Inferência em um modelo com erros de medição heteroscedásticos com observações replicadas

Willian Luís de Oliveira

Inferência em um modelo com erros de medição heteroscedásticos com observações replicadas

Willian Luís de Oliveira

Orientador: **Prof. Dr. Mário de Castro Andrade Filho**

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

USP – São Carlos
Agosto de 2011

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

O48i

Oliveira, Willian Luís
Inferência em um modelo com erros de medição heteroscedásticos com observações replicadas / Willian Luís Oliveira; orientador Mário de Castro Andrade Filho -- São Carlos, 2011.
73 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2011.

1. Modelo com erros de medição. 2. Modelo funcional. 3. Validação de métodos de medição. 4. Método SIMEX. 5. Dados desemparelhados e desbalanceados. I. Andrade Filho, Mário de Castro, orient. II. Título.

Agradecimentos

Agradeço a Deus por me dar saúde e sabedoria.

Aos meus pais Idemar e Cleusa que sempre me apoiaram e incentivaram.

À minha tia Maria pelo carinho.

Ao meu irmão Wander, minha sobrinha Tainara e cunhada Marli pelos momentos de distração.

Ao meu orientador Prof. Mário de Castro pela confiança, paciência e ensinamentos.

Aos professores Reiko Aoki e Carlos Alberto Ribeiro Diniz, membros da banca examinadora, pelas contribuições dadas.

Aos meus padrinhos Crippa e Virgílio pela motivação.

Aos meus amigos de graduação e pós-graduação pelos momentos de estudo e diversão.

Aos professores do ICMC-USP e DEs-UFSCar pela atenção dada.

Aos funcionários do ICMC-USP.

Ao CNPq pelo apoio financeiro.

Resumo

Modelos com erros de medição têm recebido a atenção de vários pesquisadores das mais diversas áreas de conhecimento. O principal objetivo desta dissertação consiste no estudo de um modelo funcional com erros de medição heteroscedásticos na presença de réplicas das observações. O modelo proposto estende resultados encontrados na literatura na medida em que as réplicas são parte do modelo, ao contrário de serem utilizadas para estimação das variâncias, doravante tratadas como conhecidas. Alguns procedimentos de estimação tais como o método de máxima verossimilhança, o método dos momentos e o método de extrapolação da simulação (SIMEX) na versão empírica são apresentados. Além disso, propõe-se o teste da razão de verossimilhanças e o teste de Wald com o objetivo de testar algumas hipóteses de interesse relacionadas aos parâmetros do modelo adotado. O comportamento dos estimadores de alguns parâmetros e das estatísticas propostas (resultados assintóticos) são analisados por meio de um estudo de simulação de Monte Carlo, utilizando-se diferentes números de réplicas. Por fim, a proposta é exemplificada com um conjunto de dados reais. Toda parte computacional foi desenvolvida em linguagem R (R Development Core Team, 2011).

Abstract

Measurement error models have received the attention of many researchers of several areas of knowledge. The aim of this dissertation is to study a functional heteroscedastic measurement errors model with replicated observations. The proposed model extends results from the literature in that replicas are part of the model, as opposed to being used for estimation of the variances, now treated as known. Some estimation procedures such as maximum likelihood method, the method of moments and the empirical simulation-extrapolation method (SIMEX) are presented. Moreover, it is proposed the likelihood ratio test and Wald test in order to test hypotheses of interest related to the model parameters used. The behavior of the estimators of some parameters and statistics proposed (asymptotic results) are analyzed through Monte Carlo simulation study using different numbers of replicas. Finally, the proposal is illustrated with a real data set. The computational part was developed in R language (R Development Core Team, 2011).

Sumário

Capítulo 1

Introdução

1.1 Motivação

Em muitos estudos nas mais diversas áreas, como por exemplo, médica, agrícola, econômica, química, física e biológica, modelos de regressão são utilizados com o objetivo de modelar (expressar) a relação existente entre duas ou mais variáveis de interesse. Nesses modelos geralmente considera-se que apenas a variável resposta é medida com erro. Entretanto, na prática nem sempre isso ocorre; muitas vezes as covariáveis envolvidas também não podem ser medidas com exatidão (Cheng & Van Ness, 1999). Esta situação é mais comum do que se imagina. Suponha que estamos relacionando, por meio de um modelo, as variáveis dosagem de uma determinada droga e nível de proteina na urina (Barnett, 1970), conteúdo de DNA por célula e quantidade de célula em um determinado tecido (Dolby *et al.*, 1987), rendimento na produção de um determinado cereal e teor de nitrogênio no solo (Fuller, 1987) ou o desempenho acadêmico ou atlético de crianças de diferentes grupos de idades em duas localizações distintas (Dolby *et al.*, 1987). Nestes casos, todas as variáveis envolvidas não podem ser medidas com exatidão. Em problemas envolvendo validação de métodos de medição, cujo ob-

jetivo é verificar a equivalência entre métodos de medição (Ripley & Thompson, 1987; Riu & Rius, 1996; Galea-Rojas *et al.*, 2003), os valores obtidos também são medidos com erros. Como aplicação à metodologia estudada nesta dissertação, utilizaremos um conjunto de dados relacionado ao problema de comparação de métodos de medição.

Modelos com erros de medição (também chamados de modelos com erros nas variáveis) estendem os modelos de regressão procurando representar as variáveis explicativas de uma forma mais realista (para muitas aplicações práticas). Valores observados destas variáveis são tratados como substitutos sujeitos a erros dos verdadeiros valores (não observáveis). A literatura sobre o assunto é extensa. Uma cobertura bem ampla de diversos tópicos referentes a estes modelos pode ser encontrada em alguns livros (por exemplo, Fuller, 1987; Cheng & Van Ness, 1999; Carroll *et al.*, 2006). Recentemente, modelos com erros de medição funcionais heteroscedásticos têm sido objeto de pesquisas (e. g., Galea-Rojas *et al.*, 2003; de Castro *et al.*, 2004; Kukush & Van Huffel, 2004; Markovsky *et al.*, 2006; de Castro & Galea, 2010; Wang *et al.*, 2010). Em diversas aplicações supõe-se que os erros de medição são descorrelacionados e suas variâncias são conhecidas e maiores do que 0, que é um cenário comumente encontrado em exemplos de áreas tais como Química Analítica (Ripley & Thompson, 1987; Riu & Rius, 1996; Galea-Rojas *et al.*, 2003), Epidemiologia (Kulathinal *et al.*, 2002), Engenharia Civil (Bertrand-Krajewski, 2004) e Botânica (Veenendaal *et al.*, 2008).

Neste trabalho os principais objetivos são: (1) Revisar os métodos inferenciais que podem ser empregados nos modelos heteroscedásticos com réplicas (Barnett, 1970; Dolby & Lipton, 1972; Dolby *et al.*, 1987; Devanarayan & Stefanski, 2002; Carroll *et al.*, 2006; Patriota, 2006), (2) Obter os estimadores de máxima verossimilhança, dos momentos e SIMEX dos parâmetros de interesse do modelo funcional heteroscedástico com dados replicados e compará-los, sendo que no modelo em questão, as réplicas podem ser desemparelhadas ou emparelhadas e desbalanceadas ou balanceadas, (3) Estudar estatísticas para testar a homoscedasticidade das variâncias dos erros de medição,

a proporcionalidade das variâncias dos erros de medição e testes envolvendo simultaneamente os coeficientes linear e angular, (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 de x e y aumenta com o aumento do número de observações, (5) Aplicar os métodos de estimação e testes de hipóteses desenvolvidos a um conjunto de dados reais.

1.2 Conceitos básicos

Nesta seção apresentamos alguns conceitos que serão muito utilizados neste trabalho. A maioria destes conceitos pode ser vista detalhadamente em Fuller (1987) e Cheng & Van Ness (1999).

Uma generalização dos modelos de regressão tradicionais, os modelos com erros de medição são caracterizados pelo fato de que há covariáveis do modelo medidas com erro. Considerando o modelo linear mais simples com erros de medição, inferências serão efetuadas sobre os parâmetros de uma reta ajustada entre as duas variáveis, levando em conta que tanto a variável resposta quanto a covariável são medidas com erro. O modelo de regressão simples é dado por

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

em que a variável independente x pode ser fixa ou aleatória e o erro ϵ tem média 0 e é descorrelacionado com x. Geralmente, os parâmetros desconhecidos β_0 e β_1 são estimados pelo método de mínimos quadrados ou através de algum outro procedimento robusto, dado um conjunto de observações independentes (Cheng & Van Ness, 1999).

Sejam n o tamanho amostral, y_i o verdadeiro valor (não observável) da variável resposta, x_i o verdadeiro valor (não observável) da variável explicativa, Y_i o valor observado da variável resposta e X_i o valor observado da variável explicativa na unidade

i, $i = 1, ..., n$. Relacionando estas variáveis, iniciamos com o modelo linear simples com erro de medição

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

$$Y_i = y_i + e_i \tag{1.2}$$

$$\text{e} \quad X_i = x_i + u_i, \tag{1.3}$$

em que u_i denota erro de medição (aditivo) e e_i denota vários erros incluindo o erro de medição (aditivo). Assumimos que x_i e os erros u_i e e_i são descorrelacionados sendo que $u_1, \ldots, u_n, e_1, \ldots, e_n$ têm médias zero e variâncias finitas.

Em algumas aplicações supomos que os erros (e_i, u_i) têm distribuição

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{e_i} & 0 \\ 0 & \sigma^2_{u_i} \end{bmatrix} \right), \quad i = 1, \ldots, n, \tag{1.4}$$

com as variâncias $\sigma^2_{e_i}$ e $\sigma^2_{u_i}$ conhecidas e maiores do que 0, o que ocorre com frequência em problemas da química analítica (Ripley & Thompson, 1987; Riu & Rius, 1996; Galea-Rojas *et al.*, 2003) e epidemiologia (Kulathinal *et al.*, 2002), entre outros em que a variabilidade dos dados geralmente está disponível de estudos passados, por exemplo.

Em muitas situações y e x são chamadas de variáveis latentes. O modelo $(1.1)-(1.4)$ também é dito modelo heteroscedástico, uma vez que as variâncias dos erros variam de observação para observação em (1.4). Um caso particular desse modelo ocorre quando $\sigma^2_{u_i} = \sigma^2_u$ e $\sigma^2_{e_i} = \sigma^2_e$, $i = 1, ..., n$, que corresponde ao modelo homoscedástico.

Na literatura encontramos três modelos principais, da forma $(1.1) - (1.4)$, dependendo das suposições formuladas sobre a variável explicativa x.

- Modelo funcional aditivo

  Neste caso, x_i's são constantes desconhecidas, $i = 1, ..., n$.

- Modelo estrutural aditivo

  Destacamos o modelo estrutural normal aditivo. Neste caso, $x_i \overset{\text{iid}}{\sim} N(\mu_x, \sigma_x^2)$ com x_i e u_i independentes entre si e independentes de e_i; $u_i \overset{\text{indep.}}{\sim} N(0, \sigma_{u_i}^2)$ e $e_i \overset{\text{indep.}}{\sim} N(0, \sigma_{e_i}^2)$ para $i = 1, ..., n$.

- Modelo ultraestrutural aditivo

  Neste caso, os x_i's são variáveis aleatórias independentes como em um modelo estrutural, mas não são identicamente distribuídas, podendo ter diferentes médias (μ_i) e variâncias iguais (σ_x^2). O modelo ultraestrutural é uma generalização dos modelos funcional e estrutural. Se $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, então o modelo ultraestrutural reduz-se ao modelo estrutural. Se $\sigma_x^2 = 0$, o modelo ultraestrutural reduz-se ao modelo funcional.

Nos casos dos modelos funcional e ultraestrutural, os x_i's ou μ_i's são parâmetros incidentais e seu número aumenta com o tamanho da amostra.

O modelo de regressão linear simples usual é um caso especial do modelo com erros de medição quando os erros, u's, são identicamente nulos. Se tentarmos escrever $(1.1) - (1.4)$ como um modelo de regressão usual, obtemos

$$\begin{aligned} Y &= y + e = \beta_0 + \beta_1 x + e = \beta_0 + \beta_1(X - u) + e \\ &= \beta_0 + \beta_1 X + (e - \beta_1 u) = \beta_0 + \beta_1 X + \zeta, \end{aligned}$$

em que $\zeta = e - \beta_1 u$. Entretanto, este não é um modelo de regressão usual, pois X é variável aleatória e é correlacionada com o termo do erro ζ. Se tentarmos usar estimadores de mínimos quadrados obtemos estimadores inconsistentes (Fuller, 1987).

Em algumas situações, réplicas das variáveis resposta e explicativa estão disponíveis podendo ser inseridas na estrutura do modelo. As réplicas são utilizadas na estimação das variâncias dos erros de medição, quando estas variâncias são desconhecidas. Nesta dissertação, trabalharemos com o modelo funcional heteroscedástico com observações

replicadas. Esse modelo será apresentado no Capítulo 2. Antes, fizemos uma revisão bibliográfica a respeito de modelos com erros de medição com e sem réplicas.

1.3 Revisão bibliográfica

Existem muitos trabalhos publicados considerando modelos com erro de medição, tanto para o caso linear quanto para o não linear. Alguns destes trabalhos utilizaram como motivação problemas das diversas áreas acima descritas.

Fuller (1987), Cheng & Van Ness (1999) e Carroll *et al.* (2006) tratam de modelos de regressão com erros de medição, apresentando conceitos básicos e investigando os efeitos desses erros sobre os estimadores de mínimos quadrados ordinários dos parâmetros. Os estimadores de máxima verossimilhança e os estimadores pelo método dos momentos também foram apresentados. Para o modelo de regressão linear simples, os autores concluíram que o estimador de mínimos quadrados para o parâmetros β_1 não é consistente. Em Carroll *et al.* (2006), as versões paramétrica e empírica do método SIMEX, que será apresentado na Subsecão 2.1.3, são detalhadas.

Barnett (1970) ajustou um modelo funcional linear heteroscedástico com observações replicadas e variâncias dos erros desconhecidas, utilizando o método de máxima verossimilhança. Aproximações das estimativas dos parâmetros são calculadas por métodos iterativos. Além disso, o autor encontra a matriz de covariâncias assintótica dos estimadores. Neste artigo, os erros de medição são considerados mutuamente descorrelacionados e normalmente distribuidos com média 0 e variâncias $\sigma^2_{u_i}$ e $\sigma^2_{e_i}$, respectivamente, sendo que a heteroscedasticidade proporcional é adotada. O número de réplicas da covariável e da variável resposta, para a i-ésima observação, $i = 1, \ldots, n$, foi considerado o mesmo. Como motivação, foi utilizado um exemplo da área médica.

Dolby & Lipton (1972) estimam os parâmetros de um modelo funcional heteroscedástico não linear geral com observações replicadas independentes. Além disso, os autores consideram que os erros das observações seguem uma distribuição normal bivariada e que são correlacionados. Novamente, um método iterativo para a obtenção de aproximações das estimativas de máxima verossimilhança dos parâmetros do modelo é proposto. Expressões em forma fechada são obtidos para a matriz de covariâncias assintótica dos estimadores dos parâmetros β_0, β_1 e dos parâmetros $\sigma^2_{u_i}$ e $\sigma^2_{e_i}$. Por fim, dados simulados para ilustrar o trabalho e uma discussão sobre a convergência estão incluídos.

Chan & Mak (1979) consideram um modelo estrutural linear homoscedástico com réplicas tanto da variável resposta quanto da variável explicativa. Os autores derivam os estimadores de máxima verossimilhança dos parâmetros do modelo adotado, obtendo sua respectiva matriz de covariâncias assintótica.

Dolby *et al.* (1987), ao contrário de Barnett (1970) que considerou dados pareados, estudaram a relação funcional linear considerando réplicas desemparelhadas e desbalanceadas. Os autores discutiram cinco modelos que se distinguem principalmente pelas suposições impostas sobre as variâncias dos erros: se são homogêneas ou heterogêneas entre os grupos i, se as variâncias $\sigma^2_{e_i}$ e $\sigma^2_{u_i}$ são proporcionais para uma mesma observação i e, se sim, se a constante de proporcionalidade pode ser identificada como sendo o quadrado do coeficiente angular. Por fim, os autores encontraram os estimadores para os parâmetros.

Kimura (1992) utilizou um modelo funcional heteroscedástico no problema da comparação das medidas considerando dois ou mais métodos distintos de medida (calibração comparativa). O autor propôs um algoritmo do tipo EM para obter aproximações das estimativas de máxima verossimilhança dos parâmetros do modelo. Além disso, propõe testes para a hipótese de os métodos de calibração serem considerados equiva-

lentes. Neste artigo, duas estruturas para a variância do erro são abordadas, sendo que na primeira $\sigma^2_{e_{ij}} = \tau_i \sigma^2_e$ e na segunda $\sigma^2_{e_{ij}} = \tau_i \sigma^2_e x_j$. Por fim, os métodos propostos são validados por simulação.

Cook & Stefanski (1994) propuseram o método SIMEX para modelos paramétricos com erros de medição considerando que as variâncias dos erros de medição da covariável medida com erro são homoscedásticas e conhecidas ou bem estimadas. Os autores realizaram um estudos de simulação e por fim aplicaram a metodologia estudada a um conjunto de dados reais.

Devanarayan & Stefanski (2002) expõem uma breve introdução à versão paramétrica do método SIMEX e em seguida, apresentaram uma variação do algoritmo SIMEX apropriado para as situações em que as variâncias dos erros de medição são desconhecidos e réplicas da variável explicativa estão disponíveis. Os pseudoerros utilizados são gerados na forma de contrastes lineares aleatórios das réplicas. Como principal característica do método proposto, temos a sua capacidade para acomodar erros de medição heteroscedásticos. Por fim, os autores aplicam o método a um conjunto de dados reais.

Galea-Rojas *et al.* (2003) utilizaram um algoritmo do tipo EM para calcular aproximações das estimativas de máxima verossimilhança dos parâmetros do modelo funcional heteroscedástico com variâncias conhecidas, utilizado na detecção de viés analítico. Além disso, mostraram que os valores do viés simulado e do erro quadrático médio simulado das estimativas de máxima verossimilhança dos parâmetros de interesse são menores do que os valores obtidos por estimativas de outros métodos. Resultados de simulações e aplicações a conjuntos de dados reais foram apresentados.

Rasekh & Fieller (2003) consideram a construção e as propriedades das funções de influência em um modelo com erros de medição funcional com dados replicados. Os autores mostram que nesse modelo as estimativas dos parâmetros podem ser afetadas tanto pelas observações individuais quanto pelas médias das observações replicadas.

Além disso, mostram também que a função de influência das médias das réplicas sobre as estimativas dos coeficientes de regressão pode ser derivada somente sob a suposição de que as variâncias dos erros são conhecidas, enquanto que para as observações individuais só pode ser obtida simultaneamente com sua função de influência sobre os estimadores das variâncias dos erros.

Patriota (2006) incorporou as réplicas ao processo de estimação dos parâmetros dos modelos funcionais heteroscedásticos com erros nas variáveis, considerando tanto a situação em que as réplicas são correlacionadas quanto a situação em que elas são descorrelacionadas e supondo também que as variâncias são desconhecidas. Entretanto, o modelo proposto se restringe às situações em que as réplicas são balanceadas. O autor também aplicou um algoritmo iterativo do tipo EM com o objetivo de obter aproximações das estimativas de máxima verossimilhança dos parâmetros de interesse. Além disso, foi proposta uma estatística de Wald para testar algumas hipóteses e foram realizadas simulações para verificar o comportamento desta estatística em determinadas situações. Por fim, o autor aplicou a técnica proposta na validação de métodos de medição, utilizando um conjunto de dados reais.

1.4 Apresentação dos capítulos

No Capítulo 1 exemplificamos algumas situações reais em que o valor da covariável não pode ser medido com exatidão. Apresentamos também conceitos básicos mas fundamentais para o desenvolvimento desta dissertação. Por fim, realizamos uma revisão de alguns trabalhos já publicados relacionados com o assunto tratado.

No Capítulo 2 apresentamos o modelo com o qual trabalharemos (modelo funcional heteroscedástico com observações replicadas) e derivamos os estimadores dos parâmetros de interesse do modelo proposto por meio do método de máxima verossimilhança

(MV), método dos momentos (MM) e método de extrapolação da simulação (SIMEX). Além disso, apresentamos as respectivas matrizes de covariâncias assintóticas dos estimadores de interesse, para os métodos de máxima verossimilhança e SIMEX. Apresentamos também a técnica *bootstrap*, utilizada para encontrar a matriz de covariâncias dos estimadores dos parâmetros de interesse no método dos momentos. Por fim, apresentamos as estatísticas da razão de verossimilhanças e de Wald para testar algumas hipóteses de interesse.

Já no Capítulo 3 realizamos um estudo de simulação utilizando a linguagem R (R Development Core Team, 2011) com o objetivo de analisar o comportamente dos estimadores obtidos pelo método de máxima verossimilhança, pelo método dos momentos e pelo método SIMEX. A situação em que o número de observações n permanece fixo e o número de réplicas cresce é analisada. O comportamento da estatística da razão de verossimilhanças e da estatística de Wald é estudado.

No Capítulo 4 aplicamos a metodologia apresentada no Capítulo 2 a um conjunto de dados reais que trata do teor de potássio (K) em cerâmicas egípcias.

Por fim, no Capítulo 5 apresentamos as conclusões e as propostas de trabalhos futuros.

