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.)

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.

```
#loading the file
load(file="./data/driving.RData")
print("data dimension")
```

```
## [1] "data dimension"
```

```
dim(data)
```

```
## [1] 1200   56
```

```
print("column names")
```

```
## [1] "column names"
```

```
colnames(data)
```

```
##  [1] "year"          "state"         "sl55"          "sl65"          "sl70"
##  [6] "sl75"          "slnone"        "seatbelt"      "minage"        "zerotol"
## [11] "gdl"           "bac10"         "bac08"         "perse"         "totfat"
## [16] "nghtfat"       "wkndfat"       "totfatpvm"     "nghtfatpvm"    "wkndfatpvm"
## [21] "statepop"      "totfatrte"     "nghtfatrte"    "wkndfatrte"    "vehicmiles"
## [26] "unem"          "perc14_24"     "sl70plus"      "sbprim"        "sbsecon"
## [31] "d80"           "d81"           "d82"           "d83"           "d84"
## [36] "d85"           "d86"           "d87"           "d88"           "d89"
## [41] "d90"           "d91"           "d92"           "d93"           "d94"
## [46] "d95"           "d96"           "d97"           "d98"           "d99"
## [51] "d00"           "d01"           "d02"           "d03"           "d04"
## [56] "vehicmilespc"
```

```
print("na.values")
```

```
## [1] "na.values"
```

```
sum(is.na(data))
```

```
## [1] 0
```

```
table(data$year)
```

```
##
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
##   48   48   48   48   48   48   48   48   48   48   48   48   48   48   48   48
## 1996 1997 1998 1999 2000 2001 2002 2003 2004
##   48   48   48   48   48   48   48   48   48
```

```
table(data$state)
```

```
##
##  1  3  4  5  6  7  8 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
## 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

2.1 Overview

This data set is compiled from the Fatality Analysis Reporting System (FARS) by NHTSA7. FARS gathers all data on all traffic crashes that results in a death of a nonmotorsit or the vehicle occupant. All data is gathered by state employees using a standard format for comparability across states.

The data set has 1200 observations across 56 columns for the years 1980 to 2004 for 48 continental states. The data has a panel structure, where each state has 25 observations in time for different variables. The data set has no missing values. Point to notes is Alaska, Hawaii, and district of Columbia is missing from the data set. This makes sense since Alaska and Hawaii are not continental states and District of Columbia is not a state. The rest of the variables can be summarized as follows.

