Econometría (II / Práctica)

Magíster en Economía

Tema 5: Variables Instrumentales (IV)

Prof. Luis Chancí
www.luischanci.com

Apuntes Econometría
Prof. Luis Chancí

$E(Y_i|X_i = x) = \int z\, f(z|X_i = x)dz$

Outline

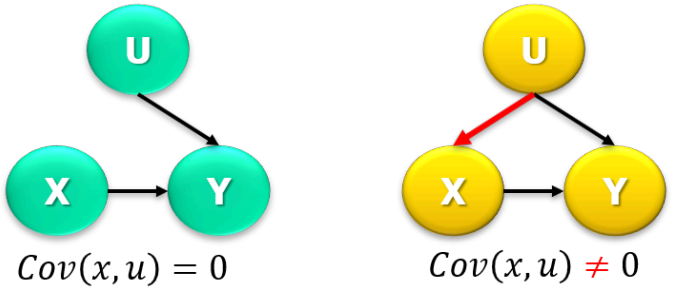
1. Introduction - Endogeneity
2. Instrumental Variables (IV)
    - Introduction
    - Finding instruments
    - The statistics when using IV
    - Consistency of 2SLS
    - Asymptotic Distribution of 2SLS
    - IV are LATE
    - Example of IV using 
    - Example of IV using Stata

Introduction - Endogeneity

Endogeneity means that an explanatory variable in a regression model is correlated with the residual or error term.

This gives biased and inconsistent parameter estimates (e.g., by OLS), weakening any causal inference conclusion drawn from an empirical model.

Using DAGs



Note: Directed Acyclic Graphs are effective in illustrating relationships between variables (particularly, the presence of endogeneity)

Introduction - Endogeneity (cont.)

Main sources:

1. Omitted Variable Bias
   When one fail to include one or more relevant variables that influence both Y and X. For instance, the true model is $Y = X_1\beta_1 + X_2\beta_2 + u_T$, where $\mathbb{E}(u_T|X_1, X_2) = 0$, and the estimated model is $Y = X_1\beta_1 + u_E$. Therefore,

$$\mathbb{E}_X(\hat{\beta}_1) = \beta_1 + (X_1'X_1)^{-1}(X_1'X_2)\beta_2 \quad \text{or for one regresor,} \quad \mathbb{E}(\hat{\beta}_1|X) \equiv \beta_1 + \beta_2 \frac{Cov(X_1, \text{Omitted variable})}{Var(X_1)}$$

2. Simultaneity or Reverse Causality
   When causality runs in both directions between Y and X). Example: Supply-Demand model, $Q^d = Q(P)$ and $P = P(Q^d)$.

3. Measurement Error
   When X is measured with error (that correlates with the true value of the variable), and this error becomes part of the residual term.

4. Sample Selection Bias
   When the sample is not randomly selected from the population and the process of selection is related to the variables of interest.
5. Dynamic Panel Bias
   When lagged dependent variables are used as regressors and can be correlated with past error terms.

Introduction - Endogeneity (cont.)

Common approaches to address endogeneity are:

- Control for Additional Variables
  Including additional relevant variables in the model to mitigate omitted variable bias.
- Instrumental Variable (IV)
  Using instruments that are correlated with the endogenous explanatory variables but uncorrelated with the error term.
- Panel Data Techniques
  Utilizing panel data methods that can account for unobserved heterogeneity and dynamic effects.
- Structural Modeling
  Building a more comprehensive model that explicitly incorporates the mechanisms causing simultaneity or selection bias.

In this section, we review Instrumental Variable.

Instrumental Variables (IV)

IV - Introduction

Instrumental Variables (IV) is a technique aimed at estimating causal relationships when controlled experiments are not feasible but there is an alternative source of information (an alternative source of exogenous variation).

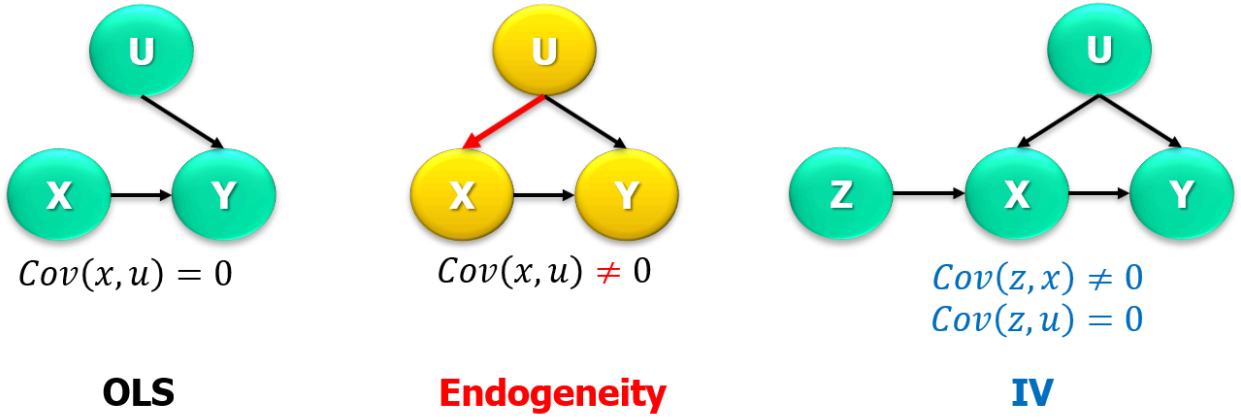
For instance, assessing the impact of a having master's degree on wage using a randomized control trial (RCT) is not feasible. There is no a random assignment or random placement in treatment (master's degree). Therefore, an estimate of β_1 by OLS raises a problem of endogeneity

$$\mathbb{E}[log(wage_i)|School_i,\ Exper_i] = \beta_0 + \beta_1 Schooling_i + \beta_2 Experience_i$$

In the paper Instrumental variables and the search for identification: From supply and demand to natural experiments (*Journal of Economic perspectives* , 2001) Angrist and Krueger state that the IV method first laid out in Wright (1982) who approached a simultaneous equation identification challenge by using "variables that appear in one equation to shift this equation and trace out the other". Thus, according to this view, instruments are variables that do the shifting.

IV - Introduction

More formally, IV is a method to estimate causal effects by using variables that influence the treatment but are not directly related to the outcome.



- The interpretation of IV estimates requires careful consideration of the instrument's validity and the assumptions underlying the IV approach.
- IV estimates are often interpreted as local average treatment effects (LATE), providing causal inference in a specific subset of the population.

IV - conditions and examples

A suitable instrument (z) for IV must satisfy:

1. Relevance ($z \rightarrow x$). The instrument must be correlated with the endogenous explanatory variable, $Cor(z, x) \neq 0$.
   Exclusion ($z \rightarrow x \rightarrow y$, $z \rightarrow /y$): $Cor(z, y|x) = 0$.
2. Exogeneity ($u \rightarrow /z$). The instrument should not be correlated with the error term in the regression model, $Cor(z, u) = 0$.

Hall of fame - examples of studies using IV:

- David Card's paper (1995) on Education and Earnings
  geographical variation in colleges' proximity as an instrument to analyze the return to education. The proximity to a college increases the likelihood of higher education, leading to higher earnings.

- Angrist's paper (1990) on the effects of compulsory military service on earnings
  used the Vietnam War draft lottery numbers as an instrument to evaluate the impact of military service on lifetime earnings.

Finding instruments

A note about finding instruments.