Capítulo 2

Modelo e Estimação dos Parâmetros

Neste capítulo inserimos as réplicas na estrutura do modelo, generalizando o modelo (1.1) − (1.4) dado na Seção 1.2. Tal modelo é dito modelo funcional heteroscedástico com observações replicadas e foi estudado por Dolby *et al.* (1987) e por Patriota (2006), que considerou um caso particular em que as réplicas são balanceadas. Apresentamos os estimadores de máxima verossimilhança, os estimadores do método dos momentos e os estimadores SIMEX dos parâmetros de interesse e as respectivas matrizes de covariâncias. Por fim, propomos testes para testar homoscedasticidade, proporcionalidade das variâncias e parâmetros tais como β_0 e β_1.

2.1 Modelo funcional heteroscedástico com observações replicadas

Suponha que dispomos de $r_i \geq 2$ e $s_i \geq 2$ réplicas das variáveis resposta e explicativa, respectivamente. Neste caso, as expressões (1.1) − (1.4) passam a ser

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

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

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

com e_{ij} independente de u_{ik}, $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_{e_i})$ para $j = 1, ..., r_i$, $u_{ik} \overset{\text{iid}}{\sim} N(0, \sigma^2_{u_i})$ para $k = 1, ..., s_i$ e $i = 1, ..., n$. As réplicas são consideradas desemparelhadas e desbalanceadas, podendo também serem pareadas (emparelhadas) e balanceadas. Assim como no modelo $(1.1) - (1.4)$, $(\beta_0, \beta_1)^\top$ são parâmetros a estimar, x_i's são constantes desconhecidas, u_{ik} são erros de medição (aditivos) e e_{ij} incorporam vários erros incluindo os erros de medição (aditivos). O modelo $(2.1) - (2.3)$ é dito modelo funcional heteroscedástico com observações replicadas.

Estimação pelo método de máxima verossimilhança neste modelo foi estudada por Dolby *et al.* (1987) e por Patriota (2006). Já os estimadores dos parâmetros do modelo (2.1)-(2.3), utilizando os métodos dos momentos e SIMEX, não foram encontrados na literatura. Desta forma, apresentamos a seguir os estimadores de máxima verossimilhança, do método dos momentos e do método SIMEX dos parâmetros de interesse do modelo (2.1)-(2.3) bem como as respectivas matrizes de covariâncias.

2.1.1 Máxima verossimilhança (MV)

Os resultados aqui apresentados podem ser encontrados em Dolby *et al.* (1987), que mostram várias situações com respeito às variâncias $\sigma^2_{e_i}$ e $\sigma^2_{u_i}$, dentre as quais podemos destacar o caso em que há heteroscedasticidade irrestrita $(\sigma^2_{e_i}, \sigma^2_{u_i})$, proporcional $(\sigma^2_{e_i}, \rho\sigma^2_{e_i})$ em que ρ é uma constante conhecida e o caso em que há homoscedasticidade (σ^2_e, σ^2_u).

Consideremos o modelo $(2.1) - (2.3)$. Neste caso, $Y_{ij} \overset{\text{iid}}{\sim} N(y_i, \sigma^2_{e_i})$ e $X_{ik} \overset{\text{iid}}{\sim} N(x_i, \sigma^2_{u_i})$ com Y_{ij} e X_{ik} independentes para $j = 1, ..., r_i$, $k = 1, ..., s_i$ e $i = 1, ..., n$. Desta

forma, definindo $\boldsymbol{\theta_i} = (\beta_0, \beta_1, \sigma^2_{u_i}, \sigma^2_{e_i}, x_i)^\top$ e $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\sigma}_u^{2^\top}, \boldsymbol{\sigma}_e^{2^\top}, \boldsymbol{x}^\top)^\top$, em que $\boldsymbol{x} = (x_1, ..., x_n)^\top$, $\boldsymbol{\sigma}^2_u = (\sigma^2_{u_1}, ..., \sigma^2_{u_n})^\top$ e $\boldsymbol{\sigma}^2_e = (\sigma^2_{e_1}, ..., \sigma^2_{e_n})^\top$, temos que a função verossimilhança para a i-ésima observação é dada por

$$\begin{aligned} L_i(\boldsymbol{\theta_i}) \;\; \propto \;\; & \frac{1}{(\sigma^2_{e_i})^{r_i/2}} \exp\left\{ -\sum_{j=1}^{r_i} \frac{1}{2\sigma^2_{e_i}} [Y_{ij} - (\beta_0 + \beta_1 x_i)]^2 \right\} \frac{1}{(\sigma^2_{u_i})^{s_i/2}} \\ & \times \exp\left\{ -\sum_{k=1}^{s_i} \frac{1}{2\sigma^2_{u_i}} (X_{ik} - x_i)^2 \right\}. \end{aligned}$$

A função verossimilhança para uma amostra de tamanho n é dada por

$$\begin{aligned} L(\boldsymbol{\theta}) \;\; = \;\; & \prod_{i=1}^{n} L_i(\boldsymbol{\theta_i}) \\ \propto \;\; & \prod_{i=1}^{n} \left\{ \frac{1}{(\sigma^2_{e_i})^{r_i/2}} \frac{1}{(\sigma^2_{u_i})^{s_i/2}} \right\} \\ & \times \exp\left\{ -\sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{1}{2\sigma^2_{e_i}} [Y_{ij} - (\beta_0 + \beta_1 x_i)]^2 - \sum_{i=1}^{n} \sum_{k=1}^{s_i} \frac{1}{2\sigma^2_{u_i}} (X_{ik} - x_i)^2 \right\}. \end{aligned}$$

Logo, a função log-verossimilhança é dada por

$$\begin{aligned} \ell(\boldsymbol{\theta}) \;\; = \;\; & const - \frac{1}{2} \sum_{i=1}^{n} r_i \log \sigma^2_{e_i} - \frac{1}{2} \sum_{i=1}^{n} s_i \log \sigma^2_{u_i} \\ & - \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma^2_{e_i}} \sum_{j=1}^{r_i} [Y_{ij} - (\beta_0 + \beta_1 x_i)]^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma^2_{u_i}} \sum_{k=1}^{s_i} (X_{ik} - x_i)^2. \end{aligned}$$

As equações de máxima verossimilhança são facilmente obtidas e formam um conjunto de $3n+2$ equações correspondentes aos parâmetros $(\beta_0, \beta_1, x_i, \sigma^2_{u_i}, \sigma^2_{e_i})$, $i = 1, ..., n$. Todas as derivadas de primeira e segunda ordem utilizadas nesta dissertação podem ser encontradas no Apêndice. Diferenciando a função log-verossimilhança com respeito aos parâmetros $\sigma^2_{u_i}, \sigma^2_{e_i}$, x_i, β_0 e β_1, e igualando a 0, segue que os estimadores de máxima verossimilhança satisfazem as equações

$$\widehat{\sigma}^2_{u_i} = \sum_{k=1}^{s_i} \frac{(X_{ik} - \widehat{x}_i)^2}{s_i}, \;\; \widehat{\sigma}^2_{e_i} = \sum_{j=1}^{r_i} \frac{[Y_{ij} - (\widehat{\beta}_0 + \widehat{\beta}_1 \widehat{x}_i)]^2}{r_i}, \tag{2.4}$$

$$\widehat{x_i} = \frac{\widehat{\sigma}^2_{u_i}\widehat{\beta}_1 r_i(\overline{Y}_i - \widehat{\beta}_0) + \widehat{\sigma}^2_{e_i} s_i \overline{X}_i}{\widehat{\beta}_1^2 \widehat{\sigma}^2_{u_i} r_i + \widehat{\sigma}^2_{e_i} s_i}, \tag{2.5}$$

$$\widehat{\beta}_0 = \widehat{\overline{Y}}_\omega - \widehat{\beta}_1 \widehat{\overline{x}_\omega} \quad \text{e} \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \widehat{\omega}_i \widehat{x}_i (\overline{Y}_i - \widehat{\overline{Y}}_\omega)}{\sum_{i=1}^{n} \widehat{\omega}_i \widehat{x}_i (\widehat{x}_i - \widehat{\overline{x}}_\omega)}, \tag{2.6}$$

em que $\widehat{\overline{x}}_\omega$, $\widehat{\overline{Y}}_\omega$ são médias gerais ponderadas estimadas dadas por

$$\widehat{\overline{x}}_\omega = \frac{\sum_{i=1}^{n} \widehat{\omega}_i \widehat{x}_i}{\sum_{i=1}^{n} \widehat{\omega}_i} \quad \text{e} \quad \widehat{\overline{Y}}_\omega = \frac{\sum_{i=1}^{n} \widehat{\omega}_i \overline{Y}_i}{\sum_{i=1}^{n} \widehat{\omega}_i}, \tag{2.7}$$

$\widehat{\omega}_i$ são pesos estimados dados por

$$\widehat{\omega}_i = \frac{r_i}{\widehat{\sigma}^2_{e_i}} \tag{2.8}$$

e $\overline{X}_i$ e $\overline{Y}_i$ são médias das réplicas em cada grupo i, $i = 1, \ldots, n$ dadas por

$$\overline{X}_i = \frac{1}{s_i} \sum_{k=1}^{s_i} X_{ik} \quad \text{e} \quad \overline{Y}_i = \frac{1}{r_i} \sum_{j=1}^{r_i} Y_{ij}. \tag{2.9}$$

Patriota (2006) também encontra os estimadores de máxima verossimilhança dos parâmetros do modelo (2.1) − (2.3). Entretanto, o autor considera o caso particular em que $r_i = s_i = m$ para $i = 1, \ldots, n$. Neste caso, a função log-verossimilhança é dada por

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= const - \frac{m}{2} \sum_{i=1}^{n} (\log \sigma^2_{e_i} + \log \sigma^2_{u_i}) - \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma^2_{e_i}} \sum_{j=1}^{m} [Y_{ij} - (\beta_0 + \beta_1 x_i)]^2 \\ &\quad - \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma^2_{u_i}} \sum_{k=1}^{m} (X_{ik} - x_i)^2. \end{aligned}$$

A equação (2.5) se reduz a

$$\widehat{x_i} = \frac{\widehat{\sigma}^2_{u_i}\widehat{\beta}_1(\overline{Y}_i - \widehat{\beta}_0) + \widehat{\sigma}^2_{e_i}\overline{X}_i}{\widehat{\beta}^2_1\widehat{\sigma}^2_{u_i} + \widehat{\sigma}^2_{e_i}} \tag{2.10}$$

e as equações (2.4) e (2.6) se mantêm com $\widehat{\omega}_i = 1/\widehat{\sigma}^2_{e_i}$.

Quando $s_i = s$ e $r_i = r$ para $i = 1, \ldots, n$, a função log-verossimilhança se reduz a

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= const - \frac{r}{2}\sum_{i=1}^{n}\log\sigma^2_{e_i} - \frac{s}{2}\sum_{i=1}^{n}\log\sigma^2_{u_i} - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma^2_{e_i}}\sum_{j=1}^{r}[Y_{ij} - (\beta_0 + \beta_1 x_i)]^2 \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma^2_{u_i}}\sum_{k=1}^{s}(X_{ik} - x_i)^2,\end{aligned}$$

e os estimadores de máxima verossimilhança para $\boldsymbol{\theta}$ novamente satisfazem as equações (2.4) − (2.6), com $\widehat{\omega}_i = 1/\widehat{\sigma}^2_{e_i}$. Quando o número de réplicas é o mesmo, isto é, $s_i = r_i$ para $i = 1, \ldots, n$, a função log-verossimilhança se reduz a

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= const - \frac{1}{2}\sum_{i=1}^{n} r_i(\log\sigma^2_{e_i} + \log\sigma^2_{u_i}) - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma^2_{e_i}}\sum_{j=1}^{r_i}[Y_{ij} - (\beta_0 + \beta_1 x_i)]^2 \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma^2_{u_i}}\sum_{k=1}^{r_i}(X_{ik} - x_i)^2.\end{aligned}$$

A equação (2.5) neste caso é dada por (2.10) e as demais equações coincidem com as encontradas no caso geral, com $\widehat{\omega}_i = r_i/\widehat{\sigma}^2_{e_i}$.

Em algumas situações, as variâncias dos erros de medição podem ser homoscedásticas ou então proporcionais. Nestes casos, podemos utilizar o modelo (2.1) − (2.3) normalmente. Suponha que as variâncias dos erros de medição são homoscedásticas, isto é, $\sigma^2_{u_i} = \sigma^2_u$ e $\sigma^2_{e_i} = \sigma^2_e$ para $i = 1, ..., n$. A função log-verossimilhança é então, dada

por

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= const - \frac{1}{2}\log\sigma_e^2\sum_{i=1}^{n} r_i - \frac{1}{2}\log\sigma_u^2\sum_{i=1}^{n} s_i \\ &\quad -\frac{1}{2}\frac{1}{\sigma_e^2}\sum_{i=1}^{n}\sum_{j=1}^{r_i}[Y_{ij} - (\beta_0 + \beta_1 x_i)]^2 - \frac{1}{2}\frac{1}{\sigma_u^2}\sum_{i=1}^{n}\sum_{k=1}^{s_i}(X_{ik} - x_i)^2.\end{aligned}$$

Os parâmetros do modelo neste caso são σ_u^2, σ_e^2, x_1, ..., x_n, β_0 e β_1 e os respectivos estimadores de máxima verossimilhança satisfazem as equações

$$\widehat{\sigma}_u^2 = \frac{\sum_{i=1}^{n}\sum_{k=1}^{s_i}(X_{ik} - \widehat{x}_i)^2}{\sum_{i=1}^{n} s_i},\ \widehat{\sigma}_e^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{r_i}[Y_{ij} - (\widehat{\beta}_0 + \widehat{\beta}_1\widehat{x}_i)]^2}{\sum_{i=1}^{n} r_i}$$

$$\text{e}\quad \widehat{x}_i = \frac{\widehat{\sigma}_u^2\widehat{\beta}_1 r_i(\overline{Y}_i - \widehat{\beta}_0) + \widehat{\sigma}_e^2 s_i\overline{X}_i}{\widehat{\beta}_1^2\widehat{\sigma}_u^2 r_i + \widehat{\sigma}_e^2 s_i},$$

e as equações (2.6) com $\widehat{\omega}_i = r_i$.

Barnett (1970) analisou o caso em que as variâncias dos erros de medição são proporcionais, isto é, $\sigma_{e_i}^2 = \rho\sigma_{u_i}^2$ para $i = 1, ..., n$, em que ρ é uma constante desconhecida. Neste caso a função log-verossimilhança é dada por

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= const - \frac{1}{2}\sum_{i=1}^{n}(r_i + s_i)\log\sigma_{u_i}^2 - \frac{1}{2}\sum_{i=1}^{n} r_i\log\rho \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\rho\sigma_{u_i}^2}\sum_{j=1}^{r_i}[Y_{ij} - (\beta_0 + \beta_1 x_i)]^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_{u_i}^2}\sum_{k=1}^{s_i}(X_{ik} - x_i)^2,\end{aligned}$$

sendo $\sigma_{u_1}^2$, ..., $\sigma_{u_n}^2$, ρ, x_1, ..., x_n, β_0 e β_1 os parâmetros do modelo. Os respectivos estimadores de máxima verossimilhança satisfazem as equações

$$\widehat{\sigma}_{u_i}^2 = \frac{\frac{1}{\widehat{\rho}}\sum_{j=1}^{r_i}[Y_{ij} - (\widehat{\beta}_0 + \widehat{\beta}_1\widehat{x}_i)]^2 + \sum_{k=1}^{s_i}(X_{ik} - \widehat{x}_i)^2}{r_i + s_i},$$

$$\widehat{\rho} = \frac{\sum_{i=1}^{n} \frac{1}{\widehat{\sigma}_{u_i}^2} \sum_{j=1}^{r_i} [Y_{ij} - (\widehat{\beta}_0 + \widehat{\beta}_1 \widehat{x}_i)]^2}{\sum_{i=1}^{n} r_i} \quad \text{e} \quad \widehat{x}_i = \frac{\widehat{\beta}_1 r_i (\overline{Y}_i - \widehat{\beta}_0) + \widehat{\rho} s_i \overline{X}_i}{\widehat{\beta}_1^2 r_i + \widehat{\rho} s_i},$$

e as equações (2.6) com $\widehat{\omega}_i = r_i / \widehat{\sigma}_{u_i}^2$.

Para obter as estimativas de máxima verossimilhança em todos os casos mencionados acima, faz-se necessário um método iterativo. O algoritmo do tipo EM (pseudo EM) proposto por Kimura (1992) fornece uma solução geral a diversos problemas de estimação por máxima verossimilhança. Assim, utilizamos esse algoritmo a fim de encontrar aproximações das estimativas para os parâmetros do modelo. Consideramos o caso geral, em que as variâncias dos erros de medição são heteroscedásticas e as réplicas são desemparelhadas e deslabanceadas. Para tal, seguimos os seguintes passos:

1. Inicie o procedimento iterativo com $a = 0$ e atribua valores iniciais $\beta_0^{(0)}$, $\beta_1^{(0)}$, $\sigma_{e_i}^{2^{(0)}}$ e $\sigma_{u_i}^{2^{(0)}}$. Como valores iniciais, podemos utilizar as estimativas encontradas por algum outro método, por exemplo, mínimos quadrados ponderados ou método dos momentos. Em problemas envolvendo comparação de métodos geralmente consideramos os valores iniciais abaixo

$$\begin{aligned} \beta_0^{(0)} &= 0,\ \beta_1^{(0)} = 1, \\ \sigma_{e_i}^{2^{(0)}} &= \frac{1}{r_i} \sum_{j=1}^{r_i} (Y_{ij} - \overline{Y}_i)^2 \quad \text{e} \quad \sigma_{u_i}^{2^{(0)}} = \frac{1}{s_i} \sum_{k=1}^{s_i} (X_{ik} - \overline{X}_i)^2, \end{aligned}$$

para $i = 1, ..., n$.

2. Calcule $\widehat{x}_i^{(a)}$ em (2.5).

3. Calcule os pesos estimados $\widehat{\omega}_i^{(a)}$ e as médias gerais ponderadas estimadas $\widehat{\overline{x}}_{\omega}^{(a)}$ e $\widehat{\overline{Y}}_{\omega}^{(a)}$ em (2.8) e (2.7), respectivamente.

4. Calcule $\widehat{\beta}_0^{(a)}$, $\widehat{\beta}_1^{(a)}$, $\widehat{\sigma}_{u_i}^{2^{(a)}}$ e $\widehat{\sigma}_{e_i}^{2^{(a)}}$, $i = 1, ..., n$ em (2.6) e (2.4), respectivamente.

5. Incremente a em uma unidade.

6. Repita os passos $2, 3, 4$ e 5 até a convergência.

O algoritmo é encerrado quando a convergência é obtida, ou seja, quando a diferença relativa entre as estimativas dos parâmetros, considerando os passos atual e o anterior, é suficientemente pequena. Seja

$$\delta = \max_{i=1,\ldots,n} \left\{ \left| \frac{\widehat{\sigma}_{u_i}^{2^{(a)}} - \widehat{\sigma}_{u_i}^{2^{(a-1)}}}{\widehat{\sigma}_{u_i}^{2^{(a-1)}}} \right|, \left| \frac{\widehat{\sigma}_{e_i}^{2^{(a)}} - \widehat{\sigma}_{e_i}^{2^{(a-1)}}}{\widehat{\sigma}_{e_i}^{2^{(a-1)}}} \right| \right\}.$$

Assim,

$$\epsilon^{(a)} = \max \left\{ \left| \frac{\widehat{\beta}_0^{(a)} - \widehat{\beta}_0^{(a-1)}}{\widehat{\beta}_0^{(a-1)}} \right|, \left| \frac{\widehat{\beta}_1^{(a)} - \widehat{\beta}_1^{(a-1)}}{\widehat{\beta}_1^{(a-1)}} \right|, \delta \right\}. \tag{2.11}$$

A matriz de covariâncias assintótica do estimador de máxima verossimilhança para os parâmetros β_0 e β_1, para a situação em que o número de observações n é fixo e o número de réplicas s_i e r_i cresce, é obtida pela inversão da matriz de informação (Dolby *et al.*, 1987; Patriota, 2006). No Apêndice encontram-se todas as derivadas necessárias. Tal matriz é dada por

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \begin{bmatrix} \sum_{i=1}^{n}(c_i/\sigma_{e_i}^2) & \sum_{i=1}^{n}(c_i x_i/\sigma_{e_i}^2) \\ \sum_{i=1}^{n}(c_i x_i/\sigma_{e_i}^2) & \sum_{i=1}^{n}(c_i x_i^2/\sigma_{e_i}^2) \end{bmatrix}^{-1}$$

em que $c_i = r_i \left(1 - \dfrac{\beta_1^2 r_i \sigma_{u_i}^2}{\beta_1^2 r_i \sigma_{u_i}^2 + s_i \sigma_{e_i}^2} \right)$.

A matriz de covariâncias assintótica é estimada consistentemente substituindo os parâmetros por aproximações das respectivas estimativas de máxima verossimilhança obtidas pelo algoritmo do tipo EM já apresentado. Sob condições de regularidade adequadas, os estimadores de máxima verossimilhança $(\widehat{\beta}_0, \widehat{\beta}_1)^\top$ obtidos nesta seção têm distribuição conjunta assintótica normal bivariada (Patriota, 2006).

