

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 Celsius 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

```
## Loading tidyverse: ggplot2
```

```
## Loading tidyverse: tibble
```

```
## Loading tidyverse: tidyr
```

```
## Loading tidyverse: readr
```

```
## Loading tidyverse: purrr
```

```
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages -----
```

```
## arrange(): dplyr, plyr
```

```
## compact(): purrr, plyr
```

```
## count(): dplyr, plyr
```

```
## failwith(): dplyr, plyr
```

```
## filter(): dplyr, stats
```

```
## id(): dplyr, plyr
```

```
## lag(): dplyr, stats
```

```
## mutate(): dplyr, plyr
```

```
## rename(): dplyr, plyr
```

```
## summarise(): dplyr, plyr
```

```
## summarize(): dplyr, plyr
```

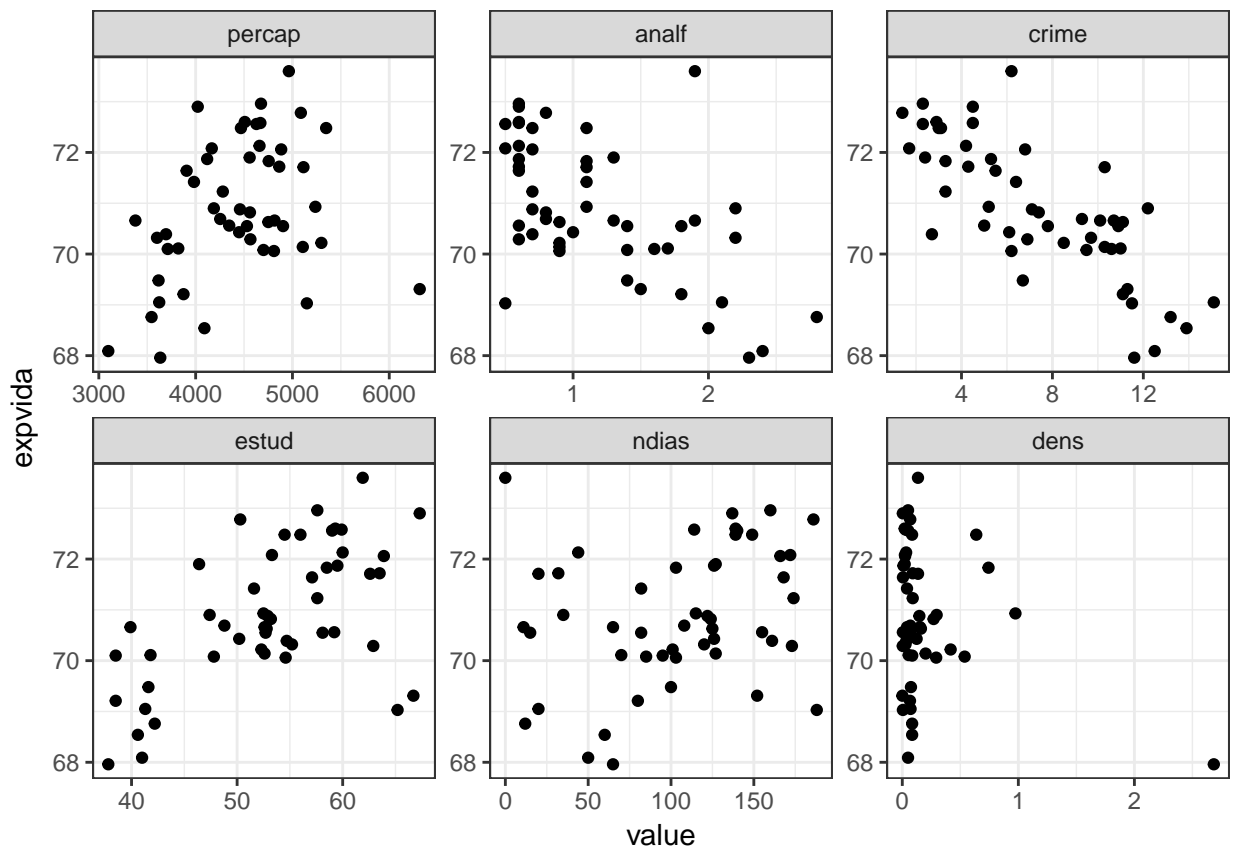
2. Análise Descritiva

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
## smiths
```

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

