



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	4
2 Metodologia	4
2.1 Variáveis	4
2.2 Amostra	4
2.3 Análise	5
3 Resultados	6
3.1 Análise Descritiva: Gráficos	6
3.1.1 Distribuição de Idades dos Pacientes	6
3.1.2 Proporção de Pacientes nos Status Socioeconômicos	6
3.1.3 Proporção de Pacientes com Casa Própria Quitada	7
3.1.4 Proporção de Pacientes nos Setores da Cidade	8
3.2 Análise Descritiva: Tabelas	9
3.3 Seleção de Variáveis	10
3.4 Modelos selecionados	11
3.4.1 Modelo 1	11
3.4.2 Modelo 2	11
3.4.3 Modelo 3	12
3.4.4 Modelo 4	12
3.5 Avaliação dos modelos	13
3.5.1 Diagnóstico do modelo	15
4 Conclusão	17
5 Apêndice	18

1 Introdução e Objetivos

A habilidade e oportunidade de poupar dinheiro podem estar relacionadas a uma variedade de fatores. Segundo Pettinger (2021), a idade, o poder aquisitivo, o desenvolvimento econômico e a inflação são possíveis questões associadas.

O objetivo deste estudo é avaliar os fatores associados à posse de conta poupança entre pacientes de uma rede hospitalar. Para alcançar esse propósito, será selecionado um modelo de regressão logística para análise.

2 Metodologia

2.1 Variáveis

A variável resposta analisada nesse estudo é qualitativa nominal binária, “Conta poupança”.

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 1 = 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 sorteados 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 amplamente utilizado para analisar casos em que a variável resposta é categórica, sendo comum o uso em situações binárias, como é o caso deste estudo, em que a variável em análise é a posse de poupança.

O funcionamento dessa técnica baseia-se em descrever e modelar a probabilidade de ocorrência de um determinado evento, no caso, a probabilidade de um paciente possuir poupança (HOSMER-JR.; LEMESHOW; STURDIVANT, 2013). Ao considerar múltiplas variáveis explicativas, como idade, poder aquisitivo e outras mencionadas anteriormente, a regressão logística permitirá avaliar a influência desses fatores na probabilidade de os pacientes e seus familiares possuírem conta poupança. Dessa forma, é possível obter *insights* importantes sobre os fatores associados à posse de poupança nesse contexto específico.

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

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

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 $j = 0, 1, 2, \dots, k - 1$.

3 Resultados

3.1 Análise Descritiva: Gráficos

3.1.1 Distribuição de Idades dos Pacientes

Na Figura 1, podemos observar a distribuição das idades em cada grupo de interesse, ou seja, aqueles que possuem ou não uma poupança. Notamos que o grupo sem poupança tende a ser composto por pessoas mais jovens, com uma mediana abaixo de 20 anos, enquanto aqueles com poupança apresentam uma mediana acima de 20 anos, chegando a valores próximos de 60 anos. Ao analisarmos a amostra como um todo, observamos que a idade mediana é ligeiramente superior a 20 anos e existem pessoas mais velhas que chegam a ser consideradas *outliers* nas idades.

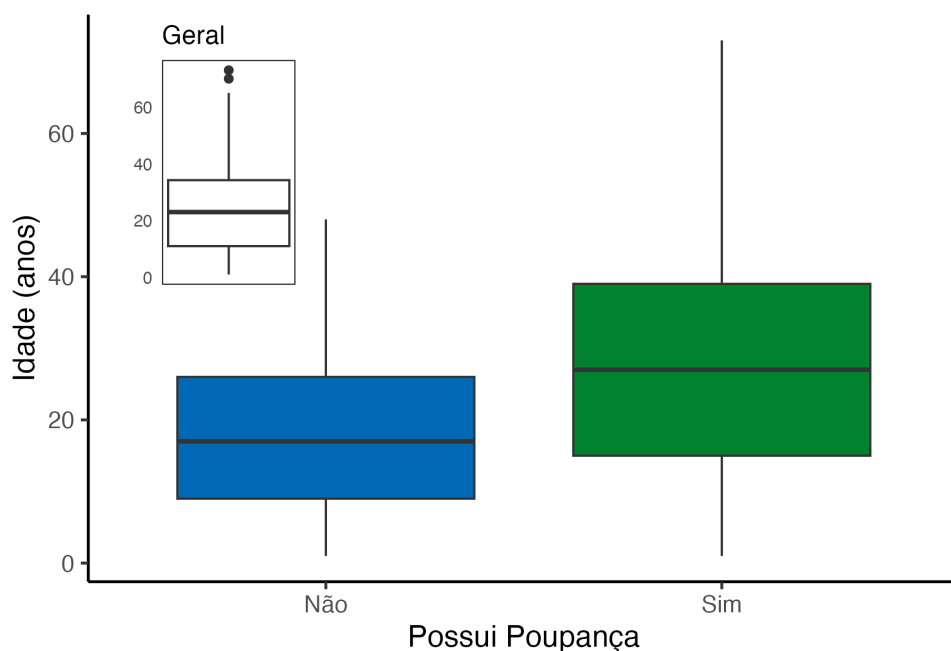


Figura 1. Diagrama de caixas para distribuição de idade entre os grupos sem (azul) e com (verde) poupança. Para comparação, a distribuição de idade geral foi adicionada no canto superior esquerdo.

3.1.2 Proporção de Pacientes nos Status Socioeconômicos

Ao analisarmos a amostra por grupo, levando em consideração a proporção de pessoas em diferentes níveis de status econômico (Figura 2), notamos que o grupo sem poupança apresenta uma maior proporção de pessoas com status econômico inferior em relação ao esperado na amostra total (representado no subgráfico “Geral”). Além disso,

o grupo com poupança tende a ser composto por pessoas com uma maior proporção de status econômico superior.

