Universidade Estadual de Campinas

Instituto de Matemática, Estatística e Computação Cientifica
Departamento de Estatística

Relatório - Parte II
Trabalho Final de ME613

Eliane Ramos de Siqueira RA:155233
Guilherme Pazian RA:160323
Henrique Capatto RA:146406
Murilo Salgado Razoli RA:150987

Professor: Caio Lucidius Naberezny Azevedo

Campinas-SP, 08 de Dezembro de 2016

1. Introdução

Este presente trabalho é parte do escopo da disciplina ME613 - Regressão Linear e visa a aplicação dos conhecimentos adquiridos em sala de aula. Grande parte do trabalho foi baseado e adaptado a partir dos materiais disponibilizados na página do curso: *http://www.ime.unicamp.br/~cnaber/Material_ME613_2S_2016.htm*.
O conjunto de dados a ser analisado está disponível na página do curso sob o nome "reg3.dat". Neste são descritas as seguintes variáveis referentes a 50 estados norte-americanos: (i) **estado** (nome do estado), (ii) **pop** (população estimada em julho de 1975), (iii) **percap** (renda percapita em 1974 em USD), (iv) **analf** (proporção de analfabetos em 1970), (v) **expvida** (expectativa de vida em anos 1969-70), (vi) **crime** (taxa de criminalidade por 100000 habitantes 1976), (vii) **estud** (porcentagem de estudantes que concluem o segundo grau 1970), (viii) **ndias** (número de dias do ano com temperatura abaixo de zero grau Celsus na cidade mais importante do estado) e (ix) **area** (área do estado em milhas quadradas). O objetivo do estudo é tentar explicar a variável **expvida** usando um modelo de regressão normal linear dadas as variáveis explicativas **percap**, **analf**, **crime**, **estud**, **ndias** e **dens**, em que **dens=**pop/area (densidade da população estimada em julho de 1975 por área do estado em milhas quadradas).
Todos os modelos neste trabalho foram ajustados via metodologia de mínimos quadrados ordinários, veja Azevedo (2016), e todas suas respectivas análises residuais foram realizadas, conforme Paula (2013). A menos que seja citado o contrário, todas as variáveis que constam na base de dados serão referidas por sua descrição completa ou pelo nome que consta no banco de dados (estes foram supracitados em negrito). Denotaremos também:

- Y_i - i-ésima observação da variável expvida
- x_{1i} - i-esima observação da variável percap
- x_{2i} - i-esima observação da variável analf
- x_{3i} - i-esima observação da variável crime
- x_{4i} - i-esima observação da variável estud
- x_{5i} - i-esima observação da variável ndias
- x_{6i} - i-esima observação da variável dens

OBS: Todas as análises presentes neste trabalho foram obtidas com auxílio dos softwares "R" e "RStudio". Ambos são gratuitos e estão disponíveis nos sites *https://cran.r-project.org/index.html* e *https://www.rstudio.com/products/rstudio/download/* respectivamente.

Para melhor entendimento dos dados, podemos observar algumas estátisticas descritivas dos dados representadas na tabela 1.

2. Análise descritiva

Tabela 1: Estatísticas Descritivas

Variável	Min.	1º quartil	Mediana	Média	3º quartil	Max.
percap	3098,0000	3993,0000	4519.0000	4436.0000	4814,0000	6315.0000
analf	0,5000	0,6250	0.9500	1.1700	1,5750	2.8000
crime	1,4000	4,3500	6.8500	7.3780	10,6800	15.1000
estud	37,8000	48,0500	53.2500	53.1100	59,1500	67.3000
ndias	0,0000	66,2500	114.5000	104.5000	139,8000	188.0000
dens	0,0006	0,0277	0.0690	0.1880	0,1443	2.6840
expvida	67,9600	70,1200	70.6800	70.8800	71,8900	73.6000

Dado a natureza quantitativa dos dados, é de grande interesse que identifiquemos as relações entre as covariáveis explicativas e a variável resposta. A fim de tornar essas relações visuais, foram construidos gráficos de dispersão entre a variável resposta e todas as covariáveis de interesse. Estes estão representados na figura 1.

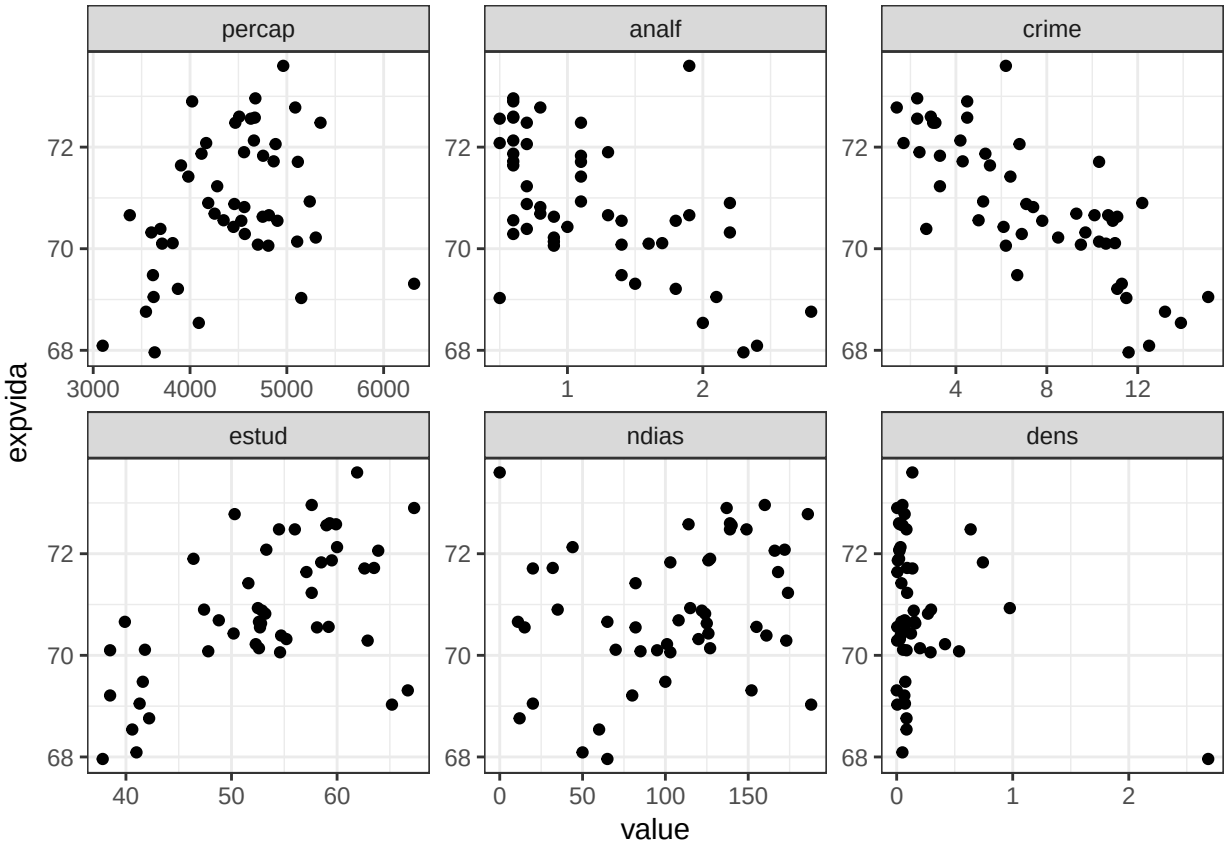

Figura 1: Gráficos de dispersão

A partir da Figura 1, podemos identificar visualmente a relação supracitada, a qual identificamos que todas as covariáveis parecem ter uma relação linear significativa. Existem relações lineares negativas entre a variável resposta “expvida” e as covariáveis “analf” e “crime”, assim como relações lineares negativas entre a variável resposta “expvida” e as covariáveis “percap”,“estud”,“ndias” e “dens” (a menos a um ponto).

