Relatório de Dengue em Bangladesh

Mariana Freitas e Aline Pires

Introdução

A dengue é uma doença infecciosa viral transmitida por mosquitos do gênero Aedes, apresentando-se como um dos principais desafios de saúde pública em regiões tropicais e subtropicais, como Bangladesh. A identificação rápida e precisa é fundamental para o controle de surtos e redução da mortalidade. Neste contexto, a análise de dados categorizados é uma ferramenta essencial para auxiliar a tomada de decisões dos orgãos públicos para controlar esse desafio, Com ela, é possível avaliar o desempenho de diferentes testes de diagnósticos e investigar os fatores associados à ocorrência da doença. O presente estudo utiliza técnicas estatísticas descritivas e inferenciais aplicadas a um banco de dados de indivíduos testados para dengue em Bangladesh, buscando compreender as relações entre características individuais, fatores ambientais e resultados dos testes de diagnósticos, bem como avaliar a precisão destes métodos na detecção da infecção.

Metodologia

Dados categorizados são conjuntos de dados cujas variáveis são categóricas, ou seja, representam característica, qualidade ou atributo. Essas variáveis categóricas podem ser nominais, quando as classes da variável não tem ordem natural (gênero, tipo sanguíneo..) ou ordinais, quando as classes apresentam ordem natural (nível de escolaridade, grau de dor,...). Uma importante técnica na análise de dados categorizados são as tabelas de contigência, ideais para organizar a frequência das interseções entre as variáveis categóricas, permitindo a observação de associação entre variáveis, cálculo de medidas de desempenho de testes diagnósticos e realização de testes de associação, simetria ou homogeneidade entre variáveis.

No contexto de testes diagnósticos - testes que identificam se um indivíduo apresenta ou não determinada doença ou condição - é possível calcular algumas medidas para aavaliar a performance dos testes, já que estão sujeitos a erros e, consequentemente, seus resultados apresentam grau de incerteza. Duas medidas muito utilizadas são a sensibilidade e especificidade. A sensibilidade é calculada como a razão de verdadeiros positivos (doentes cujo teste foi positivo) em relação à soma de falsos negativos (doentes cujo teste foi negativo) e verdadeiros positivos. Já a especificidade corresponde à razão de verdadeiros negativos (não doentes cujo teste foi negativo) em relação à soma de verdadeiros negativos e falsos positivos (não doentes cujo teste foi positivo). Assim, a sensibilidade pode ser interpretada como a probabilidade do teste ser positivo dados que o indivíduo está doente e a especificidade como a probabilidade do teste ser negativo dado que o indivíduo não está doente. Um bom teste apresenta sensibilidade e especificidade altas, sendo que podem variar de 0 a 1.

O tipo de estudo e delineamento amostral são importantes para a interpretação de resultados. Nesse caso, será justificado que se trata de um estudo transversal, mas não há informações suficientes para definir o delineamento amostral. Para esse tipo de estudo, cabe verificar se há diferenças nas proporções de doentes em diferentes classes de variáveis binárias. Para isso, é calculada a estimação pontual da diferença entre as porporções e em seguida é feito o intervalo de confiança. Se o intervalo de confiança incluir zero, não se pode afirmar que há uma diferença entre as proporções. Caso contrário, há diferença nas porporções a uma determinada nível de confiança - nesse trabalho foi utilizado 95%. Outra importante alternativa para avaliar associação em tabelas de contigência $2x2$ são as razão de chances, medida apropriada para o tipo de estudo transversal. A razão de chances se trata da razão entre a chance de uma classe apresentar a doença

e a chance da outra classe apresentar a doença. Foi feita uma estimativa da razão de chance e em seguida foi complementada por inferência estatística, com cálculo do logaritmo da razão de chances e respectivos intervalos de confiança obtidos por aproximação normal. Se o intervalo de confiança incluir 1, não se pode afirmar que há diferença entre as chances, caso contrário há diferença a um determinado nível de confiança.

Além de inferência estatística para as proporções, também foram feitos testes para avaliar independência, associação, simetria e homogeneidade. Primeiro forma feitos testes específicoa para variáveis ordinais. Para verificar a intensidade e direção da associação entre variáveis ordinais, foram aplicadas os testes Gama de Goodman e Kruskal, Tau de Kendall e Tau-b de Kendall - que consideram a ordenação das classes. A Gama de Goodman e Kruskal se baseia em pares concordantes e discordantes em tabelas de contingência, variando de -1 a 1, indicando associação perfeita negativa ou positiva, respectivamente, e desconsidera os pares empatados. As medidas Tau de Kendall e Tau-b de Kendall corrigem a Gama ao considerar empates nas margens. O Tau-b é útil para tabelas não quadradas, visto que ajusta a estatística levando em conta o número de empates nas linhas e colunas. Para testar tendência linear entre variáveis ordinais, podem ser usados os testes de Cochran-Armitage - em tabelas para verificar se a proporção de sucesso aumenta ou diminui linearmente com as categorias das variáveis ordinais - e o teste de Mantel - em tabelas sxr, avaliando a presença de uma tendência linear global entre variáveis ordinais. Aqui foi aplicado apenas o primeiro teste, já que há apenas duas variáveis ordinais.

Para analisar a associação entre duas variáveis categóricas controlando por uma terceira variável, foram construídas tabelas de contingência parciais, e calculadas razões de chances condicionais em cada classe, permitindo avaliar se a associação é condicionalmente homogênea entre as classes. A homogeneidade das razões de chances foi testada com o teste de Breslow-Day, que avalia se as ORs são estatisticamente iguais entre as classes. Quando a homogeneidade foi aceita, foi utilizado o teste de Mantel-Haenszel, que fornece uma razão de chances combinada ajustada, além de um teste de associação global. Essa técinica é importante para lidar com casos em que ocorre o Paradoxo de Simpson - quando a associação entre duas variáveis muda após o controle por uma terceira.

Para tabelas de contingência com dimensões rxs, foram aplicados alguns testes para avaliar associação e simetria. O teste de homogeneidade foi aplicado em situações nas quais uma das variáveis representa grupos e a outra representa categorias de resposta, buscando verificar se a distribuição de respostas é homogênea entre os grupos. O teste de simetria foi utilizado em tabelas quadradas para avaliar se a frequência de observações na célula (i, j) é igual a (j, i), útil para dados pareados ou classificações duplas. O teste de homogeneidade marginal, também em tabelas quadradas, verificou se as distribuições marginais das linhas e colunas são idênticas, independentemente da simetria.

Por fim, foi abordada a etapa de modelagem em tabelas de contigência. Foi ajustado um modelo de regressão logística, permitindo estimar a chance de doença como função das variáveis explicativas categóricas. Os coeficientes do modelo foram interpretados em termos do log das razões de chance. Foi feita uma seleção do modelo de regressão logística mais adequado a partir do critério de Informação Akaike (AIC) e análise do ajuste do modelo.

Todas as análises foram feitas utilizando o software R, com pacotes específicos mencionados ao longo do relatório.

Resultados

Análise exploratória

O banco de dados utilizado consiste predominantemente em variáveis categóricas, entre as quais:

- Variáveis de diagnóstico: Resultados dos testes `NS1`, `IgG` e `IgM`, que mostram se detectou ou não a dengue, além da variável resposta `Outcome` indicando a presença ou ausência dengue. Essas variáveis tem como resposta *Positive* ou *Negative*.

