Trabalho de Conclusão do Curso de Microeconometria Turma 2018

Análise de Dados Cardiológicos

Angelo Antonio Paula da Cunha

Descrição do Problema

O dataset usado neste trabalho consiste em um conjunto de dados cardiológicos contento informações dos pacientes. Aqui temos informações como Idade, Altura, Peso, Gênero, Pressão Arterial Sistólica e Diastólica, Colesterol, Glicose, Ingestão de Álcool, Fumante, Atividade Física e Presença ou Ausência de doença cardiovascular. Através desses dados busco responder, através de um modelo de regressão linear e um quantil, os determinantes do peso dos indivíduos.

Primeiramente vamos obter os dados que se encontram no site: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Dados e Manipulação

```
library(tidyverse)
library(quantreg)
library(cowplot)
library(corrplot)

cardio = read.csv2("cardio.csv", header = TRUE, sep = ";", dec=".")
```

Aqui vamos criar uma variável que será representada como idade por anos.

```
cardio<-mutate(cardio, age_year = age / 365)
```

Análise da Base de Dados

Antes de fazer o modelo é interessante a análise descritiva e exploratória dos dados. Nesse sentido será exposto aqui a análise das variáveis do conjunto. Começamos pelas médias das variáveis: Weight, Height, ap_hi e pa_lo. A média de idade é de 53 anos, peso dos paciêntes é de 74.20 quilos, da altura 164.35 cm, pressão Máx foi de 128.81 e da pressão Min de 96.63.

```
mean(cardio$height)
```

[1] 164.3592

```
mean(cardio$weight)
```

[1] 74.20569

```
mean(cardio$ap_hi)
```

[1] 128.8173

```
mean(cardio$ap_lo)
```

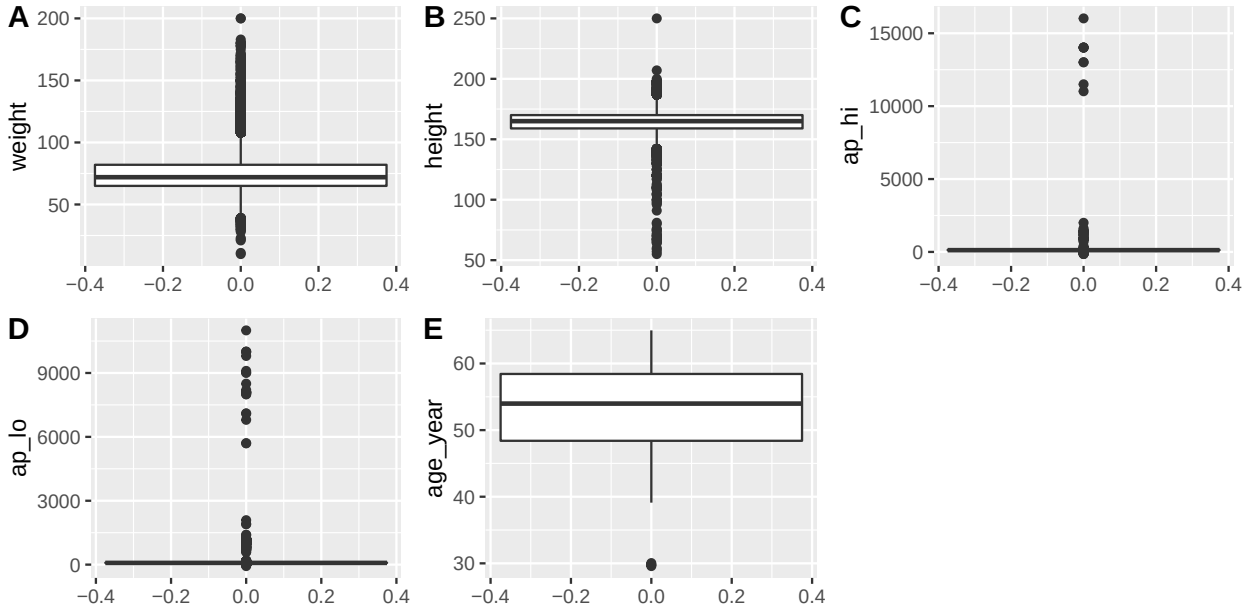
[1] 96.63041

```
mean(cardio$age_year)
```

