }
   return(pp)
}

# The definition of the d, p, q, and r functions
# pdf
dZIS <- function (x, mu = 0.5, sigma = 1, nu = 0.1, log = FALSE)
{
   if (any(mu <= 0) | any(mu >= 1))
      stop(paste("mu must be beetwen 0 and 1 ", "\n", ""))
   if (any(sigma < 0))
      stop(paste("sigma must be positive", "\n", ""))
   if (any(nu <= 0) | any(nu >= 1))
      stop(paste("nu must be beetwen 0 and 1 ", "\n", ""))
   if (any(x < 0) | any(x >= 1))
      stop(paste("x must be beetwen [0, 1)", "\n", ""))
   log.simplex <- log(dsim(x, mu = mu, sig = sigma))
   log.lik <- ifelse(x == 0, log(nu), log(1 - nu) + log.simplex)
   if (log == FALSE)
      fy <- exp(log.lik)
   else fy <- log.lik
   fy
}

# cdf
pZIS <- function (q, mu = 0.5, sigma = 1, nu = 0.1, log.p=FALSE)
{
   if (any(mu <= 0) | any(mu >= 1))
      stop(paste("mu must be beetwen 0 and 1 ", "\n", ""))
   if (any(sigma < 0))
      stop(paste("sigma must be positive", "\n", ""))
   if (any(nu <= 0) | any(nu >= 1))
      stop(paste("nu must be beetwen 0 and 1 ", "\n", ""))
   cdf <- ifelse((q > 0 & q < 1), (1 - nu) * psim2(q, mu = mu, sig=sigma), 0)
   cdf <- ifelse((q >= 1), 1, cdf)
   if (log.p == FALSE)
      cdf <- cdf
   else cdf <- log(cdf)
   cdf
}

# quantile
qZIS <- function (p, mu = 0.5, sigma = 1, nu = 0.1)
{
   if (any(mu <= 0) | any(mu >= 1))
      stop(paste("mu must be beetwen 0 and 1 ", "\n", ""))
   if (any(sigma < 0))
      stop(paste("sigma must be positive", "\n", ""))
   if (any(nu <= 0) | any(nu >= 1))
      stop(paste("nu must be beetwen 0 and 1 ", "\n", ""))
   if (any(p < 0) | any(p > 1))
      stop(paste("p must be between 0 and 1", "\n", ""))
   suppressWarnings(q <- ifelse((nu >= p), 0, qsim((p - nu)/(1 - nu), mu = mu, sig = sigma)))
   q
}

# random generated
rZIS <- function (n, mu = 0.5, sigma = 1, nu = 0.1)
{
   if (any(mu <= 0) | any(mu >= 1))
      stop(paste("mu must be between 0 and 1", "\n", ""))
   if (any(sigma < 0))
      stop(paste("sigma must be positive", "\n", ""))
   if (any(nu <= 0) | any(nu >= 1))
      stop(paste("nu must be beetwen 0 and 1 ", "\n", ""))
   if (any(n <= 0))
```

```
      stop(paste("n must be a positive integer", "\n", ""))
   n <- ceiling(n)
   p <- runif(n)
   r <- qZIS(p, mu = mu, sigma = sigma, nu = nu)
   r
}

# Distribution
ZIS <- function (mu.link = "logit", sigma.link = "log", nu.link = "logit") {
   # mu = mu
   # sigma = sqrt(sigma^2)
   # nu = alpha

   # Definition of the link function options #
   mstats <- checklink("mu.link", "ZIS", substitute(mu.link), c("logit", "probit", "cloglog", "log",
             "own"))
   dstats <- checklink("sigma.link", "ZIS", substitute(sigma.link), c("inverse", "log", "identity"))
   vstats <- checklink("nu.link", "ZIS", substitute(nu.link), c("logit", "probit", "cloglog", "log",
             "own"))

