

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE ESTATÍSTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

DIAGNÓSTICO DE INFLUÊNCIA PARA UMA
FAMÍLIA DE MODELOS DE REGRESSÃO PARA
DADOS DE TAXAS E PROPORÇÕES

FRANCISCO HILDEMAR CALIXTO DE ALENCAR

DISSERTAÇÃO DE MESTRADO



Recife
2016

FRANCISCO HILDEMAR CALIXTO DE ALENCAR

DIAGNÓSTICO DE INFLUÊNCIA PARA UMA
FAMÍLIA DE MODELOS DE REGRESSÃO PARA
DADOS DE TAXAS E PROPORÇÕES

ORIENTADOR: PROF. DR. RAYDONAL OSPINA MARTÍNEZ

COORIENTADOR: PROF. DR. JUVÊNCIO SANTOS NOBRE

Área de Concentração: Estatística Aplicada

Dissertação apresentada ao programa de Pós-graduação em Estatística do Departamento de Estatística da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de **Mestre em Estatística**.

Recife

2016

Catálogo na fonte
Bibliotecária Jane Souto Maior, CRB4-571

A368d Alencar, Francisco Hildemar Calixto de
Diagnóstico de influência para uma família de modelos de
regressão para dados de taxas e proporções / Francisco Hildemar
Calixto de Alencar. – 2016.
90 f.: il., fig., tab.

Orientador: Raydonal Ospina Martínez.
Dissertação (Mestrado) – Universidade Federal de
Pernambuco. CCEN, Estatística, Recife, 2016.
Inclui referências.

1. Análise de regressão. 2. Regressão beta. 3. Estatística
aplicada. I. Martínez, Raydonal Ospina (orientador). II. Título.

519.536

CDD (23. ed.)

UFPE- MEI 2016-021

FRANCISCO HILDEMAR CALIXTO DE ALENCAR

**DIAGNÓSTICO DE INFLUÊNCIA PARA UMA FAMÍLIA DE MODELOS DE
REGRESSÃO PARA DADOS DE TAXAS E PROPORÇÕES**

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 03 de fevereiro de 2016.

BANCA EXAMINADORA

Prof. Dr. Raydonal Ospina Martínez
UFPE

Prof. Dr. Abraão David Costa do Nascimento (Examinador Interno)
UFPE

Prof.^a Dra. Michelli Karinne Barros da Silva (Examinador Externo)
UFCG

Agradecimentos

Primeiramente agradeço a Deus, por ter permitido mais esta jornada em minha vida e por sempre cuidar de minha família.

De forma incomensurável agradeço a minha mãe, Acacia Coelho, por todo amor incondicional, por seu sempre apoio, incentivo, carinho, valiosos conselhos, por sempre ter sonhado junto comigo e por sempre me transmitir força e coragem.

À minha irmã, o anjo de minha vida, Dafinig de Paula, sou eternamente grato por seu carinho, amor, apoio e principalmente por você sempre ter sonhado meus sonhos e nunca ter desistido deles.

Ao pastor Carlos Alberto, por ter sido um pai em minha vida e nunca ter medido esforços para ajudar a mim e a minha casa, sempre lhe serei grato.

Agradeço enormemente ao professor Raydonal Ospina, por seu incentivo, comprometimento, profissionalismo e por sempre chamar minha atenção quando necessário. Agradeço também por seu constante entusiasmo, motivação e por seus conselhos que vão além da vida acadêmica.

Sou muitíssimo grato ao professor Juvêncio Nobre, por seu incentivo, apoio, paciência e por seus conselhos, que muitas vezes abriram meus olhos. Agradeço por ter acreditado em mim.

Agradeço ao meu amigo Everton Souza, sou grato por sua amizade, generosidade, companheirismo e por sempre estar disposto a servir e a ouvir seus amigos.

À minha amiga Laura, sou muito grato por seu companheirismo, apoio e incentivo.

Agradeço a todos os amigos que fiz no decorrer do curso, em especial, Jonas, Pedro, Stênio, Rodrigo goiano, Rodrigo carioca, Diego Coelho, Wanessa, Rafaela, Terezinha, Telma, Fernanda e Renata.

Agradeço aos amigos de moradia Jussieide e Leomir, pelas sempre animadas refeições de fim de semana e momentos de descontração.

À Valéria, secretária do programa, por está sempre disposta a orientar, ajudar e esclarecer as dúvidas.

À CAPES, pelo apoio financeiro. Ao membros da banca examinadora, por suas valiosas sugestões.

“Do Senhor é a terra e a sua plenitude,
o mundo e todos os que nele habitam.”

Salmos 24:1

Existem situações na modelagem estatística em que a variável de interesse é contínua e restrita no intervalo aberto $(0, 1)$, tais como taxas e proporções. Esses tipos de variáveis tipicamente apresentam características de assimetria e heteroscedasticidade, sendo assim inapropriado o uso do modelo normal linear. Kieschnick e McCullough (2003) indicaram após estudos de diferentes estratégias para modelar tais variáveis, o uso do modelo de regressão beta. Contudo, Hahn (2008) e García et al. (2011) observaram que a distribuição beta não é apropriada para o caso em que há ocorrência de eventos extremos; isto é, eventos que possam ocorrer na cauda da distribuição. Com o intuito de obter maior flexibilidade no modelo de regressão beta, Bayes et al. (2012) propuseram o modelo de regressão beta retangular considerando a distribuição beta retangular proposta por Hahn (2008). Este modelo possui como casos particulares o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) e o modelo de regressão beta com dispersão variável proposto por Smithson e Verkuilen (2006). Esta dissertação tem como proposta avaliar o uso das divergências Kullback-Leibler e χ^2 , bem como, das distâncias estocásticas Kullback-Leibler, χ^2 , Bhattacharyya, Hellinger, triangular e média-harmônica e da distância L_1 norm na detecção de observações atípicas nos modelos de regressão beta e beta retangular. Com este fim, realizamos um estudo de simulação de Monte Carlo em que ajustamos, sob o enfoque Bayesiano esses dois modelos. Nesse estudo, observamos que a divergência χ^2 demonstrou maior eficiência, que as demais medidas, na detecção de observações atípicas. A introdução dos pontos atípicos foi feita em ambas as variáveis, dependente e regressora. Por fim, apresentamos uma aplicação utilizando o conjunto de dados AIS (*Australian Institute of Sport*).

Palavras-chave: Robustez. Pontos atípicos. Distâncias estocásticas.

Abstract

There are situations in the statistical modeling where the interested variable is continuous and restricted in the open interval $(0,1)$ as rates and proportions. This type of variables typically show characteristics of asymmetry and heterocedasticity, this way unappropriated the use of the linear normal model. Kieschnick e McCullough (2003) after studies of different strategies to model variables of rates and proportions indicate the use of the regression model based on the beta distribution. However, Hahn (2008) and García et al. (2011) observed which the beta distribution is not appropriated for case where there are events in the tail of the distribution. In order to obtain greater flexibility in the beta regression model Bayes et al. (2012) proposed the rectangular beta regression model based on rectangular beta distribution proposed by Hahn (2008). This model has as particular cases the beta regression model proposed by Ferrari e Cribari-Neto (2004) and the beta regression model with variable dispersion proposed by Smithson e Verkuilen (2006). This dissertation has the purpose of evaluate the use of divergences Kullback-Leibler and χ^2 as well of stochastic distances Kullback-Leibler, χ^2 , Bhattacharyya, Hellinger, triangular and média-harmônica, and distance L_1 norm in detecting atypical points in beta regression model and beta rectangular regression model. To this end, we conducted a Monte Carlo simulation study in which fitted under the Bayesian approach, these two models. This study showed that the difference χ^2 demonstrated higher efficiency than the other measures, the detection of atypical observations. The introduction of atypical points was carried out in both variables, dependent and independent. Finally, we present an application using the set of AIS (Australian Institute of Sport).

Keywords: Robustness. Atypical points. Stochastic distances.

Lista de Figuras

2.1	Densidades beta para diferentes valores de μ (indicados no gráfico), com $\phi = 5$ gráfico à esquerda e $\phi = 100$ gráfico à direita.	19
2.2	Densidades e distribuições beta retangular com $\theta = 0$ (linha sólida), $\theta = 0.3$ (linha tracejada), $\theta = 0.6$ (linha pontilhada), e $\theta = 0.9$ (linha tracejada com pontos).	23
2.3	Região de definição de (θ, γ)	24
2.4	Densidades da distribuição beta retangular reparametrizada, com $\alpha = 0$ (linha sólida), $\alpha = 0.3$ (linha tracejada), $\alpha = 0.6$ (linha pontilhada) e $\alpha = 0.9$ (linha tracejada com pontos).	24
5.1	Diagrama de dispersão e histograma da variável resposta, gerada da distribuição beta, sem contaminação.	44
5.2	Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta, com 3% de contaminação.	45
5.3	Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta, com 6% de contaminação.	45
5.4	Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta, com 10% de contaminação.	46
5.5	Diagrama de dispersão e histograma da variável resposta, gerada da distribuição beta retangular, sem contaminação.	46
5.6	Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta retangular, com 3% de contaminação.	47
5.7	Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta retangular, com 6% de contaminação.	47
5.8	Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta retangular, com 10% de contaminação.	48
5.9	Diagrama de dispersão e histograma da variável explicativa, sendo a variável resposta gerada da distribuição beta, sem contaminação.	49

5.10	Diagramas des dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta, com 3% de contaminação.	49
5.11	Diagramas des dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta, com 6% de contaminação.	50
5.12	Diagramas des dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta, com 10% de contaminação.	50
5.13	Diagrama de dispersão e histograma da variável explicativa, sendo a variável resposta gerada da distribuição beta retangular, sem contaminação.	51
5.14	Diagramas des dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta retangular, com 3% de contaminação. . . .	51
5.15	Diagramas des dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta retangular, com 6% de contaminação. . . .	52
5.16	Diagramas des dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta retangular, com 10% de contaminação. . . .	52
5.17	Porcentagem de detecção de observações atípicas no cenário 1.	64
5.18	Porcentagem de detecção de observações atípicas no cenário 2.	65
5.19	Porcentagem de detecção de observações atípicas no cenário 3.	66
5.20	Porcentagem de detecção de observações atípicas no cenário 4.	67
5.21	Porcentagem de detecção de observações atípicas no cenário 5.	68
5.22	Porcentagem de detecção de observações atípicas no cenário 6.	69
5.23	Porcentagem de detecção de observações atípicas no cenário 7.	70
5.24	Porcentagem de detecção de observações atípicas no cenário 8.	71
6.1	Boxplot e histograma da variável percentual de gordura corporal.	73
6.2	Gráfico de dispersão entre as PGC e MCM referente aos 37 atletas que praticam remo.	73
6.3	$\widehat{CPO}$ do modelo de regressão beta ajustado.	74
6.4	Gráficos das medidas utilizadas na identificação de observações atípicas no ajuste do modelo de regressão beta.	75
6.5	$\widehat{CPO}$ do modelo de regressão beta ajustado sem a observação 13.	76
6.6	Gráficos das medidas utilizadas na identificação de observações atípicas no ajuste do modelo de regressão beta sem a observação 13.	77
6.7	Gráfico de dispersão entre as variáveis PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta ajustado com e sem a observação 13.	78
6.8	Gráficos das medidas utilizadas na identificação de observações atípicas no ajuste do modelo de regressão beta retangular.	79
6.9	$\widehat{CPO}$ do modelo de regressão beta retangular ajustado.	80
6.10	Gráfico de dispersão entre as PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta retangular ajustado com e sem a observação 13.	81

6.11	$\widehat{CPO}$ do modelo de regressão beta retangular ajustado sem a observação 13. . . .	81
6.12	Gráficos das medidas utilizadas na identificação de observações atípicas no ajuste do modelo de regressão beta retangular ajustado sem a observação 13.	82
6.13	Gráfico de dispersão entre as PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta e beta retangular ajustado com todas as observações.	84
6.14	Gráfico de dispersão entre as PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta e beta retangular ajustado sem a observação 13.	84

Lista de Tabelas

3.1	Valores de corte do fator de Bayes a favor do modelo M_1	31
4.1	Funções h e ϕ e suas respectivas distâncias.	38
5.1	Esquema de geração dos 8 cenários da simulação.	53
5.2	Expressões e valores dos pontos de corte das medidas utilizadas na detecção de observações atípicas.	54
5.3	Porcentagens das vezes que as medidas indicam observações atípicas no cenário 1.	55
5.4	Porcentagens das vezes que as medidas indicam observações atípicas no cenário 2.	56
5.5	Porcentagens das vezes que as medidas indicam observações atípicas no cenário 3.	57
5.6	Porcentagens das vezes que as medidas indicam observações atípicas no cenário 4.	58
5.7	Porcentagens das vezes que as medidas indicam observações atípicas no cenário 5.	59
5.8	Porcentagens das vezes que as medidas indicam observações atípicas no cenário 6.	60
5.9	Porcentagens das vezes que as medidas indicam observações atípicas no cenário 7.	62
5.10	Porcentagens das vezes que as medidas indicam observações atípicas no cenário 8.	63
6.1	Média <i>a posteriori</i> e intervalo de credibilidade 95% para os parâmetros dos modelos de regressão beta e beta retangular para os atletas de remo do conjunto de dados IAS usando todas as observações e sem a observação 13.	83

1	Introdução	15
1.1	Motivação	15
1.2	Estrutura da dissertação	16
2	Modelos de regressão beta e beta retangular	17
2.1	Distribuição beta	18
2.2	Modelo de regressão beta	19
2.3	Distribuição beta retangular	22
2.3.1	Reparametrização	23
2.4	Modelo de regressão beta retangular reparametrizado sob o enfoque Bayesiano	25
2.4.1	Estimação Bayesiana	25
3	Enfoque Bayesiano para avaliação de modelos paramétricos	27
3.1	Distribuição preditiva <i>a posteriori</i>	27
3.1.1	Distribuição preditiva	27
3.1.2	Distribuição preditiva <i>a posteriori</i>	28
3.2	CrITÉrios de comparação de modelos	28
3.2.1	Ordem condicional preditiva (CPO)	29
3.2.2	Fator de Bayes	30
3.2.3	CrITÉrio de informação do desvio	33
4	Distâncias estocásticas	35
4.1	Entropia	35
4.2	Distâncias estocásticas	36
5	Simulação	43
5.1	Cenário 1	54
5.2	Cenário 2	55

5.3	Cenário 3	56
5.4	Cenário 4	58
5.5	Cenário 5	59
5.6	Cenário 6	60
5.7	Cenário 7	61
5.8	Cenário 8	62
5.9	Gráficos referentes as simulações	63
6	Aplicação	72
6.1	Modelo de regressão beta	73
6.1.1	Modelo de regressão beta sem a observação 13	75
6.2	Modelo de regressão beta retangular	78
6.2.1	Modelo de regressão beta retangular sem a observação 13	80
6.3	Conclusão	82
7	Considerações finais	85
	Referências	86

1.1 Motivação

Na modelagem estatística de variáveis resposta contínuas e restritas no intervalo aberto $(0, 1)$, como é o caso das taxas e proporções, não é apropriado o uso do modelo de regressão normal linear, visto que é esperado a presença de assimetria e heteroscedasticidade. Em situações como essa, costuma-se transformar a variável resposta com a desvantagem do modelo poder produzir predições errôneas e de difícil interpretação dos seus parâmetros em termos da variável original.

Devido a essas dificuldades, Kieschnick e McCullough (2003) analisaram modelos para dados de taxas e proporções. Ao final de seus estudos, eles concluíram que o modelo com base na distribuição beta é o mais apropriado para tais casos. Ferrari e Cribari-Neto (2004) introduziram um modelo de regressão beta que estruturalmente se assemelha aos MLG's e os parâmetros podem ser interpretados em termos da resposta média.

A função densidade de probabilidade da distribuição beta pode assumir diferentes formas conforme os valores de seus parâmetros (FERRARI e CRIBARI-NETO, 2004). Hahn (2008) e García et al. (2011) observaram que essa distribuição não modela bem a ocorrência de eventos extremos. Segundo Bayes et al. (2012) esse fato pode ser um limitador para o uso dessa distribuição na modelagem de taxas e proporções. Devido a esse fato, esses autores propuseram um modelo de regressão considerando a distribuição beta retangular proposta por Hahn (2008) que pode ser visto como uma mistura de uma distribuição beta com uma distribuição uniforme. Markatou (2000) comenta que distribuições mistas são mais robustas a ocorrência de observações extremas. Desse modo, em situações em que há ocorrência de observações extremas o modelo de regressão beta retangular pode ser

mais adequado que o modelo de regressão beta.

No contexto de análise de diagnóstico, temos que a primeira técnica desenvolvida para avaliar o impacto individual de uma observação sobre o processo de estimação dos parâmetros de um modelo, se baseia na retirada individual de cada observação, e posteriormente os efeitos destas retiradas são estudados (COOK e WEISBERG, 1982). Para a classe de modelos de regressão beta, Espinheira et al. (2008a, 2008b) desenvolveram métodos de diagnóstico para avaliar a influência de observações. Bayes et al. (2012) realizam um pequeno estudo, utilizando a divergência Kullback-Leibler, para avaliar a influência de observações no ajuste dos de regressão beta e beta retangular.

Recentemente, Nascimento et al. (2010) introduziram o conceito de distâncias estocásticas para a comparação de variáveis aleatórias e consequentemente para a comparação de modelos paramétricos. Nesse trabalho focamos nossa atenção ao uso desse conceito, assim, estendendo o estudo realizado por Bayes et al. (2012), para avaliar e identificar a presença de observações atípicas nos modelos de regressão beta e beta retangular.

1.2 Estrutura da dissertação

O presente trabalho é constituído de sete capítulos. No Capítulo 2 apresentamos uma breve introdução dos modelos de regressão beta e beta retangular com o objetivo de motivar o uso de algumas medidas de divergência e distâncias estocásticas na detecção de observações atípicas. No Capítulo 3 discorremos sobre alguns critérios para avaliação de modelos paramétricos desde uma perspectiva Bayesiana. No Capítulo 4 apresentamos alguns conceitos e propriedades das medidas de divergências e distâncias estocásticas. No Capítulo 5 discutimos os resultados de um estudo simulação a fim de avaliar o desempenho das medidas de divergências e distâncias estocásticas apresentadas no contexto de diagnóstico de influência nos modelos aqui apresentados. No Capítulo 6 apresentamos aplicações em que ajustamos os modelos de regressão beta e beta retangular a um mesmo conjunto de dados e verificamos o comportamento das medidas de divergência e distâncias estocásticas na detecção de observações atípicas, em particular de observações influentes. No Capítulo 7 concluímos o trabalho com algumas considerações finais.

Modelos de regressão beta e beta retangular

O modelo normal linear é usado para investigar e modelar uma relação linear entre uma variável dependente (variável resposta) e uma ou mais variáveis explicativas (covariáveis). Para o uso desse modelo são feitas algumas suposições, tais como:

1. As fontes de variação são homoscedásticas e não correlacionadas;
2. Para análise inferencial é assumido que os erros são normalmente distribuídos.

Entretanto, em algumas situações na modelagem por regressão, a variável resposta é contínua e restrita no intervalo aberto $(0, 1)$, como por exemplo, taxas e proporções. Nesses casos a modelagem linear clássica não é apropriada, pois tipicamente a variável resposta apresenta características de assimetria e heterocedasticidade, o que não é esperado no caso do modelo normal.

Uma solução possível para tal situação é a transformação da variável dependente, para em seguida modelarmos sua média transformada através de um preditor linear. Essa abordagem, contudo, tem desvantagens, uma delas é o fato de que os parâmetros do modelo não podem ser facilmente interpretados em termos da variável resposta original. Outra desvantagem é que, dada a natureza limitada dos dados, o modelo pode conduzir a inferências equivocadas.

Neste contexto, Kieschnick e McCullough (2003) trabalharam com diferentes estratégias de modelagem para lidar com dados contidos no intervalo aberto $(0, 1)$. Esses autores abordaram os modelos normal linear, normal não-linear, normal censurado, logito e os modelos que usam a distribuição beta, simplex e de quasi-verossimilhança. Eles concluíram que dado que a variável resposta assume valores num intervalo limitado, sua média deve ser

uma função não-linear nos parâmetros de regressão com estrutura de dispersão geralmente heteroscedástica. Nesse caso, eles advogam o uso de um modelo de regressão baseado na distribuição beta, comparativamente aos modelos supracitados.

Na literatura é possível encontrar diferentes especificações de modelos de regressão que usam a distribuição beta, dentre eles podemos citar, Paolino (2001), Kieschnick e McCullough (2003), Ferrari e Cribari-Neto (2004) e Smithson e Verkuilen (2006) propuseram modelos paramétricos de regressão, em que assume-se que a resposta média é relacionada com um preditor linear por meio de uma função de ligação. O trabalho de Ferrari e Cribari-Neto (2004) apresenta algumas vantagens das quais é possível destacar: a especificação da estrutura de regressão é similar à classe dos modelos lineares generalizados, em que modela-se diretamente a média ao invés dos dois parâmetros que indexam a distribuição e a função de ligação que relaciona a resposta média ao preditor linear é geral. No trabalho de Ferrari e Cribari-Neto (2004) destaca-se o desenvolvimento de estimação pontual e intervalar, testes de hipóteses e medidas simples de diagnóstico. Smithson e Verkuilen (2006) propõem a modelagem da dispersão do modelo de regressão beta assumindo uma relação com um preditor linear através de uma adequada função de ligação. Espinheira et al. (2008a, 2008b) desenvolveram métodos de diagnóstico para avaliar a influência de observações na classe de modelos de regressão beta. Huang e Oosterlee (2008) propuseram um modelo de regressão beta generalizado misto com efeito aleatório no preditor linear. Simas et al. (2010) propuseram uma variação do modelo de regressão beta que introduz não-linearidade e dispersão variável. Implementações no software livre R do modelo de regressão beta são discutidas em Cribari-Neto e Zeileis (2010). Ospina e Ferrari (2012) propuseram modelos de regressão beta para dados inflacionados.

Com relação à abordagem Bayesiana, Branscum et al. (2007) trabalharam com um modelo de regressão beta hierárquico, Da-Silva et al. (2011) apresentam um modelo beta dinâmico Bayesiano para modelagem de séries temporais de taxas e proporções, Figueroa-Zúñiga et al. (2013) expandem o modelo proposto por Ferrari e Cribari-Neto (2004) adicionando efeitos aleatórios com distribuição normal e t -Student nas estruturas de regressão da média. Além disso, eles discutem a especificação de distribuições *a priori* e implementações computacionais do modelo por meio do amostrador de Gibbs.

2.1 Distribuição beta

Uma variável aleatória Y segue a distribuição beta se sua função densidade de probabilidade (fdp) é dada por

$$f_Y(y|\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu)\Gamma(\phi(1-\mu))} y^{\phi\mu-1} (1-y)^{\phi(1-\mu)-1} \mathbf{1}_{(0,1)}(y), \quad (2.1)$$

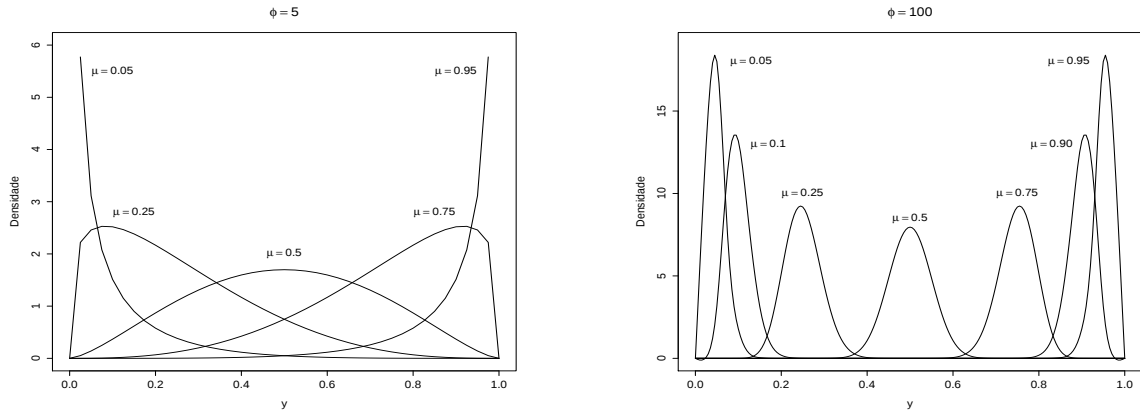
em que $0 < \mu < 1$, $\phi > 0$, $\Gamma(\cdot)$ denota a função gama e $\mathbb{1}_{(0,1)}(y)$ denota a função indicadora que é igual a 1 se $y \in (0, 1)$ e igual a 0 caso contrário. Usaremos a notação $Y \sim \mathcal{B}(\mu, \phi)$ para indicar que a variável Y tem distribuição beta de parâmetros μ e ϕ . A média e a variância da variável aleatória Y são dadas por

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

em que $V(\mu) = \mu(1 - \mu)$ denota a “função de variância” e ϕ pode ser interpretado como um parâmetro de precisão, no sentido de que, para μ fixo, se o valor de ϕ aumenta, a variância de Y diminui.

Na Figura 2.1, temos algumas densidades beta. É possível notar que as densidades podem assumir diferentes formas que dependem dos valores de μ e ϕ . Em geral, a densidade é simétrica quando $\mu = 1/2$ e assimétrica quando $\mu \neq 1/2$. Note, também, que a densidade assume formas bem variadas, acomodando distribuições simétricas, assimétricas, em formas de “J” e de “J” invertido.

Figura 2.1: Densidades beta para diferentes valores de μ (indicados no gráfico), com $\phi = 5$ gráfico à esquerda e $\phi = 100$ gráfico à direita.



Fonte: Dados próprios.

2.2 Modelo de regressão beta

Nessa seção, focaremos nossa atenção no modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004), mas sendo analisada sua estimação sob a perspectiva Bayesiana adotada por Souza (2011). Seja $Y_1, \dots, Y_n$ uma amostra aleatória em que

$$Y_t \sim \mathcal{B}(\mu_t, \phi), \quad (2.2)$$

e a componente sistemática é tal que para $x_t^\top = (x_{t1}, \dots, x_{tk})$ sendo o vetor de covariáveis associadas a y_t ,

$$g(\mu_t) = x_t^\top \boldsymbol{\beta} = \eta_t, \quad t = 1, \dots, n.$$

Para esta abordagem assumimos a existência de uma distribuição a *priori* $\mathbb{P}(\boldsymbol{\beta}, \phi)$ para $\boldsymbol{\beta}$ e ϕ , em que $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)^\top$ é o vetor dos coeficientes de regressão associados às variáveis explicativas e ϕ é o coeficiente associado à precisão do modelo. A densidade da distribuição beta para a t -ésima observação é dada por

$$f(y_t | \boldsymbol{\beta}, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_t(\boldsymbol{\beta})\phi)\Gamma((1 - \mu_t(\boldsymbol{\beta}))\phi)} y_t^{\mu_t(\boldsymbol{\beta})\phi-1} (1 - y_t)^{(1 - \mu_t(\boldsymbol{\beta}))\phi-1}, \quad (2.3)$$

em que $\mu_t(\boldsymbol{\beta}) = g^{-1}(x_t^\top \boldsymbol{\beta})$. A função de verossimilhança $L(\boldsymbol{\beta}, \phi)$ é

$$L(\boldsymbol{\beta}, \phi) = (\Gamma(\phi))^n \prod_{t=1}^n \frac{1}{\Gamma(\mu_t(\boldsymbol{\beta})\phi)\Gamma((1 - \mu_t(\boldsymbol{\beta}))\phi)} y_t^{\mu_t(\boldsymbol{\beta})\phi-1} (1 - y_t)^{(1 - \mu_t(\boldsymbol{\beta}))\phi-1}. \quad (2.4)$$

A densidade a *posteriori* dos parâmetros $\boldsymbol{\beta}$ e ϕ é obtida por

$$\begin{aligned} \mathbb{P}(\boldsymbol{\beta}, \phi | \mathbf{y}) &\propto L(\boldsymbol{\beta}, \phi) \mathbb{P}(\boldsymbol{\beta}, \phi) \\ &\propto (\Gamma(\phi))^n \prod_{t=1}^n \frac{1}{\Gamma(\mu_t(\boldsymbol{\beta})\phi)\Gamma((1 - \mu_t(\boldsymbol{\beta}))\phi)} y_t^{\mu_t(\boldsymbol{\beta})\phi-1} (1 - y_t)^{(1 - \mu_t(\boldsymbol{\beta}))\phi-1} \mathbb{P}(\boldsymbol{\beta}, \phi). \end{aligned}$$

Nesta aproximação é considerada a independência a *priori* dos parâmetros, isto é, $\mathbb{P}(\boldsymbol{\beta}, \phi) = \mathbb{P}(\boldsymbol{\beta})\mathbb{P}(\phi)$. Para esses parâmetros podemos adotar, por exemplo, $\beta_i \sim \mathcal{N}(m_i, \sigma_i^2)$ e $\phi \sim \Gamma(a, b)$, pois $\beta_k \in \mathbb{R}$ e $\phi > 0$. Os termos envolvendo $\Gamma(\mu_t(\boldsymbol{\beta})\phi)$ e $\Gamma(\phi)$ indicam que a densidade a *posteriori* de $\boldsymbol{\beta}$ e ϕ não possui forma fechada. Deste modo, para estimação desde uma perspectiva Bayesiana é necessário o uso do método de Monte Carlo via Cadeias de Markov (MCMC) para gerar amostras destas distribuições. Quando a distribuição a *posteriori* dos parâmetros não possuem forma fechada é recomendado uso do algoritmo de Metropolis Hastings (CHIB e GREENBERG, 1995) para gerar amostras dessas distribuições. Gamerman e Lopes (2006) apresentam detalhes desse algoritmo de estimação no contexto de modelos paramétricos. As distribuições condicionais completas a *posteriori* de ϕ e $\boldsymbol{\beta}$ são, respectivamente, dadas por

$$\begin{aligned} \mathbb{P}(\phi | \boldsymbol{\beta}, \mathbf{y}) &\propto L(\boldsymbol{\beta}, \phi) \mathbb{P}(\phi) \\ &\propto (\Gamma(\phi))^n \prod_{t=1}^n \frac{1}{\Gamma(\mu_t(\boldsymbol{\beta})\phi)\Gamma((1 - \mu_t(\boldsymbol{\beta}))\phi)} y_t^{\mu_t(\boldsymbol{\beta})\phi-1} (1 - y_t)^{(1 - \mu_t(\boldsymbol{\beta}))\phi-1} \exp(-b\phi) \end{aligned}$$

e

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\beta}_i | \boldsymbol{\beta}_{-i}, \phi, \mathbf{y}) &\propto L(\boldsymbol{\beta}, \phi) \mathbb{P}(\boldsymbol{\beta}_i) \\
&\propto (\Gamma(\phi))^n \prod_{t=1}^n \frac{1}{\Gamma(\mu_t(\boldsymbol{\beta})\phi) \Gamma((1 - \mu_t(\boldsymbol{\beta}))\phi)} y_t^{\mu_t(\boldsymbol{\beta})\phi - 1} (1 - y_t)^{(1 - \mu_t(\boldsymbol{\beta}))\phi - 1} \\
&\times \exp \left\{ -\frac{1}{2} \sum_{i=1}^k \left(\frac{\beta_i - m_i}{\sigma_i} \right)^2 \right\}.
\end{aligned}$$

Souza (2011) argumenta que os coeficientes de regressão são correlacionados em modelos de regressão de beta, assim distribuições *a priori* que contemplem alguma estrutura de correlação nos coeficientes podem fornecer estimativas mais precisas que as obtidas com o uso de distribuições *a priori* independentes.

Considerando as ferramentas de diagnóstico para o modelo de regressão beta, Ferrari e Cribari-Neto (2004) forneceram uma metodologia de diagnóstico desde uma perspectiva frequentista que inclui o uso de resíduos padronizados. Espinheira et al. (2008a) também propuseram resíduos para a regressão beta e Espinheira et al. (2008b) propuseram medidas de análise de influência local. No entanto, nenhum estudo foi realizado para demonstrar que a estimativa dos parâmetros da regressão não é influenciada quando há observações atípicas na cauda da distribuição da variável resposta.