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 construído gráficos 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 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 coeficientes do modelo, optamos por introduzir as covariáveis com médias e variâncias iguais, tornando-as adimensionais. Portanto, vamos iniciar a análise com o seguinte modelo:

Modelo: !!!note que pode-se resumir a interpretação de B_j/s_j caso precise de espaço!!!

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

em que $\varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$, $x_j = \frac{1}{50} \sum_{i=1}^{50} x_{ji}$ 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.

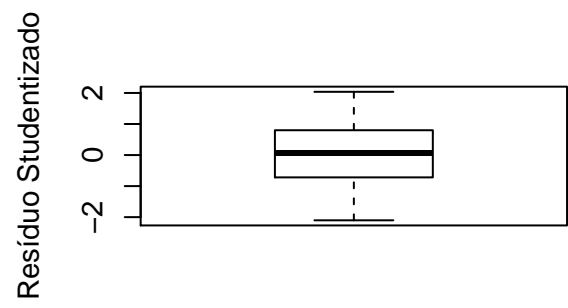
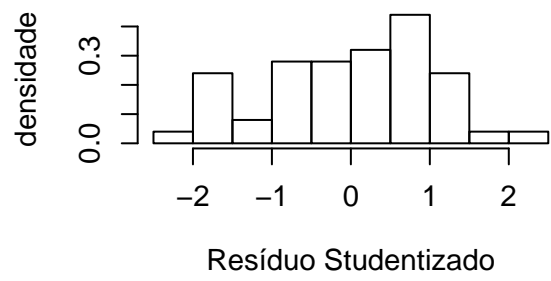
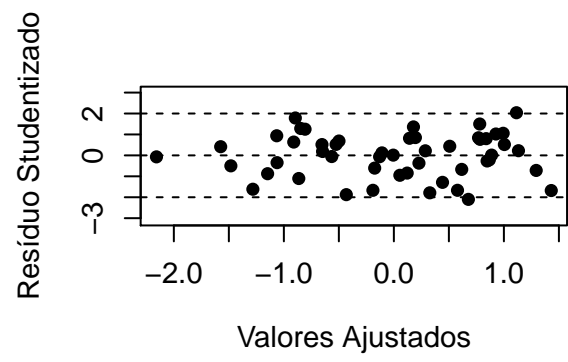
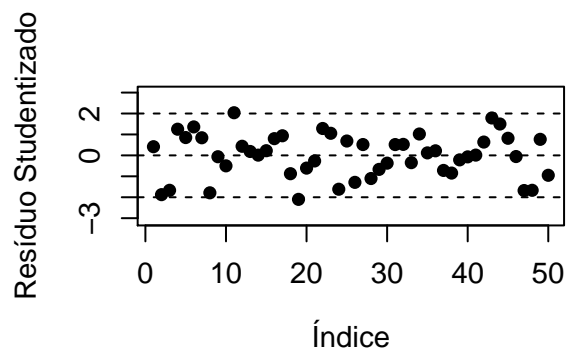
$\frac{\beta_3}{s_3}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da “taxa de criminalidade (por 100000 habitantes 1976)” em uma unidade, mantendo-se as demais covariáveis fixas.

