ECON 523: Program Evaluation for International Development
Empirical Exercise 8

Using the data set `ECON523-E8-meyersson-data.dta`, replicate Table 1, Table 2 (only the first half, on page 247, and only columns 1 through 8 of the results), and Figure 2a from "Islamic Rule and the Empowerment of the Poor and Pious" by Erik Meyersson. Upload pdfs of your finished tables and figure to gradescope, together with the file that generates them.

Extra credit: replicate Figure 2b and upload it to gradescope.

Hints and Suggestions:

1. When replicating Table 1, it will save time to write a program that produces a single row when you supply it with a variable name.

2. The code

   ```
   local l'1' : variable label '1'
   quietly putexcel A'2'="'l'1''", left
   ```

   will store a variable's label as a local macro and write it to an excel file.

3. Use the `twoway` command to make the histogram. To layer multiple plots with the same x-axis variable over each other, you can use a command like:

   ```
   twoway ///
   (plottype y2 x, options) ///
   (plottype y2 x, options), ///
   legen(off) title("I Love Making Graphs")
   ```

   where `plottype` might be `scatter` or `line` or `lpoly` (for a local linear regression like the ones in Figure 5 in the paper) or `histogram` (without a y variable) or `rarea` (with two y variables that you want to shade the area between). `twoway` allows you to easily adjust the axis lengths (using the `xlabel` and `ylabel` options) and the axis titles (using the `xtitle` and `ytitle` options). Play around with the bin width to replicate Dr. Meyersson's table to the extent possible.

4. Placing the `set scheme s1mono` command near the top of your do file will eliminate the blue background from your graphs.

5. You can install the `blindschemes` package to access the Okabe-Ito colorblind friendly palette, which is ideal for making graphs that are easy for people to read. Data visualization books recommend using blue and orange to make your graphs accessible. I like the `sea` and `vermillion` options; the most popular machine learning textbook uses `sky` and `orangebrown`.

6. You can make marker or confidence bands partially opaque by adding a percent sign followed by a number between 1 and 99. For example, I graph confidence intervals using the code (for a `twoway` layer):

```
(rarea upper lower x, sort lwidth(none) fcolor(sea%24))
```

7. You do not need to format the titles of Table 2 exactly the way Dr. Meyersson has. Since all of your columns have the same outcome variable and the same age range, you can omit those labels. You do not need to use $\hat{h}$ to indicate the Imbens-Kalyanaraman optimal bandwidth; you can just refer to it as "IK Optimal Bandwidth" or something similar.

8. The paper reports the Imbens-Kalyanaraman optimal bandwidth in approximate terms, but you can use the `rd` or `rdcv` commands to get the exact bandwidth. Alternatively, you can play around with the bandwidth in each column so that your sample sizes match the ones reported in the paper.

9. When you use the quadratic control function in an RD, you need to include both the running variable (above and below the discontinuity) and the square of those two terms. When you used the cubic control function, you need to include those terms plus cubic functions of the running variable.

10. Lilly and I are not going to help you with the extra credit, and it is a tough one. You will need to install and use Stata's `kdens` command to calculate and store estimates of the kernel density and the upper and lower bounds of the confidence interval for the kernel density. (You can rear about kernel density plots at `https://clauswilke.com/dataviz/histograms-density-plots.html` or `https://datavizcatalogue.com/methods/density_plot.html`.) You'll need to figure out which of `kdens`'s options allows you to have a sharp upper or lower boundary on the range of points for which you are estimating the kernel density. When you estimate the kernel densities separate above and below the discontinuity, they will have different scales, because the area under a kernel density is one. You will need to figure out how to rescale them so that they line up (try plotting them on top of the density estimates for the pooled sample to help you think through this). You will also need to calculate the number of observations in each of 100 bins to overlay the scatter plot. These counts will also need to be rescaled so that they reflect the same units as the density plots. Again, this is all quite challenging. Good luck to those who attempt it!