Readme File

Summary Statistics

1. Panel A
    a. To start, I selected variables related to demographics and age from the NLSY97 public dataset and imported it into Stata for panel A of the summary statistics.

    b. I gave the variables labels and renamed each one so that they are easier to locate.

    c. Unfortunately, Deza does not identify the exact variables that she selected from the NLSY97. As a result, I had to decide which ones I should use.

    d. I generated the age variable that Deza created by selected the variable CV_AGE(MONTHS) INT DATE. I do this by taking the age of participants=age_month/12 first for the year 1997. Later on, I create the age_decimal variable using the same approach but for the years 1997-2009. 1997 is when the survey first started and included the full sample of 8,984 participants. The study took place between the years 1997-2009.

    e. I start by working on the first column for panel A of the summary statistics before I reshape the data. All of the full sample column estimates are exact, except those relating to the proportion of participants who used a particular substance. Sadly, the estimates are off because many years for this specific topic are missing in the public dataset, for example, for the variable representing if a respondent ever had alcohol or not, only the years 1997, 1998, 1999, 2000, 2004, and 2005 are available.

    f. I then reshaped the data and dropped all participants lost to attrition to create the rest of the other columns for panel A of the summary statistics. All of the resulting estimates are either exact or very close. To reiterate, the only resulting estimates that were off were the one relating to the respondents who reported ever drinking alcohol.

2. Panel B
    a. For panel B, all of the estimates were very close. I reshaped the data, and the first column only included the full sample of participants. In the second column, I dropped the participants who lost to attrition. In columns 3-5, I focused on those not lost to attrition and did not avoid questions about a particular substance.

3. Panel C
    a. Unfortunately, for panel C, many years are missing for the variable involving whether or not participants ever had alcohol and I had to make a tough decision. The big problem is that the study's full sample started in 1997 and consisted of 8,984 participants.

b. The NLSY97 provides the starting year for those whoever had alcohol and whoever had marijuana, but not for cocaine. Many participants answered the most questions regarding the substance bundles they ever used in the study's earliest years. I am aware that Deza computed the probabilities for each year independently and then took the average, but this is not possible in my case.

c. Unfortunately, the earliest year for those who ever used cocaine was 1998. As a result, some of the participants are missing because they had dropped out. Since Deza provides the share of respondents who consumed a particular bundle, I took the number of participants who used a specific bundle and divided them by 8,984. This will be slightly off since the year 1997 was missing for those who ever used cocaine, but I treated it as a full sample since it's the earliest starting year.

d. Also, as I mentioned before, many years are missing. In the years that are provided, a significant number of participants either dropped out of the study or refused to answer the related interview questions, which largely biases the resulting estimates I received as indicated by a value of -5 or -1. Therefore, to create substance consumption bundles, I just matched up the earliest starting years for participants who used the selected substances mentioned above.

e. For each bundle that I created, a value of 1 indicates that a participant used a particular substance, and 0 means that it was not.

4. Panel D
   a. For panel D, Deza didn't specify whether she used the full sample or those not lost to attrition. Therefore, I decided to use the participants who were not lost to attrition. Unfortunately, once again, there were many years for those who ever used a particular substance, which caused the estimates to be off.

   b. I calculated unconditional probability as P(A and B) = P(A) * P(B) and the conditional probability as P(B/A) = P(A and B)/P(A).

Figure 1

- I generated the variable age_decimal that Deza used, which represents age_months/12 rounded to two decimal places

- The cmogram command was used to create the regression discontinuity graphs

- All of the diagrams were combined using the graph combine command

- I used the estimates that I calculated from the summary statistics, including participants not lost to attrition over the years. All of the graphs are very close. The only noticeable differences are that the y-axis of panel A averages the ID numbers given to participants. For graph F, I am not sure which variable was selected from the NLSY97 to calculate share of college attendants. I decided to choose was KEY!ENROLLED, which indicates if a respondent was enrolled in college or not. This is a potential reason why the graph had a slightly different y-axis.

Figure 2

- For the second figure that contains five graphs, unfortunately, I was unable to construct the one for panel A since the public NLSY97 dataset did not provide a variable indicating how much alcohol a participant used last year. For the other graphs, I created RDD's using the cmogram command of the participants that were not lost to attrition over the years. I used the indicators, alcohol use last month and binge drinking from the previous month from the NLSY97 dataset. As for the RDD graphs that display the percent of days, I divided the number of days by 30 days as Deza had done.

Figure 3

- Deza did not identify the exact variables she used to construct panel A and panel B of figure 3. For panel A of figure 3, I selected the number of times a participant had used hard drugs such as cocaine since the last interview.

- Unfortunately, panel B of figure 3 is slightly off because the public NLSY97 dataset did not provide a variable for the days' participants used hard drugs. I calculated the number of participants who were not lost to attrition between the years 1997-2009. Therefore, instead of not constructing a figure at all, I decided to improvise. I decided to choose if a respondent had used cocaine or hard drugs such as heroin since the last interview. As a result, the graph displays a slight downward trend rather than a slight upward trend. Unfortunately, some of the years are missing as well, which impacts the diagrams' accuracy.

Figure 4

- Creating the RDD graphs for figure 4 was a challenge since the variables relating to the starting age for using a particular substance were missing a lot of years, similar to the variables about participants who were asked if they had ever used a specific substance at all.

- To improvise, I used participants whether they were lost to attrition or not. I decided not to drop participants because I would be losing their responses that they had made in an earlier year if I removed them once they dropped out in a later year. As a result of this setback, the graphs are slightly off.