trabalho Regressão parte 02

6 de dezembro de 2016

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 em 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:

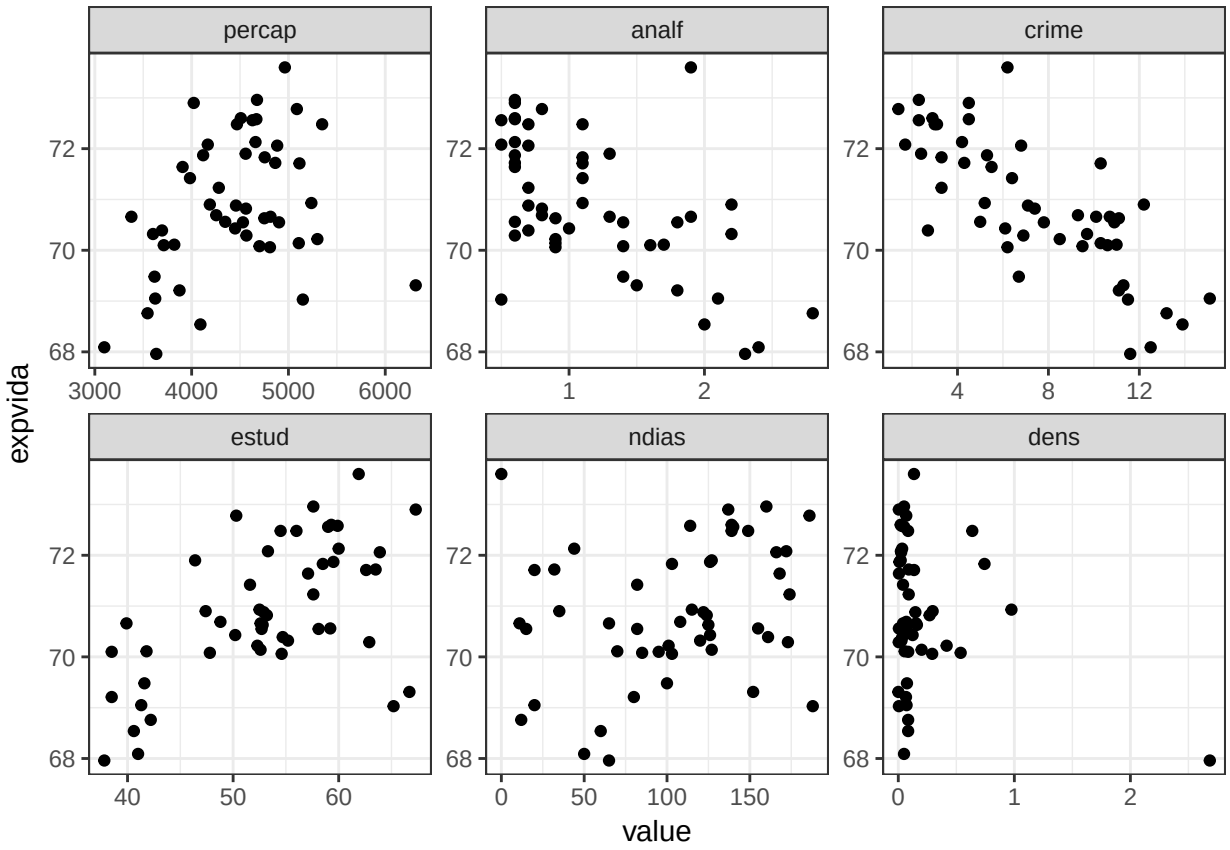
Y_i - i-ésima observação da variável expvida
x_{1i} - i-esima observação da variável percap
x_{2i} - i-esima observação da variável analf
x_{3i} - i-esima observação da variável crime
x_{4i} - i-esima observação da variável estud
x_{5i} - i-esima observação da variável ndias
x_{6i} - i-esima observação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

2.Análise Descritiva

variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
percap	3098.0000	3993.0000	4519.000	4436.000	4814.0000	6315.000
analf	0.5000	0.6250	0.950	1.170	1.5750	2.800
crime	1.4000	4.3500	6.850	7.378	10.6800	15.100
estud	37.8000	48.0500	53.250	53.110	59.1500	67.300
ndias	0.0000	66.2500	114.500	104.500	139.8000	188.000
dens	0.0006	0.0277	0.069	0.188	0.1443	2.684
expvida	67.9600	70.1200	70.680	70.880	71.8900	73.600

Tabela 1: Estatísticas Descritivas

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, foi construido graficos de dispersão entre a variável resposta e entre todas as covariáveis de interesse. Estes estão representados na figura XX.



A partir da FiguraXX, podemos identificar visualmente a relação supracitada, a qual identificamos que todas as covariáveis parecem ter uma relação linear significativa, as quais existem relações lineares negativas entre a variável resposta 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).

Análise Inferêncial

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:

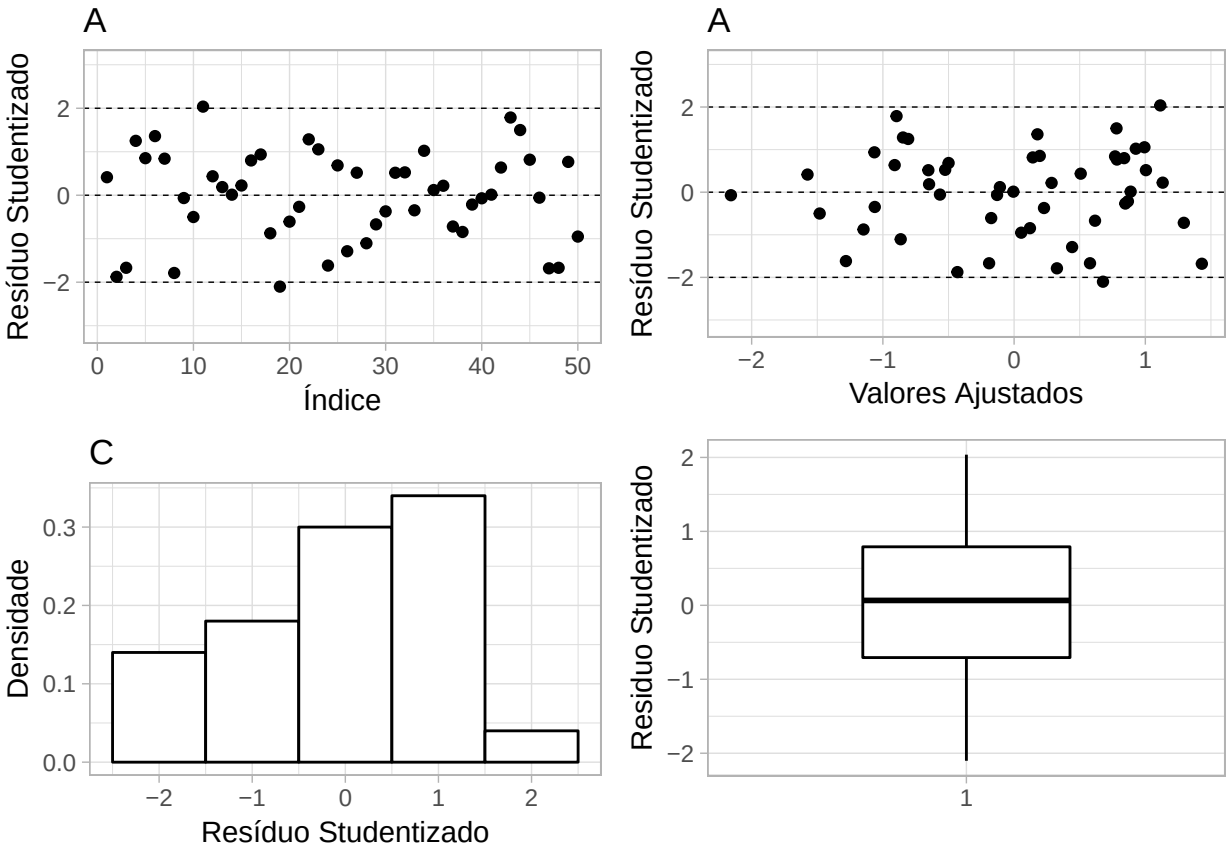
$$Y_i = \beta_0 + \sum_{j=1}^{6} \beta_j \left(\frac{x_{ji} - \bar{x_j}}{s_j} \right) + \varepsilon_i \begin{cases} i = 1, \ldots, 50 \\ j = 1, \ldots, 6 \end{cases}$$

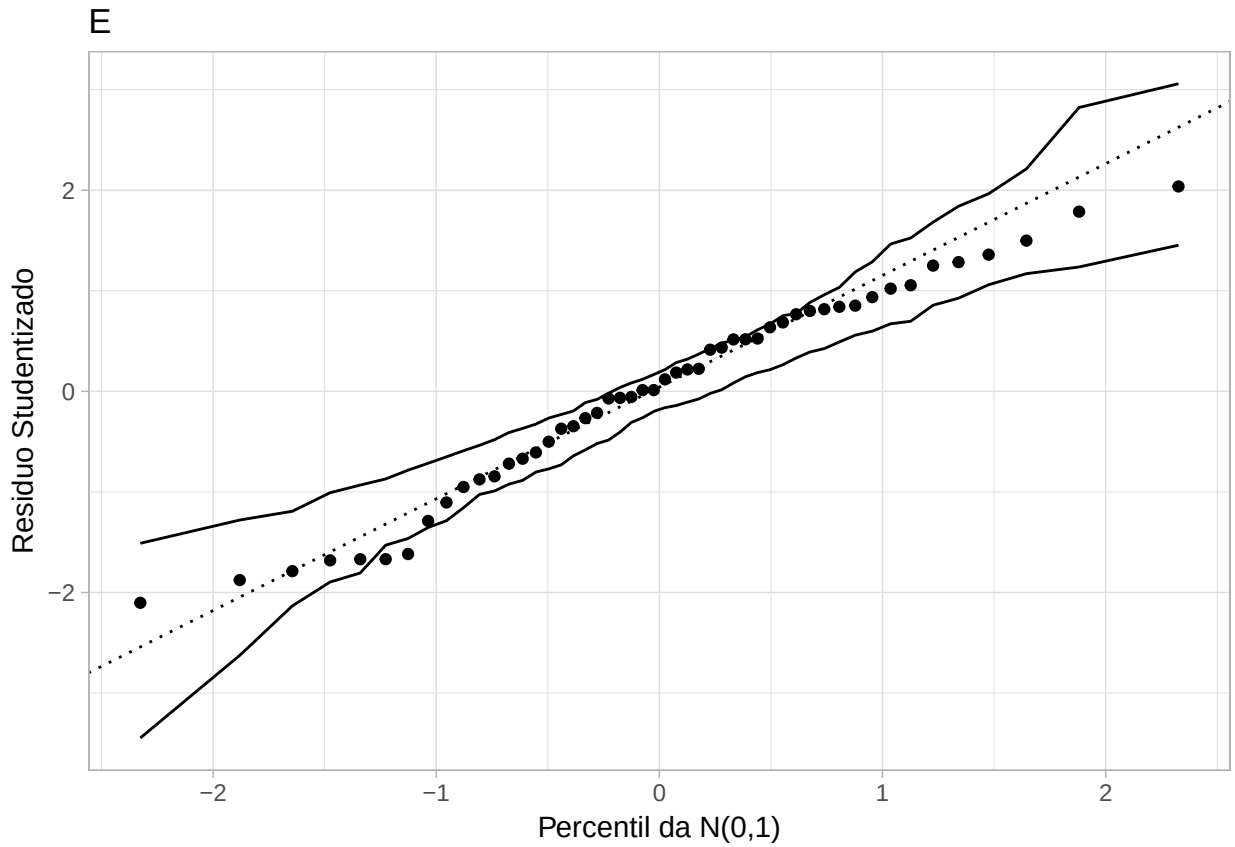
em que $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $x_j = \frac{1}{50} \sum_{i=1}^{50} x_j i$ e $s_j = \sqrt{\frac{1}{50} \sum_{i=1}^{50} (x_{ji} - \bar{x_j})^2}$

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 das figuras XX e XX que o modelo não teve um ajuste adequado. No gráfico A não observação tendências,o que pode indicar uma independência dos dados. Já no gráfico B, observa-se uma leve heterocedasticidade, dado que entre os valores -2 e -1 de Valores Ajustados vê-se uma discrepâncias entre os resíduos studentizados do que entre os resíduos compreendidos entre os valores de -1 a 2. Porém, um fator de confundimento que pode ser levado em consideração é a flutuação amostral pois o banco de dados possui apenas cinquenta observações. é possível observar uma assimetria positiva no histograma apresentado (gráfico C da figura XX), assim como uma tendencia no gráfico de envelopes (figura XX) 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. Dado o escopo do curso, procede-se com as análises posteriores.





