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(data_temp)

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

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

```
library(knitr)
library(kableExtra)

knitr::kable(cbind(Nome,Código,Descrição, Classificação),
             caption = 'Descrição dos códigos da tabela com a seguinte indentificação
             align = "lclr")%>%
  kable_styling(full_width = F)%>%
  column_spec(3,width = "0.4in")
```

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

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

.

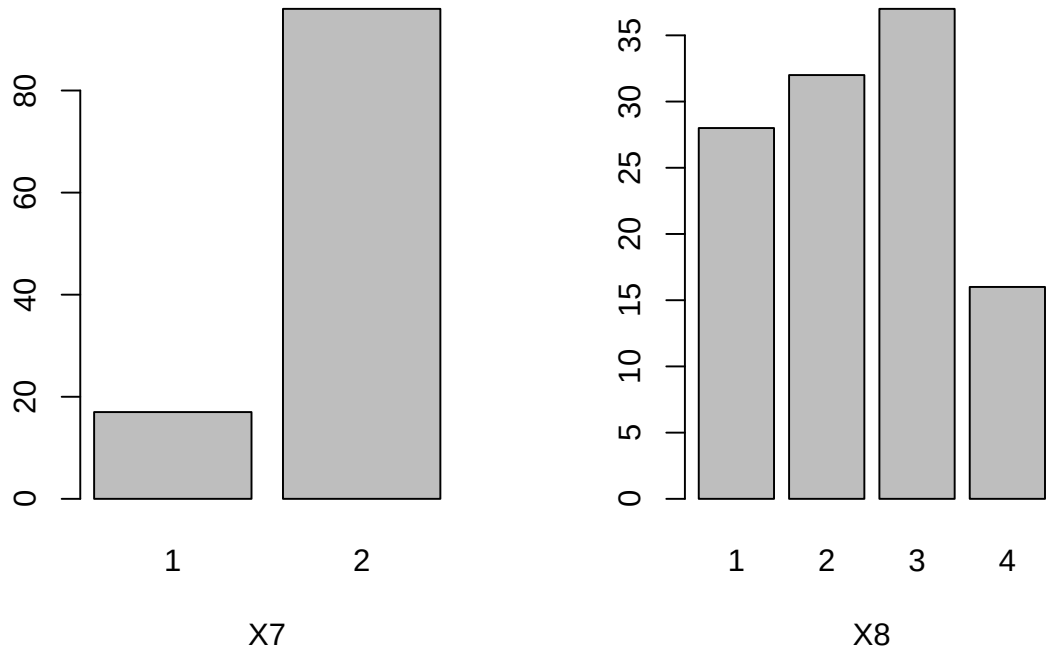
Tabela 1 – Descrição dos códigos da tabela com a seguinte indentificação da variável.

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	Quantitativa contínua

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

```
boxplot(datax_ajusdet)
```

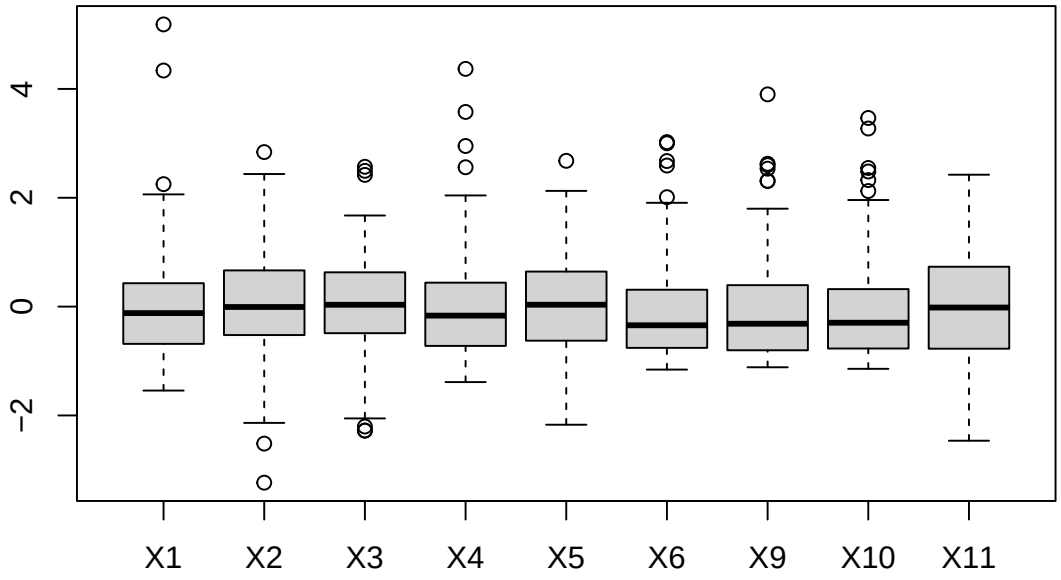


Figura 1 – Gráfico de box-plot das variaváveis dos dados.