-***Speed limit laws:*** Variables sl55, sl65, sl70, sl75, slnone, sl70plus. These indicate speeds limits of 55, 65, 70, 75 mph, no speed limit and any speed over 70mph.

-***DUI Laws:*** variables minage, zerotol, bac10, bac08, Per se law. These indicate the minimum drinking age, zero tolerance law, blood alcohol level of 0.10 and 0.08. Per Se law gives DMV the right to revoke license of individuals who refuse to take the breath test, has 0.08% blood alcohol level or is a minor with blood alcohol level of 0.01%

-***Seat Belt Laws:*** Variables seatbelt, sbprim, sbsecon. These have values 0, 1, and 2 indicating no seat belt required, primary driver seatbelt required, and secondary driver seatbelt required, respectively.

-***State statistics:*** Variables statepop, unem, perc14_24, vehicmilespc. These variables indicate state population, unemployment, percentage of drivers between 14-24, vehicle mile driven in a specific year, respectively.

-***Time in Years:*** variables d80 to d04. These indicate years 1980 to 2004.

-***Fatalities:*** totfatrte which indicates total fatalities per 100,000 population. There also other variables that explains fatalities during night and weekend.

2.2 Data Transformations

Some of the variables in data set, such as speed limit columns, blood_level columns, and so on, have binary information indicating a value of 0 or 1. However, there are some values in said columns that have values other than 0 or 1, which indicates a change of a law mid year, in which case the value is written as a decimal, indicating what portion of the year which law was in place. We decided to take the majority of the values to circumnavigate around this issues and make model implementation and interpretation easier by keeping the columns binarized. In other words, in columns with binary values, any value above 0.5 would get a score of 1 and any value below 0.5 would get a score of zero.

To be specific, the following the variables will be binarized.

-sl55, sl65, sl70, sl75, slnone, sl70plus, zerotol, gdl, bac10, bac08, perse, sl70plus, and perse

Furthermore, we will rename the columns to get a better understanding of what each column signifies. We will, also, create three new columns, that improves the data set readability. These new columns will be as follows:

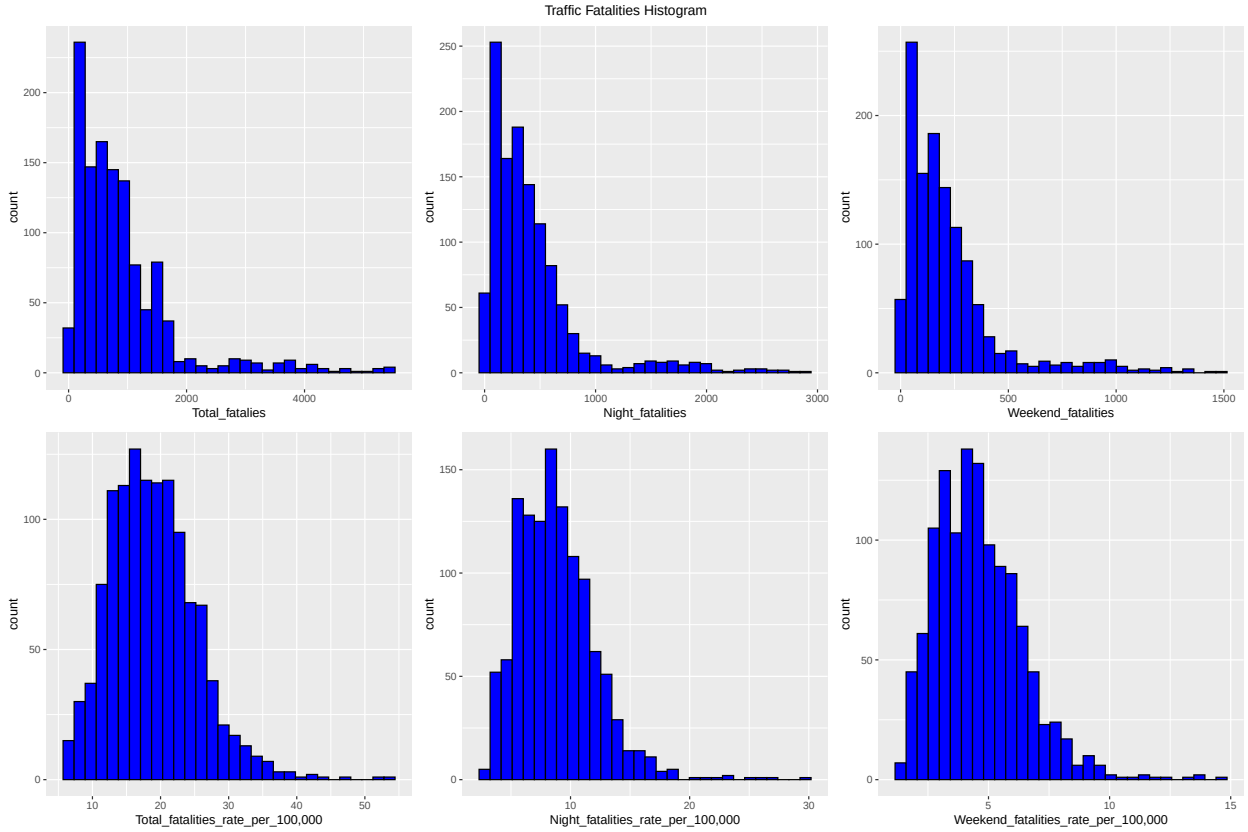
- ***speed_limit:*** This column will revert the one-hot encoded variables sl55, sl65, sl70, sl75, slnone into a single column.
- ***year_of_observation:*** This column will revert the one-hot encoded variables d80 to d04.
- ***blood_alcohol_levels:*** This column will revert the one-hot encoded variables bac10, bac08.

