Econometria e Séries Temporais
- Aula 14 -

Prof. Mestre. Omar Barroso

Instituto Brasileiro de Educação, Pesquisa e Desenvolvimento

Variáveis Instrumentais $\leftrightarrow$ Ivs (tópico especial)

- *Amplamente conhecidas pela literatura pelo seu termo em inglês (Instrumental Variables ou Ivs).*
- Modelos de regressão podem sofrer de problemas como **variáveis omitidas**, erros de medição e causalidade simultânea.
- Nesse sentido, **o termo de erro** é correlacionado com a variável de interesse e, portanto, o coeficiente correspondente é estimado de forma inconsistente.

Variáveis Instrumentais $\leftrightarrow$ IVs

- No entanto, se os fatores omitidos não puderem ser medidos ou não estiverem disponíveis por outros motivos, a regressão múltipla não pode resolver o problema.
- O mesmo problema surge se houver causalidade simultânea. Quando a causalidade vai de X para Y e vice-versa, haverá um viés de estimativa que não pode ser corrigido pela regressão múltipla.

Variáveis Instrumentais $\leftrightarrow$ IVs

- Uma técnica geral para obter um estimador consistente do coeficiente de interesse é a regressão de variáveis instrumentais (IV).
- Hoje, focaremos na ferramenta de regressão IV chamada mínimos quadrados de dois estágios (TSLS).

Variáveis Instrumentais $\leftrightarrow$ IVs

- Uma técnica geral para obter um estimador consistente do coeficiente de interesse é a regressão de variáveis instrumentais (IV).
- Hoje, focaremos na ferramenta de regressão IV chamada mínimos quadrados de dois estágios (TSLS).

O Estimador IV com um regressor e um instrumento

- Considere a seguinte regressão.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \ , \ \ i = 1, \ldots, n$$

- onde o termo de erro u_i é correlacionado com a variável Xi (X é endógeno) tal que MQO é inconsistente para o verdadeiro β_1 .
- No caso mais simples, a regressão IV usa uma única variável instrumental z para obter um estimador consistente para β_1.

Condições Para Uma IV Válida

- A variável X e Z devem ser correlacionadas, ou seja, $\rho z_i x_i \neq 0$.
- O instrumento Z não pode ser correlacionado com o termo de erro. Ou seja, $\rho z_i u_i = 0$.

O estimador de mínimos quadrados em dois estágios

- Como sugerido pelo seu nome, o processo procede em dois estágios. No primeiro estágio, a variação no regressor endógeno X é decomposta em um componente sem problemas que é explicado pelo instrumento
- Z e um componente problemático que é correlacionado com o erro ui. O segundo estágio usa o componente sem problemas da variação em x para estimar β_1.

O estimador de mínimos quadrados em dois estágios

- O Modelo de regressão em primeiro estágio pode ser representado por:

$$X_i = \pi_0 + \pi_1 Z_i + \nu_i$$

- Aonde a variável π_1 é o componente de Xi que é explicado por Z.
- v_i funciona com um termo de erro que é correlacionado com u_i.

O estimador de mínimos quadrados em dois estágios

- O Modelo de regressão em primeiro estágio pode ser representado por: $X_i = \pi_0 + \pi_1 Z_i + \nu_i$

- Utilizando os estimadores de π_0 e π_1 obtemos os valores preditivos $\hat{X}$, i = 1,..., n.
- No caso de Z ser um instrumento válido, $\hat{X}_i$ se passa como uma variável sem problemas, no sentido de $\hat{X}$ é exógeno em uma regressão de Y em $\hat{X}$ no qual é obtido em uma regressão de segundo estágio.

O estimador de mínimos quadrados em dois estágios

- O segundo estágio produz $\hat{\beta}_0$ e $\hat{\beta}_1$.
- No caso de apenas um instrumento, pode se mostrar que o estimador em dois estágios de β_1 é:

$$\widehat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Z_i - \overline{Z})},$$

- No qual é a razão da covariância amostral entre Z e Y, para a covariância amostral entre Z e X.

Vamos praticar...

Consumo de Cigarros

- Vamos trabalhar com a base de dados de consumo de cigarros nos EUA.
- A questão é quanto os impostos devem ser aumentados para atingir uma certa redução no consumo de cigarros. Economistas da saúde e microeconomistas usam elasticidades para responder a esse tipo de pergunta.