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)
library(dplyr)
pmat = dplyr::select(datax,!matches("adj")) %>% select_if(is.numeric) %>% cor_pmat()

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

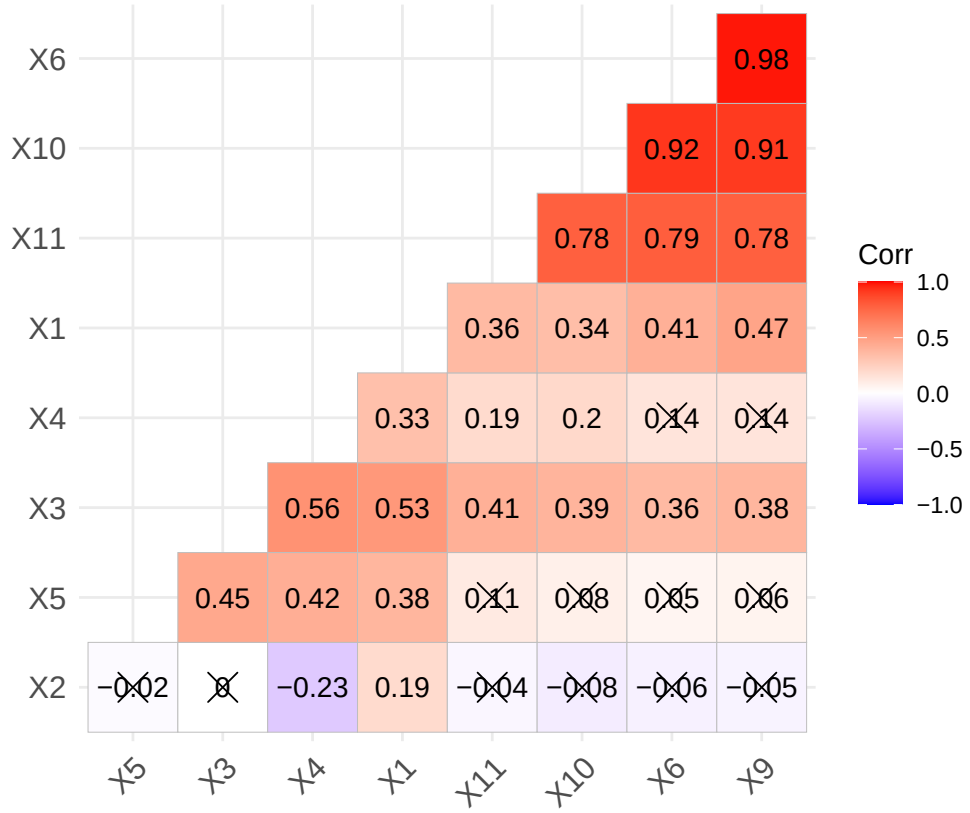


Figura 2 – Gráfico de calor da correlação entre as variaváveis dos dados.

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.

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.

1.3 Objetivo

1.3.1 Testes

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

```
library(xtable)
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()


print.xtable(xtable(table(dados_train$X8),caption = c(''),
                    caption.placement = "top") )
```

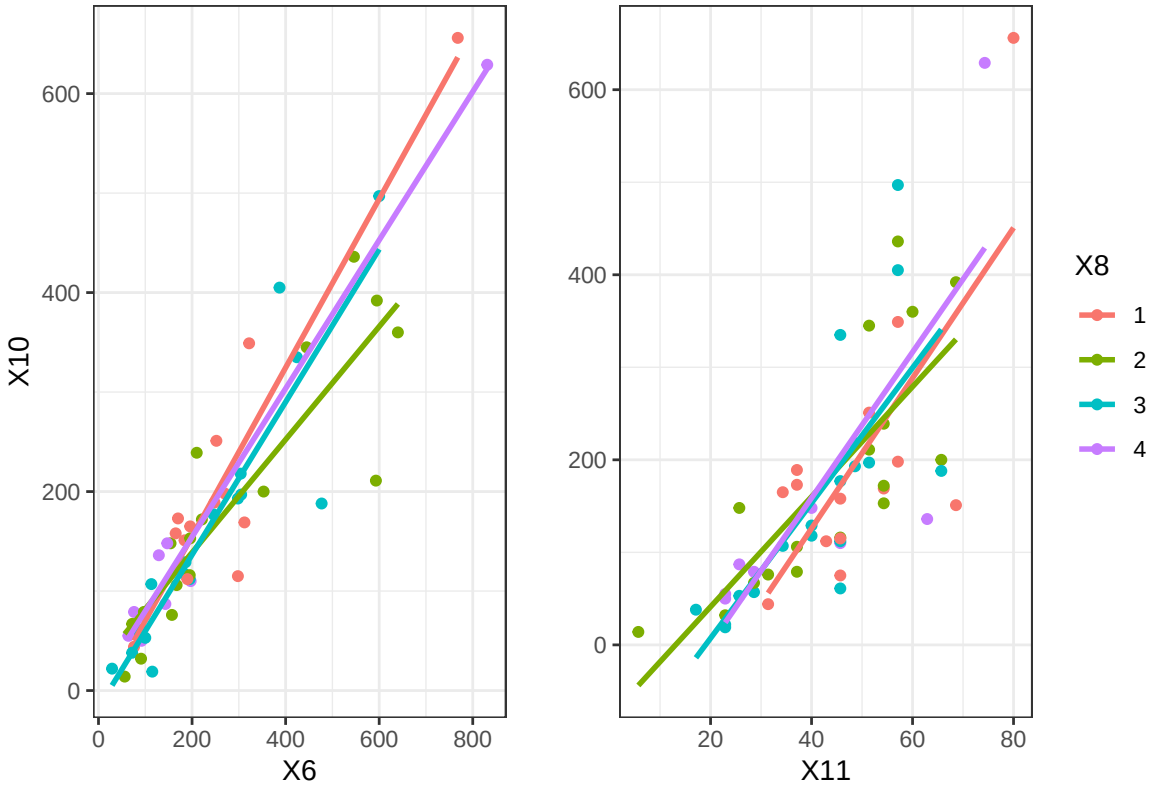
% latex table generated in R 4.2.0 by xtable 1.8-4 package % Thu May 5 19:01:47 2022