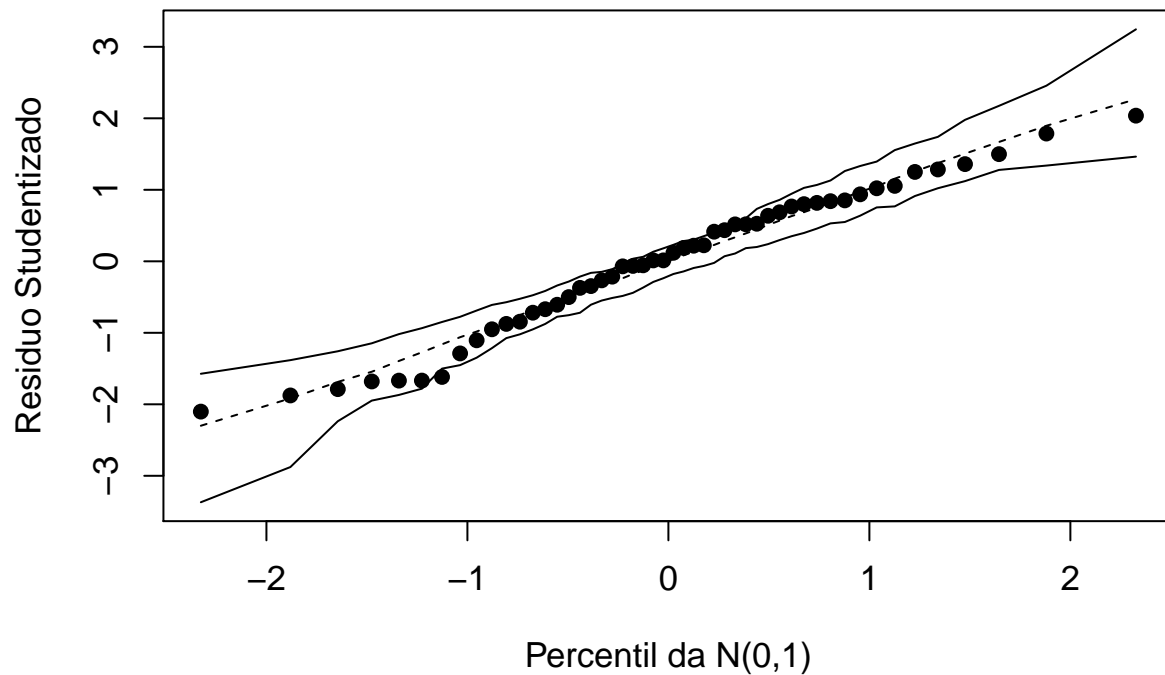
$\frac{\beta_4}{s_4}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da “porcentagem de estudantes que concluem o segundo grau (1970)” em uma unidade, mantendo-se as demais covariáveis fixas.

$\frac{\beta_5}{s_5}$: 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 Celsius 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, uma vez que se é possível observar uma assimetria positiva no histograma apresentado (gráfico YY da figura XX), assim como uma tendência no gráfico de envelopes (figura XX) o que nos dá indícios de falta de normalidade nos resíduos. **!!!eu não identifiquei heterocedasticidade, se alguém achar que têm, escreva algum argumento!!!**. 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 que somente os parâmetros relacionados às covariáveis “crime”, “ndias” e “dens” são significativos à um nível de significância = 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.

