Modelos de regressão para dados de proporção: Beta e Simplex com aplicação ao IDHM 2010

André F. B. Menezes[1], Wesley O. Furriel[1]

[1]Universidade Estadual de Maringá, Departamento de Estatística, DEs, PR, Brasil.

Resumo

O objetivo deste trabalho foi investigar o IDHM de 2010 a partir de alguns indicadores de aspecto socioeconômico e espacial, com intuito de averiguar a existência de relações entre estes. Para tal, foram aplicadas técnicas de estatística descritiva para visualização do comportamento dos dados e os modelos de regressão Beta e Simplex, que são adequados para a modelagem de taxas e proporções. Os resultados permitiram constatar que os modelos apresentaram estimativas e conclusões bastante similares, quando considerado a modelagem da média. Além disso, verificou-se uma considerável conformidade na relação entre os indicadores sociais selecionados e a variável resposta IDHM.
Palavras chave: IDHM, regressão Beta, regressão Simplex.

1 Introdução

O IDHM (Índice de Desenvolvimento Humano Municipal) é uma adaptação do IDH Global para os municípios brasileiros realizada em 2012 pelo Ipea e a Fundação João Pinheiro. Sendo amplamente utilizado para a sumarização, identificação e hierarquização do desenvolvimento humano, o indicador é composto por três dimensões: Longevidade , Educação e Renda. Seu índice varia de 0 a 1, de modo que, quanto mais próximo de 1, melhor o desenvolvimento humano, e quanto mais próximo de 0, pior o desenvolvimento humano.

É de interesse comum, nos mais diversos âmbitos do conhecimento, compreender a relação entre uma variável resposta e possíveis variáveis explicativas (covariáveis), bem como, realizar predições a partir da relação estabelecida. Nesse sentido, foi realizada a modelagem do IDHM do ano de 2010, a partir de algumas variáveis de caráter socioeconômico e espacial, conforme as seguintes : proporção de crianças extremamente pobres; população dos municípios; proporção de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados; taxa de frequência líquida ao ensino superior; mortalidade até um ano de idade; índice de Gini; e as grandes regiões do Brasil, compostas por Norte, Nordeste, Sul, Sudeste, Centro-Oeste. Para a seleção destas variáveis, foram considerados indicadores que não foram empregados diretamente na construção do IDHM, mas que nos permitiram averiguar sua consonância com demais medidas que buscam expressar aspectos da realidade social, como desigualdade, vulnerabilidade infantil, habitação e região.

Dessa forma, tendo em vista o domínio da variável resposta IDHM, restrita ao intervalo contínuo de 0 a 1, foram utilizados os modelos de regressão para dados contínuos, de proporção, Beta e Simplex, pertencentes a família exponencial biparamética. Para atingir os objetivos desejados foi utilizado o banco de dados disponibilizado pelo Atlas do Desenvolvimento Humano, que conta com mais de 200 indicadores socioeconômicos, permitindo qualificar e ampliar a análise do desenvolvimento humano nos municípios brasileiros.

2 Metodologia

A teoria dos modelos lineares generalizados (MLGs) unificou uma classe de distribuições para aplicação em modelos de regressão. Contudo, os MLGs apresentam fortes limitações para variáveis respostas cujo o suporte é limitado a um intervalo (a, b), sendo o intervalo unitário $(0, 1)$ o mais habitual. Neste contexto, diversos autores propuseram alternativas para este tipo especifico de dados. Destaca-se os seguintes

trabalhos: Song e Tan (2000), Cepeda-Cuervo (2001), Kieschnick e McCullough (2003), Ferrari e Cribari-Neto (2004), Bonat, Ribeiro e Zeviani (2013), Cepeda-Cuervo e Garrido (2015) dentre muitos outros. Na sequência apresentamos brevemente os dois modelos de regressão adotados.

2.1 Modelo de regressão Beta

O modelo de regressão Beta inicialmente proposto por Cepeda-Cuervo (2001) e mais tarde introduzido de forma independente por Ferrari e Cribari-Neto (2004) consiste em uma nova parametrização da distribuição Beta indexada por sua média e um parâmetro de precisão. A função densidade de probabilidade da distribuição Beta em sua forma reparametrizada é definida por:

$$f(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \qquad 0 < y < 1 \tag{1}$$

em que $0 < \mu < 1$ denota a esperança e $\phi > 0$ o parâmetro de precisão.