- In some studies, finding a source of exogeneous variation could be straightforward. There is a policy or a similar source of variation.
- In other cases, the researcher needs to construct more complex (perhaps 'creative') sources of variation. Some examples are:
  - Shift-share or Bartik type instruments associated with T. Bartik (1991). See, Goldsmith-Pinkham, Sorkin, and Swift (2020)
  - Simulated instruments (Borusyak and Hull, 2022)
  - Granular instruments (Gabaix and Koijen, 2023).

  Please, also check the NBER working paper 33594, by Kirill Borusyak and Peter Hull, *Optima Formula Instruments* , 2025.

Finding instruments: Bartik or Shift-Share Instruments

The 'Shift-Share' is an useful approach to identify causal relationships where local economic conditions are influenced by broader national or global trends.

It is a combination of local industry shares (the shares) with national trends (the shift) to create instruments. The idea is to exploit external/exogenous variation (i.e., national industry trends) that affects the local economy.

For instance, considering industries (indexed by k), in geographic locations ℓ at time t, a general representation would be like:

$$Z_{\ell,t} = \sum_k (\text{Share}_{\ell,ind,t} \times \text{National Trend}_{k,t})$$

In Autor, Dorn, and Hanson (2013):

- They interact local industry shares (in location ℓ) with aggregate trade flows (the growth of imports from China to other high-income countries) to examine the impact of Chinese imports on labor markets in the US.
- Thus, the exposure of a location to Chinese imports is a weighted average of how much China is exporting in general of different products ("shift"). The weights are the initial industry composition in a location ("shares").

Instrumental Variables Estimator

Simplest case: Just-identified IV. One regressor ($k = 1$), x, and one instrument ($r = 1$), z.

Employing $Cov(z, u) = 0$,

$$\begin{aligned} y &= \beta_1 x + u \\ Cov(z, y) &= Cov(z, \beta_0) + \beta_1 Cov(z, x) + Cov(z, u) \\ &= 0 + \beta_1 Cov(z, x) + 0 \\ &= \beta_1 Cov(z, x) \end{aligned}$$

therefore,

$$\hat{\beta}_{IV} = \frac{\sum (z_i - \bar{z})(y_i - \bar{y})}{\sum (z_i - \bar{z})(x_i - \bar{x})}$$

alternatively, we can obtain a similar result using the (Method of) Moments $\mathbb{E}(z_i u_i) = 0$ (a technique we will review in the following section). For ease of exposition, let's assume that $\bar{z} = 0$, thus

$$N^{-1} \sum_i z_i \hat{u}_i = 0 \quad \rightarrow \quad \sum z_i y_i - \left(\sum z_i x_i \right) \hat{\beta}_{MM} = 0 \quad \rightarrow \quad \hat{\beta}_{MM} = \hat{\beta}_{IV} = (z'x)^{-1}(z'y)$$

IV - estimation (cont.)

In the general case (more than one regressors and or instruments, $r \geq 1$), the technique is 2SLS (Two Stage Least Squares).

This approach involves the following two stages:

1. First Stage. Regression of the endogenous variable(s) on the instrumental variable(s) and other exogenous covariates.
2. Second Stage. Use the prediction values from the first stage as explanatory variables in the main regression equation.

To illustrate, let the structural equation be

$$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u$$

where X_2 is a vector or matrix representing the endogenous variable(s) and X_1 are exogenous covariates. The instrument are in Z_2. Thus,

1. The first stage would be $X_2 = X_1\Gamma_1 + Z_2\Gamma_2 + \epsilon = Z\Gamma + \epsilon$, so one can compute $\hat{X}_2 = Z\hat{\Gamma}$.
2. The second stage would be $y = X_1\beta_1 + \hat{X}_2\beta_2 + u$, hence, can get $\hat{\beta}$ by (naive) OLS.

IV - estimation (cont.)

Let's use the reduced form (the variable as function of the instruments) to get a more general expression for $\hat{\beta}_{2SLS}$.