[1] 53.33936

Para melhorar a análise das medidas descritivas será feito um gráfico boxplot para melhro vizualizar a disperção dos dados.

```
plot.wei = ggplot(cardio, aes(y = weight)) +
  geom_boxplot()

plot.hei = ggplot(cardio, aes(y = height)) +
  geom_boxplot()

plot.aphi = ggplot(cardio, aes(y = ap_hi)) +
  geom_boxplot()

plot.aplo = ggplot(cardio, aes(y = ap_lo)) +
  geom_boxplot()

plot.age = ggplot(cardio, aes(y = age_year)) +
  geom_boxplot()

plot_grid(plot.wei, plot.hei,plot.aphi, plot.aplo, plot.age, labels = "AUTO")
```



Podemos notar nos gráficos que rpincipalmente nas variáveis de pressão existem alguns pontos extremos, ainda mais quando essas variáveis se tratam de pressão arterial e vemos alguns valores bem altos.

Outra análise a ser feita será da frequência de gênero dos dados. Onde podemos observar que 1 = Mulher e 2 = Homem. Assim pelo resultado da tabela de frequência temos 65% dos paciêntes mulheres e 35% homens.

```
count.gender = table(as.factor(cardio$gender))
count.gender
```

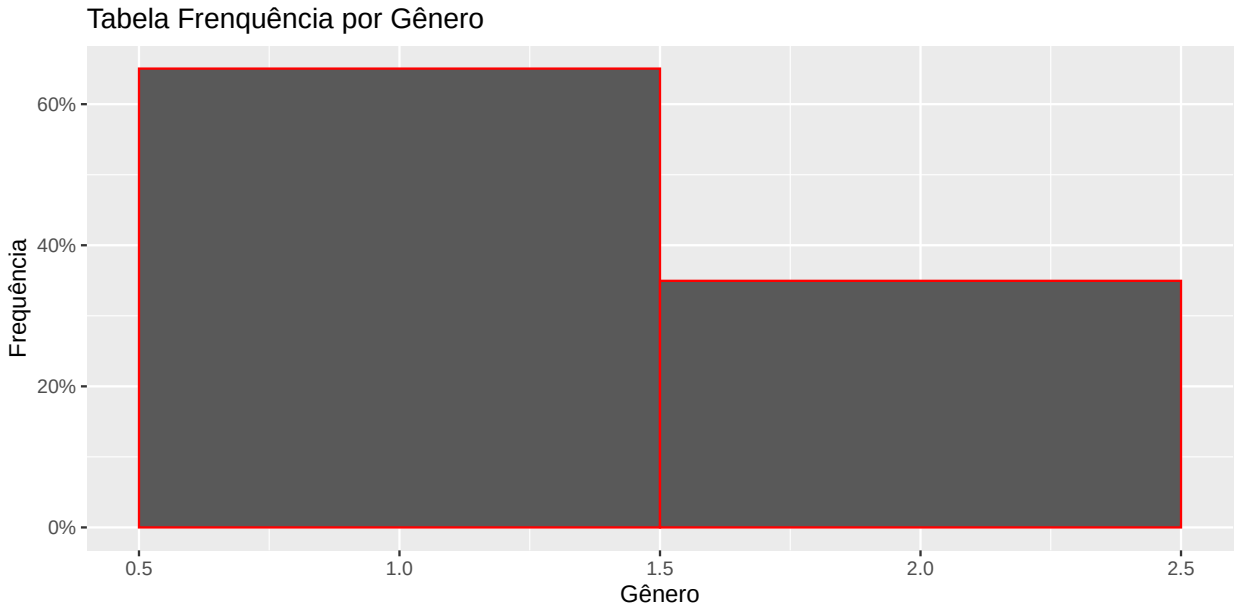
```
1     2
```

45530 24470

```
fs.gender = round(count.gender/sum(count.gender)*100)
fs.gender
```

