Universidade de Brasília

Eduardo Moreira Araújo
José Vítor Barreto Porfírio
Stefan Zurman Gonçalves

Análise de fatores associados a obtenção de conta poupança pelas famílias dos pacientes da rede hospitalar

Universidade de Brasília – UnB
Departamento de Estatística
Programa de Graduação

Professor: Maria Teresa Leão

Brasília, DF
25 de julho de 2023

Sumário

1 Introdução e Objetivos

Este trabalho tem como objetivo principal analisar fatores associados a obtenção de conta poupança pelas famílias dos pacientes da rede hospitalar. A estratégia principal consiste em descrever e avaliar a associação entre a família possuir ou não uma conta poupança com variáveis como idade do paciente, status socioeconômico, setor da cidade que habitam e a posse de casa própria. Portanto, realizaremos análise descritiva dos dados amostrais fornecidos e utilizaremos modelos de regressão logística para alcançar os objetivos da pesquisa.

Nessa perspectiva, este estudo se torna importante para que a rede hospitalar conheça com maior detalhamento seu público alvo. Além disso, é uma oportunidade para avaliar fatores que podem impactar na demanda futura por serviços médicos e nos gastos com hospital por parte das famílias de seus pacientes.

O presente relatório está estruturado da seguinte forma: introdução, metodologia, resultados, conclusão e apêndice. Na parte metodológica, discutimos em detalhes as análises e técnicas estatísticas utilizadas. Na seção sobre resultados apresentamos o conhecimento gerado pelo estudo após análise estatística descritiva e modelagem dos dados. Na aba de conclusões fazemos considerações finais e resumidas sobre o estudo. Por fim, a seção apêndice apresenta um link para acesso à listagem do programa em R utilizado.

2 Metodologia

A rede hospitalar obteve uma amostra aleatória de 196 pacientes em dois setores de uma cidade. Além da identificação dos pacientes de 1 a 196, o banco de dados fornecido conta com 5 variáveis para estudo e apresenta as seguintes informações sobre as características investigadas para a amostra selecionada:

- Conta poupança - verifica se a família tem conta poupança (0 – não; 1 – sim). Trata-se de uma variável qualitativa nominal e mensurada na escala nominal;
- Status socioeconômico - verifica qual o nível socioeconômico da família (1 = superior, 2 = médio, 3 = inferior). Trata-se de uma variável qualitativa ordinal e mensurada na escala ordinal;
- Setor da cidade - verifica qual a área que a família do paciente habita (1 = setor A; 0 = setor B). Trata-se de uma variável qualitativa nominal e mensurada na escala nominal;
- Possui casa própria - verifica se a família do paciente possui casa própria ou não (Família possui casa própria: 1 = não ou sim, mas ainda pagando financiamento, 2 = sim e quitada). Trata-se de uma variável qualitativa nominal e mensurada na escala nominal;
- Idade do paciente: medida em anos completos. Embora seja apresentada em números inteiros, a essência da variável é quantitativa contínua. A escala de mensuração é a escala de razão.

O primeiro passo foi a análise descritiva completa dos dados, na qual cada variável foi investigada individualmente. Logo depois, uma análise bivariada das variáveis foi realizada, tendo como variável resposta a obtenção de conta poupança pelas famílias dos pacientes. Tabelas e gráficos como boxplot, de colunas e barras e histograma foram gerados nessa etapa. Além disso, medidas resumo como média, mediana e desvio padrão foram calculados para a variável idade do paciente. Uma observação importante é que a variável idade foi dividida em intervalos de classes apenas para apresentação tabular, ou seja, não foi categorizada para o cálculo das medidas resumo (haveria perda de informação e a escala de razão não seria a utilizada).

Após a análise descritiva, foi feita a modelagem por regressão logística. Das 196 observações originais, foram selecionadas aleatoriamente 100 observações para serem utilizadas para a construção do modelo, sendo as demais utilizadas apenas para a validação

do modelo. Então, foram ajustados todos os modelos possíveis de combinações de variáveis, e selecionado o modelo com o menor AIC dentre eles. Foi feita uma validação prévia do ajuste do modelo por análise dos resíduos *binned*, e por um teste Hosmer-Lemeshow.

Em seguida, utilizou-se o banco de teste para fazer a validação do modelo escolhido. Foi calculado um intervalo de confiança para a acurácia e montada uma matriz de confusão para o modelo. Com isso, foram calculadas a especificidade e a sensibilidade do modelo, além da curva ROC.

Por fim, foi feita uma análise da significância de cada parâmetro do modelo final. Foram então construídos intervalos de confiança para os parâmetros, além de suas interpretações. Finalmente, foram calculadas razões de chances utilizando o modelo com intervalos de confiança, além de suas interpretações no contexto do trabalho.

3 Resultados

3.1 Análise Exploratória

Em um primeiro momento, faremos uma análise descritiva das seguintes variáveis: Conta Poupança, Idade do Paciente, Status Socioeconômico, Setor da Cidade e Possui Casa Própria. Como em nosso estudo queremos analisar fatores associados a família dos pacientes da rede hospitalar ter uma poupança ou não, também precisamos relacionar tal variável (Conta Poupança) com cada uma das outras em uma análise bivariada. Essa primeira etapa descritiva nos possibilitará obter hipóteses de trabalho fundamentais para a modelagem dos dados que será realizada em um passo seguinte.

3.1.1 Análise Descritiva: Conta Poupança

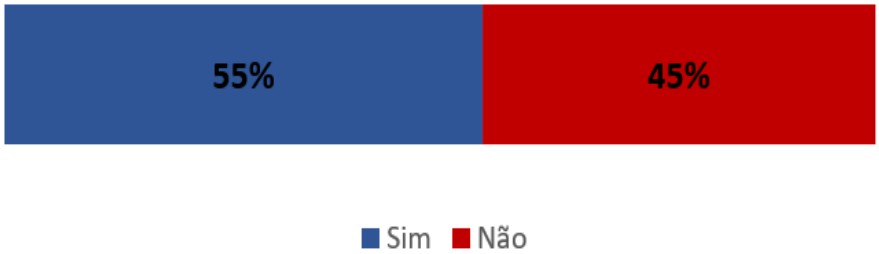
A primeira variável de interesse é denominada de Conta Poupança. Conforme discutido anteriormente, essa é a variável resposta do nosso estudo. Trata-se de uma variável qualitativa nominal, podendo ser respondida como “Sim"ou “Não"em relação a existência de conta poupança pela família do paciente.

É possível visualizar na tabela 1 abaixo que 55% das famílias dos pacientes da rede hospitalar possuem conta poupança e 45% não possuem. Apesar dessa diferença, podemos afirmar que não há uma dissimilariedade acentuada entre os valores obtidos.

Tabela 1 – Distribuição do número de pacientes segundo a obtenção de conta poupança pela família - Brasil, 2023

Conta Poupança	**Frequência Absoluta**	**Frequência Relativa (%)**
Sim	107	55 %
Não	89	45 %
Total	196	100 %

Figura 1 – Distribuição percentual do número de pacientes segundo a obtenção de conta poupança pela família - Brasil, 2023



