Contribuições em modelos de regressão com erro de medida multiplicativo

Eveliny Barroso da Silva

Contribuições em modelos de regressão com erro de medida multiplicativo

Eveliny Barroso da Silva

Orientador: ***Prof. Dr. Carlos Alberto Ribeiro Diniz***

Tese apresentada ao Departamento de Estatística da Universidade Federal de São Carlos e ao Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Doutora em Estatística no Programa Interinstitucional de Pós-graduação em Estatística - UFSCar/USP.

UFSCar - USP - São Carlos

Março de 2016

Contributions in regression models with multiplicative measurement error

Eveliny Barroso da Silva

Supervisor: ***Prof. Dr. Carlos Alberto Ribeiro Diniz***

Thesis to be submitted to the Departament of Statistics at the Federal University of São Carlos and the Institute of Mathematics and Computer Sciences, University of São Paulo, as part of the requirements for obtaining PhD degree in Statistics at Inter Program Graduate in Statistics - UFSCar / USP.

UFSCar - USP - São Carlos

March 2016

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar
Processamento Técnico
com os dados fornecidos pelo(a) autor(a)

S586c

Silva, Eveliny Barroso da
Contribuições em modelos de regressão com erro de medida multiplicativo / Eveliny Barroso da Silva. -- São Carlos : UFSCar, 2016.
98 p.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2016.

1. Erro de medida nas covariáveis. 2. Erro de medida multiplicativo. 3. Pseudo verossimilhança. 4. Análise de diagnóstico. 5. Quadratura de Gauss-Hermite. I. Título.

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado da candidata Eveliny Barroso da Silva, realizada em 04/02/2016:

Prof. Dr. Carlos Alberto Ribeiro Diniz
UFSCar

Prof. Dr. Filidor Edilfonso Vilca Labra
UNICAMP

Profa. Dra. Giovana Oliveira Silva
UFBA

Profa. Dra. Hildete Prisco Pinheiro
UNICAMP

Prof. Dr. Mário de Castro Andrade Filho
USP

Agradecimentos

Agradeço, antes de tudo, a Deus.

Agradeço também:

A meus pais e demais familiares.

Ao Thiago, pelo amor, amizade e total apoio.

A todos os amigos que sempre estiveram presentes, contribuindo com discussões, críticas e sugestões.

Ao professor Carlos Alberto Ribeiro Diniz, pela orientação segura, e pelo incentivo durante todo o curso de pós-graduação.

Aos professores Alexandre Patriota e Mário de Castro, membros da banca examinadora do exame de qualificação, pelas sugestões apresentadas.

Aos professores Filidor Edilfonso, Giovana Oliveira, Hildete Prisco e Mário de Castro, membros da banca examinadora, pelas valiosas sugestões e comentários sobre esta tese e pesquisas futuras.

Ao Departamento de Estatística da Universidade Federal de São Carlos pelo ambiente acolhedor.

A todos os colegas do Departamento de Estatística da Universidade Federal de Mato Grosso pelo apoio e incentivo durante todo o curso de doutorado.

Resumo

Em modelos de regressão em que uma covariável é medida com erro, é comum o uso de estruturas que relacionam a covariável observada com a verdadeira covariável não observada. Essas estruturas são usualmente aditivas ou multiplicativas. Na literatura existem diversos trabalhos interessantes que tratam de modelos de regressão com erro de medida aditivo, muitos dos quais são modelos lineares com covariáveis e erro de medida normalmente distribuídos. Para modelos em que o erro de medida é multiplicativo, não se encontra na literatura o mesmo desenvolvimento teórico encontrado para modelos em que o erro de medida é aditivo. O mesmo vale para situações em que as suposições de normalidade para as covariáveis e erro de medida não se aplicam. Este trabalho propõe a construção, definição, métodos de estimação e análise de diagnóstico para modelos de regressão com erro de medida multiplicativo em uma das covariáveis. Para esses modelos, consideramos que a variável resposta possa pertencer ou à classe de modelos de regressão série de potências modificadas ou à família exponencial. O rol de distribuições pertencentes à família série de potências modificada é bem abrangente, portanto, neste trabalho, desenvolvemos a teoria de estimação e validação do modelo primeiramente de forma geral e, para exemplificar, apresentamos o modelo de regressão binomial negativa com erro de medida. Para o caso em que a variável resposta pertença à família exponencial, apresentamos o modelo de regressão beta com erro de medida multiplicativo. Todos os modelos propostos foram analisados através de estudos de

simulação e aplicados a conjuntos de dados reais.

Palavras-chave: erro de medida nas covariáveis, erro de medida multiplicativo, pseudo verossimilhança, análise de diagnóstico, quadratura de Gauss-Hermite, regressão série de potências modificada.

Abstract

In regression models in which a covariate is measured with error, it is common to use structures that correlate the observed covariate with the true non-observed covariate. Such structures are usually additive or multiplicative. In the literature there are several interesting works that deal with regression models having an additive measurement error, many of which are linear models with covariate and measurement error normally distributed. For models having a multiplicative measurement error, one does not find in the literature the same theoretical amount of works as one finds for models in which the measurement error is additive. The same happens in situations where the supositions of normality for the covariates and the measurement errors do not apply. The present work proposes the construction, definition, estimation methods, and diagnostic analysis for the regression models with a multiplicative measurement error in one of the covariates. For these models it is considered that the response variable may belong either to the class of modified power series regression models or to the exponential family. The list of distributions belonging to the family modified power series is rather comprehensive; for this reason this work develops, firstly and in a general way, the models estimation and validation theory, and, as an example, presents the model of negative binomial regression with measurement error. In the case where the response variable belongs to the exponential family, the model of beta regression with multiplicative measurement error is presented. All proposed models were analysed through simulation studies

and applied to real data sets.

key-words: regression models with measurement error, multiplicative measurement error, pseudo-likelihood, diagnostic analysis, gauss-hermite quadrature, modified power series regression models.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo

1

Introdução

Modelos de regressão são ferramentas estatísticas que podem (entre outras finalidades) relacionar o valor médio da variável de interesse (variável resposta) a uma ou mais covariáveis (variáveis explicativas). Em situações práticas, pode acontecer que algumas covariáveis associadas à variável resposta sejam medidas com erro. Suponha que exista apenas uma covariável medida com erro, X_i, e que W_i seja a covariável observada no modelo. Nesta situação, supomos que existe uma estrutura que relaciona a covariável observada W_i com a verdadeira covariável não observada X_i. Essa estrutura é, usualmente, aditiva, $W_i = X_i + \varepsilon_i$, ou multiplicativa, $W_i = X_i\varepsilon_i$, $i = 1, \ldots, n$. Existem três abordagens para lidar com os erros de medida nas covariáveis em modelos de regressão: modelo funcional, estrutural e ultraestrutural. O *modelo funcional*, trata as covariáveis não observadas como parâmetros incidentais, ou seja, cada $X_i, i = 1, \ldots, n$, é uma constante desconhecida. O *modelo estrutural*, trata as covariáveis não observadas como variáveis aleatórias independentes, identicamente distribuídas e independentes dos erros de medição. O *modelo ultraestrutural* é uma generalização dos dois modelos anteriores. Essa generalização assume que as covariáveis não observadas são variáveis aleatórias não identicamente distribuídas, diferente do modelo estrutural no qual as covariáveis não observadas têm a mesma média para todas as observações (Dolby (1976) e Cheng e Van Ness (1999)). Outras propostas apresentadas por Carroll *et al.* (2006) são as

modelagens *funcional* e *estrutural*. Segundo os autores, a modelagem funcional assume que a covariável não observada X pode ser fixa ou aleatória. Caso seja aleatória, não é necessário conhecer a distribuição da covariável não observada X apenas o mínimo sobre a distribuição é suficiente, por exemplo, o primeiro ou segundo momento. Na modelagem estrutural, é necessário conhecer a distribuição da covariável não observada X.

Na literatura existem diversos trabalhos que tratam de modelos de regressão com erro de medida aditivo. Muitos deles são modelos lineares com suposição de normalidade para a covariável medida com erro e para o erro de medida. A suposição de normalidade é conveniente na construção das distribuições conjuntas e condicionais, função de verossimilhança, análise de resíduos e estimação da matriz de variâncias e covariâncias. Na literatura há várias propostas de trabalhos envolvendo modelos de regressão com erro de medida aditivo, dos quais podemos citar Guolo e Brazzale (2008), Patriota (2010), Guolo (2011), Skrondal e Kuha (2012), de Castro *et al.* (2013), Carrasco *et al.* (2014) e de Castro e Galea (2015). Uma abordagem mais abrangente referente aos modelos com erro de medida pode ser encontrada nos livros de Fuller (1987), Cheng e Van Ness (1999), Carroll *et al.* (2006) e Buonaccorsi (2010).

Em algumas situações o erro de medida relacionado à covariável pode não ser aditivo e nem normalmente distribuído. Uma aplicação nesta direção é proposta por Lyles e Kupper (1997) que ajusta um modelo de regressão linear com erro de medida multiplicativo log-normal. No trabalho de Guolo e Brazzale (2008) é apresentado um estudo de simulação no qual os autores consideram um erro de medida multiplicativo para o modelo. No trabalho de Hwang (1986), o autor considerou o caso em que o erro de medida é multiplicativo e analisou alguns dados do departamento de energia dos EUA. Nossa proposta de trabalho é apresentar a construção, estimação e diagnóstico para modelos de regressão com erro de medida multiplicativo sem supor normalidade na covariável medida com erro e no erro de medida.

Esta tese está estruturada da seguinte maneira: No Capítulo 2 apresentamos as principais definições e conceitos básicos utilizados em todo o desenvolvimento da tese.

No Capítulo 3, apresentamos de forma geral o modelo de regressão série de potências com erro de medida. A classe de distribuições série de potências modificada, proposta por

Gupta (1974), é composta por várias distribuições discretas, tais como Poisson, binomial, binomial negativa, binomial negativa generalizada, Borel, Consul e Borel-Tanner. A ideia de Gupta (1974) foi representar de uma forma geral todas essas distribuições discretas e suas propriedades. Algumas referências sobre as distribuições em série de potências modificada são Gupta (1974), Gupta (1977), Gupta *et al.* (1995) e Cordeiro *et al.* (2009). Modelos de regressão cuja variável resposta pertence à classe de distribuições série de potências modificada são discutidos em Garay *et al.* (2011), Samani (2011), Samani *et al.* (2012) e Ortega *et al.* (2015). Apesar da vasta literatura que trata de modelos de regressão série de potências, não é do nosso conhecimento nenhum trabalho que discuta estes modelos com erro de medida em uma das suas covariáveis. Neste trabalho apresentamos a construção do modelo de regressão série de potências com erro de medida, incluindo definição, representações do erro de medida, métodos de estimação e análise de diagnósticos.

No Capítulo 4 apresentamos um caso particular do modelo de regressão série de potências com erro de medida. Neste capítulo definimos o modelo de regressão binomial negativa com erro de medida, supomos o erro de medida multiplicativo log-normal e apresentamos os métodos de estimação para os parâmetros de interesse. Os parâmetros desconhecidos associados à covariável medida com erro e ao erro de medida são denominados de parâmetros perturbadores. A variância do erro de medida é estimada a partir de replicações da covariável observada via equações de estimação. Para o método de pseudo verossimilhança, o uso de equações de estimação auxilia na construção da distribuição assintótica dos estimadores dos parâmetros de interesse (Carroll *et al.*, 2006). Utilizando a abordagem de Carroll *et al.* (2006) estimamos a matriz de variâncias e covariâncias assintótica dos estimadores de pseudo verossimilhança e construímos intervalos de confiança para os coeficientes. Além disso, apresentamos a análise de diagnóstico com interesse em verificar a qualidade do ajuste. As propriedades do modelo de regressão binomial negativa com erro de medida, considerando os diferentes métodos tratados neste trabalho, foram analisadas através de um estudo de simulação. A metodologia desenvolvida neste capítulo é aplicada a um conjunto de dados reais.

No Capítulo 5 é apresentado o modelo de regressão beta com erro de medida multiplicativo

log-normal. A construção deste modelo foi motivada por um problema prático que está descrito detalhadamente no próprio capítulo. Além da construção do modelo de regressão beta com erro de medida multiplicativo, alguns métodos de estimação são estudados. Tais métodos têm como princípio a estimação por máxima verossimilhança e pseudo verossimilhança. Estimamos os parâmetros perturbadores usando dados replicados. Encontramos a distribuição assintótica dos estimadores de pseudo verossimilhança com o objetivo de construir intervalos de confiança e posteriormente verificar a significância dos coeficientes usando teste de hipóteses. Para tal, também seguimos a abordagem de Carroll *et al.* (2006). Um estudo de simulação foi realizado para avaliar as propriedades de cada método e realizamos um ajuste a um conjunto de dados reais utilizando a metodologia desenvolvida neste capítulo, incluindo análise de diagnóstico.

Capítulo

2

Fundamentos Teóricos

Os conteúdos apresentados nos Capítulos 1 e 2 são pré-requisitos fundamentais para uma melhor compreensão do restante do trabalho. Nestes capítulos iniciais encontram-se a definição de erro de medida nas covariáveis, as representações do erro de medida, os métodos de estimação, critérios de seleção de modelos e análise de diagnóstico para os modelos de regressão que estão definidos nos Capítulos 3, 4 e 5.

Os modelos com erros nas variáveis são utilizados quando as covariáveis do modelo de regressão estão sujeitas a erros de medição. Na presença de uma covariável medida com erro em um modelo de regressão é extremamente importante determinar a relação existente entre a covariável observada e a covariável não observada. Os erros podem ser classificados de duas formas: erro aditivo ou erro multiplicativo. Para cada tipo de erro também tem-se duas abordagens, a abordagem clássica e a abordagem de Berkson. A abordagem clássica é mais adequada quando os erros e incertezas são individualizados, por exemplo, medição da pressão sanguínea (Carroll *et al.*, 2006). A abordagem de Berkson é mais adequada quando os erros e incertezas são coletivos, por exemplo, trabalhadores de uma mina sujeitos à mesma exposição a poeira mesmo que a verdadeira exposição seja particular a cada indivíduo (Carroll *et al.* , 2006).

Considere um modelo de regressão cujas variáveis respostas, $Y_1, \ldots, Y_n$, sejam independentes

e com suporte Ω_Y, que pode ser discreto ou contínuo. Suponha que cada variável resposta Y_i, $i = 1, \ldots, n$, esteja associada a um vetor de covariáveis medidas sem erro $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$ e a uma única covariável medida com erro X_i, $i = 1, \ldots, n$. Neste trabalho, a distribuição da covariável X_i é sempre contínua. Definidas as variáveis a serem utilizadas nos modelos de regressão propostos nesta tese, apresentamos a seguir as principais representações dos erros de medida nas covariáveis e em seguida os métodos de estimação.

2.1 Representações do Erro de Medida

2.1.1 Erro Aditivo Clássico

O erro aditivo clássico considera que a covariável observada W_i é a soma da covariável não observada X_i e o erro de medida associado ε_i, ou seja,

$$W_i = X_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.1}$$

2.1.2 Erro Aditivo de Berkson

O erro de Berkson considera que a covariável não observada é igual à soma da covariável observada e o erro de medida associado ε_i, ou seja,

$$X_i = W_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.2}$$

2.1.3 Erro Multiplicativo Clássico

O erro multiplicativo clássico considera que a covariável observada é o produto da covariável não observada pelo erro de medida associado, ou seja,

$$W_i = X_i\varepsilon_i, \quad i = 1, \ldots, n. \tag{2.3}$$

No modelo com erro multiplicativo clássico os erros e incertezas também são individualizados, por exemplo, a renda presumida de um cliente ainda não correntista em uma instituição

financeira.

2.1.4 Erro Multiplicativo de Berkson

O erro multiplicativo de Berkson considera que a covariável não observada é o produto da covariável observada pelo erro de medida associado, ou seja,

$$X_i = W_i\varepsilon_i, \quad i = 1, \ldots, n. \tag{2.4}$$

2.2 Métodos de Estimação

Os métodos de estimação propostos neste trabalho são todos baseados nas abordagens de máxima verossimilhança e pseudo verossimilhança. Antes de descrevermos cada um destes métodos, apresentamos a notação utilizada e o conceito de equações de estimação não viesadas.

1. Notação: $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi)$ é o vetor de parâmetros de interesse, em que β_0 é o intercepto do modelo, β_1 é o coeficiente da covariável medida com erro X_i, $\boldsymbol{\gamma}^\top$ é o vetor de coeficientes associado ao vetor de covariáveis medidas sem erro $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$. $\boldsymbol{\delta} = (\sigma^2_\varepsilon, \mu_{x^*}, \sigma^2_{x^*})$ é o vetor de parâmetros perturbadores. O parâmetro σ^2_ε está associado ao erro de medida, μ_{x^*} e $\sigma^2_{x^*}$ estão associados à covariável medida com erro. $\boldsymbol{\theta}$ e $\boldsymbol{\delta}$ são desconhecidos mas podem ser estimados.

2. Para o método de máxima verossimilhança aproximada, $\boldsymbol{\zeta}$ é o vetor de parâmetros a ser estimado. Este vetor é a composição entre os parâmetros de interesse e os de perturbação. Assim, $\boldsymbol{\zeta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi, \mu_{x^*}, \sigma^2_{x^*})$.

3. **Equações de estimação não viesadas**

   Todos os estimadores descritos neste trabalho podem ser caracterizados como soluções de equações de estimação não viesadas. O uso de equações de estimação é útil, pois permite o cálculo de estimativas dos erros padrão de forma correta quando no modelo há presença de covariáveis com erro de medida (Carroll *et al.*, 2006). Seja $\tilde{\mathbf{Y}}_i, i = 1, \ldots, n$, um vetor composto pelas observações da variável resposta e covariáveis. Uma equação de estimação

para $\boldsymbol{\theta}$ é dada por

$$n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_i(\tilde{\mathbf{Y}}_i;\boldsymbol{\theta})=\mathbf{0}, \tag{2.5}$$

em que $\boldsymbol{\Psi}_i$ é uma função de $(\tilde{\mathbf{Y}}_i;\boldsymbol{\theta})$ e $\widehat{\boldsymbol{\theta}}$ é a solução da expressão (2.5) com $\boldsymbol{\theta}$ variando no conjunto de possíveis valores para o vetor de parâmetros de interesse. Neste caso, $\widehat{\boldsymbol{\theta}}$ é chamado de M-estimador (Huber, 1964). Em (2.5), a função $\boldsymbol{\Psi}_i$ é chamada de *função de estimação* e é não viesada se possuir média zero quando avaliada nos verdadeiros valores dos parâmetros, ou seja,

$$I\!E\left\{\boldsymbol{\Psi}_i(\tilde{\mathbf{Y}}_i;\boldsymbol{\theta})\right\}=\mathbf{0},\quad i=1,\ldots,n.$$

Se a equação de estimação (2.5) for não viesada, sob certas condições de regularidade (Huber, 1964) $\widehat{\boldsymbol{\theta}}$ é um estimador consistente de $\boldsymbol{\theta}$. Analogamente, uma equação de estimação para $\boldsymbol{\delta}$ é dada por

$$n^{-1}\sum_{i=1}^{n}\boldsymbol{\varphi}_i(\mathbf{W}_i;\boldsymbol{\delta})=\mathbf{0}, \tag{2.6}$$

em que $\boldsymbol{\varphi}_i$ é uma função de $(\mathbf{W}_i,\boldsymbol{\delta})$ e $\widehat{\boldsymbol{\delta}}$ é a solução da expressão (2.6) para $\boldsymbol{\delta}$ variando no conjunto de possíveis valores para o vetor de parâmetros de perturbação. A função $\boldsymbol{\varphi}_i$ é não viesada se possuir média zero quando avaliada nos verdadeiros valores dos parâmetros, ou seja,

$$I\!E\left\{\boldsymbol{\varphi}_i(\mathbf{W}_i;\boldsymbol{\theta})\right\}=\mathbf{0}, i=1,\ldots,n.$$

Se a equação de estimação (2.6) for não viesada, sob certas condições de regularidade (Huber, 1964) $\widehat{\boldsymbol{\delta}}$ é um estimador consistente de $\boldsymbol{\delta}$. Nos trabalhos de Godambe (1960), Huber (1967), Carroll e Ruppert (1988) e Godambe (1991) encontram-se mais detalhes a respeito de equações de estimação.

2.2.1 Método *Naive*

O método *naive* ignora a presença do erro de medida no modelo e substitui a covariável não observada X_i pela covariável realmente observada W_i, $i = 1, \ldots, n$. A principal conveniência em se utilizar esse método é que a regressão pode ser feita pelos métodos convencionais e a principal desvantagem é que é desconsiderada a presença do erro de medida na covariável o que pode gerar estimativas viciadas e inconsistentes.

2.2.2 Método Calibração da Regressão

O método calibração da regressão substitui a covariável não observável, X_i, pela estimativa de sua esperança condicionada a W_i, $\mathbb{E}(X_i|W_i)$, $i = 1, \ldots, n$, (Carroll *et al.*, 2006, cap.4).

2.2.3 Método de Máxima Verossimilhança

No método de máxima verossimilhança todos os parâmetros $\boldsymbol{\zeta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi, \mu_{x^*}, \sigma^2_{x^*})$ são estimados simultaneamente. Neste caso, a variância do erro σ^2_ε é considerada conhecida ou estimada via replicação da covariável observada W_i, $i = 1, \ldots, n$.

Para a construção da função de verossimilhança, temos que determinar a distribuição conjunta de $(Y_i, W_i, \mathbf{z}_i)$. Como $\mathbf{z}_i$ é fixo e conhecido, não iremos considerá-lo na notação. A função densidade de (Y_i, W_i) pode ser obtida a partir da integração da função densidade conjunta dos dados completos (Y_i, W_i, X_i), denotada por $f_{Y_i,W_i,X_i}(y_i, w_i, x_i)$ com respeito a X_i, ou seja,

$$\begin{aligned} f_{Y_i,W_i}(y_i, w_i) &= \int_{\Omega_{X_i}} f_{Y_i,W_i,X_i}(y_i, w_i, x_i)dx_i \\ &= \int_{\Omega_{X_i}} f_{Y_i,W_i|X_i}(y_i, w_i|x_i)f_{X_i}(x_i)dx_i \\ &= \int_{\Omega_{X_i}} f_{Y_i|X_i,W_i}(y_i|x_i, w_i)f_{W_i|X_i}(w_i|x_i)f_{X_i}(x_i)dx_i, \end{aligned} \tag{2.7}$$

ou

$$\begin{aligned} f_{Y_i,W_i}(y_i,w_i) &= \int_{\Omega_{X_i}} f_{Y_i|X_i,W_i}(y_i|x_i,w_i) f_{X_i|W_i}(x_i|w_i) f_{W_i}(w_i) dx_i \\ &= \int_{\Omega_{X_i}} f_{Y_i|X_i}(y_i|x_i) f_{X_i|W_i}(x_i|w_i) f_{W_i}(w_i) dx_i, \end{aligned} \tag{2.8}$$

uma vez que $f_{Y_i|X_i,W_i} = f_{Y_i|X_i}$, pois a informação contida em W_i está também contida em X_i, $i = 1, \ldots, n$, sob a suposição de que o erro de medida é não diferencial. Erro de medida não diferencial ocorre quando W_i não contém informação sobre Y_i além daquela disponível em X_i, ou seja, o erro é não diferencial se a distribuição de Y_i dado (X_i, W_i) depender apenas de X_i. Caso contrário, o erro de medida é diferencial (Carroll *et al.*, 2006, pag.36).

Dados os vetores de observações $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{w} = (w_1, \ldots, w_n)$ e $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$, $i = 1, \ldots, n$, a função log-verossimilhança, $l(\boldsymbol{\zeta}; \mathbf{y}, \mathbf{w})$, para o modelo de regressão com erro de medida em uma covariável pode ser escrita como

$$\begin{aligned} l(\boldsymbol{\zeta}; \mathbf{y}, \mathbf{w}) &= \sum_{i=1}^{n} \log f_{Y_i,W_i}(y_i,w_i), \\ &= \sum_{i=1}^{n} \log \int_{\Omega_{X_i}} f_{Y_i|X_i}(y_i|x_i) f_{W_i|X_i}(w_i|x_i) f_{X_i}(x_i) dx_i \\ &= \sum_{i=1}^{n} \log \int_{\Omega_{X_i}} f_{Y_i|X_i}(y_i|x_i) f_{X_i|W_i}(x_i|w_i) f_{W_i}(w_i) dx_i, \end{aligned} \tag{2.9}$$

em que $\boldsymbol{\zeta}$ é o vetor de parâmetros a ser estimado a partir da maximização da função de log-verossimilhança (2.9). A função densidade $f_{Y_i|X_i}(y_i|x_i) = f_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i, \mathbf{z}_i)$, representa a função de probabilidade, caso Y_i seja discreto, ou a função densidade, caso Y_i seja contínuo, da variável resposta Y_i dado as covariáveis envolvidas $(X_i, W_i, \mathbf{z}_i)$, $i = 1, \ldots, n$. $f_{W_i}(w_i)$ é a função densidade da covariável observada e a função densidade condicional $f_{X_i|W_i}(x_i|w_i)$ pode ser obtida através da função distribuição acumulada, assim como $f_{W_i|X_i}(w_i|x_i)$. É importante ressaltar que caso a distribuição de X_i fosse discreta, usaríamos o somatório no lugar da integral em cada passo mostrado nessa seção.

1. **Estimação por *naive* máxima verossimilhança (NMV)**

O método NMV tem o mesmo princípio do método *naive*, ou seja, substitui-se o valor da covariável não observável, X_i, pelo valor observado com erro, W_i. O que diferencia o método NMV do método *naive* é que no método NMV os valores das estimativas dos parâmetros de interesse são obtidos via método de máxima verossimilhança, ou seja, obtidos a partir da maximização da expressão (2.9).

2. **Estimação por calibração máxima verossimilhança (CMV)**

   O método CMV tem o mesmo princípio do método calibração da regressão, ou seja, substitui-se o valor da covariável não observável, X_i, pela estimativa de sua esperança condicionada a W_i. O que diferencia o método CMV do método calibração da regressão é que no CMV os valores das estimativas dos parâmetros de interesse são obtidos via método de máxima verossimilhança, ou seja, obtidos a partir da maximização da expressão (2.9).

2.2.4 Método de Pseudo Verossimilhança

O método de pseudo verossimilhança realiza a estimação dos parâmetros em duas etapas. Primeiramente, segundo Buonaccorsi e Tosteson (1993) e Guolo (2011), estimam-se os parâmetros perturbadores associados à covariável não observada X_i e ao erro de medida ε_i, $i = 1, \ldots, n$, e, após estes parâmetros serem estimados, estimam-se os parâmetros de interesse, substituindo-se na função de log-verossimilhança original os valores dos parâmetros perturbadores já estimados. Segundo Gong e Samaniego (1981), espera-se que o estimador de máxima pseudo verossimilhança seja eficiente e consistente quando o estimador dos parâmetros perturbadores, $\widehat{\boldsymbol{\delta}}$, for eficiente e consistente sob condições de regularidade. Além disso, a distribuição assintótica do estimador $\widehat{\boldsymbol{\theta}}$ é obtida sob condições de regularidade quando o estimador $\widehat{\boldsymbol{\delta}}$ for $\sqrt{n}$-consistente e assintoticamente normal (Gong e Samaniego, 1981).

A função log pseudo verossimilhança, l_p, para o modelo de regressão com erro de medida em uma covariável é idêntica a expressão (2.9), exceto pelo vetor de parâmetros de perturbação, $\boldsymbol{\delta}$, que aqui é previamente estimado via equações de estimação não viesadas e $\boldsymbol{\theta}$ é o vetor de parâmetros de interesse a ser estimado a partir da maximização da função de log-pseudo verossimilhança l_p.

1. **Estimação por *naive* pseudo verossimilhança (NPV)**

   O método NPV tem o mesmo princípio do método NMV exceto pelos valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}^\top$ que são estimados via método de máxima pseudo verossimilhança.

2. **Estimação por calibração pseudo verossimilhança (CPV)**

   O método CPV tem o mesmo princípio do método CMV exceto pelos valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}^\top$ que são estimados via método de máxima pseudo verossimilhança.

Distribuição do estimador de máxima verossimilhança aproximada

A distribuição assintótica do estimador de máxima verossimilhança aproximada $\widehat{\boldsymbol{\zeta}}$ do vetor de parâmetros $\boldsymbol{\zeta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi, \mu_{x^*}, \sigma^2_{x^*})$ pode ser aproximada por uma distribuição normal multivariada com média $\boldsymbol{\zeta}$ e matriz de variâncias e covariâncias $\mathbf{I}_{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\zeta}})^{-1}$, em que $\mathbf{I}_{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\zeta}})$ é a matriz de informação observada avaliada nos estimadores de máxima verossimilhança aproximada, $\widehat{\boldsymbol{\zeta}}$, ou seja,

$$n^{1/2}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \xrightarrow{d} \text{MVN}(0, \mathbf{I}_{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\zeta}})^{-1}), \tag{2.10}$$

em que $\xrightarrow{d}$ indica convergência em distribuição, $\mathbf{I}_{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\zeta}}) = \partial^2/\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^\top l(\widehat{\boldsymbol{\zeta}}; \mathbf{y}, \mathbf{w})$ e $l(\boldsymbol{\zeta}; \mathbf{y}, \mathbf{w})$ é a função log-verossimilhança aproximada definida em (2.9).

Distribuição do estimador de máxima pseudo verossimilhança

Segundo Carroll *et al.* (2006), Guolo (2011) e Gong e Samaniego (1981), a distribuição do estimador de máxima pseudo verossimilhança $\widehat{\boldsymbol{\theta}}$ do vetor de parâmetros de interesse $\boldsymbol{\theta}$ pode ser aproximada por uma distribuição normal multivariada. Carroll *et al.* (2006) mostra que, sob algumas condições de regularidade (Vide Gong e Samaniego (1981))

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2.11}$$

em que a matriz de covariâncias pode ser expressa por

$$\boldsymbol{\Sigma} \approx \mathbf{A}_{n,22}^{-1}\left\{\mathbf{B}_{n,22} - \mathbf{A}_{n,21}\mathbf{A}_{n,11}^{-1}\mathbf{B}_{n,12} - \mathbf{B}_{n,12}^{\top}\mathbf{A}_{n,11}^{-\top}\mathbf{A}_{n,21}^{\top} + \mathbf{A}_{n,21}\mathbf{A}_{n,11}^{-1}\mathbf{B}_{n,11}\mathbf{A}_{n,11}^{-\top}\mathbf{A}_{n,21}^{\top}\right\}\mathbf{A}_{n,22}^{-\top}, \tag{2.12}$$

com $\mathbf{A}_{n,11}^{-\top} = (\mathbf{A}_{n,11}^{-1})^{\top}$ e

$$\begin{aligned}
\mathbf{A}_{n,11} &= \sum_{i=1}^{n} I\!E\left\{\frac{\partial}{\partial\boldsymbol{\delta}^{\top}}\boldsymbol{\varphi}_i(\mathbf{W}_i,\boldsymbol{\delta})\right\}, & \mathbf{B}_{n,11} &= \sum_{i=1}^{n}\boldsymbol{\varphi}_i(\mathbf{W}_i,\boldsymbol{\delta})\boldsymbol{\varphi}_i^{\top}(\mathbf{W}_i,\boldsymbol{\delta}),\\
\mathbf{A}_{n,21} &= \sum_{i=1}^{n} I\!E\left\{\frac{\partial}{\partial\boldsymbol{\delta}^{\top}}\boldsymbol{\Psi}_i(\tilde{\mathbf{Y}}_i,\boldsymbol{\theta},\boldsymbol{\delta})\right\}, & \mathbf{B}_{n,12} &= \sum_{i=1}^{n}\boldsymbol{\varphi}_i(\mathbf{W}_i,\boldsymbol{\delta})\boldsymbol{\Psi}_i^{\top}(\tilde{\mathbf{Y}}_i,\boldsymbol{\theta},\boldsymbol{\delta}),\\
\mathbf{A}_{n,22} &= \sum_{i=1}^{n} I\!E\left\{\frac{\partial}{\partial\boldsymbol{\theta}^{\top}}\boldsymbol{\Psi}_i(\tilde{\mathbf{Y}}_i,\boldsymbol{\theta},\boldsymbol{\delta})\right\}, & \mathbf{B}_{n,22} &= \sum_{i=1}^{n}\boldsymbol{\Psi}_i(\tilde{\mathbf{Y}}_i,\boldsymbol{\theta},\boldsymbol{\delta})\boldsymbol{\Psi}_i^{\top}(\tilde{\mathbf{Y}}_i,\boldsymbol{\theta},\boldsymbol{\delta}),
\end{aligned}$$

em que $\boldsymbol{\varphi}_i(\mathbf{W}_i,\boldsymbol{\delta})$ e $\boldsymbol{\Psi}_i(\tilde{\mathbf{Y}}_i,\boldsymbol{\theta},\boldsymbol{\delta})$ são as equações de estimação para $\boldsymbol{\delta} = (\sigma_{\varepsilon}^2, \mu_{x^*}, \sigma_{x^*}^2)$ e $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^{\top}, \phi)$. Os termos da matriz $\boldsymbol{\Sigma}$ podem ser obtidos explicitamente ou computacionalmente e são avaliados nos estimadores de máxima pseudo verossimilhança de $\boldsymbol{\theta}$ e $\boldsymbol{\delta}$. As condições de regularidade são apresentadas de forma geral no artigo de Gong e Samaniego (1981); neste trabalho, os autores provam a consistência, a eficiência e a distribuição de normalidade do estimador de pseudo verossimilhança de uma forma bem geral. Em resumo, algumas dessas condições de regularidade envolvem a existência das derivadas de primeira, segunda e terceira ordem com respeito a $\boldsymbol{\theta}$ da função de log-pseudo verossimilhança para o i-ésimo indivíduo. Além disso, as derivadas mistas entre os parâmetros de interesse e perturbadores também devem existir, bem como intercambiar a ordem de derivação e integração nas derivadas de primeira e segunda ordem com respeito a $\boldsymbol{\theta}$ e derivadas mistas com respeito a $\boldsymbol{\theta}$ e $\boldsymbol{\delta}$.

2.3 Análise de Diagnóstico

A análise de diagnóstico consiste na verificação de afastamento das suposições iniciais feitas para o modelo. Esta verificação pode ser realizada através da análise de resíduos e da análise de

influência local. Através da análise de resíduos é possível detectar a presença de pontos extremos e avaliar se a distribuição proposta para a variável resposta está adequada (McCullagh e Nelder, 1989). Através da análise de influência local, proposta por Cook (1986), podemos verificar se o modelo é sensível a pequenas perturbações. Na literatura há diversos trabalhos envolvendo análise de diagnósticos para modelos com erro de medida. Kelly (1984) definiu uma "função de influência" para o modelo de regressão linear com erro de medida aditivo e a partir dessa função construiu a matriz de covariância assintótica dos estimadores e uma medida para detectar pontos influentes para modelos com erro de medida equivalente à proposta de Cook (1977). Miller (1990) construiu resíduos para o modelo linear multivariado com erro de medida nas covariáveis. Em seu trabalho, Miller (1990) encontrou a distribuição para os resíduos padronizados e desenvolveu alguns procedimentos de diagnósticos para o modelo proposto. As medidas de diagnósticos propostas por Miller (1990) foram teste de autocorrelação, homogeneidade de variâncias e não linearidade. Joyce e Richard (1991) definiram uma função de influência para mostrar que observações extremas afetam as estimativas para modelos lineares com erro de medida. Carroll e Spiegelman (1992) propuseram testar a heterocedasticidade dos resíduos graficamente. Zhao *et al.* (1994) desenvolveram técnicas de diagnósticos para modelos lineares generalizados com erro de medida nas covariáveis. Para modelos não-lineares com erro de medida, Zhao e Lee (1995) desenvolveram medidas de influência local para vários esquemas de perturbação. de Castro *et al.* (2007) apresentam técnicas de diagnósticos para modelos de regressão com erro de medida heterocedásticos. Xie e Wei (2009) definem técnicas de diagnósticos para o modelo de regressão poisson generalizada. Galea e de Castro (2012) desenvolveram medidas de influência local para modelos com erros nas variáveis heteroscedásticos.

2.3.1 Análise de Resíduos

Em análise de regressão, após o ajuste de um modelo aos dados, é de fundamental importância verificar se os dados se adequaram bem ao modelo proposto. Esta verificação, frequentemente, é feita via análise de resíduos. Podemos entender por resíduo como a medida que calcula a discrepância entre o modelo ajustado e os dados. Através da análise de

resíduos é possível encontrar observações discrepantes (quando houver), além de verificar se os pressupostos do modelo estão sendo cumpridos (McCullagh e Nelder, 1989). Se um gráfico dos resíduos versus ordens das observações ou valores preditos apresentar um comportamento aleatório em torno do zero o modelo pode ser considerado adequado. Existem diversos trabalhos com propostas de resíduos para modelos com erro de medida na literatura, entre eles citamos: Miller (1990), Carroll e Spiegelman (1992), Zhao *et al.* (1994), Zhao e Lee (1995) e Xie e Wei (2009).

Resíduos Ordinários Padronizados de Pearson

Os resíduos ordinários padronizados de Pearson (McCullagh e Nelder, 1989) são resíduos baseados nos valores preditos $\hat{\mu}_i$ e podem ser expressos por:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{var}}(y_i)}}, i = 1, \ldots, n, \tag{2.13}$$

em que $\hat{\mu}_i$ são os valores preditos de Y_i e $\widehat{\text{var}}(y_i)$ é uma estimativa da variância de $Y_i, i = 1, \ldots, n$.

2.3.2 Análise de Influência Local

A análise de influência local é importante para detectar a presença de observações que exerçam um efeito desproporcional ao ajuste sob pequenas perturbações. Esta análise é feita através das representações gráficas de duas medidas de influência pelos índices observados. Uma verifica quais observações são conjuntamente influentes e a outra, proposta por Lesaffre e Verbeke (1998), verifica quais observações são individualmente influentes.

Seja $\boldsymbol{\omega}$ um vetor m-dimensional de perturbações restrito a um conjunto aberto $\Omega \subset I\!R^m$. A função log-verossimilhança associada a esta perturbação será denotada por $l(\boldsymbol{\theta}|\boldsymbol{\omega})$ a qual chamaremos de função log-verossimilhança perturbada. Defina $\boldsymbol{\omega}_0 \in \Omega$ tal que $l(\boldsymbol{\theta}|\boldsymbol{\omega}_0) = l(\boldsymbol{\theta})$, para todo $\boldsymbol{\theta}$, $\boldsymbol{\omega}_0$ é conhecido como vetor de não perturbação. Dessa forma, a perturbação é feita utilizando esse vetor extra $\boldsymbol{\omega}$ que pode agir tanto no modelo quanto diretamente nos dados. Posteriormente apresentamos algumas formas de como esta perturbação pode ser feita. A influência de pequenas perturbações nos estimadores de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$ pode ser verificada através de uma medida conhecida como "Distância" ou "Afastamento

de Verossimilhanças" dada por $LD_\omega = 2\left\{l(\widehat{\boldsymbol{\theta}}) - l(\widehat{\boldsymbol{\theta}}_\omega)\right\}$, em que $\widehat{\boldsymbol{\theta}}_\omega$ denota o valor de $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi)$ que maximiza $l(\boldsymbol{\theta}|\boldsymbol{\omega})$ e $l(\hat{\boldsymbol{\theta}}_\omega)$ é o valor da função de log-verossimilhança do modelo postulado avaliada em $\widehat{\boldsymbol{\theta}}_\omega$. Por modelo postulado entende-se como o modelo proposto. Podemos interpretar LD_ω como uma medida que avalia a influência sobre a estimativa de $\boldsymbol{\theta}$ ao se variar $\boldsymbol{\omega}$ através de Ω. No entanto, Cook (1986) propõe que a influência local seja avaliada apenas analisando o comportamento local de LD_ω ao redor de ω_0 que representa a ausência de perturbação do modelo postulado. Uma das formas de avaliar o comportamento local é calcular a curvatura de $LD_{\omega_0+a\mathbf{d}}$ contra a, em que $a \in I\!R$ e $\mathbf{d}$ é um vetor de direção normalizado, ou seja $\|\mathbf{d}\| = 1$. Particularmente, têm-se interesse na direção $\mathbf{d}_{\max}$ correspondente à maior curvatura $\mathbf{C}_{\mathbf{d}_{\max}}$ (Galea e de Castro, 2012). O vetor $\mathbf{d}_{\max}$ indica a direção que provoca maior perturbação no modelo ou nos dados usando o afastamento de verossimilhanças como distância. Neste contexto, Cook (1986) mostrou que a curvatura normal na direção $\mathbf{d}$ é dada por $\mathbf{C}_\mathbf{d}(\boldsymbol{\theta}) = 2|\mathbf{d}^\top\mathbf{F}\mathbf{d}|$, em que $\boldsymbol{\Delta} = \partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial\boldsymbol{\theta}\partial\omega^\top$, $\mathbf{F} = \boldsymbol{\Delta}^\top\boldsymbol{\Sigma}\boldsymbol{\Delta}$ e $\boldsymbol{\Sigma}$ é a matriz de variâncias e covariâncias observada. Para o método de máxima verossimilhança aproximada, $\boldsymbol{\Sigma}$ é substituída pela inversa da matriz de informação observada sob o modelo postulado, $\mathbf{I}_{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\zeta}})^{-1}$. Para o método de pseudo verossimilhança, $\boldsymbol{\Sigma}$ está definida em (2.12). Ambos $\boldsymbol{\Delta}$ e $\boldsymbol{\Sigma}$ são calculados em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0$.

A curvatura máxima, $\mathbf{C}_{\mathbf{d}_{\max}}$ é duas vezes o maior autovalor em valor absoluto de $\mathbf{F} = \boldsymbol{\Delta}^\top\boldsymbol{\Sigma}\boldsymbol{\Delta}$ e a direção correspondente à curvatura máxima, $\mathbf{d}_{\max}$, é o autovetor correspondente ao maior autovalor de $\mathbf{F}$. Dessa forma, se o modelo for sensível a pequenas perturbações, um gráfico das componentes de $\mathbf{d}_{\max}$ pelos índices observados pode sugerir quais observações são conjuntamente influentes (Ospina, 2007). Para verificar quais observações são individualmente influentes avaliamos a curvatura na direção do i-ésimo indivíduo, isto é, o vetor em que o i-ésimo indivíduo assume o valor um e os demais o valor zero. Neste caso, a curvatura normal é dada por:

$$C_i = 2|\Delta_i^\top\boldsymbol{\Sigma}\Delta_i|. \tag{2.14}$$

Esta medida foi proposta por Lesaffre e Verbeke (1998). Na expressão (2.14), Δ_i é a i-ésima coluna da matriz $\boldsymbol{\Delta}$. Uma vez que C_i reflete a situação em que foi atribuído o maior valor

(o valor total) possível à i-ésima coordenada de $\mathbf{d}$, tal que $\|\mathbf{d}\| = 1$, os autores denotam tal curvatura por influência local total do i-ésimo indivíduo. Verbeke e Molenberghs (2000, seção 11.3) propõem considerar um elemento como influente na estimação se $C_i \geq 2\bar{C}$, em que $\bar{C} = \sum_{i=1}^{n} C_i/n$.

Perturbação da Função de Log-Verossimilhança ou Perturbação de casos

A perturbação de casos é realizada ponderando cada elemento da soma da função log-verossimilhança da seguinte forma:

$$l(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i l_i(\boldsymbol{\theta}), \tag{2.15}$$

em que $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$, com $\omega_i \in I\!R$ para $i = 1, \ldots, n$, e $\omega_0 = \mathbf{1}_n = (1, \ldots, 1)^\top$ é o vetor de não perturbação. Note que a função de log-verossimilhança com o i-ésimo caso completamente removido corresponde ao vetor $\boldsymbol{\omega}$ com $\omega_i = 0$ e $\omega_j = 1$ para todo $j \neq i$. Para o cálculo de $C_d(\boldsymbol{\theta}) = 2|d^\top \mathbf{F} d|$ é necessário o cálculo de $\boldsymbol{\Delta} = \partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial\boldsymbol{\theta}\partial\omega^\top$. Neste caso, a i-ésima linha da matriz $\boldsymbol{\Delta}$ é dada por:

$$\boldsymbol{\Delta}_i^\top = \left[\Delta_{i1}^\top, \Delta_{i2}^\top, \Delta_{i3}^\top, \Delta_{i4}^\top\right],$$

$$\boldsymbol{\Delta}_i^\top = \left[\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\beta_0\partial\omega_i}, \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\beta_1\partial\omega_i}, \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\boldsymbol{\gamma}\partial\omega_i}, \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\phi\partial\omega_i}\right]. \tag{2.16}$$

Os elementos da i-ésima linha da matriz $\boldsymbol{\Delta}$ serão avaliados em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e $\boldsymbol{\omega} = \omega_0$. A ponderação de casos tem sido o esquema de perturbação mais utilizado para a análise de influência e pode ser interpretado como uma perturbação na variância do i-ésimo caso (Ospina, 2007). Além disso, as derivadas obtidas pelo método de perturbação da função de log-verossimilhança coincide com o vetor escore do modelo.

Perturbação da variável resposta

A perturbação na variável resposta é realizada através da adição de um vetor $\boldsymbol{\omega}_i$ de pequenas perturbações, $i = 1, \ldots, n$ ao vetor de variáveis respostas $y = (y_1, \ldots, y_n)^\top$. Como cada y_i

apresenta uma variância diferente é usual se utilizar um fator de escala para padronizar os componentes de $\boldsymbol{\omega}_i$, por exemplo, a estimativa do desvio padrão de y_i, de forma que

$$y_i(\omega) = y_i + \omega_i s_{y_i}, i = 1, \ldots, n, \tag{2.17}$$

em que s_{y_i} é a estimativa do desvio padrão de y_i, $i = 1, \ldots, n$. Para este tipo de perturbação $\boldsymbol{\omega}_0 = (0, 0, \ldots, 0)^\top$. Os termos da matriz $\boldsymbol{\Delta}$ serão idênticos aos definidos na expressão (2.16). A função de log-verossimilhança perturbada neste caso será obtida a partir da substituição de y_i por $y_i(\omega)$ (definido em (2.17)) na função de log-verossimilhança original.

2.4 Mudança Relativa

Quando na análise de diagnósticos, observa-se a presença de *outliers* ou de observações influentes, é necessário decidir o que fazer com essas observações que se destacaram. Geralmente retira-se cada observação discrepante (uma a uma ou um grupo de observações) e reestima-se o modelo sem essa ou essas observações. Para verificar qual o impacto que a retirada dessas observações causam nas estimativas, calculamos uma medida proposta por Lee *et al.* (2006) denominada "Mudança Relativa". Sejam $\hat{\theta}_i$ e $\hat{\theta}_{(-i)}$ as estimativas de máxima pseudo verossimilhança com e sem as i-ésimas observações influentes, respectivamente. Lee *et al.* (2006) definem a quantidade que mede a diferença entre $\hat{\theta}_i$ e $\hat{\theta}_{(-i)}$ por:

$$MR_i = \left| \frac{\hat{\theta}_i - \hat{\theta}_{(-i)}}{\hat{\theta}_i} \right| \times 100\%. \tag{2.18}$$

2.5 Critérios de Seleção de Modelos

O critério de informação de Akaike (AIC) e o critério de informação Bayesiano (BIC) são critérios de seleção de modelos comumente utilizados em análise de regressão que auxiliam na escolha do melhor modelo que se ajusta a determinado conjunto de dados (Paula, 2004). O

critério de informação de Akaike (AIC) é dado por

$$AIC = 2k - 2l(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w}), \tag{2.19}$$

O critério de informação Bayesiano (BIC) é dado por

$$BIC = k\log(n) - 2l(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w}), \tag{2.20}$$

em que k refere-se ao número de parâmetros a serem estimados, n é o tamanho da amostra e $l(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w})$ representa a função log-verossimilhança do modelo que varia de acordo com o método de estimação adotado. O modelo escolhido é aquele que apresenta menor valor de AIC e BIC.

Capítulo

3

Modelos de Regressão Série de Potências com Erro de Medida

A classe de distribuições série de potências modificada, proposta por Gupta (1974), é composta por várias distribuições discretas, tais como Poisson, binomial, binomial negativa, Borel, Consul e Borel-Tanner. A ideia de Gupta (1974) foi representar de uma forma geral todas essas distribuições discretas e suas propriedades. Neste capítulo apresentamos a construção do modelo de regressão série de potências com erro de medida, incluindo definição, representações do erro de medida, métodos de estimação, análise de diagnóstico e algumas funções de ligação.

3.1 Modelo

Considere um modelo de regressão cujas variáveis respostas, $Y_1, \ldots, Y_n$, sejam independentes e com suporte discreto Ω_Y, que representa um subconjunto arbitrário não vazio dos naturais, $i = 1, \ldots, n$. Suponha que cada variável resposta Y_i esteja associada a um vetor de covariáveis medidas sem erro $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$ e a uma covariável medida com erro X_i, $i = 1, \ldots, n$. Com base no suporte da variável resposta, assumimos que $Y_i|X_i, \mathbf{z}_i$ segue uma distribuição pertencente à família de distribuições série de potências modificada (Cordeiro *et al.*, 2009). A

função de probabilidade de $Y_i|X_i, \mathbf{z}_i$ é dada por

$$f_{Y_i|X_i,\mathbf{z}_i}(y_i;\mu_i,\phi) = \frac{a(y_i,\phi)\,[q(\mu_i,\phi)]^{y_i}}{h(\mu_i,\phi)}, \quad y_i \in \Omega_{Y_i} \subseteq I\!N, \quad i = 1,\ldots,n, \tag{3.1}$$

em que $\mu_i > 0$ é a média da variável aleatória Y_i, $\phi \geq 0$ é o parâmetro de dispersão, que pode ser conhecido ou não, $a(y_i,\phi)$ é um coeficiente positivo da série de potências que não depende do parâmetro μ_i, e $h(\cdot,\cdot)$ e $q(\cdot,\cdot)$ são funções positivas, finitas e duas vezes diferenciáveis com respeito a μ_i, $\quad i = 1,\ldots,n$ e ϕ. Para relacionar a média da variável resposta com as covariáveis usamos uma função de ligação, g, tal que

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_i + \boldsymbol{\gamma}^\top \mathbf{z}_i, \quad i = 1,\ldots,n, \tag{3.2}$$

em que $\beta_0 \in I\!R$ é o intercepto, $\beta_1 \in I\!R$ é o coeficiente da covariável medida com erro X_i, $i = 1,\ldots,n$, $\boldsymbol{\gamma}^\top = (\gamma_1,\ldots,\gamma_k)$, $\gamma_j \in I\!R$, $j = 1,\ldots,k$, é o coeficiente da k-ésima covariável medida sem erro e g é uma função de ligação estritamente monótona e duplamente diferenciável.

As distribuições pertencentes à família série de potências modificada possuem algumas propriedades interessantes listadas a seguir.

- A função $h(\mu_i,\phi)$ pode ser escrita como

$$h(\mu_i,\phi) = \sum_{y_i \in \Omega_{Y_i}} a(y_i,\phi)\,[q(\mu_i,\phi)]^{y_i}\,, \tag{3.3}$$

  que, após a soma, não depende de y_i, $i = 1,\ldots,n$.

- Segundo Gupta (1974) e Cordeiro *et al.* (2009), a média e a variância de Y_i, $i = 1,\ldots,n$, podem ser obtidas por

$$I\!E(Y_i|X_i,\mathbf{z}_i) = \frac{h_i' q_i}{h_i q_i'} \quad \text{e} \quad \mathrm{Var}(Y_i|X_i,\mathbf{z}_i) = \frac{q_i}{q_i'}, \tag{3.4}$$

  sendo que h_i' e q_i' são as derivadas das funções $h(\mu_i,\phi)$ e $q(\mu_i,\phi)$, respectivamente, com respeito ao parâmetro μ_i, $i = 1,\ldots,n$, e que q e q' são funções sempre positivas.

- A Tabela 3.1 apresenta algumas distribuições pertencentes à classe série de potências modificada.

Tabela 3.1: *Algumas distribuições pertencentes à classe série de potências modificada.*

Distribuição	$h(\mu,\phi)$	$q(\mu,\phi)$	$a(y,\phi)$	Ω_Y
Poisson	e^{μ}	μ	$1/y!$	$\{0,1,2,\ldots\}$
Binomial	$\left(1+\frac{\mu}{r-\mu}\right)^{r}$	$\frac{\mu}{r-\mu}$	$\binom{r}{y}$	$\{0,1,2,\ldots,r\}$
Binomial negativa	$\left(1-\frac{\mu}{\mu+\phi}\right)^{-\phi}$	$\frac{\mu}{\mu+\phi}$	$\frac{\Gamma(\phi+y)}{y!\Gamma(\phi)}$	$\{0,1,2,\ldots\}$
Poisson generalizada	$\mathrm{e}^{\mu(1+\mu\phi)-1}$	$\frac{\mu\mathrm{e}^{-\mu\phi(1+\mu\phi)^{-1}}}{1+\mu\phi}$	$\frac{(1+\phi y)^{y-1}}{y!}$	$\{0,1,2,\ldots,r\}$

Outros exemplos de distribuições pertencentes a esta classe incluem as distribuições Borel, Consul, Binomial negativa generalizada, Borel-Tanner, Geeta e Haight.

Apesar de a covariável medida com erro X_i, $i = 1,\ldots,n$, ser não observável, algumas suposições a respeito da distribuição dessa covariável podem ser feitas. Neste trabalho assumiremos para X_i apenas distribuições contínuas. Na expressão (3.2) há a presença de uma covariável $X_i, i = 1,\ldots,n$, medida com erro.

3.2 Função de Verossimilhança

3.2.1 Métodos *Naive* e Calibração da Regressão

Para os métodos *naive* e calibração da regressão a função de log-verossimilhança será dada por:

$$l(\boldsymbol{\theta};\mathbf{y},\mathbf{w}) = \sum_{i=1}^{n} l_i(\mu_i,\phi), \quad i = 1,\ldots,n,$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros de interesse a ser estimado e $l_i(\mu_i,\phi)$ é dado por

$$l_i(\mu_i,\phi) = \log\left[a(y_i,\phi)\right] + y_i \log\left[q(\mu_i,\phi)\right] - \log\left[h(\mu_i,\phi)\right], \quad i = 1,\ldots,n. \tag{3.5}$$

A função de ligação para o método *naive* é dada por

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 W_i + \boldsymbol{\gamma}^\top \mathbf{z}_i, \quad i = 1, \ldots, n, \tag{3.6}$$

para o método calibração da regressão a função de ligação é dada por

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 I\!E(X_i|W_i) + \boldsymbol{\gamma}^\top \mathbf{z}_i, \quad i = 1, \ldots, n. \tag{3.7}$$

3.2.2 Método de Máxima Verossimilhança

Dados os vetores de observações $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{w} = (w_1, \ldots, w_n)$ e $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$, $i = 1, \ldots, n$, a função de log-verossimilhança, $l(\boldsymbol{\zeta}; \mathbf{y}, \mathbf{w})$, para o modelo de regressão com erro de medida em uma covariável e variável resposta pertencente à família de distribuições em série de potências modificada, pode ser escrita como

$$l(\boldsymbol{\zeta}; \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} \log \int_{\Omega_{X_i}} \frac{a(y_i, \phi)\,[q(\mu_i, \phi)]^{y_i}}{h(\mu_i, \phi)} f_{X_i|W_i}(x_i|w_i) f_{W_i}(w_i) dx_i, \tag{3.8}$$

em que $\boldsymbol{\zeta}$ é o vetor de parâmetros a ser estimado a partir da maximização da função de log-verossimilhança (3.8) e a função densidade $f_{Y_i|X_i}(y_i|x_i) = f_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i, \mathbf{z}_i)$, representa a equação (3.1). A função de log-verossimilhança (3.8) está em função de uma integral que geralmente não possui solução explícita. Uma forma de solucionar esse problema é fazendo uso de métodos de integração numérica, por exemplo, Monte Carlo, Laplace, Quadratura de Gauss-Hermite etc.

3.2.3 Método de Pseudo Verossimilhança

A função de log-pseudo verossimilhança, $l_p(\boldsymbol{\theta}, \widehat{\boldsymbol{\delta}}; \mathbf{y}, \mathbf{w})$, para o modelo de regressão com erro de medida em uma covariável e variável resposta pertencente à família de distribuições em série de potências modificada, pode ser escrita como

$$l_p(\boldsymbol{\theta}, \widehat{\boldsymbol{\delta}}; \mathbf{y}, \mathbf{w}) = \log \int_{\Omega_{X_i}} \frac{a(y_i, \phi)\,[q(\mu_i, \phi)]^{y_i}}{h(\mu_i, \phi)} f_{X_i|W_i}(x_i|w_i; \widehat{\boldsymbol{\delta}}) f_{W_i}(w_i) dx_i, \tag{3.9}$$

diferentemente do método de máxima verossimilhança aproximada, o vetor de parâmetros de perturbação, $\boldsymbol{\delta}$, é previamente estimado por algum método de estimação, por exemplo, equações de estimação não viesadas e $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi)$ é o vetor de parâmetros de interesse a ser estimado a partir da maximização da função de log-pseudo verossimilhança definida em (3.9).

3.3 Análise de Diagnóstico

A análise de diagnóstico deste trabalho é feita através da análise de resíduos e análise de influência local descritos no Capítulo 2.

3.3.1 Análise de Resíduos

Resíduos Ordinários Padronizados de Pearson

Para os resíduos ordinários padronizados de Pearson definidos em (2.13), a $\widehat{\text{var}}(y_i)$ está definida na expressão (3.4) e $\hat{\mu}_i$ é a estimativa da média da variável resposta (expressão (3.2)).

1. **Método *naive***

   Para o método *naive* (2.2.1), μ_i, definido na expressão (3.2), é avaliado nos estimadores de máxima verossimilhança de ϕ, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}^\top$ obtidos a partir da maximização da função de log-verossimilhança dada em (3.5) substituindo-se a covariável não observável X_i pela covariável realmente observada, W_i, $i = 1, \ldots, n$.

2. **Método calibração da regressão**

   Para o método calibração da regressão (2.2.2), μ_i, definido na expressão (3.2), é avaliado nos estimadores de máxima verossimilhança de ϕ, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}^\top$ obtidos a partir da maximização da função de log-verossimilhança dada em (3.5) substituindo-se a covariável não observável X_i por $\mathbb{E}(X_i|W_i)$, $i = 1, \ldots, n$.

3. **Método máxima verossimilhança aproximada**

   Para o método de máxima verossimilhança aproximada (2.2.3), propomos substituir em μ_i a covariável não observável X_i pela covariável observada, W_i, considerando o método *naive* máxima verossimilhança (NMV) e para o método calibração máxima verossimilhança (CMV), substitui-se a covariável não observada X_i por $\mathbb{E}(X_i|W_i)$, $i = 1, \ldots, n$. Nos dois

métodos, μ_i é avaliado nos estimadores de máxima verossimilhança de ϕ, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}^\top$ obtidos a partir da maximização da função de log-verossimilhança dada em (3.8).

4. **Método pseudo verossimilhança aproximada**

   Para o método pseudo verossimilhança (2.2.4), propomos substituir em μ_i a covariável não observável X_i pela covariável observada, W_i, considerando o método *naive* pseudo verossimilhança (NPV) e para o método calibração pseudo verossimilhança (CPV), substitui-se a covariável não observada X_i por $\mathbb{E}(X_i|W_i)$, $i = 1, \ldots, n$. Nos dois métodos, μ_i é avaliado nos estimadores de máxima verossimilhança de ϕ, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}^\top$ obtidos a partir da maximização da função de log-pseudo verossimilhança dada em (3.9).

3.3.2 Análise de Influência Local

Perturbação da Função de Log-Verossimilhança ou Perturbação de casos

Os elementos da i-ésima linha da matriz $\boldsymbol{\Delta}$ serão avaliados em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e $\boldsymbol{\omega} = \omega_0$. Para os métodos *naive* e calibração da regressão, os termos da matriz $\boldsymbol{\Delta}_i$ são dados por:

$$
\begin{aligned}
\Delta_{i1}^\top &= \frac{y_i}{q(\mu_i,\phi)}\frac{\partial q(\mu_i,\phi)}{\partial \mu_i} - \frac{1}{h(\mu_i,\phi)}\frac{\partial h(\mu_i,\phi)}{\partial \mu_i},\\
\Delta_{i2}^\top &= \frac{y_i}{q(\mu_i,\phi)}\frac{\partial q(\mu_i,\phi)}{\partial \mu_i} - \frac{1}{h(\mu_i,\phi)}\frac{\partial h(\mu_i,\phi)}{\partial \mu_i}X_i,\\
\Delta_{i3}^\top &= \frac{y_i}{q(\mu_i,\phi)}\frac{\partial q(\mu_i,\phi)}{\partial \mu_i} - \frac{1}{h(\mu_i,\phi)}\frac{\partial h(\mu_i,\phi)}{\partial \mu_i}\mathbf{z}_i,\\
\Delta_{i4}^\top &= \frac{1}{a(y_i,\phi)}\frac{\partial a(y_i,\phi)}{\partial \phi} + \frac{y_i}{q(\mu_i,\phi)}\frac{\partial q(\mu_i,\phi)}{\partial \phi} - \frac{1}{h(\mu_i,\phi)}\frac{\partial h(\mu_i,\phi)}{\partial \phi},
\end{aligned}
$$

em que no termo $\Delta_{i2}^\top$, X_i é substituído por W_i para o método *naive* e no método calibração da regressão X_i é substituído por $\mathbb{E}(X_i|W_i)$, $i = 1, \ldots, n$. Para os métodos de máxima verossimilhança e pseudo verossimilhança, os termos da matriz Δ_i dependem de integrais e somente após designarmos distribuições para as variáveis envolvidas é que é possível uma simplificação. No próximo capítulo abordamos alguns casos particulares dos modelos de regressão série de potências com erro de medida: o modelo de regressão binomial negativa, geométrica e Poisson com erro de medida multiplicativo.

Capítulo

4

Modelos de Regressão Binomial Negativa com Erro de Medida

Neste capítulo apresentamos um caso particular do modelo de regressão série de potências com erro de medida apresentado no Capítulo 3. Na Seção 4.1 definimos o modelo de regressão binomial negativa com erro de medida. Na Seção 4.2 apresentamos o modelo sob a suposição de que o erro de medida é multiplicativo e tem distribuição log-normal. Como a variância do erro de medida geralmente é desconhecida, utilizamos dados replicados para auxiliar na estimação desse parâmetro. Na Seção 4.3 mostramos os métodos de estimação propostos sob a suposição de que o erro de medida é log-normal. A análise de diagnóstico é apresentada na Seção 4.4. Um estudo de simulação e uma aplicação a um conjunto de dados reais se encontram nas Subseções 4.5.1, 4.5.2 e Seção 4.6, respectivamente.

4.1 Definição do Modelo

Considere um modelo de regressão no qual as variáveis respostas independentes $Y_1, \ldots, Y_n$, com suporte no conjunto $\Omega_{Y_i} = \{0, 1, 2, \ldots\}$, pertençam à família de distribuições série de potências definida em (3.1). Além disso, suponha que cada variável resposta Y_i esteja associada a vetores de covariáveis medidas sem erro $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$ e a uma única covariável positiva

medida com erro X_i, $i = 1, \ldots, n$. Com base no suporte da variável resposta, nesta seção supomos que $Y_i|X_i, \mathbf{z}_i$, $i = 1, \ldots, n$, segue uma distribuição binomial negativa NB-2 (Hilbe , 2011). A função densidade da binomial negativa $Y_i|X_i, \mathbf{z}_i$ é dada por:

$$f_{Y_i|X_i,\mathbf{z}_i}(y_i; \mu_i, \phi) = \frac{\Gamma(y_i + 1/\phi)}{\Gamma(y_i + 1)\Gamma(1/\phi)} \left(\frac{1}{1 + \phi\mu_i}\right)^{1/\phi} \left(\frac{\phi\mu_i}{1 + \phi\mu_i}\right)^{y_i}, \quad i = 1, \ldots, n, \tag{4.1}$$

em que $a(y_i, \phi) = \frac{\Gamma(y_i+1/\phi)}{\Gamma(y_i+1)\Gamma(\phi)}$, $q(\mu_i, \phi) = \left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)$ e $h(\mu_i, \phi) = \left(\frac{1}{1+\phi\mu_i}\right)^{1/\phi}$,

em que a média é dada por $I\!E(Y_i|X_i, \mathbf{z}_i) = \mu_i$ e a variância é dada por $\text{Var}(Y_i|X_i, \mathbf{z}_i) = \mu_i(1 + \phi\mu_i)$.

A distribuição binomial negativa é popularmente conhecida por caracterizar o número de tentativas necessárias até a ocorrência do r-ésimo sucesso. Além disso, duas distribuições importantes são casos particulares da distribuição binomial negativa. Quando na equação (4.1) o parâmetro ϕ for igual a um, tem-se a distribuição Geométrica. No caso em que o parâmetro $\phi \longrightarrow 0$, tem-se a distribuição de Poisson (Piegorsch, 1990). Para dados de contagem em que ocorre superdispersão, a distribuição binomial negativa é preferível a distribuição de Poisson, pois se ajusta melhor aos dados. Para concluir não podemos deixar de mencionar que a distribuição binomial negativa também pode ser vista como um modelo de mistura entre as distribuições Gama e Poisson.

Para relacionar a variável resposta com as covariáveis utilizamos a função de ligação logarítmica:

$$\begin{aligned} g(\mu_i) = \log(\mu_i) = \eta_i &= \beta_0 + \beta_1 X_i + \boldsymbol{\gamma}^\top \mathbf{z}_i, \\ \mu_i &= \exp(\beta_0 + \beta_1 X_i + \boldsymbol{\gamma}^\top \mathbf{z}_i), \quad i = 1, \ldots, n, \end{aligned} \tag{4.2}$$

em que $\beta_0 \in I\!R$ é o intercepto, $\beta_1 \in I\!R$ é o coeficiente da covariável X_i medida com erro, $\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_k)$, $\boldsymbol{\gamma}^\top \in I\!R^k$, é o vetor de coeficientes das covariáveis medidas sem erro e g é uma função de ligação estritamente monótona e duplamente diferenciável. Para o modelo de regressão binomial negativa a variância é dada por $\text{Var}(Y_i|X_i, \mathbf{z}_i) = g^{-1}(\eta_i)(1+g^{-1}(\eta_i)/\phi)$. Note

que, mesmo o parâmetro ϕ sendo constante as variâncias não serão constantes, pois dependem de μ_i, $i = 1, \ldots, n$. Na próxima seção apresentamos o erro de medida log-normal e a estimação de sua variância via dados replicados.

4.2 Erro de Medida Log-Normal

Admitindo o modelo estrutural com erro de medida multiplicativo clássico temos

$$W_i = X_i \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.3}$$

em que X_i corresponde a i-ésima covariável não observada e W_i corresponde a i-ésima covariável observada associada ao erro de medida ε_i, $i = 1, \ldots, n$. Além disso, a covariável não observada X_i, $i = 1, \ldots, n$, possui distribuição Log-N$(\mu_{x^*}; \sigma^2_{x^*})$, o erro de medida associado à covariável, ε_i, $i = 1, \ldots, n$, é independente de X_i e possui distribuição Log-N$\left(\frac{-\sigma^2_\varepsilon}{2}; \sigma^2_\varepsilon\right)$ e a covariável observada W_i, $i = 1, \ldots, n$, tem distribuição Log-N$(\mu_{w^*}; \sigma^2_{w^*})$, com $\mu_{w^*} = \mu_{x^*} - \frac{\sigma^2_\varepsilon}{2}$ e $\sigma^2_{w^*} = \sigma^2_{x^*} + \sigma^2_\varepsilon$. A notação Log-N$(\mu_{x^*}; \sigma^2_{x^*})$ representa uma distribuição log-normal com parâmetros μ_{x^*} e $\sigma^2_{x^*}$. Com essa parametrização adotada para o erro de medida ε_i, tem-se $I\!E(\varepsilon_i) = 1$. Para auxiliar na estimação desses parâmetros aplicamos uma transformação logarítmica à expressão (4.3) e o erro de medida multiplicativo inicial é convertido em erro de medida aditivo. Desta forma a equação (4.3) resulta em

$$\begin{aligned} \log(W_i) &= \log(X_i) + \log(\varepsilon_i), \quad i = 1, \ldots, n, \\ W_i^* &= X_i^* + \varepsilon_i^*, \quad i = 1, \ldots, n, \end{aligned} \tag{4.4}$$

em que $I\!E(\varepsilon_i^*) = 0$. É importante ressaltar que a linearização feita em (4.4) é utilizada unicamente na estimação dos parâmetros perturbadores. A construção da função de verossimilhança ainda é feita levando-se em consideração o erro de medida multiplicativo log-normal. A variância do erro de medida, σ^2_ε, pode ser estimada a partir de replicações para $W_i^*, i = 1, \ldots, n$. Suponha que existam $J_i > 1, i = 1, \ldots, n$, replicações para W_i e sejam $W^*_{i1}, \ldots, W^*_{iJ_i}$ o logaritmo dessas replicações. Além disso, $\overline{W}^*_{i\cdot} = \frac{1}{J_i}\sum_{j=1}^{J_i} W^*_{ij}$ é a média dessas

medições para o i-ésimo indivíduo. Sob essa suposição, o modelo (4.4) tem a seguinte estrutura:

$$W_{ij}^* = X_{ij}^* + \varepsilon_{ij}^*, \quad i = 1, \ldots, n, \quad j = 1, \ldots, J_i, \tag{4.5}$$

em que W_{ij}^* representa a j-ésima observação para o indivíduo i e J_i o número de réplicas do i-ésimo indivíduo. Usando o método dos momentos é possível obter $\widehat{\mu}_{w^*} = \frac{1}{n}\sum_{i=1}^{n} \overline{W}_{i.}^*$ e $\widehat{\sigma}_{w^*}^2 = s_{w^*}^2$, em que $s_{w^*}^2 = \left\{(m_1 - 1)s_{w_{.1}}^2 + (m_2 - 1)s_{w_{.2}}^2 + \ldots + (m_J - 1)s_{w_{.J}}^2\right\}/(m_1 + \ldots + m_J - J)$ é a variância amostral das replicações. Nessa expressão, $s_{w_{.j}}^2$ e m_j são a variância amostral e o tamanho amostral das j-ésimas replicações, respectivamente, com $j = 1, \ldots, n$, $i = 1, \ldots, n$ e $J = \text{máx}(J_i)$. Representando os estimadores dos parâmetros perturbadores via equações de estimação não viesadas, $\widehat{\boldsymbol{\delta}}$ é a solução da equação de estimação

$$\sum_{i=1}^{n} \boldsymbol{\varphi}_i(\mathbf{W}_i, \boldsymbol{\delta}) = \mathbf{0}, \tag{4.6}$$

em que

$$\boldsymbol{\varphi}_i(\mathbf{W}_i, \boldsymbol{\delta}) = \begin{bmatrix} \varphi_{1,i}(\mathbf{W}_i, \delta_1) \\ \varphi_{2,i}(\mathbf{W}_i, \delta_2) \\ \varphi_{3,i}(\mathbf{W}_i, \delta_3) \end{bmatrix} = \begin{bmatrix} \varphi_{1,i}(\mathbf{W}_i, \sigma_\varepsilon^2) \\ \varphi_{2,i}(\mathbf{W}_i, \mu_{x^*}) \\ \varphi_{3,i}(\mathbf{W}_i, \sigma_{x^*}^2) \end{bmatrix},$$

e

$$\begin{aligned} \varphi_{1,i}(\mathbf{W}_i, \delta_1) &= (J_i - 1)\sigma_\varepsilon^2 - \sum_{j=1}^{J_i}(W_{ij}^* - \overline{W}_{i.}^*)^2, \\ \varphi_{2,i}(\mathbf{W}_i, \delta_2) &= \frac{\mu_{x^*}}{n} - \frac{1}{nJ_i}\sum_{j=1}^{J_i} W_{ij}^* - \frac{\sum_{j=1}^{J_i}(W_{ij}^* - \overline{W}_{i.}^*)^2}{2\sum_{i=1}^{n}(J_i - 1)}, \\ \varphi_{3,i}(\mathbf{W}_i, \delta_3) &= \frac{\sigma_{x^*}^2}{n} - \frac{s_{w^*}^2}{n} + \frac{\sum_{j=1}^{J_i}(W_{ij}^* - \overline{W}_{i.}^*)^2}{\sum_{i=1}^{n}(J_i - 1)}, \end{aligned} \tag{4.7}$$

em que $\boldsymbol{\delta}^\top = (\sigma^2_\varepsilon, \mu_{x^*}, \sigma^2_{x^*})$ é o vetor de parâmetros perturbadores a ser estimado. As equações de estimação definidas em (4.7) são não viesadas, pois $I\!E\left\{\sum_{i=1}^n \boldsymbol{\varphi}_i(\mathbf{W}_i, \boldsymbol{\delta})\right\} = \mathbf{0}$. De fato,

$$\begin{aligned} I\!E\left\{\sum_{i=1}^n \varphi_{1,i}(\mathbf{W}_i, \sigma^2_\varepsilon)\right\} &= \sum_{i=1}^n (J_i - 1)\sigma^2_\varepsilon - I\!E\left\{\sum_{i=1}^n \sum_{j=1}^{J_i} (W^*_{ij} - \overline{W}^*_{i.})^2\right\} \\ &= \sum_{i=1}^n (J_i - 1)\sigma^2_\varepsilon - I\!E\left\{\sum_{i=1}^n (J_i - 1)\widehat{\sigma}^2_\varepsilon\right\} \\ &= \sum_{i=1}^n (J_i - 1)\sigma^2_\varepsilon - \sum_{i=1}^n (J_i - 1) I\!E\left\{\widehat{\sigma}^2_\varepsilon\right\} = 0. \end{aligned}$$

Analogamente,

$$\begin{aligned} I\!E\left\{\sum_{i=1}^n \varphi_{2,i}(\mathbf{W}_i, \mu_{x^*})\right\} &= \mu_{x^*} - I\!E\left\{\frac{1}{nJ_i}\sum_{i=1}^n \sum_{j=1}^{J_i} W^*_{ij}\right\} - \frac{1}{2} I\!E\left\{\frac{\sum_{i=1}^n \sum_{j=1}^{J_i} (W^*_{ij} - \overline{W}^*_{i.})^2}{\sum_{i=1}^n (J_i - 1)}\right\} \\ &= \mu_{x^*} - I\!E\left\{\widehat{\mu}_{w^*}\right\} - \frac{1}{2} I\!E\left\{\widehat{\sigma}^2_\varepsilon\right\}, \\ &= \mu_{x^*} - \mu_{w^*} - \frac{\sigma^2_\varepsilon}{2} = 0, \end{aligned}$$

já que $\mu_{w^*} = \mu_{x^*} - \sigma^2_\varepsilon/2$. Finalmente,

$$\begin{aligned} I\!E\left\{\sum_{i=1}^n \varphi_{3,i}(\mathbf{W}_i, \sigma^2_{x^*})\right\} &= \sigma^2_{x^*} - I\!E\left\{s^2_{w^*}\right\} + I\!E\left\{\frac{\sum_{i=1}^n \sum_{j=1}^{J_i} (W^*_{ij} - \overline{W}^*_{i.})^2}{\sum_{i=1}^n (J_i - 1)}\right\} \\ &= \sigma^2_{x^*} - \sigma^2_{w^*} + I\!E\left\{\widehat{\sigma}^2_\varepsilon\right\} \\ &= \sigma^2_{x^*} - \sigma^2_{w^*} + \sigma^2_\varepsilon = 0, \end{aligned}$$

já que $\sigma^2_{w^*} = \sigma^2_{x^*} + \sigma^2_\varepsilon$.

4.3 Métodos de Estimação

4.3.1 Método *Naive*

Dados os vetores de observações, $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{w} = (w_1, \ldots, w_n)$ e $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$, $i = 1, \ldots, n$, a função de log-verossimilhança é dada por

$$l(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} l_i(\mu_i, \phi), \quad i = 1, \ldots, n. \tag{4.8}$$

Para o modelo de regressão binomial negativa, $l_i(\mu_i, \phi)$ é dado por

$$l_i(\mu_i, \phi) = \log\Gamma(y_i + 1/\phi) - \log\Gamma(y_i + 1) - \log\Gamma(1/\phi) - \phi \log\left(\frac{1}{1+\phi\mu_i}\right) + y_i \log\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right), \tag{4.9}$$

em que o valor da covariável não observável, X_i, é substituído pelo valor observado com erro, W_i, na expressão (4.2), ou seja

$$\mu_i = \exp(\beta_0 + \beta_1 W_i + \boldsymbol{\gamma}^\top \mathbf{z}_i), \quad i = 1, \ldots, n. \tag{4.10}$$

Em linguagem R (R Core Team, 2014) a estimação dos parâmetros pode ser feita através da função *glm.nb* do pacote *MASS* para o modelo de regressão binomial negativa e para o modelo de regressão de Poisson através da função *glm*. É importante ressaltar que na presença de replicações, W_i é substituído por $\overline{W}_i = \frac{\sum_{j=1}^{J} W_{ij}}{J_i}, i = 1, \ldots, n$. Para a construção de intervalos de confiança para os parâmetros é possível utilizar a função *vcov* que fornece a matriz de variância e covariância dos estimadores.

4.3.2 Método Calibração da Regressão

Substituindo na expressão (4.2) o valor da covariável não observável, X_i, pela esperança condicional, $\mathbb{E}(X_i|W_i)$, temos

$$\mu_i = \exp(\beta_0 + \beta_1 \mathbb{E}(X_i|W_i) + \boldsymbol{\gamma}^\top \mathbf{z}_i), \quad i = 1, \ldots, n, \tag{4.11}$$

em que

$$\widehat{I\!E}(X_i|W_i) = W_i^{\widehat{\lambda}} \exp\left(\widehat{\alpha} + \widehat{\lambda}\frac{\widehat{\sigma}_\varepsilon^2}{2}\right), \quad i = 1, \ldots, n, \tag{4.12}$$

com $\widehat{\alpha} = \widehat{\mu}_{w^*}(1 - \widehat{\lambda}) + \frac{\widehat{\sigma}_\varepsilon^2}{2}$, $\widehat{\lambda} = \widehat{\sigma}_{x^*}^2/\widehat{\sigma}_{w^*}^2$ e na presença de replicação da covariável observada W_i, substitui-se W_i por $\overline{W}_i = \sum_{j=1}^{J_i} W_{ij}/J_i$, $i = 1, \ldots, n$. A expressão (4.12) é proposta por Carroll *et al.* (2006, pag. 74).

Para o modelo de regressão binomial negativa, μ_i definido em (4.11), é avaliado nos estimadores de máxima verossimilhança obtidos a partir da maximização da função de log-verossimilhança (4.8), com $l_i(\mu_i, \phi)$ definido em (4.9).

Os parâmetros, σ_ε^2, $\sigma_{x^*}^2$, μ_{w^*} e $\sigma_{w^*}^2$ podem ser previamente estimados via equações de estimação (vide Seção 4.2). Para a construção de intervalos de confiança para os parâmetros de interesse é necessário calcularmos os erros padrão dos seus respectivos estimadores. Utilizamos a sugestão dada por Carroll *et al.* (2006, Seção 4.6), que calcula as variâncias dos estimadores de interesse via *bootstrap* não paramétrico.

4.3.3 Método de Máxima Verossimilhança Aproximada

Para a construção da função de verossimilhança, temos que determinar a função densidade conjunta de (Y_i, W_i) já definida na equação (2.7). Como $f_{X_i}(x_i)$ é conhecida, precisamos determinar $f_{Y_i,W_i|X_i}(y_i, w_i|x_i)$. Utilizando a função distribuição acumulada de $Y_i, W_i|X_i$, temos:

$$\begin{aligned} F_{Y_i,W_i|X_i}(y_i, w_i|x_i) &= I\!P(Y_i \leq y_i, W_i \leq w_i|X_i = x_i), \\ &= I\!P(Y_i \leq y_i, X_i\varepsilon_i \leq w_i|X_i = x_i) \\ &= I\!P\left(Y_i \leq y_i, \varepsilon_i \leq \frac{w_i}{X_i}|X_i = x_i\right). \end{aligned} \tag{4.13}$$

Considerando a independência entre Y_i e ε_i dado $X_i = x_i$, a expressão (4.13) pode ser reescrita como

$$\begin{aligned} F_{Y_i,W_i|X_i}(y_i, w_i|x_i) &= I\!P(Y_i \leq y_i|X_i = x_i) I\!P\left(\varepsilon_i \leq \frac{w_i}{x_i}\right), \\ &= F_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i, \mathbf{z}_i)\, F_{\varepsilon_i}\left(\frac{w_i}{x_i}\right). \end{aligned} \tag{4.14}$$

A função densidade de $Y_i, W_i|X_i$ é dada pelo produto da densidade de $Y_i|X_i, \mathbf{z}_i$ pela densidade de ε_i, ou seja

$$f_{Y_i,W_i|X_i}(y_i, w_i|x_i) = f_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i, \mathbf{z}_i)\frac{1}{x_i} f_{\varepsilon_i}\left(\frac{w_i}{x_i}\right), \tag{4.15}$$

em que a densidade correspondente a função acumulada $F_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i, \mathbf{z}_i)$ é obtida por

$$\begin{aligned} f_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i, \mathbf{z}_i) = I\!P(Y_i = y_i|z_i, x_i) = I\!P(Y_i \leq y_i|x_i, \mathbf{z}_i) - \lim_{\delta\to 0} I\!P(Y_i \leq y_i - \delta|x_i, \mathbf{z}_i) \\ f_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i, \mathbf{z}_i) = F_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i, \mathbf{z}_i) - \lim_{\delta\to 0} F_{Y_i|X_i,\mathbf{z}_i}(y_i - \delta|x_i, \mathbf{z}_i), \end{aligned} \tag{4.16}$$

uma vez que $Y_i|X_i, \mathbf{z}_i$ tem distribuição discreta. Como W_i é uma variável aleatória contínua, derivamos $F_{\varepsilon_i}\left(\frac{w_i}{x_i}\right)$ com relação a w_i. Considerando a expressão (2.7), a densidade conjunta de (Y_i, W_i) é escrita como

$$f_{Y_i,W_i}(y_i, w_i) = \int_{\Omega_{X_i}} f_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i, \mathbf{z}_i)\frac{1}{x_i} f_{\varepsilon_i}\left(\frac{w_i}{x_i}\right) f_{X_i}(x_i)dx_i. \tag{4.17}$$

Dados os vetores de observações $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{w} = (w_1, \ldots, w_n)$ e $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$, $i = 1, \ldots, n$, a função de log-verossimilhança, $l(\boldsymbol{\zeta}; \mathbf{y}, \mathbf{w})$, para o modelo de regressão com erro de medida em uma covariável e variável resposta pertencente a família de distribuições em série

de potências modificadas, pode ser escrita como

$$\begin{aligned} l(\boldsymbol{\zeta};\mathbf{y},\mathbf{w}) &= \sum_{i=1}^{n}\log f_{Y_i,W_i}(y_i,w_i) \\ &= \sum_{i=1}^{n}\log\int_{\Omega_{X_i}} f_{Y_i,W_i|X_i}(y_i,w_i|x_i)f_{X_i}(x_i)dx_i \\ &= \sum_{i=1}^{n}\log\int_{\Omega_{X_i}} f_{Y_i|X_i,\mathbf{z}_i}(y_i|x_i,\mathbf{z}_i)\frac{1}{x_i}f_{\varepsilon_i}\left(\frac{w_i}{x_i}\right)f_{X_i}(x_i)dx_i. \end{aligned} \tag{4.18}$$

Substituindo na expressão (4.18) as distribuições assumidas na seção (4.2) para as variáveis envolvidas, a função de log-verossimilhança para o modelo de regressão binomial negativa com erro de medida é dada por

$$\begin{aligned} l(\boldsymbol{\zeta};\mathbf{y},\mathbf{w}) \;=\; & \sum_{i=1}^{n}\log\int_0^{\infty}\frac{\Gamma(y_i+1/\phi)}{\Gamma(y_i+1)\Gamma(1/\phi)}\left(\frac{1}{1+\phi\mu_i}\right)^{1/\phi}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{y_i} \\ & \times\frac{1}{w_i\sigma_\varepsilon\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left\{\log(w_i)-\log(x_i)+\frac{\sigma_\varepsilon^2}{2}\right\}^2\right) \\ & \times\frac{1}{x_i\sigma_{x^*}\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_{x^*}^2}\left\{\log(x_i)-\mu_{x^*}\right\}^2\right)dx_i, \end{aligned} \tag{4.19}$$

em que a variância do erro σ_ε^2 é conhecida ou estimada através da solução de uma das equações de estimação definidas na expressão (4.6). O objetivo em supor que a variância do erro é conhecida ou estimada no método de máxima verossimilhança é reduzir o número de parâmetros a serem estimados. A expressão (4.19) depende do vetor de parâmetros $\boldsymbol{\zeta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi, \mu_{x^*}, \sigma_{x^*}^2)$ e de uma integral intratável.

Para simplificar a notação, vamos reescrever essas expressões da seguinte forma:

$$l(\boldsymbol{\zeta};\mathbf{y},\mathbf{w}) = \sum_{i=1}^{n}\log\int_0^{\infty} f_{Y_i,W_i|X_i}(y_i,w_i|x_i)\times\frac{1}{x_i\sigma_{x^*}\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_{x^*}^2}\left\{\log(x_i)-\mu_{x^*}\right\}^2\right)dx_i. \tag{4.20}$$

Para o modelo de regressão binomial negativa com erro de medida, $f_{Y_i,W_i|X_i}(y_i,w_i|x_i)$ representa

a seguinte expressão

$$f_{Y_i,W_i|X_i}(y_i,w_i|\,x_i) = \frac{\Gamma(y_i+1/\phi)}{\Gamma(y_i+1)\Gamma(1/\phi)}\left(\frac{1}{1+\phi\mu_i}\right)^{1/\phi}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{y_i} \times \frac{1}{w_i\sigma_\varepsilon\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left\{\log(w_i)-\log(x_i)+\frac{\sigma_\varepsilon^2}{2}\right\}^2\right), \quad i=1,\ldots,n. \tag{4.21}$$

Considerando as seguintes mudanças de variáveis:

$$u_i = \frac{\log(x_i)-\mu_{x^*}}{\sigma_{x^*}\sqrt{2}} \quad \text{e} \quad du_i = \frac{dx_i}{x_i\sigma_{x^*}\sqrt{2}},$$

temos $x_i = \exp(u_i\sigma_{x^*}\sqrt{2}+\mu_{x^*})$. Assim, a equação (4.20) pode ser reescrita como:

$$l(\boldsymbol{\zeta};\mathbf{y},\mathbf{w}) = \sum_{i=1}^{n}\log\int_{-\infty}^{\infty}\frac{1}{\sqrt{\pi}}f_{Y_i,W_i|X_i}(y_i,w_i|\,x_i)\exp(-u_i^2)du_i. \tag{4.22}$$

A integral em (4.22) pode ser resolvida utilizando o método da quadratura de Gauss-Hermite (Abramowitz e Stegun, 1964) e a função de log-verossimilhança aproximada, $l(\boldsymbol{\zeta};\mathbf{y},\mathbf{w})$, dada em (4.22), agora é denotada por $l_H(\boldsymbol{\zeta};\mathbf{y},\mathbf{w})$, por se tratar do logaritmo da função de verossimilhança que depende de pontos e nós da quadratura de Hermite, ou seja,

$$l_H(\boldsymbol{\zeta};\mathbf{y},\mathbf{w}) = \sum_{i=1}^{n}\log\left(\sum_{q=1}^{Q}\frac{1}{\sqrt{\pi}}p_q f_{Y_i,W_i|X_{iq}}(y_i,w_i|\,x_{iq})\right), \tag{4.23}$$

em que Q é o número de pontos da quadratura, $r_1,\ldots,r_Q$ são as raízes do polinômio ortogonal de Hermite $H_q(x)$ (ou nós da quadratura) e $p_1,\ldots,p_Q$ são os pesos dados por

$$p_q = \frac{2^{Q-1}n!\sqrt{\pi}}{n^2\left[H_{Q-1}(u_q)\right]^2}, \quad q=1,2,\ldots,Q. \tag{4.24}$$

1. **Estimação por *Naive* Máxima Verossimilhança (NMV)**

   O método NMV consiste em encontrar as estimativas de μ_i (que serão usadas como possíveis valores preditos) substituindo os valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}'$ estimados via método de máxima verossimilhança aproximada, ou seja, obtidos a partir da maximização

da expressão (4.23), no preditor linear da regressão dado em (4.10):

$$\widehat{\mu}_i = \exp(\widehat{\beta}_0 + \widehat{\beta}_1 W_i + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i), \quad i = 1, \ldots, n. \tag{4.25}$$

2. **Estimação por Calibração Máxima Verossimilhança (CMV)**

   O método CMV consiste em encontrar as estimativas de μ_i substituindo os valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}'$ estimados via método de máxima verossimilhança aproximada, ou seja, obtidos a partir da maximização da expressão (4.23), no preditor linear da regressão dado em (4.11):

$$\widehat{\mu}_i = \exp(\widehat{\beta}_0 + \widehat{\beta}_1 \widehat{I\!E}(X_i|W_i) + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i), \quad i = 1, \ldots, n, \tag{4.26}$$

   em que $\widehat{I\!E}(X_i|W_i)$ é obtida pela expressão (4.12). Na presença de replicação da covariável W_i, $\widehat{I\!E}(X_i|W_i)$ é substituída por $\widehat{I\!E}(X_i|\overline{W}_i), i = 1, \ldots, n$.

4.3.4 Método de Pseudo Verossimilhança

Como vimos no Capítulo 2, a principal diferença entre os métodos de máxima verossimilhança e pseudo verossimilhança é que o vetor de parâmetros perturbadores $\boldsymbol{\delta} = (\sigma^2_\varepsilon, \mu_{x^*}, \sigma^2_{x^*})$ é previamente estimado via equações de estimação definidas em (4.6). Estimar os parâmetros perturbadores previamente é uma forma efetiva de reduzir o número de parâmetros a serem estimados. Assim, a função de log-pseudo verossimilhança é idêntica a expressão (4.23), em que para o modelo de regressão binomial negativa com erro de medida, $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi)$ é o único vetor de parâmetros a ser estimado dado o vetor estimado $\widehat{\boldsymbol{\delta}}$. Assim, a função de log-pseudo verossimilhança de hermite é dada por

$$l_H(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} \log \left(\sum_{q=1}^{Q} \frac{1}{\sqrt{\pi}} p_q f_{Y_i, W_i|X_{iq}} \left(y_i, w_i | \, x_{iq} \right) \right), \tag{4.27}$$

em que $x_{iq} = r_q\widehat{\sigma}_{x^*}\sqrt{2}+\widehat{\mu}_{x^*}$. Usando o método de equações de estimação, o vetor $\widehat{\boldsymbol{\theta}}$ é a solução da equação de estimação

$$\sum_{i=1}^{n} \boldsymbol{\Psi}_i(\tilde{\mathbf{Y}}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\delta}}) = 0, \tag{4.28}$$

em que

$$\boldsymbol{\Psi}_i(\tilde{\mathbf{Y}}_i; \boldsymbol{\theta}, \widehat{\boldsymbol{\delta}}) = \begin{bmatrix} \Psi_{1,i}(\tilde{\mathbf{Y}}_i; \theta_1, \widehat{\boldsymbol{\delta}}) \\ \Psi_{2,i}(\tilde{\mathbf{Y}}_i; \theta_2, \widehat{\boldsymbol{\delta}}) \\ \Psi_{3,i}(\tilde{\mathbf{Y}}_i; \theta_3, \widehat{\boldsymbol{\delta}}) \\ \Psi_{4,i}(\tilde{\mathbf{Y}}_i; \theta_4, \widehat{\boldsymbol{\delta}}) \end{bmatrix} = \begin{bmatrix} \Psi_{1,i}(\tilde{\mathbf{Y}}_i; \beta_0, \widehat{\boldsymbol{\delta}}) \\ \Psi_{2,i}(\tilde{\mathbf{Y}}_i; \beta_1, \widehat{\boldsymbol{\delta}}) \\ \Psi_{3,i}(\tilde{\mathbf{Y}}_i; \boldsymbol{\gamma}, \widehat{\boldsymbol{\delta}}) \\ \Psi_{4,i}(\tilde{\mathbf{Y}}_i; \phi, \widehat{\boldsymbol{\delta}}) \end{bmatrix},$$

e $\Psi_{j,i}(\tilde{\mathbf{Y}}_i; \theta_j, \boldsymbol{\delta}) = \partial l_H(\boldsymbol{\theta}; \boldsymbol{\delta})/\partial\theta_j$, $j = 1, \ldots, 4$, $l_H(\boldsymbol{\theta}; \boldsymbol{\delta})$ é a função de log-pseudo verossimilhança definida em (4.27). A expressão (4.28) apresenta equações de estimação para quatro parâmetros. Na Seção (4.2) mostramos que as equações de estimação para o vetor $\boldsymbol{\delta}$ são não viesadas. Desta forma, obtemos equações de estimação não viesadas para todos os parâmetros.

