



**UNIVERSIDADE FEDERAL DO PARÁ**  
**INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS**  
**PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA E ESTATÍSTICA**

**Charlene de Carvalho Silva**

**Estudo e Aplicações do Modelo de Regressão Beta**

**Orientadora: Profa. Dra. Maria Regina Madruga Tavares**

Belém-PA  
2013

**Charlene de Carvalho Silva**

**Estudo e Aplicações do Modelo de Regressão Beta**

Dissertação apresentada ao Programa de Pós-Graduação em Matemática e Estatística, da Universidade Federal do Pará, como requisito parcial para a obtenção do grau de Mestre em Estatística.

**Orientadora: Profa. Dra. Maria Regina Madruga Tavares**

Belém-PA

2013

Dados Internacionais de Catalogação-na-Publicação (CIP)  
Sistema de Bibliotecas da UFPA

---

Silva, Charlene de Carvalho, 1982-  
Estudo e aplicações do modelo de regressão  
beta / Charlene de Carvalho Silva. - 2013.

Orientadora: Maria Regina Madruga.  
Dissertação (Mestrado) - Universidade Federal  
do Pará, Instituto de Ciências Exatas e  
Naturais, Programa de Pós-Graduação em  
Matemática e Estatística, Belém, 2013.

1. Análise de regressão. 2. Estatística. 3.  
Rendimento escolar-Avaliação-Educação  
(Superior). I. Título.

CDD 22. ed. 519.536

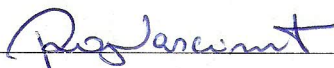
---

Charlene de Carvalho Silva

## Estudo e Aplicações do Modelo de Regressão Beta

Esta Dissertação foi julgada e aprovada, para a obtenção do grau de Mestre em Estatística, no Programa de Pós-Graduação em Matemática e Estatística, da Universidade Federal do Pará.

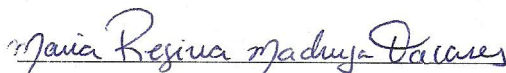
Belém, 08 de Julho de 2013.



Prof. Dra. Rúbia Gonçalves Nascimento

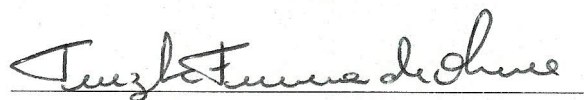
Coordenador do Programa de Pós-Graduação em Matemática e Estatística da UFPA

### Banca Examinadora



Prof. Dra. Maria Regina Madruga Tavares  
Universidade Federal do Pará

**Orientador**



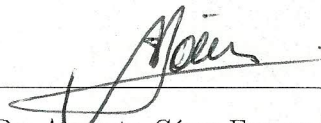
Prof. Dra. Terezinha Ferreira de Oliveira  
Universidade Federal do Pará

**Membro**



Prof. Dr. Héilton Ribeiro Tavares  
Universidade Federal do Pará

**Membro**



Prof. Dr. Augusto César Fonseca Saraiva  
Eletronorte

**Membro**

*Dedico a minha  
filha Yara Caroline.*

# Agradecimentos

- ★ A Deus por sempre me dá força, iluminar o meu caminho e me ajudar a seguir, e lutar por meus sonhos. E colocar anjos no meu caminho que me ajudaram a finalizar mais uma etapa em minha vida.
- ★ Aos meus familiares pelo apoio nos momentos oportunos. Em especial ao meu PAI que apesar de hoje está distante, sempre me deu força e acreditou em tudo que faço é um dos grandes responsáveis por esta conquista. A minha Mãe mulher de braço forte e mãos calejadas que sempre fez de tudo para nos criar.
- ★ Ao meu marido DIOGO MENDES por toda força, dedicação e por me fazer acreditar que tudo posso.
- ★ Em especial a minha orientadora Profa. Dra. Regina por todo incentivo, paciência e apoio neste trabalho.
- ★ Ao Conexões de Saberes, representado na pessoa da Profa. Dra. Maria José Avis programa que mudou minha perspectiva de vida. Ao PROJETO RENASCER (*in memoriam*) literalmente me fez renascer e realizar um sonho.
- ★ A Universidade Federal do Pará (UFPA) pela oportunidade de concluir um curso de pós-graduação.
- ★ Aos participantes da banca examinadora pelas sugestões;
- ★ À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro;

*Que Deus pode fazer*  
*Infinitamente mais*  
*Deus do impossível*  
*Tudo é possível basta crer*  
*Thalles Roberto*

---

## Resumo

---

SILVA, Charlene Carvalho. Estudo e Aplicações do Modelo de Regressão Beta. 2013. Dissertação (Mestrado em Matemática e Estatística), PPGME, UFPA, Belém, Pará, Brasil.

Neste trabalho são apresentadas algumas características e propriedades do Modelo de Regressão Beta, para o caso simples de uma única observação por indivíduo, e para o caso em que tem-se dados com medidas repetidas. Este modelo de regressão é de grande utilidade no caso em que tem-se uma variável resposta definida no intervalo  $(0, 1)$ , como por exemplo, taxas ou proporções. O objetivo desta dissertação é estudar o processo de estimação do modelo de regressão Beta e aplicá-lo a dois conjuntos de dados reais. O modelo de regressão Beta, definido por Ferrari e Cribari-Neto (2004) foi utilizado na modelagem dos dados referente as características físico-químicas e de cromatografia gasosa utilizadas para monitorar a vida útil dos transformadores, onde constatou-se que as variáveis *Cor*, *Rigidez Dielétrica*, *Teor Água*, *Tensão Interfacial*,  $O_2$ ,  $N_2$ ,  $CH_4$ ,  $CO$  e  $C_2H_4$  são estatisticamente importantes para explicar o comportamento da variável resposta Índice de neutralização. Foi observado que este tem maior chance de aumento para a variação positiva de 1 unidade com a variável *Cor* (19,73%). O modelo de regressão Beta para dados com medidas repetidas, desenvolvido por Venezuela (2008), foi utilizado para avaliar e comparar o rendimento acadêmico dos cotistas e não cotistas no curso de graduação em Medicina da UFPA, campus Belém, que ingressaram pelo primeiro Processo seletivo seriado com sistema de cotas. Verificou-se que o desempenho no vestibular influencia no rendimento acadêmico do aluno na universidade e que, apesar do curso de Medicina ser de alto prestígio e possuir demanda elevada no vestibular, o rendimento acadêmico dos cotistas e não-cotistas não difere ao longo dos nove períodos de curso observados.

**Palavras-chave:** regressão beta, índice de neutralização, rendimento acadêmico.



---

# Abstract

---

SILVA , Charlene de Carvalho. Beta Regression and applications. 2013. Dissertation (Master Science in Mathematics and Statistics), PPGME, UFPA, Belém, Pará, Brazil.

This work presents some characteristics and properties of Beta Regression Model for the simple case of a single observation per individual, and the event that has data with repeated measurements. This regression model is useful in the case that it has a variable response defined in the interval  $(0, 1)$ , such as rates or proportions. The objective of this thesis is to study the process of estimating the regression model Beta and apply it to two real data sets. Beta regression model, defined by Ferrari and Cribari-Neto (2004) was used in modeling data regarding the physico-chemical and gas chromatography used to monitor the life of transformers, where it was found that the variables *Color*, *Dielectric Strength*, *Water Content*, *Interfacial Tension*,  $O_2$ ,  $N_2$ ,  $CH_4$ ,  $CO$  and  $C_2H_4$  are statistically important in explaining the behavior of the response variable Neutralization index. It was noted that this is more likely to increase for the positive change of one unit with variable color (19,73%). Beta regression model for data with repeated measures, developed by Venezuela (2008), was used to evaluate and compare the academic performance of shareholders and non-shareholders in the course of undergraduate medical UFPA, Belém campus, who entered the first Selection Process Series with a quota system. It was found that the performance in the entrance exam influences the academic performance of the student at the university and that despite the medical school have high prestige and have high demand in the entrance exam, the academic performance of shareholders and non-shareholders do not differ over the nine periods course observed.

**Keywords:** Beta regression, Neutralization index, Academic performance.

---

# Índice

---

<b>Resumo</b>	<b>vii</b>
<b>Abstract</b>	<b>viii</b>
<b>Lista de Tabelas</b>	<b>xii</b>
<b>Lista de Figuras</b>	<b>xiv</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Aspectos Gerais . . . . .	1
1.2 Objetivos . . . . .	3
1.3 Organização da Dissertação . . . . .	3
<b>2 Modelo de Regressão Beta</b>	<b>5</b>
2.1 Introdução . . . . .	5
2.2 Distribuição Beta . . . . .	5
2.3 Modelo de Regressão Beta . . . . .	6
2.3.1 Reparametrização da distribuição Beta . . . . .	6
2.3.2 Regressão Beta . . . . .	7
2.4 Funções de ligação . . . . .	8
2.5 Função Escore e Matriz de Informação . . . . .	11
2.6 Estimação dos Parâmetros . . . . .	15
2.6.1 Processo Iterativo de Newton-Raphson . . . . .	16
2.6.2 Processo Iterativo de Escore de Fisher . . . . .	16

2.6.3	Método BFGS . . . . .	16
2.7	Teste de Hipóteses . . . . .	18
2.7.1	Teste de Wald . . . . .	18
2.8	Resíduos . . . . .	18
2.8.1	Resíduo Ordinário . . . . .	19
2.8.2	Resíduo ponderado padronizado 1 . . . . .	19
2.8.3	Resíduo ponderado padronizado 2 . . . . .	20
2.9	Diagnóstico do Modelo ajustado . . . . .	20
2.9.1	Ponto Alavanca ou Alavanca Generalizada . . . . .	21
2.9.2	Distância de Cook . . . . .	23
2.9.3	Gráfico de Probabilidade Meio-Normal com Envelope Simulado . . . . .	24
<b>3</b>	<b>Regressão Beta com medidas repetidas</b>	<b>25</b>
3.1	Introdução . . . . .	25
3.2	Equações de Estimação . . . . .	26
3.2.1	Modelagem do parâmetro de posição . . . . .	26
3.2.2	Matrizes de correlação de Trabalho . . . . .	30
3.3	Técnicas de diagnóstico . . . . .	33
3.3.1	Pontos alavanca, influentes e aberrantes . . . . .	34
3.3.2	Influência local . . . . .	35
3.4	Seleção de modelos . . . . .	36
<b>4</b>	<b>Aplicação I: Índice de Neutralização em Transformadores</b>	<b>37</b>
4.1	Introdução . . . . .	37
4.2	Sistema Isolante em Transformador . . . . .	38
4.2.1	Isolação Sólida . . . . .	38
4.2.2	Isolação Líquida . . . . .	38
4.3	Características Físico-Químicas para Monitorar a Condição do OMI . . . . .	39
4.3.1	Cor . . . . .	39
4.3.2	Tensão Interfacial . . . . .	39
4.3.3	Teor de Água . . . . .	40

4.3.4 Índice de Neutralização . . . . .	40
4.3.5 Rigidez Dielétrica . . . . .	40
4.3.6 Fator de Potência . . . . .	41
4.3.7 Densidade . . . . .	41
4.4 Cromatografia Gasosa . . . . .	41
4.5 Descrição e estudo exploratório dos dados . . . . .	42
4.6 Ajuste do modelo de Regressão Beta . . . . .	45
<b>5 Aplicação II: Rendimento acadêmico no Sistema de Cotas</b>	<b>60</b>
5.1 Introdução . . . . .	60
5.2 Implantação da política de cotas no Ensino Superior Brasileiro . . . . .	61
5.2.1 Implantação da política de cotas: o caso da UFPA . . . . .	61
5.2.2 Avaliação do Sistema de Cotas . . . . .	62
5.3 Descrição e estudo exploratório dos dados . . . . .	62
5.4 Ajuste do modelo de Regressão Beta com medidas repetidas . . . . .	66
<b>6 Considerações Finais e Recomendações</b>	<b>71</b>
6.1 Conclusões . . . . .	71
6.2 Recomendações . . . . .	72
<b>Referências Bibliográficas</b>	<b>73</b>

---

## Lista de Tabelas

---

4.1	Medidas descritivas das características físico-químicas e da cromatografia .	44
4.2	Estimativas dos parâmetros do modelo de regressão Beta . . . . .	46
4.3	Medida de impacto DQMRP e EAMRP para os valores ajustados . . . . .	50
4.4	Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem a observação 3 . . . . .	51
4.5	Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem a observação 154 . . . . .	52
4.6	Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem as observações 3 e 154 . . . . .	53
4.7	Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem a observação 159 . . . . .	54
4.8	Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem as observações 3 e 159 . . . . .	55
4.9	Erro Quadrático Médio (EQM) e Erro Absoluto Médio (EAM) do modelo de regressão Beta ajustado sem os pontos identificados no diagnóstico de influência . . . . .	57
4.10	Estimativas dos parâmetros do modelo de regressão Beta somente com as variáveis significativas após a análise de diagnóstico. . . . .	59
5.1	Estimativas e erros padrões dos parâmetros do modelo de regressão Beta com a estrutura Padrão Uniforme. Dados do rendimento acadêmico dos alunos do curso de medicina da UFPA. . . . .	68

5.2	Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem as observações (1, 2, 17, 47, 65, 82, 88, 113 e 126). Dados do rendimento acadêmico dos alunos do curso de medicina da UFPA. . . . .	70
-----	--	----

---

## Lista de Figuras

---

2.1	Densidade Beta para diferentes valores de $(p,q)$ . . . . .	7
2.2	Densidade Beta para diferentes valores de $(\mu, \phi)$ . . . . .	8
2.3	Funções de Ligação logito, log-log, clog-log e Aranda-ordaz. . . . .	10
4.1	Forma simplificada da distribuição de energia . . . . .	37
4.2	Distribuição do índice de neutralização geral e por tipo de transformador. . . . .	45
4.3	Gráficos de diagnóstico do modelo de regressão Beta ajustado para os dados de índice de neutralização. . . . .	47
4.4	Gráfico de Probabilidade Meio-Normal com Envelope Simulado. . . . .	48
4.5	Gráfico de Probabilidade Meio-Normal com Envelope Simulado . . . . .	56
4.6	Índice de neutralização observado e previsto na amostra de validação . . . . .	58
5.1	Box-plot para o CRPL por período letivo . . . . .	64
5.2	CRPL médio para não cotistas e cotistas no curso de Medicina por período letivo . . . . .	64
5.3	Correlações entre os CRPL's dos não cotistas, iniciando no período 1 (P1) ao período 5 (P5) em relação aos períodos seguintes . . . . .	65
5.4	Correlações entre os CRPL's dos cotistas, iniciando no período 1 (P1) ao período 5 (P5) em relação aos períodos seguintes . . . . .	65
5.5	Gráficos de diagnósticos com todas as observações. . . . .	69

---

## Capítulo 1

# Introdução

---

### 1.1 Aspectos Gerais

Usualmente, em diferentes áreas de aplicação da Estatística, utiliza-se o modelo de regressão Normal Linear para investigar a relação entre uma variável de interesse e um conjunto de variáveis explicativas, que afetam seu comportamento através de uma estrutura de regressão. No entanto, quando essa variável de interesse assume valores em um intervalo  $(0,1)$  restrito, tais como proporções, taxas, índices, concentrações, e outros, o uso do modelo de regressão Normal Linear não é adequado, uma vez que este modelo pressupõe que os dados sejam distribuídos normalmente, o que acaba sendo um modelo falho, pois possibilita a previsão de valores fora do suporte da variável resposta, ou seja, do intervalo  $(0,1)$ .

Na literatura encontramos algumas aplicações da distribuição beta como Janardan e Padmanabhan (1986) modelam variáveis hidrológicas usando a distribuição beta; McNally (1990) utiliza a distribuição beta num estudo referente à capacidade de reprodução em vacas; raham e Hollands (1990) e Milyutin e Yaromenko (1991) usam a distribuição beta para estudar índices relacionados a transmissão de radiação solar; Bury (1999) lista um conjunto de aplicações em engenharia. Contudo essas aplicações não envolvem situações em que o pesquisador necessita impor uma estrutura de regressão para a variável resposta. Segundo Kieschnick e McCullough (2003) e Johnson, Kotz e Balakrishnam (1995,p.235) destacam que a distribuição beta é um modelo apropriado para descrever dados distribuídos de forma contínua no intervalo  $(0,1)$  dada a grande flexibilidade de formas dessa distribuição.

Vários modelos têm sido propostos na literatura para trabalhar com esta distribuição, em uma estrutura de regressão. Segundo Smithson e Verkuilen (2006), as três primeiras aplicações de Regressão Beta vieram da literatura na área de Economia, onde Brehm e



Gates (1993) modelaram proporções usando a parametrização padrão das distribuições Beta, que complexifica a formulação do modelo de regressão dificultando a interpretação dos parâmetros. Posteriormente, Paolino (2001) comparou a Regressão Beta com modelos de regressão linear com e sem transformação logito da variável resposta. Utilizando dados simulados e reais, ele concluiu que a Regressão Beta apresenta vantagens significativas em relação à regressão linear para o estudo de variáveis respostas que assumem valores no intervalo (0,1). Kieschnick e McCullough (2003) compararam o desempenho de um modelo de Regressão Beta com ligação logito com as de diversos outros modelos, usualmente utilizados na prática a partir de uma variação do critério de informação de Akaike, proposta por McQuarrie e Tsai (1998). Em ambos os conjuntos de dados referentes a estudos de proporções, o modelo de Regressão Beta produziu um ajuste superior aos demais, segundo esse critério. Ferrari e Cribari-Neto (2004) propuseram um modelo de Regressão Beta usando uma parametrização da distribuição Beta indexada pelos parâmetros de locação e dispersão. Estes parâmetros podem ser interpretados em termos da média das observações, que é modelada usando um preditor linear que relaciona a resposta média a covariáveis e parâmetros desconhecidos através de uma função de ligação. Além de apresentar detalhes da estimação dos parâmetros por máxima verossimilhança, também desenvolveram um conjunto compreensivo de ferramentas de inferência, tais como testes de hipóteses e medidas de diagnóstico.

Após o artigo de Ferrari e Cribari-Neto (2004) vários trabalhos relacionados ao modelo de Regressão Beta foram desenvolvidos e um pacote no software R foi criado para o ajuste do modelo (CRIBARI-NETO e ZEILEIS, 2010). O modelo de Regressão Beta foi estendido por Smithson e Verkuilen (2006), para permitir que o parâmetro de precisão da distribuição Beta também varie em função de variáveis preditoras. Ospina *et al.* (2006) obtiveram estimadores de máxima verossimilhança dos parâmetros do modelo de Regressão Beta com correção do viés. Cribari-Neto e Vasconcellos (2006) (ver também OLIVEIRA, 2004) através de estudos de simulações observaram bons resultados quanto ao viés, comportamento assintótico, taxa de cobertura e assimetria para os estimadores dos parâmetros de regressão, e os resultados não foram tão satisfatórios para o parâmetro de dispersão. Ospina (2007) propôs o modelo de Regressão Beta com dispersão variável quando há independência entre todas as observações e desenvolveu medidas de diagnóstico

capazes de captar a presença dessa dispersão variável nos dados. Andrade (2007) estudou o efeito da especificação incorreta da função de ligação em modelos de Regressão Beta. O modelo de Regressão Beta foi ainda desenvolvido do ponto de vista bayesiano por Brascum *et al.* (2007). Miyashiro (2008) comparou os modelos de Regressão Beta e simplex em dois conjuntos de dados, observando que em ambos a Regressão Beta possibilitou um melhor ajuste. Já Espinheira *et al.* (2008 a,b) introduziram novos resíduos e medidas de influência local. Venezuela (2008) desenvolveu equações de estimação para os modelos de Regressão Beta com medidas repetidas e também desenvolveu algumas medidas de diagnóstico para avaliar a adequação do modelo. Souza (2011) e Melo *et al.* (2009) propuseram modelos de Regressão Beta multivariados. Simas *et al.* (2010) desenvolveram um modelo de Regressão Beta mais geral que permite não linearidade e dispersão variável. Cribari-Neto e Souza (2011) desenvolveram testes de hipóteses para os parâmetros que ajustam a média da variável resposta sem a necessidade de ajustar o parâmetro de precisão.

## 1.2 Objetivos

Os objetivos deste trabalho são relacionados a seguir:

- Estudar o processo de estimação no modelo de Regressão Beta;
- Estudar as principais medidas e critérios de diagnóstico do modelo ajustado;
- Realizar a aplicação do modelo de Regressão Beta em dois conjuntos de dados reais utilizando o modelo de Regressão Beta definido por Ferrari e Cribari-Neto (2004) e o modelo de Regressão Beta desenvolvido para dados com medidas repetidas, desenvolvido por Venezuela (2008).

## 1.3 Organização da Dissertação

A presente dissertação de mestrado está dividida em seis capítulos. No Capítulo 2 apresentamos o modelo de Regressão Beta proposto por Ferrari e Cribari-Neto (2004), algumas funções de ligação, a função escore e matriz de informação, a estimação dos parâmetros de regressão, teste de hipóteses e algumas medidas de diagnóstico. No Capítulo 3 apresentamos o modelo de Regressão Beta desenvolvido para dados com medidas repetidas, desen-

volvido por Venezuela (2008). No Capítulo 4 apresentaremos os resultados da Aplicação I, utilizando o modelo de Regressão Beta proposto por Ferrari e Cribari-Neto (2004) a um conjunto de dados reais provenientes de um estudo sobre as características físico-químicas e análise cromatográfica de gases, utilizada para monitorar a condição do óleo isolante em alguns transformadores. No Capítulo 5 apresentaremos a Aplicação 2, com a utilização do modelo de Regressão Beta para dados com medidas repetidas (VENEZUELA, 2008), com base em um conjunto de dados relativo ao acompanhamento do Coeficiente de Rendimento do Período Letivo (CRPL) de alunos ingressantes em alguns cursos da UFPA através do sistema de cotas. No Capítulo 6 apresentamos as conclusões e recomendações para trabalhos futuros.

---

## Capítulo 2

# Modelo de Regressão Beta

---

### 2.1 Introdução

O modelo de Regressão Beta desenvolvido por Ferrari e Cribari-Neto (2004) é uma ferramenta que vem sendo muito utilizada na modelagem de dados onde a variável de interesse é restrita no intervalo unitário  $(0,1)$ . O modelo proposto apresenta uma reparametrização da distribuição Beta, de modo que sua estrutura de regressão permite modelar a média de uma variável dependente em função de um conjunto de variáveis independentes e que contenha um parâmetro de precisão (ou dispersão) fixo ou variável. Assim, o modelo supõe que a distribuição da variável resposta é Beta e que a resposta média está relacionada com um preditor linear a partir de uma função de ligação. Esta apresenta vantagens sobre o modelo linear clássico, uma vez que a densidade pode assumir diferentes formas, tendo inclusive, a distribuição uniforme como um caso particular (VASCONCELLOS e CRIBARI-NETO, 2005).

### 2.2 Distribuição Beta

A família de distribuições Beta é composta de todas as distribuições com função de densidade de probabilidade (f.d.p) da forma

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(q)\Gamma(p)} y^{p-1} (1-y)^{q-1}, 0 < y < 1, p > 0, q > 0, \quad (2.1)$$

em que  $\Gamma(q)$  é a função gama definida como

$$\Gamma(p) = \int_0^{\infty} y^{p-1} e^{-y} dy. \quad (2.2)$$

A média e a variância da variável aleatória (v.a.)  $Y$  com f.d.p. dada em (2.1), são dadas, respectivamente, por

$$\mathbb{E}(Y) = \frac{p}{p+q}, \quad (2.3)$$

$$Var(Y) = \frac{pq}{(p+q)^2(p+q+1)} \quad (2.4)$$

Como a distribuição Beta possui dois parâmetros de forma, existe muita flexibilidade no formato de sua densidade.

De acordo com os valores de  $p$  e  $q$ , a densidade Beta pode assumir diferentes formas no intervalo  $(0, 1)$ . A Figura 2.1 apresenta funções densidade de probabilidade para alguns valores de  $(p, q)$ . Note que para  $p, q < 1$  a densidade tem forma de "U", quando  $p < q$  há assimetria à esquerda e quando  $p > q$  há assimetria à direita. Quando  $p = q = 1$ , obtém-se a distribuição uniforme padrão.

## 2.3 Modelo de Regressão Beta

### 2.3.1 Reparametrização da distribuição Beta

Ferrari e Cribari-Neto (2004) ressaltam a conveniência de reparametrizar (2.1) de modo que os parâmetros representem a média e a precisão da variável resposta. Para obter esta estrutura, consideram a transformação  $\mu = p/(p+q)$  e  $\phi = (p+q)$ , implicando em  $p = \mu\phi$  e  $q = (1-\mu)\phi$ . Assim, obtém-se uma nova forma para a função densidade de probabilidade para a família de distribuições Beta, dada por

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, 0 < y < 1 \quad (2.5)$$

em que  $0 < \mu < 1, \phi > 0$ . A partir de (2.3) e (2.4) segue que

$$\mathbb{E}(Y) = \mu \text{ e } Var(Y) = \frac{V(\mu)}{1+\phi},$$

em que  $V(\mu) = \mu(1-\mu)$ , de tal forma que  $\mu$  é a média da variável  $Y$  e  $\phi$  pode ser interpretado como um parâmetro de precisão no sentido que, para  $\mu$  fixado, tanto maior o valor de  $\phi$ , menor a variância de  $Y$ .

Neste modelo (2.5) é assumido que a variável dependente está restrita ao intervalo unitário  $(0, 1)$ . No entanto, este modelo pode ser utilizado em situações mais gerais, em que

a variável dependente é restrita ao intervalo  $(a, b)$ , onde  $a$  e  $b$  são constantes conhecidas, com  $a < b$ . Neste caso, em vez de modelar  $Y$  diretamente, utiliza-se a transformação  $Y^* = (Y - a)/(b - a)$  que ficará então definido no intervalo  $(0, 1)$ .

