MARIANA RESENDE

PROCEDIMENTO PARA IDENTIFICAR OUTLIERS POR MEIO DA DISTRIBUIÇÃO ACUMULADA DE MÍNIMO EM UM MODELO COM RESPOSTA GAMA

LAVRAS - MG

2016

MARIANA RESENDE

PROCEDIMENTO PARA IDENTIFICAR OUTLIERS POR MEIO DA DISTRIBUIÇÃO ACUMULADA DE MÍNIMO EM UM MODELO COM RESPOSTA GAMA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Orientador

Dr. Marcelo Ângelo Cirillo

LAVRAS - MG

2016

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Resende, Mariana.
Procedimento para identificar outliers por meio da distribuição acumulada de mínimo em um modelo com resposta gama / Mariana Resende. – Lavras : UFLA, 2016.
50 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2016.
Orientador: Marcelo Ângelo Cirillo.
Bibliografia.

1. Simulação. 2. Distância de Mahalanobis. 3. Taxa de mistura. 4. Dispersão. 5. GLM. I. Universidade Federal de Lavras. II. Título.

MARIANA RESENDE

PROCEDIMENTO PARA IDENTIFICAR OUTLIERS POR MEIO DA DISTRIBUIÇÃO ACUMULADA DE MÍNIMO EM UM MODELO COM RESPOSTA GAMA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 24 de fevereiro de 2016.

Dr. Joel Augusto Muniz UFLA

Dra. Carla Regina Guimarães Brighenti UFSJ

Dr. Marcelo Ângelo Cirillo
Orientador

LAVRAS - MG

2016

AGRADECIMENTOS

A Deus por ter me oferecido o dom da vida, por ter me regido, me iluminado e dado forças para concretização desse trabalho.

Aos meus pais, Abel e Bete, pelo amor, paciência, incentivo e apoio.

Aos meus irmãos Daniel e João Marcelo, que mesmo em silêncio sempre me apoiaram e acreditaram em mim.

Ao meu namorado, Diones, por todo amor e carinho.

Aos professores do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, pelos ensinamentos, apoio e compreensão nos momentos em que mais precisei. Em particular ao meu orientador, professor Marcelo Ângelo Cirillo, pelas orientações, ensinamentos e paciência para confecção deste e outros trabalhos. Obrigada pela compreensão constante.

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas pela oportunidade concedida para realização desse mestrado.

Ao CNPq, pela bolsa de estudos concedida, fundamental para essa conquista.

RESUMO

Esse trabalho teve por objetivo propor um procedimento fundamentado na distribuição acumulada de mínimos para identificar outliers em modelos generalizados com resposta Gama. Para validar essa metodologia utilizou-se simulação Monte Carlo, na qual considerou-se os cenários definidos pela combinação de diferentes tamanhos amostrais, taxa de contaminação e distribuições com diferentes graus de assimetria. Nesse contexto, probabilidades referentes a erros de classificação e acurácia foram obtidas em 500 realizações Monte Carlo. Conclui-se que o método é eficiente por apresentar elevadas probabilidades de acurácia. Em se tratando da aplicação, por meio do exemplo ilustrado, dada a similaridade entre a nova abordagem proposta em comparação às abordagens fundamentadas pela matriz de alavanca e distância de Cook, conclui-se que o procedimento sugerido nesse trabalho é factível de ser aplicado em respostas que envolvam a distribuição Gama.

Palavras-chave: Simulação. Distância de Mahalanobis. Taxa de mistura. Dispersão. GLM.

ABSTRACT

This study aimed to propose a procedure based on the accumulated distribution of minimums to identify generalized outlier models by using Gamma response. To validate this methodology, we used Monte Carlo simulation, considering the scenarios defined by the combination of different sample sizes, the contamination rate and the distributions with different degrees of asymmetry. In this context, probabilities related to classification and accuracy errors were obtained from 500 Monte Carlo achievements. We concluded that the method is effective for presenting high accuracy probability. In terms of implementation, through the illustrated example, given the similarity between the new proposed approach, compared to approaches based by the lever matrix and Cook's distance, we conclude that the procedure suggested in this study is feasible for implementation in response to Gamma distribution.

Keywords: Simulation. Mahalanobis distance. Mixing rate. Dispersion. GLM.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Um fato que ocorre com frequência em experimentos é a presença de observações outliers. A presença dessas observações interfere nas conclusões obtidas por meio de inferências relacionadas à qualidade de ajuste e significância das estimativas dos parâmetros. A questão que envolve maiores estudos refere-se ao fato do tratamento dessas observações como, por exemplo, detectar a presença de outliers e como lidar com eles, uma vez identificado (BARBATO et al., 2011).

Em geral a pesquisa de observações de outliers é feita por inspeção gráfica, sendo o exemplo mais clássico o gráfico *boxplot*, cuja construção em síntese, considera um conjunto de dados ordenados, assumindo uma distribuição simétrica independente do tamanho da amostra. Por meio desse gráfico, identificam-se os outliers acima ou abaixo das "linhas" externas à "caixa", previamente definidos por limites que contemplam medidas quartílicas.

Em se tratando da discriminação dos outliers feita por meio dos *boxplots*, a interpretação poderá conduzir o leitor a conclusões errôneas, uma vez que, os comprimentos das linhas externas correspondem ao efeito de cada distribuição; portanto, naturalmente, os limites de especificação proporcionarão amplitudes diferentes. Desta forma, a interpretação do *boxplot* para dados assimétricos requer certa cautela do pesquisador. Diante do exposto, enquadra-se a distribuição Gama, sendo essa, apropriada para modelar dados de natureza contínua e que mostram assimetrias (TURKMAN; SILVA, 2000).

Outra aplicação dessa distribuição dá-se no ajuste de processos estocásticos associados com o tempo, em particular, aqueles relativos às precipitações meteorológicas e aos estudos envolvendo tempos de vida de componentes (JOHNSON; KOTZ; BALAKRISHNAN, 1994).

Barbato et al. (2011) propuseram um método fundamentado na distribuição de máximos e mínimos, dos quais, limites de especificação foram determinados para discriminar outliers em distribuições assimétricas, associando uma medida de risco que mensure a probabilidade de acerto da discriminação.

Convém ressaltar que os autores não abordaram o uso de covariáveis com o formalismo de um modelo linear generalizado. Mediante o exposto, este trabalho teve por objetivo propor um procedimento para identificar outliers em modelos generalizados com resposta Gama, utilizando distribuição acumulada de mínimos.

2 REFERENCIAL TEÓRICO

2.1 Métodos de detecção de outliers

Em geral, um outlier é visto como um dado observacional destoante em relação às demais observações. Para averiguar essa discrepância, utilizam-se inúmeros métodos estatísticos, sendo os mais usuais descritos a seguir.

2.1.1 Análise exploratória de dados: estudo de outliers por meio do gráfico *boxplot*

Originado por Tukey (1977) o *boxplot*, popularmente conhecido como gráfico de caixa, tornou-se a técnica visual padrão para resumir as principais medidas estatísticas e identificar outliers. O método é fundamentado em um conjunto de dados ordenados, considerando cinco medidas, sendo elas: a mediana, primeiro quartil (Q1), terceiro quartil (Q3) e os limites superior e inferior. O limite inferior é computado como Q1-1,5(IQR), e o limite superior Q3 + 1,5(IQR), sendo o IQR o intervalo interquartil dado pela diferença entre o terceiro e o primeiro quartil (Figura 1). Tendo como referência esses limites, caso alguma observação apresente um valor não condizente com essa especificação suspeita-se que a mesma poderá ser classificada como outlier; sendo este, gerado por algum erro de mensuração ou efeito de cauda, caso a distribuição seja assimétrica (BARBATO et al., 2011; POTER et al., 2006).