3.1.2 Análise Descritiva: Idade

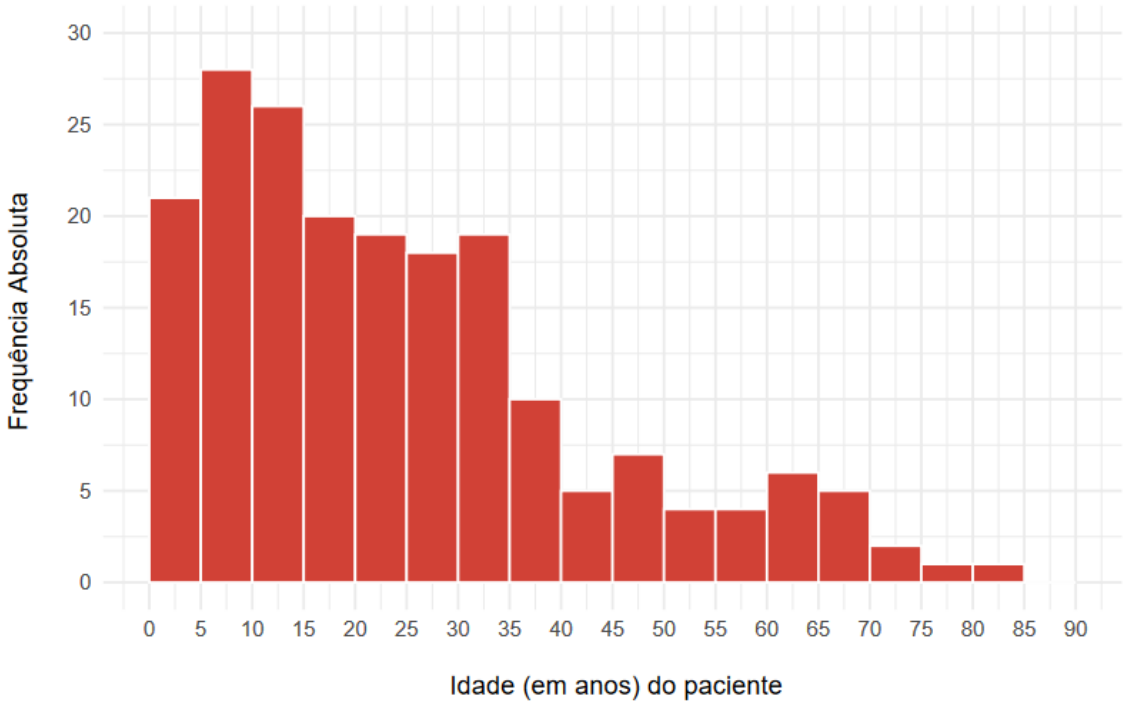
Um fator de interesse importante é a idade dos pacientes. Essa é uma variável que foi medida em anos e se caracteriza como uma variável quantitativa contínua.

Em média, a idade dos pacientes da rede hospitalar na amostra coletada é de 25,2 anos. Observamos que a mediana é de 21 anos, ou seja, 50% dos indivíduos consultados possuem 21 ou menos anos de idade. A idade mínima registrada foi de 1 ano completo e a idade máxima foi na casa dos 85 anos. Vejamos na tabela abaixo a distribuição dos pacientes a partir das faixas etárias predefinidas:

Tabela 2 – Distribuição do número de pacientes segundo a faixa etária - Brasil, 2023

Faixa Etária	**Frequência Absoluta**	**Frequência Relativa (%)**
00 ⊢ 10	48	24 %
10 ⊢ 20	44	22 %
20 ⊢ 30	36	18 %
30 ⊢ 40	32	16 %
40 ⊢ 50	11	6 %
50 ⊢ 60	8	4 %
60 ⊢ 70	11	6 %
70 ⊢ 80	5	3 %
80 ⊢ 90	1	1 %
Total	196	100 %

Figura 2 – Histograma das idades dos pacientes da rede hospitalar - Brasil, 2023



Existe uma concentração dos valores em idades mais jovens, confome verificamos na tabela e no histograma acima. Nesse sentido, cerca de 80% dos pacientes possuem menos de 40 anos de idade.

O próximo passo é realizar uma análise bivariada entre a idade do paciente e a obtenção ou não de conta poupança pela sua família. Nesse sentido, podemos verificar que, em média, a idade dos pacientes é de 30,2 quando a conta poupança familiar existe. Essa média é de 19,2 anos de idade quando a família não possui esse tipo de conta. Para o grupo sem conta poupança familiar, 50% é fomado por menores de idade. Essa análise também é reafirmada quando olhamos para a mediana, conforme verificamos abaixo:

Tabela 3 – Medidas resumo da variável idade dos pacientes segundo a obtenção de conta poupança pela família - Brasil, 2023

Medidas Resumo	**Sim**	**Não**
Média	30,2	19,2
Desvio Padrão	21,8	12,4
Coef. De Variação	0,72	0,64
Mínimo	1	1
Máximo	85	53
Mediana	27	17

Quando analisamos o desvio padrão, verificamos que no grupo de pacientes no qual a família possui poupança, há uma maior variação nas idades se comparado com o grupo que não possui poupança. Essa ideia é reafirmada quando vemos que a idade máxima é de 85 para a resposta positiva e apenas 53 para a resposta negativa.

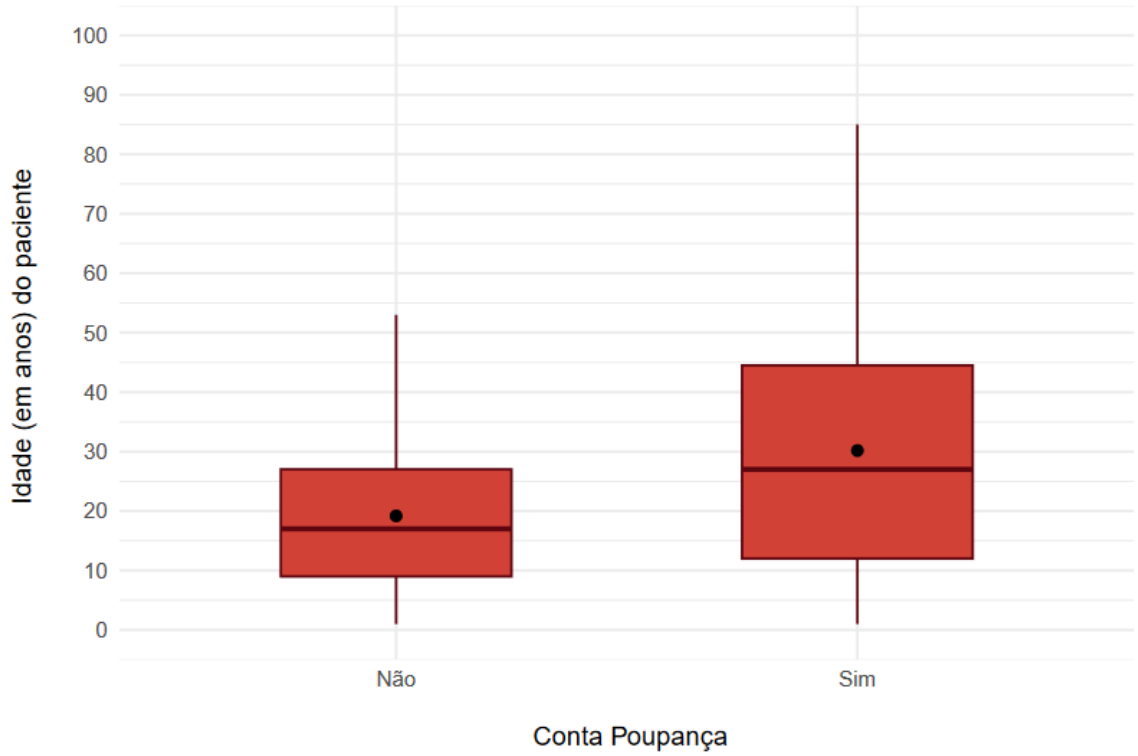
Uma hipótese de trabalho é que a idade parece ser um fator para obtenção de conta poupança pela família. Para pacientes mais novos, a família tende a ter mais gastos básicos, sobretudo, quando estamos considerando pacientes crianças. Por outro lado, quando o paciente passa para a idade adulta, há maiores chances da família ser liderada pelo próprio paciente, por exemplo. Nesse sentido, em famílias com indivíduos adultos possivelmente teremos pessoas economicamente mais ativas e com maior tendência em poupar. Para avaliar um pouco mais essa hipótese, precisamos relacionar tal ideia com renda da família, por exemplo. Como as demais variáveis englobam tal assunto, a modelagem dos dados nos permitirá verificar se essa lógica faz sentido.

Por fim, podemos observar via tabela e gráfico abaixo a existência de tendência no aumento percentual de famílias com poupança a medida que a idade do paciente avança. Por outro lado, quanto menor a idade do paciente, menos suscetível a família é em poupar dinheiro.

