UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

THAII CÉU SANTOS

TRABALHO DE MODELOS LINEARES GENERALIZADOS: Modelagem com variáveis relacionadas ao AVC

PRESIDENTE PRUDENTE
2025

THAII CÉU SANTOS

TRABALHO DE MODELOS LINEARES GENERALIZADOS: Modelagem com variáveis relacionadas ao AVC

PRESIDENTE PRUDENTE

2025

Sumário

1 INTRODUÇÃO

O acidente vascular cerebral (AVC) é uma doença recorrente entre adultos e corresponde à segunda principal causa de morte no mundo, além de ser a maior responsável pelas incapacidades que comprometem as atividades da vida diária. Segundo a Organização Mundial da Saúde, a cada ano 15 milhões de pessoas são acometidas pelo AVC; desse total, 5 milhões vêm a óbito em decorrência do evento, enquanto a maior parte dos sobreviventes apresenta sequelas tanto físicas quanto mentais. Entre os pacientes, 37% manifestam alterações discretas, 16% apresentam incapacidade moderada e 32% enfrentam comprometimento intenso ou grave da capacidade funcional — sendo que algumas pessoas tornam-se dependentes de cadeira de rodas ou permanecem acamadas. Tais sequelas representam um grande impacto econômico, social e familiar. Apenas 15% dos pacientes não apresentam qualquer dano na capacidade funcional.

Diversos são os principais fatores de risco para o aparecimento do AVC, como a hipertensão arterial, diabetes, tabagismo, consumo excessivo de álcool e outras drogas, estresse, colesterol alto, doenças cardíacas — principalmente as que causam arritmia —, sedentarismo e algumas doenças do sangue. Outros, como a idade, também representam um fator importante, sendo o AVC mais recorrente depois dos 55 anos. A origem étnica, principalmente a negra, e o histórico familiar de doenças cardíacas também aumentam o risco. Por isso, pessoas que apresentam esses riscos devem fazer um controle médico mais constante e ter maior consciência da doença e de seus sintomas.

Os métodos de regressão tornaram-se uma ferramenta essencial em qualquer análise de dados voltada a descrever a relação entre uma variável resposta e uma ou mais variáveis explicativas. Frequentemente, a variável de desfecho é discreta, assumindo dois ou mais valores possíveis. Nesses casos, o modelo de regressão mais utilizado é o da regressão logística.

Com o passar do tempo, os inquéritos populacionais de saúde vêm sendo cada vez mais usados não apenas para avaliar o funcionamento da assistência de saúde pelo ponto de vista do usuário, como também para obter informações sobre a morbidade referida e os estilos de vida da população. Quando realizados periodicamente, esses inquéritos permitem consolidar uma base de dados

populacionais, que é importante para o monitoramento de doenças crônicas e seus determinantes.

Dentro desse contexto, a Pesquisa Nacional de Saúde (PNS), como parte de um projeto do Ministério da Saúde destinado ao estudo das condições de saúde da população brasileira e à avaliação do desempenho do sistema nacional de saúde, teve como principal objetivo produzir dados, em escala nacional, sobre o estado de saúde da população, seus estilos de vida e o uso de serviços de saúde. Além disso, a PNS proporciona informações sobre o acesso e a utilização de serviços, as ações preventivas, a continuidade do cuidado e o financiamento da assistência.

A regressão logística, tem como o mesmo objetivo de qualquer outro modelo de regressão: encontrar o modelo mais adequado, parcimonioso e clinicamente interpretável, que descreva a relação entre uma variável dependente (ou resposta) e um conjunto de variáveis independentes (ou explicativas), também chamadas de covariáveis.

O exemplo mais comum de modelagem é a regressão linear, em que se assume que a variável de desfecho é contínua. A principal distinção entre a regressão logística e a linear está no tipo de variável resposta: enquanto a regressão linear lida com desfechos contínuos, a regressão logística é empregada quando a variável resposta é binária ou dicotômica.

Essa diferença entre os dois modelos se reflete tanto na estrutura matemática do modelo quanto em seus pressupostos estatísticos. No entanto, uma vez compreendida essa distinção, os procedimentos analíticos da regressão logística seguem, em geral, os mesmos princípios adotados na análise por regressão linear.

Com base nestas informações, o presente estudo propõe evidenciar e reforçar a identificação dos fatores associados ao AVC por meio do modelo de Regressão Logística, com a base de dados obtida pelo PNS.

2 METODOLOGIA

Como escrito anteriormente, iremos utilizar a base de dados disponibilizado pelo PNS de 2019, onde selecionamos algumas variáveis que possa estar correlacionado ao AVC: Diabetes (Sim, no caso de ter ; Não, caso contrário) ; Sexo (Homem ou Mulher) ; Fumar (Sim, diariamente ; Sim, menos que diariamente ; Não fumo atualmente) ; Álcool (Não bebo nunca ; Menos de uma vez por mês ; Uma vez ou mais por mês) ; Hipertensão (Sim, no caso de ter ; Não, caso contrário) ; Diabetes (Sim, no caso de ter ; Não, caso contrário) ; Cor/Raça (Branca ; Preta ; Amarela ; Parda ; Indígena) ; AVC, a variável resposta (Sim, para caso tenha ; Não, caso contrário) ; Peso (em quilogramas) e Idade (em anos). Vale ressaltar que, para esta base de dados, houve 6.326 respostas para AVC, porém identificamos alguns NA's na variável Hipertensão e Idade (no total, 15 NA's), como era poucos, resolvi retira-las da análise, ou seja, estamos trabalhando com 6.311 respostas ("403" para "sim", teve AVC, 5.908 para "não").

Utilizamos alguns métodos estatísticos que iremos explicar em diante (Regressão Logística). Para fazer as análises necessárias, utilizamos o software R.

2.1 Modelo de Regressão Logística

A regressão logística é um modelo estatístico amplamente utilizado para analisar situações em que a variável dependente é dicotômica, ou seja, assume apenas dois valores distintos, normalmente representados como 0 e 1. Esse tipo de regressão tem como objetivo modelar a probabilidade de ocorrência de um evento de interesse em função de um ou mais preditores, também chamados de variáveis explicativas ou covariáveis.

