Advanced Methods in Impact Assessment Workshop

Day 3: Measuring Program Impacts: Diff-in-Diff and Instrumental Variables

Today we will apply the information you learned this morning regarding Difference-in-Difference (DiD) and Instrumental Variables (IVs) to calculate the Average Treatment Effect (ATE) and the Local Average Treatment Effect (LATE).

There are three objectives for today's exercises:

1. Calculate the difference and then the Difference-in-Difference.
2. Examine data for potential IVs and check if those IVs are effective.
3. Implement IV estimation, using local average treatment effects.

To get started, load into **Stata** the data set `VDSA_Prod_Data_Ref.dta` from Monday. Ensure that the data has the log transformed variables. Again open a `.log` file and write your code in a `.do` file so you can reference it later.

Difference-in-Difference (DiD)

To gain a sense of how the DiD is calculated we will start by looking at simple differences. This will be similar to what we did at the end of Day 1 when we looked at the Within/Without and the Before/After comparisons.

Start by dropping survey year 2012 so that the data only contains observations from 2010 and 2011. Make sure to save this data set as a new file. We will continue to use it throughout the day.

1. Do a simple t-test (base command: `ttest lny, by (irr)`) to examine whether parcels in the irrigation program had higher yield than parcels not in the program, using observations from 2011 only. What does this measure?

In order to measure the difference in outcomes across time between the treatment and control groups we need to manipulate our data set a bit more. Below is code for how to set up the measure.

```
gen lny0 = lny if sur_yr == 2010
egen lny10 = max(lny0), by(prcl_id)
keep if sur_yr == 2011
gen lny11 = lny
gen lny1011 = lny11 - lny10
```

2. Do a t-test to compare the differences in the 2010 and 2011 log yield between parcels in the irrigation program and those not in the program. In other words, do a differences-in-differences estimate. Compare your results with those in Question 1. If they are different, explain why. Which estimate do you think is closer to the truth, and why?

Next, we're going to consider the difference-in-difference estimation in a regression context. Construct a dummy variable indicating the interaction of being in the treatment group and the year being 2011.

```
gen irr0 = irr if irr == 1 & sur_yr == 2011
egen irr11 = max(irr0), by(prcl_id)
replace irr11 = 0 if irr11 == .
xi i.sur_yr, pref(vdum)
gen irr11_yr = irr11*vdumsur_yr_2011
```

To start with, just consider an OLS regression of `lny` on `irr11 irr11_yr vdumsur_yr_2011`.

3. What do the coefficients in your regression mean?

4. Add our standard set of control variables and re-estimate the equation. Continue to just use OLS. What happens to your estimates? Why?

As a comparison to our DiD regressions, let's try to apply propensity score matching to this data set. To conduct the PSM we need to generate a new data set with only data from 2010, but with an indicator variable equal to one if the parcel was treated in 2011. Keep only the data from 2010:

```
keep if sur_yr==2010
```

Now run the `pscore` command on the irrigation data matching on our control variables, except `lnlindex` and `lndist`. Be sure to select the common support option with `comsup` and create a block id called `blockf1`.

Now we need to merge the matched households in the baseline year back into the panel data:

```
keep if blockf1!=.
keep prcl_id
sort prcl_id
merge 1:m prcl_id using <<data saved before question 1>>
keep if _merge == 3
```

You will need to recreate the DiD variables as we did prior to Question 3. With this data set carry out the DiD method as before using our standard set of control variables.

5. How do the number of observations change across the DiD and DiD using PSM approaches? Does this have an effect on the external validity of our impact assessment?
6. Comparing results between DiD and DiD using PSM. Does one have stronger internal validity? Why?

Instrumental Variables

Next, we'll move onto IVs! First, reload the data set that we save before Question 1 (this should have data only from 2010 and 2011). Next we'll deal with selecting an instrumental variable. You'll want to first use the `ssc install` command to find and install `ivendog`, `ranktest` and `ivreg2`.

7. Run an OLS regression to look at the impact of the irrigation treatment on log of yield while controlling for our standard set of exogenous control variables. What is your result for program participation? Why might it be biased?
8. Examine the data and discuss which variables would potentially make a good IV.

Next, we'll run some IV regressions. Construct an instrumental variable by interacting household land ownership with yearly rainfall in the village (`gen IV_landrain = tot_acre*rain`).

9. Why might this be a valid instrumental variable? What are the costs and benefits of using just rainfall or just land ownership? What are the requirements for a valid IV?

Run the 2SLS piecewise. Run the first stage regression: regress the endogenous treatment variable on the instrument and our standard set of exogenous control variables. Use an OLS for this first stage regression. Save the predicted value using the command `predict double irrhat`. Run the second stage regression: regress our outcome variable on the exogenous control variables and the predicted values from the first stage regression. Again, this regression should be specified as an OLS.

Run the IV regression using `ivreg`. Make sure you include the `first` option at the end of the regression command line. This will tell **Stata** to display the first stage regression:

```
ivreg lny lnl lnf lnm lnp ageH gender sizehh lnaindex lnlindex
lntotacrelndist (irr = IV_landrain), first
```

10. Compare the point estimates and standard errors between the “two-stage” approach and the “single-stage” approach. Do the coefficient estimates differ? How about the standard errors?

Now, we’ll test for a weak instrument. Test for endogeneity by typing the command `ivendog`. This calls up the Wu-Hausman test.

11. What does this test tell you? Keep in mind that the test is only valid if the IV is valid.

Conduct a simple “falsification test” in which you regress the log yield on the exogenous variables and the instrument.

```
reg lny lnl lnf lnm lnp ageH genderH sizehh lnaindex lnlindex
lntotacrelndist IV_landrain
```

12. Does the instrument have a significant impact on yields? Does this mean our instrument is valid? Why might it still not be a valid instrument?

Now, let’s test the strength of the instrument. To do this, we need two things. First, we need more than a single instrument since the test is only valid when there is more than one instrument – the equation is “over identified.” So, instead of the IV we have been using, let’s consider rainfall, land ownership, and those two variables interacted as our set of instruments. Second, we need to use the command `ivreg2`.

```
ivreg2 lny lnl lnf lnm lnp ageH genderH sizehh lnaindex lnlindex
lntotacrelndist (irr = IV_landrain rain tot_acre), first
```

If you type `help ivreg2` there is a section about “First-stage regressions, identification, and weak-id-robust inference.” This discusses the test stats that are presented as a result of the `first` option.

13. Are you instruments endogenous? Look at the results of the Sargan-Hansen test. What does this tell you? Test again using `ivendog`.

14. Interpret your IV results. Do you think that your IV estimates are better than the OLS estimates? Explain.