Note, pela tabela XX(lembrando que os valores da tabela estão arredondados em quatro casas decimais) que somente os parêmetros relacionados às covariáveis “crime”, “ndias” e “dens” significativos à um nível de significancia = 0,10. 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.

Estimativa	EP	Valor T	Valor p
0.0000	0.0774	0.0000	1
0.1007	0.1046	0.9631	0.3409
-0.0240	0.1497	-0.1601	0.8735
-0.7910	0.1137	-6.9535	<0.0001
0.1829	0.1246	1.4677	0.1495
-0.2910	0.1076	-2.7058	0.0097
-0.1571	0.0848	-1.8518	0.0709

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 com que melhor se ajusta é o modelo que leva em consideração somente as covariáveis “crime”,“estud”,“ndias” e “dens”.

Temos agora o seguinte modelo:

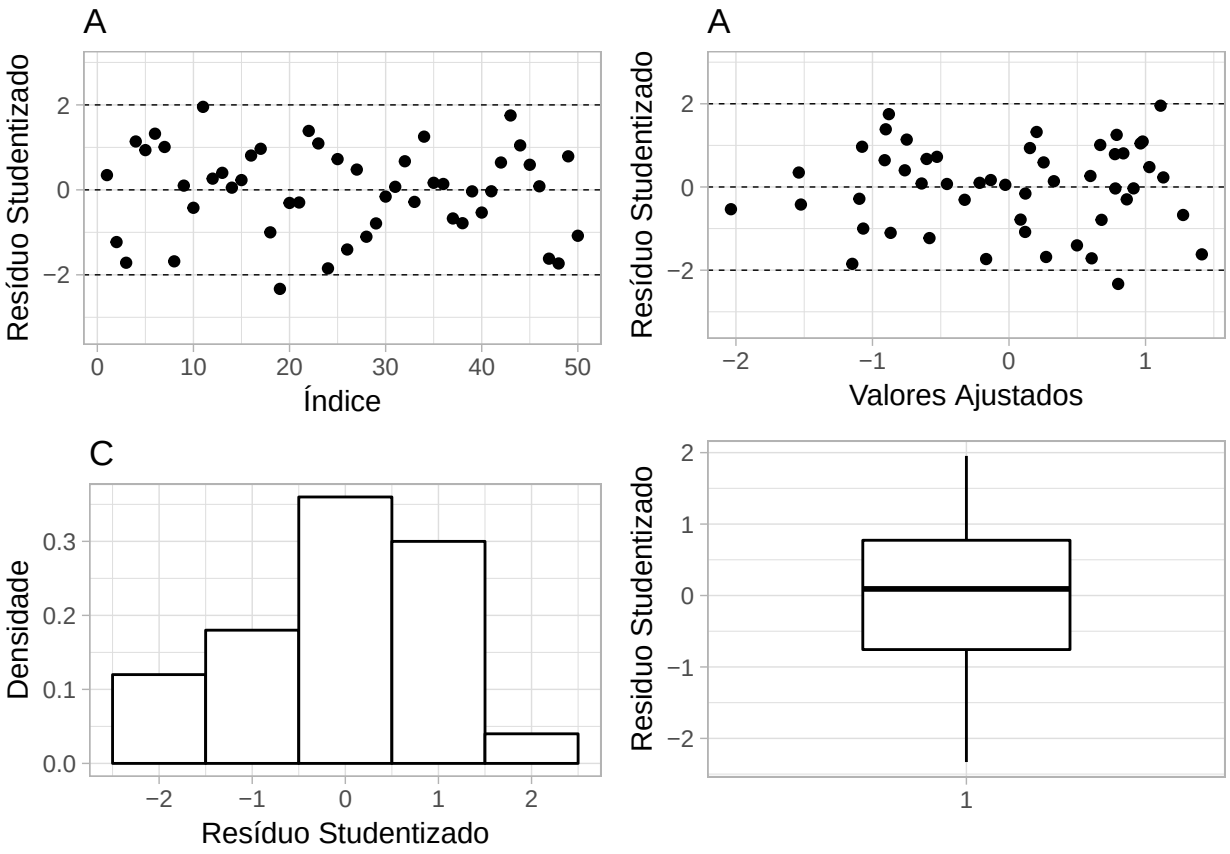
Modelo 2:

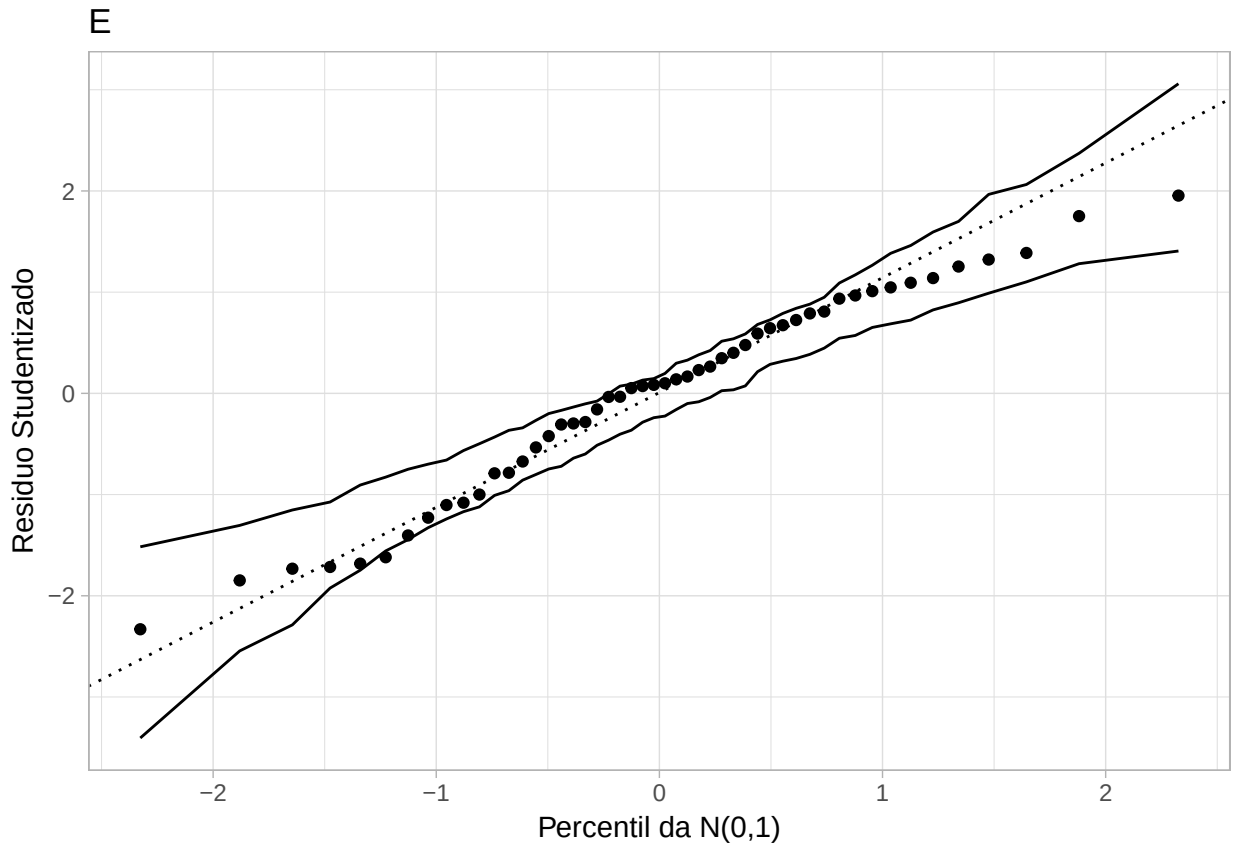
$$Y_i = \beta_0 + \sum_{j=3}^{6} \beta_j \left(\frac{x_{ji} - \bar{x_j}}{s_j} \right) + \varepsilon_i \begin{cases} i = 1, \ldots, 50 \\ j = 3, \ldots, 6 \end{cases}$$

em que $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $x_j = \frac{1}{50} \sum_{i=1}^{50} x_j i$ 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 nas figuras YY e XX indícios de mal ajuste semelhantes aos observados na análise residual para o modelo anterior. Porém uma ressalva pode ser feita de que no gráfico B pode-se observar que as discrepâncias entre os resíduos mudam conforme o valor ajustado aumenta
Neste caso, deveria-se procurar modelos alternativos que levem em consideração uma distribuição assimetrica. Novamente, dado o escopo do curso, procede-se com as análises posteriores.


A
Resíduo Studentizado
2
0
−2
0
10
20
30
40
50
Índice
A
Resíduo Studentizado
2
0
−2
−2
−1
0
1
Valores Ajustados
C
Densidade
0.3
0.2
0.1
0.0
−2
−1
0
1
2
Resíduo Studentizado
Resíduo Studentizado
2
1
0
−1
−2
1



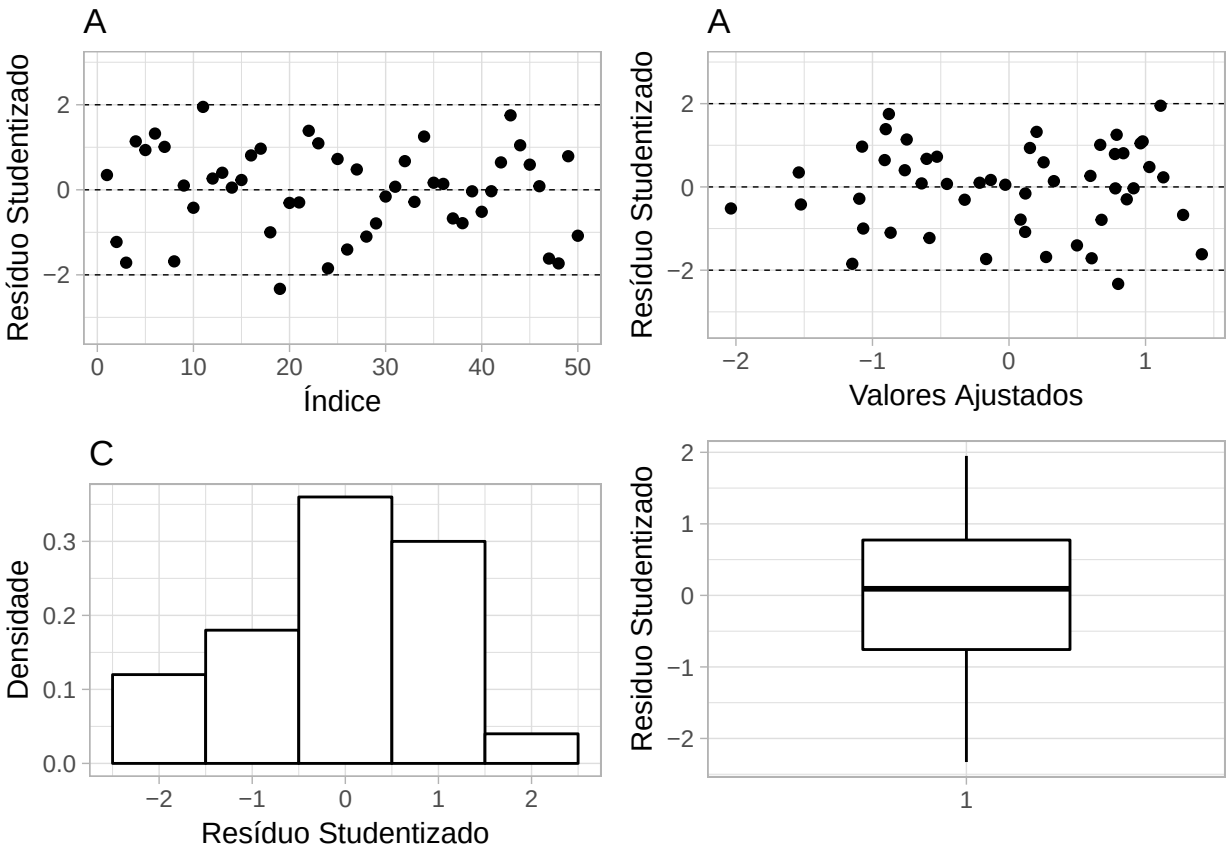
Note que a tabela acima acusa que o parâmetro β_0 não é significativo no modelo para um nível de significância de 0,10. Sendo assim, vamos ajustar o modelo anterior sem intercepto.

