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. Uma forma de visualizar essas medidas de forma gráfica é a curva ROC, que mostra a capacidade do teste de distinguir entre as duas classes.

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 probabilidade 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. Fio 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, . . .)
- falar que apenas age_less_16 e AreaType são ordinais
- Fazer análises que julgar interessantes

Avaliação de Testes Diagnósticos

- Tabela de testes diagnósticos
- Sensibilidade
- Especificidade
- Curva ROC

Tabelas de Contigência

- Explicar tabelas parciais
- Tipo de estudo
- Diferença de proporções em tabelas 2x2
- Razão de chances
- Inferência e log das razões de chance
- Intervalos de confiança para razão de chances

Inferência para Tabelas de Contigência

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

Associação em Tabelas de Contigência

- Teste de Breslow Day
- Se homogeneidade aceita: teste de Mantel-Haenszel - falar se ocorre ou não o paradoxo de simpason
- Teste qui-quadrado
- Teste de razão de verossimilhança
- Teste de homogeneidade
- Teste de simetria
- Teste de homogeneidade marginal

Regressão Logística

- Testar opções de modelos de regressão logística
- Selecionar baseado no AIC
- Interpretação do modelo

Conclusão