A distribuição beta pode ser considerada flexível uma vez que sua fdp pode assumir diferentes formas, considerando os diferentes valores dos parâmetros. No entanto, como foi observado por Hahn (2008) e García et al. (2011), a distribuição beta não modela bem a ocorrência de eventos extremos, isto é, eventos que possam distorcer o comportamento da cauda da distribuição. Além disso, a beta não permite uma maior flexibilidade na especificação da sua função de variância. Este fato pode limitar algumas aplicações na modelagem de proporções (BAYES et al., 2012).

A fim de obter maior flexibilidade no modelo de regressão beta no sentido de permitir a modelagem da dispersão e a ocorrência de eventos extremos em dados de taxas e proporções, Bayes et al. (2012) propuseram um modelo de regressão baseado na distribuição beta retangular proposta por Hahn (2008). O modelo de regressão que eles consideraram inclui como casos particulares o modelo de regressão beta e o modelo de regressão beta com dispersão variável.

A distribuição beta retangular pode ser vista como um modelo de mistura, nesta situação Markatou (2000) comenta que as distribuições mistas (ou de mistura) são mais robustas a presença de observações extremas, a variabilidade é melhor explicada e as estimativas dos parâmetros são menos afetadas pelos valores extremos. Isso indica que a regressão beta retangular é um modelo que pode ser mais robusto em comparação com o modelo de regressão beta tradicional.

2.3 Distribuição beta retangular

A distribuição beta retangular é uma mistura de uma distribuição beta, com uma distribuição uniforme. Formalmente, uma variável aleatória Y é dita ter distribuição beta retangular se sua fdp for da forma

$$f_Y(y|\mu, \phi, \theta) = \theta + (1 - \theta) \frac{\Gamma(\phi)}{\Gamma(\mu\phi, \phi(1 - \mu))} y^{\mu\phi-1} (1 - y)^{\phi(1-\mu)-1} \mathbb{1}_{(0,1)}(y). \quad (2.5)$$

em que $\mathcal{B}(\mu, \phi)$ representa a fdp da distribuição beta com parâmetros μ e ϕ . Aqui $\mu \in (0, 1)$, $\phi > 0$ e $\theta \in [0, 1]$ é o parâmetro de mistura. As distribuições uniforme e beta são obtidas, respectivamente, se $\theta = 1$ e se $\theta = 0$. Adotaremos a notação $Y \sim \mathcal{BR}(\theta, \mu, \phi)$ para indicar que Y tem distribuição beta retangular. A média e a variância de Y são dadas por

$$\mathbb{E}(Y|\mu, \phi, \theta) = \frac{\theta}{2} + (1 - \theta)\mu, \quad (2.6)$$

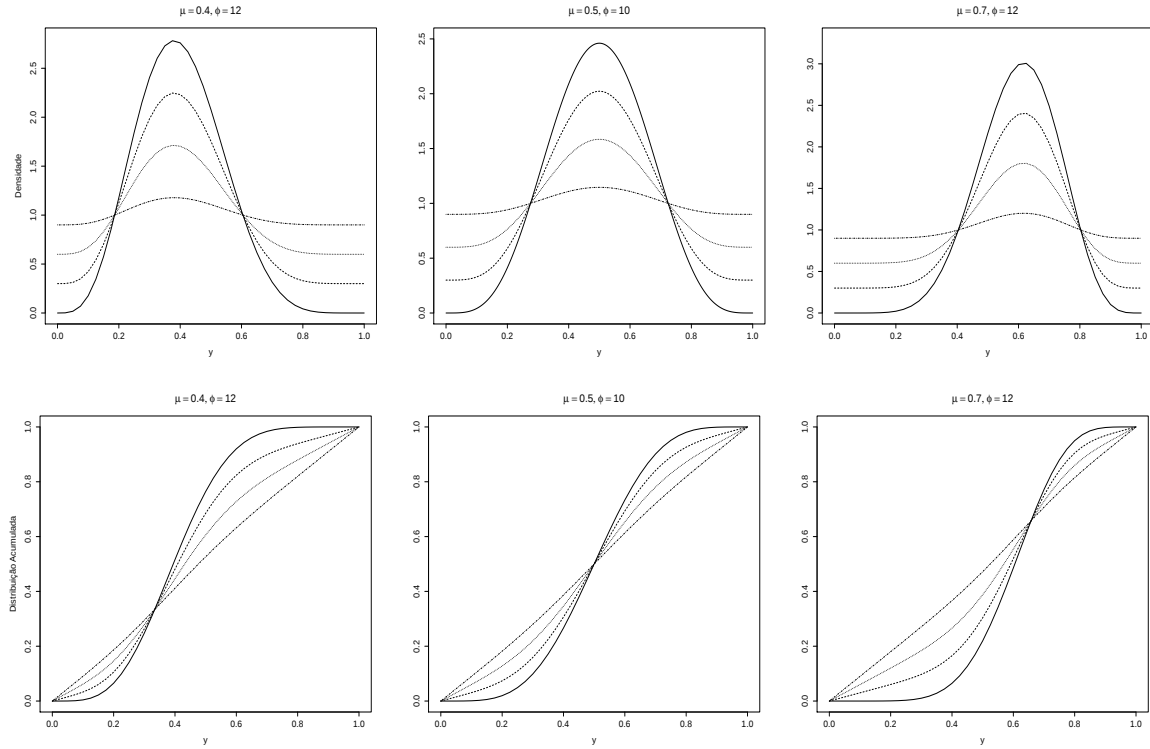
e

$$\text{Var}(Y|\mu, \phi, \theta) = \frac{V(\mu)}{1 + \phi} (1 - \theta)(1 - \theta(1 + \phi)) + \frac{\theta}{12} (4 - 3\theta). \quad (2.7)$$

Note que quando $\theta = 0$ as expressões (2.6) e (2.7) coincidem com as expressões da média e da variância da distribuição beta. E de modo semelhante, quando $\theta = 1$, as expressões coincidem com a média e variância da distribuição uniforme.

Na Figura 2.2 apresentamos algumas densidades beta retangular e também algumas funções de distribuição acumulada (fda), variando os valores de μ , ϕ e θ . Observamos que de modo geral, a densidade da distribuição beta retangular é unimodal, e para $\gamma = 0,5$ ela é simétrica e para $\gamma \neq 0,5$ a densidade é assimétrica.

Figura 2.2: Densidades e distribuições beta retangular com $\theta = 0$ (linha sólida), $\theta = 0.3$ (linha tracejada), $\theta = 0.6$ (linha pontilhada), e $\theta = 0.9$ (linha tracejada com pontos).



Fonte: Dados próprios.

2.3.1 Reparametrização

No caso da distribuição beta retangular temos que sua média é uma função dos parâmetros θ e μ . A fim de obter uma estrutura de regressão mais adequada para a média da distribuição, Bayes et al. (2012) propuseram a seguinte reparametrização

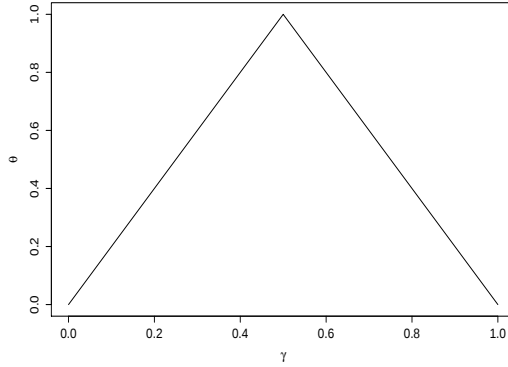
$$\mathbb{E}(Y) = \gamma = \frac{\theta}{2} + (1 - \theta)\mu \quad (2.8)$$

e parâmetro de forma

$$\alpha = \frac{\theta}{1 - (1 - \theta)|2\mu - 1|}. \quad (2.9)$$

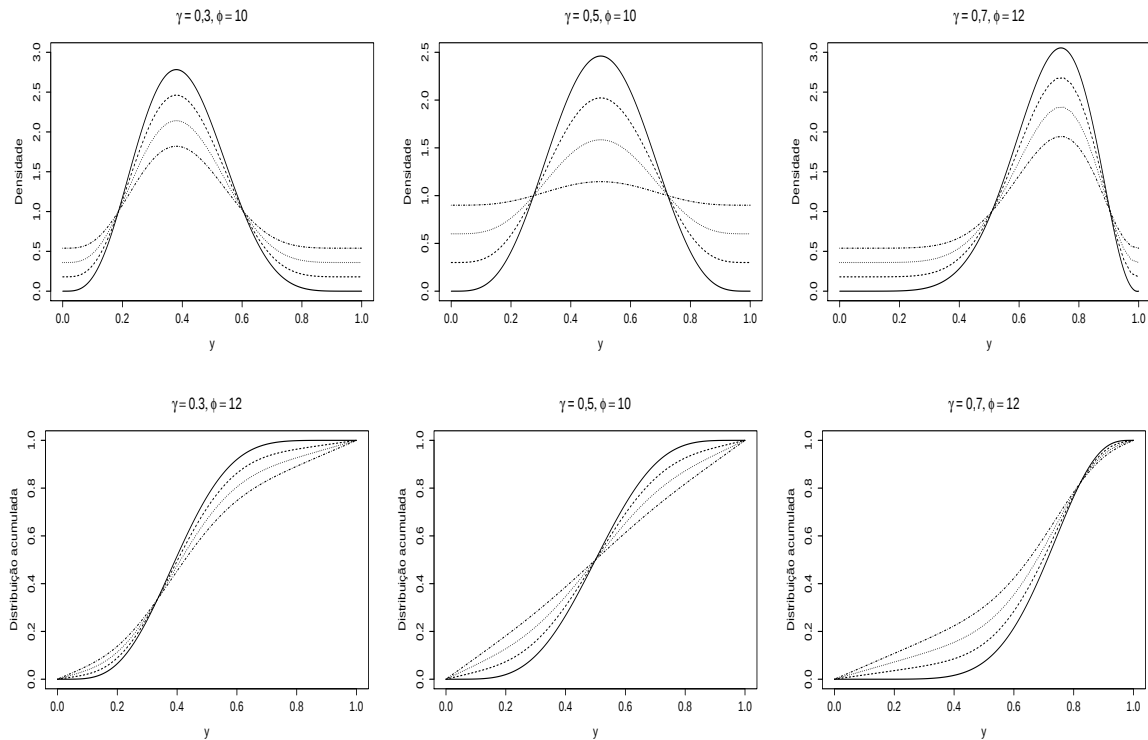
O espaço paramétrico de γ e α é o retângulo dado por $\{0 \leq \gamma \leq 1, 0 \leq \alpha \leq 1\}$, na Figura 2.3 temos a representação gráfica do espaço paramétrico de θ , que é dado por $0 < \theta < 1 - |2\gamma - 1|$. Sob esta parametrização a fdp da distribuição beta retangular fica dada por

$$\begin{aligned} g_Y(y|\gamma, \phi, \alpha) &= \alpha(1 - |2\gamma - 1|) + (1 - \alpha(1 - |2\gamma - 1|)) \\ &\times \mathcal{B}\left(\frac{\gamma - 0, 5\alpha(1 - |2\gamma - 1|)}{1 - \alpha(1 - |2\gamma - 1|)}, \phi\right) \mathbf{1}_{(0,1)}(y), \end{aligned} \quad (2.10)$$

Figura 2.3: Região de definição de (θ, γ) .

Fonte: Dados próprios.

em que $\mathcal{B}\left(\frac{\gamma-0.5\alpha(1-|2\gamma-1|)}{1-\alpha(1-|2\gamma-1|)}, \phi\right)$ representa uma fdp da distribuição beta. Vamos adotar a notação $Y \sim \mathcal{BRR}(\gamma, \phi, \alpha)$ para especificar a distribuição beta retangular reparametrizada, com média γ . Na Figura 2.4 apresentamos algumas densidades desta distribuição para diferentes valores de seus parâmetros. Pode-se notar que α é um parâmetro de forma que está associado ao peso das caudas da distribuição e ϕ é um parâmetro que controla a precisão da distribuição.

Figura 2.4: Densidades da distribuição beta retangular reparametrizada, com $\alpha = 0$ (linha sólida), $\alpha = 0.3$ (linha tracejada), $\alpha = 0.6$ (linha pontilhada) e $\alpha = 0.9$ (linha tracejada com pontos).

Fonte: Dados próprios.

2.4 Modelo de regressão beta retangular reparametrizado sob o enfoque Bayesiano

O modelo de regressão beta retangular proposto por Bayes et al. (2012) admite a seguinte especificação, seja $Y_1, \dots, Y_n$ uma amostra aleatória com $Y_t \sim \mathcal{BRR}(\gamma_t, \phi_t, \alpha)$, $t = 1, \dots, n$, deste modo temos que

$$\begin{aligned} g(\gamma_t) &= x_t^\top \boldsymbol{\beta}, \\ h(\phi_t) &= w_t^\top \boldsymbol{\delta}, \end{aligned} \quad (2.11)$$

em que $\boldsymbol{\beta}^\top = (\beta_1, \beta_2, \dots, \beta_k)$ e $\boldsymbol{\delta}^\top = (\delta_1, \delta_2, \dots, \delta_l)$ são os vetores dos parâmetros de regressão associados, respectivamente, a $x_t^\top$ e $w_t^\top$, em que $x_t^\top = (x_{t1}, \dots, x_{tk})$ e $w_t^\top = (w_{t1}, \dots, w_{tl})$ são vetores das k e l covariáveis conhecidas. Aqui $g(\cdot)$ e $h(\cdot)$ são funções estritamente monótonas e duas vezes diferenciáveis no intervalo $(0, 1)$.

Em geral, $g(\cdot)$ pode ser qualquer função quantílica correspondente a uma distribuição contínua, em que a inversa é uma função de ligação que relaciona γ_t as covariáveis $x_t^\top$. Alguns exemplos de funções de ligação para $g(\cdot)$ são, logito, probito, log-log, complemento log-log e função Cauchy. Já $h(\cdot)$ é uma função de ligação que relaciona o parâmetro de precisão ϕ_t com as covariáveis $w_t^\top$. Além disso, como os parâmetros ϕ_t devem ser estritamente positivos, uma função de ligação que pode ser usada para $h(\cdot)$ é o $\log(\phi_t) = w_t^\top \boldsymbol{\delta}$.

Temos que a função de verossimilhança do modelo definido por (2.11) é dada por

$$L(\boldsymbol{\eta}|\mathbf{Y}) = \prod_{t=1}^n f_Y(y_t|\mu_t, \phi_t, \theta_t) = \prod_{t=1}^n g_Y(y_t|\gamma_t, \phi_t, \alpha,) \quad (2.12)$$

em que $\boldsymbol{\eta}^\top = (\boldsymbol{\beta}, \boldsymbol{\delta}, \alpha)$, $\mathbf{Y} = (Y_1, \dots, Y_n)$ e sendo γ_t e ϕ_t definidos como em (2.11), já $\theta_t = \alpha(1 - |2\gamma_t - 1|)$ e $\mu_t = \frac{\gamma_t - \theta_t/2}{1 - \theta_t}$.

Note que em (2.11) quando $\alpha = 0$ e ϕ_t é constante, obtemos o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004). Note também que, se apenas $\alpha = 0$ obtemos o modelo de regressão beta com dispersão variável introduzido por Paolino (2001), Smithson e Verkuilen (2006) e por Simas et al. (2010).

2.4.1 Estimação Bayesiana

Sob o aspecto clássico, os estimadores de máxima verossimilhança dos parâmetros do modelo dado em (2.11) podem ser obtidos de uma forma semelhante à proposta por Ferrari e Cribari-Neto (2004). No entanto, Bayes et al. (2012) relatam que em algumas situações empíricas, o tamanho da amostra pode ser pequeno e, em tais casos a inferência

Bayesiana parece ser mais adequada se há informação auxiliar sobre o parâmetro, ver Congdon (2014) e Gelman et al. (2009).

Assumindo que a função de distribuição a *posteriori* dos parâmetros do modelo de regressão beta retangular é dada por

$$\mathbb{P}(\boldsymbol{\eta}|\mathbf{Y}) \propto L(\boldsymbol{\eta}|\mathbf{Y})\mathbb{P}(\boldsymbol{\eta}). \quad (2.13)$$

Em que $\mathbb{P}(\boldsymbol{\eta})$ é a distribuição a *priori* de $\boldsymbol{\eta}$, Bayes et al. (2012) assumem uma distribuição a *priori* para cada uma das quantidades desconhecidas $\boldsymbol{\eta}^\top = (\boldsymbol{\beta}, \boldsymbol{\delta}, \alpha)$, isto é,

$$\mathbb{P}(\boldsymbol{\eta}) = \mathbb{P}(\boldsymbol{\beta})\mathbb{P}(\boldsymbol{\delta})\mathbb{P}(\alpha).$$

Aqui, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ e α são assumidos independentes, e $\boldsymbol{\beta}$ e $\boldsymbol{\delta}$ seguem distribuições $\boldsymbol{\beta} \sim \mathcal{N}_k(\mathbf{a}, \mathbf{B})$, $\boldsymbol{\delta} \sim \mathcal{N}_l(\mathbf{c}, \mathbf{D})$ e $\alpha \sim \mathcal{U}(0, 1)$, respectivamente. Portanto, sob estas condições, temos a seguinte distribuição a *posteriori*

$$\begin{aligned} \mathbb{P}(\boldsymbol{\eta}|\mathbf{Y}) &\propto \prod_{i=1}^n \alpha(1 - |2g(x_i^\top \boldsymbol{\beta}) - 1|) + (1 - \alpha(1 - |2g(x_i^\top \boldsymbol{\beta}) - 1|)) \\ &\times \mathcal{B}\left(\frac{g(x_i^\top \boldsymbol{\beta}) - 0.5\alpha(1 - |2g(x_i^\top \boldsymbol{\beta}) - 1|)}{1 - \alpha(1 - |2g(x_i^\top \boldsymbol{\beta}) - 1|)}, h(w_i^\top \boldsymbol{\delta})\right) \\ &\times \psi_k(\boldsymbol{\beta}|\mathbf{a}, \mathbf{B})\psi_l(\boldsymbol{\delta}|\mathbf{c}, \mathbf{D}), \end{aligned} \quad (2.14)$$

em que $\psi_k(\cdot)$ e $\psi_l(\cdot)$ denotam a função de densidade de uma distribuição normal multivariada de ordem k e l , respectivamente, sendo $\mathbf{B}$ e $\mathbf{D}$ matrizes de posto completo.

No caso particular em que o modelo de regressão não contempla o parâmetro de dispersão, as distribuições a *priori* consideradas para $\boldsymbol{\eta} = (\boldsymbol{\beta}, \phi, \alpha)^\top$ são

$$\mathbb{P}(\boldsymbol{\eta}) = \mathbb{P}(\boldsymbol{\beta})\mathbb{P}(\phi)\mathbb{P}(\alpha),$$

em que $\phi \sim \Gamma(c, c)$ como em Branscum et al. (2007). Após explorações preliminares Bayes et al. (2012) observaram que a inferência sobre os coeficientes de regressão, que são o principal foco de interesse, são semelhantes para diferentes valores dos hiperparâmetros.

Enfoque Bayesiano para avaliação de modelos paramétricos

Para a avaliação dos modelos ajustados focaremos nossa atenção na distribuição preditiva. Box (1980) argumenta que a adequação de um modelo não pode ser avaliada a partir da distribuição *a posteriori* dos parâmetros do modelo e por isso o enfoque preditivo pode ser uma adequada alternativa. Algumas referências importantes a respeito deste assunto são: Jeffreys (1961), Geisser (1985), Ibrahim e Laud (1994), que introduziram na análise de experimentos o critério de avaliação de modelos em base na distribuição preditiva dos dados. Gelfand e Ghosh (1998) propuseram um critério preditivo com o objetivo de obter as melhores previsões dos dados observados. Ibrahim et al. (2001) apresentam os critérios preditivos, ordem condicional preditiva e a log-verossimilhança pseudo marginal, a fim de comparar modelos paramétricos.

3.1 Distribuição preditiva *a posteriori*

3.1.1 Distribuição preditiva

De forma geral, a distribuição preditiva é a distribuição conjunta dos dados e pode ser usada de vários modos a fim de avaliar a qualidade de ajuste do modelo proposto. Ver Box (1980) e Gelfand et al. (1992) para mais detalhes.

A inferência preditiva tem cálculo simples uma vez que a distribuição *a posteriori* foi obtida. Seja $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$ a densidade conjunta dos dados condicionada ao vetor de parâmetros $\boldsymbol{\theta}$, em que $\mathbf{y} = (y_1, \dots, y_n)^\top$ é o vetor de resposta. Se $\mathbf{X}$ a matriz de variáveis exploratórias de dimensão $n \times k$ com posto completo, e seja $\pi(\boldsymbol{\theta})$ a densidade *a priori* de

θ . A densidade preditiva é dada por

$$f(\mathbf{y}) = \int_{\Theta} f(\mathbf{y}|\theta, \mathbf{X})\pi(\theta)d\theta,$$

sempre que a integral exista. A densidade marginal da i -ésima observação é dada por

$$f(y_i|\mathbf{y}_{(-i)}) = \frac{f(\mathbf{y})}{f(\mathbf{y}_{(-i)})} = \int_{\Theta} f(y_i|\theta, \mathbf{y}_{(-i)}, \mathbf{X})\pi(\theta|\mathbf{y}_{(-i)})d\theta, \quad (3.1)$$

em que $\mathbf{y}_{(-i)}$ é o vetor de dados sem a i -ésima observação e $\pi(\theta|\mathbf{y}_{(-i)})$ corresponde à densidade a *posteriori* de θ sob $\mathbf{y}_{(-i)}$. Em termos de avaliação de modelos, examinar as observações $\mathbf{y}$ com respeito a $f(\mathbf{y})$ é o mesmo que avaliar $\mathbf{y}$ com respeito ao conjunto $f(y_i|\mathbf{y}_{(-i)})$.

Essa abordagem é chamada de validação cruzada, em que $f(y_i|\mathbf{y}_{(-i)})$ é chamada de densidade preditiva da validação cruzada. A validação cruzada é bem estabelecida na literatura Bayesiana, introduzida por Stone (1974).

3.1.2 Distribuição preditiva a *posteriori*

A distribuição preditiva a *posteriori* se refere à predição de um dado não observado condicionado aos dados observados. Aqui a distribuição preditiva a *posteriori* é obtida via a distribuição a *posteriori* dos parâmetros sob todas as observações. Suponha que desejamos prever o valor da variável resposta y_{n+1} . A densidade preditiva a *posteriori* do modelo avaliada neste ponto é dada por

$$f(y_{n+1}|\mathbf{y}) = \int_{\Theta} f(y_{n+1}|\theta, \mathbf{y}, \mathbf{X})\pi(\theta|\mathbf{y})d\theta. \quad (3.2)$$

Usualmente esta integral não é obtida de forma analítica. Contudo, uma aproximação para ela pode ser alcançada ao se utilizar a amostra da distribuição a *posteriori* $\pi(\theta|\mathbf{y})$ obtida pelo método de Monte Carlo via Cadeias de Markov (MCMC), por exemplo, seja $(\theta^{(1)}, \dots, \theta^{(Q)})$ uma amostra aleatória de tamanho Q da distribuição $\pi(\theta|\mathbf{y})$, assim uma aproximação para a expressão (3.2) é dada por

$$f(y_{n+1}|\mathbf{y}) = \frac{1}{Q} \sum_{q=1}^Q f(y_{n+1}|\theta^{(q)}, \mathbf{y}, \mathbf{X}).$$

3.2 Critérios de comparação de modelos

Existem muitas metodologias para comparar e selecionar modelos. Aqui, vamos considerar critérios de comparação que possam ser obtidos via o método MCMC. São eles, ordem condicional preditiva (CPO “Conditional predictive Ordinate”), Gelfand et al. (1992),

o logaritmo da verossimilhança pseudo marginal (LPML “log pseudo marginal likelihood”), Ibrahim et al. (2001), o pseudo fator de Bayes (PsBF “pseudo-Bayes factor”) Geisser e Eddy (1979), e o critério de informação do desvio (DIC “Deviance Information Criterion” Spiegelhalter et al. (2002).

3.2.1 Ordem condicional preditiva (CPO)

A ordem condicional preditiva (CPO) é uma ferramenta popular e útil para a avaliação de modelos. Esta estatística é baseada na distribuição preditiva *a posteriori* da variável resposta para a i -ésima observação dadas às demais observações. Geisser e Eddy (1979) e Gelfand et al. (1992) apresentaram uma discussão detalhada sobre o uso e aplicações do CPO.

Antes de definir a estatística CPO vamos estabelecer algumas quantidades. Seja $\mathbf{D} = (\mathbf{y}, \mathbf{X})$ todos os dados observados, $\mathbf{D}_{(-i)} = (\mathbf{y}_{(-i)}, \mathbf{X}_{(-i)})$ os dados observados sem a i -ésima observação e $\pi(\boldsymbol{\theta}|\mathbf{D}_{(-i)})$ a densidade *a posteriori* de $\boldsymbol{\theta}$ dado $\mathbf{D}_{(-i)}$. Desse modo, o CPO para a i -ésima observação é definido como

$$\begin{aligned} \text{CPO}_i &= f(y_i|\mathbf{D}_{(-i)}) = \frac{f(\mathbf{y})}{f(\mathbf{y}_{(-i)})} = \frac{\int_{\boldsymbol{\theta} \in \Theta} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \Theta} f(\mathbf{y}_{(-i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} = \\ &= \frac{\int_{\boldsymbol{\theta} \in \Theta} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \Theta} \frac{f(y_i|\boldsymbol{\theta})}{f(y_i|\boldsymbol{\theta})} f(\mathbf{y}_{(-i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\int_{\boldsymbol{\theta} \in \Theta} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \Theta} \frac{1}{f(y_i|\boldsymbol{\theta})} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} = \\ &= \left\{ \int_{\boldsymbol{\theta} \in \Theta} \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{f(y_i|\boldsymbol{\theta})} d\boldsymbol{\theta} \right\}^{-1}, \end{aligned} \quad (3.3)$$

em que grandes valores do CPO_i indicam um bom ajuste do modelo. Para os casos em que a integral dada em (3.3) não é analiticamente tratável, Ibrahim et al. (2001) apresentam uma aproximação via o método Monte Carlo. Esta aproximação é feita por amostragem da distribuição *a posteriori* $\pi(\boldsymbol{\theta}|\mathbf{D})$. Seja $(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(Q)})$ a amostra obtida de $\pi(\boldsymbol{\theta}|\mathbf{D})$ via MCMC, a aproximação para o CPO_i é dada por

$$\widehat{\text{CPO}}_i = \left\{ \frac{1}{Q} \sum_{q=1}^Q \frac{1}{f(y_i|\boldsymbol{\theta}^{(q)})} \right\}^{-1}.$$

O gráfico dos valores do $\widehat{\text{CPO}}_i$ contra o índices das observações pode fornecer indícios da qualidade do ajuste do modelo sob determinada observação, isto é, nos fornece indícios da ocorrência de observações atípicas. Como uma forma de sumarizar o CPO sob todas as observações, Ibrahim et al. (2001) apresentam a soma do logaritmo da verossimilhança

pseudo marginal que é dado por

$$\text{LPML} = \sum_{i=1}^n \log(\widehat{CPO}_i).$$

3.2.2 Fator de Bayes

Na abordagem Bayesiana um dos critérios para seleção de modelos é o fator de Bayes, introduzido por Jeffreys (1961). Suponha as hipóteses

$$H_0 : Y \sim M_1, \quad H_1 : Y \sim M_2,$$

em que $M_j, j = 1, 2$ são dois modelos especificados. O fator de Bayes é uma medida da evidência provida pelos dados a favor de uma das hipóteses (modelos) em análise. Sejam $f_j(\mathbf{y}|\boldsymbol{\theta}_j)$ a distribuição amostral dado o vetor de parâmetros e $\pi_j(\boldsymbol{\theta}_j)$ a distribuição *a priori* do vetor de parâmetros, ambos sob o modelo M_j . Define-se por $\mathbb{P}(M_j)$ a probabilidade *a priori* do modelo M_j ser o modelo verdadeiro.

Uma vez observado o conjunto de dados $\mathbf{y}$, o teorema de Bayes é utilizado para obter a probabilidade *a posteriori* de M_j ser o modelo verdadeiro, i. e.,

$$\mathbb{P}(M_j|\mathbf{y}) = \frac{\mathbb{P}_j(\mathbf{y})\mathbb{P}(M_j)}{\mathbb{P}_1(\mathbf{y})\mathbb{P}(M_1) + \mathbb{P}_2(\mathbf{y})\mathbb{P}(M_2)},$$

em que $\mathbb{P}_j(\mathbf{y}) = \int f_j(\mathbf{y}|\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j$ é a distribuição preditiva *a priori* ou marginal de $\mathbf{y}$ para M_j . O fator de Bayes a favor do modelo M_1 é definido como o quociente entre a razão das chances *a posteriori* e a razão de chances *a priori* dos modelos, isto é,

$$BF(M_1, M_2) = \frac{\mathbb{P}(M_1|\mathbf{y})}{\mathbb{P}(M_2|\mathbf{y})} / \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)} = \frac{\mathbb{P}_1(\mathbf{y})}{\mathbb{P}_2(\mathbf{y})} = \frac{\int f_1(\mathbf{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{\int f_2(\mathbf{y}|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2}.$$

Podemos dizer intuitivamente que o melhor modelo será àquele cuja distribuição preditiva *a priori* apresenta o maior valor para $\mathbf{y}$. Quando o fator de Bayes é muito grande ou muito pequeno em relação a 1, significa que temos evidências nos dados a favor de uma das hipóteses. É usual determinar um valor de corte que permita tomar uma decisão a favor de um dos modelos. Jeffreys (1961) sugeriu interpretar o fator de Bayes dividindo os possíveis valores do seu logaritmo em vários intervalos. Os valores recomendados por ele estão presentes na Tabela 3.1.

Polidoro (2013) argumenta que o fator de Bayes pode ser interpretado como a razão das médias das verossimilhanças dos parâmetros para os dois modelos, calculadas relativamente às respectivas distribuições *a priori* do vetor de parâmetros, ou seja, a razão das médias *a priori* das verossimilhanças. Pode-se dizer assim que o fator de Bayes tem alguma semelhança com o teste da razão de verossimilhanças clássico, com a diferença

de que no cálculo do fator de Bayes o vetor de parâmetros é eliminado por integração enquanto que no teste da razão de verossimilhanças o vetor de parâmetros é substituído pelos estimadores de máxima verossimilhança.

Tabela 3.1: Valores de corte do fator de Bayes a favor do modelo M_1 .

$BF(M_1, M_2)$	$\log(BF(M_1, M_2))$	Evidência a favor de M_1
< 1	< 0	negativa
1 a 3,2	0 a 0,5	insignificante
3,2 a 10	0,5 a 1	significante
10 a 100	1 a 2	forte
> 100	> 2	muito forte

Fonte: Jeffreys (1961)

Uma das limitações do fator de Bayes é o fato de depender das distribuições *a priori* do vetor de parâmetros dos dois modelos em comparação, $\pi_j(\boldsymbol{\theta}_j)$. Usualmente, são utilizadas distribuições *a priori* não informativas, o que pode gerar distribuições *a posteriori* impróprias e consequentemente as correspondentes distribuições preditivas *a priori* também podem ser impróprias, o que pode inviabilizar o próprio cálculo do fator de Bayes.

Quando as distribuições preditivas *a priori* existem, mas as integrais que as definem são difíceis de resolver analiticamente, são utilizados métodos de aproximação, tais como o método de Laplace ou o método de Monte Carlo e desta forma o fator de Bayes é aproximado.

