1. Introdução

O conjunto de dados em questão foi extraído do censo do IBGE de 2000 e apresenta, para cada unidade da federação, o número médio de anos de estudo e a renda média mensal (em reais) do chefe ou chefes do domicílio. O objetivo é estudar o relacionamento da renda média mensal em função do número médio de anos de estudo. Os dados podem ser encontrados no site (https://www.ime.usp.br/~giapaula/textoregressao.htm) sob o nome de "censo.dat". Utilizamos a metodologia dos modelos normais lineares homocedásticos, veja Draper and Smith (1998), metodologias de verificação da qualidade do ajuste e comparação de modelos apropriados, veja Paula (2013) com o suporte computacional do R, veja Faraway (2014).

2. Análise descritiva

Neste caso, por termos uma única covariável, ela ser quantitativa e termos somente um ou poucos valores da resposta para cada valor dela, podemos, apenas, fazer um gráfico de dispersão. A Figura 1 apresenta-o. Podemos perceber que a relação entre elas pode ser cabalmente representada por uma reta ou uma curva do segundo grau. Podemos observar também que a relação parece ser positiva, ou seja,

quanto maior o número médio de anos de estudo, maior a renda média do chefe ou dos chefes de família, embora esse aumento pareça maior para números médios de anos de estudo mais elevados.

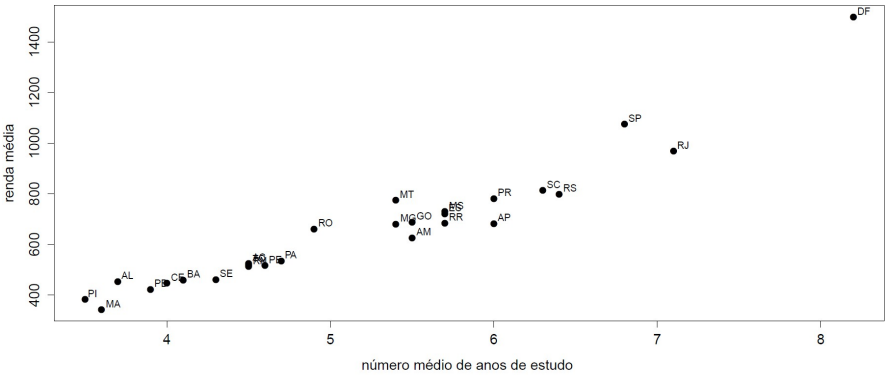

Figura 1: Gráfico de dispersão entre número médio de anos de estudo e renda média

3. Análise Inferencial

De acordo com a natureza do experimento, os objetivos em questão e devido aos resultados da análise descritiva, vamos considerar os seguintes modelos:

Modelo 1

$$Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \xi_i ; i = 1,2,\ldots,27 ; \xi_i \overset{i.\tilde{i}.d.}{\sim} N(0, \sigma^2) ; \bar{x} \approx 5{,}20$$

- Y_i: renda média mensal (em reais) do chefe ou chefes do domicílio da i-ésima unidade da federação.
- x_i : número médio de anos de estudo do chefe ou chefes do domicílio da i-ésima unidade da federação.
- β_0 : renda média mensal esperada, do chefe ou chefes do domicílio, quando 0 número médio de anos de estudo for igual à 5,20 (aproximadamente).
- β_1 : incremento na renda média mensal esperada, do chefe ou chefes do domicílio, quando o número médio de anos de estudo aumenta em uma unidade

Modelo 2

$$Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2 + \xi_i ; i = 1,2,\ldots,27 ; \xi_i \overset{i.\tilde{i}.d.}{\sim} N(0,\sigma^2) ; \bar{x} \approx 5{,}20$$

- β_0 : renda média mensal esperada, do chefe ou chefes do domicílio, quando número médio de anos de estudo for igual à 5,20 (aproximadamente).

- $\frac{-2\beta_1}{\beta_2}$: valor do número de anos de estudo para o qual a renda média esperada é máxima ou mínima.

- As outras quantidades são como definidas para o modelo 1.