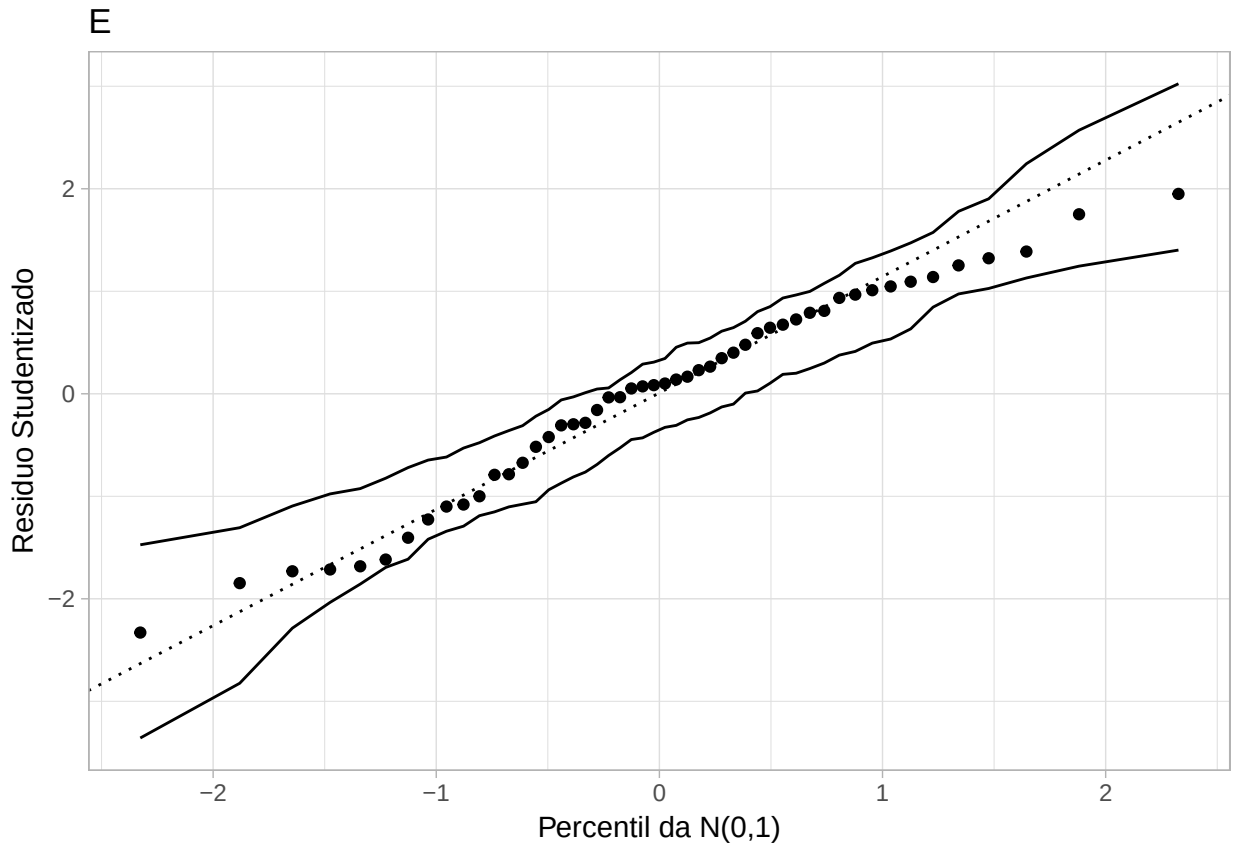
Modelo 3:

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

em que $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $x_j = \frac{1}{50} \sum_{i=1}^{50} x_j i$ 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).
$\frac{\beta_3}{s_3}$: 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.


A
Resíduo Studentizado
2
0
−2
0
10
20
30
40
50
Índice
A
Resíduo Studentizado
2
0
−2
−2
−1
0
1
Valores Ajustados
C
Densidade
0.3
0.2
0.1
0.0
−2
−1
0
1
2
Resíduo Studentizado
Resíduo Studentizado
2
1
0
−1
−2
1



A tabela XX abaixo mostra as principais medidas de comparação entre os dois últimos modelos (com e sem intercepto). Observamos nela, valores menores de AIC e BIC para o modelo sem intercepto, o que nos da uma indicação de que o modelo sem intercepto é preferível.

	Modelo 2	Modelo 3
AIC	87.25904	85.25904
BIC	98.73118	94.81915
Log-Verossimilhança	-37.62952	-37.62952
R2	0.73090	0.73090
R2 ajustado	0.70690	0.70740

Multicolinearidade

Mesmo o modelo 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 na tabela XX (abaixo) que nenhum par de covariáveis apresenta coeficiente de correlação deveras alto, assim como a natureza das covariáveis presentes no medelo não apresentarem nenhum indício de serem fontes de multicolinearidade.

