Universidade Federal de Minas Gerais
Instituto de Ciências Exatas e da Terra
Departamento de Estatística

MODELAGEM LINEAR GENERALIZADA COM EFEITO ESPACIAL: UM ESTUDO COM DADOS EDUCACIONAIS DO ESTADO DE MINAS GERAIS

Maria Luisa Gomes dos Reis

Orientador: Prof. Vinícius Diniz Mayrink

INTRODUÇÃO

- Aprimoramento do projeto de Iniciação Científica
  - Modelagem Linear Generalizada: um estudo com dados educacionais do estado de Minas Gerais
- Taxas de abandono escolar
  - Modelos Lineares Generalizados

$$\frac{\text{ABANDONARAM}}{\text{APROVADOS} + \text{REPROVADOS} + \text{ABANDONARAM}} \times 100\% = \text{TAXA DE ABANDONO}$$

- Estudo limitado a Minas Gerais
  - Estatística Espacial ➜ Estatística Bayesiana

INTRODUÇÃO

Anos escolares agrupados em blocos de ensino

- Ensino Fundamental I (EFI): 1º ao 5º ano
- Ensino Fundamental II (EFII): 6º ao 9º ano
- Ensino Médio (EM): 1ª à 3ª série

- Taxa de escolas públicas e taxa de escolas na área urbana;
- Obtenção de covariáveis explicativas referentes à cada município.

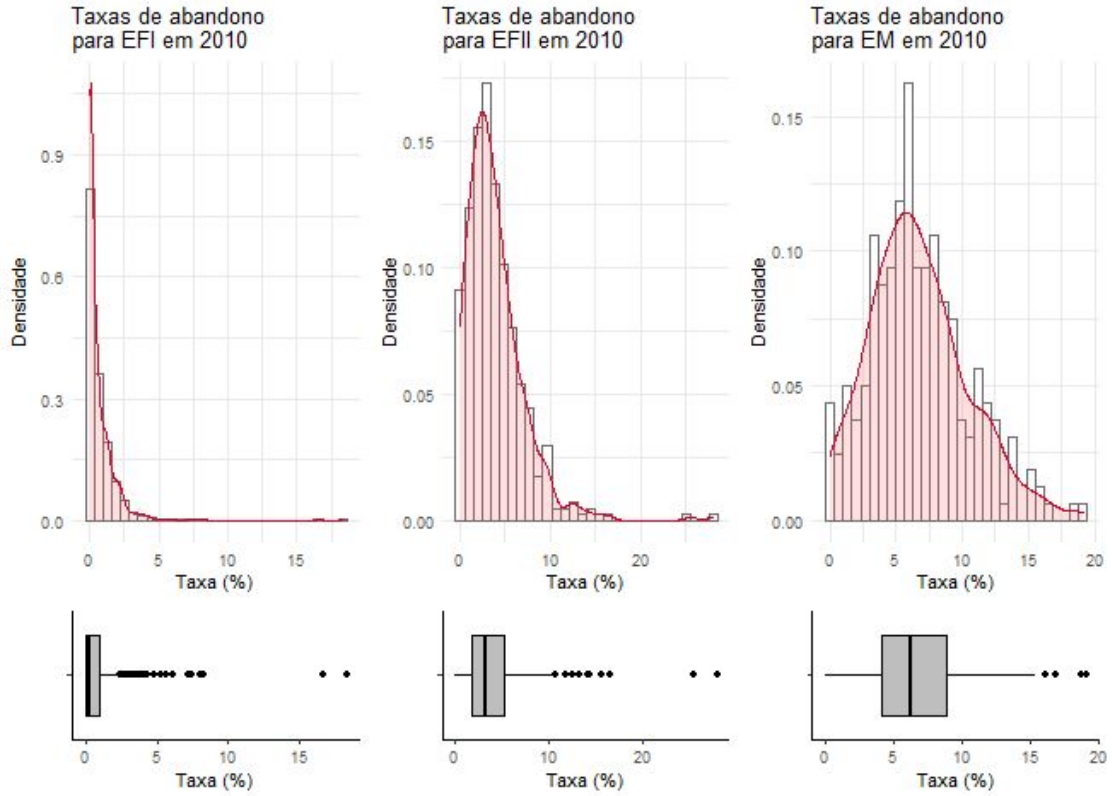
Objetivo: avaliar quais fatores influenciam na taxa de abandono em Minas Gerais

