UNIVERSIDADE FEDERAL DO PARANÁ
CURSO ESTATÍSTICA

CALEB SOUZA GRR -20149072
DENNIS LEÃO GRR - 20160239
LUAN FIORENTIN GRR - 20160219

MODELAGEM DA QUANTIDADE DE MATRÍCULAS NO ENSINO REGULAR NO ESTADO DO PARANÁ

CURITIBA
Junho de 2018

Sumário

1 Resumo

O objetivo deste estudo foi construir um modelo para estimar a quantidade de matrículas no ensino regular no Estado do Paraná, Brasil. A técnica estatística de regressão linear foi utilizada para criar uma modelo matemático que expressase o comportamento da quantidade de matrículas no ensino regular em função de diversas covariáveis obtidas no IPARDES. Dois modelos foram testados, onde o primeiro consistiu no ajuste de um modelo de Regressão Robusta e outro modelo com a variável resposta transformada. O coeficiente de determinação ajustado foi bastante semelhante para ambos os modelos, com valor próximo de 0.13. Os resíduos apresentaram tendenciosidades e com possíveis *outliers*. Os pressupostos básicos de regressão linear não foram atendidos, indicando que essa técnica não é adequada para modelar a variável resposta número de pessoas matriculadas no ensino regular.

2 Introdução

A estatística é um ramo da matemática voltada para a coleta, manipulação e análise de dados, provenientes das mais variadas áreas da ciência. Nesse contexto, é inegável a relevância da estatística na educação brasileira, desde o ensino fundamental (base) até as pesquisas de pós-graduação.

As diversas técnicas estatística permitem diferentes conclusões a respeito dos dados. Entre as principais técnicas, há a análise de regressão linear, que é um método estatístico bastante conhecido pela sua eficácia na análise de relação entre uma ou mais variáveis. Dessa forma, para cada método estatístico há premissas e hipóteses diferentes, as quais devem ser respeitadas para que os resultados obtidos sejam uteis e forneçam informações relevantes para a solução dos problemas em análise.

A partir de um modelo de regressão linear é possível estabelecer uma função matemática entre diversas variáveis aleatórias relevantes ao estudo, de modo que o modelo matemático seja capaz de descrever adequadamente a relação. Adicionalemente, a regressão linear permite que seja feita previsões de resultados de uma variável resposta a partir de um conjunto de variáveis preditoras (explicativas).

Devido a importância da educação no sistema brasileiro de ensino, esse estudo teve como objetivo aplicar técnicas de regressão linear para encontrar um modelo matemático que explique o comportamento da quantidade de matrículas no ensino regular, no estado do Paraná, Brasil.

3 Material e Métodos

A base de dados utilizado no estudo foi obtido no site oficial do Instituto Paranaense de Desenvolvimento Econômico e Social (IPARDES). O banco de dados original correspondeu a 17 variáveis referente aos 399 municipios do estado do Paraná. A variável resposta considerada foi o número de matrículas no ensido regular (V7). As demais covariáveis foram utilizadas como possíveis variáveis explicativas. A descrição de cada uma das variáveis é dada na sequência.

- **Localidade**: Nome do municipio;
- **V2**: Densidade Demográfica (hab/km^2);
- **V3**: Despesas Correntes Municipais - Pessoal e Encargos Sociais
- **V4**: Despesas Municipais - Total

- **V5**: Emprego Formal
- **V6**: Frota de Veículos - Automóvel
- **V7**: Matrículas no Ensino Regular
- **V8**: População Estimada (IBGE)
- **V9**: Receitas Correntes Municipais
- **V10**: Receitas Municipais
- **V11**: Transferências Correntes da União
- **V12**: Transferências Correntes do Estado
- **V13**: Valor Adicionado Fiscal per Capita
- **V14**: VBP - Agricultura (Valor Bruto nominal da Produção)
- **V15**: VBP - Florestais (Valor Bruto nominal da Produção)
- **V16**: VBP - Pecuária (Valor Bruto nominal da Produção)

Inicialmente, devido a elevada variabilidade dos dados e os diferentes tipos de escala, procedeu-se com a divisão de todas as variáveis pela população estimada (V8), exceto a variável que expressa o nome do município (V1 - localidade), pois é um fator com 399 níveis, e a variável densidade demográfica V2 (densidade), que já está relativizada.

Em análise preliminar do modelo de regressão linear, devido a não normalidade e a heterocedasticidade dos resíduos, optou-se por utilizar duas abordagens para construção do modelo: 1) a primeira consistiu no ajuste de um modelo robusto (Regressão Robusta) com ponderação. A ponderação foi definida pelo inverso da variável resposta, devido a forte tendência nos resíduos de aumentar a sua amplitude de variação em função do aumento da variável resposta; 2) a segunda consistiu na transformação da variável resposta como y^{-3}, definido pelo fator de potência de Box-Cox. Em seguida, a seleção das variáveis explicativas foi realizada pelo método *stepwise*, por meio do teste F de Snedecor, considerando um nível de significância de 5%.

