

Aula 2.1: Variáveis Binárias

Econometria I - IBMEC

Marcus L. Nascimento

29 de outubro de 2025

1. Introdução
2. Estimação (Máxima Verossimilhança)
3. Teste de Razão de Verossimilhança
4. Capacidade Preditiva e Qualidade do Ajuste

Introdução

- Anteriormente, estudamos modelos de regressão linear (simples e múltipla) nos quais a variável resposta Y era contínua e estava definida no conjunto dos números reais.
- Continuamos interessados em aprender sobre Y a partir de um conjunto de variáveis independentes X_1, X_2, \dots, X_p , porém estudaremos o caso em que a variável dependente é **binária** (1/0, sim/não, sucesso/fracasso, entre outras):
 - Emprego: pessoa empregada (1 = sim, 0 = não);
 - Mercado formal: carteira assinada (1 = sim, 0 = não);
 - Programa social: pessoa beneficiária (1 = sim, 0 = não).

- No caso de variáveis binárias, estamos interessados em modelar a probabilidade relacionada à resposta:

$$E(Y = 1|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p) = P(Y = 1|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p).$$

- A relação entre a variável dependente e as variáveis independentes é descrita através da seguinte classe de modelos:

$$P(Y = 1|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p) = g(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p),$$

onde $g(\cdot)$ é uma função que assume valores entre 0 e 1, $0 < g(z) < 1$ para todo z definido nos reais.

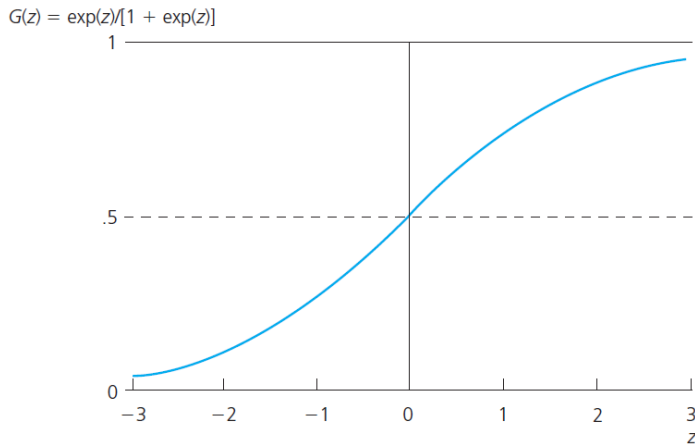
Introdução

- A função $g(\cdot)$ pode ser especificada de diferentes formas, porém estudaremos duas funções em particular: **modelo logit** e **modelo probit**.
- No modelo logit, $g(\cdot)$ é a função logística: $g(z) = \frac{\exp\{z\}}{1+\exp\{z\}}$.
 - Note que $g(\cdot)$ é a função de distribuição acumulada de uma variável aleatória logística padrão.
- No modelo probit, $g(\cdot)$ é a função de distribuição acumulada da normal padrão:

$$g(z) = \Phi(z) = \int_{-\infty}^z \phi(\nu) d\nu,$$

onde $\phi(z) = (2\pi)^{-1/2} \exp\{-z^2/2\}$ é a função densidade de probabilidade da normal padrão.

FIGURE 17.1 Graph of the logistic function $G(z) = \exp(z)/[1 + \exp(z)]$.



Introdução

- Devido à natureza não linear de $g(\cdot)$, a magnitude dos coeficientes β_j , $j = 1, 2, \dots, p$, não possui uma interpretação direta.
- Se X_j é uma variável contínua, seu efeito parcial em $p(\mathbf{x}) = P(Y = 1 | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p)$ é obtido através da derivada parcial:

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g'(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p) \beta_j, \text{ onde } g'(z) = \frac{dg(z)}{dz}.$$

- Como $g(\cdot)$ é a função de distribuição acumulada (fda) de uma variável aleatória contínua, $g'(\cdot)$ é uma função densidade de probabilidade (fdp).
- Nos casos logit e probit, $g(\cdot)$ é uma função estritamente crescente, logo $g'(\cdot) > 0 \Rightarrow$ os efeitos parciais terão o mesmo sinal que β_j .
- Os efeitos relativos de quaisquer duas variáveis independentes contínuas não dependem de \mathbf{x} : a razão dos efeitos parciais de X_{j_1} e X_{j_2} é $\beta_{j_1} / \beta_{j_2}$.

