Análise de Dados Categóricos
Regressão com resposta binária
Seções 2.2.5, 2.2.6 e 2.2.7

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Agosto, 2023

As interações e transformações ampliam a variedade de formas que podem ser propostas para se relacionar π com as variáveis explicativas.

Os termos são criados e adicionados ao preditor linear no modelo de regressão logística exatamente da mesma maneira que são usados na regressão linear.

Os termos mais comuns e interpretáveis a serem adicionados são interações em pares (two-way) e termos quadráticos. Embora outros tipos de termos também possam ser adicionados, por exemplo, interações de três termos (three-way) e transformações cúbicas, deixamos de explorar essas alternativas como exercícios.

Interações entre variáveis explicativas são necessárias quando o efeito de uma variável explicativa na probabilidade de sucesso depende do valor de uma segunda variável explicativa.

Existem algumas maneiras de incluir essas interações em um argumento de fórmula para **glm()**. Para descrevê-los, suponha que existam duas variáveis explicativas denominadas **x1** e **x2** em um arquivo de dados representando x_1 e x_2, e o objetivo é estimar o modelo

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

A inclusão de termos de interação e/ou transformações faz com que as razões de chances dependam do valor numérico de uma variável explicativa.

Por exemplo, considere o modelo

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

A razão de chances associada à mudança de x_2 por c unidades enquanto mantendo x_1 constante é

$$\frac{Odds_{x_2+c}}{Odds_{x_2}} = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 (x_2 + c) + \beta_3 x_1 (x_2 + c)\right)}{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2\right)}.$$

Assim, o aumento ou diminuição das chances de sucesso para uma mudança de unidade c em x_2 depende do nível de x_1, que decorre da definição de uma interação.

Como outro exemplo, considere o modelo

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2.$$

A razão de chances para uma mudança de unidade c em x_1 é

$$OR = \frac{Odds_{x_1+c}}{Odds_{x_1}} = \frac{\exp\left(\beta_0 + \beta_1(x_1+c) + \beta_2(x_2+c)^2\right)}{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_1^2\right)},$$

que depende do valor de x_1.

Os intervalos de confiança para essas razões de chances podem ser mais complicados de calcular, porque geralmente são baseados em combinações lineares de parâmetros de regressão, e a função genérica **confint()** não pode ser usada para a tarefa.

Para o exemplo de interação, onde $OR = \exp\left(c\left(\beta_2 + \beta_3 x_1\right)\right)$, o intervalo de confiança $(1-\alpha)100\%$ de Wald é

$$\exp\left(c\left(\widehat{\beta}_2 + \widehat{\beta}_3 x_1\right) \pm c\, Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\widehat{\beta}_2 + \widehat{\beta}_3 x_1)}\right)$$

com

$$\widehat{\text{Var}}(\widehat{\beta}_2 + \widehat{\beta}_3 x_1) = \widehat{\text{Var}}(\widehat{\beta}_2) + x_1^2\,\widehat{\text{Var}}(\widehat{\beta}_3) + 2x_1\,\widehat{\text{Cov}}(\widehat{\beta}_2, \widehat{\beta}_3).$$

Outros intervalos de Wald podem ser calculados de maneira semelhante. Geralmente, a única parte desafiadora é encontrar a variância para a combinação linear dos estimadores de parâmetros, então fornecemos código para fazer isso no próximo exemplo.

Os intervalos LR perfilados podem ser calculados para essas razões de chances usando o pacote **mcprofile**. Semelhante ao intervalo de Wald, a parte mais desafiadora é especificar a combinação linear correta dos estimadores de parâmetros. A função **wald()** deste pacote também fornece uma maneira um tanto automatizada de calcular os intervalos de confiança de Wald. Discutiremos os detalhes do cálculo no próximo exemplo.

Exemplo 2.11: Placekicking.

As variáveis explicativas categóricas são representadas no modelo de regressão logística da mesma forma que em um modelo de regressão linear. Se uma variável tiver q níveis categóricos, as $q-1$ variáveis indicadoras podem ser usadas para representá-la em um modelo.

