Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-graduação em Estatística

Inferência em modelos de regressão com erros de medição sob enfoque estrutural para observações replicadas

Lorena Yanet Cáceres Tomaya

UFSCar - São Carlos/SP
Março de 2014

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-graduação em Estatística

Inferência em modelos de regressão com erros de medição sob enfoque estrutural para observações replicadas

Lorena Yanet Cáceres Tomaya

Orientador: Prof. Dr. Mário de Castro Andrade Filho

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar como parte dos requisitos para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos/SP
Março de 2014

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

C118im

Cáceres Tomaya, Lorena Yanet.
Inferência em modelos de regressão com erros de medição sob enfoque estrutural para observações replicadas / Lorena Yanet Cáceres Tomaya. -- São Carlos : UFSCar, 2014.
82 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2014.

1. Inferência (Estatística). 2. Modelos com erros de medição. 3. Erros heteroscedásticos. 4. Máxima pseudoverossimilhança. 5. Modelo estrutural. I. Título.

CDD: 519 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br
13565-905 - SÃO CARLOS-SP - BRASIL

FOLHA DE APROVAÇÃO

Aluno(a) : Lorena Yanet Cáceres Tomaya

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA EM 10/03/2014
PELA COMISSÃO JULGADORA:

Presidente _____
Prof. Dr. Mário de Castro Andrade Filho (ICMC-USP /Orientador)

1º Examinador _____
Prof. Dr. Filidor Edilfonso Vilca Labra (UNICAMP)

2º Examinador _____
Profa. Dra. Silvia Lopes de Paula Ferrari (IME-USP)

Agradecimentos

Agradeço primeiramente a Deus pela minha vida, por me dar saúde e pelas forças necessárias de seguir em frente nos momentos difícies.

Aos meus amados pais Lucas, Agusta e Petronila, pelo carinho, motivação, e apoio que recebi durante as realizações na minha vida. Aos meus queridos irmãos Jéssica, Alexandra, Roxana, Rafael, Raúl e Ricardo, pelo carinho, comprenssão e apoio compartilhado.

Meus sinceros agradecimentos ao meu orientador, Mário de Castro, pelas orientações valiosas, pela disponibilidade de tempo, pela confiança, comprenssão e incentivo na condução e desenvolvimento deste trabalho.

Aos professores do Departamento de Estatística da Universidade Federal de São Carlos, em especial Carlos Alberto Ribeiro Diniz e Jorge Bazán do ICMC/USP pelas contribuições neste trabalho. Aos professores José Galvão, Marcio Diniz, Luis Milan, Luis Ernesto Bueno, Adriano Polpo e Vera Tomazela pelas aulas ministradas e pela contribuição na minha formação acadêmica durante o mestrado. À secretaria do programa, Isabel pelo seu dedicado e ótimo trabalho que realiza.

Aos professores membros da banca examinadora pelas críticas e sugestões para o melhoramento deste trabalho.

Aos meus amigos Katherine, Genoveva, Ricky, Claudia, Miki e Verônica por me acompanharem durante este período nos dias de descanso e diversão. De forma especial, agradeço ao Miguel pelo carinho, pelos ânimos e apoio constante. Com gratidão, às minhas amigas Patricia Hilário Tacuri e Náthali Cabrera pela acolhida ao chegar pela primeira

vez no Brasil.

De coração, agradeço a Priscila e família, a Simone e seu filho Adriel, pelo carinho e acolhida brindados durante minha estadia em Campina Grande e João Pessoa.

Meus agradecimentos aos meus colegas do mestrado, Amanda, Débora, Elizabeth, Jurandir, Valdemiro, Alexandre, Guaraci, George, Robson e Wesley pela amizade, comprenssão, carinho e na troca de ideias.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro concedido para a realização deste trabalho.

RESUMO

Um dos procedimentos usuais para estudar uma relação entre variáveis é análise de regressão. O modelo de regressão usual ajusta os dados sob a suposição de que as variáveis explicativas são medidas sem erros. Porém, em diversas situações as variáveis explicativas apresentam erros de medição. Nestes casos são utilizados os modelos com erros de medição. Neste trabalho estudamos um modelo estrutural com erros de medição para observações replicadas. A estimação dos parâmetros dos modelos propostos foi efetuada pelos métodos de máxima verossimilhança e de máxima pseudoverossimilhança. O comportamento dos estimadores de alguns parâmetros foi analisado por meio de simulações para diferentes números de réplicas. Além disso, são propostos o teste da razão de verossimilhanças, o teste de Wald, o teste escore, o teste gradiente, o teste $C(\alpha)$ de Neyman e o teste da razão de pseudoverossimilhanças com o objetivo de testar algumas hipóteses de interesse relacionadas aos parâmetros. As estatísticas propostas são avaliadas por meio de simulações. Finalmente, o modelo foi ajustado a um conjunto de dados reais referentes a medições de concentrações de elementos químicos em amostras de cerâmicas egípcias. A implementação computacional foi desenvolvida em linguagem R.

Palavras-chave: Erros heteroscedásticos, matriz de covariâncias, máxima pseudoverossimilhança, máxima verossimilhança, modelos com erros de medição, modelo estrutural.

Abstract

The usual regression model fits data under the assumption that the explanatory variable is measured without error. However, in many situations the explanatory variable is observed with measurement errors. In these cases, measurement error models are recommended. We study a structural measurement error model for replicated observations. Estimation of parameters of the proposed models was obtained by the maximum likelihood and maximum pseudolikelihood methods. The behavior of the estimators was assessed in a simulation study with different numbers of replicates. Moreover, we proposed the likelihood ratio test, Wald test, score test, gradient test, Neyman's $C(\alpha)$ test and pseudolikelihood ratio test in order to test hypotheses of interest related to the parameters. The proposed test statistics are assessed through a simulation study. Finally, the model was fitted to a real data set comprising measurements of concentrations of chemical elements in samples of Egyptian pottery. The computational implementation was developed in R language.

Keywords: Heteroscedastic errors, covariance matrix, maximum pseudolikelihood, maximum likelihood, measurement error models, structural model.

Sumário

Capítulo 1

Introdução

1.1 Motivação

O problema prático que motivou este trabalho está presente em diversas áreas do conhecimento, como por exemplo em Agronomia. Suponha que estamos relacionando, por meio de um modelo, as variáveis rendimento na produção de um determinado cereal e teor de nitrogênio no solo (Fuller, 1987). Em Medicina, podemos estudar a relação entre impedância cardiográfica (IC) e ventriculografia radioisotópica (RV), que são utilizadas para a medição do débito cardíaco em indivíduos (Carstensen *et al.*, 2012); ou estudar a relação entre as variáveis dosagem de uma determinada droga e nível de proteína na urina (Barnett, 1970) ou o desempenho acadêmico e atlético de crianças de diferentes grupos de idades em duas localizações distintas (Dolby *et al.*, 1987). Nestes casos as variáveis envolvidas não podem ser medidas com exatidão. Uma alternativa na análise para cada um dos conjuntos de dados acima seria a utilização de modelos de regressão com erros de medição ou modelos com erros nas variáveis. Estes modelos estendem os modelos de regressão usual procurando representar as variáveis explicativas de uma forma mais realista (Cheng & Van Ness, 1999). Além disso, esses erros podem acontecer devido a várias circunstâncias, como por exemplo, a leitura incorreta nos instrumentos de medição; métodos

e técnicas de registro de dados, isto é, coletas por entrevistas ou questionários, pois os erros podem ser ocasionados por desonestidade, ignorância ou por falta de cuidado e nas condições ambientais que geram a variabilidade em instrumentos de leitura (Buonaccorsi, 2010).

Em problemas envolvendo validação de métodos de medição, cuja finalidade é verificar se os procedimentos analíticos irão fornecer resultados equivalentes das medições obtidas. Em outras palavras, a validação de métodos é a verificação da equivalência entre eles. Carstensen (2010) apresentou algumas metodologias e ferramentas práticas em estudos de comparação de métodos de medições clínicas. Alguns exemplos e aplicações a conjuntos de dados reais podem ser encontrados, por exemplo, em Galea-Rojas *et al.* (2003) e Carstensen *et al.* (2012). Como aplicação da metodologia estudada nesta dissertação, utilizaremos um conjunto de dados reais relacionado ao problema em questão.

Existe uma ampla cobertura de diversos tópicos referentes aos modelos de regressão com erros de medição (vide Fuller, 1987; Cheng & Van Ness, 1999; Stefanski, 2000; Carroll *et al.*, 2006; Buonaccorsi, 2010, por exemplo). Estes modelos envolvem erros de medição nas variáveis explicativas. Se uma análise estatística é realizada ignorando a presença dos erros de medida, algumas inferências podem resultar não confiáveis. Recentemente, modelos de regressão com erros de medição heteroscedásticos sob enfoque estrutural têm sido objeto de pesquisa, como por exemplo, podemos citar Galea *et al.* (2008), Patriota *et al.* (2009) e de Castro *et al.* (2013). Em diversas aplicações supõe-se que os erros de medição são descorrelacionados e suas variâncias são conhecidas (Kulathinal *et al.*, 2002).

Nesta dissertação consideramos uma única covariável medida com erro e situações em que réplicas estão disponíveis tanto da variável explicativa quanto da variável resposta, que por sua vez estas podem ser balanceadas ou desbalanceadas. Quando as medições são replicadas, presume-se que a base do valor verdadeiro não se altera durante a replicação e não existe nenhum emparelhamento entre as medições repetidas (Sengupta, 2012, Cap. 3). Além disso, consideramos duas situações em relação à estrutura dos erros de medi-

ção, sejam heteroscedásticos ou homoscedásticos. Assim, o presente trabalho tem como objetivos:

(1) Revisar os métodos inferenciais que podem ser utilizados em modelos de regressão com erros de medição sob enfoque estrutural, tanto heteroscedásticos quanto homoscedásticos, para observações replicadas (Dolby, 1976; Chan & Mak, 1979; Gong & Samaniego, 1981; Carroll *et al.*, 1993; Artes & Botter, 2005; Carroll *et al.*, 2006; Guolo, 2011).

(2) Obter os estimadores de máxima verossimilhança (MV) e de máxima pseudoverossimilhança (MPV) dos parâmetros de interesse dos modelos propostos e compará-los.

(3) Estudar estatísticas para testar a homoscedasticidade das variâncias dos erros de medição e testes envolvendo simultaneamente os coeficientes angular e linear.

(4) Realizar um estudo de simulação sobre as propriedades assintóticas dos estimadores e estatísticas de teste desenvolvidos, considerando que o tamanho da amostra é fixo e que o número de réplicas das observações aumenta.

(5) Como ilustração, aplicamos os modelos estudados a um conjunto de dados reais referentes a medições de potássio (K) encontrado em amostras de cerâmicas egípcias, e em que foram utilizadas duas técnicas de medição (mais detalhes sobre o conjunto de dados reais encontram-se em Rasekh & Fieller, 2003).

Na Seção 1.1 apresentamos alguns conceitos básicos dos modelos com erros de medição. Resultados sobre o estimador de máxima pseudoverossimilhança e sua distribuição assintótica são descritos na Seção 1.2. Na subseção 1.3 faremos uma revisão bibliográfica a respeito de modelos com erros de medição para dados com e sem réplicas. Finalmente, na Seção 1.4 apresentamos a organização do trabalho.

1.2 Conceitos básicos

Uma extensão dos modelos de regressão usuais são os modelos de regressão com erros de medição, que são caracterizados pelo fato de que há covariáveis do modelo medidas com erros (Fuller, 1987; Cheng & Van Ness, 1999). Considerando o modelo linear usual com erros de medição, assumimos que as variáveis y e x são relacionadas por

$$y_i = \beta_0 + \beta_1 x_i, \tag{1.1}$$

para $i = 1, \ldots, n$, em que y_i e x_i não são observados diretamente. É importante destacar que a variável resposta e a covariável estão sujeitas a erros aditivos, dados por ε_i e δ_i, respectivamente, isto é, o que observamos de fato é

$$X_i = x_i + \delta_i \quad \text{e} \quad Y_i = y_i + \varepsilon_i \ , \tag{1.2}$$

em que (X_i, Y_i) independentes, $i = 1, \ldots, n$, os erros de medição δ_i e ε_i são independentes entre si e x_i é independente de $(\varepsilon_i, \delta_i)$. Esses erros têm médias nulas e variâncias finitas e não nulas.

Em algumas situações, como em Kulathinal *et al.* (2002), os erros de medição $(\delta_i, \varepsilon_i)$ têm distribuição

$$\begin{pmatrix} \delta_i \\ \varepsilon_i \end{pmatrix} \overset{ind}{\sim} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\delta_i} & 0 \\ 0 & \sigma^2_{\varepsilon_i} \end{bmatrix} \right), \quad i = 1, \ldots, n, \tag{1.3}$$

com variâncias $\sigma^2_{\delta_i}$ e $\sigma^2_{\varepsilon_i}$ conhecidas e maiores do que 0. O modelo (1.1)-(1.3) é classificado em modelo de regressão heteroscedástico, quando as variâncias dos erros variam de observação para observação em (1.3), e se estas variâncias são $\sigma^2_{\delta_i} = \sigma^2_{\delta}$ e $\sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon}$ para $i = 1, \ldots, n$, é denominado de modelo de regressão homoscedástico (caso particular do modelo anterior). Além disso, três situações de modelagem podem ser abordadas para a covariável não observada x. De acordo com as suposições encontradas na literatura, podemos classificar o modelo com erros de medição dado em (1.1)-(1.3), como

Modelo funcional: Se cada x_i é uma constante desconhecida.

Modelo estrutural: Se x_i é uma variável aleatória com média μ_x e variância σ_x^2 .

Modelo ultraestrutural: Se cada x_i é uma variável aleatória independente, mas não identicamente distribuída, podendo ter diferentes médias μ_{x_i} e variância σ_x^2 para todo i. Este modelo é uma generalização dos modelos funcional e estrutural. Assim, se $\mu_{x_1} = \cdots = \mu_{x_n} = \mu_x$, o modelo ultraestrutural se reduz ao modelo estrutural e se $\sigma_x^2 = 0$, o modelo ultraestrutural se reduz ao modelo funcional (Dolby, 1976).

Nessas condições, o modelo de regressão linear usual é um caso particular do modelo de regressão com erros de medição quando os erros δ_i's são todos nulos para $i =, 1 \ldots, n$. Se tentarmos escrever (1.1)-(1.3) como um modelo de regressão usual, temos

$$\begin{aligned} Y =& y + \varepsilon = \beta_0 + \beta_1 x + \varepsilon = \beta_0 + \beta_1(X - \delta) + \varepsilon \\ =& \beta_0 + \beta_1 X + (\varepsilon - \beta_1\delta) = \beta_0 + \beta_1 X + \varepsilon_a \end{aligned} \tag{1.4}$$

em que $\varepsilon_a = \varepsilon - \beta_1\delta$. No entanto, não se trata do modelo de regressão usual, devido a que X é uma variável aleatória e é correlacionada com o termo do ε_a.

Surgem problemas ligados à consistência dos estimadores ao estimar os parâmetros dos modelos com erros nas variáveis. Em modelos funcionais e ultraestruturais, devido à presença dos parâmetros incidentais, estimadores de máxima verossimilhança (MV) podem não existir, e caso existam, podem não ser consistentes (função verossimilhança ilimitada). Nos modelos estruturais, os problemas de inconsistência ocorrem devido à falta de identificabilidade do modelo. Modelos inidentificáveis permitem que diferentes conjuntos de valores para os parâmetros originem a mesma distribuição para X e Y (Fuller, 1987). Assim, sob o enfoque estrutural de modelos de regressão com erros nas variáveis sem réplicas assume-se alguma suposição adicional sobre os parâmetros para tornar possíveis as inferências (Cheng & Van Ness, 1999). No entanto, em diversas situações encontradas na literatura as variâncias dos erros de medição em (1.3) são estimadas utilizando réplicas

das observações X e Y. No Capítulo 2, serão apresentados os modelos de regressão com erros de medição sob enfoque estrutural para obervações replicadas.

1.2.1 Método de máxima pseudoverossimilhança

Na estimação dos parâmetros dos modelos que serão apresentados utilizamos uma alternativa baseada nas modificações do método de máxima verossimilhança, sendo esta o método de máxima pseudoverossimilhança (MPV), descrito em Gong & Samaniego (1981). Este método de estimação é utilizado na inferência em um determinado modelo que envolve apenas alguns dos parâmetros mas não todos, isto é, quando há presença de parâmetros de perturbação e de interesse. Um atrativo desta abordagem ocorre quando não é factível eliminar o parâmetro de perturbação por condicionamento ou fatoração (Gong & Samaniego, 1981; Carroll *et al.*, 2006; Guolo, 2011).

Consideramos $\mathcal{S}$ o espaço amostral sobre o qual definimos uma família de distribuições $\mathcal{D}(\boldsymbol{\theta})$ indexadas por um vetor de parâmetros desconhecidos $\boldsymbol{\theta}$, com $\dim(\boldsymbol{\theta}) = p_1 + p_2$. Além disso, consideramos uma partição para $\boldsymbol{\theta} = (\boldsymbol{\phi}^t, \boldsymbol{\lambda}^t)^t$, em que $\boldsymbol{\phi}$ e $\boldsymbol{\lambda}$ são os vetores de parâmetros de interesse e de perturbação, com dimensões p_1 e p_2, respectivamente. O símbolo "t" denota a transposta de um vetor ou matriz.

Seja $\boldsymbol{Z} = (\boldsymbol{Z}_1^t, \ldots, \boldsymbol{Z}_n^t)^t$ uma amostra aleatória da família de distribuições $\mathcal{D}(\boldsymbol{\phi}, \boldsymbol{\lambda})$, com parâmetros $(\boldsymbol{\phi}, \boldsymbol{\lambda}) \in \boldsymbol{\Phi} \times \boldsymbol{\Lambda} \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ respectivamente. Seja $f_{\boldsymbol{Z}}$ a função densidade da amostra aleatória $\boldsymbol{Z}$ e cuja função verossimilhança correspondente à amostra observada é dada por

$$L(\boldsymbol{\phi}, \boldsymbol{\lambda}) = \prod_{i=1}^{n} f_{\boldsymbol{Z}}(\boldsymbol{Z}_i; \boldsymbol{\phi}, \boldsymbol{\lambda}), \tag{1.5}$$

cujo logaritmo é dado por

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\lambda}; \boldsymbol{Z}) = \sum_{i=1}^{n} \log f_{\boldsymbol{Z}}(\boldsymbol{Z}_i; \boldsymbol{\phi}, \boldsymbol{\lambda}). \tag{1.6}$$

Definição 1.2.1 Seja $\widetilde{\boldsymbol{\lambda}}$ um estimador consistente do parâmetro $\boldsymbol{\lambda}$. Definimos

$$\mathcal{L}_p(\boldsymbol{\phi}; \boldsymbol{Z}) = \mathcal{L}(\boldsymbol{\phi}, \widetilde{\boldsymbol{\lambda}}; \boldsymbol{Z}),$$

como sendo a função logpseudoverossimilhança de $\boldsymbol{\phi}$.

Na realidade, $\mathcal{L}_p$ representa a função logverossimilhança dada em (1.6) com $\widetilde{\boldsymbol{\lambda}}$ fixado, isto é, $\mathcal{L}_p$ é agora uma função somente de $\boldsymbol{\phi}$. Em virtude dessas considerações, duas definições encontradas na literatura referentes ao estimador de máxima pseudoverossimilhança para o vetor de parâmetros de interesse são apresentadas.

Definição 1.2.2 Seja $\boldsymbol{Z} = (\boldsymbol{Z}_1^t, \ldots, \boldsymbol{Z}_n^t)^t$ uma amostra aleatória, com $\boldsymbol{Z} \sim \mathcal{D}(\boldsymbol{\phi}, \boldsymbol{\lambda})$. Seja $\widetilde{\boldsymbol{\lambda}}$ um estimador consistente do parâmetro $\boldsymbol{\lambda}$ baseado na amostra $\boldsymbol{Z}$. Se

$$\widetilde{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \mathcal{L}_p(\boldsymbol{\phi}; \boldsymbol{Z}),$$

dizemos que $\widetilde{\boldsymbol{\phi}}$ é um estimador de máxima pseudoverossimilhança do parâmetro $\boldsymbol{\phi}$.

Definição 1.2.3 Seja $\boldsymbol{Z} = (\boldsymbol{Z}_1^t, \ldots, \boldsymbol{Z}_n^t)^t$ uma amostra aleatória, em que $\boldsymbol{Z}_i = (\boldsymbol{X}_i^t,\ \boldsymbol{Y}_i^t)^t$ com $\boldsymbol{Z}_i \sim \mathcal{D}(\boldsymbol{\phi}, \boldsymbol{\lambda})$ e $\boldsymbol{X}_i \sim \mathcal{D}(\boldsymbol{\lambda})$ para $i, \ldots, n$. Seja $\widetilde{\boldsymbol{\lambda}}$ um estimador consistente do parâmetro $\boldsymbol{\lambda}$ baseado na amostra $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Se

$$\widetilde{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \mathcal{L}_p(\boldsymbol{\phi}; \boldsymbol{Z}),$$

dizemos que $\widetilde{\boldsymbol{\phi}}$ é um estimador de máxima pseudoverossimilhança do parâmetro $\boldsymbol{\phi}$.

As definições 1.1.2 e 1.1.3 são formuladas segundo os autores Gong & Samaniego (1981) e Guolo (2011), respectivamente. Além disso, um fato importante é que as funções logpseudoverossimilhanças não são únicas, elas variam de acordo com o estimador encontrado para o parâmetro de perturbação $\boldsymbol{\lambda}$.

Outro método que contorna a presença de parâmetros de perturbação é o método de máxima verossimilhança perfilada, mas este difere da abordagem tratada aqui. Para explicar brevemente este método, consideremos $\boldsymbol{\theta} = (\boldsymbol{\phi}^t, \boldsymbol{\lambda}^t)^t$ e supomos que $\boldsymbol{\phi}$ é fixo.

Reescrevemos a função logverossimilhança como $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\lambda}) = \mathcal{L}_{\boldsymbol{\phi}}(\boldsymbol{\lambda})$ para mostrar que $\boldsymbol{\phi}$ é fixo, mas $\boldsymbol{\lambda}$ varia. Assim, para estimar $\boldsymbol{\lambda}$ maximiza-se $\mathcal{L}_{\boldsymbol{\phi}}(\boldsymbol{\lambda})$ com respeito a $\boldsymbol{\lambda}$, isto é, $\widehat{\boldsymbol{\lambda}}_{\phi} = \arg\max_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \mathcal{L}_{\phi}(\boldsymbol{\lambda})$. No entanto, $\boldsymbol{\phi}$ é desconhecido, por conseguinte, para cada $\boldsymbol{\phi}$ tem-se a substituição de $\boldsymbol{\lambda}$ por $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{\phi}}$. O estimador de máxima verossimilhança perfilada de $\boldsymbol{\phi}$ é obtido por

$$\widehat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \mathcal{L}_{\phi}(\widehat{\boldsymbol{\lambda}}_{\boldsymbol{\phi}}) = \arg\max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \mathcal{L}(\boldsymbol{\phi}, \widehat{\boldsymbol{\lambda}}_{\boldsymbol{\phi}}),$$

da última expressão notemos que $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{\phi}}$ é obtida quando $\boldsymbol{\phi}$ é suposto fixo, ao contrário da abordagem do método de máxima pseudoverossimilhança, pois $\widetilde{\boldsymbol{\phi}}$ simplesmente requer um estimador consistente para $\boldsymbol{\lambda}$. Portanto, além do método de máxima verossimilhança (MV), utilizamos o método de máxima pseudoverossimilhança (MPV) na inferência sobre os parâmetros dos modelos que serão propostos no Capítulo 2.

1.2.2 Distribuição assintótica do estimador MPV

Nesta subseção utilizamos a teoria de equações de estimação para encontrar a distribuição assintótica dos estimadores de máxima pseudoverossimilhança (Carroll *et al.*, 2006; Guolo, 2011). Consideremos uma amostra aleatória $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, como na definição 1.1.2, em que cada i-ésima unidade amostral associa-se uma parcela da função de estimação, Ψ_i. O conceito de função de estimação para uma amostra $\boldsymbol{Z}$ é dado através da expressão

$$\boldsymbol{\Psi}(\boldsymbol{Z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{\Psi}_i(\boldsymbol{Z}_i, \boldsymbol{\theta}). \tag{1.7}$$

Restringimos (1.7) de modo que

$$n^{-1}\sum_{i=1}^{n} \boldsymbol{\Psi}_i(\boldsymbol{Z}_i, \widetilde{\boldsymbol{\theta}}) = \mathbf{0}, \tag{1.8}$$

é denominada de equação de estimação, com $\widetilde{\boldsymbol{\theta}}$ uma solução para (1.8) e ao mesmo tempo é um M-estimador para $\boldsymbol{\theta}$ (para mais detalhes, vide Artes & Botter, 2005, por exemplo). A equação de estimação dada em (1.8) é dita não viciada se $\mathrm{E}\{\sum_{i=1}^{n} \boldsymbol{\Psi}_i(\boldsymbol{Z}_i, \widetilde{\boldsymbol{\theta}})\} = \mathbf{0}$.

Logo, se as equações de estimação são não viciadas, então $\widetilde{\boldsymbol{\theta}}$ é um estimador consistente para $\boldsymbol{\theta}$, sob condições de regularidade (Gong & Samaniego, 1981; Delgado, 1995). No caso em que as amostras são independentes e identicamente distribuídas (*iid*), a ideia é que para cada valor de $\boldsymbol{\theta}$, (1.8) converge para sua própria esperança (isso acontece pela lei dos grandes números). Assim, pelo fato de que $\widetilde{\boldsymbol{\theta}}$ é consistente e por aproximações de Taylor, tem-se

$$0 \approx n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_i(\boldsymbol{Z}_i,\boldsymbol{\theta}) + n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}^t}\boldsymbol{\Psi}_i(\boldsymbol{Z}_i,\boldsymbol{\theta})(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}), \tag{1.9}$$

sendo $\boldsymbol{\theta}$ o verdadeiro valor do parâmetro. Daqui, aplicando novamente a lei dos grandes números ao termo entre parênteses do lado direito em (1.9), obtém-se uma aproximação da diferença entre o estimador e o verdadeiro valor do parâmetro dada por

$$\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta} \approx -\left\{\boldsymbol{A}(\boldsymbol{\theta})\right\}^{-1} n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_i(\boldsymbol{Z}_i,\boldsymbol{\theta}), \tag{1.10}$$

com

$$\boldsymbol{A}(\boldsymbol{\theta}) = n^{-1}\textstyle\sum_{i=1}^{n}\mathrm{E}\left\{\dfrac{\partial}{\partial\boldsymbol{\theta}^t}\boldsymbol{\Psi}_i(\boldsymbol{Z}_i,\boldsymbol{\theta})\right\}. \tag{1.11}$$

Utilizando as propriedades da distribuição normal multivariada, segue-se que $\widetilde{\boldsymbol{\theta}}$ tem distribuição normal assintótica com média $\boldsymbol{\theta}$ e matriz de covariâncias dada por

$$n^{-1}\{\boldsymbol{A}(\boldsymbol{\theta})^{-1}\}\{\boldsymbol{B}(\boldsymbol{\theta})\}\{\boldsymbol{A}(\boldsymbol{\theta})^{-1}\}^t, \tag{1.12}$$

em que

$$\boldsymbol{B}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\mathrm{cov}\{\boldsymbol{\Psi}_i(\boldsymbol{Z}_i,\boldsymbol{\theta})\}. \tag{1.13}$$

Consideramos uma partição $\boldsymbol{\theta} = (\boldsymbol{\phi}^t,\boldsymbol{\lambda}^t)^t$ como na Seção 1.2.1 e a estimação pelo método de máxima pseudoverossimilhança conforme apresentado anteriormente. Este método requer um estimador para o vetor de parâmetros de perturbação $\boldsymbol{\lambda}$. Com efeito, seja $\widetilde{\boldsymbol{\lambda}}$ obtido como solução de

$$\sum_{i=1}^{n}\boldsymbol{\Psi}_i(\boldsymbol{Z}_i,\boldsymbol{\lambda}) = \mathbf{0}. \tag{1.14}$$

Em seguida, estima-se o vetor de parâmetros de interesse $\boldsymbol{\phi}$, que resolve a equação de estimação obtida pela derivada primeira da função logpseudoverossimilhança, dada por

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\phi}} \mathcal{L}_{p,i}(\boldsymbol{\phi}, \widetilde{\boldsymbol{\lambda}}; \boldsymbol{Z}_i) = \mathbf{0}, \tag{1.15}$$

em que $\widetilde{\boldsymbol{\lambda}}$ é fixado em (1.15) e obtido de (1.14). Se a equação de estimação (1.15) for não viciada, temos que $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\lambda}}^t, \widetilde{\boldsymbol{\phi}}^t)^t$ é uma solução simultânea de (1.14)-(1.15) e podemos encontrar a distribuição assintótica de $(\widetilde{\boldsymbol{\lambda}}^t, \widetilde{\boldsymbol{\phi}}^t)^t$, já que podemos escrever uma única equação de estimação, isto é,

$$\sum_{i=1}^{n} \begin{pmatrix} \boldsymbol{\Psi}_i(\boldsymbol{Z}_i, \boldsymbol{\lambda}) \\ \dfrac{\partial}{\partial \boldsymbol{\phi}} \mathcal{L}_{p,i}(\boldsymbol{\phi}, \widetilde{\boldsymbol{\lambda}}; \boldsymbol{Z}_i) \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}. \tag{1.16}$$

Utilizando a aproximação dada em (1.12), particionamos as matrizes $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta})$, $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{\theta})$ e $\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-t}$, de acordo com as dimensões dos parâmetros de interesse e de perturbação, $p_1 = \dim(\boldsymbol{\phi})$ e $p_2 = \dim(\boldsymbol{\lambda})$. Por consequência, a matriz de covariâncias assintótica de $\widetilde{\boldsymbol{\phi}}$ é n^{-1} vezes a submatriz inferior direita de $\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-t}$ (Carroll *et al.*, 2006), e é dada por

$$\boldsymbol{\Sigma} = \boldsymbol{A}_{\phi\phi}^{-1}(\boldsymbol{B}_{\phi\phi} - \boldsymbol{A}_{\phi\lambda}\boldsymbol{A}_{\lambda\lambda}^{-1}\boldsymbol{B}_{\lambda\phi} - \boldsymbol{B}_{\lambda\phi}^{t}\boldsymbol{A}_{\lambda\lambda}^{-1}\boldsymbol{A}_{\phi\lambda}^{t} + \boldsymbol{A}_{\phi\lambda}\boldsymbol{A}_{\lambda\lambda}^{-1}\boldsymbol{B}_{\lambda\lambda}\boldsymbol{A}_{\lambda\lambda}^{-1}\boldsymbol{A}_{\phi\lambda}^{t})\boldsymbol{A}_{\phi\phi}^{-1}, \tag{1.17}$$

cujos blocos são

$$\begin{aligned}
&\boldsymbol{A}_{\phi\phi} = -\sum_{i=1}^{n} \mathrm{E}\left\{\frac{\partial^2 \mathcal{L}_p(\boldsymbol{\phi}, \boldsymbol{\lambda}; \boldsymbol{Z}_i)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^t}\right\}, \quad \boldsymbol{A}_{\lambda\lambda} = \sum_{i=1}^{n} \mathrm{E}\left\{\frac{\partial \boldsymbol{\Psi}_i(\boldsymbol{\lambda}; \boldsymbol{Z}_i)}{\partial \boldsymbol{\lambda}}\right\}, \\
&\boldsymbol{A}_{\phi\lambda} = -\sum_{i=1}^{n} \mathrm{E}\left\{\frac{\partial^2 \mathcal{L}_{p,i}(\boldsymbol{\phi}, \boldsymbol{\lambda}; \boldsymbol{Z}_i)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\lambda}^t}\right\}, \ \boldsymbol{B}_{\phi\phi} = \sum_{i=1}^{n} \frac{\partial \mathcal{L}_{p,i}(\boldsymbol{\phi}, \boldsymbol{\lambda}; \boldsymbol{Z}_i)}{\partial \boldsymbol{\phi}} \left\{\frac{\partial \mathcal{L}_{p,i}(\boldsymbol{\phi}, \boldsymbol{\lambda}; \boldsymbol{Z}_i)}{\partial \boldsymbol{\phi}}\right\}^t, \\
&\boldsymbol{B}_{\lambda\phi} = \sum_{i=1}^{n} \boldsymbol{\Psi}_i(\boldsymbol{\lambda}) \left\{\frac{\partial \mathcal{L}_{p,i}(\boldsymbol{\phi}, \boldsymbol{\lambda}; \boldsymbol{Z}_i)}{\partial \boldsymbol{\phi}}\right\}^t \quad \text{e} \quad \boldsymbol{B}_{\lambda\lambda} = \sum_{i=1}^{n} \boldsymbol{\Psi}_i(\boldsymbol{\lambda}; \boldsymbol{Z}_i) \left\{\boldsymbol{\Psi}_i(\boldsymbol{\lambda}; \boldsymbol{Z}_i)\right\}^t.
\end{aligned} \tag{1.18}$$

Portanto, a distribuição assintótica do estimador de máxima pseudoverossimilhança do vetor de parâmetros de interesse $\boldsymbol{\phi}$, sob condições de regularidade (Gong & Samaniego,

1981; Delgado, 1995) é

$$\sqrt{n}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{d} N_{p_1}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{1.19}$$

em que $\boldsymbol{\Sigma}$ é a matriz de covariâncias dada em (1.17).

Cada um dos blocos da matriz de covariâncias $\boldsymbol{\Sigma}$ do estimador MPV do parâmetro de interesse $\boldsymbol{\phi}$, pode ser obtido por meio do estimador sanduíche (Artes & Botter, 2005; Carroll *et al.*, 2006). Neste trabalho, adaptamos esta abordagem para o caso em que as observações são independentes, mas não identicamente distribuídas e utilizamos a teoria de equação de estimação apresentada a fim de apresentar em forma geral a distribuição assintótica do estimador de MPV de $\boldsymbol{\phi}$.

1.3 Breve revisão da literatura

Como mencionado na Seção 1.1, existem diversos trabalhos publicados sobre modelos com erro de medição (linear ou não linear), motivados por problemas das diversas áreas de conhecimento (Agronomia, Medicina, Química e Econometria, dentre outras). Referências, métodos e exemplos de conceitos básicos para esta literatura podem ser encontradas em Fuller (1987), Cheng & Van Ness (1999), Carroll *et al.* (2006), Buonaccorsi (2010).

Barnett (1970) estudou um modelo linear heteroscedástico sob enfoque funcional com observações replicadas. Em outras palavras, considerou a seguinte relação entre as variáveis observadas: $Y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}$ e $X_{ij} = x_i + \delta_{ij}$, para $j = 1, \ldots, r_i$ e os erros ε_{ij} e δ_{ij} são não correlacionados para $i = 1, \ldots, n$. O autor derivou os estimadores de máxima verossimilhança para os parâmetros β_0 e β_1, assim como também sua respectiva matriz de covariâncias assintótica. Dolby *et al.* (1987) também derivam os estimadores de máxima verossimilhança considerando o mesmo modelo de Barnett (1970), mas flexibilizando o número de réplicas para cada variável (X e Y), isto é, Y_{ij} e X_{ik} para $j = 1, \ldots, s_i$ e $k = 1, \ldots, r_i$, respectivamente.

Chan & Mak (1979) estudaram o modelo normal com erros homoscedásticos, mas com

número de réplicas constante, obtendo os estimadores de máxima verossimilhança para os parâmetros do modelo e a matriz de covariâncias assintótica dos estimadores para os parâmetros de intercepto e inclinação.

Guolo (2011) utilizou o algoritmo EM-Monte Carlo para estimar os parâmetros pelo método de máxima pseudoverossimilhança em modelos de regressão com covariáveis afetados por erros de medição. Na literatura, existem outros trabalhos publicados que consideram outras abordagens a respeito do método de MPV, por exemplo, Geys *et al.* (2006) utilizam um método de estimação por pseudoverossimilhança, mas não é o método tratado aqui. A ideia principal é substituir uma função de densidade conjunta numericamente difícil por uma função mais simples, isto é, pode ser substituída por um produto adequado de razões de verossimilhanças de subconjuntos das variáveis. Por exemplo, uma distribuição bivariada pode ser substituída pelo produto de duas funções condicionais.

1.4 Organização do trabalho

Este trabalho encontra-se dividido em seis capítulos. No Capítulo 2 apresentamos o modelo de regressão na presença de erros de medição sob enfoque estrutural para observações replicadas e segundo a natureza dos erros de medição (heteroscedásticos ou homoscedásticos). Apresentaremos a matriz de informação esperada dos parâmetros. No Capítulo 3 desenvolvemos as inferências baseadas nos métodos de máxima verossimilhança e de máxima pseudoverossimilhança. Expressões para as matrizes de covariâncias assintóticas dos estimadores de MV e MPV foram obtidas, além de estatísticas de teste para testar as hipóteses de homoscedasticidade das variâncias dos erros de medição e ausência de vieses aditivo e multiplicativo. No Capítulo 4 realizamos um estudo de simulação utilizando a linguagem R (R Core Team, 2013), para avaliar o comportamento dos estimadores e desempenho dos testes de hipóteses de interesse. A situação em que o número de observações n permanece fixo e o número de réplicas tende a infinito. O comportamento das estatísticas da razão de verossimilhanças, Wald, escore, gradiente, C(α) de Neyman e

da razão de pseudoverossimilhanças é analisado. No Capítulo 5 aplicamos a metodologia apresentada nos Capítulos 2 e 3 a um conjunto de dados reais referente a medições de potássio (K) encontrado em amostras de cerâmicas egípcias. Por último, são apresentadas as conclusões e propostas de trabalhos futuros no Capítulo 6.

Capítulo 2

Modelos

Neste capítulo inserimos as réplicas na estrutura do modelo, modificando o modelo de regressão (1.1)-(1.3) dado na Seção 1.2. Deste modo, o modelo é dito modelo de regressão com erros de medição sob enfoque estrutural para observações replicadas. Descrevemos seus casos particulares segundo as estruturas das variâncias dos erros de medição. Consideramos que o número de réplicas das observações são balanceadas e desbalanceadas. Além disso, apresentamos a matriz de informação esperada dos parâmetros para cada modelo descrito a seguir.

2.1 Modelo de regressão com erros de medição sob enfoque estrutural para observações replicadas

Supomos $r_i > 1$ e $s_i > 1$ réplicas das variáveis explicativa e resposta, X_{ik} e Y_{ij} respectivamente. Assim, a partir do modelo (1.1)-(1.3), escrevemos

$$y_i = \beta_0 + \beta_1 x_i, \tag{2.1}$$

$$Y_{ij} = y_i + \varepsilon_{ij} \tag{2.2}$$

$$\text{e} \quad X_{ik} = x_i + \delta_{ik}, \tag{2.3}$$

em que os erros de medição ε_{ij} e δ_{ik} são independentes entre si e para diferentes observações, x_i é independente de $(\varepsilon_{ij}, \delta_{ik})$, $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_{\varepsilon_i})$ para $j = 1, \ldots, s_i$ e $\delta_{ik} \overset{iid}{\sim} N(0, \sigma^2_{\delta_i})$ para $k = 1, \ldots, r_i$ e $i = 1, \ldots, n$. Sob o enfoque estrutural, assumimos neste trabalho que a distribuição da covariável não observada x_i é normal, ou seja, $x_i \overset{iid}{\sim} N(\mu_x, \sigma^2_x)$ para $i = 1, \ldots, n$, com $\mu_x \in \mathbb{R}$ e $\sigma^2_x > 0$ representando a média e a variância da i-ésima observação da variável preditora não observável. Esta especificação é conhecida como modelo estrutural normal com erros de medição para observações replicadas. A inferência, neste trabalho, está centrada no vetor de parâmetros de interesse $\boldsymbol{\phi}$, cujas componentes são β_0 e β_1, que representam os coeficientes de regressão linear e angular, respectivamente.

Por outro lado, como comentado na Seção 2.2, os modelos com erros de medição estruturais devem ser tratados levando-se em conta o problema de identificabilidade. Para evitar esse problema nestes modelos com erros de medição foram inseridas as réplicas como parte do modelo a fim de estimar todos os parâmetros do modelo tornando assim o modelo identificável.

A seguir apresentamos os modelos de regressão que surgem a partir de (2.1)-(2.3), segundo a natureza das variâncias dos erros de medição.

2.1.1 Modelo heteroscedástico

Dizemos que o modelo (2.1)-(2.3) é dito modelo heteroscedástico se as variâncias dos erros de medição $\sigma^2_{\delta_i}$ e $\sigma^2_{\varepsilon_i}$ variam de obervação para observação. Por conseguinte, obtemos $Y_{ij} \overset{ind.}{\sim} N(\beta_0 + \beta_1\mu_x,\ \beta_1^2\sigma^2_x + \sigma^2_{\varepsilon_i})$ e $X_{ik} \overset{ind.}{\sim} N(\mu_x, \sigma^2_x + \sigma^2_{\delta_i})$, com

$$\mathrm{Cov}(Y_{ij}, X_{ik}) = \beta_1\sigma^2_x \quad \text{e} \quad \mathrm{Cov}(Y_{lj}, X_{ik}) = 0, \ \ i \neq l, \ \text{para } i, l = 1, \ldots, n.$$

Desta forma definimos o vetor de parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1, \mu_x, \sigma^2_x, \boldsymbol{\sigma}^{2^t}_{\delta}, \boldsymbol{\sigma}^{2^t}_{\varepsilon})^t$, sendo $\boldsymbol{\sigma}^2_{\delta} = (\sigma^2_{\delta_1}, \ldots, \sigma^2_{\delta_n})^t$ e $\boldsymbol{\sigma}^2_{\varepsilon} = (\sigma^2_{\varepsilon_1}, \ldots, \sigma^2_{\varepsilon_n})^t$. Sejam $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ir_i})^t$, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{is_i})^t$ e $\boldsymbol{Z}_i = (\boldsymbol{X}_i^t, \boldsymbol{Y}_i^t)^t$, $i = 1, \ldots, n$. Portanto,

$$\boldsymbol{Z}_i \overset{ind}{\sim} N_{r_i+s_i}(\boldsymbol{m}_i, \boldsymbol{V}_i), \tag{2.4}$$

em que $\boldsymbol{m}_i = (\mu_x \mathbf{1}_{r_i}^t, (\beta_0 + \beta_1 \mu_x)\mathbf{1}_{s_i}^t)^t$ é o vetor de médias e $\boldsymbol{V}_i$ é a matriz de covariâncias de $\boldsymbol{Z}_i$, dada por

$$\boldsymbol{V}_i = \boldsymbol{R}_i + \sigma_x^2 \boldsymbol{b}_i \boldsymbol{b}_i^t, \tag{2.5}$$

sendo

$$\boldsymbol{R}_i = \begin{pmatrix} \sigma_{\delta_i}^2 \boldsymbol{I}_{r_i} & \mathbf{0}_{r_i,s_i} \\ \cdot & \sigma_{\varepsilon_i}^2 \boldsymbol{I}_{s_i} \end{pmatrix} \quad \text{e} \quad \boldsymbol{b}_i = (\mathbf{1}_{r_i}^t, \beta_1 \mathbf{1}_{s_i}^t), \tag{2.6}$$

em que $\mathbf{1}_r$ e $\mathbf{1}_{r,s}$ denotam o vetor e a matriz de 1's de ordem $r \times 1$ e $r \times s$, respectivamente. Portanto, a função logverossimilhança para o vetor de parâmetros $\boldsymbol{\theta}$, dada as n observações do vetor aleatório $\boldsymbol{Z}_i = (\boldsymbol{X}_i^t, \boldsymbol{Y}_i^t)^t$, $i = 1, \ldots, n$, do modelo baseado em (2.4) é dada por

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Z}) = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{Z}_i), \tag{2.7}$$

com

$$\mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{Z}_i) = \text{const.} - \frac{1}{2} \log |\boldsymbol{V}_i| - \frac{1}{2} \boldsymbol{d}_i^t \boldsymbol{V}_i^{-1} \boldsymbol{d}_i, \tag{2.8}$$

em que $\boldsymbol{d}_i = \boldsymbol{Z}_i - \boldsymbol{m}_i$, o determinante de $\boldsymbol{V}_i$ é dado por $|\boldsymbol{V}_i| = \sigma_{\delta_i}^{2r_i} \sigma_{\varepsilon_i}^{2s_i} \sigma_x^2 a_i$, a_i dada em (2.13), e a inversa da matriz $\boldsymbol{V}_i$ é

$$\begin{aligned} \boldsymbol{V}_i^{-1} &= \boldsymbol{R}_i^{-1} - \boldsymbol{R}_i^{-1} \boldsymbol{b}_i (\sigma_x^{-2} + \boldsymbol{b}_i^t \boldsymbol{R}_i^{-1} \boldsymbol{b}_i)^{-1} \boldsymbol{b}_i^t \boldsymbol{R}_i^{-1} \\ &= \begin{pmatrix} \sigma_{\delta_i}^{-2} \boldsymbol{I}_{r_i} - a_i^{-1} \sigma_{\delta_i}^{-4} \mathbf{1}_{r_i,r_i} & -a_i^{-1} \beta_1 \sigma_{\delta_i}^{-2} \sigma_{\varepsilon_i}^{-2} \mathbf{1}_{r_i,s_i} \\ \cdot & \sigma_{\varepsilon_i}^{-2} \boldsymbol{I}_{s_i} - a_i^{-1} \beta_1^2 \sigma_{\varepsilon_i}^{-4} \mathbf{1}_{s_i,s_i} \end{pmatrix}. \end{aligned} \tag{2.9}$$

Após algumas manipulações algébricas podemos expressar a função logverossimilhança (2.7) como

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Z}) = \text{const.} &- \frac{n}{2} \log \sigma_x^2 - \frac{1}{2} \sum_{i=1}^{n} r_i \log \sigma_{\delta_i}^2 - \frac{1}{2} \sum_{i=1}^{n} s_i \log \sigma_{\varepsilon_i}^2 \\ &- \frac{1}{2} \sum_{i=1}^{n} \log a_i - \frac{1}{2} \sum_{i=1}^{n} (h_i - q_i^2 a_i^{-1}), \end{aligned} \tag{2.10}$$

sendo

$$h_i = \sigma_{\delta_i}^{-2} \sum_{k=1}^{r_i} (X_{ik} - \mu_x)^2 + \sigma_{\varepsilon_i}^{-2} \sum_{j=1}^{s_i} (Y_{ij} - \beta_0 - \beta_1 \mu_x)^2, \tag{2.11}$$

$$q_i = \sigma_{\delta_i}^{-2} \sum_{k=1}^{r_i} (X_{ik} - \mu_x) + \sigma_{\varepsilon_i}^{-2} \beta_1 \sum_{j=1}^{s_i} (Y_{ij} - \beta_0 - \beta_1 \mu_x) \quad \text{e} \tag{2.12}$$

$$a_i = \boldsymbol{b}_i^t \boldsymbol{R}_i^{-1} \boldsymbol{b}_i + \sigma_x^{-2} = r_i \sigma_{\delta_i}^{-2} + s_i \beta_1^2 \sigma_{\varepsilon_i}^{-2} + \sigma_x^{-2}. \tag{2.13}$$

Ainda que a função logverossimilhança tenha sido escrita de uma forma algébrica mais simples ainda não deixa de ser complexa.

2.1.2 Modelo homoscedástico

Este modelo surge como caso particular do modelo de regressão anterior. Neste caso as variâncias dos erros de medição são $\sigma_{\delta_i}^2 = \sigma_\delta^2$ e $\sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2$ para $i = 1, \ldots, n$. Neste sentido, no modelo (2.1)-(2.3) substituímos as variâncias $\sigma_{\delta_i}^2$ e $\sigma_{\varepsilon_i}^2$ por σ_δ^2 e σ_ε^2, respectivamente, obtendo $Y_{ij} \overset{iid}{\sim} N(\beta_0 + \beta_1 \mu_x,\ \beta_1^2 \sigma_x^2 + \sigma_\varepsilon^2)$ e $X_{ik} \overset{iid}{\sim} N(\mu_x, \sigma_x^2 + \sigma_\delta^2)$ para $j = 1, \ldots, s_i$, $k = 1, \ldots, r_i$, e covariâncias

$$\mathrm{Cov}(Y_{ij}, X_{ik}) = \beta_1 \sigma_x^2 \quad \text{e} \quad \mathrm{Cov}(Y_{lj}, X_{ik}) = 0, \ \ i \neq l, \ \text{para } i, l = 1, \ldots, n.$$

Portanto, $\boldsymbol{Z}_i \sim N_{r_i + s_i}(\boldsymbol{m}_i, \boldsymbol{V}_i^*)$, com $\boldsymbol{m}_i$ como em (2.4), e $\boldsymbol{V}_i^*$ é a matriz de covariâncias $\boldsymbol{V}_i$ dada em (2.5), mas no lugar das variâncias $\sigma_{\delta_i}^2$ e $\sigma_{\varepsilon_i}^2$ são σ_δ^2 e σ_ε^2, respectivamente. Daqui em diante identificamos com $(\cdot)^*$ todas as expressões envolvidas no modelo de regressão homoscedástico.

O vetor de parâmetros é dado por $\boldsymbol{\theta}^* = (\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_\delta^2, \sigma_\varepsilon^2)^t$. Assim, a função logverossimilhança para as n observações de $\boldsymbol{Z}$ é dada por

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{Z}) = \text{const.} &- \frac{n}{2} \log \sigma_x^2 - \frac{1}{2} \log \sigma_\delta^2 \sum_{i=1}^{n} r_i - \frac{1}{2} \log \sigma_\varepsilon^2 \sum_{i=1}^{n} s_i \\ &- \frac{1}{2} \sum_{i=1}^{n} \log a_i^* - \frac{1}{2} \sum_{i=1}^{n} (h_i^* - q_i^{*2} a_i^{*-1}), \end{aligned} \tag{2.14}$$

em que h_i^*, q_i^* e a_i^* são as expressões consideradas em (2.11)-(2.13), mas com a diferença de que as variâncias $\sigma_{\delta_i}^2$ e $\sigma_{\varepsilon_i}^2$ são substituídas por σ_δ^2 e σ_ε^2, respectivamente.

Se considerarmos $r_i = r$ e $s_i = r$ (réplicas balanceadas), no modelo de regressão homoscedástico tratado aqui, este se reduz ao modelo considerado em Chan & Mak (1979).

É importante comentar que em problemas de validação de métodos de medição, cujo interesse é a equivalência de um método y em relação a outro x, isto é, (2.1) é reescrito como $y = x$. Adotamos $\beta_0 = 0$ e $\beta_1 = 1$ identificando a ausência de vieses aditivo e multiplicativo, respectivamente, que será descrito no Capítulo 3.

2.2 Matriz de informação esperada

Nesta seção, foram utilizados alguns dos resultados encontrados em Dolby (1976) e Chan & Mak (1979) para apresentar de forma explícita a matriz de informação esperada referente dos parâmetros do modelo de regressão descrito e casos particulares, heteroscedástico e homoscedástico. A matriz de informação esperada é importante, já que nos permite encontrar a matriz de covariâncias assintótica dos estimadores de MV, assim como também a construção de intervalos de confiança para os parâmetros do modelo.

Consideremos o vetor escore para o modelo baseado em (2.4), dado por

$$\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{Z}) = \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta}; \boldsymbol{Z}_i) = \sum_{i=1}^{n} \frac{\partial \mathcal{L}_i(\boldsymbol{\theta}; \boldsymbol{Z}_i)}{\partial \boldsymbol{\theta}},$$

com $\mathcal{L}_i(\cdot;\cdot)$ dada em (2.7) e os elementos do vetor $\boldsymbol{U}_i$ são

$$U_{i,\gamma} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{V}_i^{-1}(\boldsymbol{d}_i\boldsymbol{d}_i^t - \boldsymbol{V}_i)\boldsymbol{V}_i^{-1}\boldsymbol{V}_{i,\gamma}\right) - \boldsymbol{d}_{i,\gamma}^t\boldsymbol{V}_i^{-1}\boldsymbol{d}_i, \tag{2.15}$$

em que $\mathrm{tr}(\cdot)$ é o traço, $\gamma \in \{\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_{\delta_i}^2, \sigma_{\varepsilon_i}^2,\ i = 1, \ldots, n\}$, $\boldsymbol{V}_{i,\gamma} = \partial \boldsymbol{V}_i/\partial\gamma$ e $\boldsymbol{d}_{i,\gamma} = \partial \boldsymbol{d}_i/\partial\gamma$. De maneira análoga é feito para o modelo de regressão homoscedástico, trocando $\boldsymbol{\theta}$ por $\boldsymbol{\theta}^*$ e γ por $\gamma^* \in \{\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_\delta^2, \sigma_\varepsilon^2\}$.

As componentes da matriz de informação esperada do modelo de regressão com erros de medição sob enfoque estrutural para observações replicadas, tanto heteroscedástico

quanto homoscedástico, são calculadas com uma adaptação à relação de (Dolby, 1976), dada por

$$\mathcal{I}_{\varphi\gamma} = -\mathrm{E}\left\{\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{Z})}{\partial\varphi\partial\gamma}\right\} = \sum_{i=1}^{n}\left\{\frac{1}{2}\mathrm{tr}\left(\boldsymbol{V}_i^{-1}\boldsymbol{V}_{i,\varphi}\boldsymbol{V}_i^{-1}\boldsymbol{V}_{i,\gamma}\right) + \boldsymbol{d}_{i,\varphi}^t\boldsymbol{V}_i^{-1}\boldsymbol{d}_{i,\gamma}\right\}, \qquad (2.16)$$

em que $\varphi, \gamma \in \{\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_{\delta_i}^2, \sigma_{\varepsilon_i}^2,\ i = 1, \ldots, n\}$. Seguidamente, apresentamos para cada um dos casos do modelo de regressão a respectiva matriz de informação esperada.

2.2.1 Matriz de informação esperada do modelo heteroscedástico

Neste modelo lembramos que $\boldsymbol{\theta} = (\beta_0, \beta_1, \mu_x, \sigma_x^2, \boldsymbol{\sigma}_\delta^{2t}, \boldsymbol{\sigma}_\varepsilon^{2t})^t$, sendo $\boldsymbol{\sigma}_\delta^2 = (\sigma_{\delta_1}^2, \ldots, \sigma_{\delta_n}^2)^t$ e $\boldsymbol{\sigma}_\varepsilon^2 = (\sigma_{\varepsilon_1}^2, \ldots, \sigma_{\varepsilon_n}^2)^t$. Assim, o número de parâmetros é $2n + 4$. A matriz de informação esperada, a partir da relação dada em (2.16), tem a forma

$$\boldsymbol{\mathcal{I}} = \begin{pmatrix}
\mathcal{I}_{\beta_0\beta_0} & \mathcal{I}_{\beta_0\beta_1} & \mathcal{I}_{\beta_0\mu_x} & \mathcal{I}_{\beta_0\sigma_x^2} & \mathcal{I}_{\beta_0\sigma_{\delta_1}^2} & \cdots & \mathcal{I}_{\beta_0\sigma_{\delta_n}^2} & \mathcal{I}_{\beta_0\sigma_{\varepsilon_1}^2} & \cdots & \mathcal{I}_{\beta_0\sigma_{\varepsilon_n}^2} \\
 & \mathcal{I}_{\beta_1\beta_1} & \mathcal{I}_{\beta_1\mu_x} & \mathcal{I}_{\beta_1\sigma_x^2} & \mathcal{I}_{\beta_1\sigma_{\delta_1}^2} & \cdots & \mathcal{I}_{\beta_1\sigma_{\delta_n}^2} & \mathcal{I}_{\beta_1\sigma_{\varepsilon_1}^2} & \cdots & \mathcal{I}_{\beta_1\sigma_{\varepsilon_n}^2} \\
 & & \mathcal{I}_{\mu_x\mu_x} & \mathcal{I}_{\mu_x\sigma_x^2} & \mathcal{I}_{\mu_x\sigma_{\delta_1}^2} & \cdots & \mathcal{I}_{\mu_x\sigma_{\delta_n}^2} & \mathcal{I}_{\mu_x\sigma_{\varepsilon_1}^2} & \cdots & \mathcal{I}_{\mu_x\sigma_{\varepsilon_n}^2} \\
 & & & \mathcal{I}_{\sigma_x^2\sigma_x^2} & \mathcal{I}_{\sigma_x^2\sigma_{\delta_1}^2} & \cdots & \mathcal{I}_{\sigma_x^2\sigma_{\delta_n}^2} & \mathcal{I}_{\sigma_x^2\sigma_{\varepsilon_1}^2} & \cdots & \mathcal{I}_{\sigma_x^2\sigma_{\varepsilon_n}^2} \\
 & & & & \mathcal{I}_{\sigma_{\delta_1}^2\sigma_{\delta_1}^2} & \cdots & \mathcal{I}_{\sigma_{\delta_1}^2\sigma_{\delta_n}^2} & \mathcal{I}_{\sigma_{\delta_1}^2\sigma_{\varepsilon_1}^2} & \cdots & \mathcal{I}_{\sigma_{\delta_1}^2\sigma_{\varepsilon_n}^2} \\
 & & & & & \ddots & \vdots & \vdots & \cdots & \vdots \\
 & & & & & & \mathcal{I}_{\sigma_{\delta_n}^2\sigma_{\delta_n}^2} & \mathcal{I}_{\sigma_{\delta_n}^2\sigma_{\varepsilon_1}^2} & \cdots & \mathcal{I}_{\sigma_{\delta_n}^2\sigma_{\varepsilon_n}^2} \\
 & & & & & & & \mathcal{I}_{\sigma_{\varepsilon_1}^2\sigma_{\varepsilon_1}^2} & \cdots & \mathcal{I}_{\sigma_{\varepsilon_1}^2\sigma_{\varepsilon_n}^2} \\
 & & & & & & & & \ddots & \vdots \\
 \cdot & & & & & & & & & \mathcal{I}_{\sigma_{\varepsilon_n}^2\sigma_{\varepsilon_n}^2}
\end{pmatrix}. \qquad (2.17)$$

Utilizando resultados de Chan & Mak (1979) e após algumas reduções algébricas e

matriciais, conseguimos expressar a matriz $\boldsymbol{\mathcal{I}}$ como

$$\boldsymbol{\mathcal{I}} = \begin{pmatrix} \nu & \mu_x\nu & q & \mathbf{0}_{1\times(2n+1)} \\ & p_{\beta\beta} + \mu_x^2\nu & \mu_x q & \boldsymbol{t}^t \\ & & \sum_{i=1}^{n} \dfrac{a_i - \sigma_x^{-2}}{a_i\sigma_x^2} & \mathbf{0}_{1\times(2n+1)} \\ \cdot & & & \boldsymbol{T} \end{pmatrix}, \tag{2.18}$$

em que

$$p_{\beta\beta} = \sum_{i=1}^{n} 2c_i r_i s_i \left(a_i\sigma_x^2 + 2\beta_1^2 s_i \frac{\sigma_{\delta_i}^2}{r_i\sigma_{\varepsilon_i}^2} \right), \qquad c_i = (2a_i^2\sigma_{\delta_i}^2\sigma_{\varepsilon_i}^2)^{-1},$$

$$\nu = \sum_{i=1}^{n} \left(\frac{s_i}{\sigma_{\varepsilon_i}^2} - \frac{\beta_1^2 s_i^2}{a_i\sigma_{\varepsilon_i}^4} \right), \quad q = \frac{\beta_1}{\sigma_x^2} \sum_{i=1}^{n} \frac{s_i}{a_i\sigma_{\varepsilon_i}^2}, \tag{2.19}$$

$$\boldsymbol{t}^t = \left(\frac{2\beta_1}{\sigma_x^2} \sum_{i=1}^{n} r_i s_i c_i \left(1 + \frac{\beta_1^2 s_i \sigma_{\delta_i}^2}{r_i\sigma_{\varepsilon_i}^2} \right),\ \boldsymbol{t}_1^t,\ \boldsymbol{t}_2^t \right),$$

$$\boldsymbol{t}_1^t = \left(-2c_1 r_1 s_1 \beta_1 \sigma_{\delta_1}^{-2}, \ldots, -2c_n r_n s_n \beta_1 \sigma_{\delta_n}^{-2} \right),$$

$$\boldsymbol{t}_2^t = \left(\frac{2c_1 s_1 \beta_1 \sigma_{\delta_1}^2}{\sigma_{\varepsilon_1}^2} \left(a_1 - \frac{s_1\beta_1^2}{\sigma_{\varepsilon_1}^2} \right), \ldots, \frac{2c_n s_n \beta_1 \sigma_{\delta_n}^2}{\sigma_{\varepsilon_n}^2} \left(a_n - \frac{s_n\beta_1^2}{\sigma_{\varepsilon_n}^2} \right) \right),$$

$$\boldsymbol{T} = \begin{pmatrix} t_{11} & \boldsymbol{t}_{12}^t & \boldsymbol{t}_{13}^t \\ & \boldsymbol{T}_{22} & \boldsymbol{T}_{23} \\ \cdot & & \boldsymbol{T}_{33} \end{pmatrix}, \tag{2.20}$$

$$t_{11} = \sum_{i=1}^{n} c_i \sigma_{\varepsilon_i}^2 \sigma_{\delta_i}^2 \left(\frac{a_i - \sigma_x^{-2}}{\sigma_x^2} \right)^2, \quad \boldsymbol{t}_{12}^t = \left(\frac{r_1 c_1 \sigma_{\varepsilon_1}^2}{\sigma_x^4\sigma_{\delta_1}^2}, \ldots, \frac{r_n c_n \sigma_{\varepsilon_n}^2}{\sigma_x^4\sigma_{\delta_n}^2} \right),$$

$$\boldsymbol{t}_{13}^t = \left(\frac{c_1 s_1 \beta_1^2}{\sigma_x^4} \left(\frac{\sigma_{\delta_1}^2}{\sigma_{\varepsilon_1}^2} \right), \ldots, \frac{c_n s_n \beta_1^2}{\sigma_x^4} \left(\frac{\sigma_{\delta_n}^2}{\sigma_{\varepsilon_n}^2} \right) \right),$$

$$\boldsymbol{T}_{22} = \text{Diag}\left(c_i r_i \frac{\sigma_{\varepsilon_i}^2}{\sigma_{\delta_i}^2} \left(a_i^2 - \frac{2a_i}{\sigma_{\delta_i}^2} + \frac{r_i}{\sigma_{\delta_i}^4} \right), i = 1, \ldots, n \right),$$

$$\boldsymbol{T}_{23} = \text{Diag}\left(\frac{c_i r_i s_i \beta_1^2}{\sigma_{\varepsilon_i}^2\sigma_{\delta_i}^2}, i = 1, \ldots, n \right) \quad \text{e}$$

$$\boldsymbol{T}_{33} = \text{Diag}\left(c_i s_i \frac{\sigma_{\delta_i}^2}{\sigma_{\varepsilon_i}^2} \left(a_i^2 - \frac{2a_i\beta_1^2}{\sigma_{\varepsilon_i}^2} + \frac{s_i\beta_1^4}{\sigma_{\varepsilon_i}^4} \right), i = 1, \ldots, n \right).$$

Deste modo, conseguimos encontrar a matriz de informação esperada com expressões em forma explícita. Propriedades assintóticas dos estimadores são necessárias para construir intervalos de confiança e testar hipóteses de interesse. Desta forma, a matriz de informação esperada torna-se de suma importância, como veremos nos capítulos seguintes.