	V1
1	14
2	17
3	18
4	8

Tabela 3 –

```
# inbalanced data
```

1.3.2 Número de enfermeira(o)s



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

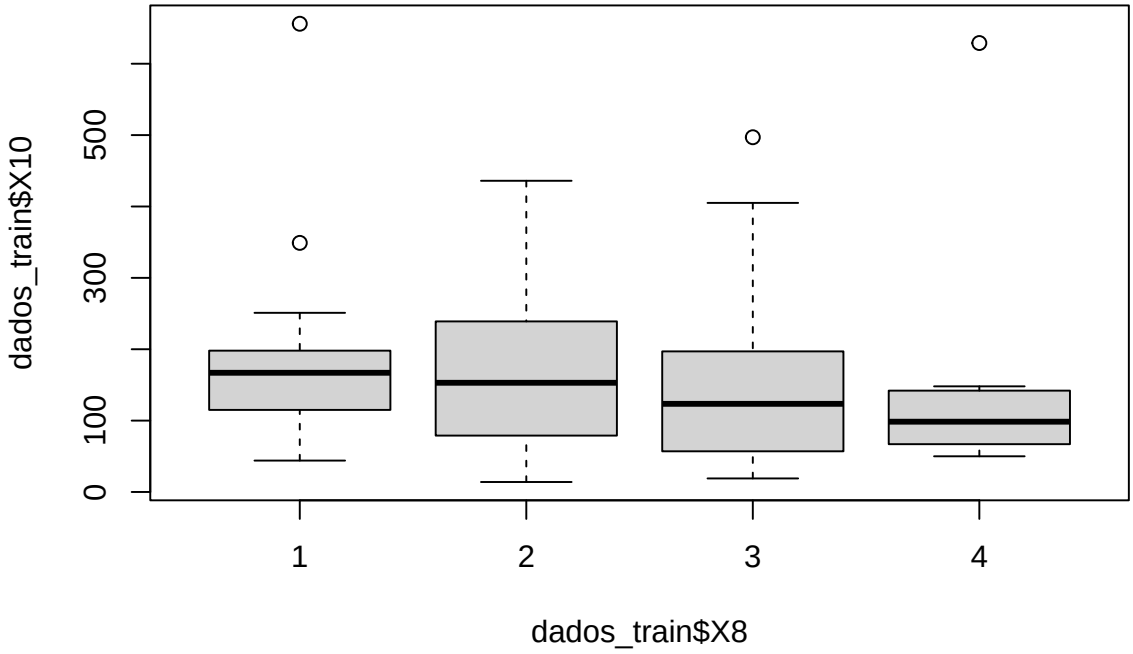
instalações:X7

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)
```




```
print.xtable(xtable(summary(aov(dados_train$X10~dados_train$X8)),caption = c(''),
                    caption.placement = "top") )
```

% latex table generated in R 4.2.0 by xtable 1.8-4 package % Thu May 5 19:01:48 2022

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dados_train$X8	3	14239.94	4746.65	0.22	0.8798
Residuals	53	1126692.10	21258.34		

Tabela 4 –

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".

```
library(xtable)
options(xtable.floating = T)
options(xtable.timestamp = "")
options(xtable.comment = FALSE)

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


print.xtable(xtable(k,caption = c('Modelo 1'),
                    caption.placement = "top") )
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1adj	1	219271.01	219271.01	104.32	0.0000
X8	3	22596.11	7532.04	3.58	0.0227
X6adj	1	727219.51	727219.51	346.00	0.0000
X11adj	1	6462.84	6462.84	3.07	0.0878
X7	1	13250.71	13250.71	6.30	0.0165
X1adj:X8	3	29330.91	9776.97	4.65	0.0074
X8:X6adj	3	15887.68	5295.89	2.52	0.0729
X8:X11adj	3	9871.51	3290.50	1.57	0.2141
X8:X7	3	19274.77	6424.92	3.06	0.0402
Residuals	37	77766.98	2101.81		

Tabela 5 – Modelo 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)


print.xtable(xtable(summary(modelo_inicial),caption = c('$X10 = X1adj*X8 + X6adj +X7*X
                    caption.placement = "top") )
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	252.5302	29.1889	8.65	0.0000
X1adj	2.7834	18.5158	0.15	0.8812
X82	-50.6189	41.1698	-1.23	0.2254
X83	-93.8268	63.6410	-1.47	0.1475
X84	11.2928	67.5629	0.17	0.8680
X6adj	121.2084	11.4402	10.59	0.0000
X72	-76.3742	32.8152	-2.33	0.0246
X1adj:X82	-30.1661	25.6811	-1.17	0.2465
X1adj:X83	71.3170	29.4059	2.43	0.0195
X1adj:X84	2.2999	41.7575	0.06	0.9563
X82:X72	30.4028	44.9525	0.68	0.5024
X83:X72	115.5296	69.3109	1.67	0.1027
X84:X72	-5.9582	82.3340	-0.07	0.9426