Vários autores sugerem modificações para o fator de Bayes. Aitkin (1991) propõe a substituição das distribuições *a priori* $\pi_j(\boldsymbol{\theta}_j)$ por suas correspondentes distribuições *a posteriori* $\pi_j(\boldsymbol{\theta}_j|\mathbf{y})$. O'Hagan (1995) critica o fator de Bayes *a posteriori* devido ao fato de os dados serem utilizados duas vezes, primeiro na determinação da distribuição *a posteriori* e depois no cálculo do fator de Bayes. Esse autor propõe o uso do fator de Bayes parcial, essa abordagem consiste em dividir a amostra em duas partes, $\mathbf{y} = (\mathbf{y}_{(1)}, \mathbf{y}_{(2)})$. Assim um subconjunto dos dados seria designada para o ajuste do modelo e utilizada para atualizar a distribuição *a priori*, ou seja, para obter a distribuição *a posteriori*, e a outra parte do conjunto de dados seria utilizada para calcular o fator de Bayes. Desta forma, o fator de Bayes parcial é dado por

$$BF_{\text{parc}}(\mathbf{y}_{(1)}, \mathbf{y}_{(2)}) = \frac{\mathbb{P}_1(\mathbf{y}_{(1)}|\mathbf{y}_{(2)})}{\mathbb{P}_2(\mathbf{y}_{(2)}|\mathbf{y}_{(1)})} = \frac{\int f_1(\mathbf{y}_{(1)}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1|\mathbf{y}_{(2)})d\boldsymbol{\theta}_1}{\int f_2(\mathbf{y}_{(2)}|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2|\mathbf{y}_{(1)})d\boldsymbol{\theta}_2}.$$

O'Hagan (1995) argumenta que o fator de Bayes parcial é menos sensível a escolha das distribuições *a priori*. Contudo, tem-se o problema de dividir de forma ótima o conjunto de dados em duas partes.

Berger e Pericchi (1996) apresentaram uma possível solução para o problema do fator de Bayes parcial, sugerindo utilizar todas as amostras de treinamento de tamanho mínimo que permita atualizar as distribuições *a priori* e assim poder determinar a média aritmética (ou geométrica) dos fatores de Bayes parciais. Isto é, utilizar todos os subconjuntos de dados de tamanho n_1 , em que n_1 representa o menor tamanho da amostra que conduz a distribuições *a posteriori* próprias, calculando posteriormente a média dos correspondentes fatores de Bayes. Essa média é denominada de fator de Bayes intrínseco (aritmético ou geométrico).

O'Hagan (1997), querendo evitar a escolha arbitrária de uma amostra de treinamento ou ter que considerar todos os subconjuntos de dados de tamanho n_1 , propôs utilizar uma proporção do conjunto de dados definida por $b = n_1/n$, em que n é o tamanho da amostra completa.

Geisser e Eddy (1979) e Gelfand et al. (1992) adotaram uma metodologia de validação cruzada para redefinir o fator de Bayes. Nesta abordagem a amostra é dividida em duas partes, $\mathbf{y} = (\mathbf{y}_{(-i)}, y_i)$, em que $\mathbf{y}_{(-i)}$ é o conjunto de observações com exceção da observação y_i . Se $y_1, \dots, y_n$ forem condicionalmente independentes dado $\boldsymbol{\theta}$, então a distribuição preditiva *a priori* $\mathbb{P}_j(\mathbf{y})$ é denominada de distribuição *pseudo preditiva*, a qual é definida por

$$\prod_{i=1}^n \mathbb{P}_j(y_i | \mathbf{y}_{(-i)}),$$

em que $\mathbb{P}_j(y_i | \mathbf{y}_{(-i)}) = \int_{\Theta} f_j(y_i | \theta_j) \pi_j(\theta_j | \mathbf{y}_{(-i)}) d\theta_j$. O valor de $\mathbb{P}_j(y_i | \mathbf{y}_{(-i)})$ é denominado de *ordenada preditiva condicional* da observação y_i para o modelo M_j e é usualmente representado por $CPO_j(y_i)$. Paulino et al. (2003) comenta que como os valores de $CPO_j(y_i)$ são indicadores da verossimilhança de cada observação dada todas as outras, valores baixos devem corresponder a observações mal ajustadas pelo modelo. Deste modo, Geisser e Eddy (1979) e Gelfand et al. (1992) propuseram o uso do pseudo fator de Bayes, definido como

$$PsBF(M_1, M_2) = \frac{\prod_{i=1}^n \mathbb{P}_1(y_i | \mathbf{y}_{(-i)})}{\prod_{i=1}^n \mathbb{P}_2(y_i | \mathbf{y}_{(-i)})} = \frac{\prod_{i=1}^n CPO_1(y_i)}{\prod_{i=1}^n CPO_2(y_i)}. \quad (3.4)$$

Polidoro (2013) argumentou que a principal motivação para o uso das distribuições preditivas condicionais é que esta é uma distribuição própria, e se $\pi(\boldsymbol{\theta} | y_i)$ também for própria, pode-se fazer uso de distribuições *a priori* não informativas.

3.2.3 Critério de informação do desvio

O critério de informação do desvio (DIC “*Deviance Information Criterion*”) foi proposto por Spiegelhalter et al. (2002). Este critério penaliza o ajuste do modelo pelo acréscimo de parâmetros.

Define-se o desvio Bayesiano de um modelo paramétrico M_j por

$$D(\boldsymbol{\theta}) = -2\log[L(\boldsymbol{\theta}|\mathbf{y})],$$

em que $L(\boldsymbol{\theta}|\mathbf{y})$ é a função de verossimilhança.

Tomando como base esta medida para analisar o ajustamento do modelo, temos que por meio da esperança a *posteriori* do desvio Bayesiano, $\mathbb{E}[D(\boldsymbol{\theta})]$, o número efetivo de parâmetros do modelo (que representa sua complexidade) é dado por

$$p_D = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[D(\boldsymbol{\theta})] - D[\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta})], \quad (3.5)$$

isto é, p_D é dado pela diferença entre a esperança da *posteriori* do desvio Bayesiano e o valor do desvio Bayesiano da esperança da *posteriori* de $\boldsymbol{\theta}$.

Assim, temos que

$$\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[D(\boldsymbol{\theta})] = D[\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta})] + p_D.$$

Segundo Polidoro (2013), a esperança a *posteriori* do desvio Bayesiano pode ser vista como uma medida de adequabilidade do modelo.

Finalmente, Spiegelhalter et al. (2002) definiram o DIC como sendo

$$DIC = 2\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[D(\boldsymbol{\theta})] - D[\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta})]. \quad (3.6)$$

A escolha a favor de um modelo M_j é feita quando este apresenta o menor valor do DIC em relação aos demais. Seja $\bar{D} = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[D(\boldsymbol{\theta})]$ e $D(\bar{\boldsymbol{\theta}}) = D[\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta})]$ temos que

$$DIC = 2\bar{D} - D(\bar{\boldsymbol{\theta}}).$$

Esta medida pode ser obtida via o método MCMC. De fato, se $(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(Q)})$ é uma amostra obtida via o método MCMC, então podemos aproximar $\bar{D}$ e $D(\bar{\boldsymbol{\theta}})$ por

$$\bar{D} = \frac{1}{Q} \sum_{q=1}^Q D(\boldsymbol{\theta}^{(q)}) \quad \text{e} \quad D(\bar{\boldsymbol{\theta}}) = D\left(\frac{1}{Q} \sum_{q=1}^Q \boldsymbol{\theta}^{(q)}\right).$$

Utilizaremos também os critérios (EAIC “Expected Akaike Information Criterion”) e (EBIC “Expected Bayesian Information Criterion”) que foram propostos por Brooks (2002) e utilizados em Bolfarine e Bazan (2010), como exemplo. Estes critérios também podem ser calculados diretamente a partir do algoritmo MCMC, da forma

$$\widehat{EAIC} = \overline{D} + 2p, \widehat{EBIC} = \overline{D} + p \log(n),$$

em que p é o número de parâmetros do modelo e n é o número total de observações.

O conceito de distância entre duas distribuições de probabilidade introduzido por Mahalanobis (1936) está intimamente relacionada com a teoria da informação (TI). Segundo Nascimento et al. (2010), este tipo de problema está relacionado à transmissão de informações entre canais de comunicação contaminados por ruídos estocásticos. Shannon (1948) introduziu as bases da TI e formalizou seus conceitos com aplicações na Teoria da Comunicação e Estatística. As chaves da TI introduzidas por Shannon (1948) recaem sobre dois conceitos: informação e entropia. Tais conceitos têm sido usados em diversas áreas; como, por exemplo, codificação e compressão de dados, telecomunicações e inferência estatística.

Neste capítulo, temos o objetivo de apresentar algumas distâncias estocásticas que serão usadas na detecção de observações atípicas no ajuste dos modelos de regressão beta e beta retangular. Inicialmente discutiremos brevemente sobre alguns conceitos importantes para a construção da definição de distância no sentido estocástico e sua relação com informação, entropia e divergência.

4.1 Entropia

A TI se relaciona intimamente com a teoria das probabilidades, pois o termo informação tem conotação de aleatoriedade, no sentido da probabilidade de ocorrência de eventos.

Seja X uma variável aleatória discreta que assume os valores $\{x_1, \dots, x_n\}$, tal que para $i = 1, \dots, n$, $\mathbb{P}(X = x_i) = p_i$ é a função de probabilidade de X . Segundo Shannon (1948)

a informação de x_i pode ser definida como

$$I(x_i) = \log_{\delta} \left(\frac{1}{\mathbb{P}(X = x_i)} \right) = -\log_{\delta}(p_i),$$

em que a base δ fornece a unidade da medida da informação, por exemplo, se $\delta = 2$ a unidade é o *bit* do inglês “*binary digit*” a menor unidade de informação, se $\delta = e$ a unidade é o *nats* do inglês “*network address translation*” e se $\delta = 10$ o conteúdo da informação é medido em *hartley* a unidade de informação de um dígito decimal. No caso de eventos equiprováveis o uso da função $\log$ pode ser explicado, por exemplo, se o número de símbolos que constituem um alfabeto é M , e se $\delta = 2$, então o número N para representar todos os M símbolos é $M = 2^N$, em que $N = \log_2 M$. Como os eventos são equiprováveis $\mathbb{P}(M = m_i) = 1/M$, então $N = \log_2(1/\mathbb{P}(m_i))$. A medida de informação satisfaz as seguintes propriedades

- a) $I(x_i) = 0$ se $\mathbb{P}(X = x_i) = 1$ ou $\mathbb{P}(X = x_i) = 0$;
- b) $I(x_i) \geq 0$ para $0 \leq \mathbb{P}(X = x_i) \leq 1$;
- c) Se $\mathbb{P}(X = x_i) > \mathbb{P}(X = x_j)$, então $I(x_i) < I(x_j)$;
- d) Se x_i e x_j são independentes, então $I(x_i, x_j) = I(x_i) + I(x_j)$.

A média da informação da variável aleatória X recebe o nome de entropia e é dada por

$$MI(X) = \sum_{i=1}^n \mathbb{P}(X = x_i) I(x_i) = -\sum_{i=1}^n p_i \log_{\delta}(p_i).$$

Para o caso em que X é uma variável aleatória contínua com função de densidade de probabilidade dada por $f_X(x)$, a entropia também chamada de entropia diferencial, é dada por

$$MI(X) = -\int_{\Omega} f_X(x) \log_{\delta}(f_X(x)) dx,$$

em que Ω é o suporte da variável X .

4.2 Distâncias estocásticas

Existe ainda a entropia relativa ou divergência como medida de informação. Essas medidas efetuam um importante papel na teoria estatística, especialmente em teste de hipóteses, tais como aqueles baseados na razão de verossimilhanças e nas estatísticas de score e Wald. A razão disto é que geralmente as divergências estão associadas a distâncias entre distribuições de probabilidade. Baseadas nos conceitos da teoria da informação, muitas medidas de divergência foram propostas desde o trabalho de Kullback

e Leibler (1951), em que os autores apresentam a divergência Kullback-Leibler. Em especial, destacamos a classe ϕ -divergência proposta por Csisz et al. (1967), que engloba muitas das principais divergências. As medidas dessa classe podem ser usadas como medidas de distância entre funções ou distribuições de probabilidade.

Sejam X e Y duas variáveis aleatórias definidas sob o mesmo suporte e com funções de densidade dadas por $f_X(x|\theta_x)$ e $f_Y(x|\theta_y)$, respectivamente. E seja $\phi : (0, \infty) \rightarrow \mathbb{R}_+$ uma função convexa tal que

$$0 \times \phi\left(\frac{0}{0}\right) = 0 \text{ e } 0 \times \phi\left(\frac{x}{0}\right) = \lim_{x \rightarrow \infty} \frac{\phi(x)}{x}.$$

A classe de ϕ -divergências proposta por Csisz et al. (1967) é dada por

$$D_\phi(X, Y) = \sum_{x \in \Omega} \phi\left(\frac{\mathbb{P}(X = x|\theta_x)}{\mathbb{P}(Y = x|\theta_y)}\right) \mathbb{P}(Y = x|\theta_y).$$

para o caso em que X e Y são variáveis aleatórias discretas. Para o caso em que X e Y são variáveis aleatórias contínuas, tem-se

$$D_\phi(X, Y) = \mathbb{E}_Y\left(\phi\left(\frac{f_X(x|\theta_x)}{f_Y(x|\theta_y)}\right)\right) = \int_{\Omega} \phi\left(\frac{f_X(x|\theta_x)}{f_Y(x|\theta_y)}\right) f_Y(x|\theta_y) dx.$$

Liese e Vajda (1987) e Vajda (1989) apresentam as propriedades da classe de medidas ϕ -divergências. Algumas medidas de divergências não pode ser obtidas pela classe ϕ -divergências, como por exemplo a divergência de Rényi (RÉNYI, 1961). Devido a esta limitação Salicrú et al. (1994) introduziram uma extensão da classe de medidas ϕ -divergência, que denominaram de classe (h, ϕ) -divergência a qual é definida para o caso discreto como

$$D_\phi^h(X, Y) = h\left(\sum_x \phi\left(\frac{\mathbb{P}(X = x|\theta_x)}{\mathbb{P}(Y = x|\theta_y)}\right) \mathbb{P}(Y = x|\theta_y)\right)$$

e

$$D_\phi^h(X, Y) = h\left(\mathbb{E}_Y\left(\phi\left(\frac{f_X(x|\theta_x)}{f_Y(x|\theta_y)}\right)\right)\right) = h\left(\int_{\Omega} \phi\left(\frac{f_X(x|\theta_x)}{f_Y(x|\theta_y)}\right) f_Y(x|\theta_y) dx\right),$$

para o caso contínuo. Em que $h : (0, \infty) \rightarrow [0, \infty)$ é uma função estritamente crescente. A escolha adequada das funções h e ϕ nos leva a medidas de divergências bem conhecidas, tais como a divergência χ^2 (TANEJA, 2006), e a divergência Kullback-Leibler.

Burbea e Rao (1982) comentam que as medidas de divergências apesar de serem ade-

quadas para comparar distribuições de probabilidade, frequentemente não satisfazem a desigualdade triangular, sendo assim, não podem ser consideradas distâncias. Como referência ao trabalho de Nascimento et al. (2010), trataremos como divergência alguma função não negativa entre duas medidas de probabilidade que satisfaz a propriedade de definitividade apresentada mais a diante. Chamaremos de distâncias às divergências que são simétricas.

Caso a divergência não seja simétrica, como o caso das divergências χ^2 e Kullback-Leibler, Nascimento et al. (2010) propõe usar a simetrização

$$d_\phi^h(X, Y) = \frac{D_\phi^h(X, Y) + D_\phi^h(Y, X)}{2}. \quad (4.1)$$

Definição 1 (Distância estocástica). *Uma função $d_\phi^h : U \times U \rightarrow \mathbb{R}$ é uma medida de distância se para todo $X, Y \in U \subset \mathcal{A}$, as seguintes propriedades se verificam*

1. Não negatividade

$$d_\phi^h(X, Y) \geq 0.$$

2. Simetria

$$d_\phi^h(X, Y) = d_\phi^h(Y, X).$$

3. Definitividade

$$d_\phi^h(X, Y) = 0 \Leftrightarrow X = Y.$$

Em que X e Y são variáveis aleatórias, $\mathcal{A}$ é uma σ -álgebra do espaço de probabilidade de X e Y , e U é um subconjunto de $\mathcal{A}$. Nascimento et al. (2010) comentam que diferentemente da definição de distância usual, as medidas de distâncias estocásticas não necessariamente satisfazem a desigualdade triangular. Na Tabela 4.1 temos as funções adequadas h e ϕ para obtermos algumas distâncias estocásticas.

Tabela 4.1: Funções h e ϕ e suas respectivas distâncias.

(h, ϕ) -distância	$h(y)$	$\phi(x)$
Kullback-Leibler	$y/2$	$(x-1) \log(x)$
χ^2	$y/4$	$\frac{(x-1)^2(x+1)}{2}$
Bhattacharyya	$-\log(1-y), 0 \leq y < 1$	$-\sqrt{x} + \frac{x+1}{2}$
Hellinger	$y/2, 0 \leq y < 2$	$(\sqrt{x}-1)^2$
Triangular	$y, 0 \leq y < 2$	$\frac{(x-1)^2}{x+1}$
Média-harmônica	$-\log(\frac{-y}{2} + 1), 0 \leq y < 2$	$\frac{(x-1)^2}{x+1}$

A seguir apresentamos as expressões das distâncias estocásticas obtidas utilizando as funções apresentadas na Tabela 4.1, e que serão utilizadas em nosso trabalho.

1. Distância estocástica χ^2

$$d_{\chi^2}(X, Y) = \frac{1}{2} \left[\int \frac{(f_X(x|\theta_x) - f_Y(x|\theta_y))^2}{2f_X(x|\theta_x)} dx + \int \frac{(f_X(x|\theta_x) - f_Y(x|\theta_y))^2}{2f_Y(x|\theta_y)} dx \right].$$

A divergência χ^2 é bastante utilizada no contexto de categorização de dados, principalmente em testes de hipóteses (BRONIATOWSKI e LEORATO, 2011).

2. Distância estocástica Kullback-Leibler

$$\begin{aligned} d_{KL}(X, Y) &= \frac{1}{2} \left[\int f_X(x|\theta_x) \log \left(\frac{f_X(x|\theta_x)}{f_Y(x|\theta_y)} \right) dx + \int f_Y(x|\theta_y) \log \left(\frac{f_Y(x|\theta_y)}{f_X(x|\theta_x)} \right) dx \right] = \\ &= \frac{1}{2} \int (f_X(x|\theta_x) - f_Y(x|\theta_y)) \log \left(\frac{f_X(x|\theta_x)}{f_Y(x|\theta_y)} \right) dx. \end{aligned}$$

Eguchi e Copas (2006) apresentam outra interpretação estatística da divergência Kullback-Leibler em termos de perda de poder do teste da razão de verossimilhança quando a distribuição errada é usado para uma das hipóteses. Elas fazem isto exploram o fato de que a divergência Kullback-Leibler é a esperança da razão de verossimilhanças, e que o lema de Neyman-Pearson é a razão de verossimilhanças. Seghouane e Amari (2007) utilizam esta distância como uma forma de corrigir o critério de informação de Akaike.

3. Distância estocástica Bhattacharyya

$$d_B(X, Y) = -\log \int \sqrt{f_X(x|\theta_x) f_Y(x|\theta_y)}.$$

Goudail e Réfrégier (2004) argumentam que essa distância é mais precisa do que a distância Kullback-Leibler em algoritmos para processamento de imagens PolSAR.

4. Distância estocástica Hellinger

$$d_H(X, Y) = 1 - \int \sqrt{f_X(x|\theta_x) f_Y(x|\theta_y)} dx.$$

Giet e Lubrano (2008) fizeram uso de métodos de estimação baseados na minimização desta distância no contexto de equações diferenciais.

5. Distância estocástica triangular

$$d_T(X, Y) = \int \frac{(f_X(x|\theta_x) - f_Y(x|\theta_y))^2}{f_X(x|\theta_x) + f_Y(x|\theta_y)} dx.$$

6. Distância estocástica média-harmônica

$$d_{MH}(X, Y) = -\log \left(\int \frac{2f_X(x|\theta_x)f_Y(x|\theta_y)}{f_X(x|\theta_x) + f_Y(x|\theta_y)} dx \right) = -\log \left(1 - \frac{d_T(X, Y)}{2} \right).$$

A distância de Hellinger é a única limitada no intervalo $[0, 1]$, as demais distâncias variam no intervalo $(0, \infty)$. Quando as distribuições pertencem à mesma família, apenas seus parâmetros são relevantes. Assim os parâmetros se tornam os argumentos das medidas de divergência e de distância. Deste modo os símbolos para variáveis aleatórias como argumentos das medidas de divergências e de distâncias são substituídos pelos parâmetros das distribuições, por exemplo, $d_{KL}(X, Y) = d_{KL}(\theta_x, \theta_y)$.

No contexto de análise de regressão estatística, mais especificamente, na análise de diagnóstico, temos que a primeira técnica desenvolvida para avaliar o impacto individual de uma observação sobre o processo de estimação dos parâmetros de um modelo, se baseia na retirada individual de cada observação, e posteriormente os efeitos destas retiradas são estudados (COOK e WEISBERG, 1982). No contexto Bayesiano, Cho et al. (2009) desenvolveram uma análise de diagnóstico de influência para as distribuições *a posteriori* conjunta e marginal dos parâmetros do modelo em estudo. Tal análise envolve a retirada individual de observações e se baseia na divergência Kullback-Leibler. Cho et al. (2009) apresentam uma expressão para calcular a divergência Kullback-Leibler entre as distribuições *a posteriori* com todas as observações e sem uma observação individual.

Seguindo Cho et al. (2009), em nosso trabalho realizaremos uma análise de diagnóstico, que também envolve a retirada individual de observações, baseada nas distâncias estocásticas apresentadas na Tabela 4.1, para os modelos de regressão beta e beta retangular. Para tal fim, reescrevemos a classe (h, ϕ) -divergência da forma

$$D_\phi^h(P, P_{(-i)}) = h \left(\int_{\Theta} \phi \left(\frac{\pi(\boldsymbol{\theta}|\mathbf{D}_{(-i)})}{\pi(\boldsymbol{\theta}|\mathbf{D})} \right) \pi(\boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta} \right) = h \left(\mathbb{E}_{\boldsymbol{\theta}|\mathbf{D}} \left(\phi \left(\frac{\pi(\boldsymbol{\theta}|\mathbf{D}_{(-i)})}{\pi(\boldsymbol{\theta}|\mathbf{D})} \right) \right) \right),$$

em que P denota a densidade *a posteriori* de $\boldsymbol{\theta}$ dado $\mathbf{D}$ e $P_{(-i)}$ denota a densidade *a posteriori* de $\boldsymbol{\theta}$ dado $\mathbf{D}_{(-i)}$.

Proposição 1. *A classe (h, ϕ) -divergência pode ainda ser escrita da forma*

$$D_\phi^h(P, P_{(-i)}) = h \left(\mathbb{E}_{\boldsymbol{\theta}|\mathbf{D}} \left(\phi \left(\frac{CPO_i}{f(y_i|\boldsymbol{\theta})} \right) \right) \right). \quad (4.2)$$

Prova 1. *Pelo teorema de Bayes temos que*

$$\begin{aligned} \frac{\pi(\boldsymbol{\theta}|\mathbf{D}_{(-i)})}{\pi(\boldsymbol{\theta}|\mathbf{D})} &= \frac{\pi(\boldsymbol{\theta}) \prod_{j \in \mathbf{D}_{(-i)}} f(y_j|\boldsymbol{\theta})}{\int_{\Theta} \pi(\boldsymbol{\theta}) \prod_{j \in \mathbf{D}_{(-i)}} f(y_j|\boldsymbol{\theta}) d\boldsymbol{\theta}} \times \frac{\int_{\Theta} \pi(\boldsymbol{\theta}) \prod_{j \in \mathbf{D}} f(y_j|\boldsymbol{\theta}) d\boldsymbol{\theta}}{\pi(\boldsymbol{\theta}) \prod_{j \in \mathbf{D}} f(y_j|\boldsymbol{\theta})} = \\ &= \frac{1}{f(y_i|\boldsymbol{\theta})} \times \frac{\int_{\Theta} \pi(\boldsymbol{\theta}) \prod_{j \in \mathbf{D}} f(y_j|\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} \frac{1}{f(y_i|\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) \prod_{j \in \mathbf{D}} f(y_j|\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{\left(\int_{\Theta} \frac{\pi(\boldsymbol{\theta}|\mathbf{D})}{f(y_i|\boldsymbol{\theta})} d\boldsymbol{\theta} \right)^{-1}}{f(y_i|\boldsymbol{\theta})} = \frac{CPO_i}{f(y_i|\boldsymbol{\theta})}, \end{aligned}$$

em que $\pi(\boldsymbol{\theta})$ é a densidade a priori de $\boldsymbol{\theta}$, $\prod_{j \in \mathbf{D}} f(y_j|\boldsymbol{\theta})$ é a função de verossimilhança para $\mathbf{D}$, $\prod_{j \in \mathbf{D}_{(-i)}} f(y_j|\boldsymbol{\theta})$ é a função de verossimilhança para $\mathbf{D}_{(-i)}$ e CPO_i é como dado em 3.3.

Uma aproximação de 4.2 pode ser obtida por amostragem de $\pi(\boldsymbol{\theta}|\mathbf{D})$ via MCMC. Seja $(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(Q)})$ a amostra obtida de $\pi(\boldsymbol{\theta}|\mathbf{D})$ a aproximação de $D_{\phi}^h(P, P_{(-i)})$ é dada por

$$\widehat{D}_{\phi}^h(P, P_{(-i)}) = h \left(\frac{1}{Q} \sum_{q=1}^Q \phi \left(\frac{\widehat{CPO}_i}{f(y_i|\boldsymbol{\theta}^{(q)})} \right) \right). \quad (4.3)$$

Segue as aproximações segundo 4.3 para as distâncias apresentadas na Tabela 4.1

1. Distância estocástica χ^2

$$\widehat{d}_{\chi^2}(P, P_{(-i)}) = \frac{1}{4Q} \sum_{q=1}^Q \frac{(\widehat{CPO}_i - f(y_i|\boldsymbol{\theta}^{(q)}))^2 (\widehat{CPO}_i + f(y_i|\boldsymbol{\theta}^{(q)}))}{f(y_i|\boldsymbol{\theta}^{(q)})^2 \widehat{CPO}_i}.$$

2. Distância estocástica Kullback-Leibler

$$\widehat{d}_{KL}(P, P_{(-i)}) = \frac{1}{2Q} \sum_{q=1}^Q \left(\frac{\widehat{CPO}_i - f(y_i|\boldsymbol{\theta}^{(q)})}{f(y_i|\boldsymbol{\theta}^{(q)})} \right) \log \left(\frac{\widehat{CPO}_i}{f(y_i|\boldsymbol{\theta}^{(q)})} \right).$$

3. Distância estocástica Bhattacharyya

$$\widehat{d}_B(P, P_{(-i)}) = -\log \left(1 - \left[\frac{1}{Q} \sum_{q=1}^Q -\sqrt{\frac{\widehat{CPO}_i}{f(y_i|\boldsymbol{\theta}^{(q)})}} + \frac{\widehat{CPO}_i + f(y_i|\boldsymbol{\theta}^{(q)})}{2f(y_i|\boldsymbol{\theta}^{(q)})} \right] \right).$$

4. Distância estocástica Hellinger

$$\widehat{d}_H(P, P_{(-i)}) = \frac{1}{2Q} \sum_{q=1}^Q \left(\sqrt{\frac{\widehat{CPO}_i}{f(y_i|\boldsymbol{\theta}^{(q)})}} - 1 \right)^2.$$

5. Distância estocástica triangular

$$\hat{d}_T(P, P_{(-i)}) = \frac{1}{Q} \sum_{q=1}^Q \frac{(\widehat{CPO}_i - f(y_i | \boldsymbol{\theta}^{(q)}))^2}{f(y_i | \boldsymbol{\theta}^{(q)}) (\widehat{CPO}_i + f(y_i | \boldsymbol{\theta}^{(q)}))}.$$

6. Distância estocástica média-harmônica

$$\hat{d}_{MH}(P, P_{(-i)}) = -\log \left(1 + \frac{\hat{d}_T(P, P_{(-i)})}{2} \right).$$

Estas são as formas das distâncias estocásticas que utilizaremos em nosso trabalho, tanto para o estudo de simulação como para a aplicação. Além destas medidas também utilizaremos a divergência Kullback-Leibler que possui função ϕ e h dadas por $\phi(x) = -\log(x)$ e $h(y) = y$ e a divergência χ^2 que tem funções dadas por $\phi(x) = (x-1)^2$ e $h(y) = y$, ver Dey e Birmiwal (1994).

CAPÍTULO 5

Simulação

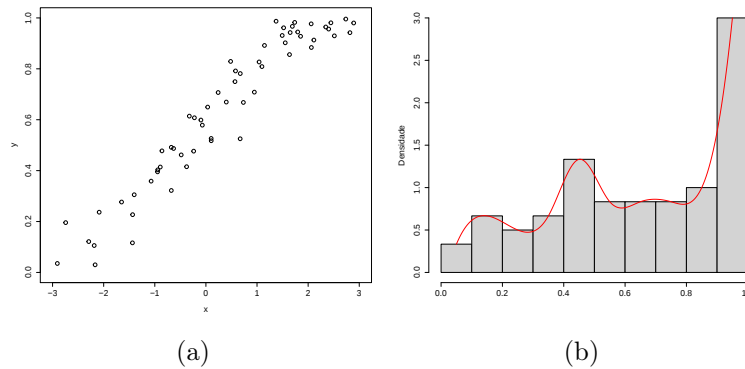
Nesse capítulo apresentamos o estudo de simulação feito com o intuito de avaliar a sensibilidade de algumas medidas de divergências e distâncias estocásticas na detecção de observações atípicas no ajuste dos modelos de regressão beta e beta retangular. Para a geração dos dados fizemos uso do software livre **R** na sua versão 3.2.2 e para os ajustes dos modelos sob a perspectiva Bayesiana, além de utilizarmos a função `jags.model` do pacote `rjags` do mesmo software, também fizemos uso do software **JAGS** na sua versão 3.4.0.

Nesse estudo fizemos uso de apenas uma covariável, diga-se X , tal que $X \sim U(-3, 3)$, sendo esta levada em consideração na geração dos conjuntos de dados das distribuições beta e beta retangular. Primeiramente geramos conjuntos de dados do modelo de regressão beta, tal que, $y_i \sim \mathcal{B}(\mu_i, \phi)$, $\text{logito}(\mu_i) = \beta_0 + \beta_1 \times x_i$, $i = 1, 2, \dots, n$, com parâmetros $\beta_0 = 0.5$, $\beta_1 = 1$ e $\phi = 15$. Posteriormente geramos conjuntos de dados do modelo de regressão beta retangular, tal que, $y_i \sim \mathcal{BRR}(\gamma_i, \phi, \alpha)$, $\text{logito}(\gamma_i) = \beta_0 + \beta_1 \times x_i$, $i = 1, 2, \dots, n$, com parâmetros $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha = 0.2$ e $\phi = 15$. Em nossa análise, consideramos quatro tamanhos de amostras $n = \{30, 60, 90, 120\}$. Para as distribuições das *prioris* dos parâmetros do modelo de regressão beta, assumimos $\beta_0 \sim \mathcal{N}(0.5, 0.01^2)$, $\beta_1 \sim \mathcal{N}(1, 0.01^2)$, e para o modelo de regressão beta retangular, temos que, $\beta_0 \sim \mathcal{N}(0.5, 0.01^2)$, $\beta_1 \sim \mathcal{N}(1, 0.01^2)$ e $\alpha \sim \mathcal{U}(0, 1)$. O percentual de observações atípicas é $p = \{3\%, 6\%, 10\%\}$. A contaminação foi feita tanto na variável resposta como na variável explicativa. Na variável resposta substituímos os valores dos dados y_i por $y_i^* = y_i \pm \Delta$, em que Δ é o incremento ou decremento dos valores de y , de tal forma que

- a) Um incremento de Δ unidades nos valores de y correspondentes aos valores mais baixos de x ;
- b) Um decremento de Δ unidades nos valores de y correspondentes aos valores mais altos de x ;
- c) Um incremento e decremento de Δ unidades nos valores de y correspondentes aos valores mais altos e mais baixos de x , respectivamente.

