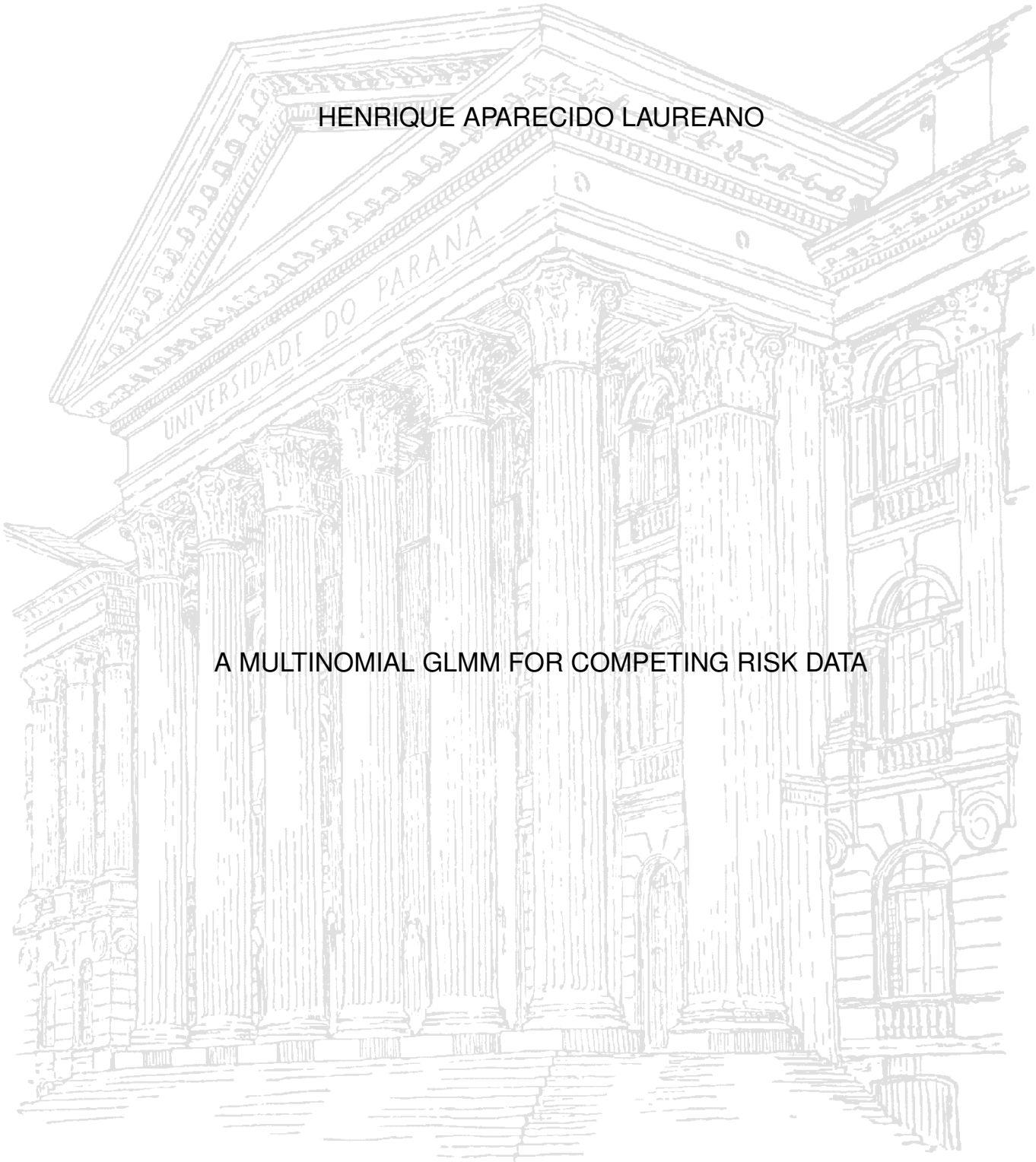
FEDERAL UNIVERSITY OF PARANÁ

HENRIQUE APARECIDO LAUREANO



A MULTINOMIAL GLMM FOR COMPETING RISK DATA

CURITIBA

2020

HENRIQUE APARECIDO LAUREANO

A MULTINOMIAL GLMM FOR COMPETING RISK DATA

Thesis presented to the Graduate Program of Numerical Methods in Engineering, Concentration Area in Mathematical Programming: Statistical Methods Applied in Engineering, Federal University of Paraná, as part of the requirements to the obtention of the Master's Degree in Sciences.