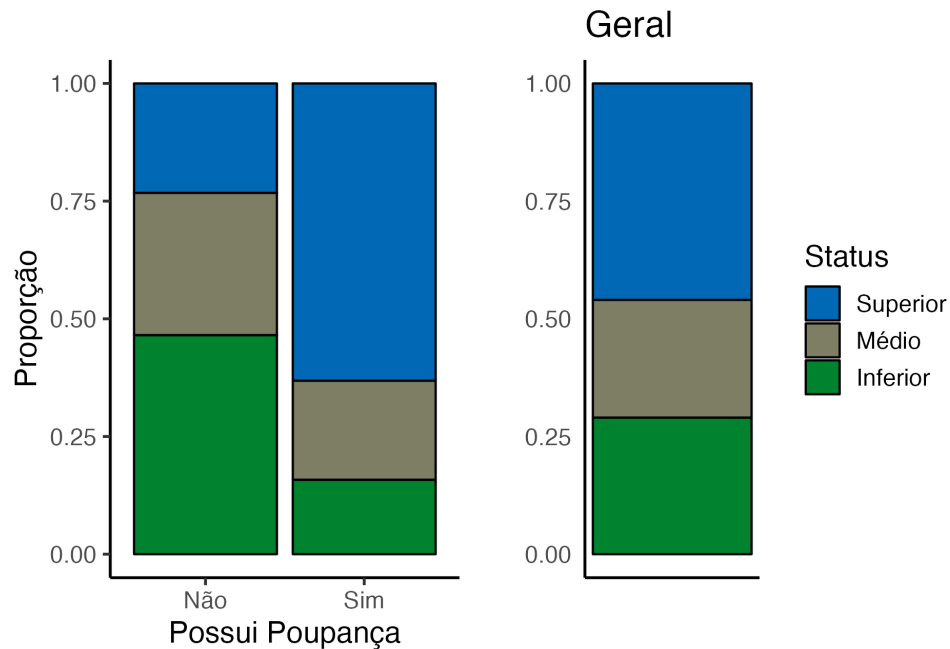


Figura 2. Proporção de pessoas com status Superior (verde), Médio (cinza) e Inferior (verde) para cada grupo (com ou sem poupança). Para comparação, a distribuição proporções geral foi adicionada ao lado direito.

3.1.3 Proporção de Pacientes com Casa Própria Quitada

Analisando a amostra por grupo, com base na proporção de pessoas que possuem ou não casa própria quitada (Figura 4), podemos observar que o grupo sem poupança apresenta uma proporção menor de pessoas com casa quitada em comparação ao grupo com poupança. De maneira geral, na amostra como um todo, constata-se que mais de 50% das pessoas não possuem casa própria quitada.

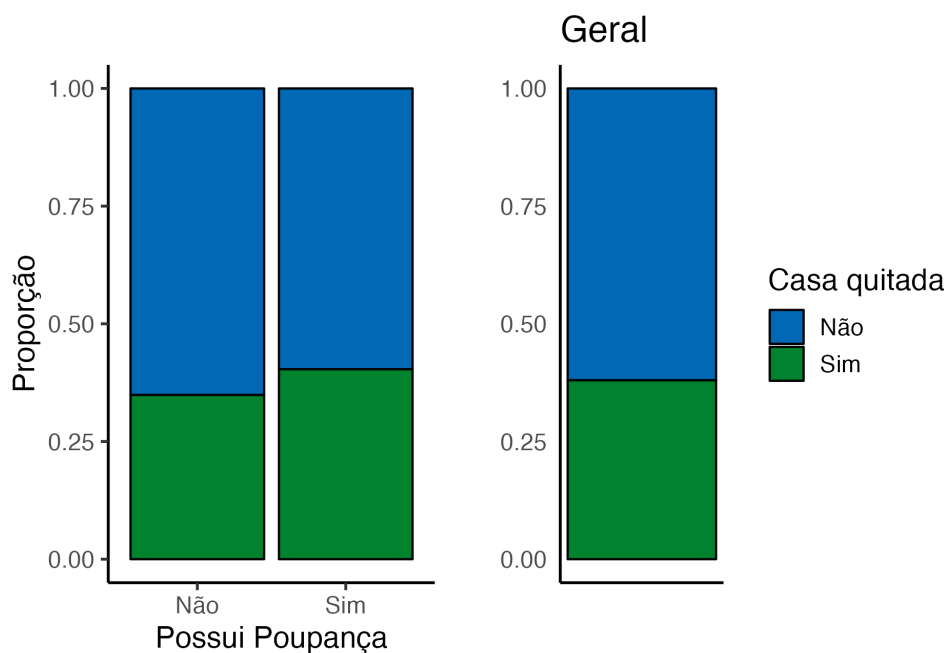


Figura 3. Proporção de pessoas com casa própria quitada (verde), ou sem casa própria/não quitada (azul) para cada grupo com ou sem poupança. Para comparação, a distribuição proporções geral foi adicionada ao lado direito.

3.1.4 Proporção de Pacientes nos Setores da Cidade

Ao analisarmos a amostra por grupo de acordo com a proporção de pessoas que residem no Setor A ou no Setor B (Figura 4), observamos que o grupo sem poupança possui uma proporção ligeiramente menor de residentes no Setor A em comparação com o grupo com poupança. De modo geral, na amostra como um todo, aproximadamente 75% das pessoas residem no Setor A.

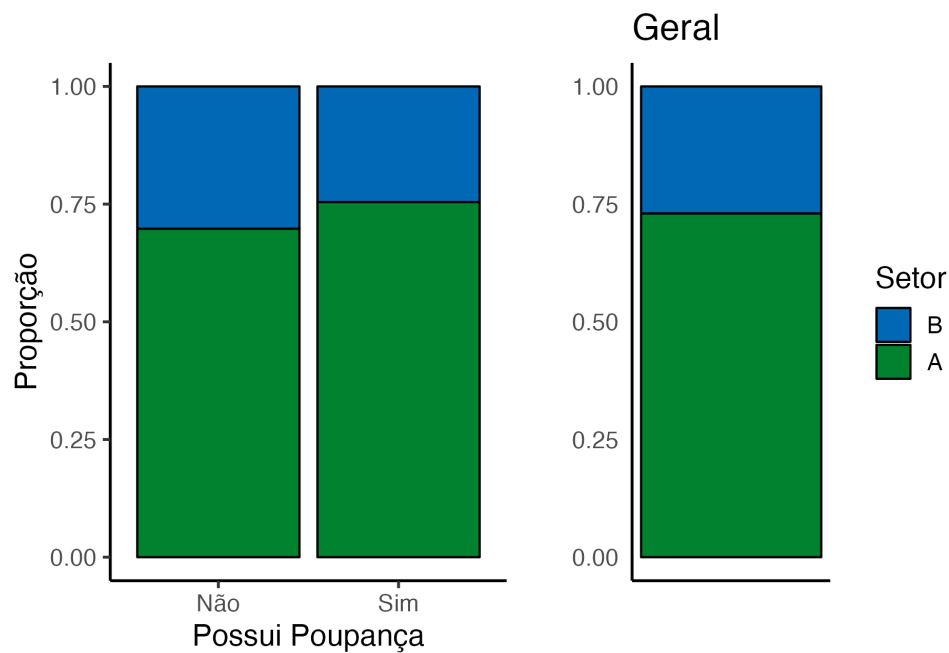


Figura 4. Proporção de pessoas que residem no setor A (verde) ou Setor B (azul), para cada grupo com ou sem poupança. Para comparação, a distribuição proporções geral foi adicionada ao lado direito.

3.2 Análise Descritiva: Tabelas

Pela Tabela 1, nota-se que, ao avaliar as associações entre cada variável explicativa com a resposta separadamente, o status socioeconômico é o único atributo que apresenta associação significativa com a posse de poupança.

Os p-valores obtidos são do Teste Qui-Quadrado, o qual testa a independência entre as variáveis como hipótese nula.

