**Implementação no software estatístico R
de modelos de regressão normal
com parametrização geral**

André Casagrandi Perette

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Estatística

Orientador: Prof. Dr. Alexandre Galvão Patriota

São Paulo, agosto de 2019

Implementação no software estatístico R de modelos de regressão normal com parametrização geral

Esta dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho realizada em 23/08/2019. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Alexandre Galvão Patriota - IME/USP
- Prof. Dr. Mário de Castro Andrade Filho - ICMC/USP
- Prof^a. Dr^a. Tatiane Ferreira do Nascimento Melo da Silva - UFG

Resumo

Este trabalho objetiva o desenvolvimento de um pacote no *software* estatístico R com a implementação de estimadores em modelos de regressão normal univariados com parametrização geral, uma particularidade do modelo definido em Patriota e Lemonte (2011). Essa classe contempla uma ampla gama de modelos conhecidos, tais como modelos de regressão não lineares e heteroscedásticos. São implementadas correções nos estimadores de máxima verossimilhança e na estatística de razão de verossimilhanças. Tais correções são efetivas quando o tamanho amostral é pequeno. Para a correção do estimador de máxima verossimilhança, considerou-se a correção do viés de segunda ordem calculada em Patriota e Lemonte (2009), enquanto que para a estatística da razão de verossimilhanças aplicou-se a correção desenvolvida em Skovgaard (2001). Todas as funcionalidades do pacote são descritas detalhadamente neste trabalho. Para avaliar a qualidade do algoritmo desenvolvido, realizaram-se simulações de Monte Carlo para diferentes cenários, avaliando taxas de convergência, erros da estimação e eficiência das correções de viés e de Skovgaard.

Palavras-chave: Correção de viés; Correção de Skovgaard; Estimador de Máxima Verossimilhança; Parametrização geral; Linguagem R.

Abstract

This work aims to develop a package in R language with the implementation of normal regression models with general parameterization, proposed in Patriota e Lemonte (2011). This model unifies important models, such as nonlinear heteroscedastic models. Corrections are implemented for the MLEs and likelihood-ratio statistics. These corrections are effective in small samples. The algorithm considers the second-order bias of MLEs solution presented in Patriota e Lemonte (2009) and the Skovgaard's correction for likelihood-ratio statistics defined in Skovgaard (2001). In addition, a simulation study is developed under different scenarios, where the convergence ratio, relative squared error and the efficiency of bias correction and Skovgaard's correction are evaluated.

Keywords: Bias Correction; General Parameterization; Maximum Likelihood Estimator; Skovgaard's Correction; Software R.

Sumário

Lista de Figuras	13
Lista de Tabelas	15
1 Introdução	17
1.1 Introdução ao modelo	17
1.2 Objetivo	18
1.3 Implementação computacional	19
1.4 Estrutura do trabalho	20
2 Modelo	23
2.1 Estimação dos parâmetros	24
2.2 Correção de viés de segunda ordem	26
2.3 Correção da estatística da razão de verossimilhanças	27
3 Descrição do pacote	31
3.1 Instalação	31
3.2 Função principal: reg_general	31
3.3 Exemplos	34
4 Simulações	39
4.1 Descrição do estudo	39
4.2 Simulação de Monte Carlo	42
4.3 Resultados - Taxa de Convergência	44
Estudo de Casos	46
4.4 Viés e EQM dos EMV	49
Estudo de Casos	53

4.5	Correção de Viés	54
4.6	Teste de hipótese	57
	Teste de razão de verossimilhanças	57
	Correção de Skovgaard	59
5	Conclusões	65
5.1	Considerações finais	65
5.2	Próximos Passos	66
	Referências Bibliográficas	69

Lista de Abreviaturas

EMV	Estimadores de máxima verossimilhança
EQM	Erro quadrático médio
EQMr	Erro quadrático médio relativo
TRV	Teste da razão de verossimilhanças
iid	Independentes e identicamente distribuídos
ind	Independentes
cor	Corrigidos

Lista de Símbolos

<code>generalReg</code>	Pacote desenvolvido. Esta fonte é adotada para todos os comandos criados nele
<code>reg_general</code>	Função principal do pacote
<code>objeto</code>	Fonte adotada para comandos, fórmulas e outros pacotes na linguagem R
$A, B, \dots$	Matrizes ou vetores
$a, b, \dots$	Vetores
$\theta, \alpha, \beta, \gamma, \dots$	Vetores de parâmetros desconhecidos
$\hat{\theta}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}, \dots$	Estimadores dos vetores de parâmetros desconhecidos
$N_q(\mu, \Sigma)$	Normal q -variada com média μ e matriz de covariâncias Σ . Quando q for suprimido, trata-se de uma Normal univariada
Θ	Espaço paramétrico
<code>vec(.)</code>	Operador que transforma matriz em vetor, empilhando suas colunas
<code>diag(.)</code>	Vetor com os elementos da diagonal da matriz
$\otimes$	Produto de Kronecker
<code>b-diag(.)</code>	Operador que retorna uma matriz bloco diagonal com as matrizes na diagonal
$\ell(.)$	O logaritmo da função de verossimilhança (log-verossimilhança)
$\mathcal{H}_0$	Hipótese nula
$\mathcal{H}_1$	Hipótese alternativa

Lista de Figuras

4.1	Gráfico de dispersão dos dados do modelo. A linha pontilhada vermelha é referente à estimação corrigida e a linha contínua azul referente à estimação não corrigida, ambas obtidas em Vasconcellos e Cordeiro (1997)	40
4.2	Gráficos de dispersão dos dados, comparando as estimativas de Vasconcellos e Cordeiro (1997) com as estimativas do pacote <code>generalReg</code> . As linhas contínuas azuis são referentes às estimações sem correção e as linhas pontilhadas vermelhas às estimações corrigidas	42
4.3	Exemplos de dados simulados para cada um dos tamanhos amostrais diferentes	43
4.4	Estudo de caso para uma simulação específica de não convergência com $n=10$ e $\kappa=0.2$, com a curva verdadeira traçada	46
4.5	Estudo de caso para uma simulação específica de não convergência com $n=10$ e $\kappa=0.2$, com a estimativas da média representada pela linha azul contínua e a curva verdadeira tracejada	47
4.6	Estudo de caso para uma simulação específica de não convergência com $n=20$ e $\kappa=0.4$, com a curva verdadeira traçada	48
4.7	Estudo de caso para uma simulação específica de não convergência com $n=20$ e $\kappa=0.4$, com as estimativas para a média (estimadas com $\kappa=0.2$) representadas pela linha azul contínua	49
4.8	Boxplots das estimativas de MV para cada parâmetro do modelo (4.2), sem correção, para diferentes tamanhos de amostra. A linha horizontal indica o verdadeiro valor do parâmetro	50
4.9	Dados simulados para um caso de estimação discrepante com as estimativas esperadas. A linha tracejada se refere a curva verdadeira, enquanto a linha contínua verde se refere a $\kappa = 0.2$ e a linha azul é para $\kappa = 1$	53

4.10	Boxplots das estimativas de MV para cada parâmetro para diferentes tamanhos de amostra. As caixas vermelhas indicam as estimativas não corrigidas, enquanto as caixas azuis indicam as estimativas corrigidas. A linha horizontal indica o verdadeiro valor de θ	54
4.11	Histogramas dos valores-p nos testes da razão de verossimilhanças sob $\mathcal{H}_0$. . .	58
4.12	Histogramas dos valores-p nos testes da razão de verossimilhanças com correção de Skovgaard sob $\mathcal{H}_0$	60
4.13	Curva de densidade estimada dos valores-p nos testes com correção de Skovgaard sob $\mathcal{H}_0$ e sem correção	61
4.14	Discrepâncias relativas do valor-p para os testes de razão de verossimilhança, separados por n e κ	63
C1	Variação do impacto da correção de viés nos EQMr dos estimadores de máxima verossimilhança para diferentes parâmetros, n e κ	78
C2	Variação do impacto da correção de viés no viés relativo dos estimadores de máxima verossimilhança para diferentes parâmetros, n e κ	79

Lista de Tabelas

4.1	Base de Dados com as informações de crescimento de trigo utilizada para os modelos	39
4.2	Estimativas dos parâmetros do modelo obtidas em Vasconcellos e Cordeiro (1997)	40
4.3	Estimativa dos parâmetros pelo pacote generalReg	41
4.4	Percentual de casos com convergência para cada n e κ	44
4.5	Média de iterações, nos casos convergentes, para cada n e κ	44
4.6	Medidas resumo das estimativas, EQMr e viés para cada parâmetro e tamanho amostral	52
4.7	EQMr e viés para cada parâmetro, tamanho amostral e parâmetro κ	52
4.8	Comparação do Erro Quadrático Médio e Viés das correções de viés, em percentual, em relação a estimativa de Máxima Verossimilhança	55
4.9	Variação do impacto da correção de viés nos EQMr e Viés dos estimadores de máxima verossimilhança, medidos para diferentes parâmetros, n e κ	56
4.10	Taxa de rejeição de $\mathcal{H}_0$ para cada n , κ e α para estatísticas da razão de verossimilhanças	59
4.11	Taxa de rejeição de $\mathcal{H}_0$ para cada n , κ e α tanto corrigidos, como não	62
C1	Erro Quadrático Médio e viés relativos para EMVs, corrigido e não, para cada tamanho amostral para os parâmetros β_1 e β_2 do modelo de regressão apresentado na equação (4.2)	76
C2	Erro Quadrático Médio e viés relativos para EMVs, corrigido e não, para cada tamanho amostral para os parâmetros σ_1 e σ_2 do modelo de regressão apresentado na equação (4.2)	77

Capítulo 1

Introdução

1.1 Introdução ao modelo

Modelos de regressão são amplamente estudados na literatura e apresentam uma variedade muito grande de estruturas e aplicações. Há trabalhos realizados ainda no século XIX utilizando modelos de regressão linear (ver Hilary, 1967, para mais detalhes), os quais, embora não tenham uma grande complexidade, têm ainda uma vasta aplicabilidade (de Carvalho *et al.*, 2011, por exemplo). Esses modelos ajustam uma relação linear entre a esperança da variável de interesse (resposta) e variáveis explicativas (covariáveis). Tipicamente, considera-se em um modelo de regressão linear homoscedástico que os erros são independentes e identicamente distribuídos, segundo uma distribuição normal com média zero e variância finita e constante. A estrutura matemática para esse modelo de regressão é $Y_i = x_i^T \beta + e_i$, com $e_i \stackrel{iid}{\sim} N(0, \sigma^2)$, sendo Y_i a variável resposta, x_i o vetor de covariáveis observáveis e β o vetor de parâmetros da reta de regressão.

O modelo de regressão linear homoscedástico nem sempre é suficiente para um bom ajuste aos dados e, por isso, modelos com estruturas de média e variância mais complexas devem ser adotados (Seber e Wild, 1989). Por exemplo, há casos nos quais as relações das variáveis não são lineares e, portanto, modelos de regressão não lineares devem ser adotados (estudados, por exemplo, em Lee, Lu e Song, 2006; Sharma e Agarwal, 2003). Basicamente, a estrutura aqui passa a ser do tipo: $Y_i = \mu(\beta, x_i) + e_i$, com $e_i \stackrel{iid}{\sim} N(0, \sigma^2)$, sendo μ uma função diferenciável pelo menos duas vezes, por meio da qual se relacionam as covariáveis (x_i) com a esperança da variável resposta. Dessa forma, é possível estimar, por exemplo, relações logarítmicas, inversas, exponenciais, porém, a depender da complexidade dessa relação e da quantidade de parâmetros a serem estimados, as estimativas não são tão facilmente obtidas,

como são nos modelos de regressão linear (ver Transtrum *et al.*, 2010, para mais detalhes).

Outra suposição que na prática pode não ser satisfeita é a de variâncias iguais para os erros (homoscedasticidade). Nessas situações, é necessário empregar uma estrutura paramétrica também para a variância: são os ditos modelos de regressão heteroscedásticos, estudados, por exemplo, por Eubank e Thomas (1993); White (1980). Uma possível estrutura desse modelo segue: $Y_i = \mu(\beta, x_i) + e_i$, com $e_i \stackrel{ind}{\sim} N(0, \sigma_i^2)$ e $\sigma_i^2 = h(\gamma, w_i)$, sendo h uma função positiva em que w_i é o vetor de covariáveis associados à variância.

Todos os modelos citados acima são considerados particularidades de uma classe elíptica de modelos de regressão com parametrização geral, proposta por Patriota e Lemonte (2011). Esse modelo necessita apenas da definição de um vetor de localização e uma matriz de escala, ambos indexados por um mesmo vetor θ , além de uma função geradora de densidades, cuja parametrização define a distribuição da variável resposta. Nesse modelo, pode-se definir erros com distribuição normal, t de Student, exponencial-potência, entre outros pertencentes à família elíptica. Uma das grandes vantagens de ter uma estrutura única para todos os modelos é utilizar a mesma metodologia para estimação dos parâmetros, independente de qual tipo de regressão esteja sendo aplicada. Essa classe de modelos inclui também modelos mistos e modelos de regressão com erro nas variáveis, entre outros (Patriota e Lemonte, 2011).

Esta dissertação considerou apenas os casos univariados com erros normais na estrutura proposta por Patriota e Lemonte (2011). Os objetivos desta dissertação são descritos a seguir.

1.2 Objetivo

Considerando o modelo de regressão proposto em Patriota e Lemonte (2011), este trabalho tem como objetivo principal implementar um pacote que disponibilize:

1. o estimador de máxima verossimilhança (EMV) do vetor de parâmetros;
2. a correção de vies de segunda ordem para o EMV;
3. a correção de Skovgaard para a estatística da razão de verossimilhanças.

Inicialmente, tentou-se implementar o algoritmo para o caso multivariado, porém, por dificuldades computacionais, decidiu-se por considerar o caso univariado sob normalidade dos erros. Dessa forma, este trabalho contempla apenas modelos de regressão não lineares heteroscedásticos, em que a média e variância podem compartilhar parâmetros. O algoritmo

aqui desenvolvido pode servir de base para implementação dos resultados em modelos cuja distribuição dos erros pertença à classe das distribuições elípticas.

A rotina desenvolvida foi validada por meio de simulações de Monte Carlo, utilizando um modelo não linear heteroscedástico já estudado em Vasconcellos e Cordeiro (1997). Consideraram-se cenários com diferentes tamanhos amostrais e variações de um parâmetro κ , que controla a variação do processo de estimação em cada iteração e será definido em (2.6). Para cada cenário, foram calculados o viés relativo e o erro quadrático médio relativo (EQMr) para os estimadores de máxima verossimilhança e sua versão corrigida; foram calculadas também a distribuição empírica da estatística da razão de verossimilhanças e da sua versão corrigida. Os resultados demonstram que as correções (de viés e da razão de verossimilhanças) são eficazes para tamanhos amostrais pequenos ($n = 10$, $n = 20$, $n = 50$, $n = 100$). Também concluiu-se que altos valores de κ (0.8 e 1) causam problemas nas estimativas e correções, principalmente em tamanhos amostrais pequenos.

1.3 Implementação computacional

Há diversos *softwares*, tais como SPSS, SAS, R, Python, Minitab, Stata, entre outros, com implementações de técnicas estatísticas, dentre elas alguns dos modelos de regressão citados na Seção 1.1. Os *softwares open-source*, ou de código aberto vêm tendo cada vez mais aderência nos últimos anos. Especificamente no âmbito de análises estatísticas, ou de análise de dados em geral, esse fenômeno é analisado em Muenchen (2017). O estudo aponta uma crescente preferência pelo *software* estatístico de código aberto R¹ (*R Development Core Team*, 2014).

Softwares de código aberto se destacam pela possibilidade de qualquer usuário poder colaborar, propondo correções, alterações e melhorias. Levine e Prietula (2013) descrevem e analisam os benefícios dessa colaboração no ambiente científico.

