Análise de Regressão Múltipla no R

Monitor PAE: Bruno Santos
Profª Silvia Nagib Elian

Instituto de Matemática e Estatística
Universidade de São Paulo

Agenda

1. **Introdução ao R**
2. **Leitura dos dados**
3. **Utilização do R para cálculos simples**
   Obtenção do EMQ no caso da regressão simples
4. **Análise descritiva dos dados**
5. **Ajuste do modelo**
6. **Análise de diagnóstico**
7. **Análise de resíduos**
8. **Links úteis**

RStudio

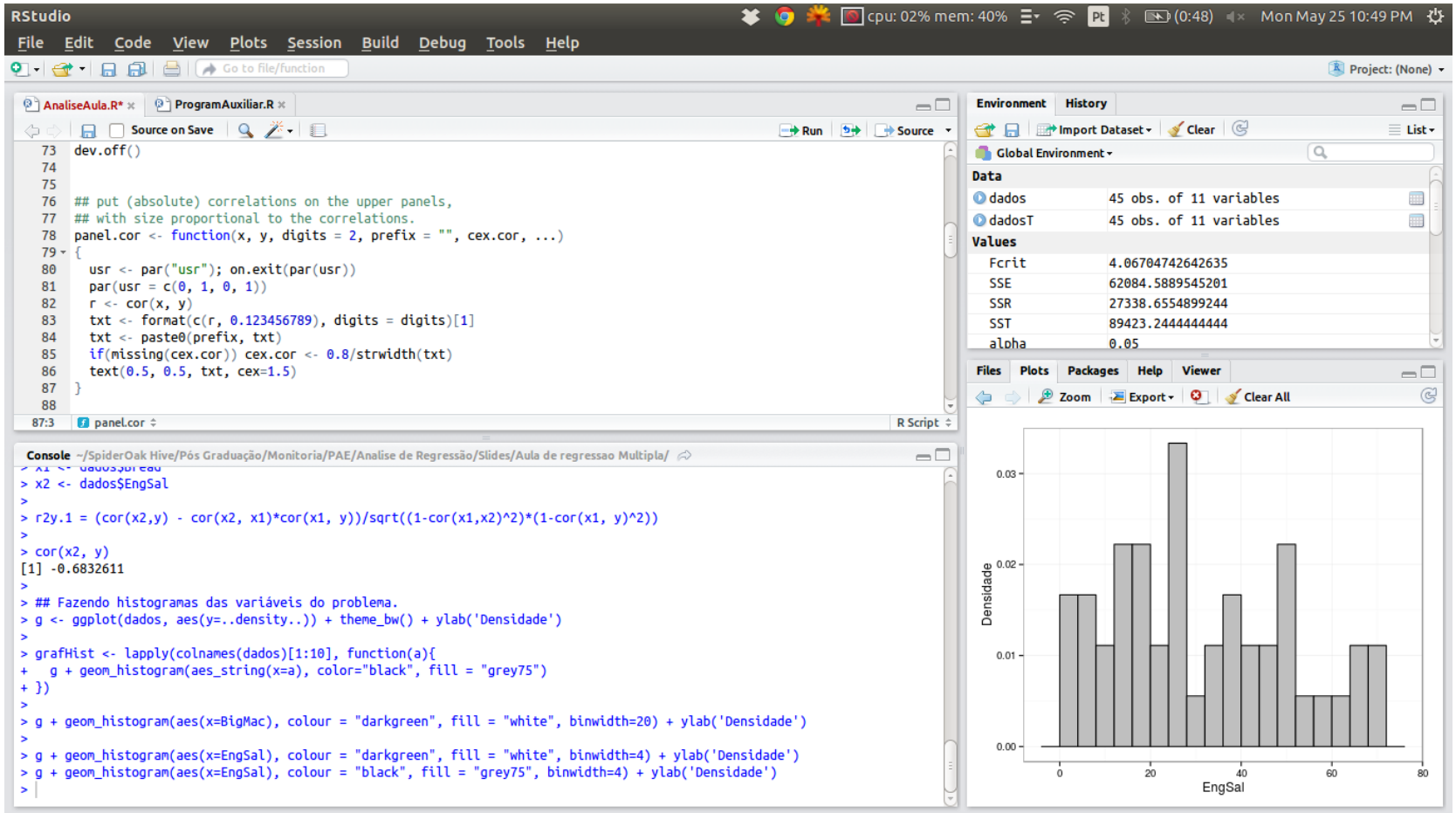
Uma sugestão para trabalhar com o R é o RStudio.

Existem versões disponíveis para Windows, Linux e Mac.

O *download* desse programa pode ser feito em `http://www.rstudio.com/`.

O RStudio está customizado para permitir a utilização de alguns pacotes como Rcpp, shiny, devtools, entre outros.

RStudio


RStudio
File Edit Code View Plots Session Build Debug Tools Help
AnaliseAula.R* ProgramAuxiliar.R
Source on Save
Run Source
73 dev.off()
74
75
76 ## put (absolute) correlations on the upper panels,
77 ## with size proportional to the correlations.
78 panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
79 {
80 usr <- par("usr"); on.exit(par(usr))
81 par(usr = c(0, 1, 0, 1))
82 r <- cor(x, y)
83 txt <- format(c(r, 0.123456789), digits = digits)[1]
84 txt <- paste0(prefix, txt)
85 if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
86 text(0.5, 0.5, txt, cex=1.5)
87 }
88
87:3 panel.cor R Script
Console ~/SpiderOak Hive/Pós Graduação/Monitoria/PAE/Analise de Regressão/Slides/Aula de regressao Multipla/
> x2 <- dados$EngSal
>
> r2y.1 = (cor(x2,y) - cor(x2, x1)*cor(x1, y))/sqrt((1-cor(x1,x2)^2)*(1-cor(x1, y)^2))
>
> cor(x2, y)
[1] -0.6832611
>
> ## Fazendo histogramas das variáveis do problema.
> g <- ggplot(dados, aes(y=..density..)) + theme_bw() + ylab('Densidade')
>
> grafHist <- lapply(colnames(dados)[1:10], function(a){
+ g + geom_histogram(aes_string(x=a), color="black", fill = "grey75")
+ })
>
> g + geom_histogram(aes(x=BigMac), colour = "darkgreen", fill = "white", binwidth=20) + ylab('Densidade')
>
> g + geom_histogram(aes(x=EngSal), colour = "darkgreen", fill = "white", binwidth=4) + ylab('Densidade')
> g + geom_histogram(aes(x=EngSal), colour = "black", fill = "grey75", binwidth=4) + ylab('Densidade')
>
Environment History
Import Dataset Clear List
Global Environment
Data
dados 45 obs. of 11 variables
dadosT 45 obs. of 11 variables
Values
Fcrit 4.06704742642635
SSE 62084.5889545201
SSR 27338.6554899244
SST 89423.2444444444
alpha 0.05
Files Plots Packages Help Viewer
Zoom Export Clear All
Densidade
0.00 0.01 0.02 0.03
0 20 40 60 80
EngSal

Leitura dos dados

Funções que podem ser utilizadas para definir a área de trabalho: `getwd()` e `setwd()`

Dados em formato Excel

```
library(xlsx)
dados <- read.xlsx('bigmac.xls', 'bigmac')
```

Dados em formato csv, sem fazer *download*

```
dados <- read.csv('
http://www.ime.usp.br/~bramos/bigmac.csv')
```

Dados disponíveis em pacote no R, sem informação da cidade.

```
library(ldr)
dados <- data(bigmac)
```

Leitura dos dados

Após a leitura dos dados, é importante se os dados foram carregados corretamente.

- `head(dados)`
- `tail(dados)`

Algumas estatísticas descritivas também podem ser úteis para verificar se os dados estão corretos.

- `summary(dados)`

Para verificar o nome das variáveis.

- `names(dados)`
- `colnames(dados)`

Descrição do banco de dados

- **BigMac** - Número de minutos trabalhados para comprar um BigMac e batatas fritas.
- **Bread** - Número de minutos trabalhados para comprar 1 kg de pão.
- **BusFare** - Menor custo de 10k de transporte público, em dólares.
- **EngSal** - Salário médio anual de um engenheiro elétrico, em milhares de dólares.
- **EngTax** - Taxa de imposto média paga por engenheiros.
- **Service** - Custo anual de 19 serviços, primariamente relevante para Europa e América do Norte.
- **TeachSal** - Salário médio anual de um professor do ensino fundamental, em milhares de dólares.
- **TeachTax** - Taxa de imposto média paga por professores de ensino fundamental.
- **VacDays** - Número médio de dias de férias por ano.
- **WorkHrs** - Número médio de horas trabalhadas por ano.

Exemplo com EMQ no caso da regressão simples

Comandos básicos para regressão simples

```
y <- dados$BigMac
x <- dados$Bread
n <- length(y)
alpha <- 0.05
beta1 <- cov(x,y)/cov(x,x)
beta0 <- mean(y) - beta1*mean(x)
y.pred <- beta0 + beta1*x
SST <- sum((y-mean(y))^2)
SSR <- sum((y.pred-mean(y))^2)
SSE <- sum((y-y.pred)^2)
estF <- SSR/(SSE/(n-2))
Fcrit <- qf(1-alpha, 1, n-2)
pvalorF <- 1 - pf(estF, 1, n-2)
```

Análise descritiva

Análise univariada das variáveis:

- boxplot.
- histograma.

Análise da relação entre a variável resposta e as outras variáveis explicativas:

- diagrama de dispersão.
- correlação.

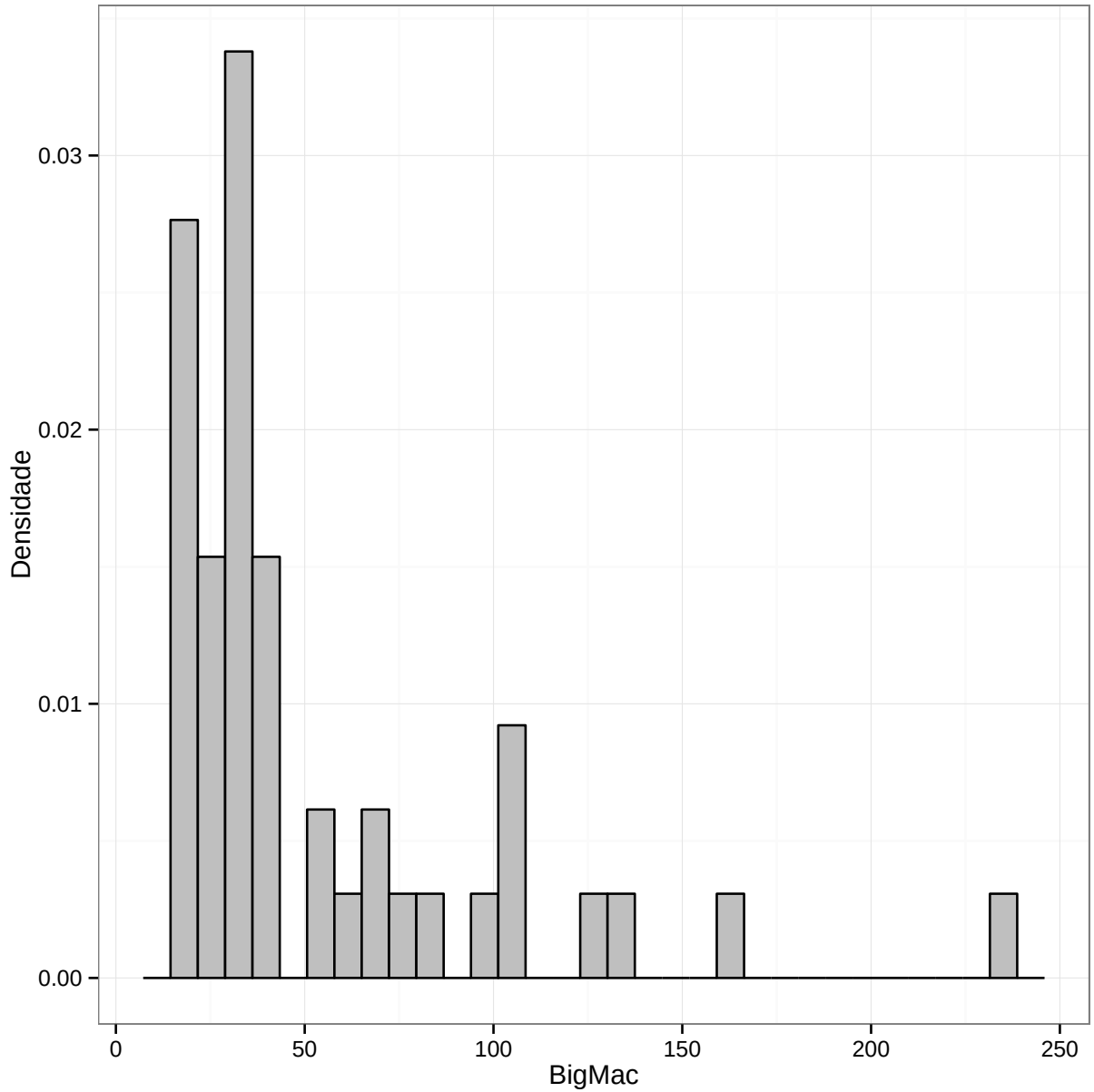
Pacote para fazer gráficos

```
library(ggplot2)
```

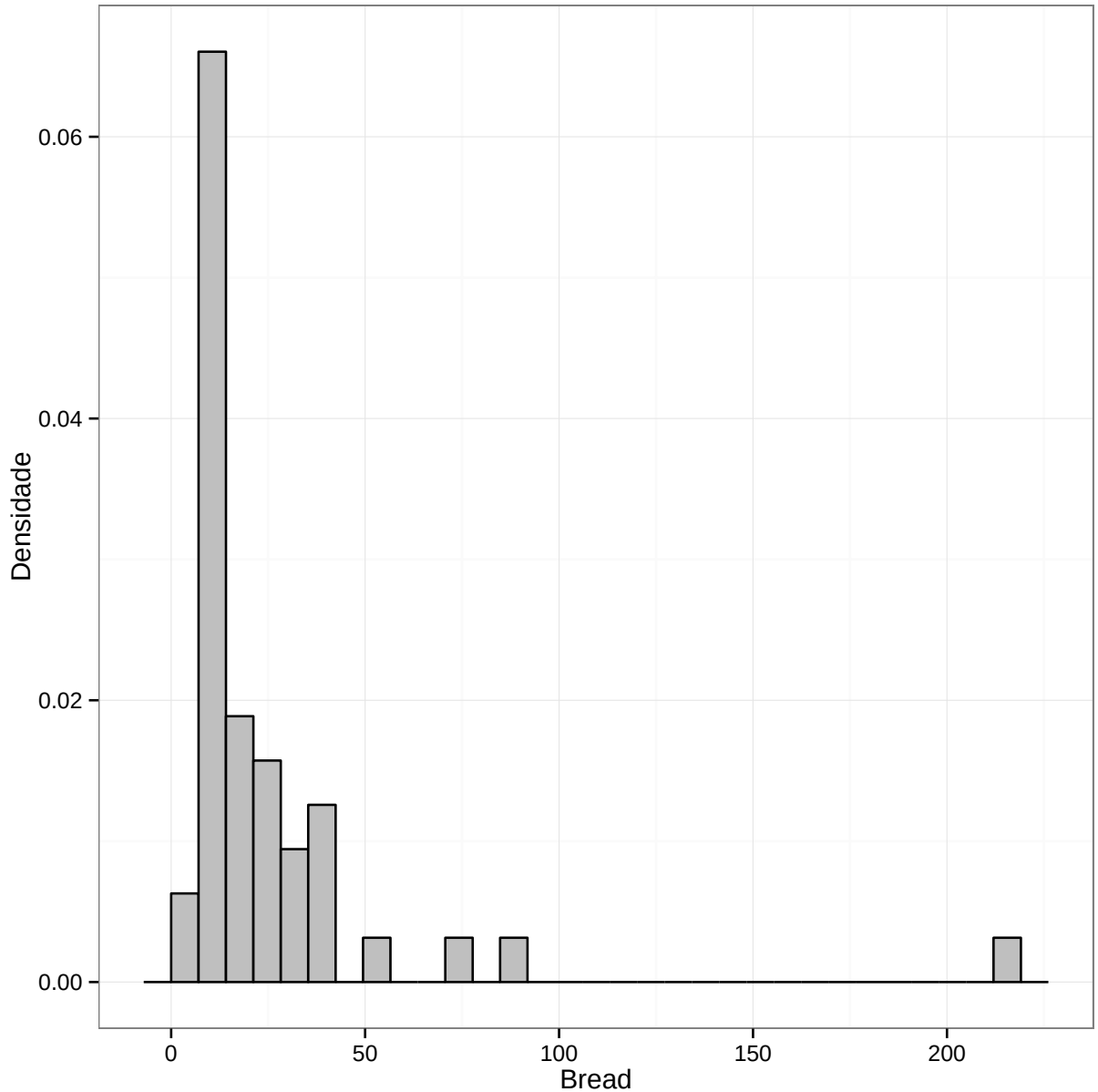
Fazendo histogramas

```
g <- ggplot(dados, aes(y=..density..))  + theme_bw()
+ ylab('Densidade')

lapply(colnames(dados)[1:10], function(a){
  g + geom_histogram(aes_string(x=a), colour =
"darkgreen", fill = "white")
})

g + geom_histogram(aes(x=BigMac), colour =
"darkgreen", fill = "white", binwidth=20) +
ylab('Densidade')

g + geom_histogram(aes(x=EngSal), colour =
"darkgreen", fill = "white", binwidth=4) +
ylab('Densidade')
```