Tabela 6 – $X10 = X1adj * X8 + X6adj + X7 * X8$

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)

print.xtable(xtable(summary(modelo_inicial),caption = c('$X10 =  X6adj+X7$'),
                    caption.placement = "top") )
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	228.1758	21.8323	10.45	0.0000
X6adj	120.8863	9.9604	12.14	0.0000
X72	-58.3778	25.0249	-2.33	0.0234

Tabela 7 – $X10 = X6adj + X7$

```
print.xtable(xtable(summary(modelo_inicial),caption = c('$X10 =  X6adj+X7$'),
                    caption.placement = "top") )
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	228.1758	21.8323	10.45	0.0000
X6adj	120.8863	9.9604	12.14	0.0000
X72	-58.3778	25.0249	-2.33	0.0234

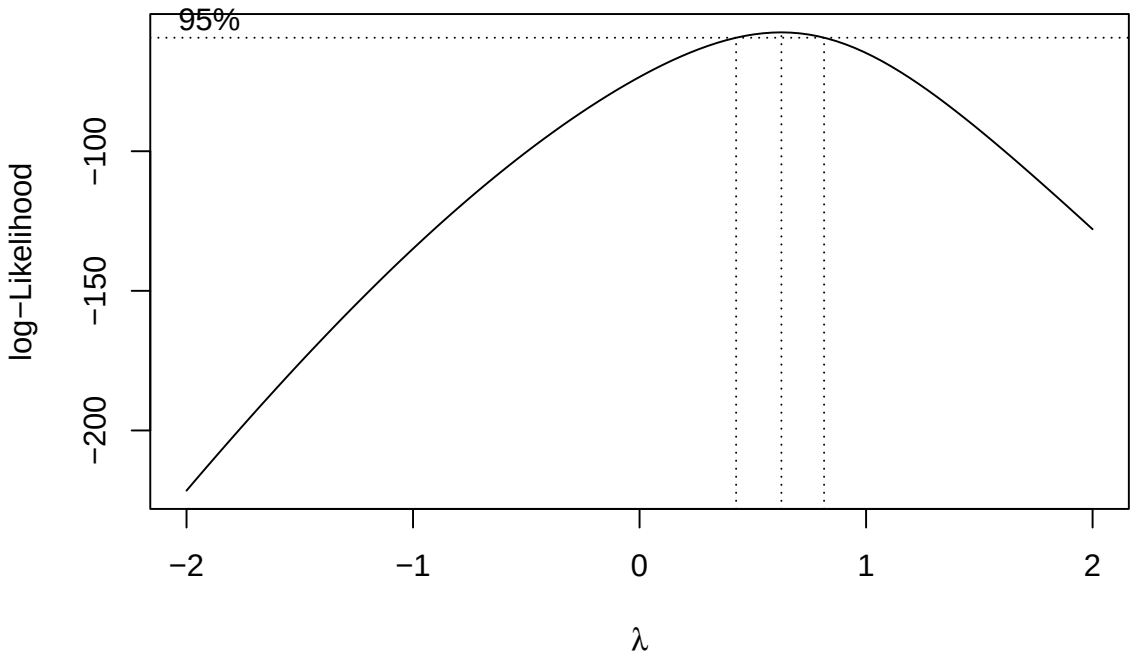
Tabela 8 – $X10 = X6adj + X7$

```
k<-shapiro.test(modelo_inicial$residuals)

k<- cbind(k$method, k$p.value)

print.xtable(xtable(k))
```

	1	2
1	Shapiro-Wilk normality test	0.0246092538237578

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

```
library(MASS)
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)

print.xtable(xtable(summary(modelo_inicial_cox),caption = c('$X10_cox ~X6adj+X7$'),
                    caption.placement = "top") )
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.3733	3.1000	13.67	0.0000
X6adj	16.3551	1.4143	11.56	0.0000
X72	-6.5013	3.5533	-1.83	0.0728