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + u \\ &= X_1\beta_1 + (X_1\Gamma_1 + Z_2\Gamma_2 + \epsilon)\beta_2 + u \\ &= Z\Gamma\beta + \xi \\ &= \Theta\beta + \xi \end{aligned}$$

where, $\Theta = Z\Gamma$, ξ, and $Z\Gamma = \begin{pmatrix} I & \Gamma_1 \\ 0 & \Gamma_2 \end{pmatrix}$.

Suppose Θ were known (we do know Z, but not Γ). Then one can get β by OLS:

$$\hat{\beta} = (\Theta'\Theta)^{-1}(\Theta'Y)$$

However, by replacing Γ by the estimator $\hat{\Gamma} = (Z'Z)^{-1}(Z'X)$, we can get $\hat{\beta}_{2SLS}$ as follows

$$\hat{\beta}_{2SLS} = (\hat{\Theta}'\hat{\Theta})^{-1}(\hat{\Theta}'Y) = (X'P_ZX)^{-1}X'P_ZY$$

where P_Z is the projection matrix $P_Z = Z(Z'Z)^{-1}Z'$.

The statistics when using IV

1. Caution!

   Be aware of the standard error calculation issue: Errors from a 'manual' (naive OLS) approach in 2SLS are incorrect!

   The OLS standard error formula assumes that X_1 is a fixed regressor, although it is actually a prediction from the first stage and has its own sampling variability.

   It's essential to account for this additional variability, which can be achieved through methods like

   - **Bootstrapping**
   - **Analytical corrections.** For instance, $V(\hat{\beta}_{IV}) = \sigma_z^2 \sigma_u^2 / cov(z,x) = (\sigma^2/\sigma_x^2) * (1/\rho_{zx}^2)$.

2. We should jointly test the instruments. For instance, using

   - **F-test**
   - **Kleibergen-Paap rk Wald F statistic.**

Consistency of 2SLS

Assumption

- (Y_i, X_i, Z_i), $i = 1..., n$, are i.i.d.
- $\mathbb{E}(Y^2) < \infty$
- $\mathbb{E}||X||^2 < \infty$
- $\mathbb{E}||Z||^2 < \infty$
- $\mathbb{E}(ZZ')$ is positive definite
- $\mathbb{E}(ZX')$ has full rank
- $\mathbb{E}(Zu) = 0$

Theorem.

Under these assumptions,

$$\hat{\beta}_{2SLS} \to_p \beta \quad \text{as} \quad n \to \infty$$

Asymptotic Distribution of 2SLS

Assumption (sufficient regularity conditions). In addition to the above mentioned assumptions, assume

- $\mathbb{E}(Y^4) < \infty$
- $\mathbb{E}||X||^4 < \infty$
- $\mathbb{E}||Z||^4 < \infty$
- $\Omega = \mathbb{E}(ZZ'u^2)$ is positive definite.

Theorem.

Under the assumptions, as $n \to \infty$,

$$\sqrt{n}\left(\hat{\beta}_{2SLS} - \beta\right) \underset{d}{\to} N(0, V_\beta)$$

where $V_\beta = (Q_{XZ}Q_{ZZ}^{-1}Q_{ZX})^{-1}(Q_{XZ}Q_{ZZ}^{-1}\Omega Q_{ZZ}^{-1}Q_{ZX})(Q_{XZ}Q_{ZZ}^{-1}Q_{ZX})^{-1}$ and, for instance, $n^{-1}X'Z \overset{p}{\to} Q_{XZ} = \mathbb{E}(XZ)$.

IV are LATE

As mentioned, IV estimates are often interpreted as local average treatment effects (LATE), providing causal inference in a specific subset of the population. Let's integrate the causal concepts — Intention-to-Treat (ITT), Reduced Form, and Local Average Treatment Effect (LATE) — into the framework of Instrumental Variables (IV), specifically the Two-Stage Least Squares (2SLS) estimation.

