Allan Victor Almeida Faria (190127180), Ananda Almeida de Sá (150117345),
Bruno Kevyn Andrade de Souza

Trabalho de Regressão Linear

Brasília, DF

21/02/2021

Allan Victor Almeida Faria (190127180), Ananda Almeida de Sá (150117345),
Bruno Kevyn Andrade de Souza

Trabalho de Regressão Linear

Trabalho de Regressão Linear de Análise de dados hospitalares.

Universidade de Brasília (UnB)

Instituto de Ciências Exatas (IE)

Departamento de Estatística (DE)

Brasília, DF

21/02/2021

Resumo

resumo aqui

Palavras-chaves: 1. Análise de dados.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
SAEB	Sistema de Avaliação da Educação Básica

Lista de símbolos

σ	Letra grega minúscula sigma
μ	Letra grega minúscula mu

Sumário

1 RESULT

1.1 Introdução

Tipo de problema, tipo de dados, proposta para contornar o problema

1.1.1 Leitura de dados

O programa utilizado para analisar os dados disponibilizados em Excel será o R Studio, versão 4.2.0, importados como um data frame (planilha), onde as colunas representam as variáveis de estudo e cada linha representa um hospital dos Estados Unidos no período de 1975-1976.

```
# Tabela de nomes X1: Nome variavel
knitr::kable(cbind(names(data),names(datax)),
             caption = 'Descrição dos códigos da tabela com a seguinte indentificação
```

Tabela 1 – Descrição dos códigos da tabela com a seguinte indentificação da variável.

Número de Identificação	ID
Duração da Internação	X1
Idade	X2
Risco de Infecção	X3
Proporção de Culturas de Rotina	X4
Proporção de Raio-X de Tórax de Rotina	X5
Número de leitos	X6
Filiação a Escola de Medicina	X7
Região	X8
Média diária de pacientes	X9
Número de enfermeiro(s)	X10
Facilidades e serviços disponíveis	X11

1.1.1.1 Descrição das variáveis

- **Duração da Internação**

A duração de internação é uma variável quantitativa contínua que representa a duração média da internação de todos os pacientes no hospital (em dias).

- **Idade**

A idade é uma variável quantitativa contínua que representa a idade média dos pacientes de cada hospital.

- **Risco de Infecção**

O risco de infecção é uma variável quantitativa contínua que representa a probabilidade média estimada de adquirir infecção no hospital (em %).

- **Proporção de Culturas de Rotina**

A proporção de culturas de rotina é uma variável quantitativa contínua que representa a razão do número de culturas realizadas com relação ao número de pacientes sem sinais ou sintomas de infeção adquirida no hospital, vezes 100.

- **Proporção de Raio-X de Tórax de Rotina**

A proporção de raio-X de tórax de rotina é uma variável quantitativa contínua que representa a razão do número de raio-X de tórax realizados com relação ao número de pacientes sem sinais ou sintomas de pneumonia, vezes 100.

- **Número de leitos**

O número de leitos é uma variável quantitativa discreta que representa o número médio de leitos no hospital durante o período de estudo.

- **Filiação a Escola de Medicina**

A filiação a escola de medicina é uma variável qualitativa ordinal onde o 1 significa que a escola tem filiação, e 2 que não tem.

- **Região**

A região é uma variáveis qualitativas nominal onde referese as regiões dos jospitais.

- **Média diária de pacientes**

O média diária de pacientes é uma variável quantitativa discreta que representa o número médio de pacientes no hospital por dia durante o período do estudo.

- **Número de enfermeiro(s)**

O número de enfermeiro(s) é uma variável quantitativa discreta que representa o Número médio de enfermeiros(as) de tempo-integral ou equivalente registrados e licenciados durante o período de estudo (número de tempos integrais+metade do número de tempo parcial).

- **Facilidades e serviços disponíveis**

A facilidades e serviços disponíveis é uma variável quantitativa contínua que representa a porcentagem de 35 potenciais facilidades e serviços que são fornecidos pelo hospital.

1.1.1.2 Estatisticass Descritivas

A tabela 2, mosta a estatistica descritivas das variáveis numéricas dos dados sem normalização, com efeito de mensuração diferentes como descrita em "Descrições de variáveis".

```
# datax2 <-datax  %>%
#   select(X5,X2,X4,X11)
# datax3 <-datax  %>%
#   select(X1,X3)
# datax1 <-datax  %>%
```

Tabela 2 – Medidas descritivas para boxplots

Variaveis	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Duração da Internação	6.700	8.340	9.420	9.648	10.470	19.560
Idade	38.80	50.90	53.20	53.23	56.20	65.90
Risco de Infecção	1.300	3.700	4.400	4.355	5.200	7.800
Proporção de Culturas de Rotina	1.60	8.40	14.10	15.79	20.30	60.50
Proporção de Raio-X de Tórax de Rotina	39.60	69.50	82.30	81.63	94.10	133.50
Número de leitos	29.0	106.0	186.0	252.2	312.0	835.0
Média diária de pacientes	20.0	68.0	143.0	191.4	252.0	791.0
Número de enfermeiro(s)	14.0	66.0	132.0	173.2	218.0	656.0
Facilidades e serviços disponíveis	5.70	31.40	42.90	43.16	54.30	80.00

```
#   select(X6,X9,X10)

# par(mfrow = c(1,3))
# boxplot(datax1)
# boxplot(datax2)
# boxplot(datax3)
boxplot(datax_ajusdet)
```

```
par(mfrow = c(1,2))
datax %>% select(X7) %>%  table(.) %>% barplot(xlab='X7')
datax %>% select(X8) %>%  table(.) %>% barplot(xlab='X8')
```

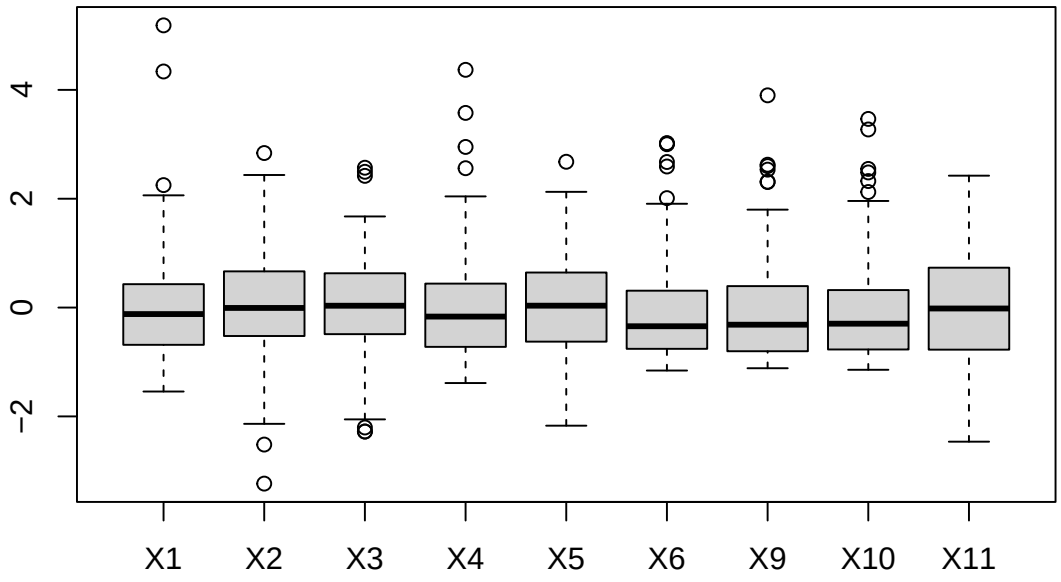
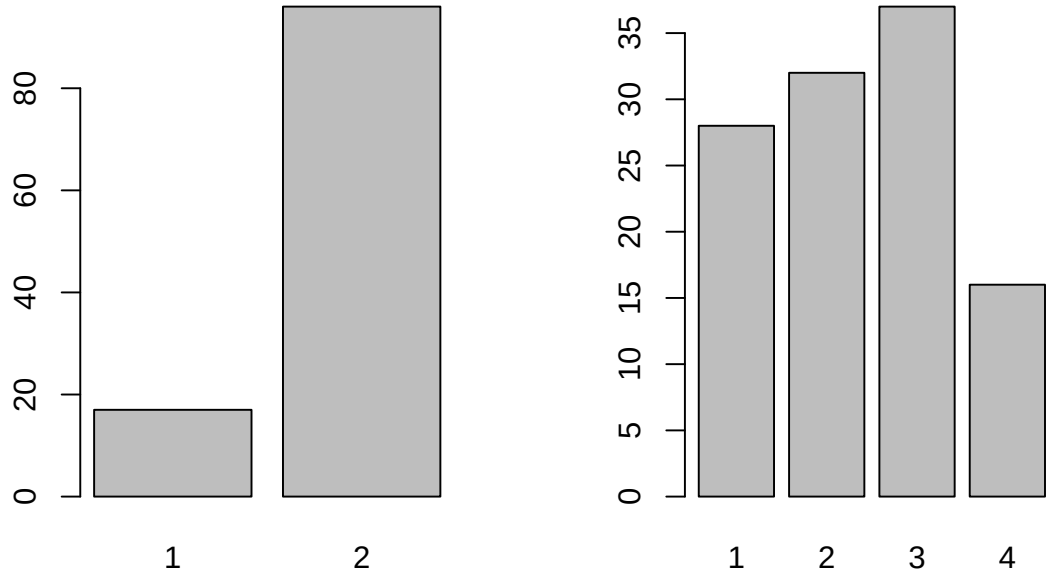


Figura 1 – Gráfico de box-plot das variaváveis dos dados.



1.1.1.3 Correlação entre as variáveis

Para verificar a natureza e a força da relação entre as variáveis e identificar lacunas e pontos discrepantes no conjunto de dados, utiliza-se a matriz de correlação aplicado no script a seguir.

```
library(ggcorrplot)
```

```
## Carregando pacotes exigidos: ggplot2
```

```
library(dplyr)

pmat = datax %>% select_if(is.numeric) %>%cor_pmat()

datax %>% select_if(is.numeric) %>% cor(.) %>%
  ggcorrplot( type = "lower", p.mat = pmat, hc.order = TRUE,lab = TRUE)
```

```
# k = datax %>% select_if(is.numeric) %>% summary()
```

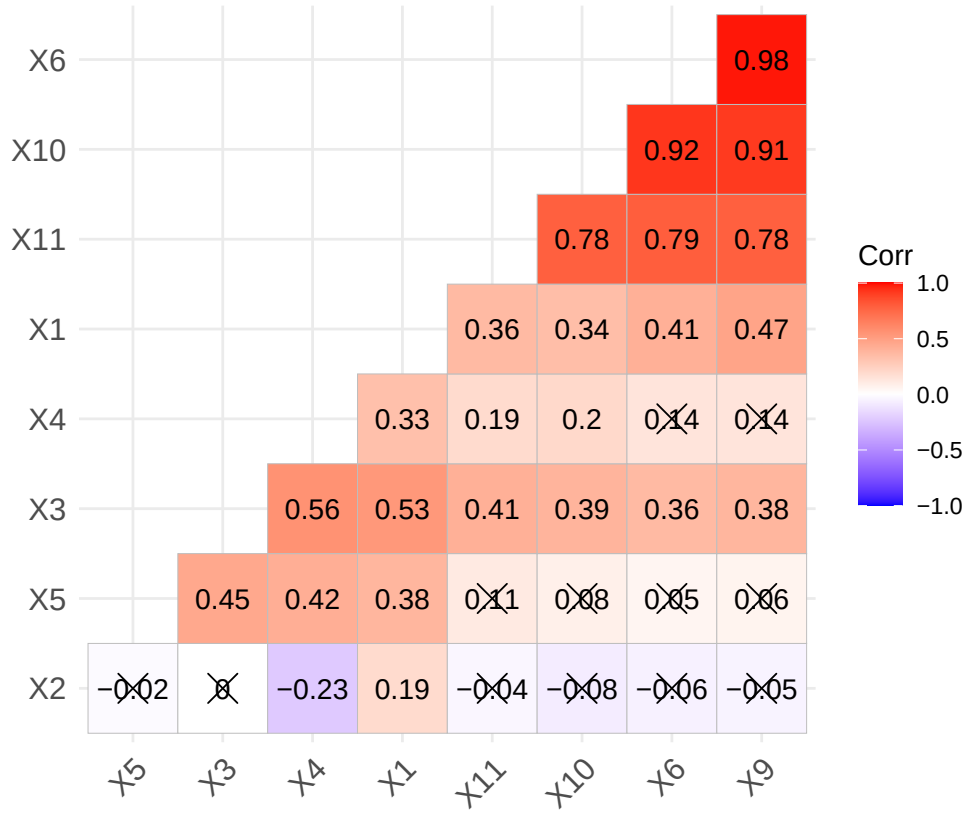


Figura 2 – Gráfico de calor da correlação entre as variaváveis dos dados.

1.2 Objetivo

1.2.1 Testes

Para efetuar um modelo, separa-se o banco em teste e treino no qual:

```
set.seed(10)
dados_train <- datax[sample(nrow(datax), 57, replace = F),] %>% data.frame()
dados_valid <- anti_join(datax, dados_train, by="ID") %>% data.frame()

# inbalanced data
table(dados_train$X8)
```

```
##
##  1  2  3  4
```

```
## 14 17 18  8
```

1.2.2 Número de enfermeira(o)s

```
library(plotly)
```

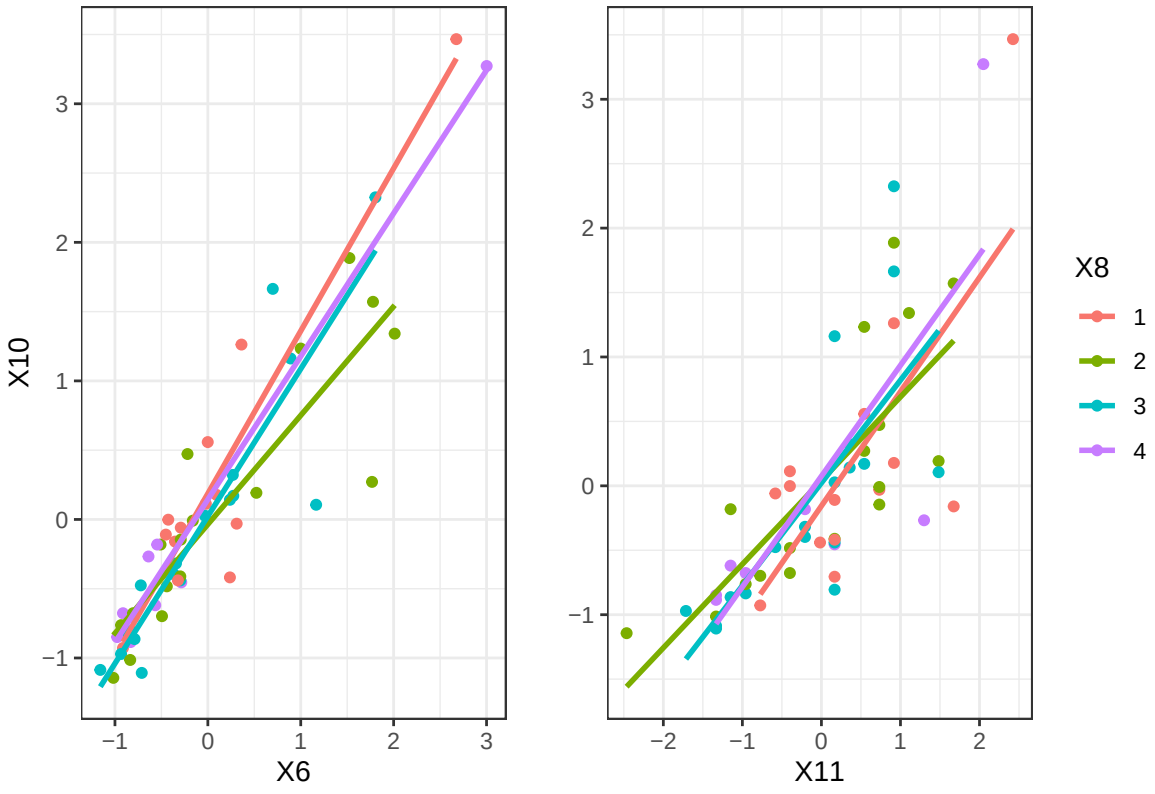
```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
require(gridExtra)
```

```
## Carregando pacotes exigidos: gridExtra

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
require(ggplot2)
library("patchwork")

g0<-ggplot(data = dados_train, aes(x=X6, X10, color = X8))+
  geom_point()+
  geom_smooth( method=lm, se=FALSE)+theme_bw()

g1<-ggplot(data = dados_train, aes(x=X11, X10, color = X8))+
  geom_point()+
  geom_smooth( method=lm, se=FALSE)+theme_bw()+ ylab("")

g0+g1+plot_layout(guides = "collect")
```

```
## `geom_smooth()` using formula 'y ~ x'

## `geom_smooth()` using formula 'y ~ x'
```



