Lab 3: Panel Models

US Traffic Fatalities: 1980 - 2004

Contents

1 U.S. traffic fatalities: 1980-2004

In this lab, we are asking you 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 the each of the variables in the dataset is also provided in the dataset.

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

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

```
##   year state sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl bac10 bac08
## 1 1980     1    1    0    0    0      0        0     18       0   0     1     0
## 2 1981     1    1    0    0    0      0        0     18       0   0     1     0
## 3 1982     1    1    0    0    0      0        0     18       0   0     1     0
## 4 1983     1    1    0    0    0      0        0     18       0   0     1     0
## 5 1984     1    1    0    0    0      0        0     18       0   0     1     0
## 6 1985     1    1    0    0    0      0        0     20       0   0     1     0
##   perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm statepop
## 1     0    940     422     236      3.20      1.437      0.803  3893888
## 2     0    933     434     248      3.35      1.558      0.890  3918520
## 3     0    839     376     224      2.81      1.259      0.750  3925218
## 4     0    930     397     223      3.00      1.281      0.719  3934109
## 5     0    932     421     237      2.83      1.278      0.720  3951834
## 6     0    882     358     224      2.51      1.019      0.637  3972527
##   totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24 sl70plus sbprim
## 1     24.14      10.84       6.06   29.37500  8.8      18.9        0      0
## 2     24.07      11.08       6.33   27.85200 10.7      18.7        0      0
## 3     21.37       9.58       5.71   29.85765 14.4      18.4        0      0
## 4     23.64      10.09       5.67   31.00000 13.7      18.0        0      0
## 5     23.58      10.65       6.00   32.93286 11.1      17.6        0      0
## 6     22.20       9.01       5.64   35.13944  8.9      17.3        0      0
##   sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91 d92 d93 d94 d95 d96
## 1       0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 2       0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 3       0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 4       0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
## 5       0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
## 6       0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0
##   d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc
## 1   0   0   0   0   0   0   0   0     7543.874
## 2   0   0   0   0   0   0   0   0     7107.785
## 3   0   0   0   0   0   0   0   0     7606.622
## 4   0   0   0   0   0   0   0   0     7879.802
## 5   0   0   0   0   0   0   0   0     8333.562
## 6   0   0   0   0   0   0   0   0     8845.614
```

```
desc
```

```
##        variable                                          label
## 1          year                              1980 through 2004
## 2         state            48 continental states, alphabetical
## 3          sl55                              speed limit == 55
## 4          sl65                              speed limit == 65
## 5          sl70                              speed limit == 70
## 6          sl75                              speed limit == 75
## 7        slnone                                 no speed limit
## 8      seatbelt     =0 if none, =1 if primary, =2 if secondary
## 9        minage                           minimum drinking age
## 10      zerotol                             zero tolerance law
## 11          gdl                  graduated drivers license law
## 12        bac10                        blood alcohol limit .10
## 13        bac08                        blood alcohol limit .08
## 14        perse administrative license revocation (per se law)
## 15       totfat                       total traffic fatalities
## 16      nghtfat                     total nighttime fatalities
## 17      wkndfat                       total weekend fatalities
## 18    totfatpvm         total fatalities per 100 million miles
## 19   nghtfatpvm     nighttime fatalities per 100 million miles
## 20   wkndfatpvm       weekend fatalities per 100 million miles
## 21     statepop                               state population
## 22    totfatrte        total fatalities per 100,000 population
## 23   nghtfatrte    nighttime fatalities per 100,000 population
## 24   wkndfatrte       weekend accidents per 100,000 population
## 25   vehicmiles               vehicle miles traveled, billions
## 26         unem                     unemployment rate, percent
## 27    perc14_24          percent population aged 14 through 24
## 28     sl70plus                           sl70 + sl75 + slnone
## 29       sbprim                     =1 if primary seatbelt law
## 30      sbsecon                   =1 if secondary seatbelt law
## 31          d80                             =1 if year == 1980
## 32          d81
## 33          d82
## 34          d83
## 35          d84
## 36          d85
## 37          d86
## 38          d87
## 39          d88
## 40          d89
## 41          d90
## 42          d91
## 43          d92
## 44          d93
## 45          d94
## 46          d95
## 47          d96
## 48          d97
## 49          d98
## 50          d99
## 51          d00
## 52          d01
## 53          d02
## 54          d03
## 55          d04                             =1 if year == 2004
## 56 vehicmilespc
```

