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 Objetivos

A fim de estudar sobre a duração da internação nos hospitais dos Estados Unidos no período de 1975-1976, foi retirada uma amostra aleatória de 113 hospitais selecionados entre 338 pesquisados, para isso foram propostas as seguintes hipóteses:

A primeira é verificar 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. Além de verificar se a mesma variável resposta mencionada anteriormente varia segundo a região.

Já a segunda é verificar se a duração da internação está associada a características do paciente, seu tratamento e do hospital.

1.1.2 Metodologia

O programa utilizado para analisar os dados disponibilizados em Excel será o R Studio, versão 4.2.0. Para uma primeira visualização dos dados, necessita-se identificar e realizar a análise descritiva das variáveis, portanto os dados estão organizados e classificados da seguinte maneira:

```
# Tabela de nomes X1: Nome variavel

Nome <- names(data)

Código <- names(datax)

Descrição <- c('1-113', 'Duração média da internação de todos os pacientes no hospital

Classificação <- c('Qualitativa ordinal', 'Quantitativa contínua', 'Quantitativa cont
```

```
library(knitr)
knitr::kable(cbind(Nome,Código,Descrição, Classificação),
             caption = 'Descrição dos códigos da tabela com a seguinte indentificação
```

```
## Warning in cbind(Nome, Código, Descrição, Classificação): number of rows of
## result is not a multiple of vector length (arg 3)
```

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

Nome	Código	Descrição	Classificação
Número de Identificação	ID	1-113	Qualitativa ordinal
Duração da Internação	X1	Duração média da internação de todos os pacientes no hospital (em dias)	Quantitativa contínua
Idade	X2	Idade média dos pacientes	Quantitativa contínua
Risco de Infecção	X3	Probabilidade média estimada de adquirir infecção no hospital (em %)	Quantitativa contínua
Proporção de Culturas de Rotina	X4	Razão do número de culturas realizadas com relação ao número de pacientes sem sinais ou sintomas de infeção adquirida no hospital, vezes 100.	Quantitativa contínua
Proporção de Raio-X de Tórax de Rotina	X5	Razão do número de Raio-X de Tórax realizados com relação ao número de pacientes sem sinais ou sintomas de pneumonia, vezes 100.	Quantitativa contínua
Número de leitos	X6	Número médio de leitos no hospital durante o período de estudo	Quantitativa contínua

Nome	Código	Descrição	Classificação
Filiação a Escola de Medicina	X7	1 – sim 2 – não	Qualitativa ordinal
Região	X8	Região Geográfica, onde: 1 – NE 2- NC 3 – S e 4 – W	Qualitativa nominal
Média diária de pacientes	X9	Número médio de pacientes no hospital por dia durante o período do estudo	Quantitativa contínua
Número de enfermeiro(s)	X10	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)	Quantitativa contínua
Facilidades e serviços disponíveis	X11	% de 35 potenciais facilidades e serviços que são fornecidos pelo hospital	Quantitativa contínua
NA	X1adj	j-113	Qualitativa ordinal
NA	X2adj	Duração média da internação de todos os pacientes no hospital (em dias)	Quantitativa contínua
NA	X3adj	Idade média dos pacientes	Quantitativa contínua
NA	X4adj	Probabilidade média estimada de adquirir infecção no hospital (em %)	Quantitativa contínua
NA	X5adj	Razão do número de culturas realizadas com relação ao número de pacientes sem sinais ou sintomas de infeção adquirida no hospital, vezes 100.	Quantitativa contínua
NA	X6adj	Razão do número de Raio-X de Tórax realizados com relação ao número de pacientes sem sinais ou sintomas de pneumonia, vezes 100.	Quantitativa contínua

Nome	Código	Descrição	Classificação
NA	X9adj	Número médio de leitos no hospital durante o período de estudo	Quantitativa contínua
NA	X10adj	1 – sim 2 – não	Qualitativa ordinal
NA	X11adj	Região Geográfica, onde: 1 – NE 2- NC 3 – S e 4 – W	Qualitativa nominal

As etapas para o estudo da internação dos hospitais foram separadas em duas maneiras, a primeira é a construção e a segunda é a validação do modelo. Para a primeira etapa, foi selecionada uma amostra aleatória simples com 57 observações, para a segundo ficou o restante das observações que compõe o banco. Para as duas hipóteses procura-se um modelo regressivo linear múltiplo do tipo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + e_i \ , \forall \ i = 1, \ldots, \ n$$

Onde tem-se,

- Y_{ij}

  - variável resposta;

- $X_{i1}, X_{i2}, \ldots, X_{ik}$

  - k variáveis explicativas ou independentes;

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$

  - parâmetros do modelo;

- e_i

  - são independentes e

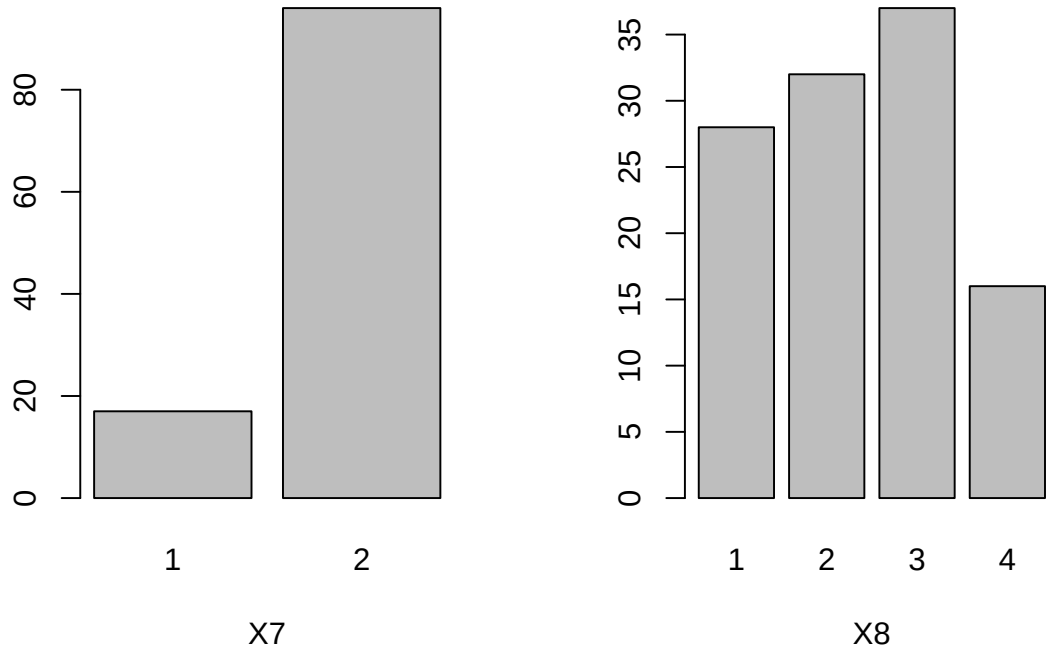
$$N(0, \sigma^2)$$

.

Para a primeira hipótese, define-se como modelo I aquele que relaciona a variável resposta, Número de enfermeiro(s) (X10), com as variáveis explicativas, instalações (X6), serviços disponíveis pelos hospitais (X11) e a região (X8).

Já o modelo II é definido como aquele que relaciona a variável resposta, Duração da internação (X1), com as variáveis explicativas, a características do paciente (X2), seu tratamento (X4 e X5) e do hospital (X3).