Por esses motivos, optou-se por construir um pacote denominado `generalReg`, escrito na linguagem R. Suas duas funções principais são:

1. `reg_general` que calcula os estimadores de máxima verossimilhança e a correção de viés de segunda ordem;
2. a função `likelihood_ratio` que calcula os testes de razão de verossimilhanças e a

¹Pode-se baixar gratuitamente no *site* oficial (<https://www.r-project.org/>)

correção de Skovgaard.

O Capítulo 3 descreve as funções com maiores detalhes.

1.4 Estrutura do trabalho

O trabalho está estruturado em 5 principais capítulos. Além deste capítulo de introdução, os próximos capítulos são ,respectivamente, Modelo, Descrição do Pacote, Simulações e Conclusões, os quais são descritos a seguir, com os detalhamentos de cada seção.

- No Capítulo 2, o modelo de regressão com parametrização geral é apresentado:
 - a Seção 2.1 detalha o processo iterativo necessário para obtenção do EMV;
 - na Seção 2.2, são apresentados os cálculos para obtenção do viés de segunda ordem do EMV;
 - a Seção 2.3 descreve os testes da razão de verossimilhanças e detalha os cálculos necessários para a correção de Skovgaard.
- O Capítulo 3 introduz todas as funcionalidades que o pacote `generalReg` disponibiliza para os usuários:
 - a Seção 3.1 indica como instalar o pacote `generalReg` e descreve os pacotes necessários em seu desenvolvimento;
 - a Seção 3.2 explica detalhadamente as funções construídas no pacote, com a descrição de todos os argumentos;
 - a Seção 3.3 apresenta exemplos de como utilizar o pacote.
- O Capítulo 4 apresenta os resultados das simulações de Monte Carlo:
 - a Seção 4.1 apresenta o modelo e os dados estudados em Vasconcellos e Cordeiro (1997), comparando os resultados obtidos no artigo com as estimativas oferecidas pelo pacote `generalReg`;
 - a Seção 4.2 define a parametrização do modelo simulado e descreve as simulações de Monte Carlo;
 - a Seção 4.3 detalha os resultados das simulações para a taxa de convergência, detalhando a influência do tamanho amostral e do parâmetro κ nesta medida. São avaliados individualmente diferentes casos que divergiram;

- a Seção 4.4 apresenta os resultados dos EMV, comparando ao valor real de θ , por meio de indicadores como EQMr e viés relativo. Também são estudados casos em que a estimativa encontra-se muito distante do valor esperado;
 - a Seção 4.5 compara os valores apresentados na seção anterior com as estimativas obtidas pelo EMV corrigido, utilizando os mesmos indicadores. Aqui conclui-se que a depender o tamanho da amostra e de κ , a eficácia da correção pode ser afetada consideravelmente;
 - a Seção 4.6 apresenta testes da razão de verossimilhanças nas simulações realizadas, e compara o resultado com os testes realizados nas estatísticas corrigidas pela correção de Skovgaard.
- O Capítulo 5 traz as principais conclusões do trabalho realizado (Seção 5.1) e dos próximos passos que podem ser desenvolvidos a partir das contribuições realizadas nesta dissertação (Seção 5.2).

Por fim, no Apêndice estão as fórmulas para a correção de Skovgaard (Apêndice A), o detalhamento do conteúdo disponível em um objeto de classe `reg_general` (Apêndice B) e as tabelas com valores mais detalhados para consulta (Apêndice C).

Embora no Brasil usualmente utiliza-se “,” como separador decimal, esta dissertação utiliza o padrão americano, ou seja, “.” é adotado em textos, tabelas e figuras para separação decimal. Optou-se por esse padrão pois esse é o utilizado na linguagem R e este trabalho apresenta diversos exemplos de códigos e saídas geradas pelos programas.

Capítulo 2

Modelo

O modelo de regressão elíptico multivariado de parametrização geral é definido em Patriota e Lemonte (2011). Os modelos de regressão desenvolvidos nesta dissertação contemplam apenas erros com distribuição normal, os quais são uma particularidade da classe de distribuições elípticas. Portanto, as formulações apresentadas a seguir são simplificações dos cálculos apresentados em Patriota e Lemonte (2011) considerando apenas distribuições normais.

Sejam $Y_1, Y_2, \dots, Y_n$ vetores aleatórios q_i -dimensionais independentes. Os modelos de regressão normal com parametrização geral são dados por

$$Y_i = \mu_i(\theta, x_i) + u_i, \quad i = 1, 2, \dots, n, \quad (2.1)$$

com $u_i \stackrel{ind}{\sim} N_{q_i}(0, \Sigma_i(\theta, \omega_i))$, em que $\mu(\theta, x_i)$ é o vetor de médias, $\Sigma(\theta, \omega_i)$ é a matriz de variâncias e covariâncias e $\theta \in \Theta$, com $\Theta \subseteq \mathbb{R}^p$, $p < \infty$, é o vetor de parâmetros do modelo, sendo x_i e ω_i os vetores de covariáveis não aleatórias associadas à i -ésima observação da variável resposta (Y_i). Ambas as funções μ_i e Σ_i têm formas conhecidas e precisam ser quatro vezes diferenciáveis em relação a $\theta \in \Theta$. Pela estrutura (2.1), temos $Y_i \stackrel{ind}{\sim} \mathcal{N}_{q_i}(\mu_i(\theta), \Sigma_i(\theta))$, para $i = 1, 2, \dots, n$. A fim de simplificar a notação, tomam-se $\mu_i = \mu_i(\theta, x_i)$ e $\Sigma_i = \Sigma_i(\theta, \omega_i)$.

A função de log-verossimilhança para o modelo de regressão normal com parametrização geral, a menos de uma constante, é dada por (Patriota e Lemonte, 2009)

$$\ell(\theta) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} \text{tr} \left\{ (Y - \mu) \Sigma^{-1} (Y - \mu)^T \right\}. \quad (2.2)$$

sendo $Y = \text{vec}(Y_1, \dots, Y_n)$, $\mu = \text{vec}(\mu_1, \dots, \mu_n)$, $\Sigma = \text{b-diag}(\Sigma_1, \dots, \Sigma_n)$.

O estimador de máxima verossimilhança (EMV) é obtido maximizando a função de log-verossimilhança dada em (2.2) em relação a θ . A maximização pode ser feita igualando a função escore, ou seja, o gradiente da função de log-verossimilhança, a zero e resolvendo as equações resultantes em θ . Patriota e Lemonte (2011) calcularam a função escore e as informações de Fisher observada e esperada. A fim de explicitar as suas respectivas formas matriciais, as seguintes matrizes auxiliares foram definidas:

$$F_i = \begin{pmatrix} D_i \\ V_i \end{pmatrix}, \quad H_i = \begin{pmatrix} \Sigma_i & 0 \\ 0 & 2\Sigma_i \otimes \Sigma_i \end{pmatrix}^{-1} \quad \text{e} \quad s_i = \begin{pmatrix} z_i \\ -\text{vec}(\Sigma_i - z_i z_i^T) \end{pmatrix}, \quad (2.3)$$

em que $z_i = Y_i - \mu_i$, $D_i = \partial\mu_i/\partial\theta^T$, $V_i = \partial\text{vec}(\Sigma_i)/\partial\theta^T$, sendo vec o operador que transforma uma matriz em um vetor, empilhando suas colunas e “ $\otimes$ ” o produto de Kronecker. Assume-se que a matriz $F = (F_1^T, F_2^T, \dots, F_n^T)^T$ tem posto p , ou seja, as funções μ_i e Σ_i precisam ser definidas de forma a atender essa condição. Note que no caso em que $\mu_i(\theta)$ e $\Sigma_i(\theta)$ não tenham nenhum parâmetro em comum, F_i será uma matriz bloco diagonal. Isto é, se $\mu_i(\theta) = \mu_i(\theta_1)$ e $\Sigma_i(\theta) = \Sigma_i(\theta_2)$, sendo $\theta = (\theta_1^T, \theta_2^T)^T$, então $F_i = \text{b-diag}\{F_i(\theta_1), F_i(\theta_2)\}$, em que b-diag é o operador que forma uma matriz bloco-diagonal.

A Seção 2.1 apresenta o vetor escore, as informações de Fisher observada e esperada e o processo iterativo por meio do qual as estimações são obtidas. A Seção 2.2 apresenta a definição e os cálculos para correção de viés de segunda ordem dos estimadores de máxima verossimilhança e, por fim, a Seção 2.3 apresenta os testes de hipótese via estatísticas da razão de verossimilhanças, assim como suas correções.

2.1 Estimação dos parâmetros

A partir das quantidades definidas em (2.3) e por meio de algumas operações matriciais (Magnus e Neudecker, 1988), a função escore e a informação esperada de Fisher são dadas por (Patriota e Lemonte, 2011)

$$U_\theta = \sum_{i=1}^n F_i^T H_i s_i \quad \text{e} \quad K_{\theta\theta} = \sum_{i=1}^n F_i^T H_i F_i. \quad (2.4)$$

Já a informação observada de Fisher é dada por:

$$J_{\theta\theta} = \sum_{i=1}^n \left\{ F_i^T H_i \ddot{M}_i H_i F_i - [s_i^T H_i] \left[\frac{\partial F_i}{\partial \theta} \right] \right\}, \quad (2.5)$$

sendo $\ddot{M}_i = \begin{pmatrix} \Sigma_i & 2z_i^T \otimes \Sigma_i \\ 2z_i \otimes \Sigma_i & 2[\Sigma_i \otimes (z_i z_i^T) + (z_i z_i^T) \otimes \Sigma_i - \Sigma_i \otimes \Sigma_i] \end{pmatrix}$. Note que $\partial F_i / \partial \theta$ é um *arranjo* de dimensão $q_i(q_i + 1) \times p \times p$. Aqui, $[.][.]$ representa o produto de uma matriz por um *arranjo* definido em Wei (1998).

O desafio para calcular U_θ e $K_{\theta\theta}$, como apresentados na fórmula (2.4), está em encontrar as matrizes F_i , $i = 1, 2, \dots, n$ e suas derivadas. Observe que a informação de Fisher esperada pode ser escrita como $K_{\theta\theta} = F^T H F$, com $H = \text{b-diag}\{H_1, H_2, \dots, H_n\}$. Dessa forma, constata-se que a informação de Fisher é uma forma quadrática positiva definida, pois F tem posto completo e Σ_i e $i = 1, 2, \dots, n$ são positivas definidas.

A estimativa de máxima verossimilhança é obtida iterativamente, por meio do método escore de Fisher, segundo a expressão:

$$(F^{(m)T} H^{(m)} F^{(m)}) \theta^{(m+1)} = F^{(m)T} H^{(m)} (F^{(m)} \theta^{(m)} + \kappa s^{(m)}), \quad m = 0, 1, \dots \quad (2.6)$$

em que $F^{(m)} = (F_1^{(m)T}, F_2^{(m)T}, \dots, F_n^{(m)T})^T$, $H^{(m)} = \text{b-diag}\{H_1^{(m)}, H_2^{(m)}, \dots, H_n^{(m)}\}$, $s^{(m)} = (s_1^{(m)T}, s_2^{(m)T}, \dots, s_n^{(m)T})^T$, $\kappa \in (0, 1]$ e m é o contador das iterações.

Quando $\kappa = 1$, a expressão se reduz àquela apresentada em Patriota e Lemonte (2011). Entretanto, a depender da complexidade do modelo e da escolha dos valores iniciais, o algoritmo pode divergir rapidamente. Nesses casos, recomenda-se diminuir o valor de κ , para que a atualização das estimativas a cada passo seja feita mais lentamente. Espera-se que para $\kappa < 1$ a taxa de convergência aumente, e, como decorrência, o número de iterações também aumenta. O impacto da variação de κ^1 na convergência e na quantidade de iterações necessárias é estudado no Capítulo 4.

Para funcionamento do processo iterativo de estimação são necessários um valor de κ , uma estimativa inicial, a saber $\hat{\theta}^{(0)}$, e um critério de parada para o processo iterativo. No algoritmo desenvolvido, o erro definido para uma m -ésima iteração é dado por

¹Pode-se implementar o uso de um valor $\kappa^{(m)}$ que maximize em cada iteração a função de máxima verossimilhança e assim não há necessidade de escolha a priori de κ . Esta dissertação não vai analisar os resultados dessa funcionalidade.

$$\epsilon_m = \max_{j=1,\dots,p} \left| \frac{\hat{\theta}_j^{(m)} - \hat{\theta}_j^{(m-1)}}{\hat{\theta}_j^{(m-1)}} \right|. \quad (2.7)$$

Portanto, encerra-se o processo iterativo quando ϵ_m for menor do que algum valor de referência, por exemplo 10^{-10} . Ao fim do processo iterativo, toma-se $\hat{\theta}_{EMV} = \theta^{(m)}$. É importante também especificar o número máximo de iterações para evitar que o algoritmo continue indefinidamente em caso de divergência. Caso o algoritmo não convirja até essa iteração, o processo iterativo é interrompido e uma mensagem de alerta é retornada para o usuário.

2.2 Correção de viés de segunda ordem

A distribuição assintótica é empregada quando não se sabe a distribuição exata do estimador de máxima verossimilhança. Dessa forma, inferências aproximadas podem ser feitas utilizando a distribuição limite. Sob algumas condições de regularidade (ver Cox e Hinkley, 1974; Loose, 2019), estimadores de máxima verossimilhança são consistentes, eficientes e têm uma distribuição assintótica normal. Como consequência, o viés dos EMVs converge para zero, sendo tipicamente de ordem $O(n^{-1})$. Em modelos de regressão não lineares heteroscedásticos, esse viés pode não ser negligenciável para pequenas amostras (Vasconcellos e Cordeiro, 1997) e uma correção de viés se torna relevante.

A seguir são listados alguns trabalhos que estudaram correções de viés em EMVs: Cox e Snell (1968) apresentaram uma fórmula geral para o viés de segunda ordem dos EMVs considerando um modelo estatístico geral. Essa fórmula foi aplicada em alguns modelos específicos, a saber, modelo de regressão não linear homoscedástico (Box, 1971), modelo linear generalizado (Cordeiro e McCullagh, 1991) e modelo de regressão não linear heteroscedástico (Vasconcellos e Cordeiro, 1997), modelo de regressão não linear com distribuição t de Student (Vasconcellos e Cordeiro, 2000), modelo de regressão beta (Vasconcellos e Cribari Neto, 2005). Patriota e Lemonte (2009) obtiveram o viés de segunda ordem dos EMVs em um modelo de regressão normal com parametrização geral, uma classe ampla de modelos, como explicado na Seção 1.1. Mais recentemente, Melo *et al.* (2018) generalizaram os resultados de Patriota e Lemonte (2009) para a classe elíptica de distribuições. Nesta dissertação, implementaram-se os resultados apresentados em Patriota e Lemonte (2009) considerando-se apenas casos univariados.

A fim de apresentar o viés de segunda ordem de forma matricial, Patriota e Lemonte (2009) definiram as seguintes matrizes auxiliares, as quais dependem das derivadas de segunda ordem de μ_i e Σ_i

$$\begin{aligned} G_{ir} &= \begin{bmatrix} d_{i1r} & \dots & d_{ipr} \\ \text{vec}(C_{i1r}) & \dots & \text{vec}(C_{ipr}) \end{bmatrix}, \\ J_{ir} &= \begin{bmatrix} 0 \\ 2(I_{ni} \otimes d_{ir})D_i \end{bmatrix} \text{ e} \\ \Phi_{ir} &= -\frac{1}{2}(G_{ir} + J_{ir}), \end{aligned} \quad (2.8)$$

com $d_{ij} = \partial\mu_i/\partial\theta_j$, $d_{ijk} = \partial^2\mu_i/\partial\theta_j\partial\theta_k$ e $C_{ijk} = \partial^2\text{vec}(\Sigma_i)/\partial\theta_j\partial\theta_k$, $i = 1, 2, \dots, n$ e $j = 1, 2, \dots, p$.

