Dados em Painel e Causalidade Usando R
- Aula 6-

Prof. Mestre. Omar Barroso Khodr

Instituto Brasileiro de Educação, Pesquisa e Desenvolvimento

- **Variáveis Instrumentais**
- **Regressão em dois Estágios**

O Estimador VI com um Único Regressor e um Único Instrumento

- Iniciamos a discussão sobre regressão de variáveis instrumentais (VI) com o caso mais simples de apenas um regressor e apenas uma variável instrumental.
- Vamos considerar o modelo de regressão linear simples para dados transversais:
- $y = \beta_0 + \beta_1 x_1 + u$
- Em regressões, muitas vezes há variáveis omitidas ou erros de medição, o que causa viés de variável omitida ou viés de simultaneidade (causalidade reversa).

O Estimador VI com um Único Regressor e um Único Instrumento

- $y = \beta_0 + \beta_1 x_1 + u$
- Por exemplo, Se uma variável explicativa (X) está correlacionada com o erro (u) na regressão o estimador de MQO (mínimos quadrados ordinários) é viesado e inconsistente.
- A ideia é encontrar uma variável instrumental (Z)...
- Ou seja, Z deve estar correlacionada com X (a variável endógena problemática).
- Nesse mesmo sentido, Z não pode ser correlacionada com o erro (u) (ou seja, afeta Y apenas por meio de X). $\hat{\beta}_1^{IV} = \frac{Cov(z,y)}{Cov(z,x)}$

Condições

- **Condição de Relevância do Instrumento:**
- X e seu instrumento Z devem estar correlacionados:
- $\rho Z_i, x_i \neq 0$
- **Condição de exogeneidade do instrumento:**
- O instrumento Z não deve estar correlacionado com o termo de erro.
- $u: \rho Z_i, u_i = 0$

Intuições

- Se Z afeta X e Z não tem efeito direto em Y (exceto via X), então podemos usar Z para "isolar" a parte de X que não é correlacionada com o erro.
- Isso permite estimar o efeito causal de X em Y sem o viés.
- Imagine que queremos medir o efeito da educação (X) no salário (Y), mas **há habilidade não observada (u)** que afeta ambos (viés).

Intuições

- Se usarmos distância até a escola (Z) como instrumento...
- Z afeta X (pessoas mais perto da escola tendem a estudar mais).
- Z não afeta Y diretamente (a distância só influencia o salário via educação).
- Assim, comparamos grupos com diferentes Z para estimar o efeito limpo de X em Y.

Exemplo R

- Nesse contexto, vamos identificar o retorno (financeiro) da educação de mulheres casadas (exemplo Wooldridge 15.1).
- Como instrumento, usaremos a educação de seu pai(s).
- Primeiramente, vamos calcular o MQO e o parâmetro de coeficiente angular do instrumento ($\hat{\beta}_1^{IV} = \frac{Cov(z,y)}{Cov(z,x)}$)...
- Nossa relação, ficaria como uma condicionante...
- $\log(salário) \sim (edu|\ edu.pai)$

Exemplo R (Resultados)

- O MQO sugere que um ano a mais de educação aumenta o salário em ~10.9% (0.109).
- Todavia, a VI estima um efeito menor: ~5.9% (0.059).
- Isso indica que o MQO provavelmente superestima o efeito devido a viés de endogeneidade (ex.: habilidade não observada correlacionada com educação e salário)...

	Dependent variable:	
	log(wage)	
	OLS	*instrumental variable*
	(1)	(2)
educ	0.109^{***}	0.059^{*}
	(0.014)	(0.035)
Constant	-0.185	0.441
	(0.185)	(0.446)
Observations	428	428
R^2	0.118	0.093
Adjusted R^2	0.116	0.091
Residual Std. Error (df = 426)	0.680	0.689
F Statistic	56.929^{***} (df = 1; 426)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Exemplo R (Resultados)

- O erro padrão do IV é maior (0.035 vs. 0.014 no OLS), reduzindo significância estatística (apenas 10% vs. 1% no OLS).
- Isso é típico em IV: trade-off entre viés e eficiência.
- O IV explica menos variação em log(wage) (R^2 = 0.093 vs. 0.118 no MQO), o que é esperado, pois IV usa apenas a variação em educ explicada pelo instrumento.