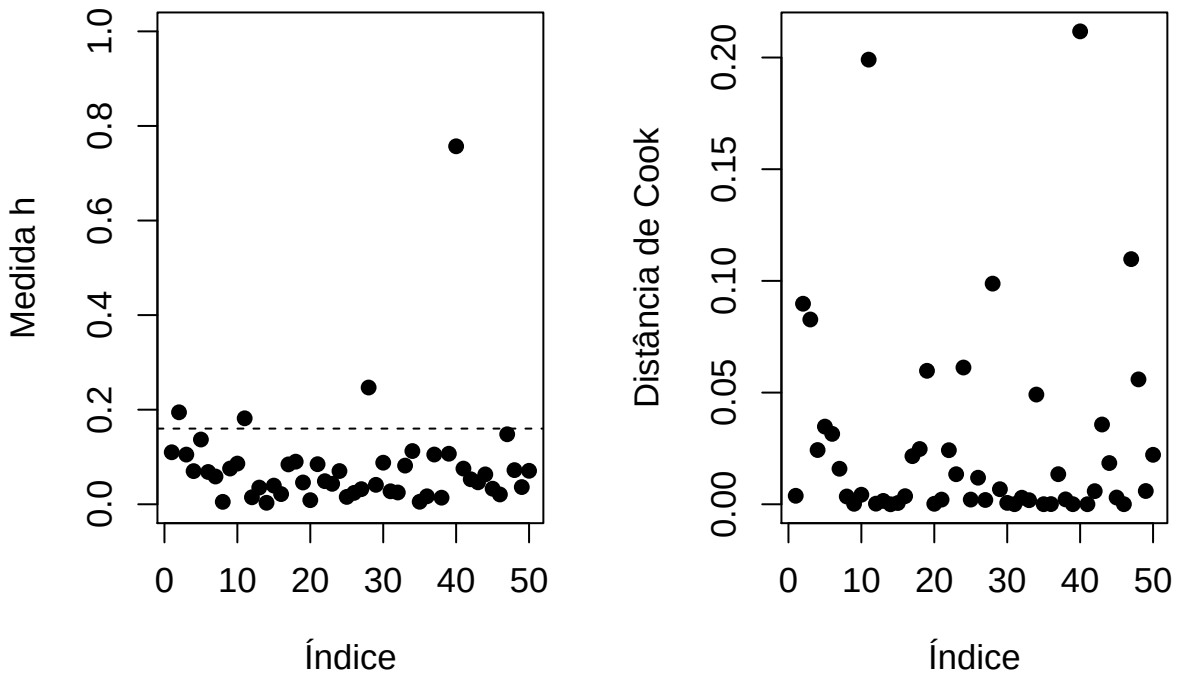
	crime	estud	ndias	dens
crime	1.0000000	-0.4879710	-0.5388834	0.1110318

	crime	estud	ndias	dens
estud	-0.4879710	1.0000000	0.3667797	-0.2667561
ndias	-0.5388834	0.3667797	1.0000000	-0.1329066
dens	0.1110318	-0.2667561	-0.1329066	1.0000000

Além disso, Sabe-se que se a razão entre o maior autovalor (λ_{max}) e o menor (λ_{min}) (o chamado índice de condição) $K = \frac{\lambda_{max}}{\lambda_{min}}$ for maior que mil, geralmente, há indicios de multicolinearidade. 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.

Alavancagem

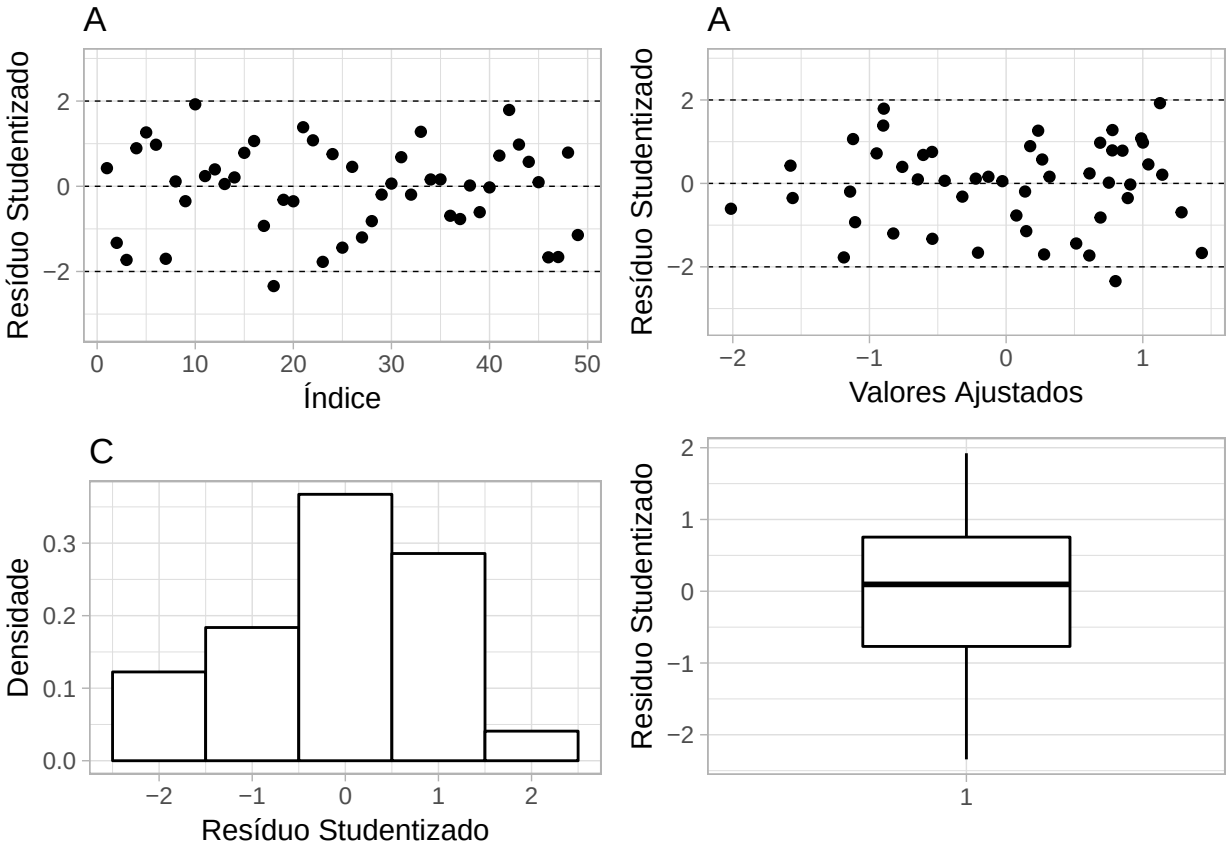
Observando a figura XX podemos identificar a existência de pontos distantes dos demais, o que é uma indicação de alavancagem.

