W271 Group Lab 3

US Traffic Fatalities: 1980 - 2004

Adam Kreitzman, Hailee Schuele, Lee Perkins, Paul Cooper

```
library(tidyverse)
library(plm)
library(readr)
library(lubridate)
library(stargazer)
library(knitr)
library(corrplot)
library(patchwork)
```

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="/home/rstudio/workspace/mnt/271/labs/MIDS271_Lab3/Lab_3/data/driving.RData")
```

(30 points, total) Build and Describe the Data

For the analysis, we needed to perform data cleaning and a few variable transformations. The first few rows of the final dataframe can be seen below.

During the transformation process, we made a few assumptions. We kept the speed limit that was used for the majority of a given year. For example, if the speed limit was 55 for 45% of the year and 65 for 55% of the year, we used 65. In cases of ties, the higher speed limit was used. We did this instead of a weighted metric because we're ultimately looking to assess policy changes, and there's not much practicality in estimating a coefficient for the speed limit of, say, 62 mph. Similar logic was applied to BAC, zero tolerance, minimum age, and per se variables. We also decided to use the year variable as the d# variables were not giving us additional information.

```
# Create states list
states_list <- c("Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado", "Connect

# Create a new column 'speed_limit' and initialize it with 0
data$speed_limit <- 0

# Assign the corresponding speed limit value to 'speed_limit' based on the true condition
```

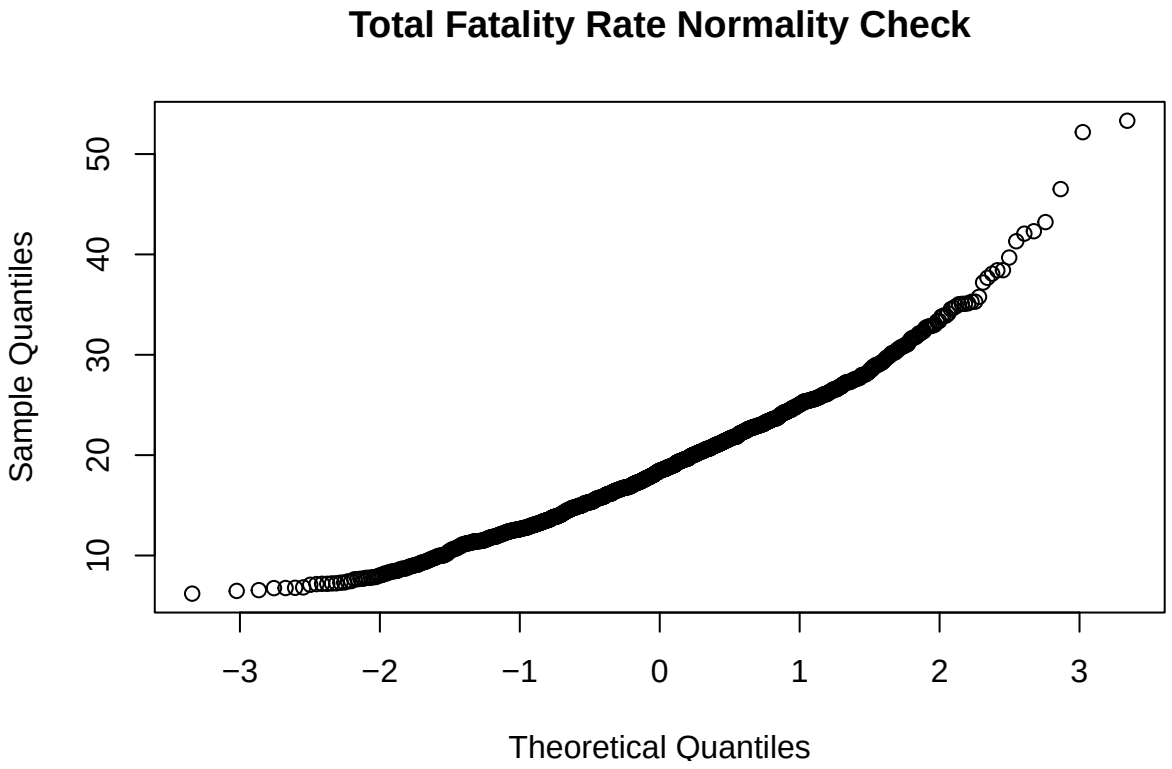
```r
data$speed_limit[data$sl55 >= 0.5] <- 55
data$speed_limit[data$sl65 >= 0.5] <- 65
data$speed_limit[data$sl70 >= 0.5] <- 70
data$speed_limit[data$sl75 >= 0.5] <- 75
data$speed_limit[data$slnone >= 0.5] <- NA
data$speed_limit <- factor(data$speed_limit)