3. Análise Inferencial

Dada a constatação visual de relações lineares entre a variável resposta “expvida” e todas as covariáveis explicativas, é razoável iniciar um modelo que contemple todas estas covariáveis. Visando poder comparar diretamente os coeficiêntes do modelo, optamos por introduzir as covariáveis com médias e variâncias iguais, tornardo-as adimensionais. Portanto, vamos iniciar a análise com o seguinte modelo:

Modelo 1:

$$Y_i = \beta_0 + \sum_{j=1}^{6} \beta_j \left(\frac{x_{ji} - \bar{x_j}}{s_j} \right) + \varepsilon_i \begin{cases} i = 1, \ldots, 50 \\ j = 1, \ldots, 6 \end{cases}$$

Onde: $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $x_j = \frac{1}{50} \sum_{i=1}^{50} x_j i$ e $s_j = \sqrt{\frac{1}{50} \sum_{i=1}^{50} (x_{ji} - \bar{x_j})^2}$

- Y_i : Expectativa de vida em anos (1969-70).
- β_0 : Expectativa de vida esperada em anos (1969-70) para valores de covariáveis iguais às suas respectivas médias.
- $\frac{\beta_1}{s_1}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "renda percapita (em 1974 em USD)" em uma unidade, mantendo-se as demais covariáveis fixas.
- $\frac{\beta_2}{s_2}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "proporção de analfabetos (em 1970)" em uma unidade, mantendo-se as demais covariáveis fixas.
- $\frac{\beta_3}{s_3}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "taxa de criminalidade (por 100000 habitantes 1976)" em uma unidade, mantendo-se as demais covariáveis fixas.
- $\frac{\beta_4}{s_4}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "porcentagem de estudantes que concluem o segundo grau (1970)" em uma unidade, mantendo-se as demais covariáveis fixas.
- $\frac{\beta_5}{s_5}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor de "número de dias do ano com temperatura abaixo de zero grau Celsus na cidade mais importante do estado" em uma unidade, mantendo-se as demais covariáveis fixas.
- $\frac{\beta_6}{s_6}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "densidade da população estimada em julho de 1975 por área do estado em milhas quadradas" em uma unidade, mantendo-se as demais covariáveis fixas.

Pode-se observar a partir da figura 1 que o modelo não teve um ajuste adequado. No gráfico A não observamos tendências,o que pode indicar uma independência dos dados. Já no gráfico B, observa-se uma leve heterocedasticidade, dado que os valores de resíduo studentizado parecem ser maiores em módulo conforme os valores ajustados aumentam. Porém, um fator de confundimento que pode ser levado em consideração é a assimetria da distribuição dos dados, já que é possível observar uma assimetria no histograma apresentado (gráfico C da figura 2), assim como uma tendencia no gráfico de envelopes (figura 3) o que nos dá indícios de falta de normalidade nos resíduos. Neste caso, deveria-se procurar modelos alternativos que levem em consideração uma distribuição assimétrica, ou um modelo que leve em consideração heterocedásticidade. Dado o escopo do curso, procede-se com as análises posteriores.

