MODELOS DE REGRESSÃO PARA RESPOSTAS NO INTERVALO UNITÁRIO: ESPECIFICAÇÃO, ESTIMAÇÃO E COMPARAÇÃO.

Wagner Hugo BONAT [1]
Paulo Justiniano RIBEIRO Jr. [1]
Walmes Marques ZEVIANI [1]

■ RESUMO: Modelos de regressão são largamente utilizados em diversas áreas de aplicação para descrever associações entre uma variável resposta e variáveis explicativas. Os modelos lineares gaussianos muito utilizados inicialmente foram gradualmente estendidos para diversos tipos de variáveis resposta. Muitas destas extensões foram posteriormente descritas como casos particulares da classe mais geral de modelos lineares generalizados (MLG) que, sob uma mesma abordagem, acomodam uma diversidade de formas para a variável resposta e funções ligando parâmetros das distribuições a um preditor linear. Desde então a estrutura dos MLG tem sido estendida em diversos desenvolvimentos subsequentes em modelagem estatística como modelos aditivos generalizados, de superdispersão, dentre outros. Variáveis respostas com valores restritos a um certo intervalo, em geral $(0, 1)$ são comuns em ciências sociais, agronomia, psicometria dentre outras áreas. As distribuições beta e simplex são usualmente adotadas, dentre outras opções na literatura. Neste artigo modelos de regressão para respostas restritas são especificados na forma de uma classe geral que inclui as formas usuais bem como permite explorar uma maior diversidade de modelos. Casos particulares são definidos pelas escolhas de três componentes: a distribuição de probabilidades para a resposta, a função de ligação de um parâmetro da distribuição escolhida e o preditor linear a uma função de transformação da resposta. São mostrados resultados das análises de quatro diferentes conjuntos de dados considerando as distribuições beta, simplex, Kumaraswamy e gaussiana, e as funções logit, probit, complemento log-log, log-log, Cauchit e Aranda-Ordaz como opções para ligação e transformação da variável resposta. Analises baseadas na verossimilhança são conduzidas de forma unificada para ajuste, comparação e escolha de modelos e códigos são disponibilizados. Os resultados mostram que não há uma forma de modelo que se destaque ilustrando a importância de se explorar uma ampla classe de modelos a cada análise.

[1]Departamento de Estatística - DEST, Laboratório de Estatística e Geoinformação - LEG, Universidade Federal do Paraná - UFPR, CEP: 81531-990, Curitiba, Paraná, Brasil, E-mail: *wagner,paulojus,walmes@leg.ufpr.br*

▪ PALAVRAS-CHAVE: máxima verossimilhança ; variáveis restritas ; proporções ; índices ; taxas.

1 Introdução

A área de modelagem por regressão estatística teve seu início com o clássico modelo linear gaussiano, adotado em diversas áreas como a principal ferramenta de análise da relação entre uma variável resposta com possíveis variáveis explicativas. Apesar de largamente utilizado, o modelo apresenta limitações para respostas não gaussianas para as quais outros métodos foram sendo desenvolvidos. Nelder e Wedderburn (1972) e McCullagh e Nelder (1989) marcam um grande avanço para modelos de regressão, unificando diversas especificações sob uma flexível classe de modelos, denominada de modelos lineares generalizados (MLG).

Sob a estrutura genérica de MLG é possível construir modelos adequados para diferentes tipos de variáveis resposta tais como binárias, contagens, politônicas e contínuas. É possível modelar parâmetros de locação e dispersão como funções de covariáveis. A partir da declaração explícita de um modelo nesta família, pode-se obter a função de verossimilhança dos parâmetros, possibilitando estimação pontual e intervalar, além da construção de testes de hipóteses e critérios para comparação de modelos e portanto, todos os elementos necessários para a prática de modelagem estatística.

Apesar da grande flexibilidade, os modelos lineares generalizados usuais apresentam limitações para variáveis resposta restrita a um intervalo (a, b), sendo o intervalo unitário $(0, 1)$ o mais usual. Tais tipos de dados surgem em diversas áreas. Índices de desenvolvimento humano, qualidade de vida, medidas de bem estar, dentre outros, são comumente medidos em ciências sociais. Nessas situações existem valores máximo e mínimo na escala de medida de uma variável observável ou latente. Em testes psicométricos, são avaliadas medidas latentes como inteligência, concordância com algum pensamento, habilidade, dentre outras. A variável resposta pode ainda ser uma proporção contínua, como no exemplo de percentual da renda total que uma família gasta com alimentação que consideraremos mais adiante. Em ciências agrárias é comum mensurar na forma de percentuais os níveis de infestação de uma doença sobre culturas ou partes afetadas de estruturas de plantas. Independentemente da área de aplicação, uma variável resposta restrita a qualquer intervalo (a, b) pode ser transladada para o intervalo unitário considerado aqui e resultados podem ser retornados para o intervalo original ao final se necessário.

Extensões de modelos lineares generalizados para tratar dados dessa natureza têm sido propostas na literatura. Kieschinick e McCullough (2003) revisam diferentes abordagens para construção de modelos para variáveis resposta restritas. Os autores consideram o clássico modelo de regressão gaussiano, que ignora a restrição e a heterocedasticidade, modelos com domínio restrito, como o modelos de regressão beta e simplex e ainda modelos semi-paramétricos estimados por quasi-verossimilhança. Os autores recomendam o uso da regressão beta como uma

abordagem padrão baseando-se na análise de exemplos com dados reais. De forma independente, Ferrari e Cribari-Neto (2004) apresentam o modelo de regressão beta com maior detalhamento matemático e computacional, incluindo ainda análise de resíduos. Na sequência, diversos avanços passaram a ser propostos na literatura. Modelagem de média e dispersão é adotada em Cepeda e Gamerman, 2005 e Simas *et al.*, 2010. Análises de resíduos e diagnósticos são apresentados por Espinheira *et al.* (2008a), Espinheira *et al.* (2008b) e Rocha e Simas (2010). Correções de viés para os estimadores de máxima verossimilhança, foram apresentadas em Vasconcellos e Cribari-Neto (2005), Ospina *et al.* (2006) e Simas *et al.* (2010). Branscum *et al.* (2007) utiliza um modelo beta e inferência bayesiana na avaliação da distância genética entre vírus. Smithson e Verkuilen (2006) adotam modelos beta em estudos psicométricos. Modelos beta de mistura são discutidos em Verkuilen e Smithson (2011). Lima (2007) adapta o teste RESET de Ramsey para especificação de modelos lineares para modelos de regressão beta.

O modelo de regressão beta é implementado no pacote *betareg* (Cribari-Neto e Zeileis, 2010) para o ambiente estatístico R (R Development Core Team, 2013). Extensões são descritas em Grün *et al.* (2011), tais como, correções de viés, particionamento recursivo e modelos de mistura finita. Alguns desenvolvimentos paralelos na análise de séries temporais foram feitos em McKenzie (1985), Grünwald *et al.* (1993) e Rocha e Simas (2010). Da-Silva *et al.* (2011) apresenta um modelo dinâmico bayesiano beta para modelagem e previsão de séries temporais com aplicação à dados de taxa de desemprego no Brasil. Bonat *et al.* (2013) apresenta um modelo beta com efeitos aleatórios, para modelar observações não independentes, como em medidas repetidas, estudos longitudinais, entre outros padrões de dependência.