Espera-se que o número de enfermeira(o)s esteja relacionado às instalações e serviços disponíveis através de um modelo de segunda ordem. Suspeita-se também que varie segundo

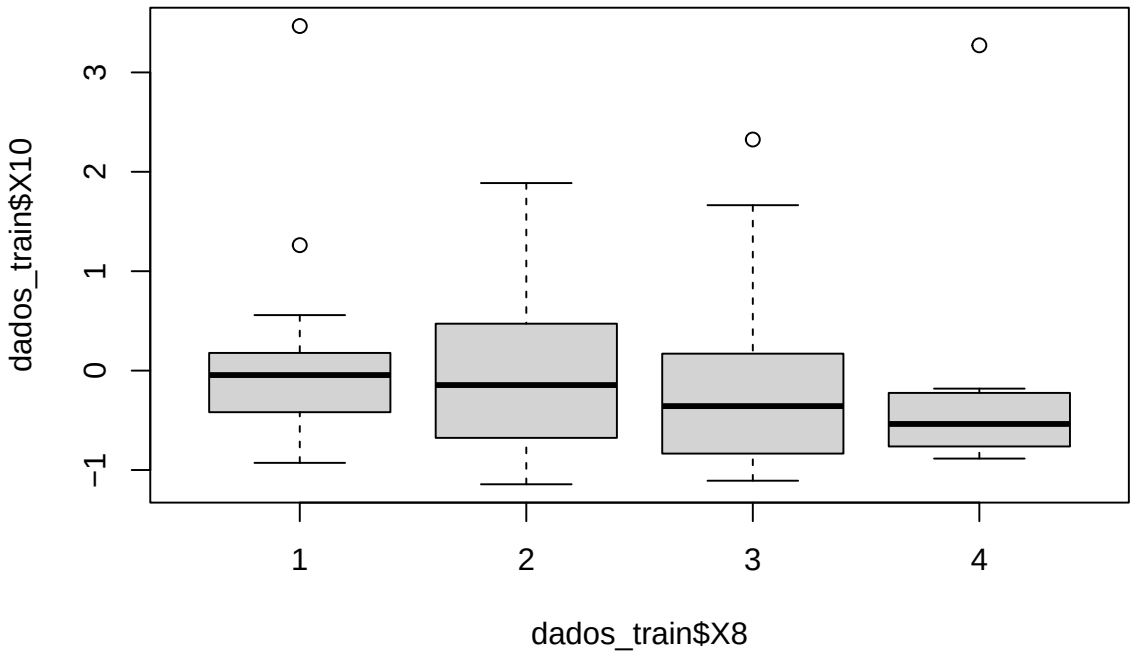
serviços disponíveis:X1,X4,X5,X6,X11

instalações:X7

região:X8

Deseja-se estudar se o número de enfermeira(o)s está relacionado às instalações, ou seja, os números de leitos do hospital, e se há diferenças entre os serviços disponíveis pelos hospitais. Neste caso, a variável resposta é o número de enfermeira(o)s e as duas outras variáveis são explicativas. Para isso, faz-se necessário a aplicação da regressão linear múltipla realizada no script a seguir:

```
boxplot(dados_train$X10~dados_train$X8)
```