```
#pacote para produzir tabela em latex
library(knitr)
```

```
# !!! depois editar!!!
#coeficientes e R^2 e R^2 ajustado
R2fitmax <- summary(fitmax)$r.squared
R2fitmax
```

```
## [1] 0.7371725
```

```
R2aj_fitmax <- summary(fitmax)$adj.r.squared
R2aj_fitmax
```

```
## [1] 0.7004989
```

```
kable(summary(fitmax)$coefficients)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0000000	0.0773952	0.0000000	1.0000000
percap	0.1007060	0.1045688	0.9630598	0.3409024
analf	-0.0239757	0.1497144	-0.1601428	0.8735186
crime	-0.7909564	0.1137496	-6.9534863	0.0000000
estud	0.1829116	0.1246228	1.4677219	0.1494569

	Estimate	Std. Error	t value	Pr(> t)
ndias	-0.2910257	0.1075544	-2.7058468	0.0097285
dens	-0.1570562	0.0848135	-1.8517829	0.0709312

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).

```
## Start:  AIC=0.99
## expvida ~ 1
##
##           Df Sum of Sq    RSS      AIC
## + crime   1   29.8763 19.124 -44.055
## + analf   1   16.9690 32.031 -18.266
## + estud   1   16.6098 32.390 -17.708
## + percap   1    5.6729 43.327  -3.162
## + ndias    1    3.3653 45.635  -0.568
## + dens     1    3.3323 45.668  -0.532
## <none>                49.000   0.990
##
## Step:  AIC=-44.05
## expvida ~ crime
##
##           Df Sum of Sq    RSS      AIC
## + estud    1    2.6032 16.521 -49.371
## + ndias     1    1.7395 17.384 -46.823
## + dens      1    1.5034 17.620 -46.149
## + percap    1    1.3345 17.789 -45.671
## <none>                19.124 -44.055
## + analf     1    0.1516 18.972 -42.453
## - crime     1   29.8763 49.000   0.990
##
## Step:  AIC=-49.37
## expvida ~ crime + estud
##
##           Df Sum of Sq    RSS      AIC
## + ndias     1    2.4410 14.080 -55.365
## + dens      1    0.7343 15.786 -49.644
## <none>                16.521 -49.371
## + analf     1    0.2452 16.275 -48.119
## + percap    1    0.0567 16.464 -47.543
## - estud     1    2.6032 19.124 -44.055
## - crime     1   15.8696 32.390 -17.708
##
## Step:  AIC=-55.37
## expvida ~ crime + estud + ndias
##
##           Df Sum of Sq    RSS      AIC
## + dens      1    0.8913 13.188 -56.635
## <none>                14.080 -55.365
## + percap    1    0.1012 13.978 -53.726
```

```

## + analf    1    0.0954 13.984 -53.705
## - ndias    1    2.4410 16.521 -49.371
## - estud    1    3.3047 17.384 -46.823
## - crime    1   18.1773 32.257 -15.915
##
## Step: AIC=-56.63
## expvida ~ crime + estud + ndias + dens
##
##           Df Sum of Sq    RSS    AIC
## <none>                13.188 -56.635
## + percap    1    0.3020 12.886 -55.793
## - dens      1    0.8913 14.080 -55.365
## + analf     1    0.0319 13.156 -54.756
## - estud     1    2.3831 15.571 -50.329
## - ndias     1    2.5980 15.786 -49.644
## - crime     1   18.5026 31.691 -14.800
##
## Start: AIC=-53.82
## expvida ~ percap + analf + crime + estud + ndias + dens
##
##           Df Sum of Sq    RSS    AIC
## - analf     1    0.0077 12.886 -55.793
## - percap     1    0.2778 13.156 -54.756
## <none>                12.879 -53.823
## - estud     1    0.6452 13.524 -53.379
## - dens      1    1.0270 13.906 -51.987
## - ndias     1    2.1928 15.071 -47.961
## - crime     1   14.4812 27.360 -18.148
##
## Step: AIC=-55.79
## expvida ~ percap + crime + estud + ndias + dens
##
##           Df Sum of Sq    RSS    AIC
## - percap     1    0.3020 13.188 -56.635
## <none>                12.886 -55.793
## - estud     1    0.7637 13.650 -54.914
## - dens      1    1.0921 13.978 -53.726
## - ndias     1    2.7034 15.590 -48.271
## - crime     1   18.8044 31.691 -12.800
##
## Step: AIC=-56.63
## expvida ~ crime + estud + ndias + dens
##
##           Df Sum of Sq    RSS    AIC
## <none>                13.188 -56.635
## - dens      1    0.8913 14.080 -55.365
## - estud     1    2.3831 15.571 -50.329
## - ndias     1    2.5980 15.786 -49.644
## - crime     1   18.5026 31.691 -14.800

```

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:

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

em que $\varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$, $x_j = \frac{1}{50} \sum_{i=1}^{50} x_{ji}$ 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.