# Drop the unnecessary speed limit columns
data <- subset(data, select = -c(sl55, sl65, sl70, sl75, slnone))

# Create a year_of_observation variable
data$year_of_observation <- factor(data$year)

# Drop the unnecessary year columns
data <- subset(data, select = -grep("^d\\d{2}$", names(data)))

# Factor state
data$state <- factor(data$state)

# Reencode one-hot variables
data$bac <- 0
data$bac[data$bac08>=0.5] <- .08
data$bac[data$bac10>=0.5] <- .1
data$bac <- factor(data$bac)

data$zeroTolerance <- 0
data$zeroTolerance[data$zerotol>=0.5] <- 1
data$zeroTolerance[data$zerotol<0.5] <- 0

data$minAge <- 0
data$minAge[data$minage>=19.5] <- 21
data$minAge[data$minage<19.5] <- 18

data$perSe <- 0
data$perSe[data$perse>=0.5] = 1
data$perSe[data$perse<0.5] = 0

data$state_str <- states_list[as.numeric(data$state)]

# Drop the unnecessary columns
data <- subset(data, select = -c(bac08, bac10, zerotol, minage, perse))

# Rename variables
data <- data %>%
  rename(total_fatality_rate = totfatrte,
         nighttime_fatality_rate  = nghtfatrte,
         weekend_fatality_rate = wkndfatrte,
```

```
          total_fatalities = totfat,
          nighttime_fatalities = nghtfat,
          weekend_fatalities = wkndfat,
          total_fatalities_per_1mmiles = totfatpvm,
          nighttime_fatalities_per_1mmiles = nghtfatpvm,
          weekend_fatalities_per_1mmiles = wkndfatpvm)