2.2.2 Matriz de informação esperada do modelo homoscedástico

Uma característica deste modelo, em contraste com o anterior, é que o número de parâmetros a serem estimados é menor, ou seja, $\boldsymbol{\theta}^* = (\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_\delta^2, \sigma_\varepsilon^2)^t$. Novamente, usando a relação dada em (2.16), temos que os elementos da matriz $\boldsymbol{\mathcal{I}}^*$ são dados por

$$\boldsymbol{\mathcal{I}}^* = \begin{pmatrix} \mathcal{I}_{\beta_0\beta_0} & \mathcal{I}_{\beta_0\beta_1} & \mathcal{I}_{\beta_0\mu_x} & \mathcal{I}_{\beta_0\sigma_x^2} & \mathcal{I}_{\beta_0\sigma_\delta^2} & \mathcal{I}_{\beta_0\sigma_\varepsilon^2} \\ & \mathcal{I}_{\beta_1\beta_1} & \mathcal{I}_{\beta_1\mu_x} & \mathcal{I}_{\beta_1\sigma_x^2} & \mathcal{I}_{\beta_1\sigma_\delta^2} & \mathcal{I}_{\beta_1\sigma_\varepsilon^2} \\ & & \mathcal{I}_{\mu_x\mu_x} & \mathcal{I}_{\mu_x\sigma_x^2} & \mathcal{I}_{\mu_x\sigma_\delta^2} & \mathcal{I}_{\mu_x\sigma_\varepsilon^2} \\ & & & \mathcal{I}_{\sigma_x^2\sigma_x^2} & \mathcal{I}_{\sigma_x^2\sigma_\delta^2} & \mathcal{I}_{\sigma_x^2\sigma_\varepsilon^2} \\ & & & & \mathcal{I}_{\sigma_\delta^2\sigma_\delta^2} & \mathcal{I}_{\sigma_\delta^2\sigma_\varepsilon^2} \\ & \cdot & & & & \mathcal{I}_{\sigma_\varepsilon^2\sigma_\varepsilon^2} \end{pmatrix}. \tag{2.21}$$

Novamente, utilizando resultados encontrados em Chan & Mak (1979), os elementos da matriz $\boldsymbol{\mathcal{I}}^*$ são

$$\boldsymbol{\mathcal{I}}^* = \begin{pmatrix} \nu^* & \mu_x\nu^* & q^* & \mathbf{0}_{1\times 3} \\ & p^*_{\beta\beta} + \mu_x^2\nu^* & \mu_x q^* & \boldsymbol{t}^* \\ & & \displaystyle\sum_{i=1}^{n} \frac{a_i^* - \sigma_x^{-2}}{a_i^*\sigma_x^2} & \mathbf{0}_{1\times 3} \\ & \cdot & & \boldsymbol{T}^* \end{pmatrix}, \tag{2.22}$$

em que

$$\begin{aligned} p^*_{\beta\beta} =& \sum_{i=1}^{n} 2c_i^* r_i s_i \left(a_i^*\sigma_x^2 + 2\beta_1^2 s_i \frac{\sigma_\delta^2}{\sigma_\varepsilon^2} \right), \quad \nu^* = \sum_{i=1}^{n} \left(\frac{s_i}{\sigma_\varepsilon^2} - \frac{\beta_1^2 s_i^2}{a_i^*\sigma_\varepsilon^4} \right), \\ a_i^* =& r_i\sigma_\delta^{-2} + s_i\beta_1^2\sigma_\varepsilon^{-2} + \sigma_x^{-2}, \; c_i^* = (2a_i^{*2}\sigma_\delta^2\sigma_\varepsilon^2)^{-1}, \quad q^* = \frac{\beta_1}{\sigma_x^2} \sum_{i=1}^{n} \frac{s_i}{a_i^*\sigma_\varepsilon^2}, \end{aligned} \tag{2.23}$$

$$\boldsymbol{t}^{*t} = \left(\frac{2\beta_1}{\sigma_x^2} \sum_{i=1}^{n} r_i s_i c_i^* \left(1 + \frac{\beta_1^2 s_i \sigma_\delta^2}{r_i \sigma_\varepsilon^2} \right), \sum_{i=1}^{n} -\frac{2c_i^* r_i s_i \beta_1}{\sigma_\delta^2}, \sum_{i=1}^{n} \frac{2c_i^* s_i \beta_1 \sigma_\delta^2}{\sigma_\varepsilon^2} \left(a_i^* - \frac{s_i \beta_1^2}{\sigma_\varepsilon^2} \right) \right) \quad \text{e}$$

$$\boldsymbol{T}^* = \begin{pmatrix} \sum_{i=1}^{n} c_i^* \sigma_\varepsilon^2 \sigma_\delta^2 \left(\frac{a_i^* - \sigma_x^{-2}}{\sigma_x^2} \right)^2 & \sum_{i=1}^{n} \frac{r_i c_i^* \sigma_\varepsilon^2}{\sigma_x^4 \sigma_\delta^2} & \sum_{i=1}^{n} \frac{c_i^* s_i \beta_1^2}{\sigma_x^4} \left(\frac{\sigma_\delta^2}{\sigma_\varepsilon^2} \right) \\ & \sum_{i=1}^{n} c_i^* r_i \frac{\sigma_\varepsilon^2}{\sigma_\delta^2} \left(a_i^{*2} - \frac{2a_i^*}{\sigma_\delta^2} + \frac{r_i}{\sigma_\delta^4} \right) & \sum_{i=1}^{n} \frac{c_i^* r_i s_i \beta_1^2}{\sigma_\varepsilon^2 \sigma_\delta^2} \\ \cdot & & \sum_{i=1}^{n} c_i^* s_i \frac{\sigma_\delta^2}{\sigma_\varepsilon^2} \left(a_i^{*2} - \frac{2a_i^* \beta_1^2}{\sigma_\varepsilon^2} + \frac{s_i \beta_1^4}{\sigma_\varepsilon^4} \right) \end{pmatrix}. \tag{2.24}$$

A matriz de informação esperada $\boldsymbol{\mathcal{I}}^*$, em comparação com $\boldsymbol{\mathcal{I}}$, apresenta uma redução na ordem, devido a que o número de parâmetros é 6. Além disso, a matriz quadrada $\boldsymbol{T}^*$ é de ordem 3×3 em contraste com a matriz $\boldsymbol{T}$, que é $(2n+1) \times (2n+1)$.

Capítulo 3

Inferência

Neste capítulo abordaremos os métodos de máxima verossimilhança (MV) e máxima pseudoverossimilhança (MPV) para estimar os parâmetros dos modelos descritos no Capítulo 2. Desta forma, apresentamos a seguir os estimadores de MV e de MPV dos parâmetros de interesse, assim como também suas respectivas matrizes de covariâncias assintóticas. Finalmente, propomos estatísticas de teste para testar homoscedasticidade das variâncias e verificar os vieses aditivo e multiplicativo, em relação a β_0 e β_1.

3.1 Estimação

As estimativas dos parâmetros pelo método de máxima verossimilhança são obtidas a partir da maximização da função logverossimilhança. O método de máxima pseudoverossimilhança é utilizado quando há parâmetros de perturbação, que são estimados previamente a partir de alguma metodologia usual, para depois serem substituídos na função logverossimilhança original da amostra (descrito no Capítulo 1). O algoritmo EM (Demspter *et al.*, 1977; Tanner, 1996) será utilizado na estimação numérica dos parâmetros dos modelos por ambos os métodos de MV e MPV.

3.1.1 Máxima verossimilhança

A função logverossimilhança dada em (2.7), para o modelo de regressão heteroscedástico, é escrita de forma alternativa como

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Z}) = \text{const.} - \frac{1}{2}\sum_{i=1}^{n} \log|\boldsymbol{V}_i| - \frac{1}{2}\sum_{i=1}^{n} \text{tr}(\boldsymbol{V}_i^{-1}\boldsymbol{D}_i), \tag{3.1}$$

com $\boldsymbol{D}_i = \boldsymbol{d}_i\boldsymbol{d}_i^t$ e $\boldsymbol{d}_i$ dada em (2.7). Assim, usando (3.1) e (2.15) a derivada da função logverossimilhança em relação a um parâmetro γ é

$$U_\gamma = \sum_{i=1}^{n} U_{i,\gamma} = \quad \frac{1}{2}\sum_{i=1}^{n} \text{tr}\left(\boldsymbol{P}_i\boldsymbol{V}_{i,\gamma}\right) - \sum_{i=1}^{n} \boldsymbol{d}_{i,\gamma}^t \boldsymbol{V}_i^{-1}\boldsymbol{d}_i, \tag{3.2}$$

em que $\boldsymbol{P}_i = \boldsymbol{V}_i^{-1}(\boldsymbol{D}_i - \boldsymbol{V}_i)\boldsymbol{V}_i^{-1}$ e $\gamma \in \{\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_{\delta_i}^2, \sigma_{\varepsilon_i}^2,\ i = 1, \ldots, n\}$.

Para encontrar as estimativas de máxima verossimilhança dos parâmetros do modelo de regressão heteroscedástico devemos resolver o sistema de equações dado por

$$\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{Z}) = (U_{\beta_0}, U_{\beta_1}, U_{\mu_x}, U_{\sigma_x^2}, U_{\sigma_{\delta_i}^2}, U_{\sigma_{\varepsilon_i}^2}, i, \ldots, n)^t = \boldsymbol{0}, \tag{3.3}$$

cujas componentes são

$$\begin{aligned}
U_{\beta_0} &= -\sum_{i=1}^{n} \boldsymbol{d}_{i,\beta_0}^t \boldsymbol{V}_i^{-1}\boldsymbol{d}_i, \\
U_{\beta_1} &= \quad \frac{1}{2}\sum_{i=1}^{n} \text{tr}\left(\boldsymbol{P}_i\boldsymbol{V}_{i,\beta_1}\right) - \sum_{i=1}^{n} \left(\boldsymbol{d}_{i,\beta_1}\right)^t \boldsymbol{V}_i^{-1}\boldsymbol{d}_i, \\
U_{\mu_x} &= -\sum_{i=1}^{n} \left(\boldsymbol{d}_{i,\mu_x}\right)^t \boldsymbol{V}_i^{-1}\boldsymbol{d}_i, \\
U_{\sigma_x^2} &= \quad \frac{1}{2}\sum_{i=1}^{n} \text{tr}\left(\boldsymbol{P}_i\boldsymbol{V}_{i,\sigma_x^2}\right), \\
U_{\sigma_{\delta_i}^2} &= \quad \frac{1}{2}\text{tr}\left(\boldsymbol{P}_i\boldsymbol{V}_{i,\sigma_{\delta_i}^2}\right) \qquad \text{e} \\
U_{\sigma_{\varepsilon_i}^2} &= \quad \frac{1}{2}\text{tr}\left(\boldsymbol{P}_i\boldsymbol{V}_{i,\sigma_{\varepsilon_i}^2}\right),\ i = 1, \ldots, n.
\end{aligned} \tag{3.4}$$

Derivando parcialmente $\boldsymbol{V}_i$ e $\boldsymbol{d}_i$ em relação a algumas das componentes de $\boldsymbol{\theta}$, obtemos as matrizes e vetores representados por

$$\boldsymbol{V}_{i,\beta_1} = \begin{pmatrix} \boldsymbol{0}_{r_i,r_i} & \sigma_x^2\boldsymbol{1}_{r_i,s_i} \\ \cdot & 2\beta_1\sigma_x^2\boldsymbol{1}_{s_i,s_i} \end{pmatrix}, \quad \boldsymbol{V}_{i,\sigma_x^2} = \begin{pmatrix} \boldsymbol{1}_{r_i,r_i} & \beta_1\boldsymbol{1}_{r_i,s_i} \\ \cdot & \beta_1^2\boldsymbol{1}_{s_i,s_i} \end{pmatrix},$$

$$\boldsymbol{V}_{i,\sigma^2_{\delta_i}} = \begin{pmatrix} \boldsymbol{I}_{r_i} & \boldsymbol{0}_{r_i,s_i} \\ \cdot & \boldsymbol{0}_{s_i,s_i} \end{pmatrix}, \quad \boldsymbol{V}_{i,\sigma^2_{\varepsilon_i}} = \begin{pmatrix} \boldsymbol{0}_{r_i,r_i} & \boldsymbol{0}_{r_i,s_i} \\ \cdot & \boldsymbol{I}_{s_i} \end{pmatrix}, \tag{3.5}$$

$$\boldsymbol{d}_{i,\beta_0} = -\begin{pmatrix} \boldsymbol{0}^t_{r_i}, & \boldsymbol{1}^t_{s_i} \end{pmatrix}^t, \ \boldsymbol{d}_{i,\beta_1} = -\begin{pmatrix} \boldsymbol{0}^t_{r_i}, & \mu_x \boldsymbol{1}^t_{s_i} \end{pmatrix}^t \text{ e } \boldsymbol{d}_{i,\mu_x} = -\begin{pmatrix} \boldsymbol{1}^t_{r_i}, & \beta_1 \boldsymbol{1}^t_{s_i} \end{pmatrix}^t.$$

Similarmente, para o modelo de regressão homoscedástico, utilizamos (3.2) mas com vetor de parâmetros $\boldsymbol{\theta}^*$ e resolver o seguinte sistema de equações dado por

$$\boldsymbol{U}^*(\boldsymbol{\theta}; \boldsymbol{Z}) = (U^*_{\beta_0}, U^*_{\beta_1}, U_{\mu_x}, U^*_{\sigma^2_x}, U^*_{\sigma^2_\delta}, U^*_{\sigma^2_\varepsilon})^t = \boldsymbol{0}, \tag{3.6}$$

cujas componentes são

$$\begin{aligned}
U^*_{\beta_0} &= -\sum_{i=1}^{n} \boldsymbol{d}^{*t}_{i,\beta_0} \boldsymbol{V}^{*-1}_i \boldsymbol{d}_i, \\
U^*_{\beta_1} &= \frac{1}{2}\sum_{i=1}^{n} \operatorname{tr}\left(\boldsymbol{P}^*_i \boldsymbol{V}^*_{i,\beta_1}\right) - \sum_{i=1}^{n} \left(\boldsymbol{d}^*_{i,\beta_1}\right)^t \boldsymbol{V}^{*-1}_i \boldsymbol{d}_i, \\
U^*_{\mu_x} &= -\sum_{i=1}^{n} \left(\boldsymbol{d}^*_{i,\mu_x}\right)^t \boldsymbol{V}^{*-1}_i \boldsymbol{d}_i, \\
U^*_{\sigma^2_x} &= \frac{1}{2}\sum_{i=1}^{n} \operatorname{tr}\left(\boldsymbol{P}^*_i \boldsymbol{V}^*_{i,\sigma^2_x}\right), \\
U^*_{\sigma^2_\delta} &= \frac{1}{2}\sum_{i=1}^{n} \operatorname{tr}\left(\boldsymbol{P}^*_i \boldsymbol{V}^*_{i,\sigma^2_\delta}\right) \qquad \text{e} \\
U^*_{\sigma^2_\varepsilon} &= \frac{1}{2}\sum_{i=1}^{n} \operatorname{tr}\left(\boldsymbol{P}^*_i \boldsymbol{V}^*_{i,\sigma^2_\varepsilon}\right),
\end{aligned} \tag{3.7}$$

em que $\boldsymbol{V}^{*-1}_i$ e $\boldsymbol{P}^*_i$ são as matrizes $\boldsymbol{V}^{-1}_i$ e $\boldsymbol{P}_i$ dadas em (2.9) e (3.2), respectivamente, com $\sigma^2_{\delta_i} = \sigma^2_\delta$ e $\sigma^2_{\varepsilon_i} = \sigma^2_\varepsilon$ para $i = 1, \ldots, n$. Logo, $\boldsymbol{V}^*_{i,\beta_0}$, $\boldsymbol{V}^*_{i,\beta_1}$, $\boldsymbol{d}^*_{i,\beta_0}$, $\boldsymbol{d}^*_{i,\beta_1}$ e $\boldsymbol{d}^*_{i,\mu_x}$ são como em (3.5), pois não dependem de σ^2_δ e σ^2_ε, e por último, $\boldsymbol{V}^*_{i,\sigma^2_\delta}$ e $\boldsymbol{V}^*_{i,\sigma^2_\varepsilon}$ são $\boldsymbol{V}_{i,\sigma^2_{\delta_i}}$ e $\boldsymbol{V}^*_{i,\sigma^2_{\varepsilon_i}}$ dadas em (3.5), mas com $\sigma^2_{\delta_i} = \sigma^2_\delta$ e $\sigma^2_{\varepsilon_i} = \sigma^2_\varepsilon$. Ao mesmo tempo, vemos que os sistemas de equações dados em (3.3) e (3.6) apresentam formas complexas, ou seja, não se consegue isolar expressões algébricas para cada um dos estimadores dos parâmetros. Assim, as equações de verossimilhança, em ambos os casos, não têm solução explícita para cada um dos sistemas acima, exigindo métodos numéricos em sua solução. Na literatura

existem procedimentos iterativos que ajudam na obtenção de estimativas numéricas dos parâmetros, por exemplo, o algoritmo de Newton-Raphson. Nesta dissertação, utilizamos um algoritmo que facilita na obtenção de estimativas tanto de MV quanto de MPV dos parâmetros para os modelos apresentados no Capítulo 2 e é descrito a seguir.

Algoritmo EM

Utilizaremos um algoritmo iterativo para obter as estimativas dos parâmetros numericamente, devido à complexidade das expressões das derivadas apresentadas em (3.4) e (3.7), pois não é simples obter os estimadores explicitamente. O algoritmo EM é muito utilizado como uma alternativa, já que em diversos problemas apresenta uma forma mais simples na obtenção de estimativas de máxima verossimilhança numericamente (Tanner, 1996). A seguir deduzimos os passos do algoritmo EM, encontrando expressões de forma algébrica em cada passo para os modelos de regressão apresentados.

A ideia do algoritmo EM, nesta dissertação, é considerar os verdadeiros valores das covariáveis (variáveis latentes) como dados perdidos, isto é, tomar como vetor de dados completos o vetor $\boldsymbol{W}_i = (x_i, \boldsymbol{X}_i^t, \boldsymbol{Y}_i^t)^t = (x_i, \boldsymbol{Z}_i^t)^t$ e com o algoritmo estimando os valores de $\mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}]$ e $\mathrm{E}[x_i^2|\boldsymbol{Z}, \boldsymbol{\theta}]$ (passo E) e maximizando a esperança condicional da função logverossimilhança dos dados completos dadas as observações (passo M).

Primeiramente, desenvolvemos o algoritmo EM para o modelo de regressão heteroscedástico. Assim, o modelo com os dados completos é

$$\boldsymbol{W}_i \overset{ind.}{\sim} N_{1+r_i+s_i}\left(\begin{pmatrix} \mu_x \\ \mu_x \mathbf{1}_{r_i} \\ (\beta_0+\beta_1\mu_x)\mathbf{1}_{s_i} \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_x^2\mathbf{1}_{r_i}^t & \beta_1\sigma_x^2\mathbf{1}_{s_i}^t \\ & \sigma_{\delta_i}^2\boldsymbol{I}_{ri}+\sigma_x^2\mathbf{1}_{r_i,r_i} & \beta_1\sigma_x^2\mathbf{1}_{r_i,s_i} \\ \cdot & & \sigma_{\varepsilon_i}^2\boldsymbol{I}_{si}+\beta_1^2\sigma_x^2\mathbf{1}_{s_i,s_i} \end{pmatrix}\right), \quad (3.8)$$

com a_i dado em (2.13), $i = 1, \ldots, n$. Seja $\boldsymbol{G}_i$ a matriz de covariâncias do modelo completo.

O determinante de $\boldsymbol{G}_i$ e sua inversa $\boldsymbol{G}_i^{-1}$ são dados por

$$|\boldsymbol{G}_i| = \sigma_x^2 \sigma_{\delta_i}^{2r_i} \sigma_{\varepsilon_i}^{2s_i} \quad \text{e} \quad \boldsymbol{G}_i^{-1} = \begin{pmatrix} a_i & -\sigma_{\delta_i}^{-2}\mathbf{1}_{r_i}^t & -\beta_1\sigma_{\varepsilon_i}^{-2}\mathbf{1}_{s_i}^t \\ & \sigma_{\delta_i}^{-2}\boldsymbol{I}_{ri} & \mathbf{0}_{r_i,s_i} \\ \cdot & & \sigma_{\varepsilon_i}^{-2}\boldsymbol{I}_{si} \end{pmatrix}. \tag{3.9}$$

É importante destacar que $|\boldsymbol{G}_i|$ não depende de β_1. Logo, a função logverossimilhança baseada no modelo completo é denotada por $\mathcal{L}_c$ e é dada por

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{W}) = \text{const.} - \frac{n}{2}\log\sigma_x^2 - \frac{1}{2}\sum_{i=1}^{n} r_i \log\sigma_{\delta_i}^2 - \frac{1}{2}\sum_{i=1}^{n} s_i \log\sigma_{\varepsilon_i}^2 - \frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(x_i - \mu_x)^2 \\ - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_{\delta_i}^2}\sum_{k=1}^{r_i}(X_{ik} - x_i)^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_{\varepsilon_i}^2}\sum_{j=1}^{s_i}(Y_{ij} - \beta_0 - \beta_1 x_i)^2. \end{aligned} \tag{3.10}$$

É conveniente utilizar a formulação em (3.10), que é mais simples do que (2.10). Logo, calculando o valor esperado condicional de $\mathcal{L}_c(\boldsymbol{W}, \boldsymbol{\theta})$ dadas as observações, obtemos

$$\begin{aligned} \mathrm{E}[\mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{W})|\boldsymbol{Z}, \boldsymbol{\theta}] = \text{const.} - \frac{n}{2}\log\sigma_x^2 - \frac{1}{2}\sum_{i=1}^{n} r_i \log\sigma_{\delta_i}^2 - \frac{1}{2}\sum_{i=1}^{n} s_i \log\sigma_{\varepsilon_i}^2 \\ -\frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(\mathrm{E}[x_i^2|\boldsymbol{Z}, \boldsymbol{\theta}] - 2\mu_x\mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}] + \mu_x^2) \\ -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_{\delta_i}^2}\sum_{k=1}^{r_i}(X_{ik}^2 - 2X_{ik}\mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}] + \mathrm{E}[x_i^2|\boldsymbol{Z}, \boldsymbol{\theta}]) \\ -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_{\varepsilon_i}^2}\sum_{j=1}^{s_i}(Y_{ij}^2 - \beta_0^2 + \beta_1^2\mathrm{E}[x_i^2|\boldsymbol{Z}, \boldsymbol{\theta}] - 2Y_{ij}\beta_0 - 2Y_{ij}\beta_1\mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}] + 2\beta_0\beta_1\mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}]). \end{aligned} \tag{3.11}$$

Para calcular $\mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}]$ e $\mathrm{E}[x_i^2|\boldsymbol{Z}, \boldsymbol{\theta}]$ utilizamos (3.8). Particionamos $\boldsymbol{W}_i = (x_i, \boldsymbol{Z}_i^t)^t$ e de acordo com as dimensões de $\boldsymbol{W}_i$ a matriz de covariâncias $\boldsymbol{G}_i$. Assim, utilizando o Teorema 3.6.1. em Graybill (1976) temos

$$x_i|\boldsymbol{Z}_i \sim N(\mu_x + \boldsymbol{\Sigma}_{i_{12}}\boldsymbol{V}_i^{-1}(\boldsymbol{Z}_i - \boldsymbol{m}_i), \sigma_x^2 - \boldsymbol{\Sigma}_{i_{12}}\boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_{i_{21}}).$$

em que $\boldsymbol{\Sigma}_{i_{12}} = (\sigma_x^2 \mathbf{1}_{r_i}^t, \beta_1 \sigma_x^2 \mathbf{1}_{s_i}^t)^t$, $\boldsymbol{\Sigma}_{i_{21}} = \boldsymbol{\Sigma}_{i_{12}}^t$. Além de serem $\boldsymbol{m}_i$ e $\boldsymbol{V}_i$ dados em (2.4).

O processo do algoritmo EM consiste estimar $\boldsymbol{\theta}$ maximizando (3.11) em relação aos parâmetros, em que os valores $\mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}]$ e $\mathrm{E}[x_i^2|\boldsymbol{Z}, \boldsymbol{\theta}]$ são substituídos pelas suas estimativas com $\widehat{\boldsymbol{\theta}}$ da iteração anterior. Em seguida, estima-se novamente $\mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}]$ e $\mathrm{E}[x_i^2|\boldsymbol{Z}, \boldsymbol{\theta}]$ utilizando as estimativas atualizadas dos parâmetros e assim sucessivamente até que se obtenha a convergência. Executando os cálculos de $\mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}]$ e $\mathrm{E}[x_i^2|\boldsymbol{Z}, \boldsymbol{\theta}]$ e calculando as derivadas de (3.11) em relação aos parâmetros obtemos os passos do algoritmo.

(1) Inicie o procedimento iterativo com $\kappa = 0$ e atribua valores iniciais a $\boldsymbol{\theta}^{(0)}$, ou seja, $\beta_0^{(0)}$, $\beta_1^{(0)}$, $\mu_x^{(0)}$, $\sigma_x^{2(0)}$, $\sigma_{\delta_i}^{2(0)}$ e $\sigma_{\varepsilon_i}^{2(0)}$, $i = \ldots, n$. Como valores iniciais, podemos usar as estimativas encontradas por algum outro método, por exemplo, mínimos quadrados ponderados e componentes de variâncias.

