**Directions:** We will reproduce a few tables and figures from Hansen (2015), which was a study of the effect of a DWI on recidivism using an RDD design. You will need to download the dataset from github repo using the following address:

<https://github.com/scunning1975/causal-inference-class/raw/master/hansen_dwi>

The outcome variable is “recidivism” or “recid” which is measuring whether the person showed back up in the data within 4 months. Use this data to answer the following questions.

**Questions**

1. In the United States, an officer can arrest a driver if after giving them a blood alcohol content (BAC) test they learn the driver had a BAC of 0.08 or higher. We will only focus on the 0.08 BAC cutoff. We will be ignoring the 0.15 cutoff for all this analysis.
   1. Create a dummy equaling 1 if **bac1**>= 0.08 and 0 otherwise in your do file or R file.
   2. The first thing to do in any RDD is look at the raw data and see if there’s any evidence for manipulation (“sorting on the running variable”). Evaluate whether you see signs of manipulation. Recreate Figure 1 using the bac1 variable as your measure of blood alcohol content or use your own density test from software. Do you find evidence for **sorting on the running variable**? Explain your results.
2. The second thing we need to do is check for **covariate balance**.
   1. Recreate Table 2 Panel A but only white, male, age and accident (acc) as dependent variables. Use your equation 1) for this.
   2. Are the covariates balanced at the cutoff? It’s okay if they are not exactly the same as Hansen’s as this is not the exact same dataset he used
3. **Recreate Figure 2 panel A-D**. You can use the -cmogram- command in Stata to do this (or another method). Fit both linear and quadratic with confidence intervals. Discuss what you find and compare it with Hansen’s paper.
4. Estimate equation (1) with recidivism (recid) as the outcome. Be careful to read exactly what you can about this equation in the article and notes. This corresponds to Table 3 column 1, but since I am missing some of his variables, your sample size will be the entire dataset of 214,558. Nevertheless, replicate Table 3, column 1, Panels A and B. Note that these are local linear regressions and Panel A uses as its bandwidth 0.03 to 0.13. But Panel B has a narrower bandwidth of 0.055 to 0.105. Your table should have three columns and two A and B panels associated with the different bandwidths:
   1. Column 1: control for the bac1 linearly
   2. Column 2: interact bac1 with cutoff linearly
   3. Column 3: interact bac1 with cutoff linearly and as a quadratic
   4. For all analysis, estimate uncertainty using heteroskedastic robust standard errors. [ed: But if you want to show off, use Kolesár and Rothe’s 2018 “honest” confidence intervals (only available in R).]
5. Recreate the top panel of Figure 3 according to the following rule:
   1. Fit linear fit using only observations with less than 0.15 bac on the bac1
   2. Fit quadratic fit using only observations with less than 0.15 bac on the bac1
6. Consider repeating this analysis using the donut hole approach.
7. Discuss what you learned from this exercise. What was the hypothesis you tested and what did you find? How confident are you in Hansen’s original conclusion? Why/why not?