Relatório de Dengue em Bangladesh

MarianaFreitas e Aline Pires

Introdução

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 rÖs, foram aplicados algins testes para avaliar associação e simetria. O teste qui-quadrado de Pearson foi utilizado para verificar a independência entre linhas e colunas. O teste de razão de verossimilhança, uma alternativa ao qui-quadrado, tem base no modelo de log-verossimilhança, sendo mais adequado em amostras pequenas ou quando os pressupostos de normalidade não são satisfeitos. 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. Pela base de dados desse trabalho se tratar de um problema com variável resposta binária, os modelos log-lineares não foram utilizados. 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

- Falar sobre os dados (variáveis, fonte, ...) Aline
- falar que apenas age_less_16 e AreaType são ordinais Aline
- Fazer análises que julgar interessantes Aline

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 tipo 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 ocorrância de dengue, não houve nenhum tipo de intervenção ou acompanhamento dos indivíduos. Esse tipo de estudo permite a 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

Gender/Outcome	Negative	Positive
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

- Teste Gama de Goodman e Kruskal (usar variáveis idade e areatype) Aline
- Teste Tau de Kendall e Tau-b de Kendall (usar variáveis idade e areatype) Aline
- Cochran-Armitage para tendência linear (usar variáveis idade e areatype) Aline

Associação em Tabelas de Contigência

- Teste de Breslow Day Aline
- Se homogeneidade aceita: teste de Mantel-Haenszel - falar se ocorre ou não o paradoxo de simpason Aline
- Teste qui-quadrado Aline
- Teste de razão de verossimilhança Aline
- Teste de homogeneidade Aline
- Teste de simetria (testar simetria entre IgG, IgM, e NS1? ou de outras variáveis?) Aline
- Teste de homogeneidade marginal Aline

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

Id	Modelo	Desvio	Graus de liberdade	Razão de verossimilhança	AIC
4	Adição de interação Age_less_16_years:AreaType	1,311.89	958	0.022	1,395.888
5	Adição de interação House-Type: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

Variável	OR	IC.2.5..	IC.97.5..
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
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

Variável	OR	IC.2.5..	IC.97.5..

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 aapontou diferença.

Conclusão

Apêndice

Códigos utilizados em R.

Copiar e colar o analises_dengue_bangladesh_script.R aqui !