Análise de Dados Categóricos
Resposta multicategórica
Seções 3.3, 3.3.1 e 3.3.2

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Outubro, 2023

- **3.3 Modelos de regressão de resposta nominal**
- **3.3.1 Odds ratios**
- **3.3.2 Tabelas de contingência**

Passamos agora ao problema de modelar as probabilidades de uma variável de resposta categórica Y com categorias de resposta $j = 1, \cdots, J$ usando variáveis explicativas $x_1, \cdots, x_p$.

Foram definidas probabilidades para uma resposta binária como $P(\text{sucesso})/P(\text{fracasso})$. De maneira mais geral, para uma resposta multinomial, podemos definir as probabilidades como uma comparação de qualquer par de categorias de resposta.

Por exemplo, $\pi_j/\pi_{j'}$ é a probabilidade da categoria j relativa a j'. Um modelo de regressão popular para respostas multinomiais é desenvolvido então selecionando uma categoria de resposta como o nível base e formando as chances das $J - 1$ categorias restantes contra este nível.

Considere que $j = 1$ represente a categoria de nível básico e forme as chances (odds) π_j/π_1 para $j = 2, \cdots, J$. Essas probabilidades são então modeladas em função de variáveis explicativas usando uma forma generalizada de regressão logística.

Observe que se $J = 2$, temos $\log(\pi_2/\pi_1) = \log\left(\pi_2/(1-\pi_2)\right)$, que é equivalente a $\log\left(\pi/(1-\pi)\right)$ como na regressão logística, a categoria de resposta 2 torna-se a categoria de "sucesso".

Especificamente, um modelo de regressão multinomial, também conhecido como modelo logit de categoria de linha de base relaciona um conjunto de variáveis explicativas a cada probabilidade logarítmica de acordo com

$$\log(\pi_j/\pi_1) = \beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p,$$

para $j = 2, \cdots, J$.

Observe que o primeiro subscrito nos parâmetros β corresponde à categoria de resposta, o que permite que os log-odds de cada resposta se relacionem com as variáveis explicativas de uma maneira diferente.

Além disso, observe que, subtraindo os log-odds apropriadas, podemos reescrever a equação acima para comparar qualquer par de categorias de resposta.

Por exemplo, para achar $\log(\pi_j/\pi_{j'})$ onde $j' \neq 1$ e $j' \neq j$, temos

$$\begin{aligned}\log(\pi_j/\pi_{j'}) &= \log(\pi_j) - \log(\pi_{j'}) \\ &= \log(\pi_j/\pi_1) - \log(\pi_{j'}/\pi_1) \\ &= (\beta_{j0} - \beta_{j'0}) + (\beta_{j1} - \beta_{j'1})x_1 + \cdots (\beta_{jp} - \beta_{j'p})x_p.\end{aligned}$$

Assim, a escolha do nível de base não é importante e pode ser feita com base na conveniência ou interpretação.

As probabilidades para cada categoria individual também podem ser encontradas em termos do modelo. Podemos reescrever

$$\pi_j = \pi_1 \exp\left(\beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p\right)$$

usando propriedades de logaritmos.

Observando que $\pi_1 + \pi_2 + \cdots + \pi_J = 1$, temos

$$\pi_1 + \pi_1 \exp\left(\beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p\right)$$

$$+ \cdots + \pi_1 \exp\left(\beta_{J0} + \beta_{J1}x_1 + \cdots + \beta_{Jp}x_p\right) = 1.$$

Fatorando o π_1 comum em cada termo, obtemos uma expressão para π_1:

$$\pi_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp\left(\beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p\right)}.$$

Isso leva a uma expressão geral para π_j:

$$\pi_j = \frac{\exp\left(\beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p\right)}{1 + \sum_{\ell=2}^{J} \exp\left(\beta_{\ell0} + \beta_{\ell1}x_1 + \cdots + \beta_{\ell p}x_p\right)},$$

para $j = 2, \cdots, J$.

Os parâmetros para o modelo são estimados usando a máxima verossimilhança. Para uma amostra de observações, y_i que denotam a categoria resposta e as correspondentes variáveis explicativas $x_{i1}, \cdots, x_{ip}$ para $i = 1, \cdots, m$, a função de verossimilhança é simplesmente o produto de m distribuições multinomiais com parâmetros de probabilidade dados anteriormente.

Procedimentos numéricos iterativos são então usados para encontrar as estimativas dos parâmetros. A função **multinom()** do pacote **nnet** executa os cálculos necessários. Conforme mencionado na ajuda desta função, redimensionar as variáveis explicativas para que fiquem entre 0 e 1 pode ajudar às vezes com a convergência das estimativas dos parâmetros.

A matriz de covariância estimada para as estimativas de parâmetros é encontrada usando procedimentos de verossimilhança padrão. Os métodos de inferência baseados nas estatísticas de Wald e LR são executados da mesma forma que os procedimentos de verossimilhança nos capítulos anteriores.

No entanto, os intervalos de razão de verossimilhança perfilada são difíceis de obter em R, porque nem uma função do método **confint()** nem no pacote **mcprofile** foram estendidos para calcular esses tipos de intervalos para objetos resultantes de **multinom()**.

Por esta razão, focamos no uso de intervalos de Wald nesta seção.

Exemplo 3.4: Grãos de trigo (Kernels).