1. **Estimação por *Naive* Pseudo Verossimilhança (NPV)**

   O método NPV consiste em encontrar as estimativas de μ_i (que serão usadas como possíveis valores preditos) substituindo os valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}^{'}$ estimados via método de máxima pseudo verossimilhança aproximada, ou seja, obtidos a partir da maximização da expressão (4.27), no preditor linear da regressão dado em (4.10):

   $$\widehat{\mu}_i = \exp(\widehat{\beta}_0 + \widehat{\beta}_1 W_i + \widehat{\boldsymbol{\gamma}}^{\top}\mathbf{z}_i), \quad i = 1, \ldots, n. \tag{4.29}$$

2. **Estimação por Calibração Pseudo Verossimilhança (CPV)**

   O método CPV consiste em encontrar as estimativas de μ_i substituindo os valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}^{'}$ estimados via método de máxima pseudo verossimilhança aproximada, ou seja, obtidos a partir da maximização da expressão (4.27), no preditor

linear da regressão dado em (4.11):

$$\widehat{\mu}_i = \exp(\widehat{\beta}_0 + \widehat{\beta}_1 \widehat{I\!E}(X_i|W_i) + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i), \quad i = 1, \ldots, n, \tag{4.30}$$

em que $\widehat{I\!E}(X_i|W_i)$ é obtida pela expressão (4.12). Na presença de replicação da covariável W_i, $\widehat{I\!E}(X_i|W_i)$ é substituída por $\widehat{I\!E}(X_i|\overline{W}_i), i = 1, \ldots, n$.

4.4 Análise de Diagnóstico

Nesta seção apresentamos os principais métodos de diagnósticos para o modelo de regressão binomial negativa com erro de medida multiplicativo log-normal, caso particular da classe descrita no Capítulo 3. Além disso, definimos os resíduos ordinários padronizados de Pearson considerando os quatro métodos propostos de estimação: *naive*, calibração da regressão, máxima verossimilhança e pseudo verossimilhança. Para a análise de influência local, que verifica se o modelo é sensível a pequenas perturbações, é utilizada apenas a pertubação de casos ou perturbação da função de log-verossimilhança.

4.4.1 Resíduos Ordinários Padronizados e Pearson

Os resíduos ordinários padronizados já foram definidos em (2.13). Para o modelo de regressão binomial negativa com erro de medida, $\widehat{\text{var}}(y_i) = \hat{\mu}_i(1 + \hat{\mu}_i/\hat{\phi})$.

1. **Método *Naive*:**

   Para o método *naive* (4.3.1), μ_i, definido na expressão (4.10), é avaliado nos estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}'$ obtidos a partir da maximização da função de log-verossimilhança dada em (4.8).

2. **Método Calibração da Regressão:**

   Para o método calibração da regressão (4.3.2), μ_i, definido na expressão (4.11), é avaliado nos estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}'$ obtidos a partir da maximização da função de log-verossimilhança dada em (4.8).

3. **Método Máxima Verossimilhança Aproximada:**

   Para o método *naive* máxima verossimilhança, NMV, substitui-se na expressão de μ_i a covariável não observável X_i por W_i e para o método calibração máxima verossimilhança, CMV, substitui-se a covariável não observada X_i por $I\!E(X_i|W_i)$, $i = 1, \ldots, n$. Nos dois métodos, μ_i é avaliado nos estimadores de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}^{'}$ obtidos a partir da maximização da função de log-verossimilhança de Hermite dada em (4.23) bem como o parâmetro ϕ.

4. **Método Pseudo Verossimilhança:**

   Para o método *naive* pseudo verossimilhança, NPV, substitui-se na expressão de μ_i a covariável não observável X_i por W_i e para o método calibração pseudo verossimilhança, CPV, substitui-se a covariável não observada X_i por $I\!E(X_i|W_i)$, $i = 1, \ldots, n$. Nos dois métodos, μ_i é avaliado nos estimadores de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}^{'}$ obtidos a partir da maximização da função de log-pseudo verossimilhança de Hermite dada em (4.27).

4.4.2 Análise de Influência Local

1. **Perturbação da Função log-verossimilhança ou Perturbação de casos**

(a) **Método *Naive*:** Para o método *naive*, a função de log-verossimilhança perturbada para o modelo de regressão binomial negativa com erro de medida multiplicativo é construída a partir da perturbação da função de log-verossimilhança dada em (4.8), que resulta em

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{\omega}) &= \sum_{i=1}^{n} \omega_i l_i(\mu_i, \phi) \\ &= \sum_{i=1}^{n} \{\omega_i \log\Gamma(y_i + 1/\phi) - \omega_i \log\Gamma(y_i + 1) - \omega_i \log\Gamma(1/\phi) - (\omega_i/\phi)\log(1 + \phi\mu_i) \\ &\quad + \omega_i y_i \log\left(\frac{\phi\mu_i}{1 + \phi\mu_i}\right)\Big\}. \end{aligned} \tag{4.31}$$

Os termos da matriz $\mathbf{\Delta}_i$, definidos em (2.16), são expressos por

$$\begin{aligned}
\Delta_{i1}^{\top} &= \frac{\phi(y_i-\mu_i)}{\mu_i(1+\phi\mu_i)},\\
\Delta_{i2}^{\top} &= \frac{\phi(y_i-\mu_i)}{\mu_i(1+\phi\mu_i)}W_i,\\
\Delta_{i3}^{\top} &= \frac{\phi(y_i-\mu_i)}{\mu_i(1+\phi\mu_i)}\mathbf{z}_i,\\
\Delta_{i4}^{\top} &= \frac{1}{\phi^2}\left\{\psi\left(1/\phi\right)-\psi\left(y_i+1/\phi\right)+\log(1+\mu_i\phi)+y_i\phi\right\}-\frac{\mu_i}{\phi}\left(\frac{y_i\phi+1}{1+\mu_i\phi}\right),
\end{aligned}$$

em que $\mathbf{z}_i = (z_{i1},\ldots,z_{ik})$, $i = 1,\ldots,n$, é o vetor de covariáveis medidas sem erro e a matriz de variâncias e covariâncias observada, $\mathbf{\Sigma}$, é obtida a partir da função *vcov* em R (R Core Team, 2014).

(b) **Método Calibração da Regressão:** Para o método calibração da regressão, a função de log-verossimilhança perturbada para o modelo de regressão binomial negativa com erro de medida multiplicativo é idêntica a (4.31), com os termos da matriz $\mathbf{\Delta}_i$, definidos em (2.16), dados por

$$\begin{aligned}
\Delta_{i1}^{\top} &= \frac{\phi(y_i-\mu_i)}{\mu_i(1+\phi\mu_i)},\\
\Delta_{i2}^{\top} &= \frac{\phi(y_i-\mu_i)}{\mu_i(1+\phi\mu_i)}I\!\!E(X_i|W_i),\\
\Delta_{i3}^{\top} &= \frac{\phi(y_i-\mu_i)}{\mu_i(1+\phi\mu_i)}\mathbf{z}_i,\\
\Delta_{i4}^{\top} &= \frac{1}{\phi^2}\left\{\psi\left(1/\phi\right)-\psi\left(y_i+1/\phi\right)+\log(1+\mu_i\phi)+y_i\phi\right\}-\frac{\mu_i}{\phi}\left(\frac{y_i\phi+1}{1+\mu_i\phi}\right),
\end{aligned}$$

em que $\mathbf{z}_i = (z_{i1},\ldots,z_{ik})$, $i = 1,\ldots,n$, é o vetor de covariáveis medidas sem erro, a esperança condicional $I\!\!E(X_i|W_i)$ está definida em (4.12) e os termos da matriz de variâncias e covariâncias, $\mathbf{\Sigma}$, são obtidos via *bootstrap* não-paramétrico.

(c) **Método de Máxima Verossimilhança Aproximada:**

A função de log-verossimilhança perturbada para o modelo de regressão binomial negativa com erro de medida multiplicativo é construída a partir da perturbação da função de

log-verossimilhança de Hermite, definida em (4.23), ou seja

$$l(\boldsymbol{\zeta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i l_H(\boldsymbol{\zeta}; \mathbf{y}, \mathbf{w}) \tag{4.32}$$
$$= \sum_{i=1}^{n} \omega_i \log\left(\sum_{q=1}^{Q} \frac{1}{\sqrt{\pi}} p_q f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)\right),$$

em que a $f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)$ está definida em (4.21) para o modelo de regressão binomial negativa com erro de medida com $x_{iq} = \exp(r_q\widehat{\sigma}_{x^*}\sqrt{2} + \widehat{\mu}_{x^*})$, Q o número de pontos da quadratura, $r_1, \ldots, r_Q$ as raízes do polinômio ortogonal de Hermite $H_q(x)$ (ou nós da quadratura) e $p_1, \ldots, p_Q$ são os pesos do polinômio já apresentados em (4.24).

(d) **Método de Pseudo Verossimilhança:**

A função de log-pseudo verossimilhança perturbada para o modelo de regressão binomial negativa com erro de medida multiplicativo é construída a partir da perturbação da função de log-pseudo verossimilhança de Hermite, dada em (4.27), que resulta em

$$l_p(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i l_H(\boldsymbol{\theta}; \widehat{\boldsymbol{\delta}}) \tag{4.33}$$
$$= \sum_{i=1}^{n} \omega_i \log\left(\sum_{q=1}^{Q} \frac{1}{\sqrt{\pi}} p_q f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)\right),$$

com $x_{iq} = \exp(r_q\widehat{\sigma}_{x^*}\sqrt{2} + \widehat{\mu}_{x^*})$.

Para o modelo de regressão binomial negativa com erro de medida multiplicativo, os termos da matriz $\boldsymbol{\Delta}_i$, definida em (2.16), são os mesmos para os métodos de máxima verossimilhança aproximada e pseudo verossimilhança, ou seja,

$$\Delta_{i1}^{\top} = \frac{1}{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)} \sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right) \frac{(y_i - \widehat{\mu}_{iq})}{(1 + \widehat{\phi}\widehat{\mu}_{iq})},$$
$$\Delta_{i2}^{\top} = \frac{1}{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)} \sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right) \frac{(y_i - \widehat{\mu}_{iq})}{(1 + \widehat{\phi}\widehat{\mu}_{iq})} x_{iq},$$
$$\Delta_{i3}^{\top} = \frac{1}{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)} \sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right) \frac{(y_i - \widehat{\mu}_{iq})}{(1 + \widehat{\phi}\widehat{\mu}_{iq})} \mathbf{z}_i,$$

$$\Delta_{i4}^{\top} = \frac{1}{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)} \sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right) \left\{ \frac{(y_i - \widehat{\mu}_{iq})}{\widehat{\phi}(1+\widehat{\phi}\widehat{\mu}_{iq})} \right.$$
$$\left. -\frac{1}{\widehat{\phi}^2}\left\{\psi\left(\frac{1}{\widehat{\phi}}\right) + \psi\left(y_i + \frac{1}{\widehat{\phi}}\right) + \log\left(\frac{1}{1+\widehat{\phi}\mu_i}\right)\right\}\right\}, \quad (4.34)$$

em que $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$, $i = 1, \ldots, n$, é o vetor de covariáveis medidas sem erro e ψ é a função digama dada por $\psi(z) = \partial \log \Gamma(z)/\partial z, z > 0$. Na matriz $\mathbf{\Delta}_i$ para o método de máxima pseudo verossimilhança, $\widehat{\mu}_{iq}$ é a função de ligação definida em (4.2) avaliada nas estimativas de pseudo verossimilhança, com X_i substituído por $x_{iq} = \exp(r_q \widehat{\sigma}_{x^*}\sqrt{2} + \widehat{\mu}_{x^*})$. Para o método de máxima verossimilhança aproximada, $\widehat{\mu}_{iq}$ é a função de ligação definida em (4.2) avaliada nas estimativas de máxima verossimilhança aproximada, com X_i substituído por $x_{iq} = \exp\left(r_q \sigma_{x^*}\sqrt{2} + \mu_{x^*}\right)$. No caso do método de máxima pseudo verossimilhança, $\mathbf{\Sigma}$ é a matriz de variâncias e covariâncias corrigida definida pela expressão (2.12). Para o método de máxima verossimilhança aproximada, $\mathbf{\Sigma}$ é a inversa da matriz de informação de fisher observada.

4.5 Estudo de Simulação

Nesta seção são apresentados os resultados de um estudo de simulação para os modelos de regressão binomial negativa, geométrica e Poisson. Todos com erro de medida multiplicativo log-normal.

Para cada método, foram realizadas $N = 1000$ simulações e tamanhos de amostra $n = 50, 100, 150$ e 200. A variância do erro σ_ε^2 é desconhecida mas é estimada utilizando duas replicações para $W_i, i = 1, \ldots, n$.

Foram calculadas as médias das estimativas nas N simulações, as raízes dos erros quadráticos médios ($\sqrt{\text{EQM}}$), as médias dos erros padrão assintóticos (EPA), os erros padrão empíricos (EPE) e as probabilidades de cobertura de 95% para os intervalos de confiança assintóticos ($\mathbb{P}$(ICA)). Para o método de pseudo verossimilhança, os erros padrão assintóticos (EPA) foram obtidos a partir da matriz de variâncias e covariâncias corrigida $\mathbf{\Sigma}$ definida em (2.12). Para o método de máxima verossimilhança aproximada, os erros padrão assintóticos (EPA) são

os elementos da diagonal principal da inversa da matriz hessiana. Além disso, os pontos de quadratura foram obtidos a partir da função *gauss.quad* do R (R Core Team, 2014), que gera pontos e nós para a quadratura de Hermite. Para o método *naive* os erros padrão assintóticos (EPA) foram obtidos a partir da matriz de variâncias e covariâncias fornecidas pela função *vcov* e para o método calibração da regressão, as variâncias dos estimadores foram obtidas via *boostrap* não-paramétrico.

Para $\boldsymbol{\theta} = (\beta_0, \beta_1, \gamma, \phi)$, a raiz do erro quadrático médio para as N simulações foi obtida a partir da expressão:

$$\sqrt{\text{EQM}(\theta_j)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{\theta}_j^{(i)} - \theta_j)^2}, j = 1, \ldots, 4. \tag{4.35}$$

Os erros padrão empíricos (EPE) são obtidos a partir da raiz quadrada da variância das N simulações que é dada por:

$$\text{Var}(\widehat{\theta}_j) = \frac{1}{N-1}\sum_{i=1}^{N}(\widehat{\theta}_j^{(i)} - \bar{\widehat{\theta}})^2, \quad i = 1, \ldots, N, \tag{4.36}$$

em que $\bar{\widehat{\theta}} = \sum_{i=1}^{N}\widehat{\theta}_i/N$ e $\widehat{\theta}_j^{(i)}$ representam a estimativa de $\widehat{\theta}_j$ na i-ésima simulação.

4.5.1 Regressão Binomial Negativa e Geométrica

Três cenários diferentes são abordados para cada modelo. Nos cenários 1 e 4, verifica-se o desempenho dos métodos de estimação: máxima verossimilhança aproximada, pseudo verossimilhança, *naive* e calibração da regressão. Nos cenários 2 e 5, estudamos a consistência dos estimadores de máxima verossimilhança aproximada e de pseudo verossimilhança e nos cenários 3 e 6, apresentamos as médias das estimativas para cada método a medida que aumentamos os valores da variância do erro.

A estimação dos parâmetros pelo método *naive* é feita utilizando a função *glm.nb*, do pacote *MASS* e a estimação via calibração da regressão, máxima verossimilhança aproximada e pseudo verossimilhança aproximada é feita utilizando a função *optim*, ambas do software R (R Core Team, 2014).

Cenário 1

Neste cenário são apresentados os resultados para um estudo de simulação considerando um tamanho de amostra fixo, $n = 200$, e a seguinte parametrização para a geração dos dados:

- $\varepsilon_1 \sim$ Log-N$(\frac{-\sigma_\varepsilon^2}{2}; \sigma_\varepsilon^2)$ com $\sigma_\varepsilon^2 = 0.15$;
- $\varepsilon_2 \sim$ Log-N$(\frac{-\sigma_\varepsilon^2}{2}; \sigma_\varepsilon^2)$ com $\sigma_\varepsilon^2 = 0.15$;
- $X \sim$ Log-N$(\mu_{x^*}; \sigma_{x^*}^2)$, $\mu_{x^*} = \mu_{w^*} + \frac{\sigma_\varepsilon^2}{2}$ e $\sigma_{x^*}^2 = \sigma_{w^*}^2 - \sigma_\varepsilon^2$, com $\mu_{w^*} = -0.15$ e $\sigma_{w^*}^2 = 0.65$;
- $W_1 = X\varepsilon_1$ e $W_2 = X\varepsilon_2$;
- $Z \sim$ Log-N$(\mu_z; \sigma_z^2)$ com $\mu_z = -0.55$ e $\sigma_z^2 = 0.10$;

Além disso, assumimos apenas uma covariável Z medida sem erro e uma covariável X medida com erro. Os valores atribuídos para ϕ, β_0, β_1 e γ foram:

$$\phi = 0.5; \quad \beta_0 = 1.0; \qquad \beta_1 = -0.5 \quad \text{e} \quad \gamma = 1.5.$$

Avaliando os resultados presentes na Tabela 4.1 destacamos os seguintes pontos:

1. Para o intercepto β_0 e o coeficiente da covariável medida com erro β_1 o método pseudo verossimilhança apresenta médias mais próximas dos valores reais, bem como probabilidades de cobertura mais próximas do valor nominal de 95%. Além disso, os valores das raízes dos erros quadráticos médios, $\sqrt{\text{EQM}}$, as médias dos erros padrão assintóticos (EPA) e os erros padrão empíricos (EPE) estão razoavelmente próximos entre si.

2. Para o coeficiente da covariável medida sem erro, γ, os resultados obtidos para todos os métodos são bons e estão bem próximos. Para os métodos *naive* e máxima verossimilhança as probabilidades de cobertura estão mais próximas de 95%.

3. Para o parâmetro de dispersão, ϕ, os resultados obtidos para todos os métodos também estão bons e bem próximos, sendo um pouco melhor para o método calibração da

Tabela 4.1: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios, $\sqrt{EQM}$, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura de 95% para os intervalos de confiança assintóticos ($\mathbb{P}$(ICA)) considerando um tamanho de amostra de $n = 200$.*

Parâmetros	Métodos	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{P}$(ICA)
β_0	*Naive*	0.9089	0.2411	0.2169	0.2234	0.9250
(1.00)	Calibração	1.0650	0.2413	0.2206	0.2325	0.9290
	Máxima-Veros	1.0848	0.2570	0.2368	0.2427	0.9300
	Pseudo-Veros	1.0522	0.2426	0.2280	0.2371	0.9330
β_1	*Naive*	-0.3708	0.1158	0.0759	0.0845	0.7710
(-0.50)	Calibração	-0.5057	0.1082	0.0836	0.1016	0.9190
	Máxima-Veros	-0.5395	0.1539	0.1229	0.1253	0.9150
	Pseudo-Veros	-0.4950	0.1101	0.1105	0.1176	0.9240
γ	*Naive*	1.5042	0.3075	0.3096	0.3076	0.9470
(1.50)	Calibração	1.4897	0.3188	0.3534	0.3187	0.9240
	Máxima-Veros	1.5047	0.3063	0.3131	0.3064	0.9540
	Pseudo-Veros	1.5034	0.3074	0.3051	0.3076	0.9370
ϕ	*Naive*	0.5035	0.0823	0.3185	0.0811	0.9990
(0.50)	Calibração	0.5045	0.0846	0.0851	0.0845	0.9320
	Máxima-Veros	0.4567	0.4461	0.3887	0.4039	0.9750
	Pseudo-Veros	0.4753	0.0849	0.0811	0.0814	0.9120

regressão. O método máxima verossimilhança obteve erros padrão bem maiores que os demais métodos para o parâmetro ϕ.

Cenário 2

Considerando a mesma parametrização adotada no cenário 1, apresentamos a seguir comparações entre quatro tamanhos de amostra, $n = 50$, $n = 100$, $n = 150$ e $n = 200$, com o interesse em verificar a consistência dos estimadores para o método máxima verossimilhança do modelo de regressão binomial negativa com erro de medida.

Os resultados da Tabela 4.2 mostram que a medida que o tamanho da amostra, n, aumenta as estimativas médias tendem a ficar mais próximas dos valores reais dando um indicativo de consistência dos estimadores. Também houve considerável redução nos valores de erros quadráticos médios e estes estão razoavelmente próximos das médias dos erros padrão assintóticos (EPA) e dos erros padrão empíricos (EPE). Todas as probabilidades de cobertura estão razoavelmente próximas do valor nominal de 95%.

Ainda com a mesma parametrização adotada no cenário 1, apresentamos a seguir

Tabela 4.2: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios,* $\sqrt{EQM}$*, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura de 95% para os intervalos de confiança assintóticos (*$\mathbb{P}$*(ICA)) considerando os tamanhos amostrais* $n = 50$*,* $n = 100$*,* $n = 150$ *e* $n = 200$ *para o método máxima verossimilhança aproximada.*

Parâmetros	n	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{P}$(ICA)
	50	1.1023	0.5099	0.4884	0.4998	0.9520
β_0	100	1.0971	0.3657	0.3381	0.3528	0.9330
(1.00)	150	1.0845	0.2838	0.2727	0.2710	0.9390
	200	1.0847	0.2569	0.2368	0.2427	0.9300
	50	-0.5686	0.2855	0.2616	0.2598	0.9620
β_1	100	-0.5587	0.2109	0.1786	0.1808	0.9350
(-0.50)	150	-0.5413	0.1737	0.1421	0.1478	0.9310
	200	-0.5395	0.1539	0.1229	0.1252	0.9150
	50	1.4934	0.6647	0.6424	0.6650	0.9420
γ	100	1.5017	0.4595	0.4456	0.4598	0.9460
(1.50)	150	1.5033	0.3659	0.3613	0.3661	0.9430
	200	1.5047	0.3063	0.3131	0.3064	0.9540
	50	0.3550	2.3247	1.6068	2.1775	0.9740
ϕ	100	0.4215	0.7666	0.6331	0.6706	0.9820
(0.50)	150	0.4446	0.5376	0.4684	0.4766	0.9870
	200	0.4567	0.4461	0.3887	0.4039	0.9750

comparações entre quatro tamanhos de amostra, $n = 50$, $n = 100$, $n = 150$ e $n = 200$, com o interesse em verificar a consistência dos estimadores para o método pseudo verossimilhança do modelo de regressão binomial negativa com erro de medida.

Tabela 4.3: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios, $\sqrt{EQM}$, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura de 95% para os intervalos de confiança assintóticos ($\mathbb{P}(ICA)$) considerando os tamanhos amostrais $n = 50$, $n = 100$, $n = 150$ e $n = 200$ para o método pseudo verossimilhança.*

Parâmetros	n	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{P}$(ICA)
	50	1.0649	0.4939	0.4568	0.4899	0.9260
β_0	100	1.0605	0.3517	0.3215	0.3467	0.9280
(1.00)	150	1.0597	0.2736	0.2618	0.2671	0.9370
	200	1.0522	0.2426	0.2280	0.2370	0.9330
	50	-0.5345	0.2608	0.2304	0.2468	0.9290
β_1	100	-0.5210	0.1840	0.1575	0.1698	0.9230
(-0.50)	150	-0.5166	0.1545	0.1278	0.1395	0.9280
	200	-0.4950	0.1301	0.1105	0.1176	0.9240
	50	1.4989	0.6694	0.6032	0.6697	0.9030
γ	100	1.5015	0.4593	0.4261	0.4595	0.9290
(1.50)	150	1.5032	0.3657	0.3485	0.3659	0.9270
	200	1.5034	0.3075	0.3051	0.3076	0.9370
	50	0.4391	0.1798	0.1575	0.1693	0.8350
ϕ	100	0.4564	0.1231	0.1117	0.1152	0.8790
(0.50)	150	0.4646	0.0988	0.0922	0.0923	0.9100
	200	0.4753	0.0849	0.0811	0.0813	0.9120

Os resultados da Tabela 4.3 mostram que a medida que o tamanho da amostra, n, aumenta as estimativas médias tendem a ficar mais próximas dos valores reais dando um indicativo de consistência dos estimadores. Também houve considerável redução nos valores de erros quadráticos médios e estes estão razoavelmente próximos das médias dos erros padrão assintóticos (EPA) e dos erros padrão empíricos (EPE). Todas as probabilidades de cobertura estão razoavelmente próximas do valor nominal de 95% e também melhoram com o aumento do tamanho da amostra. Para os coeficientes γ e ϕ as estimativas de erros padrão estão um pouco menores para o método pseudo verossimilhança comparado ao método máxima verossimilhança aproximada.

Cenário 3

Nesta subseção temos interesse em mostrar o que ocorre com as médias das estimativas dos parâmetros do modelo de regressão binomial negativa com erro de medida quando variamos a variância do erro utilizando um tamanho de amostra fixo $n = 200$ para os métodos *naive*, calibração da regressão, máxima verossimilhança aproximada e pseudo verossimilhança. Neste estudo de simulação, utilizamos a mesma parametrização adotada no Cenário 1.

Tabela 4.4: *Médias das estimativas considerando diversos valores para a variância do erro de medida,* σ_ε^2*, para um tamanho amostral de* $n = 200$*.*

Parâmetros	Métodos	$\sigma_\varepsilon^2 = 0.10$	$\sigma_\varepsilon^2 = 0.15$	$\sigma_\varepsilon^2 = 0.20$	$\sigma_\varepsilon^2 = 0.25$
β_0	*Naive*	0.9450	0.9090	0.8961	0.8607
(1.00)	Calibração	1.0392	1.0650	1.0793	1.0810
	Máxima-Veros	1.0534	1.0847	1.1543	1.2400
	Pseudo-Veros	1.0239	1.0522	1.0966	1.1271
β_1	*Naive*	-0.4009	-0.3708	-0.3508	-0.3148
(-0.50)	Calibração	-0.4746	-0.5057	-0.5398	-0.5635
	Máxima-Veros	-0.5063	-0.5395	-0.5958	-0.6713
	Pseudo-Veros	-0.4815	-0.4950	-0.5055	-0.5143
γ	*Naive*	1.4945	1.5042	1.4925	1.4848
(1.50)	Calibração	1.5037	1.4897	1.4971	1.5200
	Máxima-Veros	1.4948	1.5047	1.4932	1.4863
	Pseudo-Veros	1.4952	1.5034	1.4938	1.4900
ϕ	*Naive*	0.5057	0.5139	0.5132	0.5183
(0.50)	Calibração	0.5005	0.5049	0.5138	0.5199
	Máxima-Veros	0.4636	0.4567	0.4386	0.4214
	Pseudo-Veros	0.4840	0.4753	0.4600	0.4520

Pelos resultados apresentados na Tabela 4.4 vemos que a medida que a variância do erro aumenta, as médias das estimativas vão se distanciando dos valores reais para todos os métodos. Para o coeficiente da covariável medida com erro, β_1, o método *naive* apresenta os piores resultados e o método pseudo verossimilhança os melhores. Para o coeficiente da covariável medida sem erro, γ, os quatro métodos apresentam boas estimativas com resultados muito próximos entre si.

Cenário 4

A regressão geométrica pode ser construída a partir da distribuição binomial negativa fazendo o parâmetro $\phi = 1$. Utilizamos a mesma configuração do Cenário 1 exceto por

$\beta_1 = -0.40$, $\mu_{w^*} = -0,35$, $\sigma^2_{w^*} = 0,55$ e $\phi = 1$.

Tabela 4.5: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios, $\sqrt{EQM}$, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura de 95% para os intervalos de confiança assintóticos ($\mathbb{IP}$(ICA)) considerando um tamanho de amostra de $n = 200$.*

Parâmetros	Métodos	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{IP}$(ICA)
β_0	*Naive*	0.9301	0.2922	0.2737	0.2840	0.9280
(1.00)	Calibração	0.9997	0.2745	0.2791	0.2759	0.9700
	Máxima-Veros	1.0955	0.3180	0.3056	0.3035	0.9490
	Pseudo-Veros	1.0408	0.3118	0.2913	0.3093	0.9250
β_1	*Naive*	-0.2939	0.1380	0.1177	0.1263	0.9040
(-0.40)	Calibração	-0.3854	0.1528	0.1052	0.1494	0.9180
	Máxima-Veros	-0.4893	0.1667	0.1471	0.1408	0.9440
	Pseudo-Veros	-0.4195	0.1978	0.1775	0.1854	0.9400
γ	*Naive*	1.5196	0.4111	0.3924	0.4110	0.9310
(1.50)	Calibração	1.5402	0.3950	0.3903	0.3948	0.9200
	Máxima-Veros	1.4695	0.4209	0.4073	0.4200	0.9420
	Pseudo-Veros	1.5169	0.4068	0.3825	0.4066	0.9360
ϕ	*Naive*	0.9997	0.1321	0.1313	0.1322	0.9220
(1.00)	Calibração	0.9758	0.1333	0.1283	0.1311	0.9270
	Máxima-Veros	1.0612	0.1677	0.1506	0.1562	0.9560
	Pseudo-Veros	0.9984	0.1117	0.1481	0.1122	0.9600

Os resultados presentes na Tabela 4.5 mostram que:

1. Para o intercepto β_0 o método calibração da regressão obteve os melhores resultados.

2. Para o coeficiente da covariável medida com erro β_1, para o coeficiente da covariável medida sem erro, γ e para o parâmetro de dispersão, ϕ, o método máxima pseudo verossimilhança apresenta melhor desempenho dado que as médias estão mais próximas dos valores reais, bem como probabilidades de cobertura mais próximas do valor nominal de 95%. Além disso, os valores das raízes dos erros quadráticos médios, $\sqrt{\text{EQM}}$, as médias dos erros padrão assintóticos (EPA) e os erros padrão empíricos (EPE) estão razoavelmente próximos entre si.

Cenário 5

Utilizando a mesma parametrização adotada para a regressão geométrica no cenário 4, apresentamos a seguir comparações entre quatro tamanhos de amostra, $n = 50$, $n = 100$,

$n = 150$ e $n = 200$, com o interesse em verificar a consistência dos estimadores para o método de máxima verossimilhança aproximada.