Por exemplo, vimos anteriormente que **change** no conjunto de dados placekicking era uma variável categórica com dois níveis. Foi representado com uma variável indicadora: 0 para placekicks de troca sem chumbo e 1 para placekicks de troca de chumbo.

R trata as variáveis categóricas como um tipo de objeto de fator. Uma exceção pode ocorrer se a variável explicativa for codificada numericamente, como **change**. Discutiremos como lidar com essas situações mais adiante nesta seção.

Por padrão, R ordena os níveis usando uma ordenação numérica e, em seguida, alfabética, em que as letras minúsculas são ordenadas antes das maiúsculas. Para ver a ordenação de qualquer fator, a função **levels()** pode ser usada.

Essa ordenação é importante porque a maioria das funções em R a usa para construir variáveis indicadoras com o método de construção “definir primeiro nível como 0”.

A construção “definir último nível como 0” é usada às vezes por outros softwares estatísticos, como o SAS. Existem outras formas de construção; consulte a Seção 6.3.5 para obter um exemplo.

A tabela abaixo mostra a codificação R padrão para variáveis indicadoras de uma variável explicativa categórica com o $q = 4$ níveis codificados como **A**, **B**, **C** e **D**. O primeiro nível **A** é o nível base onde todas as variáveis indicadoras são definidas como 0.

Os níveis restantes são atribuídos a uma variável indicadora. No caso, o nível **B** é representado por x_1 onde $x_1 = 1$ para um nível observado de **B** e $x_1 = 0$ caso contrário. Os níveis **C** e **D** são definidos de maneira semelhante para x_2 e x_3, respectivamente.

A função **relevel()** em R pode ser usada para definir um novo nível base, se desejado, e isso será discutido mais detalhadamente no próximo exemplo.

Indicadoras para uma variável explicativa categórica de 4 níveis.

Nível	x_1	x_2	x_3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

Uma variável explicativa categórica é representada em um modelo de regressão logística usando todas as suas variáveis indicadoras. Por exemplo, um modelo de regressão logística com a variável explicativa categórica de 4 níveis é escrito como

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Uma representação um pouco menos formal substitui os nomes dos níveis para cada uma de suas variáveis indicadoras correspondentes; ou seja,

$$\mathit{logit}(\pi) = \beta_0 + \beta_1\,\mathrm{B} + \beta_2\,\mathrm{C} + \beta_3\,\mathrm{D}.$$

Para testar a importância de uma variável explicativa categórica, todos os seus parâmetros de regressão correspondentes devem ser iguais a 0 na hipótese nula. Por exemplo, testaríamos todas as três variáveis indicadoras na equação acima simultaneamente com

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \qquad \text{vs.} \qquad H_1 : \text{Nem todas iguais a } 0;$$

onde o LRT ou outro procedimento de teste apropriado é usado.

Exemplo 2.12: Controle do vírus da murcha do tomateiro.

Interações

Para representar uma interação entre uma variável explicativa categórica e uma variável explicativa contínua em um modelo, todos os produtos aos pares dos termos variáveis são necessários. Por exemplo, considere a equação

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

com um termo adicional de $\beta_4 v$ adicionado a ela, onde v denota genericamente uma variável explicativa contínua.

A interação entre a variável explicativa categórica de 4 níveis e v é representada pela inclusão de três produtos pareados entre v e x_1, x_2 e x_3 no modelo com parâmetros de regressão como coeficientes.

Para representar uma interação pareada entre duas variáveis explicativas categóricas, todos os produtos pareados entre os dois conjuntos de variáveis indicadoras são incluídos no modelo com parâmetros de regressão apropriados como coeficientes.

Por exemplo, suponha que haja um fator X de 3 níveis e um fator Z de 3 níveis, representados por variáveis indicadoras apropriadas x_1 e x_2 para X e z_1 e z_2 para Z.

