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 **Rstudio** the data set `VDSA_Prod_Data_Ref.csv` from Monday. Ensure that the data has the log transformed variables.

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.

1. Do a simple t-test (base command: `t.test(lny~irr, data = df)` 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.

```
# create a temporary dataframe with the output of 2010
td10 <- subset(df, sur_yr == 2010)
td10$lny10 <- td10$lny
td10 <- td10[,c("prcl_id", "lny10")]

# create dataset of 2011 observations
df11 <- subset(df, sur_yr == 2011)
df11$lny11 <- df11$lny

# merge back info from 2010 (keep only data matched in both years)
df11 <- merge(df11, td10, by = "prcl_id")

# create variable of difference
df11$lny1011 <- df11$lny11 - df11$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. In the original data frame (`df`) Construct a dummy variable indicating the interaction of being in the treatment group and the year being 2011.

```
df$irr0 <- ifelse(df$irr == 1 & df$sur_yr == 2011, df$irr, 0)
df$irr11 <- ave(df$irr0, df$prcl_id, FUN=max)
df$vdumsur_yr_2011 <- ifelse(df$sur_yr == 2011, 1, 0)
df$vdumsur_yr_2010 <- ifelse(df$sur_yr == 2010, 1, 0)
df$irr11_yr = df$irr11*df$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. Note that the variable `irr11` constructed before serves as an indicator if the parcel was treated in 2011.

```
df10 <- subset(df, sur_yr == 2010)
```

Now run the `matchit` command on the irrigation data matching on our control variables, except `lnlindex` and `lndist`. Be sure to select the common support option by setting the argument `discard = "control"`. Assign the `matchit` command to a variable called `mm`.

The `matchit` function does not handle missing variables well, so when you run it, use the following subset of the the data in the *data* argument:

```
# define list of covariates used in matching
match.covs <- c("irr11", "lnl", "lnf", "lnm", "lnp",
           "ageH", "genderH", "sizehh", "lnaindex",
           "lntot_acre", "prcl_id")

# run matching
mm <- matchit(irr11 ~ lnl + lnf + lnm + lnp +
                 ageH + genderH + sizehh +
                 lntot_acre + lnaindex,
              data = df10[,match.covs], distance = "logit",
              discard = "control", replace = T)
```

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

```
# get matched data
md <- match.data(mm)

# clean it to only include an indicator of matching
md$matched <- 1
md <- md[,c("prcl_id", "matched")]

# make matched-panel data frame (mpdf)
mpdf <- merge(df, md, by = "prcl_id")
```

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! We will use the *df* data frame in this analysis.

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 (`df$IV_landrain <- df$tot_acre*df$rain`). There

are a lot of observations with no information for rain. Make sure to exclude these lines using the command:

```
ivdf <- subset(df, is.na(IV_landrain) == FALSE)
```

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 and assign the regression to an object called `st1.`

Create a variable with predicted values using (`ivdf$irrhat <- st1$fitted.values`). 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` from AER package. Unlike **Stata**, the `ivreg` command in **R** does not have an option to automatically display the 1st stage results. Moreover, notice that each exogenous variable needs to be included as an instrument for itself in the second "block" of the function after the "|" character:

```
iv.model <- ivreg(lny  ~ irr + lnl + lnf + lni + lnm + lnp +
                    ageH + genderH + sizehh +
                    lnaindex + lnlindex + lntot_acre |
                    lnl + lnf + lni + lnm + lnp +
                    ageH + genderH + sizehh +
                    lnaindex + lnlindex + lntot_acre +
                    lndist + IV_landrain,
                  data=df)
summary(iv.model)
```

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?

The summary output of ivreg in **R** can include some base diagnostics for weak instruments. In order to display them, include the argument `diagnostics = TRUE` to the summary call

11. What does the Wu-Hausman tes results 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.

```
summary(lm(lny ~ IV_landrain +
                 lnl + lnf + lnm + lnp +
                 ageH + genderH + sizehh +
                 lnaindex + lnlindex + lntot_acre + lndist,
           data = ivdf))
```

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.

```
iv.model2 <- (ivreg(lny  ~ irr +
                        lnl + lnf + lnm + lnp +
                        ageH + genderH + sizehh +
                        lnaindex + lnlindex + lntot_acre + lndist |
                        IV_landrain + rain + tot_acre +
                        lnl + lnf + lnm + lnp +
                        ageH + genderH + sizehh +
                        lnaindex + lnlindex + lntot_acre + lndist,
                      data=ivdf))
summary(iv.model2, diagnostics = T)
```

Now, the diagnostic results should include the Sargan test.

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

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