# MFE230E Problem Set 2

## Due April 1, 2024

You may NOT use built-in regressions routines for this problem set. You need to calculate the $\mathbf{Y}$ and $\mathbf{X}$ matrices, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, standard errors, $R^2$, AIC/BIC, and any other regression diagnostics yourself. For the DF and ADF tests, your code should compute the test statistics but you can use the critical values from `statsmodels.tsa.stattools.adfuller`.

1. Assume that $x, \epsilon \sim i.i.d. N(0,1)$ and let $y = x + \epsilon$.

   (a) Consider the regression $y = \beta x + \epsilon$. Does this model satisfy the assumptions of the classical and/or asymptotic OLS models? Which OLS model is appropriate in this case? What are the theoretical properties of the OLS estimator $\hat{\beta}$? What is the theoretical standard error of $\hat{\beta}$ for a sample size $T$?

   (b) Perform the following Monte Carlo simulation for $T = 20, 50, 100, 500$.

   Step 1: Draw $T$ observations of $x$ and $\epsilon$ and compute the implied $y$ for each observation.

   Step 2: Compute the OLS regression $y = \beta x + \epsilon$.

   Step 3: Repeat steps 1. and 2. 10,000 times and save the $\hat{\beta}$ in each regression.

   Step 4: Plot the histogram of the distributions of the $\hat{\beta}$'s for each value of $T$.

   Step 5: Compare the actual distributions of $\hat{\beta}$ from the simulations to the theoretical distributions.

2. Download a monthly series of the ten-year Treasury yield from FRED (series: GS10). Use a sample period from April 1953 to February 2020. For this question, assume that the data is stationary.

   (a) Plot the data and perform a preliminary data analysis as discussed in class.

   (b) Plot the ACF and PACF. What do you learn from the ACF and PACF?

   (c) Compute the OLS regressions for the AR(1) and AR(2) models

$$x_t = \mu + \phi_1 x_{t-1} + \epsilon_t$$
$$x_t = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t$$

   Report the OLS coefficients, their standard errors assuming homoskedasticity, heteroskedasticity, and Newey-West and other standard regression output.

   (d) Compute the roots of the characteristic polynomials for both AR models.

   (e) Plot the impulse response functions for the AR(1) and AR(2) models.

   (f) Compare the two AR models using the criterion that were discussed in class. Which model is preferred? Why?

3. Now, let's investigate whether the T-bill data is stationary or not.

   (a) Use the methods discussed in class to test whether T-bill yields are stationary or not. Unless you find overwhelming evidence of stationarity, discuss alternative specifications in order to find a stationary model.

   (b) Consider various AR($p$) models for the stationary specification. What is the "best" AR($p$) model? Why?

   (c) Putting together everything you have learned about the T-bill data (including the results from the previous question), what is your preferred AR model for the T-bill data? Explain your choice.

4. Next, let's look at some spurious regressions. Repeat the following simulation exercise for $T = 200$ and $T = 1000$. For each $T$, simulate data with $\phi = 0, 0.9, 1$.

   (a) Simulate two independent AR(1) processes:

   $$y_t = \phi y_{t-1} + \epsilon_t$$
   $$x_t = \phi x_{t-1} + \eta_t$$

   where $\epsilon_t$ and $\eta_t$ are two i.i.d. independent standard normal random variables.

   (b) Compute the OLS regression

   $$y_t = \alpha + \beta x_t + u_t$$

   (c) Repeat (a) and (b) 10,000 times and save the OLS $\beta$, the $t$-test for $H_0 : \beta = 0$ and $R^2$ for each simulation.

   (d) Plot the histogram of the 10,000 $\beta$s, $t$-tests and $R^2$s.

   (e) Compute the percentage of rejections of $H_0$ at the 1%, 5%, and 10% levels. What are the theoretically correct rejection rates if the regressions were correctly specified?

   (f) Interpret the results. For example, describe how your results depend on $T$ and $\phi$.