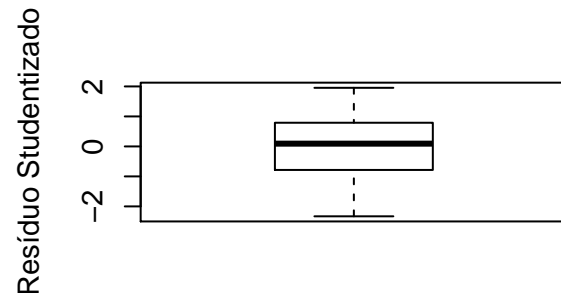
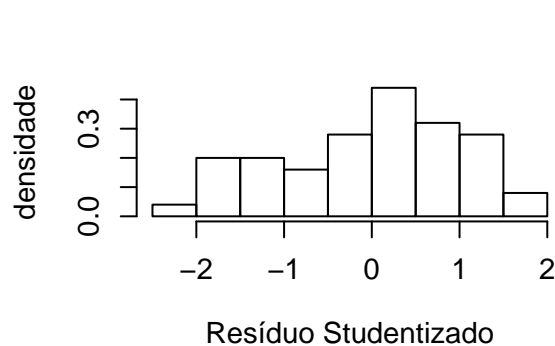
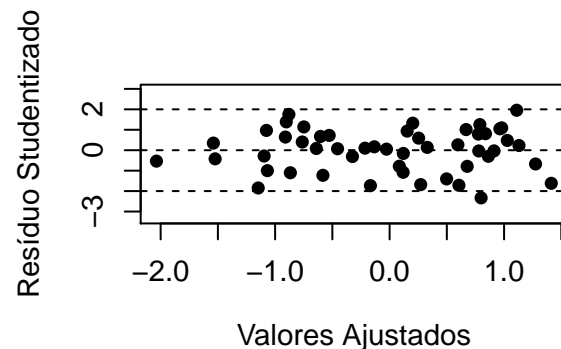
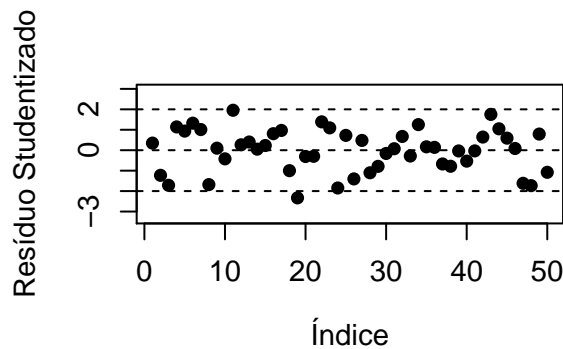
$\frac{\beta_3}{s_3}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da “taxa de criminalidade (por 100000 habitantes 1976)” em uma unidade, mantendo-se as demais covariáveis fixas.