# Log non-normal variables
data$log_fatility_rate = log(data$total_fatality_rate)
data$log_unem = log(data$unem)
data$log_vehicmilespc = log(data$vehicmilespc)
```

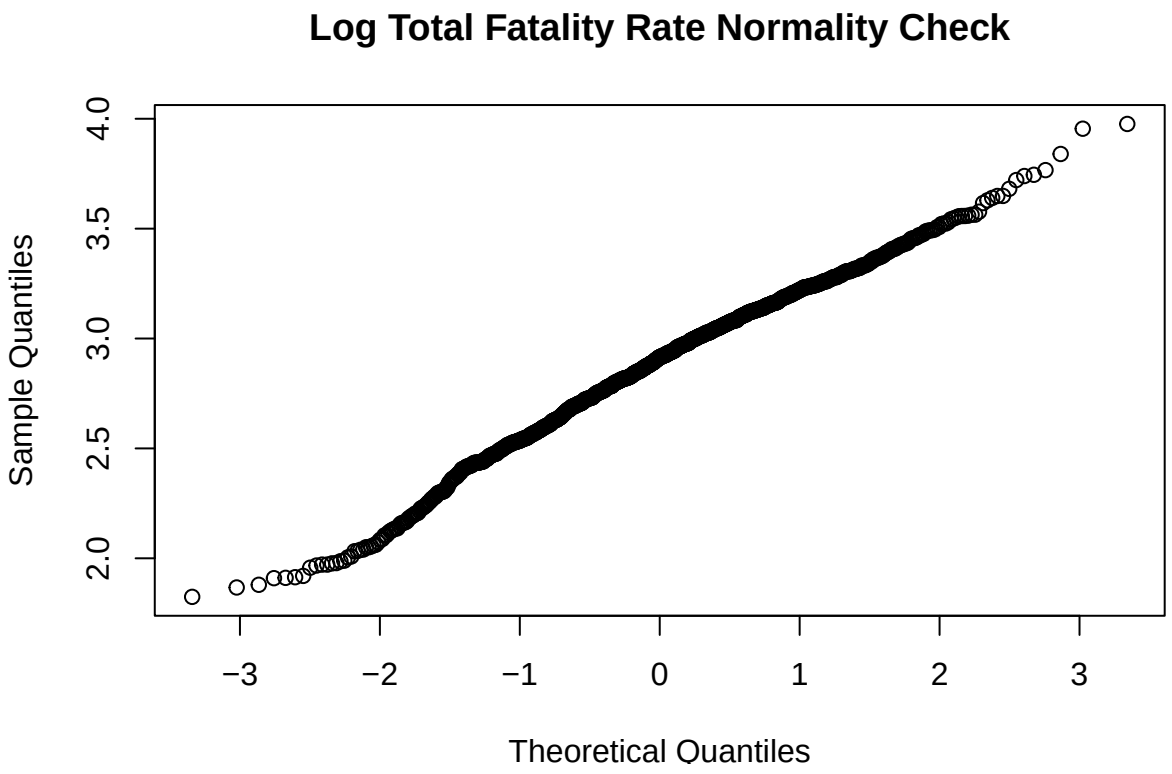
The original data has 1,200 observations and 56 columns. Each of these observations represent state-year information.

Our target variable, `total_fatality_rate`, is defined as the number of fatalities per 100,000 residents. On inspection, the fatality rate shows some left skew (see qq plot below), so to remedy this, we use the log of the rate , `log_total_fatalities_rate`, for each stage of modeling. The same was done for the `unemployment` and `vehicle miles per capita variables` (see density plots below). The dataset is a time-series that ranges from 1980-2004, tracking traffic fatality data, as well as features such as laws around BAC and speed limits. The dataset contains information for 48 states. Each of the columns represents a different variable, whereas each row represents state data from a specific year (with the year being the index.)