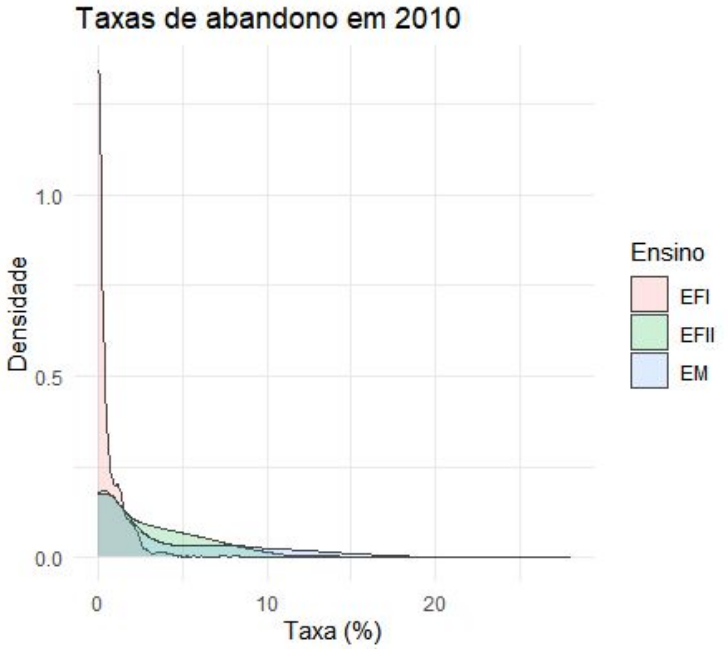
ANÁLISE DESCRITIVA

- Considerou-se três anos ao longo da mesma década;
- Foram considerados municípios com pelo menos 3 escolas válidas;
  - Escolas válidas apresentam taxas para pelo menos metade das séries escolares do bloco de ensino correspondente

Ano	Ensino	N° de municípios considerados	N° médio de escolas por município	N° médio de escolas válidas por município	N° total de escolas	N° total de escolas válidas
2020	EFI	617	12,4	9,1	10.540	7.760
	EFII	419	12,4	5,4	10.540	4.647
	EM	272	12,4	2,8	10.540	2.417
2015	EFI	652	13,2	10,2	11.222	8.664
	EFII	425	13,2	5,5	11.222	4.714
	EM	256	13,2	2,7	11.222	2.271
2010	EFI	692	14,8	11,9	12.608	10.186
	EFII	419	14,8	5,5	12.608	4.663
	EM	243	14,8	2,5	12.608	2.101

ANÁLISE DESCRITIVA





METODOLOGIA

- Suporte dos dados sugere uso de MLGs Gama ou Beta
- Estudo simulado para decidir sobre qual modelo aplicar aos dados reais
- Reparametrização das distribuições sob estudo:

$$Y \sim Gamma(\boldsymbol{\alpha},\boldsymbol{\delta}) \rightarrow \theta = \frac{\alpha}{\delta}$$

$$Y \sim Gamma(\boldsymbol{\theta},\boldsymbol{\alpha})$$

$$f(y,\theta) = \frac{\left(\frac{\alpha}{\theta}\right)^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\frac{\alpha}{\theta}y}, \quad y \geq 0,$$

$$E(Y) = \theta \qquad Var(Y) = \frac{\theta^2}{\alpha}$$

$$Y \sim Beta(\boldsymbol{\alpha},\boldsymbol{\beta}) \rightarrow \theta = \frac{\alpha}{\alpha+\beta} \qquad \zeta = \alpha + \beta$$

$$Y \sim Beta(\boldsymbol{\theta},\boldsymbol{\zeta})$$

$$f(y,\theta) = \frac{\Gamma(\zeta)}{\Gamma(\theta\zeta)\Gamma[(1-\theta)\zeta]} y^{\theta\zeta-1}(1-y)^{(1-\theta)\zeta-1},$$

$$0 \leq y \leq 1$$

$$E(Y) = \theta \qquad Var(Y) = \frac{\theta(1-\theta)}{1+\zeta}$$

METODOLOGIA - Distribuições *a priori*

- Coeficientes → **Normal multivariada**

$$m_\beta = m_\gamma = 0 \qquad S_\beta = S_\gamma = \begin{pmatrix} 10 & 0 & \cdots & 0 \\ 0 & 10 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 10 \end{pmatrix}$$

