Advanced Methods in Impact Assessment Workshop

Day 4: How to Identify a Causal Effect with Non-Experimental Data

Today we will apply the information you learned this morning regarding IVs and panel data estimators.

There are three objectives for today's exercises:

1. Examine data for potential IVs and check if those IVs are effective.
2. Implement IV estimation.
3. Implement panel data estimation using several types of panel effect and compare the outcomes.

Instrumental Variables

To get started, load into Stata the data set you have been working with. Ensure that the data set has the log transformed variables and that you have dropped data from 2012 and 2013, so that you are only working with data from 2010 and 2011. Again open a `.log` file and write your IV code in a `.do` file so you can reference it later.

First, we'll deal with selecting and instrumental variable. You'll want to first use `ssc install` to install `ivendog` and `ivreg2`.

1. 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?

Examine the data and discuss which variables would potentially make a good IV.
Next, we'll run some IV regressions.

2. Construct an instrumental variable by interacting household land ownership with average rainfall in the village. Why might this be a valid instrumental variable? What are the costs and benefits of using just average rainfall or just land ownership? What are the requirements for a valid IV?

3. Run the 2SLS piecewise:

   a. Run the first stage regression: regress the endogenous treatment variable on the instrument and our standard set of exogenous control variables.

   b. Save the predicted value using the command `predict double irrhat`

   c. Run the second stage regression: regress our outcome variable on the exogenous control variables and the predicted values from the first stage regression.

4. 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. Compare the point estimates and standard errors between the "two-stage" approach in Question 4 and the "single-stage" approach you just implemented.

Now, we'll test for the weak instrument:

5. Test for endogeneity by typing the command `ivendog`. This calls up the Wu-Hausman test. Keep in mind that the test is only valid if the IV is valid. What does this test tell you?

6. Conduct a simple "falsification test" in which you regress the log yield on the exogenous variables and the instrument. Does the instrument have a significant impact on yields? Does this mean our instrument is not valid? Why might it still be a valid instrument?

7. Now test to see if you have a weak 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`. 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.

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

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

Panel Data Techniques

For the panel data we want to use all three years of our available data. So, load into Stata the data set that contains the survey years 2010, 2011, 2012. 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. So, you should have a data set with three years of data and inverse hyperbolic sine transformed variables of per hectare inputs and outputs. Again open a `.log` file and write your panel data code in a `.do` file so you can reference it later.

Now, let's begin to implement some panel data regressions. We've already introduced `ivreg` yesterday when we used a fixed effects estimate on the data from 2010 and 2011. 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. Remember to first set the panel variable as `prcl_id` and in all of these regressions cluster the standard errors by parcel.

10. 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?

11. Create a time index (dummy variables for `sur_yr`). Run the pooled OLS again but this time include the time index. How do the results change?

12. Run the regression from Question 12 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`). How do the results change from the OLS?

13. Run again, this time using fixed effects. Why might you use fixed effects? How do the results change from OLS? From the random effects 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.

14. 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?

Now we will prepare to run a correlated random effects regression. Remember, correlated random effects 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 log(labor) log(fert) log(mech) log(pest) ageH genderH
sizehh log(aindex) log(lindex) log(tot_acre) log(dist)
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
}
```

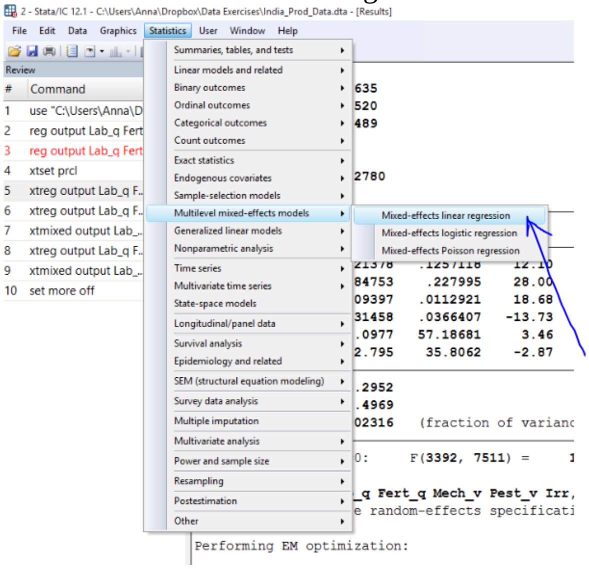
15. Run again, this time using correlated random effects. Why might you use correlated random effects? How do the results change from the four previous models?

Now we will compare the fixed effects estimates with the first difference estimates. When $t = 2$ these two methods give the same results. When $t > 2$ the results will differ. To set up the FD regression we first need to generate a time index. Run the following code:

```
gen tindex = 1 if season==1 & sur_yr==2010
replace tindex = 2 if season==2 & sur_yr==2010
replace tindex = 3 if season==1 & sur_yr==2011
replace tindex = 4 if season==2 & sur_yr==2011
replace tindex = 5 if season==1 & sur_yr==2012
replace tindex = 6 if season==2 & sur_yr==2012
xtset prcl_id tindex
```

16. Run the first difference regression. To do this, add `d.` in front of each variable of our standard regression. When might this method be appropriate? How do the results change?

Now we will introduce the Multilevel Model. Under the "Statistics" tab, go to the multilevel mixed-effects model. Under this select the "Multi-effects linear regression"

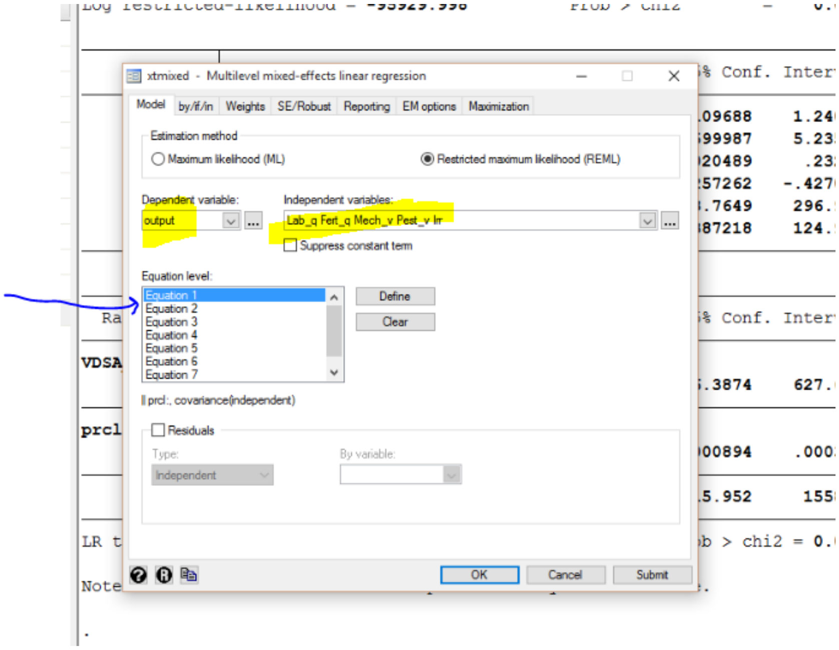


2 - Stata/IC 12.1 - C:\Users\Anna\Dropbox\Data Exercises\India_Prod_Data.dta - [Results]

#	Command
1	use "C:\Users\Anna\D
2	reg output Lab_q Fert
3	reg output Lab_q Fert
4	xtset prcl
5	xtreg output Lab_q F..
6	xtreg output Lab_q F..
7	xtmixed output Lab_..
8	xtreg output Lab_q F..
9	xtmixed output Lab_..
10	set more off

```
635
520
489

2780

21378     .1257118     12.10
84753     .227995      28.00
09397     .0112921     18.68
31458     .0366407    -13.73
.0977     57.18681      3.46
2.795     35.8062      -2.87