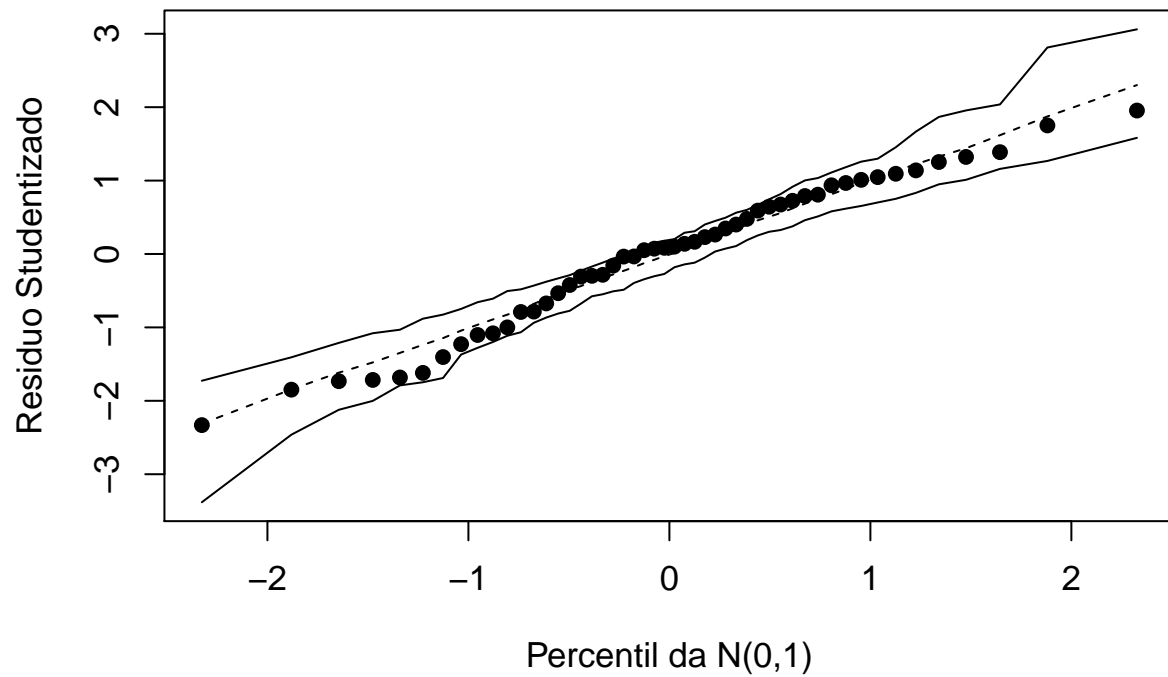
$\frac{\beta_4}{s_4}$: Incremento (positivo ou negativo) na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da “porcentagem de estudantes que concluem o segundo grau (1970)” em uma unidade, mantendo-se as demais covariáveis fixas.

$\frac{\beta_5}{s_5}$: 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 Celsius 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 XX indícios de mal ajuste semelhantes aos observados na análise residual para o modelo anterior. **!!! alguém identificou algum problema a mais??? argumentar aqui!!!** Neste caso, deveria-se procurar modelos alternativos que levem em consideração uma distribuição assimétrica. Novamente, dado o escopo do curso, procede-se com as análises posteriores.





```
#coeficientes modelo otimo
```

```
###Editar essa tabela!!!
```

```
R2fit_otimo <- summary(fit_otimo)$r.squared
R2fit_otimo
```

```
## [1] 0.7308519
```

```
R2aj_fit_otimo <- summary(fit_otimo)$adj.r.squared
R2aj_fit_otimo
```

```
## [1] 0.7069277
```

```
kable(summary(fit_otimo)$coefficients)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0000000	0.0765601	0.000000	1.0000000
crime	-0.7861414	0.0989399	-7.945644	0.0000000
estud	0.2626320	0.0921002	2.851591	0.0065462
ndias	-0.2765304	0.0928773	-2.977375	0.0046685
dens	-0.1401871	0.0803871	-1.743900	0.0880049

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 analisar os coeficientes de correlação 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. Este fato, somado ao fato das covariáveis presentes no modelo não apresentarem nenhum indício de serem fontes de multicolinearidade, descartamos então, a hipótese de presença de multicolinearidade no modelo.

```
#Análise de multicolineariedade descritivamente
```

```
#pegando só as variáveis do modelo otimo
```

```
dados_modelo_otimo <- dados[,3:6]
```

```
cor(dados_modelo_otimo)
```

	crime	estud	ndias	dens
crime	1.0000000	-0.4879710	-0.5388834	0.1110318
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

Alavancagem

Observando a figura XX podemos identificar a existência de pontos distantes dos demais, o que é uma indicação de **!!!escrever explicações, identificar os pontos e depois ajustar modelos sem estes**

pontos e compara-los via AIC e BIC!!!

