Universidade de Brasília
Instituto de Exatas
Departamento de Estatística

Análise de Dados Categorizados
Trabalho Final

Carolina Musso 18/0047850
Juliana Magalhães Rosa 18/0020935

Professor(a): Maria Teresa Leão

Brasília
1/2023

Sumário

1 Introdução e Objetivos

A habilidade e a possibilidade de poupar dinheiro pode estar relacionada a diversos fatores. Pettinger (2021) cita a idade, poder aquisitivo, desenvolvimento econômico e inflação como possíveis questões associadas.

Este trabalho visa avaliar fatores associados à posse de conta poupança por parte de pacientes de uma rede hospitalar. Isso será feito por meio da seleção de um modelo de regressão logística.

2 Metodologia

2.1 Variáveis

A variável resposta analisada nesse estudo é qualitativa nominal binária, "Conta popuança".

As variáveis explicativas (ou os fatores possivelmente associados) são:

- Idade: variável quantitativa discreta medida em anos;
- Status socioeconômico: variável qualitativa ordinal medida em 1 = superior, 2 = médio , 3 = inferior;
- Possui casa própria: variável qualitativa nominal binária, medida em 2 = não ou sim, mas ainda pagando financiamento e 2= sim e quitada;
- Setor da cidade: variável qualitativa nominal medida em 1 = setor A; 0= setor B.

2.2 Amostra

Para este trabalho, uma sub-amostra aleatória simples sem reposição de tamanho 100 foi selecionada a partir de uma amostra de 196 pacientes. Os IDS selecionados nesse trabalho foram: 2, 3, 4, 5, 6, 9, 13, 16, 18, 20, 21, 24, 27, 29, 32, 33, 35, 36, 40, 41, 42, 43, 47, 49, 53, 54, 55, 57, 58, 60, 65, 68, 69, 71, 73, 74, 76, 80, 81, 82, 83, 85, 89, 91, 92, 99, 100, 101, 102, 103, 104, 109, 110, 111, 113, 114, 115, 116, 118, 122, 128, 129, 130, 131, 134, 135, 136, 137, 138, 140, 143, 144, 146, 150, 153, 154, 158, 161, 162, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 179, 180, 182, 183, 184, 185, 187, 191, 192, 194, 196.

2.3 Análise

A regressão logística é um modelo estatístico utilizado para casos em que a variável resposta é categorizada. O mais comum é que essa variável seja binária, como é o caso do presente estudo. O funcionamento dessa técnica consiste em descrever a probabilidade de ocorrência de um evento, que nesse caso será a posse de poupança.

Assim, modela-se a média da variável resposta a partir da função Logística:

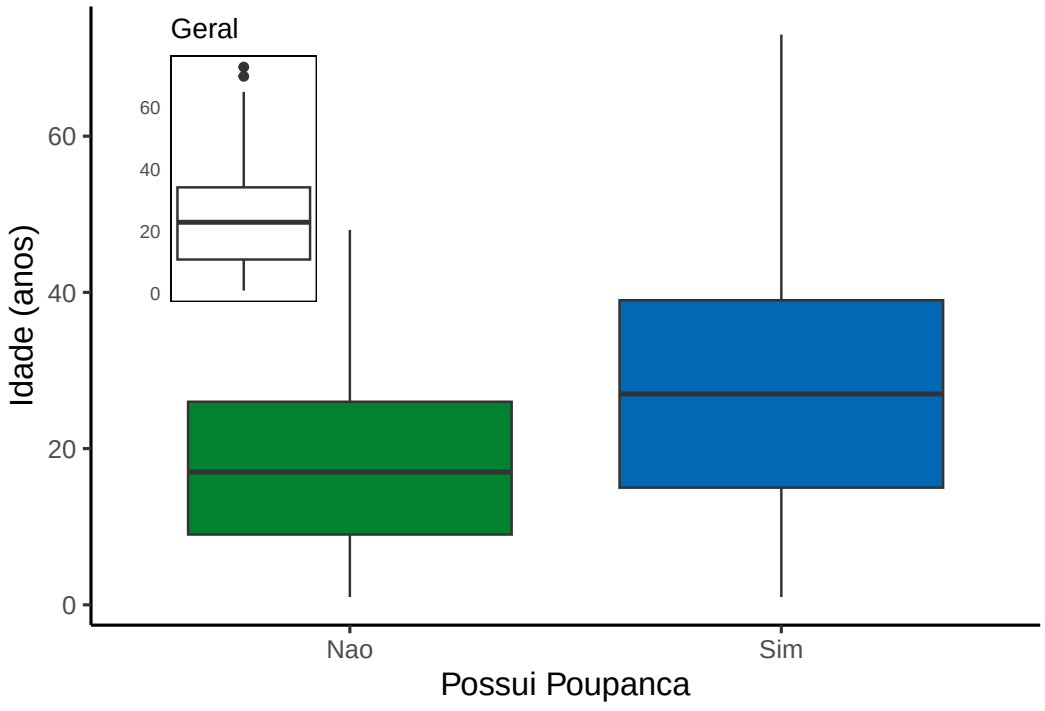
$$\pi(x_i) = \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_{k-1} x_{i(k-1)})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_{k-1} x_{i(k-1)})}$$

onde x_i é um vetor com os elementos x_{ij}, os quais representam possíveis valores das variáveis explicativas X_j; β_j é um parâmetro regressivo; k é o número de parâmetros do modelo com $j = 0, 1, 2, ..., k - 1$.

