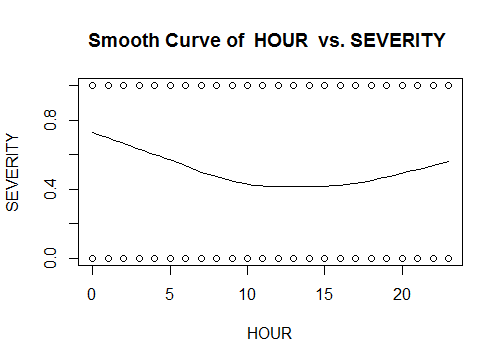
David Lowe

EDA #6

1. Tens of thousands of people in the U.S. die from motor vehicle crashes each year, despite the legislature and safeguards that are enforced by the state. Car crashes often result in serious injury and even death. Each car crash is comprised of a specific scenario that can be characterized by many variables. Factors such as seat belt use, weather conditions, blood alcohol content, and road type can all play a role into how severe the crash is. To better prevent serious or fatal accidents, the National Highway Traffic Safety Administration (NHTSA) would like to understand which of several different factors are most helpful in predicting a serious crash. From such an analysis, they can enact policy that creates a safer driving experience for the nation. Thus, the primary goal of this analysis is to understand the relationships between the factors that go into a vehicle-related crash, and the severity of the crash.
2. The data comes from a stratified sample of 8,603 crashes across the U.S. These crashes have been rated “severe” if at least one person was seriously injured or killed from the crash. Other variables included are air-bags deployed, weather conditions, hour of the day, light conditions, type of intersection, type of restrictions used, type of roadway, number of lanes, curve of the road, speed limit, and the road surface conditions. Each of these variables has a factor type response with two or more levels. Many factors are numeric, with each number corresponding to a specific condition.

Severity for each level of the included variables can be measured by the mean severity over the stratified sample. Some initial trends that are apparent is that severity goes up when someone involved in the crash had drunk alcohol. Severity by each hour of the day shows that the highest average severity of crashes occurs in the early morning (between 12:00 a.m. and 4:00 a.m.). This can be seen with the smoothed curve in. Upon further inspection of other variables, it seems that there may be interactions between different factors, such as with alcohol and the time of day.

1. Each classification of each variable may influence crash severity, so it is important to test each. Some type of analysis that allows for inference on what effects the severity of a crash is required to analyze this data. Since the data is classification type data, meaning that each variable is made up of several different levels, each level can be treated as an individual factor contributing to severity. I would suggest using logistic regression to predict a “yes” or “no” response, such as severity. Each variable can be treated as a factor, which will break the factor into several levels to estimate an effect.
2. Class variables may be modeled more effectively using some other technique than logistic regression, but it is the only type of analysis I know how to do for a binary response. Using each factor as a variable may overparameterize the model.