2.1.2 Método dos momentos (MM)

Nesta subseção, derivamos os estimadores dos parâmetros do modelo (2.1)-(2.3) pelo método dos momentos. Estes resultados não foram encontrados na literatura. Sejam

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\overline{X}_i \quad \text{e} \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i,$$

os primeiros momentos amostrais, em que $\overline{X}_i$ e $\overline{Y}_i$ são dados por (2.9) e

$$S_{\overline{X}\,\overline{X}} = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{X}_i - \overline{X})^2 \ , \ S_{\overline{Y}\,\overline{Y}} = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{Y}_i - \overline{Y})^2$$

$$\text{e} \quad S_{\overline{X}\,\overline{Y}} = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{X}_i - \overline{X})(\overline{Y}_i - \overline{Y}),$$

os segundos momentos amostrais das médias $(\overline{X}_1, \ldots, \overline{X}_n)$ e $(\overline{Y}_1, \ldots, \overline{Y}_n)$. Sejam também $E[\overline{X}]$ e $E[\overline{Y}]$ os primeiros momentos populacionais e $E[S_{\overline{X}\,\overline{X}}]$, $E[S_{\overline{X}\,\overline{Y}}]$ e $E[S_{\overline{Y}\,\overline{Y}}]$ os segundos momentos populacionais, respectivamente, em que

$$E[\overline{X}] = \overline{x},\ E[\overline{Y}] = \beta_0 + \beta_1\overline{x} = \overline{y},$$

$$E[S_{\overline{X}\,\overline{X}}] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{s_i}\sigma^2_{u_i} + S_{xx}\ ,$$

$$E[S_{\overline{X}\,\overline{Y}}] = \beta_1 S_{xx} \quad \text{e} \quad E[S_{\overline{Y}\,\overline{Y}}] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i}\sigma^2_{e_i} + \beta_1^2 S_{xx}.$$

Igualando os momentos amostrais aos seus respectivos momentos populacionais e resolvendo o sistema de equações, obtemos que os estimadores para os parâmetros β_0 e β_1 obtidos pelo método dos momentos são dados por

$$\widehat{\beta}_{0_{mom_1}} = \overline{Y} - \widehat{\beta}_{1_{mom_1}}\overline{X}$$

$$\text{e} \quad \widehat{\beta}_{1mom_1} = \frac{(S_{\overline{Y}\,\overline{Y}} - S_{\overline{X}\,\overline{X}}W) + [(S_{\overline{Y}\,\overline{Y}} - S_{\overline{X}\,\overline{X}}W)^2 + 4S^2_{\overline{X}\,\overline{Y}}W]^{\frac{1}{2}}}{2S_{\overline{X}\,\overline{Y}}}$$

com

$$W = \frac{\sum_{i=1}^{n} \frac{1}{r_i}\widehat{\sigma}^2_{e_i}}{\sum_{i=1}^{n} \frac{1}{s_i}\widehat{\sigma}^2_{u_i}},$$

em que

$$\widehat{\sigma}^2_{u_i} = \frac{1}{s_i - 1}\sum_{k=1}^{s_i}(X_{ik} - \overline{X}_i)^2 \quad \text{e} \quad \widehat{\sigma}^2_{e_i} = \frac{1}{r_i - 1}\sum_{j=1}^{r_i}(Y_{ij} - \overline{Y}_i)^2. \tag{2.12}$$

Denotaremos esse procedimento por método dos momentos 1 (MM1).

Já o estimador de momentos de x_i, $i = 1, \ldots, n$ é obtido analogamente igualando o primeiro momento amostral $\overline{X}_i$, dado por (2.9), ao primeiro momento populacional $E[\overline{X}_i]$ da amostra $(X_{i1}, \ldots, X_{is_i})$, em que $E[\overline{X}_i] = x_i$. Assim, segue que

$$\widehat{x}_{i_{mom}} = \overline{X}_i, \quad i = 1, \ldots, n.$$

Uma outra possibilidade é considerarmos apenas os momentos $S_{\overline{X}\,\overline{X}}$ e $S_{\overline{X}\,\overline{Y}}$. Desta forma, os estimadores para os parâmetros β_0 e β_1 pelo método dos momentos são dados por

$$\widehat{\beta}_{0mom_2} = \overline{Y} - \widehat{\beta}_{1mom_2}\overline{X} \quad \text{e} \quad \widehat{\beta}_{1mom_2} = \frac{S_{\overline{X}\,\overline{Y}}}{S_{\overline{X}\,\overline{X}} - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{s_i}\widehat{\sigma}^2_{u_i}},$$

em que $\widehat{\sigma}^2_{u_i}$ é dado por (2.12). Neste caso, denotaremos esse procedimento por método dos momentos 2 (MM2).

Os estimadores $(\widehat{\beta}_{0mom_1}, \widehat{\beta}_{1mom_1})^\top$ e $(\widehat{\beta}_{0mom_2}, \widehat{\beta}_{1mom_2})^\top$ são assintoticamente não viesados. Entretanto, as distribuições assintóticas destes estimadores não são tão simples de serem encontradas. Desta forma, podemos utilizar a técnica de reamostragem *bootstrap* paramétrico ou não paramétrico (Efron, 1979; Efron & Tibshirani, 1993) para estimar a matriz de covariâncias de $(\widehat{\beta}_{0mom_1}, \widehat{\beta}_{1mom_1})^\top$ e de $(\widehat{\beta}_{0mom_2}, \widehat{\beta}_{1mom_2})^\top$, con-

siderando o modelo (2.1)-(2.3). Definamos $\boldsymbol{\vartheta} = (\beta_0, \beta_1)^\top$ e seja $\widehat{\boldsymbol{\vartheta}} = (\widehat{\beta}_{0_{mom}}, \widehat{\beta}_{1_{mom}})^\top$ o respectivo estimador de momentos de $\boldsymbol{\vartheta}$. Segue que a matriz de covariâncias de $\widehat{\boldsymbol{\vartheta}}$ é dada por meio da expressão

$$\widehat{\text{var}}(\widehat{\boldsymbol{\vartheta}}) = (Q-1)^{-1} \sum_{q=1}^{Q} (\widehat{\boldsymbol{\vartheta}}^{(q)} - \overline{\boldsymbol{\vartheta}})(\widehat{\boldsymbol{\vartheta}}^{(q)} - \overline{\boldsymbol{\vartheta}})^\top, \tag{2.13}$$

sendo que Q é o número de amostras *bootstrap*, $\widehat{\boldsymbol{\vartheta}}^{(q)}$ é a estimativa obtida da q-ésima amostra, $q = 1, \ldots, Q$ e $\overline{\boldsymbol{\vartheta}}$ é a média de $\widehat{\boldsymbol{\vartheta}}^{(1)}, \ldots, \widehat{\boldsymbol{\vartheta}}^{(Q)}$.

2.1.3 Método de extrapolação da simulação (SIMEX)

Em modelos com erros de medição Carroll *et al.* (2006) denominam de modelagem funcional as técnicas que demandam poucas suposições sobre a distribuição da verdadeira variável explicativa x quando esta é aleatória, podendo x também ser constante. Por outro lado, na modelagem estrutural a distribuição de probabilidade de x deve ser completamente especificada. Este ponto de vista é diferente da distinção tradicional entre modelos funcionais e modelos estruturais, apresentada na Seção 1.2.

O método SIMEX (extrapolação da simulação), proposto por Cook & Stefanski (1994) e detalhado em Carroll *et al.* (2006), é uma técnica de modelagem funcional geral e bastante difundida. Baseado em simulação, o método tem como principal característica a correção, pelo menos aproximada, do viés induzido nos estimadores dos parâmetros quando ignoramos os erros de medição na variável explicativa no processo de estimação. Como já mencionado, utilizar métodos de estimação que não consideram os erros de medição em seus procedimentos, pode resultar em estimadores inconsistentes (Fuller, 1987; Cheng & Van Ness, 1999). O efeito do erro de medição nas estimativas dos parâmetros pode ser observado na etapa de simulação, uma das etapas do método SIMEX.

Como vantagens do método SIMEX podemos destacar o fato de não ser necessário conhecer a distribuição da covariável envolvida, a facilidade com as quais as estimativas podem ser obtidas e a simplicidade para implementar. Entretanto, o custo computacional que para determinados tamanhos de amostras se torna alto e a dificuldade em se avaliar a variabilidade nas estimativas são as desvantagens do método.

Sucintamente, o método SIMEX fundamenta-se na idéia de que os efeitos dos erros de medição (sobre procedimentos estatísticos que ignoram estes erros) podem ser determinados experimentalmente por simulações, não sendo preciso supor a distribuição da verdadeira covariável não observada x. Em seguida, resultados válidos para o modelo com erros de medição são obtidos por extrapolação, atingindo-se uma situação em que se anulam os efeitos dos erros de medição.

O método SIMEX é dividido em duas etapas: simulação e extrapolação. Na etapa de simulação, erros de medição adicionais independentes são gerados e adicionados aos dados originais, criando novos conjuntos de dados, denominados pseudorreplicações. Existem duas versões para o método SIMEX que se diferenciam basicamente pela forma com que as pseudorreplicações das observações são geradas: o SIMEX paramétrico e o SIMEX empírico.

A versão paramétrica (Devanarayan & Stefanski, 2002) é apropriada para situações em que as variâncias dos erros de medição são conhecidas ou suficientemente bem estimadas. Entretanto, os estimadores $\widehat{\beta}_{0\text{SIMEX}}$ e $\widehat{\beta}_{1\text{SIMEX}}$ obtidos por meio do método SIMEX paramétrico não são em geral assintoticamente não viesados quando as variâncias dos erros de medição não são conhecidas, sendo então estimadas (Devanarayan, 1996). Isso ocorre uma vez que os erros adicionais relacionados à estimação das variâncias dos erros de medição não são levados em conta no processo de estimação dos parâmetros β_0 e β_1. Desta forma, considerando o modelo (2.1)-(2.3), trabalhamos apenas com a versão empírica do método SIMEX (Devanarayan & Stefanski, 2002)

em que pseudorreplicações são geradas na forma de contrastes lineares aleatórios das réplicas de acordo com

$$X_{bi}(\lambda) = \overline{X}_i + \left(\frac{\lambda}{s_i}\right)^{1/2} \sum_{k=1}^{s_i} d_{bik} X_{ik}, \tag{2.14}$$

em que

$$d_{bik} = \frac{Z_{bik} - \overline{Z}_{bi}}{\{\sum_{k=1}^{s_i}(Z_{bik} - \overline{Z}_{bi})^2\}^{1/2}}, \tag{2.15}$$

com $\sum_{k=1}^{s_i} d_{bik} = 0$, $\sum_{k=1}^{s_i} d_{bik}^2 = 1$ e $Z_{bik} \overset{\text{iid}}{\sim} N(0,1)$ independentes entre si e também independentes dos dados observados (X_{ik}, Y_{ij}), $j = 1, \ldots, r_i$, $k = 1, \ldots, s_i$, $i = 1, \ldots, n$ e $b = 1, \ldots, B$, sendo que B indica o número de pseudorréplicas na etapa de simulação e λ assume valores positivos representando a magnitude (em termos de variância) dos erros adicionados.

Um cálculo simples mostra que $E[X_{bi}(\lambda)] = x_i$ e $Var[X_{bi}(\lambda)] = (1 + \lambda)\sigma_{u_i}^2/s_i = (1+\lambda)Var[\overline{X}_i]$, ou seja, a variância total do erro adicionado é $(1+\lambda)\sigma_{u_i}^2/s_i$. Desta forma, pode ser provado que $X_{bi}(\lambda)|x_i \overset{\text{indep.}}{\sim} N(x_i, (1+\lambda)\sigma_{u_i}^2/s_i)$, $i = 1, \ldots, n$ (Devanarayan & Stefanski, 2002). O ideal é um conjunto de dados sem erros de medição na covariável, o que corresponde a $(1 + \lambda)\sigma_{u_i}^2/s_i = 0$, e assim $\lambda = -1$. Este fato é a principal propriedade dos pseudodados simulados no método SIMEX (Carroll *et al.*, 2006).

Na etapa de simulação, inicialmente fixamos o número de pseudorréplicas suficientemente grande, por exemplo, $B = 200$ e escolhemos o vetor de valores de λ, denotado por $\mathbf{\Lambda} = \{\lambda_1, \ldots, \lambda_M\}$. Em muitas aplicações consideramos $M = 10$ e $\lambda_m \in (0, 2]$, $m = 1 \ldots, M$. Desta forma, em (2.14), para uma dada pseudorréplica $b \in \{1, \ldots, B\}$ e um dado valor $\lambda \in \mathbf{\Lambda}$, formamos uma pseudoamostra $(\overline{Y}_1, X_{b1}(\lambda)), \ldots, (\overline{Y}_n, X_{bn}(\lambda))$, com a qual estimamos os coeficientes em (2.1) por mínimos quadrados ponderados obtendo $\widehat{\beta}_{0b}(\lambda)$ e $\widehat{\beta}_{1b}(\lambda)$. Repetimos o processo para $b = 1, \ldots, B$. As estimativas dos parâmetros β_0 e β_1, das B pseudoamostras, são então resumidas utilizando-se a média,

resultando em $\widehat{\beta}_0(\lambda)$ e $\widehat{\beta}_1(\lambda)$. Repetindo-se estes passos com $\lambda_1, \ldots, \lambda_M$, compomos os pares de pontos

$$(0, \widehat{\beta}_0), (\lambda_1, \widehat{\beta}_0(\lambda_1)), \ldots, (\lambda_M, \widehat{\beta}_0(\lambda_M)) \quad \text{e} \quad (0, \widehat{\beta}_1), (\lambda_1, \widehat{\beta}_1(\lambda_1)), \ldots, (\lambda_M, \widehat{\beta}_1(\lambda_M)),$$

sendo que $\widehat{\beta}_0$ e $\widehat{\beta}_1$ denotam as estimativas ingênuas, que ignoram os erros de medição, como se $x_i = \overline{X}_i$, $i = 1, \ldots, n$.

Já na etapa de extrapolação, gráficos de dispersão destes pares de pontos (estimativas viesadas obtidas na etapa de simulação em função de λ) podem ser vistos como um aprendizado sobre os efeitos dos erros de medição nas estimativas de mínimos quadrados ponderados, permitindo estabelecer uma relação entre o viés e a magnitude dos erros adicionados. Desta forma, ajustamos curvas (modelo linear, quadrático ou racional linear, por exemplo) a estes pontos e mediante extrapolação chegamos a $\widehat{\beta}_{0\text{SIMEX}} = \widehat{\beta}_0(-1)$ e $\widehat{\beta}_{1\text{SIMEX}} = \widehat{\beta}_1(-1)$. A Figura 2.1 traz uma ilustração do método SIMEX.

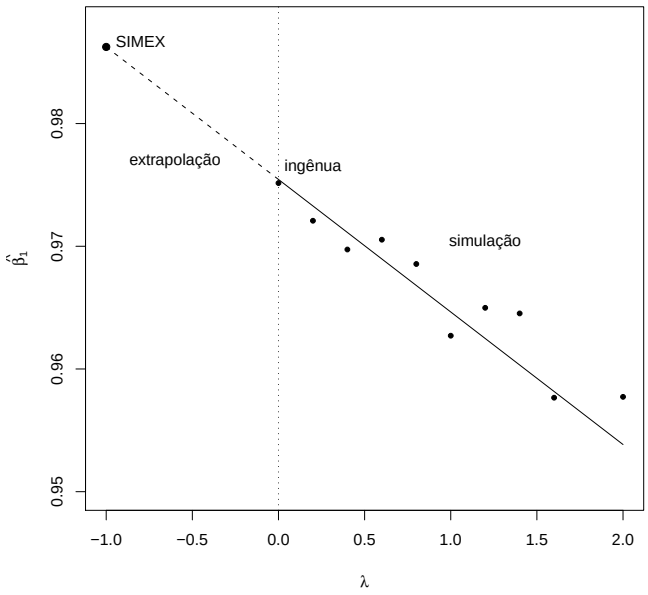

Figura 2.1: Exemplo de aplicação do método SIMEX com extrapolação por uma reta.

A matriz de covariâncias assintótica dos estimadores SIMEX pode ser obtida utilizando a abordagem de equações de estimação, como detalhada em Devanarayan (1996) e em Carroll *et al.* (2006).

Consideremos o vetor de parâmetros $\boldsymbol{\vartheta} = (\beta_0, \beta_1)^\top$ definido na Seção 2.1.2. Baseado em Devanarayan (1996), temos que uma equação de estimação para $\boldsymbol{\vartheta}$ é dada por

$$n^{-1} \sum_{i=1}^{n} \boldsymbol{\Psi}(\overline{Y}_i, x_i, \boldsymbol{\vartheta}) = \mathbf{0}. \tag{2.16}$$

Considerando o modelo (2.1) – (2.3) temos

$$\boldsymbol{\Psi}(\overline{Y}_i, x_i, \boldsymbol{\vartheta}) = \begin{pmatrix} \omega_i(\overline{Y}_i - \beta_0 - \beta_1 x_i) \\ \omega_i(\overline{Y}_i - \beta_0 - \beta_1 x_i)x_i \end{pmatrix},$$

em que $\omega_i = \dfrac{r_i}{\sigma^2_{e_i}}$. Note porém que os x_i's são inobserváveis e os pesos ω_i são desconhecidos. Neste caso, a solução de (2.16) não pode ser calculada.

Os pesos ω_i são desconhecidos uma vez que as variâncias dos erros $\sigma^2_{e_i}$ são desconhecidas. Entretanto, podemos estimá-los diretamente por meio das réplicas caso disponhamos de um grande número de replicações (Kutner *et al.*, 2005). Considerando as variâncias amostrais dadas por (2.12), estimamos os pesos ω_i substituindo $\sigma^2_{e_i}$ por $\widehat{\sigma}^2_{e_i}$, como feito na expressão (2.8). Desta forma, consideramos os pesos estimados como sendo os verdadeiros valores dos pesos ω_i.

Em relação as variâncias dos erros $\sigma^2_{u_i}$, apesar de desconhecidas não é necessário estimá-las já que na versão empírica do método SIMEX os pseudosdados são gerados usando uma combinação linear das réplicas sem a contribuição dos valores $\sigma^2_{u_i}$, $i = 1 \ldots, n$. Desta forma, não é preciso adicionar equações de estimação a (2.16). Isto torna a obtenção da matriz de covariâncias dos estimadores SIMEX empírico menos difícil (Devanarayan, 1996).

No passo de simulação, considerando um particular valor de b e λ, por uma aproximação linear obtemos

$$\sqrt{n}\{\widehat{\boldsymbol{\vartheta}}_b(\lambda) - \boldsymbol{\vartheta}(\lambda)\} \simeq -\boldsymbol{\mathcal{A}}^{-1}\{\sigma^2_{u_i}, \lambda, \boldsymbol{\vartheta}(\lambda)\} n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\Psi}(\overline{Y}_i, X_{bi}(\lambda), \boldsymbol{\vartheta}(\lambda)), \tag{2.17}$$

com

$$\boldsymbol{\mathcal{A}}\{\sigma^2_{u_i}, \lambda, \boldsymbol{\vartheta}(\lambda)\} = \frac{1}{n} \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \boldsymbol{\vartheta}} \boldsymbol{\Psi}(\overline{Y}_i, X_{bi}(\lambda), \boldsymbol{\vartheta}(\lambda))\right]. \tag{2.18}$$

Seja

$$\boldsymbol{\chi}_{Bi}\{\sigma^2_{u_i}, \lambda, \boldsymbol{\vartheta}(\lambda)\} = B^{-1} \sum_{b=1}^{B} \boldsymbol{\Psi}(\overline{Y}_i, X_{bi}(\lambda), \boldsymbol{\vartheta}(\lambda)).$$

Note que $\boldsymbol{\chi}_{Bi}(.)$, para i fixo ($i = 1, \ldots, n$), são independentes e identicamente distribuídos com média zero. Tomando a média sobre os B resultados de (2.17), obtemos a aproximação assintótica

$$\sqrt{n}\{\widehat{\boldsymbol{\vartheta}}(\lambda) - \boldsymbol{\vartheta}(\lambda)\} \simeq -\boldsymbol{\mathcal{A}}^{-1}\{\sigma^2_{u_i}, \lambda, \boldsymbol{\vartheta}(\lambda)\} n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\chi}_{Bi}\{\sigma^2_{u_i}, \lambda, \boldsymbol{\vartheta}(\lambda)\}. \tag{2.19}$$

Consideremos $\boldsymbol{\Lambda} = \{\lambda_1, \ldots, \lambda_M\}$ o vetor contendo os valores de λ selecionados na etapa de simulação e $\widehat{\boldsymbol{\vartheta}}_*(\boldsymbol{\Lambda}) = \{\widehat{\boldsymbol{\vartheta}}^\top(\lambda_1), \ldots, \widehat{\boldsymbol{\vartheta}}^\top(\lambda_M)\}^\top$ o vetor contendo as estimativas dos parâmetros para cada elemento de $\boldsymbol{\Lambda}$. Desta forma, usando a equação (2.19) temos que $\sqrt{n}\{\widehat{\boldsymbol{\vartheta}}_*(\boldsymbol{\Lambda}) - \boldsymbol{\vartheta}_*(\boldsymbol{\Lambda})\} \approx N(\mathbf{0}, \boldsymbol{\Sigma})$, com

$$\boldsymbol{\Sigma} = \boldsymbol{\mathcal{A}}_{11}^{-1}[(\sigma^2_{u_i}, \boldsymbol{\Lambda}, \boldsymbol{\vartheta}_*(\boldsymbol{\Lambda}))] \boldsymbol{C}_{11}(\sigma^2_{u_i}, \boldsymbol{\Lambda}, \boldsymbol{\vartheta}_*(\boldsymbol{\Lambda}))\{\boldsymbol{\mathcal{A}}_{11}^{-1}[(\sigma^2_{u_i}, \boldsymbol{\Lambda}, \boldsymbol{\vartheta}_*(\boldsymbol{\Lambda}))]\}^\top, \tag{2.20}$$

em que

$$\boldsymbol{\mathcal{A}}_{11}(\sigma^2_{u_i}, \boldsymbol{\Lambda}, \boldsymbol{\vartheta}_*(\boldsymbol{\Lambda})) = \text{diag}[\boldsymbol{\mathcal{A}}(\sigma^2_{u_i}, \lambda, \boldsymbol{\vartheta}(\lambda)),\ \lambda \in \boldsymbol{\Lambda}], \tag{2.21}$$

$$\boldsymbol{C}_{11}(\sigma^2_{u_i}, \boldsymbol{\Lambda}, \boldsymbol{\vartheta}_*(\boldsymbol{\Lambda})) = \text{Cov}[\boldsymbol{\Psi}_{B1(1)}(\sigma^2_{u_i}, \boldsymbol{\Lambda}, \boldsymbol{\vartheta}_*(\boldsymbol{\Lambda}))] \tag{2.22}$$

$$\text{e } \boldsymbol{\Psi}_{Bi(1)}(\sigma^2_{u_i}, \boldsymbol{\Lambda}, \boldsymbol{\vartheta}_*(\boldsymbol{\Lambda})) = \text{vec}[\boldsymbol{\chi}_{Bi}\{\sigma^2_{u_i}, \lambda, \boldsymbol{\vartheta}(\lambda)\},\ \lambda \in \boldsymbol{\Lambda}]. \tag{2.23}$$

A matriz (2.21) é estimada consistentemente por $\widehat{\boldsymbol{\mathcal{A}}}_{\mathbf{11}}(.) = \text{diag}\{\widehat{\boldsymbol{\mathcal{A}}}_{\boldsymbol{m}}(.)\}$, $m = 1, \ldots, M$, em que

$$\begin{aligned}\widehat{\boldsymbol{\mathcal{A}}}_m\{.\} &= \frac{1}{nB}\sum_{i=1}^{n}\sum_{b=1}^{B}\frac{\partial}{\partial\boldsymbol{\vartheta}}\boldsymbol{\Psi}[\overline{Y}_i, X_{bi}(\lambda_m), \widehat{\boldsymbol{\vartheta}}(\lambda_m)] \\ &= -\frac{1}{nB}\sum_{i=1}^{n}\sum_{b=1}^{B}\begin{bmatrix} \widehat{\omega}_i & \widehat{\omega}_i X_{bi}(\lambda_m) \\ \widehat{\omega}_i X_{bi}(\lambda_m) & \widehat{\omega}_i X_{bi}^2(\lambda_m) \end{bmatrix}.\end{aligned}$$

Já a matriz (2.22) é estimada consistentemente por meio da matriz de covariâncias amostral do vetor $[\boldsymbol{\Psi}_{Bi(1)}(\widehat{\sigma}^2_{u_i}, \boldsymbol{\Lambda}, \widehat{\boldsymbol{\vartheta}}_*(\boldsymbol{\Lambda})), \ldots, \boldsymbol{\Psi}_{Bn(1)}(\widehat{\sigma}^2_{u_i}, \boldsymbol{\Lambda}, \widehat{\boldsymbol{\vartheta}}_*(\boldsymbol{\Lambda}))]$.

