Problem Set 7
Economics 30331
Due at the start of class, Wednesday, April 9th

Bill Evans
Spring 2018

1. Examining the role of measurement error. Use the data set cps87.dta and construct the natural log of weekly earnings (ln_weekly) and regress this on years of education (years_educ)

   a) What is the coefficient and standard error on the years_educ variable?
   b) If you invoke the following command in stata, you will generate a standard normal random variable that has a mean of zero and a standard deviation of 1.

   ```
   gen v2=rnormal(0,1)
   ```

   Take the variable v2 and assume this is some kind of measurement error and generate a noisy measure of education by adding this to years_educ

   ```
   gen educ2=years_educ+v2
   ```

   What is the mean and standard deviation of years_educ, educ2 and v2?

   c) In our model of measurement error, think of educ2 as x^*, years_educ as x and v2 as v. Now, regress ln_weekly on educ2 instead of years_educ. What is the coefficient and standard error on educ2? Does this answer make sense? To answer this question, calculate the reliability ratio $\theta = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}$ and compare the ratio of estimates generated in parts a and c.

   d) Next, generate two new variables

   ```
   gen v3=rnormal(0,2)
   gen educ3=years_educ+v3
   ```

   v3 is now a normally distributed random variable with a mean of 0 and a standard deviation of 2 (variance of 4). What is the mean and standard deviation of years_educ, educ3 and v3? Next, repeat the steps in question c? what has happened to the coefficient on educ3 compared to the results in part a)? What is the reliability ratio in this case?

   e) Next, construct two additional variables

   ```
   gen y2=ln_weekly+v2
   gen y3=ln_weekly+v3
   ```

   Next, run three regressions: ln_weekly on years_educ, y2 on years_educ and y3 on years_educ. What is the coefficient and standard error on years_educ in these three regressions? Do these results make sense?

2. A researcher has collected data on alcohol consumption for 50 students each from 100 different colleges (50,000 observations). The outcome of interest (y_i) is number of drinks consumed in the past 30 days. The researchers have developed an index (x_i) that represents the strictness of a college's alcohol use policy with higher values meaning a more strict policy. The authors are interested in estimating an equation of the form $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$

   The researchers are concerned about measurement error in y. In particular, they believe that students at schools with stricter alcohol policies may be less likely to report actual drinking because they are not supposed to be drinking. In this case, let y_i be actual drinking and y_i^* be reported drinking where $y_i^* = y_i + u_i$. In this case, we

will continue to assume $E[u_i]=0$, $cov(x_i, \varepsilon_i)=0$, but we will assume measurement error systematic such that $cov(u_i, x_i)<0$ (students are more likely to undertstate drinking in areas with the strictest anti-drinking rules. In this case, with this form of measurement error, will the OLS estimate generated from a regression of y_i^* on x_i still be unbiased? If not, is the estimate biased up or down? (HINT: to figure out the properties of an estimate, substitute in the truth).

3. There is a useful command in stata called "sample x" that takes an x percent random sample from the original data set. So if you say sample 10 it is a 10% random sample, and sample 0.1 is a .1% random sample. Using the data set from question 1, there are almost 20,000 observations sample 0.1 will select .1% or 20 observations from the original sample.

   The program random_draws.do which is available on the web page is repeated below. The program draws random that equal .1% of the original data, run a regression of y on x, and repeats this 20 times. The command "preserve" saves the data and the command restore allow you to retrieve the data.

   a) Run this program (from the command line type do random_draws then hit return). Examine the output. What is the min and max value of the coefficient on x?
   b) Change the program to read "sample 10" which generates a 10% random sample. Run the program – what are the min and max values of the estimated coefficient on x?
   c) Change the program to read "sample 50" which generates a 50% random sample. Run the program – what are the min and max values of the estimated coefficient on x?

```
* read in stata data set cps87.dta
use cps87

gen y=ln(weekly_earn)
gen x=years_educ
preserve

forvalues i = 1/20 {
sample 0.1
reg y x
restore
preserve
}
```

4. A frequently occurring time series process is a "random walk with drift" which can be described by the equation $y_t = \alpha + y_{t-1} + \varepsilon_t$. Assume that $E[\varepsilon_t]=0$, $Var[\varepsilon_t]=\sigma_\varepsilon^2$, $\text{cov}(\varepsilon_t, \varepsilon_{t-1})=0$, and $\text{cov}(\varepsilon_{t-k}, y_{t-k-1})=0$ for all k. Let the initial value y_0 be deterministic (not random) so $E[y_0]=y_0$ so $Var[y_0]=0$. What is a) $E[y_t]$? b) $Var(y_t)$?