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



1.2 Resultado

Realizando uma breve análise descritiva das variáveis quantitativas, tem-se o boxplot com os dados normalizados:

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

```
# datax2 <-datax  %>%
#   select(X5,X2,X4,X11)
# datax3 <-datax  %>%
#   select(X1,X3)
# datax1 <-datax  %>%
#   select(X6,X9,X10)
# par(mfrow = c(1,3))
# boxplot(datax1)
# boxplot(datax2)
# boxplot(datax3)
boxplot(datax_ajusdet)
```

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.

```
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)
```

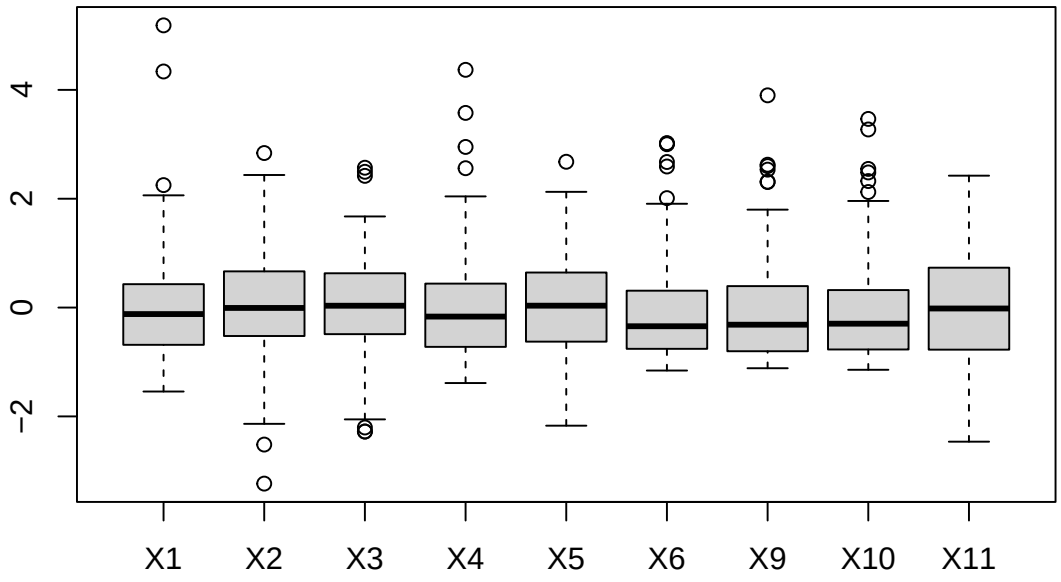


Figura 1 – Gráfico de box-plot das variaváveis dos dados.

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

Analisando o gráfico acima, tem-se que as variáveis que estão nas três extremidades externas dos dois eixos apresentam uma correlação forte, então, X10 com X11, X6 com X11 e X10 e X9 com X11, X10 e X6. A maior correlação é apresentada entre as variáveis X6 e X9, que é o número de leitos e a média diária de pacientes, respectivamente.

```
boxplot(datax)
```

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

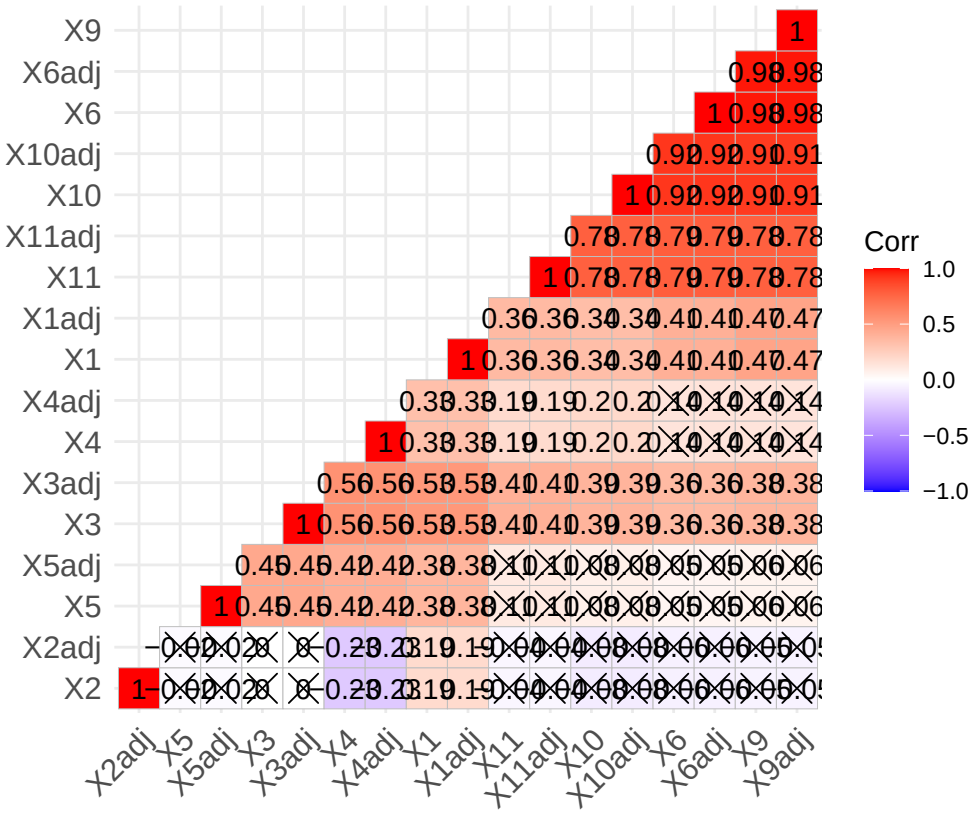
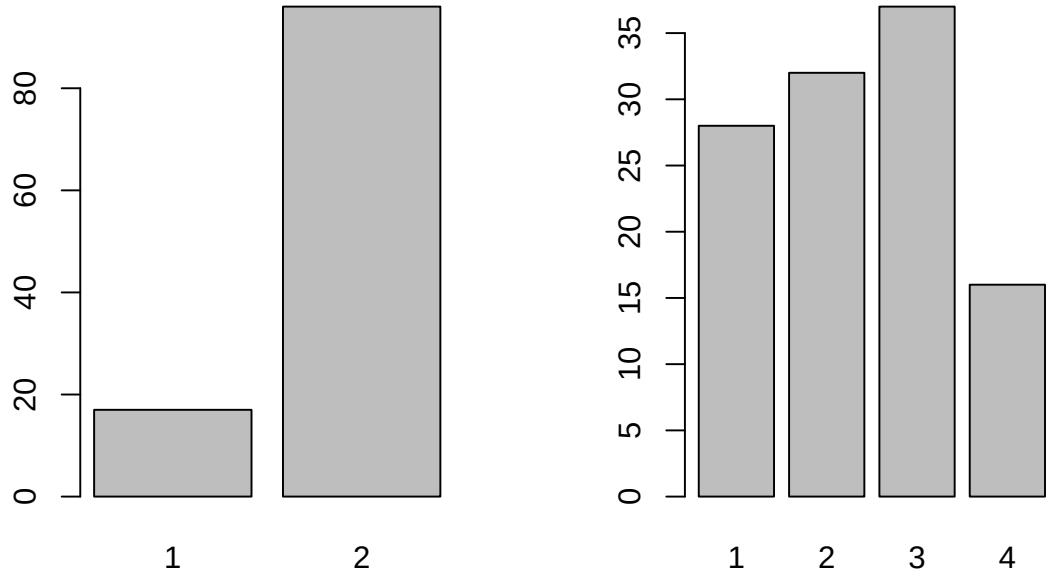


Figura 2 – Gráfico de calor da correlação entre as variaváveis dos dados.