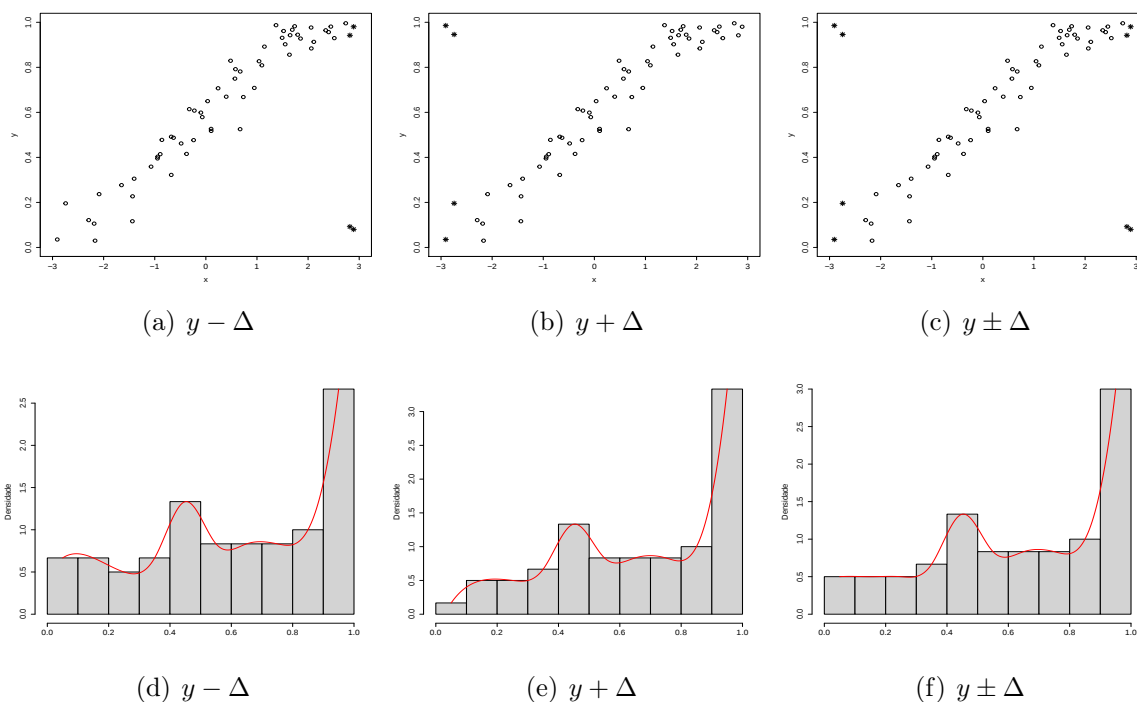
Nas Figuras que seguem, tem-se ilustrações, para tamanho de amostra igual a 60, dos diagramas de dispersão com os respectivos histogramas da variável resposta contaminada. Primeiramente, tem-se da Figura 5.1 à Figura 5.4, os esquemas referentes a geração da variável resposta a partir da distribuição beta, e das Figuras 5.5 à 5.6, os esquemas da geração da variável resposta a partir da distribuição beta retangular. Observamos em ambos os esquemas a contaminação nas caudas da distribuição da variável resposta. Também notamos que a medida que o nível de contaminação aumenta, os valores contaminados se deslocam da cauda para o meio dos histogramas, isto ocorre porque o nível de contaminação se refere ao total do tamanho da amostra, e não apenas da cauda da distribuição.

Figura 5.1: Diagrama de dispersão e histograma da variável resposta, gerada da distribuição beta, sem contaminação.



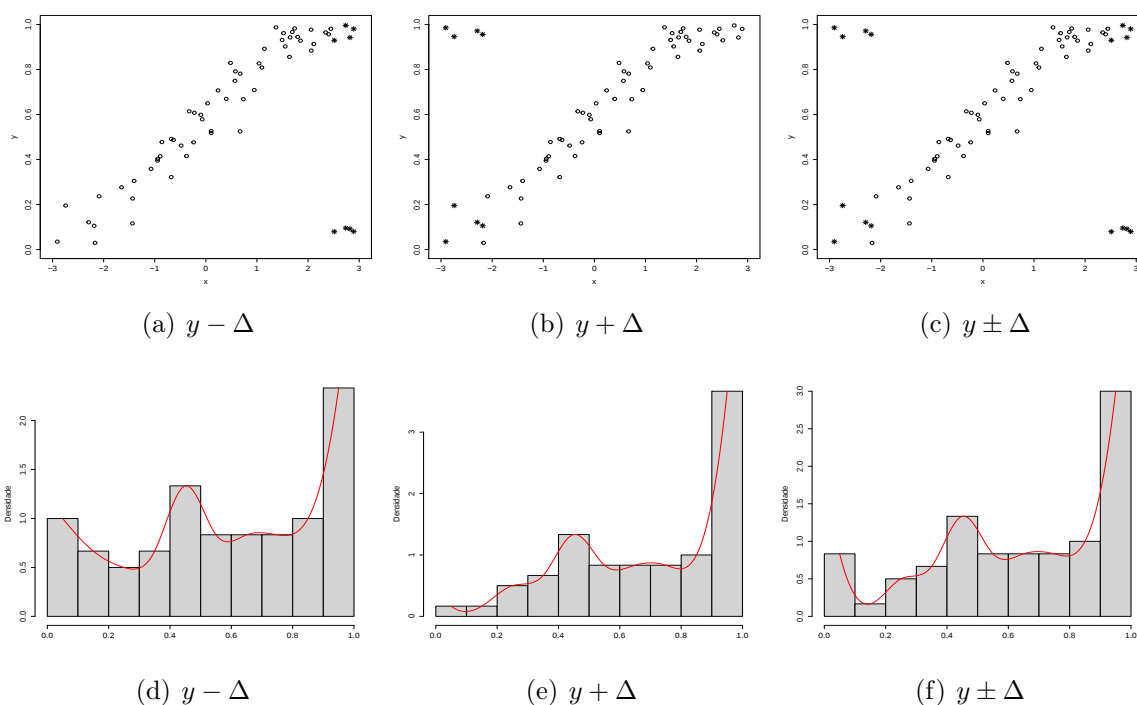
Fonte: Dados próprios.

Figura 5.2: Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta, com 3% de contaminação.



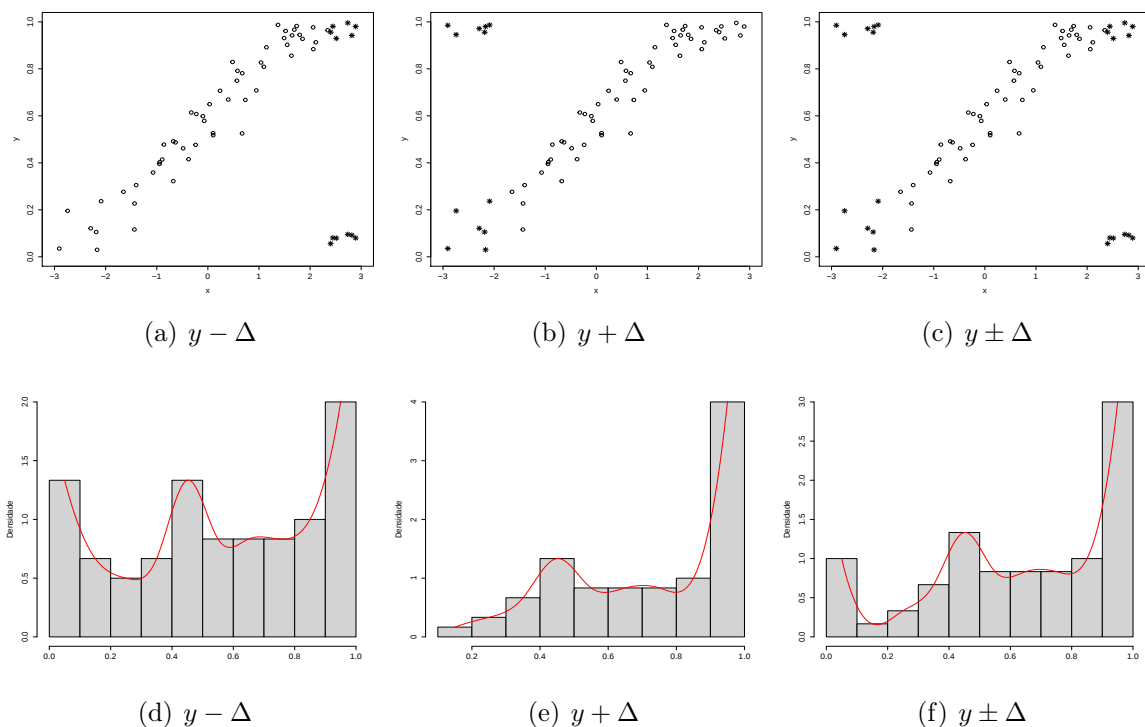
Fonte: Dados próprios.

Figura 5.3: Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta, com 6% de contaminação.



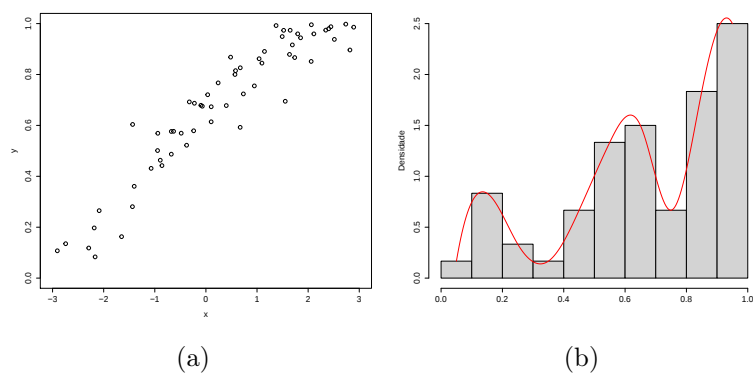
Fonte: Dados próprios.

Figura 5.4: Diagramas de dispersão e histogramas da variável resposta, gerada da distribuição beta, com 10% de contaminação.



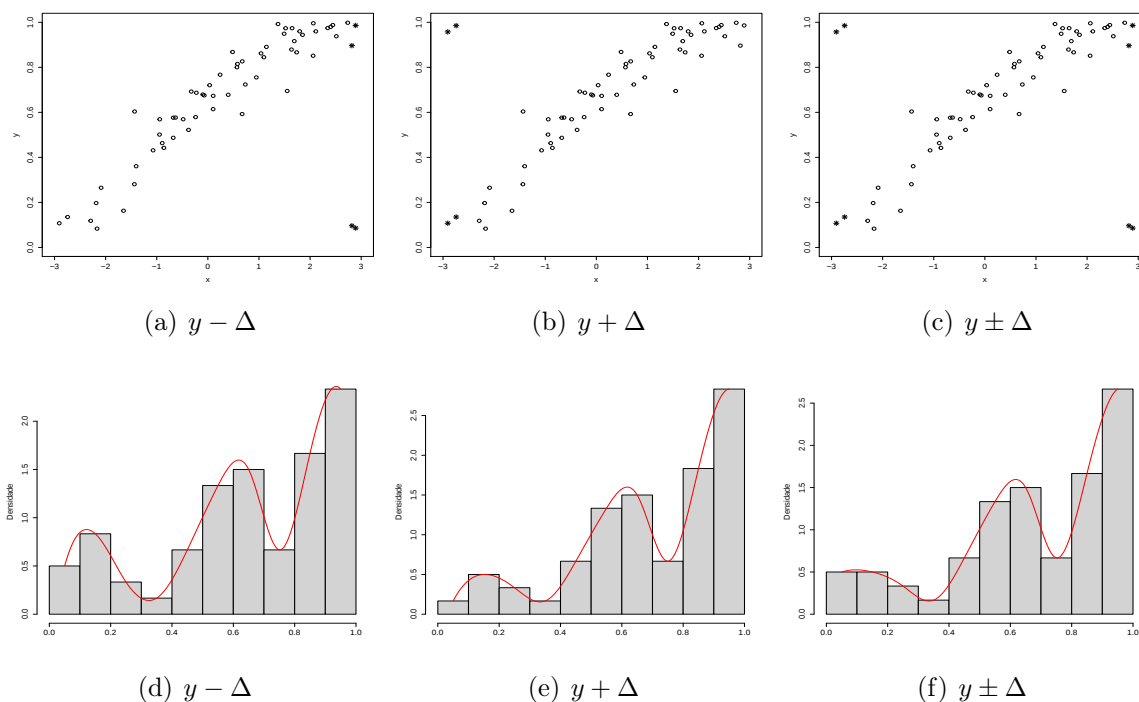
Fonte: Dados próprios.

Figura 5.5: Diagrama de dispersão e histograma da variável resposta, gerada da distribuição beta retangular, sem contaminação.



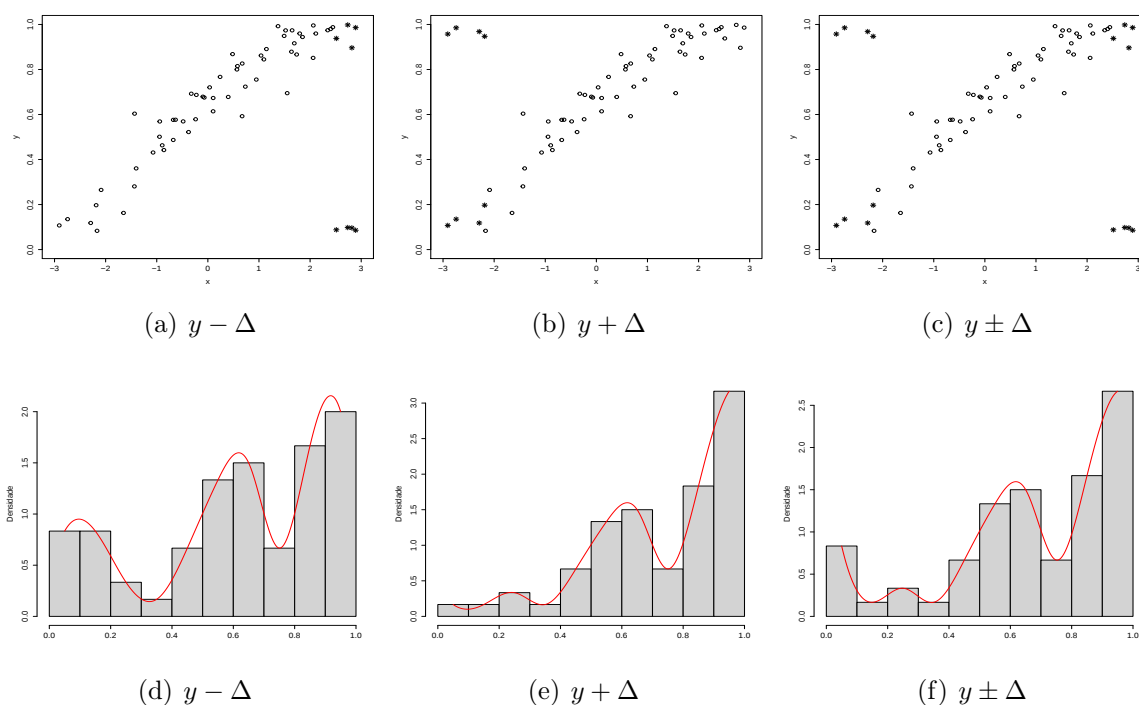
Fonte: Dados próprios.

Figura 5.6: Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta retangular, com 3% de contaminação.



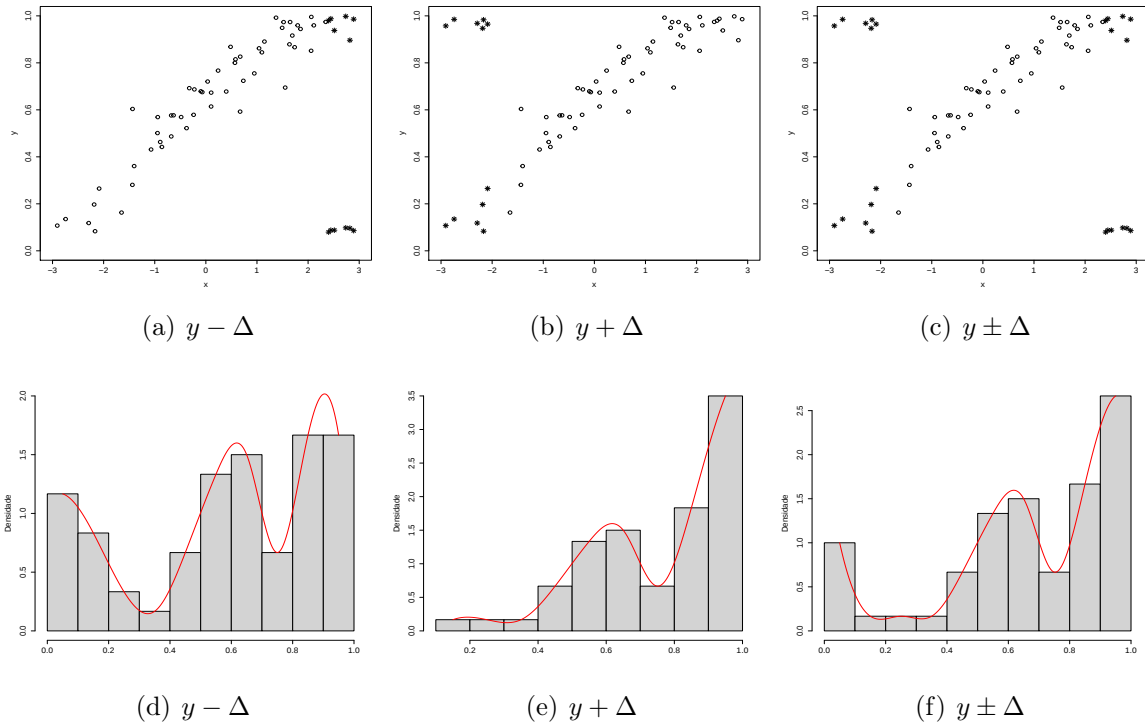
Fonte: Dados próprios.

Figura 5.7: Diagramas des dispersão e histogramas da variável resposta, gerada da distribuição beta retangular, com 6% de contaminação.



Fonte: Dados próprios.

Figura 5.8: Diagramas despersão e histogramas da variável resposta, gerada da distribuição beta retangular, com 10% de contaminação.



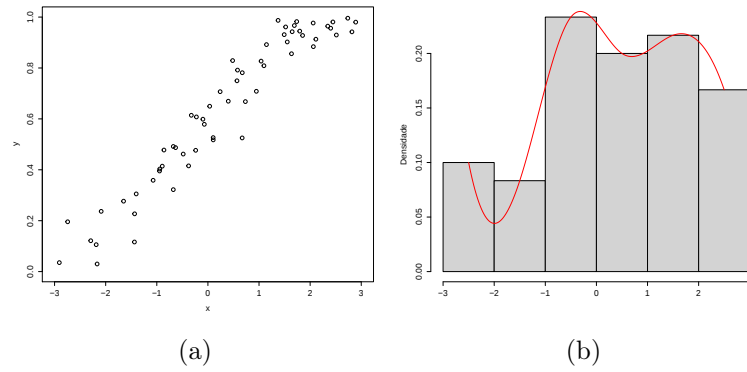
Fonte: Dados próprios.

Posteriormente, também substituímos os valores dos dados x_i por $x_i^* = x_i \pm \Delta$, de tal modo que

- Um incremento de Δ unidades nos valores de x correspondentes aos valores mais baixos de y ;
- Um decremento de Δ unidades nos valores de x correspondentes aos valores mais altos de y ;
- Um incremento e decremento de Δ unidades nos valores de x correspondentes aos valores mais baixos e altos de y , respectivamente.

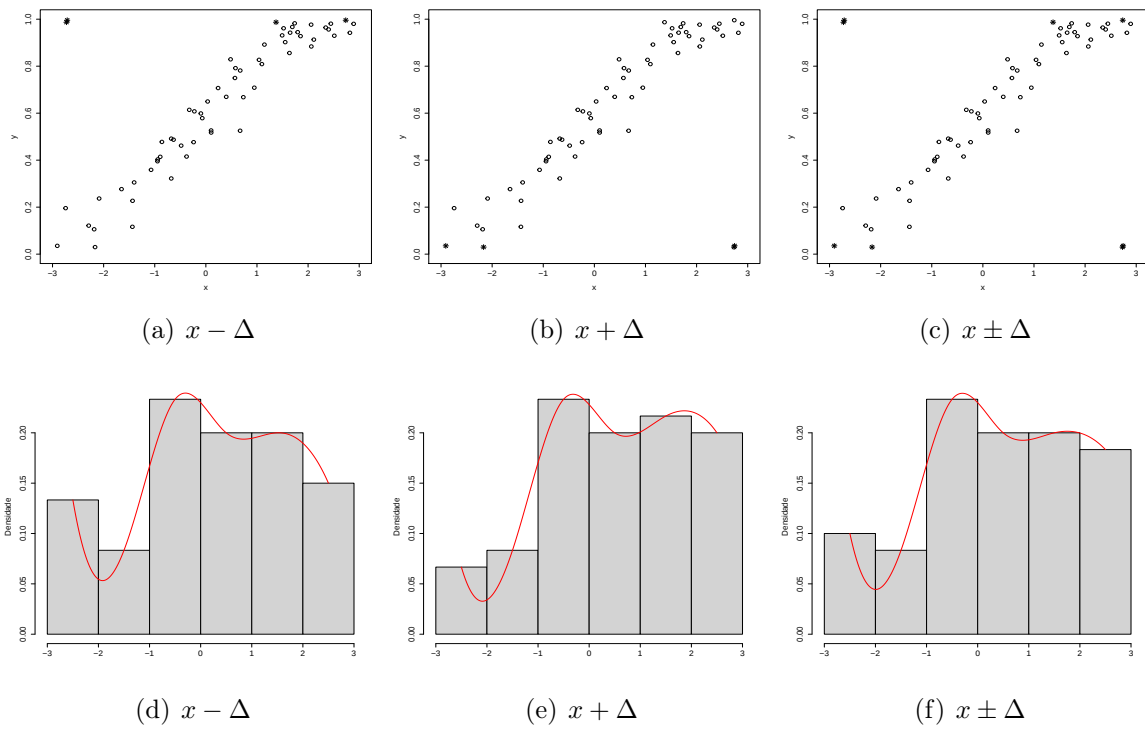
Primeiramente, tem-se da Figura 5.9 à Figura 5.12, os esquemas referentes a geração da variável resposta a partir da distribuição beta, contaminando a variável explicativa, e das Figuras 5.13 à 5.16, os esquemas da geração da variável resposta a partir da distribuição beta retangular, contaminando a variável explicativa. Notamos, novamente, em ambos os esquemas, que a medida que o nível de contaminação aumenta, os valores contaminados se deslocam da cauda para o meio dos histogramas.

Figura 5.9: Diagrama de dispersão e histograma da variável explicativa, sendo a variável resposta gerada da distribuição beta, sem contaminação.



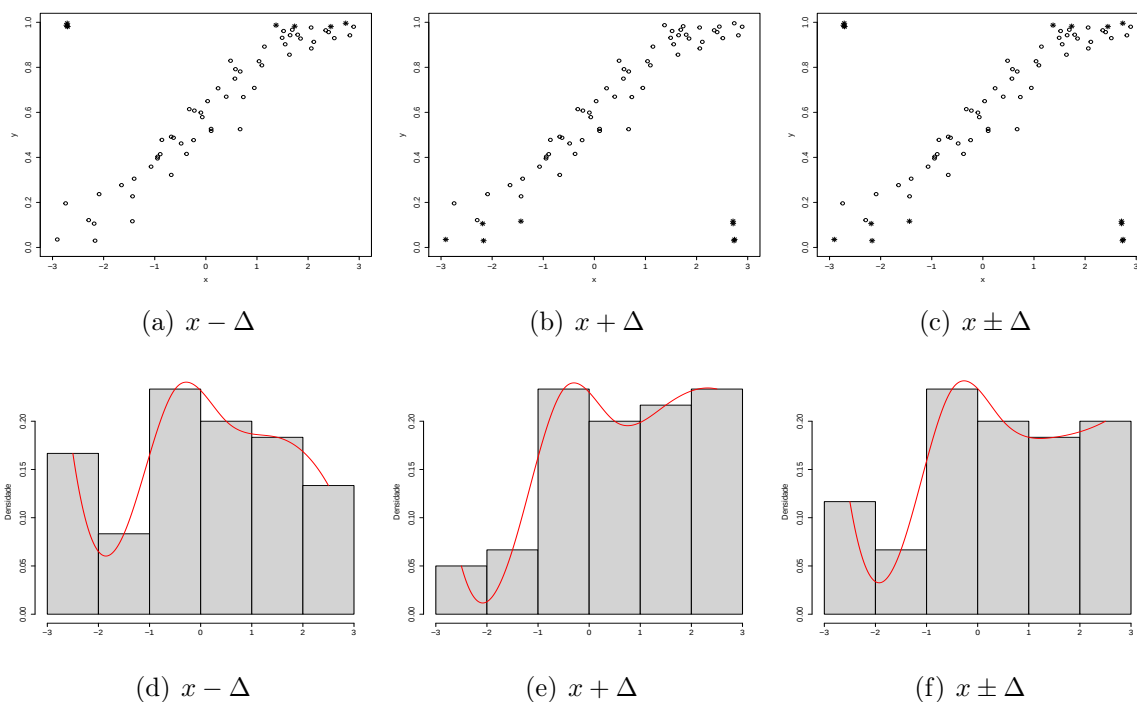
Fonte: Dados próprios.

Figura 5.10: Diagramas de dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta, com 3% de contaminação.



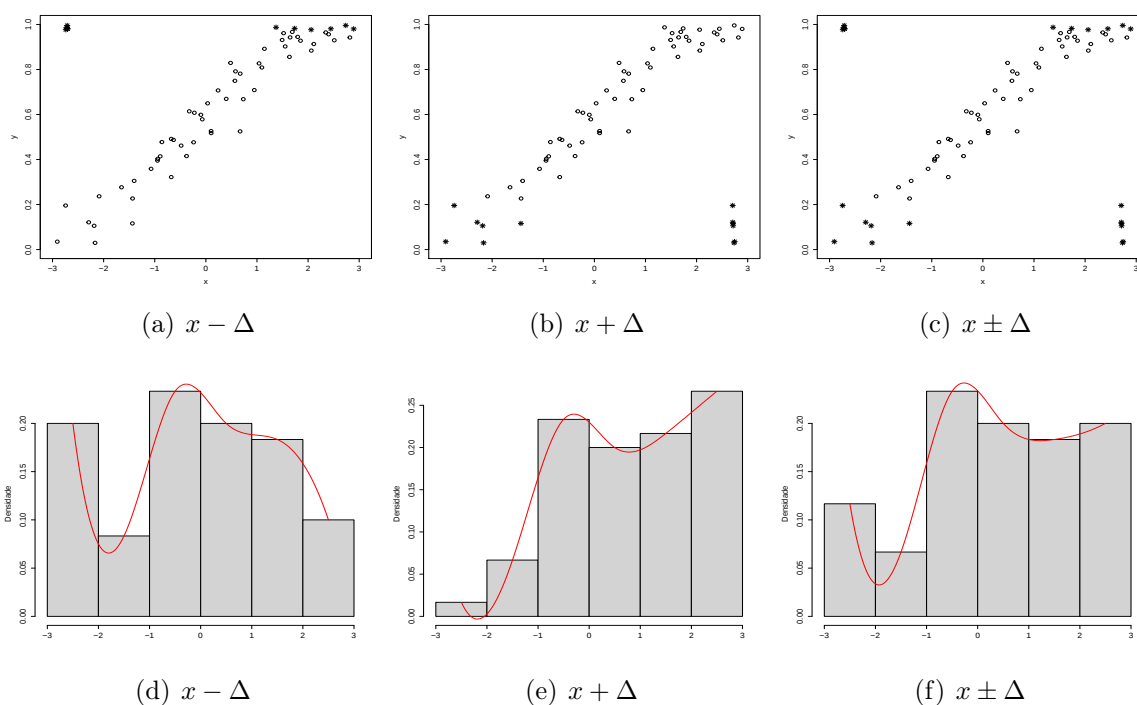
Fonte: Dados próprios.

Figura 5.11: Diagramas de dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta, com 6% de contaminação.



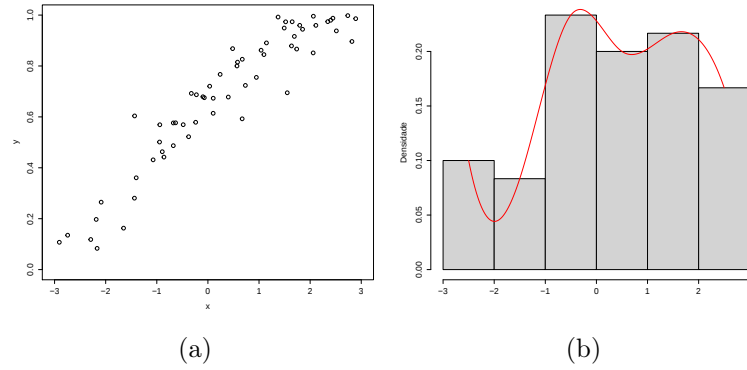
Fonte: Dados próprios.

Figura 5.12: Diagramas de dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta, com 10% de contaminação.



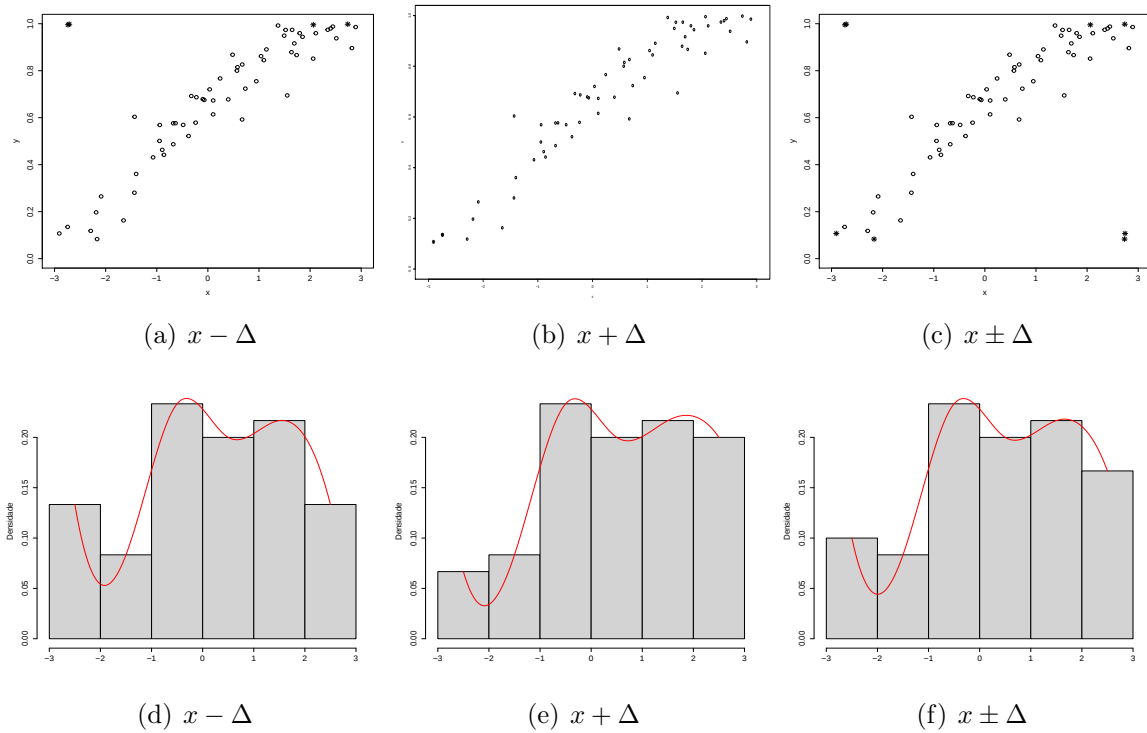
Fonte: Dados próprios.

Figura 5.13: Diagrama de dispersão e histograma da variável explicativa, sendo a variável resposta gerada da distribuição beta retangular, sem contaminação.



