Lista de Exercícios 4

Variáveis instrumentais, o preço do IV e os métodos de reamostragem

Econometria I · PIMES/UFPE

Justifique todas as respostas.

---

1. **Endogeneidade e identificação por um instrumento.**

   (a) Defina endogeneidade no modelo $y = \mathbf{x}'\beta + \varepsilon$. Contraste o **parâmetro de projeção** $\beta^* = \mathbb{E}[\mathbf{x}\mathbf{x}']^{-1}\mathbb{E}[\mathbf{x}y]$ (que o MQO sempre estima) com o **parâmetro estrutural** β, e mostre que a endogeneidade ($\mathbb{E}[\mathbf{x}\varepsilon] \neq \mathbf{0}$) é exatamente a situação $\beta^* \neq \beta$. Cite as **quatro fontes** (variável omitida, erro de medida, seleção, simultaneidade).

   (b) Explique por que a endogeneidade **não** é corrigida por aumentar a amostra — em contraste com um estimador apenas ruidoso. Qual condição populacional é violada, e por que $n \to \infty$ não a restaura?

   (c) Um instrumento $\mathbf{z}$ (dimensão ℓ) deve satisfazer **relevância** e **exclusão**. Escreva a condição de momento $\mathbb{E}[\mathbf{z}(y - \mathbf{x}'\beta)] = \mathbf{0}$ como o sistema linear $\mathbb{E}[\mathbf{z}\mathbf{x}']\beta = \mathbb{E}[\mathbf{z}y]$, e enuncie a **condição de posto** $\text{posto}(\mathbb{E}[\mathbf{z}\mathbf{x}']) = k$ e a de **ordem** $\ell \geq k$, com a tricotomia exatamente / sobre / sub-identificado.

   (d) Argumente por que a **relevância** é testável nos dados (pelo primeiro estágio), mas a **exclusão** é, em geral, um **argumento de desenho** não-testável.

2. **Os estimadores IV e 2SLS.**

   (a) No caso **exatamente identificado** ($\ell = k$), obtenha o estimador IV $\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}$ como o análogo amostral da condição de momento. Substituindo $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, mostre que $\mathbf{b}_{\text{IV}} = \beta + \left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}'\varepsilon\right)$ e conclua $\operatorname{plim}\mathbf{b}_{\text{IV}} = \beta$. Qual fator carrega a **exogeneidade** e qual a **relevância**?

   (b) No caso **sobre-identificado** ($\ell > k$), apresente o **2SLS** $\mathbf{b}_{\text{2SLS}} = (\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{Y}$, com $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, e mostre (usando $\mathbf{P}_{\mathbf{Z}}$ simétrica e idempotente) que ele é o MQO de $\mathbf{Y}$ sobre $\widehat{\mathbf{X}} = \mathbf{P}_{\mathbf{Z}}\mathbf{X}$ — isto é, o IV com $\widehat{\mathbf{X}}$ como instrumento. Descreva os **dois estágios**.

   (c) Explique o que o 2SLS **conserva** e **descarta** de $\mathbf{X}$, e por que a parte conservada ($\widehat{\mathbf{X}} = \mathbf{P}_{\mathbf{Z}}\mathbf{X}$) é a parte "limpa" (não-correlacionada com ε).

3. **Distribuição assintótica do 2SLS.**

   (a) Substituindo a forma estrutural em $\mathbf{b}_{\text{2SLS}}$ e reescalando por $\sqrt{n}$, mostre que

   $$\sqrt{n}\,(\mathbf{b}_{\text{2SLS}} - \beta) = \left(\tfrac{1}{n}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}\right)^{-1}\tfrac{1}{\sqrt{n}}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\varepsilon.$$

   (b) Supondo relevância ($\mathbf{Q}_{zx} = \mathbb{E}[\mathbf{z}\mathbf{x}']$ de posto cheio), exogeneidade e momentos finitos, argumente — via lei dos grandes números e teorema central do limite — que

o "pão" $\frac{1}{n}\mathbf{X}'\mathbf{P_Z}\mathbf{X} \xrightarrow{p} \mathbf{H} = \mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx}$ e que o escore $\frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{P_Z}\varepsilon \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$, com $\Sigma = \mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\Omega\,\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx}$ e $\Omega = \mathbb{E}[\mathbf{z}\mathbf{z}'\varepsilon^2]$. Conclua a forma **sanduíche** $\sqrt{n}(\mathbf{b}_{\text{2SLS}} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0},\ \mathbf{H}^{-1}\Sigma\,\mathbf{H}^{-1})$.