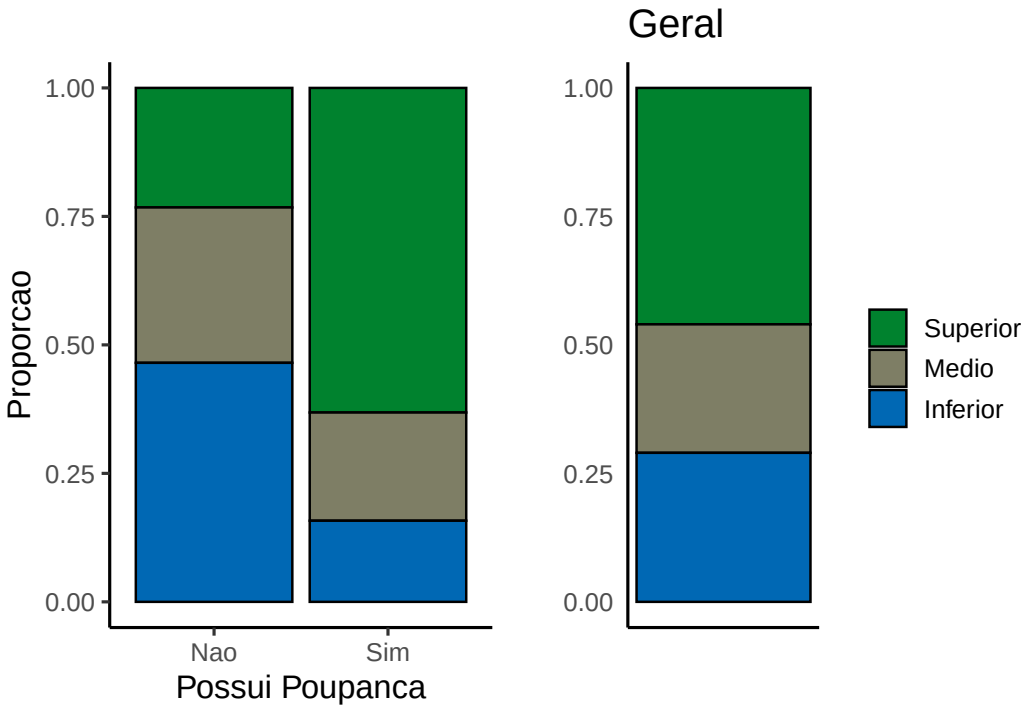
3 Resultados

3.1 Análise Descritiva Gráficos

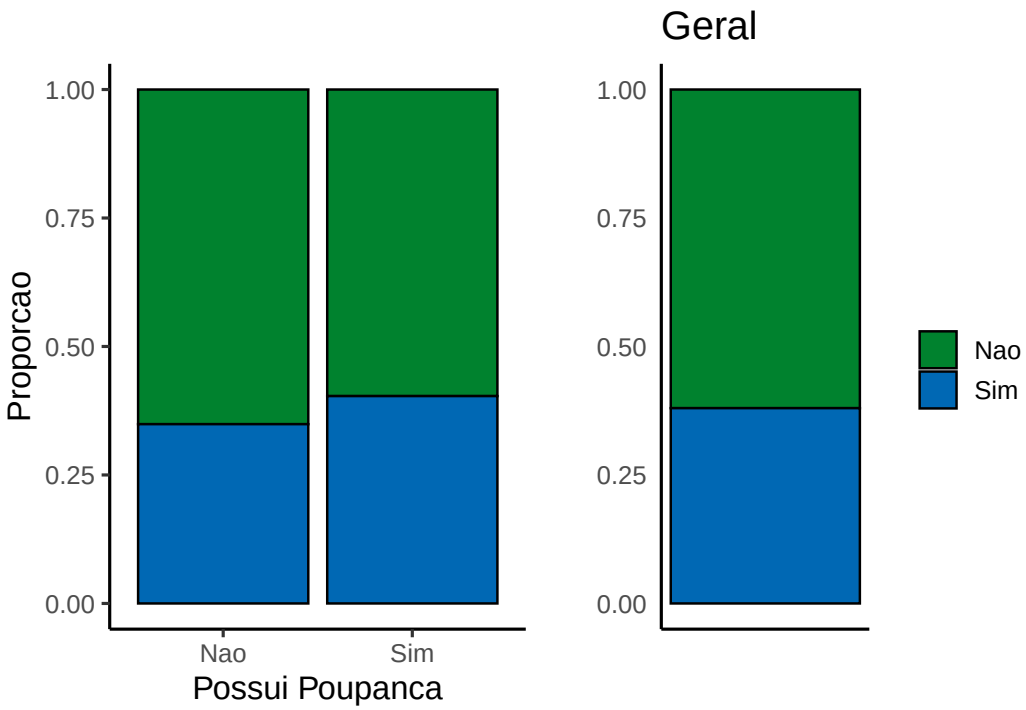
3.1.1 Distribuição de Idades dos Pacientes



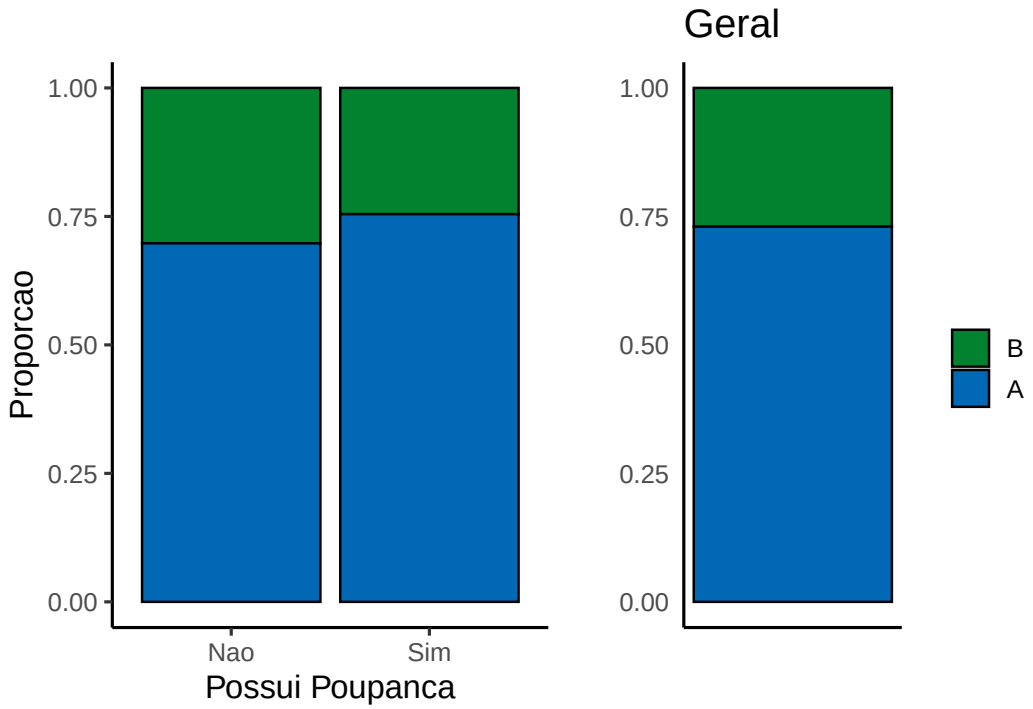
3.1.2 Proporção de Pacientes nos Status Socioeconômicos



3.1.3 Proporção de Pacientes com Casa Própria Quitada



3.1.4 Proporção de Pacientes nos Setores da Cidade



3.2 Análise Descritiva Tabelas

Tabela 1: Frequências dos Fatores Explicativos por Posse de Poupanca.

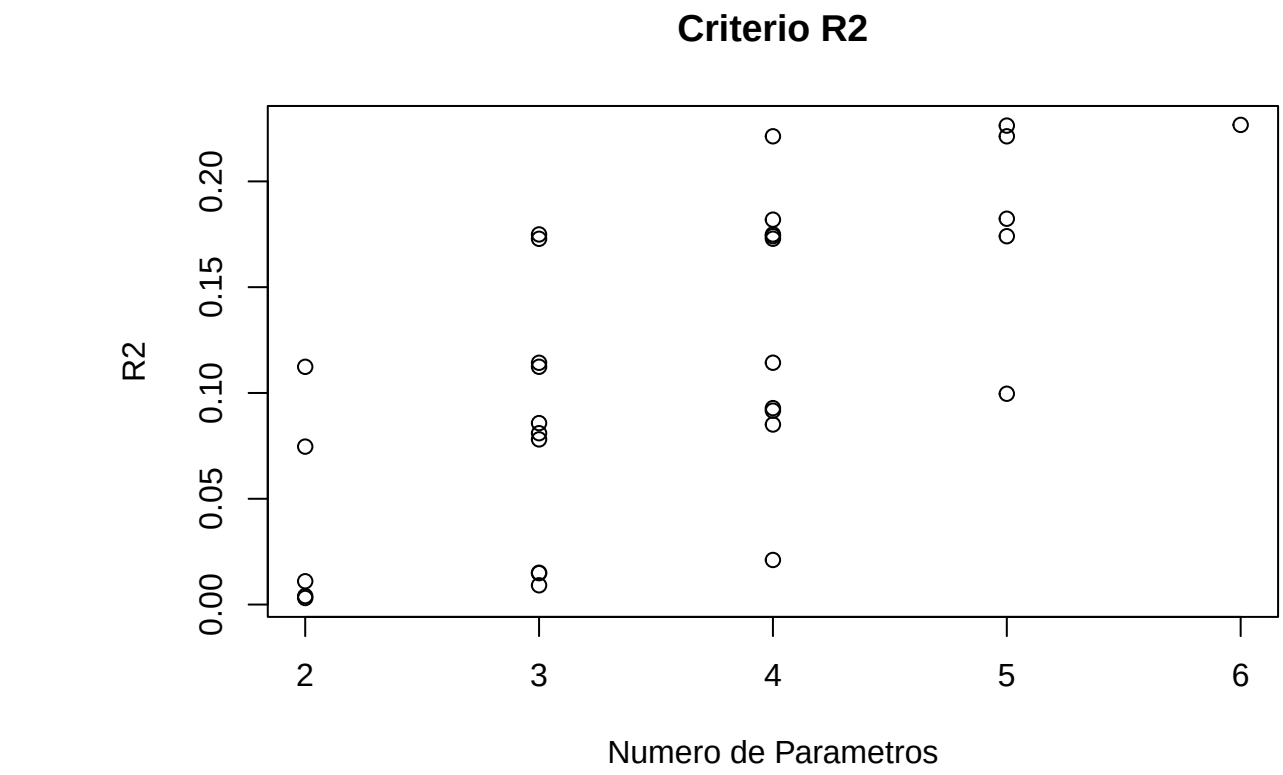
		Poupanca: Nao		Poupanca: Sim		
	Level	N	%	N	%	p.value
Status Socioeconomico	Superior	10	23.3	36	63.2	<0.001
	Medio	13	30.2	12	21.1	
	Inferior	20	46.5	9	15.8	
Casa Propria	Nao	28	65.1	34	59.6	0.678
	Sim	15	34.9	23	40.4	
Setor	B	13	30.2	14	24.6	0.650
	A	30	69.8	43	75.4	