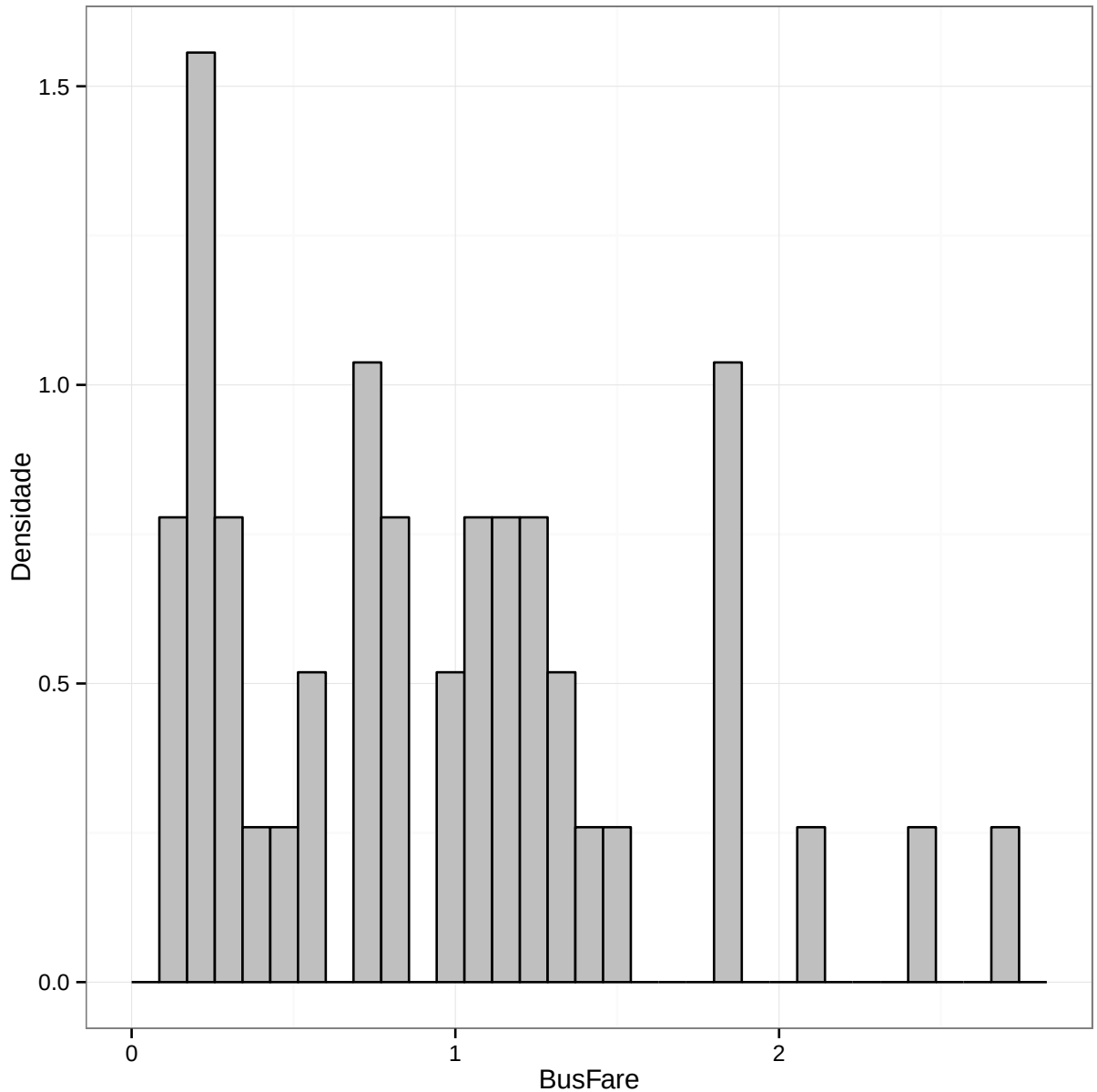
Histograma



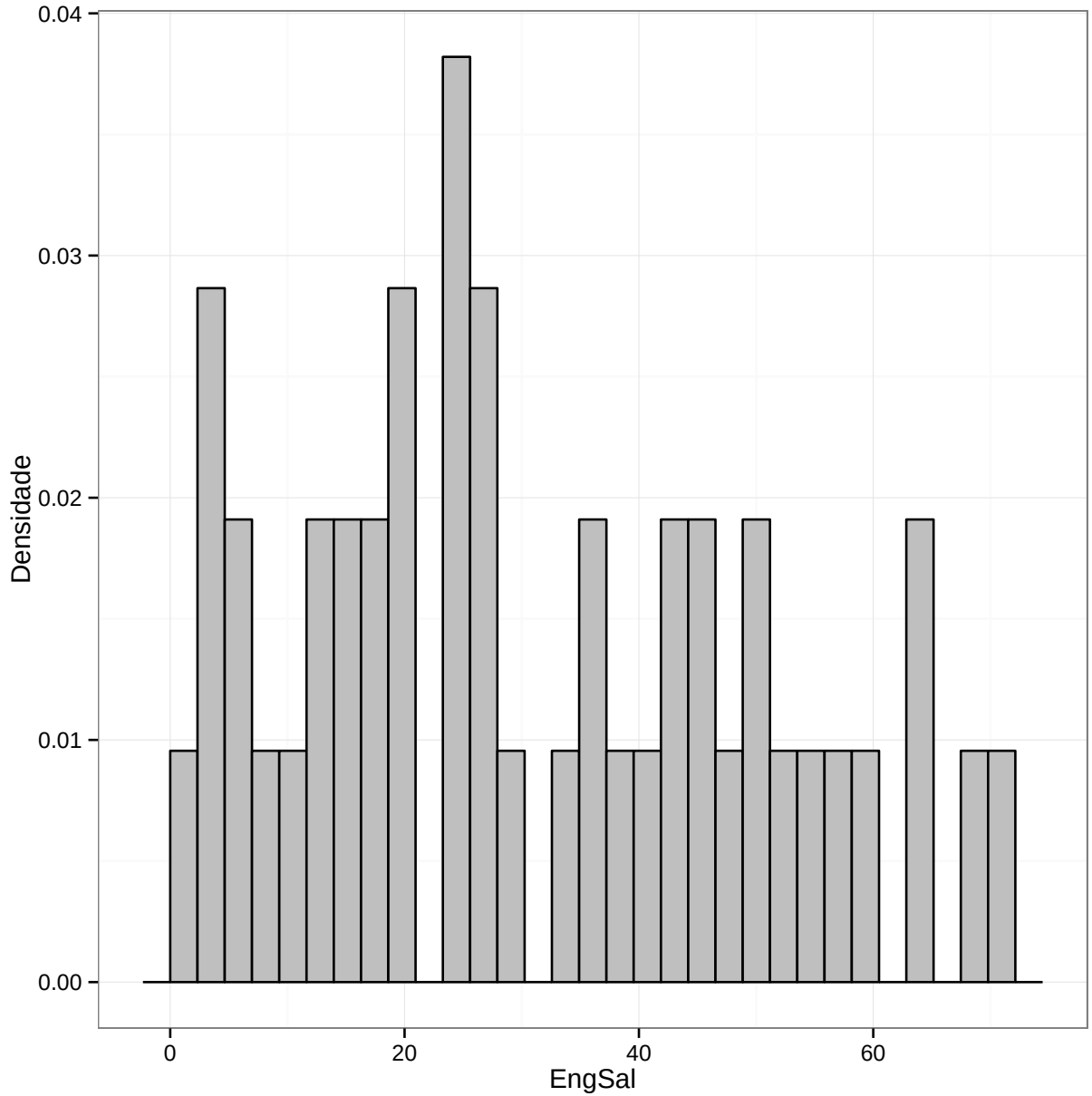
Histograma



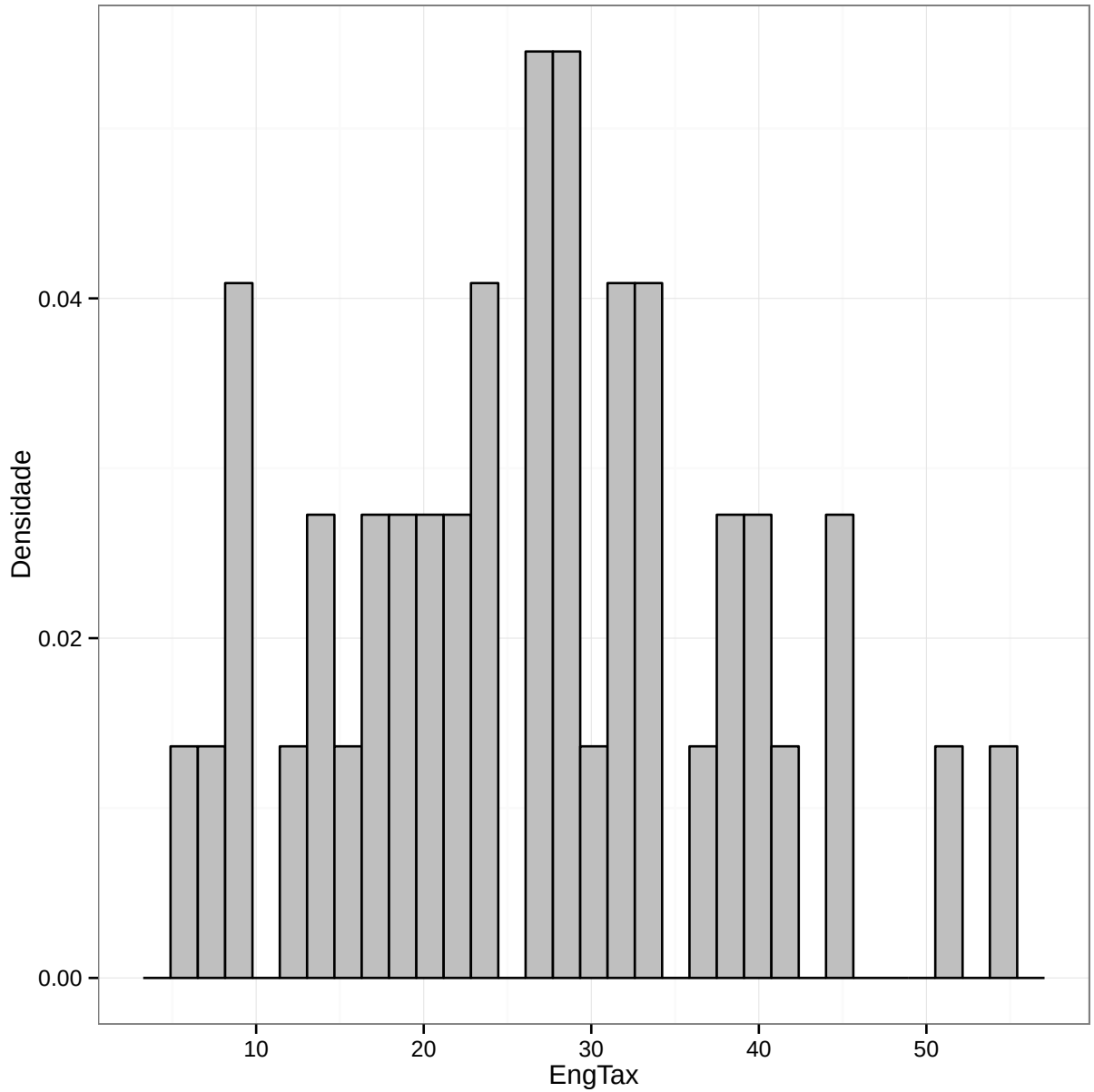
Histograma



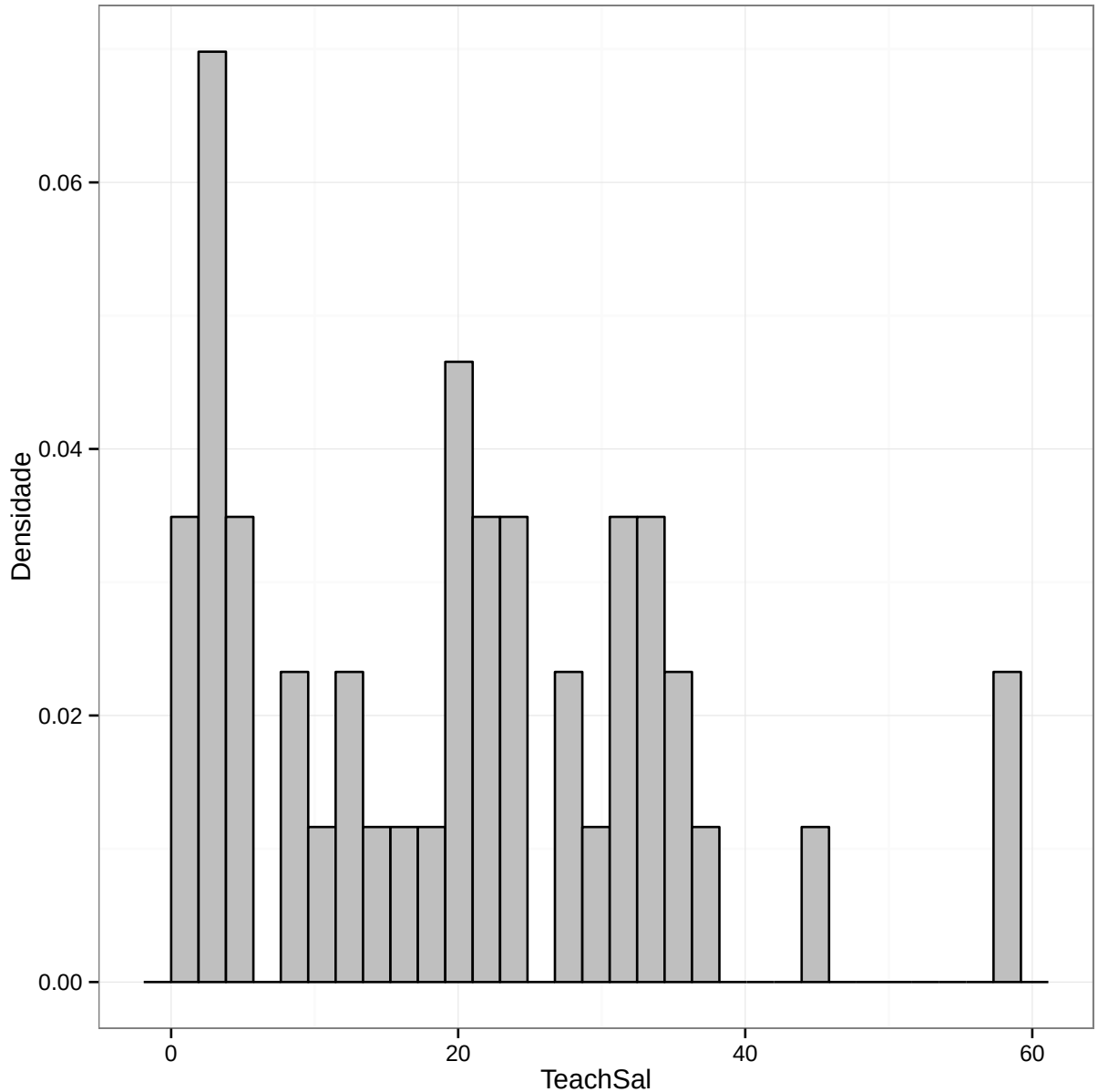
Histograma



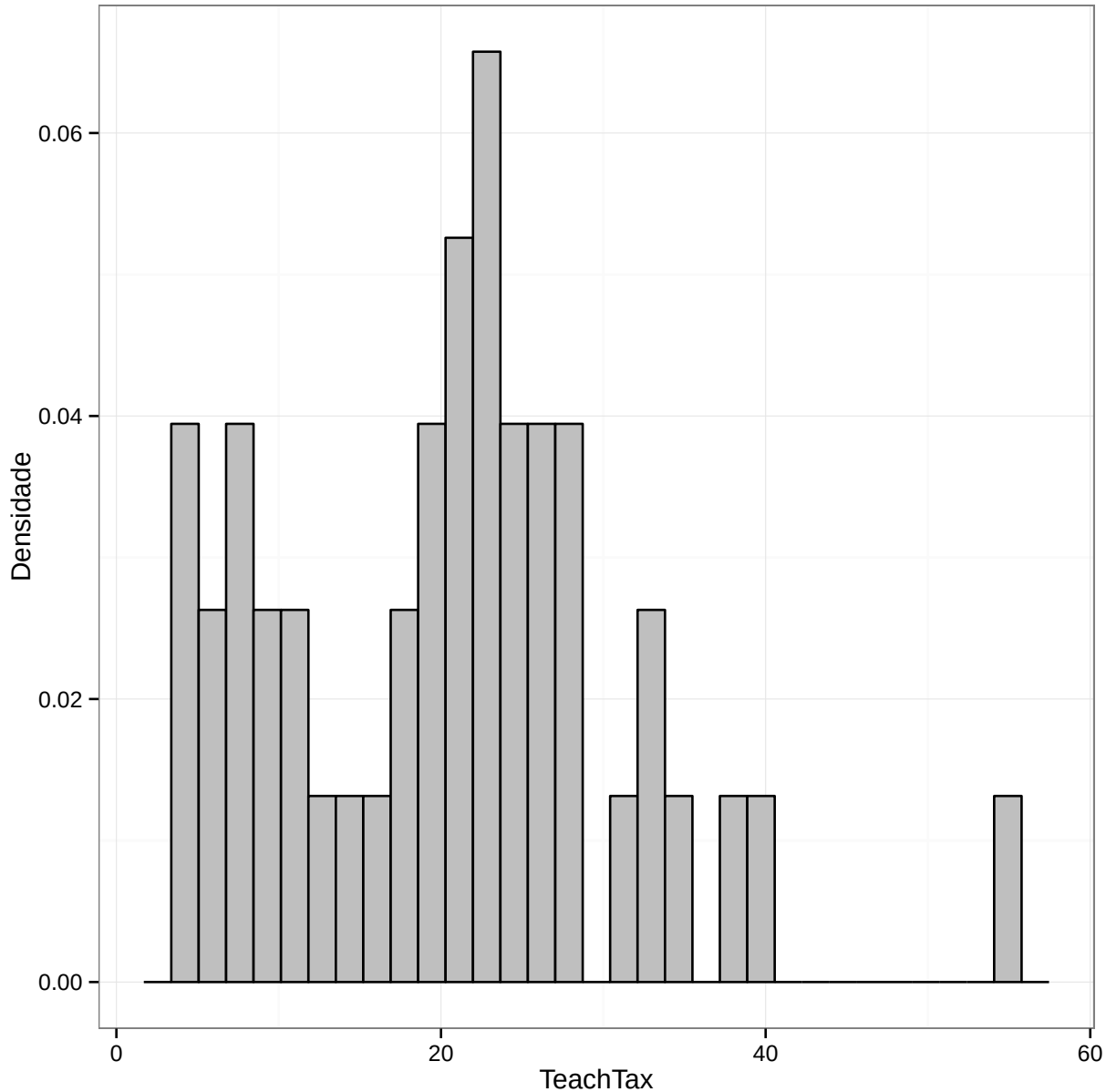
Histograma



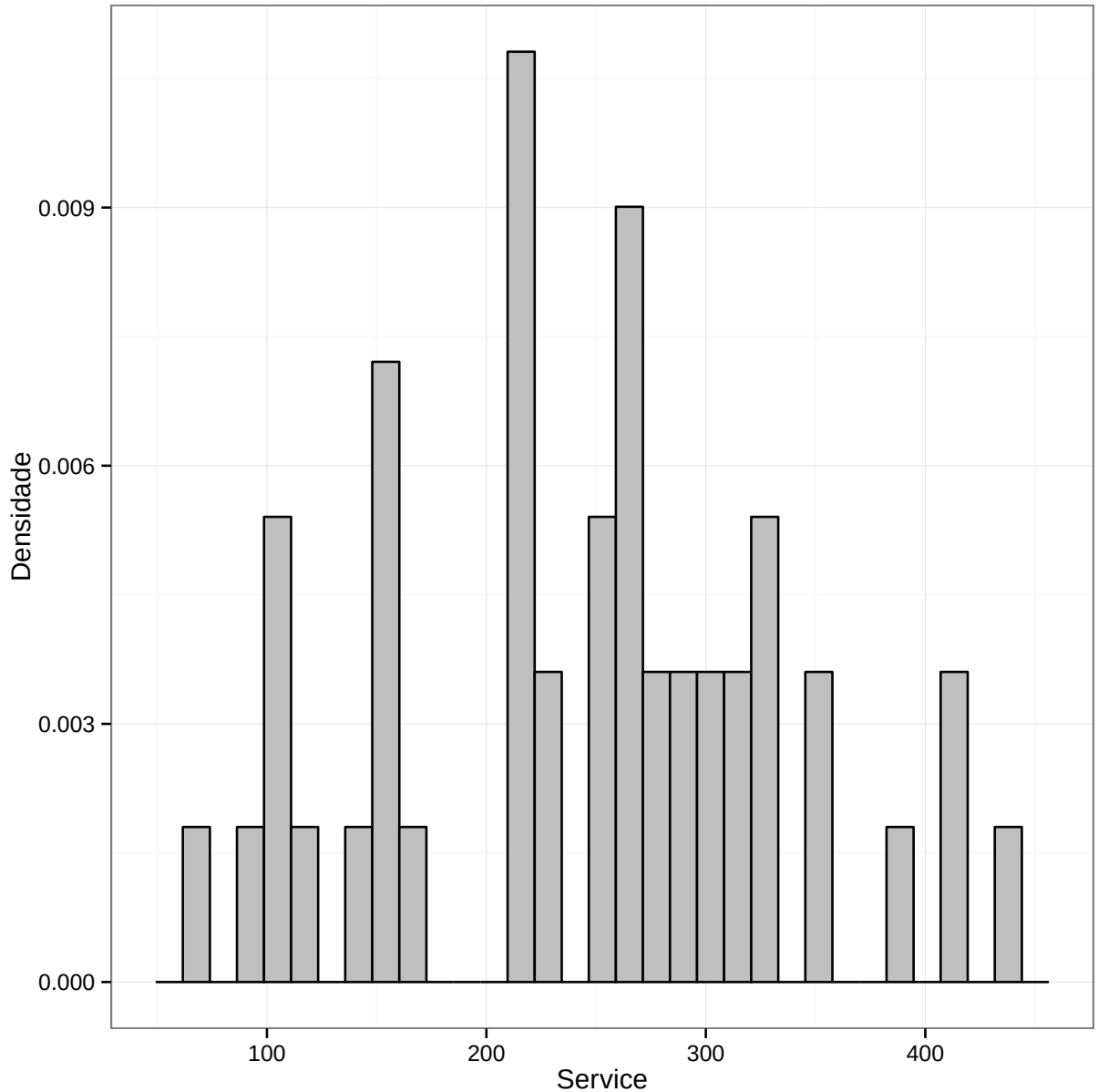
Histograma



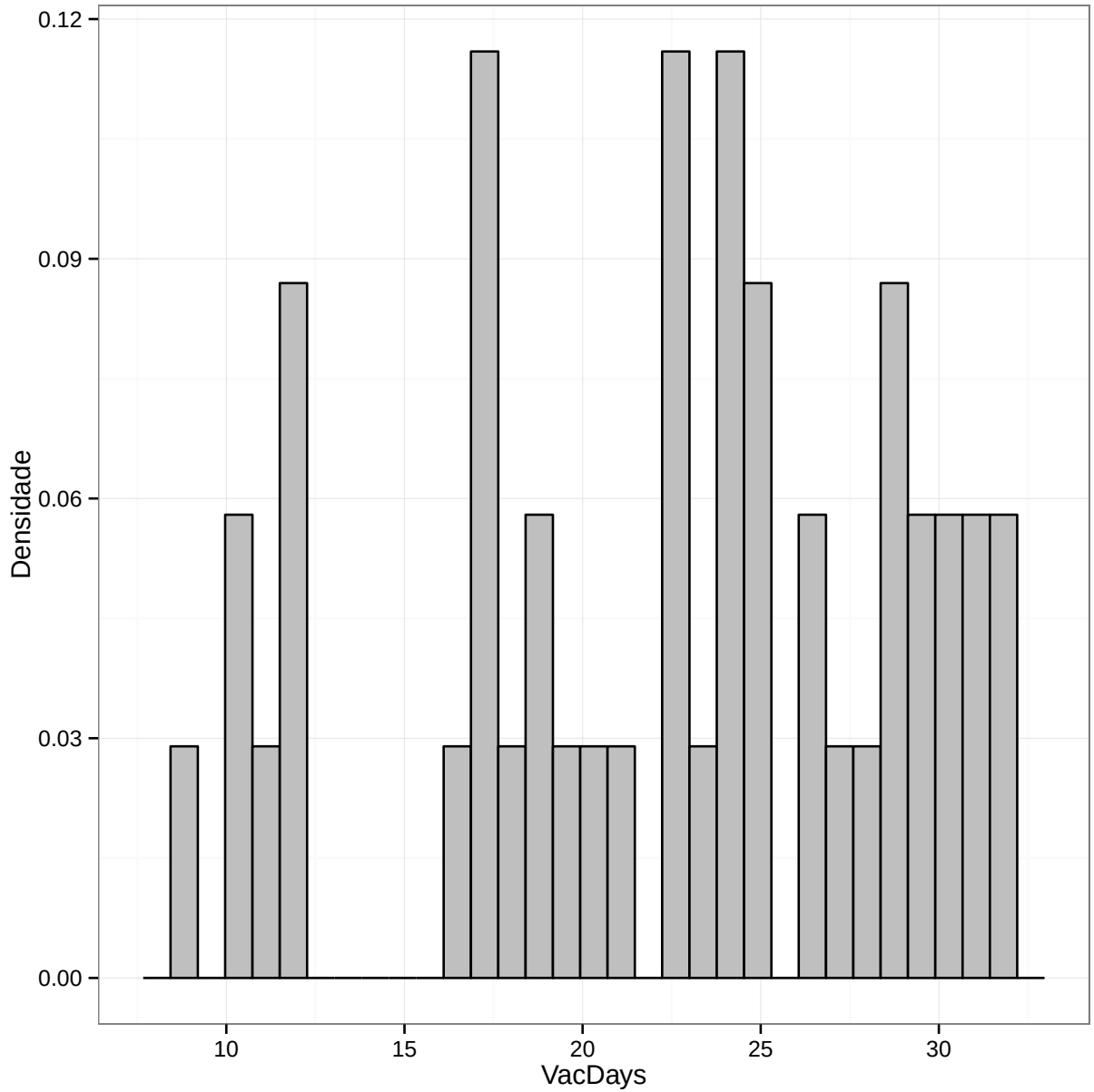
Histograma



Histograma



Histograma



Histograma

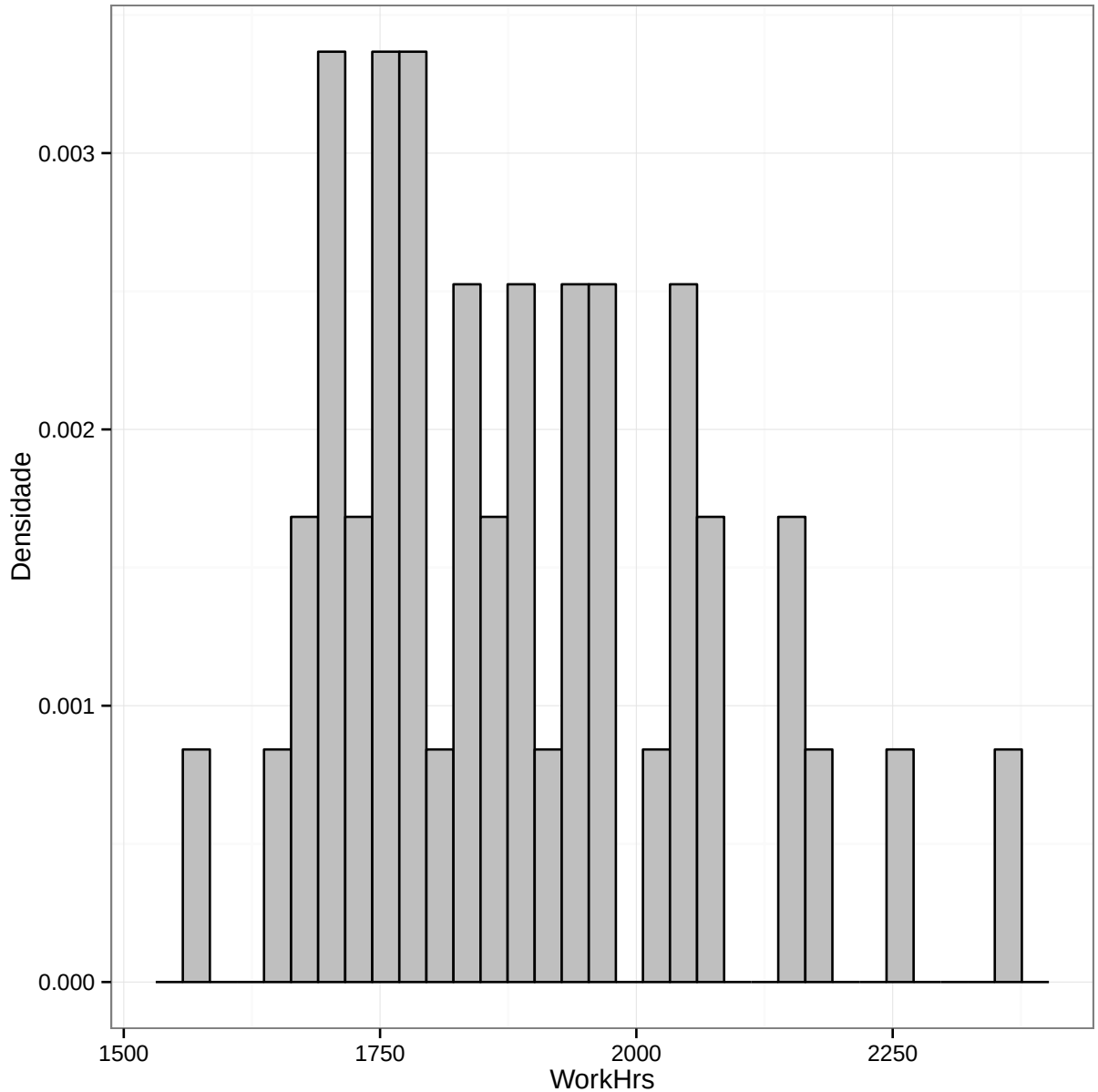