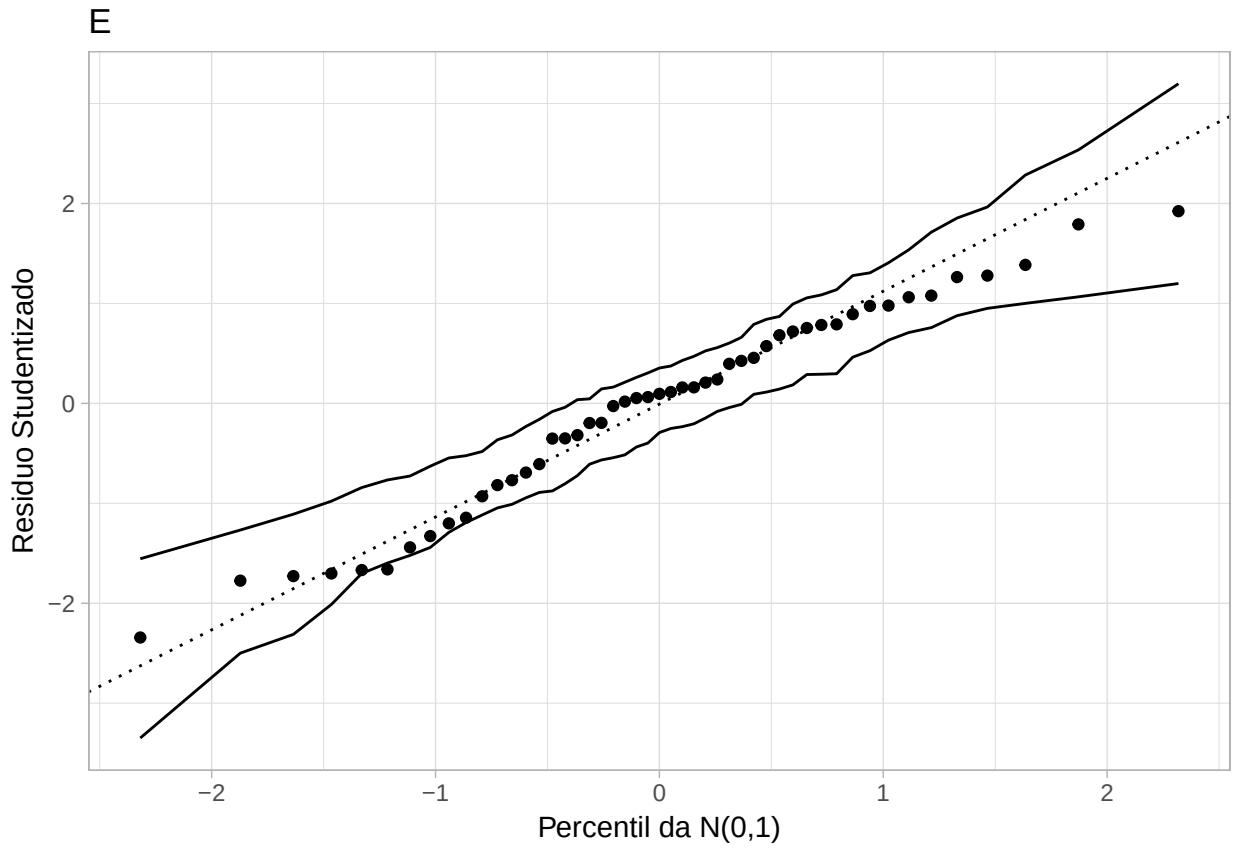


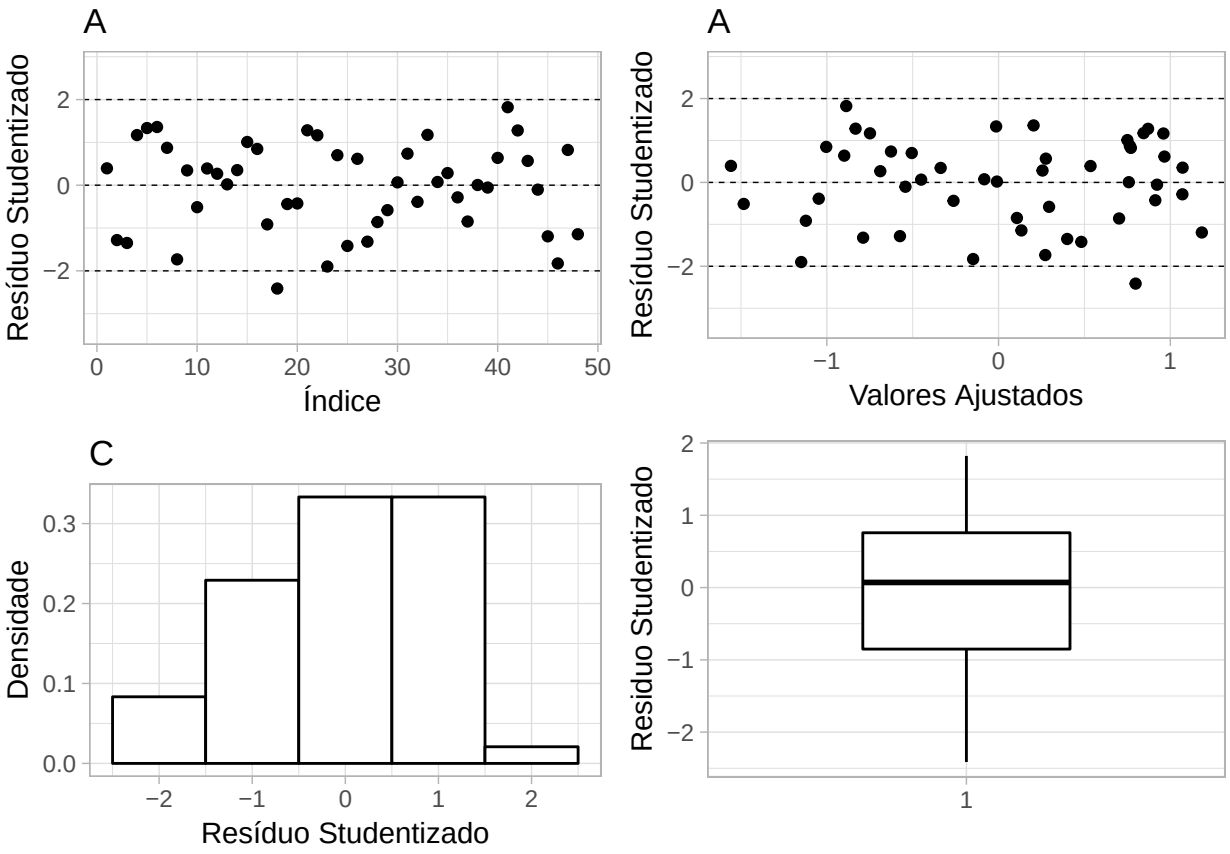
Vemos para o gráfico de Medida h, o ponto que se destaca é o de indice 40, com valor de 0.7569594. Já no de Distânca de Cook, temos dois pontos em destaque: um é o do mesmo indice,com valor 0.211707, coomparações entre os valores das duas medidas não consistentes e portanto inválidas, outro ponto é o de indice 11, com valor 0.1990533.