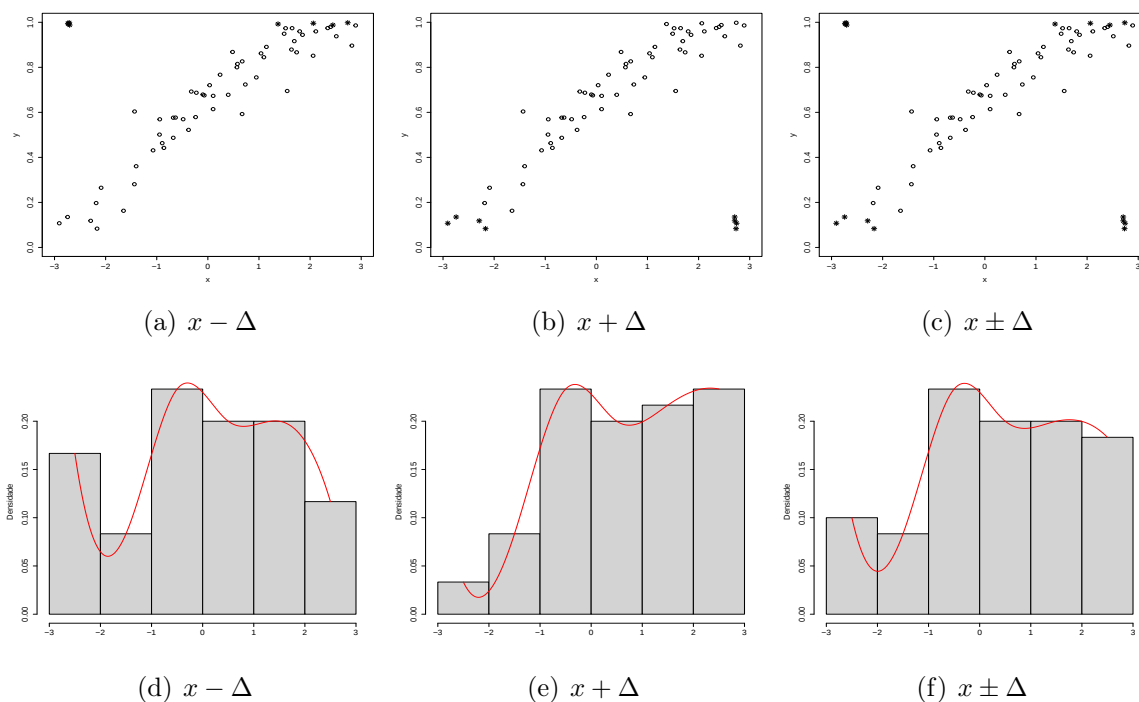
Fonte: Dados próprios.

Figura 5.14: Diagramas de dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta retangular, com 3% de contaminação.



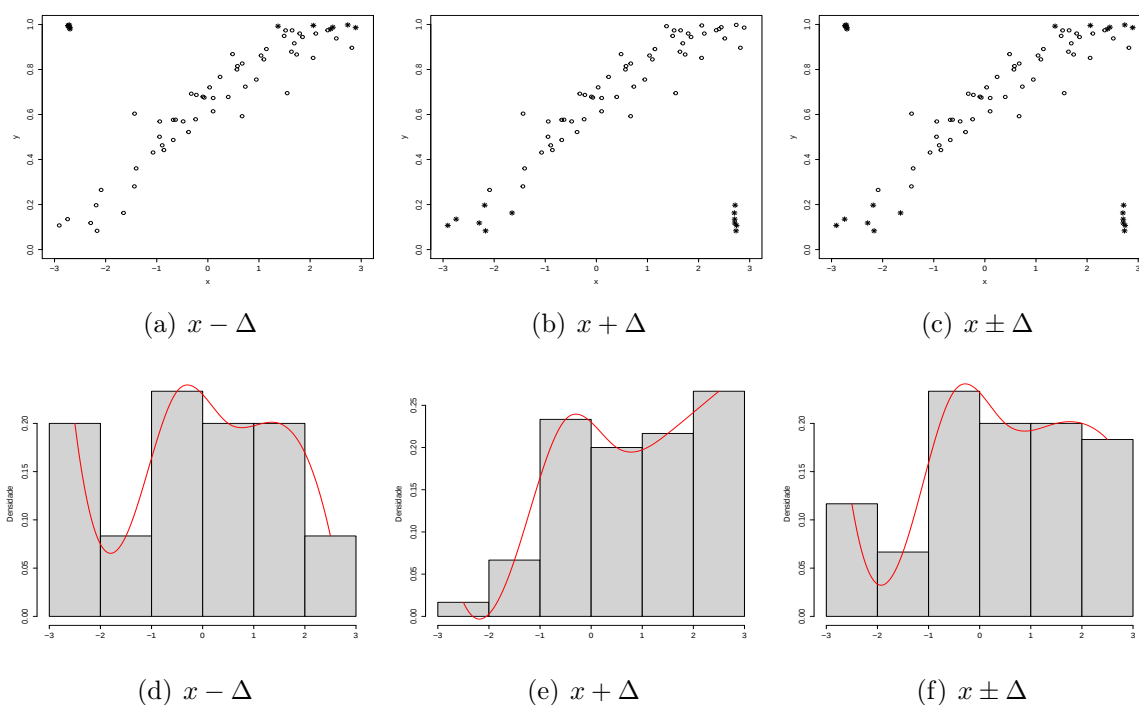
Fonte: Dados próprios.

Figura 5.15: Diagramas des dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta retangular, com 6% de contaminação.



Fonte: Dados próprios.

Figura 5.16: Diagramas des dispersão e histogramas da variável explicativa, sendo a variável resposta gerada da distribuição beta retangular, com 10% de contaminação.



Fonte: Dados próprios.

Para realizamos a contaminação dos dados assumimos para Δ o valor de 0.05, este incremento ou decremento é feito sucessivamente nos valores de y ou x , conforme o caso, até obter um valor limite de acordo com cada variável. Por exemplo, se $y = 0.98$, após aplicamos o decremento Δ sucessivamente, teremos que o valor limite para y de 0.03. No estudo em questão, alternamos tanto a distribuição de geração dos dados como também o modelo de regressão ajustado e a variável a ser contaminada, obtendo assim 8 cenários. Na Tabela 5.1 temos o esquema descrito para a obtenção dos 8 cenários. Dentro de cada cenário temos 12 combinações entre os níveis de contaminação e tamanhos de amostra. Para cada combinação utilizamos 1000 réplicas de Monte Carlo. Para os modelos ajustados usamos o método MCMC para obtermos uma cadeia de tamanho 20.000, utilizamos um *burn-in* igual a 10.000 e um *thin* igual a 10, assim obtivemos uma amostra final de tamanho 1.000 das distribuições *a posteriori* dos parâmetros dos modelos ajustados.

Tabela 5.1: Esquema de geração dos 8 cenários da simulação.

Distribuição de geração dos dados	Modelo de regressão ajustado	Variável contaminada	Cenário
Beta retangular	Beta retangular	Y	1
		X	2
	Beta	Y	3
		X	4
Beta	Beta retangular	Y	5
		X	6
	Beta	Y	7
		X	8

Para a detecção das observações atípicas, além de usarmos as distâncias estocásticas apresentadas na Tabela 4.1, também fazemos uso de duas divergências, são elas, divergência Kullback-Leibler e divergência χ^2 . Também utilizamos a distância L_1 norm Weissman et al. (2003) que tem função $\phi(x)$ igual a $0, 5|x - 1|$ e $h(y) = y$.

Consideraremos que uma observação é atípica se o valor apresentado pela medida de divergência ou distância para a referida observação, for superior a um dado ponto de corte. Para o cálculo de tal do ponto adotamos a proposta apresentada por Peng e Dey (1995) que é dada por

$$P_{\phi}^h(\theta) = \frac{h(\phi(2\theta)) + h(\phi(2(1 - \theta)))}{2},$$

em que θ representa a probabilidade de caras no lançamento de uma moeda viesada. Para o estudo de simulação em questão, adotamos $\theta = 0, 8$. Na Tabela 5.2 temos as expressões e valores para os pontos de corte das medidas utilizadas na detecção de observações atípicas.

Nas tabelas e gráficos de cada cenário a seguir, identificamos por LD ($y - \Delta; x + \Delta$) e LE ($y + \Delta; x - \Delta$) a contaminação correspondente ao lado direito e esquerdo respectivamente

dos dados. Para a contaminação em ambos os lados, adotamos a legenda CLD para nos referir a indicação de observações atípicas apenas no lado direito, CLE para nos referir a indicação de observações atípicas apenas no lado esquerdo e CAL para nos referir a indicação em ambos os lados simultaneamente. Nos gráficos identificamos as medidas de divergência e distância com as seguintes legendas, M1 - distância estocástica Kullback-Leibler, M2 - divergência Kullback-Leibler, M3 - distância estocástica χ^2 , M4 - divergência χ^2 , M5 - distância L_1 norm, M6 - distância estocástica Bhattacharyya, M7 - distância estocástica Hellinger, M8 - distância estocástica triangular e M9 - distância estocástica média-harmônica. A seguir temos os resultados obtidos para cada cenário, seus respectivos gráficos são apresentados separadamente na última seção deste capítulo.

Tabela 5.2: Expressões e valores dos pontos de corte das medidas utilizadas na detecção de observações atípicas.

Medidas	Expressões	$P_\phi^h(0, 8)$
Divergência Kullback-Leibler	$\frac{-\log(2\theta) - \log(2(1-\theta))}{4}$	0,1116
Divergência χ^2	$\frac{(2\theta-1)^2}{4}$	0,9000
Distância estocástica Kullback-Leibler	$\frac{(2\theta-1)}{4} \log\left(\frac{\theta}{1-\theta}\right)$	0,2079
Distância estocástica χ^2	$\frac{(2\theta-1)^2}{16\theta(1-\theta)}(-4\theta + 4\theta + 1)$	0,2306
Distância estocástica Bhattacharyya	$-\log\left(\frac{2\sqrt{2\theta}-2\theta+3}{2}\right) - \log\left(\frac{2\sqrt{2(1-\theta)}-2\theta-1}{2}\right)$	0,1070
Distância estocástica Hellinger	$\frac{(\sqrt{(2\theta)-1})^2 + (\sqrt{(2(1-\theta))-1})^2}{4}$	0,05130
Distância estocástica triangular	$\frac{2(2\theta-1)^2}{(2\theta+1)(3-2\theta)}$	0,1978
Distância estocástica média-harmônica	$-\log\left(-\frac{(2\theta-1)^2}{2(2\theta+1)} + 1\right) - \log\left(-\frac{(1-2\theta)^2}{2(3-2\theta)} + 1\right)$	0,1047
Distância L_1 norm	$\frac{ 2\theta-1 }{2}$	0,3000

5.1 Cenário 1

Nesse cenário geramos a variável resposta a partir de uma distribuição beta retangular e posteriormente ajustamos um modelo de regressão beta retangular, por fim alteramos a variável resposta mantendo fixa a variável regressora.

Na Tabela 5.3 temos as porcentagens das vezes que as medidas indicam as observações como sendo atípicas. Na referida tabela observamos no esquema LD que apenas a divergência Kullback-Leibler se manteve regular, isto é, ela indica 100% das vezes as observações alteradas como atípicas. Já no esquema LE, a distância e a divergência χ^2 ao nível de contaminação de 3% se destacam das demais medidas, apresentando boas porcentagens de detecção de observações atípicas. Observamos conjuntamente nos esquemas LE e LD, que grande parte das medidas apresentam boas porcentagens de detecção de pontos atípicos para o tamanho de amostra igual a 30 e nível de contaminação de 3%.

Tabela 5.3: Porcentagens das vezes que as medidas indicam observações atípicas no cenário 1.

Esquema	Medidas	3%				6%				10%			
		30	60	90	120	30	60	90	120	30	60	90	120
LD	Distância Kullback-Leibler	95,80	9,70	0,00	0,00	0,90	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância χ^2	99,00	45,30	1,20	0,00	11,50	0,00	0,00	0,00	0,20	0,00	0,00	0,00
	Divergência χ^2	100,00	97,30	41,90	1,40	77,80	0,30	0,00	0,00	1,80	0,00	0,00	0,00
	Distância L_1 norm	69,20	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	95,10	6,60	0,00	0,00	0,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	95,10	6,60	0,00	0,00	0,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	92,00	2,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	92,00	2,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
LE	Distância Kullback-Leibler	90,40	3,90	0,00	0,00	0,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	61,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	99,80	100,00	100,00	100,00	94,50	0,00	58,40	54,60	79,20	0,00	2,80	0,30
	Divergência χ^2	99,30	100,00	89,00	89,90	88,00	0,00	29,80	18,00	68,30	0,00	0,60	0,00
	Distância L_1 norm	45,80	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	88,40	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	88,40	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	81,80	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	81,80	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CLD	Distância Kullback-Leibler	0,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	11,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência χ^2	78,80	0,20	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância L_1 norm	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CLE	Distância Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	5,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência χ^2	57,70	0,20	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância L_1 norm	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CAL	Distância Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	5,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência χ^2	57,70	0,20	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância L_1 norm	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Na Figura 5.17 observamos nos esquemas CLD, CLE e CAL, que a divergência χ^2 (M4) com tamanho de amostra igual a 30 e nível de contaminação de 3% é a única que apresenta porcentagens de detecção de observações atípicas acima de 50%. Para os esquemas LD e LE percebemos que todas as medidas apresentam um bom desempenho também para o tamanho de amostra igual a 30 e nível de contaminação de 3%. Podemos destacar ainda no esquema LD, a divergência Kullback-Leibler (M2), que apresentou regularidade de indicação de 100% das observações atípicas, para todos os níveis de contaminação e tamanhos de amostra.

5.2 Cenário 2

Neste cenário geramos a variável resposta a partir de uma distribuição beta retangular e posteriormente ajustamos um modelo de regressão beta retangular, por fim alteramos a variável regressora mantendo fixa a variável resposta.

Na Tabela 5.4 vemos uma configuração semelhante a apresentada pela Tabela 5.3. Observamos nos esquemas LE e LD, que todas as medidas apresentam boas porcentagens

de observações atípicas para o tamanho de amostra igual a 30 e nível de contaminação de 3%. Para a configuração LD a distância e a divergência χ^2 se destacam das demais medidas, apresentando boas porcentagens de detecção de observações atípicas em todos os níveis de contaminação e tamanhos de amostras.

Tabela 5.4: Porcentagens das vezes que as medidas indicam observações atípicas no cenário 2.

Esquema	Medidas	3%				6%				10%			
		30	60	90	120	30	60	90	120	30	60	90	120
LD	Distância Kullback-Leibler	95,60	7,50	0,00	0,00	0,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	78,40	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	98,60	99,90	100,00	100,00	100,00	97,20	100,00	100,00	100,00	84,30	100,00	100,00
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	97,20	100,00	100,00	100,00	84,30	100,00	100,00
	Distância L_1 norm	66,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	94,60	4,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	94,60	4,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	90,50	1,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	90,50	1,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
LE	Distância Kullback-Leibler	95,70	5,00	0,00	0,00	0,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	77,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	99,30	37,50	0,90	0,00	11,30	0,00	0,00	0,00	0,20	0,00	0,00	0,00
	Divergência χ^2	99,90	96,30	28,20	0,80	77,50	0,20	0,00	0,00	2,30	0,00	0,00	0,00
	Distância L_1 norm	66,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	94,60	3,00	0,00	0,00	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	94,60	3,00	0,00	0,00	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	90,40	0,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	90,40	0,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CLD	Distância Kullback-Leibler	0,70	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	12,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência χ^2	83,50	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância L_1 norm	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	0,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	0,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CLE	Distância Kullback-Leibler	0,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	8,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência χ^2	73,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância L_1 norm	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	0,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	0,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CAL	Distância Kullback-Leibler	0,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	8,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência χ^2	73,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância L_1 norm	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	0,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	0,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Na Figura 5.18 observamos nos esquemas CLD, CLE e CAL, que para o tamanho de amostra igual a 30 e nível de contaminação de 3% a divergência χ^2 (M4) é a única a apresentar porcentagens de detecção de observações atípicas superiores a 60%. Também com tamanho de amostra igual a 30 e nível de contaminação de 3%, todas as medidas apresentam porcentagens superiores a 60%, nos esquemas LD e LE. Podemos destacar ainda no esquema LD, que a distância e divergência χ^2 (M3 e M4) tem boas porcentagens de detecção de observações atípicas para todos os tamanhos de amostra e níveis de contaminação.

5.3 Cenário 3

Neste cenário geramos a variável resposta a partir de uma distribuição beta retangular e posteriormente ajustamos um modelo de regressão beta, por fim alteramos a variável

resposta mantendo fixa a variável regressora.

Na Tabela 5.5 temos um comportamento totalmente diferente ao apresentado pelas Tabelas 5.3 e 5.4. Percebemos no esquema LD, que todas as medidas detectam melhor observação atípicas a medida que aumentamos o nível de contaminação. No esquema LE todas as medidas apresentam maiores porcentagens de detecção de observações atípicas que as apresentadas no esquema LD. Em relação aos esquemas CLD, CLE e CAL as medidas apresentam maiores porcentagens para o esquema CLE. Notamos ainda que para os esquemas CLE e LE ao nível de contaminação de 10% as porcentagens apresentadas pelas medidas diminuem conforme o tamanho da amostra aumenta.

Tabela 5.5: Porcentagens das vezes que as medidas indicam observações atípicas no cenário 3.

Esquema	Medidas	3%				6%				10%			
		30	60	90	120	30	60	90	120	30	60	90	120
LD	Distância Kullback-Leibler	59,80	2,50	0,00	0,00	91,90	58,10	0,30	0,00	99,70	98,50	84,30	52,80
	Divergência Kullback-Leibler	52,50	1,30	0,00	0,00	88,20	49,60	0,20	0,00	99,60	98,10	77,40	42,70
	Distância χ^2	77,90	9,20	0,00	0,00	96,10	73,50	2,20	0,00	99,80	99,70	92,50	73,20
	Divergência χ^2	90,60	31,30	0,10	0,00	99,10	91,40	12,90	0,80	99,90	99,90	98,60	92,70
	Distância L_1 norm	82,00	12,40	0,00	0,00	97,50	78,20	2,40	0,00	99,90	99,80	94,00	76,20
	Distância Bhattacharyya	56,30	2,20	0,00	0,00	90,00	55,70	0,30	0,00	99,60	98,40	82,50	50,00
	Distância Hellinger	56,30	2,20	0,00	0,00	90,00	55,70	0,30	0,00	99,60	98,40	82,50	50,00
	Distância triangular	49,90	1,30	0,00	0,00	87,40	48,70	0,20	0,00	99,60	97,90	77,20	43,20
	Distância média-harmônica	49,90	1,30	0,00	0,00	87,40	48,70	0,20	0,00	99,60	97,90	77,20	43,20
LE	Distância Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,50	73,30
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	98,40	60,90
	Distância χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	84,20
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	97,30
	Distância L_1 norm	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,70	76,80
	Distância Bhattacharyya	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,30	71,40
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,30	71,40
	Distância triangular	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,10	66,40
	Distância média-harmônica	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,10	66,40
CLD	Distância Kullback-Leibler	96,90	82,50	13,40	0,00	99,90	100,00	99,00	91,40	100,00	100,00	100,00	100,00
	Divergência Kullback-Leibler	94,40	76,20	7,60	0,00	99,90	99,90	97,80	82,10	100,00	100,00	100,00	100,00
	Distância χ^2	98,90	91,90	31,20	1,30	100,00	100,00	99,80	97,80	100,00	100,00	100,00	100,00
	Divergência χ^2	99,70	97,60	66,90	15,10	100,00	100,00	99,90	99,90	100,00	100,00	100,00	100,00
	Distância L_1 norm	99,00	93,60	34,80	1,40	100,00	100,00	99,80	98,40	100,00	100,00	100,00	100,00
	Distância Bhattacharyya	96,00	81,10	12,00	0,00	99,90	100,00	98,90	89,70	100,00	100,00	100,00	100,00
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância triangular	93,90	76,20	8,10	0,00	99,90	100,00	97,90	83,90	100,00	100,00	100,00	100,00
	Distância média-harmônica	93,90	76,20	8,10	0,00	99,90	100,00	97,90	83,90	100,00	100,00	100,00	100,00
CLE	Distância Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,80	39,80	0,20
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,30	25,50	0,00
	Distância χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	55,80	0,90
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	89,90	18,30
	Distância L_1 norm	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	42,80	0,20
	Distância Bhattacharyya	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,70	37,60	0,10
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,70	37,60	0,10
	Distância triangular	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,70	31,90	0,00
	Distância média-harmônica	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,70	31,90	0,00
CAL	Distância Kullback-Leibler	96,90	82,50	13,40	0,00	99,90	100,00	99,00	91,40	100,00	99,80	39,80	0,20
	Divergência Kullback-Leibler	94,40	76,20	7,60	0,00	99,90	99,90	97,80	82,10	100,00	99,30	25,50	0,00
	Distância χ^2	98,90	91,90	31,20	1,30	100,00	100,00	99,80	97,80	100,00	99,90	55,80	0,90
	Divergência χ^2	99,70	97,60	66,90	15,10	100,00	100,00	99,90	99,90	100,00	100,00	89,90	18,30
	Distância L_1 norm	99,00	93,60	34,80	1,40	100,00	100,00	99,80	98,40	100,00	99,90	42,80	0,20
	Distância Bhattacharyya	96,00	81,10	12,00	0,00	99,90	100,00	98,90	89,70	100,00	99,70	37,60	0,10
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,70	37,60	0,10
	Distância triangular	93,90	76,20	8,10	0,00	99,90	100,00	97,90	83,90	100,00	99,70	31,90	0,00
	Distância média-harmônica	93,90	76,20	8,10	0,00	99,90	100,00	97,90	83,90	100,00	99,70	31,90	0,00

Na Figura 5.19 notamos que todas as medidas parecem ser mais sensíveis a detecção de observações atípicas nos esquemas LE e CLE. Também é possível observar que as medidas apresentam configurações semelhantes dentro de cada esquema.

5.4 Cenário 4

Neste cenário geramos a variável resposta a partir de uma distribuição beta retangular e posteriormente ajustamos um modelo de regressão beta e por fim alteramos a variável regressora mantendo fixa a variável resposta.

Na Tabela 5.6 observamos no esquema LD que a distância e divergência χ^2 apresentam boas porcentagens de identificação de observações atípicas em todos os níveis de contaminação. Assim como no cenário 3, também observamos que no esquema LE todas as medidas apresentam bons desempenhos na detecção de observações atípicas em todos os níveis de contaminação. Em relação a contaminação em ambos os lados, observamos que as medidas apresentam melhor desempenho na detecção de observações atípicas no esquema CLE, apesar de que ao nível de contaminação de 10% notamos que as porcentagens diminuem a medida que aumentamos o tamanho da amostra.

Tabela 5.6: Porcentagens das vezes que as medidas indicam observações atípicas no cenário 4.

Esquema	Medidas	3%				6%				10%			
		30	60	90	120	30	60	90	120	30	60	90	120
LD	Distância Kullback-Leibler	36,90	8,10	0,10	0,00	36,20	17,80	0,00	0,00	26,10	26,70	0,70	0,00
	Divergência Kullback-Leibler	29,50	6,00	0,00	0,00	29,30	11,60	0,00	0,00	21,70	19,90	0,20	0,00
	Distância χ^2	100,00	99,80	100,00	100,00	100,00	97,60	100,00	100,00	100,00	90,00	100,00	100,00
	Divergência χ^2	100,00	99,90	100,00	100,00	100,00	98,10	100,00	100,00	100,00	94,20	100,00	100,00
	Distância L_1 norm	58,10	20,20	0,30	0,00	55,70	32,10	0,70	0,00	42,60	44,00	2,20	0,00
	Distância Bhattacharyya	34,40	7,50	0,10	0,00	33,60	16,10	0,00	0,00	24,30	25,20	0,60	0,00
	Distância Hellinger	34,40	7,50	0,10	0,00	33,60	16,10	0,00	0,00	24,30	25,20	0,60	0,00
	Distância triangular	27,90	5,80	0,00	0,00	28,70	12,40	0,00	0,00	21,50	21,30	0,30	0,00
	Distância média-harmônica	27,90	5,80	0,00	0,00	28,70	12,40	0,00	0,00	21,50	21,30	0,30	0,00
LE	Distância Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância L_1 norm	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância Bhattacharyya	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância triangular	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância média-harmônica	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
CLD	Distância Kullback-Leibler	99,90	99,60	87,90	45,20	99,50	99,70	95,10	62,00	80,70	82,30	18,00	0,30
	Divergência Kullback-Leibler	99,90	99,10	80,70	28,60	99,00	99,60	89,20	44,60	74,00	71,70	7,70	0,00
	Distância χ^2	99,90	99,80	95,40	69,30	99,80	100,00	99,00	82,10	87,30	89,70	35,90	1,30
	Divergência χ^2	100,00	100,00	99,80	96,80	100,00	100,00	99,90	99,00	96,80	98,50	82,10	22,10
	Distância L_1 norm	99,90	99,80	95,60	69,10	99,80	100,00	98,80	79,80	84,60	87,60	27,20	0,90
	Distância Bhattacharyya	99,90	99,50	86,80	42,30	99,50	99,70	94,00	58,90	79,20	80,80	15,80	0,20
	Distância Hellinger	99,90	99,50	86,80	42,30	99,50	99,70	94,00	58,90	79,20	80,80	15,80	0,20
	Distância triangular	99,90	99,20	82,10	32,00	99,30	99,70	91,40	49,20	76,40	76,20	10,30	0,00
	Distância média-harmônica	99,90	99,20	82,10	32,00	99,30	99,70	91,40	49,20	76,40	76,20	10,30	0,00
CLE	Distância Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	93,20	53,60	3,80
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,80	90,30	42,90	1,60
	Distância χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	95,30	66,40	7,90
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	98,00	88,40	35,10
	Distância L_1 norm	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	93,00	55,50	4,60
	Distância Bhattacharyya	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	92,80	51,90	3,20
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	92,80	51,90	3,20
	Distância triangular	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,80	92,00	47,60	2,50
	Distância média-harmônica	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,80	92,00	47,60	2,50
CAL	Distância Kullback-Leibler	99,90	99,60	87,90	45,20	99,50	99,70	95,10	62,00	80,60	78,00	11,60	0,10
	Divergência Kullback-Leibler	99,90	99,10	80,70	28,60	99,00	99,60	89,20	44,60	73,90	65,90	3,70	0,00
	Distância χ^2	99,90	99,80	95,40	69,30	99,80	100,00	99,00	82,10	87,30	85,80	25,80	0,20
	Divergência χ^2	100,00	100,00	99,80	96,80	100,00	100,00	99,90	99,00	96,80	96,50	73,90	9,80
	Distância L_1 norm	99,90	99,80	95,60	69,10	99,80	100,00	98,80	79,80	84,50	82,20	17,70	0,10
	Distância Bhattacharyya	99,90	99,50	86,80	42,30	99,50	99,70	94,00	58,90	79,10	76,30	9,80	0,10
	Distância Hellinger	99,90	99,50	86,80	42,30	99,50	99,70	94,00	58,90	79,10	76,30	9,80	0,10
	Distância triangular	99,90	99,20	82,10	32,00	99,30	99,70	91,40	49,20	76,30	71,20	5,90	0,00
	Distância média-harmônica	99,90	99,20	82,10	32,00	99,30	99,70	91,40	49,20	76,30	71,20	5,90	0,00

Na Figura 5.20 observamos claramente que todas as medidas apresentam melhor desempenho nas configurações do lado esquerdo, LE e CLE. No esquema LD, notamos que a distância e divergência χ^2 (M3 e M4) se destacam das demais medidas, apresentando

boas porcentagens de detecção de observações atípicas. De modo geral, assim como na Figura 5.19 notamos que as medidas apresentam configurações semelhantes dentro de cada esquema.

5.5 Cenário 5

Neste cenário geramos a variável resposta a partir de uma distribuição beta e posteriormente ajustamos um modelo de regressão beta retangular e por fim alteramos a variável dependente mantendo a variável regressora fixa.

Na Tabela 5.7 notamos uma configuração semelhante a apresentada pela tabela do cenário 1. Observamos no esquema LE que a divergência Kullback-Leibler é regular na indicação de observações atípicas, para todos os níveis de contaminação e tamanhos de amostras. No esquema LE para o nível de contaminação de 3% e todos os tamanhos de amostras, observamos que a distância e divergência χ^2 se destacam das demais medidas.

Tabela 5.7: Porcentagens das vezes que as medidas indicam observações atípicas no cenário 5.

Esquema	Medidas	3%				6%				10%			
		30	60	90	120	30	60	90	120	30	60	90	120
LD	Distância Kullback-Leibler	97,60	21,30	0,00	0,00	1,70	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90
	Distância χ^2	99,50	59,40	1,40	0,00	24,10	0,00	0,00	0,00	0,40	0,00	0,00	0,00
	Divergência χ^2	100,00	95,30	48,60	2,00	88,80	0,10	0,00	0,00	3,40	0,00	0,00	0,00
	Distância L_1 norm	88,10	0,70	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	97,30	16,00	0,00	0,00	0,70	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	97,30	16,00	0,00	0,00	0,70	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	95,80	6,70	0,00	0,00	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	95,80	6,70	0,00	0,00	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00
LE	Distância Kullback-Leibler	95,40	14,40	0,00	0,00	0,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	85,80	0,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	100,00	100,00	100,00	100,00	92,10	0,00	84,60	96,90	66,40	0,00	1,20	0,20
	Divergência χ^2	99,60	100,00	100,00	100,00	78,70	0,10	48,40	64,70	51,80	0,00	0,20	0,00
	Distância L_1 norm	78,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	94,80	10,30	0,00	0,00	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	94,80	10,30	0,00	0,00	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	93,20	4,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	93,20	4,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CLD	Distância Kullback-Leibler	2,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	25,80	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência χ^2	91,60	0,40	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância L_1 norm	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	1,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	1,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CLE	Distância Kullback-Leibler	0,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	12,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência χ^2	82,50	0,20	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância L_1 norm	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CAL	Distância Kullback-Leibler	0,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência Kullback-Leibler	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância χ^2	12,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Divergência χ^2	82,50	0,10	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância L_1 norm	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Bhattacharyya	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância Hellinger	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância triangular	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Distância média-harmônica	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Na Figura 5.21 notamos nos esquemas CLD, CLE e CLA que a divergência χ^2 (M4) se destaca das demais medidas, apresentando porcentagens acima de 75% ao nível de

A Figura 5.22 apresenta uma configuração semelhante a apresentada pela Figura 5.21, isto é, para os esquemas CLD, CLE e CAL, ao nível de contaminação de 3% e tamanho de amostra igual a 30, observamos que a divergência χ^2 (M4) apresenta porcentagem de detecção de observações atípicas superior a 75%. Para os esquemas LD e LE, temos que todas as medidas apresentam porcentagens de detecção de observações atípicas acima de 75% no nível de contaminação de 3% e tamanho de amostra igual a 30.

5.7 Cenário 7

Neste cenário geramos a variável resposta a partir de uma distribuição beta e posteriormente ajustamos um modelo de regressão beta, por fim alteramos a variável resposta mantendo fixa a variável regressora.

Na Tabela 5.9 notamos, mais uma vez, que as medidas apresentam maior porcentagem de detecção de observações atípicas no esquema LE. Vemos no esquema LD que todas as medidas apresentam bom desempenho ao nível de contaminação de 10% para todos os tamanhos de amostra. Neste mesmo esquema, observamos nos outros níveis de contaminação, que a medida em que aumentamos o tamanho da amostra, as porcentagens de detecção de observações atípicas diminuem rapidamente. No esquema CLD, destacamos o bom desempenho das medidas para os níveis de contaminação de 6% e 10% para todos os tamanhos de amostra. Ainda neste esquema, destacamos o bom desempenho das medidas ao nível de contaminação de 3% e tamanhos de amostras iguais a 30 e 60. Em relação ao esquema CLE, vemos que as porcentagens de indicação de observações atípicas não se mantêm constante apenas no nível de contaminação de 10% quanto temos tamanhos de amostra iguais a 90 e 120.

