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?"

```
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 new variables - year_of_observation, speed_limit, speed_limit_70plus, blood_alcohol_limit
#
df <- data %>%
  mutate(state = factor(state)) %>%
  rowwise() %>%
  # speed_limit
  mutate(
    speed_limit_70plus = factor(sl70plus),
    speed_limit = parse_number(
      colnames(
        select(data, starts_with("sl"))
      )[which.max(c_across(starts_with("sl")))],
      na = "slnone"
```

```
    ),
  ) %>%
  select(-starts_with("sl")) %>%
  mutate(year_of_observation = factor(year)) %>% # year_of_observation
  select(-starts_with("d")) %>%
  mutate(blood_alcohol_limit = factor(parse_number(
    colnames(
      select(data, starts_with("bac"))
    )[which.max(c_across(starts_with("bac")))]
  ) / 100)) %>% # blood_alcohol_limit
  select(-starts_with("bac")) %>%
  mutate(
    seatbelt = factor(seatbelt), # 'seatbelt' categorizes primary or secondary
    speed_limit_70plus = ifelse(speed_limit == 55 | speed_limit == 65, 0, 1)
  ) %>%
  select(-starts_with("sb"))

# rename the variables to sensible names
df <- df %>%
  dplyr::rename(
    "total_fatalities_rate" = "totfatrte",
    "minimum_drinking_age" = "minage",
    "zero_tolerance_law" = "zerotol",
    "graduated_drivers_license_law" = "gdl",
    "per_se_laws" = "perse",
    "total_traffic_fatalities" = "totfat",
    "total_nighttime_fatalities" = "nghtfat",
    "total_weekend_fatalities" = "wkndfat",
    "total_fatalities_per_100_million_miles" = "totfatpvm",
    "nighttime_fatalities_per_100_million_miles" = "nghtfatpvm",
    "weekend_fatalities_per_100_million_miles" = "wkndfatpvm",
    "nighttime_fatalities_rate" = "nghtfatrte",
    "weekend_fatalities_rate" = "wkndfatrte",
    "vehicle_miles" = "vehicmiles",
    "unemployment_rate" = "unem",
    "pct_population_14_to_24" = "perc14_24",
    "vehicle_miles_per_capita" = "vehicmilespc"
  ) %>%
  select(
    year_of_observation,
    state,
    year,
    # response variables
    total_fatalities_rate,
    nighttime_fatalities_rate,
    weekend_fatalities_rate,
    total_traffic_fatalities,
    total_nighttime_fatalities,
    total_weekend_fatalities,
    total_fatalities_per_100_million_miles,
    nighttime_fatalities_per_100_million_miles,
    weekend_fatalities_per_100_million_miles,
    # potential explanatory variables
```

```
    seatbelt,
    zero_tolerance_law,
    graduated_drivers_license_law,
    per_se_laws,
    minimum_drinking_age,
    speed_limit_70plus,
    speed_limit,
    blood_alcohol_limit,
    vehicle_miles,
    vehicle_miles_per_capita,
    # econ and demographic variables
    statepop,
    unemployment_rate,
    pct_population_14_to_24, vehicle_miles
  ) # keep the similar variables together

# check the data
# df %>% glimpse()
```

```
##     sl55  sl65  sl70 sl75 slnone sum_sl
## 8  0.542 0.458 0.000    0      0     76
## 17 0.000 0.333 0.667    0      0     76
```

Several notes about our data processing:

1) As noted in the above 2 examples, the original dataset has some percentage values in the sl55/sl65/sl70/s75/slone columns. We expect these columnes to be binary (0, 1). In our data processing, we classified the category based on the field with the max value. For example, we classified the first example as sl55, and the second example as sl70.

2) speed_limit_70plus column has values that are not 0 or 1. We reclassified these values to 0 or 1, based on speed_limit column. We als reclassified the states with no speed limit as 1 for this variable at this stage.

```
# check the values of these fields
# df$zero_tolerance_law %>%
#   table(useNA = "ifany") %>%
#   as.data.frame()
#
# df$graduated_drivers_license_law %>%
#   table(useNA = "ifany") %>%
#   as.data.frame()
#
# df$per_se_laws %>%
#   table(useNA = "ifany") %>%
#   as.data.frame()

# make binary variables
df <- df %>%
  mutate(
    zero_tolerance_law = ifelse(
      zero_tolerance_law == 0 | zero_tolerance_law == 1, zero_tolerance_law, 1
    ),
    graduated_drivers_license_law = ifelse(
      graduated_drivers_license_law == 0 | graduated_drivers_license_law == 1,
      graduated_drivers_license_law,
```

```
      1
    ),
    per_se_laws = ifelse(
      per_se_laws == 0 | per_se_laws == 1, per_se_laws, 1
    ),
    blood_alcohol_limit_binary = ifelse(
      blood_alcohol_limit == 0.1, 1, 0
    )  # set 1 for 0.1 blood_alcohol_limit and 0 for 0.08 blood_alcohol_limit
  )
```

3) We observed non-binary values in the following columns: zero_tolerance_law, graduated_drivers_license_law, per_se_laws. Since we expect these columns to have binary values (0,1) given the definition, we decided to treat all non-zero values as 1 and make it a binary variable. We also created a binary variable blood_alcohol_limit_binary to set 1 for 0.1 blood_alcohol_limit and 0 for 0.08 blood_alcohol_limit, as there are only two values (0.1 and 0.08) for this variable.

```
#check minimum_drinking_age
# df$minimum_drinking_age %>%
#   table(useNA = "ifany") %>%
#   as.data.frame()

df <- df %>%
  mutate(
    minimum_drinking_age = round(minimum_drinking_age, 0)
  )
```

4) We noticed that the minimum_drinking_age column has values that are not integers. We decided to round these values to the nearest integer.

```
# check for na speed_limit rows
check_na_speed <- df %>%
  filter(is.na(speed_limit)) %>%
  select(state, year_of_observation, speed_limit, speed_limit_70plus)

# head(check_na_speed,2)

# treat the na speed_limit rows for State 27
df <- df %>%
  mutate(
    speed_limit = ifelse(
      is.na(speed_limit) & state == 27, ifelse(
        year >= 1996 & year <= 1999, 85, 75
      ), speed_limit
    ),
    speed_limit_70plus = ifelse(
      is.na(speed_limit_70plus) & state == 27, 1, speed_limit_70plus
    )
  )
```

5) We observed that the *speed_limit* is not set for State 27, between 1996 to 2004. Our background research noted the fact that "for three years after the 1995 repeal of the increased 65 mph limit, Montana had a non-numeric"reasonable and prudent" speed limit during the daytime on most rural roads". But it doesn't mean there was no speed limit.

We decided to set the *speed_limit* to 85 for Montana between 1996 to 1999, given the legal case

of State v. Rudy Stanko (1998), who got charged for speed of 85. Effective May 28, 1999, as a result of that decision, the Montana Legislature established a speed limit of 75 mph. So we set the *speed_limit* to 75 for Montana between 2000 to 2004.

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?
     This data set is a balanced longitudinal dataset and contains traffic fatalities data for the 48 continental U.S. states from 1980 through 2004. For each year of observation, the dataset contains state-level cross sectional measurements of fatality count and rate. This data is collected and distributed by Jeffrey M. Wooldridge through this link. In this dataset, the total_fatalities_rate is defined as total fatalities per 100,000 population.

After our data processing work, the clean dataset (df) has 25 columns/fields which include: - Index variables: year_of_observation, state, year - 9 fatality variables: There are three measurements - fatality count, fatality count per 100M miles and fatality rate as defined as count per 100k population. These three measurements are provided for total, nighttime and weekend - 10 law and vehicle variables: 8 traffic laws indicators (seatbelt, zero_tolerance_law, graduated_drivers_license_law, per_se_laws, minimum_drinking_age, speed_limit_70plus, speed_limit, blood_alcohol_limit) and 2 driving variables (vehicle_miles, vehicle_miles_per_capita) - 3 Economics and demographic variables: statepop, unemployment_rate, pct_population_14_to_24

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?

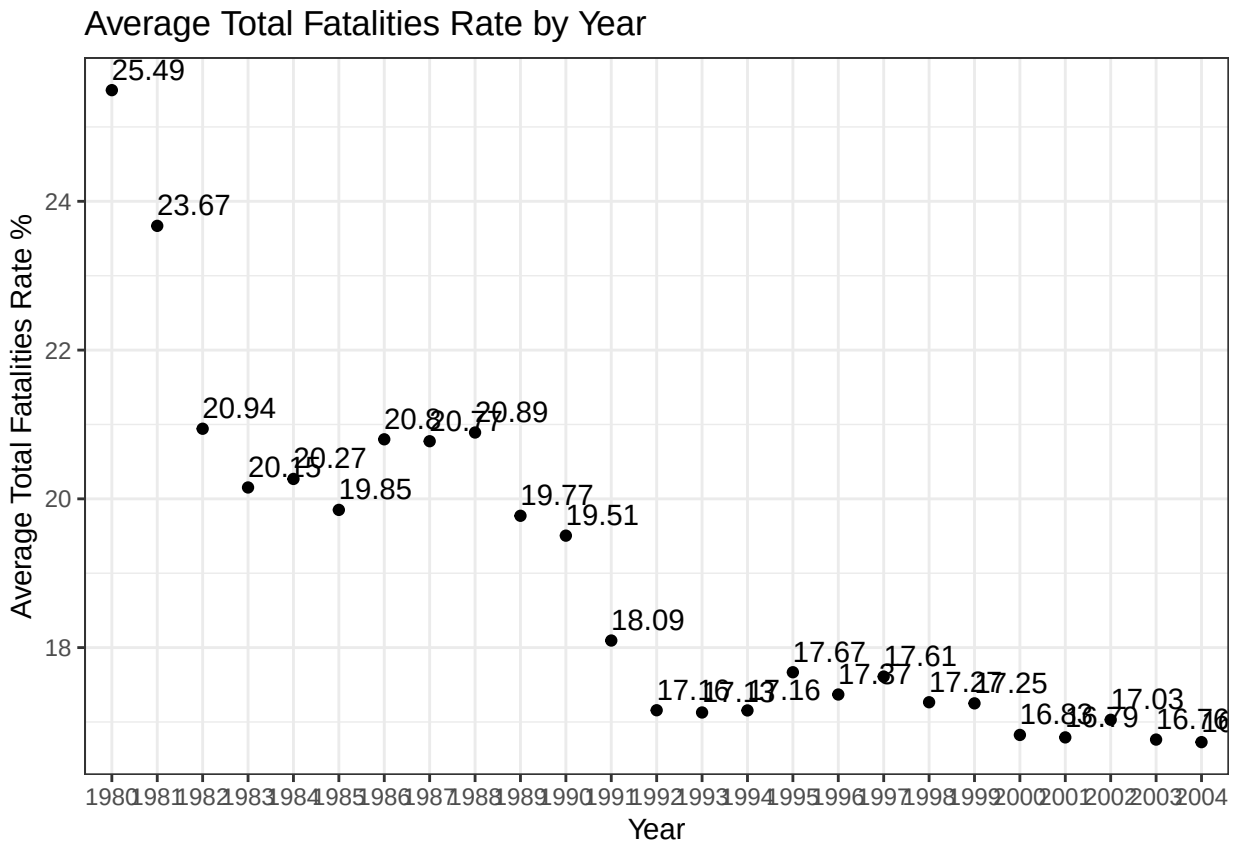
  In this dataset, the total_fatalities_rate is defined as total fatalities per 100,000 population.

- What is the average of `total_fatalities_rate` in each of the years in the time period covered in this dataset?

  See below the table and the plot.