O viés de segunda ordem para o modelo geral, da expressão (2.1), é dado por (Patriota e Lemonte, 2009)

$$B(\hat{\theta}) = (F^T H F)^{-1} F^T H \Phi, \quad (2.9)$$

em que $\Phi = (\Phi_1, \dots, \Phi_p)\text{vec}((F^T H F)^{-1})$, $\Phi_p = (\Phi_{1p}^T, \Phi_{2p}^T, \dots, \Phi_{np}^T)^T$. Observe que o cálculo do viés envolve apenas operações simples entre matrizes e vetores para qualquer estrutura não linear de média e variância, que sejam três vezes diferenciáveis em θ .

O estimador de máxima verossimilhança corrigido é dado por (Patriota e Lemonte, 2009)

$$\hat{\theta}_{cor} = \hat{\theta} - \widehat{B_{\hat{\theta}}(\hat{\theta})}. \quad (2.10)$$

Na Seção 4.5, são analisados os impactos da correção de viés em cenários simulados.

2.3 Correção da estatística da razão de verossimilhanças

Em testes de hipóteses, tipicamente são definidas duas hipóteses, uma hipótese nula $\mathcal{H}_0$ e outra alternativa $\mathcal{H}_1$. Quando as duas hipóteses são simples, Neyman e Pearson (1933) provaram que o teste de razão da verossimilhanças (TRV) é o teste mais poderoso para um

nível de significância fixado. Quando as hipóteses não são simples, ou seja, o espaço paramétrico sob a hipótese em questão tem mais de um elemento, em geral, não há um teste mais poderoso. Entretanto, sob algumas condições de regularidade, a estatística $(-2 \log)$ da razão de verossimilhanças converge em distribuição para uma qui-quadrado (Chernoff, 1954; Drton, 2009).

Considere o vetor de parâmetros desconhecidos particionado da seguinte forma $\theta = (\psi^T, \omega^T)^T \in \Theta \subseteq \mathbb{R}^p$, com $\psi \in \mathbb{R}^q$ sendo o vetor de parâmetros de interesse e $\omega \in \mathbb{R}^{(p-q)}$ o vetor de parâmetros de perturbação. As hipóteses nula e alternativa são dadas, respectivamente, por $\mathcal{H}_0 : \psi = \psi^{(0)}$ e $\mathcal{H}_1 : \psi \neq \psi^{(0)}$, sendo $\psi^{(0)}$ um vetor de dimensão q com valores conhecidos. Dessa forma, o espaço paramétrico sob a hipótese nula é dado por $\Theta_0 = \{\theta \in \Theta : \psi = \psi^{(0)}\}$. Note que a hipótese nula será composta sempre que $q < p$.

O estimador de máxima de verossimilhança sob a hipótese irrestrita é denotado por $\hat{\theta} = (\hat{\psi}^T, \hat{\omega}^T)^T$, enquanto o estimador de máxima verossimilhança restrito a hipótese nula é $\tilde{\theta} = (\tilde{\psi}^T, \tilde{\omega}^T)^T$. Todas as quantidades posteriores com os símbolos “ $\hat{\cdot}$ ” e “ $\tilde{\cdot}$ ” indicam que o valor de θ é substituído por $\hat{\theta}$ e $\tilde{\theta}$, respectivamente.

A estatística $(-2 \log)$ da razão de verossimilhanças usual para testar $\mathcal{H}_0$ contra $\mathcal{H}_1$ é dada por:

$$LR = 2\{\ell(\hat{\theta}) - \ell(\tilde{\theta})\}, \quad (2.11)$$

em que ℓ é a função de log-verossimilhança definida em (2.2). Sob condições de regularidades (Severini, 2000, Seção 3.4), LR converge em distribuição para uma χ_q^2 sob $\mathcal{H}_0$.

Para amostras pequenas, essa distribuição assintótica pode não ser adequada (vide os estudos em Brazzale e Davison, 2008; Ferrari e Pinheiro, 2014; Guolo, 2012; Lemonte e Ferrari, 2001, por exemplo), necessitando assim de alguma correção na estatística para que sua distribuição exata seja mais próxima da distribuição assintótica: Barndorff-Nielsen (1986) apresentou uma nova estatística corrigida para hipóteses unidimensionais e Skovgaard (2001) estendeu essa aplicação para hipóteses multidimensionais. Esta dissertação implementa a correção de Skovgaard para modelos de regressão com parametrização geral desenvolvida em Melo *et al.* (2017), a qual é apresentada a seguir, na fórmula (2.13). Essa correção contempla modelos de regressão multivariados com distribuições elípticas, enquanto esta dissertação considera uma particularidade desse modelos: apenas

casos univariados com distribuição normal. Devido a isso, alguns cálculos aqui tornam-se mais simples do que os apresentados em Melo *et al.* (2017).

Como discutido em Peña *et al.* (1992) e Melo *et al.* (2017), um desafio no cálculo da correção de Skovgaard é a obtenção de uma estatística ancilar a tal que sua distribuição não dependa de θ e, além disso, $(a, \hat{\theta})$ seja uma estatística suficiente para o modelo. Para modelos de regressão parametrização geral, Melo *et al.* (2017) apresentaram uma estatística assintoticamente ancilar $a = (a_1^T, \dots, a_n^T)^T$, isto é, sua distribuição assintótica independe de θ . Os autores argumentam que essa estatística pode ser usada no lugar da estatística ancilar, pois é possível mostrar, pelo teorema de Slutsky e teorema de fatorização de Neyman-Fisher, que a converge em probabilidade para uma distribuição que não depende de θ e que $(a, \hat{\theta})$ é estatística suficiente (ver Melo *et al.*, 2017). A estatística aproximadamente ancilar é dada por $a = (a_1^T, \dots, a_n^T)^T$ em que

$$a_i = \hat{P}_i^{-1}(Y_i - \hat{\mu}_i), \quad (2.12)$$

sendo que $P_i = \Sigma_i^{\frac{1}{2}}$ é obtido via decomposição de Cholesky, para $i = 1, \dots, n$.

Para cálculo da correção de Skovgaard, as seguintes matrizes são definidas:

$$\ell'(\theta; \hat{\theta}; a) = \frac{\partial \ell(\theta; \hat{\theta}; a)}{\partial \hat{\theta}}, \quad U'(\theta; \hat{\theta}; a) = \frac{\partial^2 \ell(\theta; \hat{\theta}; a)}{\partial \hat{\theta} \partial \theta^T} \quad \text{e} \quad J(\theta; \hat{\theta}; a) = -\frac{\partial^2 \ell(\theta; \hat{\theta}; a)}{\partial \theta \partial \theta^T}.$$

Note que $\ell'(\theta; \hat{\theta}; a)$ é um vetor p -dimensional, $U'(\theta; \hat{\theta}; a)$ é uma matriz $p \times p$ e $J(\theta; \hat{\theta}; a)$ é uma matriz $p \times p$.

A estatística da razão de verossimilhanças corrigida pelo método de Skovgaard é dada por (Skovgaard, 2001)

$$LR^* = LR - 2 \log(\rho), \quad (2.13)$$

em que

$$\rho = |\hat{J}|^{\frac{1}{2}} |\tilde{U}'|^{-1} |\tilde{J}_{\omega\omega}|^{\frac{1}{2}} |\tilde{\tilde{J}}_{\omega\omega}|^{-\frac{1}{2}} |\tilde{\tilde{J}}|^{\frac{1}{2}} \frac{(\tilde{U}^T \tilde{\tilde{J}}^{-1} \tilde{U})^{p/2}}{LR^{q/2-1} (\hat{\ell}' - \tilde{\ell}')^T (\tilde{U}')^{-1} \tilde{U}}, \quad (2.14)$$

$\hat{\ell}' = \ell'(\hat{\theta}; \hat{\theta}; a)$, $\tilde{\ell}' = \ell'(\tilde{\theta}; \hat{\theta}; a)$, $\tilde{U}' = U'(\tilde{\theta}; \hat{\theta}; a)$, $\tilde{\tilde{J}} = J(\tilde{\theta}; \tilde{\theta}; a)$ e os índices $\omega\omega$ em $J_{\omega\omega}$ se referem à submatriz de J para as linhas referentes aos parâmetros de perturbação. Skovgaard (2001)

prova que a estatística LR^* também converge em distribuição para uma χ_q^2 e espera-se que para pequenas amostras, a distribuição de LR^* se aproxime mais de sua distribuição limite.

Melo *et al.* (2017) desenvolveram as fórmulas para obtenção das matrizes em (2.14) para modelos de regressão elípticos multivariados com parametrização geral. Esta dissertação utiliza essas fórmulas de maneira a contemplar apenas modelos de regressão normais univariados. As formulações dessas matrizes se encontram no Apêndice A deste trabalho.

Capítulo 3

Descrição do pacote

3.1 Instalação

Os códigos desenvolvidos nesta dissertação são disponibilizados em um pacote em linguagem R, chamado de `generalReg`. O pacote está disponível no endereço <https://github.com/aperette/generalReg> e pode ser instalado por meio do pacote `devtools`, com o comando:

```
devtools::install_github("aperette/generalReg").
```

Ao instalá-lo, outros pacotes utilizados no desenvolvimento do algoritmo são instalados ou atualizados automaticamente, os quais são descritos abaixo:

- **dplyr**: pacote que facilita a manipulação de vetores, tabelas e matrizes;
- **stringr**: pacote para transformações/manipulação de *strings* (cadeias de caracteres). É necessário para traduzir os argumentos de média e escala escolhidos pelo usuário;
- **deriv**: pacote utilizado para todos os cálculos de derivadas necessários;
- **Matrix**: pacote que auxilia em operações de matrizes, e principalmente, permite a utilização de matrizes esparsas. Este último é essencial para o algoritmo funcionar de forma mais rápida.

3.2 Função principal: `reg_general`

Instalado o pacote proposto neste trabalho, o usuário tem acesso às funções desenvolvidas, cuja principal delas é a função `reg_general`. A função recebe as estruturas de média e variância do modelo e a tabela de dados para aplicar os principais cálculos (estimação dos parâmetros, predição do modelo, correção de viés, vetor escore e informações de Fisher). A

função já conta com uma documentação com o resumo do pacote e exemplos de como utilizar, que podem ser acessados por meio do comando `help(reg_general)`.

Os argumentos da função foram inspirados em pacotes de regressão em linguagem R, como `lm`, `nls`, a fim de manter uma padronização dentro do *software* e assim facilitar a utilização do pacote. Os argumentos são:

1. `formula`: estrutura da média. Deve ser no formato de **formula** (como, por exemplo, $y \sim a + xb$ ou $y \sim 1/(a + xb)$);
2. `formula_var`: estrutura da diagonal da matriz de covariâncias. Deve ser no formato de **formula** (por exemplo, $\sim a + xb$ ou $\sim \exp(a + xb)$ para garantir que a variância seja positiva);
3. `data`: tabela, no formato **data.frame** com o banco de dados. Os nomes das colunas deverão ser os mesmos utilizados nos argumentos anteriores;
4. `theta_ini`: estimativas iniciais dos parâmetros. Em geral, a convergência do algoritmo depende da escolha desses valores. Deve ser uma lista com os valores para os parâmetros declarados em `formula` e `formula_var`. Caso não seja preenchida, o algoritmo assume inicialmente todos os coeficientes com valor 1;
5. `bias_correction`: valor lógico. Caso seja **TRUE**, as estimativas são obtidas pelo estimador corrigido. Se for **FALSE**(*default*), será utilizado o EMV, sem correção;
6. `control`: lista contendo parâmetros de controle do algoritmo, que podem ser:
 - `reltol`: diferença relativa máxima das estimações entre duas iterações (equação (2.7)) para considerar convergência. Valores menores tendem a trazer maior precisão nas estimativas, porém exigem maior número de iterações e pode fazer o algoritmo não convergir. *Default*: $1e - 5$;
 - `max_it`: número máximo de iterações sem convergência até parar. *Default*: 500;
 - `kappa`: parâmetro definido em (2.6). Valores menores tendem a facilitar a convergência, porém necessitam de mais iterações. *Default*: 1;
 - `verbose`: indica se o algoritmo deve emitir mensagens durante o funcionamento. Caso seja preenchido 0, nenhuma mensagem é exibida. Para valores maiores do que 1, a função exibe mensagens a cada z iterações, sendo z o valor preenchido neste campo. *Default*: 0

A função `reg_general` retorna um objeto de classe de mesmo nome, `reg_general`. Um objeto dessa classe contém uma lista com várias informações sobre o modelo e suas estimativas. Todas essas informações estão descritas com detalhes no Apêndice B.

As vantagens de se definir uma nova classe, entre outras, é de criar particularidades para funções genéricas muito utilizadas na linguagem R. As funções desenvolvidas para a classe `reg_general` são:

- `print`: impressão das estimativas quando o objeto é inserido no console do R;
- `summary`: impressão mais completa das estimativas do modelo, com erro padrão e testes de significância dos parâmetros;
- `coef`: vetor de estimativa dos parâmetros estimados;
- `confint`: intervalo de confiança para os parâmetros;
- `vcov`: matriz de variância e covariância dos parâmetros;
- `predict`: função que retorna a média, ou a variância para a base construída, ou para uma nova tabela (com os mesmos campos inseridos no ajuste). Seus argumentos são:
 - `x`: objeto do modelo;
 - `newdata`: base de dados em que será aplicado o modelo;
 - `type`: determina se a função retorna a média ou a variância. Seu valor deve ser “var” ou “mean”.
- `likelihood_ratio`: realiza testes de hipóteses, via razão de verossimilhanças, de acordo com as estimativas do modelo. Seus argumentos são:
 - `x`: objeto do modelo;
 - `parameters`: lista nomeada de parâmetros a serem testados sob a hipótese nula. Supõe-se que os parâmetros ausentes nessa lista não têm restrição sob $\mathcal{H}_0$;
 - `correction`: valor lógico. Caso seja **TRUE**, retorna-se a estatística do teste com correção de Skovgaard. Se **FALSE**, retorna-se a estatística da razão de verossimilhanças sem correção.

Detalhadas todas as funcionalidades do algoritmo, a próxima seção apresenta exemplos de como utilizar o pacote.