Na regressão logística múltipla, considera-se um vetor de covariáveis $x = (x_1, x_2, \dots, x_p)$, associado à probabilidade $\pi(x) = P(Y = 1|\, x)$. O modelo é definido pela seguinte relação funcional:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

A expressão acima representa a função logit, ou seja, o logaritmo da razão de chances (odds) da ocorrência do evento. A transformação logit garante que a relação entre a variável resposta e os preditores seja linear no log-odds, respeitando a

natureza binária da variável resposta. A função inversa da logit é a função logística, que fornece a probabilidade estimada:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

Os coeficientes β_p são estimados por método de máxima verossimilhança, cuja função é dada por:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\pi(x_i)) + (1-y_i)\log(1-\pi(x_i))]$$

Esse processo requer algoritmos iterativos, como Newton-Raphson ou IRLS (Iteratively Reweighted Least Squares), pois não existe uma solução analítica fechada para os coeficientes.

Cada coeficiente β_p tem uma interpretação direta: representa a mudança no logaritmo da razão de chances de $Y = 1$ para um aumento unitário em x_p , mantendo as demais variáveis constantes. A razão de chances associada à variável x_p é dada por:

$$OR_j = e^{\beta_p}$$

Apesar da robustez do modelo logístico múltiplo, sua aplicação em contextos de alta dimensionalidade (quando o número de variáveis explicativas é grande ou maior que o número de observações) pode levar a problemas como super ajuste, instabilidade nas estimativas e dificuldade de interpretação.

O código do trabalho:

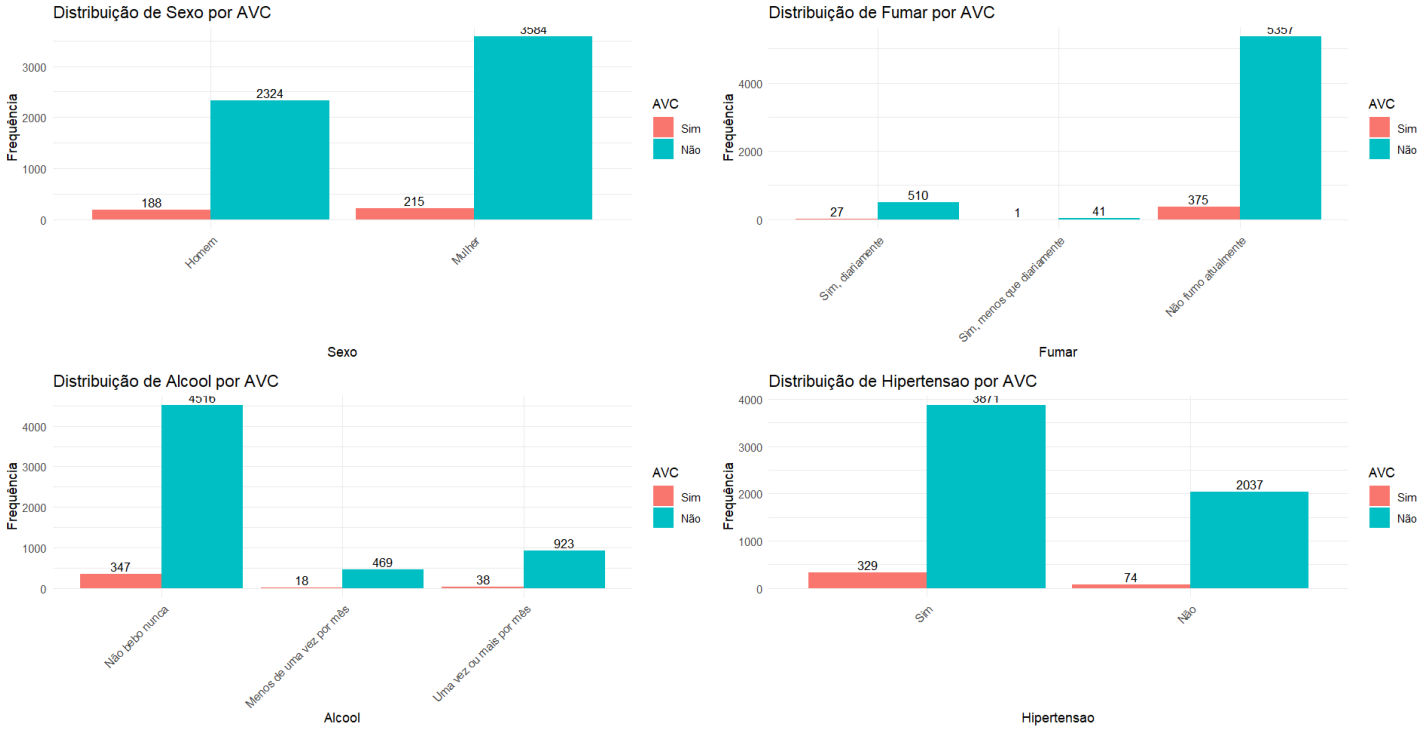
Código do trabalho (MLG).R

3 RESULTADOS

3.1 Análise Descritiva

Vamos começar com as variáveis qualitativas:

Figura 1 – Gráfico de barras da relação das variáveis com AVC



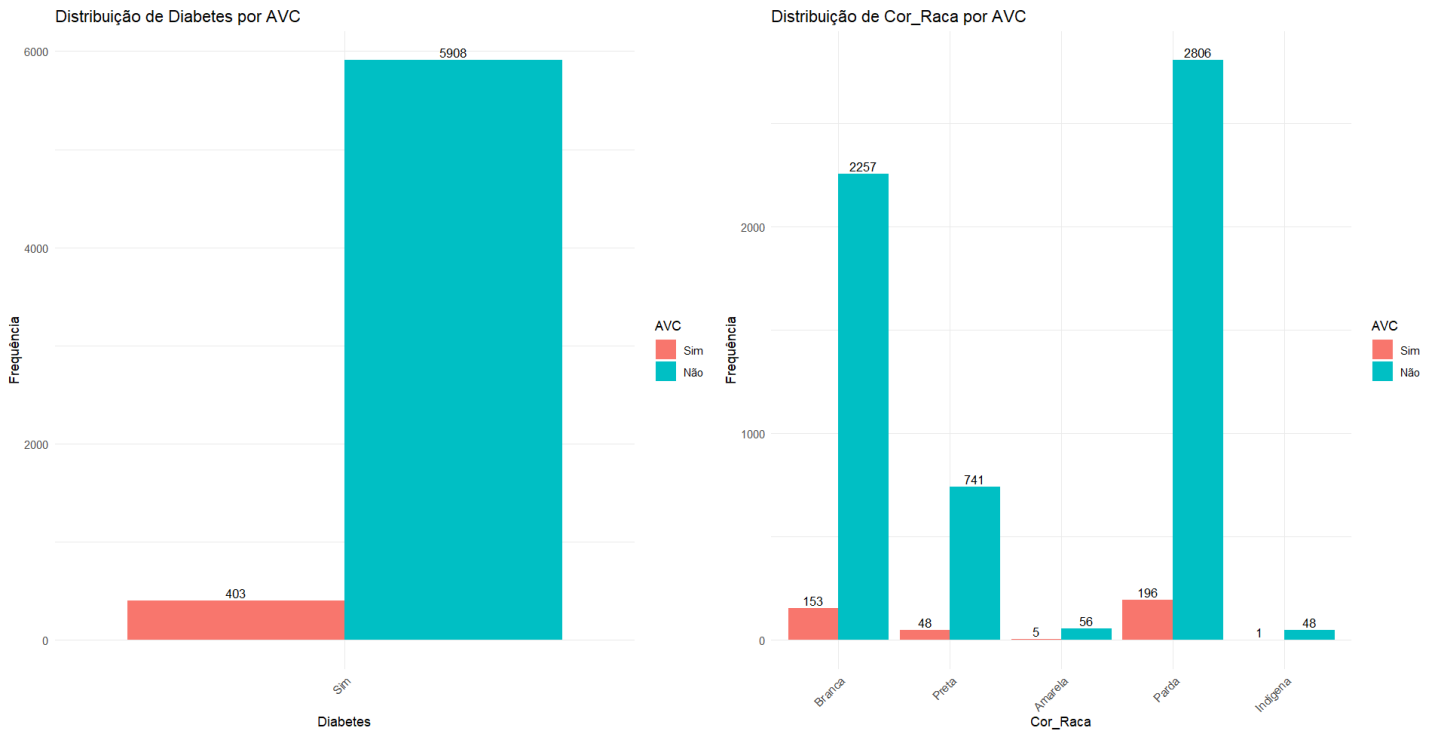
Na Figura 1, no primeiro gráfico que relaciona o sexo ao AVC, observa-se que o número de casos foi maior nas mulheres (215) em relação aos homens (188), provavelmente pelo maior tamanho da população feminina na amostra.

Em seguida, o segundo gráfico relaciona o tabagismo ao AVC. Ele revela que, entre as pessoas que fumam diariamente, houve 27 casos de AVC; entre as que fumam, mas com pouca frequência, houve 1 caso; e, entre as que não fumam atualmente, foram 375 casos. Isso significa que o AVC ocorreu tanto em pessoas que fumam quanto nas que não fumam, sendo o grupo de não-fumantes o maior em números absolutos.

Logo depois, o terceiro gráfico relaciona o consumo de álcool ao AVC. Ele revela que, entre as pessoas que declararam que não bebem, houve 347 casos de AVC; entre as que bebem menos de uma vez ao mês, foram 18 casos; e, entre as que bebem uma ou mais vezes ao mês, houve 38 casos. Dessa forma, o AVC ocorreu tanto em pessoas que consumem quanto nas que não consumem álcool, sendo o grupo que não bebe o maior.

Por fim, o quarto gráfico relaciona a doença hipertensão ao AVC. Ele revela que, entre pessoas que são hipertensas, houve 329 casos de AVC, enquanto entre as que não são hipertensas houve 74 casos. Isso evidencia uma associação importante entre a doença e o aparecimento do AVC, sendo a doença um importante fator de risco.