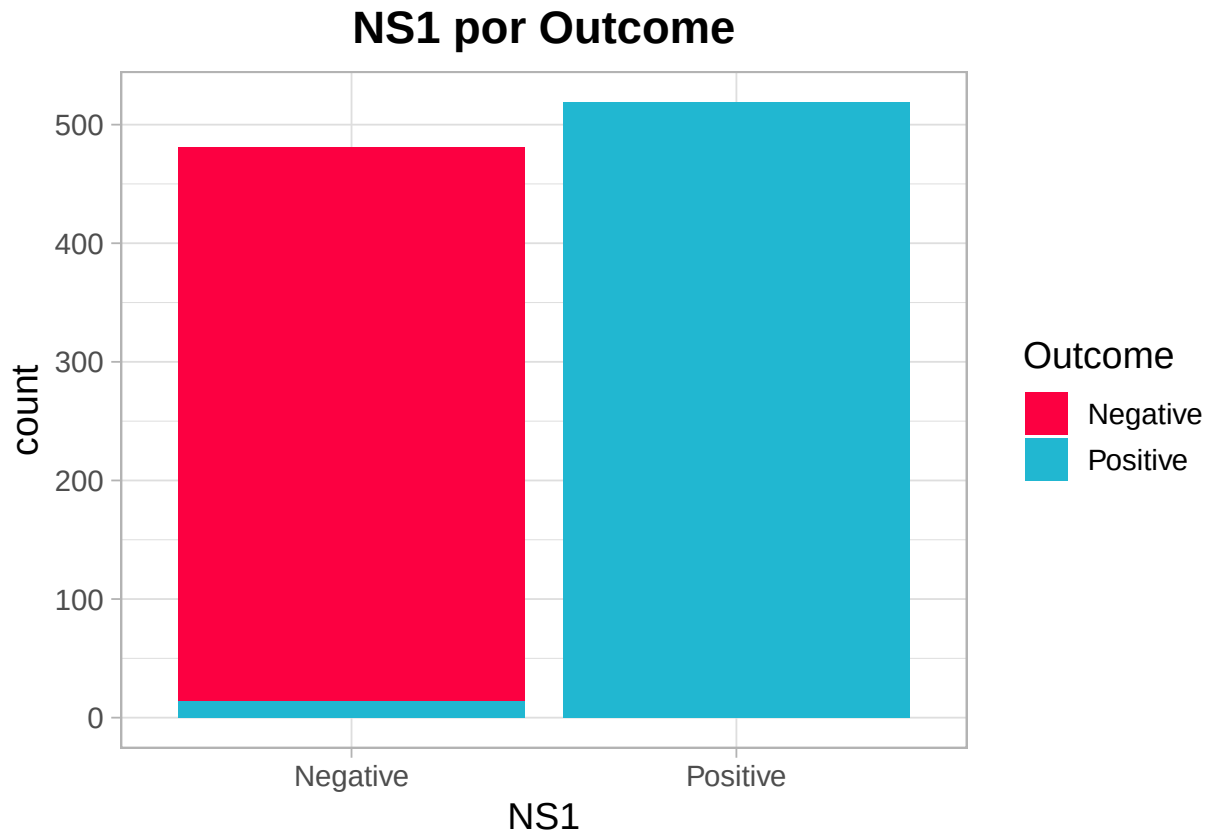
- Variáveis sociodemográficas: Incluem as variáveis `Gender`, gênero do indivíduo, classificado como male ou female e `ageless16years`, que indica se o indivíduo tem menos de 16 anos - *No* ou *Yes*.
- Variáveis ambientais: A `AreaType` representa o tipo de área de residência do indivíduo, podendo ser *Developed* ou *Undeveloped*; já a `Area` contém a área específica de moradia (diversos bairros).
- Variáveis habitacionais: Contém `HouseType` que represente o tipo de moradia, as classes são *Building*, *Tin-Shed* ou *Others*.

É importante destacar que as variáveis ordinais desse banco de dados são `AgeLess16Years` e `AreaType`, pois possuem uma ordem natural de classificação. Além disso, a variável `AgeLess16Years` foi criada a partir da variável `Age` nos dados originais, com base no valor de `Age` que maximiza a diferença do desfecho entre os grupos.

Todas essas coraviáveis listadas acima possibilitam a construção de tabelas de contingência, avaliação de medidas de sensibilidade e especificidade dos testes diagnósticos, além de análises mais avançadas, como regressão logística, para estimar fatores associados à infecção para dengue na população analisada.

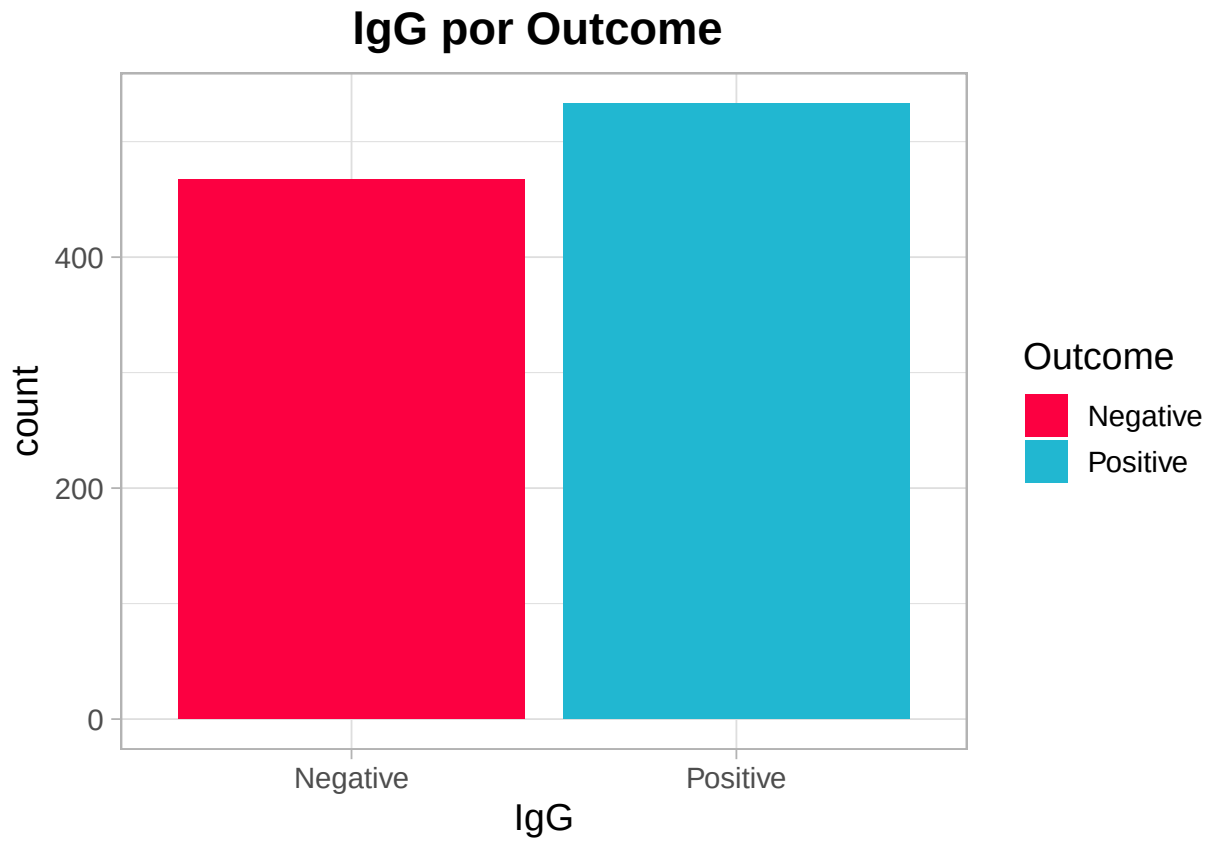
Antes de avançarmos para as avaliações estatísticas e testes de associação, é importante compreender o perfil da amostra e a distribuição das principais variáveis analisadas neste estudo.

