Economics, Problem Set #4, Structural Estimation

OSM Lab, Dr. Evans

Due Wednesday, July 19 at 8:00am

1. **Health claim amounts and the GB family of distributions.** For this problem, you will use 10,619 health claims amounts from a fictitious sample of households. These data are in a single column of the text file `clms.txt` in the `WK4_StrEst` folder. Health claim amounts are reported in U.S. dollars. For this exercise, you will need to use the generalized beta family of distributions shown in the figure in Section 7 of your MLE Jupyter notebook.

   (a) Calculate and report the mean, median, maximum, minimum, and standard deviation of monthly health expenditures for these data. Plot two histograms of the data in which the y-axis gives the percent of observations in the particular bin of health expenditures and the x-axis gives the value of monthly health expenditures. Use percentage histograms in which the height of each bar is the percent of observations in that bin (see instructions in Jupyter notebook `PythonVisualize.ipynb` in Section 1.2). In the first histogram, use 1,000 bins to plot the frequency of all the data. In the second histogram, use 100 bins to plot the frequency of only monthly health expenditures less-than-or-equal-to \$800 ($x_i \leq 800$). Adjust the frequencies of this second histogram to account for the observations that you have not displayed ($x_i > 800$). That is, the heights of the histogram bars in the second histogram should not sum to 1 because you are only displaying a fraction of the data. Comparing the two histograms, why might you prefer the second one?

   (b) Using MLE, fit the gamma $GA(x; \alpha, \beta)$ distribution to the individual observation data. Use $\beta_0 = Var(x)/E(x)$ and $\alpha_0 = E(x)/\beta_0$ as your initial guess.[1] Report your estimated values for $\hat{\alpha}$ and $\hat{\beta}$, as well as the value of the maximized log likelihood function $\ln \mathcal{L}(\hat{\theta})$. Plot the second histogram from part (a) overlayed with a line representing the implied histogram from your estimated gamma (GA) distribution.

   (c) Using MLE, fit the generalized gamma $GG(x; \alpha, \beta, m)$ distribution to the individual observation data. Use your estimates for α and β from part(b), as well as $m = 1$, as your initial guess. Report your estimated values for $\hat{\alpha}$, $\hat{\beta}$, and $\hat{m}$, as well as the value of the maximized log likelihood function $\ln \mathcal{L}$. Plot the second histogram from part (a) overlayed with a line representing the implied histogram from your estimated generalized gamma (GG) distribution.

[1] These initial guesses come from the property of the gamma (GA) distribution that $E(x) = \alpha\beta$ and $Var(x) = \alpha\beta^2$.

(d) Using MLE, fit the generalized beta 2 $GB2(x; a, b, p, q)$ distribution to the individual observation data. Use your estimates for α, β, and m from part (c), as well as $q = 10,000$, as your initial guess. Report your estimated values for $\hat{a}$, $\hat{b}$, $\hat{p}$, and $\hat{q}$, as well as the value of the maximized log likelihood function $\ln \mathcal{L}$. Plot the second histogram from part(a) overlayed with a line representing the implied histogram from your estimated generalized beta 2 (GB2) distribution.

(e) Perform a likelihood ratio test for each of the estimated in parts (b) and (c), respectively, against the GB2 specification in part (d). This is feasible because each distribution is a nested version of the GB2. The degrees of freedom in the $\chi^2(p)$ is 4, consistent with the GB2. Report the $\chi^2(4)$ values from the likelihood ratio test for the estimated GA and the estimate GG.

(f) Using the estimated GB2 distribution from part (d), how likely am I to have a monthly health care claim of more than \$1,000? How does this amount change if I use the estimated GA distribution from part (b)?