O modelo de regressão logística com a interação entre X e Z é

$$\begin{aligned} logit(\pi) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 x_1 z_1 \\ &\quad + \beta_6 x_1 z_2 + \beta_7 x_2 z_1 + \beta_8 x_2 z_2. \end{aligned}$$

Para testar a interação, a hipótese nula é que todos os parâmetros de regressão para termos de interação são 0, ou seja,

$$H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0,$$

em nosso exemplo e a alternativa é que pelo menos um desses parâmetros não seja 0.

Exemplo 2.13: Controle do vírus da murcha do tomateiro.

Odds ratios

Considere novamente o modelo dado na equação

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

A chance de sucesso no nível **B** é $\exp(\beta_0 + \beta_1)$ porque $x_1 = 1$, $x_2 = 0$ e $x_3 = 0$ e as chances de sucesso no nível **A** são $\exp(\beta_0)$ porque $x_1 = 0$, $x_2 = 0$ e $x_3 = 0$. A razão de chances resultante comparando o nível **B** com o **A** é

$$\frac{Odds_{x_1=1,\, x_2=0,\, x_3=0}}{Odds_{x_1=0,\, x_2=0,\, x_3=0}} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1).$$

Da mesma maneira, pode-se mostrar que a razão de chances comparando **C** com **A** é $\exp(\beta_2)$ e a razão de chances comparando **D** com **A** é $\exp(\beta_3)$.

Podemos usar a mesma técnica para comparar níveis não básicos. Por exemplo, a razão de chances comparando o nível **B** ao nível **C** é

$$\frac{Odds_{x_1=1,\,x_2=0,\,x_3=0}}{Odds_{x_1=0,\,x_2=1,\,x_3=0}} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0+\beta_2)} = \exp(\beta_1-\beta_2).$$

As estimativas e os intervalos de confiança são formados da mesma forma discutida nas Seções 2.2.3 e 2.2.5. Por exemplo, a razão de chances estimada comparando o nível **B** com **C** é $\exp(\widehat{\beta}_1-\widehat{\beta}_2)$, e o $(1-\alpha)100\%$ intervalo de confiança de Wald é

$$\exp\left(\widehat{\beta}_1-\widehat{\beta}_2 \pm Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\widehat{\beta}_1-\widehat{\beta}_2)}\right)$$

onde

$$\widehat{\text{Var}}(\widehat{\beta}_1-\widehat{\beta}_2) = \widehat{\text{Var}}(\widehat{\beta}_1)+\widehat{\text{Var}}(\widehat{\beta}_2)-2\widehat{\text{Cov}}(\widehat{\beta}_1,\widehat{\beta}_2).$$

Quando há duas variáveis explicativas categóricas **X** e **Z**, a comparação de dois níveis para uma das variáveis é novamente realizada pela formação das razões das chances. Quando há interação entre as variáveis, essas razões de chance dependem do nível da outra variável explicativa.

Considere novamente a equação

$$\begin{aligned} logit(\pi) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 x_1 z_1 \\ &\quad + \beta_6 x_1 z_2 + \beta_7 x_2 z_1 + \beta_8 x_2 z_2, \end{aligned}$$

que modela as probabilidades logarítmicas de sucesso em relação a duas variáveis explicativas categóricas de 3 níveis, digamos, com os níveis **A**, **B** e **C** e sua interação.

Uma análise típica com foco em X é calcular as razões de chances para comparar **B** com **A**, **C** com **A** e **B** com **C**.

Devido à interação, cada uma dessas razões de chances precisa ser calculada condicionalmente a um nível de Z. Assim, há são 9 razões de chances diferentes entre os níveis de X.

Essas razões de chances são exibidas na tabela abaixo.

Razões de chances comparando níveis de X condicionais em um nível de Z.