Apesar da disseminação do uso da distribuição beta, há poucas comparações com abordagens alternativas e menos frequentemente adotadas como a distribuição simplex (Miayshiro, 2008), Kumaraswamy proposta em Lemonte *et al.* (2013) ou mesmo com o modelo gaussiano com transformações da variável resposta. Um exemplo desse último caso é o uso da função *logit* aplicada a dados com valores no intervalo $(0, 1)$, obtendo-se uma resposta com valores nos reais que pode ser modelada com a distribuição normal.

McCullagh e Nelder (1989, pg. 378) discute brevemente diferenças entre modelos especificados com resposta transformada e modelos com diferentes distribuições de probabilidade para a variável resposta. A transformação deve ser capaz de gerar efeitos aditivos e variância constante, porém, não existem garantias que tal transformação exista, nem que seja única. Modelos lineares generalizados são mais atrativos por acomodarem tanto estruturas de efeitos de covariáveis, como de relacionamento entre média e variância determinadas pela escolha da distribuição de probabilidades assumida para a variável resposta. A clássica família de Box e Cox (1964) de transformação de dados, define transformações do tipo potência, que não são diretamente aplicáveis para dados em intervalos restritos e pode ainda ser usada como uma função de ligação em MLG.

No presente artigo descrevemos sob uma forma genérica os modelos de

regressão para resposta no intervalo unitário que permite especificar os modelos já mencionados como casos particulares. A partir da forma genérica apresentamos e comparamos ajustes para diversas especificações, disponibilizando códigos para as análises. Desta forma, destacamos que não é necessário restringir a escolha do modelo, sendo possível em uma análise de dados, inspecionar sob uma abordagem comum, de forma prática e objetiva, uma classe mais ampla de modelos. Discute-se estimação por máxima verossimilhança e compara-se as especificações em quatro conjuntos de dados reais provenientes de ciências sociais e agrárias. Cada modelo considerado é determinado pelas escolhas para os três componentes básicos da especificação genérica. As opções de escolha permitem criar uma diversidade de modelos adequados para variáveis resposta no intervalo unitário e relações com covariáveis que incluem diversos modelos não-lineares.

A forma geral dos modelos de regressão para variáveis resposta no intervalo unitário é apresentada a seguir. Nas sessões subsequentes são apresentadas análises para dados em diferentes contextos, sempre comparando formas de construção de modelos. Os principais resultados da comparação entre as abordagens são discutidos ao final com recomendações gerais para a prática de análise de dados e para trabalhos futuros.

2 Especificação do modelo

Para variáveis respostas Y_i independentes assume-se um modelo da forma:

$$\begin{aligned} T(Y_i|x_i;\lambda) &\sim d(\mu_i,\phi) \\ \mu_i &= f(x_i,\beta_x;\delta) \end{aligned} \tag{1}$$

em que cada x_i é um vetor de regressoras associado à i-ésima observação.

Para a definição do modelo é necessário especificar três componentes, as funções $d(\cdot)$, $T(\cdot)$ e $f(\cdot)$. A primeira define a distribuição de probabilidade da variável resposta é descrita por dois parâmetros. O primeiro de locação μ_i, tipicamente a esperança ou mediana da variável resposta ou, de forma mais geral, qualquer quantidade de interesse a ser relacionada com as regressoras. O segundo de dispersão ϕ, aqui tratado como um parâmetro extra na verossimilhança, embora também possa ser modelado como função de covariáveis. A escolha de $d(\cdot)$ impõe condições à escolha das demais funções para garantir modelos válidos, como descrito a seguir.

A segunda é a função de transformação $T(\cdot)$, aplicada a variável resposta e que pode ser indexada por um parâmetro de forma λ. Esta função deve ter domínio no $(0,1)$, já que Y_i é uma variável restrita a tal intervalo e com contradomínio compatível com o suporte da distribuição $d(\cdot)$. Por fim é necessário especificar a função de ligação $f(\cdot)$, que pode depender de dois conjuntos de parâmetros, β_x associado ao efeito das covariáveis x e um parâmetro de forma δ. Esta função deve ter domínio nos reais permitindo que as covariáveis possam assumir qualquer valor e com contradomínio compatível com o espaço paramétrico de μ_i. Além disso, por

simplicidade assumimos que as funções $T(\cdot)$ e $f(\cdot)$ sejam monotônicas e duplamente diferenciáveis.

Para uma amostra aleatória, a função de verossimilhança pode ser escrita na forma:

$$L(\underline{\theta}; y) = \prod_{i=1}^{n} d(f(x_i, \beta_x, \delta), \phi) \left\| \frac{\partial T(Y_i|x_i, \lambda)}{\partial Y_i} \right\|. \quad (2)$$

As estimativas dos parâmetros $\underline{\theta} = (\beta_x, \delta, \lambda, \phi)$ são obtidas pela maximização da função de verossimilhança (2). A função permite ainda obter intervalos de confiança, sejam baseados em aproximação quadrática (Wald) ou por perfilhamento da função de verossimilhança. Os ajustes de modelos de mesma dimensão podem ser comparados diretamente pelos valores maximizados da log-verossimilhança enquanto que modelos encaixados são comparados por testes da razão de verossimilhanças. Em outros casos critérios como de *Akaike* ou *BIC* podem ser utilizados na comparação e escolha de modelos.

Para cada especificação de modelo, obtém-se a partir de (2) as expressões da função escore, matriz de informação observada e, quando possível, a expressão da informação esperada. De forma geral, não é possível obter expressões fechadas para os estimadores de máxima verossimilhança de $\underline{\theta}$ e, em certos casos, nem mesmo as expressões do gradiente e hessiano. O uso de métodos numéricos é necessário para a maximização de (2) e os algoritmos devem ser escolhidos e calibrados cuidadosamente para garantir convergência.

Nas análises apresentadas aqui utilizamos algoritmos em R (R Development Core Team, 2013) seguindo procedimentos para implementação computacional descritos em Bonat et. al. (2012). Inicialmente definiu-se uma função em R para a forma geral do modelo, que calcula o valor da log-verossimilhança para um conjunto de parâmetros e para a especificação desejada do modelo. No ajuste são utilizados algoritmos para maximização implementados na função *optim*(). Em geral iniciamos utilizando o algoritmo BFGS (Byrd, 1995) em combinação com o algoritmo SANN *Simulated Annealing* (Belisle, 1995) para os casos de dificuldades com a convergência. Esta estratégia mostrou-se satisfatória para os conjuntos de dados analisados. Uma vez ajustado o modelo, são extraídas outras quantidades de interesse como erros padrão e intervalos de confiança atrevés de perfis de verossimilhança. No perfilhamento a função de verossimilhança é explorada para diversas combinações dos parâmetros. Além de fornecer intervalos mais realistas que os baseados na aproximação quadrática, também nos fornece indicação adicional de que o algoritmo numérico realmente encontrou o ponto de máximo da função de verossimilhança.

A log-verossimilhança é uma medida de compatibilidade do modelo com o particular conjunto de dados. Os valores maximizados das log-verossimilhanças permitem comparar o ajuste de modelos que tenham a mesma estrutura de covariáveis e mesmo número de parâmetros, mas que tenham diferentes formas de construção dadas pelas escolhas para as componentes $T(\cdot)$, $d(\cdot)$ e $f(\cdot)$.