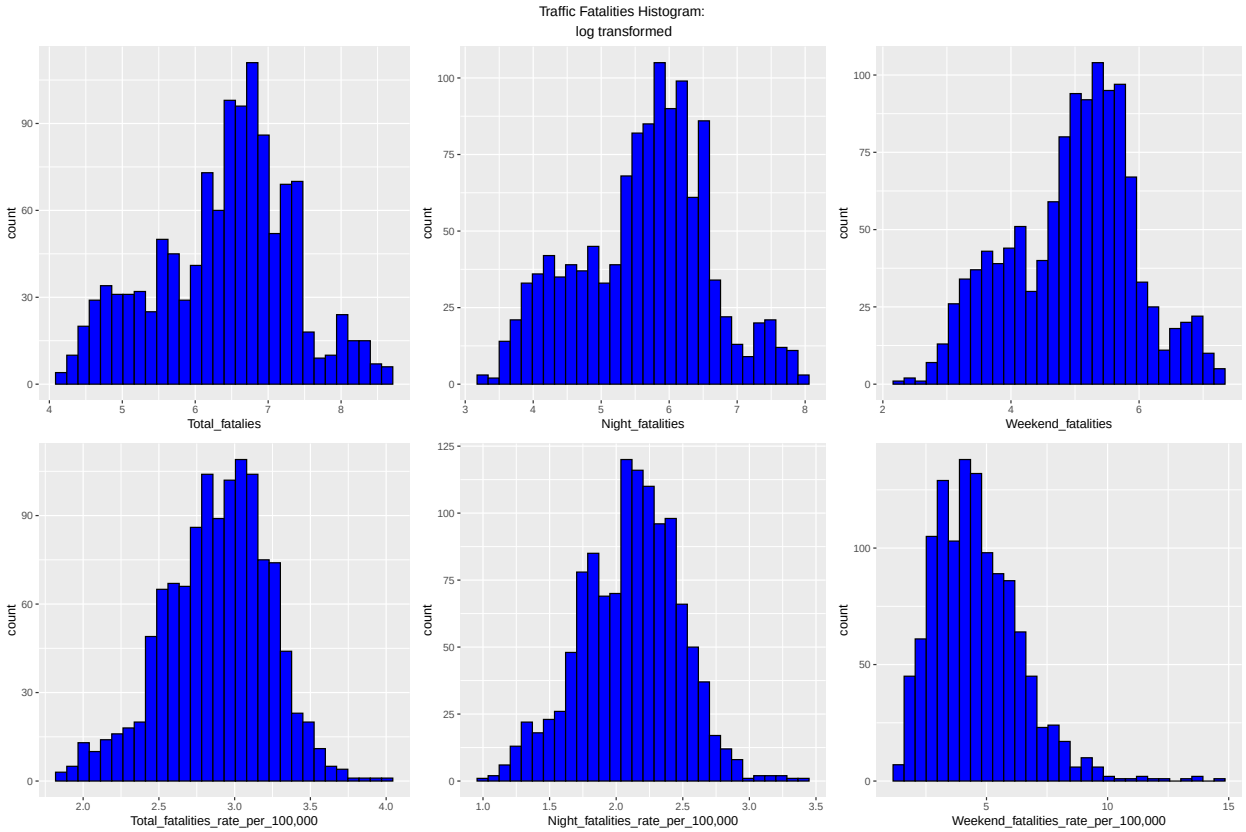
We will, then, drop the columns that have reverted from one-hot encoding and the columns totfatpvm, nghtfatpvm, and wkndfatpvm because they are derived $(totfat/(10 * vehicmiles)$.

For the state variable, the state values are numbers between 1 to 51. We will use state codes to find what number belongs to what state and assign the state abbreviations accordingly.

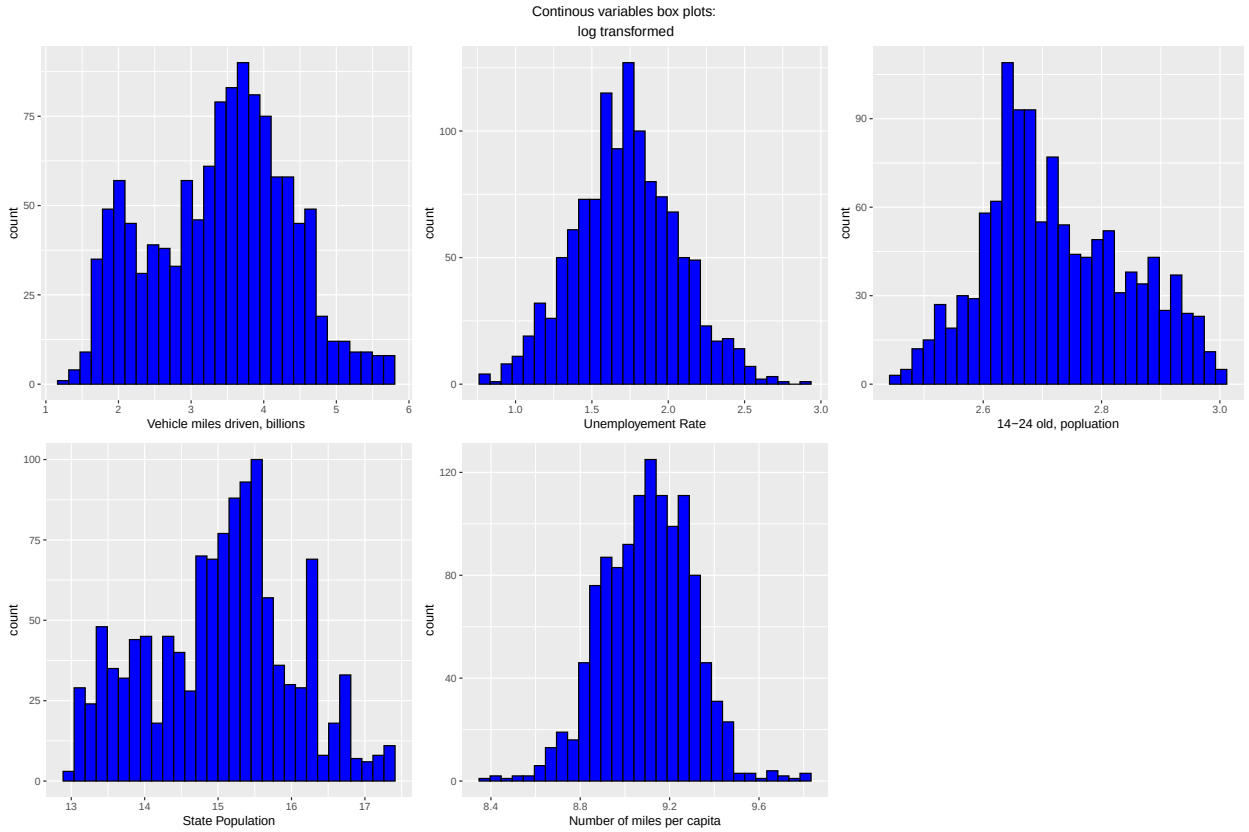
2.3 Exploratory Data Analysis

First, let's take a look at the histogram of fatality variables. As we can see in the boxplots of all the variables are skewed to the left. A log transformation may help normalize the data.

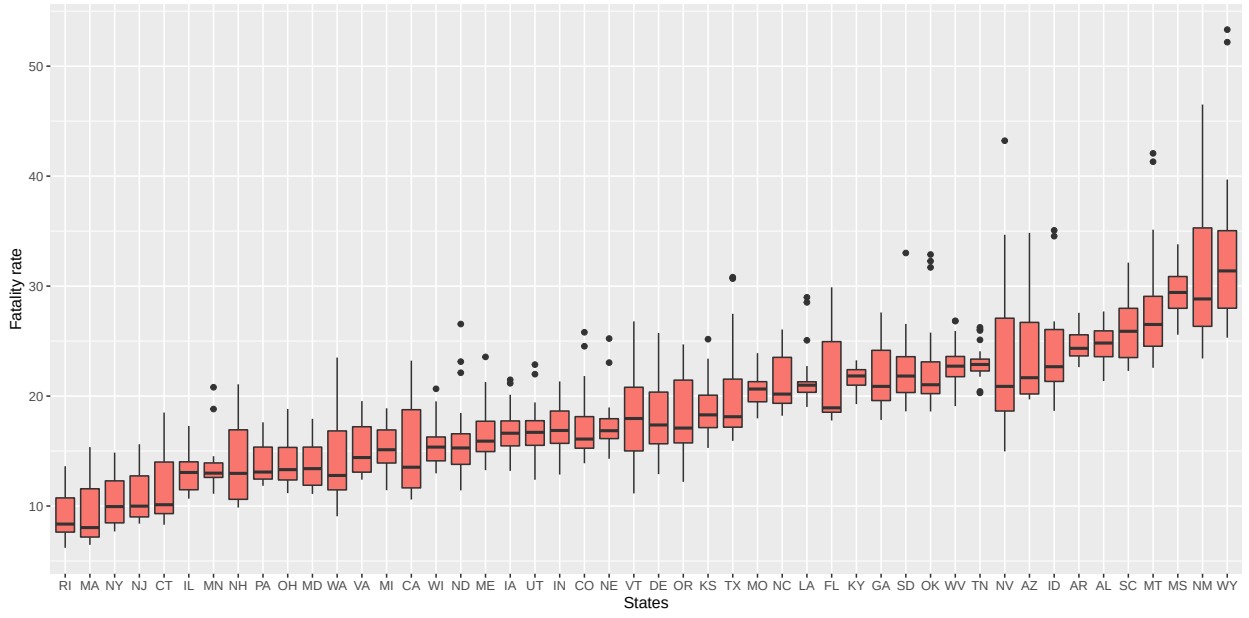




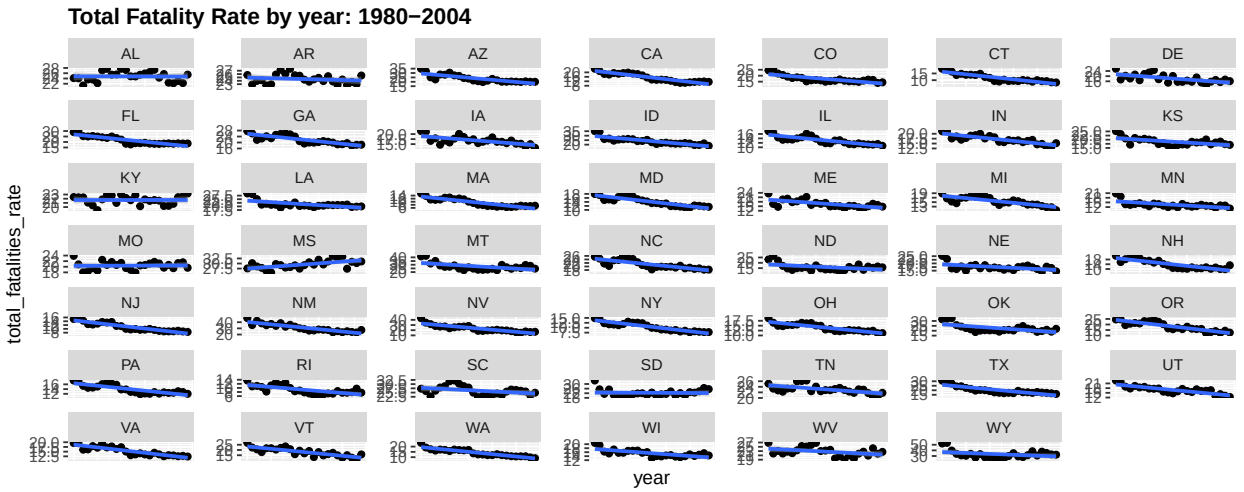
Looks like the log transformation did help normalize all the fatality variables. We can see that the nightly fatality rates is slightly higher than the weekend. Next, let's look into the other continuous variables. It appears that some of variables are right skewed again. Log transformation might help normalize the distribution. The graphs below shows log transformed graphs.



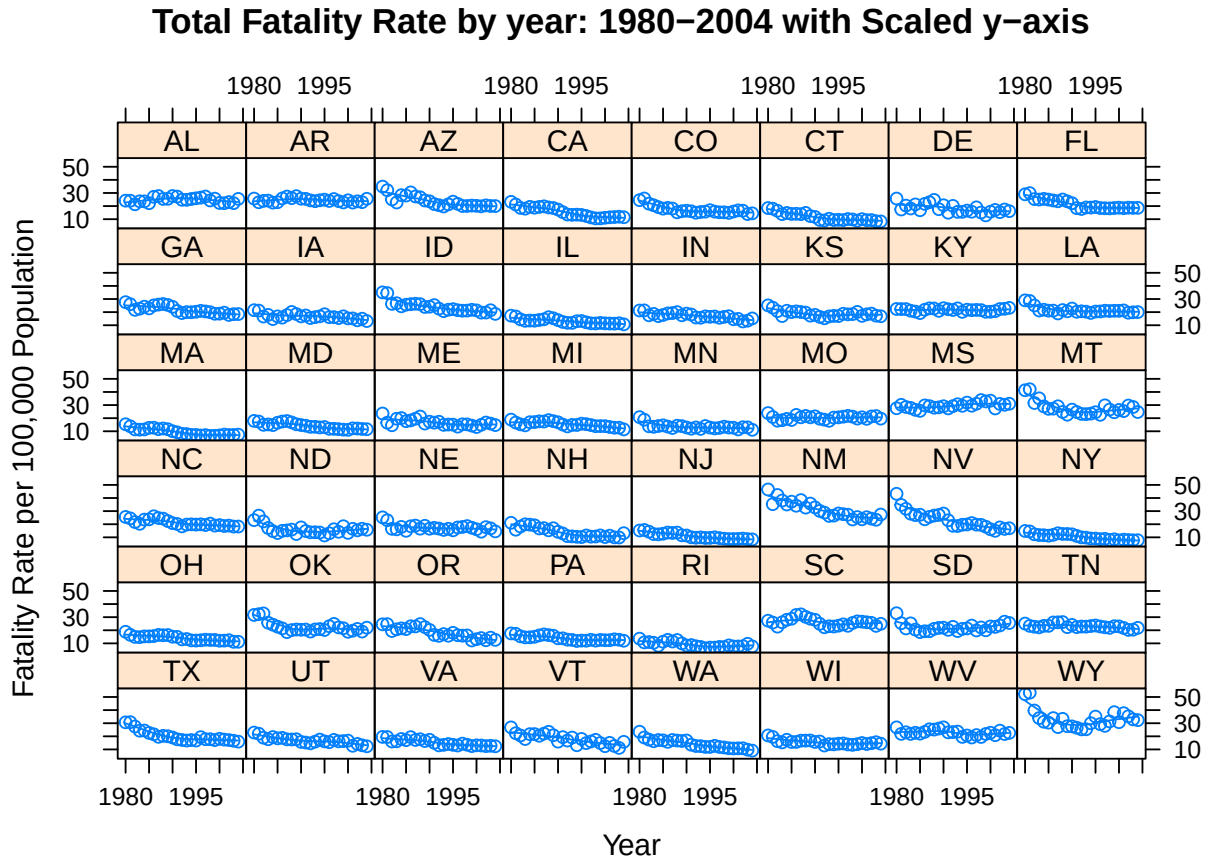
Next, let's check for heterogeneity of the fatality rate. It appears that differnet states have different fatality rates. The variance also differs across the states, though some states do have higher variance. Lowest fatalities rate state is Rhode Island(RI) and highest fatalities rate state is Wyoming(WY).



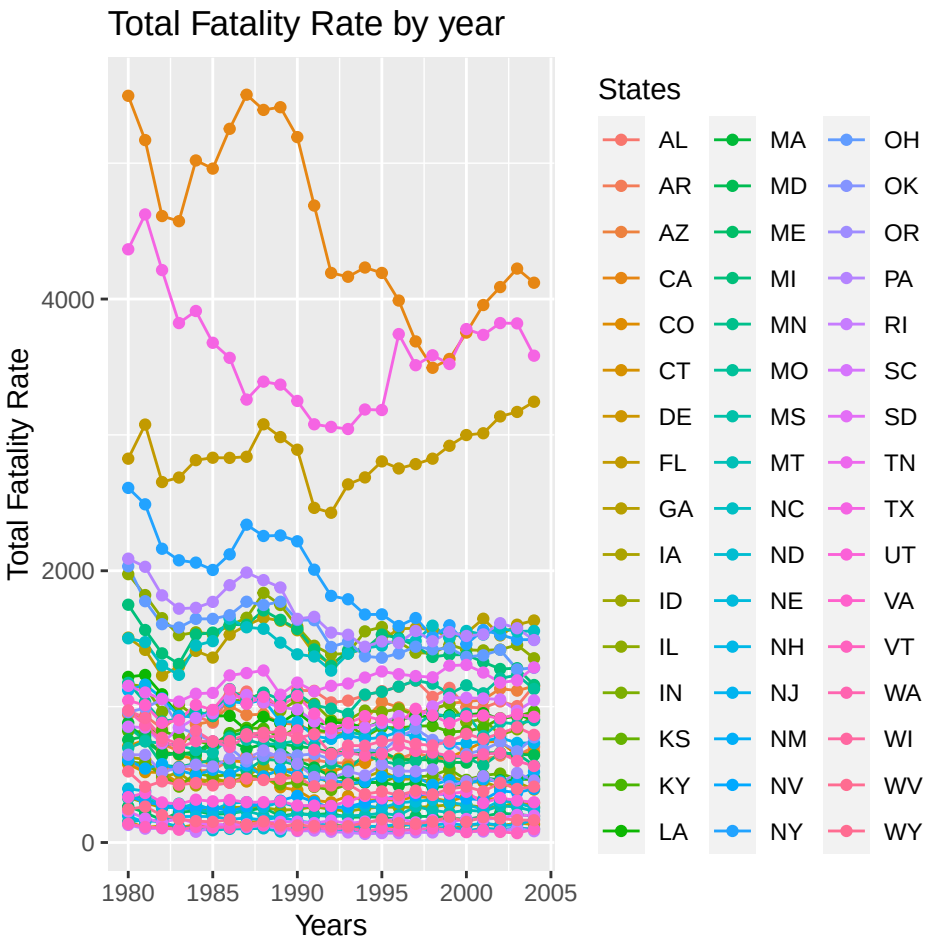
Next, let's look how the fatality rates differs across time in each state. As we can see, most states' fatality rates go down over time, a few state seems to have relatively flat fatalities rate over time, such as KY and AL. And some states have an increase in fatalities rate over time.



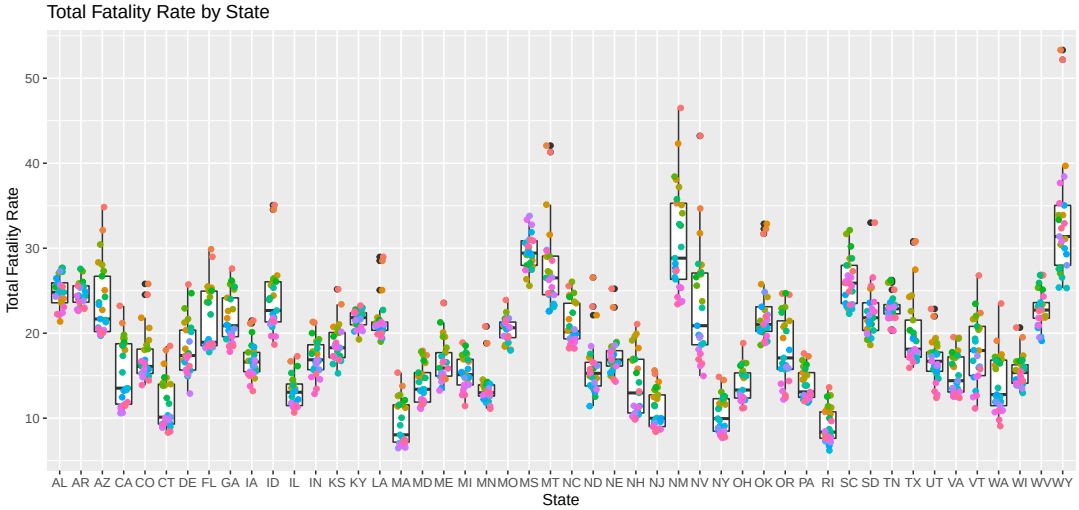
As you can see in the graph above, the y axis scales are not the same across states. In this next graphs, let's scale the y-axis and let's see if we can find new insight.Looks like scaling the y-axis doesn't really tell us any more information except that most states fatalitities rate flattens and/or decreases.



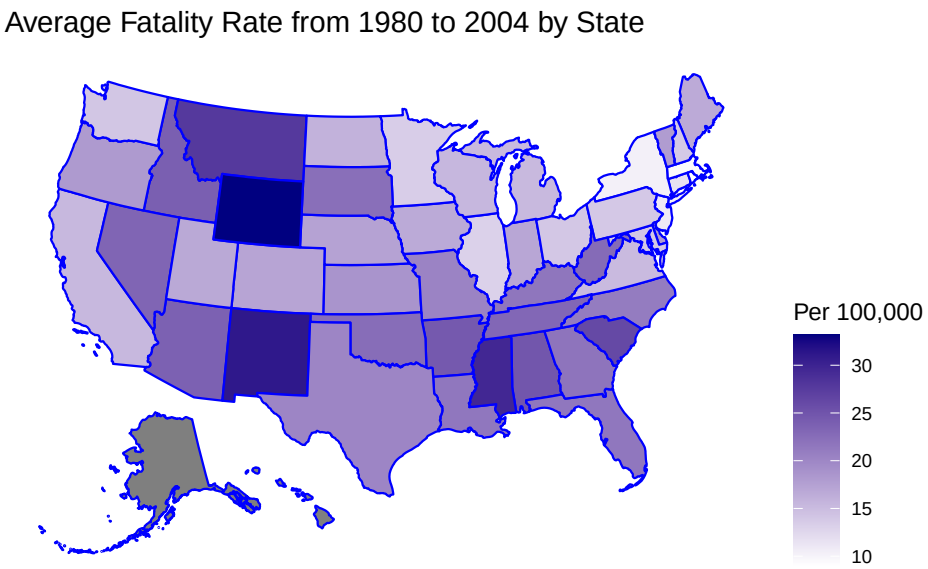
Next, let's take a look at the fatalities rate of each state as time passes by, but all in the same graph. See if anything stands out. We can immediately see that some states have much higher fatality rates across than the rest. Most states seems to have very similar fatalities rates.



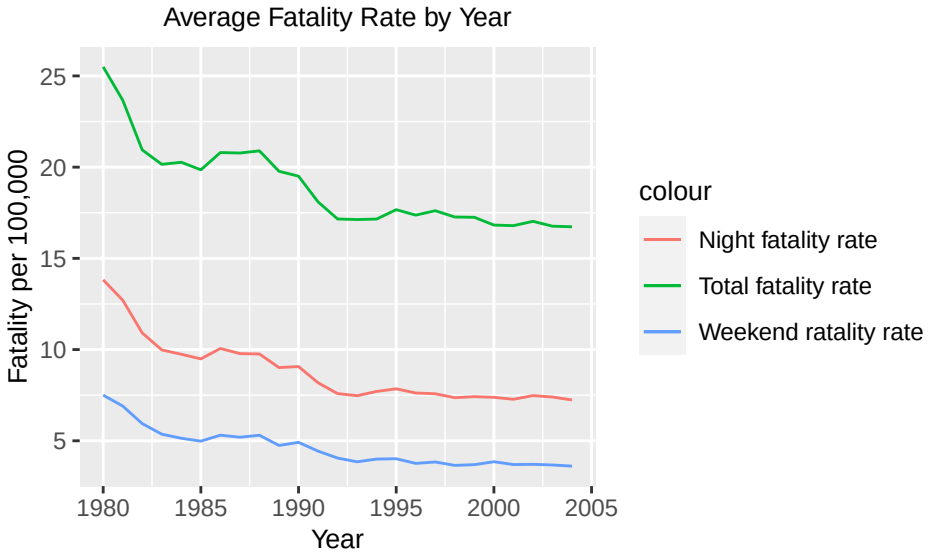
