Advanced Methods in Impact Assessment Workshop

Day 4: Panel Data Techniques and Challenges to Program Evaluation
Today we will apply the information you learned this morning regarding panel data estimators and regression discontinuity design.

There are two objectives for today's exercises:
1. Implement panel data estimation using several types of panel effect and compare the outcomes.
2. Estimate impacts based on sharp and fuzzy regression discontinuity.

Panel Data Techniques
Again open a `.log` file and write your panel data code in a `.do` file so you can reference it later. For the panel data we want to use all three years of our available data. So, load into **Stata** the data set that we created on the first day called `VDSA_Prod_Data_Ref.dta`. Ensure that the data set has the log transformed variables. If it does not, return to your `.do` file from Day 1 that contains the code for creating the transformed variables and run that code on the current data set. You should have a data set with three years of data and inverse hyperbolic sine transformed variables of per hectare inputs and outputs and controls.

Today we will run a number of other panel data estimators and we will take advantage of all three years of data. The value of using three years is that we can see the effect of the irrigation treatment for 1) those who used the treatment in 2011 and continued to use it in 2012, 2) those who used the treatment in 2011 but dis-adopted and did not use the treatment in 2012, and 3) those who did not use the treatment in 2011 but adopted and did use the treatment in 2012.

Before running any panel data regressions, the first thing you need to do is tell **Stata** which variable is your panel id variable. Since our irrigation intervention occurs at the parcel level, we want to use the parcel id variable to define our panel. So, type `xtset prcl_id`.

1. Regress log yield on the irrigation treatment and our standard set of control variables as a pooled OLS. What is the coefficient on our variable of interest? What does this mean?
2. Run the pooled OLS again but include time dummies to control for year-to-year fixed effects (simply include `i.sur_yr`). How do the results change? What is the source of variation that you are using to identify the effect of irrigation and how has it changed from the OLS regression?
3. Run the regression from Question 2 but this time use random effects. You will need to include `re` after the last independent variable. Why might you use random effects? You can test for the validity of random effects using the Breusch-Pagan Lagrangian multiplier test or the Hausman test. Try the Hausman test. To do this, you will need to rerun both models (with and without random effects) without the clustered standard errors and to save the estimates (`estimates store <<regname>>`). Are there systematic differences in the coefficients in the OLS and RE specifications?
4. You might want to control for possible time-invariant household characteristics along with the year fixed effects. Run the regression from Question 2, but `i.vdsa_hh_id` to your list of independent variables. How do your results change? How has your source of variation changed from the regression you ran in Question 2?
5. Run the regression again, this time using parcel fixed effects. You will need to include `", fe"` after the last independent variable. Why might you use parcel-level fixed effects? How do the results change from OLS? From the RE model?

Note that, except for the inclusion of observations from the year 2012, this regression is the same as the FE regression we ran as a check on the Diff-in-Diff estimator yesterday.

6. Why are the point estimates from the FE regression using only observations from 2010 and 2011 different than the most recent FE regression? Sort the data by year and then, by year, tabulate the variable `irr`. Does knowing the number of program participants in each year help explain the difference in point estimates? How?

7. Even after controlling for parcel-level FE, you might worry that the time-trends of yield in each community are different. To control for this, you might want to include village by year fixed effects. To do this, include generate an interaction term for `i.sur_yr*i.vil_id`. Include this variable and re-run the regression from Question 5. How do the results change? Are you worried that all of the village fixed effects get dropped in your regression? What effects might you capture using year by village fixed effects? Run this regression. What is your source of variation used to identify the effect of irrigation?

8. Assume you believe that irrigation is most effective during dry years. How would you test for this? There are no right answers – just think about what you might do. (Hint: an interaction term might be appropriate!) What results do you see? How do you interpret the coefficients? If you included an interaction term, how do you interpret the coefficient on this term?

Now we will prepare to run a correlated random effects (CRE) regression. Remember, CREs includes the average values of our control variables. So, first we will need to calculate the mean of these control variables for each parcel. An efficient way to do this is to define a local macro and loop to take the mean of the control variables for each parcel.

```
local z1 lnl lnf lnm lnp ageH genderH sizehh lnaindex lnlindex
lntotacre lndist
local i=1

local i=1
foreach var of varlist `z1' {
      qui egen `var'bar=mean(`var'), by(prcl_id)
      local z1bar `z11bar' `var'bar
local i=`i'+1
}
```

9. Run the regression again, this time using CRE. Note that by including the time averages we have controlled for unobserved heterogeneity and no longer have to use `xtreg` and can just use `reg`. Why might you use CREs? How do the results change from the previous models? When you include your interaction term from Question 7, how do your results change?

Regression Discontinuity Design

To conduct the RD analysis we first want to go back to our two year data set. So, drop data from 2012. (`drop if sur_yr == 2012`) RD is appropriate when there is a policy or program that has a strict cut off for eligibility. In our data, we don't have this strict cut off. To illustrate the method, let's pretend that there was an irrigation program that was only available to households that experienced less than 800 mm of rainfall. Rainfall affects yield directly, but the relationship between rainfall and yield should be smooth around the 800mm threshold for RD to be valid. The program creates a threshold at 800 mm.

