Lista 1 - Econometria II (2022)

Professora: Fabiana Rocha
Monitor: Thiago Pastorelli Rodrigues

Exercício 1. (Cameron e Trivedi, 2005) Considere um modelo de tratamento dado por $y = x'\beta + \alpha d + \epsilon$, em que d é uma variável indicadora que assume o valor 1 se o tratamento ocorre aleatoriamente e 0, caso contrário (ainda assim aleatório).

i. A aleatorização do tratamento é condição suficiente para a identificação de α?

ii. A aleatorização do tratamento é condição suficiente para a identificação de α e β?

Exercício 2. Considere o seguinte modelo de tratamento:

$$Y = \alpha + \beta T + \varepsilon,$$

em que Y é a renda mensal das família, T é o tratamento em um experimento de micro-finanças ($T = 1$ se a família pertence ao grupo de tratamento e $T = 0$ se a família não pertence ao grupo de tratamento), e ε é o termo de erro. Sob aleatorização pura, o impacto do programa pode é dado por β. Agora, o caso de aleatorização parcial, em que as unidades de tratamento e controle são escolhidas de forma aleatória mas condicionado a características observáveis X (renda da família, por exemplo):

$$Y^T = \alpha^T + \beta^T X + \varepsilon^T$$
$$Y^C = \alpha^C + \beta^C X + \varepsilon^C$$

em que a primeira equação representa o grupo que recebe tratamento e a segunda equação representa aqueles que não recebem o tratamento. Nesse caso, qual é o impacto do programa?

Exercício 3. Considere que o efeito do tratamento é o mesmo para todos os indivíduos:

a) Mostre que a equação na forma:

$$Y_i = \alpha + \rho D_i + \eta_i$$

representa:

$$\underbrace{E[Y_i|D=1] - E[Y_i|D_i=0]}_{\text{Diferença simples dos resultados observados}} = ATE + \underbrace{(E[\eta_i|D_i=1] - E[\eta_i|D_i=0])}_{\text{Viés de seleção}}$$

b) Explique como a aleatorização resolve o problema de viés de seleção.

Exercício 4. As validades externa e interna de uma amostragem são conceitos relacionados à seleção aleatória de uma população de interesse e depois uma divisão apropriada entre grupo de controle e grupo de tratamento. Defina validade interna e externa para uma amostragem qualquer.

Exercício 5. Quais dos seguintes itens podem fazer com que os estimadores de MQO sejam viesados?

i. Heterocedasticidade.
ii. Omitir uma variável importante.
iii. Um coeficiente de correlação amostral de 0,95 entre duas variáveis independentes incluídas no modelo.

Exercício 6. Suponha que a produtividade média do trabalhador da indústria (*prodmed*) dependa de dois fatores - horas médias de treinamento do trabalhador (*treinmed*) e aptidão média do trabalhador (*aptimed*):

$$prodmed = \beta_0 + \beta_1 treinmed + \beta_2 aptimed + u$$

Assuma que essa equação satisfaça as hipóteses de Gauss-Markov. Se um subsídio foi dado às empresas cujos trabalhadores têm uma aptidão menor do que a média, de modo que *treinmed* e *aptimed* sejam negativamente correlacionados, qual é o provável viés em $\tilde{\beta}_1$ obtido da regressão simples de *prodmed* sobre *treinmed*?

Exercício 7. A equação seguinte descreve o preço mediano das residências de uma comunidade em termos da quantidade de poluição (*oxn*, de óxido nitroso) e do número médio de cômodos nas residências da comunidade (*comods*):

$$log(preço) = \beta_0 + \beta_1 log(oxn) + \beta_2 comods + u$$

i. Quais são os prováveis sinais de β_1 e β_2? Qual é a interpretação de β_1? Explique.
ii. Por que *oxn* [ou, mais precisamente, $log(oxn)$] e *comods* deveriam ser negativamente correlacionados? Se esse é o caso, a regressão simples de $log(preço$ sobre $log(oxn)$ produz um estimador viesado para cima ou para baixo de β_1?
iii. Utilizando os dados disponíveis de uma pesquisa sobre preços das residências, foram estimadas as seguintes regressões:

$$\widehat{log(preço)} = 11,71 - 1,043 log(oxn),\ n = 506,\ R^2 = 0,264$$
$$\widehat{log(\text{preço})} = 9,23 - 0,718 log(oxn) + 0,306 comods,\ n = 506,\ R^2 = 0,514$$

A relação entre as estimativas da elasticidade do *preço* das regressões simples e múltipla com relação ao *oxn* é a que você esperava, tomando como base suas respostas do item (ii)? Pode-se dizer que -0,718 está claramente mais próximo da elasticidade verdadeira que -1,043?

