Lab 3: Panel Models

US Traffic Fatalities: 1980 - 2004

Contents


```
library(tidyr)
library(dplyr)
library(ggrepel)
library(ggthemes)
library(stargazer)
library(gridExtra)
library(plm)
library(knitr)
library(patchwork)
library(lubridate)
library(tsibble)
```

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.

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

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

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.)

```
year_of_observation = as.matrix(
  data[match("d80", colnames(data)):match("d04", colnames(data))]
  ) %*% c(1980:2004)

data_clean <- data %>% mutate(
  year_of_observation = factor(as.numeric(year_of_observation)),
  year = factor(year),
  state = factor(state),

  # laws
  speed_limit = factor(
    round(sl55) * 55 + round(sl65) * 65 + round(sl70) * 70 +
      round(sl75) * 75 + round(slnone) * 100
  ), # assuming 100 for no speed limit
  blood_alcohol_limit = factor(
    round(bac10) * 1 + round(bac08) * 2,
    labels = c('none', 'bac10', 'bac08')
  ),
  zero_tolerance_law = factor(round(zerotol)),
  per_se_law = factor(round(perse)),
```

```
  graduated_drivers_license_law = factor(round(gdl)),
  seat_belt = factor(seatbelt,labels=c('none','primary', 'secondary')),
  min_age = as.factor(minage),
  speed_limit_70plus = factor(round(sl70plus)),

  # demographics
  log_umeployment_rate = log(unem),
  log_vehicle_miles_per_capita = log(vehicmiles / statepop),

  # dependent variable
  log_total_fatalities_rate = log(totfatrte)
    ) %>%
  rename(
    total_fatalities_rate = totfatrte
    ) %>%
  dplyr::select(
      year,
      state,
      total_fatalities_rate,
      log_total_fatalities_rate,
      seat_belt,
      speed_limit,
      blood_alcohol_limit,
      log_vehicle_miles_per_capita,
      speed_limit_70plus,
      log_umeployment_rate,
      perc14_24,
      zero_tolerance_law,
      graduated_drivers_license_law,
      per_se_law
    )
```

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 comes from a study conducted by our textbook author. In the study, Wooldridge et al. invistigate the effects of Blood Alcohol Content laws on driving safety in the United States. The data contains observations of each state from 1980 to 2004 which means it includes both longitudinal and cross sectional aspects. Thus, the panel data is the culmination of both discrete and time series statistics and requires panel data analysis. The data was taken during a natural experiment, where the split into treatment and control groups came from state legislatures enacting laws at different times.

Is it census? how where when was it collected?

citation: Wooldridge Source: Freeman, D.G. (2007), "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws," Contemporary Economic Policy 25, 293–308. Professor Freeman kindly provided the data. Data loads lazily.

```
<!-- total traffic fatalities divided by state population -->
```

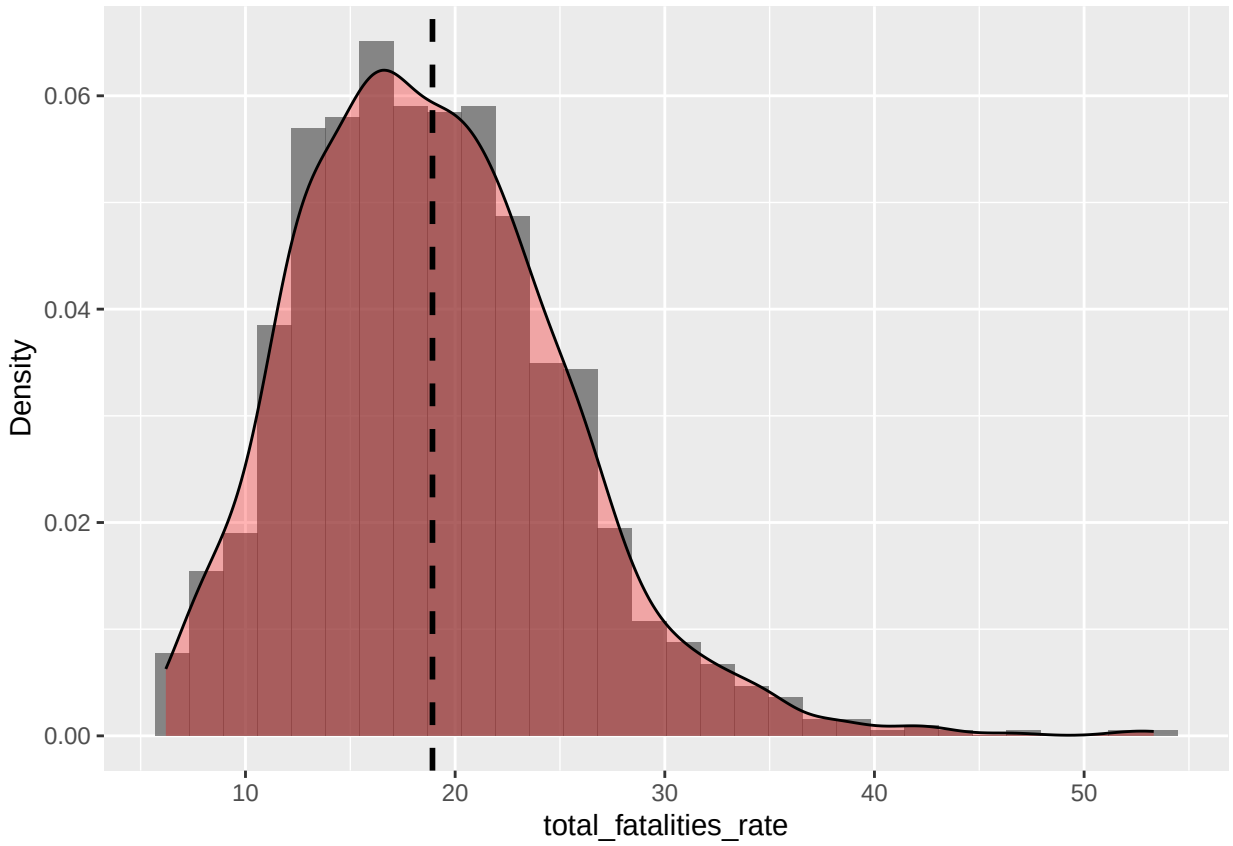
3. (20 points) Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable `total_fatalities_rate` and the potential explanatory variables. Minimally, this should include:
    - How is the our dependent variable of interest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

```
# histograms
data %>% mutate(log_total_fatalities_rate = log(totfatrte),
                total_fatalities_rate = totfatrte) %>%
ggplot(aes(x=total_fatalities_rate)) + geom_histogram(aes(y=..density..), alpha=0.7) + geom_density(alp
    labs(x="total_fatalities_rate", y = "Density")
```

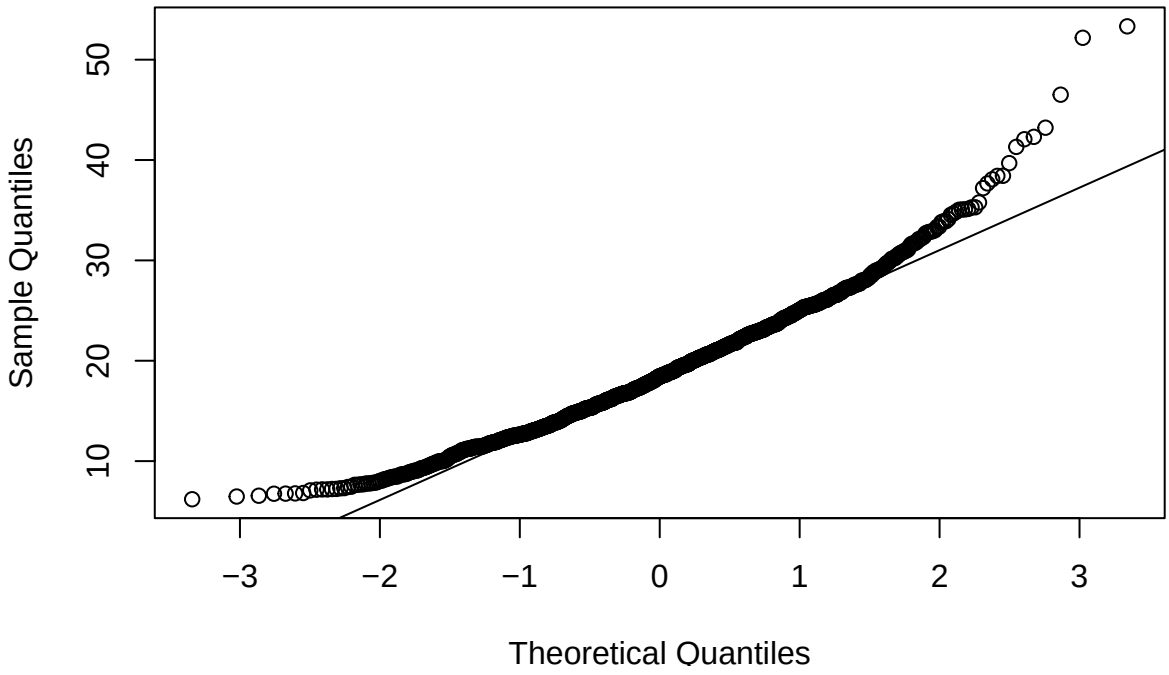
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```




```
# consider qq plot
qqnorm(data$totfatrte, main='Not Normal')
qqline(data$totfatrte)
```




```
# null hypothesis is data is normally distributed
shapiro.test(data$totfatrte) # reject null, data not normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$totfatrte
## W = 0.96832, p-value = 1.572e-15
```

```
# null hypothesis is normally distributed
ks.test(data$totfatrte, 'pnorm') # reject null, data not normal
```