Tabela 4 – Distribuição das famílias com e sem conta poupança segundo as faixas etárias dos pacientes - Brasil, 2023

Faixa Etária	**Conta Poupança**				**Total**
	Sim		**Não**		
	Freq. Absoluta	**Freq. Relativa**	**Freq. Absoluta**	**Freq. Relativa**	
00 ⊢ 10	21	44%	27	56%	48
10 ⊢ 20	19	43%	25	57%	44
20 ⊢ 30	18	50%	18	50%	36
30 ⊢ 40	19	59%	13	41%	32
40 ⊢ 50	6	55%	5	45%	11
50 ⊢ 60	7	88%	1	12%	8
60 ⊢ 70	11	100%	0	0%	11
70 ⊢ 80	5	100%	0	0%	5
80 ⊢ 90	1	100%	0	0%	1
Total	107	-	89	-	196

Figura 3 – Boxplot das idades dos pacientes segundo a presença de conta poupança familiar - Brasil, 2023



3.1.3 Análise Descritiva: Status Socioeconômico

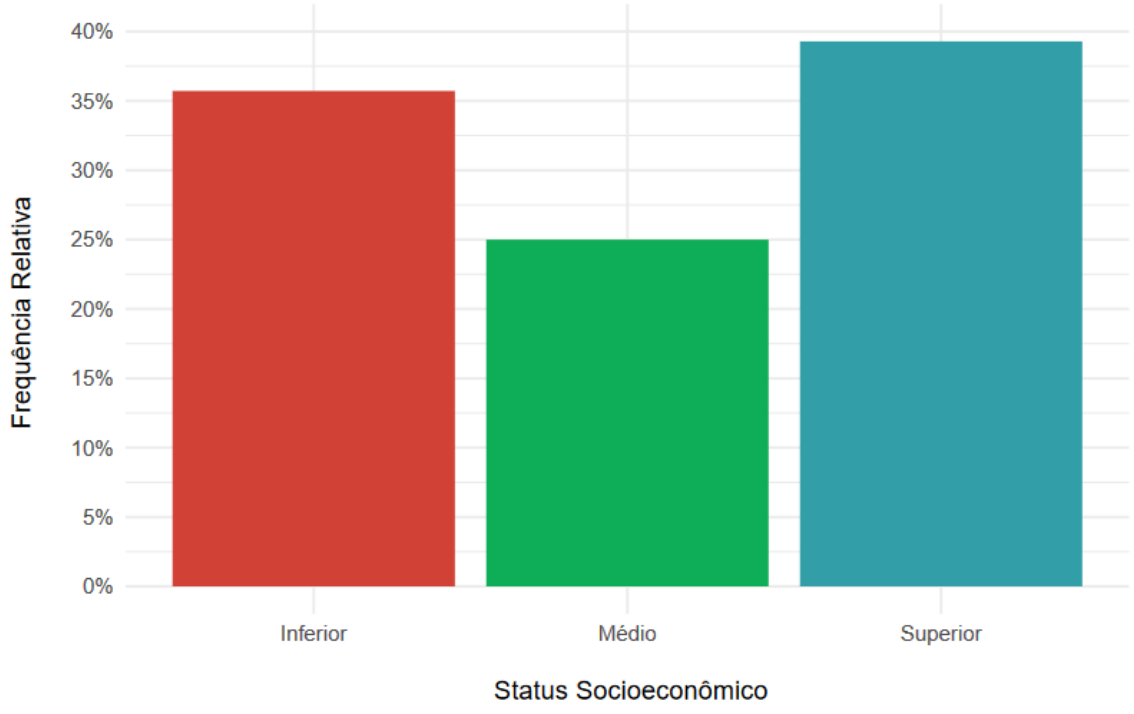
A variável em questão indica o status socioeconômico da família do paciente. Em suma, trata-se de uma variável que nos retorna informação sobre renda familiar e condições sociais que podem influenciar no objetivo da família em poupar dinheiro ou não. A variável é qualitativa e ordinal, tendo como respostas possíveis as categorias “inferior", “médio"e “superior".

Na amostra coletada, apenas 25% dos pacientes possuem como status socioeconômico familiar o nível “médio". Nesse sentido, a maior parte dos pacientes são categorizados em “inferior"ou “superior"em termos de status socioeconômico, nos quais as frequências relativas obtidas foram 36% e 39%, respectivamente.

Tabela 5 – Distribuição do número de pacientes segundo status socieconômico das famílias - Brasil, 2023

Status Socioeconômico	**Freq. Absoluta**	**Freq. Relativa (%)**
Superior	77	39%
Médio	49	25%
Inferior	70	36%
Total	196	100%

Figura 4 – Distribuição percentual do número de pacientes segundo status socieconômico das famílias - Brasil, 2023



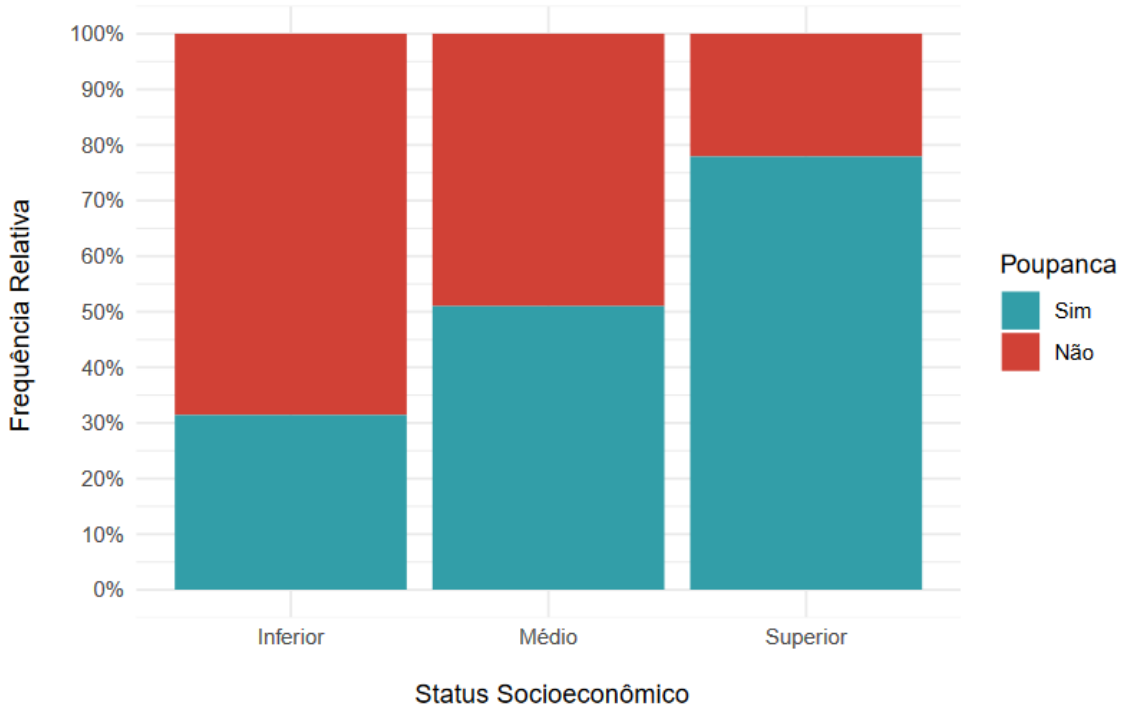
Quando associamos a variável status socioeconômico com a conta poupança, verificamos que há uma tendência em que, quanto maior o nível da categoria que indica status, maior é a proporção de famílias que possuem conta poupança. Nesse sentido, famílias com menores status tendem a responder que não possuem poupança.

A hipótese que pode ser levantada é que, como o status socioeconômico está ligado à renda, as famílias com maior poder aquisitivo possuem condições para abrir esse tipo de conta. Por outro lado, famílias que apresentam maiores dificuldades econômicas não conseguem poupar, visto que precisam usar todos seus recursos para as necessidades básicas e não sobra dinheiro.