5. A frequently occurring time series process is a "random walk with trend" which can be described by the equation $y_t = y_{t-1} + \delta t + \varepsilon_t$ where t is a time trend. Assume that $E[\varepsilon_t]=0$, $Var[\varepsilon_t]=\sigma_\varepsilon^2$, $\text{cov}(\varepsilon_t, \varepsilon_{t-1})=0$, and $\text{cov}(\varepsilon_{t-k}, y_{t-k-1})=0$ for all k. Write an equation for the "first difference" in this series, $\Delta y_t = y_t - y_{t-1}$. What is a) $E[\Delta y_t]$? b) $Var(\Delta y_t)$? Has first differencing made this series stationary?

6. Suppose a time series variable y_t can be described by the random walk process where $y_t = y_{t-1} + \varepsilon_t$
In this problem, we are going to use the basic process to predict the value of *y* in the future. Suppose we observe y_t and we are interested in predicting what *y* will be next period.

a) What is $E[y_{t+1} \mid y_t]$? HINT: Write y_{t+1} as a function of y_t then take expectations conditional on that y_t is observed.

b) What is $E[y_{t+2} \mid y_t]$? HINT: Write the equation and continually substitute until you have y_t on the right hand side and a bunch of errors, then take expectations.

c) What is $E[y_{t+3} \mid y_t]$? HINT: Write the equation and continually substitute until you have y_t on the right hand side and a bunch of errors, then take expectations.

d) Generalize the results in a) through c) and write and equation for W $E[y_{t+h} \mid y_t]$ for h>0.

7. Right below where you downloaded this problem set from the class web page is a program that is called random_random_walks.do. This program generates 5 random walks of 100 observations each. These random series are generated by "seeding" a random number generator. Select a 4-digit seed for the random number generator (the last 4 digits of your Social Security Number) and then run this program. This will produce a data set called random_series with 6 variables: a time index (t) and 5 series, y1-y5.

a) Load the random_series data again, then run 10 different regressions, all the unique possible combinations of regressions where the lowest numbers series is the dependent variable (reg y1 y2; reg y1 y3; reg y1 y4; reg y1 y5; reg y2 y3….reg y4 y5). Think of this model as $y_{at} = \beta_0 + y_{bt}\beta_1 + \varepsilon_t$. Of these 10 unique regressions, how many times did you reject the null $H_o : \beta_1 = 0$?

b) Now, construct the first difference in each series. The first difference for series 1 can be generated with the commands

```
gen yl1=y1[_n-1]
gen dy1=y1-yl1
```

Call these dy1 – dy5. Run the 10 unique regressions where the lowest numbered series is the dependent variable (reg dy1 dy2; reg dy1 dy3…; reg dy4 dy5). Think of this model as $\Delta y_{at} = \beta_0 + \Delta y_{bt}\beta_1 + \varepsilon_t$. Of these 10 unique regressions, how many times did you reject the null $H_o : \beta_1 = 0$?

8. In this problem, with a simple data set, you are asked to replicate the results in Wilcox. In the stata data set wilcox.dta I have monthly data from January 1965 through December 1985 for five variables: month, year, a time trend (time), real per capita retail sales per month (retail) and real Social Security payments per recipient (oasi).

In stata, do the following exercises. First, set the data as time series with "time" as the time series. Next, generate natural logs of retail sales and OASI payments and take the lags of both variables

```
tsset time
gen retail_ln=ln(retail)
gen oasi_ln=ln(oasi)
gen retail_ln_1=retail_ln[_n-1]
gen oasi_ln_1=oasi_ln[_n-1]
gen d_retail_ln=retail_ln - retail_ln_1
gen d_oasi_ln=oasi_ln - oasi_ln_1
gen d_oas_ln_1=d_oasi_ln[_n-1]
```

a. Regress the 1st difference in ln(retail sales) (d_retail_ln) on the 1st difference in ln(oasi) (d_oasi_ln), the lag of the first difference in ln(OASI), then a time trend – this is essentially the model estimated by Wilcox. Can you reject or not reject the null hypothesis that coefficients on d_oasi_ln =0? These numbers are not the same as in the Wilcox paper but they are in the neighborhood.

b. Test the null that the coefficients on d_oasi_ln and d_oasi_ln_1 are both equal to zero. What is the estimate for $\hat{F}$ and what is the p-value on this test?

c. In this part of the problem, we will examine what happens when we ignore the fact the outcome is a non-stationary series and estimate a model in LEVELS instead of first differences. Specifically, regress retail_ln on oasi_ln, the lag of oasi_ln, and the time trend – do not first difference the data. How do the coefficients on this regression compare to the estimates in part a? Re-do part b with this regression – what is the test statistic that ln(oasi) and its lage both have a coefficient of zero? Why are the results so different in parts a) and c)?