Problem set 3
ECON 60303
(Due: Tuesday, February 15, 2011)

Bill Evans
Spring 2011

1. In this problem, we are going to examine the roll of measurement error in OLS and fixed-effects models. Load up the data set psid1.dta, construct wages and the natural log of wages (call it wagel). The only covariate in these regressions will be the variable "tenure".

   a) Generate the variance of tenure. Call this $\hat{\sigma}_t^2$.

   b) Run an OLS model of wagel on tenure (ignore the panel nature of the data). Then with areg, run the same regression absorbing the individual "id". What are the coefficients and standard errors on tenure in these cases? Call these least-square estimates $\hat{\beta}_{w/out}^{ols}$ and $\hat{\beta}_{w/out}^{fe}$.

   c) Next, you are going to draw random errors, add them to tenure to generate measurement error, estimate the OLS and fixed-effect models, and do this 1000 times, keeping the estimates from every iteration. Do not use the bootstrap command – you will need to model this program after the "wild bootstrap" sample program.

   I want the measurement error added to tenure to be from a normal distribution with a mean of 0 and a variance of 4. Given a draw to a normal distribution, you can easily construct a draw from a standard normal distribution with a mean of zero and a variance of 1 by using a draw to a uniform distribution, take the incerse CDF, and multiplying this by 2. Call the measurement error v1 and use the following command to construct the variable:

   gen v=2*invnorm(uniform())

   Add this to tenure 2 and call the new tenure variable with measurement error tenure2

   gen tenure2=tenure+v;

   For each bootstrap iteration, keep four results a) the mean of v, b) the variance of v (call this $\hat{\sigma}_v^2(i)$ c) the OLS estimates with measurement error (call this $\hat{\beta}_{with}^{ols}(i)$), and d) the fixed-effect estimates with measurement error (call this $\hat{\beta}_{with}^{fe}(i)$).

   From the 1000 replications, generate the following values

$$\bar{\hat{\sigma}}_v^2 = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\sigma}_v^2(i)$$
$$\bar{\hat{\beta}}_{with}^{OLS} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\beta}_{with}^{OLS}(i)$$
$$\bar{\hat{\beta}}_{with}^{fe} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\beta}_{with}^{fe}(i)$$

d) What are the estimates for $\bar{\hat{\sigma}}_v^2$ $\bar{\hat{\beta}}_{with}^{OLS}$ and $\bar{\hat{\beta}}_{with}^{OLS}$

e) Construct the following measure of the reliability ratio across all 1000 trials:

$$\hat{\theta} = \frac{\hat{\sigma}_t^2}{\bar{\hat{\sigma}}_v^2 + \hat{\sigma}_t^2} ?$$

f) Compare the ratio in e) to the ratio $\frac{\bar{\hat{\beta}}_{with}^{OLS}}{\hat{\beta}_{w/out}^{ols}}$. Do these results make sense?

g) Construct the ratio of estimates in the fixed-effects model with the results from part b) $\frac{\bar{\hat{\beta}}_{with}^{fe}}{\hat{\beta}_{w/out}^{fe}}$

h) Devise a method that measures the within-panel correlation in tenure. Using this number and the reliability ratio in e), what should be the attenuation bias associated with measurement error in this case? Compare this to the ratio of estimates from part g).

2. In this problem, we will examine the problems associated with correlation in errors in panel data models, how the cluster command helps solve this problem, and the higher Type I error rate in models with small numbers of clusters.

In basic ECON 101, we teach that the burden of an excise tax is "shared" by consumers and producers: the tax should increase by an amount less than the excise tax. In this case, we examine the impact of excise tax changes on retail prices for cigarettes. This product is taxed at the state and federal level and we can examine how tax hike alter retail prices by using a fixed effect model. I have monthly sales data over a 6 year period from checkout scanners for name brand cigarettes from a sample of supermarkets in 29 states. The scanner data records the exact transactions price (retail price including any taxes minus any coupon discounts) for each pack of cigarettes sold in sampled stores. The data set has 6 years * 12 months * 29 states = 2088 observations. The data set is names cigarette_taxes.dta and the variables in the data set are defined below:

Variable	Definition
Fips	State fips code, 1-56 numeric code that identifies a state.
Year	Year (2001-2006)
Month	Month (1-12)
Time	Time trend, =1 in Jan 2001, =2 in Feb 2001, =72 in Dec 2006.
Price	Real price (in cents per pack) for a pack of cigarettes. Prices are in December 2008 values.
Tax	Total tax, state+federal, in cents per pack, for a pack of cigarettes. Prices are in December 2008 values.
transition	The variable equals 0 in all months except in a month when the tax changes in the middle of a month. In those months, the value is the new real tax rate in cents/pack. Prices are in December 2008 values.