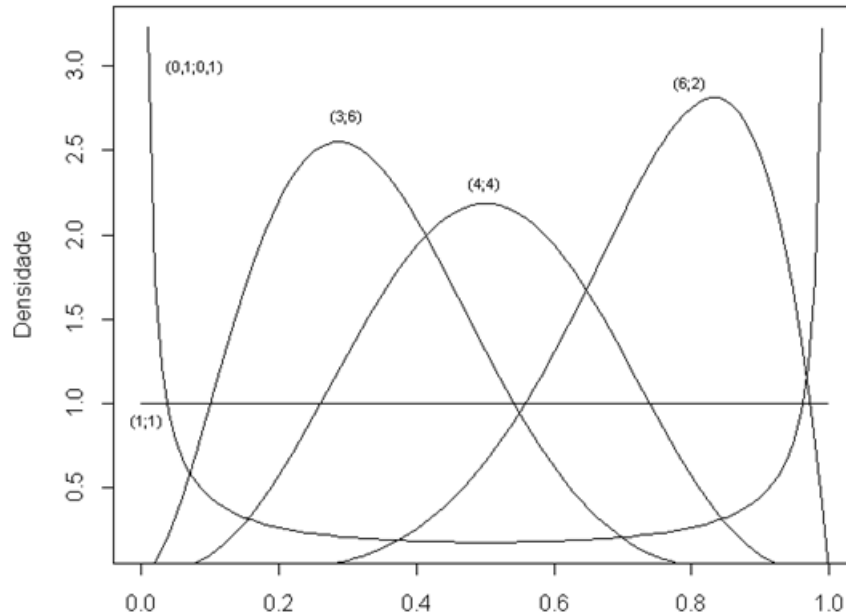


Figura 2.1 Densidade Beta para diferentes valores de  $(p, q)$

Na Figura 2.2 são apresentadas densidades Beta considerando a nova parametrização. Como se pode observar, a densidade pode assumir forma de "J", "J" invertido e de "U". Adicionalmente, ela é simétrica quando  $\mu = 1/2$  e assimétrica quando  $\mu \neq 1/2$ . Em particular, para  $\mu = 1/2$  e  $\phi = 2$  a densidade (2.5) se reduz a uniforme padrão. (FERRARI e CRIBARI-NETO, 2004).

### 2.3.2 Regressão Beta

Sejam  $Y_1, Y_2, \dots, Y_n$  variáveis aleatórias independentes, de modo que cada  $Y_t$  tenha distribuição Beta reparametrizada com densidade (2.2), isto é, com média  $\mu_t$ ,  $t = 1, \dots, n$ , e parâmetro de precisão  $\phi$  (desconhecido e constante para todo  $t$ ). Ferrari e Cribari-Neto (2004) propuseram a modelagem da média  $\mu_t$  por meio de uma função de ligação associada ao preditor linear  $\eta_t$ , de modo a satisfazer a seguinte relação funcional:

$$g(\mu_t) = \sum_{i=1}^k x_{ti} \beta_i = \eta_t, \quad (2.6)$$

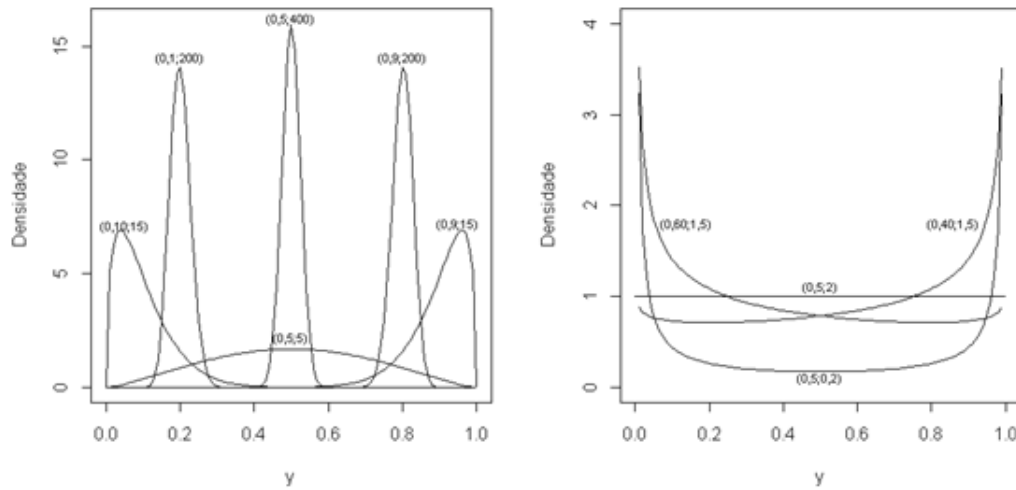


Figura 2.2 Densidade Beta para diferentes valores de  $(\mu, \phi)$ .

em que  $\eta_t = x_t^\top \beta$ , sendo  $\beta = (\beta_1, \beta_2, \dots, \beta_k)^\top$  o vetor de parâmetros desconhecidos a serem estimados, ( $\beta \in \mathfrak{R}^T$ ),  $x_t^\top = (x_{t1}, x_{t2}, \dots, x_{tk})$  representa os valores de  $k$  ( $k < n$ ) variáveis explicativas (covariáveis) que são assumidos fixos e conhecidos e  $g(\cdot)$  é uma função estritamente monótona e duplamente diferenciável (função de ligação) com domínio no intervalo  $(0, 1)$  e contradomínio  $\mathfrak{R}$ . Desta forma, tem-se que  $\mu_t = g^{-1}(\eta_t)$  e  $var(Y_t) = \mu_t(1 - \mu_t)/(1 + \phi)$ . Assim, como a variância da resposta  $Y_t$  depende de  $\eta_t$  (pela função de variância e pela inversa da função de ligação) e, portanto, não é constante para todo  $t$ , as variáveis respostas com variância não constantes (heterocedásticas) são naturalmente incorporadas no modelo.

## 2.4 Funções de ligação

O Modelo de Regressão Beta apresenta uma função que relaciona a variável dependente e a estrutura linear das variáveis independentes, denominado função de ligação  $g(\cdot)$ , e existem muitas possibilidades para sua escolha. Por exemplo, podemos utilizar a especificação logito

$$g(\mu) = \log \left\{ \frac{\mu}{1 - \mu} \right\},$$

ou a função probito

$$g(\mu) = \Phi^{-1}(\mu),$$

em que  $\Phi(\cdot)$  é a função de distribuição acumulada de uma variável aleatória Normal padrão, ou ainda, a ligação complemento *log-log*

$$g(\mu) = \log\{-\log(1 - \mu)\},$$

ou a função *log-log*

$$g(\mu) = -\log\{-\log(\mu)\},$$

ou a função Aranda-Ordaz (assimétrica) (ARANDA-ORDAZ, 1981):

$$g(\mu, \lambda) = \log\left\{(1 - \mu)^{-\lambda} - \frac{1}{\lambda}\right\}.$$

Essas cinco funções de ligação apresentadas são contínuas e estritamente crescentes no intervalo unitário  $[0, 1]$ . Na prática, é comum realizar a escolha da função de ligação que melhor ajuste a relação da estrutura linear (preditor linear) e a média da distribuição da variável dependente. De acordo com Ferrari e Cribari-Neto (2004) as funções de ligação podem assumir comportamento similar no domínio do intervalo padrão unitário.

A Figura 2.3 apresenta o comportamento de  $\eta$  como função de  $\mu$  para as funções de ligação logito, probito, complemento log-log, log-log e Aranda-Ordaz. É possível verificar que a função clog-log tem comportamento similar à logito para valores de  $\mu$  próximos de 0 ocorrendo o mesmo para log-log em relação à logito, porém para valores de  $\mu$  próximos de 1.

A função probito tende a assumir valores maiores que a logito para valores de  $\mu$  próximos de 0 e valores menores conforme o valor de  $\mu$  se aproxime de 1 com ponto de inflexão em 0.5. Andrade (2007) estudou a escolha da função de ligação e os efeitos de sua especificação incorreta. A função logito, mesmo quando está especificada incorretamente, resulta em estimativas para as respostas médias com vieses pequenos. Mas, se não há interesse na interpretação dos parâmetros, a escolha da função de ligação de Aranda-Ordaz pode ser mais adequada.



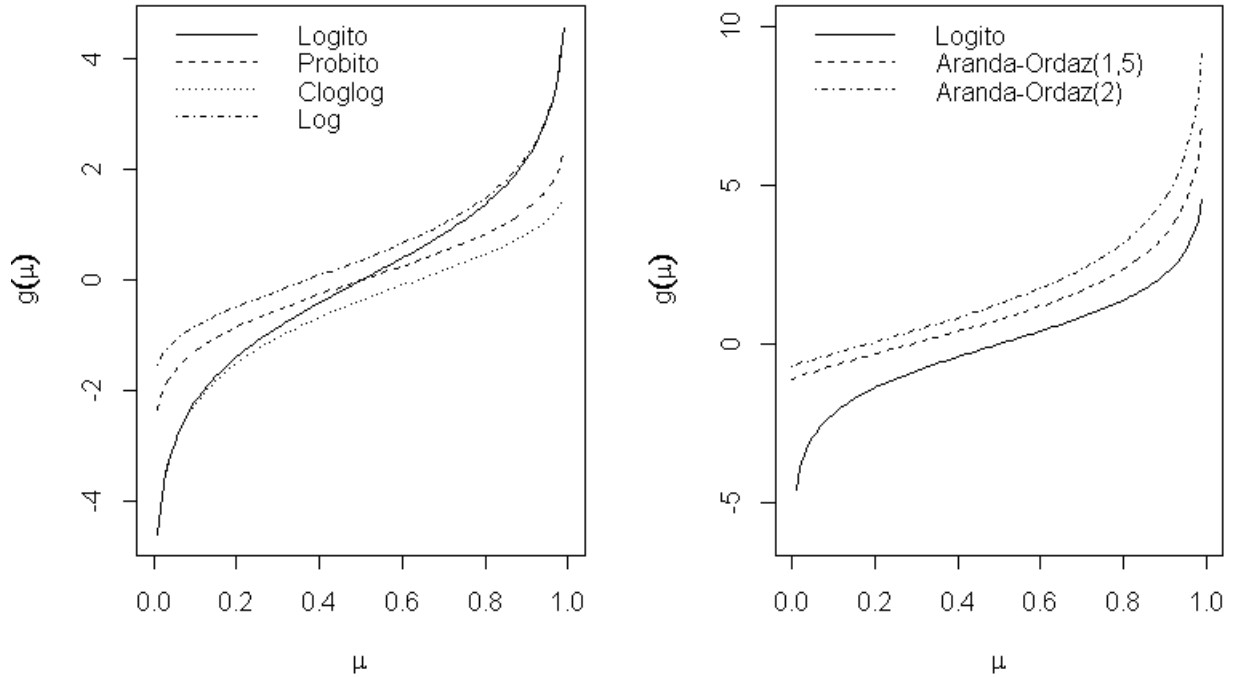


Figura 2.3 *Funções de Ligação logito, log-log, clog-log e Aranda-ordaz.*

A função de ligação logito é a mais utilizada, pois possibilita interpretação simples para os parâmetros de regressão. Utilizando a equação (2.6), tem-se

$$g(\mu_t) = \log \left\{ \frac{\mu_t}{1 - \mu_t} \right\} = x_t^\top \boldsymbol{\beta}, t = 1, \dots, n.$$

Alternativamente, o modelo pode ser escrito como

$$\frac{\mu_t}{1 - \mu_t} = \exp(x_t^\top \boldsymbol{\beta}),$$

ou ainda

$$g^{-1}(\eta_t) = \mu_t = \frac{\exp(x_t^\top \boldsymbol{\beta})}{1 + \exp(x_t^\top \boldsymbol{\beta})}.$$

Quando a função logito é utilizada os parâmetros de regressão podem ser interpretados em termos de razão de chances. Suponha que o valor da  $i$ -ésima variável regressora é aumentado por  $c$  unidades e todas as outras variáveis independentes permanecem inalteradas. Seja  $\mu^\dagger$  a média de  $Y$  sob este novo valor das covariadas, uma vez que  $\mu$  denota a

média de  $Y$  sob o valor original das covariadas. Temos então que

$$\frac{\mu^\dagger}{1 - \mu^\dagger} = \exp(x_{t1}\beta_1 + \cdots + x_{ti}\beta_i + \cdots + x_{tk}\beta_k),$$

e agora é fácil verificar que

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

isto é,  $\exp(c\beta_i)$  é a razão de chances (*oddsratio*).

## 2.5 Função Escore e Matriz de Informação

Nesta seção, apresentamos a função *escore* e a *matriz de Informação* de Fisher para o vetor de parâmetros  $\boldsymbol{\beta}$  e, também, para o parâmetro de precisão  $\phi$ , que são obtidas através das derivadas do logaritmo natural da função de verossimilhança com respeito a  $\boldsymbol{\beta}$  e  $\phi$ , respectivamente.

O logaritmo natural da função de verossimilhança baseada em uma amostra de  $n$  observações independentes é

$$l(\boldsymbol{\beta}, \phi) = \log\left(\prod_{i=1}^n f(\mu_t, \phi)\right) = \sum_{i=1}^n l_t(\mu_t, \phi), \quad (2.7)$$

em que

$$\begin{aligned} l_t(\boldsymbol{\beta}, \phi) &= \log\Gamma(\phi) - \log\Gamma(\mu_t\phi) - \log\Gamma((1 - \mu_t)\phi) \\ &+ (\mu_t\phi - 1)\log y_t + \{(1 - \mu_t)\phi - 1\}\log(1 - y_t), \end{aligned}$$

com  $\mu_t = g^{-1}(\eta_t)$  definido em (2.6). A função escore é obtida derivando o logaritmo da função de verossimilhança com respeito aos parâmetros desconhecidos. Segue que para  $i = 1, \dots, k$ ,

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_i} = \sum_{i=1}^n \frac{\partial l(\mu_t, \phi)}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \beta_i} \quad (2.8)$$

$$\frac{\partial l_t(\mu_t, \phi)}{\partial \mu_t} = \phi \left[ \log \frac{y_t}{1 - y_t} - \{\psi(\mu_t\phi) - \psi((1 - \mu_t)\phi)\} \right] \quad (2.9)$$

Onde  $\psi(\cdot)$  é a função digama, definida como

$$\psi(a) = \frac{d \log \Gamma(a)}{da} = \frac{\Gamma'(a)}{\Gamma(a)}.$$

Note que  $d\mu_t/d\eta_t = 1/g'(\mu_t)$  e também  $\partial\eta_t/\partial\beta_i = x_{ti}$ . Antes de obter a expressão final para (2.8), sejam

$$y_t^* = \log \left\{ \frac{y_t}{1 - y_t} \right\} \text{ e } \mu_t^* = \psi(\mu_t\phi) - \psi((1 - \mu_t)\phi) \quad (2.10)$$

Assim, a expressão (2.9) pode ser reescrita como

$$\frac{\partial l(\mu_t, \phi)}{\partial \mu_t} = \phi(y_t^* - \mu_t^*)$$

e, finalmente (2.8) reduz-se a

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_i} = \phi \sum_{t=1}^n (y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} x_{ti} \quad (2.11)$$

Note também que assumindo condições usuais de regularidade (ver, SEN e SINGER, 1993, Capítulo 7) apud Oliveira(2004), o valor esperado de  $\partial l(\mu_t, \phi)/\partial \mu_t$  é igual a zero. Mas,

$$E \left[ \frac{\partial l(\mu_t, \phi)}{\partial \mu_t} \right] = 0 \Leftrightarrow E[\phi(y_t^* - \mu_t^*)] = 0 \Leftrightarrow E[y_t^*] = \mu_t^*,$$

ou seja, o valor esperado da variável transformada  $y_t^*$  iguala-se a  $\mu_t^*$ . A partir da primeira derivada do logaritmo da função de verossimilhança em relação a  $\beta_i, i = 1, \dots, k$ , obtemos a função escore para  $\boldsymbol{\beta}$ , o vetor coluna  $U_\beta(\boldsymbol{\beta}, \phi)$  de dimensão  $k$ , pode ser escrito na seguinte forma matricial:

$$U_\beta(\boldsymbol{\beta}, \phi) = \phi X^\top \mathbf{G}(y^* - \boldsymbol{\mu}^*) \quad (2.12)$$

em que  $X$  é uma matriz  $(n \times k)$  cuja  $t$ -ésima linha é  $x_t^\top$ ,  $\mathbf{G} = \text{diag} \left\{ \frac{1}{g'(\mu_1)}, \dots, \frac{1}{g'(\mu_n)} \right\}$ ,  $\mathbf{y}^* = (y_1^*, \dots, y_n^*)^\top$  e  $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_n^*)^\top$ .

Analogamente, a derivada de 1ª ordem do logaritmo da função de verossimilhança em relação a  $\phi$  é dada por

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \phi} = \sum_{t=1}^n \frac{\partial l_t(\mu_t, \phi)}{\partial \phi}, \quad (2.13)$$

com

$$\begin{aligned} \frac{\partial l_t(\boldsymbol{\beta}, \phi)}{\partial \phi} &= \mu_t \left[ \log \frac{y_t}{1 - y_t} - \psi(\mu_t\phi) + \psi((1 - \mu)\phi) \right] \\ &+ \log(1 - y_t) - \psi((1 - \mu)\phi) + \psi(\phi) \end{aligned} \quad (2.14)$$

Portanto, a função escore para o parâmetro de precisão  $\phi$  é dada por

$$U_\phi(\beta, \phi) = \sum_{t=1}^n \{\mu_t(y_t^* - \mu_t^*) + \log(1 - y_t) - \psi((1 - \mu_t)\phi) + \psi(\phi)\}, \quad (2.15)$$

sendo  $U_\phi(\beta, \phi)$  um escalar.

O nosso próximo passo é obter a *matriz de informação* de Fisher para  $(\beta, \phi)$ . De 2.8, as derivadas de 2ª ordem de  $l(\beta, \phi)$  com respeito a  $\beta_i$  e  $\beta_j$  são dadas por

$$\frac{\partial^2 l(\beta, \phi)}{\partial \beta_i \partial \beta_j} = \sum_{t=1}^n \left\{ \frac{\partial^2 l(\mu_t, \phi)}{\partial \mu_t^2} \frac{d\mu_t}{d\eta_t} + \frac{\partial l_t(\mu_t, \phi)}{\partial \mu_t} \frac{\partial}{\partial \mu_t} \left( \frac{d\mu_t}{d\eta_t} \right) \right\} \frac{d\mu_t}{d\eta_t} x_{ti} x_{tj}, \quad (2.16)$$

como  $\mathbb{E}(\partial l_t(\mu_t, \phi)/\partial \mu_t) = 0$ , tem-se

$$\mathbb{E} \left( \frac{\partial^2 l(\beta, \phi)}{\partial \beta_i \partial \beta_j} \right) = \sum_{t=1}^n \mathbb{E} \left[ \frac{\partial^2 l_t(\mu_t, \phi)}{\partial \mu_t^2} \right] \left( \frac{d\mu_t}{d\eta_t} \right)^2 x_{ti} x_{tj}.$$

Agora, de (2.9) temos

$$\frac{\partial^2 l_t(\mu_t, \phi)}{\partial \mu_t^2} = -\phi^2 \{\psi'(\mu_t \phi) - \psi'((1 - \mu_t)\phi)\},$$

sendo  $\psi(\cdot)$  a função trigama definida como

$$\psi'(\lambda) = \frac{d\psi(\lambda)}{d\lambda} = \frac{d^2 \log \Gamma(\lambda)}{d\lambda^2} = \frac{\Gamma''(\lambda)\Gamma(\lambda) - [\Gamma'(\lambda)]^2}{[\Gamma(\lambda)]^2}, \lambda > 0$$

e, assim,

$$\mathbb{E} \left( \frac{\partial^2 l(\beta, \phi)}{\partial \beta_i \partial \beta_j} \right) = -\phi \sum_{t=1}^n w_t x_{ti} x_{tj},$$

com

$$w_t = \phi \{\psi'(\mu_t \phi) - \psi'((1 - \mu_t)\phi)\} \frac{1}{\{g'(\mu_t)\}^2}.$$

Em forma matricial, tem-se

$$\mathbb{E} \left( \frac{\partial^2 l(\beta, \phi)}{\partial \beta \partial \beta^\top} \right) = -\phi \mathbf{X}^\top \mathbf{W} \mathbf{X}, \quad (2.17)$$

com  $\mathbf{W} = \text{diag}\{w_1, \dots, w_n\}$ .

De (2.11) a deriva de 2ª ordem de  $l(\beta, \phi)$  com respeito a  $\beta_i$  e  $\phi$  pode ser escrita como

$$\frac{\partial^2 l(\beta, \phi)}{\partial \beta_i \partial \phi} = \sum_{t=1}^n \left[ (y_t^* - \mu_t^*) - \phi \frac{\partial \mu_t^*}{\partial \phi} \right] \frac{1}{g'(\mu_t)} x_{ti}.$$

Como  $\mathbb{E}(y_t^*) = \mu_t^*$  e  $\partial\mu_t^*/\partial\phi = \psi'(\mu_t\phi)\mu_t - \psi'((1-\mu_t)\phi)(1-\mu_t)$ , chega-se a

$$\mathbb{E}\left(\frac{\partial^2 l(\boldsymbol{\beta}, \phi)}{\partial\beta_i \partial\phi}\right) = \sum_{t=1}^n c_t \frac{1}{g'(\mu_t)} x_{ti},$$

em que  $c_t = \phi\{\psi'(\mu_t\phi)\mu_t - \psi'((1-\mu_t)\phi)(1-\mu_t)\}$ . Em notação matricial,

$$\mathbb{E}\left(\frac{\partial^2 l(\boldsymbol{\beta}, \phi)}{\partial\boldsymbol{\beta} \partial\phi}\right) = -\mathbf{X}^\top \mathbf{G} \mathbf{c}$$

com  $\mathbf{c} = (c_1, \dots, c_n)^\top$ .

Finalmente, temos que  $\partial^2 l(\boldsymbol{\beta}, \phi)/\partial\phi^2$  pode ser obtida diferenciando a expressão em (2.13) em relação a  $\phi$ , ou seja,

$$\frac{\partial^2 l(\boldsymbol{\beta}, \phi)}{\partial\phi^2} = \sum_{t=1}^n \frac{\partial^2 l_t(\boldsymbol{\beta}, \phi)}{\partial\phi^2} \quad (2.18)$$

$$= -\sum_{t=1}^n [\psi'(\mu_t\phi)\mu_t^2 - \psi'((1-\mu_t)\phi)(1-\mu_t)^2 - \psi'(\phi)]. \quad (2.19)$$

Seja  $d_t = [\psi'(\mu_t\phi)\mu_t^2 - \psi'((1-\mu_t)\phi)(1-\mu_t)^2 - \psi'(\phi)]$ . Logo,

$$\mathbb{E}\left(\frac{\partial^2 l(\boldsymbol{\beta}, \phi)}{\partial\phi^2}\right) = -\sum_{t=1}^n d_t,$$

que em notação matricial, pode ser escrito como

$$\mathbb{E}\left(\frac{\partial^2 l(\boldsymbol{\beta}, \phi)}{\partial\phi^2}\right) = -tr(\mathbf{D}),$$

em que  $\mathbf{D} = diag\{d_1, \dots, d_n\}$  e  $tr(\mathbf{D})$  indica o *traço* da matriz  $\mathbf{D}$ .

Finalmente, a *matriz de informação* de Fisher para  $(\boldsymbol{\beta}, \phi)$  é dada por

$$\mathbf{K} = \mathbf{K}(\boldsymbol{\beta}, \phi) = \begin{pmatrix} \mathbf{K}_{\beta\beta} & \mathbf{K}_{\beta\phi} \\ \mathbf{K}_{\phi\beta} & \mathbf{K}_{\phi\phi} \end{pmatrix}, \quad (2.20)$$

em que  $\mathbf{K}_{\beta\beta} = \phi \mathbf{X}^\top \mathbf{W} \mathbf{X}$ ,  $\mathbf{K}_{\beta\phi} = \mathbf{K}_{\phi\beta}^\top = \mathbf{X}^\top \mathbf{G} \mathbf{c}$ , e  $\mathbf{K}_{\phi\phi} = tr(\mathbf{D})$ .

Sob condições gerais de regularidade (ver SEN e SINGER (1993)), quando o tamanho da amostra é grande, a distribuição aproximada dos estimadores de máxima verossimilhança é dada por

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\phi} \end{pmatrix} \xrightarrow{D} \mathcal{N}_{k+1} \left( \begin{pmatrix} \boldsymbol{\beta} \\ \phi \end{pmatrix}, \mathbf{K}^{-1} \right) \quad (2.21)$$

no qual  $\hat{\beta}$  e  $\hat{\phi}$  são estimadores de máxima verossimilhança de  $\beta$  e  $\phi$ , respectivamente.