1 2 65 35

```
ggplot(cardio, aes(x = gender)) +
  geom_histogram(aes(y = stat(count.gender) / sum(count.gender)), bins = 2, color="red") +
    scale_y_continuous(labels = scales::percent)+
      labs(title = 'Tabela Frenquência por Gênero',
       y = 'Frequência', x = 'Gênero')
```



Aqui será análisada a variável cholesterol. Podemos observar que 75% dos paciêntes têm cholesterol do tipo 1, 14% do tipo 2 e 12% do tipo 3.

```
count.cholesterol=table(as.factor(cardio$cholesterol))
count.cholesterol
```

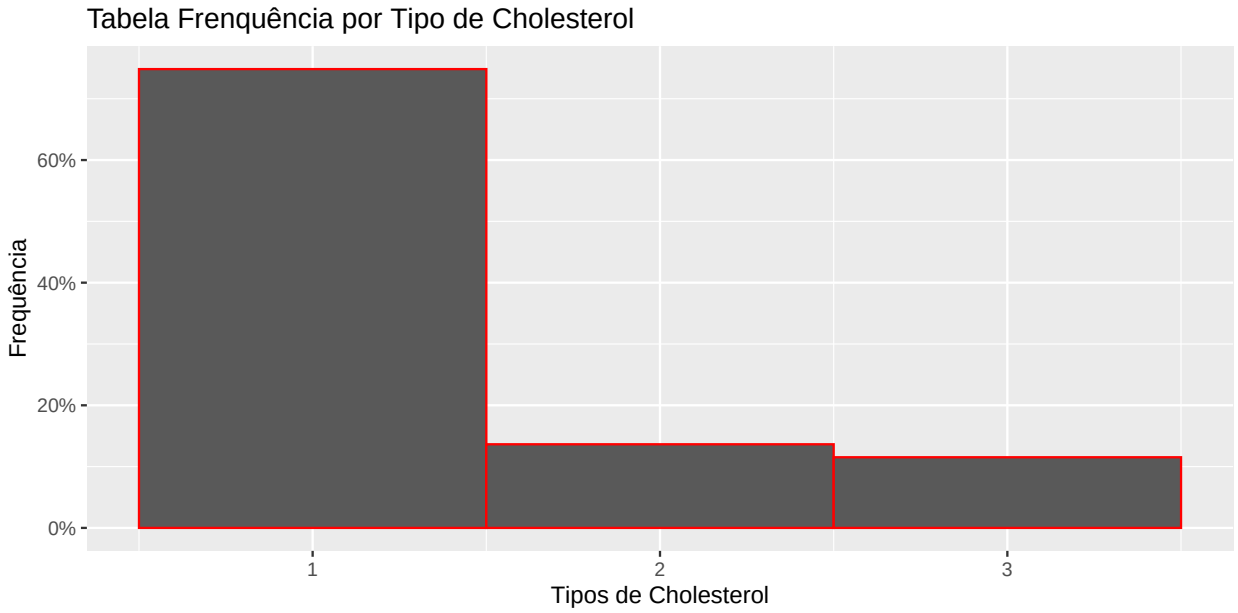
```
1     2     3
```

52385 9549 8066

```
fs.cholesterol = round(count.cholesterol/sum(count.cholesterol)*100)
fs.cholesterol
```

1 2 3 75 14 12

```
ggplot(cardio, aes(x = cholesterol)) +
  geom_histogram(aes(y = stat(count.cholesterol) / sum(count.cholesterol)), bins = 3, color="red") +
  scale_y_continuous(labels = scales::percent)+
  labs(title = 'Tabela Frenquência por Tipo de Cholesterol',
       y = 'Frequência', x = 'Tipos de Cholesterol')
```



Outra variável que investigaremos é a Glicose (gluc). Observamos que 85% dos pacientes tem diabetes do tipo 1.

```
count.gluc=table(as.factor(cardio$gluc))
count.gluc
```