Supervisor: Prof. PhD Wagner Hugo Bonat

Co-supervisor: Prof. PhD Paulo Justiniano Ribeiro Jr

CURITIBA

2020

HENRIQUE APARECIDO LAUREANO

A MULTINOMIAL GLMM FOR COMPETING RISK DATA

Thesis presented to the Graduate Program of Numerical Methods in Engineering, Concentration Area in Mathematical Programming: Statistical Methods Applied in Engineering, Federal University of Paraná, as part of the requirements to the obtention of the Master's Degree in Sciences.

Master thesis approved. XXX XX, 2020.

Prof. PhD Wagner Hugo Bonat
Supervisor

Prof. PhD Paulo Justiniano Ribeiro Jr
Co-supervisor

Prof. PhD ...
Internal Examinator - PPGMNE

Prof. PhD ...
Internal Examinator - PPGMNE

Prof. PhD ...
External Examiner -

CURITIBA
2020

To Celita and Olivio

ACKNOWLEDGEMENTS

...

"A simplicidade é o último grau de sofisticação".
(Leonardo da Vinci)

ABSTRACT

...

Keywords: . . . . . .

RESUMO

...

Palavras-chave: . . . . . .

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUÇÃO

Em diversas áreas de pesquisa é comum investigar a relação entre uma variável de interesse com outras variáveis que compõem o estudo. Para tanto, faz-se uso da técnica estatística de modelos de regressão, uma vez que se pode estudar o relacionamento entre uma variável resposta (variável dependente) com possíveis variáveis explicativas (covariáveis) (MONTGOMERY; PECK; VINING, 2012). A aplicação desta técnica estatística é ampla, abrangendo diversas áreas do conhecimento como medicina, engenharias, agronomia, ciências sociais dentre outras. Nesse contexto, um dos principais modelos de regressão e sem dúvida um dos mais usados por usuários de estatística aplicada é o clássico modelo de regressão linear (Gaussiano). No entanto, para uso desse modelo alguns pressupostos devem ser atendidos, tais como erros independentes e identicamente distribuídos segundo a distribuição normal com média zero e variância constante (DRAPER; SMITH, 2014). Na prática, isso nem sempre acontece e a má especificação desse modelo pode gerar erros padrões inconsistentes, além de outros problemas que invalidam todo o processo de inferência (MYERS et al., 2010; MONTGOMERY; PECK; VINING, 2012). Apesar de amplamente utilizado, o modelo de regressão linear não é adequado para respostas binárias, politômicas, contagens ou limitadas.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Propor um modelo de regressão para análise de variáveis respostas limitadas multivariada.

1.1.2 Objetivos específicos

1. Estudar o desempenho do algoritmo NORTA (*NORmal To Anything*) para simular variáveis aleatórias beta correlacionadas.
2. Especificar o modelo usando suposições de primeiro e segundo momentos.
3. Usar as funções de estimação quase-score e Pearson para estimar os parâmetros de regressão e dispersão, respectivamente.
4. Delinear estudos de simulação para explorar a flexibilidade do modelo para lidar com dados limitados em estudos longitudinais, além de checar propriedades dos estimadores em estudos com múltiplas respostas correlacionadas.

5. Adaptar técnicas de diagnóstico para o modelo proposto, como DFFITS, DFBETAS, distância de Cook e o gráfico de probabilidade meio-normal com envelope simulado.

6. Aplicar o modelo proposto em dois conjuntos de dados.

1.2 JUSTIFICATIVA

1.3 LIMITAÇÕES

Este trabalho se restringe a propor um novo modelo de regressão para análise de variáveis respostas limitadas multivariada. Para motivar o novo modelo, serão apresentadas aplicações em dois conjuntos de dados, que não são facilmente manipulados pelos métodos estatísticos existentes. Portanto, testes de hipóteses e de comparações múltiplas multivariados não serão desenvolvidos no decorrer deste trabalho.

