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 **Rstudio** and save an `.R` script 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.csv`. Ensure that the data set has the log transformed variables. If it does not, return to your `.R` script 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.

The package we will use for panel analysis in **R** is called *plm*. The detailed package manual can be found at https://cran.r-project.org/web/packages/plm/vignettes/plm.pdf. The first thing we will need to do to perform the panel analysis using the *plm* package is to tell **R** the index attribute that describe the individual and time dimensions of the data. We do so with the following command:

```
df <- read.csv("VDSA_Prod_Data_Ref.csv")
pdf <- pdata.frame(df, index=c("prcl_id","sur_yr"))
```

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.

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?

To run a pooled OLS using *plm*, everything is written exactly as in a usual *lm* call, however you need to specify the *pooled model* option:

```
pm1 <- plm(lny ~ ... , data = pdf, model = "pooling")
```

To get the summary table with clustered standard errors at the individual parcel level, you can use the following extended summary command:

```
summary(pm1, vcov=vcovHC(pm1,type="HC0",cluster="group"))
```

2. Run the pooled OLS again but include time dummies to control for year-to-year effects. You can use year as a categorical variable using the command `factor(sur_yr)` when you add it in the regression formula. 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 can do so by changing the model argument to "random". 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 can write: `phtest(pm2, pm3)`. 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. Add vdsa_hh_id to your list of independent variables (you will need to transform it into a categorical *factor* variable). 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 change the model argument to "within". Why might you use 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? Tabulate the variable irr by year. 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, use the commands:

```
df$sur_yr.vil_id <- paste(df$sur_yr, df$vil_id)
pdf2 <- pdata.frame(df, index=c("prcl_id","sur_yr.vil_id"))
```

   You can then use the "within" model again, using the new data frame *pdf2*. 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? 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? What results do you see when you include this interaction term? How do you interpret the coefficient on this interaction 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 the following loop command.

```
controls <- c("lnl", "lnf", "lnm", "lnp",
              "ageH", "genderH", "sizehh",
              "lnaindex", "lnlindex", "lntot_acre", "lndist")
mean_controls <- paste0(controls, "_bar")

for(i in 1:length(controls))
{
    pdf[,mean_controls[i]] <- ave(pdf[,controls[i]],
                                  pdf[,"prcl_id"], FUN=mean)
}
```

9. Run the regression again, this time using CRE. Note that by including the time averages we have controlled for unobserved heterogeneity and could run the model as a pooled OLS. Why might you use CREs? How do the results change from the four previous models?

Regression Discontinuity Design

To conduct the RD analysis we first want to go back to our two year data set. So run the command `df <- subset(df, 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. So first, create a subset data frame using the code below:

```
sdf <- subset(df, (irr == 1 & rain < 800) | (irr == 0 & rain > 800))
```

Graph the data to make sure we have a sharp discontinuity around 800mm of rainfall.

```
plot(x = sdf$rain, y = sdf$irr)
```

Let's check to see whether the relationship between rainfall and yield is smooth around the threshold:

```
plot(x = sdf$rain, y = sdf$lny)
```

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

```
sdf$lnr <- asinh(sdf$rain)
```

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

```
rd10 <- RDestimate(lny ~ rain, data = sdf, cutpoint = 800, bw = 350)
summary(rd9)
```

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 1150. What happens if you make that range smaller around the cut- off point, comparing the range 600 to 800 to 800 to 1000?

OPTIMAL BANDWIDTH?

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 run the following fuzzy discontinuity regression with the original *df* data frame:

```
rd11 <- RDestimate(lny ~ rain + irr,
                   data = df,
                   cutpoint = 800, bw = 350)
summary(rd11)
```

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?