- Parâmetros de dispersão/ruído e variância do efeito espacial → **Gamma**

$$a_\tau = a_{\tau,\theta} = a_{\tau,\zeta} = b_\tau = b_{\tau,\theta} = b_{\tau,\zeta} = 0{,}1$$

- Efeitos aleatórios espaciais → **Normal multivariada**

$$m_\Delta = m_{\Delta,\theta} = m_{\Delta,\zeta} = 0$$

S_Δ, $S_{\Delta,\theta}$ e $S_{\Delta,\zeta}$ = **matrizes de vizinhança**

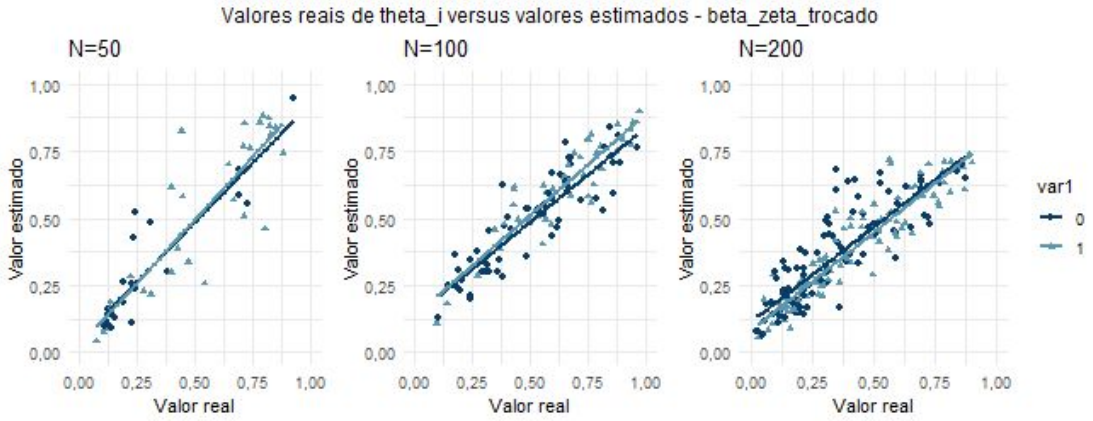
METODOLOGIA

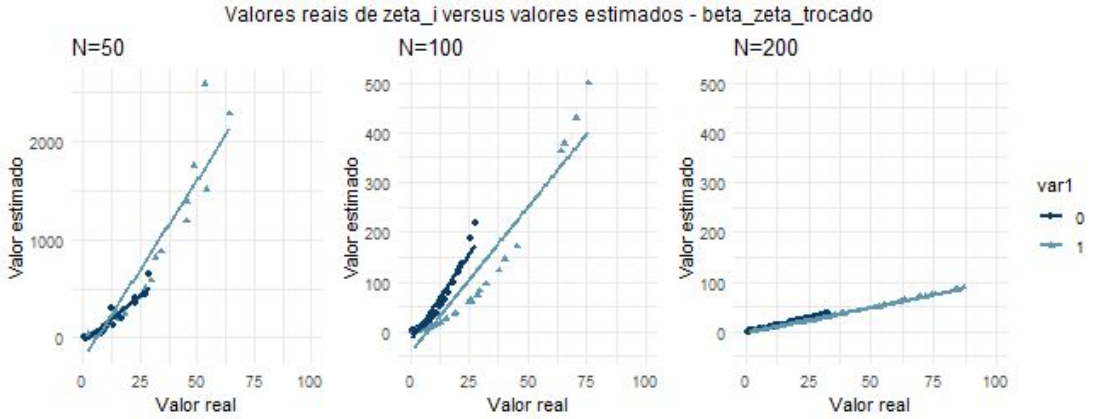
- Ajustou-se os dados de acordo com a estrutura que os gerou
- Modelos trocados para avaliar o impacto da má especificação do modelo:
  - ***beta zeta trocado*** ➔ dados beta_zeta ajustados com o modelo beta_zeta_delta
  - ***beta zeta delta trocado*** ➔ dados beta_zeta_delta ajustados com o modelo beta_zeta

