Análise de Dados Categóricos
Regressão com resposta binária
Seções 2.1, 2.2 e 2.2.1

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Setembro, 2023

A Seção 1.1 discute como estimar e fazer inferências sobre uma única probabilidade de sucesso π. A Seção 1.2 generaliza essa discussão para a situação de duas probabilidades de sucesso que agora dependem de um nível de grupo.

Agora completamos a generalização para uma situação em que há muitas possibilidades diferentes de sucesso para estimar e realizar inferências. Além disso, quantificamos como uma variável explicativa com muitos níveis possíveis, talvez contínuo em vez de categórico, afeta a probabilidade de sucesso.

Essas generalizações são feitas através do uso de modelos de regressão binária, também chamados de modelos de regressão binomial.

Vamos revisar os modelos de regressão linear normal. Seja Y_i a variável de resposta para observações $i = 1, \cdots, n$. Além disso, considere a situação com p variáveis explicativas $x_{i1}, \cdots, x_{ip}$ que são medidos na i-ésima observação.

Relacionamos as variáveis explicativas com a variável resposta através de um modelo linear:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

onde ϵ_i para $i = 1, \cdots, n$ são independentes e cada um tem uma distribuição de probabilidade normal com média 0 e variância σ^2.

Isso leva a cada Y_i para $i = 1, \cdots, n$ sendo independentes com distribuições normais.

Os $\beta_0, \cdots, \beta_p$ são parâmetros de regressão que quantificam as relações entre as variáveis explicativas e Y_i. Por exemplo, se $\beta_1 = 0$, isso indica que não existe uma relação linear entre a primeira variável explicativa e a variável de resposta dadas as outras variáveis do modelo.

Alternativamente, se $\beta_1 > 0$, existe uma relação positiva, um aumento na variável explicativa leva a um aumento na variável resposta e se $\beta_1 < 0$, existe uma relação negativa, um aumento na variável explicativa leva a uma diminuição na a variável de resposta.

Tomando a esperança de Y_i, também podemos escrever o modelo como

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

onde assumimos que as variáveis explicativas são condicionadas ou simplesmente constantes.

Essas equações do modelo permitem que diferentes valores possíveis de Y_i ou $E(Y_i)$ sejam funções de um conjunto de variáveis explicativas.

Assim, se tivermos $x_{i1}, \cdots, x_{ip}$ e de alguma forma conhecer todos os valores dos parâmetros $\beta_0, \cdots, \beta_p$, poderíamos encontrar o que Y_i seria em média.

Claro, não sabemos $\beta_0, \cdots, \beta_p$ em aplicações reais, mas podemos estimar esses parâmetros usando estimação de mínimos quadrados ou outros procedimentos mais complicados, se desejado.

No contexto de respostas binárias, a quantidade que queremos estimar é a probabilidade de sucesso, π.

Sejam Y_i variáveis de resposta binárias independentes para observações $i = 1, \cdots, n$, onde um valor de 1 denota um sucesso e um valor de 0 denota uma falha.

A distribuição normal obviamente não é mais um modelo apropriado para Y_i. Em vez disso, semelhante à Seção 1.1, uma distribuição de Bernoulli descreve Y_i muito bem, mas agora permitimos que o parâmetro de probabilidade de sucesso π_i seja diferente para cada observação.

Assim,

$$P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \qquad \text{para} \qquad y_i = 0 \quad \text{ou} \quad 1$$

é a função de probabilidade (PMF) para Y_i.

Para encontrar o estimador de máxima verossimilhança de π_i, a função de verossimilhança é

$$L(\pi_1, \cdots, \pi_n | y_1, \cdots, y_n) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

Observe que existem n parâmetros diferentes para as n observações diferentes. Isso significa que o modelo está saturado: há tantos parâmetros quanto observações.

As estimativas de parâmetros neste caso são as mesmas que os dados, $\widehat{\pi}_i = 0$ ou 1, então não há ganho trabalhando com este modelo.

Suponha, em vez disso, que temos $x_{i1}, \cdots, x_{ip}$ as variáveis explicativas como anteriormente, e desejamos relacionar os π_i's a essas variáveis.

Podemos propor uma função matemática, digamos

$$\pi_i = (x_{i1}, \cdots, x_{ip}),$$

para descrever a relação.

Por exemplo, em uma configuração simples, esta função pode designar dois valores possíveis de $\pi(x_{i1})$ dependendo de um único valor binário para x_{i1}, x_{i1} poderia denotar o número do grupo para a observação i como fizemos na Seção 1.2.

Em seguida, substituiríamos $\pi(x_{i1}, \cdots, x_{ip})$ na equação acima para estimar seus parâmetros.

Isso nos forneceria um modelo que pode ser usado para estimar uma probabilidade de sucesso em função de quaisquer valores possíveis para um conjunto de variáveis explicativas.

A próxima seção mostra que uma escolha apropriada da função $\pi(x_{i1}, \cdots, x_{ip})$ leva a uma gama completa de procedimentos úteis e convenientes de estimação e inferência.

