Estudos em Regressão Logística

Projeto Supervisionado – MS777 – 2º Semestre 2011 – Prof. Alberto Saa

Diego Peterlevitz Frota - 070641

Introdução

O uso da regressão logística tem estado presente nas duas últimas décadas para estimar a probabilidade de eventos dicotômicos, com aplicações em economia, medicina, análise de risco e tomadas de decisão. Dado um conjunto de dados $y_1, y_2, y_3, \dots, y_m$, podemos citar como exemplos de respostas binárias:

- $y_i = 1 \Rightarrow paciente\ está\ doente, e\ y_i = 0 \Rightarrow paciente\ não\ está\ doente$
- $y_i = 1 \Rightarrow não\ pagamento\ de\ uma\ dívida, e\ y_i = 0 \Rightarrow pagamento$
- $y_i = 1 \Rightarrow fraude\ em\ operação\ financeira, e\ y_i = 0 \Rightarrow operação\ ok$

1. Primeiros Conceitos

Tenho um conjunto de dados independentes (x_i, y_i), $i = 1,2,3, \dots, m$, onde $y_i \in \{0,1\}$ e $x \in \mathbb{R}$. Diremos que y_i é a *variável resposta* (ou variável dependente), e x_i é a variável independente a qual pretende explicar o fenômeno descrito por y_i.

Estamos interessados em saber o valor da variável resposta Y dado um valor da variável independente x. Esse valor é a esperança condicional $E(Y|x)$.

Para tal, sendo $\beta_0, \beta_1 \in \mathbb{R}$, definiremos as funções reais:

$$E(Y|x) = \pi(x) := \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

$$g(x) := \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_i$$

Definimos $\pi(x)$ desta maneira pois uma função que retorne apenas valores entre 0 e 1, e de forma que possamos aproximar os valores dessa função com os valores da nossa base de dados, inputados os valores da variável dependente.

A função $g(x)$ é importante pois tem muitas propriedades importantes de um modelo de regressão linear. Portanto, vamos visitar a construção de um modelo linear e sua maneira de estimar o parâmetro $\beta = (\beta_0, \beta_1)$ para verificar relações entre a regressão linear e a regressão logística.

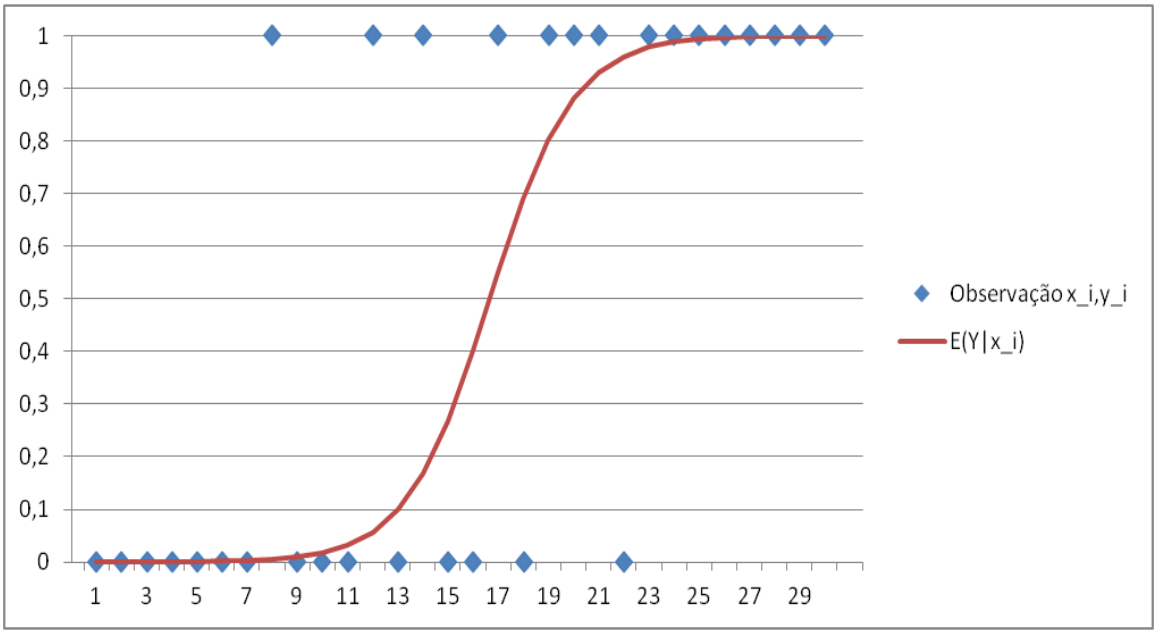

Gráfico 1: Curva dada pelo ajuste de π aos dados

2. Regressão Linear

Sejam os pares (x_i, y_i), $i = 1,2,3, \dots, m$ uma série de dados, $x_i, y_i \in \mathbb{R}$, com $\{ (x_i, y_i) \}_i$ linearmente independente.

