Econometria I

Variáveis instrumentais

Luis A. F. Alvarez

23 de abril de 2025

Instrumento

- Considere um modelo linear causal da forma:

$$Y = X\beta + U\,,$$

  onde X é uma causa observada escalar, U são causas não observadas, e β é o efeito causal de X.
- Considere uma situação em que não é razoável supor que $\text{cov}(X, U) = 0$.
  - Nesse caso, a inclinação de X no melhor preditor linear de Y em X e um intercepto não identificará β, de modo que o estimador de MQO de Y em X e um intercepto não estimará consistentemente o efeito causal de interesse.
- Suponha que observemos uma variável Z que satisfaz:
  - (Relevância) $\text{cov}(Z, X) \neq 0$.
  - (Exogeneidade ou exclusão) $\text{cov}(Z, U) = 0$
- À variável Z damos o nome de instrumento:
  - Trata-se de variável que exibe associação, na população, com X (relevância), e cuja **única associação com Y se dá através de X** (exogeneidade ou exclusão).

Estimando de Wald e Identificação sob variável instrumental

- Defina **o estimando de Wald** como o **parâmetro**:

$$\gamma_{\text{Wald}} := \frac{\text{cov}(Z, Y)}{\text{cov}(Z, X)}\,,$$

- Sob as hipóteses de relevância e exogeneidade do instrumento, estimando identifica β, i.e.:

$$\gamma_{\text{Wald}} = \beta\,.$$

Estimação e Inferência

- Dada uma amostra aleatória $(Y_i, X_i, Z_i) \sim (Y, X, Z)$, resultado anterior sugere que estimemos o efeito causal como:

$$\hat{\gamma}_{\text{Wald}} = \frac{\widehat{\text{cov}(Z, Y)}}{\widehat{\text{cov}(Z, X)}} = \frac{\hat{b}_{Y,Z}}{\hat{b}_{X,Z}},$$

onde $\hat{b}_{S,Z}$ é o estimador de MQO para o coeficiente de Z numa regressão de S num intercepto e Z.

- Sob condições de regularidade da aula anterior, vimos que, com $n \to \infty$

$$\hat{b}_{Y,Z} \xrightarrow{p} \frac{\text{cov}(Y, Z)}{\mathbb{V}(Z)}$$

$$\hat{b}_{X,Z} \xrightarrow{p} \frac{\text{cov}(X, Z)}{\mathbb{V}(Z)}$$

de onde segue, por aplicação do teorema do mapa contínuo que, sob as condições de relevância e exogeneidade:

$$\hat{\gamma}_{\text{Wald}} \xrightarrow{p} \gamma_{\text{Wald}} = \beta\,.$$

Testando relevância

- Note que, como estimador de $\hat{b}_{X,Z}$ é consistente para $\frac{\text{cov}(X,Z)}{\mathbb{V}(Z)}$, podemos utilizar esse estimador para realizar um teste da nula de que que o instrumento **não** é relevante.
- Por outro lado, em nosso ambiente com uma variável instrumental e um modelo causal linear, a hipótese de exclusão é **intestável**.
  - Devemos suplementar a análise empírica com uma argumentação de porquê o único mecanismo através do qual variações em Z produzem variações em Y é através de X.

Caso geral

- Vamos agora considerar um modelo causal geral da forma:

$$Y = X'\beta + U$$

onde X é um vetor de k causas observadas.

- Vamos supor a existência de um vetor Z de l instrumentos que satisfaz as hipóteses:

Hipótese (H1-Relevância)

$\mathbb{E}[ZX']$ *tem posto* k.

Hipótese (H2-Exogeneidade)

$\mathbb{E}[ZU] = 0$

- Note que a hipótese de exogeneidade é equivalente a $\text{cov}(Z, U) = 0$ quando X inclui um intercepto, pois nesse caso podemos supor que U tem média zero "absorvendo sua média" ao intercepto.
  - Note, no entanto, que isso muda a interpretação do intercepto, que supomos não ser de interesse direto no caso.

