Aula Prática - Variável Dependente Limitada

Introdução

Nesta aula prática, replicaremos os principais resultados do Capítulo 11 de Stock e Watson (2020), que trata de modelos de regressão com variável dependente limitada A aplicação empírica central explora a seguinte pergunta: **existe discriminação racial no mercado de crédito imobiliário nos Estados Unidos?**

Para responder a essa pergunta, utilizamos os dados do **HMDA (Home Mortgage Disclosure Act)**, coletados pelo Federal Reserve de Boston. A base contém 2.380 observações de solicitações individuais de hipotecas para residências unifamiliares feitas em 1990 na Grande Boston. A variável dependente é `deny`, uma variável binária igual a 1 se a solicitação foi negada e 0 caso contrário.

O conjunto de dados possui várias informações financeiras e características dos solicitantes, o que nos permite verificar se, após controlar por esses fatores, ainda existe evidência de discriminação racial.

Pacotes e Dados

Os pacotes utilizados nesta aula são:

A base de dados HMDA está disponível diretamente no pacote `AER`. Após carregá-la, transformamos as variáveis categóricas em variáveis numéricas binárias e criamos indicadores para as categorias da razão empréstimo/valor. Para ver a definição de cada variável, consulte a página de documentação do pacote no R.[1][2]

```
Rows: 2,380
Columns: 22
$ deny       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,…
$ pirat      <dbl> 0.221, 0.265, 0.372, 0.320, 0.360, 0.240, 0.350, 0.280, 0.3…
$ hirat      <dbl> 0.221, 0.265, 0.248, 0.250, 0.350, 0.170, 0.290, 0.220, 0.2…
$ lvrat      <dbl> 0.8000000, 0.9218750, 0.9203980, 0.8604651, 0.6000000, 0.51…
$ chist      <dbl> 5, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 1, 1, 1, 1, 2, 1, 2, 2, 2,…
$ mhist      <dbl> 2, 2, 2, 2, 1, 1, 2, 2, 2, 1, 2, 2, 2, 1, 1, 1, 1, 1, 2, 2,…
$ phist      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ unemp      <dbl> 3.9, 3.2, 3.2, 4.3, 3.2, 3.9, 3.9, 1.8, 3.1, 3.9, 3.1, 4.3,…
$ selfemp    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ insurance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,…
$ condomin   <fct> no, no, no, no, no, no, yes, no, no, no, yes, no, no, no, n…
$ afam       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ single     <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1,…
$ hschool    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ lvrat_med  <dbl> 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,…
$ lvrat_high <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
```

[1] No pacote `AER`, as variáveis `deny`, `afam`, `phist`, `insurance`, `selfemp`, `single` e `hschool` são fatores com níveis `"no"` e `"yes"`. As variáveis `chist` e `mhist` são fatores ordenados. A conversão abaixo as transforma em variáveis numéricas 0/1 para uso nos modelos de regressão.

[2] Os dados também pode ser obtidos no site do livro. Para isso, faça download da base hmda e leia o dicionário de variáveis. Note que neste caso, as variáveis tem que ser construídas a partir do dicionário do HMDA. Por isso, para facilitar, vamos utilizar a versão disponível no pacote `AER`.

```
$ chist_3    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ chist_4    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ chist_5    <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ chist_6    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ mhist_3    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ mhist_4    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
```

Modelo de Probabilidade Linear

Formulação

O **Modelo de Probabilidade Linear (LPM)** aplica o estimador de MQO diretamente à variável binária `deny`. Como $E(Y \mid X) = \Pr(Y = 1 \mid X)$ quando Y é binária, o valor previsto pelo LPM é interpretado como a probabilidade condicional de negação:

$$\Pr(\text{deny} = 1 \mid X) = \beta_0 + \beta_1 \cdot \text{pirat}$$

Utilizamos erros-padrão robustos à heterocedasticidade (HC1) por meio do pacote `sandwich`, pois quando a variável dependente é binária, a heterocedasticidade é inerente ao modelo.[3]

Regressão simples: `deny ~ pirat`

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -0.079910   0.031967 -2.4998   0.01249 *
pirat        0.603535   0.098483  6.1283 1.036e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretação: O coeficiente de `pirat` é positivo e estatisticamente significativo. Um aumento de 0,1 unidade na razão P/I (equivalente a 100 pontos percentuais) está associado a um aumento de aproximadamente $0,604 \times 0,1 = 0,06$ na probabilidade de negação, ou seja, 6 pontos percentuais.

Calculando as probabilidades previstas para P/I = 0,3 e P/I = 0,4:

```
        1         2
0.1011508 0.1615043
```

O efeito de aumentar a razão P/I de 0,3 para 0,4 é elevar a probabilidade de negação em cerca de 6 pontos percentuais.