Na Figura 5.23 notamos, de modo geral, que as medidas apresentam configurações semelhantes dentro de cada esquema. Observamos que as medidas parecem ser mais sensíveis a detecção de observações atípicas nos esquemas LE e CLE. Notamos ainda, que as Figuras 5.23 e 5.19 apresentam configurações semelhantes.

Tabela 5.9: Porcentagens das vezes que as medidas indicam observações atípicas no cenário 7.

Esquema	Medidas	3%				6%				10%			
		30	60	90	120	30	60	90	120	30	60	90	120
LD	Distância Kullback-Leibler	84,80	20,90	0,30	0,00	98,70	86,30	8,00	0,50	100,00	99,90	98,30	92,70
	Divergência Kullback-Leibler	79,60	13,30	0,10	0,00	97,70	80,60	4,70	0,10	100,00	99,70	97,10	88,40
	Distância χ^2	93,70	38,70	0,50	0,00	99,70	94,20	21,30	1,90	100,00	100,00	99,60	97,20
	Divergência χ^2	98,20	69,80	5,90	0,10	100,00	98,70	51,10	14,60	100,00	100,00	100,00	99,50
	Distância L_1 norm	95,40	45,00	0,90	0,00	99,80	95,40	23,90	2,20	100,00	100,00	99,80	97,90
	Distância Bhattacharyya	83,10	19,40	0,20	0,00	98,30	84,80	7,10	0,50	100,00	99,90	98,20	91,70
	Distância Hellinger	83,10	19,40	0,20	0,00	98,30	84,80	7,10	0,50	100,00	99,90	98,20	91,70
	Distância triangular	78,40	13,30	0,10	0,00	97,30	80,10	4,60	0,10	100,00	99,70	97,20	88,70
LE	Distância média-harmônica	78,40	13,30	0,10	0,00	97,30	80,10	4,60	0,10	100,00	99,70	97,20	88,70
	Distância Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	97,70
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	94,30
	Distância χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,60
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância L_1 norm	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	98,50
	Distância Bhattacharyya	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	97,50
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	97,50
CLD	Distância triangular	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	96,60
	Distância média-harmônica	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	96,60
	Distância Kullback-Leibler	99,90	95,20	46,70	5,00	100,00	100,00	100,00	99,90	100,00	100,00	100,00	100,00
	Divergência Kullback-Leibler	99,50	93,00	34,30	2,00	100,00	100,00	100,00	99,60	100,00	100,00	100,00	100,00
	Distância χ^2	99,90	98,40	67,60	15,90	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Divergência χ^2	100,00	99,50	91,10	53,10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância L_1 norm	100,00	98,70	71,40	16,40	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância Bhattacharyya	99,90	94,50	43,80	4,10	100,00	100,00	100,00	99,80	100,00	100,00	100,00	100,00
CLE	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância triangular	99,40	93,10	35,40	2,40	100,00	100,00	100,00	99,70	100,00	100,00	100,00	100,00
	Distância média-harmônica	99,40	93,10	35,40	2,40	100,00	100,00	100,00	99,70	100,00	100,00	100,00	100,00
	Distância Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	76,40	8,30
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	63,00	2,40
	Distância χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	87,50	16,40
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,30	61,10
	Distância L_1 norm	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	80,10	10,30
CAL	Distância Bhattacharyya	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	75,20	7,30
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	75,20	7,30
	Distância triangular	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	71,30	5,00
	Distância média-harmônica	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	71,30	5,00
	Distância Kullback-Leibler	99,90	95,20	46,70	5,00	100,00	100,00	100,00	99,90	100,00	100,00	76,40	8,30
	Divergência Kullback-Leibler	99,50	93,00	34,30	2,00	100,00	100,00	100,00	99,60	100,00	100,00	63,00	2,40
	Distância χ^2	99,90	98,40	67,60	15,90	100,00	100,00	100,00	100,00	100,00	100,00	87,50	16,40
	Divergência χ^2	100,00	99,50	91,10	53,10	100,00	100,00	100,00	100,00	100,00	100,00	99,30	61,10

5.8 Cenário 8

Neste cenário geramos a variável resposta a partir de uma distribuição beta e posteriormente ajustamos um modelo de regressão beta, por fim alteramos a variável regressora mantendo fixa a variável resposta.

Na Tabela 5.10 observamos inicialmente que no esquema LE todas as medidas indicam 100% das observações atípicas em todos os níveis de contaminação e tamanhos de amostras. No esquema LD notamos que a distância e divergência χ^2 se destaca das demais medidas, por manter uma certa regularidade entre os níveis de contaminação e tamanhos de amostras. Todas as medidas apresentam boas porcentagens de indicação de observações atípicas nos esquemas CLD e CLE, contudo, notamos que em ambos os esquemas, ao nível de contaminação de 10%, a medida que aumentamos o tamanho da amostra as porcentagens diminuem rapidamente.

Tabela 5.10: Porcentagens das vezes que as medidas indicam observações atípicas no cenário 8.

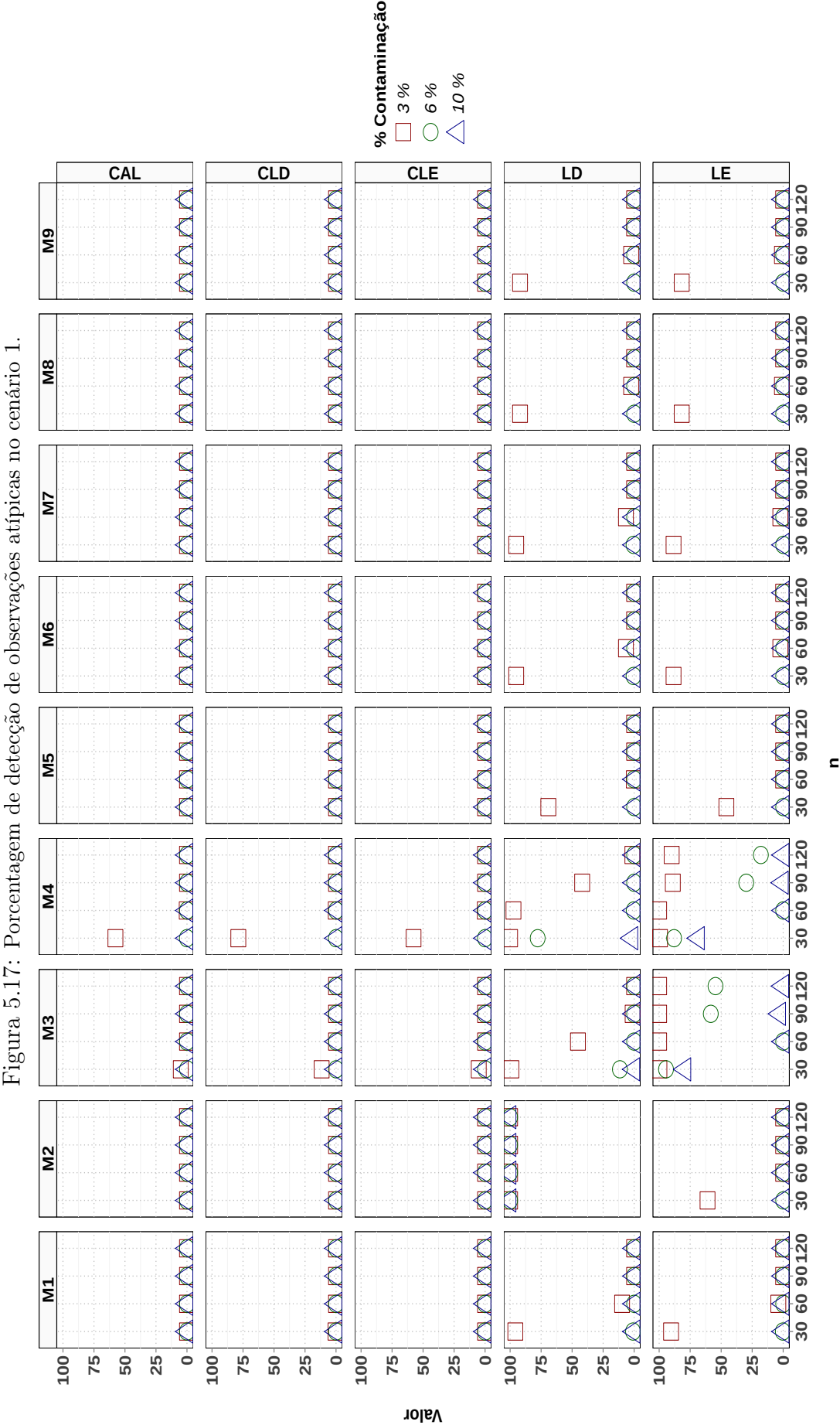
Esquema	Medidas	3%				6%				10%			
		30	60	90	120	30	60	90	120	30	60	90	120
LD	Distância Kullback-Leibler	81,60	67,60	19,20	2,10	80,70	79,00	24,70	3,60	77,40	88,00	44,80	7,70
	Divergência Kullback-Leibler	77,30	58,50	13,00	1,50	74,20	72,00	16,30	1,40	72,70	82,40	31,20	4,00
	Distância χ^2	100,00	99,80	100,00	100,00	100,00	99,00	100,00	100,00	100,00	99,00	100,00	100,00
	Divergência χ^2	100,00	99,90	100,00	100,00	100,00	99,90	100,00	100,00	100,00	99,70	100,00	100,00
	Distância L_1 norm	93,40	83,80	35,00	6,50	92,50	91,00	41,30	10,20	89,90	95,10	66,60	17,30
	Distância Bhattacharyya	80,70	65,30	17,60	2,00	78,70	77,10	22,70	3,00	76,60	86,70	41,70	6,90
	Distância Hellinger	80,70	65,30	17,60	2,00	78,70	77,10	22,70	3,00	76,60	86,70	41,70	6,90
	Distância triangular	76,40	58,80	13,30	1,50	74,00	72,30	17,10	1,60	72,40	83,50	33,90	4,90
LE	Distância média-harmônica	76,40	58,80	13,30	1,50	74,00	72,30	17,10	1,60	72,40	83,50	33,90	4,90
	Distância Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância L_1 norm	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância Bhattacharyya	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
CLD	Distância triangular	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância média-harmônica	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Distância Kullback-Leibler	100,00	100,00	99,90	97,90	99,90	100,00	100,00	98,40	95,10	95,40	56,80	4,40
	Divergência Kullback-Leibler	100,00	100,00	99,70	94,60	99,90	100,00	100,00	95,90	93,00	91,10	40,20	2,00
	Distância χ^2	100,00	100,00	100,00	99,40	100,00	100,00	100,00	99,90	96,60	97,60	72,20	13,80
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,40	99,90	96,00	62,40
	Distância L_1 norm	100,00	100,00	100,00	99,30	99,90	100,00	100,00	99,80	95,70	96,30	65,50	8,80
	Distância Bhattacharyya	100,00	100,00	99,90	97,50	99,90	100,00	100,00	98,10	94,80	94,70	54,40	3,40
CLE	Distância Hellinger	100,00	100,00	99,90	97,50	99,90	100,00	100,00	98,10	94,80	94,70	54,40	3,40
	Distância triangular	100,00	100,00	99,80	95,60	99,90	100,00	100,00	97,00	94,10	93,20	46,80	2,60
	Distância média-harmônica	100,00	100,00	99,80	95,60	99,90	100,00	100,00	97,00	94,10	93,20	46,80	2,60
	Distância Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	95,90	67,70	9,30
	Divergência Kullback-Leibler	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,80	93,80	56,30	4,30
	Distância χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	97,40	77,10	18,20
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,10	93,60	53,20
	Distância L_1 norm	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	95,70	68,90	10,60
CAL	Distância Bhattacharyya	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	95,60	66,10	8,20
	Distância Hellinger	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	95,60	66,10	8,20
	Distância triangular	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	95,20	61,80	6,10
	Distância média-harmônica	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,90	95,20	61,80	6,10
	Distância Kullback-Leibler	100,00	100,00	99,90	97,90	99,90	100,00	100,00	98,40	95,00	91,60	41,50	0,50
	Divergência Kullback-Leibler	100,00	100,00	99,70	94,60	99,90	100,00	100,00	95,90	92,80	86,40	24,80	0,00
	Distância χ^2	100,00	100,00	100,00	99,40	100,00	100,00	100,00	99,90	96,50	95,30	58,10	3,50
	Divergência χ^2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,40	99,00	90,50	36,60
CAL	Distância L_1 norm	100,00	100,00	100,00	99,30	99,90	100,00	100,00	99,80	95,60	92,30	49,00	1,10
	Distância Bhattacharyya	100,00	100,00	99,90	97,50	99,90	100,00	100,00	98,10	94,70	90,80	38,70	0,40
	Distância Hellinger	100,00	100,00	99,90	97,50	99,90	100,00	100,00	98,10	94,70	90,80	38,70	0,40
	Distância triangular	100,00	100,00	99,80	95,60	99,90	100,00	100,00	97,00	94,00	89,10	31,70	0,10
	Distância média-harmônica	100,00	100,00	99,80	95,60	99,90	100,00	100,00	97,00	94,00	89,10	31,70	0,10
	Distância Kullback-Leibler	100,00	100,00	99,90	97,90	99,90	100,00	100,00	98,40	95,00	91,60	41,50	0,50
	Divergência Kullback-Leibler	100,00	100,00	99,70	94,60	99,90	100,00	100,00	95,90	92,80	86,40	24,80	0,00
	Distância χ^2	100,00	100,00	100,00	99,40	100,00	100,00	100,00	99,90	96,50	95,30	58,10	3,50

Na Figura 5.24 vemos uma configuração muito semelhante a encontrada na Figura 5.20. Isto é, as medidas se mostram mais sensíveis a detecção de observações atípicas nos esquemas LE e CLE, as configurações apresentadas pelas medidas são semelhantes dentro de cada esquema e destacamos a regularidade na indicação de observações atípicas da distância e divergência χ^2 (M3 e M4) no esquema LD.

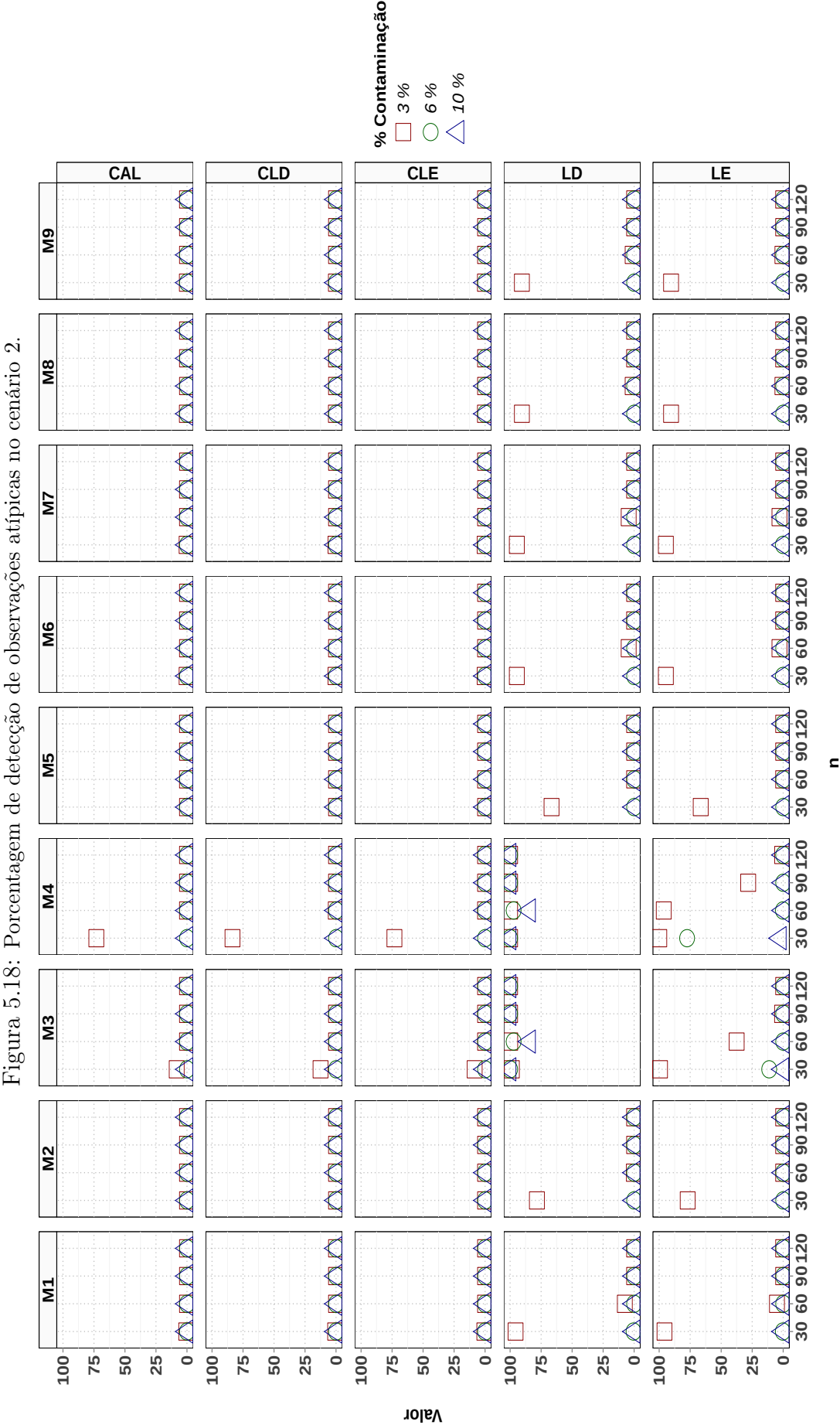
De modo geral, as configurações das tabelas e gráficos estão se repetindo dependendo do modelo ajustado. Nesse estudo, observamos que a divergência χ^2 demonstrou maior eficiência que as demais medidas na detecção de observações atípicas. Percebemos que o modelo de regressão beta retangular é mais apropriado que o modelo de regressão beta, no caso em que temos a presença de observações atípicas. Notamos também que as medidas apresentam melhor desempenho na detecção de observações atípicas quando o modelo ajustado não corresponde a verdadeira distribuição geradora dos dados.

5.9 Gráficos referentes as simulações

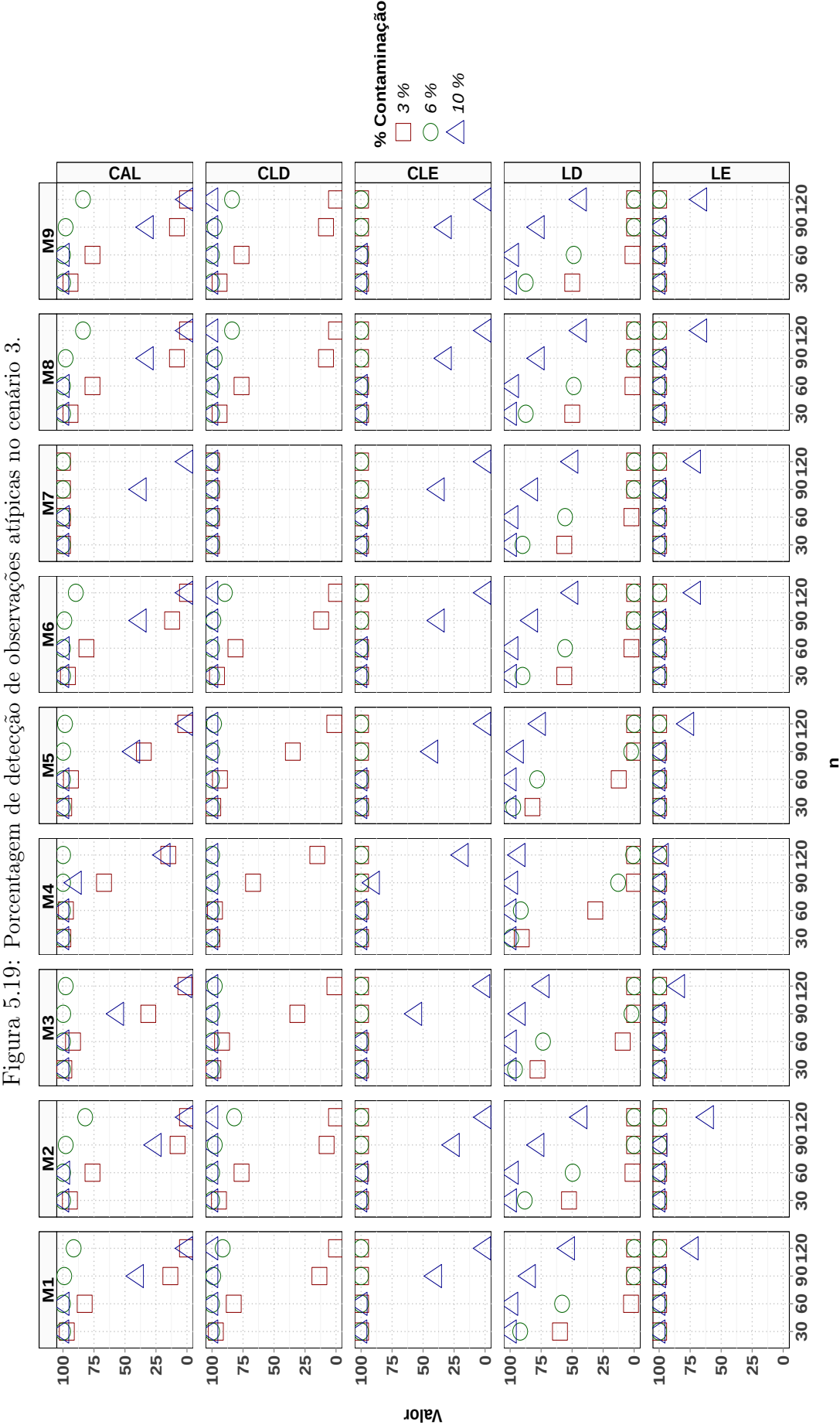
Nesta seção nós dispomos todos os gráficos referentes aos 8 cenários tratados na simulação.



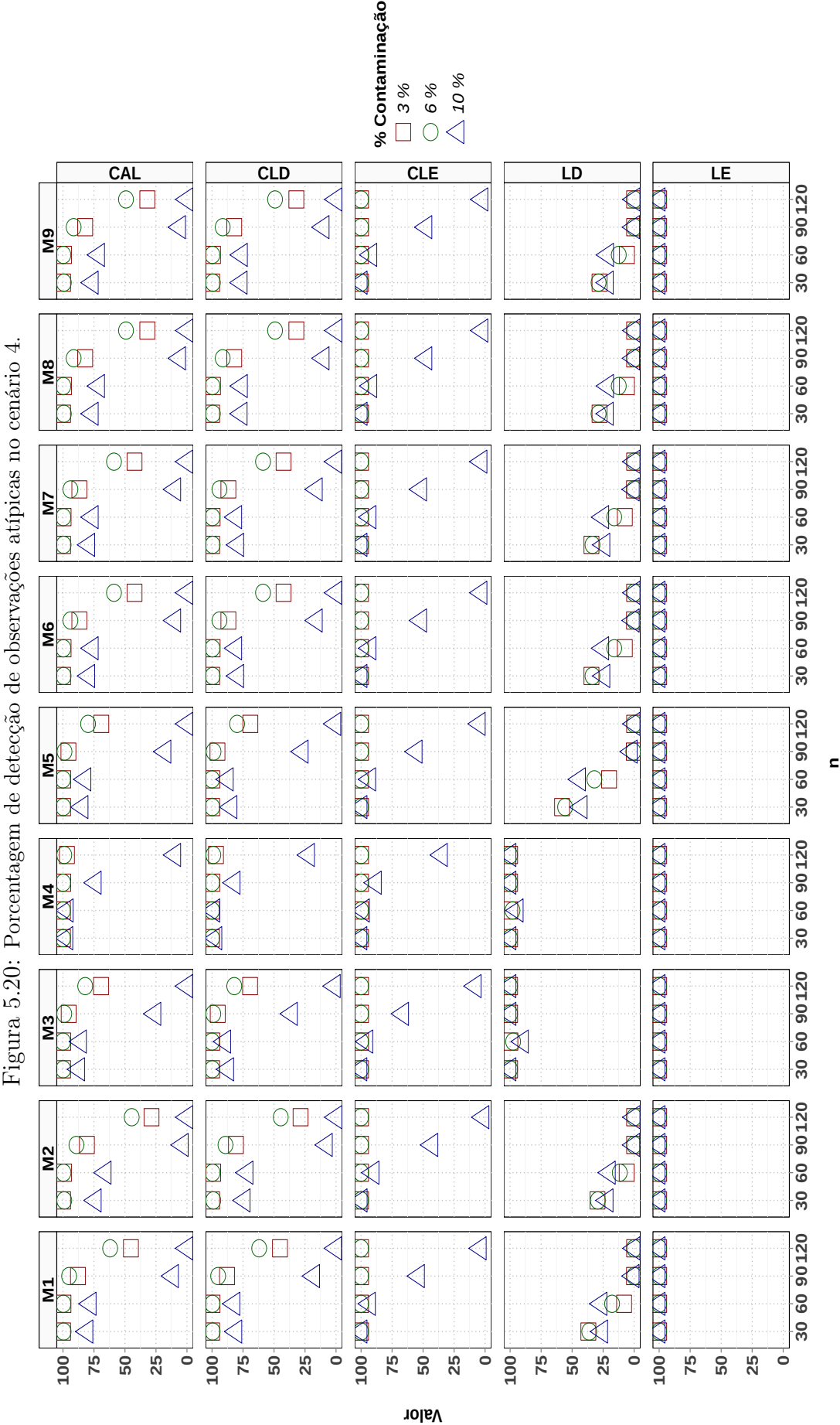
Fonte: Dados próprios.



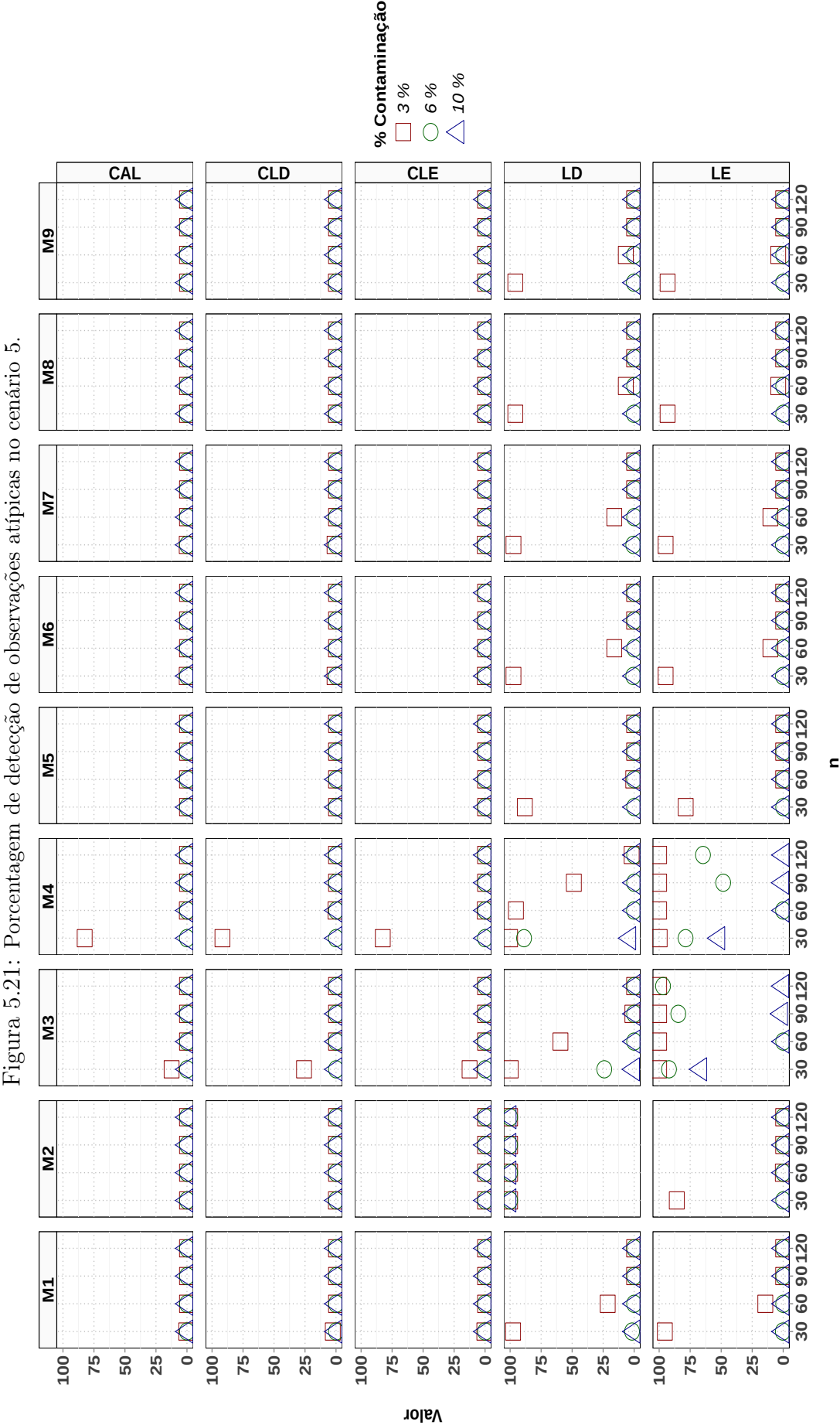
Fonte: Dados próprios.



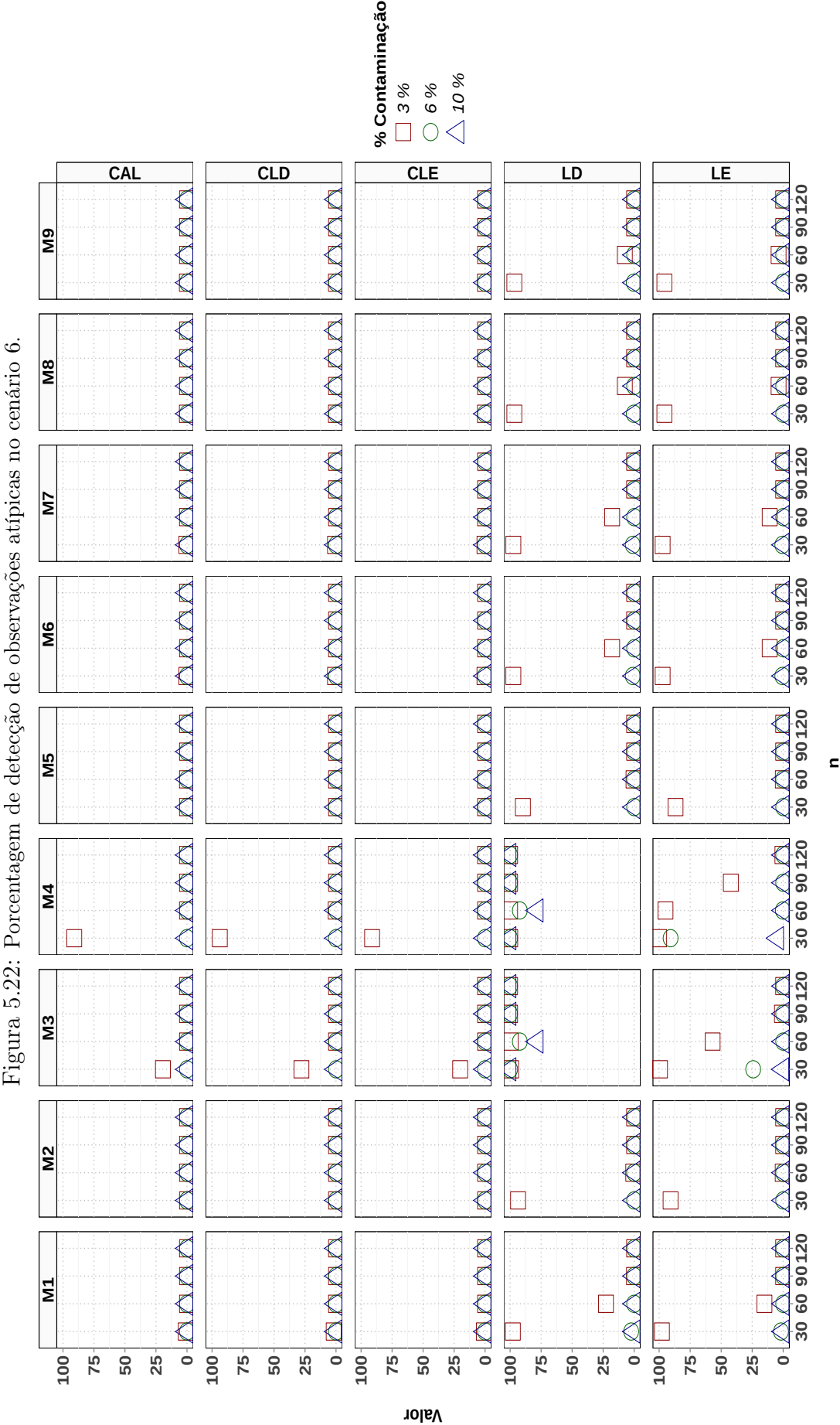
Fonte: Dados próprios.