Considerando a reparametrização da distribuição Beta em função de μ e ϕ podemos definir o modelo de regressão Beta com modelagem conjunta dos parâmetros média e precisão (CEPEDA-CUERVO, 2001). De forma geral, assume-se uma amostra aleatória $Y_i \sim \text{Beta}(\mu_i, \phi_i), i = 1, \ldots, n$, onde ambos os parâmetros média e precisão seguem a estrutura de regressão definida por:

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{2}$$
$$h(\phi_i) = \mathbf{z}_i^\top \boldsymbol{\gamma}, \tag{3}$$

em que , $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^\top$ e $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_k)^\top$ são, respectivamente, os vetores de parâmetros associados a média e precisão do modelo de regressão, $\mathbf{x}_i$ e $\mathbf{z}_i$ são os vetores de covariáveis associados a média e precisão do modelo na i-ésima observação, respectivamente e $g(\cdot)$ e $h(\cdot)$ são funções de ligações apropriadas, duas vezes diferenciáveis em relação aos parâmetros de regressão.

2.2 Modelo de regressão Simplex

Proposta por Barndorff-Nielsen e Jørgensen (1991) a distribuição Simplex também é parametrizada por sua esperança e um parâmetro extra de dispersão. Sua função densidade de probabilidade pode ser expressa por:

$$f(y \mid \mu, \phi) = \left[2\,\pi\,\phi^2\{y\,(1-y)\}^3\right]^{-1/2} \exp\left\{-\frac{1}{2\,\phi^2}\left[\frac{(y-\mu)^2}{y\,(1-y)\,\mu^2(1-\mu)^2}\right]\right\}, \qquad 0 < y < 1 \tag{4}$$

em que $0 < \mu < 1$ e $\phi > 0$. Na realidade, a distribuição Simplex faz parte dos modelos de dispersão propostos por Jørgensen (1997) e que estendem os MLGs.

Assim como no modelo Beta os parâmetros de média e dispersão na regressão Simplex tem estrutura definida nas equações (2) e (3). Uma importante propriedade do modelo de regressão Simplex é que os parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ são ortogonais. A ortogonalidade dos parâmetros fornece muitas vantagens nos resultados inferências, por exemplo, assintoticamente as estimativas de máxima verossimilhança são independentes (COX; REID, 1987).

Ressalta-se que para ambos os modelos as estimativas dos parâmetros foram obtidas seguindo o paradigma da máxima verossimilhança, isto é, pela maximização numérica da função de log-verossimilhança. Para isso as bibliotecas `betareg` (CRIBARI-NETO; ZEILEIS, 2010) e `simplexreg` (ZHANG; QIU; SHI, 2016) do *software* R foram utilizadas. As estimativas também foram conferidas utilizando a *procedure* `NLIMIXED` (SAS, 2010) do SAS.

3 Resultados e Discussões

Para ambos os modelos, Beta e Simplex, assumimos o seguinte modelo de regressão para ambos os parâmetros μ e ϕ:

$$\begin{aligned}\text{logit}(\mu_i) &= \beta_0 + \beta_1\text{PINDCRI}_i + \beta_2\text{AGUAESGOTO}_i + \beta_3\text{TFLSUPER}_i + \beta_4\text{MORT1}_i + \beta_5\text{GINI}_i \\ &+ \beta_6\text{POP}_i + \beta_7\text{CENTROESTE}_i + \beta_8\text{NORDESTE}_i + \beta_9\text{NORTE}_i + \beta_{10}\text{SUDESTE}_i \end{aligned} \tag{5}$$

$$\begin{aligned}\log(\phi_i) &= \gamma_0 + \gamma_1\text{PINDCRI}_i + \gamma_2\text{AGUAESGOTO}_i + \gamma_3\text{TFLSUPER}_i + \gamma_4\text{MORT1}_i + \gamma_5\text{GINI}_i \\ &+ \gamma_6\text{POP}_i + \gamma_7\text{CENTROESTE}_i + \gamma_8\text{NORDESTE}_i + \gamma_9\text{NORTE}_i + \gamma_{10}\text{SUDESTE}_i, \end{aligned} \tag{6}$$

$i = 1, \ldots, 4063$.

É importante destacar que quanto as grandes regiões, foi selecionada a região Sul como referência, na construção da matriz de variáveis *dummy* empregada no modelo.