	Dependent variable:	
	log(wage)	
	OLS	*instrumental variable*
	(1)	(2)
educ	0.109***	0.059*
	(0.014)	(0.035)
Constant	-0.185	0.441
	(0.185)	(0.446)
Observations	428	428
R^2	0.118	0.093
Adjusted R^2	0.116	0.091
Residual Std. Error (df = 426)	0.680	0.689
F Statistic	56.929*** (df = 1; 426)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Mínimos Quadrados em Dois Estágios

- Mínimos Quadrados em Dois Estágios (MQ2) é uma abordagem geral para estimativa de VI quando temos um ou mais regressores endógenos e pelo menos o mesmo número de variáveis instrumentais adicionais. Considere o modelo de regressão:
- $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1$
- Se os regressores y_2 e y_3 são potencialmente correlacionados com o termo de erro u_1, assumimos que os regressores z_1, z_2 e z_3 são exógenos.
- Como temos dois regressores endógenos, precisamos de pelo menos duas variáveis instrumentais adicionais (z_4e z_5).

Mínimos Quadrados em Dois Estágios

- O nome "mínimos quadrados em dois estágios" vem do fato de que ele pode ser realizado em dois estágios de regressões MQO:
- **Primeiro Estágio (1):**
- Regredir y_2 em z_1 até z_5 e regredir y_3 em z_1 até z_5 . Assim obtendo os valores estimados $\hat{y}_2$ e $\hat{y}_3$.
- $y_2 = \beta_0 + \beta_2 z_1 + \cdots + \beta_7 z_5 + u_1 \rightarrow \hat{y}_2$
- $y_3 = \beta_0 + \beta_2 z_1 + \cdots + \beta_7 z_5 + u_1 \rightarrow \hat{y}_3$

Mínimos Quadrados em Dois Estágios

- **Segundo Estágio (2):**
- Regredir y_1 em $\hat{y}_2$, $\hat{y}_3$, e z_1 até z_3. Assim obtendo os valores estimados.
- $y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 \hat{y}_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1$
- Se os instrumentos forem válidos, isso fornecerá estimativas consistentes dos parâmetros b0 a b5.

Dois Estágios no R

- Podemos facilmente realizar as estimativas com o R....
- Continuaremos analisando o exemplo anterior....
- Desta vez, queremos investigar os resultados do retorno sob educação de mulheres, considerando os instrumentos: educação paterna e materna.
- Primeiro, fazemos ambos os estágios manualmente, incluindo a educação ajustada como ajustada (estágio 1) como regressores no segundo estágio.

Exemplo 2SLS

- Avaliando o primeiro estágio...
- Coluna (1)...
- Cada ano adicional de educação da mãe aumenta a educação do filho(a) em 0.158 anos.
- Cada ano adicional de educação do pai aumenta a educação do filho(a) em 0.190 anos.
- Experiência não tem efeito claro na educação. Nesse mesmo contexto, a experiência ao quadrado.
- Os Instrumentos (motheduc e fatheduc) explicam 21.1% da variação em educ.

	Dependent variable:	
	educ	log(wage)
	(1)	(2)
fitted(stage1)		0.061*
		(0.033)
exper	0.045	0.044***
	(0.040)	(0.014)
I(exper2)	-0.001	-0.001**
	(0.001)	(0.0004)
motheduc	0.158***	
	(0.036)	
fatheduc	0.190***	
	(0.034)	
Constant	9.103***	0.048
	(0.427)	(0.420)
Observations	428	428
R^2	0.211	0.050
Note:	*p<0.1; **p<0.05; ***p<0.01	

Exemplo 2SLS

- Avaliando o segundo estágio...
- Coluna (2)...
- Cada ano adicional de educação (predito pelos instrumentos) aumenta o salário em 6.1% (efeito causal).
- Cada ano de experiência aumenta o salário em 4.4%.
- Efeito decrescente da experiência ao quadrado (curva convexa).
- Modelo explica apenas 5% da variação em salários (comum em VI, pois usa apenas variação exógena).

	Dependent variable:	
	educ	log(wage)
	(1)	(2)
fitted(stage1)		0.061^{*}
		(0.033)
exper	0.045	0.044^{***}
	(0.040)	(0.014)
I(exper2)	-0.001	-0.001^{**}
	(0.001)	(0.0004)
motheduc	0.158^{***}	
	(0.036)	
fatheduc	0.190^{***}	
	(0.034)	
Constant	9.103^{***}	0.048
	(0.427)	(0.420)
Observations	428	428
R^2	0.211	0.050
Note:	*p<0.1; **p<0.05; ***p<0.01	

Testando a Exogeneidade dos Regressores

- Há outra maneira de obter as mesmas estimativas dos parâmetros do VI que nos dois estágios. Na mesma configuração anterior, esta "função de controle" também consiste em dois estágios:
- (1) Como nos dois estágios, regredir y_2 e y_3 em z_1 a z_5 . Obter os resíduos $\widehat{v_2}$ e $\widehat{v_3}$ em vez dos valores ajustados para os y.
- (2) Regredir y_1 em y_2, y_3 , z_1 , z_2 , z_3 e os resíduos do primeiro estágio $\widehat{v_2}$ e $\widehat{v_3}$.