Next, let's try to see if there are any interesting insights in fatalities across states but focusing on the year variable. As we can see in the graph, the fatalites rates seem be higher across all states between 1980-1985 and seem to be lower across all states in 2000 and onwards.



Next, let's see the average fatalities across all states from 1980 and 2004. As we can see, it appears that the Wyoming and New Mexico have the highest average fatality rates in the nation.

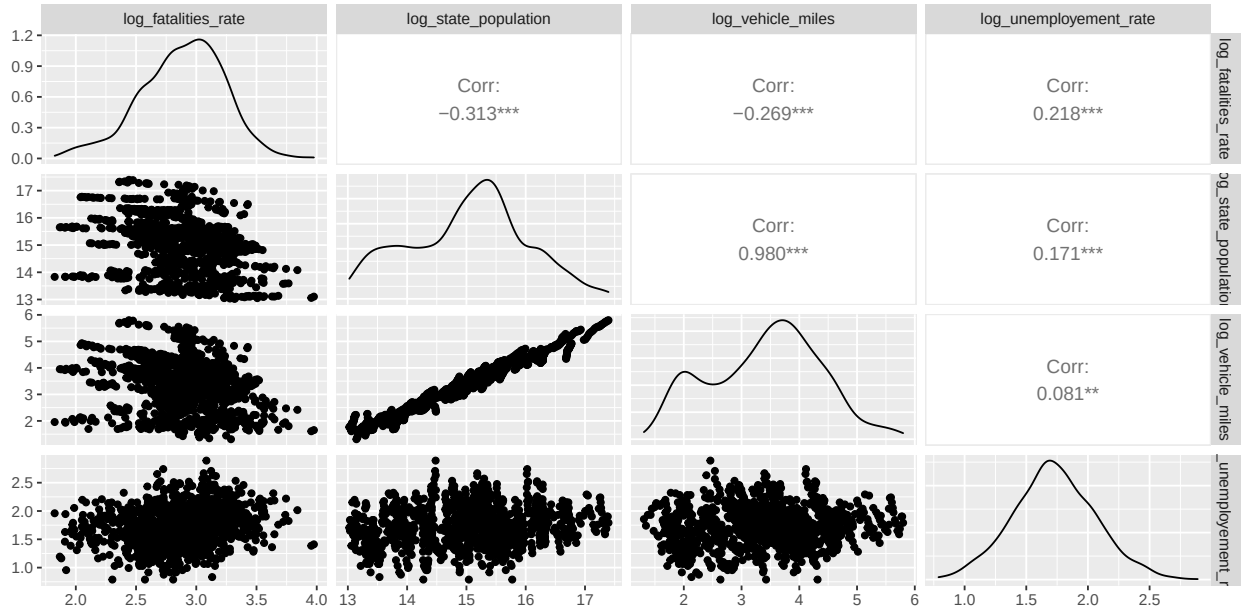


Let's also look at the average fatalities rate for the three fatality variables across time. It does appear all the average fatality rates are indeed trending down as time passes.

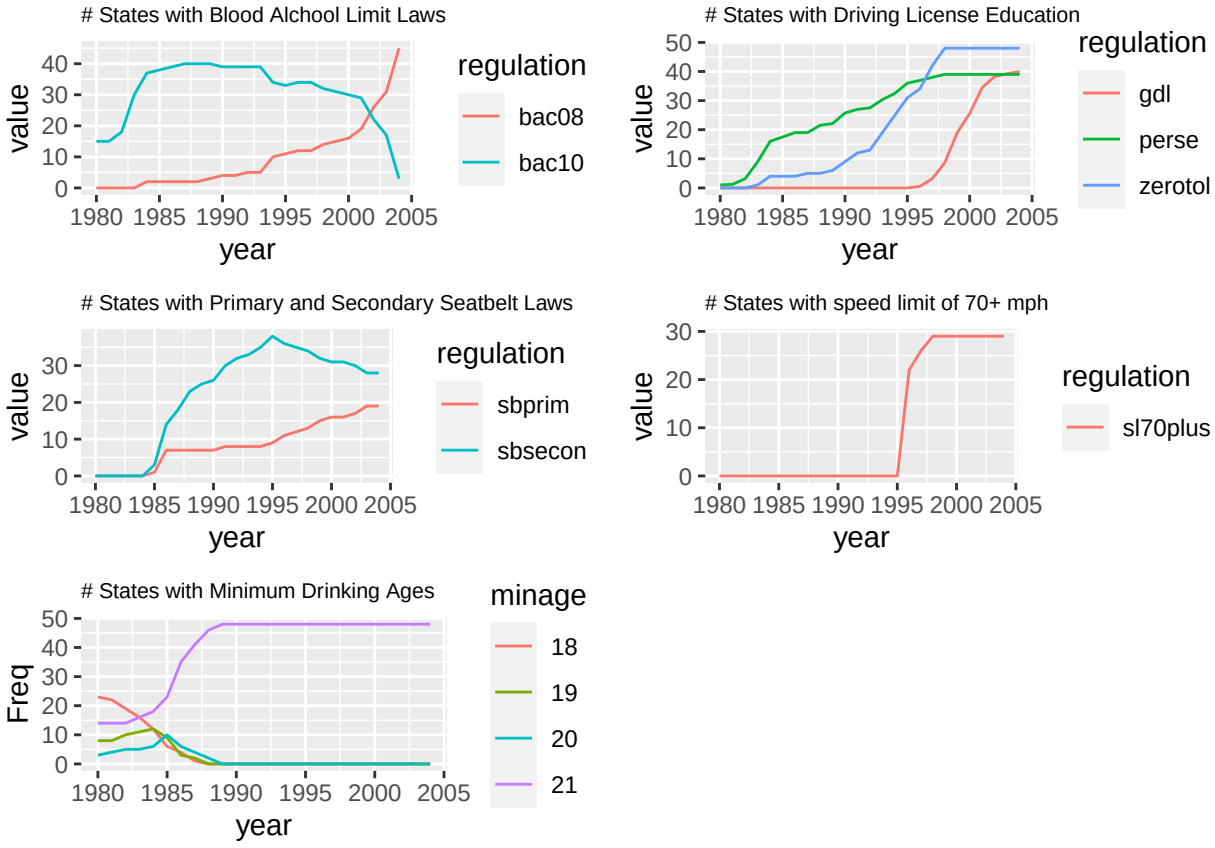


Now, let's shift the focus on continues variables. We will look at the scatter Plot Matrix for the continuous variables including total fatalities rate variables and see if anything stands out. We decided to have the variables in a log format since from the histogram analysis earlier, we had seen that a log transformation was beneficial. As we can see, the log of fatalities rate variable does has negative correlations with log of state

population and log vehicle miles driven and a negative correlation with log of unemployment rate. Vehicle miles driven and state population seems to have a linear relationship with a very high correlation.



Finally, we will use the original data set to take a look at blood alcohol laws, driving licence education, seat belt laws, speed limits above 70 mph, and minimum drinking age law. As we can see, the as time passes by, more states have become stricter about seat belt laws and blood alcohol laws. Also zero tolerance law, gdl and perse laws have become more prevalent in states as well as time has passed by. Interestingly, it appears that around 1987, minimum drinking age was changed to 21 across all states. Speed limit of 70+ mph has also become prevalent across states.



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:

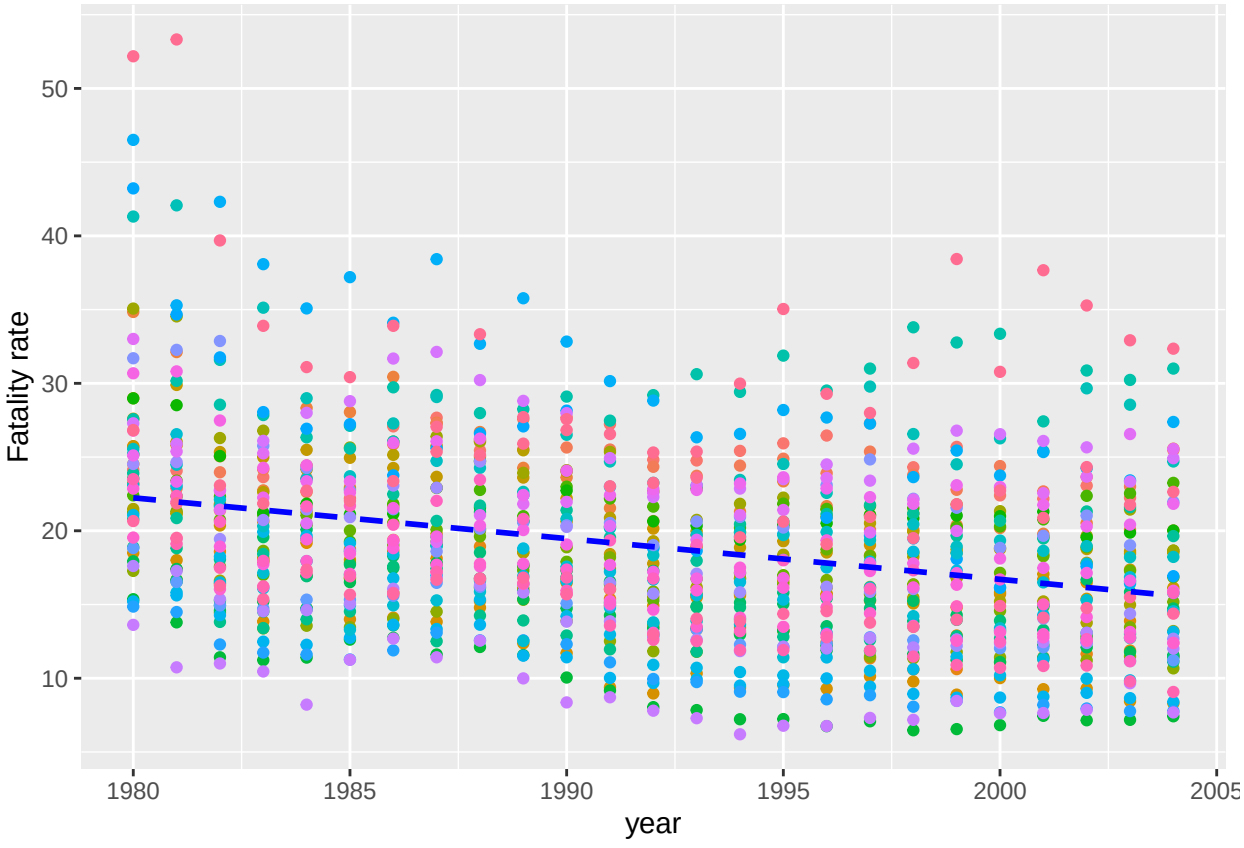
- Why is fitting a linear model a sensible starting place?
- What does this model explain, and what do you find in this model?
- Did driving become safer over this period? Please provide a detailed explanation.
- What, if any, are the limitation of this model. In answering this, please consider **at 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?
- a time based linear model or pooled OLS is a good starting point because as an exploratory analysis, it gives some information whether there is a time trend to the data. If there is a time trend, we will need to consider controlling for time when estimating effect of other treatments.