Tabela 1. Frequências relativas e absolutas das variáveis qualitativas por grupo Com ou Sem poupança. P-valor proveniente do teste de χ^2 de associação para cada variável.

Variável	Nível	Sem Poupança		Com Poupança		p-valor
		N	%	N	%	
Status Socioeconômico	Superior	10	23.3	36	63.2	0.0001
	Médio	13	30.2	12	21.1	
	Inferior	20	46.5	9	15.8	
Casa Própria	Não	28	65.1	34	59.6	0.727
	Sim	15	34.9	23	40.4	
Setor	B	13	30.2	14	24.6	0.686
	A	30	69.8	43	75.4	

Já a Tabela 2 indica associação significativa entre idade e posse de poupança, com o p-valor obtido a partir do Teste de Wilcoxon para comparar a média das idades para os grupos com e sem poupança (hipótese nula é de que essas médias não diferem).

Tabela 2. Distribuição de idade grupo Com ou Sem poupança. P-valor proveniente do teste não paramétrico de Wilcoxon.

Variável	Tem Poupança	N	Média	DP	Min	Q1	Mediana	Q3	Max	p-valor
Idade	Não	43	19.30	12.59	1	9	17	26	48	0.017
	Sim	57	28.84	19.36	1	15	27	39	73	

3.3 Seleção de Variáveis

A Figura 5 apresenta todos os modelos logísticos possíveis (busca exaustiva). Os três modelos destacados são os que possuem os menores valores para o AIC e incluem as variáveis idade e status como explicativas. Um dos modelos também inclui a posse de casa própria enquanto o outro inclui o setor da cidade.

Esses três modelos serão ajustados e avaliados.

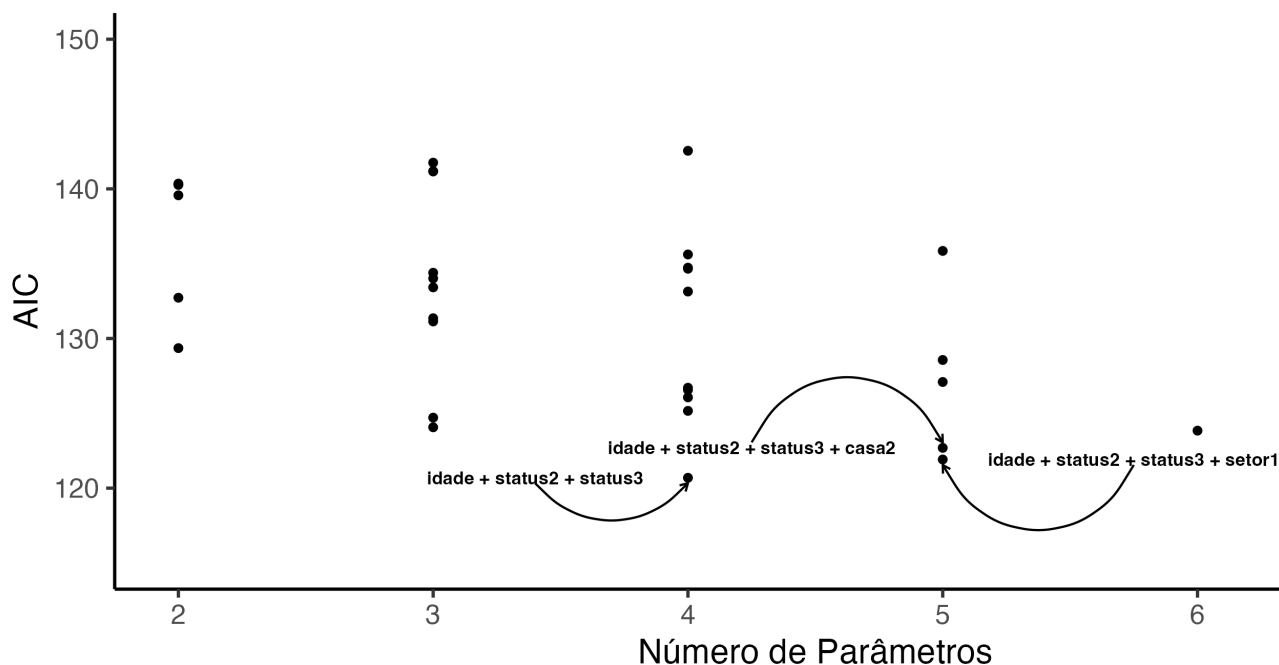


Figura 5. Critério de Informação de Akaike (AIC) por número de parâmetros no modelo, para cada modelo possível. Informações obtidas por seleção exaustiva de variáveis.

3.4 Modelos selecionados

3.4.1 Modelo 1

Este modelo utiliza apenas o status econômico e a idade como variáveis explicativas. Conforme demonstrado na Tabela 3, a idade apresenta uma relação positiva com a posse de poupança: a medida que a idade aumenta, a probabilidade de possuir poupança também aumenta. Por outro lado, os coeficientes das variáveis dummy para status econômico indicam que há uma relação negativa entre essa variável e a resposta: a medida que se avança nos níveis socioeconômicos, a probabilidade de possuir poupança diminui. Vale ressaltar que os níveis mais altos de status socioeconômico são indexados por a valores/níveis menores.

De acordo com a razão de chances, observa-se que a cada ano que uma pessoa envelhece, sua chance de possuir poupança aumenta em 3%. Além disso, caso um paciente com status socioeconômico superior passe a ter status médio, sua probabilidade de possuir poupança reduz em 72%. Já se esse paciente passar a ter status inferior, sua chance de possuir poupança sofrerá uma redução de 87%.

Tabela 3. Estimativas dos parâmetros e das razões de chances para o modelo 1.

Parâmetro	Coeficiente	OR	SE	p-value
(Intercept)	0,450	1,568	0,487	0,355
idade	0,033	1,034	0,014	0,019
factor(status)2	-1,257	0,285	0,549	0,022
factor(status)3	-2,073	0,126	0,561	0,000

3.4.2 Modelo 2

Este modelo mantém as mesmas variáveis do anterior, mas agora inclui também o setor de habitação (A ou B) na cidade. As estimativas dos parâmetros estão apresentadas na Tabela 4 e são próximas àquelas já apresentadas no modelo anterior para idade e status.

Quanto ao setor de habitação, observa-se uma associação negativa com a resposta, indicando que os pacientes do setor B têm menor probabilidade de possuir poupança. De fato, a chance de ter poupança para os habitantes do setor B representa apenas 63% da chance para os moradores do setor A. Entretanto, esse parâmetro não foi significativo.

Tabela 4. Estimativas dos parâmetros e das razões de chances para o modelo 2.

