

## Exercícios sobre Estimação por MQO

No que segue, sempre suponha que as variáveis aleatórias relevantes estão definidas no mesmo espaço de probabilidade.

**Exercício 1** Seja  $S$  um vetor aleatório, e  $X$  um conjunto de variáveis aleatórias. Mostre a lei da variância total, i.e:

$$\mathbb{V}[S] = \mathbb{E}[\mathbb{V}[S|X]] + \mathbb{V}[\mathbb{E}[S|X]].$$

**Exercício 2** Sejam  $A$  e  $B$  duas matrizes simétricas positiva definidas de dimensão  $n$ . Mostre que, se  $A - B$  é positiva semidefinida, então  $B^{-1} - A^{-1}$  é positiva semidefinida.

**Exercício 3** Considere o seguinte modelo linear para uma amostra de  $n$  observações:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\epsilon},$$

onde  $\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}_1, \mathbf{X}_2] = 0$ ,  $\mathbb{V}[\boldsymbol{\epsilon}|\mathbf{X}] = \sigma^2 \mathbb{I}_{n \times n}$ ,  $[\mathbf{X}_1 \ \mathbf{X}_2]$  tem posto cheio, e  $\mathbf{X}_1 = [\mathbf{1}_{n \times 1} \ \mathbf{Z}]$ , onde  $\mathbf{1}_{n \times 1}$  é um vetor de uns e  $\mathbf{Z}$  é independente de  $\mathbf{X}_2$ .

- a Mostre que o estimador de MQO de  $\mathbf{y}$  em  $\mathbf{X}_1$ , denotado por  $\tilde{b}_1$ , é não viciado (condicionalmente a  $\mathbf{X}_1$ ) para  $\beta_1$ .
- b Mostre que a variância condicional de  $\tilde{b}_1$  se escreve como:

$$\mathbb{V}[\tilde{b}_1|\mathbf{X}_1] = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbb{V}[\mathbf{X}_2 \beta_2 | \mathbf{X}_1] \mathbf{X}_1' (\mathbf{X}_1 \mathbf{X}_1')^{-1} + \sigma^2 (\mathbf{X}_1' \mathbf{X}_1)^{-1}$$

- c Mostre que a variância do estimador de MQO de  $\beta_1$  da regressão que inclui  $\mathbf{X}_1$  e  $\mathbf{X}_2$ , denotado por  $\hat{b}_1$ , é não viciado para  $\beta_1$  (condicionalmente a  $\mathbf{X}_1$  e  $\mathbf{X}_2$ ) e que sua variância condicional é dada por:

$$\mathbb{V}[\hat{b}_1|\mathbf{X}_1, \mathbf{X}_2] = \sigma^2 (\mathbf{X}_1' M_2 \mathbf{X}_1)^{-1},$$

onde  $M_2$  é a residualizadora de  $\mathbf{X}_2$ .

- d Mostre que  $(\mathbf{X}_1' M_2 \mathbf{X}_1)^{-1} - (\mathbf{X}_1' \mathbf{X}_1)^{-1}$  é positiva semidefinida.
- e Mostre que  $(\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbb{V}[\mathbf{X}_2 \beta_2 | \mathbf{X}_1] \mathbf{X}_1' (\mathbf{X}_1' \mathbf{X}_1)^{-1}$  é positiva semidefinida.
- f Usando a lei da variância total e os dois resultados anteriores, mostre que a diferença entre as variâncias *incondicionais* de  $\tilde{b}_1$  e  $\hat{b}_1$  se fatoram em dois termos,  $A$  e  $B$ , em que  $A$  é a esperança de uma matriz aleatória positiva semidefinida e  $B$  a esperança de uma matriz negativa semidefinida. Proveja uma intuição para esse resultado. Se o interesse reside em  $\beta_1$ , quando fará sentido incluir  $\mathbf{X}_2$  na regressão?
- g Mostre que se  $\beta_2 = 0$ , então  $\mathbb{V}[\hat{\beta}_1] - \mathbb{V}[\tilde{\beta}_1]$  é positiva semidefinida. Esse resultado contradiz o teorema de Gauss-Markov? Por quê?

