W271 Lab 3

```
# Insert the function to *tidy up* the code when they are printed out
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
# Start with a clean R environment
rm(list = ls())
```

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
# str(data)
# desc
```

Observation: 1200 observations, include 48 states from 1980 to 2004.

(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`; By checking the data structure and detail, we found speed-limit-related variables are percentage ranging from 0 to 1. Therefore, we produced a new variable, speed_limit, where speed limit was determined by the percentage. For example, if the speed limit was 55 for 0.542(>0.5) for the year, we used 55 for that year. But if the spped limit was 55 for 0.5 and 65 for 0.5, we used higher tier 65.

```
# # Convert numbers to state abru.
state_ls <- c("al", "az", "ar", "ca", "co", "ct", "de", "fl", "ga", "id", "il", "in",
              "ia", "ks", "ky", "la", "me", "md", "ma", "mi", "mn", "ms", "mo", "mt",
              "ne", "nv", "nh", "nj", "nm", "ny", "nc", "nd", "oh", "ok", "or", "pa",
              "ri", "sc", "sd", "tn", "tx", "ut", "vt", "va", "wa", "wv", "wi", "wy")
state_num_ls <- c(1, 3:8, 10:11, 13:51)
stateNumToName <- data.frame(state_num = state_num_ls, state = state_ls)
data$state_num <- data$state
data$state <- NULL
data <- merge(data, stateNumToName, by="state_num", all.x = T)
data$state_num <- NULL

# Produce new variable called speed limit
df <- data %>% mutate(speed_limit=case_when(sl55 > 0.5 ~ "55",
                                            sl65 >= 0.5 ~ "65",
                                            sl70 >= 0.5 ~ "70",
                                            sl75 >= 0.5 ~ "75",
                                            slnone > 0.5 ~ NA))
df$speed_limit = factor(df$speed_limit)

# Drop unused columns to make data cleaner
df <- df %>% dplyr::select(-c("sl55", "sl65", "sl70", "sl75", "slnone"))
```

- Produce a new variable, called `year_of_observation` that re-encodes the data that is in `d80`, `d81`

Observation: By using table function, we found each state have 25 year's observation, no missing year. So we rename column year as year_of_observation and dropped binary indicators, such as 'd80','d81'... till 'd04'.

```
table(df$year)
```

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

```
table(df$state)
```

```
##
## al ar az ca co ct de fl ga ia id il in ks ky la ma md me mi mn mo ms mt nc nd
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
## ne nh nj nm nv ny oh ok or pa ri sc sd tn tx ut va vt wa wi wv wy
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

```
df <- df %>% dplyr::rename(year_of_observation=year)
df$year_of_observation = factor(df$year_of_observation)

#drop unused columns to make data cleaner
df<-df %>% dplyr::select(-c(which(colnames(df)=="d80"):which(colnames(df)=="d04")))
```

- Produce a new variable for each of the other variables that are one-hot encoded (i.e. `bac*` variable

Observation:We found two variables related to blood alcohol limits, which have similar situation as speed limit. So we will apply similar transformation strategy to blood alchol limits.
Other variables, including `miniage`, `zerotol`, `gdl`, and, `perse`, also have some non-integer values among their ranges. Therefore, we decided to recode them into two categories variables. For example, for minimal drinking age, we decoded any age smaller than 20 is coded as 18, larger or equals to 20 is coded as 21.

```
# Renames the categories of categorical variables
df<-df%>%mutate(blood_alcohol_limit=case_when(bac08>0.5 ~ "8",
                                              bac10 >= 0.5 ~ "10",
                                              TRUE ~ "NA"),
                minimal_drink_age=case_when(minage < 20 ~ "18",
                                            minage >= 20 ~ "21"),
                zero_tolerance_law=case_when(zerotol < 0.5 ~"0",
                                             zerotol >= 0.5 ~ "1"),
                graduate_driver_law=case_when(gdl < 0.5 ~"0",
                                              gdl >= 0.5~"1"),
                admin_license_revoke=case_when(perse < 0.5~"0",
                                               perse >= 0.5~"1"))

# Change the data type to factor
df$blood_alcohol_limit=factor(df$blood_alcohol_limit)
df$minimal_drink_age=factor(df$minimal_drink_age)
df$zero_tolerance_law=factor(df$zero_tolerance_law)
df$graduate_driver_law=factor(df$graduate_driver_law)
df$admin_license_revoke=factor(df$admin_license_revoke)
df$state=factor(df$state)
df$seatbelt=factor(df$seatbelt)
df$sbprim=factor(df$sbprim)
df$sbsecon=factor(df$sbsecon)

# Drop unused columns to make data cleaner
df<-df%>% dplyr::select(-c("bac08", "bac10","minage","zerotol","gdl","perse"))
```

- Rename these variables to sensible names that are legible to a reader of your analysis. For example,

```
# Rename variables
df <- df %>%
  dplyr::rename(total_fatalities = totfat,
         nighttime_fatalities = nghtfat,
         weekend_fatalities = wkndfat,
         total_fatalities_per_100million_miles = totfatpvm,
         nighttime_fatalities_per_100million_miles = nghtfatpvm,
         weekend_fatalities_per_100million_miles = wkndfatpvm,
         total_fatality_rate = totfatrte,
         nighttime_fatality_rate  = nghtfatrte,
         weekend_fatality_rate = wkndfatrte,
         vehicle_miles_billion=vehicmiles,
         unemp_rate=unem,
         percent_pop_14_24=perc14_24,
         vehicle_miles_per_capita=vehicmilespc)
```

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?

Answer:
The data came from Dr. Donald G Freeman's study in paper, Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws(Donald G. Freeman, 2007. "Drunk Driving Legislation And Traffic Fatalities: New Evidence On Bac 08 Laws," Contemporary Economic Policy, Western Economic Association International, vol. 25(3), pages 293-308, July.), which primarily focus on examine the effectiveness of blood alcohol content(BAC) laws in reducing traffic fatalities. The fatality data is collected from the Fatality Analysis Reporting System(FARS), which is created by the National Highway Traffic Safety Administration(NHTSA). The economic and demographic data was collected through US Bureau of Labour Statistics and Bureau of the Census.
This dataset is composed by 1200 observations, which can be divided into repeated measurements for 48 states in 25 years from 1980 to 2004.Each observation represents one state's measurements from a specific year. All states were index by alphabetic order and presented by their index number instead of name. This data have 56 columns. Besides a few economics and demographic variables, it also provided some measurements about speed limit, seat belt, and drinking driving.
Considering the total number of states is a constant(n=48, excluding Alaska and Hawaii), the data represents the entire population. The dependent variable of interest, total_fatality_rate, is defined as the number of fatalities per 100,000 population.

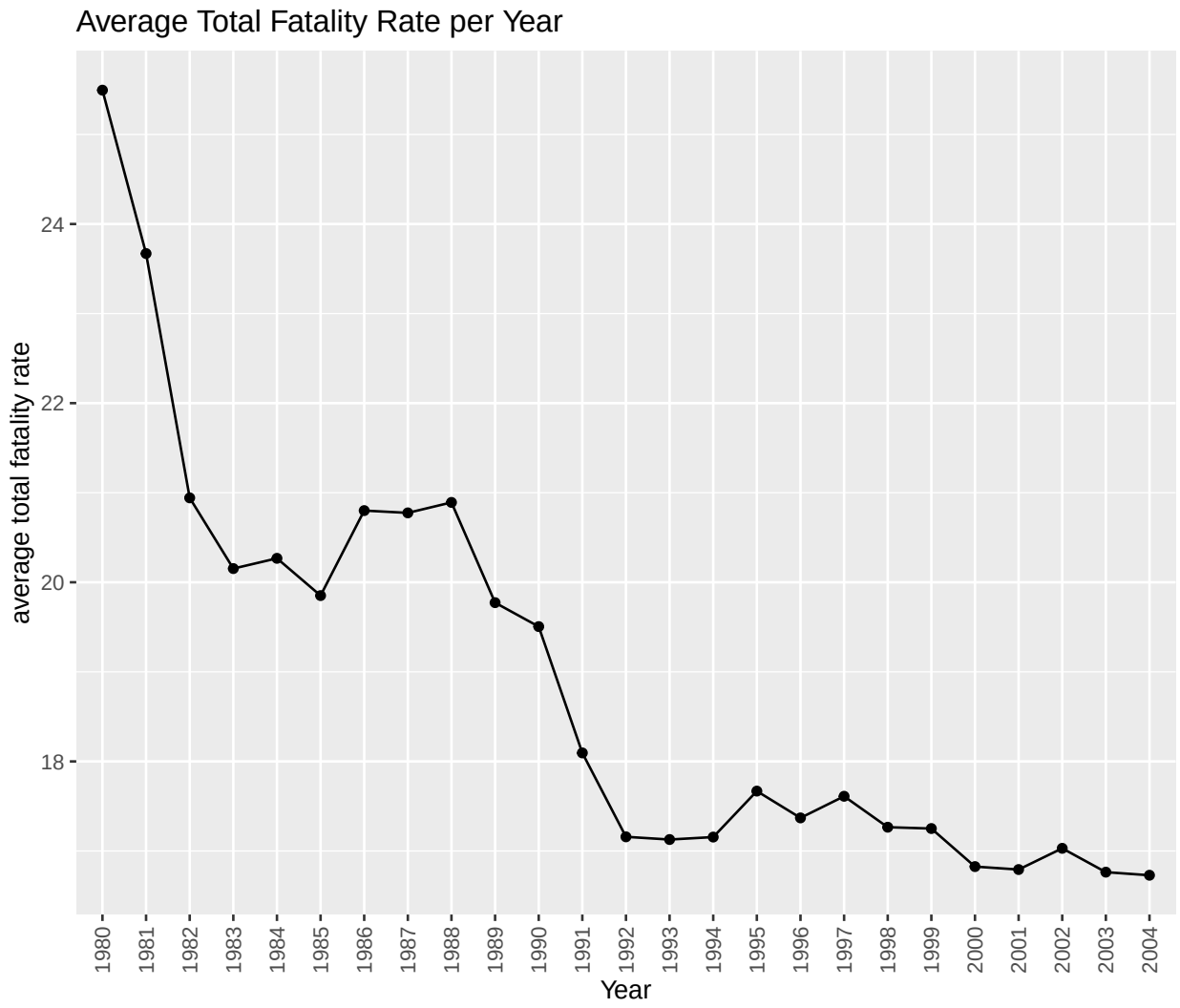
3. (20 points) Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable `total_fatalities_rate` and the potential explanatory variables. Minimally, this should include:
    - How is the our dependent variable of interest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

Answer: As discussed above, the dependent variable, total_fatalities_rate, is calculated by dividing the FARS traffic fatalities count by the state population and then standardized to per 100,000 people.