```
linear.mod <- lm(total_fatalities_rate ~ year, data = final_data)
summary(linear.mod)
```

```
##
## Call:
## lm(formula = total_fatalities_rate ~ year, data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -12.9201  -4.3576  -0.7668   3.6596  31.3606
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 569.58781   48.24189   11.81   <2e-16 ***
## year         -0.27644    0.02422  -11.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.05 on 1198 degrees of freedom
## Multiple R-squared:  0.09809,    Adjusted R-squared:  0.09734
## F-statistic: 130.3 on 1 and 1198 DF,  p-value: < 2.2e-16
```

- The linear model explains that there is a negative correlation between fatality rate and time (year), meaning as years increased there were less car accidental fatality. However, the model doesn't explain why there was a less fatality, only that there is negative time trend and time by itself cannot cause fatality to drop. To better understand the cause of less fatality over time, we need to control for other omitted variable such as regulations, alcohol consumption and unemployment rate and individual state difference. If we still observe significant negative time trend after controlling for these omitted variable bias, then we know there is something else outside of the data that is causing the fatality rate to drop.



- Above is all of the state's observation with the best fit linear line. The color on the graph represent each state. As we can see, The main problem is that in this linear model we do not observe ($state_i$), which is constant over time varies across individuals. Hence if we estimate the model in levels using OLS then $state_i$ will go into the error term: $\epsilon_{it} = state_i + u_{it}$.
- If $state_i$ is correlated with $year$, then putting $state_i$ in the error term can cause serious problems. This, of course, is **an omitted variable problem**. For this single regressor model:

$$plim\widehat{\beta_{ols}} = \beta + \frac{cov(year, state_i)}{\sigma_x^2}$$

- which shows that the OLS estimator is inconsistent unless $cov(year, state_i) = 0$. In the graph above, we do see covariance between state and year. Therefore our point estimate is both **inconsistent** and **biased**.
- Even if $state_i$ is uncorrelated with $year$, then $state_i$ is just another unobserved factor making up the residual, then, this OLS will not be efficient (smallest variance) because the error term ϵ_{it} is serially correlated. Therefore the standard formula for calculating the standard errors is wrong. As a result, our uncertainty estimate (standard errors) is also incorrect.

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?
- In EDA section, we saw some of the variables are quite skewed. Therefore, in this expanded linear model, we decided to log some skewed variables to make them more normally distributed. This helps with the first assumption of a linear model that says that explanatory and dependent variables are linear, because skewed data are less linear.
- We also factor the blood_alcohol_levels wit BAC08 (0.1%) is the baseline, as a result BAC08 (0.08%) and other(0.00%) becomes dummy variables.

