Elisabeth Sadoulet and Erin Kelley

ENV ECON 118 / IAS 118 - Introductory Applied Econometrics
Assignment 5
Due Tuesday November 17 at the beginning of class

Note: You have to turn in a log file with all Stata output for this problem set. However, you will not receive credit for numbers that are not separately reported in your answer, ("see log file" is not acceptable). And remember **SSS** whenever you're asked to interpret an estimated parameter.

Exercise 1: Gasoline Taxes and Consumer Behavior - Panel Regression

In this exercise, we will reproduce a policy analysis that was published in the American Economic Journal, Applied Economics in 2014 ("Gasoline Taxes and Consumer Behavior" by Shanjun Li; Joshua Linn; Eric Muehlegger, Vol. 6, No. 4). The paper examines how gasoline tax changes affect automobile use and gasoline consumption. In contrast to the literature, the author's analysis estimates "consumer responses to gasoline taxes explicitly by decomposing retail gasoline prices into tax and tax-exclusive components".

To this end, they use a panel dataset on gasoline consumption, gasoline prices, and state and federal gasoline taxes by state-year from 1966 to 2008. We will use a subset of data from 1970 to 2004 from 34 states. The data are taken from annual issues of the Highway Statistics, published by the Federal Highway Administration. Tax-inclusive retail gasoline prices are from the Energy Information Administration State Energy Price Reports.

Variable Name	Label
id	State id
state	State Name
year	Year
fsize	Average Family Size from CPS and American Community Survey
lngca	Log gasoline consumption per adult (gallon per adult)
lngpinc	Log tax-inclusive retail price ($ per gallon)
lngp	Log tax-exclusive gasoline price ($ per gallon)
lntr	Log state and federal gasoline tax ratio (cents per gallon)
lnrma	Log road miles per adult
lncarscap	Log number of registered cars per capita
lntrkscap	Log number of registered trucks per capita
lndriverscap	Log number of licensed drivers per capita
lnincpop	Log real income per capita
railpop	Fraction of population living in metro areas with rail transport
urbanization	Fraction of the population living in metro areas

1. Briefly describe the data that you have : a) How many states? b) How many years? Report in a nicely presented table the mean, sd, min and the max of the following variables: log gasoline consumption per adult, log number of registered cars per capita, and the log tax inclusive retail price (`tabstat` command is useful for this).

2. Generate year dummy variables ($y1971, y1972, \cdots y2004$) and estimate the following equation for gasoline consumption.

$$lngca_{it} = \beta_0 + \beta_1(lngpinc) + \delta_{1971} y1971_t + \delta_{1972} y1972_t + \cdots + \delta_{2004} y2004_t + u_{it} \quad (1)$$

where i indicates a state and t indicates a year

(a) Give the meaning (economic interpretation) of β_0, β_1, δ_{1971}

(b) Report your estimates for β_0, β_1, δ_{1971} (remember SSS)

(c) Why is the year 1970 dummy excluded?

3. Consider now the following (unobserved) fixed effects model:

$$lngca_{it} = \beta_0 + \beta_1(lngpinc) + \delta_t + a_i + u_{it} \tag{2}$$

(a) What is the interpretation of the fixed effects terms a_i?

(b) Why are we adding these fixed effects, as opposed to estimating model (1)? In other words, what do these fixed effects control for in the regression?

(c) How does the estimated coefficient $\hat{\beta}_1$ change? Explain.

4. Now consider the following (unobserved) fixed effects model:

$$\begin{aligned} lngca_{it} = \beta_0 + \beta_1(lngpinc)_{it} + \beta_2 fsize_{it} + \beta_3 lnrma_{it} + \beta_4 lnincpop_{it} + \beta_5 lncarscap_{it} + \beta_6 lntrkscap_{it} \\ + \beta_7 lndriverscap_{it} + \beta_8 urbanization_{it} + \beta_9 railpop_{it} + \delta_t + a_i + u_{it} \end{aligned} \tag{3}$$

(a) Why do we add the additional covariates here?

(b) Could we add a covariate for the geographic size of the state in square kilometers?

(c) Give a very precise interpretation of the true β_4, β_9. Then report (SSS) for both parameters.

5. What are the assumptions necessary for the parameters of model (3) to be unbiased? Do you think they are likely to hold? Whatever position you take, give your argument.

6. Lastly, consider the following model:

$$lngca_{it} = \beta_0 + \beta_1(lngp) + \beta_2(lntr) + \delta_1 y1971_t + \delta_2 y1972_t + \cdots + \delta_{34} y2004_t + a_i + u_{it} \tag{4}$$

Mirroring the paper, we separate the gasoline price into the tax-exclusive price (price of gasoline net of tax) and tax components.

(a) Report your results for $\hat{\beta}_1$, $\hat{\beta}_2$. Compare the magnitude of these two estimates - is it in the direction you might expect?