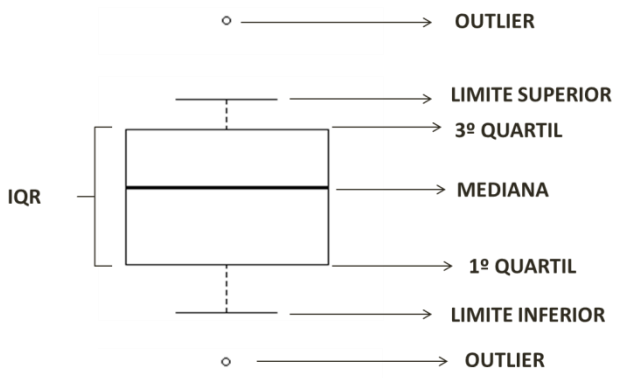

Figura 1 Layout das medidas estatísticas que compõem a estruturação do gráfico *boxplot*

Segundo mencionado por Barbato et al. (2011), o gráfico *boxplot* é um método bastante simples e robusto devido à baixa sensibilidade à distorção decorrente a outliers, uma vez que apenas a parte central da distribuição é considerada para calcular IQR. Além disso, não há restrições em relação ao tamanho amostral, enquanto outros métodos inferenciais apresentam algumas restrições.

Em decorrência desses problemas, modificações foram propostas como uma melhoria para o *boxplot* padrão. Podemos citar os intervalos de confiança em torno da mediana, a incorporação da densidade de informação, além dos cinco parâmetros descritivos no *boxplot* padrão. Outras soluções incluem um *histplot*, em que a função densidade de probabilidade subjacente é estimada na mediana e o *boxpercentile*, onde as informações sobre a distribuição empírica dos dados cumulativos são usadas em conjunto com o *boxplot* sendo, portanto, uma combinação de um *boxplot* com o percentil.

Embora o *boxplot* de Tukey seja uma ferramenta amplamente utilizada para detecção de outliers, ele pode ser modificado para melhorar o desempenho. Shevlyakov et al. (2013) propuseram uma modificação do *boxplot* de Tukey para

melhorar o seu desempenho através da incorporação de uma estimativa robusta de escala altamente eficiente, mantendo a estrutura simples do *boxplot*.

Essa estimativa de escala é baseada na visualização das áreas de cauda e presença de anomalias nos dados. Neste caso, o intervalo interquartil (IQR) não é a melhor escolha uma vez que sua robustez pode ser melhorada consideravelmente.

Robustez de uma estimativa é medida pelo ponto de ruptura do erro total sendo $0 \le \varepsilon^* \le 0{,}5$, que é a maior fração de erros totais (anomalias) nos dados que ainda mantêm a tendência de um estimador limitado. Considerando que o alcance interquartil apresenta um valor moderado de ponto de ruptura $\varepsilon^* = 0{,}25$ e o desvio absoluto médio $MAD_n x = med_i \mid x_i - med\, x \mid$ tem o valor máximo $\varepsilon^* = 0{,}5$ (SHEVLYAKOV et al., 2013).

Uma vez que o intervalo interquartil é menos resistente a outliers do que o desvio médio absoluto $MAD_n x$, uma regra de construção mais robusta para os extremos *boxplot* pode ser dada por:

$$x_L = \max\left\{ x_{(1)}, LQ - k_{MAD} MAD_n \right\},$$
$$x_U = \max\left\{ x_{(n)}, UQ + k_{MAD} MAD_n \right\},$$

onde k_{MAD} é um coeficiente de limiar escolhido a partir de considerações adicionais.

Embora o desvio absoluto médio $MAD_n x$ seja uma estimativa de escala muito robusta com o valor máximo do ponto de ruptura $\varepsilon^* = 0{,}5$, a sua eficiência é de apenas 0,37 na distribuição normal (SHEVLYAKOV et al., 2013).

Diante disso, outras estimativas robustas de escala altamente eficientes foram propostas, mas algumas apesar de apresentarem alta eficiência e ponto de ruptura de $\varepsilon^* = 0{,}5$ demandam grande tempo de execução computacional. Dentre essas estimativas eficientes, a FQ_n, função de influência da estimativa Q_n (quartil inferior das diferenças absolutas entre pares) apresenta destaque, uma vez que necessita de pouco tempo de execução computacional e apresentam valor máximo do ponto de ruptura $\varepsilon^* = 0{,}5$ e sua eficiência é de 0,81.

Com base na estimativa de escala robusta altamente eficiente FQ_n, Shevlyakov et al. (2013) propuseram uma nova regra para os extremos *boxplot* definidos como:

$$x_L = \max\left\{ x_{(1)}, LQ - k_{FQ} FQ_n \right\},$$
$$x_U = \max\left\{ x_{(n)}, UQ + k_{FQ} FQ_n \right\},$$

A estatística FQ_n é consistente sob a suposição de normalidade quando é multiplicada pela constante de 1,483.

2.1.2 Estudo de outliers utilizando distribuições de valores extremos

Barbato et al. (2011) descrevem métodos para detectar outliers considerando distribuições acumuladas de valores extremos máximo (F_{max}) e mínimo (F_{min}). O entendimento dessas distribuições dar-se-á no conceito de estatística de ordem, em que, pressupõe-se uma amostra aleatória x_1, x_2,...,x_n ordenada, cuja notação a ser utilizada para considerar a ordenação é dada em (1).

$$
\begin{aligned}
&X_{(1)} = \text{menor em } \{X_1, X_2, ..., X_n\} \\
&X_{(2)} = \text{segundo menor em } \{X_1, X_2, ..., X_n\} \\
&... \\
&X_{(j)} = \text{j - ésimo em } \{X_1, X_2, ..., X_n\} \\
&... \\
&X_{(n)} = \text{maior em } \{X_1, X_2, ..., X_n\}
\end{aligned} \tag{1}
$$

De um modo geral, considera-se a representação de amostra aleatória ordenada representada por $X_{(1)}, X_{(2)}, ..., X_{(n)}$, proveniente de uma distribuição $F_X(x)$. Com essa especificação, Casella e Berger (2002) deduzem em (2) resultando em (3).

$$F_{X_{(j)}}(x) = P[X_{(j)} \leq x] = P[Y \geq j] = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k} \tag{2}$$

$$F_{X_{(n)}}(x) = \sum_{k=n}^{n} \binom{n}{n} [F_X(x)]^n [1 - F_X(x)]^{n-n} = [F_X(x)]^n \tag{3}$$

De forma análoga, tem-se que $F_X(1)$ dado em (4) como principal resultado obtido em (5).

$$
\begin{aligned}
F_{X_{(1)}}(x) &= \sum_{k=1}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k} \\
&= \sum_{k=0}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k} - \binom{n}{0} [F_X(x)]^0 [1 - F_X(x)]^{n-0} \\
&= [F_X(x) + 1 - F_X(x)]^n - [1 - | F_X(x)]^n
\end{aligned} \tag{4}
$$

$$F_{X_{(1)}}(x) = 1 - [1 - F_X(x)]^n \tag{5}$$

Logo, procede considerar que as distribuições de mínimo e máximos por Barbato et al. (2011) correspondem a

$$F_{Min}(x) = 1 - \left[1 - F(x)\right]^n, \;\; F_{Max}(x) = \left[F(x)\right]^n \tag{6}$$

Seguindo essas especificações, os autores sugerem algumas distribuições de F_{max} e F_{min}, conforme descreve a Tabela 1.

Tabela 1 Distribuição de F_{Min} e F_{Max} para algumas distribuições padronizadas

Tipo	Limites	F_x	$F_{Min}(x)$	$F_{Max}(x)$
U-shaped	$-\sqrt{2} < x < \sqrt{2}$	$\frac{1}{\pi} arcsen \frac{x}{\sqrt{2}} + \frac{1}{2}$	$1 - \left[\frac{1}{2} - \frac{1}{\pi} arcsen \frac{x}{\sqrt{2}}\right]^n$	$\left[\frac{1}{\pi} arcsen \frac{x}{\sqrt{2}} + \frac{1}{2}\right]^n$
Uniforme	$-\sqrt{3} < x < \sqrt{3}$	$\frac{1}{2\sqrt{3}} x + \frac{1}{2}$	$1 - \left[\frac{1}{2} - \frac{1}{2\sqrt{3}} x\right]^n$	$\left[\frac{1}{2\sqrt{3}} x + \frac{1}{2}\right]^n$
Triangular	$-\sqrt{6} < x < \sqrt{6}$	$\frac{1}{12} x^2 - \frac{\text{sgn}(x)}{\sqrt{6}} x + \frac{1}{2}$	$1 - \left[-\frac{1}{12} x^2 + \frac{\text{sgn}(x)}{\sqrt{6}} x + \frac{1}{2}\right]^n$	$\left[\frac{1}{12} x^2 - \frac{\text{sgn}(x)}{\sqrt{6}} x + \frac{1}{2}\right]^n$

Em conexão a essas distribuições Barbato et al. (2011) propuseram a mensuração de um risco de rejeitar como outlier um valor pertencente à região inferior definida por α_L e, de forma semelhante, α_U para a região superior. Para isso, os limites inferior e superior, respectivamente dados em L_L e L_U, são especificados de forma que:

$$\alpha_L = 1 - \left[1 - F(L_L)\right]^n, \;\; 1 - \alpha_U = \left[F(L_U)\right]^n \tag{7}$$

Os valores para esses limites L_L e L_U podem facilmente ser obtidos por aproximações numéricas, obtidas em função da distribuição amostral.