Tabela 4.6: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios,* $\sqrt{EQM}$*, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura de 95% para os intervalos de confiança assintóticos (*$\mathbb{P}$*(ICA)) considerando os tamanhos amostrais* $n = 50$*,* $n = 100$*,* $n = 150$ *e* $n = 200$ *para o método de* ***máxima verossimilhança aproximada****.*

Parâmetros	Amostras	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{P}$(ICA)
β_0 (1.00)	$n = 50$	1.1433	0.7062	0.6532	0.6919	0.9340
	$n = 100$	1.0892	0.4780	0.4457	0.4699	0.9330
	$n = 150$	1.0873	0.3778	0.3601	0.3678	0.9420
	$n = 200$	1.0661	0.3227	0.3083	0.3160	0.9370
β_1 (-0.40)	$n = 50$	-0.5055	0.5115	0.4550	0.4875	0.9610
	$n = 100$	-0.4756	0.3271	0.2935	0.3022	0.9550
	$n = 150$	-0.4543	0.2552	0.2372	0.2330	0.9550
	$n = 200$	-0.4504	0.2237	0.2009	0.1999	0.9370
γ (1.50)	$n = 50$	1.4349	0.8697	0.8210	0.8677	0.9240
	$n = 100$	1.5007	0.6052	0.5760	0.6055	0.9480
	$n = 150$	1.4827	0.4695	0.4633	0.4694	0.9560
	$n = 200$	1.5169	0.4068	0.3986	0.4067	0.9430
ϕ (1.00)	$n = 50$	1.2186	0.4932	0.3788	0.4423	0.9760
	$n = 100$	1.0853	0.2333	0.2104	0.2172	0.9640
	$n = 150$	1.0558	0.1793	0.1643	0.1705	0.9540
	$n = 200$	1.0451	0.1516	0.1399	0.1449	0.9640

Os resultados da Tabela 4.6 mostram que a medida que o tamanho da amostra aumenta as estimativas médias tendem a ficar mais próximas dos valores reais dando um indicativo de consistência dos estimadores. Também observamos que a medida que o tamanho da amostra aumenta os valores dos erros quadráticos médios, das médias dos erros padrão assintóticos (EPA) e os erros padrão empíricos (EPE) ficam mais próximos entre si. Todas as probabilidades de cobertura estão razoavelmente próximas do valor nominal de 95%.

Apresentamos a seguir comparações entre quatro tamanhos de amostra, $n = 50$, $n = 100$, $n = 150$ e $n = 200$, desta vez com o interesse em verificar a consistência dos estimadores para o método pseudo verossimilhança.

Tabela 4.7: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios,* $\sqrt{EQM}$*, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura de 95% para os intervalos de confiança assintóticos (*$\mathbb{P}$*(ICA)) considerando os tamanhos amostrais* $n = 50$*,* $n = 100$*,* $n = 150$ *e* $n = 200$ *para o método pseudo verossimilhança.*

Parâmetros	Amostras	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{P}$(ICA)
β_0 (1.00)	$n = 50$	1.1051	0.6694	0.5876	0.6614	0.9060
	$n = 100$	1.0604	0.4619	0.4143	0.4582	0.9190
	$n = 150$	1.0608	0.3645	0.3397	0.3595	0.9340
	$n = 200$	1.0408	0.3118	0.2913	0.3093	0.9250
β_1 (-0.40)	$n = 50$	-0.4577	0.4339	0.3705	0.4206	0.9240
	$n = 100$	-0.4073	0.2510	0.2904	0.2447	0.9100
	$n = 150$	-0.4215	0.2260	0.2097	0.2145	0.9370
	$n = 200$	-0.4195	0.1978	0.1775	0.1854	0.9400
γ (1.50)	$n = 50$	1.4347	0.8692	0.7519	0.8672	0.8940
	$n = 100$	1.5006	0.6051	0.5415	0.6055	0.9220
	$n = 150$	1.4824	0.4696	0.4419	0.4695	0.9280
	$n = 200$	1.5169	0.4068	0.3825	0.4066	0.9360
ϕ (1.00)	$n = 50$	0.9042	0.2783	0.2460	0.2614	0.8530
	$n = 100$	0.9587	0.1905	0.1804	0.1861	0.9130
	$n = 150$	0.9725	0.1549	0.1505	0.1524	0.9220
	$n = 200$	0.9984	0.1117	0.1481	0.1122	0.9600

Os resultados da Tabela 4.7 mostram que a medida que o tamanho da amostra aumenta as estimativas médias tendem a ficar mais próximas dos valores reais dando um indicativo de consistência dos estimadores. Também observamos que a medida que o tamanho da amostra aumenta os valores dos erros quadráticos médios, das médias dos erros padrão assintóticos (EPA) e os erros padrão empíricos (EPE) ficam mais próximos entre si. Todas as probabilidades de cobertura estão razoavelmente próximas do valor nominal de 95% e também melhoram com o aumento do tamanho da amostra. Para os coeficientes β_1 e ϕ o método pseudo verossimilhança apresentou resultados melhores que o método máxima verossimilhança aproximada.

Cenário 6

Nesta subseção temos interesse em mostrar o que acontece com as médias das estimativas dos parâmetros do modelo de regressão geométrica com erro de medida quando variamos a variância do erro utilizando um tamanho de amostra de $n = 100$ para os métodos *naive*, calibração da regressão, máxima verossimilhança aproximada e pseudo verossimilhança. Neste estudo de simulação, utilizamos a mesma parametrização adotada no cenário 5. Os resultados se encontram na Tabela 4.8.

Tabela 4.8: *Médias das estimativas considerando diversos valores para a variância do erro de medida,* σ^2_ε*, para um tamanho amostral de* $n = 100$*.*

Parâmetros	Métodos	$\sigma^2_\varepsilon = 0.10$	$\sigma^2_\varepsilon = 0.15$	$\sigma^2_\varepsilon = 0.20$	$\sigma^2_\varepsilon = 0.25$
β_0	*Naive*	0.9861	0.9506	0.9158	0.8979
(1.00)	Calibração	1.0334	1.0579	1.0940	1.1039
	Máxima-Veros	1.0647	1.0892	1.1536	1.2566
	Pseudo-Veros	1.0535	1.0604	1.0718	1.1011
β_1	*Naive*	-0.3166	-0.2983	-0.2731	-0.2371
(-0.40)	Calibração	-0.3976	-0.4400	-0.4546	-0.4724
	Máxima-Veros	-0.4123	-0.4756	-0.5495	-0.6594
	Pseudo-Veros	-0.3946	-0.4073	-0.4200	-0.4655
γ	*Naive*	1.4465	1.4769	1.5006	1.4850
(1.50)	Calibração	1.4972	1.4538	1.4752	1.4556
	Máxima-Veros	1.4518	1.5007	1.4807	1.4861
	Pseudo-Veros	1.4496	1.5006	1.5016	1.4852
ϕ	*Naive*	0.9817	0.9798	0.9857	0.9988
(1.00)	Calibração	0.9822	0.9763	0.9767	0.9817
	Máxima-Veros	1.0826	1.0853	1.1020	1.1441
	Pseudo-Veros	0.9654	0.9587	0.9517	0.9550

Pelos resultados apresentados na Tabela 4.8 vemos que a medida que a variância do erro aumenta, as médias das estimativas vão se distanciando dos valores reais para todos os métodos. Para o intercepto β_0 e para o coeficiente da covariável medida com erro, β_1, o método pseudo verossimilhança obteve os melhores resultados. Para o coeficiente da covariável medida sem erro, γ, os quatro métodos apresentam boas estimativas com resultados muito próximos entre si. Para o parâmetro de dispersão ϕ, o método *naive* apresenta os melhores resultados.

4.5.2 Regressão Binomial Negativa e Poisson

Nesta seção apresentamos os resultados de um estudo de simulação no qual são abordados dois cenários. Para cada cenário foram realizadas $N = 1000$ simulações. A parametrização adotada é a mesma utilizada no Cenário 1 da Seção 4.5.1.

No Cenário 1 apresentamos os resultados do estudo de simulação para os modelos de regressão Poisson e regressão binomial negativa com erro de medida. No Cenário 2, apresentamos as médias de AIC e BIC para as seguintes situações:

a) variável resposta é gerada via distribuição de Poisson mas o modelo é ajustado tanto pela regressão Poisson quanto pela regressão binomial negativa;

b) variável resposta é gerada via distribuição binomial negativa mas o modelo é ajustado em ambos, pela regressão binomial negativa e pela regressão Poisson. Além disso, apresentamos as taxas de seleção dos modelos via AIC e BIC para cada situação levando em consideração todas simulações.

Cenário 1

Neste cenário são apresentados os resultados dos ajustes dos modelos de regressão Poisson e binomial negativa com erro de medida log-normal. Os valores atribuídos para β_0, β_1 e γ foram: $\beta_0 = 1.00$, $\beta_1 = -0.50$, $\gamma = 1.50$ e $\sigma_\varepsilon^2 = 0.10$.

Tabela 4.9: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios, $\sqrt{EQM}$, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura de 95% para os intervalos de confiança assintóticos ($\mathbb{P}$(ICA)) considerando um tamanho de amostra de $n = 100$.*

Modelo	Parâmetros	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{P}$(ICA)
Poisson	β_0	0.9895	0.2911	0.3232	0.2911	0.9360
	β_1	-0.5158	0.1522	0.1512	0.1514	0.9340
	γ	1.4184	0.4106	0.4238	0.4026	0.9210
Binomial Negativa	β_0	1.0368	0.3294	0.3664	0.3275	0.9380
	β_1	-0.5327	0.1501	0.1884	0.1466	0.9440
	γ	1.4816	0.4648	0.4534	0.4647	0.9240
	ϕ	0.4741	0.1168	0.4022	0.1140	0.9450

A partir da Tabela 4.9 vemos que os dois modelos apresentam bom desempenho, pois as estimativas estão próximas dos valores reais e as probabilidades de cobertura estão

razoavelmente próximas do valor nominal de 95%. Os valores da $\sqrt{\text{EQM}}$, EPA e EPE do coeficiente da covariável medida com erro estão mais próximos entre si para o modelo de regressão Poisson.

Cenário 2

Tabela 4.10: *Valores médios de AIC e BIC para 1000 simulações considerando um tamanho de amostra de* $n = 100$.

Modelo Ajustado	Modelo Original			
	Poisson		Binomial Negativa	
	AIC	BIC	AIC	BIC
Poisson	586.81	596.48	761.10	771.52
Binomial Negativa	636.61	647.03	702.09	712.51

Analisando os resultados da Tabela 4.10 vemos que os modelos ajustados estão coerentes com os dados simulados.

O percentual de vezes em que estes mesmos critérios selecionam o modelo correto em 1000 simulações são apresentados na Tabela 4.11.

Tabela 4.11: *Taxas de aceitação de AIC e BIC para 1000 simulações considerando um tamanho de amostra de* $n = 100$.

Modelo Ajustado	Modelo Original	
	Poisson	Binomial Negativa
Poisson	82%	0%
Binomial Negativa	18%	100%

A Tabela 4.11 mostra que em 1000 simulações, os critérios AIC e BIC escolheram corretamente os modelos Poisson e binomial negativa em 82% e 100% das vezes, respectivamente. Pelos resultados do estudo de simulação presente nessa seção vemos que o modelo de regressão binomial negativa apresenta melhor desempenho com relação ao modelo de regressão Poisson.

4.6 Aplicação a Dados Reais

Em um determinado hospital do Estado de São Paulo, 257 pacientes portadores de uma doença na aorta foram submetidos a um procedimento cirúrgico e o número de dias de internação

na UTI foi observado. Também foram medidos alguns fatores pré e pós-operatório. Para o nosso estudo, a variável resposta de interesse é o número de dias de internação na UTI e as covariáveis de interesse associadas a esta variável resposta serão *peso do paciente, última medida de creatina no sangue, presença ou não de diabetes no pré-operatório, se o paciente é fumante ou não e se teve AVC antes da cirurgia*. Por se tratar de uma medida laboratorial, consideramos que a covariável *última medida de creatina no sangue* seja medida com erro. Assim, as covariáveis medidas sem erro serão

- Z_{i1} - peso do paciente no pré-operatório;
- Z_{i2} - paciente foi avaliado com diabetes (ou não) no pré-operatório;
- Z_{i3} - paciente foi classificado como fumante (ou não) no pré-operatório;
- Z_{i4} - paciente teve AVC (ou não) no pré-operatório;

e a covariável observada medida com erro, W_i, será *a última medida de creatina no sangue*. Além disso, temos duas medidas de creatina do mesmo paciente. Uma medida foi avaliada por um técnico de um laboratório A e a outra foi dada por um técnico de um laboratório B, disponibilizando duas replicações desta covariável. As estatísticas descritivas dos dados estão na Tabela 4.12.

Tabela 4.12: *Estatísticas Descritivas.*

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Dias da UTI	0.000	2.000	3.000	4.253	4.000	39.000
Creatina 1	0.500	0.900	1.000	1.126	1.300	2.400
Creatina 2	0.450	0.810	0.900	1.013	1.170	2.160

A Figura 4.1 apresenta a relação entre a variável resposta, número de dias de internação na UTI, e a média da covariável observada, última medida de creatina no sangue.

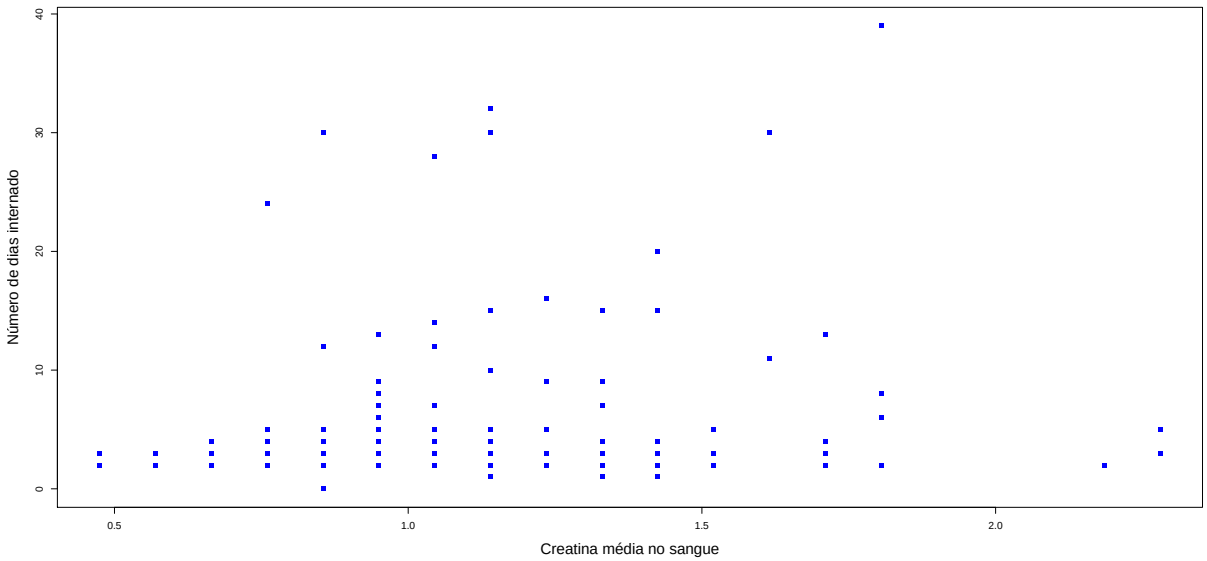

Figura 4.1: Número de dias de internação na UTI *versus* Creatina média no sangue.

Os dados são ajustados via regressão binomial negativa com erro de medida e via Poisson com erro de medida. A estimação dos parâmetros perturbadores, $\mu_{x^*}, \sigma^2_{x^*}$ e σ^2_ε, é feita via equações de estimação (ver Seção 4.2). As análises são feitas utilizando o método pseudo verossimilhança, método que obteve a melhor performance no estudo de simulação apresentado na seção anterior.

4.6.1 Ajuste via modelo de regressão binomial negativa com erro de medida

Os resultados para o modelo de regressão binomial negativa encontram-se na Tabela 4.13.

Tabela 4.13: *Estimativas, Erros Padrão, Intervalos de Confiança de 95% para os parâmetros de interesse.*

Parâmetros	Estimativas	Erros Padrão	IC(95%)
β_0	2.42980	0.28228	(1.87651, 2.98308)
β_1	-0.10194	0.04898	(-0.19795,-0.00594)
γ_1	-0.01205	0.00232	(-0.01660,-0.00750)
γ_2	0.32561	0.15955	(0.01287, 0.63834)
γ_3	-0.00481	0.11973	(-0.23950, 0.22986)
γ_4	-0.25363	0.01745	(-0.28784,-0.21942)
ϕ	0.42666	0.01172	(0.40367, 0.44965)

Analisando os resultados apresentados na Tabela 4.13 observamos que os erros padrão dos estimadores estão razoavelmente pequenos para todos os coeficientes e os intervalos de confiança

não incluem o valor zero exceto para o coeficiente γ_3. Estes resultados mostram que quase todas as covariáveis são significativas a um nível de 5% exceto a covariável Z_{i3} que define se o paciente foi classificado como fumante (ou não) no pré-operatório. Para verificarmos a qualidade desse ajuste realizamos uma análise de resíduos e diagnóstico para o modelo. Os resultados são apresentados a seguir.

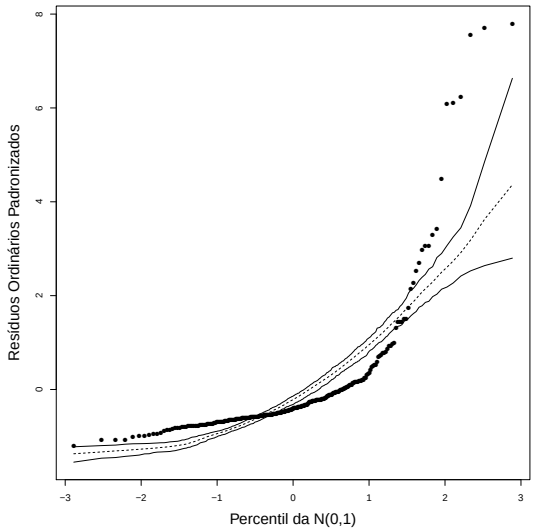

Figura 4.2: Gráfico de envelope quantil normal para os resíduos ordinários padronizados.

A Figura 4.2 mostra que os resíduos não são normalmente distribuídos.

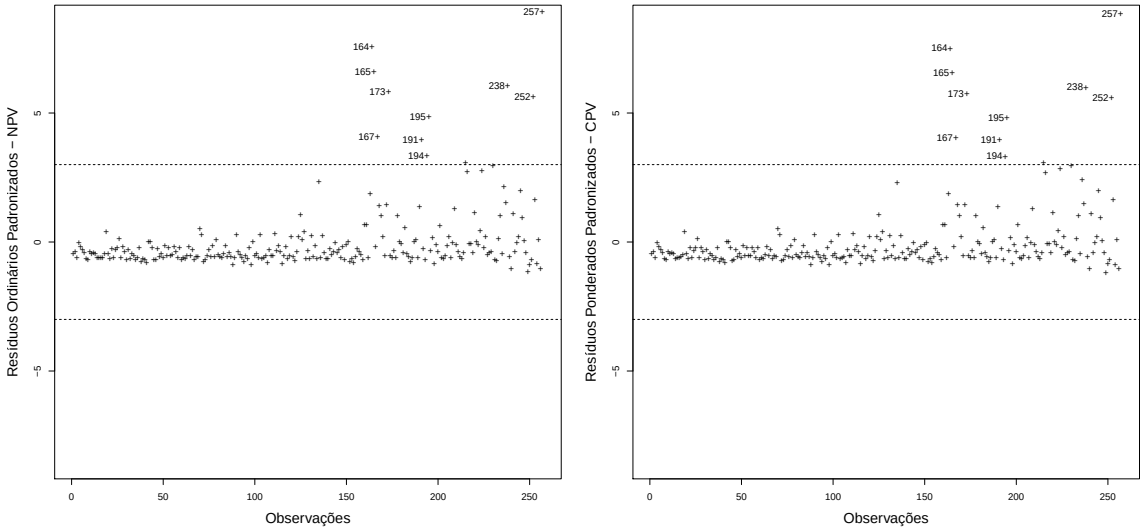

Figura 4.3: Resíduos versus índices das observações para os métodos Calibração Pseudo Verossimilhança (CPV) e *Naive* Pseudo Verossimilhança (NPV).

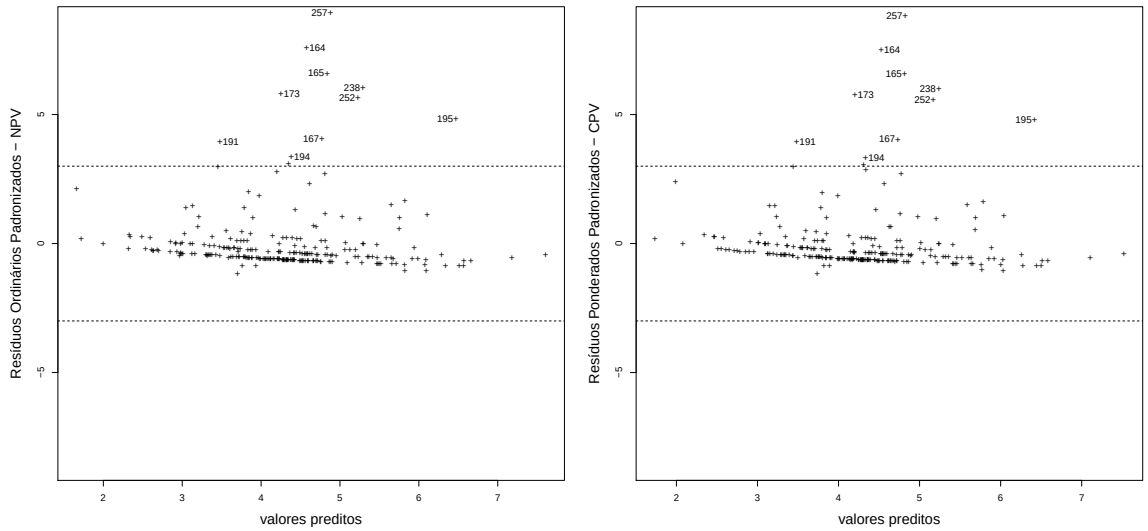

Figura 4.4: Resíduos versus valores preditos para os métodos Calibração Pseudo Verossimilhança (CPV) e *Naive* Pseudo Verossimilhança (NPV).

Analisando os gráficos das Figuras 4.3 e 4.4, observamos que os métodos NPV e CPV para os resíduos ordinários padronizados destacou as observações 164, 165, 167, 173, 191, 194, 195, 238, 252 e 257 como *outliers*, em que um ponto é definido como *outlier* quando se encontra fora do intervalo (-3,3). Esse intervalo foi determinado após estudo de simulação dos resíduos nos dados artificiais apresentados na seção (4.5.1). Antes de decidirmos o que fazer com estes pontos vamos analisar os gráficos de influência local.

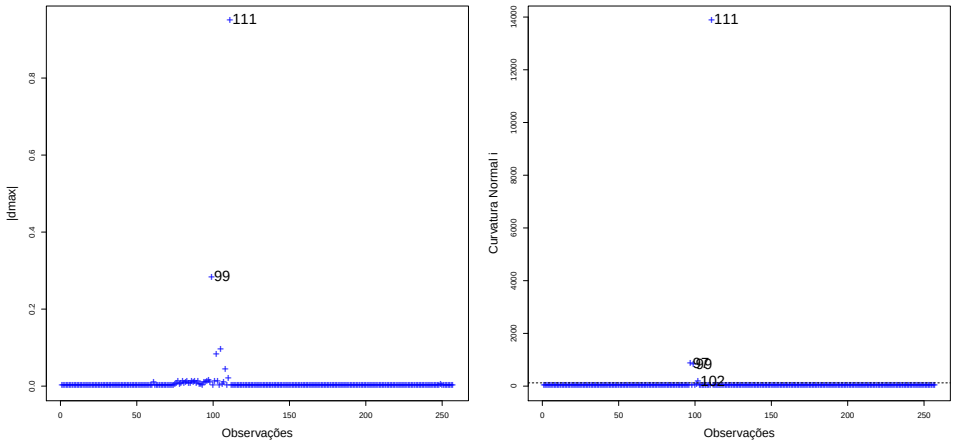

Figura 4.5: Gráficos de *dmax* e C_i pelos índices observados usando o método de pseudo verossimilhança para o esquema de perturbação de casos.

Para os gráficos da Figura 4.5, vemos que para o método pseudo verossimilhança tem-se as

observações 97, 99 e 111 como influentes. Vamos reanalisar os dados retirando dois grupos de observações. Para tal, consideramos duas situações a analisar:

Caso 1: Dados sem as observações: 164, 165 e 257. Observações mais distantes das demais nas Figuras 4.3 e 4.4;

Caso 2: Dados sem as observações: 97, 99 e 111. Observações classificadas como influentes da Figura 4.5;

Tabela 4.14: *Mudanças Relativas (MR), Estimativas dos parâmetros via pseudo verossimilhança, Erros Padrão e Intervalos de Confiança após a remoção das observações discrepantes.*

		Caso 1		
Parâmetros	MR(%)	Estimativas	Erros Padrão	IC(95%)
β_0	15.69	2.04867	0.20037	(1.65593, 2.44140)
β_1	14.88	-0.08677	0.03369	(-0.15282,-0.02072)
γ_1	35.60	-0.00776	0.00255	(-0.01276,-0.00275)
γ_2	30.51	0.42494	0.12828	(0.17349, 0.67639)
γ_3	2714.97	-0.13540	0.07918	(-0.29062, 0.01980)
γ_4	31.30	-0.17697	0.14576	(-0.46267, 0.10873)
ϕ	26.34	0.31429	0.03012	(0.25524, 0.37333)
		Caso 2		
Parâmetros	MR(%)	Estimativas	Erros Padrão	IC(95%)
β_0	0.13	2.43307	1.01886	(0.43609, 4.43005)
β_1	2.18	-0.10416	0.10144	(-0.30300, 0.09467)
γ_1	0.17	-0.01203	0.01145	(-0.03448, 0.01041)
γ_2	2.54	0.33388	0.43875	(-0.52606, 1.19384)
γ_3	145.94	-0.01183	0.59520	(-1.17843, 1.15476)
γ_4	7.40	-0.27239	0.26225	(-0.78642, 0.24162)
ϕ	1.60	0.43351	0.30349	(-0.16132, 1.02835)

Vamos analisar os resultados da Tabela 4.14 por casos. Analisando o caso 1 vemos que os erros padrão diminuíram para todos os coeficientes exceto para os coeficientes γ_1, γ_3 e γ_4. Os comprimentos dos intervalos ficaram menores para quase todos os coeficientes exceto para γ_1 e ϕ. As porcentagens de mudanças relativas (MR(%)) foram pequenas para todos os coeficientes exceto o coeficiente γ_3. Os intervalos de confiança para os parâmetros β_1, γ_1, γ_3 e γ_4 incluem o valor zero, ou seja, são não significativos para o modelo. Portanto, a retirada do grupo de observações no caso 1 causa bastante impacto nos resultados do ajuste do modelo e aparentemente não de forma favorável.

Para o caso 2, os erros padrão aumentaram para todos os coeficientes bem como os comprimentos dos intervalos de confiança comparando aos resultados obtidos com toda a amostra expostos na Tabela 4.13. Apenas o coeficiente β_0 possui intervalo de confiança que não inclui o valor zero. As porcentagens de mudanças relativas, MR(%), foram pequenas para todos os coeficientes exceto para o coeficiente γ_3. A covariável Z_{i3}, que define se o paciente foi classificado como fumante (ou não) no pré-operatório, apresentou-se não significativa em todas as situações, caso 1, caso 2 e análise com todos os dados (ver Tabela 4.13), além de altas taxas de mudança relativa para os casos 1 e 2, logo, temos fortes indicativos que esta covariável pode ser retirada do modelo. A seguir apresentamos os resultados do ajuste para o modelo de regressão Poisson.

4.6.2 Ajuste via modelo de regressão Poisson com erro de medida

Para o modelo de regressão Poisson os resultados encontram-se na Tabela 4.15.

Tabela 4.15: *Estimativas, Erros Padrão, Intervalos de Confiança de 95% para os parâmetros de interesse.*

Parâmetros	Estimativas	Erros Padrão	IC(95%)
β_0	2.54921	0.42008	(1.72583, 3.37258)
β_1	-0.21141	0.09630	(-0.40017,-0.02266)
γ_1	-0.01162	0.00444	(-0.02033,-0.00292)
γ_2	0.32167	0.21789	(-0.10540, 0.74875)
γ_3	-0.02566	0.17639	(-0.37140, 0.32007)
γ_4	-0.28310	0.18447	(-0.64466, 0.07846)

Analisando os resultados apresentados na Tabela 4.15 observamos que os erros padrão dos estimadores estão razoavelmente pequenos para todos os coeficientes mas são maiores do que os obtidos via regressão binomial negativa com erro de medida. Os intervalos de confiança não incluem o valor zero para os coeficientes β_0 e β_1, coeficiente da covariável medida com erro, mas incluem para γ_2, γ_3 e γ_4. Estes resultados tornam as covariáveis Z_{i2}, Z_{i3} e Z_{i4} não significativas a um nível de 5%. Para verificarmos a qualidade desse ajuste realizamos uma análise de resíduos e diagnóstico para o modelo. A Figura 4.6 mostra que os resíduos não são normalmente distribuídos.

Analisando os gráficos das Figuras 4.7 e 4.8, observamos que os métodos NPV e CPV para

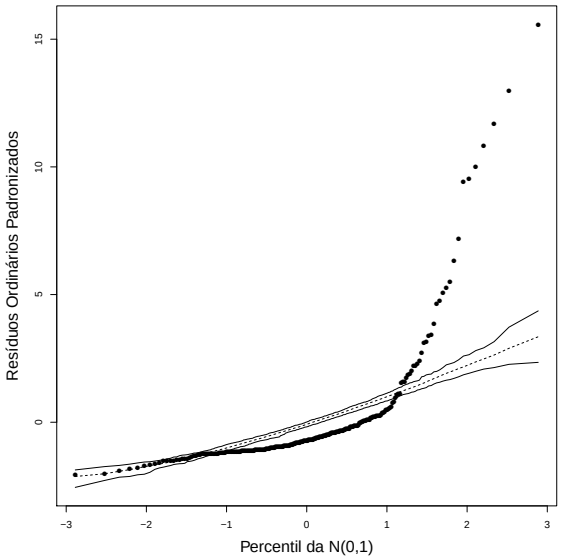

Figura 4.6: Gráfico de envelope quantil normal para os resíduos ordinários padronizados.