2 (30 points, total) Build and Describe the Data

1. (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 `sl55`, `sl65`, `sl70`, `sl75`, and `slnone`;
    - 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.)
2. (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.

```r
#head(data)
```

```r
#speed limit
data$speed_limit <- with(data,
  ifelse(as.integer(sl55 >= 0.5), 55,
    ifelse(as.integer(sl65 >= 0.5), 65,
      ifelse(as.integer(sl70 >= 0.5), 70,
        ifelse(as.integer(sl75 >= 0.5), 75,
          ifelse(as.integer(slnone >= 0.5), NA, NA))))))

#year of observation
year_vars <- grep("^d\\d{2}$", names(data), value = TRUE)
data$year_of_observation <- apply(data[, year_vars], 1, function(x) 1980 + which(x == 1) - 1)

data <- data[, !(names(data) %in% c(year_vars, "sl55", "sl65", "sl70", "sl75", "slnone"))]

#make sure they are indicating 0 and 1 respectively
data$bac10 <- round(data$bac10)
data$bac08<- round(data$bac08)
data$sbprim <- round(data$sbprim)
data$sbsecon <- round(data$sbsecon)

data$minage <- round(data$minage)
data$perse <- round(data$perse)
data$zerotol <- round(data$zerotol)
data$gdl <- round(data$gdl)
data$sl70plus <- round(data$sl70plus)

#rename for clarity
names(data)[names(data) == "totfatpvm"] <- "total_fatalities_pvm"
names(data)[names(data) == "nghtfatpvm"] <- "night_fatalities_pvm"
names(data)[names(data) == "wkndfatpvm"] <- "weekend_fatalities_pvm"
names(data)[names(data) == "totfatrte"] <- "total_fatalities_rate"
names(data)[names(data) == "nghtfatrte"] <- "night_fatalities_rate"
names(data)[names(data) == "wkndfatrte"] <- "weekend_fatalities_rate"
names(data)[names(data) == "unem"] <- "unemployment_rate"

names(data)[names(data) == "gdl"] <- "graduated_drivers_license_law"
names(data)[names(data) == "zerotol"] <- "zero_tolerance_law"
names(data)[names(data) == "totfat"] <- "total_fatalities"
names(data)[names(data) == "nghtfat"] <- "total_nighttime_fatalities"
names(data)[names(data) == "wkndfat"] <- "total_weekend_fatalities"
```

```r
data$state[data$state == 1] <- 'al'
data$state[data$state == 3] <- 'az'
```

```r
data$state[data$state == 4] <- 'ar'
data$state[data$state == 5] <- 'ca'
data$state[data$state == 6] <- 'co'
data$state[data$state == 7] <- 'ct'
data$state[data$state == 8] <- 'de'
data$state[data$state == 10] <- 'fl'
data$state[data$state == 11] <- 'ga'
data$state[data$state == 13] <- 'id'
data$state[data$state == 14] <- 'il'
data$state[data$state == 15] <- 'in'
data$state[data$state == 16] <- 'ia'
data$state[data$state == 17] <- 'ks'
data$state[data$state == 18] <- 'ky'

data$state[data$state == 19] <- 'la'
data$state[data$state == 20] <- 'me'
data$state[data$state == 21] <- 'md'
data$state[data$state == 22] <- 'ma'
data$state[data$state == 23] <- 'mi'
data$state[data$state == 24] <- 'mn'
data$state[data$state == 25] <- 'ms'
data$state[data$state == 26] <- 'mo'
data$state[data$state == 27] <- 'mt'
data$state[data$state == 28] <- 'ne'
data$state[data$state == 29] <- 'nv'
data$state[data$state == 30] <- 'nh'
data$state[data$state == 31] <- 'nj'


data$state[data$state == 32] <- 'nm'
data$state[data$state == 33] <- 'ny'
data$state[data$state == 34] <- 'nc'
data$state[data$state == 35] <- 'nd'
data$state[data$state == 36] <- 'oh'
data$state[data$state == 37] <- 'ok'
data$state[data$state == 38] <- 'or'
data$state[data$state == 39] <- 'pa'
data$state[data$state == 40] <- 'ri'
data$state[data$state == 41] <- 'sc'


data$state[data$state == 42] <- 'sd'
data$state[data$state == 43] <- 'tn'
data$state[data$state == 44] <- 'tx'
data$state[data$state == 45] <- 'ut'
data$state[data$state == 46] <- 'vt'
data$state[data$state == 47] <- 'va'
data$state[data$state == 48] <- 'wa'
data$state[data$state == 49] <- 'wv'
data$state[data$state == 50] <- 'wi'
data$state[data$state == 51] <- 'wy'
```