**Exercício 4** Suponha que você esteja interessado em estudar o efeito de infraestrutura escolar no desempenho escolar médio dos alunos. Você possui acesso a uma amostra aleatória de municípios, para os quais você observa o desempenho escolar médio dos alunos de  $C$  escolas,  $c = 1, \dots, C$ , em cada um dos  $M$  municípios,  $m = 1, \dots, M$ , amostrados ( $Y_{c,m}$ ), e uma medida de conectividade da escola à Internet, em *megabytes por segundo* (mbps) ( $X_{c,m}$ ). Você então postula o seguinte modelo causal linear para  $Y_{c,m}$

$$Y_{c,m} = \gamma X_{c,m} + \epsilon_{c,m}, \quad c = 1, \dots, C, \quad m = 1, \dots, M,$$

onde  $\gamma$  é o efeito causal de se aumentar a conectividade da escola em 1mbps sobre o desempenho médio dos alunos, e  $\epsilon_{c,m}$  são as demais causas não observadas do desempenho escolar.

- a Sob qual condição a inclinação do melhor preditor linear de  $Y_{c,m}$  num intercepto e  $X_{c,m}$  identifica  $\gamma$ ? Você acredita que essa condição seria satisfeita na prática? Por quê?

Em virtude da dificuldade de se acreditar na hipótese anteriormente postulada, um colega seu sugere que você explicitamente acomode a presença de causas não observadas no nível municipal que afetam desempenho municipal, considerando a decomposição:

$$\epsilon_{c,m} = \psi_m + \nu_{c,m},$$

onde  $\psi_m$  é um conjunto de causas comuns ao município, e  $\nu_{c,m}$  são causas idiossincráticas às escolas. Seu colega lhe diz que é possível levar em conta (“controlar”) as causas  $\psi_m$  na estimação, mesmo que estas não sejam observadas, considerando uma regressão de  $Y_{c,m}$  em  $X_{c,m}$  e um conjunto de  $M$  *dummies* municipais, i.e.  $D_{c,m}(l) = \mathbf{1}_{\{l\}}(m)$ ,  $l = 1 \dots, m$ .

- b Usando o teorema de Frisch-Waugh-Lovell, mostre que o estimador do coeficiente de  $X_{c,m}$  sugerido pelo seu colega é algebricamente idêntico a se estimar  $\gamma$  a partir de uma regressão (sem intercepto) de  $Y_{c,m} - \bar{Y}_m$  em  $X_{c,m} - \bar{X}_m$ , onde  $\bar{Z}_m = \frac{\sum_{i=1}^C Z_{i,m}}{C}$ . Em outras palavras, o estimador sugerido pelo seu colega é equivalente ao estimador de MQO do modelo transformado.

$$Y_{c,m} - \bar{Y}_m = \gamma(X_{c,m} - \bar{X}_m) + \epsilon_{c,m} - \bar{\epsilon}_m.$$

- c Proveja condições suficientes para que o estimador de MQO sugerido seja consistente para o parâmetro de interesse, num regime assintótico em que  $M \rightarrow \infty$ . Qual a interpretação dessas condições?
- d Derive a variância assintótica do estimador, sob as hipóteses do item anterior, permitindo que haja correlação arbitrária entre os  $\epsilon_{c,m}$  intramunicipais. Sugira um estimador consistente para a variância assintótica.

**Exercício 5** O arquivo `cps_union_data.csv` contém uma amostra **12.834 indivíduos na força de trabalho norte-americana**, extraída do suplemento anual de 2019 da *Pesquisa de População Atual dos EUA* (*Current Population Survey*). Seu objetivo é **estimar o efeito causal da filiação/cobertura sindical** (variável `union`) sobre os **rendimentos semanais** (variável `earnings`). O conjunto de dados contém diversas outras variáveis que podem ser utilizadas como controles (verifique o dicionário do conjunto de dados, disponível em `dictionary.xlsx`).

1. Como ponto de partida, compare a **média dos rendimentos** entre os indivíduos com cobertura sindical (`union == 1`) e os indivíduos sem essa cobertura (`union == 0`). Qual é a diferença estimada? Usando um teste  $t$  para comparação de médias entre duas populações, teste a hipótese nula de que a remuneração média na população filiada é igual à remuneração média na população não filiada, contra a alternativa bilateral. Você rejeita a hipótese nula a 5% de significância? E a 1%? diferença é estatisticamente significativa? Você acredita que essa diferença é uma estimativa crível do impacto causal da cobertura sindical? Por quê? *Dica:* utilize o comando `t.test` do pacote básico do R.
2. Mostre analiticamente que a diferença de médias estimada anteriormente pode ser obtida através de uma regressão de `earnings` em um intercepto e `union`, e que a estatística  $t$  do coeficiente associado a `union` baseada em erros padrão robustos à heterocedasticidade é idêntica, a não ser por correção de graus de liberdade, à estatística  $t$  utilizada no teste anterior.
3. Para melhorar e/ou avaliar a credibilidade dos seus resultados anteriores, você decide considerar um modelo linear causal da forma:

$$\text{earnings}_i = \beta_0 + \beta_1 \text{union}_i + \gamma' Z_i + \epsilon_{it} \quad (1)$$

onde  $Z_i$  representa um conjunto de variáveis de controle.