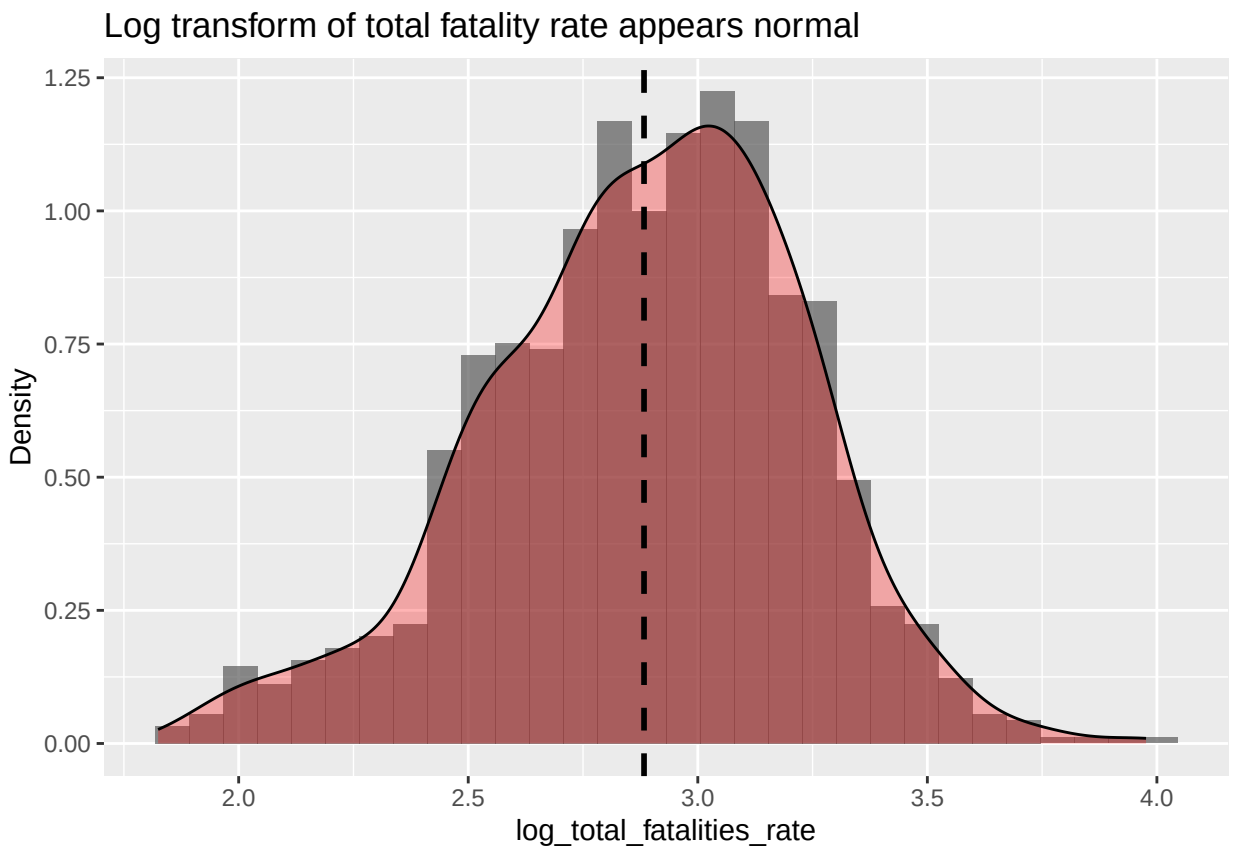
```
## Warning in ks.test.default(data$totfatrte, "pnorm"): ties should not be present
## for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  data$totfatrte
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

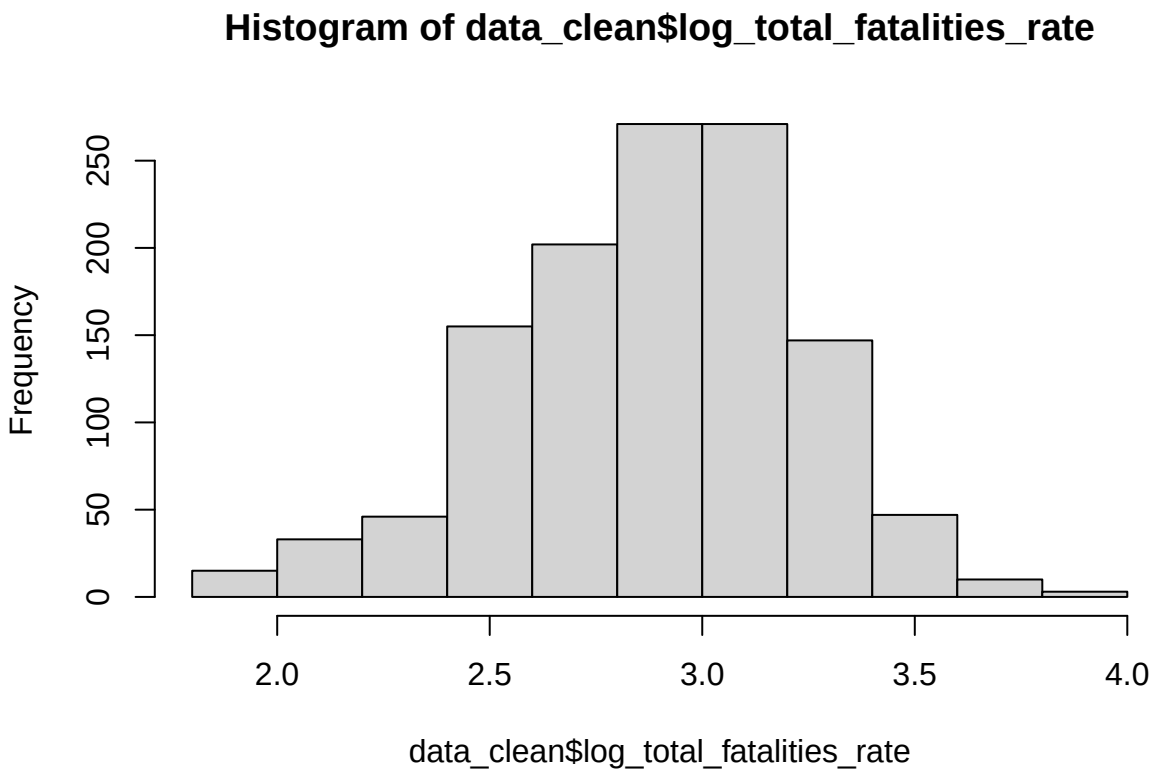
We start by examining the dependent variable - total fatality rate. The total fatality rate represents the number of deaths from traffic accidents per 1000 (FIX ME). The histogram of values suggests that the data is not normally distributed because of the long right tail. We can confirm the lack of normality using the Shapiro-Wilk test. The Shapiro-Wilk normality test has the null hypothesis that the data is normally distributed. With a p-value of 1.5e-15, we reject the null hypothesis and conclude that the data needs transformation to improve the linearity between the independent and dependent variables.

