Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Pós-Graduação em Estatística

Ensaios sobre modelos de regressão com dispersão variável

Tatiene Correia de Souza

TESE DE DOUTORADO

Recife
11 de fevereiro de 2011

Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Tatiene Correia de Souza

Ensaios sobre modelos de regressão com dispersão variável

Trabalho apresentado ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Estatística.

Orientador: *Francisco Cribari-Neto*

Recife
11 de fevereiro de 2011

Catalogação na fonte
Bibliotecária Joana D'Arc L. Salvador, CRB 4-572

Souza, Tatiene Correia de.
Ensaios sobre modelos de regressão com dispersão variável / Tatiane Correia de Souza - Recife: O Autor, 2011.
xxii, 116 p.; fig. tab.

Orientador: Francisco Cribari Neto.
Tese (Doutorado) - Universidade Federal de Pernambuco. CCEN. Estatística, 2011.

Inclui bibliografia e apêndice.

1.Análise de regressão. 2.Eleição. 3.Inteligência. 4.Religião. I. Cribari Neto, Francisco (orientador). II. Título.

519.536 (22.ed.) MEI 2011-014

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

11 de fevereiro de 2011
(data)

Nós recomendamos que a tese de doutorado de autoria de

Tatiene Correia de Souza

intitulada

"Ensaios sobre modelos de regressão com dispersão variável"

seja aceita como cumprimento parcial dos requerimentos para o grau de Doutora em Estatística.

Coordenador da Pós-Graduação em Estatística

Banca Examinadora:

Francisco Cribari Neto orientador

Raydonal Ospina

Klaus Leite Pinto Vasconcellos

Silvia Lopes de Paula Ferrari (USP)

Verônica Maria Cadena Lima (UFBA)

Este documento será anexado à versão final da dissertação.

Dedico esta tese...

Ao meu orientador, Francisco Cribari Neto,
a Nina, amor da minha vida e
aos meu pais,
com muito carinho.

Agradecimentos

Inicialmente, gostaria de agradecer ao meu orientador e mentor, pessoa que, além de mim, mais se dedicou a esta tese, Francisco Cribari–Neto, pelo imenso apoio intelectual, seriedade, dedicação, estímulo, confiança e incentivo constante ao aprofundamento dos meus estudos e da minha qualificação profissional.

Aos meus pais, pelo amor incondicional, orientação e paciência permanente, aos quais dedico todas as minhas conquistas.

À minha querida e amada Nina, a quem sinto um amor incondicional, por ser minha fonte de energia e alegria e por me tornar uma pessoa mais feliz a cada dia.

A Antonio Carlos, pelo carinho, compreensão, paciência, dedicação, estímulo, confiança, por ser uma pessoa tão doce e especial, por me fazer sorrir até mesmo nos momentos mais difíceis, por estar presente em todos os momentos. Por seu imenso amor incondicional, a minha eterna gratidão.

A toda minha família, em especial, às minhas avós, Zizi e Rilsa, ao meu avô Severino, aos meus tios, Aninha e Herton, e à minha prima Katy, pelo incentivo, confiança e por estarem sempre presentes na minha vida.

Aos meus irmãos que tanto amo, Helinho e Humberto.

À melhor pessoa que conheci em toda minha vida, Tarciana, por seu companheirismo inquestionável, por sua lealdade, por contornar os problemas que sugiram durante a nossa caminhada, por toda sua dedicação, por ser tão solidária e caridosa com todos, por ter cuidado de mim em todos os momentos que precisei, pelos momentos de alegria e de cumplicidade, enfim, por ser a minha grande amiga.

À minha querida amiga e secretária do Programa de Pós-Graduação em Estatística, Valéria, por ser motivadora, eficiente, organizada, prestativa, companheira e alegre, por ser alguém tão especial que tornou os meus dias mais felizes e divertidos e, principalmente, por ser alguém por quem sinto profunda admiração, carinho e respeito.

Ao meu querido amigo e grande parceiro Fábio, por estar sempre disposto a ajudar-me, por toda sua simplicidade e por sua enorme amizade.

À pessoa mais especial que conheci em Salvador, Lia, por me envolver com muito amor e carinho.

Ao meu amigo Getúlio, por sua amizade, carinho e companheirismo.

Ao meu amigo Klaus, pelos momentos de confidências, por sua grande amizade e por estar sempre disposto a me ajudar, minha eterna gratidão.

A Verônica, pela amizade sincera, carinho e lealdade.

Aos meus amigos Fernando, Luz e Marcela, pelos momentos inesquecíveis e divertidos que passamos juntos.

A Alejandro, por quem sinto profunda admiração, carinho e respeito.

A Raydonal, pelas valiosas sugestões.

Aos colegas do doutorado, em especial, a Tarciana, Fábio, Hemílio, Jane, Abrãao, Manoel, Luz, Laércio, Basílides e Izabel, pelas experiências e momentos compartilhados.

Aos colegas do mestrado, em especial, a Fernando, Marcela, Marcelo e Lutemberg pelas amizades conquistadas.

Aos amigos que conquistei na Plataforma, em especial aos meus professores e amigos Marcelo e Michele, que me ajudaram a alcançar meus limites físicos.

Aos meus grandes amigos, Tarciana, Valéria, Cribari, Klaus, Getúlio, Ernando, Manu, Lia e Verônica, pelo amor, carinho e apoio que sempre me ofereceram.

Às minhas queridas e eternas professoras Cristina e Claudia, que sempre estiveram dispostas a me ajudar.

A Renato J. Cintra, por tudo que me ensinou.

Aos professores do Departamento de Estatística da Universidade Federal de Pernambuco pela socialização dos seus conhecimentos, em especial, aos professores Cristina Raposo, Claudia Regina, Francisco Cribari, Getúlio Amaral, Klaus Vasconcellos, Manoel Senna e Renato J. Cintra.

Aos colegas do Departamento de Estatística da Universidade Federal da Bahia pela compreensão, apoio e incentivo dados durante o doutorado. Em especial aos professores Denise, Gecynalda, Gilênio, Giovana, Leila, Lia, Maristela e Verônica pela confiança em mim depositada e

pelo apoio que foi essencial para conclusão deste curso.

À FAPESB e à CAPES pelo apoio financeiro.

Aos participantes da banca examinadora pelas sugestões.

Quando mais nada resistir que valha
a pena de viver e a dor de amar
e quando nada mais interessar
(nem o torpor do sono que se espalha),

quando pelo desuso da navalha
a barba livremente caminhar
e até Deus em silêncio se afastar
deixando-te sozinho na batalha

a arquitetar na sombra a despedida
do mundo que te foi contraditório,
lembra-te que afinal te resta a vida

com tudo que é insolvente e provisório
e de que ainda tens uma saída:
entrar no acaso e amar o transitório.
—CARLOS PENA FILHO

Resumo

A análise de regressão é uma das técnicas estatísticas mais usadas. Nesta tese são abordados modelos de regressão linear e modelos de regressão beta ambos com dispersão variável. Na primeira etapa da tese, apresentamos o modelo de regressão linear, e na segunda, o modelo de regressão beta. Sob as pressuposições usuais estabelecidas para o modelo linear, a estimação dos parâmetros é usualmente feita pelo método de mínimos quadrados ordinários (MQO). Esse método fornece estimadores com propriedades desejáveis, como não-viciosidade, consistência e eficiência. Entretanto, sob heteroscedasticidade, os estimadores de MQO tornam-se ineficientes e o estimador usual de sua matriz de covariâncias não é consistente. Vários autores propuseram estratégias para estimar de forma consistente a matriz de covariâncias dos estimadores dos parâmetros do modelo de regressão, geralmente baseadas em resíduos de MQO. Estes resíduos, porém, podem ser fortemente influenciados pela presença de pontos de alavanca. Avaliamos os comportamentos dos estimadores HC0, HC3 e HC4 da matriz de covariâncias do estimador de MQO quando resíduos oriundos de regressões robustas (menor mediana dos quadrados, mínimos quadrados podados e mínimos quadrados ponderados) são usados em substituição aos resíduos de MQO, no modelo de regressão com e sem restrição sobre os parâmetros. Os resultados revelaram que o teste construído a partir do estimador HC0 apresenta melhor desempenho quando resíduos oriundos de regressão robusta são utilizados em substituição aos resíduos de MQO, na presença de pontos de alavanca. O desempenho do teste baseado nesse estimador melhora significativamente quando estimação restrita é utilizada, principalmente com resíduos robustos. A mesma conclusão foi obtida na maioria dos cenários estudados sobre o teste baseado no estimador HC3, contudo considerando estimação irrestrita. O teste baseado no estimador HC4 apresenta desempenho superior quando resíduos oriundos do estimador MQO e mínimos quadrados ponderados são utilizados.

Diferentemente do modelo de regressão linear, o modelo de regressão beta que foi proposto por Ferrari & Cribari–Neto (2004) possui aplicabilidade na modelagem de variáveis do tipo taxas ou proporções. A resposta média é relacionada com um preditor linear que envolve covariáveis e parâmetros de regressão desconhecidos através de uma função de ligação, sendo ainda o modelo indexado por um parâmetro de precisão. Smithson & Verkuilen (2006) apresentaram o modelo de regressão beta em que há estrutura de regressão para o parâmetro de precisão. Em termos práticos, contudo, há certa dificuldade em modelar a precisão. Considerando tal dificuldade, um dos nossos interesses consiste em propor estimadores do tipo sanduíche para modelos de regressão beta em que a estrutura de regressão para o parâmetro de dispersão (o recíproco da precisão) é negligenciada. Adicionalmente, consideramos o caso em que há estrutura de regressão para dispersão, mas a modelamos de forma incorreta através dos preditores e das funções de ligação dos submodelos da média e dispersão. Através de simulações de Monte Carlo, nós

avaliamos os desempenhos dos testes baseados nos estimadores sanduíche e comparamos com os desempenhos dos testes z que utilizam os estimadores usuais da matriz de covariâncias. Para ilustrar nossos resultados, apresentamos uma aplicação a dados reais.

Considerando o modelo de regressão beta com dispersão variável, um outro objetivo consiste em explicar a diferença entre as proporções de votos válidos do presidente Lula nos segundos turnos das eleições de 2006 e 2002. Adicionalmente, calculamos os impactos dos gastos em programas assistenciais e do crescimento da economia sobre o resultado da eleição presidencial de 2006. Comparando os gastos com programas assistenciais em 2006 e 2002, vale ressaltar que em 2006 o gasto em tais programas foi muito maior do que em 2002. Nós estimamos que, se os gastos com programas assistenciais em 2006 fossem mantidos nos níveis de 2002, haveria uma redução de aproximadamente 7 milhões na votação do ex-presidente Lula. Adicionalmente, estimamos que sem o crescimento da economia, haveria uma redução na votação do ex-presidente Lula de cerca de 2 milhões de votos.

Por fim, objetivamos modelar a proporção de ateus e a proporção de pessoas que não acreditam que religião é importante para suas vidas cotidianas em diferentes países. Para isso, consideramos algumas covariáveis, tais como quociente de inteligência, renda nacional bruta ajustada pela paridade no poder de compra, predominância de muçulmanos, grau de abertura da economia e expectativa de vida em cada uma das nações consideradas. Estimamos o impacto do quociente de inteligência sobre a concentração média de ateus e de pessoas que não valorizam religião. Os resultados revelaram que há forte associação positiva entre inteligência e descrença religiosa. Dado que as duas variáveis dependentes consideradas são proporções, aqui também utilizamos o modelo de regressão beta.

Palavras-chave: Ateísmo, eleição presidencial, estimadores sanduíche, heteroscedasticidade, inteligência, modelo de regressão beta, modelo de regressão linear, programas assistenciais, resíduos robustos e religião.

Abstract

Regression analysis is a commonly used tool for analyzing data. This thesis deals with variable dispersion linear and beta regressions. In the first part of the thesis, we consider the linear regression model and in what follows we work with the class of beta regressions. In the linear regression model, parameter estimation is typically carried out by ordinary least squares (OLS). The estimator is unbiased, consistent and efficient under the standard assumptions. Under heteroskedasticity (i.e., different error variances) the estimator remains unbiased and consistent, but becomes inefficient. The usual estimator of its covariance matrix becomes inconsistent. Several consistent covariance matrix estimators have been proposed in the literature. They are typically based on OLS residuals. Such residuals, nonetheless, can be quite sensitive to leverage data points. We construct alternative estimators that employ restricted and unrestricted robust residuals. The numerical evidence shows that, under leveraged data, the HC0 test performs better when the associated estimator uses robust residuals, especially those obtained by imposing the null hypothesis when estimating the regression parameters. The HC3 test tends to perform better when it is based on unrestricted robust residuals. The HC4 test displays superior finite sample performance when the standard error is obtained using unrestricted OLS or weighted least squares residuals.

Unlike the linear regression model, the class of beta regressions is useful for modeling responses that continuously assume values on the standard unit interval, $(0,1)$, such as rates and proportions. The mean response is related to a linear predictor, which includes regressors and parameters, through a (monotonic, twice-differentiable) link function. The model is also indexed by a precision parameter. Smithson & Verkuilen (2006) used a general class of beta regressions in which there is a submodel for the precision parameter, i.e., this parameter is not constant for all observations and is modelled in a fashion that is similar to the mean parameter. In practice, however, it is typically easier to model mean than dispersion effects. We propose sandwich-like covariance matrix estimators for the mean regression parameter estimators that can be used to perform hypothesis testing on these parameters without estimating a submodel for the dispersion effects. In our numerical analysis, the finite-sample behavior of such quasi-z tests is compared to those of standard z tests. The latter are carried out under both correct and incorrect model specification. Several sources of misspecifications are considered, such as incorrectly chosen link functions and incorrect choice of covariates. Overall, the results favor the quasi-z tests, especially when one takes into consideration that modeling dispersion effects can introduce misspecification into the model. We also present an application to a real (nonsimulated) data set.

Next, we use the beta regression model to analyze data related to the outcome of the 2006 presidential election in Brazil. Our response is the difference between the proportions of valid

votes for Luiz Inácio Lula da Silva in the 2006 and 2002 elections. We estimate the impacts of the increased spending on assistance programs and of economic growth on the election outcome. We conclude that the candidate Luiz Inácio Lula da Silva would lose approximately 7 million votes in the 2006 runoff election had the spending on assistance programs remained at the 2002 level. Under no economic growth between the two elections, we estimate that he would lose around 2 million votes.

A second empirical analysis used the beta regression model to analyze an international cross-sectional dataset on religion disbelief (atheism). The response is the proportion of atheists in over one hundred countries. We also model the proportion of people in different countries who do not feel that religion is important for their daily lifes. We use several covariates, such as average intelligence, gross national income (adjusted by purchasing power parity), an indicator as to whether the country is muslim, life expectation and a measure of the economic openess to trade. Our chief goal is to estimate the impact of intelligence on religion disbelief. We conclude that such an impact is highly positive, i.e., greater intelligence positively impacts the proportion on atheists in a given country.

Keywords: Assistance programs, atheism, beta regression model, heteroskedasticity, intelligence, linear regression model, presidential election, religion, robust residuals and sandwich estimators.

Sumário

1 Introdução **1**

2 Inferência sob heteroscedasticidade de forma desconhecida **5**
2.1 Introdução 5
2.2 Modelos e estimadores 6
2.3 Novos estimadores consistentes 9
2.4 Estatísticas quasi-*t* e quasi-*F* 11
2.5 Avaliação numérica I 12
2.6 Avaliação numérica II 22
2.7 Aplicação 25
2.8 Conclusões 33

3 Inferência no modelo de regressão beta **35**
3.1 Introdução 35
3.2 O modelo de regressão beta 36
3.3 Os estimadores sanduíche 42
3.4 Avaliação numérica 45
3.5 Aplicação 53
3.6 Conclusões 55

4 Eleição presidencial de 2006 no Brasil **57**
4.1 Introdução 57
4.2 O modelo de regressão beta 59
4.2.1 Medidas de diagnóstico e adequabilidade para modelos de regressão beta 63
4.3 Descrição dos dados 64
4.4 Especificação do modelo 67
4.4.1 Diferença entre as proporções de votos válidos do ex-presidente Lula 68
4.5 Conclusões 76

5 Ateísmo e inteligência **77**
5.1 Introdução 77
5.2 Descrição dos dados 78
5.3 Modelo de regressão beta 86
5.4 Especificação dos modelos 89
5.5 Conclusões 101

6 Considerações finais **103**
6.1 Conclusões 103
6.2 Trabalhos futuros 105

A Poderes dos testes quasi-*t* **107**

B Teste escore **111**

Referências bibliográficas **113**

Lista de Figuras

2.1 Ajustes do modelo de regressão $y_t = \beta_0 + \beta_1 x_t + e_t$, $t = 1, \ldots, 20$. 11

4.1 Histograma e gráfico *box-plot*. 69
4.2 Gráfico dos resíduos e da probabilidade normal com envelopes simulados. 70
4.3 Gráficos das distâncias de Cook e da alavancagem generalizada. 71

5.1 Mapa do quociente médio de inteligência nas nações. 82
5.2 Mapa da proporção de ateus nas nações. 83
5.3 Mapa da renda per capita ajustada pela paridade do poder de compra. 84
5.4 Gráfico da probabilidade normal com envelopes simulados. 91
5.5 Distâncias de Cook. Proporção de ateus em 124 nações. 92
5.6 Impacto do quociente de inteligência sobre a prevalência de ateus I. 94
5.7 Impacto do quociente de inteligência sobre a prevalência de ateus II. 94
5.8 Impacto da inteligência sobre a prevalência de ateus em três cenários. 97
5.9 Distâncias de Cook e gráfico de envelopes simulados. 98
5.10 Impacto da inteligência sobre a variável *RELIG*. 99
5.11 Impacto do quociente de inteligência sobre a prevalência de ateus III. 100

Lista de Tabelas

2.1 Taxas de rejeição de testes quasi-*t* baseados nos estimadores HC0_k. 16
2.2 Taxas de rejeição de testes quasi-*t* baseados nos estimadores HCR0_k. 17
2.3 Taxas de rejeição de testes quasi-*t* baseados nos estimadores HC3_k. 18
2.4 Taxas de rejeição de testes quasi-*t* baseados nos estimadores HCR3_k. 19
2.5 Taxas de rejeição de testes quasi-*t* baseados nos estimadores HC4_k. 20
2.6 Taxas de rejeição de testes quasi-*t* baseados nos estimadores HCR4_k. 21
2.7 Poderes de testes quasi-*t* baseados nos estimadores HCj_k. 23
2.8 Poderes de testes quasi-*t* baseados nos estimadores HCRj_k. 24
2.9 Taxas de rejeição de testes quasi-*F* baseados nos estimadores HC0_k. 26
2.10 Taxas de rejeição de testes quasi-*F* baseados nos estimadores HCR0_k. 27
2.11 Taxas de rejeição de testes quasi-*F* baseados nos estimadores HC3_k. 28
2.12 Taxas de rejeição de testes quasi-*F* baseados nos estimadores HCR3_k. 29
2.13 Taxas de rejeição de testes quasi-*F* baseados nos estimadores HC4_k. 30
2.14 Taxas de rejeição de testes quasi-*F* baseados nos estimadores HCR4_k. 31
2.15 Inferência quasi-*t*, *p*-valores. 33
2.16 Erros-padrão para b_2. 33

3.1 Descrição dos cenários considerados no estudo de simulação de Monte Carlo. 47
3.2 Taxas de rejeição sob $\mathcal{H}_0 : \beta_2 = 0$ dos testes z_v, z_f e τ_j. 49
3.3 Poderes dos teste z_v, z_f e τ_j. 51
3.4 Taxas de rejeição sob $\mathcal{H}_0 : \beta_2 = 0$ dos testes τ_{h_s}. 52
3.5 Poderes dos testes híbridos τ_{h_s}. 52
3.6 Taxas de rejeição sob $\mathcal{H}_0 : \beta_2 = 0$ dos testes τ_{h_s}. 53
3.7 Poderes dos testes híbridos τ_{h_s}. 53
3.8 Estimativas pontuais e *p*-valores (dispersão variável). 54
3.9 Estimativas pontuais e *p*-valores (dispersão fixal). 55
3.10 *p*-valores de testes quasi-*z* baseados em diferentes estimadores sanduíche. 55

4.1 Descrição das variáveis utilizadas. 65
4.2 Estatísticas descritivas das variáveis utilizadas. 66
4.3 Estimativas pontuais e *p*-valores. 70
4.4 Estimativas pontuais e *p*-valores (estimador sanduíche S_2) 72
4.5 Impactos estimados da diferença entre gastos nos programas assistenciais. 74
4.6 Intervalo de confiança bootstrap ao nível de 95% de confiança (gastos). 74
4.7 Impactos estimados da diferença do PIB per capita em 2006 e 2002. 75
4.8 Intervalos de confiança bootstrap ao nível de 95% de confiança (PIB). 75

5.1 Descrição das variáveis utilizadas nas modelagens. 79
5.2 Estatísticas descritivas das variáveis utilizadas. 81
5.3 Correlações entre $ATEU$ e as covariáveis. 85
5.4 Correlações entre $RELIG$ e as covariáveis. 85
5.5 Estimativas pontuais e p-valores. Proporção de ateus nas 124 nações. 90
5.6 Variações percentuais nas estimativas dos parâmetros. 93
5.7 Impactos estimados da inteligência sobre a proporção média de ateus. 95
5.8 Erros-padrão bootstrap para os impactos estimados da inteligência. 96
5.9 Intervalos de confiança bootstrap para o impacto da inteligência. 96
5.10 Estimativas da proporções médias de ateus. 98
5.11 Estimativas pontuais e p-valores (modelo de regressão beta com dispersão fixa 100

A.1 Poderes de testes quasi-t baseados nos estimadores HCj_k. 108
A.2 Poderes de testes quasi-t baseados nos estimadores HCRj_k. 109

B.1 Tamanho do teste escore. 112
B.2 Poder do teste escore. 112

Capítulo 1

Introdução

O modelo clássico de regressão teve origem nos trabalhos de astronomia elaborados por Gauss no período de 1809 a 1821; trata-se de uma técnica adequada para estudar a relação entre variáveis explicativas e uma variável resposta. Comumente utilizamos o estimador de mínimos quadrados ordinários na estimação dos parâmetros do modelo de regressão. Em particular, ao se utilizar o modelo de regressão clássico é usual assumir que os erros têm variância constante, caso em que se diz haver homoscedasticidade. Contudo, tal suposição não é verificada em muitas situações práticas. Quando as variâncias dos erros não são constantes, isto é, na presença de heteroscedasticidade, a estimativa de MQO da matriz de covariâncias dos estimadores dos parâmetros de regressão torna-se imprecisa, podendo conduzir a conclusões incorretas. Isso ocorre já que os testes *t* associados requerem estimativas confiáveis das variâncias dos estimadores dos parâmetros. Uma possível solução é usar estimadores consistentes da matriz de covariâncias que funcionam bem tanto sob homoscedasticidade quanto sob heteroscedasticidade (Long & Ervin, 2000). Dado que em um grande número de aplicações práticas os erros são heteroscedásticos, é importante considerar estimadores mais confiáveis para variância do estimador de MQO.

Em termos práticos há várias situações em que a variável resposta é do tipo taxa ou proporção, ou seja, está definida no intervalo $(0,1)$. Nessas situações o modelo de regressão linear não é adequado. Além da imposição de que a função de esperança condicional é linear e de que a variância não é necessariamente função da média, há o fato de que proporções no intervalo $(0,1)$ não estão definidas sobre todos os reais, que é o suporte da distribuição normal.

A distribuição beta é uma das distribuições mais utilizadas para modelar experimentos aleatórios que produzem resultados no intervalo $(0,1)$. Ferrari & Cribari–Neto (2004) propuseram uma parametrização alternativa para densidade beta. Nessa parametrização, a resposta média é relacionada com um preditor linear que envolve covariáveis e parâmetros de regressão desconhecidos através de uma função de ligação, sendo ainda o modelo indexado por um parâmetro de dispersão (ou seu recíproco, um parâmetro de precisão). A modelagem proposta por Ferrari & Cribari–Neto (2004) não requer transformação na variável resposta e permite interpretação dos parâmetros em termos da variável original (sem ser transformada). Smithson & Verkuilen (2006) apresentaram o modelo de regressão beta em que há estrutura de regressão para o parâmetro de dispersão. Em termos práticos, contudo, há certa dificuldade em modelar a dispersão. Nesta tese, apresentamos um modelo de regressão beta em que o parâmetro de dispersão varia com as observações, havendo assim uma estrutura heteroscedástica. Vale ressaltar aqui que, mesmo que a dispersão seja constante ao longo das observações, a variância da variável resposta não é constante, pois depende das médias desconhecidas que variam de acordo com a estrutura de regressão. Assim sendo, o conceito de heteroscedasticidade no modelo de

regressão beta difere daquele empregado no modelo de regressão linear.

Esta tese tem como objetivo principal a apresentação de resultados teóricos e práticos na classe dos modelos de regressão com dispersão variável. No que se restrige ao enfoque teórico, o objetivo desta tese consiste em propor estimadores para matrizes de covariâncias no modelo de regressão com dispersão variável. Para propor tais estimadores, consideramos o modelo de regressão linear e o modelo de regressão beta. Sobre o modelo de regressão linear, temos que vários autores propuseram estratégias para estimar de forma consistente a matriz de covariâncias dos estimadores dos parâmetros do modelo de regressão, geralmente baseadas em resíduos de MQO, todavia esses resíduos podem ser fortemente influenciados pela presença de pontos de alavanca. Os novos estimadores aqui propostos fazem uso de resíduos oriundos de regressão robusta no modelo de regressão linear com e sem restrição nos parâmetros. Em relação ao modelo de regressão beta, nós consideramos o cenário em que há estrutura de regressão para média e para a dispersão. Em termos práticos, há dificuldade em modelar a dispersão, portanto, o nosso enfoque consiste em propor estimadores para modelos de regressão beta em que a estrutura de regressão para o parâmetro de dispersão é negligenciada.

Em relação aos resultados práticos apresentados nesta tese, direcionamos o nosso objetivo em dois diferentes enfoques. No primeiro, o nosso interesse consiste em explicar a diferença entre as proporções de votos recebidos pelo ex-presidente Lula nas eleições presidenciais de 2006 e 2002, e no segundo, a nossa meta é explicar a proporção de ateus e de pessoas que não consideram religião importante para suas vidas cotidianas em diferentes nações. A grande motivação para estudar a eleição presidencial de 2006 foi baseada na mudança do perfil da votação do candidato reeleito em 2006, Luiz Inácio Lula da Silva. Vale ressaltar que apesar da mudança no perfil dos eleitores, a votação do ex-presidente Lula foi praticamente a mesma nas eleições de 2002 e 2006, em torno de 61% dos votos válidos. Nas eleições presidenciais anteriores, Lula foi bem sucedido em regiões mais urbanizadas e desenvolvidas do Brasil, enquanto que, em 2006, a base eleitoral do candidato migrou para as regiões menos desenvolvidas do país. Para tal mudança, várias hipóteses foram levantadas, entre elas o programa Bolsa-Família e o crescimento econômico. Considerando tais hipóteses, o interesse aqui consiste em explicar a diferença da proporção de votos do ex-presidente Lula nas eleições de 2006 e 2002. Adicionalmente, temos interesse em estimar o impacto do programa Bolsa-Família e do crescimento da economia sobre as votações 2006 e 2002. Considerando que a variável resposta é uma proporção, utilizamos o modelo de regressão beta. Por fim, em relação ao segundo enfoque prático apresentado nesta tese, a motivação é baseada na evidência de que há um declínio da crença religiosa nos últimos 150 anos ou mais e ao mesmo tempo tem ocorrido aumento da inteligência média da população. O declínio da crença religiosa tem sido demostrado pelas estatísticas de frequência à igreja e de crença em Deus registradas em pesquisas de opinião. As pesquisas mostraram que nos Estados Unidos 95,5% das pessoas acreditavam em Deus em 1948 (Argyle, 1958), mas em 2004 este percentual caiu para 89,5% (Zuckerman, 2007). Com base nessa informação, o nosso objetivo é modelar a proporção de ateus e também a proporção de pessoas que não acreditam que religião é importante para suas vidas cotidianas em diferentes países.

A presente tese encontra-se organizada em seis capítulos. Os Capítulos 2, 3, 4 e 5 são autosuficientes, ou seja, cada capítulo pode ser lido separadamente, em qualquer ordem. Desta forma, algumas notações e resultados são apresentados mais de uma vez. No Capítulo 2, apre-

sentamos os modelos de regressão linear irrestrito e restrito. Sob as pressuposições usuais estabelecidas para o modelo linear, a estimação dos parâmetros é usualmente feita pelo método de mínimos quadrados ordinários (MQO). Esse método fornece estimadores com propriedades desejáveis, como não-viciosidade, consistência e eficiência. Entretanto, sob heteroscedasticidade, os estimadores de MQO tornam-se ineficientes e o estimador usual de sua matriz de covariâncias não é consistente. Vários autores propuseram estratégias para estimar de forma consistente a matriz de covariâncias dos estimadores dos parâmetros do modelo de regressão, geralmente baseadas em resíduos de MQO. Estes resíduos, porém, podem ser fortemente influenciados pela presença de pontos de alavanca. O objetivo do presente capítulo é avaliar o comportamento dos estimadores HC0, HC3 e HC4 da matriz de covariâncias do estimador de MQO quando resíduos oriundos de regressões robustas (menor mediana dos quadrados, mínimos quadrados podados e mínimos quadrados ponderados) são usados em substituição aos resíduos de MQO, no modelo de regressão com e sem restrição sobre os parâmetros.

No Capítulo 3, introduzimos o modelo de regressão beta com dispersão variável. No entanto, em termos práticos, há certa dificuldade em modelar a dispersão, portanto, um dos objetivos é propor estimadores do tipo sanduíche para modelos de regressão beta em que a estrutura de regressão para o parâmetro de dispersão é negligenciada. Consideramos também o caso em que há estrutura de regressão para a dispersão, mas a modelamos de forma incorreta através dos preditores e das funções de ligação dos submodelos da média e da dispersão. Através de simulações de Monte Carlo nós avaliamos, em amostras de tamanho finito, os desempenhos de testes z e quasi-z através dos tamanhos e dos poderes dos testes. No primeiro teste, consideramos o estimador usual da matriz de covariâncias e, no segundo teste, utilizamos estimadores do tipo sanduíche. Para tal avaliação consideramos dados gerados de duas formas distintas. Na primeira, geramos dados com dispersão variável ao longo das observações e consideramos dois contextos de estimação, a saber: no primeiro, estimamos o modelo supondo dispersão variável, mas a modelamos de forma incorreta através dos preditores e das funções de ligação dos submodelos da média e da dispersão; no segundo contexto, estimamos o modelo supondo dispersão constante. Na segunda forma de geração, geramos dados com dispersão fixa, mas estimamos o modelo supondo dispersão variável.

No Capítulo 4, avaliamos o impacto de programas assistenciais e do crescimento da economia sobre o resultado da eleição presidencial de 2006. Dividimos a nossa análise em duas partes. Na primeira, nosso interesse consiste em explicar a diferença entre as proporções de votos válidos nos segundos turnos das eleições presidenciais de 2006 e 2002. Esta avaliação é realizada considerando o modelo de regressão beta com dispersão variável. Para explicar a diferença da proporção de votos do presidente Lula nas eleições de 2006 e 2002, consideramos variáveis geográficas, políticas e eleitorais e características socioeconômicas dos eleitores, além da diferença entre os gastos per capita em programas assistenciais em 2006 e em 2002 e a diferença dos níveis do PIB per capita de 2006 e 2002. Na segunda parte, estimamos o impacto dos gastos com programas assistenciais e do crescimento da economia sobre o resultado das eleições presidenciais de 2006. O nosso interesse aqui reside em estimar a quantidade de votos que o presidente Lula teria em duas diferentes situações. Primeira, se os gastos com programas assistenciais fossem mantidos nos níveis de 2002, e segunda, se o PIB per capita de 2006 fosse o mesmo de 2002.

No Capítulo 5, o nosso objetivo consiste em modelar a proporção de ateus e a proporção de pessoas que não acreditam que religião é importante para suas vidas cotidianas em diferentes países. Para isso, consideramos algumas covariáveis, tais como quociente de inteligência, renda nacional bruta ajustada pela paridade no poder de compra, predominância de muçulmanos, grau de abertura da economia e expectativa de vida em cada uma das nações consideradas. Dado que as duas variáveis dependentes consideradas são proporções, nós utilizamos o modelo de regressão beta. Adicionalmente, estimamos o impacto do quociente de inteligência sobre a concentração média de ateus e de pessoas que não valorizam religião.

Por fim, no último capítulo apresentamos as considerações finais do trabalho que incluem conclusões e possíveis direcionamentos para trabalhos futuros.

As avaliações numéricas realizadas no Capítulo 2 foram implementadas na linguagem de programação `Ox` (Doornik, 2006). Os resultados computacionais dos demais capítulos foram produzidos utilizando a liguagem de programação `R`, que se encontra disponível gratuitamente em `http://www.r-project.org`.

CAPÍTULO 2

Inferência em modelos lineares de regressão irrestritos e restritos sob heteroscedasticidade de forma desconhecida

2.1 Introdução

A análise de regressão é um dos métodos de análise de dados da teoria estatística mais utilizados na pesquisa científica. O modelo clássico de regressão teve origem nos trabalhos de astronomia elaborados por Gauss no período de 1809 a 1821; trata-se de uma técnica adequada para estudar o efeito que variáveis explicativas exercem sobre uma variável resposta. Técnicas clássicas de regressão linear assumem que os erros do modelo têm variância constante, ou seja, geralmente assume-se *homoscedasticidade*. Porém, esta suposição é bastante forte e, em uma relevante parte dos problemas práticos, muito pouco razoável.

Sob as suposições tipicamente feitas para este modelo, a estimação dos parâmetros é comumente realizada utilizando o método de mínimos quadrados ordinários (MQO). Este método, através de operações matriciais de fácil implementação computacional, fornece estimadores de simples interpretação e que possuem propriedades desejáveis, como não-viés, consistência, eficiência e normalidade assintótica. Quando a suposição de homoscedasticidade é violada, ou seja, quando as variâncias dos erros não são constantes, dizemos que há *heteroscedasticidade* no modelo. A presença de heteroscedasticidade ocorre com frequência quando se trabalha com dados de corte transversal.

Sob heteroscedasticidade, o estimador de MQO mantém algumas propriedades desejáveis, tornando-se, contudo, ineficiente na classe dos estimadores lineares não-viesados, i.e., não é mais o melhor estimador linear não-viesado (*Best Linear Unbiased Estimator, BLUE)*. O estimador usual de sua matriz de covariâncias torna-se viesado e não-consistente, tornando pouco confiáveis estimativas intervalares e testes de hipóteses nele baseados. Uma maneira de contornar este problema é obter estimadores alternativos ao estimador de MQO. Torna-se razoavelmente fácil estimar os parâmetros de um modelo de regressão quando os erros são heteroscedásticos com padrão de heteroscedasticidade que é determinado por uma função cedástica conhecida, ou mesmo quando os parâmetros da função cedástica são desconhecidos, mas com forma de heteroscedasticidade conhecida. Contudo, quase sempre o processo gerador das diferentes variâncias é desconhecido.

Um procedimento de estimação muito usado na prática para o modelo de regressão linear na presença de heteroscedasticidade de forma desconhecida consiste em utilizar para o vetor de parâmetros β o estimador de mínimos quadrados ordinários, que permanece não-viesado e

consistente, juntamente com um estimador para sua matriz de covariâncias que possua propriedades assintóticas desejáveis. Esse estimador consistente da matriz de covariâncias do estimador de MQO possibilita a realização de inferências no modelo heteroscedástico sem que seja necessário fazer suposições sobre a forma da heteroscedasticidade.

O objetivo deste capítulo é estudar o comportamento de testes baseados em estimadores consistentes da matriz de covariâncias do estimador de mínimos quadrados orninários quando resíduos oriundos de regressões robustas são usados em substituição a resíduos de mínimos quadrados ordinários, sob heteroscedasticidade de forma desconhecida, usando tanto estimação irrestrita quanto restrita. Objetiva-se, ainda, investigar o efeito da presença de pontos de alta alavancagem sobre os comportamentos dos testes considerados. Essa idéia foi inicialmente explorada por Furno (1996), que considerou apenas resíduos robustos oriundos de mínimos quadrados ponderados. Neste capítulo, consideraremos três estimadores robustos: menor mediana dos quadrados dos resíduos (LMS), mínimos quadrados podados (LTS), mínimos quadrados ponderados (WLS).

2.2 Modelos e estimadores

O modelo de interesse é o modelo de regressão linear, definido como

$$y = X_1\beta_1 + X_2\beta_2 + e = X\beta + e, \tag{2.2.1}$$

em que y é um vetor $n\times 1$ de observações de uma variável dependente, $X = (X_1, X_2)$ é uma matriz $n\times k$ ($k = k_1 + k_2 < n$) de regressores fixos com posto$(X) = k$, $\beta = (\beta_1^\top, \beta_2^\top)^\top$ é um vetor $k\times 1$ de parâmetros desconhecidos e e é um vetor $n\times 1$ de erros aleatórios não-correlacionados, cada um tendo média zero e variância σ_t^2, $t = 1,\ldots,n$. A matriz de covariâncias de e é dada por $\Psi = \text{diag}\{\sigma_1^2,\ldots,\sigma_n^2\}$. A Equação (2.2.1) representa o modelo irrestrito. Sob $\mathcal{H}_0 : \beta_2 = 0$, o modelo restrito é

$$y = X_1\beta_1 + e. \tag{2.2.2}$$

Os estimadores de mínimos quadrados ordinários dos coeficientes dos modelos de regressão irrestrito e restrito são denotados por $\widehat{\beta} = (\widehat{\beta}_1^\top, \widehat{\beta}_2^\top)^\top$ e $\tilde{\beta}_1$, respectivamente. Sob homoscedasticidade, temos que $\sigma_t^2 = \sigma^2 > 0$, i.e., $\Psi = \sigma^2 I_n$, em que I_n é a matriz identidade de ordem n. Então, a matriz de covariâncias do estimador de mínimos quadrados ordinários é dada por $\sigma^2(X^\top X)^{-1}$ e pode ser facilmente estimada por $\widehat{\sigma}^2(X^\top X)^{-1}$, em que $\widehat{\sigma}^2 = \widehat{e}^\top\widehat{e}/(n-k)$, $\widehat{e} = [I - X(X^\top X)^{-1}X^\top]y = My$ considerando o Modelo (2.2.1) e $\tilde{e} = (I - X_1(X_1^\top X_1)^{-1}X_1^\top)y = M_1 y$ para o Modelo (2.2.2)

Na presença de heteroscedasticidade, o estimador $\widehat{\sigma}^2(X^\top X)^{-1}$ não é mais consistente nem não-viesado para a matriz de covariâncias de $\widehat{\beta}$. Como em um grande número de aplicações os erros são heteroscedásticos, é importante considerar estimadores mais confiáveis para a variância de $\widehat{\beta}$. Um estimador consistente da matriz de covariâncias de $\widehat{\beta}$ muito utilizado foi proposto por Halbert White (1980), sendo aqui denotado por HC0. O estimador de White é obtido substituindo-se o t-ésimo elemento diagonal de Ψ pelo t-ésimo resíduo ao quadrado, resultando em

$$\text{HC0} = (X^\top X)^{-1} X^\top \widehat{\Psi}_0 X (X^\top X)^{-1},$$

em que $\widehat{\Psi}_0 = \text{diag}\{\widehat{e}_1^2, \ldots, \widehat{e}_n^2\}$.

Utilizando o Modelo (2.2.2), i.e., impondo a restrição de que $\beta_2 = 0$, temos o estimador HCR0, que segue a mesma estrutura do estimador HC0, mas utiliza resíduos obtidos de regressão restrita. Para este estimador a matriz $\widehat{\Psi}_0$ é substituída por $\widehat{\Psi}_{0R} = \text{diag}\{\tilde{e}_1^2, \ldots, \tilde{e}_n^2\}$.

Apesar de ser um estimador consistente da matriz de covariâncias de $\widehat{\beta}$ tanto sob heteroscedasticidade como sob homoscedasticidade, o estimador HC0 tende a subestimar, em amostras finitas, a variância do estimador de MQO, sendo tipicamente bastante viesado quando o tamanho da amostra é pequeno. Através de simulações de Monte Carlo, Long & Ervin (2000) constataram que este estimador não é confiável quando o tamanho da amostra é menor ou igual a 250. Adicionalmente, Cribari–Neto & Zarkos (1999), fazendo uso de estudos de simulação, mostraram que o estimador HC0 pode ser muito viesado quando o tamanho da amostra não é grande.

Assim como o tamanho amostral, a estrutura da matriz de regressores X, é outro fator que também interfere no desempenho do estimador de White em amostras finitas. Quando a matriz de regressores X contém observações de alta alavancagem, isto é, observações que possuem potencial para exercer grande influência sobre os valores a serem preditos, o estimador HC0 tende a ser muito viesado. Dessa maneira, o uso de estimativas obtidas a partir de HC0 pode levar a inferências errôneas.

Davidson & MacKinnon (1993) definiram um estimador, denotado por HC3, que é uma modificação do estimador de White, cujo comportamento é similar ao do estimador 'jackknife' (MacKinnon & White, 1985). O estimador HC3 é dado por

$$\text{HC3} = (X^\top X)^{-1} X^\top \widehat{\Psi}_3 X (X^\top X)^{-1},$$

em que

$$\widehat{\Psi}_3 = \text{diag}\left\{\frac{\widehat{e}_1^2}{(1-h_1)^2}, \ldots, \frac{\widehat{e}_n^2}{(1-h_n)^2}\right\},$$

h_t denotando o t-ésimo elemento diagonal da matriz $H = X(X^\top X)^{-1}X^\top$.

Utilizando resíduos obtidos de regressão restrita, obtemos o estimador HCR3, que tem a mesma estrutura do estimador HC3, mas a matriz $\widehat{\Psi}_3$ é substituída por

$$\widehat{\Psi}_{3R} = \text{diag}\left\{\frac{\tilde{e}_1^2}{(1-g_1)^2}, \ldots, \frac{\tilde{e}_n^2}{(1-g_n)^2}\right\}.$$

No estimador HCR3, cada elemento h_t é substituído por g_t, em que g_t é o t-ésimo elemento diagonal de $X_1(X_1^\top X_1)^{-1}X_1^\top$.

Cribari–Neto (2004) propôs uma modificação do estimador HC3, denominada HC4, que leva em consideração o impacto de observações com alta alavancagem em amostras finitas, incorporando fatores de desconto definidos pela razão entre o grau individual de alavancagem e o grau médio de alavancagem. A importância de considerar o efeito da alta alavancagem de algumas observações pode ser verificada através dos resultados em Cribari–Neto & Zarkos (2001),

que mostram que a presença de pontos de alavanca na estrutura da matriz dos regressores é mais decisiva para o comportamento de estimadores consistentes da matriz de covariâncias de $\widehat{\beta}$ em amostras finitas do que o próprio grau de heteroscedasticidade. O estimador HC4 é dado por

$$\text{HC4} = (X^\top X)^{-1} X^\top \widehat{\Psi}_4 X (X^\top X)^{-1},$$

em que

$$\widehat{\Psi}_4 = \text{diag}\left\{ \frac{\widehat{e}_1^2}{(1-h_1)^{\delta_1}}, \ldots, \frac{\widehat{e}_n^2}{(1-h_n)^{\delta_n}} \right\}$$

e

$$\delta_t = \min\left\{4, \frac{h_t}{\overline{h}}\right\} = \min\left\{4, \frac{nh_t}{\sum_{j=1}^n h_j}\right\},$$

com $\overline{h} = n^{-1}\sum_{j=1}^n h_j$, i.e., $\overline{h}$ é a média dos h_t's. Ou seja,

$$\delta_t = \min\left\{4, \frac{nh_t}{k}\right\}.$$

O expoente controla o nível de desconto do t-ésimo resíduo ao quadrado e é determinado pela razão entre h_t e a média dos h_t's, $\overline{h}$. Como $0 < 1 - h_t < 1$ e $\delta_t > 0$, segue que $0 < (1-h_t)^{\delta_t} < 1$. O t-ésimo resíduo ao quadrado deverá ser tanto mais fortemente inflacionado quanto maior for h_t relativamente a $\overline{h}$. Este desconto linear é truncado em 4, que equivale a duas vezes o grau de desconto usado pelo estimador HC3. Ou seja, usa-se $\delta_t = 4$ quando $h_t > 4\overline{h} = 4k/n$. Os resultados numéricos obtidos por Cribari–Neto (2004) mostram que este estimador tem comportamento superior ao do estimador HC3 em amostras finitas no que diz respeito a testes quasi-t associados.