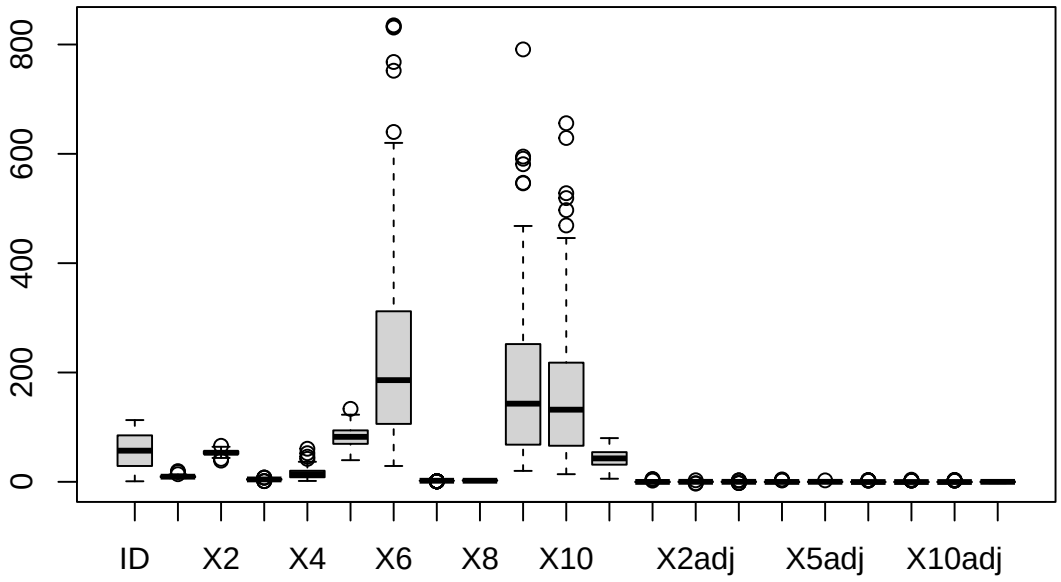


Figura 3 – Gráfico de box-plot das variaváveis dos dados.

1.2.0.1 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)
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)
```

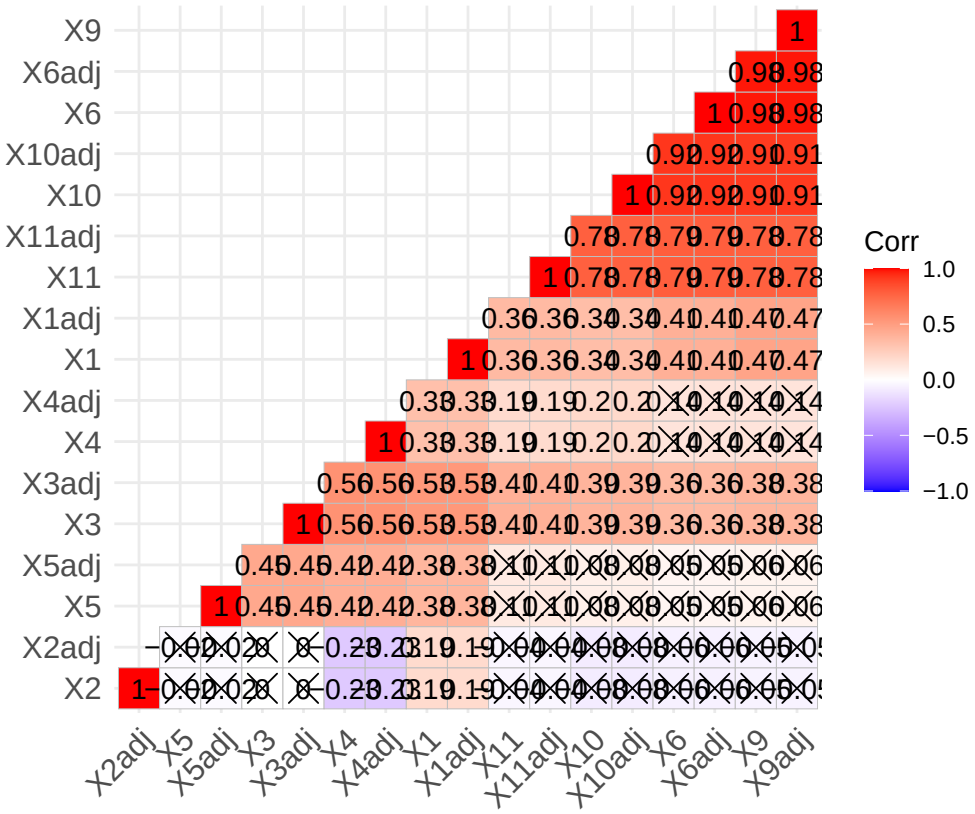


Figura 4 – Gráfico de calor da correlação entre as variaváveis dos dados.

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

1.3 Objetivo

1.3.1 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
```

1.3.2 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
```

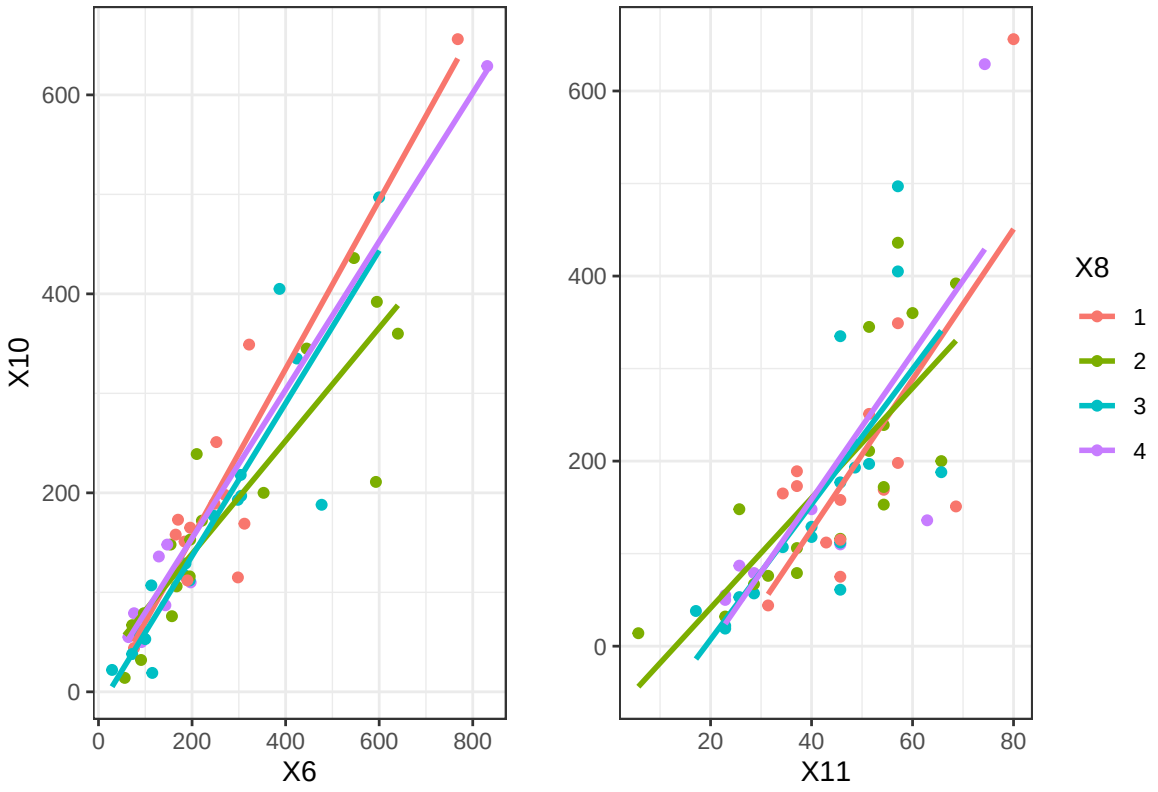
```
##
## Attaching package: 'gridExtra'
```