Simply re-define the system of equations involved in IV as follows:

- Structural Equation (captures the causal effect of treatment D on outcome Y):

$$Y_i = \alpha + \delta D_i + \varepsilon_i$$

- First Stage Equation (captures the effect of the instrument Z on treatment D):

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

- Reduced Form Equation (captures the total effect of the instrument Z on outcome Y):

$$Y_i = \gamma_0 + \gamma_1 Z_i + \eta_i$$

where, γ_1 is the Reduced Form coefficient; π_1 is the First Stage coefficient; and δ is the causal effect of interest.

IV are LATE

Intention-To-Treat (ITT) Effect

- As mentioned at the begining of the semester, the ITT effect is the effect of being assigned to treatment — whether or not the treatment is actually taken. In the IV context, this is captured by the Reduced Form:

$$\text{ITT} = \gamma_1 = \frac{\partial Y}{\partial Z}$$

- This effect is useful in RCTs with noncompliance as it estimates how outcomes change when the instrument (treatment assignment) is varied.

First Stage: Compliance Effect

- The first stage regression estimates how much the treatment responds to the instrument:

$$\pi_1 = \frac{\partial D}{\partial Z}$$

- It captures the compliance behavior (how being assigned to treatment influences actual uptake).

IV are LATE

Local Average Treatment Effect (LATE)

- The IV estimator in the just-identified case is:

$$\hat{\delta}_{IV} = \frac{\gamma_1}{\pi_1} = \frac{\text{Reduced Form}}{\text{First Stage}} = \frac{\text{ITT}}{\text{Compliance}}$$

This ratio estimates the Local Average Treatment Effect (LATE). That is, the causal effect for compliers, i.e., individuals who take the treatment only when encouraged by the instrument.

In summary,

Name	Equation	Interpretation
Structural Model	$Y = \alpha + \delta D + \varepsilon$	Causal effect of treatment on outcome
First Stage	$D = \pi_0 + \pi_1 Z + \nu$,	Instrument affects treatment
Reduced Form	$Y = \gamma_0 + \gamma_1 Z + \eta$,	Instrument affects outcome (ITT)
IV Estimator (LATE),	$\delta = \gamma_1 / \pi_1$	Causal effect for compliers

Example of IV using R

Let's use a dataset from Wooldridge (Wage2), which contains information on wages, schooling, among other related variables.

We are interested in the returns ($y = wage$) to education ($x = schooling$).

OLS with will likely be biased,

$$\text{Wage}_i = \beta_0 + \beta_1 \text{schooling}_i + u_i$$

In particular, **the OLS results**

Variable	Coeff.	S.error	t.stat.	p-value
(Intercept)	176.504	89.152	1.98	0.0481
schooling	**58.594**	**6.439**	**9.10**	**<0.001**

```r
# R code:

# Select the variables to use
wage_df <- wage2 %>%
            transmute(wage,
                      schooling = educ,
                      education_dad = feduc,
                      education_mom = meduc) %>%
              na.omit() %>% as_tibble()

head(wage_df)
## # A tibble: 6 × 4
##    wage schooling education_dad education_mom
##   <int>     <int>         <int>         <int>
## 1   769        12             8             8
## 2   808        18            14            14
## 3   825        14            14            14
## 4   650        12            12            12
## 5   562        11            11             6
## 6   600        10             8             8
```

Example of IV using R (cont.)

Using mother's education as an instrument.

