Ibmec

Variável dependente binária: modelo de probabilidade linear

IBM0288 - 2026.2

Prof. Raphael Gouvea

1 Para reflexão



Domenico Ferraro • 2º
Associate Professor @ Arizona State University | PhD in E...
1 m •

+ Seguir

Love this line from Wold (1954), "Causality and Econometrics":

"Causal relation and causality are theoretical concepts, not empirical."

A powerful reminder that data alone cannot deliver causal conclusions — theory must guide identification and interpretation.

Exibir tradução

WOLD (1954)

▶ "Causality and Econometrics" (cont'd):

Causal relation and causality are theoretical concepts, not empirical. This is obvious in the nonexperimental case; and in the experimental case the typical situation is one in which the causal relation enters as a hypothesis to be tested or demonstrated. Logically, the concept is therefore the same in the two cases; and in distinguishing between them our intention is mainly to stress that the test of the causal hypothesis is more direct and indisputable in the experimental case. We note further that our definition conforms to the usage of causality and related words in ordinary language. The controlled experiment is, in fact, nothing but a systematic way of collecting experience, and from a logical point of view the causal notions involved are of the same type as those which are all-pervading even in the earliest stages of prescientific experience, for example, "the water will boil if you put it over the fire" or "the crop will be larger if you manure the soil."

Ibmec

2 Empréstimos bancários e raça

- Duas pessoas idênticas, exceto pela raça, entram em um banco para solicitar uma hipoteca a fim de comprar casas idênticas.
- Por lei, ambas devem receber tratamento igual no processo de solicitação.
- A questão central é: o banco realmente as trata da mesma forma? As chances de aprovação são iguais para ambas?
- Essa é uma grande preocupação para os reguladores bancários: saber se a igualdade de tratamento ocorre **na prática**, e não apenas **no papel**.

3 Discriminação ou outros motivos?

- Empréstimos são concedidos ou negados por diversas razões legítimas: por exemplo, o grau de comprometimento da renda mensal do solicitante.
- Erros honestos também ocorrem: a negação do empréstimo a um único solicitante de um grupo minoritário não comprova discriminação!
- Por isso, muitos estudos buscam **evidências estatísticas** de discriminação.

4 Como avaliar causalidade?

- Uma abordagem inicial é comparar a **taxa de negação** entre solicitantes brancos e minoritários.
- Exemplo: pedidos individuais de hipotecas para imóveis unifamiliares realizados em 1990 em Boston:
  - **28%** dos solicitantes negros tiveram a hipoteca negada
  - Apenas **9%** dos solicitantes brancos tiveram a hipoteca negada
- No entanto, essa comparação não responde à pergunta central, pois os solicitantes negros e brancos **não eram necessariamente idênticos**, exceto pela raça.
- É necessário, portanto, um método que compare as taxas de negação **mantendo constantes as demais características** dos solicitantes.

5 Variáveis Dependentes Binárias: o que muda?

Mas não usamos variáveis dummy anteriormente?

Sim, mas elas eram variáveis explicativas e não de resultado. Até agora, a variável dependente (Y) era contínua: nota média do distrito escolar; taxa de mortalidade no trânsito.

E se Y for binária?

- Y = entrar na faculdade ou não; X = notas no ensino médio, pontuação no SAT, variáveis demográficas
- Y = a pessoa fuma ou não; X = imposto sobre cigarro, renda, variáveis demográficas
- Y = pedido de hipoteca aprovado ou não; X = raça, renda, características do imóvel, estado civil

6 Revisão Estatística

Seja Y uma variável aleatória discreta que assume os valores 0 e 1.

- Y é chamada de variável aleatória de Bernoulli.

$$Y = \begin{cases} 1, \text{com probabilidade } & p, \\ 0, \text{com probabilidade } & 1-p, \end{cases}$$

Valor Esperado e Variância da Bernoulli

O valor esperado e a variância da distribuição Bernoulli são:

$$E(Y) = p \qquad \text{e} \qquad \text{Var}(Y) = p(1-p)$$

A variância depende de p: é máxima em $p = 0,5$ e zero nos extremos.