Figura 2 – Gráfico de barras da relação das variáveis com AVC

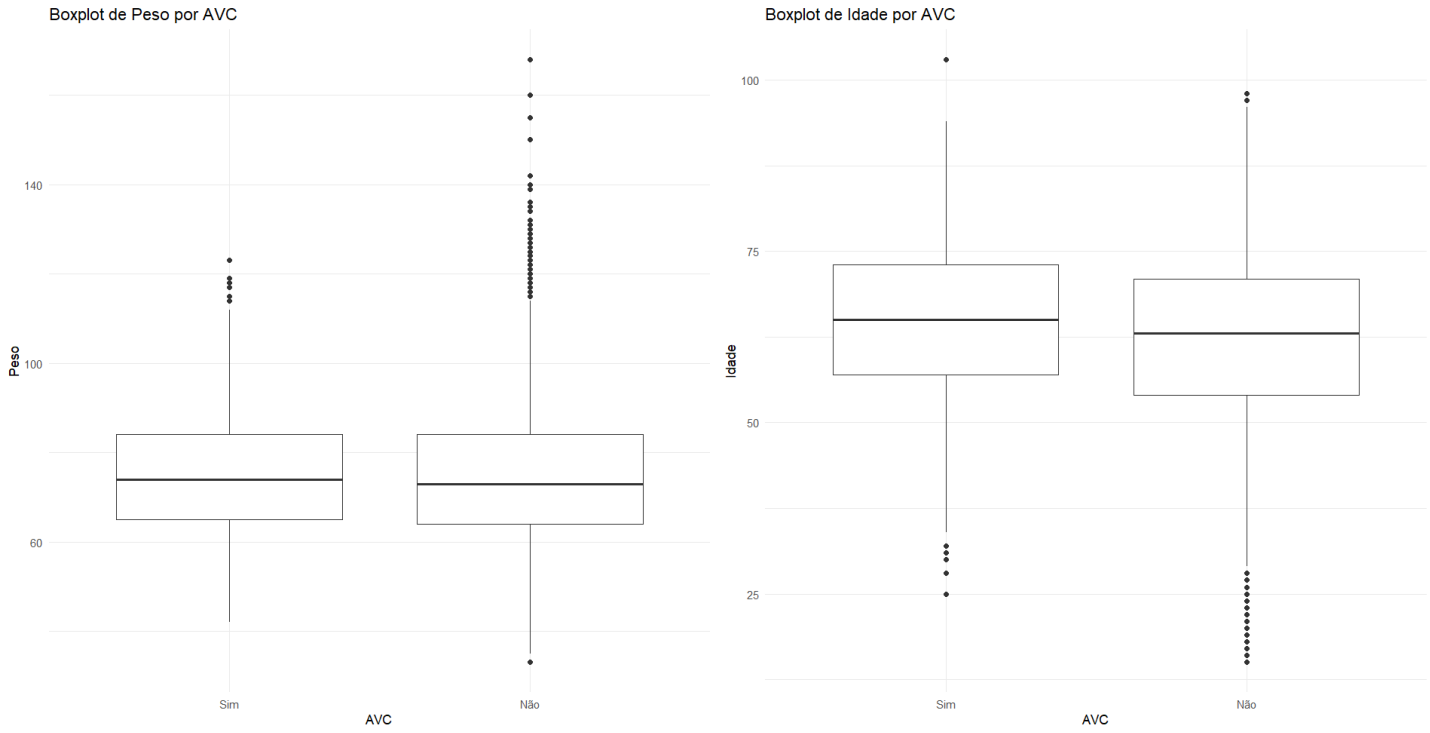


Na Figura 2, primeiramente, no gráfico analisando a diabete, conseguimos notar que não houve pessoas que não tinha diabete, ou seja, todas as pessoas da amostra continham diabete, onde 403 delas tiveram AVC, e as outras 5908, não tiveram. Para esta variável, não iremos continuar para o modelo, já que a resposta foi unanime

Em seguida, no gráfico que relaciona a cor ou raça ao AVC, verifica-se que, entre as pessoas brancas, houve 153 casos de AVC; nas pretas, foram 48; nas amarelas, 5; nas pardas, 196; e nas indígenas, 148. Em todos esses grupos, o número de pessoas que não sofreram AVC é maior, sendo 2.257 brancas, 741 pardas, 2.806 indígenas, 56 amarelas e 5 pretas. Dessa forma, o AVC ocorreu em todos os grupos populacionais, sendo o maior número de casos observado nas pessoas pardas, provavelmente pelo tamanho maior dessa população na amostra.

Partindo para as variáveis quantitativas:

Figura 3 – Boxplot da relação das variáveis com AVC



Na Figura 3, no gráfico que relaciona o peso ao AVC, verifica-se que tanto o grupo que teve AVC quanto o que não teve apresenta uma mediana de peso relativamente próxima, em torno de 70 a 80 quilos. No entanto, o grupo que não teve AVC apresenta maior dispersão, sendo que existem algumas pessoas com pesos muito baixos (próximo de 40) e outras com pesos muito altos (acima de 140), enquanto o grupo que teve AVC apresenta uma variação um pouquinho menor.

Em seguida, no boxplot que relaciona a idade ao AVC, observa-se que o grupo que teve AVC apresenta uma mediana de idade maior, próximo aos 75 anos, enquanto o grupo que não teve AVC apresenta uma mediana um pouco mais baixa com quem teve AVC. Além disso, o grupo que não teve AVC apresenta uma maior dispersão, sendo formado tanto por pessoas muito jovens quanto muito idosas, enquanto o grupo que teve AVC apresenta uma concentração maior de pessoas na meia-idade e na terceira idade.

3.2 Modelo de Regressão Logística

Avaliando primeiramente o modelo completo, lembrando que tiramos a variável “Diabetes” pois não apresentava variações (Todos os que responderam a pesquisa, tem diabete). Para as variáveis numéricas (Peso e Idade) normalizamos estas variáveis para ficar mais fácil na modelagem:

Tabela 1 – Estatísticas das variáveis (Modelo Completo)

Variável	Estimativa (β)	Erro-Padrão	Valor-p
Intercepto	2.21331	0.22668	< 2e-16 ***
Fumar (Sim, menos que diariamente)	0.72273	1.03613	0.48547
Fumar (Não fumo atualmente)	- 0.11245	0.20803	0.58881
Álcool (Menos de 1x/mês)	0.69643	0.24977	0.00530 **
Álcool (1x/mês ou mais)	0.72314	0.18279	7.62e-05 ***
Peso	- 0.07087	0.05659	0.21045
Sexo (Mulher)	0.48667	0.11238	1.49e-05 ***
Hipertensão (Não)	0.79396	0.13624	5.62e-09 ***
Idade	- 0.16035	0.05940	0.00695 **
Raça (Preta)	0.01752	0.17282	0.91924
Raça (Amarela)	- 0.35394	0.47965	0.46057
Raça (Parda)	- 0.05350	0.11406	0.63904
Raça (Indígena)	1.05742	1.01690	0.29841

Avaliando os resultados da Tabela 1, em relação ao hábito de fumar, o modelo revela que fumar menos que diariamente (β = 0,72; p = 0,49) e não fumar atualmente (β = -0,11; p = 0,59) não apresenta uma associação estatisticamente significativa com o AVC.

Com relação ao álcool, tanto o consumo de menos de uma vez ao mês (β = 0,7; p = 0,005) quanto o de uma vez ao mês ou mais (β = 0,72; p < 0,0001) estão associados a um maior risco de AVC, sendo esses resultados estatisticamente significativos.

Em relação ao peso, o parâmetro é negativo (β = -0,07), porém não apresenta significância (p = 0,21), sendo assim o peso não parece ter um papel relevante neste modelo.

Com relação ao sexo, ser mulher (β = 0,49; p < 0,0001) apresenta uma associação positiva e significativa ao AVC.

Ainda sobre a hipertensão, a ausência de doença (β = 0,79; p < 0,0001) relaciona-se a um maior risco de AVC, o que é um resultado inesperado e que deve ser interpretado com cuidado.

Com relação à idade, verifica-se uma associação negativa (β = -0,16; p = 0,007), mostrando que o risco de AVC diminui levemente a cada aumento de um ponto na idade.