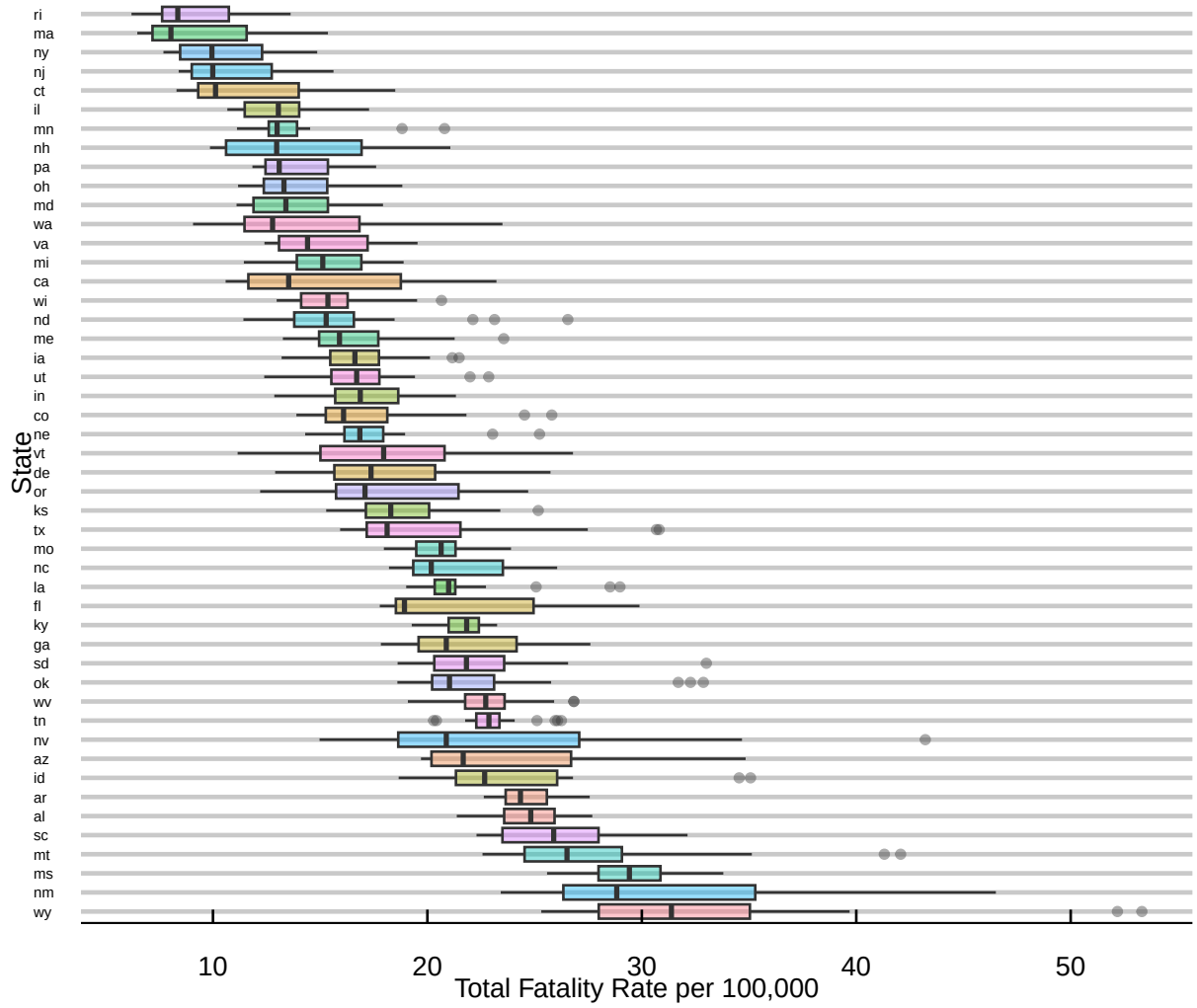
```
- What is the average of `total_fatalities_rate` in each of the years in the time period covered in this
```

Answer: The time series of average of total fatalities rate shows a general decreasing trend with 4 stages. The first stage has higher decreasing rate in early days between 1980 to 1982,followed by second stage,a stable period, between 1982 to 1988, then third stage, another decreasing period, from 1988 to 1992, and finally the last stage, a relative stable period with fluctuations, from 1992 to 2004.

```r
# Average of total fatalities rate in each years
avg_total_fatalities_rate <- df %>%
    group_by(year_of_observation) %>%
    summarise(avg_total_fatalities_rate = mean(total_fatality_rate))

avg_total_fatalities_rate %>%
  ggplot(aes(x = year_of_observation, y = avg_total_fatalities_rate,group=1)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Total Fatality Rate per Year",
       x = "Year",
       y = "average total fatality rate"
    ) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

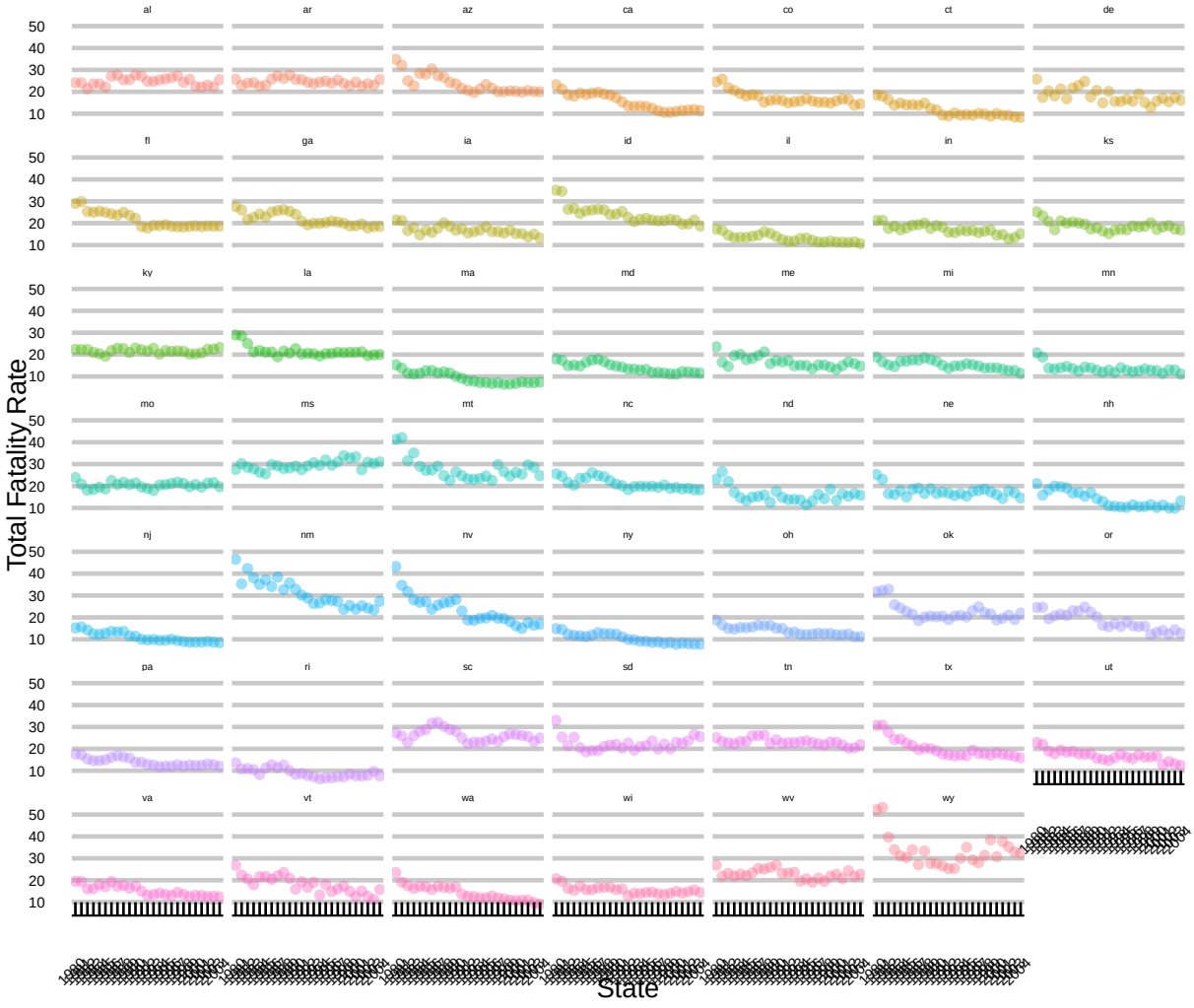



```
# Boxplot of total fatality rate per state
df %>%
  ggplot(aes(reorder(state, desc(total_fatality_rate)),total_fatality_rate, fill=state))+
  geom_boxplot(alpha=0.4)+
  theme_economist_white(gray_bg=F)+
  theme(legend.position="none", axis.text.y=element_text(size=6))+
  scale_y_continuous()+
  xlab("State")+
  ylab("Total Fatality Rate per 100,000")+
  coord_flip()
```

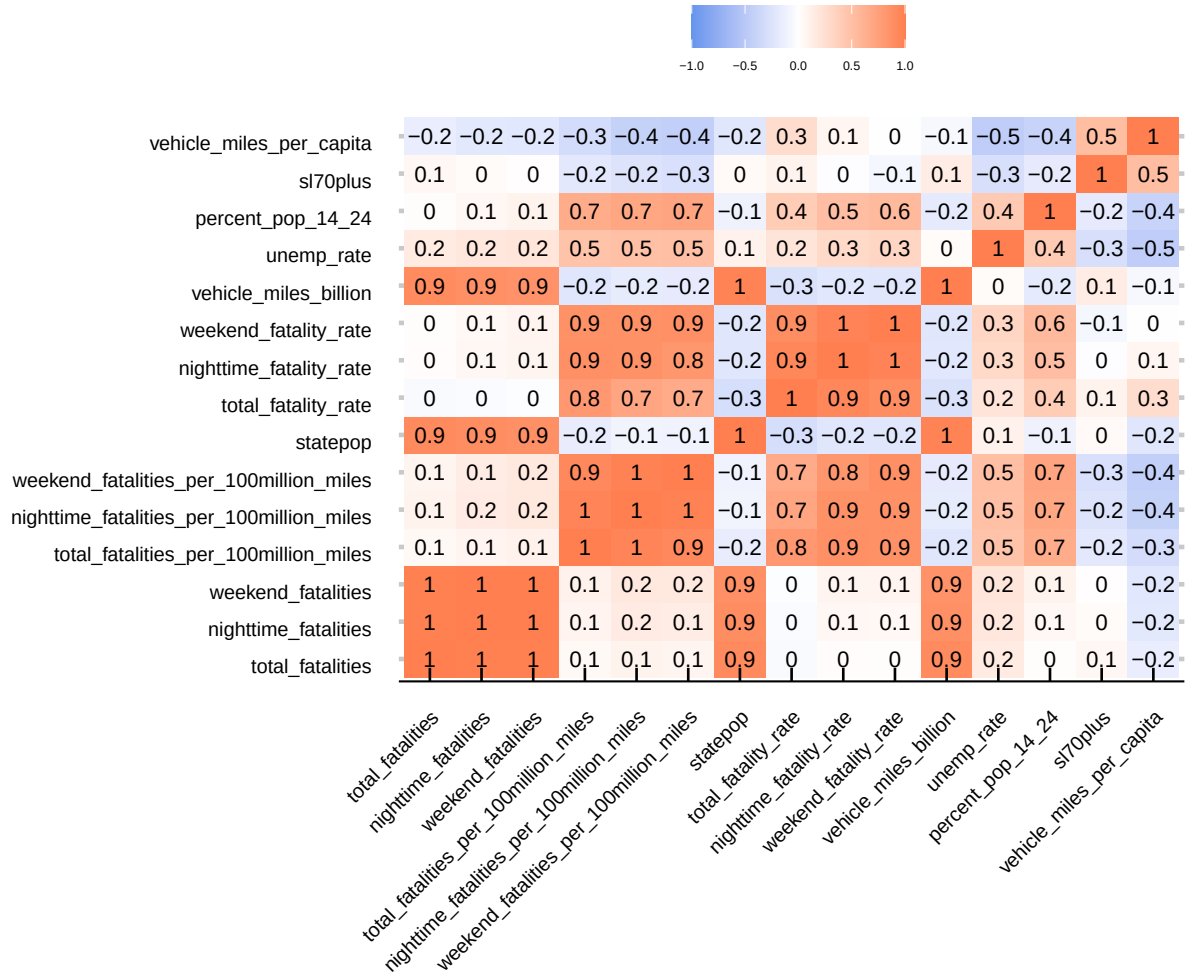