```
final_data$blood_alcohol_levels <- relevel(factor(final_data$blood_alcohol_levels), ref = "BAC10")
linear.mod.exp <- lm(log(total_fatalities_rate) ~ year + as.factor(blood_alcohol_levels) + per_se_law +
                       + sl70plus + grad_driver_license + percent_popul_14_to_24
                       + log(unemployement_rate) + log(vehicmilespc)
                     , data = final_data)
summary(linear.mod.exp)
```

```
##
## Call:
## lm(formula = log(total_fatalities_rate) ~ year + as.factor(blood_alcohol_levels) +
##     per_se_law + as.factor(seatbelt) + sl70plus + grad_driver_license +
##     percent_popul_14_to_24 + log(unemployement_rate) + log(vehicmilespc),
##     data = final_data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62716 -0.13373 -0.00002  0.13161  0.67609
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         69.074157   4.788028  14.426  < 2e-16 ***
## year                                -0.040566   0.002438 -16.639  < 2e-16 ***
## as.factor(blood_alcohol_levels)BAC08 -0.022433   0.018299  -1.226   0.2205
## as.factor(blood_alcohol_levels)other  0.021577   0.018997   1.136   0.2563
## per_se_law                          -0.028728   0.014802  -1.941   0.0525 .
## as.factor(seatbelt)1                 0.017403   0.023899   0.728   0.4666
## as.factor(seatbelt)2                 0.030164   0.020321   1.484   0.1380
## sl70plus                             0.219258   0.020364  10.767  < 2e-16 ***
## grad_driver_license                  0.019902   0.021952   0.907   0.3648
## percent_popul_14_to_24               0.026697   0.005567   4.795 1.83e-06 ***
## log(unemployement_rate)              0.205288   0.021509   9.544  < 2e-16 ***
## log(vehicmilespc)                    1.516835   0.044300  34.240  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2077 on 1188 degrees of freedom
## Multiple R-squared:  0.6402, Adjusted R-squared:  0.6368
## F-statistic: 192.1 on 11 and 1188 DF,  p-value: < 2.2e-16
```

- As expected the time effect has dropped significantly after controlling other factors. However it is still statistically significant, which means we most likely still need it in the model because it serve as a control for factors that is not observed in the model
- In terms of blood alcohol level, when compared to the baseline alcohol level 0.1%, there is no statistical significant relationship between more strict alcohol level laws and fatality rate. This is counter intuitive, which suggests that the current model specification is incorrect. A fixed effect model may be more appropriate.
- As expected, **per se** law do have a statistically significant negative effect on fatality rate.
- Compared to baseline of no seat belt law, both primary and secondary seat belt law requirement have no statistical significant relationship to fatality rate. This is also counter intuitive, which suggests that the current model specification is incorrect. A fixed effect model may be more appropriate.
- Based on these counter intuitive coefficients, we think a normal linear model will not work here. Rather we need a state-level and time-level fixed effects model

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?