- a) Especifique o modelo linear da Equação (1) apresentando um conjunto de covariáveis  $Z$  a serem incluídas como controles. Justifique a escolha dessas variáveis. Qual é a interpretação de  $\beta_1$  no seu modelo?
- b) Estime o modelo especificado. Qual é a sua estimativa de  $\beta_1$ ? Usando erros padrão robustos à heterocedasticidade, ela é estatisticamente significativa? Comente brevemente os seus resultados.
- c) Usando o teorema de Frisch-Waugh-Lovell, mostre que o estimador de MQO para  $\beta_1$  tem a seguinte forma:

$$\hat{\beta}_1 = \sum_{i=1}^N \text{union}_i \cdot \omega_i \cdot \text{earnings}_i - \sum_{i=1}^N (1 - \text{union}_i) \cdot \omega_i \cdot \text{earnings}_i \quad (2)$$

onde os pesos  $\omega_i$  são definidos como:

$$\omega_i = \frac{\hat{\xi}_i (2 \cdot \text{union}_i - 1)}{\text{SSR}_{\text{union}, Z}} \quad (3)$$

com  $\hat{\xi}_i$  sendo o resíduo da observação  $i$  de uma regressão linear de `unioni` sobre  $Z$  (incluindo intercepto); e  $\text{SSR}_{\text{union}, Z}$  é a soma dos quadrados dos resíduos dessa regressão auxiliar.

- d) Calcule os pesos para sua especificação usando a fórmula acima. Apresente estatísticas descritivas da distribuição dos pesos nos grupos de controle e tratamento.
  - Os pesos somam 1 no grupo de controle?
  - E no grupo de tratamento?
  - Há valores negativos?
  - E outliers?

Como esses pesos diferem com relação ao estimador de diferença de médias para o tratamento? Proveja uma intuição para essa diferença.

**Exercício 6** O objetivo deste exercício consiste em entender as propriedades do estimador da variância robusta a *cluster*. Para isso, usaremos os dados do resultado da aplicação de um teste lógico em alunos de um conjunto de escolas primárias, disponível em `dados.csv`.

a Construa uma função em R que:

1. Sorteie 500 escolas,  $c = 1, \dots, 500$ , **com reposição**, do conjunto de dados. Note que haverá escolas repetidas, mas elas serão tratadas como distintas para os fins da simulação (de modo a refletir amostragem aleatória de uma população de escolas).
2. Para cada uma das  $c = 1, \dots, 500$  escolas sorteadas, gere um tratamento fictício no nível escolar,  $X_c \sim \text{Uniforme}[0, 1]$ , de forma independente entre as escolas e dos dados.
3. Com base na base de dados construída, estime por MQO o modelo linear:

$$\text{logico}_{i,c} = \alpha + \beta X_c + \gamma \text{sexo}_{i,c} + \phi \text{idade}_{i,c} + \epsilon_{i,c},$$

4. Guarde as estatísticas  $t$  do teste da hipótese nula de que  $\beta = 0$  baseados em erros padrão robustos à heterocedasticidade e a *cluster* no nível da escola.
- b Aplique a função 1000 vezes, e calcule a proporção de casos em que cada um dos dois tipos de testes  $t$ , a 5% de significância, rejeita a hipótese nula. Se o estimador da variância estiver “correto”, qual a proporção esperada de casos em que se rejeita a nula? Por quê? O que você obteve na prática? Por quê?
- c Agora adapte a função anterior, permitindo que o tratamento, agora denotado por  $\tilde{X}_{i,c} \sim \text{Uniforme}[0, 1]$ , varie no nível individual de forma independente. Aplique a nova função mil vezes. Reporte a proporção de casos em que se rejeitou a hipótese nula. O que ocorreu? Por quê?
- d Adapte mais uma vez a função, agora permitindo que o tratamento exerça um efeito heterogêneo entre escolas. Especificamente, você gerará  $\widetilde{\text{logico}}_{i,c} = \text{logico}_{i,c} + \tau_c \tilde{X}_{i,c}$ , onde  $\tau_c = 1$  com probabilidade 1/2 e -1 com probabilidade 1/2, e independente das demais variáveis. Nós continuaremos rodando a regressão linear de  $\widetilde{\text{logico}}_{i,c}$  em  $\tilde{X}_{i,c}$  e controles. Aplique a função nova 1000 vezes. Qual a proporção de casos rejeitados agora? Por quê?