Em relação à raça/cor, o modelo revela que preto (β = 0,02; p = 0,92), amarelo (β = -0,35; p = 0,46), pardo (β = -0,05; p = 0,64) e indígena (β = 1,06; p = 0,29) não apresentam associação estatisticamente significativa com o AVC. Uma explicação para estes resultados, pode ser por conta da amostragem feita pelo PNS, ou as nossa variável resposta, pois temos muitos não's do que sim's, visto anteriormente.

Utilizando o método de seleção de variáveis (StepWise):

Tabela 2 – Estatísticas das variáveis (Modelo StepWise)

Variável	**Estimativa (β)**	**Erro-Padrão**	**Valor-p**
Intercepto	2.05699	0.09182	< 2e-16 ***
Álcool (Menos de 1x/mês)	0.69545	0.24929	0.00528 **
Álcool (1x/mês ou mais)	0.71692	0.18098	7.45e-05 ***
Sexo (Mulher)	0.52643	0.10762	1.00e-06 ***
Hipertensão (Não)	0.82159	0.13460	1.03e-09 ***
Idade	- 0.14349	0.05655	0.01118 *

As interpretações para as variáveis selecionadas na Tabela 2, são as mesmas da Tabela 1, nos deixando curioso com os resultados. Acredito que seja pela amostragem feita pela PNS, ou a quantidade de não's que temos. Avaliando as OR's:

Tabela 3 – OR's das variáveis (Modelo StepWise)

Variável	**OR**	**IC 95% (Inferior - Superior)**
Intercepto	7.82	6.55 - 9.40
Álcool (Menos de 1x/mês)	2.00	1.27 - 3.38
Álcool (1x/mês ou mais)	2.05	1.45 - 2.96
Sexo (Mulher)	1.69	1.37 - 2.088
Hipertensão (Não)	2.27	1.76 - 2.98
Idade	0.87	0.78 - 0.97

Na Tabela 3, com relação ao consumo de álcool, pessoas que bebem menos de uma vez ao mês apresentam o dobro de chances de AVC (OR ≈ 2,0) em relação ao grupo de referência. Aquelas que bebem uma ou mais vezes ao mês também apresentam um risco maior (OR ≈ 2,05), sendo quase 2 vezes maior em relação ao grupo que não faz uso de álcool.

Em relação ao sexo, as mulheres apresentam aproximadamente 1,7 vezes a chance de ter um AVC comparadas ao grupo de referência (os homens), sendo essa associação estatisticamente significativa.

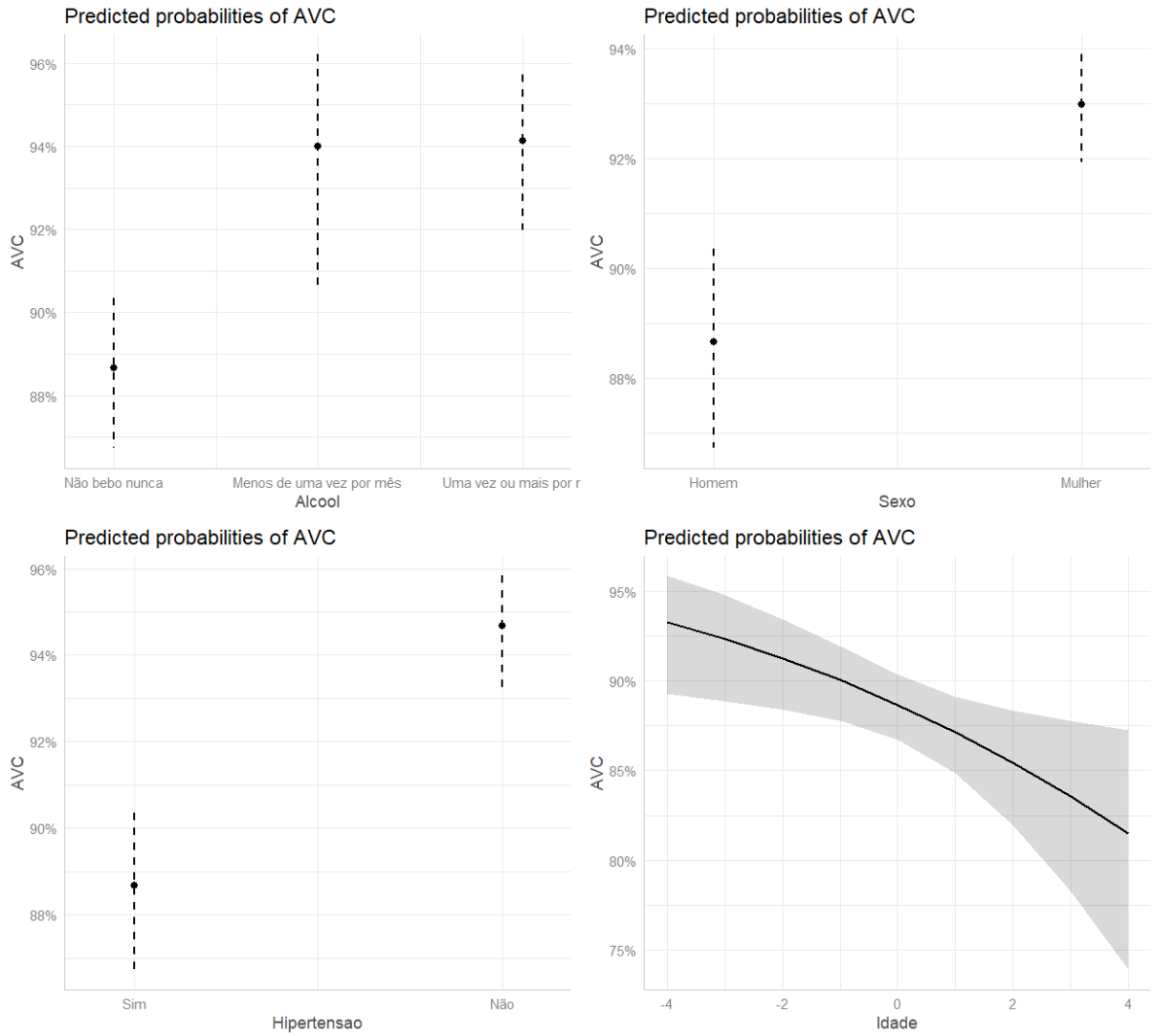
Com relação à hipertensão, o resultado revela que pessoas que não são hipertensas apresentam um risco 2,27 vezes maior de AVC.