Exercício 8. Seja *mate10* a porcentagem de aprovação em um teste padrão de matemática de estudantes de uma escola de ensino médio. Estamos interessados em estimar o efeito do gasto por estudante no desempenho em matemática. Um modelo simples é:

$$mate10 = \beta_0 + \beta_1 log(gasto) + \beta_2 log(matricl) + \beta_3 pobreza + u$$

onde *pobreza* é a porcentagem de estudantes vivendo em condições de pobreza.

i. A variável *prgalm* é a porcentagem de estudantes qualificados para o programa de merenda escolar financiado pelo governo federal. Por que ela é uma variável *proxy* razoável de *pobreza*?

ii. A tabela abaixo contém as estimativas de MQO, com e sem *prgalm* como uma variável explicativa.

Variáveis Independentes	**(1)**	**(2)**
log(gasto)	11,13 (3,30)	7,75 (3,04)
log(matricl)	0,022 (0,615)	-1,26 (0,58)
prgalm	-	-0,324 (0,036)
Intercepto	-69,24 (26,72)	-23,14 (24,99)
Observações R^2	428 0,0297	428 0,1893

Explique por que o efeito dos gastos sobre *mate10* é menor na coluna (2) do que na coluna (1). O efeito na coluna (2) ainda é estatisticamente maior que zero?

iii. Parece que as taxas de aprovação são menores em escolas maiores, com outros fatores sendo iguais?

iv. Interprete o coeficiente de *prgalm* na coluna (2).

v. O que você deduz do substancial aumento de R^2 da coluna (1) para a coluna (2)?

Exercício 9. A equação seguinte explica o número de horas por semana que uma criança passa assistindo televisão, em termos da idade da criança, educação da mãe, educação do pai e número de irmãos:

$$tvhoras* = \beta_0 + \beta_1 idade + \beta_2 idade^2 + \beta_3 educm + \beta_4 educp + \beta_5 irms + u$$

Estamos preocupados com a possibilidade de que *tvhoras** tenha sido medida com erro em nossa pesquisa. Seja *tvhoras* o número de horas por semana que se gasta assistindo televisão.

i. O que as hipóteses de erro clássico nas variáveis (CEV) requerem nesta aplicação?

ii. Você acha que as hipóteses CEV têm possibilidades de se manter? Explique.

Exercício 10. Considere a seguinte citação:

"Econometricians want too much from the data and hence tend to be disappointed by the answers, because the data are incomplete and imperfect. In part it is our fault, the appetite grows with eating. As we get larger samples, we keep adding variables and expanding our models, until on the margin, we come back to the same insignificance levels. - Griliches, American Economic Review, 1985."

Logo no início da citação, o autor aponta para o problema de não disponibilidade de dados. Com base nisso, responda os itens abaixo:

a) Quando a omissão de uma variável torna os estimadores de uma regressão inconsistentes? Explique.
b) Suponha que o modelo verdadeiro é $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, mas o modelo estimado é $y_i = \gamma_0 + \gamma_1 x_{1i} + \epsilon$. Derive o viés de variável omitida. Derive em quais casos este viés é positivo, negativo ou nulo.
c) Mostre que $Var(\hat{\gamma}_1|X) \leq Var(\hat{\beta}_1|X)$.

Exercício 11. Suponha que estamos tentando entender como os indivíduos escolhem quantas quilocalorias consumir diariamente em função da porcentagem de gordura corporal e atividade física praticada:

$$calorias = \beta_0 + \beta_1 gordura + \beta_2 exercício + u$$

onde *calorias* representa as quilocalorias ingeridas em um dia, *gordura* é o percentual de gordura corporal e *exercício* são as horas de exercício praticadas por semana. No entanto, os dados disponíveis não traziam a porcentagem de gordura dos indivíduos. Uma saída encontrada foi usar o Índice de Massa Corporal (IMC) como variável independente no modelo. Sabe-se também que:

$$gordura = \delta_0 + \delta_1 IMC + \varepsilon$$

Com essas informações, responda:
a) Qual a relação entre as variáveis gordura e IMC.
b) Escreva o modelo estimado e faça sua equivalência com o modelo verdadeiro.
c) É necessária alguma hipótese para que o modelo seja válido?

Exercício 12. Suponha um modelo para explicar a determinação da poupança das famílias, dado por:

$$poup = \beta_0 + \beta_1 RendaFam + \beta_2 TamFam + \beta_3 Educ + \beta_4 Idade + u$$

em que $RendaFam$ é a renda do domicílio, $TamFam$ é o tamanho da família, $Educ$ e $Idade$ são educação e idade do chefe do domicílio, respectivamente. Assuma que o termo de erro é tal que $E(u|RendaFam, TamFam, Educ, Idade) = 0$. Suponha ainda que conseguimos uma amostra aleatória de famílias brasileiras, suficientemente grande para que as propriedades assintóticas dos estimadores sejam válidas. Com essas informações, responda:

a) Estimando este modelo pelo método de Mínimos Quadrados Ordinários (MQO) é possível obter estimadores não-viesados para os coeficientes β_j? Explique.
b) Qual sinal você esperaria encontrar para cada um dos coeficientes? Explique.
c) Vamos agora restringir a nossa amostra, usando apenas casais sem filhos. Neste caso, quais parâmetros do modelo populacional podem ser estimados? Explique.
d) Vamos voltar a usar nossa amostra completa. No entanto, considere que seja estimado o seguinte modelo:

$$poup = \gamma_0 + \gamma_1 RendaFam + \gamma_2 TamFam + \gamma_3 Idade + u$$

Qual hipótese deveríamos assumir para garantir que os estimadores obtidos sejam não-viesados? Explique porque essa hipótese é improvável.
e) No caso do item anterior, qual deveria ser a direção do viés? Explique.

Exercício 13. Considere o seguinte modelo de equações simultâneas:

$$\text{Demanda:} \quad Q^D = \alpha_0 + \alpha_1 P + \alpha_2 Y + u_D$$
$$\text{Oferta:} \quad Q^O = \beta_0 + \beta_1 P + u_O$$
$$\text{Market Clearing:} \quad Q^D = Q^O$$

onde Q^D é a quantidade demandada, Q^O é a quantidade ofertada, P é o preço e Y o produto, u_D e u_O são termos de erro aleatórios, com média zero e variância constante.

a) Encontre a forma reduzida de Q e P.
b) Quais condições sobre os parâmetros do modelo devem ser impostas para que possamos encontrar a forma reduzida das equações?
c) Em que termo da forma reduzida é possível identificar a fonte do viés de simultaneidade?

Exercício 14. (ANPEC 2005 - Questão 8 Modificada) Considere o modelo de equações simultâneas:

$$\text{Demanda:} \quad Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 X_t + \varepsilon_{1t}$$
$$\text{Oferta:} \quad Q_t = \beta_0 + \beta_1 P_t + \varepsilon_{2t}$$

Onde Q_t é a quantidade de equilíbrio, P_t é o preço de equilíbrio, X_t é uma variável exógena e ε_{1t} e ε_{2t} são termos de erro aleatórios, com média zero e variância constante. São corretas as seguintes afirmativas:
a) As equações da demanda e da oferta são exatamente identificadas.
b) Os parâmetros estruturais do modelo são consistentemente estimados por Mínimos Quadrados Ordinários.
c) As equações na forma reduzida são: $P_t = \pi_0 + \pi_1 X_t + v_t$ e $Q_t = \pi_2 + \pi_3 X_t + w_t$, em que: $\pi_0 = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1}$; $\pi_1 = -\frac{\alpha_2}{\alpha_1 - \beta_1}$; $v_t = \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\alpha_1 - \beta_1}$; $\pi_2 = \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1}$; $\pi_3 = -\frac{\alpha_2 \beta_1}{\alpha_1 - \beta_1}$; $w_t = \frac{\alpha_1 \varepsilon_{2t} - \beta_1 \varepsilon_{1t}}{\alpha_1 - \beta_1}$.
d) As estimativas dos parâmetros da forma reduzida descritos no item anterior, por Mínimos Quadrados Ordinários, são consistentes.

Exercício 15. Adaptado de Greene (2018) Considere o seguinte modelo de duas equações:

$$y_1 = \gamma_1 y_2 + \beta_{11} x_1 + \beta_{21} x_2 + \beta_{31} x_3 + \varepsilon_1$$
$$y_2 = \gamma_2 y_1 + \beta_{12} x_1 + \beta_{22} x_2 + \beta_{31} x_3 + \varepsilon_2$$

i. Verifique que da forma enunciada nenhuma equação é identificada.

ii. Verifique se as seguintes condições são suficientes para identificar (total e parcialmente) o modelo:

   (a) $\beta_{21} = \beta_{32} = 0$
   (b) $\gamma_1 = 0$

Exercício 16. Considere o seguinte modelo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Suponha que z_i é uma variável *dummy* que serve de instrumento para x_i.

a) Mostre que o estimador IV de β_1 pode ser escrito como:

$$\hat{\beta}_1 = \frac{\bar{y_1} - \bar{y_0}}{\bar{x_1} - \bar{x_0}}$$

onde $\bar{y_1}$ e $\bar{x_1}$ são as médias de y_i e x_i na parte da amostra em que $z_i = 1$, e $\bar{y_0}$ e $\bar{x_0}$ são as médias de y_i e x_i para a parte da amostra em que $z_i = 0$.

b) Como devemos interpretar $\hat{\beta}_1$?

