Problem Set #4

MACS 30100, Dr. Evans
Due Monday, Feb. 6 at 11:30am

1. **Some income data, lognormal distribution, and SMM (10 points).** For this problem, you will use the same 200 data points from Problem Sets 2 and 3 of annual incomes of students who graduated in 2018, 2019, and 2020 from the University of Chicago M.A. Program in Computational Social Science. These data are in a single column of the text file `incomes.txt` in the PS4 folder. Incomes are reported in U.S. dollars. For this exercise, you will need to use the log normal distribution.

$$
\begin{aligned}
(LN): \quad & f(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{[\ln(x)-\mu]^2}{2\sigma^2}} \\
& \text{for} \quad 0 \leq x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0
\end{aligned}
$$

The function $f(x|\mu,\sigma^2)$ is a probability density function in that $f(x|\mu,\sigma^2) > 0$ for all x and $\int f(x|\mu,\sigma^2)dx = 1$. Note that x must be nonnegative in the lognormal distribution and σ must be strictly positive.

(a) Plot a histogram of percentages of the `income.txt` data with 30 bins. Make sure that the bins are weighted using the `normed=True` option. Make sure your plot has correct x-axis and y-axis labels as well as a plot title.

(b) Write your own function for the lognormal PDF above called `LN_pdf()`. Have this function take as inputs `xvals`, `mu`, and `sigma`, where `xvals` is either a one-dimensional $(N,)$ vector or an $N \times S$ matrix of simulated data where each column is a simulation of the N data points on income, μ is the mean of the normal distribution on which the lognormal is based, and σ is the standard deviation of the normal distribution on which the lognormal is based. This function should return an array `pdf_vals` that is the same size as `xvals` and represents the lognormal PDF values of each element of `xvals` given the parameters μ and σ. Test your function by inputting the matrix `xvals = np.array([[200.0, 270.0], [180.0, 195.5]])` with parameter values $\mu = 5.0$ and $\sigma = 1.0$.

(c) Estimate the parameters of the lognormal distribution by simulated method of moments (SMM). Create $S = 300$ simulations, each with $N = 200$ observations on income from the lognormal distribution above. Use the average income and standard deviation of income as your two moments. Use the identity matrix as your weighting matrix $\hat{\boldsymbol{W}}$. Plot your estimated lognormal PDF against the histogram from part (a). Report the value of your SMM criterion function at the estimated parameter values. Report and compare your two data moments against your two model moments at the estimated parameter values. To draw your $N \times S$ matrix of incomes from the lognormal distribution within your criterion function, set the seed to 1234 in the random number generator by typing `np.random.seed(seed=1234)`. Then draw an $N \times S$ matrix of values from a normal distribution with mean μ and standard deviation σ. Then transform those draws into the lognormal distribution by exponentiating them `numpy.exp()`.

(d) Perform the two-step SMM estimator by using your estimates from part (c) with two moments to generate an estimator for the variance covariance matrix $\hat{\boldsymbol{\Omega}}_{2step}$, which you then use to get the two-step estimator for the optimal weighting matrix $\hat{\boldsymbol{W}}_{2step}$. Report your estimates as well as the criterion function value at these estimates. Plot your estimated lognormal PDF against the histogram from part (a) and the estimated PDF from part (c). Report and compare your two data moments against your two model moments at the estimated parameter values.