Na etapa inicial, buscou-se compreender a distribuição das principais variáveis categóricas em relação ao desfecho (Outcome). Foram construídos gráficos de barras comparando os resultados dos testes diagnósticos (NS1, IgG e IgM) e a variável sociodemográfica Age_less_16_years estratificada por gênero.

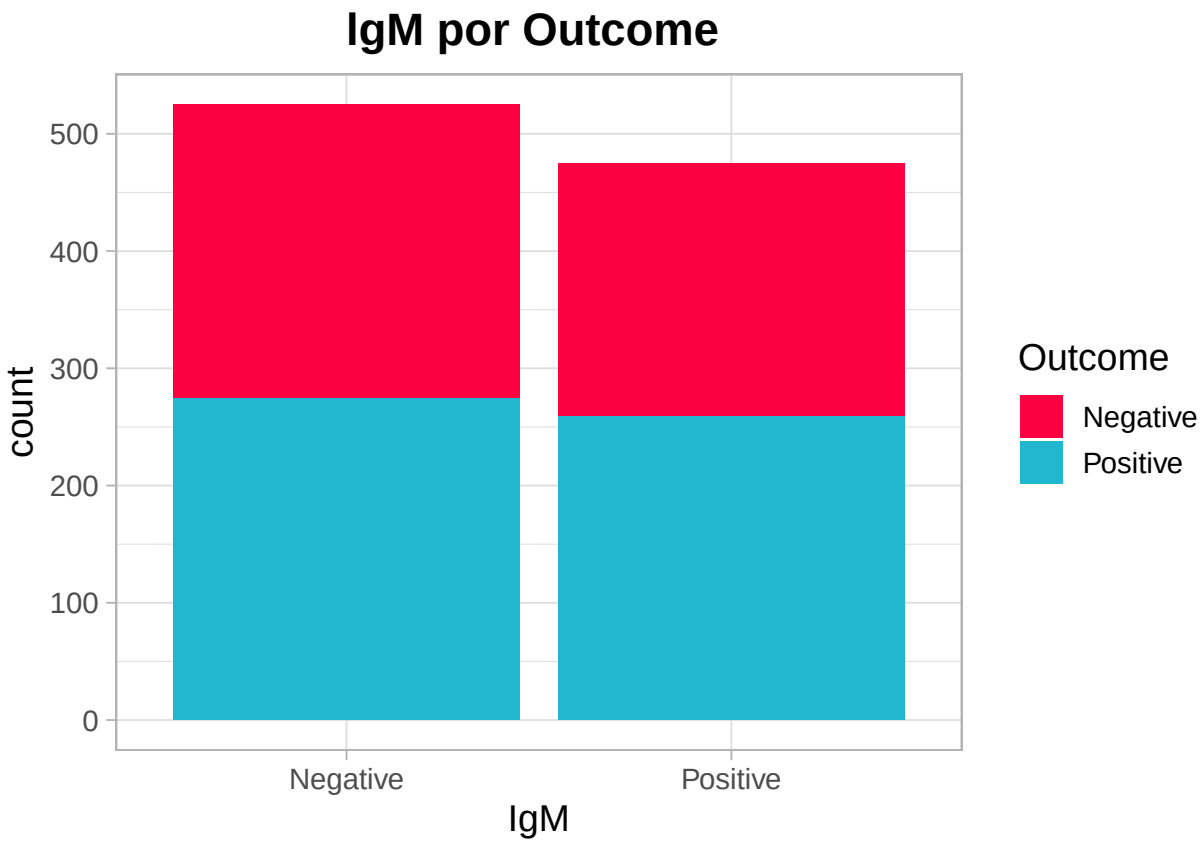


No teste NS1, observa-se uma clara separação entre os grupos: a grande maioria dos indivíduos que tiveram resultado positivo no teste também apresentaram Outcome positivo (dengue), enquanto praticamente não houve falsos positivos. Já no grupo com resultado negativo no NS1, predominaram os indivíduos sem a

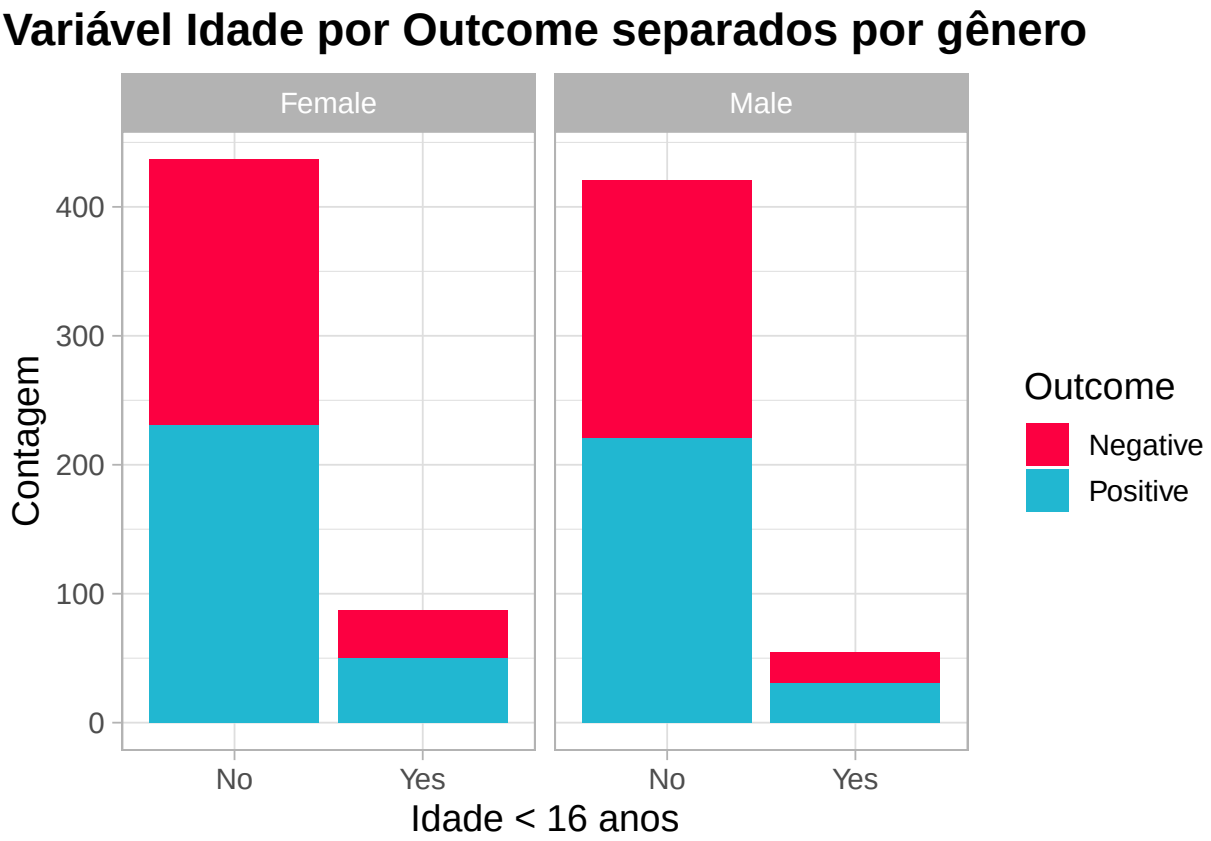
doença, com poucos casos classificados como falso negativo. Esse padrão indica alta concordância do NS1 com o desfecho clínico.