1.4 ORGANIZAÇÃO DO TRABALHO

Esta dissertação contém seis capítulos incluindo esta introdução. O chapter 2 descreve os dois conjuntos de dados que serão usados como exemplos de aplicação no novo modelo. O chapter 3 apresenta a revisão bibliográfica que motivou este trabalho, introduz o modelo de regressão beta (univariado), apresenta o algoritmo NORTA (*NORmal To Anything*) usado nos estudos de simulação e discute brevemente as medidas de bondade de ajuste usadas no trabalho. O chapter 4 propõe o modelo de regressão quase-beta multivariado, apresenta o método usado para estimação e inferência e adapta técnicas de diagnóstico. No chapter 5 são apresentados os resultados de três estudos de simulação, além da análise dos dados apresentados no chapter 2. Finalmente, o chapter 6 discute as principais contribuições desta dissertação, além de apresentar as conclusões seguidas por sugestões para futuros trabalhos.

2 CONJUNTOS DE DADOS

Este Capítulo descreve os dois conjuntos de dados que serão usados como exemplos de aplicação no novo modelo de regressão, proposto no chapter 4. O primeiro conjunto se refere ao índice de qualidade da água de reservatórios de usinas hidrelétricas operadas pela COPEL no Estado do Paraná. Já o segundo conjunto de dados corresponde ao percentual de gordura corporal de indivíduos avaliados no Hospital de Clínicas da Universidade Federal do Paraná.