```r
# test log transform
data_clean %>%
  ggplot(aes(x=log_total_fatalities_rate)) + geom_histogram(aes(y=..density..), alpha=0.7) + geom_densi
    labs(x="log_total_fatalities_rate", y = "Density", title='Log transform of total fatality rate appea
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
hist(data_clean$log_total_fatalities_rate)
```




```
shapiro.test(data_clean$log_total_fatalities_rate)
```

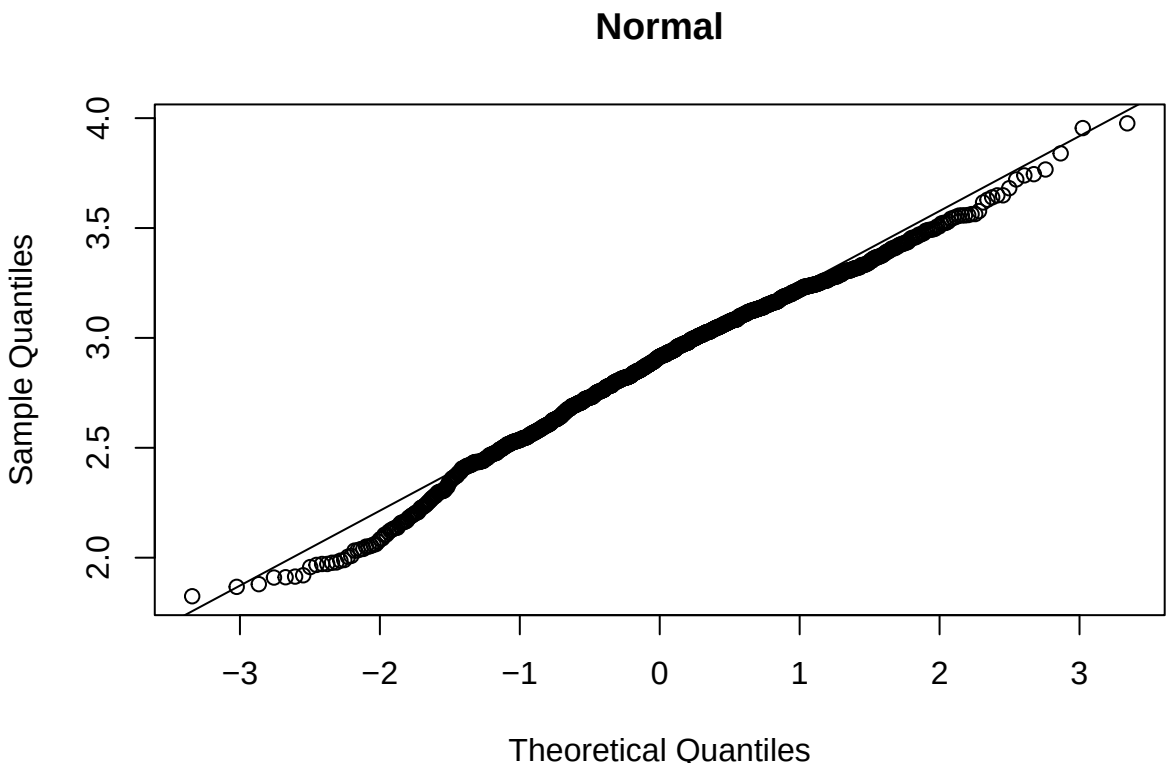
```
##
##  Shapiro-Wilk normality test
##
## data:  data_clean$log_total_fatalities_rate
## W = 0.99093, p-value = 9.448e-07
```

```
ks.test(data_clean$log_total_fatalities_rate, 'pnorm')
```

```
## Warning in ks.test.default(data_clean$log_total_fatalities_rate, "pnorm"): ties
## should not be present for the Kolmogorov-Smirnov test
```

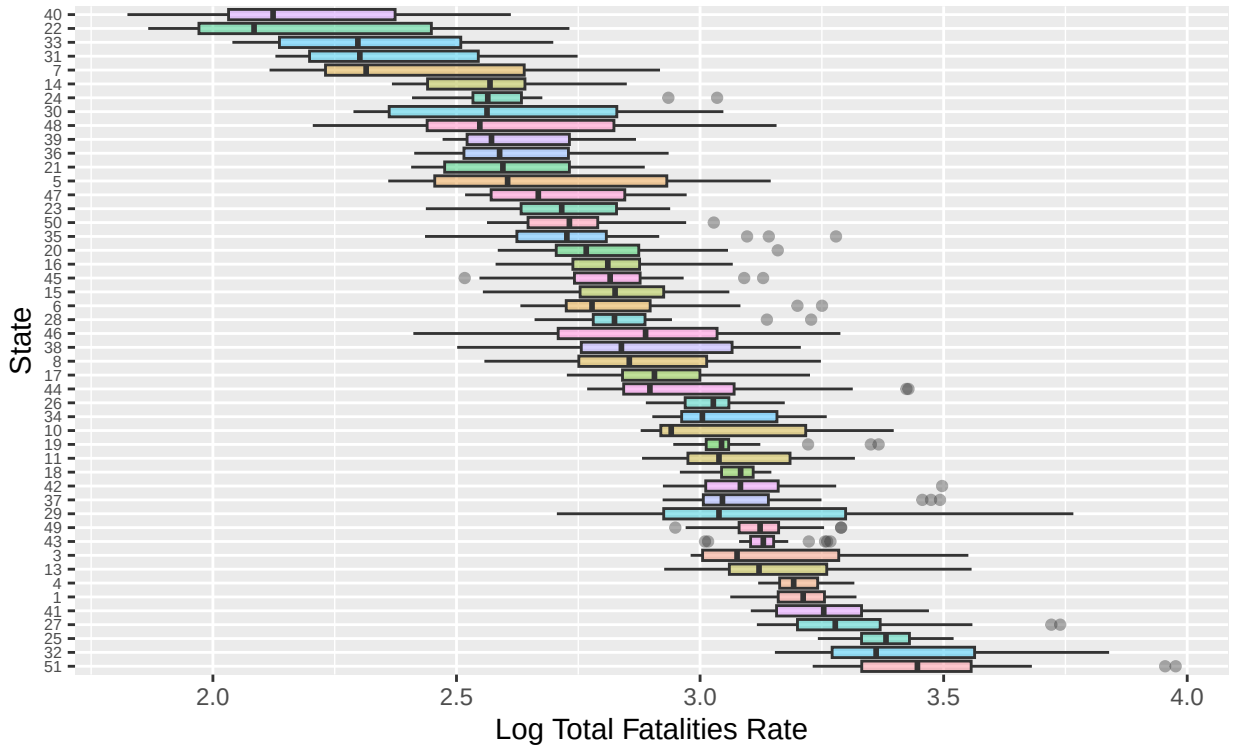
```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  data_clean$log_total_fatalities_rate
## D = 0.9694, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
qqnorm(data_clean$log_total_fatalities_rate, main='Normal')
qqline(data_clean$log_total_fatalities_rate)
```

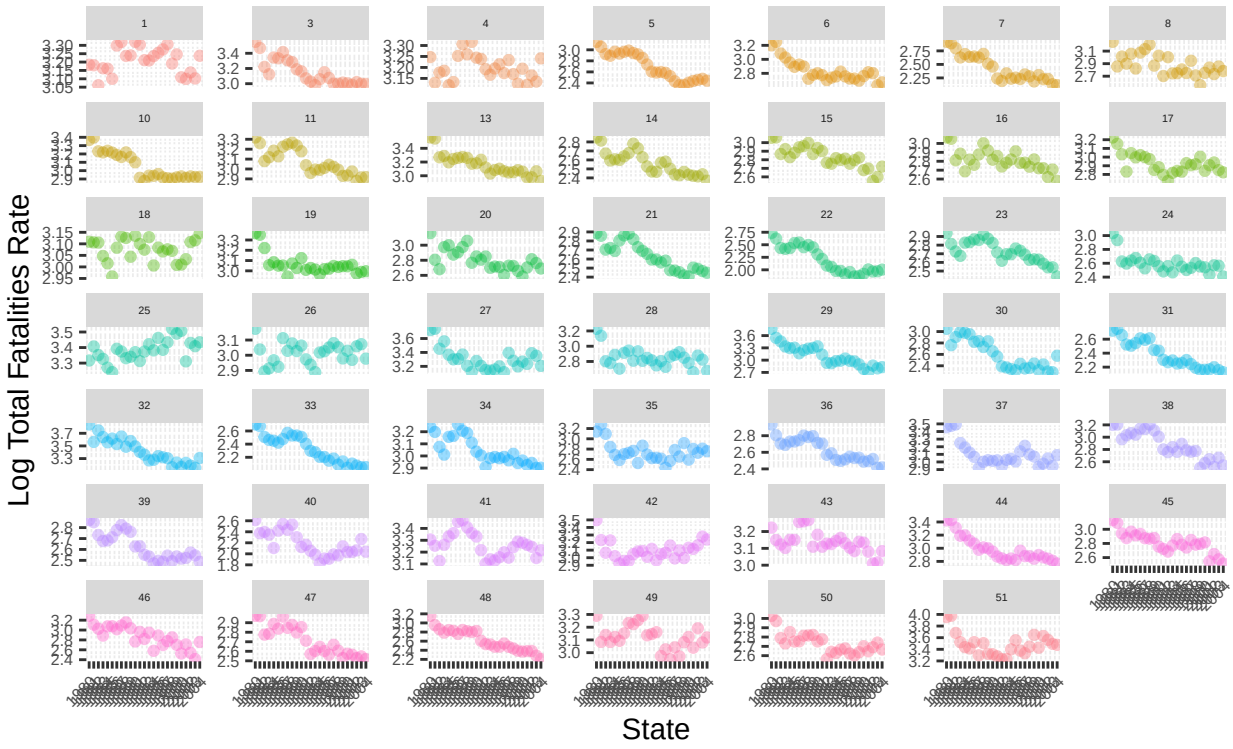


After transformation, the dependent variable is now `log_total_fatality_rate` and it appears normally distributed.

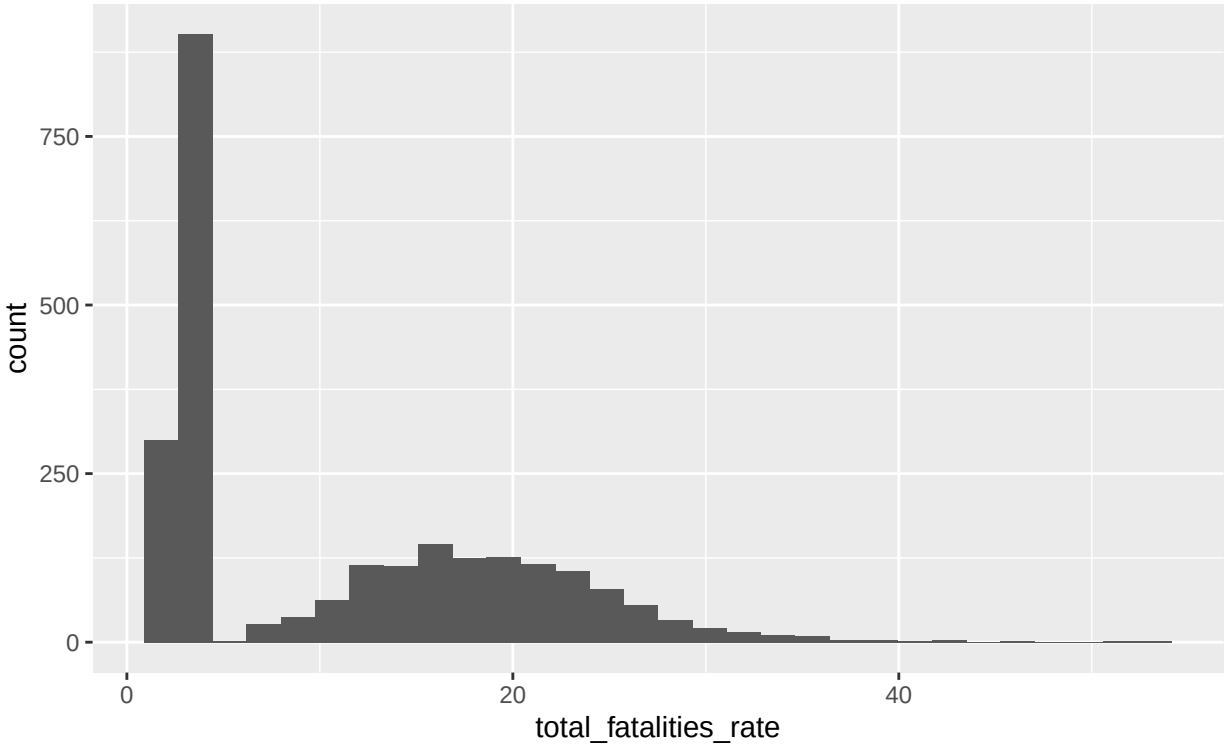
```
# boxplot over time
data_clean %>% ggplot(aes(reorder(state, desc(
  log_total_fatalities_rate
)), log_total_fatalities_rate,
fill = state)) +
  geom_boxplot(alpha = 0.4) +
  # theme_economist_white(gray_bg=F) +
  theme(legend.position = "none", axis.text.y = element_text(size = 6)) +
  scale_y_continuous() +
  xlab("State") +
  ylab("Log Total Fatalities Rate") +
  coord_flip()
```




