Trabalho apresentado no XXXVIII CNMAC, Campinas - SP, 2018.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Estudo e aplicação de Regressão Logística usando R

Mariella A. Bogoni[1]
Instituto de Ciências Exatas, UFF, Volta Redonda, RJ
Daiana G. de Menezes[2]
Instituto de Ciências Exatas, UFF, Volta Redonda, RJ
Marina S. D. de Freitas[3]
Instituto de Ciências Exatas, UFF, Volta Redonda, RJ

1 Introdução

Em muitos problemas, dada uma amostra de observações, busca-se relacionar uma variável, geralmente denotada por Y e chamada de variável dependente ou variável resposta, em função de outras variáveis, geralmente denotadas por x e chamadas de variáveis independentes ou preditoras. Os problemas que possuem variável resposta categórica são chamados de problemas de classificação. As categorias da variável predita são denominadas grupos de classificação. Existem muitas técnicas estatísticas para estudar tais problemas e neste trabalho pretende-se estudar um método chamado de regressão logística. Diferente da regressão tradicional, que usa um modelo descrito como $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$, ($n$ é o número de variáveis preditoras), esta técnica retorna as probabilidades de que uma dada observação pertença às categorias de uma variável categórica predita em vez de um valor numérico estimado. Essas probabilidades são estimadas através de uma função logística sobre os valores assumidos pelas variáveis preditoras. No caso mais simples, a variável predita Y é categórica binária, o método é chamado de regressão logística binomial e Y possui uma distribuição de Bernoulli, para a qual $Y = 1$ denota o sucesso e $Y = 0$ denota o fracasso. A interpretação de sucesso está associada a uma das categorias. No caso mais geral, chamado de regressão logística multinomial, a variável predita Y é categórica e assume múltiplas categorias e denota-se $Y = 1, 2, \ldots, J$, onde J são as classes. Neste caso, dada uma amostra de observações $(x_i, Y_i), i = 1, \ldots, n$, cada variável Y_i pode ser representada pelo conjunto $Y_{i1}, Y_{i2}, \ldots, Y_{iJ}$ de seus valores binários, onde $Y_{ij} = 1$ se $Y_i = j$ e $Y_{ij} = 0$ se $Y_i \neq j$. Para estimar o modelo, deve-se adotar uma categoria como referência e proceder a regressão das demais categorias em relação a essa. Para mais detalhes, consultar [2].

[1]mariellabogoni@id.uff.br
[2]daianagomes@id.uff.br
[3]msdias@id.uff.br

2 Regressão Logística

As respostas $Y_1, \ldots, Y_n$ são assumidas serem variáveis aleatórias independentes, não identicamente distribuídas mas cada uma com uma distribuição da mesma família, isto é, $Y_i \sim \text{Bernoulli}(\pi_i)$. Lembre-se de que $\text{E}Y_i = \pi_i = P(Y_i = 1)$. Na regressão logística, π_i é assumido estar relacionado com x_i por

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i. \tag{1}$$

O lado esquerdo é o logaritmo das chances de sucesso para Y_i. O modelo assume que este log de probabilidades ou transformações *logit* é uma função linear do preditor x.

A equação (1) pode ser reescrita como

$$\pi_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \tag{2}$$

Note que $0 < \pi(x) < 1$, o que é apropriado já que $\pi(x)$ é uma probabilidade. A derivada de $\pi(x)$ pode ser escrita como

$$\frac{d\pi(x)}{dx} = \beta\pi(x)(1-\pi(x)). \tag{3}$$

Como o termo $\pi(x)(1-\pi(x))$ é sempre positivo, a derivada de $\pi(x)$ é positiva, 0, ou negativa conforme β é positivo, 0 ou negativo. Se β é positivo, $\pi(x)$ é uma função estritamente crescente de x; se β é negativo, $\pi(x)$ é uma função estritamente decrescente de x; se $\beta = 0, \pi(x) = e^{\alpha}/(1+e^{\alpha})$ para todo x e não existe relação entre π e x.

Os parâmetros α e β têm significados similares àqueles da regressão linear simples. Se $x = 0$ em (1) então α representa o logaritmo da chance de sucesso em Y quando $x = 0$ e β é a mudança no logaritmo da chance de sucesso correspondendo a uma unidade de aumento em x, pois calculando (1) em x. Para estimar os parâmetros do modelo, o método mais usado é a máxima verossimilhança. Trata-se de um método estatístico usado para estimar parâmetros de uma função estatística de modo a maximizar a probabilidade de uma amostra.

A regressão logística possui inúmeras aplicações em diferentes áreas. Em instituições financeiras, em modelos de risco de crédito; em Medicina, permite determinar os fatores que caracterizam um grupo de indivíduos doentes em relação a indivíduos sãos. Outras aplicações podem ser encontradas em [1] e [3].

Referências

[1] Casella, G., Berger, R. L. *Statistical Inference*. Duxbury Resource Center, 2002.

[2] Loesch, C., Hoeltgebaum, M. *Métodos Estatísticos Multivariados*. Editora Saraiva, 2012.

[3] Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning; data mining, inference, and prediction*. Springer, 2001