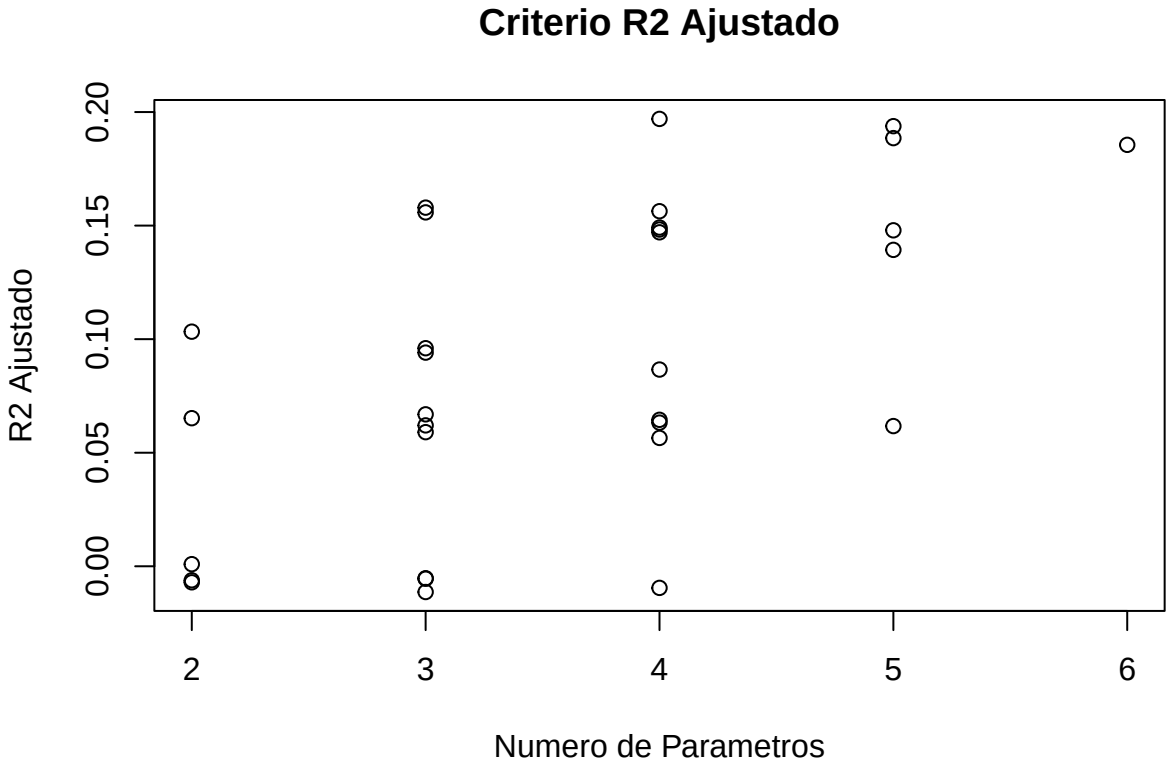
Tabela 2: Medidas Descritivas para Idade dos Pacientes por Posse de Poupanca.

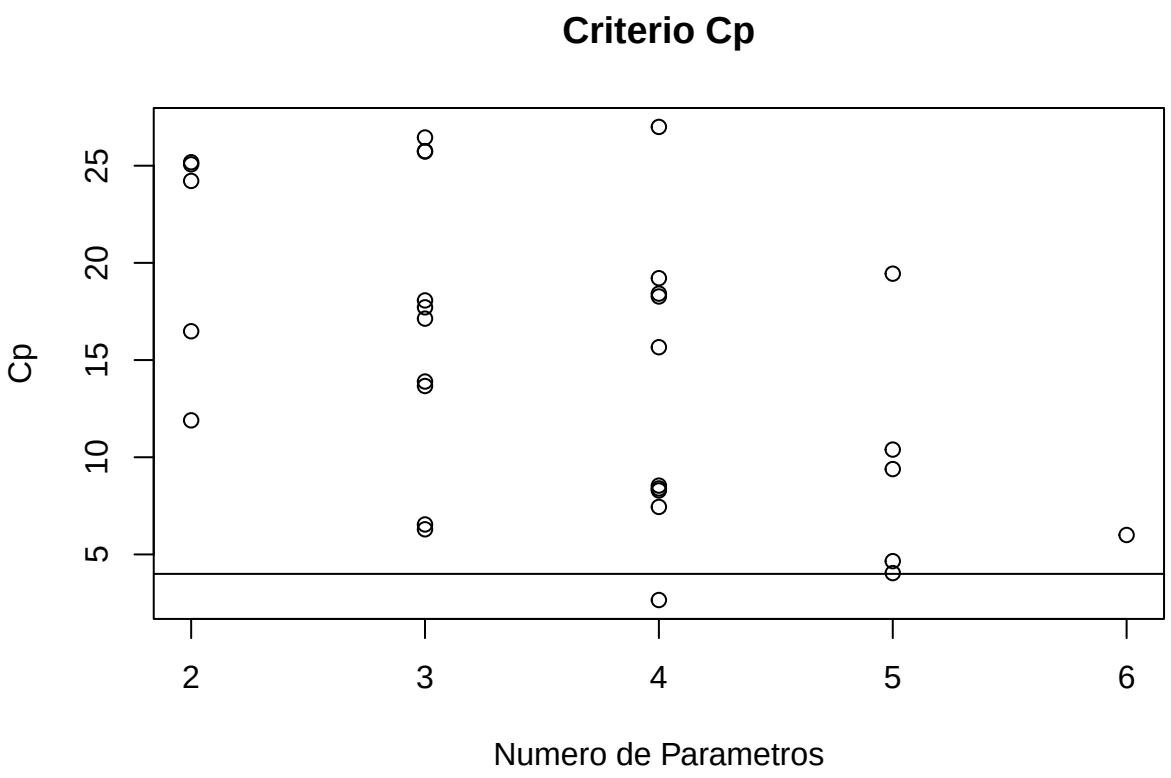
	Poupanca	N	Mean	SD	Min	Q1	Median	Q3	Max	p.value
Idade	Nao	43	19.30	12.59	1.00	9.00	17.00	26.00	48.00	0.004
	Sim	57	28.84	19.36	1.00	15.00	27.00	39.00	73.00	