```r
# lineplot over time
data_clean %>%
  ggplot(aes(year, log_total_fatalities_rate, color = state)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  facet_wrap( ~ state, scales = "free_y") +
  theme(
    legend.position = "none",
    axis.text.x = element_text(
      angle = 45,
      hjust = 1,
      vjust = 1,
      size = 6
    ),
    axis.text.y = element_text(size = 6)
  ) +
  theme(strip.text = element_text(size = 4)) +
  scale_y_continuous() +
  xlab("State") +
  ylab("Log Total Fatalities Rate")
```




```
data %>% mutate(log_total_fatalities_rate = log(totfatrte),
                total_fatalities_rate = totfatrte) %>%
  ggplot() + geom_histogram(aes(total_fatalities_rate)) +
  geom_histogram(aes(log_total_fatalities_rate))
```



```
plot_multi_histogram <- function(df, feature, label_column) {
    plt <- ggplot(df, aes(x=eval(parse(text=feature)), fill=eval(parse(text=label_column)))) +
    geom_histogram(alpha=0.7, position="identity", aes(y = ..density..), color="black") +
    geom_density(alpha=0.7) +
    geom_vline(aes(xintercept=mean(eval(parse(text=feature)))), color="black", linetype="dashed", size=
    labs(x=feature, y = "Density")
    plt + guides(fill=guide_legend(title=label_column))
}

# data %>% mutate(log_total_fatalities_rate = log(totfatrte),
#                 total_fatalities_rate = totfatrte) %>%
# plot_multi_histogram()
```

> EDA shows that the driver fatality rate has been trending down but is different on a state by state basis. figure 1 shows States have varying degrees of both median and spread of fatality rates over time. figure 2 shows that fatality rate over time has decreased for almost every state

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.

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? 1. explain log model (a one unit increase in x becomes a 1 degree increase in y). changing from 1980 t 2. only includes the change in years, not cross sectional features - 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?

> In this section, we pool the state effects together and investigate total fatality rates over time.

> Linear models are a sensible place to start because they are simple and easy to interpret. In our case, this linear model can show how much the average total fatalities rate changed compared with the year 1980.

```
preliminary_model  <- lm(log_total_fatalities_rate ~ year, data = data_clean)
summary(preliminary_model)
```

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