(2) Calcule

$$x_i^{(\kappa+1)} = \mathrm{E}[x_i|\boldsymbol{Z}, \boldsymbol{\theta}] = \frac{\dfrac{\mu_x^{(\kappa)}}{\sigma_x^{2(\kappa)}} + \dfrac{\displaystyle\sum_{k=1}^{r_i} X_{ik}}{\sigma_{\delta_i}^{2(\kappa)}} + \beta_1^{(\kappa)} \dfrac{\displaystyle\sum_{j=1}^{s_i} Y_{ij} - s_i \beta_0^{(\kappa)}}{\sigma_{\varepsilon_i}^{2(\kappa)}}}{\dfrac{1}{\sigma_x^{2(\kappa)}} + \dfrac{r_i}{\sigma_{\delta_i}^{2(\kappa)}} + \dfrac{s_i \beta_1^{2(\kappa)}}{\sigma_{\varepsilon_i}^{2(\kappa)}}} \quad \text{e} \tag{3.12}$$

$$(x_i^2)^{(\kappa+1)} = \mathrm{E}[x_i^2|\boldsymbol{Z}, \boldsymbol{\theta}] = (x_i^{(\kappa+1)})^2 + \frac{1}{\dfrac{1}{\sigma_x^{2(\kappa)}} + \dfrac{r_i}{\sigma_{\delta_i}^{2(\kappa)}} + \dfrac{s_i \beta_1^{2(\kappa)}}{\sigma_{\varepsilon_i}^{2(\kappa)}}}, \tag{3.13}$$

$i = 1, \ldots, n$.

(3) Calcule

$$\mu_x^{(\kappa+1)} = \frac{1}{n} \sum_{i=1}^{n} x_i^{(\kappa+1)}, \quad \sigma_x^{2(\kappa+1)} = \frac{1}{n} \sum_{i=1}^{n} \left[(x_i^2)^{(\kappa+1)} - 2\mu_x^{(\kappa+1)} x_i^{(\kappa+1)} + \mu_x^{2(\kappa+1)} \right],$$

$$\sigma_{\varepsilon_i}^{2(\kappa+1)} = \frac{1}{s_i} \sum_{j=1}^{s_i} [Y_{ij}^2 + \beta_0^2 + \beta_1^2 (x_i^2)^{(\kappa+1)} - 2\beta_0 Y_{ij} - 2\beta_1 Y_{ij} x_i^{(\kappa+1)} + 2\beta_0 \beta_1 x_i^{(\kappa+1)}],$$

$$\sigma_{\delta_i}^{2(\kappa+1)} = \frac{1}{r_i}\sum_{\kappa=1}^{r_i}[X_{ik}^2 - 2X_{ik}x_i^{(\kappa+1)} + (x_i^2)^{(\kappa+1)}], \quad w_i^{(\kappa+1)} = \frac{s_i}{\sigma_{\varepsilon_i}^{2(\kappa+1)}},\ i, \ldots, n,$$

$$\beta_0^{(\kappa+1)} = \frac{\sum_{i=1}^{n} w_i^{(\kappa+1)}\overline{Y}_{i.}}{\sum_{i=1}^{n} w_i^{(\kappa+1)}} - \beta_1^{(\kappa)}\frac{\sum_{i=1}^{n} w_i^{(\kappa+1)}x_i^{(\kappa+1)}}{\sum_{i=1}^{n} w_i^{(\kappa+1)}} \quad \text{e} \tag{3.14}$$

$$\beta_1^{(\kappa+1)} = \frac{\sum_{i=1}^{n} w_i^{(\kappa+1)}x_i^{(\kappa+1)}(\overline{Y}_{i.} - \beta_0^{(\kappa+1)})}{\sum_{i=1}^{n} w_i^{(\kappa+1)}(x_i^2)^{(\kappa+1)}}. \tag{3.15}$$

(4) Incremente κ em uma unidade.

(5) Repita os passos 2, 3 e 4 até a convergência.

Observe que a maximização da esperança condicional da função logverossimilhança dos dados completos em relação a $\boldsymbol{\theta}$ é mais simples e tem forma explícita. Desta forma, a obtenção das estimativas de máxima verossimilhança mediante o algoritmo EM é bastante simples e apresenta menor custo computacional.

Os passos do algoritmo EM, para o modelo de regressão homoscedástico, são resultado de uma adaptação do algoritmo EM descrito anteriormente. Especificamos que a esperança condicional da função logverossimilhança dos dados completos para este modelo é dada por (3.11), mas com as variâncias dos erros de medição substituídas por $\sigma_{\delta_i}^2 = \sigma_\delta^2$ e $\sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2$ (passo E). Logo, por analogia, o passo M é obtido calculando as derivadas desta função logverossimilhança em relação às componentes de $\boldsymbol{\theta}^* = (\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_\delta^2, \sigma_\varepsilon^2)^t$. Assim, obtemos os passos do algoritmo.

(1) Inicie o procedimento iterativo com $\kappa = 0$ e atribua valores iniciais a $\boldsymbol{\theta}^{*(0)}$, ou seja, $\beta_0^{(0)}$, $\beta_1^{(0)}$, $\mu_x^{(0)}$, $\sigma_x^{2(0)}$, $\sigma_\delta^{2(0)}$ e $\sigma_\varepsilon^{2(0)}$. Como valores iniciais, podemos usar as estimativas encontradas por algum outro método, por exemplo, mínimos quadrados ponderados e componentes de variâncias.

(2) Calcule

$$x_i^{*(\kappa+1)} = \mathrm{E}[x_i|\boldsymbol{Z},\boldsymbol{\theta}] = \frac{\dfrac{\mu_x^{(\kappa)}}{\sigma_x^{2(\kappa)}} + \dfrac{\displaystyle\sum_{k=1}^{r_i} X_{ik}}{\sigma_\delta^{2(\kappa)}} + \beta_1^{(\kappa)} \dfrac{\displaystyle\sum_{j=1}^{s_i} Y_{ij} - s_i\beta_0^{(\kappa)}}{\sigma_\varepsilon^{2(\kappa)}}}{\dfrac{1}{\sigma_x^{2(\kappa)}} + \dfrac{r_i}{\sigma_\delta^{2(\kappa)}} + \dfrac{s_i\beta_1^{2(\kappa)}}{\sigma_\varepsilon^{2(\kappa)}}} \quad \text{e} \tag{3.16}$$

$$(x_i^{*2})^{(\kappa+1)} = \mathrm{E}[x_i^2|\boldsymbol{Z},\boldsymbol{\theta}] = (x_i^{*(\kappa+1)})^2 + \frac{1}{\dfrac{1}{\sigma_x^{2(\kappa)}} + \dfrac{r_i}{\sigma_\delta^{2(\kappa)}} + \dfrac{s_i\beta_1^{2(\kappa)}}{\sigma_\varepsilon^{2(\kappa)}}}, \tag{3.17}$$

para $i, \ldots, n$.

(3) Calcule

$$\mu_x^{(\kappa+1)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{*(\kappa+1)}, \quad \sigma_x^{2(\kappa+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i^{*2})^{(\kappa+1)} - 2\mu_x^{(\kappa+1)}x_i^{*(\kappa+1)} + (\mu_x^{(\kappa+1)})^2\right],$$

$$\sigma_\varepsilon^{2(\kappa+1)} = \frac{1}{\displaystyle\sum_{i=1}^{n} s_i}\sum_{i=1}^{n}\sum_{j=1}^{s_i}\left[Y_{ij}^2 + \beta_0^2 + \beta_1^2(x_i^{*2})^{(\kappa+1)} - 2\beta_0 Y_{ij} - 2\beta_1 Y_{ij}x_i^{*(\kappa+1)} + 2\beta_0\beta_1 x_i^{*(\kappa+1)}\right],$$

$$\sigma_\delta^{2(\kappa+1)} = \frac{1}{\displaystyle\sum_{i=1}^{n} r_i}\sum_{i=1}^{n}\sum_{k=1}^{r_i}\left[X_{ik}^2 - 2X_{ik}x_i^{*(\kappa+1)} + (x_i^{*2})^{(\kappa+1)}\right],$$

$$\beta_0^{(\kappa+1)} = \frac{\displaystyle\sum_{i=1}^{n} s_i\overline{Y}_{i.}}{\displaystyle\sum_{i=1}^{n} s_i} - \beta_1^{(\kappa)}\frac{\displaystyle\sum_{i=1}^{n} s_i x_i^{*(\kappa+1)}}{\displaystyle\sum_{i=1}^{n} s_i} \quad \text{e} \quad \beta_1^{(\kappa+1)} = \frac{\displaystyle\sum_{i=1}^{n} s_i x_i^{*(\kappa+1)}(\overline{Y}_{i.} - \beta_0^{(\kappa+1)})}{\displaystyle\sum_{i=1}^{n} s_i (x_i^{*2})^{(\kappa+1)}}. \tag{3.18}$$

(4) Incremente κ em uma unidade.

(5) Repita os passos 2, 3 e 4 até a convergência.

Neste trabalho, a convergência é obtida quando a diferença relativa entre as estimativas dos parâmetros, considerando os passos atual e o anterior, é suficientemente pequena. Seja

$$\zeta^{(\kappa+1)} = \max\left\{\left|\frac{\widehat{\theta}_\iota^{(\kappa+1)} - \widehat{\theta}_\iota^{(\kappa)}}{\widehat{\theta}_\iota^{(\kappa)}}\right|, \iota = 1, \ldots, 2n+4\right\}, \tag{3.19}$$

em que $\widehat{\theta}_\iota^{(\kappa)}$ é a estimativa de máxima verossimilhança da ι-ésima componente do vetor $\boldsymbol{\theta} = (\beta_0, \beta_1, \mu_x, \sigma_x^2, \boldsymbol{\sigma}_\delta^{2t}, \boldsymbol{\sigma}_\varepsilon^{2t})^t$ na κ-ésima iteração. No caso do modelo de regressão homoscedástico às componentes de $\boldsymbol{\theta}^* = (\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_\delta^2, \sigma_\varepsilon^2)^t$.

3.1.2 Máxima pseudoverossimilhança

O método de máxima verossimilhança é um dos mais conhecidos e utilizados na estimação dos parâmetros. Além de apresentar boas propriedades assintóticas como consistência forte, eficiência assintótica e distribuição normal assintótica. No entanto, a aplicação deste método pode enfrentrar dificuldades em diversas situações, e especialmente na presença de parâmetros de perturbação. Existem vários procedimentos alternativos que têm sido propostos, como por exemplo, o método de máxima verossimilhança perfilada, máxima verossimilhança marginal (Berger *et al.*, 1999) e máxima pseudoverossimilhança (Gong & Samaniego, 1981). Em geral, o objetivo de todas estas abordagens é a eliminação desses parâmetros de perturbação. A seguir utilizamos o método de máxima pseudoverossimilhança com o intuito de contornar a presença de parâmetros de perturbação como descrito na Seção 1.2. Deste modo, consideramos a partição do vetor $\boldsymbol{\theta}$ ($\boldsymbol{\theta}^*$) em $\boldsymbol{\phi}$ ($\boldsymbol{\phi}^*$) e $\boldsymbol{\lambda}$ ($\boldsymbol{\lambda}^*$), vetores de parâmetros de interesse e perturbação, respectivamente, segundo o modelo de regressão heteroscedástico (homoscedástico).

O método de máxima pseudoverossimilhança requer encontrar uma estimativa $\widetilde{\boldsymbol{\lambda}}$ ($\widetilde{\boldsymbol{\lambda}}^*$) para $\boldsymbol{\lambda}$ ($\boldsymbol{\lambda}^*$) por meio de algum método de estimação alternativo. Neste trabalho, $\widetilde{\boldsymbol{\lambda}}$ foi obtida por meio de componentes de variância. Guolo (2011) apresentou outra alternativa para encontrar uma estimativa para $\boldsymbol{\lambda}$, obtida pela maximização de uma função logverossimilhança reduzida ($\mathcal{L}_R$) baseada na distribuição das covariáveis de $\boldsymbol{X}$, mas não será

apresentada neste trabalho.

Estimação do parâmetro de perturbação

Em Searle (1971, Cap. 10, Seção 8) encontramos estimadores de componentes de variância. A obtenção de um estimador para $\boldsymbol{\lambda} = (\mu_x, \sigma_x^2, \sigma_{\delta_i}^2, \sigma_{\varepsilon_i}^2, i = 1, \ldots, n)^t$ no modelo de regressão heteroscedástico utilizou componentes de variância levando em conta que o número de réplicas podem ser balanceadas ou desbalanceadas. No modelo de regressão homoscedástico usamos as estimativas de componentes de variância para $\boldsymbol{\lambda}^* = (\mu_x, \sigma_x^2, \sigma_\delta^2, \sigma_\varepsilon^2)^t$, encontradas em Carroll *et al.* (2006, Cap. 4, Seção 4.2). Denotamos por $\widetilde{\boldsymbol{\lambda}}$ e $\widetilde{\boldsymbol{\lambda}}^*$ como os estimadores dos parâmetros $\boldsymbol{\lambda}$ e $\boldsymbol{\lambda}^*$, respectivamente. Assim, as componentes de $\widetilde{\boldsymbol{\lambda}}$ são dadas pelas expressões

$$\begin{gathered}\widetilde{\mu}_x = \overline{X}_{..} = \frac{\sum_{i=1}^{n} r_i \overline{X}_{i.}}{\sum_{i=1}^{n} r_i}, \qquad \widetilde{\sigma}_{\delta_i}^2 = \frac{1}{r_i - 1} \sum_{k=1}^{r_i} (X_{ik} - \overline{X}_{i.})^2, \\ \widetilde{\sigma}_{\varepsilon_i}^2 = \frac{1}{s_i - 1} \sum_{j=1}^{s_i} (Y_{ij} - \overline{Y}_{i.})^2, \ i = 1, \ldots, n, \quad \text{e} \\ \widetilde{\sigma}_x^2 = \left[\sum_{i=1}^{n} r_i (\overline{X}_{i.} - \widetilde{\mu}_x)^2 - \sum_{i=1}^{n} \frac{N - r_i}{N} \widetilde{\sigma}_{\delta_i}^2 \right] \frac{N}{N^2 - S_2},\end{gathered} \tag{3.20}$$

em que

$$N = \sum_{i=1}^{n} r_i, \ \ S_2 = \sum_{i=1}^{n} r_i^2, \ \ \overline{X}_{i.} = \frac{1}{r_i} \sum_{k=1}^{r_i} X_{ik} \ \text{ e } \ \overline{Y}_{i.} = \frac{1}{s_i} \sum_{j=1}^{s_i} Y_{ij}. \tag{3.21}$$

Para $\boldsymbol{\lambda}^*$, temos $\widetilde{\boldsymbol{\lambda}}^* = (\widetilde{\mu}_x^*, \widetilde{\sigma}_x^{2*}, \widetilde{\sigma}_\delta^{2*}, \widetilde{\sigma}_\varepsilon^{2*})$, sendo $\widetilde{\mu}_x^*$ igual a $\widetilde{\mu}_x$ dada em (3.20) e os demais estimadores são

$$\begin{gathered}\widetilde{\sigma}_\delta^{2*} = \frac{1}{\sum_{i=1}^{n}(r_i - 1)} \sum_{i=1}^{n} \sum_{k=1}^{r_i} \left(X_{ik} - \overline{X}_{i.}\right)^2, \ \widetilde{\sigma}_\varepsilon^{2*} = \frac{1}{\sum_{i=1}^{n}(s_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{s_i} \left(Y_{ij} - \overline{Y}_{i.}\right)^2 \\ \text{e} \quad \widetilde{\sigma}_x^{2*} = \left[\sum_{i=1}^{n} r_i (\overline{X}_{i.} - \widetilde{\mu}_x)^2 - (n-1)\widetilde{\sigma}_\delta^{*2} \right] \frac{N}{N^2 - S_2}.\end{gathered} \tag{3.22}$$

Em algumas situações, as estimativas $\widetilde{\sigma}_x^2$ e $\widetilde{\sigma}_x^{*2}$ podem resultar em valores negativos. Nesses casos, Searle (1971) recomenda truncar em zero, mas o estimador obtido por componentes de variância seria viesado. No Capítulo 4 as amostras que apresentaram valores negativos para $\widetilde{\sigma}_x^2$ e $\widetilde{\sigma}_x^{*2}$ foram descartadas. Note que as dimensões dos vetores de perturbação $\boldsymbol{\lambda}$ e $\boldsymbol{\lambda}^*$ são $2n+2$ e 4, respectivamente.

Estimação do parâmetro de interesse por MPV

Após encontrar estimadores para $\boldsymbol{\lambda}$, a etapa seguinte consiste em substituir $\widetilde{\boldsymbol{\lambda}}$ na função logverossimilhança dada em (2.7). Assim, a função logpseudoverossimilhança para o modelo de regressão heteroscedástico é

$$\mathcal{L}(\boldsymbol{\phi}; \boldsymbol{Z}, \widetilde{\boldsymbol{\lambda}}) = \text{const.} - \frac{1}{2}\sum_{i=1}^{n} \log|\boldsymbol{V_i}(\boldsymbol{\phi}, \widetilde{\boldsymbol{\lambda}})| - \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{d}_i(\boldsymbol{\phi}, \widetilde{\boldsymbol{\lambda}})^t \boldsymbol{V}_i(\boldsymbol{\phi}, \widetilde{\boldsymbol{\lambda}})^{-1} \boldsymbol{d}_i(\boldsymbol{\phi}, \widetilde{\boldsymbol{\lambda}}), \quad (3.23)$$

em que $\boldsymbol{V}_i$ e $\boldsymbol{d}_i$ estão em função de $(\boldsymbol{\phi}, \widetilde{\boldsymbol{\lambda}})$, com $\widetilde{\boldsymbol{\lambda}}$ fixado. Conservando a notação utilizada no Capítulo 1, para a função logpseudoverossimilhança, temos

$$\mathcal{L}_p(\boldsymbol{\phi}; \boldsymbol{Z}) = \text{const.} - \frac{1}{2}\sum_{i=1}^{n} \log|\boldsymbol{V_i}(\boldsymbol{\phi})| - \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{d}_i(\boldsymbol{\phi})^t \boldsymbol{V}_i(\boldsymbol{\phi})^{-1} \boldsymbol{d}_i(\boldsymbol{\phi}). \quad (3.24)$$

Para encontrar os estimadores de MPV para $\boldsymbol{\phi} = (\beta_0, \beta_1)^t$, derivamos (3.24) em relação a cada componente, e igualando a 0, obtemos o sistema de equações

$$\begin{aligned} \frac{\partial \mathcal{L}_p(\boldsymbol{\phi}; \boldsymbol{Z})}{\partial \beta_0} &= -\sum_{i=1}^{n} \boldsymbol{d}_{i,\beta_0}(\boldsymbol{\phi})^t \boldsymbol{V}_i(\boldsymbol{\phi})^{-1} \boldsymbol{d}_i(\boldsymbol{\phi}) = 0 \quad \text{e} \\ \frac{\partial \mathcal{L}_p(\boldsymbol{\phi}; \boldsymbol{Z})}{\partial \beta_1} &= \frac{1}{2}\sum_{i=1}^{n} \text{tr}\left(\boldsymbol{P}_i(\boldsymbol{\phi}) \boldsymbol{V}_{i,\beta_1}(\boldsymbol{\phi})\right) - \sum_{i=1}^{n} \boldsymbol{d}_{i,\beta_1}(\boldsymbol{\phi})^t \boldsymbol{V}_i(\boldsymbol{\phi})^{-1} \boldsymbol{d}_i(\boldsymbol{\phi}) = 0, \end{aligned} \quad (3.25)$$

em que $\boldsymbol{d}_i$ e $\boldsymbol{V}_i^{-1}$ são dadas em (2.8) e (2.9), respectivamente, $\boldsymbol{V}_{i,\beta_1}$, $\boldsymbol{d}_{i,\beta_0}$ e $\boldsymbol{d}_{i,\beta_1}$ são como em (3.5), mas com $\boldsymbol{\lambda} = \widetilde{\boldsymbol{\lambda}}$ dado em (3.20). Se segue por analogia que para o modelo de regressão homoscedástico, as expressões citadas acima são substituídas por suas correspondentes matrizes e vetores envolvidos e $\boldsymbol{\lambda}^*$ substituído por $\widetilde{\boldsymbol{\lambda}}^*$ dada em (3.22). Agora, vemos que o número de equações é reduzido somente a duas equações

de pseudoverossimilhança, diminuindo a dimensão do espaço paramétrico e torna-se mais simples encontrar suas soluções, mas ainda assim as equações não têm solução explícita. Novamente recorremos ao algoritmo EM descrito na Seção 3.1.1, que será adaptado ao método de MPV. É importante comentar que para o modelo de regressão heteroscedástico, considerando que μ_x, σ_x^2 e os vetores $\boldsymbol{\sigma}_\delta^2$ e $\boldsymbol{\sigma}_\varepsilon^2$ são estimados por componentes de variância, o número de possíveis estimadores de MPV é $2^4 - 1 = 15$. Agora, se considerarmos os elementos dos vetores $\boldsymbol{\sigma}_\delta^2$ e $\boldsymbol{\sigma}_\varepsilon^2$, o número de estimadores MPV é muito maior, passando a ser $2^{2(n+1)} - 1$. Por exemplo, se $n = 10$, temos 4194303 diferentes estimadores de MPV de $\boldsymbol{\phi}$.

Algoritmo EM para o método de MPV

Em primeiro lugar, considerando o modelo de regressão heteroscedástico, o algoritmo EM para o método de MPV é baseado em algumas adaptações do algoritmo descrito na Seção 3.1.1. Assim, obteremos $\widetilde{\boldsymbol{\phi}} = (\widetilde{\beta}_0, \widetilde{\beta}_1)^t$, uma estimativa numérica de MPV para $\boldsymbol{\phi} = (\beta_0, \beta_1)^t$. Os passos do algoritmo são:

(1) Inicie o procedimento iterativo em $k = 0$ e atribua valores iniciais a $\beta_0^{(0)}$ e $\beta_1^{(0)}$. Como valores iniciais, podemos usar as estimativas encontradas por algum outro método (mínimos quadrados ponderados, método dos momentos, etc).

(2) Calcule $\widetilde{x}_i^{(\kappa+1)}$ e $(\widetilde{x}_i^2)^{(\kappa+1)}$ dadas em (3.12), mas com $\boldsymbol{\lambda} = \widetilde{\boldsymbol{\lambda}}$ dado em (3.20).

(3) Calcule $\widetilde{\beta}_0^{(\kappa+1)}$ e $\widetilde{\beta}_1^{(\kappa+1)}$ dadas pelas expressões de $\beta_0^{(\kappa+1)}$ e $\beta_1^{(\kappa+1)}$ em (3.14), mas com $\boldsymbol{\lambda} = \widetilde{\boldsymbol{\lambda}}$ dado em (3.20).

(4) Incremente κ em uma unidade.

(5) Repita os passos 2, 3 e 4 até a convergência.

Para a obtenção de estimativas de MPV utilizando o algoritmo EM para o modelo de regressão homoscedástico, são dados os seguintes passos:

(1) Inicie o procedimento iterativo em $\kappa = 0$ e atribua valores iniciais a $\beta_0^{(0)}$ e $\beta_1^{(0)}$. Como valores iniciais, podemos usar as estimativas encontradas por algum outro método alternativo, como mínimos quadrados ponderados, método dos momentos, etc.

(2) Calcule $\widetilde{x}_i^{*(\kappa+1)}$ e $(\widetilde{x}_i^{*2})^{(\kappa+1)}$ dadas em (3.16), mas substituídas em $\boldsymbol{\lambda} = \widetilde{\boldsymbol{\lambda}}^*$ dada em (3.22).

(3) Calcule $\widetilde{\beta}_0^{*(\kappa+1)}$ e $\widetilde{\beta}_1^{*(\kappa+1)}$ dadas pelas expressões de $\beta_0^{*(\kappa+1)}$ e $\beta_1^{*(\kappa+1)}$ em (3.18), mas com $\boldsymbol{\lambda} = \widetilde{\boldsymbol{\lambda}}^*$ dado em (3.22).

(4) Incremente κ em uma unidade.

(5) Repita os passos 2,3 e 4 até a convergência.

Neste caso, a convergência é verificada pelo erro relativo, dado por

$$\zeta^{(\kappa+1)} = \max\left\{\left|\frac{\widetilde{\beta}_0^{(\kappa+1)} - \widetilde{\beta}_0^{(\kappa)}}{\widetilde{\beta}_0^{(\kappa)}}\right|, \left|\frac{\widetilde{\beta}_1^{(\kappa+1)} - \widetilde{\beta}_1^{(\kappa)}}{\widetilde{\beta}_1^{(\kappa)}}\right|\right\},$$

até atingir um valor pequeno. Da mesma maneira, para as componentes de $\widetilde{\boldsymbol{\phi}}^* = (\widetilde{\beta}_0^*, \widetilde{\beta}_1^*)^t$. No Capítulo 4, estudos de simulação serão realizados para avaliar algumas propriedades dos estimadores MV e MPV para ambos os modelos propostos.

3.2 Matriz de covariâncias assintótica

Propriedades assintóticas dos estimadores são necessárias para construir intervalos de confiança e testar hipóteses. Nesta seção apresentamos as matrizes de covariâncias assintóticas dos estimadores obtidos pelos métodos de estimação estudados na Seção 3.1 para os parâmetros de interesse.

3.2.1 Estimador MV

Em geral, sob algumas condições de regularidade (Lehmann & Casella, 1998, Cap. 7, Seção 5), no modelo de regressão com erros de medição sob enfoque estrutural para observações replicadas e casos particulares, as matrizes de covariâncias assintóticas dos estimadores de máxima verossimilhança dos vetores de parâmetros $\boldsymbol{\theta}$ e $\boldsymbol{\theta}^*$, para a situação em que o número de observações n é fixo e o número de réplicas r_i e s_i cresce, é obtida pela inversão das matrizes de informação esperada $\boldsymbol{\mathcal{I}}$ e $\boldsymbol{\mathcal{I}}^*$, respectivamente (Chan & Mak, 1979; Dolby, 1976).

Em primeiro lugar, particionamos a matriz $\boldsymbol{\mathcal{I}}$ de acordo com as dimensões dos vetores $\boldsymbol{\theta} = (\boldsymbol{\phi}^t, \boldsymbol{\lambda}^t)^t$ sendo $\boldsymbol{\phi} = (\beta_0, \beta_1)^t$ e $\boldsymbol{\lambda} = (\mu_x, \sigma_x^2, \mu_x, \boldsymbol{\sigma}_\delta^{2t}, \boldsymbol{\sigma}_\varepsilon^{2t})^t$ os parâmetros de interesse e perturbação, respectivamente, pois estamos interessados especialmente na matriz de covariâncias assintótica dos estimadores dos parâmetros de interesse $\boldsymbol{\phi}$. Assim, tal partição de $\boldsymbol{\mathcal{I}}$ é efetuada como

$$\boldsymbol{\mathcal{I}} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{11} & \boldsymbol{\mathcal{I}}_{12} \\ \cdot & \boldsymbol{\mathcal{I}}_{22} \end{pmatrix}, \tag{3.26}$$

sendo

$$\begin{gathered} \boldsymbol{\mathcal{I}}_{11} = \begin{pmatrix} \nu & \mu_x \nu \\ \cdot & p_{\beta\beta} + \mu_x^2 \nu \end{pmatrix}, \quad \boldsymbol{\mathcal{I}}_{12} = \begin{pmatrix} q & \mathbf{0}_{1\times(2n+1)} \\ \mu_x q & \boldsymbol{t}^t \end{pmatrix} \quad \text{e} \\ \boldsymbol{\mathcal{I}}_{22} = \begin{pmatrix} \sum_{i=1}^{n} \dfrac{a_i - \sigma_x^{-2}}{a_i \sigma_x^2} & \mathbf{0}_{1\times(2n+1)} \\ \mathbf{0}_{(2n+1)\times 1} & \boldsymbol{T} \end{pmatrix}, \end{gathered} \tag{3.27}$$

em que a_i é dado em (2.13), ν, $p_{\beta\beta}$, $\boldsymbol{t}$ e $\boldsymbol{T}$ são dados em (2.19) e (2.20).

A matriz $\boldsymbol{\mathcal{I}}^{-1}$ tem blocos

$$\boldsymbol{\mathcal{I}}^{-1} = \begin{pmatrix} \boldsymbol{\mathcal{I}}^{11} & \boldsymbol{\mathcal{I}}^{12} \\ \cdot & \boldsymbol{\mathcal{I}}^{22} \end{pmatrix}. \tag{3.28}$$

O nosso propósito é obter a submatriz $\boldsymbol{\mathcal{I}}^{11}$. A partir da forma de $\boldsymbol{\mathcal{I}}$ em (3.26) e usando propriedades de partição de matrizes (Magnus & Neudecker, 2007, Cap. 1, Seção

11), temos

$$\boldsymbol{\mathcal{I}}^{11} = \left(\boldsymbol{\mathcal{I}}_{11} - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}^{t}\right)^{-1} = \begin{pmatrix} \nu - \eta & \mu_x(\nu - \eta) \\ \cdot & p_{\beta\beta} + \mu_x^2(\nu - \eta) - \boldsymbol{t}^t\boldsymbol{T}^{-1}\boldsymbol{t} \end{pmatrix}^{-1}, \tag{3.29}$$

em que

$$\eta = \frac{\beta_1^2}{\sigma_x^4}\left(\sum_{i=1}^{n}\frac{a_i - \sigma_x^{-2}}{a_i\sigma_x^2}\right)^{-1}\left(\sum_{i=1}^{n}\frac{s_i}{a_i\sigma_{\varepsilon_i}^2}\right)^2, \tag{3.30}$$

com $\boldsymbol{t}$ dado em (2.18) e no Apêndice A encontra-se a matriz $\boldsymbol{T}^{-1}$.