Testando a Exogeneidade dos Regressores

- Esta abordagem é simples de implementar, assim como os dois estágios, e também resultará nas mesmas estimativas de parâmetros e erros-padrão de MQO inválidos no segundo estágio (a menos que os regressores duvidosos y2 e y3 sejam de fato exógenos).
- Após esta regressão do segundo estágio, podemos testar a exogeneidade de forma simples, assumindo que os instrumentos são válidos.
- Precisamos apenas realizar um teste t ou F da hipótese nula de que os parâmetros dos resíduos do primeiro estágio são iguais a zero. Se rejeitarmos essa hipótese, isso indica endogeneidade de y2 e y3.

Continuando para o Exemplo Anterior

- Continuamos utilizando a educação dos pais como instrumentos.
- O primeiro estágio continua idêntico ao exemplo anterior.
- O segundo estágio adiciona os resíduos do primeiro estágio para lista original dos regressores.

Resultado

- Incluimos os resíduos do primeiro estágio (resid(est.1)) como uma variável adicional no segundo estágio.
- O coeficiente de resid(est.1) é marginalmente significativo (p-valor = 0.095).
- Se p < 0.05: Rejeita-se a hipótese nula de exogeneidade (ou seja, educ é endógeno).
- Se p ≥ 0.05: Não há evidência forte para rejeitar a exogeneidade.

```
t test of coefficients:

                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   0.04810030  0.39457526  0.1219 0.9030329
educ          0.06139663  0.03098494  1.9815 0.0481824 *
exper         0.04417039  0.01323945  3.3363 0.0009241 ***
I(exper^2)   -0.00089897  0.00039591 -2.2706 0.0236719 *
resid(est.1)  0.05816661  0.03480728  1.6711 0.0954406 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resultado

- p = 0.095 (10% de significância), o que sugere:
- Evidência fraca de endogeneidade em educ.
- O modelo OLS pode não ser severamente viesado, mas o IV ainda é preferível para estimativas causais.

```
t test of coefficients:

                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   0.04810030  0.39457526  0.1219 0.9030329
educ          0.06139663  0.03098494  1.9815 0.0481824 *
exper         0.04417039  0.01323945  3.3363 0.0009241 ***
I(exper^2)   -0.00089897  0.00039591 -2.2706 0.0236719 *
resid(est.1)  0.05816661  0.03480728  1.6711 0.0954406 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resultado

- No modelo MQO, o coeficiente de educ era 0.109 (superestimado).
- No modelo VI/2SLS, o coeficiente cai para 0.061, indicando que o OLS sofria de viés para cima devido a endogeneidade (ex.: habilidade não observada).

```
t test of coefficients:

                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   0.04810030  0.39457526  0.1219 0.9030329
educ          0.06139663  0.03098494  1.9815 0.0481824 *
exper         0.04417039  0.01323945  3.3363 0.0009241 ***
I(exper^2)   -0.00089897  0.00039591 -2.2706 0.0236719 *
resid(est.1)  0.05816661  0.03480728  1.6711 0.0954406 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Variáveis Instrumentais com Dados em Painel

- VI’s podem ser utilizadas em dados em painel. Dessa maneira, podemos retirar a constante de tempo [sob heterogeneidade].
- Fazemos isso, diferenciando ou sob transformações e assim ‘fixamos’ os problemas de endogeneidade com VI’s.
- Conforme sabemos o método MQO e IV o mesmo é aplicado neste contexto.

Variáveis Instrumentais com Dados em Painel

- Exemplo: Retorno de treinamento [linha de produção] sob **taxa de sucateamento** (A porcentagem de materiais ou produtos que são descartados devido a defeitos, erros ou retrabalho durante a produção).
- Nesse exemplo específico, queremos ver essa relação para os anos 1987 e 1988 na base de dados.
- Utilizaremos as variáveis [índices] fcode [código de barra] e year[ano].
- Assim, podemos estimar os parâmetros utilizando o primeiro estágio com VI grant [financiamento].

Variáveis Instrumentais com Dados em Painel

- **Objetivo:** Analisar a relação entre horas de treinamento por empregado (hrsemp) e log do índice de desperdício (log(scrap)) em empresas, usando financiamento (grant) como instrumento para hrsemp.
- **Dados:** Painel desbalanceado (47 empresas, anos 1987-1988, 45 observações usadas).
- **Método:**
- Primeiras Diferenças (FD): Controla efeitos fixos não observados (ex.: características constantes das empresas).
- Variável Instrumental (VI): grant (financiamento) é usado para corrigir endogeneidade em hrsemp.

Variáveis Instrumentais com Dados em Painel

- coeficiente de hrsemp:
- Estimativa: -0.014
- Interpretação: Um aumento de 1 hora em treinamento por empregado reduz o desperdício em 1.4% (em média, nos anos 1987-1988).
- Significância: Marginal (p-valor = 0.074, significativo a 10%).
- Sugere que o efeito é estatisticamente relevante, mas não robusto a níveis convencionais (5%).