Para se obter os erros padrões assintóticos dos estimadores de máxima verossimilhança é interessante escrever  $\mathbf{K}^{-1}$  de forma explícita. Utilizando a fórmula para inversa de matrizes subdivididas dada por Rao (1973, p. 33), obtém-se

$$\mathbf{K}^{-1} = \mathbf{K}^{-1}(\beta, \phi) = \begin{pmatrix} \mathbf{K}^{\beta\beta} & \mathbf{K}^{\beta\phi} \\ \mathbf{K}^{\phi\beta} & \mathbf{K}^{\phi\phi} \end{pmatrix}, \quad (2.22)$$

em que

$$\mathbf{K}^{\beta\beta} = \frac{1}{\phi} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \left\{ \mathbf{I}_k + \frac{\mathbf{X}^\top \mathbf{G} \mathbf{c} \mathbf{c}^\top \mathbf{G}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}}{\gamma \phi} \right\}$$

, com

$$\gamma = \text{tr}(\mathbf{D}) - \phi^{-1} \mathbf{c}^\top \mathbf{G}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{G} \mathbf{c},$$

$$\mathbf{K}^{\beta\phi} = (\mathbf{K}^{\phi\beta})^\top = -\frac{1}{\gamma \phi} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{G} \mathbf{c},$$

e  $\mathbf{K}^{\phi\phi} = \gamma^{-1}$ . Aqui  $\mathbf{I}_k$  é a matriz identidade de ordem  $k$ .

## 2.6 Estimação dos Parâmetros

Os estimadores de máxima verossimilhança dos parâmetros,  $\hat{\beta}$  e  $\hat{\phi}$ , são obtidos através da solução do sistema

$$U_\beta(\beta, \phi) = 0$$

$$U_\phi(\beta, \phi) = 0$$

Este sistema não possui solução em forma fechada e as estimativas de máxima verossimilhança são obtidas por meio da maximização numérica da função de log-verossimilhança. A maximização deve ser realizada usando um algoritmo de otimização não linear, tais como algoritmos baseados no método de Newton-Raphson, *escore* de Fisher ou um algoritmo Quasi-Newton, conhecido como método BFGS (BROYDEN, FLETCHER, GOLDFARB E SHANNO, ver NOCEDAL e WRIGT, 1999). Na próxima seção descreveremos estes três métodos iterativos.

### 2.6.1 Processo Iterativo de Newton-Raphson

Seja  $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi)^\top$ , o vetor de parâmetros e  $U(\boldsymbol{\theta}) = (U_\beta(\boldsymbol{\beta}, \phi)^\top, U_\phi(\boldsymbol{\beta}, \phi))^\top$ , o vetor das funções escores de dimensão  $(K + 1) \times 1$ . O processo iterativo de Newton-Raphson para obtenção da estimativa de máxima verossimilhança do vetor  $\boldsymbol{\theta}$  é definido expandindo-se a função escore  $U(\boldsymbol{\theta})$  em torno de um valor inicial  $\boldsymbol{\theta}^{(0)}$ , tal que

$$U(\boldsymbol{\theta}) \approx U(\boldsymbol{\theta}^{(0)}) + U'(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})$$

Em que  $U'(\boldsymbol{\theta})$  denota a derivada de 1ª ordem de  $U(\boldsymbol{\theta})$  com respeito a  $\boldsymbol{\theta}^\top$ . Fazendo  $U(\boldsymbol{\theta}) = 0$  e repetindo o procedimento acima, chega-se ao seguinte processo iterativo

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \{-U'(\boldsymbol{\theta}^{(m)})\}^{-1}U(\boldsymbol{\theta}^{(m)}), m = 0, 1, \dots \quad (2.23)$$

O aspecto mais trabalhoso desse esquema iterativo é a inversão da matriz  $U'(\boldsymbol{\theta})$ .

### 2.6.2 Processo Iterativo de Escore de Fisher

A expressão (2.23) apresenta uma forma alternativa equivalente assintoticamente, pois, pela Lei dos Grandes Números,  $U'(\boldsymbol{\theta})$  converge para a matriz  $\mathbf{K}$  definida em (2.20), quando  $n \rightarrow \infty$ , substituindo a informação observada pela esperada. Neste caso obtemos o procedimento iterativo

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \{-K^{(m)}\}^{-1}U(\boldsymbol{\theta}^{(m)}), m = 0, 1, \dots$$

Esse procedimento é denominado método *escore* de Fisher. Semelhantemente ao método de Newton-Raphson o aspecto mais moroso desse esquema iterativo é a inversão da matriz  $\mathbf{K}$ .

### 2.6.3 Método BFGS

O método BFGS foi desenvolvido por Broyden, Fletcher, Goldfarb e Shanno (ver NOCEDAL e WRIGT, 1999). Este utiliza o mesmo princípio do método de Newton-Raphson, diferenciando-se pelo fato de utilizar uma sequência de matrizes simétricas e positivas definidas  $\mathbf{B}^{(m)}$  no lugar da matriz  $\{U'(\boldsymbol{\theta}^{(m)})\}^{-1}$ . Comumente, toma-se como matriz inicial,  $\mathbf{B}^{(0)}$ , a matriz identidade de mesma ordem. A forma recursiva para obter as demais

matrizes é dada por

$$\mathbf{B}^{(m+1)} = \mathbf{B}^{(m)} - \frac{\mathbf{B}^{(m)}\mathbf{S}^{(m)}(\mathbf{S}^{(m)})^\top\mathbf{B}^{(m)}}{(\mathbf{S}^{(m)})^\top\mathbf{B}^{(m)}\mathbf{S}^{(m)}} + \frac{\mathbf{Y}^{(m)}(\mathbf{Y}^{(m)})^\top}{(\mathbf{Y}^{(m)})^\top\mathbf{S}^{(m)}}, m = 0, 1, \dots,$$

em que  $\mathbf{S}^{(m)} = \boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}$  e  $\mathbf{Y} = U(\boldsymbol{\theta}^{(m+1)}) - U(\boldsymbol{\theta}^{(m)})$ . De forma análoga aos métodos anteriores, o máximo é obtido pela recorrência

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \alpha^{(m)}\mathbf{B}^{(m)}U(\boldsymbol{\theta}^{(m)}), m = 0, 1, \dots,$$

em que  $\alpha^{(m)}$  é um escalar determinado por algum procedimento de busca linear partir de  $\boldsymbol{\theta}^{(m)}$  na direção  $-\mathbf{B}^{(m)}U(\boldsymbol{\theta}^{(m)})$  de forma que  $f(y; \boldsymbol{\theta}^{(m)})$  cresça nessa direção.

Os algoritmos de otimização requerem a especificação de valores iniciais a serem usados no esquema iterativo. Ferrari e Cribari-Neto (2004) sugerem utilizar como valores iniciais para  $\boldsymbol{\beta}$  a estimativa de mínimos quadrados ordinários obtida da regressão linear das respostas transformadas  $(g(y_1), \dots, g(y_n))$  com relação a  $\mathbf{X}$ , ou seja,  $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{z}$ , onde  $\mathbf{z} = (g(y_1), \dots, g(y_n))^\top$ . Em relação ao parâmetro de precisão, eles sugerem um valor inicial para  $\phi$  baseado no fato de que

$$\text{Var}(Y_t) = \frac{\mu_t(1 - \mu_t)}{1 + \phi}$$

de forma alternativa, podemos escrever

$$\phi = \frac{\mu_t(1 - \mu_t)}{\text{Var}(Y_t)} - 1$$

e ao expandir até a primeira ordem a função  $g(y_t)$  em série de Taylor ao redor do ponto  $\mu_t$  e tomando a variância, temos que

$$\text{Var}\{g(y_t)\} \approx \text{Var}\{g(\mu_t) + (y_t - \mu_t)g'(\mu_t)\} = \text{Var}(y_t)\{g'(\mu_t)\}^2,$$

isto é,  $\text{Var}(y_t) \approx \text{Var}\{g(y_t)\}\{g'(\mu_t)\}^{-2}$ . Portanto, o valor sugerido é

$$\phi^{(0)} = \frac{1}{n} \sum_{t=1}^n \frac{\check{\mu}_t(1 - \check{\mu}_t)}{\check{\sigma}_t^2} - 1,$$

em que  $\check{\mu}_t$  é obtido aplicando  $g^{-1}(\cdot)$  para o  $t$ -ésimo valor ajustado da regressão linear de  $(g(y_1), \dots, g(y_n))$  em  $\mathbf{X}$ , isto é,

$$\check{\mu}_t = g^{-1}(\mathbf{x}_t^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{z}) \text{ e } \check{\sigma}_t^2 = \check{\mathbf{e}}^\top\check{\mathbf{e}}/[(n - k)\{g'(\check{\mu}_t)\}^2].$$

Aqui,  $\check{\mathbf{e}} = \mathbf{z} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{z}$  é o vetor de resíduos de mínimos quadrados ordinários de uma regressão linear com variável resposta transformada. Oliveira (2004) relata que estes valores iniciais proporcionaram resultados satisfatórios em seus experimentos.



## 2.7 Teste de Hipóteses

Teste de hipóteses é um método estatístico para verificar se as estimativas dos parâmetros são compatíveis com alguma hipótese, podendo muitas vezes sugerir sua não validade. Na classe dos modelos lineares generalizados os testes de hipóteses usuais para os parâmetros dos modelos são os teste da Razão de Verossimilhanças, Escore e Wald. Analogamente, Ferrari e Cribari-Neto (2004) desenvolveram estes testes para os parâmetros do modelo de regressão Beta para averiguar sua significância estatística. Entre estes, nesta dissertação será visto o teste de Wald descrito abaixo.

### 2.7.1 Teste de Wald

O teste de Wald é utilizado para realizar inferências assintóticas acerca do vetor de parâmetros  $\beta$ . A estatística do teste de  $\mathcal{H}_0 : \beta_1 = \beta_1^{(0)}$  versus  $\mathcal{H}_1 : \beta_1 \neq \beta_1^{(0)}$  é

$$w_3 = (\hat{\beta}_1 - \beta_1^{(0)})^\top (\hat{\mathbf{K}}_{11}^{\beta\beta})^{-1} (\hat{\beta}_1 - \beta_1^{(0)}),$$

em que  $\hat{\mathbf{K}}_{11}^{\beta\beta}$  é igual a  $\mathbf{K}_{11}^{\beta\beta}$  avaliado no estimador de máxima verossimilhança sem restrição, e  $\hat{\beta}_1$  é o estimador de máxima verossimilhança de  $\beta_1$ . Sob condições gerais de regularidade e supondo  $\mathcal{H}_0$  verdadeira, temos que  $w_3 \xrightarrow{D} \chi_m^2$ , onde  $m$  é a dimensão do espaço paramétrico sob a hipótese  $\mathcal{H}_0$ . Em particular, para testar a significância do  $i$ -ésimo parâmetro regressor ( $\beta_i$ ),  $i = 1, \dots, k$ , podemos utilizar a raiz quadrada sinalizada da estatística de Wald, isto é,

$$\frac{\hat{\beta}_i}{EP(\hat{\beta}_i)},$$

no qual  $EP(\beta_i)$  é o erro padrão assintótico do estimador de máxima verossimilhança de  $\beta_i$ , obtido da inversa da matriz de informação de Fisher avaliada nos estimadores de máxima verossimilhança. A distribuição nula limite da estatística deste teste é Normal padrão.

## 2.8 Resíduos

Vários tipos de resíduos estão disponíveis na literatura para a avaliação do modelo de regressão Beta. Alguns trabalhos a respeito de resíduos e análise de diagnóstico em

regressão Beta são encontrados em Espinheira *et al.* (2008a,b) e Ferrari *et al.* (2011). Dentre os diversos resíduos propostos o resíduo ponderado padronizado 2 proposto por Espinheira *et al.* (2008) é o que apresenta as melhores propriedades, além de ser o resíduo mais eficiente na identificação de observações que tem maior influência nas estimativas dos parâmetros. Nesta seção, apresentamos os resíduos proposto por Espinheira *et al.* (2008) e Ferrari e Cribari-Neto (2004). Para definição desses resíduos foi considerado o modelo (2.2) com  $\phi$  conhecido.

### 2.8.1 Resíduo Ordinário

O resíduo ordinário padronizado foi definido por Ferrari e Cribari-Neto (2004) como

$$r_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\widehat{Var}(y_t)}}$$

em que  $\widehat{Var}(Y_t) = \hat{\mu}_t(1 - \hat{\mu}_t)/(1 + \hat{\phi})$ ,  $\hat{\mu}_t = g^{-1}(\mathbf{x}_t^\top \hat{\boldsymbol{\beta}})$ , e  $\hat{\boldsymbol{\beta}}$  e  $\hat{\phi}$  são os estimadores de máxima verossimilhança de  $\boldsymbol{\beta}$  e  $\phi$ , respectivamente.

### 2.8.2 Resíduo ponderado padronizado 1

Este resíduo foi baseado no método *score* de Fisher para estimação de  $\beta$ , com  $\phi$  conhecido. O processo iterativo é dado por

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{X}^\top \mathbf{W}^{(m)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{G}^{(m)}(y^* - \boldsymbol{\mu}^{*(m)}) \quad (2.24)$$

com  $\mathbf{y}^*$ ,  $\boldsymbol{\mu}^*$ ,  $\mathbf{G}$  e  $\mathbf{W}$  definidos em (2.10), (2.12) e (2.20). Reescrevendo o processo iterativo (2.24) em termos de regressão de mínimos quadrados ponderados:

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^\top \mathbf{W}^{(m)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(m)} \mathbf{z}^{(m)},$$

sendo  $\mathbf{z}^{(m)} = \boldsymbol{\eta}^{(m)} + \mathbf{W}^{-1(m)} \mathbf{G}^{(m)}(y^* - \boldsymbol{\mu}^{*(m)})$ , com  $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)^\top = \mathbf{X}^\top \boldsymbol{\beta}$ . Na convergência, segue que

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^\top \hat{\mathbf{W}} \mathbf{z} \quad (2.25)$$

com

$$\mathbf{z} = \hat{\boldsymbol{\eta}} + \mathbf{W}^{-1(m)} \hat{\mathbf{G}}(y^* - \hat{\boldsymbol{\mu}}^{*(m)}),$$

onde  $\hat{\mathbf{W}}$ , e  $\hat{\mathbf{G}}$  são as matrizes  $\mathbf{W}$  e  $\mathbf{G}$  avaliadas no estimador de máxima verossimilhança. Assim, usando as definições  $\mathbf{G}$  e  $\mathbf{W}$  em (2.12), Espinheira et al. (2008) definem o resíduo ponderado como

$$r_t^* = \frac{y_t^* - \mu_t^*}{\sqrt{\phi v_t}}, \quad (2.26)$$

com  $v_t = \psi'(\hat{\mu}_t\phi) + \psi'((1 - \hat{\mu}_t)\phi)$ . A mesma autora propõe utilizar apenas a variância de  $y_t^*$  ( $v_t$ ) no denominador do resíduo ponderado dando origem ao resíduo ponderado padronizado 1 dado por

$$r_t^p = \phi^{1/2} r_t^* = \frac{y_t^* - \mu_t^*}{\sqrt{v_t}}, \quad (2.27)$$

### 2.8.3 Resíduo ponderado padronizado 2

Outro resíduo definido por Espinheira *et al.* (2008) se baseia na variância de  $\mathbf{z}$ . Para isso, reescrevemos (2.25) como  $(\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \hat{\mathbf{W}} \mathbf{z}$ . Note que  $\hat{\boldsymbol{\beta}}$  é distribuído como  $\mathcal{N}_k(\boldsymbol{\beta}, \mathbf{K}_{\beta\beta}^{-1})$  quando  $\phi$  é conhecido (fixo), com  $\mathbf{K}_{\beta\beta} = \phi(\mathbf{X}^\top \mathbf{W} \mathbf{X})$ , ou seja,  $\text{cov}(\hat{\boldsymbol{\beta}}) = \phi^{-1}(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$ . Segue que  $\text{cov}(\mathbf{z}) \approx \phi^{-1} \hat{\mathbf{W}}^{-1}$ . Assim, obtém-se  $\text{cov}(\mathbf{r}^*) \approx \phi^{-1}(\mathbf{I}_n - \mathbf{H})$ , onde  $\mathbf{I}_n$  é a matriz de identidade de ordem  $n$  e  $\mathbf{H} = \hat{\mathbf{W}}^{1/2} \mathbf{X}(\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^\top \hat{\mathbf{W}}^{1/2}$ . O resíduo padronizado ponderado 2 é dado por

$$r_t^{pp} = \frac{r_t^*}{\sqrt{\phi^{-1}(1 - h_{tt})}} = \frac{r_t^p}{\sqrt{(1 - h_{tt})}} = \frac{y_t^* - \mu_t^*}{\sqrt{v_t(1 - h_{tt})}} \quad (2.28)$$

em que  $h_{tt}$  é o  $t$ -ésimo elemento da diagonal de  $\mathbf{H}$  (veja a seção 2.9.1). Na prática, para calcularmos esses resíduos, devemos substituir  $\phi$  pela sua estimativa de máxima verossimilhança  $\hat{\phi}$ .

## 2.9 Diagnóstico do Modelo ajustado

Uma etapa essencial na análise do ajuste dos modelos de regressão é a verificação da adequação do modelo aos dados. Esta etapa, conhecida como diagnóstico do modelo ajustado constitui um conjunto de critérios e ferramentas gráficas na avaliação da adequação do modelo ajustado aos dados. Algumas medidas de diagnóstico fornecem subsídios para detectar: possível violação das suposições feitas para o modelo, especialmente para a componente aleatória ( $y$ ) e a função de ligação ( $g(\cdot)$ ), aleatoriedade dos dados, presença de

pontos extremos (*outliers*), adequação da distribuição de probabilidade proposta para a variável dependente e observação de pontos influentes.

Nesta seção serão apresentadas ferramentas gráficas para detecção de desvio do modelo e medidas de análise de influência, como a alavanca generalizada proposta por Wei, Hu e Fung (1998) e a distancia de Cook (COOK, 1977).

### 2.9.1 Ponto Alavanca ou Alavanca Generalizada

É um componente importante na análise de influência em modelos de regressão. A alavanca generalizada (GL) no modelo de regressão Beta definido por Ferrari e Cribari-Neto (2004) apresenta-se sob dois enfoques. Primeiramente, sob a suposição de que o parâmetro de precisão  $\phi$  é conhecido e, num segundo momento, admitindo que  $\phi$  é desconhecido.

Considerando  $\phi$  conhecido, Ferrari e Cribari-Neto (2004) obtiveram uma expressão de forma fechada para  $GL(\beta)$  no modelo de regressão Beta proposto. Assim, a alavanca generalizada fica dada por

$$GL(\beta) = D_{\beta} \left( -\frac{\partial^2 l(\beta, \phi)}{\partial \beta \partial \beta^{\top}} \right)^{-1} \frac{\partial^2 l(\beta, \phi)}{\partial \beta \partial \mathbf{y}^{\top}}$$

em que

$$D_{\beta} = \frac{\partial \boldsymbol{\mu}}{\partial \beta^{\top}} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}^{\top}} \frac{\partial \boldsymbol{\eta}}{\partial \beta^{\top}} = \mathbf{GX} \quad (2.29)$$

e de (2.16)

$$\begin{aligned} -\frac{\partial^2 l(\beta, \phi)}{\partial \beta_i \partial \beta_j} &= \sum_{t=1}^n \frac{\partial}{\partial \mu_t} \left\{ \frac{\partial^2 l(\mu_t, \phi)}{\partial \mu_t^2} \frac{d\mu_t}{d\eta_t} + \frac{\partial l_t(\mu_t, \phi)}{\partial \mu_t} \frac{\partial}{\partial \mu_t} \left( \frac{d\mu_t}{d\eta_t} \right) \right\} \frac{d\mu_t}{d\eta_t} x_{ti} x_{tj} \quad (2.30) \\ &= -\phi \sum_{t=1}^n \left\{ \phi \{ \psi'(\mu_t \phi) + \psi'((1 - \mu_t)\phi) \} + (y_t^* - \mu_t^*) \frac{g''(\mu_t)}{g'(\mu_t)} \right\} \frac{x_{ti} x_{tj}}{\{g'(\mu_t)\}^2} \end{aligned}$$

que em forma matricial, é dado por

$$-\frac{\partial^2 l(\beta, \phi)}{\partial \beta \partial \beta^{\top}} = \phi \mathbf{X}^{\top} \mathbf{Q} \mathbf{X},$$

em que  $\mathbf{Q} = \text{diag}\{q_1, \dots, q_n\}$  com

$$q_t = \left\{ \phi \{ \psi'(\mu_t \phi) + \psi'((1 - \mu_t)\phi) \} + (y_t^* - \mu_t^*) \frac{g''(\mu_t)}{g'(\mu_t)} \right\} \frac{1}{\{g'(\mu_t)\}^2}, t = 1, \dots, n.$$

Adicionalmente, temos que

$$\frac{\partial^2 l(\boldsymbol{\beta}, \phi)}{\partial \beta_i \partial \mathbf{y}_t} = \phi \sum_{t=1}^n \left( \frac{1}{y_t(1-y_t)} \frac{1}{g'(\mu_t)} x_{ti} \right),$$

que em forma matricial, fica

$$\frac{\partial^2 l(\boldsymbol{\beta}, \phi)}{\partial \beta_i \partial \mathbf{y}_t} = \phi \mathbf{X}^\top \mathbf{G} \mathbf{M},$$

em que  $\mathbf{M} = \{m_1, \dots, m_n\}$  com  $m_t = 1/\{y_t(1-y_t)\}$ ,  $t = 1, \dots, n$ . Finalmente, a alavanca generalizada pode ser escrita como

$$GL(\boldsymbol{\beta}) = \mathbf{G} \mathbf{X} (\mathbf{X}^\top \mathbf{Q} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{G} \mathbf{M}. \quad (2.31)$$

Ferrari e Cribari-Neto (2004) notaram que ao substituir a informação observada,  $-\partial^2 l(\boldsymbol{\beta}, \phi)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top$ , pela informação esperada,  $\mathbb{E}(-\partial^2 l(\boldsymbol{\beta}, \phi)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top)$ , a expressão para  $GL(\boldsymbol{\beta})$  fica como dada em (2.31), mas com a matriz  $\mathbf{Q}$  substituída por  $\mathbf{W}$ , sendo dada por  $GL^*(\boldsymbol{\beta})$ . Pode-se notar que os elementos da diagonal de  $GL^*(\boldsymbol{\beta})$  são os mesmos de

$$\mathbf{M}^{1/2} \mathbf{G} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{G} \mathbf{M}^{1/2},$$

em que  $\mathbf{M}^{1/2} \mathbf{G}$  é uma matriz diagonal cujo t-ésimo elemento diagonal é dado por  $\{g'(\mu_t) V(y_t)^{1/2}\}^{-1}$ . Outro fato importante aqui é notar que existe uma conexão estreita entre os elementos da diagonal de  $GL^*(\boldsymbol{\beta})$  e aqueles da “matriz chapéu” usual de modelos lineares generalizados,

$$\mathbf{H} = \mathbf{M}^{1/2} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}^{1/2},$$

quando  $\phi$  é grande. A relação origina-se do fato de que, quando o parâmetro de precisão é grande, o t-ésimo elemento da diagonal de  $\mathbf{W}^{(1/2)}$  é aproximadamente igual a

$$[g'(\mu_t) V(\mu_t)^{1/2}]^{-1},$$

como pode ser verificado na expressão

$$\omega \approx \phi \left\{ \frac{1}{\mu\phi} + \frac{1}{(1-\mu)\phi} \right\} \frac{1}{\{g'(\mu)\}^2} = \frac{1}{(1-\mu)} \frac{1}{\{g'(\mu)\}^2}.$$

Agora, considerando  $\phi$  desconhecido, temos que  $\boldsymbol{\theta}^\top = (\boldsymbol{\beta}^\top, \phi)$ . Segue, então que,

$$D_\theta = \begin{bmatrix} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^\top} & \frac{\partial \boldsymbol{\mu}}{\partial \theta^\top} \end{bmatrix} = \begin{bmatrix} \mathbf{T} \mathbf{X} & \mathbf{0} \end{bmatrix},$$

em  $\mathbf{0}$  é um vetor  $n$ -dimensional de zeros. A quantidade  $-\partial^2 l / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top$  é dada por (2.20) com  $\mathbf{W}$  substituído por  $\mathbf{Q}$  e  $\mathbf{c}$  substituído por  $\mathbf{f}$ , em que  $\mathbf{f} = (f_1, \dots, f_n)^\top$  com  $f_t = \{c_t - (y_t^* - \mu_t^*)\}$ ,  $t = 1, \dots, n$ . Assim a inversa de  $-\partial^2 l / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top$  será dada por (2.22) com  $\mathbf{W}$  e  $\mathbf{c}$  substituídos por  $\mathbf{Q}$  e  $\mathbf{f}$ , respectivamente. Adicionalmente,

$$\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \mathbf{y}^\top} = \begin{pmatrix} \phi \mathbf{X}^\top \mathbf{G} \mathbf{M} \\ \mathbf{b}^\top \end{pmatrix}$$

em que  $\mathbf{b} = (b_1, \dots, b_n)^\top$  com  $b_t = -(y_t - \mu_t) / \{y_t(1 - y_t)\}$ ,  $t = 1, \dots, n$ . Finalmente, a alavanca generalizada é dada por

$$GL(\boldsymbol{\beta}, \phi) = GL(\boldsymbol{\beta}) + \frac{1}{\gamma \phi} \mathbf{G} \mathbf{X} (\mathbf{X}^\top \mathbf{Q} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{G} f (f^\top \mathbf{G} \mathbf{X} (\mathbf{X}^\top \mathbf{Q} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{G} \mathbf{M} - \mathbf{b}^\top),$$

onde  $GL(\boldsymbol{\beta})$  é dado em (2.31). Quando  $\phi$  é suficientemente grande,  $GL(\boldsymbol{\beta}, \phi) \approx GL(\boldsymbol{\beta})$ .

Um ponto alavanca possui um perfil diferente dos demais em relação aos valores das variáveis explicativas. Na prática, construímos um gráfico de  $h_{tt}$  versus o índice das observações  $t$ ,  $t = 1, \dots, n$ . Um valor grande de  $h_{tt}$  ( $t$ -ésimo elemento da diagonal principal de  $\mathbf{H}$ ). Quando comparado com os valores das demais observações pode indicar que um ponto é alavanca.