The traffic fatalities data originates from the Fatality Analysis Reporting System (FARS), managed by the National Highway Traffic Safety Administration (NHTSA). This system gathers data on every traffic crash across the 48 contiguous United States that results in the death of a vehicle occupant or a nonmotorist. The data collection process, conducted by state employees, employs a standardized format to ensure consistency and comparability across different states.

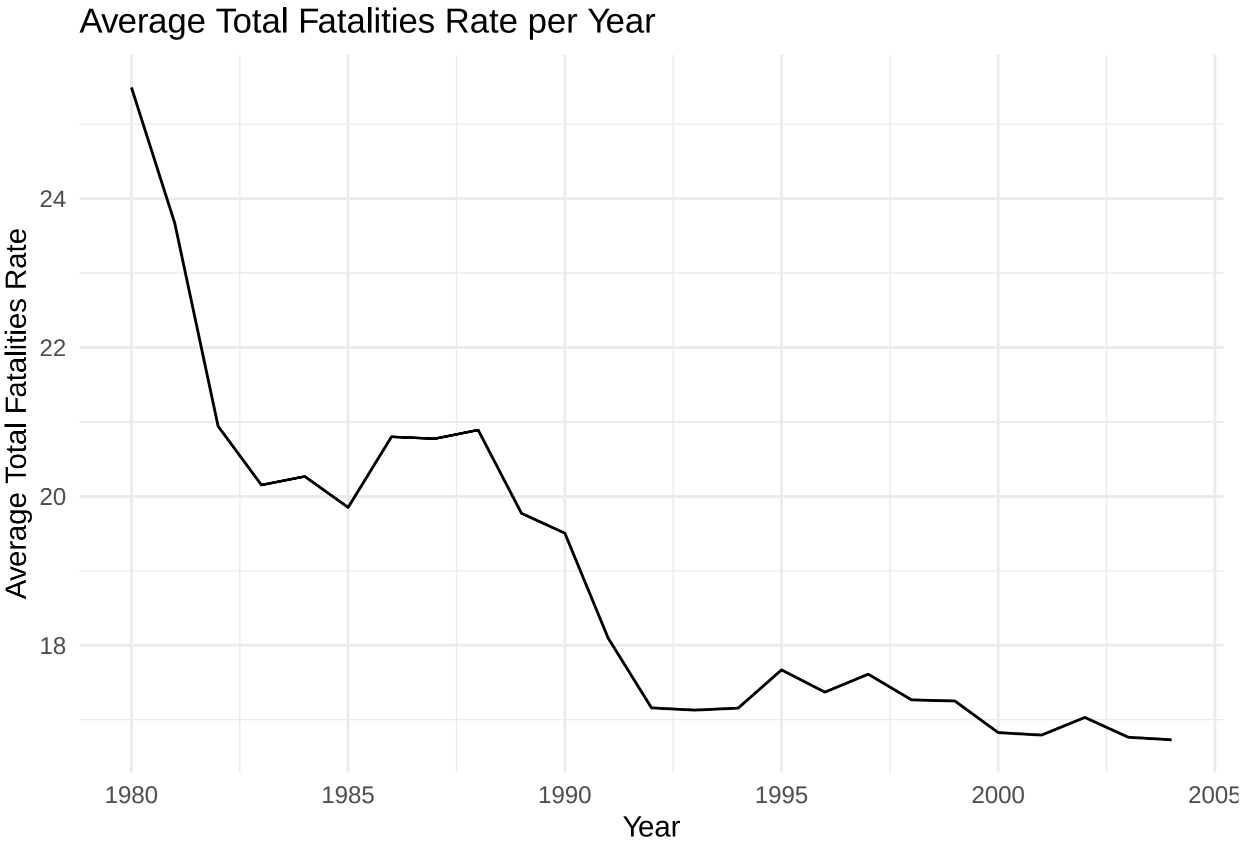
The dependent variable of interest, total__fatalities__rate, is defined as the number of traffic fatalities per 100,000 population at the state level over over the years 1980-2004. The dataset consists of approximately 1200 records, reflecting 48 records for each of the 25 years covered.