Tabela 6 – Distribuição das famílias com e sem conta poupança segundo o status socioeconômico das famílias dos pacientes - Brasil, 2023

Status	**Conta Poupança**				**Total**
	Sim		**Não**		
	Freq. Absoluta	**Freq. Relativa**	**Freq. Absoluta**	**Freq. Relativa**	
Superior	60	78%	17	22%	87
Médio	25	51%	24	49%	49
Inferior	22	31%	48	69%	60
Total	107	-	89	-	196

Figura 5 – Distribuição percentual do número de contas poupança segundo o status socioeconômico das famílias dos pacientes - Brasil, 2023



3.1.4 Análise Descritiva: Setor da Cidade

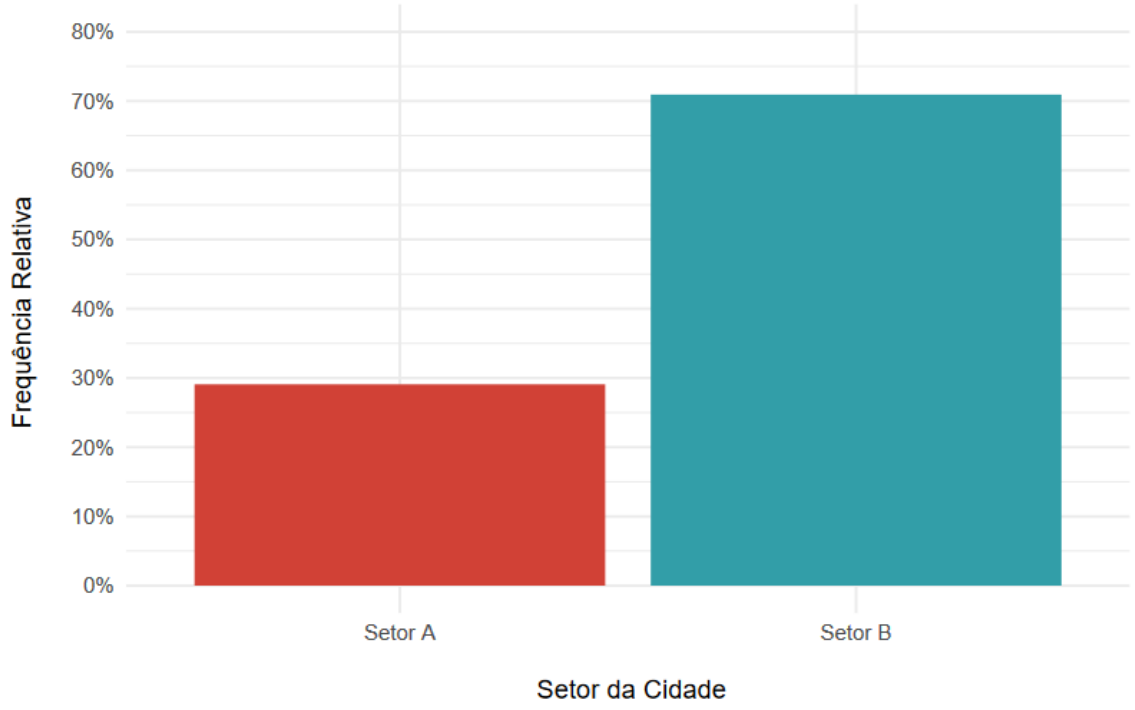
A variável setor da cidade indica a área em que a família do paciente reside. Trata-se de uma variável nominal com duas categorias: “Setor A"e “Setor B".

Em síntese, 71% das famílias dos pacientes residem no setor B da cidade e apenas 29% no setor A. Apesar dessa baixa porcentagem, a amostra de indivíduos que habitam no setor A não é pequena em termos absolutos.

Tabela 7 – Distribuição do número de pacientes segundo o setor da cidade em que as famílias residem - Brasil, 2023

Setor da Cidade	**Freq. Absoluta**	**Freq. Relativa (%)**
Setor A	57	29%
Setor B	139	71%
Total	196	100%

Figura 6 – Distribuição percentual do número de pacientes segundo o setor da cidade em que as famílias residem - Brasil, 2023



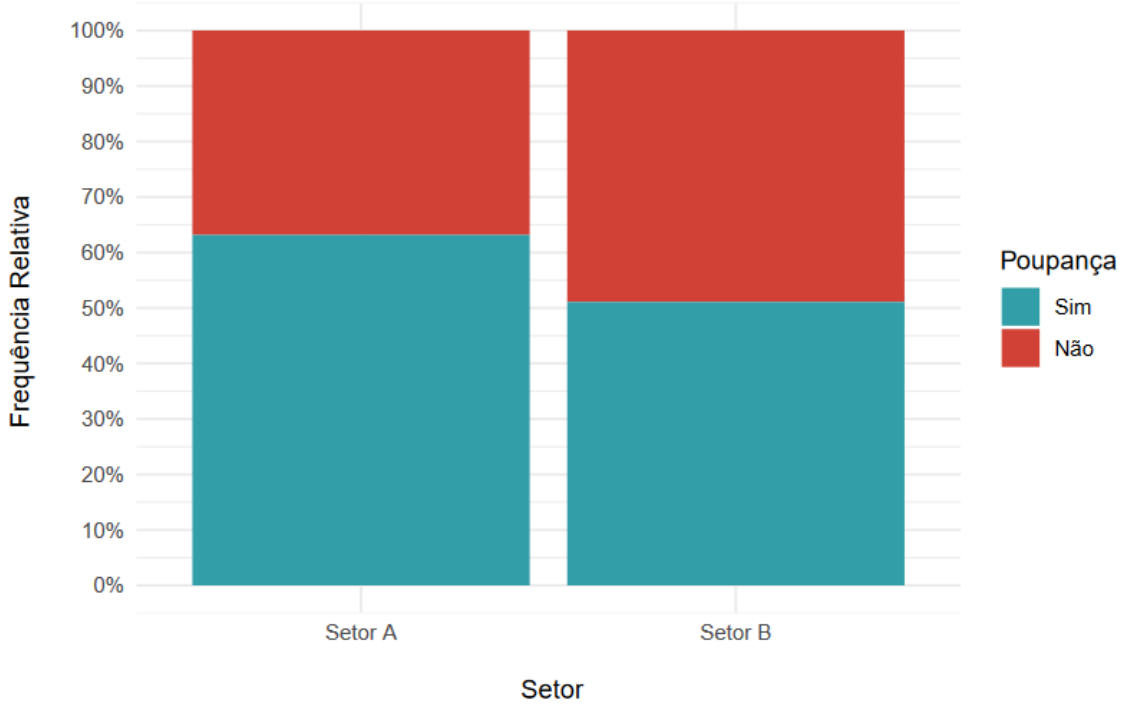
A relação entre os setores de habitação e a conta poupança também foi investigada. Para indivíduos do setor B, observamos que não houve diferença significativa na proporção de famílias com ou sem conta poupança. Por outro lado, esse cenário muda ao vermos que 63% das famílias do setor A afirmam ter conta poupança.

Não conseguimos definir uma hipótese clara, visto que não foram forncidas informações das características dos setores A e B da cidade. Caso o setor A seja a área mais nobre, essa informação é condizente com as análises anteriores que envolvem status socioeconômico, por exemplo, uma vez que novamente famílias com melhores condições tendem a poupar mais.