   # Fitting information #
   structure(
      list(family = c("ZIS", "Zero Inflated Simplex"),
      parameters = list(mu = TRUE, sigma = TRUE, nu = TRUE),
      nopar = 3,
      type = "Mixed",
      mu.link = as.character(substitute(mu.link)),
      sigma.link = as.character(substitute(sigma.link)),
      nu.link = as.character(substitute(nu.link)),
      mu.linkfun = mstats$linkfun,
      sigma.linkfun = dstats$linkfun,
      nu.linkfun = vstats$linkfun,
      mu.linkinv = mstats$linkinv,
      sigma.linkinv = dstats$linkinv,
      nu.linkinv = vstats$linkinv,
      mu.dr = mstats$mu.eta,
      sigma.dr = dstats$mu.eta,
      nu.dr = vstats$mu.eta,
      dldm = function(y, mu, sigma) {
         amu <- (y - mu)*(y - 2*mu*y + mu^2)
         bmu <- sigma^2 * y*(1-y) * (mu^3)*((1-mu)^3)
         dldm <- ifelse((y == 0), 0, amu/bmu)
         dldm
      },
      d2ldm2 = function(y, mu, sigma) {
         cmu <- (3*sigma^2)/(mu*(1-mu))
         dmu <- 1/( (mu^3)*( (1-mu)^3 ) )
         d2ldm2 <- ifelse((y == 0), 0, -(1/sigma^2) * (cmu + dmu))
         d2ldm2
      },
      dldd = function(y, mu, sigma) {
         emu <- (y-mu)^2
         fmu <- y*(1-y)*(mu^2)*((1-mu)^2)
         dldd <- ifelse((y == 0), 0, -(1/(2*sigma^2)) + (1/(2*sigma^4))*(emu/fmu) )
         dldd
      },
      d2ldd2 = function(y, mu, sigma) {
         d2ldd2 <- ifelse((y == 0), 0, -1/(2*(sigma^4)))
         d2ldd2
      },
      dldv = function(y, nu) {
         dldv <- ifelse(y == 0, 1/nu, -1/(1 - nu))
         dldv
      },
      d2ldv2 = function(nu) {
         d2ldv2 <- -1/(nu * (1 - nu))
         d2ldv2
      },
      d2ldmdd = function(y, mu, sigma) {
         d2ldmdd <- rep(0, length = y)
```

```
        d2ldmdd
      },
      d2ldmdv = function(y) {
        d2ldmdv <- rep(0, length = y)
        d2ldmdv
      },
      d2ldddv = function(y) {
        d2ldddv <- rep(0, length = y)
        d2ldddv
      },
      G.dev.incr = function(y, mu, sigma, nu, ...) {
        -2 * dZIS(y, mu, sigma, nu, log = TRUE)
      },
      rqres = expression({
      uval <- ifelse(y == 0, nu * runif(length(y), 0, 1), (1 - nu) * pZIS(y, mu, sigma, nu))
      rqres <- qnorm(uval)
      }),
      mu.initial = expression(mu <- (y + mean(y))/2),
      sigma.initial = expression(sigma <- rep(1, length(y))),
      nu.initial = expression(nu <- rep(length(y[y==0])/length(y), length(y))),
      mu.valid = function(mu) all(mu > 0 & mu < 1),
      sigma.valid = function(sigma) all(sigma > 0),
      nu.valid = function(nu) all(nu > 0 & nu < 1),
      y.valid = function(y) all(y >= 0 & y < 1)),
      class = c("gamlss.family", "family")
   )
}

## Simulating ZIS-RE model ##

set.seed(6581) # seed

n=500 # sample size

# Generating the linear preditor of mu
x1=runif(n,min=0,max=1) # covariate of mu
eta.mu=-1.5+1.5*x1 # linear predictor of mu
mu=exp(eta.mu)/(1+exp(eta.mu)) # inverse of the link function of mu

# Generating the linear preditor of nu
z1=runif(n,min=0,max=1) # covariate of nu
eta.nu=-1+.5*x3 # linear predictor of nu
nu=exp(eta.nu)/(1+exp(eta.nu)) # inverse of the link function of nu

# precision parameter (sigma)
sigma = 1 # inverse of the link function of sigma

# Generating the response variable
y = mapply(rZIS, 1, mu, sigma, nu)

# Proportion of zeros
zeros = length(which(y==0))
zeros/n # proportion of zeros

# Adjusting ZIS-RE model
fit = gamlss(y~x1, sigma.formula=~1,nu.formula=~z1, family=ZIS(mu.link = "logit", sigma.link = "identity",
      nu.link = "logit"))

summary(fit)

******************************************************************
Family:  c("ZIS", "Zero Inflated Simplex")

Call:  gamlss(formula = y ~ x1, sigma.formula = ~1, nu.formula = ~z1,
family = ZIS(mu.link = "logit", sigma.link = "identity",
nu.link = "logit"))

Fitting method: RS()

------------------------------------------------------------------
```

```
Mu link function:  logit
Mu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.51353    0.04074  -37.15   <2e-16 ***
x1           1.49886    0.07480   20.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Sigma link function:  identity
Sigma Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.98399    0.03791   25.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Nu link function:  logit
Nu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.0066     0.1959  -5.137 4.02e-07 ***
z1            0.5522     0.3317   1.665   0.0966 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
No. of observations in the fit:  500
Degrees of Freedom for the fit:  5
Residual Deg. of Freedom:  495
at cycle:  2