```
qqnorm(data$total_fatality_rate, main = "Total Fatality Rate Normality Check")
```

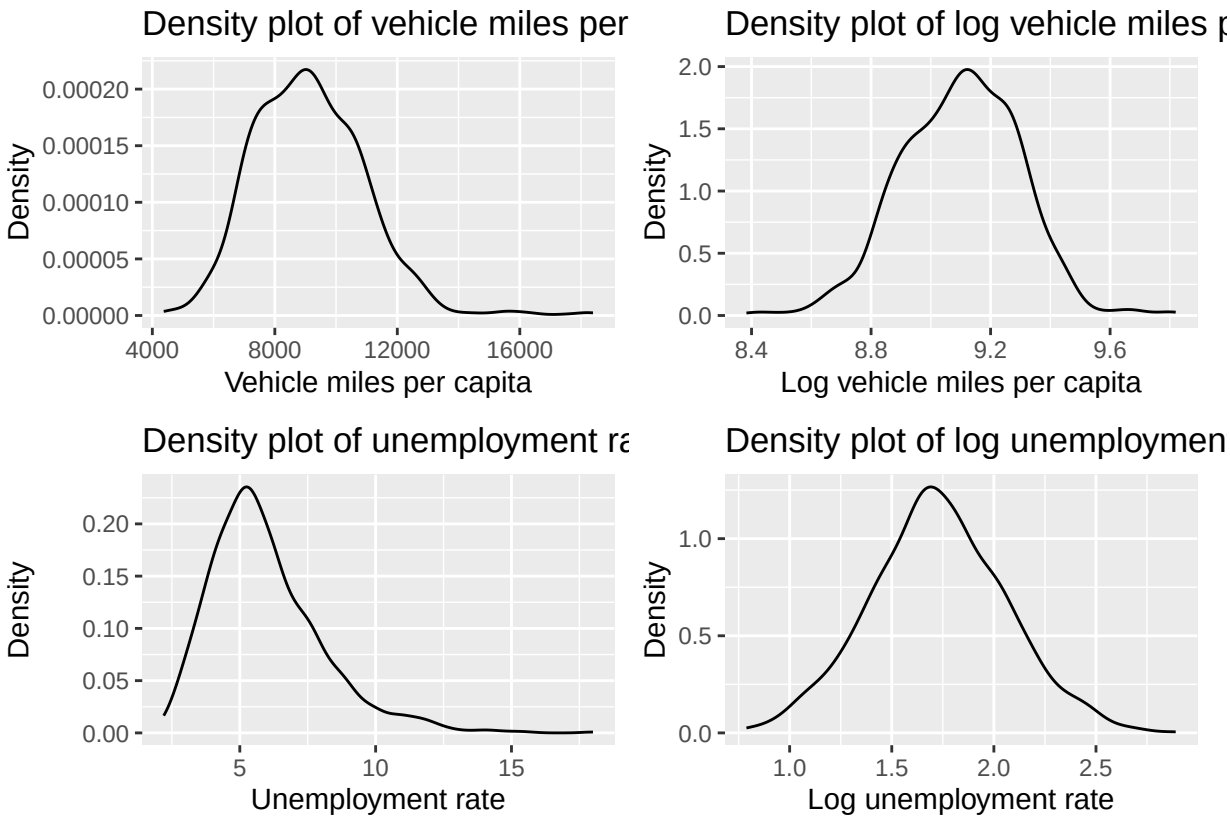


```r
qqnorm(data$log_fatility_rate,  main = "Log Total Fatality Rate Normality Check")
```




```r
vehicle_density_plot <- data %>%
    ggplot(aes(x = vehicmilespc)) +
    geom_density() +
    labs(
        title = "Density plot of vehicle miles per capita",
        x = "Vehicle miles per capita",
        y = "Density"
    )

log_vehicle_density_plot <- data %>%
    ggplot(aes(x = log(vehicmilespc))) +
    geom_density() +
    labs(
        title = "Density plot of log vehicle miles per capita",
        x = "Log vehicle miles per capita",
        y = "Density"
    )

unem_density_plot <- data %>%
    ggplot(aes(x = unem)) +
    geom_density() +
    labs(
        title = "Density plot of unemployment rate",
        x = "Unemployment rate",
```

```
        y = "Density"
    )

log_unem_density_plot <- data %>%
    ggplot(aes(x = log(unem))) +
    geom_density() +
    labs(
        title = "Density plot of log unemployment rate",
        x = "Log unemployment rate",
        y = "Density"
    )