3.3 Exemplos

A seguir é apresentado um exemplo de ajuste utilizando o pacote. Ajusta-se um modelo de regressão linear simples homoscedástico, no banco de dados `mtcars` do pacote básico do R, que traz informações sobre automóveis nos Estados Unidos. A coluna `mpg` (milhas por galão de combustível) é a variável resposta e `disp` (distância percorrida) é a covariável:

```
fit <- generalReg::reg_general(
  formula = mpg~alfa + beta*disp,
  formula_var = ~sigma,
  data=mtcars)
```

Esse exemplo ilustra como o usuário deve inserir as informações para o algoritmo funcionar. A acurácia das estimativas do modelo é estudada no Capítulo 4. Note que, nesse exemplo, não foi necessário fornecer valores iniciais para θ , pois o algoritmo procura identificar os nomes dos parâmetros a partir dos argumentos de `formula` e `formula_var` e atribui valor 1 para todos os coeficientes. Porém, na maioria das vezes, isso faz o algoritmo divergir, como por exemplo:

```
fit <- generalReg::reg_general(
  formula = mpg~ alfa+1/(beta*disp),
  formula_var = ~(sigma),
  data=mtcars)
```

```
## Error: Non convergence
```

Quando o processo não é executado com sucesso, a função retorna uma mensagem de erro. Neste caso, como o modelo é mais complexo, é necessário introduzir valores iniciais razoavelmente próximos da EMV para o algoritmo convergir:

```
fit <- generalReg::reg_general(
  formula = mpg~ alfa+1/(beta*disp),
  formula_var = ~(sigma),
  data=mtcars,
  start=list(alfa=10.75,beta=0.001,sigma=5))
```

Para exibição dos resultados do modelo, escreve-se o nome do objeto (`fit`) ou utiliza-se a função `print`. Para resultados mais detalhados, como o erro padrão e valores-p dos testes

de significância dos coeficientes utiliza-se o comando **summary**:

```
fit
```

```
## Call:
## generalReg::reg_general(formula = mpg ~ alfa + 1/(beta * disp),
##   formula_var = ~(sigma), data = mtcars, start = list(alfa = 10.75,
##   beta = 0.001, sigma = 5))
##
##
## Coefficients:
##      alfa      beta      sigma
## 10.75202  0.00064  4.93749
##
## Corrected coefficients:
##      alfa      beta      sigma
## 10.75202  0.00064  5.24608
```

```
summary(fit)
```

```
##
## Call:
## generalReg::reg_general(formula = mpg ~ alfa + 1/(beta * disp),
##   formula_var = ~(sigma), data = mtcars, start = list(alfa = 10.75,
##   beta = 0.001, sigma = 5))
##
##
## Residuals:
##      Min.      1st Qu.      Median      3rd Qu.      Max.
## -3.73826571 -1.82157641 -0.07511027  1.04865993  4.61048610
##
##
## Coefficients:
##      Estimate Std. Error  t.value      Pv
## alfa  10.75202    0.77402 13.89115 2.896361e-23
## beta   0.00064    0.00005 14.00217 2.300772e-23
```

```
## sigma 4.93749 1.23437 4.00000 3.132911e-09
```

Para utilização da correção de viés, utiliza-se o argumento `bias_correction=TRUE`.

```
fit <- generalReg::reg_general(
  formula = mpg~ alfa+1/(beta*disp),
  formula_var = ~(sigma),
  data=mtcars, bias_correction = TRUE,
  start=list(alfa=10.75,beta=0.001,sigma=5))
fit

## Call:
## generalReg::reg_general(formula = mpg ~ alfa + 1/(beta * disp),
##   formula_var = ~(sigma), data = mtcars, start = list(alfa = 10.75,
##     beta = 0.001, sigma = 5), bias_correction = TRUE)
##
##
## Coefficients:
##      alfa      beta      sigma
## 10.75202  0.00064  4.93749
##
## Corrected coefficients:
##      alfa      beta      sigma
## 10.75202  0.00064  5.24608
```

Além de estimar os parâmetros, uma funcionalidade muito importante é a de predição dos valores esperados. Para isso, utiliza-se o comando `predict`, com o objeto do modelo como primeiro argumento. O segundo argumento, `newdata`, define a base de dados em que o modelo será aplicado. Caso este segundo argumento não seja preenchido, a saída considera os dados utilizados na construção do modelo, equivalente a utilizar o comando `x$fitted.values`:

```
fit$fitted.values

## [1] 20.48748 20.48748 25.17492 16.78951 15.07889 17.67501 15.07889
## [8] 21.37011 21.81504 20.04602 20.04602 16.39986 16.39986 16.39986
## [15] 14.05217 14.13827 14.29219 30.54457 31.32895 32.66023 23.72183
## [22] 15.65036 15.87595 15.20251 14.64620 30.46941 23.70026 27.13134
```

```
## [29] 15.18984 21.49460 15.92701 23.62536
```

```
predict(fit)
```

```
## [1] 20.53739 20.53739 25.24887 16.82047 15.10107 17.71050 15.10107
## [8] 21.42454 21.87176 20.09366 20.09366 16.42881 16.42881 16.42881
## [15] 14.06909 14.15563 14.31034 30.64604 31.43444 32.77255 23.78832
## [22] 15.67548 15.90221 15.22533 14.66617 30.57049 23.76664 27.21531
## [29] 15.21259 21.54967 15.95354 23.69135
```

Para aplicar as estimativas do modelo em uma base de dados nova, basta preencher o campo **newdata** com uma tabela que contenha os campos utilizados como covariáveis no modelo. O argumento **type** define se a saída será a média estimada (preenchimento padrão) ou a variância estimada. Se o argumento **type** for preenchido com alguma entrada diferente, o algoritmo acusa um erro.

```
mtcars2 <- data.frame(disp=c(100, 50))
predict(fit,mtcars2)
```

```
## [1] 26.40861 42.06521
```

```
predict(fit,mtcars2,type="var")
```

```
## [1] 5.246081 5.246081
```

```
predict(fit,mtcars2,type="desvio padrao")
```

```
## Error: type must be 'mean' or 'var'
```

Feitos os modelos, agora pode-se realizar testes de hipótese para os coeficientes com a função **likelihood_ratio**:

```
generalReg::likelihood_ratio(fit,correction = T,
                             parameters = list(alfa=10.5,beta=0.0007,sigma=5.

## Likelihood ratio test:
## LR:7.742 p-value:5.17e-02
## Skovgaard correction:
## LR*:7.322 p-value:6.23e-02
```

No caso acima, a hipótese nula é uma hipótese simples, ou seja, o espaço paramétrico

associado contém apenas um ponto. Portanto não é necessário estimar o EMV restrito, pois este é o próprio ponto do espaço paramétrico sob $\mathcal{H}_0$. Quando restringe-se apenas parte dos coeficientes, a estimação sob $\mathcal{H}_0$ é necessária.

```
generalReg::likelihood_ratio(fit, correction = T
                             , parameters = list(beta=0.0007))

## MLE under H0:
## alfa:11.526  sigma:5.145

## Likelihood ratio test:
## LR:1.319 p-value:2.51e-01
## Skovgaard correction:
## LR*:0.538 p-value:4.63e-01
```

Este capítulo apresentou a estrutura do pacote e exemplos de como utilizá-lo. No próximo capítulo, apresentam-se os resultados das estimações em cenários simulados.

Capítulo 4

Simulações

4.1 Descrição do estudo

Para testar tanto a metodologia como o funcionamento do algoritmo, são simulados alguns casos sob diferentes cenários. Para dar maior veracidade aos testes realizados, este trabalho baseia-se em modelagens já apresentadas na literatura e a escolha foi feita por um modelo complexo dentre os modelos que o pacote contempla (um modelo não linear heteroscedástico).

Foram escolhidos dados estudados em Vasconcellos e Cordeiro (1997) para ajuste de um modelo de regressão não linear heteroscedástico. O banco de dados em questão apresenta o crescimento de trigo de inverno (peso de matéria seca, em quilogramas) e informação de calor (em uma escala de graus-dias), os quais são apresentados na Tabela 4.1.

Tabela 4.1: Base de Dados com as informações de crescimento de trigo utilizada para os modelos

Calor (x)	Peso de matéria seca (Y)
405.65	90.5, 113.383
498.75	161.6, 207.65
567.25	246.743, 309.514
618.30	422.936, 460.686
681.45	868.662, 972.383, 999.633, 1034, 1047, 1072.022, 1133.287, 1141.883, 1266.29, 1169.767

A estrutura proposta por Vasconcellos e Cordeiro (1997) foi de um modelo de regressão heteroscedástico com relações não lineares para a média e também para a variância, necessária para que essa seja positiva. O modelo foi definido por

$$\begin{aligned}\mu_i(\theta) &= \beta_1 \exp(\beta_2 x_i), \\ \Sigma_i(\theta) &= \sigma_1 \exp(\sigma_2 x_i),\end{aligned}\tag{4.1}$$

em que x_i é a medida de calor (escala de graus-dias) e Y_i é o peso de matéria seca e variável resposta do modelo, com $i = 1, 2, \dots, 18$ e $\theta = (\beta_1, \beta_2, \sigma_1, \sigma_2)^T$.

A estimação dos parâmetros feita em Vasconcellos e Cordeiro (1997) utilizou um algoritmo descrito em Cordeiro e Paula (1989). Os resultados, com e sem correção de viés, são apresentados na Tabela 4.2 e, a partir deles, Vasconcellos e Cordeiro (1997) concluem que o viés na variância era maior do que na média.

Na Figura 4.1 traçaram-se as médias esperadas, com e sem correção de viés, no gráfico de dispersão entre x e Y , a fim de verificar se o modelo foi bem ajustado. Nota-se que as estimativas do artigo não parecem se ajustar muito bem aos pontos, ainda que a correção de viés aproxime-se mais dos pontos observados. Observando apenas os pontos do gráfico, percebe-se uma tendência não linear entre as variáveis, assim como a dispersão parece aumentar para maiores valores de x , ou seja, visualmente o modelo escolhido parece adequado.

Tabela 4.2: Estimativas dos parâmetros do modelo obtidas em Vasconcellos e Cordeiro (1997)

Parâmetros	Estimativa	Erro Padrão	Viés	Estimativa Corrigida
β_1	1.1985	0.4132	-0.1033	1.3018
β_2	0.0108	0.0034	0.0019	0.0089
σ_1	8.2716	2.6931	3.0524	5.2192
σ_2	0.0127	0.0085	-0.0011	0.0138

Figura 4.1: Gráfico de dispersão dos dados do modelo. A linha pontilhada vermelha é referente à estimação corrigida e a linha contínua azul referente à estimação não corrigida, ambas obtidas em Vasconcellos e Cordeiro (1997)

No que se segue, os parâmetros são estimados via pacote `generalReg` para comparação dos resultados obtidos. A sintaxe utilizada nessa estimação é indicada a seguir, com os

valores iniciais iguais às estimativas da 4.2.

```
data=data.frame(
  X=c(405.65, 498.75, 567.25, 618.3, 681.45,
      405.65, 498.75, 567.25, 618.3, 681.45,
      681.45, 681.45, 681.45, 681.45, 681.45,
      681.45, 681.45, 681.45),
  Y=c(90.5, 161.6, 246.743, 422.936, 868.662,
      113.383, 207.65, 309.514, 460.686, 972.383,
      999.633, 1034, 1047, 1072.022, 1133.287,
      1141.883, 1266.290, 1169.767))

fit = reg_general(
  Y~beta1*exp(beta2*X),
  ~sigma1*exp(sigma2*X),
  data=data,
  start = list(beta1=1, beta2=0.01, sigma1=8.3, sigma2=0.01))
```

Tabela 4.3: Estimativa dos parâmetros pelo pacote *generalReg*

Parâmetro	EMV		EMV Corrigido	
	Estimativas	Erro Padrão	Estimativas	Erro Padrão
β_1	0.8113	0.6570	0.8113	0.6570
β_2	0.0105	0.0012	0.0105	0.0012
σ_1	69.2572	16.9413	69.2572	16.9413
σ_2	0.0075	0.0046	0.0075	0.0046

As estimativas apresentadas na Tabela 4.3 apresentam resultados bem divergentes em relação aos encontrados em Vasconcellos e Cordeiro (1997). Como os estimadores propostos neste trabalho são de máxima verossimilhança, devem ser comparados seus respectivos valores na função log-verossimilhança e aquele com maior valor deve ser escolhido. O valor da função log-verossimilhança encontrado no estimador do pacote foi de -104.9145 e pode ser obtido por `fit$loglike`. O valor da função log-verossimilhança para as estimativas da Tabela 4.2 é de -196.6235. Portanto, o maior valor de verossimilhança foi obtido pelo pacote *generalReg*.

Na Figura 4.2, são traçadas as projeções das médias no gráfico de dispersão dos dados, comparando os resultados obtidos neste trabalho com as estimativas corrigidas do artigo. De forma clara, o ajuste via pacote *generalReg* está mais próximo aos pontos e, portanto, as simulações feitas na sequência do trabalho serão baseadas nestas estimativas. Nota-se que

as estimativas dos parâmetros apresentados em Vasconcellos e Cordeiro (1997) não indicam um bom ajuste aos dados, principalmente sem correção. Algumas hipóteses para este efeito podem ser alguma escolha ruim de estimativas iniciais, resultando em convergência para um máximo local, uma não convergência até o número de iterações realizados ou simplesmente um erro de digitação, possivelmente para $\hat{\beta}_1$. Os autores foram consultados e a causa está sendo analisada.

Figura 4.2: Gráficos de dispersão dos dados, comparando as estimativas de Vasconcellos e Cordeiro (1997) com as estimativas do pacote `generalReg`. As linhas contínuas azuis são referentes às estimações sem correção e as linhas pontilhadas vermelhas às estimações corrigidas

4.2 Simulação de Monte Carlo

Definidos a estrutura do modelo e valores dos parâmetros, são simulados alguns cenários para testar tanto a metodologia como o funcionamento do algoritmo. O algoritmo é aplicado para estimar os parâmetros e suas estimativas são comparadas com os valores de θ que geraram os dados. Os resultados são obtidos via Método de Monte Carlo, repetindo esse processo $M = 10000$ vezes. É importante lembrar que a depender da complexidade das estruturas e dos valores iniciais, o processo iterativo pode divergir. Em caso de divergência, uma solução proposta é utilizar $\kappa < 1$.

O modelo de regressão univariado não linear e heteroscedástico é definido de forma semelhante a (4.1)

$$\begin{aligned}\mu_i(\theta) &= \exp(\beta_1 + \beta_2 x_i), \\ \Sigma_i(\theta) &= \exp(\sigma_1 + \sigma_2 x_i),\end{aligned}\tag{4.2}$$

em que $\beta_1 = -0.21$, $\beta_2 = 0.01$, $\sigma_1 = 4.2$, $\sigma_2 = 0.008$ para $i = 1, 2, \dots, n$.

A alteração utilizada aqui foi na parametrização de β_1 e σ_1 , para evitar restrição nos valores de θ (na Estrutura (4.1), havia a restrição de $\sigma_1 > 0$). Essa prática também auxilia na taxa de convergência. Vale lembrar que as estimativas obtidas (Tabela 4.3) foram de $\hat{\beta}_1 = 0.81$ e $\hat{\sigma}_1 = 69$ e para adaptar à nova parametrização, na Estrutura (4.2), é necessário aplicar o logaritmo nas estimativas de $\hat{\beta}_1$ e $\hat{\sigma}_1$. Por fim, os valores foram arredondados para melhor exibição dos resultados.

Os valores de x_i foram gerados segundo uma distribuição $U(400, 700)$ para manter a escala parecida aos dados em Vasconcellos e Cordeiro (1997). São testados tamanhos amostrais $n = 10$, $n = 20$, $n = 50$, $n = 100$ e $n = 500$, e valores de κ (Equação (2.6)) iguais a 0.2, 0.4, 0.5, 0.8 e 1. As variáveis de controle do algoritmo foram definidas da seguinte forma:

- Número máximo de iterações: 5000;
- Diferença máxima utilizada entre dois passos para convergência: 10^{-10} .

Figura 4.3: Exemplos de dados simulados para cada um dos tamanhos amostrais diferentes

São apresentados exemplos de amostras para cada um dos tamanhos amostrais simulados com o intuito de verificar se o comportamento esperado pelo modelo se reflete nos dados simulados. A Figura 4.3 traz gráficos de dispersão entre a variável x e a variável resposta Y para cada n . Nota-se que a média cresce em uma curva exponencial em função de x , como esperado, enquanto a heteroscedasticidade é um pouco mais sutil de se perceber, principalmente para amostras menores. No geral, a estrutura proposta parece estar refletida nos dados.

A seguir são apresentados os resultados de convergência do processo iterativo, ou seja, quando o algoritmo consegue estimar sem retornar nenhuma mensagem de erro.

4.3 Resultados - Taxa de Convergência

A não convergência costuma ser causada por valores iniciais de θ distantes do que os dados indicam, se alguma condição de regularidade não foi cumprida, ou se atingiu-se o número máximo de iterações. Para cada uma das 10000 repetições, avaliou-se se o algoritmo convergiu. A seguir, na Tabela 4.4, são apresentadas as taxas de simulações com convergência para os diferentes valores de κ e n , bem como seus resultados marginais. A Tabela 4.5 apresenta o valor médio de iterações necessários para convergência (descartando os casos sem convergência).