Supondo que todos os pontos exerçam a mesma influência sobre os valores ajustados pode-se esperar que cada valor de  $h_{tt}$  esteja próximo de  $tr(\mathbf{H})/n = k/n$ . Dessa forma, sugere-se examinar os pontos para os quais  $h_{tt} \geq 2k/n$ , conhecidos como pontos de alavanca. Como os valores de  $h_{tt}$  dependem da matriz de pesos  $\hat{\mathbf{W}}$ , observações com valores grandes de  $h_{tt}$  precisam ser examinados com cautela, pois nem sempre são considerados ponto de alavanca.

### 2.9.2 Distância de Cook

A distância de Cook (Cook, 1977) é uma medida bastante utilizada para detectar a influência de cada observação nas estimativas dos parâmetros do modelo de regressão. Esta medida identifica a influência da retirada da  $i$ -ésima observação sobre as estimativas dos parâmetros do modelo, sendo definida por

$$\mathbf{K}^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(t)})^\top \mathbf{X}^\top \mathbf{W} \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(t)}),$$

em que  $\hat{\beta}_{(t)}$ , é a estimativa do parâmetro sem a  $t$ -ésima observação. Essa quantidade mede a distância quadrática entre  $\hat{\beta}$  e  $\hat{\beta}_{(t)}$ . Para evitar ajustar o modelo  $(n+1)$  vezes, utilizamos a aproximação usual para a distância de Cook, dada por

$$C_t = \frac{h_{tt}r_t^2}{k(1-h_{tt})^2},$$

que combina alavanca e resíduos. Portanto  $C_t$  será grande quando o  $t$ -ésimo ponto fornecer  $r_t$  grande ou quando  $h_{tt}$  for próximo de um. Usualmente constrói-se um gráfico de  $C_t$  contra a ordem das observações ( $t$ ) para detectar pontos suspeitos de serem influentes.

### 2.9.3 Gráfico de Probabilidade Meio-Normal com Envelope Simulado

Segundo Atkinson (1985) como a distribuição dos resíduos não é conhecida, gráficos de probabilidade meio-normal com envelope simulado são ferramentas de diagnóstico muito úteis. A proposta desta medida é acrescentar ao gráfico de probabilidade normal usual um envelope simulado que pode ser usado para decidir se as observações são consistentes com o modelo ajustado.

Este gráfico é construído com base nos resíduos padronizados. Detalhes sobre sua construção podem ser encontrados em Neter, Kutner, Nachtleim e Wasserman (1996). O envelope simulado é a banda de confiança. Tendências não aleatórias dos resíduos dentro do envelope gerado podem indicar especificação incorreta da distribuição dos dados, da parte sistemática do modelo ou da função de ligação.

Ressalta-se que o modelo de regressão Beta descrito neste capítulo é apropriado para respostas no intervalo aberto  $(0, 1)$ . Entretanto, em algumas situações é possível ter respostas  $y_i = 0$  ou  $y_i = 1$  e nesta situação pode-se utilizar diferentes estratégias. Uma delas é substituir os valores nulos por um número muito próximo de zero  $\epsilon > 0$  e valores unitários por  $1 - \epsilon$ . Alternativamente, quando o máximo e mínimos teóricos,  $\alpha_1$  e  $\alpha_2$ , são conhecidos, pode-se utilizar a seguinte transformação

$$y_i = \frac{(n-1)(y_i - \alpha_1)}{(\alpha_2 - \alpha_1)n} + \frac{1}{2n} \quad (2.32)$$

em que  $n$  é o comprimento do vetor  $\mathbf{y}$ ,  $i = 1, 2, \dots, n$ . Estas aproximações foram consideradas por Smithson e Verkuilen (2006), Zimprich (2010), Verkuilen e Smithson (2011) e Eskelson, Madsen, Hagar e Temesgen (2011).

---

## Capítulo 3

# Regressão Beta com medidas repetidas

---

### 3.1 Introdução

Estudos com unidades experimentais acompanhadas ao longo do tempo são comuns em várias áreas do conhecimento. Em Estatística, esses estudos são denominados *longitudinais*, levando à observação de dados caracterizados por medidas repetidas na mesma unidade experimental. Nesses experimentos, as repetições podem ocorrer em situações ou tempos fixos, ou então, em tempos não necessariamente controlados (AFIUNE *et al.*, 2005). Neste contexto, medidas repetidas na mesma unidade experimental podem estar correlacionadas, não sendo possível a utilização de Modelos Lineares Generalizados (NELDER e WEDDERBURN, 1972) para o tratamento destes dados. Algumas metodologias tem sido propostas para modelagem na presença de correlação, tais como os **modelos lineares generalizados mistos** para modelagem de dados com distribuição na família exponencial, **modelos lineares mistos elípticos** vêm sendo empregados para dados com distribuições elípticas simétricas, **modelos lineares mistos assimétricos** para dados com distribuições elípticas assimétricas. Detalhes podem ser vistos em Arellano-Vale, Bolfarine e Lachos (2005), Lachos, Ghosh e Arellano-Valle (2010), Pinheiro, Liu e Wu (2001) e Savalli, Paula e Cysneiros (2006), entre outros.

No caso em que as respostas estão limitadas em um intervalo finito, como taxas e proporções, e as respostas são independentes, entre as opções para modelagem deste tipo de dado está a distribuição Beta, cujo modelo de regressão foi descrito no capítulo 2, e suas variações: a distribuição simplex (BARNDORFF-JØRGENSEN, 1991) que faz parte dos modelos de dispersão (Jorgensen, 1997) e, mais recentemente, o **modelo de Regressão Beta retangular** (BAYES, BAZÁN E GARCIA, 2012).

Para medidas repetidas com respostas no intervalo aberto  $(0, 1)$ , Venezuela (2008)



propôs uma modelagem via Equações de Estimação Generalizadas (EEG) para as distribuições Beta e Simplex, desenvolvendo também medidas de diagnóstico e de influência local. Manco (2013) propôs modelos de Regressão Beta com efeitos aleatórios normais e não-normais para dados longitudinais, cujos métodos de estimação dos parâmetros e dos efeitos aleatórios são por máxima verossimilhança e pelo método do melhor preditor bayesiano empírico, respectivamente.

A vantagem da utilização da metodologia EEG proposta inicialmente por Liang e Zeger (1986) para a modelagem de medidas repetidas, quando o interesse é apenas na estimação dos efeitos fixos, é que esta exige somente os dois primeiros momentos das distribuições marginais. Em outras palavras, não há necessidade de especificar a distribuição conjunta dos resíduos de cada unidade experimental. Nesta linha, Prentice e Zhao (1991) também propuseram um método para lidar com medidas repetidas utilizando EEG, sendo este método chamado de EEG2. A proposta de Liang e Zeger (1986) passou a ser chamada de EEG1. Ambos decompõe a matriz de variância-covariância de cada unidade experimental isolando os parâmetros de correlação, que são estimados separadamente. A diferença entre EEG1 e EEG2 está na estimação dos parâmetros de correlação. No EEG1 os efeitos fixos são estimados via equações de estimação generalizadas e os parâmetros de correlação são estimados por método dos momentos, enquanto no EEG2 ambos os parâmetros são estimados via equações de estimação generalizadas. Uma discussão completa sobre as duas metodologias encontra-se em Hardin e Hilbe (2003).

Neste capítulo são apresentados o método EEG1 para estudos com medidas repetidas, cujas respostas tem marginalmente uma distribuição Beta com parâmetro de precisão comum a todas as observações, a partir da parametrização de Ferrari e Cribari-Neto (2004). Mais detalhes do método descrito nas seções seguintes podem ser encontrados em Venezuela (2008).

## 3.2 Equações de Estimação

### 3.2.1 Modelagem do parâmetro de posição

Seja  $y_i = (y_{i1}, y_{i2}, \dots, y_{it_i})^\top$  o vetor de respostas da  $i$ -ésima unidade experimental, com  $i = 1, 2, \dots, n$ , e assumamos que a distribuição marginal da variável resposta tenha distribuição

Beta, ou seja,  $y_{ij} \sim \text{Beta}(\mu_{ij}, \phi)$ , com f.d.p. dada por

$$f(y_{ij}; \mu_{ij}, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma((1-\mu_{ij})\phi)} y_{ij}^{\mu_{ij}\phi-1} (1-y_{ij})^{(1-\mu_{ij})\phi-1}. \quad (3.1)$$

No modelo acima o parâmetro de precisão  $\phi$  é constante para todas as observações, semelhante ao caso homocedástico do modelo de Regressão Beta clássico, adicionalmente o modelo admite unidades experimentais com vetor de respostas de dimensões diferentes.

As médias são modeladas por  $g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta}$ , sendo  $g(\cdot)$  umas das funções de ligação estudadas na Seção 2.4,  $\eta_{ij}$  é o preditor linear,  $\mathbf{x}_{ij}^\top$  um vetor de covariáveis referente à  $j$ -ésima observação da  $i$ -ésima unidade experimental, com  $i = 1, 2, \dots, n$  e  $j = 1, \dots, t_i$  e  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ ,  $p < n$ , um vetor de parâmetros desconhecidos a serem estimados.

Uma função de estimação é uma função dos dados e dos parâmetros de interesse, e será dita *ótima* no caso unidimensional se suas raízes tiverem variância assintótica mínima. No caso em que o vetor de parâmetros for multidimensional, este conceito é estendido por meio da introdução de alguma ordenação nas matrizes de covariâncias assintóticas (VENEZUELA, 2008). Para os modelos de Regressão Beta com medidas repetidas, a função de estimação ótima é obtida a partir da definição de vetores

$$\mathbf{b}_i = \mathbf{b}_i(\mathbf{y}_i; \boldsymbol{\beta}) = \mathbf{y}_i^* - \boldsymbol{\mu}_i^*,$$

com  $\mathbf{y}_i^* = (y_{i1}^*, \dots, y_{it_i}^*)^\top$ ,  $\boldsymbol{\mu}_i^* = (\mu_{i1}^*, \dots, \mu_{it_i}^*)^\top$  e  $i = 1, \dots, n$ , com média zero, mutualmente independentes e satisfazendo as mesmas propriedades das funções de estimação regulares. Uma importante vantagem desta definição é que ela tem como caso particular da independência das observações na mesma unidade experimental, a função *score* apresentada por Ferrari e Cribari-Neto (2004).

As equações de estimação para modelos de Regressão Beta considerando a densidade marginal de  $y_{ij}$  em (3.1) e o componente sistemático  $\mathbf{x}_{ij}^\top \boldsymbol{\beta}$  são dadas por

$$\mathbb{E} \left( \frac{\partial \mathbf{b}_i}{\partial \boldsymbol{\beta}^\top} \right)^\top = -\mathbf{X}_i^\top \boldsymbol{\Lambda}_i$$

e

$$\text{Cov}(\mathbf{b}_i) = \text{Var}(\mathbf{y}_i^*)^{1/2} \mathbf{R}(\mathbf{y}_i^*) \text{Var}(\mathbf{y}_i^*)^{1/2} = \mathbf{A}_i^{1/2} \mathbf{R}(\mathbf{y}_i^*) \mathbf{A}_i^{1/2},$$

sendo  $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{it_i})^\top$ ,  $\boldsymbol{\Lambda}_i = \phi \mathbf{G}_i \mathbf{A}_i$ ,  $\mathbf{A}_i = \text{diag}(a_{i1}, \dots, a_{it_i})$  e  $\mathbf{R}(\mathbf{y}_i^*)$  a verdadeira

matriz de correlação de  $\mathbf{y}_i^*$ , com  $a_{ij} = \psi'(\mu_{ij}\phi) + \psi'((1 - \mu_{ij})\phi)$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, t_i$  e  $\mathbf{G}_i = \text{diag}(\partial g^{-1}(\eta_{i1})/\partial \eta_{i1}, \dots, \partial g^{-1}(\eta_{it_i})/\partial \eta_{it_i})$ .

A função de estimação linear ótima de  $\boldsymbol{\beta}$ , com a verdadeira matriz de correlação e  $\phi$  conhecido é equivalente a

$$\Psi_1^o(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{X}_i^\top \boldsymbol{\Lambda}_i \text{Cov}(\mathbf{b}_i)^{-1} (\mathbf{y}_i^* - \boldsymbol{\mu}_i^*). \quad (3.2)$$

Sob condições gerais (VENEZUELA, 2008, Teorema 1, p.13),  $\hat{\boldsymbol{\beta}}$ , solução de  $\Psi_1^o(\boldsymbol{\beta}) = 0$ , é um estimador consistente de  $\boldsymbol{\beta}$  e que

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}_p \left( \mathbf{0}, \lim_{n \rightarrow \infty} n \left\{ \sum_{i=1}^n \mathbf{X}_i^\top \boldsymbol{\Lambda}_i \text{Cov}(\mathbf{b}_i)^{-1} \boldsymbol{\Lambda}_i \mathbf{X}_i \right\}^{-1} \right).$$

Na prática, a verdadeira matriz de correlação  $\mathbf{R}(\mathbf{y}_i^*)$  geralmente é desconhecida. Nesse caso, de acordo com Liang e Zeger (1986) substituindo  $\mathbf{R}(\mathbf{y}_i^*)$  por uma matriz  $\mathbf{R}(\boldsymbol{\alpha})$ , denominada de *matriz de correlação de trabalho*, com as mesmas propriedades de uma matriz de correlação de dimensão  $(t \times t)$  e completamente caracterizada por  $\boldsymbol{\alpha}$ , um vetor  $(s \times 1)$ , não necessariamente igual a verdadeira matriz de correlação. Com isso, a função de estimação generalizada de  $\boldsymbol{\beta}$  é dada por

$$\boldsymbol{\Psi}_1(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{X}_i^\top \boldsymbol{\Lambda}_i \boldsymbol{\Omega}_i (\mathbf{y}_i^* - \boldsymbol{\mu}_i^*) = \sum_{i=1}^n \mathbf{X}_i^\top \mathbf{W}_i \boldsymbol{\Lambda}_i^{-1} \mathbf{b}_i, \quad (3.3)$$

sendo  $\boldsymbol{\Omega}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$  e  $\mathbf{W}_i = \boldsymbol{\Lambda}_i \boldsymbol{\Omega}_i^{-1} \boldsymbol{\Lambda}_i$ .

A função de estimação definida em (3.2) deixa de ser ótima, na presença da matriz  $\mathbf{R}_i(\boldsymbol{\alpha})$ . Nesse caso é necessário que um estimador  $\hat{\boldsymbol{\alpha}}$  de  $\boldsymbol{\alpha}$  seja determinado, de modo que o estimador obtido a partir de (3.3) seja consistente e assintoticamente normal. Além disso, comumente  $\phi$  é desconhecido, e portanto, é necessário estimá-lo.

Assumindo condições gerais e que  $\hat{\boldsymbol{\alpha}}$  seja um estimador  $\sqrt{n}$ -consistente de  $\boldsymbol{\alpha}$ , dados  $\boldsymbol{\beta}$  e  $\phi$ , e  $\hat{\phi}$  é um estimador  $\sqrt{n}$ -consistente de  $\phi$ , dado  $\boldsymbol{\beta}$ , tem-se que

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow \mathcal{N}_p(0, \mathbf{J}^{-1})$$

com  $\mathbf{J} = \lim_{n \rightarrow \infty} \mathbf{J}_n/n$ , sendo  $\mathbf{J}_n$  a matriz de informação de Godambe de  $\boldsymbol{\beta}$  associada à  $\boldsymbol{\Psi}_1$  dada por

$$\mathbf{J}_n = \left\{ \sum_{i=1}^n \mathbf{S}_i \right\} \left\{ \sum_{i=1}^n \mathbf{V}_i \right\}^{-1} \left\{ \sum_{i=1}^n \mathbf{S}_i \right\},$$

em que

$$\sum_{i=1}^n \mathbf{S}_i = \mathbb{E} \left[ \frac{\partial}{\partial \boldsymbol{\beta}^\top} \boldsymbol{\Psi}_1(\boldsymbol{\beta}) \right] = \sum_{i=1}^n \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i$$

$$\sum_{i=1}^n \mathbf{V}_i = \mathbb{E}[\boldsymbol{\Psi}_1(\boldsymbol{\beta}) \boldsymbol{\Psi}_1^\top(\boldsymbol{\beta})] = \sum_{i=1}^n \mathbf{X}_i^\top \boldsymbol{\Lambda}_i \boldsymbol{\Omega}_i^{-1} \text{Cov}(\mathbf{b}_i) \boldsymbol{\Omega}_i^{-1} \boldsymbol{\Lambda}_i \mathbf{X}_i$$

Quando a matriz de correlação de trabalho  $\mathbf{R}_i(\boldsymbol{\alpha})$  é a verdadeira matriz de correlação das respostas transformadas  $\mathbf{y}_i^*$ , então  $\boldsymbol{\Omega}_i = \text{Cov}(\mathbf{b}_i)$ . Assim, um estimador consistente para a matriz de covariâncias de  $\hat{\boldsymbol{\beta}}$ , denominado de *naive* ou *model-based*, é dado por

$$\hat{\mathbf{J}}_n = - \left\{ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\mathbf{W}}_i \mathbf{X}_i \right\}^{-1} \quad (3.4)$$

Caso contrário, estima-se consistentemente a matriz de covariâncias de  $\hat{\boldsymbol{\beta}}$ , usando

$$\hat{\mathbf{J}}_n = \left\{ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\mathbf{W}}_i \mathbf{X}_i \right\}^{-1} \left\{ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\boldsymbol{\Lambda}}_i \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^\top \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\boldsymbol{\Lambda}}_i \mathbf{X}_i \right\} \left\{ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\mathbf{W}}_i \mathbf{X}_i \right\}^{-1}. \quad (3.5)$$

Sendo que  $\hat{\mathbf{W}}_i, \hat{\boldsymbol{\Lambda}}_i, \hat{\boldsymbol{\Omega}}_i$  são obtidos a partir das estimativas de  $\boldsymbol{\beta}, \boldsymbol{\alpha}$  e  $\phi$ . Este estimador é denominado *robusto*, *empírico* ou *sanduiche*.

Para calcular  $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}$  e  $\hat{\phi}$  utiliza-se o método *score* de Fisher para estimar  $\boldsymbol{\beta}$  e o método dos momentos para estimar  $\boldsymbol{\alpha}$  e  $\phi$ . O processo iterativo para o cálculo de  $\hat{\boldsymbol{\beta}}$  a partir da expansão da equação (3.3) em torno de um valor inicial  $\hat{\boldsymbol{\beta}}^{(0)}$  é dado por

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \left\{ \left[ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\mathbf{W}}_i \mathbf{X}_i \right]^{-1} \left[ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\mathbf{W}}_i \hat{\boldsymbol{\Lambda}}_i^{-1} \hat{\mathbf{b}}_i \right] \right\}^{(m)} \quad (3.6)$$

sendo  $m = 0, 1, 2, \dots$  o número de iterações. O índice  $m$  do lado direito da equação acima indica que as matrizes e vetores são atualizados pelas estimativas de  $\boldsymbol{\beta}, \boldsymbol{\alpha}$  e  $\phi$  da  $m$ -ésima iteração. A estimativa  $\hat{\boldsymbol{\beta}}^{(0)}$  é obtida pelo modelo de Regressão Beta assumindo independência entre as observações. O processo iterativo finaliza obtendo-se  $\hat{\boldsymbol{\beta}}$  quando a diferença  $\hat{\boldsymbol{\beta}}^{(m+1)} - \hat{\boldsymbol{\beta}}^{(m)} < \epsilon$  para algum  $\epsilon > 0$  arbitrariamente pequeno, ou quando  $m$  exceder o número máximo de iterações.

Reescrevendo a equação (3.6) na forma de um processo iterativo de mínimos quadrados ponderados, tem-se

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \left\{ \left[ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\mathbf{W}}_i \mathbf{X}_i \right]^{-1} \left[ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\mathbf{W}}_i \mathbf{z}_i \right] \right\}^{(m)} \quad (3.7)$$

sendo  $\mathbf{z}_i = \hat{\boldsymbol{\eta}}_i + \hat{\boldsymbol{\Lambda}}_i^{-1} \hat{\mathbf{b}}_i$ . Demonstra-se que o estimador pelo método dos momentos para  $\phi$  no m-ésimo passo do processo iterativo é dado por

$$\hat{\phi}^{(m)} = \left( \left\{ \sum_{i=1}^n \sum_{j=1}^{t_i} \left( \frac{y_{ij} - \hat{\mu}_{ij}^{(m)}}{\sqrt{\hat{\mu}_{ij}^{(m)}(1 - \hat{\mu}_{ij}^{(m)})}} \right)^2 / \left( \sum_{i=1}^n t_i - p \right) \right\} - 1 \right). \quad (3.8)$$

Para estimar  $\boldsymbol{\alpha}$  é necessário previamente especificar a estrutura da *matriz de correlação de trabalho*  $\mathbf{R}(\boldsymbol{\alpha})$ . Liang e Zeger (1986) introduziram algumas estruturas para  $\mathbf{R}(\boldsymbol{\alpha})$  e apresentaram o estimador  $\hat{\boldsymbol{\alpha}}$  para cada caso. Venezuela (2008), no contexto do modelo de Regressão Beta com medidas repetidas, propôs estimadores via métodos de momentos para as seguintes matrizes: padrão uniforme, auto-regressiva de ordem 1 e não estruturada, que serão descritas na seção seguinte.

### 3.2.2 Matrizes de correlação de Trabalho

Nesta seção são descritas as possíveis estruturas da matriz de correlação de trabalho. Como  $\mathbf{R}(\boldsymbol{\alpha})$  representa a correlação de um mesmo sujeito/grupo, ajustada pelas variáveis explicativas presentes no modelo, os valores que os componentes de  $\boldsymbol{\alpha}$  podem assumir estão no intervalo  $[-1, +1]$  e a dimensão dessa matriz é determinada pelo número de observações feitas em cada sujeito/grupo. A seguir são apresentadas algumas das principais estruturas com seus respectivos estimadores:

- **Padrão uniforme** (*Exchangeable*): também chamada de **permutável**, cujas correlações são as mesmas entre observações no mesmo sujeito/grupo.

$$\text{Corr}(b_{ij}, b_{il}) = \begin{cases} 1, & \text{se } j = l \\ \alpha, & \text{se } j \neq l \end{cases}$$

Como exemplo, para  $t_i = t = 4$ , tem-se

$$\mathbf{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{bmatrix}$$

Para a matriz de correlação padrão uniforme a estimativa de  $\alpha$  no passo  $m$  é dada por

$$\hat{\alpha}^{(m)} = \left\{ \frac{\left( \sum_{i=1}^n \sum_{j>l}^t \hat{b}_{ij} \hat{b}_{il} \right)}{\left( \sum_{i=1}^n \sum_{j=1}^t \hat{b}_{ij}^2 \right)} \frac{2}{(t-1)} \right\}^{(m)}.$$

• **Matriz autoregressiva de ordem M (AR-M)**: estrutura utilizada comumente para repetições ao longo do tempo. Impõe que a correlação é maior entre as observações que distam M repetições, e diminui conforme a distância entre os tempos se afastam de M. Para M=1, ou seja, **AR-1**, a matriz de correlações de trabalho é definida da seguinte forma:

$$\text{Corr}(b_{ij}, b_{il}) = \alpha^{|j-l|}$$

Como exemplo para  $t_i = t = 4$  repetições, tem-se

$$\mathbf{R}(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{bmatrix}$$

Para a matriz de correlação autoregressiva de primeira ordem, AR-1, um estimador simples de  $\alpha$  no passo  $m$  é dado por

$$\hat{\alpha}^{(m)} = \left\{ \frac{\left( \sum_{i=1}^n \sum_{j=1}^{t-1} \hat{b}_{ij} \hat{b}_{i,j+1} \right)}{\left( \sum_{i=1}^n \sum_{j=1}^{t-1} \hat{b}_{ij}^2 \sum_{i=1}^n \sum_{j=2}^t \hat{b}_{ij}^2 \right)^{1/2}} \right\}^{(m)}.$$

Uma deficiência desse estimador é que ele usa somente a informação sobre a dependência de observações consecutivas de  $b_{ij}$ .

• **Não estruturada (Unstructured)**: considera que entre observações distintas no sujeito/grupo as correlações são diferentes. A matriz de correlações é dada por

$$\text{Corr}(b_{ij}, b_{il}) = \begin{cases} 1, & \text{se } j=l \\ \alpha_{jl}, & \text{se } j \neq l \end{cases}$$

Como exemplo para  $t_i = t = 4$  repetições, tem-se

$$\mathbf{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} \\ \alpha_{14} & \alpha_{24} & \alpha_{34} & 1 \end{bmatrix}$$

O estimador de momentos para  $\boldsymbol{\alpha} = (\alpha_{12}, \alpha_{13}, \dots, \alpha_{t-1,t})^\top$ , um vetor com  $t(t-1)/2$  componentes em que  $\alpha_{jl}$  é a correlação entre  $b_{ij}$  e  $b_{il}$ , com  $i = 1, \dots, n$ ,  $j, l = 1, \dots, t$ , e  $j < l$  é dado por

$$\hat{\alpha}_{jl}^{(m)} = \left\{ \frac{\left( \sum_{i=1}^n \hat{b}_{ij} \hat{b}_{il} \right)}{\left( \sum_{i=1}^n \hat{b}_{ij}^2 \right)^{1/2} \left( \sum_{i=1}^n \hat{b}_{il}^2 \right)^{1/2}} \right\}^{(m)}.$$

Ressalta-se que a matriz  $\mathbf{R}(\boldsymbol{\alpha})$  pode ser uma matriz fixa especificada pelo pesquisador. Littel *et al.*(2000) sugerem que para o ajuste das estruturas de variância-covariância deve-se iniciar com estruturas mais parcimoniosas. No caso da  $i$ -ésima unidade experimental não possuir as  $t$  observações utiliza-se a matriz  $\mathbf{R}_i(\boldsymbol{\alpha})$  obtida a partir da matriz  $\mathbf{R}(\boldsymbol{\alpha})$ , retirando-se a linha e a coluna correspondentes às observações faltantes. Quando a unidade experimental é observada em todas as  $t$  ocasiões, tem-se  $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{R}(\boldsymbol{\alpha})$ .

