Análise de Dados Categóricos
Regressão com resposta binária
Seções 2.3 e 2.3.1

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Setembro, 2023

Os modelos de regressão logística se enquadram em uma família de modelos chamados modelos lineares generalizados (GLMs). Cada modelo linear generalizado tem três partes diferentes:

- **Componente Aleatório**. A distribuição de Y. Na regressão logística, Y tem distribuição de Bernoulli ou binomial.
- **Componente Sistemático**. Especifica a combinação linear dos parâmetros de regressão com as variáveis explicativas,

$$\eta = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p.$$

- **Função de ligação**. Especifica como $\mathrm{E}(Y)$ está vinculado ao componente sistemático. Na regressão logística, temos

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

$\mathrm{E}(Y) = \pi$ e a transformação logit é a função de ligação.

Funções de ligação usadas com modelos de regressão binária

Outros modelos lineares generalizados também são usados para modelar respostas binárias. Esses modelos de regressão binária têm os mesmos componentes aleatórios e sistemáticos do modelo de regressão logística, mas suas funções de ligação são diferentes do logit.

O aspecto mais importante da função ligação nesses casos é que sua inversa deve garantir que $\mathrm{E}(Y)$ esteja entre 0 e 1. Por exemplo, vimos que

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

está sempre entre 0 e 1.

Essa garantia é obtida usando uma função de distribuição acumulada (CDF) como o inverso da função de ligação. De fato, o CDF de uma distribuição de probabilidade logística é usada para a regressão logística, o que resulta no nome do modelo.

Para revisar CDFs, suponha que X seja uma variável aleatória contínua com uma função de densidade de probabilidade (PDF) $f(x)$. O CDF $F(x)$ fornece a área sob o PDF à esquerda de x. Mais formalmente, o CDF de X é

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\mathrm{d}u,$$

$f(u)$ em vez de $f(x)$ é usado como o integrando para evitar confusão entre o que está sendo integrado e os limites da integração.

Como todas as probabilidades estão entre 0 e 1,

$$0 \leq F(x) \leq 1$$

para $-\infty < x < +\infty$.

Já usamos CDFs muitas vezes em R através do uso de funções como **pnorm()** e **pchisq()**. Por exemplo, **pnorm(q = 1.96)** pode ser expresso como $F(1.96)$ onde $F(\cdot)$ é a CDF de uma distribuição normal padrão. Equivalentemente, $Z_{0.975} = 1.96$.

Como o CDF sempre produz um valor entre 0 e 1, outros CDFs inversos são usados como funções de ligação para modelos de regressão binária. Embora a função de ligação logit seja a mais usada para regressão binária, existem duas outras que são comuns.

Regressão Probit — O CDF inverso de uma distribuição normal padrão é usado como função de ligação. Se denotarmos o CDF de uma distribuição normal padrão como $\Phi(\cdot)$, o modelo pode ser escrito como

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

ou

$$\Phi^{-1}(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Comumente, **probit( )** é usado para denotar $\phi^{-1}(\pi)$ levando ao modelo escrito como

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

Observe que **probit( )** é equivalente a Z_π, o quantil normal padrão na probabilidade π.

Regressão log-log complementar — A CDF inversa de uma distribuição Gumbel, também conhecida como distribuição de valores extremos é usada para formar esta função de ligação.

O CDF é

$$F(x) = \exp\Big(-\exp\big(-(x-\mu)/\sigma\big)\Big),$$

$-\infty < x < \infty$ e parâmetros $-\infty < \mu < \infty$, $\sigma > 0$.

Em vez de definir igual ao CDF como feito para os modelos de regressão logística e probit, o modelo log-log complementar π é igual a $1-F(x)$, que é essencialmente a probabilidade resultante de um complemento de um evento.

Podemos escrever nosso modelo então como

$$\pi = 1 - \exp\Big(- \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)\Big)$$

ou

$$\log\Big(- \log(1 - \pi)\Big) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Fazendo π igual a $1 - F(x)$, podemos manter as convenções padrão do que uma relação positiva e negativa significa em relação à variável de resposta. Se fosse igual a $F(x)$, o modelo teria uma interpretação não padronizada. Ou seja, aumentos em uma variável explicativa, digamos x_r, onde seu correspondente β_r é positivo, levaria a diminuições na probabilidade de sucesso. Por outro lado, diminuições em x_r com β_r negativo levariam a aumentos na probabilidade de sucesso.

Exemplo 2.21: Comparando três modelos de regressão binária

Estimação e inferência para modelos de regressão binária

Os modelos probit e log-log complementares são estimados da mesma forma que o modelo de regressão logística. A diferença é que π_i é representado na função de log-verossimilhança pela correspondente especificação do modelo probit ou log-log complementar.

Procedimentos numéricos iterativos são novamente necessários para encontrar as estimativas dos parâmetros. Uma vez encontradas as estimativas dos parâmetros, os mesmos procedimentos de inferência utilizados para o modelo de regressão logística estão disponíveis para os modelos probit e log-log complementares. Estes incluem os testes de Wald e LRTs da Seção 2.2.1 e os intervalos Wald e LR perfilada das Seções 2.2.3 e 2.2.4.

Por exemplo, podemos usar a função **anova()** para LRTs envolvendo parâmetros de regressão e a função **predict()** para intervalos de confiança de Wald para π. Uma das diferenças mais importantes entre os modelos de regressão logística, probit e log-log complementar surge ao calcular as razões de chance.

Na Seção 2.2.3, mostramos que a razão de chances para x no modelo $logit(\pi) = \beta_0 + \beta_1 x$ é $OR = e^{c\,\beta_1}$ para um aumento de c-unidades em x. Um aspecto muito importante dessa razão de chances é que ela é a mesma para qualquer valor de x. Infelizmente, isso não ocorre para modelos probit e log-log complementares.

Por exemplo, considere o modelo $probit(\pi) = \beta_0 + \beta_1 x$. As chances de sucesso são então

$$Odds_x = \frac{\Phi(\beta_0 + \beta_1 x)}{1 - \Phi(\beta_0 + \beta_1 x)},$$

em um determinado valor da variável explicativa x. Se x for aumentado em $c > 0$ unidades, as chances de sucesso se tornam

$$Odds_{x+c} = \frac{\Phi(\beta_0 + \beta_1 x + \beta_1 c)}{1 - \Phi(\beta_0 + \beta_1 x + \beta_1 c)}.$$

Quando a razão dessas duas chances é tomada, o resultado não tem uma forma fechada e depende do valor de x. As razões de chances do modelo log-log complementar também dependem de x. Essa é uma das principais razões pelas quais os modelos de regressão logística são mais usados entre os modelos de regressão binária.

Exemplo 2.22: Placekicking.