Os dois modelos foram ajustados via metodologia de mínimos quadrados ordinários, veja Azevedo (2019) e análises residuais foram realizadas, conforme Paula (2013), veja as Figuras de 2 a 5. Podemos ver que o modelo 2 apresentou um melhor ajuste, embora este, também, não seja satisfatório. Para o modelo 1, vemos uma observação que se destaca muito em relação as demais, e que ela induz uma assimetria positiva nos resíduos. Além disso, o gráfico de envelopes acusa um mal ajuste, devido ao comportamento sistemático, além do que percebemos heterocedasticidade (pelo gráfico de resíduos x valore ajustado), uma vez que a variabilidade dos resíduos tende a aumentar com o aumento do valor predito. Para o modelo 2, os aspectos mencionados acima continuam presentes (embora o ajuste tenha melhorado) pelos mesmos motivos anteriormente apresentados, com exceção do ponto que se destacava, o qual não mais se destaca. Concluímos, pois, que nenhum dos dois modelos se ajustou bem. Contudo, devido ao fato de não podermos escolher modelos para além da classe dos modelos lineares normais homocedásticos, continuaremos a compará-los.

As estatísticas de comparação dos dois modelos foram AIC = 315,26, BIC = 319,15 (modelo 1), AIC = 298,66, BIC = 303,85 (modelo 2), resultado este que favorece, novamente, ao modelo 2. Na Tabela 1 apresentamos os principais

resultados relativos à estimação pontual e intervalar dos dois modelos. Podemos ver que todos os parâmetros são diferentes de zero, para qualquer nível de significância usual (0,01 à 0,10). Novamente, devido à significância do parâmetro β_2 (modelo 2), temos evidências em favor dele. Finalmente, na Figura 6, apresentamos os valores individuais observados e preditos, veja Azevedo (2019). A melhor predição é obtida através do modelo 2 o qual, inclusive, consegue acomodar o ponto que se destaca, correspondente ao Distrito Federal. Portanto, utilizaremos tal modelo para realizar a análise dos dados.

Pela Tabela 1 vemos que a renda média, para chefe ou chefes de domicílio, com anos de estudo de aproximadamente 5,20 é da ordem de [588,73;645,33]. Como estamos com um modelo quadrático, sabemos que o incremento na renda, para o aumento de uma unidade nos anos de estudo, depende deste. Com efeito, este incremento é igual à $(\beta_1 + 2(x- \bar{x}) \beta_2 + \beta_2)$. O comportamento de tal incremento, em função dos anos de estudo, pode ser visto na Figura 7. Podemos ver que o incremento na renda (para o aumento em uma unidade nos anos de estudo) só se torna positivo com nove anos de estudo. Isto pode ser devido ao fato de que não se observou domicílios cujos chefes tivessem menos do que 3,5 nem mais do que 8,2 anos médios de estudo, o mal ajuste do modelo, o aumento não linear observado na

Figura 1, bem como a outros fatores que devem ser investigados. Note-se, também, que estamos trabalhando com um incremento de um ano de estudo.