The independent variables include indicator variables for control legislation, including blood alcohol limit, graduated drivers license law, seat belt, and speed limit laws. Other controls include continuous variables for mileage traveled and demographic characteristics.

This dataset represents a census of traffic fatalities, rather than a sample. It documents every recorded instance of traffic-related fatalities within its scope and time frame, making it a complete record for the study period and not a subset of the total incidents.

```r
average_fatalities_per_year <- aggregate(total_fatalities_rate ~ year_of_observation, data, mean)

ggplot(average_fatalities_per_year, aes(x = year_of_observation, y = total_fatalities_rate)) +
    geom_line() +
    labs(title = "Average Total Fatalities Rate per Year",
         x = "Year",
         y = "Average Total Fatalities Rate") +
    theme_minimal()
```



The graph shows a significant downward trend in the average total fatalities rate per year from 1980 to 2004. From the early 1990s onwards, the rate shows some fluctuations but generally remains at a lower rate than at the start of the period observed. This suggests that road safety may have improved, possibly due to policy changes, improved vehicle safety, or other factors.

```r
average_by_age <- aggregate(total_fatalities_rate ~ minage, data = data, FUN = mean)
#print(average_by_age)
```

Table 1: Average Total Fatalities by Minimum Drinking Age

Minimum Drinking Age	18	19	20	21
Average Total Fatalities Rate	23.96	21.33	19.76	18.20