Consumo de Cigarros

- Normalmente, desejamos modelar a demanda por cigarros (quantidades de cigarros consumidos) como uma função do preço.
- Mas como o preço é endógeno, usaremos impostos (taxs) como um instrumento para o preço (price).
- Também pode ser necessário transformar ou criar novas variáveis, como o preço por maço e a renda per capita, para uma melhor especificação do modelo.

Consumo de Cigarros

- Usamos o conjunto de dados CigarettesSW que vem com o pacote AER. É um conjunto de dados de painel que contém observações sobre o consumo de cigarros e vários indicadores econômicos para todos os 48 estados federais continentais dos EUA de 1985 a 1995. Seguindo as especificações de livros e artigos que antecedem essa pesquisa, consideramos dados para a seção transversal de estados apenas em 1995.

Consumo de Cigarros

- Estamos interessados em estimar β_1 em:

$$\log(Q_i^{cigarettes}) = \beta_0 + \beta_1 \log(P_i^{cigarettes}) + u_i$$

- Aonde Qi é o número dos maços de cigarro per capita vendidos.
- Pi é o preço real do maço por estado i.
- A variável instrumental que utilizaremos será o regressor endógeno log(pi) no qual representa a parcela dos impostos sobre cigarros proveniente do imposto geral sobre vendas mensurado por dólar por maço.

Consumo de Cigarros

- Estamos interessados em estimar β_1 em:

$$\log(Q_i^{cigarettes}) = \beta_0 + \beta_1 \log(P_i^{cigarettes}) + u_i$$

- A ideia é que o imposto de venda é um instrumento relevante que é incluso na média dos preços do maço após o imposto.
- Também, é possível que os impostos de venda sejam exógenos dado que conforme os impostos de venda anão influenciam a quantidade vendida diretamente sobre o preço.

Consumo de Cigarros

- Realizamos algumas transformações para obter dados de seção transversal deflacionados para o ano de 1995.
- Também calculamos a correlação amostral entre o imposto sobre vendas e o preço por maço. A correlação amostral é um estimador consistente da correlação populacional.
- A estimativa de aproximadamente de 0,614 indica que o imposto sobre vendas e Pi exibem correlação positiva, o que atende às nossas expectativas: **impostos sobre vendas mais altos levam a preços mais altos.** No entanto, uma análise de correlação como essa não é suficiente para verificar se o instrumento é relevante. Mais tarde, voltaremos à questão de verificar se um instrumento é relevante e exógeno.

Consumo de Cigarros

- A primeira regressão é:

$$\log(P_i^{cigarettes}) = \pi_0 + \pi_1 SalesTax_i + \nu_i$$

- Estimamos esse modelo no qual o segundo estágio rodamos uma regressão LogQi em LogPi para obter os betas estimados em segundo estágio.
- O resultados da primeira regressão ficam da determinada maneira.

$$\log(\widehat{P_i^{cigarettes}}) = \underset{(0.03)}{4.62} + \underset{(0.005)}{0.031} SalesTax_i$$

- No qual o modelo sugere que os impostos de venda são variáveis com significância estatística.

Consumo de Cigarros

- A variação de Pi é explicada por quanto pelos impostos de venda?
- Podemos utilizar o R2 para verificar essa relação, no qual 47% da variação em preços após o imposto são explicados pela variação de impostos em diferentes estados.

Consumo de Cigarros

- Logo após rodamos a regressão em 2º estágio que nos dá os valores que procuramos.

$$\log(\widehat{Q_i^{cigarettes}}) = \underset{(1.70)}{9.72} - \underset{(0.36)}{1.08} \log(P_i^{cigarettes})$$

Resultado IV

- Desta forma, aparentemente o modelo realiza o ajuste necessário automaticamente. conforme utilizamos a regressão IV.

```
> # executar regressão em dois estágios usando 'ivreg()'
> cig_ivreg <- ivreg(log(packs) ~ log(rprice) | salestax, data = c1995)
> coeftest(cig_ivreg, vcov = vcovHC, type = "HC1")

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  9.71988    1.52832  6.3598 8.346e-08 ***
log(rprice) -1.08359    0.31892 -3.3977  0.001411 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testes de Diagnóstico

- Coeficientes:
- Intercepto: A estimativa do intercepto (9,7199) é altamente significativa (p < 0,001). Isso mostra o log estimado de maços de cigarro vendidos quando o log do preço real (rprice) é zero.
- Log do preço real (log(rprice)): A estimativa do coeficiente para log(rprice) é -1,0836, o que significa que um aumento de 1% no preço real dos cigarros está associado a uma redução de 1,08% na quantidade de maços de cigarro vendidos (já que é em termos de log). Essa relação é estatisticamente significativa (p = 0,00131).