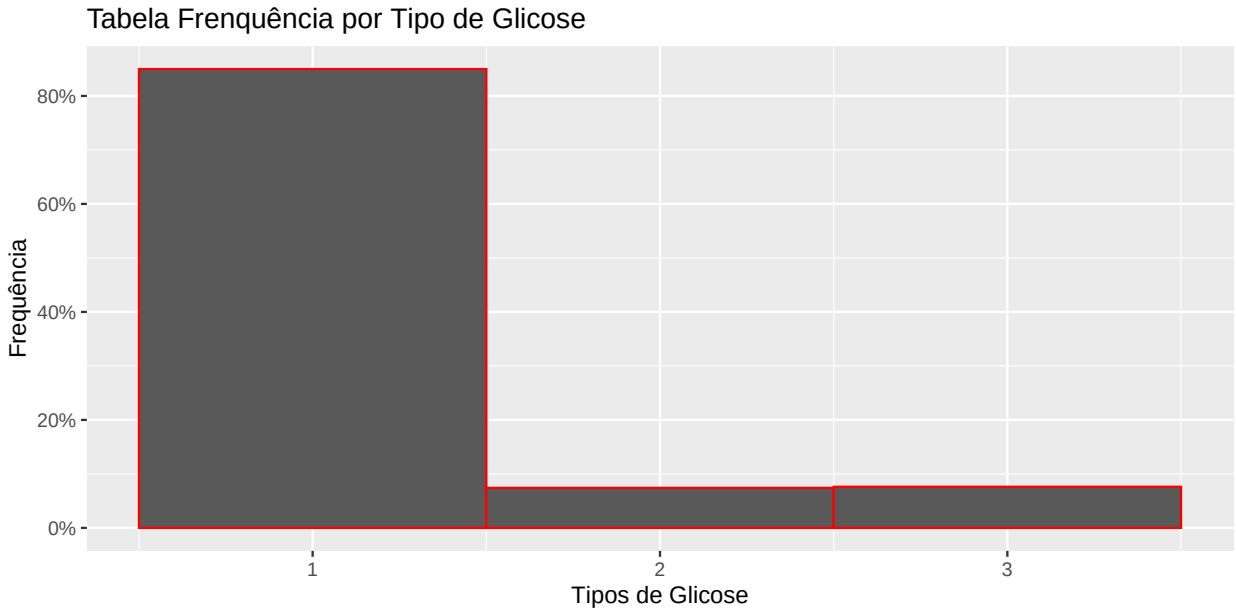
```
1     2     3
```

59479 5190 5331

```
fs.gluc = round(count.gluc/sum(count.gluc)*100)
fs.gluc
```

1 2 3 85 7 8

```
ggplot(cardio, aes(x = gluc)) +
  geom_histogram(aes(y = stat(count.gluc) / sum(count.gluc)), bins = 3, color="red") +
  scale_y_continuous(labels = scales::percent)+
  labs(title = 'Tabela Frenquência por Tipo de Glicose',
       y = 'Frequência', x = 'Tipos de Glicose')
```



Na análise do alcool, vemos pelo gráfico que apenas 5% dos paciêntes ingerem alcool.

```
count.alco=table(as.factor(cardio$alco))
count.alco
```