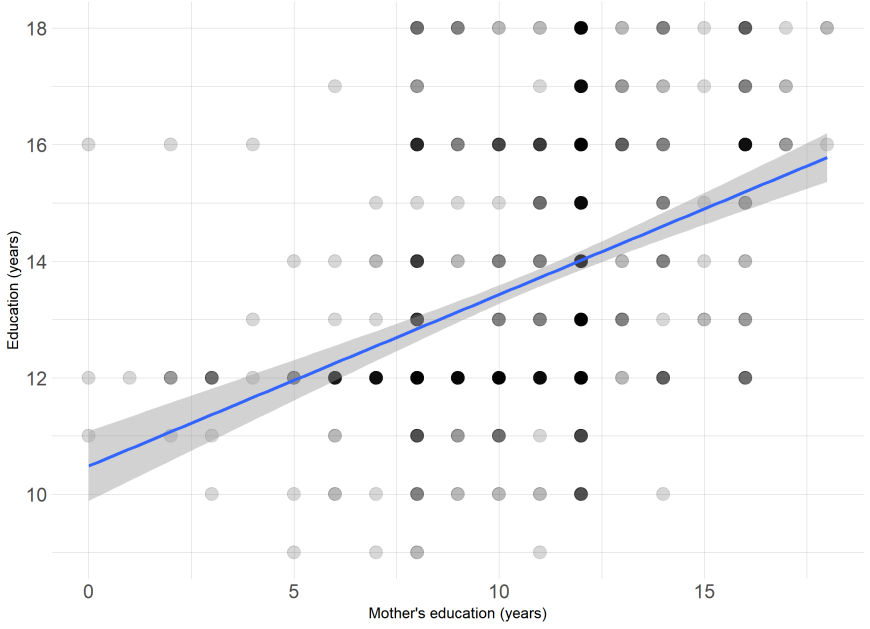
Checking the instrument: considering that z must (i) only affect y (wages) through x (schooling), and (ii) be uncorrelated with other factors that affect y, does the mother's education provide a valid instrument?

First-stage: The effect of z on x,

$$\text{Education}_i = \gamma_0 + \gamma_1 (\text{Mother's Education})_i + v_i$$

Variable	Coeff.	S.error	t.stat.	p-value
(Intercept)	10.487	0.306	34.32	<0.001
education_mom	**0.294**	**0.027**	**10.75**	**<0.001**

The p-value suggest the variable is an important predictor.



Example of IV using R (cont.)

Second stage: using $\hat{x}$ from the first stage, the estimate of β in the second stage (using OLS) is:

```
# store the predicted values
schooling_hat <- reg_fs$fitted.values

reg_ss     <- lm(wage ~ schooling_hat, wage_df)
```

Variable	Beta	s.e. ;	t.stat	p-value
Intercept	-501.474	244.201	-2.05	0.0404
Schooling	**108.214**	**17.840**	**6.07**	**<0.0001**

Note:
$\hat{\beta}_1$ can also be computed from $\hat{\beta}_1 = \hat{\beta}_{IV} = \hat{\phi}_1 / \hat{\gamma}_1$, where $\hat{\gamma}_1$ is computed in the first stage and ϕ_1 from $y_i = \phi_0 + \phi_1 z + \xi$ (reduced form).

Example of IV using R (cont.)

In empirical works, R users rely on functions like `ivreg()` or `iv_robust()`, as they compute heteroskedasticity-robust standard errors. The former is from the `AES` package and the later from `estimatr`.

The functions work quite similar to the `lm` command: `(y ~ x1 + x2 + ... | z1 + z2 + ... , data)`. Also, they compute heteroskedasticity-robust standard errors.

```
# IV regression ( instrumento: educ de la madre)
#iv_est <- iv_robust(wage ~ schooling | education_mom, data = wage_df)
iv_est <- ivreg(wage ~ schooling | education_mom, data = wage_df)
```

Variable	Coeff.	S.error	t.stat.	p-value
(Intercept)	-501.474	246.684	-2.03	0.0424
schooling	**108.214**	**18.021**	**6.00**	**<0.001**

