Hw#4 Article Questions

Marialexia Alfano

October 25, 2017

**1. In the in-course exercises, we have been analyzing data with accident as the observation unit. This study uses a different observation unit. What is the unit of observation in the Brady and Li study?**

Drivers who died within 1 hr of crash.

**When you download the FARS data for a year, you get a zipped folder with several different datasets. Which of the FARS datasets provides information at this observation level (and so will be the one you want to use for this analysis)?**

The “person” dataset.

**2. This study only analyzes a subset of the available FARS data. Enumerate all of the constraints that are used by the study to create the subset of data they use in their study (e.g., years, states, person type, injury type).**

Time (years) – January 1, 1999 to December 31, 2010, States (narrowed down to 6 that routinely test drivers for drugs) – California, Hawaii, Illinois, New Hampshire, Rhode Island, West Virginia Person type – driver of the motor vehicle, Injury type – death, Drug (including alcohol) testing data must be available, Time of death data must be available

**Go through the FARS documentation and provide the variable names for the variables you will use to create filter statements in R to limit the data to this subset. Provide the values that you will want to keep from each variable.**

Variable names for which filter statements are used (in all caps): “PER\_TYP”: keep only “1” which is “driver” “INJ\_SEV”: keep only “4” which is “fatal”  
“STATE”: keep only 6, 15, 17, 33, 44, 54 which are the states California, Hawaii, Illinois, New Hampshire, Rhode Island, and West Virginia “LAG\_HRS”: keep only the values that are “less than (<) 1 hr” or equal to 1 hour if “LAG\_MIN” is also equal to 0 minutes

**3. The study gives results stratified by age category, year, year category (e.g., 1999–2002), alcohol level, non-alcohol drug category, and sex. For each of these stratifications, give the variable name for the FARS variable you could use to make the split (i.e., the column name you would use in a group\_by statement to create summaries within each category or the column name you would mutate to generate a categorical variable).**

Age category – “AGE” variable

Year and year category – the “DEATH\_YR” variable can be used stratify by year of when the crashes occurred

Alcohol level – “ALC\_RES” variable for result of test for alcohol

Non-Alcohol drug category – “DRUG\_RES” variable for result of test for different drugs

Sex – “SEX” variable which shows if male or female (or unknown/not reported)

**Describe how each of these variables are coded in the data. Are there any values for missing data that you’ll need to mutate to NA values in R? Are there any cases where coding has changed over the study period?**

The variables are coded numerically and in a system, which usually incorporates a range of numbers to record several types of results that are possible in a scenario. Some of these numerical values might represent the data exactly as is (i.e. age in years is represented by the number given) but also according to the FARS coding system, if you were 100 years old at the time of the crash, the value would be recorded as 97, so the numerical values standing for their actual value may not hold up throughout the system even within just one variable’s coding system. For NA or missing or not reported results, the coding seems to be a series of repeated values (i.e. 999 or 99 etc.) relatively consistently for a lot of the variables in the dataset.

Age category – “AGE” variable coded (2009 onward): 0 = less than one year old, 1-120 = age of the individual in years, 998 = not reported, 999 = unknown

Year and year category – the “DEATH\_YR” variable can be used stratify by year of when the crashes occurred coded: 0 or 8888 Not Applicable (Non-Fatal), xxxx Year of the Death (4 number format), 9999 Unknown

Alcohol level – “ALC\_RES” variable for result of test for alcohol 0-93 = actual value of BAC Test, 94 = 0.94 or greater, 95 = test refused, 96 = none given, 97 = test performed, results unknown, 98 = Positive reading but no actual value, 99 = unknown if tested

Non-Alcohol drug category – “DRUG\_RES” variable for result of test for different drugs 0 Not Tested for Drugs, 1 No Drugs Reported/Negative, 95 Not Reported, 100-295 Narcotic*, 300-395 Depressant*, 400-495 Stimulant*, 600-695 Cannabinoid*, 996 Other Drugs, 997 Tested for Drugs, Results Unknown. 998 Tested for Drugs, Drugs Found, Type Unknown/Positive, 999 Unknown if Tested/Not Reported, 999 Unknown if Tested

Sex – “SEX” variable coded: 1 = Male, 2 = Female, 8 = not reported, 9 = unknown

There are certainly values for missing data that were mutated to NA values in R. For example, in the AGE variable, the not reported or unknown data points were coded as 998 and 999, respectively. These values were then mutated to represent “NA”. Within the AGE variable, the coding did change over the study period. Any year that was 2008 or earlier had just “99” as the code for “unknown” values. Then from 2009 onward, the “unknown” values were coded as “999”. Also from 2009 onward, the range for the age of individual in years was also expanded from “1-96” to “1-120”, eliminating the need for “97” to represent all people 97 and older.