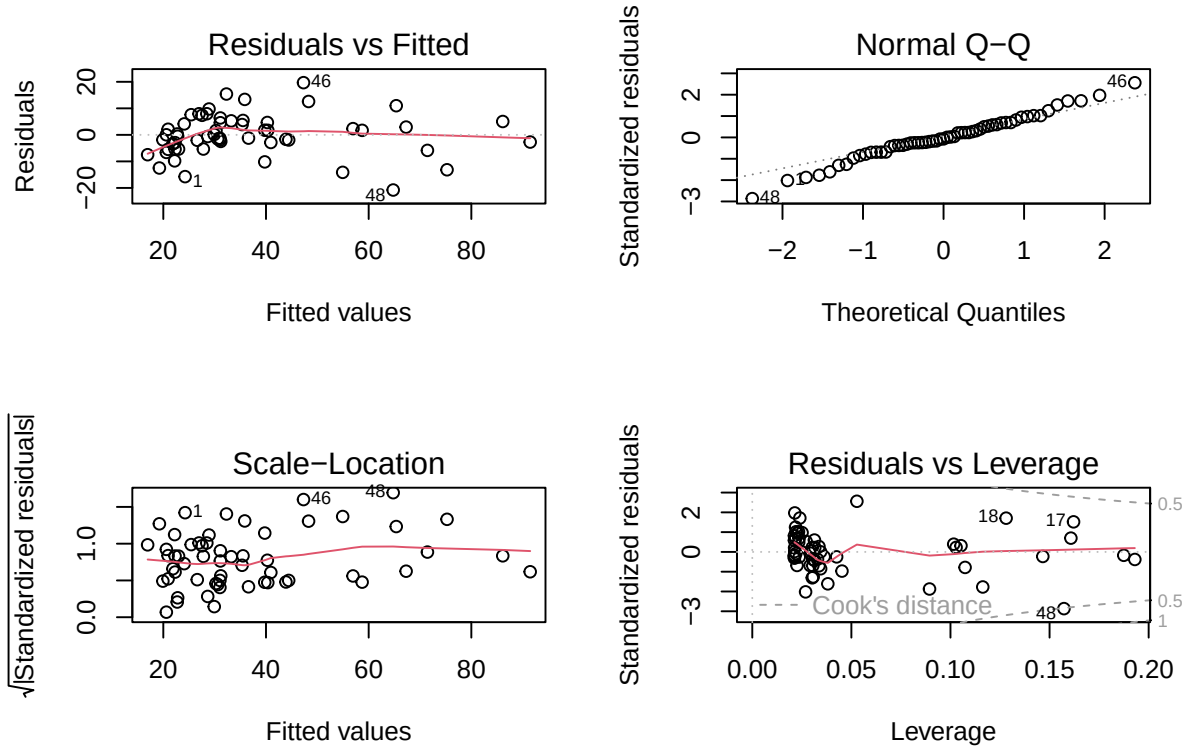
Tabela 9 – $X10_cox\ X6adj + X7$

```
k<-shapiro.test(modelo_inicial_cox$residuals)

k<- cbind(k$method, k$p.value)

print.xtable(xtable(k))
```

	1	2
1	Shapiro-Wilk normality test	0.85952093816942

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)



# 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),])



print.xtable(xtable(summary(modelo_inicial_cox),caption = c('$X10_cox =  X6adj+X7$'),
                    caption.placement = "top") )
```

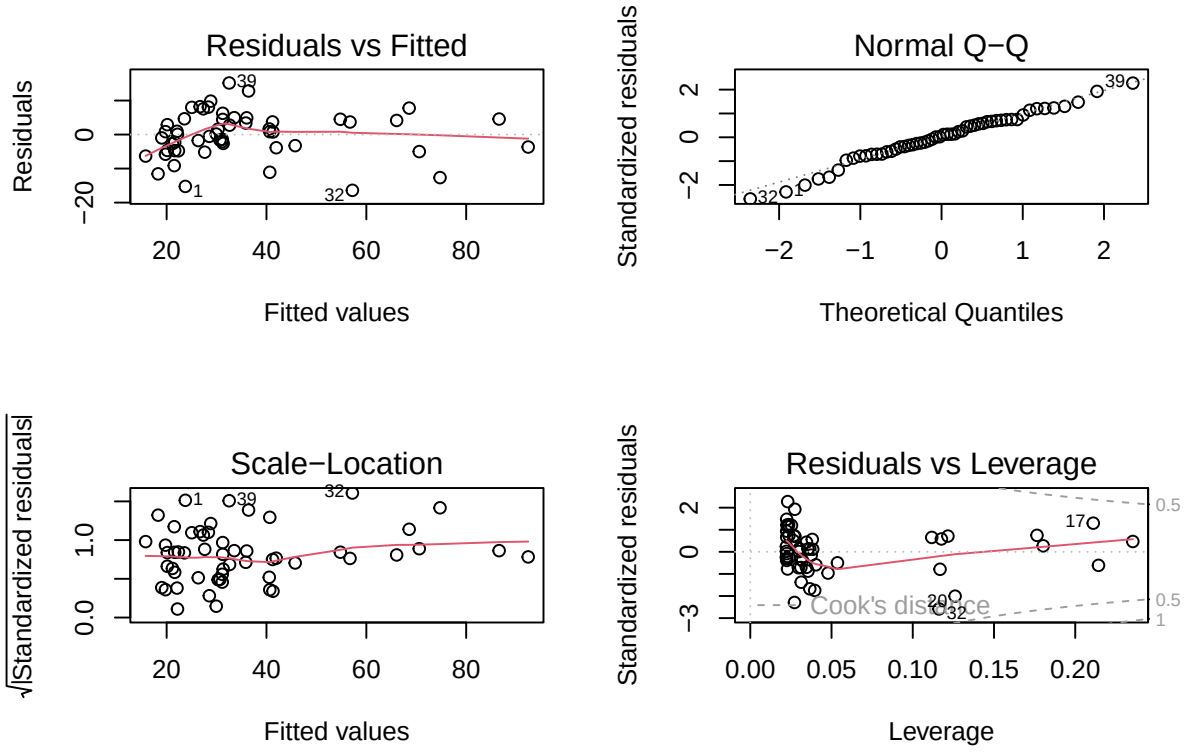
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.9123	2.9434	13.22	0.0000
X6adj	17.8263	1.3548	13.16	0.0000
X72	-2.4858	3.4301	-0.72	0.4719

Tabela 10 – $X10_cox = X6adj + X7$