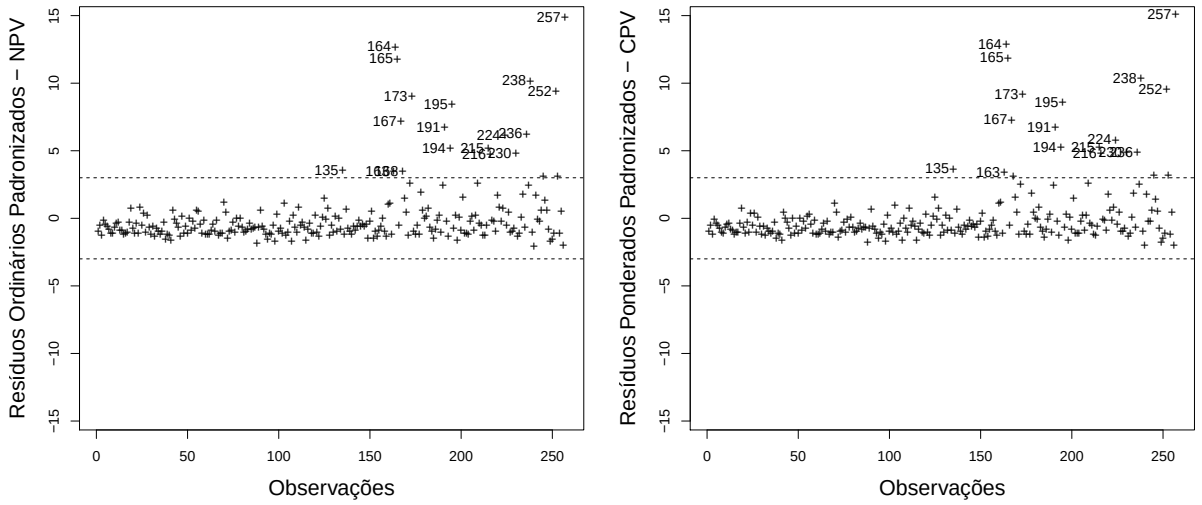

Figura 4.7: Resíduos versus índices das observações para os métodos Calibração Pseudo Verossimilhança (CPV) e *Naive* Pseudo Verossimilhança (NPV).

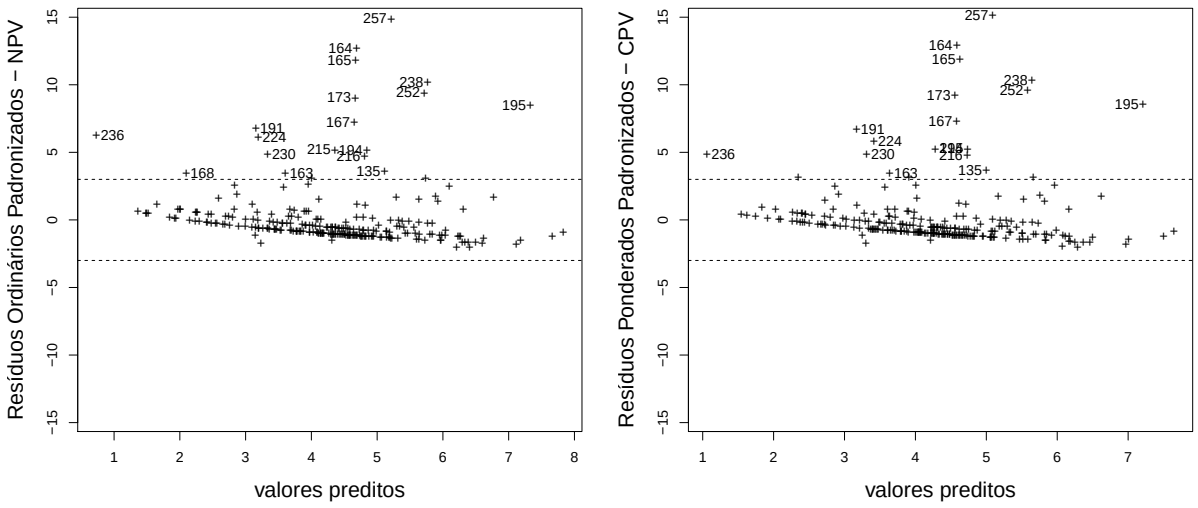

Figura 4.8: Resíduos versus valores preditos para os métodos Calibração Pseudo Verossimilhança (CPV) e *Naive* Pseudo Verossimilhança (NPV).

os resíduos ordinários padronizados destacou as observações 135, 163,164, 165, 167, 173, 191, 194, 195, 215, 216, 224, 230, 236, 238, 252 e 257 como *outliers*.

Antes de decidirmos o que fazer com estes pontos vamos analisar os gráficos de influência local.

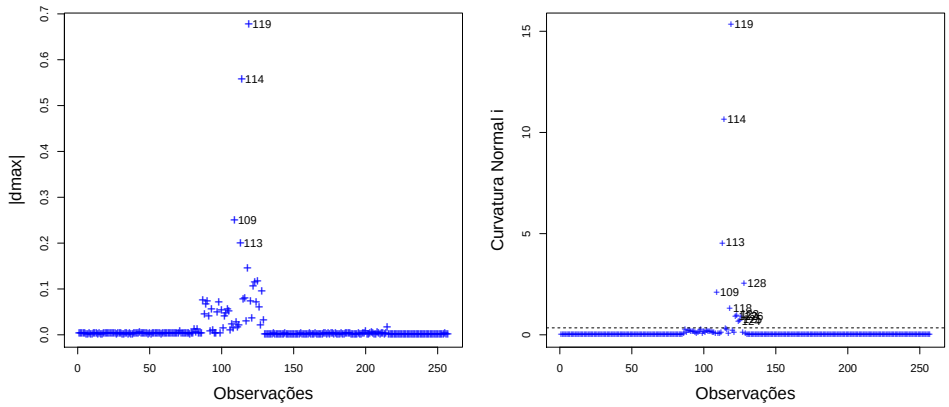

Figura 4.9: Gráficos de *dmax* e C_i pelos índices observados usando o método de pseudo verossimilhança para o esquema de perturbação de casos.

Para os gráficos da Figura 4.9, vemos que para o método pseudo verossilhança tem-se as observações 109, 113, 114, 119 e 128 como influentes. Consideramos duas situações a analisar:

Caso 1: Dados sem as observações: 164, 165 e 257. Observações mais distantes das demais nas

Figuras 4.7 e 4.8;

Caso 2: Dados sem as observações: 114 e 119. Observações classificadas como influentes na Figura 4.9 e estão mais afastadas das demais;

A Tabela 4.16 apresenta os valores de mudança relativa (definida em (2.18)) e os resultados das estimativas sem as observações especificadas para os casos 1 e 2.

Tabela 4.16: *Mudanças Relativas (MR), Estimativas dos parâmetros via pseudo verossimilhança, Erros Padrão e Intervalos de Confiança após a remoção das observações discrepantes.*

		Caso 1		
Parâmetros	MR(%)	Estimativas	Erros Padrão	IC(95%)
β_0	21.63	1.90420	0.32358	(1.26997, 2.53844)
β_1	100.92	0.00097	0.05478	(-0.10640, 0.10833)
γ_1	37.26	-0.00756	0.00402	(-0.01545, 0.00033)
γ_2	26.59	0.41220	0.20861	(0.00330, 0.82109)
γ_3	2586.47	-0.12922	0.15267	(-0.42847, 0.17003)
γ_4	23.84	-0.19316	0.17309	(-0.53242, 0.14610)
		Caso 2		
Parâmetros	MR(%)	Estimativas	Erros Padrão	IC(95%)
β_0	5.45	2.29743	0.52550	(1.26744, 3.32741)
β_1	84.03	-0.01628	0.37590	(-0.75306, 0.72049)
γ_1	1.49	-0.01187	0.00507	(-0.02182,-0.00192)
γ_2	4.08	0.28475	0.21505	(-0.13674, 0.70626)
γ_3	183.58	0.00402	0.20636	(-0.40045, 0.40850)
γ_4	10.24	-0.27961	0.18577	(-0.64373, 0.08449)

Vamos analisar os resultados da Tabela 4.16 por casos. Analisando o caso 1 vemos que os erros padrão aumentaram para todos os coeficientes bem como os comprimentos dos intervalos de confiança comparando aos resultados obtidos com toda a amostra apresentados na Tabela 4.13. As porcentagens de mudanças relativas foram razoavelmente altas para todos os coeficientes sendo maiores para os coeficientes β_1 e γ_3. Todos os intervalos de confiança incluem o valor zero exceto para os coeficientes β_0 e γ_2. Portanto, a retirada do grupo de observações no caso 1 causa bastante impacto nos resultados do ajuste do modelo mas não de forma favorável.

Para o caso 2, os erros padrão para todos os coeficientes aumentaram, assim como os comprimentos dos intervalos de confiança comparando aos resultados obtidos com toda a amostra (ver Tabela 4.13). Apenas os coeficientes β_0 e γ_1 possuem intervalos de confiança

que não incluem o valor zero. As porcentagens de mudanças relativas foram pequenas para todos os coeficientes exceto para os coeficientes β_1 e γ_3. Portanto, a retirada do grupo de observações no caso 2 também causa bastante impacto nos resultados do ajuste do modelo mas não de forma favorável.

Ajustamos os dados reais para os dois modelos, Poisson e Binomial Negativa. Além das análises apresentadas nas Subseções 4.6.1 e 4.6.2 utilizamos os critérios de seleção de modelos AIC e BIC para decidir qual modelo melhor se ajusta aos dados. Os resultados estão na Tabela 4.17.

Tabela 4.17: *Valores médios de AIC e BIC para os dados reais.*

Modelo	AIC	BIC
Poisson	2576.67	2590.86
Binomial Negativa	2265.69	2279.89

Analisando os resultados de AIC e BIC presentes na Tabela 4.17, vemos que os critérios apontam o modelo de regressão binomial negativa com erro de medida como mais adequado aos dados reais. Este resultado já era esperado, visto que o modelo de regressão binomial negativa apresentou um melhor desempenho no ajuste comparado ao modelo Poisson.

Neste capítulo apresentamos a construção, estimação e análise de diagnóstico para um modelo de regressão cuja variável resposta é discreta e o erro de medida é multiplicativo. No próximo capítulo apresentamos a construção, estimação e análise de diagnósticos para um modelo de regressão cuja variável resposta é contínua e o erro de medida é multiplicativo.

Capítulo

5

Modelo de Regressão Beta com Erro de Medida Multiplicativo

A construção do modelo de regressão beta com erro de medida log-normal foi motivado por problemas que ocorrem frequentemente na área financeira. Instituições financeiras têm a necessidade de determinar a proporção de gasto do limite a ser oferecido no cheque especial para futuros clientes. O conhecimento desta proporção permite à instituição alocar um montante de capital para fazer frente a um possível risco de inadimplência, além de ter a possibilidade de calcular o lucro futuro oriundo dos juros que serão pagos. Um dado extremamente importante para a determinação desta proporção é a renda do cliente. Por se tratar de clientes ainda não correntistas, informações sobre a renda real podem não estar disponíveis no cadastro. Por essa razão, a instituição obtém no mercado uma renda presumida do cliente, ou seja, uma renda medida com erro. Nesse contexto, a variável explicativa X representaria a renda real do potencial cliente não disponível para a instituição, uma vez que esse cliente ainda não é correntista, e W a variável explicativa realmente observada, representaria a renda presumida desse cliente através de um modelo de renda presumida disponibilizado no mercado. Inspirados por essa ideia sugerimos assumir para a covariável X e para o erro de medida ε a distribuição log-normal que é uma distribuição estritamente positiva além de oferecer boas propriedades

aritméticas.

Neste capítulo, apresentamos o modelo de regressão beta com erro de medida multiplicativo log-normal. Alguns métodos de estimação são estudados. Tais métodos têm como princípio a estimação por máxima verossimilhança e pseudo verossimilhança. Estimamos os parâmetros perturbadores usando dados replicados. Um estudo de simulação foi feito para ilustrar os resultados das estimações para cada método. Além do estudo de simulação, realizamos uma análise a um conjunto de dados reais e análise de diagnóstico.

5.1 Introdução

Modelos de regressão são ferramentas estatísticas que relacionam o valor médio da variável de interesse (variável resposta) a uma ou mais covariáveis (variáveis explicativas). Quando a variável de interesse assume valores no intervalo unitário $(0, 1)$, tais como proporções e taxas, o modelo de regressão beta é uma boa opção para modelar tal variável. Como descrito anteriormente, assumiremos que a estrutura que relaciona a covariável observada W_i com a verdadeira covariável não-observada X_i é multiplicativa, $W_i = X_i\varepsilon_i$, $i = 1, \ldots, n$, em que os erros, ε_i, são independentes e identicamente distribuídos. Além disso, assumimos para a covariável não observada X_i e o erro associado ε_i, $i = 1, \ldots, n$, distribuição log-normal.

Na literatura, há várias propostas de trabalhos envolvendo o modelo de regressão beta sem considerar erro nas covariáveis, alguns destes: Ferrari e Cribari-Neto (2004), Ospina *et al.* (2006), Espinheira *et al.* (2008a), Espinheira *et al.* (2008b) e Cribari-Neto e Zeileis (2010). Para modelos de regressão com variável resposta contínua e com erro nas covariáveis, temos Fuller (1987), Carroll *et al.* (2006), Cheng e Van Ness (1999) e Buonaccorsi (2010). Para o modelo de regressão beta com erro de medida aditivo tem-se como referência Carrasco *et al.* (2014). Além de considerar o erro de medida aditivo, o trabalho de Carrasco *et al.* (2014) consideram que as covariáveis e erro de medida são normalmente distribuídos e estimam os parâmetros perturbadores em dois processos de maximização no método de pseudo verossimilhança. No nosso trabalho os parâmetros perturbadores são estimados via solução de equações de estimação simples e para as covariáveis e erro de medida do modelo assumimos distribuição log-normal. A escolha da distribuição log-normal para as covariáveis e erro de

medida está diretamente relacionada ao comportamento das variáveis explicativas dos dados reais utilizados nesse capítulo.

O objetivo principal deste capítulo é apresentar e comparar quatro diferentes métodos de estimação para o modelo de regressão beta com erro de medida multiplicativo. O trabalho está dividido em 9 seções. A descrição do modelo de regressão beta com erro de medida multiplicativo é feita na Seção 5.2. A introdução ao erro de medida log-normal e suas propriedades, além da estimação dos parâmetros perturbadores são apresentados na Seção 5.3. Os quatro diferentes métodos de estimação, *naive*, calibração, máxima verossimilhança aproximada e pseudo verossimilhança aproximada, são abordados na Seção 5.4. A análise de diagnósticos é apresentada na Seção 5.5. Um estudo de simulação é apresentado na Seção 5.6. Na Seção 5.7, uma aplicação a um conjunto de dados reais e, finalmente, a conclusão é descrita na Seção 5.8.

5.2 Modelo

Considere um modelo de regressão no qual as variáveis respostas independentes $Y_1, \ldots, Y_n$, com suporte no intervalo $(0, 1)$, estejam associadas a vetores de covariáveis medidas sem erro $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$, $i = 1, \ldots, n$. Além disso, cada variável resposta Y_i, também está associada a uma única covariável positiva medida com erro X_i, $i = 1, \ldots, n$. Com base no suporte da variável resposta, neste trabalho supomos que $Y_i|X_i, \mathbf{z}_i$ segue uma distribuição Beta(p_i, q_i), $i = 1, \ldots, n$. Reparametrizando a distribuição de $Y_i|X_i, \mathbf{z}_i$ pela sua média (Ferrari e Cribari-Neto , 2004), é possível determinar, através do uso de uma função de ligação, uma relação funcional entre os parâmetros da distribuição beta e os dados. Sejam $\mu_i = p_i/(p_i + q_i)$ e $\phi = p_i + q_i$, em que ϕ é fixo. Assim, $I\!E(Y_i|X_i, \mathbf{z}_i) = \mu_i$ e $\text{Var}(Y_i|X_i, \mathbf{z}_i) = \mu_i(1 - \mu_i)/(1 + \phi)$, em que μ_i é um parâmetro de locação enquanto ϕ pode ser considerado um parâmetro de precisão. Consequentemente, $Y_i|X_i, \mathbf{z}_i \sim \text{Beta}\,(\mu_i\phi, (1 - \mu_i)\phi)$. A função densidade para $Y_i|X_i, \mathbf{z}_i$ é dada por

$$f_{Y_i|X_i,\mathbf{z}_i}(y_i; \mu_i, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma((1 - \mu_i)\phi)} y_i^{\mu_i\phi-1}(1 - y_i)^{(1-\mu_i)\phi-1}, \quad 0 < y_i < 1, \tag{5.1}$$

em que $\phi > 0$ e $0 < \mu_i < 1$, $i = 1, \ldots, n$. Para relacionar a variável resposta com as covariáveis utilizamos a função de ligação logito:

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 X_i + \boldsymbol{\gamma}^\top \mathbf{z}_i = \eta_i,$$

$$\mu_i = \frac{\exp(\beta_0 + \beta_1 X_i + \boldsymbol{\gamma}^\top \mathbf{z}_i)}{1 + \exp(\beta_0 + \beta_1 X_i + \boldsymbol{\gamma}^\top \mathbf{z}_i)}, \quad i = 1, \ldots, n, \tag{5.2}$$

em que $\beta_0 \in \mathbb{R}$ é o intercepto, $\beta_1 \in \mathbb{R}$ é o coeficiente da covariável X medida com erro, $\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_k)$, $\boldsymbol{\gamma}^\top \in \mathbb{R}^k$, é o vetor de coeficientes do vetor de covariáveis medidas sem erro e $g(\cdot) : (0,1) \to \mathbb{R}$ é uma função de ligação estritamente monótona e duplamente diferenciável. Note que $\mu_i = g^{-1}(\eta_i)$ e $\mathrm{Var}(Y_i|X_i, \mathbf{z}_i) = g^{-1}(\eta_i)(1 - g^{-1}(\eta_i))/(1+\phi)$, ou seja, mesmo o parâmetro de dispersão ϕ sendo constante as variâncias não serão constantes, pois dependem de μ_i (Ospina *et al.*, 2006). Outras funções de ligação poderiam ser utilizadas como por exemplo a probito, $g(\mu_i) = \Phi(\eta_i)$, mas a logito é mais fácil de ser interpretada.

5.3 Métodos de Estimação

Os métodos de estimação propostos nesta seção também são baseados no princípio da máxima verossimilhança e pseudo verossimilhança, já apresentados nos capítulos 2 e 3.

5.3.1 Método *Naive*

O método *naive* consiste em substituir na expressão (5.1) o valor da covariável não observável, X_i, pelo valor observado com erro, W_i, no modelo de regressão beta e, em seguida, realizar a regressão pelo método usual. Desta forma, ao invés de $Y_i|X_i, \mathbf{z}_i$ teremos $Y_i|W_i, \mathbf{z}_i$ e a função log-verossimilhança será dada por:

$$l(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} l_i(\mu_i, \phi), \quad i = 1, \ldots, n, \tag{5.3}$$

em que

$$l_i(\mu_i,\phi) = \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) + (\mu_i\phi - 1)\log y_i + \{(1-\mu_i)\phi - 1\}\log(1-y_i),$$

e

$$\mu_i = \frac{\exp(\beta_0 + \beta_1 W_i + \boldsymbol{\gamma}^\top \mathbf{z}_i)}{1 + \exp(\beta_0 + \beta_1 W_i + \boldsymbol{\gamma}^\top \mathbf{z}_i)}, \quad i = 1, \ldots, n. \tag{5.4}$$

No software R (R Core Team, 2014) a estimação dos parâmetros pode ser feita através do pacote *betareg* (Cribari-Neto e Zeileis, 2010). É importante ressaltar que na presença de replicações, W_i é substituído por $\overline{W}_i = \frac{\sum_{j=1}^{J} W_{ij}}{J_i}, i = 1, \ldots, n$. Para a construção de intervalos de confiança para os parâmetros é possível utilizar a função *vcov* do pacote *betareg* que fornece a matriz de variâncias e covariâncias dos estimadores.

5.3.2 Método Calibração da Regressão

O método calibração da regressão consiste em substituir na expressão (5.1) o valor da variável não observável, X_i, pela esperança condicional, $\mathbb{E}(X_i|W_i)$, ou seja,

$$\mu_i = \frac{\exp(\beta_0 + \beta_1 \mathbb{E}(X_i|W_i) + \boldsymbol{\gamma}^\top \mathbf{z}_i)}{1 + \exp(\beta_0 + \beta_1 \mathbb{E}(X_i|W_i) + \boldsymbol{\gamma}^\top \mathbf{z}_i)}, \quad i = 1, \ldots, n. \tag{5.5}$$

A função de log-verossimilhança será idêntica a expressão (5.3) com μ_i definido por (5.5). E a esperança condicional, $\mathbb{E}(X_i|W_i)$, é a mesma definida no capítulo 3, expressão (4.12). Para a construção de intervalos de confiança para os parâmetros de interesse é necessário calcularmos os erros padrão dos seus respectivos estimadores. Estimamos via *bootstrap* não-paramétrico.

5.3.3 Método de Máxima Verossimilhança

A função de verossimilhança do modelo de regressão beta com erro de medida multiplicativo é idêntica a expressão (4.18), definida no capítulo 3, exceto pela função densidade de $Y_i|X_i, \mathbf{z}_i$ que é dada pela equação (5.1).

Desta forma, dado os vetores de observações $\mathbf{y} = (y_1, ..., y_n)$, $\mathbf{z}_i = (z_{i1}, ..., z_{in})$ e $\mathbf{w} = (w_1, ..., w_n)$, as distribuições assumidas para X e para ε são as mesmas definidas na Seção

4.2, a função de log-verossimilhança para o modelo de regressão beta com erro de medida multiplicativo pode ser escrita como:

$$l(\boldsymbol{\zeta};\mathbf{y},\mathbf{w}) = \sum_{i=1}^{n} \log \int_0^{\infty} \frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma((1-\mu_i)\phi)} y_i^{\mu_i\phi-1}(1-y_i)^{(1-\mu_i)\phi-1} \times \frac{1}{w_i\sigma_\varepsilon\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left\{\log(w_i)-\log(x_i)+\frac{\sigma_\varepsilon^2}{2}\right\}^2\right) \times \frac{1}{x_i\sigma_{x^*}\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_{x^*}^2}\left\{\log(x_i)-\mu_{x^*}\right\}^2\right) dx_i, \tag{5.6}$$

em que $\boldsymbol{\zeta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi, \mu_{x^*}, \sigma_{x^*}^2)$ é o vetor de parâmetros desconhecidos a serem estimados e a variância do erro, σ_ε^2, é supostamente conhecida ou previamente estimada via dados replicados. Utilizando as mesmas mudanças de variáveis definidas no capítulo 3, a integral em (5.6) pode ser resolvida utilizando o método da Quadratura de Gauss-Hermite (Abramowitz e Stegun, 1964) e a função de log-verossimilhança, $l(\boldsymbol{\zeta};\mathbf{y},\mathbf{w})$, dada em (5.6), agora é denotada por $l_H(\boldsymbol{\zeta};\mathbf{y},\mathbf{w})$, pois trata-se de uma função log-verossimilhança que depende de pontos e nós da Quadratura de Hermite, ou seja:

$$l_H(\boldsymbol{\zeta};\mathbf{y},\mathbf{w}) = \sum_{i=1}^{n} \log\left(\sum_{q=1}^{Q} \frac{1}{\sqrt{\pi}} p_q f_{Y_i,W_i|X_{iq}}\left(y_i, w_i \middle| x_{iq}\right)\right), \tag{5.7}$$

em que $x_{iq} = \exp(r_q\sigma_{x^*}\sqrt{2} + \mu_{x^*})$, Q é o número de pontos da quadratura, $r_1, \ldots, r_Q$ são as raízes do polinômio ortogonal de Hermite $H_q(x)$ (ou nós da quadratura) e $p_1, \ldots, p_Q$ são os pesos dados por (4.24).

Estimação por *naive* máxima verossimilhança (NMV)

Substituindo no preditor linear da regressão dado em (5.4), os valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}^\top$ estimados via máxima verossimilhança aproximada, ou seja, obtidos a partir da maximização da expressão (5.7), temos

$$\widehat{\mu}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 W_i + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 W_i + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}, \quad i = 1, \ldots, n.$$

Estimação por calibração máxima verossimilhança (CMV)

Substituindo no preditor linear da regressão dado em (5.5), os valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}^\top$ estimados via máxima verossimilhança aproximada, ou seja, obtidos a partir da maximização da expressão (5.7), temos

$$\widehat{\mu}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 \widehat{\mathbb{E}}(X_i|W_i) + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 \widehat{\mathbb{E}}(X_i|W_i) + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}, \quad i = 1, \ldots, n. \tag{5.8}$$

5.3.4 Método de Pseudo Verossimilhança

Como vimos nos capítulos anteriores, a principal diferença entre os métodos de máxima verossimilhança e pseudo verossimilhança é que o vetor de parâmetros perturbadores $\boldsymbol{\delta} = (\sigma^2_\varepsilon, \mu_{x^*}, \sigma^2_{x^*})$ é previamente estimado via equações de estimação definidas em (4.6). Estimar os parâmetros perturbadores previamente é uma forma efetiva de reduzir o número de parâmetros a serem estimados. Assim, a função de log pseudo verossimilhança será idêntica a expressão (5.7), em que $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\gamma}^\top, \phi)$ serão os únicos parâmetros a serem estimados dado o vetor estimado $\widehat{\boldsymbol{\delta}}$, ou seja,

$$l_H(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} \log \left(\sum_{q=1}^{Q} \frac{1}{\sqrt{\pi}} p_q f_{Y_i, W_i|X_{iq}} \left(y_i, w_i | \, x_{iq} \right) \right), \tag{5.9}$$

em que $x_{iq} = \exp(r_q \widehat{\sigma}_{x^*} \sqrt{2} + \widehat{\mu}_{x^*})$. Usando o método de equações de estimação, o vetor $\widehat{\boldsymbol{\theta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\boldsymbol{\gamma}}^\top, \widehat{\phi})^\top$ é a solução da equação de estimação (4.28), com $\Psi_{j,i}(\tilde{Y}_i, \theta_j, \boldsymbol{\delta}) = \partial l_H(\boldsymbol{\theta}; \boldsymbol{\delta}) / \partial \theta_j$, $j = 1, \ldots, 4$, e $l_H(\boldsymbol{\theta}; \boldsymbol{\delta})$ é a função log-pseudo verossimilhança definida em (5.9). A distribuição assintótica dos estimadores de pseudo verossimilhança para o modelo de regressão beta com erro de medida multiplicativo é a mesma já definida em (2.2.4).

Estimação por *naive* pseudo verossimilhança (NPV)

Substituindo no preditor linear da regressão dado em (5.4), os valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}^\top$ estimados via método pseudo verossimilhança aproximada, ou seja, obtidos a partir da

maximização da expressão (5.9), temos

$$\widehat{\mu}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 W_i + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 W_i + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}, \quad i = 1, \ldots, n.$$

Estimação por calibração pseudo verossimilhança (CPV)

Substituindo no preditor linear da regressão dado em (5.5), os valores dos coeficientes β_0, β_1 e $\boldsymbol{\gamma}^\top$ estimados via método pseudo verossimilhança aproximada, ou seja, obtidos a partir da maximização da expressão (5.9), temos

$$\widehat{\mu}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 \widehat{I\!E}(X_i|W_i) + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 \widehat{I\!E}(X_i|W_i) + \widehat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}, \quad i = 1, \ldots, n, \tag{5.10}$$

em que $\widehat{I\!E}(X_i|W_i)$ já foi definida em (4.12).

5.4 Análise de Diagnóstico

No Capítulo 2, Seção 2.3 explicamos o que representa no ajuste do modelo de regressão uma análise de diagnósticos. Nesta seção mostramos de forma mais sucinta os principais resultados para o modelo de regressão beta com erro de medida multiplicativo.

Na literatura, há diversos trabalhos envolvendo análise de diagnósticos. Para o modelo de regressão beta, podemos citar as seguintes referências: Ferrari e Cribari-Neto (2004), Espinheira *et al.* (2008b), Espinheira *et al.* (2008a) e Ferrari *et al.* (2011). Enquanto que para modelos com erro de medida, temos: Kelly (1984), Miller (1990), Carroll e Spiegelman (1992), Zhao *et al.* (1994), Zhao e Lee (1995), de Castro *et al.* (2007), Xie e Wei (2009) e Galea e de Castro (2012).

5.4.1 Análise de Resíduos

Nesta seção, apresentamos os resíduos ordinários e ponderados padronizados para o modelo de regressão beta com erro de medida multiplicativo usando os métodos de estimação *naive*, calibração da regressão, máxima verossimilhança aproximada e pseudo verossimilhança. Os resíduos ordinários padronizados já foram definidos no Capítulo 2, expressão (2.13).

Para o método *naive*, $\hat{\mu}_i$ é dado por

$$\hat{\mu}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)},$$

em que $\hat{\phi}$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ e $\hat{\boldsymbol{\gamma}}^\top$ são as estimativas de máxima verossimilhança de ϕ, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}^\top$, respectivamente, obtidas via maximização da função de log-verossimilhança dada em (5.3) e $\widehat{\text{var}}(y_i) = \hat{\mu}_i(1-\hat{\mu}_i)/(1+\hat{\phi})$. Para o método de calibração da regressão $\hat{\mu}_i$ é dado por

$$\hat{\mu}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \widehat{I\!E}(X_i|W_i) + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \widehat{I\!E}(X_i|W_i) + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)},$$

em que $\widehat{I\!E}(X_i|W_i)$ foi definido na expressão (4.12). No caso de replicação das covariáveis W_i, $i = 1, \ldots, n$, W_i é substituído por $\overline{W}_i$, $i = 1, \ldots, n$.