The impact of the tax hike can easily be captured within a fixed effects model – regress real prices on state, month, and year effects plus the real tax and the transition variable (this variable is needed because the timing of some tax changes is not at the start of the month and this sort of controls for it.) The coefficient of interest is on the tax variable.

a. Using the xi command, construct fixed effects for state, year, and month. Next, run a fixed-effect model regressing price on the fixed-effects, the transition and the tax. What is the coefficient/standard error on the tax coefficient? Call this coefficient $\hat{\beta}_{tax}$. What is the p-value on the hypothesis $H_0 : \beta_{tax} = 0$? What is the p-value on the hypothesis $H_0 : \beta_{tax} = 1$?

b. Output the residuals from the fixed-effect model, sort the data by state and time, then generate the lag of the residual. If the name of the residual is titled "error" then you can generate these results by using the code

```
predict error, resid
sort state time
by state: gen errorlag=error[_n-1]
```

Note that in generating within panel lags, you lose the 1st observation in every panel.

Next, regress error on error1, without a constant, and obtain an estimate of the following equation

$$\hat{\varepsilon}_{it} = \rho\hat{\varepsilon}_{it-1} + \xi_{it}$$

This can be estimated by the stata command

```
reg error errorlag, noc
```

Call the coefficient on `errorlag rhohat`. This demonstrates that there is tremendous autocorrelation in the data and therefore, we anticipate that the estimates from equation (a) may be subject to high Type I error rates. You can correct for AR(1) easily by "rho differencing" the data. In this case, we have a bunch of fixed effects that do not need to be differenced), the dependent variable y and the two x's. Use the text beow, fill in your estimate of rhohat

```
gen rhohat= /*(type in your answer)*/
by state:  pricelag=price[_n-1]
by state: transitionlag=transition[_n-1]
by state: taxlag=tax[_n-1]
gen priced=price-rhohat*pricelag
gen transd=transition-rhohat*transitionlag
gen taxd=tax-rhohat*taxlag
reg priced _I* transd taxd
```

This AR(1) correction in panel data can easily be obtained by the following two lines

xtset state time
xtregar price _Im* _Iy* transition tax, fe twostep rhotype(reg)

The results from this example should be identical to the answer from above. Note also that the standard error from this model on tax should increase over the value in part a). With the results from the xtregar model to generate the p-value on the tests of the null hypothesis $H_0: \beta_{tax} = 0$ and $H_0: \beta_{tax} = 1$.

c. Returning to the model in part a) now re-estimate the basic panel data model and cluster the standard errors on state. Note that there is a large increase in the size of the standard error on tax over that in both part a) and b). Generate the p-value on the tests of the null hypothesis $H_0: \beta_{tax} = 0$ and $H_0: \beta_{tax} = 1$.

d. The problem with the estimate in part c) is that there are a limited number of states (29) and the basic Huber-white sandwich correction has poor properties when the number of clusters is small. Take the wild boostrap program from class, and re-write the code so that you can test the null hypothesis that $H_0: \beta_{tax} = 0$ using this procedure. Remember that you need to generate the errors assuming the null so run the model with tax missing from the model. Compare the p-value of the test statistic $H_0: \beta_{tax} = 0$ with the estimates from parts a) b) and c).

e. In this problem, you are to re-write your program from d) and test the null that $H_0: \beta_{tax} = 1$. Compare the p-values on these statistics from parts a) b) and c).

After part e) you should be able to fill in the following table.

P-values of tests of hypothesis, 29 state sample

Hypothesis	OLS	AR(1)	Custer	Wild Bootstrap
$H_0: \beta_{tax} = 0$				
$H_0: \beta_{tax} = 1$				

f. Now, delete the odd numbered fips code variables. This should leave you with only 11 states. Re-estimate parts a) c) d) and e) and fill in the following

P-values of tests of hypothesis, 29 state sample

Hypothesis	OLS	Custer	Wild Bootstrap
$H_0: \beta_{tax} = 0$			
$H_0: \beta_{tax} = 1$			