```
k<-shapiro.test(modelo_inicial_cox$residuals)

k<- cbind(k$method, k$p.value)

print.xtable(xtable(k))
```

	1	2
1	Shapiro-Wilk normality test	0.632227200344572

```
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)



k<-data.frame(
  RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
  R2 = R2(predictions, ((dados_valid$X10)^lambda-1)/lambda)
  )

print.xtable(xtable(k,caption = c(''),
                    caption.placement = "top") )
```

	RMSE	R2
1	8.09	0.83

Tabela 11 –

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

```
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



print.xtable(xtable(summary(modfim),caption = c('summary(modfim)'),
                    caption.placement = "top") )


k<-shapiro.test(modfim$residuals)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	39.6234	2.5543	15.51	0.0000
X6adj	5.5375	4.4381	1.25	0.2186
X72	-3.2001	2.9536	-1.08	0.2844
X3adj	3.8744	1.0969	3.53	0.0010
X2adj	2.3237	1.0954	2.12	0.0394
X11adj	1.8227	1.1943	1.53	0.1340
X9adj	10.4575	4.6942	2.23	0.0309
X1adj	-2.6161	1.3893	-1.88	0.0662
X5adj	1.4966	0.8374	1.79	0.0807

Tabela 12 – summary(modfim)

```
k<- cbind(k$method, k$p.value)


print.xtable(xtable(k))
```

	1	2
1	Shapiro-Wilk normality test	0.403973043275297

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

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

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	51	2320.39				
2	45	1328.60	6	991.78	5.60	0.0002

```
 # modelo 2 é melhor
print.xtable(xtable(AIC(modelo_inicial_cox,modfim)))
```

	df	AIC
modelo_inicial_cox	4.00	364.31
modfim	10.00	346.20

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


print.xtable(xtable(as.data.frame(car::vif(modfim))
))
```

	car::vif(modfim)
X6adj	32.07
X7	2.22
X3adj	1.99
X2adj	1.29
X11adj	2.72
X9adj	34.91
X1adj	2.11
X5adj	1.19

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

```
# quanto menoor melhor



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

print.xtable(xtable(summary(modsem9)))
```

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

```
print.xtable(xtable(as.data.frame(car::vif(modsem9))))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	40.6245	2.6205	15.50	0.0000
X6adj	14.8783	1.5160	9.81	0.0000
X72	-4.4636	3.0210	-1.48	0.1463
X3adj	3.8477	1.1431	3.37	0.0015
X2adj	2.5953	1.1346	2.29	0.0268
X11adj	1.9966	1.2420	1.61	0.1148
X1adj	-1.5640	1.3616	-1.15	0.2567
X5adj	1.3020	0.8680	1.50	0.1405

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.0721	0.8310	44.61	0.0000
X6adj	17.4011	0.9314	18.68	0.0000
X3adj	3.2619	0.9251	3.53	0.0009

	car::vif(modsem9)
X6adj	1.13
X3adj	1.13

```
k<-shapiro.test(modsem9$residuals)

k<- cbind(k$method, k$p.value)

print.xtable(xtable(k))
```

	1	2
1	Shapiro-Wilk normality test	0.195743839981802

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


print.xtable(xtable(k,caption = c(''),
                    caption.placement = "top") )
```

	RMSE	R2
1	8.12	0.84

Tabela 13 –

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	39.5260	2.5685	15.39	0.0000
X9adj	15.9911	1.5478	10.33	0.0000
X72	-3.0875	2.9701	-1.04	0.3040
X3adj	3.8678	1.1035	3.50	0.0010
X2adj	2.1837	1.0963	1.99	0.0523
X11adj	2.0215	1.1907	1.70	0.0963
X1adj	-3.0601	1.3510	-2.26	0.0283
X5adj	1.5822	0.8397	1.88	0.0659

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.2650	0.8272	45.05	0.0000
X9adj	17.8166	0.9479	18.80	0.0000
X3adj	2.8030	0.9286	3.02	0.0040

```
print.xtable(xtable(as.data.frame(
car::vif(modsem6))))
```

	car::vif(modsem6)
X9adj	1.15
X3adj	1.15

```
k<-shapiro.test(modsem6$residuals)

k<- cbind(k$method, k$p.value)

print.xtable(xtable(k))
```

	1	2
1	Shapiro-Wilk normality test	0.00772145754984382

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


print.xtable(xtable(k,caption = c(''),
                    caption.placement = "top") )
```