```
# Scatter plot of total fatality rate per state
df %>%
  ggplot(aes(year_of_observation, total_fatality_rate, color=state))+
  geom_point(alpha=0.4)+
  geom_smooth(method="lm")+
  facet_wrap(~state)+
  theme_economist_white(gray_bg = F)+
  theme(legend.position="none",
        axis.text.x=element_text(size=6, angle=45),
        axis.text.y=element_text(size=6))+
  theme(strip.text = element_text(size=4))+
  scale_y_continuous()+
  xlab("State")+
  ylab("Total Fatality Rate")
```

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

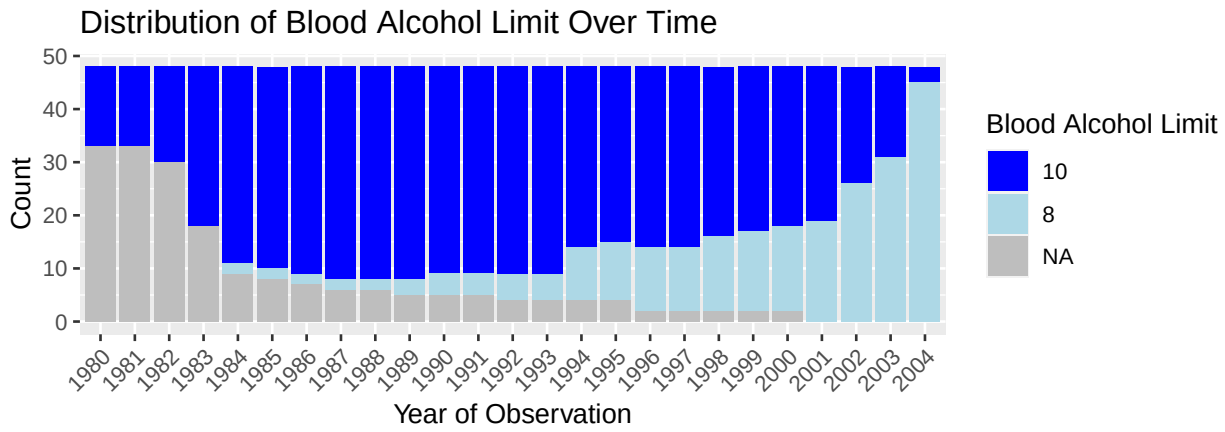



```
#Correlation among numeric independent variables and DV
col_numeric <- colnames(select_if(df, is.numeric))
df[col_numeric]%>%
  cor() %>%
  melt() %>%
  ggplot(aes(Var1,Var2, fill=value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 1)), size = 3) +
  theme_economist_white(gray_bg = F)+
  theme(legend.title = element_blank(),legend.text=element_text(size=5),
        axis.text.x = element_text(size=8,angle=45, hjust=1, vjust=1),
        axis.text.y = element_text(size=8,hjust=1, vjust=1))+
  scale_fill_gradient2(low="cornflowerblue",high="coral",mid="white",
                       midpoint=0,limit=c(-1,1))+
  xlab("")+
  ylab("")
```



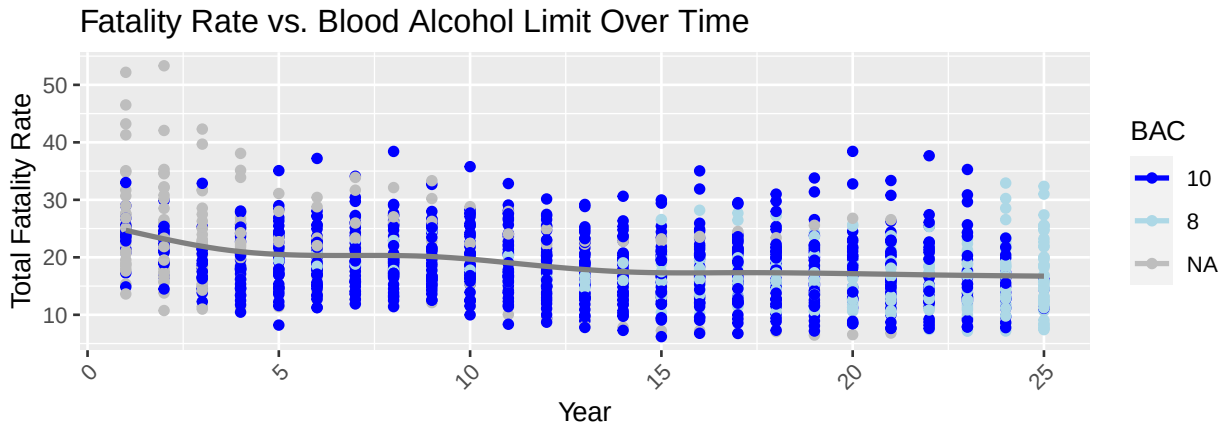

```r
# Scatter plot: linearity
lm_bac <- ggplot(df, aes(x = as.numeric(year_of_observation), y = total_fatality_rate, color = blood_al
  geom_point() +
  geom_smooth(method = "auto", se = FALSE, aes(color = "black")) +
  labs(title = "Fatality Rate vs. Blood Alcohol Limit Over Time",
       x = "Year",
       y = "Total Fatality Rate",
       color = "BAC") +
  scale_color_manual(values = c("10" = "blue", "8" = "light blue", "NA" = "grey")) +
  theme(axis.text.x = element_text(angle=45, hjust=1, vjust=1))

dist_bac <- ggplot(df, aes(x = year_of_observation, fill = blood_alcohol_limit)) +
  geom_bar(position = "stack") +  # Use geom_bar() to create a stacked bar chart
  labs(title = "Distribution of Blood Alcohol Limit Over Time",
         x = "Year of Observation",
         y = "Count",
         fill = "Blood Alcohol Limit") +
  scale_fill_manual(values = c("10" = "blue", "8" = "light blue", "NA" = "grey")) +
  theme(axis.text.x = element_text(angle=45, hjust=1, vjust=1))
```