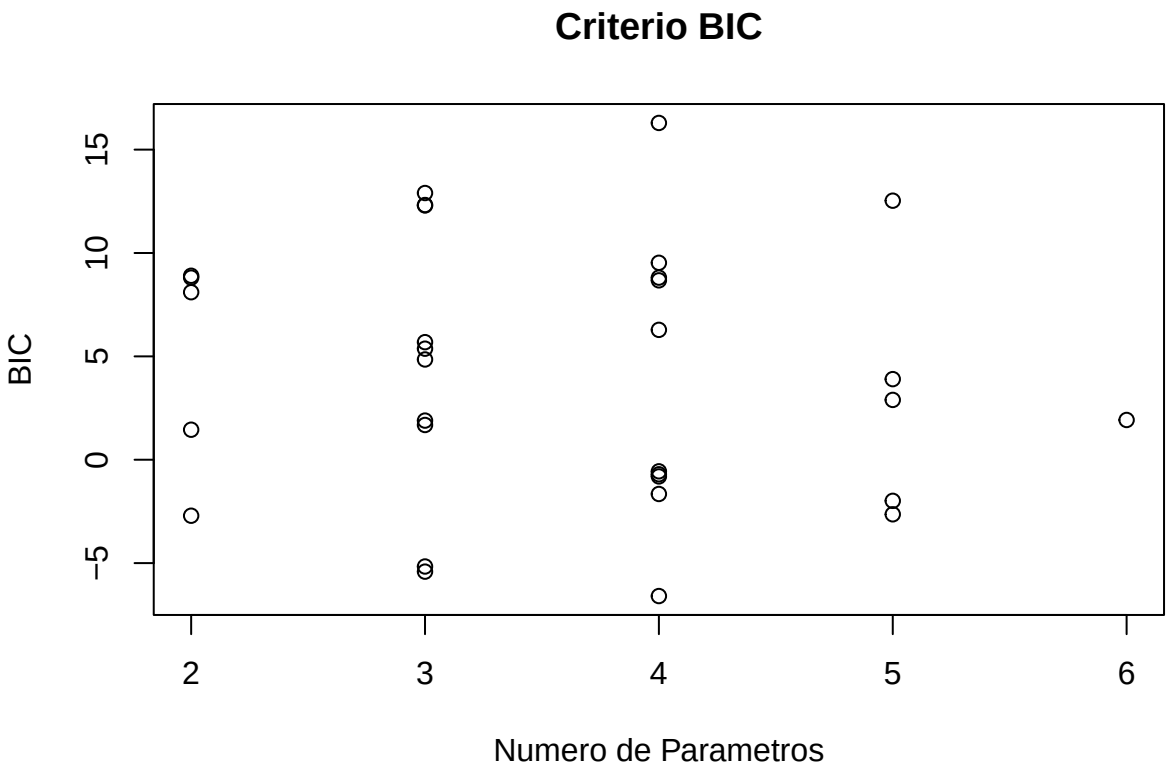
Parametros e interpretacao Modelo e residuos