Tabela 4.4: Percentual de casos com convergência para cada n e κ

κ	n					
	10	20	50	100	500	Total
0.2	99.8	100.0	100.0	100.0	100	100.0
0.4	99.7	99.9	100.0	100.0	100	99.9
0.6	97.9	99.2	99.9	100.0	100	99.4
0.8	94.3	96.0	99.2	100.0	100	97.9
1	88.6	90.8	97.0	99.3	100	95.1
Total	96.1	97.2	99.2	99.9	100	98.5

Tabela 4.5: Média de iterações, nos casos convergentes, para cada n e κ

κ	n					
	10	20	50	100	500	Total
0.2	259.0	175.8	139.8	129.4	117.9	164.4
0.4	148.1	95.7	73.4	67.6	61.9	89.3
0.6	106.2	65.5	50.6	46.4	42.4	62.1
0.8	83.5	49.7	38.5	35.2	31.9	47.3
1	77.3	51.5	31.9	28.1	24.1	41.7
Total	136.9	88.7	67.1	61.4	55.6	81.5

A taxa de convergência se manteve alta para todo valor de n e κ : 98.4% de todas as simulações realizadas convergiram e a combinação de n e κ com pior taxa de convergência é de 88.6%. Analisando a influência de n , nota-se claramente que quanto maior o tamanho da

amostra, maior a taxa de convergência: para $n = 10$, 96% dos casos convergiram, 97.2% para $n = 20$, 99.2% para $n = 50$, 99.8% para $n = 100$, até 100% para $n = 500$. O número médio de iterações (Tabela 4.5) também segue interpretação parecida: quanto maior o n , mais rápida é a convergência, pois a quantidade média varia de 136 ($n = 10$) para 55 ($n = 500$) iterações necessárias.

Analisando os impactos do parâmetro κ , tem-se que ele é inversamente proporcional a taxa de convergência: para $\kappa = 1$, a taxa de casos convergentes é de 95.1%, enquanto para $\kappa = 0.2$ a taxa cresce para 99.96% de sucesso. Em compensação, valores menores de κ necessitam de mais iterações, variando de 41 ($\kappa = 1$) para 164 ($\kappa = 0.2$). Ou seja, menores valores de κ aumentam a taxa de retorno sem erros, porém há um custo computacional nesse incremento de iterações, como esperado.

Para tamanhos amostrais maiores, o impacto de κ é menor, ou seja, para $n = 500$ o algoritmo converge independentemente de κ , enquanto para $n = 10$ as taxas de convergência vão de 88.6% ($\kappa = 0.2$) para 99.8% ($\kappa = 1$). Portanto, quando a amostra é pequena os resultados sugerem que deve-se utilizar valores menores de κ .

Na sequência são estudados casos específicos sem convergência para entender suas diferentes causas.

Estudo de Casos

Caso 1: $n = 10$ e $\kappa = 0.2$

Figura 4.4: Estudo de caso para uma simulação específica de não convergência com $n=10$ e $\kappa=0.2$, com a curva verdadeira traçada

Na Figura 4.4, apresenta-se o gráfico de dispersão de x por Y para um caso específico de não convergência. A linha tracejada em vermelho indica a curva para os valores verdadeiros de θ utilizados para gerar as informações, e nota-se que os pontos estão próximos da média, indicando que a limitação não é causada por alguma simulação distante do esperado.

```
fit = reg_general(Y~exp(beta1+beta2*X),
                  ~exp(sigma1+sigma2*X),
                  data=data1,
                  start = list(beta1=0.2,beta2=0.01,
                               sigma1=2,sigma2=0.01),
                  control=list(reltol = 1e-10,max_it = 5000,kappa=0.2))

## Error: Não convergiu após o número máximo de iterações
```

O algoritmo para este caso retorna a mensagem de erro: “Error: Não convergiu após o número máximo de iterações”, ou seja, a configuração utilizada de 5000 iterações não foi suficiente para convergir. Dois possíveis caminhos para solucionar esse problema são: aumentar o

limite de iterações até que o algoritmo consiga convergir ou aumentar o valor de κ para uma convergência mais rápida. A primeira solução, que consiste em aumentar o termo `max_it` até convergir, foi feita inicialmente adotando uma quantidade máxima de 10000 iterações, porém a mensagem de erro persiste. Aumentando para um máximo de 15000, o algoritmo converge no passo 11628. A segunda solução, alterando o argumento κ da função, foi utilizada com um valor de 0.8 e resultou numa convergência no passo 3198 (essas quantidades podem ser obtidas com o comando `fit$it`). Ambos os resultados apresentaram estimativas muito parecidas, nas quais só foi possível diferenciá-las a partir da sétima casa decimal. As estimativas obtidas foram: $\hat{\beta}_1 = -0.91$, $\hat{\beta}_2 = 0.011$, $\hat{\sigma}_1 = 13.95$ e $\hat{\sigma}_2 = -0.0102$.

Traçando os valores estimados, observado pela curva contínua na Figura 4.5, nota-se um bom ajuste aos pontos. Portanto, neste caso, a alteração de κ trouxe uma convergência três vezes mais rápida com resultados semelhantes. Todos os outros casos de não convergência para $\kappa = 0.2$ apresentaram o mesmo problema e puderam ser solucionados da mesma forma.

Figura 4.5: Estudo de caso para uma simulação específica de não convergência com $n=10$ e $\kappa=0.2$, com a estimativas da média representada pela linha azul contínua e a curva verdadeira tracejada

Caso 2: $n = 20$ e $\kappa = 0.4$

Figura 4.6: Estudo de caso para uma simulação específica de não convergência com $n=20$ e $\kappa=0.4$, com a curva verdadeira traçada

Outro caso específico de não convergência é observado na Figura 4.6, em que nota-se a maior parte dos pontos próximos da curva estimada para θ , porém a observação com $x \approx 700$ aparece distante do esperado.

```
fit = reg_general(
  Y~exp(beta1+beta2*X),
  ~exp(sigma1+sigma2*X),
  data=data2,
  theta_ini = list(beta1=0.2,beta2=0.01,sigma1=2,sigma2=0.01),
  control=list(reltol = 1e-10,max_it = 5000,kappa=0.4))
```

```
## Error: Não convergiu
```

O erro retornado aqui foi diferente do caso anterior, ou seja, o algoritmo divergiu antes do número de iterações máximo. Isso acontece quando as estimativas iniciais se distanciam muito do que os dados indicam. Duas soluções possíveis seriam: diminuir o parâmetro κ para que a variação em cada iteração seja menor ou utilizar valores iniciais mais adequados. Na primeira solução, ao utilizar $\kappa = 0.2$ o modelo passa a convergir e os resultados são

apresentados na Figura 4.7. As estimativas obtidas foram $\hat{\beta}_1 = 1.69$, $\hat{\beta}_2 = 0.0064$, $\hat{\sigma}_1 = 12.96$ e $\hat{\sigma}_2 = -0.0107$.

Nas simulações, todos os casos sem convergência semelhantes a esse se resolveram diminuindo o valor de κ . Na prática, nem sempre essa solução resolve o problema. Ao persistir o erro, deve-se escolher valores iniciais mais adequados aos dados, contudo esta solução pode não ser tão simples. Neste trabalho foram estudados métodos de escolha de valores iniciais automática, como o método *Particle swarm optimization* (Kennedy e Eberhart, 1995), porém os resultados ainda não foram satisfatórios o bastante para serem implementados.

Figura 4.7: Estudo de caso para uma simulação específica de não convergência com $n=20$ e $\kappa=0.4$, com as estimativas para a média (estimadas com $\kappa=0.2$) representadas pela linha azul contínua

A seguir são comparados os resultados das estimativas nos parâmetros com os valores de θ que geraram os dados, desconsiderando os casos sem convergência.

4.4 Viés e EQM dos EMV

Todos os resultados desta seção se referem às estimativas de máxima verossimilhança não corrigidas. As estimativas corrigidas são analisadas na Seção 4.5.

Figura 4.8: Boxplots das estimativas de MV para cada parâmetro do modelo (4.2), sem correção, para diferentes tamanhos de amostra. A linha horizontal indica o verdadeiro valor do parâmetro

Na Figura 4.8, são apresentadas as estimativas para cada n e cada um dos parâmetros separadamente na forma de *boxplot*, que permite ver a dispersão das estimativas. Apenas por questão de visualização dos gráficos, foram excluídos os valores extremos tanto máximos quanto mínimos (1% maiores e 1% menores). Pelo gráfico, os resultados se comportam como

esperado, ou seja, conforme aumenta-se o valor de n , mais próximas do esperado estão concentradas as estimativas e menor é a amplitude de valores. A mediana, indicada pelo traço central das caixas, parece indicar que os resultados são mais viesados para as estimativas dos parâmetros de variância σ_1 e σ_2 em amostras pequenas.

Para se ter maior precisão nas análises, são calculados alguns indicadores de erro para resumir as estimativas em uma medida de acerto. Os indicadores utilizados para os parâmetros são o erro quadrático médio relativo (EQMr) e o Viés relativo (Viés):

$$\begin{aligned} EQMr_j &= \frac{1}{M} \sum_{i=1}^M \left(\frac{\theta_j - \hat{\theta}_{j(i)}}{\theta_j} \right)^2 \quad \text{e} \\ Viés_j &= \frac{\frac{1}{M} \sum_{i=1}^M \hat{\theta}_{j(i)}}{\theta_j} - 1 \end{aligned} \quad (4.3)$$

com $j = 1, 2, \dots, p$ e $\hat{\theta}_{j(i)}$ se referindo à estimativa do parâmetro θ_j proveniente da i -ésima simulação de Monte Carlo.

A Tabela 4.6 ilustra com mais detalhes as medidas resumo das estimativas, assim como os indicadores de erro (Viés e EQMr). No geral, o valor de viés parece seguir o esperado: quanto maior o tamanho amostral, menor o viés, porém os valores de EQMr não seguem essa característica, ou seja, há casos em que aumenta-se o erro mesmo em amostras maiores. Por exemplo, para β_1 , os vieses em amostras de $n = 10$, $n = 20$ e $n = 50$ é dado respectivamente por -7.15, -5.09 e -2.69, enquanto o EQMr é de 86.24, 97.89 e 118.54. Essas divergências nos erros podem ser causadas por valores discrepantes: o máximo estimado para β_1 , por exemplo, chega a 2000, enquanto o valor esperado é -0.21. As medianas, que são medidas mais robustas, apresentam-se bem próximas do esperado, confirmando essa percepção. Por isso, esses casos discrepantes foram analisados individualmente, e notou-se que a frequência aumenta conforme cresce o valor de κ . Portanto, o valor de κ pode influenciar não apenas a taxa de convergência, como também a qualidade da estimação.

Para verificação dessa possível influência, foram geradas as medidas de Viés e EQMr para diferentes valores de κ , apresentadas na Tabela 4.7. Nela, nota-se claramente que a variação de κ influencia no erro, no qual valores menores trazem estimativas mais precisas. Ainda analisando β_1 , as medidas de EQMr para $\kappa = 0.2$ são de 6.45, 3.41, 1.92, 1.33 e 0.60, gradualmente de $n = 10$ a $n = 500$, enquanto para $\kappa = 1$ os valores são de 163.55, 217.60, 267.88, 12.51 e 0.58. Ou seja, para altos valores de κ a escala dos erros é muito maior e há

Tabela 4.6: Medidas resumo das estimativas, EQMr e viés para cada parâmetro e tamanho amostral

Parâmetros	Real	n	Mínimo	Mediana	Média	Máximo	EQMr	Viés
β_1	-0.210	10	-12.528	1.292	-0.283	702.379	86.249	-7.153
		20	-6.724	0.860	-0.270	1262.240	97.895	-5.094
		50	-2.147	0.356	-0.229	2038.657	118.541	-2.695
		100	-1.373	-0.211	-0.222	89.443	5.712	0.003
		500	-0.733	-0.213	-0.214	0.265	0.587	0.015
β_2	0.010	10	-1.646	0.006	0.010	0.033	4.250	-0.368
		20	-3.142	0.007	0.010	0.020	5.024	-0.269
		50	-4.981	0.009	0.010	0.013	6.121	-0.142
		100	-0.209	0.010	0.010	0.012	0.290	-0.001
		500	0.009	0.010	0.010	0.011	0.020	0.001
σ_1	4.200	10	-103.424	4.421	4.739	67.898	2.051	0.053
		20	-18.610	4.339	4.461	24.062	0.708	0.033
		50	-3.395	4.248	4.277	10.916	0.347	0.011
		100	-0.623	4.220	4.233	7.864	0.227	0.005
		500	2.521	4.209	4.212	6.043	0.098	0.002
σ_2	0.008	10	-0.115	0.007	0.006	0.193	1.998	-0.184
		20	-0.031	0.007	0.007	0.051	0.697	-0.080
		50	-0.005	0.008	0.008	0.025	0.335	-0.028
		100	0.001	0.008	0.008	0.020	0.216	-0.013
		500	0.005	0.008	0.008	0.011	0.093	-0.004

Tabela 4.7: EQMr e viés para cada parâmetro, tamanho amostral e parâmetro κ

Parâmetros	n	EQMr					Viés				
		κ					κ				
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
β_1	10	6.45	26.25	37.61	102.53	163.55	0.94	-1.53	-4.16	-10.53	-22.30
	20	3.41	5.58	18.93	57.45	217.60	0.34	0.17	-0.76	-5.28	-21.40
	50	1.92	1.99	1.94	10.73	267.88	0.12	0.14	0.16	-0.24	-13.97
	100	1.33	1.40	1.33	1.37	12.51	0.06	0.08	0.11	0.06	-0.30
	500	0.60	0.59	0.57	0.59	0.58	0.03	0.01	0.01	0.01	0.01
β_2	10	0.24	1.31	1.83	5.07	8.05	0.03	-0.09	-0.22	-0.53	-1.11
	20	0.11	0.25	0.96	2.95	11.17	0.01	0.00	-0.05	-0.28	-1.10
	50	0.06	0.07	0.07	0.53	13.83	0.00	0.00	0.01	-0.02	-0.73
	100	0.04	0.05	0.04	0.05	0.64	0.00	0.00	0.00	0.00	-0.02
	500	0.02	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.00	0.00
σ_1	10	1.71	1.97	2.02	2.09	2.45	0.23	0.11	0.02	0.00	-0.11
	20	0.67	0.68	0.71	0.71	0.77	0.06	0.06	0.02	0.03	-0.01
	50	0.34	0.34	0.35	0.34	0.37	0.01	0.01	0.00	0.02	0.01
	100	0.22	0.23	0.23	0.22	0.23	0.00	0.01	0.00	0.01	0.00
	500	0.10	0.10	0.10	0.10	0.10	0.00	0.00	0.00	0.01	0.00
σ_2	10	1.69	1.95	1.97	2.02	2.34	-0.35	-0.24	-0.16	-0.13	-0.02
	20	0.64	0.66	0.69	0.71	0.78	-0.11	-0.11	-0.08	-0.08	-0.03
	50	0.32	0.32	0.33	0.33	0.37	-0.03	-0.03	-0.02	-0.03	-0.03
	100	0.21	0.22	0.22	0.21	0.22	-0.01	-0.02	-0.01	-0.02	-0.01
	500	0.09	0.09	0.09	0.09	0.09	0.00	0.00	0.00	-0.01	0.00

inconsistências em relação a n , ou seja, para $n = 50$ o EQMr é muito maior do que para $n = 10$.

De maneira geral, não há inconsistências em relação a n nos EQMr para pequenos valores de κ (0.2 e 0.4), enquanto para altos valores (0.8 e 1) os erros são bem descontrolados.