Modelo/dados	Estrutura da média com covariáveis	Estrutura da dispersão com covariáveis	Parâmetro ruído/dispersão	Efeito espacial na estrutura da média	Efeito espacial na estrutura da dispersão
gama	x		x	x	
beta	x		x	x	
beta_zeta	x	x		x	
beta_zeta_delta	x	x		x	x

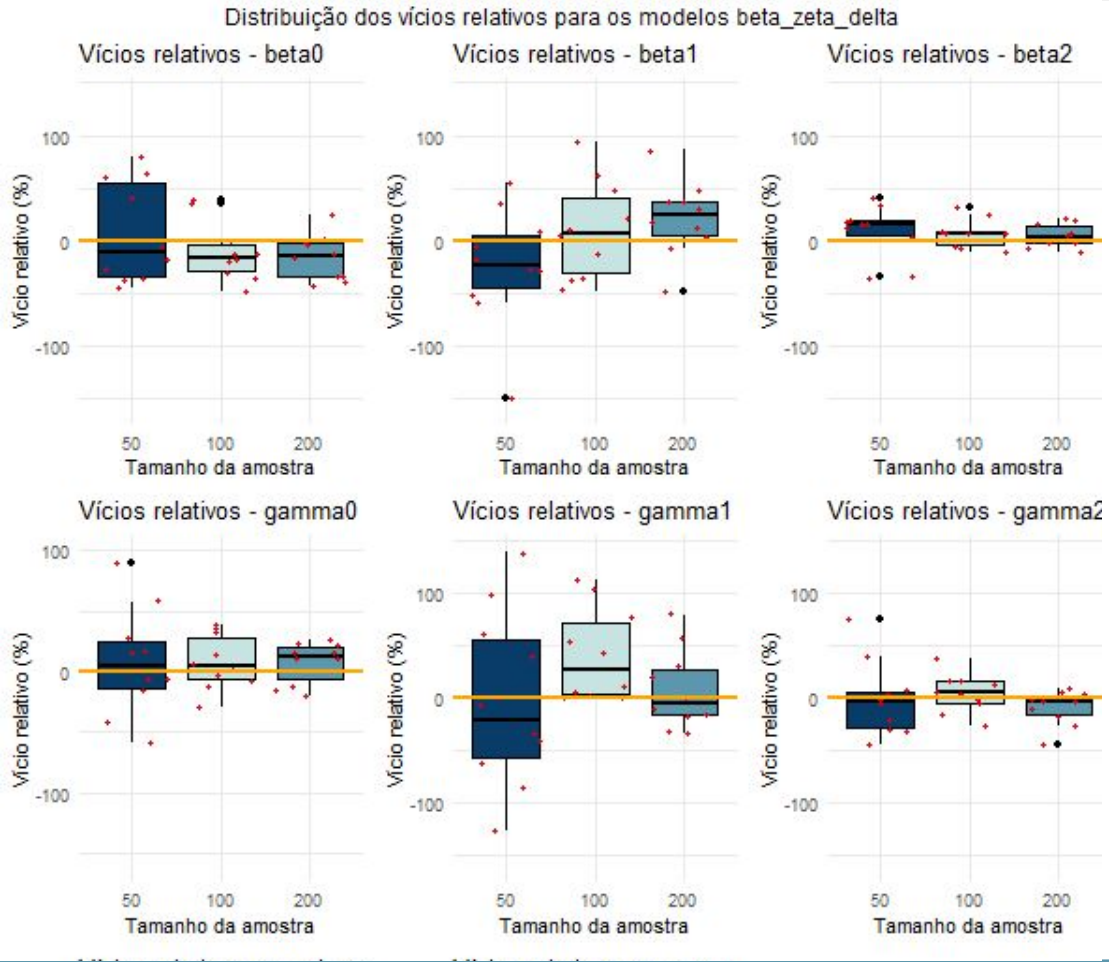
ESTUDO SIMULADO

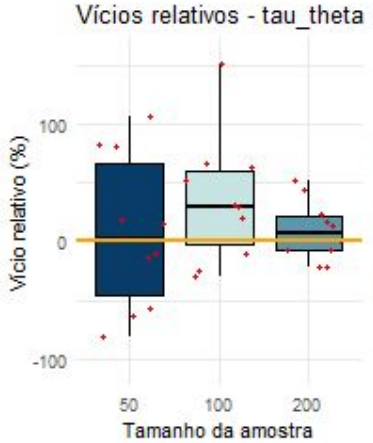
- Bom ajuste para os dados simulados;
- Aumento do tamanho amostral fornece estimativas mais acuradas;
- Ajustar dados trocados indicou que dados sem estrutura espacial podem ser modelados considerando essa estrutura.

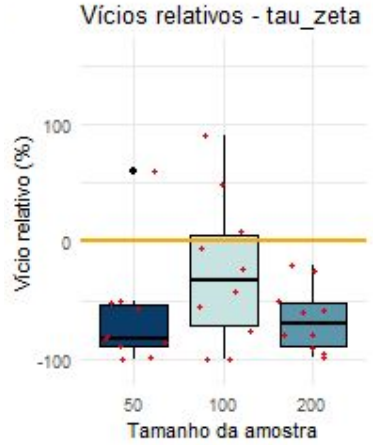




ESTUDO SIMULADO







ESTUDO SIMULADO

- Ajustou-se os dados beta_zeta_delta sem qualquer estrutura espacial;
    - Teste I de Moran indicou associação espacial nos resíduos.

Amostra	Valor-p
N=50	< 0,01
N=100	< 0,01
N=200	< 0,01

- Ajustar uma estrutura espacial não presente nos dados é menos prejudicial que ignorar uma estrutura existente.

Optou-se por ajustar as taxas de abandono com o modelo beta zeta delta

APLICAÇÃO REAL - Amostras originais

- Coeficientes β_i significativos coincidiram para a maioria dos anos:
  - Taxa de escolas públicas contribui para o aumento da taxa de abandono;
- Coeficientes γ_i significativos não coincidiram:
  - Apenas em 2015 e 2020 a taxa de escolas na área urbana contribui significativamente para o aumento da dispersão.

Amostra original

Ano	Coeficiente	Variável	Média	D.P.	Intervalo HPD
2020	β_i	Intercepto	-1,72	2,049	(-5,603; 2,492)
		Taxa urbana	-0,755	0,473	(-1,699; 0,136)
		Taxa pública	1,983	0,477	**(1,068; 2,900)**
		IDHM Renda	-2,11	2,426	(-6,733; 2,658)
		log(População)	0,052	0,144	(-0,212; 0,339)
		log(PIB per capita)	-0,167	0,258	(-0,659; 0,331)
	γ_i	Intercepto	1,654	2,141	(-2,297; 5,940)
		Taxa urbana	1,675	0,536	**(0,656; 2,766)**
		Taxa pública	0,347	0,500	(-0,618; 1,319)
		IDHM Renda	-1,435	2,579	(-6,314; 3,612)
		log(População)	0,096	0,164	(-0,242; 0,397)
		log(PIB per capita)	-0,035	0,289	(-0,577; 0,545)

APLICAÇÃO REAL - Amostras sem taxa zero

- Na maioria dos anos, a taxa de escolas na área urbana contribui para diminuição da média da taxa de abandono;
- Taxa de escolas públicas contribui para o aumento da média da taxa de abandono;
- Coeficientes γ_i significativos não coincidiram, apresentando sinais opostos para as mesmas covariáveis.