3.3 Modelagem



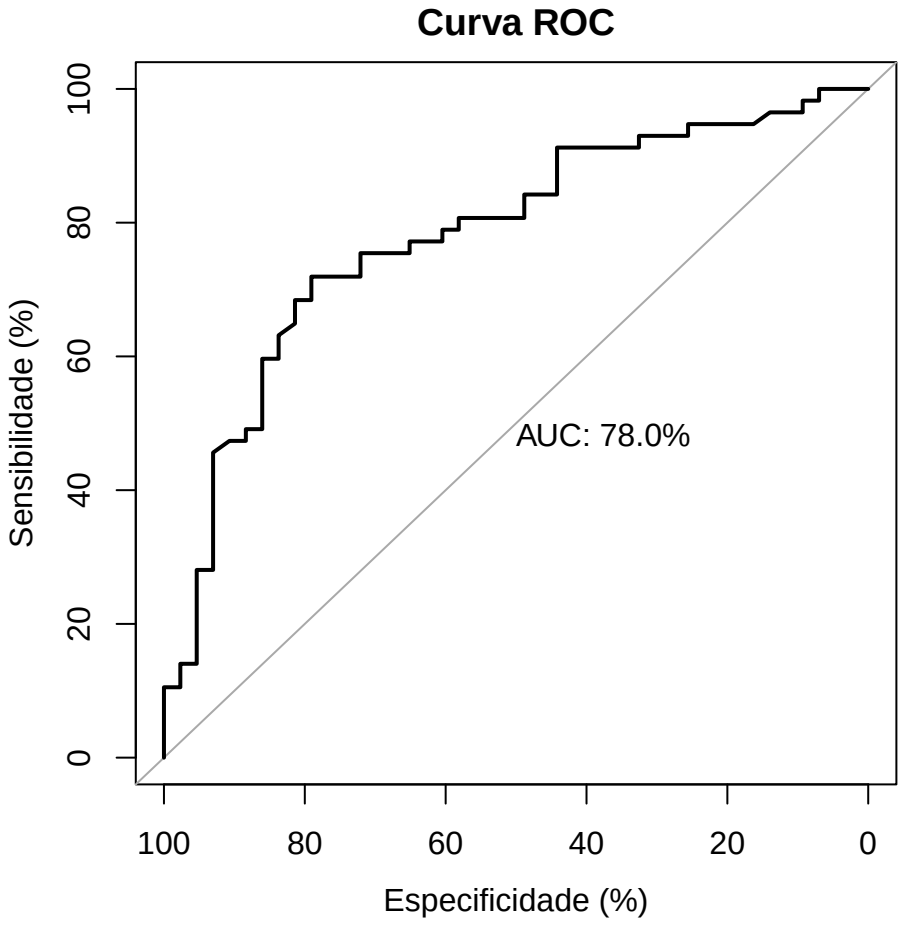


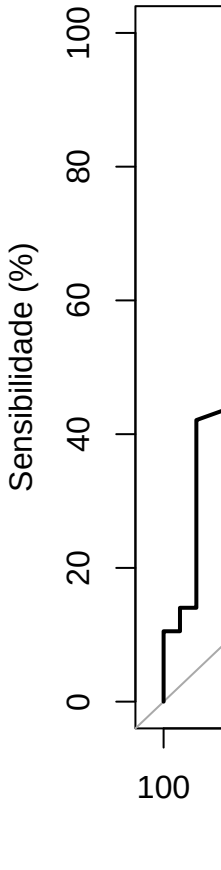




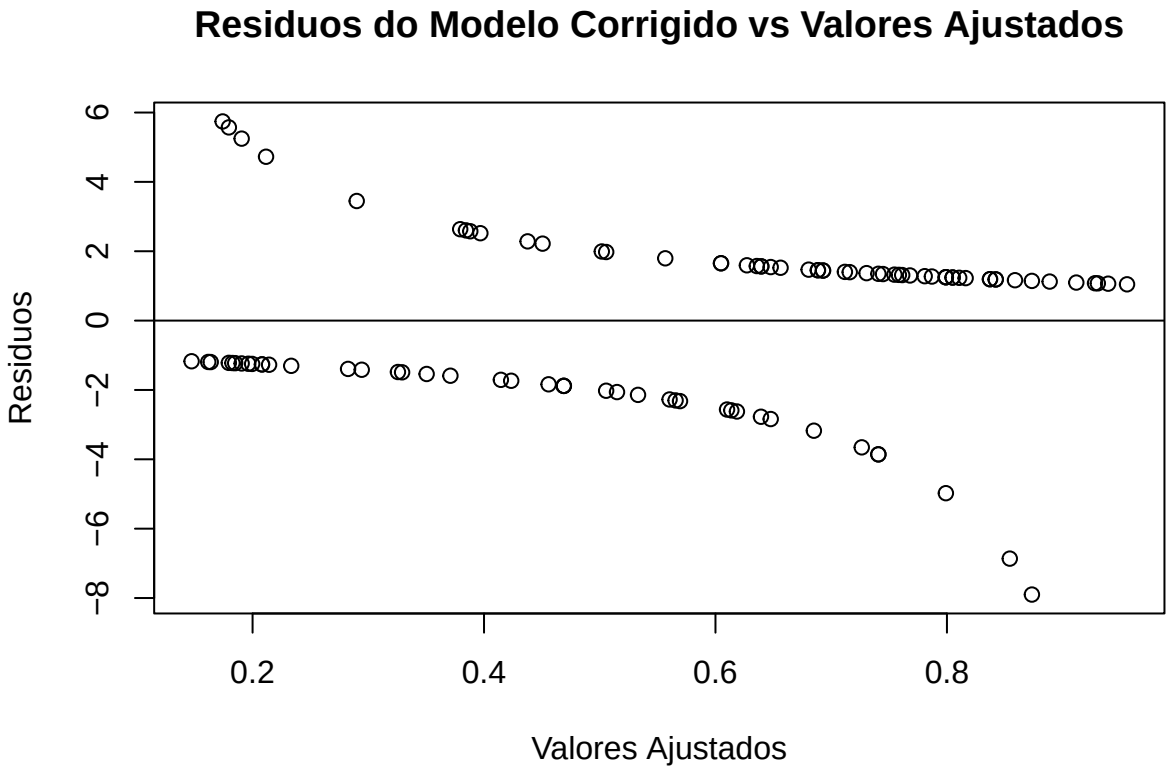

```
##                          Idade    ‘Status Socioeconomico‘Medio
##                            "*"                             " "
```

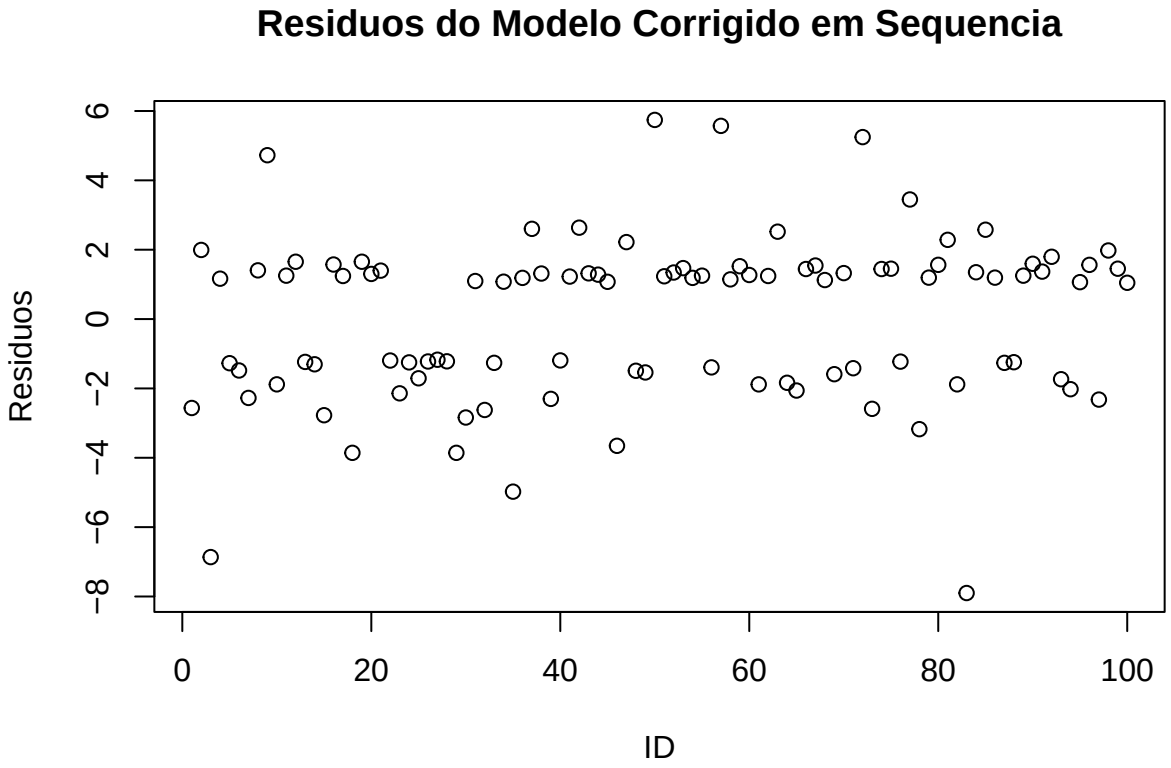
```
## ‘Status Socioeconomico‘Inferior                ‘Casa Propria‘Sim
##                                "*"                               " "
##                              SetorA
##                                " "

##                               Idade     ‘Status Socioeconomico‘Medio
##                                "*"                               "*"
## ‘Status Socioeconomico‘Inferior                ‘Casa Propria‘Sim
##                                "*"                               " "
##                              SetorA
##                                " "

## (Intercept)       idade      status        casa       setor
##  1.30615163  0.03673061 -1.02775114  0.11031818 -0.49326807

## (Intercept)       idade      status
##  1.42771976  0.03374204 -1.04880728
```