O desempenho dos modelos foi avaliado pelo coeficiente de determinação ajustado (R^2_{aj}). A homogeneidade dos resíduos studentizados foi avaliada por meio de gráficos de dispersão de resíduos sob a variável resposta estimada. A normalidade dos resíduos foi avaliada pelo teste de Shapiro-Wilk, considerando 5% de significância, além de gráficos qqplot, do pacote car(). Ainda, para avaliar possíveis pontos (observações) influentes, o gráfico da distância de Cook foi construído.

4 Resultados e Discussão

O resultado da estatísica descritiva para as variáveis V2, V4, V7, V11, V13, V14, V15 e V16 está apresentado na Tabela 1. É possível notar que a variável V2, densidade demográfica, possui maior coeficiente de variação (%), enquanto a variável V4, total de despesas municipais, possui o menor valor.

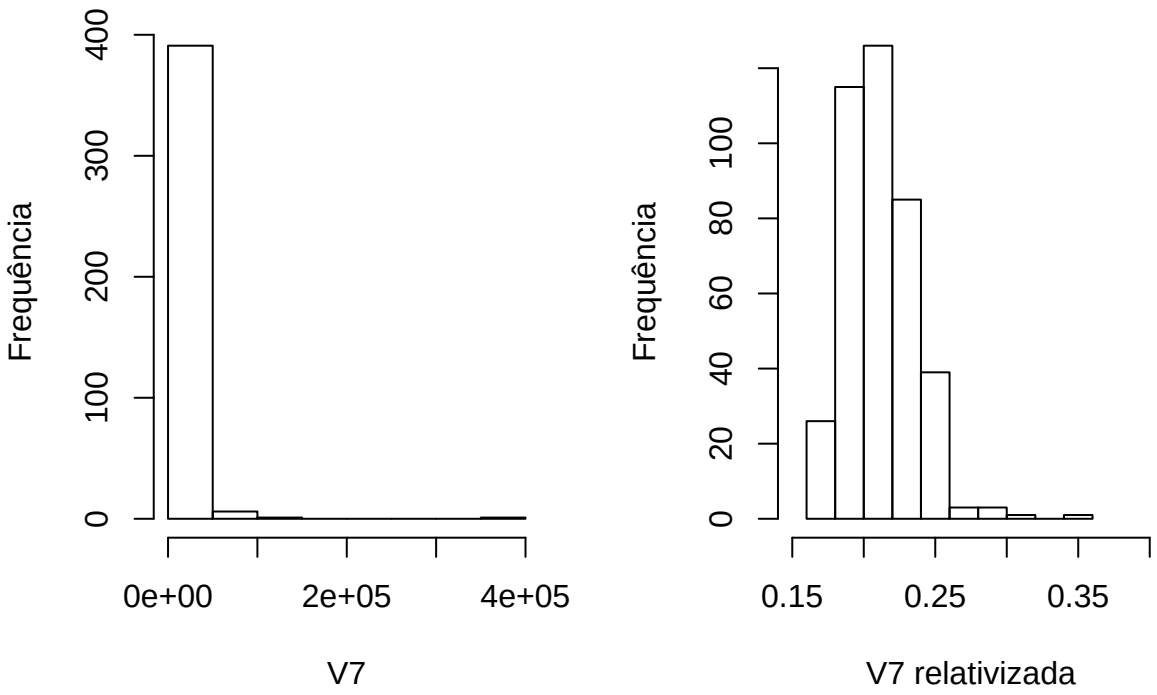

Figura 1: Distribuição de frequência da variável resposta número de matrículas no ensino regular (V7)

Tabela 1: Estatística descritiva das variáveis V2, V4, V7, V11, V13, V14, V15 e V16

	V2	V4	V7	V11	V13	V14	V15	V16
Mínimo	0.0005	0.0022	0.0003	0.0002	0.0001	0.0018	0.0001	0.0012
Média	5.1301	3.8579	4.0762	7.0509	8.6367	6.9357	3.4724	4.0317
Máximo	47.4300	18.1285	23.0900	63.1500	89.2600	28.0300	22.1500	16.2500
Desv. Padrão	12.0686	4.8228	6.5204	16.3299	22.9051	9.4776	7.1006	5.4483
Coef. Var.	235.2525	125.0098	159.9629	231.6012	265.2062	136.6494	204.4900	135.1343

A Figura 1 mostra o comportamento da variável número de matrículas no ensino regular (V7) na escala original e com a variável relativizada pela população estimada (V8). Nota-se que na escala original há um padrão claro de distribuição Exponencial Negativa bastante forte, mas se relativizar a variável resposta a distribuição aproxima-se de alguma outra distribuição assimétrica, para valores entre 0 e 1, como a distribuição simplex e beta. Portanto, percebe-se que a regressão linear com distribuição normal da variável resposta não é o modelo mais adequado.