Para um modelo inicial temos que

$$\hat{y} = \beta_0 + \beta_{X8}X8 + \beta_{X6}X6 + \beta_{X11}X11$$

analizando ANOVA do modelo, percebemos que o modelo tem apenas interação com $X6, X11$, e o resto das variáveis nao foram significantes

```
# Avaliando quais variaveis tem significância
summary(aov(X10 ~ X8*X6*X11*X7, data=dados_train))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## X8             3   0.73    0.24   1.932 0.1450
## X6             1  49.17   49.17 388.115 <2e-16 ***
## X11            1   0.31    0.31   2.433 0.1290
## X7             1   0.68    0.68   5.340 0.0277 *
## X8:X6          3   1.49    0.50   3.921 0.0175 *
## X8:X11         3   0.65    0.22   1.703 0.1868
## X6:X11         1   0.01    0.01   0.083 0.7756
## X8:X7          3   0.54    0.18   1.433 0.2519
## X6:X7          1   0.44    0.44   3.474 0.0719 .
## X11:X7         1   0.00    0.00   0.025 0.8752
## X8:X6:X11      3   0.16    0.05   0.411 0.7459
## X8:X6:X7       1   0.36    0.36   2.812 0.1036
## X8:X11:X7      1   0.00    0.00   0.017 0.8968
## X6:X11:X7      1   0.31    0.31   2.454 0.1274
## X8:X6:X11:X7   1   0.05    0.05   0.365 0.5501
## Residuals     31   3.93    0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

agora construindo um modelo de regressão linear com esta configuração temos que

```
modelo_inicial <- lm(X10 ~ X6+X7, data=dados_train)
summary(modelo_inicial)
```

```
##
## Call:
```

```
## lm(formula = X10 ~ X6 + X7, data = dados_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2383 -0.1577 -0.0397  0.1697  1.0820
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.39441    0.15677   2.516   0.0149 *
## X6           0.86803    0.07152  12.137   <2e-16 ***
## X72         -0.41918    0.17969  -2.333   0.0234 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3995 on 54 degrees of freedom
## Multiple R-squared:  0.8535, Adjusted R-squared:  0.8481
## F-statistic: 157.3 on 2 and 54 DF,  p-value: < 2.2e-16
```

com valor do F-statistics, percebe-se que o teste de regressão é significativo, indicando que há regressão nesses dados, e analizando o modelo, x11 e x6 não tem diferença significativa, podendo descartar acabando com um modelo do tipo

```
modelo_inicial <- lm(X10 ~  X6+X7, data=dados_train)
summary(modelo_inicial)
```

```
##
## Call:
## lm(formula = X10 ~ X6 + X7, data = dados_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2383 -0.1577 -0.0397  0.1697  1.0820
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.39441    0.15677   2.516   0.0149 *
```

```
## X6           0.86803    0.07152  12.137   <2e-16 ***
## X72         -0.41918    0.17969  -2.333   0.0234 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3995 on 54 degrees of freedom
## Multiple R-squared:  0.8535, Adjusted R-squared:  0.8481
## F-statistic: 157.3 on 2 and 54 DF,  p-value: < 2.2e-16
```

agora avalindo este modelo temos que o erro medio das previsões é baixo e o R2 no banco de teste é alto, assim sendo um bom modelo para começar e avaliar com as suposições do hospital

```
library(caret)
```

```
## Carregando pacotes exigidos: lattice
```

```
# predições
predictions <- modelo_inicial %>% predict(dados_valid)
data.frame(
  RMSE = RMSE(predictions, dados_valid$X10),
  R2 = R2(predictions, dados_valid$X10)
  )
```