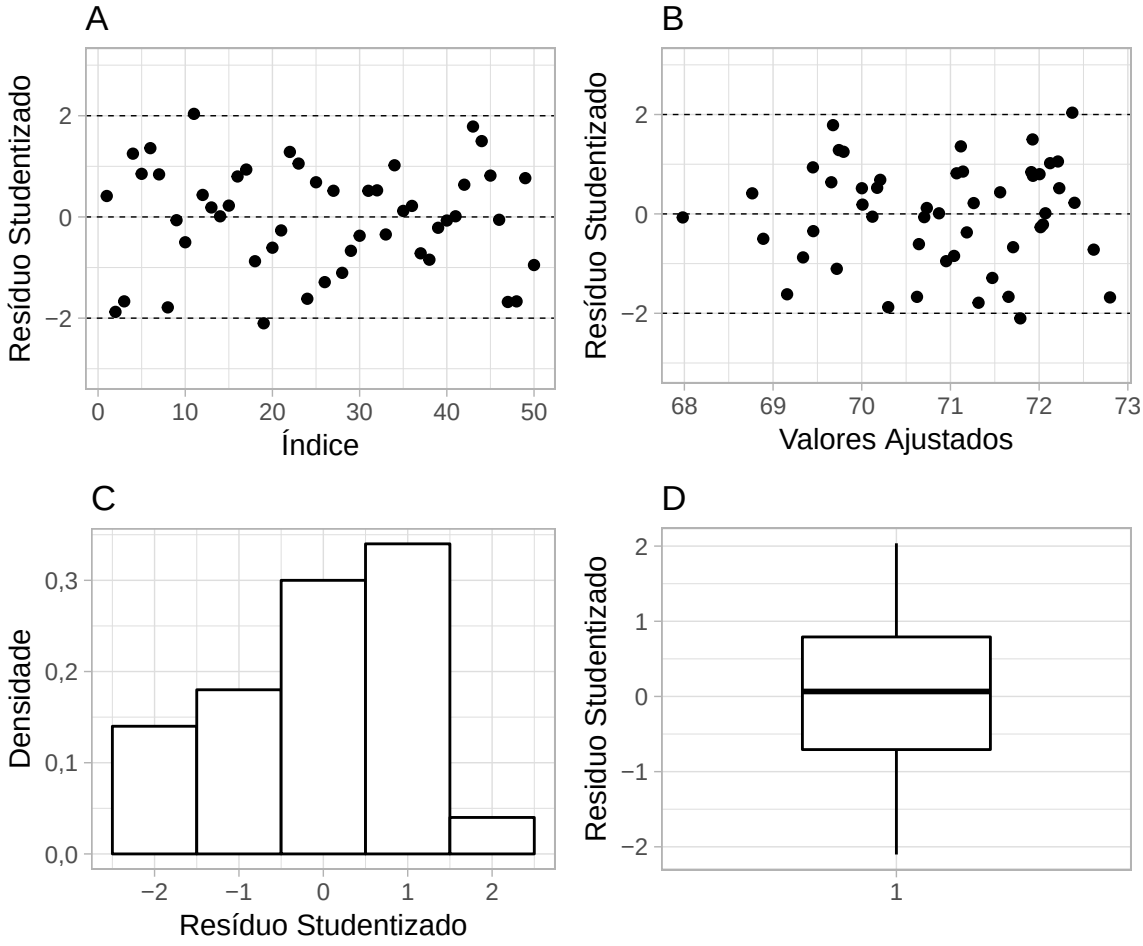

Figura 2: Análise de Resíduo para o modelo 1

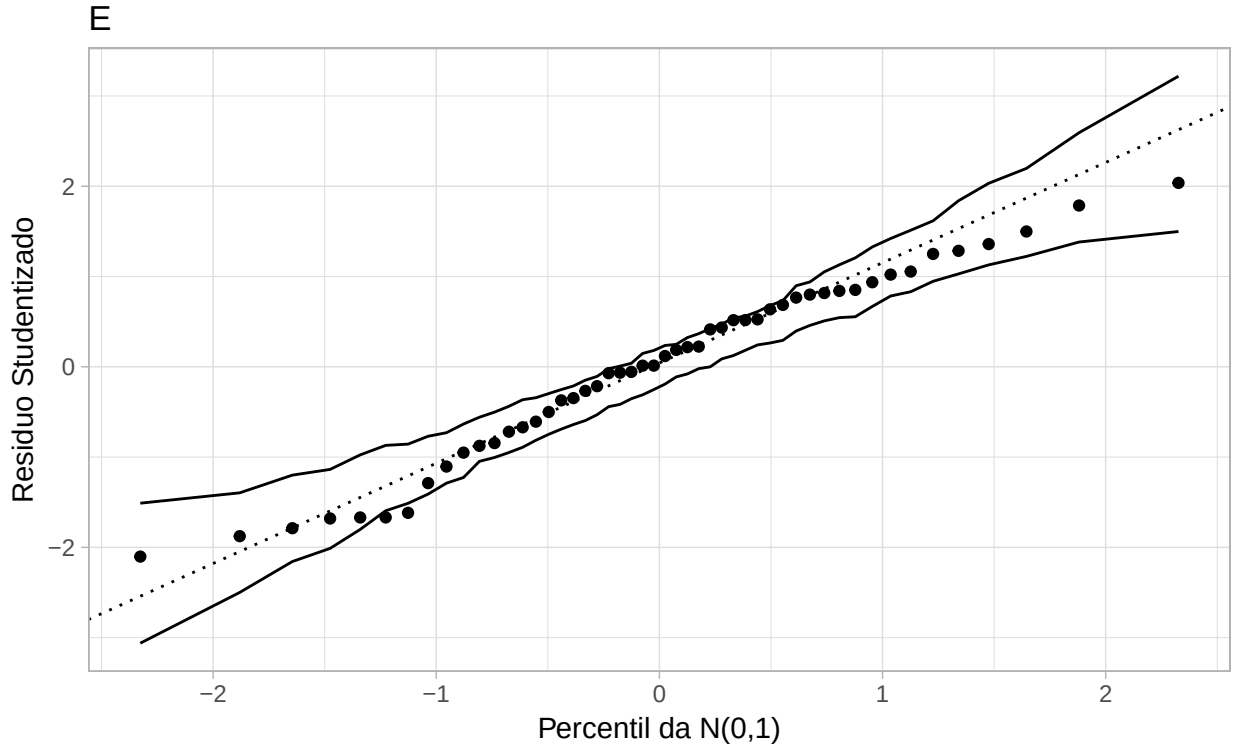

Figura 3: Gráfico de envelope para os resíduos studentizado para o modelo 1

Note, pela tabela 2, que somente o intercepto (β_0) e os parêmetros relacionados às covariáveis "crime", "ndias" e "dens" (β_3,β_5 e β_6) são significativos à um nível de significância de 10%. A partir desse resultado, temos a indicação de que um modelo reduzido pode ser mais apropriado. Contudo, desejamos também, obter o modelo que melhor se ajusta aos dados.

Tabela 2: Estimativas dos parâmetros, intervalo de confiança e teste de nulidade para o modelo 1

Parâmetro	Estimativa	EP	IC (95%)	Valor T	Valor p
β_0	70,8786	0,1039	[70,6691 ; 71,0881]	682,2147	<0,0001
β_1	0,1352	0,1404	[-0,1479 ; 0,4183]	0,9631	0,3409
β_2	-0,0322	0,2010	[-0,4375 ; 0,3731]	-0,1601	0,8735
β_3	-1,0618	0,1527	[-1,3697 ; -0,7538]	-6,9535	<0,0001
β_4	0,2455	0,1673	[-0,0918 ; 0,5829]	1,4677	0,1495
β_5	-0,3907	0,1444	[-0,6818 ; -0,0995]	-2,7058	0,0097
β_6	-0,2108	0,1139	[-0,4404 ; 0,0188]	-1,8518	0,0709

3.1 Seleção dos modelos

A técnica escolhida para nos auxiliar a selecionar o modelo normal linear homocedastico que melhor se ajusta aos dados foi a técnica stepwise, veja Azevedo (2016).

A aplicação da metodologia stepwise, começando com o modelo só com o intercepto ou começando com o modelo completo, indicou, em ambos os casos, que o modelo o qual melhor se ajusta é o modelo que leva em consideração somente as

covariáveis “crime”,“estud”,“ndias” e “dens”.

Temos agora o seguinte modelo:

Modelo 2:

$$Y_i = \beta_0 + \sum_{j=3}^{6} \beta_j \left(\frac{x_{ji} - \bar{x_j}}{s_j} \right) + \varepsilon_i \begin{cases} i = 1, \ldots, 50 \\ j = 3, \ldots, 6 \end{cases}$$

Onde $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $x_j = \frac{1}{50} \sum_{i=1}^{50} x_j i$ e $s_j = \sqrt{\frac{1}{50} \sum_{i=1}^{50} (x_{ji} - \bar{x_j})^2}$

- Y_i : Expectativa de vida em anos (1969-70).
- β_0 : Expectativa de vida esperada em anos (1969-70) para valores de covariáveis iguais às suas respectivas médias.
- $\frac{\beta_3}{s_3}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "taxa de criminalidade (por 100000 habitantes 1976)" em uma unidade, mantendo-se as demais covariáveis fixas.
- $\frac{\beta_4}{s_4}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "porcentagem de estudantes que concluem o segundo grau (1970)" em uma unidade, mantendo-se as demais covariáveis fixas.
- $\frac{\beta_5}{s_5}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor de "número de dias do ano com temperatura abaixo de zero grau Celsus na cidade mais importante do estado" em uma unidade, mantendo-se as demais covariáveis fixas.
- $\frac{\beta_6}{s_6}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "densidade da população estimada em julho de 1975 por área do estado em milhas quadradas" em uma unidade, mantendo-se as demais covariáveis fixas.