```
Call:
ivreg(formula = log(packs) ~ log(rprice) | salestax, data = c1995)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64619 -0.07732  0.02981  0.11283  0.41904

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.7199     1.5141   6.420 6.79e-08 ***
log(rprice)  -1.0836     0.3166  -3.422  0.00131 **

Diagnostic tests:
                 df1 df2 statistic  p-value
Weak instruments   1  46    40.956 7.27e-08 ***
Wu-Hausman         1  45     0.314    0.578
Sargan             0  NA        NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1904 on 46 degrees of freedom
Multiple R-Squared: 0.4011,	Adjusted R-squared: 0.3881
Wald test: 11.71 on 1 and 46 DF,  p-value: 0.001313
```

Testes de Diagnóstico

- Teste de instrumento fraco:
- Estatística F = 40,956, valor p = 7,27e-08: O teste de instrumento fraco verifica se o instrumento (neste caso, salestax) está suficientemente correlacionado com a variável endógena (log(rprice)). Uma estatística F de primeiro estágio acima de 10 geralmente indica que o instrumento é forte.
- Interpretação: Como a estatística F está bem acima do limite convencional de 10 (com um valor p muito pequeno), o instrumento (salestax) é forte. Isso significa que salestax é um instrumento válido e relevante para log(rprice).

```
Call:
ivreg(formula = log(packs) ~ log(rprice) | salestax, data = c1995)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64619 -0.07732  0.02981  0.11283  0.41904

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.7199     1.5141   6.420 6.79e-08 ***
log(rprice)  -1.0836     0.3166  -3.422  0.00131 **

Diagnostic tests:
                 df1 df2 statistic  p-value
Weak instruments   1  46    40.956 7.27e-08 ***
Wu-Hausman         1  45     0.314    0.578
Sargan             0  NA        NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1904 on 46 degrees of freedom
Multiple R-Squared: 0.4011,     Adjusted R-squared: 0.3881
Wald test: 11.71 on 1 and 46 DF,  p-value: 0.001313
```

Testes de Diagnóstico

- Teste de Wu-Hausman:
- Estatística Wu-Hausman = 0,314, valor p = 0,578: Este teste avalia se o regressor endógeno (log(rprice)) está correlacionado com o termo de erro, o que violaria as suposições do modelo de mínimos quadrados ordinários (OLS).
- Interpretação: O alto valor de p (0,578) sugere que não há evidência de endogeneidade. Isso significa que, estatisticamente, o modelo OLS teria produzido estimativas semelhantes, e a endogeneidade pode não ser um grande problema neste caso específico. No entanto, já estamos usando regressão IV, ainda é uma boa prática prosseguir com o método IV se a endogeneidade for uma preocupação teórica.

```
Call:
ivreg(formula = log(packs) ~ log(rprice) | salestax, data = c1995)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64619 -0.07732  0.02981  0.11283  0.41904

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.7199     1.5141   6.420 6.79e-08 ***
log(rprice)  -1.0836     0.3166  -3.422  0.00131 **

Diagnostic tests:
                 df1 df2 statistic  p-value
Weak instruments   1  46    40.956 7.27e-08 ***
Wu-Hausman         1  45     0.314    0.578
Sargan             0  NA        NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1904 on 46 degrees of freedom
Multiple R-Squared: 0.4011,     Adjusted R-squared: 0.3881
Wald test: 11.71 on 1 and 46 DF,  p-value: 0.001313
```

Referências

- Hlavac, Marek. 2022. Stargazer: Well-Formatted Regression and Summary Statistics Tables. Bratislava, Slovakia: Social Policy Institute. https://CRAN.R-project.org/package=stargazer.
- Kleiber, Christian, and Achim Zeileis. 2008. Applied Econometrics with R. New York: Springer-Verlag. https://CRAN.R-project.org/package=AER.