```
##        RMSE        R2
## 1 0.4212077 0.8287177
```

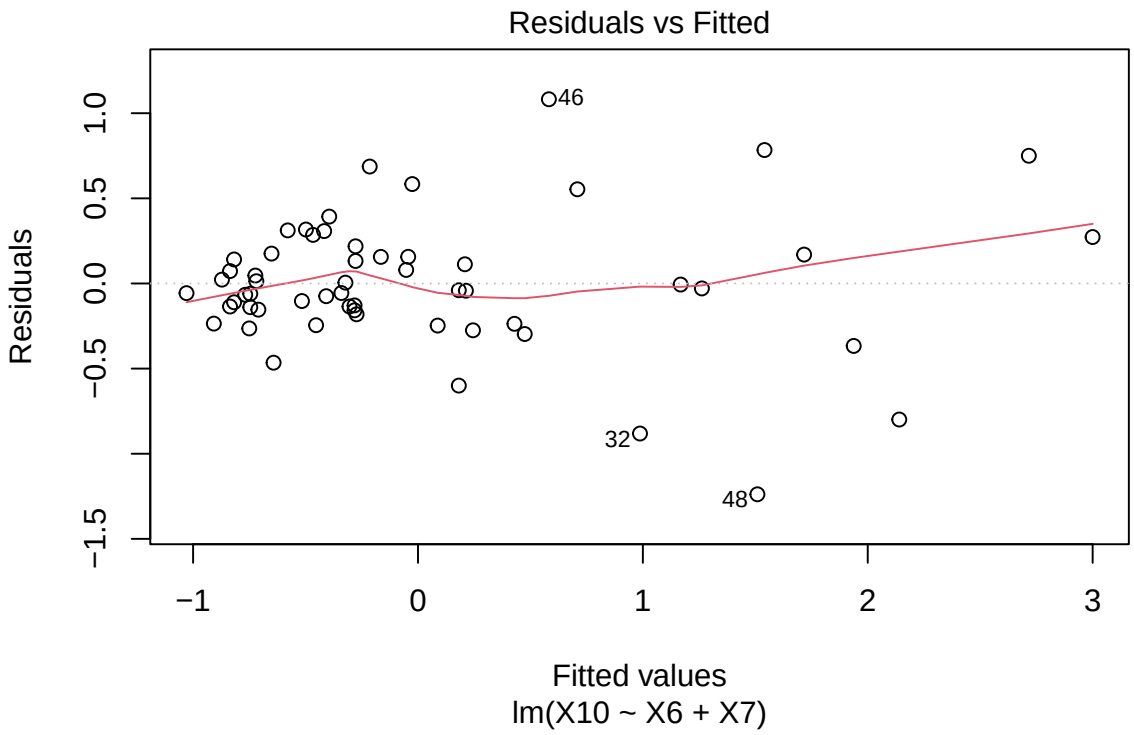
```
# Teste de multicolinearidade Gif (>1 indica multicolinearidade)
car::vif(modelo_inicial)
```

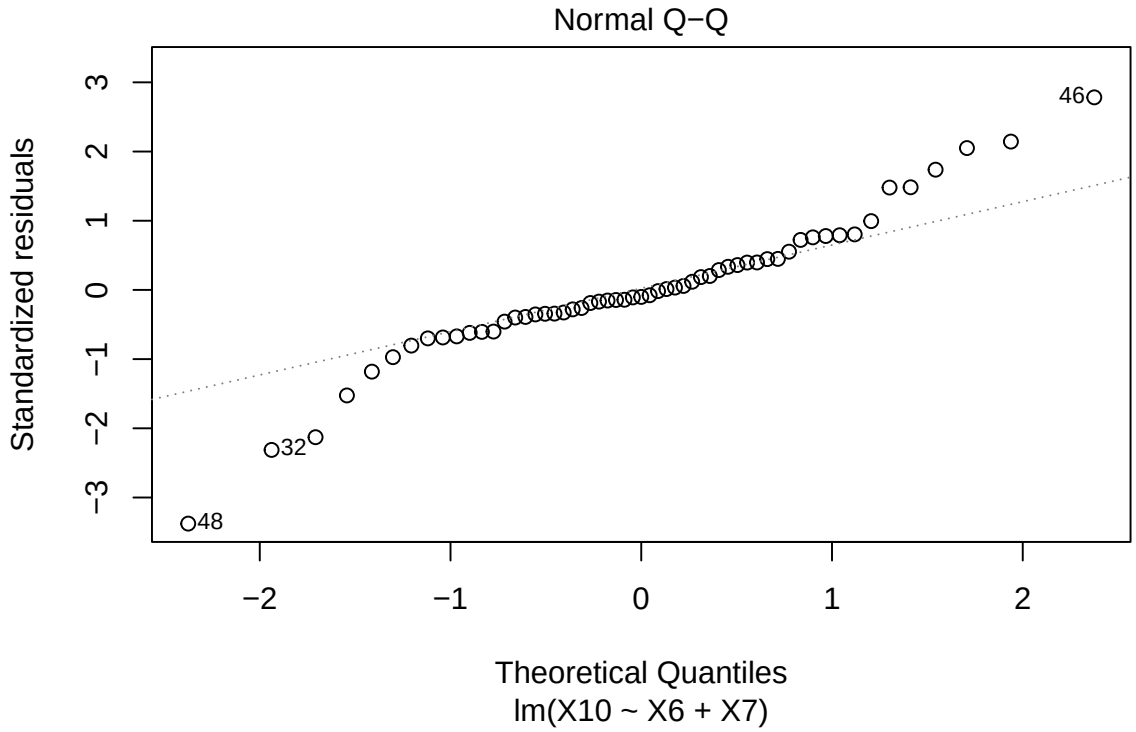
```