2.1 Construções avaliadas

Na estrutura genérica (1), adotamos componentes usuais na literatura que, quando combinados, geram desde modelos tradicionais como o modelo de regressão beta com ligação *logit*, até modelos mas recentemente propostos como o Kumaraswamy com função de ligação *complemento log-log*.

Para a distribuição de probabilidades $d(\cdot)$ da resposta consideramos quatro opções: gaussiana, beta, simplex e Kumaraswamy (Kw) com expressões nas quais o parâmetro de escala μ será associado às covariáveis. Para as três primeiras $E[Y_i] = \mu_i$ e para última $md[Y_i] = \mu_i$. As expressões das densidades sâo:

- *gaussiana*: $d(y;\mu,\phi) = \frac{1}{\sqrt{2\pi\phi}} \exp\{-\frac{1}{2\phi^2}(y-\mu)^2\}$;
- *beta*: $d(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$;
- *simplex*: $d(y;\mu,\phi) = (2\pi\phi^2\{y(1-y)^3\})^{-1/2} \exp\left\{-\frac{1}{2\phi^2}\left\{\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}\right\}\right\}$;
- *Kumaraswamy*: $d(y;\mu,\phi) = \frac{\phi \ln(1-0.5)}{\ln(1-\mu^\phi)} y^{\phi-1}(1-y^\phi)^{\frac{\ln(1-0.5)}{\ln(1-\mu^\phi)}-1}$.

Para as distribuições beta, simplex e Kumaraswamy tem-se $y \in (0,1)$, $\mu \in (0,1)$. Para o caso da distribuição gaussiana tem-se $y, \mu \in \Re$. Para todas as distribuições $\phi > 0$. A construção com a distribuição gaussiana não é estritamente adequada pois, apesar da função $f(\cdot)$ mapear o intervalo unitário, o espaço paramétrico da média de uma distribuição gaussiana é toda a reta real. Portanto a resposta assume valores restritos ao intervalo unitário mas modelo gaussiano tem suporte em toda a reta real. Mesmo assim tal construção foi explorada por ser comum na literatura, como por exemplo, no ajuste de modelos de crescimento de incidência de doenças em plantas.

Selecionamos seis formas funcionais para $T(\cdot)$ $(0,1) \mapsto \mathbb{R}$ com as seguintes expressões:

- *logit*: $T(y) = log\left(\frac{y}{1-y}\right)$;
- *probit*: $T(y) = \Phi(y)$, onde $\Phi(\cdot)$ é a acumulada da distribuição Normal;
- *complemento log-log*: $T(y) = \ln(-\ln(1-y))$;
- *log-log*: $T(y) = -\ln(-\ln(y))$;
- *cauchit*: $T(y) = \tan(\pi \cdot (y - 1/2))$;
- *Aranda-Ordaz*: $T(y;\lambda) = \ln\left\{\frac{(1-y)^{-\lambda}-1}{\lambda}\right\}$, com $\lambda > 0$.

Embora $T(\cdot)$ seja aplicada à variável resposta e (a inversa de) $f(\cdot)$ seja aplicada ao parâmetro μ_i, é possível utilizar a mesma forma funcional para ambas. Dessa forma inversas dessas mesmas funções são as opções consideradas para função de ligação $f(\cdot)$.

Tabela 1 - Log-verossimilhança para modelos ajustados sob diferentes distribuições e funções de ligação ou transformação (última linha) - IQVC.

Distribuições	Funções de ligação ou transformação					
	Logit	Probit	Cloglog	Loglog	Cauchit	Aranda
beta	56,84	56,14	54,18	57,75	59,34	58,47
gaussiana	57,19	56,74	55,40	57,78	58,71	58,20
Kw	60,49	60,04	58,48	61,36	62,17	62,16
simplex	51,57	50,76	48,73	52,94	55,87	54,29
trans-gaussiana	54,79	54,95	52,95	53,01	46,68	54,85

Com escolhas de $d(\cdot)$, $T(\cdot)$, $f(\cdot)$ definimos 30 modelos a serem ajustados. Os seis primeiros são os modelos de resposta transformada definidos com as seis opções para $T(\cdot)$ combinadas com a distribuição gaussiana para $d(\cdot)$ e função identidade como ligação. Os demais 24 modelos são obtidos pelas combinações das seis opções para as ligações $f(\cdot)$ com as quatro opções de $d(\cdot)$. Nesses casos a função $T(\cdot)$ é a identidade. Todos esses modelos são ajustados e comparados nas análises de quatro conjuntos de dados apresentadas a seguir.

3 Resultados

3.1 Índice de Qualidade de Vida Intraurbana em Curitiba - IQVC

O Índice de Qualidade de Vida Intraurbano de Curitiba (IQVC) resulta da composição de 18 indicadores, separados em 5 áreas temáticas: habitação, saúde, educação, segurança e transporte. O método para a sua construção segue as premissas do Índice de Desenvolvimento Humano (IDH), preconizado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD)[1]. Por construção, tal índice resulta em um número restrito ao intervalo unitário. Quanto maior o valor, melhor é a qualidade de vida da população residente avaliada no bairro. Os dados que compõe o índice são provenientes da base de microdados do Censo 2000, disponibilizado pelo IBGE e trabalhados pelo Instituto de Pesquisa e Planejamento Urbano de Curitiba (IPPUC).

O objetivo desta análise é relacionar o IQVC com a renda média do bairro, mensurada em salários minímos vigentes na época da pesquisa. Tem-se portanto, uma variável resposta restrita ao intervalo unitário e uma covariável contínua. A amostra tem tamanho 75 que corresponde ao número de bairros da cidade. Avaliamos diferentes construções ajustando os 30 modelos mencionados na Sessão 2. O valor da log-verossimilhança para cada um dos modelos é apresentado na Tabela 1 para as diferentes opções de distribuição $d(\cdot)$ nas linhas da tabela. As colunas da tabela correspondem às diferentes opções funções de ligação $f(\cdot)$, exceto pela última linha onde essas correspondem a funções de transformação $T(\cdot)$.

[1] http://hdr.undp.org/en/humandev/

Tabela 2 - Estimativas pontuais e erro padrões para os modelos com *cauchit* como função de ligação ou de transformação (última coluna) - IQVC.

Efeitos	Distribuições de probabilidade				
	beta	gaussiana	Kw	simplex	trans-gaussiana
Intercepto	-0.53(0.11)	-0.57(0.10)	-0.59(0.10)	-0.57(0.09)	-0.56(0.14)
Renda	7.82(0.93)	8.51(1.08)	8.67(1.15)	8.99(1.19)	8.99(1.08)

Todos os modelos possuem a mesma quantidade de parâmetros tornando os valores de verossimilhança diretamente comparáveis, com excessão dos definidos com a função Aranda-Ordaz que possuem um parâmetro adicional. Nestes casos a verossimilhança deve ser maior ou igual às que utilizam a função *logit*, que é um caso particular da Aranda-Ordaz com parâmetro de forma igual a 1.

De acordo com os resultados apresentados na Tabela 1, de forma geral, o modelo que apresenta a maior log-verossimilhança é o que define $d(\cdot)$ pela distribuição Kw e função de ligação $f(\cdot)$ pela *cauchit*. Nota-se ainda que a função de ligação *cauchit* foi a melhor em combinação com as quatro suposições distribucionais e a distribuição Kw foi superior para todas escolhas de funções de ligação. Nesse exemplo, a escolha das funções de ligação do parâmetro pouco afeta a qualidade do ajuste.

