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

Tabela 1 – Medidas descritivas para boxplots 10

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 8
1.1 Introdução 8
1.1.1 Objetivos 8
1.1.2 Metodologia 8
1.2 Resultado 10
1.2.0.1 Correlação entre as variáveis 11
1.3 Objetivo 12
1.3.1 Testes 12
1.3.2 Número de enfermeira(o)s 12
1.3.2.1 Pressupostos para um modelo inicial 13
1.3.2.2 modelo inicial com o metodo de step wise 15
1.3.2.3 modelo hospital assumptions 17
1.3.3 Duração da internação 19

REFERÊNCIAS 20

ANEXOS 21

ANEXO A – AMOSTRA 22

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

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

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 1 – 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)
# 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)
```

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')
```

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:

```
# 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)
```

1.3.2 Número de enfermeira(o)s

```
# library(plotly)
# require(gridExtra)
# 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")

### ANANDA CODE
# 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))
```

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

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

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))
```

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 variávei $X1$ 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)
#
# modelo_inicial <- lm(X10 ~ X1adj*X8 + X6adj +X7*X8, data=dados_train)
# summary(modelo_inicial)
```

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

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$X10^lambda-1)/lambda
#
# modelo_inicial_cox <- lm(X10_cox ~X6adj+X7, data=dados_train)
# summary(modelo_inicial_cox)
# shapiro.test(modelo_inicial_cox$residuals)
```

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),])
# summary(modelo_inicial_cox)
# shapiro.test(modelo_inicial_cox$residuals)
#
# 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)
#   )
```

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, dat
#
# modfim  <- step(modmin, scope=list(lower=modmin, upper=modcompl), direction="both
# summary(modfim)
# shapiro.test(modfim$residuals)
```

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
# AIC(modelo_inicial_cox,modfim)
```

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)
```

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
# summary(modsem9)
# modsem9<-lm(X10_cox ~ X6adj+X3adj, data=dados_train[-c(18,48,46),])
# summary(modsem9)
# car::vif(modsem9)
# shapiro.test(modsem9$residuals)
#
# predictions <- modsem9 %>% predict(dados_valid)
# data.frame(
#   RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
```

```
#   R2 = R2(predictions, (dados_valid$X10^lambda-1)/lambda)
#   )
#
#
#
# modsem6<-lm(X10_cox ~ X9adj+X7+X3adj+X2adj+X11adj+X1adj+X5adj, data=dados_train[-c
# summary(modsem6)
# modsem6<-lm(X10_cox ~ X9adj+X3adj, data=dados_train[-c(18,48,46),])
# summary(modsem6)
# car::vif(modsem6)
# shapiro.test(modsem6$residuals)
#
#
# 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)
#   )
#
#
#
# anova(modfim,modsem6)
# anova(modfim,modsem9)
# AIC(modsem9,modsem6)
```

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

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

```
# car::vif(mod_sec)
# shapiro.test(mod_sec$residuals)
#
# predictions <- mod_sec %>% predict(dados_valid)
#
# data.frame(
#   RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
#   R2 = R2(predictions, (dados_valid$X10^lambda-1)/lambda)
#   )
#
# predictions <- modsem9 %>% predict(dados_valid)
# data.frame(
#   RMSE = RMSE(predictions, (dados_valid$X10^lambda-1)/lambda),
#   R2 = R2(predictions, (dados_valid$X10^lambda-1)/lambda)
#   )
#
```

```
# anova(modsem9,mod_sec)
# AIC(modsem9,mod_sec)
```

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