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.

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.)

```
my_panel <- data %>%
  mutate(speed_limit = ifelse(sl55 >=0.5, "55",
                       ifelse(sl65 >=0.5, "65",
                       ifelse(sl70 >=0.5, "70",
                       ifelse(sl75 >=0.5, "75",
                       ifelse(slnone >=0.5, "nolim", "NAN"))))) %>%
  select(-sl55, -sl65, -sl70, -sl75, -slnone) %>% # Drop unused columns
  mutate(speed_limit = as.factor(speed_limit))    # Convert to factor
```

```
# Create an empty vector to store the names of non-zero columns for each row
result_vector <- character(nrow(my_panel))
year_names <- names(my_panel[,26:50])
# Loop through each row and find the non-zero column name (if any)
for (i in 1:nrow(my_panel)) {
  year_level <-  sub(".", "", year_names[my_panel[i,26:50] != 0])
  year_level <- paste(ifelse(as.numeric(year_level) < 80, "20", "19"), year_level, sep = "")
  result_vector[i] <- year_level
}
# Add the new column to the dataframe

my_panel$year_of_observation <- relevel(as.factor(result_vector), ref = "1980")
# Drop unneeded columns
my_panel <- cbind(my_panel[,1:25], my_panel[,51:53])
```

```
my_panel <- my_panel %>%
  mutate(blood_alc_lmt = ifelse(bac10 >=0.5, "1.0",
                         ifelse(bac08 >=0.5, "0.8", "None"))) %>%
  select(-bac10, -bac08, -sbprim, -sbsecon) %>%
  mutate(blood_alc_lmt = as.factor(blood_alc_lmt),
         seatbelt = as.factor(seatbelt),
         zerotol = as.logical(ifelse(zerotol >=0.5, T, F)),
         gdl = as.logical(ifelse(gdl >=0.5, T, F)),
         perse = as.logical(ifelse(perse >=0.5, T, F)))
```

```
# Create vectors for US states and their abbreviations
us_states <- c(
  "Alabama", "Alaska", "Arizona", "Arkansas", "California",
  "Colorado", "Connecticut", "Delaware", "District Columbia", "Florida", "Georgia",
  "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa",
  "Kansas", "Kentucky", "Louisiana", "Maine", "Maryland",
  "Massachusetts", "Michigan", "Minnesota", "Mississippi", "Missouri",
  "Montana", "Nebraska", "Nevada", "New Hampshire", "New Jersey",
  "New Mexico", "New York", "North Carolina", "North Dakota", "Ohio",
  "Oklahoma", "Oregon", "Pennsylvania", "Rhode Island", "South Carolina",
  "South Dakota", "Tennessee", "Texas", "Utah", "Vermont",
  "Virginia", "Washington", "West Virginia", "Wisconsin", "Wyoming")

state_abbreviations <- c(
  "AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DE", "DC", "FL", "GA",
  "HI", "ID", "IL", "IN", "IA", "KS", "KY", "LA", "ME", "MD",
```

```
  "MA", "MI", "MN", "MS", "MO", "MT", "NE", "NV", "NH", "NJ",
  "NM", "NY", "NC", "ND", "OH", "OK", "OR", "PA", "RI", "SC",
  "SD", "TN", "TX", "UT", "VT", "VA", "WA", "WV", "WI", "WY")

# Create a dataframe with US states and their abbreviations
us_states_df <- data.frame(State_name = us_states, Abbreviation = state_abbreviations, state = seq(1:51)
# Merge the names to the main panel
my_panel <- merge(my_panel, us_states_df, by.x = "state", by.y = "state", all.x = TRUE)
```

```
names(my_panel) <- c("state", "year", "seat_belt", "min_age", "zero_tolerance",
   "gradual_dl", "no_proof_guilt",
   "total_fatal", "night_fatal", "weekend_fatal",
   "total_fatal_per_miles", "night_fatal_per_miles", "weekend_fatal_per_miles",
   "state_pop",
   "total_fatal_per_pop", "night_fatal_per_pop", "weekend_fatal_per_pop",
   "tot_miles_driven", "unemployment", "percent_young", "speed_limit_above_70",
   "miles_per_capita", "dum_speed_limit", "dum_year", "dum_blood_alc",
   "state_name", "state_abbr")
```

2.1 Description of Dataset

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?