Dentre os modelos transformados, o melhor ajuste foi obtido com $T(\cdot)$ dada pela função *probit*, porém com resultados próximo ao obtido com a *Aranda-Ordaz* e seu caso particular a *logit*. Os modelos transformados apresentaram verossimilhanças nitidamente inferiores em todos os casos.

A Tabela 2 apresenta as estimativas pontuais e erros padrão para os coeficientes do modelo obtidos por cada uma das especificações de distribuição em $d(\cdot)$ usando a *cauchit* como função de ligação e também como de transformação para comparação dos coeficientes.

Os resultados na Tabela 2 mostram que as estimativas pontuais e os erros padrão são muito próximas independentemente da suposição distribucional. A distribuição beta apresenta a maior diferença em relação as estimativas do intercepto e do efeito da covariável renda. Os gráficos na Figura 1 mostram o ajuste desses modelos com as bandas de predição, sobreposto aos dados observados. A renda em salários mínimos foi dividida por 100 para facilitar a visualização dos algarismos significativos dos coeficientes.

Os ajustes apresentados na Figura 1, são visualmente semelhantes. As bandas de confiança para o modelo transformado são mais simétricas do que no demais modelos, porém a diferença é pequena. Concluímos que, neste caso, apesar de existirem diferenças nos valores das log-verossimilhanças, na comparação de estimativas pontuais, seus respectivos erros padrões e valores preditos não são detectadas diferenças relevantes entre os modelos.

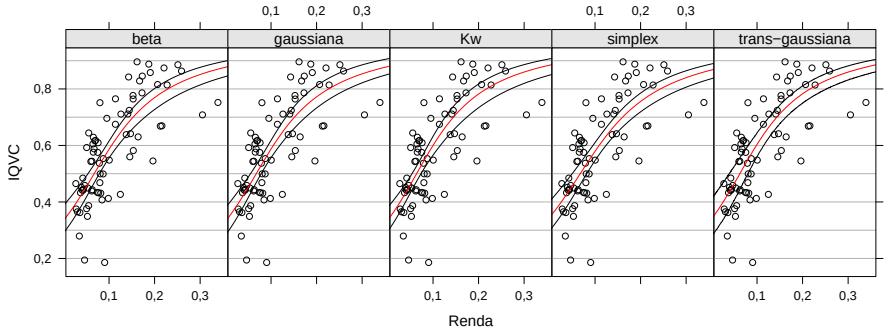

Figura 1 - Predição do IQVC como função da renda sobreposta aos dados originais para os modelos da Tabela 2.

Tabela 3 - Log-verossimilhança para modelos ajustados sob diferentes distribuições e funções de ligação ou transformação (última linha) - *Food Expenditure*.

Distribuições	Funções de ligação ou transformação					
	Logit	Probit	Cloglog	Loglog	Cauchit	Aranda
beta	45,33	45,09	45,77	44,55	46,96	47,51
gaussiana	45,80	45,50	46,34	44,78	47,68	47,87
Kw	48,88	48,65	49,25	48,10	50,19	50,51
simplex	45,60	45,43	45,90	45,04	46,78	47,40
trans-gaussiana	45,22	45,27	45,00	44,77	41,41	45,22

3.2 Percentual da renda gasta com alimentação

Nesse segundo exemplo são apresentados os resultados das análises de dados do percentual da renda gasto com alimentação em uma amostra de 38 economias domésticas de uma grande cidade dos Estados Unidos. Este conjunto de dados foi analisado em Ferrari e Cribari-Neto (2004) ao apresentarem o modelo de regressão beta e é disponibilizado no objeto *FoodExpenditure* do pacote *betareg* (Cribari-Neto e Zeileis, 2010). A Tabela 3 apresenta o valor da log-verossimilhança maximizada para cada uma das construções consideradas neste texto. O modelo tem como resposta o percentual da renda gasta com alimentação e como covariáveis a renda e o número de pessoas que compõe a família (economia doméstica).

De acordo com os resultados apresentados na Tabela 3 a distribuição Kw é a que apresenta os maiores valores de log-verossimilhança. Dentre as funções de ligação a *cauchit* é a que fornece os melhores ajustes. A combinação de Kw com *cauchit* é a melhor dentre todas as avaliadas.

A suposição distribucional é a que tem maior impacto na qualidade do ajuste medida pela log-verossimilhança. Desta forma, as estimativas pontuais e erros padrões para as diferentes escolhas de $d(\cdot)$ em combinação com a função de ligação

Tabela 4 - Estimativas pontuais e erro padrões para os modelos com *cauchit* como função de ligação ou de transformação (última coluna) - *Food Expenditure*.

Efeitos	Distribuições de probabilidade				
	beta	gaussiana	Kw	simplex	trans-gaussiana
Intercepto	-0.50(0.21)	-0.53(0.21)	-0.71(0.19)	-0.54(0.24)	-0.51(0.31)
Renda	-14.15(3.23)	-14.96(3.19)	-10.54(2.36)	-12.68(3.19)	-12.85(4.12)
Pessoas	13.50(3.58)	15.17(3.48)	14.48(2.86)	12.55(3.69)	10.13(4.83)

cauchit podem ser comparadas na Tabela 4.

Os resultados mostram que o intercepto do modelo foi maior sob a distribuição Kw. O efeito da covariável renda foi semelhante em todas as distribuições, porém o menor erro padrão foi obtido a distribuição Kw. Os coeficientes da covariável *número de pessoas residentes* apresentaram a maior variabilidade entre as suposições distribucionais. No modelo transformado este efeito foi mensurado como sendo $10,131$ enquanto que no modelo gaussiano este efeito foi de $15,174$, uma diferença de $49,75\%$. Sob a distribuição Kw este efeito foi de $14,482$ e com o menor erro padrão de todas as distribuições $(2,857)$, e portanto, a menor incerteza associada aos efeitos das covariáveis.

Os ajustes dos modelos sobrepostos aos dados observados para cada uma das suposições distribucionais e cada um dos 7 valores da covariável *número de pessoas* são mostrados na Figura 2. Os ajustes são para a escolha da função *cauchit* como $T(\cdot)$ no caso do modelo transformado e como $f(\cdot)$ nos demais casos.

O modelo com a distribuição Kw captura melhor o comportamento dos dados quando comparado as outras abordagens em particular quando o número de residentes é 6. Além disso, as bandas de confiança são ligeiramente mais estreitas com este modelo refletindo as menores estimativas de erros padrão. A covariável renda foi dividida por 100 nas análises para facilitar a visualização e comparação dos dígitos significativos das estimativa dos coeficientes.

3.3 Índice de Qualidade de Vida dos Trabalhadores da Indústria Brasileira - IQVT

O Índice de Qualidade de Vida do Trabalhador da Indústria Brasileira - IQVT é composto por 25 indicadores separados em oito áreas temáticas: habitação, saúde, educação, saúde integral e segurança no trabalho, desenvolvimento de competências, atribuição de valor ao trabalho e orientação a participação e ao desempenho. A metodologia de construção segue as premissas do Índice de Desenvolvimento Humano - IDH, preconizado pela ONU em diversos países. O IQVT é expresso em valores no intervalo $(0,1)$, sendo que, quanto mais próximo de 1 melhor a qualidade de vida dos trabalhadores de uma determinada indústria.