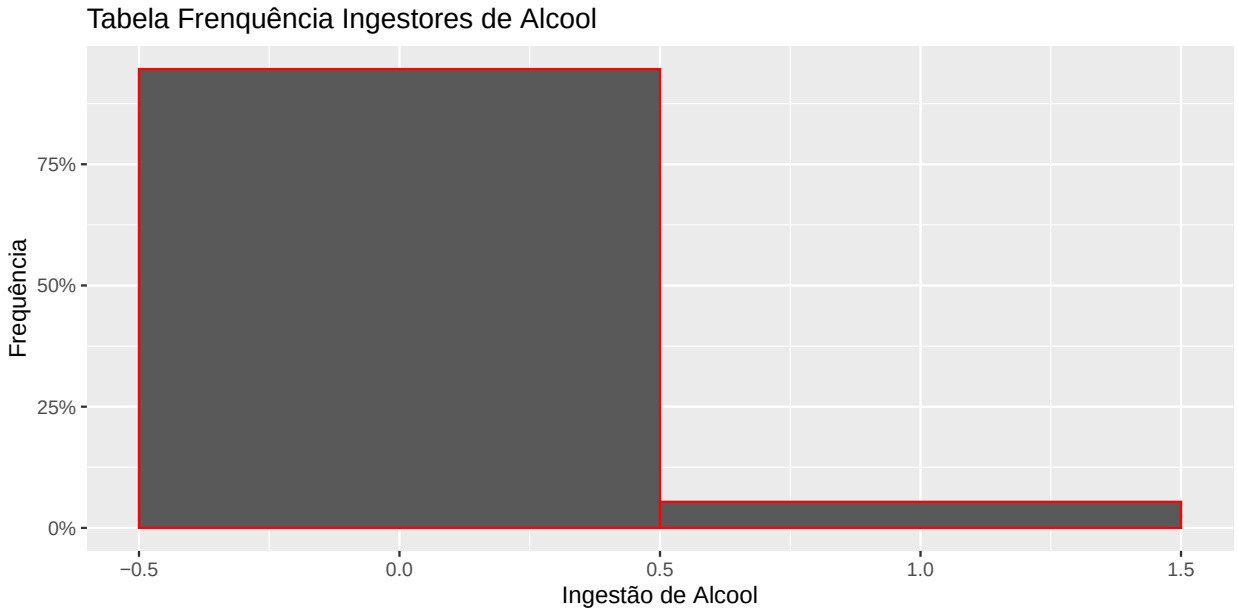
```
0     1
```

66236 3764

```
fs.alco = round(count.alco/sum(count.alco)*100)
fs.alco
```

0 1 95 5

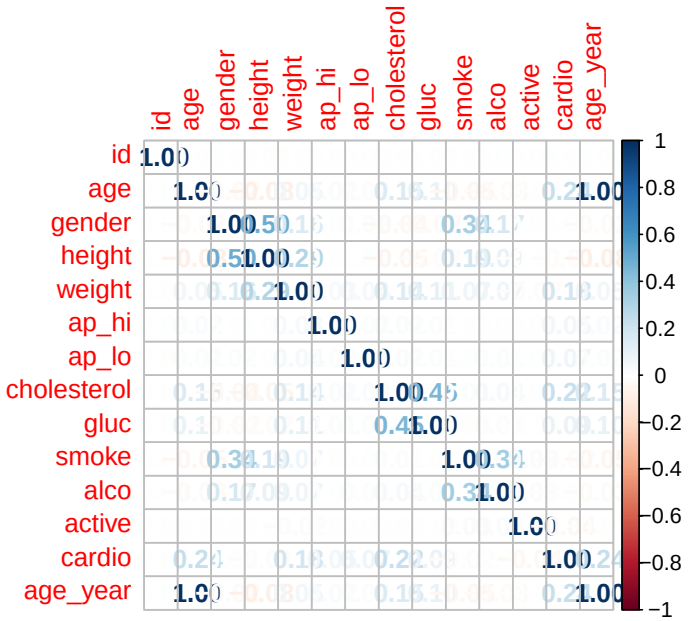
```
ggplot(cardio, aes(x = alco)) +
  geom_histogram(aes(y = stat(count.alco) / sum(count.alco)), bins = 2, color="red") +
  scale_y_continuous(labels = scales::percent)+
  labs(title = 'Tabela Frenquência Ingestores de Alcool',
       y = 'Frequência', x = 'Ingestão de Alcool')
```



Algumas Correlações

Afim de saber mais sobre as relações entre as variáveis aqui seram feitas algumas correlações de variáves explicativas com a nossa variável dependente.

```
corrplot(cor(cardio), method ='number')
```



Observamos que nossa variável dependente apresenta maior grau de correlação positiva com as varipaveis height (0.29), cardio(0.18) e gender(0.16).