O teste IgG apresentou comportamento ainda mais acentuado: todos os indivíduos classificados como positivos no IgG também tiveram Outcome positivo, enquanto aqueles com resultado negativo concentraram os casos sem dengue. Esse padrão sugere desempenho quase perfeito desse teste em relação ao desfecho.



Em contraste, o teste IgM apresentou maior dispersão. Tanto no grupo IgM negativo quanto no positivo foram observadas quantidades consideráveis de indivíduos com Outcome positivo e negativo. Isso indica que o IgM tem menor poder discriminatório para predizer a presença de dengue, corroborando a expectativa de baixa sensibilidade e especificidade observada posteriormente.



Do ponto de vista sociodemográfico, avaliou-se a variável idade (menor que 16 anos) separada por gênero. A maior parte da amostra é composta por indivíduos com 16 anos ou mais, tanto no grupo feminino quanto no masculino. Entre os menores de 16 anos, a quantidade de positivos e negativos é relativamente equilibrada, sem indicar uma diferença marcante por faixa etária. Além disso, não se observa grande discrepância entre os gêneros: tanto homens quanto mulheres apresentam proporções semelhantes de casos positivos e negativos.

Esses achados exploratórios permitem uma primeira visão sobre o perfil da amostra e a efetividade dos testes diagnósticos, destacando a maior acurácia do IgG e NS1 em comparação ao IgM. Ademais, sugerem que fatores sociodemográficos simples, como idade e gênero, não apresentam diferenças substanciais na proporção de casos de dengue.

Avaliação de Testes Diagnósticos

O conjunto de dados apresenta três testes para detectar a dengue: denotados por NS1, IgG e IgM. Considerando apenas as informações sobre os resultados desses três testes e a presença ou não da doença nos indivíduos testados, são obtidas as tabelas 1, 2 e 3 para NS1, IgG e IgM, respectivamente.

Tabela 1: Distribuição dos desfechos segundo teste NS1.

NS1/Outcome	Negative	Positive
Negative	467	14
Positive	0	519

Tabela 2: Distribuição dos desfechos segundo teste IgG.

IgG/Outcome	Negative	Positive
Negative	467	0
Positive	0	533

Tabela 3: Distribuição dos desfechos segundo teste IgM.

IgM/Outcome	Negative	Positive
Negative	251	274
Positive	216	259

Apenas observando as tabelas parciais é possível notar que o teste mais preciso parece ser o IgG, enquanto o de menor eficácia seria o IgM. No entanto, essa intuição pode ser formalizada utilizando as medidas de sensibilidade e especificidade apresentadas na tabela 4.

Tabela 4: Medidas de avaliação dos testes diagnósticos.

Teste Diagnóstico	Sensibilidade	Especificidade
NS1	0,974	1
IgG	1	1
IgM	0.486	0.538

Concluões sobre os testes:

- O teste NS1 apresentou boa performance no geral, classificando corretamente todos os que não tinham dengue e também com alta sensibilidade - indicando que classificou grande parte dos indivíduos com a diença corretamente.
- O teste IgG classificou corretamente todos os indivíduos.
- O teste IgM teve performance bem ruim, com ambas as medidas baixas.

Tabelas de Contigência

Para decidir que ferramentas serão usadas na análise de uma tabela de contigência, é importante antes de qualquer coisa entender qual é o tipo de estudo. A partir das informações fornecidas pela fonte dos dados, é possível inferir que se trata de um estudo transversal, pois os dados são colhidos em um ponto específico no tempo após a ocorrência ou não de dengue; não houve nenhum tipo de intervenção ou acompanhamento dos indivíduos. Esse tipo de estudo permite verificar se há diferenças nas proporções de doentes nas classes de variáveis binárias, a partir da criação de tabelas $2x2$ considerando a variável explicativa binária de interesse e a variável resposta. Nas tabelas 5, 6 e 7 são mostradas as tabelas $2x2$ para as variáveis binárias `Gender`, `AreaType` e `Age_less_16_years`.

Tabela 5: Distribuição dos desfechos segundo gênero.

Gender/Outcome	Negative	Positive
Female	243	281
Male	224	252

Tabela 6: Distribuição dos desfechos segundo tipo de área.

AreaType/Outcome	Negative	Positive
Developed	244	257
Undeveloped	223	276

Tabela 7: Distribuição dos desfechos segundo classificação da idade.

Age_less_16_years/Outcome	Negative	Positive
No	406	452
Yes	61	81

Inicialmente não é possível concluir muito apenas observando as tabelas, então é útil calcular as estimações pontuais e intervalos de confiança para as diferenças de proporção de doentes entre as classes de cada variável. Os resultados estão apresentados na tabela 8.

Tabela 8: Resultados para diferenças de proporções.

Variável	Diferença na proporção de doentes	Estimação pontual	IC (95%)
Gender	Female - Male	0.006848	[-0.055, 0.069]
AreaType	Undeveloped - Developed	0.040132	[-0.021, 0.102]
Age_less_16_years	No - Yes	-0.04361	[-0.132, 0.044]

Todos os intervalos de confiança, a um nível de 95% contém o valor 0. Dessa forma, assumimos que não há diferença nas proporções de doentes entre as classes das variáveis `Gender`, `AreaType` e `Age_less_16_years`. As estimações pontuais também foram bem próximas de zero, o que reafirma que as diferenças são fruto da aleatoriedade, não das classes.

Uma medida importante para verificar se há associação entre a ocorrência de doença e as variáveis binárias mencionadas é a razão de chance. suas estimativas pontuais e intervalos de confiança calculados a partir da exponencial dos logs das razões de chance constam na tabela 9.

Tabela 9: Resultados para razões de chances.

Variável	Razão das chances de apresentar dengue	Estimação pontual	IC (95%)
Gender	Male/Female	0.973	[0.759 , 1.248]
AreaType	Developed/Undeveloped	0.851	[0.664 , 1.091]
Age_less_16_years	Yes/No	1.19	[0.834 , 1.707]

Visto que todos os intervalos de confiança para a razão de chance contém 1, pode-se afirmar que não há diferença causada pelas classes entre as chances de ter dengue. As estimações pontuais também estão bem próximas de 1, reforçando essa ideia.

Inferência para Tabelas de Contigência

Para complementando os resultados descritivos apresentados anteriormente, realizamos testes estatísticos para avaliar a associação e possíveis tendências entre as variáveis ordinais "idade < 16 anos" (Ageless16years) e "tipo de área" (AreaType) em relação à ocorrência de dengue. São empregados testes que consideram a ordenação natural dessas variáveis.

Primeiro é aplicado o teste de Gama de Goodman e Kruskal, para medir a força e a direção da associação entre duas variáveis ordinais, e o teste Tau-b que também avalia associação entre variáveis ordinais mas considerando os possíveis empates.