Por fim, a idade apresenta um OR de 0,87, mostrando que o aumento de um ponto na idade relaciona-se a uma redução de cerca de 13% nas chances de AVC.

Mais uma vez, os resultados obtidos pela Hipertensão e Idade, podem ser fatores associados a má amostragem feita, ou a quantidade de não's que temos na variável AVC.

Analisando os gráficos dos efeitos:

Figura 4 – Gráfico dos efeitos (Modelo StepWise)



Na Figura 4, no primeiro gráfico, observa-se que o consumo de álcool apresenta um gradiente de risco, sendo maior para pessoas que bebem pelo menos uma vez ao mês em relação às que não bebem ou bebem raramente.

Na parte superior direita, o gênero também revela uma diferença: as mulheres apresentam maior probabilidade de AVC do que os homens.

Na parte inferior esquerda, verifica-se que a ausência de hipertensão se relaciona a um risco de AVC menor, enquanto a presença de hipertensão eleva consideravelmente essa probabilidade.

Por fim, na parte inferior direita, observa-se uma relação inversa da idade com o AVC — quanto maior a idade, menor a taxa de AVC — mostrando uma tendência decrescente. Observando em geral, tudo bate com as interpretações feitas com a OR

e as estatísticas do p-valor, mesmo algumas variáveis (como Hipertensão e Idade) não batendo com a nossa realidade.

Observando a Figura 5 e os resultados da Tabela 4, a matriz de confusão revela que o modelo apresenta uma taxa de acertos relativamente baixa, com acurácia de 47,81%.

Apenas 32,01% dos casos positivos ("Sim") foram classificados corretamente (sensibilidade), enquanto 48,83% dos casos negativos ("Não") também foram classificados corretamente (especificidade).

O valor preditivo positivo (PPV) é baixo (cerca de 4,097%), mostrando que, quando o modelo prediz "Sim" (isto é, AVC), ele acerta pouquíssimas vezes. Por outro lado, o valor preditivo negativo (VPN) é alto (91,34%) — ou seja, ele acerta na maior parte das vezes que prediz "Não".

A prevalência de casos positivos na base é baixa (6,39%) — o que torna o modelo particularmente propenso ao desbalanceamento de classes.
A acurácia balanceada (40,45%) revela que o modelo não consegue distinguir bem entre as duas classes.

O Kappa (-0,0458), sendo negativo, evidencia que o modelo faz pior do que o acaso.

Por fim, o p-valor do Teste de McNemar (< 2e-16) revela que existem diferenças significativas nas proporções de acertos e erros, reforçando que o modelo apresenta um desempenho particularmente pobre na classificação da doença (AVC).

Figura 5 – Matriz de Confusão (Modelo StepWise)