Variáveis Endógenas e Exógenas

- A condição de relevância implica que necessitamos de $l \geq k$ variáveis que exibam suficiente variação com as causas X, e que não exibam associação com as causas U.
    - Se uma entrada X_j é exógena por hipótese, i.e. $\mathbb{E}[X_j U] = 0$, podemos incluí-la entre os instrumentos Z.
    - Por outro lado, para cada variável X_s endógena, i.e. tal que potencialmente $\mathbb{E}[X_s U] \neq 0$, precisamos de pelo menos um instrumento que exiba associação com as causas observadas, mas não com as causas não observadas.

Identificação dos efeitos no modelo linear geral

- Sob as hipóteses H1 e H2, para qualquer matriz A $k \times l$ de posto cheio, temos que:

$$\gamma_A := (\mathbb{E}[AZX'])^{-1}\mathbb{E}[AZY] = \beta$$

- Note que, o resultado anterior sugere que, para uma amostra aleatória $(Y_i, X_i, Z_i) \sim (Y, X, Z)$ e uma dada matriz A, podemos estimar β por:

$$\hat{\gamma}_A = \left(\sum_{i=1}^{n} \hat{A} Z_i X_i'\right)^{-1} \sum_{i=1}^{n} \hat{A} Z_i Y_i = (\hat{A}\boldsymbol{Z}'\boldsymbol{X})^{-1}\hat{A}\boldsymbol{Z}'\boldsymbol{y}\,,$$

onde $\hat{A}$ é um estimador da matriz A e:

$$\boldsymbol{Z} = \begin{bmatrix} Z_1' \\ Z_2' \\ \vdots \\ Z_n' \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Mínimos Quadrados em Dois Estágios

- Note que, em termos de **identificação**, temos alguma liberdade na escolha de $\hat{A}$.
- Em termos **inferenciais**, uma escolha natural é tomar a combinação linear dos instrumentos $\hat{A}$ que "maximiza"o poder explicativo de $\boldsymbol{Z}$ sobre $\boldsymbol{X}$. Isto é, tomamos:

$$\hat{A}' = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{X})$$

- Essa escolha produz o estimador de MQO em dois estágios, que denotamos por $\hat{b}_{2\text{SLS}}$:

$$\hat{b}_{2\text{SLS}} = (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = \left(\boldsymbol{X}'P_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'P_{\boldsymbol{Z}}\boldsymbol{y}$$

onde $P_{\boldsymbol{Z}}$ é a matriz de projeção de $\boldsymbol{Z}$.
- Estimador pode ser obtido em dois estágios
  - Primeiro, regredimos cada uma das $v = 1\,2\ldots,k$ variáveis em$\boldsymbol{X}$, $X_{i,v}$, em Z_i, e guardamos os valores preditos $\hat{X}_{i,v} = \hat{c}_v' Z_i$.
    - Note que $\hat{X}_{i,v} = X_{i,v}$ se X_v é variável inclusa em Z.
  - Usando os valores preditos, regredimos Y_i nos $\hat{X}_{i,v}, v = 1\,l\ldots,k$. O estimador de $\hat{b}_{2\text{SLS}}$ será o resultado dessa regressão.

Propriedades Assintóticas do 2SLS

Para a análise do estimador de MQO em dois estágios requeremos:

Hipótese (H3-Posto)

$\mathbb{E}[ZZ']$ *tem posto cheio.*

Proposição

Sob as hipóteses H1-H3, amostragem aleatória $(Y_i, X_i, Z_i) \overset{iid}{\sim} (Y, X, Z)$ *e segundos momentos finitos de* (Y, X, Z), $\hat{b}_{2SLS} \xrightarrow{p} \beta$.

Proposição

Sob as hipóteses H1-H3, amostragem aleatória $(Y_i, X_i, Z_i) \overset{iid}{\sim} (Y, X, Z)$ *e* ***quartos*** *momentos finitos de* (X, Y, Z),

$$\sqrt{n}(\hat{b}_{2SLS} - \beta) \xrightarrow{d} \mathcal{N}(0, BMB)\,,$$

$$B = \left(\mathbb{E}[XZ']\mathbb{E}[ZZ']^{-1}\mathbb{E}[X'Z]\right)^{-1}\,,$$

$$M = \mathbb{E}[XZ']\mathbb{E}[ZZ']^{-1}\mathbb{E}[U^2 ZZ']\mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX']$$

Eficiência do MQO em dois estágios

A escolha de pesos A implícita ao estimador de MQO em dois estágios possui propriedades de eficiência assintótica, na classe de estimadores baseados em combinações lineares dos instrumentos, se uma versão da hipótese de homocedasticidade valer.