Tabela 2: Parâmetros estimados dos modelos de regressão testados

X	Coeficientes	Erro Padrão	Significância	X	Coeficientes	Erro Padrão	Significância
b0	0.1953	4.11e-03	**	b0	146.6798	7.8741	**
b1	1.88e-05	8.84e-06	**	b1	-0.0199	0.00130	**
b2	8.64e-03	1.41e-03	**	b2	-15.9285	2.7761	**
b3	-1.1251	0.4619	**	b3	1845.0476	731.5038	**
b4	-7.54e-04	1.61e-04	**	b4	0.9031	0.5084	**
b5	1.77e-03	6.29e-04	**	b5	1.1101	0.2645	**
b6	-4.77e-04	1.33e-04	**	b6	-2.8394	0.9927	**
-	-	-	-	b7	0.7134	0.2148	**

O modelo final para cada abordagem apresentou as seguintes variáveis:

- Regressão Robusta:

$$y_{=}\beta_0 + \beta_1 v_2 + \beta_2 v_4 + \beta_3 v_{11} + \beta_4 v_{14} + \beta_5 v_{15} + \beta_6 v_{16}$$

Regressão com transformação:

$$y_{=}\beta_0 + \beta_1 v_2 + \beta_2 v_4 + \beta_3 v_{11} + \beta_4 v_{13} + \beta_5 v_{14} + \beta_6 v_{15} + \beta_7 v_{15}$$

A Tabela 2 apresenta os parâmetros estimados dos modelos testados. Nota-se que os coeficientes foram todos significativos, considerando nível de significância de 5%. Além disso, o coeficiente de determinação ajustado apresentou valor baixo e muito semelhante para ambos os modelos testados, os quais foram aproximadamente 0.13. Esse resultado pode ser atribuído ao fato de que os dados apresentaram elevada variabilidade, e mais covariáveis deveriam ser utilizadas para explicar a variação total da variável resposta.

Na Figura 2 é possível observar os gráficos de resíduos para cada modelo testado e o gráfico qqplot. Nota-se que há um padrão diferente nos resíduos para cada modelo, mas em ambos há tendências de subestimativas nos menores valores da variável resposta. O Gráfico qqplot ainda indicou que há problemas no extremos da distribuição, devido a existência de observações fora do envelope de confiança. Além disso, o teste de Shapiro-Wilk indicou que os resíduos não possuem distribuição normal, considerando 5% de significância, para ambos os modelos.

Na Figura 3 está apresentado a distância de Cook de cada observação para cada modelo testado. Observa-se que o modelo com transformação na resposta indicou maior quantidade de observações classificadas como influentes, enquanto a Regressão Robusta indicou poucas observações e com menores distâncias.

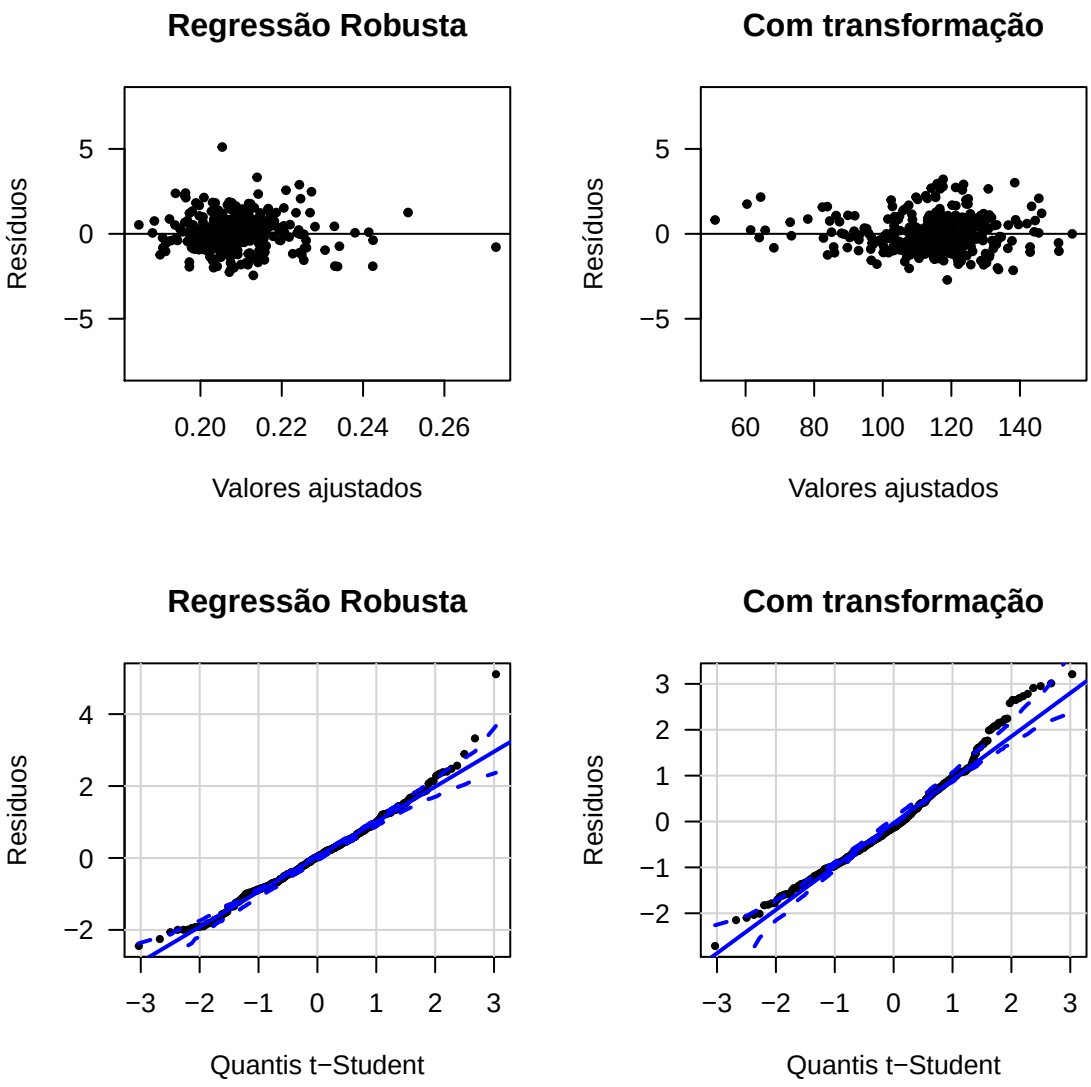

Figura 2: Gráficos de dispersão de resíduos para a variável resposta

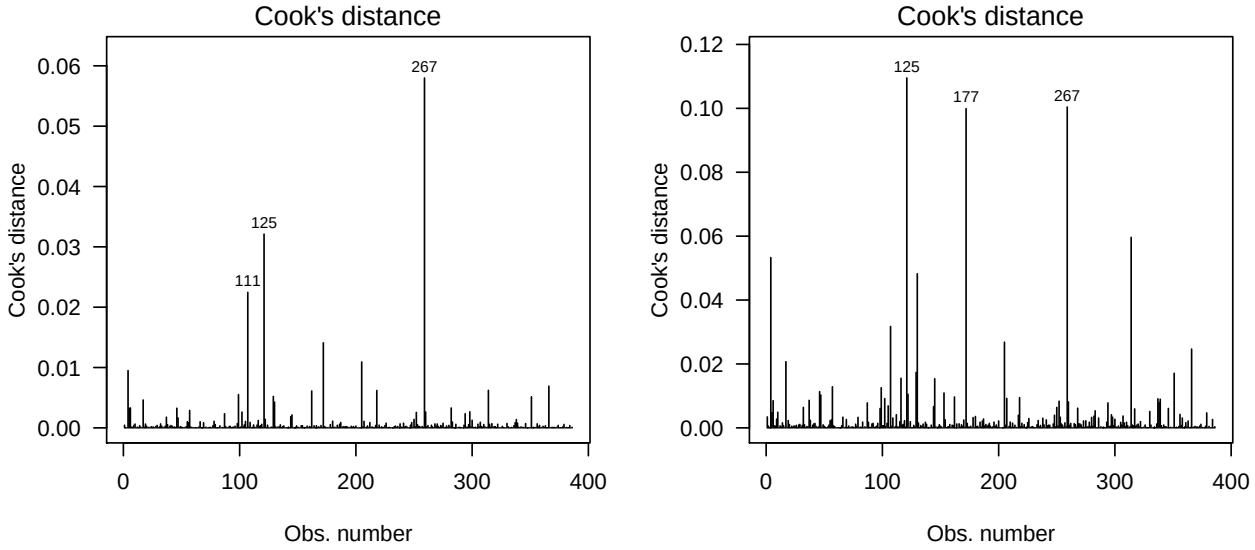

Figura 3: Gráficos da distância de cook

5 Conclusão

As técnicas de regressão linear para os dados utilizados nesse trabalho foram insuficientes devido ao não atendimento dos pressupostos básicos. A falta de normalidade dos resíduos aliada a heterocedasticidade dificultaram o ajuste dos modelos de maneira satisfatória. Devido ao comportamento da distribuição de probabilidade do número de matrículas no ensino regular, melhores resultados devem ser obtidos se modelar a variável resposta com outra distribuição de probabilidade.