Portanto, a matriz de covariâncias dos estimadores de máxima verossimilhança dos parâmetros de interesse do modelo de regressão heteroscedástico, é obtida por

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{1}{\Delta}\begin{pmatrix} p_{\beta\beta} + \mu_x^2(\nu - \eta) - \boldsymbol{t}^t\boldsymbol{T}^{-1}\boldsymbol{t} & -\mu_x(\nu - \eta) \\ \cdot & \nu - \eta \end{pmatrix}, \tag{3.31}$$

com $\Delta = (p_{\beta\beta} - \boldsymbol{t}^t\boldsymbol{T}^{-1}\boldsymbol{t})(\nu - \eta)$.

De acordo com o procedimento anterior, podemos encontrar a matriz de covariâncias do estimador de máxima verossimilhança para o vetor de parâmetros $\boldsymbol{\phi}^*$ no modelo de regressão homoscedástico. Neste caso, a partição de $\boldsymbol{\theta}^*$ tem a forma $\boldsymbol{\phi}^* = (\beta_0, \beta_1)^t$ e $\boldsymbol{\lambda}^* = (\mu_x, \sigma_x^2, \sigma_\delta^2, \sigma_\varepsilon^2)^t$. Depois de manipulações matriciais, a matriz correspondente é dada por

$$\mathrm{Cov}(\widehat{\beta}_0^*, \widehat{\beta}_1^*) = \frac{1}{\Delta^*}\begin{pmatrix} p_{\beta\beta}^* + \mu_x^2(\nu^* - \eta^*) - \boldsymbol{t}^{*t}\boldsymbol{T}^{*-1}\boldsymbol{t}^* & -\mu_x(\nu^* - \eta^*) \\ \cdot & \nu^* - \eta^* \end{pmatrix}, \tag{3.32}$$

sendo

$$\eta^* = \frac{\beta_1^2}{\sigma_x^4}\left(\sum_{i=1}^{n}\frac{a_i^* - \sigma_x^{-2}}{a_i^*\sigma_x^2}\right)^{-1}\left(\sum_{i=1}^{n}\frac{s_i}{a_i^*\sigma_\varepsilon^2}\right)^2,$$

e $\Delta^* = (p_{\beta\beta}^* - \boldsymbol{t}^{*t}\boldsymbol{T}^{*-1}\boldsymbol{t}^*)(\nu^* - \eta^*)$, em que ν^*, $p_{\beta\beta}^*$ e $\boldsymbol{t}^*$ são dados em (2.24) e (2.23). A matriz $\boldsymbol{T}^{*-1}$ encontra-se no Apêndice A.

A matriz de covariâncias assintótica dada em (3.31) é estimada consistentemente substituindo os parâmetros por seus estimadores $\widehat{\boldsymbol{\lambda}}$ e $\widehat{\boldsymbol{\phi}}$. Sob condições de regularidade adequadas, os estimadores de máxima verossimilhança para $\widehat{\boldsymbol{\phi}} = (\widehat{\beta}_0, \widehat{\beta}_1)^t$ obtido na Seção 3.1.1

têm distribuição conjunta assintótica normal bivariada. Analogamente, $\widehat{\boldsymbol{\phi}}^* = (\widehat{\beta_0}^*, \widehat{\beta_1}^*)^t$ é substituída em (3.32) e como resultado teremos um estimador consistente para sua matriz de covariâncias assintótica.

3.2.2 Estimador MPV

Para encontrar a matriz de covariâncias assintótica do estimador de máxima pseudoverossimilhança para $\boldsymbol{\phi}$ ($\boldsymbol{\phi}^*$ no caso homoscedástico), utilizaremos a teoria de equações de estimação, semelhante à abordagem descrita na Seção 2.1. No modelo de regressão heteroscedástico as observações são independentes, mas não são identicamente distribuídas para $i = 1, \ldots, n$, pois existem parâmetros específicos para cada i, como por exemplo, $\sigma^2_{\delta_i}$ é a variância do erro de medida δ_i para cada i, assim como também existem parâmetros comuns entre as observações (β_0, β_1, μ_x e σ^2_x). Além disso, a nossa situação é para o número de observações n mantido fixo e o número de réplicas r_i e s_i tendendo ao infinito. A seguir, apresentaremos as matrizes de covariâncias para os estimadores de máxima pseudoverossimilhança para $\boldsymbol{\phi}$ e $\boldsymbol{\phi}^*$.

Modelo heteroscedástico

A partir do estimador $\widetilde{\boldsymbol{\lambda}}$ dado em (3.20), podemos obter a parcela da função de estimação para a i-ésima observação como

$$\boldsymbol{\Psi}_i(\boldsymbol{\lambda}, \boldsymbol{Z}_i) = \begin{pmatrix} \Psi_{i1}(\mu_x, \boldsymbol{Z}_i) \\ \Psi_{i2}(\mu_x, \sigma^2_x, \sigma^2_{\delta_i}, \boldsymbol{Z}_i) \\ \boldsymbol{\Psi}_{i3}(\boldsymbol{\sigma}^2_\delta, \boldsymbol{Z}_i) \\ \boldsymbol{\Psi}_{i4}(\boldsymbol{\sigma}^2_\varepsilon, \boldsymbol{Z}_i) \end{pmatrix}, \tag{3.33}$$

em que

$$\begin{aligned} \Psi_{i1}(\mu_x, \boldsymbol{Z}_i) =& \sum_{k=1}^{r_i} (\mu_x - X_{ik}), \\ \Psi_{i2}(\mu_x, \sigma^2_x, \sigma^2_{\delta_i}, \boldsymbol{Z}_i) =& r_i \frac{N^2 - S_2}{N^2} \sigma^2_x - r_i (\overline{X}_{i\cdot} - \mu_x)^2 + \frac{N - r_i}{N} \sigma^2_{\delta_i}, \end{aligned} \tag{3.34}$$

$$\Psi_{i3_l}(\boldsymbol{\sigma}^2_{\delta_i}, \boldsymbol{Z}_i) = \begin{cases} (r_i - 1)\sigma^2_{\delta_i} - \sum_{k=1}^{r_i}(X_{ik} - \overline{X}_{i.})^2, & \text{se} \quad i = l, \\ 0 \qquad\qquad\qquad\quad , & \text{c.c.} \end{cases}$$

e

$$\Psi_{i4_m}(\boldsymbol{\sigma}^2_{\varepsilon_i}, \boldsymbol{Z}_i) = \begin{cases} (s_i - 1)\sigma^2_{\varepsilon_i} - \sum_{j=1}^{s_i}(Y_{ij} - \overline{Y}_{i.})^2, & \text{se} \quad i = m, \\ 0 \qquad\qquad\qquad\quad , & \text{c.c.} \end{cases} \tag{3.35}$$

para $l = 1, \ldots, n$ e $m = 1, \ldots, n$, respectivamente.

Por sua vez, a função de estimação para as n observações no modelo de regressão heteroscedástico é

$$\sum_{i=1}^{n} \boldsymbol{\Psi}_i(\boldsymbol{\lambda}, \boldsymbol{Z}_i) = \begin{pmatrix} \sum_{i=1}^{n} r_i(\mu_x - X_{ik}) \\ \frac{N^2 - S_2}{N}\sigma^2_x - \sum_{i=1}^{n} r_i(\overline{X}_{i.} - \mu_x)^2 + \sum_{i=1}^{n} \frac{N - r_i}{N}\sigma^2_{\delta_i} \\ (r_1 - 1)\sigma^2_{\delta_1} - \sum_{k=1}^{r_1}(X_{1k} - \overline{X}_{1.})^2 \\ \vdots \\ (r_n - 1)\sigma^2_{\delta_n} - \sum_{k=1}^{r_n}(X_{nk} - \overline{X}_{n.})^2 \\ (s_1 - 1)\sigma^2_{\varepsilon_1} - \sum_{j=1}^{s_1}(Y_{1j} - \overline{Y}_{1.})^2 \\ \vdots \\ (s_n - 1)\sigma^2_{\varepsilon_n} - \sum_{j=1}^{s_n}(Y_{nj} - \overline{Y}_{n.})^2 \end{pmatrix}. \tag{3.36}$$

Neste contexto, a matriz de covariâncias assintótica do estimador de máxima pseudo-verossimilhança para $\boldsymbol{\phi}$ com $\boldsymbol{\lambda}$ substituído por $\widetilde{\boldsymbol{\lambda}}$ é

$$\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}_{\phi\phi}(\boldsymbol{B}_{\phi\phi} - \boldsymbol{A}_{\phi\lambda}\boldsymbol{A}^{-1}_{\lambda\lambda}\boldsymbol{B}_{\lambda\phi} - \boldsymbol{B}^t_{\lambda\phi}\boldsymbol{A}^{-1}_{\lambda\lambda}\boldsymbol{A}^t_{\phi\lambda} + \boldsymbol{A}_{\phi\lambda}\boldsymbol{A}^{-1}_{\lambda\lambda}\boldsymbol{B}_{\lambda\lambda}\boldsymbol{A}^{-1}_{\lambda\lambda}\boldsymbol{A}^t_{\phi\lambda})\boldsymbol{A}^{-1}_{\phi\phi}, \tag{3.37}$$

cujos blocos são dados em (1.18). Entretanto, Parker (1986) mostra que a matriz $\boldsymbol{B}_{\lambda\phi}$ em (3.37) é a matriz nula. Além disso, utilizamos o resultado encontrado em Cordeiro (1999, Cap. 1, Seção 3) sobre $\mathrm{E}\{\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{Z})\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{Z})^t\} = -\mathrm{E}\{\partial^2\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Z})/\partial\boldsymbol{\theta}^2\}$ com $\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{Z})$ dado

em (3.3), que por conseguinte utilizando esse fato na função de logpseudoverossimilhança, teremos

$$\boldsymbol{B}_{\phi\phi} = \sum_{i=1}^{n} \frac{\partial \mathcal{L}_{p,i}(\boldsymbol{\phi}, \boldsymbol{\lambda})}{\partial \boldsymbol{\phi}} \left\{ \frac{\partial \mathcal{L}_{p,i}(\boldsymbol{\phi}, \boldsymbol{\lambda})}{\partial \boldsymbol{\phi}} \right\}^t = -\sum_{i=1}^{n} \mathrm{E} \left\{ \frac{\partial^2 \mathcal{L}_{p,i}(\boldsymbol{\phi}; \boldsymbol{Z}_i, \boldsymbol{\lambda})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^t} \right\} = \boldsymbol{A}_{\phi\phi}, \tag{3.38}$$

sob algumas condições de regularidade adequadas. Logo, considerando (3.38) e o resultado de Parker (1986) temos que a matriz de covariâncias assintótica do estimador de MPV para $\boldsymbol{\phi}$, com $\boldsymbol{\lambda}$ substituído por $\widetilde{\boldsymbol{\lambda}}$, é reescrita como

$$\boldsymbol{\Sigma} = \boldsymbol{A}_{\phi\phi}^{-1} (\boldsymbol{A}_{\phi\phi} + \boldsymbol{A}_{\phi\lambda} \boldsymbol{A}_{\lambda\lambda}^{-1} \boldsymbol{B}_{\lambda\lambda} \boldsymbol{A}_{\lambda\lambda}^{-1} \boldsymbol{A}_{\phi\lambda}^{t}) \boldsymbol{A}_{\phi\phi}^{-1}, \tag{3.39}$$

sendo

$$\boldsymbol{A}_{\phi\phi} = -\sum_{i=1}^{n} \mathrm{E} \left\{ \frac{\partial^2 \mathcal{L}_{p,i}(\boldsymbol{\phi}, \boldsymbol{\lambda})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^t} \right\} = \begin{pmatrix} \nu & \mu_x \nu \\ \cdot & p_{\beta\beta} + \mu_x^2 \nu \end{pmatrix}, \tag{3.40}$$

$$\boldsymbol{A}_{\phi\lambda} = -\sum_{i=1}^{n} \mathrm{E} \left\{ \frac{\partial^2 \mathcal{L}_{p,i}(\boldsymbol{\phi}, \boldsymbol{\lambda})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\lambda}^t} \right\} = \begin{pmatrix} q & \mathbf{0}_{1\times(2n+1)} \\ \mu_x q & \boldsymbol{t}^t \end{pmatrix}, \tag{3.41}$$

$$\boldsymbol{A}_{\lambda\lambda} = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\Psi}_i(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \quad \text{e} \quad \boldsymbol{B}_{\lambda\lambda} = \sum_{i=1}^{n} \boldsymbol{\Psi}_i(\boldsymbol{\lambda}) \left\{ \boldsymbol{\Psi}_i(\boldsymbol{\lambda}) \right\}^t, \tag{3.42}$$

com $\boldsymbol{\Psi}_i(\boldsymbol{\lambda}; \boldsymbol{Z}_i)$ dada em (3.33), obtemos

$$\boldsymbol{A}_{\lambda\lambda} = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\Psi}_i(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \sum_{i=1}^{n} \begin{pmatrix} \dfrac{\partial(\Psi_{1i}, \Psi_{2i})}{\partial \boldsymbol{\lambda}_1} & \dfrac{\partial(\Psi_{1i}, \Psi_{2i})}{\partial \boldsymbol{\lambda}_2} \\ \dfrac{\partial(\boldsymbol{\Psi}_{3i}, \boldsymbol{\Psi}_{4i})}{\partial \boldsymbol{\lambda}_1} & \dfrac{\partial(\boldsymbol{\Psi}_{3i}, \boldsymbol{\Psi}_{4i})}{\partial \boldsymbol{\lambda}_2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_{\lambda\lambda 1} & \boldsymbol{A}_{\lambda\lambda 2} \\ \mathbf{0}_{2n\times 2} & \boldsymbol{A}_{\lambda\lambda 3} \end{pmatrix},$$

em que

$$\boldsymbol{A}_{\lambda\lambda 1} = \begin{pmatrix} N & 0 \\ 2\sum_{i=1}^{n} r_i (\overline{X}_{i\cdot} - \mu_x) & \dfrac{N^2 - S_2}{N} \end{pmatrix}, \quad \boldsymbol{A}_{\lambda\lambda 2} = \begin{pmatrix} \mathbf{0}_{1\times n} & \mathbf{0}_{1\times n} \\ \left(\dfrac{N - r_i}{N}, i = \ldots, n \right)^t & \mathbf{0}_{1\times n} \end{pmatrix} \tag{3.43}$$

$$\text{e} \quad \boldsymbol{A}_{\lambda\lambda 3} = \mathrm{Diag}\left(r_1 - 1 \ldots, r_n - 1, s_1 - 1, \ldots, s_n - 1\right),$$

em que ν, $p_{\beta\beta}$, q e $\boldsymbol{t}$ são como em (2.18) e (2.19), mas com $\boldsymbol{\lambda}$ substituído por $\widetilde{\boldsymbol{\lambda}}$. Denotamos $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^t, \boldsymbol{\lambda}_2^t)^t$ com $\boldsymbol{\lambda}_1 = (\mu_x, \sigma_x^2)^t$ e $\boldsymbol{\lambda}_2 = (\boldsymbol{\sigma}_\delta^{2t}, \boldsymbol{\sigma}_\varepsilon^{2t})^t$. Seja $\partial(\Psi_{1i}, \Psi_{2i})/\partial \boldsymbol{\lambda}_1$ a matriz

jacobiana das funções de estimação Ψ_{1i} e Ψ_{2i} dadas em (3.34) em relação aos parâmetros de μ_x e σ_x^2, em outras palavras, a primeira linha representa as derivadas parciais da função Ψ_{1i} em relação a cada componente de $\boldsymbol{\lambda}_1$ (μ_x e σ_x^2). A segunda linha representa as derivadas parciais de Ψ_{2i} em relação a $\boldsymbol{\lambda}_1$. De forma análoga seguem as outras componentes da matriz $\boldsymbol{A}_{\lambda\lambda}$. Além disso, a inversa da matriz $\boldsymbol{A}_{\lambda\lambda}$ pode ser encontrada no Apêndice A.

É notável observar que para $\boldsymbol{\Sigma}$ foi possível encontrar cada bloco em forma fechada. A matriz $\boldsymbol{\Sigma}$ é estimada consistentemente substituindo os parâmetros de interesse por seus estimadores $\widetilde{\boldsymbol{\phi}}$. Sob condições de regularidade apresentadas por Gong & Samaniego (1981) e Delgado (1995), sabemos que a distribuição assintótica de $(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi})$ é normal com vetor de média $\mathbf{0}$ e matriz de covariâncias $\boldsymbol{\Sigma}$ dada em (3.39).

Modelo homoscedástico

O procedimento utilizado para encontrar a matriz de covariâncias no modelo de regressão heteroscedástico, pode ser utilizado quando estamos adotando o modelo de regressão homoscedástico, pois trocando $\boldsymbol{\theta}$ por $\boldsymbol{\theta}^*$, sendo $\boldsymbol{\theta}^* = (\boldsymbol{\phi}^{*t}, \boldsymbol{\lambda}^{*t})$ com $\boldsymbol{\phi}^* = (\beta_0, \beta_1)^t$ e $\boldsymbol{\lambda}^* = (\mu_x, \sigma_x^2, \sigma_\delta^2, \sigma_\varepsilon^2)^t$ vetores de parâmetros de interesse e perturbação do modelo. Além disso, ao invés da função de estimação (3.44), consideramos a função de estimação

$$\sum_{i=1}^{n} \boldsymbol{\Psi}_i^*(\boldsymbol{\lambda}, \boldsymbol{Z}_i) = \begin{pmatrix} \displaystyle\sum_{i=1}^{n} r_i(\mu_x - X_{ik}) \\ \dfrac{N^2 - S_2}{N}\sigma_x^2 - \displaystyle\sum_{i=1}^{n} r_i(\overline{X}_{i.} - \mu_x)^2 + (n-1)\sigma_\delta^2 \\ \displaystyle\sum_{i=1}^{n}[(r_i - 1)\sigma_\delta^2 - \sum_{k=1}^{r_i}(X_{ik} - \overline{X}_{i.})^2] \\ \displaystyle\sum_{i=1}^{n}[(s_i - 1)\sigma_\varepsilon^2 - \sum_{j=1}^{s_i}(Y_{ij} - \overline{Y}_{i.})^2] \end{pmatrix}. \tag{3.44}$$

Desse modo, o processo é análogo. Estudos de simulação serão realizados para analisar as propriedades assintóticas, já que são muito importantes na validação teórica e serão apresentados no Capítulo 4.

3.3 Testes de hipóteses

Na Seção 3.1 estudamos abordagens para estimar os parâmetros do modelo de regressão considerando seus casos particulares. Em geral, realizam-se também testes de hipóteses sobre parâmetros de interesse. Os testes assintóticos mais frequentes utilizados na literatura são da razão de verossimilhanças (ξ_{RV}), Wald (W) e escore (S_R) (Fahrmeir, 1988). Por outro lado, existem outros testes assintóticos equivalentes aos mencionados acima, como por exemplo, a estatística gradiente (S_T) (Terrell, 2002), $C(\alpha)$ de Neyman (Bera & Bilias, 2001) e da razão de pseudoverossimilhanças (ξ_{RPV}) (Self & Liang, 1996). Uma propriedade em comum destes testes é que são assintoticamente equivalentes sob a hipótese nula, exceto a estatística de teste da razão de pseudoverossimilhanças (ξ_{RPV}).

Self & Liang (1996) mostraram que a distribuição assintótica nula da estatística da razão de pseudoverossimilhanças é dada por uma soma ponderada de variáveis independentes qui-quadrado. Além disso, os autores sugerem uma aproximação para os pesos.

A seguir iremos apresentar de forma breve alguns conceitos referentes às estatísticas de gradiente S_T, $C(\alpha)$ de Neyman e da razão de pseudoverossimilhanças ξ_{RPV}.

Consideremos o vetor $\boldsymbol{Z} = (\boldsymbol{Z}_1^t, \ldots, \boldsymbol{Z}_n^t)^t$ composto de n observações independentes mas não necessariamente identicamente distribuídas. Seja $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Z})$ a função logverossimilhança, dado $\boldsymbol{Z}$, que depende de um vetor de parâmetros desconhecido $\boldsymbol{\theta}$ com $p_1 + p_2$ componentes. Sejam $\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{Z}) = \partial\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Z})/\partial\boldsymbol{\theta}$ e $\boldsymbol{\mathcal{I}} = -\mathrm{E}\{\partial^2\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Z})/\partial\boldsymbol{\theta}^2\}$, a função escore e a matriz de informação esperada para $\boldsymbol{\theta}$, respectivamente. O problema em questão é o de testar uma hipótese nula composta

$$H_0 : \boldsymbol{\phi} = \boldsymbol{\phi}_0, \tag{3.45}$$

contra a hipótese alternativa $H_1 : \boldsymbol{\phi} \neq \boldsymbol{\phi}_0$, em que $\boldsymbol{\theta} = (\boldsymbol{\phi}^t, \boldsymbol{\lambda}^t)^t$, $\boldsymbol{\phi}$ e $\boldsymbol{\lambda}$ são os vetores de parâmetros de interesse e perturbação com dimensões p_1 e p_2, respectivamente, e sendo $\boldsymbol{\phi}_0$ um vetor p_1-dimensional de valores especificados. A partição de $\boldsymbol{\theta}$ induz as correspondentes partições em $\boldsymbol{U}(\boldsymbol{\theta}) = (\boldsymbol{U}_1(\boldsymbol{\theta})^t, \boldsymbol{U}_2(\boldsymbol{\theta})^t)^t$ a partir de (3.3), para $\boldsymbol{\mathcal{I}}$ e $\boldsymbol{\mathcal{I}}^{-1}$

consideram-se as partições em (3.26) e (3.28), respectivamente.

Definição 3.3.1 (Terrell, 2002) A estatística gradiente, S_T, para testar a hipótese nula composta $H_0 : \boldsymbol{\phi} = \boldsymbol{\phi}_0$ contra a alternativa $H_1 : \boldsymbol{\phi} \neq \boldsymbol{\phi}_0$ é definida como

$$S_T = \boldsymbol{U}_1(\check{\boldsymbol{\theta}})^t(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0), \tag{3.46}$$

em que $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\phi}}^t, \widehat{\boldsymbol{\lambda}}^t)^t$ e $\check{\boldsymbol{\theta}} = (\boldsymbol{\phi}_0^t, \check{\boldsymbol{\lambda}}^t)^t$ são os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ sob H_1 e H_0, respectivamente.

A estatística gradiente é uma estatística muito simples de ser calculada, não envolvendo nenhum cálculo matricial como pode-se observar na Definição 3.3.1. Um atrativo desta estatística é que não há necessidade de se obter a matriz de informação esperada.

A estatística de teste de $C(\alpha)$ de Neyman (Bera & Bilias, 2001) para testar a hipótese (3.45) é dada por

$$C(\alpha) = \boldsymbol{U}(\widehat{\boldsymbol{\theta}}^{**})^t\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}^{**})^{-1}\boldsymbol{U}(\widehat{\boldsymbol{\theta}}^{**}) - \boldsymbol{U}_2(\widehat{\boldsymbol{\theta}}^{**})^t\boldsymbol{\mathcal{I}}_{22}(\widehat{\boldsymbol{\theta}}^{**})^{-1}\boldsymbol{U}_2(\widehat{\boldsymbol{\theta}}^{**}), \tag{3.47}$$

sendo $\widehat{\boldsymbol{\theta}}^{**} = (\boldsymbol{\phi}_0^t, \widehat{\boldsymbol{\lambda}}^{**t})^t$ e $\widehat{\boldsymbol{\lambda}}^{**}$ é um estimador consistente de $\boldsymbol{\lambda}$. Vale lembrar que $\widehat{\boldsymbol{\theta}}^{**}$ não necessariamente é um estimador de máxima verossimilhança, para evitar confusão. Algumas vezes é difícil encontrar as expressões para as matrizes envolvidas em (3.47), mas neste trabalho foram obtidas estas matrizes com $\boldsymbol{\mathcal{I}}$ e $\boldsymbol{\mathcal{I}_{22}}$ dadas em (3.27) e (3.26), respectivamente.

Por fim, consideramos a estatística de teste da razão de pseudoverossimilhanças (RPV), que é dada por $\xi_{RPV} = -2\{\mathcal{L}_p(\boldsymbol{\phi}_0) - \mathcal{L}_p(\widetilde{\boldsymbol{\phi}})\}$, em que $\boldsymbol{\phi}_0$ é um ponto interior do espaço paramétrico (Self & Liang, 1996). A seguir enunciamos duas conjecturas que surgem a partir de estudos teóricos apresentados em Self & Liang (1996).

Conjectura 1 Suponha que

$$\left.\begin{pmatrix} \dfrac{\partial\mathcal{L}_p(\boldsymbol{\phi};\boldsymbol{Z})}{\partial\boldsymbol{\phi}} \\ \widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \end{pmatrix}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0,\boldsymbol{\lambda}=\boldsymbol{\lambda_0}} \xrightarrow{d} N_{p_1+p_2}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{A}_{\phi\phi} & \mathbf{0}\\ \cdot & \boldsymbol{\Sigma}_{\lambda\lambda}\end{pmatrix}\right), \tag{3.48}$$

em que $\boldsymbol{\phi}_0$ e $\boldsymbol{\lambda}_0$ são os verdadeiros valores de $\boldsymbol{\phi}$ e $\boldsymbol{\lambda}$, respectivamente, tal que

$$\mathrm{E}\left[-\frac{\partial^2 \mathcal{L}_p(\boldsymbol{\phi}_0; \boldsymbol{Z})}{\partial \boldsymbol{\phi}^2}; \boldsymbol{\phi}_0\right] \to \boldsymbol{A}_{\phi\phi},$$

quando r_i e s_i crescem na mesma taxa para $i = 1, \ldots, n$. Então, sob condições de regularidade sob $\mathcal{L}_p(\boldsymbol{\phi}; \boldsymbol{Z})$ dada em (3.24), temos

$$\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi} \xrightarrow{d} N_{p_1}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

sendo

$$\boldsymbol{\Sigma} = \boldsymbol{A}_{\phi\phi}^{-1}(\boldsymbol{A}_{\phi\phi} + \boldsymbol{A}_{\phi\lambda}\boldsymbol{\Sigma}_{\lambda\lambda}\boldsymbol{A}_{\phi\lambda}^t)\boldsymbol{A}_{\phi\phi}^{-1} = \boldsymbol{A}_{\phi\phi}^{-1}\boldsymbol{\mathcal{B}}_{\phi\phi}\boldsymbol{A}_{\phi\phi}^{-1} \tag{3.49}$$

e

$$\mathrm{E}\left[-\frac{\partial^2 \mathcal{L}_p(\boldsymbol{\phi}_0, \widetilde{\boldsymbol{\lambda}})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\lambda}^t}\right] \to \boldsymbol{A}_{\phi\lambda}.$$

Observe que $\boldsymbol{\Sigma}_{\lambda\lambda} = \boldsymbol{A}_{\lambda\lambda}^{-1}\boldsymbol{B}_{\lambda\lambda}\boldsymbol{A}_{\lambda\lambda}$ é a matriz de covariâncias de $(\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)$ e cujos blocos são dados em (3.42).

Conjectura 2 Sob as mesmas condições da Conjectura 1, a distribuição de ξ_{RPV} sob $H_0 : \boldsymbol{\phi} = \boldsymbol{\phi}_0$, quando r_i e s_i incrementam-se na mesma taxa para n fixo, é $\chi = \sum_{p=1}^{P} \Lambda_p \chi_p$, sendo os χ_p's variáveis independentes qui-quadrado com um grau de liberdade e $\Lambda_1 \geq \Lambda_2 \geq \cdots \geq \Lambda_P$ são os P autovalores positivos de $\boldsymbol{\mathcal{B}}_{\phi\phi}\boldsymbol{A}_{\phi\phi}^{-1}$ dado em (3.49).

Prova: A prova desta conjectura pode ser análoga à encontrada em Self & Liang (1996), mas não será apresentada aqui.

Esta Conjectura 2 é útil se os Λ_p's forem estimados. Self & Liang (1996) apresentaram uma aproximação para χ, estimando consistentemente $\boldsymbol{\mathcal{B}}_{\phi\phi}$ e $\boldsymbol{A}_{\phi\phi}^{-1}$ que são funções de $(\boldsymbol{\phi}_0, \boldsymbol{\lambda}_0)$ por meio de $\widetilde{\boldsymbol{\mathcal{B}}}_{\phi\phi}$ e $\widetilde{\boldsymbol{A}}_{\phi\phi}^{-1}$ em função de $(\boldsymbol{\phi}_0, \widetilde{\boldsymbol{\lambda}})$, respectivamente. Desse modo, χ é aproximada por $\widetilde{\Lambda}\chi_P^2$ em que

$$\widetilde{\Lambda} = \frac{1}{P}\mathrm{tr}\left(\widetilde{\boldsymbol{\mathcal{B}}}_{\phi\phi}\widetilde{\boldsymbol{A}}_{\phi\phi}^{-1}\right) \tag{3.50}$$

em que P é o número de autovalores positivos e $\mathrm{tr}(\cdot)$ denota o traço.

É importante ressaltar que estamos considerando o caso em que n é fixo, $r_i \to \infty$ e $s_i \to \infty$, sendo que a taxa de crescimento é a mesma para r_i e s_i.