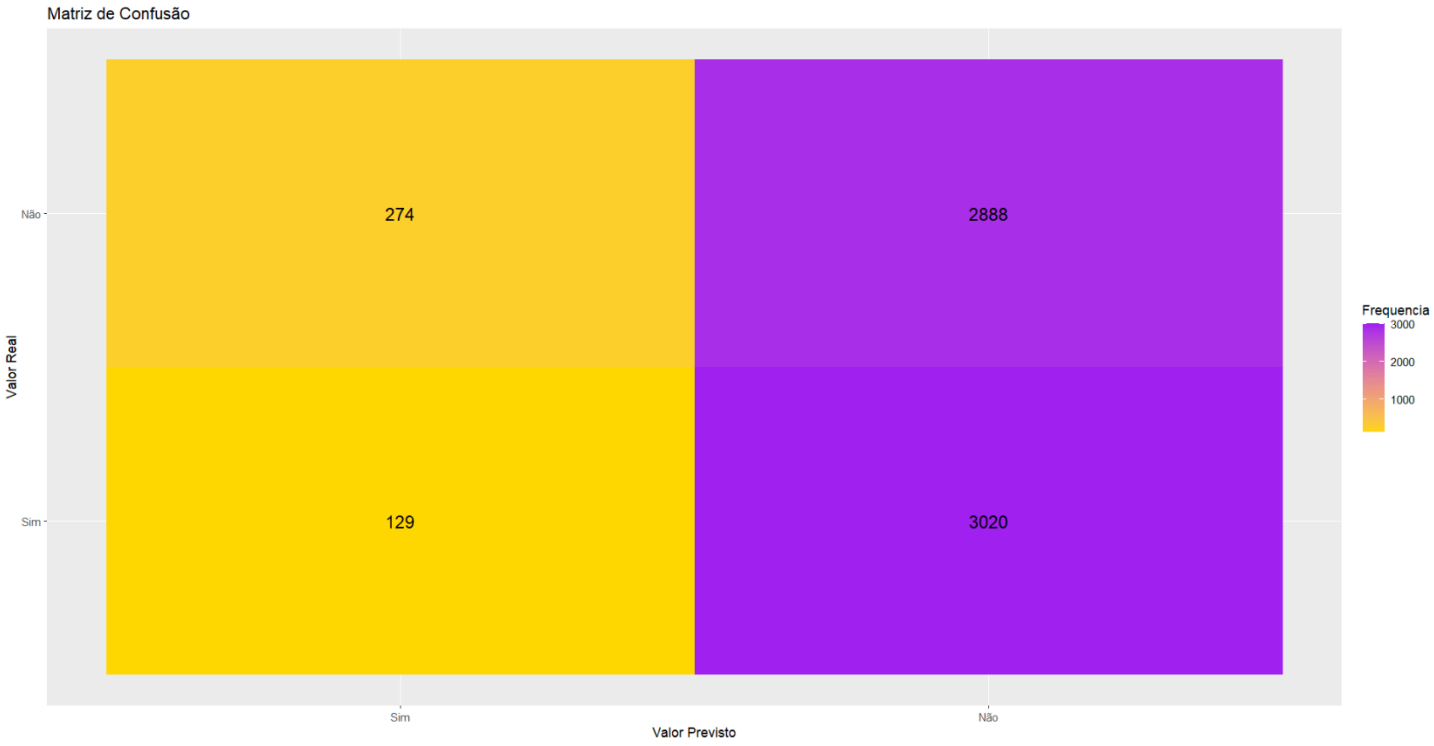


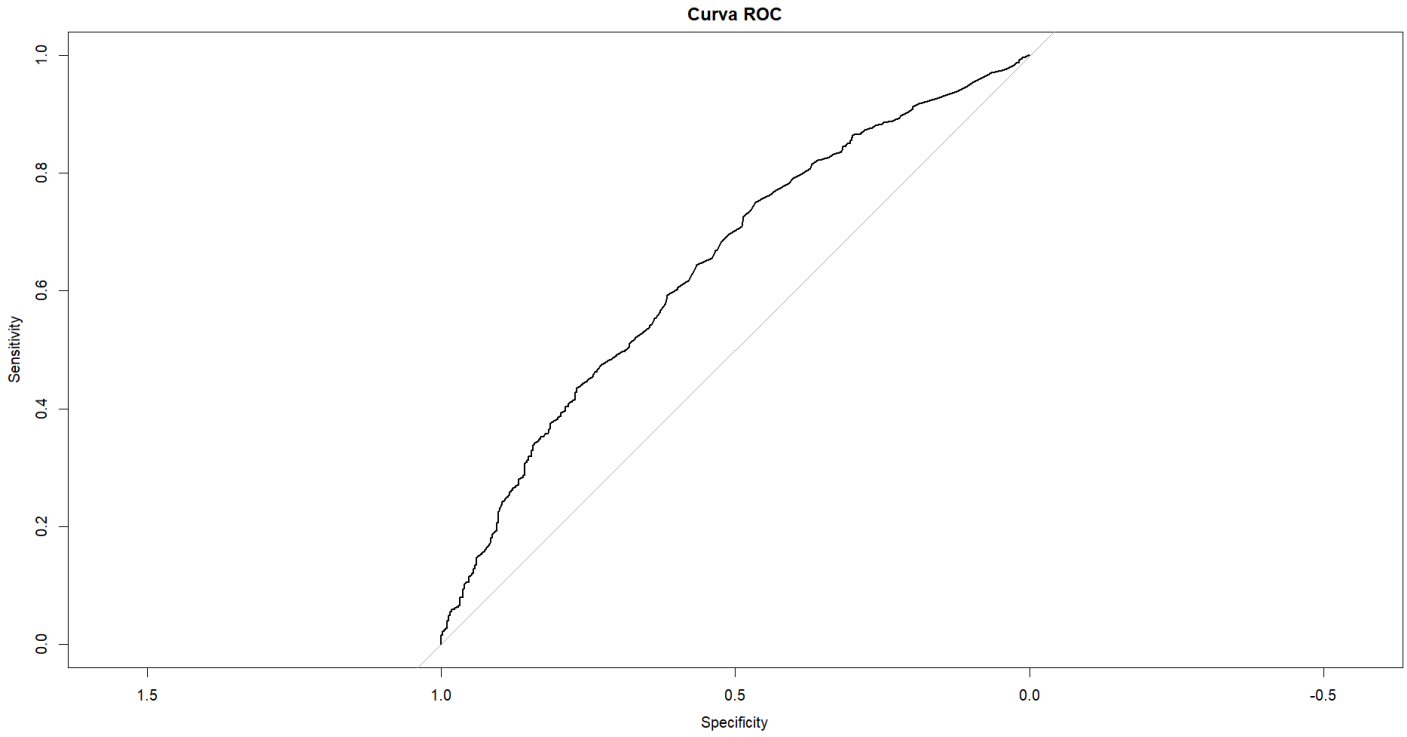
Tabela 4 – Estatísitcas do Modelo StepWise

Estatística	Valor
Acurácia	47,81%
Sensibilidade (Recall)	32,01%
Especificidade	48,88%
Valor Preditivo Positivo (PPV)	4,097%
Valor Preditivo Negativo (VPN)	91,34%
Prevalência	6,39%
Detecção	2,04%

Estatística	Valor
Kappa	-0,0458
Balanced Accuracy	40,45%
p-valor (McNemar)	< 2e-16