Amostra sem taxas zero

Ano	Coeficiente	Variável	Média	D.P.	Intervalo HPD
2020	β_i	Intercepto	-2,841	1,835	(-6,44; 0,711)
		Taxa urbana	-0,988	0,342	**(-1,732; -0,352)**
		Taxa pública	0,333	0,411	(-0,432; 1,137)
		IDHM Renda	0,672	2,321	(-4,003; 5,083)
		log(População)	-0,053	0,106	(-0,251; 0,151)
		log(PIB per capita)	0,015	0,206	(-0,368; 0,42)
	γ_i	Intercepto	0,87	2,315	(-3,784; 5,085)
		Taxa urbana	0,54	0,56	(-0,501; 1,669)
		Taxa pública	1,06	0,542	**(0,044; 2,129)**
		IDHM Renda	-0,195	2,857	(-5,653; 5,48)
		log(População)	0,165	0,169	(-0,158; 0,501)
		log(PIB per capita)	-0,056	0,312	(-0,694; 0,514)

APLICAÇÃO REAL

- A variabilidade dos efeitos aleatórios espaciais diminuiu para as amostras sem as taxas zero.
    - Quanto maior os valores τ_θ e τ_ζ estimados, maior dispersão e variabilidade dos efeitos aleatórios espaciais estimados.