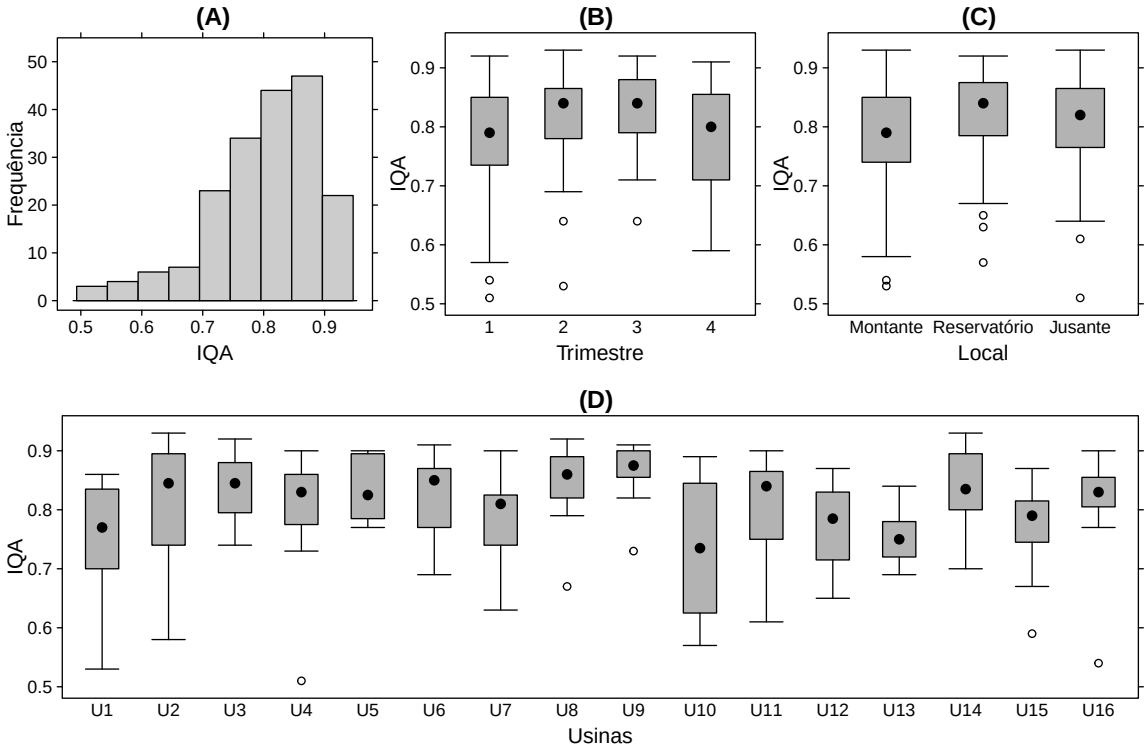
2.1 CONJUNTO DE DADOS I: ÍNDICE DE QUALIDADE DA ÁGUA

TABLE 1 – ANÁLISE DESCRITIVA PARA O IQA POR TRIMESTRE E LOCAL

Trimestre	Local		
	Montante	Reservatório	Jusante
1	$0,75 \pm 0,11$	$0,80 \pm 0,10$	$0,78 \pm 0,10$
2	$0,79 \pm 0,10$	$0,83 \pm 0,06$	$0,83 \pm 0,07$
3	$0,81 \pm 0,07$	$0,85 \pm 0,05$	$0,83 \pm 0,06$
4	$0,76 \pm 0,10$	$0,81 \pm 0,08$	$0,79 \pm 0,09$

FONTE: O autor (2018).

FIGURE 1 – HISTOGRAMA (A) E BOXPLOTS PARA O ÍNDICE DE QUALIDADE DA ÁGUA (IQA) POR TRIMESTRE (B), LOCAL (C) E USINAS (D)



FONTE: O autor (2018).

Por fim, os resultados apresentados na Figure 1 (D) mostram que o IQA não é homogêneo entre as usinas, com um destaque maior para as usinas 1, 2 e 10. É importante ressaltar que os resultados apresentados na Table 1 e Figure 1 se referem apenas a análise descritiva e exploratória dos dados, onde são criadas hipóteses que serão confirmadas somente após ajuste do modelo de regressão proposto no chapter 4. No **??** são apresentados gráficos boxplots para o IQA separado por trimestre e local em função das usinas.