Estimação (Máxima Verossimilhança)

Estimação por Máxima Verossimilhança

- No caso em que a variável resposta é contínua e definida nos reais, aplicamos o método de mínimos quadrados ordinários para estimar os parâmetros do modelo.
 - Nenhuma hipótese sobre a distribuição condicional de Y dados X_1, X_2, \dots, X_p é necessária.
- Devido à natureza não-linear de $E(Y|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p)$, aplicar a estimação por máxima verossimilhança torna-se mais usual.
 - Método baseado na distribuição condicional de Y dados X_1, X_2, \dots, X_p .

Estimação por Máxima Verossimilhança

Suponha uma amostra aleatória de tamanho n . Para obtermos o estimador de máxima verossimilhança, é necessária a densidade Y_i dados $X_{i1}, X_{i2}, \dots, X_{ip}$:

$$\begin{aligned} f(y_i | x_{i1}, x_{i2}, \dots, x_{ip}) &= [g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})]^{y_i} \\ &\times [1 - g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})]^{(1-y_i)}, \quad y_i \in \{0, 1\}. \end{aligned}$$

A função **log-verossimilhança** para observação i é uma função dos parâmetros e dos dados (y_i, x_i) obtida ao tomarmos o logaritmo da equação anterior:

$$\begin{aligned} \ell_i(\beta_0, \beta_1, \dots, \beta_p) &= y_i \cdot \ln [g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})] \\ &+ (1 - y_i) \cdot \ln [1 - g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})] \end{aligned}$$

Estimação por Máxima Verossimilhança

A log-verossimilhança para uma amostra de tamanho n é, portanto, a soma em todas as observações:

$$\begin{aligned}\ell(\beta_0, \beta_1, \dots, \beta_p) &= \sum_{i=1}^n \ell_i(\beta_0, \beta_1, \dots, \beta_p) \\ &= \sum_{i=1}^n y_i \cdot \ln [g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})] \\ &\quad + \sum_{i=1}^n (1 - y_i) \cdot \ln [1 - g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})]\end{aligned}$$

O estimador de máxima verossimilhança de β_j , $j = 0, 1, \dots, p$, denotado por $\hat{\beta}_j$, é obtido através da maximização da log-verossimilhança.

Estimação por Máxima Verossimilhança

OBSERVAÇÕES:

- Devido à natureza não-linear do problema de maximização, não é possível obter fórmulas fechadas para os estimadores nos casos logit e probit.
 - Métodos numéricos são aplicados para encontrar as soluções do problema.
- Também, devido à natureza não linear do problema de maximização, a teoria estatística dos modelos logit e probit é mais difícil do que no método de mínimos quadrados ordinários.
- Sob condições bastante gerais, o estimador de máxima verossimilhança é
 - Consistente;
 - Assintoticamente normal;
 - Assintoticamente eficiente.
- A partir dos resultados de normalidade assintótica, é possível testar a hipótese $H_0 : \beta_j = 0$ de maneira usual utilizando uma estatística $t \hat{\beta}_j / \text{se}(\hat{\beta}_j)$.

Estimação por Máxima Verossimilhança

- Através de $\hat{\beta}_j$, $j = 1, 2, \dots, p$, podemos estimar os efeitos de X_j em $p(\mathbf{x}) = P(Y = 1 | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p)$. Se X_j é uma variável contínua, temos:

$$\Delta \hat{p}(\mathbf{x}) \approx [g'(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p) \hat{\beta}_j] \Delta x_j$$

para “pequenas” variações em x_j .

- Para $\Delta x_j = 1$, temos que a variação na probabilidade de sucesso estimada é $g'(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p) \hat{\beta}_j$.
- Sumarizar os efeitos parciais, no entanto, pode ser complicado considerando que $g'(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p) \hat{\beta}_j$ depende de todas as variáveis independentes.

Estimação por Máxima Verossimilhança

- Uma alternativa para obter, de forma resumida, as magnitudes dos efeitos parciais consiste em considerar uma especificação única para $g'(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p)$ e multiplicá-la por cada $\hat{\beta}_j$.
- Neste caso, a maneira mais comum é substituir cada variável independente por sua média amostral:

$$g'(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \dots + \hat{\beta}_p \bar{x}_p). \quad (1)$$

- Ao multiplicarmos (1) por $\hat{\beta}_j$, obtemos o efeito parcial estimado de X_j na pessoa “média” da amostra;
- Efeito parcial na média (PEA).