Gráfico de dispersão

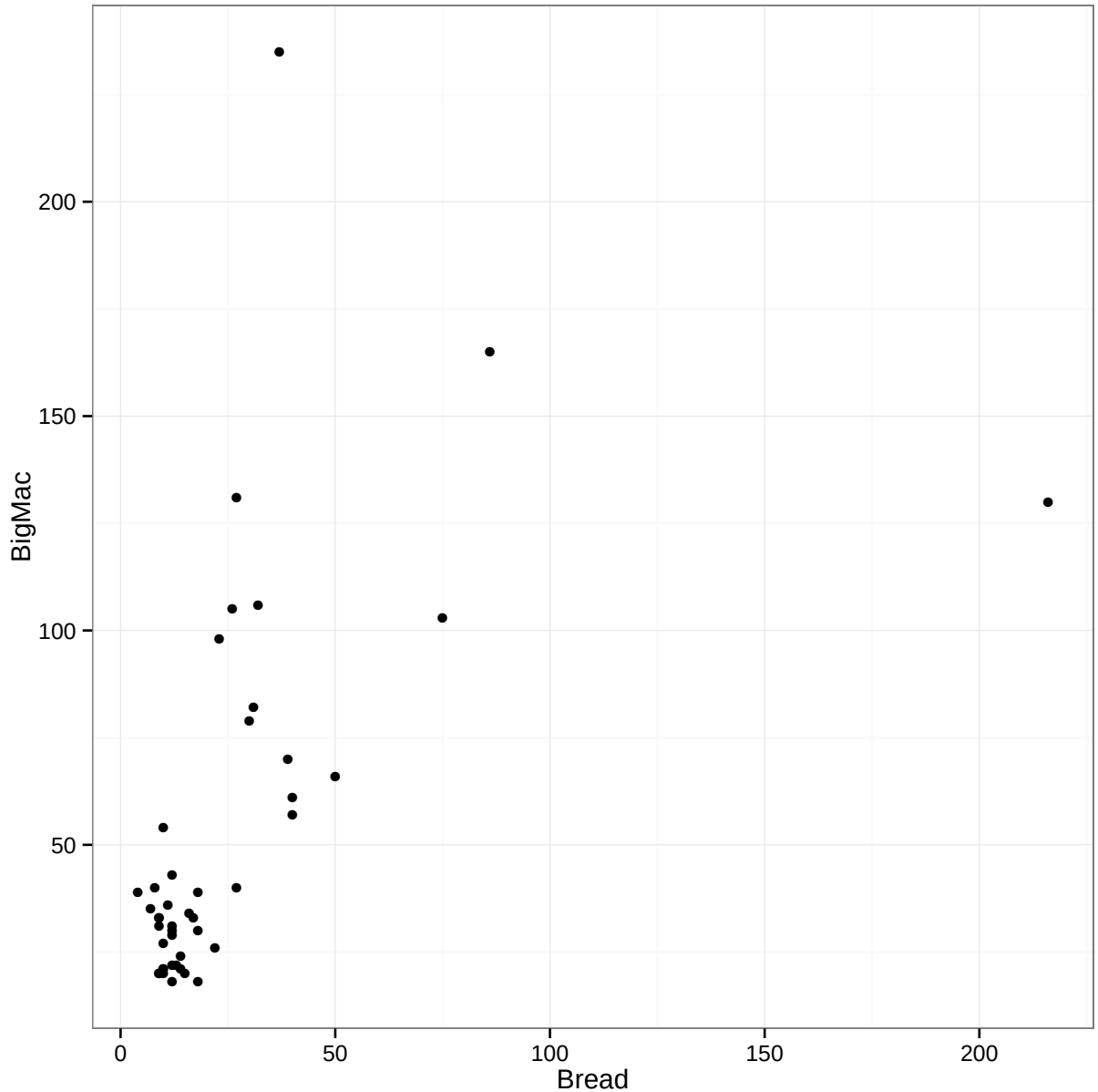


Gráfico de dispersão

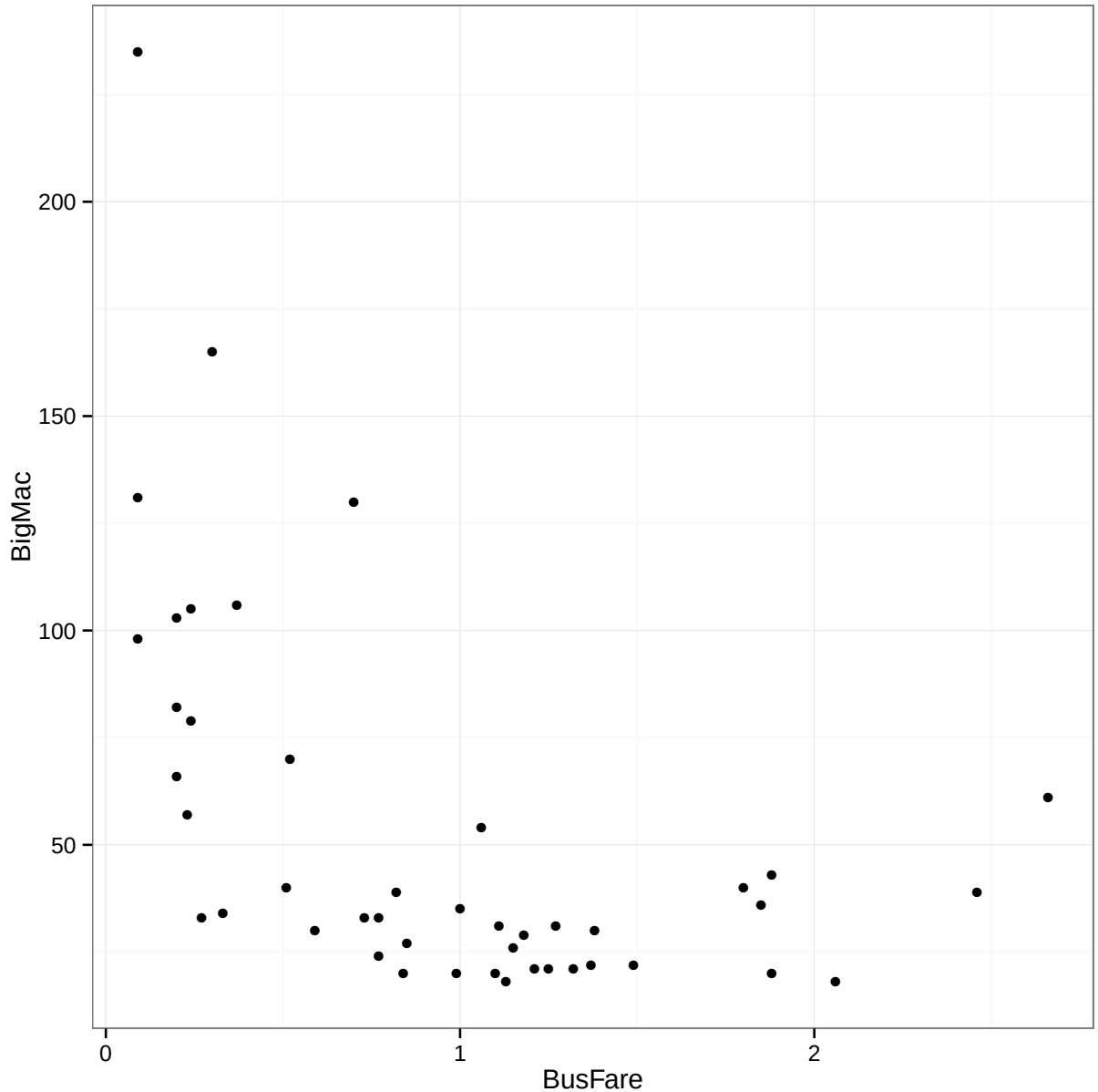


Gráfico de dispersão

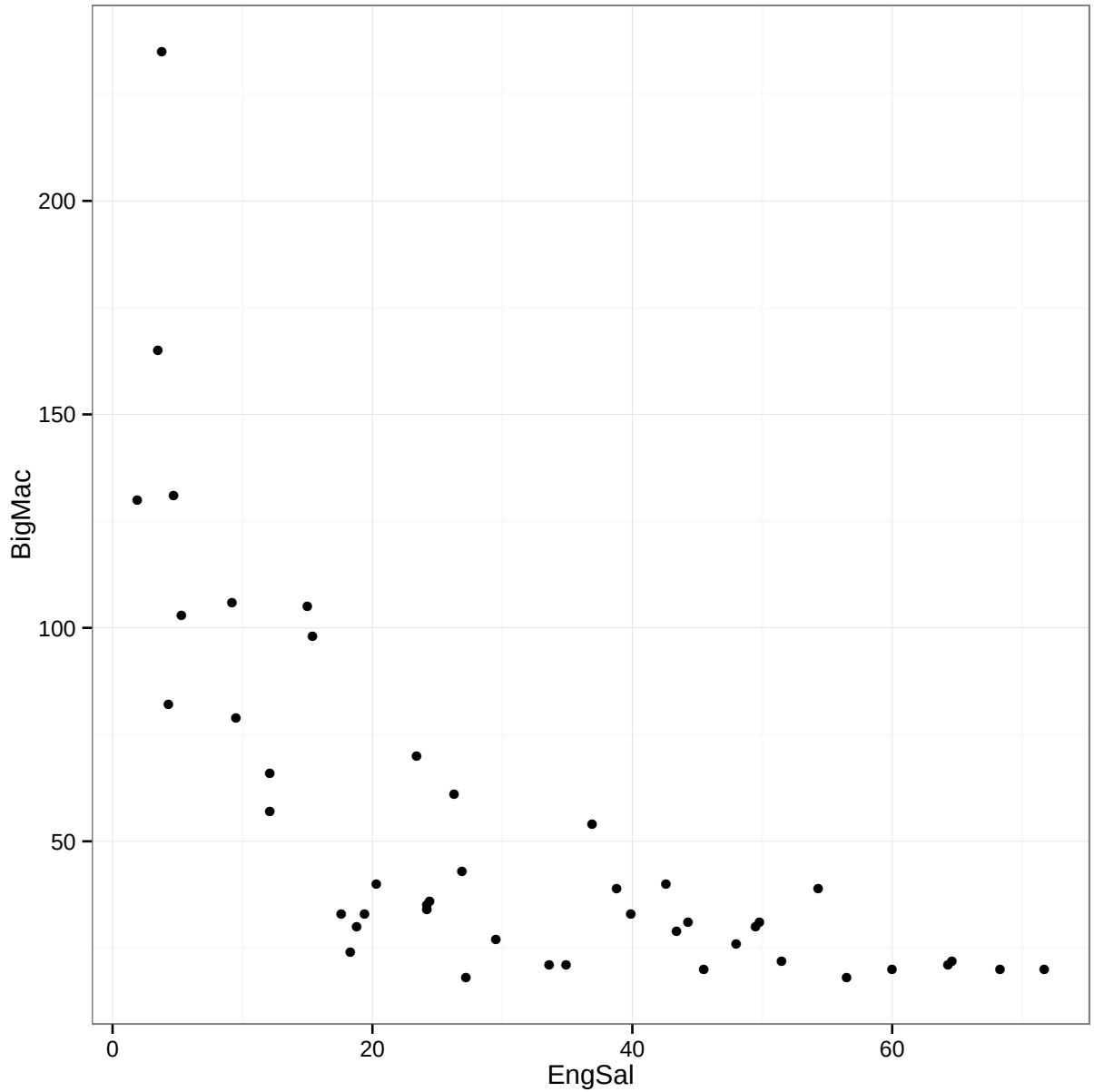


Gráfico de dispersão

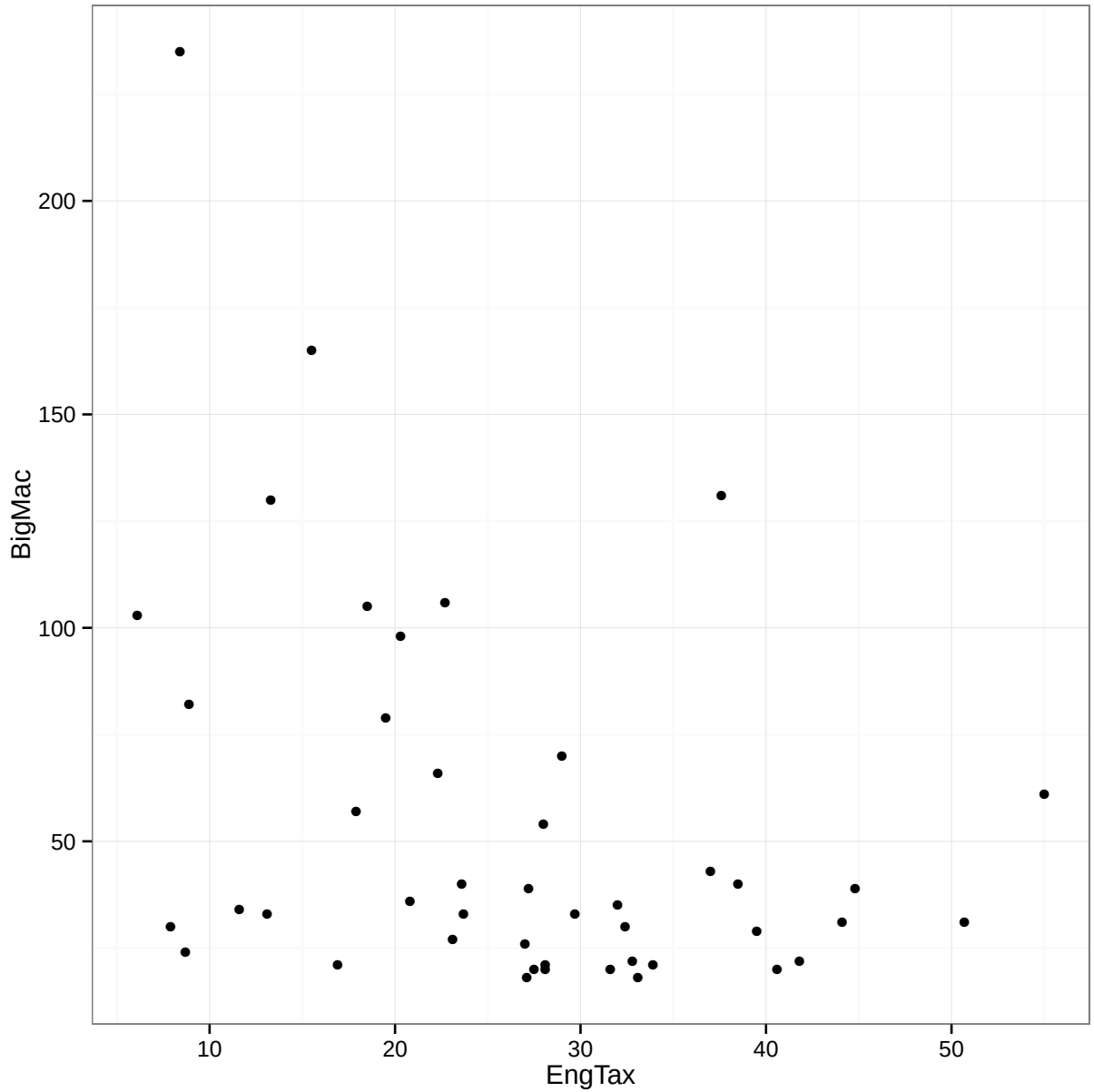


Gráfico de dispersão

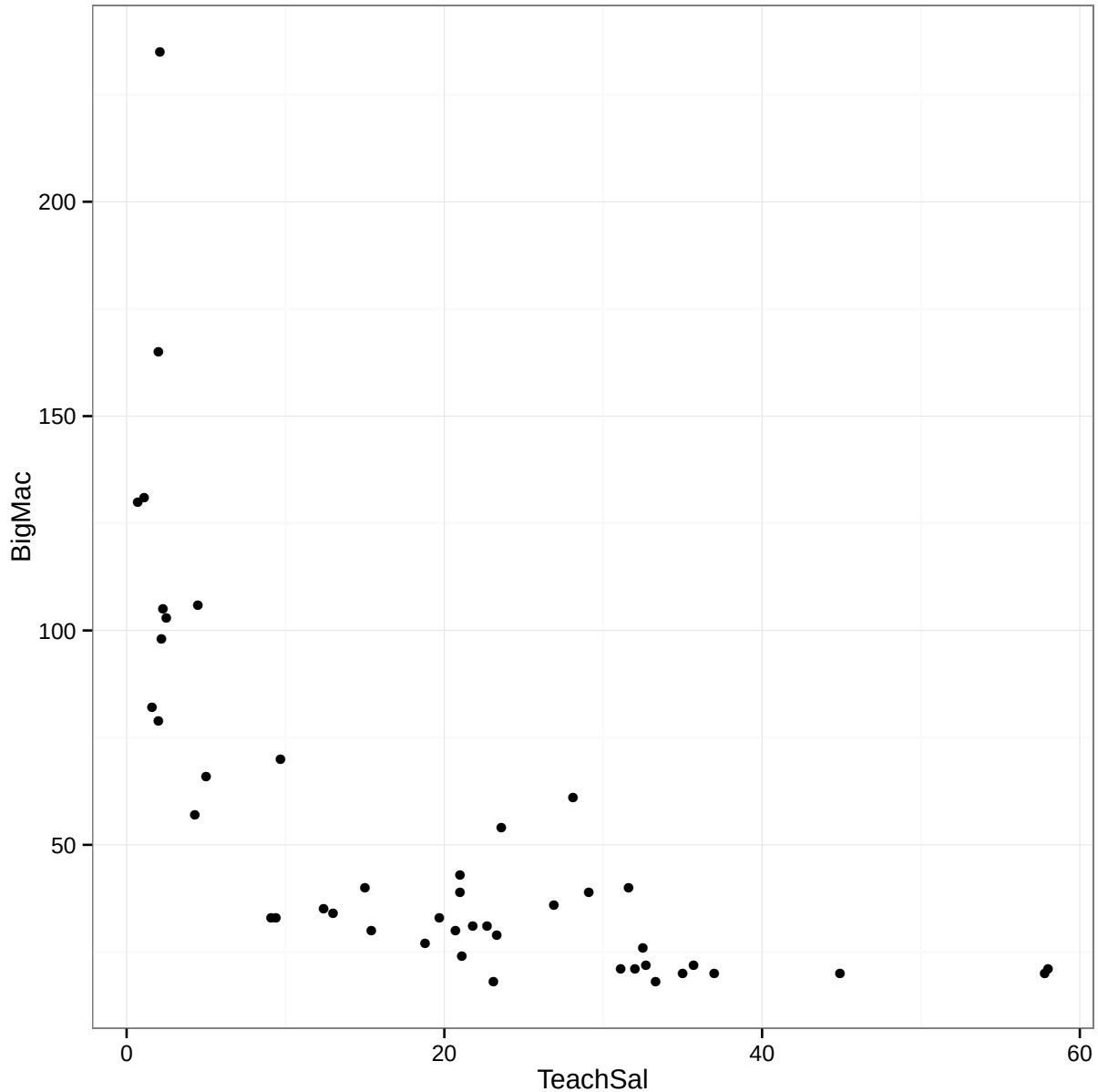


Gráfico de dispersão

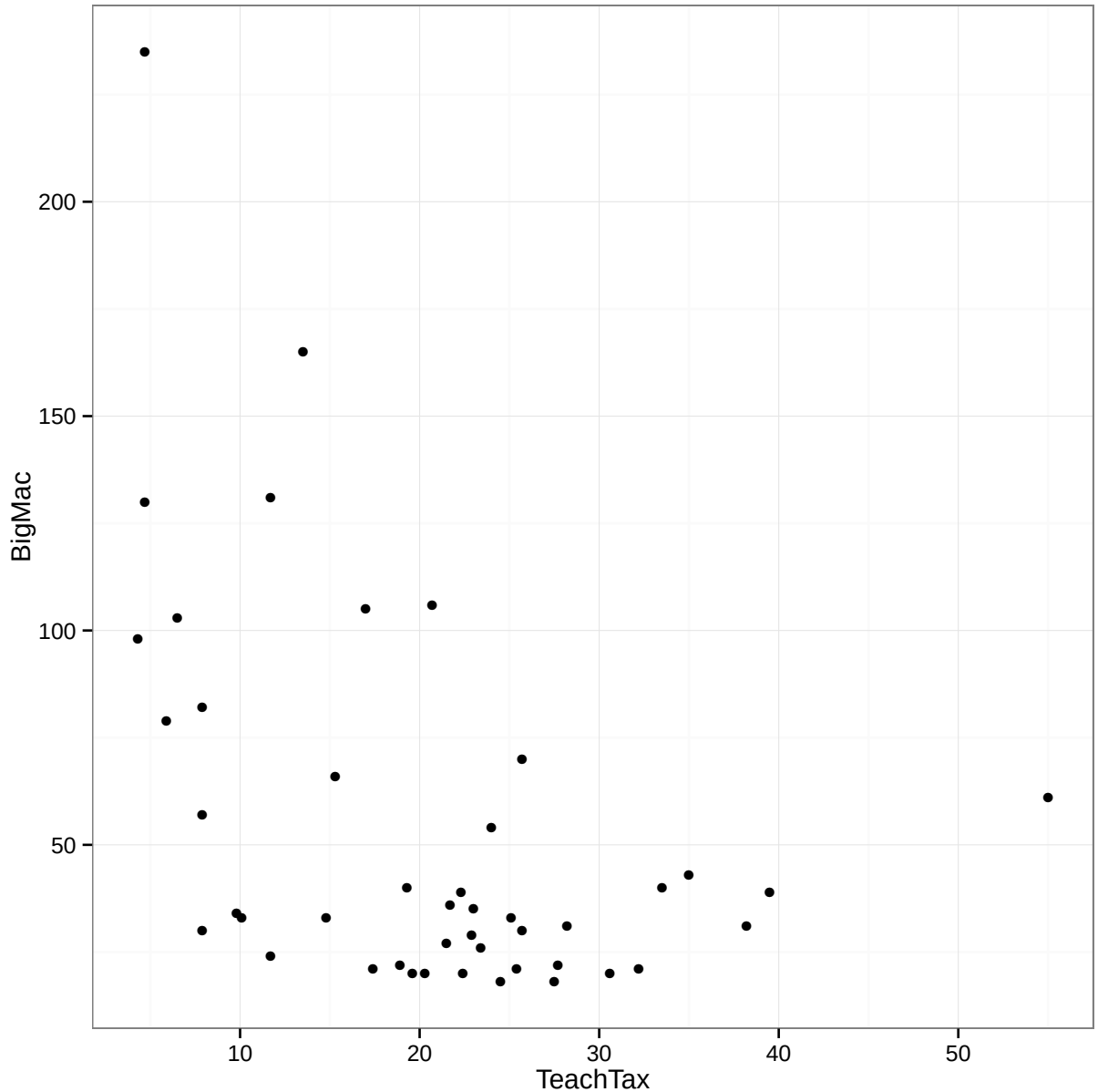


Gráfico de dispersão

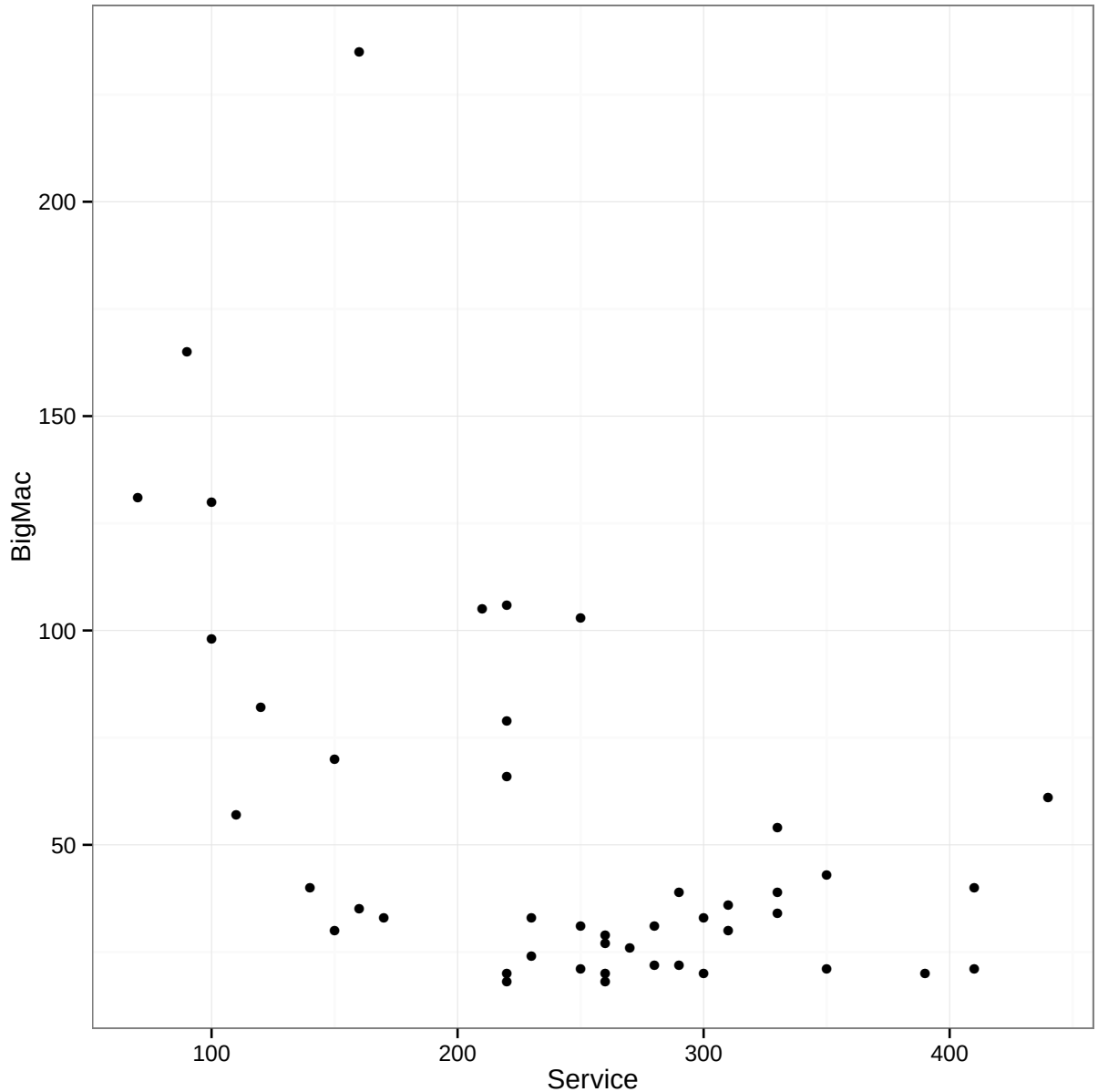