Fazendo uso dos resíduos obtidos de regressão restrita, obtemos o estimador HCR4 que segue a mesma estrutura do estimador HC4, mas a matriz $\widehat{\Psi}_4$ é substituída por

$$\widehat{\Psi}_{4R} = \text{diag}\Big\{ \frac{\tilde{e}_1^2}{(1-g_1)^{\delta_{1R}}}, \ldots, \frac{\tilde{e}_n^2}{(1-g_n)^{\delta_{nR}}} \Big\},$$

em que $\delta_{tR} = \min\left\{4, \frac{g_t}{\overline{g}}\right\}$. Note que h_t e $\overline{h} = k/n$ são substituídos por g_t e $\overline{g} = k_1/n$, respectivamente, em que g_t é o t-ésimo elemento diagonal de $X_1(X_1^\top X_1)^{-1}X_1^\top$.

Com o objetivo de reduzir os vieses dos estimadores HC0 e HC3, Furno (1996) sugere o uso de regressão robusta para limitar o efeito de pontos de alavanca e estimar HC0 e HC3 usando resíduos robustos. O estimador robusto da matriz de covariâncias proposto pela autora é do tipo HC0:

$$\text{HC0}_{\text{WLS}} = (X^\top W X)^{-1} X^\top W \widehat{\Psi}_{WLS} W X (X^\top W X)^{-1}, \tag{2.2.3}$$

em que W é uma matriz diagonal de ordem n, sendo o t-ésimo elemento diagonal dado por $w_t = \min(1, c/h_t)$, em que $c = 1.5k/n$. Aqui, $\widehat{\Psi}_{WLS} = \text{diag}\{\check{e}_1^2, \ldots, \check{e}_n^2\}$, $\check{e}_t$ sendo o t-ésimo resíduo obtido da regressão de mínimos quadrados ponderados. No caso do estimador robusto da matriz de covariâncias do tipo HC3, HC3_{WLS}, a matriz $\widehat{\Psi}_{WLS}$ na Equação (2.2.3) é substituída por $\widehat{\Psi}_{3WLS} = \text{diag}\{(\check{e}_1/(1-h_1^*))^2, \ldots, (\check{e}_n/(1-h_n^*))^2\}$, em que h_t^* é o t-ésimo elemento diagonal de

$\sqrt{W}X(X^\top WX)^{-1}X^\top\sqrt{W}$. Por fim, é possível usar a idéia proposta pela autora para se obter um estimador robusto do tipo HC4, HC4_{WLS}. Neste sentido, deve-se substituir $\widehat{\Psi}_{WLS}$ em (2.2.3) por

$$\widehat{\Psi}_{4WLS} = \text{diag}\left\{\frac{\check{e}_1^2}{(1-h_1^*)^{\delta_1^*}},\ldots,\frac{\check{e}_n^2}{(1-h_n^*)^{\delta_n^*}}\right\},$$

em que $\delta_t^* = \min\left\{4, \frac{h_t^*}{\bar{h}^*}\right\}$.

2.3 Novos estimadores consistentes

Os estimadores consistentes da estrutura de covariâncias de $\widehat{\beta}$ (EMQO de β) são baseados em resíduos de mínimos quadrados. Os ajustes propostos ao estimador HC0 tentaram compensar o fato de que pontos de alavanca tendem a conduzir a resíduos correspondentes relativamente pequenos em valores absolutos e o estimador interpreta resíduos pequenos como indicação de que as correspondentes variâncias também são pequenas (Chester & Jewitt, 1987, p. 1219).

A nossa proposta é usar resíduos obtidos de métodos de estimação robustos em substituição aos resíduos de mínimos quadrados na definição dos estimadores da matriz de covariâncias de $\widehat{\beta}$, dado que tais resíduos são menos sensíveis à presença de observações distoantes e, dessa forma, os erros-padrão resultantes devem ser menos distorcidos em amostras finitas. Essa idéia foi inicialmente explorada por Lima et al. (2010), que considerou resíduos oriundos de regressão robusta em substiuição a resíduos oriundos de mínimos quadrados ordinários, contudo considerou-se apenas resíduos obtidos de regressão irrestrita.

Furno (1996) propôs um estimador que atribui ponderações diferentes a diferentes observações. Em particular, observações de maior alavancagem recebem ponderações menores no processo de estimação. Todavia, seu estimador permanece baseado no critério de mínimos quadrados, não sendo assim robusto à presença de observações extremas, incluindo *outliers*. A nossa proposta é a utilização de resíduos obtidos a partir de métodos de estimação robustos, com alto ponto de ruptura. [1] Em particular, propomos a utilização de resíduos obtidos via estimação LMS e LTS.

Nas Seções 5 e 6 nós avaliaremos os estimadores consistentes da matriz de covariâncias de $\widehat{\beta}$ que usam resíduos obtidos de regressão irrestrita ($\widehat{e_t}$) e restritos ($\tilde{e}_t$) ambos oriundos do estimador MQO e dos estimadores robustos de menor mediana dos quadrados dos resíduos (LMS) e mínimos quadrados podados (LTS), que foram propostos por Rousseeuw (1984), além do de mínimos quadrados ponderados (WLS), proposto por Furno (1996). Os estimadores LMS e LTS podem ser descritos como segue. Para qualquer vetor de estimativas $\widehat{\beta}$, o t-ésimo resíduo é dado por $\widehat{e}_t = y_t - x_t^\top\widehat{\beta}$, em que x_t é a t-ésima linha de X. As funções LMS e LTS são definidas, respectivamente, como $\text{med}(\widehat{e}_1^2,\widehat{e}_2^2,\ldots,\widehat{e}_n^2)$ e $\sum_{t=1}^r \widehat{e}_{(t)}^2$, em que $\widehat{e}_{(t)}^2$ corresponde à t-ésima estatística de ordem dos quadrados dos resíduos. O valor ótimo de r está relacionado com o ponto de ruptura do estimador de LTS. De acordo com Rousseeuw & Hubert (1997), o valor ótimo de r, denotado por r^*, é dado por $r^* = \lfloor(n+k+1)/2\rfloor$, em que $\lfloor.\rfloor$ denota a parte inteira. Como

[1] Ponto de ruptura de um estimador mede qual seria a maior porcentagem de contaminação que um estimador poderia suportar e ainda assim fornecer informação confiável sobre o parâmetro considerado.

não existem fórmulas fechadas para esses estimadores de β, um algoritmo de estimação foi proposto por Rousseeuw & Leroy (1987). Esse algoritmo baseia-se na obtenção de subamostras de m diferentes observações, em que $m \geq k+1$. Sejam tais subamostras, y_l^{sub} e $X_l^{\text{sub}}, m \times k$ $l = 1, \ldots, n_S$, em que $n_S = \binom{n}{m}$ é o número total de subamostras de tamanho m. Alternativamente, dependendo do valor de n (tamanho amostral) e de k (número de parâmetros) é possível considerar apenas um subconjunto aleatório de n_S (ver Rousseeuw & Hubert (1997), Tabela 4). Os parâmetros β_l são estimados por mínimos quadrados ordinários para cada par $(y_l^{\text{sub}}, X_l^{\text{sub}})$. Para cada β_l, os valores das funções objetivo, em relação a LMS e LTS, são determinados com relação às n observações. Os estimadores $\widehat{\beta}_{\text{LMS}}$ e $\widehat{\beta}_{\text{LTS}}$ são aqueles que minimizam as funções objetivo LMS e LTS, respectivamente.

A desvantagem do estimador HC0 reside no fato de que os resíduos de mínimos quadrados tendem a subestimar os verdadeiros erros (Judge et al., 1985, p. 367). Este fato pode tornar HC0 um estimador muito viesado quando o tamanho amostral é pequeno. Para ilustrar este fato, a Figura 2.1 apresenta os ajustes do modelo de regressão $y_t = \beta_0 + \beta_1 x_t + e_t, t = 1, \ldots, 20$, obtidos via MQO, LMS, LTS e WLS. Dois cenários foram considerados, um balanceado (sem pontos de alavanca) e outro não-balanceado (com pontos de alavanca). Os valores da covariável x correspondem a valores igualmente espaçados entre 0 e 1, esse é o que chamamos de cenário 1 (dados sem pontos de alavanca). No segundo cenário, cenário 2, substituímos a última observação da covariável por 4 a fim de introduzir um ponto de alavanca nos dados, ou seja, para que o último elemento diagonal da matrix $X(X^\top X)^{-1}X^\top$ ultrapasse $3k/n = 0.30$, o valor limite comumente utilizado na identificação de pontos de alavanca. Nos dois cenários tomamos $\beta_0 = 10$ e $\beta_1 = 5$. Fizemos o ajuste do modelo considerando os estimadores MQO, LMS, LTS e WLS, como apresenta a Figura 2.1. Observamos que no cenário 1 ($a = 1$), os ajustes dos modelos são semelhantes, apesar dos resíduos correspondentes à última observação obtidos através dos estimadores LMS e LTS serem um pouco maiores, em valores absolutos, do que o resíduo obtido via estimador MQO. No cenário 2 ($a = 4$), há um ponto de alta alavancagem ($h_{\max} = 0.888$). O valor predito da regressão obtida a partir do estimador MQO se encontra muito próximo da observação de alta alavancagem e o resíduo correspondente é menor, em valor absoluto, do que os resíduos obtidos via LMS, LTS e WLS. Vale ressaltar que os ajustes MQO e WLS não diferem muito, como pode ser visto na Figura 2.1. Os resíduos referentes à última observação, em valor absoluto, obtidos via MQO e WLS são 10.26 e 11.95, respectivamente; quando utilizamos os estimadores LMS e LTS esse resíduo aumenta, passando a ser 17.88 e 18.27, respectivamente.

Em suma, o resíduo de MQO correspondente ao ponto de alavanca tende a ser demasiadamente pequeno (em valor absoluto) o que termina por distorcer a estimativa HC0, já que o estimador de White toma resíduos pequenos (em valor absoluto) como indicação de que as variâncias correspondentes são pequenas. Isso aumenta o viés do estimador e o torna demasiadamente 'otimista' em amostras finitas. Os resíduos robustos LMS e LTS, por outro lado, são maiores (em termos absolutos) e sinalizam que a variância do erro no ponto de alavanca é alta. Ou seja, eles emitem a sinalização correta ao estimador da estrutura de covariâncias de $\widehat{\beta}$.

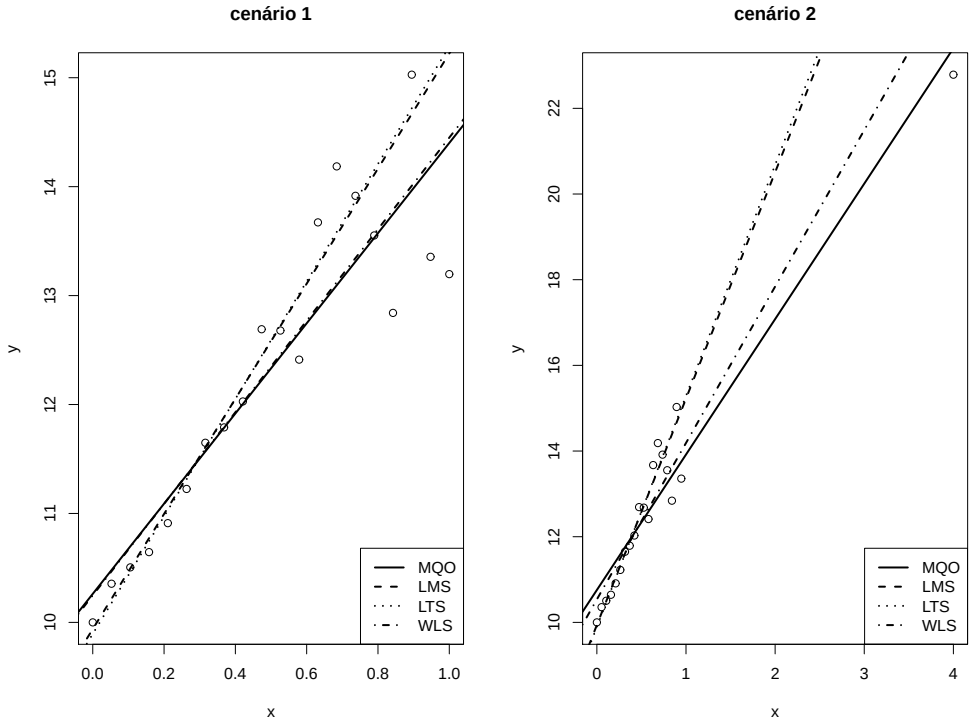

Figura 2.1 Ajustes do modelo de regressão $y_t = \beta_0 + \beta_1 x_t + e_t, t = 1, \ldots, 20$, considerando os estimadores MQO, LMS, LTS e WLS nos cenários 1 e 2.

2.4 Estatísticas quasi-t e quasi-F

Considere o teste da hipótese nula $\mathcal{H}_0 : \beta_r = \beta_r^{(0)}$ versus $\mathcal{H}_1 : \beta_r \neq \beta_r^{(0)}$ ou mais geralmente, $\mathcal{H}_0 : c^\top \beta = a$ versus $\mathcal{H}_1 : c^\top \beta \neq a$, em que c é um vetor $k \times 1$, a é uma constante e β_r é o r-ésimo elemento do vetor de parâmetros β. Sob as suposições do modelo de regressão linear, a estatística

$$t = \frac{c^\top b - a}{\sqrt{\widehat{\text{var}}(c^\top b)}} \tag{2.4.1}$$

tem, sob $\mathcal{H}_0$, distribuição t de Student com $n-k$ graus de liberdade. A estimativa da variância usada no denominador é a estimativa usual de mínimos quadrados ordinários. Quando há heteroscedasticidade, contudo, a estatística t não possui distribuição nula t de Student exata e nem mesmo distribuição assintótica nula normal padrão. Todavia, podemos definir a estatística de teste com base em uma das estimativas consistentes da variância de $c^\top b$ apresentadas anteriormente. Quando definimos t desta forma, temos que, sob $\mathcal{H}_0$, t converge em distribuição para a distribuição normal padrão.

Suponha agora que o interesse reside em testar mais de uma restrição, i.e., $\mathcal{H}_0 : R\beta - q = 0$ versus $\mathcal{H}_1 : R\beta - q \neq 0$, em que R é uma matriz $k_2 \times k$ de posto k_2, k_2 sendo o número de restrições em teste e q sendo um vetor $k_2 \times 1$. A estatística de teste nesse caso é dada por

$$W = \widehat{m}^\top \widehat{\text{var}}(\widehat{m})^{-1} \widehat{m}, \tag{2.4.2}$$

em que $\widehat{m} = R\widehat{\beta} - q$. Para que haja robustez contra heteroscedasticidade, a estatística (2.4.2) deve ser escrita como

$$W = \widehat{m}^\top R(X^\top X)^{-1} X^\top \widehat{\Psi}_l X (X^\top X)^{-1} R^\top \widehat{m}, \quad l = 0, 3, 4.$$

Sob $\mathcal{H}_0$, $W \overset{d}{\to} \chi^2_{k_2}$, em que $\overset{d}{\to}$ denota convergência em distribuição.

2.5 Avaliação numérica I

O primeiro experimento de Monte Carlo é baseado no modelo de regressão linear

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t, \quad t = 1, \ldots, n.$$

As covariáveis x_1 e x_2 correspondem à ocupação média diária de uma instalação e ao número de alas construídas na instalação, respectivamente, com um total de 25 observações (fonte: Myers, 1990, p. 218). Cada uma das observações de x_1 e x_2 é repetida duas, três e quatro vezes para formar amostras de tamanhos $n = 50, 75, 100$, que são mantidas constantes em todo experimento. Isto garante que $\lambda = \max(\sigma^2_t)/\min(\sigma^2_t)$, uma medida do grau de heteroscedasticidade, mantém-se constante para os diferentes tamanhos amostrais, e que os resultados da simulação são apenas afetados pelo crescente número de observações. Em nossos estudos de simulação utilizamos $\beta_0 = \beta_1 = 1$ e $\beta_2 = 0$. Sob homoscedasticidade $\lambda = 1$ e sob heteroscedasticidade o valor de λ depende da especificação da função cedástica, tendo-se $\lambda > 1$.

Consideramos o conjunto de dados denominado 'dados1', que é o conjunto de dados completos mencionado anteriormente. A função cedástica foi especificada como $\sigma^2_t = \exp\{c_1 x_{t1}\}$. Simulações sob homoscedasticidade foram realizadas com $c_1 = 0$. Para dados heteroscedásticos, consideramos um grau de heteroscedasticidade próximo de 50. Os desempenhos dos vários estimadores da matriz de covariâncias considerados no experimento de Monte Carlo foram avaliados com base em: (a) tamanho I: tamanho do teste quasi-*t* associado da hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ contra $\mathcal{H}_1 : \beta_2 \neq 0$; (b) poder: o poder do teste. Os experimentos de simulação são baseados em 100,000 réplicas de Monte Carlo. Todas as simulações foram realizadas usando a linguagem de programação `Ox` (Doornik, 2006).

Para as duas formas de avaliação citadas, foram utilizados resíduos oriundos dos estimadores de mínimos quadrados ordinários e dos estimadores robustos, LMS, LTS e WLS. Doze experimentos de Monte Carlo foram realizados, seis sob homoscedasticidade e seis sob heteroscedasticidade para cada conjunto de resíduos. O conjunto de dados usado (para os regressores) é uma amostra de tamanho 25 retirada de Myers (1990, Tabela 5.2, p. 218). Isto é o que chamamos de 'dados1'.

A seguir, apresentamos os resultados de simulações relativos às taxas de rejeição dos diferentes testes quasi-*t*, aos níveis nominais de 10%, 5% e 1% (ou seja, as probabilidades nominais de rejeição da hipótese nula quando esta é de fato verdadeira), sob homoscedasticidade e sob heteroscedasticidade. Esses resultados encontram-se nas Tabelas 2.1 a 2.6. As nomenclaturas utilizadas nessas tabelas são:

- HCj_1, $j = 0,3,4$: resíduos oriundos do estimador MQO;
- HCj_2, $j = 0,3,4$: resíduos oriundos do estimador LMS;
- HCj_3, $j = 0,3,4$: resíduos oriundos do estimador LTS;
- HCj_4, $j = 0,3,4$: resíduos oriundos do estimador WLS;
- HCj_5, $j = 0,3,4$: resíduos LMS usados no estimador consistente WLS (estimador de Furno);
- HCj_6, $j = 0,3,4$: resíduos LTS usados no estimador consistente WLS (estimador de Furno).

A adição da letra R aos nomes dos estimadores acima citados indica o uso de resíduos obtidos de regressão restrita. Os tamanhos dos testes quasi-t baseados nos estimadores HC0_k e HCR0_k, $k = 1,\ldots,6$, estão apresentados nas Tabelas 2.1 e 2.2, respectivamente. As taxas de rejeição dos testes baseados nos estimadores HC3_k e HCR$3_k, k = 1,\ldots,6$, estão nas Tabelas 2.3 e 2.4. Por fim, os tamanhos dos testes baseados nos estimadores HC4_k e HCR$4_k, k = 1,\ldots,6$, encontram-se nas Tabelas 2.5 e 2.6. As principais conclusões extraídas estão resumidas a seguir.

Nós restringimos a análise aos estimadores HC0, HC3, HC4 e ao estimador proposto por Furno (1996). O estimador HC0 é o mais conhecido e mais utilizado, os resultados numéricos apresentados por Long & Ervin (2000) fornecem testes de hipótese baseados no estimador HC3 e os resultados de simulação apresentados por Cribari–Neto(2004) indicam que testes baseados no estimador HC4 são competitivos com testes realizados via bootstrap duplo. Não consideramos o estimador HC5 porque seu desempenho é similar ao do teste HC4. Também não consideramos os estimadores HC1 e HC2 dado apresentarem desempenho em amostras finitas inferior àquele alcançado pelo estimador HC3.

Primeiro, o teste baseado nos estimadores HC$0_k, k = 1,\ldots 6$, não apresenta comportamento confiável, tanto sob homoscedasticidade quanto sob heteroscedasticidade, principalmente quando os resíduos usados são de MQO. Por exemplo, sob heteroscedasticidade ($\lambda \approx 50$), $n = 50$ e $\alpha = 5\%$, as taxas de rejeição dos testes HC$0_k, k = 1,\ldots,6$, são 23.89%, 11.14%, 10.23%, 12.76%, 8.18% e 7.63%, respectivamente, (ver Tabela 2.1). Quando utilizamos resíduos oriundos de estimações robustas o desempenho do teste se torna mais confiável. Ao impor a restrição de que $\beta_2 = 0$, considerando o mesmo cenário, as taxas ficam mais próximas do nível nominal considerado. Por exemplo, o teste baseado no estimador HC0_4 rejeita a hipótese nula 12.76% das vezes (Tabela 2.1); o teste baseado no estimador HCR0_4, por outro lado, rejeita $\mathcal{H}_0$ 6.01% (Tabela 2.2), taxa que é quase duas vezes menor do que a anterior e muito mais próxima do nível nominal do teste (5%). Assim, o desempenho do teste baseado no estimador de White melhora consideravelmente quando impomos restrição nos parâmetros, sob homoscedasticidade e sob heteroscedasticidade.

Segundo, o desempenho do teste HC0 melhora significativamente quando utilizamos resíduos robustos em substituição a resíduos MQO, sob homoscedasticidade e sob heteroscedasticidade. Por exemplo, quando $n = 50$, $\lambda \approx 50$ e $\alpha = 10\%$, a taxa de rejeição do teste com resíduos MQO é mais do que o dobro das taxas obtidas (em quase todos os casos) quando resíduos robustos são utilizados. Esta conclusão pode ser estendida para estimadores restritos e irrestritos. Para ilustrar a conclusão anterior, considere o seguinte cenário: $n = 25$, $\lambda \approx 50$ e $\alpha = 10\%$. O teste baseado no estimador HC0_1, ou seja, usando resíduos de MQO, rejeita a hipótese nula 59.84% das vezes, enquanto o teste baseado no estimador HC0_6 rejeita a mesma

hipótese apenas 16.96% das vezes, como está apresentado na Tabela 2.1. Ao considerarmos o modelo de regressão restrito, as taxas de rejeição dos testes baseados nesses estimadores são 27.22% e 7.13%, respectivamente, como pode ser visto na Tabela 2.2.

Terceiro, os testes baseados nos estimadores $HC3_k, k = 1,\ldots,6$, apresentam taxas de rejeição mais próximas do nível nominal do que as taxas apresentadas pelos testes correspondentes baseados nos estimadores $HC0_k, k = 1,\ldots,6$. Por exemplo, para $n = 25$, $\alpha = 1\%$, $\lambda \approx 50$ e utilizando resíduos oriundos do estimador MQO, o teste baseado no estimador $HC0_1$ rejeita a hipótese em avaliação 41.18% das vezes (Tabela 2.1), enquanto que o teste baseado no estimador $HC3_1$ rejeita a hipótese nula apenas 4.49% das vezes (ver Tabela 2.3). É importante comentar que, apesar dos testes quasi-t associados ao estimador HC3 terem apresentado menores distorções de tamanho em relação ao teste baseado no estimador HC0, suas taxas de rejeição ficaram sempre acima dos níveis de significância nominais correspondentes na presença de pontos de alavanca, sendo estes testes, portanto, liberais, exceto para $n = 25$ em alguns casos.

Quarto, os testes construídos a partir de $HCR3_k, k = 1,\ldots 6$, apresentam melhores desempenhos do que os testes $HC3_k, k = 1,\ldots,6$, apenas quando resíduos MQO são utilizados, exceto para n pequeno, i.e., $n = 25$. Por exemplo, quando $\lambda \approx 50$, $n = 50$ e $\alpha = 5\%$, o teste baseado no estimador $HC3_1$ rejeita a hipótese nula 10.07% das vezes (Tabela 2.3) enquanto que ao se impor a restrição o teste baseado em $HCR3_1$ rejeita a hipótese nula apenas 2.69% das vezes; ver Tabela 2.4.

Quinto, o desempenho do teste baseado no estimador $HCR3_1$ é superior ao dos testes baseados nos estimadores $HCR3_k, k = 2,\ldots 6$, quando resíduos robustos são utilizados. Por exemplo, quando $\lambda \approx 50$, $n = 50$ e $\alpha = 10\%$, o teste baseado no estimador $HCR3_1$ rejeita a hipótese nula 9.79% das vezes; com resíduos robustos as taxas de rejeição são 2.19%, 2.17%, 2.87% e 2.74%, respectivamente, como pode ser visto na Tabela 2.4. O desempenho do teste baseado no estimador $HCR3_4$ é similar ao do teste baseado no estimador $HCR3_1$. Por exemplo, sob heteroscedasticidade, $n = 75$ e $\alpha = 10\%$, o teste baseado no estimador $HCR3_1$ rejeita a hipótese nula 10.20% das vezes, enquanto que o teste baseado no estimador $HCR3_4$ rejeita a hipótese nula 9.63% das vezes, ver Tabela 2.4.

Sexto, quando usamos resíduos robustos, os testes baseados nos estimadores $HCR3_k$, $k = 2,\ldots,6$, apresentam taxas de rejeição abaixo dos níveis nominais considerados, exceto para $n = 25$. O mesmo não ocorre quando resíduos oriundos do estimador de MQO são utilizados. Por exemplo, quando $\lambda \approx 50$, $n = 75$ e $\alpha = 10\%$, o teste baseado no estimador $HC3_1$ rejeita a hipótese nula 12.64% das vezes; considerando este mesmo cenário, mas utilizando resíduos robustos estas taxas são menores e estão mais distantes dos níveis nominais considerados, apresentado na Tabela 2.3.

Sétimo, os testes baseados nos estimadores $HC4_k$, $k = 1,\ldots,6$, apresentam taxas de rejeição mais próximas dos níveis nominais quando resíduos de MQO são utilizados, sob homoscedasticidade e sob heteroscedasticidade. Apesar de apresentar porcentagens de rejeição mais baixas do que os níveis nominais considerados, os testes baseados nesses estimadores têm desempenho superior aos dos testes baseados nos estimadores $HC0_k, k = 1,\ldots,6$, em todos os casos, sendo superiores também aos dos testes baseados em $HC3_k, k = 1,\ldots,6$, em alguns casos. Por exemplo, sob heteroscedasticidade, $n = 75$ e $\alpha = 5\%$, as taxas de rejeição dos testes baseados nos estimadores $HC0_1$, $HC3_1$ e $HC4_1$ são, respectivamente, 16.26% (Tabela 2.1), 7.98%

(Tabela 2.3) e 3.29% (Tabela 2.5).

Objetivando avaliar os comportamentos das principais estatísticas apresentadas anteriormente quanto à habilidade de conduzir baixas probabilidades do erro do tipo II, isto é, baixas probabilidades de não rejeitar a hipótese nula quando ela não é verdadeira, foram realizadas simulações de poder. O número de réplicas de Monte Carlo foi 100,000. Os resultados da investigação de poder sob homoscedasticidade e sob heteroscedasticidade, considerando $n = 50$ encontram-se nas Tabelas 2.7 (estimador irrestrito) e 2.8 (estimador restrito). Aqui, foram usados valores críticos exatos (estimados via simulação) já que alguns testes são demasiadamente liberais. A hipótese nula em teste é $\mathcal{H}_0 : \beta_2 = 0$ e os dados foram gerados utilizando $\beta_2 = 0.07$.

Os resultados obtidos nas simulações realizadas sob a hipótese alternativa revelam que, sob homoscedasticidade, os poderes dos testes baseados nos estimadores estudados são semelhantes, considerando os seis tipos de resíduos utilizados que foram oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS e WLS-LTS, mas vale ressaltar que os testes mais poderosos são os que utilizam resíduos oriundos dos estimadores de mínimos quadrados ordinários e mínimos quadrados ponderados, sob homoscedasticidade. Por exemplo, na situação em que $\lambda = 1$ e $\alpha = 5\%$, os poderes dos testes baseados nos estimadores $HC4_1$ e $HC4_4$ são 93.72% e 94.11%, respectivamente, enquanto que quando são usados resíduos oriundos dos estimadores LMS e LTS, os poderes dos testes são 71.99% e 64.79%, respectivamente, como pode ser visto na Tabela 2.7. Considerando este mesmo cenário, mas utilizando o estimador restrito, os poderes dos testes são menores que os poderes apresentados anteriormente, como pode ser visto na Tabela 2.8.

Adicionalmente, o grau de heteroscedasticidade afeta decisivamente os poderes de todos os testes considerados. Os poderes dos testes diminuem com o aumento da intensidade da heteroscedasticidade, i.e., aumentam as probabilidades de erro do tipo II. Tome como exemplo a situação em que $\alpha = 10\%$ e a estatística quasi-t que utiliza o estimador $HC3_2$ (resíduos oriundos do estimador LMS). Neste caso, os poderes dos testes quando $\lambda = 1$ (homoscedasticidade) e $\lambda \approx 50$ (heteroscedasticidade) são, respectivamente, 87.73% e 37.10%, ou seja, o poder quando há forte heteroscedasticidade é quase duas vezes menor do que sob homoscedastidade (Tabela 2.7). Esta mesma conclusão pode ser estendida para o estimador com restrição; ver Tabela 2.8.

No que diz respeito aos melhores desempenhos quanto ao poder sob heteroscedasticidade, quando os resíduos oriundos do estimador WLS são considerados, o poder do teste é maior do que quando os outros resíduos são utilizados, considerando estimadores irrestrito e restrito. Por exemplo, sob heteroscedasticidade e ao nível de 1%, o poder do teste baseado no estimador $HC0_4$ é quase seis vezes maior do que quando utilizamos o estimador $HC0_3$, como pode ser visto na Tabela 2.7. O poder do teste baseado no estimador $HC0_4$ é maior do que os poderes dos testes baseados nos estimadores $HC3_4$ e $HC4_4$. Por exemplo, ao nível de significância de 5%, os poderes dos testes baseados nos estimadores $HC0_4$, $HC3_4$ e $HC4_4$ são, respectivamente, 58%, 43.79% e 30.12%, como apresenta a Tabela 2.7. Quando consideramos o estimador restrito, as mesmas conclusões são obtidas, todavia os poderes dos testes baseados nestes estimadores são semelhantes. Por exemplo, considere o mesmo cenário descrito anteriormente; os poderes dos testes baseados nos estimadores $HCR0_4$, $HCR3_4$ e $HCR4_4$ são 38.90%, 38.63% e 37.98%, respectivamente, ver Tabela 2.8.

Por fim, fixamos o tamanho amostral em 50 e estimamos, sob heteroscedasticidade, os

Tabela 2.1 Porcentagens de rejeição de testes quasi-t baseados nos estimadores $HC0_k$, $k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim (0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n = 25, 50, 75, 100$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

λ	n	α	$HC0_1$	$HC0_2$	$HC0_3$	$HC0_4$	$HC0_5$	$HC0_6$
1	25	10	34.56	18.14	18.92	13.67	8.54	8.10
		5	26.83	14.10	15.05	7.75	5.00	4.77
		1	16.04	8.55	9.56	2.27	1.73	1.66
	50	10	22.30	14.94	13.93	11.62	9.12	8.29
		5	15.50	10.16	9.60	6.23	4.81	4.37
		1	7.36	4.81	4.81	1.51	1.23	1.06
	75	10	18.02	13.45	12.17	11.03	9.65	8.68
		5	11.94	8.67	7.90	5.72	5.03	4.42
		1	5.07	3.60	3.30	1.30	1.16	1.00
	100	10	15.91	12.42	11.55	10.67	9.72	9.04
		5	9.93	7.54	6.98	5.37	4.92	4.51
		1	3.82	2.78	2.55	1.21	1.11	1.00
≈ 50	25	10	59.84	26.31	25.91	28.40	17.71	16.96
		5	53.22	22.76	22.23	20.15	13.08	12.46
		1	41.18	17.84	17.98	9.66	7.37	7.10
	50	10	31.03	15.75	14.50	19.84	13.19	12.15
		5	23.89	11.14	10.23	12.76	8.18	7.63
		1	14.48	6.10	5.52	4.91	3.25	2.96
	75	10	22.86	13.77	12.29	16.47	12.13	11.04
		5	16.26	8.83	7.76	10.13	7.19	6.43
		1	8.14	3.93	3.41	3.56	2.50	2.17
	100	10	19.08	12.37	11.27	14.77	11.41	10.54
		5	12.70	7.53	6.73	8.75	6.48	5.89
		1	5.65	2.82	2.47	2.77	1.96	1.72

Tabela 2.2 Porcentagens de rejeição de testes quasi-t baseados nos estimadores $HCR0_k$, $k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim (0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n = 25, 50, 75, 100$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

λ	n	α	$HCR0_1$	$HCR0_2$	$HCR0_3$	$HCR0_4$	$HCR0_5$	$HCR0_6$
1	25	10	10.17	6.09	6.07	10.01	7.10	6.77
		5	3.65	2.37	2.42	3.89	2.83	2.73
		1	0.18	0.25	0.23	0.23	0.25	0.24
	50	10	10.33	7.31	6.97	10.22	8.38	7.88
		5	3.04	2.71	2.60	4.69	3.71	3.46
		1	0.33	0.29	0.30	0.55	0.47	0.45
	75	10	10.47	8.71	7.97	10.12	9.20	8.60
		5	4.42	3.49	3.14	4.81	4.25	3.87
		1	0.43	0.35	0.33	0.67	0.63	0.55
	100	10	10.37	9.06	8.54	10.05	9.26	8.91
		5	4.48	3.87	3.52	4.72	4.34	4.07
		1	0.48	0.43	0.40	0.74	0.67	0.62
≈ 50	25	10	27.22	5.32	4.91	13.76	7.49	7.13
		5	9.83	1.95	1.74	4.45	2.52	2.36
		1	0.26	0.21	0.22	0.20	0.18	0.18
	50	10	19.55	8.09	7.78	13.62	8.93	8.53
		5	9.63	2.29	2.17	6.01	3.67	3.44
		1	0.93	0.12	0.12	0.55	0.34	0.31
	75	10	16.14	9.58	8.74	12.80	9.59	8.93
		5	8.35	3.60	3.25	6.19	4.33	3.85
		1	1.17	0.18	0.16	0.77	0.51	0.45
	100	10	14.33	9.64	8.96	12.15	9.67	9.02
		5	7.43	4.07	3.60	5.96	4.40	3.98
		1	1.21	0.33	0.27	0.89	0.63	0.52

Tabela 2.3 Porcentagens de rejeição de testes quasi-t baseados nos estimadores $HC3_k$, $k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim (0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n = 25, 50, 75, 100$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

λ	n	α	$HC3_1$	$HC3_2$	$HC3_3$	$HC3_4$	$HC3_5$	$HC3_6$
1	25	10	7.34	5.95	11.78	4.41	4.78	4.90
		5	4.83	3.82	8.74	1.99	2.70	2.81
		1	2.16	1.67	4.57	0.43	0.83	0.89
	50	10	11.40	7.76	7.65	7.47	6.08	5.46
		5	7.19	4.87	5.03	3.43	2.92	2.62
		1	2.80	1.99	2.20	0.69	0.66	0.57
	75	10	11.21	8.16	7.45	8.31	7.39	6.56
		5	6.74	4.89	4.53	3.99	3.58	3.10
		1	2.38	1.72	1.60	0.75	0.71	0.63
	100	10	10.77	8.14	7.53	8.73	7.98	7.38
		5	6.20	4.63	4.26	4.14	3.79	3.45
		1	2.01	1.45	1.35	0.78	0.74	0.65
≈ 50	25	10	12.08	13.76	18.27	9.39	10.92	10.73
		5	8.42	11.47	16.36	5.63	8.07	7.99
		1	4.49	7.89	12.79	2.15	4.45	4.52
	50	10	14.70	6.25	5.68	11.16	7.48	6.91
		5	10.07	4.31	3.89	6.29	4.36	3.99
		1	4.91	2.28	2.10	1.88	1.60	1.43
	75	10	12.64	6.39	5.60	10.08	7.86	7.06
		5	7.98	3.81	3.30	6.15	4.32	3.85
		1	3.38	1.54	1.34	1.75	1.35	1.15
	100	10	11.63	6.66	5.90	10.75	8.03	7.39
		5	6.99	3.61	3.16	5.86	4.22	3.81
		1	2.61	1.18	0.97	1.53	1.14	0.96

Tabela 2.4 Porcentagens de rejeição de testes quasi-t baseados nos estimadores HCR3$_k$, $k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim (0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para n = 25, 50, 75, 100. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

λ	n	α	HCR3$_1$	HCR3$_2$	HCR3$_3$	HCR3$_4$	HCR3$_5$	HCR3$_6$
1	25	10	0.97	1.95	3.37	5.11	2.05	2.23
		5	0.11	0.53	1.11	1.36	0.60	0.67
		1	0.00	0.01	0.04	0.02	0.02	0.03
	50	10	6.48	4.06	4.13	8.10	4.91	4.61
		5	1.88	1.34	1.40	3.38	1.75	1.69
		1	3.57	0.10	0.13	0.28	0.14	0.13
	75	10	8.51	6.20	5.66	8.88	6.65	6.12
		5	3.02	2.18	1.96	3.94	2.68	2.44
		1	0.19	0.18	0.16	0.46	0.28	0.25
	100	10	8.71	7.24	6.74	9.11	7.43	7.10
		5	3.49	2.81	2.56	4.16	3.20	2.97
		1	0.29	0.25	0.24	0.58	0.39	0.36
≈ 50	25	10	0.29	1.44	2.34	3.54	1.01	1.08
		5	0.04	0.39	0.81	0.77	0.28	0.34
		1	0.00	0.01	0.04	0.02	0.01	0.02
	50	10	9.79	2.19	2.17	8.29	2.87	2.74
		5	2.69	0.46	0.45	2.75	0.80	0.75
		1	0.03	0.03	0.03	0.14	0.03	0.03
	75	10	10.20	4.05	3.69	9.63	4.90	4.43
		5	4.36	1.14	1.00	4.01	1.69	1.51
		1	0.29	0.03	0.03	0.33	0.12	0.09
	100	10	9.98	5.28	4.76	9.89	5.99	5.43
		5	4.57	1.79	1.52	4.40	2.27	1.99
		1	0.52	0.07	0.06	0.50	0.21	0.18

Tabela 2.5 Porcentagens de rejeição de testes quasi-t baseados nos estimadores $HC4_k$, $k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim (0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n = 25, 50, 75, 100$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

λ	n	α	$HC4_1$	$HC4_2$	$HC4_3$	$HC4_4$	$HC4_5$	$HC4_6$
1	25	10	0.40	0.39	1.97	1.54	4.23	5.15
		5	0.25	0.22	1.24	0.69	2.30	2.99
		1	0.10	0.08	0.52	0.14	0.66	1.00
	50	10	4.46	3.38	4.00	5.36	5.01	4.53
		5	2.44	1.93	2.50	2.27	2.42	2.15
		1	0.82	0.67	1.05	0.41	0.52	0.46
	75	10	6.11	4.43	4.22	7.13	6.58	5.79
		5	3.34	2.44	2.33	3.23	3.13	2.65
		1	0.99	0.76	0.79	0.56	0.60	0.51
	100	10	6.71	5.00	4.66	7.95	7.38	6.73
		5	3.60	2.61	2.42	3.64	3.41	3.10
		1	1.02	0.73	0.75	0.63	0.64	0.56
≈ 50	25	10	0.52	1.86	7.04	3.12	9.72	10.42
		5	0.35	1.28	5.33	1.82	7.20	7.96
		1	0.20	0.64	3.16	0.72	3.88	4.62
	50	10	4.94	2.46	2.24	6.04	4.75	4.31
		5	2.96	1.68	1.56	3.09	2.75	2.49
		1	1.17	0.94	0.91	0.82	1.03	0.90
	75	10	5.77	2.68	2.29	7.37	5.27	4.69
		5	3.29	1.51	1.31	3.81	2.81	2.49
		1	1.13	0.55	0.51	0.92	0.83	0.71
	100	10	6.24	3.14	2.73	7.93	5.79	5.22
		5	3.38	1.56	1.36	3.95	2.84	2.51
		1	1.01	0.44	0.37	0.91	0.71	0.62

Tabela 2.6 Porcentagens de rejeição de testes quasi-t baseados nos estimadores HCR4$_k$, $k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim (0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para n = 25, 50, 75, 100. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

λ	n	α	HCR4$_1$	HCR4$_2$	HCR4$_3$	HCR4$_4$	HCR4$_5$	HCR4$_6$
1	25	10	0.19	0.51	1.54	3.64	1.20	1.84
		5	0.03	0.15	0.44	0.90	0.32	0.57
		1	0.00	0.01	0.01	0.01	0.01	0.03
	50	10	3.74	2.81	3.10	7.62	3.66	3.50
		5	0.74	0.86	1.04	2.99	1.20	1.19
		1	0.01	0.05	0.10	0.21	0.08	0.09
	75	10	6.72	4.75	4.36	8.70	5.69	5.23
		5	2.19	1.53	1.44	3.80	2.19	2.01
		1	0.09	0.12	0.12	0.42	0.21	0.18
	100	10	7.73	6.11	5.62	9.03	6.75	6.37
		5	2.93	2.22	2.05	4.08	2.79	2.60
		1	0.21	0.18	0.17	0.55	0.32	0.29
≈ 50	25	10	0.07	0.29	1.08	1.75	0.52	0.81
		5	0.01	0.07	0.25	0.43	0.14	0.26
		1	0.00	0.00	0.01	0.01	0.01	0.01
	50	10	2.58	1.05	1.11	4.99	1.08	1.06
		5	0.20	0.22	0.26	1.43	0.27	0.25
		1	0.00	0.01	0.02	0.07	0.01	0.01
	75	10	5.98	1.79	1.65	7.36	2.50	2.16
		5	2.15	0.47	0.41	2.63	0.71	0.63
		1	0.04	0.01	0.01	0.16	0.03	0.03
	100	10	6.69	2.76	2.43	8.16	3.61	3.21
		5	2.76	0.76	0.65	3.30	1.18	1.05
		1	0.21	0.02	0.03	0.30	0.09	0.07

poderes dos testes para diferentes valores de β_2 (0.05, 0.10, 0.20). Os resultados encontram-se no Apêndice A.

2.6 Avaliação numérica II

O experimento de Monte Carlo da segunda avaliação numérica é baseado no modelo de regressão linear

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + e_t, \quad t = 1, \ldots, n.$$

O interesse na presente avaliação numérica reside na avaliação de teste de que $\beta_3 = 0, \beta_4 = 0$ e $\beta_5 = 0$, ou seja, $\mathcal{H}_0 : R\beta - q = 0$ versus $\mathcal{H}_1 : R\beta - q \neq 0$, em que

$$R = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} \text{ e } \quad q = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

As covariáveis correspondem a uma medida de entrada de trabalho (LAB) e a um índice de capital estocado (CAP). As covariáveis utilizadas no modelo acima apresentado são $x_{t1} = \log(\text{LAB})$, $x_{t2} = \log(\text{CAP})$, $x_{t3} = x_{t1}^2$, $x_{t4} = x_{t2}^2$ e $x_{t5} = x_{t1}x_{t2}$, havendo um total de 27 observações (fonte: Greene, 2003). Cada uma das observações de x_1, x_2, x_3, x_4 e x_5 é repetida duas, três e quatro vezes para formar amostras de tamanhos $n = 54, 81, 108$, que são mantidas constantes no experimento. Em nossos estudos de simulação utilizamos $\beta_0 = \beta_1 = \beta_2 = 1$ e $\beta_3 = \beta_4 = \beta_5 = 0$. Sob homoscedasticidade $\lambda = 1$ e sob heteroscedasticidade o valor de λ depende da especificação da função cedástica, tendo-se $\lambda > 1$.

Consideramos o conjunto de dados extraídos de Greene (2003, Tabela F6.1) denominado 'dados2'. A função cedástica foi especificada como $\sigma_t^2 = \exp\{c_1 x_{t1}\}$. As simulações sob homoscedasticidade foram realizadas com $c_1 = 0$. Para dados heteroscedásticos, consideramos $\lambda \approx 50$. No experimento de Monte Carlo foram avaliados os tamanhos dos testes quasi-F associados à hipótese nula $\mathcal{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0$. Os experimentos de simulação são baseados em 100,000 réplicas de Monte Carlo. Os principais resultados encontram-se resumidos a seguir.