```
## -0.96324 -0.22134  0.01005  0.23221  0.86830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.19577    0.04697  68.035  < 2e-16 ***
## year1981    -0.07878    0.06643  -1.186 0.235904
## year1982    -0.19957    0.06643  -3.004 0.002719 **
## year1983    -0.23523    0.06643  -3.541 0.000414 ***
## year1984    -0.22585    0.06643  -3.400 0.000697 ***
## year1985    -0.24301    0.06643  -3.658 0.000265 ***
## year1986    -0.19681    0.06643  -2.963 0.003111 **
## year1987    -0.19871    0.06643  -2.991 0.002836 **
## year1988    -0.18885    0.06643  -2.843 0.004547 **
## year1989    -0.24815    0.06643  -3.735 0.000196 ***
## year1990    -0.26785    0.06643  -4.032 5.89e-05 ***
## year1991    -0.34372    0.06643  -5.174 2.69e-07 ***
## year1992    -0.40229    0.06643  -6.056 1.88e-09 ***
## year1993    -0.40257    0.06643  -6.060 1.83e-09 ***
## year1994    -0.40798    0.06643  -6.142 1.12e-09 ***
## year1995    -0.38492    0.06643  -5.794 8.79e-09 ***
## year1996    -0.39949    0.06643  -6.014 2.42e-09 ***
## year1997    -0.38596    0.06643  -5.810 8.03e-09 ***
## year1998    -0.40954    0.06643  -6.165 9.67e-10 ***
## year1999    -0.41450    0.06643  -6.240 6.11e-10 ***
## year2000    -0.43694    0.06643  -6.578 7.18e-11 ***
## year2001    -0.43521    0.06643  -6.552 8.50e-11 ***
## year2002    -0.42672    0.06643  -6.424 1.93e-10 ***
## year2003    -0.43978    0.06643  -6.620 5.44e-11 ***
## year2004    -0.44853    0.06643  -6.752 2.29e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3254 on 1175 degrees of freedom
## Multiple R-squared:  0.126,  Adjusted R-squared:  0.1081
## F-statistic: 7.057 on 24 and 1175 DF,  p-value: < 2.2e-16
```

This model describes the log of total fatality rate conditional on the year the driving occured. Since each year is a binary variable, year 1980 represents the intercept term in the model and years 1981 through 2004 are represented with an indicator variable. The coefficients on all non intercept terms represent the change in the log fatality rate compared to the 1980 baseline.

$$log(total.fatality.rate_i) = \beta_{1980} + \Sigma_{i=1981}^{2004} \beta_i * I(year_i)$$

To achieve the total fatality rate, we require the inverse transform for the log function.

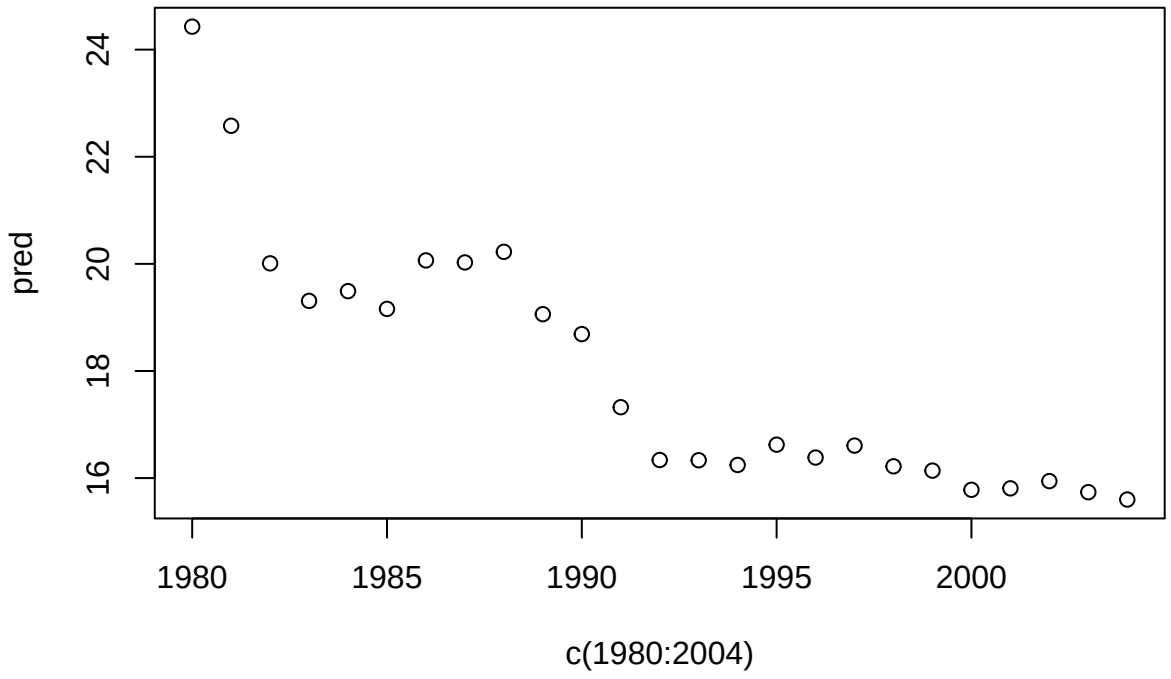
$$total.fatality.rate_i = exp(\beta_{1980} + \Sigma_{i=1981}^{2004} \beta_i * I(year_i))$$

Initially, our model estimates that the total fatality rate of 1980 is 24.41. With a p-value near 0, this estimate is statistically significant. The additional beta's of the model are the estimatations of each year's impact on the fatality rate from the baseline of 1980. The coefficients over time are negative and decreasing, meaning that the model suggests driving got safer over the

period because the fatality rate decreased. Almost all of these estimates, save for 1981 only, are statistically significant.

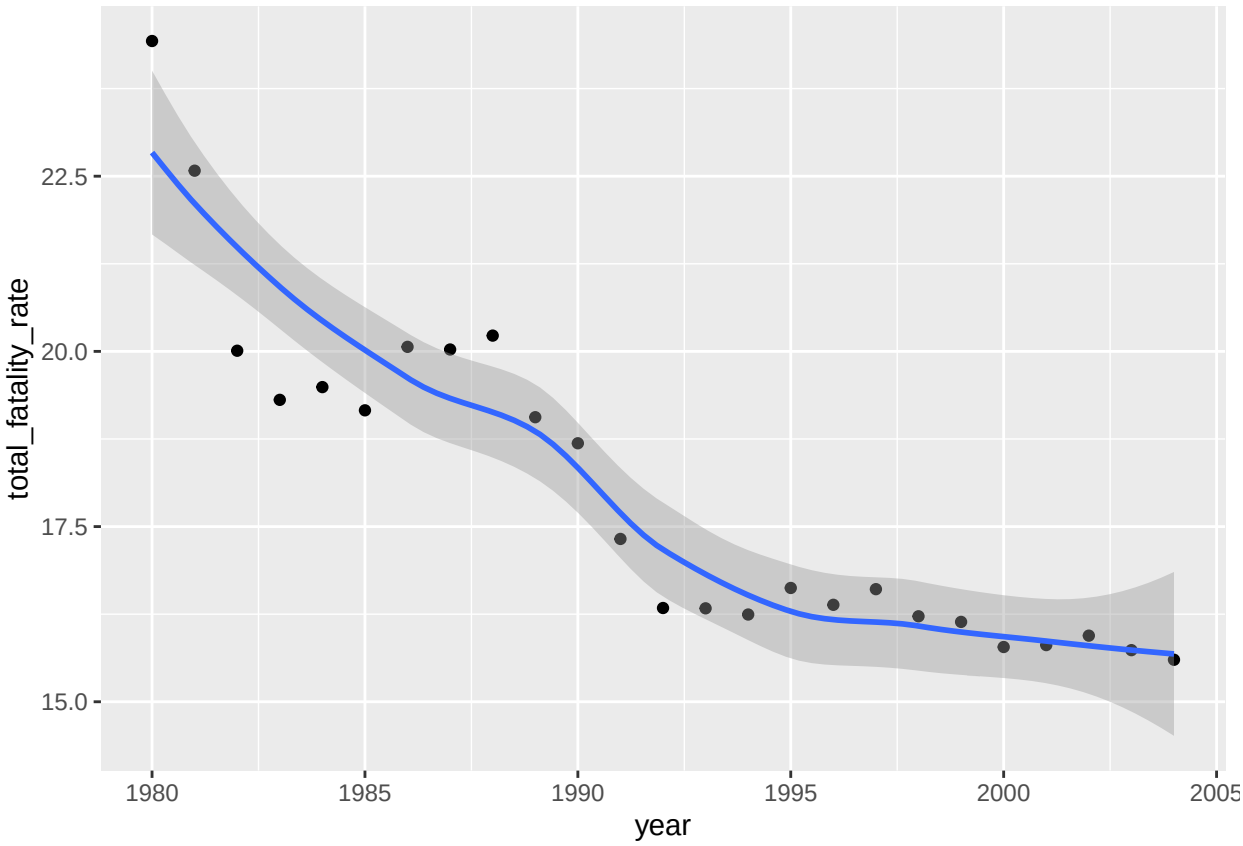
```
#model estimated fatality rate over time
new_data <- data_clean %>% filter(state==1)
pred <- predict(preliminary_model,newdata = new_data)
pred <- exp(pred)

plot(c(1980:2004), pred)
```




```
data.frame(year=c(1980:2004), total_fatality_rate=pred) %>% ggplot() +
  geom_point(aes(x=year, y=total_fatality_rate)) + geom_smooth(aes(x=year, y=total_fatality_rate))
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



An important feature of the dataset is that it has both longitudinal (recorded fatality rates for years 1980-2004) and cross sectional (the state at which the driving occured) features which are hallmarks of panel data. In this way, each state has multiple records in the dataset corresponding to different years, such that Alabama has 24 repeated observations in the dataset, one for each year. This feature introduces omitted variable bias because the states likely have effects on fatality rates that we do not see but will violate our model assummptions based on repeated sampling. Specifically, this linear model regressing year on log total fatality rate violates the IID assumption. For example the state variable represents the location in the US that the drivers are in. the state is repeated across years which means that each data point is not idependent and identically distributed. We take the results of this model with a grain of salt because linear models are not reliable with panel data since they violate iid.

pooled data (not independent) has to do with the way the data is strucutured not the sampling variability

Need to comment on consistency, bias, reliability of estimates

4 (15 points) Expanded Model

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

consider when laws were passed, if eda shows groupings of states, years, and laws, and consider alpha corrections, like bonferoni do stargazer on simple model, full model, best fit model using anova discuss significance, and parameter values refactor bac into one variable, with 3 factors 0, 1, 2, with 0 as no law, 1 as bac08 law, 2 as bac10 law

- 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.