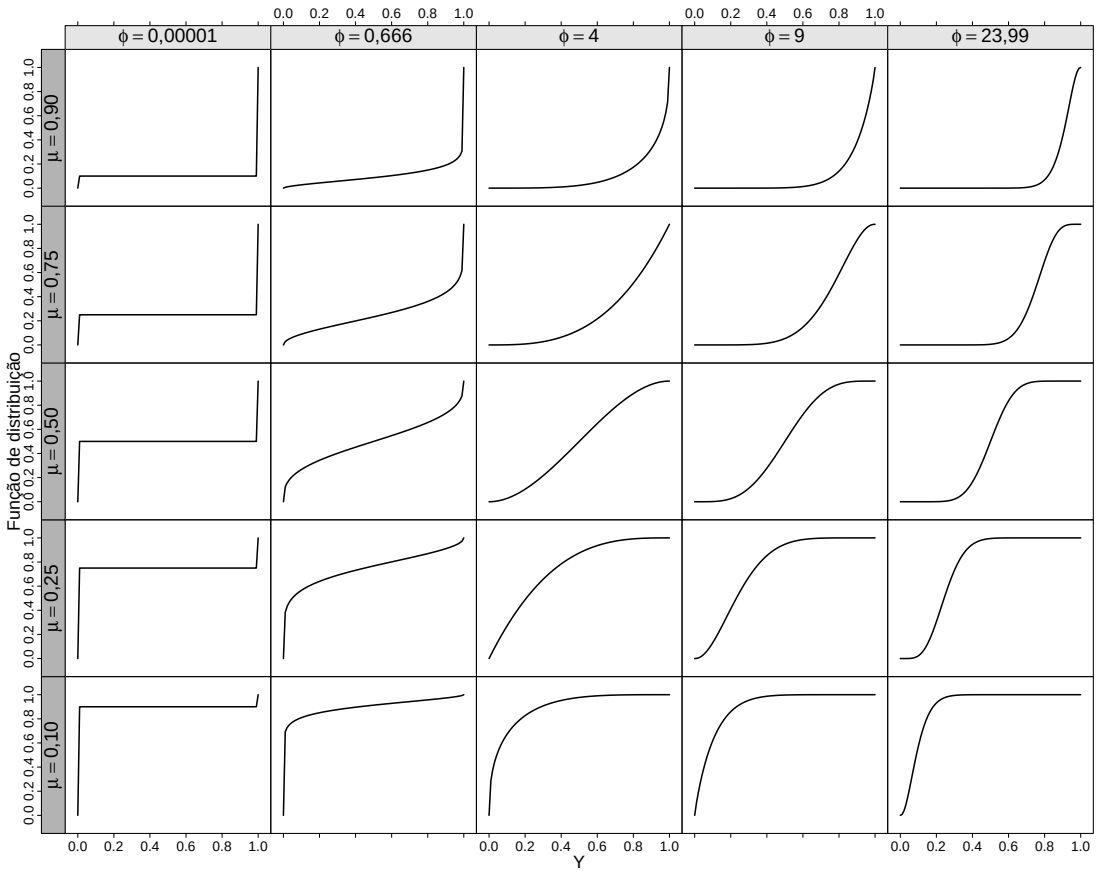
3 FUNDAMENTAÇÃO TEÓRICA

Este Capítulo apresenta a fundamentação teórica que será usada nesta dissertação. A section 3.1 apresenta um breve resumo dos principais trabalhos relacionados ao assunto. A distribuição de probabilidade beta e suas propriedades encontram-se na section 3.2. A **??** apresenta o algoritmo NORTA, que será usado para simular variáveis aleatórias beta correlacionadas. A **??** introduz o modelo de regressão beta (univariado). Por fim, a **??** apresenta brevemente as medidas de bondade de ajuste usadas na comparação entre os modelos.

3.1 REVISÃO DA LITERATURA

3.2 DISTRIBUIÇÃO DE PROBABILIDADE BETA

FIGURE 2 – FUNÇÃO DE DISTRIBUIÇÃO BETA PARA DIFERENTES VALORES DE μ COMBINADOS COM $\phi = (0{,}00001;\ 0{,}666;\ 4;\ 9;\ 23{,}99)$



FONTE: O autor (2018).

FIGURE 3 – CÓDIGOS EM LINGUAGEM R PARA GERAÇÃO DE VARIÁVEIS ALEATÓRIAS BETA CORRELACIONADAS

```
R = 1000  # tamanho da amostra
mu = 0.5  # parâmetro de média
phi = 9  # parâmetro de dispersão
cor_matrix <- matrix(c(1.0,0.75,0.75,1.0),2,2) # matriz de correlação
require(MASS)  # carrega o pacote com a função mvrnorm()
Z <- mvrnorm(n = R, mu = c(0,0), Sigma = cor_matrix)  # passo 1
Y <- qbeta(pnorm(Z), shape1 = mu*phi, shape2 = (1 - mu)*phi) # passo 2
```

FONTE: O autor (2018).

4 MODELO DE REGRESSÃO MULTIVARIADO

Este Capítulo apresenta o novo modelo de regressão usado para análise de variáveis respostas limitadas multivariada, o qual será chamado por modelo de regressão quase-beta multivariado. A section 4.1 apresenta a estrutura do modelo, enquanto a **??** apresenta o método proposto para estimação dos parâmetros de regressão e dispersão. Por fim, a **??** adapta técnicas de diagnóstico para o modelo proposto.