```
## 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'
```




```
# Avaliando quais variaveis tem significância
# library("tidyverse")
# library("repurrrsive")
# summary(aov(X10 ~ X8*X6*X11*X7, data=dados_train))
# lista <- map_df(, extract, c("Df", "Sum Sq", "Mean Sq", "F value", "Pr(>F)"))

#
# knitr::kable(teste)
```

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,X9,X11

instalações:X7

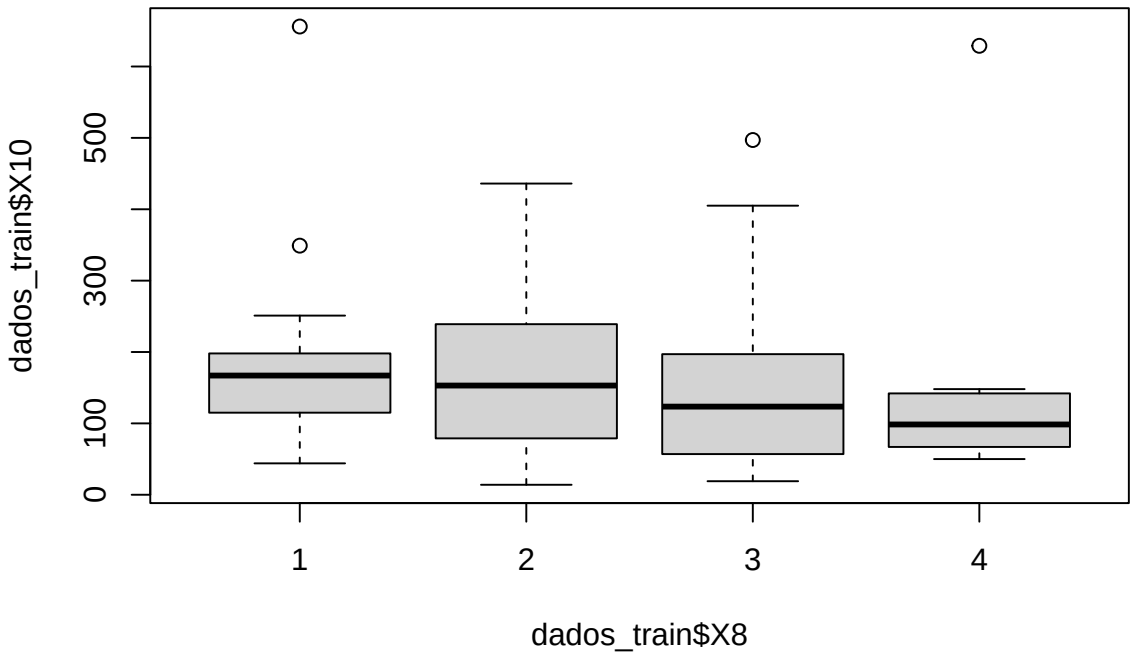
região:X8

\ 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. No qual avaliando o gráfico da dispersão de ordem da variável região $X8$ e o número de enfermeiros $X10$, verifica-se que não possui diferença significatíva na dispersão destes valores.

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




```
summary(aov(dados_train$X10~dados_train$X8))
```

```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## dados_train$X8  3   14240    4747   0.223   0.88
## Residuals      53 1126692   21258
```

1.3.2.1 Pressupostos para um modelo inicial

Agora presumindo um modelo inicial para explicar a variável de número de enfermeiros $X10$ é dada por

$$\hat{y}_{X10} = \beta_0 + \beta_{X1}X1 + \beta_{X6}X6 + \beta_{X8}X8 + \beta_{X11}X11 + \beta_{X1,X8}(X1X8) + \beta_{X6,X8}(X6X8) + \beta_{X7,X8}(X7X8) + \beta_{X11,X8}(X11X8)$$

no qual presume que o modelo é explicado pela "duração da internação" ($X1$), "Número de leitos" ($X6$), "Facilidades e serviços disponiveis" ($X11$) com a "Região".

```
# Avaliando quais variaveis tem significância
summary(aov(X10 ~ X1adj*X8+X6adj*X8+X11adj*X8+X7*X8, data=dados_train))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## X1adj        1 219271  219271 104.325 2.57e-12 ***
## X8           3  22596    7532   3.584   0.0227 *
## X6adj        1 727220  727220 345.997  < 2e-16 ***
## X11adj       1   6463    6463   3.075   0.0878 .
## X7           1  13251   13251   6.304   0.0165 *
## X1adj:X8     3  29331    9777   4.652   0.0074 **
## X8:X6adj     3  15888    5296   2.520   0.0729 .
## X8:X11adj    3   9872    3291   1.566   0.2141
## X8:X7        3  19275    6425   3.057   0.0402 *
## Residuals   37  77767    2102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

agora os resultados obtidos pela anova, temos que pelos testes, deu significativo as variáveis explicativas sem interação e a interação com da região $X8$ com a variável $X1$ e as outras variáveis foram descartadas por estar perto do limite do p-value 0.05.

Agora construindo um novo modelo de regressão

$$\hat{y}_{X11} = \beta_0 + \beta_{X1}X1 + \beta_{X6}X6 + \beta_{X7}X7 + \beta_{X8}X8 + \beta_{X1,X8}(X1X8)$$

temos que

```
table(dados_train$X8)
```

```
##
##  1  2  3  4
## 14 17 18  8
```

```
modelo_inicial <- lm(X10 ~ X1adj*X8 + X6adj +X7*X8, data=dados_train)
summary(modelo_inicial)
```

```
##
## Call:
## lm(formula = X10 ~ X1adj * X8 + X6adj + X7 * X8, data = dados_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -141.234  -26.732   -2.642   34.468  112.310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  252.530     29.189   8.652 4.77e-11 ***
## X1adj          2.783     18.516   0.150   0.8812
## X82          -50.619     41.170  -1.230   0.2254
## X83          -93.827     63.641  -1.474   0.1475
## X84           11.293     67.563   0.167   0.8680
## X6adj        121.208     11.440  10.595 1.09e-13 ***
## X72          -76.374     32.815  -2.327   0.0246 *
## X1adj:X82    -30.166     25.681  -1.175   0.2465
## X1adj:X83     71.317     29.406   2.425   0.0195 *
## X1adj:X84      2.300     41.757   0.055   0.9563
## X82:X72       30.403     44.952   0.676   0.5024
## X83:X72      115.530     69.311   1.667   0.1027
## X84:X72       -5.958     82.334  -0.072   0.9426
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.48 on 44 degrees of freedom
## Multiple R-squared:  0.8938, Adjusted R-squared:  0.8648
## F-statistic: 30.85 on 12 and 44 DF,  p-value: < 2.2e-16
```

com valor do F-statistics, para o teste linear geral, percebe-se que o teste de regressão é significativo, indicando que há regressão nesses dados, e analizando o modelo, apenas x6 tem diferenças significativas, podendo descartar acabando com um modelo do tipo, no qual rejeitamos a normalidade, assim transformando a variável através do boxcox

```
modelo_inicial <- lm(X10 ~  X6adj+X7, data=dados_train)
summary(modelo_inicial)
```

```
##
## Call:
## lm(formula = X10 ~ X6adj + X7, data = dados_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -172.453  -21.961   -5.528   23.631  150.681
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   228.18      21.83  10.451  1.4e-14 ***
## X6adj         120.89       9.96  12.137  < 2e-16 ***
## X72           -58.38      25.02  -2.333   0.0234 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.63 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
```

