

HW 11: US Traffic Fatalities: 1980 - 2004

Part-1

Contents

1 (30 points, total) Build and Describe the Data	1
2 (15 points) Preliminary Model	2
3 (15 points) Expanded Model	2

In this two-part project (HW-11 and HW-12), you are required to answer the following **causal** question:

“Do changes in traffic laws affect traffic fatalities?”

To answer this question, please complete the tasks specified below using the data provided in `data/driving.Rdata`. This data includes 25 years of data that cover changes in various state drunk driving, seat belt, and speed limit laws.

Specifically, this data set contains data for the 48 continental U.S. states from 1980 through 2004. Various driving laws are indicated in the data set, such as the alcohol level at which drivers are considered legally intoxicated. There are also indicators for “per se” laws—where licenses can be revoked without a trial—and seat belt laws. A few economics and demographic variables are also included. The description of each of the variables in the dataset is also provided in the dataset.

```
load(file="./data/driving.RData")

## please comment these calls in your work
#glimpse(data)
#desc
```

1 (30 points, total) Build and Describe the Data

- (5 points) Load the data and produce useful features. Specifically:
 - Produce a new variable, called `speed_limit` that re-encodes the data that is in `s155`, `s165`, `s170`, `s175`, and `s1none`;
 - Produce a new variable, called `year_of_observation` that re-encodes the data that is in `d80`, `d81`, `...`, `d04`.
 - Produce a new variable for each of the other variables that are one-hot encoded (i.e. `bac*` variable series).
 - Rename these variables to sensible names that are legible to a reader of your analysis. For example, the dependent variable as provided is called, `totfatrte`. Pick something more sensible, like, `total_fatalities_rate`. There are few enough of these variables to change, that you should change them for all the variables in the data. (You will thank yourself later.)
- (5 points) Provide a description of the basic structure of the dataset. What is this data? How, where, and when is it collected? Is the data generated through a survey or some other method? Is the data that is presented a sample from the population, or is it a *census* that represents the entire population? Minimally, this should include:
 - How is the our dependent variable of interest `total_fatalities_rate` defined?

3. (20 points) Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable `total_fatalities_rate` and the potential explanatory variables. Minimally, this should include:
 - How is the our dependent variable of interest `total_fatalities_rate` defined?
 - What is the average of `total_fatalities_rate` in each of the years in the time period covered in this dataset?

As with every EDA this semester, the goal of this EDA is not to document your own process of discovery – save that for an exploration notebook – but instead it is to bring a reader that is new to the data to a full understanding of the important features of your data as quickly as possible. In order to do this, your EDA should include a detailed, orderly narrative description of what you want your reader to know. Do not include any output – tables, plots, or statistics – that you do not intend to write about.

2 (15 points) Preliminary Model

Estimate a linear regression model of *totfatrt* on a set of dummy variables for the years 1981 through 2004 and interpret what you observe. In this section, you should address the following tasks:

- Why is fitting a linear model a sensible starting place?
- What does this model explain, and what do you find in this model?
- Did driving become safer over this period? Please provide a detailed explanation.
- What, if any, are the limitation of this model. In answering this, please consider **at least**:
 - Are the parameter estimates reliable, unbiased estimates of the truth? Or, are they biased due to the way that the data is structured?
 - Are the uncertainty estimate reliable, unbiased estimates of sampling based variability? Or, are they biased due to the way that the data is structured?

3 (15 points) Expanded Model

Expand the **Preliminary Model** by adding variables related to the following concepts:

- Blood alcohol levels
- Per se laws
- Primary seat belt laws (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)
- Secondary seat belt laws
- Speed limits faster than 70
- Graduated drivers licenses
- Percent of the population between 14 and 24 years old
- Unemployment rate
- Vehicle miles driven per capita.

If it is appropriate, include transformations of these variables. Please carefully explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed.

- How are the blood alcohol variables defined? Interpret the coefficients that you estimate for this concept.
- Do *per se laws* have a negative effect on the fatality rate?
- Does having a primary seat belt law?