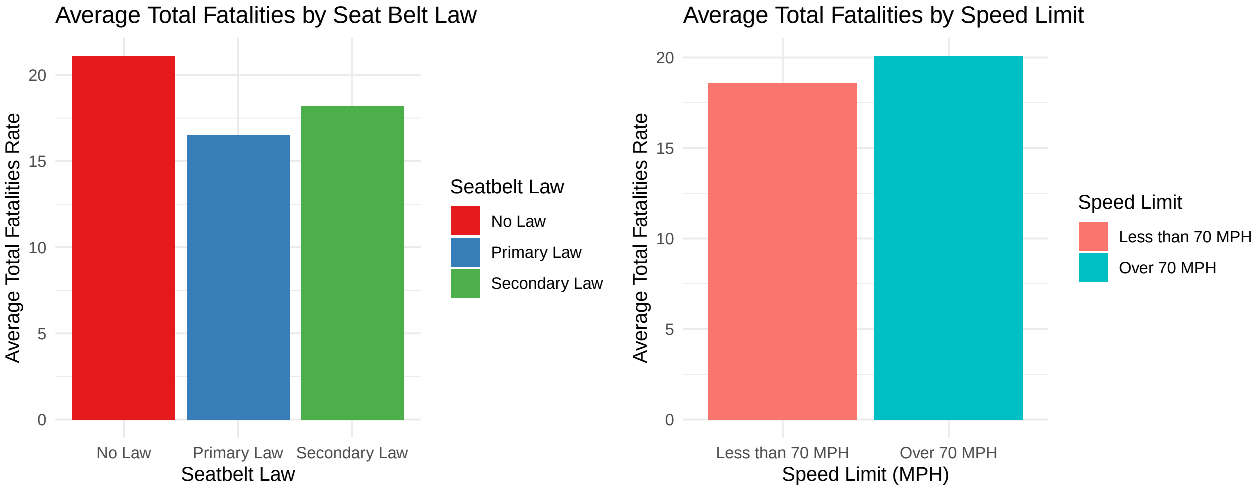
This table suggests that as the minimum legal drinking age increases, the average total fatalities rate decreases. This trend may imply that higher drinking age laws could be contributing to improved road safety, possibly by reducing alcohol consumption among younger drivers.

```r
p1_bar <- ggplot(data, aes(x=factor(seatbelt), y=total_fatalities_rate, fill=factor(seatbelt))) +
  geom_bar(stat="summary", fun=mean) +
  labs(title="Average Total Fatalities by Seat Belt Law",x="Seatbelt Law", y="Average Total Fatalities 
  theme_minimal() +
  scale_fill_brewer(palette="Set1", labels=c("No Law", "Primary Law", "Secondary Law")) +
  scale_x_discrete(labels=c("No Law", "Primary Law", "Secondary Law"))


p2_bar <- ggplot(data, aes(x=factor(sl70plus), y=total_fatalities_rate, fill=factor(sl70plus))) +
  geom_bar(stat="summary", fun=mean) +
  labs(title="Average Total Fatalities by Speed Limit", x="Speed Limit (MPH)", y="Average Total Fataliti
  theme_minimal() +
  scale_fill_brewer(palette="Set2",labels=c("0"="Less than 70 MPH", "1"="Over 70 MPH")) +
  scale_x_discrete(labels=c("0"="Less than 70 MPH", "1"="Over 70 MPH")) +
  scale_fill_discrete(labels=c("0"="Less than 70 MPH", "1"="Over 70 MPH"))
```

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```

```r
p1_bar | p2_bar
```