Seja $\mathcal{G}^*(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}^*) = (\mathcal{G}(\lambda_1, \boldsymbol{\Gamma}^*)^\top, \ldots, \mathcal{G}(\lambda_M, \boldsymbol{\Gamma}^*)^\top)^\top$ em que $\mathcal{G}(\lambda_m, \boldsymbol{\Gamma}^*)$, $m = 1, \ldots, M$ é o modelo ajustado para cada um dos parâmetros $\widehat{\boldsymbol{\vartheta}}(\lambda)$ na etapa de extrapolação, calculado em $\lambda_m \in \boldsymbol{\Lambda}$. $\boldsymbol{\Gamma}^*$ é o vetor de parâmetros do modelo ajustado. No modelo de extrapolação linear simples, por exemplo, temos

$$\mathcal{G}(\lambda_m, \boldsymbol{\Gamma}^*) = \begin{pmatrix} \Gamma_0 + \Gamma_1\lambda_m \\ \Gamma_2 + \Gamma_3\lambda_m \end{pmatrix},$$

com $\boldsymbol{\Gamma}^* = (\Gamma_0, \Gamma_1, \Gamma_2, \Gamma_3)^\top$. Seja também $\boldsymbol{R}(\boldsymbol{\Gamma}^*) = \widehat{\boldsymbol{\vartheta}}_*(\boldsymbol{\Lambda}) - \mathcal{G}^*(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}^*)$ o resíduo do modelo ajustado. Podemos obter $\widehat{\boldsymbol{\Gamma}}^*$ pela soma dos quadrados dos resíduos do modelo ajustado, $\boldsymbol{R}(\boldsymbol{\Gamma}^*)^\top\boldsymbol{R}(\boldsymbol{\Gamma}^*)$, utilizando a equação de estimação $\boldsymbol{s}(\boldsymbol{\Gamma}^*)\boldsymbol{R}(\boldsymbol{\Gamma}^*) = \mathbf{0}$, em que $\boldsymbol{s}^\top(\boldsymbol{\Gamma}^*) = \{\partial/\partial(\boldsymbol{\Gamma}^*)^\top\}\boldsymbol{R}(\boldsymbol{\Gamma}^*)$. A teoria assintótica mostra que

$$n^{-1/2}(\widehat{\boldsymbol{\Gamma}}^* - \boldsymbol{\Gamma}^*) \approx N\{\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\Gamma}^*)\}, \tag{2.24}$$

em que $\boldsymbol{\Sigma}(\boldsymbol{\Gamma}^*) = \boldsymbol{D}^{-\mathbf{1}}(\boldsymbol{\Gamma}^*)\boldsymbol{s}(\boldsymbol{\Gamma}^*)\boldsymbol{\Sigma}\boldsymbol{s}^\top(\boldsymbol{\Gamma}^*)\boldsymbol{D}^{-\mathbf{1}}(\boldsymbol{\Gamma}^*)$, $\boldsymbol{D} = \boldsymbol{s}(\boldsymbol{\Gamma}^*)\boldsymbol{s}^\top(\boldsymbol{\Gamma}^*)$ e $\boldsymbol{\Sigma}$ é dada por (2.20).

Sendo $\widehat{\boldsymbol{\vartheta}}_{\text{SIMEX}} = \mathcal{G}^*(-1, \widehat{\boldsymbol{\Gamma}}^*)$ a estimativa do vetor de parâmetros $\boldsymbol{\vartheta}$ obtida na ausência de erro de medição por meio do método SIMEX, a teoria assintótica mostra que

$\sqrt{n}(\widehat{\boldsymbol{\vartheta}}_{\text{SIMEX}} - \boldsymbol{\vartheta}) \approx N(\mathbf{0}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\vartheta}}_{\text{SIMEX}}})$, ou seja, os estimadores $(\widehat{\beta}_{0\text{SIMEX}}, \widehat{\beta}_{1\text{SIMEX}})$ têm distribuição conjunta assintoticamente normal bivariada. A matriz de covariâncias assintótica $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\vartheta}}_{\text{SIMEX}}}$ é obtida pelo método delta e é dada por

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\vartheta}}_{\text{SIMEX}}} = \mathcal{G}^*_{\boldsymbol{\Gamma}^*}(-1, \boldsymbol{\Gamma}^*)\boldsymbol{\Sigma}(\boldsymbol{\Gamma}^*)\{\mathcal{G}^*_{\boldsymbol{\Gamma}^*}(-1, \boldsymbol{\Gamma}^*)\}^\top, \tag{2.25}$$

com $\mathcal{G}^*_{\boldsymbol{\Gamma}^*}(\lambda, \boldsymbol{\Gamma}^*) = \dfrac{\partial}{\partial(\boldsymbol{\Gamma}^*)^\top}\mathcal{G}^*(\lambda, \boldsymbol{\Gamma}^*)$.

Considerando o modelo linear simples na etapa de extrapolação, temos que o resíduo do modelo ajustado é dado por

$$\boldsymbol{R}(\widehat{\boldsymbol{\Gamma}}^*) = \begin{pmatrix} \widehat{\beta}_0(\lambda_1) - (\widehat{\Gamma}_0 + \widehat{\Gamma}_1\lambda_1) \\ \widehat{\beta}_1(\lambda_1) - (\widehat{\Gamma}_2 + \widehat{\Gamma}_3\lambda_1) \\ \vdots \\ \widehat{\beta}_0(\lambda_M) - (\widehat{\Gamma}_0 + \widehat{\Gamma}_1\lambda_M) \\ \widehat{\beta}_1(\lambda_M) - (\widehat{\Gamma}_2 + \widehat{\Gamma}_3\lambda_M) \end{pmatrix}.$$

Desta forma, segue que

$$\boldsymbol{s}^\top(\boldsymbol{\Gamma}^*) = -\begin{bmatrix} 1 & \lambda_1 & 0 & 0 \\ 0 & 0 & 1 & \lambda_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \lambda_M & 0 & 0 \\ 0 & 0 & 1 & \lambda_M \end{bmatrix},$$

e portanto, as estimativas para $\boldsymbol{\Gamma}^*$ são obtidas por meio da equação de estimação

$$\begin{pmatrix} \sum_{m=1}^{M} (\widehat{\beta}_0(\lambda_m) - \widehat{\Gamma}_0 - \widehat{\Gamma}_1\lambda_m) \\ \sum_{m=1}^{M} [\lambda_m(\widehat{\beta}_0(\lambda_m) - \widehat{\Gamma}_0 - \widehat{\Gamma}_1\lambda_m)] \\ \sum_{m=1}^{M} (\widehat{\beta}_1(\lambda_m) - \widehat{\Gamma}_2 - \widehat{\Gamma}_3\lambda_m) \\ \sum_{m=1}^{M} [\lambda_m(\widehat{\beta}_1(\lambda_m) - \widehat{\Gamma}_2 - \widehat{\Gamma}_3\lambda_m)] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Encontradas as estimativas $\widehat{\boldsymbol{\Gamma}}^*$, estimamos a matriz $\boldsymbol{\Sigma}(\boldsymbol{\Gamma}^*)$ dada em (2.24) substituindo $\boldsymbol{\Gamma}^*$ por $\widehat{\boldsymbol{\Gamma}}^*$. Sendo

$$\mathcal{G}^*_{\boldsymbol{\Gamma}^*}(\lambda, \boldsymbol{\Gamma}^*) = \begin{bmatrix} 1 & \lambda & 0 & 0 \\ 0 & 0 & 1 & \lambda \end{bmatrix},$$

obtemos a matriz $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\vartheta}}_{\text{SIMEX}}}$ por (2.25).

Observação: Existem outros métodos de estimação para modelos com erros de medição tais como o método de mínimos quadrados generalizado e o método de mínimos quadrados modificado (Cheng & Van Ness, 1999). Entretanto, não podemos utilizar estes métodos no modelo (2.1) − (2.3), já que neste caso temos réplicas das variáveis resposta e explicativa e as variâncias dos erros de medição são heteroscedásticas e desconhecidas.

2.2 Testes de hipóteses

Quando consideramos o modelo (2.1) − (2.3), estamos supondo que as variâncias dos erros de medição são heteroscedásticas. Entretanto, em algumas situações as variâncias dos erros podem ser homoscedásticas ou proporcionais (caso particular de variâncias heteroscedásticas). Nesses casos, podemos utilizar o modelo (2.1) − (2.3) normalmente, como visto na Seção 2.1.1, com a vantagem de que o número de parâmetros a serem

estimados é menor do que no caso em que as variâncias são heteroscedásticas. A homoscedasticidade das variâncias $\sigma^2_{u_i}$ e $\sigma^2_{e_i}$ pode ser testada considerando

$$\text{Hipóteses 1} = \begin{cases} H_0 : \sigma^2_{u_i} = \sigma^2_u \text{ e } \sigma^2_{e_i} = \sigma^2_e \text{ para } i = 1, ..., n \text{ contra} \\ H_1 : \sigma^2_{u_i} \neq \sigma^2_{u_j} \text{ ou } \sigma^2_{e_i} \neq \sigma^2_{e_j} \text{ para pelo menos um } i \neq j,\ i, j \in \{1, ..., n\}. \end{cases}$$

Utilizamos o teste da razão de verossimilhanças (RV). Assim, temos que a estatística do teste é dada por $\xi_{RV_1} = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widetilde{\boldsymbol{\theta}})\}$, sendo que $\ell(\boldsymbol{.})$ é a função log-verossimilhança dada na Seção 2.1.1, $\widehat{\boldsymbol{\theta}}$ é o estimador de máxima verossimilhança (EMV) irrestrito de $\boldsymbol{\theta}$ e $\widetilde{\boldsymbol{\theta}}$ é o EMV de $\boldsymbol{\theta}$ sob H_0. Tanto $\widehat{\boldsymbol{\theta}}$ quanto $\widetilde{\boldsymbol{\theta}}$ são obtidos por algoritmos do tipo EM (Kimura, 1992), por exemplo, e são apresentados na Seção 2.1.1. Neste caso, ξ_{RV_1} tem distribuição assintótica χ^2_{2n-2}, quando $s_i \to \infty$, $r_i \to \infty$, sendo que a taxa de crescimento é a mesma para $i = 1, ..., n$. Assim, rejeitamos H_0 com um nível de significância α se o valor de ξ_{RV_1} é maior do que o quantil $(1-\alpha)$ da distribuição χ^2_{2n-2}.

A proporcionalidade das variâncias também pode ser testada. Consideremos

$$\text{Hipóteses 2} = \begin{cases} H_0 : \sigma^2_{e_i} = \rho\sigma^2_{u_i} \text{ para } i = 1, ..., n \text{ contra} \\ H_1 : \sigma^2_{e_i} \neq \rho\sigma^2_{u_i} \text{ para pelo menos um } i \in \{1, ..., n\}, \end{cases}$$

sendo ρ uma constante conhecida. Utilizamos novamente o teste da razão de verossimilhanças (RV), cuja estatística de teste denotamos por ξ_{RV_2}. Neste caso, ξ_{RV_2} tem distribuição assintótica χ^2_{n-1}, quando $s_i \to \infty$, $r_i \to \infty$, sendo que a taxa de crescimento é a mesma para $i = 1, ..., n$. Assim, rejeitamos H_0 com um nível de significância α se o valor de ξ_{RV_2} é maior do que o quantil $(1 - \alpha)$ da distribuição χ^2_{n-1}.

Em várias aplicações as observações em (2.2) e (2.3) dizem respeito a medições de uma mesma quantidade desconhecida x efetuadas utilizando dois métodos, conforme apresentado por Ripley & Thompson (1987), Riu & Rius (1996) e Galea-Rojas *et al.* (2003), por exemplo. Neste contexto, β_0 e β_1 representam vícios de um método (y)

em relação ao outro (x), de modo que o teste da ausência de vícios, traduzida pela hipótese $H_0: \ (\beta_0, \beta_1)^\top = (0,1)^\top$, constitui questão de interesse. A estatística de Wald para testar

$$\text{Hipóteses 3} = \begin{cases} H_0 : (\beta_0, \beta_1)^\top = (\beta_{00}, \beta_{10})^\top \ \text{ contra} \\ H_1 : (\beta_0, \beta_1)^\top \neq (\beta_{00}, \beta_{10})^\top, \end{cases}$$

em que β_{00} e β_{10} são constantes conhecidas, é dada por

$$\xi_{W_{\beta_0,\beta_1}} = \begin{pmatrix} \widehat{\beta}_0 - \beta_{00} \\ \widehat{\beta}_1 - \beta_{10} \end{pmatrix}^\top \left[\widehat{\text{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1)\right]^{-1} \begin{pmatrix} \widehat{\beta}_0 - \beta_{00} \\ \widehat{\beta}_1 - \beta_{10} \end{pmatrix}.$$

Note que $(\widehat{\beta}_0, \widehat{\beta}_1)$ são as estimativas dos parâmetros (β_0, β_1) obtidas pelo método de máxima verossimilhança, método dos momentos ou método SIMEX e $\widehat{\text{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1)$ é um estimador da matriz de covariâncias de $(\widehat{\beta}_0, \widehat{\beta}_1)$. A matriz de covariâncias assintótica dos estimadores dos parâmetros (β_0, β_1) considerando os métodos de máxima verossimilhança e SIMEX foram apresentadas nas Seções 2.1.1 e 2.1.3, respectivamente. Já a matriz de covariâncias considerando o método dos momentos 1 ou 2 é obtida por meio da técnica *bootstrap* e foi apresentada na Seção 2.1.2.

Como visto nas Seções 2.1.1 e 2.1.3, $(\widehat{\beta}_0, \widehat{\beta}_1)$ obtidas via MV e SIMEX têm distribuição conjunta assintótica normal bivariada. Assim, sob H_0 temos que $\xi_{W_{\beta_0,\beta_1}} \xrightarrow{D} \chi^2_2$. É importante enfatizar que estamos considerando o caso em que n é fixo, $s_i \to \infty$ e $r_i \to \infty$, sendo que a taxa de crescimento de s_i e r_i é a mesma. Desta forma, rejeitamos H_0 com nível de significância α se o valor de $\xi_{W_{\beta_0,\beta_1}}$ é maior do que o quantil $(1-\alpha)$ da distribuição χ^2_2.

Os resultados assintóticos apresentados nesta seção podem não ser satisfatórios para determinados números de réplicas, tornando-se inadequados. Desta forma, podemos utilizar simultaneamente a técnica *bootstrap* paramétrica e não paramétrica (Davidson & MacKinnon, 2000) na construção de testes de hipóteses para testar as Hipóteses 1, 2 e

3. No método de máxima verossimilhança e no método dos momentos, encontramos os estimadores de todos os parâmetros do modelo $(2.1) - (2.3)$. Assim, podemos utilizar o método *bootstrap* paramétrico. Já no método SIMEX em que não encontramos os estimadores do parâmetro x_i, $i = 1 \ldots, n$, utilizamos o método *bootstrap* não paramétrico. Para determinadas quantidades de réplicas, a versão não paramétrica pode ficar comprometida uma vez que a geração de muitas réplicas fica limitada.

Nos testes de hipóteses por meio do método *bootstrap*, calcula-se o valor p empírico (Davidson & MacKinnon, 2000) encontrando a proporção de estatísticas $t^*_{(q)}$, $q = 1, \ldots, Q$ que sejam maiores do que a estatística t_0, sendo que Q é o número de amostras *bootstrap* geradas (sob H_0), $t^*_{(q)}$ é o valor da estatística de teste calculada na q-ésima amostra, $q = 1, \ldots, Q$ e t_0 é a estatística de teste calculada no conjunto de dados original. Por fim, compara-se o valor p empírico com o nível nominal adotado, rejeitando-se H_0 caso o valor p empírico seja menor do que o nível nominal.

Capítulo 3

Simulações

Neste capítulo realizamos um estudo de simulação com o objetivo de comparar o comportamento dos estimadores dos parâmetros β_0 e β_1, em relação ao viés e à raiz quadrada do erro quadrático médio. Consideramos todos os métodos de estimação apresentados no Capítulo 2. Além disso, analisamos os testes da razão de verossimilhanças e de Wald, dados na Seção 2.2, em relação às taxas de rejeição 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, próximos dos esperados.

Para todas as situações tratadas neste capítulo, no método de máxima verossimilhança adotamos que o erro $\epsilon^{(a)}$ dado pela expressão (2.11) deve ser menor do que 10^{-3} para que o algoritmo se encerre. Já no método SIMEX, consideramos $B = 200$ réplicas SIMEX e $M = 10$ com valores de λ igualmente espaçados no intervalo $(0, 2]$. Além disso, utilizamos a média das $B = 200$ estimativas SIMEX para obter a estimativa SIMEX correspondente a cada $\lambda \in (0, 2]$. Já na etapa de extrapolação, utilizamos o modelo linear simples por ser simples e apresentar um bom ajuste para este modelo.

Procurando criar um cenário semelhante à condição dos dados reais com o qual tra-

balharemos em detalhes no Capítulo 4, adotamos a distribuição uniforme com valores entre $0,5$ e $2,5$ para gerar o verdadeiro valor da covariável x_i, isto é, $x_i \sim U(0,5; 2,5)$. Além disso, consideramos que $\beta_0 = 0$ e $\beta_1 = 1$, que é uma situação razoável em problemas de comparação de métodos. Já as variâncias dos erros $\sigma^2_{u_i}$ e $\sigma^2_{e_i}$ são geradas respectivamente pelas expressões $\sigma^2_{u_i} = (0,5 \times x_i + 0,16)^2$ e $\sigma^2_{e_i} = (0,5 \times y_i + 0,29)^2$, sendo que neste caso, $y_i = x_i$. A escolha destas expressões foi baseada nas relações encontradas entre os desvios padrão amostrais e as concentrações, nos dados reais (cerâmica egípcia). Como os resultados assintóticos apresentados nesta dissertação consideram que o tamanho da amostra é fixo e o número de réplicas aumenta com o aumento do número de observações, fixamos $n = 21$ como nos dados reais. Em relação ao número de réplicas, consideramos algumas situações: (1) número mínimo de réplicas para x e y, ou seja, $s_i = r_i = 2$, (2) situação semelhante às condições dos dados reais utilizados no Capítulo 4 em que o número de réplicas de x e y varia de 2 a 18, sendo neste caso as réplicas desemparelhadas e desbalanceadas, (3) número de réplicas de x e y igual ao dobro do número existente nos dados reais, ou seja, número de réplicas variando entre 4 e 36, (4) número de réplicas de x e y constantes e igual a 18, que é o valor máximo de número de réplicas no conjunto de dados reais, (5) número de réplicas constante e igual a 40 e por fim, (6) número de réplicas constante e igual a 80, sendo que esta última situação foi abordada apenas na Seção 3.2, para os testes da razão de verossimilhanças e de Wald considerando o método de máxima verossimilhança. Já para o teste de Wald considerando o método SIMEX, não utilizamos a situação (6) devido ao custo computacional. É importante destacar que começamos o estudo de simulação adotando o número mínimo de réplicas ($s_i = r_i = 2$) e aumentamos essas quantidades sem que o tamanho da amostra aumentasse. Em cada caso analisado, geramos 5000 conjuntos de dados.

