Análise de Dados Categóricos
Regressão com resposta binária
Seções 2.2.2, 2.2.3 e 2.2.4

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Novembro, 2022

Testes de hipóteses podem ser usados para avaliar a importância de variáveis explicativas em um modelo. Por exemplo, um teste de $H_0 : \beta_r = 0$ vs. $H_1 : \beta_r \neq 0$ avalia o r-ésimo termo da variável explicativa no modelo

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + + \cdots + \beta_p x_p.$$

Se $\beta_r = 0$, isso significa que o termo correspondente é excluído do modelo. Se $\beta_r \neq 0$, significa que o termo correspondente está incluído no modelo.

De forma equivalente, podemos enunciar as hipóteses como

$$H_0 : logit(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{r-1} x_{r-1} + \beta_{r+1} x_{r+1} + \cdots + \beta_p x_p$$
$$H_1 : logit(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \cdots + \beta_p x_p.$$

Em geral, as variáveis explicativas no modelo na hipótese nula, também conhecido como modelo reduzido, devem estar todas no modelo de hipótese alternativa, também conhecido como modelo completo.

Usaremos um teste Wald ou um LRT para realizar esses testes. Agora nos concentramos nas especificidades da configuração de regressão logística.

Teste Wald

A estatística Wald

$$Z_0 = \frac{\widehat{\beta}_r}{\sqrt{\widehat{\text{Var}}(\widehat{\beta}_r)}}$$

é usada para testar $H_0 : \beta_r = 0$ vs. $H_1 : \beta_r \neq 0$.

Se a hipótese nula for verdadeira, Z_0 tem uma distribuição normal padrão aproximada para uma amostra grande e rejeitamos a hipótese nula se Z_0 tiver um valor observado incomum para esta distribuição. Definimos incomum por $|Z_0| > Z_{1-\alpha/2}$. O p-valor é $2P(Z > |Z_0|)$ onde Z tem uma distribuição normal padrão.

Mais de um parâmetro pode ser testado para igualdade a 0 na hipótese nula. No entanto, renunciamos à discussão sobre isso porque os procedimentos de inferência de Wald aqui geralmente encontram os mesmos tipos de problemas discutidos no Capítulo 1.

No contexto de testes de hipóteses, isso significa que a taxa de erro do tipo I declarada para um teste de hipóteses, ou seja α, não é a mesma que a taxa de erro real do tipo I.

O teste da razão de verossimilhanças (LRT) tipicamente tem um desempenho melhor do que o teste de Wald, então nos concentramos neste procedimento a seguir.

Teste da razão de verossimilhanças (LRT)

A estatística LRT pode ser escrita informalmente como

$$\Lambda = \frac{\text{Máximo da função de verossimilhança sob } H_0}{\text{Máximo da função de verossimilhança sob } H_0 \text{ ou } H_1}.$$

No contexto de testar a igualdade dos parâmetros de regressão a 0, o denominador na equação acima (Λ) é a função de verossimilhança avaliada nas MLEs para o modelo contendo todos os $p + 1$ parâmetros de regressão.

O numerador desta equação é a função de verossimilhança avaliada nas MLEs para o modelo que exclui aquelas variáveis cujos parâmetros de regressão são definidos como 0 na hipótese nula.

Por exemplo, para testar $H_0 : \beta_r = 0$ vs. $H_1 : \beta_r \neq 0$, β_r seria mantida igual a 0, o que significa que essa variável explicativa correspondente seria excluída do modelo de hipótese nula.

É claro que as estimativas para os parâmetros restantes não precisam ser as mesmas do modelo de hipóteses alternativas devido às diferenças entre os dois modelos.

Se q parâmetros de regressão forem definidos como 0 na hipótese nula e se isso for verdade, a estatística $-2\log(\Lambda)$ tem uma distribuição aproximada de χ^2_q para uma amostra grande e rejeitamos a hipótese nula se $-2\log(\Lambda)$ tem um valor observado incomumente grande para essa distribuição.

Por exemplo, se $\alpha = 0.05$ para o teste de

$$H_0 : \beta_r = 0 \quad \text{vs.} \quad H_1 : \beta_r \neq 0,$$

a região de rejeição é > 3.84, onde $\chi^2_{1,0.95} = 3.84$ é o quantil 0.95 a partir de uma distribuição qui-quadrado. O *p*-valor é

$$P(A > -2\log(\Lambda))$$

onde A tem uma distribuição χ^2_1 quando $q = 1$.