Na Tabela 1 apresentamos, para ambos os modelos, as estimativas, os erros padrão assintóticos e valor-p do teste para os coeficientes do modelo. Antes de seguir para a análise dos resultados é preciso ressaltar que, o parâmetro γ, foi estimado em escalas distintas, tendo em vista a parametrização dos modelos. Desse modo, no caso do modelo Beta temos as estimativas para o parâmetro de precisão e na Simplex do parâmetro de dispersão.

Tabela 1: Estimativas dos parâmetros, erro padrão, valor-p para os modelos Beta e Simplex

Parâmetro	Beta			Simplex		
	Estimativas	E.P.	Valor-p	Estimativas	E.P.	Valor-p
β_0	0.6767	0.0189	< 0.0001	0.6775	0.0189	< 0.0001
β_1(PINDCRI)	-0.9643	0.0226	< 0.0001	-0.9555	0.0226	< 0.0001
β_2(AGUAESGOTO)	-0.0211	0.0194	0.2769	-0.0228	0.0194	0.2397
β_3(TFLSUPER)	1.5176	0.0397	< 0.0001	1.5129	0.0397	< 0.0001
β_4(MORT1)	-1.1895	0.0429	< 0.0001	-1.1727	0.0427	< 0.0001
β_5(GINI)	0.4639	0.0363	< 0.0001	0.4584	0.0364	< 0.0001
β_6(POP)	0.4803	0.0423	< 0.0001	0.5499	0.0470	< 0.0001
β_7(CENTROESTE)	-0.0238	0.0068	0.0005	-0.0254	0.0068	0.0002
β_8(NORDESTE)	-0.0162	0.0082	0.0472	-0.0213	0.0082	0.0093
β_9(NORTE)	-0.0760	0.0113	< 0.0001	-0.0807	0.0113	< 0.0001
β_{10}(SUDESTE)	-0.0082	0.0059	0.1668	-0.0096	0.0060	0.1094
γ_0	5.6994	0.2407	< 0.0001	-2.7466	0.2412	< 0.0001
γ_1(PINDCRI)	-0.5464	0.3037	0.0720	0.1447	0.3044	0.6344
γ_2(AGUAESGOTO)	-0.7373	0.2603	0.0046	0.7583	0.2606	0.0036
γ_3(TFLSUPER)	-0.3060	0.4845	0.5277	1.5428	0.4856	0.0015
γ_4(MORT1)	0.8896	0.6173	0.1496	-1.3817	0.6181	0.0254
γ_5(GINI)	0.1665	0.4547	0.7143	0.1218	0.4560	0.7894
γ_6(POP)	-1.0074	0.1137	< 0.0001	1.9907	0.1409	< 0.0001
γ_7(CENTROESTE)	0.5658	0.0949	< 0.0001	-0.6030	0.0950	< 0.0001
γ_8(NORDESTE)	0.4380	0.1050	< 0.0001	-0.5140	0.1053	< 0.0001
γ_9(NORTE)	0.0315	0.1367	0.8176	-0.1243	0.1370	0.3643
γ_{10}(SUDESTE)	-0.1487	0.0681	0.0291	0.1296	0.0683	0.0576

Diante dos resultados da Tabela 1 várias conclusões podem ser retiradas. Por questões de espaço destacamos as seguintes. No que tange a modelagem das médias é possível verificar que ambos os modelos apresentam estimativas bastante próximas, bem como, as mesmas conclusões. Dessa forma, das covariáveis selecionadas apenas as estimativas de β_2 e β_{10} não foram significativas a um nível de significância nominal de 5%. Além disso, nota-se que as estimativas para β_1 e β_4 apresentaram relações negativas com o IDHM, enquanto β_3, β_5 e β_6 positivas, isto é, elevações nestas variáveis implicam em aumento no IDHM. Quanto as regiões, constatamos que quando comparadas ao Sul, todas as regiões apresentam IDHM inferiores.

No que diz respeito aos parâmetros de precisão e dispersão, não é possível comparar as estimativas. Porém, é possível comparar as conclusões acerca dos resultados, isto é, a rejeição da hipótese $H_0 : \gamma_k = 0$. Assim, constatamos que para as estimativas dos parâmetros γ_3, γ_4 e γ_{10} ocorreram diferenças ao rejeitar H_0, considerando-se, um nível de significância de 5%.

Por fim, ressalta-se que a relação entre os indicadores socioeconômicos e o IDHM se mostrou consonante, ou seja, indicadores que apontaram para melhores condições sociais e econômicas, se mostraram positivamente relacionado ao IDHM.