Gráfico de dispersão

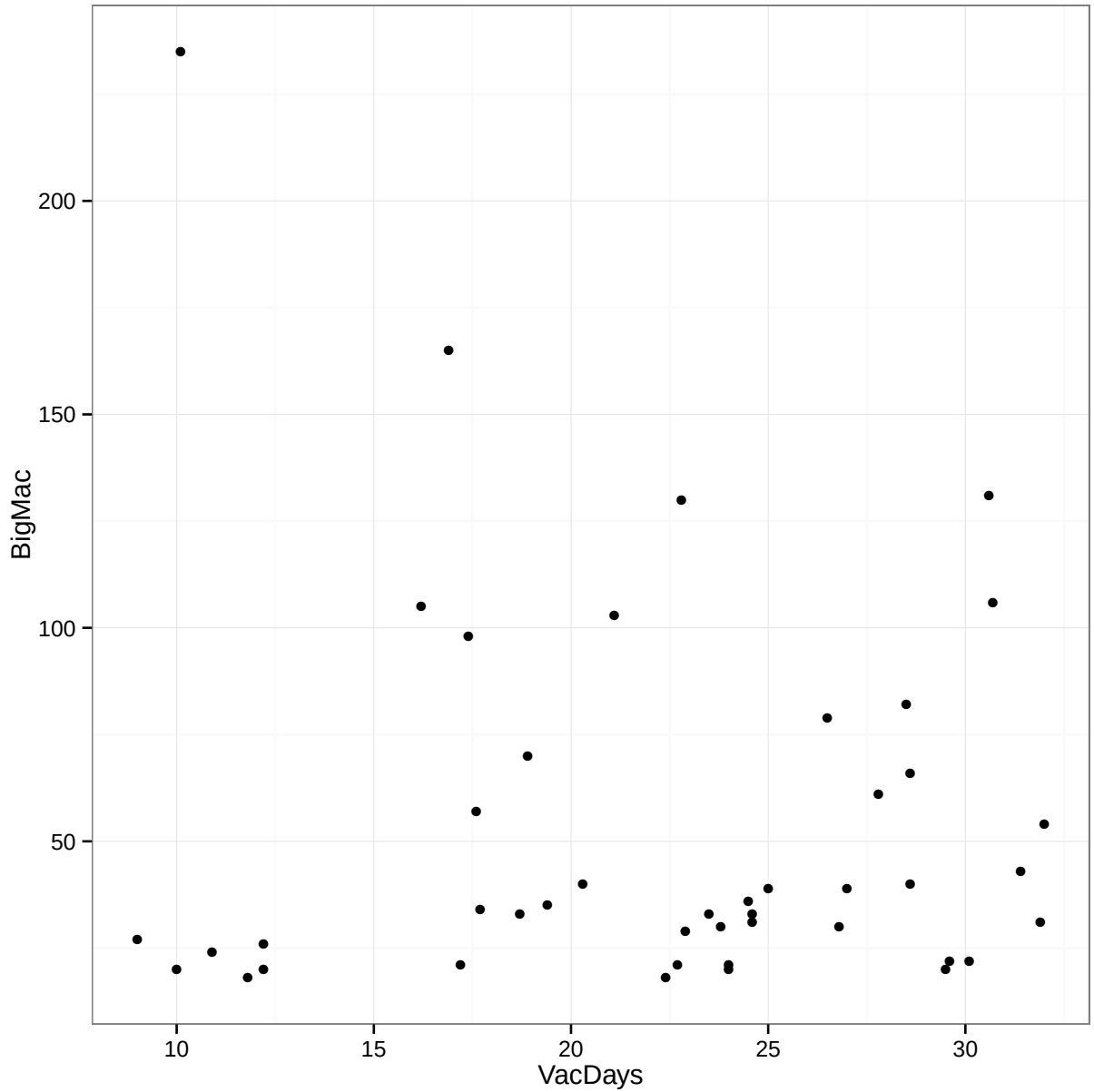
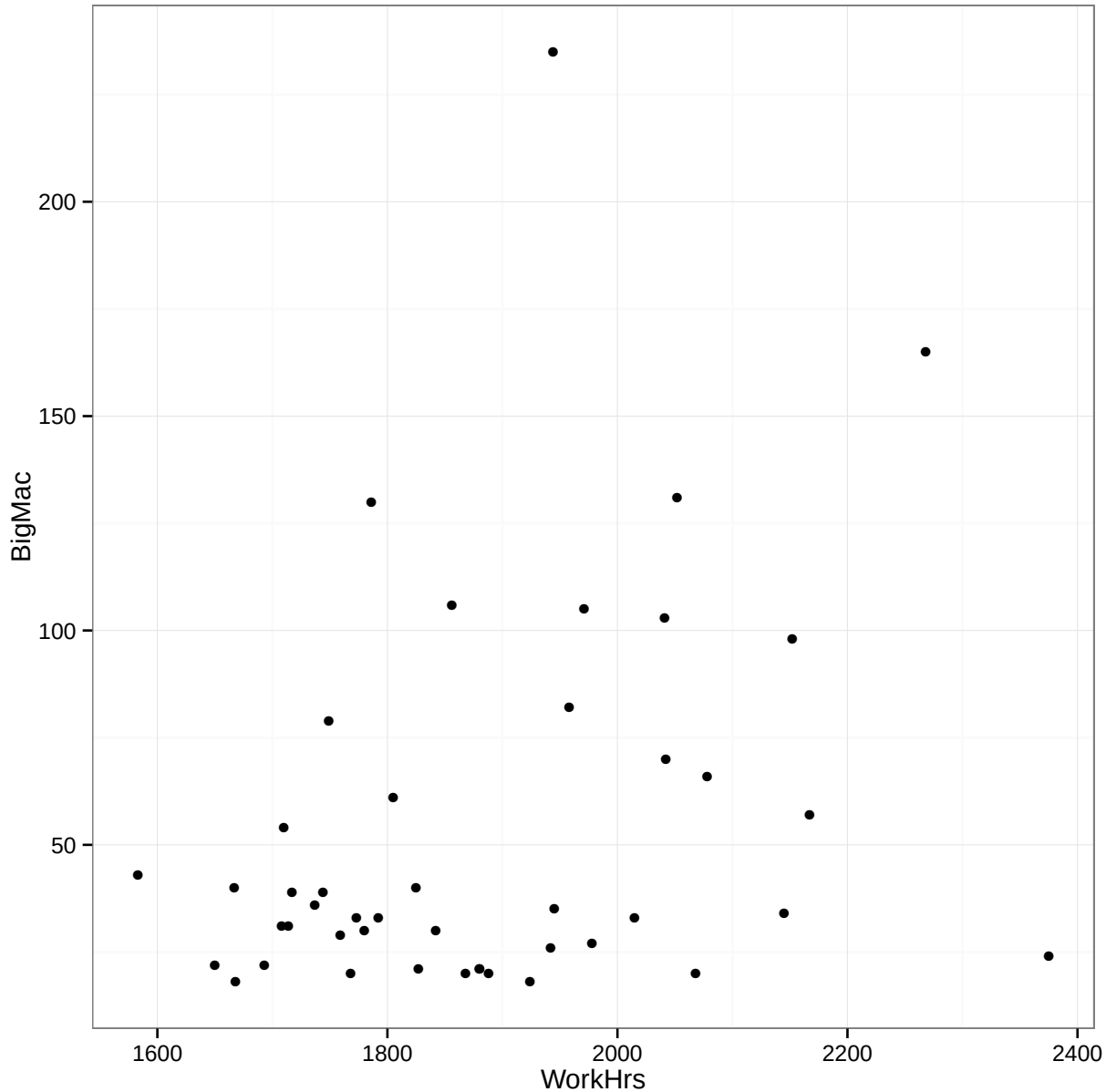


Gráfico de dispersão

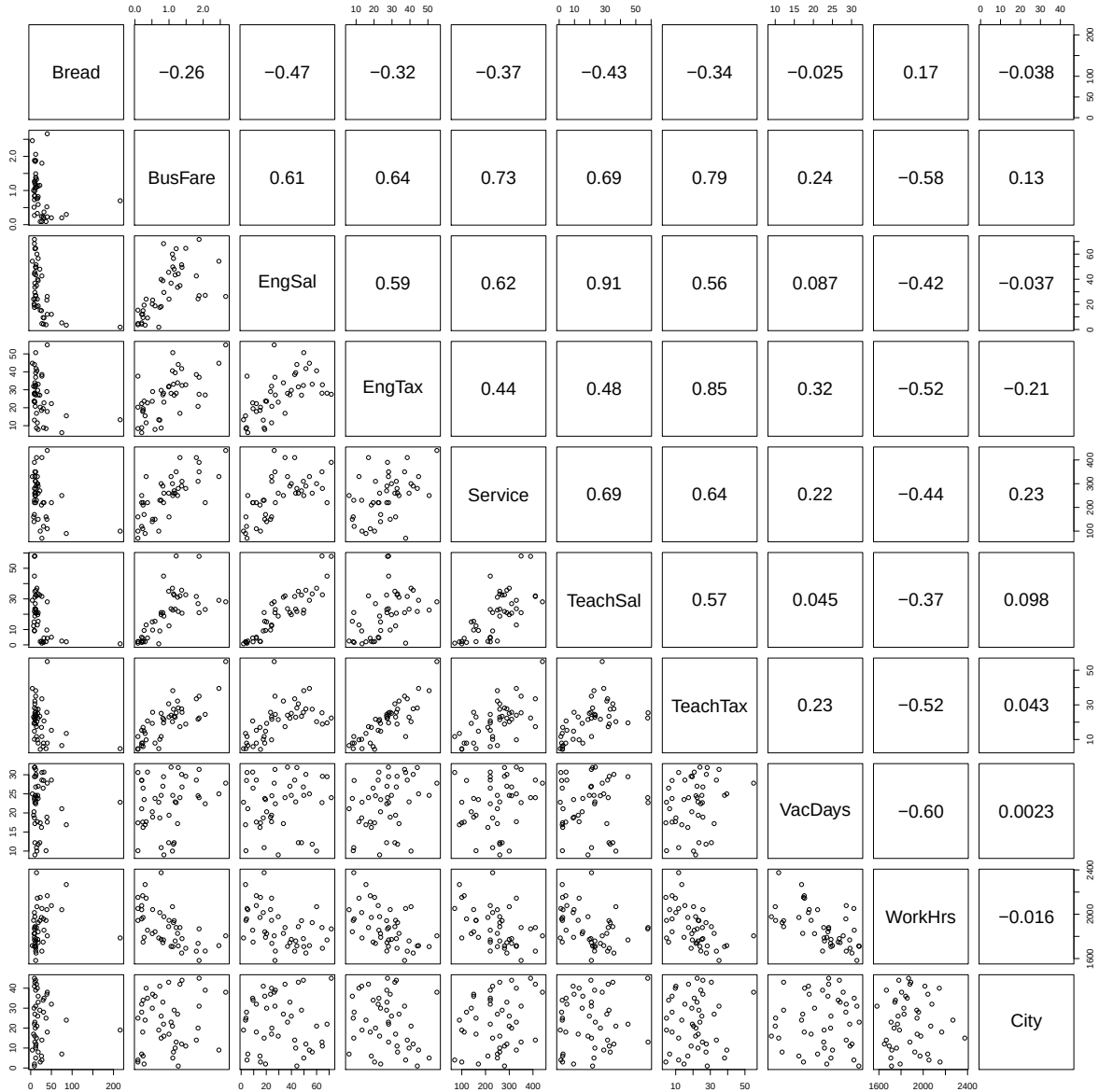


Correlação

Gráfico de dispersão com coeficiente de correlação

```
panel.cor <- function(x, y, digits = 2, prefix = ,
cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- cor(x, y)
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex=1.5)
}

pairs(dados[,-1], upper.panel = panel.cor)
```