3.1 EQM e viés das estimativas $(\widehat{\beta}_0, \widehat{\beta}_1)$

Nesta seção, calculamos o viés simulado e a raiz quadrada do erro quadrático médio ($\sqrt{\text{EQM}}$) simulado. O viés simulado é calculado fazendo a diferença entre a média das estimativas do parâmetro e o verdadeiro valor do parâmentro em questão. Já a raiz quadrada do erro quadrático médio ($\sqrt{\text{EQM}}$) simulado é dado pela expressão

$$\sqrt{\text{EQM}} = \left[\sum_{i=1}^{R}(\widehat{\gamma}_i - \gamma)^2/R\right]^{1/2},$$

em que R é o número de simulações e γ é o valor verdadeiro do parâmetro do modelo.

Na Tabela 3.1 apresentamos o viés simulado e a raiz quadrada do EQM simulado das estimativas $\widehat{\beta}_0$ e $\widehat{\beta}_1$ obtidas pelos métodos de estimação abordados nesta dissertação, considerando o cenário descrito anteriormente. Já os histogramas com os valores das estimativas $\widehat{\beta}_0$ e $\widehat{\beta}_1$ são dados respectivamente pelas Figuras 3.2 e 3.3, complementando os resultados vistos na Tabela 3.1. Os resultados apresentados não foram encontrados na literatura.

Tabela 3.1: Viés simulado e $\sqrt{\text{EQM}}$ simulado das estimativas $(\widehat{\beta}_0, \widehat{\beta}_1)$ obtidas pelo método de máxima verossimilhança, pelo método dos momentos 1 e 2 e pelo método SIMEX - $x_i \sim U(0,5; 2,5)$ com $n = 21$.

Número de réplicas s_i	Número de réplicas r_i	Método de estimação	Viés		$\sqrt{\text{EQM}}$	
			$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
		MV	$0,80538$	$-0,48460$	$2,00640$	$1,17985$
mínimo: 2	mínimo: 2	MM1	$\mathbf{0,20757}$	$\mathbf{-0,10453}$	$54,27245$	$30,34556$
médio: 2	médio: 2	MM2	$2,37874$	$-1,25097$	$150,13865$	$78,77016$
máximo: 2	máximo: 2	SIMEX	$1,04030$	$-0,61944$	$\mathbf{1,86172}$	$\mathbf{1,05682}$
		MV	$\mathbf{0,01117}$	$\mathbf{-0,01084}$	$0,35634$	$0,38121$
mínimo: 2	mínimo: 2	MM1	$-0,03876$	$0,03477$	$\mathbf{0,32862}$	$\mathbf{0,27913}$
médio: $7,4$	médio: $7,5$	MM2	$-0,10304$	$0,08611$	$1,02645$	$0,84760$
máximo: 18	máximo: 18	SIMEX	$0,17694$	$-0,17497$	$0,48423$	$0,42994$
		MV	$\mathbf{-0,00871}$	$\mathbf{0,00863}$	$\mathbf{0,14919}$	$\mathbf{0,12658}$
mínimo: 4	mínimo: 4	MM1	$-0,01483$	$0,01414$	$0,22010$	$0,17520$
médio: $14,8$	médio: $15,0$	MM2	$-0,04425$	$0,03469$	$0,28328$	$0,21694$
máximo: 36	máximo: 36	SIMEX	$0,03679$	$-0,03195$	$0,17489$	$0,15003$
		MV	$\mathbf{-0,01109}$	$\mathbf{0,00897}$	$\mathbf{0,17191}$	$\mathbf{0,13986}$
mínimo: 18	mínimo: 18	MM1	$-0,01356$	$0,01036$	$0,19520$	$0,15351$
médio: 18	médio: 18	MM2	$-0,02894$	$0,02118$	$0,21742$	$0,16861$
máximo: 18	máximo: 18	SIMEX	$0,04038$	$-0,03545$	$0,17532$	$0,14259$
		MV	$\mathbf{-0,00104}$	$\mathbf{0,00169}$	$\mathbf{0,08156}$	$\mathbf{0,08065}$
mínimo: 40	mínimo: 40	MM1	$-0,00246$	$0,00282$	$0,09211$	$0,08815$
médio: 40	médio: 40	MM2	$-0,00666$	$0,00622$	$0,09580$	$0,09085$
máximo: 40	máximo: 40	SIMEX	$0,00685$	$-0,00690$	$0,08288$	$0,08184$

Ilustramos na Figura 3.1 o gráfico de dispersão entre as médias das réplicas para cada observação e os respectivos desvios padrão estimados, considerando um dos conjuntos de dados simulados.

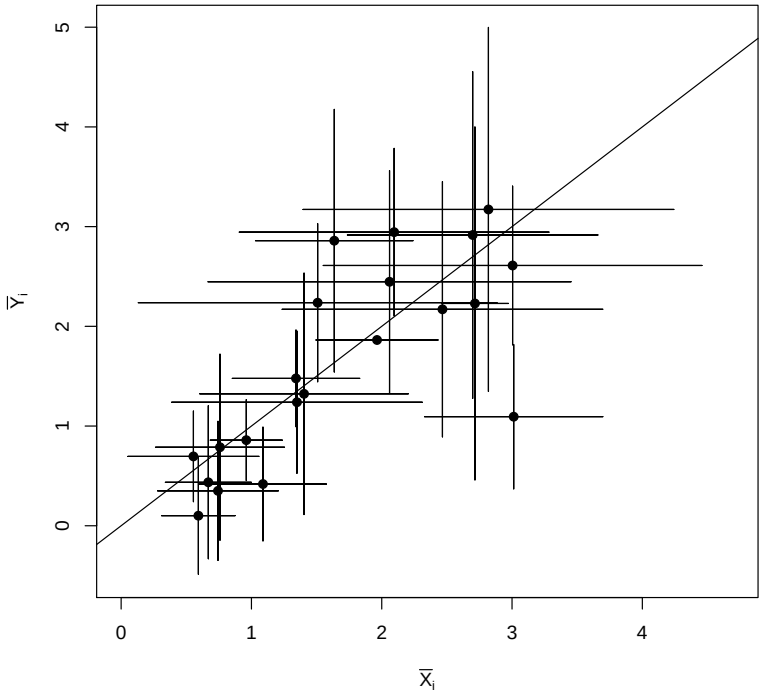

Figura 3.1: Gráfico de dispersão entre as médias das réplicas para cada observação com a reta identidade.

Podemos observar pela Figura 3.1 que os valores dos desvios padrão estimados em geral aumentam com o aumento dos valores das médias das réplicas, dando um indicativo da heteroscedasticidade dos dados.

Nas Figuras 3.2 e 3.3 apresentamos os histogramas das estimativas $\widehat{\beta}_0$ e $\widehat{\beta}_1$ considerando todos os métodos de estimação apresentados no Capítulo 2. Podemos notar em ambas as figuras que à medida que aumentamos o número de réplicas, a dispersão nos valores encontrados para as estimativas $\widehat{\beta}_0$ e $\widehat{\beta}_1$ se torna menor. Além disso, analisando o comportamento dos gráficos temos indícios de que as estimativas seguem uma distribuição normal.

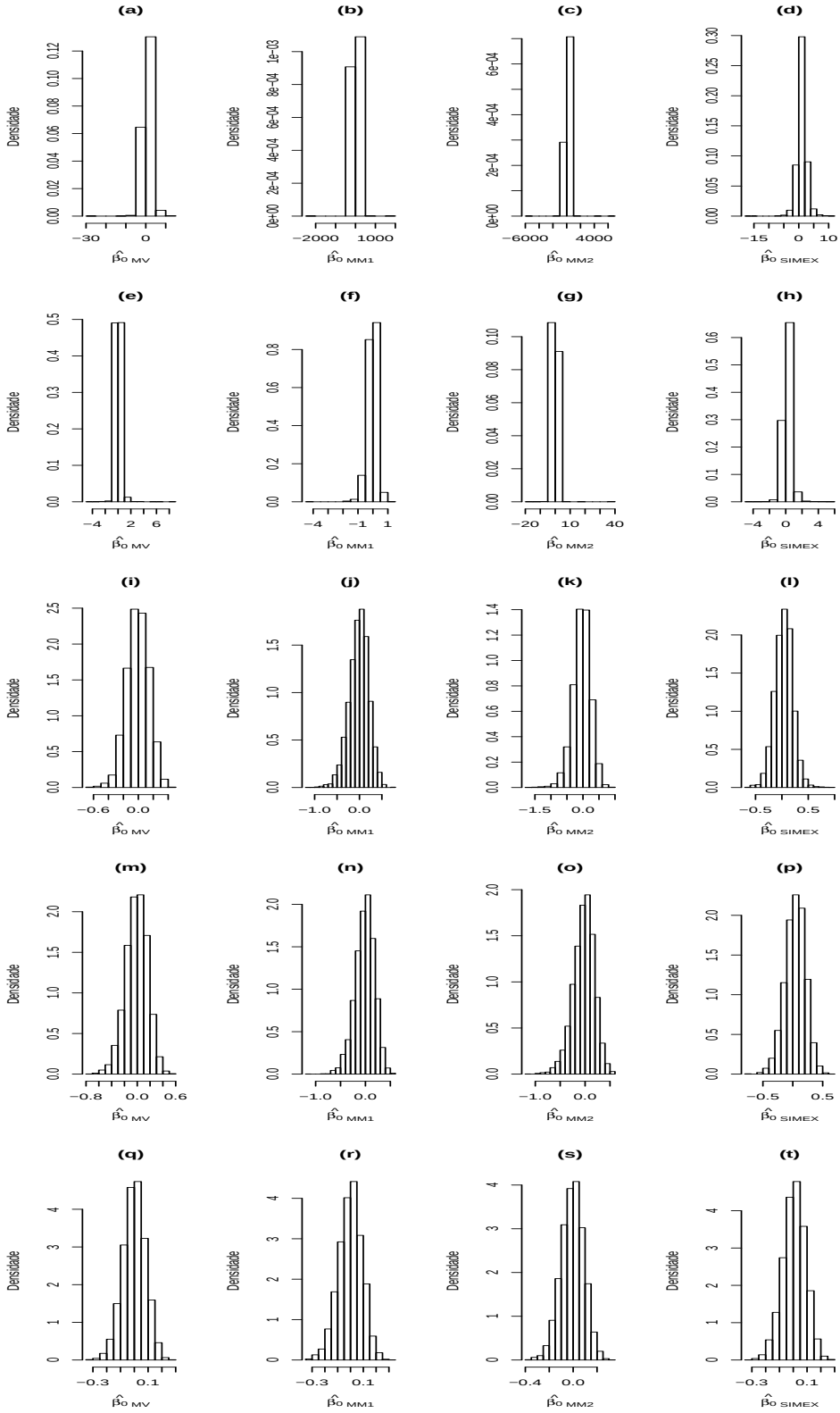

Figura 3.2: Histogramas das 5000 estimativas $\widehat{\beta}_0$ - (a), (b), (c) e (d): $s_i = r_i = 2$; (e), (f), (g) e (h): s_i, r_i entre 2 e 18; (i), (j), (k) e (l): s_i, r_i entre 4 e 36; (m), (n), (o) e (p): $s_i = r_i = 18$ e (q), (r), (s) e (t): $s_i = r_i = 40$ - $x_i \sim U(0,5; 2,5)$ com $n = 21$.

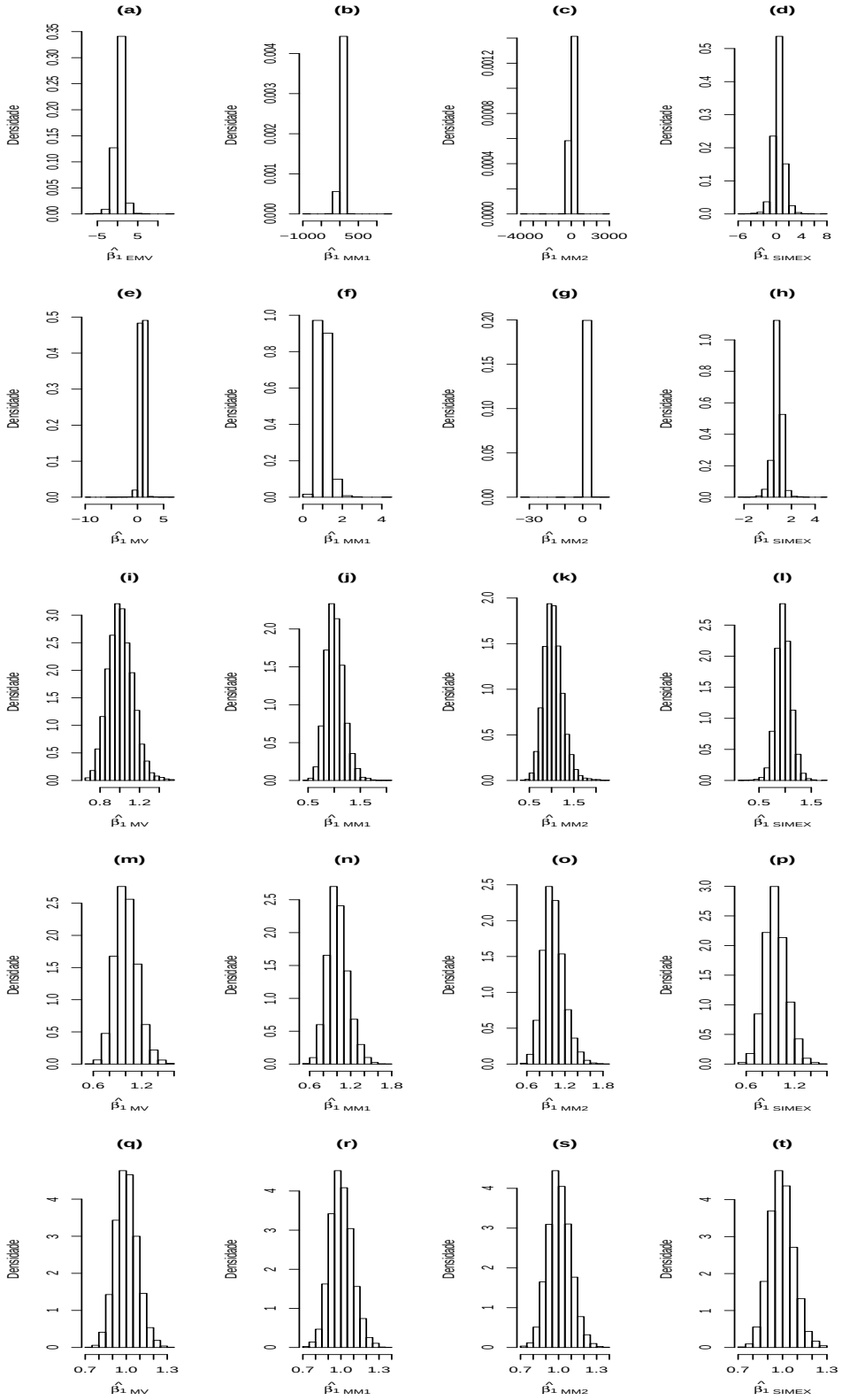

Figura 3.3: Histogramas das 5000 estimativas $\widehat{\beta}_1$ - (a), (b), (c) e (d): $s_i = r_i = 2$; (e), (f), (g) e (h): s_i, r_i entre 2 e 18; (i), (j), (k) e (l): s_i, r_i entre 4 e 36; (m), (n), (o) e (p): $s_i = r_i = 18$ e (q), (r), (s) e (t): $s_i = r_i = 40$ - $x_i \sim U(0,5; 2,5)$ com $n = 21$.

Observamos pela Tabela 3.1 que quando dispomos de um número mínimo de réplicas de x e y, isto é, $s_i = r_i = 2$ para $i = 1, \ldots, n$, o método dos momentos 1 apresenta melhores resultados, em termos de viés simulado de $\widehat{\beta}_0$ e de $\widehat{\beta}_1$, do que os demais métodos. Entretanto, a variabilidade nas estimativas encontradas por este método é a mais alta entre os métodos analisados, o que pode ser explicado pela quantidade de réplicas disponível. O método SIMEX e o método de máxima verossimilhança, apesar de não apresentarem os menores valores de viés simulado dos estimadores dos parâmetros β_0 e β_1, produzem os menores valores da raiz quadrada do erro quadrático médio como visto nas Figuras 3.2 e 3.3. Desta forma, para $s_i = r_i = 2$ optamos pelo método dos momentos 1 apesar dos altos valores encontrados para $\sqrt{\text{EQM}}$.

Para quantidade de réplicas semelhante à existente no conjunto de dados estudado no Capítulo 4, isto é, s_i e r_i entre 2 e 18, o método de máxima verossimilhança mostra-se melhor em termos de viés simulado de $\widehat{\beta}_0$ e $\widehat{\beta}_1$ já que apresenta valores abaixo dos encontrados pelos demais métodos. Já o método SIMEX apresenta o maior valor de viés simulado dos estimadores de ambos os parâmetros. Em relação à $\sqrt{\text{EQM}}$, o método de máxima verossimilhança e o método dos momentos 1 são mais eficientes do que o método SIMEX, sendo os valores obtidos semelhantes. O método dos momentos 2 apresenta os maiores valores de $\sqrt{\text{EQM}}$ para os estimadores de ambos os parâmetros, o que também pode ser notado nas Figuras 3.2 (g) e 3.3 (g). Portanto, neste caso optamos pelo método de máxima verossimilhança. O método dos momentos 1 também pode ser utilizado.

Em relação aos demais casos abordados, em que o número de réplicas s_i e r_i é fixo e igual a 18, varia entre 4 e 36 ou é fixo e igual a 40, observamos que entre os métodos analisados, o método de máxima verossimilhança mostra-se melhor em termos de valores de viés simulado e de $\sqrt{\text{EQM}}$ simulado, apresentando os menores valores tanto para o estimador $\widehat{\beta}_0$ quanto para $\widehat{\beta}_1$. Já o método dos momentos 2 em todos os casos é o método que tem o pior desempenho em relação à $\sqrt{\text{EQM}}$. O método SIMEX consegue

reduzir o viés produzindo valores semelhantes aos encontrados pelos demais métodos. Entretanto, o método SIMEX não elimina o viés totalmente, sendo que os valores de viés obtidos ainda são maiores do que os obtidos pelos métodos de máxima verossimilhança e dos momentos 1. Já os valores de $\sqrt{\text{EQM}}$ obtidos no método SIMEX são próximos dos valores obtidos no método de máxima verossimilhança. Assim, optamos pelo método de máxima verossimilhança sendo o método dos momentos 1 e SIMEX boas alternativas.

Verificamos que o viés dos estimadores dos parâmetros β_0 e β_1 em todos os métodos diminui à medida que a quantidade de réplicas aumenta, tendendo a 0. Isto ocorre uma vez que $\widehat{\beta}_0$ e $\widehat{\beta}_1$ são assintoticamente não viesados para todos os métodos utilizados, como visto no Capítulo 2. Não podemos esquecer que nesta dissertação o tamanho da amostra é fixo e os resultados assintóticos são relacionados à quantidade de réplicas. Além disso, devemos destacar que em todos os casos tratados neste capítulo o método dos momentos 1 (MM1) teve melhor desempenho do que o método dos momentos 2 (MM2), em termos de valores de viés simulado e $\sqrt{\text{EQM}}$, considerando $\widehat{\beta}_0$ e $\widehat{\beta}_1$. Isso pode ser explicado pelo fato de que no método dos momentos 1, o segundo momento amostral $S_{\overline{Y}\,\overline{Y}}$ é considerado, o que não ocorre no método dos momentos 2. Na sequência desta dissertação consideraremos apenas o método dos momentos 1.

Desta forma, concluímos que o método de máxima verossimilhança, em geral, tem melhor comportamento em relação aos demais métodos, em termos de viés simulado e raiz quadrada do erro quadrático médio simulado considerando o cenário descrito na introdução do Capítulo 3. O método SIMEX, apesar de reduzir o viés, não o elimina totalmente, podendo ser utilizado simultaneamente ao método de máxima verossimilhança. O método dos momentos 1 se mostrou uma boa alternativa, apesar da grande variabilidade das estimativas para o caso em que $s_i = r_i = 2$.

3.2 Taxa de rejeição dos testes da razão de verossimilhanças (RV) e Wald

Nesta seção analisamos os desempenhos dos testes da razão de verossimilhanças e de Wald, apresentados na Seção 2.2, em relação às taxas de rejeição sob H_0 e sob H_1, considerando o cenário descrito na introdução do Capítulo 3. Em todos os testes analisados, a probabilidade do erro tipo I foi estimada gerando amostras sob a hipótese H_0 e calculando a proporção de simulações em que o teste rejeitou a hipótese nula. Já a taxa de rejeição sob H_1 foi calculada gerando amostras sob a hipótese H_1 e calculando a proporção de simulações em que o teste rejeitou a hipótese nula. Como já mencionado, geramos em cada caso 5000 conjuntos de dados.

3.2.1 Homoscedasticidade das variâncias

Sejam as Hipóteses 1 apresentadas na Seção 2.2. Consideramos que sob H_0, as variâncias dos erros $\sigma^2_{u_i}$ e $\sigma^2_{e_i}$ são dadas respectivamente por $\sigma^2_{u_i} = (0,5 \times x_{12} + 0,16)^2$ e $\sigma^2_{e_i} = (0,5 \times y_{12} + 0,29)^2$, $i = 1, ..., n$, em que $y_{12} = \beta_0 + \beta_1 x_{12} = x_{12}$ neste caso.