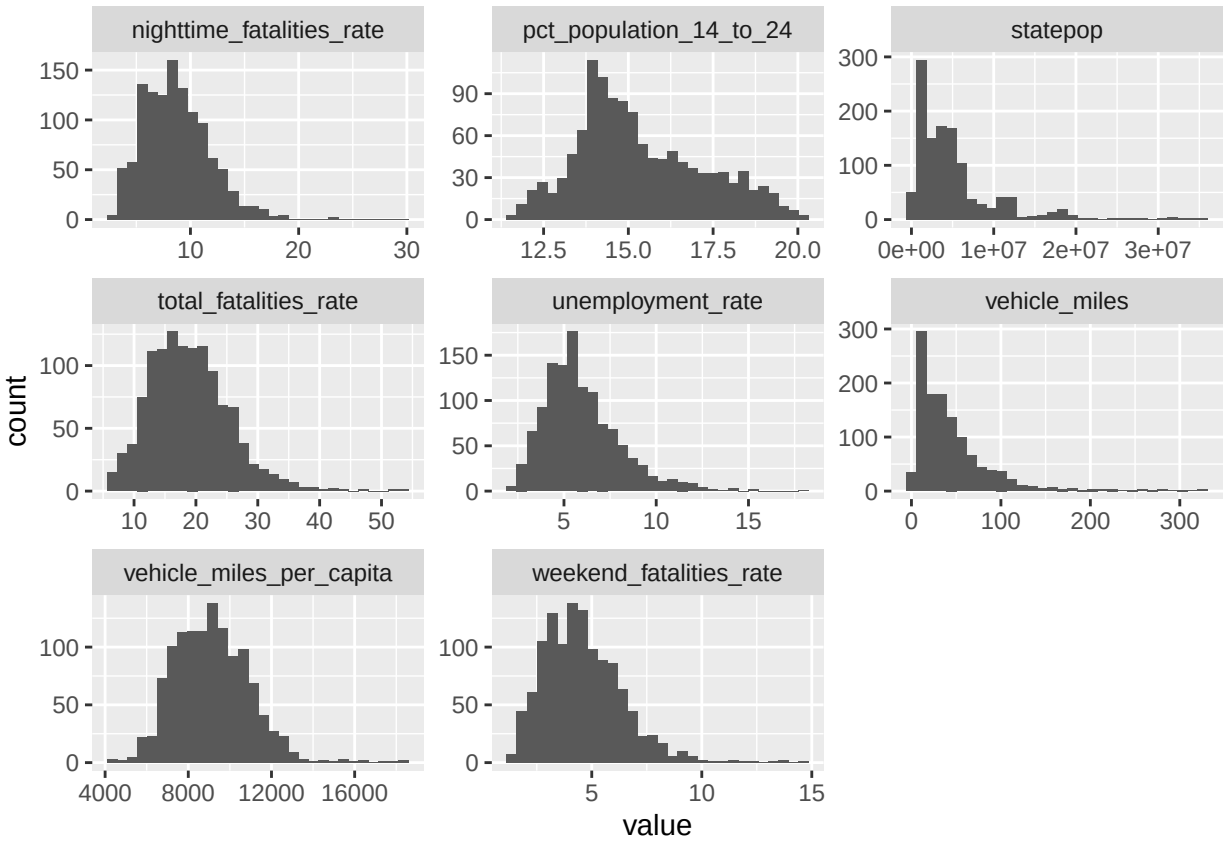
```
df_avg <- df %>%
  group_by(year_of_observation) %>%
  summarise(avg_total_fatalities_rate = mean(total_fatalities_rate))

# average fatality by year
years <- unique(df$year_of_observation)
avg_df <- data.frame(
  year = years,
  avg_fatality_rate = round(df_avg$avg_total_fatalities_rate, 2)
) %>%
  knitr::kable()

# plot fatality by year
df_avg %>%
  ggplot(aes(year_of_observation, avg_total_fatalities_rate,
    label = round(avg_total_fatalities_rate, 2)
  )) +
  geom_point() +
  geom_text(hjust = 0, vjust = -0.5) +
```

```
  theme_bw() +
  labs(title = "Average Total Fatalities Rate by Year") +
  xlab("Year") +
  ylab("Average Total Fatalities Rate %")
```



Next, check the distribution of the continuous variables. We noted that the distributions of the variables are mostly right-skewed.

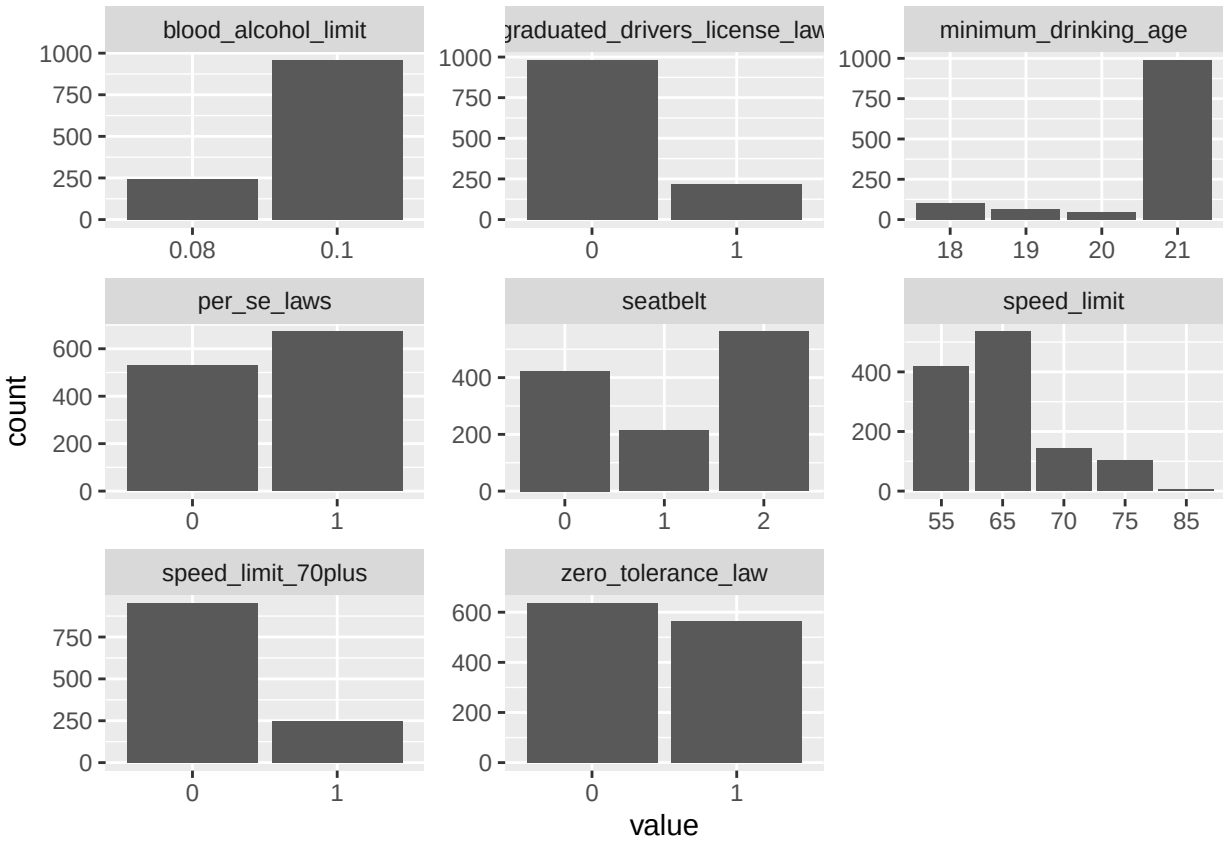
```
df %>%
  select(
    total_fatalities_rate,
    nighttime_fatalities_rate,
    weekend_fatalities_rate,
    vehicle_miles,
    vehicle_miles_per_capita,
    statepop,
    unemployment_rate,
    pct_population_14_to_24
  ) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~key, scales = "free") +
  geom_histogram()
```




```
# hist(df$pct_population_14_to_24)
# hist(log(df$pct_population_14_to_24))
```

We also examined the count distribution of the categorical variables.

```
df %>%
  select(
    seatbelt,
    zero_tolerance_law,
    graduated_drivers_license_law,
    per_se_laws,
    minimum_drinking_age,
    speed_limit_70plus,
    speed_limit,
    blood_alcohol_limit
  ) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~key, scales = "free") +
  geom_bar()
```



We observed the following distribution for these categorical variables - Blood alcohol limit: Most states have the limit of 0.1. - Minimum drinking age: Most states have 21. - Graduated drivers license law, most states have 0 - Speed limit: Most states have less than 70 miles - per-se_laws, seatbelt and zero_tolence_law are more evenly distributed among the categories

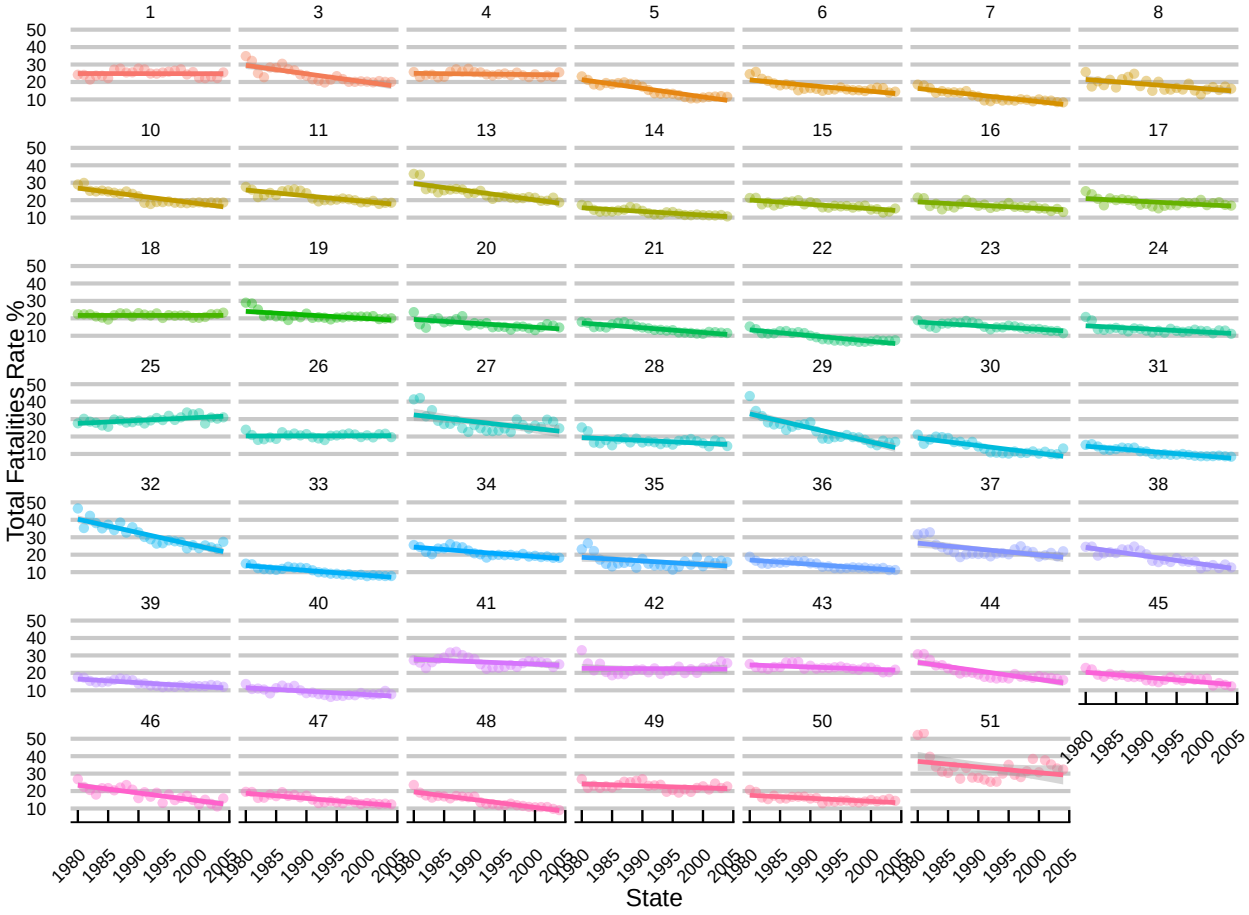
We plotted total_fatalities_rate by state below to understand the fixed effects by state

We see strong differences in total traffic fatalities rates across states, suggesting that fixed effects are important for controlling for unobserved differences.