```
dist_bac
```



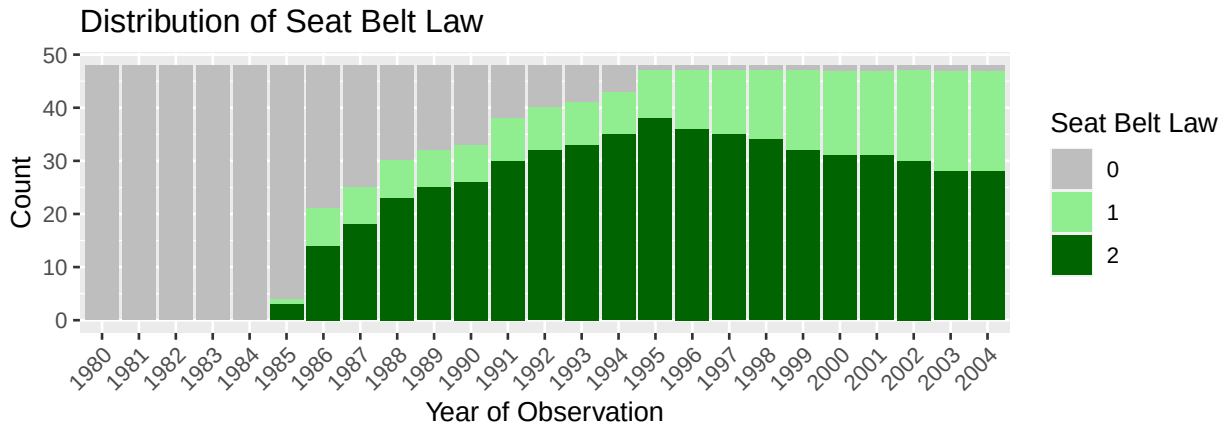

```
lm_bac
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



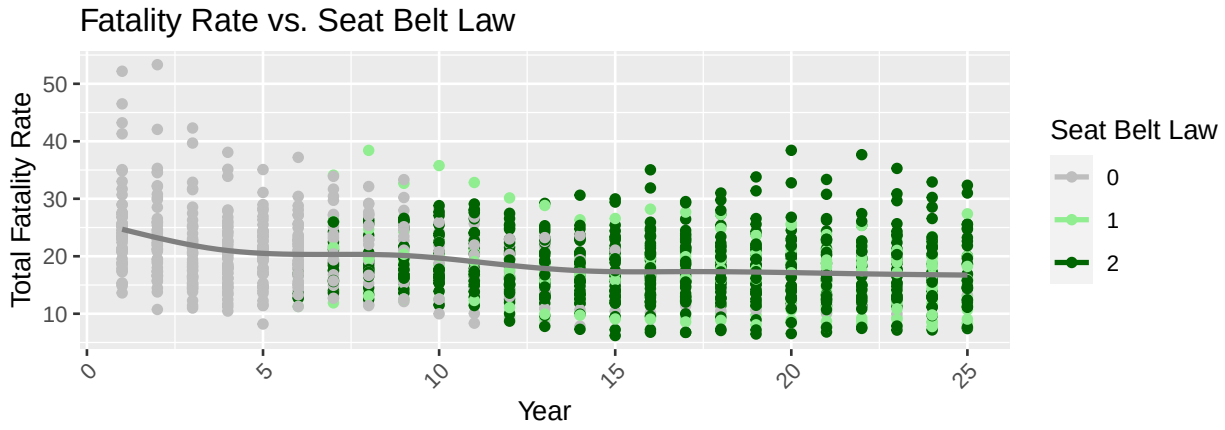

```
# Scatter plot: linearity
lm_seatbelt <- ggplot(df, aes(x = as.numeric(year_of_observation),
                              y = total_fatality_rate, color = seatbelt)) +
  geom_point() +
  geom_smooth(method = "auto", se = FALSE, aes(color = "black")) +
  labs(title = "Fatality Rate vs. Seat Belt Law",
       x = "Year",
       y = "Total Fatality Rate",
       color = "Seat Belt Law") +
  scale_color_manual(values = c("0" = "grey", "1" = "light green", "2" = "darkgreen")) +
  theme(axis.text.x = element_text(angle=45, hjust=1, vjust=1))

dist_seatbelt <- ggplot(df, aes(x = year_of_observation, fill = seatbelt)) +
  geom_bar(position = "stack") +  # Use geom_bar() to create a stacked bar chart
  labs(title = "Distribution of Seat Belt Law",
       x = "Year of Observation",
```

```
       y = "Count",
       fill = "Seat Belt Law") +
  scale_fill_manual(values = c("0" = "grey", "1" = "light green", "2" = "darkgreen")) +
  theme(axis.text.x = element_text(angle=45, hjust=1, vjust=1))

dist_seatbelt
```




```
lm_seatbelt
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Answer:
1) The box plots shows strong differences in total fatality rate across states, indicating that fixed effects are important for controlling for unobserved differences.
2) The scatter plots shows that some states had more similar fatality rate throughout entire period, some states have consistently decreasing trends, while some states have up and downs. We can observe the fatality rate decreases from year 1980 to 2004 in most of the states but the rate of decreasing is different across states.
3) The correlation plot shows that certain variables, state population and vehicle miles billion seem highly related to each other.
4) The blood alcohol concentration (BAC) distribution plot suggests that most states did not enforce blood alcohol limits in the early 1980s. Since 1985, however, there has been an increase in the number of states enforcing BAC limits of .1% or lower. The scatter plot between BAC and total fatality rate reveals a

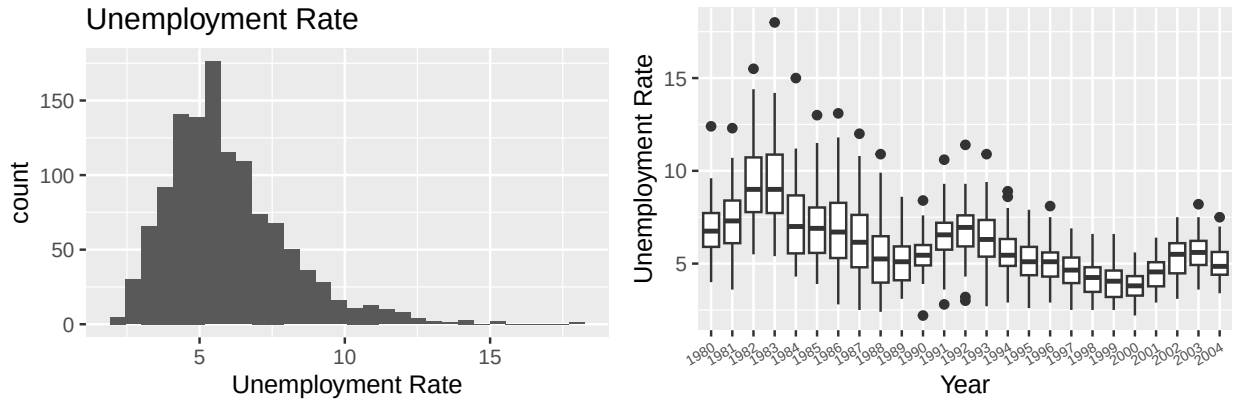
positive association between these variables. The bar plot of blood alcohol limit (BAC) indicates that more states start to utilize .08% instead of 0.1% as the blood alcohol limit in recent years. In other words, the limit becomes stringent in more states after 1985.
5) The seat belt law distribution plot indicates the presence of both primary and secondary seat belt laws across states. Starting in 1985, four states began enforcing seat belt laws, and by the end of the decade, nearly all states had implemented such laws, with an increasing number of states adopting primary seat belt laws. The implementation of seat belt laws may give rise to the observed decrease in average total fatality rate.

```
### Unemployment ###
# Histogram
hist_unemploy <- df %>%
  ggplot() +
  geom_histogram(aes(x = unemp_rate)) +
  labs(title = "Unemployment Rate",
       x = "Unemployment Rate") +
  theme(legend.position = c(.2, .8))

# Boxplot
bp_unemploy <- df %>%
  ggplot() +
  geom_boxplot(aes(x = year_of_observation, y = unemp_rate)) +
  labs(y = "Unemployment Rate",
       x = "Year") +
  theme(legend.position = c(.2, .8),
        axis.text.x = element_text(size=6, angle=30, hjust=1, vjust=1))

# Scatter plot: linearity
lm_unemploy <- ggplot(df, aes(x = unemp_rate, y = total_fatality_rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Fatality Rate vs. Unemployment Rate",
       x = "Unemployment Rate",
       y = "Total Fatality Rate")

### Vehicle ###
# Histogram
hist_vehicle <- df %>%
  ggplot() +
  geom_histogram(aes(x = vehicle_miles_per_capita)) +
  labs(title = "Vehicle Miles Traveled per Capita",
       x = "Vehicle Miles Traveled per Capita") +
  theme(legend.position = c(.2, .8))

# Boxplot
bp_vehicle <- df %>%
  ggplot() +
  geom_boxplot(aes(x = year_of_observation, y = vehicle_miles_per_capita)) +
  labs(y = "Vehicle Miles Traveled per Capitae",
       x = "Year") +
  theme(legend.position = c(.2, .8),
        axis.text.x = element_text(size=6, angle=30, hjust=1, vjust=1))

# Scatter plot: linearity
```

```r
lm_vehicle <- ggplot(df, aes(x = vehicle_miles_per_capita, y = total_fatality_rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Fatality Rate vs. Vehicle Miles Traveled per Capita",
       x = "Vehicle Miles Traveled per Capita",
       y = "Total Fatality Rate")


### Pct of 14 to 24 ###
# Histogram
hist_pop12 <- df %>%
  ggplot() +
  geom_histogram(aes(x = percent_pop_14_24)) +
  labs(title = "Percentage of Population Aged 14 to 24",
       x = "Percentage of Population Aged 14 to 24") +
  theme(legend.position = c(.2, .8))

# Boxplot
bp_pop12 <- df %>%
  ggplot() +
  geom_boxplot(aes(x = year_of_observation, y = percent_pop_14_24)) +
  labs(y = "Percentage of Population Aged 14 to 24",
       x = "Year") +
  theme(legend.position = c(.2, .8),
        axis.text.x = element_text(size=6, angle=30, hjust=1, vjust=1))

# Scatter plot: linearity
lm_pop12 <- ggplot(df, aes(x = percent_pop_14_24, y = total_fatality_rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Fatality Rate vs. Percentage of Population Aged 14 to 24",
       x = "Percentage of Population Aged 14 to 24",
       y = "Total Fatality Rate")


#### Distribution Plot ###
dist_sl70plus <- df %>%
    mutate(sl70plus = ifelse(sl70plus == "1", "1", NA)) %>%  # Transforming sl70plus column
    ggplot(aes(x = year_of_observation, fill = sl70plus)) +  # Create plot
    geom_bar(position = "stack") +  # Create stacked bar plot
    geom_smooth(aes(x = year_of_observation, y = total_fatality_rate),
                method = "auto", se = FALSE) +  # Add smoothing line
    labs(title = "Fatality Rate vs. Speed Limit 70 Plus",
         x = "Year of Observation",
         y = "Total Fatality Rate") +
    scale_fill_manual(values = c("1" = "orange")) +
  theme(legend.position = c(.2, .8),
        axis.text.x = element_text(size=6, angle=30, hjust=1, vjust=1))

dist_grad <- df %>%
    ggplot(aes(x = year_of_observation, fill = graduate_driver_law)) +  # Create plot
    geom_bar(position = "stack") +  # Create stacked bar plot
    geom_smooth(aes(x = year_of_observation, y = total_fatality_rate),
```

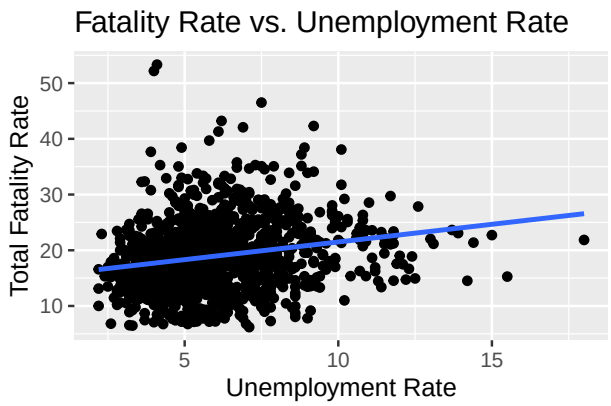
```
                method = "auto", se = FALSE) +  # Add smoothing line
    labs(title = "Fatality Rate vs. Graduate Driver Law",
         x = "Year of Observation",
         y = "Total Fatality Rate") +
    scale_fill_manual(values = c("1" = "green")) +
  theme(legend.position = c(.2, .8),
        axis.text.x = element_text(size=6, angle=30, hjust=1, vjust=1))

dist_adm <- df %>%
    ggplot(aes(x = year_of_observation, fill = admin_license_revoke)) +  # Create plot
    geom_bar(position = "stack") +  # Create stacked bar plot
    geom_smooth(aes(x = year_of_observation, y = total_fatality_rate), method = "auto", se = FALSE) +
    labs(title = "Fatality Rate vs. Administration License Revocation",
         x = "Year of Observation",
         y = "Total Fatality Rate") +
    scale_fill_manual(values = c("1" = "pink")) +
  theme(legend.position = c(.2, .8),
        axis.text.x = element_text(size=6, angle=30, hjust=1, vjust=1))

grid.arrange(hist_unemploy, bp_unemploy, nrow = 1, ncol = 2)
```

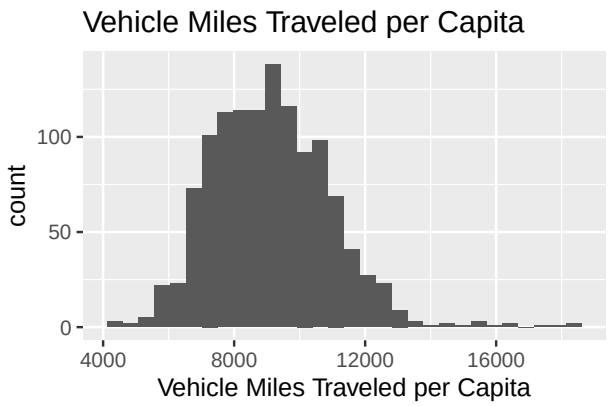
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
grid.arrange(lm_unemploy, hist_vehicle, nrow = 1, ncol = 2)
```