Global Deviance:     -75.65782
AIC:     -65.65782
SBC:     -44.58478
******************************************************************
```

A.7 Maximum likelihood estimation of Zero and One Inflated Simplex regression model (ZOIS-RE)

```
# Maximum likelihood estimation of Zero Inflated Simplex regression model (ZIS-RE)

# Packages
require(simplexreg)
require(gamlss)

# Adapted distribution function of simplex
psim2 = function (q, mu, sig)
{
   ll <- length(q)
   pp <- rep(0, ll)
   for (i in 1:ll) {
      dsimp <- function(x) {
         1/sqrt(2 * pi * sig[i]^2 * (x * (1 - x))^3) * exp(-1/2/sig[i]^2 *
         (x - mu[i])^2/(x * (1 - x) * mu[i]^2 * (1 - mu[i])^2))
      }
   if (sig[i] < 0.001 | (1 - mu[i]) * sig[i] < 0.01) {
      pp[i] <- psim.norm(q[i], mu[i], sig[i])
      }
   else {
      tem <- integrate(Vectorize(dsimp), lower = 1e-100, upper = q[i])
      pp[i] <- tem$value
      }
   }
   return(pp)
}
```

```
# The definition of the d, p, q, and r functions
# pdf
dZIS <- function (x, mu = 0.5, sigma = 1, nu = 0.1, log = FALSE)
{
   if (any(mu <= 0) | any(mu >= 1))
      stop(paste("mu must be beetwen 0 and 1 ", "\n", ""))
   if (any(sigma < 0))
      stop(paste("sigma must be positive", "\n", ""))
   if (any(nu <= 0) | any(nu >= 1))
      stop(paste("nu must be beetwen 0 and 1 ", "\n", ""))
   if (any(x < 0) | any(x >= 1))
      stop(paste("x must be beetwen [0, 1)", "\n", ""))
   log.simplex <- log(dsim(x, mu = mu, sig = sigma))
   log.lik <- ifelse(x == 0, log(nu), log(1 - nu) + log.simplex)
   if (log == FALSE)
      fy <- exp(log.lik)
   else fy <- log.lik
   fy
}

# cdf
pZIS <- function (q, mu = 0.5, sigma = 1, nu = 0.1, log.p=FALSE)
{
   if (any(mu <= 0) | any(mu >= 1))
      stop(paste("mu must be beetwen 0 and 1 ", "\n", ""))
   if (any(sigma < 0))
      stop(paste("sigma must be positive", "\n", ""))
   if (any(nu <= 0) | any(nu >= 1))
      stop(paste("nu must be beetwen 0 and 1 ", "\n", ""))
   cdf <- ifelse((q > 0 & q < 1), nu + (1 - nu) * psim2(q, mu = mu, sig=sigma), 0)
   cdf <- ifelse((q == 0), nu, cdf)
   cdf <- ifelse((q >= 1), 1, cdf)
   if (log.p == FALSE)
      cdf <- cdf
   else cdf <- log(cdf)
      cdf
}

# quantile
qZIS <- function (p, mu = 0.5, sigma = 1, nu = 0.1)
{
   if (any(mu <= 0) | any(mu >= 1))
      stop(paste("mu must be beetwen 0 and 1 ", "\n", ""))
   if (any(sigma < 0))
      stop(paste("sigma must be positive", "\n", ""))
   if (any(nu <= 0) | any(nu >= 1))
      stop(paste("nu must be beetwen 0 and 1 ", "\n", ""))
   if (any(p < 0) | any(p > 1))
      stop(paste("p must be between 0 and 1", "\n", ""))
   suppressWarnings(q <- ifelse((nu >= p), 0, qsim((p - nu)/(1 - nu), mu = mu, sig = sigma)))
   q
}

# random generated
rZIS <- function (n, mu = 0.5, sigma = 1, nu = 0.1)
{
   if (any(mu <= 0) | any(mu >= 1))
      stop(paste("mu must be between 0 and 1", "\n", ""))
   if (any(sigma < 0))
      stop(paste("sigma must be positive", "\n", ""))
   if (any(nu <= 0) | any(nu >= 1))
      stop(paste("nu must be beetwen 0 and 1 ", "\n", ""))
   if (any(n <= 0))
      stop(paste("n must be a positive integer", "\n", ""))
   n <- ceiling(n)
   p <- runif(n)
   r <- qZIS(p, mu = mu, sigma = sigma, nu = nu)
   r
}
```

```
# Distribution
ZIS <- function (mu.link = "logit", sigma.link = "log", nu.link = "logit") {
   # mu = mu
   # sigma = sqrt(sigma^2)
   # nu = alpha