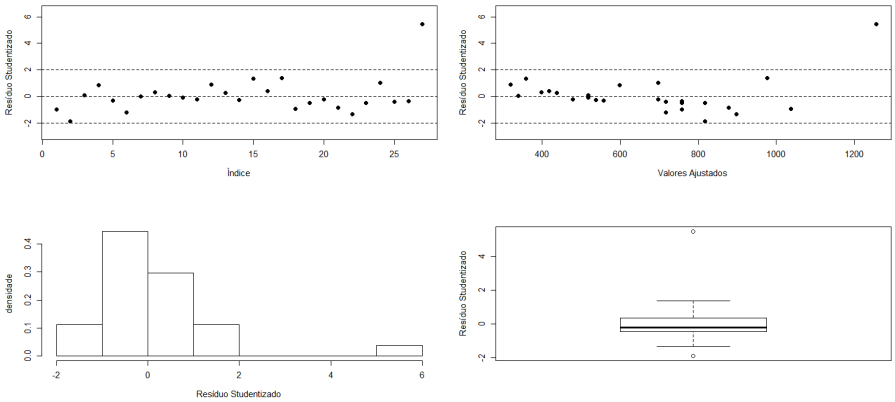

Figura 2: Análise residual para o modelo 1

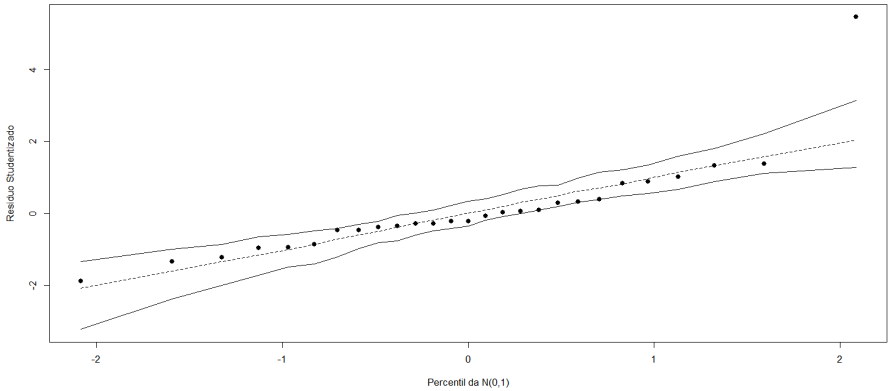

Figura 3: Gráfico de envelope para o resíduos studentizado para o modelo 1

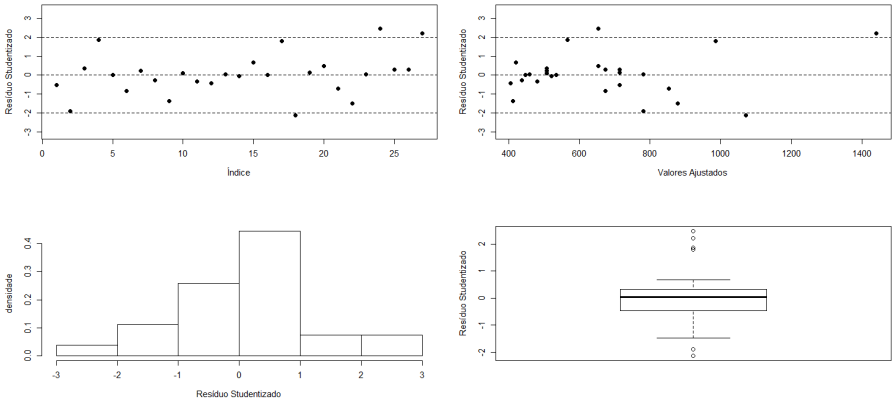

Figura 4: Análise residual para o modelo 2

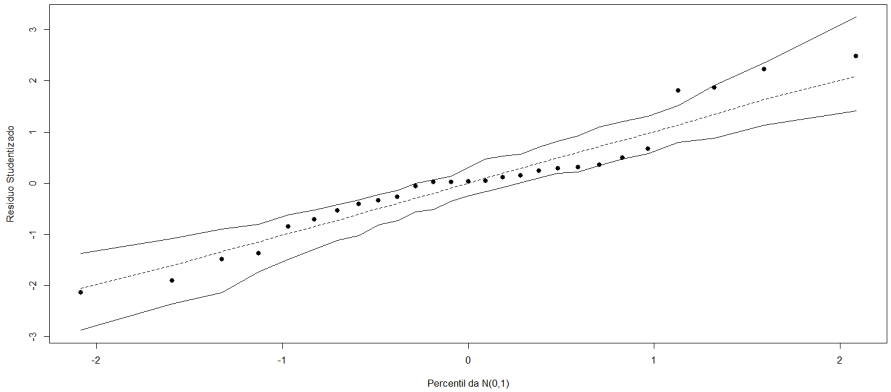

Figura 5: Gráfico de envelope para o resíduos studentizado para o modelo 2

Tabela 1: Estimativas dos parâmetros, intervalos de confiança e teste de nulidade: modelo

Modelo	Parâmetro	Estimativa	EP	IC (95%)	Estat. t	p-valor
1	β_0	658,66	14,87	[627,94;689,16]	44,32	< 0,0001
	β_1	199,83	13,03	[172,99;226,57]	15,34	< 0,0001
2	β_0	617,03	13,72	[588,73;645,33]	45,00	< 0,0001
	β_1	179,54	10,30	[158,26;200,80]	17,43	< 0,0001
	β_2	32,92	6,54	[18,41;45,21]	4,88	0,0001

