Spring 2023 Gregory A. Bruich, Ph.D.
Economics 1123 Harvard University

Section 6: Instrumental Variables Regression

1 Introduction

Suppose we are interested in estimating the causal effect of X on Y, $\beta_1 = E\left[Y_i(1) - Y_i(0)\right]$, using the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, but we think that OLS will be biased due to omitted variables, measurement error, or simultaneous causality. Instrumental variables regression is a way of solving this problem.

1. **How did we solve the OVB problem before the midterm?**

2. Instruments: In order for an IV approach to work, the instrument(s) Z has to satisfy the following conditions:

   - Z has to be relevant. **What does that mean in math? What does that mean in pictures?**

   - Z has to be exogenous. **What does that mean in math? What does that mean in pictures?**

2 Two stage least squares: Equivalent methods

One way of implementing this idea is through a 2SLS regression.

1. The first way of implementing 2SLS involves a first stage regression and a reduced form regression. The ratio of the coefficients from these two regressions forms the causal estimate.

   - First stage: Regress treatment X on instrument Z.

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

   - Reduced form: Regress outcome variable Y on instrument Z.

   $$Y_i = \alpha_0 + \alpha_1 Z_i + v_i$$

   - **What is the interpretation of α_1? What assumption underlies this interpretation?**

   - Putting it together: From these two regressions, the estimated causal impact of X on Y is

   $$\beta^{2SLS} = \frac{\alpha_1}{\pi_1}.$$

   More generally, the ratio of the reduced form coefficient over the first stage coefficient will give us the causal estimate of interest.

   - Intuitively, we rescale so that the effect of instrument on the outcome is in the units of the treatment.

2. A second way of implementing 2SLS involves two components: a first stage regression and a second-stage regression based off the first stage's predicted values.

   - First stage: decompose X to get the "problem-free" part of X that is uncorrelated with u by regressing X on Z.

     $$X_i = \pi_0 + \pi_1 Z_i + v_i.$$

     The problem-free part of X is the predicted value of the first stage regression:

     $$\hat{X}_i = \pi_0 + \pi_1 Z_i$$

   - **Why does this give us the problem-free part of X?**

   - Second stage: use the problem-free part of X to estimate $\beta_1$by regressing Y_i on $\hat{X}_i$

     $$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i.$$

   - **Show mathematically that this is the same as $\beta^{2SLS} = \frac{\alpha_1}{\pi_1}$.**

   In Stata, one can run code of the following form to execute the 2SLS method described above, *but we never do it this way.*

```
reg x z, r
predict pred_x
reg y pred_x, r
```

   Why don't we do it this way?

3. It turns out that there is a third equivalent way. Instead of using $\hat{X}_i$, instead we can generate residuals $\widehat{v}_i = X_i - \hat{X}_i = X_i - \{\pi_0 + \pi_1 Z_i\}$, and control for these directly:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \widehat{v}_i + u_i.$$

Intuitively, this works because we are controlling for the "problematic" part of X_i.

In Stata, one can run code of the following form to execute the 2SLS method described above, *but we never do it this way.*

```
reg x z, r
predict uhat, resid
reg y x uhat, r
```

3 Example: Returns to schooling

1. Suppose we run the following OLS regression to study the return to schooling.

$$\ln(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \beta_4 black_i + \beta_5 south_i + u_i$$

- **What is the coefficient we are interested in? What is our "X" variable? What are our "W" variables?**

- **Why might the OLS estimate of β_1 be biased?**

2. One potential instrumental variable exploits geographical differences in accessibility of college – the distance to one's closest four-year college.

- **Is this instrument plausibly relevant?**

- **Is this instrument plausibly exogenous?**

4 Implementing two stage least squares

To simplify, we begin with no controls and only one instrument.

Note The commands `ivreg` and `ivregress 2sls` tell Stata to do different things for computing standard errors and test statistics. Use `ivregress 2sls`.

```
use card.dta, clear

*Reg 1: First stage
reg educ nearc2, r
outreg2 using table1.xls, cttop(OLS) replace


Linear regression                               Number of obs     =      3,010
                                                F(1, 3008)        =       6.72
                                                Prob > F          =     0.0096
                                                R-squared         =     0.0022
                                                Root MSE          =     2.6744

------------------------------------------------------------------------------
             |               Robust
        educ | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      nearc2 |   .2552584   .0984531     2.59   0.010     .0622162    .4483006
       _cons |   13.15092     .06451   203.86   0.000     13.02443    13.27741
------------------------------------------------------------------------------


*Reg 2: Reduced form
reg lwage nearc2, r
outreg2 using table1.xls, cttop(OLS)

Linear regression                               Number of obs     =      3,010
                                                F(1, 3008)        =      28.55
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0096
                                                Root MSE          =     .44173

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      nearc2 |   .0876235   .0163978     5.34   0.000     .0554715    .1197756
```

```
       _cons |   6.223202   .0103073   603.77   0.000     6.202992    6.243412
------------------------------------------------------------------------------
```

```
*Reg 3: 2SLS
ivregress 2sls lwage (educ=nearc2), r
outreg2 using table1.xls, cttop(2SLS)

Instrumental variables 2SLS regression          Number of obs   =      3,010
                                                Wald chi2(1)    =       7.23
                                                Prob > chi2     =     0.0072
                                                R-squared       =          .
                                                Root MSE        =      .8859

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .3432739   .1276252     2.69   0.007     .0931332    .5934146
       _cons |   1.708834     1.6926     1.01   0.313    -1.608601     5.02627
------------------------------------------------------------------------------
Instrumented: educ
 Instruments: nearc2
```

Exercise. How can we visualize two stage least squares?

Exercise. Compute the ratio of the coefficient on `nearc2` in the second (reduced form) regression to the coefficient on `nearc2` in the first (first-stage) regression. Is this ratio greater than, less than, or equal to the coefficient on `educ` in the third (2SLS) regression? Explain, either algebraically or using an intuitive explanation.

How do we tell if the Z_is are valid instruments? What are the two things we want to test?

Relevance: F-test of the coefficients on the Z_is in the first stage regression Our first stage model was:

$$X_i = \pi_0 + \pi_1 Z_i + ... + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + ... + \pi_{m+k} W_{ki} + \epsilon_i,$$

So to test for relevance, we test:

$$H_0 : \pi_1 = \pi_2 = ... = \pi_m = 0.$$

What is the interpretation of this test? Why does it make sense for a test of relevance?

Exogeneity: Can we test for exogeneity?