Correlação entre as variáveis explicativas

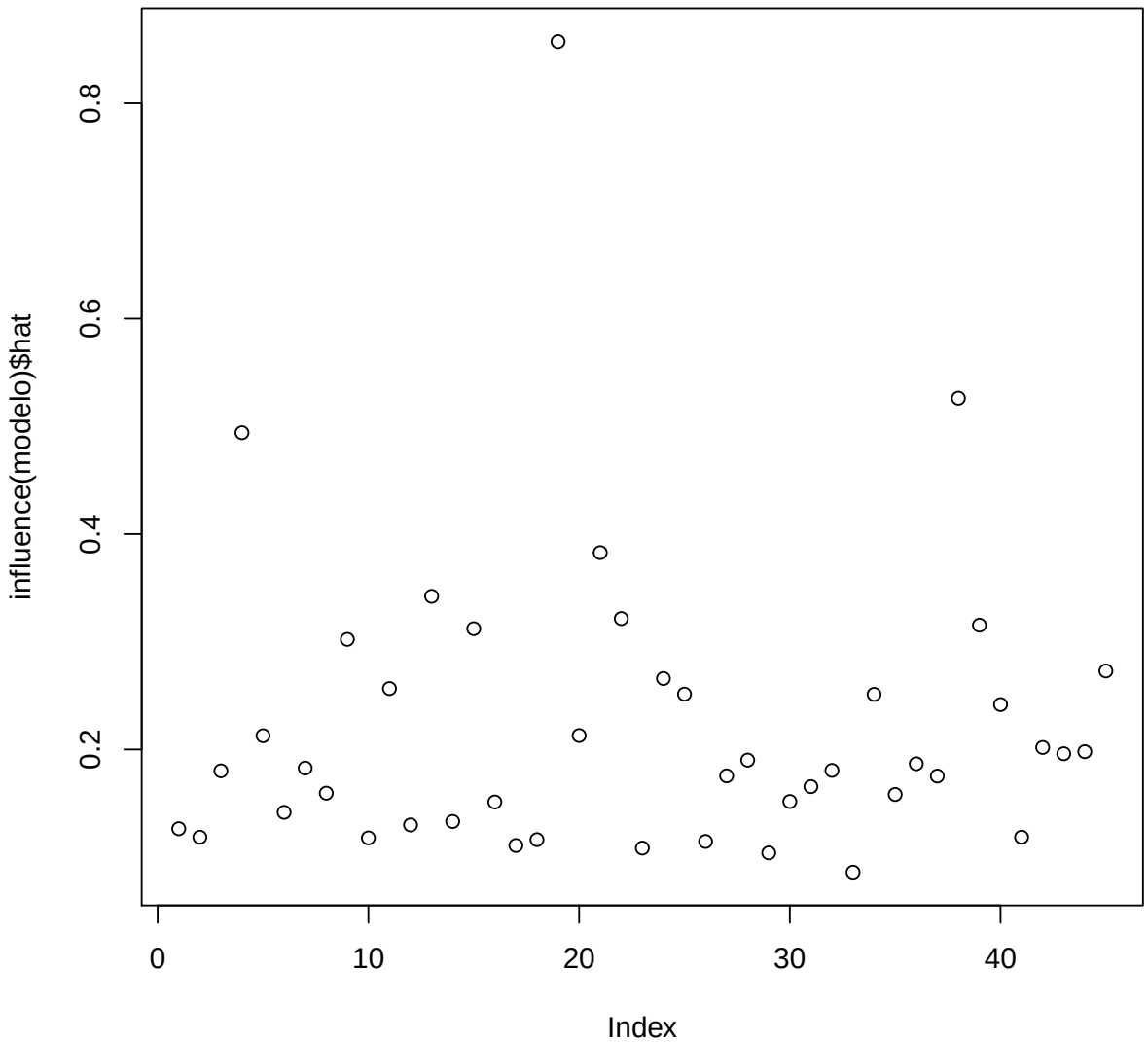


Ajuste do modelo

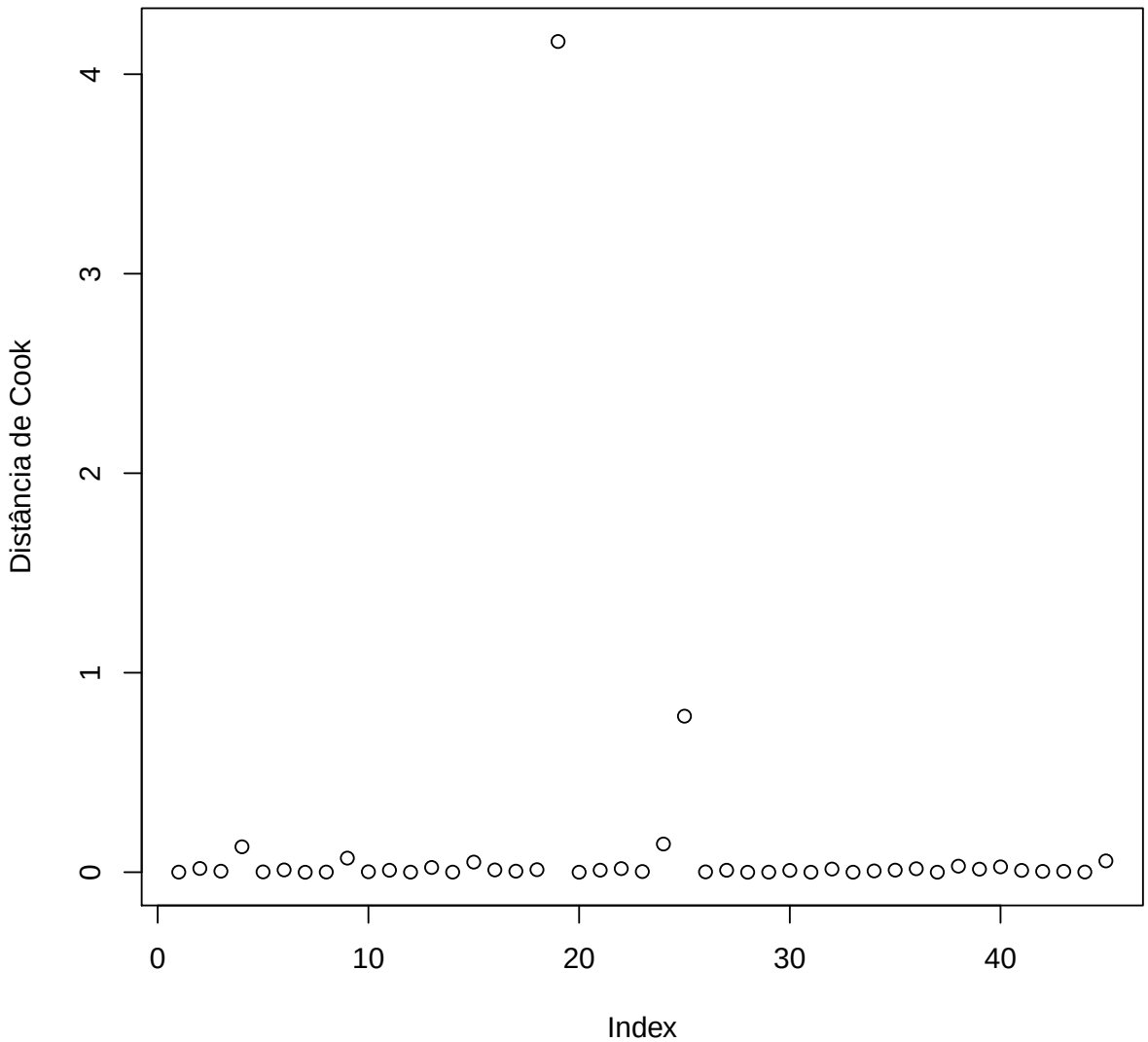
Código para ajuste do modelo

```
modelo <- lm(BigMac ~ Bread + BusFare + EngSal +
EngTax + TeachSal + TeachTax + Service + VacDays +
WorkHrs, data=dados)

summary(modelo)

vif(modelo)

plot(modelo$fitted.values, rstandard(modelo))

plot(modelo$fitted.values, rstudent(modelo))

plot(influence(modelo)$hat)

plot(cooks.distance(modelo))

library(car)
qqPlot(modelo, line='none', xlab='Quantis teóricos',
ylab='Quantis amostrais')
```

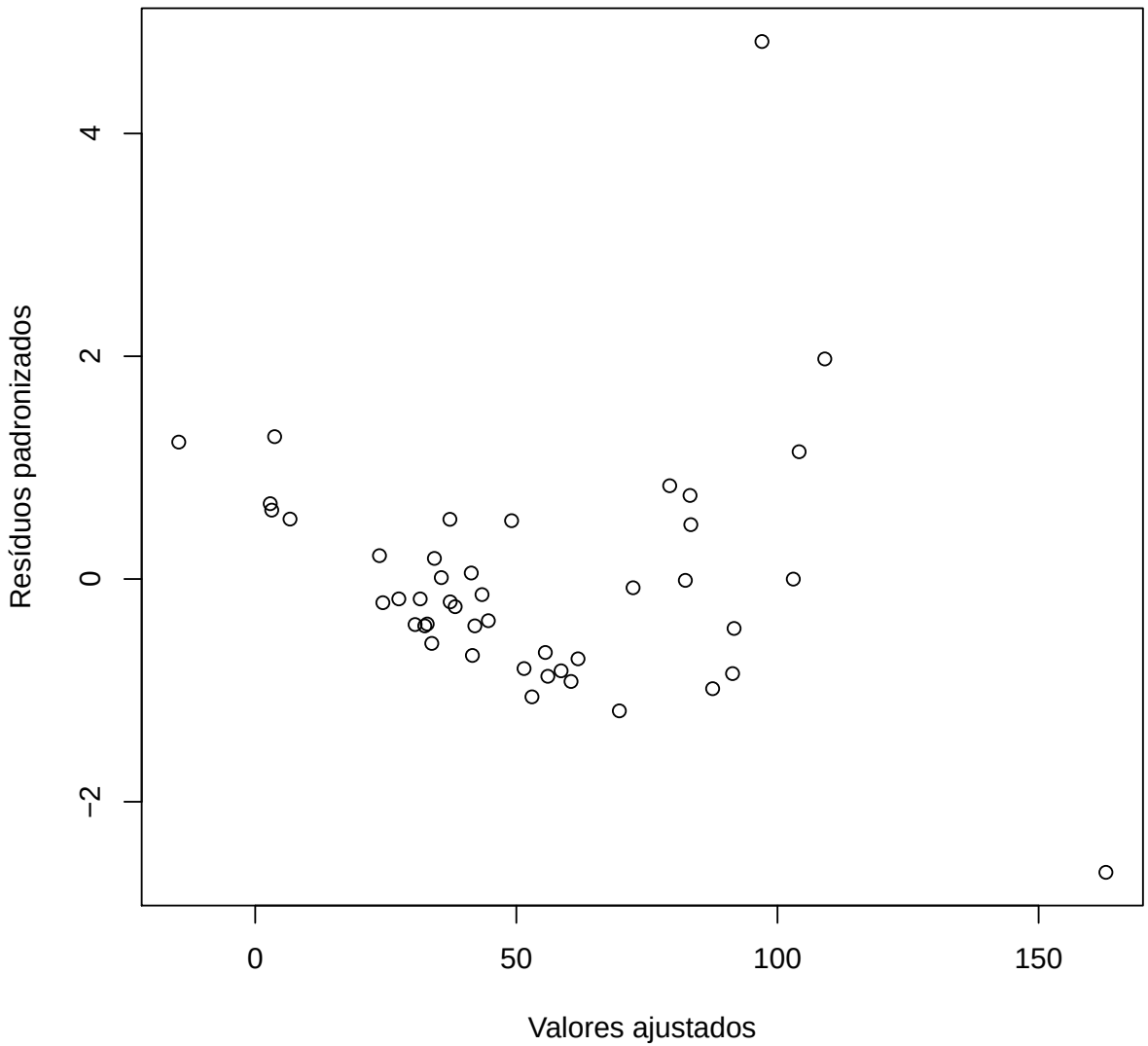
Análise de diagnóstico



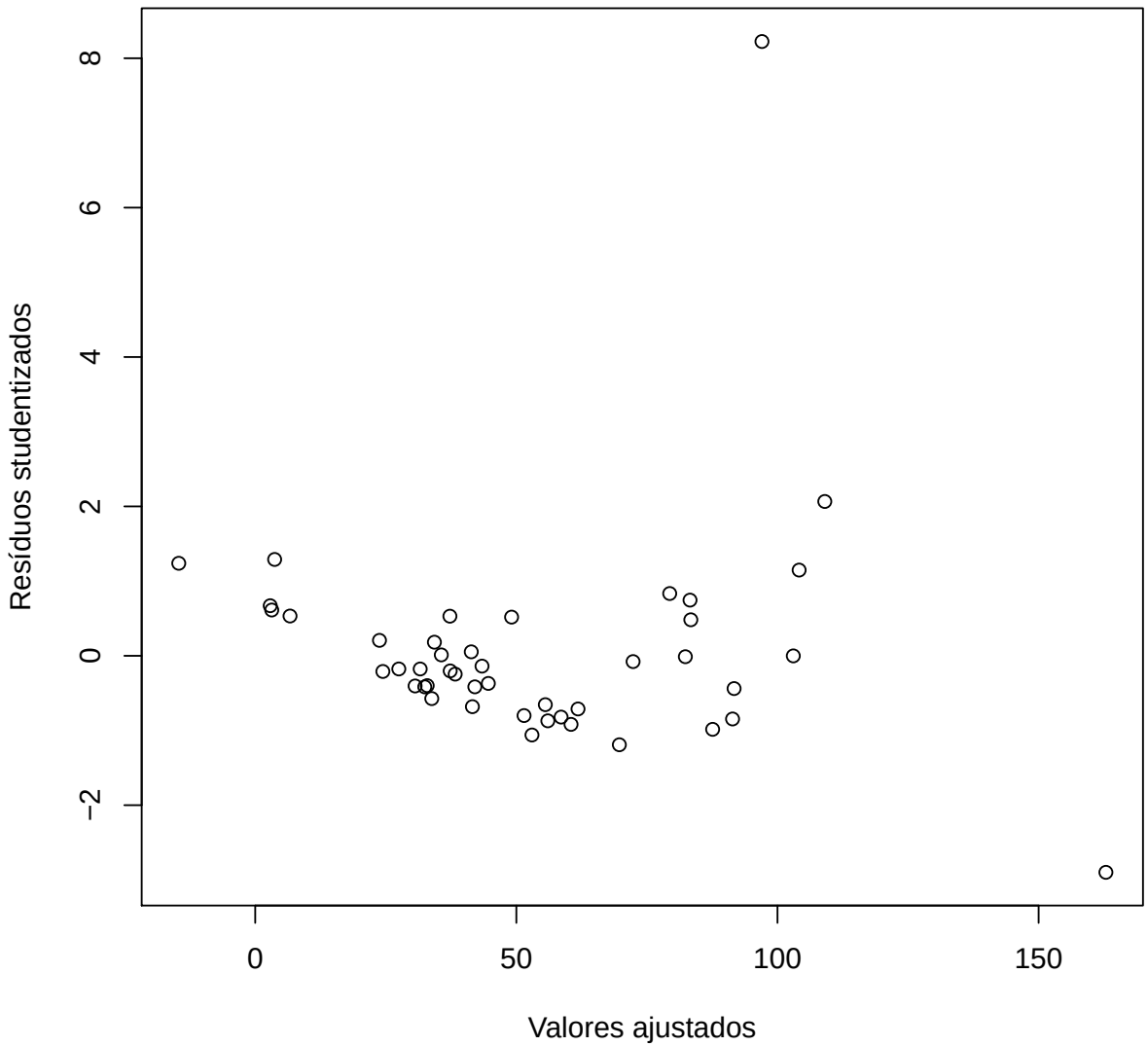
Distância de Cook



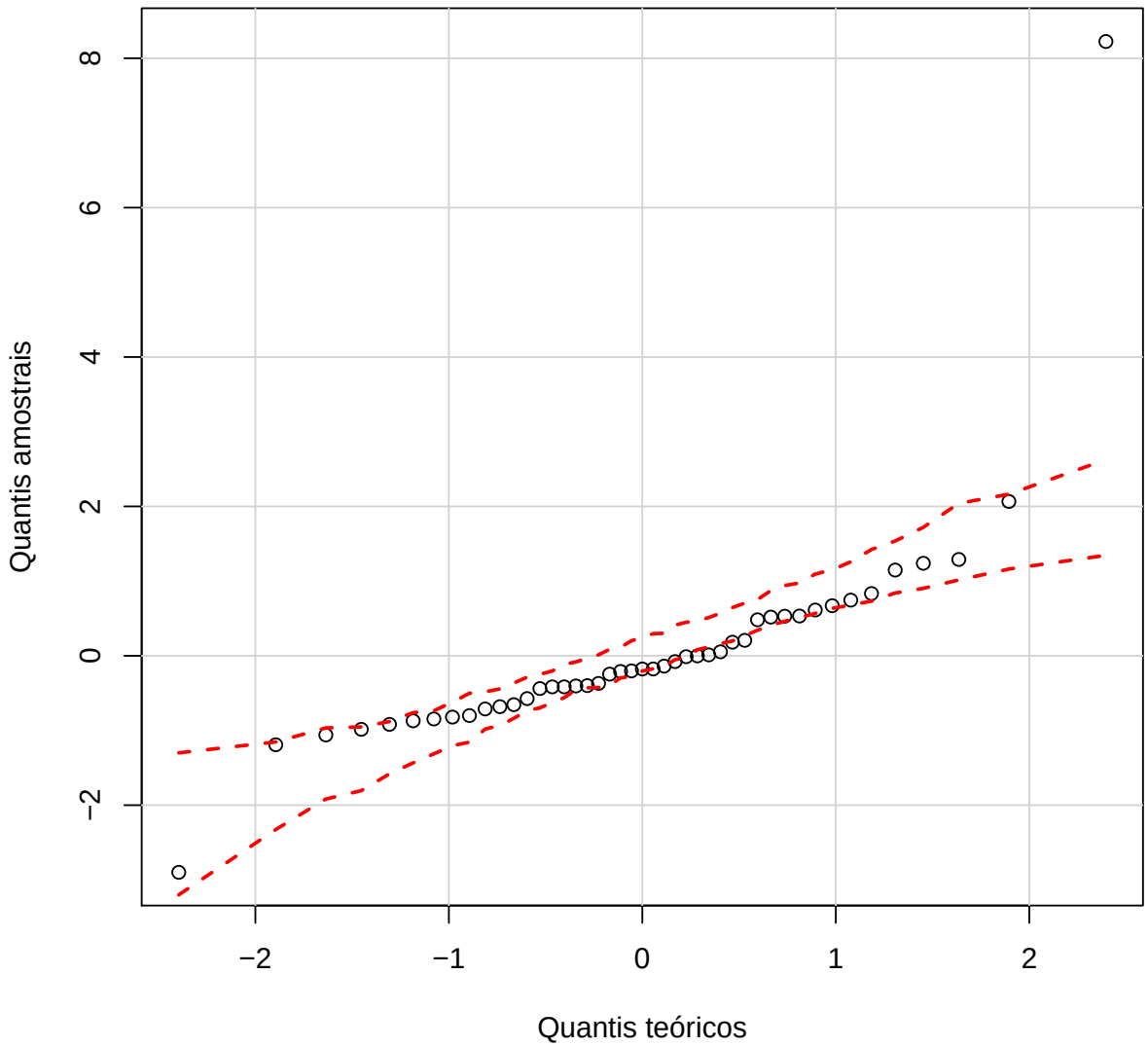
Análise de resíduos



Análise de resíduos



Análise de resíduos



Retirando observação da amostra

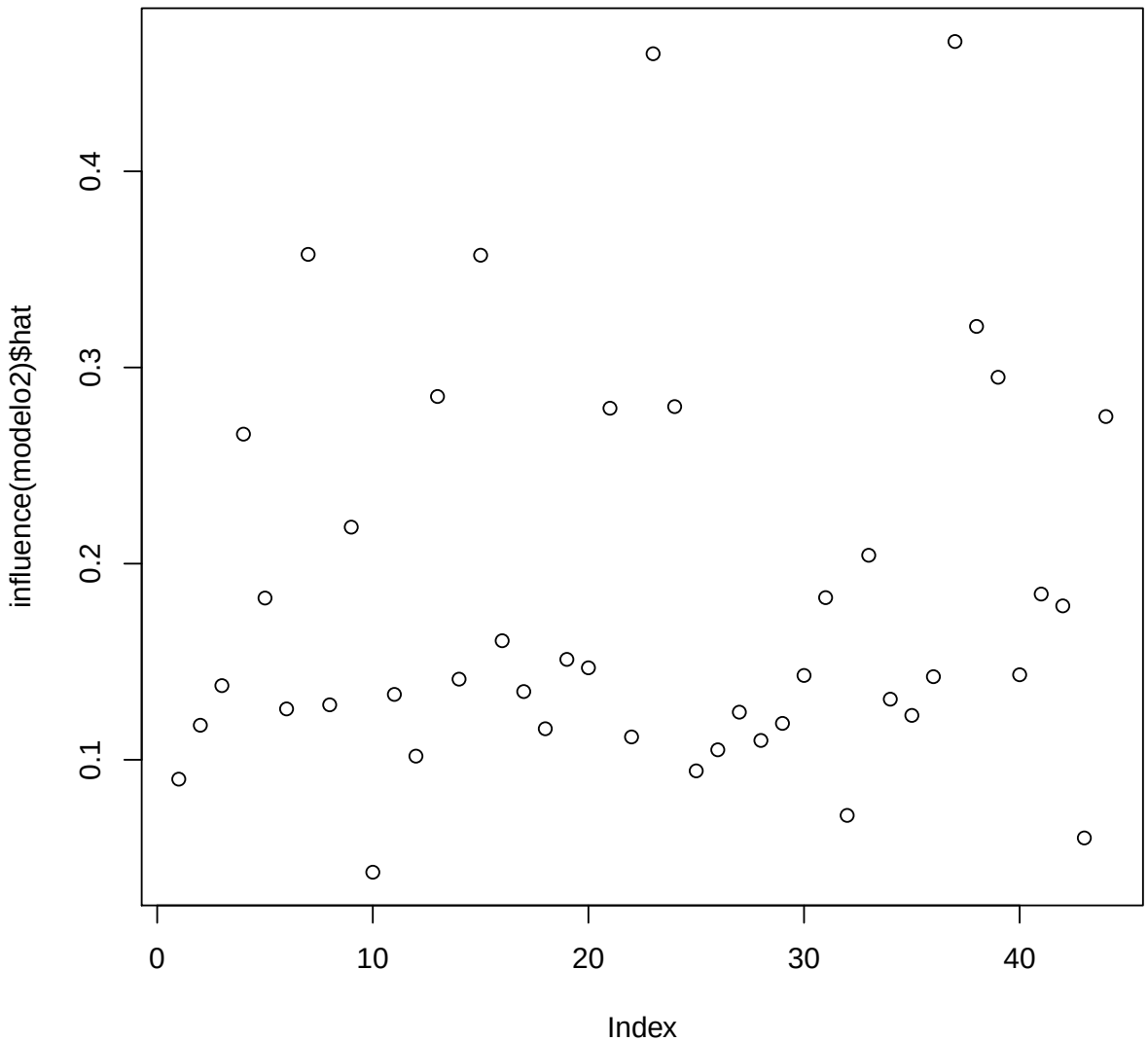
Observação número 19

```
# Para verificar valores da linha 19.
dados[19,]

dados2 <- dados[-19, ]
```

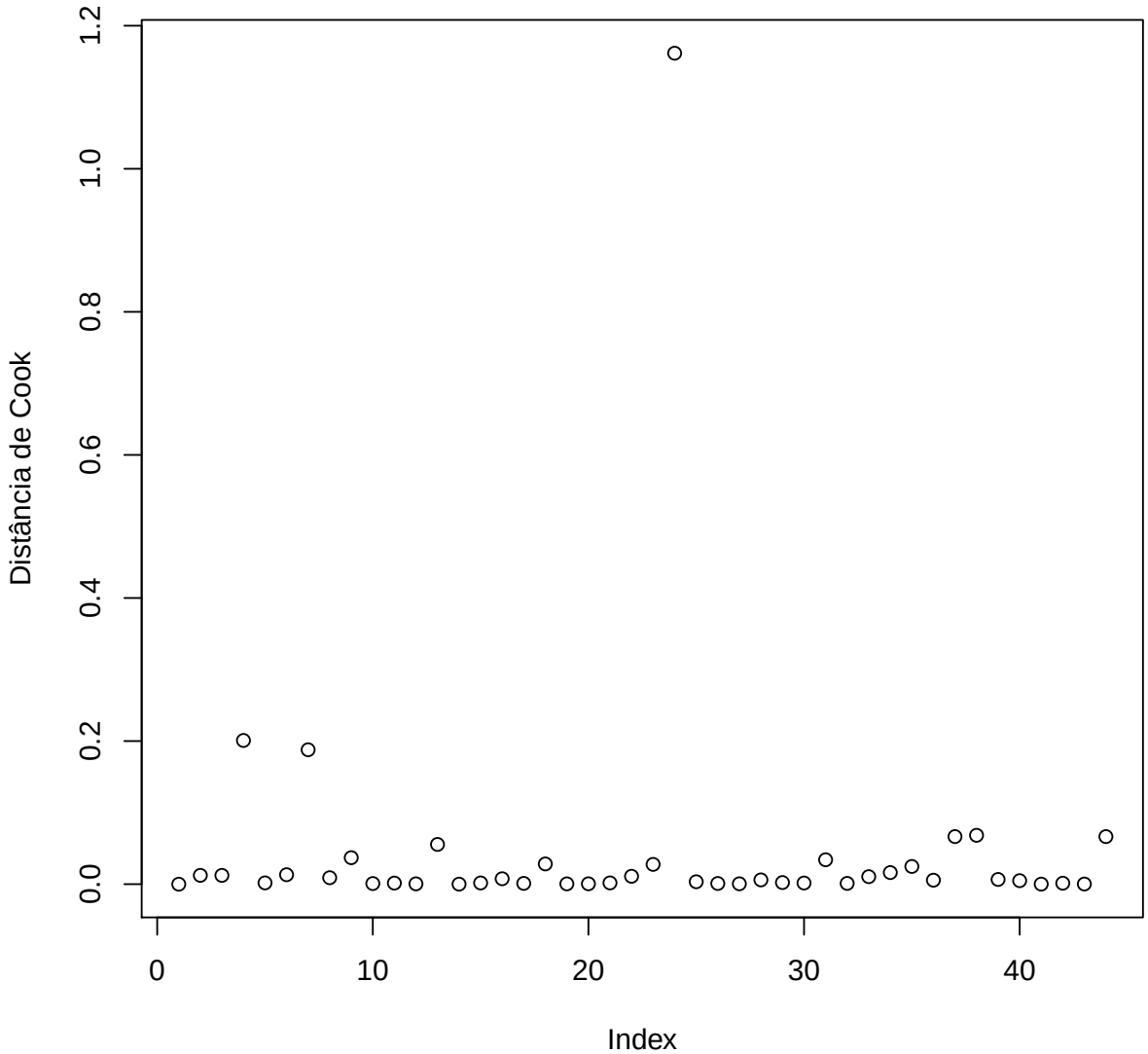
Ajustando modelo novamente

```
modelo2 <- lm(BigMac ~ Bread + BusFare + Service +
TeachSal + TeachTax + VacDays + WorkHrs, data=dados2)
```