```
## convert data frame to pdata.frame
pfatalities <- pdata.frame(final_data, index=c("state", "year"))
pdim(pfatalities)
```

```
## Balanced Panel: n = 48, T = 25, N = 1200
```

```
fixed_model  <- plm(log(total_fatalities_rate) ~ as.factor(blood_alcohol_levels) + per_se_law +  as.fac
                      + sl70plus + grad_driver_license + percent_popul_14_to_24
                      + log(unemployement_rate) + log(vehicmilespc) + year
                    , data = pfatalities, index = c("state", "year"), effect = "individual", model = "w
summary(fixed_model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(total_fatalities_rate) ~ as.factor(blood_alcohol_levels) +
##     per_se_law + as.factor(seatbelt) + sl70plus + grad_driver_license +
##     percent_popul_14_to_24 + log(unemployement_rate) + log(vehicmilespc) +
##     year, data = pfatalities, effect = "individual", model = "within",
##     index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.3802381 -0.0517304  0.0043269  0.0542630  0.2914374
##
## Coefficients:
##                                        Estimate Std. Error  t-value  Pr(>|t|)
## as.factor(blood_alcohol_levels)BAC08 -0.0047132  0.0106922  -0.4408  0.659434
## as.factor(blood_alcohol_levels)other  0.0122743  0.0113559   1.0809  0.279988
## per_se_law                           -0.0548790  0.0098408  -5.5767 3.072e-08
## as.factor(seatbelt)1                 -0.0407267  0.0149882  -2.7172  0.006685
## as.factor(seatbelt)2                  0.0057103  0.0109996   0.5191  0.603770
## sl70plus                              0.0727124  0.0113691   6.3956 2.345e-10
## grad_driver_license                  -0.0227631  0.0121942  -1.8667  0.062202
## percent_popul_14_to_24                0.0197386  0.0041592   4.7458 2.346e-06
## log(unemployement_rate)              -0.1921317  0.0171783 -11.1846 < 2.2e-16
## log(vehicmilespc)                     0.6773776  0.0508010  13.3339 < 2.2e-16
## year1981                             -0.0631965  0.0180574  -3.4998  0.000484
## year1982                             -0.1349548  0.0189770  -7.1115 2.045e-12
## year1983                             -0.1677706  0.0197471  -8.4960 < 2.2e-16
## year1984                             -0.2065477  0.0205847 -10.0340 < 2.2e-16
## year1985                             -0.2317961  0.0215368 -10.7628 < 2.2e-16
## year1986                             -0.1950506  0.0230823  -8.4502 < 2.2e-16
## year1987                             -0.2410092  0.0250684  -9.6141 < 2.2e-16
## year1988                             -0.2715432  0.0274335  -9.8982 < 2.2e-16
## year1989                             -0.3458468  0.0292566 -11.8212 < 2.2e-16
## year1990                             -0.3557098  0.0304076 -11.6981 < 2.2e-16
## year1991                             -0.3926681  0.0311000 -12.6260 < 2.2e-16
## year1992                             -0.4530067  0.0321542 -14.0886 < 2.2e-16
## year1993                             -0.4709716  0.0327550 -14.3786 < 2.2e-16
## year1994                             -0.5029506  0.0336870 -14.9301 < 2.2e-16
```

```
## year1995                              -0.5032369  0.0347122 -14.4974 < 2.2e-16
## year1996                              -0.5582722  0.0369069 -15.1265 < 2.2e-16
## year1997                              -0.5789513  0.0378291 -15.3044 < 2.2e-16
## year1998                              -0.6318512  0.0386767 -16.3367 < 2.2e-16
## year1999                              -0.6486873  0.0392334 -16.5340 < 2.2e-16
## year2000                              -0.6809932  0.0398211 -17.1013 < 2.2e-16
## year2001                              -0.6499253  0.0400917 -16.2110 < 2.2e-16
## year2002                              -0.6117420  0.0403066 -15.1772 < 2.2e-16
## year2003                              -0.6146403  0.0404704 -15.1874 < 2.2e-16
## year2004                              -0.6518494  0.0415766 -15.6783 < 2.2e-16
##
## as.factor(blood_alcohol_levels)BAC08
## as.factor(blood_alcohol_levels)other
## per_se_law                            ***
## as.factor(seatbelt)1                  **
## as.factor(seatbelt)2
## sl70plus                              ***
## grad_driver_license                   .
## percent_popul_14_to_24                ***
## log(unemployement_rate)               ***
## log(vehicmilespc)                     ***
## year1981                              ***
## year1982                              ***
## year1983                              ***
## year1984                              ***
## year1985                              ***
## year1986                              ***
## year1987                              ***
## year1988                              ***
## year1989                              ***
## year1990                              ***
## year1991                              ***
## year1992                              ***
## year1993                              ***
## year1994                              ***
## year1995                              ***
## year1996                              ***
## year1997                              ***
## year1998                              ***
## year1999                              ***
## year2000                              ***
## year2001                              ***
## year2002                              ***
## year2003                              ***
## year2004                              ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    31.924
## Residual Sum of Squares: 8.6746
## R-Squared:      0.72828
## Adj. R-Squared: 0.70859
## F-statistic: 88.131 on 34 and 1118 DF, p-value: < 2.22e-16
```

- The absolute value of blood alcohol level shrunk, and it is still statistically insignificant. This suggest having more strict blood alcohol level (less than 0.1%) has no effect on fatality rate.
- The coefficient of per se law becomes more negative and more statistically significant. This suggest after controlling for state difference per se law actually have a significant negative effect on fatality rate.
- The primary seat belt law went from having no statistical significant effect to having a statistically significant negative effect on the fatality after controlling for state level difference.
- We trust the fixed effect model more than the linear model because we know from both EDA and earlier model's diagnostic graph that state do perform differently. Therefore we should control for their difference.

5.0.1 model assumptions

For linear models have the following assumption: 1- **Linearity**: the model is linear in parameters

2- **i.i.d.** : The observations are independent across individuals but not necessarily across time. This is guaranteed by random sampling of individuals.

3- **Indentifiability**: the regressors, including a constant, are not perfectly collinear, and all regressors (but the constant) have non-zero variance and not too many extreme values.

4- x_it is uncorrelated with idiosyncratic error term u_{it} and individual-specific effect γ_i

a)

$$E(u_{it}x_{it}) = 0$$

b)

$$E(x_{it}, \gamma_i) = 0$$

While first 3 assumptions are reasonable, the 4th assumption is violated, because we know individual (state) is correlated with x_it, which are regulations in our case.

For fixed effect model, the first three assumptions are the same, but 4th assumption, see below is different.

4- **Zero conditional means (strict exogeneity)**

$E(x_{it}, u_{is}) = 0$ for $s = 1, 2, 3, ..., T$

If the above assumptions holds, we can use the Fixed Effects (FE) estimators to obtain consistent estimates of β. However we cannot be certain that feedback from past u_{is} has no effect on the current x_{it}. One implication of this is that estimators will not yield consistent estimates if x_{it} depends on lagged dependent variables $(y_{it-1}, y_{it-2}, ...)$ as in the case of a VAR model.

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?
- The following are the assumptions of a random effects model according to "Introductory Econometrics" by Wooldridge:

1) There are no perfect linear relationships among the explanatory variables.
2) For each t, the expected value of the idiosyncratic error given the explanatory variables in all time periods and the unobserved effect is zero: $E(u_{it}|X_i, a_i) = 0$.
3) The expected value of a_i, given all explanatory variables, is constant: $E(a_i|X_i) = \beta_0$.
4) $Var(u_{it}|X_i, a_i) = Var(u_{it}) = \sigma_u^2$ for all $t = 1, ..., T$.
5) The variance of a_i, given all explanatory variables, is constant: $Var(a_i|X_i) = \sigma_a^2$.
6) For all $t \neq s$, the idiosyncratic errors are uncorrelated (conditional on all explanatory variables and a_i): $Cov(u_{it}, u_{is}|X_i, a_i) = 0$.

Assumption 3 is likely not satisfied, as there are unobserved time-constant effects correlated with the explanatory variables. A few examples of the violation of this assumption are: 1) Oil producing states like Texas have lower gas prices, which are likely correlated to more miles driven per capita. 2) Whether a state has more rural or urban areas might affect appropriate speed limit laws. For example, a 70 m.p.h. limit in a highly urban state might be too lax, while it might be acceptable for states with more rural roads and highways. 3) Public perception of drinking in a state might be correlated with the implementation of DUI laws: states where the majority of the population view drinking favorably might foster the implementation of DUI laws.

The violation of the assumption of unobserved effects being uncorrelated explanatory variables leads to the widely known omitted variable problem, which results in biased coefficient estimates. The direction of correlation between the unobserved effects and the explanatory variables will result in either underestimating or overestimating the coefficients of the dependent variables. The tradeoff with respect to the fixed-effect model is efficiency, producing smaller standard errors.