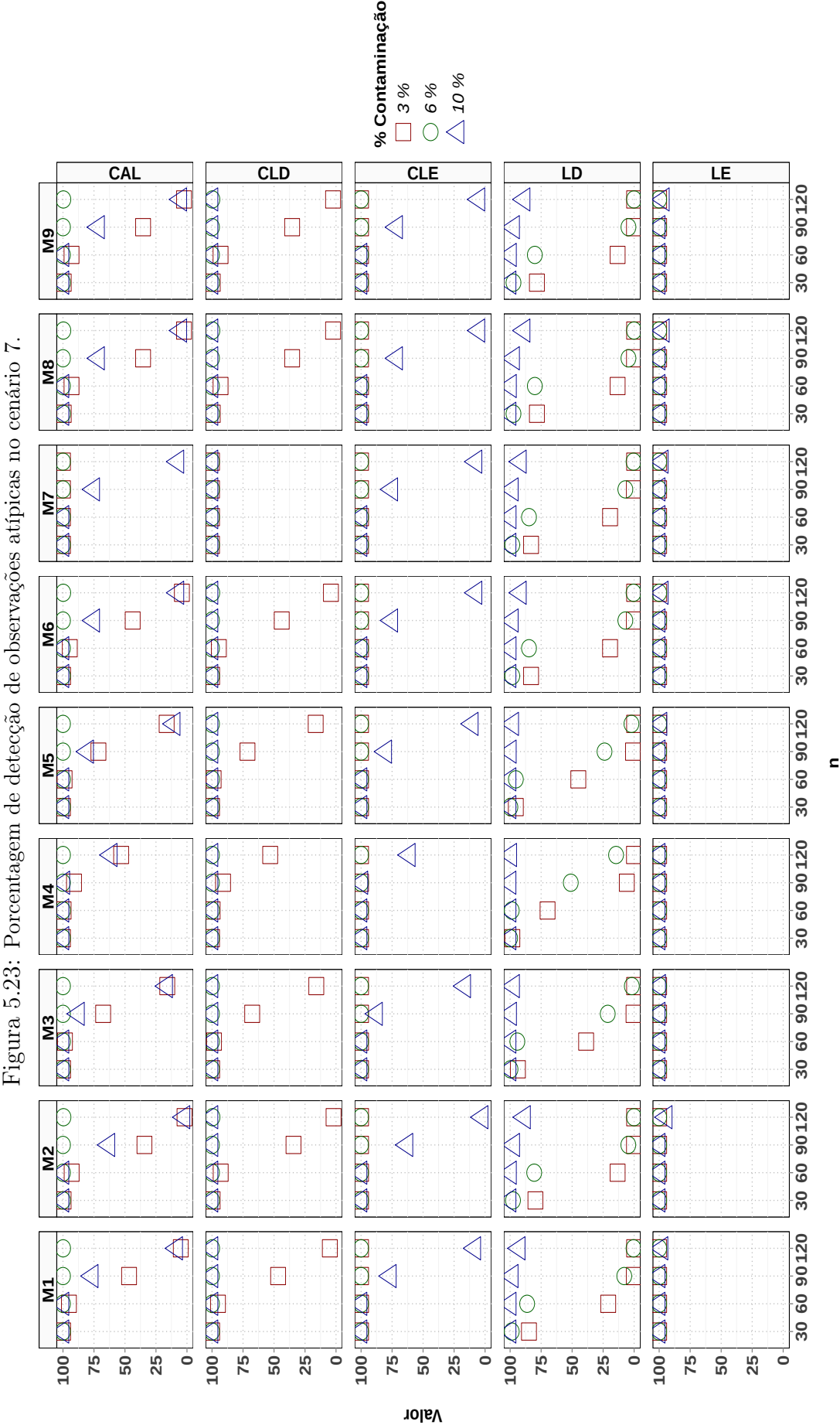
Fonte: Dados próprios.



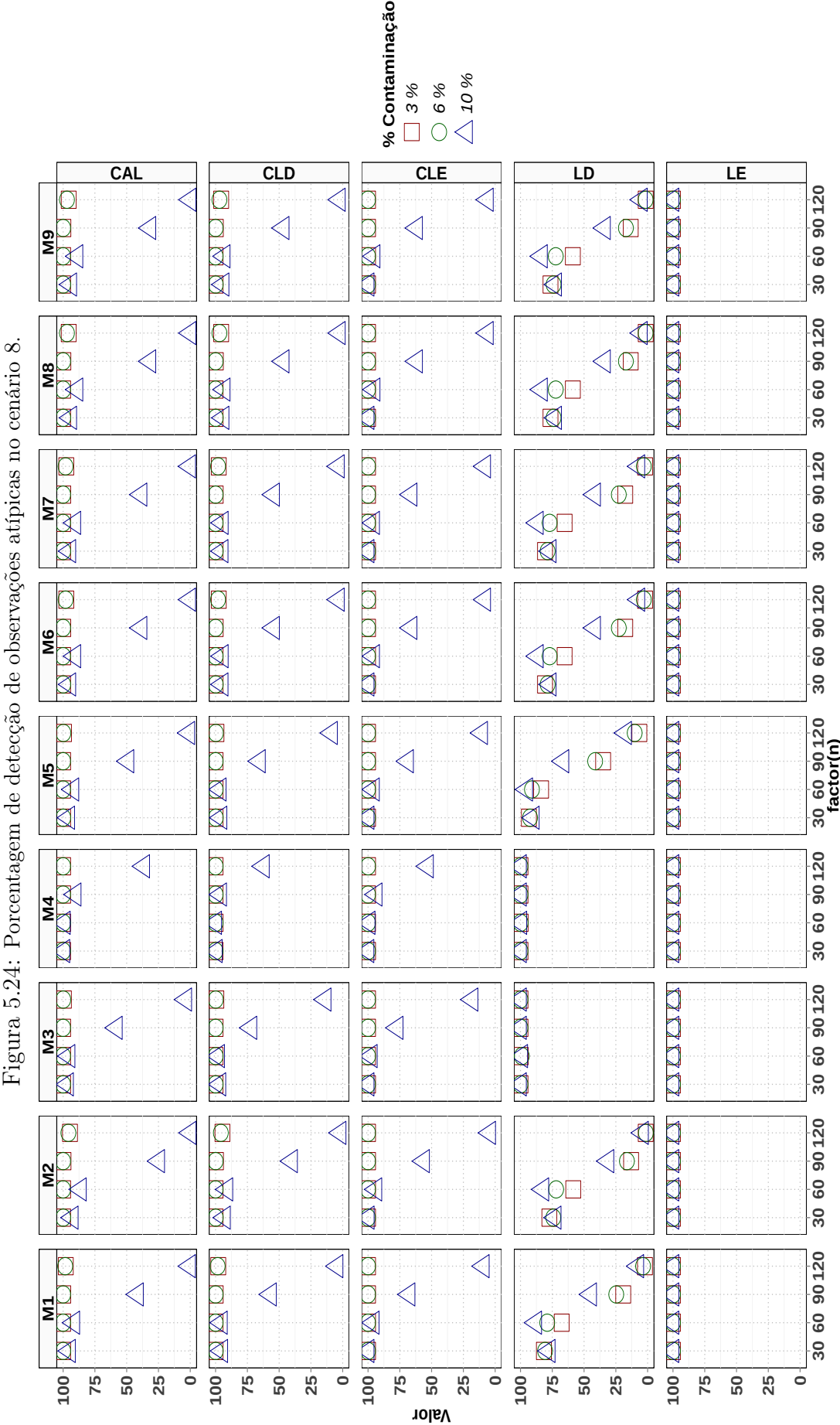
Fonte: Dados próprios.



Fonte: Dados próprios.



Fonte: Dados próprios.



Fonte: Dados próprios.

CAPÍTULO 6

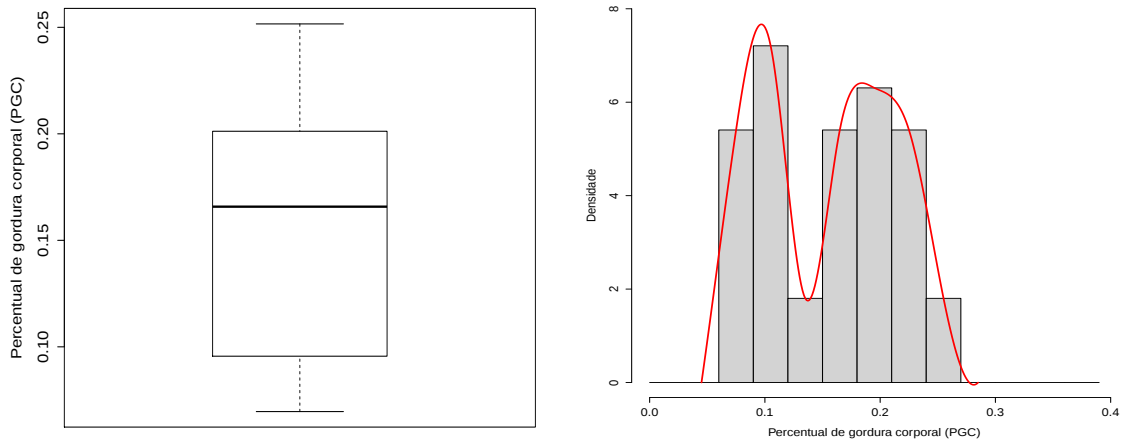
Aplicação

Neste capítulo apresentamos uma aplicação do uso das divergências Kullback-Leibler e χ^2 e das distâncias estocásticas Kullback-Leibler, χ^2 , Bhattacharyya, Hellinger, triangular e média-harmônica, bem como da distância L_1 norm na detecção de observações atípicas nos ajustes dos modelos de regressão beta e beta retangular. O banco de dados utilizado se refere a um estudo esportivo conduzido pelo Instituto Australiano de Esportes (AIS) no qual são analisadas 13 variáveis em 202 atletas praticantes de diversos esportes. Tal banco se encontra disponível no pacote `sn` do software estatístico R.

Os ajustes dos modelos de regressão beta e beta retangular serão feitos segundo a abordagem Bayesiana, para os modelos ajustados usamos o método MCMC para obtermos uma cadeia de tamanho 100.000, utilizando um *burn-in* igual a 50.000 e um *thin* igual a 10, obtivemos uma amostra final de tamanho 5.000 das distribuições *a posteriori* dos parâmetros dos modelos. Consideraremos como variável resposta o percentual de gordura corporal (PGC) e como variáveis explicativas a massa corporal magra (MCM) e o sexo referente aos 37 atletas que praticam remo. Na Figura 6.1 apresentamos o gráfico box-plot (à esquerda) e o histograma (à direita) da variável resposta, por ambos os gráficos observamos que a distribuição da variável é assimétrica no intervalo (0, 1).

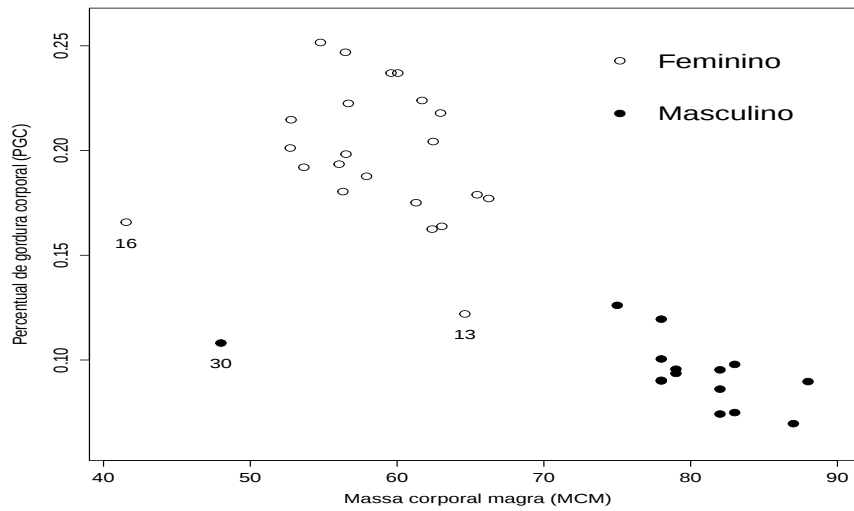
Na Figura 6.2 temos o gráfico de dispersão entre a variável resposta PGC e a variável explicativa MCM, podemos observar que estas variáveis apresentam um grau de correlação negativa, isto é, a medida que os valores da variável MCM aumentam os valores da variável PGC diminuem. Na mesma figura podemos ver a presença de dois grupos na nuvem de pontos, como também a presença de 3 observações que se destacam dos grupos. As observações 16 e 30 apresentam os menores valores da variável MCM. A observação 13 apresenta o menor valor da variável PGC referente as observações do sexo feminino.

Figura 6.1: Boxplot e histograma da variável percentual de gordura corporal.



Fonte: Dados próprios.

Figura 6.2: Gráfico de dispersão entre as PGC e MCM referente aos 37 atletas que praticam remo.



Fonte: Dados próprios.

A seguir ajustaremos o modelo de regressão beta.

6.1 Modelo de regressão beta

Nesta seção, adotaremos um modelo de regressão beta com dispersão constante, da forma

$$PGC_i \sim \mathcal{B}(\mu_i, \phi), \text{logito}(\mu_i) = \beta_0 + \beta_1 \times MCM_i + \beta_2 \times \text{sexo}_i, i = 1, 2, \dots, 37, \quad (6.1)$$

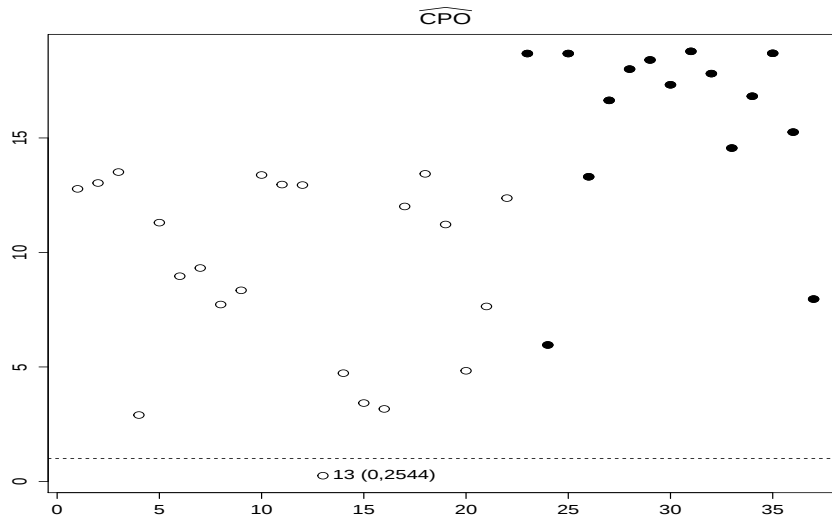
em que $\text{sexo}_i = 1$ se i representar uma observação do sexo masculino e $\text{sexo}_i = 0$ caso contrário. Conforme indicado pela literatura, adotamos $\phi \sim \mathcal{G}(0, 01; 0, 01)$ como distribuição

a *priori* para ϕ e $\beta_i \sim \mathcal{N}(\mu_i^*, \sigma^2)$ para os β_i 's.

Para uma escolha inicial dos hiperparâmetros das distribuições dos β_i 's adotamos $\beta_i \sim \mathcal{N}(0; 10^2)$. Posteriormente, ajustamos o modelo com tais *prioris*, e usamos as médias e desvios-padrão das distribuições a *posteriori* dos β_i 's como novos hiperparâmetros. Continuamos a fazer isto sucessivamente até não observarmos significantes mudanças nas estimativas dos parâmetros em relação ao modelo ajustado anteriormente. No final do ajuste de 12 modelos nós adotamos as seguintes distribuições a *priori* para os β_i 's, são elas, $\beta_0 \sim \mathcal{N}(-1,04644; 0,02515^2)$, $\beta_1 \sim \mathcal{N}(-0,00597; 0,00039^2)$ e $\beta_2 \sim \mathcal{N}(-0,75024; 0,0305^2)$. Na Tabela 6.1 temos as médias das distribuições a *posteriori* dos parâmetros do modelo proposto, bem como o intervalo de credibilidade de 95% e os valores dos critérios de comparação de modelos adotados. Vemos pela tabela que todos os parâmetros são significativos para o modelo.

Na Figura 6.3 temos o gráfico do $\widehat{CPO}$, em que os pontos sem preenchimento se referem as observações do sexo feminino, enquanto que os pontos com preenchimento se referem as observações do sexo masculino. Em destaque vemos que a observação 13 tem o menor valor desta medida, isto nos fornece indícios que esta observação não está sendo bem ajustada pelo modelo.

Figura 6.3: $\widehat{CPO}$ do modelo de regressão beta ajustado.

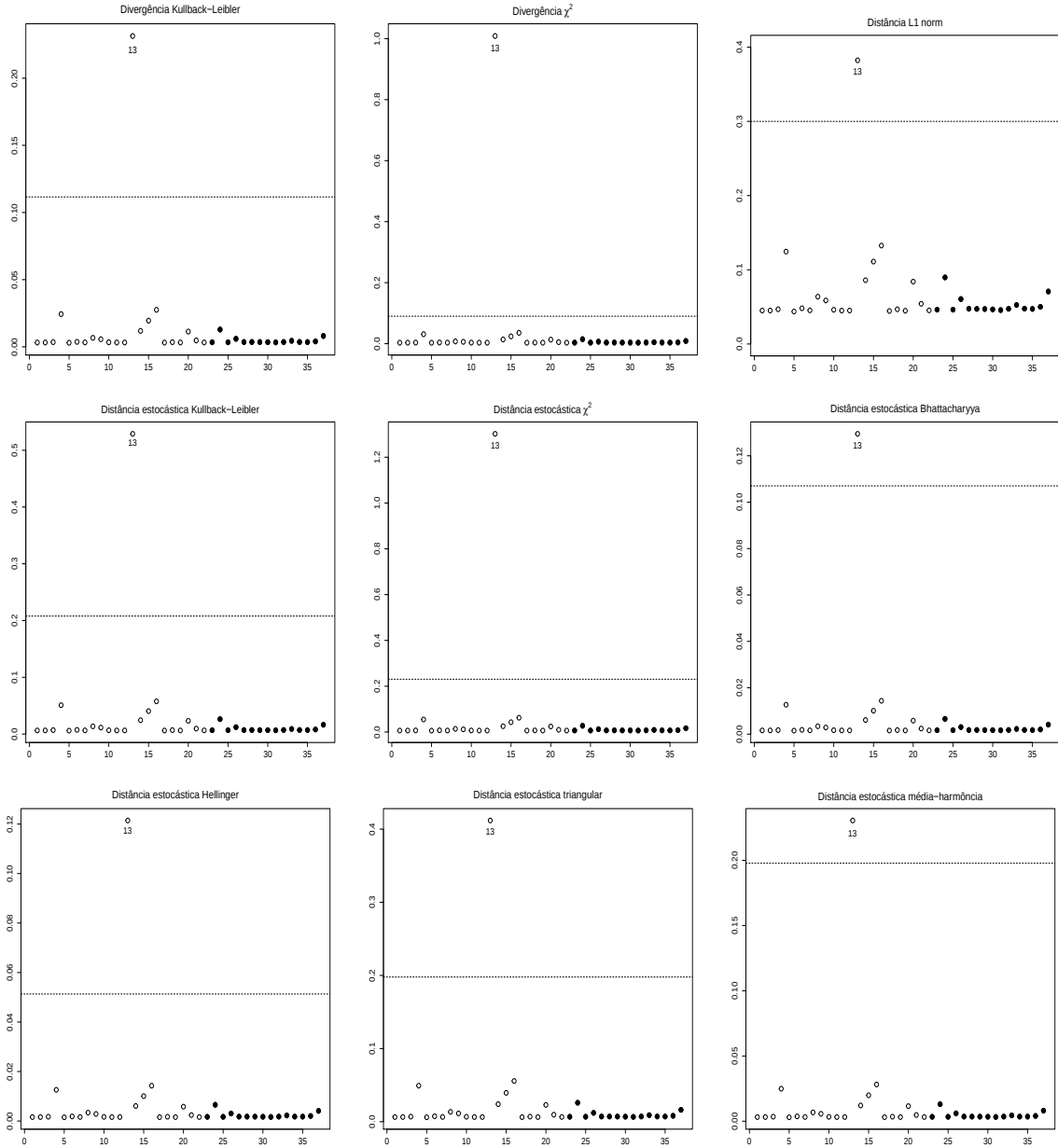


Fonte: Dados próprios.

Na Figura 6.4 apresentamos os gráficos das medidas utilizadas na identificação de observações atípicas. Estamos usando, assim como nas simulações, $\theta = 0,8$ para o cálculo do ponto de corte. Os pontos sem preenchimento se referem as observações do sexo feminino, enquanto que os pontos com preenchimento se referem as observações do sexo masculino. Notamos que a observação 13 é indicada por todas as medidas como sendo atípica. Por este indício, posteriormente ajustaremos mais uma vez o modelo de regressão

beta sem a presença dessa observação, com o intuito de verificarmos seu impacto sobre o modelo ajustado.

Figura 6.4: Gráficos das medidas utilizadas na identificação de observações atípicas no ajuste do modelo de regressão beta.



Fonte: Dados próprios.

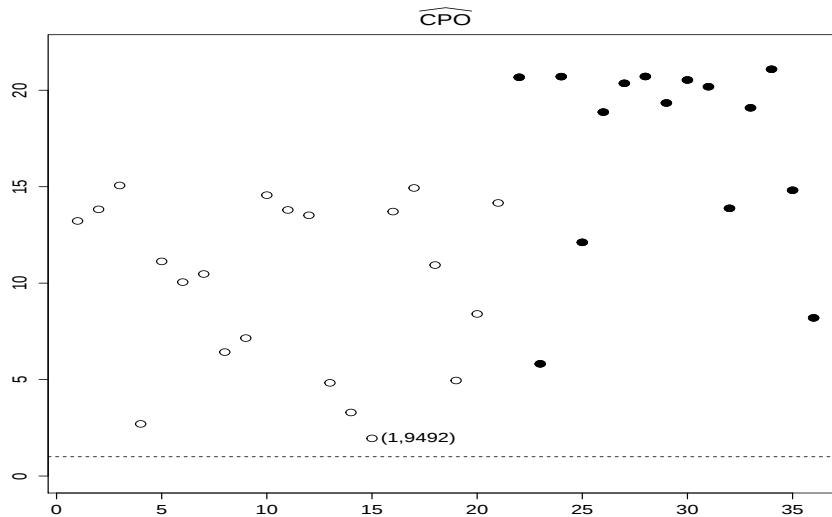
6.1.1 Modelo de regressão beta sem a observação 13

A escolha dos hiperparâmetros das distribuições dos β_i 's foi feita de mesmo modo que no ajuste do modelo de regressão beta com todas as observações, assim nos adotamos, $\beta_0 \sim \mathcal{N}(-1,03996; 0,01438^2)$, $\beta_1 \sim \mathcal{N}(-0,00587; 0,00023^2)$ e $\beta_2 \sim \mathcal{N}(-0,75128; 0,02201^2)$. Na Tabela 6.1 observamos as médias das distribuições a *posteriori* dos parâmetros do

novo modelo ajustado bem como o intervalo de credibilidade de 95%. Notamos que com a retirada desta observação as inferências com relação aos parâmetros do modelo não mudaram e que as estimativas dos parâmetros β_0 e ϕ são mais fortemente afetadas que as dos demais parâmetros. Em relação aos critérios de comparação de modelos, vemos que estes critérios nos indicam que com a retirada da observação 13 o modelo se ajusta melhor aos dados. Assim podemos concluir que este ponto tem influência sobre as estimativas dos parâmetros do modelo.

Na Figura 6.5 temos o gráfico do $\widehat{CPO}$, notamos que com a retirada da observação 13, nenhuma outra observação aparece em destaque, isto nos fornece indícios de que todas as observações estão sendo bem ajustadas pelo modelo.

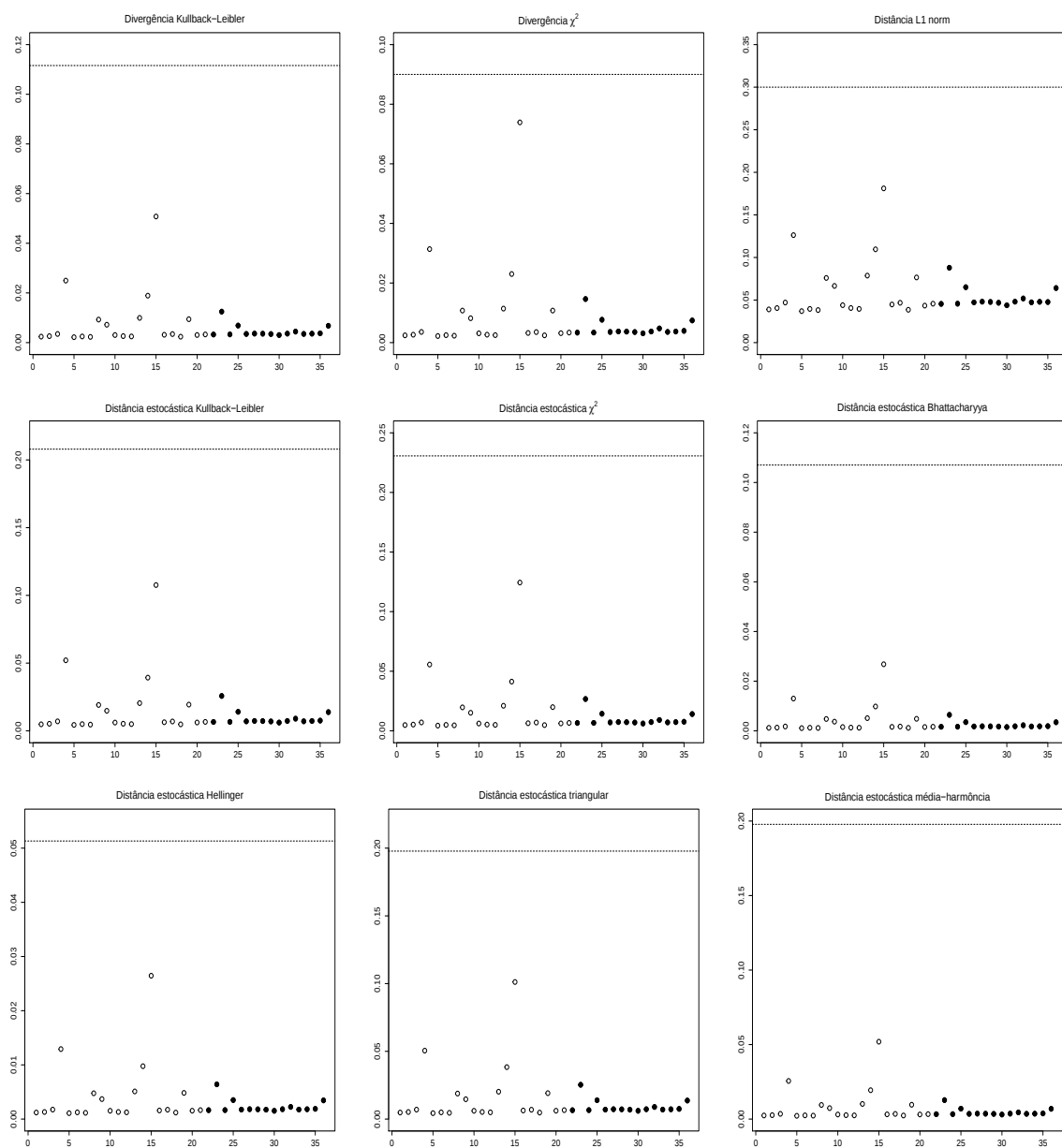
Figura 6.5: $\widehat{CPO}$ do modelo de regressão beta ajustado sem a observação 13.



Fonte: Dados próprios.

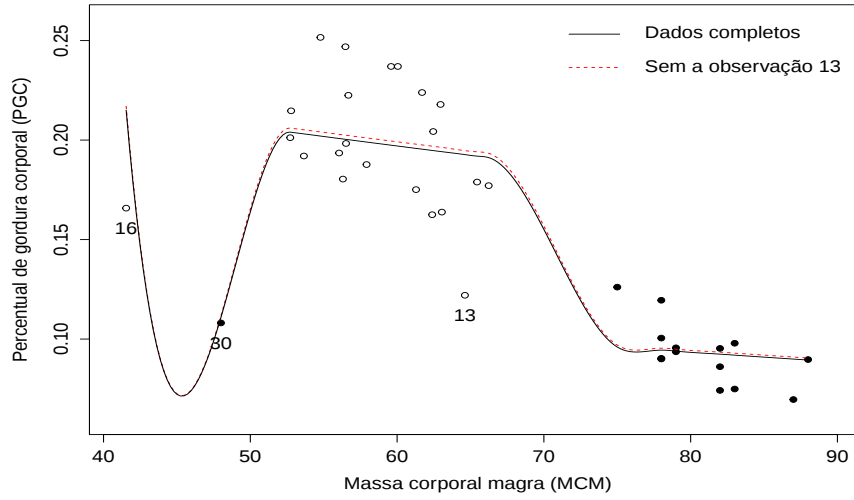
Na Figura 6.6 vemos os gráficos das medidas utilizadas na identificação de observações atípicas. Observamos nesta figura que com a retirada da observação 13 nenhuma divergência ou distância indica pontos atípicos. Na Figura 6.7 temos o gráfico de dispersão entre as variáveis PGC e MCM, com a linha de regressão da média da variável PGC do modelo de regressão beta ajustado com e sem a observação 13. Notamos por esta figura uma suave diferença entre as linhas de regressão da média da variável PGC. Contudo, como comentado anteriormente, com a retirada da observação 13 observamos significantes mudanças nas estimativas dos parâmetros β_0 e ϕ .

Figura 6.6: Gráficos das medidas utilizadas na identificação de observações atípicas no ajuste do modelo de regressão beta sem a observação 13.



Fonte: Dados próprios.

Figura 6.7: Gráfico de dispersão entre as variáveis PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta ajustado com e sem a observação 13.



Fonte: Dados próprios.

Bayes et al. (2012) indicam que o modelo de regressão beta retangular pode ser uma melhor alternativa ao modelo de regressão beta para situações em que há ocorrência de eventos extremos. Como a observação 13 é um evento extremo, ajustaremos a seguir o modelo de regressão beta retangular. Nosso intuito com isto é comparar os ajustes deste modelo com os modelos de regressão beta ajustados anteriormente.