3.3.1 Teste de homoscedasticidade

Quando consideramos o modelo na Seção 2.1.1, supomos que as variâncias dos erros de medição são heteroscedásticas. Entretanto, em algumas situações as variâncias dos erros podem ser homoscedásticas. Nesses casos, podemos utilizar o modelo visto na Seção 2.1.2, com a vantagem de que o número de parâmetros a estimar é menor do modelo heteroscedástico. A homoscedasticidade das variâncias $\sigma^2_{\delta_i}$ e $\sigma^2_{\varepsilon_i}$ pode ser testada considerando

$$\text{Hipóteses 1:} \begin{cases} \text{H}_0 : \sigma^2_{\delta_i} = \sigma^2_{\delta} \text{ e } \sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon}, \text{ para } i = 1, \ldots, n, \quad \text{contra} \\ \text{H}_1 : \sigma^2_{\delta_i} \neq \sigma^2_{\delta_j} \text{ ou } \sigma^2_{\varepsilon_i} \neq \sigma^2_{\varepsilon_j} \text{ para pelo menos um } i \neq j,\ i, j \in \{1, \ldots, n\}. \end{cases} \tag{3.51}$$

Utilizando o teste da razão de verossimilhanças, temos que a estatística de teste é dada por $\xi_{RV} = 2\{\mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{Z}) - \mathcal{L}(\widehat{\boldsymbol{\theta}}^*; \boldsymbol{Z})\}$, sendo $\mathcal{L}(\cdot;\cdot)$ é a função logverossimilhança dada em (2.7), $\widehat{\boldsymbol{\theta}}$ é o estimador de máxima verossimilhança (EMV) irrestrito de $\boldsymbol{\theta}$ e $\widehat{\boldsymbol{\theta}}^*$ o EMV sob H_0. Tanto $\widehat{\boldsymbol{\theta}}$ quanto $\widehat{\boldsymbol{\theta}}^*$ são obtidos por algoritmos EM. Neste caso, ξ_{RV} tem uma distribuição assintótica χ^2_{2n-2}, quando $s_i \to \infty$ e $r_i \to \infty$, sendo a taxa de crescimento a mesma para $i = 1, \ldots, n$. Assim, rejeitamos H_0 com um nível de significância α se o valor ξ_{RV} é maior do que o quantil $1 - \alpha$ da distribuição χ^2_{2n-2}.

3.3.2 Testes de vieses aditivo e multiplicativo

Em diversas aplicações as observações em (2.2) e (2.3) dizem respeito a medições de uma mesma quantidade desconhecida x utilizando dois métodos de medição, conforme apresentado em Riu & Rius (1996) e Galea-Rojas *et al.* (2003), por exemplo. Desse modo, outras hipóteses de relevância neste trabalho, envolvem os vícios aditivo e multiplicativo de um método em relação ao outro, desta forma a hipótese formulada como $H_0 : (\beta_0, \beta_1)^t =$

$(0,1)^t$ verifica a ausência de vício, pois constitui questão de interesse. Lembremos que $\boldsymbol{\phi} = (\beta_0, \beta_1)^t$ é o parâmetro de interesse e $\boldsymbol{\lambda} = (\mu_x, \sigma_x^2, \boldsymbol{\sigma}_\delta^{2t}, \boldsymbol{\sigma}_\varepsilon^{2t})^t$ representa parâmetro de perturbação. Assim, para testar

$$\text{Hipóteses 2:} \begin{cases} \text{H}_0 : (\beta_0, \beta_1)^t = (\beta_{00}, \beta_{10})^t \quad \text{contra} \\ \text{H}_1 : (\beta_0, \beta_1)^t \neq (\beta_{00}, \beta_{10})^t. \end{cases} \tag{3.52}$$

em que β_{00} e β_{10} são constantes conhecidas, utilizamos as estatísticas dadas por

$$\begin{aligned} &\text{RV: } \xi_{RV} = -2\{\mathcal{L}(\check{\boldsymbol{\theta}}; \boldsymbol{Z}) - \mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{Z})\}, \\ &\text{Wald: } W = (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)^t \{\boldsymbol{\mathcal{I}}_{11}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\mathcal{I}}_{12}(\widehat{\boldsymbol{\theta}}) \boldsymbol{\mathcal{I}}_{12}^t(\widehat{\boldsymbol{\theta}})^{-1} \boldsymbol{\mathcal{I}}_{22}(\widehat{\boldsymbol{\theta}})\} (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0), \\ &\text{Escore: } S_R = \boldsymbol{U}_1(\check{\boldsymbol{\theta}})^t \boldsymbol{\mathcal{I}}^{11}(\check{\boldsymbol{\theta}}) \boldsymbol{U}_1(\check{\boldsymbol{\theta}}), \\ &\text{Gradiente: } S_T = \boldsymbol{U}_1(\check{\boldsymbol{\theta}})^t (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0), \\ &C(\alpha) \text{ de Neyman: } C(\alpha) = \boldsymbol{U}(\widehat{\boldsymbol{\theta}}^{**})^t \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}}^{**}) \boldsymbol{U}(\widehat{\boldsymbol{\theta}}^{**}) - \boldsymbol{U}_2(\widehat{\boldsymbol{\theta}}^{**})^t \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}}^{**}) \boldsymbol{U}_2(\widehat{\boldsymbol{\theta}}^{**}) \\ &\text{e} \quad \text{RPV: } \xi_{RPV} = -2\{\mathcal{L}_p(\boldsymbol{\phi}_0; \boldsymbol{Z}) - \mathcal{L}_p(\widetilde{\boldsymbol{\phi}}; \boldsymbol{Z})\}, \end{aligned} \tag{3.53}$$

em que $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\phi}}^t, \widehat{\boldsymbol{\lambda}}^t)$ e $\check{\boldsymbol{\theta}} = (\boldsymbol{\phi}_0^t, \check{\boldsymbol{\lambda}}^t)$ são os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ sob H_1 e H_0, respectivamente. Além disso, $\widehat{\boldsymbol{\theta}}^{**} = (\boldsymbol{\phi}_0^t, \widehat{\boldsymbol{\lambda}}^{**t})^t$, em que $\widehat{\boldsymbol{\lambda}}^{**}$ é um estimador consistente de $\boldsymbol{\lambda}$, $\widetilde{\boldsymbol{\phi}}$ é o estimador de máxima pseudoverossimilhança de $\boldsymbol{\phi}$ sob H_1. Assim, sob H_0 temos que ξ_{RV}, W, S_R, S_T, e $C(\alpha) \xrightarrow{d} \chi_2^2$. Além disso, $\xi_{RPV} \xrightarrow{d} \widetilde{\Lambda}\chi_2^2$ com $\widetilde{\Lambda}$ obtido em (3.50). Novamente, enfatizamos que tratamos do caso em que n é fixo e o número de réplicas r_i e s_i cresce na mesma taxa. Logo, rejeitamos H_0 com um nível de significância α se o valor ξ_{RV} é maior do que o quantil $1-\alpha$ da distribuição χ_2^2 e por analogia seguem as demais.

Similarmente ao procedimento utilizado para testar as Hipóteses 2 no modelo de regressão heteroscedástico, é considerado também para o modelo dada na Seção 2.1.2 substituindo o parâmetro $\boldsymbol{\theta}$ por $\boldsymbol{\theta}^*$, sendo $\boldsymbol{\theta}^* = (\boldsymbol{\phi}^{*t}, \boldsymbol{\lambda}^{*t})$ com $\boldsymbol{\phi}^* = (\beta_0, \beta_1)^t$ e $\boldsymbol{\lambda}^* = (\mu_x, \sigma_x^{2*}, \sigma_\delta^{2*}, \sigma_\varepsilon^{2*})^t$.

É importante comentar, que testar $\beta_0 = 0$ e $\beta_1 = 1$ separadamente podem-se reduzir a dois modelos diferentes e conhecidos.

Capítulo 4

Estudos de simulação

Neste capítulo realizamos um estudo de simulação com o objetivo de avaliar o comportamento dos estimadores dos parâmetros de interesse, β_0 e β_1, dos modelos de regressão com erros de medição sob enfoque estrutural heteroscedástico e homoscedástico para observações replicadas. São utilizadas as medidas do viés e da raiz quadrada do erro quadrático médio para os estimadores obtidos através dos métodos de máxima verossimilhança e de máxima pseudoverossimilhança como foram abordados no Capítulo 3, assim como também avaliar o comportamento do desvio padrão amostral e os erros padrão assintóticos. Após isso, verificamos o desempenho dos estimadores de β_0 e β_1 apresentando simulações da amplitude média e taxas de cobertura dos intervalos de confiança. Além disso, analisamos os testes da razão de verossimilhanças, Wald, escore, gradiente, $C(\alpha)$ de Neyman e da razão pseudoverossimilhanças, descritos na Seção 3.3, em relação às taxas de rejeição de H_0 sob H_0 e sob H_1. Diferentes números de réplicas foram utilizados, em que procuramos encontrar a quantidade mínima de replicações necessárias para que o nível nominal e poder dos testes propostos sejam satisfatórios.

Em uma situação semelhante ao conjunto dos dados reais com o qual trabalharemos no Capítulo 5, geramos $B = 5000$ conjuntos de amostras com a escolha de verdadeiros valores dos parâmetros baseada nas estimativas obtidas pelos métodos de máxima verossimilhança

e máxima pseudoverossimilhança. Selecionamos os seguintes parâmetros que permanecem fixos para todas as simulações: $\mu_x = 1,3$, $\sigma_x^2 = 0,1$, $\beta_0 = -0,1$, $\beta_1 = 1,1$, pois foram encontrados a partir do modelo da Seção 2.1.1. Assim, amostramos os valores verdadeiros de $x_i \sim N(\mu_x, \sigma_x^2)$ e e calculamos $y_i = \beta_0 + \beta_1 x_i$. As observações replicadas são geradas das equações (2.2) e (2.3) e de acordo com a natureza das variâncias dos erros de medição, ou seja, as variâncias podem ser tanto heteroscedásticas quanto homoscedásticas resultando nos modelos dados nas Seções 2.1.1 e 2.1.2, respectivamente.

Cabe enfatizar que nesta dissertação definimos uma alta ou baixa variância dos erros de medição considerando os coeficientes de confiabilidade dados por $k_{x_i} = \sigma_x^2/(\sigma_x^2 + \sigma_{\delta_i}^2)$ e $k_{y_i} = \sigma_y^2/(\sigma_y^2 + \sigma_{\varepsilon_i}^2)$ com $\sigma_y^2 = \beta_1^2\sigma_x^2$, e por conseguinte, as variâncias dos erros de medição são geradas por $\sigma_{\delta_i}^2 = \sigma_x^2(1 - k_{x_i})/k_{x_i}$ e $\sigma_{\varepsilon_i}^2 = \beta_1^2\sigma_x^2(1 - k_{y_i})/k_{y_i}$ para $i = 1, \ldots, n$, pois ambas expressões dependem da variância σ_x^2. Adotamos a distribuição uniforme (U) no intervalo (a', b') para gerar os valores de k_{x_i} e k_{y_i}. No modelo de regressão heteroscedástico analisamos dois cenários. No Cenário 1, quando consideramos $k_{x_i} \sim \mathrm{U}(0,5; 0,6)$ e $k_{y_i} \sim \mathrm{U}(0,55; 0,65)$ tem-se maiores valores para as variâncias dos erros de medição. No Cenário 2, quando $k_{x_i} \sim \mathrm{U}(0,7; 0,9)$ e $k_{y_i} \sim \mathrm{U}(0,75; 0,95)$ obtém-se menores valores para as variâncias dos erros de medição. Em seguida, no modelo de regressão homoscedástico, consideramos somente um cenário, em que k_x é igual à média dos k_{x_i}'s e k_y igual à média dos k_{y_i}'s com k_{x_i} e k_{y_i} referentes ao Cenário 2 descrito acima. Em relação ao número de réplicas consideramos as seguintes situações: (1) $r_i = s_i = 2$ (número mínimo de réplicas), (2) a escolha baseada no conjunto dos dados reais, ou seja, em que r_i e s_i estão entre 2 a 18, sendo neste caso réplicas desbalanceadas, (3) consideramos o dobro do número de réplicas existentes dos dados reais, ou seja, r_i e s_i variam entre 4 a 36, (4) o número de réplicas balanceadas, dado por $r_i = s_i = 20$, (5) quando o número de réplicas é $r_i = s_i = 40$, e por fim (6) o número de réplicas balanceadas $r_i = s_i = 200$.

4.1 Viés e REQM

Como medidas de avaliação utilizamos o viés e a raiz do erro quadrático médio (REQM) simulados. O viés simulado do estimador foi obtido pela diferença entre a média das estimativas do parâmetro e o valor verdadeiro do parâmetro em questão. A raiz quadrada do erro quadrático médio simulado (REQM), é calculada pela expressão

$$\text{REQM} = \left\{ \frac{1}{B} \sum_{b=1}^{B} (\widehat{\gamma}_b - \gamma)^2 \right\}^{1/2},$$

em que B é o número de simulações e γ é o verdadeiro valor do parâmetro do modelo. Enfatizamos que nesta dissertação, fixamos o tamanho da amostra em $n = 21$ e o número de réplicas cresce. Como pode ser visto na Tabela 4.1.

Tabela 4.1: Cenário 1 - Viés simulado e REQM simulado das estimativas de β_0 e β_1 obtidas pelos método de MV e pelo MPV - $k_{x_i} \sim \text{U}(0,5; 0,6)$ e $k_{y_i} \sim \text{U}(0,55; 0,65)$.

Situação	Número de réplicas	Método de estimação	Viés		REQM	
			β_0	β_1	β_0	β_1
	$r_i = 2$	MV	0,407	-0,314	0,875	0,662
1	$s_i = 2$	MPV	0,135	-0,099	0,835	0,641
	$r_i = 2$ a 18	MV	0,238	-0,179	0,411	0,299
2	$s_i = 2$ a 18	MPV	0,032	-0,026	0,270	0,198
	$r_i = 4$ a 36	MV	0,079	-0,059	0,139	0,102
3	$s_i = 4$ a 36	MPV	0,009	-0,008	0,105	0,076
	$r_i = 20$	MV	0,055	-0,042	0,111	0,083
4	$s_i = 20$	MPV	-0,001	0,000	0,098	0,073

Podemos observar na Tabela 4.1 que quando dispomos de um número mínimo de réplicas de x e y, ou seja, na situação (1), o método MPV apresenta melhores resultados do que o método MV, em termos de viés simulado das estimativas de β_0 e β_1 e analogamente acontece nas situações (2), (3) e (4). Em relação à REQM, o método de MPV mostra melhor desempenho do que pelo método MV quando consideramos o número de réplicas mínimas (situação (1)), assim como também acontece nas situações (2) e (3). Por fim,

na última situação $r_i = s_i = 20$, o método de MPV mostrou-se com melhor desempenho em contraste com o método de MV. As medidas de viés simulado (em módulo) e REQM diminuem à medida que o número de réplicas aumenta com o tamanho de amostra $n = 21$ fixo, conforme esperado. A seguir apresentamos os resultados das medidas de viés e REQM simulados correspondentes ao Cenário 2, descrito no início deste capítulo, dados na Tabela 4.2.

Tabela 4.2: Cenário 2 - Viés simulado e REQM simulado das estimativas de β_0 e β_1 obtidas pelos métodos de MV e pelo MPV - $k_{x_i} \sim \mathrm{U}(0,7; 0,9)$ e $k_{y_i} \sim \mathrm{U}(0,75; 0,95)$.

Situação	Número de réplicas	Método de estimação	Viés		REQM	
			β_0	β_1	β_0	β_1
	$r_i = 2$	MV	0,118	-0,117	0,202	0,170
1	$s_i = 2$	MPV	0,045	-0,040	0,257	0,197
	$r_i = 2$ a 18	MV	0,015	-0,013	0,092	0,070
2	$s_i = 2$ a 18	MPV	0,010	-0,009	0,113	0,081
	$r_i = 4$ a 36	MV	0,002	-0,001	0,056	0,040
3	$s_i = 4$ a 36	MPV	0,002	-0,002	0,057	0,041
	$r_i = 20$	MV	0,004	-0,003	0,051	0,038
4	$s_i = 20$	MPV	0,002	-0,001	0,052	0,039

Notamos que os vieses das estimativas dos parâmetros β_0 e β_1 no Cenário 2 do modelo de regressão heteroscedástico, apresentados na Tabela 4.2, têm melhor desempenho quando utilizamos o método MPV em contraste com o método de MV, nas situações (1), (2) e (4). Já na situação (3) o método de MV mostrou os menores valores de viés em módulo. Além disso, na Tabela 4.2, nas situações de réplicas (1) e (2), o método de MV apresentou melhor comportamento do que o método de MPV em relação à REQM. Ainda com relação a esta métrica, os resultados obtidos nas situações (3) e (4) têm melhor desempenho pelo método MV, porém, nota-se que os valores são bem próximos a aqueles obtidos pelo método MPV. Verificamos também, que o viés simulado (em módulo) dos estimadores dos parâmetros β_0 e β_1 pelos dois métodos diminui à medida que o número de réplicas aumenta, tendendo a 0, o qual acontece devido a que os estimadores de β_0 e β_1

são consistentes. Após analisar os resultados obtidos para os Cenários 1 e 2 do modelo de regressão heteroscedástico, foram analisados os vieses e a REQM simulados para o modelo de regressão homoscedástico, cujos resultados calculados são apresentados na Tabela 4.3.

Tabela 4.3: Modelo homoscedástico - Viés simulado e REQM simulado das estimativas de β_0 e β_1 obtidas pelos método de MV e pelo MPV - $k_x \sim \mathrm{U}(0,7; 0,9)$ e $k_y \sim \mathrm{U}(0,75; 0,95)$.

Situação	Número de réplicas	Método de estimação	Viés		REQM	
			β_0	β_1	β_0	β_1
	$r_i = 2$	MV	-0,001	0,001	0,174	0,131
1	$s_i = 2$	MPV	-0,004	0,004	0,187	0,141
	$r_i = 2$ a 18	MV	0,005	-0,004	0,079	0,058
2	$s_i = 2$ a 18	MPV	0,009	-0,008	0,082	0,059
	$r_i = 4$ a 36	MV	0,002	-0,001	0,059	0,042
3	$s_i = 4$ a 36	MPV	0,003	-0,002	0,061	0,044
	$r_i = 20$	MV	0,004	-0,003	0,052	0,039
4	$s_i = 20$	MPV	0,001	-0,001	0,055	0,041

Para o modelo de regressão homoscedástico, conforme Tabela 4.3, observamos que, em módulo, os vieses dos estimadores de β_0 e β_1, nas situações em (1), (2) e (3), o método de MV mostrou melhores resultados do que o método MPV. Para $r_i = s_i = 20$, o método de MPV apresentou melhor desempenho em termo de viés simulado (em módulo), porém, os valores da REQM foram maiores do que aqueles encontradas pelo método de MV. Para as situações (1),(2) e (3), notamos que em relação à REQM, pelos métodos de MV e MPV, os resultados obtidos são bem próximos, mas sendo o método MV com melhor desempenho, além disso, diminuem conforme o número de réplicas cresce, como esperado.

4.2 Desvio padrão e erro padrão

Como medidas de dispersão das estimativas, apresentamos o desvio padrão (DP) e o erro padrão (EP). O desvio padrão amostral do estimador $\widehat{\gamma}$ é obtido por

$$\mathrm{DP}(\widehat{\gamma}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\gamma}_b - \overline{\widehat{\gamma}})^2 \right\}^{1/2}, \tag{4.1}$$

em que $\overline{\widehat{\gamma}}$ é a média dos B valores $\widehat{\gamma}$. O erro padrão (EP) estimado é a média dos erros padrão assintóticos obtidos das matrizes $\boldsymbol{\mathcal{I}}$ e $\boldsymbol{\Sigma}$ dadas em (3.31) e (3.39), respectivamente, quando as variâncias dos erros de medida são heteroscedásticas e analogamente são calculadas para o caso de variâncias homoscedásticas.

Tabela 4.4: Cenário 1 - Desvio e erro padrão simulados das estimativas de β_0 e β_1 obtidas pelos métodos de MV e pelo MPV - $k_{x_i} \sim \mathrm{U}(0,5; 0,6)$ e $k_{y_i} \sim \mathrm{U}(0,55; 0,65)$.

Situação	Número de réplicas	Método de estimação	DP		EP	
			β_0	β_1	β_0	β_1
1	$r_i = 2$	MV	0,774	0,583	0,124	0,093
	$s_i = 2$	MPV	0,824	0,633	0,250	0,188
2	$r_i = 2$ a 18	MV	0,335	0,240	0,111	0,080
	$s_i = 2$ a 18	MPV	0,268	0,197	0,119	0,086
3	$r_i = 4$ a 36	MV	0,115	0,083	0,091	0,066
	$s_i = 4$ a 36	MPV	0,105	0,076	0,093	0,067
4	$r_i = 20$	MV	0,097	0,072	0,088	0,066
	$s_i = 20$	MPV	0,098	0,073	0,090	0,067

As Tabelas 4.4 e 4.5, para o modelo de regressão heteroscedástico, nos mostram o DP amostral simulado e o EP assintótico dos estimadores de $\boldsymbol{\phi} = (\beta_0, \beta_1)^t$, para dois diferentes cenários, ou seja, valores diferentes para $\sigma^2_{\delta_i}$ e $\sigma^2_{\varepsilon_i}$ para $i = 1, \ldots, n$, as variâncias dos erros de medição e diferentes números de réplicas. O estudo de simulação tem o objetivo de avaliar o comportamento das matrizes de covariâncias assintóticas, apresentadas nas Seções 3.2.1 e 3.2.2, e determinou que os erros padrão assintóticos dos estimadores MV de

β_0 e β_1, são menores que os respectivos erros padrão assintóticos dos estimadores de MPV em relação à Tabela 4.4, mas apresentaram resultados bem próximos. Além disso, os resultados do EP calculados pelo método MPV apresentaram valores menores de aqueles encontradas pelo método MV como são mostrados na Tabela 4.5, mas sendo também próximos. Cabe ressaltar que quando dispomos de valores altos para as variâncias dos erros de medição (Cenário 1) e em que a quantidade de réplicas é mínima (situação (1)), o valor do EP assintótico encontrado pelo método MV apresentou menor valor, pode ser útil nestas condições.

Tabela 4.5: Cenário 2 - Desvio e erro padrão simulados das estimativas de β_0 e β_1 obtidas pelos método de MV e pelo MPV - $k_{x_i} \sim \mathrm{U}(0,7; 0,9)$ e $k_{y_i} \sim \mathrm{U}(0,75; 0,95)$.

Situação	Número de	Método de	DP		EP	
	réplicas	estimação	β_0	β_1	β_0	β_1
	$r_i = 2$	MV	0,129	0,096	0,086	0,064
1	$s_i = 2$	MPV	0,186	0,134	0,073	0,055
	$r_i = 2$ a 18	MV	0,078	0,057	0,074	0,055
2	$s_i = 2$ a 18	MPV	0,099	0,069	0,065	0,047
	$r_i = 4$ a 36	MV	0,055	0,040	0,056	0,042
3	$s_i = 4$ a 36	MPV	0,056	0,040	0,051	0,037
	$r_i = 20$	MV	0,051	0,038	0,050	0,037
4	$s_i = 20$	MPV	0,051	0,038	0,048	0,036

É possível notar a redução do valor do EP assintótico dos estimadores de MV e MPV dos parâmetros, β_0 e β_1, para a situação em que o número de réplicas aumenta e $n = 21$ fixo. No Cenário 1 e 2 dadas nas Tabelas 4.4 e 4.5, o DP simulado manteve o mesmo comportamento do EP assintótico em relação a β_0 e β_1, quando dispomos de número de réplicas variando entre 4 e 36 e quando ambas réplicas são constantes, ou seja, $r_i = s_i = 20$. Para $r_i = s_i = 2$, os valores de DP simulado das estimativas de MV foram menores do que os valores obtidos por MPV.

Por outro lado, na Tabela 4.6 mostra-se o DP amostral e EP assintótico dos esti-

madores de $\boldsymbol{\phi}^* = (\beta_0, \beta_1)^t$ na situação em que as variâncias dos erros de medição são homoscedásticas. Notamos que o comportamento dos desvios padrão das estimativas de MV de $\boldsymbol{\phi}^*$, ao contrário dos desvios padrão das estimativas de MPV, são menores em todas as situações, ou seja, para todas as situações do número das réplicas. Em relação ao EP, os valores obtidos pelo método de MPV foram menores do que os valores obtidos pelo método de MV nas situações de réplicas (2), (3) e (4), porém, apresentaram valores muito próximos. Além disso, as medidas de DP e EP obtidas pelos métodos de MV e MPV mantiveram um comportamento assintótico próximos, assim como também a redução dos erros padrão dos estimadores de β_0 e β_1.

Tabela 4.6: Modelo homoscedástico - Desvio e erro padrão das estimativas de β_0 e β_1 obtidas pelos método de MV e pelo MPV - $k_x \sim \mathrm{U}(0,7; 0,9)$ e $k_y \sim \mathrm{U}(0,75; 0,95)$.

Situação	Número de réplicas	Método de estimação	DP		EP	
			β_0	β_1	β_0	β_1
1	$r_i = 2$	MV	0,174	0,131	0,177	0,133
	$s_i = 2$	MPV	0,187	0,141	0,179	0,135
2	$r_i = 2$ a 18	MV	0,079	0,057	0,086	0,065
	$s_i = 2$ a 18	MPV	0,081	0,059	0,078	0,057
3	$r_i = 4$ a 36	MV	0,058	0,042	0,063	0,047
	$s_i = 4$ a 36	MPV	0,061	0,044	0,057	0,041
4	$r_i = 20$	MV	0,052	0,039	0,054	0,040
	$s_i = 20$	MPV	0,055	0,041	0,053	0,039

4.3 Amplitude média e probabilidade de cobertura dos intervalos de confiança

Nesta seção adicionamos a situação (5) na qual as quantidades das réplicas serão $r_i = s_i = 200$. Nas Tabelas (4.7)-(4.9) apresentamos as probabilidades de cobertura (PC)

dos intervalos de confiança para os parâmetros de interesse, isto é, β_0 e β_1, com coeficiente de confiança nominal 95%. Esta medida foi obtida por

$$\mathrm{PC} = \frac{\#\{\phi_\iota \in \mathrm{IC}[\phi_\iota; 1-\alpha]\}_{b=1}^{B}}{B}, \tag{4.2}$$

em que ϕ_ι é o ι-ésimo elemento de $\boldsymbol{\phi} = (\beta_0, \beta_1)^t$, B é o número de repetições e $\mathrm{IC}[\phi_\iota; 1-\alpha]$ é o intervalo de confiança para ϕ_ι com coeficiente nominal $1-\alpha$. Na Tabela 4.7 são apresentados os resultados referentes a amplitude média (AM) e a probabilidade de cobertura (PC) dos intervalos de confiança nas situações (1)-(4) e (6) considerando o modelo de regressão heteroscedástico.

Tabela 4.7: Cenário 1 - Amplitude média e probabilidade de cobertura dos intervalos com 95% de confiança. Métodos de MV e MPV - $k_{x_i} \sim U(0,5; 0,6)$ e $k_{y_i} \sim U(0,55; 0,65)$.

Situação	Número de réplicas	Método de estimação	AM		PC (%)	
			β_0	β_1	β_0	β_1
1	$r_i = 2$	MV	0,485	0,364	26,86	26,70
	$s_i = 2$	MPV	0,980	0,738	45,64	45,90
2	$r_i = 2$ a 18	MV	0,436	0,315	51,10	49,98
	$s_i = 2$ a 18	MPV	0,466	0,338	78,82	79,78
3	$r_i = 4$ a 36	MV	0,358	0,259	79,96	79,38
	$s_i = 4$ a 36	MPV	0,364	0,264	91,58	91,42
4	$r_i = 20$	MV	0,344	0,258	86,60	86,38
	$s_i = 20$	MPV	0,351	0,263	93,32	93,32
6	$r_i = 200$	MV	0,112	0,084	93,64	93,54
	$s_i = 200$	MPV	0,112	0,084	94,64	94,52

Notamos que os valores da medida de amplitude média (AM) dos intervalos de confiança, em relação de β_0 e β_1, obtidos pelo método de MV foi menor em contraste com o método MPV nas quatro primeiras situações. Observe que na situação (5), as medidas de AM obtidas por ambos os métodos de estimação são bem semelhantes. Em relação à probabilidade de cobertura dos IC's, o método de MPV mostrou probabilidades maiores do que o método de MV nas situações (1)-(4). Cabe enfatizar que na situação (1) a PC

encontrada pelo método de MV mostrou-se uma taxa com péssimo desempenho. A fim de encontrar o número de réplicas mínimo para atingir o coeficiente de confiança nominal foi dada pela situação (6). Neste sentido, o método MPV apresentou probabilidade de cobertura mais próxima no nível nominal (95%). Similarmente, observamos na Tabela 4.8 um desempenho semelhante dos estimadores.

Convém ressaltar que na Tabela (4.7) as amplitudes médias dos IC's são maiores do que na Tabela (4.8), isso pelo fato de considerar valores para as variâncias dos erros de medição maiores e menores, respectivamente.

Tabela 4.8: Cenário 2 - Amplitude média e probabilidade de cobertura dos intervalos com 95% de confiança. Métodos de MV e MPV - $k_{x_i} \sim U(0,7; 0,9)$ e $k_{y_i} \sim U(0,75; 0,95)$.