Estudo de Casos

Para entender a divergência em um caso particular, foi escolhida a simulação que apresenta o maior $\hat{\beta}_1$, com $n = 50$ e $\kappa = 1$. Para comparação, estimou-se também para $\kappa = 0.2$. As estimativas obtidas dos parâmetros para $\kappa = 1$ e $\kappa = 0.2$ foram respectivamente de $\hat{\beta}_1 = 2038$, $\hat{\beta}_2 = -4.98$, $\hat{\sigma}_1 = -0.40$, $\hat{\sigma}_2 = 0.02$ e $\hat{\beta}_1 = -0.19$, $\hat{\beta}_2 = -0.01$, $\hat{\sigma}_1 = 4.29$, $\hat{\sigma}_2 = 0.008$. Os resultados no segundo caso são muito mais próximos de θ , e isso pode ser confirmado pelos valores na função de log-verossimilhança, os quais são -340.771 e -288.7426, respectivamente. O gráfico de dispersão, na Figura 4.9 com as curvas estimadas (linha cheia) ilustra como as primeiras estimativas não fazem sentido. Portanto, o uso de menores valores κ , mesmo havendo um custo de processamento maior, trouxe resultados melhores, tanto para convergência quanto para qualidade das estimativas. Nos valores mais altos, pode ser que o algoritmo esteja encontrando pontos de máximo local e por isso não esteja obtendo o valor correto do máximo da função de verossimilhança. Para estes caso específicos, o algoritmo não retorna nenhuma mensagem de erro, pois ele converge. Uma melhoria posterior para o pacote seria retornar algum aviso de que as estimativas estão potencialmente incorretas, porém o desafio é definir um critério que identifique isso.

Figura 4.9: Dados simulados para um caso de estimação discrepante com as estimativas esperadas. A linha tracejada se refere a curva verdadeira, enquanto a linha contínua verde se refere a $\kappa = 0.2$ e a linha azul é para $\kappa = 1$

A seguir, são analisados os resultados para a correção do viés de segunda ordem, a qual foi calculada pela expressão (2.10) em comparação ao estimador sem correção.

4.5 Correção de Viés

Figura 4.10: Boxplots das estimativas de MV para cada parâmetro para diferentes tamanhos de amostra. As caixas vermelhas indicam as estimativas não corrigidas, enquanto as caixas azuis indicam as estimativas corrigidas. A linha horizontal indica o verdadeiro valor de θ

Os resultados do estimador corrigido são analisados utilizando gráficos de caixa (box-plots) da mesma forma dos estimadores não corrigidos. A Figura 4.10 apresenta os resultados considerando o estimador corrigido à direita, em azul, e o estimador não corrigido, à esquerda, em vermelho. Também foram calculadas medidas de variação do EQMr e do viés relativos, dadas por:

$$EQMr_{var} = 1 - EQMr_{cor}/EQMr_{EMV} \quad e$$

$$Viés_{var} = 1 - Viés_{cor}/Viés_{EMV},$$

em que $EQMr$ e $Viés$ foram definidos na expressão (4.3). Assim, esperam-se valores negativos nesses indicadores quando a correção de viés deixa a estimativa mais próxima do valor verdadeiro.

Tabela 4.8: Comparação do Erro Quadrático Médio e Viés das correções de viés, em percentual, em relação a estimativa de Máxima Verossimilhança

Parâmetro	n	$EQMr_{EMV}$	$EQMr_{cor}$	$EQMr_{var}$	$Viés_{EMV}$	$Viés_{cor}$	$Viés_{var}$
$\beta_1 = -0.21$	10	8624.86	8980965.90	104028.79	-715.27	103827.19	14415.73
	20	9789.54	565305.69	5674.59	-509.38	9358.44	1737.23
	50	11854.14	4799.52	-59.51	-269.46	-62.43	-76.83
	100	571.22	402.92	-29.46	0.30	3.07	917.39
	500	58.69	58.66	-0.05	1.53	1.00	-34.37
$\beta_2 = 0.01$	10	425.00	438034.67	102965.80	-36.76	5080.26	13721.30
	20	502.41	29069.69	5686.08	-26.86	479.02	1683.35
	50	612.06	249.44	-59.25	-14.21	-3.36	-76.34
	100	28.98	20.00	-30.99	-0.13	0.08	-36.53
	500	1.96	1.96	-0.05	0.05	0.04	-28.64
$\sigma_1 = 4.2$	10	205.15	187.24	-8.73	5.27	3.42	-35.12
	20	70.82	66.85	-5.60	3.31	1.47	-55.60
	50	34.70	34.00	-2.02	1.14	0.42	-63.34
	100	22.70	22.50	-0.86	0.47	0.12	-73.66
	500	9.81	9.79	-0.17	0.22	0.16	-29.73
$\sigma_2 = 0.008$	10	199.80	181.77	-9.02	-18.40	-6.94	-62.28
	20	69.70	65.28	-6.35	-7.98	-1.53	-80.88
	50	33.53	32.70	-2.48	-2.82	-0.29	-89.71
	100	21.56	21.33	-1.07	-1.32	-0.07	-94.59
	500	9.29	9.27	-0.24	-0.37	-0.12	-67.12

Na Figura 4.10, os *boxplots* para correção de viés indicam uma melhor aproximação dos valores medianos para o valor verdadeiro de θ , representado pela linha horizontal, porém a Tabela 4.8 indica resultados de correção de viés bem ruins, principalmente para os parâmetros β_1 e β_2 . Essa diferença de interpretação pode ser novamente explicada por pontos discrepantes nas estimações, como já foi verificado na Tabela 4.6 da seção anterior.

Assim como concluído na Seção 4.4, existe influência de κ na qualidade das estimações e, portanto, analisou-se os resultados para cada valor de κ . Os indicadores de variação de $EQMr$ e $Viés$ estão na Tabela 4.9 (para obter a tabela completa, com os valores de $EQMr$

e Viés de ambos os estimadores, consultar o Apêndice C).

Tabela 4.9: Variação do impacto da correção de viés nos EQMr e Viés dos estimadores de máxima verossimilhança, medidos para diferentes parâmetros, n e κ .

Parâmetros	n	Variação EQMr (em %)					Variação Viés (em %)				
		κ					κ				
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
$\beta_1 = -0.21$	10	-3.07	-4.06	167.63	1551.72	127752.17	-22.90	-40.36	-96.18	471.27	24856.18
	20	-1.12	-2.52	21.40	2588.79	5864.82	-35.60	-69.59	-90.55	951.00	1963.53
	50	-0.44	-0.49	-0.51	-9.08	-59.66	-43.48	-37.02	-31.35	4.82	-77.02
	100	-0.23	-0.26	-0.30	-0.25	-31.15	-39.67	-33.38	-24.18	-42.48	-81.86
	500	-0.08	-0.05	-0.06	-0.04	-0.04	-15.68	-44.76	-34.98	-76.32	-56.47
$\beta_2 = 0.01$	10	-3.79	-3.93	178.27	1555.83	126535.12	-20.72	-41.90	-99.24	460.08	24344.86
	20	-0.98	-2.90	22.83	2618.84	5874.16	-30.56	-95.12	-99.44	916.98	19444.43
	50	-0.38	-0.43	-0.45	-9.40	-59.38	-38.03	-31.35	-27.31	-3.75	-77.02
	100	-0.19	-0.23	-0.26	-0.22	-31.75	-33.93	-26.51	-20.14	-36.18	-78.26
	500	-0.07	-0.05	-0.05	-0.03	-0.04	-13.69	-31.39	-28.71	-78.65	-45.25
$\sigma_1 = 4.2$	10	-8.87	-8.86	-8.37	-9.29	-8.41	-17.67	-12.16	-81.37	7677.19	-0.62
	20	-4.79	-5.11	-5.22	-6.39	-6.35	-34.33	-29.92	-84.46	-60.90	99.59
	50	-1.78	-1.82	-1.79	-1.99	-2.67	-91.02	-49.14	43.18	-45.28	-45.46
	100	-0.87	-0.86	-0.80	-0.86	-0.91	196.05	-42.98	552.83	-35.21	-69.72
	500	-0.18	-0.18	-0.16	-0.18	-0.17	-58.97	-18.44	146.39	-13.59	-52.26
$\sigma_2 = 0.008$	10	-9.76	-9.11	-8.68	-9.47	-8.45	-38.24	-46.36	-72.56	-88.20	256.06
	20	-5.82	-6.14	-5.95	-7.06	-6.64	-61.52	-60.31	-87.39	-83.48	35.50
	50	-2.18	-2.26	-2.14	-2.52	-3.13	-94.46	-82.54	-86.72	-74.80	-92.66
	100	-1.03	-1.11	-0.93	-1.18	-1.14	-76.86	-78.43	-51.33	-66.53	-98.67
	500	-0.22	-0.26	-0.20	-0.30	-0.21	-89.89	-54.85	-93.97	-40.39	-92.13

Nota-se que a correção de viés também é influenciada pela variação de κ . Para $\kappa = 0.2$ e $\kappa = 0.4$, as estimativas já estavam mais próximas do esperado e a correção de viés reduziu ainda mais o EQMr e Viés para todos os parâmetros. Além disso, a diminuição do erro é menor conforme aumenta o tamanho da amostra, como já era esperado (ver Patriota e Lemonte, 2009). Por exemplo, para β_1 e $\kappa = 0.2$ a correção de viés diminui em 3.06% o erro quadrático médio relativo em amostras com $n = 10$, enquanto essa diminuição é de 0.078% para $n = 500$.

Já para altos valores de κ , que induzem a estimativas mais imprecisas (Tabela 4.7), a correção aumenta o erro, principalmente para as menores amostras. Por exemplo, para $\kappa = 0.6$ e $n = 10$, a correção aumenta o EQMr em 178%, mas para $n = 50$, a correção já passa a diminuir o EQMr, ainda que pouco, em 0.44%.

Conclui-se que a correção de viés funciona muito bem quando o estimador, de fato, é o de máxima verossimilhança. Para casos em que a estimativa não é boa a correção tende a se afastar ainda mais dos valores reais.

Na Seção seguinte são realizados testes de hipóteses, com correção de Skovgaard para a estatística de teste, considerando as mesmas simulações desta seção.

4.6 Teste de hipótese

Teste de razão de verossimilhanças

Os testes considerados nesta seção foram realizados por meio da estatística da razão de verossimilhanças. Foram realizados diferentes testes para cada nível de significância, dados por $\alpha = 1\%, 5\%, 10\%$.

As hipóteses nula e alternativa são:

$$\mathcal{H}_0 : \theta = (-0.21, 0.01, 4.2, 0.008) \quad \times \quad \mathcal{H}_1 : \theta \neq (-0.21, 0.01, 4.2, 0.008),$$

sendo a hipótese nula correspondente aos verdadeiros valores sob os quais as simulações foram conduzidas. É esperado que a distribuição dos valores-p dos testes tenha distribuição uniforme sob $\mathcal{H}_0$. Para avaliar se isso de fato ocorre, são gerados histogramas dos valores-p calculados para cada κ e n indicados na Figura 4.11.

Para cálculo do teste de razão de verossimilhanças no pacote, utiliza-se a função `likelihood_ratio` aplicada ao objeto do modelo, como por exemplo:

```
test=likelihood_ratio(
  fit,
  parameters = list(
    beta1=-0.21,beta2=0.01,
    sigma1=4.2,sigma2=0.008))

test

## $LR
## [1] 20.19409
##
## $p_value
## [1] 0.0004572097
```

Para cálculo a correção de Skovgaard, utiliza-se o argumento `correction=TRUE`. Caso $\mathcal{H}_0$ esteja relacionada a um subconjunto de θ , basta entrar com a lista dos parâmetros que possuam restrição. Por exemplo, caso queira testar apenas $\beta_1 = -0.21$ e $\beta_2 = 0.01$, deve-se fornecer o argumento escrito da forma `list(beta1=-0.21,beta2=0.01)`. Dessa forma, o pacote entende que os parâmetros σ_1 e σ_2 não têm restrição sob $\mathcal{H}_0$.

Figura 4.11: Histogramas dos valores-p nos testes da razão de verossimilhanças sob $\mathcal{H}_0$

Os histogramas indicam que quanto maior o tamanho amostral mais uniforme tende a ser a distribuição do valor-p. Outras amostras com $n = 1000$ e $n = 5000$ foram estudadas para avaliar a taxa de rejeição e a tendência é se aproximar cada vez mais da taxa de rejeição esperada (α). Para $n = 10$ e $n = 20$ os valores-p estão indicando maior rejeição do que é

Tabela 4.10: Taxa de rejeição de $\mathcal{H}_0$ para cada n , κ e α para estatísticas da razão de verossimilhanças

κ	n	α			% Sucesso
		1%	5%	10%	
0.2	10	4.7%	13.5%	21.2%	99.7%
	20	1.7%	7.2%	13.2%	100%
	50	1.1%	6%	11.2%	100%
	100	1%	5%	9.7%	100%
	500	1.4%	5.5%	10.5%	100%
0.4	10	5.3%	13%	21.1%	98.4%
	20	1.7%	7.8%	14.1%	99.8%
	50	1.4%	6.3%	11.5%	100%
	100	1%	5.5%	11.4%	100%
	500	1.3%	5.6%	10.8%	100%
0.6	10	4.6%	14.3%	23.2%	94.6%
	20	2%	7.4%	14.2%	98.5%
	50	1.4%	5.8%	11.3%	99.9%
	100	1.2%	5.9%	10.5%	100%
	500	1%	5%	8.9%	100%
0.8	10	4.8%	13.9%	21.9%	90.5%
	20	1.6%	7.2%	13.8%	94.1%
	50	1.4%	5.6%	10.9%	99%
	100	1%	4.8%	9.6%	100%
	500	0.8%	5.1%	9.8%	100%
1.0	10	6.6%	14.7%	23.2%	83.2%
	20	1.9%	8.8%	15.1%	87.5%
	50	1.2%	5.2%	10.2%	95.9%
	100	1.4%	5%	11.1%	99.2%
	500	1.1%	4.6%	10%	100%

esperado e a correção de Skovgaard é utilizada para melhorar a aproximação.

Conforme apresentado na Seção 4.4, algumas estimativas com convergência não são EMV. Em parte delas, a função de log-verossimilhança sob $\mathcal{H}_0$ é maior do que o valor a função de log-verossimilhança para a estimativa $LR < 0$ e conseqüentemente não é possível realizar o TRV. Nesses casos, a função retorna uma mensagem de erro. Na simulação, esses casos foram descartados e a taxa de casos restantes é apresentada na última coluna da Tabela 4.10. Nota-se que para $\kappa \geq 0.80$ a taxa de sucesso é menor, indicando problemas na estimação, como já havia sido observado na Seção 4.4.

A seguir são apresentados os resultados com o TRV corrigido.

Correção de Skovgaard

A Figura 4.12 traz os histogramas dos valores-p, da mesma forma que foi feita nos testes sem correção, enquanto a Figura 4.13 compara a curva de densidade estimada para o teste corrigido e não corrigido. Percebe-se pelas densidades, principalmente para $n = 10$, uma melhor aproximação para a distribuição uniforme, representada no gráfico pela linha

horizontal que cruza o eixo y no valor 1, sinalizando que a correção de Skovgaard (Seção 2.3) é eficaz.

Figura 4.12: Histogramas dos valores-p nos testes da razão de verossimilhanças com correção de Skovgaard sob $\mathcal{H}_0$

Figura 4.13: Curva de densidade estimada dos valores- p nos testes com correção de Skovgaard sob $\mathcal{H}_0$ e sem correção

Outra forma de avaliar os resultados da correção é comparar as taxas de rejeição para os diferentes níveis de significância. O teste corrigido é eficaz se sua taxa de rejeição for mais próxima do nível de significância do que o teste sem correção. A Tabela 4.11 indica que o teste LR^* (equação (2.13)) altera bastante os resultados, principalmente nos tamanhos amostrais

menores: para $n = 10$, $\alpha = 10\%$ e $\kappa = 0.2$, 21.19% dos casos foram rejeitados para o teste não corrigido, enquanto para o teste corrigido a taxa é 13.49%. Para $\kappa = 1$, as taxas são 21.78% e 14.07%, respectivamente. Para tamanhos amostrais grandes, a variação das taxas de rejeição é mínima.