   # Definition of the link function options #
   mstats <- checklink("mu.link", "ZIS", substitute(mu.link), c("logit", "probit", "cloglog", "log", "own"))
   dstats <- checklink("sigma.link", "ZIS", substitute(sigma.link), c("inverse", "log", "identity"))
   vstats <- checklink("nu.link", "ZIS", substitute(nu.link), c("logit", "probit", "cloglog", "log", "own"))

   # Fitting information #
   structure(
      list(family = c("ZIS", "Zero Inflated Simplex"),
      parameters = list(mu = TRUE, sigma = TRUE, nu = TRUE),
      nopar = 3,
      type = "Mixed",
      mu.link = as.character(substitute(mu.link)),
      sigma.link = as.character(substitute(sigma.link)),
      nu.link = as.character(substitute(nu.link)),
      mu.linkfun = mstats$linkfun,
      sigma.linkfun = dstats$linkfun,
      nu.linkfun = vstats$linkfun,
      mu.linkinv = mstats$linkinv,
      sigma.linkinv = dstats$linkinv,
      nu.linkinv = vstats$linkinv,
      mu.dr = mstats$mu.eta,
      sigma.dr = dstats$mu.eta,
      nu.dr = vstats$mu.eta,
      dldm = function(y, mu, sigma) {
         amu <- (y - mu)*(y - 2*mu*y + mu^2)
         bmu <- sigma^2 * y*(1-y) * (mu^3)*((1-mu)^3)
         dldm <- ifelse((y == 0), 0, amu/bmu)
         dldm
      },
      d2ldm2 = function(y, mu, sigma) {
         cmu <- (3*sigma^2)/(mu*(1-mu))
         dmu <- 1/( (mu^3)*( (1-mu)^3 ) )
         d2ldm2 <- ifelse((y == 0), 0, -(1/sigma^2) * (cmu + dmu))
         d2ldm2
      },
      dldd = function(y, mu, sigma) {
         emu <- (y-mu)^2
         fmu <- y*(1-y)*(mu^2)*((1-mu)^2)
         dldd <- ifelse((y == 0), 0, -(1/(2*sigma^2)) + (1/(2*sigma^4))*(emu/fmu) )
         dldd
      },
      d2ldd2 = function(y, mu, sigma) {
         d2ldd2 <- ifelse((y == 0), 0, -1/(2*(sigma^4)))
         d2ldd2
      },
      dldv = function(y, nu) {
         dldv <- ifelse(y == 0, 1/nu, -1/(1 - nu))
         dldv
      },
      d2ldv2 = function(nu) {
         d2ldv2 <- -1/(nu * (1 - nu))
         d2ldv2
      },
      d2ldmdd = function(y, mu, sigma) {
         d2ldmdd <- rep(0, length(y))
         d2ldmdd
      },
      d2ldmdv = function(y) {
         d2ldmdv <- rep(0, length(y))
         d2ldmdv
      },
      d2ldddv = function(y) {
         d2ldddv <- rep(0, length(y))
```

```
        d2ldddv
      },
      G.dev.incr = function(y, mu, sigma, nu, ...) {
        -2 * dZIS(y, mu, sigma, nu, log = TRUE)
      },
      rqres = expression({
        uval <- ifelse(y == 0, nu * runif(length(y), 0, 1), (1 - nu) * pZIS(y, mu, sigma, nu))
        rqres <- qnorm(uval)
      }),
      mu.initial = expression(mu <- (y + mean(y))/2),
        sigma.initial = expression(sigma <- rep(.5, length(y))),
      nu.initial = expression(nu <- rep(length(y[y==0])/length(y), length(y))),
      mu.valid = function(mu) all(mu > 0 & mu < 1),
      sigma.valid = function(sigma) all(sigma > 0),
      nu.valid = function(nu) all(nu > 0 & nu < 1),
      y.valid = function(y) all(y >= 0 & y < 1)),
      class = c("gamlss.family", "family")
   )
}

## Simulating ZIS-RE model ##

set.seed(6581) # seed

n=500 # sample size

# Generating the linear preditor of mu
x1=runif(n,min=0,max=1) # covariate of mu
eta.mu=-1.5+1.5*x1 # linear predictor of mu
mu=exp(eta.mu)/(1+exp(eta.mu)) # inverse of the link function of mu

# Generating the linear preditor of nu
z1=runif(n,min=0,max=1) # covariate of nu
eta.nu=-1+.5*z1 # linear predictor of nu
nu=exp(eta.nu)/(1+exp(eta.nu)) # inverse of the link function of nu