Proposição

Sob as hipóteses H1-H3, amostragem aleatória $(Y_i, X_i, Z_i) \overset{iid}{\sim} (Y, X, Z)$*, quartos momentos finitos de* (X, Y, Z)*, e a condição:*

$$\mathbb{E}[U^2|Z] = \sigma^2 \, .$$

Então o estimador de 2SLS exibe a menor variância assintótica, relativamente a qualquer outro estimador $\hat{\gamma}_C$ *baseado numa combinação linear* $\hat{C}\boldsymbol{Z}$ *dos instrumentos, onde* $\hat{C} \xrightarrow{p} C$*. Em outras palavras:*

$$\text{Avar}(\sqrt{n}(\hat{\gamma}_C - \beta)) - \text{Avar}(\sqrt{n}(\hat{b}_{2SLS} - \beta)) \textit{ é positiva semidefinida}$$

Teste de relevância

- Assim como no estimador de Wald, é possível usar o primeiro estágio para testar a hipótese de relevância de que $\mathbb{E}[ZX']$ tem posto k.
  - Se só há somente uma variável endógena em X, só há efetivamente um primeiro estágio a ser rodado (por quê?), e podemos testar relevância realizando um teste de Wald de que os coeficientes de primeiro estágio associados aos instrumentos excluídos (isto é, que não constam entre as variáveis exógenas em X) são todos iguais a zero.
  - Se há mais de uma endógena, temos de fazer um teste de Wald multi-equação, que pode ser construído a partir do "empilhamento" dos diferentes primeiros estágios (veja o livro do Wooldridge para mais discussão).

Teste de endogeneidade

- Se $Z \neq X$, é possível construir um teste da hipótese nula:

$$\text{cov}(X, U) = 0$$

contra a alternativa de que $\text{cov}(X, U) \neq 0$.

- Ideia é basear teste em estatística de Wald que meça a "magnitude" de:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - X_i'\hat{b}_{2SLS})$$

  - Observe que, se $Z = X$, $\hat{b}_{2SLS} = \hat{b}_{MQO}$, e essa magnitude sempre seria zero (teste teria poder trivial).

- Não é difícil ver que a magnitude da covariância entre o resíduo do estimador de 2SLS e X depende crucialmente da diferença:

$$\hat{b}_{MQO} - \hat{b}_{2SLS}\,.$$

  - Aplicação do princípio de Hausman para a construção de testes, em que comparamos dois estimadores, sendo que um deles é consistente sob a nula e inconsistente sob a alternativa (MQO), enquanto o outro é consistente sob a nula e a alternativa (2SLS) (ver Wooldridge).

Teste de restrições sobreidentificadoras

- Quando $l > k$, é possível construir um teste para a hipótese nula:

$$\mathbb{E}[Zu] = 0$$

contra a alternativa de que

$$\mathbb{E}[Zu] \neq 0$$

com base em uma estatística de Wald que meça a magnitude de:

$$\frac{1}{n} \sum_{i=1}^{n} Z_i (Y_i - X_i' \hat{b}_{2SLS})$$

  - Se $l = k$, essa magnitude é igual a zero, e poder do teste seria trivial.
- Teste baseado no fato de que há mais instrumentos que o necessário para a identificação (restrições sobreidentificadoras).
  - Rejeição da nula não aponta **qual** instrumento é inválido.
  - Ver Wooldridge para a construção do teste.