The data originates from the study by Freeman, D.G. (2007), "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws" published in Contemporary Economic Policy 25, 293–308. Author of the study made the data publicly accessible. It appears to describe an entire population of the US, although the details of data collection are behind a pay wall.

This data set is organized as a long table, with a set of observations for each US state (except Alaska, Hawaii, District of Columbia that are missing) for each year from 1980 to 2004. The years are coded as a numeric values as well as categorical variables. For each state-year category there are observations related to the state's traffic laws, traffic fatalities, and population/driving statistics. The data is balanced.

To make the data set suitable for our analysis, we applied a set of transformations (see above) and the rest of the description will discuss the transformed data set.

Some of the variable that can only be binary (zero tollerance that can either be implemented or not) sometimes have fractional value. Presumably, that indicates the years when the variable changed its value. To simplify interpretations, we rounded these values to the nearest integer.

2.1.1 Traffic Laws and Drinking Regulations

Speed limit is represented by two categorical variables:

- Max speed limit in the state: *dum_speed_limit* with categories *sl*55, *sl*65, *sl*70, *sl*75, *slnone*
- Weather max speed limit in the state exceeds 70 mph: $speed\_limit\_above_7 0$

Seat belt rules are represented by a categorical variable `seat_belt` with categories "0" if there is no law, '1' if not wearing a seat belt is a sufficient reason for a ticket, "2" if seat belt ticket can only be issued in conjunction with another violation.

Drivers licencing process is represented by `gradual_dl` indicating if there is a step-wise process to obtaining a full licence.

Drinking regulations are represented by the following variables:

- *min_age* is a number representing the legal drinking age
- *zero_tolerance* is a logical variable showing if zero alcohol tolerance is in force
- *blood_alc_lmt* is a categorical with three levels: 'bac08' for 0.8‰ 'bac10' for 1.0‰ and 'none' for no upper limit.
- no_proof_guilt is a logical variable that indicates if drivers licence can be suspended without positive medical proof of driving under influence.

2.1.2 Fatalities

Fatality is given as $total\_fatal$, total fatalities at this year and state and additional $night\_fatal$ and $weekend\_fatal$ variables giving tempral details. Data normalized to 100,000 people and to 100 milion miles driven is also presented in two groups of three variable.

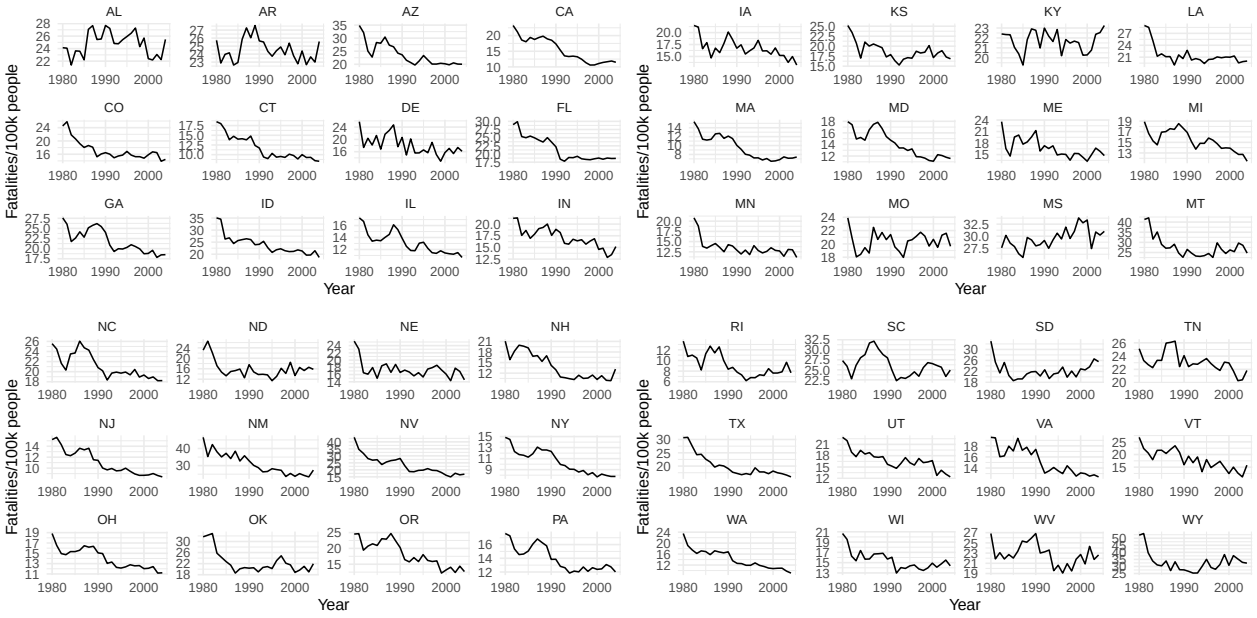
2.1.2.1 State demographics Additional information for each year and each state is presented in numeric variables:

- $state_pop$ is the state population
- *unemployment* is the percent of unemployment
- *percent_young* is the precent of poipulation 18 to 24 years old
- *miles_per_capita* is the number of miles driven per person

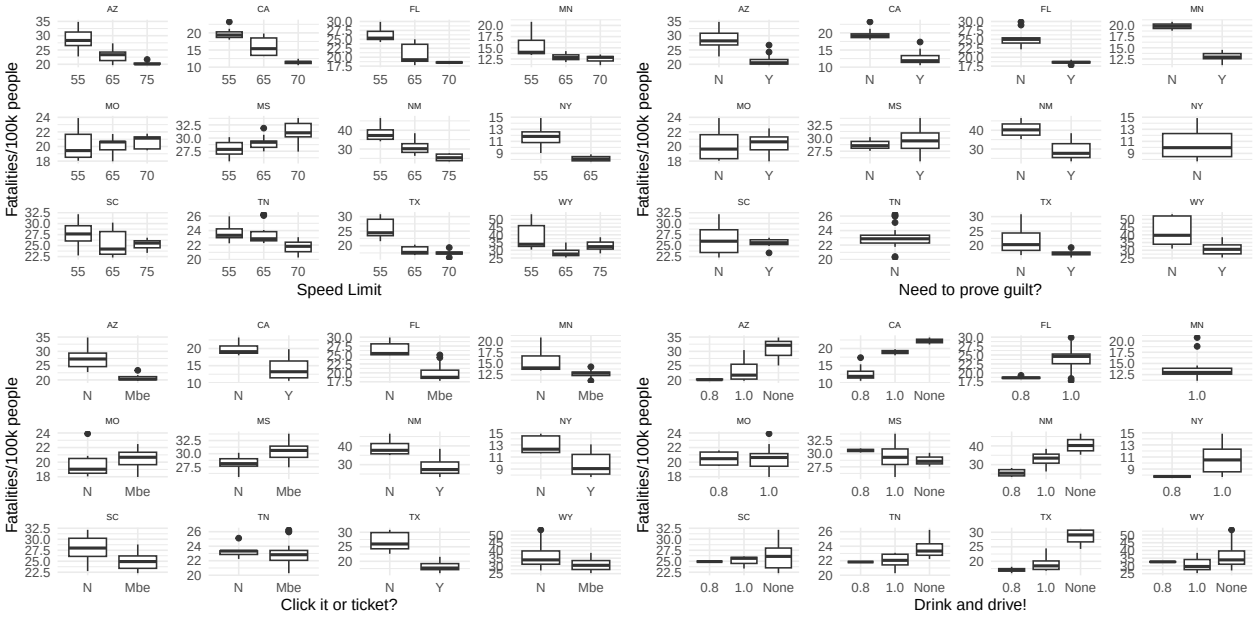
2.2 EDA

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.