Por uma questão de simplicidade, assume-se que o risco bilateral α é simétrico, portanto, $\alpha_L = \alpha_U = \alpha/2$, os limites inferior e superior podem ser facilmente expressas em termos de α (Tabela 2).

Tabela 2 Valores dos limites superior (L_U) e inferior (L_L) expressos em função de α (risco de erro aceitável) com os limites relevantes

Tipo	Limite Inferior L_L		Limite Superior L_U	
	Condição	Limite Inferior	Condição	Limite Superior
U-shaped	$\sqrt{2}sen\left[\pi\left(\frac{1}{2}-\frac{1}{n}\sqrt{1-\frac{\alpha}{2}}\right)\right]$	$-\sqrt{2}$	$-\sqrt{2}sen\left[\pi\left(\frac{1}{2}-\frac{1}{n}\sqrt{1-\frac{\alpha}{2}}\right)\right]$	$\sqrt{2}$
Uniforme	$2\sqrt{3}\left(\frac{1}{2}-\frac{1}{n}\sqrt{1-\frac{\alpha}{2}}\right)$	$-\sqrt{3}$	$-2\sqrt{3}\left(\frac{1}{2}-\frac{1}{n}\sqrt{1-\frac{\alpha}{2}}\right)$	$\sqrt{3}$
Triangular	$-\sqrt{6}+2\sqrt{3}\sqrt{1-\left(1-\frac{\alpha}{2}\right)^{1/n}}$	$-\sqrt{6}$	$\sqrt{6}-2\sqrt{3}\sqrt{1-\left(1-\frac{\alpha}{2}\right)^{1/n}}$	$\sqrt{6}$

2.2 Distribuição Gama

Seja y a realização de uma variável aleatória distribuída por uma $Gama(\mu,\delta)$ com função densidade definida por

$$f(y \mid \mu,\delta) = \frac{1}{\Gamma(\mu)\delta^{\mu}} y^{\mu-1} e^{-y/\delta}, \mu > 0; \delta > 0 \tag{8}$$

μ e δ, nomeados respectivamente por parâmetro de forma e escala. $\Gamma(\mu)$ corresponde à função Gama definida por

$$\Gamma(\mu) = \int_0^{\infty} y^{\mu-1} e^{-y} dy \tag{9}$$

A integração por partes de $\Gamma(\mu)$ permite verificar que (ROSS, 2003):

$$\begin{aligned}\Gamma(\mu) &= -e^{-y}y^{\mu-1}\Big|_0^{\infty} + \int_0^{\infty} e^{-y}(\mu-1)y^{\mu-2}dy \\ &= (\mu-1)\int_0^{\infty} e^{-y}y^{\mu-2}dy \\ &= (\mu-1)\Gamma(\mu-1)\end{aligned} \tag{10}$$

Para valores inteiros de μ, exemplificando por $\mu = n$, por meio de um processo recursivo podemos concluir

$$\begin{aligned}\Gamma(n) &= (n-1)\Gamma(n-1) \\ &= (n-1)(n-2)\Gamma(n-2) \\ &= \ldots = (n-1)(n-2)\ldots 3.2\Gamma(1)\end{aligned}$$

Logo, substituindo $\Gamma(1)$em (10), uma vez que, $\Gamma(1) = \int_0^{\infty} e^{-y}dy = 1$ temos que $\Gamma(n) = (n-1)!$.

Decorrente à expressão (1), nota-se que ao integrar o núcleo de uma função de densidade de probabilidade sabemos que, para qualquer $\delta, \mu > 0$ temos

$$\int_0^{\infty} y^{\mu-1}e^{-y/\delta}dy = \Gamma(\mu)\delta^{\mu}, \tag{11}$$

Logo, desenvolvendo $E(Y)$ temos

$$
\begin{aligned}
E(Y) &= \frac{1}{\Gamma(\mu)\delta^{\mu}} \int_0^{\infty} y y^{\mu-1} e^{-y/\delta} dy \\
E(Y) &= \frac{1}{\Gamma(\mu)\delta^{\mu}} \Gamma(\mu+1)\delta^{\mu+1} \\
E(Y) &= \frac{\mu\Gamma(\mu)\delta}{\Gamma(\mu)} \\
E(Y) &= \mu\delta
\end{aligned}
\tag{12}
$$

De forma análoga, a variância da distribuição Gama (μ, δ). Em particular, ao calcular $E(Y)^2$ lidamos com o núcleo de uma distribuição Gama $(\mu+2, \delta)$,cujo resultado é $Var(Y) = \mu\delta^2$, facilmente verificado em Casella e Berger (2002).

Outras parametrizações da distribuição Gama são encontradas na literatura, resumidas na Tabela 3.

Tabela 3 Algumas parametrizações da distribuição Gama

Autor	Parametrização	Densidade $f(y \mid \mu, \delta)$ e Verossimilhança $L((\alpha, \lambda) \mid y)$	$E(y)$ e $Var(y)$
Rizzo (2008)	$Y \sim G(\alpha, \lambda)$ $\alpha; \lambda = \frac{1}{\beta}$	$f(y \mid \alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\lambda y\}$ $L((\alpha, \lambda) \mid y) = \left(\frac{\lambda^{\alpha}}{\Gamma(\alpha)}\right)^n \prod_{i=1}^{n} y_i^{\alpha-1} \exp\{-\lambda y_i\}$	$E(y) = \frac{\alpha}{\lambda}$ $Var(y) = \frac{\alpha}{\lambda^2}$
Aitkin (2010)	$Y \sim G(\alpha, \mu)$ $\alpha; \lambda = \alpha\beta$	$f(y \mid \alpha, \mu) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)\mu^{\alpha}} y^{\alpha-1} \exp\{-\alpha y / \mu\}$ $L((\alpha, \mu) \mid y) = \left(\frac{\alpha^{\alpha}}{\Gamma(\alpha)\mu^{\alpha}}\right)^n \prod_{i=1}^{n} y_i^{\alpha-1} \exp\{-\alpha y_i / \mu\}$	$E(y) = \mu$ $Var(y) = \frac{\mu^2}{\alpha}$

Segundo menciona Ribeiro Júnior et al. (2011 p. 72), a parametrização sugerida por Aitkin (2010) apresenta como vantagem a ortogonalidade na matriz de informação entre *r* e μ. Usualmente μ refere-se ao parâmetro de interesse para inferências e *r* o parâmetro de perturbação, ou seja, "nuisance". Em uma modelagem estatística, na qual se pressupõe um modelo de regressão para a média μ, como por exemplo, modelos lineares generalizados, essa parametrização poderá ser recomendável.