Parâmetro	Coeficiente	OR	SE	p-value
(Intercept)	0,507	1,660	0,492	0,303
idade	0,036	1,037	0,015	0,014
factor(status)2	-1,246	0,288	0,551	0,024
factor(status)3	-2,051	0,129	0,564	0,000
factor(setor)1	-0,460	0,631	0,523	0,379

3.4.3 Modelo 3

Por fim, foi realizado o ajuste do modelo que incorpora a idade, o status socioeconômico e uma variável indicadora referente à posse de casa própria quitada.

Os valores estimados para os atributos previamente modelados mantêm-se próximos aos apresentados anteriormente (conforme Tabela 5). Entretanto, a variável “casa” apresentou uma razão de chances próxima de 1, indicando independência entre a posse de casa própria e a constituição de uma poupança, entretanto esse parâmetro não foi significativo.

Tabela 5. Estimativas dos parâmetros e das razões de chances para o modelo 3.

Parâmetro	Coeficiente	OR	SE	p-value
(Intercept)	0,448	1,565	0,532	0,400
idade	0,033	1,034	0,014	0,019
factor(status)2	-1,257	0,285	0,549	0,022
factor(status)3	-2,072	0,126	0,569	0,000
factor(casa)2	0,003	1,003	0,480	0,994

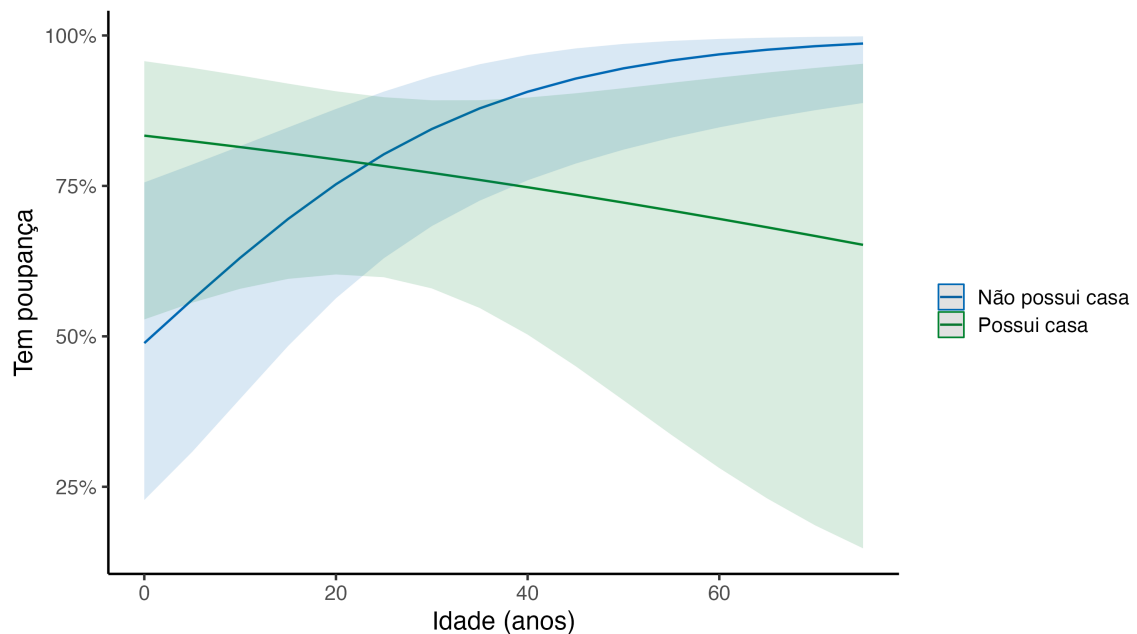
3.4.4 Modelo 4

Para avaliar a possibilidade de refinamento dos modelos, realizamos testes com versões que incluíam interações. Nessa etapa, observamos que somente a interação entre a idade do paciente e a posse de casa própria quitada apresentou um efeito significativo na modelagem da probabilidade de possuir poupança. Os coeficientes deste modelo ajustado estão detalhados na Tabela 6. Neste caso, não devemos mais analisar os fatores separadamente, mas sim a interação entre eles.

Tabela 6. Estimativas dos parâmetros e das razões de chances para o modelo 4.

Parâmetro	Coeficiente	OR	SE	p-value
(Intercept)	-0,0450	0,9560	0,5992	0,9400
idade	0,0579	1,0596	0,0195	0,0030
factor(status)2	-1,3297	0,2646	0,5612	0,0180
factor(status)3	-2,2627	0,1041	0,6080	0,0001
factor(casa)2	1,6559	5,2379	0,8671	0,0560
idade:factor(casa)2	-0,0710	0,9314	0,0306	0,0200

Para uma melhor compreensão dessa interação, analisemos a Figura 6, na qual é possível observar que a relação entre idade e posse de poupança é positiva para os moradores sem casa quitada, a probabilidade de ter poupança cresce com a idade. Já para pessoas com casa quitada, a probabilidade de ter poupança decresce com a idade.

**Figura 6.** Gráfico de interação entre posse de casa própria quitada e idade do paciente.

3.5 Avaliação dos modelos

Para essa avaliação utilizou-se a curva ROC (Receiver Operating Characteristic), que é uma representação gráfica utilizada para avaliar o desempenho de um modelo de classificação binária. Essa curva é criada plotando a taxa de verdadeiros positivos (sensibilidade) em função da taxa de falsos positivos (1 - especificidade) em diferentes pontos

de corte do modelo. Associado a ela, apresentamos a área sob a curva ROC, conhecida como AUC (Area Under the Curve), é um valor numérico que varia de 0 a 1 e resume a habilidade do modelo em distinguir entre as classes. Quanto maior o valor do AUC, melhor é o desempenho do modelo em realizar a classificação correta. Um AUC de 0,5 indica que o modelo é equivalente a um classificador aleatório, enquanto um AUC de 1 representa um classificador perfeito.

Na Figura 7 podemos observar a curva ROC e as respectiva áreas sob a curva desses três modelos. Observa-se que por essa métrica os três modelos performaram de forma muito semelhante.