O tipo mais simples de função para $\pi(x_{i1}, \cdots, x_{ip})$ é o modelo linear

$$\pi_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

No entanto, este modelo pode levar a valores de π_i menores que 0 ou maiores que 1, dependendo dos valores das variáveis explicativas e dos parâmetros de regressão.

Obviamente, isso é bastante indesejável ao estimar uma probabilidade. Felizmente, estão disponíveis muitas expressões não lineares que forçam π_i a ficar entre 0 e 1. A expressão mais comumente usada é o modelo de regressão logística:

$$\pi_i = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)}.$$

Observe que $\exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)$ é sempre positivo e o numerador da equação acima é menor que o denominador, o que leva a $0 < \pi_i < 1$.

O modelo de regressão logística também pode ser escrito como

$$\log\left(\frac{\pi}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

através de algumas manipulações algébricas.

O lado esquerdo da equação acima é o logaritmo natural para as chances de sucesso, que exploraremos ao interpretar o modelo. Essa transformação de π_i é frequentemente chamada de transformação logit, assim chamada em analogia ao modelo "probit" e é simplesmente denotada como $\mathit{logit}(\pi_i)$ ao escrever o lado esquerdo do modelo.

O lado direito da equação acima é uma combinação linear dos parâmetros de regressão com as variáveis explicativas, e isso é frequentemente chamado de preditor linear.

Frequentemente escreveremos os modelos sem os subscritos *i* como

$$\pi = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}$$

e

$$\mathit{logit}(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

quando não estamos nos referindo a observações particulares.

Ficará claro com base no contexto quando a probabilidade de sucesso for uma função de variáveis explicativas e não como π foi usado na Seção 1.1.1, mesmo que o símbolo π não as mencione explicitamente.

Gráfico do modelo de regressão logística.

O objetivo deste exemplo é examinar a forma do modelo de regressão logística quando há uma única variável explicativa x_1.

Considere o modelo

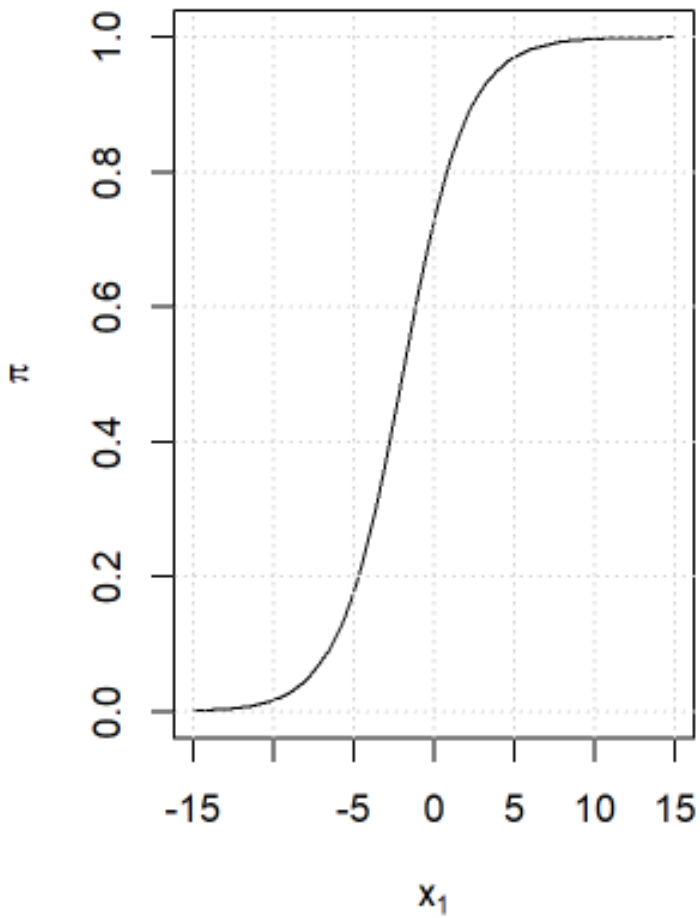
$$\pi = \frac{\exp(\beta_0 + \beta_1 x1)}{1 + \exp(\beta_0 + \beta_1 x_1)},$$

que é equivalentemente expresso como

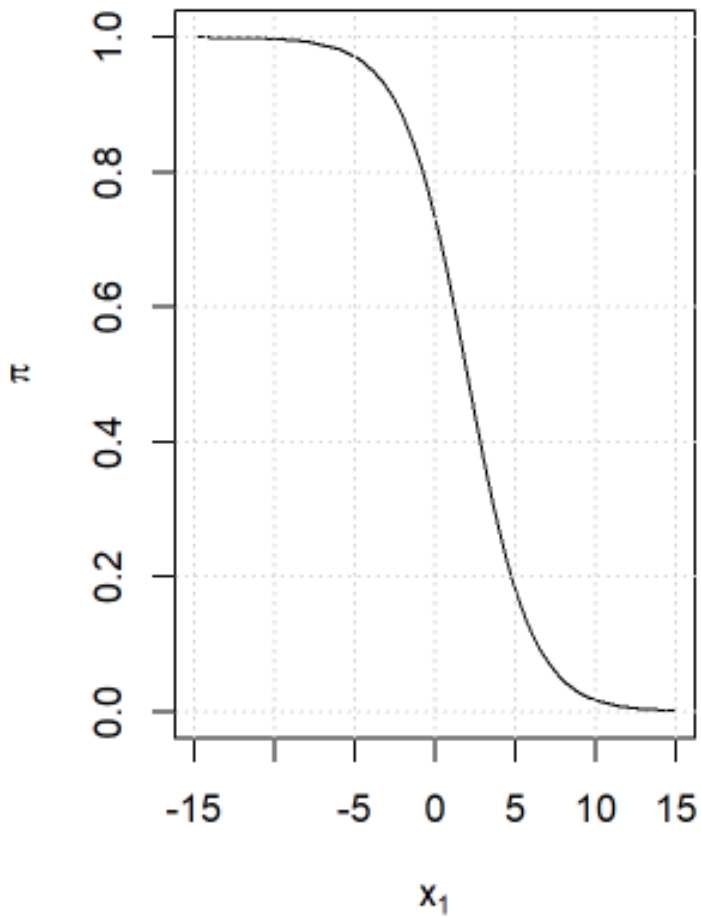
$$logit(\pi) = \beta_0 + \beta_1 x_1.$$

Suponha que $\beta_0 = 1$ e $\beta_1 = 0.5$. A figura abaixo mostra este modelo à esquerda. O gráfico à direita é o mesmo, mas com $\beta_1 = -0.5$.

$$\pi = \frac{e^{1+0.5x_1}}{1+e^{1+0.5x_1}}$$



$$\pi = \frac{e^{1-0.5x_1}}{1+e^{1-0.5x_1}}$$



Podemos fazer as seguintes generalizações examinando o modelo e esses gráficos:

- Quando $\beta_1 > 0$; existe uma relação positiva entre x_1 e π. Quando $\beta_1 < 0$; existe uma relação negativa entre x_1 e π.
- A forma da curva é um pouco semelhante à letra s, essa forma é chamada de "sigmoidal".
- A inclinação da curva depende do valor de x_1. Podemos mostrar isso matematicamente tomando a derivada de π em relação a x_1: $\partial\pi/\partial x_1 = \beta_1\pi(1-\pi)$.
- Acima $\pi = 0.5$ é uma imagem espelhada de abaixo $\pi = 0.5$.

A estimação por máxima verossimilhança é usada para estimar os parâmetros de regressão $\beta_0, \cdots, \beta_p$ do modelo de regressão logística. Substituindo nosso modelo por π_i e tomando o logaritmo natural para obter a função de log-verossimilhança:

$$\log L(\beta_0, \cdots, \beta_p | y_1, \cdots, y_n) = \log \left(\prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \right)$$

$$= \sum_{i=1}^{n} y_i \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \right)$$

$$- \log \left(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \right).$$

Derivando a expressão acima em relação a $\beta_0, \cdots, \beta_p$, iguale-os a 0 e resolva-os simultaneamente para obter as estimativas dos parâmetros $\widehat{\beta}_0, \cdots, \widehat{\beta}_p$.

Quando as estimativas dos parâmetros são incluídas no modelo, podemos obter a probabilidade estimada de sucesso como

$$\widehat{\pi} = \frac{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}\right)}{1 + \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}\right)}.$$

Infelizmente, existem apenas alguns casos simples em que as estimativas de parâmetros têm soluções de forma fechada; ou seja, geralmente não podemos escrever as estimativas dos parâmetros em termos dos dados observados como poderíamos para a estimativa de probabilidade única $\widehat{\pi}$ na Seção 1.1.2.

Em vez disso, usamos procedimentos numéricos iterativos para encontrar sucessivamente estimativas dos parâmetros de regressão que aumentam a função de log-verossimilhança.

Quando as estimativas mudam de forma insignificante para iterações sucessivas, isso sugere que atingimos o pico da função de log-verossimilhança e dizemos que elas convergiram.

Se as estimativas continuarem a mudar visivelmente até um número máximo de iterações selecionado, o procedimento numérico iterativo não convergiu e essas estimativas de parâmetros finais não devem ser usadas. Discutiremos convergência e não convergência com mais detalhes na Seção 2.2.7.

Dentro do R o algoritmo dos mínimos quadrados reponderados iterativamente (IRLS) é o procedimento numérico iterativo usado para encontrar as estimativas dos parâmetros.

Este procedimento utiliza o critério dos mínimos quadrados ponderados, que é comumente utilizado para modelos de regressão linear normal quando há variância não constante. O algoritmo IRLS alterna entre atualizar os pesos e as estimativas dos parâmetros de forma iterativa até que a convergência seja alcançada.

A função **glm()**, "glm" significa "modelo linear generalizado", dentro de R implementa este procedimento de estimativa de parâmetros e mostramos como usar esta função no próximo exemplo.

Exemplo: Placekicking.