.2952
.4969
02316    (fraction of varianc

0:       F(3392, 7511) =      1

_q Fert_q Mech_v Pest_v Irr,
e random-effects specificati

Performing EM optimization:
```

In the window, fill in the dependent variable of interest and include the parcel level inputs as independent variables. These are marked in yellow in the image below. Next, you'll want to click on "Define" after selecting Equation 1.



xtmixed - Multilevel mixed-effects linear regression

Estimation method: ○ Maximum likelihood (ML) ● Restricted maximum likelihood (REML)

Dependent variable: output

Independent variables: Lab_q Fert_q Mech_v Pest_v Irr

□ Suppress constant term

Equation level: Equation 1 (selected), Equation 2, Equation 3, Equation 4, Equation 5, Equation 6, Equation 7

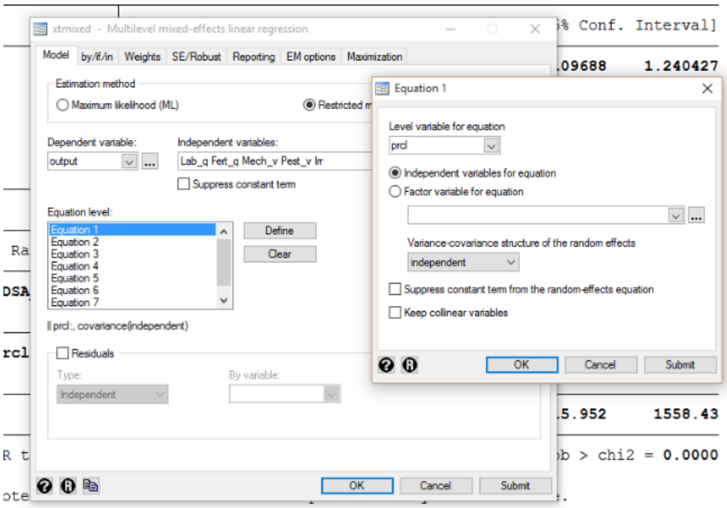
|| prcl:, covariance(independent)

□ Residuals

Type: Independent

By variable:

Finally, you'll want to set the level, in the new window. This should be parcel (`prcl_id`).



Equation 1

Level variable for equation: prcl

● Independent variables for equation

○ Factor variable for equation

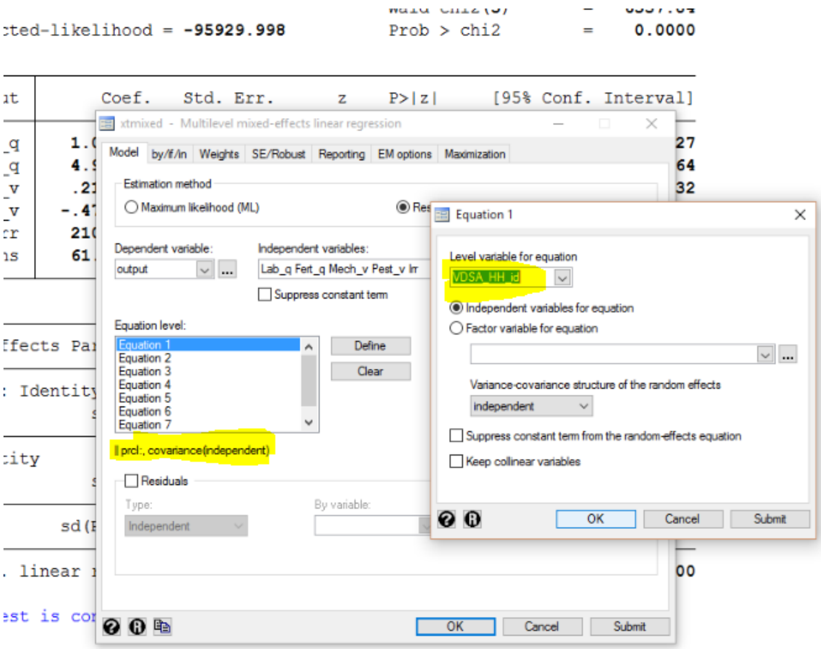
Variance-covariance structure of the random effects: independent

□ Suppress constant term from the random-effects equation

□ Keep collinear variables

17. Run this single level model, with our level of interest being the parcel. Compare these results with a fixed effect regression that includes only the parcel level inputs?

Finally, we'll run a two-level model where our levels of interest are parcel and household. This procedure is the same as above, except that you will need to define a level for both Equation 1 and Equation 2. Equation 1 should be the level with the smaller number of observations (in this case, that is the household variable (`vdsa_hh_id`)). Equation two should be the larger number (in this case, parcel (`prcl_id`)).



```
:ted-likelihood = -95929.998                      Prob > chi2       =     0.0000

ıt         |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
_q         |   1.(                                                         27
_q         |   4.!                                                         64
_v         |    .2]                                                        32
_v         |  -.47
:r         |   21(
ıs         |   61.

Effects Pa

: Identity

:ity

        sd(F

. linear 1                                                                 00

est is co
```

xtmixed - Multilevel mixed-effects linear regression

Model | by/if/in | Weights | SE/Robust | Reporting | EM options | Maximization

Estimation method: ○ Maximum likelihood (ML) ● Res

Dependent variable: output

Independent variables: Lab_q Fert_q Mech_v Pest_v Ir

□ Suppress constant term

Equation level:
Equation 1
Equation 2
Equation 3
Equation 4
Equation 5
Equation 6
Equation 7

|| prcl:, covariance(independent)

□ Residuals

Type: Independent

By variable:

Equation 1

Level variable for equation: VDSA_HH_id

● Independent variables for equation

○ Factor variable for equation

Variance-covariance structure of the random effects: independent

□ Suppress constant term from the random-effects equation

□ Keep collinear variables

18. Run this two level model. What does this model tell us that the single level model did not? Does this seem like a more effective way of isolating the effects of different variables?