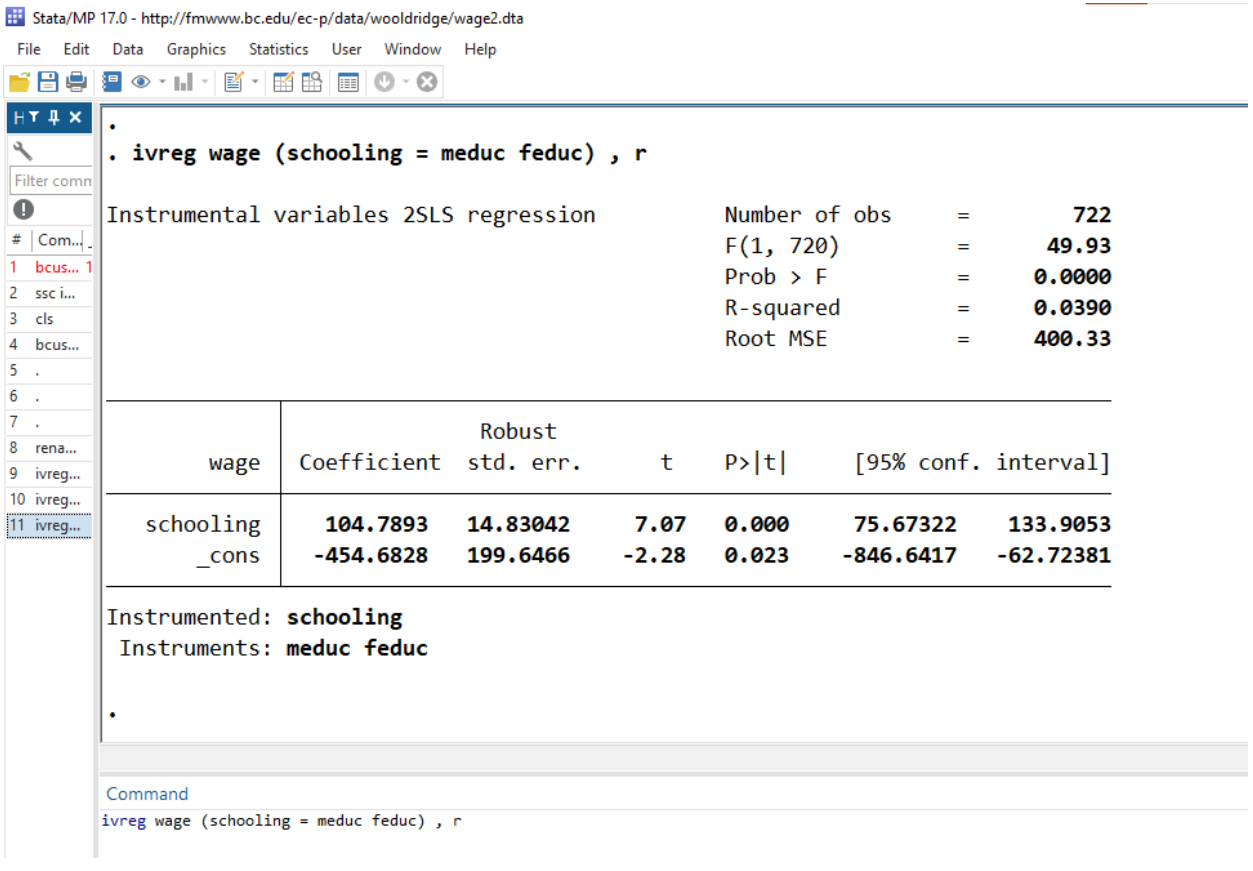
Or using the mother's and father's education as instruments:

```
# IV regression (usando ivreg y dos instrumentos)
ivreg(wage ~ schooling | education_mom + education_dad, data = wage_df)
```

Variable	Coeff.	S.error	t.stat.	p-value
(Intercept)	-454.683	201.201	-2.26	0.0241
schooling	**104.789**	**14.685**	**7.14**	**<0.001**

Example of IV using Stata

Or in `Stata` it would be `ivreg y (x = z1 z2 ...) , r`



Cierre

¿Preguntas?

O vía E-mail: lchanci1@binghamton.edu