##       X6       X7
## 1.668569 1.668569
```

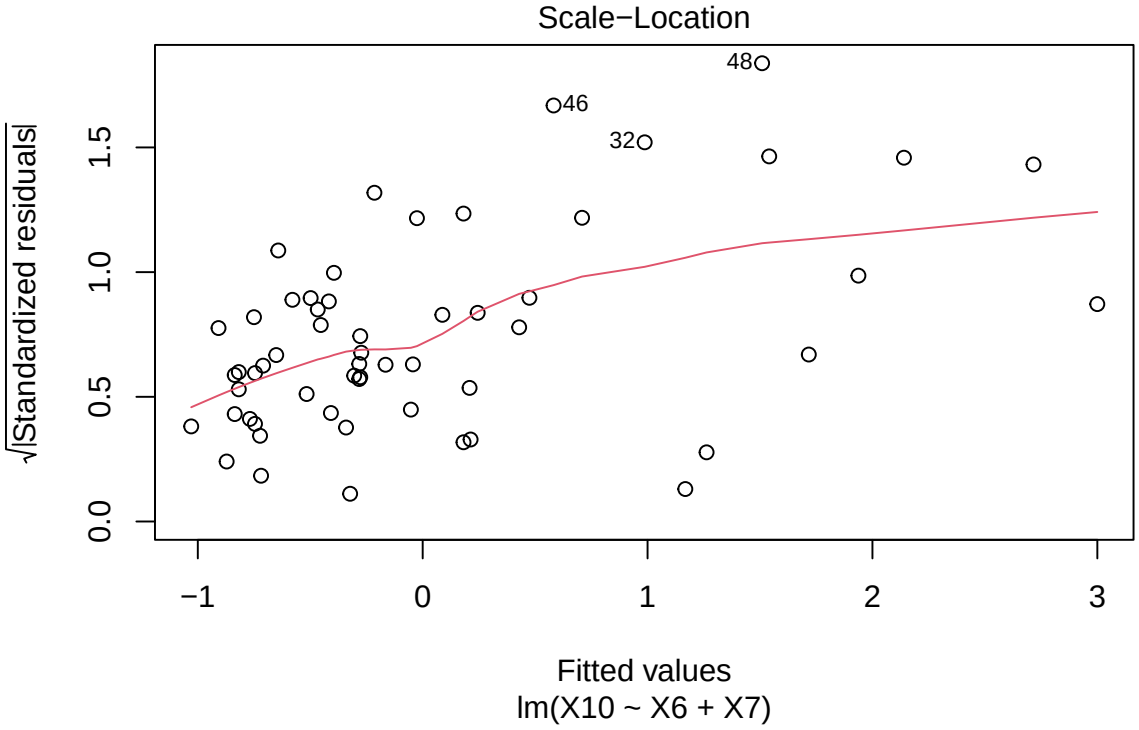
```
shapiro.test(modelo_inicial$residuals)
```

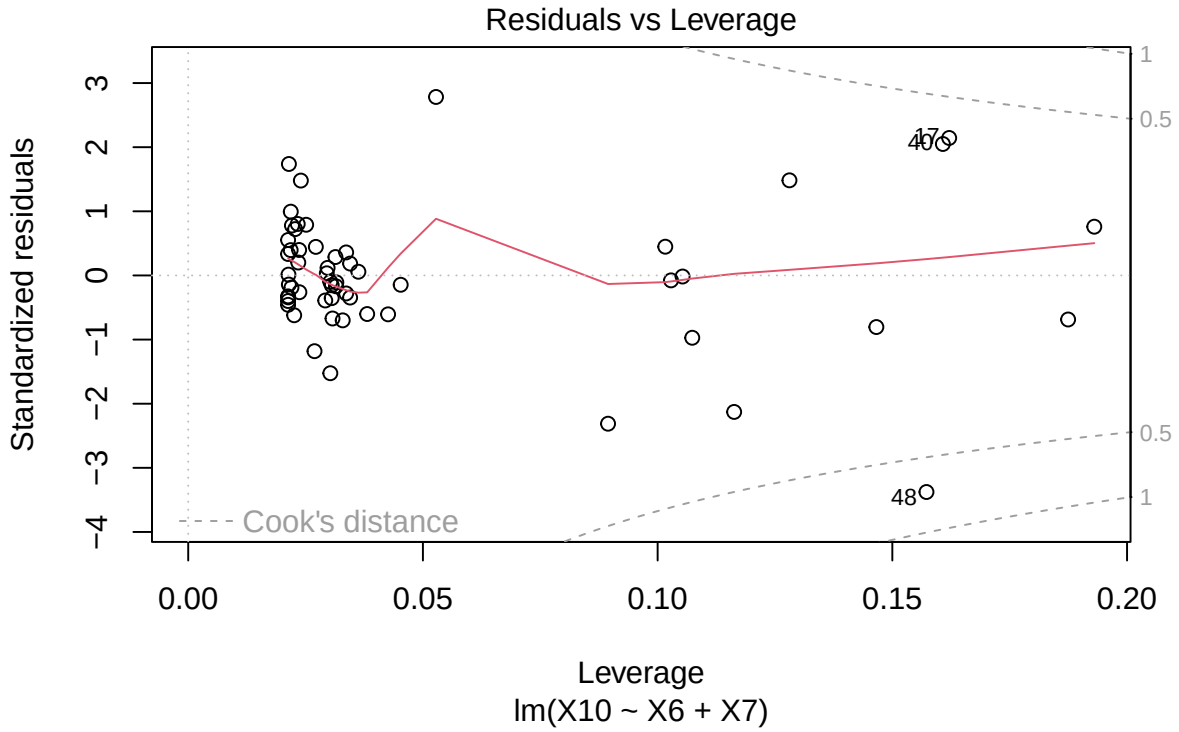
```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_inicial$residuals
## W = 0.95214, p-value = 0.02461
```

