Financial Econometrics
FINN/ECON 6219, Fall 2023

Problem Set 1
Due September 6 at the beginning of class
(include the names of all group members on solutions)

You may work in a group of 2 or 3 students

Attach a copy of your Stata code at the end of your solutions

Question 1.

This question uses data from Federal Reserve Bank of St. Louis (FRED). Install the `freduse` utility in Stata (see lecture note 1) and complete the following.

a. Load the monthly seasonally-adjusted index for U.S. Industrial Production into Stata (search for "industrial production" in FRED and select "United States of America" from the left-side menu to find it). The `daten` variable is a daily date that is used to indicate the month and year of the observation. Use the `mofd` command to generate a new variable called `datem` that contains only the month and year, and use the `tsset` command to declare `datem` as a monthly time index. Plot the industrial production index for 1972m1 to 2017m12 with the `tsline` command (by, for example, using `if tin( )` to restrict the sample period). Label the vertical and horizontal axes appropriately and give the graph a suitable title. Copy and paste the graph into your solutions.

b. Repeat part (a) using monthly non-seasonally-adjusted data for "Industrial Production: Manufacturing (NAICS)." Copy and paste the graph into your solutions.

c. Plot the logarithm of each series (Industrial Production and Industrial Production: Manufacturing) for 1972m1 to 2017m12 on the same graph. There are many ways to do this. For example, you can use `drop _all` to get rid of all previously defined variables, use a single `freduse` command to bring in both series at the same time, and construct a monthly time index as in parts (a) and (b). Copy and paste the graph into your solutions.

Question 2.

Let x denote a random variable that has expected value $E[x] = \mu$. Let $\hat{x}$ denote a forecast of x that is constructed from sample information. The mean squared error (MSE) of the forecast, which is defined as $MSE = E[(x - \hat{x})^2]$, can be expressed as $MSE = Var(x) + Var(\hat{x}) + (E(\hat{x}) - \mu)^2$. The first term captures the fundamental uncertainty generated by random variation in x. It is always present. The second and third terms are the variance and squared bias of the forecast (i.e., $bias = E(\hat{x}) - \mu$). We often compare competing forecasts in terms of variance and bias.

Now consider a simple Stata program.

```
program findmean, rclass
      drop _all
      set obs 100
      gen x = rnormal(-1,sqrt(2))
      quietly summarize x, detail
      return scalar mu = r(mean)
end
```

It generates a random sample of 100 observations from a N(-1,2) distribution and returns the sample mean. If you copy the program into a do-file and then issue the commands `findmean` and `display r(mu)` afterwards, then Stata will display the sample mean.

a. Use a Monte Carol experiment with 5000 repetitions to investigate the variance and bias of the sample mean. The syntax is `simulate xmean=r(mu), reps(5000): findmean`. First, set the seed to 6219 (if you don't then your results will not match mine). Next, issue the `simulate` command. Finally, use the `summarize, detail` command to find the variance and bias of the sample mean. Report the variance and bias in your solutions.
b. Repeat part (a) using the sample median instead of the sample mean as the forecast. Note that you will have to modify the program to return the sample median (see the "Stored results" section of the Stata documentation on the `summarize` command to figure out how). Based on the results, would you rather use the sample mean or sample median as your forecast of x? Explain you answer.
c. Modify the program of part (a) to return the sample mean of y=exp(x). The expected value of y is 1. Follow the same steps as in part (a) to find the variance and bias of the sample mean of y. Report these values. Follow the same steps as in part (b) to find the variance and bias of the sample median of y. Report these values. Based on the results, would you rather use the sample mean or sample median as your forecast of y? Explain you answer. If it differs from that in part (a), then explain why.

Question 3.

The R-squared produced by a regression model is often interpreted in inappropriate ways. Copy the following link into your browser and read the article.

http://econbrowser.com/archives/2014/01/on_rsquared_and

Download the Shiller data from Canvas, which contains monthly observations on the S&P 500 index starting in 1871m1, and complete the following. You can create a monthly time index for the data with the commands `gen time=tm(1871m1)+_n-1, format time %tm`, and `tsset time`.

a. Use the `regress` command to replicate the two S&P index regressions (equations (1) and (2)) for the sample period examined in the article. Copy and paste the output from the `regress` command into your solutions. Do the results match those reported in the article?
b. Use the `regress` command to replicate the two interest rate regressions (equations (3) and (4)) for the sample period examined in the article (note that the data are from FRED). Copy and paste the output from the `regress` command into your solutions. Do the results match those reported in the article?
c. The regressions show that the R-squared falls dramatically if we use differences rather than levels as the dependent variable in the regressions. But it's important to think about how we should interpret the R-squared for a regression that uses differences. We can use simulations to gains some insights. Consider a simple Stata program.

```
program findrsq, rclass
      drop _all
      set obs 100
      gen time= _n
      tsset time
      gen y = rnormal(0,1)
      quietly regress y L.y
      return scalar rsq = e(r2)
      quietly regress D.y L.y
      return scalar rsqdiff = e(r2)
end
```

It generates *y* by randomly sampling 100 observations from a N(0,1) distribution and returns the R-squared for (i) a regression of *y* on the lagged value of *y*, and (ii) a regression of the difference in *y* on the lagged value of *y*. Use a Monte Carol experiment with 1000 repetitions to investigate the average R-squared values produced by the regressions. The syntax is `simulate r2=r(rsq) r2diff=r(rsqdiff), reps(1000): findrsq`. First, set the seed to 6219 (if you don't then your results will not match mine). Next, issue the `simulate` command. Finally, use the `summarize` command to find the average R-squared values. Report them in your solutions. How do you explain your results?