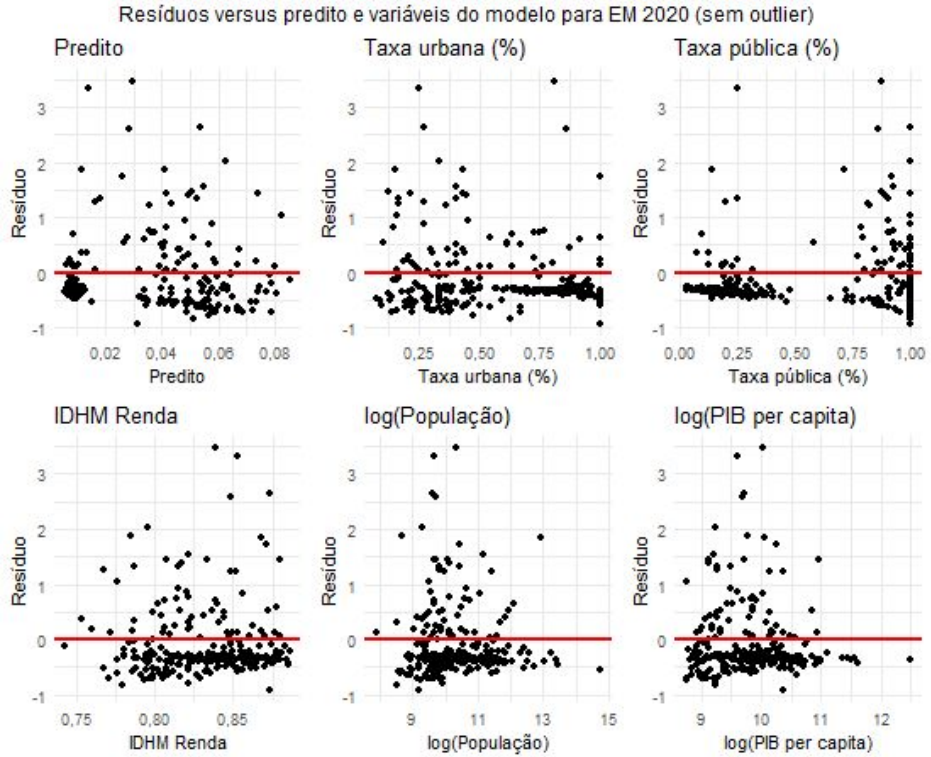
Amostras originais

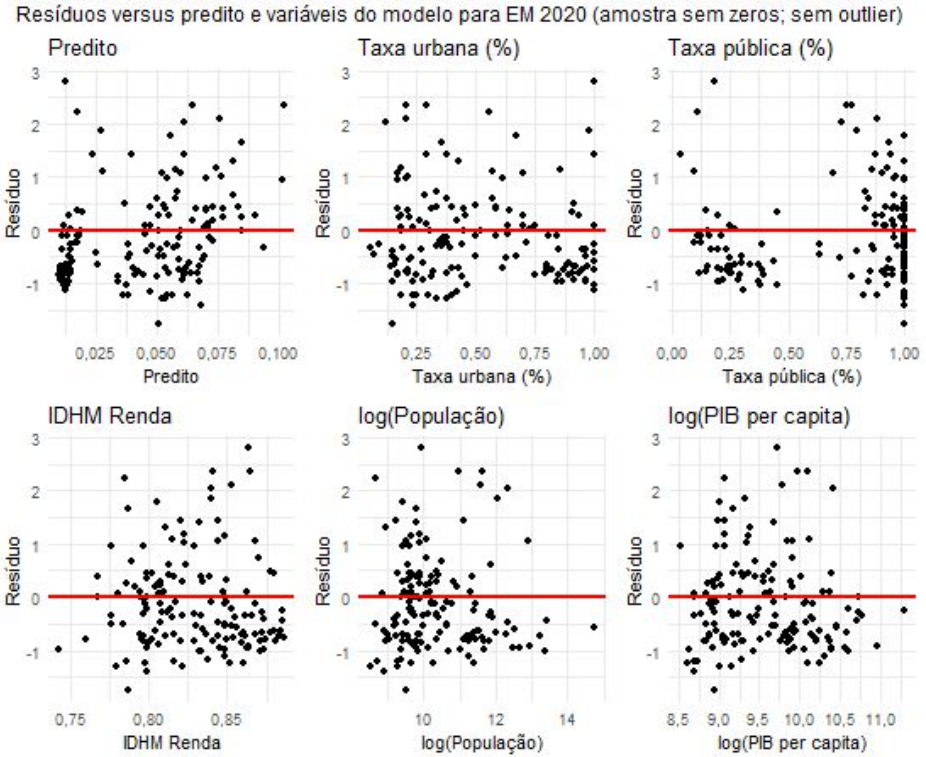
Ano	Parâmetro	Média	D.P.
2020	τ_θ	0,140	0,191
	τ_ζ	1,019	0,964
2015	τ_θ	0,175	0,178
	τ_ζ	2,896	1,028
2010	τ_θ	0,060	0,084
	τ_ζ	0,174	0,296

Amostras sem taxas zero

Ano	Parâmetro	Média	D.P.
2020	τ_θ	0,115	0,217
	τ_ζ	0,106	0,397
2015	τ_θ	1,151	0,852
	τ_ζ	0,602	1,118
2010	τ_θ	0,172	0,191
	τ_ζ	0,391	0,672

APLICAÇÃO REAL - Resíduos





APLICAÇÃO REAL - Resíduos

- Teste I de Moran foi aplicado aos resíduos de Pearson dos modelos;
- Não há evidência de associação espacial nos resíduos:
  - Modelos foram capazes de modelar a estrutura espacial dos dados.

Amostra	Valor-p
EM 2020	0,877
EM 2015	0,316
EM 2010	0,202
EM 2020 (sem zeros)	0,736
EM 2015 (sem zeros)	0,382
EM 2010 (sem zeros)	0,598

CONCLUSÃO

- Ao longo da década passada, os fatores que influenciaram na média da taxa de abandono escolar para o Ensino Médio dos municípios de Minas Gerais foram:
  - Taxa de escolas na rede pública de ensino ➜ **aumenta a taxa de abandono;**
  - Taxa de escolas na área urbana do município ➜ **diminui a taxa de abandono.**