Concept	Variable in raw data	Round	Log transformation	Description
Blood alcohol levels	bac08, bac10	yes		Combine bac08 and bac10 into a ordinal variabe: None=no limit; bac10=limit at 0.1; bac08=limit at 0.08
Per se laws	perse	yes		0=no per se law, 1=per se law
Primary seat belt laws	seatblet	yes		0=no seat belt law; 1=primary, 2=secondary
Secondary seat belt laws	as above	yes		as above
Speed limits faster than 70	sl70plus	yes		0=speed limit <70; 1=speed limit >=70 or no limit
Graduated drivers licenses	gdl	yes		0=no graduated drivers license law
Percent of the population between 14 and 24	perc14_24			it's a continous variable with no obvious skewness, so no transformation needed
Unemployment rate	unem		yes	it's a continuous variable and obviously right skewed, so log transformation is preferred
Vehicle miles driven per capita	vehicmiles, statepop		yes	we use the ratio of these two variables. Since the ratio is right-skewed, we perform log transformation

(MAYBE CUT THIS) Blood alcohol variables describe if the legal blood alcohol content (BAC) is for a driver. If the Driver's BAC is over the legal limit, then they are illegally driving while intoxicated. There are two columns in the driving dataset, bac08 and bac10. The bac08 variable describes if the state had a 0.08 BAC limit that year, with 0 being no law and 1 meaning having a law. If the law went into effect for a partial year, then the value is factional for the amount of time it was in effect (a law starting in June would have 6/12 months = 0.5 value). The feature bac10 has the same properties - the value is the fractional amount of time the law was in effect that year (0 not in effect, 1 in effect). We engineer these features using rounding such that our new feature shows if the BAC law was effective for a majority of the year.

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. > BAC laws put a limit on alcohol consumption for drivers where BAC is measured in a percentage of alcohol in the blood. Before Congress enacted a national BAC limit in the year 2000, each state was able to enact their own laws with varying legal limits on BAC. Our data set identifies

these laws as a state having no BAC limit, a BAC limit of .08, and a BAC limit of .10. We label the state based on the law that was in place for the majority of the year.

The effect of having no BAC law was added to the intercept term, which also includes the effects of the year 1980, no seatbelt laws, and more. The estimate for having a BAC08 law was -0.062 with a p-value of 0.01. That means that states enacting a BAC law limit of 0.08 see a decrease in fatality rate of 1.0639623 all else being equal. States that enact a more lenient level of 0.10 BAC see less of an effect. Those states see a decrease in fatality rates of 1.0171453, but with a pvalue of 0.34 the effect is not statistically significant.

- Do *per se laws* have a negative effect on the fatality rate?

  Per se laws strive to improve driver safety by allowing judges to suspend the license of an arrest drunk driver before they are convicted of a crime. Unfortunately, the pooled model does not support the effect. With a coefficient -0.01 and a p-value 0.19, there is not enough evidence to support per se laws significantly improving safety via log fatality rates.

- Does having a primary seat belt law?

  Seat belts help reduce injuries and fatalities in car crashes. In an attempt to improve seat belt usage, states enacted primary and secondary seat belt laws. Primary seat belt laws allow law enforcement to stop drivers and give tickets to drivers not wearing a seatbelt. Secondary laws allow police performing a traffic stop for another reason, say speeding, to give out tickets for not wearing a seat belt. Secondary laws would be considered more stringent.

  The coefficients for primary and secondary seatbelt laws are 0.009 and 0.02042 respectively. Unfortunately, both p-values are much larger than the critical value at 0.96 and 0.34 respectively, meaning that there is not enough evidence to suggest that there is a significant relationship between seatbelt laws and log fatality rates, all else equal.

Answer

My way of adding variables related to the following concepts: (some of the vairables need log transformation because they are skewed)

- Blood alcohol levels: as we talked about
- Per se laws: similarly, round and factor
- 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.): actually there's no fractional number, so we can just factorize the *seatbelt* variable.
- Secondary seat belt laws: combined with the primary seat belt laws.
- Speed limits faster than 70: round and factorize the vairable *sl70plus*
- Graduated drivers licenses: round and factorize the vairable *gdl*
- Percent of the population between 14 and 24 years old: just use the variable *perc14_24*
- Unemployment rate: use the *unem* **after log transformation**
- Vehicle miles driven per capita: log(vehicmiles/statepop)

```
expanded_model = plm(
  log_total_fatalities_rate ~ year + blood_alcohol_limit + per_se_law +
    seat_belt + speed_limit_70plus + graduated_drivers_license_law + perc14_24 +
    log_umeployment_rate + log_vehicle_miles_per_capita,
  data = data_clean,index = c("state", "year"),
  model = "pooling"
)
summary(expanded_model)
```

```
## Pooling Model
##
## Call:
## plm(formula = log_total_fatalities_rate ~ year + blood_alcohol_limit +
##     per_se_law + seat_belt + speed_limit_70plus + graduated_drivers_license_law +
##     perc14_24 + log_umeployment_rate + log_vehicle_miles_per_capita,
##     data = data_clean, model = "pooling", index = c("state",
##         "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -0.58513248 -0.12787021 -0.00044488  0.14047406  0.62366622
##
## Coefficients:
##                                   Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)                     20.63563710  0.55396704  37.2507 < 2.2e-16 ***
## year1981                        -0.09191767  0.04120595  -2.2307  0.025892 *
## year1982                        -0.29560610  0.04206306  -7.0277 3.566e-12 ***
## year1983                        -0.35201682  0.04268541  -8.2468 4.352e-16 ***
## year1984                        -0.29970230  0.04346254  -6.8956 8.772e-12 ***
## year1985                        -0.33705447  0.04443389  -7.5855 6.748e-14 ***
## year1986                        -0.31327035  0.04639835  -6.7518 2.299e-11 ***
## year1987                        -0.34922535  0.04839286  -7.2165 9.595e-13 ***
## year1988                        -0.36018880  0.05093625  -7.0714 2.639e-12 ***
## year1989                        -0.44511076  0.05291498  -8.4118 < 2.2e-16 ***
## year1990                        -0.50526138  0.05406779  -9.3450 < 2.2e-16 ***
## year1991                        -0.62071784  0.05516748 -11.2515 < 2.2e-16 ***
## year1992                        -0.72660029  0.05626552 -12.9138 < 2.2e-16 ***
## year1993                        -0.71845554  0.05694197 -12.6173 < 2.2e-16 ***
## year1994                        -0.70553914  0.05801736 -12.1608 < 2.2e-16 ***
## year1995                        -0.68180924  0.05949771 -11.4594 < 2.2e-16 ***
## year1996                        -0.80797025  0.06161353 -13.1135 < 2.2e-16 ***
## year1997                        -0.81695095  0.06266096 -13.0376 < 2.2e-16 ***
## year1998                        -0.86535992  0.06380804 -13.5619 < 2.2e-16 ***
## year1999                        -0.86709969  0.06459498 -13.4236 < 2.2e-16 ***
## year2000                        -0.87936828  0.06568439 -13.3878 < 2.2e-16 ***
## year2001                        -0.93494948  0.06610085 -14.1443 < 2.2e-16 ***
## year2002                        -0.97941190  0.06653375 -14.7205 < 2.2e-16 ***
## year2003                        -1.00267287  0.06679307 -15.0116 < 2.2e-16 ***
## year2004                        -0.98389153  0.06853273 -14.3565 < 2.2e-16 ***
## blood_alcohol_limitbac10        -0.01695683  0.01798634  -0.9428  0.345998
## blood_alcohol_limitbac08        -0.06189810  0.02432588  -2.5445  0.011070 *
## per_se_law1                     -0.01882352  0.01463685  -1.2860  0.198686
## seat_beltprimary                 0.00094193  0.02456297   0.0383  0.969417
## seat_beltsecondary               0.02042971  0.02143944   0.9529  0.340837
## speed_limit_70plus1              0.22192294  0.02162505  10.2623 < 2.2e-16 ***
## graduated_drivers_license_law1 -0.02129173  0.02528907  -0.8419  0.399998
## perc14_24                        0.01779098  0.00611155   2.9110  0.003671 **
## log_umeployment_rate             0.26728200  0.02414277  11.0709 < 2.2e-16 ***
## log_vehicle_miles_per_capita     1.54069734  0.04431782  34.7647 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Total Sum of Squares:    142.38
## Residual Sum of Squares: 47.303
## R-Squared:      0.66777
## Adj. R-Squared: 0.65807
## F-statistic: 68.8702 on 34 and 1165 DF, p-value: < 2.22e-16
```

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?