Seja a matriz $X \in \mathbb{R}^{m \times n}$ tal que $X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_m \end{pmatrix}_{m \times n}$. Chamaremos de X_i as colunas de X.

Queremos, com o modelo de regressão linear:

$$\min_{\beta} \|X\beta - y\|_2^2$$

Para tal, basta que o resíduo da minimização seja ortogonal ao subespaço gerado pelas colunas de X, ou seja, $X\beta - y \perp X_i$, o que implica que $(X\beta - y)^T X_i = 0$.
Graficamente, podemos ver:

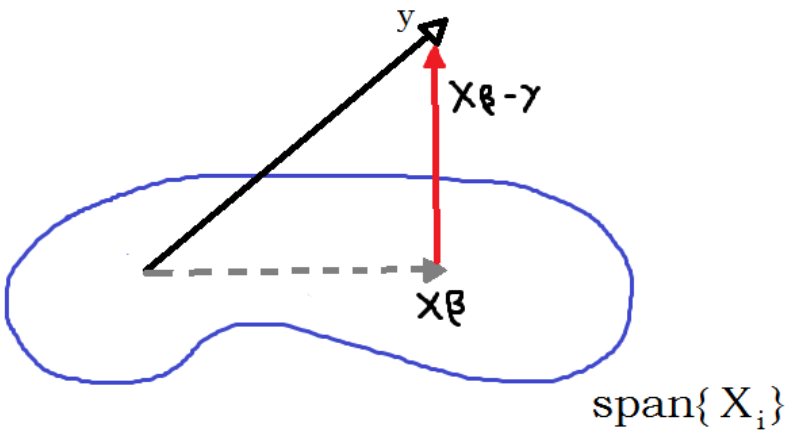

Figura 1: Ortogonalidade entre resíduo e X

Para uma prova, consulte [1].

Daí, temos:

$$X^T(X\beta - y) = 0$$

$$X^T X\beta - X^T y = 0$$

$$X^T X\beta = X^T y$$

$$\beta = (X^T X)^{-1} X^T y$$

A inversa $(X^T X)^{-1}$ sempre existe, pois vamos supor que as colunas de X são Linearmente Independentes.

3. Regressão Logística

Queremos encontrar os parâmetros que maximizam a probabilidade de obter os dados observados (x_i, y_i), $i = 1,2,3, \dots, m$. Utilizando $x = x_i$ temos que $\pi(x)$ definido em (1.) nos dá a probabilidade que Y seja igual a 1 dado x. Ou seja:

$$\pi(x) = P(Y = 1 \mid x)$$

Uma implicação simples é que:

$$1 - \pi(x) = P(Y = 0 \mid x)$$

Definiremos a contribuição de verossimilhança de (x_i, y_i):

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

Assim, definimos a função de verossimilhança (ou *likehood function*):

$$l(\beta) = \prod_i \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

Definimos também:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{m} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$

Queremos, com o objetivo de regressão logística:

$$\max_{\beta} L(\beta)$$

Daí:

$$\frac{\partial L}{\partial \beta_0} = \sum y_i \frac{1}{\pi(x_i)} \frac{\partial \pi(x_i)}{\partial \beta_0} + (1 - y_i) \frac{1}{1 - \pi(x_i)} \frac{\partial [1 - \pi(x_i)]}{\partial \beta_0}$$

Repare que:

$$\begin{aligned} \frac{\partial \pi(x_i)}{\partial \beta_0} &= e^{\beta_0 + \beta_0 x_i} \frac{1}{(1 + e^{\beta_0 + \beta_0 x_i})} + \frac{e^{\beta_0 + \beta_0 x_i}(-1)(e^{\beta_0 + \beta_0 x_i})}{(1 + e^{\beta_0 + \beta_0 x_i})^2} \\ &= \frac{e^{\beta_0 + \beta_0 x_i}}{(1 + e^{\beta_0 + \beta_0 x_i})} - \left[\frac{e^{\beta_0 + \beta_0 x_i}}{(1 + e^{\beta_0 + \beta_0 x_i}}\right]^2 = \pi(x_i) - \pi(x_i)^2 \\ &= \pi(x_i)[1 - \pi(x_i)] \end{aligned}$$