Incluindo raça como regressor: `deny ~ pirat + afam`

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
```

[3] Em um LPM, $\text{Var}(u_i \mid X_i) = \Pr(Y_i = 1 \mid X_i)[1 - \Pr(Y_i = 1 \mid X_i)]$, que varia com X_i por construção. Portanto, erros-padrão robustos são sempre necessários no LPM.

```
(Intercept) -0.090514   0.028600 -3.1649  0.001571 **
pirat        0.559195   0.088666  6.3067 3.387e-10 ***
afam         0.177428   0.024946  7.1124 1.502e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretação: O coeficiente de `afam` é positivo e estatisticamente significativo. Controlando pela razão P/I, um solicitante negro tem probabilidade de negação aproximadamente 17,7 pontos percentuais maior do que um solicitante branco com a mesma razão P/I.

Probabilidades previstas para um solicitante com P/I = 0,3, por raça:

```
# A tibble: 2 × 2
  Raça   `Prob. prevista (LPM)`
  <chr>                   <dbl>
1 Negro                  0.255
2 Branco                 0.0772
```

Limitações do LPM

O LPM apresenta duas limitações importantes:

1. **Efeito constante:** O modelo impõe que o efeito de `pirat` sobre a probabilidade de negação é o mesmo para qualquer valor de `pirat`. Isso não é realista: intuitivamente, o efeito deveria ser menor quando a probabilidade já está próxima de 0 ou de 1.
2. **Probabilidades fora de [0, 1]:** O LPM pode prever probabilidades negativas ou maiores que 1, o que é teoricamente inconsistente.

```
[1] -0.0905136  1.5870702
```

Os modelos probit e logit, apresentados a seguir, corrigem essas limitações ao impor que as probabilidades previstas estejam sempre dentro do intervalo $[0, 1]$.

Modelo Probit

Formulação

O modelo probit modela a probabilidade de negação usando a **função de distribuição normal padrão acumulada**, Φ:

$$\Pr(\text{deny} = 1 \mid X) = \Phi(\beta_0 + \beta_1 \cdot X_1 + ... + \beta_k \cdot X_k)$$

O argumento $z = \beta_0 + \beta_1 X_1 + ...$ é chamado de **índice-z**. O coeficiente β_1 representa a variação no índice-z associada a uma variação unitária em X_1, e **não** representa diretamente a variação na probabilidade. Para obter os efeitos sobre a probabilidade, precisamos fazer o calculo em etapas.

O probit é estimado por **máxima verossimilhança (MV)** e os estimadores têm distribuição normal assintótica, permitindo inferência pelo procedimento usual.

Regressão simples: `deny ~ pirat`

```

Call:
```

```
glm(formula = deny ~ pirat, family = binomial(link = "probit"),
    data = hmda)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.1941     0.1378 -15.927  < 2e-16 ***
pirat         2.9679     0.3858   7.694 1.43e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1744.2  on 2379  degrees of freedom
Residual deviance: 1663.6  on 2378  degrees of freedom
AIC: 1667.6

Number of Fisher Scoring iterations: 6
```

Probabilidades previstas para P/I = 0,3 e P/I = 0,4:

```
         1          2
0.09615344 0.15696777
```

O efeito de aumentar a razão P/I de 0,3 para 0,4 é elevar a probabilidade de negação de aproximadamente 9,7% para 15,9%, ou seja, 6,2 pontos percentuais. Note que neste caso, o resultado obtido é próximo ao resultado do LPM, mas em um modelo em que as probabilidades respeitam o intervalo [0, 1] e com efeito não constante.[4]

Regressão com raça: `deny ~ pirat + afam`

```

Call:
glm(formula = deny ~ pirat + afam, family = binomial(link = "probit"),
    data = hmda)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.25879    0.13669 -16.525  < 2e-16 ***
pirat        2.74178    0.38047   7.206 5.75e-13 ***
afam         0.70816    0.08335   8.496  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1744.2  on 2379  degrees of freedom
Residual deviance: 1594.3  on 2377  degrees of freedom
```

[4]Isto acontece porque estamos avaliando o impacto em valores próximos da média de `pirate` não nos extremos. Se quiser testar, você pode alterar o código e verifique a diferença para pontos no extremo e compare com o previsto pelo LPM.

```
AIC: 1600.3

Number of Fisher Scoring iterations: 5
```

Probabilidades previstas para P/I = 0,3, por raça:

```
# A tibble: 2 × 4
  Raça   `Prob. prevista` `IC 95% – inf.` `IC 95% – sup.`
  <chr>             <dbl>           <dbl>           <dbl>
1 Branco           0.0755          0.0642          0.0882
2 Negro            0.233           0.190           0.282
```

Para um solicitante com P/I = 0,3, a probabilidade prevista de negação é de cerca de 7,5% para brancos e 23,3% para negros, uma diferença de aproximadamente 15,8 pontos percentuais.