Exercício 17. Considere o seguinte modelo:

$$y = X\beta + u$$

Em que y é um vetor $n \times 1$, X é uma matriz $n \times k$ contendo todas as variáveis explicativas (inclusive o intercepto), e u é um vetor $n \times 1$. Considere que $E(X'u) \neq 0$. Considere também que

você possui uma variável instrumental para cada variável explicativa endógena, sendo Z a matriz de instrumentos.

a) Qual a dimensão e quais termos estão contidos na matriz Z? Explique.
b) Derive o estimador de variáveis instrumentais na forma matricial.
c) Encontre a variância do estimador derivado no item acima (também na forma matricial). Compare esta variância com a do estimador de MQO.

Exercício 18. Suponha que você queira estimar o efeito da frequência escolar sobre o desempenho dos alunos. Um modelo básico é:

$$respad = \beta_0 + \beta_1 taxafreq + \beta_2 nmgradp + \beta_3 tac + u$$

onde $respad$ é o resultado padronizado de um exame, $taxafreq$ é a taxa de frequência escola e $nmgradp$ é o desempenho dos alunos no passado.

a) Defina $dist$ como distância da residência do aluno até o local de estudos. Você considera que $dist$ não é correlacionada com u?
b) Assumindo que $dist$ e u sejam não correlacionados, quais outras hipóteses $dist$ deve satisfazer para ser uma IV válida de $taxafreq$?
c) Suponha que adicionamos um termo de interação $nmgradp \cdot taxafreq$:

$$respad = \beta_0 + \beta_1 taxafreq + \beta_2 nmgradp + \beta_3 tac + \beta_4 nmgradp \cdot taxafreq + u$$

Se $taxafreq$ é correlacionada com u, então, em geral a interação $nmgradp \cdot taxafreq$ também será. O que poderíamos usar como uma boa IV para $nmgradp \cdot taxafreq$?

Exercício 19. Considere o seguinte modelo estrutural:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

Considere então que temos uma variável explicativa endógena, uma variável explicativa exógena e uma variável exógena adicional, de modo que a forma reduzida para y_2 é dada por: $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$.

a) Escreva a forma reduzida para y_1 e encontre os α_j em termos de β_j e π_j.
b) Encontre a forma reduzida do erro v_1, em termos de u_1, v_2 e os parâmetros.
c) Como você estimaria consistentemente os α_j?

Exercício 20. (ANPEC 2013 - Questão 14) Usando uma base de dados que contém informações sobre 65.000 indivíduos, estimamos o retorno da educação utilizando educação da mãe do indivíduo i como instrumento para educação do indivíduo i, obtendo o seguinte resultado:

$$\hat{Y_i} = \underset{(220,75)}{-320,89} + \underset{(38,68)}{67,21} X_i + \underset{(1,60)}{5,49} W_i, \quad R^2 = 0,46$$

onde Y_i representa a renda mensal do indivíduo i, X_i o número de anos de estudo do indivíduo i e W_i a idade do indivíduo i. Ainda, considere que Z_i representa a educação da mãe do indivíduo i. O termo entre parênteses representa o desvio padrão respectivo. Baseado nessas informações, julgue as afirmativas:

a) Para a educação da mãe (Z_i) ser um bom instrumento para a educação do filho (X_i), ele deve atender duas condições: (1) $Cov(X_i, Z_i) \neq 0$ e (2) $Cov(X_i, u_i) = 0$.
b) Com base nos resultados acima, podemos testar a condição (1) $Cov(X_i, Z_i) \neq 0$, isto é, que a educação da mãe é correlacionada com a educação do filho.
c) Com base nos resultados acima, é possível rejeitar a hipótese de que a educação da mãe gtem um efeito parcial significativo na renda mensal do indivíduo ao nível de significância de 5%.
d) Suponha que educação do pai seja correlacionada com educação da mãe e tenha uma correlação não-nula com a renda mensal do indivíduo. Neste caso, educação da mãe contínua sendo um instrumento válido para educação do indivíduo.
e) Se houver uma correlação positiva entre idade e educação da mãe, educação da mãe deixa de ser um instrumento válido para educação do indivíduo.