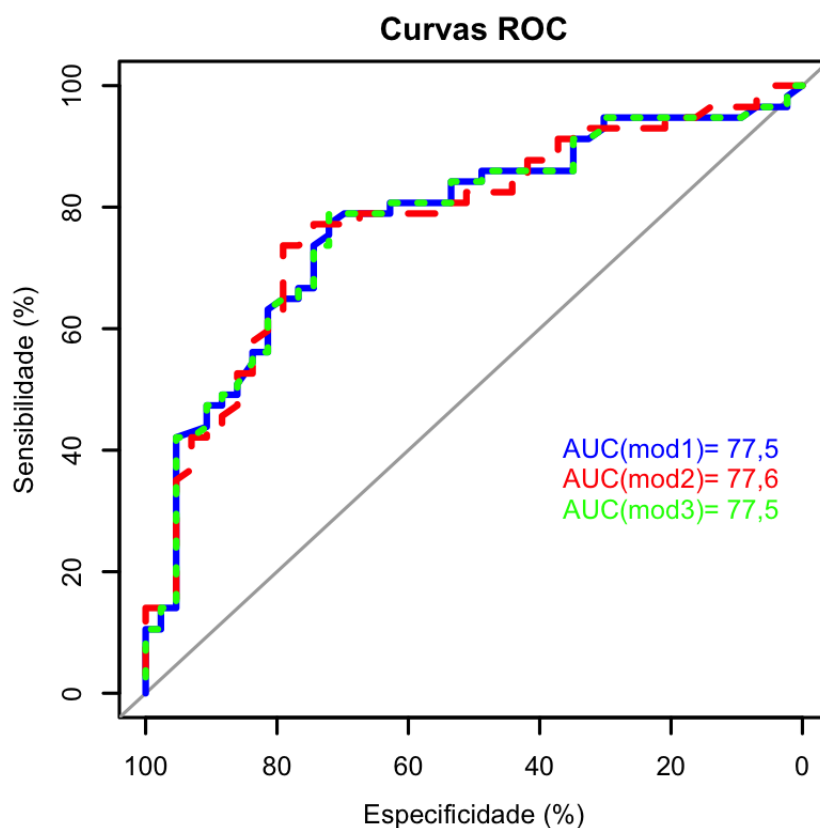


Figura 7. Diagrama de caixas para distribuição de idade entre os grupos sem (azul) e com (verde) poupança. Para comparação, a distribuição de idade geral foi adicionada no canto superior esquerdo.

Após tentativas de refinamento do modelo, selecionamos mais um modelo, agora com interação. Abaixo (Figura 8) comparamos esse modelo com o seu equivalente (sem interação), vemos que houve uma melhora significativa do AUC com a inclusão desse parâmetro.

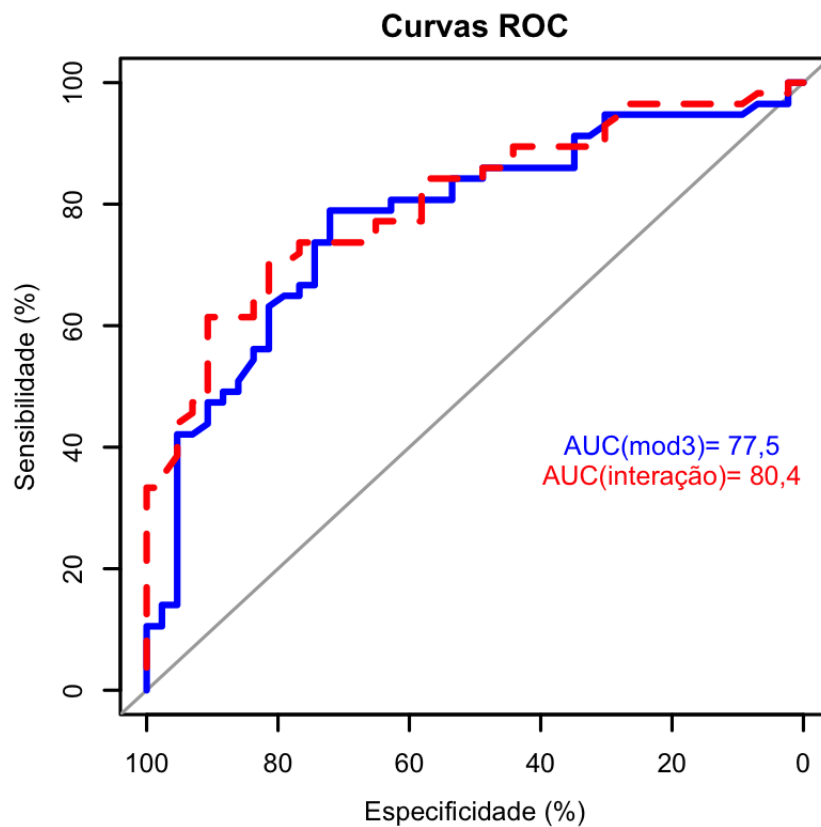


Figura 8. Diagrama de caixas para distribuição de idade entre os grupos sem (azul) e com (verde) poupança. Para comparação, a distribuição de idade geral foi adicionada no canto superior esquerdo.

3.5.1 Diagnóstico do modelo

Vemos que para os quatro modelos ajustados há um bom ajustamento, uma vez que o teste de Hosmer-Lemeshow não rejeita a hipótese de bom ajustamento para nenhum dos modelos. Também apresentamos os valores nominais do Critério de Informação de Akaike e AUC para os três modelos (Tabela 7). O modelo escolhido foi o 4, que apresentou o menor AIC e o maior AUC.

Tabela 7. Teste de Hosmer, AIC e AUC dos modelos ajustados

Modelo	p-valor Hosmer	AIC	AUC
Modelo 1	0,72	120,69	77,5
Modelo 2	0,24	121,91	77,6
Modelo 3	0,70	122,69	77,5
Modelo 4 (3 c/ interacção)	0,67	119,20	80,4

Para esse modelo, observou-se o comportamento dos resíduos. A Figura (9) apresenta os resíduos pelos valores ajustados e está de acordo com o que seria esperado para um modelo de Regressão Logística: para valores ajustados maiores, os resíduos positivos tendem a diminuir em magnitude, enquanto os negativos tendem a aumentar em magnitude.

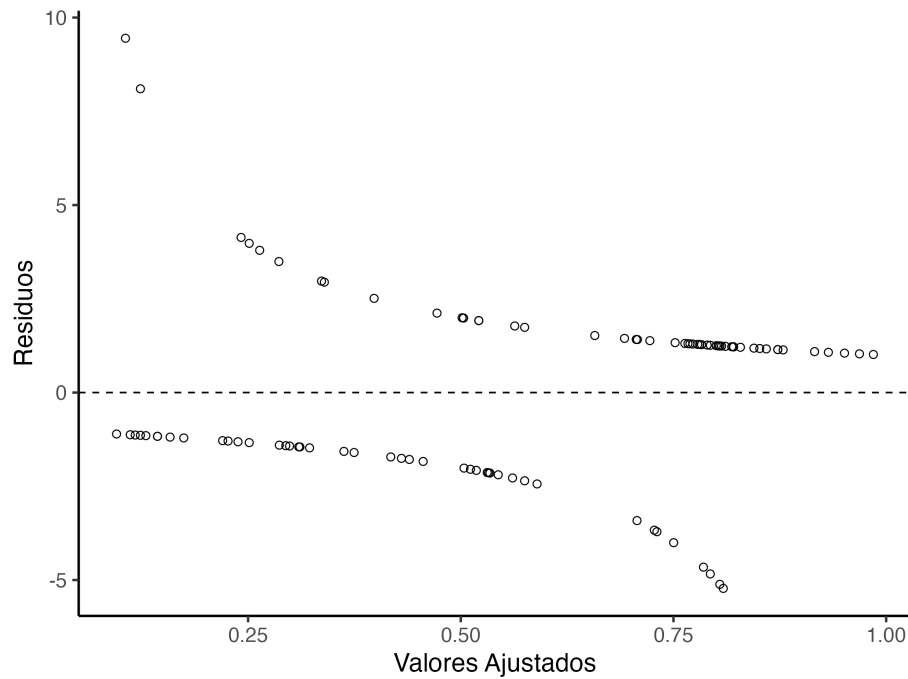


Figura 9. Gráfico de resíduos por valores ajustados do modelo selecionado.

