Workshop on Econometric Methods for Program Evaluation
Day 1 Exercise: Mystery of the Vanishing Benefits

This exercise builds on the hypothetical program evaluation example presented by Ravallion (2001), discussed earlier in the day. You will be supplied with household survey data for the fictional country of Labas (the Labas Living Standards Survey, LSS). Our goal is to walk through ways of looking at the data, from basic summary statistics and graphics to more sophisticated ways of controlling for observed and unobserved characteristics that threaten to confound program effects.

1. To load the dataset, which is called mystery1.dta, either
   - Click File → Open, and browse to find the dataset; or
   - Alternatively type[1] `use` "*path*/mystery1.dta", `clear` , where *path* is the full name of the directory in which the dataset is stored.
2. Familiarize yourself with the dataset:
   - `describe` tells you what variables are contained.
   - `summarize` [*varlist*] gives counts, means, and other descriptive statistics of the variables you specify in *varlist.* If you leave the variable list out, it will provide this information for all variables.
3. 'Naive' tests of the effect of PROSCOL: Are average years of schooling (the variable S) higher among PROSCOL recipients? Are average test scores (the variable *score*) higher?
   - You can test this directly with Stata's `anova` command:
     `anova` S P
   - Try running a regression of S on P or of *score* on P:
     `regress` S P
     To see where the `anova` results come from, you can perform an F-test of the hypothesis that PROSCOL doesn't matter for the outcome variable.[2] After the regression, type
     `test` $P = 0$

[1] Throughout, Stata commands are written in `typewriter font`, options that you can specify are enclosed in [brackets], and generic terms that you should replace depending on the specific task at hand are written in *italics*.

[2] The F-test is based on comparing how much of the variation in the data you can explain under the restriction you're testing versus under an unrestricted model:

$$F = \frac{(RRSS - URSS)/r}{URSS/(n-k-1)} \tag{1}$$

where

URSS = unrestricted residual sum of squares
RRSS = restricted residual sum of squares obtained by imposing the restrictions of the hypothesis

- Is Proscol reaching its target population? How would you test this using the commands above?
- You might also try a graphical representation of the outcome variables, as a way of getting a feel for the data:
  `graph bar` *score* , `over`(P)

4. Add controls in the cross-section: multivariate regression.

   The simplest explanation for what may be going on here is that there are omitted variables that are correlated with *both* Proscol treatment and the outcomes of interest. Mrs. Tangential Economiste suggests you run a regression of the form

   $$S_i = a + bP_i + cX_i + \varepsilon_i \tag{2}$$

   Try including controls for household income, parents' education, and age of child. How do the results change?

   `regress` *S P hhincome feduc meduc age*
   `regress` *score P hhincome feduc meduc age*

5. Fixed effects in a cross section

   As noted by Professor Chisquare, propensity score matching is a valid approach to estimation of treatment effects under the assumption of *conditional independence*: that no unobserved determinants of schooling are also important factors determining access to Proscol. For example, if

   $$\begin{aligned} S_i &= a + bP_i + cX_i + \varepsilon_i & (3)\\ P^* &= \gamma_0 + \gamma_z Z + \nu_i & (4)\\ P &= 1[P^* > 0] & (5) \end{aligned}$$

   where Z are the observed determinants of Proscol (they may or may not be the same variables as appear in X). Then conditional independence requires that ε_i is independent of ν_i.

   An alternative approach is to suppose that there are unobserved characteristics of, say, schools that are correlated with both the likelihood of receiving Proscol treatment and schooling outcomes. If these troublesome unobservables are constant across time, *or across observed groups in a cross section*, then we can use a fixed effects estimator.

r = number of restrictions imposed by the hypothesis
n = number of observations
k = number of explanatory variables

This has a known distribution if the null hypothesis is true. Both `anova` and `test` report the probability that the value for the F test is observed under the hypothesis of no program impact.

For example, it may be the case that schools with more pro-active head teachers have both better schooling outcomes and more students receiving PROSCOL. If this is the case, then we can use a fixed effects approach that exploits only the variation in outcomes *within* schools.

6. Panel data

   Good news! The Labas Bureau of Statistics had the foresight to conduct a baseline survey. This means we can use a difference-in-differences estimator that will control for any such troublesome unobservables at the *individual* level—much stronger than just the schools unobservables controlled for above.

   - Load the complete dataset, called "*mystery2.dta*".
   - Tell Stata that these are panel data:

     `tsset` hhid wave

     The variable hhid marks the unit of observation, and $wave \in \{0, 1\}$ denotes the wave of the survey.

     Use `tab` *wave* to see that the panel is 'balanced': all observations appear in each round.
   - Now we need to use Stata's difference operator (`D.`) to generate first-differences of the variables we're interested in. For example,

     `generate` Dscore = D.score

     creates the first difference of the outcome *score*. Note that for variables that don't change across waves (for example, mother's education) these differences will be identically zero. Such variables have no effect
   - Now you can simply regress

     `regress` *Dscore Dhhincome*
   - How would you allow for growth-rate effects here? (Hint: Stata's lag operator is `L.`). Does this affect the results?

References

RAVALLION, M. (2001): "The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation," *World Bank Economic Review*, 15(1), 115–140.