- Fatores influentes nas taxas de abandono escolar em Minas Gerais não mudaram drasticamente ao longo dos anos estudados.

REFERÊNCIAS

Bolstad, W. M. (2007) **Introduction to Bayesian Statistics**, 2ed. John Wiley and Sons.

Banerjee, S., Carlin, B.P., Gelfand, A.E. (2014) **Hierarchical Modeling and Analysis for Spatial Data**. 2 ed. Chapman and Hall/CRC.

Dobson, A. J., Barnett, A. G. (2008) **An Introduction to Generalized Linear Models**, 3rd ed., Boca Raton: Chapman & Hall/CRC.

Ferrari, S., & Cribari-Neto, F. (2004). **Beta regression for modelling rates and proportions**. *Journal of applied statistics*, *31*(7), 799-815.

R Core Team (2021). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stan Development Team (2021). **Stan Modeling Language Users Guide and Reference Manual**, 2.18. https://mc-stan.org.

Getis, A., Ord, K. (1992). **The Analysis of Spatial Association by Use of Distance Statistics**. Geographical Analysis. 24. 189 - 206.

Stan Development Team (2020). **RStan: the R interface to Stan**. R package version 2.21.2. http://mc-stan.org/.

Jonah Gabry and Rok Cesnovar (2021). **cmdstanr: R Interface to 'CmdStan'**. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.