2.2.1 Distribuições resultantes da Gama

A distribuição Gama apresenta alguns casos particulares, de acordo com os valores assumidos para seus parâmetros de forma e escala, ela se reduz a outras distribuições (TURKMAN; SILVA, 2000):

A distribuição Qui - quadrado, umas das mais usadas em processos de inferência estatística, é um caso particular da distribuição Gama, para μ = υ/2 e δ = 2 a distribuição Gama (μ, δ) se reduz a uma distribuição χ2 (υ), dada por:

$$f(y) = \frac{1}{\Gamma\left(\frac{\upsilon}{2}\right)2^{\upsilon/2}} y^{\frac{\upsilon}{2}-1} e^{-y/2}$$

A distribuição Erlang (μ,δ) é uma distribuição de probabilidade contínua, muito utilizada em processos estocásticos e também se reduz à distribuição Gama (μ,δ) para μ inteiro.

$$f(y) = \frac{\delta^{\mu}}{(\mu-1)!} y^{\mu-1} e^{-y/\delta}$$

Outra distribuição com relação estreita com a distribuição Gama é a distribuição exponencial, uma vez que adotando μ = 1, a distribuição Gama (μ,δ) se reduz a uma distribuição Exponencial (δ):

$$f(y)=\frac{1}{\delta}e^{-y/\delta}$$

2.3 Modelo Generalizado Gama

Basicamente, um modelo linear generalizado (G.L.M.) passa pela especificação de três componentes descrita a seguir:

a) Componente aleatória – consiste em um conjunto de *n* observações, que são identicamente distribuídas, com distribuições univariadas diferentes, porém pertencentes à família exponencial, mencionada por Paulino e Singer (2006).

$$f(y_{i|}\,|\,\alpha_i,\phi)=\exp\left[\frac{y_i\alpha_i-b(\alpha_i)}{a(\phi)}\right] \tag{13}$$

Com funções apropriadas, $a(.), b(.)$ e $h(.)$. Importante enfatizar que, ϕ é denominado parâmetro de dispersão comum à distribuição de todas as observações.

b) Componente sistemática – corresponde a uma combinação linear especificada pelo pesquisador, a título de ilustração, considerando *p* variáveis preditoras e *n* observações, temos:

$\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p;\forall\in\{1,\ldots,n\}$

c) Função de ligação: indica uma função $g(.)$ monótona e diferenciável, ligando cada média com as variáveis explicativas por meio de um preditor, descrito por:

$$g(\mu) = \eta = X\beta \tag{14}$$

Dada a ênfase da distribuição Gama a ser utilizada nesse trabalho, a conexão entre as três componentes, na estimação dos parâmetros de um G.L.M. com resposta Gama, é ilustrada por Azzalini (1996) ao considerar $Y \sim Gamd(w, w/\mu_i)$, sendo uma *w*, e seu valor médio μ_i. A função densidade $f(y)$ é dada a seguir:

$$\begin{aligned} f(y) &= \frac{1}{\Gamma(\phi)}\left(\frac{\phi y}{\mu}\right)^{\phi} \exp\left(\frac{-\phi y}{\mu}\right)\frac{1}{y} \\ f(y) &= \exp\left\{\phi\left(-\frac{y}{\mu} - \ln\mu\right) - \ln\Gamma(\phi) + \phi\ln(\phi y) - \ln(y)\right\} \\ f(y) &= \exp\{\phi(\theta y + \ln(-\theta)) + c(y,\phi)\} \end{aligned} \tag{15}$$

para $y > 0$, sendo $\theta = -\frac{1}{\mu}$ e $c(y,\phi) = -\ln\Gamma(\phi) + \phi\ln(\phi y) - \ln(y)$

A função de ligação canônica é definida por:

$$-\frac{1}{\mu_i} = \theta_i = x_i^T\beta \tag{16}$$

e neste caso

$$l(\beta) = \exp\left\{\phi\left[\left(\sum_i x_i y_i\right)^T \beta + \sum_i \ln\left(x_i^T \beta\right)\right] + \sum_i c\left(y_i \phi\right)\right\} \tag{17}$$

Confirmando que $\sum_i x_i y_i$, é uma estatística suficiente sendo um valor conhecido $(\phi = 1)$ implica em uma distribuição exponencial. Caso o valor de *w* seja não conhecido, então $\sum_i x_i y_i$, é um componente das estatísticas suficientes mínimas $\left(\sum_i x_i y_i, \sum_i \ln y_i\right)$ (AZZALINI, 1996).

Por considerar a função de ligação canônica, alguns resultados são factíveis de serem deduzidos. Nesse contexto, mencionamos a matriz de informação de Fisher, definida por:

$$\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} = \frac{\phi_i x_{ij}}{\phi} \frac{\partial \mu_i}{\partial \beta_k} \tag{18}$$

que não depende das observações. Portanto, temos que,

$$\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} = E\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right] \tag{19}$$

A título de ilustração, consideremos um exemplo encontrado em Azzalini (1996) que descreve as equações necessárias para obtenção da matriz

de informação de Fisher, assumindo y_i um valor obtido a partir de uma amostra Y_i, variável aleatória exponencial com parâmetro de escala ρ_i, de tal forma que:

$$\ln \rho_i = \alpha + \beta x_i \qquad (i = 1,...,n) \tag{20}$$

em que x_i são constantes conhecidas não sendo todas iguais, e α e β são parâmetros reais a serem estimados, sem perda de generalidade, assumiremos $\sum x_i = 0$, sob independência das observações, o log da verossimilhança é

$$\begin{aligned} \ln(\alpha, \beta) &= \sum_{i=1}^{n} \left\{ \alpha + \beta x_i - y_i e^{(\alpha + \beta x_i)} \right\} \\ &= n\alpha - e^{\alpha} \sum y_i e^{\beta x_i} \end{aligned} \tag{21}$$

Para maximização da verossimilhança, considere

$$\frac{\partial l}{\partial \alpha} = n - e^{\alpha} \sum y_i e^{\beta x_i} \tag{22}$$

$$\frac{\partial l}{\partial \beta} = -e^{\alpha} \sum x_i y_i e^{\beta x_i} \tag{23}$$

Fazendo $\frac{\partial l}{\partial \alpha} = 0$, resulta em

$$\hat{\alpha}(\beta) = \ln\left(\frac{n}{\sum y_i e^{\hat{\beta} x_i}} \right) \tag{24}$$

Em relação à solução da segunda equação com a substituição de $\alpha(\hat{\beta})$, temos

$$g(\hat{\beta}) = \sum_i x_i y_i e^{\hat{\beta} x_i} = 0 \tag{25}$$

A matriz Hessiana da log-verossimilhança é estimada por

$$I(\beta^*) = \begin{pmatrix} -e^{\hat{\alpha}} \sum y_i e^{\hat{\beta} x_i} & -e^{\hat{\alpha}} \sum x_i y_i e^{\hat{\beta} x_i} \\ -e^{\hat{\alpha}} \sum x_i y_i e^{\hat{\beta} x_i} & -e^{\hat{\alpha}} \sum x_i^2 y_i e^{\hat{\beta} x_i} \end{pmatrix}$$

Cujos elementos diagonais são negativos, e o determinante é

$$e^{2\hat{\alpha}} \left\{ \sum y_i e^{\hat{\beta} x_i} \sum x_i^2 y_i e^{\hat{\beta} x_i} - \left(\sum x_i y_i e^{\hat{\beta} x_i} \right)^2 \right\}$$

2.4 Técnicas de diagnósticos de outliers em modelos generalizados com resposta Gama

Reportando de Paula (2011), a especificação de uma amostra aleatória $Y_i \sim G(\mu_i, \phi)$, ou seja, observações provenientes de uma Gama com médias diferentes, porém, o mesmo parâmetro de dispersão e considerando $g(\mu_i) = \eta$, sendo, $\eta_i = \mathbf{X_i^T \beta}$, onde $\mathbf{X_i} = (x_{i1}, ..., x_{ip})^T$ e $\beta = (\beta_1, ..., \beta_p)^T$, o vetor paramétrico a ser estimado por um processo iterativo, tendo como exemplo, o mínimo quadrado reponderado, especificado por:

$$\beta^{(m+1)} = \left(X^T W^{(m)} X \right)^{-1} X^T W^{(m)} Z^{(m)} \tag{26}$$

Sendo $m = 0,1,...$ o indexador que identifica a variável dependente modificado $Z = \boldsymbol{\eta} + \mathbf{W}^{-1/2}\mathbf{V}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$ a cada iteração, sendo, $\boldsymbol{\eta} = (\eta_1,...,\eta_n)^Y$; $\mathbf{y} = (y_1,...,y_n)^Y$; $\boldsymbol{\mu} = (\mu_1,...,\mu_n)^Y$; $V = diag(\mu_1,...,\mu_n)$ e $W = diag(w_1,...,w_n)$ com $w_i = \left(\frac{d\mu_i}{d\eta_i}\right) \Big/ \mu_i$. Seguindo essas especificações e considerando a função de ligação recíproca, cabe ressaltar que, os elementos da matriz **V** e **W** são dados por $V(Y_i/\eta_i) = \phi\mu_i$ e $w_i = \mu_i^2/\mu_i^2 = 1$.