```
random_effects_model  <- plm(log(total_fatalities_rate) ~ as.factor(blood_alcohol_levels) + per_se_law +
                        + sl70plus + grad_driver_license + percent_popul_14_to_24
                        + log(unemployement_rate) + log(vehicmilespc) + year
                    , data = pfatalities, index = c("state", "year"), effect = "individual", model = "ra
```

Below we run a Hausman test to verify analytically whether the random effects model is applicable to the data.

```
phtest(fixed_model, random_effects_model)
```

```
##
##  Hausman Test
##
## data:  log(total_fatalities_rate) ~ as.factor(blood_alcohol_levels) +  ...
## chisq = 77.8, df = 34, p-value = 2.774e-05
## alternative hypothesis: one model is inconsistent
```

Using the Hausman test, we can reject the null hypothesis that the random effects model is appropriate, and thus we should use the fixed effects model instead.

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.

We have obtained vehicle miles driven per month from January 2018 to May 2022, from the U.S. Federal Highway Administration [https://fred.stlouisfed.org/series/TRFVOLUSM227NFWA].

```
miles_driven_2018_2022 = read.csv("./data/TRFVOLUSM227NFWA.csv")

miles_driven_2018_2022 = miles_driven_2018_2022 %>%
  mutate(year = year(DATE),
         month = month(DATE))

miles_driven_2018 = miles_driven_2018_2022[miles_driven_2018_2022$year == 2018,]
miles_driven_2020 = miles_driven_2018_2022[miles_driven_2018_2022$year == 2020,]
miles_driven_2021 = miles_driven_2018_2022[miles_driven_2018_2022$year == 2021,]

miles_driven_diff = data.frame(month = miles_driven_2018$month,
                               miles_2018 = miles_driven_2018$TRFVOLUSM227NFWA,
                               miles_2020 = miles_driven_2020$TRFVOLUSM227NFWA,
                               miles_2021 = miles_driven_2021$TRFVOLUSM227NFWA)

miles_driven_diff = miles_driven_diff %>%
  mutate(diff_2020 = miles_2020 - miles_2018,
         diff_2021 = miles_2021 - miles_2018)

miles_driven_diff
```

```
##    month miles_2018 miles_2020 miles_2021 diff_2020 diff_2021
## 1      1     244736     260847     231030     16111    -13706
## 2      2     227759     242695     213038     14936    -14721
## 3      3     270705     226638     269426    -44067     -1279
## 4      4     275127     167617     259189   -107510    -15938
## 5      5     283713     221006     284326    -62707       613
## 6      6     282648     250330     286898    -32318      4250
## 7      7     290989     265550     296458    -25439      5469
## 8      8     284989     265060     287409    -19929      2420
## 9      9     267434     257531     277998     -9903     10564
## 10    10     281382     266596     285755    -14786      4373
## 11    11     260473     238300     267749    -22173      7276
## 12    12     270370     241451     268420    -28919     -1950
```

We compare the difference in driven miles between 2018 and March 2020 to December 2020, and 2018 with 2021, which corresponds to the strongest years of the pandemic (2020-2021). We exclude the months of January and February 2020, as the official declaration of Covid-19 as a pandemic was announced by the WHO in March 11th, 2020 (https://pubmed.ncbi.nlm.nih.gov/32191675/#:~:text=The%20World%20Health%20Organization%20(WHO,a%20global%20pandemic%20(1).)

The largest decrease in driving during the pandemic occurred in April 2020, where there was a decrease of 107,510 miles in comparison with April 2018. This constitutes a decrease of 39.07% ($(167617 - 275127)/275127 = -0.3907$).

The largest increase in driving during the pandemic occurred in September 2021, where there was an increase of 10,564 miles in comparison with September 2018. This constitutes an increase of 3.8% ((277998 –

267434)/267434 = 0.0395).

Recall that the coefficient of the log of the vehicle miles driven per capita in our fixed effect model is $\beta = 0.677$. Since both the dependent and independent variable are log transformed, all else being equal, a one percent increase in miles driven corresponds to a $100 * (1.01^{0.677} - 1) = 0.68$ percent increase in fatalities.

Therefore, during the boom in driving in Covid, in September 2021, the 3.95% increase in miles driven would have resulted in a $3.95 * 0.68 = 2.69\%$ increase in fatalities, all else being equal, with respect to the number of fatalities in the baseline of September 2018.

In contrast, during the bust in driving during Covid, in April 2020, the 39.07% decrease in miles driven would have resulted in a $39.07 * 0.68 = 26.56\%$ decrease in fatalities, all else being equal, with respect to the number of fatalities in the baseline of April 2018. This is a considerable effect in terms of saving lives, but unfortunately at the large cost of the millions of lives lost due to the pandemic itself.

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?

Heteroskedasticity in the errors violates one of the assumptions of the OLS regression. In the presence of heteroskedasticity the estimates will be less precise and the standard errors are smaller than they should be, which might result in spurious statistical significance of coefficients.

Serial correlation in idiosyncratic errors will also bias the standard errors and will produce less efficient estimates.

We perform a Breusch-Pagan test to determine if there's heteroskedasticity in the Fixed Effects model.

```
pcdtest(fixed_model, test = "lm")
```

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  log(total_fatalities_rate) ~ as.factor(blood_alcohol_levels) +     per_se_law + as.factor(sea
## chisq = 2732.8, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

We reject the null hypothesis of homoskedasticity in the fixed effects model.

We now perform a Durbin-Watson test to determine if there's serial correlation in the errors of the fixed effects Model.

```
pdwtest(fixed_model)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  log(total_fatalities_rate) ~ as.factor(blood_alcohol_levels) +  ...
## DW = 1.2288, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

The test rejects the null hypothesis of no serial correlation in the idiosyncratic errors.