Sob H_1, analisamos três situações denotadas por $H_{1(1)}$, $H_{1(2)}$ e $H_{1(3)}$. Em $H_{1(1)}$ consideramos que as variâncias dos erros $\sigma^2_{u_i}$ e $\sigma^2_{e_i}$ são geradas por meio das expressões $\sigma^2_{u_i} = [0,97 + N(0;0,4)]^2$ e $\sigma^2_{e_i} = [0,85 + N(0;0,4)]^2$, em que $N(0;0,4)$ indica que geramos para cada i, $i = 1 \ldots, n$, um valor de uma variável aleatória com distribuição normal com média 0 e variância $0,4$. Já em $H_{1(2)}$, adotamos que $\sigma^2_{u_i} = [0,97 + N(0;0,8)]^2$ e $\sigma^2_{e_i} = [0,85 + N(0;0,8)]^2$ com $N(0;0,8)$ análogo ao caso anterior mas com variância igual a $0,8$ e em $H_{1(3)}$, que as variâncias dos erros $\sigma^2_{u_i}$ e $\sigma^2_{e_i}$ são dadas pelas expressões $\sigma^2_{u_i} = [0,97 + N(0;1,2)]^2$ e $\sigma^2_{e_i} = [0,85 + N(0;1,2)]^2$, respectivamente. A diferença entre $H_{1(1)}$, $H_{1(2)}$ e $H_{1(3)}$ está no valor adotado para a variância da distribuição nor-

mal. Desta forma, espera-se que os valores gerados para as variâncias dos erros de medição em $H_{1(1)}$ sejam mais próximos uns dos outros do que os valores gerados para as variâncias dos erros de medição em $H_{1(3)}$, sendo $H_{1(2)}$ o caso intermediário.

Sob H_0, em $0,06\%$ das amostras geradas para $s_i = r_i = 2$ e em $1,10\%$ das amostras geradas para s_i, r_i entre 2 e 18, o valor da estatística da razão de verossimilhanças foi negativo. Nestes casos, descartamos as respectivas amostras da análise. Os resultados obtidos são apresentados na Tabela 3.2 e não foram encontrados na literatura. Gráficos de quantis para a estatística de teste da razão de verossimilhanças também foram apresentados na Figura 3.4.

Tabela 3.2: Taxas de rejeição (%) das hipóteses H_0, $H_{1(1)}$, $H_{1(2)}$ e $H_{1(3)}$ - teste RV - homoscedasticidade das variâncias - $x_i \sim U(0,5; 2,5)$ com $n = 21$ para um nível de significância $\alpha = 0,05$.

Número de réplicas s_i	Número de réplicas r_i	Sob H_0	Sob $H_{1(1)}$	Sob $H_{1(2)}$	Sob $H_{1(3)}$
mínimo: 2 médio: 2 máximo: 2	mínimo: 2 médio: 2 máximo: 2	97,29	99,76	99,92	99,98
mínimo: 2 médio: 7,4 máximo: 18	mínimo: 2 médio: 7,5 máximo: 18	56,80	96,66	100,00	100,00
mínimo: 4 médio: 14,8 máximo: 36	mínimo: 4 médio: 15,0 máximo: 36	19,06	99,14	100,00	100,00
mínimo: 18 médio: 18 máximo: 18	mínimo: 18 médio: 18 máximo: 18	9,36	100,00	100,00	100,00
mínimo: 40 médio: 40 máximo: 40	mínimo: 40 médio: 40 máximo: 40	6,68	100,00	100,00	100,00
mínimo: 80 médio: 80 máximo: 80	mínimo: 80 médio: 80 máximo: 80	5,60	100,00	100,00	100,00

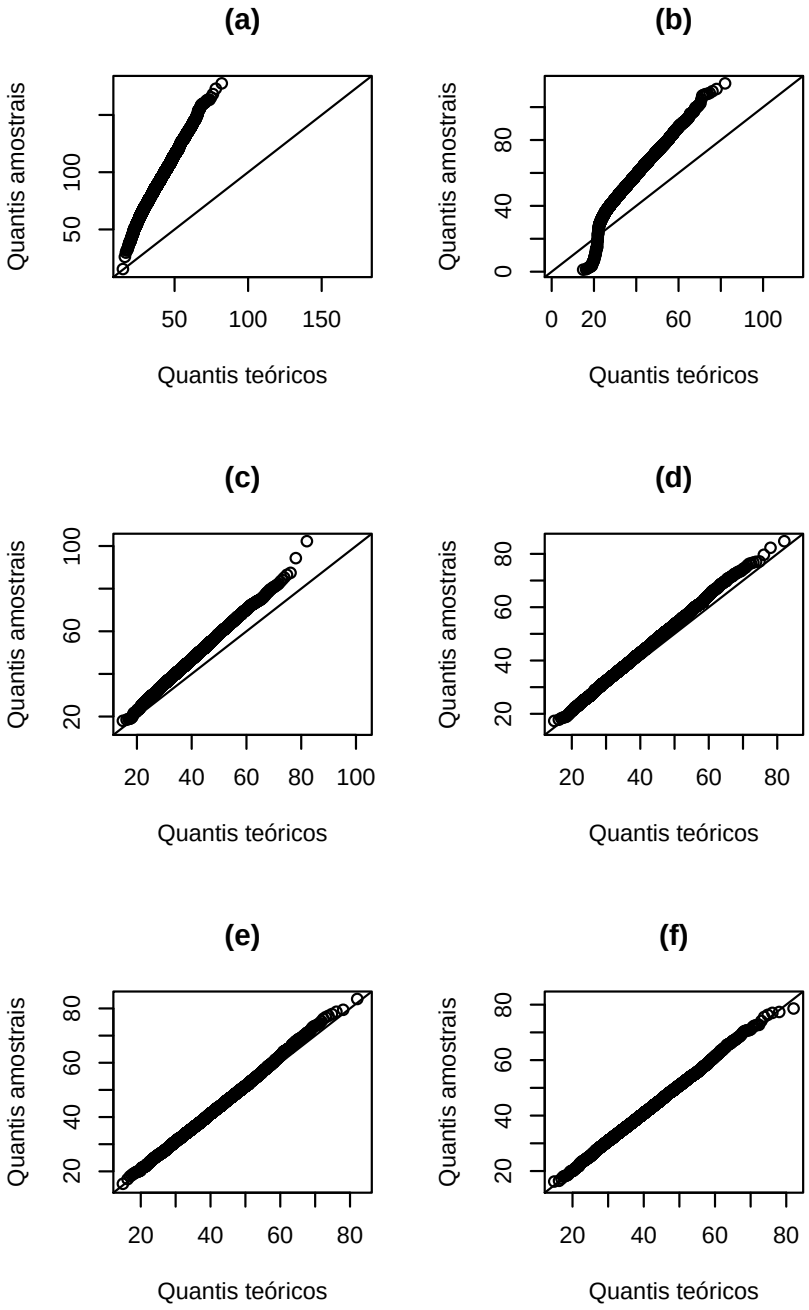

Figura 3.4: Gráficos de quantis da distribuição χ^2_{2n-2} - teste RV - homoscedasticidade (a) $s_i = r_i = 2$, (b) s_i, r_i entre 2 e 18, (c) s_i, r_i entre 4 e 36, (d) $s_i = r_i = 18$, (e) $s_i = r_i = 40$ e (f) $s_i = r_i = 80$ - $x_i \sim U(0,5; 2,5)$ com $n = 21$ para um nível de significância $\alpha = 0,05$.

Pela Figura 3.4 verificamos que quando dispomos de um número mínimo de réplicas, isto é, $s_i = r_i = 2$, ou um número de réplicas entre 2 e 18, não há indícios de que a

estatística da razão de verossimilhanças tenha distribuição χ^2_{2n-2}. Nestes casos vemos pela Tabela 3.2 que a taxa de rejeição sob H_0 é alta, sendo os valores encontrados distantes do valor nominal de 5%. Entretanto, fixando-se o tamanho da amostra e aumentando o número de réplicas, percebemos que a taxa de rejeição sob H_0 diminui, o que também é notado na Figura 3.4 em que os quantis teóricos e amostrais se aproximam com o aumento do número de réplicas. Percebemos que no cenário analisado, para número de réplicas igual a 80 a taxa de rejeição sob H_0 é semelhante ao nível nominal de 5%, indicando que para esta quantidade de réplicas, o resultado assintótico é satisfatório. Para as situações (número de réplicas) em que a taxa de rejeição sob H_0 foi alta, podemos utilizar alternativamente o teste de hipóteses *bootstrap* apresentado na Seção 2.2 a fim de testar as Hipóteses 1.

Em relação à taxa de rejeição sob H_1, notamos que os valores encontrados em todos os casos analisados ($H_{1(1)}$, $H_{1(2)}$ e $H_{1(3)}$) são altos, como desejados.

Logo, 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 80, no cenário analisado. Nestes casos, podemos utilizar alternativamente a técnica teste de hipóteses *bootstrap* dada na Seção 2.2.

3.2.2 Proporcionalidade das variâncias

Em relação às Hipóteses 2, dadas na Seção 2.2, consideramos que sob H_0, $\sigma^2_{u_i} = (0,5 \times x_i + 0,16)^2$ e $\rho = 1,5$ e portanto, $\sigma^2_{e_i} = 1,5 \times \sigma^2_{u_i}$ para $i = 1, ..., n$. É importante lembrar que ρ é a razão entre as variâncias $\sigma^2_{e_i}$ e $\sigma^2_{u_i}$ e que sob H_0, ρ é constante. Sob H_1, para cada observação i, geramos um valor para ρ de modo que $\sigma^2_{e_i}/\sigma^2_{u_i}$ seja não constante. Analisamos três situações denotadas por $H_{1(1)}$, $H_{1(2)}$ e $H_{1(3)}$. Em $H_{1(1)}$ consideramos que $\sigma^2_{u_i} = (0,5 \times x_i + 0,16)^2$ e $\sigma^2_{e_i} = \sigma^2_{u_i} \times |N(1,5;1)|$, isto é, geramos para cada i um número aleatório de uma distribuição normal com média $1,5$

e variância 1 e consideramos o valor absoluto do número gerado. Em $H_{1(2)}$ adotamos $\sigma^2_{u_i} = (0,5 \times x_i + 0,16)^2$ e $\sigma^2_{e_i} = \sigma^2_{u_i} \times |N(1,5;2)|$ e em $H_{1(3)}$ adotamos $\sigma^2_{u_i} = (0,5 \times x_i + 0,16)^2$ e $\sigma^2_{e_i} = \sigma^2_{u_i} \times |N(1,5;3)|$. Logo, espera-se que os valores gerados para ρ em $H_{1(1)}$ sejam mais próximos uns dos outros do que os valores gerados em $H_{1(3)}$ já que a variância da distribuição de ρ (distribuição normal) em $H_{1(1)}$ é menor. $H_{1(2)}$ seria a etapa intermediária, em que espera-se que os valores gerados para ρ sejam menos próximos do que os gerados em $H_{1(1)}$ e mais próximos do que os gerados em $H_{1(3)}$.

Sob H_0, em $0,04\%$ das amostras geradas para $s_i = r_i = 2$ e em $0,52\%$ das amostras geradas para s_i, r_i entre 2 e 18, o valor da estatística da razão de verossimilhanças foi negativo. Desta forma, assim como na Subseção 3.2.1, descartamos as respectivas amostras da análise. Os resultados obtidos são apresentados na Tabela 3.3 e não foram encontrados na literatura. Já os gráficos de quantis para a estatística de teste da razão de verossimilhanças são apresentados na Figura 3.5.

Tabela 3.3: Taxas de rejeição (%) das hipóteses H_0, $H_{1(1)}$, $H_{1(2)}$ e $H_{1(3)}$ - teste RV - proporcionalidade das variâncias - $x_i \sim U(0,5; 2,5)$ com $n = 21$ para um nível de significância $\alpha = 0,05$.

Número de réplicas s_i	Número de réplicas r_i	Sob H_0	Sob $H_{1(1)}$	Sob $H_{1(2)}$	Sob $H_{1(3)}$
mínimo: 2 médio: 2 máximo: 2	mínimo: 2 médio: 2 máximo: 2	99,60	99,52	99,84	99,84
mínimo: 2 médio: 7,4 máximo: 18	mínimo: 2 médio: 7,5 máximo: 18	82,22	97,22	99,88	99,46
mínimo: 4 médio: 14,8 máximo: 36	mínimo: 4 médio: 15,0 máximo: 36	32,00	99,76	100,00	100,00
mínimo: 18 médio: 18 máximo: 18	mínimo: 18 médio: 18 máximo: 18	11,70	99,94	100,00	100,00
mínimo: 40 médio: 40 máximo: 40	mínimo: 40 médio: 40 máximo: 40	7,00	99,96	100,00	100,00
mínimo: 80 médio: 80 máximo: 80	mínimo: 80 médio: 80 máximo: 80	5,40	100,00	100,00	100,00

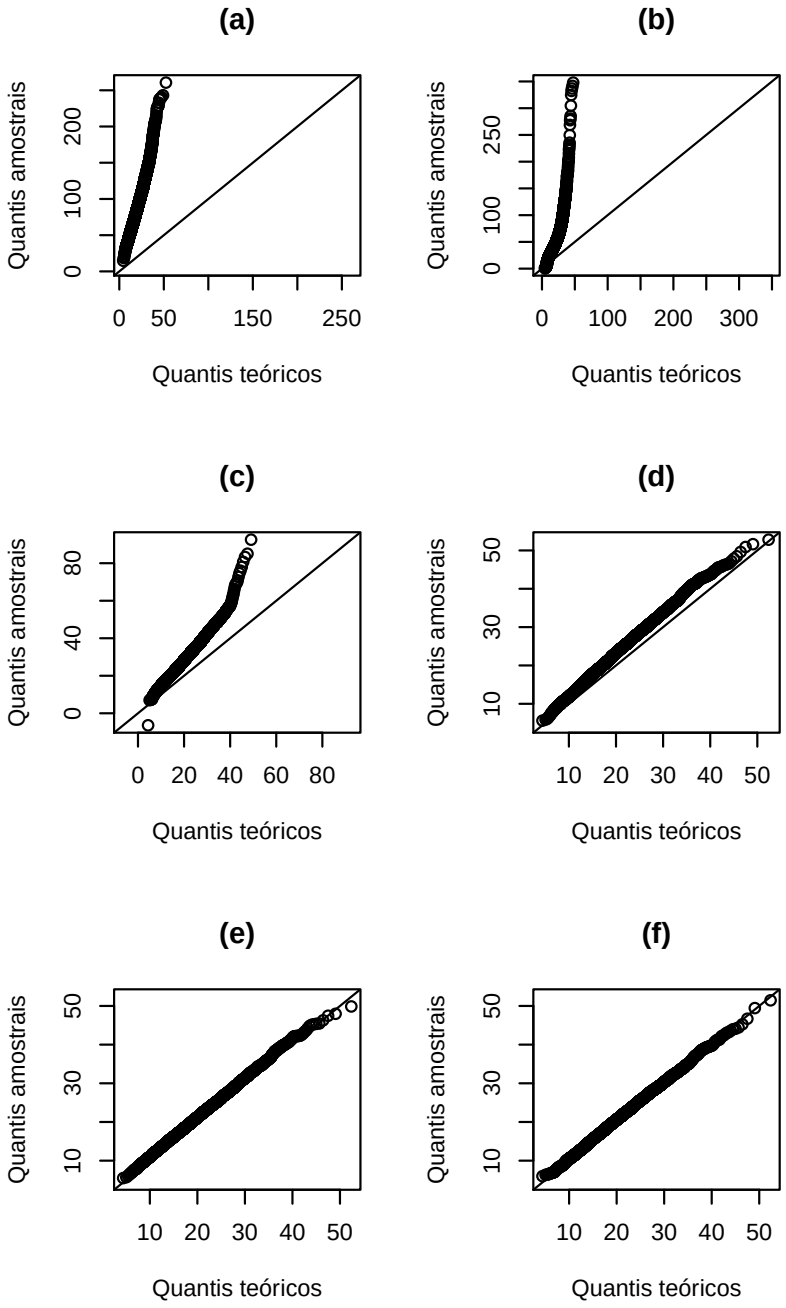

Figura 3.5: Gráficos de quantis da distribuição χ^2_{2n-2} - teste RV - proporcionalidade (a) $s_i = r_i = 2$, (b) s_i, r_i entre 2 e 18, (c) s_i, r_i entre 4 e 36, (d) $s_i = r_i = 18$, (e) $s_i = r_i = 40$ e (f) $s_i = r_i = 80$ - $x_i \sim U(0,5; 2,5)$ com $n = 21$ para um nível de significância $\alpha = 0,05$.

Assim como visto na Subseção 3.2.1, observamos na Figura 3.5 e na Tabela 3.3 que quando dispomos de uma quantidade mínima de réplicas ($s_i = r_i = 2$) ou então

dispomos de uma quantidade de réplicas entre 2 e 18, semelhante à condição do conjunto de dados com o qual trabalharemos no Capítulo 4, a estatística da razão de verossimilhanças para testar a proporcionalidade das variâncias novamente não aparenta ter distribuição χ^2_{n-1}. Além disso, os valores da taxa de rejeição sob H_0 são altos e distantes do nível nominal adotado, por exemplo, $\alpha = 5\%$. Todavia, quando fixamos o tamanho da amostra e aumentamos a quantidade de réplicas, percebemos que a taxa de rejeição sob H_0 diminui tendendo ao nível nominal. Para quantidade de réplicas igual a 80, a taxa de rejeição sob H_0 é semelhante a $\alpha = 5\%$, tornando-se o teste da razão de verossimilhanças satisfatório. Para as situações (número de réplicas) em que a taxa de rejeição sob H_0 foi alta, podemos utilizar alternativamente o teste de hipóteses *bootstrap* apresentado na Seção 2.2 a fim de testar as Hipóteses 2.

Em relação à taxa de rejeição sob H_1, nos três casos analisados os valores obtidos são altos e próximos de 100%.

Logo, no cenário descrito na introdução do Capítulo 3, concluímos que o teste RV para testar as Hipóteses 2 produz resultados satisfatórios para quantidade de réplicas igual a 80, sendo neste caso, apropriado. Para as demais situações analisadas, podemos utilizar alternativamente o teste de hipóteses *bootstrap* dado na Seção 2.2.

3.2.3 Viés aditivo e multiplicativo

Por fim, analisamos o comportamento da estatística de Wald para testar as Hipóteses 3, no cenário considerado. Como visto na Seção 2.2, testar as Hipóteses 3 é o grande interesse em problemas de comparação de métodos de medição já que por meio deste teste verificamos vícios aditivos e multiplicativos de um método em relação ao outro método analisado. Levamos em conta nesta análise apenas as estimativas $(\widehat{\beta}_0, \widehat{\beta}_1)$ obtidas por meio dos métodos de máxima verossimilhança e SIMEX uma vez que nesta dissertação, as distribuições assintóticas dos estimadores $(\widehat{\beta}_0, \widehat{\beta}_1)$ considerando os métodos

dos momentos 1 (MM1) e 2 (MM2) não foram obtidas, impossibilitando assim, utilizar resultados assintóticos na estatística de Wald.

Como já mencionado no início deste capítulo, as variâncias $\sigma^2_{e_i}$ são geradas de acordo com os valores de β_0 e β_1. Desta forma, considerando as Hipóteses 3, adotamos que sob H_0, $\beta_0 = 0$ e $\beta_1 = 1$ e então $\sigma^2_{e_i} = (0,5 \times y_i + 0,29)^2 = (0,5 \times (0 + 1x_i) + 0,29)^2$. Já sob H_1, analisamos duas situações: $H_{1(1)} : (\beta_0, \beta_1)^\top = (0,3; 0; 8)^\top$ e $H_{1(2)} : (\beta_0, \beta_1)^\top = (0,5; 2)^\top$. Os resultados obtidos são apresentados na Tabela 3.4 e não foram encontrados na literatura. Para complementar os resultados, apresentamos os gráficos de quantis para a estatística de Wald. Tais gráficos são apresentados nas Figuras 3.6 e 3.7.

Tabela 3.4: Taxas de rejeição (%) das hipóteses H_0, $H_{1(1)}$ e $H_{1(2)}$ - Teste de Wald - $x_i \sim U(2, 40)$ com $n = 20$ para um nível de significância $\alpha = 0,05$.