Modelo de Regressão Múltipla

O primeiro modelo a ser usado será o de regressão múltipla, onde vamos observar as variáveis que mais impactam no peso dos pacientes.

```
lm_cardio<-lm(weight~height+cholesterol+gluc+smoke+alco+active+age_year,data=cardio)
summary(lm_cardio)
```

```
##
## Call:
## lm(formula = weight ~ height + cholesterol + gluc + smoke + alco +
##     active + age_year, data = cardio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -69.534  -9.164  -2.359   7.080 143.039
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.967526   1.158696 -19.822  < 2e-16 ***
## height        0.524073   0.006373  82.232  < 2e-16 ***
## cholesterol   2.634356   0.085113  30.951  < 2e-16 ***
## gluc          1.265061   0.100285  12.615  < 2e-16 ***
## smoke         0.077984   0.194793   0.400    0.689
## alco          2.258916   0.241593   9.350  < 2e-16 ***
## active       -0.586941   0.128927  -4.552 5.31e-06 ***
## age_year      0.116761   0.007694  15.175  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.54 on 69992 degrees of freedom
## Multiple R-squared:  0.1157, Adjusted R-squared:  0.1156
## F-statistic:  1308 on 7 and 69992 DF,  p-value: < 2.2e-16
```

Pelo resultado da regressão observamos que a única variável que não é estatisticamente significante é smoke. O intercepto foi de -22.96, as variáveis Height e age_year foram de, respectivamente, 0.5240 e 0.1167 o que nos permite interpretar que em média o peso aumenta com a altura e com a maior idade dos indivíduos. As outras variáveis por serem binárias nos permite uma interpretação de por exemplo: o indivíduo que apresenta Cholesterol terá aumento de em média 2.6343 no seu peso. Vale chamar atenção que a única variável que representa redução de peso é quando o indivíduo pratica atividade fisíca.

Modelo de Regressão Quantílica

```
cardio_rq<-rq(lm_cardio, tau = c(0.25,0.75), data=cardio)
summary(cardio_rq)
```

```
##
## Call: rq(formula = lm_cardio, tau = c(0.25, 0.75), data = cardio)
##
## tau: [1] 0.25
##
## Coefficients:
##             Value     Std. Error t value   Pr(>|t|)
## (Intercept) -51.37078   0.95143  -53.99345   0.00000
## height        0.65263   0.00520  125.38859   0.00000
## cholesterol   1.75219   0.08645   20.26837   0.00000
## gluc          0.46828   0.09887    4.73650   0.00000
```

```
## smoke        -0.21137   0.18539   -1.14017   0.25422
## alco          1.39710   0.22325    6.25809   0.00000
## active       -0.03902   0.10761   -0.36258   0.71692
## age_year      0.11234   0.00612   18.34840   0.00000
##
## Call: rq(formula = lm_cardio, tau = c(0.25, 0.75), data = cardio)
##
## tau: [1] 0.75
##
## Coefficients:
##             Value     Std. Error t value   Pr(>|t|)
## (Intercept) -13.64405   1.84153   -7.40907   0.00000
## height        0.49075   0.01006   48.79793   0.00000
## cholesterol   3.60262   0.13276   27.13676   0.00000
## gluc          1.89125   0.16974   11.14211   0.00000
## smoke         0.16443   0.27339    0.60145   0.54755
## alco          2.64183   0.33168    7.96505   0.00000
## active       -0.77847   0.20624   -3.77452   0.00016
## age_year      0.14055   0.01216   11.56206   0.00000
```

Os resultados da regressão quantil não são muito diferentes da regressão multipla. As variáveis continuaram com os sinais de impacto positivo, de uamento do peso.

Gráfico dos Resultados

```
cardio_rqsq<-rq(lm_cardio, tau = seq(0.05, 0.95, by = 0.05), data=cardio)
cardio_rqs<-summary(cardio_rqsq)
plot(cardio_rqs)
```