Tabela 4.11: Taxa de rejeição de $\mathcal{H}_0$ para cada n , κ e α tanto corrigidos, como não

κ	n	$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 10\%$		% Sucesso
		LR	LR*	LR	LR*	LR	LR*	
0.2	10	4.7%	2.4%	13.5%	7.5%	21.2%	13.5%	92.8%
	20	1.7%	1.8%	7.2%	6.2%	13.2%	10.7%	99.5%
	50	1.1%	1.1%	6%	5.2%	11.2%	9.9%	99.9%
	100	1%	1%	5%	4.6%	9.7%	9.4%	99.9%
	500	1.4%	1.3%	5.5%	5.6%	10.5%	10.3%	100%
0.4	10	5.2%	2.6%	12.8%	8.9%	20.9%	13.6%	90.4%
	20	1.7%	1.5%	7.8%	6.9%	14.1%	12.6%	99.2%
	50	1.4%	1.1%	6.3%	6.1%	11.5%	11.2%	99.8%
	100	1%	1.2%	5.5%	5.6%	11.4%	10.9%	100%
	500	1.3%	1.4%	5.6%	5.4%	10.8%	10.6%	100%
0.6	10	4.4%	2.7%	13.8%	7.5%	22.4%	13.4%	88%
	20	2%	1.5%	7.4%	6.2%	14.1%	11.2%	97.7%
	50	1.4%	1.2%	5.8%	5.6%	11.3%	11.1%	99.8%
	100	1.2%	1.3%	5.9%	5.2%	10.5%	10.2%	100%
	500	1%	0.8%	5%	4.8%	8.9%	9.3%	100%
0.8	10	4.6%	2.2%	13.3%	6.8%	21%	13.1%	84.5%
	20	1.6%	1.7%	7%	5.6%	13.5%	10.3%	93.5%
	50	1.4%	1.2%	5.6%	4.7%	10.8%	10%	98.7%
	100	1%	0.9%	4.8%	4.5%	9.6%	9.2%	99.9%
	500	0.8%	0.8%	5.1%	5.1%	9.8%	10%	100%
1.0	10	6.2%	3%	13.8%	8.6%	21.8%	14.1%	77%
	20	1.9%	1.5%	8.5%	6.7%	14.5%	12.8%	86.9%
	50	1.2%	1.1%	5.2%	4.5%	10.1%	9.3%	95.8%
	100	1.4%	1.6%	5%	5.1%	11.1%	10%	99.2%
	500	1.1%	1.1%	4.6%	4.7%	10%	9.9%	100%

Em alguns casos, não foi possível obter a estatística corrigida, fazendo com que o algoritmo retorne uma mensagem de erro. Isso acontece quando o valor de ρ na equação (2.14) é menor do que zero. Para contabilizar esses casos, a Tabela 4.11 traz a taxa de sucesso da correção, definida pelo total de casos em que houve estatística corrigida sem erro, dividido pelo total de casos. Analisando as falhas com maior detalhe, encontrou-se basicamente dois tipos de inconsistências: a primeira se dá quando o valor da função de verossimilhança sob $\mathcal{H}_0$ é maior do que o valor para a estimativa, concluindo que não era de fato uma estimativa de máxima verossimilhança; o segundo tipo se dá quando os determinantes das matrizes $\tilde{U}'$ ou $\tilde{\tilde{J}}$ são negativos. Neste trabalho não se encontrou justificativa da causa deste segundo tipo de problema, que poderá ser analisado com mais detalhamento em estudos futuros.

Os resultados comprovam menor taxa de sucesso nas combinações com estimativas com

maior erro, ou seja, pequenas amostras com altos valores de κ . Por exemplo, para $n = 10$ e $\kappa = 1$, 92.8% de casos obtiveram sucesso contra 77% de casos para $n = 10$ e $\kappa = 0.2$. Portanto, a causa da correção de Skovgaard não poder ser calculada pode ser consequência de imprecisões nas estimações, como em casos exemplificados na Seção 4.4.

Figura 4.14: Discrepâncias relativas do valor- p para os testes de razão de verossimilhança, separados por n e κ .

Para mensurar o impacto da correção ponderado pelo valor-p assintótico obtido por LR^* , calcula-se uma medida de discrepância que relaciona este valor-p com o valor-p exato, aproximado pela distribuição empírica de LR obtido via Simulação de Monte Carlo. O valor-p exato indica a média de casos simulados que apresentaram valor da estatística acima da observada. A medida de discrepância é dada por: $(vp_{exato} - vp_{assint})/vp_{assint}$, sendo vp_{exato} e vp_{assint} os valores-p exatos e assintóticos, respectivamente. Quanto mais próximo do 0, melhor é a aproximação assintótica (Melo *et al.*, 2017).

Na Figura 4.14 são exibidas as comparações da discrepância pelo valor-p assintótico, comparando o teste corrigido com o teste não corrigido para cada n e κ . Nota-se, novamente, que a correção de Skovgaard é bastante eficaz para amostras pequenas nas quais há maior descolamento entre as linhas. Aparentemente, a influência de κ não é muito grande aqui, lembrando que os casos em que não houve cálculo da correção foram descartados.

Embora os resultados das correções estejam compatíveis com o que se espera, os resultados obtidos em Melo *et al.* (2017) indicam correções mais impactantes. Uma comparação em um cenário mais parecido pode ser feito para verificar se a causa da divergência é justificada apenas pelas diferenças de modelo, ou se houve alguma divergência nos cálculos.

Capítulo 5

Conclusões

5.1 Considerações finais

Neste trabalho foi desenvolvido com sucesso o pacote `generalReg` em linguagem R, para ajuste de um modelo de regressão normal univariado com parametrização geral. As simulações de Monte Carlo, utilizadas como forma de validar tanto a implementação do algoritmo quanto a eficácia dos cálculos, trouxeram resultados esperados: quanto maior o tamanho amostral, maior a taxa de retorno do algoritmo sem erros, mais próxima do valor verdadeiro é a estimativa e menos passos são necessários até a convergência.

Esta dissertação apontou também a relevância do parâmetro de controle κ (ver equação (2.6)) na obtenção das estimativas de máxima verossimilhança. Valores menores de κ exigem mais iterações até a convergência do algoritmo, entretanto, as taxas de convergência, EQMr e Viés relativo das estimativas são consideravelmente melhores. Observou-se que a escolha de κ é mais influente nos resultados para amostras pequenas ($n \leq 20$ principalmente) justamente nos casos em que a estimação tende a ser mais complicada.

A correção de viés de segunda ordem para o EMV mostrou-se eficaz quando a estimativa é de fato a de máxima verossimilhança, isto é, em casos com tamanho amostral grande ou estimados com pequenos valores de κ . Em simulações de pequenas amostras com valores de κ próximos de 1, os quais induzem a estimações imprecisas, a estimativa corrigida aumenta ainda mais o EQMr e o Viés relativo.

Já a correção de Skovgaard em testes da razão de verossimilhanças, indicou-se bastante relevante para pequenos tamanhos amostrais (principalmente $n = 10$), aproximando mais os valores-p assintóticos dos valores-p exatos. No geral, para os casos em que foi possível calcular

a correção, a influência do parâmetro κ não é relevante nos testes. O impacto de κ é notado para os casos nos quais não é possível calcular a correção de Skovgaard, causada por valores negativos de ρ (equação (2.14)), ou seja, pequenas amostras com alto valor de κ apresentam uma taxa de falha maior em calcular a correção. Portanto, para todas as funcionalidades implementadas, o parâmetro κ é fundamental na qualidade dos resultados, principalmente quando a amostra é pequena.

5.2 Próximos Passos

Estudos com outras simulações variando a quantidade de covariáveis, utilizando relações mais complexas de parametrização e da influência da variação das estimativas iniciais nos resultados, podem ser feitos para analisar o comportamento do algoritmo em mais cenários. Algumas melhorias podem ser feitas visando retornar mensagens de erros mais específicas e emitir mais avisos para o usuário entender com mais detalhes quando houver algum resultado que possivelmente não é de máxima verossimilhança.

Estudou-se também neste trabalho, formas de estimações que não dependam de uma escolha de valores iniciais. Métodos como a função `optim` no pacote base do R e otimização por enxames (Kennedy e Eberhart, 1995) foram testados. Os maiores desafios foram minimizar o tempo de processamento e conseguir uma solução mais geral possível, pois funcionavam apenas quando os parâmetros estavam reduzidos a um intervalo fixado.

A implementação proposta nesta dissertação trata apenas de distribuições normais univariadas. Porém, a classe definida em Patriota e Lemonte (2011) considera uma variedade de modelos muito maior: modelos de regressão multivariada com distribuições elípticas, nas quais incluem-se distribuições tais como normal, t de Student, Cauchy, exponencial potência. Próximos trabalhos podem ser desenvolvidos para ampliar a estrutura construída aqui.

Durante este trabalho iniciou-se uma implementação de modelos de regressão multivariada, porém não se mostrou eficiente em termos de desempenho, já que o algoritmo necessitava de cálculos em listas de matrizes, ao invés de cálculos em vetores, tornando a convergência muito lenta. Assim, a escolha foi aprofundar-se em um escopo mais limitado de modelos, mas que garanta eficiência na utilização para um usuário. A forma do usuário definir a parametrização em modelos de regressão multivariados também foi um desafio, principalmente em estruturas muito complexas, em que seria necessário preencher os parâmetros para cada elemento do vetor de média e da matriz de variância, podendo ser inviável para o usuário em

problemas com amostra muito grande.

Referências Bibliográficas

- Aldrich, J. (1997) R. A. Fisher and the making of maximum likelihood. *Statistical Science*, **12** 162-176
- Barndorff-Nielsen, O.E. (1986) Inference on full or partial parameters, based on the standardized signed log likelihood ratio. *Biometrika*, **73** 307-322
- Box, M.J. (1971) Bias in nonlinear estimation (with discussion). *Journal of the Royal Statistical Society B.*, **33** 171-201
- Brazzale, A.R.; Davison, A.C. (2008) Accurate parametric inference for small samples. *Statistical Science*, **23** 465-484
- Chernoff, H. (1954) On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, **25** 573-578
- Cordeiro, G.M.; McCullagh, P. (1991) Bias correction in generalized linear models. *Journal of the Royal Statistical Society B.*, **53** 629-643
- Cordeiro, G.M.; Paula, G.A. (1989) Fitting non-exponential family nonlinear models in GLIM by using the offset facility. *Lecture Notes in Statistics*, **57** 105-114
- Cox, D.R.; Hinkley, D.V. (1974). Theoretical Statistics. *London: Chapman and Hall*
- Cox, D.R.; Snell, E. (1968) A general definition of residuals (with discussion). *Journal of the Royal Statistical Society*, **30** 248-275
- de Carvalho C.; Gomes D.; de Souza J.; Agoulmine N. (2011) Multiple linear regression to improve prediction accuracy in wsn data reduction. *7th Latin American Network Operations and Management Symposium*, 1-8
- Drton, M. (2009) Likelihood ratio tests and singularities. *The Annals of Statistics*, **37** 979-1012

- Eubank, R.L.; Thomas, W. (1993) Detecting heteroscedasticity in nonparametric regression. *Journal of the American Statistical Association*, **55** 145-155
- Ferrari, S.L.P.; Pinheiro, E.C. (2014) Small-sample likelihood inference in extremevalue regression models. *Journal of Statistical Computation and Simulation*, **84** 582-595
- Fisher, R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, **52** 399-433
- Fraser, D.A.S.; Reid, N.; Wu, J. (1999) A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika*, **86**, 249-264
- Guolo, A. (2012) Higher-order likelihood inference in meta-analysis and meta-regression. *Statistics in Medicine*, **31**, 313-327
- Hilary L.S. (1967) The historical development of the Gauss linear model. *Biometrika*, **54**, 1-24
- Kennedy, J.; Eberhart, R. (1995) Particle Swarm Optimization. *Proceedings of IEEE International Conference on Neural Networks IV* 1942-1948
- Kulathinal, S.B.; Kuulasmaa, K.; Gasbarra, D. (2002) Estimation of an errors-in-variables regression model when the variances of the measurement error vary between the observations. *Statistics in Medicine*, **21**, 1089-1101
- Lee, S.Y.; Lu, B.; Song, X.Y. (2006) Assessing local influence for nonlinear structural equation models with ignorable missing data. *Computational Statistics and Data Analysis*, **50** 1356-1377
- Lemonte, A. J.; Ferrari, S.L.P. (2011) Signed likelihood ratio tests in the BirnbaumSaunders regression model. *Journal of Statistical Planning and Inference*, **141** 1031-1040
- Levine, S.S.; Prietula, M.J. (2013) Open Collaboration for Innovation: Principles and Performance. *Organization Science*, 2-3
- Loose, L.H. (2019) Condições de regularidade para o modelo de regressão com parametrização geral. 82 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo
- Magnus, J.R.; Neudecker, H. (1988) Matrix differential calculus with applications. *Wiley: New York*
- Melo, T.F.N.; Ferrari, S.L.P.; Patriota, A.G. (2017) Improved hypothesis testing in a gene-

- ral multivariate elliptical model. *Journal of Statistical Computation and Simulation*, **87**, 1416–1428
- Melo, T.F.N.; Ferrari, S.L.P.; Patriota, A.G. (2018) Improved estimation in a general multivariate elliptical model. *Brazilian Journal of Probability and Statistics*, **32(1)** 44–68
- Muenchen, R. (2017) The Popularity of Data Analysis *software*. <http://r4stats.com/>
- Neyman, J.; Pearson, E. S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, **231** 289–337
- Patriota, A.G.; Lemonte, A.J. (2009) Bias correction in a multivariate normal regression model with general parameterization. *Statistics & Probability Letters*, **79** 1655–1662
- Patriota, A.G.; Lemonte, A.J. (2011) Multivariate elliptical models with general parameterization. *Statistical Methodology*, **8** 389–400
- Peña, E.A.; Rohatgi, V.K.; Szekely, G.J. (1992) On the non-existence of ancillary statistics. *Statistics and Probability Letters*, **15** 357–360
- R Development Core Team (2014) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, ISBN 3-900051-07-0
- Seber, G.A.F.; Wild, C.J. (1989) Nonlinear Regression. *New York: Wiley*
- Severini, T.A. (2000) Likelihood Methods in Statistics. *Oxford University Press*
- Sharma, M.; Agarwal, R. (2003) Maximum likelihood method for parameter estimation in non-linear models with below detection data. *Environmental and Ecological Statistics*, **10**, 445–454
- Skovgaard, I.M. (2001) Likelihood asymptotics. *Scandinavian Journal of Statistics*, **28**, 3–32
- Transtrum, M.K.; Machta, B.B.; Sethna, J.P. (2010) Why are nonlinear fits to data so challenging? *American Physical Society*
- Vasconcellos, K.L.P.; Cordeiro, G.M (1997) Approximate bias for multivariate nonlinear heteroscedastic regressions. *Brazilian Journal of Probability and Statistics*, **11**, 141–159
- Vasconcellos, K.L.P.; Cordeiro, G.M. (2000) Bias corrected estimates in multivariate Student t regression models. *Communications in Statistics, Theory and Methods*, **29** 797–822
- Vasconcellos, K.L.P.; CribariNeto, F. (2005) Improved maximum likelihood estimation in a

- new class of beta regression models. *Brazilian Journal of Probability and Statistics*, **19** 13-31
- Wei, B.C. (1998) Exponential Family Nonlinear Models. *Singapore: Springer*
- White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48** 817-838