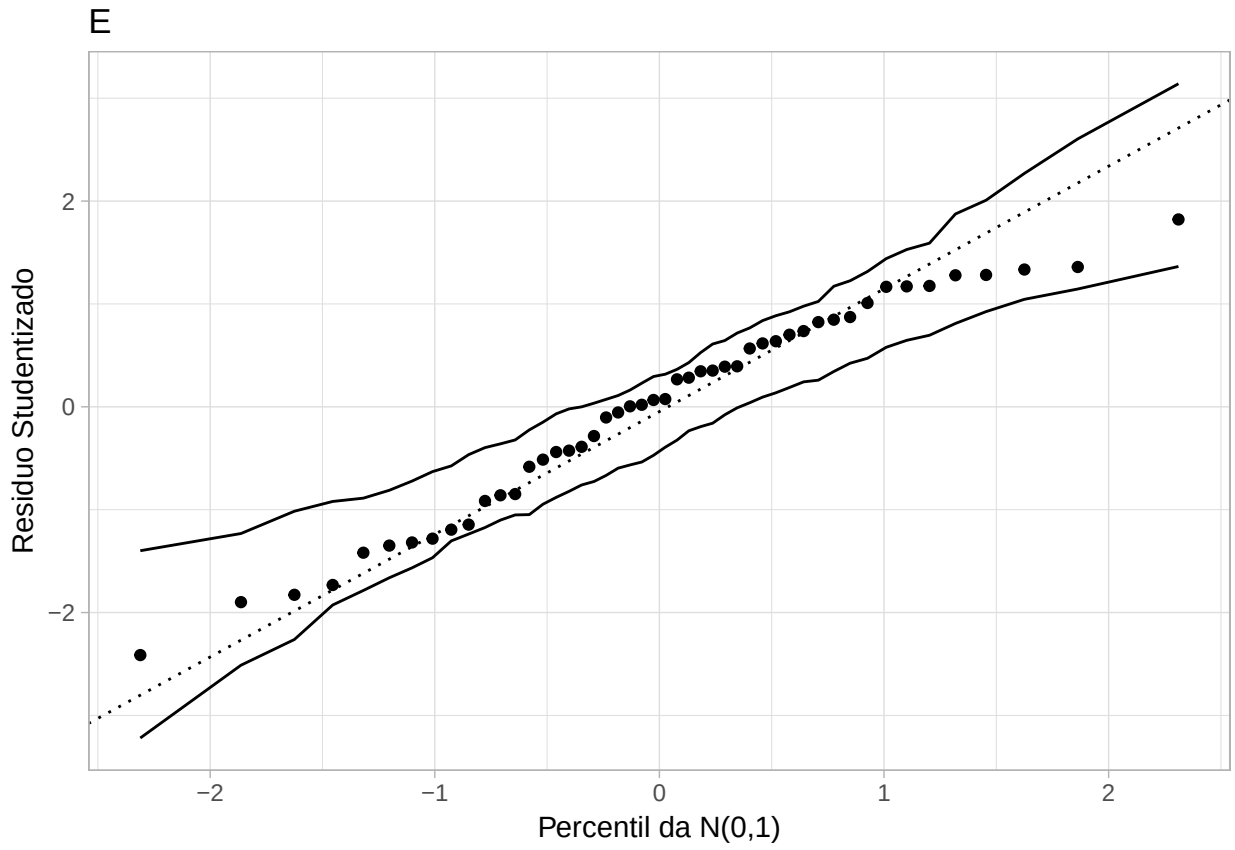
Logo, como identificamos esses pontos, vamos montar um modelo reduzido, como feito acima mas excluindo primeiramente as observações de indice 40 pois nos dois caso é o ponto mais elevado, possívelmente o que mais influência no modelo. caso ocorra uma melhora de ajuste, faremos para o outro identificado pela Distância de Cook.

Segue abaixo o ajuste do modelo sem a observação 40




E
Residuo Studentizado
2
0
−2
−2
−1
0
1
2
Percentil da N(0,1)


A
Resíduo Studentizado
Índice
A
Resíduo Studentizado
Valores Ajustados
C
Densidade
Resíduo Studentizado
Resíduo Studentizado



	Modelo.1	Modelo.2	Modelo.3	Modelo.4	Modelo.5
AIC	90.07085	87.25904	85.25904	83.35105	79.76038
BIC	105.36704	98.73118	94.81915	92.81015	89.11639
Log-Verossimilhança	-37.03543	-37.62952	-37.62952	-36.67552	-34.88019
R2	0.73720	0.73090	0.73090	0.73830	0.70070
R2 ajustado	0.70050	0.70690	0.70740	0.71500	0.67350

Conclusões

Como visto, nenhum modelo destes propostos teve um bom ajuste ao conjunto de dados não obstante utilizaremos o modelo quatro pois é o que apresenta as melhores estimativas de comparação de modelo, ou seja baixo AIC E BIC, alta Log-Verossimilhança. Podemos observar também que há uma relaç