Como as probabilidades log-odds ou logaritmos da probabilidades de chances são modeladas diretamente em um modelo de regressão multinomial, as razões de chances são úteis para interpretar a relação de uma variável explicativa com a resposta.

Conforme descrito na Seção 2.2.3 para modelos de regressão logística, as razões de chances para variáveis explicativas numéricas representam a mudança nas chances correspondentes a um aumento de c unidades em uma variável explicativa específica.

A única diferença com os modelos multinomiais é que as probabilidades agora são formadas como uma comparação entre duas das J categorias de resposta.

Por exemplo, suponha que os termos na equação

$$\log(\pi_j/\pi_1) = \beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p,$$

sejam variáveis explicativas distintas, não transformações ou interações.

As chances de uma resposta da categoria j versus uma resposta da categoria 1 são $\exp(\beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p)$.

Essas probabilidades mudam por $\exp(c\beta_{jr})$ vezes para cada c unidade de aumento em x_r, mantendo as outras variáveis constantes.

Da mesma forma, a partir da equação

$$\log(\pi_j/\pi_{j'}) = (\beta_{j0} - \beta_{j'0}) + (\beta_{j1} - \beta_{j'1})x_1 + \cdots + (\beta_{jp} - \beta_{j'p})x_p,$$

as chances de uma categoria j vs. uma resposta de categoria j', $j \neq j', j > 1$ e $j' > 1$, mudam por $\exp\left(c(\beta_{jr} - \beta_{j'r})\right)$ vezes para cada c unidades de aumento em x_r mantendo as outras variáveis no modelo constantes.

As estimativas de máxima verossimilhança das razões de chances são obtidas substituindo os parâmetros de regressão por suas estimativas correspondentes. Os métodos de inferência baseados em Wald e LR para odds ratios são usados da mesma forma que discutimos nos capítulos anteriores.

Exemplo 3.5: Grãos de trigo (Kernels).

O modelo de regressão multinomial fornece uma maneira conveniente de realizar o teste da razão de verossimilhanças (LRT) para independência descrito na Seção 3.2.3. Podemos tratar a variável de linha X como uma variável explicativa categórica (consulte a Seção 2.2.6) construindo $I - 1$ variáveis indicadoras $x_2, \cdots, x_I$ representando os níveis $2, \cdots, I$ de X.

Observe que excluímos o rótulo x_1 por conveniência, para que os índices das variáveis indicadoras correspondam aos níveis de X. Usando Y como variável de resposta com probabilidades de categorias $\pi_1, \cdots, \pi_J$, o modelo de regressão multinomial para a tabela é

$$\log\left(\pi_j/\pi_1\right) = \beta_{j0} + \beta_{j2}x_2 + \cdots + \beta_{jI}x_I,$$

para $= j = 2, \cdots, J$.

Este modelo implica que

$$\log\left(\pi_j/\pi_1\right) = \beta_{j0}$$

quando $X = 1$ e

$$\log\left(\pi_j/\pi_1\right) = \beta_{j0} + \beta_{j1}$$

quando $X = i$ para $i = 2, \cdots, I$. Portanto, o log-odds entre duas colunas depende da linha da tabela. Observe que existem $I(J-1)$ parâmetros de regressão no total. Porque $\sum_{j=1}^{J} \pi_j = 1$ para cada X, ou seja, apenas as $J-1$ probabilidades de resposta precisam ser estimadas em cada linha, o modelo está saturado.

A independência entre X e Y remove essa dependência de linha, de modo que o log-odds entre duas colunas seja constante entre as linhas. Isso implica no modelo simplificado

$$\log\left(\pi_j/\pi_1\right) = \beta_{j0}.$$

Assim, testar a independência é equivalente a testar

$$H_0 : \beta_{j2} = \cdots = \beta_{jI} = 0,$$

para cada $j = 2, \cdots, J$ vs.

$$H_1 : \text{Pelo menos um } \beta_{ji} \neq 0 \text{ para alguns } j \text{ e } i.$$

Este teste é facilmente realizado usando o teste da razão de verossimilhanças (LRT). Quando a hipótese nula é rejeitada, o próximo passo é investigar as fontes de dependência.

Como na Seção 3.2.3, o cálculo das razões de chances para seções 2×2 selecionadas da tabela de contingência pode às vezes ajudar a explicar a natureza da associação.

Essas razões de chances são facilmente obtidas a partir dos parâmetros do modelo de regressão. Por exemplo, as chances estimadas para comparar as categorias de resposta $Y = j$ a $Y = 1$ são $\exp\left(\widehat{\beta}_{ji}\right)$ vezes maiores para $X = i$ do que $X = 1$. As razões de chances (odds ratios) envolvendo linhas e colunas diferentes da primeira são encontradas como exponenciais de combinações lineares de parâmetros de regressão. Por exemplo, as probabilidades estimadas para comparar as categorias de resposta 2 e 3 são

$$\exp\left((\widehat{\beta}_{24} - \widehat{\beta}_{34}) - (\widehat{\beta}_{25} - \widehat{\beta}_{35})\right)$$

vezes maiores na linha 4 do que na linha 5. Os intervalos de confiança podem ser formados para essas razões de chances, conforme descrito na Seção 3.3.1.

Exemplo 3.6: Crackers enriquecidos com fibra.