Número de réplicas s_i	Número de réplicas r_i	Método de estimação	Sob H_0	Sob $H_{1(1)}$	Sob $H_{1(2)}$
mínimo: 2	mínimo: 2	MV	80,16	75,80	95,25
médio: 2	médio: 2	SIMEX	91,04	78,93	97,32
máximo: 2	máximo: 2				
mínimo: 2	mínimo: 2	MV	21,30	78,52	97,50
médio: 7,4	médio: 7,5	SIMEX	53,16	79,09	98,90
máximo: 18	máximo: 18				
mínimo: 4	mínimo: 4	MV	9,74	82,09	99,89
médio: 14,8	médio: 15,0	SIMEX	33,42	86,13	99,80
máximo: 36	máximo: 36				
mínimo: 18	mínimo: 18	MV	8,70	80,77	100,00
médio: 18	médio: 18	SIMEX	48,48	86,35	100,00
máximo: 18	máximo: 18				
mínimo: 40	mínimo: 40	MV	6,70	87,64	100,00
médio: 40	médio: 40	SIMEX	26,61	95,82	100,00
máximo: 40	máximo: 40				
mínimo: 80	mínimo: 80	MV	5,62	95,91	100,00
médio: 80	médio: 80	SIMEX	–	–	–
máximo: 80	máximo: 80				

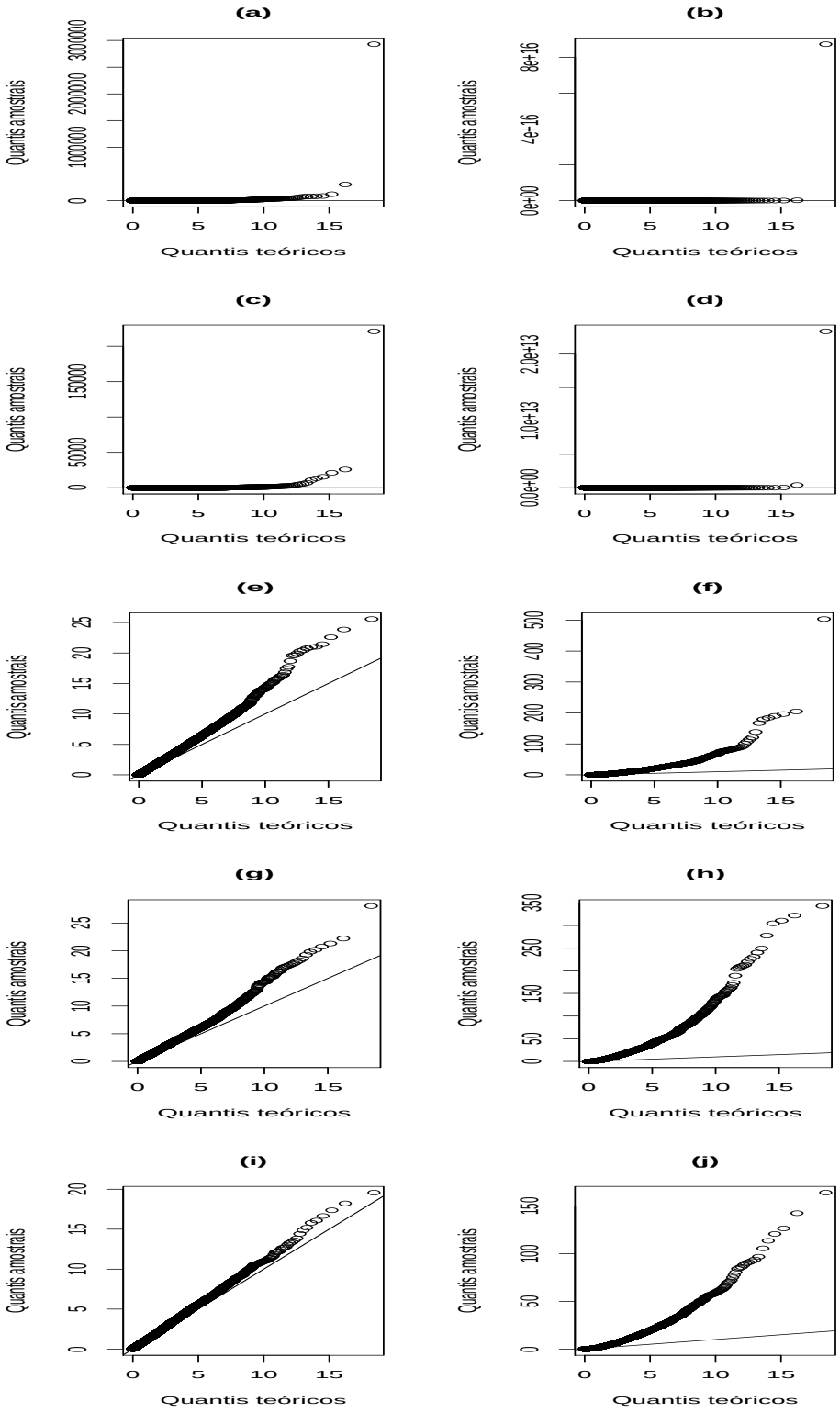

Figura 3.6: Gráficos de quantis da distribuição χ^2_2 - teste de Wald (a), (c), (e), (g) e (i) MV com $s_i = r_i = 2$; s_i, r_i entre 2 e 18; s_i, r_i entre 4 e 36; $s_i = r_i = 18$ e $s_i = r_i = 40$, respectivamente e (b), (d), (f), (h) e (j) SIMEX com $s_i = r_i = 2$; s_i, r_i entre 2 e 18; s_i, r_i entre 4 e 36; $s_i = r_i = 18$ e $s_i = r_i = 40$, respectivamente.

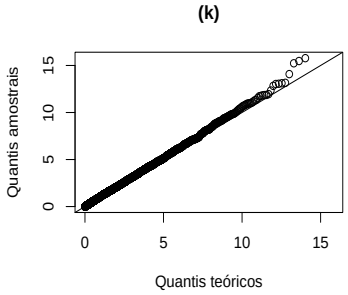

Figura 3.7: Gráfico de quantis da distribuição χ^2_2 - teste de Wald (k) MV com $s_i = r_i = 80$.

Analisando a Tabela 3.4, verificamos que para número de réplicas mínimo, isto é, $s_i = r_i = 2$, as taxas de rejeição sob H_0 tanto para o método de máxima verossimilhança quanto para o método SIMEX são altas. Além disso, notamos pelas Figuras 3.6 (a) e (b) que não há indício de que a estatística de Wald tenha distribuição χ^2_2. Desta forma, podemos utilizar alternativamente o teste de hipóteses *bootstrap* detalhado na Seção 2.2. Para quantidade de réplicas entre 2 e 18, condição semelhante ao conjunto de dados analisado no Capítulo 4, os valores da taxa de rejeição sob H_0 são menores do que os valores encontrados no caso anterior. Entretanto, as taxas ainda são altas, sendo que o valor obtido considerando o método de máxima verossimilhança foi menor do que o valor obtido considerando o método SIMEX. Novamente, uma alternativa seria utilizar o teste de hipóteses *bootstrap* como dado na Seção 2.2.

Verificamos que a taxa de rejeição sob H_0 diminui à medida que aumentamos a quantidade de réplicas, como esperado. Isto também é observado pelos gráficos de quantis dados na Figura 3.6 em que os quantis amostrais e teóricos se aproximam conforme aumentamos o número de réplicas. Além disso, observamos que para o método de máxima verossimilhança é necessário uma menor quantidade de réplicas, em relação ao método SIMEX, para que o valor da taxa de rejeição sob H_0 se aproxime do nível nominal $\alpha = 5\%$, isto é, para que os resultados do teste de Wald sejam satisfatórios. Isto pode ser notado para quantidade de réplicas igual a 40 em que o valor da taxa de

rejeição sob H_0 considerando o método de máxima verossimilhança é próximo de $6,70$ enquanto que considerando o método SIMEX fica em torno de $26,61\%$.

Pela Tabela 3.4 e pelas Figuras 3.6 e 3.7 verificamos que para quantidade de réplicas igual a 80, o valor da taxa de rejeição sob H_0 é próximo do nível nominal $\alpha = 5\%$. Não encontramos o valor da taxa de rejeição sob H_0, na mesma condição, considerando o método SIMEX por causa do alto custo computacional. Desta forma, concluímos que o teste de Wald considerando o método de máxima verossimilhança se mostrou mais eficiente do que considerando o método SIMEX, em termos de valor de taxa de rejeição sob H_0, para as quantidades de réplicas analisadas nesta dissertação. O teste de Wald utilizando o método SIMEX se mostrou insatisfatório para todas as situações analisadas.

Em relação ao poder do teste, vimos que tanto a taxa de rejeição sob $H_{1(1)}$ quanto a taxa de rejeição sob $H_{1(2)}$ aumenta à medida que aumentamos o número de réplicas, como esperado. Notamos também que os valores da taxa de rejeição sob $H_{1(1)}$ são menores do que os encontrados sob $H_{1(2)}$. Isto ocorreu uma vez que os valores para os parâmetros β_0 e β_1 em $H_{1(1)}$ são mais próximos dos verdadeiros valores do que os valores em $H_{1(2)}$, sendo assim o teste mais propício a erros. Apesar de os valores da taxa de rejeição sob H_1 considerando ambos os métodos serem semelhantes em cada caso analisado, os valores encontrados para o método SIMEX, em geral foi ligeiramente maior.

Portanto, concluímos que para quantidade de réplicas inferior a 80, o teste de Wald considerando os métodos de máxima verossimilhança e SIMEX não é satisfatório no cenário analisado. Para quantidade de réplicas igual a 80, o teste de Wald considerando apenas o método de máxima verossimilhança se mostrou satisfatório. Como na prática dispomos de uma quantidade inferior a 80, podemos utilizar simultaneamente o teste de hipóteses *bootstrap* dado na Seção 2.2, complementando os resultados assintóticos.

Capítulo 4

Aplicação

É comum aplicar técnicas de regressão em problemas de comparação de métodos analíticos. Como visto no Capítulo 1, nestes problemas o principal objetivo é detectar possíveis vícios aditivo e multiplicativo de um método em relação ao outro, o que é feito testando-se as Hipóteses 3 dadas na Seção 2.2. Desta forma, neste capítulo ajustamos o modelo (2.1) − (2.3) apresentado no Capítulo 2 a um conjunto de dados relacionado a problemas de comparação de métodos de medição, aplicando os métodos de estimação desenvolvidos nas Seções 2.1.1, 2.1.2 e 2.1.3 e realizando o teste de Wald proposto na Seção 2.2 a fim de verificar vícios aditivo e multiplicativo de um método em relação ao outro. Os testes da razão de verossimilhanças para testar as Hipóteses 1 e 2 (homoscedasticidade e proporcionalidade das variâncias, respectivamente), apesar de não serem os principais objetivos em problemas de comparação de métodos analíticos, também foram aplicados. O ajuste do modelo foi verificado por meio de gráficos de envelopes (Atkinson, 1985).

O conjunto de dados aqui descrito foi utilizado por Rasekh & Fieller (2003). Em uma extensa pesquisa arqueológica da produção e distribuição de cerâmica, medições do conteúdo do elemento químico (elemento mineral) potássio (K), presente em algumas

cerâmicas, foram realizadas utilizando duas diferentes técnicas 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 cerâmica possui um código de construção que ajuda na identificação.

As cerâmicas com o mesmo código de construção e da mesma proveniência são consideradas repetições. Desta forma, o conjunto de dados foi dividido em 21 grupos sendo que em cada grupo o número de cerâmicas (réplicas) varia de 2 a 18. Adotamos Y_{ij} como sendo o j-ésimo valor observado do i-ésimo grupo, utilizando a técnica NAA e X_{ik} como sendo o k-ésimo valor observado do i-ésimo grupo, utilizando a técnica ICP. Note que no conjunto de dados em questão, $i = 1, \ldots, 21$, $k = 1, \ldots, s_i$ e $j = 1 \ldots, r_i$.

As médias e os desvios padrão das medições do teor de K em cada um dos 21 grupos, considerando os métodos NAA e ICP, podem ser observados na Figura 4.1. Há indícios de que em média, os valores do teor de K obtidos pela técnica NAA são semelhantes aos valores obtidos pela técnica ICP. Além disso, pela Figura 4.1 nota-se também que os desvios padrão assumem valores distintos indicando uma possível heteroscedasticidade das variâncias. Esta hipótese será testada na sequência.

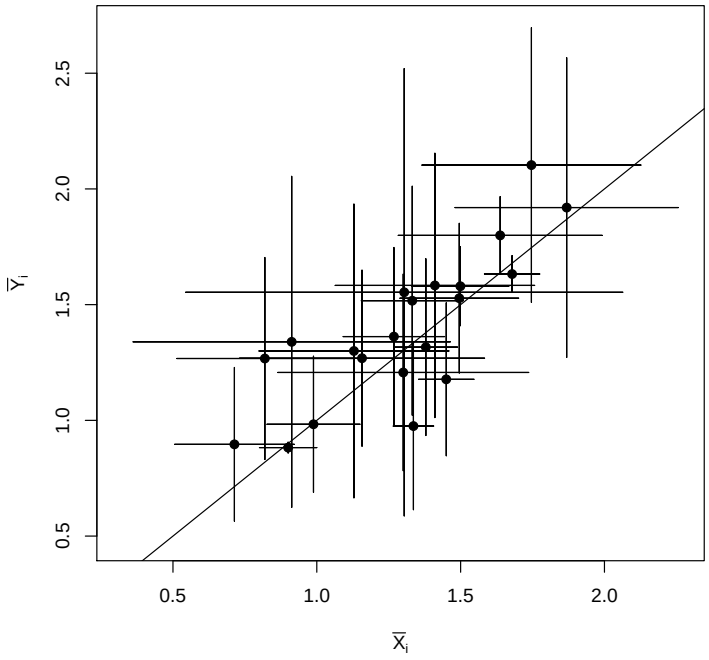

Figura 4.1: Médias de cada grupo i, $i = 1 \ldots, 21$, considerando os métodos NAA ($\overline{Y}_i$) e ICP ($\overline{X}_i$) e seus respectivos desvios padrão com a reta identidade.

Aplicamos o teste da razão de verossimilhanças com o objetivo de testar as Hipóteses 1 e 2 apresentadas na Seção 2.2. Simultaneamente, utilizamos o teste de hipóteses *bootstrap* também mostrado na Seção 2.2. Os valores da estatística da razão de verossimilhanças para as Hipóteses 1 e 2, seus respectivos valores p, utilizando os resultados assintóticos dados na Seção 2.2 e os valores p empíricos considerando uma amostra *bootstrap* de tamanho $Q = 10000$, são apresentados na Tabela 4.1.

Tabela 4.1: Valores da estatística da razão de verossimilhanças para as Hipóteses 1 e 2, respectivos valores p e valores p empíricos - Teor de K nas cerâmicas egípcia.

Hipóteses 1			Hipóteses 2		
Estatística	Valor p	Valor p empírico	Estatística	Valor p	Valor p empírico
$156,469$	$< 0,001$	$< 0,001$	$38,799$	$0,007$	$0,589$

Em relação às Hipóteses 1, observamos pela Tabela 4.1 que tanto o valor p quanto o valor p empírico foram inferiores ao nível de significância adotado $\alpha = 5\%$, nos levando

à rejeição da hipótese nula, ou seja, a rejeição da homoscedasticidade das variâncias. Já em relação às Hipóteses 2, o valor p considerando os resultados assintóticos nos leva à rejeição da hipótese nula enquanto que o valor p empírico nos conduz à não rejeição de H_0. É importante destacar que no estudo de simulação realizado no Capítulo 3 nas mesmas condições dos dados reais, os resultados do teste da razão de verossimilhanças para testar as Hipóteses 1 e 2 não foram satisfatórios. Desta forma, consideramos o valor p empírico e portanto, concluímos que as variâncias são proporcionais e consequentemente, heteroscedásticas, como suspeitado na Figura 4.1.

Desta forma, ajustamos o modelo (2.1) − (2.3) ao conjunto de dados referente ao teor de K nas cerâmicas egípcias aplicando o método de máxima verossimilhança, método dos momentos 1 e o método SIMEX e então verificamos possíveis vícios aditivo e multiplicativo da técnica NAA em relação à técnica ICP, o grande interesse neste caso. No caso da estimação pelo método de máxima verossimilhança, o algoritmo foi encerrado quando o valor de $\epsilon^{(a)}$, dado por (2.11), foi menor do que 10^{-3}. No método dos momentos 1, utilizamos o método *bootstrap* não paramétrico para a obtenção da matriz de covariâncias dos estimadores $(\widehat{\beta}_{0_{mom_1}}, \widehat{\beta}_{1_{mom_1}})^\top$, adotando $Q = 10000$. Na estimação SIMEX, na etapa de simulação adotamos $B = 200$ e $M = 10$ com $\lambda \in (0, 2]$ e na etapa de extrapolação consideramos o modelo linear simples, como no Capítulo 3.

O ajuste do modelo (2.1) − (2.3), considerando o método de máxima verossimilhança, pode ser verificado por meio do gráfico de envelopes simulados. A técnica foi desenvolvida por Atkinson (1985). Para utilizar esta ferramenta gráfica, inicialmente encontramos as estimativas de máxima verossimilhança considerando os dados reais e então, calculamos

$$\delta_i = \sum_{j=1}^{r_i} \frac{Y_{ij} - (\beta_0 + \beta_1 x_i)}{\sigma_{e_i}} + \sum_{k=1}^{s_i} \frac{X_{ik} - x_i}{\sigma_{u_i}}, \quad i = 1 \ldots, n, \tag{4.1}$$

com $\delta_i \overset{\text{iid}}{\sim} N(0, 1)$, substituindo os parâmetros pelas suas respectivas estimativas. Desta

forma, obtemos n valores $\widehat{\delta}_i$. Ordenamos esses valores obtendo $\widehat{\delta}_{(1)} \leq \cdots \leq \widehat{\delta}_{(n)}$ e então, representamos os pontos $(\Phi^{-1}((i-3/8)/(n+1/4)), \widehat{\delta}_{(i)})$ em um gráfico, em que Φ^{-1} é a função quantil da distribuição normal padrão. Já para obter os limites do envelope, simulamos $G = 1000$ 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 $\widehat{\delta}_{g\ i}$, $i = 1, \ldots, n$ substituindo os parâmetros da expressão dada em (4.1) pelas estimativas encontradas. Ordenamos os valores encontrados e por fim, representamos os pontos $(\Phi^{-1}((i-3/8)/(n+1/4)), \min_{g=1}^{G} \widehat{\delta}_{g(i)})$ e $(\Phi^{-1}((i-3/8)/(n+1/4)), \max_{g=1}^{G} \widehat{\delta}_{g(i)})$ para $i = 1, \ldots, n$, representando os limites inferior e superior do envelope. Também desenhamos no gráfico os pontos $(\Phi^{-1}((i-3/8)/(n+1/4)), \sum_{g=1}^{G} \widehat{\delta}_{g(i)}/G)$. A técnica é aplicada ao conjunto de dados em questão, como visto na Figura 4.2.

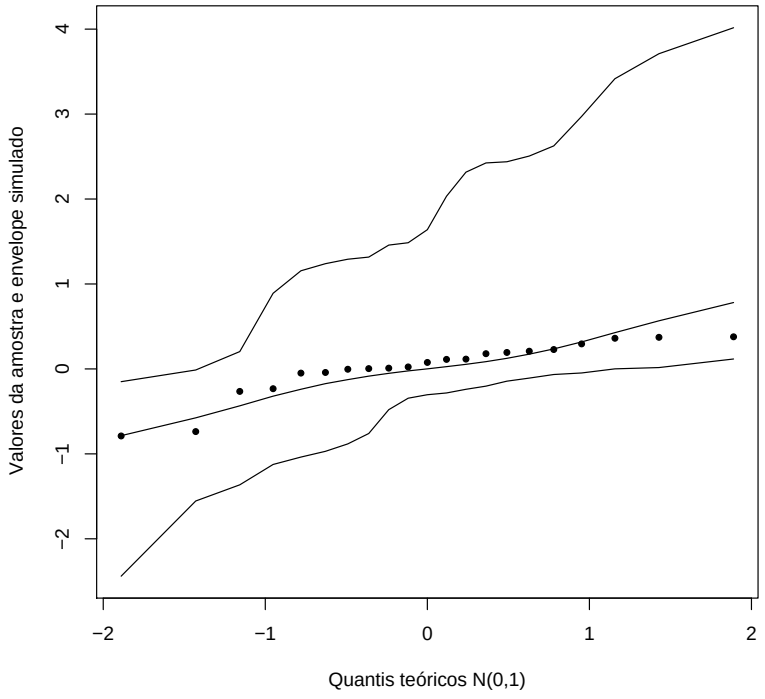

Figura 4.2: Gráficos de quantis da distribuição $N(0,1)$ e envelope simulado - método de máxima verossimilhança - ajuste do modelo $(2.1) - (2.3)$ - teor de K nas cerâmicas egípcia.

Como visto na Figura 4.2, o modelo (2.1) − (2.3) se ajustou bem ao conjunto de dados do teor de K nas cerâmicas egípcias. Não construímos gráficos de envelope para os métodos dos momentos 1 e SIMEX.

Na Tabela 4.2, são apresentadas as estimativas dos parâmetros β_0 e β_1 calculadas pelos três métodos de estimação e os respectivos erros padrão estimados.

Tabela 4.2: Estimativas dos parâmetros β_0 e β_1 obtidas pelos métodos de máxima verossimilhança, momentos 1 e SIMEX e respectivos erros padrão estimados - Teor de K nas cerâmicas egípcia.

Parâmetro	Método de estimação	Estimativa	Erro padrão
β_0	MV	$-0,01420$	$0,11804$
	MM1	$0,12645$	$0,22621$
	SIMEX	$-0,07079$	$0,05034$
β_1	MV	$1,02989$	$0,09059$
	MM1	$0,97105$	$0,16350$
	SIMEX	$1,06579$	$0,06645$

Analisando a Tabela 4.2, observamos que os valores do erro padrão estimado das estimativas dos parâmetros β_0 e β_1 obtidas pelo método SIMEX foram os menores entre todos os métodos abordados. Já os valores obtidos pelo método dos momentos 1 foram os maiores. Não podemos esquecer que a matriz de covariâncias dos estimadores $\widehat{\beta}_{0_{mom_1}}$, $\widehat{\beta}_{1_{mom_1}}$ são obtidas pela técnica *bootstrap*, enquanto que para os demais métodos utilizamos as matrizes de covariâncias assintóticas, o que pode explicar a diferença nos valores encontrados.

Na Tabela 4.3 apresentamos os intervalos de confiança *bootstrap* das estimativas $\widehat{\beta}_0$ e $\widehat{\beta}_1$, obtidas pelo método de máxima verossimilhança e pelo método dos momentos 1. Não calculamos o intervalo de confiança *bootstrap* para as estimativas SIMEX devido ao alto custo computacional. Para o cálculo do intervalo de confiança *bootstrap*, consideramos 5000 amostras *bootstrap*. Os intervalos foram obtidos utilizando os quantis 2,5% e 97,5% da amostra das estimativas *bootstrap* encontradas.