We also plotted total_fatalities_rate by state below to understand the trend over the years.

```
# we do have 48 states -> missing states 2, 9, and 12.
# Thus, the ordering is a little bit off
df %>%
  ggplot(aes(year, total_fatalities_rate, color = state)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  facet_wrap(~state) +
  theme_economist_white(gray_bg = FALSE) +
  theme(
    legend.position = "none",
    axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1, size = 8),
    axis.text.y = element_text(size = 8),
    strip.text = element_text(size = 8)
  ) +
  scale_y_continuous() +
  xlab("State") +
  ylab("Total Fatalities Rate %")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

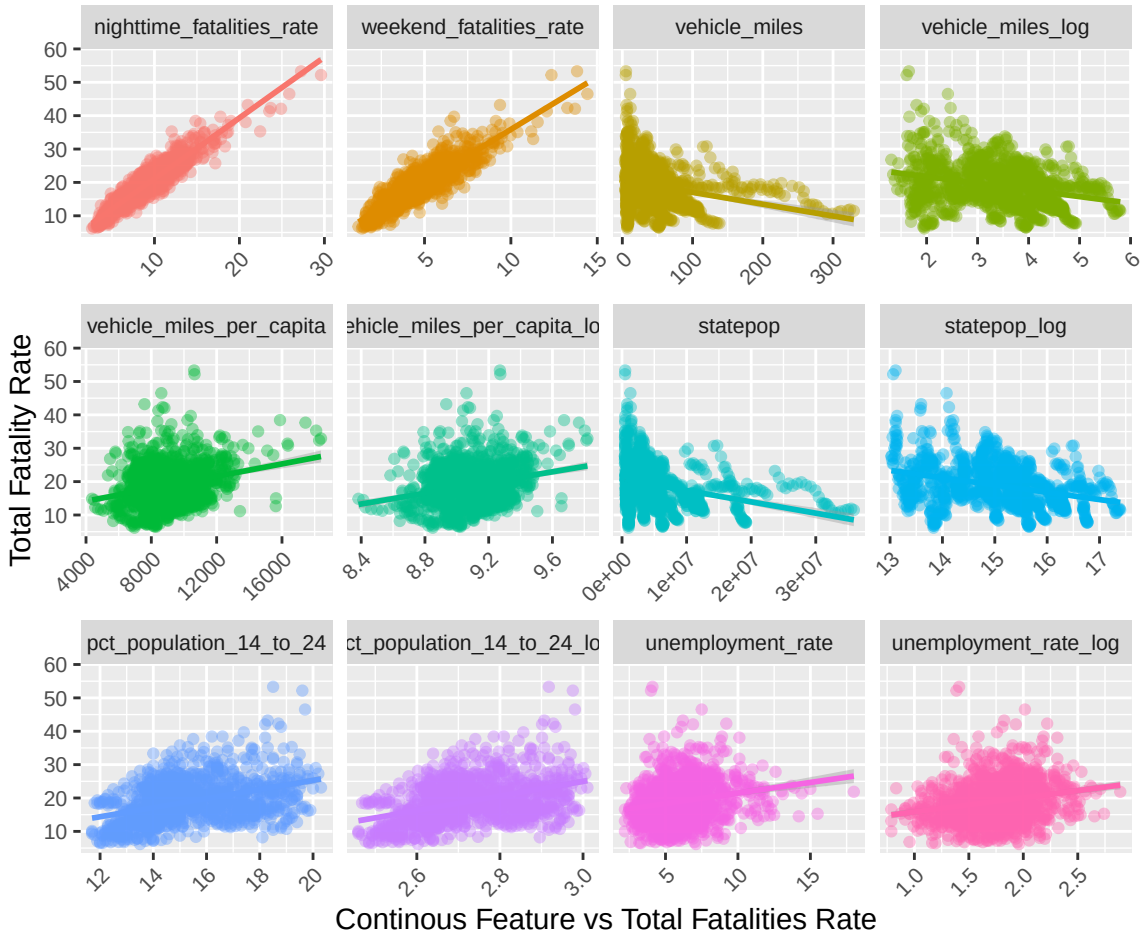


We noted that the plots are sequentially ordered, despite of missing state number 2, 9, and 12. Most states have a downward trend in total fatalities rate over the years, however there are trend variations by state.

```
df %>%
  mutate(
    vehicle_miles_log = log(vehicle_miles),
    vehicle_miles_per_capita_log = log(vehicle_miles_per_capita),
    statepop_log = log(statepop),
    pct_population_14_to_24_log = log(pct_population_14_to_24),
    unemployment_rate_log = log(unemployment_rate)
  ) %>%
  select(
    total_fatalities_rate,
    nighttime_fatalities_rate,
    weekend_fatalities_rate,
    vehicle_miles,
    vehicle_miles_log,
    vehicle_miles_per_capita,
    vehicle_miles_per_capita_log,
    statepop,
    statepop_log,
    pct_population_14_to_24,
    pct_population_14_to_24_log,
```

```
    unemployment_rate,
    unemployment_rate_log
  ) %>%
  melt(id.vars = c("total_fatalities_rate")) %>%
  ggplot(aes(value, total_fatalities_rate, color = variable)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  facet_wrap(~variable, scales = "free_x") +
  # theme_economist_white(gray_bg=F) +
  theme(
    legend.position = "none",
    axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1, size = 8),
    axis.text.y = element_text(size = 8),
    strip.text = element_text(size = 8)
  ) +
  # scale_y_continuous(label=percent) +
  xlab("Continous Feature vs Total Fatalities Rate") +
  ylab("Total Fatality Rate")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We explored log transformation on several variables: vehicle_miles, vehicle_miles_per_capita, statepop, pct_population_14_to_24, unemployment_rate. We visualized both the original and log transformed variables. The log transformation seemed to improve the relationship between the

explanatory variable and the respond variable (total fatality rate) visually. Therefore we decided to use log transformation for our interpretation for vehicle_miles, vechicle_miles_per_capita, statepop and unemployment_rate.

Total fatalities rate is positively correlated with unemployment_rate and percentage of population aged 14 through 24.

Total fatalities rate is negatively correlated with vehicle miles and state population which is not intuitive. We interpret this as population and vehicle miles are both increasing with time, so is the other driving forces of fatalities rate (quality of the car, technology of the car, road conditions, etc.), so the relationship between total fatalities rate vs vehicle miles and state population is potentially spurious, and thus inconsistent with our background knowlege.

Total fatalities rate is positively correlated with vehicle miles per capita, as what we expect. As the density of the population increases, there is expected to be more severe traffic incidents that leads to higher fatalities rate. This is consistent with our background knowledge. Therefore, it makes sense to use vehicle miles per capita which is normalized for the population.

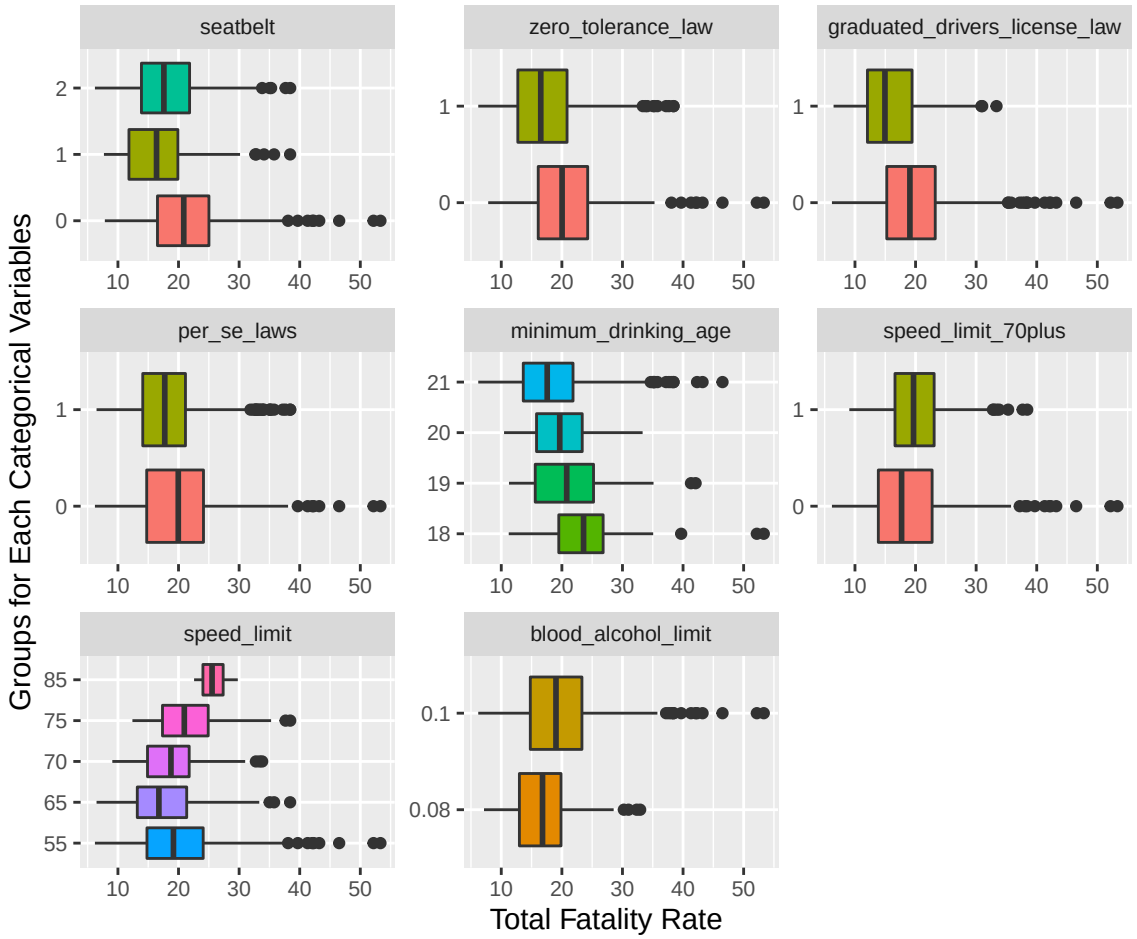
```
df <- df %>%
  mutate(
    vehicle_miles_log = log(vehicle_miles),
    vehicle_miles_per_capita_log = log(vehicle_miles_per_capita),
    statepop_log = log(statepop),
    unemployment_rate_log = log(unemployment_rate)
  )
```

We used boxplots to visually examine the differences in total fatalities rate for the groups within each law categorical variable. We observed that stricter law requirements tend to associate with lower total fatalities rate, see below the detailed comments for each law variable.

