hw4questions

Kate Huebner

10/20/2017

To access the documentation for the FARS dataset, go to this [link](https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812447) To access the Li and Brown paper, go to this [link](http://aje.oxfordjournals.org/content/early/2014/01/27/aje.kwt327.full.pdf+html)

# Question 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? 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)?

There are two levels of observation in this dataset (person and person-drug combination), which results in lots of repeated information. Their objective was to evaluate the time trends in alcohol and nonalcohol drugs detected in drivers who died within 1 hour of a motor vehicle crash in 6 states in the United States that routinely test such drivers for drugs. Thus, the 'person\_year data file' is used for thisanalysis.

# Question 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). 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.

The variables used in the analysis include st\_case, veh\_no, per\_no, state, per\_typ, lag\_hrs, lag\_mins, inj\_sev, age, alc\_res, contains("drugres"), and sex.

* st-case: unique identifier for the state case (consecutive)
* veh\_no: unique identifier for the vehicle identification
* per\_no: unique identifer for the drver. Will limit the values equal to 1, which is "driver"
* state: state number corresponding to the state code. Will limit the values to California (code = 6), Hawaii (code = 15), Illinois (code = 17), New Hampshire (code = 33), Rhode Island (code = 44), and West Virginia (code = 54))
* per\_typ: describes the role of the person involved in the crash
* lag\_hrs:records the hours between the time of crash and time of death
* lag\_mis: the minutes in addition to hours betwen time of crash adn time of death
* inj\_sev: severity of the injury to the person in the crash, using KABCO scale. Will limit the values equal to 4, which is a "fatal injury""
* age: the person’s age at the time of the crash, in years, with respect to their last birthday
* alc\_res: alcohol test result
* contains("drugres"): identifies the drug test result for this person. this incorportates drugres1,2 or 3 categories
* sex: identifies the sex of this person involved in the crash. will limit the values equal to 1,2, and 9 (male, female and unknown)

# Question 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). 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?

\*Age: The age variable will be broken up into 4 stratifications and renamed age\_cat. This variable is coded numerically, where 0 is < 1 year, 1-120 is the age in years, 998 is unreported, and 999 is unknown. Since missing values are coded as 998 and 999, they wil need to be saved as NA in the code.

\*Year:

\*Year Category:

\*Alcohol Level: The alc\_res variable will need to be coded such that Alc\_res > 94 is converted to NA because it represented unknown results, no test given, or a different test was performed. In 2010 and later the value of 95 was added to code an unreported test result. The remaining integer values can be divided by 10 to signify the blood alcohol content. To turn this into a categorical variable, this continuous variable would need to be converted into bins before grouping by it.

\*Sex: