Problem Set #5

MACS 30150, Dr. Evans
Due Wednesday, Feb. 12 at 11:30am

1. **Some income data, lognormal distribution, and GMM (7 points).** For this problem, you will use the same 200 data points from Problem Set 4 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 `PS5/data` folder. Incomes are reported in U.S. dollars. For this exercise, you will need to use the log normal distribution.

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

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 `density=True` option. Make sure your plot has correct x-axis and y-axis labels as well as a plot title.

   (b) Estimate the parameters of the lognormal distribution by generalized method of moments. 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 GMM criterion function at the estimated parameter values. Report and compare your two data moments against your two model moments at the estimated parameter values.

   (c) Perform the two-step GMM estimator by using your estimates from part (b) 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 (b). Report and compare your two data moments against your two model moments at the estimated parameter values.

   (d) Now estimate the lognormal PDF to fit the data by GMM using different moments. Use percent of individuals who earn less than \$75,000, percent of individuals who earn between \$75,000 and \$100,000, and percent of individuals who earn more than \$100,000 as your three moments. Use the identity matrix as your estimator for the optimal weighting matrix. Plot

your estimated lognormal PDF against the histogram from part (a). Report the value of your GMM criterion function at the estimated parameter values. Report and compare your three data moments against your three model moments at the estimated parameter values.

(e) Perform the two-step GMM estimator by using your estimates from part (d) with three 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 (d). Report and compare your three data moments against your three model moments at the estimated parameter values.

(f) Which of the four estimations from parts (b), (c), (d), and (e) fits the data best? Justify your answer.

2. **Linear regression and GMM (3 points).** Ordinary least squares (OLS) is the most common estimator for a linear regression. However, in this exercise, you will implement the more general GMM estimator of the linear regression. Assume the following linear regression model for determining what effects the number of weeks that an individual i is sick during the year ($sick_i$).

$$sick_i = \beta_0 + \beta_1 age_i + \beta_2 children_i + \beta_3 temp\_winter_i + \varepsilon_i$$

The parameters $(\beta_0, \beta_1, \beta_2, \beta_3,)$ are the parameters of the model that we want to estimate. Note that we don not have to make any assumptions about the distribution of the error terms ε_i. The variable age_i gives the age of individual i at the end of 2016 (including fractions of a year). The variable $children_i$ states how many children individual i had at the end of 2016. And the variable $temp\_winter_i$ is the average temperature during the months of January, February, and December 2016 for individual i. The data for this model are in the file `sick.txt` in the `PS5/data` folder, which contains comma-separated values of 200 individuals for four variables $(sick_i, age_i, children_i, temp\_winter_i)$ with variable labels in the first row.

(a) Estimate the parameters of the model $(\beta_0, \beta_1, \beta_2, \beta_3)$ by GMM by solving the minimization problem of the GMM criterion function. Use the identity matrix as the estimator for the optimal weighting matrix. Treat each of the 200 values of the variable $sick_i$ as your data moments $m(x_i)$ (200 data moments). Treat the predicted or expected sick values from your model as your model moments (200 model moments),

$$m(x_i|\beta_0, \beta_1, \beta_2, \beta_3) = \beta_0 + \beta_1 age_i + \beta_2 children_i + \beta_3 temp\_winter_i$$

where x_i is short hand for the data. Let the error function of the moments be the simple difference (not percent difference) of the data moments from

the model moments. These are equal to the error terms of the linear regression equation

$$\begin{aligned} e(x_i|\beta_0, \beta_1, \beta_2, \beta_3) &= sick_i - \beta_0 - \beta_1 age_i - \beta_2 children_i - \beta_3 temp\_winter_i \\ &= \varepsilon_i \end{aligned}$$

Use these error functions in your criterion function to estimate the model parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$ by GMM. This is a more general version of what OLS does. It minimizes the distance between the model moments and the data moments. It minimizes the sum of squared error terms. Report your estimates and report the value of your GMM criterion function. In this case, the GMM criterion function value evaluated at the optimal parameter values is simply the sum of squared errors.