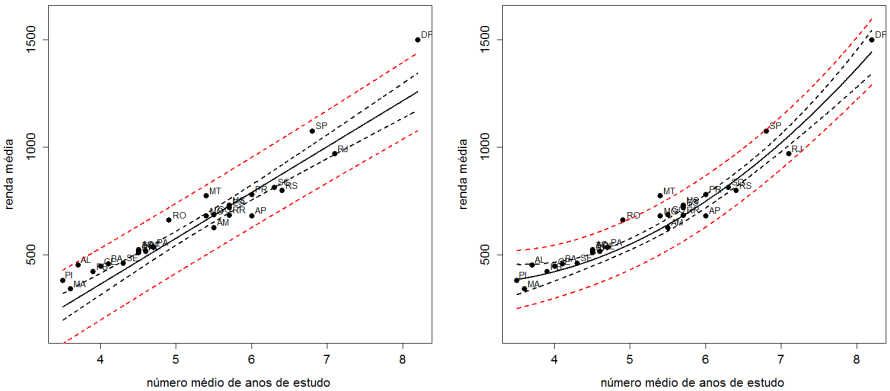

Figura 6: Valores observados e preditos da renda para o modelo 1 (esquerda) e modelo 2 (direita)

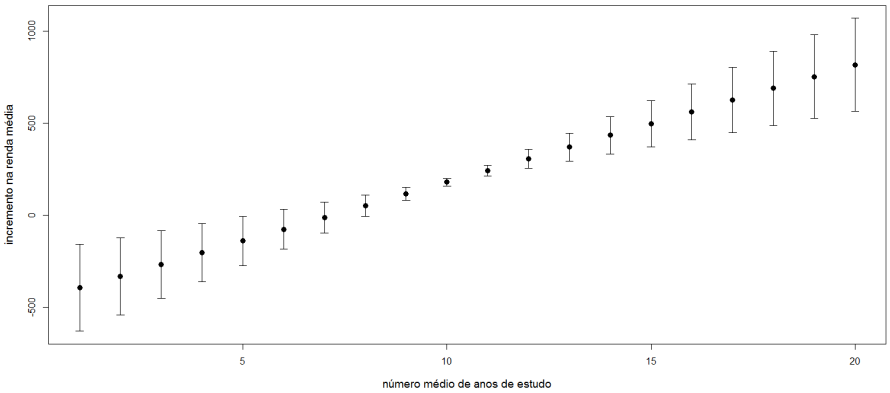

Figura 7: Estimativas pontuais e intervalares do incremento na renda média em função dos anos de estudo.

4. Conclusões

Nenhum dos dois modelos se ajustou bem ao conjunto de dados. Contudo, devido às restrições, optou-se por aquele com o ajuste menos ruim. O modelo revelou que há um aumento na renda em função dos anos de estudo, embora esse modelo aumento dependa dos anos de estudo. Para valores menores da variável explicativa, o aumento em uma unidade causa uma diminuição na renda, embora para aumentos de maior magnitude e/ou para anos de estudos mais elevados, observa-se um aumento na renda. Tal fato pode ser devido ao espectro de valores observados para os anos de estudo, mal ajuste do modelo, outros fatores que não foram considerados no experimento e/ou aspectos que devem ser considerados por especialistas da área de estudo.

5. Referências Bibliográficas

- Azevedo, C. L. N (2019). Notas de aula sobre Análise de regressão, http://www.ime.unicamp.br/~cnaber/Material_ME613_1S_2019.htm

- Draper, N. R. and Smith, H. (1998). Applied regression analysis, third edition. New York, NY: John Wiley & Sons.

- Faraway, J. J. (2014). Linear Models with R, Second Edition, Chapman & Hall/CRC Texts in Statistical Science

- Paula, G. A. (2013). Modelos de regressão com apoio computacional, ver são pré-eliminar, https://www.ime.usp.br/~giapaula/texto_2013.pdf