Então, ficamos:

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} &= \sum y_i \frac{1}{\pi(x_i)} \pi(x_i)[1 - \pi(x_i)] - (1 - y_i) \frac{1}{1 - \pi(x_i)} \pi(x_i)[1 - \pi(x_i)] \\ &= \sum y_i [1 - \pi(x_i)] - (1 - y_i)\pi(x_i) \\ &= \sum y_i [1 - \pi(x_i)] - (1 - y_i)\pi(x_i) = \sum y_i - \pi(x_i) \end{aligned}$$

Similarmente, calculamos $\frac{\partial L}{\partial \beta_0}$. Para maximizar $L(\beta)$ igualando suas derivadas parciais a 0, ficando com:

$$\sum_i y_i - \pi(x_i) = 0$$

$$\sum_i x_i [y_i - \pi(x_i)] = 0$$

Esse sistema, que nos fornece o $\beta = (\beta_0, \beta_1)$ que melhor ajusta o modelo no critério de máxima verossimilhança, não possui solução fechada e deve ser resolvido usando-se algum método iteratico. Maiores detalhes podem ser vistos em [2].

Análise de Erro

Após estimar os coeficientes, nossa principal análise do modelo é se as variáveis utilizadas são significantes. Este conceito pode ser resumido com uma pergunta: "o modelo que inclui determinada variável nos mais sobre a variável-resposta do que um modelo que não inclui tal variável?". A resposta para essa pergunta é respondido comparando as variáveis-resposta dos modelos com e sem a variável em questão.

Para entender tal comparação, vamos considerar a variável-resposta de um *modelo saturado*. Um modelo saturado é um modelo tal que temos tantos parâmetros quanto dados observados (em regressão linear, seria encontrar a reta que passa por dois pontos).

A comparação entre valores observados e previstos usando a likelihood function é baseada na seguinte equação:

$$D = -2\ ln\left[\frac{(likelihood\ do\ modelo\ estimado)}{(likelihood\ do\ modelo\ saturado)}\right]$$

Chamaremos D de *deviance*. No caso em que os valores da variável resposta são 0 ou 1, a likelihood do modelo saturado é 1. Isso segue da definição de modelo saturado, onde temos $\hat{\pi}(x_i) = y_i$. Assim:

$$D = -2\ln(likelihood\ do\ modelo\ estimado)$$

Podemos pensar em deviance nos mesmos termos que pensamos em resíduo da soma dos quadrados em regressão linear no contexto de teste de significância do modelo ajustado, considerando valores próximos de 0 como bons ajustes.

Para analisar a significância de uma variável, basta considerar o valor de D com e sem tal variável. Para tal, temos:

$$\begin{aligned} G &= D(modelo\ sem\ a\ variável) - D(modelo\ com\ a\ variável) \\ &= -2\ ln\left[\frac{(likelihood\ sem\ a\ variável)}{(likelihood\ com\ a\ variável)}\right] \end{aligned}$$

Considerando o caso em que temos apenas uma variável independente, a estatística G obedece uma distribuição Qui-Quadrado com 1 grau de liberdade (supondo apenas que temos uma amostra grande de dados). Assim, para quantizar a significância da variável, basta calcularmos o p-valor associado à $P[\mathcal{X}^2(1) > G]$, onde valores pequenos indicam uma boa significância.

4. Conclusões

Neste trabalho vimos uma boa abordagem para regressão quando a variável resposta possui informação binária, considerando semelhanças entre maximação de verossimilhança e minimização de erro quadrático. Deduzimos qual o sistema de equações que nos dá a solução ótima, e por fim analisamos o erro através da discussão de abordagens de comparação entre modelos.

5. Bibliografia

- [1] Stewart, G.W., Afternotes goes to Graduate School: Lecture on Advance Numerical Analysis. SIAM, 1998.
- [2] McCullagh, Peter; Nelder, John (1989). Generalized Linear Models, Second Edition. Boca Raton: Chapman and Hall/CRC
- [3] Hosmer, David W.; Stanley Lemeshow (2000). Applied Logistic Regression, 2nd ed.. New York; Chichester, Wiley