Curso de Especialização em *Data Science* e Estatística Aplicada
Módulo IV - *Análise de Regressão*
Atividade Avaliativa

Cynthia Tojeiro

25-02-2025

Instruções

- O desenvolvimento desta atividade deve ser realizada de forma individual ou em dupla.
- Deve-se completar o arquivo Rmd enviado na atividade.
- É necessário devolver o arquivo em Rmd e em pdf.
- Valor da atividade: 10 pontos.
- Use o código em anexo como base.
- A atividade deve conter todas as etapas abaixo com conclusões.

- O conjunto de dados trata-se de 11 características clínicas utilizadas para a previsão de possíveis eventos relacionados a pressão arterial.

- A hipertensão arterial (HA) representa o principal fator de risco para desenvolvimento de doenças cardiovasculares e mortalidade em todo o mundo. É uma doença multifatorial, caracterizada e diagnosticada por níveis elevados e sustentados de pressão arterial (PA), possuindo, como critério clínico, em indivíduos maiores de 18 anos, níveis tensionais iguais ou maiores a 140 mmHg $\times$ 90 mmHg1.
- Há diversos fatores que podem ser responsáveis pelo desenvolvimento da doença. Entre eles podemos citar: idade, sexo, colesterol, açúcar no sangue em jejum, doença cardíaca, entre outros. O objetivo nesse trabalho é verificar dentre as variáveis disponíveis quais delas podem contribuir para a hipertensão arterial(RestingBP).
- Nesse estudo foram utilizados dados relacionados à variável dependente (pressão arterial) e às seguintes variáveis independentes: idade, sexo, colesterol sérico, açúcar no sangue em jejum, resultados de eletrocardiograma, doenças cardíacas , angina induzida por exercícios, frequência cardíaca, "OldPeak" e "ST" que é relacionada a inclinação do segmento ST do exercício de pico.

- Questões para a atividade avaliativa:

1. (0.5 pts.) Realizar a Análise Descritiva dos Dados: Para se familiarizar com os dados, faça uma boa caracterização de todas as variáveis quantitativas e qualitativas do arquivo, uma a uma, usando medidas resumo adequados a cada variável. Comente!
2. (0.5 pts.) Faça uma análise descritiva gráfica dos dados usando boxplots e diagramas de dispersão. Comente!

3. (0.5 pts.) Calcule correlações lineares de Pearson entre as variáveis contínuas (faça um gráfico de correlações), comente.
4. (0.5 pt.) Proponha um modelo de regressão normal linear com todas as variáveis explicativas.Comente!
5. (0.8 pts.) Faça uma análise de Multicolinearidade. Comente!
6. (0.8 pts.) Aplique o método do StepWise para verificar qual o melhor ajuste. Comente!
7. (0.8 pts.) No modelo escolhido pelo StepWise faça a interpretação dos parâmetros.
8. (0.8 pts.)Faça uma anova entre os 2 modelos, com todas as variáveis e aquele com as variáveis dependentes escolhidas pelo StepWise. Comente!
9. (0.8 pts.) Para o modelo escolhido em 5) aplicar as análises de resíduos verificando as condições de normalidade e heterocedasticidade (Usar os programas postados na segunda aula).Comente todas elas!
10. (0.8 pts.) Caso algum pressuposto do modelo de regressão normal linear falhe, tente fazer alguma transformação (vistas em aula) com o objetivo de atingir os pressupostos do modelo. Comente!
11. (0.8 pts.) Caso nenhuma transformação funcione, tente utilizar Box-Cox.Comente!
12. (0.8 pts.) Eliminar individualmente os pontos mais discrepantes (se existirem), e verificar se houve mudança inferencial. Comente!
13. (0.8 pts.) Tente atingir a normalidade ao nível de significância de 10%. Comente!
14. (0.8 pts.) Elaborar a conclusão.