Decorrente a essa definição, as técnicas de diagnóstico utilizadas para discriminar os outliers são fundamentadas nos componentes do desvio padronizado, obtidos por:

$$t_{D_i} = \frac{\pm\sqrt{2\hat{\phi}}}{\sqrt{1-\hat{h}_{ii}}}\left\{\log\left(\hat{\mu}_i \Big/ y_i\right) - (y_i - \hat{\mu}_i)\Big/\hat{\mu}_i\right\}^{1/2} \tag{27}$$

Sendo, $y_i > 0$ e h_{ii} o i-ésimo elemento da diagonal principal da matriz $H = W^{1/2}X\left(X^T W X\right)H^{-1}X^T W^{1/2}$. Paula (2011) evidencia que em particular quando há um intercepto em η_i, t_{D_i} é simplificado em:

$$t_{D_i} = \frac{\pm\sqrt{2\hat{\phi}}}{\sqrt{1-\hat{h}_{ii}}}\left\{\log\left(\hat{\mu}_i \Big/ y_i\right)\right\}^{1/2} \tag{28}$$

Outra medida alternativa é definida em relação à exclusão da i-ésima observação ao considerar a distância de Cook, cuja expressão é dada em (29).

$$L_{D_i} = \frac{\hat{\phi}\hat{h}_{ii}}{\left(1-\hat{h}_{ii}\right)^2}\frac{\left(y_i-\hat{\mu}_i\right)^2}{\hat{\mu}_i} \tag{29}$$

Por fim, gráficos descritivos utilizando as medidas (27) e (28) versus $\hat{h}_{ii}$ são usualmente confeccionadas para discriminação de outliers.

3 METODOLOGIA

Em concordância com os objetivos propostos, o procedimento computacional utilizado para validar os resultados foi feito nas seguintes etapas: 3.1 Simulação de amostras provenientes de modelo Gama com resposta contaminada e 3.2 Incorporação da distribuição acumulada de mínimo na discriminação de outliers.

3.1 Simulação de amostras provenientes de modelo Gama com resposta contaminada

O modelo Gama contaminado foi gerado inicialmente considerando o termo linear definido por $\eta_i=\beta_{1i}x_{1i}+\beta_{2i}x_{2i}$, para X_1 e X_2 ~ Uniforme (0,1), sendo, i=1,...,n, em que n correspondeu ao tamanho amostral, fixado nos valores de n=15, 30 e 50. Seguindo essas especificações, considerou-se a função de ligação recíproca (30).

$$g^{-1}(\eta_i)=\frac{1}{\mu_i}; \text{ dado } g(\mu_i)=\eta_i \tag{30}$$

A contaminação da amostra por observações outliers foi feita pela geração da distribuição H(y), utilizando a regra (31).

$$H(y)=(1-\lambda)\times G(\mu_i,\delta)+\lambda\times G(.);\ 0\leq\lambda\leq 1 \tag{31}$$

Sendo G(.) as distribuições das observações outliers, com as densidades representadas na Figura 2.

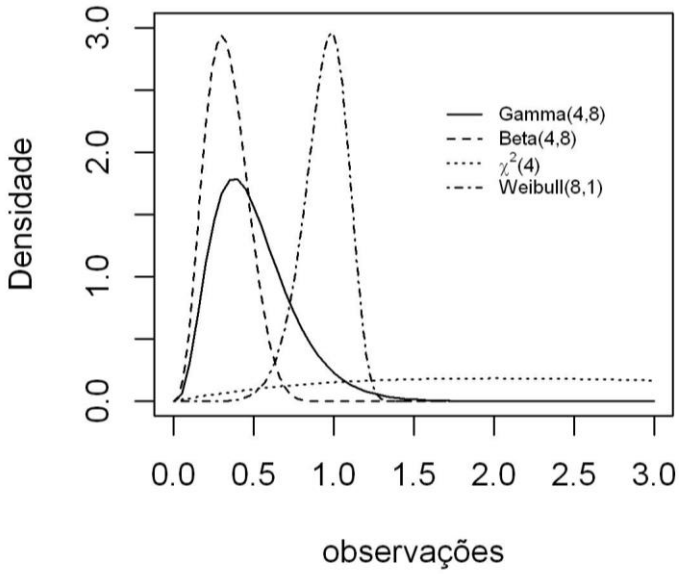

Figura 2 Densidades das observações outliers geradas por simulação Monte Carlo

As taxas de contaminação foram especificadas em λ = 0,05; 0,10 e 0,20 e o parâmetro de escala foi arbitrariamente definido em (32), uma vez na geração das amostras considerou-se a dispersão entre as variáveis explicativas do modelo.

$$\delta = \frac{1}{\sqrt{\det\left(X'X\right)}} \tag{32}$$

em que, X indicou a matriz de delineamento com as variáveis descritas em η_i

Após a geração das amostras, procedeu-se com a obtenção das estimativas de máxima verossimilhança para os parâmetros do modelo, bem como, o parâmetro de dispersão, obtido pelo método dos momentos e diretamente pela relação da variância das observações geradas, representada em

var(h_i) e função de variância $V(\mu_i)=\mu_i$. Convém ressaltar que, por se tratar de uma observação proveniente de um modelo Gama contaminado, essa relação é dada de forma aproximada, conforme seguem os resultados descritos em (33).

$$\begin{aligned} &\operatorname{var}\left(h_i\right) \approx \phi^{-1} V\left(\mu_i\right) \\ &\phi \approx \frac{\mu_i^2}{\operatorname{var}(h_i)} \rightarrow \hat{\phi} \approx \frac{1}{\left[CV(h_i)\right]^2}, \end{aligned} \tag{33}$$

Sendo, CV(h_i) o coeficiente de variação das observações geradas pelo modelo Gama contaminado. Com esse procedimento, estimativas do parâmetro de dispersão estimado em (33) pelo método dos momentos, denotado por $\hat{\varphi}$. Assim, ambas as estimativas foram confrontadas e verificadas a similaridade, pode-se então validar o procedimento de simulação em 500 amostras de simulação Monte Carlo.

3.2 Incorporação da distribuição acumulada de mínimos na discriminação de outliers

Com o propósito de adequar a distribuição acumulada de mínimos em relação aos valores preditos pelo modelo, utilizando a distribuição acumulada obtida a partir da estatística de ordem conforme menciona Barbato et al. (2011), a distribuição acumulada de mínimos foi dada em (34).

$$H_{min}(\hat{\mu}_i) = 1 - \left[1 - H(\hat{\mu}_i)\right]^n \tag{34}$$

A representação gráfica das observações supostamente classificadas como outliers foi feita utilizando a distância de Mahalanobis, tendo como referência a matriz de covariância estruturada em (35).

$$S = \begin{bmatrix} \text{var}(\hat{\mu}_i) & \text{cov}(\hat{\mu}_i, H_{min}(\hat{\mu}_i)) \\ \text{cov}(\hat{\mu}_i, H_{min}(\hat{\mu}_i)) & \text{var}(H_{min}(\hat{\mu}_i)) \end{bmatrix}, \text{ logo} \qquad (35)$$

A distância de Mahalanobis para cada observação foi obtida por D_h (36)

$$D_h^2 = \sqrt{\left(\underset{\sim}{H}_{min}(\hat{\mu}_i) - \underset{\sim}{\hat{\mu}}_i\right)^T S^{-1} \left(\underset{\sim}{H}_{min}(\hat{\mu}_i) - \underset{\sim}{\hat{\mu}}_i\right)}, \text{em que}$$
$$\underset{\sim}{H}_{min}(\hat{\mu}_i) = \left(H_{min}(\hat{\mu}_1), ..., H_{min}(\hat{\mu}_n)\right)^T \; e \; \underset{\sim}{\hat{\mu}}_i = \left(\hat{\mu}_1, ..., \hat{\mu}_n\right)^T, \qquad (36)$$

Portanto, plotando D_h^2 versus $H_{min}(\hat{\mu}_i)$ dado que $D_h^2 \sim \chi^2_{p=2;1-\alpha}$ tornou-se possível identificar os pontos localizados acima desse quantil como observações discrepantes, sendo possível computar as frequências de observações e outliers classificados incorretamente (Tabela 4).

Tabela 4 Frequências das observações e outliers classificadas corretamente e incorretamente

	Classificação		Total
	Observação	Outlier	
Observação	A	b	$n_1 = a + b$
Outlier	C	d	$n_2 = c + d$

De acordo com as frequências, procedeu-se com um estudo de falso positivo e negativo, de modo que, as probabilidades obtidas, descritas na

Tabela 5, fossem interpretadas conjuntamente como medidas de desempenho do método proposto.