	RMSE	R2
1	8.90	0.82

Tabela 14 –

```
print.xtable(xtable(anova(modfim,modsem6)))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	45	1328.60				
2	51	1865.05	-6	-536.45	3.03	0.0142

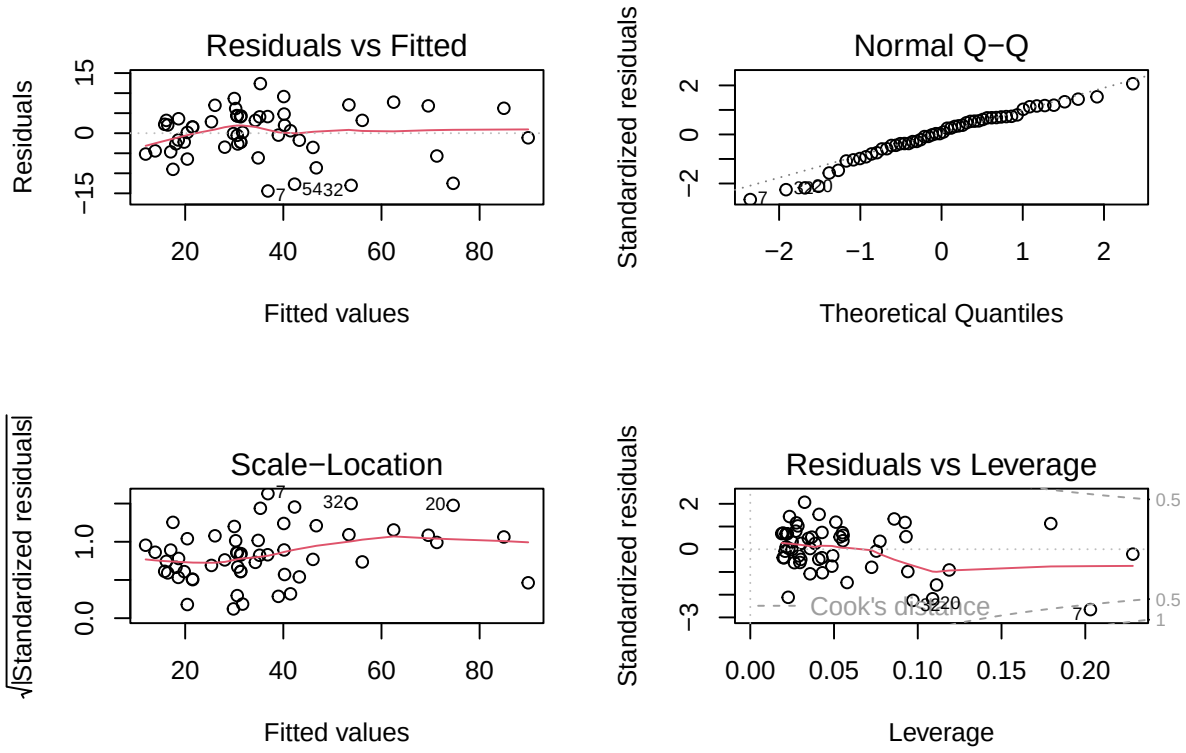
```
print.xtable(xtable(anova(modfim,modsem9)))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	45	1328.60				
2	51	1884.78	-6	-556.18	3.14	0.0117

```
print.xtable(xtable(AIC(modsem9,modsem6)))
```

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

	df	AIC
modsem9	4.00	353.08
modsem6	4.00	352.52



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,

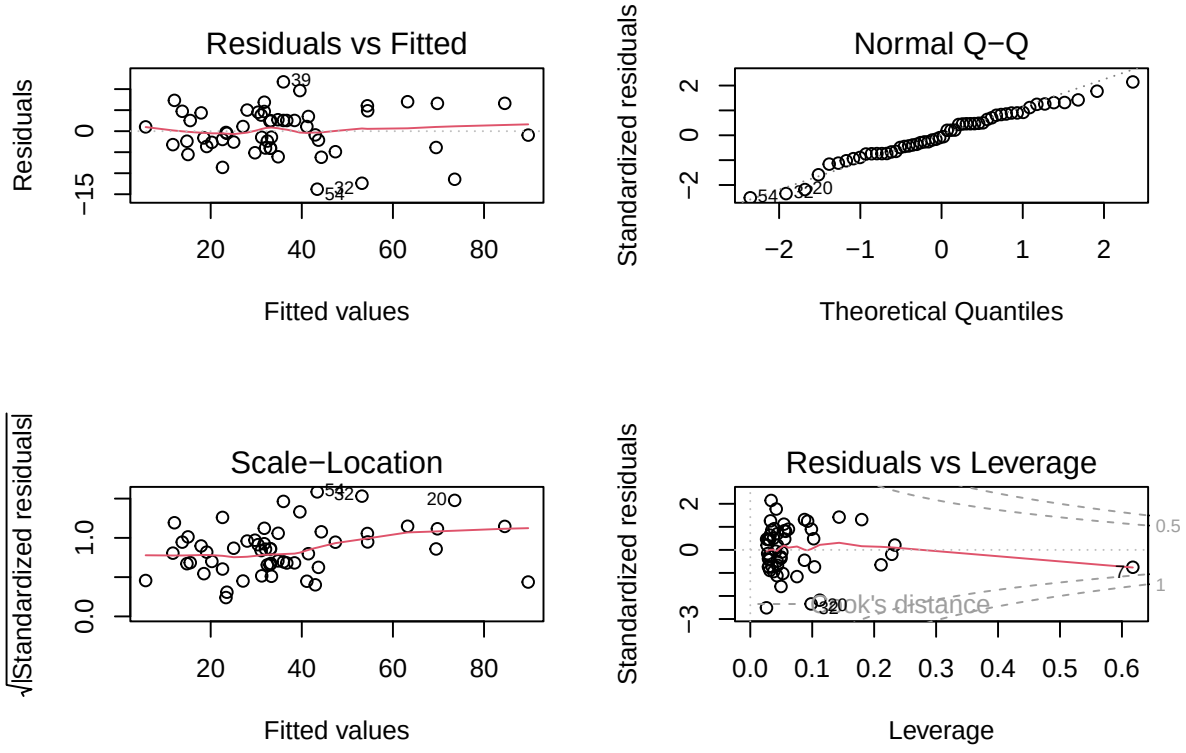
```
mod_sec<- lm(formula = X10_cox ~ X6adj+I(X6adj^2) + X3adj+I(X3adj^2)+X8 , data = dados

print.xtable(xtable(summary(mod_sec)))
```