7 O que acontece com a regressão?

- Nos modelos de MQO que estudamos anteriormente, a função de regressão populacional é dada por:

$$E(Y \mid X_1, X_2, ..., X_k)$$

- Quando $Y_i \in \{0, 1\}$, o valor esperado de Y é p, ou seja, a probabilidade de $Y = 1$:

$$E(Y) = 0 \times Pr(Y = 0) + 1 \times Pr(Y = 1) = Pr(Y = 1)$$

- Portanto, para uma variável binária Y, temos:

$$E(Y \mid X_1, X_2, ..., X_k) = Pr(Y = 1 \mid X_1, X_2, ..., X_k)$$

8 O Modelo de Probabilidade Linear

Um ponto de partida natural é o modelo de regressão linear (*Linear Probability Model - LPM*) com um único regressor:

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

- O que a reta $\beta_0 + \beta_1 X$ significa quando Y é binária?
- O que β_1 significa quando Y é binária? $\beta_1 = \frac{\Delta Y}{\Delta X}$?
- O que o valor predito $\hat{Y}$ significa quando Y é binária?

9 O que a reta significa?

Modelo de Probabilidade Linear: $Y_i = \beta_0 + \beta_1 X_i + u_i$

Quando Y é binária:

$$E(Y|X) = 1 \cdot \Pr(Y = 1|X) + 0 \cdot \Pr(Y = 0|X) = \Pr(Y = 1|X)$$

Sob a hipótese $E(u_i|X_i) = 0$:

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

e assim:

$$\Pr(Y = 1|X) = \beta_0 + \beta_1 X_i$$

10 O que o valor predito significa?

No LPM, o valor predito de Y é interpretado como a probabilidade predita de que $Y = 1$:

$$\hat{Y}_i = \widehat{\Pr}(Y = 1|X) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

11 O que β_1 significa?

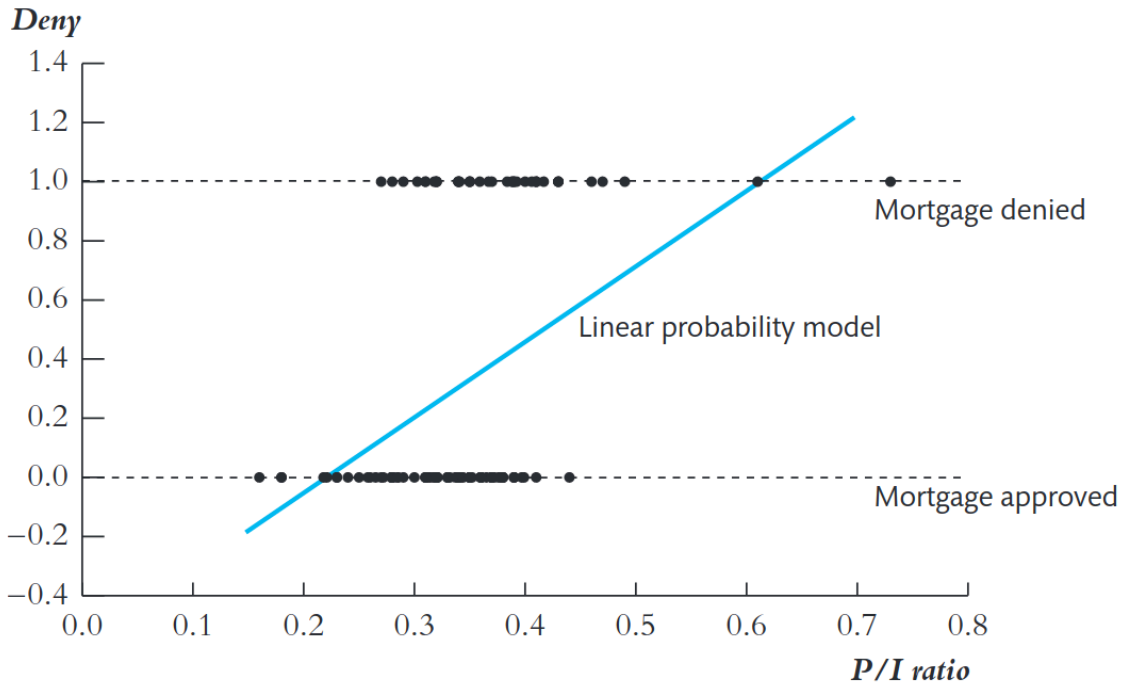
β_1 é a mudança nessa probabilidade para uma variação em X:

$$\beta_1 = \frac{\Pr(Y = 1|X = x + \Delta x) - \Pr(Y = 1|X = x)}{\Delta x}$$

> Interpretação
>
> Como β_1 é a mudança de uma probabilidade, os resultados devem ser interpretados como variação em pontos percentuais (p. p.). Ou seja, dado um aumento de $\Delta x = 1$, a $\Pr(Y = 1|X)$ varia β_1 pontos percentuais.

12 O caso das hipotecas

- **Contexto:** o agente bancário avalia se o solicitante conseguirá pagar o empréstimo.
- **Indicador-chave:** a razão **P/I** = pagamento mensal / renda mensal.
- **Variável dependente:** `deny = 1` se a solicitação foi negada, `0` se aprovada.
- **Padrão observado:**
  - Quando **P/I < 0,3**, quase nenhum pedido é negado.
  - Quando **P/I > 0,4**, a maioria dos pedidos é negada.



Obs.: gráfico com apenas 127 observações para facilitar a visualização.

13 Resultados

$$\widehat{deny} = \underset{(0.032)}{-0.080} + \underset{(0.098)}{0.604}\, P/I\ ratio. \qquad (11.1$$

Valor predito para a razão P/I = 0,3:

$$\hat{Y} = -0,080 + 0,604 \times 0,3 = 0,101$$

Calculando efeitos: aumento da razão P/I de 0,3 para 0,4:

$$\Delta\hat{Y} = 0,604 \times (0,4 - 0,3) = 0,061$$

> O efeito de aumentar a razão P/I de 0,3 para 0,4 é elevar a probabilidade de negação em **0,061** (ou seja, **6,1 pontos percentuais**).

14 Evidências de discriminação?

$$\widehat{deny} = \underset{(0.029)}{-0.091} + \underset{(0.089)}{0.559}\, P/I\ ratio + \underset{(0.025)}{0.177}\, black. \qquad (11.3)$$

- Probabilidade predita de negação para um candidato **negro** com P/I = 0,3:

$$Pr(deny = 1) = -0,091 + 0,559 \times 0,3 + 0,177 \times 1 = 0,254$$

- Probabilidade predita de negação para um candidato **branco** com P/I = 0,3:

$$Pr(deny = 1) = -0,091 + 0,559 \times 0,3 = 0,077$$

- **Diferença = 0,177 = 17,7 pontos percentuais**
- O coeficiente de `negro` é significativo ao nível de 5%.

Qual sua conclusão? Existe evidência de discriminação racial nesse mercado?

15 Resumo LPM

Vantagens:

- Simples de estimar e interpretar
- A inferência é a mesma da regressão múltipla (é necessário usar erros-padrão robustos à heterocedasticidade)
- Mesmas hipóteses em relação à causalidade e aos efeitos de tratamento

Desvantagens:

- O LPM supõe que a mudança na probabilidade predita é a mesma para todos os valores de X (linearidade!)
- Probabilidades preditas pelo LPM podem ser < 0 ou > 1!

Essas desvantagens podem ser superadas por um **modelo de probabilidade não linear**: Probit e Logit (próxima aula).