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

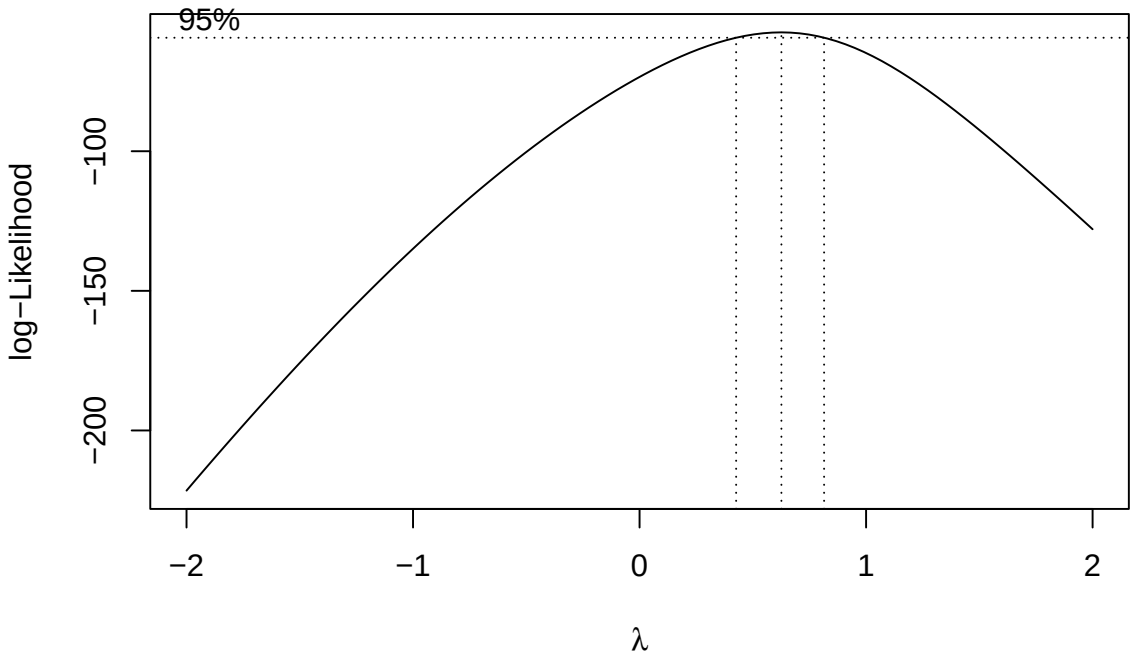
```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_inicial$residuals
## W = 0.95214, p-value = 0.02461
```

como foi rejeitada o teste de normalidade, utilizamos uma transformação boxcox para criar o novo modelo, onde seque se que

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:patchwork':
##
##     area

## The following object is masked from 'package:plotly':
##
##     select

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

```
k<-boxcox(modelo_inicial)
```




```
lambda<- k$x[which.max(k$y)]

dados_train['X10_cox'] <- (dados_train$X10^lambda-1)/lambda

modelo_inicial_cox <- lm(X10_cox ~X6adj+X7, data=dados_train)
summary(modelo_inicial_cox)
```

```
##
## Call:
## lm(formula = X10_cox ~ X6adj + X7, data = dados_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.7865  -3.3753  -0.5417   4.6651  19.6757
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   42.373      3.100   13.67  < 2e-16 ***
## X6adj         16.355      1.414   11.56 3.11e-16 ***
## X72           -6.501      3.553   -1.83   0.0728 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.899 on 54 degrees of freedom
## Multiple R-squared:  0.8351, Adjusted R-squared:  0.829
## F-statistic: 136.7 on 2 and 54 DF,  p-value: < 2.2e-16
```

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

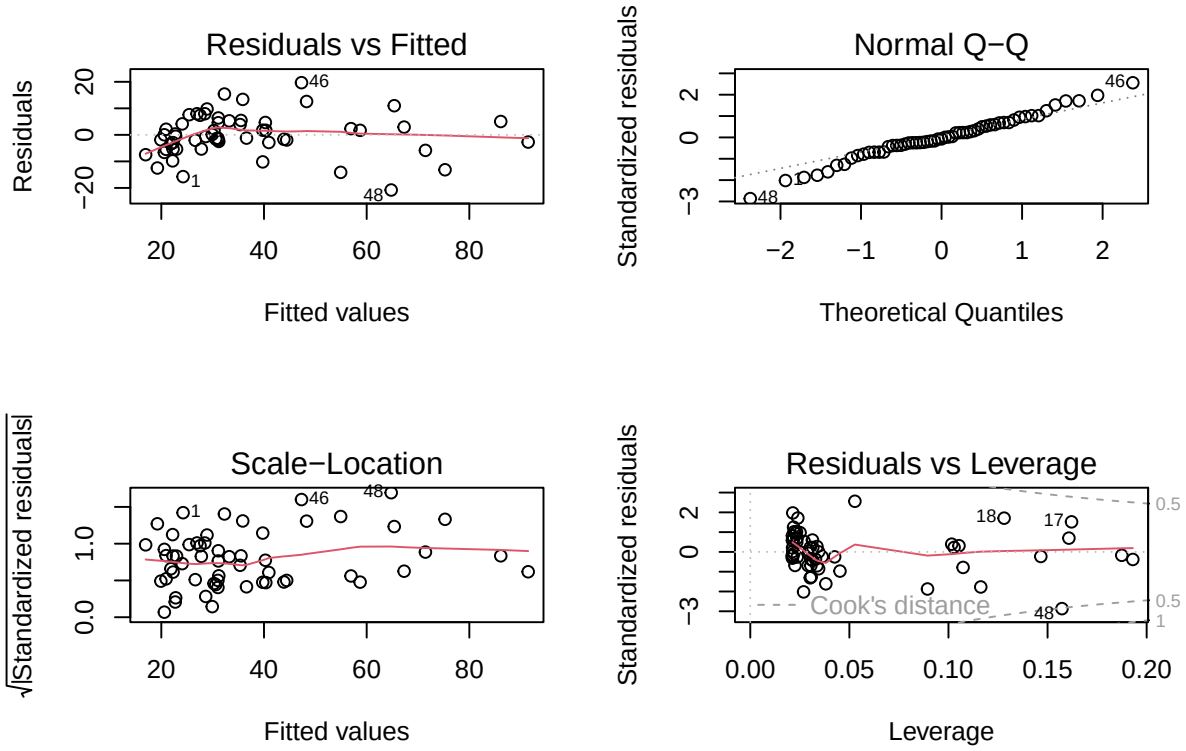
```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_inicial_cox$residuals
## W = 0.98839, p-value = 0.8595
```