Situação	Número de réplicas	Método de estimação	AM		PC (%)	
			β_0	β_1	β_0	β_1
	$r_i = 2$	MV	0,231	0,173	29,68	29,18
1	$s_i = 2$	MPV	0,240	0,180	36,82	36,64
	$r_i = 2$ a 18	MV	0,232	0,167	55,46	55,14
2	$s_i = 2$ a 18	MPV	0,238	0,172	73,66	74,82
	$r_i = 4$ a 36	MV	0,166	0,120	80,70	79,12
3	$s_i = 4$ a 36	MPV	0,166	0,120	87,60	86,30
	$r_i = 20$	MV	0,162	0,121	89,90	88,50
4	$s_i = 20$	MPV	0,163	0,122	92,56	91,70
	$r_i = 200$	MV	0,050	0,038	94,86	94,38
6	$s_i = 200$	MPV	0,050	0,038	94,90	94,56

No modelo de regressão homoscedástico, com base na Tabela 4.9, as amplitudes médias dos IC's para β_0 e β_1, obtidas por ambos os métodos de estimação são próximos, mas ainda assim pelo método de MPV apresentam valores abaixo do que o método MV. Em relação ao método de MV, ressaltamos que as probabilidades de cobertura se mostraram mais próximos do coeficiente de confiança nominal 95% quando consideramos as quantidades de réplicas $r_i = s_i = 200$.

Tabela 4.9: Modelo homoscedástico - Amplitude média e probabilidade de cobertura dos intervalos com 95% de confiança. Métodos de MV e MPV - $k_x \sim \mathrm{U}(0,7; 0,9)$ e $k_y \sim \mathrm{U}(0,75; 0,95)$.

Situação	Número de réplicas	Método de estimação	AM		PC (%)	
			β_0	β_1	β_0	β_1
1	$r_i = 2$	MV	0,546	0,411	93,96	93,82
	$s_i = 2$	MPV	0,541	0,407	93,18	93,04
2	$r_i = 2$ a 18	MV	0,293	0,220	95,94	96,62
	$s_i = 2$ a 18	MPV	0,265	0,192	93,70	93,72
3	$r_i = 4$ a 36	MV	0,213	0,160	96,08	97,04
	$s_i = 4$ a 36	MPV	0,192	0,139	93,22	93,06
4	$r_i = 20$	MV	0,182	0,137	94,84	94,68
	$s_i = 20$	MPV	0,177	0,133	94,46	94,14
6	$r_i = 200$	MV	0,080	0,060	95,28	95,06
	$s_i = 200$	MPV	0,078	0,058	94,62	94,40

É importante notar que foram observados os seguintes comportamentos esperados dos estimadores de MV e MPV, comuns a todas as tabelas desta seção:

- à medida que aumentamos o número de réplicas com n fixo, as medidas amplitude média dos IC's de β_0 e β_1, diminuem;
- as probabilidades de cobertura dos IC's obtidos pelos métodos de MV e MPV convergem para o coeficiente de confiança nominal conforme aumentamos os valores de r_i e s_i para o tamanho amostral n fixado.

4.4 Taxas de rejeição dos testes

Nesta seção analisamos o desempenho dos testes da razão de verossimilhanças (RV), razão de pseudoverossimilhanças (RPV), Wald (W), escore (S_R), gradiente (S_T) e $C(\alpha)$ de Neyman em relação às taxas de rejeição de H_0 sob H_0 e H_1 considerando os cenários dos valores das variâncias dos erros de medição descritos no inicio do Capítulo 4. Enfatizamos

que os testes mencionados acima foram utilizados para testar as Hipóteses 2 em (3.52) e, no caso, das Hipóteses 1 dada em (3.51) foi utilizado somente o teste da razão de verossimilhanças. Utilizamos a probabilidade do erro de tipo I, que foi estimada gerando amostras sob a hipótese H_0 e calculando a proporção de simulações em que o teste rejeitou a hipótese nula. A taxa de rejeição de H_0 sob H_1 foi obtida gerando amostras sob a hipótese H_1 e calculando a proporção de simulações em que o teste rejeitou a hipótese nula. Geramos 5000 conjuntos de dados, como foi mencionado no inicio do Capítulo 4.

4.4.1 Homoscedasticidade

Sejam as Hipóteses 1 apresentadas em (3.51). Consideramos que sob H_0, as variâncias dos erros $\sigma^2_{\delta_i}$ e $\sigma^2_{\varepsilon_i}$ são dadas respectivamente por $\sigma^2_{\delta_i} = \sigma^2_x(1-k_x)/k_x$ e $\sigma^2_{\varepsilon_i} = \beta_1^2\sigma^2_x(1-k_y)/k_y$, $i = 1,\ldots,n$, em que k_x e k_y são as médias dos valores dos k_{x_i}'s e k_{y_i}'s correspondentes no Cenário 2 do modelo heteroscedástico. Sob H_1, analisamos a situação em que as variâncias dos erros de medição são geradas por meio das expressões $\sigma^2_{\delta_i} = \sigma^2_x(1-k_{x_i})/k_{x_i}$ e $\sigma^2_{\varepsilon_i} = \sigma^2_y(1-k_{y_i})/k_{y_i}$ para $i = 1,\ldots,n$, e consideramos os valores de k_{x_i} e k_{y_i} como no Cenário 1, ou seja, $k_{x_i} \sim \mathrm{U}(0,5;0,6)$ e $k_{y_i} \sim \mathrm{U}(0,55;0,65)$.

Tabela 4.10: Taxas de rejeição (%) das hipóteses H_0 e H_1 - Teste RV - homoscedasticidade das variâncias com $n = 21$ para um nível de significância $\alpha = 0,05$.

Situação	Número de réplicas	Sob H_0	Sob H_1
	$r_i = 2$		
1	$s_i = 2$	38,72	51,82
	$r_i = 2$ a 18		
2	$s_i = 2$ a 18	14,32	71,22
	$r_i = 4$ a 36		
3	$s_i = 4$ a 36	8,26	95,54
	$r_i = 20$		
4	$s_i = 20$	6,18	100,00
	$r_i = 40$		
5	$s_i = 40$	5,16	100,00

É importante comentar que sob H_0, em $0,44\%$ das amostras geradas para o número de réplicas $r_i = s_i = 2$, o valor da estatística da razão de verossimilhanças foi negativo. Nesse caso, essas amostras foram descartadas da análise. Gráficos de quantis para a estatística do teste da razão de verossimilhanças também foram apresentados na Figura (4.1).

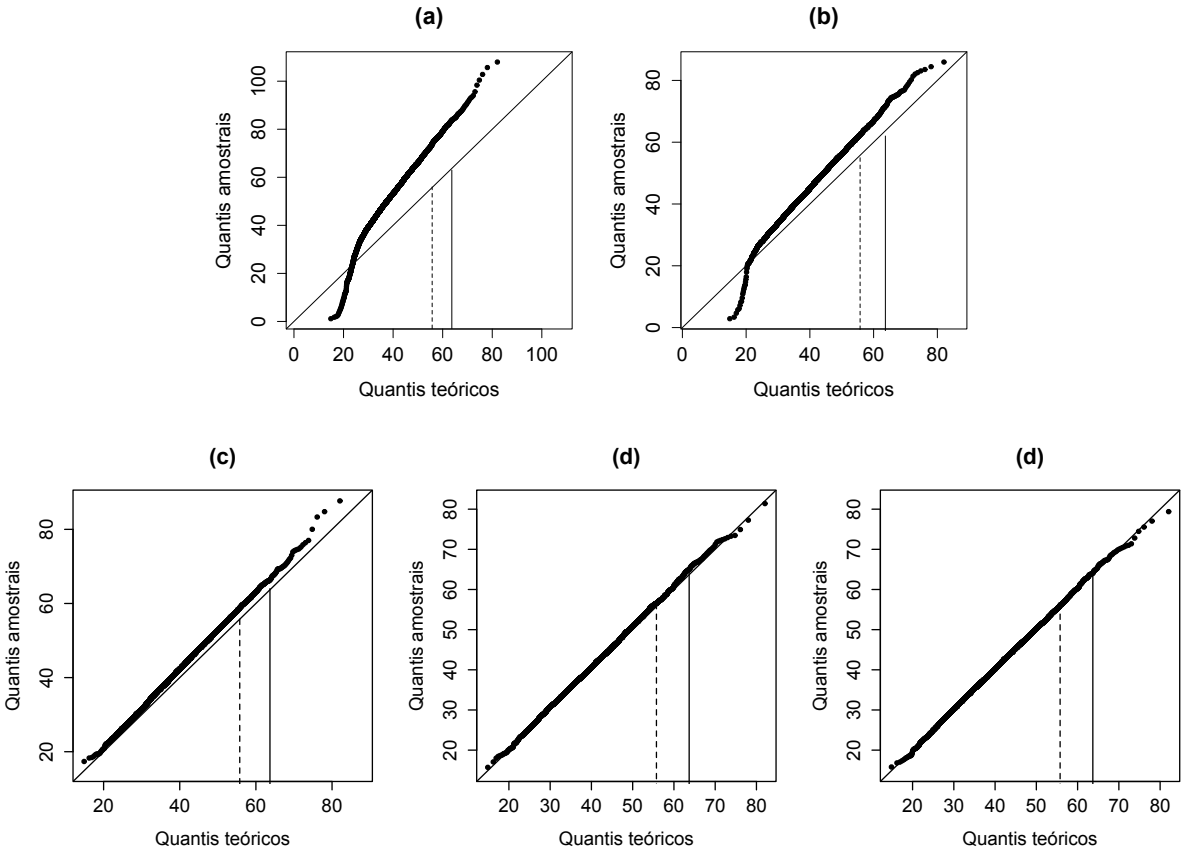

Figura 4.1: Gráfico de quantis da estatística de teste ξ_{RV} - homoscedasticidade (a) $r_i = s_i = 2$, (b) r_i, s_i entre 2 a 18, (c) r_i, s_i entre 4 a 36, (d) $r_i = s_i = 20$ e (e) $r_i = s_i = 40$ com $n = 21$ para um nível de significância $\alpha = 0,05$, quantil $1 - \alpha = 0,95$ (linha vertical tracejada) e $1 - \alpha = 0,99$ (linha vertical cheia).

Verificamos na Figura (4.1) (a) que quando o número de réplicas é $r_i = s_i = 2$, não há indícios de que a estatística da RV segue uma distribuição χ^2_{2n-2}, ou seja, uma distribuição χ^2_{40}. Na Tabela (4.10) em relação à estatística da ξ_{RV}, vemos que as taxas de rejeição

sob H_0 são altas, sendo os valores distantes do valor nominal dado por 5%. Entretanto, observamos que fixando o tamanho amostral $n = 21$ e aumentando o número de réplicas, percebemos que as taxas de rejeição sob H_0 diminuem. Na Figura (4.1) observamos que os quantis teóricos e amostrais se aproximam com o aumento do número de réplicas. Percebemos que na situação (4), com o número de réplicas $r_i = s_i = 20$, a taxa de rejeição de H_0 sob H_0 é próxima ao nível nominal 5% e o resultado melhora quando o número de réplicas é incrementada para $r_i = s_i = 40$, como pode ser visto na Figura 4.1 (e), indicando que para esta quantidade de réplicas o resultado assintótico é satisfatório. Em relação às taxas de rejeição de H_0 sob H_1, notamos que os valores encontrados em todas as situações foram altos, conforme esperado. Portanto, concluímos que o teste da razão de verossimilhanças para testar a homoscedasticidade das variâncias é insatisfatório para as situações em que dispomos de quantidade de réplicas inferior a 40, no cenário analisado. O teste escore também poderia ser utilizado para testar as Hipóteses 1, mas não será apresentado neste trabalho.

4.4.2 Vieses aditivo e multiplicativo

Em seguida, como mencionado na Seção 3.3, analisamos o teste das Hipóteses 2, já que é de interesse em problemas de comparação de métodos de medição, em que por meio deste teste dada em (3.52) verificamos vícios aditivo e multiplicativo de um método de medição em relação ao outro método analisado. Sob H_0, $\beta_0 = 0$, $\beta_1 = 1$ e as variâncias dos erros de medição são iguais às variâncias dos cenários descritos no inicio do Capitulo 4, tanto para o modelo de regressão heteroscedástico quanto homoscedástico. Sob H_1, como o espaço paramétrico de (β_0, β_1) é ilimitado, esses parâmetros podem tomar infinitos valores sob H_1. Dessa forma, consideramos a situação H_1: $(\beta_0, \beta_1)^t = (-0,08;\ 0,9)^t$.

Para o modelo de regressão heteroscedástico, os resultados obtidos são apresentados na Tabela (4.11), segundo os testes da ξ_{RV}, W, S_R, S_T, $C(\alpha)$ e da ξ_{RPV}, e a fim de completar os resultados, apresentamos os gráficos de quantis para todas as estatísticas

mencionadas acima. Tais gráficos referem-se apenas ao Cenário 1, uma vez que para o Cenário 2 a situação é análoga, e com situação de réplicas (4) (Figura 4.2).

Tabela 4.11: Modelo heteroscedástico - Taxas de rejeição (%) das hipóteses H_0 - Testes RV, W, S_R, S_T, $C(\alpha)$ e RPV - para um nível de significância $\alpha = 0,05$.

Situação	Número de réplicas	Estatísticas de teste	Cenário 1		Cenário 2	
			Sob H_0	Sob H_1	Sob H_0	Sob H_1
1	$r_i = 2$ $s_i = 2$	RV	30,64	85,70	28,68	99,14
		W	63,32	96,86	67,68	99,94
		S_R	3,32	45,76	3,16	85,38
		S_T	11,62	77,86	11,64	98,80
		$C(\alpha)$	82,94	97,46	79,52	99,90
		RPV	72,86	99,96	74,70	99,94
2	$r_i = 2$ a 18 $s_i = 2$ a 18	RV	8,18	99,86	8,16	100,00
		W	11,90	99,98	13,64	100,00
		S_R	5,74	99,82	6,26	100,00
		S_T	6,72	99,96	7,28	100,00
		$C(\alpha)$	26,34	99,56	25,56	99,94
		RPV	23,80	99,92	28,00	100,00
3	$r_i = 4$ a 36 $s_i = 4$ a 36	RV	6,40	100,00	6,06	100,00
		W	6,82	100,00	6,48	100,00
		S_R	6,76	100,00	7,02	100,00
		S_T	6,10	100,00	5,88	100,00
		$C(\alpha)$	8,74	100,00	7,78	100,00
		RPV	8,52	100,00	9,10	100,00
4	$r_i = 20$ $s_i = 20$	RV	5,92	100,00	6,30	100,00
		W	7,04	100,00	7,54	100,00
		S_R	5,04	100,00	5,34	100,00
		S_T	5,72	100,00	6,06	100,00
		$C(\alpha)$	7,48	100,00	7,18	100,00
		RPV	6,28	100,00	6,50	100,00

Analisando a Tabela 4.11, verificamos que para o número de réplicas mínimo, ou seja, $r_i = s_i = 2$, as taxas de rejeição sob H_0 referentes aos Cenários 1 e 2 foram altas exceto

para a estatística de teste S_R que teve um comportamento conservador. Para o número de réplicas nas situações (2) e (3), observa-se que, em relação ao teste escore, as taxas de rejeição sob H_0 são menores do que as taxas encontradas pelos outros testes. No entanto, as taxas seguem sendo altas, mas o desempenho obtido pelo teste escore, S_R, se mostrou bem melhor na situação (2) e a estatística gradiente apresentou a menor taxa na situação (3), mas ainda ambas estatísticas não atingem ao valor nominal 5%, como pode ser observado nos Cenários 1 e 2. À medida que aumentamos o número de réplicas com $n = 21$ fixo, observamos que para a situação (4) a estatística S_R teve o melhor desempenho em relação às taxas de rejeição sob H_0, seguido pelos testes S_T e RV, assim como também pode ser visto na Figura 4.2 (a), (c) e (d), respectivamente. Nessa mesma situação, as estatísticas de teste W, $C(\alpha)$ e RPV apresentam valores que não são próximos ao valor nominal, o que requer o aumento de número de réplicas assim como também pode ser visto através do gráfico de quantis dado na Figura 4.2 (b), (e) e (f), respectivamente. Portanto, logramos verificar que todas as estatísticas de teste mostradas na Tabela 4.11 apresentam melhorias quando acrescentamos o número das réplicas, conforme esperado. Cabe ressaltar que o o teste escore apresentou melhor comportamento em relação às taxas de rejeição atingindo próximo do nível nominal.

Para o modelo de regressão homoscedástico, as taxas de rejeição de H_0, sob H_0 e H_1, utilizando as estatísticas descritas na Seção 3.3.2, são apresentadas na Tabela 4.12. Sob H_0, notamos que na situação (1), a estatística de teste RV apresentou melhor desempenho em relação as taxas de rejeição de H_0. Nas situações (2) e (3), as estatísticas de teste RV e RPV se mostraram com melhor desempenho próximos do valor nominal 5%. Entretanto, a estatística $C(\alpha)$ apresentou altas taxas de rejeição nas situações $(1) - (4)$. Por fim, na situação (4) notamos que os valores das taxas de rejeição sob H_0 utilizando as estatísticas RV, W, S_T e RPV se aproximam ao valor nominal (5%). Em relação ao poder do teste, vimos que as taxas de rejeição sob H_1, tanto para o modelo heteroscedástico quanto no modelo homoscedástico em todas as situações, foram altas as taxas de rejeição, como

esperado.

Na Figura (4.3) foram apresentados os gráficos de quantis para a situação (4), isto é, $r_i = s_i = 20$. Observe que também incluímos os quantis do 95% e 99% que são pelos valores $5,99$ e $9,21$, respectivamente. Notamos que na Figura (4.3) (a), (d) e (f) os quantis amostrais são próximos aos quantis teóricos.

Tabela 4.12: Modelo homoscedástico - Taxas de rejeição (%) das hipóteses de H_0 - Testes RV, W, S_R, S_T, $C(\alpha)$ e RPV para um nível de significância $\alpha = 0.05$

Situação	Número de réplicas	Estatísticas de teste	Sob H_0	Sob H_1
1	$r_i = 2$ $s_i = 2$	RV	4,62	82,64
		W	7,10	82,56
		S_R	13,46	31,32
		S_T	14,10	64,48
		$C(\alpha)$	19,40	62,50
		RPV	6,20	51,98
2	$r_i = 2$ a 18 $s_i = 2$ a 18	RV	5,10	99,68
		W	11,82	100,00
		S_R	8,58	92,70
		S_T	7,50	99,90
		$C(\alpha)$	14,00	99,10
		RPV	5,20	99,48
3	$r_i = 4$ a 36 $s_i = 4$ a 36	RV	4,66	100,00
		W	6,72	100,00
		S_R	8,10	100,00
		S_T	5,62	100,00
		$C(\alpha)$	13,40	100,00
		RPV	4,82	100,00
4	$r_i = 20$ $s_i = 20$	RV	4,68	100,00
		W	5,36	100,00
		S_R	5,98	100,00
		S_T	5,60	100,00
		$C(\alpha)$	8,10	100,00
		RPV	5,02	100,00

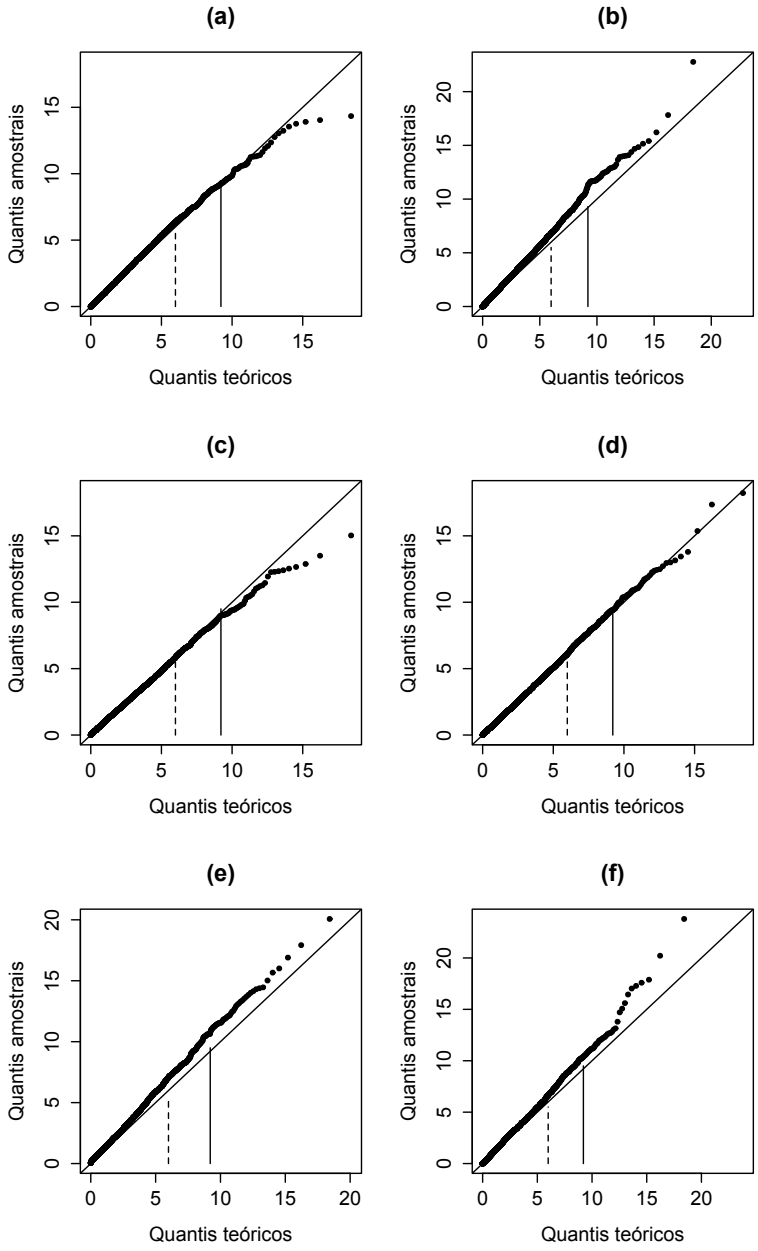

Figura 4.2: Cenário 1 - Gráfico de quantis da distribuição χ^2_2 para réplicas $r_i = s_i = 20$ - Testes (a) RV, (b) Wald (W), (c) escore S_R (d) gradiente S_T, (e) $C(\alpha)$ e (f) RPV para um nível de significância $\alpha = 0,05$, quantil $1 - \alpha = 0,95$ (linha vertical tracejada) e $1 - \alpha = 0,99$ (linha vertical cheia).

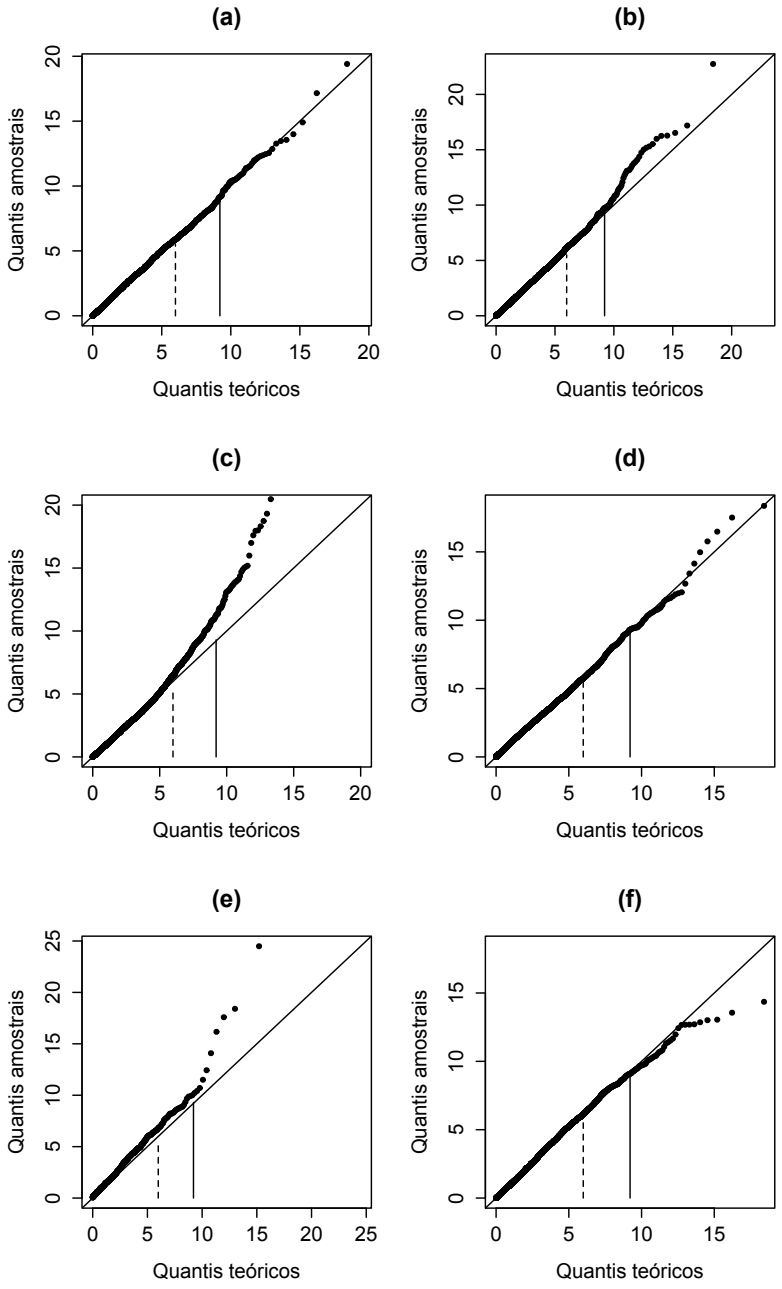

Figura 4.3: Modelo homoscedástico - Gráfico de quantis da distribuição χ^2_2 para réplicas r_i,$s_i = 20$ - Testes (a) RV, (b) Wald (W), (c) escore S_R (d) gradiente S_T, (e) $C(\alpha)$ e (f) RPV para um nível de significância $\alpha = 0,05$, quantil $1 - \alpha = 0,95$ (linha vertical tracejada) e $a - \alpha = 0,99$ (linha vertical cheia).

Capítulo 5

Aplicação

Em problemas de comparação de métodos analíticos é comum aplicar técnicas de regressão, como visto no Capítulo 1. Nestes problemas o principal objetivo é detectar possíveis vícios aditivo e multiplicativo de um método em relação ao outro, que é feito testando-se as Hipóteses 2 dadas em (3.52). Neste capítulo ajustamos o modelo (2.1)-(2.3) a um conjunto de dados reais relacionado a problemas de comparação de métodos de medição. Aplicamos os métodos de estimação desenvolvidos no Capítulo 3 e os testes da razão de verossimilhanças e escore para testar as Hipóteses 1 em (3.51) e as Hipóteses 2 em (3.52), respectivamente. As Hipóteses 2 permitem verificar a ausência de vícios aditivo e multiplicativo de um método em relação a outro. A adequação do ajuste do modelo foi verificada por meio do gráfico de envelope (Atkinson, 1985).

O conjunto de dados utilizado aqui foi retirado de Rasekh & Fieller (2003). Os dados referem-se a medições de concentrações de potássio (K), presente em amostras de cerâmicas egípcias. As medições foram realizadas utilizando duas técnicas diferentes denominadas análise de ativação de nêutrons (NAA) e espectrometria por plasma indutivamente acoplado (ICP). A unidade das medições não foi informada no artigo. Os potes foram coletados de diferentes localidades ao redor da antiga cidade egípcia de Amarna e cada peça possui um código de construção que será de utilidade na identificação. Assim, as

peças da mesma procedência e com o mesmo código são consideradas repetições. Desta forma, o conjunto de dados foi dividido em 21 grupos, em que cada grupo contém um certo número de peças. A quantidade de réplicas varia de 2 a 18. Convencionamos que Y_{ij} é o j-ésimo valor observado no i-ésimo grupo utilizando a técnica NAA, ao passo que X_{ik} denota o k-ésimo valor observado no i-ésimo grupo utilizando a técnica ICP, $j = 1, \ldots, s_i$, $k = 1, \ldots, r_i$ e $i = 1, \ldots, 21$.

Inicialmente realizamos uma breve análise descritiva das variáveis envolvidas no problema. A Tabela 5.1 mostra que estamos na situação em que as quantidades das réplicas são desbalanceadas, pois as médias de r_i e s_i são $7,5$ e $7,4$, respectivamente. Algumas medidas resumo das medições do teor de potássio em cada um dos 21 grupos, considerando as técnicas NAA e ICP, mostram que são bem próximos. As medidas Q_1 e Q_3 são o primeiro quartil e o terceiro quartil, respectivamente.

Tabela 5.1: Medidas resumo das variáveis.

Variáveis (Métodos)	Número de réplicas	Mínimo	Q_1	Mediana	Média	Q_3	Máximo
Y (NAA)	min: 2 média:7,5 max: 18	0,882	1,207	1,339	1,390	1,580	2,103
X (ICP)	min: 2 média: 7,4 max: 18	0,714	1,129	1,332	1,301	1,495	1,869

No gráfico da Figura 5.1 observa-se a dispersão das réplicas e de suas médias notando que uma relação linear parece fornecer uma boa aproximação.

O gráfico na Figura 5.2 mostra as médias e os desvios padrão das medições de concentrações do elemento K. Pode-se notar que em média, os valores do teor de K obtidos por ambas técnicas são bem semelhantes, mas os desvios padrão assumem valores distintos apresentando uma possível heteroscedasticidade das variâncias, que será verificada mais adiante.