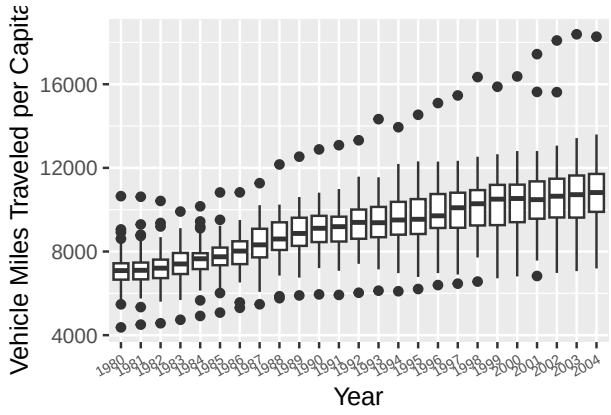
```
## `geom_smooth()` using formula = 'y ~ x'
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

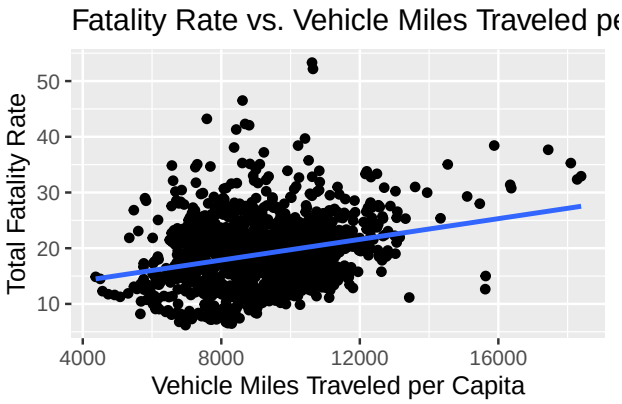





```
grid.arrange(bp_vehicle, lm_vehicle, nrow = 1, ncol = 2)
```

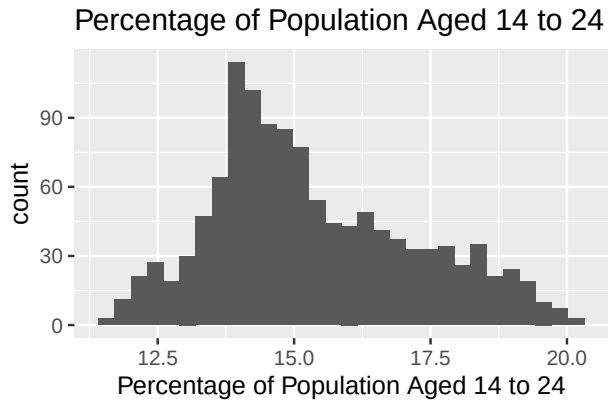
```
## `geom_smooth()` using formula = 'y ~ x'
```

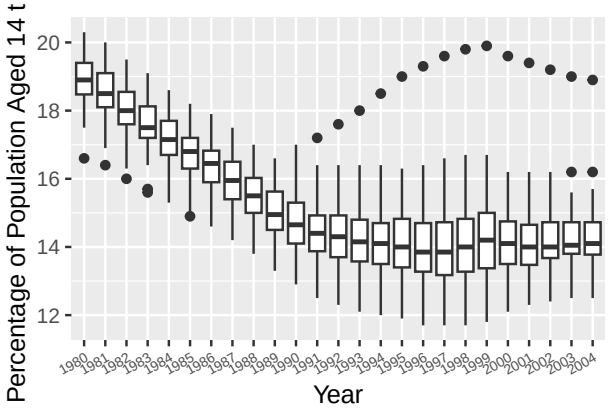





```
grid.arrange(hist_pop12, bp_pop12, nrow = 1, ncol = 2)
```

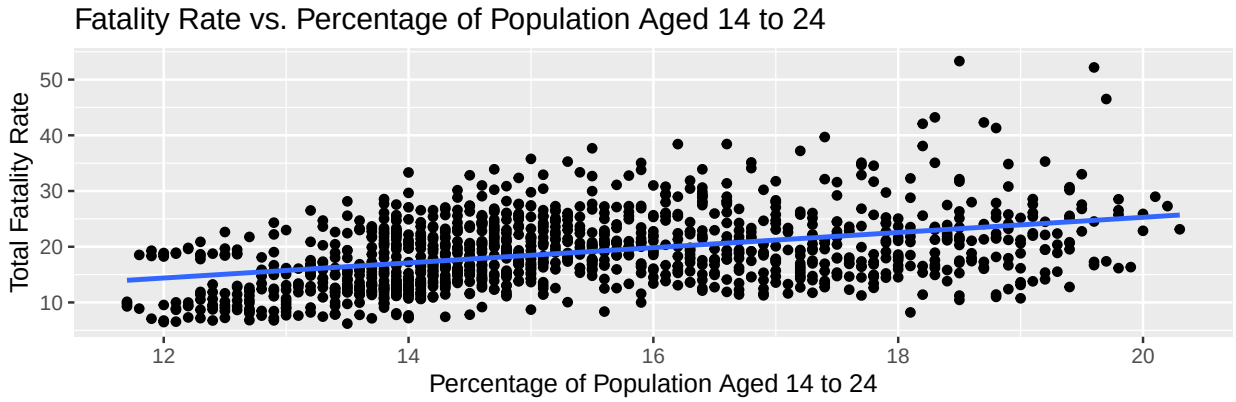
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





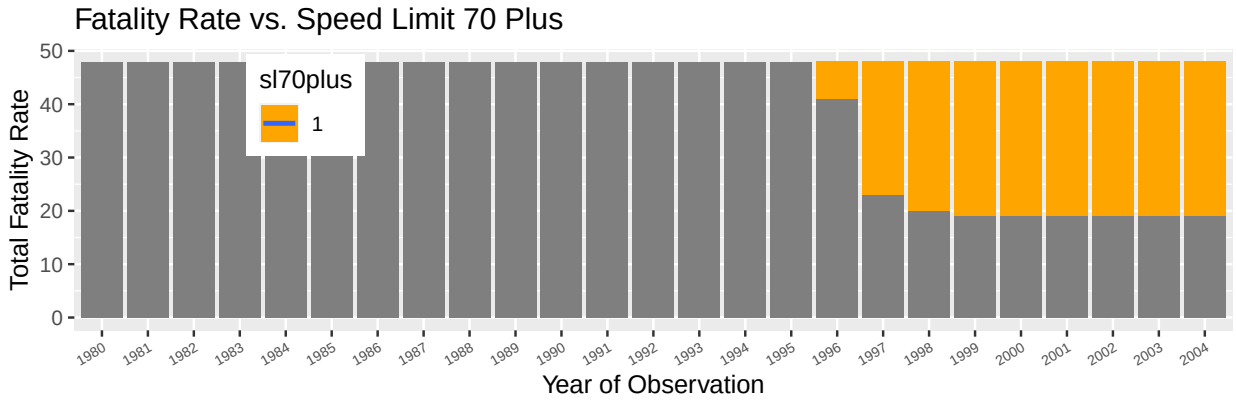
```
grid.arrange(lm_pop12, nrow = 1, ncol = 1)
```

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



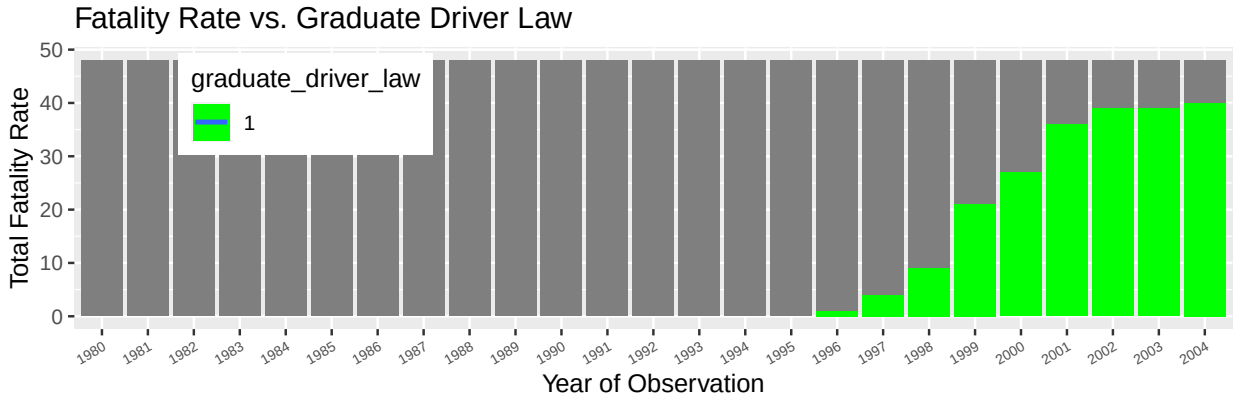

```
grid.arrange(dist_sl70plus, nrow = 1, ncol = 1)
```

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



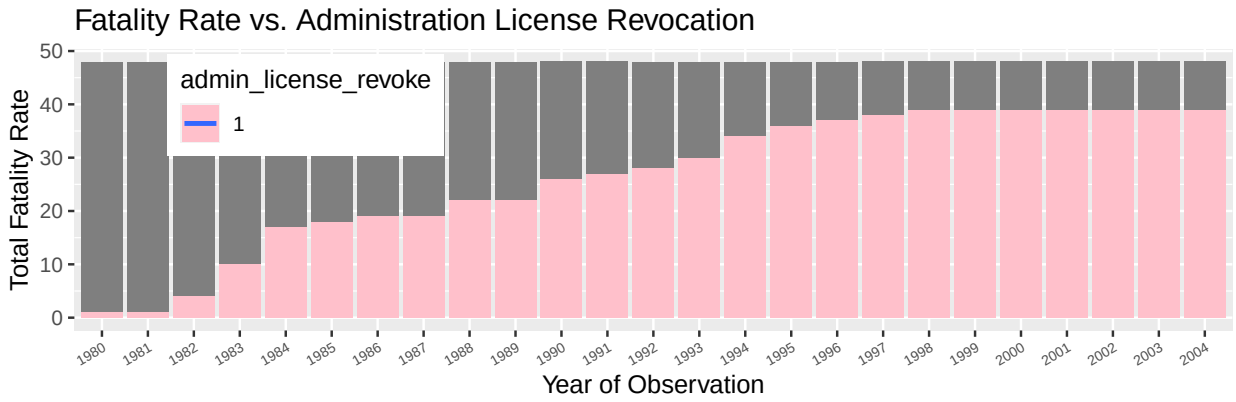

```
grid.arrange(dist_grad, nrow = 1, ncol = 1)
```

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



```
grid.arrange(dist_adm, nrow = 1, ncol = 1)
```

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



Answer:
1) We first plot the histogram of the three numeric variables included in our expanded model: `Percentage of Population Aged 14 to 24`, `Vehicle Miles Traveled per Capit`, and `Unemployment Rate`. The histograms of all three numeric variables are heavily right skewed, and we decided to apply log transformation.
2) The unemployment rate fluctuates between 2.2% and 18%, demonstrating a downward trend over the years. Additionally, the variance in the yearly unemployment rate shows a decreasing pattern over time. A linear regression analysis reveals a positive association between total fatality rate and unemployment rate.
3) The per capita vehicle miles traveled exhibits an upward trajectory, with the variance across states increasing over time. A linear regression model indicates a positive correlation between per capita vehicle miles traveled and total fatality rate.
4) The overall trend in the percentage of the population aged 14 to 24 years old has been declining until 1997. Despite this trend, there have been higher outlines in earlier years until 1997. This demographic's proportion shows a positive relationship with the total fatality rate.
5) The administrative license revocation became more common from 1990. It seems to have negative correlation with the fatality rate.
6) The graduate driver license law starts to be common from 1995 and can affect the total fatality rate.
7) Around half of the states have applied speed limit higher than 70 mph since 1887, which accounts for the decreasing trend of the total fatality rate.

(15 points) Preliminary Model

Estimate a linear regression model of *totfatrte* (total_fatality_rate) 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:

```
mod.linear <- lm(total_fatality_rate ~ year_of_observation - 1, data = df)
summary(mod.linear)
```

```
##
## Call:
## lm(formula = total_fatality_rate ~ year_of_observation - 1, data = df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## year_of_observation1980  25.4946     0.8671   29.40   <2e-16 ***
## year_of_observation1981  23.6702     0.8671   27.30   <2e-16 ***
## year_of_observation1982  20.9425     0.8671   24.15   <2e-16 ***
## year_of_observation1983  20.1529     0.8671   23.24   <2e-16 ***
## year_of_observation1984  20.2675     0.8671   23.37   <2e-16 ***
## year_of_observation1985  19.8515     0.8671   22.89   <2e-16 ***
## year_of_observation1986  20.8004     0.8671   23.99   <2e-16 ***
## year_of_observation1987  20.7748     0.8671   23.96   <2e-16 ***
## year_of_observation1988  20.8917     0.8671   24.09   <2e-16 ***
## year_of_observation1989  19.7723     0.8671   22.80   <2e-16 ***
## year_of_observation1990  19.5052     0.8671   22.49   <2e-16 ***
## year_of_observation1991  18.0948     0.8671   20.87   <2e-16 ***
## year_of_observation1992  17.1579     0.8671   19.79   <2e-16 ***
## year_of_observation1993  17.1277     0.8671   19.75   <2e-16 ***
## year_of_observation1994  17.1552     0.8671   19.78   <2e-16 ***
## year_of_observation1995  17.6685     0.8671   20.38   <2e-16 ***
## year_of_observation1996  17.3694     0.8671   20.03   <2e-16 ***
## year_of_observation1997  17.6106     0.8671   20.31   <2e-16 ***
## year_of_observation1998  17.2654     0.8671   19.91   <2e-16 ***
## year_of_observation1999  17.2504     0.8671   19.89   <2e-16 ***
## year_of_observation2000  16.8256     0.8671   19.40   <2e-16 ***
## year_of_observation2001  16.7927     0.8671   19.37   <2e-16 ***
## year_of_observation2002  17.0296     0.8671   19.64   <2e-16 ***
## year_of_observation2003  16.7635     0.8671   19.33   <2e-16 ***
## year_of_observation2004  16.7290     0.8671   19.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.008 on 1175 degrees of freedom
## Multiple R-squared:  0.9113, Adjusted R-squared:  0.9094
## F-statistic: 482.9 on 25 and 1175 DF,  p-value: < 2.2e-16
```

- Why is fitting a linear model a sensible starting place?

Answer: Linear regression model provides a straight forward way to assess the relationship between years and the total fatality rate. With it, we can see a general trend over the years and how each year compares to the base year. There is linear relationshipt between the numeric explanatory variables and total fatality rate over time.

- What does this model explain, and what do you find in this model?

Answer: The linear model explains the variance in the total fatality rate with an R^2 value of 0.9113. The high value suggests that the model is capturing a significant amount of variability in the total fatality rate over the years even though it is a simple model.

- Did driving become safer over this period? Please provide a detailed explanation.

Answer: We can observe that the coefficients from 'mod.linear' indicates a trend over time. The coefficients for year 1980 is 25.49 with the subsequent years' coefficients being lower. The consistent decrease in the estimated coef suggest that the total fatality rate has decreased from 1980 to 2004 so the driving has become safer overall in this period and is confirmed by the significant p-value for all of the years.

- 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?

Answer: The parameter estimates are statistically significant and provide insight into the trend over time. However, their reliability as unbiased estimates of the true effect of time might be compromised by the exclusion of other relevant variables and the potential for serial correlation in the errors, given the panel nature of the data.

```
- Are the uncertainty estimate reliable, unbiased estimates of sampling based variability? Or, are they
```

Answer: The uncertainty estimates in this linear regression model, which include the standard errors and confidence intervals, are likely biased due to the structure of the data, since panel data consists of multiple observations across time for the same states. This inherently introduces correlation among observations within each state across different years.

(15 points) Expanded Model

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

- Blood alcohol levels
- Per se laws
- Primary seat belt laws (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)
- Secondary seat belt laws
- Speed limits faster than 70
- Graduated drivers licenses
- Percent of the population between 14 and 24 years old
- Unemployment rate
- Vehicle miles driven per capita.

If it is appropriate, include transformations of these variables. Please carefully explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed.

```
mod.expanded <- lm(total_fatality_rate ~ year_of_observation +
                                         I(blood_alcohol_limit == "8") +
                                         I(blood_alcohol_limit == "10") +
                                         admin_license_revoke +
                                         sbprim + sbsecon +
                                         I(sl70plus == "1") + graduate_driver_law +
                                         log(percent_pop_14_24) +
                                         log(unemp_rate) +
                                         log(vehicle_miles_per_capita),
                                         data = df)

summary(mod.expanded)
```

```
##
## Call:
## lm(formula = total_fatality_rate ~ year_of_observation + I(blood_alcohol_limit ==
##     "8") + I(blood_alcohol_limit == "10") + admin_license_revoke +
##     sbprim + sbsecon + I(sl70plus == "1") + graduate_driver_law +
##     log(percent_pop_14_24) + log(unemp_rate) + log(vehicle_miles_per_capita),
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9656  -2.6813  -0.2624   2.3800  20.6741
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -244.75378    8.67419 -28.216  < 2e-16 ***
## year_of_observation1981            -2.14051    0.80745  -2.651  0.00814 **
## year_of_observation1982            -6.47265    0.82335  -7.861 8.60e-15 ***
## year_of_observation1983            -7.39827    0.83887  -8.819  < 2e-16 ***
## year_of_observation1984            -6.23734    0.85170  -7.323 4.50e-13 ***
## year_of_observation1985            -7.02968    0.86633  -8.114 1.23e-15 ***
## year_of_observation1986            -6.42247    0.90027  -7.134 1.71e-12 ***
## year_of_observation1987            -7.02188    0.93638  -7.499 1.27e-13 ***
## year_of_observation1988            -7.21387    0.98364  -7.334 4.18e-13 ***
## year_of_observation1989            -8.80063    1.02080  -8.621  < 2e-16 ***
## year_of_observation1990            -9.82403    1.04338  -9.416  < 2e-16 ***
## year_of_observation1991           -12.02732    1.06620 -11.281  < 2e-16 ***
## year_of_observation1992           -13.82179    1.08821 -12.701  < 2e-16 ***
## year_of_observation1993           -13.64771    1.10232 -12.381  < 2e-16 ***
## year_of_observation1994           -13.13620    1.12525 -11.674  < 2e-16 ***
## year_of_observation1995           -12.55917    1.15251 -10.897  < 2e-16 ***
## year_of_observation1996           -13.59736    1.17505 -11.572  < 2e-16 ***
## year_of_observation1997           -14.42449    1.21883 -11.835  < 2e-16 ***
## year_of_observation1998           -14.98977    1.23750 -12.113  < 2e-16 ***
## year_of_observation1999           -14.84599    1.26067 -11.776  < 2e-16 ***
## year_of_observation2000           -15.01380    1.28067 -11.723  < 2e-16 ***
## year_of_observation2001           -16.02059    1.29391 -12.382  < 2e-16 ***
## year_of_observation2002           -16.77800    1.29778 -12.928  < 2e-16 ***
## year_of_observation2003           -17.23163    1.30025 -13.253  < 2e-16 ***
## year_of_observation2004           -16.62211    1.33194 -12.480  < 2e-16 ***
## I(blood_alcohol_limit == "8")TRUE   -2.35875    0.51411  -4.588 4.96e-06 ***
## I(blood_alcohol_limit == "10")TRUE  -1.21722    0.38038  -3.200  0.00141 **
## admin_license_revoke1               -0.63269    0.28600  -2.212  0.02714 *
## sbprim1                             -0.07814    0.48219  -0.162  0.87129
## sbsecon1                             0.05747    0.42009   0.137  0.89121
## I(sl70plus == "1")TRUE               3.11003    0.43335   7.177 1.27e-12 ***
## graduate_driver_law1                -0.83145    0.49194  -1.690  0.09127 .
## log(percent_pop_14_24)               3.49095    1.81076   1.928  0.05411 .
## log(unemp_rate)                      5.37971    0.47363  11.358  < 2e-16 ***
## log(vehicle_miles_per_capita)       28.23622    0.87393  32.309  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.95 on 1165 degrees of freedom
## Multiple R-squared:  0.626,  Adjusted R-squared:  0.6151
```