Em anuência ao ajuste do modelo anterior, este modelo também não apresentou um ajuste adequado, uma vez que se observa na figura 4 indícios de mal ajuste semelhantes aos observados na análise residual para o modelo anterior. A tendência no gráfico B indica heterocedasticidade dos resíduos e as tendências observadas nos gráfico C (Assimetria da distribuição) e na figura 5 (tendência nos envelopes) indicam a falta de normalidade nos resíduos.
Neste caso, deveria-se procurar modelos alternativos que levem em consideração uma distribuição assimetrica, ou um modelo que leve em consideração heterocedásticidade. Novamente, dado o escopo do curso, procede-se com as análises posteriores.

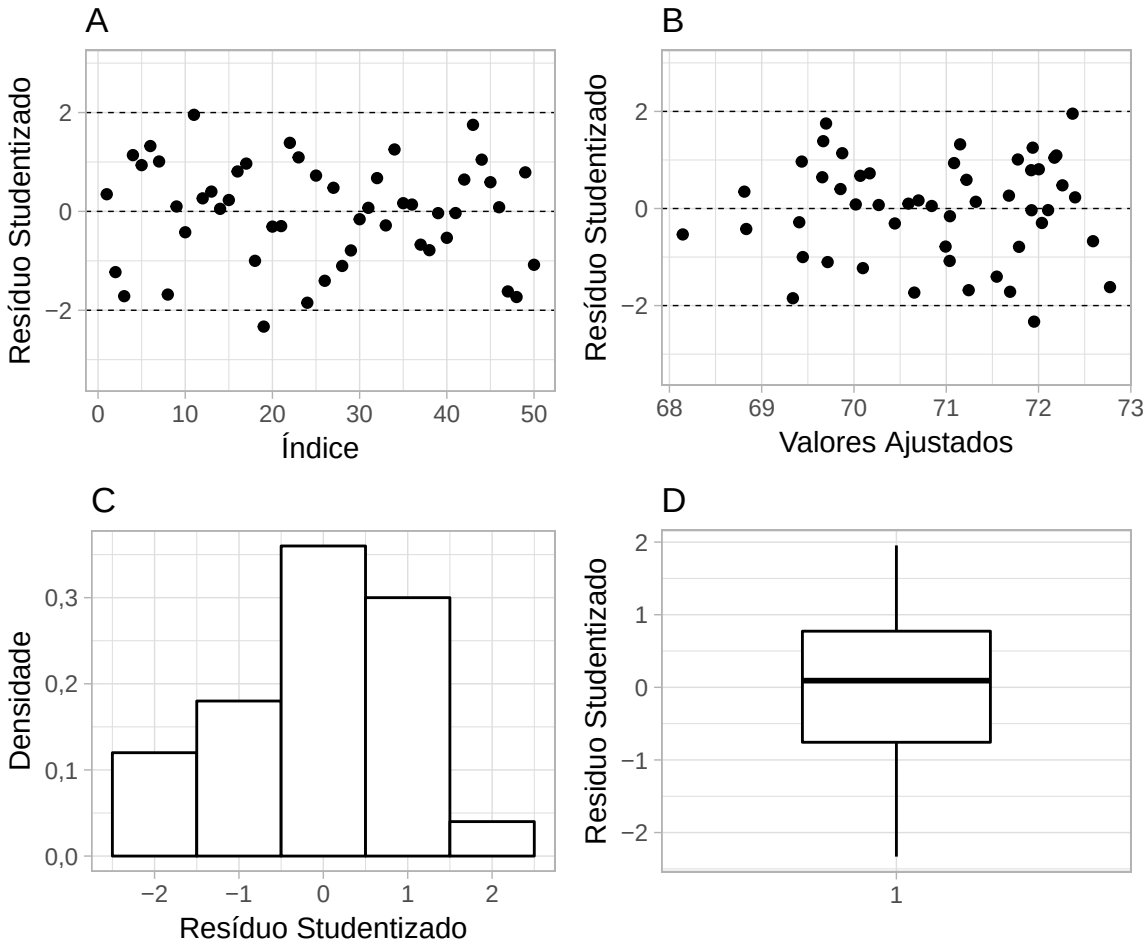

Figura 4: Análise residual para o modelo 2

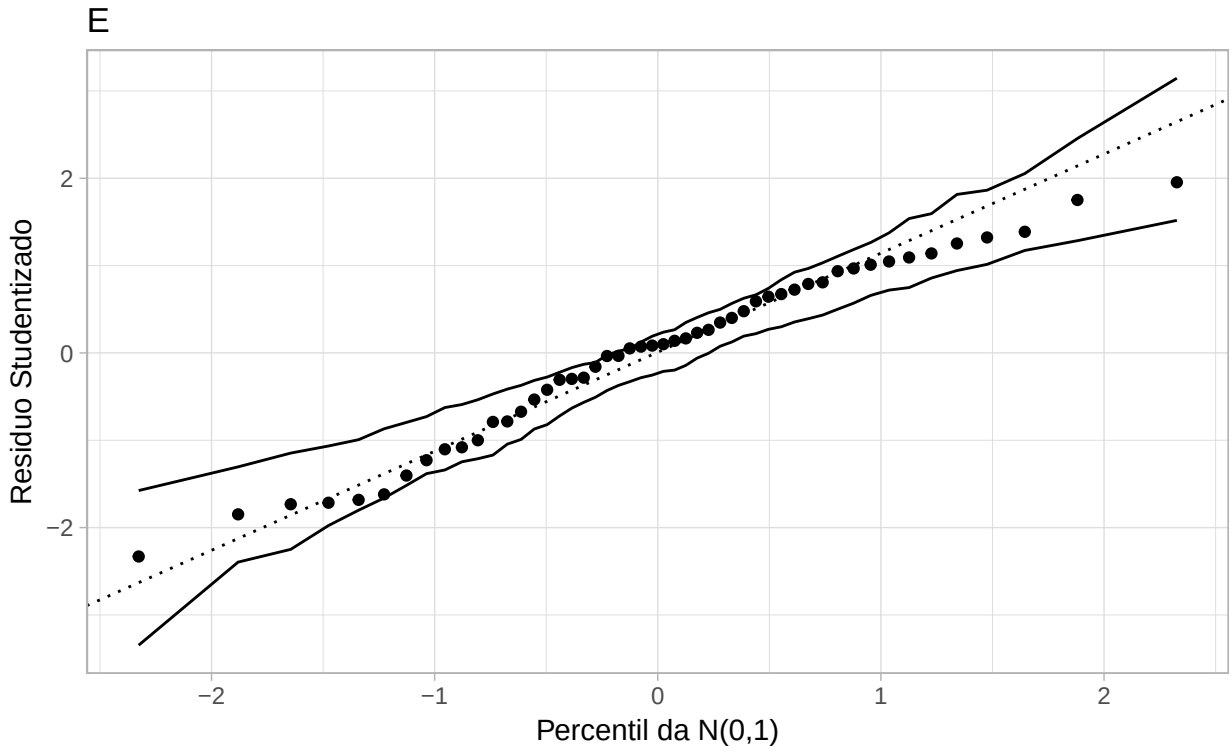

Figura 5: Gráfico de envelope para os resíduos studentizado para o modelo 2

Note, pela tabela 3 que todos os parâmetros do modelo são significativos a qualquer nível de significância usual (0,01 a 0,1). Portanto, temos indícios de que este é o modelo mais reduzido que, neste caso, pode-se ter sem perder poder preditivo do modelo. Vamos então, a partir deste ponto, analisar outros aspectos relacionados ao modelo que teve melhor ajuste.