5.1 Answer:

From the comparison below we can see:

1. The coefficients for *blood alcohol limit* is -0.02 for limit at 0.08 and -0.16 for limit at 0.1, both less negative than those in the pooled model. Unlike the pooled model, the blood_alcohol_limit are not significant.
2. The coefficient for *per se law* is -0.053, which becomes significant and more negative in the FE model, suggesting adopting the per se law helps reduce the total fatalities rate by nearly 5%. (note that we approximate the log change with percent change)
3. The coefficient for *primary seat belt law* is -0.04, which also becomes significant and more negative in the FE model, implying the primary seat belt law can help reduce the total fatalities rate by nearly 4%.

```
fixed_effect_model = plm(
  log_total_fatalities_rate ~ year + blood_alcohol_limit + per_se_law +
    seat_belt + speed_limit_70plus + graduated_drivers_license_law + perc14_24 +
    log_umeployment_rate + log_vehicle_miles_per_capita,
  data = data_clean,
  index = c("state", "year"),
  model = "within"
)
stargazer(
  expanded_model,
  fixed_effect_model,
```

```
  type = "text",
  omit = c("year"),
  omit.labels = c("year dummy variables"),
  column.labels = c("Pooled", "Fixed Effect")
)
```

```
##
## ================================================================================
##                                          Dependent variable:
##                                 ---------------------------------------------
##                                      log_total_fatalities_rate
##                                   Pooled                 Fixed Effect
##                                     (1)                      (2)
## --------------------------------------------------------------------------------
## blood_alcohol_limitbac10          -0.017                    -0.016
##                                   (0.018)                   (0.010)
##
## blood_alcohol_limitbac08          -0.062**                  -0.020
##                                   (0.024)                   (0.014)
##
## per_se_law1                       -0.019                    -0.053***
##                                   (0.015)                   (0.010)
##
## seat_beltprimary                   0.001                    -0.040***
##                                   (0.025)                   (0.015)
##
## seat_beltsecondary                 0.020                     0.006
##                                   (0.021)                   (0.011)
##
## speed_limit_70plus1               0.222***                  0.072***
##                                   (0.022)                   (0.011)
##
## graduated_drivers_license_law1    -0.021                    -0.015
##                                   (0.025)                   (0.012)
##
## perc14_24                         0.018***                  0.020***
##                                   (0.006)                   (0.004)
##
## log_umeployment_rate              0.267***                  -0.193***
##                                   (0.024)                   (0.017)
##
## log_vehicle_miles_per_capita      1.541***                  0.678***
##                                   (0.044)                   (0.051)
##
## Constant                          20.636***
##                                   (0.554)
##
## --------------------------------------------------------------------------------
## year dummy variables                 Yes                       Yes
## --------------------------------------------------------------------------------
## Observations                        1,200                     1,200
## R2                                  0.668                     0.727
## Adjusted R2                         0.658                     0.708
```

```
## F Statistic                    68.870*** (df = 34; 1165) 87.786*** (df = 34; 1118)
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

For pooled OLS to be the appropriate estimator, we need to assume:

1- **Linearity**: the model is linear in parameters

2- **I.I.D.** : The observations are independent across individuals but not necessarily across time.

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- The independent variables is **uncorrelated** with idiosyncratic error term and individual-specific effect.

The pooled OLS estimator is consistent under assumptions 1-4. We also need to assume **homoskedasticity** and **no serial correlation** in the data to do inference based on the conventional OLS estimator of the covariance matrix.

The main issues of pooled OLS is when the unobserved individual-specific effects are correlated with the independent variables, the model will suffer from an omitted variable bias. In this case, the fixed effect(FE) model is preferred because it eliminates the unobserved time-invariant effects by de-mean procedures. The rest of the assumptions for the FE model are the same.

We can run a test to see whether the pooled OLS model is better than the FE model as below. The null hypothesis is rejected, suggesting the significance of individual fixed effects. **Therefore, the FE model provides better estimates**.

```
pFtest(fixed_effect_model, expanded_model)
```

```
##
##  F test for individual effects
##
## data:  log_total_fatalities_rate ~ year + blood_alcohol_limit + per_se_law +  ...
## F = 105.56, df1 = 47, df2 = 1118, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

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?

6.1 Answer:

The random effects(RE) model assumes the time-invariant unobserved effect is uncorrelated with the explanatory variables. Other than that, the rest assumptions are the same with the FE model. Whether the

time-invariant unobserved effect is uncorrelated with the explanatory variables can be tested with a Hausman test as below. **We reject the null hypothesis, suggesting the assumption is not met.** [1] Since the individual specific effects exist and they are correlated with the explanatory variables, the random effect model will suffer from biased coefficient estimates and standard error estimates.

```
random_effect_model = plm(
  log_total_fatalities_rate ~ year + blood_alcohol_limit + per_se_law +
    seat_belt + speed_limit_70plus + graduated_drivers_license_law + perc14_24 +
    log_umeployment_rate + log_vehicle_miles_per_capita,
  data = data_clean,
  index = c("state", "year"),
  model = "random"
)
phtest(fixed_effect_model,random_effect_model)
```

```
##
##  Hausman Test
##
## data:  log_total_fatalities_rate ~ year + blood_alcohol_limit + per_se_law +  ...
## chisq = 81.408, df = 34, p-value = 9.181e-06
## alternative hypothesis: one model is inconsistent
```

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.

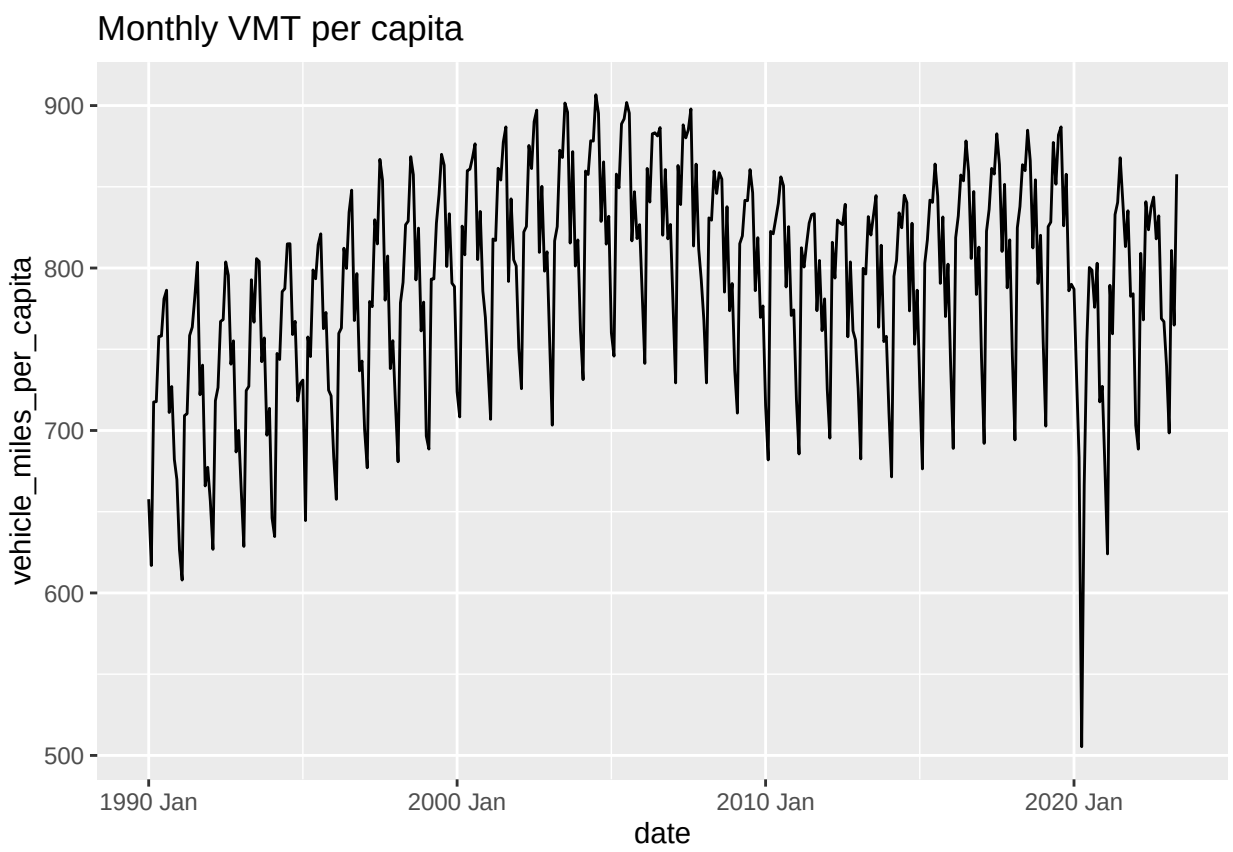
7.1 answer

We download both the monthly *Vehicle Miles Traveled* and *Population* data of the U.S. from FRED and use their ratio as the vehicle miles traveled(VMT) per capita, as below. The time series plot shows it has strong seasonality and changed dramatically during the pandemic.