A Figura 10 mostra os resíduos em sequência, os quais se distribuem em duas faixas em torno do zero, sem apresentar padrões de variação no decorrer do sequenciamento. Sendo assim, também está de acordo com o esperado.

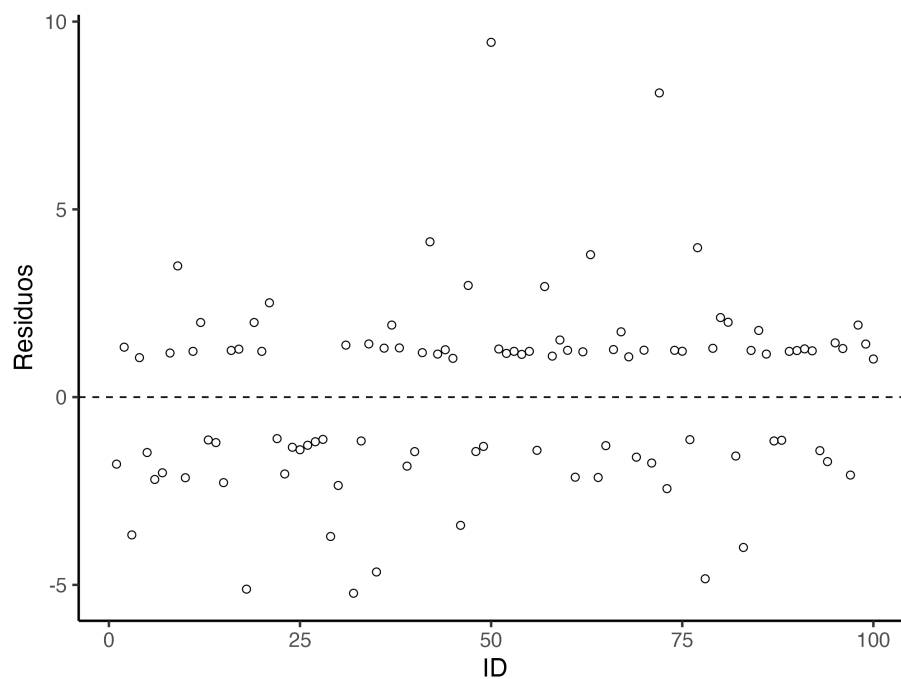


Figura 10. Gráfico de resíduos sequenciais para o modelo selecionado.

4 Conclusão

Em resumo, a posse de poupança pelos pacientes da rede hospitalar está correlacionada ao status socioeconômico, sendo mais comum entre as classes sociais mais altas. A idade também exerce influência nessa relação, inicialmente sugerindo uma associação direta: quanto mais velho o paciente, maior a probabilidade de ter poupança. No entanto, descobriu-se que essa relação é válida somente para os pacientes sem casa própria quitada. Na verdade, a relação entre idade e poupança depende da posse de casa pelo paciente. O único atributo que não demonstrou influência significativa na presença de poupança foi o setor habitacional onde os pacientes residem.

Com base nessas descobertas, o modelo mais adequado para descrever a associação entre essas variáveis e ajustar as probabilidades de possuir poupança foi uma regressão logística envolvendo as três variáveis significativas e a interação idade-casa. Após avaliação e análise residual, concluiu-se que o modelo estava bem ajustado e, portanto, é apropriado para fazer previsões em relação às poupanças dos pacientes.

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::pload(tidyverse, dplyr, rio, paperR, patchwork,
              kableExtra, pROC, ExhaustiveSearch, scales,
              sjPlot, sjmisc, performance, lmtest, stringr)

# 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", "Médio", "Inferior")),
         casa=factor(casa, labels=c("Não", "Sim")),
         setor=factor(setor, levels=c(1,0), labels=c("B", "A")),
         save=factor(save, labels=c("Não", "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( "#0068B4", "#00822E"))+
```

```
#geom_jitter(color="black", size=0.4, alpha=0.9) +
theme_classic(base_size = 14)+
theme(legend.position = "none")+
labs(x= "Possui Poupança", 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")

box_idade <- idade_by + inset_element(idade, left = 0.01,
                                       bottom = 0.45,
                                       right = 0.25, top = 1)

ggsave(plot=box_idade, filename = "img/idade.png")

status_by <- amostra_trat %>%
  ggplot(aes(x=save,fill=factor(status)))+
  geom_bar(position="fill", color="black")+
  scale_fill_manual(values=c("#0068B4", "#7E7E65", "#00822E"),
                    labels=c("Inferior", "Médio", "Superior"))+
  #geom_jitter(color="black", size=0.4, alpha=0.9) +
  theme_classic(base_size = 14)+
  theme(legend.position = "none")+
  labs(x= "Possui Poupança", y= "Proporção")
```

```

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

bar_plot_status <- status_by + status + plot_layout(widths = c(2, 1))

ggsave(plot= bar_plot_status, filename = "img/status.png")

casa_by <- amostra_trat %>%
  ggplot(aes(x=save,fill=factor(casa)))+
  geom_bar(position="fill", color="black")+
  scale_fill_manual("", values=c( "#0068B4", "#00822E"),
                    labels=c("Não", "Sim"))+
  #geom_jitter(color="black", size=0.4, alpha=0.9) +
  theme_classic(base_size = 14)+
  theme(legend.position = "none")+
  labs(x= "Possui Poupança", y= "Proporção")

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

bar_plot_casa <- casa_by + casa + plot_layout(widths = c(2, 1))
ggsave(plot= bar_plot_casa, filename = "img/casa.png")

setor_by <- amostra_trat %>%
  ggplot(aes(x=save,fill=factor(setor)))+

```

```

geom_bar(position="fill", color="black")+
scale_fill_manual("", values=c( "#0068B4", "#00822E"))+
#geom_jitter(color="black", size=0.4, alpha=0.9) +
theme_classic(base_size = 14)+
theme(legend.position = "none")+
labs(x= "Possui Poupança", y= "Proporção")

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

bar_plot_setor <- setor_by + setor + plot_layout(widths = c(2, 1))
ggsave(plot= bar_plot_setor, filename = "img/setor.png")

names(amostra_trat) <- c("ID", "Idade", "Status Socioeconômico", "Casa Própria", "S