Primeiro, os testes construídos a partir dos estimadores $\text{HC0}_k, k = 1, \ldots, 6$, não apresentam comportamento confiável tanto sob homoscedasticidade quanto sob heteroscedasticidade. Por exemplo, quando $\lambda \approx 50$, $n = 27$, $\alpha = 10\%$, os tamanhos dos testes $\text{HC0}_k, k = 1, \ldots, 6$, são 47.55%, 34.85%, 33.93%, 37.24%, 28.43% e 27.40% (Tabela 2.9). Contudo, ao impor a restrição de que $\beta_3 = \beta_4 = \beta_5 = 0$ considerando o mesmo cenário, as taxas ficam muito mais próximas dos níveis nominais considerados. Considerando o mesmo cenário, os testes baseados nos estimadores HCR0_1 e HCR0_4 rejeitam $\mathcal{H}_0$, respectivamente, 9.60% e 10.61% (Tabela 2.10) das vezes, ou seja, podemos concluir que o desempenho do teste baseado no estimador de White melhora acentuadamente quando impomos a restrição em teste sobre os parâmetros, principalmente quando utilizamos resíduos de MQO e de WLS, sob homoscedasticidade e sob

Tabela 2.7 Poderes de testes quasi-t baseados nos estimadores $\mathrm{HC}j_k$, $j = 0,3,4, k = 1,\ldots,6$ no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim \mathcal{N}(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n = 50$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'. A hipótese em teste é $\mathcal{H}_0 : \beta_2 = 0$ e os dados foram gerados usando $\beta_2 = 0.07$. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

λ	α	$\mathrm{HC0}_1$	$\mathrm{HC0}_2$	$\mathrm{HC0}_3$	$\mathrm{HC0}_4$	$\mathrm{HC0}_5$	$\mathrm{HC0}_6$
	10	99.52	91.83	89.13	98.97	96.91	96.13
	5	98.30	84.82	80.84	97.79	93.81	92.37
	1	89.12	67.28	62.04	92.95	82.08	79.78
	α	$\mathrm{HC3}_1$	$\mathrm{HC3}_2$	$\mathrm{HC3}_3$	$\mathrm{HC3}_4$	$\mathrm{HC3}_5$	$\mathrm{HC3}_6$
	10	99.56	87.73	83.95	98.06	95.77	94.57
1	5	99.05	78.37	72.86	96.06	91.92	89.90
	1	97.03	56.97	50.13	87.52	78.82	79.02
	α	$\mathrm{HC4}_1$	$\mathrm{HC4}_2$	$\mathrm{HC4}_3$	$\mathrm{HC4}_4$	$\mathrm{HC4}_5$	$\mathrm{HC4}_6$
	10	98.38	83.97	79.15	97.17	93.74	91.99
	5	93.72	71.99	64.79	94.11	89.06	86.36
	1	70.88	47.31	37.55	82.35	74.68	70.66
	α	$\mathrm{HC0}_1$	$\mathrm{HC0}_2$	$\mathrm{HC0}_3$	$\mathrm{HC0}_4$	$\mathrm{HC0}_5$	$\mathrm{HC0}_6$
	10	39.82	39.04	38.98	66.75	49.66	49.26
	5	27.08	24.06	23.70	58.00	35.80	35.57
	1	10.72	7.03	6.84	41.16	15.08	14.57
	α	$\mathrm{HC3}_1$	$\mathrm{HC3}_2$	$\mathrm{HC3}_3$	$\mathrm{HC3}_4$	$\mathrm{HC3}_5$	$\mathrm{HC3}_6$
	10	46.92	37.10	37.34	54.14	47.08	46.63
≈ 50	5	41.28	21.58	21.63	43.79	33.01	32.65
	1	33.76	5.58	5.24	26.20	13.01	12.99
	α	$\mathrm{HC4}_1$	$\mathrm{HC4}_2$	$\mathrm{HC4}_3$	$\mathrm{HC4}_4$	$\mathrm{HC4}_5$	$\mathrm{HC4}_6$
	10	36.07	35.74	36.09	40.66	43.28	43.32
	5	23.53	20.20	20.22	30.12	29.14	28.85
	1	8.06	4.52	4.15	15.59	10.74	10.62

Tabela 2.8 Poderes de testes quasi-*t* baseados nos estimadores HCRj_k, $j = 0,3,4, k = 1,\ldots,6$ no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim \mathcal{N}(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n = 50$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'. A hipótese em teste é $\mathcal{H}_0 : \beta_2 = 0$ e os dados foram gerados usando $\beta_2 = 0.07$. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

λ	α	HCR0_1	HCR0_2	HCR0_3	HCR0_4	HCR0_5	HCR0_6
1	10	95.43	88.79	86.20	95.88	91.07	90.03
	5	73.69	63.89	60.58	89.11	78.48	76.53
	1	10.52	10.31	8.55	59.17	42.40	39.08
	α	HCR3_1	HCR3_2	HCR3_3	HCR3_4	HCR3_5	HCR3_6
	10	97.44	78.15	75.33	94.64	82.48	80.62
	5	82.29	60.96	58.17	85.95	65.80	63.88
	1	16.10	7.48	5.99	50.47	27.09	24.88
	α	HCR4_1	HCR4_2	HCR4_3	HCR4_4	HCR4_5	HCR4_6
	10	98.76	67.54	65.40	92.84	69.55	68.05
	5	90.70	55.95	52.76	81.07	50.02	49.37
	1	27.64	9.69	7.23	38.34	13.88	12.50
≈ 50	α	HCR0_1	HCR0_2	HCR0_3	HCR0_4	HCR0_5	HCR0_6
	10	25.78	37.68	36.55	52.25	48.98	48.10
	5	11.67	22.50	21.74	38.90	33.22	32.46
	1	1.66	4.33	3.99	15.91	11.54	11.11
	α	HCR3_1	HCR3_2	HCR3_3	HCR3_4	HCR3_5	HCR3_6
	10	31.17	34.24	33.34	52.10	45.34	44.50
	5	17.62	19.71	19.40	38.63	29.59	29.03
	1	3.76	4.15	3.57	14.94	8.95	8.55
	α	HCR4_1	HCR4_2	HCR4_3	HCR4_4	HCR4_5	HCR4_6
	10	34.69	29.33	28.79	51.95	40.54	39.87
	5	23.87	16.59	15.97	37.98	24.74	24.26
	1	8.45	3.95	3.52	13.60	6.80	6.33

heteroscedasticidade. Quando resíduos robustos são usados, os desempenhos dos testes baseados nos estimadores $HC0_j$, $j = 1,\ldots,6$, melhoram significativamente, sob homoscedasticidade e sob heteroscedasticidade. Por exemplo, quando $n = 54$, $\lambda \approx 50$ e $\alpha = 5\%$, a taxa de rejeição do teste baseado no estimador $HC0_1$ é 20.91%, enquanto o teste baseado no estimador $HC0_6$ rejeita a hipótese nula 11.51% das vezes, como pode ser visto na Tabela 2.9.

Segundo, os testes baseados nos estimadores $HC3_k, k = 1,\ldots,6$, apresentam taxas de rejeição mais próximas dos níveis nominais considerados do que os testes baseados nos estimadores $HC0_k, k = 1,\ldots,6$, apenas quando não impomos restrições sobre os parâmetros. Por exemplo, para $n = 27$, $\alpha = 5\%$ e $\lambda \approx 50$, o teste baseado no estimador $HC0_1$ rejeita $\mathcal{H}_0$ 38.52% das vezes (Tabela 2.9), enquanto o teste baseado no estimador $HC3_1$ rejeita a hipótese nula apenas 8.31% das vezes, ver Tabela 2.11. Todavia, o mesmo não acontece quando o estimador restrito é utilizado. Considerando o mesmo cenário, o teste baseado no estimador $HCR0_1$ rejeita a hipótese nula 2.76% das vezes (Tabela 2.10), enquanto que o teste baseado no estimador $HCR3_1$ rejeita $\mathcal{H}_0$ 0.10% das vezes (Tabela 2.12), ou seja, seu tamanho efetivo está muito distante do nível nominal de 5%. Os testes construídos a partir do estimador $HCR3_k, k = 1,\ldots,6$, apresentam taxas de rejeição próximas dos níveis nominais apenas quando n é grande, mas vale ressaltar que os desempenhos dos testes baseados nos estimadores $HCR3_k, k = 1,\ldots,6$, (Tabela 2.12) não são superiores aos dos testes baseados nos estimadores $HC3_k, k = 1,\ldots,6$ (Tabela 2.11). Por exemplo, quando $\lambda \approx 50$, $n = 54$ e $\alpha = 10\%$, o teste baseado no estimador $HC3_2$ rejeita a hipótese nula 10.49% das vezes (Tabela 2.11) enquanto o teste baseado no estimador $HCR3_2$ (em que se impõe a restrição) rejeita a hipótese nula 4.35% das vezes, como apresenta a Tabela 2.12.

Terceiro, sob homoscedasticidade, o teste baseado no estimador $HC4_1$ apresenta taxas de rejeição mais próximas dos níveis nominais considerados do que os testes baseados em $HC4_k, k = 2,\ldots,6$. Para ilustrar a conclusão anterior, considere $\lambda = 1$, $n = 27$ e $\alpha = 5\%$. O teste baseado no estimador $HC4_1$ rejeita a hipótese nula 5.12% das vezes, enquanto o teste baseado no estimador $HC4_5$ rejeita a hipótese nula 9.16% das vezes (Tabela 2.13). Sob heteroscedasticidade, o teste baseado no estimador $HC4_1$ apresenta bom desempenho quando n é pequeno. Ao considerar o estimador restrito, os tamanhos dos testes baseados nos estimadores $HCR4_k$, $k = 1,\ldots,6$, situam-se distantes e abaixo dos níveis nominais considerados, sob homoscedasticidade e sob heteroscedasticidade (Tabela 2.14). Os testes baseados nos estimadores $HC4_k$, $k = 1,\ldots,6$, apresentam desempenhos superiores aos dos testes baseados nos estimadores $HC0_k$ e $HC3_k, k = 1,\ldots,6$. Por exemplo, sob heteroscedasticidade, $n = 54$ e $\alpha = 5\%$, as taxas de rejeição dos testes baseados nos estimadores $HC0_3$, $HC3_3$, $HC4_3$ são 14.28% (Tabela 2.9), 6.46% (Tabela 2.11) e 5.88% (Tabela 2.13), respectivamente. Todavia, ao considerar o estimador restrito, os testes baseados nos estimadores $HCR0_k$, $k = 1,\ldots,6$, apresentam comportamentos superiores aos dos testes baseados nos estimadores $HCR3_k$ e $HCR4_k$, $k = 1,\ldots,6$, como pode ser visto nas Tabelas 2.10, 2.12 e 2.14.

2.7 Aplicação

Nesta seção apresentamos uma aplicação baseada em dados reais. O nosso objetivo é apresentar um cenário em que, entre os estimadores consistentes da matriz de covariâncias, apenas o

Tabela 2.9 Porcentagens de rejeição de testes quasi-F baseados nos estimadores $HC0_k, k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + e_t$, $e_t \sim \mathcal{N}(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n = 27$, 54, 81, 108. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ - 'dados2'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS e WLS-LTS.

λ	n	α	$HC0_1$	$HC0_2$	$HC0_3$	$HC0_4$	$HC0_5$	$HC0_6$
1	27	10	41.65	30.98	29.80	32.70	24.50	23.45
		5	33.05	24.53	23.62	24.77	18.53	17.67
		1	20.20	15.30	14.90	13.47	10.67	10.17
	54	10	25.31	19.19	18.12	20.58	16.76	15.19
		5	17.60	12.81	12.14	13.53	10.30	9.61
		1	7.99	5.59	5.26	5.37	3.90	3.60
	81	10	19.98	15.71	14.62	16.76	13.65	12.65
		5	12.83	9.65	8.95	10.19	8.09	7.41
		1	4.93	3.56	3.22	3.44	2.61	2.37
	108	10	17.12	13.87	13.09	14.86	12.46	11.74
		5	10.53	8.20	7.68	8.64	7.03	6.55
		1	3.70	2.69	2.43	2.59	2.03	1.86
≈ 50	27	10	47.55	34.85	33.93	37.24	28.43	27.40
		5	38.52	28.29	27.38	28.36	21.93	21.18
		1	24.39	18.57	18.04	16.19	13.16	12.67
	54	10	29.28	21.35	20.49	23.33	18.29	17.59
		5	20.91	14.93	14.28	15.72	12.15	11.51
		1	10.32	7.15	6.86	6.76	5.15	4.88
	81	10	22.56	17.00	16.21	18.63	15.23	14.43
		5	15.22	10.94	10.44	11.72	9.24	8.76
		1	6.44	4.41	4.31	4.29	3.32	3.08
	108	10	19.14	8.51	8.03	16.29	13.59	12.89
		5	12.18	4.70	4.40	9.75	7.99	7.50
		1	4.70	1.45	1.34	3.26	2.54	2.35

Tabela 2.10 Porcentagens de rejeição de testes quasi-F baseados nos estimadores $\mathrm{HCR0}_k, k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + e_t$, $e_t \sim \mathcal{N}(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n =$ 27, 54, 81, 108. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados2'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS e WLS-LTS.

λ	n	α	$\mathrm{HCR0}_1$	$\mathrm{HCR0}_2$	$\mathrm{HCR0}_3$	$\mathrm{HCR0}_4$	$\mathrm{HCR0}_5$	$\mathrm{HCR0}_6$
1	27	10	8.98	6.07	5.58	9.73	6.24	5.72
		5	2.61	1.81	1.67	3.08	1.93	1.77
		1	0.08	0.05	0.05	0.11	0.06	0.05
	54	10	9.98	7.81	7.29	10.22	7.99	7.41
		5	3.93	3.07	2.86	4.29	4.29	3.30
		1	0.30	0.30	0.28	0.39	0.31	0.28
	81	10	10.05	8.40	7.71	10.04	8.49	7.69
		5	4.39	3.66	3.38	4.61	3.79	3.38
		1	0.48	0.45	0.43	0.59	0.48	0.43
	108	10	10.04	8.60	8.10	10.03	8.64	8.05
		5	4.54	3.86	3.59	4.59	3.85	3.49
		1	0.61	0.52	0.49	0.68	0.56	0.51
≈ 50	27	10	9.60	6.86	6.59	10.61	7.16	6.67
		5	2.76	2.18	2.12	3.25	2.22	2.12
		1	0.06	0.08	0.07	0.10	0.07	0.07
	54	10	11.04	8.42	8.04	11.53	8.80	8.25
		5	4.47	3.38	3.21	4.89	3.59	3.33
		1	0.38	0.35	0.36	0.49	0.38	0.37
	81	10	11.12	8.78	8.38	11.27	9.05	8.49
		5	4.89	3.73	3.62	5.13	3.90	3.64
		1	0.58	0.45	0.45	0.66	0.53	0.49
	100	10	11.01	8.88	8.48	10.97	9.05	8.59
		5	5.05	3.92	3.72	5.20	4.11	3.88
		1	0.67	0.52	0.53	0.78	0.60	0.58

Tabela 2.11 Porcentagens de rejeição de testes quasi-F baseados nos estimadores $HC3_k, k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + e_t,$, $e_t \sim \mathcal{N}(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para n = 27, 54, 81, 108. Os níveis nominais considerados são α=0.10,0.05,0.01 (com pontos de alavanca) - 'dados2'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS e WLS-LTS.

λ	n	α	$HC3_1$	$HC3_2$	$HC3_3$	$HC3_4$	$HC3_5$	$HC3_6$
1	27	10	11.16	13.42	14.16	9.51	10.51	10.28
		5	7.35	9.61	10.19	5.90	7.15	7.09
		1	3.18	4.85	5.38	2.31	3.37	3.37
	54	10	11.99	9.04	8.61	10.25	8.04	7.44
		5	7.26	5.42	5.14	5.74	4.40	4.07
		1	2.50	1.91	1.78	1.69	1.31	1.21
	81	10	11.46	8.64	7.94	10.09	8.10	7.40
		5	6.57	4.82	4.40	5.46	4.31	3.87
		1	2.00	1.43	1.31	1.48	1.10	0.98
	108	10	10.92	8.53	8.00	9.95	8.28	7.67
		5	6.22	4.65	4.30	5.31	4.20	3.90
		1	1.77	1.23	1.14	1.36	1.04	0.93
≈ 50	27	10	12.08	16.14	16.74	10.85	12.63	12.38
		5	8.31	12.08	12.69	6.89	9.00	8.87
		1	3.98	7.02	7.39	2.98	4.78	4.74
	54	10	13.90	10.49	10.07	11.72	9.49	8.96
		5	8.71	6.65	6.46	6.83	5.63	5.36
		1	3.44	2.83	2.74	2.39	2.04	1.94
	81	10	13.19	9.59	9.12	11.32	9.09	8.64
		5	7.87	5.61	5.29	6.39	5.16	4.77
		1	2.69	1.95	1.80	1.90	1.54	1.43
	108	10	12.29	9.24	8.72	10.93	9.08	8.57
		5	7.19	5.16	4.84	6.08	4.87	4.53
		1	2.32	1.63	1.48	1.72	1.32	1.25

Tabela 2.12 Porcentagens de rejeição de testes quasi-F baseados nos estimadores $\text{HCR3}_k, k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + e_t$, $e_t \sim \mathcal{N}(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para n = 27, 54, 81, 108. Os níveis nominais considerados são α=0.10,0.05,0.01 (com pontos de alavanca) - 'dados2'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS e WLS-LTS.

λ	n	α	HCR3_1	HCR3_2	HCR3_3	HCR3_4	HCR3_5	HCR3_6
1	27	10	0.98	1.14	1.21	1.76	1.42	1.33
		5	0.11	0.18	0.18	0.26	0.24	0.22
		1	0.00	0.00	0.00	0.00	0.00	0.00
	54	10	4.89	4.03	3.81	6.02	4.66	4.28
		5	1.43	1.31	1.24	1.94	1.56	1.38
		1	0.05	0.08	0.08	0.09	0.10	0.08
	81	10	6.75	5.70	5.24	7.38	6.17	5.57
		5	2.54	2.22	2.08	2.92	2.47	2.20
		1	0.19	0.20	0.19	0.26	0.23	0.21
	108	10	7.53	6.46	6.06	7.96	6.83	6.37
		5	3.05	2.62	2.45	3.33	2.84	2.56
		1	0.30	0.26	0.27	0.41	0.33	0.29
≈ 50	27	10	1.02	1.41	1.55	1.74	1.62	1.61
		5	0.10	2.18	0.28	0.23	0.32	0.33
		1	0.00	0.08	0.00	0.00	0.00	0.00
	54	10	5.45	4.35	4.19	6.64	5.03	4.72
		5	1.60	1.41	1.40	2.22	1.71	1.64
		1	0.06	0.10	0.11	0.12	0.12	0.12
	81	10	7.41	5.80	5.57	8.17	6.43	8.49
		5	2.75	2.18	2.10	3.28	2.54	3.64
		1	0.28	0.20	0.20	0.32	0.25	0.49
	108	10	8.25	6.67	6.31	8.77	7.18	6.71
		5	3.42	2.68	2.57	3.80	3.01	2.78
		1	0.34	0.29	0.28	0.47	0.36	0.33

Tabela 2.13 Porcentagens de rejeição de testes quasi-F baseados nos estimadores $HC4_k, k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + e_t$, $e_t \sim N(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n =$ 27, 54, 81, 108. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados2'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS e WLS-LTS.

λ	n	α	$HC4_1$	$HC4_2$	$HC4_3$	$HC4_4$	$HC4_5$	$HC4_6$
1	27	10	7.82	11.32	12.74	11.07	12.91	12.88
		5	5.12	8.05	9.24	7.16	9.16	9.18
		1	2.25	4.07	4.86	2.93	4.61	4.65
	54	10	10.23	8.08	7.81	11.19	8.94	8.31
		5	6.12	4.87	4.69	6.43	5.10	4.72
		1	2.08	1.69	1.64	2.00	1.61	1.47
	81	10	10.24	7.76	7.15	10.74	8.66	7.99
		5	5.76	4.28	3.95	5.92	4.72	4.25
		1	1.74	1.26	1.14	1.65	1.25	1.20
	108	10	9.98	7.84	7.27	10.44	8.75	8.09
		5	5.65	4.23	3.86	5.65	4.48	4.16
		1	1.56	1.08	0.99	1.47	1.13	1.02
≈ 50	27	10	8.78	14.79	16.11	12.62	15.35	15.21
		5	6.02	11.19	12.33	8.39	11.32	11.21
		1	2.95	6.57	7.36	3.81	6.28	6.21
	54	10	11.60	9.37	9.20	12.68	10.55	10.02
		5	7.16	6.01	5.88	7.58	6.42	6.07
		1	2.74	2.59	2.56	2.72	2.42	2.84
	81	10	11.69	8.67	8.25	12.02	9.80	9.34
		5	6.86	5.01	4.79	6.84	5.59	5.23
		1	2.27	1.76	1.63	2.13	1.73	1.62
	108	10	11.23	8.51	8.03	11.47	9.60	9.07
		5	6.46	4.70	4.40	6.41	5.21	4.88
		1	2.03	1.45	1.34	1.85	1.45	1.37

Tabela 2.14 Porcentagens de rejeição de testes quasi-F baseados nos estimadores HCR4$_k$, $k = 1,\ldots,6$, no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + e_t$, $e_t \sim \mathcal{N}(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para n = 27, 54, 81, 108. Os níveis nominais considerados são α=0.10,0.05,0.01 (com pontos de alavanca) - 'dados2'. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS e WLS-LTS.

λ	n	α	HCR4$_1$	HCR4$_2$	HCR4$_3$	HCR4$_4$	HCR4$_5$	HCR4$_6$
1	27	10	1.27	1.59	1.73	3.15	2.38	2.29
		5	0.18	0.30	0.32	0.59	0.53	0.48
		1	0.00	0.00	0.01	0.01	0.01	0.01
	54	10	5.07	4.26	4.08	7.05	5.55	5.02
		5	1.51	1.42	1.36	2.48	1.94	1.76
		1	0.06	0.09	0.08	0.14	0.15	0.12
	81	10	6.82	5.82	5.37	8.03	6.76	6.14
		5	2.56	2.30	2.16	3.34	2.80	2.52
		1	0.20	0.20	0.21	0.33	0.30	0.26
	108	10	7.58	6.54	6.13	8.52	7.31	6.81
		5	3.09	2.67	2.50	3.67	3.11	2.80
		1	0.30	0.28	0.29	0.48	0.39	0.35
≈ 50	27	10	1.36	1.92	2.16	3.17	2.71	2.70
		5	0.17	0.41	0.49	0.55	0.64	0.68
		1	0.00	0.01	0.01	0.00	0.01	0.11
	54	10	5.64	4.65	4.50	7.81	5.92	5.64
		5	1.70	1.55	1.55	2.79	2.13	2.03
		1	0.06	0.11	0.13	0.17	0.17	0.18
	81	10	7.54	6.01	5.77	8.99	7.07	6.67
		5	2.82	2.27	2.19	3.71	2.89	2.64
		1	0.24	0.22	0.22	0.40	0.31	0.26
	108	10	8.33	6.78	6.43	9.30	7.67	7.17
		5	3.48	2.75	2.64	4.17	3.27	3.06
		1	0.35	0.29	0.30	0.53	0.41	0.39

estimador proposto por Halbert White, HC0, e suas novas versões propostas nesse capítulo podem ser utilizadas. Sendo assim, apresentamos uma situação em que o estimador HC0 e suas modificações têm vantagem em relação os demais estimadores consistentes da matriz de covariâncias. Vale destacar que uma das principais diferenças entre o estimador HC0 e os demais estimadores apresentados na Seção 2 é o uso do t-ésimo elemento diagonal da matriz $H = X(X^\top X)^{-1}X^\top$ chamado h_t.

A aplicação é baseada nos dados apresentados por Greene (2003) e sua fonte original é o Departamento de Comércio dos Estados Unidos. A variável dependente, y, representa o gasto per capita em escolas públicas e as variáveis independentes (x, z) correspondem à renda per capita por estado, em 1979, nos Estados Unidos, e seus valores quadrados, respectivamente. Os valores para a covariável 'renda' estão reescalonados por 10^{-4}. O modelo proposto para descrever a relação entre a variável dependente e os regressores é da forma

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + e_t, \tag{2.7.1}$$

em que $t = 1, \ldots, 50$. O modelo linear em (2.7.1) são estimados pelo método dos mínimos quadrados ordinários. As estimativas pontuais são $b_0 = 832.91$, $b_1 = -1834.20$, $b_2 = 1587.04$ e o R^2 é 0.655.

Uma análise de influência das observações correspondentes aos estados revela que a observação correspondente ao Estado do Alasca é um ponto de alta alavancagem, uma vez que, para o Modelo (2.7.1), o valor de referência $3p/n$ é igual a 0.180 e o elemento diagonal da matriz H referente a esta observação é 0.651. Para examinar o impacto que a observação correspondente ao estado do Alasca (observação 2) tem na inferência resultante, Cribari–Neto (2004) estimou o Modelo (2.7.1) 50 vezes, cada vez retirando uma observação distinta. O grande impacto que esta observação (referente ao Estado do Alasca) tem nas estimativas é evidente. Quando esta observação não está na amostra a estimativa de β_2 se torna negativa (-314.139). Nos outros casos, as estimativas variam entre 1526.776 e 2113.170, com média 1603.681. Isto revela que a relação entre a média de y e x é provavelmente linear, e que a rejeição da hipótese nula de que β_2 é igual a zero pelo teste que usa o estimador HC0 é determinada por uma única observação. As inferências realizadas através de testes construídos a partir de outros estimadores consistentes, por outro lado, não são determinadas por uma única observação e apontam para uma relação linear entre as duas variáveis (Souza, 2003, p. 56).

Com base no grande impacto que a observação referente ao Estado do Alasca tem nas estimativas, propomos um outro modelo de regressão para explicar o gasto per capita em escolas públicas. Neste modelo não consideramos a covariável z, mas incluímos uma variável dummy, d, que assume valor 1 para identificar a observação de mais alta alavancagem (observação 2) e 0 para as demais observações. O modelo proposto para descrever a relação entre a variável dependente e os regressores é da forma

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 d_t + e_t, \tag{2.7.2}$$

em que $t = 1, \ldots, 50$.

Considerando o modelo proposto em (2.7.2), temos que as estimativas pontuais são b_0 =-26.81 , $b_1 = 518.31$, $b_2 = 285.39$ e o R^2 é 0.722. Nosso interesse consiste em testar a hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ contra a hipótese alternativa bicaudal. Uma vez que β_2 é o coeficiente da

covariável d_t, o nosso interesse consiste em testar se a covariável dummy é ou não relevante. É importante destacar que apenas os estimadores $HC0_k$ e $HCR0_k, k = 1,2,3,4,5,6$ podem ser utilizados, dado que os demais estimadores consistentes da matriz de covariâncias (HC3-HC4) e suas modificações propostas nesse capítulo fazem uso dos elementos $h_t's$, e neste caso o elemento h correspondente a observação 2 é 1, portanto o fator $\frac{1}{(1-h_t)}$ tem denominador nulo. Verifica-se que apenas os testes cujas estatísticas utilizam os estimadores $HC0_1$ e $HC0_4$ rejeitam a hipótese nula, sugerindo, então, que a covariável d_t é relevante para explicar o gasto per capita em escolas públicas, mas os testes cujas estatísticas utilizam as demais diferentes versões do estimador HC0 não rejeitam a hipótese nula, como apresenta a Tabela 2.15.

Na Tabela 2.16 encontram-se os erros-padrão dos elementos de $b = (b_0, b_1, b_2)$, obtidos a partir do estimador HC0 e das diferentes versões deste estimador propostas nesse capítulo apresentadas anteriormente. Os erros-padrão dos elementos de b_2 diferem bastante quando calculados com base no estimadores $HC0_1$ e $HC0_4$. Por exemplo, o erro padrão de b_2 obtido a partir do estimador $HC0_3$ é quase dez vezes maior do que aquele obtido a partir do estimador $HC0_1$.

Tabela 2.15 Inferência quasi-t, p-valores.

Teste	$HC0_1$	$HC0_2$	$HC0_3$	$HC0_4$	$HC0_5$	$HC0_6$
p-valor	<0.05	0.256	0.289	<0.05	0.255	0.287
Teste	$HCR0_1$	$HCR0_2$	$HCR0_3$	$HCR0_4$	$HCR0_5$	$HCR0_6$
p-valor	0.208	0.180	0.283	0.268	0.179	0.282

Tabela 2.16 Erros-padrão para b_2 em um modelo de regressão linear da forma $y_t = \beta_0 + \beta_1 x_t + \beta_2 d_t + e_t$, $t = 1, \ldots, 50$.

Estimador	$HC0_1$	$HC0_2$	$HC0_3$	$HC0_4$	$HC0_5$	$HC0_6$
$\sqrt{\widehat{\text{var}}(b_2)}$	26.935	250.91	269.39	26.496	250.53	269.21
Estimador	$HC0_1$	$HC0_2$	$HC0_3$	$HC0_4$	$HC0_5$	$HC0_6$
$\sqrt{\widehat{\text{var}}(b_2)}$	226.720	212.920	265.720	257.650	212.150	265.590

2.8 Conclusões

Os testes baseados nos estimadores HC0, HC3, HC4, HCR0, HCR3 e HCR4 foram avaliados utilizando seis tipos de resíduos diferentes, são eles: resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, resíduo oriundo do estimador LMS usado no estimador WLS e resíduo oriundo do estimador LTS usado no estimador WLS.

O teste construído a partir do estimador HC0 não apresenta comportamento confiável em pequenas amostras, tanto sob homoscedasticidade quanto sob heteroscedasticidade. Ao se im-

por a restrição sobre os parâmetros, o desempenho do teste baseado no estimador de White melhora notavelmente. Adicionalmente, vimos que há um ganho significativo em se utilizar resíduos robustos em substituição a resíduos MQO, sob homoscedasticidade e sob heteroscedasticidade.

O teste baseado no estimador HC3 apresenta desempenho superior ao do teste baseado no estimador HC0. Todavia, o mesmo não acontece quando o estimador restrito é utilizado. O teste construído a partir do estimador HCR3 apresenta taxas de rejeição próximas aos níveis nominais apenas quando n é grande, mas vale ressaltar que o desempenho do teste baseado no estimador HCR3 não é superior ao do teste baseado no estimador HC3.

Em relação aos tamanhos do teste baseado no estimador HC4, os resultados revelaram que este teste apresenta taxas de rejeição mais próximas dos níveis nominais quando resíduos oriundos dos estimadores MQO e WLS são utilizados, sob homoscedasticidade e heteroscedasticidade. Apesar de apresentarem taxas de rejeição inferiores aos níveis nominais considerados, o teste baseado no estimador HC4 apresenta desempenho superior ao do teste baseado no estimador HC0 em todos os casos e superior ao do teste baseado no estimador HC3 em alguns casos. Ao considerar o estimador restrito, o tamanho do teste baseado no estimador HCR4 situa-se distante e abaixo dos níveis nominais considerados, sob homoscedasticidade e sob heteroscedasticidade.

Os resultados obtidos nas simulações de poder revelaram que os poderes dos testes baseados nos estimadores estudados são semelhantes, mas vale ressaltar que os testes mais poderosos são os que utilizam resíduos de MQO e WLS. Em relação ao poder do teste baseado nos estimadores HCR0, HCR3 e HCR4, temos que em todos os casos esses testes são menos poderosos do que os testes baseados nos estimadores HC0, HC3 e HC4 (sem restrição). Vimos ainda que na presença de heteroscedasticidade os poderes de todos os testes considerados diminuem quando comparados com os resultados obtidos sob homoscedasticidade.

Em resumo, concluímos que os testes quasi-t e quasi-F baseados no estimador HC0 apresentam desempenho superior aos testes baseados nos estimadores HC3 e HC4 quando resíduos oriundos do modelo regressão restrito e robusto são considerados. Em quase todos os cenários considerados, os testes baseados no estimador HC3 apresentaram menores distorções de tamanhos quando resíduos robustos foram utilizados, contudo, essa mesma conclusão não se aplica aos testes baseados no estimador HC4, portanto, ao considerar o estimador HC4, recomendamos a utilização de resíduos de mínimos quadrados ordinários. Adicionalmente, sugerimos a utilização de resíduos oriundos do modelo de regressão restrito apenas no que pertine aos testes baseados no estimador HC0.

CAPÍTULO 3

Inferência no modelo de regressão beta com dispersão variável

3.1 Introdução

A distribuição beta é uma das distribuições mais utilizadas para modelar experimentos aleatórios que produzem resultados no intervalo $(0,1)$. Paolino (2001), Kieschnick & McCullough (2003), Ferrari & Cribari–Neto (2004), Ospina et al. (2006), Smithson & Verkuilen (2006) e Espinheira et al. (2008a, b) utilizaram modelos de regressão para situações em que a variável resposta segue a distribuição beta. Em tais modelos, assume-se que a resposta média é relacionada com um preditor linear por meio de uma função de ligação. O preditor linear envolve covariáveis e parâmetros de regressão desconhecidos. Estes modelos também são indexados por um parâmetro de dispersão, que em certas situações pode variar ao longo das observações (Smithson & Verkuilen, 2006; Espinheira et al. 2008a, 2008b; Simas et al., 2009).

A variável aleatória y possui distribuição beta se sua densidade é da forma

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, 0 < y < 1,$$

em que $p > 0$, $q > 0$ e $\Gamma(\cdot)$ é a função gama. A média e a variância de y, são, respectivamente,

$$\mathbb{E}(y) = \frac{p}{p+q} \tag{3.1.1}$$

e

$$\text{var}(y) = \frac{pq}{(p+q)^2(p+q+1)}. \tag{3.1.2}$$

Os parâmetros p e q são parâmetros de ajuste e através da escolha dos valores de p e q podem ser obtidas diferentes distribuições no intervalo $(0,1)$. A distribuição uniforme é um caso particular da distribuição beta quando $p = q = 1$. Quando $p = q = 1/2$, a distribuição beta se reduz à distribuição 'arco seno', que é utilizada no estudo de passeios aleatórios. Distribuições beta para as quais $p + q = 1$ com $p \neq 1/2$ são conhecidas como distribuições 'arco seno generalizadas'.

O nosso objetivo neste capítulo é avaliar os desempenhos de testes z e quasi-z no modelo de regressão beta através de seus tamanhos e poderes. No primeiro teste, consideramos o estimador usual da matriz de covariâncias e no segundo teste utilizamos estimadores do tipo sanduíche. No que tange aos testes z, o nosso objetivo é avaliar as consequências de má especificação no modelo de regressão beta com dispersão fixa e variável. Em relação aos testes

quasi-*z*, o nosso objetivo é avaliar o desempenho dos testes baseados nos estimadores sanduíche sem modelar a dispersão. Para tal avaliação geramos dados de duas formas distintas. Na primeira, geramos dados com dispersão variável ao longo das observações e consideramos dois contextos de estimação, a saber: estimamos o modelo supondo dispersão variável, mas a modelamos de forma incorreta através dos preditores e das funções de ligação do submodelo da média e da dispersão, e no segundo contexto, estimamos o modelo supondo dispersão constante. Na segunda forma, geramos dados com dispersão fixa, mas estimamos o modelo supondo dispersão variável e também estimamos o modelo com dispersão fixa. Vale ressaltar que nos dois contextos de geração (dispersão fixa e variável), consideramos o caso de especificação correta.

Este capítulo está dividido em seis seções. Na Seção 2, apresentamos o modelo de regressão beta. Na Seção 3, apresentamos estimadores para matriz de covariâncias de $\hat{\beta}$. Na Seção 4, apresentamos o experimento de Monte Carlo em que avaliamos os desempenhos de testes *z* e quasi-*z*, considerando diversos cenários. Na Seção 5, apresentamos uma aplicação a dados reais. A Seção 6 conclui o capítulo.

3.2 O modelo de regressão beta

Ferrari & Cribari–Neto (2004) propuseram uma parametrização diferente para a densidade beta que permite a modelagem da média da resposta através de uma estrutura de regressão e que envolve também um parâmetro de precisão. Sejam $\mu = p/(p+q)$ e $\phi = p+q$, i.e., $p = \mu\phi$ e $q = (1-\mu)\phi$, logo, as equações (3.1.1) e (3.1.2) são dadas por $\mathbb{E}(y) = \mu$ e $\text{var}(y) = \frac{V(\mu)}{1+\phi}$, respectivamente, sendo $V(\mu) = \mu(1-\mu)$ a 'função de variância', em que μ é a média da variável resposta e ϕ pode ser interpretado como um parâmetro de precisão. A densidade de y pode ser escrita, nesta parametrização, como

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, 0<y<1, \tag{3.2.1}$$

em que $0<\mu<1$ e $\phi>0$. Neste capítulo utilizaremos uma outra parametrização para o parâmetro de precisão. Escrevemos $\sigma^2 = \frac{1}{(1+\phi)}$, ou seja, $\phi = \frac{1-\sigma^2}{\sigma^2}$; logo a densidade (3.2.1) pode ser escrita como

$$f(y;\mu,\sigma) = \frac{\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right)}{\Gamma\left(\mu\frac{1-\sigma^2}{\sigma^2}\right)\Gamma\left((1-\mu)\frac{1-\sigma^2}{\sigma^2}\right)} y^{\mu\left(\frac{1-\sigma^2}{\sigma^2}\right)-1}(1-y)^{(1-\mu)\left(\frac{1-\sigma^2}{\sigma^2}\right)-1}, \tag{3.2.2}$$

em que $0<\mu<1$ e $0<\sigma<1$. Nessa parametrização os dois parâmetros que indexam a densidade assumem valores no intervalo unitário padrão $(0,1)$.

Assumiremos que a variável resposta está restrita ao intervalo $(0,1)$. Quando a resposta assume valores no intervalo (a,b), com a e b conhecidos ($-\infty<a<b<\infty$), podemos modelar a variável $(y-a)/(b-a)$, que é uma padronização que permite trabalhar com a densidade (3.2.2) no intervalo $(0,1)$.

Sejam $y_1,\ldots,y_n$ variáveis aleatórias independentes, em que cada y_t, $t=1,\ldots,n$, possui densidade (3.2.2), a média μ_t e o parâmetro de dispersão σ_t sendo desconhecidos. O modelo

proposto por Ferrari & Cribari–Neto (2004) é obtido assumindo que a média de y_t pode ser escrita como

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t,$$

em que $\beta = (\beta_1, \ldots, \beta_k)^\top$ é um vetor de parâmetros desconhecidos ($\beta \in \mathbb{R}^k$) e $x_{t1}, \ldots, x_{tk}$ são observações de k covariáveis conhecidas. Por fim, $g(\cdot)$, a função de ligação $g : (0,1) \to \mathbb{R}$, é estritamente monótona e duas vezes diferenciável. Assim, temos que $\mu_t = g^{-1}(\eta_t)$ e $\mathrm{var}(y_t) = \mu_t(1-\mu_t)\sigma_t^2$. Admitimos que o parâmetro de dispersão σ_t é dado por

$$h(\sigma_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \upsilon_t,$$

em que $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ é um vetor de parâmetros desconhecidos, $z_{t1}, \ldots, z_{tq}$ são observações de q covariáveis ($q + k < n$), assumidas fixas e conhecidas, e $h(\cdot)$ é uma função estritamente monótona e duas vezes diferenciável. Há várias possíveis escolhas para as funções de ligação $g(\cdot)$ e $h(\cdot)$, por exemplo, podemos utilizar as especificações logit e probit, entre outras. Para maiores detalhes sobre estas funções de ligação, ver McCullagh & Nelder (1989).

O logaritmo da função de verossimilhança a partir de uma amostra de n observações independentes é

$$\ell(\beta, \gamma) = \sum_{t=1}^{n} \ell_t(\mu_t, \sigma_t),$$

em que

$$\begin{aligned}\ell_t(\mu_t, \sigma_t) = {} & \log\Gamma\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \log\Gamma\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \log\Gamma\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right) \\ & + \left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2} - 1\right)\log y_t + \left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2} - 1\right)\log(1-y_t).\end{aligned}$$

Logo, para $i = 1 \ldots, k$, a função escore para β_i é dada por

$$\frac{\partial \ell(\beta, \gamma)}{\partial \beta_i} = \sum_{t=1}^{n} \frac{\partial \ell_t(\mu_t, \sigma_t)}{\mu_t}\frac{d\mu_t}{d\eta_t}\frac{\partial \eta_t}{\partial \beta_i}, \tag{3.2.3}$$

em que $d\mu_t/d\eta_t = 1/g'(\mu_t)$ e

$$\frac{\partial \ell_t(\mu_t, \sigma_t)}{\partial \mu_t} = \frac{1-\sigma_t^2}{\sigma_t^2}(y_t^* - \mu_t^*), \tag{3.2.4}$$

com $y_t^* = \log\frac{y_t}{1-y_t}$,

$$\mu^* = \psi\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \psi\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right),$$

em que $\psi(\cdot)$ é a função digama. Logo,

$$\frac{\partial\ell(\beta,\gamma)}{\partial\beta_i} = \sum_{t=1}^{n}\frac{1-\sigma_t^2}{\sigma_t^2}(y_t^* - \mu_t^*)\frac{1}{g'(\mu_t)}x_{ti}.$$

Podemos escrever o vetor $U_\beta(\beta,\gamma)$, a função escore de $\beta = (\beta_1,\ldots,\beta_k)^\top$, em forma matricial como

$$U_\beta(\beta,\gamma) = X^\top\Sigma T(y^* - \mu^*),$$

em que X é uma matriz $n\times k$ cuja t-ésima linha é $x_t^\top$, $\Sigma = \text{diag}\left\{\frac{1-\sigma_1^2}{\sigma_1^2},\ldots,\frac{1-\sigma_n^2}{\sigma_n^2}\right\}$, $y^* = (y_1^*,\ldots,y_n^*)^\top$, $T = \text{diag}\{1/g'(\mu_1),\ldots,1/g'(\mu_n)\}$ e $\mu^* = (\mu_1^*,\ldots,\mu_n^*)^\top$.

Considerando as derivadas em relação aos parâmetros que modelam a precisão, γ_j, $j = 1\ldots,q$, temos que

$$\frac{\partial\ell(\beta,\gamma)}{\partial\gamma_j} = \sum_{t=1}^{n}\frac{\partial\ell_t(\mu_t,\sigma_t)}{\partial\sigma_t}\frac{d\sigma_t}{d\upsilon_t}\frac{\partial\upsilon_t}{\partial\gamma_j}, \tag{3.2.5}$$

em que $d\sigma_t/d\upsilon_t = 1/h'(\sigma_t)$ e

$$\frac{\partial\ell_t(\mu_t,\sigma_t)}{\partial\sigma_j} = -\frac{2}{\sigma_t^3}\left[\mu_t(y_t^* - \mu_t^*) + \log(1-y_t) + \psi\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \psi\left((1-\mu_t)\frac{(1-\sigma_t^2)}{\sigma_t^2}\right)\right].$$

Então,

$$\frac{\partial\ell(\beta,\gamma)}{\partial\gamma_j} = \sum_{t=1}^{n}a_t\frac{1}{h'(\sigma_t)}z_{tj},$$

em que

$$a_t = -\frac{2}{\sigma_t^3}\left[\mu_t(y_t^* - \mu_t^*) + \log(1-y_t) + \psi\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \psi\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right].$$

Podemos escrever o vetor $U_\gamma(\beta,\gamma)$, a função escore de $\gamma = (\gamma_1,\ldots,\gamma_k)^\top$, matricialmente como

$$U_\gamma(\beta,\gamma) = Z^\top Ha,$$

em que Z é uma matriz $n\times q$ cuja t-ésima linha é $z_t^\top$, $H = \text{diag}\{1/h'(\sigma_1),\ldots,1/h'(\sigma_n)\}$ e $a = (a_1,\ldots,a_n)^\top$.

Com a finalidade de obter a matriz de informação de Fisher conjunta dos vetores β e γ, calcularemos as segundas derivadas do logaritmo da função de verossimilhança. A partir de (3.2.3), obtemos

$$\frac{\partial^2 \ell(\beta,\gamma)}{\partial\beta_i\beta_p} = \sum_{t=1}^{n} \frac{\partial}{\partial\mu_t}\left(\frac{\partial\ell_t(\mu_t,\sigma_t)}{\partial\mu_t}\frac{d\mu_t}{d\eta_t}\right)\frac{d\mu_t}{d\eta_t}\frac{\partial\eta_t}{d\beta_j}x_{ti}$$
$$= \sum_{t=1}^{n}\left(\frac{\partial^2\ell_t(\mu_t,\sigma_t)}{\partial^2\mu_t^2}\frac{d\mu_t}{d\eta_t} + \frac{\partial\ell_t(\mu_t,\sigma_t)}{\partial\mu_t}\frac{\partial}{\partial\mu_t}\frac{d\mu_t}{d\eta_t}\right)\frac{d\mu_t}{d\eta_t}x_{ti}x_{tp},$$

em que $d\mu_t/d\eta_t = 1/g'(\mu_t)$ e de (3.2.4) temos ainda que

$$\frac{\partial^2\ell_t(\mu_t,\sigma_t)}{\partial\mu_t^2} = -\frac{(1-\sigma_t^2)^2}{\sigma_t^4}\left[\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) + \psi'\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right],$$

em que $\psi'(\cdot)$ é a função trigama.

Considerando que σ_t varia ao longo das observações, a densidade (3.2.2) escrita na forma da família exponencial biparamétrica canônica é dada por

$$f(y_t;\mu_t,\sigma_t) = \exp\{\tau_1 T_1 + \tau_2 T_2 - \mathcal{A}(\tau)\}(1/y_t(1-y_t)),$$

em que $\tau = (\tau_1,\tau_2) = \left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}, \frac{1-\sigma_t^2}{\sigma_t^2}\right)$, $(T_1,T_2) = (\log\{y_t/(1-y_t)\}, \log(1-y_t))$ e

$$\mathcal{A}(\tau) = -\log\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right) + \log\Gamma\left(\mu_t\frac{1-\sigma^2}{\sigma^2}\right) + \log\Gamma\left((1-\mu_t)\frac{1-\sigma^2}{\sigma^2}\right).$$

Assim, segue que

$$\mathbb{E}(T_1) = \partial\mathcal{A}(\tau)/\partial\tau_1 = \psi\left(\mu_t\frac{1-\sigma^2}{\sigma^2}\right) - \psi\left((1-\mu_t)\frac{1-\sigma^2}{\sigma^2}\right) = \mu_t^* \tag{3.2.6}$$

e

$$\mathbb{E}(T_2) = \partial\mathcal{A}(\tau)/\partial\tau_2 = -\psi\left(\frac{1-\sigma^2}{\sigma^2}\right) + \psi\left((1-\mu_t)\frac{1-\sigma^2}{\sigma^2}\right). \tag{3.2.7}$$

Utilizando os resultados apresentados em (3.2.6) e (3.2.7), temos então que $\mathbb{E}(\partial\ell_t(\mu_t,\sigma_t)/\partial\mu_t) = 0$. Portanto,

$$\mathbb{E}\left(\frac{\partial^2\ell(\beta,\gamma)}{\partial\beta_i\partial\beta_p}\right) = \sum_{t=1}^{n}\mathbb{E}\left(\frac{\partial^2\ell_t(\mu_t,\sigma_t)}{\partial\mu_t^2}\right)\left(\frac{d\mu_t}{d\eta_t}\right)^2 x_{ti}x_{tp}.$$

Seja

$$w_t = \frac{1-\sigma_t^2}{\sigma_t^2}\left[\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) + \psi'\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right]\left(\frac{1}{g'(\mu_t)}\right)^2.$$

Logo,

$$\mathbb{E}\left(\frac{\partial^2\ell(\beta,\gamma)}{\partial\beta_i\partial\beta_p}\right) = -\sum_{t=1}^{n}\frac{(1-\sigma_t^2)}{\sigma_t^2}w_t x_{ti}x_{tp}.$$

Em termos matriciais, temos

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial\beta\partial\beta^\top}\right) = -X^\top \Sigma W X,$$

$W = \text{diag}\{w_1,\ldots,w_n\}$.

Adicionalmente, segue de (3.2.3) que as segundas derivadas de $\ell(\beta,\gamma)$ com respeito a β_i e γ_j, para $i = 1\ldots,k$ e $j = 1\ldots,q$, são dadas por

$$\begin{aligned}
\frac{\partial^2 \ell(\beta,\gamma)}{\partial\beta_i\partial\gamma_j} &= \sum_{t=1}^{n} \frac{\partial}{\partial\sigma_t}\left\{\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right)\left\{(y_t^* - \mu_t^*)\frac{1}{g'(\mu_t)}x_{ti}\right\}\frac{d\sigma_t}{d\upsilon_t}\frac{d\upsilon_t}{d\gamma_j}\right. \\
&= -\frac{2}{\sigma_t^3}\left\{y_t^* - \mu_t^* - \left(\frac{1-\sigma_t^2}{\sigma_t^2}\right)\left[\mu_t\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right.\right. \\
&- \left.\left.(1-\mu_t)\psi'\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\frac{1}{g'(\mu_t)}\frac{1}{h'(\sigma_t)}x_{ti}z_{tj}\right]\right\}.
\end{aligned}$$

Logo,

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial\beta_i\partial\gamma_j}\right) = -\sum_{t=1}^{n} c_t \frac{1}{g'(\mu_t)}\frac{1}{h'(\sigma_t)}x_{ti}z_{tj},$$

em que

$$c_t = \frac{2}{\sigma_t^3}\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right)\left[\mu_t\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) - (1-\mu_t)\psi'\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right].$$

Em forma matricial, temos

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial\beta_i\partial\gamma_j}\right) = -X^\top C T H Z,$$

em que $C = \text{diag}\{c_1,\ldots,c_n\}$.

A partir de (3.2.5) podemos obter as segundas derivadas de $\ell(\beta,\gamma)$ com respeito a γ_j e γ_l, para $j,l = 1,\ldots,q$, que são dadas por

$$\frac{\partial^2 \ell(\beta,\gamma)}{\partial\gamma_j\partial\gamma_l} = \sum_{t=1}^{n}\left(\frac{\partial^2 \ell_t(\mu_t,\sigma_t)}{\partial^2\sigma_t^2}\frac{d\sigma_t}{d\upsilon_t} + \frac{\partial\ell_t(\mu_t,\sigma_t)}{\partial\sigma_t}\frac{\partial}{\partial\sigma_t}\frac{d\sigma_t}{d\upsilon_t}\right)\frac{d\sigma_t}{d\upsilon_t}z_{tl}z_{tj},$$

em que

$$\begin{aligned}\frac{\partial^2 \ell_t(\mu_t,\sigma_t)}{\partial^2 \sigma_t^2} = -\frac{4}{\sigma_t^6}\Bigg[&-\psi'\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right)+\mu_t^2\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right)\\ &+(1-\mu_t)^2\psi'\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\Bigg]\\ &+\frac{6}{\sigma_t^4}\Bigg[\mu_t(y_t^*-\mu_t^*)+\log(1-y_t)+\psi\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right)\\ &-\psi\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\Bigg].\end{aligned}$$

Logo,

$$\mathbb{E}\left(\frac{\partial^2 \ell_t(\beta,\gamma)}{\partial\gamma_j\partial\gamma_l}\right)=\sum_{t=1}^{n}\mathbb{E}\left(\frac{\partial^2 \ell_t(\mu_t,\sigma_t)}{\partial\sigma_t^2}\right)\left(\frac{d\sigma_t}{d\upsilon_t}\right)^2 z_{tl}x_{tj},$$

pois $\mathbb{E}\left(\frac{\partial \ell_t(\mu_t,\sigma_t)}{\partial\sigma_t}\right)=0$.

Temos ainda que

$$\begin{aligned}\mathbb{E}\left(\frac{\partial^2 \ell_t(\mu_t,\sigma_t)}{\partial^2 \sigma_t^2}\right) = -\frac{4}{\sigma_t^6}\Bigg[&-\psi'\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right)+\mu_t^2\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right)\\ &+(1-\mu_t)^2\psi'\left((1-\mu_t)\frac{1-\sigma^2}{\sigma_t^2}\right)\Bigg].\end{aligned}$$

Portanto,

$$\begin{aligned}\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial\gamma_j\partial\gamma_l}\right) = -\sum_{t=1}^{n}\frac{4}{\sigma_t^6}\Bigg[&-\psi'\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right)+\mu_t^2\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right)\\ &+(1-\mu_t)^2\psi'\left((1-\mu_t)\frac{1-\sigma^2}{\sigma_t^2}\right)\Bigg]\frac{1}{\{h'(\sigma_t)\}^2}z_{tl}z_{tj}.\end{aligned}$$

Essa expressão pode ser escrita em notação matricial como

$$\mathbb{E}\left(\frac{\partial^2 \ell(\beta,\gamma)}{\partial\gamma_j\partial\gamma_l}\right)=-Z^\top D^* Z,$$

em que $D^*=\text{diag}\{d_1^*,\ldots,d_n^*\}$, com

$$d_t^*=\frac{4}{\sigma_t^6}\Bigg[-\psi'\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right)+\mu_t^2\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right)+(1-\mu_t)^2\psi'\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\Bigg]\frac{1}{\{h'(\sigma_t)\}^2}.$$

Portanto, a matriz da informação de Fisher para o vetor paramétrico $\theta=(\beta^\top,\gamma^\top)^\top$ é dada por

$$K(\theta) = K(\beta, \gamma) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\gamma} \\ K_{\gamma\beta} & K_{\gamma\gamma} \end{pmatrix},$$

em que $K_{\beta\beta} = X^\top \Sigma W X$, $K_{\beta\gamma} = (K_{\gamma\beta})^\top = X^\top C T H Z$ e $K\gamma\gamma = Z^\top D^* Z$.

Os estimadores de máxima verossimilhança de β e γ são obtidos pela solução do sistema formado por $U_\beta(\beta, \gamma) = 0$ e $U_\gamma(\beta, \gamma) = 0$. Tais estimadores não possuem forma fechada. Assim, eles têm que ser obtidos numericamente pela maximização da função de log-verossimilhança com auxílio de um algoritmo de otimização não-linear.

Usando a expressão padrão para a inversa de matrizes particionadas, temos que a inversa da matriz da informação de Fisher é dada por

$$K(\theta)^{-1} = K(\beta, \gamma)^{-1} = \begin{pmatrix} K_*^{\beta\beta} & K_*^{\beta\gamma} \\ K_*^{\gamma\beta} & K_*^{\gamma\gamma} \end{pmatrix},$$

em que

$$K_*^{\beta\beta} = (X^\top \Sigma W X - X^\top C T H Z (Z^\top D^* Z)^{-1} Z^\top H T C X)^{-1},$$

$$K_*^{\beta\gamma} = (K_*^{\gamma\beta})^\top = -K_*^{\beta\beta} X^\top C T H Z (Z^\top D^* Z)^{-1}$$

e

$$K_*^{\gamma\gamma} = (Z^\top D^* Z)^{-1} \{ I_q + (Z^\top H T C X) K_*^{\beta\beta} X^\top C T H Z (Z^\top D^* Z)^{-1} \}.$$

Sob certas condições de regularidade, temos que, para tamanhos de amostras grandes, a distribuição aproximada conjunta de $\hat{\beta}$ e $\hat{\gamma}$ é aproximadamente normal $(k+q)$-multivariada, de forma que

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix}, K(\theta)^{-1} \right),$$

aproximadamente, sendo $\hat{\beta}$ e $\hat{\gamma}$ os estimadores de máxima verossimilhança de β e γ, respectivamente.

3.3 Os estimadores sanduíche

Uma abordagem para modelos de regressão paramétricos é obter estimativas de parâmetros desconhecidos através de funções de estimação. Inferência para esses modelos é tipicamente baseada no Teorema Central do Limite em que a matriz de covariâncias é do tipo sanduíche: uma fatia de 'carne' entre duas fatias de 'pão'. Zeleis (2004) apresentou o estimador sanduíche para modelos de regressão linear, que está implementado no pacote `sandwich` do software estatístico R (Kleiber & Zeileis, 2008). Zeleis (2006) generaliza o estimador sanduíche para a classe dos modelos lineares generalizados e apresenta aplicações a dados reais para modelos probit e tobit, entre outros.

Num contexto de regressão, assumimos que temos o par $(y_t, x_t), t = 1, \ldots, n$, e um vetor θ de parâmetros k-dimensional. Em muitas situações, uma função de estimação $\xi(\cdot)$ está disponível para este tipo de modelo tal que $\mathbb{E}[\xi(y, x, \theta)] = 0$. Então, sob certas condições fracas de regularidade (White, 1994), θ pode ser estimado usando um estimador M, isto é, $\hat{\theta}$ é tal que

$$\sum_{t=1}^{n} \xi(y_t, x_t, \hat{\theta}) = 0,$$

$t = 1, \ldots, n$. Um caso particular é quando a função de estimação é a derivada de uma função objetivo $\Xi(\cdot)$:

$$\xi(y, x, \theta) = \frac{\partial \Xi(y, x, \theta)}{\partial \theta}.$$

Inferências sobre θ são tipicamente realizadas utilizando o Teorema Central do Limite, ou seja,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, S(\theta)),$$

em que $\xrightarrow{d}$ denota convergência em distribuição e $S(\theta)$ é a variância assintótica do estimador $\hat{\theta}$ (variância do tipo sanduíche), que tem a forma

$$S(\theta) = B(\theta)M(\theta)B(\theta),$$

com $B(\theta) = (\mathbb{E}[-\xi'(y, x, \theta)])^{-1}$ e $M(\theta) = \text{var}[\xi(y, x, \theta)]$; para maiores detalhes, ver White (1994). A 'carne', $M(\theta)$, do estimador sanduíche é a variância da função de estimação e o 'pão', $B(\theta)$, é a inversa do valor esperado da primeira derivada de ξ' com relação a θ.

Note que no caso de estimação por máxima verossimilhança e quando a especificação do modelo está *c*orreta, temos que

$$\mathbb{E}(-\xi'(y, x, \theta)) = \mathbb{E}(\xi(y, x, \theta)\xi(y, x, \theta)^\top),$$

em que $\xi(\cdot, \cdot, \cdot)$ denota o vetor escore. Adicionalmente, deve-se notar que

$$\mathbb{E}(\xi(y, x, \theta)\xi(y, x, \theta)^\top) = \text{var}(\xi(y, x, \theta)).$$

Assim, sob especificação correta, a estrutura de covariância se reduz a

$$B(\theta)M(\theta)B(\theta) = K(\theta)^{-1}K(\theta)K(\theta)^{-1} = K(\theta)^{-1},$$

em que $K(\cdot)$ denota a matriz de informação de Fisher. É dessa matriz, após a substituição dos parâmetros por estimativas de máxima verossimilhança, que são usualmente obtidos erros-padrão, variâncias e covariâncias estimadas para a realização de inferências, em particular testes de hipóteses (testes z). O nosso enfoque, todavia, será distinto. A proposta reside em estimar os três componentes das estrutura de covariâncias ('pão', 'carne', 'pão'), separadamente, sem lançar mão da simplificação descrita acima, que é fortemente dependente da especificação correta do modelo.

Para estimar $B(\theta)$ utilizamos a matriz jacobiana, ou seja, a média da derivada da função estimada

$$\hat{B} = \left(\frac{1}{n} \sum_{t=1}^{n} -\xi'(y_t, x_t, \hat{\theta}) \right)^{-1}.$$

Enquanto $B(\theta)$ é tipicamente estimado através da matriz jacobiana, diferentes tipos de estimadores estão disponíveis para estimar a 'carne', $M(\theta)$. Muitos desses estimadores são baseados nos valores empíricos das funções de estimação. Um estimador simples para $M(\theta)$ é o produto das funções empíricas de estimação:

$$\hat{M} = \frac{1}{n} \sum_{t=1}^{n} \xi(y_t, x_t, \hat{\theta}) \xi(y_t, x_t, \hat{\theta})^\top.$$

Este estimador corresponde ao estimador Eicker-Huber-White (Eicker, 1963; Huber, 1967; White, 1980) e também é conhecido como produto do estimador gradiente. Na prática, um ajuste dos graus de liberdade é frequentemente utilizado, isto é, o produto das funções empíricas estimadas é multiplicado por $n/(n-k)$. Com base nesta correção pelos graus de liberdade, nós propomos quatro outras modificações:

$$\hat{M}_r = \left(\frac{n}{n-p} \right)^r \hat{M},$$

em que $r = 1, 2, 3, 4$ e para $r = h$ temos que

$$\hat{M}_h = \left(\frac{n}{n-p} \right)^{1/(1-h_{\max})} \hat{M},$$

em que $h_{\max}$ correspondendo ao maior elemento diagonal da matriz de alavancagem generalizada, $\hat{W}^{1/2} X (X^\top \hat{W} X)^{-1} X^\top \hat{W}^{1/2}$, em que $\hat{W}$ denota que os parâmetros desconhecidos em W foram substituídos por suas estimativas de máxima verossimilhança. Sendo assim, temos seis variações para o estimador sanduíche: os estimadores $\hat{M}$ e $\hat{M}_1$ apresentado por Zeleis (2006) e quatro modificações que propomos para o estimador $\hat{M}$. Os estimadores do tipo sanduíche são dados por

$$\begin{aligned}
S &= \hat{B}\hat{M}\hat{B}, \\
S_1 &= \hat{B}\hat{M}_1\hat{B}, \\
S_2 &= \hat{B}\hat{M}_2\hat{B}, \\
S_3 &= \hat{B}\hat{M}_3\hat{B}, \\
S_4 &= \hat{B}\hat{M}_4\hat{B}, \\
S_h &= \hat{B}\hat{M}_h\hat{B}.
\end{aligned}$$

Nosso interesse central consiste em avaliar os desempenhos de testes de hipótese quando os estimadores do tipo sanduíche são utilizados em substituição ao estimador usual da matriz de covariâncias no modelo de regressão beta tanto sob dispersão constante quanto sob dispersão variável.

3.4 Avaliação numérica

Através de simulações de Monte Carlo nós avaliamos, em amostras de tamanho finito, os desempenhos de testes *z* (erros-padrão obtidos de forma usual) e quasi-*z* (erros-padrão sanduíche) em dois diferentes cenários num modelo de regressão beta. No primeiro cenário, cenário 1, o verdadeiro processo gerador de dados tem dispersão variável. Alguns erros de especificação no modelo foram cometidos, entre eles: (i) especificação incorreta no submodelo da média; (ii) especificação incorreta no submodelo da dispersão; (iii) especificação incorreta nos dois submodelos; (iv) a dispersão é variável, mas estimamos o modelo de regressão com dispersão fixa. No segundo cenário, cenário 2, o verdadeiro processo gerador de dados possui dispersão fixa. Neste cenário são considerados modelos de regressão beta com dispersão fixa e variável. O procedimento computacional foi desenvolvido utilizando o pacote `GAMLSS` 'generalized additive models for location, scale and shape' (Stasinopoulos et al., 2006) do software estatístico `R`. O número de réplicas de Monte Carlo foi $10,000$.

A geração de dados foi realizada sob dispersão fixa e variável. Em ambos os casos, o modelo foi estimado sob os pressupostos de que a dispersão era fixa e variável. Computamos testes *z* e quasi-*z*. Com a dispersão variável e fixa, o modelo foi estimado com a especificação correta e incorreta. Em particular, consideramos sete tipos erros de especificação do modelo: (i) o preditor linear do submodelo da dispersão está incorretamente especificado, (ii) erro na função de ligação do submodelo da média, (iii) erro na função de ligação do submodelo da dispersão, (iv) erros nas funções de ligação dos submodelos da média e da dispersão, (v) erro na função de ligação e no preditor linear do submodelo da média, (vi) erros na função de ligação e no preditor linear do submodelo da dispersão, (vii) erros nas funções de ligação dos submodelos da média e da dispersão e no preditor linear do submodelo da dispersão.

Na geração dos dados no modelo de regressão beta, consideramos cinco modelos, a saber:

$$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t \quad \text{e} \quad \text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t, \tag{3.4.1}$$

$$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t \quad \text{e} \quad \text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t, \tag{3.4.2}$$

$$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t \quad \text{e} \quad \text{cloglog}(\sigma_t) = \gamma_0 + \gamma_1 x_t, \tag{3.4.3}$$

$$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t \quad \text{e} \quad \text{cloglog}(\sigma_t) = \gamma_0 + \gamma_1 x_t, \tag{3.4.4}$$

$$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t \quad \text{e} \quad \text{logit}(\sigma_t) = \gamma_0, \tag{3.4.5}$$

$t = 1, \ldots, n$. Para estas investigações geramos x_t e z_t aleatoriamente da distribuição $\mathcal{U}(0,1)$. Medimos a intensidade da não-constância da dispersão dos dados através da seguinte quantidade:

$$\lambda = \max(\sigma_t)/\min(\sigma_t),$$

$t = 1, \ldots, n$. Os Modelos (3.4.1), (3.4.2), (3.4.3) e (3.4.4) foram utilizados no cenário 1 (dados foram gerados com dispersão variável) e o Modelo (3.4.5) foi considerado no cenário 2 (dados foram gerados com dispersão fixa). Note que $\lambda = 1$ indica que a dispersão é constante para todas as observações (cenário 2). No primeiro cenário, $\lambda \approx 8$. Nos Modelos (3.4.1), (3.4.3) e (3.4.5), os verdadeiros valores dos parâmetros são $\beta_0 = 0.5$, $\beta_1 = 0.5$ e $\beta_2 = 0$; nos Modelos (3.4.2) e (3.4.4), os verdadeiros valores dos parâmetros são $\beta_0 = 0.1$, $\beta_1 = 0.1$ e $\beta_2 = 0$. Quando a dispersão é variável, $\gamma_0 = -2.2$ e $\gamma_1 = -2.2$.

Consideramos $n = 25, 50, 75, 100$, sendo que são geradas 25 observações das variáveis x_t e z_t, que são replicadas duas, três e quatro vezes, respectivamente, para os tamanhos amostrais $n = 50, 75, 100$. Este procedimento de replicação de valores das covariáveis assegura que o grau de heterogeneidade da dispersão dos dados mantém-se constante à medida que o tamanho da amostra cresce. O nosso interesse inicial consiste em avaliar os tamanhos dos testes z e quasi-z da hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ contra $\mathcal{H}_1 : \beta_2 \neq 0$. A Tabela 3.1 apresenta com detalhes os dois cenários estudados para avaliar os desempenhos dos testes z e quasi-z quando erros de especificação no modelo são cometidos.

A seguir, apresentamos os resultados de simulações relativos às taxas de rejeição sob a hipótese nula dos diferentes testes z_v, z_f e quasi-z, aos níveis nominais de 10% e 5%. Esses resultados encontram-se na Tabela 3.2. Aqui, z_v e z_f denotam os testes z computados a partir de modelos com dispersão variável e fixa. As principais conclusões extraídas estão resumidas a seguir. Primeiro, considerando o teste z_v, no modelo de dispersão variável corretamente especificado, temos as taxas de rejeição estão acima dos níveis nominais considerados. Por exemplo, considerando $n = 50$ e $\alpha = 10\%$, a taxa de rejeição é 12.72%, como mostra a Tabela 3.2.

Segundo, nos casos em que os modelos estimados foram especificados de forma incorreta, mas em que se modela o parâmetro de dispersão (cenário 1), os nossos resultados revelam que é menos grave errar na especificação da função de ligação do que na especificação do preditor linear. Por exemplo, com $n = 75$ e $\alpha = 5\%$, o teste z_v baseado no modelo estimado com erro no preditor rejeita a hipótese nula 13.71% das vezes, enquanto que os testes baseados no modelo estimado com erro na função de ligação da média, da dispersão e de ambos (ligação e preditor) rejeitam esta hipótese 6.79%, 6.90% e 7.16% das vezes, respectivamente, como mostra a Tabela 3.2.

Terceiro, quando os dois erros de especificação são cometidos simultaneamente, ou seja, erro no preditor e na função de ligação, as taxas de rejeição apresentadas pelo teste z_v são ainda mais distantes dos níveis nominais considerados. Por exemplo, se $n = 50$ e $\alpha = 10\%$, as taxas de rejeição do teste z_v sob (i) erro na função de ligação do submodelo da média e no preditor linear do submodelo da dispersão, (ii) erro na função de ligação do submodelo da dispersão e no preditor linear do submodelo da dispersão, e (iii) erro nas funções de ligação do submodelo da média e dispersão simultaneamente e no preditor linear do submodelo da dispersão, as taxas de rejeição do teste z_v são 21.72%, 22.77% e 21.91%, respectivamente. Aqui, vale ressaltar que as taxas de rejeição do teste z_v são mais que o dobro do nível nominal considerado (10%).

Tabela 3.1 Descrição dos cenários considerados no estudo de simulação de Monte Carlo.

Cenário	Teste	Geração	
		μ	σ
1	z_f (dispersão fixa, modelo incorretamente especificado)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v(dispersão variável, modelo corretamente especificado)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erro no preditor linear)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erro na função de ligação da média)	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erro na função de ligação da dispersão)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 + x_t + \beta_2 z_t$	$\text{cloglog}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erros nas funções de ligação da média e da dispersão)	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{cloglog}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erros na função de ligação da média e no preditor linear da disp.)	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erros na função de ligação da dispersão e no preditor linear disp.)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{cloglog}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erros nas funções de lig. da média e da disp. e no pred. da disp.)	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{cloglog}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
2	z_f (dispersão fixa, modelo corretamente especificado)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0$
	z_v (dispersão variável, modelo incorretamente especificado)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0$
Cenário	Teste	Estimação	
		μ	σ
1	z_f (dispersão fixa, modelo incorretamente especificado)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0$
	z_v(dispersão variável, modelo corretamente especificado)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erro no preditor linear)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 z_t$
	z_v (erro na função de ligação da média)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erro na função de ligação da dispersão)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erros nas funções de ligação da média e da dispersão)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 x_t$
	z_v (erros nas funções de ligação da média e no preditor linear da disp.)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 z_t$
	z_v (erros na função de ligação da dispersão e no preditor linear disp.)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 z_t$
	z_v (erros nas funções de lig. da média e da disp. e no pred. da disp.)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 z_t$
2	z_f (dispersão fixa, modelo corretamente especificado)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0$
	z_v (dispersão variável, modelo incorretamente especificado)	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t + \beta_2 z_t$	$\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 z_t$

Quarto, em todos os casos estudados notamos que as taxas de rejeição ficam mais distantes dos níveis nominais quando cometemos erro de especificação no preditor do que quando estimamos o modelo desconsiderando a estrutura de regressão no parâmetro de dispersão. Para ilustrar esta conclusão, considere na Tabela 3.2, $n = 25$ e $\alpha = 5\%$. As taxas de rejeição dos testes z_v (erro no preditor linear, cenário 1) e z_f (dispersão fixa, especificação incorreta, cenário 1) são, respectivamente, 20.09% e 11.55%.

Quinto, os testes quasi-z baseados em estimadores do tipo sanduíche apresentam desempenho superior aos testes z_v na situação em que o modelo é especificado de forma incorreta. Por exemplo, considere a situação em que $n = 50$ e $\alpha = 10\%$. O teste z_v (erro na função de ligação do submodelo da dispersão e no preditor do submodelo da dispersão) baseado no estimador usual da matriz de covariâncias rejeita a hipótese nula 22.77% das vezes, enquanto que o teste baseado no estimador sanduíche S_1 rejeita esta hipótese 12.71% das vezes.

Sexto, comparando os desempenhos dos testes quasi-z baseados nos diferentes estimadores sanduíche, notamos que os testes que utilizam os estimadores S_2 e S_3 apresentam as taxas de rejeição mais próximas dos níveis nominais. Por exemplo, quando $n = 25$ e $\alpha = 5\%$, os testes baseados nesses estimadores rejeitam a hipótese nula 6.81% e 4.93% das vezes, respectivamente, enquanto os testes baseados nas demais versões dos estimadores sanduíche apresentam taxas de rejeição mais distantes do nível nominal considerado.

Sétimo, um dos resultados que mais chama a atenção na Tabela 3.2 são os desempenhos dos testes z_v na situação em que o modelo está corretamente especificado e dos teste quasi-z. Notamos que os testes baseados nos estimadores sanduíche apresentam taxas de rejeição mais próximas dos níveis nominais (em quase todos os casos) do que os testes z_v do modelo corretamente especificado. Para ilustrar, considere $n = 25$ e $\alpha = 10\%$. Temos que a taxa de rejeição do teste z_v é 16.75%, enquanto que a do teste baseado no estimador S_2 é 9.14%, como mostra a Tabela 3.2.

Oitavo, na situação em que modelamos a dispersão (σ_t) quando na realidade a dispersão é fixa (cenário 2), temos que o teste z_v apresenta taxas de rejeição mais distantes dos níveis nominais do que os testes z_v apresentados no cenário 1. Para ilustrar, considere $n = 25$ e $\alpha = 10\%$; as taxas de rejeição do teste z_v são 18.77% (dispersão fixa, especificação incorreta, cenário 2) e 16.75% (dispersão variável, especificação correta, cenário 1).

Nono, o teste quasi-z apresenta taxas de rejeição mais próximas dos níveis nominais no cenário 2 (dispersão fixa) do que no cenário 1 (dispersão variável). Por exemplo, quando $n = 25$, $\lambda = 1$ (cenário 2) e $\alpha = 10\%$, o teste τ_h rejeita a hipótese nula 10.88% das vezes, mas ao considerar $\lambda \approx 8$ (cenário 1) a taxa de rejeição do teste τ_h é 13.35%.

Objetivando avaliar os comportamentos das principais estatísticas apresentadas anteriormente quanto à habilidade de conduzir a baixas probabilidades do erro do tipo II, isto é, baixas probabilidades de não rejeitar a hipótese nula quando ela não é verdadeira, foram realizadas simulações de poder. O número de réplicas de Monte Carlo foi o mesmo usado anteriormente, ou seja, 10,000 (dez mil). Os resultados da investigação de poder, considerando $n = 25, 50, 75, 100$, encontram-se na Tabela 3.3. Foram usados valores críticos exatos (estimados via simulação). Todos os testes comparados têm, assim, tamanho correto. Para as simulações de poder consideramos os modelos de regressão beta apresentados anteriormente. A hipótese nula em teste é $\mathcal{H}_0 : \beta_2 = 0$ e os dados foram gerados utilizando $\beta_2 = 0.1$ para os Modelos (3.4.1), (3.4.3) e

Tabela 3.2 Taxas de rejeição sob $\mathcal{H}_0 : \beta_2 = 0$ dos testes z_v, z_f e τ_j ($j = 0,1,2,3,4,h$) nos Modelos (3.4.1), (3.4.2), (3.4.3), (3.4.4) e (3.4.5).

		$n = 25$		$n = 50$		$n = 75$		$n = 100$	
Cenário	Teste	10%	5%	10%	5%	10%	5%	10%	5%
1	z_f (dispersão fixa, modelo incorretamente especificado)	18.95	11.55	15.75	9.01	14.93	8.89	15.17	8.60
	z_v (dispersão variável, modelo corretamente especificado)	16.75	10.43	12.72	7.16	11.57	6.45	10.74	5.75
	z_v (erro no preditor linear)	28.12	20.09	22.17	14.43	20.71	13.71	19.82	12.47
	z_v (erro na função de ligação da média)	16.87	10.29	12.63	6.82	11.94	6.79	11.13	5.67
	z_v (erro na função de ligação da dispersão)	19.44	12.71	13.37	8.20	12.49	6.90	12.08	6.52
	z_v (erros nas funções da média e da dispersão)	19.53	12.70	13.84	7.85	12.51	7.16	11.89	6.35
	z_v (erros na função de lig. da média e no pred. da disp.)	27.11	19.85	21.72	14.39	20.20	13.28	19.39	12.34
	z_v (erros na função de lig. da disp. e no pred. da disp.)	27.40	19.83	22.77	15.44	21.42	13.73	20.05	13.28
	z_v (erros nas funções de lig. da média e da disp. e no pred. da disp.)	27.79	20.26	21.91	14.81	20.92	13.79	20.04	13.28
	τ_0(teste baseado no estimador S_0)	17.38	11.32	14.29	8.40	12.05	6.67	12.27	6.96
	τ_1(teste baseado no estimador S_1)	14.14	8.97	12.71	7.32	11.20	6.04	11.75	6.41
	τ_2(teste baseado no estimador S_2)	11.36	6.81	11.32	6.22	10.28	5.40	10.97	5.87
	τ_3(teste baseado no estimador S_3)	9.01	4.93	9.90	5.14	9.41	4.74	10.10	5.38
	τ_4(teste baseado no estimador S_4)	6.83	3.52	8.67	4.21	8.54	4.13	9.59	4.95
	τ_h(baseado no estimador S_h)	13.35	8.34	12.50	7.16	11.11	5.98	11.71	6.35
2	z_f(dispersão fixa, modelo corretamente especificado)	13.61	8.19	11.73	6.55	11.59	6.01	11.08	5.65
	z_v (dispersão variável, modelo incorretamente especificado)	18.77	11.84	13.66	8.00	12.75	6.92	11.89	6.22
	τ_0(teste baseado no estimador S_0)	14.88	9.13	12.50	6.74	11.74	6.36	11.16	5.98
	τ_1(teste baseado no estimador S_1)	11.76	6.85	10.93	5.71	10.71	5.82	10.51	5.43
	τ_2(teste baseado no estimador S_2)	9.14	4.80	9.51	4.64	9.90	5.21	9.79	5.05
	τ_3(teste baseado no estimador S_3)	6.90	3.48	8.18	3.77	9.06	4.55	9.18	4.61
	τ_4(teste baseado no estimador S_4)	4.82	2.26	6.97	3.17	8.22	4.04	8.62	4.18
	τ_h(teste baseado no estimador S_h)	10.88	6.30	10.74	5.59	10.59	5.77	10.49	5.39

(3.4.5), e $\beta_2 = 0.05$ para os Modelos (3.4.2) e (3.4.4).

Os resultados obtidos nas simulações realizadas sob a hipótese alternativa revelam que o poder do teste z_v (modelo está corretamente especificado, cenário 1) é substancialmente mais poderoso do que os dos testes quasi-z (estimadores sanduíche, cenário 1). Adicionalmente, podemos notar que dependendo do erro de especificação cometido, os poderes do testes quasi-z são maiores do que os poderes dos testes z_v sob especificação incorreta do submodelo da dispersão. Por exemplo, quando $n = 50$ e $\alpha = 10\%$, os poderes dos testes z_v sob especificação correta do modelo e quando a função de ligação do submodelo da média e o preditor linear do submodelo da dispersão estão incorretamente especificados são 84.98% e 45.64%, respectivamente, os poderes dos testes quasi-z são cerca de 56%; quando consideramos o cenário 2 (dispersão constante), os poderes dos testes quasi-z ficam próximos de 20%. Os resultados apresentados na Tabela 3.3 mostram que má especificação do modelo pode aumentar muito a probabilidade do erro tipo II. Em algumas situações, o poder do teste pode ser muito baixo, como resultado de má especificação do modelo. Inferência via teste quasi-z oferece um bom equilíbrio entre tamanho e poder.

Uma alternativa é usar uma estratégia inferencial híbrida. Nesse enfoque, primeiro realiza-se um teste da hipótese nula de dispersão fixa. Se tal hipótese nula de dispersão fixa não for rejeitada, então testa-se a hipótese envolvendo os parâmetros de regressão através do teste z usual (no modelo estimado com dispersão fixa); se, por outro lado, o teste indicar que há dispersão variável, então faz-se inferências sobre os β's usando-se ou o teste z no modelo em que há um submodelo da dispersão ou um dos testes quasi-z propostos no presente capítulo. Convém notar que a modelagem da dispersão no segundo estágio, se feita, pode conter ou não erro de especificação.

A seguir, nós apresentamos resultados de simulação em que avaliamos os desempenhos (tamanho e poder) em amostras finitas de diferentes estratégias híbridas. Os testes serão denotados por $\tau_{h_s}, s = 1,\ldots,5$. Em todos eles, usa-se o teste z usual (do modelo estimado com dispersão fixa) quando o teste inicial indicar que não há dispersão variável. As diferenças ocorrem quando a hipótese nula de dispersão fixa é rejeitada. Na nossa avaliação consideramos cinco testes. Neles, as estatísticas de teste utilizadas na segunda etapa inferencial (quando há rejeição da hipótese nula de dispersão fixa) são: as estatísticas z do modelo que incorpora estrutura de regressão para o parâmetro de dispersão em que esse submodelo está corretamente especificado ($s = 1$), em que há erro no preditor da dispersão ($s = 2$) e em que há erro de especificação na função de ligação e no preditor da dispersão ($s = 3$); por fim, usamos as estatísticas quasi-z baseadas nos estimadores S_2 ($s = 4$) e S_3 ($s = 5$). A hipótese nula de dispersão fixa (primeira etapa do processo inferencial) é testada através de um teste escore (ver Apêndice B).

A Tabela 3.4 apresenta os resultados de simulações relativos às taxas de rejeição sobre a hipótese nula de $\mathcal{H}_0 : \beta_2 = 0$, aos níveis nominais de 10% e 5%. Considerando o Modelo (3.4.1), os verdadeiros valores dos parâmetros são $\beta_0 = 0.5$, $\beta_1 = 0.5$, $\beta_2 = 0$, $\gamma_0 = -2.2$ e $\gamma_1 = -2.2$. A dispersão aqui é variável ($\lambda \approx 8$). Os testes híbridos τ_{h_4} e τ_{h_5}, baseados em estimadores do tipo sanduíche, apresentam desempenho superior aos testes híbridos τ_{h_1}, τ_{h_2} e τ_{h_3}. Por exemplo, quando $n = 25$ e $\alpha = 10\%$, o teste híbrido, τ_{h_5}, que usa uma estatística quasi-z, rejeita a hipótese nula 15.37% das vezes, enquanto o teste τ_{h_1}, cuja estatística z é obtida modelando a dispersão de forma correta, rejeita esta hipótese 18.31% das vezes; se considerarmos o teste τ_{h_2}, esta taxa

Tabela 3.3 Poderes dos teste z_v, z_f e τ_j ($j = 0,1,2,3,4,h$).

		$n = 25$		$n = 50$		$n = 75$		$n = 100$	
Cenário	Teste	10%	5%	10%	5%	10%	5%	10%	5%
1	z_f (dispersão fixa, modelo incorretamente especificado)	38.26	26.14	56.97	47.05	73.71	62.14	83.63	74.91
	z_v(dispersão variável, modelo corretamente especificado)	56.97	43.39	84.98	75.96	95.08	90.48	98.58	96.82
	z_v (erro no preditor linear)	33.54	22.66	55.39	41.57	69.66	56.41	80.37	70.82
	z_v (erro na função de ligação da média)	47.78	34.31	75.53	64.19	88.86	80.43	95.32	91.71
	z_v (erro na função de ligação da dispersão)	51.89	35.14	81.42	71.11	93.83	87.72	98.03	95.62
	z_v (erros nas funções de ligação da média e da dispersão)	41.67	26.37	53.80	41.14	65.45	54.36	75.75	66.18
	z_v (erros na função de lig. da média e no pred. da disp.)	28.23	17.98	45.64	32.18	58.18	44.33	69.00	56.65
	z_v (erros na função de lig. da disp. e no pred. da disp.)	32.03	20.99	50.88	37.36	66.61	53.32	76.70	66.59
	z_v (erros nas funções de lig. da média e da disp. e no pred. da disp.)	26.93	16.28	42.79	28.87	54.56	40.22	65.23	52.17
	τ_0(teste baseado no estimador S_0)	35.96	23.82	56.71	43.48	73.82	62.25	83.61	72.73
	τ_1(teste baseado no estimador S_1)	36.07	23.86	56.80	43.34	73.82	62.21	83.58	72.76
	τ_2(teste baseado no estimador S_2)	36.32	23.78	56.96	43.28	73.78	62.21	83.58	72.76
	τ_3(teste baseado no estimador S_3)	36.25	23.92	56.82	43.39	73.77	62.25	83.58	72.84
	τ_4(teste baseado no estimador S_4)	36.18	23.78	56.75	43.15	73.84	61.96	83.63	72.76
	τ_h(teste baseado no estimador S_h)	36.32	23.98	56.87	43.32	73.95	62.01	83.55	73.05
2	z_f(dispersão fixa, modelo corretamente especificado)	14.88	8.40	19.52	11.43	21.99	14.83	27.23	17.91
	z_v (dispersão variável, modelo incorretamente especificado)	14.78	7.95	19.04	11.33	21.75	14.39	26.56	17.45
	τ_0(teste baseado no estimador S_0)	15.52	8.61	19.97	12.20	22.74	13.26	26.96	17.47
	τ_1(teste baseado no estimador S_1)	15.50	8.49	19.99	12.12	22.54	13.18	27.12	17.47
	τ_2(teste baseado no estimador S_2)	15.51	8.48	19.73	12.28	22.74	13.25	27.07	17.47
	τ_3(teste baseado no estimador S_3)	15.59	8.39	19.97	12.16	22.64	13.21	26.98	17.50
	τ_4(teste baseado no estimador S_4)	15.75	8.52	19.93	12.27	22.52	13.20	27.26	17.55
	τ_h(teste baseado no estimador S_h)	15.47	8.53	19.66	12.10	22.76	13.26	27.10	17.66

é 26.27%.

Os resultados referentes aos poderes dos testes híbridos utilizando valores críticos exatos (estimados nas simulações de tamanho) estão apresentados na Tabela 3.5. A hipótese nula em teste é $\mathcal{H}_0 : \beta_2 = 0$ e os dados foram gerados utilizando $\beta_2 = 0.1$ e considerando o Modelo (3.4.1). Os resultados revelam que quando o submodelo da dispersão está corretamente especificado, o poder do teste híbrido τ_{h_1} é alto, mas quando especificamos a estrutura de regressão da dispersão de forma incorreta os poderes dos testes híbridos são baixos. Para ilustrar, considere $n = 50$ e $\alpha = 10\%$. O poder do teste τ_{h_1} é bem maior do que os poderes dos testes híbridos τ_{h_2} e τ_{h_3}. Os poderes dos testes τ_{h_4} e τ_{h_5} são um pouco mais altos do que os poderes dos testes τ_{h_2} e τ_{h_3} (modelagem incorreta da dispersão).

Tabela 3.4 Taxas de rejeição sob $\mathcal{H}_0 : \beta_2 = 0$ dos testes τ_{h_s}, $s = 1, \ldots, 5$, no Modelo (3.4.1).

	$n = 25$		$n = 50$		$n = 100$	
	10%	5%	10%	5%	10%	5%
τ_{h_1}	18.31	11.41	14.30	8.36	12.45	6.89
τ_{h_2}	26.27	18.26	21.38	13.69	19.48	12.27
τ_{h_3}	25.41	18.17	22.17	14.92	19.74	13.09
τ_{h_4}	16.61	10.19	13.78	7.88	12.64	6.89
τ_{h_5}	15.37	9.33	12.95	7.32	12.19	6.55

Tabela 3.5 Poderes dos testes híbridos τ_{h_s}, $s = 1, \ldots, 5$, no Modelo (3.4.1).

	$n = 25$		$n = 50$		$n = 100$	
	10%	5%	10%	5%	10%	5%
τ_{h_1}	43.51	31.96	72.20	62.08	93.67	89.92
τ_{h_2}	32.22	21.29	53.14	40.07	79.55	69.34
τ_{h_3}	30.44	19.60	49.26	36.04	75.57	64.82
τ_{h_4}	35.09	23.44	56.27	43.44	81.56	72.35
τ_{h_5}	34.32	23.17	56.33	42.70	81.15	71.98

Além das taxas de rejeições e dos poderes dos testes híbridos apresentados anteriormente considerando a dispersão variável, nós avaliamos os desempenhos destes testes no cenário em que a dispersão é fixa. O modelo considerado nesta avaliação é dado por (3.4.5). A Tabela 3.6 apresenta os resultados de simulações relativos às taxas de rejeição sobre a hipótese nula de $\mathcal{H}_0 : \beta_2 = 0$, aos níveis nominais de 10% e 5%. Para o Modelo (3.4.5), os verdadeiros valores dos parâmetros são $\beta_0 = 0.5, \beta_1 = 0.5, \beta_2 = 0$ e $\gamma_0 = -1.8$. A dispersão aqui é fixa. Neste cenário, os testes híbridos apresentam taxas de rejeição mais distantes dos níveis nominais considerados quando comparados com os testes híbridos nos cenários em que há dispersão variável. Para ilustrar, considere $n = 25$ e $\alpha = 10\%$. O teste τ_{h_4} rejeita a hipótese nula 13.06% das vezes

(ver Tabela 3.6), enquanto que quando consideramos a dispersão variável, este teste rejeita a hipótese nula 16.61% das vezes, como mostra a Tabela 3.4.

Os resultados referentes aos poderes dos testes híbridos utilizando valores críticos exatos e considerando o Modelo (3.4.5) estão apresentados na Tabela 3.7. Notamos que os poderes dos testes híbridos no cenário em que a dispersão é fixa são relativamente baixos quando comparados aos poderes dos mesmos testes no cenário em que a dispersão é variável. Por exemplo, quando $n = 100$ e $\alpha = 5\%$, o poder do teste τ_{h_1} (Tabela 3.7) é mais do que quatro vezes menor do que poder do teste τ_{h_1} quando há dispersão variável (Tabela 3.5). Ao compararmos os poderes dos testes híbridos (dispersão fixa) com os poderes dos testes z e quasi-z, notamos que os testes híbridos são menos poderosos. Para ilustrar, considere $n = 50$ e $\alpha = 10\%$, o poder do teste z_v (erro no preditor linear, cenário 1) τ_{EP} é 55.39% (Tabela 3.3), ao passo que poder do teste τ_{h_2} é apenas 18.26% (Tabela 3.7).

Tabela 3.6 Taxas de rejeição sob $\mathcal{H}_0 : \beta_2 = 0$ dos testes τ_{h_s}, $s = 1,\ldots,5$, no modelo (3.4.5).

	$n = 25$		$n = 50$		$n = 100$	
	10%	5%	10%	5%	10%	5%
τ_{h_1}	14.03	8.52	11.87	6.66	11.12	5.69
τ_{h_2}	15.86	9.86	12.78	7.33	11.48	5.91
τ_{h_3}	15.91	9.88	12.87	7.22	11.10	5.86
τ_{h_4}	13.06	7.76	11.50	6.30	10.90	5.55
τ_{h_5}	12.69	7.52	11.34	6.11	10.74	5.45

Tabela 3.7 Poderes dos testes híbridos τ_{h_s}, $s = 1,\ldots,5$, no Model (3.4.5).

	$n = 25$		$n = 50$		$n = 100$	
	10%	5%	10%	5%	10%	5%
τ_{h_1}	14.25	7.69	18.31	10.43	25.99	16.66
τ_{h_2}	13.35	7.56	18.26	10.33	25.49	16.34
τ_{h_3}	12.99	6.77	17.08	10.33	23.66	15.04
τ_{h_4}	14.04	7.62	18.01	10.48	25.92	16.88
τ_{h_5}	13.84	7.79	17.90	10.31	26.12	17.13

3.5 Aplicação

Nesta seção apresentamos uma aplicação baseada nos dados apresentados por Smithson & Verkuilen (2006). Trata-se de um estudo sobre a habilidade em leitura de um grupo de 44 crianças realizado na Faculdade de Psicologia da Universidade Nacional da Austrália. A variável

de interesse, y, é o escore em um teste de habilidade em leitura aplicado a um grupo de 44 crianças e as covariáveis são ausência ou presença de dislexia (x_1), o escore padronizado de QI não-verbal (x_2), aqui chamadas de dislexia e QI, respectivamente, e a interação entre essas duas covariáveis (x_3). Em Espinheira, Ferrari & Cribari–Neto (2008a, b) e Smithson & Verkuilen (2006), a covariável x_1 assume os valores 1 e -1 indicando se a criança é ou não disléxica, respectivamente. Aqui, contudo, tal variável dummy assume valores 1 e 0 (presença e ausência de dislexia, respectivamente).

Realizamos o teste escore (Espinheira, 2007, p. 94) em que a hipótese nula é $\mathcal{H}_0 : \sigma_1 = \ldots = \sigma_n = \sigma$, ou seja, testamos a hipótese de que a dispersão é constante. O teste escore rejeita a hipótese nula aos níveis usuais de significância, ou seja, podemos concluir que a dispersão é variável. [1]

Smithson & Verkuilen (2006) consideraram um modelo de regressão beta com ligação logit para a média com as covariáveis citadas anteriormente e concluíram que apenas a variável dislexia se mostrava significativa ao nível de significância de 5%. Os autores argumentaram que esta conclusão pode ser enganosa, já que o efeito das covariáveis na dispersão dos dados não está sendo considerado. Sendo assim, ajustamos um modelo de regressão beta com dispersão variável em que as covariáveis para média são: dislexia, QI e a interação e para dispersão consideramos apenas as covariáveis dislexia e QI, sem a interação. O modelo considerado encontra-se apresentado a seguir:

$$\begin{aligned} \text{logit}(\mu_t) &= \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3}, \\ \text{logit}(\sigma_t) &= \gamma_0 + \gamma_1 x_{t1} + \gamma_2 x_{t2}, \end{aligned}$$

em que $t = 1, \ldots, 44$. As estimativas de máxima verossimilhança dos parâmetros e os correspondentes p-valores para este modelo são apresentados na Tabela 3.8. Todas as covariáveis são significativas a níveis usuais de significância e $\hat{\lambda} = 17.70$.

Tabela 3.8 Estimativas pontuais e p-valores (modelo de regressão beta com dispersão variável).

	β_0	β_1	β_2	β_3	γ_0	γ_1	γ_2
estimativa	1.914	-1.533	1.012	-1.107	-0.365	-2.103	-0.751
p-valor	$< 4\times10^{-12}$	$< 4\times10^{-8}$	$< 6\times10^{-5}$	$< 1\times10^{-5}$	0.101	$< 1\times10^{-1}$	$< 4\times10^{-6}$

Ajustamos o modelo de regressão beta sem modelar a dispersão. As estimativas de máxima verossimilhança dos parâmetros e os correspondentes p-valores para este modelo são apresentados na Tabela 3.9. Temos que, ao considerar o estimador usual da matriz de covariâncias na construção da estatística de teste, ao nível de 10%, a covariável QI deixa ser relevante para o modelo considerado. Contudo, ao substituirmos o estimador usual da matriz de covariâncias pelos estimadores sanduíche $S_0, \ldots, S_4, S_h$ apresentados neste capítulo, temos que, ao nível

[1] Smithson & Verkuilen (2006), Espinheira, Ferrari & Cribari-Neto (2008a, b) também encontraram evidências de de dispersão não-constante.

nominal de 10%, todas as covariáveis são significativas, exceto quando consideramos os estimadores S_3 e S_4; neste caso, a covariável QI deixa ser relevante para o modelo considerado, como mostra a Tabela 3.10. As conclusões dos testes de hipótese utilizando os estimadores sanduíche revelam que chegamos às mesmas inferências acerca da relevância das covariáveis obtidas a partir do modelo de regressão beta com dispersão variável, exceto quando consideramos os estimadores S_3 e S_4.

Tabela 3.9 Estimativas pontuais e p-valores (modelo de regressão beta com dispersão fixa).

Parâmetro	β_0	β_1	β_2	β_3	γ_0
Estimativa	2.306	-1.946	0.379	-0.437	-0.909
p-valor	$< 1\times10^{-16}$	$< 3\times10^{-9}$	0.135	0.091	$< 1\times10^{-7}$

Tabela 3.10 p-valores de testes quasi-z baseados em diferentes estimadores sanduíche.

	β_0	β_1	β_2	β_3
S_0	$< 1\times10^{-15}$	$< 1\times10^{-13}$	0.059	0.034
S_1	$< 1\times10^{-15}$	$< 2\times10^{-12}$	0.075	0.046
S_2	$< 1\times10^{-15}$	$< 4\times10^{-11}$	0.094	0.060
S_3	$< 7\times10^{-14}$	$< 5\times10^{-10}$	0.117	0.077
S_4	$< 1\times10^{-12}$	$< 5\times10^{-9}$	0.137	0.096
S_h	$< 2\times10^{-16}$	$< 1\times10^{-11}$	0.084	0.053

As razões entre o efeito da variável QI (x_2) sobre a resposta média ($\partial\mu/\partial x_2$) nos casos em há e em que não há dislexia (ambas avaliadas nas correspondentes estimativas dos parâmetros e nos QIs médios de cada grupo) são 3.17 e 3.42 quando se modela σ_t e quando não se modela σ_t, respectivamente. Ou seja, esse quociente é aproximadamente o mesmo nos dois cenários.

3.6 Conclusões

Considerando o modelo de regressão beta, avaliamos os desempenhos de testes z e quasi-z em situações em que o modelo estimado foi corretamente e incorretamente especificado. Como esperávamos, para o caso em que o modelo estava corretamente especificado, as taxas de rejeição ficaram próximas dos níveis nominais considerados. Apesar de termos bons resultados neste cenário, vale ressaltar que em situações práticas é difícil modelar a dispersão de forma adequada.

Consideramos dois cenários diferentes ao estimar o modelo de regressão beta de forma incorreta. No primeiro, consideramos que havia estrutura de regressão para a dispersão, mas cometemos erros na função de ligação da média da variável resposta e na da dispersão, no preditor da estrutura de regressão da dispersão e a junção dos dois erros, i.e., função de ligação e

preditor simultaneamente. No total, sete tipos de erros de especificação foram cometidos. No que pertine aos testes *z*, nossos resultados revelaram que é mais grave errar na especificação do preditor linear da estrutura de regressão da dispersão do que na função de ligação da média da variável resposta ou na da dispersão. No segundo cenário, consideramos que não havia estrutura de regressão para a dispersão, quando na verdade deveria existir. Aqui, construímos o teste *z* utilizando o estimador usual da matriz de covariâncias e e o teste quasi-*z* utilizando os estimadores do tipo sanduíche. Percebemos que ao utilizarmos o estimador usual da matriz de covariâncias, as taxas de rejeição são bem maiores do que os níveis nominais, ou seja, o teste baseado neste estimador rejeita a hipótese nula incorretamente muito mais frequentemente do que deveria. Por outro lado, ao utilizarmos os estimadores do tipo sanduíche, as taxas de rejeição sob $\mathcal{H}_0$ ficam próximas dos níveis nominais, principalmente quando estes testes são baseados nos estimadores S_2 e S_3.

Em relação às simulações de poder, os resultados revelam que os poderes dos testes quando o modelo está corretamente especificado são altos. Quando há erros de especificação, os poderes dos testes *z* são muito mais baixos. Adicionalmente os poderes dos testes quasi-*z* são baixos comparados aos poderes dos teste *z* no modelo corretamente especificado.

Os nossos testes quasi-*z* apresentaram bom desempenho em termos de tamanho e superaram os testes *z* no cenário em que a especificação incorreta é considerada. É importante notar que os testes quasi-*z* foram inclusive competitivos com o teste *z* sob especificação correta no que tange ao controle da probabilidade de erro do tipo I.

Capítulo 4

Estimativas dos impactos dos gastos com programas assistenciais e do crescimento da economia sobre a eleição presidencial de 2006 no Brasil

4.1 Introdução

Havia em 2002 no Brasil alguns programas sociais que beneficiavam cerca de cinco milhões de famílias, como, por exemplo, o Bolsa-Escola, vinculado ao Ministério da Educação, o Auxílio-Gás, vinculado ao Ministério de Minas e Energia e o Cartão-Alimentação, vinculado ao Ministério da Saúde. O Programa Bolsa-Família consistiu na unificação desses programas sociais num único programa social, com cadastro e administração centralizados no Ministério do Desenvolvimento Social e Combate à Fome.

O Bolsa-Escola foi um programa educacional brasileiro idealizado pelo prefeito de Campinas, no Estado de São Paulo, José Roberto Magalhães Teixeira, implantado naquele município durante sua gestão no ano de 1994. O objetivo era pagar uma bolsa às famílias de jovens e crianças de baixa renda como estímulo para que essas frequentassem a escola regularmente. O programa Bolsa-Escola federal foi implementado em 2001 pelo governo do então presidente Fernando Henrique Cardoso. O Auxílio-Gás foi um programa de distribuição de renda implementado pelo Governo Federal brasileiro em 2001 para atender aos beneficiários da Rede de Proteção Social. O Cartão-Alimentação foi um programa de transferência de renda para aquisição de alimentos por meio de cartão magnético destinado às famílias atendidas pelo Programa Fome Zero. O objetivo final deste programa era garantir a emancipação socioeconômica das famílias assistidas. Por fim, o programa Bolsa-Alimentação foi criado por José Serra e teve início, em módulo piloto, em julho de 2001. Foi lançado nacionalmente em setembro de 2001. Esse programa era voltado para famílias com gestantes, mães em amamentação e crianças até 6 anos de idade, de baixa renda e em risco nutricional, sendo de caráter complementar à Bolsa-Escola, que era voltada às famílias com crianças em idade escolar.

O programa Bolsa-Família foi instituído pela Medida Provisória nº 132, de 20 de outubro de 2003, posteriormente convertida na Lei nº 10.836, de 9 de janeiro de 2004. O Bolsa-Família unificou os seguintes programas de transferência de renda do Governo Federal: Bolsa-Escola, Auxílio-Gás, Cartão-Alimentação e Bolsa-Alimentação.

O Programa Bolsa-Família oferece às famílias quatro tipos de benefícios, são eles: básico, variável, variável para jovem e variável de caráter extraordinário. O benefício básico é concedido às famílias em situação de extrema pobreza. Os benefícios variável e variável para

jovem são destinados às famílias que se encontram em situação de pobreza ou extrema pobreza, contudo, famílias que apresentam em sua composição gestante, nutrizes, crianças ou adolescentes com até 15 anos de idade recebem o benefício variável, enquanto famílias que tenham adolescentes entre 16 e 17 anos de idade recebem o benefício chamado variável para jovem. Por fim, o benefício variável de caráter extraordinário é pago nos casos em que a migração dos programas Auxílio-Gás, Bolsa-Escola, Bolsa-Alimentação e Cartão-Alimentação para o Bolsa-Família causa perdas financeiras. O Bolsa-Família é executado pelos municípios. Cabe à prefeitura realizar o cadastramento das famílias, por meio do Cadastro Único dos Programas Sociais do Governo Federal, o *CadÚnico*. A seleção das famílias elegíveis, no entanto, é feita pelo Ministério do Desenvolvimento Social e Combate à Fome (MDS).

O ex-presidente do Brasil, que é filiado ao Partido dos Trabalhadores (PT), Luiz Inácio Lula da Silva, foi cinco vezes candidato à presidência da República do Brasil. Ele foi derrotado nas três primeiras tentativas, em 1989, 1994 e 1998. Em 1989, foi vencido pelo candidato Fernando Collor de Mello e nas duas vezes seguintes, pelo candidato Fernando Henrique Cardoso. Nas eleições presidenciais de 2002 e 2006, Lula disputou em segundo turno com José Serra e Geraldo Alckmin, respectivamente, ambos filiados ao Partido da Social Democracia Brasileira (PSDB). Lula se tornou presidente da República no segundo turno da eleição presidencial de 2002 e foi reeleito no segundo turno da eleição de 2006.

Com a maior votação já recebida por um candidato no Brasil até então, em 2002, Lula foi eleito presidente da República. Sua votação foi quase o dobro da obtida por Fernando Henrique Cardoso em cada uma das vezes em que foi eleito. Partindo de uma elevada popularidade, durante o governo de 2002, Lula viu sua aceitação ruir em meio a denúncias de corrupção. No início de 2006, Lula voltou a ter altos índices de popularidade. Em 29 de outubro de 2006, Lula foi reeleito com mais de 58 milhões de votos, a maior votação recebida até então por um candidato em eleições no país. Vale destacar a diferença entre o percentual de votos de Lula e Geraldo Alckmin no segundo turno da eleição presidencial de 2006 foi de aproximadamente 22%. Comparando as eleições de 2002 e 2006, em termos de percentuais, a votação em Lula no segundo turno foi praticamente a mesma, em torno de 61%.

O desempenho eleitoral do presidente causou uma grande discussão sobre o comportamento eleitoral. Notou-se que o perfil dos eleitores de Lula havia sofrido alteração. Na sua primeira vitória (2002), Lula teve melhor desempenho nas regiões mais urbanizadas e desenvolvidas enquanto que em 2006 o voto migrou para lugares mais atrasados e menos urbanizados (Canêdo–Pinheiro, 2010). Hunter & Power (2007) concluíram que grandes disparidades regionais e a baixa penetração de revistas e jornais em muitas localidades fizeram com que parte dos eleitores, em especial, os mais pobres, não tivesse conhecimento das denúncias de corrupção que ocorreram durante o governo do presidente Lula. Segundo Zucco (2008), há evidências de que no Brasil o candidato a presidente ligado ao governo sempre obtém desempenho eleitoral relativamente melhor em regiões menos desenvolvidas.

No que tange à relação entre gastos em programas assistenciais e o desempenho do ex-presidente Lula na eleição de 2006, há uma grande discussão sobre o impacto do Bolsa-Família no resultado da eleição presidencial de 2006. Os resultados ainda são díspares. Marques et al. (2009) afirmam que tal programa teve papel decisivo na vitória do presidente Lula, enquanto que Shikida et al. (2009) sugerem que o programa não foi determinante para a reeleição de

Lula. Um outro fator que contribuiu para a mudança do padrão eleitoral do ex-presidente Lula em 2006 foi o desempenho da economia. Alguns autores defendem que os ganhos de bem-estar ligados à melhora das condições de mercado de trabalho para a parcela mais pobre da população e ao impacto do câmbio valorizado e da inflação controlada na cesta de consumo dos estratos de renda mais baixos teriam contribuído para a reeleição do ex-presidente Lula em 2006. Cânedo–Pinheiro (2010) avaliou o impacto do Bolsa-Família e o efeito do crescimento da economia na eleição presidencial de 2006 e concluiu que, tomando o Brasil como um todo, o efeito do programa Bolsa-Família na votação do ex-presidente Lula em 2006 foi de 3%, enquanto que o efeito do crescimento da economia foi de 0.34%.

Com a finalidade de avaliar o impacto dos programas assistenciais e do crescimento da economia sobre o resultado da eleição presidencial de 2006, dividimos a nossa análise em duas partes. Na primeira, nosso interesse é explicar a diferença entre as proporções de votos nos segundos turnos das eleições presidenciais de 2006 e 2002. Aqui se faz necessário o uso de modelos adequados para situações em que a variável resposta é do tipo proporção, dado que os modelos lineares usuais não são adequados. Nessa modelagem, utilizamos o modelo de regressão beta proposto por Ferrari & Cribari–Neto (2004). Na literatura há alguns autores que fazem uso de modelos de regressão linear para descrever variáveis resposta que assumem a forma de proporções, por exemplo, Zucco (2008), Shikida et al. (2009) e Marques et al. (2009). No modelo de regressão beta utilizado para explicar a diferença das proporções de votos do ex-presidente Lula nas eleições de 2006 e 2002, além de considerar variáveis geográficas, políticas e eleitorais e características socioeconômicas dos eleitores, consideramos a diferença entre os gastos per capita nos programas assistenciais de 2006 e 2002 e a diferença dos níveis do PIB per capita de 2006 e 2002. Na segunda parte, estimamos o impacto dos gastos com programas assistenciais e do crescimento da economia sobre eleições presidenciais de 2006. O nosso interesse aqui reside em estimar a quantidade de votos que o ex-presidente Lula teria em duas diferentes situações. Primeira, se os gastos com programas assistenciais fossem mantidos nos níveis de 2002, e segunda, se o PIB per capita de 2006 fosse o mesmo de 2002.

Além desta introdução, este capítulo está dividido em cinco seções. Na Seção 4.2, apresentamos o modelo de regressão beta, bem como algumas medidas de avaliação da qualidade do ajuste. Na Seção 4.3, descrevemos os dados utilizados. Na Seção 4.4, especificamos o modelo de regressão beta considerado, bem como as inferências associadas. Por fim, a seção 4.5 conclui o capítulo.

4.2 O modelo de regressão beta

Ferrari & Cribari–Neto (2004) propuseram uma parametrização alternativa para a densidade beta que permite a modelagem da média da resposta através de uma estrutura de regressão e que envolve também um parâmetro de precisão. A função de densidade beta nessa reparametrização tem a forma

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0<y<1, \tag{4.2.1}$$

em que $0<\mu<1$ e $\phi>0$. Aqui, $\mathbb{E}(y)=\mu$ e $\text{var}(y)=\frac{\mu(1-\mu)}{1+\phi}$, sendo $V(\mu)=\mu(1-\mu)$ a 'função

de variância', em que μ é a média da variável resposta e ϕ pode ser interpretado como um parâmetro de precisão.

Neste capítulo utilizaremos uma outra parametrização para o parâmetro de precisão. Nós escrevemos $\sigma^2 = \frac{1}{(1+\phi)}$, ou seja, $\phi = \frac{1-\sigma^2}{\sigma^2}$. Note que σ é um parâmetro de *dispersão*. Logo, a densidade em (4.2.1) é dada por

$$f(y;\mu,\sigma) = \frac{\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right)}{\Gamma\left(\mu\frac{1-\sigma^2}{\sigma^2}\right)\Gamma\left((1-\mu)\frac{1-\sigma^2}{\sigma^2}\right)} y^{\mu\left(\frac{1-\sigma^2}{\sigma^2}\right)-1}(1-y)^{(1-\mu)\left(\frac{1-\sigma^2}{\sigma^2}\right)-1}, \tag{4.2.2}$$

em que $0 < \mu < 1$ e $0 < \sigma < 1$. Nessa parametrização os dois parâmetros que indexam a densidade assumem valores no intervalo unitário padrão $(0,1)$.

Assumiremos que a variável resposta está restrita ao intervalo $(0,1)$. Quando a resposta assume valores no intervalo (a,b), com a e b conhecidos ($-\infty < a < b < \infty$), podemos modelar a variável $(y-a)/(b-a)$, que é uma padronização que permite trabalhar com a densidade (4.2.2) no intervalo $(0,1)$.

Sejam $y_1,\ldots,y_n$ variáveis aleatórias independentes, em que cada y_t, $t = 1,\ldots,n$, possui densidade (4.2.2), a média μ_t e o parâmetro de dispersão σ_t sendo desconhecidos. O modelo de regressão beta (Ferrari & Cribari–Neto, 2004) é obtido assumindo que a média de y_t pode ser escrita como

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t,$$

em que $\beta = (\beta_1,\ldots,\beta_k)^\top$ é um vetor de parâmetros desconhecidos ($\beta \in \mathbb{R}^k$) e $x_{t1},\ldots,x_{tk}$ são observações de k covariáveis. Por fim, a função de ligação $g : (0,1) \to \mathbb{R}$ é estritamente monótona e duas vezes diferenciável. Assim, temos que $\mu_t = g^{-1}(\eta_t)$ e $\mathrm{var}(y_t) = \mu_t(1-\mu_t)\sigma_t^2$. Admitimos que o parâmetro de dispersão σ_t é dado por

$$h(\sigma_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \nu_t,$$

em que $\gamma = (\gamma_1,\ldots,\gamma_q)^\top$ é um vetor de parâmetros desconhecidos, $z_{t1},\ldots,z_{tq}$ são observações de q covariáveis ($q + k < n$), assumidas fixas e conhecidas, e $h : (0,1) \to \mathbb{R}$ é uma função estritamente monótona e duas vezes diferenciável. Uma interpretação interessante dos parâmetros que definem a estrutura de regressão beta é obtida quando a função de ligação é logit. Neste caso, a média de y_t, μ_t, pode ser escrita como

$$\mu_t = \frac{e^{x_t^\top\beta}}{1+e^{x_t^\top\beta}},$$

em que $x_t^\top = (x_{t1},\ldots,x_{tk}), t = 1,\ldots,n$. Neste contexto, os parâmetros do modelo de regressão beta têm importantes interpretações. Suponha que o valor da t-ésima covariável aumenta em c unidades e que todas as outras covariáveis independentes permanecem constantes. Seja $\mu^\dagger$ a média da variável y sob o novo valor da covariável e seja μ a média da variável y sob os valores originais das covariáveis. Sendo assim, a razão de chances (*odds ratio*) é dada por

$$e^{c\beta_i} = \frac{\mu^\dagger/(1-\mu^\dagger)}{\mu/(1-\mu)},$$

ou seja, $e^{c\beta_i}$ é igual à razão de chances.

O logaritmo da função de verossimilhança a partir de uma amostra de n observações independentes é

$$\ell(\beta,\gamma) = \sum_{t=1}^{n} \ell_t(\mu_t,\sigma_t),$$

em que

$$\ell_t(\mu_t,\sigma_t) = \log\Gamma\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \log\Gamma\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \log\Gamma\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)$$
$$+\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}-1\right)\log y_t + \left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}-1\right)\log(1-y_t).$$

Logo, para $i = 1,\ldots,k$, a função escore para β_i é dada por

$$\frac{\partial\ell(\beta,\gamma)}{\partial\beta_i} = \sum_{t=1}^{n} \frac{\partial\ell_t(\mu_t,\sigma_t)}{\mu_t}\frac{d\mu_t}{d\eta_t}\frac{\partial\eta_t}{\partial\beta_i},$$

em que $d\mu_t/d\eta_t = 1/g'(\mu_t)$ e

$$\frac{\partial\ell_t(\mu_t,\sigma_t)}{\partial\mu_t} = \frac{1-\sigma_t^2}{\sigma_t^2}(y_t^* - \mu_t^*),$$

com $y_t^* = \log\frac{y_t}{1-y_t}$,

$$\mu_t^* = \psi\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \psi\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right),$$

em que $\psi(\cdot)$ é a função digama. Logo,

$$\frac{\partial\ell(\beta,\gamma)}{\partial\beta_i} = \sum_{t=1}^{n} \frac{(1-\sigma_t^2)}{\sigma_t^2}(y_t^* - \mu_t^*)\frac{1}{g'(\mu_t)}x_{ti}.$$

Podemos escrever o vetor score $U_\beta(\beta,\gamma)$, a função escore de $\beta = (\beta_1,\ldots,\beta_k)^\top$, em forma matricial como

$$U_\beta(\beta,\gamma) = X^\top\Sigma T(y^* - \mu^*),$$

em que X é uma matriz $n\times k$ cuja t-ésima linha é $x_t^\top$, $\Sigma = \text{diag}\left\{\frac{1-\sigma_1^2}{\sigma_1^2},\ldots,\frac{1-\sigma_n^2}{\sigma_n^2}\right\}$, $T = \text{diag}\{1/g'(\mu_1),\ldots,1/g'(\mu_n)\}$, $y^* = (y_1^*,\ldots,y_n^*)^\top$ e $\mu^* = (\mu_1^*,\ldots,\mu_n^*)^\top$.

Considerando as derivadas em relação aos parâmetros que modelam a dispersão, $\gamma_j, j = 1,\ldots,q$, temos que

$$\frac{\partial \ell(\beta,\gamma)}{\partial \gamma_j} = \sum_{t=1}^{n} \frac{\partial \ell_t(\mu_t,\sigma_t)}{\partial \sigma_t}\frac{d\sigma_t}{d\nu_t}\frac{\partial \nu_t}{\partial \gamma_j},$$

em que $d\sigma_t/d\nu_t = 1/h'(\sigma_t)$ e

$$\frac{\partial \ell_t(\mu_t,\sigma_t)}{\partial \sigma_j} = -\frac{2}{\sigma_t^3}\left[\mu_t(y_t^* - \mu_t^*) + \log(1-y_t) + \psi\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \psi\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right].$$

Então,

$$\frac{\partial \ell(\beta,\gamma)}{\partial \gamma_j} = \sum_{t=1}^{n} a_t \frac{1}{h'(\sigma_t)} z_{tj},$$

em que

$$a_t = -\frac{2}{\sigma_t^3}\left[\mu_t(y_t^* - \mu_t^*) + \log(1-y_t) + \psi\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \psi\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right].$$

Podemos escrever o vetor $U_\gamma(\beta,\gamma)$, a função escore de $\gamma = (\gamma_1,\ldots,\gamma_q)^\top$, matricialmente como

$$U_\gamma(\beta,\gamma) = Z^\top H a,$$

em que Z é uma matriz $n \times q$ cuja t-ésima linha é $z_t^\top$, $H = \text{diag}\{1/h'(\sigma_1),\ldots,1/h'(\sigma_n)\}$ e $a = (a_1,\ldots,a_n)^\top$.

A matriz de informação de Fisher conjunta dos vetores β e γ é dada pelo negativo dos valores esperados das segundas derivadas do logaritmo da função de verossimilhança. Para o modelo de regressão beta na parametrização usada aqui, a matriz da informação de Fisher para o vetor paramétrico $\theta = (\beta^\top, \gamma^\top)^\top$ pode ser escrita como

$$K(\theta) = K(\beta,\gamma) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\gamma} \\ K_{\gamma\beta} & K_{\gamma\gamma} \end{pmatrix},$$

em que $K_{\beta\beta} = X^\top \Sigma W X$, onde $W = \text{diag}\{w_1,\ldots,w_n\}$, com

$$w_t = \frac{1-\sigma_t^2}{\sigma_t^2}\left[\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) + \psi'\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right]\left(\frac{1}{g'(\mu_t)}\right)^2,$$

$K_{\beta\gamma} = (K_{\gamma\beta})^\top = X^\top C T H Z$, onde $C = \text{diag}\{c_1,\ldots,c_n\}$, com

$$c_t = \frac{2}{\sigma_t^3}\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right)\left[\mu_t\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) - (1-\mu_t)\psi'\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right]$$

e $K\gamma\gamma = Z^\top D Z$, onde $D = \text{diag}\{d_1,\ldots,d_n\}$, com

$$d_t = \frac{4}{\sigma_t^6}\left[-\psi'\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right) + \mu_t^2\psi'\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) + (1-\mu_t)^2\psi'\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)\right]\frac{1}{\{h'(\sigma_t)\}^2}.$$

Os estimadores de máxima verossimilhança de β e γ são obtidos a partir da solução do sistema formado por $U_\beta(\beta,\gamma) = 0$ e $U_\gamma(\beta,\gamma) = 0$. Tais estimadores não possuem forma fechada. Assim, eles têm que ser obtidos numericamente pela maximização da função de log-verossimilhança com auxílio de um algoritmo de otimização não-linear.

Usando a expressão padrão para a inversa de uma matriz particionada, temos que a inversa da matriz da informação de Fisher é dada por

$$K(\theta)^{-1} = K(\beta,\gamma)^{-1} = \begin{pmatrix} K_*^{\beta\beta} & K_*^{\beta\gamma} \\ K_*^{\gamma\beta} & K_*^{\gamma\gamma} \end{pmatrix},$$

em que

$$K_*^{\beta\beta} = (X^\top \Sigma W X - X^\top C T H Z (Z^\top D Z)^{-1} Z^\top H T C X)^{-1},$$

$$K_*^{\beta\gamma} = (K_*^{\gamma\beta})^\top = -K_*^{\beta\beta} X^\top C T H Z (Z^\top D Z)^{-1}$$

e

$$K_*^{\gamma\gamma} = (Z^\top D Z)^{-1}\{I_q + (Z^\top H T C X) K_*^{\beta\beta} X^\top C T H Z (Z^\top D Z)^{-1}\}.$$

Sob certas condições de regularidade, temos que, para tamanhos de amostras grandes, a distribuição conjunta de $\hat\beta$ e $\hat\gamma$ é aproximadamente normal $(k+q)$-multivariada:

$$\begin{pmatrix} \hat\beta \\ \hat\gamma \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix}, K(\theta)^{-1}\right),$$

aproximadamente, sendo $\hat\beta$ e $\hat\gamma$ os estimadores de máxima verossimilhança de β e γ, respectivamente.

4.2.1 Medidas de diagnóstico e adequabilidade para modelos de regressão beta

A análise de diagnóstico teve início com a análise de resíduos objetivando detectar pontos mal ajustados ou aberrantes e avaliar indícios de afastamento das suposições sobre o modelo. A análise de resíduos pode basear-se nos resíduos ordinários ou em suas várias padronizações. Técnicas gráficas baseadas em resíduos são frequentemente adotadas para análise de diagnósticos.

Espinheira et al. (2008a) propuseram um resíduo para o modelo de regressão beta e constataram que o mesmo tipicamente apresenta desempenho superior ao resíduo proposto por Ferrari & Cribari-Neto (2004), especialmente no sentido de identificar observações influentes para as estimativas das médias. O resíduo ponderado padronizado 2 proposto por esses autores, r^{pp}, é dado por

$$r_t^{pp} = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{\hat{v}_t(1-h_{tt}^*)}}, \tag{4.2.3}$$

em que h_{tt}^* é o t-ésimo elemento da diagonal de $(\widehat{W}\widehat{\Sigma})^{1/2}X(X^\top\widehat{\Sigma}\widehat{W}X)^{-1}X^\top(\widehat{\Sigma}\widehat{W})^{1/2}$ em que $\widehat{W}$ denota que os parâmetros desconhecidos em W foram substituídos por suas estimativas de máxima verossimilhança, $\hat{v}_t = \psi'\left(\hat{\mu}_t\frac{1-\hat{\sigma}_t^2}{\hat{\sigma}_t^2}\right) + \psi'\left((1-\hat{\mu}_t)\frac{1-\hat{\sigma}_t^2}{\hat{\sigma}_t^2}\right)$ e $\widehat{\Sigma} = \text{diag}\left\{\frac{1-\hat{\sigma}_1^2}{\hat{\sigma}_1^2},\ldots,\frac{1-\hat{\sigma}_n^2}{\hat{\sigma}_n^2}\right\}$. Aqui, $\hat{\mu}_t = g^{-1}(\hat{\eta}_t) = g^{-1}(x_t^\top\hat{\beta})$ e $\hat{\sigma}_t^2 = h^{-1}(\hat{v}_t) = h^{-1}(z_t^\top\hat{\gamma})$, em que x_t e z_t são as t-ésimas linhas de X e Z, respectivamente.

Para verificar a qualidade global do ajuste de um modelo é útil ter uma medida análoga ao coeficiente de determinação, R^2, utilizado em modelos lineares de regressão. Ferrari e Cribari-Neto (2004) propuseram uma medida global chamada de pseudo R^2, que é definida como o quadrado do coeficiente de correlação entre $\hat{\eta}$ e $g(y)$. Na evolução dos métodos de diagnóstico uma etapa que se mostrou relevante foi a detecção de observações que exercem efeito desproporcional no ajuste. Neste contexto, encontram-se a distância de Cook e a alavancagem generalizada.

A distância de Cook (1977) objetiva quantificar o impacto de uma observação particular nas estimativas dos coeficientes de regressão a partir de sua exclusão do conjunto de dados. Este tipo de diagnóstico pode ser feito utilizando o método de influência local desenvolvido por Cook (1986). A análise de influência local visa avaliar conjuntamente o impacto das observações sob pequenas perturbações no modelo ou nos dados. A proposta de influência local apresentada por Cook (1986) tem sido vastamente utilizada em modelagens de regressão.

A distância de Cook para o modelo de regressão beta definido em (4.2.1) apresentada em Espinheira et al. (2008a) é dada por

$$LD_t = \frac{h_{tt}^*}{1-h_{tt}^*}(r_t^{pp})^2,$$

em que r_t^{pp} é o resíduo ponderado padronizado 2 definido em (4.2.3).

4.3 Descrição dos dados

A Tabela 4.1 apresenta uma breve descrição das variáveis utilizadas. As fontes dos dados são Zucco (2008), o Tribunal Superior Eleitoral (TSE) e o Instituto Brasileiro de Geografia e Estatística (IBGE).

Na Tabela 4.2 apresentamos algumas estatísticas descritivas: mínimo, primeiro quartil ($Q_{1/4}$), mediana, média, terceiro quartil ($Q_{3/4}$) e máximo das variáveis utilizadas. Tais estatísticas descritivas para as variáveis são calculadas com base em 4864 observações. Algumas conclusões podem ser extraídas da Tabela 4.2. Considerando as variáveis $GASTO2002$ e $GASTO2006$, temos que há grande disparidade nos gastos com os programas assistenciais em 2002 e 2006; o valor médio dos gastos em 2002 foi de aproximadamente 109 mil reais, enquanto que em 2006 o gasto médio foi quase um milhão e meio de reais. A variável $\Delta GASTO$

Tabela 4.1 Descrição das variáveis utilizadas.

Variável	Definição
$LULA2006$	Proporção de votos válidos de Lula no segundo turno da eleição de 2006
$LULA2002$	Proporção de votos válidos de Lula no segundo turno da eleição de 2002
Δy ($\Delta VOTOS$)	Diferença entre a proporção de votos válidos de Lula
$GASTO2002$	Gastos, em milhões de reais, nos programas assistenciais em 2002 [a]
$GASTO2006$	Gastos, em milhões de reais, nos programas assistenciais em 2006
$\Delta GASTO$	Diferença per capita dos gastos em programas assistenciais em 2006 e 2002
URB	Proporção da população vivendo em áreas urbanas em 2000
$ANALF$	Proporção de analfabetos entre pessoas acima de 15 anos de idade em 2000
GOV	Variável dummy: 1 se governador do PT em 2006, 0 caso contrário
$PREF$	Variável dummy: 1 se prefeito do PT em 2006, 0 caso contrário
$RENDA$	Renda per capita em 2000
$PENT$	Proporção da população de orientação religiosa pentecostal
$POP2002$	População estimada do município em 2002
$POP2006$	População estimada do município em 2006
DU	Variável dummy: 1 se $POP2006$>50000 habitantes, 0 caso contrário
$CRESC$	Taxa de crescimento do PIB per capita entre 2006 e 2002

[a] Os gastos de 2002 encontram-se a preços constantes de 2006.

representa a diferença do quociente entre gasto em programas assistenciais em 2006 e a população estimada em 2006 e o quociente entre gasto em programas assistenciais em 2002 e a população estimada em 2002, portanto, $\Delta GASTO$ é definida pela diferença entre os gastos per capita de 2006 e 2002. O valor médio da diferença entre os gastos per capita de 2006 e 2002 é de aproximadamente 59 reais.

Através da variável $CRESC$ é possível avaliar o crescimento desempenho da economia de 2002 para 2006. Essa variável é obtida através do quociente entre a diferença entre os PIB's per capita de 2006 e 2002 e o PIB per capita 2002. O crescimento econômico médio dos municípios brasileiros foi de 0.57%.

Em 50% dos municípios brasileiros o percentual de votos válidos do ex-presidente Lula no segundo turno da eleição de 2002 foi quase a 54% (Tabela 4.2), enquanto que em 2006 esse percentual foi superior a 61%. A menor renda per capita da amostra é de 28.38 reais e a maior, 954.60 reais. O percentual de analfabetismo nos municípios mais desenvolvidos é inferior a 1%; nos menos desenvolvidos este percentual chega a ultrapassar 50%. O terceiro quartil da variável $ANALF$ é 0.315. Isto significa que 75% dos municípios têm preponderância de analfabetos igual ou menor que 31.50%.

Segundo a revista britânica *The Economist*, o Bolsa-Família contribuiu para a redução dos índices de pobreza do Brasil, seu impacto sendo mais pronunciado nas zonas rurais. Assim sendo, incorporamos nas nossas análises a variável *dummy* DU com o objetivo de diferenciar municípios urbanos e rurais. Em nosso estudo o município é classificado como urbano se sua população estimada em 2006 exceder 50 mil habitantes ($DU = 1$) e é classificado como rural caso contrário. Em nossa análise 531 municípios foram classificados como urbanos. A população estimada nesses 531 municípios corresponde a aproximadamente 66% da população total.

Tabela 4.2 Estatísticas descritivas das variáveis utilizadas.

Variável	mínimo	$Q_{1/4}$	mediana	média	$Q_{3/4}$	máximo
$LULA2002$	0.125	0.453	0.538	0.539	0.623	0.934
$LULA2006$	0.196	0.486	0.614	0.618	0.757	0.963
Δy ($\Delta VOTOS$)	−0.443	−0.070	0.051	0.079	0.213	0.705
$GASTO2002$	124	22500	51710	109000	117500	7707000
$GASTO2006$	4851	264300	644500	1452000	1529000	130500000
$\Delta GASTO$	1.74	28.11	49.92	58.72	87.57	378.10
URB	0.016	0.414	0.601	0.595	0.785	1.000
$ANALF$	0.009	0.115	0.176	0.214	0.315	0.607
$RENDA$	28.38	88.78	169.90	173.70	235.00	954.60
$PENT$	0.001	0.039	0.072	0.082	0.112	0.461
$POP2002$	808	5272	10900	32230	22600	10660000
$POP2006$	828	5293	11130	34140	23570	11020000
$CRESC$	0.001	0.384	0.522	0.570	0.678	12.93

Destacamos que a menor proporção de votos que o ex-presidente Lula recebeu em 2002 foi registrada no município de Juruá, no Estado do Amazonas, e a maior, no Estado da Bahia, no município de São Francisco do Conde. Em 2006, a menor e a maior proporções de votos foram registradas nos municípios de São Marcos (Rio Grande do Sul) e Bela Água (Maranhão), respectivamente. Vale destacar que na eleição presidencial de 2002 a proporção de votos válidos obtidos pelo ex-presidente Lula foi um pouco menor do que em 2006.

Em 2002, existiam três programas assistenciais, a saber: Bolsa-Escola, Bolsa-Alimentação e Auxílio-Gás. O valor máximo gasto nos programas assistenciais em 2002 foi registrado no município de Salvador, na Bahia. Os gastos considerados em 2002 foram deflacionados utilizando o IPCA - Índice de Preços ao Consumidor Amplo. Tais dados se encontram, assim, a preços constantes de 2006. Em 2006, além dos três programas assistenciais existentes em 2002, também existiam o Cartão-Alimentação e o Bolsa-Família. No município de São Paulo foram gastos mais de 130 milhões de reais enquanto que no município de Águas de São Pedro foram gastos aproximadamente 4.9 mil reais.

O município de Jordão, no Estado do Acre, apresentou a maior taxa de analfabetismo, e a menor taxa de analfabetismo foi registrada no município de São João do Oeste, em Santa Catarina. Nesse último município em 2006 o ex-presidente Lula teve aproximadamente 40% de votos a menos do que obteve em 2002. No município com maior taxa de analfabetismo, a votação do ex-presidente em 2006 foi aproximadamente 13% maior do que em 2002.

Dos 4864 municípios considerados nesse estudo, em 368 municípios o prefeito era filiado ao Partido dos Trabalhadores em 2006. Em termos percentuais, em quase 8% dos municípios estudados o prefeito era do mesmo partido do ex-presidente Lula. Apenas os Estados do Acre, Piauí e Mato Grosso do Sul eram governados pelo PT em 2006.

Em Campos Verdes (Goiás) foi registrada a maior diferença entre os gastos per capita de 2006 e 2002. Essa diferença foi de aproximadamente R$378. A menor diferença foi menor do que R$ 2. Vale destacar que no município de Campos Verdes, em 2006, o ex-presidente Lula obteve 28% de votos a mais do que em 2002, enquanto que no município em a diferença entre os gastos per capita foi menor, o ex-presidente Lula teve em 2006 12% de votos a menos do que em 2002.

Em relação ao crescimento econômico, podemos destacar que o município de Rubiácea em São Paulo apresentou o menor crescimento econômico, e o maior ocorreu em Canaã dos Carajás no Pará.

Um dos nossos interesses neste capítulo reside em quantificar o impacto dos gastos em programas assistenciais e do crescimento da economia sobre o resultado da eleição presidencial de 2006. Vale salientar que em 2006 foram gastos aproximadamente 6.535 bilhões de reais a mais do que em 2002 em programas assistenciais.

4.4 Especificação do modelo

Nesta seção apresentamos uma modelagem empírica relacionada à eleição presidencial no Brasil. O nosso interesse é explicar a diferença entre as proporções de votos válidos do ex-presidente Lula nos segundos turnos nas eleições de 2006 e 2002. Nós utilizamos o modelo de regressão beta (Ferrari & Cribari–Neto, 2004), que é apropriado para dados dessa natureza. O

procedimento computacional foi desenvolvido utilizando o software estatístico R.

Na seleção das covariáveis utilizadas para explicar a diferença entre as proporções de votos válidos do ex-presidente Lula nos segundos turnos nas eleições de 2006 e 2002 utilizamos o critério de seleção de modelo AIC (*Akaike's information criterion*), que foi proposto por Akaike (1974).

Inicialmente, ao usar o modelo de regressão beta, nosso interesse reside em determinar se a dispersão é fixa, ou seja, se há ou não estrutura de regressão para o parâmetro de dispersão. Para tanto, empregamos o teste score (Espinheira, 2007) e rejeitamos a hipótese nula de dispersão fixa ao nível nominal de 1% (p-valor < 0.01). Sendo assim, além de modelar a média, há necessidade de modelar a dispersão.

4.4.1 Diferença entre as proporções de votos válidos do ex-presidente Lula

A variável de interesse, Δy, é a diferença entre as proporções de votos válidos do ex-presidente Lula nos segundos turnos das eleições de 2006 e 2002. As covariáveis selecionadas foram $\Delta GASTO$, $RENDA$, GOV, $PREF$, DU, URB, $PENT$, $ANALF$, $CRESC$ e $\Delta GASTO^2$. A covariável $\Delta GASTO$ dada pela soma dos gastos per capita com programas assistenciais e com Bolsa-Família em 2006 menos a soma dos gastos per capita com programas assistenciais em 2002.

Da forma como definimos a variável de interesse, não podemos garantir um dos pressupostos do modelo de regressão beta, ou seja, que Δy está restrita ao intervalo $(0,1)$. Para contornar tal situação, fizemos uma transformação na variável resposta. A transformação foi $\Delta y^{\ddagger} = (\Delta y - a)/(b-a)$, com $a = -1$ e $b = 1$. Esta padronização nos garante que a variável resposta está restrita ao intervalo $(0,1)$. A variável resposta neste caso apresenta valor mínimo igual a 0.279, valor máximo igual a 0.883 e mediana igual a 0.539. Na Figura 4.1 estão apresentados o histograma e o *box-plot* da variável dependente $\Delta y^{\ddagger}$. É possível notar que a distribuição dessa variável é assimétrica, portanto, é necessário considerar um modelo adequado, i.e., um modelo que capture a assimetria. Assim sendo, utilizamos o modelo de regressão beta.

O modelo selecionado, após exaustiva seleção de modelos, foi

$$\begin{aligned}
\text{logit}(\mu_t^{\ddagger}) &= \beta_0 + \beta_1 \Delta GASTO_t + \beta_2 RENDA_t + \beta_3 GOV_t + \beta_4 PREF_t + \beta_5 DU_t \\
&\quad + \beta_6 URB_t + \beta_7 PENT_t + \beta_8 ANALF_t + \beta_9 CRESC_t + \beta_{10} \Delta GASTO_t^2, \\
\text{logit}(\sigma_t) &= \gamma_0 + \gamma_1 RENDA_t + \gamma_2 DU_t + \gamma_3 URB_t + \gamma_4 ANALF_t,
\end{aligned}$$

$t = 1, \ldots, 4864$.

Através da análise dos coeficientes estimados para o modelo selecionado (ver Tabela 4.3) é possível verificar que a covariável $\Delta GASTO$ influencia positivamente a diferença entre as proporções de votos de Lula em 2006 e 2002. As covariáveis DU, URB, $PENT$, $ANALF$ e $CRESC$ também exercem efeito positivo sobre a diferença das proporções de votos, ao passo que, as covariáveis $RENDA$, GOV e $PREF$ exercem efeito negativo. Os municípios mais urbanizados tendem a apresentar maiores diferenças entre as proporções de votos em 2006 e 2002. Considerando a estrutura de regressão para o parâmetro de dispersão, temos que à medida que a covariável $RENDA$ diminui a dispersão também diminui, ou seja, os municípios

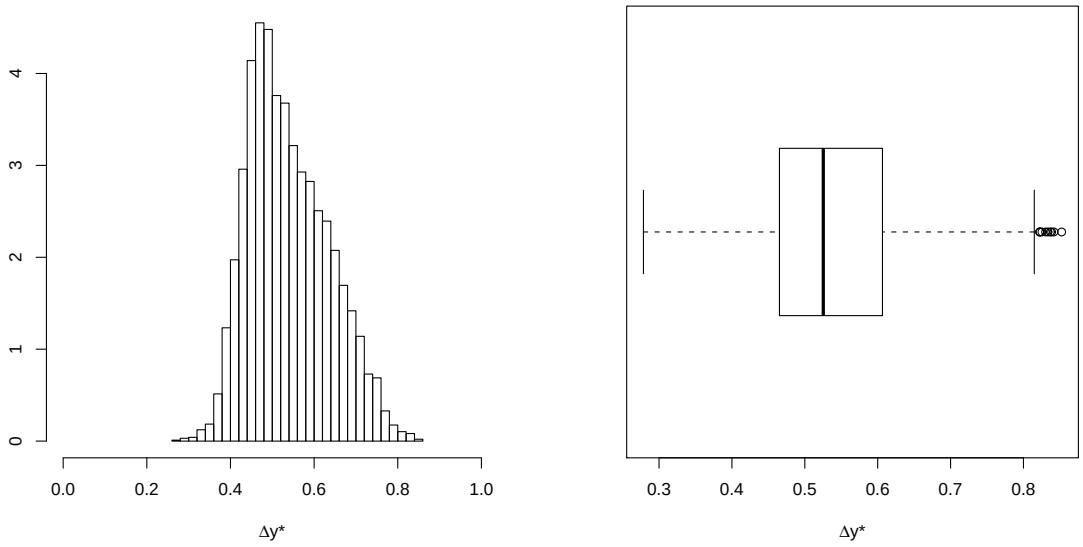

Figura 4.1 Histograma e gráfico *box-plot* da diferença entre as proporções de votos válidos do ex-presidente Lula nas eleições de 2006 e 2002.

que apresentam menores rendimentos tendem a apresentar resposta menos dispersa. Medimos a intensidade da não-constância da dispersão dos dados através da seguinte quantidade: $\hat{\lambda} = \max(\hat{\sigma}_t)/\min(\hat{\sigma}_t)$. Para o modelo de regressão beta considerado temos $\hat{\lambda} = 8.24$. Correspondentemente, a razão entre a maior e a menor precisão estimadas é 70.28. O pseudo R^2 obtido foi 0.67.

Com o intuito de verificar possíveis afastamentos das suposições feitas para o modelo, a Figura 4.2 apresenta os gráficos dos resíduos ponderados padronizados 2 versus os índices das observações e também o gráfico da probabilidade normal com envelopes simulados. O modelo de regressão parece estar bem ajustado, dado que apenas 31 resíduos se encontram fora do intervalo $(-3,3)$; isto corresponde a menos de 1% do número total de municípios. Adicionalmente, notamos que uma vez que os resíduos, em geral, permanecem dentro das bandas de confiança dos envelopes simulados, não há indícios de afastamento da suposição de que o modelo de regressão beta fornece uma boa representação para os dados.

Com o objetivo de complementar a análise de resíduos realizada anteriormente, construímos os gráficos das distâncias de Cook versus os valores preditos e da alavancagem generalizada versus valores preditos, que estão apresentados na Figura 4.3. Notamos que no gráfico das distâncias de Cook e da alavancagem generalizada a observação 58, que corresponde ao município de Águas de São Pedro, encontra-se destacada das demais. Vale destacar que nesse município foi registrado o menor patamar de gastos com programas assistenciais e com Bolsa-Família em 2006. Além disso, o percentual de votos do ex-presidente Lula em 2006 foi aproximadamente

Tabela 4.3 Estimativas pontuais e *p*-valores.

	$\hat{\beta}$	*p*-valor	$\hat{\gamma}$	*p*-valor
INTERCEPTO	-5.084×10^{-1}	$<1\times10^{-15}$	-1.762	$<1\times10^{-15}$
$\Delta GASTO$	4.003×10^{-3}	$<1\times10^{-4}$	—	—
$RENDA$	-1.819×10^{-4}	$<1\times10^{-7}$	-0.002	$<1\times10^{-14}$
GOV	-1.221×10^{-1}	$<4\times10^{-14}$	—	—
$PREF$	-4.053×10^{-2}	$<1\times10^{-5}$	—	—
DU	2.797×10^{-2}	$<1\times10^{-3}$	0.107	$<7\times10^{-3}$
URB	1.181×10^{-1}	$<1\times10^{-14}$	-0.410	$<5\times10^{-12}$
$PENT$	2.359×10^{-1}	$<2\times10^{-5}$	—	—
$ANALF$	1.896	$<1\times10^{-15}$	0.665	$<1\times10^{-6}$
$CRESC$	2.271×10^{-2}	$<8\times10^{-4}$	—	—
$\Delta GASTO^2$	-7.667×10^{-6}	$<4\times10^{-6}$	—	—

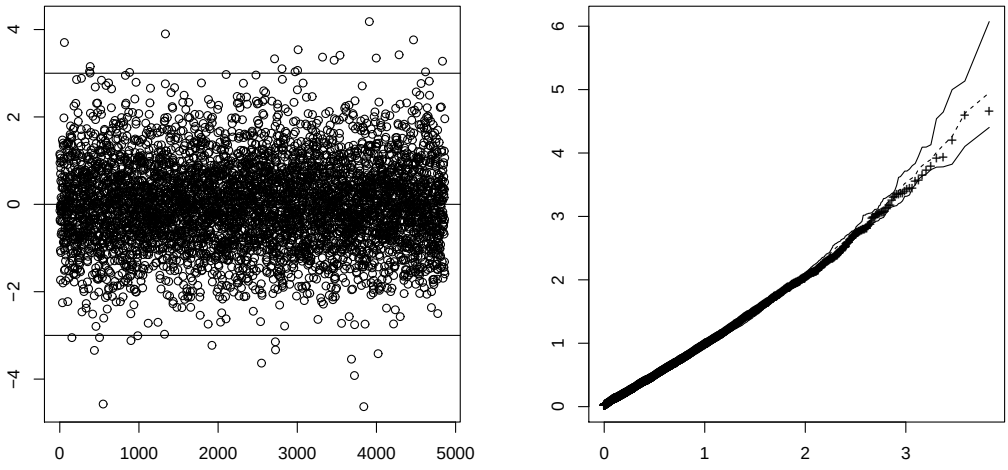

Figura 4.2 Gráfico dos resíduos ponderados padronizados 2 versus os índices das observações e gráfico da probabilidade normal com envelopes simulados.

16% menor do que em 2002. Excluímos individualmente a observação 58, contudo as variações percentuais nas estimativas dos parâmetros foram pequenas.

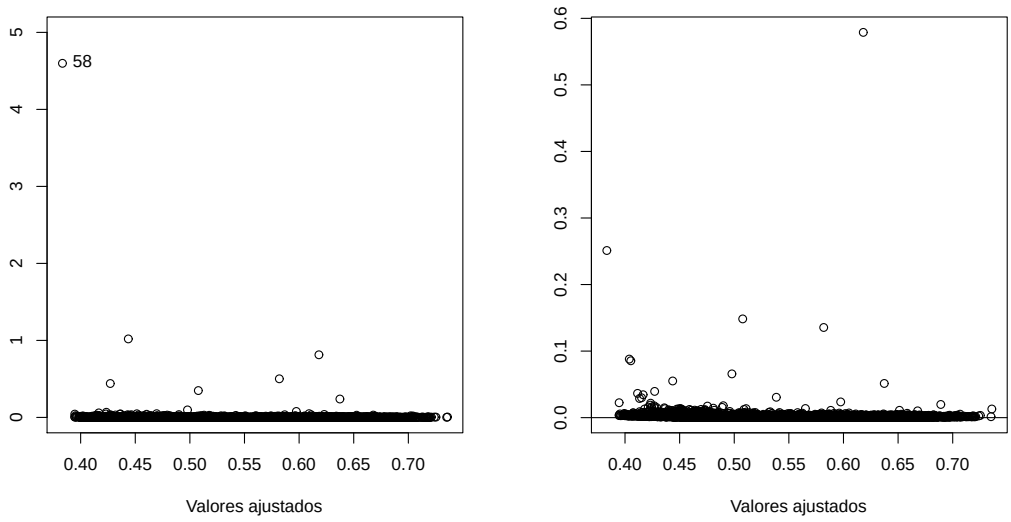

Figura 4.3 Gráficos das distâncias de Cook e da alavancagem generalizada.

Um dos objetivos deste capítulo é quantificar o impacto dos gastos em programas assistenciais e do crescimento da economia sobre o resultado da eleição presidencial de 2006. Vale salientar que em 2006 foram gastos aproximadamente 6.535 bilhões de reais a mais do que em 2002 em programas assistenciais.

Para obter o impacto da diferença entre gastos com programas assistenciais nos anos de 2006 e 2002, calculamos os valores ajustados para cada $t = 1, \ldots, n$, considerando $\Delta GASTO = 0$, $\Delta GASTO^2 = 0$ e as outras covariáveis iguais aos seus valores em cada observação. Assim obtemos $\widehat{\Delta\mu_t^{\ddagger}}$ condicional a $\Delta GASTO = 0$ em todos os municípios. Como foi citado anteriormente, fizemos uma transformação na variável resposta de tal forma que $\Delta y^{\ddagger} \in (0,1)$, assim sendo, a relação entre $\Delta y^{\ddagger}$ e Δy é dada por $\Delta y^{\ddagger} = \frac{\Delta y+1}{2}$, ou ainda $\Delta y = 2\Delta y^{\ddagger} - 1$. Sabemos que $\mu^{\ddagger} = \mathbb{E}(\Delta y^{\ddagger}) = \frac{1}{2}\mathbb{E}(\Delta y) + \frac{1}{2}$. O nosso interesse é obter $\widehat{\Delta\mu_t}$, em que $\widehat{\Delta\mu_t} = 2 \times \widehat{\Delta\mu_t^{\ddagger}} - 1$. Depois de obter $\widehat{\Delta\mu_t}$, calculamos o somatório dos produtos entre $\widehat{\Delta\mu_t}$ e a correspondente população votante. O resultado dessa soma nos fornece a quantidade de votos (em número absolutos) predita pelo modelo do ex-presidente Lula condicional aos gastos sociais de 2006 mantidos nos níveis de 2002. Os nossos resultados revelam que se os gastos em programas assistenciais fossem mantidos nos níveis de 2002, haveria uma redução na votação do ex-presidente de 7.116 milhões de votos. De forma análoga calculamos o impacto do crescimento da economia sobre a votação do ex-presidente Lula, mas nesse caso obtemos $\widehat{\Delta\mu_t^{\ddagger}}$ condicional a $CRESC = 0$

em todos os municípios, ou seja, assumimos que não há diferença entre os PIB's per capita de 2006 e 2002. Os nossos resultados revelaram que se não houvesse crescimento econômico, a votação do ex-presidente Lula teria uma redução de aproximadamente 2 milhões de votos. Portanto, concluímos que o aumento dos gastos com programas assistenciais contribuiu mais para votação do ex-presidente Lula do que o crescimento da economia.

Adicionalmente, consideramos o modelo de regressão beta sem modelar a dispersão e substituímos o estimador usual da matriz de covariâncias pelo estimador sanduíche S_2 proposto nesta tese (Capítulo 3). Ao nível nominal de 5%, todas as covariáveis são significativas, como pode ser visto na Tabela 4.4. Considerando o modelo de regressão beta com dispersão fixa, concluímos que, se os gastos com programas assistenciais fossem mantidos nos níveis de 2002, haveria uma redução na votação do ex-presidente Lula de 7.516 milhões de votos, enquanto que se não houvesse o crescimento econômico, a votação teria uma redução de 2.974 milhões de votos. Note que esses resultados são condizentes com aqueles obtidos a partir do modelo de regressão beta em que se modela a dispersão, especialmente no que pertine ao impacto do Programa Bolsa-Família.

Tabela 4.4 Estimativas pontuais e *p*-valores obtidos de testes quasi-*z* baseados no estimador sanduíche S_2.

	$\hat{\beta}$	*p*-valor	$\hat{\gamma}$	*p*-valor
INTERCEPTO	-3.445×10^{-1}	$<1\times10^{-15}$	-2.081	$<1\times10^{-15}$
$\Delta GASTO$	3.416×10^{-3}	$<1\times10^{-15}$	—	—
$RENDA$	-6.090×10^{-4}	$<3\times10^{-16}$	—	—
GOV	-1.347×10^{-1}	$<1\times10^{-13}$	—	—
$PREF$	-6.912×10^{-2}	$<3\times10^{-8}$	—	—
DU	3.964×10^{-2}	$<5\times10^{-5}$	—	—
URB	1.056×10^{-1}	$<3\times10^{-9}$	—	—
$PENT$	1.708×10^{-1}	$<6\times10^{-3}$	—	—
$ANALF$	1.698	$<1\times10^{-15}$	—	—
$CRESC$	1.576×10^{-2}	$<3\times10^{-2}$	—	—
$\Delta GASTO^2$	-5.645×10^{-6}	$<7\times10^{-4}$	—	—

As Tabelas 4.5 e 4.7 apresentam os impactos estimados de $\Delta GASTO$ e $CRESC$ sobre a diferença entre as votações do ex-presidente Lula em 2006 e 2002 e seus respectivos erros-padrão obtidos via bootstrap. Para mensurar este impacto, consideramos algumas situações diferentes, são elas: a covariável $RENDA$ é fixada no primeiro, segundo e terceiro quartis, enquanto que as outras covariáveis são fixadas na mediana, exceto as covariáveis $\Delta GASTO$, $\Delta GASTO^2$, GOV, $PREF$ e DU. Consideramos o método bootstrap paramétrico para cada obtenção do erro-padrão associado a cada impacto. Para este cálculo utilizamos 1000 réplicas de bootstrap. Resumimos a seguir o procedimento bootstrap adotado.

1. Para cada $t = 1,\ldots,n$, geramos $y_t^{\dagger}$ da densidade beta com parâmetros $\hat{\mu}_t$ e $\hat{\sigma}_t$.

2. Estimamos a regressão beta usando como resposta $y^\dagger = (y_1^\dagger, \ldots, y_n^\dagger)^\top$ conjuntamente com os regressores originais.

3. Considerando as estimativas obtidas no item anterior, calculamos o impacto da covariável $\Delta GASTO$, ou seja, estimamos

$$\mathbb{E}(\Delta y_{t,b}) = \mathbb{E}(2\Delta y_{t,b}^\dagger - 1) = 2\mathbb{E}(\Delta y_{t,b}^\dagger) - 1.$$

4. Repetimos os passos (1) a (3) 1000 vezes.

5. Considerando os 1000 impactos estimados, calculamos a raiz quadrada de sua variância, ou seja, o erro-padrão.

(O procedimento baseado no regresssor $CRESC$ é análogo.)

As covariáveis $\Delta GASTO$ e $\Delta GASTO^2$ foram fixadas em zero, e as três covariáveis *dummies* foram fixadas em zero e um, alternadamente. Várias conclusões podem ser extraídas da Tabela 4.5. Se todos os estados e municípios não fossem governados pelo PT, e se todos fossem localizados em áreas rurais (GOV, $PREF$ e DU assumissem valor zero) e fixando a renda per capita no primeiro quartil (R$ 28.38), encontramos que se os gastos assistenciais de 2006 fossem mantidos a níveis de 2002, haveria uma redução de aproximadamente 4.6% na votação do ex-presidente Lula. Considerando esse mesmo cenário, mas tomando todos os municípios como sendo urbanos ($DU = 1$), essa redução de votos do ex-presidente teria sido em torno de 3%. Os erros-padrão desses dois impactos são de apenas 0.007 e 0.009, respectivamente. Em contrapartida, se em todos os municípios os prefeitos fossem filiados ao PT e os respectivos estados fossem governados pelo PT, e se todos os municípios fossem rurais, a redução de votos do ex-presidente seria de aproximadamente 12%. Além disso, podemos destacar que em todos os casos estudados a redução de votos do ex-presidente teria sido maior na zona rural do que na urbana se os gastos assistenciais fossem mantidos a preços constantes de 2002. Em todos os casos considerados, podemos notar que o erro-padrão obtido através do método bootstrap é pequeno. A Tabela 4.6 apresenta intervalos de confiança bootstrap ao nível de 95% de confiança para os impactos. Como podemos notar, o valor zero não pertence a qualquer dos intervalos de confiança apresentados, sendo assim os impactos obtidos são estatisticamente diferentes de zero ao nível nominal de 5%, ou seja, não são estatisticamente nulos.

Considerando as mesmas situações apresentadas anteriormente para calcular o impacto da covariável $\Delta GASTO$ sobre a eleição do ex-presidente Lula, de forma análoga calculamos o impacto da covariável $CRESC$. Portanto, a covariável $CRESC$ foi fixada em zero e as três covariáveis *dummies* foram fixadas em zero e um, alternadamente. Várias conclusões podem ser extraídas da Tabela 4.7. Se todos os estados e municípios fossem governados pelo PT, e se todos fossem localizados em áreas rurais e fixando a renda per capita no terceiro quartil (R$ 235.00), concluímos que se o crescimento da economia tivesse sido nulo, haveria uma redução de quase 6% na votação do ex-presidente Lula. Considerando o mesmo cenário, mas tomando todos municípios como sendo urbanos, essa redução seria um pouco menor (4.2%). Por outro lado, se todos os estados e municípios não fossem governados pelo PT, encontramos que sem o crescimento da economia, haveria um acréscimo na votação do ex-presidente Lula independente de ser na zona rural ou urbana. Por exemplo, se os prefeitos de todos os municípios não

Tabela 4.5 Impactos estimados da diferença entre gastos nos programas assistenciais em 2006 e 2002 sobre a resposta média e respectivos erros-padrão obtidos via bootstrap.

	RENDA=$Q_{1/4}$	RENDA=$Q_{2/4}$	RENDA=$Q_{3/4}$
$GOV=0, PREF=0, DU=0$	−0.046 (0.007)	−0.053 (0.006)	−0.059 (0.005)
$GOV=1, PREF=1, DU=0$	−0.126 (0.011)	−0.133 (0.010)	−0.139 (0.010)
$GOV=1, PREF=0, DU=0$	−0.106 (0.010)	−0.113 (0.009)	−0.120 (0.009)
$GOV=0, PREF=1, DU=0$	−0.066 (0.008)	−0.073 (0.008)	−0.079 (0.007)
$GOV=0, PREF=0, DU=1$	−0.032 (0.009)	−0.039(0.008)	−0.045 (0.007)
$GOV=1, PREF=1, DU=1$	−0.113 (0.012)	−0.119 (0.011)	−0.126 (0.011)
$GOV=1, PREF=0, DU=1$	−0.093 (0.011)	−0.100 (0.011)	−0.106 (0.010)
$GOV=0, PREF=1, DU=1$	−0.052 (0.010)	−0.059 (0.009)	−0.065 (0.008)

Tabela 4.6 Intervalo de confiança bootstrap ao nível de 95% de confiança para o impactos da diferença entre gastos nos programas assistenciais em 2006 e 2002 sobre a resposta média.

	RENDA=$Q_{1/4}$	RENDA=$Q_{2/4}$	RENDA=$Q_{3/4}$
$GOV=0, PREF=0, DU=0$	(−0.058; −0.032)	(−0.064; −0.040)	(−0.069; −0.048)
$GOV=1, PREF=1, DU=0$	(−0.147; −0.106)	(−0.153; −0.113)	(−0.159; −0.119)
$GOV=1, PREF=0, DU=0$	(−0.126; −0.087)	(−0.131; −0.094)	(−0.137; −0.101)
$GOV=0, PREF=1, DU=0$	(−0.082; −0.050)	(−0.088; −0.057)	(−0.094; −0.065)
$GOV=0, PREF=0, DU=1$	(−0.048; −0.015)	(−0.053; −0.023)	(−0.058; −0.031)
$GOV=1, PREF=1, DU=1$	(−0.138; −0.089)	(−0.143; −0.097)	(−0.148; −0.105)
$GOV=1, PREF=0, DU=1$	(−0.114; −0.069)	(−0.120; −0.077)	(−0.126; −0.085)
$GOV=0, PREF=1, DU=1$	(−0.071; −0.034)	(−0.077; −0.042)	(−0.082; −0.049)

fossem filiados ao PT ($PREF = 0$), na zona rural ($DU = 0$) e fixando a renda per capita no primeiro quartil, concluímos que sem o crescimento da economia, haveria um acréscimo de aproximadamente 3.8% na votação do ex-presidente Lula. Considerando esse mesmo cenário, todavia tomando todos os municípios como sendo urbanos ($DU = 1$), o acréscimo de votos do ex-presidente seria aproximadamente 5%. Os erros-padrão desses dois impactos apresentados são de apenas 0.003 e 0.005, respectivamente. A Tabela 4.8 apresenta intervalos de confiança bootstrap ao nível de 95% de confiança para os impactos. Como podemos notar em 4 dos 28 intervalos de confiança apresentados, o zero pertence ao intervalo de confiança, sendo assim os impactos obtidos são estatisticamente nulos ao nível nominal de 5%. Nos demais intervalos apresentados, ou seja, nos outros 24 intervalos, o valor zero não pertence ao intervalo, portanto, nessas situações o impacto é estatisticamente diferente de zero.

Tabela 4.7 Impactos estimados da diferença do PIB per capita em 2006 e 2002 sobre a resposta média e respectivos erros-padrão obtidos via bootstrap.

	RENDA=$Q_{1/4}$	RENDA=$Q_{2/4}$	RENDA=$Q_{3/4}$
$GOV = 0, PREF = 0, DU = 0$	0.038 (0.003)	0.031 (0.003)	0.025 (0.003)
$GOV = 1, PREF = 1, DU = 0$	−0.025 (0.009)	−0.050 (0.009)	−0.056 (0.009)
$GOV = 1, PREF = 0, DU = 0$	−0.023 (0.008)	−0.029 (0.008)	−0.036 (0.008)
$GOV = 0, PREF = 1, DU = 0$	0.018 (0.006)	0.011 (0.005)	0.005 (0.005)
$GOV = 0, PREF = 0, DU = 1$	0.052 (0.005)	0.046(0.004)	0.039 (0.042)
$GOV = 1, PREF = 1, DU = 1$	−0.037 (0.010)	−0.036 (0.010)	−0.042 (0.095)
$GOV = 1, PREF = 0, DU = 1$	−0.009 (0.009)	−0.006 (0.009)	−0.022 (0.009)
$GOV = 0, PREF = 1, DU = 1$	0.032 (0.007)	0.025 (0.006)	0.019 (0.0062)

Tabela 4.8 Intervalos de confiança bootstrap ao nível de 95% de confiança para o impacto da diferença do PIB per capita em 2006 e 2002 sobre a resposta média.

	RENDA=$Q_{1/4}$	RENDA=$Q_{2/4}$	RENDA=$Q_{3/4}$
$GOV = 0, PREF = 0, DU = 0$	(0.032; 0.045)	(0.027; 0.037)	(0.020; 0.031)
$GOV = 1, PREF = 1, DU = 0$	(−0.060; −0.024)	(−0.066; −0.031)	(−0.073; −0.037)
$GOV = 1, PREF = 0, DU = 0$	(−0.038; −0.005)	(−0.045; −0.013)	(−0.052; −0.020)
$GOV = 0, PREF = 1, DU = 0$	(0.007; 0.030)	(0.009; 0.022)	(−0.006; 0.016)
$GOV = 0, PREF = 0, DU = 1$	(0.048; 0.068)	(0.043; 0.060)	(0.037; 0.054)
$GOV = 1, PREF = 1, DU = 1$	(−0.047; −0.008)	(−0.054; −0.015)	(−0.060; −0.021)
$GOV = 1, PREF = 0, DU = 1$	(−0.025; 0.011)	(−0.032; 0.003)	(−0.038; 0.004)
$GOV = 0, PREF = 1, DU = 1$	(0.020; 0.046)	(0.013; 0.038)	(0.007; 0.031)

4.5 Conclusões

Neste capítulo consideramos uma estratégia de modelagem para avaliar o efeito do aumento nos gastos com programas assistenciais e do crescimento da economia sobre a reeleição do presidente Lula em 2006. Nesse enfoque, utilizamos o modelo de regressão beta, dado que a variável resposta é uma proporção. Modelamos a diferença entre as proporções de votos válidos do ex-presidente Lula nos segundos turnos das eleições de 2006 e 2002. Durante o primeiro mandato houve uma queda de popularidade do ex-presidente, mas no início de 2006 sua popularidade voltou a subir e o presidente foi reeleito no segundo turno da eleição de 2006. Contudo, estudos mostram que o perfil dos eleitores do presidente reeleito mudou. Na sua primeira vitória, Lula apresentou melhor desempenho nas regiões mais urbanizadas e mais desenvolvidas enquanto em 2006 o voto migrou para lugares mais atrasados e menos urbanizados. No segundo turno da eleição de 2006, Lula disputou com Geraldo Alckmin, candidato do PSDB, que obteve maior votação em apenas sete estados.

Em 2006, foram gastos pouco mais do que 7.065 bilhões de reais em programas assistenciais remanescentes e no Bolsa-Família enquanto que em 2002 gastou-se um pouco mais de 530 milhões de reais. Além disso, em 2002 só existiam três programas assistenciais: Bolsa-Escola, Auxílio-Gás e Cartão-Alimentação. Assim sendo, calculamos o impacto da diferença entre os gastos totais em programas assistenciais em 2006 e 2002 sobre a resposta média, que neste caso foi a diferença entre as proporções de votos em 2006 e 2002 do presidente Lula. Considerando o cenário em que se em todos os municípios os prefeitos fossem filiados ao PT e os respectivos estados fossem governados pelo PT, e se todos os municípios fossem rurais, encontramos que se os gastos nos programas assistenciais fossem mantidos a níveis de 2002 a redução de votos do presidente seria de aproximadamente 12%. Em relação ao impacto do crescimento da economia, concluímos que sem crescimento econômico entre 2002 e 2006, o presidente Lula teria uma redução de votos apenas nas situações em que o governador era filiado ao PT. Vale ressaltar que o impacto do crescimento da economia é menor do que o impacto do aumento dos gastos assistenciais. Considerando o Brasil como um todo, nossos resultados revelam que se os gastos com programas assistenciais fossem mantidos nos níveis de 2002, a votação do presidente Lula teria uma redução de aproximadamente 7.116 milhões de votos, enquanto que, sem crescimento econômico essa redução seria de apenas 2 milhões de votos. Vale ressaltar que mesmo com tais reduções de votos, o ex-presidente Lula seria reeleito, dado que a diferença entre os votos de Lula e de Geraldo Alckmin foi de 22% que corresponde a aproximadamente 21 milhões de votos.

Capítulo 5

Ateísmo e inteligência

5.1 Introdução

A determinação do percentual de uma sociedade que acredita ou não em Deus (ou deuses) envolve vários obstáculos. Primeiro, poucas pessoas respondem a este tipo de pergunta, e se a quantidade de respostas obtidas for escassa, não é possível fazer generalizações para toda a sociedade. Segundo, se a amostragem adotada para a coleta de tal informação não for aleatória, assim como no caso anterior, não se pode proceder a generalizações. Terceiro, nas sociedades em que a religião é imposta pelo governo, os indivíduos relutam em admitir que não têm fé religiosa. Mesmo em sociedades democráticas, muitos afirmam ser religioso simplesmente porque esta resposta é socialmente desejável ou culturalmente apropriada (Zuckerman, 2007).

A designação de ateu é estigmatizada em muitas sociedades. Muitas pessoas que não acreditam em Deus, evitam a auto-designação de 'ateu'. Greeley (2003) constatou que 41% dos noruegueses, 48% dos franceses e 54% dos checos alegaram não acreditar em Deus, mas apenas 10%, 19% e 20%, respectivamente, dos entrevistados se identifica como ateu.

Os ateus estão mais concentrados em países economicamente desenvolvidos. Em países subdesenvolvidos, a preponderância de ateus é muito baixa. Por exemplo, na maioria dos países da África o percentual de ateus não ultrapassa 1% (Zuckerman, 2007), ao contrário de países como a Suécia, Dinamarca, França e Alemanha, que apresentam elevados percentuais de ateus: 64%, 48%, 44% e 42%, respectivamente.

Howells (1928) mostrou que há correlação negativa entre inteligência e crença religiosa entre estudantes universitários usando diferentes medidas de crença religiosa. Argyle (1958) concluiu que alunos inteligentes são menos propensos a aceitar crenças religiosas. Algumas evidências adicionais apontam ainda para a existência de uma relação negativa entre inteligência e crença religiosa. Dentre elas podemos destacar: menor preponderância de crenças religiosas entre as elites relativamente à população geral, declínio da crença religiosa com o aumento da idade entre crianças e adolescentes e, por fim, declínio da intensidade de crença religiosa e aumento da inteligência da população durante o século XX. Há evidência de declínio da crença religiosa nos últimos 150 anos ou mais; ao mesmo tempo, a inteligência média da população tem aumentado. O declínio da crença religiosa tem sido demostrado pelas estatísticas de frequência a igrejas e outros locais de liturgia religiosa e também pelo declínio na crença em Deus, que vem sendo registrado em pesquisas de opinião. Pesquisas mostraram que 95.5% da população dos Estados Unidos acreditavam em Deus em 1948 (Argyle, 1958), mas em 2004 este percentual caiu para 89.5% (Zuckerman, 2007).

Lynn et al. (2009) afirmam que pessoas com QI's (quociente de inteligência) mais altos são menos propensas a manter crenças religiosas. Os autores notaram que em apenas 23 de

um total de 137 países a porcentagem da população que não acredita em Deus ultrapassa 20% e que tais países são, na maioria, os que apresentam os maiores QI's. Adicionalmente, os autores dividiram os países em dois grupos. Os países cujo QI's médios variam entre 64 e 86 pontos participaram do primeiro grupo. Nesse grupo, a média de ateus é de apenas 1.95%. No segundo grupo foram considerados países em que o QI médio variava entre 87 a 108 pontos. Nesse grupo, cerca de 17% da população não possui crença em Deus. A correlação entre QI e ateísmo nos dois grupos é positiva, mas vale ressaltar que no segundo grupo de países essa correlação é mais de três vezes maior do que a correlação encontrada no primeiro grupo. Os autores argumentam contudo, que há algumas exceções para a conclusão de que o QI mais alto se associa a taxas mais altas de ateísmo.

Um dos objetivos deste capítulo é modelar a proporção de ateus em um amplo conjunto de países. Para isto, consideramos três cenários diferentes. A motivação para que consideremos diferentes cenários reside no fato de que é importante verificar a robustez dos resultados à escolha dos dados utilizados. Ou seja, é importante analisar as conclusões obtidas a partir de diferentes escolhas de conjuntos de dados e países a fim de estabelecer se tais conclusões são alteradas quando há alterações na amostra. Como veremos, isso não ocorre. Ou seja, as conclusões que obtemos são aproximadamente as mesmas independentemente da amostra utilizada. No primeiro cenário, consideramos o quociente de inteligência como variável explicativa e usamos outras variáveis, tais como renda nacional bruta ajustada pela paridade do poder de compra, grau de abertura da economia e predominância de muçulmanos. No segundo cenário, consideramos apenas os países em que o percentual de descrentes é igual ou maior do que 1%. Por fim, no último cenário, consideramos os cinquenta países que apresentam as maiores proporções de ateus. O número de países em cada um dos cenários foi 124, 81 e 50, respectivamente. Em nossa análise, estimamos o impacto exercido pela inteligência sobre a preponderância de ateus nos diferentes países. Os resultados revelam que há forte associação positiva, ou seja, em média maior inteligência está associada a maior descrença religiosa, mesmo quando outros fatores são levados em consideração. Um outro objetivo reside na modelagem da proporção de pessoas que não consideram religião importante para suas vidas cotidianas. Essa análise foi baseada em dados de 84 nações. Vale ressaltar que nesse último estudo e nos três cenários descritos anteriormente a variável de interesse é uma proporção. Sendo assim, faz-se necessário o uso de modelos adequados. Nas modelagens, utilizamos o modelo de regressão beta proposto por Ferrari & Cribari-Neto (2004).

Além desta introdução, o presente capítulo contém mais quatro seções. Na Seção 5.2, descrevemos os dados utilizados. Na Seção 5.3, apresentamos o modelo de regressão beta, bem como algumas medidas de avaliação da qualidade do ajuste. Na Seção 5.4, apresentamos os resultados empíricos de nossa análise. Por fim, a seção 5.5 conclui o capítulo.

5.2 Descrição dos dados

A Tabela 5.1 apresenta uma breve descrição das variáveis utilizadas. As fontes dos dados são Lynn et al. (2009), Banco Mundial e as páginas web: `http://www.qran.org/a/a-world.htm` e `http://www.gallup.com` (informações oriundas de diferentes fontes).

Lynn et al. (2009) apresentam as porcentagens de ateus em 137 países (que agregam mais

Tabela 5.1 Descrição das variáveis utilizadas nas modelagens.

Variável	Definição
$ATEU$	Proporção de ateus em 2004
$RELIG$	Proporção de pessoas que não consideram religião importante para suas vidas cotidianas em 2009
QI	Quociente médio de inteligência da população
$RENDA$	Renda per capita ajustada pela paridade do poder de compra em 2008, em mil dólares
$ABERT$	Quociente entre a soma das importações e exportações e o Produto Interno Bruto em 2008
$MULC$	Variável dummy: 1 se o percentual de muçulmanos é maior do que 50% e 0 caso contrário
$MULCP$	Percentual de muçulmanos
$EXPEC$	Expectativa de vida da população em 2007 (em anos)

de 95% da população mundial) extraídas de Zuckerman (2007). Estes dados foram coletados a partir de diversos artigos publicados, na sua maioria em 2004. Além das porcentagens de ateus, Lynn et al. (2009) apresentam os quocientes médios de inteligência das 137 nações.

Estes quocientes foram calculados por Lynn et al. (2006), que utilizaram os resultados da TIMSS (*Trends in International Mathematics and Science Study*) em 2003, que consiste de testes de amostras representativas de alunos da quarta e oitava séries com idades de 10 e 14 anos, respectivamente. Duas etapas foram consideradas no processo de amostragem. Na primeira etapa, 150 escolas públicas e privadas foram selecionadas em cada país. A segunda etapa teve como objetivo selecionar escolas que possuíam duas ou mais classes de alunos com as idades exigidas. O número médio de alunos testados em cada país foi 4498 para alunos da quarta série e 4777 para alunos da oitava série. Esses alunos foram avaliados em questões de matemática e ciências. No total, havia 152 itens nas provas aplicados a alunos com 10 anos de idade e 189 itens nas provas resolvidas por estudantes de 14 anos de idade. Os testes incluíram as mesmas questões em todos os países. Os dados relacionados ao quociente médio de inteligência foram utilizados em vários estudos, como, e.g., Rushton et al. (2009), Lynn et al. (2007), Kanazawa (2009), entre outros. Para detalhes sobre a coleta e o processamento dos dados de inteligência, ver Lynn et al. (2006).

O nosso interesse reside, inicialmente em modelar as proporções de descrentes nas diferentes nações e, a seguir, as proporções de pessoas que não acreditam que religião é importante para suas vidas cotidianas. Para tal, além da covariável *QI* consideramos outras covariáveis, a saber: renda per capita ajustada pela paridade no poder de compra, grau de abertura da economia, expectativa de vida e preponderância de muçulmanos. As duas primeiras covariáveis correspondem ao ano de 2008 e as demais, aos anos de 2007 e 2004, respectivamente.

Os ateus estão mais concentrados em países economicamente desenvolvidos. Em países subdesenvolvidos, a preponderância de ateus é tipicamente baixa. As razões que contribuem para que as igrejas percam força nos países desenvolvidos podem ser resumidas em termos de forças de mercado. Com o desenvolvimento da ciência, da segurança nas instituições e com famílias menores, há menos medo e insegurança na vida cotidiana das pessoas e, consequentemente, menos espaço para religião. Com o objetivo de levar em consideração o grau de desenvolvimento econômico de cada país, incluímos a covariável *RENDA* para explicar a proporção de ateus. Como se sabe, há uma tendência da diminuição do grau de religiosidade de um país à medida em que ele se torna mais rico, próspero e desenvolvido. A renda interna bruta ajustada pela paridade do poder de compra é construída a partir de uma cesta única internacional de mercadorias e serviços que é periodicamente arbitrada a partir de pesquisas de preços e composição de gastos nos diferentes países analisados pelo Programa de Comparações Internacionais das Nações Unidas (Kilsztajn, 2000). O ajuste pela paridade do poder de compra é importante a fim de neutralizar — ou ao menos minimizar — distorções introduzidas por desequilíbrios de ordem cambial. Uma outra covariável importante é dada pela razão entre a soma das importações e exportações e o PIB (Produto Interno Bruto) de cada país (*ABERT*). Vale ressaltar que *ABERT* foi incluída para levar em consideração o fato de que alguns países que apresentam alto grau de ateísmo são consideravelmente fechados para o resto do mundo e sua população, em consequência, tem pouca exposição a religiões preponderantes em outras localidades geográficas, o que reduz sua capacidade de discernimento e opção religiosa. Tal

efeito é mais intenso em regimes socialistas (e.g., Cuba). Países mais desenvolvidos tendem a apresentar maiores valores para *RENDA* e *ABERT* do que os menos desenvolvidos. Adicionalmente, consideramos também a expectativa de vida. Por fim, consideramos ainda se o país é ou não preponderantemente muçulmano, ou seja, se mais de 50% da população pertence à religião muçulmana. Tal religião tem crescido nos últimos anos e está presente em todos os continentes, mas a maior parte de seguidores do islamismo encontra-se em países árabes do Oriente Médio e em países do norte da África. A religião muçulmana é monoteísta, ou seja, cultua apenas um deus: Alá. Seus seguidores se dividem em dois grupos principais: sunitas e xiitas. Em países classificados como muçulmanos, a proporção de ateus tende a ser pequena. A motivação para levar em conta o efeito possivelmente diferenciado sobre o grau de ateísmo nos países muçulmanos reside na menor separação entre estado e religião prevalente em tais países e no maior grau de literalismo em relação ao texto sagrado que fundamenta o islamismo (Corão). Assim, é possível antever que o comportamento da variável resposta nos países muçulmanos possa ser distinto daquele verificado em países de outras orientações religiosas.

Na Tabela 5.2 apresentamos algumas estatísticas descritivas: mínimo, primeiro quartil ($Q_{1/4}$), mediana, média, terceiro quartil ($Q_{3/4}$) e máximo das variáveis utilizadas.[1] Algumas conclusões podem ser extraídas da Tabela 5.2. Em 75% das nações a proporção de ateus não excede 0.15. A maior proporção de ateus é 0.81 e a menor, 0.005. Considerando a variável *QI*, temos que 50% das nações apresentam quociente de inteligência médio igual ou superior a 87. A menor expectativa de vida é 42 anos e a maior, 83 anos de idade. Espera-se viver, em média, aproximadamente 68 anos. O maior valor entre o quociente da soma de importações e exportações e o PIB (grau de abertura da economia) é 361.60, que é quase quinze vezes maior do que o menor valor. A proporção média de pessoas que não consideram religião importante para suas vidas cotidianas é 0.25. O maior valor registrado da variável *RELIG* é 0.820. As Figuras 5.1, 5.2 e 5.3 apresentam os mapas do QI médio, proporção de ateus e a da renda per capita ajustada pela paridade do poder de compra das 124 nações, respectivamente.

Tabela 5.2 Estatísticas descritivas das variáveis utilizadas.

Variável	mínimo	$Q_{1/4}$	mediana	média	$Q_{3/4}$	máximo
ATEU	0.005	0.005	0.024	0.105	0.150	0.810
QI	64	72.75	87.00	85.60	96.25	108
RENDA	0.30	2.12	7.14	12.59	18.26	58.50
ABERT	24.17	47.25	66.67	74.55	57.86	361.60
EXPEC	42.00	61.00	71.50	68.26	76.00	83.00
MULCP	0.00	0.28	3.00	23.22	38.00	99.00
RELIG	0.0005	0.050	0.150	0.250	0.480	0.820

Lynn et al. (2009) calcularam a correlação entre as variáveis *ATEU* e *QI* considerando dois grupos. No primeiro grupo, estão as nações cujos quocientes de inteligência variam entre 64 e

[1] As estatísticas descritivas das sete primeiras variáveis utilizadas foram calculadas com base em 124 países, enquanto que para última variável (*RELIG*), 84 países foram considerados.

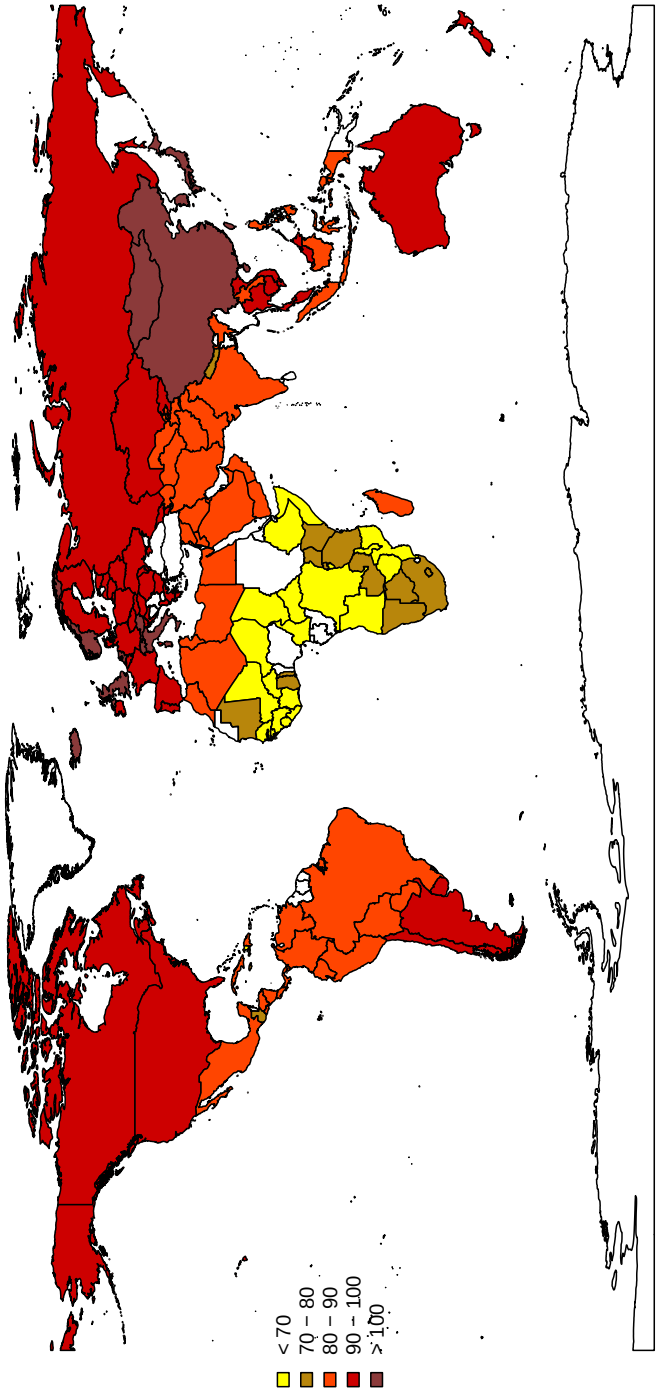

Figura 5.1 Mapa do quociente médio de inteligência nas nações.

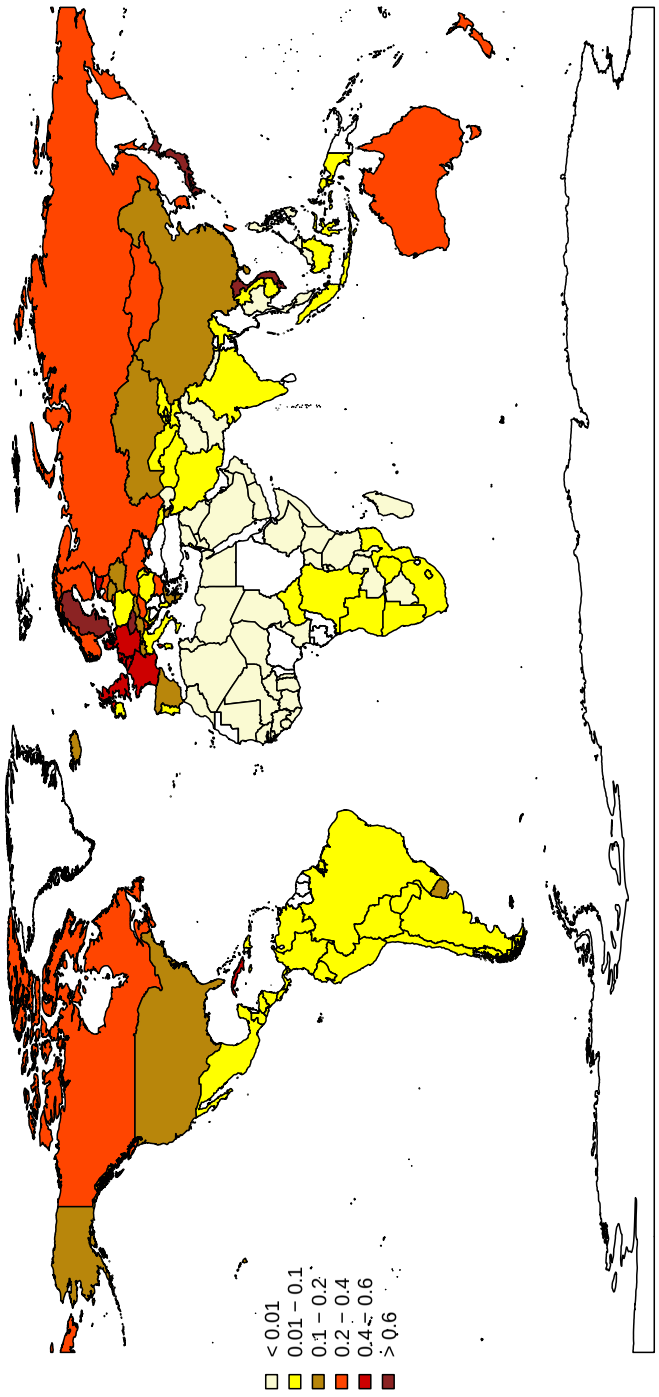

Figura 5.2 Mapa da proporção de ateus nas nações.

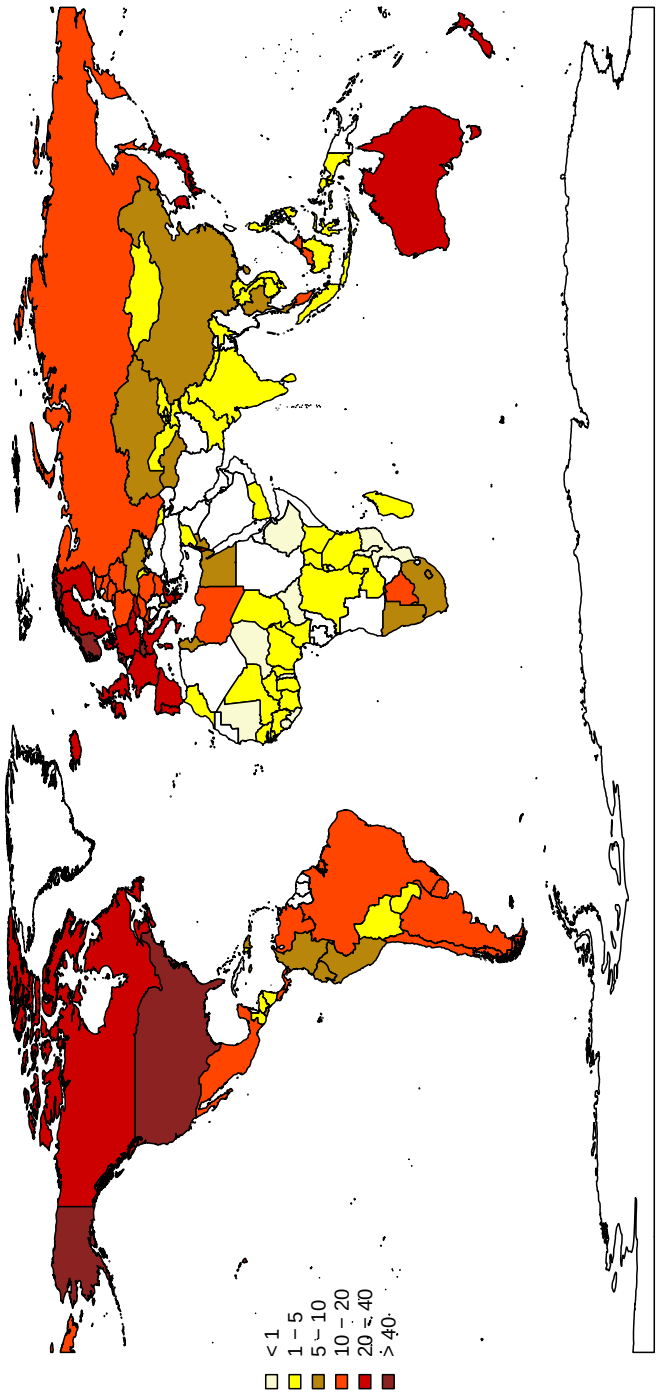

Figura 5.3 Mapa da renda per capita ajustada pela paridade do poder de compra nas nações.

86 pontos, e no segundo grupo, as nações que se situam entre 87 e 108. Os autores observaram que nos países do segundo grupo, isto é, aqueles de maior inteligência média, a correlação entre *QI* e *ATEU* é alta, 0.54, ao passo que no primeiro grupo tal correlação é de apenas 0.16. Adotamos a mesma classificação apresentada por Lynn et al. (2009) e calculamos as correlações entre *ATEU* e as covariáveis descritas anteriormente (ver Tabela 5.3). Adicionalmente, na Tabela 5.4 apresentamos as correlações entre *RELIG* e as covariáveis listadas na Tabela 5.1.

De acordo com a Tabela 5.3, podemos observar que *ATEU* se correlaciona positivamente com todas as covariáveis, exceto com *MULCP*. Notamos que as maiores correlações (absolutas) de *ATEU* são registradas com as covariáveis *QI* e *RENDA*. Também observamos que a maior correlação entre *ATEU* e *QI* está nos países em que o *QI* médio está entre 87 e 108. Por outro lado, a correlação entre *RENDA* e a proporção de ateus é maior nos países em que o *QI* está entre 64 e 86. É importante destacar ainda que a correlação entre a proporção descrente da população e a variável *MULCP* nos países que apresentam maiores valores para o quociente de inteligência é -0.329 enquanto que, considerando o outro grupo, esta correlação é -0.295.

Tabela 5.3 Correlações entre *ATEU* e as covariáveis *QI*, *RENDA*, *ABERT*, *MULCP* e *EXPEC*.

Grupos	*QI*	*RENDA*	*ABERT*	*MULCP*	*EXPEC*
64 – 108	0.609	0.611	0.259	−0.353	0.488
64 – 86	0.119	0.509	0.228	−0.295	0.105
87 – 108	0.529	0.428	0.160	−0.329	0.395

Como mostra a Tabela 5.4, podemos notar que *RELIG* se correlaciona positivamente com todas as covariáveis, exceto *MULCP* e *ABERT* (segundo grupo). Assim como *ATEU*, a variável *RELIG* apresenta as maiores correlações com as covariáveis *QI* e *RENDA*. Notamos ainda que a maior correlação entre *RELIG* e *QI* ocorre nos países em que o *QI* médio está entre 87 e 108. A correlação entre a proporção de pessoas que não acreditam que religião é importante para suas vidas e a variável *EXPEC* é maior nos países que apresentam níveis médios de inteligência mais elevados, enquanto que considerando o outro grupo esta correlação é 0.429.

Tabela 5.4 Correlações entre *RELIG* e as covariáveis *QI*, *RENDA*, *ABERT*, *MULCP* e *EXPEC*.

Grupos	*QI*	*RENDA*	*ABERT*	*MULCP*	*EXPEC*
64 – 108	0.733	0.720	0.165	−0.361	0.582
64 – 86	0.375	0.683	0.063	−0.525	0.429
87 – 108	0.577	0.527	−0.002	−0.252	0.443

O Vietnã, país asiático comunista, apresenta a maior proporção de ateus e é um dos países que apresentam maior proporção de pessoas que não consideram religião importante para suas vidas cotidianas. Outros países estão bem perto dessa proporção (ateus), como Japão, Suécia e República Checa. Vale ressaltar que nenhum desses países citados é classificado como muçulmano, dado que para um país ser muçulmano em nossa definição é necessário que mais de

50% da população pertença à religião islâmica. A correlação entre *ATEU* e *MULCP* é negativa (Tabela 5.3). Dos 124 países estudados, 29 foram considerados muçulmanos. O maior percentual de muçulmanos foi registrado na Argélia, seguido de Marrocos e Tunísia.

Os maiores valores que a variável *QI* assume foram registrados em dois países que pertencem ao continente asiático, são eles: Cingapura, o menor país do Sudoeste Asiático, e Coreia do Sul. Os quocientes médios de inteligência nesses dois países são quase o dobro dos observados nos seis países do continente africano que apresentaram os menores valores do quociente médio de inteligência.

O maior quociente entre a soma de importações e exportações e o PIB, isto é, *ABERT*, também foi registrado em Cingapura, que é o quinto país mais rico do mundo em termos do PIB per capita de acordo com dados de 2007 do Fundo Monetário Internacional. Em 2009, a *Economist Intelligence Unit* classificou Cingapura como a décima cidade mais cara do mundo para se viver e a terceira mais cara da Ásia, depois de Tóquio e Osaka. Cingapura tem uma economia pujante voltada para as exportações. Juntamente com Hong Kong, Coreia do Sul e Taiwan, Cingapura é um dos quatro 'Tigres Asiáticos'.

O país com a maior expectativa de vida é o Japão. Espera-se que, em média, as pessoas vivam até os 82 anos no Japão. Alguns outros países estão bem perto deste índice, como a Suíça, Itália, Austrália, Islândia e França, entre outros. O país com a menor expectativa de vida é Moçambique, na África, onde se espera que uma pessoa viva apenas até os 42 anos. Em outros países africanos, tais como República Centro-Africana, Congo e Zâmbia, a expectativa de vida ao nascer não passa dos 46 anos.

A correlação entre a proporção de ateus e a proporção de pessoas que não consideram religião importante para suas vidas cotidianas é 0.86, ou seja, as duas variáveis em questão são altamente correlacionadas. As nações que apresentam as maiores proporções de pessoas que não valorizam religião são a Suíça, Dinamarca e Estônia. Vale destacar que essas nações apresentam altas proporções de ateus: 0.64, 0.48 e 0.49, respectivamente. Os quocientes médios de inteligência dessas nações são altos. Na Suiça e na Estônia, o *QI* médio é 99, na Dinamarca é 98. Por fim, notamos que apesar dos Estados Unidos apresentarem *QI* médio alto (98), a proporção de pessoas que não consideram religião importante é baixa (0.34) frente ao elevado nível médio de inteligência.

5.3 Modelo de regressão beta

Ferrari & Cribari–Neto (2004) propuseram uma parametrização alternativa para a densidade beta que permite a modelagem da média da resposta através de uma estrutura de regressão e que envolve também um parâmetro de precisão. A função de densidade beta nessa reparametrização tem a forma

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0<y<1, \tag{5.3.1}$$

em que $0<\mu<1$, $\phi>0$ e $\Gamma(\cdot)$ é a função gama. Aqui, $\mathbb{E}(y)=\mu$ e $\text{var}(y)=\frac{V(\mu)}{1+\phi}$, sendo $V(\mu)=\mu(1-\mu)$ a 'função de variância', em que μ é a média da variável resposta e ϕ pode ser interpretado como um parâmetro de precisão.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, em que cada y_t, $t = 1, \ldots, n$, possui densidade (5.3.1), com média μ_t e parâmetro de precisão ϕ_t desconhecidos.[2] O modelo proposto por Ferrari & Cribari–Neto (2004) é obtido assumindo que a média de y_t pode ser escrita como

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t,$$

em que $\beta = (\beta_1, \ldots, \beta_k)^\top$ é um vetor de parâmetros desconhecidos ($\beta \in \mathbb{R}^k$) e $x_{t1}, \ldots, x_{tk}$ são observações de k covariáveis. Por fim, $g(\cdot)$, a função de ligação $g : (0,1) \to \mathbb{R}$, é estritamente monótona e duas vezes diferenciável. Portanto, $\mu_t = g^{-1}(\eta_t)$ e $\mathrm{var}(y_t) = \mu_t(1-\mu_t)/(1+\phi))$. Admitimos, como Simas et al. (2010), que o parâmetro de precisão ϕ_t é dado por

$$h(\phi_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \nu_t,$$

em que $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ é um vetor de parâmetros desconhecidos, $z_{t1}, \ldots, z_{tq}$ são observações de q covariáveis ($k + q < n$), assumidas fixas e conhecidas, e $h(\cdot)$ é uma função estritamente monótona e duas vezes diferenciável que mapeia os reais positivos na reta.

Segue de (5.3.1) que o logaritmo da função de verossimilhança é

$$\ell(\beta, \gamma) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi_t),$$

em que

$$\begin{aligned} \ell_t(\mu_t, \phi_t) &= \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) + (\mu_t\phi_t - 1)\log y_t \\ &+ \{(1-\mu_t)\phi_t - 1\}\log(1-y_t). \end{aligned}$$

Logo, para $i = 1, \ldots, k$, a função escore para β_i é dada por

$$\frac{\partial \ell(\beta,\gamma)}{\partial \beta_i} = \sum_{t=1}^{n} \frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \mu_t}\frac{d\mu_t}{d\eta_t}\frac{\partial \eta_t}{\partial \beta_i},$$

em que $d\mu_t/d\eta_t = 1/g'(\mu_t)$ e

$$\frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \mu_t} = \phi_t(y_t^* - \mu_t^*),$$

com $y_t^* = \log\frac{y_t}{1-y_t}$ e $\mu_t^* = \psi(\mu_t\phi_t) - \psi((1-\mu_t)\phi_t)$, em que $\psi(\cdot)$ é a função digama. Logo,

$$\frac{\partial \ell(\beta,\gamma)}{\partial \beta_i} = \phi_t(y_t^* - \mu_t^*)\frac{1}{g'(\mu_t)}x_{ti}.$$

[2]Note que estamos permitindo que o parâmetro de precisão varie ao longo das observações.

Podemos escrever o vetor $U_\beta(\beta,\gamma)$, a função escore de $\beta = (\beta_1,\ldots,\beta_k)^\top$, em forma matricial como

$$U_\beta(\beta,\phi) = X^\top \Phi T(y^* - \mu^*),$$

em que X é uma matriz $n \times k$ cuja t-ésima linha é $x_t^\top$, $T = \text{diag}\{1/g'(\mu_1),\ldots,1/g'(\mu_n)\}$, $\Phi = \text{diag}\{\phi_1,\ldots,\phi_n\}$, $y^* = (y_1^*,\ldots,y_n^*)^\top$ e $\mu^* = (\mu_1^*,\ldots,\mu_n^*)^\top$.

Consideraremos agora as derivadas do logaritmo da função de verossimilhança em relação aos parâmetros que modelam a precisão, $\gamma_j, j = 1\ldots,q$. Temos que

$$\frac{\partial \ell(\beta,\gamma)}{\partial \gamma_j} = \sum_{t=1}^{n} \frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \phi_t}\frac{d\phi_t}{d\nu_t}\frac{\partial \nu_t}{\partial \gamma_j},$$

em que $d\phi_t/d\nu_t = 1/h'(\phi_t)$, e

$$\frac{\partial \ell_t(\mu_t,\phi_t)}{\partial \phi_t} = \mu_t(y_t^* - \mu_t^*) + \log(1-y_t) - \psi((1-\mu_t)\phi_t) + \psi(\phi_t).$$

Então,

$$\frac{\partial \ell(\mu_t,\phi_t)}{\partial \gamma_j} = \sum_{t=1}^{n} a_t \frac{1}{h'(\sigma_t)} z_{tj},$$

em que $a_t = \mu_t(y_t^* - \mu_t^*) + \log(1-y_t) - \psi((1-\mu_t)\phi_t) + \psi(\phi_t)$.

Podemos escrever o vetor $U_\gamma(\beta,\gamma)$, a função escore de $\gamma = (\gamma_1,\ldots,\gamma_q)^\top$, matricialmente como

$$U_\gamma(\beta,\gamma) = Z^\top H a,$$

em que Z é uma matriz $n \times q$ cuja t-ésima linha é $z_t^\top$, $H = \text{diag}\{1/h'(\phi_1),\ldots,1/h'(\phi_n)\}$ e $a = (a_1,\ldots,a_n)^\top$.

A matriz de informação de Fisher conjunta dos vetores β e γ é formada pelo negativo dos valores esperados das segundas derivadas do logaritmo da função de verossimilhança. Para o modelo de regressão beta na parametrização aqui usada, a matriz da informação de Fisher para o vetor paramétrico $\theta = (\beta^\top,\gamma^\top)^\top$ pode ser escrita como

$$K(\theta) = K(\beta,\gamma) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\gamma} \\ K_{\gamma\beta} & K_{\gamma\gamma} \end{pmatrix},$$

em que $K_{\beta\beta} = X^\top \Phi W X$, sendo $W = \text{diag}\{w_1,\ldots,w_n\}$, com

$$w_t = \phi[\psi'(\mu_t\phi_t) + \psi'((1-\mu_t)\phi_t)]\frac{1}{\{g'(\mu_t)\}^2},$$

$K_{\beta\gamma} = (K_{\gamma\beta})^\top = X^\top C T H Z$, em que $C = \text{diag}\{c_1,\ldots,c_n\}$, com

$$c_t = \phi_t[\psi'(\mu_t\phi_t) - \psi'((1-\mu_t)\phi_t)(1-\mu_t)]$$

e $K\gamma\gamma = Z^\top D Z$. Adicionalmente, $D = \text{diag}\{d_1,\ldots,d_n\}$, com

$$d_t = [\psi'(\mu_t\phi_t)\mu_t^2 + \psi'((1-\mu_t)\phi_t)(1-\mu_t)^2 - \psi'(\phi_t)]\frac{1}{\{h'(\phi_t)\}^2}.$$

Os estimadores de máxima verossimilhança de β e γ são obtidos pela solução do sistema formado por $U_\beta(\beta,\gamma) = 0$ e $U_\gamma(\beta,\gamma) = 0$. Tais estimadores não possuem forma fechada. Assim, eles têm que ser obtidos numericamente pela maximização da função de log-verossimilhança com auxílio de um algoritmo de otimização não-linear.

Usando a expressão padrão para a inversa de matrizes particionadas, temos que a inversa da matriz da informação de Fisher é dada por

$$K(\theta)^{-1} = K(\beta,\gamma)^{-1} = \begin{pmatrix} K_*^{\beta\beta} & K_*^{\beta\gamma} \\ K_*^{\gamma\beta} & K_*^{\gamma\gamma} \end{pmatrix},$$

em que

$$K_*^{\beta\beta} = (X^\top \Phi W X - X^\top C T H Z (Z^\top D Z)^{-1} Z^\top H T C X)^{-1}.$$

Adicionalmente,

$$K_*^{\beta\gamma} = (K_*^{\gamma\beta})^\top = -K_*^{\beta\beta} X^\top C T H Z (Z^\top D Z)^{-1}$$

e

$$K_*^{\gamma\gamma} = (Z^\top D Z)^{-1}\{I_q + (Z^\top H T C X) K_*^{\beta\beta} X^\top C T H Z (Z^\top D Z)^{-1}\}.$$

Sob certas condições de regularidade, temos que, para tamanhos de amostras grandes, a distribuição aproximada conjunta de $\hat\beta$ e $\hat\gamma$ é normal $(k+1)$-variada:

$$\begin{pmatrix} \hat\beta \\ \hat\gamma \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix}, K^{-1}\right),$$

aproximadamente, sendo $\hat\beta$ e $\hat\gamma$ os estimadores de máxima verossimilhança de β e γ, respectivamente.

5.4 Especificação dos modelos

Nesta seção apresentaremos três modelagens empíricas relacionadas à prevalência de ateísmo nos diferentes países. Nas três modelagens o nosso interesse reside em explicar a proporção de descrença. No primeiro modelo consideramos, além da covariável *QI*, outras covariáveis: *RENDA*, *MULC* e *ABERT*. No ajuste do segundo modelo consideramos apenas os países em que o percentual de ateus é igual ou maior do que 1%. Por fim, no último ajuste, consideramos os cinquenta países que apresentam as maiores proporções de descrentes. Adicionalmente, apresentamos uma modelagem em que a variável resposta é a proporção de pessoas que consideram religião importante para suas vidas cotidianas. Utilizamos em nossas modelagens o modelo de regressão beta (Ferrari & Cribari–Neto, 2004). O procedimento computacional foi

desenvolvido utilizando o pacote `betareg` (Cribari–Neto & Zeleis, 2010) do software estatístico R. Na seleção das covariáveis utilizadas para cada um dos cenários acima utilizamos o critério de seleção de modelo AIC (*Akaike's information criterion*), que foi proposto por Akaike (1974).

Inicialmente, ao usar o modelo de regressão beta, nosso interesse reside em determinar se a precisão é fixa, ou seja, se há ou não estrutura de regressão para o parâmetro de precisão. Para tanto, empregamos o teste score (Espinheira, 2007) e concluímos que a precisão é variável para todos os cenários estudados. Sendo assim, além de modelar a média, há a necessidade de modelar a precisão. É importante destacar que as funções de ligação utilizadas para o primeiro modelo de regressão beta foram loglog e log para média e precisão, respectivamente. Nos dois últimos modelos de regressão foram consideradas a ligação logit para média, e para a precisão, log e raiz quadrada, respectivamente. Essas foram as ligações que conduziram aos melhores ajustes.

Na primeira modelagem, o modelo selecionado, após exaustiva seleção, foi

$$\begin{aligned} \text{loglog}(\mu_t) &= \beta_0 + \beta_1 QI_t + \beta_2 QI_t^2 + \beta_3 MULC_t + \beta_4 RENDA_t + \beta_5 \log ABERT_t \\ \log(\phi_t) &= \gamma_0 + \gamma_1 QI_t, \end{aligned}$$

$t = 1, \ldots, 124$.

A Tabela 5.5 apresenta os coeficientes estimados e seus respectivos p-valores. É possível verificar que as covariáveis *RENDA* e *ABERT* influenciam positivamente a proporção de ateus nas 124 nações, ao contrário da covariável *MULC*. Em relação à estrutura de regressão para o parâmetro de precisão, assim como foi apresentado no caso anterior, à medida que a covariável *QI* aumenta, a precisão diminui, ou seja, os países que apresentam maiores quocientes de inteligência tendem a apresentar respostas menos dispersas. Todas as covariáveis são significativas ao nível de 5%. A intensidade da não-constância da precisão é $\hat{\lambda} = \max(\hat{\phi}_t)/\min(\hat{\phi}_t) = 180.97$. O pseudo R^2 é 0.64.

Tabela 5.5 Estimativas pontuais e p-valores. Proporção de ateus nas 124 nações.

	$\hat{\beta}$	p-valor	$\hat{\gamma}$	p-valor
INTERCEPTO	6.042	$< 1e-04$	13.446	$< 1e-04$
QI	−0.222	$< 1e-04$	−0.118	$< 1e-04$
QI^2	0.002	$< 1e-04$	—	—
MULC	−0.097	0.0165	—	—
RENDA	0.007	0.0493	—	—
log*ABERT*	0.101	0.0169	—	—

Com o intuito de identificar possíveis afastamentos das suposições feitas para o modelo, a Figura 5.4 apresenta o gráfico da probabilidade normal com envelopes simulados e o gráfico dos valores observados versus valores preditos. O modelo de regressão parece estar bem ajustado, uma vez que os resíduos, em geral, permanecem dentro das bandas de confiança dos envelopes

simulados. Não há indícios de afastamento da suposição de que o modelo de regressão beta estimado é adequado para os dados.

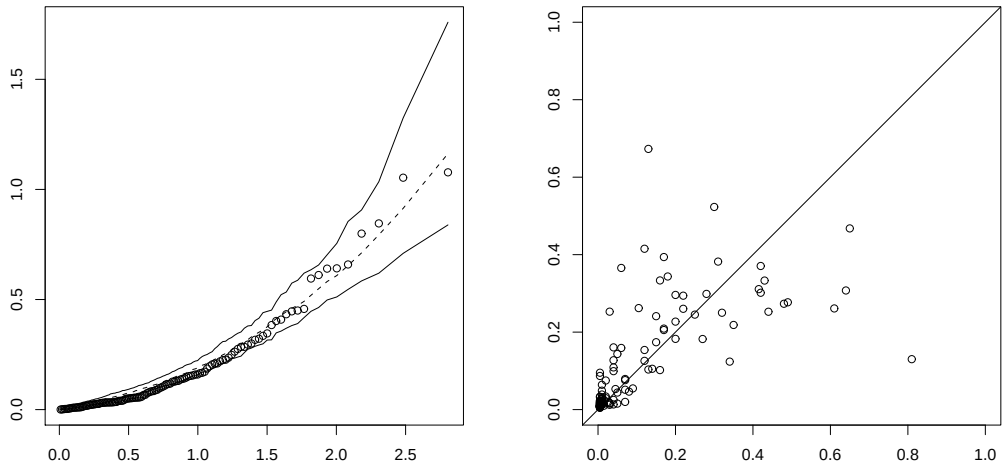

Figura 5.4 Gráfico da probabilidade normal com envelopes simulados (esquerda) e gráfico dos valores observados versus valores preditos (direita). Proporção de ateus em 124 nações.

Com o objetivo de complementar a análise realizada anteriormente, construímos o gráfico da distância de Cook versus os valores preditos, que está apresentado na Figura 5.5. Notamos que três nações se encontram destacadas das demais, são elas: Cingapura, Vietnã e Moçambique. Cingapura é uma nação que merece destaque em relação às demais porque, além de apresentar o maior *QI*, também tem o maior valor da covariável *ABERT*. Assim como Cuba, o Vietnã é uma exceção em relação aos demais países porque apresenta um *QI* mediano ($QI = 94$) e a maior proporção de ateus. Por fim, Moçambique é um país com baixo quociente médio de inteligência, e além disso, faz parte do grupo dos países que apresentam proporção de ateus maior do que a mediana, 0.024 (Tabela 5.2).

Além de Cuba e Vietnã, um outro país que Lynn et al. (2009) destacam como exceção são os Estados Unidos. Lá, a proporção de ateus (0.105) é baixa frente ao elevado nível médio de inteligência (98). Uma possível explicação pode estar no fato de que há um grande fluxo de imigrantes de países católicos, como por exemplo, o México, onde a proporção de ateus é apenas 0.045. Uma outra explicação reside na completa separação constitucional entre estado e religião, que facilitou o surgimento, a evolução e a adaptação de diversas novas religiões e vertentes religiosas. A competição entre elas por fiéis vem resultando em adaptação de discursos, pregações e práticas, o que, por sua vez, contribui para sua prosperidade. É como se houvesse um mercado competitivo em que religiões de diferentes denominações competem por fiéis. Como na seara econômica, competição conduz a dinamismo o que conduz, por sua vez, a prosperidade.

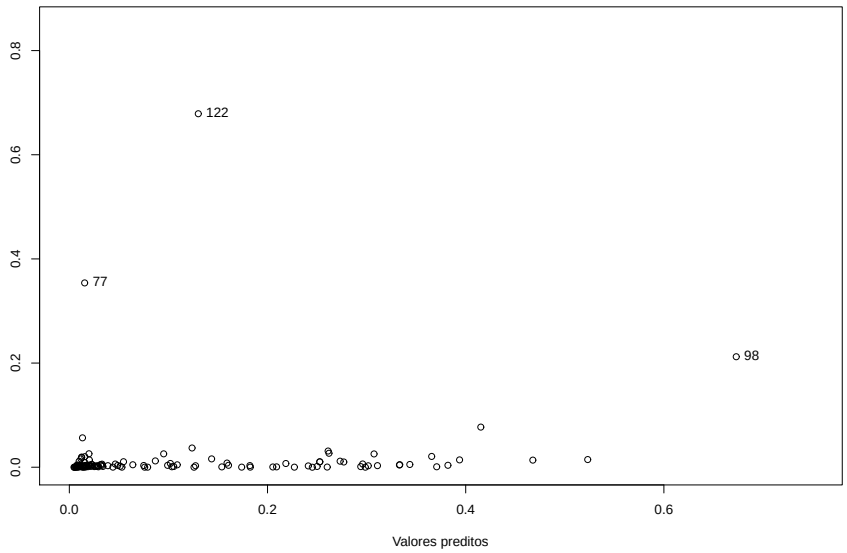

Figura 5.5 Distâncias de Cook. Proporção de ateus em 124 nações.

Excluímos individualmente e conjuntamente as observações 77, 98 e 122. As variações percentuais nas estimativas dos parâmetros encontram-se na Tabela 5.6. Notamos que a exclusão dos três casos é influente para as estimativas dos dois sub-modelos, o da média e o da precisão. O caso 122 (Vietnã) é o mais atípico, pois apresenta o maior valor da variável resposta (0.81) e um valor mediano para a variável *QI*. Ao excluir a observação 122, a variação percentual na estimativa de β_4 é consideravel (37.44%), dado o baixo valor da variável *RENDA* no Vietnã, ao passo que, a estimativa de β_5 cai cerca de em torno de 15%. A variação percentual na estimativa de γ_1 evidencia a influência que a observação 122 exerce sobre a estimativa da precisão. A exclusão dessa observação conduz a um decréscimo considerável na precisão, a estimativa de γ_1 cai 8.13%, enquanto que ao retirar a observação 77 a estimativa de γ_1 aumenta 17.62%. A observação 77 corresponde a Moçambique, que está entre os países com menor *QI* (64), mas onde a proporção de ateus está acima da mediana. Por fim, ao excluir a observação 98 (Cingapura), temos que as variações percentuais são menores comparadas aos casos 77 e 122, apesar dos maiores valores de *QI* e *ABERT*. Temos ainda que as observações 98 e 122 influenciam positivamente a modelagem da precisão, uma vez que suas exclusões diminuem o grau de heterogeneidade, enquanto que a observação 77 influencia negativamente essa precisão. Vale destacar que nos três países a predominância dos muçulmanos é menor do que 50%.

Um dos objetivos do presente capítulo é estimar o impacto exercido pelo quociente de inteligência sobre a preponderância média de ateus nos diferentes países. Esse impacto é obtido da seguinte forma:

$$\frac{\partial \mathbb{E}(y_t)}{\partial QI_t} = \frac{\partial \mu_t}{\partial QI_t}, \tag{5.4.1}$$

Tabela 5.6 Variações percentuais nas estimativas dos parâmetros ao se retirar observações influentes. Proporção de ateus em 124 nações.

Casos	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\lambda}$
77	−6.72	−4.38	−3.66	−13.04	20.06	−9.78	14.64	17.62	149.95
98	3.50	2.69	2.42	−0.73	2.83	5.24	−0.69	−1.07	−25.12
122	−6.65	−6.28	−6.94	−4.36	37.44	−15.22	−4.32	−8.13	−34.48
77, 98, 122	−9.92	−7.90	−8.08	−16.59	54.58	−17.88	8.61	6.97	11.56

em que

$$\mu_t = g^{-1}(\beta_0 + \beta_1 QI_t + \beta_2 QI_t^2 + \beta_3 MULC_t + \beta_4 RENDA_t + \beta_5 \log ABERT_t).$$

Considerando que a função de ligação para média é loglog, o impacto em (5.4.1) pode ser escrito como

$$\begin{aligned}\frac{\partial \mathbb{E}(\delta y_t)}{\partial QI_t} = \exp(-\exp(-(\beta_0 + \beta_1 QI_t + \beta_2 QI_t^2 + \beta_3 MULC_t + \beta_4 RENDA_t \\ + \beta_5 \log ABERT_t))) \times (-\exp(-(\beta_0 + \beta_1 QI_t + \beta_2 QI_t^2 + \beta_3 MULC_t \\ + \beta_4 RENDA_t + \beta_5 \log ABERT_t))) \times (-(\beta_1 + 2\beta_2 QI_t)).\end{aligned}$$

Para avaliar este impacto, consideramos duas situações diferentes, são elas: primeira, a covariável *RENDA* é fixada no primeiro, segundo e terceiro quartis, enquanto a covariável *ABERT* é fixada na mediana, e a covariável *MULC*, *dummy*, é fixada em um. Na segunda situação, as covariáveis *RENDA* e *ABERT* são fixadas em suas respectivas medianas e a covariável *MULC* é fixada em zero e em um, alternadamente. As duas situações consideradas podem ser visualizadas nas Figuras 5.6 e 5.7, respectivamente, de onde se nota que o efeito do impacto do *QI* é positivo e assume forma acelerada para valores de $QI < 107$. Para valores de *QI* maiores do que 107, o efeito continua sendo positivo, contudo apresenta forma desacelerada.

Para valores de $QI < 103$, temos que o impacto do *QI* sobre a preponderância média de ateus é maior na situação em que a *RENDA* é fixada no terceiro quartil. Para valores de *QI* superiores a 103, o impacto é mais acentuado nos países em que a *RENDA* é fixada no primeiro quartil, como mostra a Figura 5.6. Considerando a *RENDA* fixada na mediana, o maior impacto é 0.037.

A Figura 5.7 mostra que, para valores *QI* menores do que 104, o impacto de inteligência é maior entre os países classificados como não-muçulmanos, todavia, essa conclusão se inverte quando valores de *QI* maiores do que 104 são considerados.

Na segunda modelagem, consideramos apenas os países em que a proporção de descrença é igual ou maior do que 1%. Nesse ajuste, foram considerados 81 países. Consideramos as ligações logit e log para a média e para a precisão, respectivamente. As covariáveis selecionadas para a média foram *QI*, *RENDA* e *ABERT* e a covariável *QI* foi usada para modelar a precisão. Assim como no ajuste apresentado anteriormente, as covariáveis *QI*, *RENDA* e

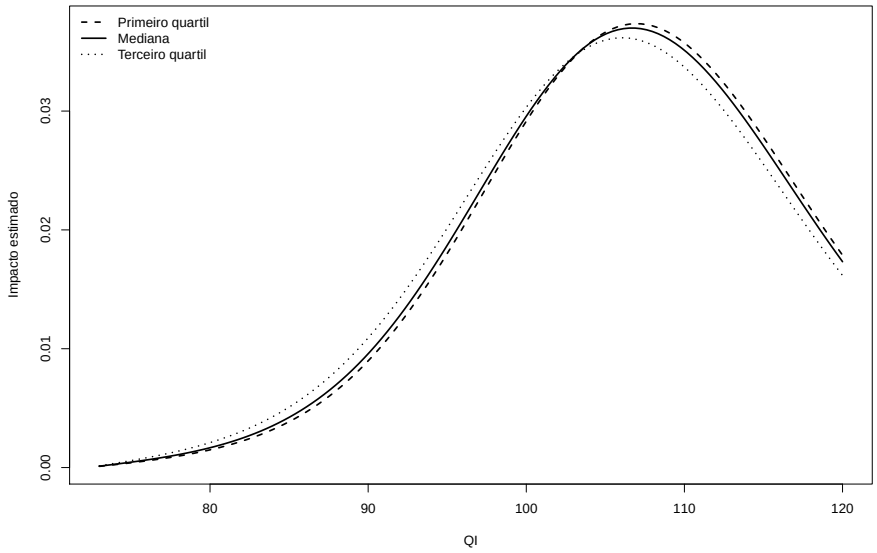

Figura 5.6 Impacto do quociente de inteligência sobre a prevalência de ateus fixando-se a covariável *RENDA* no primeiro, segundo e terceiro quartis. Proporção de ateus em 124 nações.

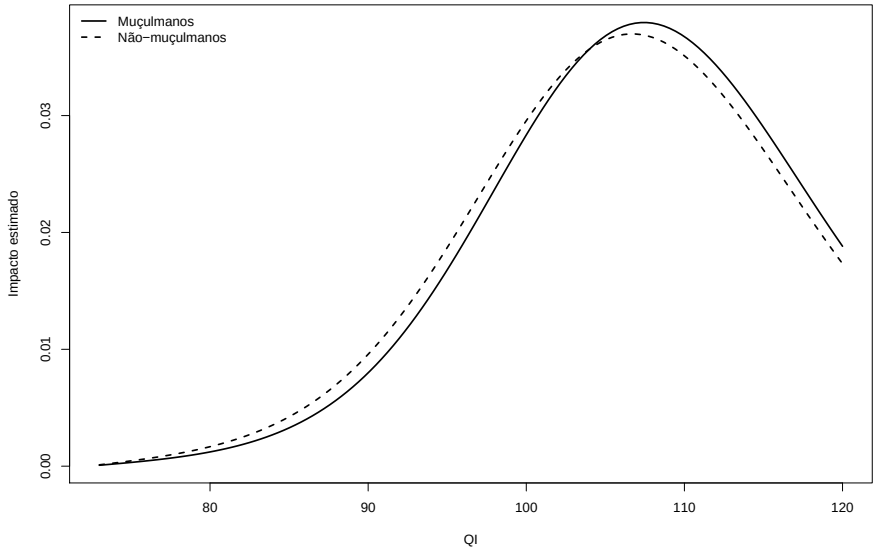

Figura 5.7 Impacto do quociente de inteligência sobre a prevalência de ateus entre os países muçulmanos e não-muçulmanos. Proporção de ateus em 124 nações.

$\log ABERT$ influenciam positivamente a proporção de ateus nas 81 nações. A intensidade de não-constância da precisão ($\hat{\lambda}$) foi 159.59.

Dado que na segunda modelagem consideramos a função de ligação logit, a expressão (5.4.1), considerando as covariáveis selecionadas para o segundo ajuste é dada por:

$$\frac{\partial \mathbb{E}(y_t)}{\partial QI_t} = (\beta_1 + 2\beta_2 QI_t)\frac{\exp(\beta_0 + \beta_1 QI_t + \beta_2 QI_t^2 + \beta_3 RENDA_t + \beta_4 \log ABERT_t}{(1+\exp(\beta_0 + \beta_1 QI_t + \beta_2 QI_t^2 + \beta_3 RENDA_t + \beta_4 \log ABERT_t))^2}.$$

Na terceira modelagem, consideramos os cinquenta países que apresentam as maiores prevalências de ateísmo. O menor valor registrado da proporção de ateus é 0.05 e o maior, 0.81. Dos cinquenta países, há dois em que o percentual de muçulmanos ultrapassa 50%, são eles: Albânia e República do Quirguistão. Entre os 50 países considerados, apenas Moçambique apresenta quociente de inteligência menor do que 82, o valor do QI médio neste país é 64. Diferentemente das outras modelagens apresentadas, aqui a covariável $EXPEC$, que exerce influência positiva sobre a proporção de ateus, foi significativa. Utilizamos as ligações logit e raiz quadrada para a média e para a precisão, respectivamente. Além das covariáveis utilizadas na segunda modelagem para a média, a covariável $EXPEC$ e a interação entre as covariáveis $RENDA$ e $\log ABERT$ foram selecionadas. Para modelar a precisão, consideramos as covariáveis QI, $RENDA$ e $EXPEC$. A intensidade da não-constância da precisão é $\hat{\lambda} = 125.75$.

A Tabela 5.7 apresenta os impactos estimados do quociente de inteligência sobre a proporção média de ateus. Esses impactos foram obtidos considerando os três modelos de regressão beta descritos anteriormente. Fixamos a covariável QI em 85, 95 e 110. As demais covariáveis foram fixadas na mediana, exceto $MULC$, que foi fixada em zero (não-muçulmanos). Os impactos estimados do QI são maiores quando consideramos o modelo 1 (cenário 1), exceto quando $QI = 110$, nesse caso o impacto calculado a partir do modelo 2 (cenário 2) é mais intenso do que os demais. Através dos impactos estimados, podemos notar que à medida que o QI aumenta, o impacto estimado do QI sobre a proporção média de ateus se acentua. Por exemplo, no cenário 2 com $QI = 85$, o impacto estimado é 0.0040 contudo, ao considerar $QI = 110$, o impacto estimado passa a ser 0.0448, ou seja, quase onze vezes maior.

Tabela 5.7 Impactos estimados da inteligência sobre a proporção média de ateus.

QI	cenário 1	cenário 2	cenário 3
85	0.0043	0.0040	0.0008
95	0.0192	0.0180	0.0140
110	0.0349	0.0448	0.0423

Consideramos o método bootstrap paramétrico para a obtenção do erro-padrão associado a cada impacto estimado apresentado na Tabela 5.7. Utilizamos 1000 réplicas de bootstrap. Resumimos a seguir o procedimento bootstrap adotado.

1. Para cada $t = 1, \ldots, n$, geramos $y_t^{\dagger}$ da densidade beta com parâmetros $\hat{\mu}_t$ e $\hat{\sigma}_t$.

2. Estimamos a regressão beta usando como resposta $y = (y_1^\dagger, \ldots, y_n^\dagger)^\top$ conjuntamente com os regressores originais.

3. Considerando as estimativas obtidas no passo anterior, calculamos o impacto da covariável QI, ou seja, estimamos

$$\frac{\partial \mathbb{E}(y_{t,b})}{\partial QI_t} = \frac{\partial \mu_{t,b}}{\partial QI_t}.$$

4. Repetimos os passos (1) a (3) 1000 vezes.

5. Calculamos o desvio-padrão dos 1000 impactos estimados.

Utilizamos também o procedimento bootstrap paramétrico para construir intervalos de confiança ao nível 95% para os impactos. Para tanto, tomamos como limites inferior e superior os quantis 0.025 e 0.975 dos 1000 impactos calculados via bootstrap.