Para o caso da indústria brasileira, foi realizada pelo Serviço Social da Indústria (SESI), no ano de 2010, uma pesquisa em 365 empresas distribuídas no Distrito

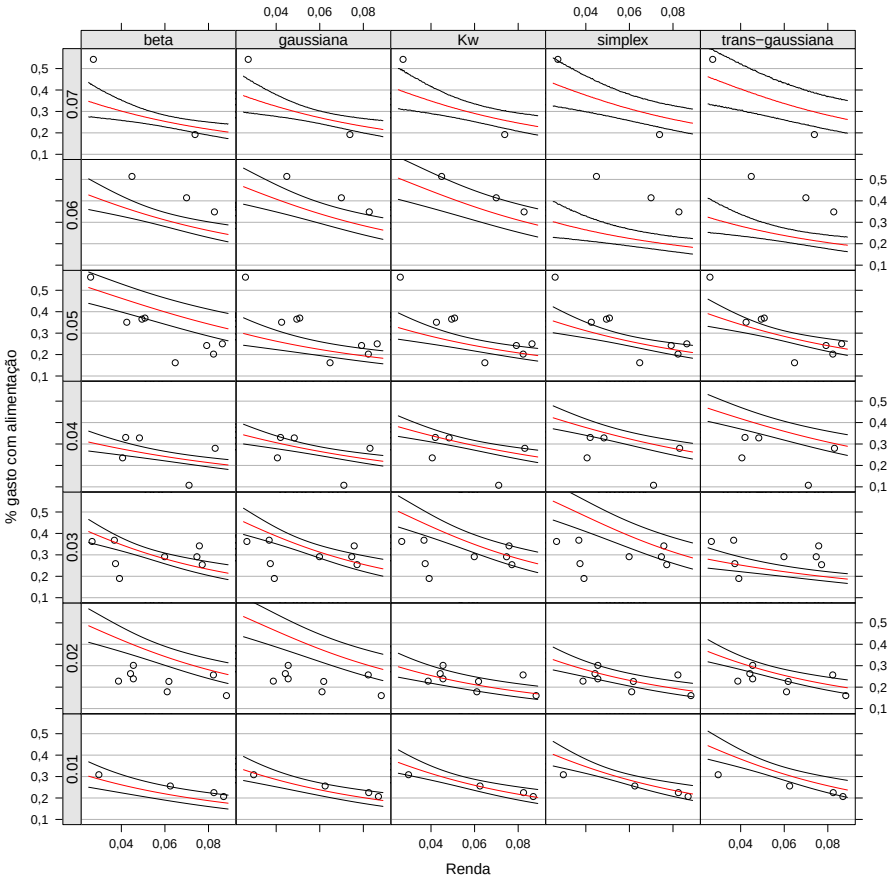

Figura 2 - Predição do percentual de renda gasto com alimentação, como função da renda e número de pessoas residentes, sobreposta aos dados originais para os modelos da Tabela 4.

Tabela 5 - Log-verossimilhança para modelos ajustados sob diferentes distribuições e funções de ligação ou transformação (última linha) - IQVT.

Distribuições	Funções de ligação ou tranformação					
	Logit	Probit	Cloglog	Loglog	Cauchit	Aranda
beta	567,03	566,87	566,23	567,39	567,77	567,68
gaussiana	564,34	564,20	563,63	564,65	564,98	564,88
Kw	564,81	564,67	564,06	565,27	565,71	565,79
simplex	568,72	568,53	567,77	569,15	569,63	569,50
trans-gaussiana	568,63	567,64	561,31	572,28	573,90	574,54

Federal e em oito dos 26 estados brasileiros. Foram entrevistados trabalhadores das indústrias, selecionadas segundo um plano amostral. O índice foi calculado a partir dos questionários de cada empresa pertencente à amostra. Além dos trabalhadores, a empresa foi também entrevistada sobre como trata a questão da qualidade de vida, seus gastos com benefícios sociais, dentre outros indicadores.

Para a presente análise separamos duas características de particular interesse a serem relacionadas com a qualidade de vida dos trabalhadores: *a renda média da empresa* e e o *estado* em que a empresa atua. A primeira indica a capacidade dos trabalhadores em suprir suas necessidades básicas, com alimentação, saúde, habitação, educação entre outras. A segunda é considerada uma *proxy* para as condições sociais determinadas por políticas públicas estaduais, que podem ser diferentes entre os estados.

Os modelos são ajustados com variável resposta *Índice de Qualidade de Vida dos Trabalhadores (IQVT)* e como covariáveis a *renda média da empresa* e o *estado da federação* em que a empresa atua. Essa última é categórica para a qual foi fixado como nível de referência o estado do Amazonas (AM). A Tabela 5 apresenta os valores das log-verossimilhanças para as construções de modelo consideradas.

Os resultados mostram que os modelos de resposta transformada apresentam as maiores log-verossimilhança com as transformações *log-log*, *cauchit* e *Aranda-Ordaz*. Para demais distribuições, os maiores valores aparecem em combinação com a distribuição simplex. A transformação *complemento log-log* apresenta a menor log-verossimilhança dentre as abordagens consideradas.

Os valores das log-verossimilhanças para os diferentes modelos apresentam maior variabilidade nesse conjunto de dados em comparação com os exemplos anteriores, variando de $561,30$ até $574,54$. A qualidade do ajuste é portanto aqui mais sensível às opções para construção dos modelos.

O modelo transformado com a função Aranda-Ordaz é o que apresenta a maior log-verossimilhança $(574,54)$. Porém, este modelo tem um parâmetro a mais que os modelos com as demais funções para $f(\cdot)$ ou $T(\cdot)$. Opta-se então pelo modelo com transformação *cauchit* apresenta uma log-verossimilhança de $573,90$ e portanto uma diferença na log-verossimilhança de apenas $0,64$. A *cauchit* teve ainda uma maior log-verossimilhança em combinação com as distribuições beta, simplex e gaussiana.

Para uma melhor comparação entre os modelos, a Tabela 6 apresenta as

Tabela 6 - Estimativas pontuais e erro padrões para os modelos com *cauchit* como função de ligação ou de transformação (última coluna) - IQVT.

Coeficientes	Distribuições de probabilidade				
	beta	gaussiana	Kw	simplex	trans-gaussiana
Intercepto	-0.02(0.04)	-0.03(0.04)	0.01(0.03)	-0.02(0.04)	-0.01(0.04)
Renda	3.27(0.29)	3.31(0.31)	3.11(0.26)	3.24(0.27)	3.26(0.26)
CE	0.03(0.04)	0.03(0.04)	0.02(0.03)	0.03(0.04)	0.03(0.04)
DF	-0.24(0.04)	-0.24(0.04)	-0.20(0.03)	-0.24(0.04)	-0.25(0.04)
MT	-0.10(0.04)	-0.10(0.04)	-0.08(0.03)	-0.10(0.04)	-0.10(0.04)
MS	0.02(0.04)	0.02(0.04)	0.01(0.03)	0.01(0.04)	0.02(0.04)
PA	0.11(0.04)	0.11(0.04)	0.09(0.03)	0.11(0.04)	0.11(0.04)
PR	0.01(0.03)	0.01(0.03)	0.01(0.03)	0.01(0.03)	0.01(0.03)
RO	-0.20(0.05)	-0.20(0.05)	-0.11(0.04)	-0.20(0.05)	-0.19(0.05)
RR	-0.15(0.05)	-0.15(0.05)	-0.14(0.04)	-0.15(0.05)	-0.16(0.06)

estimativas pontuais e os respectivos erros padrões dos coeficientes das covariáveis para a escolha da *cauchit* como função de ligação $d(\cdot)$ ou transformação $T(\cdot)$. A covariável renda é expressa em milhares de reais.