```
 summary(plm(log(scrap)~ hrsemp| grant, model = "fd", data = jtrain.p))
Oneway (individual) effect First-Difference Model
Instrumental variable estimation
   (Balestra-Varadharajan-Krishnakumar's transformation)

Call:
plm(formula = log(scrap) ~ hrsemp | grant, data = jtrain.p, model = "fd")

Unbalanced Panel: n = 47, T = 1-2, N = 92
Observations used in estimation: 45

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-2.3088293 -0.2188848 -0.0089255  0.2674362  2.4305637

Coefficients:
             Estimate Std. Error z-value Pr(>|z|)
(Intercept) -0.0326684  0.1269512 -0.2573  0.79692
hrsemp      -0.0141532  0.0079147 -1.7882  0.07374 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    17.29
Residual Sum of Squares: 17.015
R-Squared:      0.061927
Adj. R-Squared: 0.040112
Chisq: 3.19767 on 1 DF, p-value: 0.073743
> View(jtrain.p)
>
```

Variáveis Instrumentais com Dados em Painel

- Estimativa: -0.033 (não significativa, p = 0.797).
- Indica que, sem horas de treinamento, não há mudança sistemática no desperdício.
- 3. Diagnósticos:
- R^2: 0.062 (baixo poder explicativo, comum em modelos FD/VI).
- Teste de Significância Global (Chisq): p-valor = 0.074 → Modelo é marginalmente significativo.

```
 summary(plm(log(scrap)~ hrsemp| grant, model = "fd", data = jtrain.p))
Oneway (individual) effect First-Difference Model
Instrumental variable estimation
   (Balestra-Varadharajan-Krishnakumar's transformation)

Call:
plm(formula = log(scrap) ~ hrsemp | grant, data = jtrain.p, model = "fd")

Unbalanced Panel: n = 47, T = 1-2, N = 92
Observations used in estimation: 45

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-2.3088293 -0.2188848 -0.0089255  0.2674362  2.4305637

Coefficients:
             Estimate Std. Error z-value Pr(>|z|)
(Intercept) -0.0326684  0.1269512 -0.2573  0.79692
hrsemp      -0.0141532  0.0079147 -1.7882  0.07374 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    17.29
Residual Sum of Squares: 17.015
R-Squared:      0.061927
Adj. R-Squared: 0.040112
Chisq: 3.19767 on 1 DF, p-value: 0.073743
> View(jtrain.p)
>
```

Variáveis Instrumentais com Dados em Painel

- Pressuposto: grant afeta log(scrap) apenas via hrsemp (não diretamente).
- Se grant influenciar desperdício por outros canais (ex.: infraestrutura), o IV é inválido.
- Se o MQO e VI diferirem, há evidência de endogeneidade em hrsemp (ex.: empresas com mais desperdício investem mais em treinamento, viés de simultaneidade).

```
 summary(plm(log(scrap)~ hrsemp| grant, model = "fd", data = jtrain.p))
Oneway (individual) effect First-Difference Model
Instrumental variable estimation
   (Balestra-Varadharajan-Krishnakumar's transformation)

Call:
plm(formula = log(scrap) ~ hrsemp | grant, data = jtrain.p, model = "fd")

Unbalanced Panel: n = 47, T = 1-2, N = 92
Observations used in estimation: 45

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-2.3088293 -0.2188848 -0.0089255  0.2674362  2.4305637

Coefficients:
             Estimate Std. Error z-value Pr(>|z|)
(Intercept) -0.0326684  0.1269512 -0.2573  0.79692
hrsemp      -0.0141532  0.0079147 -1.7882  0.07374 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    17.29
Residual Sum of Squares: 17.015
R-Squared:      0.061927
Adj. R-Squared: 0.040112
Chisq: 3.19767 on 1 DF, p-value: 0.073743
> View(jtrain.p)
>
```

Bibliografia

- Wooldridge, J.M. (2013) Introductory econometrics: a modern approach. 5th ed. Michigan State University.