Tabela 8 – Distribuição das famílias com e sem conta poupança segundo o setor da cidade em que os pacientes residem - Brasil, 2023

Setor	**Conta Poupança**				**Total**
	Sim		**Não**		
	Freq. Absoluta	**Freq. Relativa**	**Freq. Absoluta**	**Freq. Relativa**	
Setor A	36	63%	21	37%	57
Setor B	71	51%	68	49%	139
Total	107	-	89	-	196

Figura 7 – Distribuição percentual do número de contas poupança segundo o setor da cidade em que os pacientes residem - Brasil, 2023



3.1.5 Análise Descritiva: Casa Própria

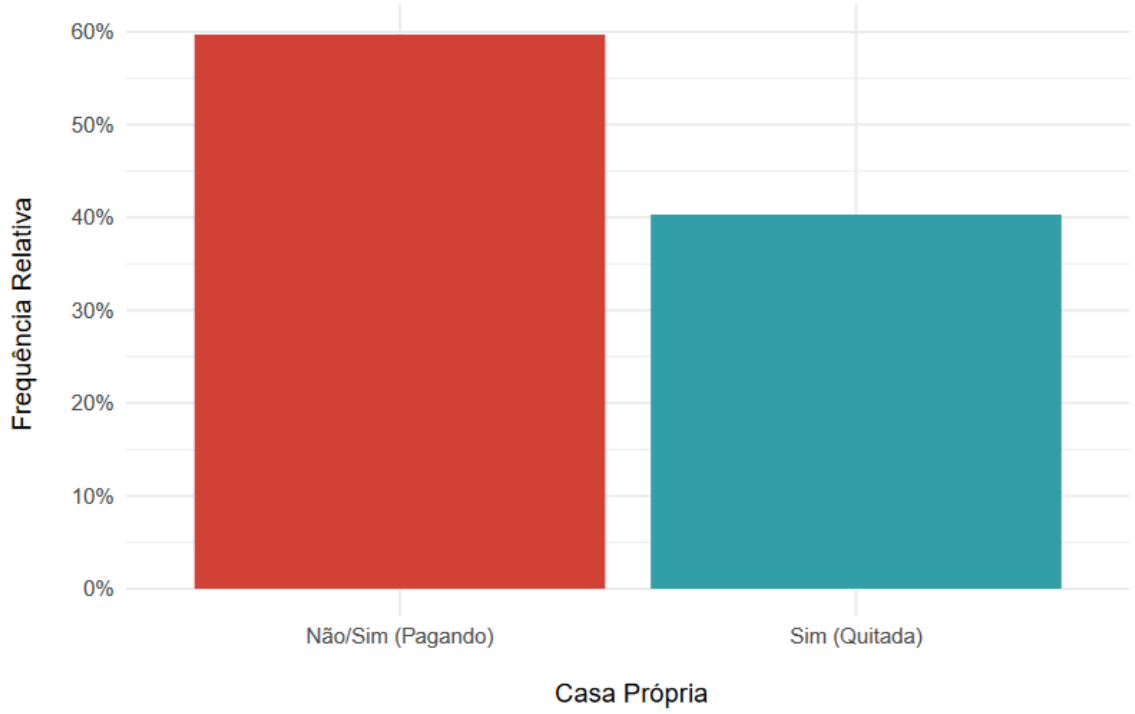
A variável casa própria pode ser caracterizada como nominal e está dividida em duas categorias: “Não ou Sim (pagando)"e “Sim (quitada)". No primeiro caso, temos as famílias que não possuem casa própria ou possuem, mas o financiamento está sendo pago. O segundo grupo é formado pelas famílias que possuem casa própria totalmente quitada. Trata-se de um variável, em certo grau, que nos informa sobre renda familiar.

Em síntese, 60% das famílias dos pacientes estão na categoria “Não ou Sim (pagando)"e 40% na categoria “Sim (quitada)". Em certa medida, essas proporções fazem sentido quando comparamos com os resultados obtidos na análise para o setor da cidade e status socioeconômico.

Tabela 9 – Distribuição do número de pacientes segundo a aquisição de casa própria pelas famílias - Brasil, 2023

Casa Própria	**Freq. Absoluta**	**Freq. Relativa (%)**
Não/Sim (Pagando)	117	60%
Sim (Quitada)	79	40%
Total	196	100%

Figura 8 – Distribuição percentual do número de pacientes segundo a aquisição de casa própria pelas famílias - Brasil, 2023



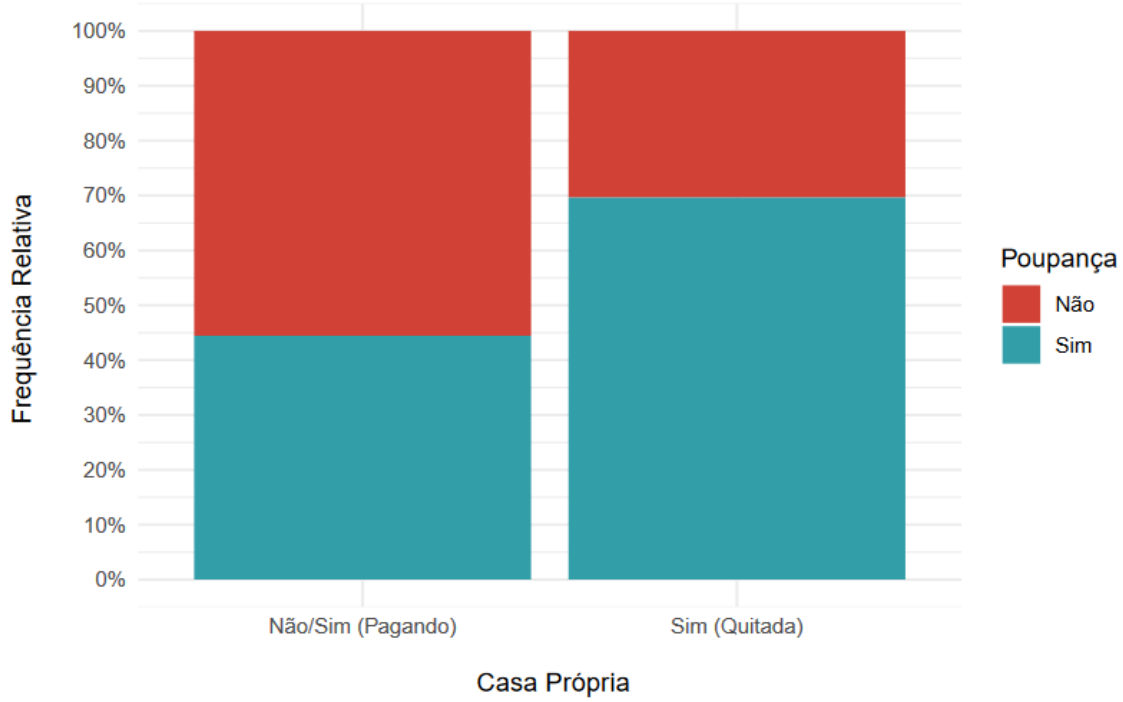
Quando avaliamos o grupo de pessoas na categoria “Não ou Sim (pagando)"para a casa própria, vemos que 56% não possuem conta poupança. Por outro lado, 70% das famílias com casa própria quitada possuem conta poupança. Portanto, parece existir uma associação entre as variáveis conta poupança e aquisição da casa própria.

Em certa medida, a hipótese é que as famílias com casa quitada possuem recursos disponíveis para abrir poupança.

Tabela 10 – Distribuição das famílias com e sem conta poupança segundo a aquisição de casa própria pela família dos pacientes - Brasil, 2023

Casa Própria	**Conta Poupança**				**Total**
	Sim		**Não**		
	Freq. Absoluta	**Freq. Relativa**	**Freq. Absoluta**	**Freq. Relativa**	
Não/Sim (Pagando)	52	44%	65	56%	117
Sim (Quitada)	55	70%	24	30%	79
Total	107	-	89	-	196

Figura 9 – Distribuição percentual do número de contas poupança segundo a aquisição de casa própria pela família dos pacientes - Brasil, 2023



Obs: em estudos futuros, seria importante separar a primeira categoria entre quem não tem casa própria e quem tem, mas ainda paga financiamento. Para famílias sem casa própria, a não existência de conta poupança deve ser, possivelmente, o padrão. Por outro lado, pessoas que ainda estão pagando financiamento eventualmente precisam de conta poupança no processo de pagamento. Além disso, não sabemos quantas famílias estão em cada situação, o que pode representar um viês nos resultados obtidos.

3.2 Modelagem

3.2.1 Seleção de Modelos

Uma vez que a análise exploratória aponta que existem relações entre o chefe da família ter poupança e as demais variáveis, a escolha de um modelo de regressão logística aparenta ser adequada para modelar e facilitar a obtenção de explicações e predições sobre as relações entre as variáveis preditoras e a variável resposta.

Para saber se de fato o modelo proposto pode ser bem ajustado pelos dados disponíveis, foram ajustados dois modelos, um modelo que contém apenas o intercepto e outro com um coeficiente para cada uma das variáveis preditoras disponíveis, o modelo nulo e o modelo saturado, respectivamente. Então foi feito o teste da Razão de Verossimilhança para saber se o modelo saturado era de fato adequado e pelo menos um dos coeficientes da regressão deveria ser não nulo.