Análise de diagnóstico



Distância de Cook



Retirando observação influente

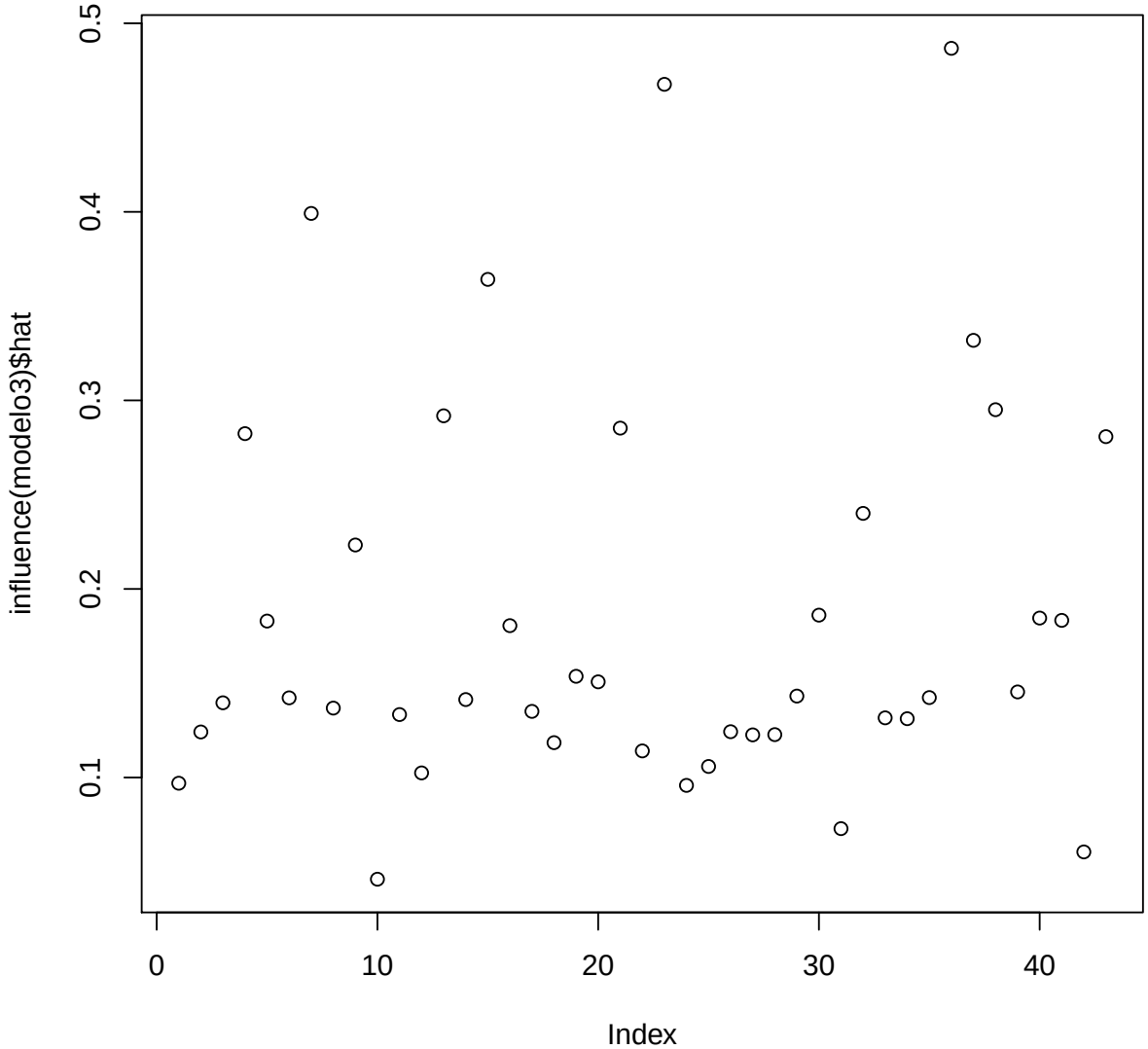
Observação número 19

```
# Para verificar valores da linha 19.
dados2[24,]

dados3 <- dados2[-24, ]
```

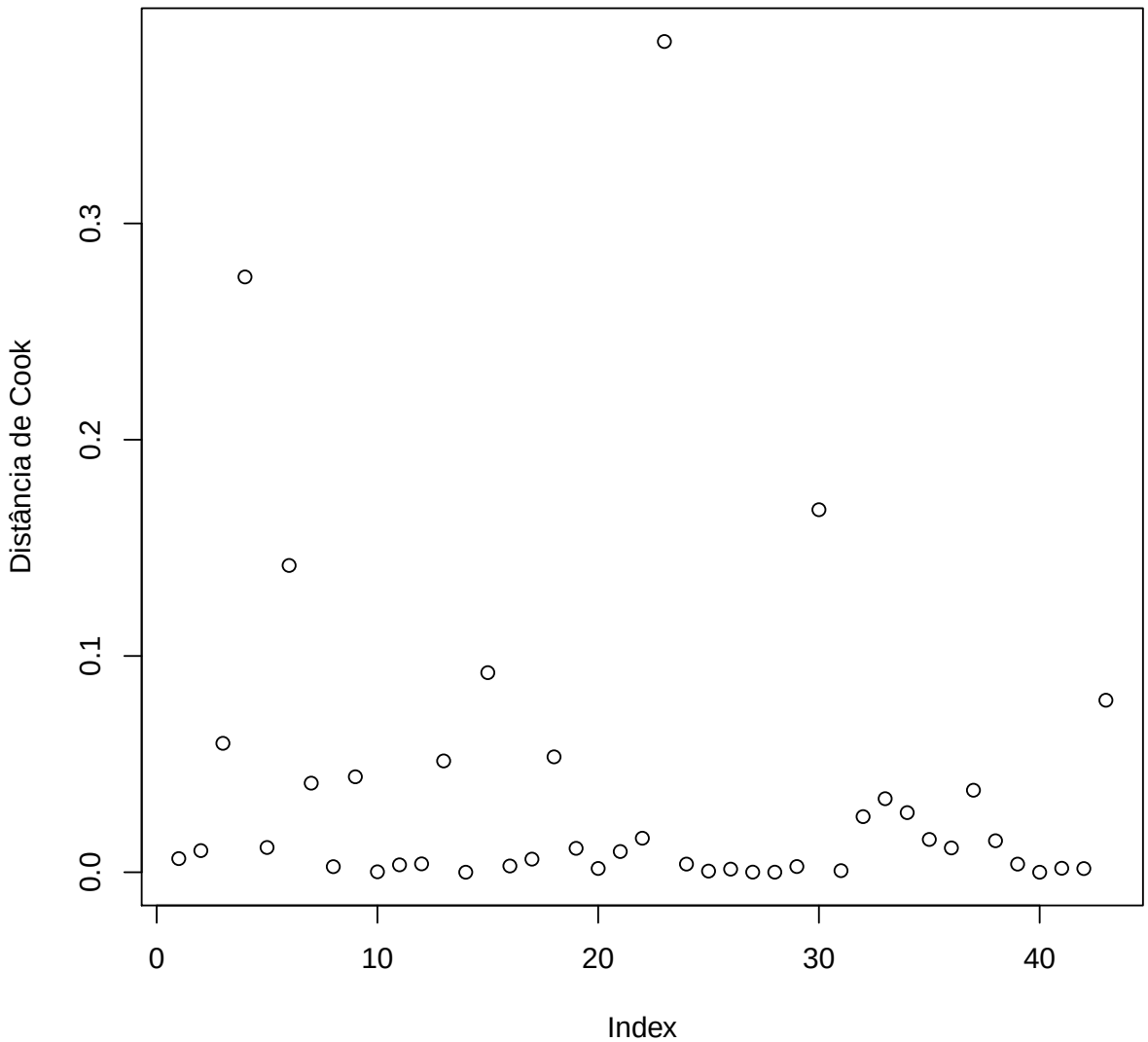
Ajustando modelo novamente

```
modelo3 <- lm(BigMac ~ Bread + BusFare + Service +
TeachSal + TeachTax + VacDays + WorkHrs, data=dados3)
```