(c) Mostre que, sob **homocedasticidade** ($\Omega = \sigma^2\mathbf{Q}_{zz}$), o sanduíche colapsa em $\sigma^2\mathbf{H}^{-1}$, e que no caso exatamente identificado ($\ell = k$) os fatores $\mathbf{Q}_{zz}$ se cancelam. Qual a consequência prática para os erros-padrão que se deve reportar?

4. **Instrumentos fracos (um instrumento, um regressor endógeno).**

   Modelo escalar $y_i = \beta x_i + \varepsilon_i$ com um instrumento z_i e primeiro estágio $x_i = \pi z_i + v_i$.

   (a) Mostre que $\hat{\beta}_{\text{IV}} - \beta = \dfrac{\frac{1}{n}\sum_i z_i\varepsilon_i}{\frac{1}{n}\sum_i z_i x_i}$. Identifique o papel do **numerador** (exogeneidade) e do **denominador** (relevância), e explique por que, se o denominador estiver próximo de zero, o estimador fica **instável**.

   (b) No primeiro estágio, mostre que $\text{se}(\hat{\pi}) = \sigma_v/\sqrt{\sum_i (z_i - \bar{z})^2} = \kappa/\sqrt{n}$, com $\kappa = \sigma_v/\sqrt{\text{Var}(z)}$, e conclua que a estatística t do primeiro estágio é $t \approx \sqrt{n}\,\pi/\kappa \propto \sqrt{n}\,\pi$.

   (c) Conclua que "instrumento fraco" é a situação em que $\sqrt{n}\,\pi$ é **pequeno** — propriedade de π **e** de n em conjunto. Explique, a partir da razão em (a), por que um primeiro estágio fraco torna a distribuição amostral do IV dispersa e **não-normal** em amostra finita, e por que "mais dados" não resolve quando π é genuinamente minúsculo no n que se tem.

5. **Monte Carlo como método de investigação.**

   (a) Descreva o que é um experimento de Monte Carlo e o **procedimento** que o define — desenhar um DGP **conhecido**, gerar R amostras independentes dele, calcular o estimador em cada uma e resumir a distribuição resultante. Em que sentido ele torna **visível** uma propriedade já provada em teoria?

   (b) Explique como estimar, por Monte Carlo, três grandezas: **(i)** o viés de um estimador; **(ii)** a sua variância; e **(iii)** a **taxa de cobertura** de um intervalo de confiança de nível nominal 95% — a fração das R réplicas em que o intervalo contém o parâmetro verdadeiro.

6. **O bootstrap: princípio do plug-in, variantes e a média.**

   (a) Enuncie o **princípio do plug-in**: o bootstrap substitui a distribuição populacional desconhecida F por qual objeto? Em que sentido a reamostragem **imita** a amostragem original?

   (b) Contraste os bootstraps de **pares**, **residual** e **selvagem** (*wild*): que estrutura do DGP cada um preserva ou impõe, e por que o *wild* é o apropriado sob heterocedasticidade?

   (c) Defina o **erro-padrão bootstrap** e os intervalos **percentil** e **percentil-t**. Por que o percentil-t costuma ter cobertura mais acurada?

   (d) Sejam $X_1, \ldots, X_n$ i.i.d. com média $\theta = \mathbb{E}[X]$ e variância σ^2, $\hat{\theta} = \bar{X}$. O bootstrap não-paramétrico reamostra com reposição da distribuição empírica $\widehat{F}_n$; sejam $\mathbb{E}^*, \text{Var}^*$ a esperança e variância sobre a reamostragem, dado o dado. Mostre que $\text{Var}^*[\bar{X}^*] = \frac{1}{n}\cdot\frac{1}{n}\sum_i (X_i - \bar{X})^2 = \hat{\sigma}^2/n$, conclua que o erro-padrão bootstrap de $\bar{X}$ é $\sqrt{\hat{\sigma}^2/n}$, e explique em que sentido isso **é** o princípio do plug-in — e por que o bootstrap **não conserta** um estimador inconsistente.