# precision parameter (sigma)
sigma = 1 # inverse of the link function of sigma

# Generating the response variable
y = mapply(rZIS, 1, mu, sigma, nu)

# Proportion of zeros
zeros = length(which(y==0))
zeros/n # proportion of zeros

# Adjusting ZIS-RE model
fit = gamlss(y~x1, sigma.formula=~1,nu.formula=~z1, family=ZIS(mu.link = "logit", sigma.link = "identity", nu.link = "log

summary(fit)

******************************************************************
Family:  c("ZIS", "Zero Inflated Simplex")

Call:  gamlss(formula = y ~ x1, sigma.formula = ~1, nu.formula = ~z1,
family = ZIS(mu.link = "logit", sigma.link = "identity",
nu.link = "logit"))

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.51353    0.04074  -37.15   <2e-16 ***
x1           1.49886    0.07480   20.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
```

```
Sigma link function:  identity
Sigma Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.98397    0.03791   25.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Nu link function:  logit
Nu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.0066     0.1959  -5.137 4.02e-07 ***
z1            0.5522     0.3317   1.665   0.0966 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
No. of observations in the fit:  500
Degrees of Freedom for the fit:  5
Residual Deg. of Freedom:  495
at cycle:  2

Global Deviance:     -75.65783
AIC:     -65.65783
SBC:     -44.58479
******************************************************************
```

APPENDIX B

Appdences of Chapter 3

B.1 Residuals for the zero or one inflated simplex regression

Using the general definition of residuals given by Cox & Snell (1968) the standard residuals can be defined for the zero or one inflated simplex regression model. If σ^2 is known, Fisher' scoring algorithm is given by

$$\begin{aligned}\varphi^{(m+1)} &= \varphi^{(m)} + (Z^\top Q^{(m)} Z)^{-1} Z^\top P^{(m)} G^{(m)} (y^c - \alpha^{*(m)}) \\ &= (Z^\top Q^{(m)} Z)^{-1} Z^\top Q^{(m)} z^{(m)},\end{aligned} \tag{B.1}$$

$m = 0, 1, 2, \ldots$, where, for $t = 1, \ldots, n$,

$$\begin{aligned} z^{(m)} &= Z\varphi^{(m)} + (Q^{(m)})^{-1} P^{(m)} G^{(m)} (y^c - \alpha^{*(m)}) \\ Q &= \text{diag}\{q_1, \ldots, q_n\},\ q_t = -p_t[1/h'(\alpha_t)]^2, \\ P &= \text{diag}\{1/[\alpha_1(1-\alpha_1)], \ldots, 1/[\alpha_n(1-\alpha_n)]\} \\ G &= \text{diag}\{1/h'(\alpha_1), \ldots, 1/h'(\alpha_n)\} \\ y^c &= (\mathbb{1}_{\{c\}}(y_1), \cdots, \mathbb{1}_{\{c\}}(y_1))^\top. \end{aligned}$$

Fisher' scoring method to obtain the MLE of β can be expressed as

$$\begin{aligned}\beta^{(m+1)} &= \beta^{(m)} + (X^\top \Delta^{(m)} A^{(m)} X)^{-1} X^\top T^{(m)} H^{(m)} u^{(m)} \\ &= (X^\top \Delta^{(m)} A^{(m)} X)^{-1} X^\top \Delta^{(m)} A^{(m)} z_1^{(m)},\end{aligned} \tag{B.2}$$

where

$$\begin{aligned} z_1^{(m)} &= X\beta^{(m)} + (\Delta^{(m)} A^{(m)})^{-1} T^{(m)} H^{(m)} u^{(m)}, \\ \Delta &= \text{diag}\{1-\alpha_1, \ldots, 1-\alpha_n\}, \\ A &= \text{diag}\{a_1, \ldots, a_n\}, \\ a_t &= \left(\frac{3\sigma^2}{\mu_t(1-\mu_t)} + \frac{1}{\mu_t^3(1-\mu_t)^3}\right)\left(\frac{1}{g'(\mu_t)}\right)^2, \\ T &= \text{diag}\{1/g'(\mu_1), \ldots, 1/g'(\mu_n)\}, \\ H &= \text{diag}\{1 - \mathbb{1}_{\{c\}}(y_1), \cdots, 1 - \mathbb{1}_{\{c\}}(y_n)\}, \\ u^\top &= (u_1, \ldots, u_n)^\top, \\ u_t &= \frac{(y_t - \mu_t)(y_t - 2\mu_t y_t + \mu_t^2)}{y_t(1-y_t)\mu_t^3(1-\mu)^3}. \end{aligned}$$