INTERPRETAÇÃO SOB HETEROGENEIDADE

- Os resultados de identificação apresentados anteriormente suponham um modelo causal linear.
- Imbens e Angrist (1994) estudam a interpretação do estimando de Wald sob um modelo causal heterogêneo.
- Vamos considerar um resultado de interesse Y, causa binária $D \in \{0, 1\}$ cujo efeito queremos aferir, e um instrumento binário $Z \in \{0, 1\}$.
- Definimos os seguintes resultados potenciais, associados a manipulações hipotéticas das causas e dos instrumentos:

$$Y(d, z), \quad (d, z) \in \{0, 1\}^2$$
$$D(z), \quad z \in \{0, 1\}^2$$

- As variáveis observáveis são dadas por:

$$D = ZD(1) + (1 - Z)D(0)\,,$$

$$Y = ZDY(1,1) + (1-Z)DY(1,0) + Z(1-D)Y(0,1) + (1-Z)(1-D)Y(0,0)$$

Hipóteses sobre o modelo causal heterogêneo

- Imbens e Angrist (1994) supõem as seguintes hipóteses sobre o modelo causal:
  1. **Exclusão:**
  $$Y(d,0) = Y(d,1) =: Y(d)\,, \quad d \in \{),1\}$$
  .
  2. **Independênca** Z é independente de $\{Y(0), Y(1), D(1), D(0)\}$.
  3. **Relevância**:
  $$\frac{\mathrm{cov}(D,Z)}{\mathrm{var}(Z)} = \mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0] \neq 0$$
  .
  4. Monotonicidade: ou $\mathbb{P}[D(1) \geq D(0)] = 1$, ou $\mathbb{P}[D(1) \leq D(0)] = 1$.
  - Hipótese de monotonicidade requer que tratamento exerça um efeito unidirecional sobre o tratamento D na população de interesse.
    - Ou manipulação de $Z = 0$ para $Z = 1$ aumenta (fracamente) o valor de D para todos os indivíduos na população, ou diminui.

Imbens e Angrist (1994)

Proposição

Sob as hipótese 1-4, estimando de Wald:

$$\gamma_{Wald} = \frac{\text{cov}(Y, Z)}{\text{cov}(D, Z)} = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]}$$

é igual ao efeito médio do tratamento na subpopulação induzida ao tratamento pelo instrumento (compliers), isto é:

$$\gamma_{Wald} = \frac{\mathbb{E}[(Y(1) - Y(0))\mathbf{1}_{D(1)\neq D(0)}]}{\mathbb{P}[D(1) \neq D(0)]} = \mathbb{E}[Y(1) - Y(0)|D(1) \neq D(0)]$$

- Estimando identifica um efeito médio de tratamento local (LATE).
- Existem extensões para instrumentos discretos (Imbens e Angrist, 1994), tratamentos ordenados (Angrist e Imbens, 1995), tratamentos contínuos Angrist, Graddy e Imbens (2000), especificações com controles (2SLS) (Słoczyński, 2022; Blandhol et al., 2022), instrumentos contínuos (Alvarez e Toneto, 2024).

Alvarez, Luis A.F. e Rodrigo Toneto (2024). “The interpretation of 2SLS with a continuous instrument: A weighted LATE representation”. Em: *Economics Letters* 237, p. 111658. ISSN: 0165-1765. DOI: https://doi.org/10.1016/j.econlet.2024.111658. URL: https://www.sciencedirect.com/science/article/pii/S0165176524001411.

Angrist, Joshua D., Kathryn Graddy e Guido W. Imbens (2000). “The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish”. Em: *The Review of Economic Studies* 67.3, pp. 499–527. ISSN: 00346527, 1467937X. URL: http://www.jstor.org/stable/2566964 (acesso em 16/10/2023).

Angrist, Joshua D. e Guido W. Imbens (1995). “Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity”. Em: *Journal of the American Statistical Association* 90.430, pp. 431–442. DOI: 10.1080/01621459.1995.10476535.

Blandhol, Christine et al. (2022). *When is TSLS actually late?* Rel. técn. National Bureau of Economic Research.

Imbens, Guido W. e Joshua D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". Em: *Econometrica* 62.2, pp. 467–475. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2951620 (acesso em 13/10/2023).

Słoczyński, Tymon (2022). *When Should We (Not) Interpret Linear IV Estimands as LATE?* arXiv: 2011.06695 [econ.EM].