```
plot(modelo_inicial)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(X10 ~ X6 + X7)
46
32
48


Scale−Location
√|Standardized residuals|
Fitted values
lm(X10 ~ X6 + X7)
48
46
32



Agora avaliando através do steepwise, temos que o modelo que converge sobre o uso de mais variaveis

```
modelo_inicial <- lm(X10 ~ X6+X7, data=dados_train)
summary(modelo_inicial)
```

```
##
## Call:
## lm(formula = X10 ~ X6 + X7, data = dados_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2383 -0.1577 -0.0397  0.1697  1.0820
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.39441    0.15677   2.516   0.0149 *
## X6           0.86803    0.07152  12.137   <2e-16 ***
```

```
## X72         -0.41918    0.17969  -2.333   0.0234 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3995 on 54 degrees of freedom
## Multiple R-squared:  0.8535, Adjusted R-squared:  0.8481
## F-statistic: 157.3 on 2 and 54 DF,  p-value: < 2.2e-16
```

```
modmin<-lm(X10 ~ 1, data=dados_train)
modcompl<-lm(X10~ X1+X2+X3+X4+X5+X6+X8+X9+X11, data=dados_train)

modfim  <- step(modmin, scope=list(lower=modmin, upper=modcompl), direction="both",dat
```

```
## Start:  AIC=3.8
## X10 ~ 1
##
##        Df Sum of Sq    RSS      AIC
## + X9    1    49.718  9.109 -100.527
## + X6    1    49.342  9.485  -98.221
## + X11   1    33.638 25.189  -42.549
## + X1    1    11.306 47.521   -6.367
## + X3    1    10.532 48.294   -5.447
## + X4    1     3.608 55.218    2.190
## <none>              58.827    3.798
## + X5    1     1.534 57.293    4.292
## + X2    1     0.002 58.825    5.796
## + X8    3     0.734 58.092    9.082
##
## Step:  AIC=-100.53
## X10 ~ X9
##
##        Df Sum of Sq    RSS      AIC
## + X11   1     0.476  8.633 -101.587
## + X4    1     0.391  8.718 -101.029
## + X5    1     0.377  8.732 -100.934
## + X3    1     0.326  8.783 -100.605
```

```
## <none>                9.109 -100.527
## + X8    3     0.871  8.237 -100.259
## + X6    1     0.256  8.853 -100.152
## + X2    1     0.053  9.056  -98.858
## + X1    1     0.004  9.105  -98.551
## - X9    1    49.718 58.827    3.798
##
## Step:  AIC=-101.59
## X10 ~ X9 + X11
##
##        Df Sum of Sq     RSS      AIC
## <none>                8.6327 -101.587
## + X5    1    0.2891  8.3436 -101.528
## + X4    1    0.2513  8.3814 -101.271
## + X3    1    0.1887  8.4440 -100.847
## + X6    1    0.1833  8.4494 -100.811
## - X11   1    0.4761  9.1088 -100.527
## + X8    3    0.6907  7.9420 -100.340
## + X2    1    0.0390  8.5937  -99.846
## + X1    1    0.0033  8.6294  -99.609
## - X9    1   16.5563 25.1890  -42.549
```

```
summary(modfim)
```

```
##
## Call:
## lm(formula = X10 ~ X9 + X11, data = dados_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30613 -0.16346 -0.01179  0.13399  0.80067
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04952    0.05328   0.929   0.3568
## X9           0.87702    0.08618  10.177 3.67e-14 ***
```

```
## X11          0.13898    0.08053   1.726   0.0901 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3998 on 54 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8478
## F-statistic:   157 on 2 and 54 DF,  p-value: < 2.2e-16
```

agora com o teste linear geral, temos que existe diferença significatifva entre os modelos e acabamos com um moelo mais parcimanioso sem multicolineariade que é o caso do modtest

```
anova(modelo_inicial,modfim) # modelo 2 é melhor
```

```
## Analysis of Variance Table
##
## Model 1: X10 ~ X6 + X7
## Model 2: X10 ~ X9 + X11
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     54 8.6165
## 2     54 8.6327  0 -0.016185
```

```
AIC(modelo_inicial,modfim) # quanto menoor melhor
```

```
##                df      AIC
## modelo_inicial  4 62.06488
## modfim          4 62.17184
```

```
BIC(modelo_inicial,modfim)
```

```
##                df      BIC
## modelo_inicial  4 70.23708
## modfim          4 70.34405
```

```
car::vif(modfim)
```

```
##       X9      X11
## 2.357255 2.357255
```

```
shapiro.test(modfim$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modfim$residuals
## W = 0.92676, p-value = 0.001982
```

```
modtest<- lm(X10 ~ X8+X4+X6, data=dados_train)
car::vif(modtest)
```

```
##        GVIF Df GVIF^(1/(2*Df))
## X8 1.150084  3        1.023579
## X4 1.179353  1        1.085980
## X6 1.056460  1        1.027842
```

```
anova(modfim,modtest)
```

```
## Analysis of Variance Table
##
## Model 1: X10 ~ X9 + X11
## Model 2: X10 ~ X8 + X4 + X6
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     54 8.6327
## 2     51 8.7650  3  -0.13234
```

```
predictions <- modtest %>% predict(dados_valid)
data.frame(
  RMSE = RMSE(predictions, dados_valid$X10),
```

```
  R2 = R2(predictions, dados_valid$X10)
  )
```

```
##        RMSE        R2
## 1 0.4026966 0.8498328
```

```
predictions <- modfim %>% predict(dados_valid)
data.frame(
  RMSE = RMSE(predictions, dados_valid$X10),
  R2 = R2(predictions, dados_valid$X10)
  )
```

```
##        RMSE        R2
## 1 0.4319206 0.8343771
```

1.2.3 Duração da internação

A duração da internação está associada a características do paciente, seu tratamento e do hospital

características do paciente:X2, seu tratamento:X4,X5 hospital:X3,X6,X7,X9,X10, X11

Deseja-se estudar se a Duração da internação está associada a características do paciente, seu tratamento e do hospital, ou seja, a duração da internação, e se há diferenças entre os serviços disponíveis pelos hospitais. Neste caso, a variável resposta é o número de enfermeira(o)s e as duas outras variáveis são explicativas. Para isso, faz-se necessário a aplicação da regressão linear múltipla realizada no script a seguir:

Referências

Anexos

ANEXO A – Amostra