Tab1 <- paperR::summarize(amostra_trat[,-1],
                          type = "factor", group = "Poupança", test="chisq.test")

names(Tab1 ) <- c("Variável", "Nível", "a", "N ", "%", "b", "N", " % ", "c", "p-val

Tab1 <- Tab1 %>% as_tibble() %>% select(-c("a", "b", "c")) %>%
  mutate(`p-valor`= str_replace(`p-valor`, "<0.001", "0.0001"))

knitr::kable(Tab1,format="latex", booktabs = T,
              linesep="",
              align="c", caption= "Frequências relativas e absolutas das variáveis qualit
              kable_styling(latex.options = "HOLD_position", position="center",
              full_width = F) %>%

```

```

row_spec(c(3,5),hline_after = TRUE) %>%
  add_header_above(c(" ", " ", "Sem Poupança" = 2,
                     "Com Poupança" = 2, "" ))

Tab2 <- paper::summarize(amostra_trat[,-1],
                        type = "numeric", group = "Poupança",
                        test="wilcox.test")

names(Tab2)<-c("Variável", "Tem Poupança", "a", "N", "b", "Média",
              "DP", "c", "Min", "Q1", "Mediana", "Q3", "Max", "d",
              "p-valor")

Tab2 <- Tab2 %>% as_tibble() %>% select(-c("a", "b", "c", "d")) %>%
  mutate(`p-valor`= str_replace(`p-valor`, "<0.001", "0.0001"))

knitr::kable(Tab2,format="latex", booktabs = T,
              linesep="",
              align="c",
              caption= "Distribuição de idade grupo Com ou Sem poupança.
              P-valor proveniente do teste não paramétrico de Wilcoxon.") %>%
  kable_styling(latex_options = "HOLD_position", position="center",
                full_width = F)

selecao_exaustiva <- ExhaustiveSearch(save ~ idade +
                                     factor(status) +
                                     factor(setor) +
                                     factor(casa),
                                     data = amostra[,-1],
                                     family = "binomial",
                                     performanceMeasure = "AIC")

plot_AIC_npar <- ExhaustiveSearch::resultTable(selecao_exaustiva) %>%
  mutate(n_par=str_count(Combination, "\\++")+2)

plot_AIC_npar_labels <- plot_AIC_npar %>%
  head(3) %>%
  mutate(label=str_replace_all(Combination, "factor\\(", ""),
         label=str_replace_all(label, "\\)", ""))

```

```
label=str_squish(label))

plot_AIC <- plot_AIC_npar %>%
  ggplot(aes(x=n_par, y=AIC))+
  geom_point()+
  theme_classic(base_size = 16)+
  scale_x_continuous(limits = c(2, 7))+
  scale_y_continuous(limits = c(115, 150))+
  labs(x="Número de Parâmetros", y="AIC")+
  geom_text(data=plot_AIC_npar_labels,
            aes(label=label), size=3, nudge_x = c(-0.6, 0.75, -0.75),
            fontface="bold")+
  geom_curve(data=plot_AIC_npar_labels[1,],
             aes(x =n_par*0.85, y = AIC*0.997, xend = n_par,
                 yend = AIC*0.997),
             arrow = arrow(length = unit(0.015, "npc")),
             curvature = 0.5)+
  geom_curve(data=plot_AIC_npar_labels[3,],
             aes(x =n_par*0.85, y = AIC*1.003, xend = n_par,
                 yend = AIC*1.003),
             arrow = arrow(length = unit(0.015, "npc")),
             curvature = -0.7)+
  geom_curve(data=plot_AIC_npar_labels[2,],
             aes(x =n_par*1.15, y = AIC*0.997, xend = n_par,
                 yend = AIC*0.997),
             arrow = arrow(length = unit(0.015, "npc")),
             curvature = -0.7)
ggsave("img/AIC.png", width=10)

mod1 <- glm(save ~ idade + factor(status),
            data = amostra[,-1], family=binomial(link="logit"))

mod2 <- glm(save ~ idade + factor(status) +
            factor(setor), data = amostra[,-1],
```

```

    family=binomial(link="logit"))

mod3 <- glm(save ~ idade + factor(status) +
            factor(casa), data = amostra[, -1],
            family=binomial(link="logit"))
#modelo selecionado
prettify(summary(mod1)) %>%
  as_tibble() %>%
  select(-c("CI (lower)", "CI (upper)", "z value", " " )) %>%
  rename(`p-value` = `Pr(>|z|)`) %>%
  mutate(`p-value` = as.numeric(str_replace(`p-value`,
                                           "<0.001", "0.0001"))) %>%
  mutate_if(is.numeric, ~comma(., decimal.mark = ",",
                                accuracy=0.001)) %>%
  kable(format="latex", booktabs = T,
        linesep="",
        align="c",
        col.names = c("Parâmetro", "Coeficiente", "OR", "SE",
                      "p-value"),
        caption="Estimativas dos parâmetros e das razões de
chances para o modelo 1.") %>%
  kable_styling(latex_options = "HOLD_position", position="center",
                full_width = F)
#modelo saturado
prettify(summary(mod2)) %>%
  as_tibble() %>%
  select(-c("CI (lower)", "CI (upper)", "z value", " " )) %>%
  rename(`p-value` = `Pr(>|z|)`) %>%
  mutate(`p-value` = as.numeric(str_replace(`p-value`, "<0.001",
                                           "0.0001"))) %>%
  mutate_if(is.numeric, ~comma(., decimal.mark = ",",
                                accuracy=0.001)) %>%
  kable(format="latex", booktabs = T,
        linesep="",
        align="c",
        col.names = c("Parâmetro", "Coeficiente", "OR", "SE",
                      "p-value"),
        caption="Estimativas dos parâmetros e das razões de chances

```

```

    para o modelo 2.") %>%
  kable_styling(latex_options = "HOLD_position", position="center",
    full_width = F)
#modelo saturado
prettify(summary(mod3)) %>%
  as_tibble() %>%
  select(-c("CI (lower)","CI (upper)", "z value", " " )) %>%
  rename(`p-value`=`Pr(>|z|)`) %>%
  mutate(`p-value`=as.numeric(str_replace(`p-value`, "<0.001",
    "0.0001")) %>%

  mutate_if(is.numeric, ~comma(., decimal.mark = ",",
    accuracy=0.001)) %>%

  kable(format="latex", booktabs = T,
    linesep="",
    align="c",
    col.names = c("Parâmetro", "Coeficiente", "OR", "SE",
      "p-value"),
    caption="Estimativas dos parâmetros e das razões de
    chances para o modelo 3.") %>%
  kable_styling(latex_options = "HOLD_position", position="center",
    full_width = F)

mod3_iter <- glm(save ~ idade+factor(status)+
  factor(casa)+idade:factor(casa), data = amostra[,-1],
  family=binomial(link="logit"))