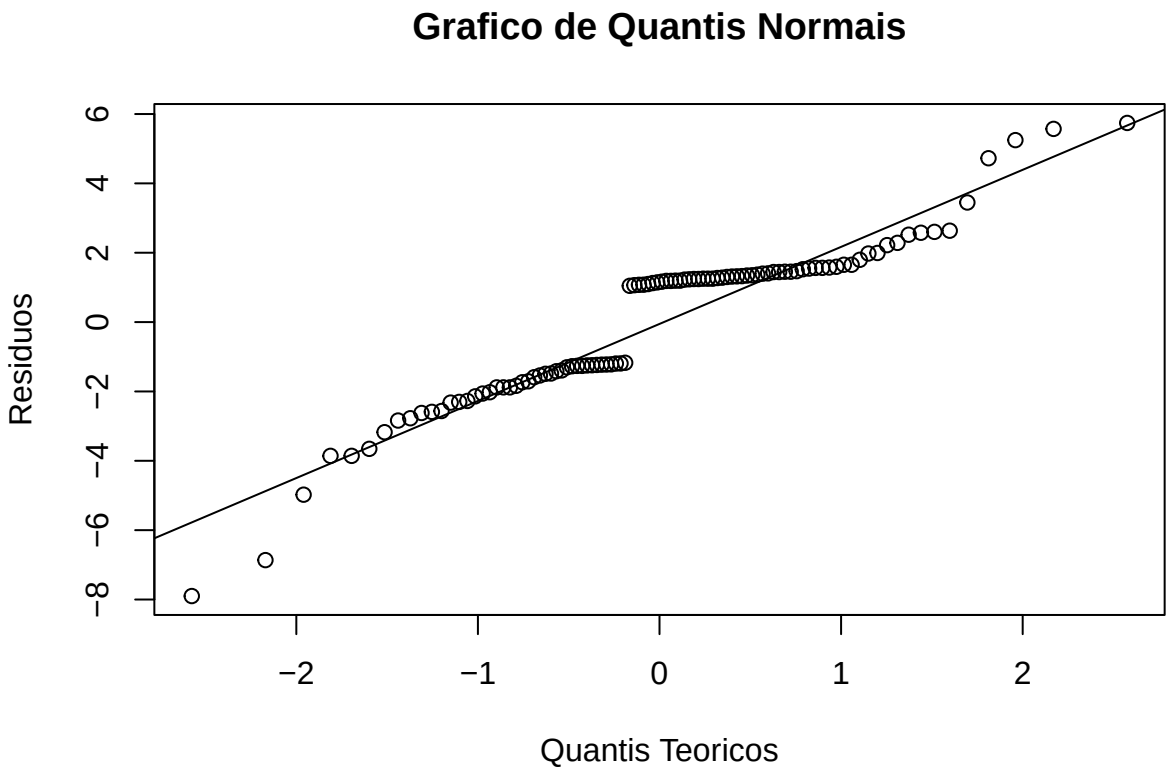




Como funciona a análise de resíduos para a regressão logística?
Fiz igual à linear, mas acho que não é assim...
De toda forma, falta reproduzir para o modelo 2 ainda.








```
##
##  studentized Breusch-Pagan test
```

```
##
## data:  mod1
## BP = 6.5784, df = 4, p-value = 0.1599

##
##  Shapiro-Wilk normality test
##
## data:  mod1$residuals
## W = 0.92427, p-value = 2.388e-05
```