Tabela 5 Probabilidades referente ao erro de classificação e acurácia das observações classificadas por meio da distância de Mahalanobis computada em função dos vetores $\underset{\sim}{H}_{min}(\hat{\mu}_i)$ *e* $\underset{\sim}{\hat{\mu}}_i = (\hat{\mu}_1, ..., \hat{\mu}_n)^T$

Medida	Interpretação
$E_1 = \frac{b}{b+a}$	Erro 1: probabilidade condicional de uma observação ser classificada como outlier, dado que ela é uma observação original.
$E_2 = \frac{c}{c+d}$	Erro 2: probabilidade condicional de uma observação ser classificada como observação original, dado que ela é um outlier.
$A = \frac{a+d}{n_1+n_2}$	Acurácia: probabilidade de acerto

Fonte: Ramos (2013).

Após a obtenção dos resultados, para fins de comparação, realizou-se uma aplicação a dados reais, considerando a identificação dos outliers e observações discrepantes utilizando a matriz de alavanca e distância de Cook em comparação à metodologia proposta que concerne na distância de Mahalanobis calculada entre a distribuição acumulada de mínimos e valores preditos pelo modelo Gama contaminado.

Para obtenção dos resultados simulados e aplicação aos dados, elaboraram-se rotinas computacionais no software R Core Team (2014).

4 RESULTADOS E DISCUSSÃO

4.1 Distribuições com domínio (0,∞)

Tendo como referência as amostras geradas via Monte Carlo, a eficiência do método proposto para identificar outliers considerando a distribuição acumulada de mínimos e os valores preditos obtidos em modelo com diferentes taxas de contaminação são resumidos na Tabela 6, na qual encontram-se as distribuições cujo domínio é igual ao domínio da distribuição Gama.

Tabela 6 Medidas de desempenho do método proposto para identificar observações outliers

n = 15					
Distribuição	λ	Qtde de simulações válidas	E_1	E_2	A
Gama (4,8)	0,05	300	0,1190	0,6210	0,8494
	0,10	300	0,0899	0,8046	0,8321
	0,20	200	0,0647	0,8186	0,7778
χ^2_4	0,05	200	0,1178	0,8317	0,8421
	0,10	300	0,0928	0,9075	0,8250
	0,20	200	0,1058	0,8999	0,7350
Weibull (8,1)	0,05	200	0,0811	0,9182	0,8710
	0,10	500	0,1318	0,8709	0,7937
	0,20	500	0,1163	0,9045	0,7260
n = 30					
Distribuição	λ	Qtde de simulações válidas	E_1	E_2	A
Gama (4,8)	0,05	500	0,0751	0,6750	**0,8919**
	0,10	500	0,0556	0,8833	**0,8549**
	0,20	500	0,0575	0,9452	**0,7654**
χ^2_4	0,05	500	0,0972	**0,8726**	0,8599
	0,10	500	0,0558	**0,9322**	0,8557
	0,20	500	0,0611	**0,9568**	0,7583
Weibull (8,1)	0,05	500	0,0882	0,8907	0,8673
	0,10	500	0,0515	0,7191	0,8796
	0,20	500	0,0736	0,9599	0,7459
n = 50					
Distribuição	λ	Qtde de simulações válidas	E_1	E_2	A
Gama (4,8)	0,05	500	0,0613	0,9432	0,8895
	0,10	500	0,0504	0,9318	0,8579
	0,20	500	0,0444	0,9401	0,7754
χ^2_4	0,05	500	0,0913	0,8747	0,8679
	0,10	500	0,0465	0,9244	0,8639
	0,20	500	0,0921	0,8816	0,2478
Weibull (8,1)	0,05	500	0,0368	0,9588	0,9152
	0,10	500	0,0566	0,9582	0,8516
	0,20	500	0,0529	0,9412	0,7682

Por meio da Tabela 6, inicialmente pode-se observar que ao considerar amostras menores (n=15), o efeito dos outliers gerados por meio da distribuição Gama (4,8) e χ_4^2 produziu amostras inválidas, por não convergirem para as estimativas de máxima verossimilhança. Tal fato refere-se mais ao desempenho numérico aplicado à distribuição Gama, o qual é corroborado por Noufaily e Jones (2013) ao mencionar que computacionalmente as estimativas de máxima verossimilhança, além de serem sensíveis à parametrização, apresenta situações de distribuições pertencentes à família exponencial e uma simples perturbação no parâmetro de forma ou escala resultara em verossimilhanças não côncavas, ou em um máximo local (COORAY; ANANDA, 2008; SONG, 2008). Nas situações nas quais as perturbações causadas pelos outliers resultam em distribuições assimétricas ou multimodais Wiper, Insua e Ruggeri (2001) recomendam que as misturas sejam feitas por distribuições pertencentes à família exponencial.

Diante desse problema, a discussão dos resultados em relação à eficiência do método proposto nesse trabalho para identificar outliers foi feita considerando as situações em que as amostras apresentaram convergência nas 500 realizações Monte Carlo. Desta forma, tal discussão restringiu-se aos tamanhos amostrais n=30 e 50. Assim, em ambas as situações, aumentando a taxa de contaminação (λ) e o tamanho amostral (n) para a situação na qual os outliers foram gerados pela distribuição Gama (4,8), nota-se que as probabilidades relacionadas ao E_2, ou seja, o outlier a ser classificado como uma observação original, considerando as flutuações devido ao processo Monte Carlo, mantiveram-se praticamente constantes.

Naturalmente, com o aumento da taxa de contaminação (λ=0,20), a acurácia correspondente ao acerto global foi reduzida. Entretanto, ao comparar com a distribuição Qui-quadrado, o aumento da taxa de contaminação repercutiu em uma forte redução, proporcionando uma estimativa incoerente.

Supostamente, essa acurácia pôde ser explicada pela distribuição Qui-quadrado ser mais assimétrica em relação à Gama (4,8), isso implica um excesso de curtose, na qual os resultados obtidos evidenciaram uma redução na acurácia.

Em se tratando dos outliers terem sido gerados pela Weibull (4,8), percebeu-se que as probabilidades referentes ao erro E_1, isto é, uma observação original ser classificada como outlier apresentou menores probabilidades, ao comparar o aumento do tamanho amostral. Em relação ao erro E_2, dado um n=50, as probabilidades mantiveram-se próximas independente da taxa de contaminação de outliers, porém, uma redução na acurácia foi identificada.

Considerando as afirmações dadas por Cousineau e Chartier (2010), de que a influência dos outliers é mais importante em relação ao tamanho amostral, podemos ressaltar que, o aumento do tamanho amostral proporcionou probabilidades do E_2 bem próximas, quando os outliers são provenientes da distribuição Qui-quadrado, sendo essa com efeito de curtose mais pronunciado. No caso da acurácia, as probabilidades foram próximas ao considerar outliers gerados pela Gama (4,8).

Em se tratando da estimativa do parâmetro de dispersão, obtido no ajuste do modelo Gama contaminado, os resultados descritos na Tabela 7 evidenciaram considerar a estimativa do parâmetro de dispersão obtida pela relação entre a função de variância e coeficiente de variação $\hat{\phi}$. Assim, nota-se que, o aumento da taxa de contaminação não afetou as estimativas, uma vez que os valores de $\hat{\varphi}$ foram similares. Contudo, ao considerar a estimativa obtida pelo método dos momentos $\hat{\varphi}$ observou-se que, na maioria dos casos, as estimativas foram reduzidas conforme a taxa de contaminação foi incrementada; o que de certa forma era esperado, uma vez que os outliers influenciam a qualidade de ajuste do modelo.

Convém ressaltar que em todas as situações, consideraram-se as estimativas do parâmetro de dispersão constantes, ou seja, a dispersão não foi

modelada. Nesse contexto, uma alternativa que poderia ser utilizada é verificada por meio da inferência bayesiana que considera os outliers gerados por uma distribuição normal (MARRON; WAND, 1992; RICHARDSON; GREEN, 1997; ROEDER; WASSERMAN, 1997).

Com base nessas referências, entende-se que a estimativa do parâmetro de dispersão obtida com o ajuste do modelo de fato é influenciada pela distribuição na qual os outliers foram gerados.