Tabela 3: Estimativas dos parâmetros, intervalo de confiança e teste de nulidade para o modelo 2

	Estimativa	EP	IC (95%)	Valor T	Valor p
β_0	70.8786	0.1028	[70.6716 ; 71.0856]	689.6565	<0.0001
β_3	-1.0553	0.1328	[-1.3228 ; -0.7878]	-7.9456	<0.0001
β_4	0.3526	0.1236	[0.1035 ; 0.6016]	2.8516	0.0065
β_5	-0.3712	0.1247	[-0.6223 ; -0.1201]	-2.9774	0.0047
β_6	-0.1882	0.1079	[-0.4055 ; 0.0292]	-1.7439	0.088

3.2 Multicolinearidade

Mesmo o modelo 2 não apresentando muitas covariáveis, é interessante, a fim de assegurar a validade dos resultados obtidos (desconsidere que o modelo não teve um bom ajuste), verificar também a possibilidade de se ter multicolinearidade presente no modelo ajustado. Com o propósito de identificar alguma indicação de multicolinearidade no modelo, podemos análisar os coeficiêntes de corelação linear entre as covariáveis “crime”, “estud”, “ndias” e “dens”. Porém, observamos pela tabela 4 e figura 6,abaixo, que nenhum par de covariáveis apresenta coeficiente de correlação deveras alto. Aliás, a natureza das covariáveis presentes no medelo não apresentarem nenhum indício de serem fontes de multicolinearidade.

Tabela 4: Tabela dos coeficientes de correlação linear entre as covariáveis

	crime	estud	ndias	dens
crime	1,0000	-0,4880	-0,5389	0,1110
estud	-0,4880	1,0000	0,3668	-0,2668
ndias	-0,5389	0,3668	1,0000	-0,1329
dens	0,1110	-0,2668	-0,1329	1,0000

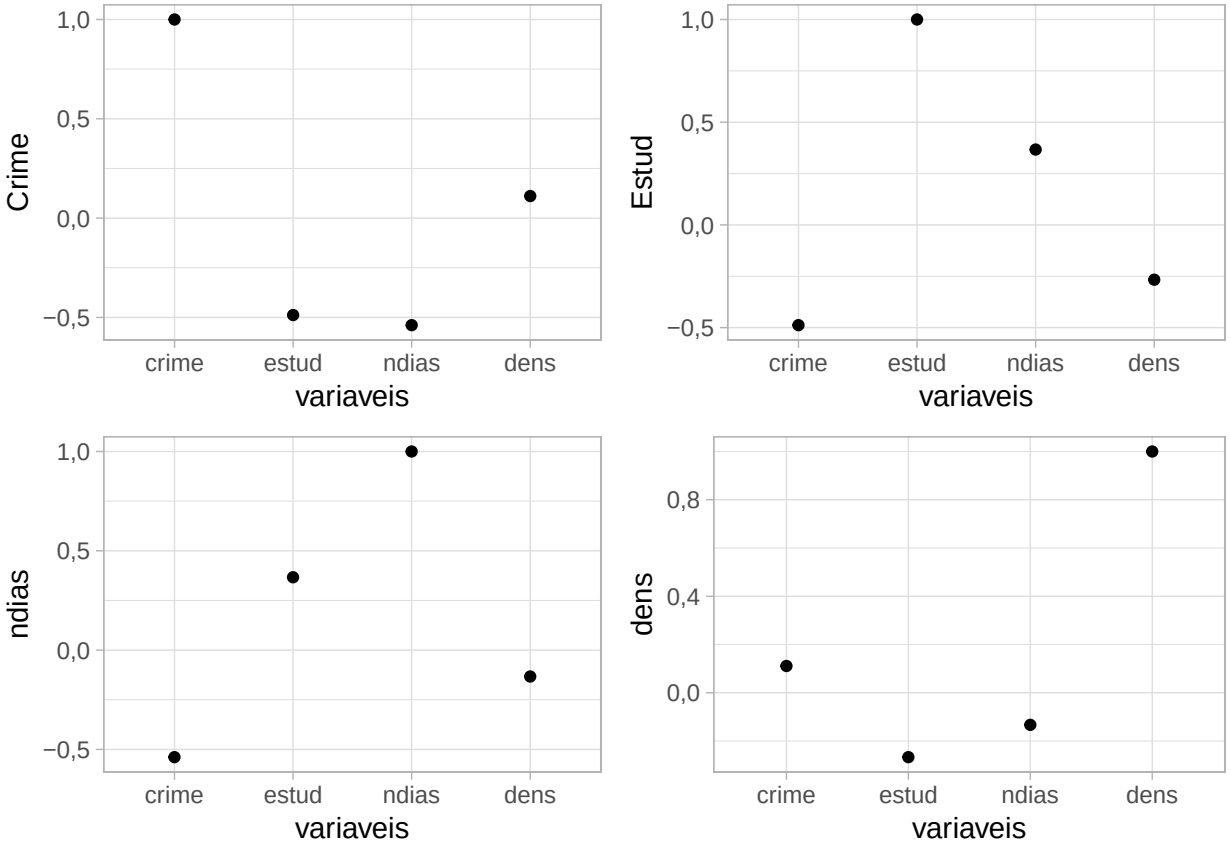

Figura 6: Gráfico de correlação

Além disso, sabe-se que se a razão entre o maior autovalor (λ_{max}) e o menor autovalor (λ_{min}) da matriz X^tX (o chamado índice de condição) $K = \frac{\lambda_{max}}{\lambda_{min}}$ for maior que mil, geralmente, há indicios de multicolinearidade (veja Azevedo(2016)). Fazendo isso para o caso do problema em questão, tem-se que: $K = 4{,}8048$. Como $K < 1000$, portanto descartaremos a hipótese de multicolinearidade dos dados.

3.3 Alavancagem

Observando a figura 7 podemos identificar a existência de pontos distantes dos demais, o que é uma indicação de alavancagem.

