FARS Writing Assignment

Preston Phillips

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?

**In our previous exercises the data has focused on crash characteristics and environmental conditions at the time of the crash. This study focuses on the characteristics of the humans involved in the crash. The FARS data provides information on age, sex, injury severity, alcohol test results and other drug results. This allows the study to assess the trends in alcohol and other drugs detected in drivers.**

* + 1. 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)?
    2. **The file needed for this assignment is the PERSON.dbf file.**

1. 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.

* **Drivers (PER\_TYP==1)**
* **Only fatal deaths (INJ\_SEV==4)**
* **Died in less than an hour (60\*LAG\_HRS+LAG\_MINS <61)**
* **Only in CA, HI, NH, RI, WV (ST=c(6,15,17,33,44,54))**

1. 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 category (AGE)**
* **year (YEAR)**
* **year category (YEAR)**
* **alcohol level (ALCRES)**
* **non-alcohol drug category (DRUGRES1, DRUGRES2,DRUGRES3)**
* **sex (SEX)**

**For AGE, for FARS PERSON files for 1975-2008, 97 means the person was 97 or over and 99 means unknown. For 2009 and 2010, the age goes up to 120 and the unknown is 999.**

**For SEX, for FARS PERSON files from 1975-2009 there is no option for the sex not reported. In 2010 there is value defined for when the sex was not reported.**