Na prática, o processo de estimação de  $\boldsymbol{\beta}$ ,  $\phi$ , e  $\boldsymbol{\alpha}$  seguirá três etapas:

1. Supondo independência entre observações da mesma unidade experimental, ajusta-se um modelo de regressão linear para as respostas transformadas  $g(\mathbf{y})$  sobre  $\mathbf{X}$  via métodos de mínimos quadrados ordinários, com  $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_n)^\top$  e  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^\top$ . A premissa de independência elimina os parâmetros  $\boldsymbol{\alpha}$  e  $\phi$  do processo de estimação.
2. Ainda, sob independência entre as observações da mesma unidade experimental, ou seja, definindo  $\mathbf{R}(\boldsymbol{\alpha})$  igual a uma matriz identidade, considerou-se as estimativas dos parâmetros de regressão da etapa anterior como valores iniciais,  $(\boldsymbol{\beta}^{(0)})$ , para estimar  $\boldsymbol{\beta}$  e  $\phi$  com as equações (3.7) e (3.8) respectivamente. Essas duas estimativas são calculadas em cada passo do processo iterativo até a convergência de  $\boldsymbol{\beta}$ . Nesta etapa, obtemos as estimativas de  $\boldsymbol{\beta}$  e  $\phi$  considerando um modelo de Regressão Beta para observações independentes da mesma unidade experimental.

3. A partir das estimativas de  $\beta$  e  $\phi$  obtidas na etapa anterior, especifica-se a matriz  $\mathbf{R}(\alpha)$  para ser utilizada na modelagem dos dados. Considerando como  $\hat{\beta}^{(0)}$  e  $\hat{\phi}^{(0)}$  as estimativas de  $\beta$  e  $\phi$  obtidas na etapa 2, calcula-se os parâmetros de correlação ( $\alpha^{(0)}$ ). No passo seguinte do processo iterativo, calcula-se os parâmetros de regressão ( $\hat{\beta}^{(m)}$ ) utilizando a equação (3.7), o parâmetro de precisão ( $\hat{\phi}^{(m)}$ ) e os parâmetros de correlação  $\hat{\alpha}^{(m)}$ . Repete-se este passo até a convergência de  $\beta$ .

As etapas 1 e 2, podem ser substituídas pelo seguinte procedimento: ajustar o modelo de Regressão Beta descrito no capítulo 2, supondo independência das observações na mesma unidade experimental, em que o tempo é considerado como uma covariável, neste caso o método de otimização utilizado será o BFGS combinado com o método *scoring* de Fisher, disponíveis nos principais pacotes estatísticos como SAS e R.

### 3.3 Técnicas de diagnóstico

Após a especificação do modelo faz-se necessário verificar se este é eficiente para descrever a relação entre as covariáveis e a variável resposta. Técnicas de diagnóstico para Modelos Lineares Generalizados (MLG) com observações independentes são apresentadas em Paula (2010). Para detectar pontos de alavanca, observações influentes e discrepantes, são utilizados os elementos da diagonal principal da matriz de projeção (matriz chapéu), a distância de Cook (COOK, 1997) e os resíduos padronizados, além dos gráficos de probabilidade meio normal com envelope simulado.

Venezuela *at al.* (2007) entendeu as técnicas de diagnóstico para MLG com medidas repetidas sob o enfoque das EEG. Venezuela (2008) desenvolveu técnicas de diagnóstico para modelos de Regressão Beta e Simplex com medidas repetidas e a influência local para estes e os modelos da família exponencial, ambos sob a abordagem das EEG. Os desenvolvimentos apresentados nesta seção seguiram de perto os resultados obtidos por Venezuela (2008) para o modelo de Regressão Beta com medidas repetidas sob a homogeneidade da dispersão. Na Seção 3.4 os critérios para seleção de modelos são apresentados.



### 3.3.1 Pontos alavanca, influentes e aberrantes

Para definir o resíduo ordinário, Venezuela (2008) partiu da convergência do processo iterativo dado em (3.7) para estimar  $\boldsymbol{\beta}$ , ou seja,

$$\hat{\boldsymbol{\beta}} = \left[ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\mathbf{W}}_i \mathbf{X}_i \right]^{-1} \left[ \sum_{i=1}^n \mathbf{X}_i^\top \hat{\mathbf{W}}_i \mathbf{z}_i \right] \quad (3.9)$$

sendo  $\mathbf{z}_i = \hat{\boldsymbol{\eta}}_i + \hat{\boldsymbol{\Lambda}}_i^{-1} \hat{\mathbf{b}}_i$ . Para eliminar o somatório em (3.9), define-se  $\hat{\mathbf{W}} = (\hat{\mathbf{W}}_1, \dots, \hat{\mathbf{W}}_n)^\top$  como uma matriz bloco diagonal de dimensão  $(N \times N)$ , sendo  $N = \sum_{i=1}^n t_i$ . Nesse caso, (3.9) pode ser reescrita como

$$\hat{\boldsymbol{\beta}} = \left( \mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X} \right)^{-1} \mathbf{X} \hat{\mathbf{W}} \mathbf{z} \quad (3.10)$$

em que  $\mathbf{X}$  é a matriz de planejamento de dimensão  $(N \times p)$  e  $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)^\top$ .

Desta forma, o resíduo ordinário, que é a diferença entre o valor observado e o ajustado, fica dado por

$$\mathbf{r}_O = \hat{\mathbf{W}}^{1/2} (\mathbf{z} - \hat{\boldsymbol{\eta}}) = (\mathbf{I} - \mathbf{H} \mathbf{W}^{1/2} \mathbf{z}), \quad (3.11)$$

sendo  $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}_1^\top, \dots, \hat{\boldsymbol{\eta}}_n^\top)^\top$ ,  $\mathbf{I}$  a matriz identidade. A equação (3.10) permite mostrar que  $\hat{\boldsymbol{\beta}}$  pode ser interpretado como a solução de mínimos quadrados da regressão linear de  $\hat{\mathbf{W}}^{1/2} \mathbf{z}$  tendo com matriz de planejamento  $\mathbf{W}^{1/2} \mathbf{X}$ . A partir daí, define-se a matriz chapéu ou a matriz de projeção ortogonal, como sendo a matriz bloco diagonal  $(N \times N)$  dada por  $\mathbf{H} = \text{diag}(\mathbf{H}_1, \dots, \mathbf{H}_n)$ , com

$$\mathbf{H}_i = \hat{\mathbf{W}}_i^{1/2} \mathbf{X}_i (\mathbf{X}_i^\top \hat{\mathbf{W}}_i)^{-1} \mathbf{X}_i^\top \hat{\mathbf{W}}_i^{1/2}, \quad (3.12)$$

tendo dimensão  $t_i \times t_i$ ,  $i = 1, \dots, n$ . A matriz  $\mathbf{H}$  pode ser utilizada para detectar pontos de alavanca, tal como em Paula (2010) e Venezuela (2007), para MLG com medidas repetidas. Considere  $\mathbf{h}$  um vetor de dimensão  $(N \times 1)$  cujos elementos  $h_{ij}$ ,  $i = 1, \dots, n$ ;  $j = 1, \dots, t_i$  são os elementos da diagonal principal da matriz  $\mathbf{H}$ . Desta forma, a  $i$ -ésima unidade experimental pode ser um ponto alavanca nos ajustes dos parâmetros  $\boldsymbol{\beta}$  avaliando  $h_{i.} = t_i^{-1} \sum_{j=1}^{t_i} h_{ij}$  com  $i = 1, \dots, n$ . Supondo que a influência de cada ponto seja a mesma, espera-se que  $\text{tr}(\mathbf{H})/N = p/N$ , pois  $\mathbf{H}$  é idempotente. Assim, uma regra prática é considerar um ponto como potencialmente influente se  $h_{ij} \geq 2p/N$ . Analogamente, a  $i$ -ésima unidade experimental é uma observação influente se  $h_{i.} \geq 2p/N$ . Ressalta-se que

os valores  $h_{ij}$  dependem da matriz de pesos  $\hat{\mathbf{W}}_i$ , logo observações com valores altos de  $h_{ij}$  nem sempre são pontos alavanca.

A definição de *resíduo padronizado* associado a observação  $y_{ij}$  com adaptações a partir de Venezuela (2008) é dada por

$$(r_{PD})_{ij} = \frac{\mathbf{e}_{(ij)}^\top \hat{\mathbf{W}}(\mathbf{z} - \hat{\boldsymbol{\eta}})}{\sqrt{1 - h_{ij}}} = \frac{\mathbf{e}_{(ij)}^\top \hat{\mathbf{W}} \hat{\boldsymbol{\Lambda}}^{-1} \hat{\mathbf{b}}}{\sqrt{1 - h_{ij}}} \quad (3.13)$$

em que  $\hat{\boldsymbol{\Lambda}} = (\hat{\boldsymbol{\Lambda}}_1, \dots, \hat{\boldsymbol{\Lambda}}_n)^\top$ ,  $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1^\top, \dots, \hat{\mathbf{b}}_n^\top)^\top$ ,  $e_{ij}$  é um vetor de tamanho conveniente com valor 1 referente a observação  $y_{ij}$  e 0 nas demais posições, e  $h_{ij}$  é o  $j$ -ésimo elemento da diagonal principal de  $\mathbf{H}_i$ , com  $i = 1, \dots, n$  e  $j = 1, \dots, t_i$ . Um ponto é aberrante quando possui um elevado *resíduo padronizado* com um valor relativamente baixo na matriz de projeção  $\mathbf{H}$ . Desta forma, um mesmo ponto dificilmente será aberrante e alavanca.

Um *ponto influente* é definido como aquele que tem grande impacto no valor estimado dos parâmetros do modelo e apresenta um valor alto na matriz  $\mathbf{H}$ . Esta influencia pode ser detectada pela distância de Cook, quando se exclui a observação  $y_{ij}$ , sendo definida por

$$(DC)_{ij} = (r_{PD})_{ij}^2 \frac{h_{ij}}{p(1 - h_{ij})}. \quad (3.14)$$

Desta forma, a distância de Cook para a  $i$ -ésima unidade experimental é dada por  $(DC)_i = \sum_{j=1}^{t_i} (DC)_{ij}$ , com  $i = 1, \dots, n$ . De acordo com Venezuela (2008) essas medidas podem não ter acurácia se a matriz de correlação de trabalho,  $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ , divergir demasiadamente da verdadeira.

### 3.3.2 Influência local

Um método muito conhecido na detecção de pontos influentes no modelo é o método de influência local proposto por Cook (1986), que avalia a sensibilidade dos estimadores de máxima verossimilhança com respeito a perturbações no modelo ou nos dados. Perturbação é qualquer modificação nas suposições do modelo ou nos dados que causam diferenças substanciais nos resultados da análise. Os esquemas mais conhecidos são: perturbação de casos ponderados, perturbação aditiva na resposta, perturbação aditiva no preditor, entre outros.

### 3.4 Seleção de modelos

Pan (2001) propôs uma medida denominada *QIC* que é uma modificação na medida *AIC*. Assim esta medida é útil tanto para seleção de modelos como para escolha da estrutura de uma matriz de correlação de trabalho  $\mathbf{R}$ . Pode ser escrita, com base no logaritmo da função de verossimilhança, da seguinte forma:

$$QIC(\mathbf{R}) = -2l\left(\hat{\boldsymbol{\beta}}(\mathbf{R})\right) + 2tr(\hat{S}_{\mathbf{I}}\hat{\mathbf{J}}_{n\mathbf{R}}^{-1}),$$

sendo  $l(\cdot)$  o logaritmo da função de verossimilhança que gera a função *escore* equivalente a equação de estimação quando assumimos independência entre as observações,  $\hat{S}_{\mathbf{I}}$  a matriz de sensibilidade sob a estrutura de independência e  $\hat{\mathbf{J}}_{n\mathbf{R}}^{-1}$  o estimador robusto sob a estrutura  $\mathbf{R}$  proposta e são avaliadas em  $\hat{\boldsymbol{\beta}}(\mathbf{R})$  que é a estimativa de  $\boldsymbol{\beta}$  obtida com a matriz de correlação de trabalho  $\mathbf{R}$ .

Mais recentemente, Hin e Wang (2009) propuseram uma modificação do *QIC*, o *Critério de Informação de Correlação* (*Correlation Information Criterion - CIC*) para aperfeiçoar seu desempenho na escolha da estrutura de correlação de trabalho.

---

## Capítulo 4

# Aplicação I: Índice de Neutralização em Transformadores

---

### 4.1 Introdução

Transformadores ou Trafos são equipamentos de extrema importância nos sistemas de conversão e distribuição de energia elétrica e estão presentes desde a planta geradora, elevando a tensão para níveis adequados à transmissão de longas distâncias, até a distribuição, reduzindo a tensão para níveis de consumo residencial (ilustrado na Figura 4.1). Tal aplicabilidade faz dos transformadores os equipamentos mais importantes do sistema elétrico de potência. Um componente básico do transformador de corrente elétrica é o óleo, que promove a refrigeração e o isolamento dos circuitos elétricos e magnéticos.

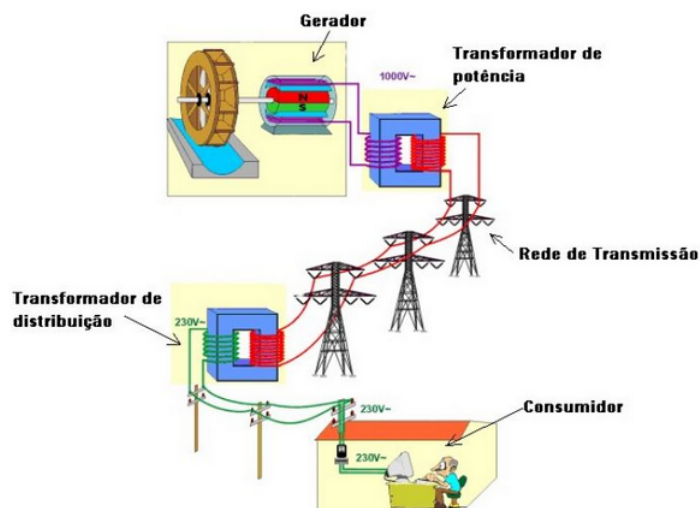


Figura 4.1 *Forma simplificada da distribuição de energia*

## 4.2 Sistema Isolante em Transformador

Segundo Morais (2004), a isolação líquida e a isolação sólida de um transformador desempenham papel de fundamental importância para as condições ideais de operacionalidade do equipamento. O isolamento líquido, em sua grande maioria, é constituído pelo óleo mineral isolante-OMI e o isolamento sólido pelo papel isolante. Tanto o isolamento sólido quanto o líquido apresentam processos normais de degradação, ou seja, a partir do momento que o transformador é energizado, inicia-se o envelhecimento, ou a degradação, do sistema de isolação óleo-papel (FERNANDES, 1987).

### 4.2.1 Isolação Sólida

O isolante sólido mais usado em transformadores é de natureza celulósica, ou seja, de papel, os principais materiais celulósicos utilizados são o papel Kraft, que é feito de fibra de madeira; o papel manilha, feito de fibras de madeira e cânhamo; o papelão kraft, também de fibra de madeira e o pressboard, que é um papelão com fibra de algodão. O papel Kraft tem sido muito utilizado nos transformadores como isolante, por ter boas propriedades elétricas e mecânicas quando imerso em OMI. Além disso, é de fácil obtenção e seu custo é baixo. O envelhecimento, ou a degradação, do papel provoca a perda das características mecânicas e dielétricas do equipamento. Este envelhecimento da isolação sólida se inicia a partir do momento em que o transformador é energizado. Produtos de caráter ácido juntamente com água provocam a degradação acelerada da celulose, e a borra originada deste processo prejudica a transferência de calor do papel para o óleo (FERNANDES, 1987). Na maioria dos casos, o papel isolante é o elemento responsável pela definição da vida útil do equipamento elétrico, uma vez que, diferente do isolante líquido, não pode ser regenerado ou substituído ao longo da operação do transformador (BATISTA *et al.*, 2005).

### 4.2.2 Isolação Líquida

O isolante líquido mais utilizado nos transformadores é o OMI, obtido pela destilação do petróleo natural, na fração de 300 a 400 °C. Este óleo é uma mistura de compostos na qual a maioria das moléculas é constituída por carbono e hidrogênio, denominados de

hidrocarbonetos. Sua origem pode ser parafínica ou naftênica (CASTRO JR., 2005). Esse fluido mineral desempenha uma função importante no isolamento e no resfriamento das partes ativas de um transformador, e, para seu perfeito funcionamento, suas características físico-químicas devem ser preservadas.

### **4.3 Características Físico-Químicas para Monitorar a Condição do OMI**

Nos transformadores, o líquido isolante circula em contato com as diversas partes do equipamento e, desta forma, a análise de suas características físico-químicas permite realizar medidas preventivas sobre o óleo e sobre o próprio transformador, já que, alterações destas características, podem indicar não só o envelhecimento do líquido isolante, como sua contaminação. Os parâmetros monitorados para controlar a condição do OMI em uso são a Tensão interfacial (dina/cm), o Teor de água (ppm), a Cor, o Fator de potência (%), o Índice de neutralização (mg KOH/g), a Rigidez dielétrica (kV) e a Densidade. Estes parâmetros, apresentados a seguir, traduzem as condições imediatas do óleo em uso.

#### **4.3.1 Cor**

O óleo isolante novo tradicionalmente é amarelo pálido e límpido, isento de materiais em suspensão. A cor é geralmente aceita como um índice do grau de refino. À medida que o óleo vai deteriorando, sua cor muda tornando-se mais escura. O número referente à cor, estando elevado, representa envelhecimento, contaminação, deterioração, presença de decomposição de arcos elétricos.

#### **4.3.2 Tensão Interfacial**

A tensão interfacial é uma medida indireta da concentração de substâncias polares presentes no OMI. Estas substâncias prejudicam as propriedades dielétricas do óleo, além de contribuir para seu envelhecimento precoce (MILASCH, 1984). A tensão interfacial no sistema água-óleo visa exatamente inferir sobre a quantidade de substâncias polares presentes no óleo, uma vez que tais substâncias tendem a concentrar-se na interface do

sistema. Quanto maior a concentração de substâncias polares, menor o valor da tensão interfacial do OMI.

### **4.3.3 Teor de Água**

O teor de água é medido em partes por milhão (ppm), ou seja, expressa a quantidade de litros de água para cada um milhão de litros de óleo (MILASCH,1984) A quantidade de umidade presente nos líquidos isolantes depende da forma como este líquido foi manipulado e de sua estrutura química, isto é, fluido polar, como, por exemplo, o óleo vegetal isolante apresenta maior afinidade com a água do que o fluido apolar, neste caso o OMI (WILHELM, 2008).

### **4.3.4 Índice de Neutralização**

O índice de neutralização é a medida da quantidade de hidróxido de potássio (KOH) necessária para neutralizar os constituintes ácidos presentes em 1g de óleo. Os compostos ácidos a partir de certa concentração reagem com materiais de construção do equipamento, principalmente com o papel do isolamento sólido. Estes compostos podem ainda formar borra, dificultando a transferência de calor para o meio ambiente, depositando-se na parte ativa ou nos trocadores de calor. Isto faz com que diminua a vida útil do equipamento. Para o óleo em uso, o índice de neutralização é uma medida indireta do grau de oxidação do óleo.

### **4.3.5 Rigidez Dielétrica**

A rigidez elétrica é a tensão alternada medida em kV, na qual ocorre a descarga disruptiva na camada de óleo situada entre dois eletrodos e em condições perfeitamente determinadas. Mede a capacidade de um líquido isolante resistir ao impacto elétrico sem falhar. É considerada uma medida indireta das impurezas contidas no óleo (MILASCH, 1984). No caso de óleo novo pode ser usado para avaliar o processo de secagem, ao qual o óleo foi submetido, sendo um indicativo das condições de manuseio, transporte e armazenagem. Para óleo em serviço, a verificação da rigidez é muito importante para avaliar a função isolante do líquido.

### 4.3.6 Fator de Potência

O fator de potência do óleo aumenta com a temperatura e com a quantidade de substâncias polares provenientes da deterioração do óleo. O fator de potência dá uma ideia da intensidade da corrente que flui pelo óleo, e que é uma medida de sua contaminação e de sua deterioração (MILASCH,1984). No OMI as perdas dielétricas estão relacionadas com a presença de compostos polarizáveis, oriundos dos diversos materiais constituintes do equipamento ou do processo oxidativo no meio hidrocarbônico. Sendo assim, o acompanhamento da variação do fator de potência pode determinar a qualidade do sistema em termos do grau de envelhecimento do óleo. O fator de perdas dielétricas é uma medida de qualidade do dielétrico. Quanto menor, mais próximo este se encontra de um sistema ideal.

### 4.3.7 Densidade

É a relação de massas de uma determinada substância OMI e outra substância do mesmo volume (em geral, água à temperatura de 4°C). Indica uma característica intrínseca do óleo isolante.

## 4.4 Cromatografia Gasosa

A análise cromatográfica dos gases dissolvidos no óleo isolante, denominada como cromatografia Gasosa (análise dos gases dissolvidos em óleo) é utilizada para determinar previamente defeitos ou falhas em desenvolvimento em equipamentos imersos em óleo isolante. Desta forma, primeiro passo para o diagnóstico de anomalias em transformadores é feito a partir da amostra de óleo isolante, fundamental para subsidiar decisões de operação e manutenção de unidades transformadoras.

- A ocorrência de um arco elétrico no óleo isolante dá origem a grandes quantidades de Hidrogênio ( $H_2$ ) e acetileno ( $C_2H_2$ ), com pequenas quantidades de metano ( $CH_4$ ) e etileno ( $C_2H_4$ ).
- Descargas elétricas de baixa energia produzem hidrogênio e metano com pequenas quantidades de etano ( $C_2H_6$ ) e etileno ( $C_2H_4$ ).



- Caso a falha atinja a celulose, em ambos os casos também haverá a formação de monóxido e dióxido de carbono (CO e  $CO_2$ )

## 4.5 Descrição e estudo exploratório dos dados

O conjunto de dados utilizado nesta seção é constituído de 825 observações, e refere-se a um estudo sobre as características físico-químicas e a cromatografia gasosa, utilizadas para monitorar a condição do óleo isolante nos transformadores. Estes dados foram cedidos pela Eletrobrás Eletronorte (Centrais Elétricas do Norte do Brasil S.A. Eletronorte), uma concessionária de serviço público que fornece energia elétrica para nove estados da Amazônia Legal, dentre eles o Estado do Pará.

Analizou-se esse conjunto de dados fazendo uso do modelo de regressão Beta descrito no Capítulo 3. Como a variável resposta, o índice de neutralização, apresentou valores discrepantes (ver Figura 4.1), selecionou-se uma amostra aleatória estratificada por tipo de transformador com tamanho amostral ( $n = 222$ ) observações, ao nível de confiança de 95% baseado na tabela de tamanhos amostrais publicada por Arkin e Colton (1961), *apud* Gil, que fornece o tamanho da amostra adequada para um nível de confiança de 95%, o restante das observações foi utilizado para avaliar a capacidade preditiva dos modelos em estudo, ou seja, construiu-se um esquema de validação dos modelos. As seguintes características serão utilizadas para a modelagem, acompanhadas de suas representações (em negrito) no modelo definido na seção seguinte:

- Características físico-químicas
  - *Índice de neutralização* (em MG KOH/g óleo)
  - Indicadora do tipo autotransformador -  $\mathbf{x}_1$

$$AT_t = \begin{cases} 1 & \text{se a } t\text{-ésima observação é um autotransformador} \\ 0 & \text{caso contrário;} \end{cases}$$

- Indicadora de reator -  $\mathbf{x}_2$

$$RE_t = \begin{cases} 1 & \text{se a } t\text{-ésima observação é um reator} \\ 0 & \text{caso contrário;} \end{cases}$$

- *Cor* -  $\mathbf{x}_3$
- *Rigidez Dielétrica* (em kV) -  $\mathbf{x}_4$
- *Teor de Água* (em partes por milhão-ppm) -  $\mathbf{x}_5$
- *Densidade* -  $\mathbf{x}_6$
- *Tensão Interfacial* -  $\mathbf{x}_7$
  
- Cromatografia gasosa
  - *FP25* (Fator de potência a 25°C) -  $\mathbf{x}_8$
  - *FP100* (Fator de potência 100°C) -  $\mathbf{x}_9$
  - $H_2$  (Gás hidrogênio) -  $\mathbf{x}_{10}$
  - $O_2$  (Gás oxigênio) -  $\mathbf{x}_{11}$
  - $N_2$  (Gás nitrogênio) -  $\mathbf{x}_{12}$
  - $CH_4$  (Gás metano) -  $\mathbf{x}_{13}$
  - $CO$  (Monóxido de carbono) -  $\mathbf{x}_{14}$
  - $CO_2$  (Dióxido de carbono) -  $\mathbf{x}_{15}$
  - $C_2H_4$  (Gás Etileno) -  $\mathbf{x}_{16}$
  - $C_2H_6$  (Gás etano) -  $\mathbf{x}_{17}$
  - $C_2H_2$  (Gás acetileno) -  $\mathbf{x}_{18}$

Recomenda-se, antes de qualquer modelagem, um estudo exploratório dos dados. Observa-se na Tabela 4.1 que somente as variáveis *Rigidez Dielétrica*, *Tensão Interfacial*, *Densidade* e  $N_2$  tem coeficiente de variação (C.V.) menor que 70%, indicando que existe alta variabilidade nas outras variáveis, particularmente na variável  $C_2H_2$  que apresenta C.V. acima 1.000

Tabela 4.1 *Medidas descritivas das características físico-químicas e da cromatografia*

Variável	Mínimo	1º Quartil	Média	C.V.(%)	3º Quartil	Máximo
Cor	0,10	0,30	1,32	82,64	1,90	6,50
Neutralização	0,00	0,01	0,02	119,89	0,02	0,16
Rigidez Dielétrica	12,60	65,63	77,53	31,07	95,90	100,20
FP25	0,01	0,02	0,07	176,95	0,07	1,20
FP100	0,07	0,30	1,79	175,15	1,92	29,89
Tensão Interfacial	15,60	33,03	38,06	18,83	43,98	51,30
Teor de Água	3,00	8,25	16,97	74,95	21,75	77,00
Densidade	0,85	0,87	0,88	1,12	0,88	0,89
$H_2$	0	0	41,29	432,29	27,75	2446,00
$O_2$	233,00	2520,50	7980,74	94,90	11145,75	48767,00
$N_2$	2603,00	34317,50	49192,21	51,48	64697,25	180062,00
$CH_4$	0	2,00	37,68	196,85	47,00	872,00
CO	2,00	66,75	288,93	120,38	373,25	2709,00
$CO_2$	12,00	706,00	2699,94	92,58	4022,25	13592,00
$C_2H_4$	0	0	15,25	365,01	10,75	757,00
$C_2H_6$	0	0	38,68	226,11	35,50	676,00
$C_2H_2$	0	0	1,81	1113,88	0	299,00

A partir da Figura 4.2 tem-se que a distribuição do índice de neutralização é assimétrica à direita, assumindo valores no intervalo  $(0, 1)$ , e pode ser modelado por uma distribuição Beta. Analisando sua distribuição por tipo de transformador, a assimetria persiste com possíveis pontos discrepantes principalmente nos transformadores.

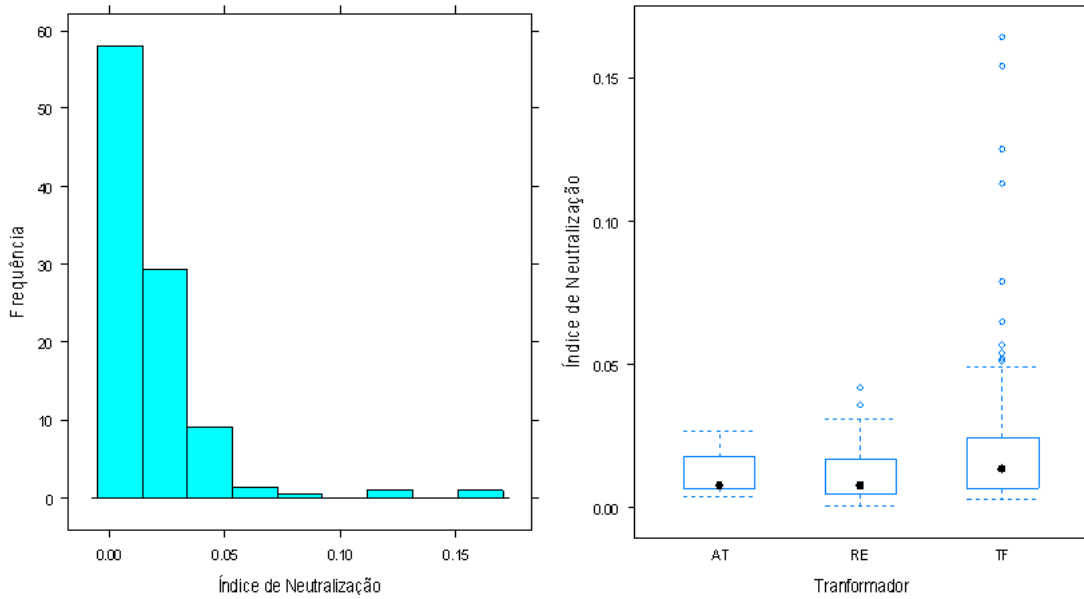


Figura 4.2 Distribuição do índice de neutralização geral e por tipo de transformador.

## 4.6 Ajuste do modelo de Regressão Beta

As observações  $y_t, t = 1, 2, \dots, 222$ , referentes aos índices de neutralização, serão consideradas independentes e seguindo uma distribuição Beta com média  $\mu_t$  e parâmetro de precisão  $\phi$  desconhecido. Inicialmente, modela-se a média segundo a relação

$$g(\mu_t) = \mathbf{x}_t^\top \boldsymbol{\beta}, \quad (4.1)$$

onde  $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \dots, \beta_{18})$  e  $\mathbf{x}_t^\top = (1, x_{t1}, x_{t2}, \dots, x_{t18})$ . Para o ajuste do modelo de regressão Beta especificado em (4.1), utilizou-se a função de ligação logito. Os resultados da estimação são apresentados na Tabela 4.2. Observa-se que o nível descritivo marginal das variáveis explicativas: *Cor*, *Rigidez Dielétrica*, *Teor Água*, *Tensão Interfacial*,  $O_2$ ,  $N_2$ ,  $CH_4$ ,  $CO$  e  $C_2H_4$  estão próximos de zero, indicando que essas variáveis são estatisticamente importantes para explicar o comportamento da variável resposta “Índice de neutralização”. Em todos os testes foram considerados níveis de significância de 5%.

Tabela 4.2 *Estimativas dos parâmetros do modelo de regressão Beta*

Coeficientes	Estimativa	EP	Valor p
$\beta_0$ - Intercepto	-5,027	2,773	0,070
$\beta_1$ - Tipo - RE	-0,105	0,125	0,399
$\beta_2$ - Tipo - TF	0,012	0,117	0,920
$\beta_3$ - Cor	0,079	0,038	0,037
$\beta_4$ - Rigidez dielétrica	0,008	0,002	0,000
$\beta_5$ - Teor de água	0,014	0,003	0,000
$\beta_6$ - Densidade	3,134	3,206	0,328
$\beta_7$ - Tensão interfacial	-0,078	0,006	0,000
$\beta_8$ - FP25	0,435	0,731	0,552
$\beta_9$ - FP100	0,016	0,031	0,613
$\beta_{10}$ - $H_2$	0,000	0,000	0,229
$\beta_{11}$ - $O_2$	0,000	0,000	0,018
$\beta_{12}$ - $N_2$	0,000	0,000	0,004
$\beta_{13}$ - $CH_4$	0,003	0,001	0,023
$\beta_{14}$ - CO	0,001	0,000	0,000
$\beta_{15}$ - $CO_2$	0,000	0,000	0,254
$\beta_{16}$ - $C_2H_4$	-0,004	0,001	0,002
$\beta_{17}$ - $C_2H_6$	0,001	0,000	0,306
$\beta_{18}$ - $C_2H_2$	0,000	0,002	0,816
$\phi$	291,901	28,442	0,000

Aplicando algumas técnicas de diagnóstico, foram construídos os gráficos apresentados na Figura 4.3. Na Figura 4.3 (a) é apresentado o gráfico dos resíduos padronizados ponderados versus os índices das observações, enquanto que na Figura 4.3(b) temos o gráfico dos  $C_t$  versus  $t$ . O gráfico dos resíduos padronizados ponderados versus o preditor linear ( $\hat{\eta}_t$ ) é apresentado na Figura 4.3 (c), enquanto que na Figura 4.3 (d) temos o gráfico dos elementos da diagonal de  $GL(\hat{\beta}, \hat{\phi})$  versus  $\mu_t$ .

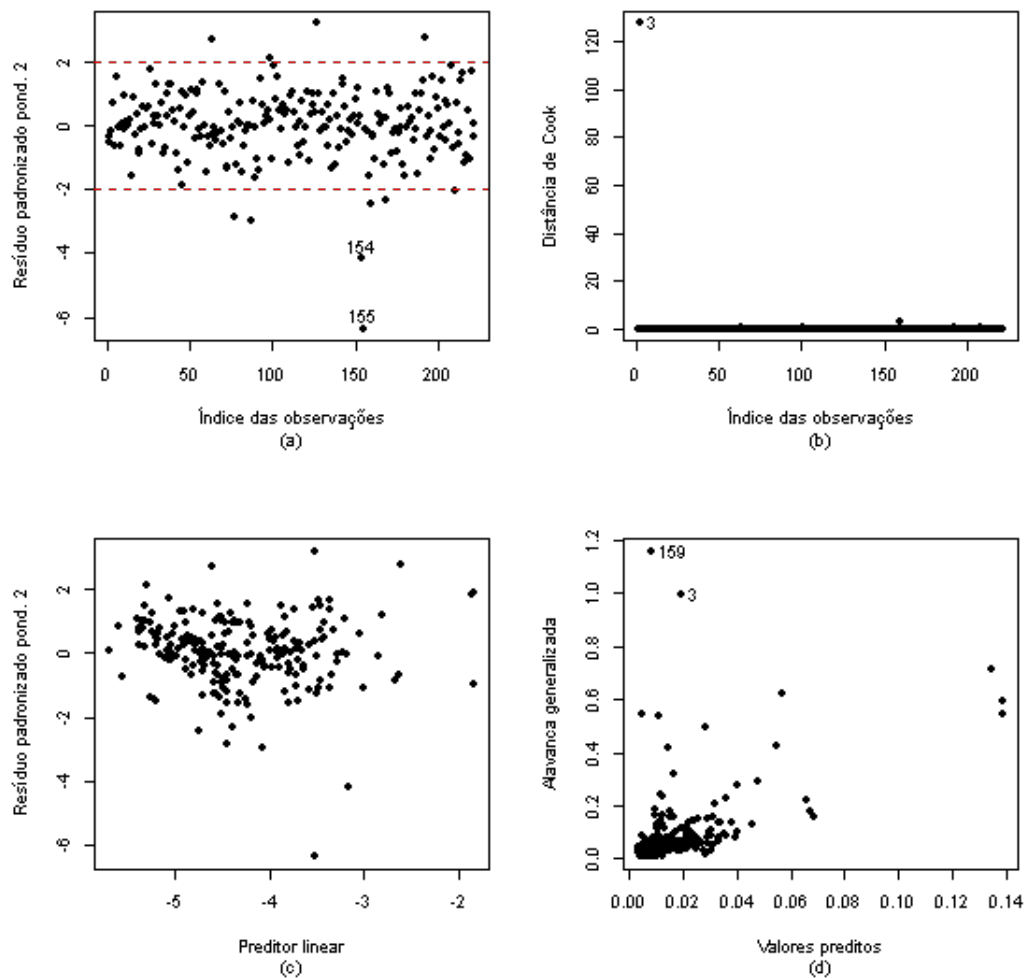


Figura 4.3 Gráficos de diagnóstico do modelo de regressão Beta ajustado para os dados de índice de neutralização.

Uma inspeção na Figura 4.3 (a) mostra que as observações 154 e 155 aparecem como possíveis valores aberrantes. Em relação à distância de Cook temos que  $C_3$  pode ser considerado um possível ponto influente. No gráfico de alavanca generalizada, nota-se que as observações 3 e 159 aparecem como possíveis pontos de alavanca.

Na Figura 4.4, nota-se a ocorrência de alguns pontos fora da banda de confiança, indicando a necessidade de avaliação da influencia de tais pontos no ajuste do modelo.

As observações influentes detectadas pelas técnicas de diagnóstico correspondem ao Tipo

autotransformador-AT (observação 3) e ao Tipo Transformador-TF (observações 154, 155 e 159).

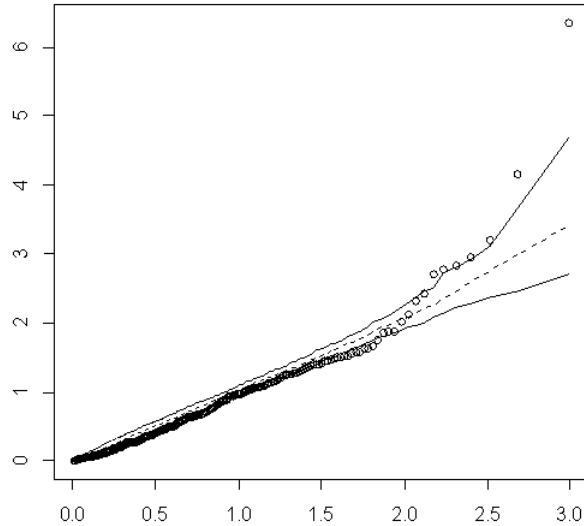


Figura 4.4 Gráfico de Probabilidade Meio-Normal com Envelope Simulado.

Vamos analisar a influencia destas observações nas estimativas dos parâmetros, retirando uma a uma, conjuntamente e combinadas entre si. Para avaliar a magnitude do impacto exercido pelas observações utilizamos as seguintes medidas (MARTINEZ, 2008):

**Desvio relativo percentual (DRP):**

$$DRP = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{\theta}} \times 100\%,$$

onde  $\hat{\theta}$  é o estimador de  $\theta$  obtido usando todas as observações e  $\hat{\theta}^*$  é o estimador obtido ao retirar uma ou várias (dependendo do caso) observações atípicas do conjunto de todas as observações.

**Desvio quadrático médio relativo percentual (DQMRP):**

$$DQMRP = \frac{DQM^* - DQM}{DQM} \times 100\%,$$

em que  $DQM = \sqrt{(1/n) \sum_{j=1}^n (y_j - \hat{y}_j)^2}$  e  $DQM^*$  é o valor de  $DQM$  obtido ao retirar

uma ou várias (dependendo do caso) observações atípicas. Com esta medida avalia-se o ganho relativo de poder preditivo do modelo sem as observações, tomando como base o modelo com todas as observações, note que se  $DQMRP$  for negativo isto indica que o modelo possui um desvio quadrático médio menor e, portanto, por este critério é preferível.

**Erro absoluto médio relativo percentual (EAMRP):**

$$EAMRP = \frac{EAM^* - EAM}{EAM} \times 100\%,$$

em que  $EAM = (1/n) \sum_{j=1}^n |y_j - \hat{y}_j|$  e  $DQM^*$  e  $EAM^*$  é o valor de  $EAM$  obtido ao retirar uma ou várias (dependendo do caso) observações atípicas. A interpretação é a mesma da medida anterior ( $DQMRP$ ), todavia o Erro absoluto Médio ( $EAM$ ) é, menos sensível às observações atípicas no conjunto de dados.

A Tabela 4.3 mostra as medidas de impacto  $DQMRP$  e  $EAMRP$  para os valores ajustados, considerando a retirada individual das observações atípicas 3, 154, 155 e 159, e a retirada conjunta, considerando todas as combinações possíveis. Nota-se que a retirada individual das observações 3, 154 e 159 e das combinações duas a duas, (3 e 159) e (3 e 154), levam aos menores valores das medidas de  $DQMRP$  e  $EAMRP$ .

Entre os pontos retirados, a observação 3 foi a que menos impactou nos resíduos do modelo, indicando que o poder preditivo do modelo é pouco influenciado por este ponto. O próximo passo é estudar a qualidade e poder preditivo dos 5 modelos sob estudo.

A Tabela 4.4 apresenta as estimativas, erros-padrão, valores  $p$  e o desvio relativo percentual ( $DRP$ ) dos parâmetros estimados retirando a observação 3. Nota-se que as estimativas dos parâmetros não sofrem grande impacto, exceto o coeficiente de  $C_2H_2$  (possível valor discrepante detectado na análise descritiva) e os resultados inferenciais do modelo não são alterados.



Tabela 4.3 *Medida de impacto DQMRP e EAMRP para os valores ajustados*

Observações retiradas	DQMRP (%)	EAMRP (%)
( 1 ) 3	-0,25	-0,35
( 2 ) 154	3,47	3,10
( 3 ) 155	4,64	4,50
( 4 ) 159	1,69	2,90
( 5 ) 3 154	3,22	2,74
( 6 ) 3 155	4,43	4,15
( 7 ) 3 159	1,47	2,56
( 8 ) 154 155	9,14	7,71
( 9 ) 154 159	5,37	5,89
( 10 ) 155 159	6,13	7,48
( 11 ) 3 154 155	8,92	7,37
( 12 ) 3 154 159	5,14	5,54
( 13 ) 3 155 159	5,97	7,19
( 14 ) 154 155 159	10,87	10,88
( 15 ) 3 154 155 159	10,71	10,60

Tabela 4.4 *Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem a observação*

3

Coeficientes	Estimativa	EP	Valor	DRP (%)
$\beta_0$ - Constante	-5,05E+00	2,78E+00	0,07	0,56
$\beta_1$ - Tipo-RE	-1,06E-01	1,25E-01	0,40	0,99
$\beta_2$ - Tipo-TF	1,24E-02	1,17E-01	0,92	5,23
$\beta_3$ - Cor	7,94E-02	3,82E-02	0,04	0,50
$\beta_4$ - Rigidez dielétrica	7,91E-03	2,12E-03	0,00	-0,21
$\beta_5$ - Teor de Água	1,35E-02	3,38E-03	0,00	-0,10
$\beta_6$ - Densidade	3,16E+00	3,21E+00	0,32	0,94
$\beta_7$ - Tensão interfacial	-7,77E-02	6,02E-03	< 0,01	-0,06
$\beta_8$ - FP25	4,32E-01	7,33E-01	0,56	-0,60
$\beta_9$ - FP100	1,60E-02	3,14E-02	0,61	0,68
$\beta_{10}$ - $H_2$	-2,75E-04	2,29E-04	0,23	0,13
$\beta_{11}$ - $O_2$	1,50E-05	6,34E-06	0,02	0,70
$\beta_{12}$ - $N_2$	-5,98E-06	2,07E-06	< 0,01	0,22
$\beta_{13}$ - $CH_4$	2,71E-03	1,19E-03	0,02	1,34
$\beta_{14}$ - CO	4,88E-04	1,28E-04	< 0,01	-0,05
$\beta_{15}$ - $CO_2$	-1,76E-05	1,56E-05	0,26	-0,78
$\beta_{16}$ - $C_2H_4$	-4,14E-03	1,35E-03	< 0,01	1,65
$\beta_{17}$ - $C_2H_6$	4,47E-04	4,42E-04	0,31	-1,05
$\beta_{18}$ - $C_2H_2$	4,06E-03	2,04E-02	0,84	1.026,67
$\phi$	2,91E+02	2,84E+01	< 0,01	-0,45

Na Tabela 4.5, para o modelo sem a observação 154, a variável  $O_2$  deixa de ser significativa. As maiores variações relativas (DRP) nas estimativas foram nos coeficientes de FP100 (-69,72%) e  $C_2H_6$  (-49,85%), ambas não estatisticamente significativas com ou sem este ponto.

Tabela 4.5 *Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem a observação 154*

Coeficientes	Estimativa	EP	Valor	DRP (%)
$\beta_0$ - Constante	-4,91E+00	2,68E+00	0,07	-2,28
$\beta_1$ - Tipo-RE	-9,78E-02	1,22E-01	0,42	-6,94
$\beta_2$ - Tipo-TF	9,32E-03	1,14E-01	0,93	-20,91
$\beta_3$ - Cor	9,24E-02	3,69E-02	0,01	16,87
$\beta_4$ - Rigidez dielétrica	6,72E-03	2,06E-03	< 0,01	-15,17
$\beta_5$ - Teor de Água	1,24E-02	3,27E-03	< 0,01	-8,12
$\beta_6$ - Densidade	3,18E+00	3,10E+00	0,31	1,42
$\beta_7$ - Tensão interfacial	-8,01E-02	5,86E-03	< 0,01	3,03
$\beta_8$ - FP25	6,35E-01	7,03E-01	0,37	46,03
$\beta_9$ - FP100	4,80E-03	3,02E-02	0,87	-69,72
$\beta_{10}$ - $H_2$	-3,03E-04	2,24E-04	0,18	10,07
$\beta_{11}$ - $O_2$	1,15E-05	6,17E-06	0,06	-22,59
$\beta_{12}$ - $N_2$	-4,31E-06	2,04E-06	0,03	-27,83
$\beta_{13}$ - $CH_4$	3,08E-03	1,14E-03	0,01	15,30
$\beta_{14}$ - CO	4,49E-04	1,24E-04	< 0,01	-7,96
$\beta_{15}$ - $CO_2$	-2,20E-05	1,52E-05	0,15	24,02
$\beta_{16}$ - $C_2H_4$	-4,42E-03	1,28E-03	< 0,01	8,67
$\beta_{17}$ - $C_2H_6$	2,27E-04	4,38E-04	0,60	-49,85
$\beta_{18}$ - $C_2H_2$	3,44E-04	1,51E-03	0,82	-4,48
$\phi$	3,10E+02	3,02E+01	< 0,01	6,16

Retirando as observações 3 e 154 da amostra, a variável  $O_2$  deixa de ser significativa conforme a Tabela 4.6. As maiores variações relativas (DRP) nas estimativas quando retirada as observações 154 e 3 foram  $FP100$  e  $C_2H_6$  respectivamente.

Tabela 4.6 *Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem as observações 3 e 154*

Coeficientes	Estimativa	EP	Valor	DRP (%)
$\beta_0$ - Constante	-4,95E+00	2,69E+00	0,07	-1,53
$\beta_1$ - Tipo-RE	-9,93E-02	1,22E-01	0,42	-5,54
$\beta_2$ - Tipo-TF	1,02E-02	1,14E-01	0,93	-13,57
$\beta_3$ - Cor	9,29E-02	3,71E-02	0,01	17,56
$\beta_4$ - Rigidez dielétrica	6,70E-03	2,07E-03	0,00	-15,49
$\beta_5$ - Teor de Água	1,24E-02	3,28E-03	0,00	-8,29
$\beta_6$ - Densidade	3,22E+00	3,11E+00	0,30	2,71
$\beta_7$ - Tensão interfacial	-8,01E-02	5,87E-03	0,00	2,98
$\beta_8$ - FP25	6,32E-01	7,05E-01	0,37	45,27
$\beta_9$ - FP100	4,94E-03	3,02E-02	0,87	-68,85
$\beta_{10}$ - $H_2$	-3,03E-04	2,25E-04	0,18	10,31
$\beta_{11}$ - $O_2$	1,17E-05	6,21E-06	0,06	-21,65
$\beta_{12}$ - $N_2$	-4,32E-06	2,04E-06	0,03	-27,53
$\beta_{13}$ - $CH_4$	3,13E-03	1,16E-03	0,01	17,20
$\beta_{14}$ - CO	4,49E-04	1,25E-04	0,00	-8,03
$\beta_{15}$ - $CO_2$	-2,18E-05	1,52E-05	0,15	22,98
$\beta_{16}$ - $C_2H_4$	-4,52E-03	1,32E-03	0,00	11,03
$\beta_{17}$ - $C_2H_6$	2,20E-04	4,39E-04	0,62	-51,34
$\beta_{18}$ - $C_2H_2$	5,73E-03	2,00E-02	0,77	1.490,56
$\phi$	3,09E+02	3,02E+01	0,00	5,70

Retirando somente a observação 159 da amostra, a Cor deixa ser significativa para explicar o *Índice de Neutralização* e  $H_2$  passa a ser significativo, ou seja, as conclusões inferenciais foram alteradas, conforme a Tabela 4.7. Os coeficientes de regressão mais afetados pela retirada deste ponto são:  $H_2$ ,  $C_2H_6$  e  $C_2H_2$ .

Tabela 4.7 *Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem a observação 159*

Coeficientes	Estimativa	EP	Valor	DRP (%)
$\beta_0$ - Constante	-6,59E+00	2,78E+00	0,02	31,10
$\beta_1$ - Tipo-RE	-2,67E-02	1,25E-01	0,83	-74,61
$\beta_2$ - Tipo-TF	7,22E-02	1,16E-01	0,53	513,16
$\beta_3$ - Cor	6,87E-02	3,76E-02	0,07	-13,03
$\beta_4$ - Rigidez dielétrica	8,32E-03	2,10E-03	0,00	4,95
$\beta_5$ - Teor de Água	1,39E-02	3,35E-03	0,00	3,20
$\beta_6$ - Densidade	4,87E+00	3,22E+00	0,13	55,41
$\beta_7$ - Tensão interfacial	-7,99E-02	5,93E-03	0,00	2,70
$\beta_8$ - FP25	5,27E-01	7,20E-01	0,46	21,26
$\beta_9$ - FP100	1,33E-02	3,08E-02	0,66	-15,82
$\beta_{10}$ - $H_2$	1,71E-03	5,50E-04	< 0,01	-722,00
$\beta_{11}$ - $O_2$	1,63E-05	6,25E-06	0,01	9,84
$\beta_{12}$ - $N_2$	-6,13E-06	2,04E-06	< 0,01	2,72
$\beta_{13}$ - $CH_4$	3,25E-03	1,16E-03	0,01	21,56
$\beta_{14}$ - CO	4,84E-04	1,27E-04	< 0,01	-0,88
$\beta_{15}$ - $CO_2$	-2,20E-05	1,54E-05	0,15	23,81
$\beta_{16}$ - $C_2H_4$	-4,55E-03	1,30E-03	< 0,01	11,91
$\beta_{17}$ - $C_2H_6$	-9,63E-06	4,42E-04	0,98	-102,13
$\beta_{18}$ - $C_2H_2$	-3,68E-03	1,84E-03	0,05	-1.119,85
$\phi$	3,01E+02	2,94E+01	< 0,01	3,09

Na Tabela 4.8, tem-se as estimativas dos parâmetros, erros padrões, valores p e DRP a partir da retiradas das observações 3 e 159. As conclusões são as mesmas do modelo sem a observação 159, a variável Cor torna-se não significativa ao nível de 5% e  $H_2$  passa a ser significativa neste nível de significância.

Tabela 4.8 *Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem as observações 3 e 159*

Coeficientes	Estimativa	EP	Valor	DRP (%)
$\beta_0$ - Constante	-6,59E+00	2,79E+00	0,02	31,02
$\beta_1$ - Tipo-RE	-2,64E-02	1,26E-01	0,83	-74,90
$\beta_2$ - Tipo-TF	7,22E-02	1,16E-01	0,54	512,59
$\beta_3$ - Cor	6,86E-02	3,78E-02	0,07	-13,15
$\beta_4$ - Rigidez dielétrica	8,32E-03	2,10E-03	< 0,01	4,93
$\beta_5$ - Teor de Água	1,39E-02	3,35E-03	< 0,01	3,19
$\beta_6$ - Densidade	4,87E+00	3,22E+00	0,13	55,24
$\beta_7$ - Tensão interfacial	-7,98E-02	5,95E-03	< 0,01	2,65
$\beta_8$ - FP25	5,27E-01	7,21E-01	0,46	21,27
$\beta_9$ - FP100	1,34E-02	3,09E-02	0,67	-15,79
$\beta_{10}$ - $H_2$	1,71E-03	5,51E-04	< 0,01	-721,97
$\beta_{11}$ - $O_2$	1,63E-05	6,29E-06	0,01	9,63
$\beta_{12}$ - $N_2$	-6,12E-06	2,05E-06	< 0,01	2,61
$\beta_{13}$ - $CH_4$	3,24E-03	1,18E-03	0,01	21,25
$\beta_{14}$ - CO	4,84E-04	1,27E-04	< 0,01	-0,92
$\beta_{15}$ - $CO_2$	-2,20E-05	1,55E-05	0,16	23,89
$\beta_{16}$ - $C_2H_4$	-4,54E-03	1,33E-03	< 0,01	11,56
$\beta_{17}$ - $C_2H_6$	-8,85E-06	4,44E-04	0,98	-101,96
$\beta_{18}$ - $C_2H_2$	-4,31E-03	2,03E-02	0,83	-1.295,97
$\phi$	3,00E+02	2,93E+01	< 0,01	2,61

Na Figura 4.5 foram plotados os resíduos com envelope para checar a adequação do

modelo sem os pontos já mencionados. Nota-se que para todos os pontos retirados os resíduos plotados contra seus quantis normais estão fora da banda de confiança, entretanto não parece haver desvios de magnitude elevada destes, sendo o modelo sem a 3<sup>o</sup> observação na amostra o que apresentou menos desvios.

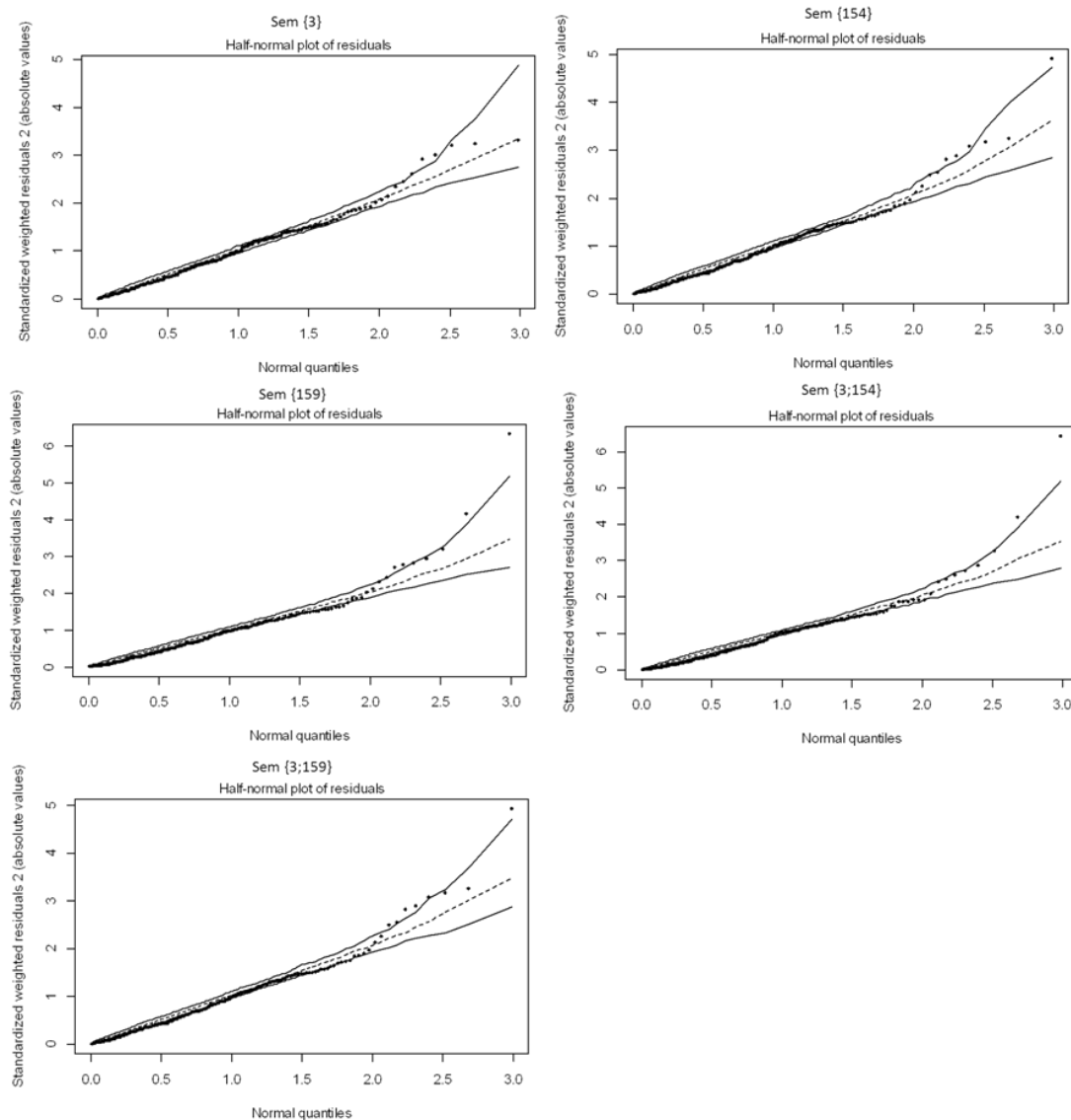


Figura 4.5 *Gráfico de Probabilidade Meio-Normal com Envelope Simulado*

Foi realizada uma validação no modelo de regressão Beta ajustado sem os pontos 3, 154, 159, 3; 154 e 3; 159 e somente com as variáveis significativas em cada caso, e utilizou-se

as 603 observações não selecionadas na amostragem. Para isto, foram obtidos os erros de previsão para cada uma destas observações, considerando o modelo sem pontos já mencionados, e em seguida calculou-se o Erro Quadrático Médio (EQM) e o Erro Absoluto Médio (EAM) dispostos na Tabela 4.9. O modelo ajustado sem a observação 3 possui o menor EQM e EAM, ou seja, em termos preditivos este modelo teve o melhor desempenho.

Adicionalmente, na Figura 4.6, foram plotados os valores observados do *índice de neutralização* e os ajustes do modelo sem as observações influentes. É desejável que para os modelos com alto poder preditivo os pontos no diagrama de dispersão estejam próximos da reta traçada. Observa-se que o modelo ajustado sem o ponto 3 tem o melhor alinhamento dos pontos, indicando que o modelo de regressão sem esta observação ajustou-se melhor aos dados de validação, resultando evidentemente em EQM e EAM menores.

Considerando somente o modelo sem a observação 3, que não alterou as conclusões inferenciais e teve o maior poder preditivo nos resultados da validação, reajustou-se o modelo mas agora somente com as variáveis significativas, os resultados estão na Tabela 4.10.

Para calcular a chance de aumento do *índice de neutralização* usou-se a expressão  $(\exp(\hat{\beta}_i) - 1)\%$  que retorna a chance relativa de um aumento no índice para 1 unidade de aumento na variável explicativa correspondente ao coeficiente estimado  $\hat{\beta}_i$ . O *índice de Neutralização* tem maior chance de aumento para a variação positiva de 1 unidade com a variável *Cor* (19,73%) e a chance de aumento do índice reduz com o incremento de 1 unidade da *Tensão Interfacial* (-7,65%) do transformador.

Tabela 4.9 *Erro Quadrático Médio (EQM) e Erro Absoluto Médio (EAM) do modelo de regressão Beta ajustado sem os pontos identificados no diagnóstico de influência*

Modelo	EQM	EAM
Sem a obs. 3	1,95E-03	2,62E-02
Sem a obs. 154	3,55E-03	2,68E-02
Sem a obs. 159	6,40E-03	3,29E-02
Sem a obs. 3 e 154	5,53E-03	2,98E-02
Sem a obs. 3 e 159	6,41E-03	3,29E-02



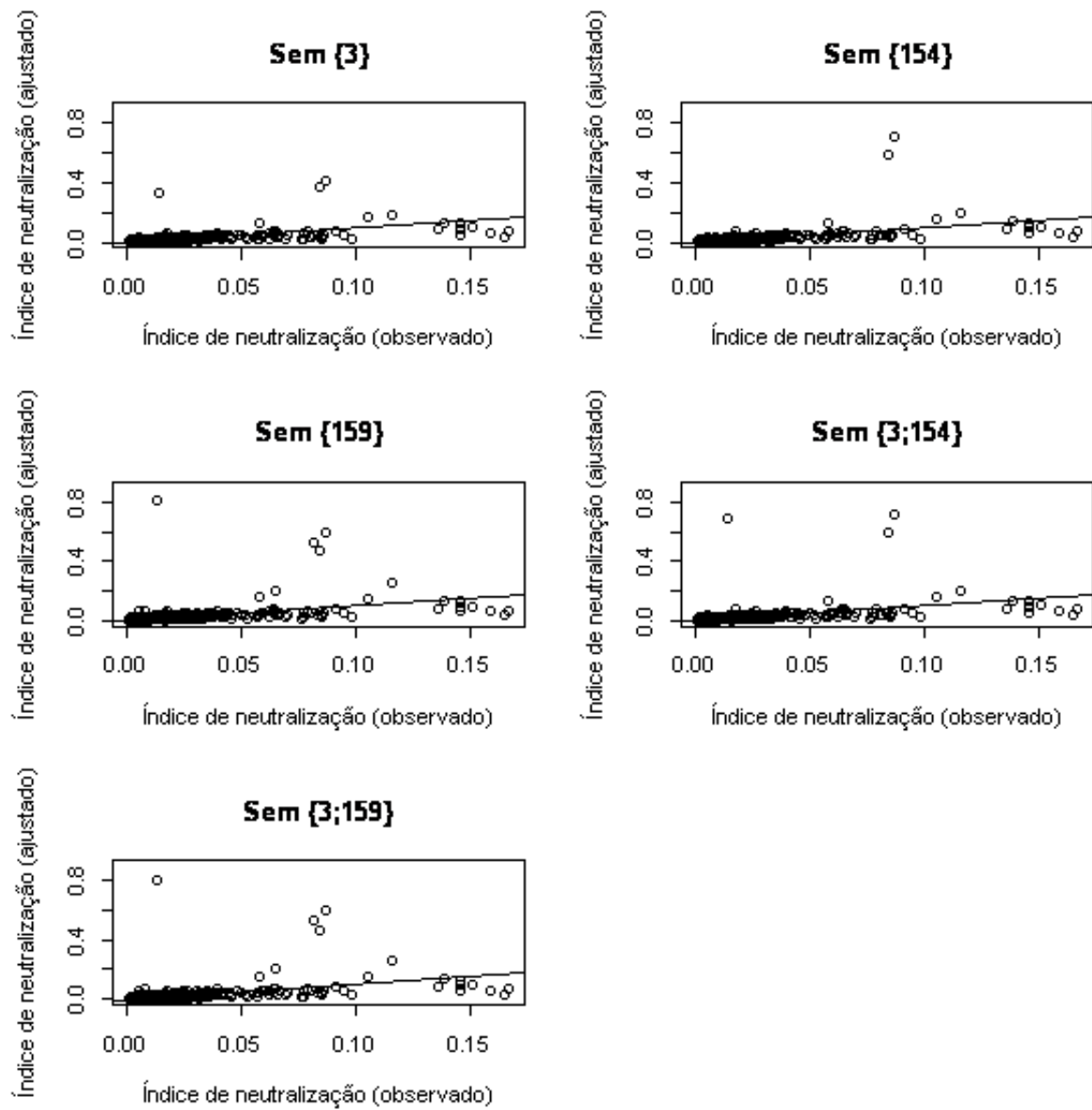


Figura 4.6 *Índice de neutralização observado e previsto na amostra de validação*

Tabela 4.10 *Estimativas dos parâmetros do modelo de regressão Beta somente com as variáveis significativas após a análise de diagnóstico.*

Parâmetros	Estimativa	EP	valor p	$(\exp(\hat{\beta}_i) - 1)\%$
$\beta_0$ - Intercepto	-2,3293E+00	3,1017E-01	< 0, 01	-
$\beta_1$ - Cor	1,8008E-01	3,0687E-02	< 0, 01	19,7313
$\beta_2$ - Rigidez Dielétrica	7,4946E-03	2,0847E-03	0, 01	0,7523
$\beta_3$ - Teor de Água	1,4419E-02	3,3345E-03	< 0, 01	1,4523
$\beta_4$ - Tensão Interfacial	-7,9535E-02	5,8125E-03	0	-7,6454
$\beta_5$ - $O_2$	1,8280E-05	6,2500E-06	< 0, 01	0,0018
$\beta_6$ - $N_2$	-6,4000E-06	2,0400E-06	< 0, 01	-0,0006
$\beta_7$ - $CH_4$	3,9357E-03	8,9616E-04	< 0, 01	0,3943
$\beta_8$ - CO	5,3839E-04	1,3086E-04	< 0, 01	0,0539
$\beta_9$ - $C_2H_4$	-2,2900E-05	1,5580E-05	< 0, 01	-0,4759
$\phi$	-4,7705E-03	1,1859E-03	< 0, 01	-

Desta forma, com base nos resultados da Tabela 4.10 o modelo final ajustado é dado por

$$\begin{aligned} \log\left(\frac{\hat{\mu}_t}{1 - \hat{\mu}_t}\right) = & -2,3293 + 1,8008E-01 \times (Cor)_t + 7,4946E-03 \times (\text{Rigidez Dielétrica})_t \\ & + 1,4419E-02 \times (\text{Teor Água})_t - 7,9535E-02 \times (\text{Tensão Interfacial})_t \\ & + 1,8280E-05 \times (O_2)_t - 6,4000E-06 \times (N_2)_t + 3,9357E-03 \times (CH_4)_t \\ & + 5,3839E-04 \times (CO)_t - 2,2900E-05 \times (C_2H_4)_t. \end{aligned}$$

---

## Capítulo 5

# Aplicação II: Rendimento acadêmico no Sistema de Cotas

---

### 5.1 Introdução

O termo *ação afirmativa* (*affirmative action*) surgiu nos Estados Unidos, na legislação trabalhista de 1935. As políticas de Ações Afirmativas no direito americano foram implementadas no âmbito do mercado de trabalho, na educação superior e nos contratos governamentais (WEDDERBURN, 2005 e MEDEIROS, 2005). Em pouco tempo, experiências similares podiam ser observadas em países como Índia, Malásia, Canadá, Nigéria, África do Sul, Argentina, Cuba, dentre outros. Na Europa ela ficou conhecida como “Ação ou Discriminação Positiva”.

No Brasil, uma das primeiras ações com propósitos de reparação foi a inserção na Constituição Federal de 1988 de um ponto no **“Título II - Dos direitos e garantias fundamentais, em seu Capítulo I - Dos direitos e deveres individuais e coletivos - Artigo 5º - que diz: Todos são iguais perante a lei, sem distinção de qualquer natureza, garantindo-se aos brasileiros e aos estrangeiros residentes no País a inviolabilidade do direito à vida, à liberdade, à igualdade, à segurança e à propriedade”**. O termo “Ação Afirmativa” chegou ao conhecimento da maior parte população, principalmente após a Conferência Mundial de Combate ao Racismo, Discriminação Racial, Xenofobia e Intolerância Correlata, realizada em setembro de 2001 em Durban, na África do Sul; onde os países signatários firmaram o compromisso de eliminar quaisquer formas de discriminação.

Segundo Piovesan (2005) as Ações Afirmativas são políticas que visam ampliar o acesso de grupos historicamente discriminados a todos o setores sociais. A ideia central é que

as mesmas possam servir para reparar danos causados a esses grupos e lhes oferecer uma oportunidade de equiparação; tais políticas têm um caráter transitório, tendo como objetivo maior aceleração dessa reparação criando um cenário consistente de grupos que superam suas condições.

## **5.2 Implantação da política de cotas no Ensino Superior Brasileiro**

No Brasil, as universidades pioneiras em adotar Ações Afirmativas e implantar cotas raciais no vestibular para o nível superior de ensino foram a Universidade Estadual da Bahia (Uneb) e a Universidade Estadual do Rio de Janeiro-Uerj. A universidade Nacional de Brasília-UnB foi a primeira universidade federal a adotar o sistema de cotas para negros, através do “plano de metas da UNB”, aprovado em 06 de junho de 2003, e entrando em vigor no segundo vestibular de 2004.

Ferreira (2008), no Foro Latino-americano de Políticas Educacionais, faz um mapa das ações afirmativas na Educação Superior com base em pesquisa realizada pelo Laboratório de Políticas Públicas da UERJ, no qual se constata que 72 instituições de ensino superior promovem algum tipo de ação afirmativa, havendo variações quanto ao tipo de política adotada (sistema de cotas, sistema de bonificação por pontos, reserva de vagas) e quanto à identificação dos sujeitos beneficiados com a ação afirmativa adotada: negros, indígenas, pessoas com deficiência, alunos da rede pública, pobres, mulheres negras. etc. (FERREIRA, 2008, p.3).

### **5.2.1 Implantação da política de cotas: o caso da UFPA**

A Universidade Federal do Pará (UFPA) foi um das primeiras instituições do Norte e Nordeste a implementar a política de cotas no vestibular. A decisão foi regulamentada pela Resolução nº 3.361 de 5 de agosto de 2005, do Conselho Superior de Ensino, Pesquisa e Extensão da UFPA (CONSEPE), no qual disponibilizou 50% das vagas de cada curso por um período de cinco (5) anos aos candidatos que estudaram todo o ensino médio em escola pública (cotas sociais). Dentro desse percentual, no mínimo 40% serão para os estudantes que se autodeclararem pretos ou pardos (cotas raciais) e optarem pelo sistema de cotas. Os demais candidatos, que ingressarem sem nenhuma cota, são considerados

não-cotistas. O primeiro ano de implementação desta política foi no Processo Seletivo Seriado da UFPA -PSS 2008.

A partir da Lei Federal nº 12.711/2012 todas as universidades devem adotar o sistema de reserva de vagas em seus vestibulares, sendo 50% das vagas ofertadas nos cursos de graduação para estudantes que cursaram todo o ensino médio na rede pública de ensino e, destas, 50% para alunos cuja renda familiar per capita seja de até um salário mínimo e meio, o que está sendo chamado de cota renda. A UFPA adaptou seu sistema de cotas à nova legislação, instituindo a cota renda .

### **5.2.2 Avaliação do Sistema de Cotas**

A característica principal das Ações Afirmativas, a exemplo da política de cotas, consiste em uma transitoriedade na sua aplicação, ou seja, sua duração pressupõe um período limitado. Nesse sentido, se faz necessária o acompanhamento dessa política nas universidades, pois é fundamental dispor de dados para que a avaliação do novo sistema seja deslocado do “achismo” e das representações ideológicas presentes no debate, e amparado no discurso da meritocracia. Ou seja, é necessário avaliar se haverá queda na qualidade de ensino, aumento da evasão e uma formação deficiente de profissionais.

Segundo o Instituto de Pesquisa Econômica Aplicada (Ipea) no universo de 54 universidades públicas que nos últimos oito anos adotaram o sistema de cotas no País, em ao menos quatro, distribuídas pelos principais Estados, os cotistas apresentaram desempenho próximo, similar ou até melhor em relação aos não-cotistas. Resultados iniciais do aproveitamento de cotistas na Unicamp, Universidade Federal da Bahia (UFBa), Universidade de Brasília (UnB) e Universidade do Estado do Rio de Janeiro (Uerj)

## **5.3 Descrição e estudo exploratório dos dados**

Os dados analisados neste capítulo aborda um estudo de caso referente aos alunos do curso de Medicina, campus Belém da UFPA, que ingressaram pelo primeiro PSS - Processo Seletivo Seriado da Instituição, que incluiu o sistema de cotas. Duas fontes de informações foram utilizadas nesse estudo:

- I. Banco de dados disponibilizado pelo CEPS/UFPA, contendo informações sobre o desempenho no vestibular dos candidatos que ingressaram em 2008, a idade e o sexo;
- II. Banco de dados disponibilizado pelo CIAC/UFPA, contendo o *Coefficiente de Rendimento do Período Letivo (CRPL)* dos alunos para 10 períodos letivos, compreendendo o período de 2008 a 2012.

O objetivo principal, quando se estuda dados do rendimento acadêmico de alunos cotistas, é comparar seu desempenho com o desempenho dos alunos não cotistas. Como os alunos cotistas, ingressantes no curso de Medicina na UFPA pelo PSS 2008, tiveram sua entrada apenas no segundo semestre letivo, foram considerados neste estudo apenas o *CRPL* de nove períodos letivos, a fim de ter boa representatividade dos dois grupos nestes períodos.

O *CRPL* é uma medida de desempenho acadêmico, atribuída pela instituição, que leva em conta fatores como o número de disciplinas cursadas no período, a carga horária destas disciplinas e o conceito obtido. Esta medida é apresentada com valores limitados ao intervalo  $[0, 10]$ .

A Figura 5.1 apresenta o box-plot do *CRPL* ao longo dos nove períodos de acompanhamento, nota-se que o *CRPL* mediano está entre 6 e 8 em todos os períodos e que a distribuição do *CRPL* torna-se mais assimétrica a partir de 5º período de curso, fato que reforça a aplicação da modelagem com medidas repetidas a partir da distribuição Beta.

Na Figura 5.2 tem-se os rendimentos médios de alunos cotistas e não cotistas ao longo dos nove primeiros períodos letivos. O *CRPL* médio dos não cotistas é inferior nos três primeiros períodos letivos avaliados, e ocorre uma relação inversa nos períodos seguintes.

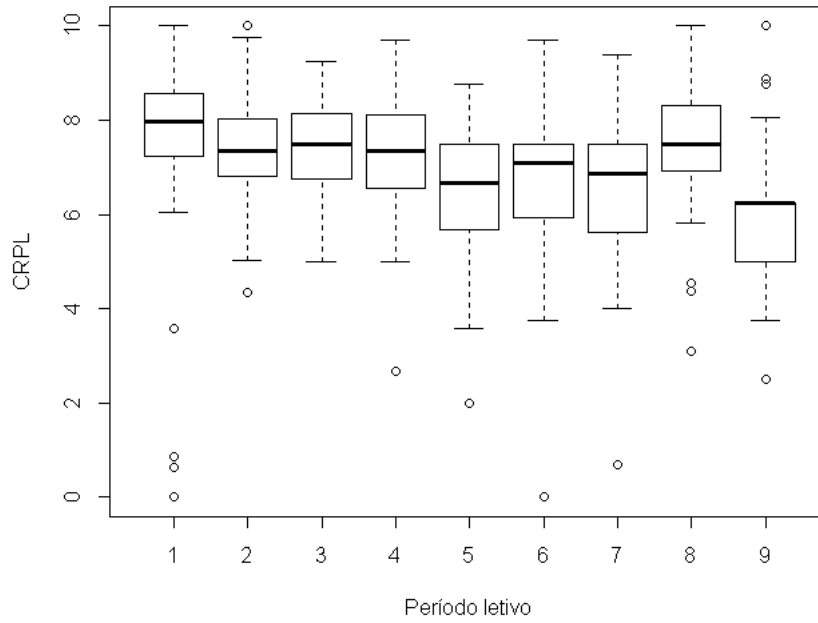


Figura 5.1 *Box-plot para o CRPL por período letivo*

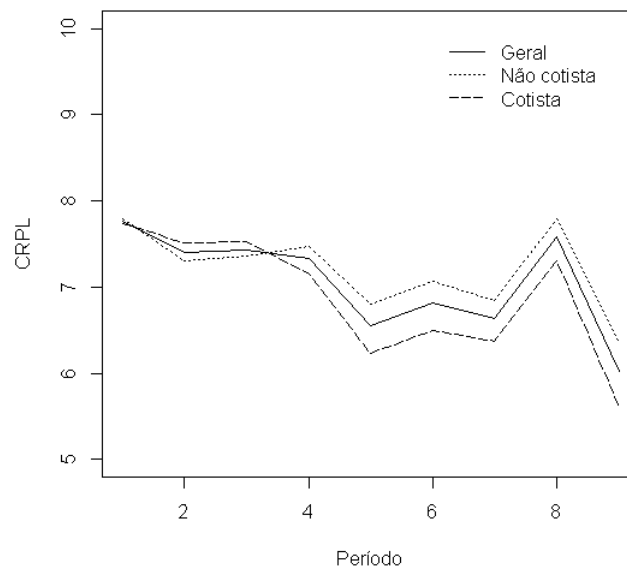


Figura 5.2 *CRPL médio para não cotistas e cotistas no curso de Medicina por período letivo*

A fim de propor uma estrutura para a matriz de correlação de trabalho, necessária no processo de estimação dos parâmetros do modelo de regressão, foram calculadas e plotadas as correlações para o *CRPL* entre os períodos letivos. As tendências a partir desse gráfico podem sugerir estruturas mais adequadas para a modelagem. Nas Figuras 5.3 e 5.4, buscou-se verificar a possibilidade de haver matrizes de correlações diferentes para não cotistas e cotistas, o que não foi observado. Além disso, alguns períodos como P1 e P4 alternam entre aumento e decréscimo da correlação com os períodos seguintes, indicando que talvez a estrutura AR-1 não seja a mais adequada. No entanto, foram utilizadas as matrizes padrão uniforme e AR-1 em detrimento da matriz não estruturada, por serem mais parcimoniosas. Os resultados do ajuste do modelo de regressão Beta com medidas repetidas será apresentado na próxima seção.