Tabela 10: Medidas de associação de variáveis ordinais em relação à ocorrência de dengue

Medidas/Variáveis	Age_less_16_years	AreaType
Gama de Goodman e Kruskal	0.0878	-0.0805
Tau-b	0.0305	-0.0402

Ambos os coeficientes indicam associação positiva muito fraca entre ter menos de 16 anos e ser positivo para dengue, indicando ausência de associação relevante entre essa característica e a ocorrência da doença. Além disso, os coeficientes associados a variável `AreaType` também apontam associação negativa, mas, como o valor absoluto é muito baixo, não se observa relação relevante na prática.

Por fim, para verificar se existe tendência linear, ou seja, se a proporção de dengue aumenta ou diminui de acordo com a idade (<16 anos) ou tipo de área utiliza-se o teste de Cochran-Armitage.

P-valor/Variáveis	Age_less_16_years	AreaType
p-valor	0.3346	0.2034

O teste não encontrou tendência linear significativa (p>0.05) entre ser menor de 16 anos e chance de ter dengue e nem entre o tipo de área que o indivíduo mora.

Associação em Tabelas de Contigência

Para verificar se as distribuições das frequências observadas indicam relações estatísticas significativas entre as variáveis, considerando também possíveis efeitos de variáveis de controle, é necessário aplicar diferentes testes para avaliar a associação, homogeneidade, simetria e possíveis paradoxos nas tabelas de contingência construídas.

Primeiro aplica-se o Teste de Breslow Day a fim de verificar se a associação entre o desfecho de dengue (`Outcome`) e idade (`Age_less_16_years`) é homogênea ao longo das categorias do tipo de área (`AreaType`). Ele indicou homogeneidade das razões de chances entre os estratos (p-valor = 0,43), assim, foi possível prosseguir com o teste de Mantel-Haenszel para estimar a associação ajustada entre ser menor de 16 anos e o desfecho de dengue, controlando pelo tipo de área de residência.

Porém o teste de Mantel-Haenszel não evidenciou associação significativa (p-valor = 0,40) entre essas variáveis, indicando que, após ajuste pelo estrato, a chance comum de apresentar dengue para indivíduos com menos de 16 anos é aproximadamente 1,19 (IC 95%: 0,83 a 1,70). Como este intervalo inclui 1 e o p-valor é elevado, não há evidências de associação estatística.

Além disso, a ausência de variação significativa nas razões de chances entre os estratos, evidenciada pelo teste de Breslow-Day, indica que não ocorre o paradoxo de Simpson neste caso — ou seja, o ajuste pelo tipo de área não alterou a direção ou a magnitude da associação entre idade e dengue.

Assim, conclui-se que, na amostra avaliada, não há associação relevante entre ser menor de 16 anos e a ocorrência de dengue, independentemente do tipo de área onde o indivíduo reside.

Ao avaliar a associação entre gênero e desfecho de dengue estratificada pelas 36 áreas específicas do estudo, o teste de Breslow-Day indicou heterogeneidade significativa nas razões de chances entre os estratos (p = 0,02). Essa rejeição da homogeneidade sugere que a intensidade e/ou direção da associação entre gênero e ocorrência de dengue variam conforme a área de residência. Diante disso, o teste de Mantel-Haenszel foi utilizado para estimar a razão de chances comum ajustada pelo estrato, mas não evidenciou associação estatisticamente

significativa entre gênero e dengue na amostra combinada (0,75), com razão comum estimada em 0,95 (IC 95%: 0,74 a 1,23).

A presença de heterogeneidade nas razões de chances, aliada à ausência de associação significativa no teste combinado, sugere a ocorrência do paradoxo de Simpson e classifica-se a área de residência como um fator de confusão para essa análise.

Além dos testes de Breslow-Day e Mantel-Haenszel já apresentados para avaliar a associação ajustada entre as variáveis, foram aplicados testes adicionais para aprofundar a análise da relação entre as variáveis categóricas no estudo.

Para investigar a associaçõe simples entre tipo de área e desfecho, aplicou-se o teste qui-quadrado de independência e o teste de razão de verossimilhança (G-test) que deu um p-valor = 0,23 pelo qui-quadrado, indicando ausência de associação direta nessas análises simples.

No entanto, a análise estratificada por área revelou uma heterogeneidade significativa na associação entre gênero e desfecho de dengue, como vimos anteriomente. Então, ao fazer os Testes qui-quadrado aplicados em cada uma das 36 áreas específicas mostraram que, em algumas áreas como Adabor (p = 0,017) e Sutrapur (p = 0,02), há associação estatisticamente significativa, enquanto nas demais áreas a associação não foi significativa.

Para os testes diagnósticos IgG, IgM e NS1, foram aplicados testes de simetria pelo teste de McNemar, cujo resultado indicou ausência de simetria significativa entre os testes, com p-valores menores que 0,05 em todas as comparações: IgG versus NS1 (p = 0,0005), IgG versus IgM (p = 0,01) e IgM versus NS1 (p = 0,049). Esses resultados indicam discordâncias sistemáticas entre os testes diagnósticos avaliados, sugerindo que os diferentes métodos podem fornecer informações variadas sobre a presença de dengue.

O teste de homogeneidade marginal foi aplicado para avaliar se as distribuições do desfecho dengue são homogêneas entre diferentes grupos das variáveis área específica e gênero. Ao analisar a associação entre a área de residência e o desfecho, o teste qui-quadrado indicou rejeição da hipótese de homogeneidade (p-valor = 0,0015), o que significa que as distribuições do desfecho dengue variam significativamente entre as áreas avaliadas. Por outro lado, ao testar a homogeneidade entre gênero e desfecho, não houve evidência de diferença nas distribuições marginais (p = 0,88), indicando que a proporção de casos de dengue não difere de forma significativa entre os gêneros na amostra. Logo, esses resultados reforçam a importância de considerar a área de residência como fator relevante que influencia a ocorrência do desfecho dengue, enquanto o gênero, isoladamente, não apresenta variação significativa na distribuição do desfecho, corroborando as análises de associação e heterogeneidade previamente feitas.

Regressão Logística

A regressão logística é uma utilizada para modelar a relação entre uma variável resposta dicotômica e as variáveis explicativas. Quando as variáveis explicativas são categorizadas, elas são incluídas no modelo por meio de variáveis chamadas de *dummies*, permitindo avaliar o efeito de cada categoria em comparação a uma referência. Esse modelo estima a chance do evento de interesse ocorrer a partir do logaritmo da razão de chance, o que possibilita interpretar o impacto de cada categoria na chance do desfecho. Asim, é bastante útil em contextos onde o objetivo é compreender como diferentes categorias das variáveis influenciam a ocorrência de uma doença, no caso da dengue Possibilita também compreender a influência de interações entre as variáveis explicativas no desfecho.

Nesse caso, para a modelagem não foram utilizadas as variáveis IgG, IgM e NS1, visto que os testes explicam quase por completo o desfecho, fazendo com que não seja possível os coeficientes do modelo convergirem. Todas as outras variáveis foram utilizadas.