A fim de avaliar o ajuste de ambos os modelos os resíduos padronizado ponderados foram utilizados (ver Figura 1). Para o modelo de regressão Beta este resíduo foi proposto por Espinheira, Ferrari e Cribari-Neto (2008), já sob a suposição de distribuição simplex foi Miyashiro (2008) que o introduziu. Finalmente, para discriminar os modelos o AIC e o teste de Voung (VUONG, 1989) foram empregados. Os valores obtidos para o AIC são -18518.31 e -18499.35 para os modelos Beta e Simplex, respectivamente. Corroborando com o critérios de informação, a hipótese nula do teste de Voung, de que não existe diferenças significativa do ajuste entre o modelo Simplex e Beta foi rejeitada, favorecendo a escolha do modelo Simplex.

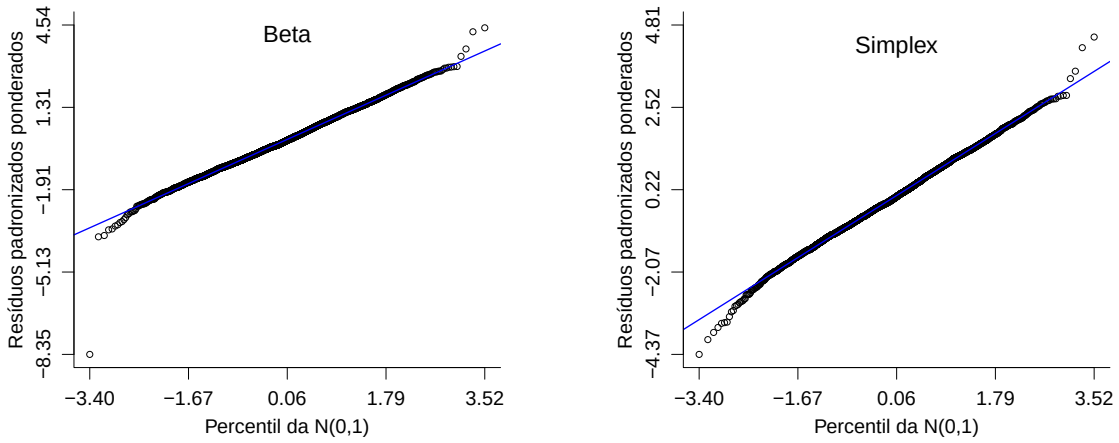

Figura 1: Gráficos dos resíduos padronizado ponderados versus percentil da Normal padrão.

Referências

BARNDORFF-NIELSEN, O.; JØRGENSEN, B. Some parametric models on the Simplex. *Journal of Multivariate Analysis*, v. 39, n. 1, p. 106–116, 1991. ISSN 0047-259X.

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

CEPEDA-CUERVO, E. *Variability modeling in generalized linear models*. Tese (Doutorado) — Mathematics Institute, Universidade Federal do Rio de Janeiro, 2001.

CEPEDA-CUERVO, E.; GARRIDO, L. Bayesian Beta regression models with joint mean and dispersion modeling. *Monte Carlo Methods and Applications*, v. 21, n. 1, p. 49–58, 2015.

COX, D. R.; REID, N. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 49, n. 1, p. 1–39, 1987.

CRIBARI-NETO, F.; ZEILEIS, A. Beta regression in R. *Journal of Statistical Software*, v. 34, n. 2, p. 1–24, 2010. ISSN 1548-7660.

ESPINHEIRA, P. L.; FERRARI, S. L. P.; CRIBARI-NETO, F. On Beta regression residuals. *Journal of Applied Statistics*, Taylor & Francis, v. 35, n. 4, p. 407–419, 2008.

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

JØRGENSEN, B. *The Theory of Dispersion Models*. [S.l.]: Chapman & Hall/CRC, 1997.

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Statistical Modelling*, v. 3, n. 3, p. 193–213, 2003.

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

SAS. *The NLMIXED Procedure, SAS/STAT® User's Guide, Version 9.4*. Cary, NC: SAS Institute Inc.: [s.n.], 2010. 4967–5062 p.

SONG, P. X.-K.; TAN, M. Marginal models for longitudinal continuous proportional data. *Biometrics*, v. 56, n. 2, p. 496–502, 2000.

VUONG, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, [Wiley, Econometric Society], v. 57, n. 2, p. 307–333, 1989.

ZHANG, P.; QIU, Z.; SHI, C. simplexreg: An R package for regression analysis of proportional data using the Simplex distribution. *Journal of Statistical Software*, v. 71, n. 11, p. 1–21, 2016.