- Seatbelt: No seatbelt requirements tend to have higher fatalities rate, in comparison to having primary and secondary seatbelt requirements.
- For zero_tolerance_law,graduated_drivers_license_law and per_se_laws: 0 (stricter laws) tends to have lower fatalities rate, in comparison to having some level of flexibility/tolerance in law requirements.
- minimum_drinking_age: higher legally-eligible age tends to have lower fatalities rate, in comparison to lowering the age legal requirement
- speed_limit_70plus: below 70 speed limit tends to have lower fatalities rate, in comparison to allowing 70+ speed limit
- blood_alcohol_limit: 0.08 blood alcohol limit tends to have have lower fatalities rate, in comparison to 0.1 blood alcohol limit

```
df %>%
  select(
    total_fatalities_rate,
    seatbelt,
    zero_tolerance_law,
    graduated_drivers_license_law,
    per_se_laws,
    minimum_drinking_age,
    speed_limit_70plus,
    speed_limit,
    blood_alcohol_limit
  ) %>%
  melt(id.vars = c("total_fatalities_rate")) %>%
  ggplot(aes(value, total_fatalities_rate)) +
  geom_boxplot(aes(fill = factor(value))) +
  coord_flip() +
```

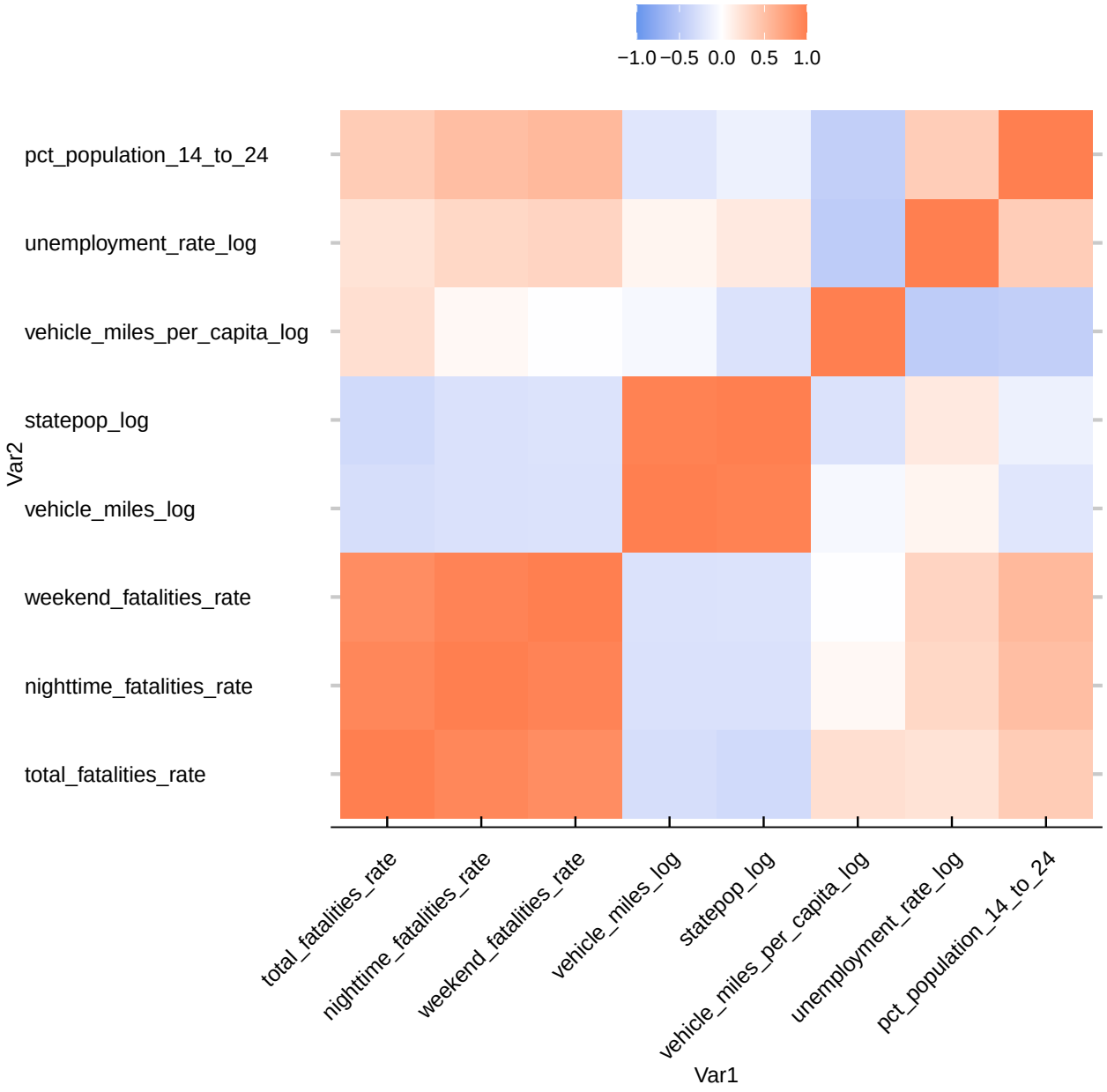
```
  facet_wrap(~variable, scales = "free") +
  theme(
    legend.position = "none",
    axis.text.x = element_text(size = 8),
    axis.text.y = element_text(size = 8),
    strip.text = element_text(size = 8)
  ) +
  xlab("Groups for Each Categorical Variables ") +
  ylab("Total Fatality Rate")
```



We examined the correlation matrix of the continuously variables, so we can eliminate including the explanatory variables with perfect correlation with each other.

```
df %>%
  mutate(seatbelt_num = as.numeric(seatbelt)) %>%
  select(
    total_fatalities_rate,
    nighttime_fatalities_rate,
    weekend_fatalities_rate,
    vehicle_miles_log,
    statepop_log,
    vehicle_miles_per_capita_log,
    unemployment_rate_log,
    pct_population_14_to_24,
```

```
) %>%
mutate_all(as.numeric) %>%
cor() %>%
melt() %>%
ggplot(aes(Var1, Var2, fill = value)) +
geom_tile() +
theme_economist_white(gray_bg = FALSE) +
theme(
  legend.title = element_blank(),
  legend.text = element_text(size = 10),
  axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1)
) +
scale_fill_gradient2(
  low = "cornflowerblue", high = "coral", mid = "white",
  midpoint = 0, limit = c(-1, 1)
)
```



The state population and vehicle miles are almost perfectly correlated. To avoid the colinearity problem, we will only use vehicle_miles_per_capita in our model development. In our exploratory work above, we also noted that state population and vehicle miles seem to have spurious relationship with total fatalities rate, so it makes sense to exclude them in the model.

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.

To summarize our data exploratory work (as detailed in the above plots and comments), we noted the following key takeaways:

- 1) We pre-processed the variables in the original dataset, including creating required new variables, cleaning up categorical variables for consistency and renaming all variables with sensible names.

Our final dataset in the df file have 25 variables, including index variables (year/year-of-observations and state), the fatality rates variables and the potential explanatory variables (law, economics and demographic variables). We noted 48 states are included in this dataset, which are number 1-51 except for missing state number 2, 9 and 12.

- 2) We conducted comprehensive data exploratory work by using extensive plots to visualize and examine the distribution of each variable, relationship between the total fatalities rate and the potential explanatory variables (including the log form of several variables), as well as the correlation among the variables.
- 3) Based on our exploratory work, we noted a declining trend of the average of total fatalities rates over the period, with trend variations at the state level. However, that does not mean time by itself caused the fatalities to go down, based on our experience. There could be many other factors that caused the total fatalities rates to go down over the years. We had two key observations that could explain the changes in total fatalities rates over the period: a) stricter traffic laws (as indicated in seatbelt, zero_tolerance_law, graduated_drivers_license_law, per_se_laws, minimum_drinking_age, speed_limit_70plus, blood_alcohol_limit) tend to associate with lower total fatalities rate. b) there is a sensible and positive relationship between total fatalities rate and each of the three economics/demographic variables (pct_population_14_to_24, unemployment_rate_log and vehicle_miles_per_capita_log). These observations provide important insight and guidance to our subsequent modeling work.

3 (15 points) Preliminary Model

```
mod.lm1 <- lm((total_fatalities_rate) ~ year_of_observation, data = df)
# summary(mod.lm1)
# par(mfrow = c(2, 2))
# plot(mod.lm1)
```

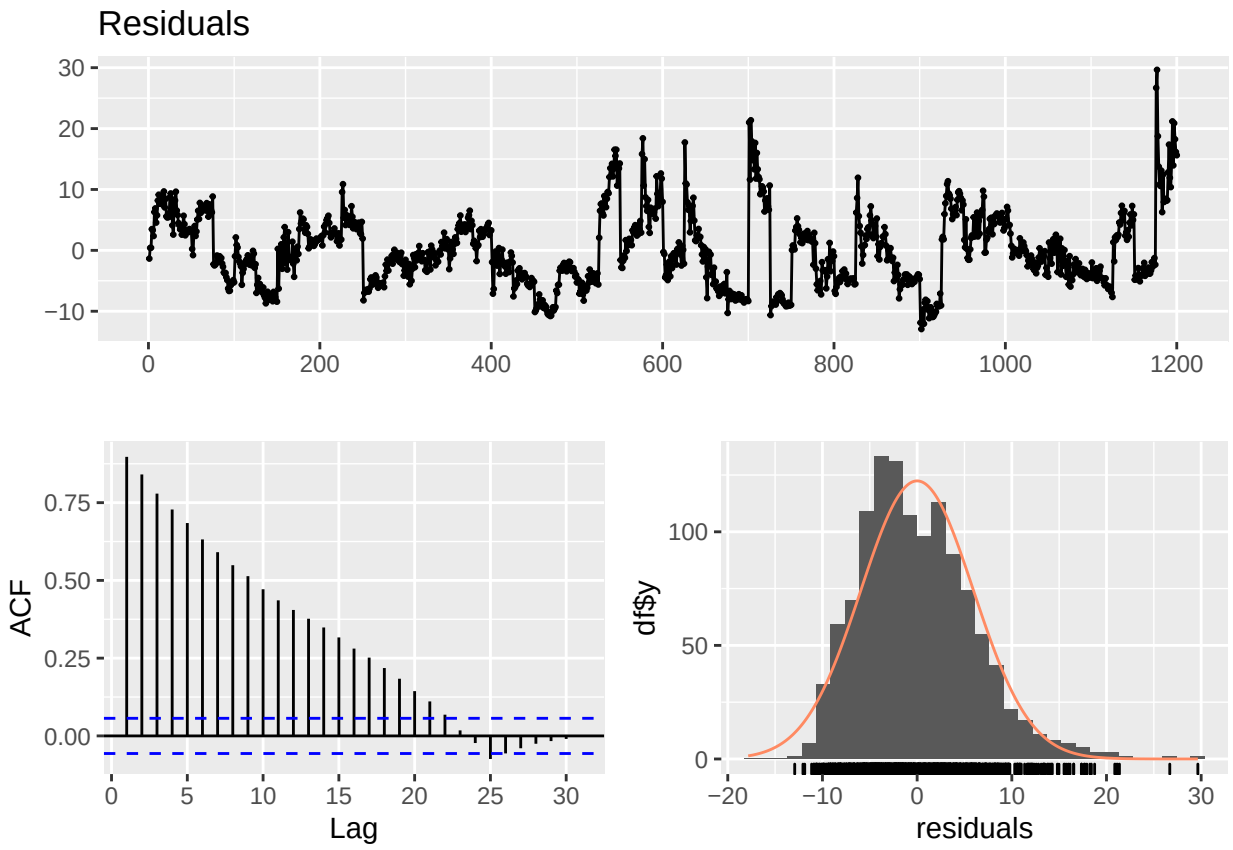
```
coeftest(mod.lm1, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##                          Estimate Std. Error t value  Pr(>|t|)
## (Intercept)               25.4946     1.1573 22.0295 < 2.2e-16 ***
## year_of_observation1981   -1.8244     1.6151 -1.1296 0.2588825
## year_of_observation1982   -4.5521     1.5100 -3.0147 0.0026276 **
## year_of_observation1983   -5.3417     1.4551 -3.6710 0.0002524 ***
## year_of_observation1984   -5.2271     1.4068 -3.7155 0.0002123 ***
## year_of_observation1985   -5.6431     1.3988 -4.0344 5.829e-05 ***
## year_of_observation1986   -4.6942     1.4185 -3.3092 0.0009640 ***
## year_of_observation1987   -4.7198     1.4233 -3.3162 0.0009404 ***
## year_of_observation1988   -4.6029     1.3900 -3.3114 0.0009564 ***
## year_of_observation1989   -5.7223     1.3946 -4.1030 4.359e-05 ***
## year_of_observation1990   -5.9894     1.4130 -4.2386 2.425e-05 ***
## year_of_observation1991   -7.3998     1.3843 -5.3455 1.082e-07 ***
## year_of_observation1992   -8.3367     1.3862 -6.0142 2.410e-09 ***
## year_of_observation1993   -8.3669     1.3777 -6.0733 1.690e-09 ***
## year_of_observation1994   -8.3394     1.4013 -5.9513 3.506e-09 ***
## year_of_observation1995   -7.8260     1.4533 -5.3851 8.738e-08 ***
## year_of_observation1996   -8.1252     1.4223 -5.7129 1.406e-08 ***
## year_of_observation1997   -7.8840     1.4357 -5.4913 4.887e-08 ***
## year_of_observation1998   -8.2292     1.4435 -5.7009 1.506e-08 ***
```

```
## year_of_observation1999  -8.2442     1.4812 -5.5659 3.228e-08 ***
## year_of_observation2000  -8.6690     1.4473 -5.9898 2.789e-09 ***
## year_of_observation2001  -8.7019     1.4418 -6.0356 2.120e-09 ***
## year_of_observation2002  -8.4650     1.4681 -5.7658 1.038e-08 ***
## year_of_observation2003  -8.7310     1.4445 -6.0443 2.012e-09 ***
## year_of_observation2004  -8.7656     1.4699 -5.9634 3.263e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
forecast::checkresiduals(mod.lm1)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```