Pacotes Necessários

```
library(readr)
library(car)
library(tidyverse)
library(robustbase) #Boxplot robusto das variaveis
library(dplyr)       # Manipulação de dados
library(ggplot2)     # Visualização de dados
library(MASS)# necessário para análise de multicolinearidade
library(alr4)
library(xtable)
library("ggcorrplot")
library(visdat)#visualização "mais elegante" dos dados
library(skimr) #análise exploratória mais geral
library(stargazer)#Tabelas
library(plotly)#Box-plots com informações
library(corrplot)
library(lmtest)#teste de Breush-Pagan
```

Carregando o dataset

```r
heart <- read.csv("C:\\datascience\\heart.csv")
attach(heart)

# Inspecionando os dados
summary(heart)  # Estatísticas descritivas das variáveis
str(heart)      # Estrutura do dataset
```

#É necessário realizar algumas transformações, para transformar as variáveis #categóricas com os seus respectivos fatores.

```r
Sex<- factor(Sex, levels = c("M","F"), labels = c("Masculino", "Feminino"))
ChestPainType <- factor(ChestPainType, levels = c("TA", "ATA", "NAP", "ASY"),

labels = c("Angina Típica", "Angina Atípica", "Dor Não Anginosa", "Assintomática"))

FastingBS <- factor(FastingBS, levels = c(0,1),labels = c("C.C", "JejumBS > 120 mg/dl"))

RestingECG <- factor(RestingECG, levels = c("Normal", "ST", "LVH"),
labels = c("Normal", "anormalidade da onda", "hipertrofia ventricular"))

ExerciseAngina <- factor(ExerciseAngina, levels = c("N", "Y"), labels = c("Não", "Sim"))

inclinacao <- factor(ST_Slope, levels = c("Up", "Flat", "Down"),
labels = c("Ascendente", "Plano", "Descendente"))

HeartDisease <- factor(HeartDisease, levels = c(0,1),labels = c("Normal", "Doença cardiaca"))
```

#Comentário: Com o gráfico abaixo, podemos analisar se existe informação #ausente, como NA e em quais variáveis encontra-se a observação ausente, se #assim existir:

visdat::vis_miss(heart)

#Outra maneira de ver se existem informações ausentes no banco de dados:

is.na(heart) %>% colSums()

#Tratamento de valores ausentes (substituindo zeros por médias nas variáveis contínuas (Não faz sentido ter colesterol=0 e pressão=0))

```r
heart_clean <- heart %>%
  mutate(across(c(RestingBP, Cholesterol,), ~
ifelse(. == 0, mean(., na.rm = TRUE), .)))

# Confirmando o tratamento
summary(heart_clean)

#Gráfico que classifica as variáveis:

visdat::vis_dat(heart)
```

#Análise exploratória mais completa

```r
skim(heart)
```

#Exemplos de Box-plot (Fazer para todas as variáveis sempre diferenciando indivíduos cardíacos e não cardíacos)

```
plot_ly(heart, x =RestingBP, color = cardiaco, type = "box") %>%
layout(title = "Boxplot da Pressão de acordo com a presença ou ausência de doença cardíaca")

 plot_ly(heart, x = Colesterol, color = cardiaco, type = "box") %>%
layout(title = "Boxplot da Colesterol de acordo com a presença ou ausência de doença cardíaca")

plot_ly(heart, x = Idade, color = cardiaco, type = "box") %>%
layout(title = "Boxplot da Idade de acordo com a presença ou ausência de doença cardíaca")
```

#Matriz de Correlações

```
cor_matrix <- cor(heart_clean[,-c(2,3,6,7,9,10,11,12)], method = "pearson")
corrplot(cor(heart_clean[,-c(2,3,6,7,9,10,11,12)],method = "pearson"), method = "number")
```

#Ajuste do modelo ...

#Exemplo de eliminação de observações discrepantes:

```
heart.2 = heart_clean[-obs1,]

fit.model<-ajuste1<-lm(log(RestingBP) ~ Age + Cholesterol + ExerciseAngina
+ Oldpeak + ST_Slope, data=heart.2)

heart.3 = heart_clean[-c(obs1,obs2),]

fit.model<-ajuste2<-lm(log(RestingBP) ~ Age + Cholesterol + ExerciseAngina
+ Oldpeak + ST_Slope, data=heart.3)
```

#Exemplo de verificação do modelo com e sem as observações:

```
stargazer(ajuste com todas as observações, ajuste com as observações retiradas, type = "text")
```

#Programas para verificação dos pressupostos dos modelos postados na segunda aula

```
(source("C:\\datascience\\Programas\\Diag2.norm.r")\\
source("C:\\datascience\\Programas\\Envel_norm.r")\\
source("C:\\datascience\\Programas\\anainflu_norm.r"))
```

BOA SORTE NA ATIVIDADE AVALIATIVA!