Para selecionar se seria incluída a seleção de alguma interação entre variáveis no modelo, foi utilizado o método de seleção *stepwise*, que, de forma iterativa, avalia se a remoção ou adição de interações entre variáveis acarreta em uma diminuição do AIC. O objetivo é encontrar um modelo mais simples e com bom ajuste, evitando o excesso de variáveis que não contribuem significativamente para o modelo. O AIC é uma

boa métrica para alcançar essa meta, já que penaliza a complexidade do modelo para evitar problemas de *overfitting*, mas também considera se o modelo se ajusta bem aos dados, cumprindo então o princípio da parcimônia. A seguir, é possível observar os resultados do primeiro passo do método *stepwise*. Além do AIC, há informação sobre desvio do modelo, graus de liberdade e teste de razão de verossimilhança.

Id	Modelo	Desvio	Graus de liberdade	Razão de verossimilhança	AIC
1	Independente	1,311.90	959		1,393.900
2	Adição de interação Gender:AreaType	1,311.53	958		1,395.529
3	Adição de interação Gender:Age_less_16_years	1,311.72	958	-0.561	1,395.719
4	Adição de interação Age_less_16_years:AreaType	1,311.89	958	0.022	1,395.888
5	Adição de interação HouseType:Age_less_16_years	1,310.11	957	1.950	1,396.106
6	Adição de interação Gender:Housetype	1,311.38	957	-3.059	1,397.384
7	Adição de interação HouseType:AreaType	1,311.58	957	1.079	1,397.582
8	Adição de interação Gender:Area	1,255.92	924	55.855	1,407.924
9	Adição de interação Area:Age_less_16_years	1,272.44	925	-72.177	1,422.444
10	Adição de interação Area:AreaType	1,279.15	924	9.816	1,431.148
11	Adição de interação Area:HouseType	1,249.09	958	36.762	1,471.089

É possível observar que o modelo independente, ou seja, aquele que não considera interações entre as variáveis, apresenta desvio de 1311,9, 959 graus de liberdade e um AIC de 1393,9. Os modelos 2 a 7 apresentam mudança no desvio bem pequena, indicando que a adição das interações não contribui significativamente para o modelo - o que é embasado também pelos resultados do teste de razão de verossimilhança. Já nos modelos 8 a 11, há uma maior queda no desvio e valores maiores também para o teste de razão de verossimilhança. No entanto, os valores de AIC são maiores, indicando que o ganho no ajuste pode não compensar a complexidade do modelo. Assim, observando o menor valor de AIC, a melhor decisão nesse caso é não adicionar interações entre as variáveis e adotar o modelo independente.

O modelo de regressão logística pode ser interpretado a partir das razões de chance entre as categorias de variáveis considerando uma categoria de referência. Além disso, também é possível obter um intervalo de confiança para essa razão de chances. Uma razão de chances maior que 1 indica aumento na chanece de ocorrência do evento, enquanto menor que 1 indica diminuição dessa chance Essas informações - considerando o modelo escolhido - são apresentadas abaixo.

Variável	OR	IC.2.5..	IC.97.5..
(Intercept)	1.705	0.703	4.410
GenderMale	0.958	0.736	1.246
AreaBadda	0.684	0.216	2.087
AreaBanasree	0.289	0.087	0.901
AreaBangshal	1.355	0.425	4.281
AreaBiman Bandar	0.385	0.119	1.177
AreaBosila	0.473	0.140	1.527
AreaCantonment	0.864	0.244	3.047
AreaChawkbazar	0.561	0.163	1.861
AreaDemra	1.220	0.395	3.691
AreaDhanmondi	0.575	0.185	1.711
AreaGendaria	0.395	0.112	1.310
AreaGulshan	0.566	0.175	1.757
AreaHazaribagh	0.515	0.155	1.646
AreaJatrabari	2.523	0.763	8.606
AreaKadamtali	0.696	0.224	2.084
AreaKafrul	1.031	0.309	3.421
AreaKalabagan	0.674	0.206	2.136
AreaKamrangirchar	0.354	0.105	1.125
AreaKeraniganj	0.400	0.127	1.199
AreaKhilgaon	0.860	0.252	2.897
AreaKhilkhet	0.651	0.204	1.996
AreaLalbagh	0.778	0.207	2.916
AreaMirpur	0.702	0.226	2.099
AreaMohammadpur	1.382	0.414	4.647
AreaMotijheel	0.702	0.215	2.223
AreaNew Market	0.396	0.125	1.187
AreaPallabi	0.856	0.251	2.884
AreaPaltan	0.459	0.144	1.390
AreaRamna	0.283	0.085	0.878
AreaRampura	0.491	0.154	1.497
AreaSabujbagh	1.058	0.308	3.645
AreaShahbagh	0.478	0.141	1.544

Variável	OR	IC.2.5..	IC.97.5..
AreaSher-e-Bangla Nagar	0.357	0.102	1.174
AreaShyampur	0.531	0.147	1.846
AreaSutrapur	0.273	0.082	0.842
AreaTejgaon	1.887	0.561	6.518
AreaTypeUndeveloped	1.174	0.902	1.527
HouseTypeOther	0.911	0.664	1.251
HouseTypeTinshed	0.984	0.713	1.358
Age_less_16_yearsYes	1.172	0.805	1.713

Aqui serão pontuadas algumas interpretações relevantes:

- A razão de chances para os gêneros masculino e feminino é bem próxima de 1, indicando que há pouca diferença na chance de contrair dengue entre homens e mulheres em Bangladesh;
- Os intervalos de confiança para a razão de chances das variáveis referentes ao tipo de casa, tipo de área e idade todos contém 1, indicando também que bão podemos afirmar que há diferença na chance de ter dengue nos diferentes grupos;
- Para a variável referente à área, a área de referência utilizada foi Adabor. Algumas áreas apresentaram diferença na chance de testar positivo para dengue ao observar o intervalo de confiança, como Banasree, Ramna e Sutrapur. No entanto, o restante não apontou diferença.

Conclusão

Este estudo avaliou dados categorizados de indivíduos testados para dengue em Bangladesh, com ênfase na performance dos testes diagnósticos e nos fatores associados ao desfecho da doença.

Os resultados mostraram que os testes IgG e NS1 apresentaram alta acurácia, com destaque para o IgG, que atingiu sensibilidade e especificidade perfeitas. Em contrapartida, o teste IgM apresentou baixo desempenho, indicando que sua utilização isolada não é adequada para o diagnóstico confiável da dengue.

As análises de associação revelaram que idade, gênero, tipo de área e tipo de habitação não se associaram de forma significativa à ocorrência da doença. Entretanto, observou-se heterogeneidade nas razões de chances em algumas áreas específicas, sugerindo que a localização geográfica desempenha papel importante na distribuição dos casos.

Assim, conclui-se que os testes IgG e NS1 são os mais indicados para o diagnóstico da dengue nesta população, enquanto fatores individuais não explicam a ocorrência da doença. A área de residência, no entanto, mostrou-se relevante e deve ser considerada em políticas públicas para ações para combater a dengue em Bangladesh.