```
##
##  Breusch-Godfrey test for serial correlation of order up to 28
##
## data:  Residuals
## LM test = 1009.4, df = 28, p-value < 2.2e-16
```

- Why is fitting a linear model a sensible starting place? > As a start, fitting a linear model helps identify significant explanatory variables and evaluate whether the linear relationship exists and how strong the linear relationship is.
- What does this model explain, and what do you find in this model? > As noted above, we expect the fatalities rate to decline over the years given our EDA observations. This model shows whether a given year has a linear relationship with total fatalities rate, with Year 1980 as the base value in the linear model. There is strong statistical evidence that all the years (except for 1981) are negatively related to

total fatalities rate at the significance level of 0. The negative relationship aligns with the decreasing trend observed in EDA. Year 1981 is not a significant variable in this model.

- Did driving become safer over this period? Please provide a detailed explanation. > As noted in our EDA work, the average total fatalities rate was 24% in 1980 and by 2004 the average total fatalities decreased down to 16%. This linear model shows statistically significant relationship between time and total fatalities rates. However, based on our expericence, time by itself does not necessarily have a causal effect on the fatalities rates. There are many other factors (e.g. law and economics/demographic variables) that could explain the changes in total fatalities rates over the period, as we summarized as the key take-aways in the last part of our EDA section.
- 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?

    Based on the residual diagnostic analysis and plots, the residuals have large fluctuations for each year and are not constant with zero mean expectation. This violates one of the key classical linear model assumptions - linear conditional mean. So the parameter estimates are biased and not reliable.

    The parameter estimates are biased because the model recognizes year as the sole explanatory variable for the total fatalities rate. This introduces omitted variable bias because we have not accounted for other factors such as law, economics and demographic explanatory variables, as we discussed in the EDA section.
- 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 our residual analysis above, the residuals' distribution is not normal. variance is not constant (heteroskedasticity issues). The Breusch-Godfrey test for serial correlation has a p-value much less than 0.05. There is strong evidience of serial correlation in the model residuals. The ACF plot shows significant positive autocorrelation of many lags. With positive serial correlation, the standard errors (uncertain estimate of our parameter coefficients) are understated and the statistical inferences are not reliable.

  We pulled heteroskedasticity-robust standard errors using coeftest(mod.lm1, vcov. = vcovHC, type = "HC1") above. The HC errors are higher than the standard error (1.2263) reported in the linear model. This dataset is structured by state and each state has 25 rows representing 25 years. Using this data structure and year as the sole variable, this model has omitted variables bias as noted above.

4 (15 points) Expanded Model

In the EDA section, we included detailed codes and explanation of our data pre-processing work, in order to clean up and standardize our variables to prepare for our exploratory work and model development.

We also applied the log transformation to total_fatalities_rate, unemployment_rate, and vehicle miles per capita to normalize the distribution given the original skewed distribution.

In the summary take-aways part of our EDA section (the last part of EDA), we included the set of key variables that can explain total fatalities rate, which assign with the variables that we are instructed to use in this expanded model.

- Below is the result of the expanded model (mod.lm2).

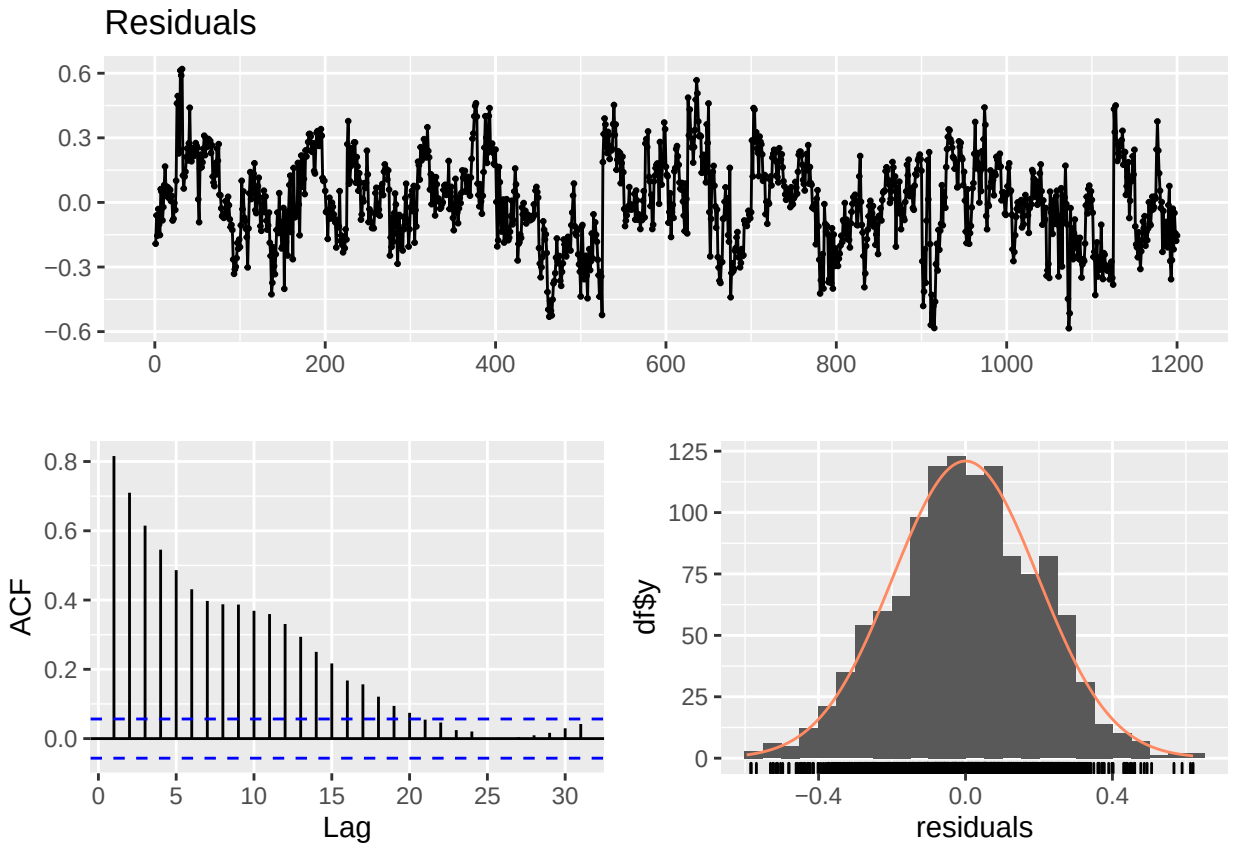
```
#df <- pdata.frame(df, index = c('state', 'year_of_observation'))
```

```
mod.lm2 <- lm(
  log(total_fatalities_rate) ~ year_of_observation
    + factor(blood_alcohol_limit)
    + factor(per_se_laws)
    + factor(seatbelt)
    + factor(speed_limit_70plus)
    + factor(graduated_drivers_license_law)
    + pct_population_14_to_24
    + unemployment_rate_log
    + vehicle_miles_per_capita_log,
  data = df

)
#summary(mod.lm2)
```

Looking at the coefficients estimates, all the years coefficients are negative, which aligns with our EDA observations. However, some of the coefficients have unintuitive signs and not aligned with our EDA work and our experience. For example, seatbelt 2 (secondary seatbelt) has a positive coefficient compared to no seatbelt requirement, which is not sensible. unemployment_rate_log has a positive coefficient, which is not what we expect as well.

We analyzed the model residuals below

```
forecast::checkresiduals(mod.lm2)
```




```
##
##  Breusch-Godfrey test for serial correlation of order up to 37
##
```

```
## data:  Residuals
## LM test = 833.32, df = 37, p-value < 2.2e-16
```

The model residuals improved over the prelimary linear model version based on the plots. The Normal Q-Q plot is closer to the normal distribution, better than the preliminary linear model. However, the residuals still have a lot of variance fluctuations, and serial correlation is a major problem based on the Breusch-Godfrey test and the ACF plot. Serial correlation makes standard error estimates and statistical inferences not reliable.

We will need to use robust standard errors.