Utilizando os resíduos propostos por Ferrari *et al.* (2011) definimos os resíduos ponderados padronizados para o modelo de regressão beta com erro de medida multiplicativo e parâmetro de dispersão constante por

$$r_i^{pp} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\widehat{\upsilon}_i(1-\hat{h}_{ii}^*)}}, \quad i = 1, \ldots, n. \tag{5.11}$$

em que $y_i^* = \log(y_i/(1-y_i))$, para o método *naive*, $\hat{\mu}_i^*$, $\hat{\upsilon}_i$ and $\hat{h}_{ii}^*$ são dados respectivamente por

$$\begin{aligned}
\widehat{\mu}_i^* &= \psi\left(\frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}\widehat{\phi}\right) - \psi\left(\frac{1}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}\widehat{\phi}\right), \\
\widehat{\upsilon}_i &= \psi'\left(\frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}\widehat{\phi}\right) + \psi'\left(\frac{1}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\boldsymbol{\gamma}}^\top \mathbf{z}_i)}\widehat{\phi}\right),
\end{aligned} \tag{5.12}$$

em que $\psi(\cdot)$ é a função *digama* dada por $\psi(z) = \partial \log\Gamma(z)/\partial z, z > 0$ e $\psi'(\cdot)$ é a função *trigama* dada por $\psi'(z) = \partial^2 \log\Gamma(z)/\partial z^2, z > 0$, e $\hat{h}_{ii}^*$ são os elementos da diagonal da matriz

$$\mathbf{H}^* = (\hat{\mathbf{M}}\Phi)^{1/2}\mathbf{C}(\mathbf{C}^\top \Phi \hat{M} \mathbf{C})^{-1}\mathbf{C}^\top(\Phi\hat{\mathbf{M}})^{1/2}, \tag{5.13}$$

em que $\widehat{\Phi} = \hat{\phi}I$, $\hat{\mathbf{M}} = \text{diag}\{\hat{m}_1, \ldots, \hat{m}_n\}$, $\hat{m}_i = (\hat{\phi}_i\widehat{v}_i)/(g'(\widehat{\mu}_i))^2$ e $\mathbf{C}$, uma matriz $n \times p$ dada por

$$\mathbf{C} = \begin{bmatrix} 1 & W_1 & z_{11} & z_{21} & \ldots & z_{k1} \\ 1 & W_2 & z_{12} & z_{22} & \ldots & z_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & W_n & z1n & z_{2n} & \ldots & z_{kn} \end{bmatrix}, \tag{5.14}$$

em que z_{ki} representa o k-ésimo valor da covariável medida sem erro para o i-ésimo indivíduo, $k = 1, \ldots, p$, $i = 1, \ldots, n$ e a covariável não observada X_i é substituída pela covariável observada W_i, $i = 1, \ldots, n$. Para o método Calibração da Regressão, $\hat{\mu}_i^*$, υ_i e $\hat{h}_{ii}^*$ são dados por,

$$\begin{aligned} \widehat{\mu}_i^* &= \psi\left(\frac{\exp(\hat{\beta}_0 + \hat{\beta}_1\widehat{I\!E}(X_i|W_i) + \hat{\boldsymbol{\gamma}}^\top\mathbf{z}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\widehat{I\!E}(X_i|W_i) + \hat{\boldsymbol{\gamma}}^\top\mathbf{z}_i)}\widehat{\phi}\right) - \psi\left(\frac{1}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\widehat{I\!E}(X_i|W_i) + \hat{\boldsymbol{\gamma}}^\top\mathbf{z}_i)}\widehat{\phi}\right), \\ \widehat{\upsilon}_i &= \psi'\left(\frac{\exp(\hat{\beta}_0 + \hat{\beta}_1\widehat{I\!E}(X_i|W_i) + \hat{\boldsymbol{\gamma}}^\top\mathbf{Z}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\widehat{I\!E}(X_i|W_i) + \hat{\boldsymbol{\gamma}}^\top\mathbf{Z}_i)}\widehat{\phi}\right) + \psi'\left(\frac{1}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\widehat{I\!E}(X_i|W_i) + \hat{\boldsymbol{\gamma}}^\top\mathbf{Z}_i)}\widehat{\phi}\right) \end{aligned} \tag{5.15}$$

e $\hat{h}_{ii}^*$ são os elementos da diagonal da matriz

$$\mathbf{H}^* = (\hat{\mathbf{M}}\Phi)^{1/2}\mathbf{C}(\mathbf{C}^\top\Phi\hat{\mathbf{M}}\mathbf{C})^{-1}\mathbf{C}^\top(\Phi\hat{\mathbf{M}})^{1/2}, \tag{5.16}$$

em que $\widehat{\Phi} = \hat{\phi}I$, $\hat{M} = \text{diag}\{\hat{m}_1, \ldots, \hat{m}_n\}$, $\hat{m}_i = (\hat{\phi}_i\widehat{v}_i)/(g'(\widehat{\mu}_i))^2$ e $\mathbf{C}$ é uma matriz $n \times p$ dada por

$$\widehat{\mathbf{C}} = \begin{bmatrix} 1 & \widehat{I\!E}(X_1|W_1) & z_{11} & z_{21} & \ldots & z_{k1} \\ 1 & \widehat{I\!E}(X_2|W_2) & z_{12} & z_{22} & \ldots & z_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \widehat{I\!E}(X_n|W_n) & z_{1n} & z_{2n} & \ldots & z_{kn} \end{bmatrix}. \tag{5.17}$$

Os resíduos ordinários e ponderados padronizados para os métodos *naive* máxima verossimilhança (NMV) e calibração máxima verossimilhança (CMV) são expressos de forma idêntica as equações (2.13) e (5.11) em que $\hat{\phi}$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ e $\hat{\boldsymbol{\gamma}}^\top$ são as estimativas de máxima

verossimilhança aproximada de ϕ, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}^\top$ obtidas a partir da maximização da função de log-verossimilhança de Hermite dada por (5.7).

Os resíduos ordinários e ponderados padronizados para os métodos *naive* pseudo verossimilhança (NPV) e calibração pseudo verossimilhança (CPV) são expressos de forma idêntica as equações (2.13) e (5.11) em que $\hat{\phi}$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ e $\hat{\boldsymbol{\gamma}}^\top$ são as estimativas de máxima pseudo verossimilhança de ϕ, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}^\top$ obtidas a partir da maximização da função de log-pseudo verossimilhança de Hermite dada por (5.9).

5.4.2 Análise de Influência Local

No Capítulo 2, Seção 2.3.2, falamos sobre a importância de uma análise de influência local para avaliar a qualidade do ajuste de um modelo de regressão. Neste capítulo mostramos os principais resultados para o modelo de regressão beta com erro de medida multiplicativo de forma mais resumida.

Perturbação da Função Log-Verossimilhança ou Perturbação de casos

(a) **Método *Naive*:** Para o método *naive*, a função de log-verossimilhança perturbada para o modelo de regressão beta com erro de medida multiplicativo é construída a partir da perturbação da função de log-verossimilhança dada em (5.3), que resulta em

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{\omega}) &= \sum_{i=1}^{n} \omega_i l_i(\mu_i, \phi) \\ &= \sum_{i=1}^{n} \{\omega_i \log\Gamma(\phi) - \omega_i \log\Gamma(\mu_i\phi) + \omega_i(\mu_i\phi - 1)\log y_i \\ &+ \omega_i \{(1-\mu_i)\phi - 1\}\log(1-y_i)\}\,. \end{aligned} \tag{5.18}$$

Os termos da matriz $\boldsymbol{\Delta}_i$, definida em (2.16), serão dados por:

$$\begin{aligned}
\Delta_{i1}^{\top} &= \phi(y_i^* - \mu_i^*),\\
\Delta_{i2}^{\top} &= \phi(y_i^* - \mu_i^*)W_i,\\
\Delta_{i3}^{\top} &= \phi(y_i^* - \mu_i^*)\mathbf{z}_i,\\
\Delta_{i4}^{\top} &= \mu_i(y_i^* - \mu_i^*) + \psi(\phi) - \psi((1-\mu_i)\phi) + \log(1-y_i),
\end{aligned} \tag{5.19}$$

em que $y_i^* = \log(y_i/(1-y_i))$, μ_i^* é dada em (5.12) e $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$, $i = 1, \ldots, n$, é o vetor de covariáveis medidas sem erro.

(b) **Método Calibração da Regressão:** Para o método calibração da regressão, a função de log-verossimilhança perturbada para o modelo de regressão beta com erro de medida multiplicativo será idêntica a (5.18), com os termos da matriz $\boldsymbol{\Delta}_i$, definida em (2.16) dados por:

$$\begin{aligned}
\Delta_{i1}^{\top} &= \phi(y_i^* - \mu_i^*),\\
\Delta_{i2}^{\top} &= \phi(y_i^* - \mu_i^*)I\!E(X_i|W_i),\\
\Delta_{i3}^{\top} &= \phi(y_i^* - \mu_i^*)\mathbf{z}_i,\\
\Delta_{i4}^{\top} &= \mu_i(y_i^* - \mu_i^*) + \psi(\phi) + \psi((1-\mu_i)\phi) + \log(1-y_i),
\end{aligned} \tag{5.20}$$

em que a esperança condicional $I\!E(X_i|W_i)$ está definida em (4.12).

(c) **Método de Máxima Verossimilhança:** Para o método de máxima verossimilhança aproximada, a função de log-verossimilhança perturbada para o modelo de regressão beta com erro de medida multiplicativo é construída a partir da perturbação da função de

log-verossimilhança de Hermite dada em (5.7), que resulta em

$$l(\boldsymbol{\zeta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i l_H(\boldsymbol{\zeta}; \widehat{\boldsymbol{\delta}}) \\ = \sum_{i=1}^{n} \omega_i \log\left(\sum_{q=1}^{Q} \frac{1}{\sqrt{\pi}} p_q f_{Y_i,W_i|X_i}\left(y_i, w_i|\, x_q\right)\right), \tag{5.21}$$

em que $f_{Y_i,W_i|X_i}\left(y_i, w_i|\, x_q\right)$ é dada em (4.15), $x_{iq} = \exp(r_q\sigma_{x^*}\sqrt{2} + \mu_{x^*})$, Q é o número de pontos da quadratura, $r_1, \ldots, r_Q$ são as raízes do polinômio ortogonal de Hermite $H_q(x)$ (ou nós da quadratura) e $p_1, \ldots, p_Q$ são os pesos do polinômio já apresentados em (4.24).

(d) **Método de Pseudo Verossimilhança:** Para o método de pseudo verossimilhança, a função de log-verossimilhança perturbada para o modelo de regressão beta com erro de medida multiplicativo é construída a partir da perturbação da função de log-verossimilhança de Hermite dada em (5.9), que resulta em

$$l(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i l_H(\boldsymbol{\theta}; \widehat{\boldsymbol{\delta}}) \\ = \sum_{i=1}^{n} \omega_i \log\left(\sum_{q=1}^{Q} \frac{1}{\sqrt{\pi}} p_q f_{Y_i,W_i|X_i}\left(y_i, w_i|\, x_q\right)\right), \tag{5.22}$$

em que $f_{Y_i,W_i|X_i}\left(y_i, w_i|\, x_q\right)$ é dada em (4.15), $x_{iq} = \exp\left(r_q\widehat{\sigma}_{x^*}\sqrt{2} + \widehat{\mu}_{x^*}\right)$, Q é o número de pontos da quadratura, $r_1, \ldots, r_Q$ são as raízes do polinômio ortogonal de Hermite $H_q(x)$ (ou nós da quadratura) e $p_1, \ldots, p_Q$ são os pesos do polinômio já apresentados em (4.24).

Os termos da matriz $\boldsymbol{\Delta}_i$, definida em (2.16), são os mesmos para os métodos de máxima

verossimilhança aproximada e pseudo verossimilhança, ou seja,

$$\Delta_{i1}^{\top} = \frac{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)\left[\widehat{\phi}(y_i^* - \widehat{\mu}_i^*)\right]\{\widehat{\mu}_i(1-\widehat{\mu}_i)\}}{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)},$$

$$\Delta_{i2}^{\top} = \frac{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)\left[\widehat{\phi}(y_i^* - \widehat{\mu}_i^*)\right]\{\widehat{\mu}_i(1-\widehat{\mu}_i)x_{iq}\}}{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)},$$

$$\Delta_{i3}^{\top} = \frac{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)\left[\widehat{\phi}(y_i^* - \widehat{\mu}_i^*)\right]\{\widehat{\mu}_i(1-\widehat{\mu}_i)\mathbf{z}_i\}}{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)},$$

$$\Delta_{i4}^{\top} = \frac{1}{\sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)} \sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i,W_i|X_{iq}}\left(y_i, w_i|\, x_{iq}\right)\Big\{\psi(\widehat{\phi}) - \psi((1-\widehat{\mu}_i)\widehat{\phi}) + \widehat{\mu}_i(y_i^* - \widehat{\mu}_i^*) + \log(1-y_i)\Big\}, \tag{5.23}$$

Para o método *naive* pseudo verossimilhança (NPV) (ver 5.3.4) substitui-se a covariável não observada X_i por W_i. Para o método calibração pseudo verossimilhança (CPV) (ver 5.3.4) substitui-se a covariável não observada X_i por $\mathbb{E}(X_i|W_i)$, $i = 1, \ldots, n$, dada em (4.12), com os estimadores de $\boldsymbol{\theta}$ obtidos via maximização da função log-verossimilhança de Hermite definida em (5.9). O mesmo ocorre para o método máxima verossimilhança aproximada mas com os estimadores de $\boldsymbol{\zeta}$ obtidos via maximização da log-verossimilhança de Hermite definida em (5.7).

Perturbação da variável resposta

A perturbação na variável resposta é realizada através da adição de um vetor $\boldsymbol{\omega}_i$ de pequenas perturbações, $i = 1, \ldots, n$, ao vetor de variáveis respostas $\boldsymbol{y} = (y_1, \ldots, y_n)^{\top}$. Como cada y_i apresenta uma variância diferente é usual se utilizar um fator de escala para padronizar os

componentes de $\boldsymbol{\omega}_i$, por exemplo, a estimativa do desvio padrão de y_i, de forma que

$$y_i(\omega) = y_i + \omega_i s_{y_i}, i = 1, \ldots, n, \tag{5.24}$$

em que $s_{y_i} = \sqrt{\widehat{\mu}_i(1-\widehat{\mu}_i)/(1+\widehat{\phi})}$. Para esse tipo de perturbação $\boldsymbol{\omega}_0 = (0, 0, \ldots, 0)^\top$. Para o caso de perturbação da variável resposta, os termos da matriz $\boldsymbol{\Delta}_i$ foram calculados numericamente no software R (R Core Team, 2014).

5.5 Estudo de Simulação

Nesta seção são apresentados os resultados de um estudo de simulação para o modelo de regressão beta com erro de medida multiplicativo log-normal. Três cenários são explorados. No Cenário 1, comparamos os métodos de estimação: máxima verossimilhança aproximada, máxima pseudo verossimilhança, *naive* e calibração da regressão. No Cenário 2, comparamos os resultados para os métodos de máxima verossimilhança aproximada e pseudo verossimilhança considerando diversos tamanhos de amostra e finalmente, no Cenário 3, comparamos os métodos de estimação: máxima verossimilhança aproximada, máxima pseudo verossimilhança, *naive* e calibração da regressão para diversos valores da variância do erro para um tamanho de amostra fixado. Em todos os cenários foram realizadas $N = 1000$ simulações.

Para os métodos máxima verossimilhança aproximada e pseudo verossimilhança, a variância do erro é considerada desconhecida e é estimada através de duas replicações para $W_i, i = 1, \ldots, n$. A estimação dos parâmetros pelos métodos *naive* e calibração da regressão é feita utilizando a função *betareg*, do pacote *betareg* e as estimações via máxima verossimilhança aproximada e máxima pseudo verossimilhança aproximada é feita utilizando a função *optim*, ambas do software R (R Core Team, 2014). Os pontos de quadratura são obtidos a partir da função *gauss.quad* do R (R Core Team, 2014), que gera pontos e nós para a quadratura de Hermite.

Foram calculadas as médias das estimativas nas N simulações, as raízes dos erros quadráticos médios ($\sqrt{\text{EQM}}$), as médias dos erros padrão assintóticos (EPA), os erros padrão empíricos (EPE) e as probabilidades de cobertura de 95% para os intervalos de confiança assintóticos

($I\!P$(ICA)).

Para o método *naive* usamos a matriz de variâncias e covariâncias fornecidas pela função *vcov* do pacote *betareg* e para o método calibração da regressão, a variância dos estimadores foram obtidas via *boostrap* não-paramétrico. Para $\boldsymbol{\theta} = (\beta_0, \beta_1, \gamma, \phi)$, a raíz do erro quadrático médio para as N simulações foi obtida a partir da expressão:

$$\sqrt{\text{EQM}}(\theta_j) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{\theta}_j^{(i)} - \theta_j)^2}, j = 1, \ldots, 4. \tag{5.25}$$

Os erros padrão empíricos (EPE) são obtidos a partir da raíz quadrada da variância das N simulações que é dada por:

$$\text{Var}(\widehat{\theta}_j) = \frac{1}{N-1}\sum_{i=1}^{N}(\widehat{\theta}_j^{(i)} - \bar{\widehat{\theta}})^2, \quad i = 1, \ldots, N, \tag{5.26}$$

em que $\bar{\widehat{\theta}} = \sum_{i=1}^{N}\widehat{\theta}_i/N$ e $\widehat{\theta}_j^{(i)}$ representam a estimativa de $\widehat{\theta}_j$ na i-ésima simulação.

5.5.1 Cenário 1

Os dados foram gerados com as seguintes características:

- $\varepsilon_1 \sim$ Log-N$(\frac{-\sigma_\varepsilon^2}{2}; \sigma_\varepsilon^2)$ com $\sigma_\varepsilon^2 = 0.15$;
- $\varepsilon_2 \sim$ Log-N$(\frac{-\sigma_\varepsilon^2}{2}; \sigma_\varepsilon^2)$ com $\sigma_\varepsilon^2 = 0.15$;
- $X \sim$ Log-N$(\mu_{x^*}; \sigma_{x^*}^2)$, $\mu_{x^*} = \mu_{w^*} + \frac{\sigma_\varepsilon^2}{2}$ e $\sigma_{x^*}^2 = \sigma_{w^*}^2 - \sigma_\varepsilon^2$, com $\mu_{w^*} = -0.15$ and $\sigma_{w^*}^2 = 0.65$;
- $W_1 = X\varepsilon_1$ e $W_2 = X\varepsilon_2$;
- $Z \sim$ Log-N$(\mu_z; \sigma_z^2)$ com $\mu_z = -0.65$ and $\sigma_z^2 = 0.10$;

Além disso, assumimos apenas uma covariável Z medida sem erro e uma covariável X medida com erro. Os valores atribuídos para ϕ, γ e $\boldsymbol{\beta} = (\beta_0, \beta_1)$ foram:

$$\phi = 3.5; \quad \beta_0 = 1.0; \qquad \beta_1 = -0.5 \quad \text{e} \quad \gamma = -0.5.$$

A Tabela 5.1 mostra os resultados obtidos para os parâmetros do modelo considerando um tamanho de amostra de $n = 300$ em 1000 simulações.

Tabela 5.1: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios, $\sqrt{EQM}$, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura para os intervalos de confiança assintóticos de 95% ($\mathbb{P}$(ICA)) considerando um tamanho de amostra de $n = 300$.*

Parâmetros	Métodos	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{P}$(ICA)
β_0	*Naive*	0.5432	0.4934	0.1609	0.1865	0.2220
(1.00)	Calibração	1.0203	0.2065	0.2028	0.2056	0.9480
	Máxima-Veros	1.0531	0.2163	0.2088	0.2098	0.9440
	Pseudo-Veros	1.0253	0.2084	0.2022	0.2070	0.9470
β_1	*Naive*	-0.2781	0.2286	0.0292	0.0554	0.0020
(-0.50)	Calibração	-0.5167	0.0878	0.0838	0.0862	0.9390
	Máxima-Veros	-0.5359	0.0978	0.0915	0.0910	0.9490
	Pseudo-Veros	-0.5093	0.0864	0.0815	0.0859	0.9470
γ	*Naive*	-0.4253	0.2308	0.2252	0.2186	0.9500
(-0.50)	Calibração	-0.4824	0.3141	0.3058	0.3138	0.9350
	Máxima-Veros	-0.4985	0.3103	0.3116	0.3104	0.9510
	Pseudo-Veros	-0.4984	0.3100	0.3048	0.3102	0.9410
ϕ	*Naive*	3.1135	0.4679	0.2293	0.2639	0.5730
(3.50)	Calibração	3.4265	0.2650	0.2584	0.2548	0.9340
	Máxima-Veros	3.6404	0.3286	0.2951	0.2972	0.9420
	Pseudo-Veros	3.6318	0.3234	0.2919	0.2955	0.9180

Analisando os resultados da Tabela 5.1, vemos que o método de máxima pseudo verossimilhança obteve os melhores resultados para as estimativas médias de todos os coeficientes e menor valor de EQM para os coeficientes β_0, β_1 e γ. Além disso, os valores das raízes dos erros quadráticos médios ($\sqrt{\text{EQM}}$), erros padrão assintóticos (EPA) e erros padrão empíricos (EPE) estão bem próximos entre si para os métodos de máxima verossimilhança aproximada e pseudo verossimilhança. As probabilidades de coberturas referentes aos intervalos de confiança assintótico médio, $\mathbb{P}$(ICA), também estão melhores para os métodos de máxima verossimilhança aproximada e pseudo verossimilhança, ou seja, mais próximas do valor nominal 95%. O método máxima verossimilhança aproximada apresenta bons resultados mas é computacionalmente mais demorado que o método pseudo verossimilhança e só funciona corretamente dada a estimativa da variância do erro, caso contrário ele apresenta problemas em estimar os erros padrão dos estimadores.

5.5.2 Cenário 2

Considerando a mesma parametrização adotada no Cenário 1, a tabela a seguir apresenta comparações entre os tamanhos amostrais, $n = 50$, $n = 100$, $n = 200$ e $n = 300$, com o interesse em verificar a consistência dos estimadores de máxima verossimilhança aproximada.

Tabela 5.2: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios, $\sqrt{EQM}$, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura para os intervalos de confiança assintóticos de 95% ($\mathbb{P}$(ICA)) considerando os tamanhos amostrais $n = 50$, $n = 100$, $n = 200$ e $n = 300$ para o método máxima verossimilhança aproximada.*

Parâmetros	Amostras	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{P}$(ICA)
β_0 (1.00)	$n = 50$	1.0812	0.5473	0.5315	0.5415	0.9510
	$n = 100$	1.0724	0.3697	0.3667	0.3627	0.9470
	$n = 200$	1.0524	0.2631	0.2565	0.2579	0.9500
	$n = 300$	1.0531	0.2163	0.2087	0.2098	0.9440
β_1 (-0.50)	$n = 50$	-0.5680	0.2682	0.2448	0.2596	0.9500
	$n = 100$	-0.5439	0.1787	0.1639	0.1733	0.9530
	$n = 200$	-0.5416	0.1173	0.1133	0.1098	0.9600
	$n = 300$	-0.5359	0.0978	0.0915	0.0910	0.9490
γ (-0.50)	$n = 50$	-0.4843	0.7855	0.7779	0.7858	0.9500
	$n = 100$	-0.5016	0.5250	0.5240	0.5253	0.9390
	$n = 200$	-0.4873	0.3940	0.3829	0.3940	0.9410
	$n = 300$	-0.4985	0.3103	0.3116	0.3104	0.9510
ϕ (3.50)	$n = 50$	4.0715	1.1613	0.8615	1.0114	0.9610
	$n = 100$	3.7672	0.6200	0.5377	0.5597	0.9600
	$n = 200$	3.6889	0.4192	0.3679	0.3744	0.9490
	$n = 300$	3.6404	0.3286	0.2951	0.2972	0.9420

A Tabela 5.2 mostra que a medida que o tamanho da amostra aumenta as estimativas médias ficam bem mais próximas dos valores reais dando um indicativo de consistência dos estimadores. Também houve considerável redução nos valores de erros quadráticos médios e estes estão bem próximos dos EPA e EPE. Todas as probabilidades de cobertura estão próximas do valor nominal de 95% e também melhoram com o aumento do tamanho da amostra.

Ainda com a mesma parametrização adotada no Cenário 1, a tabela a seguir apresenta comparações entre os tamanhos amostrais, $n = 50$, $n = 100$, $n = 200$ e $n = 300$, com o interesse em verificar a consistência dos estimadores de pseudo verossimilhança.

A Tabela 5.3 mostra que a medida que o tamanho da amostra aumenta as estimativas médias ficam bem mais próximas dos valores reais dando um indicativo de consistência dos

Tabela 5.3: *Médias das estimativas dos parâmetros, raízes dos erros quadráticos médios,* $\sqrt{EQM}$*, médias dos erros padrão assintóticos (EPA), erros padrão empíricos (EPE), probabilidades de cobertura para os intervalos de confiança assintóticos de 95% (*$\mathbb{P}$*(ICA)) considerando os tamanhos amostrais* $n = 50$*,* $n = 100$*,* $n = 200$ *e* $n = 300$ *para o método pseudo verossimilhança.*

Parâmetros	Amostras	Médias	$\sqrt{\text{EQM}}$	EPA	EPE	$\mathbb{P}$(ICA)
β_0 (1.00)	$n = 50$	1.0388	0.5286	0.4960	0.5275	0.9270
	$n = 100$	1.0400	0.3574	0.3531	0.3554	0.9440
	$n = 200$	1.0234	0.2551	0.2495	0.2541	0.9450
	$n = 300$	1.0253	0.2083	0.2021	0.2070	0.9470
β_1 (-0.50)	$n = 50$	-0.5276	0.2341	0.2010	0.2326	0.9020
	$n = 100$	-0.5130	0.1613	0.1429	0.1609	0.9260
	$n = 200$	-0.5138	0.1034	0.1009	0.1029	0.9450
	$n = 300$	-0.5093	0.0860	0.0815	0.0859	0.9470
γ (-0.50)	$n = 50$	-0.4838	0.7848	0.7327	0.7850	0.9280
	$n = 100$	-0.5016	0.5250	0.5240	0.5253	0.9390
	$n = 200$	-0.4870	0.3938	0.3766	0.3938	0.9390
	$n = 300$	-0.4984	0.3100	0.3048	0.3102	0.9410
ϕ (3.50)	$n = 50$	4.0463	1.1202	0.8144	0.9784	0.9310
	$n = 100$	3.7545	0.6083	0.5187	0.5527	0.9520
	$n = 200$	3.6797	0.4129	0.3634	0.3720	0.9460
	$n = 300$	3.6318	0.3232	0.2919	0.2955	0.9390

estimadores. Também houve considerável redução nos valores de erros quadráticos médios e estes estão bem próximos dos EPA e EPE. Todas as probabilidades de cobertura estão próximas do valor nominal de 95% e também melhoram com o aumento do tamanho da amostra.

5.5.3 Cenário 3

Considerando a mesma parametrização adotada no Cenário 1, apresentamos o que acontece com as médias das estimativas dos parâmetros quando variamos a variância do erro utilizando um tamanho de amostra de $n = 300$ para os métodos *naive*, calibração da regressão, máxima verossimilhança aproximada e pseudo verossimilhança.

Pelos resultados apresentados na Tabela 5.4 vemos que:

- Para o intercepto e o coeficiente da covariável medida com erro, β_1, o método *naive* apresenta os piores resultados e o método pseudo verossimilhança os melhores.

- Para o coeficiente da covariável medida sem erro, γ, todos os métodos apresentam boas estimativas com resultados próximos entre si mas o método pseudo verossimilhança ainda fica um pouco melhor.

Tabela 5.4: *Médias das estimativas considerando diversos valores para a variância do erro de medida, σ_ε^2, para um tamanho amostral de $n = 300$.*

Parâmetros	Métodos	$\sigma_\varepsilon^2 = 0.10$	$\sigma_\varepsilon^2 = 0.15$	$\sigma_\varepsilon^2 = 0.20$	$\sigma_\varepsilon^2 = 0.25$
β_0	*Naive*	0.8797	0.8581	0.8336	0.8015
(1.00)	Calibração	0.9800	1.0203	1.0646	1.1080
	Máxima-Veros	1.0028	1.0531	1.1228	1.2115
	Pseudo-Veros	0.9620	1.0253	1.0500	1.0808
β_1	*Naive*	-0.4030	-0.3806	-0.3612	-0.3375
(-0.50)	Calibração	-0.4828	-0.5167	-0.5484	-0.5891
	Máxima-Veros	-0.4993	-0.5359	-0.5944	-0.6732
	Pseudo-Veros	-0.4612	-0.5093	-0.5272	-0.5554
γ	*Naive*	-0.4681	-0.4905	-0.4949	-0.4945
(-0.50)	Calibração	-0.4661	-0.4824	-0.4900	-0.4864
	Máxima-Veros	-0.4847	-0.4985	-0.5046	-0.5032
	Pseudo-Veros	-0.4757	-0.4984	-0.5040	-0.5020
ϕ	*Naive*	3.4352	3.4077	3.3785	3.3790
(3.50)	Calibração	3.4484	3.4265	3.4161	3.3751
	Máxima-Veros	3.5996	3.6404	3.6903	3.7746
	Pseudo-Veros	3.5569	3.6318	3.6538	3.7121

Analisando os resultados apresentados nas Tabelas 5.1, 5.2, 5.3 e 5.4 concluímos que o método pseudo verossimilhança obteve melhor desempenho em comparação ao método de máxima verossimilhança aproximada, pois é computacionalmente mais rápido e as estimativas médias ficaram mais próximas dos valores reais com menores valores de $\sqrt{\text{EQM}}$, EPA e EPE. Por esta razão a análise de dados reais é feita apenas via pseudo verossimilhança.

5.6 Aplicação

Nesta seção apresentamos os resultados do ajuste do modelo de regressão beta com erro multiplicativo log-normal aos dados provenientes de uma instituição financeira do Brasil. Utilizando uma amostra de $n = 200$, o conjunto de dados é composto pelas seguintes variáveis:

- Y_i, variável resposta de interesse. Representa a proporção de gastos do limite no cheque especial, $i = 1, 2, \ldots, n$.
- X_i, covariável não observada. Representa a renda real do futuro cliente, $i = 1, 2, \ldots, n$.
- W_{ij}, covariável observada. Representa a renda presumida do cliente i obtida a partir de um dado modelo de renda presumida j disponível no mercado, $i = 1, 2, \ldots, n$ e $j = 1, 2$.

- Z_i, covariável binária observada que representa o gênero do cliente, em que 0 significa que o cliente é do sexo masculino e 1 significa que o cliente é do sexo feminino, $i = 1, 2, \ldots, n$.

Os gráficos apresentados em (5.1) e (5.2) mostram o comportamento dos dados.

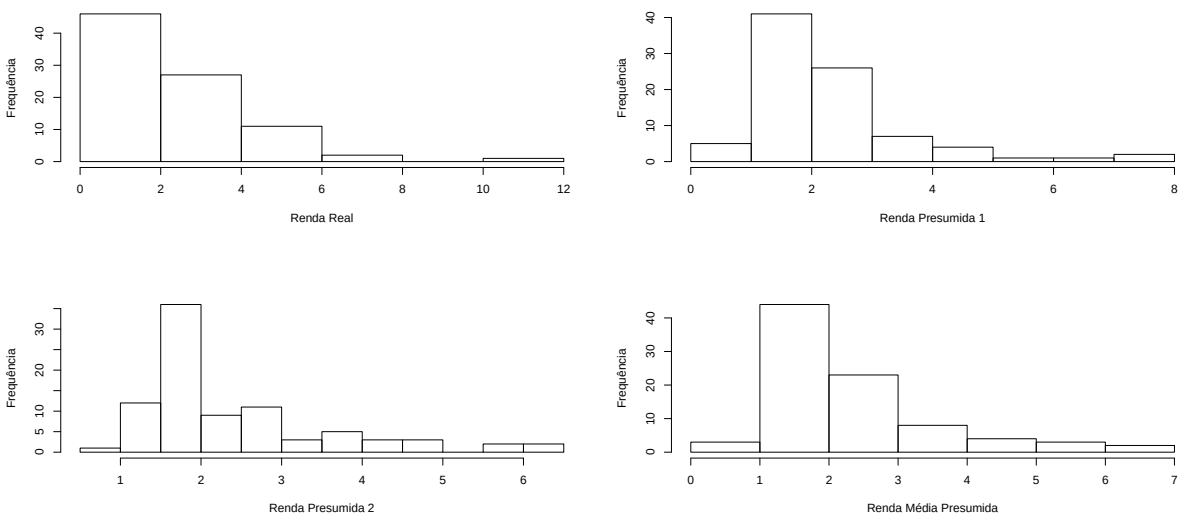

Figura 5.1: Renda real, renda presumida 1, renda presumida 2 e , renda média presumida.