Estimação por Máxima Verossimilhança

- Há dois problemas potenciais em utilizar PEAs para resumir os efeitos parciais das variáveis independentes:
 - Nos casos em que a variável independente é discreta, a média não representa ninguém na amostra (ou na população);
 - Nos casos em que uma transformação não-linear é realizada em uma variável contínua, não é claro se aplicamos a média na função não-linear ou a função não-linear na média.
 - Por exemplo, $\overline{\log(\text{vendas})}$ ou $\log(\overline{\text{vendas}})$.
- Alternativamente, podemos calcular o efeito parcial médio (APE):

$$\left[n^{-1} \sum_{i=1}^n g'(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}) \right] \hat{\beta}_j.$$

Teste de Razão de Verossimilhança

Teste de Razão de Verossimilhança

- Nos casos em que é possível estimar os modelos probit e logit com e sem restrição, o **teste de razão de verossimilhança** pode ser aplicado para testar hipóteses acerca de um grupo de variáveis:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \dots = \beta_p = 0$$

$$H_1 : H_0 \text{ não é verdadeira,}$$

onde $q < p$ é o tamanho do grupo a ser testado.

- Como o estimador de máxima verossimilhança maximiza a função log-verossimilhança, o teste se baseia no fato de que retirar variáveis geralmente diminui (ou pelo menos mantém) o valor da log-verossimilhança.
 - A questão é se a diminuição na log-verossimilhança é grande o suficiente para concluir se o grupo de variáveis é relevante.

Teste de Razão de Verossimilhança

- A estatística de razão de verossimilhança é dada por:

$$LR = 2[\ell_{ur} - \ell_r]$$

onde ℓ_{ur} é o valor da log-verossimilhança para o modelo irrestrito e ℓ_r é o valor da log-verossimilhança para o modelo restrito.

- Função log-verossimilhança é sempre um número negativo;
- LR é não negativa e usualmente estritamente positiva ($\ell_{ur} > \ell_r$);
- A multiplicação por 2 é necessária para que LR possua aproximadamente uma distribuição χ^2 . Se estamos testando a exclusão de q variáveis, LR terá distribuição χ^2_q .

Capacidade Preditiva e Qualidade do Ajuste

Capacidade Preditiva

- Diremos que um preditor binário para y_i assume valor 1 se $g(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip})$ é pelo menos 0,5 e 0 caso contrário:

$$\tilde{y}_i = \begin{cases} 1, & \text{se } g(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}) \geq 0,5 \\ 0, & \text{se } g(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}) < 0,5 \end{cases}$$

- Considerando $\{\tilde{y}_i : i = 1, 2, \dots, n\}$ podemos verificar a capacidade preditiva do modelo entre as observações y_i , $i = 1, 2, \dots, n$.
- Para cada par (y_i, \tilde{y}_i) há quatro desfechos possíveis.
 - Quando ambos são 0 ou ambos são 1, temos o modelo prevendo corretamente.

Valores Reais

		Positivo (1)	Negativo (0)
Valores Previstos	Positivo (1)	TP (Verdadeiro Positivo)	FP (Falso Positivo)
	Negativo (0)	FN (Falso Negativo)	TN (Verdadeiro Negativo)

- Taxa (%) de classificações de corretos:
 - Sensibilidade (S): $S = p(\tilde{y} = 1|y = 1) = \frac{VP}{VP+FN}$;
 - Especificidade (E): $E = p(\tilde{y} = 0|y = 0) = \frac{VN}{FP+VN}$.
- Avaliação:
 - $S \geq 80\%$ e $E \geq 80\%$: alta capacidade preditiva;
 - $50\% < S \leq 80\%$ e $50\% < E \leq 80\%$: média capacidade preditiva;
 - $S \leq 50\%$ e $E \leq 50\%$: baixa capacidade preditiva.

Qualidade do Ajuste (Pseudo- R^2)

- Há diferentes alternativas de Pseudo- R^2 para os casos nos quais a variável dependente é binária. Em particular veremos o Pseudo- R^2 de McFadden:

$$\begin{aligned} R_{MF}^2 &= 1 - \frac{\ell_{ur}}{\ell_0} \\ &= \frac{\ell_0 - \ell_{ur}}{\ell_0}, \end{aligned}$$

onde ℓ_{ur} é o valor da log-verossimilhança para o modelo irrestrito e ℓ_0 é o valor da log-verossimilhança para o modelo com o intercepto apenas.

OBSERVAÇÕES:

- $0 \leq R_{MF}^2 < 1$;
- Ganho de informação estimada pelo modelo completo.