(vehicle_density_plot + log_vehicle_density_plot) /
(unem_density_plot + log_unem_density_plot)
```

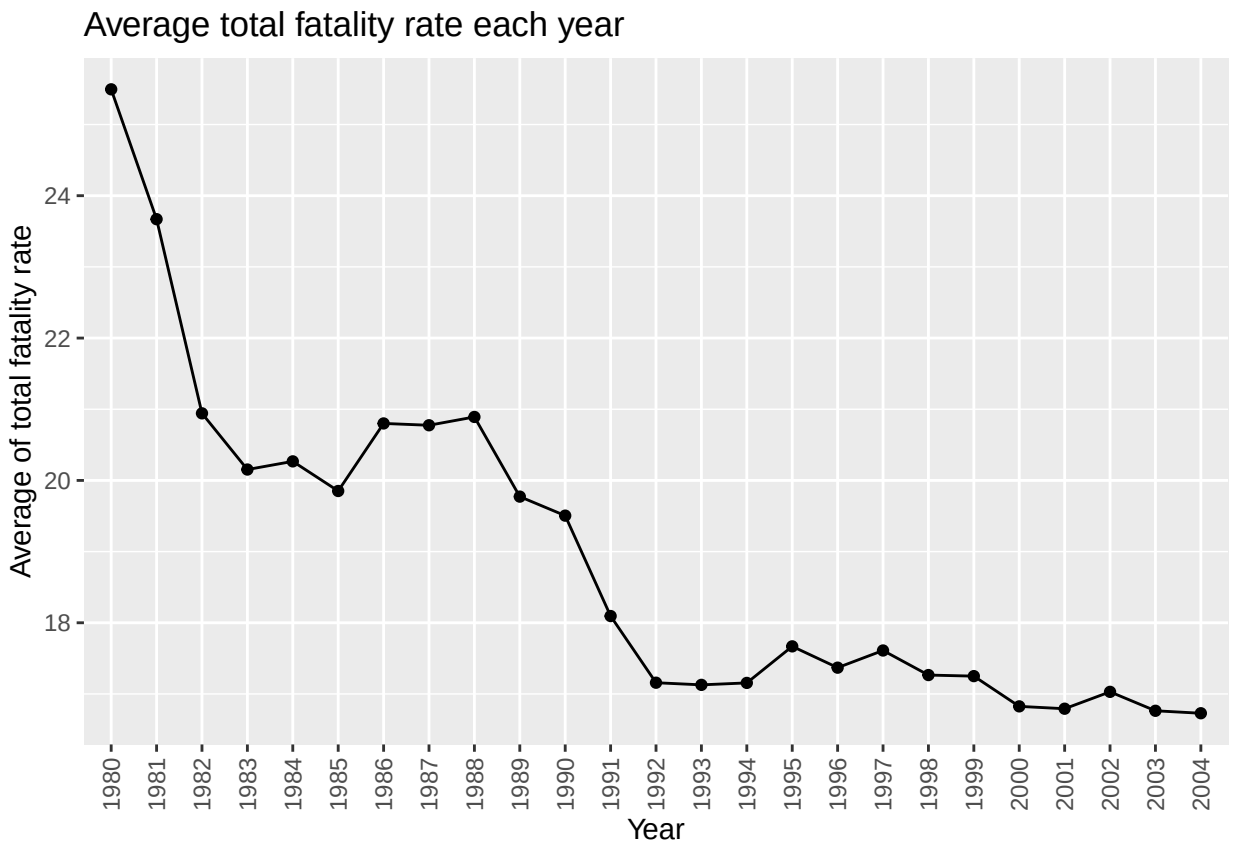


The data is collected through a survey. This would be similar to census data, as traffic fatalities are carefully recorded. Features such as laws and population are also carefully recorded.

As can be seen in the plot below, the average total fatalities decreases over the years with a bump up in 1986 and dropping back down after 1988.

```
avg_total_fatalities_rate <- data %>%
    group_by(year_of_observation) %>%
    summarise(avg_total_fatalities_rate = mean(total_fatality_rate))
```

```r
# Plot the average of 'total_fatalities_rate' in each year
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 each year",
        x = "Year",
        y = "Average of total fatality rate"
    ) +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```




```r
avg_total_fatalities_rate %>%
    mutate(avg_total_fatalities_rate = round(avg_total_fatalities_rate, 2)) %>%
    kable(
        caption = "Average total fatality rate each year",
        col.names = c("Year", "Average of total fatality rate")
    )
```