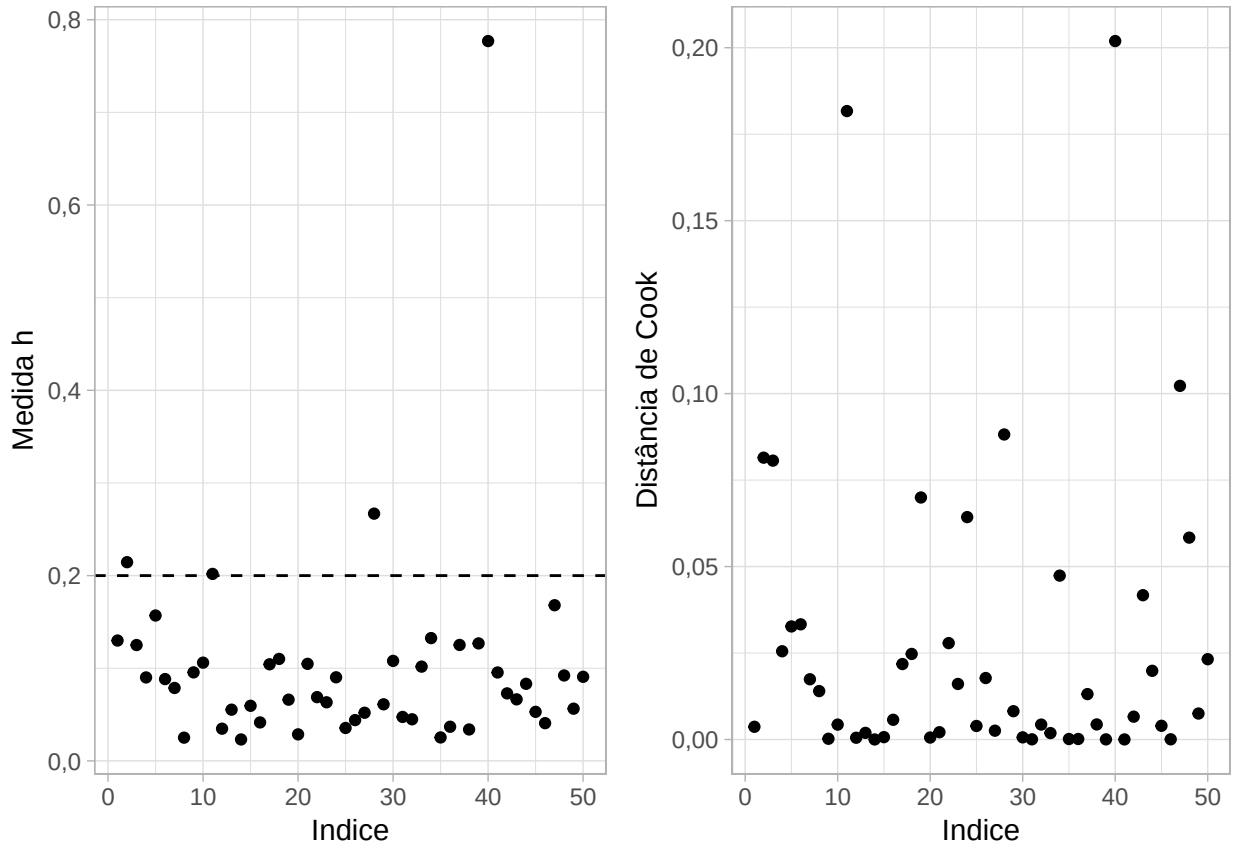

Figura 7: Pontos alavanca e distância de Cook

Vemos para o gráfico de Medida h, o ponto que se destaca é o de indice 40, com valor de 0,777. Já no de Distânca de Cook, temos dois pontos em destaque: um é o do mesmo indice, com valor 0,2019, outro ponto é o de indice 11, com valor 0,1817.

Logo, como identificamos esses pontos, vamos montar um modelo excluindo, primeiramente a observação mais discrepante, e depois excluindo ambas as observações discrepantes, ou seja, repetimos o modelo anterior, excluindo-se da base de dados a observação de indice 40, pois esta foi identificada como a mais discrepante, a qual possívelmente têm influência significativa no modelo. Caso ocorra uma melhora de ajuste, faremos o mesmo, excluíndo-se agora também o outro ponto de indice 11, identificado

pela Distância de Cook. Intitularemos o modelo sem a observação 40 como “Modelo 3” e o modelo sem as observações 40 e 11 como “Modelo 4”.

Segue abaixo os gráficos para análise residual do modelo sem a observação 40 (“Modelo 3”).

Note, pelas figuras 7 e 8 que a análise de resíduos, salvo à mínimos detalhes, continua a mesma da análise apresentada para o Modelo 2, ou seja, mesmo retirando a observação de indice 40 da base de dados, o modelo continua com o mesmo ajuste inadequado.