```
## F-statistic: 57.36 on 34 and 1165 DF,  p-value: < 2.2e-16
```

Data Transformation **Answer**: From the correlation heat map, we can observe the weak correlation between 'total_fatality_rate' and the three variables 'percent_pop_14_24', 'unemp_rate', and 'vehicle_miles_per_capita' resulting in a log transformation of these variables to better see nonlinear relationship. On the other hand, variables such as blood_alcohol_limit, admin_license_revoke, sbprim, sbsecon, and sl70plus did not undergo transformations as they are binary indicators reflecting the presence or absence of specific laws or regulations.

- How are the blood alcohol variables defined? Interpret the coefficients that you estimate for this concept. **Answer**: The blood alcohol variables are defined using dummy variables to indicate the presence of specific blood alcohol concentration (BAC) limits: I(blood_alcohol_limit == "8")TRUE for whether the legal BAC limit is 0.08% and I(blood_alcohol_limit == "10")TRUE for whether the legal BAC limit is 0.10% in a given state-year. For BAC limit of 0.08%, the coefficient is -2.35875, with a p-value < 0.00001, suggesting that having a BAC limit of 0.08% is significantly associated with a decrease in the total fatality rate by approximately 2.36 fatalities per 100,000 population. For BAC limt of 0.10% the coefficient is -1.21722, with a p-value of 0.00141, indicating that having a BAC limit of 0.10% is also associated with a decrease in the total fatality rate but to a lesser extent compared to a BAC limit of 0.08%.
- Do *per se laws* have a negative effect on the fatality rate? **Answer**: The coefficient for this variable, admin_license_revoke1, is -0.63269 with a p-value of 0.02714, indicating that the presence of per se laws is associated with a statistically significant reduction in the total fatality rate, decreasing fatalities by roughly 0.63 per 100,000 population. Suggesting that per se laws have a negative effect on the fatality rate, contributing to safer driving conditions by possibly deterring driving under the influence.
- Does having a primary seat belt law? **Answer**: The coefficient for primary seat belt laws (sbprim) is -0.07814 with a p-value of 0.87129. This result is not statistically significant, suggesting that the data does not provide strong evidence that primary seat belt laws, as modeled here, have a significant impact on reducing the total fatality rate within the scope of this study.

(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?
  Answer:
  For the 0.08% limit (blood_alcohol_limit == "8"):
  panel linear within: -1.836
  OLS expanded model: -2.359
  For the 0.10% limit (blood_alcohol_limit == "10"):
  panel linear within: -1.412
  OLS expanded model: -1.217
  The coefficients on blood alcohol limit variables change slightly between the two models. The coefficients of both models are statistially significant. Both models find negative coefficients for both BAC limits, suggesting that a stricter limit on BAC is associated with a decrease in total fatality rate. The effects are slightly stronger in the panel linear within model compared to the OLS expanded model.

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

For administrative license revocation (admin_license_revoke1):
panel linear within: -1.588
OLS expanded model: -0.633
The coefficients on the administrative license revocation variable decrease from the panel linear within model to the OLS expanded model, but both coefficients remain negative and are statistically significant. This suggests that administrative license revocation is associated with a decrease in total fatality rate.

- What do you estimate for coefficients on primary seat-belt laws? How do the coefficients on primary seatbelt laws change, if at all?
  Answer:
  The primary seat belt law (sbprim1) coefficients in the two models are:
  panel linear within: -1.838
  OLS expanded model: -0.078
  The panel linear within model estimates a significant negative effect of primary seat belt laws on total fatality rate. However, in the OLS expanded model, the coefficient is very close to zero and not significant.

- Which set of estimates do you think is more reliable? Why do you think this?
  Answer:
  The Within model can account for more variation within each state over time and can control for unobserved, state-specific fixed effects. This can lead to more reliable estimates as it controls for some confounding variables. OLS Expanded Model includes more variables and may suffer from omitted variable bias or multicollinearity issues. It may be less reliable due to its broader approach that could introduce complexities in the model. Thus, the panel linear within estimates are likely more reliable because they control for fixed effects and may better account for unobserved, state-specific differences over time. Besides, we conducted `pFtest()` for pooled model and FE model, shown in the following analysis, which reinforced that individual fixed effects are significant in our context.

- What assumptions are needed in each of these models?

Answer:
Assumption of OLS:
1) Linearity
2) Independence of errors
3) Homoscedasticity (Constant Variance)
4) Normality of errors
5) No perfect multicollinearity

Assumption of FE:
The fixed effects model is commonly applied to remove omitted variable bias in the case of unobserved individual characteristics. By estimating changes within a specific group over time, all time-invariant differences between entities such as states are controlled for. The fixed effect model could be estimated using three estimation methods:
1) Least Squares Dummy Variable Estimation
2) First-difference Estimator
3) Fixed Effect or Within-groups Estimator

All fixed effect estimation methods are consistent under the following assumptions:
1) Linearity: the model is linear in parameters
2) i.i.d. : The observations are independent across individuals but not necessarily across time. This is

guaranteed by random sampling of individuals.
3) Indentifiability: the regressors, including a constant, are not perfectly collinear, and all regressors (but the constant) have non-zero variance and not too many extreme values.
4) Zero conditional means (strict exogeneity)

- Are these assumptions reasonable in the current context?

Answer:
1) Linearity: It's reasonable to assume linearity in the current context. Based on scatter plots showing explicit linear relationships between the numeric independent variables and the dependent variable. This implies that the linear model is a suitable choice for modeling the data.
2) Independence of Observations (iid.): In the context of analyzing data from the same 48 states over time, the assumption of independence of observations is not reasonable. This is because observations from the same state may be correlated due to shared characteristics, policies, or external factors. To address this issue, panel data techniques, which explicitly account for the dependencies among observations within the same state over time, are more appropriate. These techniques allow for modeling both within-state and between-state variations.
3) No Collinearity: It's reasonable to assume that the explanatory numeric variables are not collinear with each other, as multicollinearity can lead to unstable parameter estimates and inflated standard errors. Nevertheless, We address serial correlation and heterogeneity by employing robust error estimation techniques tailored for panel data analysis. Heterogeneity is a reasonable assumption, because it facilitates valid statistical inference.
4) This is reasonable because fixed effects models include individual-specific dummy variables or equivalent transformations to control for unobserved time-invariant individual-specific effects. These effects capture all the individual-specific characteristics that do not change over time. By including these fixed effects, the model effectively removes the correlation between the individual-specific effects and the independent variables, making the assumption of zero conditional mean plausible.

```
lsdv_model <- plm(total_fatality_rate ~  state +
                                         I(blood_alcohol_limit == "8") +
                                         I(blood_alcohol_limit == "10") +
                                         admin_license_revoke +
                                         sbprim + sbsecon +
                                         I(sl70plus == "1") + graduate_driver_law +
                                         log(percent_pop_14_24) +
                                         log(unemp_rate) +
                                         log(vehicle_miles_per_capita),
                  data = df,
                  index = c("state", "year_of_observation"),
                  effect = "individual",
                  model = "pooling")
```

```
first_diff_model <- plm(total_fatality_rate ~ I(blood_alcohol_limit == "8") +
                                         I(blood_alcohol_limit == "10") +
                                         admin_license_revoke +
                                         sbprim + sbsecon +
                                         I(sl70plus == "1") + graduate_driver_law +
                                         log(percent_pop_14_24) +
                                         log(unemp_rate) +
                                         log(vehicle_miles_per_capita),
                        data = df,
                        index = c("state", "year_of_observation"),
```

```
                        effect = "individual", model = "fd")
within_model <- plm(total_fatality_rate ~ I(blood_alcohol_limit == "8") +
                                          I(blood_alcohol_limit == "10") +
                                          admin_license_revoke +
                                          sbprim + sbsecon +
                                          I(sl70plus == "1") + graduate_driver_law +
                                          log(percent_pop_14_24) +
                                          log(unemp_rate) +
                                          log(vehicle_miles_per_capita),
                        data = df,
                        index = c("state", "year_of_observation"),
                    effect = "individual", model = "within")

summary(within_model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = total_fatality_rate ~ I(blood_alcohol_limit ==
##     "8") + I(blood_alcohol_limit == "10") + admin_license_revoke +
##     sbprim + sbsecon + I(sl70plus == "1") + graduate_driver_law +
##     log(percent_pop_14_24) + log(unemp_rate) + log(vehicle_miles_per_capita),
##     data = df, effect = "individual", model = "within", index = c("state",
##         "year_of_observation"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -6.658204 -1.213996 -0.059761  1.123123 14.425501
##
## Coefficients:
##                                    Estimate Std. Error t-value  Pr(>|t|)
## I(blood_alcohol_limit == "8")TRUE  -1.83555    0.37289 -4.9225 9.801e-07 ***
## I(blood_alcohol_limit == "10")TRUE -1.41200    0.26090 -5.4120 7.590e-08 ***
## admin_license_revoke1              -1.58754    0.23996 -6.6158 5.666e-11 ***
## sbprim1                            -1.83766    0.34822 -5.2772 1.569e-07 ***
## sbsecon1                           -0.88182    0.24976 -3.5307 0.0004311 ***
## I(sl70plus == "1")TRUE             -1.12405    0.23960 -4.6913 3.042e-06 ***
## graduate_driver_law1               -0.63477    0.22623 -2.8058 0.0051040 **
## log(percent_pop_14_24)             15.33946    1.13901 13.4673 < 2.2e-16 ***
## log(unemp_rate)                    -3.15848    0.32567 -9.6985 < 2.2e-16 ***
## log(vehicle_miles_per_capita)       3.74356    1.01489  3.6886 0.0002361 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 5618.7
## R-Squared:      0.53695
## Adj. R-Squared: 0.51384
## F-statistic: 132.425 on 10 and 1142 DF, p-value: < 2.22e-16
```

```
#### add pool model and conduct statistical test for FE ####
pooled_ols <- plm(total_fatality_rate ~ I(blood_alcohol_limit == "8") +
                                       I(blood_alcohol_limit == "10") +
                                       admin_license_revoke +
                                       sbprim + sbsecon +
                                       I(sl70plus == "1") + graduate_driver_law +
                                       log(percent_pop_14_24) +
                                       log(unemp_rate) +
                                       log(vehicle_miles_per_capita),
                     data = df,
                     index = c("state", "year_of_observation"),
                  effect = "individual", model = "pooling")
pFtest(within_model, pooled_ols)
```

```
##
##  F test for individual effects
##
## data:  total_fatality_rate ~ I(blood_alcohol_limit == "8") + I(blood_alcohol_limit ==  ...
## F = 75.103, df1 = 47, df2 = 1142, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
# stargazer(lsdv_model, first_diff_model, within_model, mod.expanded, type = "text",
#           omit.stat = c("ser","f","adj.rsq"), dep.var.labels = "",
#           column.labels = c("LSDV", "FD", "Within", "Expanded Mod"))
stargazer(within_model, mod.expanded, type = "text",
          omit.stat = c("ser","f","adj.rsq"), dep.var.labels = "",
          column.labels = c("Within", "Expanded Mod"))
```

```
##
## ===============================================================
##                                    Dependent variable:
##                                ---------------------------
##
##                                    panel           OLS
##                                    linear
##                                    Within     Expanded Mod
##                                     (1)            (2)
## ---------------------------------------------------------------
## year_of_observation1981                           -2.141***
##                                                    (0.807)
##
## year_of_observation1982                           -6.473***
##                                                    (0.823)
##
## year_of_observation1983                           -7.398***
##                                                    (0.839)
##
## year_of_observation1984                           -6.237***
##                                                    (0.852)
##
## year_of_observation1985                           -7.030***
##                                                    (0.866)
```

```
##
## year_of_observation1986                    -6.422***
##                                             (0.900)
##
## year_of_observation1987                    -7.022***
##                                             (0.936)
##
## year_of_observation1988                    -7.214***
##                                             (0.984)
##
## year_of_observation1989                    -8.801***
##                                             (1.021)
##
## year_of_observation1990                    -9.824***
##                                             (1.043)
##
## year_of_observation1991                   -12.027***
##                                             (1.066)
##
## year_of_observation1992                   -13.822***
##                                             (1.088)
##
## year_of_observation1993                   -13.648***
##                                             (1.102)
##
## year_of_observation1994                   -13.136***
##                                             (1.125)
##
## year_of_observation1995                   -12.559***
##                                             (1.153)
##
## year_of_observation1996                   -13.597***
##                                             (1.175)
##
## year_of_observation1997                   -14.424***
##                                             (1.219)
##
## year_of_observation1998                   -14.990***
##                                             (1.238)
##
## year_of_observation1999                   -14.846***
##                                             (1.261)
##
## year_of_observation2000                   -15.014***
##                                             (1.281)
##
## year_of_observation2001                   -16.021***
##                                             (1.294)
##
## year_of_observation2002                   -16.778***
##                                             (1.298)
##
## year_of_observation2003                   -17.232***
##                                             (1.300)
```

```
##
## year_of_observation2004                          -16.622***
##                                                    (1.332)
##
## I(blood_alcohol_limit == "8")      -1.836***       -2.359***
##                                     (0.373)         (0.514)
##
## I(blood_alcohol_limit == "10")     -1.412***       -1.217***
##                                     (0.261)         (0.380)
##
## admin_license_revoke1              -1.588***       -0.633**
##                                     (0.240)         (0.286)
##
## sbprim1                            -1.838***        -0.078
##                                     (0.348)         (0.482)
##
## sbsecon1                           -0.882***         0.057
##                                     (0.250)         (0.420)
##
## I(sl70plus == "1")                 -1.124***       3.110***
##                                     (0.240)         (0.433)
##
## graduate_driver_law1               -0.635***        -0.831*
##                                     (0.226)         (0.492)
##
## log(percent_pop_14_24)             15.339***        3.491*
##                                     (1.139)         (1.811)
##
## log(unemp_rate)                    -3.158***       5.380***
##                                     (0.326)         (0.474)
##
## log(vehicle_miles_per_capita)       3.744***       28.236***
##                                     (1.015)         (0.874)
##
## Constant                                           -244.754***
##                                                    (8.674)
##
## ------------------------------------------------------------
## Observations                         1,200           1,200
## R2                                   0.537           0.626
## ============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

The pooled estimator ignores any of the panel data structure and just runs basic OLS. pFtest() will run this with the null hypothesis that the pooled OLS model is better than the FE model, i.e., individual intercepts are zero. The null hypothesis is rejected in favor of individual fixed effects being significant.