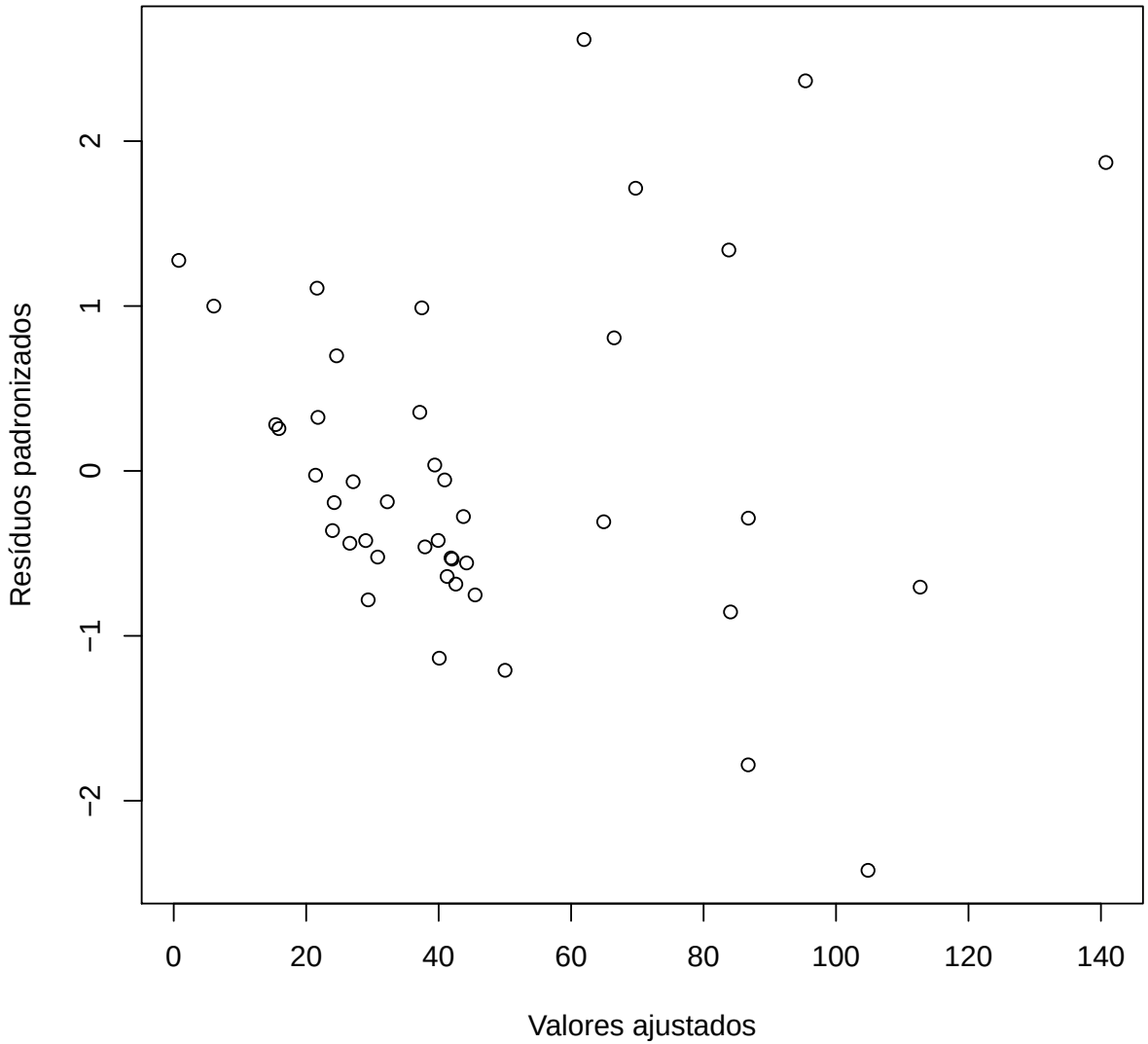
Análise de diagnóstico



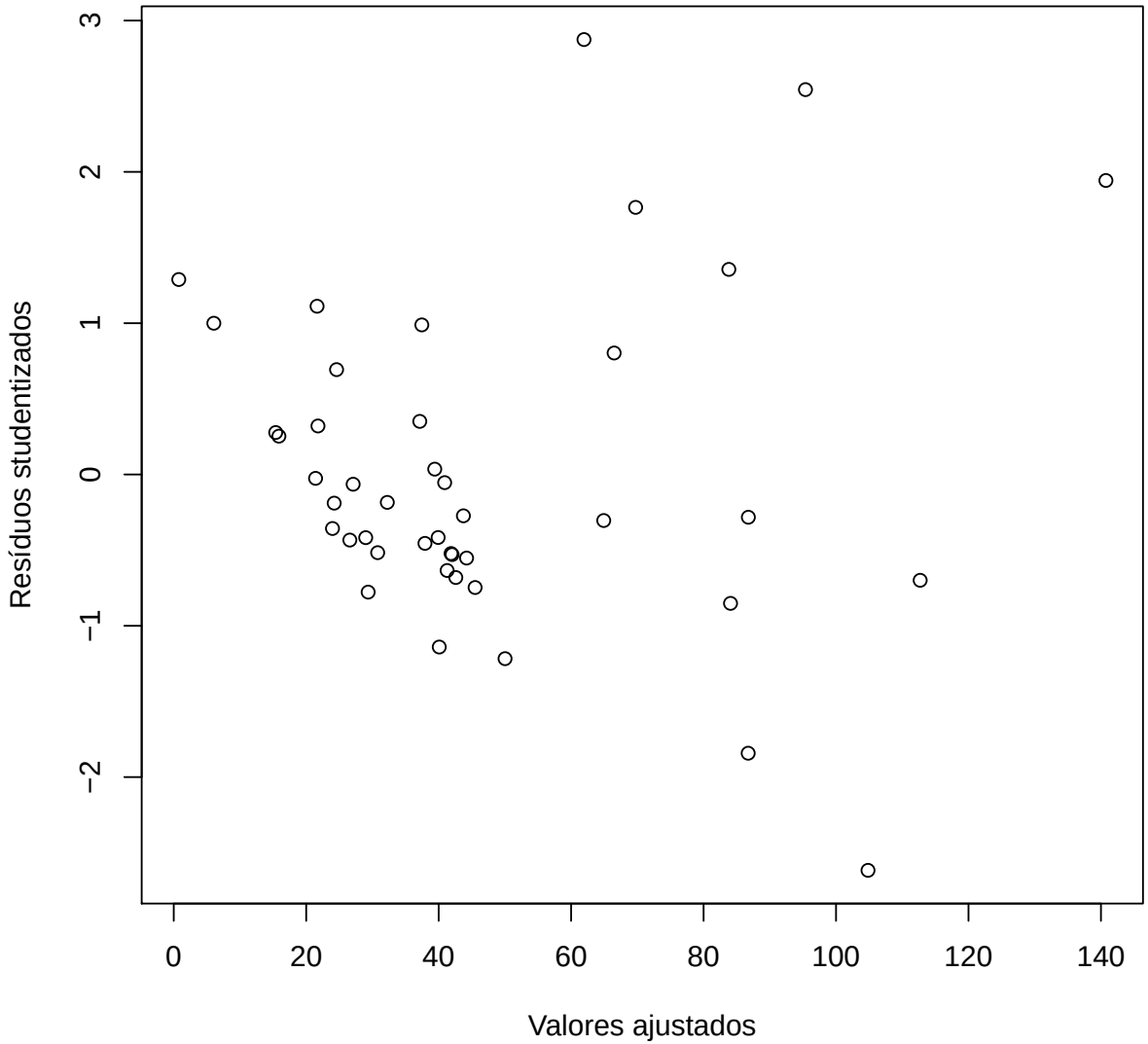
Distância de Cook



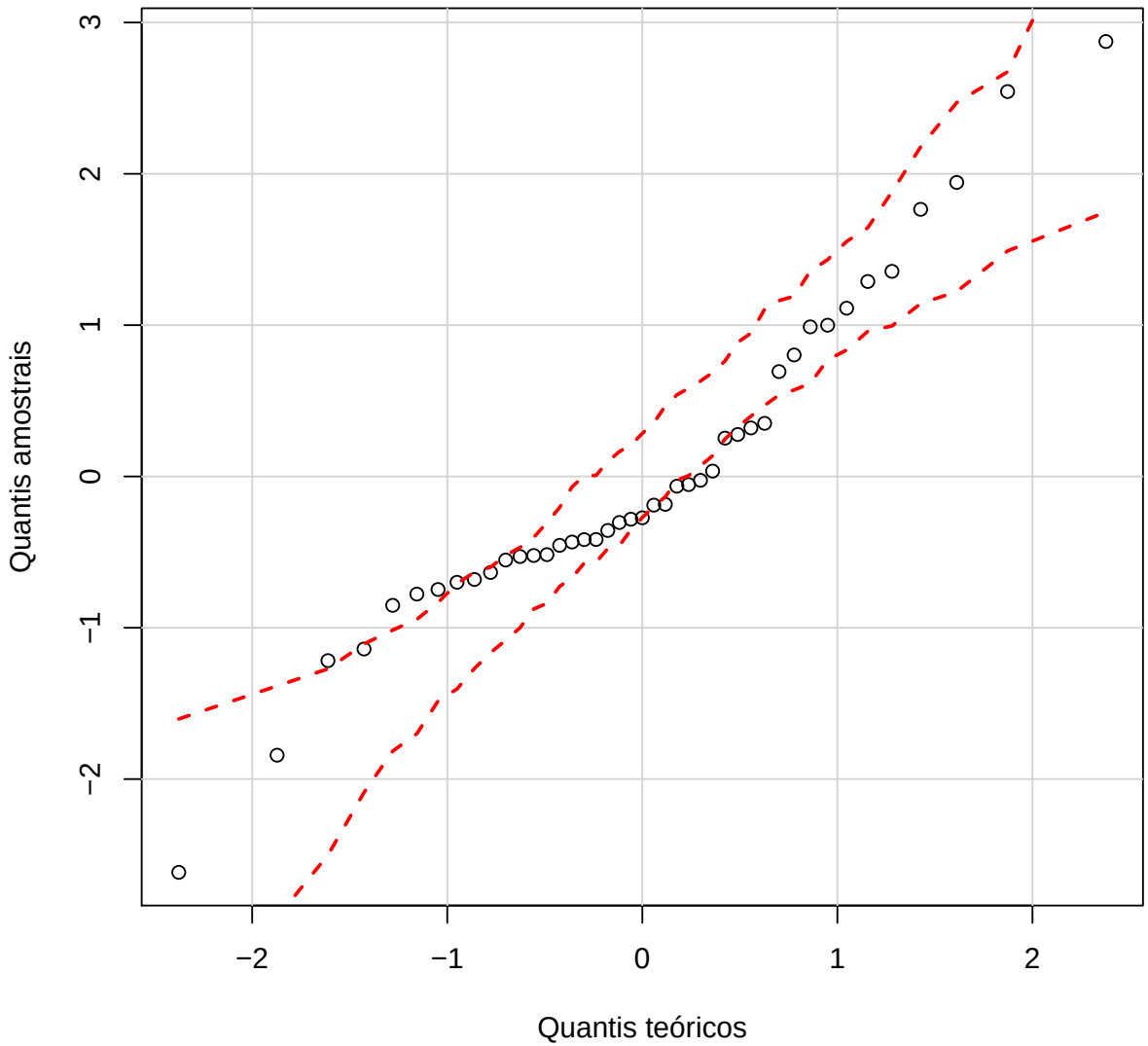
Análise de resíduos - 3



Análise de resíduos - 3



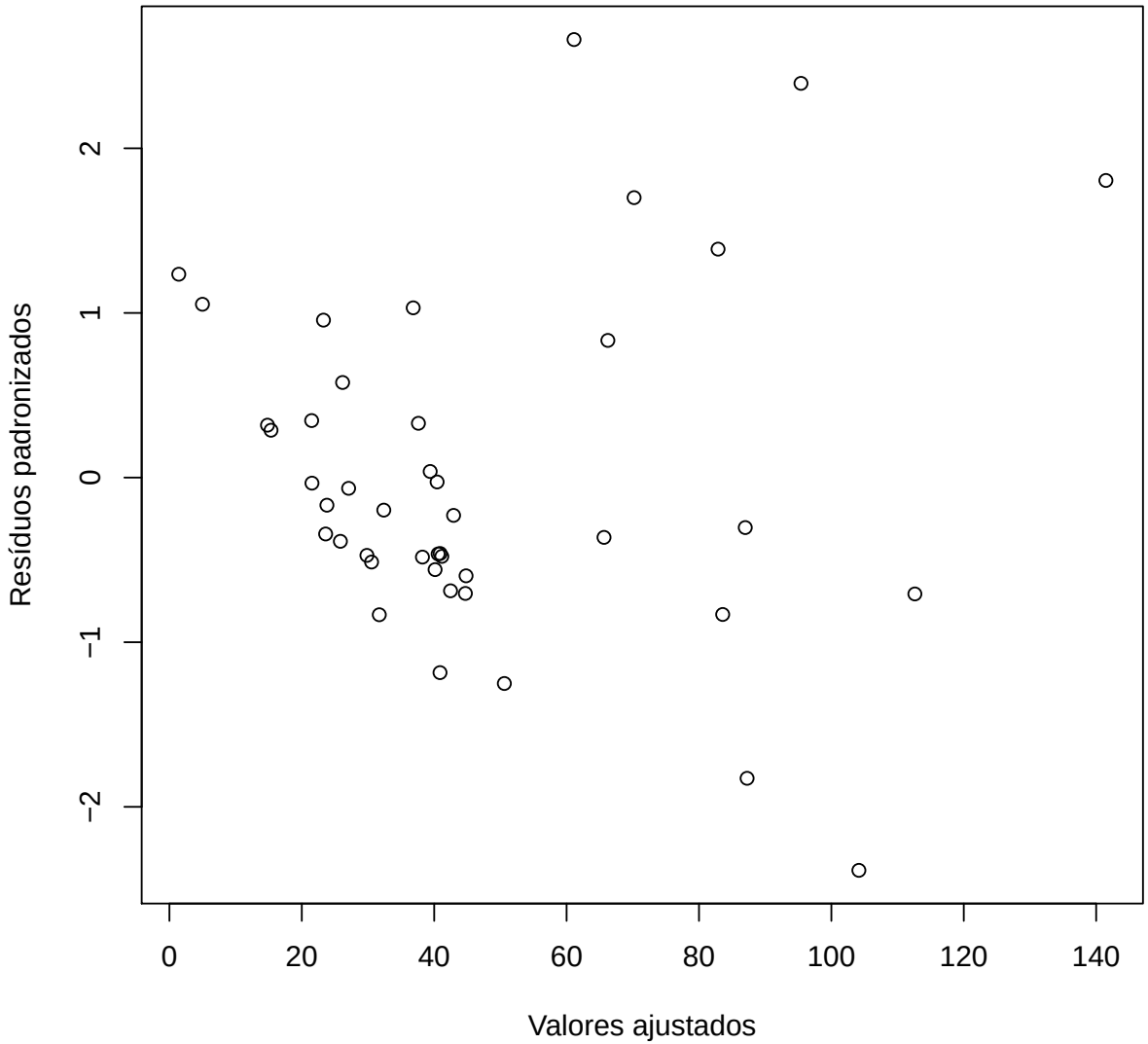
Análise de resíduos - 3



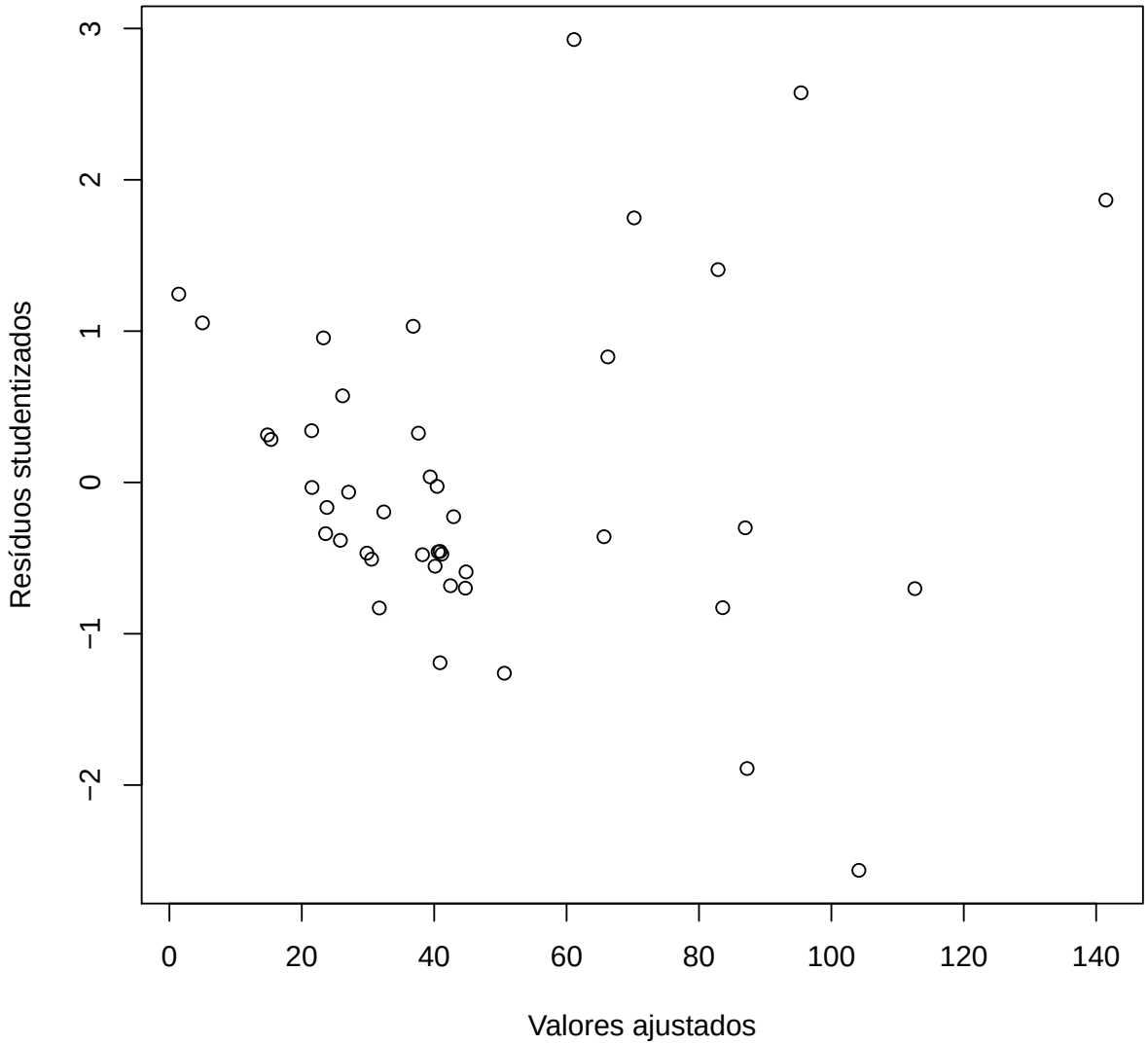
Ajustando modelo novamente

```
modelo4 <- lm(BigMac ~ Bread + Service + TeachSal +
TeachTax + VacDays + WorkHrs, data=dados3)
```

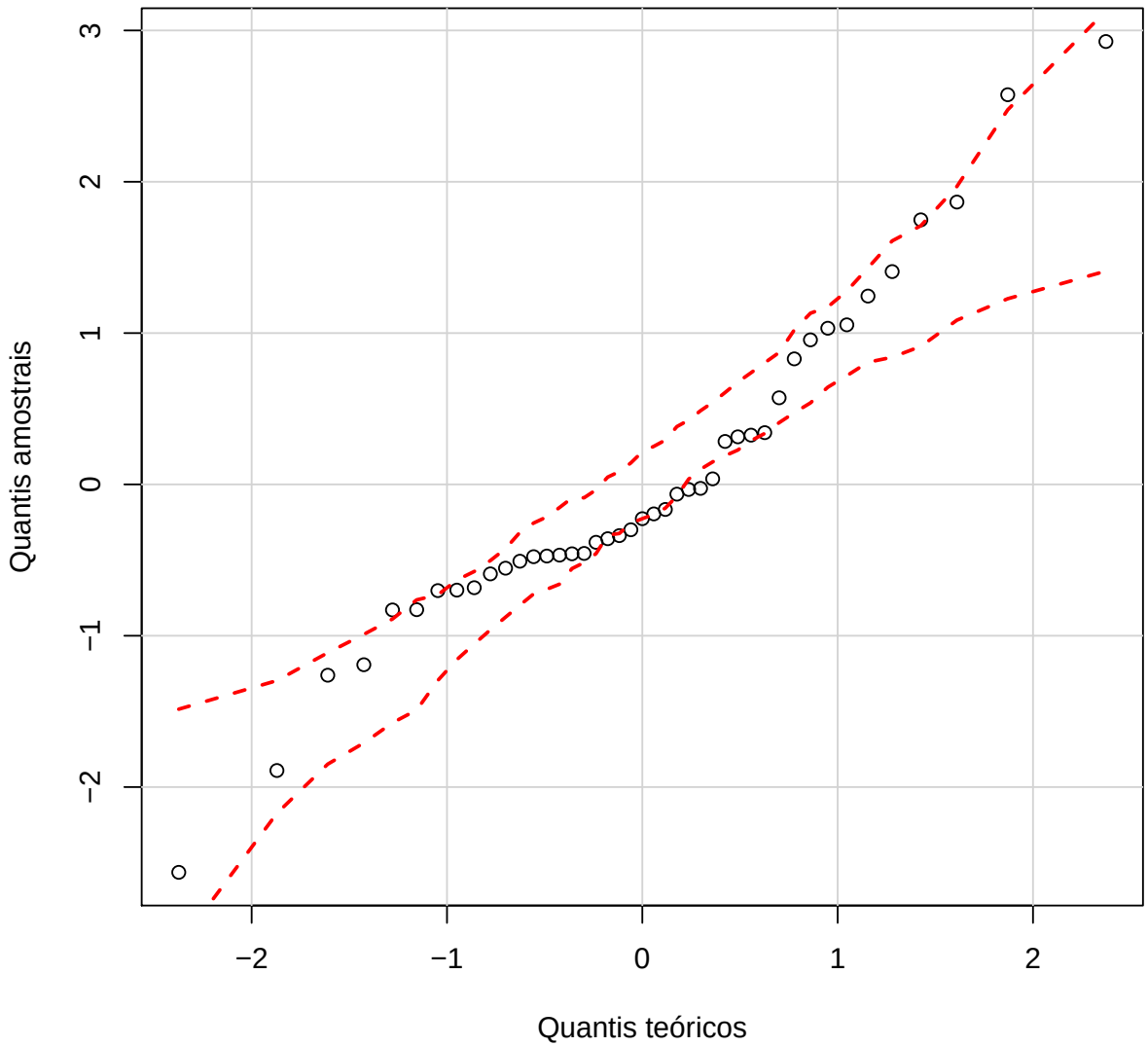
Análise de resíduos - 4



Análise de resíduos - 4



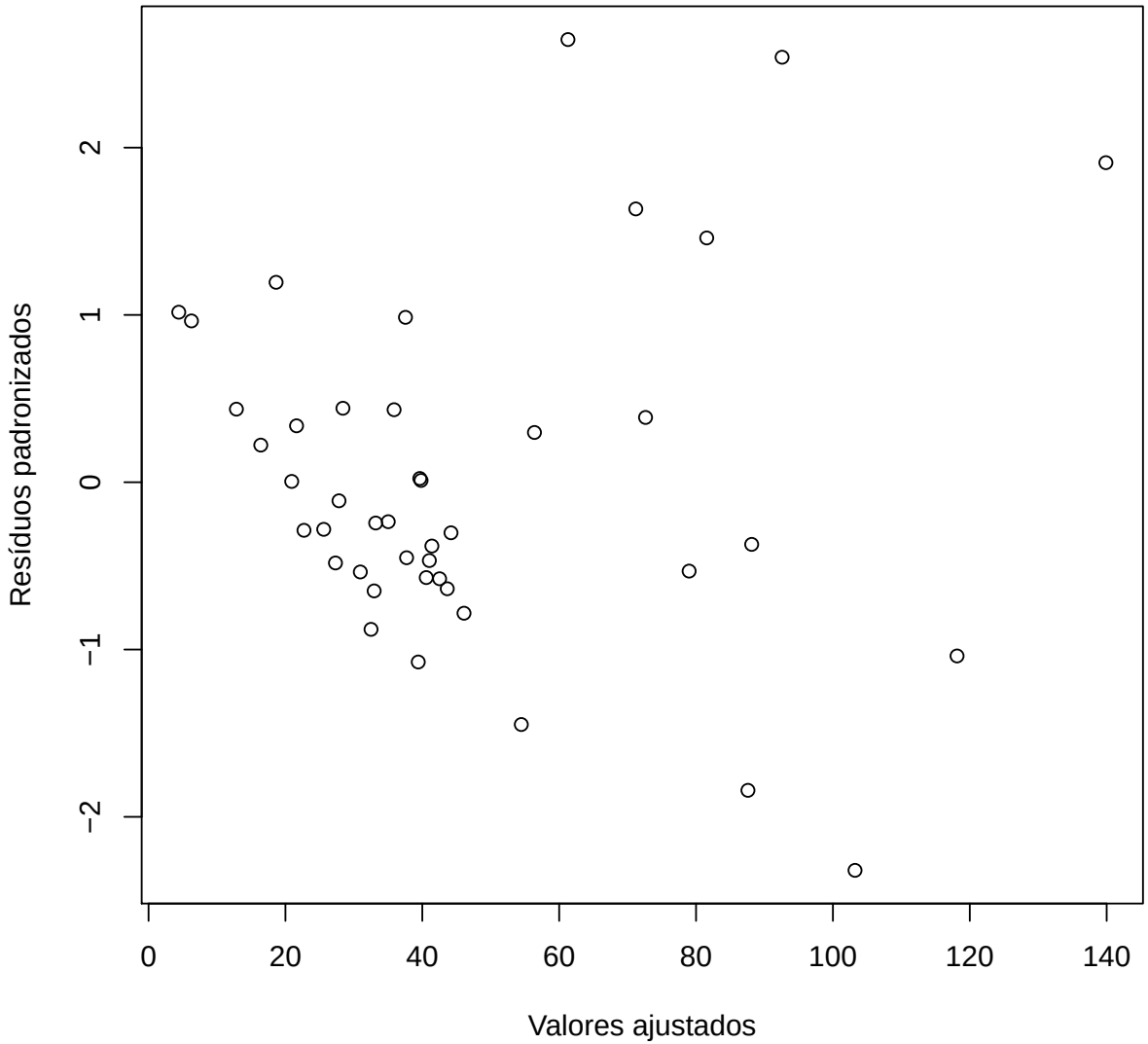
Análise de resíduos - 4



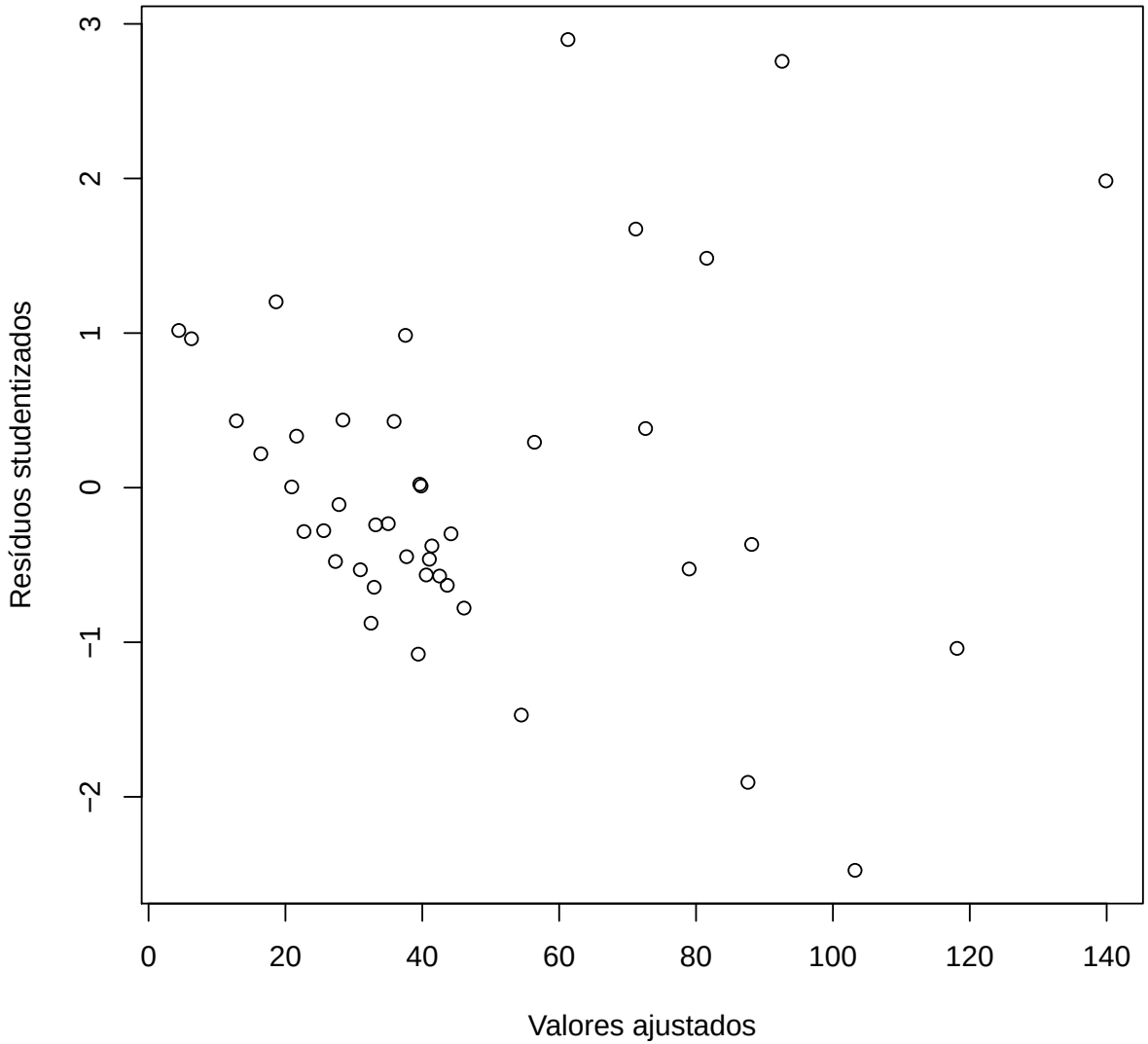
Ajustando modelo novamente

```
modelo5 <- lm(BigMac ~ Bread + Service + TeachSal +
VacDays + WorkHrs, data=dados3)
```