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	39.7257	1.7727	22.41	0.0000
X6adj	16.8905	1.6225	10.41	0.0000
I(X6adj^2)	-0.0079	0.8695	-0.01	0.9928
X3adj	2.5140	1.0105	2.49	0.0165
I(X3adj^2)	-1.8322	0.6104	-3.00	0.0043
X82	-0.9517	2.2007	-0.43	0.6674
X83	-1.8955	2.2114	-0.86	0.3958
X84	-1.3230	2.7475	-0.48	0.6324

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.6812	0.9035	42.81	0.0000
X6adj	16.8500	0.8687	19.40	0.0000
X3adj	2.7070	0.8632	3.14	0.0029
I(X3adj^2)	-1.8592	0.5636	-3.30	0.0018

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



```
print.xtable(xtable(as.data.frame(
car::vif(mod_sec))))
```

	car::vif(mod_sec)
X6adj	1.17
X3adj	1.17
I(X3adj^2)	1.12

```
k<-shapiro.test(mod_sec$residuals)

k<- cbind(k$method, k$p.value)

print.xtable(xtable(k))
```

	1	2
1	Shapiro-Wilk normality test	0.504228264019901

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

k<-data.frame(
  RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
  R2 = R2(predictions, (dados_valid$X10^lambda-1)/lambda)
  )

print.xtable(xtable(k,caption = c(''),
                    caption.placement = "top") )
```

	RMSE	R2
1	7.39	0.86

Tabela 15 –

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

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


print.xtable(xtable(k,caption = c(''),
                    caption.placement = "top") )
```

	RMSE	R2
1	8.12	0.84

Tabela 16 –

```
print.xtable(xtable(anova(modsem9,mod_sec)))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	51	1884.78				
2	50	1547.87	1	336.91	10.88	0.0018

```
print.xtable(xtable(AIC(modsem9,mod_sec)))
```

	df	AIC
modsem9	4.00	353.08
mod_sec	5.00	344.45

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:

1.3.4 MODELO POR HIPOTESE

```
# summary(aov(X1 ~ X2+X3+X4+X5+X6+X7+X9+X10+X11, data=dados_train))
#
#
# modelo_inicial <- lm(X1 ~  X3 + X6 + X9  ,data=dados_train)
#
#
# car::vif(modelo_inicial) #multicolinearidade
#
#
#
#
# pmat = datax %>% select_if(is.numeric) %>%cor_pmat()
# datax %>% select_if(is.numeric) %>% cor(.) %>%
#   ggcorrplot( type = "lower", p.mat = pmat, hc.order = TRUE)
#
#
#
# #
# # cor(X3,X9)
# #
# #
# # cor(X3,X6)
# #
# #
# # cor(X9,X6)
# #
# #
# # cor(X1,X9)
# #
```

```
# #
# # cor(X1,X6)
# #
#
# modelo_inicial <- lm(X1 ~  X3  + X9  ,data=dados_train)
# car::vif(modelo_inicial) #multicolinearidade
#
#
# shapiro.test(modelo_inicial$residuals)#normal
#
#
# predictions <- modelo_inicial %>% predict(dados_valid)
# RMSE(predictions, dados_valid$X1)# modelo bom
#
# R2(predictions, dados_valid$X1)  # Fraco
#
# ##Forward##
# modmin<-lm(X1 ~ 1, data=dados_train)
# step(modmin, direction='forward', scope=( ~ X2+X3+X4+X5+X6+X7+X8+X9+X10+X11))
#
#
#
# ##Backward##
# modcompl<-lm(X1 ~ X2+X3+X4+X5+X6+X7+X8+X9+X10+X11, data=dados_train)
# step(modcompl, direction = 'backward')
#
#
# # stepwise
# modmin<-lm(X1 ~ 1, data=dados_train)
# modcompl<-lm(X1 ~ X2+X3+X4+X5+X6+X7+X8+X9+X10+X11, data=dados_train)
# step(modmin, scope=list(lower=modmin, upper=modcompl),
#      direction="both",data=dados_train)
#
#
# #todos os modelos foram iguais
```

```
# modstep <- lm(formula = X1 ~ X3 + X8 + X9 + X11, data = dados_train)
# summary(modstep)
#
#
# shapiro.test(modstep$residuals)#Normal
#
# par(mfrow = c(4,1))
# plot(modelo_inicial)
#
#
# predictions <- modstep  %>% predict(dados_valid)
# RMSE(predictions, dados_valid$X1)#modelo bom
#
# AIC(modelo_inicial,modstep)
```

Referências

Anexos

ANEXO A – Amostra