When σ^2 is known, upon convergence

$$\begin{aligned}\widehat{\varphi} &= (Z^\top \widehat{Q} Z)^{-1} Z^\top \widehat{Q} \widehat{\tau}_1,\\ \widehat{\beta} &= (X^\top \widehat{\Delta}\widehat{A} X)^{-1} X^\top \widehat{\Delta}\widehat{A} \widehat{\tau}_2,\end{aligned}$$

where $\widehat{\tau}_1 = \widehat{\zeta} + G^{-1}(y^c - \widehat{\alpha})$ and $\widehat{\tau}_2 = \widehat{\eta} + (\widehat{\Delta}\widehat{A})^{-1}\widehat{T}\widehat{u}$, where $\widehat{\zeta} = (\widehat{\zeta}_1, \ldots, \widehat{\zeta}_n) = Z\widehat{\varphi}$ and $\widehat{\eta} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_n) = X\widehat{\beta}$. Then, let $Z^* = \widehat{Q}^{1/2} Z$, $X^* = (\widehat{\Delta}\widehat{A})^{1/2} X$, $\tau_1^* = \widehat{Q}^{1/2}\widehat{\tau}_1$ and $\tau_2^* = (\widehat{\Delta}\widehat{A})^{1/2}\widehat{\tau}_2$. Then, let

$$\begin{aligned}\widehat{\varphi} &= (Z^{*\top} Z^*)^{-1} Z^{*\top} \tau_1^*,\\ \widehat{\beta} &= (X^{*\top} X^*)^{-1} X^{*\top} \tau_2^*.\end{aligned} \tag{B.3}$$

Thus, $\widehat{\varphi}$ is the least square solution of a linear regression of $\tau*_1$ on the columns of Z^*. Similarly, $\widehat{\beta}$ is the least square solution of a linear regression of τ_2^* on the columns of X^*. From (B.3) we have

$$\begin{aligned}\widehat{\tau}_1^* &= Z^*\widehat{\varphi} = \widehat{Q}^{1/2} Z (Z^\top \widehat{Q} Z)^{-1} Z^\top \widehat{Q}^{1/2} \widehat{Q}^{1/2}\widehat{\tau}_1\\ &= \widehat{H}_1^* \tau_1^*,\\ \widehat{\tau}_2^* &= X^*\widehat{\beta} = (\widehat{\Delta}\widehat{A})^{1/2} X (X^\top \widehat{\Delta}\widehat{A} X)^{-1} X^\top (\widehat{\Delta}\widehat{A})^{1/2} (\widehat{\Delta}\widehat{A})^{1/2}\widehat{\tau}_2\\ &= \widehat{H}_2^* \tau_2^*,\end{aligned}$$

where $\widehat{H}_1^* = \widehat{Q}^{1/2} Z (Z^\top \widehat{Q} Z)^{-1} Z^\top \widehat{Q}^{1/2}$ and $\widehat{H}_2^* = (\widehat{\Delta}\widehat{A})^{1/2} X (X^\top \widehat{\Delta}\widehat{A} X)^{-1} X^\top (\widehat{\Delta}\widehat{A})^{1/2}$ are projection matrices. From (B.3), the ordinary residuals can be expressed as

$$\mathrm{e}_1^* = \tau_1^* - \widehat{\tau}_1^* = (I_n - \widehat{H}_1^*)\tau_1^*, \tag{B.4}$$

$$\mathrm{e}_2^* = \tau_2^* - \widehat{\tau}_2^* = (I_n - \widehat{H}_2^*)\tau_2^*, \tag{B.5}$$

where I_n is the identity matrix of dimension $n \times n$.

The asymptotic covariance matrices of residuals e_1^* and e_2^*, evaluated at the true parameters are

$$\begin{aligned}\mathrm{Var}(e_1^*) &= \mathrm{Var}((I_n - H_1^*)\tau*_1)\\ &= (I_n - H_1^*)\mathrm{Var}(\widehat{Q}^{1/2}\tau_1)(I_n - H_1^*)\\ &= (I_n - H_1^*)\widehat{Q}^{1/2}\mathrm{Var}(\tau_1)\widehat{Q}^{1/2}(I_n - H_1^*)\\ &= (I_n - H_1^*)\widehat{Q}^{1/2}\widehat{Q}^{-1}\widehat{Q}^{1/2}(I_n - H_1^*)\\ &= (I_n - H_1^*)\end{aligned} \tag{B.6}$$

and

