Análise de Regressão Logística

Ernesto F. L. Amaral
Magna M. Inácio

02 de setembro de 2010
Tópicos Especiais em Teoria e Análise Política:
Problema de Desenho e Análise Empírica (DCP 859B4)

Regressão Logística

O modelo de regressão não linear logístico é utilizado quando a variável resposta é qualitativa com dois resultados possíveis:

0 = Votou no mesmo candidato no 1º e 2º turnos.

1 = Mudou de voto entre 1º e 2º turnos (Helena-Lula e Alckmin-Lula)

Este modelo pode ser estendido quando a variável resposta qualitativa tem mais do que duas categorias. Por exemplo, posicionamento ideológico: esquerda, centro, direita.

Interpretação da função de resposta quando a variável resposta é binária

Vamos considerar o modelo de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \begin{cases} 1 \\ 0 \end{cases}$$

A resposta esperada é dada por:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Na regressão logística, Y_i possui uma distribuição de probabilidade:

$$Y_i = 1 \rightarrow P(Y_i = 1) = \pi_i$$

$$Y_i = 0 \rightarrow P(Y_i = 0) = 1 - \pi_i$$

Definição do valor esperado

Pela definição de valor esperado, obtemos:

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i$$

Assim, a resposta média, quando a variável resposta é uma variável binária (1 ou 0), representa a probabilidade de *Y* = 1, para o nível da variável independente X_i.

Função logística com uma variável independente

Considerações teóricas e práticas sugerem que quando a variável resposta é binária, a forma da função resposta será frequentemente curvilínea.

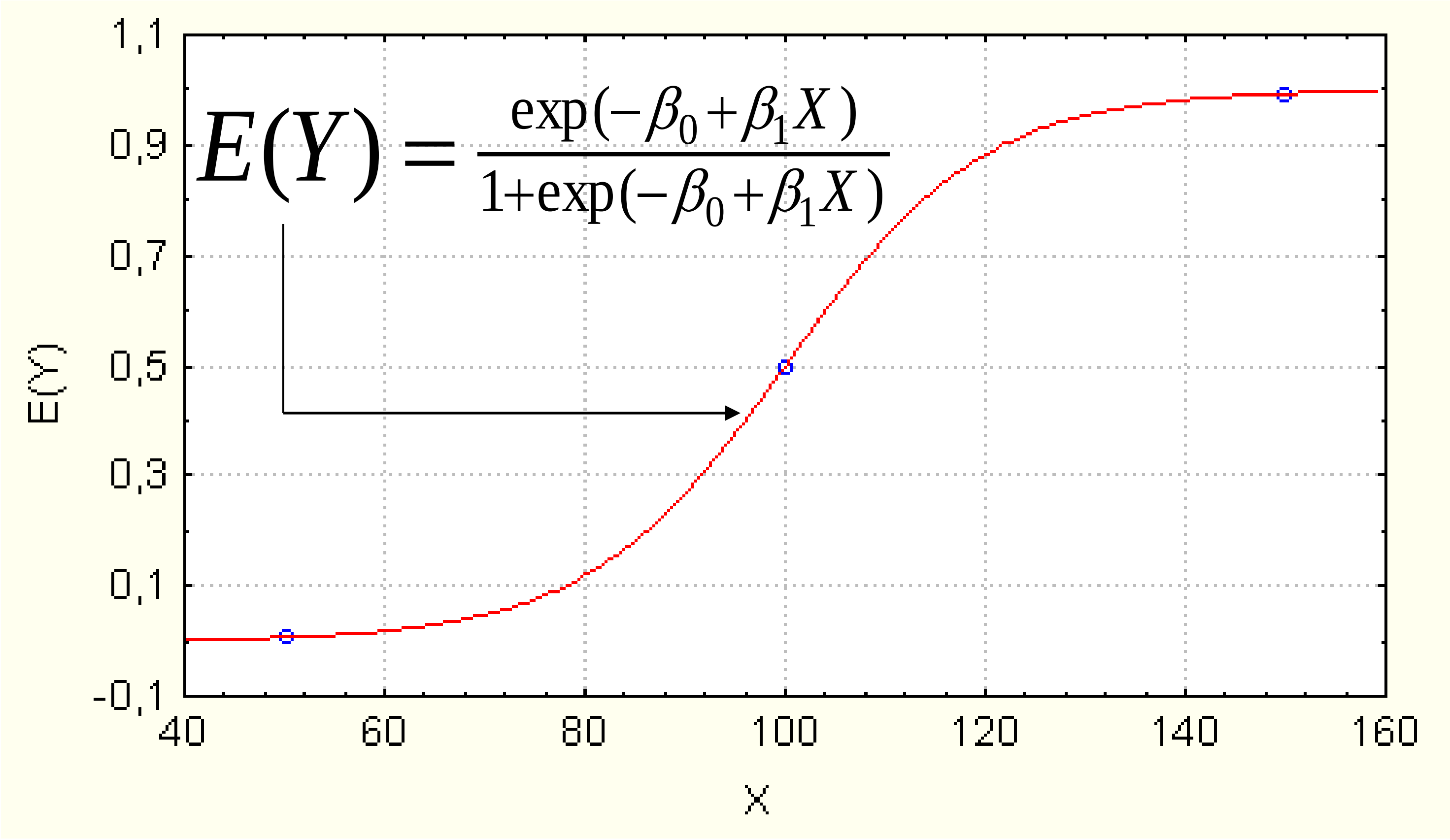
As funções respostas das figuras são denominadas funções logísticas, cuja expressão é:

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

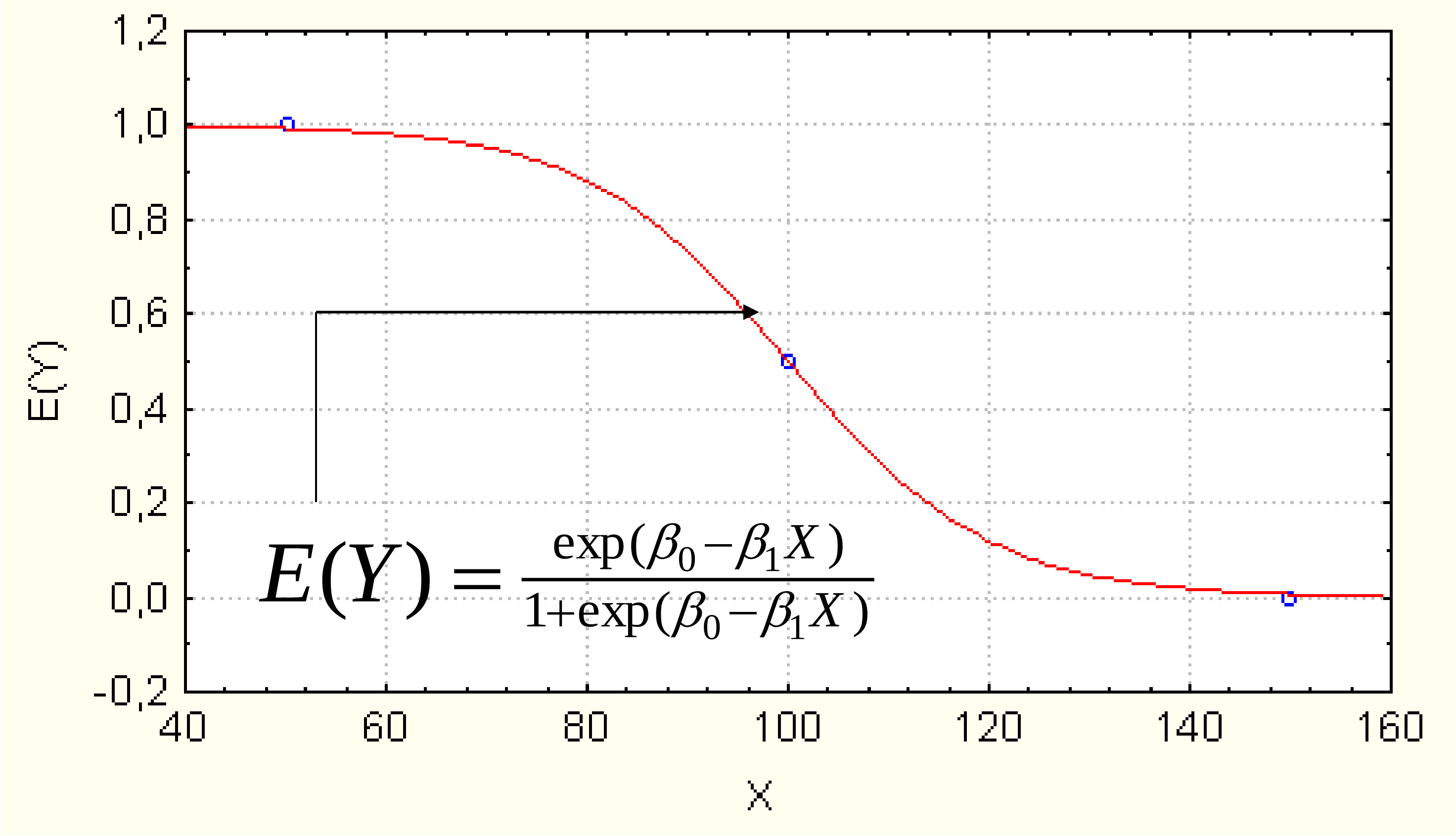
Forma equivalente:

$$E(Y) = \left[1 + \exp\left(-\beta_0 - \beta_1 X\right)\right]^{-1}$$

Variável dependente estimada pela variável independente observada



Variável dependente estimada pela variável independente observada



$$E(Y) = \frac{\exp(\beta_0 - \beta_1 X)}{1 + \exp(\beta_0 - \beta_1 X)}$$

Propriedade da função logística

Uma propriedade interessante é que a função logística pode ser linearizada, fazendo-se a transformação:

$$\pi' = \log_e\left(\frac{\pi}{1-\pi}\right) \quad \text{obtemos:} \quad \pi' = \beta_0 + \beta_1 X$$

Esta transformação é chamada de *transformação logit* da probabilidade π.

A razão *π/(1- π)* na transformação logit é chamada de ***Odds (Chance)****.*

A função resposta transformada é denominada como *função resposta logit,* e π' é denominada de *resposta média logit.*

Observe que: $-\infty \leq \pi' \leq \infty$ para $-\infty \leq X \leq \infty$.

Estimadores de máxima verossimilhança

Não existe uma solução analítica para os valores β_0 e β_1 que maximizam a função de verossimilhança.

Métodos numéricos são necessários para encontrar as estimativas de máxima verossimilhança, b_0 e b_1.

Encontradas as estimativas b_0 e b_1, substitui-se esses valores para encontrar os valores ajustados.

O valor ajustado para o *i-ésimo* valor é dado por:

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)}$$

Se usarmos a transformação *logit*, a função é:

$$\hat{\pi} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}$$

A função de resposta ajustada é dada por:

$$\hat{\pi}' = b_0 + b_1 X \qquad \text{onde:} \qquad \hat{\pi}' = \log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$$

Regressão logística com mais de uma variável independente

Função com uma variável independente:

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

Função com uma série de variáveis independentes:

$$E(Y) = \frac{\exp(\boldsymbol{\beta}^{'}\mathbf{X})}{1 + \exp(\boldsymbol{\beta}^{'}\mathbf{X})}$$

Uma forma equivalente é dada por:

$$E(Y) = (1 + \exp(-\boldsymbol{\beta}^{'}\mathbf{X}))^{-1}$$

A transformação *logit* resulta em:

$$\pi^{'} = \boldsymbol{\beta}^{'}\mathbf{X}$$

Ajustando o modelo

A função log-verossimlhança estende-se diretamente para o modelo de regressão logística múltipla, dada por:

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\boldsymbol{\beta}'\mathbf{X}_i) - \sum_{i=1}^{n} \log_e(1+\exp(\boldsymbol{\beta}'\mathbf{X}_i))$$

Métodos numéricos devem ser utilizados para encontrar os valores de β_0, β_1,..., β_{p-1} para maximizar a expressão.

As estimativas de máxima verossimilhança serão denotadas por b_0, b_1,...,b_{p-1}.

A função resposta logística ajustada e os valores ajustados são dados por:

$$\hat{\pi} = \frac{\exp(\mathbf{b}'\mathbf{X})}{1+\exp(\mathbf{b}'\mathbf{X})} = (1+\exp(-\mathbf{b}'\mathbf{X}))^{-1}$$

$$\hat{\pi}_i = \frac{\exp(\mathbf{b}'\mathbf{X}_i)}{1+\exp(\mathbf{b}'\mathbf{X}_i)} = (1+\exp(-\mathbf{b}'\mathbf{X}_i))^{-1}$$