[1] We notice that if we use the original total fatalities rate instead of its logged version, the test will fail to reject the null hypothesis. Therefore, the data transformation can also influence the validity of the model choice.

```r
if(!"fredr"%in%rownames(installed.packages())) {install.packages("fredr")}
library(fredr)
fredr_set_key("cd565a10e83d56f9f1150d5a2c067e2a")
# Vehicle Miles Traveled are in Millions of Miles, Not Seasonally Adjusted
vmt=fredr(
  series_id = "TRFVOLUSM227NFWA",
  observation_start = as.Date("1990-01-01"),
  observation_end = as.Date("2023-05-01")
) %>% dplyr::select(date,value) %>% as_tsibble(index=date)

# population are in Thousands, Not Seasonally Adjusted
us_pop=fredr(
  series_id = "POPTHM",
  observation_start = as.Date("1990-01-01"),
  observation_end = as.Date("2023-05-01")
)%>% dplyr::select(date,value) %>% as_tsibble(index=date)
# vmt_per_capita are in miles per capita
vmt_per_capita <- vmt %>% mutate(value=value/us_pop$value*1000,date=yearmonth(date)) %>%
  rename(vehicle_miles_per_capita=value)
vmt_per_capita %>% ggplot(aes(x=date,y=vehicle_miles_per_capita))+geom_line()+
  ggtitle("Monthly VMT per capita")
```



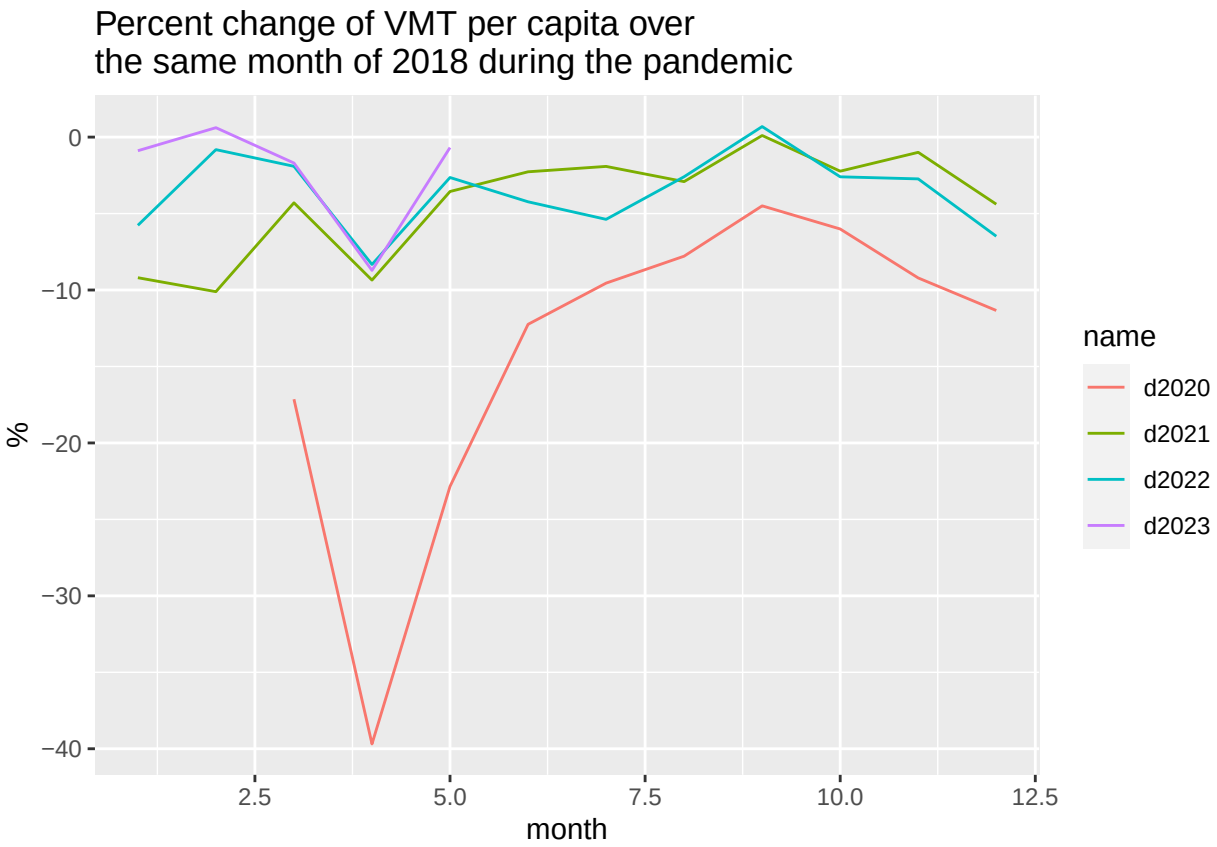
We compare monthly miles driven during pandemic with the same month of 2018. As the table and plot below show:[we exclude the data of Jan 2020 and Feb 2020 because the pandemic hadn't started in the US then.]

- Apr 2020 demonstrated the largest decrease in driving by 39.69%

- Sep 2022 demonstrated the largest increase in driving by 0.69%

```
vmt_pivot = vmt_per_capita %>% mutate(year = year(date), month = month(date)) %>% as_tibble() %>%
  pivot_wider(id_cols = -date,
              names_from = year,
              values_from = vehicle_miles_per_capita)
comparison = vmt_pivot %>% mutate(
  d2020 = 100 * (.data[["2020"]] / .data[["2018"]] - 1),
  d2021 = 100 * (.data[["2021"]] / .data[["2018"]] - 1),
  d2022 = 100 * (.data[["2022"]] / .data[["2018"]] - 1),
  d2023 = 100 * (.data[["2023"]] / .data[["2018"]] - 1),
) %>% dplyr::select(month, d2020, d2021, d2022, d2023)
comparison[1:2,2]=NA
comparison %>% pivot_longer(cols = c(d2020, d2021, d2022, d2023)) %>%
  ggplot(aes(x = month, y = value, color = name)) + geom_line()+
  ggtitle("Percent change of VMT per capita over\nthe same month of 2018 during the pandemic")+ylab("%")
```

```
## Warning: Removed 9 rows containing missing values (`geom_line()`).
```




```
kable(comparison, digits = 2, caption = "Percent change of VMT per capita over\nthe same month of 2018
```

\begin{table}

\caption{Percent change of VMT per capita over the same month of 2018 (%)}

month	d2020	d2021	d2022	d2023
1	NA	-9.19	-5.77	-0.89
2	NA	-10.11	-0.82	0.61
3	-17.14	-4.30	-1.91	-1.70
4	-39.69	-9.35	-8.33	-8.72
5	-22.85	-3.56	-2.64	-0.68
6	-12.24	-2.27	-4.23	NA
7	-9.55	-1.92	-5.38	NA
8	-7.79	-2.91	-2.59	NA
9	-4.50	0.10	0.69	NA
10	-6.01	-2.23	-2.60	NA
11	-9.21	-1.00	-2.73	NA
12	-11.34	-4.39	-6.48	NA

\end{table}

In the FE model above, the coefficient of *log_vehicle_miles_per_capita* is 0.678, which means a unit increase in *log_vehicle_miles_per_capita* will cause the *log_total_fatalities_rate* to increase by 0.678. Therefore, we can calculate the changes to the total fatalities rate based on the changes of *log_vehicle_miles_per_capita* during boom and bust in the pandemic, as below. However, since each state has different population and fatalities rate, we cannot calculate a specific number of fatalities based on the data above.

```
comparison_log = vmt_pivot %>% mutate(
  d2020 = log(.data[["2020"]]) - log(.data[["2018"]]),
  d2021 = log(.data[["2021"]]) - log(.data[["2018"]]),
  d2022 = log(.data[["2022"]]) - log(.data[["2018"]]),
  d2023 = log(.data[["2023"]]) - log(.data[["2018"]]),
) %>% dplyr::select(month, d2020, d2021, d2022, d2023)
result = data.frame(
  scenario = c("boom", "bust"),
  time = c("2022M09", "2020M04"),
  change_in_log_vmt_per_capita = c(as.numeric(comparison_log[9, "d2022"]), as.numeric(comparison_log[4,
) %>% mutate(change_in_log_total_fatalities_rate = change_in_log_vmt_per_capita *
               0.678)
kable(result, digits = 2)
```

scenario	time	change_in_log_vmt_per_capita	change_in_log_total_fatalities_rate
boom	2022M09	0.01	0.00
bust	2020M04	-0.51	-0.34

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?

8.1 answer

Serial correlation or heteroskedasticity will not influence the consistence of the estimators but will lead to inaccurate estimate of their standard errors.

We can use Breusch Pagan test the FE model (the only valid one) for heteroskedasticity. As shown below, the test rejects the null hypothesis of homoskedasticity, in favor of heteroskedasticity. Therefore, we should use robust standard error estimates.

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

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  log_total_fatalities_rate ~ year + blood_alcohol_limit + per_se_law +     seat_belt + speed_li
## chisq = 2713.5, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

We perform both Durbin Watson test and Breusch-Godfrey test for the FE model. As shown below, both tests suggest serial correlation in errors. This confirms that we should use robust standard error estimates.

```
pdwtest(fixed_effect_model)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  log_total_fatalities_rate ~ year + blood_alcohol_limit + per_se_law +  ...
## DW = 1.2192, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(fixed_effect_model, order = 2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  log_total_fatalities_rate ~ year + blood_alcohol_limit + per_se_law +  ...
## chisq = 202.72, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```