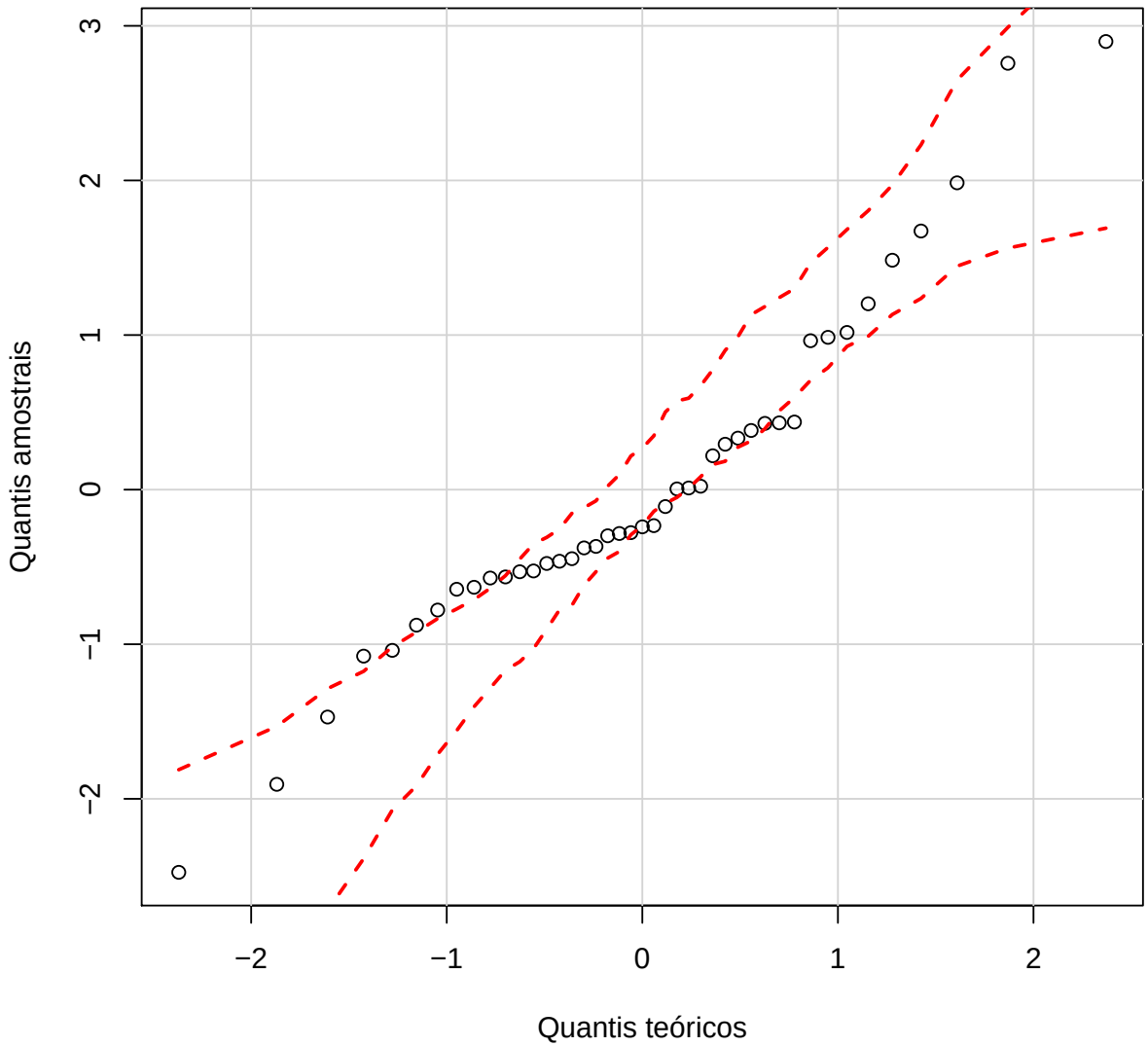
Análise de resíduos - 5



Análise de resíduos - 5



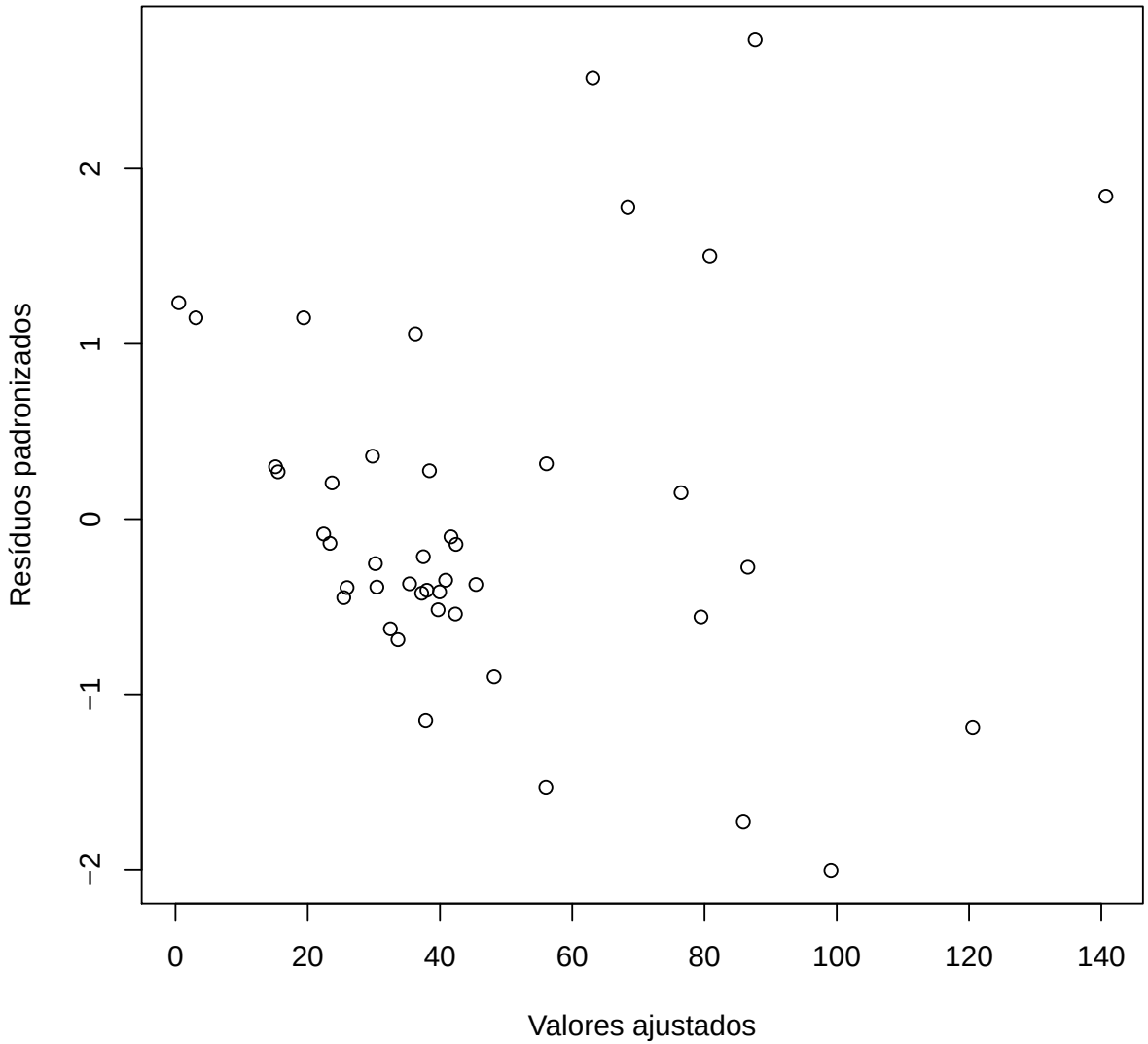
Análise de resíduos - 5



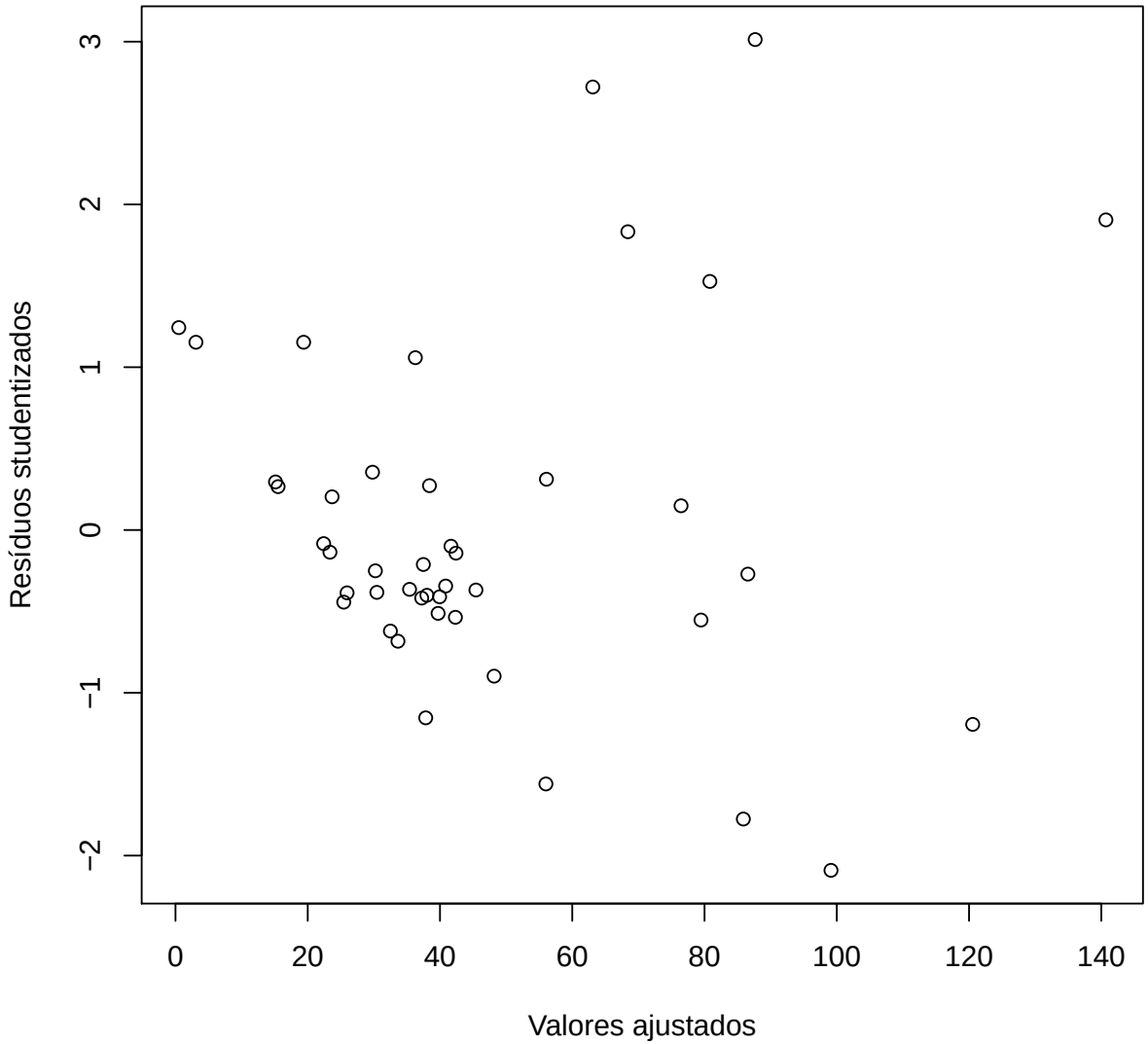
Ajustando modelo novamente

```
modelo6 <- lm(BigMac ~ Bread + Service + TeachSal +
VacDays, data=dados3)
```

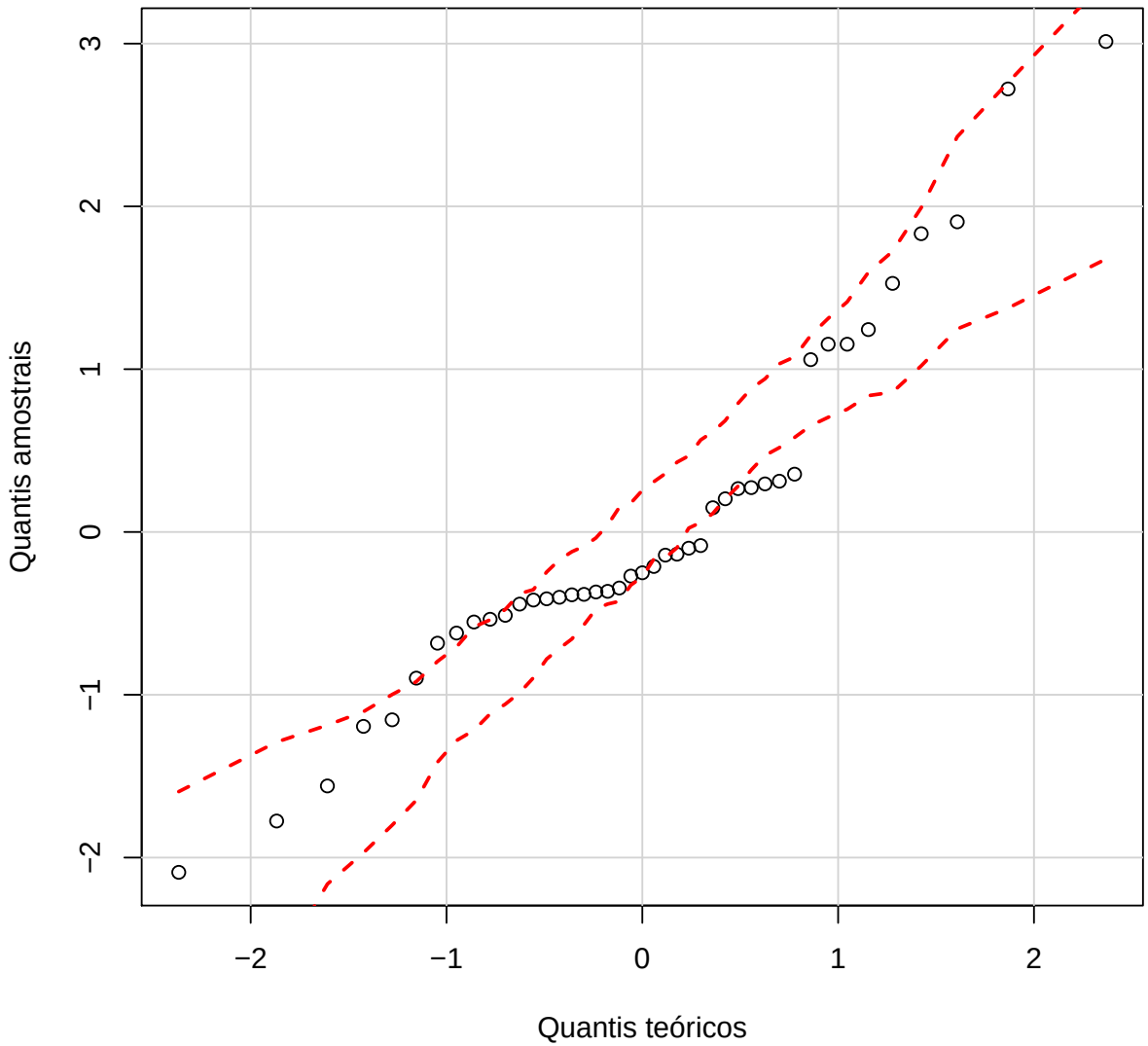
Análise de resíduos - 6



Análise de resíduos - 6



Análise de resíduos - 6

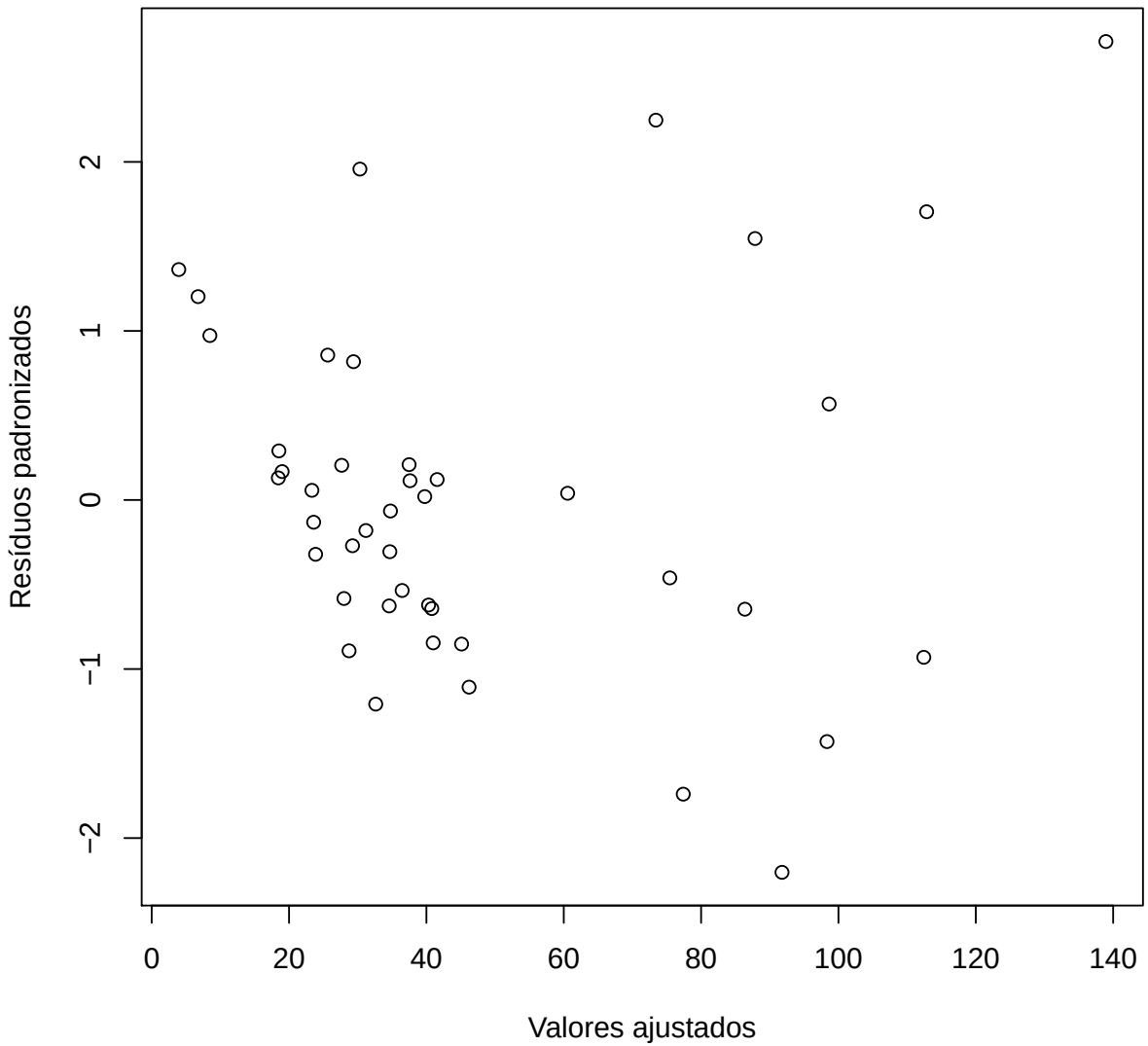


Considerando transformações em variáveis explicativas

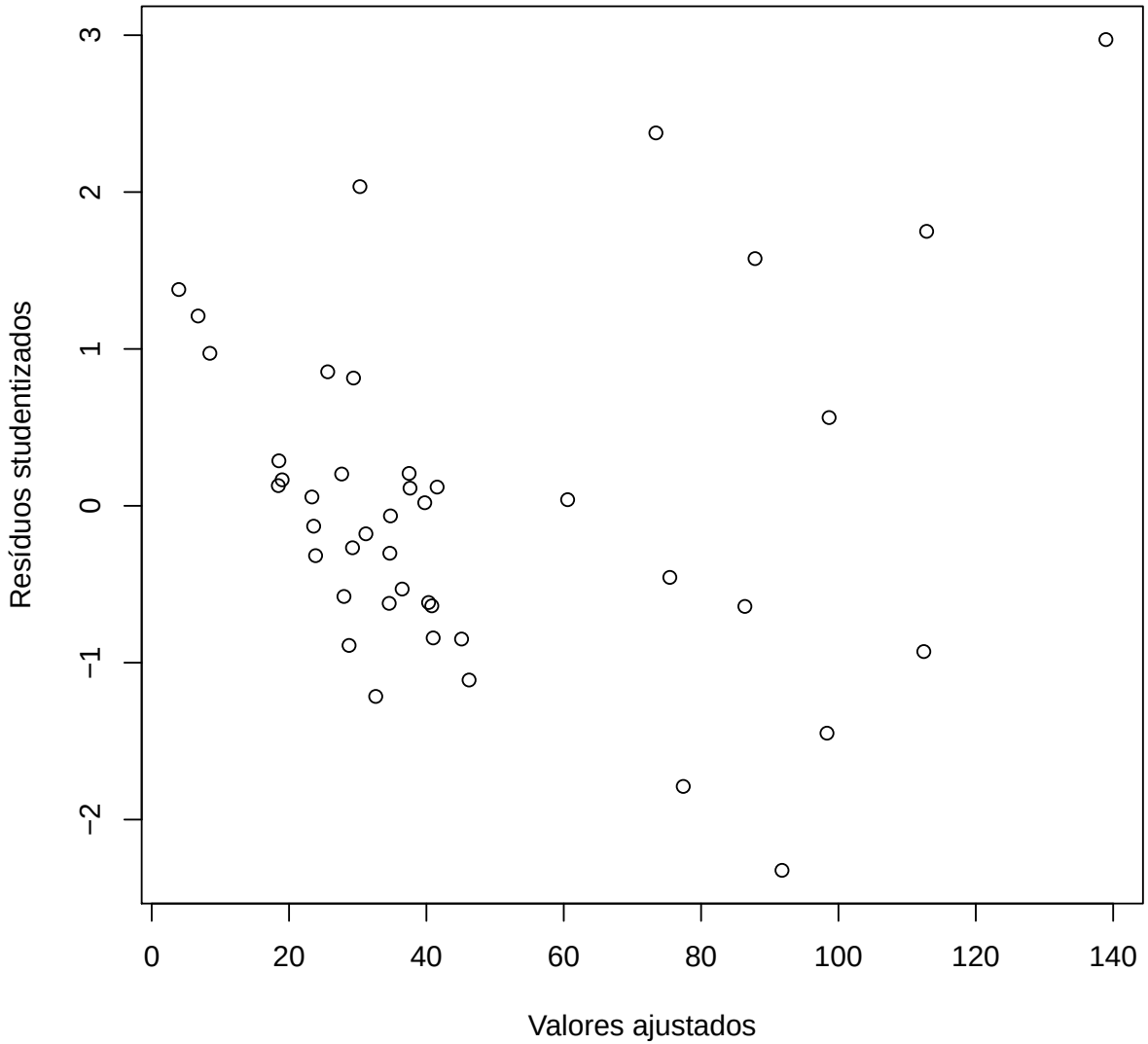
Ajustando modelo novamente

```
modeloAlt <- lm(BigMac ~ Bread + log(TeachSal) +
log(TeachTax), data=dados3, data=dados3)
```

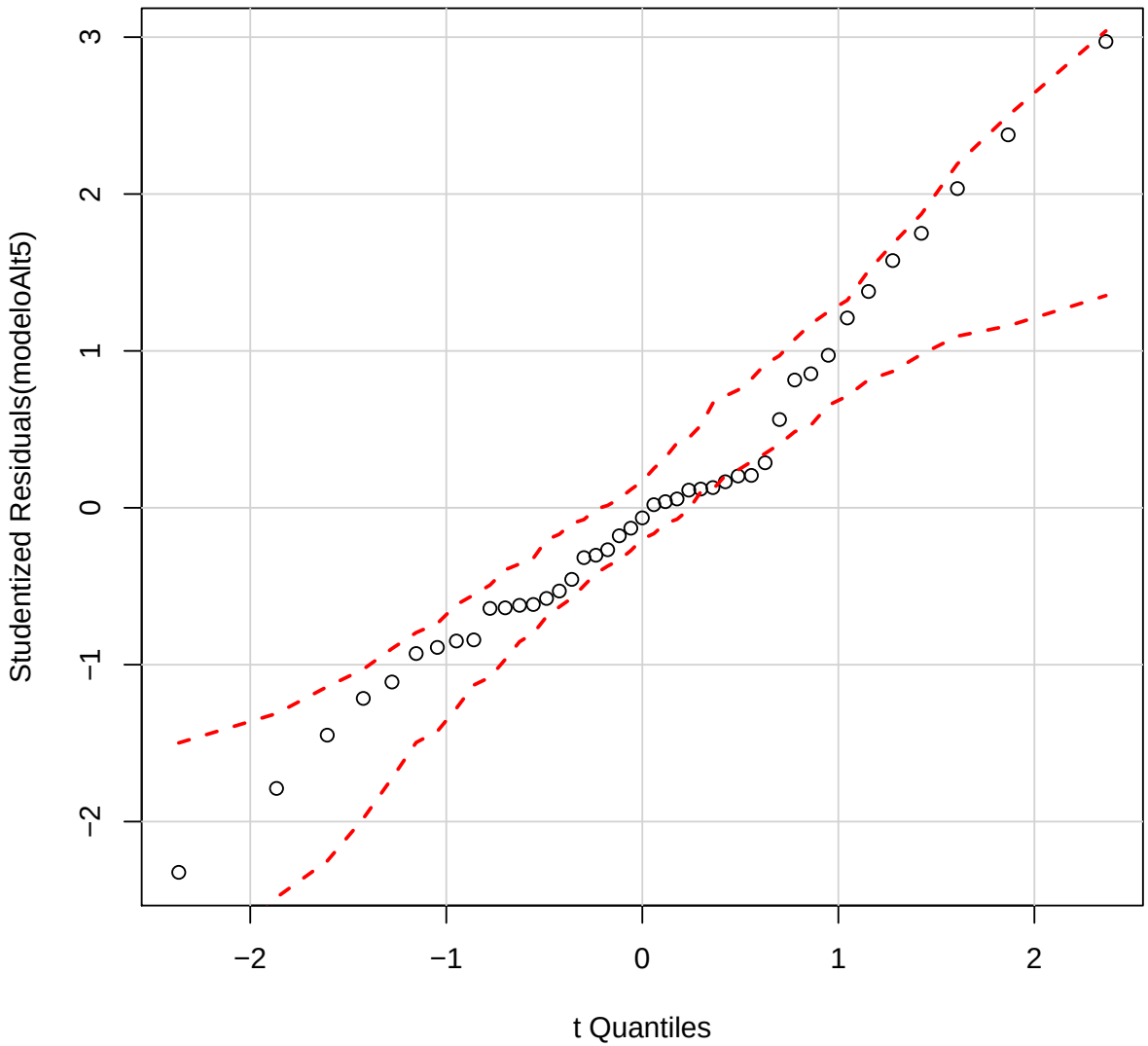
Análise de resíduos - 6



Análise de resíduos - 6



Análise de resíduos - 6



Estatística F-parcial

Ajustar modelo completo

```
modeloC <- lm(BigMac ~ Bread + BusFare + Service +
TeachSal + TeachTax + VacDays + WorkHrs, data=dados3)
```

Ajustar modelo reduzido

```
modeloR <- lm(BigMac ~ Bread + BusFare + Service +
TeachSal, data=dados3)
```

Estatística de teste F-parcial

```
anova(modeloR, modeloC)
```

Links úteis

Lista de comandos úteis na análise de regressão.

- http://cran.r-project.org/doc/contrib/Ricci-refcard-regression.pdf

Pacote ggplot2.

- http://docs.ggplot2.org/current/

Pacote Rcpp para utilização de C++ em conjunto com o R.

- http://www.rcpp.org/.
- http://adv-r.had.co.nz/Rcpp.html.

Pacote shiny.

- http://shiny.rstudio.com/.

Pacote knitr (*Reproducible research*).

- http://yihui.name/knitr/.