Z	X	OR
A	B com A	$\exp(\beta_1)$
	C com A	$\exp(\beta_2)$
	B com C	$\exp(\beta_1 - \beta_2)$
B	B com A	$\exp(\beta_1 + \beta_5)$
	C com A	$\exp(\beta_2 + \beta_7)$
	B com C	$\exp(\beta_1 - \beta_2 + \beta_5 - \beta_7)$
C	B com A	$\exp(\beta_1 + \beta_6)$
	C com A	$\exp(\beta_2 + \beta_8)$
	B com C	$\exp(\beta_1 - \beta_2 + \beta_6 - \beta_8)$

Por exemplo, a comparação dos níveis **B** e **C** de X no nível **B** de Z é encontrada por

$$\frac{Odds_{x_1=1,\,x_2=0,\,z_1=1,\,z_2=0}}{Odds_{x_1=0,\,x_2=1,\,z_1=1,\,z_2=0}} = \frac{\exp(\beta_0+\beta_1+\beta_3+\beta_5)}{\exp(\beta_0+\beta_2+\beta_3+\beta_7)}$$

$$= \exp(\beta_1-\beta_2+\beta_5-\beta_7).$$

As razões de chances para comparar níveis de Z em cada nível de X podem ser encontradas da mesma maneira, se necessário.

Observe que se não houver interação, então os parâmetros que definem a interação entre X e Z, β_5, β_6, β_7 e β_8 são todos 0.

É então evidente na tabela acima que as razões de chances comparando os níveis de X são as mesmas em todos os níveis de Z e, portanto, precisam ser computadas apenas uma vez.

Exemplo 2.14: Controle do vírus da murcha do tomateiro.

A função **glm()** usa procedimentos computacionais que iteram até que a convergência seja alcançada ou até que o número máximo de iterações seja alcançado. O critério utilizado para determinar se há convergência é a mudança nos valores sucessivos da deviance residual, ao invés da mudança nas estimativas sucessivas dos parâmetros de regressão. Isso cria um critério único e compacto que equilibra equitativamente a convergência de todas as estimativas de parâmetros simultaneamente.

Se consideramos $G^{(k)}$ denotar o desvio residual na iteração k, então a convergência ocorre quando

$$\frac{|G^{(k)} - G^{(k-1)}|}{0.1 + |G^{(k)}|} < \epsilon,$$

onde ϵ é algum número pequeno especificado maior que 0.

O numerador $|G^{(k)} - G^{(k-1)}|$ dá uma medida geral de quanto

$$\widehat{\pi}_1, \cdots, \widehat{\pi}_n$$

mudam, assim, quanto $\widehat{\beta}_0, \cdots, \widehat{\beta}_p$ mudam da iteração anterior para a iteração atual. O denominador $0.1 + |G^{(k)}|$ converte isso aproximadamente em uma mudança proporcional.

A função **glm()** fornece algumas maneiras de controlar como a convergência é determinada. Primeiro, seu argumento **epsilon** permite que o usuário declare o valor de ϵ. O valor padrão é **epsilon** = 10^{-8}. Segundo, o argumento **maxit** indica o número máximo de iterações permitidas para o procedimento numérico, onde o padrão é **maxit = 25**. Terceiro, o valor do argumento **trace = TRUE** pode ser usado para ver os valores reais de $G^{(k)}$ para cada iteração. O padrão é **trace = FALSE**.

Exemplo 2.15: Placekicking.

Quando a convergência não ocorre, a primeira solução possível é tentar um número maior de iterações. Por exemplo, se as 25 iterações padrão não forem suficientes, tente 50 iterações. Se isso não funcionar, pode haver algum problema fundamental com os dados ou o modelo que está interferindo nos procedimentos numéricos iterativos.

O problema mais comum ocorre quando uma variável explicativa separa perfeitamente os dados entre os valores de $y_i = 0$ e 1; isso é muitas vezes referido como separação completa.

Além de uma mensagem de aviso de convergência, **glm()** também pode relatar **"glm.fit: fitted probabilities numerically 0 or 1 occurred"**; no entanto, esta afirmação por si só nem sempre é indicativa de problemas de ajuste de modelo. Ilustramos o problema de separação completa no próximo exemplo.

Exemplo 2.16: Separação completa.