(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.

Answer:
1) Linearity: the model is linear in parameters
2) i.i.d. : The observations are independent across individuals but not necessarily across time. This is guaranteed by random sampling of individuals.
3) Indentifiability: the regressors, including a constant, are not perfectly collinear, and all regressors (but the constant) have non-zero variance and not too many extreme values.
4) The unobserved individual-specific effect (time-invariant) is independent of all explanatory variables in all time periods.

- 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.

Answer:

```
re.model <- plm(total_fatality_rate ~ I(blood_alcohol_limit == "8") +
                                      I(blood_alcohol_limit == "10") +
                                      admin_license_revoke +
                                      sbprim + sbsecon +
                                      I(sl70plus == "1") + graduate_driver_law +
                                      log(percent_pop_14_24) +
                                      log(unemp_rate) +
                                      log(vehicle_miles_per_capita),
                    data = df,
                    index = c("state", "year_of_observation"),
                  effect = "individual", model = "random")
summary(re.model)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = total_fatality_rate ~ I(blood_alcohol_limit ==
##     "8") + I(blood_alcohol_limit == "10") + admin_license_revoke +
##     sbprim + sbsecon + I(sl70plus == "1") + graduate_driver_law +
##     log(percent_pop_14_24) + log(unemp_rate) + log(vehicle_miles_per_capita),
##     data = df, effect = "individual", model = "random", index = c("state",
##         "year_of_observation"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Effects:
##                  var std.dev share
## idiosyncratic 4.920   2.218 0.374
## individual    8.228   2.869 0.626
## theta: 0.8472
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -5.86055 -1.40991 -0.23674  0.99205 16.52957
##
## Coefficients:
##                                      Estimate Std. Error z-value  Pr(>|z|)
```

```
## (Intercept)                          -75.50491   10.82595 -6.9744 3.071e-12 ***
## I(blood_alcohol_limit == "8")TRUE     -2.09935    0.38482 -5.4554 4.885e-08 ***
## I(blood_alcohol_limit == "10")TRUE    -1.53916    0.26995 -5.7017 1.186e-08 ***
## admin_license_revoke1                 -1.49140    0.24660 -6.0479 1.467e-09 ***
## sbprim1                               -1.94148    0.35784 -5.4255 5.778e-08 ***
## sbsecon1                              -1.01617    0.25833 -3.9337 8.366e-05 ***
## I(sl70plus == "1")TRUE                -1.11302    0.24873 -4.4748 7.648e-06 ***
## graduate_driver_law1                  -0.79329    0.23527 -3.3718 0.0007469 ***
## log(percent_pop_14_24)                16.63114    1.16179 14.3151 < 2.2e-16 ***
## log(unemp_rate)                       -2.60775    0.33491 -7.7863 6.900e-15 ***
## log(vehicle_miles_per_capita)          6.27100    1.00451  6.2429 4.296e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12986
## Residual Sum of Squares: 6385.8
## R-Squared:      0.50826
## Adj. R-Squared: 0.50412
## Chisq: 1228.94 on 10 DF, p-value: < 2.22e-16
```

```
phtest(within_model, re.model)
```

```
##
##  Hausman Test
##
## data:  total_fatality_rate ~ I(blood_alcohol_limit == "8") + I(blood_alcohol_limit ==  ...
## chisq = 207.81, df = 10, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

The p-value of Hausman Test is statistically significant, it suggests that there is a correlation between the unobserved individual-specific effect and the explanatory variables, indicating that the fixed effects model is more appropriate. In other words, the assumptions of Random Effect model is not satisfied.

- 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?

Answer:
1, Biased Coefficient Estimates: If the unobserved individual-specific effect is correlated with the explanatory variables, the estimates of the Random Effect model coefficients could be biased and inconsistent. This is because the model does not consider about the endogeneity between the random effect and the explanatory variables.
2. Omitted Variable Bias: By assuming that the unobserved individual-specific effect (time-invariant) is independent of all explanatory variables in all time periods, we can omit important variables that explain variations in the data, which results in omitted variable bias.
3. Misleading Inference: If the random effects model is used inappropriately, the conclusions drawn from the analysis may be invalid. For example, the result may indicate that certain variables have a significant impact when they do not, or vice versa.

(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.

Answer:

```
# Read the Excel file
file_path <- "./data/10316_vmt_fluctuation_1-12-24.xlsx"
#new_data <- read_excel(file_path, range = "B3:N17", col_names = FALSE)

# Read column names from row 3
col_names <- read_excel(file_path, range = "C3:N3", col_names = FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
```

```
col_names <- as.character(col_names)

# Assign "Year" as the name for the first column
col_names <- c("Year", col_names)

# Read data from the specified range excluding the first row
new_data <- read_excel(file_path, range = "B4:N17", col_names = FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
```

```
# Set column names
colnames(new_data) <- col_names

# Read the "Year" column from the Excel file
year_column <- read_excel(file_path, range = "B4:B17", col_names = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
# Assign the "Year" column to the data frame
new_data$Year <- year_column

# Display the resulting data frame
#print(new_data)
```

```
years_of_interest <- c(2018, 2020)
filtered_data <- new_data[new_data$Year$...1 %in% years_of_interest, ]

# Extract monthly data for 2018 and 2020
miles_2018 <- filtered_data[filtered_data$Year$...1 == 2018, -1]
miles_2020 <- filtered_data[filtered_data$Year$...1 == 2020, -1]

# Calculate the percentage difference between 2018 and 2023
percentage_difference <- ((miles_2020 - miles_2018) / miles_2018) * 100

# Find the month associated with the most positive difference
most_positive_month <- colnames(percentage_difference)[which.max(percentage_difference)]

# Find the month associated with the most negative difference
most_negative_month <- colnames(percentage_difference)[which.min(percentage_difference)]

# Print the results
cat("Month with the most positive difference:", most_positive_month,percentage_difference$Feb, "%\n")
```

```
## Month with the most positive difference: Feb 3.039648 %
```

```
cat("Month with the most negative difference:", most_negative_month, percentage_difference$Apr,  "%\n")
```

```
## Month with the most negative difference: Apr -39.23077 %
```

```
percentage_decrease <- 39.23077/100
percentage_increase <- 3.039648/100
coefficient <- 3.744
std_error <- 1.015

# Calculate the change in the logarithm of vehicle miles per capita
delta_log_increase <- log(1 + percentage_increase)
delta_log_decrease <- log(1 - percentage_decrease)

# Estimate the impact on the fatality rate for increase and decrease scenarios
impact_increase <- coefficient * delta_log_increase
impact_decrease <- coefficient * delta_log_decrease

# Calculate the margin of error for the confidence interval
margin_of_error <- 1.96 * std_error

# Calculate the confidence interval for increase and decrease scenarios
lower_bound_increase <- impact_increase - margin_of_error
upper_bound_increase <- impact_increase + margin_of_error
lower_bound_decrease <- impact_decrease - margin_of_error
upper_bound_decrease <- impact_decrease + margin_of_error

# Create a data frame for the results
results <- data.frame(
  "Scenario" = c("Increase", "Decrease"),
  "Estimate" = c(impact_increase, impact_decrease),
  "Lower Bound (95% CI)" = c(lower_bound_increase, lower_bound_decrease),
  "Upper Bound (95% CI)" = c(upper_bound_increase, upper_bound_decrease)
)

# Print the results
print(results)
```

```
##   Scenario   Estimate Lower.Bound..95..CI. Upper.Bound..95..CI.
## 1 Increase  0.1121091           -1.877291           2.1015091
## 2 Decrease -1.8648363           -3.854236           0.1245637
```

The greatest increase in the miles driven when comparing between 2018 and 2020 (the year where we felt the effects of the pandemic) was in February, where there was about a 3% increase. The greatest drop in miles driven was in April with about 39% drop. This makes sense as this was around the time when the lock down was imposed. Holding all else constant in the within model, a 3% increase would result in the total fatality rate 0.11. However, when looking at the confidence interval we can see at 95% confidence 0 becomes part of our interval, meaning this estimate is not significant. Similarly, we see that approximately 39% drop in the miles driven per captia would result in the total fatality rate decreasing by 1.16. But we face the same challenge here, where at 95% confidence, 0 becomes part of our confidence interval. As such this result is also not significant.

(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?

If there were serial correlation or heteroskedasticity in the idiosyncratic errors of the model then we would have biased estimates. Which means that the estimates would not be a true representation of relationships of the variables in the model. Furthermore, we would not have efficient and consistent estimates, because the standard errors could under or over estimated, and as the sample size changes, the estimates would not hold. This would consequently result in undermining the validity of our model.

Answer:
To check for heteroskedasticity, we perform the Breusch Pagan Test, which has a null hypothesis of homoskedasticity.

```
pcdtest(within_model, test = 'lm')
```

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  total_fatality_rate ~ I(blood_alcohol_limit == "8") + I(blood_alcohol_limit ==     "10") + adu
## chisq = 3400.9, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

The p-value of the Breusch Pagan test is significant, less than 0.05, meaning we reject the null hypothesis of homoskedasticity.

Next we check for serial correlation using the Breusch-Godfrey test, which has a null hypothesis of no serial correlation.

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

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatality_rate ~ I(blood_alcohol_limit == "8") + I(blood_alcohol_limit ==  ...
## chisq = 339.98, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

The p-value of the Bresuch-Godfrey test with two lags is less than 0.05. Thus we reject the null hypothesis of no serial correlation. The analysis shows that the residuals in our model are both heteroskedastic and have serial correlation. This suggests that we employ techniques to correct for serial correlation and heteroskedasticity, such as using robust standard errors.