6.2 Modelo de regressão beta retangular

Nesta seção, para estudar a relação entre a variável resposta PGC e as variáveis explicativas MCM e sexo, ajustaremos o modelo de regressão beta retangular proposto por Bayes et al. (2012). O modelo que ajustaremos é dado por

$$PGC_i \sim \mathcal{BRR}(\gamma_i, \phi, \alpha), \logit(\gamma_i) = \beta_0 + \beta_1 \times MCM_i + \beta_2 \times sexo_i, i = 1, 2, \dots, 37. (6.2)$$

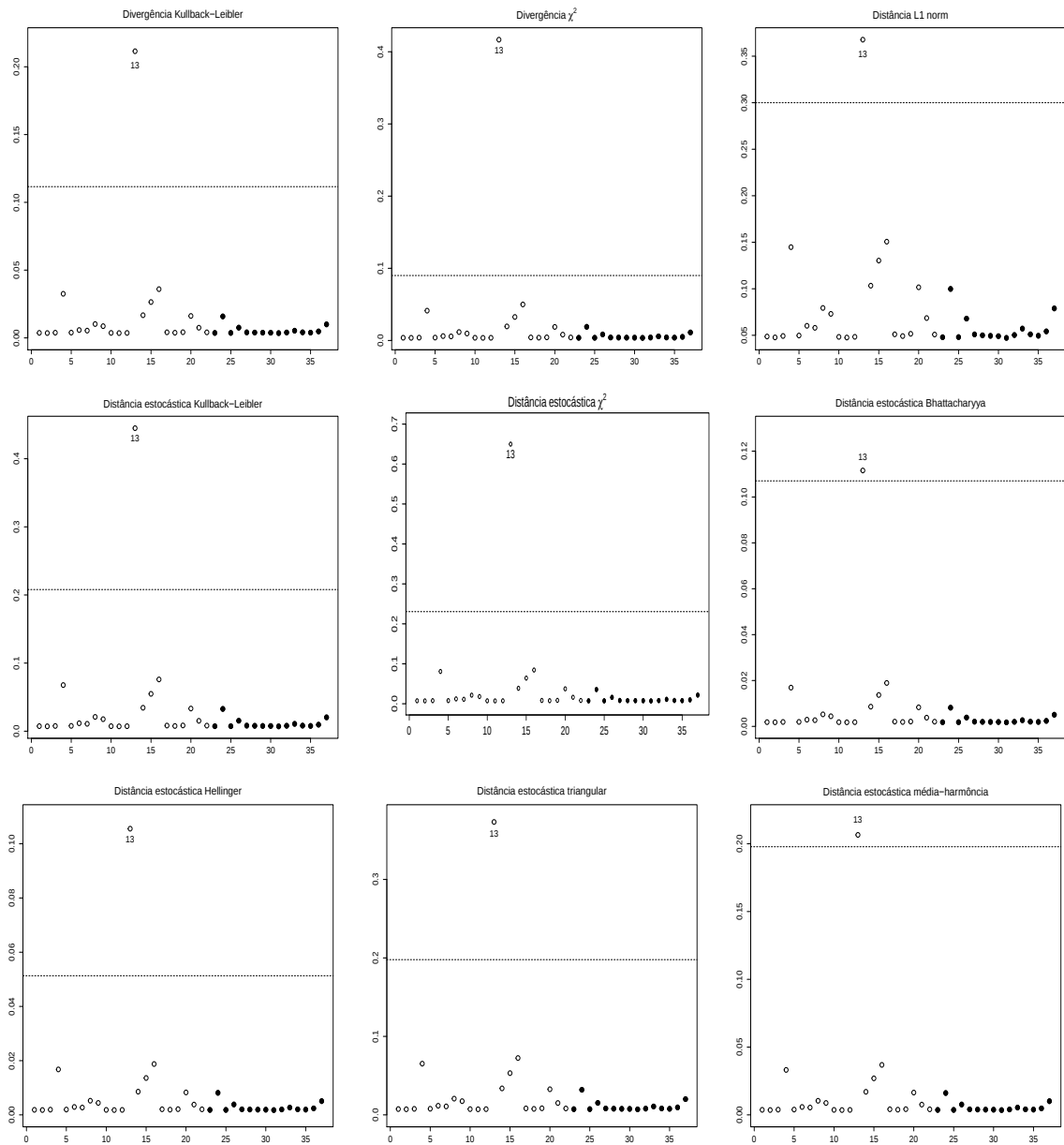
Para as distribuições *a priori* dos β_i 's, ϕ e α nos adotamos $\beta_i \sim \mathcal{N}(\mu_i^*, \sigma^2)$, $\phi \sim \mathcal{G}(0, 01; 0, 01)$ e $\alpha \sim \mathcal{U}(0; 1)$, conforme indicado pela literatura. A escolha dos hiperparâmetros das distribuições dos β_i 's foi feita de mesmo modo que nos ajustes dos modelos de regressão beta anteriores, assim nos adotamos, $\beta_0 \sim \mathcal{N}(-1, 02780; 0, 03294^2)$, $\beta_1 \sim \mathcal{N}(-0, 00544; 0, 00054^2)$ e $\beta_2 \sim \mathcal{N}(-0, 75266; 0, 02997^2)$.

Na Tabela 6.1 observamos as médias das distribuições *a posteriori* dos parâmetros do modelo ajustado bem como o intervalo de credibilidade de 95%. Pelos valores dos critérios de comparação de modelos, temos que o modelo de regressão beta ajustado com todas as observações é mais adequado para este caso que o modelo de regressão beta retangular

ajustado.

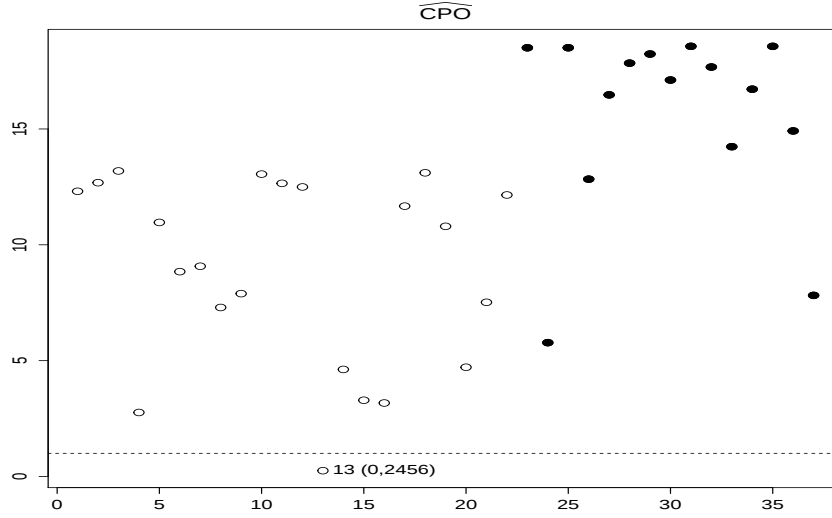
Na Figura 6.8 vemos os gráficos das medidas utilizadas na identificação de observações atípicas. Assim como no caso do ajuste do modelo beta, todas as medidas indicam apenas a observação 13 como atípica. Posteriormente ajustamos o modelo de regressão beta retangular sem esta observação.

Figura 6.8: Gráficos das medidas utilizadas na identificação de observações atípicas no ajuste do modelo de regressão beta retangular.



Fonte: Dados próprios.

Na Figura 6.9 temos os gráfico do $\widehat{CPO}$, podemos notar, assim como no caso do ajuste do modelo de regressão beta com todas as observações, que a observação 13 se destaca com o menor valor do $\widehat{CPO}$, isto nos fornece indícios de que esta observação não está sendo bem ajustada pelo modelo.

Figura 6.9: $\widehat{CPO}$ do modelo de regressão beta retangular ajustado.

Fonte: Dados próprios.

A seguir ajustaremos mais uma vez o modelo de regressão beta retangular, mas agora sem a observação 13. Com isso estamos interessados em verificar o impacto desta observação sobre o ajuste do modelo.

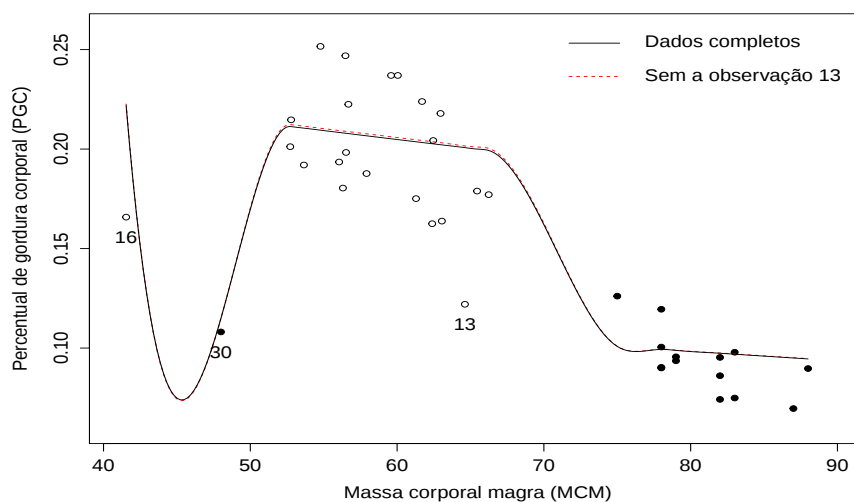
6.2.1 Modelo de regressão beta retangular sem a observação 13

Para as distribuições dos β'_i s, conforme feito anteriormente, adotamos, $\beta_0 \sim \mathcal{N}(-1,02514; 0,02133^2)$, $\beta_1 \sim \mathcal{N}(-0,00542; 0.00033^2)$ e $\beta_2 \sim \mathcal{N}(-0,75494; 0,02155^2)$. Assumimos $\phi \sim \mathcal{G}(0,01; 0,01)$ e $\alpha \sim \mathcal{U}(0; 1)$.

Na Tabela 6.1 temos as médias das distribuições *a posteriori* dos parâmetros do modelo ajustado e seus respectivos intervalos de 95% de credibilidade. Notamos que com a retirada da observação 13 as inferências com relação aos parâmetros do modelo não mudaram, e que as estimativas dos β'_i s são fracamente afetadas, apenas os parâmetros ϕ e α tem suas estimativas mais fortemente afetadas. Os critérios de comparação de modelos nos indicam que o modelo de regressão beta ajustado sem a observação 13 é mais indicado ao ajuste dos dados do que o modelo de regressão beta retangular ajustado também sem esta observação.

Na Figura 6.10 temos o gráfico de dispersão entre as PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta retangular ajustado com e sem a observação 13. Notamos nesta figura uma suave diferença entre as linhas de regressão da média da variável PGC. Contudo, como comentado anteriormente, com a retirada da observação 13 observamos significantes mudanças nas estimativas dos parâmetros α e ϕ .

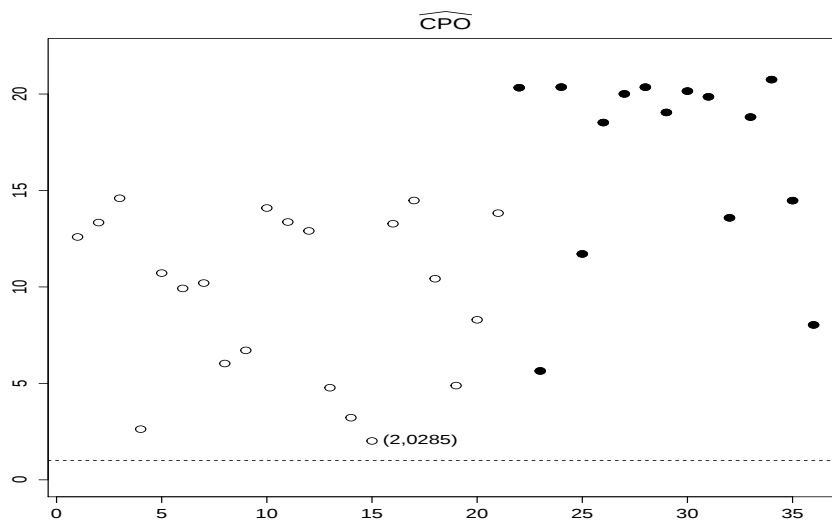
Figura 6.10: Gráfico de dispersão entre as PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta retangular ajustado com e sem a observação 13.



Fonte: Dados próprios.

Na Figura 6.11 temos o gráfico do $\widehat{CPO}$, observamos que não há nenhuma observação em destaque, isto nos fornece indícios de que todas as observações estão sendo bem ajustadas pelo modelo.

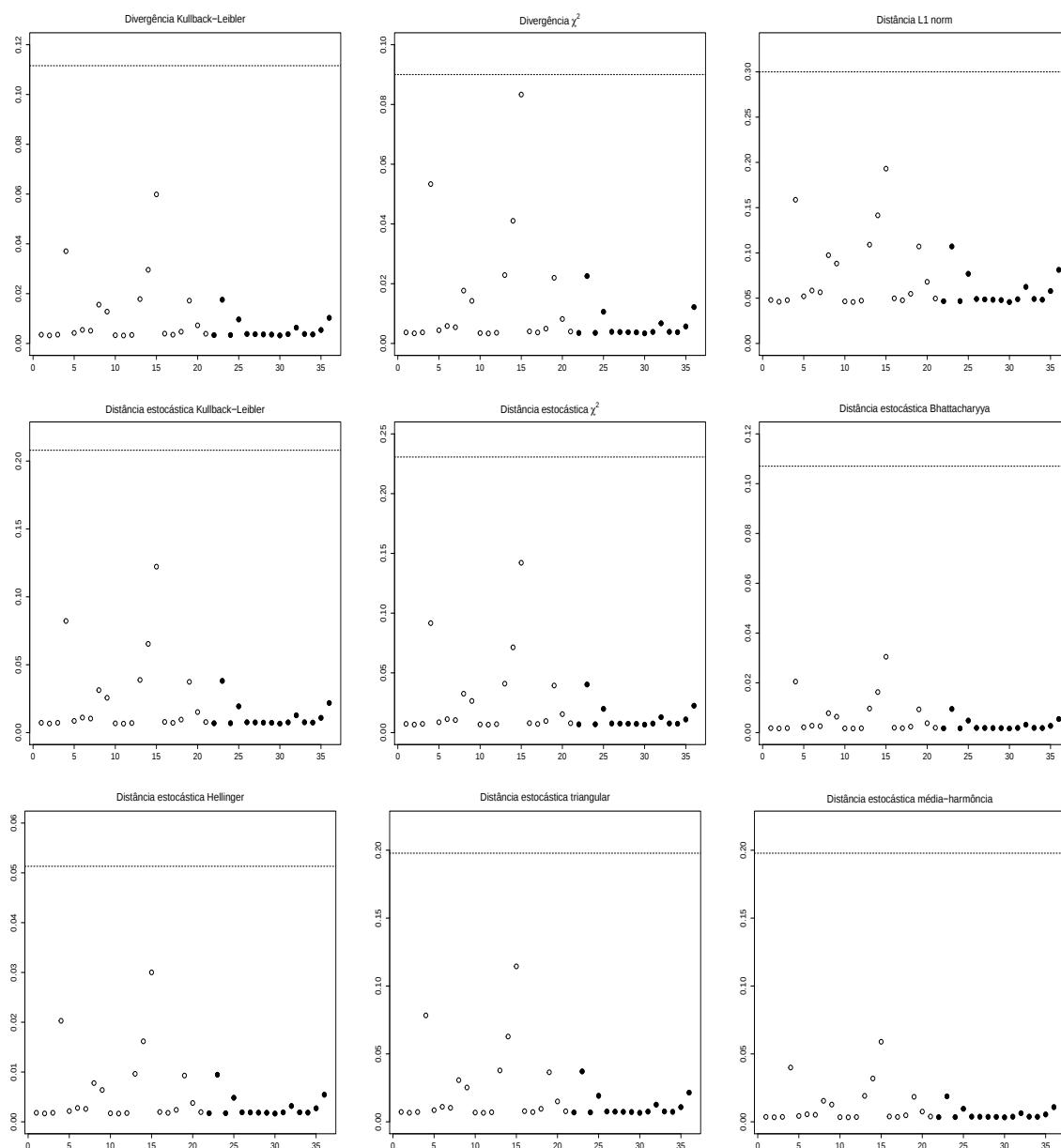
Figura 6.11: $\widehat{CPO}$ do modelo de regressão beta retangular ajustado sem a observação 13.



Fonte: Dados próprios.

Na Figura 6.12 vemos os gráficos das medidas utilizadas na identificação de observações atípicas. Assim como no caso do ajuste do modelo beta, nenhuma observação é indicada pelas medidas como sendo atípica.

Figura 6.12: Gráficos das medidas utilizadas na identificação de observações atípicas no ajuste do modelo de regressão beta retangular ajustado sem a observação 13.



Fonte: Dados próprios.

6.3 Conclusão

Em nossa aplicação verificamos pelos critérios de comparação de modelos que em ambos os casos, com todas as observações e sem a observação 13, o modelo de regressão beta é mais indicado que o modelo de regressão beta retangular. Contudo, as estimativas dos parâmetros do modelo de regressão beta retangular são menos afetados que os parâmetros do modelo de regressão beta, e a estimativa de α é próxima de zero em ambos os casos,

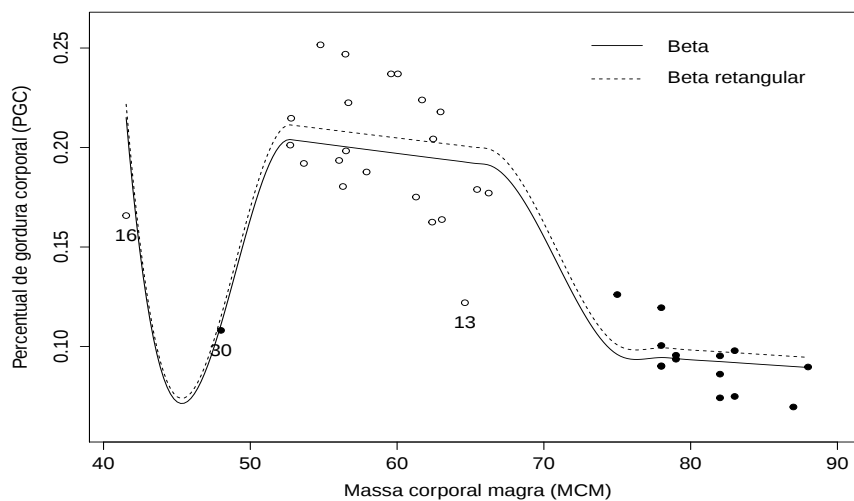
deste modo advogamos a favor do uso do modelo de regressão beta retangular para ajustar aos dados. Em ambos os modelos, observamos também que as inferências em relação aos seus parâmetros não se alteram com a retirada da observação 13. Notamos que as estimativas dos parâmetros β_0 e ϕ são significativamente afetadas quando ajustamos o modelo de regressão beta com a retirada da observação 13. No ajuste do modelo de regressão beta retangular não observamos significativas mudanças nas estimativas dos parâmetros $\beta's$, apenas nas estimativas dos parâmetros α e ϕ . Notamos nas Figuras 6.7 e 6.10, que com a retirada da observação 13 o ajuste do modelo de regressão beta retangular não sofre significativa alteração em relação ao modelo de regressão beta, notamos mais claramente a alteração causada pela retirada da observação 13 no ajuste do modelo de regressão beta. Em relação as divergências e distâncias utilizadas na detecção de observações atípicas, podemos dizer que tais medidas foram eficientes, visto que observamos que todas as medidas, no ajuste de ambos os modelos, indicam a observação 13 como atípica.

Na Tabela 6.1 observamos as médias das distribuições *a posteriori* dos parâmetros dos modelos de regressão beta e beta retangular ajustados com e sem a observação 13, bem como o intervalo de credibilidade de 95% e os critérios de comparação de modelos. Nas Figuras 6.13 e 6.14 temos os gráfico de dispersão entre as PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta e beta retangular ajustado com e sem a observação 13, respectivamente.

Tabela 6.1: Média *a posteriori* e intervalo de credibilidade 95% para os parâmetros dos modelos de regressão beta e beta retangular para os atletas de remo do conjunto de dados IAS usando todas as observações e sem a observação 13.

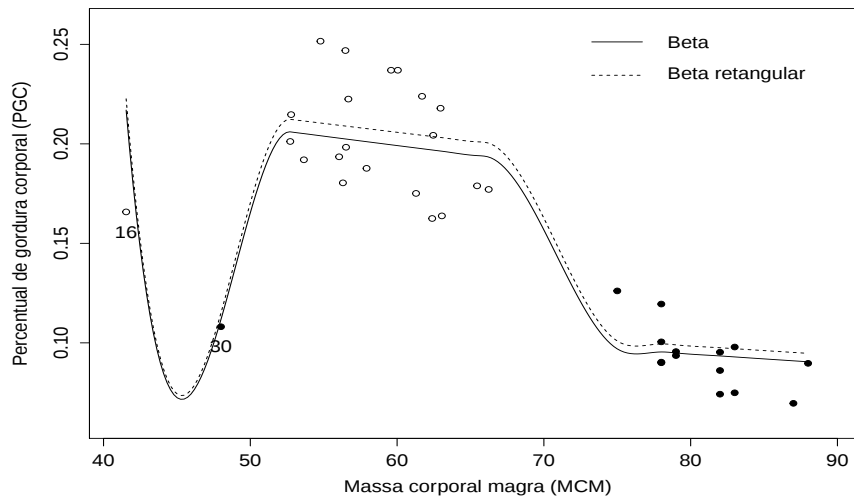
Observações	Parâmetros	Regressão beta			Regressão beta retangular		
		Média	2,5%	97,5%	Média	2,5%	97,5%
Todas	β_0	-1.04625	-1.08916	-1.00418	-1.02842	-1.08512	-0.96989
	β_1	-0.00598	-0.00664	-0.00530	-0.00546	-0.00638	-0.00454
	β_2	-0.74801	-0.80281	-0.69423	-0.74984	-0.80383	-0.69598
	ϕ	194.53601	117.63917	292.24876	197.35608	114.76043	302.89466
	α	-	-	-	0.06027	0.00285	0.15619
	LPML		84,50131			83,60112	
	$\widehat{DIC}$		-169,05313			-167,20000	
	$\widehat{EAIC}$		-162,72300			-159,20545	
	$\widehat{EBIC}$		-156,27932			-151,15083	
	$PsBF(\mathcal{B} \mathcal{BRR})$	2,260					
Sem a observação 13	β_0	-1.03972	-1.06527	-1.01377	-1.02536	-1.06385	-0.98703
	β_1	-0.00586	-0.00627	-0.00545	-0.00542	-0.00601	-0.00481
	β_2	-0.75191	-0.79335	-0.71194	-0.75548	-0.79542	-0.71622
	ϕ	242.52801	144.16676	369.13141	239.87716	141.13864	365.70748
	α	-	-	-	0.05012	0.00251	0.13066
	LPML		86,96935			86,02235	
	$\widehat{DIC}$		-173,45062			-171,51798	
	$\widehat{EAIC}$		-166,88150			-163,38013	
	$\widehat{EBIC}$		-160,54754			-155,46255	
	$PsBF(\mathcal{B} \mathcal{BRR})$	2,1195					

Figura 6.13: Gráfico de dispersão entre as PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta e beta retangular ajustado com todas as observações.



Fonte: Dados próprios.

Figura 6.14: Gráfico de dispersão entre as PGC e MCM com a linha de regressão da média da variável PGC para o modelo de regressão beta e beta retangular ajustado sem a observação 13.



Fonte: Dados próprios.

Considerações finais

Neste trabalho discutimos sobre modelos de regressão beta e beta retangular sob uma perspectiva Bayesiana, sobre alguns critérios de comparação de modelos e sobre o uso e algumas divergências e distâncias estocásticas. Como dito anteriormente nosso objetivo é avaliar o uso das divergências Kullback-Leibler e χ^2 , e das distâncias estocásticas Kullback-Leibler, χ^2 , Bhattacharyya, Hellinger, triangular e média-harmônica, e da distância L_1 norm na identificação de observações atípicas nos ajustes dos modelos de regressão beta e beta retangular. Para tal análise fizemos um estudo de simulação que, de modo geral, demonstrou ser eficiente o uso das citadas medidas para alcançarmos o objetivo proposto. Também observamos que o modelo de regressão beta retangular se mostrou mais apto que o modelo de regressão beta, para se trabalhar com dados atípicos. Percebemos que o modelo de regressão beta retangular é mais apropriado que o modelo de regressão beta, no caso em que temos a presença de observações atípicas. Notamos também que as medidas apresentam melhor desempenho na detecção de observações atípicas quando o modelo ajustado não corresponde a verdadeira distribuição geradora dos dados. Em destaque das demais medidas utilizadas, temos a divergência χ^2 , que se mostrou a melhor medida para detectar pontos atípicos. Por último apresentamos uma aplicação a fim de evidenciar a utilização das citadas medidas na indicação de observações atípicas nos ajustes dos modelos de regressão beta e beta retangular. Como proposta para trabalhos futuros podemos pensar na criação de novos pontos de corte.

Referências

- AITKIN, M. Posterior bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 111–142, 1991.
- BAYES, C. L.; BAZÁN, J. L.; GARCÍA, C. A new robust regression model for proportions. *Bayesian Analysis*, International Society for Bayesian Analysis, v. 7, n. 4, p. 841–866, 2012.
- BERGER, J. O.; PERICCHI, L. R. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, Taylor & Francis, v. 91, n. 433, p. 109–122, 1996.
- BOLFARINE, H.; BAZAN, J. L. Bayesian estimation of the logistic positive exponent irt model. *Journal of educational and behavioral statistics*, SAGE Publications, v. 35, n. 6, p. 693–713, 2010.
- BOX, G. E. Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, JSTOR, p. 383–430, 1980.
- BRANSCUM, A. J.; JOHNSON, W. O.; THURMOND, M. C. Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics*, Wiley Online Library, v. 49, n. 3, p. 287–301, 2007.
- BROOKS, S. *Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde*. NJ, USA: Wiley & Blackwell, 2002.
- BURBEA, J.; RAO, C. On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, IEEE, v. 28, n. 3, p. 489–495, 1982.

- CHO, H. et al. Bayesian case influence diagnostics for survival models. *Biometrics*, Wiley Online Library, v. 65, n. 1, p. 116–124, 2009.
- CONGDON, P. *Applied bayesian modelling*. West Sussex, England: John Wiley & Sons, 2014.
- CRIBARI-NETO, F.; ZEILEIS, A. Beta regression in R. *Journal of Applied Statistics*, v. 2, n. 34, p. 1–24, 2010.
- CSISZ, I. et al. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, v. 2, p. 299–318, 1967.
- DA-SILVA, C.; MIGON, H. S.; CORREIA, L. Dynamic bayesian beta models. *Computational Statistics & Data Analysis*, Elsevier, v. 55, n. 6, p. 2074–2089, 2011.
- DEY, D. K.; BIRMIWAL, L. R. Robust bayesian analysis using divergence measures. *Statistics & Probability Letters*, Elsevier, v. 20, n. 4, p. 287–294, 1994.
- EGUCHI, S.; COPAS, J. Interpreting kullback–leibler divergence with the neyman–pearson lemma. *Journal of Multivariate Analysis*, Elsevier, v. 97, n. 9, p. 2034–2040, 2006.
- ESPINHEIRA, P. L.; FERRARI, S. L.; CRIBARI-NETO, F. Influence diagnostics in beta regression. *Computational Statistics & Data Analysis*, Elsevier, v. 52, n. 9, p. 4417–4431, 2008.
- ESPINHEIRA, P. L.; FERRARI, S. L.; CRIBARI-NETO, F. On beta regression residuals. *Journal of Applied Statistics*, Taylor & Francis, v. 35, n. 4, p. 407–419, 2008.
- FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.
- FIGUEROA-ZÚÑIGA, J. I.; ARELLANO-VALLE, R. B.; FERRARI, S. L. Mixed beta regression: A bayesian perspective. *Computational Statistics & Data Analysis*, Elsevier, v. 61, p. 137–147, 2013.
- GAMERMAN, D.; LOPES, H. F. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Londres, England: Chapman & Hall, 2006.
- GARCÍA, C.; PÉREZ, J. G.; VAN DORP, J. R. Modeling heavy-tailed, skewed and peaked uncertainty phenomena with bounded support. *Statistical Methods & Applications*, Springer, v. 20, n. 4, p. 463–486, 2011.
- GEISSER, S. *On the prediction of observables: a selective update*. North-Holland, Amsterdam: University of Minnesota, School of Statistics, 1985.

- GEISSER, S.; EDDY, W. F. A predictive approach to model selection. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 74, n. 365, p. 153–160, 1979.
- GELFAND, A. E.; DEY, D. K.; CHANG, H. Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics*, v. 2, p. 147–167, 1992.
- GELFAND, A. E.; GHOSH, S. K. Model choice: a minimum posterior predictive loss approach. *Biometrika*, Biometrika Trust, v. 85, n. 1, p. 1–11, 1998.
- GELMAN, A. et al. *Bayesian data analysis*. New York, USA: Chapman & Hall, 2009.
- GIET, L.; LUBRANO, M. A minimum hellinger distance estimator for stochastic differential equations: An application to statistical inference for continuous time interest rate models. *Computational statistics & data analysis*, Elsevier, v. 52, n. 6, p. 2945–2965, 2008.
- GOUDAIL, F.; RÉFRÉGIER, P. Contrast definition for optical coherent polarimetric images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 26, n. 7, p. 947–951, 2004.
- HAHN, E. D. Mixture densities for project management activity times: A robust approach to pert. *European Journal of Operational Research*, Elsevier, v. 188, n. 2, p. 450–459, 2008.
- HUANG, X.; OOSTERLEE, C. W. Generalized beta regression models for random loss-given-default. *The Journal of Credit-Risk*, v. 7, p. 1–26, 2008.
- IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. Criterion-based methods for bayesian model assessment. *Statistica Sinica*, v. 11, n. 2, p. 419–444, 2001.
- IBRAHIM, J. G.; LAUD, P. W. A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 89, n. 425, p. 309–319, 1994.
- JEFFREYS, H. *Theory of probability*, Harold Jeffreys. England: Clarendon. Oxford. GB, 1961. 459 p.
- KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical modelling*, SAGE Publications, v. 3, n. 3, p. 193–213, 2003.
- KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *The annals of mathematical statistics*, JSTOR, p. 79–86, 1951.
- LIESE, F.; VAJDA, I. *Convex statistical distances*. Teubner, 1987.

- MAHALANOBIS, P. C. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, v. 2, p. 49–55, 1936.
- MARKATOU, M. Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, Wiley Online Library, v. 56, n. 2, p. 483–486, 2000.
- NASCIMENTO, A. D.; CINTRA, R. J.; FRERY, A. C. Hypothesis testing in speckled data with stochastic distances. *Geoscience and Remote Sensing, IEEE Transactions on*, IEEE, v. 48, n. 1, p. 373–385, 2010.
- O'HAGAN, A. Fractional bayes factors for model comparison (with discussion). *J Roy Statist Soc Ser B*, v. 57, p. 99–138, 1995.
- O'HAGAN, A. Properties of intrinsic and fractional bayes factors. *Test*, Springer, v. 6, n. 1, p. 101–118, 1997.
- OSPINA, R.; FERRARI, S. L. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, Elsevier, v. 56, n. 6, p. 1609–1623, 2012.
- PAOLINO, P. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, SPM-PMSAPSA, v. 9, n. 4, p. 325–346, 2001.
- PENG, F.; DEY, D. K. Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics*, Wiley Online Library, v. 23, n. 2, p. 199–213, 1995.
- POLIDORO, M. J. F. P. *Metodologia bayesiana e adequação de modelos*. 150 p. Tese (Doutorado) — Universidade de Lisboa, 2013.
- SALICRÚ, M. et al. On the applications of divergence type measures in testing statistical hypotheses. *Journal of Multivariate Analysis*, Elsevier, v. 51, n. 2, p. 372–391, 1994.
- SEGHOUANE, A.-K.; AMARI, S.-I. The aic criterion and symmetrizing the kullback–leibler divergence. *Neural Networks, IEEE Transactions on*, IEEE, v. 18, n. 1, p. 97–106, 2007.
- SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, p. 379–423, 1948.
- SIMAS, A. B.; BARRETO-SOUZA, W.; ROCHA, A. V. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, Elsevier, v. 54, n. 2, p. 348–366, 2010.

- SMITHSON, M.; VERKUILEN, J. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, American Psychological Association, v. 11, n. 1, p. 54, 2006.
- SOUZA, D. *Regressão beta multivariada com aplicações em pequenas áreas*. Tese (Doutorado) — Tese de doutorado em estatística, Instituto de Matemática-Universidade Federal do Rio de Janeiro. 3, 2011.
- SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002.
- STONE, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 111–147, 1974.
- VAJDA, I. *Theory of statistical inference and information*. [S.l.]: Kluwer Academic Pub, 1989.
- WEISSMAN, T. et al. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.