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

```
require(MASS)
library(caret)
```

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

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

par(mfrow=c(2,2))
plot(modelo_inicial_cox)
```



Retirando os outliers temos que

```
modelo_inicial_cox <- lm(X10_cox ~  X6adj+X7, data=dados_train[-c(18,48,46),])
summary(modelo_inicial_cox)
```

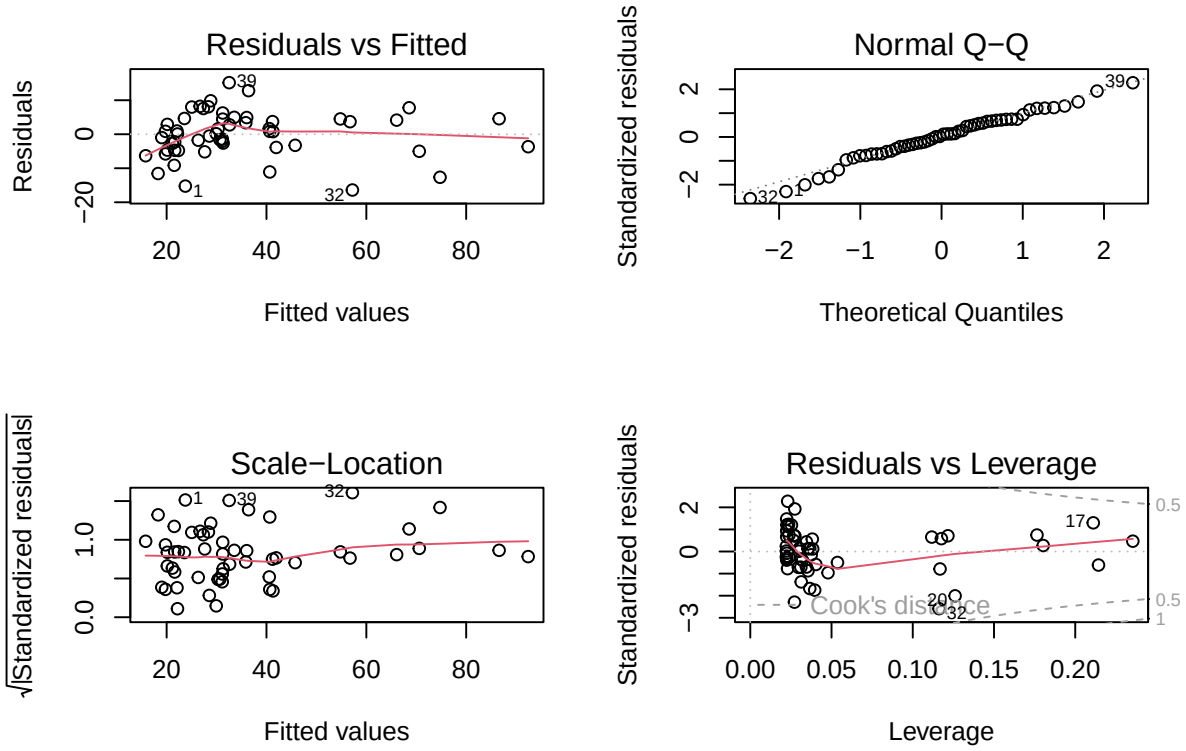
```
##
## Call:
## lm(formula = X10_cox ~ X6adj + X7, data = dados_train[-c(18,
##     48, 46), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.397  -3.828   0.453   4.488  15.163
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.912      2.943  13.220   <2e-16 ***
## X6adj         17.826      1.355  13.158   <2e-16 ***
```

```
## X72           -2.486      3.430  -0.725    0.472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.745 on 51 degrees of freedom
## Multiple R-squared:  0.8767, Adjusted R-squared:  0.8718
## F-statistic: 181.3 on 2 and 51 DF,  p-value: < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_inicial_cox$residuals
## W = 0.98291, p-value = 0.6322
```

```
par(mfrow=c(2,2))
plot(modelo_inicial_cox)
```



Agora avalindo o modelo no banco de teste, temos que a raiz do erro quadratico médio e dado por

```
# predições
predictions <- modelo_inicial_cox %>% predict(dados_valid)
data.frame(
  RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
  R2 = R2(predictions, ((dados_valid$X10)^lambda-1)/lambda)
  )
```

```
##       RMSE        R2
## 1 8.086683 0.8341232
```

1.3.2.2 modelo inicial com o metodo de step wise

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

```r
modmin<-lm(X10_cox ~ X6adj+X7, data=dados_train[-c(18,48,46),])

modcompl<-lm(X10_cox~ X1adj+X2adj+X3adj+X4adj+X5adj+X6adj+X7+X8+X9adj+X11adj, data=da

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

```
##
## Call:
## lm(formula = X10_cox ~ X6adj + X7 + X3adj + X2adj + X11adj +
##     X9adj + X1adj + X5adj, data = dados_train[-c(18, 48, 46),
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2762  -2.2828   0.1701   3.7534  10.6444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6234     2.5543  15.513  < 2e-16 ***
## X6adj         5.5375     4.4381   1.248 0.218584
## X72          -3.2001     2.9536  -1.083 0.284389
## X3adj         3.8744     1.0969   3.532 0.000966 ***
## X2adj         2.3237     1.0954   2.121 0.039442 *
## X11adj        1.8227     1.1943   1.526 0.133955
## X9adj        10.4575     4.6942   2.228 0.030942 *
## X1adj        -2.6161     1.3893  -1.883 0.066159 .
## X5adj         1.4966     0.8374   1.787 0.080660 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.434 on 45 degrees of freedom
## Multiple R-squared:  0.9294, Adjusted R-squared:  0.9168
## F-statistic: 74.04 on 8 and 45 DF,  p-value: < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  modfim$residuals
## W = 0.97759, p-value = 0.404
```

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

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

```
## Analysis of Variance Table
##
## Model 1: X10_cox ~ X6adj + X7
## Model 2: X10_cox ~ X6adj + X7 + X3adj + X2adj + X11adj + X9adj + X1adj +
##     X5adj
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     51 2320.4
## 2     45 1328.6  6    991.78 5.5987 0.0002097 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(modelo_inicial_cox,modfim)
```

```
##                    df      AIC
## modelo_inicial_cox  4 364.3127
## modfim             10 346.2019
```