H_0) O modelo nulo é o mais adequado; H_1) O modelo saturado é o mais adequado

$$\text{(Estatística do teste) } G^2 = -2(L_0 - L_1) \sim \chi^2_{(4)}$$
$$G^2_{obs} \approx 26.6498 \Rightarrow \text{p-valor} = P\left(G^2 \geq G^2_{obs}\right) < 0.0001$$

Logo, como o nível de significância estabelecido para este estudo é de 5%, rejeita-se a hipótese nula. Em outras palavras assume-se que o modelo nulo não é mais adequado que o modelo saturado, portanto, pelo menos um dos coeficientes da regressão deve ser não nulo, o que é uma evidência significativa de que existe regressão entre a variável resposta e as preditoras.

Analisando todos os modelos possíveis com pelo menos uma das variáveis preditoras obtém-se a figura 10, da qual é possível observar que há um decrescimento do AIC para alguns modelos que incluem mais parâmetros, atingindo o vale em 4 parâmetros, o que indica que o modelo selecionado apresenta no máximo 4 parâmetros. Analogamente para a acurácia, que apresenta resultado similar ao do AIC.

Da figura 10 é possível perceber que o AIC muda relativamente pouco de 3 parâmetros para 4 parâmetros, em outras palavras, espera-se que o modelo mais adequado tenha 3 ou 4 parâmetros contando com o intercepto.

Como há muito modelos possíveis, o teste da razão de verossimilhança foi repetido para todos eles, cujas hipóteses são

H_0) O modelo em estudo é o mais adequado; H_1) O modelo saturado é o mais adequado

Os modelos cujo teste não rejeitou que pudessem ser os modelos mais adequados estão apresentados na tabela 11, juntamente às respectivas estatísticas do teste, graus de liberdade e o p-valor obtido.

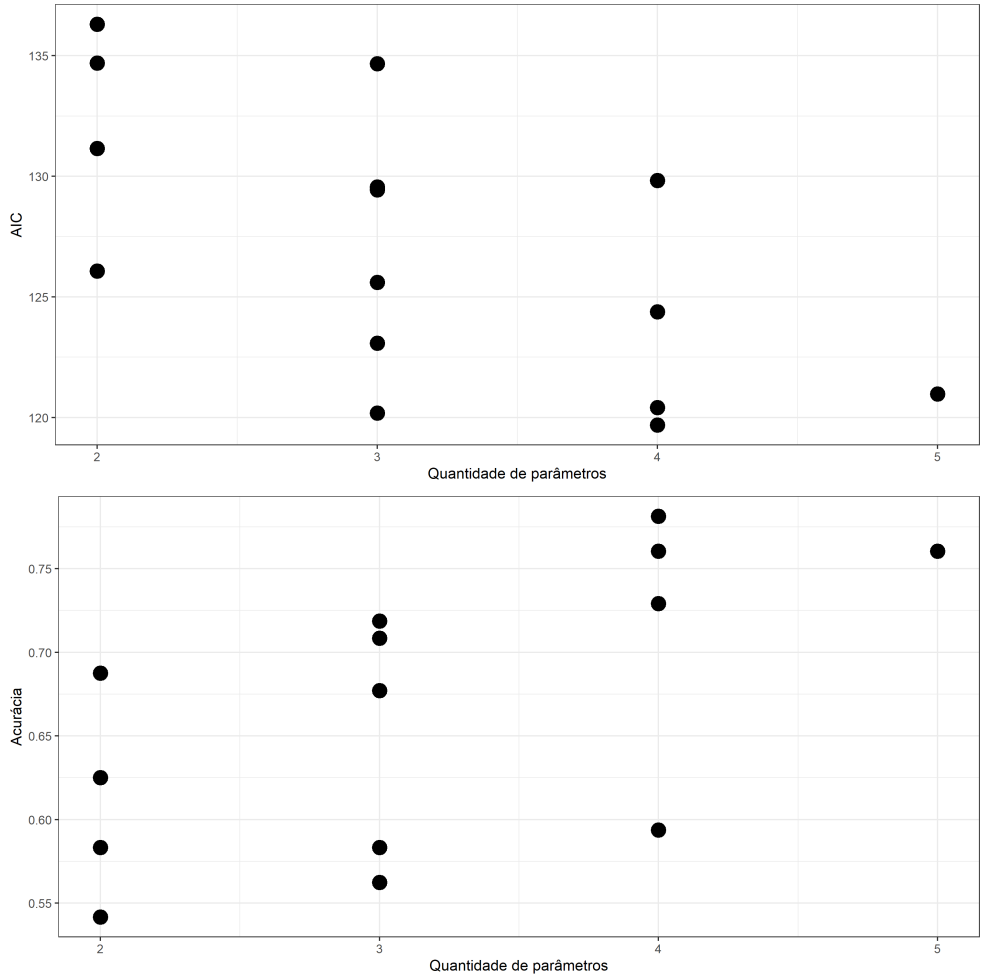


Figura 10 – Critérios de seleção de modelos para todos os modelos possíveis

Tabela 11 – Modelos não rejeitados pelo teste da razão de verossimilhança

Modelo	G^2	g.l.	p-valor	AIC	ACC
Poupanca ~ Idade + Socioecon	3.211	2	0.201	120.189	0.719
Poupanca ~ Idade + Socioecon + Casa	1.439	1	0.230	120.417	0.781
Poupanca ~ Idade + Socioecon + Setor	0.705	1	0.401	119.683	0.760
Poupanca ~ Idade + Socioecon + Casa + Setor	-	-	-	120.978	0.760

Da tabela 11 percebe-se que o modelo com menor AIC utiliza 4 parâmetros contando com intercepto e as variáveis Idade, Socioecon e Setor, esse mesmo modelo também foi o que apresentou maior p-valor para o teste da razão de verossimilhança e tem uma acurácia próxima ao modelo com melhor acurácia da tabela, que utiliza a variável Casa invés da variável Setor. Portanto, o modelo escolhido é o modelo da forma:

$$\textbf{Poupanca} \sim \textbf{Idade} + \textbf{Socioecon} + \textbf{Setor}$$

Ou, de maneira extensa:

$$Y_i \sim Bernoulli(\pi(\mathbf{x}_i))$$
$$\pi(\mathbf{x}_i) = \frac{exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})}{1 + exp(\beta_0 + \beta_1 x_1 i + \beta_2 x_{2i} + \beta_3 x_{3i})}$$

Onde Y é a variável poupança, x_1 é a variável idade, x_2 é a variável status socioeconômico e x_3 é a variável setor.

3.2.2 Análise de Resíduos

Tipicamente os resíduos dados pela definição de valor real menos valor ajustado não são muito informativos para a regressão logística (GELMAN; HILL, 2006), portanto a análise de resíduos do modelo selecionado foi feita por meio de um *binned plot*, pois esse gráfico evidencia de forma mais informativa padrões e possíveis problemas com os resíduos (WEBB, 2017).

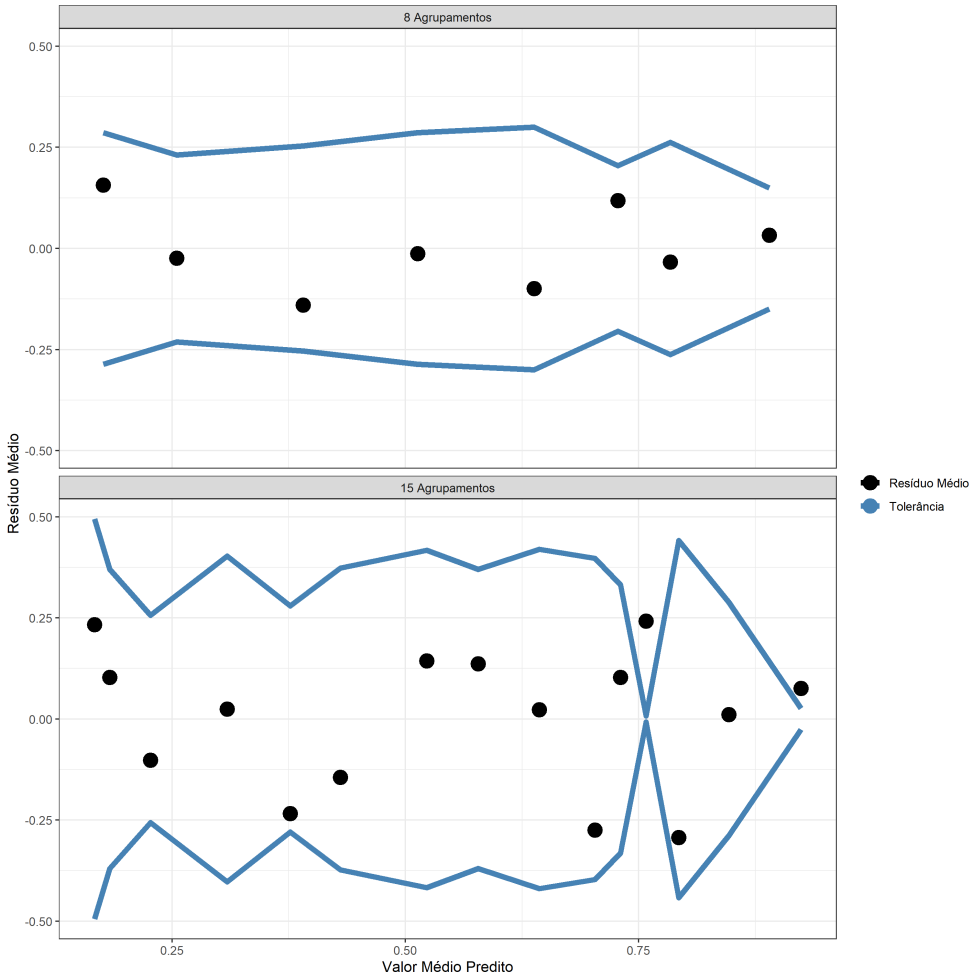


Figura 11 – binned plot com 8 e 15 agrupamentos para os resíduos do modelo ajustado, respectivamente

Da figura 11 conclui-se que os resíduos do modelo apresentam bom comportamento, pois os resíduos médios estão contidos dentro das faixas de tolerância de 2 desvios-padrões

para todas os agrupamentos. No gráfico com 15 agrupamentos, dois dos pontos estão além das faixas de tolerância, porém a faixa ficou muito próxima de 0, indicando que poucos resíduos foram utilizados naquele agrupamento e que um número menor de grupos de fato é o mais adequado para essa visualização.

3.2.3 Análise de Falta de Ajustamento

Para análisar a possibilidade do modelo selecionado estar com falta de ajsutamento, foi feito o teste de Hosmer Lemeshow, dividindo os valores preditos em 10 grupos a partir dos decis, os grupos estão apresentados na tabela 12.

Tabela 12 – Grupos obtidos para o teste de Hosmer e Lemeshow para o modelo selecionado

Grupo	Tamanho	Observado	Esperado
1	10	4	1.740
2	10	1	2.191
3	11	2	3.709
4	10	4	4.273
5	10	7	5.529
6	10	5	6.425
7	10	7	7.180
8	10	9	7.583
9	11	8	9.023
10	8	8	7.347

O teste segue com as hipóteses

H_0) Os valores ajustados e esperados não divergem nos grupos obtidos

H_1) Os valores ajustados e esperados divergem nos grupos obtidos

A estatística do teste é dada por

$$\chi^2 = \sum_{i=1}^{10} \sum_{j=0}^{1} \frac{(f_{ji} - fe_{ji})^2}{fe_{ji}} \sim \chi^2_{(6)}$$

O valor obtido para a estatística do teste foi de 9.83 e seu p-valor foi de 13.21%, acima do nível de significância de 5%. Portanto não se pode rejeitar H_0, isto é, considera-se que os valores observados e esperados obtidos para os grupos não divergem significativamente, caracterizando que não há falta de ajustamento do modelo.

3.2.4 Análise de Desempenho Preditivo

Como já visto da tabela 11 tem-se que a acurácia de teste do modelo escolhido é de cerca de 76%, com ponto de arredondamento em aproximadamente 0.5417. O intervalo

de confiança de 95% para a acurácia é dado por

$$IC(ACC,\ 95\%) \approx (65.34\%,\ 83.12\%)$$

O intervalo de confiança para a acurácia está estritamente acima de 50%, o que é um ótimo sinal, pois isso indica que o modelo alcançado é melhor que um palpite ao acaso e pode ser utilizado para predição.

Mais detalhadamente pode-se estudar a matriz de confusão da tabela 13, que indica que as respostas estão balanceadas, de forma que a quantidade de falsos negativos e falsos positivos está bem próxima.

Tabela 13 – Matriz de confusão para o modelo selecionado

		Resposta Verdadeira		
		Sim	Não	Total
Predição	Sim	34	13	47
	Não	10	39	49
	Total	44	52	96

Sob o mesmo ponto de corte, no conjunto de teste o modelo apresenta sensibilidade de 75% e especificidade pouco acima de 77%, um resultado condizente com a tabela 13, sendo assim mais um sinal de que o ajuste do modelo foi bom e que o ponto de corte escolhido de fato está adequado, pois criou um bom balanço entre sensibilidade e especificidade.

Da figura 12 pode-se observar que a área sob a curva (AUC) foi 77.2%, reforçando que o ajuste do modelo apresenta uma boa performance para o conjunto de teste, porém o ponto de corte que máximiza a soma de especificidade e sensibilidade na verdade é de cerca de 0.569 (arredondado para 0.6 no gráfico), indicando que pode ser um ponto importante a se considerar rever o ponto de corte escolhido para estudos futuros com objetivos mais bem definidos, este estudo segue assumindo que o mesmo ponto de corte de 0.5417.

Obs: Caso a curva ROC fosse feita para o conjunto de treino, então o ponto ótimo de corte para curva coincidiria com o ponto ótimo de corte para a acurácia.

3.2.5 Interpretação do Modelo

Os parâmetros do modelo ajustado e seus respectivos testes de significância estão dispostos na tabela 14, em que as hipóteses dos testes são dadas por

$$H_0)\beta_i = 0; \quad H_1)\beta_i \neq 0$$

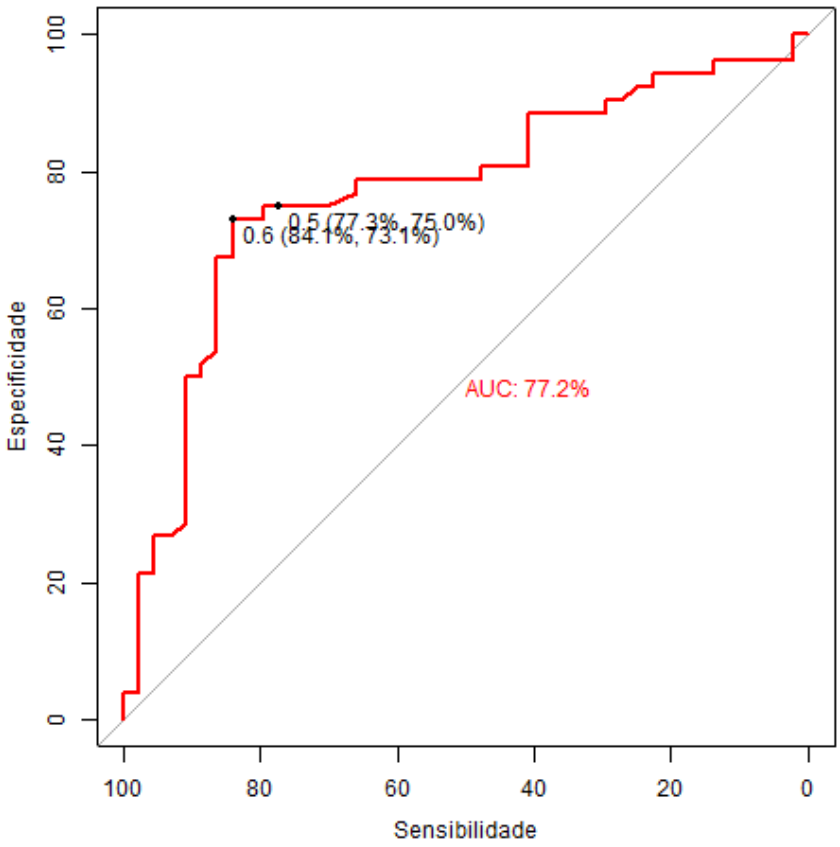


Figura 12 – Curva ROC para o modelo selecionado sob o conjunto de teste

e estatística do teste é dada por

$$Z = \hat{\beta}_i / ASE(\hat{\beta}_i) \sim N(0,1)$$

onde $\beta_i, i = 0,1,2,3$ é a estimativa para o i-ésimo parâmetro da regressão e ASE($\hat{\beta}_i$) é o seu respectivo erro padrão.

