Problem Set #4

MACS 40200, Dr. Evans
Due Tuesday, Jan. 31 at 12:00pm

1. **Matching the U.S. income distribution by GMM (10 points).** 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) (2 points) Plot the histogram implied by the moments in the tab-delimited text file `usincmoms.txt`.[1] 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 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) (3 points) 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.

[1] 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.

(c) (3 points) 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.[2] 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) (1 point)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) (1 point) 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)?

References

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

[2] 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)