Assim, o modelo 2 apresenta melhor desenpenho considerando o RSS, e o teste linear geral possui diferença significante, ou seja, o modelos são diferentes, agora avalindo este modelo `modfim`, temos que

```
 # quanto menoor melhor
car::vif(modfim)
```

```
##     X6adj        X7     X3adj     X2adj    X11adj     X9adj     X1adj     X5adj
## 32.074398  2.216131  1.986354  1.291405  2.718891 34.913365  2.110575  1.189607
```

para os parametros do $X6$ e $X9$, encontrou grande correlação entre elas, e para avaliar que o modelo não possua colinearidade, temos que

```
modsem9<-lm(X10_cox ~ X6adj+X7+X3adj+X2adj+X11adj+X1adj+X5adj, data=dados_train[-c(18
summary(modsem9)
```

```
##
## Call:
## lm(formula = X10_cox ~ X6adj + X7 + X3adj + X2adj + X11adj +
##     X1adj + X5adj, data = dados_train[-c(18, 48, 46), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5084  -2.9763   0.2571   4.0507  11.7946
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.624      2.620  15.503  < 2e-16 ***
## X6adj         14.878      1.516   9.814 7.36e-13 ***
## X72           -4.464      3.021  -1.478  0.14634
## X3adj          3.848      1.143   3.366  0.00155 **
## X2adj          2.595      1.135   2.287  0.02682 *
## X11adj         1.997      1.242   1.608  0.11477
## X1adj         -1.564      1.362  -1.149  0.25665
## X5adj          1.302      0.868   1.500  0.14046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.663 on 46 degrees of freedom
```

```
## Multiple R-squared:  0.9216, Adjusted R-squared:  0.9097
## F-statistic: 77.25 on 7 and 46 DF,  p-value: < 2.2e-16
```

```
modsem9<-lm(X10_cox ~ X6adj+X3adj, data=dados_train[-c(18,48,46),])
summary(modsem9)
```

```
##
## Call:
## lm(formula = X10_cox ~ X6adj + X3adj, data = dados_train[-c(18,
##     48, 46), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4114  -3.2837   0.4073   4.1360  12.3687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.0721     0.8310  44.612  < 2e-16 ***
## X6adj        17.4011     0.9314  18.683  < 2e-16 ***
## X3adj         3.2619     0.9251   3.526 0.000901 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.079 on 51 degrees of freedom
## Multiple R-squared:  0.8998, Adjusted R-squared:  0.8959
## F-statistic: 229.1 on 2 and 51 DF,  p-value: < 2.2e-16
```

```
car::vif(modsem9)
```

```
##    X6adj    X3adj
## 1.128576 1.128576
```

```
shapiro.test(modsem9$residuals)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  modsem9$residuals
## W = 0.97015, p-value = 0.1957
```

```
predictions <- modsem9 %>% predict(dados_valid)
data.frame(
  RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
  R2 = R2(predictions, (dados_valid$X10^lambda-1)/lambda)
  )
```

```
##       RMSE        R2
## 1 8.121198 0.8406073
```

```
modsem6<-lm(X10_cox ~ X9adj+X7+X3adj+X2adj+X11adj+X1adj+X5adj, data=dados_train[-c(18
summary(modsem6)
```

```
##
## Call:
## lm(formula = X10_cox ~ X9adj + X7 + X3adj + X2adj + X11adj +
##     X1adj + X5adj, data = dados_train[-c(18, 48, 46), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2179  -2.4119   0.1548   3.5366  10.4195
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.5260     2.5685  15.389  < 2e-16 ***
## X9adj        15.9911     1.5478  10.331 1.43e-13 ***
## X72          -3.0875     2.9701  -1.040  0.30400
## X3adj         3.8678     1.1035   3.505  0.00103 **
## X2adj         2.1837     1.0963   1.992  0.05233 .
## X11adj        2.0215     1.1907   1.698  0.09633 .
## X1adj        -3.0601     1.3510  -2.265  0.02827 *
```

```
## X5adj         1.5822     0.8397   1.884  0.06585 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.466 on 46 degrees of freedom
## Multiple R-squared:  0.9269, Adjusted R-squared:  0.9158
## F-statistic: 83.38 on 7 and 46 DF,  p-value: < 2.2e-16
```

```
modsem6<-lm(X10_cox ~ X9adj+X3adj, data=dados_train[-c(18,48,46),])
summary(modsem6)
```

```
##
## Call:
## lm(formula = X10_cox ~ X9adj + X3adj, data = dados_train[-c(18,
##     48, 46), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5728  -1.8658   0.6328   3.2096  11.0777
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2650     0.8272  45.052  < 2e-16 ***
## X9adj        17.8166     0.9479  18.796  < 2e-16 ***
## X3adj         2.8030     0.9286   3.018  0.00396 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.047 on 51 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.897
## F-statistic: 231.8 on 2 and 51 DF,  p-value: < 2.2e-16
```

```
car::vif(modsem6)
```

```
##   X9adj   X3adj
## 1.14937 1.14937
```

```
shapiro.test(modsem6$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modsem6$residuals
## W = 0.93803, p-value = 0.007721
```

```
predictions <- modsem6 %>% predict(dados_valid)
data.frame(
  RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
  R2 = R2(predictions, (dados_valid$X10^lambda-1)/lambda)
  )
```

```
##       RMSE        R2
## 1 8.898686 0.8175485
```

```
anova(modfim,modsem6)
```

```
## Analysis of Variance Table
##
## Model 1: X10_cox ~ X6adj + X7 + X3adj + X2adj + X11adj + X9adj + X1adj +
##     X5adj
## Model 2: X10_cox ~ X9adj + X3adj
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     45 1328.6
## 2     51 1865.0 -6   -536.45 3.0283 0.01424 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modfim,modsem9)
```

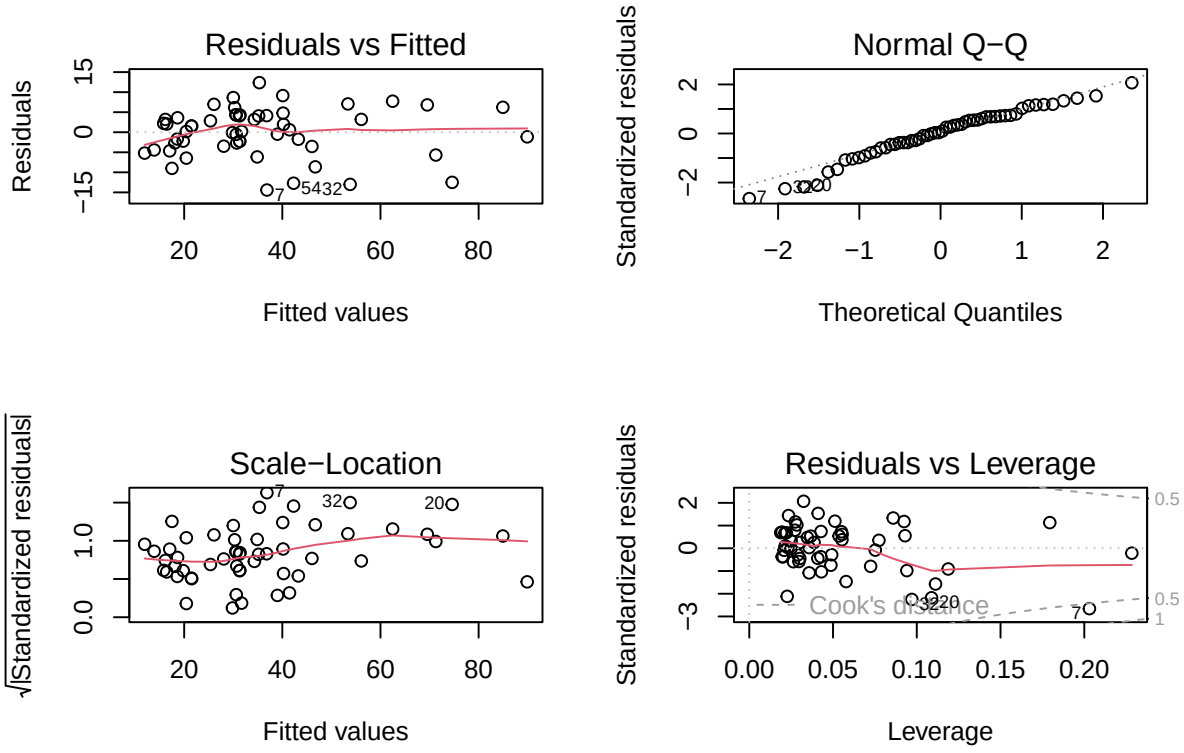
```
## Analysis of Variance Table
##
## Model 1: X10_cox ~ X6adj + X7 + X3adj + X2adj + X11adj + X9adj + X1adj +
```

```
##     X5adj
## Model 2: X10_cox ~ X6adj + X3adj
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     45 1328.6
## 2     51 1884.8 -6   -556.18 3.1396 0.01174 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(modsem9,modsem6)
```

```
##         df      AIC
## modsem9  4 353.0849
## modsem6  4 352.5166
```

```
par(mfrow=c(2,2))
plot(modsem9)
```



assim, no final foi escolhido o modelo `modsem9` no qual os pressupostos são atendidos e possui valores mais consistentes na predição do número de enfermeiros