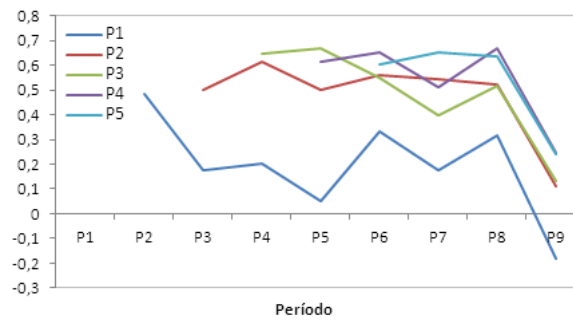


Figura 5.3 *Correlações entre os CRPL's dos não cotistas, iniciando no período 1 (P1) ao período 5 (P5) em relação aos períodos seguintes*

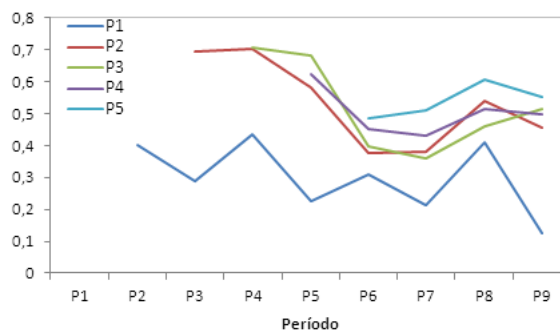


Figura 5.4 *Correlações entre os CRPL's dos cotistas, iniciando no período 1 (P1) ao período 5 (P5) em relação aos períodos seguintes*



## 5.4 Ajuste do modelo de Regressão Beta com medidas repetidas

Para o ajuste do modelo, considerou-se o *Coefficiente de Rendimento do Período Letivo* do  $i$ -ésimo aluno, no  $j$ -ésimo período ( $CRPL_{ij}$ ). A variável resposta  $y_{ij}$  é dada por

$$y_{ij} = \begin{cases} 0,0001 & \text{se } CRPL_{ij} = 0, \\ CRPL_{ij}/10 & \text{se } 0 < CRPL_{ij} < 10, \\ 0,9999 & \text{se } CRPL_{ij} = 10. \end{cases} \quad (5.1)$$

Conforme citado anteriormente a transformação (5.1) é necessária pois o espaço paramétrico da distribuição Beta é o intervalo unitário aberto  $(0,1)$ . Adotando a função de ligação logito e sob homogeneidade da dispersão, o modelo de regressão Beta com medidas repetidas proposto foi o seguinte:

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \sum_{k=2}^9 \beta_{1k} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{4ij}^2 + \beta_6 x_{5ij},$$

onde

$\mu_{ij}$ : Média da variável resposta  $y_{ij}$ ;

$\beta_{1k}$ : Coeficiente associado à variável indicadora do  $j$ -ésimo período, com

$$x_{1k} = \begin{cases} 1 & \text{se } k=j\text{-ésimo período} \\ 0 & \text{caso contrário;} \end{cases}$$

$\beta_2$ : Coeficiente associado ao gênero, com

$$x_{2ij} = \begin{cases} 1 & \text{se o aluno } i \text{ é homem} \\ 0 & \text{caso contrário;} \end{cases}$$

$\beta_3$ : Coeficiente associado ao cotista, com

$$x_{3ij} = \begin{cases} 1 & \text{se o aluno } i \text{ é cotista} \\ 0 & \text{caso contrário;} \end{cases}$$

$\beta_4$ : Coeficiente associado a *Idade* do aluno ( $x_{4ij}$ ),  $\forall j$ ;

$\beta_5$ : Coeficiente associado ao quadrado da *Idade* do aluno ( $x_{4ij}^2$ ),  $\forall j$ ;

$\beta_6$ : Nota padronizada final (NPF) do aluno  $i$  no PSS 2008 corrigida pela média ( $x_{5ij}$ ),  $\forall j$ .

Para a escolha da matriz de correlação de trabalho utilizou-se o critério QIC para as

estruturas *Padrão Uniforme* (QIC = -1603,03) e AR-1 (QIC = -1600,73). A estrutura Padrão Uniforme foi escolhida como a mais adequada para explicar a correlação entre as respostas da mesma unidade experimental, pois apresentou QIC menor.

As estimativas dos parâmetros do modelo estão apresentadas na Tabela 5.1. Pode-se observar que o desempenho no Vestibular (*NPF*) foi significativo, sendo um fator de influência no rendimento do aluno na universidade. O sinal negativo dos coeficientes associados aos períodos letivos indica que relativamente ao primeiro período letivo, o rendimento médio dos alunos do curso de Medicina foi menor, redução estatisticamente não significativa somente no 2º e 8º período. Verificou-se também que o fator *Gênero* não foi significativo, ou seja, não há diferenças entre as médias do *CRPL* dos alunos do gênero masculino e feminino.

O efeito da política de cotas não foi significativo, ou seja, o rendimento acadêmico dos alunos cotistas e não cotistas do curso de Medicina não difere significativamente. Entretanto, após o ajuste deste modelo inicial, é necessário utilizar as técnicas de diagnósticos, a fim de verificar se o modelo está bem ajustado. A Figura 5.5 apresenta a alavanca por observação (a), alavanca por aluno (b), distância de Cook (c) e o resíduo padronizado (d).

Uma inspeção na Figura 5.5 (d) mostra que várias observações aparecem como possíveis valores aberrantes. Em relação à distância de Cook, na Figura 5.5 (c) temos que os pontos (126;9), (47;6) e (65;1) podem ser considerados possíveis pontos influentes. No gráfico de alavanca por observação Figura 5.5 (a) nota-se que as observações 876, 877 e 878 aparecem como possíveis pontos de alavanca.

As observações influentes detectadas pelas técnicas de diagnóstico correspondem a três alunos cotistas e 4 alunos não cotistas que apresentaram o valor máximo do *CRPL* (10). Apenas dois alunos não cotistas apresentaram valores mínimos do *CRPL* (0). O passo seguinte é analisar a influência destas observações nas estimativas dos parâmetros.

Tabela 5.1 *Estimativas e erros padrões dos parâmetros do modelo de regressão Beta com a estrutura Padrão Uniforme. Dados do rendimento acadêmico dos alunos do curso de medicina da UFPA.*

Coeficiente-Variável	Estimativa	Erro padrão		Valor p
		<i>Robusto</i>	<i>Naive</i>	
$\beta_0$ - Intercepto	1,8131	0,5777	0,4493	0,0017
$\beta_{12}$ - Período2	-0,2833	0,1452	0,0645	0,0511
$\beta_{13}$ - Período3	-0,3272	0,1377	0,0643	0,0175
$\beta_{14}$ - Período4	-0,3654	0,1375	0,0641	0,0079
$\beta_{15}$ - Período5	-0,7395	0,1436	0,0624	< 0,001
$\beta_{16}$ - Período6	-0,6638	0,1512	0,0627	< 0,001
$\beta_{17}$ - Período7	-0,6997	0,1421	0,0626	< 0,001
$\beta_{18}$ - Período8	-0,1873	0,1466	0,0652	0,2013
$\beta_{19}$ - Período9	-0,9016	0,1598	0,0620	< 0,001
$\beta_2$ - Gênero	0,1188	0,0812	0,0490	0,1436
$\beta_3$ - Cota	0,0307	0,0845	0,0593	0,7162
$\beta_4$ - Idade	-0,0342	0,0431	0,0346	0,4272
$\beta_5$ - $Idade^2$	0,0004	0,0007	0,0006	0,5059
$\beta_6$ - NPF	0,1191	0,0362	0,0293	< 0,001
$\phi$ - Precisão	15,3882			
$\rho$ - Correlação	0,1645			

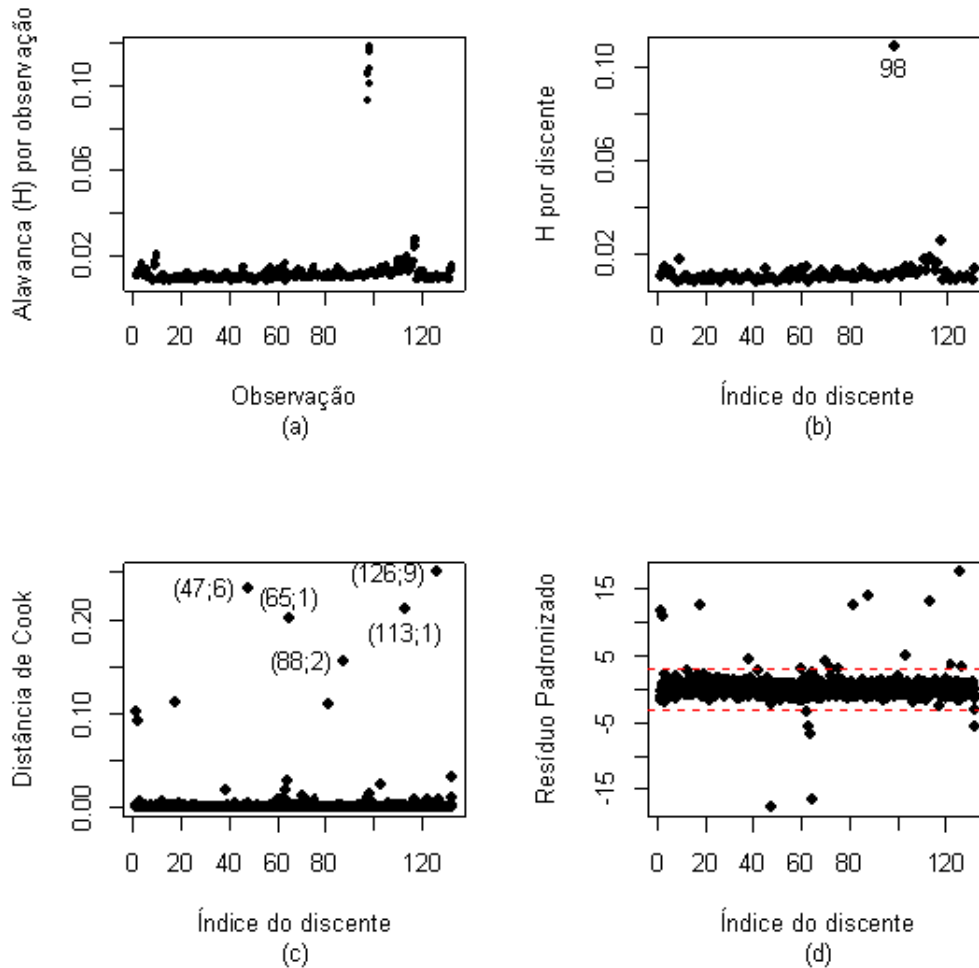


Figura 5.5 Gráficos de diagnósticos com todas as observações.

Na Tabela 5.2, para o modelo sem as observações (1, 2, 17, 47, 65, 82, 88, 113 e 126), somente as variáveis período2, sexo e idade passam a ser significativas (ao nível de significância de 5%). As maiores variações relativas (DRP) nas estimativas foram nos coeficientes *cota* (100,12%), que permaneceu não significativo e *idade* (68,61%).

Outras modelagens foram realizadas retirando somente o ponto alavanca correspondente ao aluno 98, e este aluno conjuntamente com as observações retiradas anteriormente com base nos resíduos padronizados e na distância de Cook, entretanto não houve mudanças no sinal e nem na significância estatística da variável *Cota*. Conclui-se, então, que não há

evidências de que o desempenho acadêmico de cotistas e não cotistas no curso de Medicina da UFPA difere significativamente com base no modelo proposto.

Tabela 5.2 *Estimativas dos parâmetros e o desvio relativo percentual (DRP) sem as observações (1, 2, 17, 47, 65, 82, 88, 113 e 126). Dados do rendimento acadêmico dos alunos do curso de medicina da UFPA.*

Coeficiente-Variável	Estimativa	Erro padrão		Valor p	DRP (%)
		<i>Robusto</i>	<i>Naive</i>		
$\beta_0$ - Intercepto	1,9736	0,3996	0,3403	< 0,001	8,8473
$\beta_{12}$ - Período2	-0,2009	0,0611	0,0571	0,0010	-29,0617
$\beta_{13}$ - Período3	-0,1890	0,0666	0,0571	0,0046	-42,2499
$\beta_{14}$ - Período4	-0,2201	0,0659	0,0569	< 0,001	-39,7681
$\beta_{15}$ - Período5	-0,6040	0,0630	0,0553	0	-18,3227
$\beta_{16}$ - Período6	-0,4628	0,0604	0,0559	< 0,001	-30,2923
$\beta_{17}$ - Período7	-0,5645	0,0624	0,0554	< 0,001	-19,3132
$\beta_{18}$ - Período8	-0,0971	0,0641	0,0577	0,1298	-48,1603
$\beta_{19}$ - Período9	-0,8325	0,0714	0,0548	< 0,001	-7,66082
$\beta_2$ - <i>Gênero</i>	0,1230	0,0530	0,0361	0,0203	3,5590
$\beta_3$ - <i>Cota</i>	0,0615	0,0618	0,0506	0,3197	100,1259
$\beta_4$ - <i>Idade</i>	-0,0577	0,0290	0,0260	0,0468	68,6137
$\beta_5$ - <i>Idade</i> <sup>2</sup>	0,0009	0,0005	0,0005	0,0780	95,1913
$\beta_6$ - <i>NPF</i>	0,1169	0,0283	0,0244	< 0,001	-1,9079
$\phi$ - <i>Precisão</i>	17,9890				
$\rho$ - <i>Correlação</i>	0,2388				

Assim, este estudo de caso envolvendo o curso de Medicina da UFPA apresenta resultados que contrariam o discurso de que os alunos que ingressam através do sistema de cotas apresentariam desempenho inferior e, conseqüentemente, teriam algum déficit no rendimento das disciplinas (FILHO *et al.*, 2005).

---

## Capítulo 6

# Considerações Finais e Recomendações

---

### 6.1 Conclusões

Nesse trabalho apresentamos o modelo de Regressão Beta proposto por Ferrari e Cribari-Neto (2004). Mais especificamente, foram estudadas as características e propriedades do modelo no que diz respeito a sua definição, funções de ligação, função escore, matriz de informação, estimação dos parâmetros de regressão e algumas medidas de diagnóstico. Apresentamos, também, a proposta da Venezuela (2008) que desenvolveu equações de estimação para os modelos de Regressão Beta com medidas repetidas. Foram apresentadas duas novas aplicações do modelo de regressão Beta a dados reais.

Na primeira aplicação utilizou-se o modelo de Regressão Beta simples, usado para identificar quais as características físico-químicas e de cromatografia gasosa que influenciam significativamente na vida útil de transformadores, através da avaliação do seu índice de neutralização. Foram observadas significativas as seguintes variáveis: *Cor*, *Rigidez Dielétrica*, *Teor Água*, *Tensão Interfacial*, *O<sub>2</sub>*, *N<sub>2</sub>*, *CH<sub>4</sub>*, *CO* e *C<sub>2</sub>H<sub>4</sub>*. Observou-se, também, que o *índice de Neutralização* tem maior chance de aumento para a variação positiva de 1 unidade com a variável *Cor* (19,73%) e a chance de aumento do índice reduz com o incremento de 1 unidade da *Tensão Interfacial* (-7,65%) do transformador.

Na segunda aplicação utilizou-se o modelo de Regressão Beta com medidas repetidas para avaliar e comparar o rendimento acadêmico de alunos cotistas e não cotistas da UFPA, através do CRPL observado ao longo de nove períodos letivos, em um estudo de caso envolvendo alunos do curso de Medicina da UFPA. Foram observadas significativas as seguintes variáveis: *Sexo*, *Idade*, *Nota padronizada Final* no PSS e os *períodos letivos* (menos o período 8), quando comparados ao período 1. Observou-se, nestes períodos, que o rendimento acadêmico médio dos alunos decresceu com o tempo. A variável cota não foi

significativa, indicando que não foram detectadas diferenças significativas no rendimento acadêmico dos cotistas e não cotistas do curso de medicina da UFPA que ingressaram pelo PSS 2008, resultados que contribuem para a discussão e avaliação da eficiência da política de cotas como instrumento de minimização das desigualdades no acesso ao ensino superior no Brasil. Este estudo mostra a necessidade de se fazer este acompanhamento nos demais cursos da UFPA.

Por fim, nessas aplicações mostrou-se que a distribuição Beta e seus modelos de regressão associados surgem como novas propostas para ajustar dados que têm seu limite de variação no intervalo unitário  $(0,1)$ .

## 6.2 Recomendações

A fim de dar continuidade no estudo destes modelos, sugerimos para trabalhos futuros:

- Modelar conjuntamente a média e a dispersão nos dados de Índices de Neutralização;
- Para os dados de CRPL dos alunos que ingressaram pelo PSS 2008, aplicar as técnicas de influência local generalizada conforme Venezuela (2008), considerando se for o caso a modelagem conjunta da média e da dispersão;
- Ampliar o leque de estruturas de correlações de trabalho seguindo a proposta de Artes (1997);
- Utilizar outras medidas de seleção de estruturas de correlação possivelmente mais acuradas, uma alternativa é utilizar o Critério de Informação de Correção (CIC) (Hin e Wuang, 2009) para o modelo de regressão Beta com medidas repetidas.
- Desenvolver as equações de estimação generalizadas para medidas repetidas a partir do modelo de regressão Beta retangular (BAYES, BÁZAN e GARCIA, 2012) que mostrou-se mais robusto na presença de pontos aberrantes para o caso de independência das observações;
- Utilização de um modelo de regressão Beta misto.

---

## Referências Bibliográficas

---

- Artes, R. (1997). Extensões da teoria das equações de estimação generalizadas a dados circulares e modelos de dispersão, Tese de doutorado.
- Arellano-Vale, R.; Bolfarine, H.; Lachos, V.H. (2005). Skew-normal linear mixed models, *Jornal of data Science* 3: 415-438
- Andrade, A. C. G. (2007), Efeitos de especificação incorreta da função de ligação no modelo de regressão beta, Dissertação de Mestrado, IME-USP, São Paulo.
- Bury, K. (1999). *Statistical Distributions in Engineering*. New York: Cambridge University Press.
- Batista, D. A. (2005). Desenvolvimento de Metodologia Colorimétrica Dosagem de Compostos Furânicos em Óleo Mineral Isolante. Curitiba: UFPR.
- Bayes, L. C.; Bazán, L. J.; García, Catarina (2012). A New Robust Regression Model for Proportions **7**, 4: 841-866.
- Branscum, A. J.; Johnson, W. O. e Thurmond; M. C. (2007). “Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease viruses” *Australian and New Zealand Journal of Statistics* 49(3), 287-301.
- Cribari-Neto, F.; Zeileis, A. (2010). “Beta regression in r” *jornal of statistical Software* 34 (2), 1-24.
- Cribari-Neto, F.; Souza, T. C. (2011), “Testing inference in variable dispersion beta regressions”, *Journal of Stat.*



- 
- Espinheira, P. L.; Ferrari, S. L. P; Cribari-Neto, F. (2008a). "Influence diagnostics in beta regression", *Computational Statistics and Data Analysis* 52(9), 4417-4431. 3, 56.
- Espinheira, P. L.; Ferrari, S. L. P; Cribari-Neto, F. (2008b). "On beta regression residuals", *Journal of Applied Statistics* 35(4), 407-419. 3, 49, 53, 54.
- Fernandes, P. O (1987). *Apostila: Manutenção de transformadores através do controle do óleo isolante mineral: notas de aula do curso de manutenção de transformadores. Westinghouse do Brasil – Divisão de Serviços Industriais*, 17p.
- Ferrari, S.; Cribari-Neto, F.. Beta Regression for Modelling Rates and Proportions, *Journal of Applied Statistics* 31 (7) (2004) 799-815.
- Graham, V. A.; Hollands, K. G. T. (1990). Method to generate synthetic hourly solar radiation globally. *Solar Energy*, 44, 333-341.
- Gil, A. C.. *Como elaborar projetos de pesquisa*. 4<sup>a</sup> ed. São Paulo: Atlas, 2002.
- Hardin, J.; Hilbe J. (2003). *Generalized Estimating Equations*, Chapman e Hall/CRC.
- Hin L-Y; Wang Y-G. Working-correlation-structure identification in generalized estimating equations. *Stat Med*. 2009 Feb;28(4)642-58.
- Janardan, K. G.; Padmanabhan, G. (1986). Double bounded beta distribution for hydrologic variables. *Proceedings of 17 th Annual Pittsburgh Conference*, 17, 1107-1111.
- Johnson, N. L.; Kotz, S.; Balakrishnan, N. (1995). *Continuous Univariate Distributions*, vol. 2. 2<sup>a</sup> edição. New York: Wiley.
- Littell, R.C. *et al.* *SAS System for Mixed Models*. 1<sup>a</sup> ed. North Carolina: SAS Institute Inc., 2000, 615 p.
- Lachos, V. H.; Ghosh, P.; Arellano-Valle, R. B. (2010). Likelihood based inference for skew-normal independent linear mixed models, *Statistica Sinica* **20**: 303-322.
- Myers, S. D.; Kelly, J.; Parrish, R. (1982). *A guide to transformer maintenance*. Ohio.: Transformer Maintenance Institute.

- 
- Milash, M. (1984). *Manutenção de Transformadores em Líquido Isolante*. Edgard Blücher Ltda.
- McNally, R. J. (1990). Maximum likelihood estimation of the parameters of the prior distributions of three variables that strongly influence reproductive performance in cows. *Biometrics*, 446, 501-514.
- Milyutin, E. R.; Yaromenko, Y. I. (1991). Statistical characteristics of atmospheric transparency index over tilted routes. *Meteorologiya I Gidrologiya*, 12, 72-76
- Morais, D. R. Ferramenta inteligente para detecção de falhas incipientes em transformadores baseada na análise de gases dissolvidos no óleo isolante. Programa de pós-graduação em engenharia elétrica-UFSC-2004.
- Manco, U. C. O. Modelos de regressão beta com efeitos aleatórios normais e não normais para dados longitudinais. 2013. p. Tese (Doutorado em estatística)-Instituto de Matemática e Estatística Universidade de São Paulo, São Paulo. 2007.
- Miyashiro, E. S. (2008), Modelos de regressão beta e simplex para análise de proporções, Dissertação de mestrado em estatística, Instituto de Matemática e Estatística-Universidade de São Paulo 2, 3.
- Melo, T. F. N.; Vasconcellos, K. L. P.; Lemonte, A. J. (2009), “Some restriction tests in a new class of regression models for proportions”, *Computational Statistics and Data Analysis* 53(12), 3972-3979. 3.
- Nocedal, J.; Wright, S. J. (1999), *Numerical optimization*, Springer-Verlag, New York.
- Oliveira, M. S. (2004), Um modelo de regressão beta: teoria e aplicações, Dissertação de Mestrado, IME-USP, São Paulo.
- Ospina, R.; Cribari-Neto, F.; Vasconcellos, K. L. P (2006), “Improved point and interval estimation for a beta regression model”, *Computational Statistics and Data Analysis* 51(2), 960-981. 3.
- Paolino, P.. Maximum Likelihood Estimation of Models with Beta-Distributed Dependent

---

Variables, *Political Analysis* 9 (4) (2001) 325-346.

Pinheiro, J. C.; Liu, C. & Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate-t distribution, *Journal of Computational and Graphical Statistics* 10: 249-276.

Savalli, C.; Paula, G. A.; Cysneiros, F. J. A. (2006). Assessment of variance components in elliptical linear mixed models, *Statistical Modelling* 6: 59-76.

Sen, P.; Singer, J. M. (1993), *Large Sample Methods in Statistics: An Introduction With applications*, Chapman and Hall, . New York.

Smithson M; Verkuilen J. (2006). "A better lemon-squeezer? Maximum likelihood regression with beta-distributed dependent variables". *Psychological Methods*, 11(1),54-71.

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, 348-366. 2010.

Souza, D. F.. *Regressão beta multivariada com aplicações em pequenas áreas*, Tese de doutorado em estatística Instituto de Matemática-Universidade Federal do Rio de Janeiro. 2011.

Venezuela, M. K. *Equação de estimação generalizada e influência local para modelos de regressão beta com medidas repetidas*. 2008. 153 p. Tese (Doutorado em estatística) - Instituto de Matemática e Estatística Universidade de São Paulo, São Paulo. 2008, 3, 56.

Wilhelm, H. M.; Stocco, G. B. *Determinação da Estabilidade á Oxidação de óleos Vegetais Isolantes pelo Método Condutivimétrico*. 2008, 6 p. Relatório Técnico apresentado no CIGRÊ.