Na maioria das vezes, usaremos as funções genéricas **anova()** do pacote **stats** e **Anova()** do pacote **car** para realizar esses testes, funções comumente usadas por outros incluem a função **drop1()** do pacote **stats** e a função **lmtest()** do pacote **lmtest**.

Essas funções testam hipóteses de uma estrutura semelhante àquelas vistas em uma análise de variância (ANOVA), mas podem realizar LRTs em vez dos típicos testes *F* ANOVA.

As duas funções são baseadas em testes de comparação de modelos de diferentes tipos. Quando usado com um objeto de ajuste de modelo resultante de **glm()**, a função **anova()** calcula testes tipo 1 (sequenciais), enquanto **Anova()** calcula testes tipo 2 (parciais).

A diferença entre os tipos está nos modelos de hipótese nula usados para cada termo testado. Com os testes do tipo 2, cada modelo de hipótese nula consiste em todas as outras variáveis listadas no lado direito do argumento **formula**, ignorando quaisquer interações de ordem superior que contenham o termo.

Para testes do tipo 1, o modelo de hipótese nula contém apenas as variáveis listadas no argumento **formula** antes do termo testado.

Geralmente, os testes do tipo 2 são preferidos, a menos que haja algum motivo específico para considerar uma sequência de testes, como em modelos polinomiais. Ver Milliken and Johnson (2004) para mais detalhes.

A estatística LRT transformada $-2\log(\Lambda)$ tem uma forma simplificada. Suponha que a probabilidade estimada de sucesso nos modelos sob a hipótese nula e alternativa sejam denotados como $\widehat{\pi}_i^{(0)}$ e $\widehat{\pi}_i^{(1)}$, respectivamente. Da mesma forma, definimos um vetor das estimativas dos parâmetros da regressão como $\widehat{\beta}^{(0)}$ e $\widehat{\beta}^{(1)}$.

Temos então

$$\begin{aligned}
-2\log(\Lambda) &= -2\log\left(\frac{L\left(\widehat{\beta}^{(0)}|y_1,\cdots,y_n\right)}{L\left(\widehat{\beta}^{(1)}|y_1,\cdots,y_n\right)}\right) \\
&= -2\left(\log\left(L\left(\widehat{\beta}^{(0)}|y_1,\cdots,y_n\right)\right)\right. \\
&\quad \left.\log\left(L\left(\widehat{\beta}^{(1)}|y_1,\cdots,y_n\right)\right)\right) \\
&= -2\sum_{i=1}^{n} y_i\log\left(\widehat{\pi}_i^{(0)}\right)+(1-y_i)\log\left(1-\widehat{\pi}_i^{(0)}\right) \\
&\qquad -y_i\log\left(\widehat{\pi}_i^{(1)}\right)-(1-y_i)\log\left(\widehat{\pi}_i^{(1)}\right).
\end{aligned}$$

Exemplo: Placekicking.

Em um modelo linear onde $E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, o parâmetro de regressão β_r é interpretado como a mudança na resposta média para cada aumento de 1 unidade em x_r, mantendo as outras variáveis no modelo constantes.

Em um modelo de regressão logística, a interpretação dos parâmetros de regressão precisa levar em conta o fato de que eles estão relacionados à probabilidade de sucesso através de

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Mantendo as outras variáveis constantes, um aumento de 1 unidade em x_r faz com que $logit(\pi)$ mude por β_r. Essa mudança constante nas probabilidades logarítmicas de um sucesso leva a uma maneira conveniente de usar as probabilidades na interpretação.

Para ver como isso funciona, considere um modelo de regressão logística com apenas uma variável explicativa x. As chances de um sucesso em um determinado valor de x são

$$Odds_x = \exp(\beta_0 + \beta_1 x)\cdot$$

Se x for aumentado em $c > 0$ unidades, as chances de sucesso se tornam

$$Odds_{x+c} = \exp\left(\beta_0 + \beta_1(x + c)\right)\cdot$$

Para determinar o quanto as chances de sucesso mudaram por esse aumento de c unidades, encontramos a razão de chances:

$$OR = \frac{Odds_{x+c}}{Odds_x} = \frac{\exp\left(\beta_0 + \beta_1(x + c)\right)}{\exp\left(\beta_0 + \beta_1 x\right)} = \exp(c\,\beta_1)\cdot$$

Curiosamente, o valor original da variável explicativa x é cancelado na simplificação; apenas a quantidade de aumento c e o coeficiente β_1 importam. A interpretação padrão desta razão de chances é

> As chances de sucesso mudam em $\exp(c\,\beta_1)$ vezes para cada c unidades aumentadas em x.

Também é comum dizer "aumentar" ao invés de "mudar" quando $\exp(c\,\beta_1) > 1$ e "diminuir" quando $\exp(c\,\beta_1) < 1$.

Quando x é binário com codificação de 0 e 1, c é sempre 1. A interpretação da razão de chances ou odds ratio então compara as chances de sucesso no nível codificado "x = 1" com as chances em "x = 0".

As estimativas dos parâmetros do modelo podem ser substituídas por seus parâmetros correspondentes em *OR* para estimar a razão de chances. A razão de chances estimada torna-se

$$\widehat{OR} = \exp(c\,\widehat{\beta}_1),$$

e sua interpretação é que as chances estimadas de um sucesso mudam por $\exp(c\,\widehat{\beta}_1)$ vezes para cada c unidades de aumento em x.

Como a razão de chances estimada é uma estatística, ela varia de amostra para amostra. Portanto, precisamos encontrar um intervalo de confiança para *OR* a fim de fazer inferências com um determinado nível de confiança. Os procedimentos baseados em verossimilhanças fornecem a base para os intervalos de confiança discutidos a seguir.

Para encontrar um intervalo de confiança de Wald para *OR*, primeiro encontramos um intervalo de confiança para $c\,\beta_1$ e, em seguida, usamos a função exponencial com os pontos finais do intervalo.

Para este fim, extraímos $\widehat{\text{Var}}(\widehat{\beta}_1)$ da matriz de covariâncias estimada para os estimadores dos parâmetros e formamos o intervalo de confiança $(1-\alpha)100\%$ de Wald para $c\,\beta_1$ como

$$c\,\widehat{\beta}_1 \pm cZ_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\widehat{\beta}_1)},$$

onde utilizamos $\widehat{\text{Var}}(c\,\widehat{\beta}_1) = c^2\widehat{\text{Var}}(\widehat{\beta}_1)$.

O $(1-\alpha)100\%$ intervalo de confiança de Wald para *OR* torna-se

$$\exp\left(c\,\widehat{\beta}_1 \pm cZ_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_1)}\right).$$

O intervalo de confiança de Wald geralmente tem um nível de confiança verdadeiro próximo ao intervalo de confiança declarado somente quando há amostras grandes.

Quando o tamanho da amostra não é grande, os intervalos de confiança LR perfilada geralmente apresentam melhor desempenho. Para esta configuração, encontramos o conjunto de valores de β_1 tais que

$$-2\log\left(\frac{L(\widehat{\beta}_0,\beta_1\,|\,y_1,\cdots,y_n)}{L(\widehat{\beta}_0,\widehat{\beta}_1\,|\,y_1,\cdots,y_n)}\right) < \chi^2_{1,1,-\alpha}$$

seja satisfeito, onde $\widehat{\beta_0}$ é o MLE de β_0 seguindo a especificação de um valor de β_1.

Na maioria das configurações, não há soluções de forma fechada para os limites inferior e superior, portanto, são necessários procedimentos numéricos iterativos para encontrá-los.

Uma vez que os limites do intervalo de confiança para β_1 são encontrados, digamos, "lower" e "upper", usamos a função exponencial e levamos em consideração um valor de c para encontrar o intervalo de confiança LR perfilada $(1-\alpha)100\%$ para OR:

$$\exp(c \times lower) < OR < \exp(c \times upper).$$

Exemplo: Placekicking.

Uma vez que um modelo de regressão logística é estimado, é interessante estimar a probabilidade de sucesso para um conjunto de valores de variáveis explicativas.

Isso pode ser feito simplesmente substituindo as estimativas dos parâmetros no modelo:

$$\widehat{\pi} = \frac{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p\right)}{1 + \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p\right)}.$$

Como $\widehat{\pi}$ é uma estatística, ela varia de amostra para amostra. Portanto, precisamos encontrar um intervalo de confiança para fazer inferências com um determinado nível de confiança. Os intervalos Wald e LR perfilados serão discutidos nesta seção.

Para ajudar a explicar como o intervalo de Wald é calculado, considere novamente um modelo de regressão logística com apenas uma variável explicativa.

A probabilidade estimada de sucesso é então

$$\widehat{\pi} = \frac{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right)}{1 + \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right)}.$$

A distribuição normal é uma melhor aproximação para a distribuição de $\widehat{\beta}_0$ e $\widehat{\beta}_1$ do que para $\widehat{\pi}$, então procedemos de maneira semelhante a encontrar o intervalo de Wald para a razão de chances.

Primeiro encontramos um intervalo de confiança para o preditor linear, $\mathit{logit}(\pi) = \beta_0 + \beta_1 x$, e então transformamos as extremidades desse intervalo em um intervalo usando a transformação $\exp(\cdot)/\big(1 + \exp(\cdot)\big)$.

O intervalo de confiança $(1 - \alpha)100\%$ de Wald para $\beta_0 + \beta_1 x$ é

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm Z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}\big(\widehat{\beta}_0 + \widehat{\beta}_1 x\big)}.$$

A variância estimada é encontrada por

$$\widehat{\mathrm{Var}}\big(\widehat{\beta}_0 + \widehat{\beta}_1 x\big) = \widehat{\mathrm{Var}}\big(\widehat{\beta}_0\big) + x^2 \widehat{\mathrm{Var}}\big(\widehat{\beta}_1\big) + 2x\widehat{\mathrm{Cov}}\big(\widehat{\beta}_0, \widehat{\beta}_1\big),$$

onde cada variância e o termo de covariância estão disponíveis a partir da matriz de covariância estimada das estimativas de parâmetros.

Esta é outra aplicação do seguinte resultado:

$$\text{Var}(aU + bV) = a^2\text{Var}(U) + b^2\text{Var}(V) + 2ab\text{Cov}(U, V),$$

onde U e V são variáveis aleatórias e a e b são constantes.

Usando os limites de intervalo para $\beta_0 + \beta_1 x$, o intervalo de confiança $(1 - \alpha)100\%$ Wald para π é

$$\frac{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x \pm Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right)}\right)}{1 + \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x \pm Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right)}\right)}.$$

Observe que o limite inferior (superior) para π usa o sinal de menos (mais) na parte $\pm$ da equação acima.

Quando existem p variáveis explicativas no modelo, o intervalo de Wald para π é encontrado da mesma maneira.

O intervalo é

$$\frac{\exp\left(\widehat{\eta} \pm Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\widehat{\eta})}\right)}{1+\exp\left(\widehat{\eta} \pm Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\widehat{\eta})}\right)},$$

sendo $\widehat{\eta} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p$, onde a expressão de variância é encontrada de maneira semelhante ao caso de variável única:

$$\widehat{\text{Var}}(\widehat{\eta}) = \sum_{i=0}^{p} x_i^2 \widehat{\text{Var}}(\widehat{\beta}_i) + 2\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} x_i x_j \widehat{\text{Cov}}(\widehat{\beta}_i, \widehat{\beta}_j),$$

com $x_0 = 1$.

Embora essa expressão de variância possa ser longa, ela é calculada automaticamente pela função **predict()** conforme ilustrado no próximo exemplo.

Os intervalos LR perfilada também são calculados primeiro para $\mathit{logit}(\pi)$ e depois transformados em intervalos para π.

Por exemplo, em um modelo com uma variável explicativa,

$$\mathit{logit}(\pi) = \beta_0 + \beta_1 x,$$

é uma combinação linear de β_0 e β_1. O numerador de $-2\log(\Lambda)$ envolve maximizar a função de verossimilhança com uma restrição para esta combinação linear, que é mais difícil do que restringir apenas um parâmetro.

O problema se torna ainda mais complicado quando há várias variáveis explicativas e, em alguns casos, os procedimentos numéricos iterativos podem demorar muito para serem executados ou até mesmo não convergir.

Supondo que a computação seja bem-sucedida e um intervalo para $\beta_0 + \beta_1 x$ seja encontrado, realizamos a transformação

$$\exp(\cdot)/\big(1 + \exp(\cdot)\big),$$

para obter um intervalo para π.

Usamos o pacote **mcprofile**, um pacote de contribuição de usuários que não está na instalação padrão do R, para calcular os intervalos de verossimilhança perfilada para π e para a maioria das outras combinações lineares de parâmetros que serão discutidas aqui.

Usamos este pacote porque é o pacote mais geral disponível para calcular esses tipos de intervalos. Sugerimos o seguinte curso de ação ao calcular os intervalos de confiança LR perfilada com **mcprofile**:

- Calcule um intervalo de Wald.
- Calcule um intervalo LR perfilada com o pacote **mcprofile**.
- Use o intervalo LR perfilada, desde que não seja muito diferente do Wald e não haja mensagens de aviso fornecidas pelo R ao calcular o intervalo. Caso contrário, use o intervalo Wald.

Exemplo: Placekicking.