```
# use the robust standard errors
coeftest(mod.lm2, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##                                          Estimate  Std. Error  t value
## (Intercept)                            -1.1304e+01  4.3808e-01 -25.8029
## year_of_observation1981                -9.1458e-02  4.5747e-02  -1.9992
## year_of_observation1982                -2.9488e-01  4.5472e-02  -6.4848
## year_of_observation1983                -3.5158e-01  4.3525e-02  -8.0775
## year_of_observation1984                -3.0565e-01  4.3908e-02  -6.9611
## year_of_observation1985                -3.4394e-01  4.4931e-02  -7.6550
## year_of_observation1986                -3.2076e-01  4.8890e-02  -6.5608
## year_of_observation1987                -3.5700e-01  4.9161e-02  -7.2619
## year_of_observation1988                -3.6729e-01  5.1127e-02  -7.1839
## year_of_observation1989                -4.5246e-01  5.5107e-02  -8.2105
## year_of_observation1990                -5.1193e-01  5.9466e-02  -8.6088
## year_of_observation1991                -6.2769e-01  6.0371e-02 -10.3973
## year_of_observation1992                -7.3378e-01  6.3545e-02 -11.5474
## year_of_observation1993                -7.2507e-01  6.2049e-02 -11.6854
## year_of_observation1994                -7.1109e-01  6.2311e-02 -11.4120
## year_of_observation1995                -6.8866e-01  6.4406e-02 -10.6926
## year_of_observation1996                -8.1354e-01  6.4378e-02 -12.6369
## year_of_observation1997                -8.2170e-01  6.5436e-02 -12.5574
## year_of_observation1998                -8.6859e-01  6.5567e-02 -13.2475
## year_of_observation1999                -8.6643e-01  6.5720e-02 -13.1837
## year_of_observation2000                -8.7678e-01  6.8490e-02 -12.8016
## year_of_observation2001                -9.3025e-01  6.8896e-02 -13.5023
## year_of_observation2002                -9.7436e-01  7.0564e-02 -13.8082
## year_of_observation2003                -9.9693e-01  7.1519e-02 -13.9395
## year_of_observation2004                -9.7953e-01  7.3202e-02 -13.3811
## factor(blood_alcohol_limit)0.08        -4.5393e-02  1.6517e-02  -2.7484
## factor(per_se_laws)1                   -2.1973e-02  1.5010e-02  -1.4639
## factor(seatbelt)2                       1.9363e-02  2.2848e-02   0.8475
## factor(seatbelt)1                      -6.7146e-04  2.4832e-02  -0.0270
## factor(speed_limit_70plus)1             2.2112e-01  2.0923e-02  10.5684
## factor(graduated_drivers_license_law)1 -3.4342e-02  2.4528e-02  -1.4001
## pct_population_14_to_24                 1.7801e-02  6.6061e-03   2.6946
## unemployment_rate_log                   2.6728e-01  2.3813e-02  11.2243
## vehicle_miles_per_capita_log            1.5413e+00  4.7854e-02  32.2092
##                                         Pr(>|t|)
## (Intercept)                           < 2.2e-16 ***
## year_of_observation1981                0.045817 *
## year_of_observation1982               1.309e-10 ***
```

```
## year_of_observation1983                1.637e-15 ***
## year_of_observation1984                5.622e-12 ***
## year_of_observation1985                4.038e-14 ***
## year_of_observation1986                8.031e-11 ***
## year_of_observation1987                6.960e-13 ***
## year_of_observation1988                1.206e-12 ***
## year_of_observation1989                5.788e-16 ***
## year_of_observation1990                < 2.2e-16 ***
## year_of_observation1991                < 2.2e-16 ***
## year_of_observation1992                < 2.2e-16 ***
## year_of_observation1993                < 2.2e-16 ***
## year_of_observation1994                < 2.2e-16 ***
## year_of_observation1995                < 2.2e-16 ***
## year_of_observation1996                < 2.2e-16 ***
## year_of_observation1997                < 2.2e-16 ***
## year_of_observation1998                < 2.2e-16 ***
## year_of_observation1999                < 2.2e-16 ***
## year_of_observation2000                < 2.2e-16 ***
## year_of_observation2001                < 2.2e-16 ***
## year_of_observation2002                < 2.2e-16 ***
## year_of_observation2003                < 2.2e-16 ***
## year_of_observation2004                < 2.2e-16 ***
## factor(blood_alcohol_limit)0.08         0.006082 **
## factor(per_se_laws)1                    0.143489
## factor(seatbelt)2                       0.396917
## factor(seatbelt)1                       0.978432
## factor(speed_limit_70plus)1            < 2.2e-16 ***
## factor(graduated_drivers_license_law)1  0.161748
## pct_population_14_to_24                 0.007149 **
## unemployment_rate_log                  < 2.2e-16 ***
## vehicle_miles_per_capita_log           < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- How are the blood alcohol variables defined? Interpret the coefficients that you estimate for this concept.

```
#coefficient of blood_alcohol_limit0.08
blood_alcohol_limit0.08_coef = mod.lm2$coefficients[26]

# The effect on the total fatalities rate, if changing blood_alcohol_limit from 0.1% to 0.08%
# Since the response variable is the log form of the total fatalities rate, we will take the exponent o
exp(blood_alcohol_limit0.08_coef) - 1
```

```
## factor(blood_alcohol_limit)0.08
##                     -0.04437858
```

Every state in the United States has laws making it illegal to operate a motor vehicle while under the influence of alcohol. In each state, the blood alcohol content (BAC) has a legal limit of 0.08% or 0.1% for ordinary, non-commercial vehicles. In this dataset, blood_alcohol_limit variable represents the BAC legal limit for each state, either 0.1 and 0.08.

In this model version, blood_alcohol_limit of 0.1 is the default value for the base model. The estimated coefficient on *blood alcohol limit 0.08* is -0.045924 and significantly different from zero at 5%. The interpretation of this coefficent is that changing the legal blood alcohol limit from 0.10% to 0.08% will cause the traffic fatalities rate to be 4.4% lower.

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

```r
#coefficient of per_se_laws
per_se_laws_coef = mod.lm2$coefficients[27]

# The effect on the total fatalities rate, if per_se_laws from 0 to 1
# Since the response variable is the log form of the total fatalities rate, we will take the exponent o
exp(per_se_laws_coef) - 1
```

```
## factor(per_se_laws)1
##          -0.02173334
```

The estimated coefficient on *per se laws* is -0.022. The negative coefficient aligns with our EDA observation). However, this coefficient is not significantly different from zero. This is a surprise, as we expect per se laws to have a meaningful impact on the total fatalities rate.

Its interpretation is that changing the per se laws from no (0) to yes (1) will cause traffic fatalities rate to decrease by 2.17%, however, this effect is not statistically significant in this model.

- Does having a primary seat belt law?

```r
#coefficient of primary seat belt
seatbelt1_coef = mod.lm2$coefficients[29]

# The effect on the total fatalities rate, if changing seatbelt law from none to primary
# Since the response variable is the log form of the total fatalities rate, we will take the exponent o
exp(seatbelt1_coef) - 1
```

```
## factor(seatbelt)1
##     -0.0006712344
```

The estimated coefficient on *primary seatbelt laws* (seatbelt1) is negative (-0.0006714598). The negative sign aligns with our EDA observation, but is not significantly different from zero in this model. This is a surprise, as we expect primary seatbelt law requiremnets to have a material impact on the total fatalities rate.

Its interpretation is that changing the seatbelt law from no requirement to primary requirement will cause traffic fatalities rate to decrease marginally (0.067%), not statistically or practically significant.

5 (15 points) State-Level Fixed Effects

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

```r
# estimate the fixed effects regression with plm()
mod.fe <- plm(
  log(total_fatalities_rate) ~ year_of_observation
    + factor(blood_alcohol_limit)
    + factor(per_se_laws)
    + factor(seatbelt)
    + factor(speed_limit_70plus)
    + factor(graduated_drivers_license_law)
    + pct_population_14_to_24
    + unemployment_rate_log
    + vehicle_miles_per_capita_log
    + factor(state),
  data = df,
  index = c("state", "year_of_observation"),
  model = "within",
```

```
  effect="individual"
)

#summary(mod.fe)
# print summary using robust standard errors
coeftest(mod.fe, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##                                          Estimate Std. Error t value  Pr(>|t|)
## year_of_observation1981                -0.0620639  0.0168586 -3.6814  0.000243
## year_of_observation1982                -0.1341301  0.0180606 -7.4267 2.203e-13
## year_of_observation1983                -0.1664768  0.0216291 -7.6969 3.051e-14
## year_of_observation1984                -0.2124940  0.0213503 -9.9528 < 2.2e-16
## year_of_observation1985                -0.2379207  0.0262729 -9.0557 < 2.2e-16
## year_of_observation1986                -0.2011135  0.0336844 -5.9705 3.172e-09
## year_of_observation1987                -0.2472163  0.0383805 -6.4412 1.757e-10
## year_of_observation1988                -0.2776718  0.0480131 -5.7832 9.499e-09
## year_of_observation1989                -0.3510217  0.0535312 -6.5573 8.349e-11
## year_of_observation1990                -0.3606396  0.0586036 -6.1539 1.051e-09
## year_of_observation1991                -0.3984518  0.0627482 -6.3500 3.124e-10
## year_of_observation1992                -0.4587240  0.0666755 -6.8799 9.945e-12
## year_of_observation1993                -0.4755859  0.0678098 -7.0135 4.014e-12
## year_of_observation1994                -0.5087505  0.0670278 -7.5901 6.711e-14
## year_of_observation1995                -0.5089585  0.0725202 -7.0182 3.888e-12
## year_of_observation1996                -0.5600408  0.0752837 -7.4391 2.014e-13
## year_of_observation1997                -0.5828591  0.0770172 -7.5679 7.898e-14
## year_of_observation1998                -0.6356498  0.0772402 -8.2295 5.183e-16
## year_of_observation1999                -0.6510148  0.0789708 -8.2437 4.634e-16
## year_of_observation2000                -0.6816718  0.0792438 -8.6022 < 2.2e-16
## year_of_observation2001                -0.6497400  0.0831842 -7.8109 1.301e-14
## year_of_observation2002                -0.6105826  0.0810339 -7.5349 1.005e-13
## year_of_observation2003                -0.6127145  0.0835637 -7.3323 4.329e-13
## year_of_observation2004                -0.6504450  0.0875255 -7.4315 2.127e-13
## factor(blood_alcohol_limit)0.08        -0.0048883  0.0176342 -0.2772  0.781673
## factor(per_se_laws)1                   -0.0554823  0.0163076 -3.4022  0.000692
## factor(seatbelt)2                       0.0046718  0.0162378  0.2877  0.773621
## factor(seatbelt)1                      -0.0413988  0.0248524 -1.6658  0.096035
## factor(speed_limit_70plus)1             0.0727041  0.0222040  3.2744  0.001091
## factor(graduated_drivers_license_law)1 -0.0311878  0.0195683 -1.5938  0.111265
## pct_population_14_to_24                 0.0192910  0.0105815  1.8231  0.068556
## unemployment_rate_log                  -0.1940177  0.0235478 -8.2393 4.799e-16
## vehicle_miles_per_capita_log            0.6678486  0.1374350  4.8594 1.345e-06
##
## year_of_observation1981                ***
## year_of_observation1982                ***
## year_of_observation1983                ***
## year_of_observation1984                ***
## year_of_observation1985                ***
## year_of_observation1986                ***
## year_of_observation1987                ***
## year_of_observation1988                ***
## year_of_observation1989                ***
```

```
## year_of_observation1990                ***
## year_of_observation1991                ***
## year_of_observation1992                ***
## year_of_observation1993                ***
## year_of_observation1994                ***
## year_of_observation1995                ***
## year_of_observation1996                ***
## year_of_observation1997                ***
## year_of_observation1998                ***
## year_of_observation1999                ***
## year_of_observation2000                ***
## year_of_observation2001                ***
## year_of_observation2002                ***
## year_of_observation2003                ***
## year_of_observation2004                ***
## factor(blood_alcohol_limit)0.08
## factor(per_se_laws)1                   ***
## factor(seatbelt)2
## factor(seatbelt)1                      .
## factor(speed_limit_70plus)1            **
## factor(graduated_drivers_license_law)1
## pct_population_14_to_24                .
## unemployment_rate_log                  ***
## vehicle_miles_per_capita_log           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#coefficient of blood_alcohol_limit0.08
blood_alcohol_limit0.08_coef_fe = mod.fe$coefficients[25]
#blood_alcohol_limit0.08_coef_fe = mod.fe$coefficients[1]

# The effect on the total fatalities rate, if changing blood_alcohol_limit from 0.1% to 0.08%
# Since the response variable is the log form of the total fatalities rate, we will take the exponent o
exp(blood_alcohol_limit0.08_coef_fe) - 1
```

```
## factor(blood_alcohol_limit)0.08
##                    -0.004876388
```

- What do you estimate for coefficients on the blood alcohol variables? How do the coefficients on the blood alcohol variables change, if at all? > In this State-level fixed effects model, the estimated coefficient on blood alcohol limit 0.08 is -0.0236, less negative than the estimated coefficient of -0.045924 in the expanded linear model (mod.lm2).