#modelo saturado
prettify(summary(mod3_iter)) %>%
  as_tibble() %>%
  select(-c("CI (lower)","CI (upper)", "z value", " " )) %>%
  rename(`p-value`=`Pr(>|z|)`) %>%
  mutate(`p-value`=as.numeric(str_replace(`p-value`, "<0.001",
    "0.0001")) %>%

  mutate_if(is.numeric, ~comma(., decimal.mark = ",",
    accuracy=0.0001)) %>%

  kable(format="latex", booktabs = T,

```



```

    linesep="",
    align="c",
    col.names = c("Parâmetro", "Coeficiente", "OR", "SE",
                  "p-value"),
    caption="Estimativas dos parâmetros e das razões de
    chances para o modelo 4.") %>%
kable_styling(latex_options = "HOLD_position", position="center",
              full_width = F)

p <- plot_model(mod3_iter, type="pred", terms=c("idade", "casa"),
               se=FALSE)+
  scale_color_manual("", labels=c("Não possui casa", "Possui casa"),
                    values=c( "#0068B4", "#00822E"))+
  scale_fill_manual("", labels=c("Não possui casa", "Possui casa"),
                   values=c( "#0068B4", "#00822E"))+
  theme_classic(base_size = 16)+
  labs(x="Idade (anos)", y="Tem poupança")+
  ggtitle("")

ggsave(plot=p, filename = "img/interacao.png", width=10, height=6)
png('img/roc.png', pointsize=6, width=850, height=800, res=300)
#ROC --> avaliação do modelo
roc1 <- roc( amostra$save,as.vector(fitted.values(mod1)) ,
            grid=TRUE, percent=T)
auc1<-comma(as.numeric(roc1$auc),
            decimal.mark = ",", accuracy=0.1)

roc2 <- roc( amostra$save,as.vector(fitted.values(mod2)) ,
            grid=TRUE, percent=T)
auc2 <-comma(as.numeric(roc2$auc),
            decimal.mark = ",", accuracy=0.1)

roc3 <- roc( amostra$save,as.vector(fitted.values(mod3)) ,
            grid=TRUE, percent=T)
auc3 <- comma(as.numeric(roc3$auc),
            decimal.mark = ",", accuracy=0.1)

```

```
plot(roc1 ,xlim=c(100,0),ylim=c(0,100), asp = NA, legacy.axes = F,
     xlab="Especificidade (%)", ylab="Sensibilidade (%)",
     percent=T, print.auc = F, main="Curvas ROC", col="blue")

lines(roc2 ,xlim=c(100,0),ylim=c(0,100), asp = NA, legacy.axes = F,
     xlab="Especificidade (%)", ylab="Sensibilidade (%)",
     percent=T, print.auc = F, main="Curva ROC", col="red",
     lty = 2)

lines(roc3 ,xlim=c(100,0),ylim=c(0,100), asp = NA, legacy.axes = F,
     xlab="Especificidade (%)", ylab="Sensibilidade (%)",
     percent=T, print.auc = F, main="Curva ROC", col="green", lty = 3)

text(20, 40, paste("AUC(mod1)=", auc1), col="blue")
text(20, 35, paste("AUC(mod2)=", auc2), col="red")
text(20, 30, paste("AUC(mod3)=", auc3), col="green")

png('img/roc_inter.png', pointsize=6, width=850, height=800, res=300)
#ROC --> avaliação do modelo

roc3_inter <- roc( amostra$save,as.vector(fitted.values(mod3_iter)) ,
                 grid=TRUE, percent=T)
auc3_inter <- comma(as.numeric(roc3_inter$auc),
                  decimal.mark = ",", accuracy=0.1)

plot(roc3 ,xlim=c(100,0),ylim=c(0,100), asp = NA, legacy.axes = F,
     xlab="Especificidade (%)", ylab="Sensibilidade (%)",percent=T,
     print.auc = F, main="Curvas ROC", col="blue")

lines(roc3_inter ,xlim=c(100,0),ylim=c(0,100), asp = NA, legacy.axes = F,
     xlab="Especificidade (%)", ylab="Sensibilidade (%)",
     percent=T, print.auc = TRUE, main="Curva ROC", col="red",
     lty = 2)
```

```

text(20, 40, paste("AUC(mod3)=", auc1), col="blue")
text(20, 35, paste("AUC(interacção)=", auc3_inter), col="red")

mod1_hosmer <- performance_hosmer(mod1)
mod2_hosmer <- performance_hosmer(mod2)
mod3_hosmer <- performance_hosmer(mod3)
mod4_hosmer <- performance_hosmer(mod3_iter)

mod1_aic <- mod1$aic
mod2_aic <- mod2$aic
mod3_aic <- mod3$aic
mod3I_aic <- mod3_iter$aic

modelos <- c("Modelo 1", "Modelo 2", "Modelo 3",
            "Modelo 4 (3 c/ interacção)")
hosmer <- c(mod1_hosmer$p.value,
            mod2_hosmer$p.value, mod3_hosmer$p.value,
            mod4_hosmer$p.value )
aic <- c(mod1_aic, mod2_aic, mod3_aic, mod3_iter$aic)
auc <- c(auc1, auc2, auc3, auc3_inter)

ajuste_modelos <- cbind(modelos, hosmer, aic, auc) %>%
  as_tibble() %>%
  mutate(across(c("hosmer", "aic"), ~as.double(.x))) %>%
  mutate(across(c("hosmer", "aic"), ~comma(., decimal.mark = ",",
                                         accuracy=0.01)))

ajuste_modelos %>%
  kable(format="latex", booktabs = T,
        linesep="",
        align=c("c"),
        col.names=c("Modelo", "p-valor Hosmer", "AIC", "AUC"),
        caption="Teste de Hosmer, AIC e AUC dos modelos ajudtados ") %>%
  kable_styling(latex_options = "HOLD_position", position="center",
               full_width = F)

## MODELO 1

```

```
graf1_res <- ggplot()+
  geom_point(aes(mod3_iter$fitted.values, mod3_iter$residuals), size=2,
             shape=1)+
  geom_hline(yintercept=0, linetype = 'dashed')+
  labs(x="Valores Ajustados", y="Resíduos")+
  theme_classic(base_size = 16)

ggsave(plot=graf1_res, filename = "img/residuo1.png", width=8, height=6)

graf2_res <- data.frame(x=1:100, y=mod3_iter$residuals ) %>%
ggplot()+
  geom_point(aes(x,y), size=2, shape=1)+
  geom_hline(yintercept=0, linetype = 'dashed')+
  labs(x="ID", y="Resíduos")+
  theme_classic(base_size = 16)

ggsave(plot=graf2_res, filename = "img/residuo2.png", width=8,
        height=6)
```

Referências

HOSMER-JR., D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. 3rd edition. ed. [S.l.]: Wiley, 2013. ISBN 978-0470582473.

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