$$\begin{aligned}\mathrm{Var}(e_2^*) &= \mathrm{Var}((I_n - H_2^*)\tau*_2)\\ &= (I_n - H_2^*)\mathrm{Var}((\widehat{\Delta}\widehat{A})^{1/2}\tau_2)(I_n - H_2^*)\\ &= (I_n - H_2^*)(\widehat{\Delta}\widehat{A})^{1/2}\mathrm{Var}(\tau_2)(\widehat{\Delta}\widehat{A})^{1/2}(I_n - H_2^*)\\ &= (I_n - H_2^*)(\widehat{\Delta}\widehat{A})^{1/2}(\widehat{\Delta}\widehat{A})^{-1}(\widehat{\Delta}\widehat{A})^{1/2}(I_n - H_2^*)\\ &= (I_n - H_2^*).\end{aligned} \tag{B.7}$$

Furthermore, note that if the quantities are evaluated at the true values, $E(e_1^*) = 0$ and $E(e_2^*) = 0$.

B.2 Residuals for the zero and one inflated simplex regression

The following results are similar to the obtained by Ospina & Ferrari (2012). Consider the zero or one inflated simplex regression model with σ^2 constant. Upon convergence of the iterative process of Fisher' scoring algorithm for $\Upsilon = (\rho^\top, \varphi^\top)^\top$ we have

$$\widehat{\Upsilon} = (\widetilde{Z}^\top \widehat{Q} \widetilde{Z})^{-1} Z^\top \widehat{Q}\widehat{\tau}_d,$$

where $\widehat{\tau}_d = \widetilde{Z}\widehat{\Upsilon} + \widehat{Q}^{-1}(y_d - \widehat{\delta}_d)$, where $y_d = (y_{\{0\}}^\top, y_{\{1\}}^\top)^\top$, $\delta_d = (\delta_0^\top, \delta_1^\top)^\top$ are vectors with dimension $2n \times 1$. Here, $\delta_0 = (\delta_{01}, \ldots, \delta_{0n})^\top$, $\delta_0 = (\delta_{11}, \ldots, \delta_{1n})^\top$, $y_{\{0\}} = (\mathbb{1}_{\{0\}}(y_1), \ldots, \mathbb{1}_{\{0\}}(y_n))^\top$ and $y_{\{1\}} = (\mathbb{1}_{\{1\}}(y_1), \ldots, \mathbb{1}_{\{1\}}(y_n))^\top$ are

vectors of dimension $n \times 1$. Also,

$$\widetilde{Z} = \begin{pmatrix} V & 0 \\ 0 & Z \end{pmatrix}, \qquad Q = \begin{pmatrix} Q_1 & Q_3 \\ Q_3 & Q_2 \end{pmatrix},$$

are, respectively, matrices of dimension $(2n \times (k_0 + k_1))$ and $2n \times 2n$. V and Z are $n \times k_0$ and $n \times k_1$ matrices of regressors whose tth lines are ν_t and z_t, respectively. The elements of Q are $Q_1 = V^\top \text{diag}\{\delta_{01}(1-\delta_{01}), \ldots, \delta_{0n}(1-\delta_{0n})\}V$, $Q_2 = Z^\top \text{diag}\{\delta_{11}(1-\delta_{11}), \ldots, \delta_{1n}(1-\delta_{1n})\}Z$ and $Q_3 = Z^\top \text{diag}\{-\delta_{01}\delta 11 \ldots, -\delta_{0n}\delta_{1n}\}V$.

Let $\tau_d^* = Q^{1/2}\widehat{\tau}_d$ and $Z^* = Q^{1/2}\widetilde{Z}$. Then,

$$\widehat{\Upsilon} = (Z^{*\top} Z^*)^{-1} Z^{*\top} \tau_d^*. \tag{B.8}$$

Therefore, $\widehat{\Upsilon}$ is the least square solution of τ_d^* on Z^*. From (B.8) we have

$$\begin{aligned} \widehat{\tau}_d^* &= Z^* \widehat{\Upsilon} = \widehat{Q}^{1/2}\widetilde{Z}(\widetilde{Z}^\top \widetilde{Q} \widetilde{Z})^{-1} \widetilde{Z}^\top \widehat{Q}^{1/2} \widehat{Q}^{1/2} \widehat{\tau}_d \\ &= \widehat{H}_d^* \tau_d^*. \end{aligned}$$

The ordinary residual of the regression (B.8) can be expressed as

$$\text{e}_d^* = \tau_d^* - \widehat{\tau}_d^* = \tau_d^* - \widehat{H}_d^* \tau_d^* = (I_{2n} - \widehat{H}_d^*)\tau_d^*, \tag{B.9}$$

where I_{2n} is the $2n \times 2n$ identity matrix and $\widehat{H}_d^* = \widehat{Q}^{1/2}\widetilde{Z}(\widetilde{Z}^\top \widehat{Q}\widetilde{Z})^{-1}\widetilde{Z}^\top \widehat{Q}^{1/2}$ is the projection matrix. Note that e_d^* can also be written as

