Problem Set 2

Prof. Conlon

Due date: March 6

Your answers should be produced in LaTeX, and should include all relevant graph and code. Code should be in the appropriate verbatim environment and properly documented. You are allowed to work in groups but you must turn in your own writeup. Submit your assignment via dropbox link to cconlon@stern.nyu.edu.

Description of Nielsen Data

This problem set uses simulated data that is made to look like responses of households to the Nielsen Consumer Panel Survey. An important feature of this dataset (and many other datasets) is that consumers report income in bands (between \$40,000 and \$50,000) rather than exact values. This can make estimating the overall distribution of income complicated. To make matters worse, there are also two regimes of coding income into bins. From 2006-2009 there are three extra bins for high earners. From 2010-onwards these bins are eliminated. Our goal is to construct a distribution of income for the provided panelists.

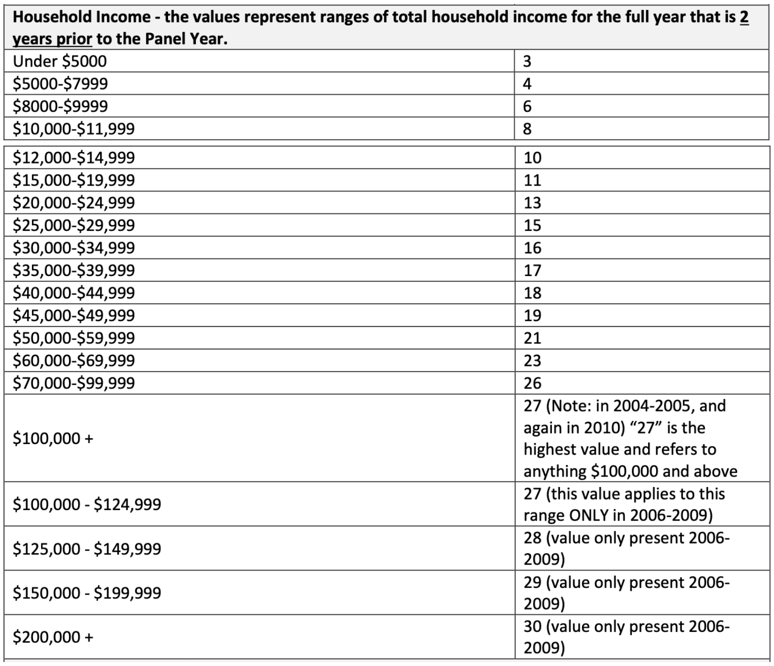
Each panelist is described by:

household_code a unique panelist identifier

hh_income an income "bin" coded as below

year_regime either 'early' or 'late' which denotes which bin assignments to use.

projection_weight an (optional) weight for each household to make the overall sample more demographically representative. [Note: these weights do not sum to one, but you should ensure that they do].



Household Income - the values represent ranges of total household income for the full year that is 2 years prior to the Panel Year.	
Under \$5000	3
\$5000-\$7999	4
\$8000-\$9999	6
\$10,000-\$11,999	8
\$12,000-\$14,999	10
\$15,000-\$19,999	11
\$20,000-\$24,999	13
\$25,000-\$29,999	15
\$30,000-\$34,999	16
\$35,000-\$39,999	17
\$40,000-\$44,999	18
\$45,000-\$49,999	19
\$50,000-\$59,999	21
\$60,000-\$69,999	23
\$70,000-\$99,999	26
\$100,000 +	27 (Note: in 2004-2005, and again in 2010) "27" is the highest value and refers to anything \$100,000 and above
\$100,000 - \$124,999	27 (this value applies to this range ONLY in 2006-2009)
\$125,000 - \$149,999	28 (value only present 2006-2009)
\$150,000 - \$199,999	29 (value only present 2006-2009)
\$200,000 +	30 (value only present 2006-2009)

You will want to produce two versions of all results, one using the **weighted sample** and one using the **unweighted sample**.

For each method, we will want to estimate the lognormal parameters under three assumptions:

(a) Assuming each household earns the lowest income in each bin.

(b) Assuming each household earns the highest income in each bin.

(c) Assuming each household earns the mean income in each bin.

Because income is top coded (b) and (c) are tricky. The goal here is to make reasonable assumptions, not ridiculous ones.

Part 1: Method of Moments

Let's begin by estimating a log-normal distribution of income. If income is distributed as $y_i \sim LN(\mu, \sigma)$ then $\ln y_i \sim N(\mu, \sigma)$. We will start with a simple *method of moments* estimator. Here the idea is to compute moments from our sample (means or variances y_i) and/or $\ln y_i$ and to use those to solve for parameters. You may want to look up the moments of the lognormal distribution on Wikipedia or in a textbook.

1. Estimate (μ, σ) working with y_i.
2. Estimate (μ, σ) working with $\ln y_i$.
3. Discuss how you would compute an estimate of the standard errors of (μ, σ). For extra credit: compute them.

Part 2: Maximum Likelihood

Now let's consider a *maximum likelihood estimator*. Again assume that the distribution of income is lognormal.

4. Write the likelihood for the entire sample: $Pr(y_1, \ldots, y_n|\mu, \sigma)$ or $Pr(\ln y_1, \ldots, \ln y_n|\mu, \sigma)$
   (Note: be careful that the coding of bins changes after 2010. Also be-careful to allow for weighted observations w_i.)
5. Write the log-likelihood
6. Write the score of the log-likelihood for each i and the overall gradient.
7. Estimate (μ, σ) for each assumption about coding income using the dataset.
8. Estimate the standard errors of $(\widehat{\mu}_{MLE}, \widehat{\sigma}_{MLE})$ using the Fisher Information/Hessian Matrix.
9. Discuss your results so far. How sensitive are they to assumptions about the income bins? How sensitive are they to using the weighted vs. unweighted sample?

Part 3: Generalized Method of Moments

Now let's write a GMM estimator that correctly handles the fact that we observe income in bins rather than directly observing income. Again we can assume that income is log-normally distributed and define the contribution of each observation as. The idea is to match the probability that an individual falls into a particular income bin in the data, to the probability that an individual falls in a particular income bin under the model:

$$g(y_i|\mu, \sigma) = Pr(40,000 \leq y_i \leq 44,999|\mu, \sigma) - \frac{1}{N}\sum_{i=1}^{N} I(40,000 \leq y_i \leq 44,999|\mu, \sigma)$$

- You will have at least one moment per bin.
- You will also need to think about how you want to handle the fact that the income coding regime is not the same for the entire sample.

10. Write the Jacobian of the moment conditions with respect to the parameters $\frac{\partial g(y_i|\mu,\sigma)}{\partial \theta}$.
11. Estimate (μ, σ) using GMM. You will want to report the first stage estimates and the 2nd stage estimates.
12. Put all of your MoM, MLE, and GMM estimates into a table. Also compute the mean income in dollars and the 10th percentile and 90th percentile of the corresponding income distribution. How do they compare? How do standard errors compare?

Part 4: Nonparametric Estimates

13. Plot a histogram of the income data. (It may be difficult to arrange the data into bins).
14. Plot your lognormal densities together on a single plot (Upper bound MLE, Lower Bound MLE, Mean MLE, and 2-step GMM estimates).
15. Estimate a kernel density plot for income (using the upper bound, the lower bound, and the mean of the bin) and compare that to your estimated lognormal densities.
16. Explain how the plots are similar and how they are different.