Em se tratando da distribuição acumulada dos mínimos em relação aos valores preditos do modelo, observou-se em todos os cenários avaliados a existência de uma correlação negativa. A importância desse resultado é justificada por utilizar a distância de Mahalanobis, que considera a informação das estatísticas de ordem e valores preditos na identificação de outliers.

Tabela 7 Efeito dos outliers na estimativa do parâmetro de dispersão no ajuste do modelo Gama contaminado e correlação entre a média ajustada e distribuição acumulada de mínimos

Distribuição	n	λ	$\text{cor}\left(\hat{\mu}_i, H_{min}(\hat{\mu}_i)\right)$	$\hat{\varphi}$	$\hat{\phi}$
Gama (4,8)	15	0,05	-0,818	11,309	8,075
		0,10	-0,682	10,217	4,623
		0,20	-0,623	11,645	3,742
	30	0,05	-0,753	9,663	7,148
		0,10	-0,763	10,909	5,366
		0,20	-0,625	9,190	4,408
	50	0,05	-0,542	8,413	4,723
		0,10	-0,757	7,787	6,999
		0,20	-0,616	8,365	6,097
χ^2_4	15	0,05	-0,743	11,495	4,929
		0,10	-0,893	11,528	6,198
		0,20	-0,768	12,779	4,477
	30	0,05	-0,762	8,981	5,431
		0,10	-0,623	9,276	1,700
		0,20	-0,633	10,994	1,204
	50	0,05	-0,736	8,270	6,992
		0,10	-0,637	8,500	0,709
		0,20	-0,620	8,635	2,183
Weibull (8,1)	15	0,05	-0,666	12,146	6,684
		0,10	-0,823	9,639	6,248
		0,20	-0,771	9,343	6,102
	30	0,05	-0,615	9,566	7,752
		0,10	-0,853	9,172	12,845
		0,20	-0,754	10,700	7,650
	50	0,05	-0,840	9,841	12,061
		0,10	-0,801	9,935	11,209
		0,20	-0,775	8,916	13,860

4.2 Distribuições com domínio (0,1)

Seguindo a mesma metodologia computacional para validar o procedimento de identificação das observações outliers utilizando a distribuição acumulada dos mínimos e os valores preditos obtidos por um modelo Gama

contaminado, para todos os cenários resultantes da combinação dada pela taxa de contaminação e tamanho amostral, foram simuladas amostras geradas pela distribuição Beta (4,8) (Tabela8).

Tabela 8 Medidas de desempenho do método proposto para identificar observações outliers com amostras geradas pela distribuição Beta (4,8)

n = 15					
Distribuição	Λ	Qtde de amostras válidas	E_1	E_2	A
Beta (4,8)	0,05	200	0,0926	0,6667	0,8683
	0,10	300	0,0769	0,7651	0,8430
	0,20	300	0,0679	0,7412	0,7918
n = 30					
Distribuição	Λ	Qtde de amostras válidas	E_1	E_2	A
Beta (4,8)	0,05	500	0,0878	0,4114	0,8925
	0,10	500	0,0776	0,6581	0,8559
	0,20	500	0,0719	0,8651	0,7693
n = 50					
Distribuição	Λ	Qtde de amostras válidas	E_1	E_2	A
Beta (4,8)	0,05	500	0,0485	0,9237	0,9031
	0,10	500	0,0581	0,6955	0,8741
	0,20	500	0,0202	0,9739	0,7823

Os resultados evidenciados na Tabela 8 indicam que ao considerar o tamanho amostral n=15, observou-se menor número de experimentos que não convergiram, supostamente, tal fato ocorreu por gerar observações próximas a zero, implicando em uma singularidade numérica ocasionada pelo método Newton – Raphson ao computar a matriz Hessiana no procedimento iterativo. Contudo, para amostras maiores, as probabilidades referentes ao desempenho do

teste foram similares, bem como, os resultados relacionados às estimativas do parâmetro de dispersão e correlação entre a distribuição acumulada de mínimos e valores preditos, conforme descreve a Tabela 9.

Tabela 9 Efeito dos outliers na estimativa do parâmetro de dispersão no ajuste do modelo Gama contaminado com observações geradas pela Beta (4,8) e correlação entre a média ajustada e distribuição acumulada de mínimos

n	Λ	$\mathrm{cor}\left(\hat{\mu}_i, H_{min}(\hat{\mu}_i)\right)$	$\hat{\varphi}$	$\hat{\phi}$
15	0,05	-0,792	9,916	4,145
	0,10	-0,697	10,649	4,293
	0,20	-0,706	11,451	3,640
30	0,05	-0,799	9,026	7,426
	0,10	-0,788	8,468	5,022
	0,20	-0,700	8,377	3,857
50	0,05	-0,735	8,540	5,932
	0,10	-0,695	9,225	6,746
	0,20	-0,555	8,379	4,716

4.3 Aplicação

Com o propósito de ilustrar a metodologia proposta, consideraram-se os dados descritos na Tabela 10, referentes aos tempos de sobrevivência em semanas de 17 pacientes com leucemia aguda. Para cada paciente registrou-se a contagem de glóbulos brancos (WBC) com seus correspondentes logaritmos na base 10.

Tabela 10 Tempo de sobrevivência dos pacientes com leucemia aguda

Observação	Tempos	$\log_{10}$(WBC)	Tempos	$\log_{10}$(WBC)
1	65	3,36	143	3,85
2	156	2,88	56	3,97
3	100	3,63	26	4,51
4	134	3,41	22	4,54
5	16	3,78	1	5,00
6	108	4,02	1	5,00
7	121	4,0	5	4,72
8	4	4,23	65	5,0
9	39	3,73		

Fonte: Lawless (1982)

O modelo Gama ajustado foi dado por:

$$\hat{\mu} = \frac{1}{-0,0346 + 0,0135 \times \log_{10}(\text{WBC})}$$

Tendo como referência o nível nominal de significância fixado em 5%, a probabilidade da falta de ajuste obtida por meio da deviance foi 0,1382. Portanto, há evidências estatísticas de que o modelo está bem ajustado aos dados sem a necessidade de incorporar termos de maior grau, como por exemplo, quadrático ou cúbico. Seguindo essa validação, procedeu-se com uma análise de diagnósticos de outliers e observações influentes, conforme ilustra a Figura 3.

Os gráficos para o diagnóstico de outliers e observações influentes foram computados usualmente seguindo a matriz de alavanca e distância de Cook, tendo como fundamentação a influência nas estimativas dos parâmetros e afastamento da verossimilhança. Desta forma, nota-se que, a segunda observação foi diagnosticada como uma observação influente e discrepante. Tal fato também foi diagnosticado considerando a distância de Mahalanobis computada em função das correlações dos valores preditos e da distribuição

acumulada dos mínimos (Fmin). Assim, percebe-se que a metodologia proposta concerne em identificar observações outliers e influentes em concordantes com os métodos usuais de diagnóstico utilizados na análise de um modelo Gama.

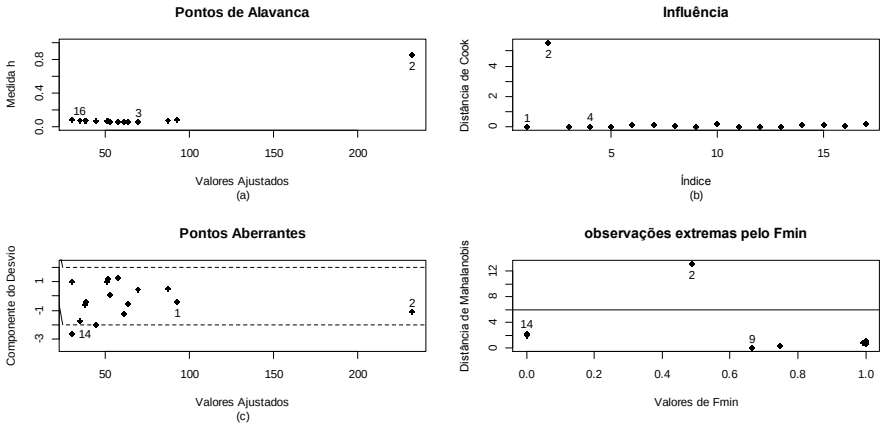

Figura 3 Métodos de diagnósticos de observações outliers e discrepantes no modelo Gama proposto para a aplicação.