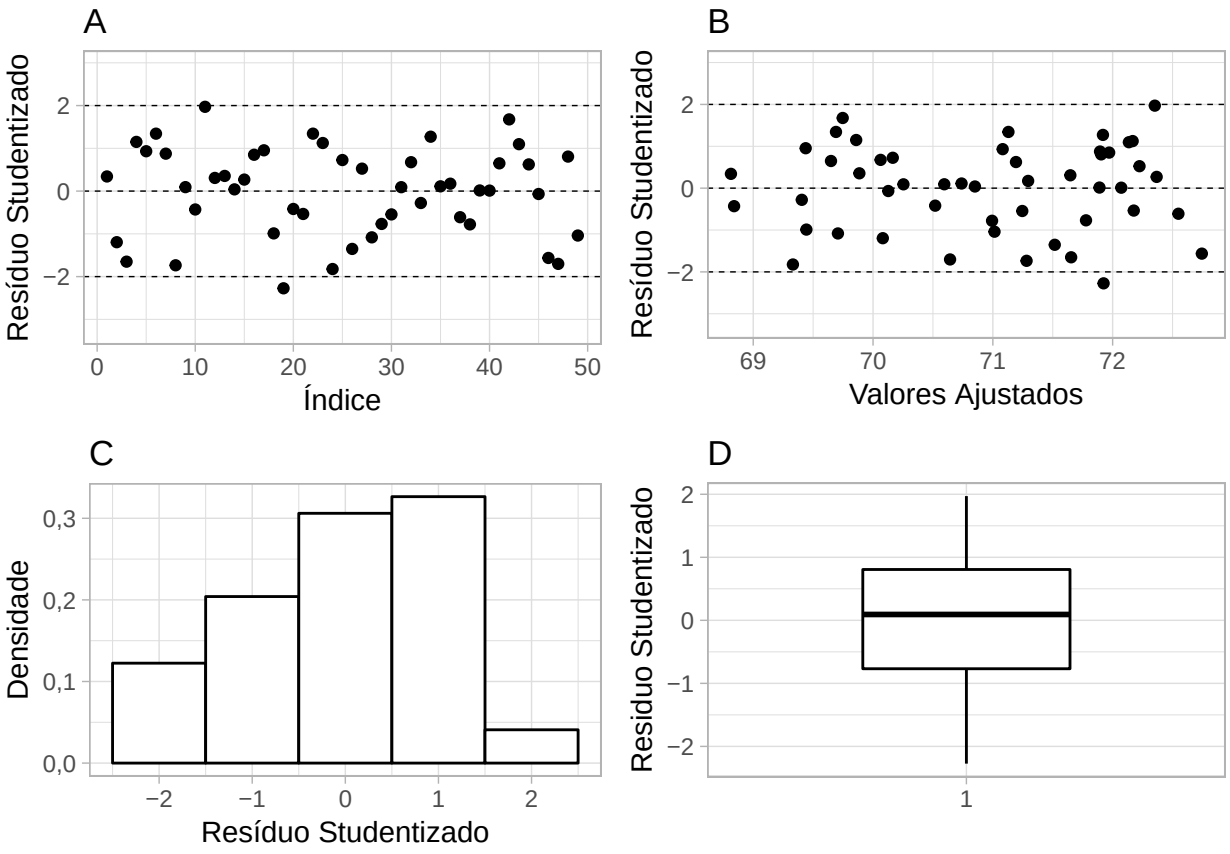

Figura 8: Análise residual para o modelo 3

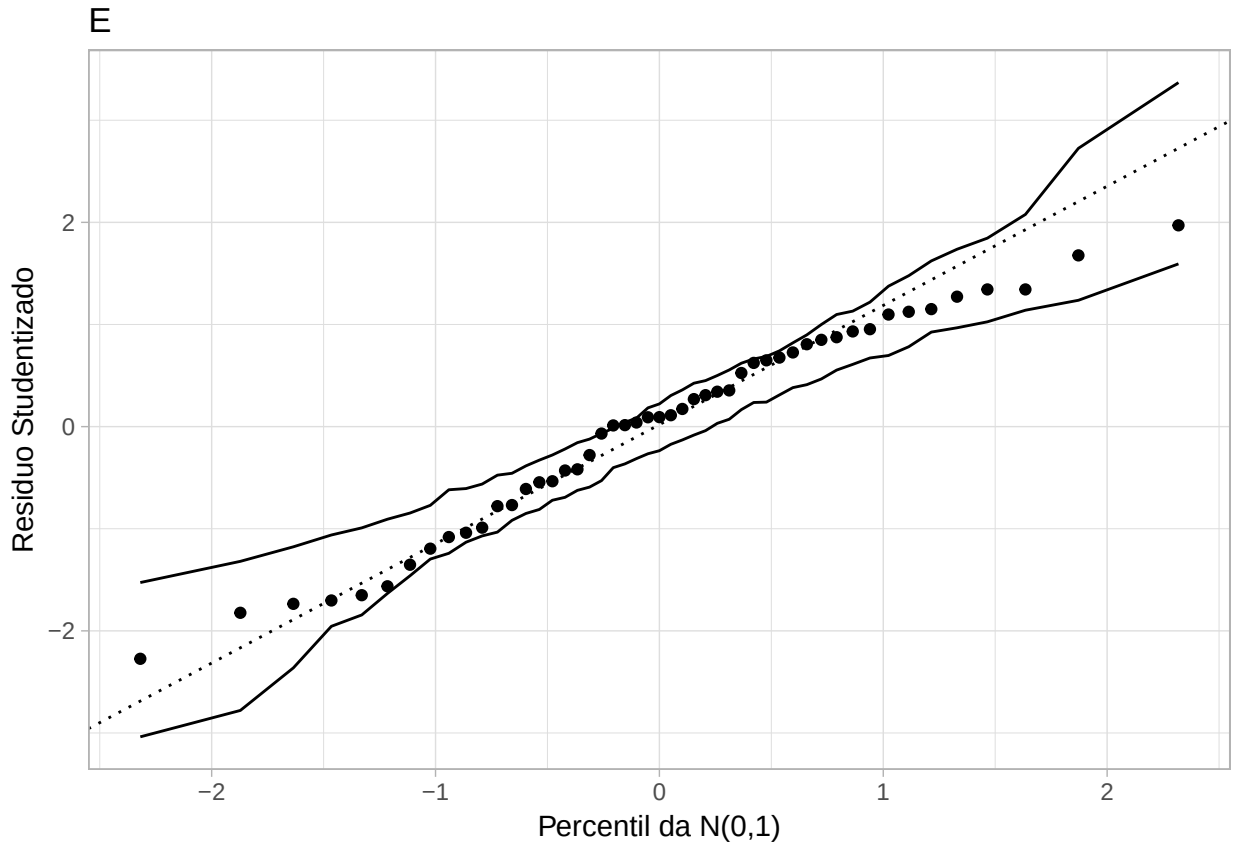

Figura 9: Gráfico de envelope para os resíduos studentizado para o modelo 3

Como se pode observar na tabela 5, ocorreu uma mudança significativa nos valores das medidas apresentadas na tabela, ou seja, existe indícios de que a observação 40 têm grande influência no modelo. Vamos então ajustar um modelo agora sem as observações 40 e 11 para avaliar as mudanças causadas em relação a este modelo (modelo 3).

Tabela 5: Tabela de Comparação dos modelos 1, 2 e 3

	Modelo 1	Modelo 2	Modelo 3
AIC	119,5163	116,7045	110,0140
BIC	134,8125	128,1766	121,2412
Log-Verossimilhança	-51,7581	-52,3522	-49,0070
R2	0,7372	0,7309	0,7008
R2 ajustado	0,7005	0,7069	0,6729

Segue abaixo os gráficos para análise residual do modelo sem as observações 40 e 11 (“Modelo 4”).

Note, pelas figuras 9 e 10 que a análise de resíduos, salvo à mínimos detalhes, continua a mesma da análise apresentada

para o Modelo 2, ou seja, mesmo retirando as observação de indices 40 e 11 da base de dados, o modelo continua com o mesmo ajuste inadequado.