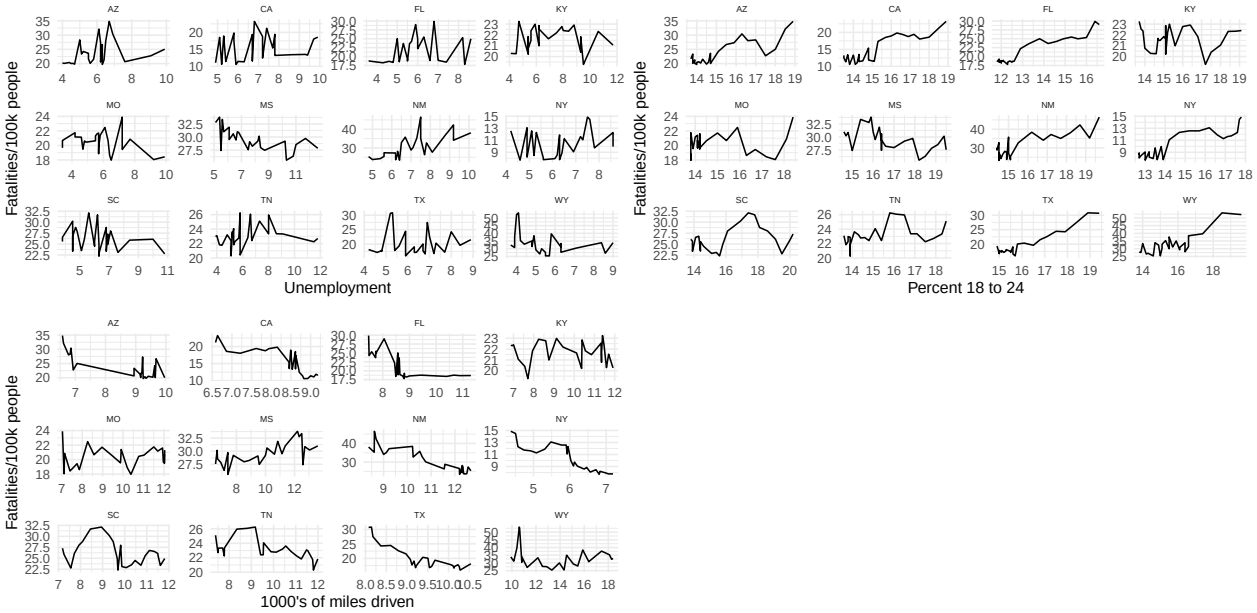
Plots above show how fatality per person changes over time for each state represented in the data set. There is a general downward trend that is particularly visible in case of the more populous states such as Texas, California and Florida. Only South Dakota is trending up consistently, but that might be a statistical artifact caused by small population. Some states, like Texas, Utah and Colorado, exhibit a very consistent downward trend, that would be hard to explain by the influence of any one-time decision. A large group of states (Oregon, Pennsylvania and Maryland, to name a few) exhibit strong increase in the fatality rate in the late 80's early 90's before sudden drop to almost modern levels. This patters would be consistent with the influence of some policy change that became popular in the late 80's.



Visualizing all 50 states would be visually overwhelming and for the next set of graphs we decided to use the most populous states and the states with the most interesting fatality-time patterns as representatives of the entire country.

The group of plots above show correlation of four policies on the road fatality rates: changes in speed limit, changes in seat belt laws, institution of no-proof guilt when it comes to DUI and blood alcohol limits. Unexpectedly, the speed limit seems to be inversely correlated with the fatality: the higher the limit the

lower the fatality. This is likely the case of spurious correlation. More stringent seat belt laws are also related to the lower fatality rate more often than not. Introduction of *per se* laws is also assotiated with the reduced fatality. Unsurprisingly, there is strong negative association between allowed alcohol concentration in the Having a mixture of expected and unexpected results, these plots are good reminders that correlation does not equal causation.



The graphs above demonstrate correlation of factors that are outside of the legislator's control and change gradually, the factors that could cause gradual change in the fatality that we observed above. Unemployment does not seem to have a clear correlation with the fatality, while percentage of the young population has clear positive corrrelation. Given the car insurance rates for young driver, we suspect that this effect is well known in the insurance world. Unexpectedly, the number of miles driven per person has negative correlation with the fatality: as the millage goes up the fatality goes down in most populous states. This effect is hard to rationalize and we expect to shed some light on it with more sophisticated analysis.