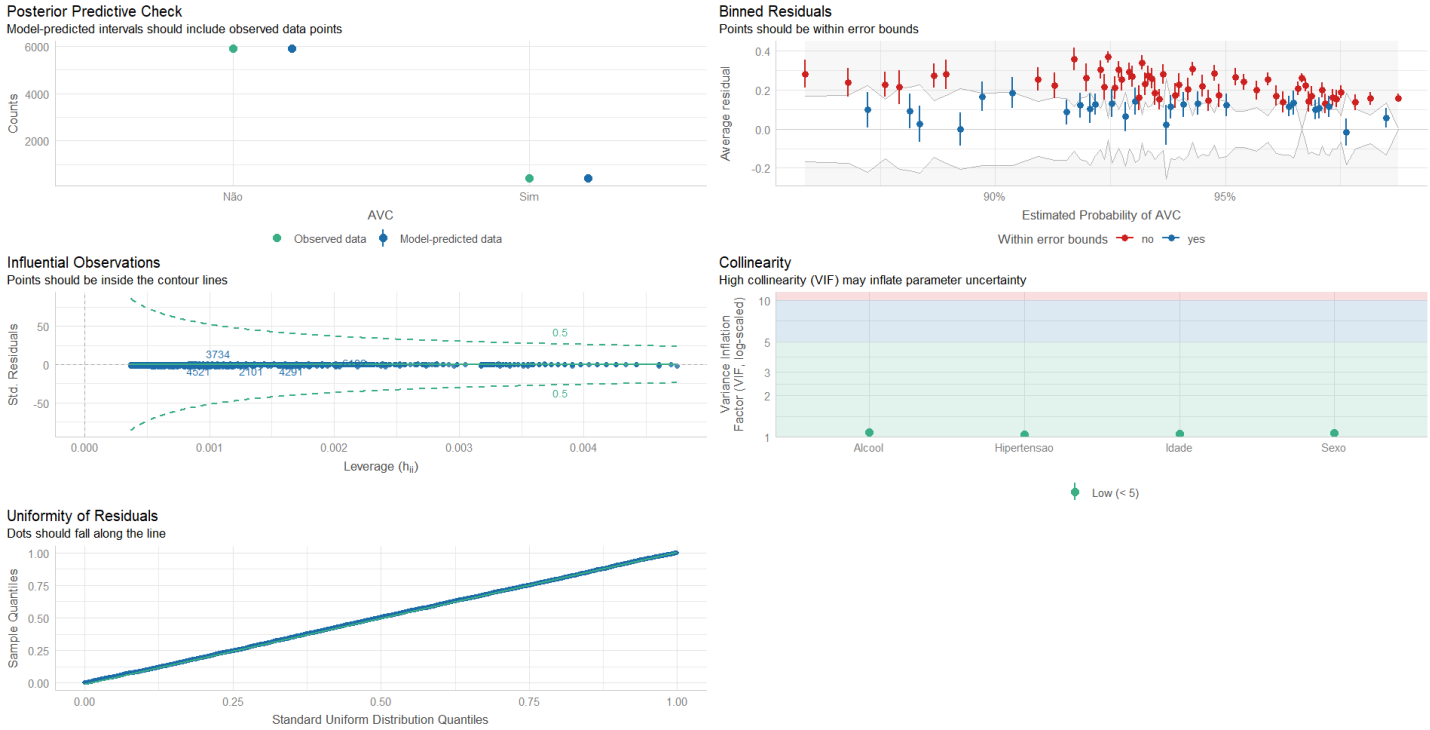
Observando a Figura 6, e o valor do AUC de, o modelo consegue distinguir pacientes com e sem a doença (neste caso, AVC) cerca de 64% das vezes, ou seja, AUC de aproximadamente 0,64 revela um poder discriminatório baixo, sendo melhor que o acaso, mas ainda longe de um modelo robusto para apoiar a tomada de decisão clínica.

Figura 6 – Curva ROC (Modelo StepWise)



Observando a análise residual:

Figura 7 – Análise Residual (Modelo StepWise)



Na Figura 7, no primeiro gráfico é mostrado pontos verdes, que representam os dados observados, e os pontos azuis, que representam os valores preditos pelo modelo. A proximidade entre eles revela que o modelo consegue reproduzir relativamente bem a distribuição de casos de AVC presentes na base de dados.

Continuando para o segundo gráfico, o modelo verifica o comportamento médio dos resíduos ao longo das probabilidades previstas. Os pontos vermelhos representam o grupo que não teve AVC, enquanto os pontos azuis representam o grupo que teve AVC. De forma geral, os resíduos estão distribuídos em torno de zero, embora seja perceptível uma concentração de resíduos positivos para determinados valores de probabilidade, principalmente para o grupo que não teve AVC. Isso pode rebelar uma leve falta de ajuste, sendo que o modelo superestima o risco de algumas pessoas que na realidade não sofreram AVC. Observando o teste de Hosmer – Lemeshow, é perceptível isso, pois nosso p-valor foi muito abaixo de 0.05 (2.2e-16).

Seguindo, o terceiro gráfico é o de alavancagem, cada ponto corresponde a uma observação. A maior parte das observações apresenta alavancagem baixa e resíduos padronizados próximos de zero.

Para o gráfico de multicolineariedade, todas estão muito abaixo de 5, o que significa que não existem problemas neste quesito. Isso quer dizer que as variáveis são relativamente independentes entre si e que o modelo consegue estimar seus coeficientes de forma estável.

Para o Q-Q Plot, os pontos seguem exatamente a linha de referência, mostrando que a distribuição de resíduos é aproximadamente normal. Isso fortalece a ideia de que o modelo se ajusta relativamente bem e que ele não apresenta problemas de especificação relacionados à forma da distribuição de erro.

4 CONCLUSÃO

É notório que, o modelo não foi o melhor possível, com conclusões de algumas variáveis questionáveis, pois não bate com a nossa realidade (como vimos com a Hipertensão e Idade). Talvez com a exclusão das duas variáveis que não bateram com a realidade, o modelo fique melhor, acertando mais.

É perceptível também, uma falta das outras variáveis que, na nossa realidade, faria a diferença (Como, por exemplo, a variável “Fumar”). Acredito que o que aconteceu foi uma amostragem não tão boa assim, ou até mesmo o fato da nossa variável resposta “AVC” ter mais resultados negativos do que positivos.

Porém, creio que acabei entendendo melhor como funciona uma regressão logística, conseguindo aproveitar bastante o trabalho para tentar avaliar melhor o meu modelo. A única dúvida que fiquei, foi sobre os pontos influentes, e o que fazer quando o modelo não tem um bom ajuste (talvez tenha faltado um pouco de aprofundamento meu neste quesito).

Acredito também, que com modelos mais robustos, como Floresta Aleatória ou Árvore de Decisão, consiga obter resposta melhores.

REFERÊNCIAS

BRASIL. **Ministério da Saúde**. Acidente Vascular Cerebral (AVC). Disponível em: https://bvsms.saude.gov.br/avc-acidente-vascular-cerebral/. Acesso em: 15 jun. 2025.

BRASIL. **Fiocruz**. Pesquisa Nacional de Saúde (PNS). Disponível em: https://www.pns.icict.fiocruz.br/. Acesso em: 15 jun. 2025.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009. Disponível em: https://hastie.su.domains/ElemStatLearn/. Acesso em: 5 jun. 2025.

HOSMER, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. *Applied Logistic Regression*. 3. ed. Hoboken: John Wiley & Sons, 2013. Disponível em: https://books.google.com/books/about/Applied_Logistic_Regression.html?id=Po0RLQ7USIMC. Acesso em: 5 jun. 2025.

SCIELO. Disponível em: https://www.scielo.br/j/ape/a/mHYgZZ5BGngmHnkTKfhzQkS/. Acesso em: 15 jun. 2025.