(b) Compared with model (2), which model is a better fit to the data ?

(c) Give an explanation as to why we might be want to separate the gasoline price in this way?

Exercise 2: Police Enforcement and Traffic Fatalities - Double difference

In this exercise, we will reproduce a policy analysis that was published in the American Economic Journal: Economic Policy in 2014 ("Life and Death in the Fast Lane: Police Enforcement and Traffic Fatalities" by Gregory DeAngelo and Benjamin Hansen, Vol. 6, No. 2). The paper estimates the "causal effect of highway patrol officers on traffic fatalities and serious injuries by exploiting a mass layoff of state police in Oregon that reduced the likelihood that speeders were apprehended by police." For this study, Idaho and Washington serve as states proximate to Oregon which have similar demographics, economic trends, and weather patterns" We refer to this mass layoff as a policy change or policy reform.

We are using a selected sub-sample of the original data set for Idaho, Washington and Oregon spanning 2000-2004. The policy change was enacted in 2003. As a result, we have two years of "post" data from the years after the policy change (2003 and 2004) as well as two years of "pre" data from the years before the policy change (2001 and 2002). Finally, we have one year of data from before the pre-period (2000) There are a total of 180 observations and the data is recorded on a monthly basis for each of the three states.

Variable Name	Label
lndrunk_fatals	Log Traffic Fatalities related to drunk driving per year
post	= 1 for the 2 years after the Policy Layoff ($years = 2003$ & 2004)
	= 0 for the 2 years before the Policy Layoff ($years = 2001$ & 2002)
	= . for the year 2000
pre	= 1 for the 2 years before the Policy Layoff ($years = 2001$ & 2002)
	= 0 for the year 2000
	= . for the 2 years after the Policy Layoff ($years = 2003$ & 2004)
Oregon	= 1 if Oregon
Idaho	= 1 if Idaho
Washington	= 1 if Washington
month	Month
year	Year
temp	Temperature
precip	Precipitation
max_speed	Maximum Speed Limit
un_rate	Unemployment Rate

1. Generate a summary table with two columns and 3 rows. There should be two columns: one for Oregon (Treatment column) and one for non-Oregon states (Control column). There should be three rows: one for the pre pre-period (year 2000), one for the pre-period (years 2001-2002) and one for the post-period (years 2003-2004). Within each cell compute the mean of log traffic fatalities related to drunk driving ($lndrunk\_fatals$)

2. We want to compare mean log fatalities related to drunk driving pre and post the policy reform in Oregon. Is there a statistically significant difference in log traffic fatalities? We also want to compare the mean log traffic fatatlities related to drunk driving pre and post the policy reform outside Oregon (Washington and Idaho combined). Is there a statistically significant difference in log traffic fatalities?

   Stata Tip: Here it's necessary to do t-tests comparing means between two groups (pre and post) but only for the "treated" group or only for the "control" group. The command for this kind of t-test is

```
ttest var1 if var2==n, by(groupvar)
```

where $var1$ is the variable whose means we are comparing; the if statement $var2 == n$ restricts the test to those in a specific treatment or control group; and *groupvar* specifies the groups you want to compare the means across (e.g pre/post).

3. Let's proceed with a difference-in-difference estimation:
   (a) Write an equation that will give you the Difference-in-Differences estimator for the impact of the policy on log traffic fatalities related to drunk driving per year.
   (b) Give a verbal interpretation of the estimator.
   (c) Perform the estimation
   (d) Interpret (economic meaning) $\beta_0, \beta_1, \beta_2, \beta_3$
   (e) From the results of your estimation, what do you conclude?
4. What key assumption do you need to make for your regression in question 2 to estimate the causal effect of the introduction of the benefits policy on traffic fatalities?
5. Let's think about the difference-in-difference assumption:
   (a) How would you provide evidence for this assumption? What data would you need to complete this test?
   (b) Complete this test and conclude.
6. Draw a graph that shows that illustrates that both a) the key assumption is satisfied and b) the difference- in -difference estimator is significant. Comment.
7. Add dummy variables for the different months. Specify a test for whether or not the month variables help you explain fatalities and perform the test without the Stata command (write out all 5 steps of hypothesis testing that we have used in class, including the equation for your test statistic, using the 5% significance level).
8. Add controls for *temp*, *precip* max_speed, un_rate. Report the estimated coefficients on these variables and interpret (SSS) for max_speed and *temp* only. Would you say that the number of fatalities depends on temperature, max_speed, or both?
9. Compare the estimated impact from the regressions in questions (3), (7), and (8):
   (a) Comment on how the coefficient on $oregon * post$ changes between <u>each</u> model
   (b) Comment on what might explain the change in the coefficient on $oregon*post$ from the regression in equation (3) to (7)
   (c) Conclude on the robustness of the estimated impact of the policy.