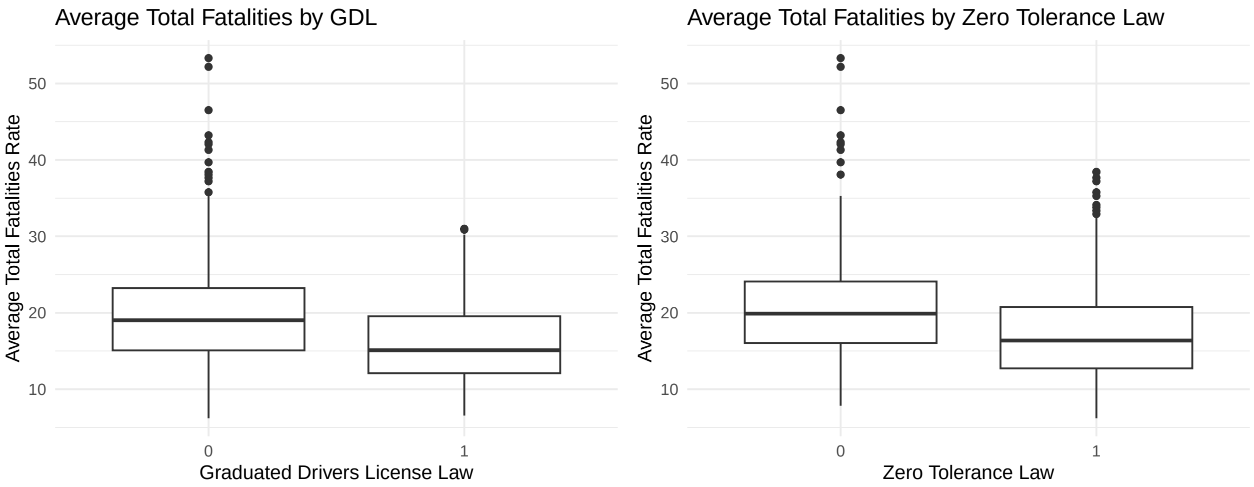


The seat belt law graph shows that regions with no seatbelt law (0) have the highest average fatality rate, while those with a primary enforcement law (1) show a lower rate, and secondary enforcement (2) has the lowest. This indicates that stricter seatbelt laws could be associated with reduced fatalities. The speed limit graph shows that when the speed limit is over 70 the average total fatalities rate is a bit higher, but interestingly not by much.

```r
p1_box <- ggplot(data, aes(x=factor(graduated_drivers_license_law), y=total_fatalities_rate)) +
  geom_boxplot() +
  labs(title="Average Total Fatalities by GDL",
       x="Graduated Drivers License Law",
       y="Average Total Fatalities Rate") +
  theme_minimal()

p2_box <- ggplot(data, aes(x=factor(zero_tolerance_law), y=total_fatalities_rate)) +
  geom_boxplot() +
  labs(title="Average Total Fatalities by Zero Tolerance Law",
       x="Zero Tolerance Law",
       y="Average Total Fatalities Rate") +
  theme_minimal()

p1_box | p2_box
```



The box plots comparing average total fatalities rates against the presence of Graduated Drivers License Law and Zero Tolerance Law show a lower median fatality rate when these laws are in place (indicated by ‘1’) as opposed to when they are not (indicated by ‘0’).

3 (15 points) Preliminary Model

Estimate a linear regression model of *totfatrte* 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?

4 (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?

5 (15 points) State-Level Fixed Effects

Re-estimate the **Expanded Model** using fixed effects at the state level.

- What do you estimate for coefficients on the blood alcohol variables? How do the coefficients on the blood alcohol variables change, if at all?
- What do you estimate for coefficients on per se laws? How do the coefficients on per se laws change, if at all?
- What do you estimate for coefficients on primary seat-belt laws? How do the coefficients on primary seatbelt laws change, if at all?

Which set of estimates do you think is more reliable? Why do you think this?

- What assumptions are needed in each of these models?

- Are these assumptions reasonable in the current context?

6 (10 points) Consider a Random Effects Model

Instead of estimating a fixed effects model, should you have estimated a random effects model?

- Please state the assumptions of a random effects model, and evaluate whether these assumptions are met in the data.
- If the assumptions are, in fact, met in the data, then estimate a random effects model and interpret the coefficients of this model. Comment on how, if at all, the estimates from this model have changed compared to the fixed effects model.
- If the assumptions are **not** met, then do not estimate the data. But, also comment on what the consequences would be if you were to *inappropriately* estimate a random effects model. Would your coefficient estimates be biased or not? Would your standard error estimates be biased or not? Or, would there be some other problem that might arise?

7 (10 points) Model Forecasts

The COVID-19 pandemic dramatically changed patterns of driving. Find data (and include this data in your analysis, here) that includes some measure of vehicle miles driven in the US. Your data should at least cover the period from January 2018 to as current as possible. With this data, produce the following statements:

- Comparing monthly miles driven in 2018 to the same months during the pandemic:
  - What month demonstrated the largest decrease in driving? How much, in percentage terms, lower was this driving?
  - What month demonstrated the largest increase in driving? How much, in percentage terms, higher was this driving?

Now, use these changes in driving to make forecasts from your models.

- Suppose that the number of miles driven per capita, increased by as much as the COVID boom. Using the FE estimates, what would the consequences be on the number of traffic fatalities? Please interpret the estimate.
- Suppose that the number of miles driven per capita, decreased by as much as the COVID bust. Using the FE estimates, what would the consequences be on the number of traffic fatalities? Please interpret the estimate.

8 (5 points) Evaluate Error

If there were serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors? Is there any serial correlation or heteroskedasticity?