A Tabela 6 mostram resultados próximos para as estimativas pontuais para as diferentes construções. As maiores diferenças aparecem com a distribuição Kw que apresenta o intercepto positivo enquanto as demais indicam negativo. O efeito da renda é apenas ligeiramente menor. Sob a distribuição Kw também se obtém menores valores para os efeitos dos estados, em comparação com as demais. Já para o erro padrão os resultados são mais diversos. Por exemplo, sob a Kw o erro padrão foi de $-0,1145$ para o estado RO e em torno de $-0,19$ a $-0,20$ para as demais, uma razoável diferença. Para qualquer das especificações, as conclusões sobre significância dos efeitos são as mesmas.

Na Figura 3 são mostradas as curvas ajustadas sob as diversas especificações sobrepostas aos dados observados . Apesar das maiores diferenças nos valores das log-verossimilhanças, os ajustes dos modelos são próximos. É interessante notar que para esse conjunto de dados, diferentemente dos exemplo anteriores, a escolha de $T(\cdot)$ no caso dos modelos transformados e a de $f(\cdot)$ possui maior impacto no ajuste do que a escolha de $d(\cdot)$.

3.4 Incidência de doenças em plantas

O progresso temporal da incidência da ferrugem do pessegueiro (*Tranzschelia discolor*), foi avaliado em 11 cultivares de pessegueiro, com uma planta por cultivar, avaliando-se seis ramos por planta. Os dados são parte da tese de doutorado de Alves (2012). A incidência corresponde ao número de folhas com sintomas em relação ao total de folhas do ramo. Realizaram-se avaliações quinzenais entre novembro e abril de duas safras. A covariável *dia* expressa o número de dias entre o início do acompanhamento e a avaliação.

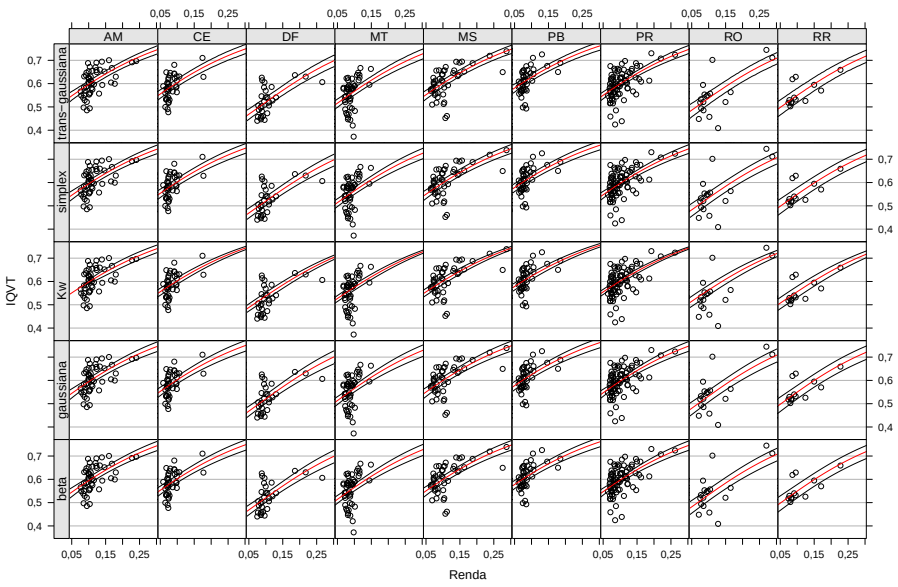

Figura 3 - Predição do IQVT como função da renda e estado de atuação sobreposta aos dados originais para os modelos da Tabela 6.

Usualmente, são ajustados modelos não-lineares gaussianos para descrever o crescimento da doença ao longo do tempo. Entre exemplos de modelos langamente utilizados estão o logístico, monomolecular e Gompertz (Spósito, M.B.; Bassanezi e Amorim, 2004). Nessa abordagem ignora-se que a resposta é restrita ou intervalo unitário representando níveis de 0 a 100% incidência ou severidade da doença. Tais modelos podem ser reexpressos sob a formulação adotada aqui. Por exemplo, o modelo logístico é obtido definindo-se a distribuição gaussiana para $d(\cdot)$ e *logit* como a função de ligação $f(\cdot)$ adequada. Entretanto, a formulação adotada aqui permite ampliar largamente as opções de modelos para curvas de progressos de doenças e respeitando as restrições nos valores da resposta na obtenção de curvas estimadas e suas bandas de confiança.

Os 30 modelos considerados aqui foram ajustados para cada um dos cultivares e optamos por apresentar os valores das log-verossimilhanças na Figura 4.

Chama atenção nos resultados o fato de que os modelos não lineares gaussianos, que são largamente utilizados na prática, terem um comportamento claramente inferior às demais alternativas. Isso possivelmente se deve à suposição de homoscedasticidade, inadequada para dados dessa natureza. Quando a grande maioria das observações encontra-se ao redor de $0,5$ como no caso dos exemplos anteriores, a suposição de homoscedasticidade é menos importante. Porém, quando os dados cobrem todo o intervalo os resultados são mais sensíveis à esta suposição, o que explica as diferenças no ajuste.

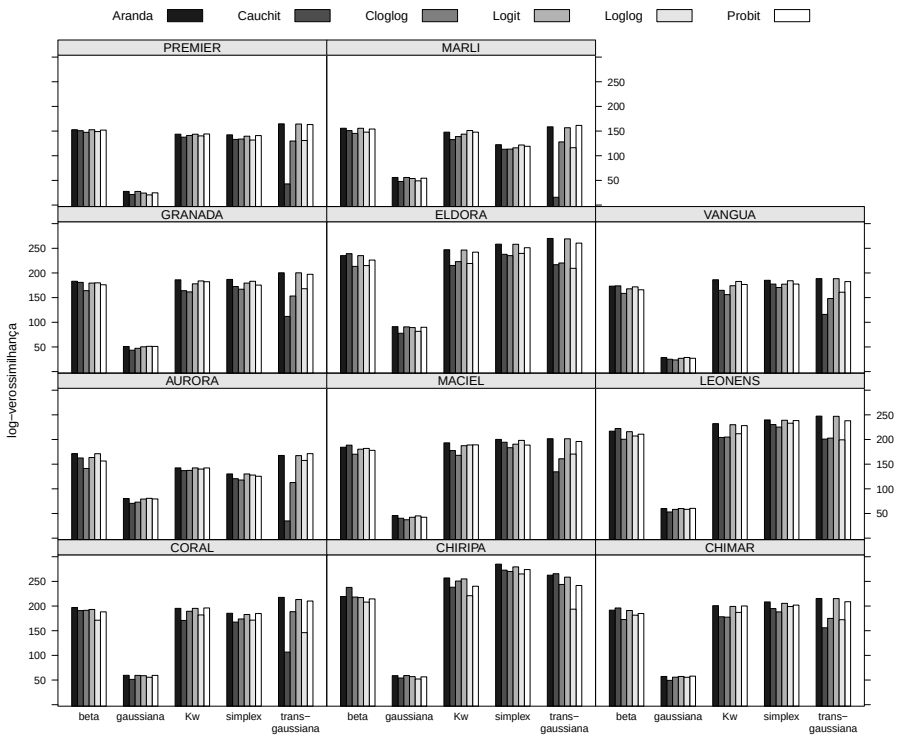

Figura 4 - Valores da log-verossimilhança das diversas especificações dos modelos, para cada cultivar.

Para as demais distribuições os valores da log-verossimilhança variam pouco e podemos afirmar que a escolha da função de ligação, que define a curva de progresso da doença, é de menor importância. A ressalva é para os modelos transformados, mais afetados pela escolha de $T(\cdot)$, sendo que, a transformação *cauchit*, que se destacou positivamente nos exemplos anteriores, mostrou ser aqui a menos indicada.

A transformação Aranda-Ordaz, que dentre as funções apresentadas deveria ser a mais flexível por contar com um parâmetro extra, não mostrou um desempenho uniformemente superior às demais funções de ligação. De forma geral, sempre uma das outras opções foi capaz de fornecer um ajuste de qualidade similar, ainda que com um parâmetro a menos.

Comparando as diferentes suposições distribucionais não é possível identificar nenhuma que se destaque como sendo a de melhor ajuste para a maioria dos conjuntos de dados. As diferenças não são de grande magnitude. Destacamos que, apesar de pouco utilizado, o modelo com dados transformados aprsentou melhor ajuste para vários conjuntos de dados. O modelo transformado seria o escolhido em 9 das 11 cultivares analisadas ainda que pequenas as diferenças em log-verossimilhança com outros modelos.

Os ajustes apresentam estimativas pontuais e erros padrões próximos para todas as especificações sem diferenças que impactem nas conclusões práticas das análises.

Conclusões

Este artigo apresenta uma forma geral para especificação de modelos de regressão para variáveis aleatórias no intervalo unitário. A estrutura apresentada mostrou-se bastante flexível e incluindo como casos particulares uma diversidade de modelos apresentados na literatura.

Análises são feitas sob uma abordagem comum para a especificação de modelos e estimação dos parâmetros. O método de máxima verossimilhança é sempre utilizado para a estimação de parâmetros e construção de intervalos de confiança, sejam eles assintóticos ou perfilhados. Os valores da log-verossimilhança maximizada são usados em cada análise para comparar a adequação de cada um dos modelos.

A implementação computacional foi feita de forma genérica podendo ser prontamente estendida para outras escolhas dos componentes do modelo. Os algoritmos numéricos utilizados (BFGS e SANN) mostraram-se satisfatórios para encontrar as estimativas de máxima verossimilhança, bem como, seus respectivos erros padrões e verossimilhanças perfilhadas.

Os modelos transformados são os únicos que dispõem de forma fechada para os estimadores de máxima verossimilhança, sendo portanto os mais simples de serem ajustados, podendo até mesmo serem estimados pelo método de mínimos quadrados ordinários, seja qual for a função de transformação utilizada. Isso é uma vantagem sobre as demais construções, já que, estas exigem o uso de algoritmos de maximização numérica que podem enfrentar problemas de convergência, exigindo cuidados no seu uso. Por outro lado, para a obtenção de valores preditos da

variável resposta na escala original sob tais modelos, é necessário o uso de integração numérica ou métodos baseados em simulação. Este último passo ainda é tipicamente mais barato computacionalmente do que o uso de algoritmos de maximização numérica.

Os resultados para os quatro conjuntos de dados mostram que as especificações de modelos levaram a diferenças substanciais em termos de log-verossimilhança, com diferenças por vezes maiores que 10 unidades. Em três dos quatro casos examinados, as diferenças são mais substanciais entre as diferentes suposições distribucionais. A escolha da função de ligação do parâmetro parece ser de segunda ordem, no sentido de que se a suposição distribucional for adequada a escolha função de ligação do parâmetro terá menor impacto.

O mesmo não de aplica aos modelos transformados, que mostram ser mais sensíveis a escolha da função de transformação. Esperava-se que a função Aranda-Ordaz seja como ligação do parâmetro ou função de transformação, assumisse formas comparáveis com as demais, levando a melhores ou pelo menos bons ajustes em todos os casos. Isto não se confirmou nas análises para as quais outras funções mostraram ajustes melhores ou equivalentes, mesmo com um parâmetro a menos.

A função *cauchit* teve um bom desempenho como escolha para função de ligação, para as três primeiras análises apresentadas, sugerindo que seja considerada quando se busca um melhor ajuste, ainda que funções de ligação como a logit levem a interpretações mais simples.

Em todas as análises de dados apresentadas a interpretação dos efeitos das covariáveis não foi afetada pela especificação do modelo, ainda que verificadas diferenças nas medidas de ajuste. As diferenças nos valores das estimativas pontuais ou dos erros padrão associados foram de pouca relevância para o entendimento geral do fenômeno estudado. Sob esta ótica há vantagens na adoção de modelos transformados, que levam a resultados semelhantes aos das demais especificações, porém com a simplicidade computacional associada à distribuição gaussiana, sendo o único caso que os estimadores estão disponíveis de forma analítica. As mesmas interpretações práticas acerca das covariáveis puderam ser obtidas com o menor esforço computacional. Isto sugere que tais modelos devem ser usados ao menos em estágios iniciais das análises em situações que requerem-se um grande volume de análises.

Possíveis tópicos para pesquisas futuras incluem a realização de estudos de simulação para verificar se os resultados obtidos aqui com a análises de dados reais se confirma quando temos controle sobre o mecanismo gerador de dados. Isto permitiria identificar situações em que haja clara vantagem em adotar um modelo com suposição distribucional diferente da gaussiana composta com as diversas funções de transformação aqui apresentadas. Outro possível tópico inclui verificar se estes resultados se mantém sob inclusão de efeitos aleatórios.

Agradecimentos

Ao IPPUC (Instituto de Pesquisa e Planejamento Urbano de Curitiba) pelos os

dados de Índice de Qualidade de Vida Intraurbana de Curitiba. À Sonia Beraldi de Magalhães da regional Paraná e Milton Matos de Souza do Departamento Nacional do Serviço Social da Indústria (SESI) pelos dados de Índices de Qualidade de Vida dos Trabalhadores da Indústria. A Giselda Alves e Larissa May de Mio pelos dados da ferrugem do pessegueiro. A dois revisores anônimos pelos comentários que contribuiram para aprimorar o texto.

BONAT, W. H.; RIBEIRO Jr, P. J.; ZEVIANI, W.M. Regression models for responses in the unit interval: specification, estimation and comparison. *Rev. Mat. Estat.,* São Paulo, v.xx, n.x, p.xx-xx, 2013. *Rev. Bras. Biom.* (São Paulo), v. 20, n.1, p. 1-10, 2013.