  Additionally, the expanded linear model (mod.lm2) showed that this variable is statistically significant, but the State-level fixed effects model shows that this variable's coefficient is not significantly different from zero using the robust standard errors.

  Its interpretation is that changing the blood alcohol limit from 0.10% to 0.08% causes traffic fatalities rate to decrease marginally by 2.3%, which is less than the 4.4% decrease estimated by the expanded liniear model.

- What do you estimate for coefficients on per se laws? How do the coefficients on per se laws change, if at all?

```
#coefficient of per_se_laws
per_se_laws_coef_fe = mod.fe$coefficients[26]
#per_se_laws_coef_fe = mod.fe$coefficients[2]
```

```
# The effect on the total fatalities rate, if per_se_laws from 0 to 1
# Since the response variable is the log form of the total fatalities rate, we will take the exponent o
exp(per_se_laws_coef_fe) - 1
```

```
## factor(per_se_laws)1
##          -0.05397123
```

The estimated coefficient on *per se laws* is -0.086 and significantly different from zero at 0% significance level, using the robust standard errors. In the prior expanded linear model (mod.lm2), this variable has a coefficient of -0.022 and was not statistically significant.

Its interpretation is that change per se laws from no (0) to yes (1) will cause traffic fatalities rate to decrease by 8.2%. This is significant both statistically and practically. Compared to the expanded model, the impact magnitude of per se laws estimated by the fixed effects model is much larger and is sensible.

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

```
#coefficient of primary seat belt
seatbelt1_coef_fe = mod.fe$coefficients[28]
#seatbelt1_coef_fe = mod.fe$coefficients[4]

# The effect on the total fatalities rate, if changing seatbelt law from none to primary
# Since the response variable is the log form of the total fatalities rate, we will take the exponent o
exp(seatbelt1_coef_fe) - 1
```

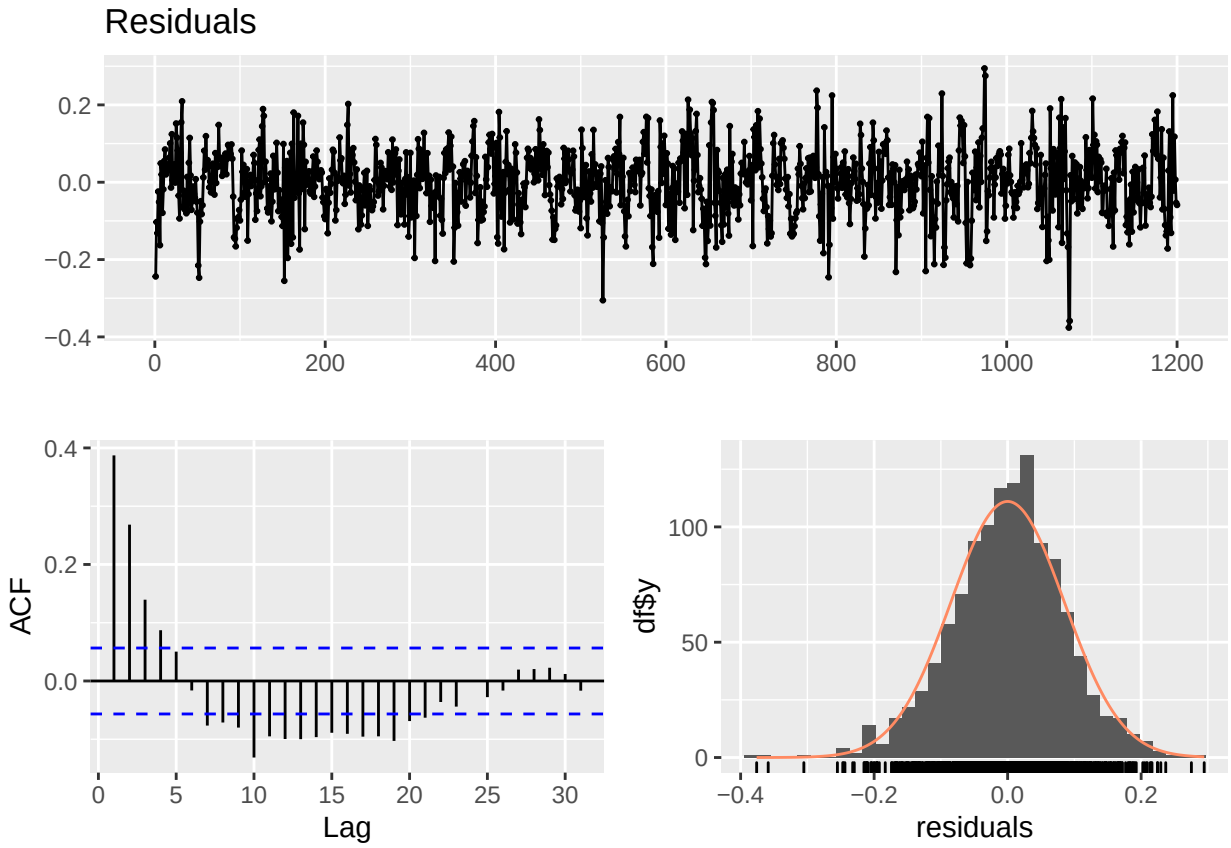
```
## factor(seatbelt)1
##       -0.04055359
```

The estimated coefficient on *primary seatbelt laws* is -0.09 and has strong evidence that it is difference from zero at 0.05% significance level, using the robust standard errors. The prior expanded model (mod.lm2) noted that this variable was not statistically significant, which was not aligned with our experience.

Its interpretation is that chaning seatbelt law requirement from none to primary will cause traffic fatalities rate to decrease by 8.7%, which is significant both statistically and practically. This effect makes much more sense than the marginal effect estimated by the expanded model.

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

```
forecast::checkresiduals(mod.fe)
```



For the State-level Fixed Effects models, the "within" setting eliminated the omitted variable bias for time-invariant fixed effects and improved the model performance over the linear models. The linear models have omitted variable bias which caused the parameter estimates to be biased.

We performed residuals diagnostic analysis for all models. For the State-level Fixed Effects model, the model residuals have a conditional mean around zero, with a normal distribution. Since this model uses the within setting to fit a fixed effect model at the state level, the effect of unobserved individual heterogeneity is eliminated. The conditinoal mean of zero assumption is met.

Variance fluctulations are much improved compared to the linear models. The ACF plot for the model residuals shows autocorrelation but decays quickly (this is much improved compared to the significant but non-decaying autocorrelation in the expanded model). We used the robust standard errors to account for the group heteroskedasticity and serial correlation by allowing correlation across time in groups.

Based on the residuals analysis, the State-level Fixed Effects model will generate more reliable estimates than the linear models.

Also, the parameter estimates' sign and magnitude in the State-level Fixed Effects model make much sense than the expanded model, based on our EDA work and our experience.

- What assumptions are needed in each of these models? > Fixed Effect Model Assumptions: -**Linearity**: the model is linear in parameters -**i.i.d.** : The observations are independent across individual states but not necessarily across time. This is guaranteed through random sampling. -**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.
  -**Zero conditional means (strict exogeneity)**
- Are these assumptions reasonable in the current context? > -**Linearity**: This assumption is met since each of the explanatory variable is shown to have a linear relationship with the response variable

in the models. -**i.i.d.** : This assumption is met since the cross-sectional variables measured across state/year on a randomized population. Since we have 48 different states, we assume that each state has independent legislative procedures and has the state-specific economics and demographic situations. -**Indentifiability**: In our EDA work, we excluded the variables that have perfect collinearity. Also any variables with perfect collinearity would be dropped out from the model. So this condition is met. Using robust standard error on the fitted model, each explanatory variables resulted in a less than 1 standard deviation, which, upon raising it to the power of two, resulted in low non-zero variance. This suggests there are not too many extreme values in the explanatory variables. Therefore, this condition is met. -**Zero conditional means (strict exogeneity)**: Since we're using within estimator to fit a fixed effect model on the state, the effect of unobserved individual heterogeneity is eliminated. Thus this condition is met by the State-level Fixed Effect Model. This condition is not met by the two linear models.

6 (10 points) Consider a Random Effects Model

- Please state the assumptions of a random effects model, and evaluate whether these assumptions are met in the data.

  Random Effect Model Assumptions: - All assumptions under Fixed Effect model - The unobservbed effect term ai is independent of all explanatory variables in all the time periods in the model

- 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.

  The first set of assumption is already met under the fixed effect analsysis. To test for the last assumption, we will perform Hausman Test for Fixed vs. Random Effects. In the null hypothesis, both estimates are consistent but only the random effects model estimates are efficient (minimum variance). The alternative hypothes is that only the coefficients of the fixed effects model are consistent, and the coefficients of the random effects model are not consistent. The alternative hypothesis means that there is correlation between residuals and predictors.

  If Hausman test p-value is less than the significance level, we reject the null hypothesis and deem that the fixed-effects model should be preferred instead of a random-effects model.

  Our test result below shows a p-value of 2.649e-7, which is significantly lower than the 95% confidence level, alpha = 0.05. Thus, there are sufficient evidence to reject the null hypothesis that a random effects model is appropriate, suggesting that we should use the fixed effect models. The random effects model is not likely to be consistent in this case, which means that the paramater estimates won't get closer to the the true parameter values even with a large sample size.

  Based on the above notes, we don't think it is appropriate to use the random effects model for this dataset.

```
re.model <- plm(
  log(total_fatalities_rate) ~ year_of_observation
    + factor(blood_alcohol_limit)
    + factor(per_se_laws)
    + factor(seatbelt)
    + factor(speed_limit_70plus)
    + factor(graduated_drivers_license_law)
    + pct_population_14_to_24
    + unemployment_rate_log
    + vehicle_miles_per_capita_log,
  data = df,
  index = c("state", "year_of_observation"),
  model = "random"
```

```
)

phtest(mod.fe, re.model)
```

```
##
##  Hausman Test
##
## data:  log(total_fatalities_rate) ~ year_of_observation + factor(blood_alcohol_limit) +  ...
## chisq = 90.825, df = 33, p-value = 2.649e-07
## alternative hypothesis: one model is inconsistent
```

```
#summary(re.model)
```