3 (15 points) Preliminary Model

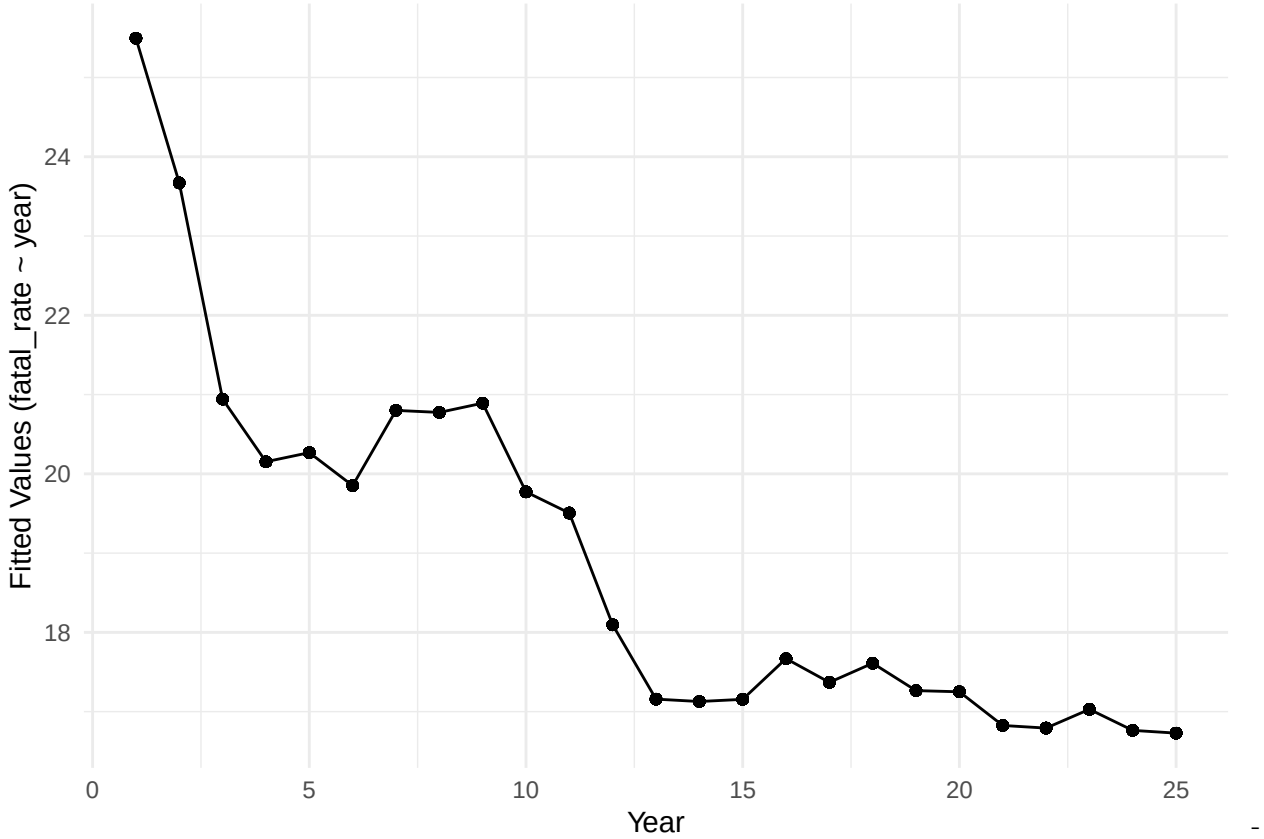
Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004 and interpret what you observe. In this section, you should address the following tasks:

- 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?

```
lsdv_model <- plm(total_fatal_per_pop ~ dum_year, data = my_panel,
                  effect = "individual", model = "pooling")
# summary(lsdv_model)

ggplot(data = broom::augment(lsdv_model), aes(x = as.numeric(dum_year), y = .fitted)) +
  geom_point() +
  geom_line() +
```

```
  labs(x = "Year", y = "Fitted Values (fatal_rate ~ year)",
       color = "state") +
  theme(legend.position = "none")
```



-

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?