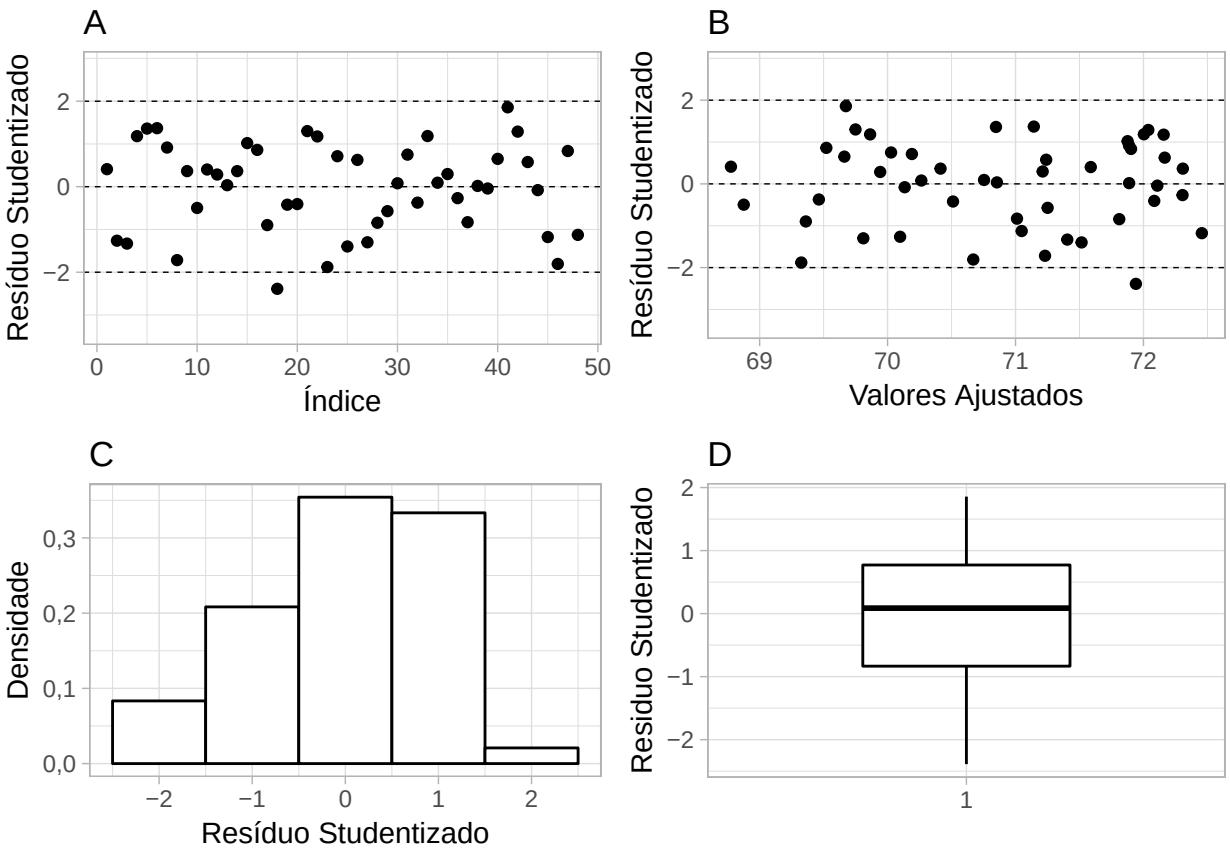

Figura 10: Análise residual para o modelo 4

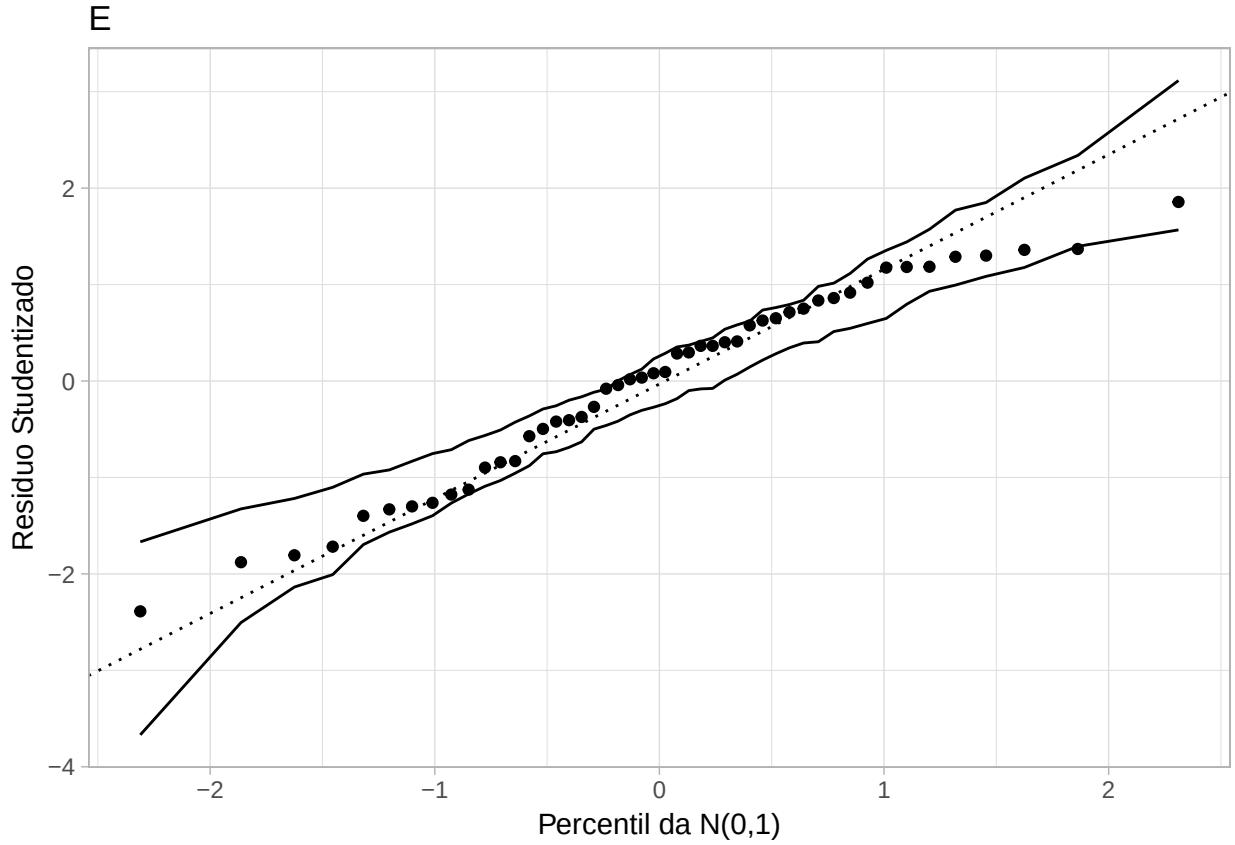

Figura 11: Gráfico de envelope para os resíduos studentizado para o modelo 4

Note agora, que não ocorreram mudanças significativas nos valores das medidas apresentadas pela tabela 5 deste modelo (modelo 4) com relação aos valores das medidas do modelo anterior (modelo 3), portanto, não existem indícios de que a observação 11 tenha grande influência no modelo.

Tabela 6: Tabela de Comparação dos modelos 1, 2, 3 e 4

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
AIC	119,5163	116,7045	110,0140	110,0140
BIC	134,8125	128,1766	121,2412	121,2412
Log-Verossimilhança	-51,7581	-52,3522	-49,0070	-49,0070
R2	0,7372	0,7309	0,7008	0,7008
R2 ajustado	0,7005	0,7069	0,6729	0,6729

4. Conclusões

Como visto, nenhum modelo destes propostos teve um bom ajuste ao conjunto de dados, não obstante, dado o escopo do curso, e dados os resultados mensionados anteriormente, consideramos que o modelo 3 foi o que teve ajuste mais satisfatório, pois é o que apresenta as melhores estimativas de comparação de modelo, ou seja valores baixos para AIC e BIC, e alto valor de Log-Verossimilhança, embora tenha valores de R2 e R2 ajustado menores que os valores apresentados para os modelos 1 e 2. Quanto aos valores das estatísticas apresentados na tabela 5, os modelos 3 e 4 são similares, porém o modelo 4 exclui uma observação desnecessariamente. Porém, levando-se em consideração que não estamos em contato com o experimentador, e por isso não teremos um entendimento maior quanto aos dados, não sabemos se a exclusão da observação 40 é válida, ou não, para produzir um modelo mais adequado.

5. Referências Bibliográficas

- Azevedo, C. L. N (2016). Notas de aula sobre planejamento e análise de experimentos,http://www.ime.unicamp.br/~cnaber/Material_ME613_2S_2016.htm
- Faraway, J. J. (2014). Linear Models with R, Second Edition,Chapman e Hall/CRC Texts in Statistical Science
- Draper, N. R. and Smith, H. (1998). Applied regression analysis, third edition. New York, NY: John Wiley e Sons.
- Paula, G. A. (2013). Modelos de regressão com apoio computacional, versão pré-eliminar https://www.ime.usp.br/~giapaula/texto_2013.pdf