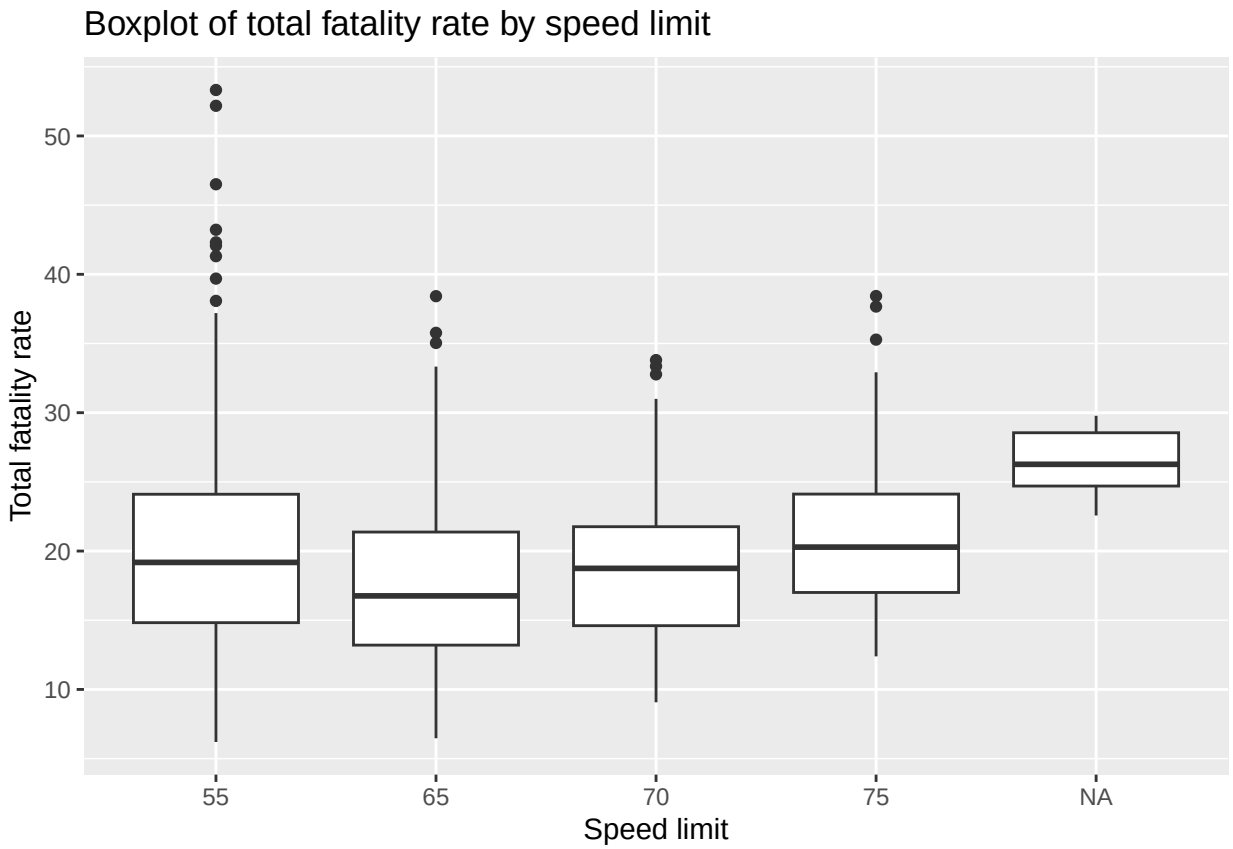
Table 1: Average total fatality rate each year

Year	Average of total fatality rate
1980	25.49

Year	Average of total fatality rate
1981	23.67
1982	20.94
1983	20.15
1984	20.27
1985	19.85
1986	20.80
1987	20.77
1988	20.89
1989	19.77
1990	19.51
1991	18.09
1992	17.16
1993	17.13
1994	17.16
1995	17.67
1996	17.37
1997	17.61
1998	17.27
1999	17.25
2000	16.83
2001	16.79
2002	17.03
2003	16.76
2004	16.73

Looking at a boxplot of the total fatality rate by speed limit we can see that there is a larger distribution of fatalities at 55mph. It is more common to see areas with a speed limit with 55mph so there will be a larger spread. What is interesting is the tighter and slightly higher distribution of the total fatality rate at 75mph. This shows that at higher speeds the total fatality rate could also be higher. The NA speed limits in this plot could range from any other speed limit that is not 55 to 75, areas without a posted speed limit, or incomplete data.