1.3.2.3 modelo hospital assumptions

Agora como o modelo formulado pelo hospital temos que,

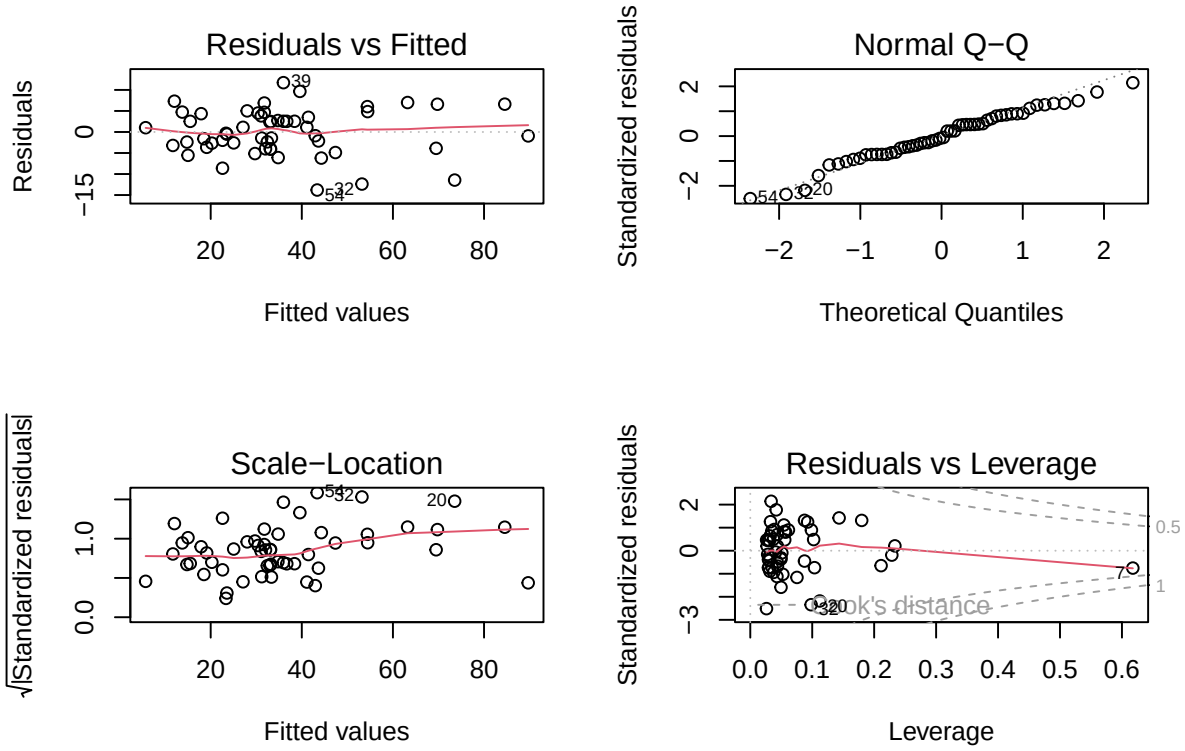
```
mod_sec<- lm(formula = X10_cox ~ X6adj+I(X6adj^2) + X3adj+I(X3adj^2)+X8 , data = dados
summary(mod_sec)
```

```
##
## Call:
## lm(formula = X10_cox ~ X6adj + I(X6adj^2) + X3adj + I(X3adj^2) +
##     X8, data = dados_train[-c(18, 48, 46), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7958  -3.1296  -0.1209   3.8364  11.7547
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.725721   1.772682  22.410  < 2e-16 ***
## X6adj       16.890543   1.622481  10.410 1.12e-13 ***
## I(X6adj^2)  -0.007856   0.869457  -0.009  0.99283
## X3adj        2.514039   1.010543   2.488  0.01654 *
## I(X3adj^2)  -1.832165   0.610414  -3.002  0.00433 **
## X82         -0.951667   2.200682  -0.432  0.66744
## X83         -1.895478   2.211356  -0.857  0.39580
## X84         -1.323005   2.747512  -0.482  0.63242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.753 on 46 degrees of freedom
## Multiple R-squared:  0.9191, Adjusted R-squared:  0.9068
## F-statistic: 74.64 on 7 and 46 DF,  p-value: < 2.2e-16
```

```
mod_sec<- lm(formula = X10_cox ~ X6adj + X3adj+I(X3adj^2) , data = dados_train[-c(18,4
summary(mod_sec)
```

```
##
## Call:
## lm(formula = X10_cox ~ X6adj + X3adj + I(X3adj^2), data = dados_train[-c(18,
##     48, 46), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8032  -3.5342  -0.4144   4.2161  11.7274
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.6812     0.9035  42.811  < 2e-16 ***
## X6adj        16.8500     0.8687  19.397  < 2e-16 ***
## X3adj         2.7070     0.8632   3.136  0.00287 **
## I(X3adj^2)   -1.8592     0.5636  -3.299  0.00179 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.564 on 50 degrees of freedom
## Multiple R-squared:  0.9177, Adjusted R-squared:  0.9128
## F-statistic: 185.9 on 3 and 50 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(mod_sec)
```




```
car::vif(mod_sec)
```

```
##      X6adj      X3adj I(X3adj^2)
##   1.171908   1.173127   1.117543
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_sec$residuals
## W = 0.98009, p-value = 0.5042
```

```
predictions <- mod_sec %>% predict(dados_valid)

data.frame(
```

```
  RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
  R2 = R2(predictions, (dados_valid$X10^lambda-1)/lambda)
  )
```

```
##       RMSE        R2
## 1 7.387265 0.8631546
```

```
predictions <- modsem9 %>% predict(dados_valid)
data.frame(
  RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
  R2 = R2(predictions, (dados_valid$X10^lambda-1)/lambda)
  )
```

```
##       RMSE        R2
## 1 8.121198 0.8406073
```

```
anova(modsem9,mod_sec)
```

```
## Analysis of Variance Table
##
## Model 1: X10_cox ~ X6adj + X3adj
## Model 2: X10_cox ~ X6adj + X3adj + I(X3adj^2)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     51 1884.8
## 2     50 1547.9  1    336.91 10.883 0.001793 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(modsem9,mod_sec)
```

```
##         df      AIC
## modsem9  4 353.0849
## mod_sec  5 344.4506
```

Agora avaliando o teste linear geral e o AIC, temos que o modelo proposto com diferença significativa, e assim, o modelo escolhido foi o que possui ordem quadrática e consegue explicar boa parte da variabilidade do número de enfermeiros.

1.3.3 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