EXERCISE ON WEIGHT CORRECTIONS FOR SELECTION BIAS

A data file containing individual log income and sampling weights is available at `http://sims.princeton.edu/yftp/ApplEmet17/psid07.csv`. This is a comma-separated-values text files, which should be readable by most spreadsheet or statistical programs. There is also a .RData version which loads a data frame called `ywd` containing the same data. (R can read .csv files with **`read.csv`**`()`).

I have done (1)-(3) below myself, so I know that much works out sensibly. The rest might contain mistakes or be impossible, so complain if it starts to seem to you that this is the case.

(1) Modeling the log income data alone as distributed with a pdf whose kernel is

$$\frac{1}{e^{-a(y-\mu)} + e^{b(y-\mu)}}$$

Find the maximum likelihood estimator of the (a, b, μ) parameter vector. Since you only have an explicit form for the kernel, you will have to evaluate the integral of the kernel for each possible parameter vector when you evaluate the likelihood for that parameter vector. This is probably best done by a simple numerical integration, discretizing the argument and summing. To keep results comparable across the class, everyone should use the range [3, 16] as the space of possible values for log income when doing the numerical integration.

[I said in class that a similar kernel gave nonsense results. That resulted from using α and $1-\alpha$ in place of $e^{a\mu}$ and $e^{-b\mu}$ in the kernel formula. While both parameterizations describe the same family of distributions, the one using alpha misbehaves numerically when the center of the distribution is far from zero.]

(2) Plot your MLE estimate of the density function against a histogram or a kernel-smoothed estimate of the density (**`hist`**`()` or **`density`**`()` in R). Be sure your MLE estimate and the histogram or kernel-smoothed estimate have the same integral before you plot them.

(3) Form a kernel-smoothed estimate of the density using the weights, and plot that along with both the unweighted density estimate and your MLE density. In R a kernel-smoothed estimate with weights can be formed automatically by using the **`weights`** argument to **`density`**`()`. To do kernel-smoothing with weights, you estimate the density at x as proportional to the sum of $k(w_i(x - x_i))$ over the sample x_i values, with w_i the weight on observation i. The simplest form of this makes k the indicator function for a small interval.

(4) Make MCMC draws from the posterior on (a, b, μ). To be sure the posterior is integrable, you will need a proper prior. A flat one on $(0, 10) \times (0, 10) \times (0, 20)$ should work. Assess convergence for your draws. Use them to form a 90% hpd interval for

Date: February 15, 2017.

the proportion of income (not log income) earned by the top 1% of incomes. Compare your interval to the point estimate of the same thing emerging from the kernel and/or histogram estimates of the density with and without weights.

(5) Plot a scatter of log income against weight and, on the same plot, the regression line from a regression of income on a second order polynomial in weight. (Since the scatter will have a lot of points, you may want to use small plotting characters, e.g. `pch="."` in R.)

(6) Fit a parametric model to the joint distribution of weights and log income. The weight distribution is more lumpy than that of income. One possibility would be a mixture of three gamma's for the marginal of the weights, and the a, b, μ distribution you have already used as the conditional distribution of income given weight, but with μ, and possibly also a and b, as linear or quadratic functions of weight. This gets to be a big model, so you might start with a single gamma distribution for the weights and μ alone dependent on weights and work up if possible to something that fits better.

(7) Construct the 90% hpd interval for the proportion of income earned by the top 1% from your joint distribution estimate. Note that your joint distribution is the joint distribution of the sampled log income, with selection bias, and what we want here is a 90% band for the "true" distribution of income corrected for selection bias. If $p(y, w)$ is the joint pdf of a draw of log income and weight, including selection bias, what you are after is

$$\frac{\int_{y>\bar{y}} yw\, p(y,w)\, dy\, dw}{\int yw\, p(y,w)\, dy\, dw} \quad \text{subject to}$$

$$\frac{\int_{y>\bar{y}} w\, p(w,y)\, dy\, dw}{\int w\, p(w,y)\, dy\, dw} = .01\,.$$

(8) It could be that even for this share-of-the-one-percent function of the parameters there is not much difference between weighted and unweighted calculations and not much uncertainty. If so, you could see what happens with a random sample of 400 from the original data set. But this is not required.