Apêndice

Códigos utilizados em R.

```r
# Pacotes utilizados
library(readr)
library(dplyr)
library(janitor)
library(tidyr)
library(maxstat)
library(data.table)
library(caret)
library(PropCIs)

# Carregando dados
dados_aux <-  fread("https://www.kaggle.com/api/v1/datasets/download/kawsarahmad/dengue-dataset-banglad

# Pontos de corte para a idade
cutpoint <- maxstat.test(Outcome ~ Age, data = dados_aux)
cutpoint$estimate


dados <- dados_aux |>
  mutate(
    Age_less_16_years = case_when(Age < 16 ~ "Yes",
                                  Age >= 16 ~ "No"),
    NS1 = case_when(NS1 == 1 ~ "Positive",
                    NS1 == 0 ~ "Negative"),
    IgG = case_when(IgG == 1 ~ "Positive",
                    IgG == 0 ~ "Negative"),
    IgM = case_when(IgM == 1 ~ "Positive",
                    IgM == 0 ~ "Negative"),
    Outcome = case_when(Outcome == 1 ~ "Positive",
                    Outcome == 0 ~ "Negative")
  ) |>
  select(-c(Age, District))

# Número de variáveis faltantes
sum(is.na(dados))

# Transformando todas as variáveis em fator

dados <- dados |>
  mutate(
    Gender = factor(Gender),
    IgG = factor(IgG),
    NS1 = factor(NS1),
    IgM = factor(IgM),
    Area = factor(Area),
    AreaType = factor(AreaType,
                      levels = c("Undeveloped", "Developed")),
    HouseType = factor(HouseType,
                       levels = c("Tinshed", "Building", "Other")),
    Age_less_16_years = factor(Age_less_16_years),
    Outcome = factor(Outcome)
  )
```

```r
write.csv(dados, "dados_tratados.csv")

#--------------- Análise de teste diagnósticos ----------------------

# métricas para NS1
confusionMatrix(as.factor(dados$NS1),
                as.factor(dados$Outcome), positive = "Positive")
#métricas para IgG
confusionMatrix(as.factor(dados$IgG),
                as.factor(dados$Outcome), positive = "Positive")
#métricas para IgM
confusionMatrix(as.factor(dados$IgM),
                as.factor(dados$Outcome), positive = "Positive")

#--------------- Diferença de duas proporções ----------------------


comparar_proporcoes <- function(data, grupo_var, desfecho_var,
                                positivo_label = "Positive",
                                conf.level = 0.95) {
  # Extrair os dois níveis do grupo
  grupo_niveis <- unique(data[[grupo_var]])
  if (length(grupo_niveis) != 2) stop("A variável do grupo deve ter exatamente dois níveis.")

  g1 <- grupo_niveis[1]
  g2 <- grupo_niveis[2]

  # Totais
  n1 <- nrow(dplyr::filter(data, !!rlang::sym(grupo_var) == g1))
  n2 <- nrow(dplyr::filter(data, !!rlang::sym(grupo_var) == g2))

  # Positivos
  x1 <- nrow(dplyr::filter(data, !!rlang::sym(grupo_var) == g1, !!rlang::sym(desfecho_var) == positivo_
  x2 <- nrow(dplyr::filter(data, !!rlang::sym(grupo_var) == g2, !!rlang::sym(desfecho_var) == positivo_

  # Proporções
  p1 <- x1 / n1
  p2 <- x2 / n2
  diff <- p1 - p2

  # Erro padrão e IC
  z <- qnorm(1 - (1 - conf.level)/2)
  se <- sqrt( (p1 * (1 - p1)) / n1 + (p2 * (1 - p2)) / n2 )
  lower <- diff - z * se
  upper <- diff + z * se

  # Retorno
  result <- list(
    grupo_1 = g1,
    grupo_2 = g2,
    total_g1 = n1,
    total_g2 = n2,
    positivos_g1 = x1,
```

```
    positivos_g2 = x2,
    prop_g1 = p1,
    prop_g2 = p2,
    diff_prop = diff,
    ic_95 = c(lower, upper)
  )

  return(result)
}

# Para Gender
res_gender <- comparar_proporcoes(dados, grupo_var = "Gender",
                                  desfecho_var = "Outcome")
print(res_gender)

# Para AreaType
res_areatype <- comparar_proporcoes(dados, grupo_var = "AreaType",
                                    desfecho_var = "Outcome")
print(res_areatype)

# Para Age_less_16_years
res_age <- comparar_proporcoes(dados, grupo_var = "Age_less_16_years",
                               desfecho_var = "Outcome")
print(res_age)

#--------------- Análises de Razão de Chances -----------------------

#### Gender

# Contagens
a <- nrow(filter(dados, Gender == "Male" &
                   Outcome == "Positive"))
b <- nrow(filter(dados, Gender == "Male" &
                   Outcome == "Negative"))
c <- nrow(filter(dados, Gender == "Female" &
                   Outcome == "Positive"))
d <- nrow(filter(dados, Gender == "Female" &
                   Outcome == "Negative"))

# OR, log(OR), SE
or_gender <- (a * d) / (b * c)
log_or_gender <- log(or_gender)
se_log_or_gender <- sqrt(1/a + 1/b + 1/c + 1/d)

# IC 95%
z <- qnorm(0.975)
ic_log_gender <- log_or_gender + c(-1, 1) * z * se_log_or_gender
ic_gender <- exp(ic_log_gender)

# Resultado
cat(" Gender:\n")
cat("  OR =", round(or_gender, 3), "\n")
cat("  log(OR) =", round(log_or_gender, 3), "\n")
```

```r
cat("  IC 95% OR = [", round(ic_gender[1], 3), ",", round(ic_gender[2], 3), "]\n\n")


#### AreaType

a <- nrow(filter(dados, AreaType == "Developed" &
                   Outcome == "Positive"))
b <- nrow(filter(dados, AreaType == "Developed" &
                   Outcome == "Negative"))
c <- nrow(filter(dados, AreaType == "Undeveloped"
                 & Outcome == "Positive"))
d <- nrow(filter(dados, AreaType == "Undeveloped"
                 & Outcome == "Negative"))

or_area <- (a * d) / (b * c)
log_or_area <- log(or_area)
se_log_or_area <- sqrt(1/a + 1/b + 1/c + 1/d)

ic_log_area <- log_or_area + c(-1, 1) * z * se_log_or_area
ic_area <- exp(ic_log_area)

cat(" AreaType:\n")
cat("  OR =", round(or_area, 3), "\n")
cat("  log(OR) =", round(log_or_area, 3), "\n")
cat("  IC 95% OR = [", round(ic_area[1], 3), ",",
    round(ic_area[2], 3), "]\n\n")

#### Age_less_16_years

a <- nrow(filter(dados, Age_less_16_years == "Yes" &
                   Outcome == "Positive"))
b <- nrow(filter(dados, Age_less_16_years == "Yes" &
                   Outcome == "Negative"))
c <- nrow(filter(dados, Age_less_16_years == "No" &
                   Outcome == "Positive"))
d <- nrow(filter(dados, Age_less_16_years == "No" &
                   Outcome == "Negative"))

or_age <- (a * d) / (b * c)
log_or_age <- log(or_age)
se_log_or_age <- sqrt(1/a + 1/b + 1/c + 1/d)

ic_log_age <- log_or_age + c(-1, 1) * z * se_log_or_age
ic_age <- exp(ic_log_age)

cat(" Age < 16:\n")
cat("  OR =", round(or_age, 3), "\n")
cat("  log(OR) =", round(log_or_age, 3), "\n")
cat("  IC 95% OR = [", round(ic_age[1], 3), ",",
    round(ic_age[2], 3), "]\n\n")

# ------------- Modelo de Regressão Logística
```

```r
##### Escolha de interações no modelo

# Preparar dados (removendo IgG, que prevê perfeitamente o Outcome)
dados_modelo <- dados |>
  select(-c(IgG, IgM, NS1))