A Tabela 5.8 apresenta os erros-padrão para os impactos estimados da inteligência sobre a proporção média de ateus e a Tabela 5.9 apresenta os intervalos de confiança obtidos ao nível de 95% de confiança via bootstrap para o impacto da inteligência sobre a proporção média de ateus. Podemos observar que os erros-padrão obtidos via bootstrap são pequenos. No que tange aos intervalos de confiança, notamos que o valor zero pertence a apenas um dos intervalos apresentados, sendo assim o impacto estimado nesse caso ($QI = 85$ no cenário 3) é estatisticamente nulo. Nos demais casos apresentados, os impactos estimados são estatisticamente diferentes de zero.

Tabela 5.8 Erros-padrão bootstrap para os impactos estimados da inteligência sobre a proporção média de ateus.

QI	cenário 1	cenário 2	cenário 3
85	0.0006	0.0005	0.0031
95	0.0031	0.0037	0.0042
110	0.0014	0.0051	0.0064

Tabela 5.9 Intervalos de confiança bootstrap para o impacto da inteligência sobre a proporção média de ateus.

QI	cenário 1	cenário 2	cenário 3
85	(0.0031;0.0054)	(0.0029;0.0050)	(−0.0074;0.0044)
95	(0.0130;0.0256)	(0.0114;0.0258)	(0.0052;0.0216)
110	(0.0313;0.0368)	(0.0307;0.0503)	(0.0089;0.0384)

A Figura 5.8 apresenta os impactos estimados do *QI* obtidos a partir dos modelos 1, 2 e 3, ou seja, cenários 1, 2 e 3, respectivamente, sobre a proporção média de ateus. Notamos que, para valores de *QI* menores do que 96, o impacto estimado a partir do modelo 1 (cenário 1) é maior do que o estimado a partir do modelo 2 (cenário 2). Para valores de *QI* entre 96 e 114, a conclusão a respeito dos impactos obtidos a partir dos modelos 1 e 2 se inverte. Adicionalmente, podemos observar que para valores de *QI* menores do que 103, o impacto estimado a partir do modelo 1 é maior do que o obtido através do modelo 3 (cenário 3). Vale ressaltar que ao se considerar o modelo 3, o maior valor do impacto estimado é 0.04365, ao passo que ao se considerar o cenário 1 o valor máximo é 0.03698.

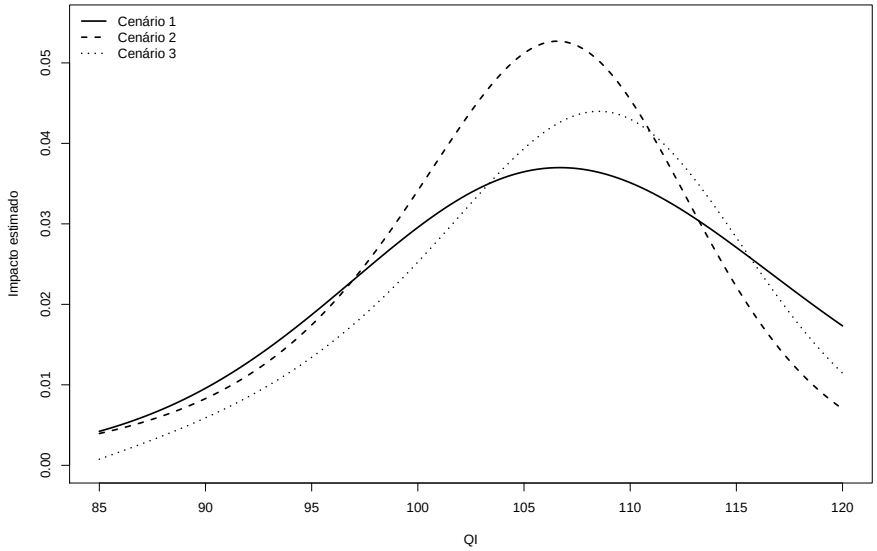

Figura 5.8 Impacto da inteligência sobre a prevalência de ateus em três cenários.

Adicionalmente, obtemos as estimativas das proporções médias de ateus fixando a covariável *QI* em 80, 90, 100 e 110, e fixando as demais covariáveis em suas medianas, exceto *MULC* que foi fixada em zero (não-muçulmanos). As estimativas estão apresentados na Tabela 5.10. De acordo com os resultados apresentados observamos que, à medida que aumentamos o valor da covariável *QI*, a estimativa da proporção média de ateus aumenta rapidamente. Por exemplo, considerando o cenário 1 e $QI = 90$, a estimativa da proprção de ateus é 0.0643; ao considerar $QI = 100$, essa estimativa é quase quatro vezes maior (0.2545). As miores estimativas da proporção de ateus foram obtidas quando consideramos o cenário 2, exceto quando $QI = 110$.

Além das três modelagens empíricas apresentadas nessa seção, ajustamos um modelo de regressão beta em que a variável resposta é a proporção de pessoas que não consideram religião importante para suas vidas cotidianas. Para esse modelo consideramos as seguintes covariáveis: QI, QI^2, $RENDA$, $ABERT$ e $MULC$. Assim como na primeira modelagem empírica, as funções de ligação utilizadas foram loglog e log para a média e precisão, respectivamente. Essa modelagem é baseada em 84 nações das quais 19 foram classificadas como muçulmanas. A intensidade da não-constância da precisão é $\hat{\lambda} = 59.41$. O pseudo R^2 obtido foi 0.64.

Tabela 5.10 Estimativas da proporções médias de ateus.

QI	cenário 1	cenário 2	cenário 3
80	0.0174	0.1500	0.0357
90	0.0643	0.1582	0.0789
100	0.2545	0.2997	0.2663
110	0.6050	0.6758	0.7382

A Figura 5.9 apresenta o gráfico das distâncias de Cook versus valores preditos e da probabilidade normal com envelopes simulados. A partir do gráfico das distâncias de Cook podemos observar que três nações se encontram destacadas das demais, são elas: República do Uzbequistão, Albânia e Cingapura. Na República do Uzbequistão e Albânia, o percentual de muçulmanos ultrapassa 50%. O modelo de regressão parece estar bem ajustado, dado que os resíduos permanecem dentro das bandas de confiança dos envelopes simulados.

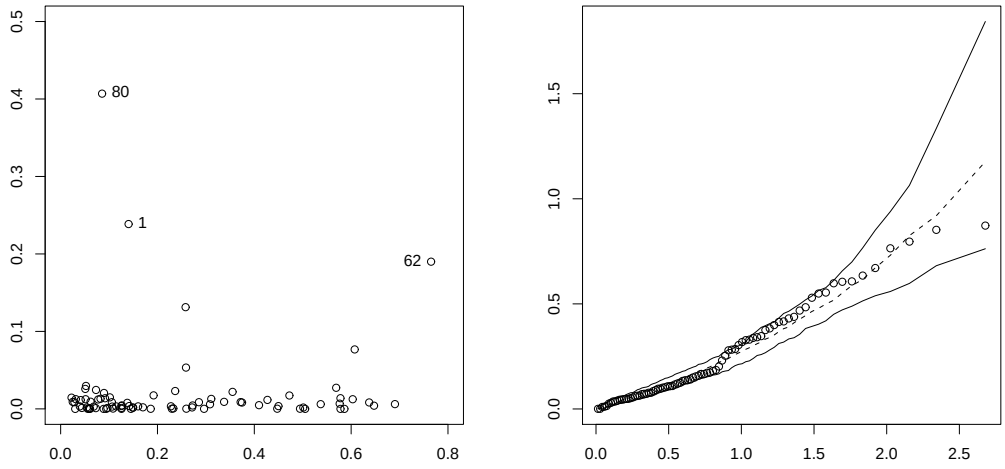

Figura 5.9 Gráficos das distâncias de Cook (esquerda) e da probabilidade normal com envelopes simulados (direita). Proporção de pessoas que não consideram religião importante para suas vidas cotidianas em 84 nações.

Assim como nos casos anteriores, calculamos o impacto do quociente de inteligência sobre a proporção de pessoas que não consideram religião importante nas 84 nações (Figura 5.10). Notamos que o efeito do impacto do *QI* é positivo e assume forma acelerada para valores de $QI < 109$. Para valores de *QI* maiores do que 109, o efeito continua sendo positivo, contudo apresenta forma desacelerada.

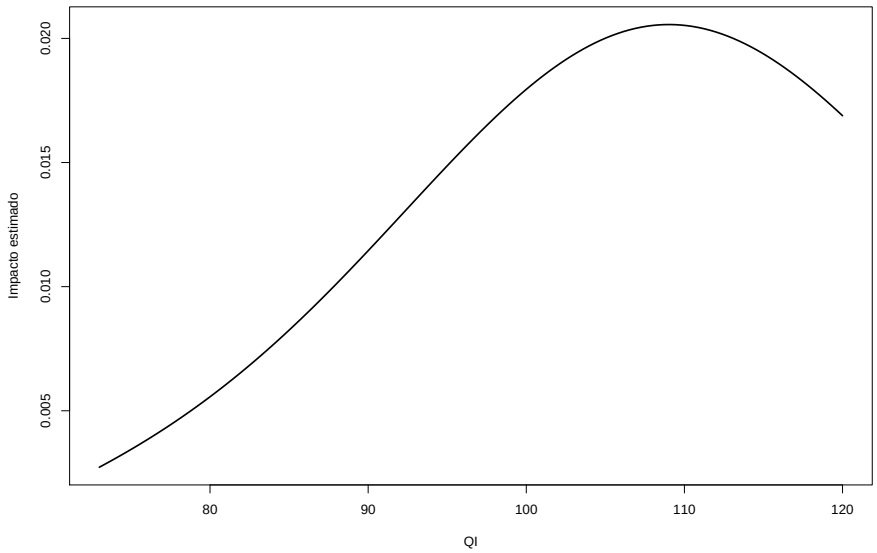

Figura 5.10 Impacto da inteligência sobre a proporção de pessoas que não consideram religião importante em 84 nações.

As quatro modelagens apresentadas que utilizaram duas variáveis resposta diferentes conduzem à mesma conclusão, qual seja: existe impacto positivo da inteligência sobre descrença religiosa e se esse impacto se dá de forma acelerada até níveis médios de inteligência ligeiramente abaixo de 110, apresentando desaceleração a partir desse nível. Esse efeito existe mesmo quando outras variáveis condicionantes são levadas em consideração.

Por fim, consideramos a estimação do modelo de regressão beta a partir do cenário 1, mas assumindo dispersão fixa. Os testes de significância das variáveis condicionantes são feitos a partir de erros-padrão robustos ('estimadores sanduíche'); ver Capítulo 3. Utilizamos o estimador denominado S_2 naquele capítulo. As estimativas pontuais e os p-valores encontram-se na Tabela 5.11. Assim como foi apresentado na Tabela 5.5 (modelo de regressão beta com dispersão variável), todas as covariáveis foram significativas aos níveis usuais de significância, exceto a covariável $\log ABERT$, que perde a significância. O pseudo R^2 desse modelo é 0.62, que é muito próximo do modelo em que a dispersão é variável (0.64).

Estimamos o impacto exercido pelo quociente de inteligência sobre a preponderância média de ateus nos diferentes países considerando o modelo de regressão beta com dispersão fixa, a fim de comparar tal impacto com aquele obtido sob dispersão variável. Para avaliar tal impacto, fixamos as covariáveis $RENDA$ e $ABERT$ em suas medianas, e fixamos a covariável $MULC$ em zero (não-muçulmanos). As estimativas do impacto da inteligência obtidas a partir dos modelos de regressão beta com dispersão variável e fixa sobre a preponderância média de ateus podem ser visualizadas na Figura 5.11. Note que o efeito do impacto do QI é positivo considerando os dois modelos de regressão beta, contudo o impacto do QI é mais intenso quando o modelo de regressão beta com dispersão variável é considerado. Tal conclusão, todavia, é invertida para os níveis mais elevados de inteligência (a partir de aproximadamente 115).

Tabela 5.11 Estimativas pontuais e *p*-valores (modelo de regressão beta com dispersão fixa). Proporção de ateus nas 124 nações.

	$\hat{\beta}$	*p*-valor	$\hat{\gamma}$	*p*-valor
INTERCEPTO	4.195	$< 6e-03$	2.294	$< 1e-15$
QI	-0.153	$< 1e-04$	—	—
QI^2	0.002	0.030	—	—
$MULC$	-0.092	$< 7e-03$	—	—
$RENDA$	0.009	$< 7e-05$	—	—
$\log ABERT$	0.064	0.382	—	—

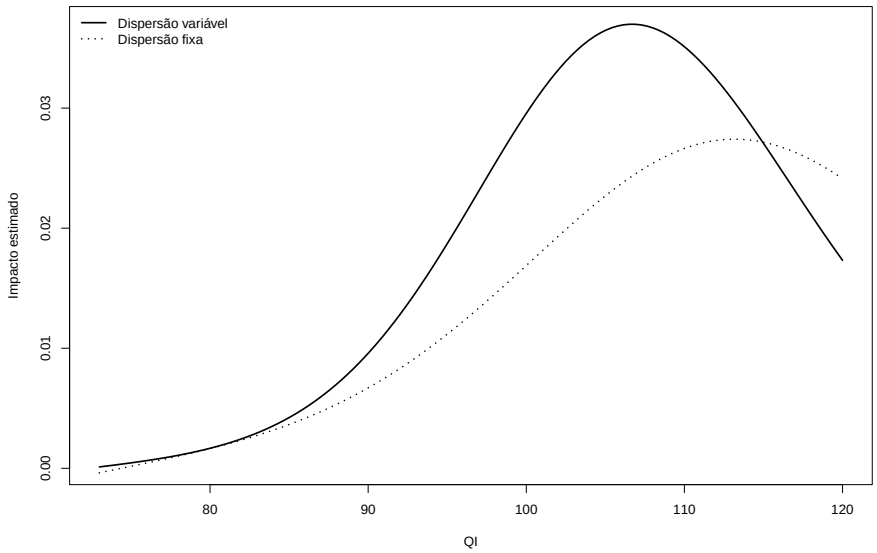

Figura 5.11 Impacto do quociente de inteligência estimado a partir do modelo de regressão com dispersão variável e fixa sobre a prevalência de ateus fixando-se as covariáveis $RENDA$ e $ABERT$ na mediana. Proporção de ateus em 124 nações.

5.5 Conclusões

Assim como Lynn et al. (2009), notamos que há uma forte associação positiva entre inteligência e ateísmo. Essa associação é mais intensa em países cujo quociente médio de inteligência está entre 87 e 108. Há algumas exceções nesse padrão, como por exemplo, Cuba e Vietnã, onde as proporções de ateus são 0.40 e 0.81, respectivamente, embora apresentem níveis medianos de inteligência. Os Estados Unidos também merecem destaque, pois apresentam *QI* alto (98), mas a proporção de ateus é baixa (0.105).

Através do modelo de regressão beta com dispersão variável, modelamos a proporção de ateus considerando três cenários diferentes. No primeiro cenário, ajustamos o modelo de regressão considerando o quociente de inteligência como variável explicativa e outros regressores: renda nacional bruta ajustada pela paridade do poder de compra, grau de abertura da economia e predominância de muçulmanos. A utilização da renda nacional como covariável se destina a levar em consideração o fato de que há maior prevalência de ateus em países desenvolvidos. No segundo cenário, consideramos apenas os países em que o percentual de descrentes é igual ou maior do que 1%. Por fim, no último cenário, consideramos os cinquenta países que apresentam as maiores proporções de descrentes.

Estimamos o impacto exercido pelo quociente médio de inteligência sobre a prevalência de ateus nos diferentes países. Os resultados revelaram que os maiores impactos estimados foram obtidos quando consideramos os países em que a proporção de descrença é igual ou maior do que 1%. Através dos impactos estimados, podemos notar que o impacto do *QI* é positivo e assume forma acelerada para valores de $QI < 107$. Para valores de *QI* maiores do que 107, o efeito continua sendo positivo, todavia de forma desacelerada.

Consideramos nesse capítulo um outro modelo de regressão beta em que a variável resposta é a proporção de pessoas que não consideram religião importante para suas vidas cotidianas. Os resultados revelaram que o impacto do *QI* é positivo e assume forma acelerada para valores de $QI < 109$. Para valores de *QI* maiores do que 109, o efeito continua sendo positivo, contudo de forma desacelerada. Vale ressaltar aqui que o impacto estimado obtido a partir do modelo regressão beta em que a variável resposta é proporção de pessoas que não consideram religião importante para suas vidas cotidianas é menor que os impactos obtidos a partir dos outros modelos em que a variável resposta é a proporção de ateus.

Capítulo 6

Considerações finais

6.1 Conclusões

Ao longo desta tese desenvolvemos alguns aspectos de inferência e ilustramos a aplicabilidade prática de modelos de regressão com dispersão variável. As principais contribuições e conclusões deste trabalho encontram-se resumidas a seguir.

1. No Capítulo 2, avaliamos, através dos tamanhos e poderes de testes quasi-t e quasi-F, os comportamentos dos estimadores HC0, HC3 e HC4 da matriz de covariâncias do estimador de MQO quando resíduos oriundos de regressões robustas (menor mediana dos quadrados, mínimos quadrados podados e mínimos quadrados ponderados) são usados em substituição aos resíduos de MQO, no modelo de regressão linear com e sem restrição sobre os parâmetros. Os resultados revelaram que o teste construído a partir do estimador HC0 apresenta melhor desempenho quando resíduos oriundos de regressão robusta são utilizados em substituição aos resíduos de MQO, na presença de pontos de alavanca. A mesma conclusão foi obtida na maioria dos cenários estudados sobre o teste baseado no estimador HC3. O teste baseado no estimador HC4 apresenta desempenho superior quando resíduos oriundos do estimador MQO e mínimos quadrado ponderados são utilizados. O desempenho do teste baseado no estimador HC0 melhora significativamente quando estimação restrita é utilizada, principalmente com resíduos robustos.

2. No Capítulo 3, propomos estimadores do tipo sanduíche para modelos de regressão beta em que a estrutura de regressão para o parâmetro de dispersão é negligenciada. Avaliamos os desempenhos de testes z que utilizam os erros-padrão obtidos da forma usual e testes quasi-z que utilizam erros-padrão dos estimadores sanduíches propostos. Para tal avaliação nós consideramos dois cenários diferentes ao estimar o modelo de regressão beta. No primeiro, consideramos que o verdadeiro processo gerador de dados tem dispersão variável, contudo modelamos a estrutura de regressão da dispersão de forma incorreta. No total, sete tipos de erros de especificação foram cometidos. No segundo cenário, geramos dados com dispersão fixa, mas estimamos o modelo supondo dispersão variável. Vale ressaltar que nos dois cenários consideramos o caso de especificação correta, ou seja, o modelo foi estimado considerando a forma como foi gerado. Em relação aos testes z, nossos resultados revelaram que é mais grave errar na especificação do preditor linear da estrutura de regressão da dispersão do que na função de ligação do submodelo da média ou do submodelo da dispersão. Adicionalmente, percebemos que ao utilizarmos o estimador usual da matriz de covariâncias, as taxas de rejeição excedem consideravelmente os níveis nominais, ou seja, o teste baseado neste estimador rejeita a

hipótese nula incorretamente muito mais frequentemente do que deveria. Em contrapartida, ao utilizarmos os estimadores do tipo sanduíche propostos, os resultados se tornam muito mais interessantes, ou seja, as taxas de rejeição sob $\mathcal{H}_0$ ficam próximas dos níveis nominais, principalmente quando estes testes são baseados nos estimadores S_2 e S_3. Como esperávamos, para o caso em que o modelo estava corretamente especificado, as taxas de rejeição do teste *z* ficaram próximas dos níveis nominais considerados. Apesar de termos bons resultados neste cenário, vale ressaltar que em situações práticas é difícil modelar a dispersão de forma adequada. Os resultados obtidos nas simulações de poder revelaram que os poderes dos testes quando o modelo está corretamente especificado são altos. Adicionalmente os nossos resultados revelaram que os poderes dos testes quasi-*z* são baixos comparados aos poderes dos teste *z* no modelo corretamente especificado. Quando há erros de especificação, os poderes dos testes *z* são muito mais baixos. Os nossos testes quasi-*z* apresentaram bom desempenho em termos de tamanho e superaram os testes *z* no cenário em que a especificação incorreta é considerada. Concluímos ainda que os testes quasi-*z* foram inclusive competitivos com o teste *z* sob especificação correta no que tange ao controle da probabilidade de erro do tipo I.

3. No Capítulo 4, consideramos uma estratégia de modelagem para avaliar o efeito do aumento nos gastos com programas assistenciais e do crescimento da economia sobre a reeleição do presidente Lula em 2006. Através do modelo de regressão beta com dispersão variável, modelamos a diferença entre as proporções de votos válidos do ex-presidente Lula nos segundos turnos das eleições de 2006 e 2002. Adicionalmente, calculamos o impacto da diferença entre os gastos totais em programas assistenciais em 2006 e 2002 sobre a diferença da proporção de votos válidos do ex-presidente Lula em 2006 e 2002. Nossos resultados revelaram que se em todos os municípios os prefeitos fossem filiados ao PT e os respectivos estados fossem governados pelo PT, se todos os municípios fossem rurais, e se os gastos nos programas assistenciais fossem mantidos a níveis de 2002 a redução de votos do ex-presidente seria de aproximadamente 12%. Também calculamos o impacto do crescimento da economia sobre a diferença das proporções de votos do ex-presidente Lula em 2006 e 2002. Concluímos que se o crescimento da economia tivesse sido nulo, se todos os estados e municípios fossem governados pelo PT e se todos fossem localizados na zona rural, haveria uma redução de quase 6% na votação do ex-presidente Lula. Por outro lado, se todos os estados e municípios não fossem governados pelo PT encontramos que sem o crescimento da economia haveria um acréscimo na votação do ex-presidente Lula independente do município ser localizado na zona rural ou urbana. Considerando o Brasil como um todo, concluímos que se os gastos com programas assistenciais fossem mantidos nos níveis de 2002, a votação do ex-presidente Lula teria uma redução de aproximadamente 7 milhões de votos, enquanto que, sem crescimento econômico, essa redução seria de apenas 2 milhões de votos, mas mesmo com tais reduções de votos, o ex-presidente Lula seria reeleito, dado que a diferença entre os votos de Lula e de Geraldo Alckmin foi de 22% que corresponde a aproximadamente 21 milhões de votos.

4. No Capítulo 5, nossos resultados revelaram que há associação positiva entre inteligência

e ateísmo. Verificamos que essa associação é mais intensa em países cujo quociente médio de inteligência é mais alto. Contudo, há algumas exceções nesse padrão, como por exemplo Cuba, Vietnã e os Estados Unidos. Através do modelo de regressão beta com dispersão variável, modelamos duas variáveis respostas, são elas: a proporção de ateus e a proporção de pessoas que não consideram religião importante para suas vidas cotidianas. No ajuste em que consideramos como variável resposta a proporção de ateus, três cenários diferentes foram avaliados. No primeiro cenário, ajustamos o modelo de regressão considerando o quociente de inteligência como variável explicativa e outros regressores: renda nacional bruta ajustada pela paridade do poder de compra, grau de abertura da economia e predominância de muçulmanos. No segundo cenário, consideramos apenas os países em que o percentual de descrentes é igual ou maior do que 1%. Por fim, no último cenário, consideramos os cinquenta países que apresentam as maiores proporções de descrentes. Adicionalmente, estimamos o impacto exercido pelo quociente médio de inteligência sobre a prevalência de ateus nos diferentes países e a proporção de pessoas que não consideram religião importante para suas vidas. Os resultados revelaram que as quatro modelagens apresentadas conduzem à mesma conclusão, qual seja: existe impacto positivo da inteligência sobre a descrença religiosa e esse impacto se dá de forma acelerada até níveis médios de inteligência ligeiramente abaixo de 110; para valores de *QI* acima desse nível o efeito continua sendo positivo, todavia assume forma desacelerada. Vale ressaltar que os maiores impactos estimados foram obtidos quando consideramos os países em que a proporção de descrença é igual ou maior do que 1%.

6.2 Trabalhos futuros

Algumas linhas de pesquisa devem ser desenvolvidas. Por exemplo, serão foco de nossas pesquisas futuras:

1. Estimação intervalar em regressões lineares a partir de estimações pontuais e resíduos robustos.

2. Estimadores sanduíche para regressão beta não-linear.

3. Analisar a mais recente eleição presidencial no Brasil, onde foi eleita a presidente da República a candidata do PT, Dilma Rousseff.

4. Utilização de regressão quantílica nas análises empíricas dos Capítulos 4 e 5.

Apêndice A

Poderes de testes quasi-*t* sob heteroscedasticidade

Neste apêndice apresentamos os resultados de simulação de poder sob heteroscedasticidade para $n = 50$ e diferentes valores de β_2 (0.05, 0.10, 0.20). A hipótese nula em teste é $\mathcal{H}_0 : \beta_2 = 0$ e os dados foram gerados utilizando $\beta_2 = 0.05, 0.10, 0.20$. Aqui objetivamos avaliar os comportamentos das principais estatísticas apresentadas no Capítulo 2 quanto à habilidade de conduzir baixas probabilidades do erro tipo II considerando diferentes valores para β_2. O número de réplicas de Monte Carlo foi 100,000. Os resultados revelam que para valores de β_2 mais distantes de zero, maiores são os poderes dos testes. Por exemplo, gerando os dados com $\beta_2 = 0.05$, ao nível nominal de 10% o poder do teste baseado no estimador $HC0_4$ é 35.66%, todavia, gerando os dados com $\beta_2 = 0.20$, o poder do teste baseado neste estimador é 99.51%.

Tabela A.1 Poderes de testes quasi-t baseados nos estimadores $\mathrm{HC}j_k$, $j = 0,3,4, k = 1,\ldots,6$ no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim \mathcal{N}(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n = 50$. Os níveis nominais considerados são $\alpha = 0.10, 0.05$ (com pontos de alavanca) - 'dados1'. A hipótese em teste é $\mathcal{H}_0 : \beta_2 = 0$ e os dados foram gerados usando $\beta_2 = 0.05, 0.10, 0.20$. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

β_2	α	$\mathrm{HC0}_1$	$\mathrm{HC0}_2$	$\mathrm{HC0}_3$	$\mathrm{HC0}_4$	$\mathrm{HC0}_5$	$\mathrm{HC0}_6$
	10	26.60	25.91	26.02	35.66	32.78	32.63
	5	16.74	15.23	14.83	24.91	21.47	21.32
	α	$\mathrm{HC3}_1$	$\mathrm{HC3}_2$	$\mathrm{HC3}_3$	$\mathrm{HC3}_4$	$\mathrm{HC3}_5$	$\mathrm{HC3}_6$
	10	25.45	24.89	25.03	34.79	31.38	30.85
0.05	5	15.64	13.72	13.74	24.19	20.03	19.78
	α	$\mathrm{HC4}_1$	$\mathrm{HC4}_2$	$\mathrm{HC4}_3$	$\mathrm{HC4}_4$	$\mathrm{HC4}_5$	$\mathrm{HC4}_6$
	10	24.52	24.13	12.96	33.73	28.91	28.81
	5	14.77	24.22	12.99	22.86	18.10	17.87
	α	$\mathrm{HC0}_1$	$\mathrm{HC0}_2$	$\mathrm{HC0}_3$	$\mathrm{HC0}_4$	$\mathrm{HC0}_5$	$\mathrm{HC0}_6$
	10	60.75	59.64	59.56	76.85	72.74	72.61
	5	45.69	40.88	40.36	66.51	59.03	58.82
	α	$\mathrm{HC3}_1$	$\mathrm{HC3}_2$	$\mathrm{HC3}_3$	$\mathrm{HC3}_4$	$\mathrm{HC3}_5$	$\mathrm{HC3}_6$
	10	57.28	56.86	57.16	74.79	69.70	69.14
0.10	5	41.90	36.33	36.40	63.85	54.51	54.25
	α	$\mathrm{HC4}_1$	$\mathrm{HC4}_2$	$\mathrm{HC4}_3$	$\mathrm{HC4}_4$	$\mathrm{HC4}_5$	$\mathrm{HC4}_6$
	10	54.81	54.80	55.37	71.93	64.43	64.67
	5	39.04	33.71	33.81	58.80	48.05	47.93
	α	$\mathrm{HC0}_1$	$\mathrm{HC0}_2$	$\mathrm{HC0}_3$	$\mathrm{HC0}_4$	$\mathrm{HC0}_5$	$\mathrm{HC0}_6$
	10	96.57	95.23	95.40	99.51	98.67	98.69
	5	90.32	86.27	86.16	98.64	96.40	96.45
	α	$\mathrm{HC3}_1$	$\mathrm{HC3}_2$	$\mathrm{HC3}_3$	$\mathrm{HC3}_4$	$\mathrm{HC3}_5$	$\mathrm{HC3}_6$
	10	94.69	94.16	94.42	99.24	97.99	97.97
0.20	5	85.96	81.66	82.06	97.87	94.40	94.45
	α	$\mathrm{HC4}_1$	$\mathrm{HC4}_2$	$\mathrm{HC4}_3$	$\mathrm{HC4}_4$	$\mathrm{HC4}_5$	$\mathrm{HC4}_6$
	10	92.90	93.08	93.58	98.69	96.45	96.63
	5	82.20	77.95	78.58	96.26	90.10	90.43

Tabela A.2 Poderes de testes quasi-t baseados nos estimadores HCj_k, $j = 0,3,4, k = 1,\ldots,6$ no modelo $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$, $e_t \sim \mathcal{N}(0, \exp\{c_1 x_{t1}\})$, $t = 1,\ldots,n$, para $n = 50$. Os níveis nominais considerados são $\alpha = 0.10, 0.05$ (com pontos de alavanca) - 'dados1'. A hipótese em teste é $\mathcal{H}_0 : \beta_2 = 0$ e os dados foram gerados usando $\beta_2 = 0.05, 0.10, 0.20$. São usados resíduos oriundos dos estimadores MQO, LMS, LTS, WLS, WLS-LMS, WLS-LTS.

β_2	α	HCR0$_1$	HCR0$_2$	HCR0$_3$	HCR0$_4$	HCR0$_5$	HCR0$_6$
	10	18.91	25.31	24.81	34.68	32.60	31.58
	5	8.94	14.14	13.99	23.29	19.75	19.58
	α	HCR3$_1$	HCR3$_2$	HCR3$_3$	HCR3$_4$	HCR3$_5$	HCR3$_6$
	10	21.89	23.44	23.16	34.57	30.96	30.57
0.05	5	12.09	12.74	12.72	23.13	19.04	18.99
	α	HCR4$_1$	HCR4$_2$	HCR4$_3$	HCR4$_4$	HCR4$_5$	HCR4$_6$
	10	24.14	20.79	20.63	34.56	29.61	28.70
	5	15.52	11.40	10.90	22.74	17.75	17.39
λ	α	HCR0$_1$	HCR0$_2$	HCR0$_3$	HCR0$_4$	HCR0$_5$	HCR0$_6$
	10	36.81	56.90	55.48	75.70	72.59	71.14
	5	15.23	37.51	36.53	63.28	55.06	53.96
	α	HCR3$_1$	HCR3$_2$	HCR3$_3$	HCR3$_4$	HCR3$_5$	HCR3$_6$
	10	46.12	51.67	50.53	75.55	69.71	68.73
0.10	5	27.00	32.35	32.04	62.84	20.69	20.16
	α	HCR4$_1$	HCR4$_2$	HCR4$_3$	HCR4$_4$	HCR4$_5$	HCR4$_6$
	10	51.18	66.00	64.27	75.38	70.02	69.81
	5	37.69	46.98	45.70	61.88	58.09	57.51
λ	α	HCR0$_1$	HCR0$_2$	HCR0$_3$	HCR0$_4$	HCR0$_5$	HCR0$_6$
	10	68.08	93.50	92.76	99.48	99.10	98.95
	5	21.10	80.53	79.33	98.36	95.07	94.56
	α	HCR3$_1$	HCR3$_2$	HCR3$_3$	HCR3$_4$	HCR3$_5$	HCR3$_6$
	10	84.90	88.01	86.99	99.45	92.26	91.83
0.20	5	57.87	68.73	68.51	98.23	75.92	75.83
	α	HCR4$_1$	HCR4$_2$	HCR4$_3$	HCR4$_4$	HCR4$_5$	HCR4$_6$
	10	87.68	73.84	73.82	99.42	69.72	69.39
	5	75.65	47.46	46.92	97.89	42.20	42.61

Apêndice B

Avaliação do desempenho do teste escore

Neste apêndice apresentamos a estatística de teste escore para testar se a dispersão depende de covariáveis e parâmetros desconhecidos, ou seja, se a dispersão é variável no modelo de regressão beta. Considere a hipótese nula $\mathcal{H}_0 : \sigma_1 = \cdots = \sigma_n = \sigma$. Testar esta hipótese equivale a testar

$$\mathcal{H}_0 : \gamma_{q-1} = 0, \tag{B.0.1}$$

em que $\gamma_{q-1} = (\gamma_2, \ldots, \gamma_q)^\top$ no Modelo definido em (3.2) com $z_{t1} = 1$ para $t = 1, \ldots, n$.

Seja $U_{(q-1)\gamma}$ o vetor coluna $(q-1)$-dimensional contendo os $(q-1)$ elementos finais da função escore de γ e seja $K^{\gamma\gamma}_{*(q-1)\times(q-1)}$ a matriz $(q-1)\times(q-1)$ formada pelas $(q-1)$ últimas linhas e as pelas $(q-1)$ últimas colunas da matriz $K(\beta,\gamma)^{-1}$. Então, a estatística escore para testar a hipótese em (B.0.1) é dada por

$$\omega = \widetilde{U}^\top_{(q-1)\gamma} \widetilde{K}^{\gamma\gamma}_{*(q-1)\times(q-1)} \widetilde{U}_{(q-1)\gamma},$$

em que til indica que as quantidades são avaliadas no estimador de máxima verossimilhança restrito. Sob as condições usuais de regularidade e sob $\mathcal{H}_0$, ω converge em distribuição para $\chi^2_{(q-1)}$.

Através de simulações de Monte Carlo nós avaliamos o desempenho do teste escore (tamanho e poder). Na geração dos dados no modelo de regressão beta, consideramos o seguinte modelo:

$$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_t \quad \text{e} \quad \text{logit}(\sigma_t) = \gamma_0 + \gamma_1 z_t, \tag{B.0.2}$$

$t = 1, \ldots, n$. Para o Modelo (B.0.2), os verdadeiros valores para os parâmetros são $\beta_0 = 1$, $\beta_1 = 1$ e $\gamma_0 = -1.5$. Nas avaliações de tamanho usamos $\gamma_1 = 0$ e nas simulações de poder foi usado $\gamma_1 = 2.0$. Para estas investigações geramos x_t e z_t aleatoriamente da distribuição $\mathcal{U}(0,1)$. Consideramos $n = 25, 50, 100$, sendo que são geradas 25 observações das variáveis x_t e z_t, que são replicadas duas e quatro vezes, respectivamente, para os tamanhos amostrais $n = 50, 100$. O nosso interesse consiste em testar $\mathcal{H}_0 : \gamma_1 = 0$ contra $\mathcal{H}_1 : \gamma_1 \neq 0$.

A Tabela B.1 apresenta as taxas de rejeição sob a hipótese nula do teste escore correspondentes aos níveis nominais de 10% e 5%. O número de réplicas de Monte Carlo é 10 mil. Observamos que as taxas de rejeição se aproximam dos respectivos níveis nominais à medida que o tamanho amostral aumenta. Por exemplo, quando $n = 25$ e $\alpha = 10\%$, o teste rejeita a hipótese nula 8.57% das vezes; todavia, quando $n = 100$, esta taxa passa ser 9.47%.

Além das simulações relativas ao tamanho do teste, fizemos simulações de poder para avaliar a capacidade do teste de conduzir a baixas probabilidades do erro do tipo II. Aqui, utilizamos o mesmo número de réplicas de Monte Carlo considerado anteriormente (dez mil). Os resultados da investigação do poder encontram-se na Tabela B.2. Foram usados valores críticos exatos (estimados nas simulações de tamanho). A hipótese nula em teste é $\mathcal{H}_0 : \gamma_1 = 0$ e os dados foram gerados utilizando $\text{logit}(\sigma_t) = \gamma_0 + \gamma_1 z_t$, em que $\gamma_0 = -1.5$ e $\gamma_1 = 2.0$. Os resultados obtidos sob a hipótese alternativa revelam que o poder do teste aumenta à medida que o tamanho amostral cresce, como esperado.

Tabela B.1 Tamanho do teste escore.

$n = 25$		$n = 50$		$n = 100$	
10%	5%	10%	5%	10%	5%
8.57	3.91	9.34	4.54	9.47	4.69

Tabela B.2 Poder do teste escore.

$n = 25$		$n = 50$		$n = 100$	
10%	5%	10%	5%	10%	5%
65.34	57.05	94.59	92.11	99.98	99.95

Referências Bibliográficas

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

[2] Argyle, M. *Religious Behaviour.* London: Routledge and Kegan Paul, 1958.

[3] Canêdo-Pinheiro, M. Bolsa família ou desempenho da economia? Determinantes da reeleição de Lula em 2006. In *Anais do XXXVII Encontro Nacional de Economia* (2009).

[4] Chesher, A.; Jewitt, I. The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica 55* (1987), 1217–1222.

[5] Cook, R. Assessment of local influence (with discussion). *Journal of the Royal Statistical Society B 48* (1968), 133–169.

[6] Cook, R. Detection of influential observations in linear regression. *Technometrics 19* (1977), 15–18.

[7] Cribari-Neto, F. Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis 45* (2004), 215–233.

[8] Cribari-Neto, F.; Zarkos, S. Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing. *Econometric Reviews 18* (1999), 211–228.

[9] Cribari-Neto, F.; Zarkos, S. Heteroskedasticity-consistent covariance matrix estimation: White's estimator and the bootstrap. *Journal of Statistical Computation and Simulation 68* (2001), 391–411.

[10] Cribari-Neto, F.; Zeleis, A. Beta regression in R. *Journal of Statistical Software 34(2)* (2010), 1–24.

[11] Davidson, R.; MacKinnon, J. *Estimation and Inference in Econometrics*. Oxford University Press, New York, 1993.

[12] Doornik, J. *Ox: An Object-Oriented Matrix Language*. Timberlake Consultants & Oxford, London, 2006. `http://www.doornik.com`.

[13] Eicker, F. Asymptotic normality and consistency of the least squares estimator for families of linear regressions. *Annals of Mathematical Statistics 34* (1963), 447–465.

[14] Espinheira, P. L. *Regressão Beta*. Tese de doutorado, Departamento de Estatística, Universidade de São Paulo, 2007.

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

[16] Espinheira, P. L.; Ferrari, S. L.; Cribari-Neto, F. Influence diagnostics in beta regression. *Computational Statistics and Data Analysis 52* (2008), 4417–4431.

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

[18] Furno, M. Small sample behavior of a robust heteroskedasticity consistent covariance matrix estimator. *Journal of Statistical Computation and Simulation 54* (1996), 115–128.

[19] Greeley, A. *Religion in Europe at the end of the Second Millennium*. New Brunswick, NJ: Transaction Publishers, 2003.

[20] Greene, W. H. *Econometric Analysis*. Prentice Hall, New York, 2003.

[21] Harrell, F. J. *Regression Modeling Strategies*. New York: Springer, 2001.

[22] Horn, S. D.; Horn, R.; Duncan, D. Estimating heteroskedastic variances in linear models. *Journal of the American Statistical Association 70* (1975), 380–385.

[23] Howells, T. H. A comparataive study of those who accept as against those who reject religious authority. *University of Iowa Studies of Character 2* (1928).

[24] Huber, P. J. The behavior of maximum likelihood estimation under non-standard conditions. *In Proceedings of the Fifth Berkeley Symposium on Mathematics Statistics and Probability* (1967), 221–233.

[25] Hunter, W.; Power, T. J. Rewarding Lula: executive power, social policy, and the brazilian elections of 2006 Lula. *Latin American Politics and Society 49* (2007), 1–30.

[26] Judge, G. C.; Hill, R. C.; Griffiths, W. E.; Lutkepohl, H.; Lee, T. C. *The Theory and Pratice of Econometrics*. New York: Wiley, 1985.

[27] Kanazawa, S. Iq and the values of nations. *Journal of Biosocial Science 41* (2009), 537–556.

[28] Kieschnick, R.; McCullough, B. D. Regression analysis of variates observed on $(0, 1)$ percentages, proportions and fractions. *Statistical Modelling 3* (2003), 193–213.

[29] Kilsztajn, S. Paridade de poder de compra, renda per capita e outros indicadores econômicos. *Pesquisa & Debate 11* (2000), 93–102.

[30] Kleiber, C.; Zeileis, A. *Applied Econometrics with R*. New York: Springer, 2008.

[31] Koenker, R. A note on studentizing a test for heteroscedasticity. *Journal of Econometrics 17* (1981), 107–112.

[32] Lima, V.; Souza, T.; Cribari-Neto, F.; Fernandes, G. Heteroskedasticity-robust inference in linear regressions. *Simulation and Computation 39* (2010), 194–206.

[33] Long, J.; Ervin, L. Using heteroskedasticity-consistent standard errors in the linear regression model. *American Statistician 54* (2000), 217–224.

[34] Lynn, R.; Mikk, J. National differences in intelligence and educational attainment. *Intelligence 35* (2007), 115–121.

[35] Lynn, R.; Harvey, J. N. H. Average intelligence predicts atheism rates across 137 nations. *Intelligence 37* (2009), 11–15.

[36] Lynn, R.; Vanhanen, T. *IQ and Global Inequality*. Washington Summit Publishers, 2006.

[37] MacKinnon, J. G.; White, H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite-sample properties. *Journal of Econometrics 29* (1985), 305–325.

[38] Marques, R. M.; Leite, M.; Mendes, A.; Ribeiro, J. Discutindo o papel do programa bolsa família na decisão das eleições presidenciais brasileiras de 2006. *Revista de Economia Política 29* (2009), 114–132.

[39] McCullagh, P.; Nelder, J. *Generalized Linear Models*. Chapman and Hall, 1989.

[40] Myers, R. H. *Classical and Modern Regression with Applications*. Duxbury Press, Belmont, 1990.

[41] Ospina, R.; Cribari-Neto, F.; Vasconcellos, K. L. P. Improved point and interval estimation for a beta regression model. *Computational Statistics and Data Analysis 51* (2006), 960–981.

[42] Paolino, P. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis 9* (2001), 325–346.

[43] Rocha, A. V.; Simas, A. B. Influence diagnostics in a general class of beta regression models. *TEST* (2010), 1–25.

[44] Rousseeuw, P.; Hubert, M. Recent developments in PROGRESS. In *L1-Statistical Procedures and Related Topics*, vol. 31. Institute of Mathematical Statistics Lecture Notes-Monograph Series, Hayward, California, 1997, pp. 201–214.

[45] Rousseeuw, P. J. Least median of squares regression. *Journal of American Statistical Association 79* (1984), 871–880.

[46] Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[47] Rushton, J. P.; Templer, D. I. National differences in intelligence, crime, income, and skin color. *Intelligence 37* (2009), 341–346.

[48] Shikida, C. D.; Monasterio, L. M.; Araujo, A. F.; Carraro, A.; Damião, O. M. It is the economy, companheiro!": an empirical analysis of Lula's re-election based on municipal data. *Economics Bulletin 29 (2)* (2009), 976–991.

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

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

[51] Souza, T. C. Inferência em modelos heteroscedásticos na presença de pontos de alavanca. Dissertação de mestrado, Departamento de Estatística, Universidade Federal de Pernambuco, Brazil, 2003.

[52] Stasinopoulos, D. M.; Rigby, R. A.; Akantziliotou, C. *Instructions on how to use the GAMLSS package in R. Accompanying documentation in the current GAMLSS help files*. (http://www.gamlss.com), 2006.

[53] White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica 48* (1980), 817–838.

[54] White, H. *Estimation, Inference and Specification Analysis*. Cambridge University Press, 1994.

[55] Zeleis, A. Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software 11* (2004), 1–17.

[56] Zeleis, A. Object-oriented computation of sandwich estimators. *Journal of Statistical Software 16* (2006), 1–16.

[57] Zucco, C. The president's 'new' constituency: Lula and the pragmatic vote in Brazil's 2006 presidential elections. *Journal of Latin American Studies 40* (2008), 29–39.

[58] Zuckerman, P. *Atheism: Contemporay Numbers and Patterns*. University of Cambridge Press, 2007.