■ *ABSTRACT: Regression models are widely used on a diversity of application areas to describe associations between explanatory and response variables. The initially and most adopted form of Gaussian linear models was gradually extended to accomodate different kinds of response variables. Several of such models were described as particular cases of the class of the generalized linear models (GLM) which allows, under the same framework, a diversity of the formats for the response variable and functions linking the parameters of the distribution to a linear predictor. Since then GLM's structure became benchmark for several further extensions and developments is statistical modelling such as generalised additive models, overdispersion, among others. Response variables with values restricted to an interval most often* $(0, 1)$*, are usual in social sciences, agronomy, psicometrics among other areas. The Beta or the Simplex distributions are often adopted although other options are mentioned in the literature. A set of regression models for restriced response variables which includes the usualy adopted formats and also allows for a wider range of models are declared here under a general specification. Individual models are defined by the choices for the three components, the probability distribution for the response, the link function between a parameter of the distribution of choice and the linear predictor and the transformation function for the response. We report results of the analysis for four different datasets considering Beta, Simplex, Kumaraswamy and Gaussian distributions, and logit, probit, complementary log-log, log-log, Cauchit and Aranda-Ordaz as options for the link and transformation funtions. Likelihood based analysis for model fitting, comparison and choice are carried out on a unified way and computer code is made available. Results shows there is no proeminent model within the class illustrating the importance of tools for exploring a wide class of models at each analysis.*

■ *KEYWORDS: maximum likelihood ; restricted variables ; proportions ; indexes ; rates.*

Referências

ALVES, G. *Características fitotécnicas e comportamento de cultivares de pessegueiro em relação à podridão parda e à ferrugem na Lapa/PR*, 2012. Tese Doutorado - UFPR - Universidade Federal do Paraná, Curitiba, 2012.

BELISLE, C. J. P. *Convergence theorems for a class of simulated annealing algorithms on Rd*. Applied Probability, v.29, p.885-895, 1992.

BYRD, R. H. ; LU, P. ; NOCEDAL, J. ; ZHU, C. *A limited memory algorithm for bound constrained optimization*. SIAM - Journal on Scientific Computing, v.16, p.1190-1208, 1995.

BONAT, W. H. ; RIBEIRO Jr, P. J. ; ZEVIANI, W. M. *Likelihood analysis for a class beta mixed models*. Relatório Técnico - LEG, 2013.

BONAT, W. H. ; KRAINSKI, E. T. ; RIBEIRO Jr, P. J. ; ZEVIANI, W. M. *Métodos computacionais para inferência com aplicações em R*. João Pessoa: 20º Simpósio Brasileiro de Probabilidade e Estatística - SINAPE, 2012. 260p.

BOX, G. E. P. ; COX, D. R. *An analysis of transformations*. Journal of the Royal Statistical Society, Series B (Methodological), v.26(2), p.211-252, 1964.

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 and New Zealand Journal of Statistics, v.49(3), p.287-301, 2007.

CEPEDA, C. E. ; GAMERMAN, D. *Bayesian methodology for modeling parameters in the two parameter exponential family*. Psychological Methods, v.57(1), p.93-105, 2005.

CRIBARI-NETO, F. ; ZEILEIS, A. *Beta regression in R*. Journal of Statistical Software, v.34(2), p.1-24, 2010.

Da-SILVA, C. Q. ; MIGON, H. S. ; CORREIA, L. T. *Dynamic Bayesian beta models*. Computational Statistics and Data Analysis, v.55(6), p.2074-2089, 2011.

ESPINHEIRA, P. ; FERRARI, S. ; CRIBARI-NETO, F. *Influence diagnostics in beta regression*. Computational Statistics and Data Analysis, v.52(9), p.4417-4431, 2008a.

ESPINHEIRA, C. Q. ; FERRARI, S. ; CRIBARI-NETO, F. *On beta regression residuals*. Journal of Applied Statistics, v.35(4), p.407-419, 2008b.

FERRARI, S. ; CRIBARI-NETO, F. *Beta regression for modelling rates and proportions*. Journal of Applied Statistics, v.31(7), p.799-815, 2004.

GRÜN, B. ; KOSMIDIS, I. ; ZEILEIS, A. *Extended beta regression in R: Shaken, stirred, mixed, and partitioned*. Journal of Statistical Software, v.48(11), p.1-25, 2012.

GRUNWALD, G. K. ; RAFTERY, A. E. *Times series of continuous proportions*. Journal of the Royal Statistical Society: Series B, v.9(4), p.586-597, 1993.

KIESCHINICK, R. ; McCULLOUGH, B. D. *Regression analysis of variates observed on (0,1): percentages, proportions and fractions*. Statistical Modelling, v.3(3), p.193-213, 2003.

LEMONTE, A. J. ; BARRETO-SOUZA, W. CORDEIRO, G. *The exponentiated Kumaraswamy distribution and its log-transform*. Brazilian Journal of Probability and Statistics, v.27(1), p.31-53, 2013.

LIMA, L.B. *Um teste de especificação correta em modelos de regressão beta*. Dessertação, Universidade Federal de Pernambuco, 2007. 107p.

McCULLAGH, P.; NELDER, J. A. *Generalized linear models*. 2.ed. London: Chapman and Hall, 1989. 511p.

McKENZIE, E. *An autoregressive process for beta random variables*. Management Sciences, v.31(8), p.988-997, 1985.

MIYASHIRO, E. S. *Modelos de regressão Beta e Simplex para a análise de proporções*, 2008. 84p. Dissertação de Mestrado - USP - Universidade de São Paulo, São Paulo, 2008.

NELDER, J. A. ; WEDDERBURN, W. M. *Generalized linear models*. Journal of the Royal Statistical Society. Series A, v.135(3), p.370-384, 1972.

OSPINA, R. ; CRIBARI-NETO, F. ; VASCONCELLOS, K. L. P. *Improved point and interval estimation for a beta regression model*. Computational Statistics and Data Analysis, v.51(2), p.960-981, 2006.

OSPINA, R. CRIBARI-NETO, F. ; VASCONCELLOS, K. L. P. *Erratum: "Erratum to Improved point and interval estimation for a beta regression model"*. Computational Statistics and Data Analysis, v.55(7), p.2445, 2011.

R Development Core Team. 2012. *R: A language and environment for statiscal computing*. R Foundation for Statistical Computing, Vienna, Austria.

ROCHA, A. V. ; SIMAS, A. B. *Influence diagnostics in a general class of beta regression models*. Test, v.20(1), p.95-119, 2010.

SIMAS, A. B. ; BARRETO-SOUZA, W. ; ROCHA, A. V. *Improved estimators for a general class of beta regression models*. Computational Statistics and Data Analysis, v.54(2), p.348-366, 2010.

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

SPÓSITO, M.B. ; BASSANEZI, R.B. ; AMORIM, L. *Resistência à mancha preta dos citros avaliada por curvas de progresso da doença*. Fitopatologia Brasileira, v.29(5), p.532-537, 2004.

VASCONCELLOS, K. L. P. ; CRIBARI-NETO, F. *Improved maximum likelihood estimation in a new class of beta regression models*. Brazilian Journal of Probability and Statistics, v.19, p.13-31, 2005.

VERKUILEN, J. ; SMITHSON, M. *Mixed and misture regression models for continuous bounded responses using beta distribution*. Journal of Educational and Behavioral Statistics, v.37(1), p.82-113, 2012.

Received in 01.01.2013.

Approved after revised in 01.01.2013.