Problem Set #2

MACS 40200, Dr. Evans
Due Monday, Jan. 27 at 1:30pm

1. **Health claim amounts and the GB family of distributions (5 points).** 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 PS2 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) (0.5 points) 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) (1 point) 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) (1 point) 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) (1 point) 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) (1 point) 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) (0.5 points) 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 (5 points).** 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 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.

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

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

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

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

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

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) (2 points) 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) (2 points) 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) (1 point) 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 $z_t > z^*$.]

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.