5 CONCLUSÃO

Em função da validação do procedimento proposto para identificar observações outliers utilizando distribuição acumulada dos mínimos em um modelo Gama contaminado dada pela técnica Monte Carlo concluiu-se que o método é eficiente por apresentar elevadas probabilidades de acurácia, partindo do pressuposto que a distribuição dos outliers pertence à família exponencial. Em se tratando da aplicação, por meio do exemplo, decorrente a similaridade entre os métodos usuais, conclui-se que o método é factível de ser aplicado em modelos generalizado com respostas Gama.

REFERÊNCIAS

AITKIN, M. **Statistical Inference**: an integrated bayesian likelihood approach. London: CRC Press, 2010. 254 p.

AZZALINI, A. **Statistical inference based on the likelihood**. London: CRC Press, 1996. 352 p.

BARBATO, G. et al. Features and performance of some outliers detection methods. **Journal of Applied Statistics**, Abingdon, v. 38, n. 10, p. 2133-2149, Oct. 2011.

CASELLA, G.; BERGER, R. L. **Statistical inference**. 2nd ed. Boston: Thomson Learning, 2002. 660 p.

COORAY K.; ANANDA, M. A. A Generalization of the Half-Normal distribution with applications to lifetime data. **Communications in Statistics: Theory and Methods**, New York, v. 37, n. 9, p. 1323-1337, 2008.

COUSINEAU, D.; CHARTIER, S. Outliers detection and treatment: a review. **International Journal of Psychological Research**, Medellín, v. 3, n. 1, p. 58-67, 2010.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous univariate distributions**. 2nd ed. New York: Wiley, 1994. v. 1, 784 p.

LAWLESS, J. F. **Statistical models and methods for mifetime**. New York: Wiley, 1982. 664 p.

MARRON, J. S.; WAND, M. P. Exact mean integrated squared error. **The Annals of Statistics**, Hayward, v. 20, n. 2, p. 712-736, June 1992.

NOUFAILY, A.; JONES, M. C. On maximization of the likelihood for the generalized gamma distribution. **Computational Statistics**, Heidelberg, v. 28, n. 2, p. 505-517, Apr. 2013.

PAULA, G. A. **Modelos de regressão com apoio computacional**. São Paulo: IME/USP. 2011. 396 p.

PAULINO, C. D.; SINGER, J. M. **Análise de dados categorizados**. São Paulo: E. Blucher, 2006. 648 p.

POTTER, K. et al. Methods for presenting statistical information: the box plot. In: HAGEN, H.; KERREN, A.; DANNENMANN, P. (Ed.). **Visualization of Large and unstructured data sets**. Dagstuhl Castle: Gesellschaft für Informatik, 2006. p. 97-106.

RAMOS, M. F. **Técnica de mineração de dados na discriminação sensorial da qualidade do café arábica e o meio físico**. 2013. 70 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2013.

R CORE TEAM. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria: Computing RFS, 2014. Disponível em: <http://www.R-project.org/.>. 2014. Acesso em: 17 dez. 2015.

RIBEIRO JUNIOR, P. J. et al. **Métodos computacionais em inferência estatística**. Curitiba: UFRP, 2012. 284 p. Minicurso.

RICHARDSON, S.; GREEN, P. On bayesian analysis of mixtures with an unknown number of components. **Journal of the Royal Statistical Society, Series B: Statistical Methodology**, Oxford, v. 59, n. 4, p. 731-792, 1997.

RIZZO, M. L. **Statistical computing with R**. Boca Raton: Chapman & Hall, 2008. 399 p.

ROEDER, K.; WASSERMAN, L. Practical bayesian density estimation using mixtures of normals. **Journal of the American Statistical Association**, New York, v. 92, n. 439, p. 894-902, Sept. 1997.

ROSS, S. M. Peirce's criterion for the elimination of suspect experimental data. **Journal of Engineering Technology Management**, Amsterdam, v. 20, n. 2, p. 38-41, Sept. 2003.

SHEVLYAKOV, G. et al. Robust versions of the Tukey boxplot with their application to detection of outliers. In: ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASS); IEEE INTERNATIONAL CONFERENCE ON, 2103, Vancouver. **Proceedings**… Vancouver: IEE, 2013. p. 6506-6510.

SONG, K. Globally convergent algorithms for estimating generalized gamma distributions in fast signal and image processing. **IEEE Transactions on Image Processing**, New York, v. 17, n. 8, p. 1233-1250, Aug. 2008.

TUKEY, J. W. **Exploratory data analysis**. London: Pearson, 1977. 607 p.

TURKMAN, M. A. A.; SILVA, L. S. **Modelos lineares generalizados**: da teoria a prática. Lisboa: Universidade Técnica de Lisboa, 2000. 139 p.

WIPER, M.; INSUA, D. R.; RUGGERI, F. Mixtures of gamma distributions with applications. **Journal of Computational and Graphical Statistics**, Alexandria, v. 10, n. 3, p. 440-454, Sept. 2001.

ANEXOS

ANEXO A – Script utilizando na simulação Monte Carlo para obtenção das amostras contaminadas, probabilidades, Distância de Mahalanobis e Distribuição acumulada dos mínimos.

```
med_ajuste=matrix(0,nsim,6)

colnames(med_ajuste) <- c("Dev.res","gl_res","p-valor",
"cor(Fmin,pred)", "disp_est", "disp_desc")

aux_taxa=matrix(0,1,6)

colnames(aux_taxa) <- c("E1","E2","A","Er","Tx_out","Tx_obs")

nsim=500


################ inicio do programa ################## #


 for (sim in 1:nsim)

{

   teste1=simula_obs(n,x,beta,alfa,a,b)

   cv=sd(teste1$ymisto[,1])/mean(teste1$ymisto[,1])

   disp=1/((cv)^2)

   sgama=rgamma(n,shape=teste1$mu,scale=disp)

   mod_out <-glm((sgama ~ -1 + x[,1] + x[,2]),family = Gamma

  (link="inverse"))

   yajus <- mod_out$fitted


# Estimativa do parâmetro de dispersão - método dos momentos #
```

```
  est_disp=gamma.shape(mod_out)
  fi=as.numeric(est_disp[1])
  teste2=Ac_Fmin(teste1$ymisto,yajus,teste1$mu,fi)
  Li = qchisq(0.95,2)  # ### Alterar o limite #### #
  teste3 = t (as.matrix (taxas (n, teste2$DH,
  teste1$ymisto[,2],Li)))
```

########## Resumo dos resultados ################## #

```
 med_ajuste[sim,1]=mod_out$deviance
  med_ajuste[sim,2]=mod_out$df.residual
  med_ajuste[sim,3]=1-
  pchisq(mod_out$deviance,mod_out$df.residual)
  med_ajuste[sim,4]=teste2$R[1,2]
  med_ajuste[sim,5]=fi
  med_ajuste[sim,6]=disp
  aux_taxa=rbind(aux_taxa,teste3) }
```

############## Fim do programa ################## #

```
      summary(med_ajuste)  ;     summary(aux_taxa)
```

ANEXO B - Script utilizando na simulação Monte Carlo para aplicação aos dados reais

.

```
dados <- read.table ("dados.txt", h=T)
attach(dados)
fit.model <- glm(tempo ~ wbc,
family=Gamma(link="inverse"))
source("diag_gama.txt")
identify(fitted(fit.model),h,n=3)

# #### Identificação dos outliers utilizando Fmin #####

n=nrow(dados)
beta=coef(fit.model)
mu <- 1/(beta[1] + beta[2]*wbc)# ### Valor predito ####

# ## ajuste do modelo com valores em torno da média ## #

cv=sd(tempo)/mean(tempo)
disp=1/((cv)^2)
Fac=pgamma(tempo,mu,disp, lower=T)
Fmin=(1-(1-(Fac))^n)

# ###### Distância de Mahalanobis ##### #

D=cbind(Fmin,mu)
S=cov(D)
DH=mahalanobis(D,colMeans(D),S)
Li=qchisq(0.95,2)
plot(Fmin,DH,pch=16,main="observações extremas pelo
Fmin",xlab = "Valores de Fmin", ylab = "Distância de
Mahalanobis" )
abline(h =Li , lty = 1)
```