4.1 MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO

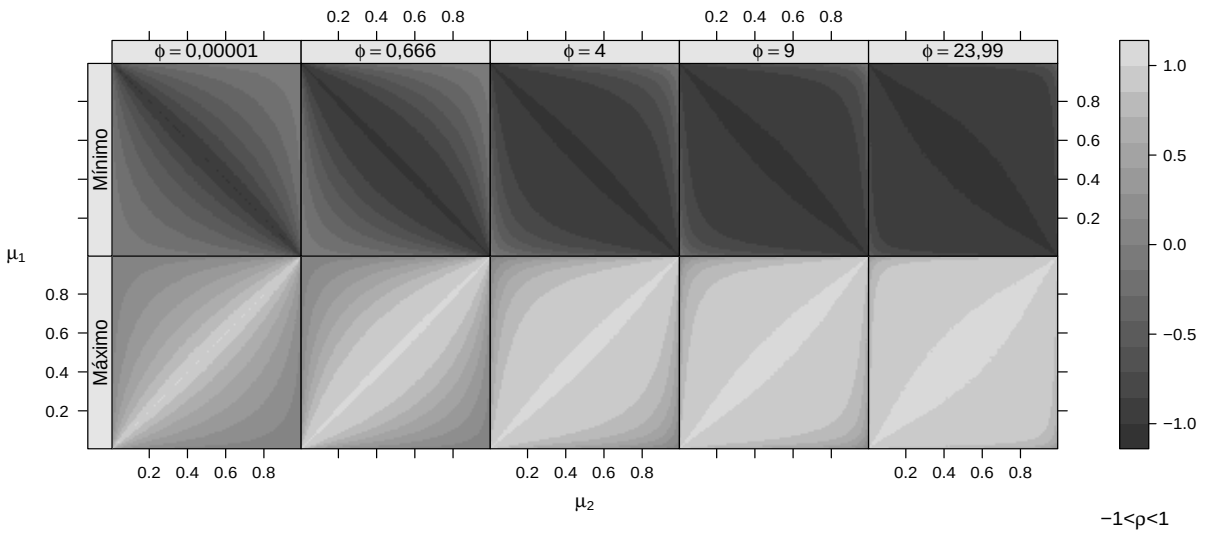
5 RESULTADOS

Neste Capítulo são apresentados os resultados de três estudos de simulação, além da análise dos dados apresentados no chapter 2. O primeiro estudo de simulação foi conduzido para investigar o comportamento do algoritmo NORTA (NOR*mal To Anything*) na simulação de variáveis aleatórias beta correlacionadas (subsection 5.1.1). O segundo visou checar propriedades dos estimadores para os parâmetros de dispersão, no contexto de análise de dados longitudinais (**??**). E o terceiro foi delineado para explorar a flexibilidade dos estimadores para lidar com múltiplas respostas correlacionadas (**??**). Por fim, a **??** apresenta os resultados da análise dos dados referente ao índice de qualidade da água (IQA), enquanto a **??** apresenta os resultados correpondentes ao percentual de gordura corporal.

5.1 ESTUDOS DE SIMULAÇÃO

5.1.1 Comportamento do algoritmo NORTA

FIGURE 4 – VALORES MÍNIMOS E MÁXIMOS PARA A CORRELAÇÃO ENTRE DUAS VARIÁVEIS ALEATÓRIAS BETA EM FUNÇÃO DAS MÉDIAS MARGINAIS E DIFERENTES VALORES DO PARÂMETRO ϕ



FONTE: O autor (2018).

$$g(\mu_{jki}) = \beta_0 + \beta_{1j}\,\texttt{local}_{ji} + \beta_{2k}\,\texttt{trimestre}_{ki}, \tag{5.1}$$

Na sequência, ajustou-se o modelo de regressão quase-beta multivariado aos dados do IQA, considerando as quatro estruturas acima mencionadas além de especificar a função de ligação *logit* para o preditor linear (Equation 5.1).

6 CONSIDERAÇÕES FINAIS

O objetivo geral desta dissertação foi propor um novo modelo de regressão para análise de variáveis respostas limitadas multivariada. O modelo foi especificado usando apenas suposições de primeiro e segundo momentos. Para estimação dos parâmetros, adotou-se uma abordagem que combina as funções de estimação quase-score e Pearson para estimação dos parâmetros de regressão e dispersão, respectivamente. Assim, o modelo proposto nesta dissertação segue o estilo de quase-verossimilhança apresentado por Wedderburn (1974), onde a especificação do modelo é feita pela combinação da função de variância da distribuição binomial com as tradicionais funções de ligação para dados binários.

6.1 FUTUROS TRABALHOS

BIBLIOGRAPHY

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. New York: John Wiley & Sons, 2014. Cited on page 11.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. New York: John Wiley & Sons, 2012. v. 821. Cited on page 11.

MYERS, R. et al. *Generalized Linear Models: With Applications in Engineering and the Sciences: Second Edition*. New Jersey: John Wiley and Sons Inc., 2010. 496 p. Cited on page 11.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, v. 61, n. 3, p. 439–447, 1974. Cited on page 19.