dados_modelo$Outcome <- as.factor(dados_modelo$Outcome)

# Modelo independente
modelo_independente <- glm(Outcome ~ Gender + Area + AreaType + HouseType + Age_less_16_years,
                   data = dados_modelo,
                   family = binomial)

# Modelo completo com interações
modelo_completo <- glm(
  Outcome ~ Gender*Area*AreaType*HouseType*Age_less_16_years,
  data = dados_modelo,
  family = binomial
)

# Stepwise
modelo_final <- step(
  object = modelo_independente,
  scope = list(lower = modelo_independente, upper = modelo_completo),
  direction = "both",
  trace = TRUE
)

summary(modelo_final)

# Analisando como as variáveis dimunuem o desvio

fit0 <-glm(Outcome ~ Gender + Area + AreaType + HouseType  + Age_less_16_years,
           data = dados_modelo, family = binomial)

fit1 <-glm(Outcome ~ Gender*AreaType + Area + AreaType + HouseType  + Age_less_16_years,
           data = dados_modelo, family = binomial)

fit2 <-glm(Outcome ~ Gender*Age_less_16_years + Area + AreaType + HouseType,
           data = dados_modelo, family = binomial)

fit3 <-glm(Outcome ~ Gender + Area + HouseType  + Age_less_16_years*AreaType,
           data = dados_modelo, family = binomial)

fit4 <-glm(Outcome ~ Gender + Area + AreaType  + Age_less_16_years*HouseType,
           data = dados_modelo, family = binomial)

fit5 <-glm(Outcome ~  Area + AreaType + HouseType*Gender  + Age_less_16_years,
           data = dados_modelo, family = binomial)

fit6 <-glm(Outcome ~ Gender + Area + AreaType*HouseType  + Age_less_16_years,
           data = dados_modelo, family = binomial)
```

```r
fit7 <-glm(Outcome ~ Gender*Area + AreaType + HouseType  + Age_less_16_years,
           data = dados_modelo, family = binomial)

fit8 <-glm(Outcome ~ Gender + AreaType + HouseType  + Age_less_16_years*Area,
           data = dados_modelo, family = binomial)

fit9 <-glm(Outcome ~ Gender + Area*AreaType + HouseType  + Age_less_16_years,
           data = dados_modelo, family = binomial)

fit10 <-glm(Outcome ~ Gender + AreaType + HouseType*Area  + Age_less_16_years,
            data = dados_modelo, family = binomial)



a <- anova(fit0, fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fit9, fit10, test = "Chisq")


desvio <- c(deviance(fit0), deviance(fit1), deviance(fit2), deviance(fit3), deviance(fit4),
            deviance(fit5), deviance(fit6), deviance(fit7), deviance(fit8),
            deviance(fit9), deviance(fit10))

aic <- c(AIC(fit0), AIC(fit1), AIC(fit2), AIC(fit3), AIC(fit4),
         AIC(fit5), AIC(fit6), AIC(fit7), AIC(fit8), AIC(fit9),
         AIC(fit10))

# Criar a estatística do teste LR
lr_stat <- c(NA, round(diff(a$Deviance), 3))

# Odds ratios e ICs
exp(cbind(OR = round(coef(modelo_final), 3), round(confint(modelo_final), 3)) )


# Medidas de assosiação de variaveis ordinais

library (DescTools)

tab <- table(dados_tratados$Outcome, dados_tratados$Age_less_16_years)
GoodmanKruskalGamma (tab)
KendallTauB (tab)


dados_tratados$AreaType <- factor(dados_tratados$AreaType, levels = c("Undeveloped", "Developed"))
tab2 <- table(dados_tratados$Outcome, dados_tratados$AreaType)
GoodmanKruskalGamma (tab2)
KendallTauB (tab2)

# Teste de tendencia linear de variaveis ordinais
library(coin)
CochranArmitageTest(tab)
CochranArmitageTest(tab2)


# Teste de Breslow day
```

```r
# Tabela estratificada Outcome x Age_less_16_years por AreaType
tab_age_area <- with(dados_tratados, table(Age_less_16_years,
                                           Outcome, AreaType))
tab_age_area
# Aplicar o teste Breslow-Day e de Mantel
BreslowDayTest(tab_age_area)
mantelhaen.test(tab_age_area)


# Tabela estratificada Outcome x Age_less_16_years por AreaType
tab3 <- with(dados_tratados, table(Gender, Outcome, Area))
# Aplicar o teste Breslow-Day e de Mantel
BreslowDayTest(tab3)
mantelhaen.test(tab3)

library(dplyr)
# Criar uma função para aplicar o qui-quadrado por área
chi_sq_by_area <- dados_tratados %>%
  group_by(Area) %>%
  summarise(
    p_value = chisq.test(table(Gender, Outcome))$p.value,
    statistic = chisq.test(table(Gender, Outcome))$statistic
  )

print(chi_sq_by_area, n = 36)

# associação simples entre AreaType e Outcome
tab_area_outcome <- with(dados_tratados, table(AreaType, Outcome))
chi_sq_area <- chisq.test(tab_area_outcome)
chi_sq_area

# Teste de simetria usando o teste de McNemar
tab_diagnostics <- with(dados_tratados, table(IgG, NS1))
mcnemar.test(tab_diagnostics)

# Teste de simetria usando o teste de McNemar
tab_diagnostics <- with(dados_tratados, table(IgG, IgM))
mcnemar.test(tab_diagnostics)

# Teste de simetria usando o teste de McNemar
tab_diagnostics <- with(dados_tratados, table(IgM, NS1))
mcnemar.test(tab_diagnostics)


# Teste de Homogeneidade Marginal para duas variáveis categóricas
tab_gender_outcome <- with(dados_tratados, table(Gender, Outcome))
homogeneidade_test <- chisq.test(tab_gender_outcome)
print(homogeneidade_test)

tab_area_outcome <- with(dados_tratados, table(Area, Outcome))
homogeneidade_test_2 <- chisq.test(tab_area_outcome)
print(homogeneidade_test_2)
```

```r
# Graficos da analise exploratória
ggplot(dados_tratados) +
  aes(x = NS1, fill = Outcome) +
  geom_bar() +
  scale_fill_manual(values = c("Negative" = "#FC0041",
                               "Positive" = "#21B7D1")) +
  labs(title = "NS1 por Outcome") +
  theme_light(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5,
                                  face = "bold"))

ggplot(dados_tratados) +
  aes(x = IgG, fill = Outcome) +
  geom_bar() +
  scale_fill_manual(values = c("Negative" = "#FC0041",
                               "Positive" = "#21B7D1")) +
  labs(title = "lgG por Outcome ") +
  theme_light(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5,
                                  face = "bold"))

ggplot(dados_tratados) +
  aes(x = IgM, fill = Outcome) +
  geom_bar() +
  scale_fill_manual(values = c("Negative" = "#FC0041",
                               "Positive" = "#21B7D1")) +
  labs(title = "lgM por Outcome ") +
  theme_light(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5,
                                  face = "bold"))

ggplot(dados_tratados) +
  aes(x = Age_less_16_years, fill = Outcome) +
  geom_bar() +
  scale_fill_manual(values = c("Negative" = "#FC0041",
                               "Positive" = "#21B7D1")) +
  theme_light(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  facet_wrap(vars(Gender)) +
  labs(
    title = "Variável Idade por Outcome separados por gênero",
    x = "Idade < 16 anos",
    y = "Contagem"
  )
```