$$\begin{aligned} \text{e}_d^* = (I_{2n} - \widehat{H}_d^*)\widehat{\tau}_d &= \widehat{Q}^{1/2}\widehat{\tau}_d - \widehat{Q}^{1/2}\widetilde{Z}(\widetilde{Z}^\top \widehat{Q}\widetilde{Z})^{-1}\widetilde{Z}^\top \widehat{Q}^{1/2}\widehat{Q}^{1/2}\widehat{\tau}_d, \\ &= \widehat{Q}^{1/2}\widehat{Q}^{-1}(y_d - \widehat{\Delta}_d) \\ &= \widehat{Q}^{1/2}(y_d - \widehat{\Delta}_d). \end{aligned} \tag{B.10}$$

Note that if the quantities are evaluated in the true values of the parameters $E(e_d^*) = 0$ and

$$\text{Var}(e_d^*) = \widehat{Q}^{-1/2}\text{Var}(y_d)\widehat{Q}^{-1/2}.$$

On the other hand,

$$\begin{aligned} \text{Var}(e_d^*) &= \text{Var}((I_{2n} - \widehat{H}_d^*)\tau *_d) \\ &= (I_{2n} - \widehat{H}_d^*)\text{Var}(\tau_d^*)(I_{2n} - \widehat{H}_d^*) \\ &= (I_{2n} - \widehat{H}_d^*)\widehat{Q}^{1/2}\text{Var}(\widehat{\tau}_d)\widehat{Q}^{1/2}(I_{2n} - \widehat{H}_d^*). \end{aligned}$$

Asymptotically, $\text{Var}(\widehat{\Upsilon}) = (\widetilde{Z}^\top \widehat{Q}\widetilde{Z})^{-1}$, then $\text{Var}(\widehat{\tau}_d) = \widehat{Q}^{-1}$ and

$$\begin{aligned} \text{Var}(e_d^*) &= (I_{2n} - \widehat{H}_d^*)\widehat{Q}^{1/2}\widehat{Q}^{-1}(\widehat{\tau}_d)\widehat{Q}^{1/2}(I_{2n} - \widehat{H}_d^*) \\ &= (I_{2n} - \widehat{H}_d^*). \end{aligned}$$

Futhermore,

$$\text{Var}(y_d) = \widehat{Q}^{-1/2}(I_{2n} - \widehat{H}_d^*)\widehat{Q}^{-1/2}.$$

Then, standardized residual for the discrete component of the zero and one inflated simplex regression model (i.e., for $y = 0$ and $y = 1$) can be defined as

$$r_i^d = \frac{y_{d_i} - \widehat{\Delta}_{d_i}}{\sqrt{\widehat{q}_{ii}(1 - \widehat{h}_{d_{ii}})}}, \; i = 1, 2, \ldots, 2n,$$

where $y_d = (y_{\{0\}}^\top, y_{\{1\}}^\top)^\top$,$\Delta_d = (\delta\top_0, \delta\top_1)^\top$ are $2n \times 1$ vectors. Here, $\delta_0 = (\delta_{01}, \ldots, \delta_{0n})^\top$, $\delta_1 = (\delta_{11}, \ldots, \delta_{1n})^\top$, $y_{\{0\}} = (\mathbb{1}_{\{0\}}(y_1), \ldots, \mathbb{1}_{\{0\}}(y_n))^\top$ and $y_{\{1\}} = (\mathbb{1}_{\{1\}}(y_1), \ldots, \mathbb{1}_{\{1\}}(y_n))^\top$ are $n \times 1$ vectors. Furthermore, $\widehat{q}_{ii}$ is the ith diagonal element of matrix $\widehat{Q}$, and $\widehat{h}_{d_{ii}}$ is the ith diagonal element of the projection matrix.

For $i = 1, \ldots, n$, residual r_i^d is the standard residual of the submodel of that models the probability of occurence of zeros and for $i = n+1, \ldots, 2n$, r_i^d is the standard residual of the submodel that models the probability of occurence of ones. For this reason, the graph containing r_i^d against $\widehat{\Delta}_{d_i}$ should be separated: the first for $i = 1, \ldots, n$, and the second for $i = n+1, \ldots, 2n$. Both plots can reveal outliers in each submodel.

We can define the projection matrices as

$$H^{\{0\}} = \Psi_0^\top H_d^* \Psi_0,$$
$$H^{\{1\}} = \Psi_1^\top H_d^* \Psi_1,$$

where $\Psi_0 = (I_n, 0_n)^\top$ and $\Psi_1 = (0_n, I_n)^\top$, I_n is the $n \times n$ identity matrix and 0_n is a $n \times n$ matrix containing zeros. $H^{\{0\}}$ and $H^{\{1\}}$ are the projection matrices in the submodel that models zeros and that models ones, respectively.