Tabela 14 – Estimativas dos parâmetros para o modelo selecionado

	Estimativa	Erro Padrão	Z	p-valor
(Intercepto)	0.985	0.669	1.472	0.141
Idade	0.032	0.014	2.235	0.025
Socioecon	-0.894	0.271	-3.301	0.001
Setor	0.818	0.525	1.558	0.119

Da tabela 14 tem-se que as estimativas dos parâmetros correspondentes à variável Setor e ao Intercepto não apresentaram valores significativamente diferentes de 0, ou seja, não pode-se descartar a possibilidade de ambos não estarem presentes no modelo.

É importante lembrar que toda a análise desde o processo de seleção de modelos até o momento sustenta que a presença do intercepto e da variável Setor são adequadas

para o modelo, pois os resíduos apresentaram bom comportamento e o modelo obteve bom desempenho preditivo.

Como o intercepto teve o p-valor mais elevado no teste de significância, a decisão tomada foi de ajustar o modelo sem intercepto para verificar que impactos essa mudança traria ao modelo.

A tabela 15 aponta que em termos de direção as estimativas para os parâmetros continuas as mesmas, por conta da preservação dos sinais, entretanto, o p-valor do teste de significância para o coeficiente da variável setor continua não rejeitando que o coeficiente pode ser nulo, porém com um p-valor muito próximo ao nível de significância de 5%.

Tabela 15 – Estimativas dos parâmetros para o modelo sem intercepto

	Estimativa	Erro Padrão	Z	p-valor
Idade	0.041	0.013	3.222	0.001
Socioecon	-0.582	0.164	-3.548	<0.001
Setor	0.945	0.505	1.870	0.061

Também foi considerado o modelo sem intercepto e sem a variável Setor. De forma análoga a tabela 16 aponta que os sinais dos coeficientes são preservados, porém ambas as variáveis rejeitam a possibilidade de serem nulas pelo teste de significância.

Tabela 16 – Estimativas dos parâmetros para o modelo sem intercepto e sem a variável Setor

	Estimativa	Erro Padrão	Z	p-valor
Idade	0.048	0.012	3.916	<0.001
Socioecon	-0.538	0.157	-3.417	0.001

Em termos de desempenho preditivo observa-se que da tabela 17 que o modelo selecionado com intercepto e com a variável Setor com certeza apresenta o melhor desempenho preditivo em termos de acurácia, sensibilidade e especificidade. Embasado nesse ganho de desempenho preditivo e a perfeita adequação do modelo vista pela análise de resíduos, tomou-se a decisão de manter o modelo selecionado.

Tabela 17 – Comparação de desempenho preditivo entre os modelos possíveis.

Modelo	Acurácia	Sensibilidade	Especificidade
Poupanca ~ Idade + Socioecon + Setor	76%	75%	77%
Poupanca ~ 0 + Idade + Socioecon + Setor	72%	71%	73%
Poupanca ~ 0 + Idade + Socioecon	68%	64%	73%

Os coeficientes do modelo selecionados estão apresentados na tabela 18, juntamente aos seus respectivos limites do intervalo de confiança de 95%.

Tabela 18 – Estimativas pontuais e limites de 95% de confiança para os coeficientes da regressão

	Estimativa Pontual	Limite Inferior	Limite Superior
(Intercepto)	0.985	-0.326	2.296
Idade	0.032	0.004	0.059
Socioecon	-0.894	-1.425	-0.363
Setor	0.818	-0.211	1.847

Sabe-se da regressão logística que interpretar os parâmetros de forma direta é uma tarefa difícil, por tanto a interpretação fica por conta do *odds ratio*, obtido tomando a exponencial de todos os valores da tabela 18 e apresentados na tabela 19 (AGRESTI, 2007; GELMAN; HILL, 2006)

Tabela 19 – Estimativas pontuais e limites de 95% de confiança para o efeito multiplicativo no odds ratio para cada coeficiente sob aumento de uma unidade

	Estimativa Pontual	Limite Inferior	Limite Superior
Idade	1.032	1.004	1.061
Socioecon	0.409	0.241	0.695
Setor	2.266	0.810	6.338

Os resultados da tabela 18 apontam que o Intercepto é o coeficiente com maior variação em sua estimativa, além de atingir o maior limite superior entre os parâmetros. Note que no contexto do problema, esse parâmetro tem apenas a interpretação de ser o intercepto do modelo.

Da tabela 19, especificamente para a idade, interpreta-se que o aumento de um ano de idade do paciente deve aumentar a chance do chefe da família em cerca de 1.032 vezes, variando de 1.004 a 1.061 com 95% de confiança, mantido as demais variáveis constantes. É importante notar que a variável Idade foi a única que apresentou limite inferior acima de 1 para o *odds ratio*, isso significa que um aumento da idade sempre contribui para o aumento da chance do chefe da família possuir poupança.

Para as variáveis Socioecon e Setor, uma vez que são variáveis categóricas, sua interpretação é um pouco diferente. Interpreta-se que a descida de status socioeconômico sempre diminui a razão de chances do chefe da família possuir poupança, mantido as demais variáveis constantes, o que é um resultado condizente com a intuição, se o chefe da família está em um maior status socioeconômico, então possui mais dinheiro guardado. Para o Setor interpreta-se que a mudança do Setor A para o Setor B aumenta a razão de chances em 2.266 vezes variando de 0.810 vezes a 6.338 vezes com 95% de confiança, mantido as demais variáveis constantes.

4 Conclusão

Na análise exploratória dos dados, percebemos que as variáveis idade do paciente, status socioeconômico e setor de habitação aparentam ser fatores significativos associados à obtenção de conta poupança pelas famílias dos pacientes da rede hospitalar. Nesse sentido, o modelo selecionado entrou em acordo com a análise exploratória e suas estimativas apontam que quanto maior a idade do paciente, status da família e moradia no setor A da cidade, maior a proporção de respostas "Sim"para a existência de conta poupança familiar. Com relação à variável casa própria, observamos que famílias com imóveis quitados possuem maior tendência em adquirir conta poupança, entretanto, a análise é menos informativa por não haver separação entre quem não tem casa própria e quem tem, mas ainda paga financiamento.

Para a modelagem dos dados, selecionamos um modelo logístico que apresenta excelente ajustamento, resíduos condizentes e eficiente desempenho preditivo. Em suma, reafirmamos que a variável conta poupança pode ser explicada pelas variáveis idade do paciente, status socioeconômico e setor de habitação da família. A variável casa própria não foi considerada significativa no modelo, entretanto, cabe ressaltar que em futuras análises é fundamental redefinir as suas categorias para melhor avaliação do estudo.

5 Apêndice

Código fonte disponível em https://github.com/Voz-bonita/Dados_Categorizados

Referências

AGRESTI, A. *An Introduction to Categorical Data Analysis*. [S.l.]: John Wiley Sons, 2007. Citado na página 23.

GELMAN, A.; HILL, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. [S.l.]: Cambridge University Press, 2006. Citado 2 vezes nas páginas 18 e 23.

WEBB, J. *Course Notes for IS 6489, Statistics and Predictive Analytics*. [S.l.]: University of Utah's David Eccles School of Business, 2017. Citado na página 18.