Apêndice

Apêndice A: Cálculos para a correção de Skovgaard

Para cálculo dos elementos presentes em (2.14), são necessárias algumas matrizes auxiliares. As decomposições são:

$$\begin{aligned}\hat{\ell}' &= \hat{R}^T \hat{\Sigma}^{-1} \hat{z}, & \tilde{\ell}' &= \hat{R}^T \tilde{\Sigma}^{-1} \tilde{z}, & \tilde{U}' &= \tilde{T}^T \tilde{\Sigma}^{-1} \hat{R}^T, \\ \tilde{J} &= \tilde{T}^T \tilde{\Sigma}^{-1} \tilde{D} + \tilde{G}, & \hat{J} &= \hat{T}^T \hat{\Sigma}^{-1} \hat{D} + \hat{G} & \text{e} & \tilde{U} = F^T H \tilde{s},\end{aligned}\tag{5.1}$$

em que, como definido na Seção 2.3, “ $\hat{\cdot}$ ” e “ $\tilde{\cdot}$ ” representam expressões calculadas para $\hat{\theta}$ e $\tilde{\theta}$ respectivamente. Os termos definidos posteriormente por “ $\approx$ ” são relativos ao termo $(\tilde{\theta}; \tilde{\theta}, a)$. As matrizes $\tilde{D}, F, H$ e $\tilde{s}$ já foram definidas em (2.3), enquanto as matrizes G, T e R são matrizes auxiliares, de dimensões $p \times p$, $n \times p$ e $n \times p$ respectivamente, definidas a seguir: Cada elemento das matrizes G e T , dado por G_{rs} e $T_{i(r)}$, é dado por:

$$\begin{aligned}T_{i(r)} &= d_{i(r)} + z_i^T \Sigma^{-1} V_{i(r)}, \\ G_{rs} &= \sum_{i=1}^n [B_{i(r)} A_{i(s)} + E_{i(rs)}],\end{aligned}$$

com

$$\begin{aligned}A_{i(r)} &= -\Sigma_i V_{i(r)} \Sigma_i, \\ B_{i(r)} &= -z_i d_{i(r)}^T - \frac{1}{2} V_{i(r)}, \\ E_{i(rs)} &= -\frac{1}{2} [A_{i(sr)} (\Sigma_i - z_i^T z_i)] - z_i^T \Sigma_i d_{i(sr)},\end{aligned}$$

em que $A_{i(sr)} = \partial A_{i(s)} / \partial \theta_r = -2A_{i(r)} V_{i(s)} \Sigma_i^{-1} - \Sigma_i^{-1} C_{i(sr)} \Sigma_i^{-1}$ e os elementos já definidos no Capítulo 2: $z_i = Y_i - \mu_i$, $V_i = \partial \text{vec}(\Sigma_i) / \partial \theta^T$, $d_{i(r)} = \partial \mu_i / \partial \theta_r$, $d_{i(sr)} = \partial^2 \mu_i / \partial \theta_s \partial \theta_r$, $C_{i(rs)} = \partial^2 \text{vec}(\Sigma_i) / \partial \theta_r \partial \theta_s$.

Para a matriz $\hat{R}$, presente na decomposição da derivada da informação de Fisher observada e da função de log-verossimilhança, cada elemento ($\hat{R}_{i(r)}$) é dado por:

$$\hat{R}_{i(r)} = \hat{P}_{i(r)} a_i + \hat{d}_{i(r)},$$

sendo $\hat{P}_{i(r)} = \partial \hat{P}_i / \partial \hat{\theta}_r$ e $\hat{d}_{i(r)} = \partial \hat{\mu}_i / \partial \hat{\theta}_r$. A notação matricial é dada da forma $\hat{R}^T = (\hat{R}_1^T, \dots, \hat{R}_n^T)$, $\hat{R}_i = (\hat{R}_{i(1)}, \dots, \hat{R}_{i(p)})$, $z^* = -(z_1^T, \dots, z_n^T)^T$, $T^T = (T_1^T, \dots, T_n^T)$, $T_i = (T_{i(1)}, \dots, T_{i(p)})$. Os resultados são análogos aos apresentados acima, ou seja,

$$\tilde{G}_{rs} = \sum_{i=1}^n [\tilde{B}_{i(r)} \tilde{A}_{i(s)} + \tilde{E}_{i(rs)}],$$

$$\tilde{T}_{i(r)} = \tilde{d}_{i(r)} + a_i^T \tilde{P}_i^T \tilde{\Sigma}^{-1} \tilde{V}_{i(r)},$$

$$\tilde{B}_{i(r)} = -\tilde{P}_i a_i \tilde{d}_{i(r)} - \frac{1}{2} \tilde{V}_{i(r)} \quad \text{e}$$

$$\tilde{E}_{i(rs)} = -\frac{1}{2} [\tilde{A}_{i(sr)} (\tilde{\Sigma}_i - (a_i \tilde{P}_i)^T (a_i \tilde{P}_i))] - a_i^T \tilde{P}_i^T \tilde{\Sigma}_i \tilde{d}_{i(sr)}.$$

Apêndice B: Funcionalidades da classe `reg_general`

Para ter acesso, por exemplo, ao valor do vetor escore estimado de um objeto `x` da classe `reg_general`, basta escrever: `x$score`

1. `parameters`: vetor com as estimativas do vetor θ ;
2. `fitted.values`: vetor com as estimativas da variável resposta na tabela **data** referenciada no ajuste;
3. `var`: matriz estimada de covariâncias Σ para o conjunto **data** referenciado no ajuste;
4. `score`: vetor escore U estimado;
5. `Fisher`: informação de Fisher esperada;
6. `Fisher_obs`: matriz com a informação de Fisher observada;
7. `loglike`: valor da função de log-verossimilhança obtido com as estimativas;
8. `it`: iteração em que o algoritmo convergiu;
9. `data`: tabela de dados inputados no algoritmo;

10. `target`: vetor de respostas utilizado;
11. `call`: expressão exata usada para chamar a função `reg_general`;
12. Os itens a seguir indicam funções usadas para gerar as principais matrizes auxiliares utilizada nos cálculos. Todas são parametrizadas para um θ .
 - `function_mu`: vetor $\mu(\hat{\theta})$;
 - `function_sigma`: matriz diagonal $\Sigma(\hat{\theta})$;
 - `function_l`: valor do log da função de verossimilhança l ;
 - `function_F`: matriz F definida em (2.3);
 - `function_H`: matriz H definida em (2.3);
 - `function_s`: vetor s definido em (2.3);
 - `function_D`: matriz $n \times p$ dada por $\partial\mu/\partial\theta^T$;
 - `function_V`: matriz $n \times p$ dada por $V = \partial\text{diag}(\Sigma(\hat{\theta}))/\partial\theta^T$;
 - `function_G`: matriz definida em (2.8);
 - `function_C`: matriz $n \times (p \times p)$ dada por $\partial^2\text{diag}(\Sigma(\hat{\theta}))/\partial(\theta^T)^2$;
 - `function_D2`: matriz $n \times (p \times p)$ dada por $\partial^2\mu/\partial(\theta^T)^2$.

Apêndice C: Tabelas auxiliares

Tabela C1: Erro Quadrático Médio e viés relativos para EMVs, corrigido e não, para cada tamanho amostral para os parâmetros β_1 e β_2 do modelo de regressão apresentado na equação (4.2)

Parâmetros	κ	n	$EQMr_{EMV}$	$EQMr_{cor}$	$EQMr_{var}$	$Viés_{EMV}$	$Viés_{cor}$	$Viés_{var}$
β_1	0.2	10	135.4%	131.2%	-3.1%	94.1%	72.6%	-22.9%
		20	71.5%	70.7%	-1.1%	33.8%	21.8%	-35.6%
		50	40.4%	40.2%	-0.4%	11.6%	6.6%	-43.5%
		100	27.9%	27.9%	-0.2%	6.5%	3.9%	-39.7%
		500	12.6%	12.6%	-0.1%	3.3%	2.8%	-15.7%
	0.4	10	551%	529%	-4.1%	-153.4%	-91.5%	-40.4%
		20	117.2%	114.2%	-2.5%	17.3%	5.3%	-69.6%
		50	41.7%	41.5%	-0.5%	13.9%	8.8%	-37%
		100	29.5%	29.4%	-0.3%	7.8%	5.2%	-33.4%
		500	12.5%	12.4%	0%	1.2%	0.6%	-44.8%
	0.6	10	790%	2114%	167.6%	-416.4%	15.9%	-96.2%
		20	397%	483%	21.4%	-75.9%	7.2%	-90.6%
		50	40.8%	40.6%	-0.5%	16%	11%	-31.4%
		100	28%	27.9%	-0.3%	10.7%	8.1%	-24.2%
		500	12%	12%	-0.1%	1.5%	1%	-35%
	0.8	10	2153%	35563%	1552%	-1052.8%	6015%	471%
		20	1206%	32439%	2589%	-528.4%	5554%	951%
		50	225%	205%	-9.1%	-23.7%	-24.8%	4.8%
		100	28.7%	28.6%	-0.2%	6.1%	3.5%	-42.5%
		500	12.4%	12.4%	0%	0.7%	0.2%	-76.3%
	1.0	10	3434%	4391031%	127752%	-2229.7%	556445%	24856%
		20	4570%	272573%	5865%	-2140.3%	44165%	1964%
		50	5626%	2269%	-59.7%	-1397.2%	-321.1%	-77%
		100	263%	180.9%	-31.2%	-29.7%	-5.4%	-81.9%
		500	12.2%	12.2%	0%	0.9%	0.4%	-56.5%
β_2	0.2	10	0.2%	0.2%	-3.8%	2.9%	2.3%	-20.7%
		20	0.1%	0.1%	-1%	1.1%	0.8%	-30.6%
		50	0.1%	0.1%	-0.4%	0.4%	0.2%	-38%
		100	0%	0%	-0.2%	0.2%	0.1%	-33.9%
		500	0%	0%	-0.1%	0.1%	0.1%	-13.7%
	0.4	10	1.3%	1.3%	-3.9%	-8.9%	-5.2%	-41.9%
		20	0.2%	0.2%	-2.9%	0.4%	0%	-95.1%
		50	0.1%	0.1%	-0.4%	0.5%	0.3%	-31.3%
		100	0%	0%	-0.2%	0.3%	0.2%	-26.5%
		500	0%	0%	0%	0%	0%	-31.4%
	0.6	10	1.8%	5.1%	178.3%	-22.1%	0.2%	-99.2%
		20	1%	1.2%	22.8%	-4.6%	0%	-99.4%
		50	0.1%	0.1%	-0.4%	0.5%	0.4%	-27.3%
		100	0%	0%	-0.3%	0.4%	0.3%	-20.1%
		500	0%	0%	0%	0.1%	0%	-28.7%
	0.8	10	5.1%	84%	1556%	-53.4%	299%	460%
		20	2.9%	80.2%	2619%	-28.3%	287%	917%
		50	0.5%	0.5%	-9.4%	-1.5%	-1.5%	-3.7%
		100	0%	0%	-0.2%	0.2%	0.1%	-36.2%
		500	0%	0%	0%	0%	0%	-78.7%
	1.0	10	8.1%	10198%	126535%	-111.4%	27225%	24345%
		20	11.2%	667%	5874%	-110.5%	2258%	1944%
		50	13.8%	5.6%	-59.4%	-72.5%	-16.7%	-77%
		100	0.6%	0.4%	-31.8%	-1.7%	-0.4%	-78.3%
		500	0%	0%	0%	0%	0%	-45.3%

Tabela C2: Erro Quadrático Médio e viés relativos para EMVs, corrigido e não, para cada tamanho amostral para os parâmetros σ_1 e σ_2 do modelo de regressão apresentado na equação (4.2)

Parâmetros	κ	n	$EQMr_{EMV}$	$EQMr_{cor}$	$EQMr_{var}$	$Viés_{EMV}$	$Viés_{cor}$	$Viés_{var}$
σ_1	0.2	10	716%	653%	-8.9%	22.6%	18.6%	-17.7%
		20	280%	266%	-4.8%	6%	4%	-34.3%
		50	141.1%	138.6%	-1.8%	0.8%	0.1%	-91%
		100	94.3%	93.5%	-0.9%	0.1%	-0.3%	196.1%
		500	41%	40.9%	-0.2%	0.1%	0%	-59%
	0.4	10	829%	756%	-8.9%	10.5%	9.2%	-12.2%
		20	287%	272%	-5.1%	5.8%	4.1%	-29.9%
		50	142.6%	140%	-1.8%	1.4%	0.7%	-49.1%
		100	95.7%	94.9%	-0.9%	0.8%	0.5%	-43%
		500	41.4%	41.4%	-0.2%	0.4%	0.3%	-18.4%
	0.6	10	850%	779%	-8.4%	2.2%	0.4%	-81.4%
		20	297%	282%	-5.2%	2.3%	0.4%	-84.5%
		50	146.1%	143.4%	-1.8%	0.3%	-0.4%	43.2%
		100	95.8%	95%	-0.8%	-0.1%	-0.4%	553%
		500	40.8%	40.8%	-0.2%	0%	0%	146.4%
	0.8	10	876%	795%	-9.3%	0%	-2%	7677%
		20	300%	281%	-6.4%	3.3%	1.3%	-60.9%
		50	144.4%	141.5%	-2%	1.7%	1%	-45.3%
		100	94.4%	93.6%	-0.9%	1%	0.7%	-35.2%
		500	40.9%	40.8%	-0.2%	0.5%	0.4%	-13.6%
	1.0	10	1028%	942%	-8.4%	-11.2%	-11.1%	-0.6%
		20	324%	304%	-6.4%	-1.4%	-2.8%	99.6%
		50	154.6%	150.4%	-2.7%	1.4%	0.8%	-45.5%
		100	96.5%	95.6%	-0.9%	0.5%	0.1%	-69.7%
		500	41.8%	41.7%	-0.2%	0.1%	0.1%	-52.3%
σ_2	0.2	10	1.3%	1.2%	-9.8%	-35.3%	-21.8%	-38.2%
		20	0.5%	0.5%	-5.8%	-10.9%	-4.2%	-61.5%
		50	0.3%	0.3%	-2.2%	-2.7%	-0.1%	-94.5%
		100	0.2%	0.2%	-1%	-1%	0.2%	-76.9%
		500	0.1%	0.1%	-0.2%	-0.3%	0%	-89.9%
	0.4	10	1.6%	1.4%	-9.1%	-23.5%	-12.6%	-46.4%
		20	0.5%	0.5%	-6.1%	-10.6%	-4.2%	-60.3%
		50	0.3%	0.3%	-2.3%	-3%	-0.5%	-82.5%
		100	0.2%	0.2%	-1.1%	-1.6%	-0.3%	-78.4%
		500	0.1%	0.1%	-0.3%	-0.4%	-0.2%	-54.8%
	0.6	10	1.6%	1.4%	-8.7%	-15.8%	-4.3%	-72.6%
		20	0.6%	0.5%	-6%	-7.5%	-1%	-87.4%
		50	0.3%	0.3%	-2.1%	-2.2%	0.3%	-86.7%
		100	0.2%	0.2%	-0.9%	-0.8%	0.4%	-51.3%
		500	0.1%	0.1%	-0.2%	-0.2%	0%	-94%
	0.8	10	1.6%	1.5%	-9.5%	-13.2%	-1.6%	-88.2%
		20	0.6%	0.5%	-7.1%	-7.9%	-1.3%	-83.5%
		50	0.3%	0.3%	-2.5%	-3.5%	-0.9%	-74.8%
		100	0.2%	0.2%	-1.2%	-1.9%	-0.6%	-66.5%
		500	0.1%	0.1%	-0.3%	-0.6%	-0.4%	-40.4%
	1.0	10	1.9%	1.7%	-8.5%	-2.1%	7.5%	256%
		20	0.6%	0.6%	-6.6%	-2.6%	3.5%	35.5%
		50	0.3%	0.3%	-3.1%	-2.7%	-0.2%	-92.7%
		100	0.2%	0.2%	-1.1%	-1.3%	0%	-98.7%
		500	0.1%	0.1%	-0.2%	-0.3%	0%	-92.1%

Figura C1: Variação do impacto da correção de viés nos EQMr dos estimadores de máxima verossimilhança para diferentes parâmetros, n e κ

Figura C2: Variação do impacto da correção de viés no viés relativo dos estimadores de máxima verossimilhança para diferentes parâmetros, n e κ