Modelo Logit

O modelo logit substitui a função normal acumulada Φ pela **função de distribuição logística** $F(z) = \frac{1}{1+e^{-z}}$:

$$\Pr(\text{deny} = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{ pirat } + \beta_2 \cdot \text{ afam})}}$$

As duas funções têm forma semelhante em "S", e na prática os resultados são muito próximos.

```

Call:
glm(formula = deny ~ pirat + afam, family = binomial(link = "logit"),
    data = hmda)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1256     0.2684 -15.370  < 2e-16 ***
pirat         5.3704     0.7283   7.374 1.66e-13 ***
afam          1.2728     0.1462   8.706  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1744.2  on 2379  degrees of freedom
Residual deviance: 1591.4  on 2377  degrees of freedom
AIC: 1597.4

Number of Fisher Scoring iterations: 5
```

Probabilidades previstas para P/I = 0,3, por raça:

```
# A tibble: 2 × 4
  Raça   `Prob. prevista` `IC 95% – inf.` `IC 95% – sup.`
  <chr>             <dbl>           <dbl>           <dbl>
```

```
1 Branco          0.0749          0.0639          0.0875
2 Negro           0.224           0.182           0.272
```

As probabilidades previstas pelo logit (branco ≈ 7,5%, negro ≈ 22,5%) são muito próximas às do probit (branco ≈ 7,5%, negro ≈ 23,3%). Esse resultado ilustra um padrão geral: **para fins empíricos, probit e logit produzem probabilidades previstas muito semelhantes**, e a escolha entre eles raramente altera as conclusões substantivas.

Exercício E11.1 Stock e Watson (4a Edição)

Como atividade prática, iremos discutir uma aplicação de modelos com variável dependente limitada na área de macroeconomia.

Em abril de 2008, a taxa de desemprego nos Estados Unidos era de 5,0%. Em abril de 2009, ela havia subido para 9,0% e, em outubro de 2009, subiu ainda mais, chegando a 10,0%. Alguns grupos de trabalhadores tiveram maior probabilidade de perder o emprego do que outros durante a Grande Recessão? Por exemplo, trabalhadores jovens tiveram maior probabilidade de perder o emprego do que trabalhadores de meia-idade? E quanto a trabalhadores com diploma universitário em comparação com aqueles sem diploma, ou mulheres em comparação com homens?

No site do livro, você encontrará o arquivo de dados Employment_08_09, que contém uma amostra aleatória de 5.440 trabalhadores entrevistados em abril de 2008 e que declararam estar empregados em tempo integral. O dicionário de variáveis está disponível em Employment_08_09_Description, também no site. Esses trabalhadores foram entrevistados novamente um ano depois, em abril de 2009, e perguntados sobre sua situação ocupacional (empregado, desempregado ou fora da força de trabalho). A base de dados também inclui diversas características demográficas de cada indivíduo. Use esses dados para responder às seguintes perguntas.

a. Que fração dos trabalhadores da amostra estava empregada em abril de 2009? Use sua resposta para calcular um intervalo de confiança de 95% para a probabilidade de um trabalhador estar empregado em abril de 2009, condicional a estar empregado em abril de 2008.

b. Estime a regressão de `Employed` em `Age` e $\texttt{Age}^2$, usando um modelo de probabilidade linear.

   i. Com base nessa regressão, a idade foi um determinante estatisticamente significativo do emprego em abril de 2009?

   ii. Há evidência de um efeito não linear da idade sobre a probabilidade de estar empregado?

   iii. Calcule a probabilidade prevista de emprego para um trabalhador de 20 anos, um de 40 anos e um de 60 anos.

c. Repita o item (b) usando uma regressão probit.

d. Repita o item (b) usando uma regressão logit.

e. Há diferenças importantes entre suas respostas nos itens (b)–(d)? Explique.

f. A base de dados inclui variáveis que medem a escolaridade, o sexo, a raça, o estado civil, a região do país e o rendimento semanal dos trabalhadores em abril de 2008.

   i. Construa uma tabela semelhante à Tabela 11.2 do livro para investigar se as conclusões sobre o efeito da idade sobre o emprego obtidas em (b)–(d) são afetadas por viés de variável omitida.

   ii. Use as regressões da sua tabela para discutir as características dos trabalhadores mais prejudicados pela Grande Recessão.

g. Os resultados em (a)–(f) basearam-se na probabilidade de estar empregado. Trabalhadores que não estão empregados podem estar (i) desempregados ou (ii) fora da força de trabalho. As conclusões obtidas em (a)–(f) também valem para os trabalhadores que ficaram desempregados? (*Dica:* use a variável binária `Unemployed` em vez de `Employed`.)

h. Esses resultados tratam das transições no mercado de trabalho durante a Grande Recessão, mas e quanto às transições em tempos normais? No site do livro, você encontrará o arquivo de dados Employment_06_07, que contém as mesmas variáveis, mas para os anos de 2006–2007 (o dicionário de variáveis é o mesmo da base anterior). Analise esses dados e comente as diferenças nas transições de emprego em períodos de recessão e em tempos normais.