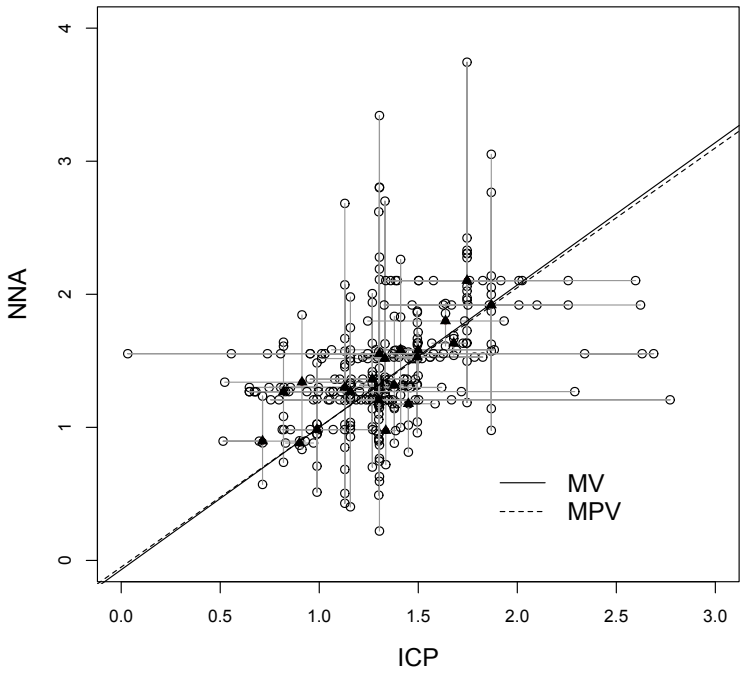

Figura 5.1: Gráfico de dispersão das réplicas (○) e suas médias (▲) com as retas segundo as estimativas de MV e MPV.

Desta forma, ajustamos os modelos de regressão dados na Seção 2.1.1 e 2.1.2 apresentado no Capítulo 2 ao conjunto de dados reais descrito no início deste capítulo, aplicando os métodos de máxima verossimilhança e de máxima pseudoverossimilhança, obtendo as estimativas dos parâmetros de interesse β_0 e β_1, assim como também utilizamos o método de reamostragem *bootstrap* paramétrico (Efron, 1993) para realizar inferências.

O ajuste do modelo de regressão (2.1)-(2.3) e seus casos particulares, pode ser verificado por meio do gráfico de envelope simulado. A técnica foi desenvolvida por Atkinson (1985). Para utilizar esta ferramenta gráfica, inicialmente encontramos as estimativas de máxima verossimilhança considerando os valores amostrais. De (2.4) podemos utilizar

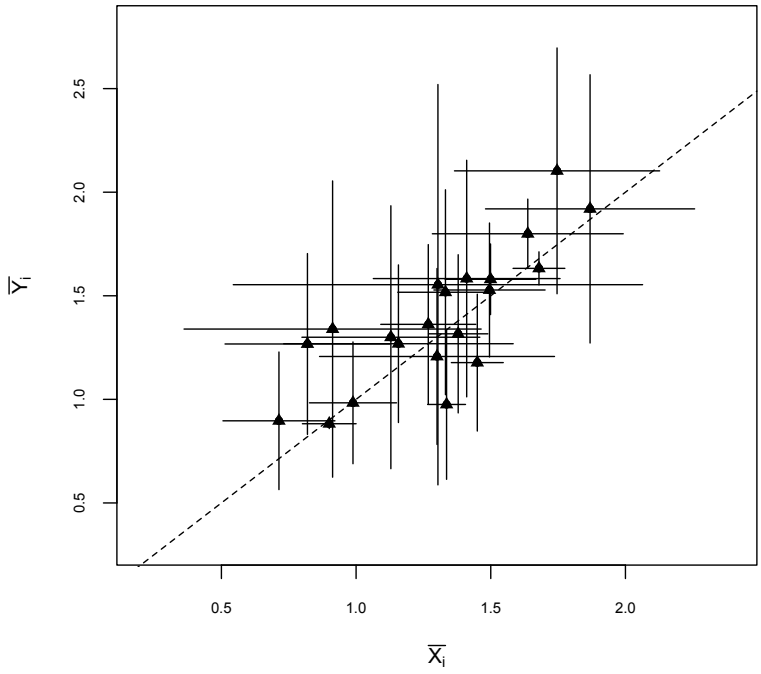

Figura 5.2: Médias de cada grupo (▲) i, $i = 1, \ldots, 21$, considerando NAA($\overline{Y}_{i.}$) e ICP ($\overline{X}_{i.}$) e seus respectivos desvios padrões com a reta tracejada identidade.

que sob o modelo,

$$(\boldsymbol{Z}_i - \boldsymbol{m}_i)^t \boldsymbol{V}^{-1} (\boldsymbol{Z}_i - \boldsymbol{m}_i) \sim \chi^2_{r_i+s_i} \quad \text{para } i = 1, \ldots, n.$$

Aplicando a transformação de Wilson-Hilferty (Johnson *et al.*, 1994), calculamos

$$\tau_i = 3\sqrt{\frac{r_i + s_i}{2}} \left\{ \left[\frac{(\boldsymbol{Z}_i - \boldsymbol{m}_i)^t \boldsymbol{V}_i^{-1} (\boldsymbol{Z}_i - \boldsymbol{m}_i)}{r_i + s_i} \right]^{1/3} - 1 + \frac{2}{9(r_i + s_i)} \right\} \; i = 1, \ldots, n, \quad (5.1)$$

com $\tau_i \overset{iid}{\sim} N(0,1)$, aproximadamente. Substituindo os parâmetros pelas suas respectivas estimativas de MV. Deste modo, obtemos n valores $\widehat{\tau}_i$. Ordenamos esses valores obtendo $\widehat{\tau}_{(1)} < \cdots < \widehat{\tau}_{(p)}$ e representamos os pontos $(\Phi^{-1}((i - 3/8)/(n + 1/4)), \widehat{\tau}_{(i)})$ em um gráfico,

em que Φ^{-1} é a função quantil da distribuição normal padrão. Para obter os limites do envelope, simulamos $G = 100$ conjuntos de dados, e para cada conjunto g gerado, encontramos as estimativas de máxima verossimilhança dos parâmetros de interesse. Então, novamente encontramos os valores de $\widehat{\tau}_{gi}$, $i = 1, \ldots, n$, substituindo os parâmetros da expressão dada em (5.1) pelas estimativas encontradas. Ordenamos os valores encontrados e por fim, representamos os pontos $(\Phi^{-1}((i-3/8)/(n+1/4)), \min_{g=1}^{G} \widehat{\tau}_{g(i)})$ e $(\Phi^{-1}((i-3/8)/(n+1/4)), \max_{g=1}^{G} \widehat{\tau}_{g(i)})$ para $i = 1, \ldots, n$, correspondendo aos limites inferior e superior do envelope. Os pontos $(\Phi^{-1}((i-3/8)/(n+1/4)), \sum_{g=1}^{G} \widehat{\tau}_{g(i)}/G)$ também são representados no gráfico.

A técnica descrita acima é aplicada ao conjunto de dados em questão, considerando os modelos de regressão homoscedástico e heteroscedástico, como visto na Figura 5.3. Note que o gráfico na Figura 5.3-(b) não apresenta afastamentos sérios em contraste com o gráfico em (a). Entretanto, o modelo de regressão homoscedástico não apresentou um ajuste adequado para os dados em questão. Como visto no gráfico na Figura 5.3-(b). Portanto, o modelo de regressão heteroscedástico se ajustou bem ao conjunto de dados de concentrações de K nas peças egípcias.

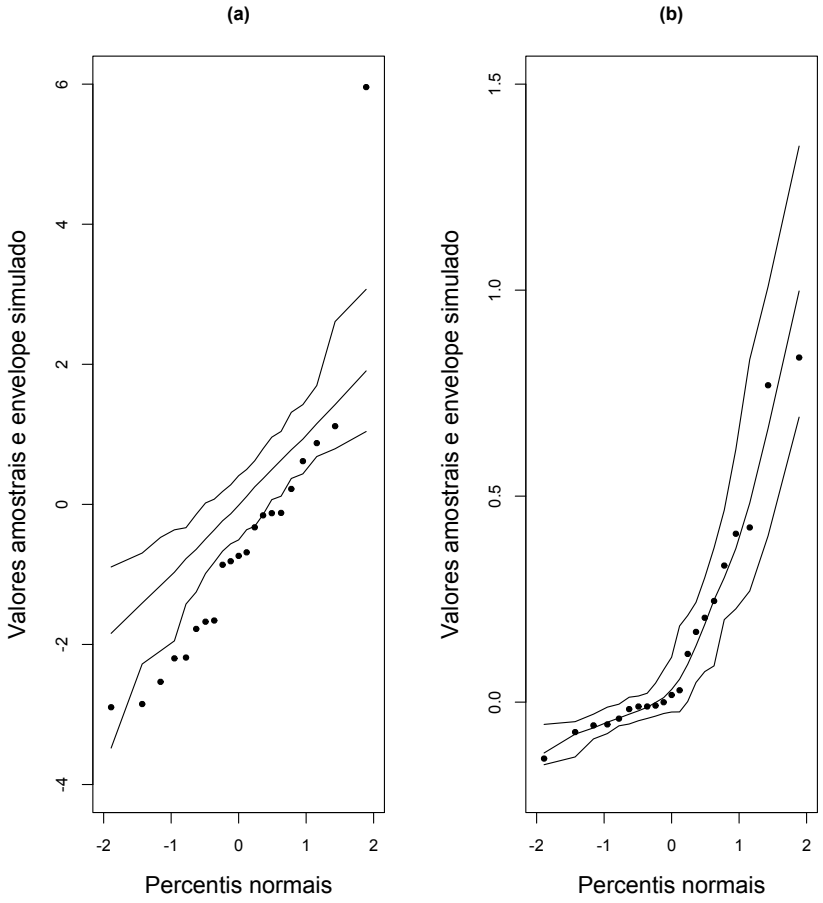

Figura 5.3: Gráfico normal de probabilidades com envelope simulado. Ajuste dos modelos de regressão (a) homoscedástico e (b) heteroscedástico com $n = 21$ fixo.

Na Tabela 5.2 são apresentadas as estimativas dos parâmetros β_0 e β_1 considerando o modelo de regressão heteroscedástico. Note que, pelos métodos MV e MPV, as estimativas de β_0 e β_1 são bem próximas, assim como também as estimativas *bootstrap* se comportam de modo similar. Não podemos esquecer que através da matriz de covariâncias assintótica dos estimadores de MV e MPV foi obtido o erro padrão (EP) assintótico, respectivamente. Observamos que, na Tabela 5.2, os valores dos EP assintóticos das estimativas dos parâmetros β_0 e β_1 obtidos pelo método de MV foram menores se comparando com o método

Tabela 5.2: Estimativas dos parâmetros pelos métodos de MV e MPV com seus respectivos erros padrões estimados.

Parâmetro	Método de	Amostra		*bootstrap*	
	estimação	Estimativa	EP	Estimativa	EP_b
β_0	MV	-0,068	0,162	-0,069	0,127
	MPV	-0,052	0,170	-0,064	0,165
β_1	MV	1,073	0,119	1,078	0,100
	MPV	1,057	0,125	1,075	0,130

de MPV. Mas, observe-se que os valores dos erros padrão, foram bem próximos. Além disso, os erros padrão *bootstrap*, EP_b, apesar de apresentarem valores menores do que os EP assintóticos, foram próximos. As retas ajustadas segundo os métodos de MV e MPV são mostradas no gráfico da Figura 5.1. A implementação computacional foi desenvolvida em linguagem R (R Core Team, 2013).

Na Tabela 5.3 apresentamos os intervalos de confiança *bootstrap* para β_0 e β_1, obtidos pelos método de máxima verossimilhança e pelo método de máxima pseudoverossimilhança. Para o cálculo do intervalo de confiança *bootstrap*, consideramos 5000 amostras *bootstrap*. Os intervalos foram obtidos utilizando as estimativas corrigidas *bootstrap* e seus respectivos erros padrão (EP_b), apresentados na Tabela 5.2, com 95% de confiança.

Tabela 5.3: Intervalos de confiança assintótico e *bootstrap* dos parâmetros de β_0 e β_1 obtidas pelos métodos de máxima verossimilhança e máxima pseudoverossimilhança - Teor de K nas cerâmicas egípcia.

Parâmetro	Método de	Assintótico	*bootstrap*
	estimação	(LI,LS)	(LI,LS)
β_0	MV	(-0,386; 0,249)	(-0,317; 0,180)
	MPV	(-0,385; 0,281)	(-0,388; 0,260)
β_1	MV	(0,840; 1,306)	(0,883; 1,273)
	MPV	(0,812; 1,302)	(0,820; 1,330)

A seguir aplicamos o teste da razão de verossimilhanças e o teste escore com o objetivo

de testar as Hipóteses 1 e 2 dadas em (3.51) e (3.52), respectivamente. Os valores da estatística da razão verossimilhanças e da estatística escore para as Hipóteses 1 e 2, respectivamente, com seus respectivos valores-p são apresentados na Tabela 5.4.

Tabela 5.4: Estatísticas da razão de verossimilhanças e escore para as Hipóteses 1 e 2, e respectivos valores-p.

Hipóteses 1		Hipóteses 2	
Estatística ξ_{RV}	Valor p	Estatística S_R	Valor p
$129,910$	$< 0,001$	$0,873$	$0,771$

Para as Hipóteses 1, observamos que o valor-p foi inferior ao nível de significância adotado (5%), levando à rejeição da hipótese nula de homoscedasticidade das variâncias. Em relação às Hipóteses 2, observamos pela Tabela 5.4 que o valor-p foi maior do que $\alpha = 5\%$, levando à não rejeição da hipótese nula, ou seja, a ausência de vícios aditivo e multiplicativo não é rejeitada. É importante destacar que no estudo de simulação realizado no Capítulo 4, em condições semelhantes às dos dados reais, a estatística escore para testar as Hipóteses 2 apresentou melhor desempenho em comparação com as outras estatísticas de teste abordadas na Seção 3.3.

Em síntese, pela análise realizada neste capítulo, concluímos que não há vícios aditivo e multiplicativo da técnica NAA em relação à técnica ICP, sendo como um de nossos objetivos principais a verificação da equivalência de um método de medição em relação ao outro, e por conseguinte em geral produzem medições semelhantes.

Capítulo 6

Considerações finais

6.1 Conclusão

Nesta dissertação apresentamos, como contribuição deste trabalho, um modelo com erros de medição sob enfoque estrutural para observações replicadas e seus casos particulares, isto é, foram apresentados os modelos de regressão heteroscedástico e homoscedástico. Neste sentido, estes modelos são modificações flexibles aos modelos considerados em Chan & Mak (1979) e Kulathinal *et al.* (2002). Estes modelos são utilizados, por exemplo, em diversas aplicações relacionadas com problemas de comparação de métodos de medição. Portanto, nesses casos o interesse é verificar a existência de vícios aditivo e multiplicativo de um método em relação ao outro.

Os métodos de máxima verossimilhança e máxima pseudoverossimilhança foram utilizados para estimar os parâmetros dos modelos propostos. O algoritmo EM foi desenvolvido para ambas abordagens de estimação logrando-se encontrar as expressões das estimativas em forma simples. Avaliamos numericamente o comportamento dos estimadores dos parâmetros de interesse calculando o viés e raiz do erro quadrático médio de 5000 conjuntos de amostras. As estimativas de máxima verossimilhança e máxima pseudoverossimilhança dos parâmetros β_0 e β_1, calculadas utilizando o algoritmo EM, mostraram que à medida

que o número de réplicas aumenta com tamanho amostral n fixo, o viés (em módulo) e a REQM simulados diminuem, conforme esperado. Em relação à probabilidade de cobertura dos intervalos de confiança, o método de máxima pseudoverossimilhança mostrou-se com melhor desempenho do que o método de máxima verossimilhança no modelo de regressão heteroscedástico. Em termos computacionais o método de máxima pseudoverossimilhança requer menos esforço computacional e é mais simples, do que o método de máxima verossimilhança. É recomendável utilizar o método de máxima pseudoverossimilhança como alternativa de estimação na presença de parâmetros de perturbação, pois apresenta boas propriedades assintóticas, como foram verificadas nas avaliações numéricas.

As matrizes de covariâncias assintóticas dos estimadores de máxima verossimilhança e de máxima pseudoverossimilhança foram encontradas em forma fechada, assim como também foram avaliadas numericamente. Em geral, os resultados apresentados do DP e EP assintótico mostraram que à medida que o número de réplicas r_i e s_i cresce com n mantido fixo, o DP e o EP diminuem, como esperado.

Utilizamos a estatística de teste da razão de verossimilhanças para testar as Hipóteses 1 (homoscedasticidade das variâncias dos erros de medição) e as estatísticas de teste da razão de verossimilhanças, Wald, escore, gradiente, $C(\alpha)$ de Neyman e razão de pseudoverossimilhanças para testar as Hipóteses 2 (Vieses aditivo e multiplicativo). Em relação aos testes de homoscedasticidade das variâncias dada na Seção 3.3, o teste da RV mostrou-se insatisfatório para situações em que a quantidade de réplicas disponível é inferior a 40.

Para os testes abordados na Seção 3.3, ou seja, RV, W, S_R, S_T, $C(\alpha)$ e RPV para testar as hipóteses de vícios aditivo e multiplicativo mostraram-se insatisfatórios quando o número de réplicas foi inferior a 20, exceto o teste escore S_R. Cabe ressaltar que o teste escore mostrou-se com o melhor desempenho em todas as situações dos Cenários 1 e 2, no modelo de regressão heteroscedástico. Para o único cenário do modelo de regressão homoscedástico, os testes da RV e da RPV mostraram-se com melhor desempenho em contraste com os testes score, gradiente, Wald e $C(\alpha)$ de Neyman.

No Capítulo 5 consideramos um conjunto de dados reais para avaliar as metodologias propostas nos Capítulos 2 e 3. Utilizamos o método de *bootstrap* como alternativa para realizar inferências.

6.2 Propostas de trabalhos futuros

Como possíveis propostas de trabalho futuros listamos:

1.- Considerar em (2.1) erro na equação e então, aplicar os métodos descritos no Capítulo 3.

2.- Estender o modelo (2.1)-(2.3) para as situações em que os erros de medição são correlacionados.

3.- Utilizar o enfoque estrutural supondo outra distribuição para a covariável x não observada.

4.- Provar as conjecturas descritas na Seção 3.3 sobre a distribuição assintótica do teste da razão de pseudoverossimilhanças para o caso em que o tamanho amostral n é fixo e o número de réplicas cresce.

5.- Aperfeiçoamento de testes como da razão de verossimilhanças e da razão de pseudoverossimilhanças, quando considera-se pequenas amostras.

Apêndice A

Neste Apêndice, apresentamos as inversões de algumas matrizes utilizadas neste trabalho já que permitiram a implementação computacional como são obtidas a seguir.

Inversa da matriz $\boldsymbol{T}$

Nesta seção apresentamos a inversa da matriz $\boldsymbol{T}$ dada em (2.20). O uso desta matriz facilita a obtenção da matriz de covariâncias assintótica de ambos os estimadores, MV e MPV, do modelo de regressão heteroscedástico. Usamos a seguinte propriedade da inversa de uma matriz em blocos (Magnus & Neudecker, 2007, Cap. 1). temos que

$$\begin{aligned}\begin{pmatrix} \boldsymbol{F}_{11} & \boldsymbol{F}_{12} \\ \cdot & \boldsymbol{F}_{22} \end{pmatrix}^{-1} &= \begin{pmatrix} \boldsymbol{F}^{11} & \boldsymbol{F}^{12} \\ \cdot & \boldsymbol{F}^{22} \end{pmatrix} \\ &= \begin{pmatrix} (\boldsymbol{F}_{11}-\boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{12}^{t})^{-1} & -(\boldsymbol{F}_{11}-\boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{12}^{t})^{-1}\boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1} \\ \cdot & \boldsymbol{F}_{22}^{-1}+\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{12}^{t}(\boldsymbol{F}_{11}-\boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{12}^{t})^{-1}\boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1} \end{pmatrix}. \end{aligned} \tag{6.1}$$

Utilizando a partição dada em (6.1) em $\boldsymbol{T}$ dada em (2.20). Tomamos $\boldsymbol{F}_{11} = t_{11}$, $\boldsymbol{F}_{12} = (\boldsymbol{t}_{12}, \boldsymbol{t}_{13})^t$, $\boldsymbol{F}_{21} = \boldsymbol{F}_{12}^t$ e

$$\boldsymbol{F}_{22} = \begin{pmatrix} \boldsymbol{T}_{22} & \boldsymbol{T}_{23} \\ \cdot & \boldsymbol{T}_{33} \end{pmatrix},$$

em que

$$\boldsymbol{T}_{22} = \text{Diag}\left(\frac{c_i r_i \sigma^2_{\varepsilon_i}}{\sigma^2_{\delta_i}} \left(a_i^2 - \frac{2a_i}{\sigma^2_{\delta_i}} + \frac{r_i}{\sigma^4_{\delta_i}} \right), i = 1, \ldots, n \right),$$

$$\boldsymbol{T}_{23} = \text{Diag}\left(\frac{c_i r_i s_i \beta_1^2}{\sigma_{\varepsilon_i}^2 \sigma_{\delta_i}^2}, i = 1, \ldots, n\right) \quad \text{e}$$

$$\boldsymbol{T}_{33} = \text{Diag}\left(\frac{c_i s_i \sigma_{\delta_i}^2}{\sigma_{\varepsilon_i}^2}\left(a_i^2 - \frac{2a_i\beta_1^2}{\sigma_{\varepsilon_i}^2} + \frac{s_i\beta_1^4}{\sigma_{\varepsilon_i}^4}\right), i = 1, \ldots, n\right).$$

Inicialmente calculamos a inversa do bloco $\boldsymbol{F}_{22}$ dada por

$$\boldsymbol{F}_{22}^{-1} = \begin{pmatrix} \boldsymbol{T}^{22} & \boldsymbol{T}^{23} \\ \cdot & \boldsymbol{T}^{33} \end{pmatrix},$$

em que

$$\boldsymbol{T}^{22} = \text{Diag}\left(\frac{\sigma_{\varepsilon_i}^4 \sigma_{\delta_i}^4 \omega_i}{\lambda_i r_i c_i g_i}, i = 1, \ldots, n\right), \quad \boldsymbol{T}^{23} = \text{Diag}\left(\frac{\beta_1^2 \sigma_{\varepsilon_i}^2 \sigma_{\delta_i}^2}{c_i g_i}, i = 1, \ldots, n\right)$$

$$\text{e} \quad \boldsymbol{T}^{33} = \text{Diag}\left(\frac{\lambda_i \sigma_{\delta_i}^4 \sigma_{\varepsilon_i}^4 z_i}{c_i s_i g_i}, i = 1, \ldots, n\right),$$

com

$$\lambda_i = \frac{\sigma_{\varepsilon_i^2}}{\sigma_{\delta_i}^2},\ z_i = a_i^2 - \frac{2a_i}{\sigma_{\delta_i}^2} + \frac{r_i}{\sigma_{\delta_i}^4}, \quad \omega_i = a_i^2 - \frac{2a_i\beta_1^2}{\sigma_{\varepsilon_i}^2} + \frac{s_i\beta_1^4}{\sigma_{\varepsilon_i}^4} \quad \text{e} \quad g_i = \sigma_{\delta_i}^4 \sigma_{\varepsilon_i}^4 z_i \omega_i - r_i s_i \beta_1^4.$$

De maneira análoga é encontrada para a matriz $\boldsymbol{T}^*$ dada em (2.24).

Inversa da matriz $\boldsymbol{A}_{\lambda\lambda}$

Apresentamos a inversa da matriz $\boldsymbol{A}_{\lambda\lambda}$ dada em (3.42), como segue:

$$\boldsymbol{A}_{\lambda\lambda}^{-1} = \begin{pmatrix} \boldsymbol{A}^{\lambda\lambda 1} & \boldsymbol{A}^{\lambda\lambda 2} \\ \boldsymbol{0}_{2n\times 2} & \boldsymbol{A}^{\lambda\lambda 3} \end{pmatrix},$$

sendo

$$\boldsymbol{A}^{\lambda\lambda 1} = \begin{pmatrix} \dfrac{1}{N} & 0 \\ -\dfrac{2\displaystyle\sum_{i=1}^{n} r_i(\overline{X}_{i.} - \mu_x)}{N^2 - S_2} & \dfrac{N}{N^2 - S_2} \end{pmatrix},$$

$$\boldsymbol{A}^{\lambda\lambda 2} = \begin{pmatrix} \boldsymbol{0}_{1\times n} & \boldsymbol{0}_{1\times n} \\ \left(-\dfrac{N - r_i}{(r_i - 1)(N^2 - S_2)}, i = \ldots, n\right)^t & \boldsymbol{0}_{1\times n} \end{pmatrix} \quad \text{e}$$

$$\boldsymbol{A}^{\lambda\lambda 3} = \text{Diag}\left(\frac{1}{r_1 - 1}, \ldots, \frac{1}{r_n - 1}, \frac{1}{s_1 - 1}, \ldots, \frac{1}{s_n - 1}\right).$$

Referências bibliográficas

Artes, R. & Botter, D. A. (2005). *Funções de estimação em modelos de regressão*. Associação Brasileira de Estatística, São Paulo - SP.

Atkinson, A. C. (1985). *Plots, transformations and regression*. Oxford, Claredon.

Barnett, V. D. (1970). Fitting straight lines - the linear functional relationship with replicated observations. *Journal of the Royal Statistical Society C*, **19**, 135–144.

Bera, A. K. & Bilias, Y. (2001). R's score, Neyman's C(α) and Silvey's lm test: an essay on historical developments and some new results. *Journal of Statistical Planning and inference*, **97**, 9–44.

Berger, O. M., Liseo, B. & Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, **4**, 1–28.

Buonaccorsi, J. P. (2010). *Measurement error models methods and applications*. Chapman & Hall/CRC Press, Boca Raton.

Carroll, R. J., Gail, M. H. & Lubin, J. H. (1993). Case-control studies with errors in covariates. *Journal of the American Statistical Association*, **88**, 185–199.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Chapman & Hall/CRC Press, Boca Raton, second edition.

Carstensen, B. (2010). *Comparing clinical measurement methods: A practical guide*. Wiley, Chichester.

Carstensen, B., Gurrin, L. & Ekstrom, C. (2012). *MethComp: Functions for analysis of method comparison studies*. R package version 1.15.

Chan, L. K. & Mak, T. K. (1979). Maximum likelihood estimation of a linear structural relationship with replication. *Journal of the Royal Statistical Society B*, **41**, 263–268.

Cheng, C. L. & Van Ness, J. W. (1999). *Statistical regression with measurement error*. Arnold, London.

Cordeiro, G. M. (1999). *Introdução à teoria assintótica*. IMPA, Rio de Janeiro - RJ.

de Castro, M., Bolfarine, H. & Galea, M. (2013). Bayesian inference in measurement error models for replicated data. *Environmetrics*, **24**, 22–30.

Delgado, J. J. (1995). *Estimação por pseudo Máxima verossimilhança*. Dissertação, Universidade de São Paulo - SP.

Demspter, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1–38.

Dolby, G. R. (1976). The ultrastructural relation: A synthesis of the functional and structural relations. *Biometrika*, **63**, 39–50.

Dolby, G. R., Cormack, R. M. & Sinclair, D. F. (1987). On fitting bivariate functional relationships to unpaired and unequally replicated data. *Biometrika*, **74**, 393–399.

Efron, B.; Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall, London, second edition.

Fahrmeir, L. (1988). A note on asymptotic testing theory for nonhomogeneous observations. *Stochastic Processes and their Applications*, **28**, 267–273.

Fuller, W. A. (1987). *Measurement error models*. Wiley, New York.

Galea, M., de Castro, M. & Bolfarine, H. (2008). Hypothesis testing in an errors-in-variables model with heteroscedastic measurement errors. *Statistical Medicine*, **27**, 5217–5234.

Galea-Rojas, M., de Castilho, M. V., Bolfarine, H. & de Castro, M. (2003). Detection of analytical bias. *The Analyst*, **128**, 1073–1081.

Geys, H., Molenberghs, G. & Louise, M. R. (2006). *Pseudo-likelihood estimation*. Chapman & Hall/CRC Press, Boca Raton.

Gong, G. & Samaniego, F. J. (1981). Pseudo maximun likelihood estimation: Theory and applications. *The Annals of Statistics*, **9**, 861–869.

Graybill, F. A. (1976). *Theory and application of the linear model*. Duxbury Press, North Scituate.

Guolo, A. (2011). Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica*, **21**, 1639–1663.

Johnson, N. L., Kotz, S. & Balakrishman, N. (1994). *Continuous univariate distributions*. Oxford, Claredon, second edition.

Kulathinal, S. B., Kuulasmaa, K. & Gasbarra, D. (2002). Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between observations. *Statistics in Medicine*, **21**, 1089–1101.

Lehmann, E. L. & Casella, G. (1998). *Theory of point estimation*. Springer-Verlag, New York, third edition.

Magnus, J. R. & Neudecker, H. (2007). *Matrix differential calculus with applications in statistics and econometrics*. Wiley, Chichester, third edition.

Parker, W. R. (1986). Pseudo maximun likelihood estimation: The asymptotic distribution. *The Annals of Statistics*, **14**, 355–357.

Patriota, A. C., Bolfarine, H. & de Castro, M. (2009). Heteroscedastic structural errors-in-variables model with equation error. *Statistical Methodology*, **6**, 408–423.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasekh, A. R. & Fieller, N. (2003). Influence functions in functional measurement error model with replicate data. *Statistics*, **37**, 169–178.

Riu, J. & Rius, F. X. (1996). Assessing the accuracy of analytical methods using linear regression with errors in both axes. *Analytical Chemistry*, **68**, 1851–1857.

Searle, S. R. (1971). *Linear models*. Wiley, New York.

Self, S. G. & Liang, K.-Y. (1996). On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society B*, **58**, 785–796.

Sengupta, B. (2012). *A robust linear mixed effects model with application to method comparison studies*. Ph.D. thesis, The University of Texas, Dallas.

Stefanski, L. A. (2000). Measurement error models. *Journal of the American Statistical Association*, **95**, 1353–1358.

Tanner, M. A. (1996). *Tools for statistical inference*. Springer, New York, third edition.

Terrell, G. R. (2002). The gradient statistic. *Computing Science and Statistics*, **34**, 206–215.