Pela Figura (5.1), vemos que as covariáveis podem se ajustar a uma distribuição log-normal.

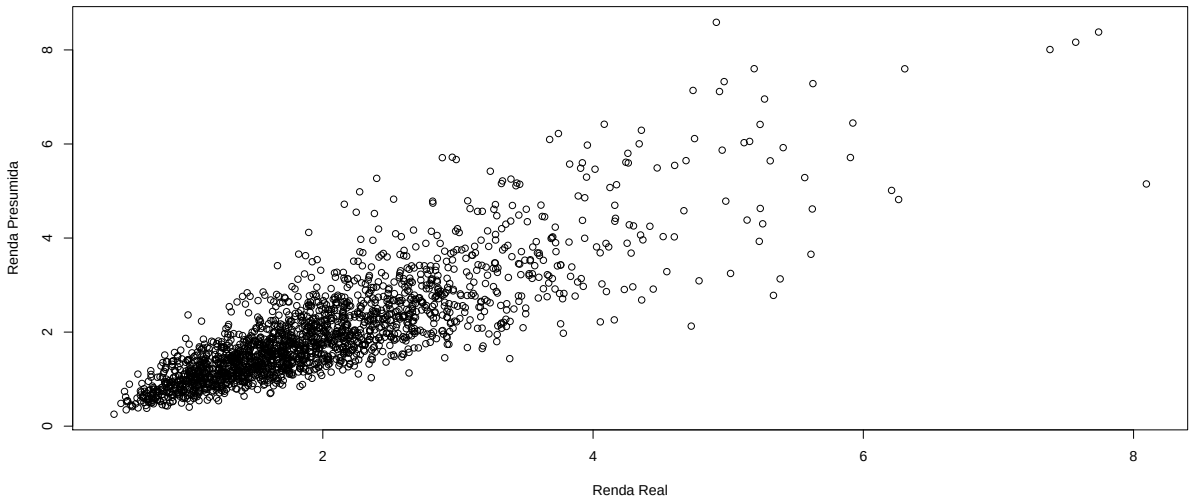

Figura 5.2: Renda presumida *versus* Renda real.

Pela Figura (5.2), vemos que a medida que aumenta a renda dos indivíduos os pontos vão ficando mais distantes entre si, ou seja, proporcionalmente erra-se menos com quem ganha

menos e erra-se mais com quem ganha mais.

A tabela a seguir mostra os resultados obtidos para o ajuste dos dados reais utilizando o método de pseudo verossimilhança, já que o mesmo apresentou melhor desempenho no estudo de simulação.

Tabela 5.5: *Estimativas dos parâmetros, Erros Padrão e Intervalos de Confiança.*

Parâmetros	Estimativas	Erros Padrão	IC(95%)
β_0	1.1302	0.2232	(0.6926, 1.5679)
β_1	-0.4956	0.1411	(-0.7722,-0.2189)
γ	-0.3010	0.0124	(-0.3254,-0.2766)
ϕ	3.9694	0.0679	(3.8361, 4.1024)

Analisando os resultados da Tabela 5.5 vemos que os valores dos erros padrão são razoavelmente pequenos e os coeficientes da regressão são significativos a 5%, dado que o zero não pertence a nenhum dos intervalos de confiança. Para verificarmos a qualidade deste ajuste apresentamos a seguir alguns gráficos com representações dos resíduos e análise de influência local.

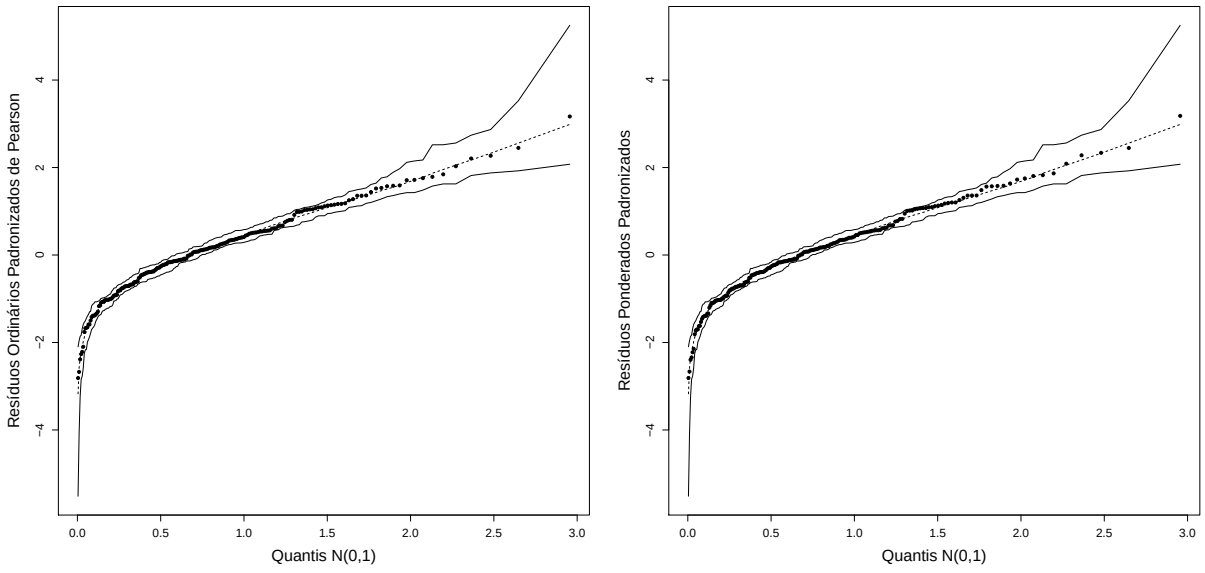

Figura 5.3: Gráfico de envelope quantil normal para os resíduos ordinários padronizados e ponderados padronizados.

Utilizando a teoria de gráfico envelope apresentada em Ferrari e Cribari-Neto (2004), construímos gráficos para os resíduos ordinários padronizados e ponderados padronizados. A

Figura 5.3 mostra uma indicação que os resíduos são normalmente distribuídos e estão entre -3 e 3. Para confirmar se realmente são normalmente distribuídos realizamos um teste de normalidade nos resíduos (*Shapiro-Wilk*) e obtivemos um p-valor de 0.08672 para os resíduos ordinários padronizados e um p-valor de 0.8848 para os resíduos ponderados padronizados. Logo, os resíduos são normalmente distribuídos e o modelo aparentemente está bem ajustado aos dados.

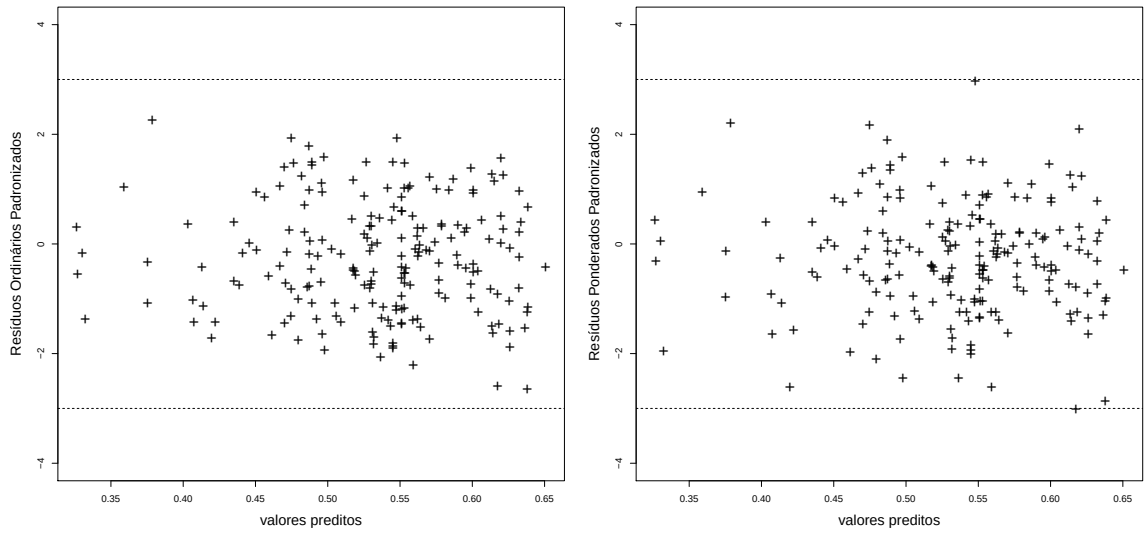

Figura 5.4: Resíduos versus valores preditos via método calibração pseudo verossimilhança.

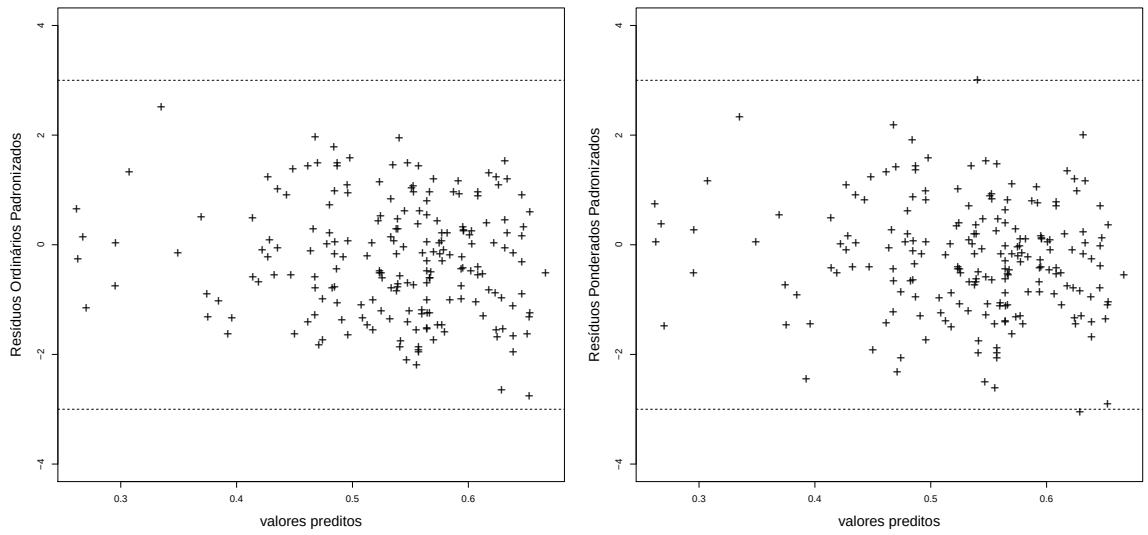

Figura 5.5: Resíduos versus valores preditos via método *naive* pseudo verossimilhança.

Analisando os gráficos das Figuras 5.4 e 5.5, não vemos a presença de *outliers* e os resíduos se distribuem aleatoriamente em torno do zero. A seguir vamos analisar os gráficos de influência local. Para o método de perturbação de casos temos:

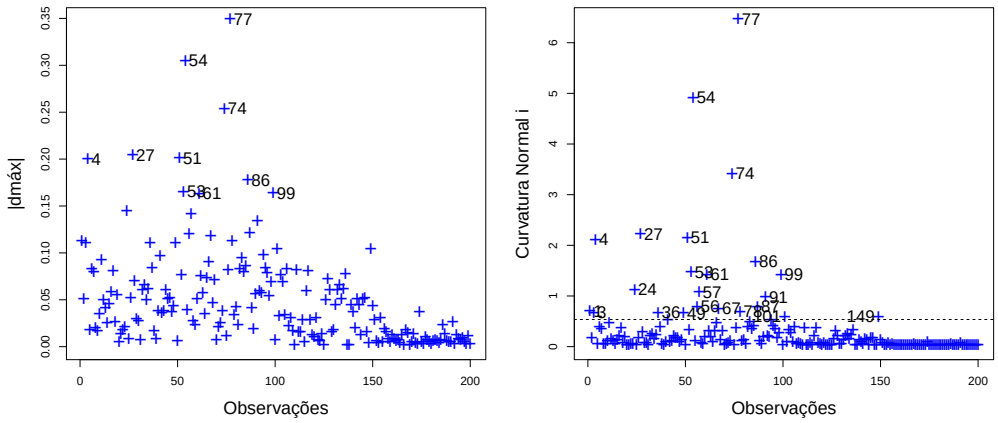

Figura 5.6: Gráficos de *dmax* e C_i pelos índices observados usando o método de pseudo verossimilhança para o esquema de perturbação de casos.

O gráfico de C_i pelos índices observados mostra quais observações são individualmente influentes enquanto que um gráfico de *dmax* pelos índices observados mostra quais observações são conjuntamente influentes. Analisando a Figura 5.6, vemos que as observações que mais se distanciaram das outras foram $\{54, 74, 77\}$. Estas são possíveis candidatas a observações influentes. Para a perturbação da variável resposta temos:

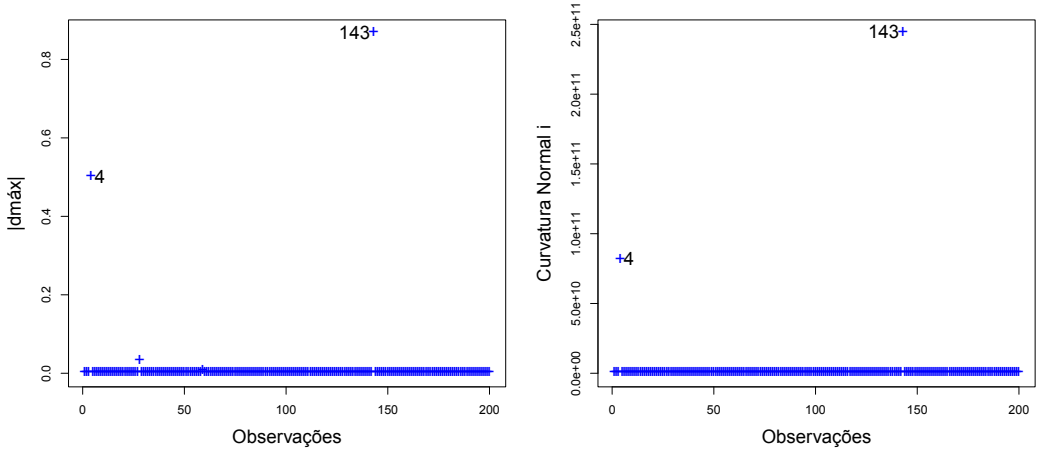

Figura 5.7: Gráficos de *dmax* e C_i pelos índices observados usando o método de pseudo verossimilhança para o esquema de perturbação da variável resposta.

Analisando os gráficos apresentados na Figura 5.7, vemos que apenas as observações 4 e 143 merecem atenção. Por conta disso, refizemos as análises retirando as observações que mais se afastam das demais. Consideramos duas situações a analisar:

Caso 1: Dados sem as observações: 4 e 143. Observações discrepantes da Figura 5.7;

Caso 2: Dados sem as observações: 54, 74 e 77. Observações mais afastadas das demais na Figura 5.6;

Tabela 5.6: *Mudanças Relativas (MR), Estimativas dos parâmetros via pseudo verossimilhança, Erros Padrão e Intervalos de Confiança após a remoção das observações discrepantes.*

		Caso 1		
Parâmetros	MR(%)	Estimativas	Erros Padrão	IC(95%)
β_0	54.31	1.7442	0.1148	(1.5190, 1.9694)
β_1	50.52	-0.7460	0.0752	(-0.8934,-0.5986)
γ	89.07	-0.5691	0.0105	(-0.5896,-0.5486)
ϕ	4.15	4.1341	0.0484	(4.0391, 4.2290)
		Caso 2		
Parâmetros	MR(%)	Estimativas	Erros Padrão	IC(95%)
β_0	12.67	1.2734	0.3070	(0.6715, 1.8753)
β_1	25.46	-0.6218	0.1981	(-1.0102,-0.2335)
γ	3.42	-0.2907	0.0112	(-0.3128,-0.2686)
ϕ	5.35	4.1818	0.1539	(3.8801, 4.4835)

Analisando os resultados da Tabela 5.6 vemos que as mudanças relativas dos coeficientes β_0, β_1 e γ indicam que as estimativas de máxima pseudo verossimilhança destes parâmetros são sensíveis a presença dos pontos discrepantes no Caso 1. Para o Caso 2 os valores de MR não são tão altos. Além disso, observe que os erros padrão aumentaram para três coeficientes no Caso 2 enquanto que para o Caso 1 estes valores diminuíram para todos os coeficientes. Os coeficientes são significativos a 5% em ambos os casos, pois o zero não pertence a nenhum dos intervalos. A partir destes resultados concluímos que a retirada das observações influentes do Caso 1 melhora as estimativas obtidas para os dados reais.

Capítulo

6

Conclusões e Propostas de Trabalhos Futuros

Neste trabalho propomos a construção, definição, métodos de estimação e análise de diagnósticos para os modelos de regressão com erro de medida multiplicativo em uma das covariáveis. No Capítulo 1, apresentamos alguns conceitos básicos referentes a teoria de modelos de regressão com erro de medida nas covariáveis bem como uma breve revisão bibliográfica dos principais trabalhos na área. No Capítulo 2, apresentamos os fundamentos teóricos necessários para o desenvolvimento do nosso trabalho. No Capítulo 3, apresentamos a construção do modelo de regressão série de potências com erro de medida, incluindo definição, representações do erro de medida, métodos de estimação e análise de diagnósticos. No Capítulo 4, um caso particular do modelo de regressão série de potências com erro de medida é apresentado. Neste capítulo definimos o modelo de regressão binomial negativa com erro de medida, apresentamos a construção do erro de medida multiplicativo log-normal e os métodos de estimação para os parâmetros de interesse. Dentre os métodos de estimação apresentados, o método de pseudo verossimilhança obteve melhor desempenho. Os parâmetros perturbadores são estimados utilizando replicações da covariável observada e representados via equações de estimação. Utilizar equações de estimação auxilia na construção da distribuição assintótica para os estimadores dos parâmetros de interesse. Para a construção de intervalos de confiança, para os coeficientes obtidos via pseudo verossimilhança, encontramos a distribuição assintótica

utilizando a abordagem de Carroll *et al.* (2006). Para a verificação da qualidade do ajuste, uma análise de diagnósticos também é apresentada. As propriedades do modelo de regressão binomial negativa com erro de medida para cada método foram analisadas através de um estudo de simulação. Aplicamos a metododologia desenvolvida neste capítulo a um conjunto de dados reais. No Capítulo 5 é feita a construção do modelo de regressão beta com erro de medida multiplicativo log-normal. A construção deste modelo foi motivado por um problema prático descrito detalhadamente no próprio capítulo. Além da construção do modelo de regressão beta com erro de medida multiplicativo, alguns métodos de estimação são estudados. Tais métodos têm como princípio a estimação por máxima verossimilhança e pseudo verossimilhança. Dentre os métodos de estimação apresentados, o método de pseudo verossimilhança também obteve melhor desempenho. Estimamos os parâmetros perturbadores usando a mesma ideia de dados replicados utilizada no Capítulo 4. Apresentamos a distribuição assintótica dos estimadores de pseudo verossimilhança com interesse em construir intervalos de confiança para os coeficientes. Um estudo de simulação foi feito para ilustrar os resultados das estimações para cada método. Além do estudo de simulação, realizamos uma análise a um conjunto de dados reais e análise de diagnóstico.

Como propostas de trabalhos futuros pretendemos utilizar outras distribuições para o erro de medida e para a covariável medida com erro. Também pretendemos realizar outros esquemas de perturbação para a análise de influência local nos modelos estudados bem como análise de influência global.

Referências Bibliográficas

Abramowitz e Stegun (1964) M. Abramowitz e I.A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Dover Publications, New York. ISBN 9780486612720. URL http://books.google.com.br/books?id=hLqv8pHoO0AC. Citado na pág. 36, 72

Buonaccorsi (2010) J.P. Buonaccorsi. *Measurement error: models, methods, and applications*. Chapman and Hall, Boca Raton, FL. Citado na pág. 2, 68

Buonaccorsi e Tosteson (1993) J.P. Buonaccorsi e T.D. Tosteson. Correcting for non linear measurements error in the dependent variable in the linear general models. *Communications in Statistics - Theory and Methods.*, 22:2687–2702. Citado na pág. 11

Carrasco *et al.* **(2014)** Jalmar M.F. Carrasco, S.L.P. Ferrari e R.B. Arellano-Valle. Errors-in-variables beta regression models. *Journal of Applied Statistics*, 41:1530–1547. Citado na pág. 2, 68

Carroll e Ruppert (1988) Raymond J Carroll e David Ruppert. *Transformation and Weighting in Regression*. Chapman and Hall, London. Citado na pág. 8

Carroll *et al.* **(2006)** Raymond J Carroll, David Ruppert, Leonard A Stefanski e Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective; 2nd ed.* Chapman and Hall, Boca Raton, FL. URL https://cds.cern.ch/record/1012043. Citado na pág. 1, 2, 3, 4, 5, 7, 9, 10, 12, 33, 68, 94

Carroll e Spiegelman (1992) R.J. Carroll e C.H. Spiegelman. Diagnostics for nonlinearity and heterocedasticity in errors in variables regression. *Technometrics.*, 34:186–196. Citado na pág. 14, 15, 74

Cheng e Van Ness (1999) C.L. Cheng e J.W. Van Ness. *Statistical regression with measurement error*. Arnold, London. Citado na pág. 1, 2, 68

Cook (1986) R. Cook. Assessment of local influence. *Journal of the Royal Statistical Society B.*, 48:133–169. Citado na pág. 14, 16

Cook (1977) R.D. Cook. Detection of influential observations in linear regression. *Technometrics.*, 19:15–18. Citado na pág. 14

Cordeiro *et al.* **(2009)** G. Cordeiro, M. Andrade e M. de Castro. Power series generalized nonlinear models. *Computational Statistics & Data Analysis.*, 53:1155–1166. Citado na pág. 3, 21, 22

Cribari-Neto e Zeileis (2010) F. Cribari-Neto e A. Zeileis. Beta regression in R. *Journal of Statistical Software.*, 34:1–24. Citado na pág. 68, 71

de Castro e Galea (2015) M. de Castro e M. Galea. Inference in a structural heteroskedastic calibration model. *Statistical Papers.*, 56:479–494. Citado na pág. 2

de Castro *et al.* **(2007)** M. de Castro, M. Galea e H. Bolfarine. Local influence assessment in heteroscedastic measurement error models. *Computational Statistics & Data Analysis.*, 52: 1132–1142. Citado na pág. 14, 74

de Castro *et al.* **(2013)** M. de Castro, H. Bolfarine e M. Galea. Bayesian inference in measurement error models for replicated data. *Environmetrics.*, 24:22–30. Citado na pág. 2

Dolby (1976) G.R. Dolby. The ultrastructural relation: A synthesis of the functional and structural relations. *Biometrika.*, 63:39–60. Citado na pág. 1

Espinheira *et al.* **(2008a)** P.L. Espinheira, S.L.P. Ferrari e F. Cribari-Neto. Influence diagnostics in beta regression. *Computational Statistics & Data Analysis.*, 52:4417–4431. Citado na pág. 68, 74

Espinheira *et al.* **(2008b)** P.L. Espinheira, S.L.P. Ferrari e F. Cribari-Neto. On beta regression residuals. *Journal of Applied Statistics .*, 35:407–419. Citado na pág. 68, 74

Ferrari e Cribari-Neto (2004) S.L.P. Ferrari e F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics.*, 31:799–815. Citado na pág. 68, 69, 74, 88

Ferrari *et al.* **(2011)** S.L.P. Ferrari, P.L. Espinheira e F. Cribari-Neto. Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica.*, 65:337–351. Citado na pág. 74, 75

Fuller (1987) W.A. Fuller. *Measurement error models.* Wiley Series in Probability and Statistics. Wiley. ISBN 9780470317334. Citado na pág. 2, 68

Galea e de Castro (2012) M. Galea e M. de Castro. Influence Assessment in an Heteroscedastic Errors-in-Variables Model. *Communications in Statistics - Theory and Methods.*, 41:8:1350–1363. Citado na pág. 14, 16, 74

Garay *et al.* **(2011)** A.M. Garay, E.M. Hashimoto, E.M.M. Ortega e V.H. Lachos. On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics and Data Analysis.*, 55:1304–1318. Citado na pág. 3

Godambe (1960) V.P. Godambe. An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics.*, 31:1208–1211. Citado na pág. 8

Godambe (1991) V.P. Godambe. *Estimating Functions.* Oxford science publications. Clarendon Press. ISBN 9780198522287. URL https://books.google.com.br/books?id=td1mj1ppjngC. Citado na pág. 8

Gong e Samaniego (1981) G. Gong e F.J. Samaniego. Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics.*, 9:861–869. Citado na pág. 11, 12, 13

Guolo (2011) A. Guolo. Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica.*, 21:1639–1663. Citado na pág. 2, 11, 12

Guolo e Brazzale (2008) A. Guolo e A.R. Brazzale. A simulation based comparison of techniques to corrent for measurement error in matched case-control studies. *Statistics in Medicine.*, 27:3755–3775. Citado na pág. 2

Gupta *et al.* **(1995)** P.L. Gupta, R.C. Gupta e R.C. Tripathi. Inflated modified power series distributions with applications. *Communications in Statistics - Theory and Methods.*, 24: 2355–2374. Citado na pág. 3

Gupta (1974) R.C. Gupta. Modified power series distribution and some of its applications. *Sankhyã.*, B36:288–298. Citado na pág. 3, 21, 22

Gupta (1977) R.C. Gupta. Minimum variance unbiased estimation in a modified power series distribution and some of its applications. *Communications in Statistics - Theory and Methods.*, 6:10:977–991. Citado na pág. 3

Hilbe (2011) J.M. Hilbe. *Negative binomial regression.* Cambridge University Press. Cambridge. ISBN 9780521198158. Citado na pág. 28

Huber (1964) P.J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101. Citado na pág. 8

Huber (1967) P.J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium*, 1:221–233. Citado na pág. 8

Hwang (1986) J.T. Hwang. Multiplicative errors-in-variables models with applications to recent data released by the US Department of Energy. *Journal of the American Statistical Association*, 81:680–688. Citado na pág. 2

Joyce e Richard (1991) M. W. Joyce e F. Richard. Influence Diagnostics for Linear Measurement Error Models. *Biometrika*, 78:373–380. Citado na pág. 14

Kelly (1984) G. Kelly. The influence function in the errors in variable problem. *The Annals of Statistics*, 12:87–100. Citado na pág. 14, 74

Lee *et al.* **(2006)** S.Y. Lee, B Lu e X.Y. Song. Assessing local influence for nonlinear structural equation models with ignorable missing data. *Computational Statistics and Data Analysis*, 50:1356–1377. Citado na pág. 18

Lesaffre e Verbeke (1998) E. Lesaffre e G. Verbeke. Local influence in linear mixed models. *Biometrics*, 54:570–582. Citado na pág. 15, 16

Lyles e Kupper (1997) R.H. Lyles e L.L. Kupper. A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology. *Biometrics*, 53:1008–53. Citado na pág. 2

McCullagh e Nelder (1989) P. McCullagh e J. A. Nelder. *Generalized linear models (Second edition)*. London: Chapman & Hall. Citado na pág. 14, 15

Miller (1990) S.M. Miller. Analysis of residuals from measurement error models. *Contemporary Mathematics.*, 112. Citado na pág. 14, 15, 74

Ortega *et al.* (2015) E.M.M. Ortega, G.M. Cordeiro, A.K. Campelo, M.W. Kattand e V.G. Cancho. A power series beta Weibull regression model for predicting breast carcinoma. *Statistics in Medicine.*, 34:1366–1388. Citado na pág. 3

Ospina (2007) Patricia Leone Espinheira. Ospina. *Regressão Beta.* Tese (doutorado em estatística), IME-USP, São Paulo. URL http://www.teses.usp.br/teses/disponiveis/45/45133/tde-15052007-110020/. Acesso em: 2014-01-17. Citado na pág. 16, 17

Ospina *et al.* (2006) R. Ospina, F. Cribari-Neto e K.L.P. Vasconcellos. Improved point and interval estimation for a beta regression model. *Computational Statistics & Data Analysis.*, 51:960–981. Citado na pág. 68, 70

Patriota (2010) Alexandre Galvão. Patriota. *Modelos heterocedásticos com erros nas variáveis.* Tese (doutorado em estatística), IME-USP, São Paulo. URL http://www.teses.usp.br/teses/. Acesso em: 2014-02-22. Citado na pág. 2

Paula (2004) G.A. Paula. *Modelos de Regressão com apoio computacional.* Citado na pág. 18

Piegorsch (1990) W.W. Piegorsch. Maximum Likelihood Estimation for the Negative Binomial Dispersion Parameter. *Biometrics*, 46:863–867. Citado na pág. 28

R Core Team (2014) R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/. Citado na pág. 32, 41, 44, 71, 81

Samani (2011) E.B. Samani. A missing inflated power series model for regression analysis of the British Household Panel Survey (BHPS) data. *Australian Journal of Basic and Applied Sciences*, 5:325–331. Citado na pág. 3

Samani *et al.* (2012) E.B. Samani, Y. Amirian e A. Ganjali. Likelihood estimation for longitudinal zero-inflated power series regression models. *Journal of Applied Statistics*, 39:9: 1965–1974. Citado na pág. 3

Skrondal e Kuha (2012) A. Skrondal e J. Kuha. Improved regression calibration. *Psychometrika.*, 77:649–669. Citado na pág. 2

Verbeke e Molenberghs (2000) G. Verbeke e G. Molenberghs. *Linear mixed models for longitudinal data.* Springer-Verlag, New York. Citado na pág. 17

Xie e Wei (2009) Feng Chang Xie e Bo Cheng Wei. Diagnostics for generalized poisson regression models with errors in variables. *Journal of Statistical Computation and Simulation*, 79(7):909–922. doi: 10.1080/00949650802050365. URL http://dx.doi.org/10.1080/00949650802050365. Citado na pág. 14, 15, 74

Zhao e Lee (1995) Y. Zhao e A.H. Lee. Assessment of influence in non-linear measurement error models. *Journal of Applied Statistics.*, 22:215–225. Citado na pág. 14, 15, 74

Zhao *et al.* (1994) Y. Zhao, A.H. Lee e Y.V. Hui. Influence diagnostics for generalized Linear measurement error models. *Biometrics.*, 50:1117–1128. Citado na pág. 14, 15, 74