To demonstrate sharp discontinuity, we are going to drop non-adopting households in low rainfall areas as well as adopting households in high rainfall areas. Therefore, we are going to make the assumption that everybody who was eligible for this program adopted irrigation. Also, no households who were ineligible for the program adopted irrigation. First, save the current file as `RD.dta` so that we can return to this data. Then, run the code below:

```
drop if (irr == 0 & rain < 800) | (irr == 1 & rain > 800)
```

Graph the data to make sure we have a sharp discontinuity around 800mm of rainfall. You can graph in Stata using the `twoway` command:

```
twoway (scatter rain irr)
```

Let's check to see whether the relationship between rainfall and our yield variable is smooth around the threshold. Again, use the `twoway` command:

```
twoway (scatter rain lny)
```

We also now want to take the Inverse Hyperbolic Sine of rainfall:

```
gen lnr = asinh(rain)
```

Now, we're going to work on a program for a sharp discontinuity. Depending on your version of **Stata** you may need to download the `locpoly` (kernel-weighted local polynomial regression) command. If so, type `findit locpoly` Look for the documentation of `locpoly` and then click on "click here to install."

To run the sharp discontinuity regression, use the following code:

```
*****Program for Sharp Discontinuity
capture prog drop rd_sharp
prog rd_sharp, rclass
  args outcome
  confirm var `outcome'
  tempname outrd1 outrd0 outcome1 outcome0
  lpoly `outcome' lnr if rain<800, gen(`outrd1') at(lnr) nogr tri w(3)
d(1)
  lpoly `outcome' lnr if rain>=800, gen(`outrd0') at(lnr) nogr tri w(3)
d(1)
  sum `outrd1' if rain>=450 & rain<800, meanonly
  scalar `outcome1'=r(mean)
  sum `outrd0' if rain>=800 & rain<1100, meanonly
  scalar `outcome0'=r(mean)
  return scalar diff_outcome=`outcome1'-`outcome0'
end

****Participation
set seed 12345
bootstrap "rd_sharp lny" impact_sharp=r(diff_outcome), reps(100) nowarn
gen t_impact_sharp=_b[impact_sharp]/_se[impact_sharp]
sum t_impact_sharp
```

10. What is the impact of adoption of irrigation on log of yields using the sharp RD technique? How does this compare with your previous results using Diff-in-Diff, IV, and FE?
11. The program is set up to look at the mean of the estimators in the range where rainfall is between 450 and 800 and to compare it to the mean of the estimators in the range where land is between 800 and 1100. What happens if you make that range smaller around the cut- off point, comparing the range 500 to 800 to 800 to 1000?

Now, let's explore fuzzy discontinuity regression. Assume that the irrigation program involved extension visits to those farm households that had less than 800mm of rain. Some of those households adopted irrigation and others did not. Also, some households with rainfall above 800mm were able to adopt irrigation, and others were not. Now re-load the `RD.dta` file

(you may need to calculate the log of rainfall again). Now, run the following fuzzy discontinuity regression:

```
*Program for Fuzzy Discontinuity
capture prog drop rd_fuzzy
prog rd_fuzzy, rclass
  args treatment outcome
  confirm var `treatment'
  confirm var `outcome'
  tempname treatrd1 treatrd0 outrd1 outrd0 treat1 treat0 outcome1
outcome0
  lpoly `treatment' lnr if rain<800, gen(`treatrd1') at(lnr) nogr tri
w(3) d(1)
  lpoly `treatment' lnr if rain>=800, gen(`treatrd0') at(lnr) nogr tri
w(3) d(1)
  lpoly `outcome' lnr if rain<800, gen(`outrd1') at(lnr) nogr tri w(3)
d(1)
  lpoly `outcome' lnr if rain>=800, gen(`outrd0') at(lnr) nogr tri w(3)
d(1)
  sum `treatrd1' if rain>=450 & rain<=1000, meanonly
  scalar `treat1'=r(mean)
  sum `treatrd0' if rain>=450 & rain<=1000, meanonly
  scalar `treat0'=r(mean)
  sum `outrd1' if rain>=450 & rain<=1000, meanonly
  scalar `outcome1'=r(mean)
  sum `outrd0' if rain>=450 & rain<=1000, meanonly
  scalar `outcome0'=r(mean)

  return scalar impact=(`outcome1'-`outcome0')/(`treat1'-`treat0')
end

*Participation
set seed 12345
bootstrap "rd_fuzzy irr lny" impact_fuzzy=r(impact), reps(100) nowarn
gen t_impact_fuzzy=_b[impact_fuzzy]/_se[impact_fuzzy]
sum t_impact_fuzzy
```

12. What do you conclude about the impact of microcredit on expenditure based on the two regression discontinuity estimators? How are the estimates different and why might that be?

Now that you have used 4 different techniques (Diff-in-Diff, IV, Panel, and RD) to estimate the impact of irrigation on yields, go back and look at your previous problem sets.

13. What do you think is the most appropriate technique, and why?