Tabela 4.3: Intervalos de confiança *bootstrap* das estimativas $\widehat{\beta}_0$ e $\widehat{\beta}_1$ obtidas pelos métodos de máxima verossimilhança e momentos 1 - Teor de K nas cerâmicas egípcia.

Parâmetro	Método de estimação	Limite inferior	Limite superior
	MV	$-0,38892$	$0,24745$
β_0	MM1	$-0,60100$	$0,66919$
	MV	$0,82469$	$1,31753$
β_1	MM1	$0,57847$	$1,50238$

Como já dito anteriormente, nosso principal interesse nesta dissertação é verificar possíveis vícios aditivo e multiplicativo de um método em relação ao outro, testando as Hipóteses 3 dadas na Seção 2.2. Desta forma, os valores da estatística de Wald considerando os métodos de estimação máxima verossimilhança e SIMEX e os respectivos valores p são apresentados na Tabela 4.4. Além disso, também apresentamos o valor p empírico obtido por meio do teste de hipóteses *bootstrap* paramétrico, considerando o método de máxima verossimilhança. O valor da estatística de Wald e o respectivo valor p para o método dos momentos 1 não foram calculados uma vez que não conhecemos a distribuição de $(\widehat{\beta}_{0_{mom_1}}, \widehat{\beta}_{1_{mom_1}})^\top$, sendo neste caso o uso de resultados assintóticos para a estatística de Wald inadequado, como visto na Seção 2.2. Além disso, também não obtemos os valores p empíricos considerando os métodos dos momentos 1 e SIMEX já que é inviável por causa do tempo de processamento.

Tabela 4.4: Estatística de Wald e respectivos valores p - métodos de máxima verossimilhança e SIMEX e valor p empírico - método de máxima verossimilhança - Hipóteses 3 (viés aditivo e multiplicativo) - Teor de K nas cerâmicas egípcias.

Método de Estimação	Estatística de Wald	Valor p	Valor p empírico
MV	$0,753$	$0,686$	$0,888$
SIMEX	$3,939$	$0,140$	-

Analisando a Tabela 4.4, notamos que para ambos os métodos o teste de Wald não nos leva à rejeição da hipótese nula uma vez que tanto o valor p quanto o valor p empírico são maiores do que o nível de significância adotado $\alpha = 5\%$. Desta forma,

concluímos que com um nível de confiança de 95%, as técnicas NAA e ICP não se diferem de maneira significativa para medições do teor de K, mostrando que neste caso não há vícios aditivo e multiplicativo da técnica NAA em relação à técnica ICP e portanto, que em geral as técnicas produzem medições semelhantes.

Capítulo 5

Conclusão

Nesta dissertação apresentamos o modelo funcional heteroscedástico com observações replicadas em que as réplicas são inseridas na estrutura do modelo. Esse modelo é bastante utilizado principalmente em problemas de comparação de métodos de medição, foco de nosso estudo. Nestes casos, o principal interesse está em verificar possíveis vícios aditivo e multiplicativo de um método em relação ao outro.

Utilizamos três métodos no processo de estimação dos parâmetros de interesse: o método de máxima verossimilhança, tradicional e com importantes propriedades; o método dos momentos, simples e eficiente e o método SIMEX, promissor na redução do viés. Considerando a situação proposta nesta dissertação, em que o tamanho da amostra é fixo e as quantidades de réplicas das variáveis resposta e explicativa aumentam com o aumento do número de observações, verificamos que em geral, os métodos propostos apresentam bons resultados em questão de viés e $\sqrt{\text{EQM}}$, tendo o método de máxima verossimilhança se destacado. O método SIMEX como esperado, reduziu o viés mas não o eliminou totalmente, sendo os valores do viés simulado encontrados nesse método, em algumas situações, próximos dos obtidos pelo método de máxima verossimilhança e pelo método dos momentos 1.

Em relação aos testes propostos na Seção 2.2, tanto o teste da razão de verossimilhanças quanto o teste de Wald se mostraram insatisfatórios para as situações analisadas em que a quantidade de réplicas disponível é inferior a 80. Para o teste de Wald considerando o método SIMEX, os resultados não foram satisfatórios em nenhum dos casos tratados no Capítulo 3, sendo as taxas de rejeição sob H_0 obtidas bem acima do nível nominal $\alpha = 5\%$. Já considerando o método de máxima verossimilhança, o teste de Wald se mostrou satisfatório para quantidade de réplicas igual a 80.

Como propostas de trabalhos futuros destacamos:

1. Estender o modelo (2.1) − (2.3) para as situações em que os erros são correlacionados;

2. Encontrar a matriz de covariâncias assintótica dos estimadores dos parâmetros $(\beta_0, \beta_1)^\top$ obtidos por meio do método dos momentos;

3. Incluir estatísticas de teste tais como razão de verossimilhanças e de escore para testar as Hipóteses 3 dadas na Seção 2.2;

4. Considerar em (2.1) erro na equação e então, aplicar os métodos descritos no Capítulo 2 a esse caso;

5. Estudar influência local e

6. Estudar modelos com erros de medição com distribuições diferentes da distribuição normal.

Apêndice A

Estimadores de Máxima Verossimilhança

Apresentamos as derivadas de 1ª e 2ª ordem da função log-verossimilhança, considerando o modelo (2.1) − (2.3) dado no Capítulo 2. Essas derivadas são necessárias para a obtenção dos estimadores de máxima verossimilhança e da respectiva matriz de covariâncias. O procedimento matemático utilizado na obtenção da matriz de covariâncias dos estimadores de máxima verossimilhança também é apresentado. Considerando o modelo (2.1) − (2.3) temos que a função log-verossimilhança para uma amostra de tamanho n é dada por

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log L(\boldsymbol{\theta}) = const - \frac{1}{2}\sum_{i=1}^{n} r_i \log \sigma^2_{e_i} - \frac{1}{2}\sum_{i=1}^{n} s_i \log \sigma^2_{u_i} \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma^2_{e_i}}\sum_{j=1}^{r_i}[Y_{ij} - (\beta_0 + \beta_1 x_i)]^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma^2_{u_i}}\sum_{k=1}^{s_i}(X_{ik} - x_i)^2. \end{aligned}$$

As derivadas de primeira ordem da função log-verossimilhança com respeito aos

parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\sigma}_u^{2^\top}, \boldsymbol{\sigma}_e^{2^\top}, \boldsymbol{x}^\top)^\top$ são

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_0} = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{Y_{ij} - (\beta_0 + \beta_1 x_i)}{\sigma_{e_i}^2}, \tag{A.1}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_1} = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{[Y_{ij} - (\beta_0 + \beta_1 x_i)]x_i}{\sigma_{e_i}^2}, \tag{A.2}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial x_i} = \sum_{j=1}^{r_i} \frac{[Y_{ij} - (\beta_0 + \beta_1 x_i)]\beta_1}{\sigma_{e_i}^2} + \sum_{k=1}^{s_i} \frac{(X_{ik} - x_i)}{\sigma_{u_i}^2}, \tag{A.3}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma_{u_i}^2} = -\frac{1}{2} s_i \frac{1}{\sigma_{u_i}^2} + \frac{1}{2} \sum_{k=1}^{s_i} \frac{(X_{ik} - x_i)^2}{(\sigma_{u_i}^2)^2} \tag{A.4}$$

$$\text{e } \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma_{e_i}^2} = -\frac{1}{2} r_i \frac{1}{\sigma_{e_i}^2} + \frac{1}{2} \sum_{j=1}^{r_i} \frac{[Y_{ij} - (\beta_0 + \beta_1 x_i)]^2}{(\sigma_{e_i}^2)^2}. \tag{A.5}$$

Já as derivadas de segunda ordem da função log-verossimilhança com respeito aos parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1, \boldsymbol{\sigma}_u^{2^\top}, \boldsymbol{\sigma}_e^{2^\top}, \boldsymbol{x}^\top)^\top$ são

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_0^2} = -\sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{1}{\sigma_{e_i}^2} = -\sum_{i=1}^{n} \frac{r_i}{\sigma_{e_i}^2}, \tag{A.6}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_1} = -\sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{x_i}{\sigma_{e_i}^2} = -\sum_{i=1}^{n} \frac{r_i x_i}{\sigma_{e_i}^2}, \tag{A.7}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_1^2} = -\sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{x_i^2}{\sigma_{e_i}^2} = -\sum_{i=1}^{n} \frac{r_i x_i^2}{\sigma_{e_i}^2}, \tag{A.8}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial x_i^2} = -\sum_{j=1}^{r_i} \frac{\beta_1^2}{\sigma_{e_i}^2} - \sum_{k=1}^{s_i} \frac{1}{\sigma_{u_i}^2} = -\frac{r_i \beta_1^2}{\sigma_{e_i}^2} - \frac{s_i}{\sigma_{u_i}^2}, \tag{A.9}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_0 \partial x_i} = -\sum_{j=1}^{r_i} \frac{\beta_1}{\sigma_{e_i}^2} = -\frac{r_i \beta_1}{\sigma_{e_i}^2}, \tag{A.10}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_1 \partial x_i} = \sum_{j=1}^{r_i} \frac{Y_{ij}}{\sigma_{e_i}^2} - \sum_{j=1}^{r_i} \frac{\beta_0}{\sigma_{e_i}^2} - \sum_{j=1}^{r_i} \frac{\beta_1 2 x_i}{\sigma_{e_i}^2} = \frac{r_i}{\sigma_{e_i}^2} (\overline{Y}_i - \beta_0 - 2\beta_1 x_i), \tag{A.11}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_0 \partial \sigma^2_{u_i}} = 0, \tag{A.12}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_0 \partial \sigma^2_{e_i}} = -\sum_{j=1}^{r_i} \frac{Y_{ij} - (\beta_0 + \beta_1 x_i)}{(\sigma^2_{e_i})^2} = -\frac{r_i(\overline{Y}_i - \beta_0 - \beta_1 x_i)}{(\sigma^2_{e_i})^2}, \tag{A.13}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_1 \partial \sigma^2_{u_i}} = 0, \tag{A.14}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_1 \partial \sigma^2_{e_i}} = -\sum_{j=1}^{r_i} \frac{[Y_{ij} - (\beta_0 + \beta_1 x_i)]x_i}{(\sigma^2_{e_i})^2} = -\frac{r_i(\overline{Y}_i - \beta_0 - \beta_1 x_i)x_i}{(\sigma^2_{e_i})^2}, \tag{A.15}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial (\sigma^2_{u_i})^2} = \frac{s_i}{2}\frac{1}{(\sigma^2_{u_i})^2} - \frac{\sum_{k=1}^{s_i}(X_{ik} - x_i)^2}{(\sigma^2_{u_i})^3}, \tag{A.16}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2_{u_i} \partial x_i} = -\frac{\sum_{k=1}^{s_i}(X_{ik} - x_i)}{(\sigma^2_{u_i})^2} = -\frac{s_i(\overline{X}_i - x_i)}{(\sigma^2_{u_i})^2}, \tag{A.17}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial (\sigma^2_{e_i})^2} = \frac{r_i}{2}\frac{1}{(\sigma^2_{e_i})^2} - \frac{\sum_{j=1}^{r_i}[Y_{ij} - (\beta_0 + \beta_1 x_i)]^2}{(\sigma^2_{e_i})^3} \tag{A.18}$$

$$\text{e } \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2_{e_i} \partial x_i} = -\frac{\sum_{j=1}^{r_i}[Y_{ij} - (\beta_0 + \beta_1 x_i)]\beta_1}{(\sigma^2_{e_i})^2} = -\frac{r_i(\overline{Y}_i - \beta_0 - \beta_1 x_i)\beta_1}{(\sigma^2_{e_i})^2}. \tag{A.19}$$

A matriz de covariâncias assintótica dos estimadores pode ser obtida por meio da inversa da matriz de informação observada, usando resultados padrão de inversão de matrizes particionadas (Dolby *et al.*, 1987). A matriz de informação observada é dada por

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \mathbf{0}_{(2\times 2n)} & \boldsymbol{B} \\ \mathbf{0}_{(2n\times 2)} & \boldsymbol{C} & \mathbf{0}_{(2n\times n)} \\ \boldsymbol{B}^\top & \mathbf{0}_{(n\times 2n)} & \boldsymbol{D} \end{bmatrix}$$

em que

$$\boldsymbol{A} = \begin{bmatrix} \sum_{i=1}^{n} \frac{r_i}{\sigma^2_{e_i}} & \sum_{i=1}^{n} \frac{r_i x_i}{\sigma^2_{e_i}} \\ \sum_{i=1}^{n} \frac{r_i x_i}{\sigma^2_{e_i}} & \sum_{i=1}^{n} \frac{r_i x_i^2}{\sigma^2_{e_i}} \end{bmatrix}, \quad \boldsymbol{B} = \beta_1 \begin{bmatrix} \frac{r_1}{\sigma^2_{e_1}} & \cdots & \frac{r_n}{\sigma^2_{e_n}} \\ \frac{x_1 r_1}{\sigma^2_{e_1}} & \cdots & \frac{x_n r_n}{\sigma^2_{e_n}} \end{bmatrix} \text{ e}$$

$$\boldsymbol{D} = diag\left(\frac{\beta_1^2 r_1}{\sigma^2_{e_1}} + \frac{s_1}{\sigma^2_{u_1}}, \ldots, \frac{\beta_1^2 r_n}{\sigma^2_{e_n}} + \frac{s_n}{\sigma^2_{u_n}} \right).$$

Note que não mostraremos a matriz $\boldsymbol{C}$, que é o bloco correspondente aos parâmetros $(\sigma^2_{u_1}, \ldots, \sigma^2_{u_n}, \sigma^2_{e_1}, \ldots, \sigma^2_{e_n})$, uma vez que a matriz de covariâncias de $(\widehat{\beta}_0, \widehat{\beta}_1)$ não depende de $\boldsymbol{C}$.

Dividindo a matriz $\boldsymbol{M}$ em blocos, obtemos

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{M}_2 \\ \boldsymbol{M}_2^\top & \boldsymbol{M}_3 \end{bmatrix},$$

sendo que

$$\boldsymbol{M}_1 = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0}_{(2\times 2n)} \\ \boldsymbol{0}_{(2n\times 2)} & \boldsymbol{C} \end{bmatrix}, \quad \boldsymbol{M}_2^\top = \begin{bmatrix} \boldsymbol{B}^\top & \boldsymbol{0}_{(n\times 2n)} \end{bmatrix} \text{ e } \boldsymbol{M}_3 = \boldsymbol{D}.$$

Nosso interesse é na matriz de covariâncias de $(\widehat{\beta}_0, \widehat{\beta}_1)$. Assim, invertemos $\boldsymbol{M}$ e selecionamos a submatriz referente às duas primeiras linhas e duas primeiras colunas. Seja

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{M}_1^* & \boldsymbol{M}_2^* \\ \boldsymbol{M}_2^{*\top} & \boldsymbol{M}_3^* \end{bmatrix}$$

a inversa da matriz $\boldsymbol{M}$. Utilizando resultados de inversão de matrizes (Rao, 1973)

temos que

$$\boldsymbol{M}_1^* = (\boldsymbol{M}_1 - \boldsymbol{M}_2\boldsymbol{M}_3^{-1}\boldsymbol{M}_2^\top)^{-1} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^\top & \mathbf{0}_{2\times 2n} \\ \mathbf{0}_{2n\times 2} & \boldsymbol{C} \end{bmatrix}^{-1}.$$

Assim, obtemos

$$\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^\top = \begin{bmatrix} \sum_{i=1}^{n} \frac{r_i}{\sigma_{e_i}^2}\left(1 - \frac{\beta_1^2 r_i \sigma_{u_i}^2}{\sigma_{u_i}^2\beta_1^2 r_i + \sigma_{e_i}^2 s_i}\right) & \sum_{i=1}^{n} \frac{x_i r_i}{\sigma_{e_i}^2}\left(1 - \frac{\beta_1^2 r_i \sigma_{u_i}^2}{\sigma_{u_i}^2\beta_1^2 r_i + \sigma_{e_i}^2 s_i}\right) \\ & \sum_{i=1}^{n} \frac{x_i^2 r_i}{\sigma_{e_i}^2}\left(1 - \frac{\beta_1^2 r_i \sigma_{u_i}^2}{\sigma_{u_i}^2\beta_1^2 r_i + \sigma_{e_i}^2 s_i}\right) \end{bmatrix}.$$

Definindo $c_i = r_i\left(1 - \frac{\beta_1^2 r_i \sigma_{u_i}^2}{\sigma_{u_i}^2\beta_1^2 r_i + \sigma_{e_i}^2 s_i}\right)$, segue que a matriz de covariâncias assintótica de $(\widehat{\beta}_0, \widehat{\beta}_1)$ é dada por

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \begin{bmatrix} \sum_{i=1}^{n} \frac{c_i}{\sigma_{e_i}^2} & \sum_{i=1}^{n} \frac{x_i c_i}{\sigma_{e_i}^2} \\ & \sum_{i=1}^{n} \frac{x_i^2 c_i}{\sigma_{e_i}^2} \end{bmatrix}^{-1}.$$

Bibliografica

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. Series C*, **19**(2), 135–144.

Bertrand-Krajewski, J. L. (2004). Tss concentration in sewers estimated from turbidity measurements by means of linear regression accounting for uncertainties in both variables. *Water Science and Technology*, **50**(11), 81–88.

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

Chan, L. K. & Mak, T. K. (1979). Maximum likelihood estimation of a linear structural relationship with replication. *Journal of the Royal Statistical Society. Series B (Methodological)*, **41**(2), pp. 263–268.

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

Cook, J. & Stefanski, L. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, **89**(428), 1314–1328.

Davidson, R. & MacKinnon, J. (2000). Bootstrap tests: how many bootstraps? *Econometric Reviews*, **19**(1), 55–68.

de Castro, M. & Galea, M. (2010). Robust inference in an heteroscedastic measurement error model. *Journal of the Korean Statistical Society*, **39**(4), 439–447.

de Castro, M., Galea-Rojas, M., Bolfarine, H. & de Castilho, M. V. (2004). Detection of analytical bias when comparing two or more measuring devices. *Journal of Chemometrics*, **18**(2), 431–440.

de Castro, M., Galea-Rojas, M. & Bolfarine, H. (2007). Local influence assessment in heteroscedastic measurement error models. *Computational Statistics & Data Analysis*, **52**(2), 1132–1142.

Devanarayan, V. (1996). *Simulation extrapolation method for heteroscedastic measurement error models with replicate measurements*. Ph.D. thesis, North Carolina State University, Raleigh, North Carolina, USA.

Devanarayan, V. & Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics and Probability Letters*, **59**(4), 219–225.

Dolby, G. R. & Lipton, S. (1972). Maximum likelihood estimation of the general nonlinear functional relationship with replicated observations and correlated errors. *Biometrika*, **59**(1), 121–129.

Dolby, G. R., Cormack, R. M. & Sinclair, D. F. (1987). On fitting bivariate functional relationships to unpaired and unequally replicated data. *Biometrika*, **74**(2), 393–399.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC.

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

Galea-Rojas, M., de Castilho, M. V., Bolfarine, H. & de Castro, M. (2003). Detection of analytical bias. *The Analyst*, **128**(8), 1073–1081.

Kimura, D. K. (1992). Functional comparative calibration using an EM algorithm. *Biometrics*, **48**(4), 1263–1271.

Kukush, A. & Van Huffel, S. (2004). Consistency of elementwise-weighted total least squares estimator in a multivariate errors-in-variables model AX = B. *Metrika*, **59**(1), 75–97.

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 the observations. *Statistics in Medicine*, **21**(8), 1089–1101.

Kutner, M., Nachtsheim, C., Neter, J. & Li, W. (2005). *Applied linear statistical models*. McGraw-Hill Irwin Boston.

Mak, T. (1982). Estimation in the presence of incidental parameters. *Canadian Journal of Statistics*, **10**(2), 121–132.

Markovsky, I., Rastello, M.-L., Premoli, A., Kukush, A. & Huffel, S. V. (2006). The element-wise weighted total least squares problem. *Computational Statistics & Data Analysis*, **50**(1), 181–209.

Patefield, W. (1978). The unreplicated ultrastructural relation: large sample properties. *Biometrika*, **65**(3), 535.

Patriota, A. G. (2006). *Modelo funcional heterocedástico com erro nas variáveis: uma abordagem para medidas repetidas*. Dissertação de mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo-SP.

R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. (1973). *Linear statistical inference and its applications*. J. Wiley and Sons, New York, second edition.

Rasekh, A. R. & Fieller, N. R. J. (2003). Influence functions in functional measurement error models with replicated data. *Statistics*, **37**(2), 169–178. Cited By (since 1996): 1.

Ripley, B. D. & Thompson, M. (1987). Regression techniques for the detection of analytical bias. *The Analyst*, **112**(4), 377–383.

Riu, J. & Rius, F. X. (1996). Assessing the accuracy of analytical methods using linear regression with errors in both axes. *Analytical Chemistry*, **68**(11), 1851–1857.

Shalabh, Paudel, C. M. & Kumar, N. (2009). Consistent estimation of regression parameters under replicated ultrastructural model with non-normal errors. *Journal of Statistical Computation and Simulation*, **79**(3), 251–274.

Stefanski, L. A. (2000). Measurement error models. *Journal of the American Statistical Association*, **95**(452), 1353–1358.

Veenendaal, E. M., Mantlana, K. B., Pammenter, N. W., Weber, P., Huntsman-Mapila, P. & Lloyd, J. (2008). Growth form and seasonal variation in leaf gas exchange of Colophospermum mopane savanna trees in northwest Botswana. *Tree Physiology*, **28**(3), 417–424.

Wang, X., Fan, Z. & Wang, B. (2010). Estimating smooth distribution function in the presence of heteroscedastic measurement errors. *Computational Statistics & Data Analysis*, **54**(1), 25–36.