While we don't think it is appropriate to use the random effect model, we included the random effect model in the comparison below for information only.

```
#calculate the robust standard errors for the models
cov_exp        <- vcovHC(mod.lm2, type = "HC1")
robust_se_exp    <- sqrt(diag(cov_exp))

cov_fe        <- vcovHC(mod.fe, type = "HC1")
robust_se_fe    <- sqrt(diag(cov_fe))

cov_rm        <- vcovHC(re.model, type = "HC1")
robust_se_rm    <- sqrt(diag(cov_rm))

#compare the three models, show robust standard errors in the comparison
stargazer(mod.lm2, mod.fe, re.model,
          type="latex",
          omit.stat=c("adj.rsq","f"),
          se = list(robust_se_exp, robust_se_fe, robust_se_rm),
          column.labels = c("expand model", "Fixed Effects","Random Effects"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Apr 17, 2023 - 10:33:11 PM

- 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?

  Since the null hypothesis for the Hausman test is rejected, the last assumption is not met. Thus we should not use the random effect model to estimate the coefficients since the estimates will not likely to be consistent.

  The random effects model assumes the individual unobserved heterogeneity is uncorrelated with the independent variables. However, the Hausman test shows strong evidence that correlation exists. Therefore, using the random effects model for this dataset could produce bias in parameters estimates. This bias does not arise in the fixed effects model, because the fixed effects model assumes that the omitted fixed effects can be correlated with the other variables. Fixed effects models remove the effect of time-invariant variables so we can assess the net effect of the other predictors on the response variable (total traffic fatalities rate).

  Given the Hausman test result, the random effects model is also not likely to be efficient, which means its variance would not be minimum or stable causing the standard errors to be biased and test statistics to be not reliable.

Table 1:

	Dependent variable:		
	log(total_fatalities_rate)		
	OLS	*panel linear*	
	expand model	Fixed Effects	Random Effects
	(1)	(2)	(3)
year_of_observation1981	−0.091**	−0.062***	−0.063***
	(0.046)	(0.017)	(0.017)
year_of_observation1982	−0.295***	−0.134***	−0.141***
	(0.045)	(0.018)	(0.018)
year_of_observation1983	−0.352***	−0.166***	−0.176***
	(0.044)	(0.022)	(0.021)
year_of_observation1984	−0.306***	−0.212***	−0.219***
	(0.044)	(0.021)	(0.022)
year_of_observation1985	−0.344***	−0.238***	−0.246***
	(0.045)	(0.026)	(0.026)
year_of_observation1986	−0.321***	−0.201***	−0.211***
	(0.049)	(0.034)	(0.033)
year_of_observation1987	−0.357***	−0.247***	−0.259***
	(0.049)	(0.038)	(0.038)
year_of_observation1988	−0.367***	−0.278***	−0.290***
	(0.051)	(0.048)	(0.047)
year_of_observation1989	−0.452***	−0.351***	−0.365***
	(0.055)	(0.054)	(0.053)
year_of_observation1990	−0.512***	−0.361***	−0.378***
	(0.059)	(0.059)	(0.058)
year_of_observation1991	−0.628***	−0.398***	−0.419***
	(0.060)	(0.063)	(0.062)
year_of_observation1992	−0.734***	−0.459***	−0.482***
	(0.064)	(0.067)	(0.066)
year_of_observation1993	−0.725***	−0.476***	−0.498***
	(0.062)	(0.068)	(0.067)
year_of_observation1994	−0.711***	−0.509***	−0.530***
	(0.062)	(0.067)	(0.067)
year_of_observation1995	−0.689***	−0.509***	−0.530***
	(0.064)	(0.073)	(0.072)
year_of_observation1996	−0.814***	−0.560***	−0.584***
	(0.064)	(0.075)	(0.075)
year_of_observation1997	−0.822***	−0.583***	−0.607***
	(0.065)	(0.077)	(0.076)

From a practical perspective, fixed effects models are relatively straighforward. However, random effects models require additional mathematical assumptions with added-complexity, which strenthens the argument of not using the random effects model in this case.

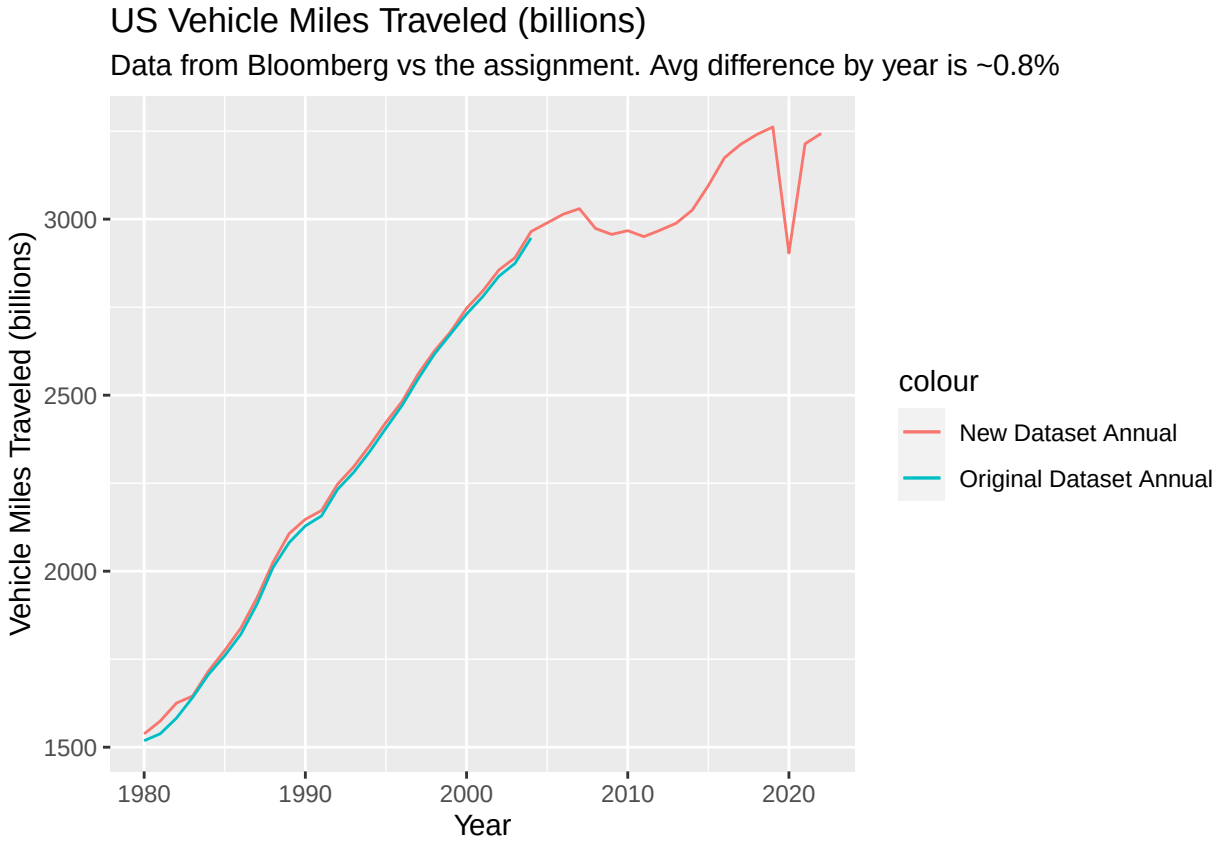
7 (10 points) Model Forecasts

We have collected the United States Motor Vehicle Miles Traveled Total (Millions) data from the US Department of Transformation from Jan 1980 to January 2023. This data is available at here [https://www.fhwa.dot.gov/policyinformation/statistics/2021/vm202.cfm].

We have downloaded the data that is compiled by Bloomberg (under ticker: VMTDVCLE Index) and saved it in the file `data/VMTDVCLE.csv`.

```
#load the monthly new panel data
vehicle_miles_month <- read.csv("./data/VMTDVCLE.csv",
  sep = ",",
  skip = 5
) %>%
  as_tibble() %>%
  mutate(
    vehicle_miles = PX_LAST / 1000, # convert to billions
    Date = as.Date(Date, format = "%m/%d/%Y"),
    year = year(Date),
    month = month(Date)
  ) %>%
  select(year, month, vehicle_miles) %>%
  arrange(year, month)

#create the annual data by grouping the new monthly data by year
vehicle_miles_annual <- vehicle_miles_month %>%
  filter(year < 2023) %>% # remove the 2023 data since it's not complete
  group_by(year) %>%
  summarise(total = sum(vehicle_miles))

# create the annual data by grouping the original data by year
df_annual <- df %>%
  group_by(year) %>%
  summarise(total = sum(vehicle_miles))

# calculate the avg difference between the two data
#
combined_mile_annual <- vehicle_miles_annual %>%
  filter(year <= 2004) %>%
  dplyr::rename(
    "new" = "total",
    "new_year" = "year") %>%
  # rename(new = total, new_year = year) %>%
  cbind(df_annual) %>%
  mutate(diff = (new - total) / total)

avg_diff <- round(mean(combined_mile_annual$diff) * 100, 1)


# visualize the new vs the old vehicle miles data
```

```r
ggplot() +
  geom_line(
    data = vehicle_miles_annual,
    aes(x = year, y = total, color = "New Dataset Annual")
  ) +
  geom_line(
    data = df_annual,
    aes(x = year, y = total, color = "Original Dataset Annual")
  ) +
  labs(
    title = "US Vehicle Miles Traveled (billions)",
    subtitle = paste0(
      "Data from Bloomberg vs the assignment. Avg difference by year is ~",
      avg_diff, "%"
    ),
    x = "Year",
    y = "Vehicle Miles Traveled (billions)"
  )
```



We can see the new vehicle milage data is very similar to the old data, and the new data is always higher than the old data. The average difference between the two data is 0.8%.

This is expected as the new vehicle miles data from Bloomberg include the entire United States, while the old data only include 48 states.

But we can believe the new data is a good replacement for the old data in our analysis.

```
vehicle_miles_month %>%
  filter(year >= 2018 & year < 2023) %>%
  group_by(month) %>%
  summarise(
    year = year,
    chg_vs_2018 = round((vehicle_miles / vehicle_miles[year == 2018] - 1) * 100, 1)
    # vehicle_miles_yoy = round((vehicle_miles / lag(vehicle_miles) - 1) * 100, 1) # yoy
  ) %>%
  filter(year > 2018) %>% # remove the first row
  spread(year, chg_vs_2018) %>%
  knitr::kable()
```

month	2019	2020	2021	2022
1	1.7	6.6	-8.1	-4.3
2	1.8	6.6	-6.5	3.5
3	0.6	-16.3	-0.5	2.5
4	-0.6	-39.1	-5.8	-4.4
5	2.1	-22.1	0.2	1.5
6	-0.5	-11.4	1.5	-0.2
7	0.2	-8.7	1.9	-1.5
8	2.9	-7.0	0.8	1.5
9	2.2	-3.7	4.0	5.0
10	0.9	-5.3	1.5	1.6
11	-0.1	-8.5	2.8	1.4
12	-3.2	-10.7	-3.5	-5.2

The largest decrease in driving during COVID bust is in April 2020 (-39.1% compared with April 2018), and the largest increase in driving post-COVID boom is in September 2022 (+5.0% compared with January 2018), which is technically before COVID.

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.

  Holding everything else constant, if the number of miles driven per capita increased by 5.0%, the number of traffic fatalities would increase by 5. Given the fatalities

- 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.

  Holding everything else constant, if the number of miles driven per capita decreased by 39.1%, the number of traffic fatalities would increase by -39.1. We should be cautious here since the change in miles driven per capita is very large, so the log transformation interpretation might not be accurate.

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?

```
pcdtest(mod.fe, test = "lm")
```

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
```

```
##
## data:  log(total_fatalities_rate) ~ year_of_observation + factor(blood_alcohol_limit) +     factor(p
## chisq = 2737.8, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

```
pbgtest(mod.fe, order = 2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  log(total_fatalities_rate) ~ year_of_observation + factor(blood_alcohol_limit) +  ...
## chisq = 203.82, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, the standard errors or uncertainty of the estimators will be underestimated and statistical inferences are not reliable. However, serial correlation does not cause bias in the regression coefficient estimates of the estimators.

In this case, when we apply the Breusch Pagan Test on homoskedasticity to the FE model, we obtained a p-value of 2.2e-16, which is significantly less than 0.05, suggesting that there is sign of heteroskedasticity from the idiosyncratic errors. We also performed the Breusch-Godfrey test for serial correlation and obtained a p-value of 2.2e-16, which is significantly less than 0.05, suggesting that we have do have serially correlation in idiosyncratic errors. In the ACF plot, we can also visually detect the autocorrelation in the model residuals.

Therefore, the FE model has serial correlation issues and standard errors are underestimated. We should use the robust standard errors to address heteroskedasticity and serial correlation issues by allowing correlation across time in groups.