4 Conclusão

5 Apêndice

```
knitr::opts_chunk$set(echo = FALSE,
                      warning = FALSE,
                      message = FALSE)

#rm(list = ls()) #will clear all objects includes hidden objects.
#options(rstudio.help.showDataPreview = FALSE)
# Carregando bibliotecas ---------
pacman::p_load(tidyverse, dplyr, rio, papeR, patchwork, kableExtra, pROC)


# Bases -----

dados <- import ("data/dados_trabalho.xlsx")

# amostra
set.seed(42)
amostra <- slice_sample(dados, n=100)

## Tratamento ----

names(amostra) <- c("ID", "idade", "status", "casa", "setor", "save" )

amostra_trat <- amostra %>%
  mutate(status=factor(status,
                       labels=c("Superior", "Medio", "Inferior")),
         casa=factor(casa, labels=c("Nao", "Sim")),
         setor=factor(setor, levels=c(1,0), labels=c("B", "A")),
        save=factor(save, labels=c("Nao", "Sim"))) %>%
  as.data.frame()
sort(amostra$ID)
idade_by <- amostra_trat %>%
  ggplot(aes(x=save, y=idade, fill=save))+
   geom_boxplot()+
   scale_fill_manual(values=c("#00822E", "#0068B4"))+
   #geom_jitter(color="black", size=0.4, alpha=0.9) +
   theme_classic(base_size = 14)+
```

```
   theme(legend.position = "none")+
   labs(x= "Possui Poupanca", y= "Idade (anos)")



idade <- amostra_trat %>%
   ggplot(aes(y=idade))+
   geom_boxplot()+
  #geom_jitter( color="black", size=0.4, alpha=0.9) +
  theme_classic(base_size = 10)+
   theme(legend.position = "none",
         axis.text.x = element_blank(),
         axis.ticks = element_blank(),
         axis.line.x = element_blank(),
          axis.line.y = element_blank(),
         panel.border = element_rect(colour = "black", fill=NA))+
   labs(x= "", y= "")+
  ggtitle("Geral")
 idade_by +  inset_element(idade, left = 0.01, bottom = 0.45, right = 0.25, top = 1)
status_by <- amostra_trat %>%
  ggplot(aes(x=save,fill=factor(status)))+
   geom_bar(position="fill", color="black")+
   scale_fill_manual(values=c("#00822E", "#7E7E65", "#0068B4"))+
   #geom_jitter(color="black", size=0.4, alpha=0.9) +
   theme_classic(base_size = 14)+
   theme(legend.position = "none")+
   labs(x= "Possui Poupanca", y= "Proporcao")



status <-   amostra_trat %>%
  ggplot(aes(x=1,fill=factor(status)))+
  geom_bar(position="fill", color="black")+
  scale_fill_manual("",values=c("#00822E", "#7E7E65", "#0068B4"))+
  theme_classic(base_size = 14)+
  theme( axis.text.x = element_blank(),
        axis.ticks = element_blank())+
  labs(x= "", y= "", title="Geral")
```

```
 status_by +  status+  plot_layout(widths = c(2, 1))
casa_by <- amostra_trat %>%
  ggplot(aes(x=save,fill=factor(casa)))+
   geom_bar(position="fill", color="black")+
   scale_fill_manual("", values=c("#00822E", "#0068B4"))+
   #geom_jitter(color="black", size=0.4, alpha=0.9) +
   theme_classic(base_size = 14)+
   theme(legend.position = "none")+
   labs(x= "Possui Poupanca", y= "Proporcao")



casa<-   amostra_trat %>%
  ggplot(aes(x=1,fill=factor(casa)))+
  geom_bar(position="fill", color="black")+
  scale_fill_manual("", values=c("#00822E", "#0068B4"))+
  theme_classic(base_size = 14)+
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank())+
  labs(x= "", y= "")+
  ggtitle("Geral")
 casa_by +  casa +   plot_layout(widths = c(2, 1))
setor_by <- amostra_trat %>%
  ggplot(aes(x=save,fill=factor(setor)))+
   geom_bar(position="fill", color="black")+
   scale_fill_manual("", values=c("#00822E", "#0068B4"))+
   #geom_jitter(color="black", size=0.4, alpha=0.9) +
   theme_classic(base_size = 14)+
   theme(legend.position = "none")+
   labs(x= "Possui Poupanca", y= "Proporcao")



setor <-   amostra_trat %>%
  ggplot(aes(x=1,fill=factor(setor)))+
  geom_bar(position="fill", color="black")+
  scale_fill_manual("", values=c("#00822E", "#0068B4"))+
  theme_classic(base_size = 14)+
```

```
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank())+
  labs(x= "", y= "")+
  ggtitle("Geral")
 setor_by +  setor + plot_layout(widths = c(2, 1))
names(amostra_trat) <- c("ID", "Idade", "Status Socioeconomico", "Casa Propria", "Setor"


Tab1 <- papeR::summarize(amostra_trat[,-1],
                 type = "factor", group = "Poupanca")

xtable(Tab1, caption="Frequências dos Fatores Explicativos por Posse de Poupanca.")
Tab2 <- papeR::summarize(amostra_trat[,-1],
                 type = "numeric", group = "Poupanca")

xtable(Tab2, width = rep("0.5in", 11), caption="Medidas Descritivas para Idade dos Pacie
#Seleção exaustiva
library(leaps)
sele1<-regsubsets(Poupanca~.,data=amostra_trat[, -1], nbest=10)
n_parametros<-as.numeric(rownames(summary(sele1)$which))+1
s <- summary(sele1)

#Critério R^2
plot(n_parametros,summary(sele1)$rsq, main="Criterio R2", xlab="Numero de Parametros", y

#Critério R^2 ajustado
plot(n_parametros,summary(sele1)$adjr2, main="Criterio R2 Ajustado", xlab="Numero de Par

#Critério Cp
plot(n_parametros,summary(sele1)$cp, main="Criterio Cp", xlab="Numero de Parametros", yl
#plot(n_parametros, summary(sele1)£cp, ylim=c(0, 10), main="Critério Cp Ampliado", xla
abline(h=4)

#Critério bic
plot(n_parametros,summary(sele1)$bic, main="Criterio BIC", xlab="Numero de Parametros",

#Modelo com 4 parâmetros (3 variáveis explicativas) parece melhor
#O de 3 parâmetros poderia ser bom também
```

```
#Modelo 6 é o melhor com 3 parâmetros (2 variáveis explicativas)
#Modelo 16 é o melhor com 4 parâmetros (2 variáveis explicativas)
s$outmat[6, ]
s$outmat[16, ]

#com 3 parâmetros fica só com uma das dummies de status socioeconômico
#faz mais sentido pegar de 4 parâmetros

#o melhor modelo de 4 parâmetros é o que inclui status (as 2 dummies) e idade -->
mod1 <- glm(data=amostra, save ~ idade+status+casa+setor, binomial(link = "logit"))
mod2 <- glm(data=amostra, save ~ idade+status, binomial(link = "logit"))
#modelo saturado
mod1$coefficients
#modelo selecionado
mod2$coefficients
#ROC --> avaliação do modelo
r <- roc( amostra$save,as.vector(fitted.values(mod1)) ,
          grid=TRUE, print.auc = TRUE, percent=T)


plot(r ,xlim=c(100,0),ylim=c(0,100), asp = NA, legacy.axes = F, xlab="Especificidad

rr <- roc( amostra$save,as.vector(fitted.values(mod2)) ,
          grid=TRUE, print.auc = TRUE, percent=T)
plot(rr ,xlim=c(100,0),ylim=c(0,100), asp = NA, legacy.axes = F, xlab="Especificida
## MODELO 1
#gráficos residuais
plot(mod1$fitted.values,mod1$residuals, main="Residuos do Modelo Corrigido vs Valor
abline(h=0)
plot(mod1$residuals, main="Residuos do Modelo Corrigido em Sequencia", xlab="ID", y
qqnorm(mod1$residuals, main="Grafico de Quantis Normais", xlab="Quantis Teoricos",
qqline(mod1$residuals)
#testes para resíduos
library(lmtest)
bptest(mod1)
shapiro.test(mod1$residuals) #rejeita
```

Referências

PETTINGER, T. *Factors that influence saving levels*. 2021. ⟨https://www.economicshelp.org/blog/146244/economics/factors-that-influence-saving-levels/⟩. Accessed: 2023-07-10.