2. **MLE estimation of simple macroeconomic model.** You can observe time series data in an economy for the following variables: $(c_t, k_t, w_t, r_t$. Data on (c_t, k_t, w_t, r_t) can be loaded from the file `MacroSeries.txt`. This file is a comma separated text file with no labels. The variabls are ordered as (c_t, k_t, w_t, r_t). These data have 100 periods, which are quarterly (25 years). Suppose you think that the data are generated by a process similar to the Brock and Mirman (1972). A simplified set of characterizing equations of the Brock and Mirman model are the following.

$$(c_t)^{-1} - \beta E\left[r_{t+1}(c_{t+1})^{-1}\right] = 0 \tag{1}$$

$$c_t + k_{t+1} - w_t - r_t k_t = 0 \tag{2}$$

$$w_t - (1-\alpha)e^{z_t}(k_t)^{\alpha} = 0 \tag{3}$$

$$r_t - \alpha e^{z_t}(k_t)^{\alpha-1} = 0 \tag{4}$$

$$z_t = \rho z_{t-1} + (1-\rho)\mu + \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim N(0, \sigma^2) \tag{5}$$

The variable c_t is aggregate consumption in period t, k_{t+1} is total household savings and investment in period t for which they receive a return in the next period (this model assumes full depreciation of capital). The wage per unit of labor in period t is w_t and the interest rate or rate of return on investment is r_t. Total factor productivity is z_t, which follows an AR(1) process given in (5). The rest of the symbols in the equations are parameters that must be estimated $(\alpha, \beta, \rho, \mu, \sigma)$. The constraints on these parameters are the following.

$$\alpha, \beta \in (0,1), \quad \mu, \sigma > 0, \quad \rho \in (-1,1)$$

Assume that the first observation in the data file variables is $t = 1$. Let k_1 be the first observation in the data file for the variable k_t. Assume that $z_0 = \mu$ so that $z_1 = \mu$. Assume that the discount factor is known to be $\beta = 0.99$.

(a) Use the data (w_t, k_t) and equations (3) and (5) to estimate the four parameters $(\alpha, \rho, \mu, \sigma)$ by maximum likelihood. Given a guess for the parameters $(\alpha, \rho, \mu, \sigma)$, you can use the two variables from the data (w_t, k_t) and (3) to back out a series for z_t. You can then use equation (5) to compute the probability of each $z_t \sim N\Big(\rho z_{t-1} + (1-\rho)\mu, \sigma^2\Big)$. The maximum likelihood estimate $(\hat{\alpha}, \hat{\rho}, \hat{\mu}, \hat{\sigma})$ maximizes the likelihood function of that normal distribution of z_t's. Report your estimates and the inverse hessian variance-covariance matrix of your estimates.

(b) Now we will estimate the parameters another way. Use the data (r_t, k_t) and equations (4) and (5) to estimate the four parameters $(\alpha, \rho, \mu, \sigma)$ by maximum likelihood. Given a guess for the parameters $(\alpha, \rho, \mu, \sigma)$, you can use the two variables from the data (r_t, k_t) and (4) to back out a series for z_t. You can then use equation (5) to compute the probability of each $z_t \sim N\Big(\rho z_{t-1} + (1-\rho)\mu, \sigma^2\Big)$. The maximum likelihood estimate $(\hat{\alpha}, \hat{\rho}, \hat{\mu}, \hat{\sigma})$ maximizes the likelihood function of that normal distribution of z_t's. Report your estimates and the inverse hessian variance-covariance matrix of your estimates.

(c) According to your estimates from part (a), if investment/savings in the current period is $k_t = 7,500,000$ and the productivity shock in the previous period was $z_{t-1} = 10$, what is the probability that the interest rate this period will be greater than $r_t = 1$. That is, solve for $Pr(r_t > 1 | \hat{\theta}, k_t, z_{t-1})$. [HINT: Use equation (4) to solve for the $z_t = z^*$ such that $r_t = 1$. Then use (5) to solve for the probability that $z_t > z^*$.]

3. **Matching the U.S. income distribution by GMM.** In this problem set, you will use the tab-delimited data file `usincmoms.txt`, which contains the 42 moments listed in Table 1 along with the midpoints of each bin. The first column in the data file gives the percent of the population in each income bin (the third column of Table 1). The second column in the data file has the midpoint of each income bin. So the midpoint of the first income bin of all household incomes less than \$5,000 is \$2,500.

(a) Plot the histogram implied by the moments in the tab-delimited text file `usincmoms.txt`.[2] The centers of each bin are in the second column of the data file `usincmoms.txt`. List the dollar amounts on the x-axis as thousands of dollars. That is, divide them by 1,000 to put them in units of thousands of dollars (\$000s). The bin cutoffs are given in Table 1. Even though the top bin is all incomes of \$250,000 and up, only graph the histogram up to the maximum income of \$350,000. (It doesn't look very

[2] As a reminder, a histogram is a bar chart, in which each of the bars represents the percent of observations in a particular x-axis bin (income, in this case). As such, the bars should be touching each other because each edge of each bar represents the cutoff level of each income category bin.

good graphing it between 0 and ∞.) In summary, your histogram should have 42 bars. The first 40 bars for the lowest income bins should be the same width. However, the last two bars should be different widths from each other and from the rest of the bars. Because the 41st bar is 10 times bigger (fatter) than the first 40 bars, divide its height by 10. Because the 42nd bar is 20 times bigger (fatter) than the first 40 bars, divide its height by 20. This is analogous to dividing the last two bars into 10 and 20 bars, respectively, and spreading frequency of each evenly among its divisions.

(b) Using GMM, fit the lognormal $LN(x; \mu, \sigma)$ distribution defined in the MLE notebook to the distribution of household income data using the moments from the data file. Make sure to try various initial guesses. (HINT: $\mu_0 = \ln(avg.inc.)$ might be good.) For your weighting matrix $\boldsymbol{W}$, use a 42×42 diagonal matrix in which the diagonal elements are the moments from the data file. This will put the most weight on the moments with the largest percent of the population. Report your estimated values for $\hat{\mu}$ and $\hat{\sigma}$, as well as the value of the minimized criterion function $\boldsymbol{e}(\boldsymbol{x}|\hat{\boldsymbol{\theta}})^T \boldsymbol{W} \boldsymbol{e}(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$. Plot the histogram from part (a) overlayed with a line representing the implied histogram from your estimated lognormal (LN) distribution. Each point on the line is the midpoint of the bin and the implied height of the bin. Do not forget to divide the values for your last two moments by 10 and 20, respectively, so that they match up with the histogram.

(c) Using GMM, fit the gamma $GA(x; \alpha, \beta)$ distribution defined in the MLE notebook to the distribution of household income data using the moments from the data file. Use $\alpha_0 = 3$ and $\beta_0 = 20,000$ as your initial guess.[3] Report your estimated values for $\hat{\alpha}$ and $\hat{\beta}$, as well as the value of the minimized criterion function $\boldsymbol{e}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})^T \boldsymbol{W} \boldsymbol{e}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$. Use the same weighting matrix as in part (b). Plot the histogram from part (a) overlayed with a line representing the implied histogram from your estimated gamma (GA) distribution. Do not forget to divide the values for your last two moments by 10 and 20, respectively, so that they match up with the histogram.

(d) Plot the histogram from part (a) overlayed with the line representing the implied histogram from your estimated lognormal (LN) distribution from part (b) and the line representing the implied histogram from your estimated gamma (GA) distribution from part (c). What is the most precise way to tell which distribution fits the data the best? Which estimated distribution—LN or GA—fits the data best?

(e) Repeat your estimation of the GA distribution from part (c), but use the two-step estimator for the optimal weighting matrix $\hat{\boldsymbol{W}}_{twostep}$. Do your estimates for α and β change much? How can you compare the goodness of fit of this estimated distribution versus the goodness of fit of the estimated distribution in part (c)?

[3] These initial guesses come from the property of the gamma (GA) distribution that $E(x) = \alpha\beta$ and $Var(x) = \alpha\beta^2$.

Table 1: Distribution of Household Money Income by Selected Income Class, 2011

Income class	# households (000s)	households %
All households	121,084	100.0
Less than $5,000	4,261	3.5
$5,000 to $9,999	4,972	4.1
$10,000 to $14,999	7,127	5.9
$15,000 to $19,999	6,882	5.7
$20,000 to $24,999	7,095	5.9
$25,000 to $29,999	6,591	5.4
$30,000 to $34,999	6,667	5.5
$35,000 to $39,999	6,136	5.1
$40,000 to $44,999	5,795	4.8
$45,000 to $49,999	4,945	4.1
$50,000 to $54,999	5,170	4.3
$55,000 to $59,999	4,250	3.5
$60,000 to $64,999	4,432	3.7
$65,000 to $69,999	3,836	3.2
$70,000 to $74,999	3,606	3.0
$75,000 to $79,999	3,452	2.9
$80,000 to $84,999	3,036	2.5
$85,000 to $89,999	2,566	2.1
$90,000 to $94,999	2,594	2.1
$95,000 to $99,999	2,251	1.9
$100,000 to $104,999	2,527	2.1
$105,000 to $109,999	1,771	1.5
$110,000 to $114,999	1,723	1.4
$115,000 to $119,999	1,569	1.3
$120,000 to $124,999	1,540	1.3
$125,000 to $129,999	1,258	1.0
$130,000 to $134,999	1,211	1.0
$135,000 to $139,999	918	0.8
$140,000 to $144,999	1,031	0.9
$145,000 to $149,999	893	0.7
$150,000 to $154,999	1,166	1.0
$155,000 to $159,999	740	0.6
$160,000 to $164,999	697	0.6
$165,000 to $169,999	610	0.5
$170,000 to $174,999	617	0.5
$175,000 to $179,999	530	0.4
$180,000 to $184,999	460	0.4
$185,000 to $189,999	363	0.3
$190,000 to $194,999	380	0.3
$195,000 to $199,999	312	0.3
$200,000 to $249,999	2,297	1.9
$250,000 and over	2,808	2.3
Mean income	$69,677	
Median income	$50,054	

Source: 2011 Current Population Survey household income count data Current Population Survey (2012, Table HINC-01)

4. **Estimating the Brock and Mirman (1972) model by GMM.** You can observe time series data in an economy for the following variables: (c_t, k_t, w_t, r_t). As in Problem 2, data on (c_t, k_t, w_t, r_t) can be loaded from the file `MacroSeries.txt`. This file is a comma separated text file with no labels. The variables are ordered as (c_t, k_t, w_t, r_t). These data have 100 periods, which are quarterly (25 years). Suppose you think that the data are generated by a process similar to the Brock and Mirman (1972). A simplified set of characterizing equations of the Brock and Mirman model are the following.

$$(c_t)^{-1} - \beta E\left[r_{t+1}(c_{t+1})^{-1}\right] = 0 \tag{1}$$
$$c_t + k_{t+1} - w_t - r_t k_t = 0 \tag{2}$$
$$w_t - (1-\alpha)e^{z_t}\left(k_t\right)^{\alpha} = 0 \tag{3}$$
$$r_t - \alpha e^{z_t}\left(k_t\right)^{\alpha-1} = 0 \tag{4}$$
$$z_t = \rho z_{t-1} + (1-\rho)\mu + \varepsilon_t \quad \text{where} \quad E\left[\varepsilon_t\right] = 0 \tag{5}$$

The variable c_t is aggregate consumption in period t, k_{t+1} is total household savings and investment in period t for which they receive a return in the next period (this model assumes full depreciation of capital). The wage per unit of labor in period t is w_t and the interest rate or rate of return on investment is r_t. Total factor productivity is z_t, which follows an AR(1) process given in (5). The rest of the symbols in the equations are parameters that must be estimated $(\alpha, \beta, \rho, \mu)$. The constraints on these parameters are the following.

$$\alpha, \beta \in (0,1), \quad \mu, \sigma > 0, \quad \rho \in (-1,1)$$

Assume that the first observation in the data file variables is $t = 1$. Let k_1 be the first observation in the data file for the variable k_t.

(a) Estimate α, β, ρ, and μ by GMM using the unconditional moment conditions that $E[\varepsilon_t] = 0$ and $E[\beta r_{t+1} c_t / c_{t+1} - 1] = 0$. Use the identity matrix $I(4)$ as your estimator of the optimal weighting matrix. Use the following four moment conditions to estimate the four parameters.

$$E\Big[z_{t+1} - \rho z_t - (1-\rho)\mu\Big] = 0 \tag{6}$$
$$E\Big[\Big(z_{t+1} - \rho z_t - (1-\rho)\mu\Big) z_t\Big] = 0 \tag{7}$$
$$E\left[\beta\alpha e^{z_{t+1}} k_{t+1}^{\alpha-1} \frac{c_t}{c_{t+1}} - 1\right] = 0 \tag{8}$$
$$E\left[\left(\beta\alpha e^{z_{t+1}} k_{t+1}^{\alpha-1} \frac{c_t}{c_{t+1}} - 1\right) w_t\right] = 0 \tag{9}$$

The estimation inside each iteration of the minimizer of the GMM objective function is the following.

- Given a guess for $(\alpha, \beta, \rho, \mu)$ and data (c_t, k_t, w_t, r_t), use (4) to back out an implied series for z_t.
- Given z_t, parameters $(\alpha, \beta, \rho, \mu)$ and data (c_t, k_t, w_t, r_t), calculate four empirical analogues of the moment conditions (6), (7), (8), and (9).
- Update guesses for parameters $(\alpha, \beta, \rho, \mu)$ until minimum criterion value is found.

Report your estimated parameter values $(\hat{\alpha}, \hat{\beta}, \hat{\rho}, \hat{\mu})$ and the value of your minimized criterion function.

5. **Estimating the Brock and Mirman (1972) model by SMM.** One nice property of the Brock and Mirman (1972) model is that the household decision has a known analytical solution in which the optimal savings decision k_{t+1} is a function of the productivity shock today z_t and the amount of capital today k_t.

$$k_{t+1} = \alpha\beta e^{z_t} k_t^{\alpha} \tag{10}$$

With this solution, it is straightforward to simulate the data of the Brock and Mirman (1972) model given parameters $(\alpha, \beta, \rho, \mu, \sigma)$. First, assume that $z_1 = \mu$ and that $k_1 = \text{mean}(k_t)$ from the data. These are initial values that will not change across simulations. Next, draw $T = 100$ normally distributed values of $\varepsilon_t \sim N(0, \sigma)$. Note that for SMM we have to return to fully specifying the distributional assumptions. Then, you can use equation (5) to calculate the simulated series for z_t. Now, you can use the policy function for savings (10) recursively to solve for the entire k_t series. With the entire k_t and z_t simulated series, you can use (3) to solve for the w_t series and (4) to solve for the r_t series. Lastly, you use the budget constraint (2) to solve for the c_t series.

(a) Estimate the five parameters of the Brock and Mirman (1972) model $(\alpha, \beta, \rho, \mu, \sigma)$ described by equations (1) through (5) by SMM. Choose the five parameters to match the following six moments from the 100 periods of empirical data $\{c_t, k_t, w_t, r_t\}_{t=1}^{100}$ in `MacroSeries.txt`: mean(c_t), mean(k_t), var(c_t), var(k_t), corr(c_t, k_t), and corr(k_t, k_{t+1}). In your simulations of the model, set $T = 100$ and $S = 1,000$. Start each of your simulations from $k_1 = \text{mean}(k_t)$ from the `MacroSeries.txt` file and $z_1 = \mu$. Input the bounds to be $\alpha, \beta \in [0.01, 0.99]$, $\rho \in [-0.99, 0.99]$, $\mu \in [-0.5, 1]$, and $\sigma \in [0.001, 1]$. Also, use the identity matrix as your weighting matrix $\boldsymbol{W}$. Report your solution $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}, \hat{\rho}, \hat{\mu}, \hat{\sigma})$, the vector of moment differences at the optimum, and the criterion function value.

References

Brock, William A. and Leonard J. Mirman, "Optimal economic growth and uncertainty: The discounted case," *Journal of Economic Theory*, June 1972, *4* (3), 479–513.

Current Population Survey, "2012 Annual Social and Economic (ASEC) Supplement," Technical Report, Bureau of the Census and Bureau of Labor Statistics 2012.