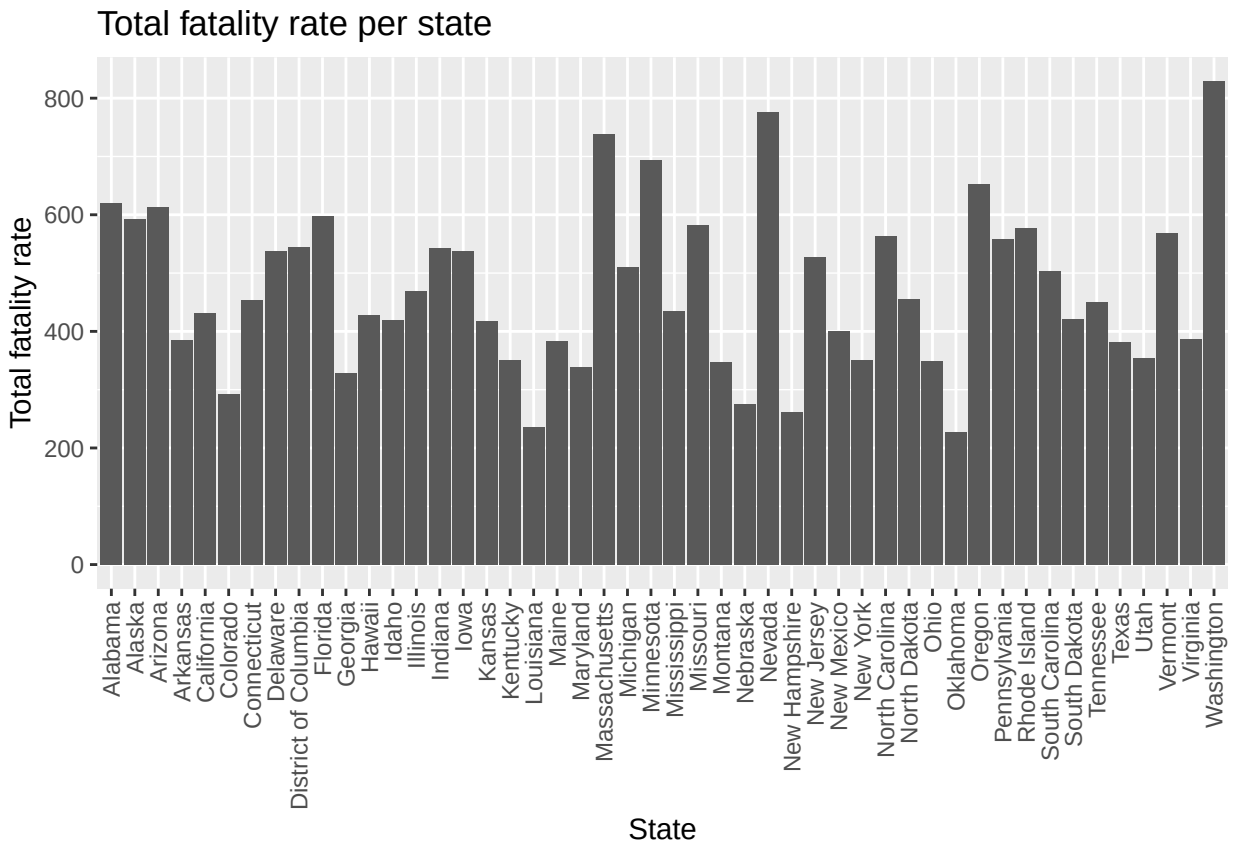
```
data %>%
    ggplot(aes(x = speed_limit, y = total_fatality_rate)) +
    geom_boxplot() +
    labs(
        title = "Boxplot of total fatality rate by speed limit",
        x = "Speed limit",
        y = "Total fatality rate"
    )
```



The bar plot below shows us the total fatality rate by state. What is interesting is how some states with dense traffic have a low total fatality rate compared to states with less traffic. Take California and Wyoming for example, California has busy and congested cities with drivers compared to Wyoming that is spread out and less congested.

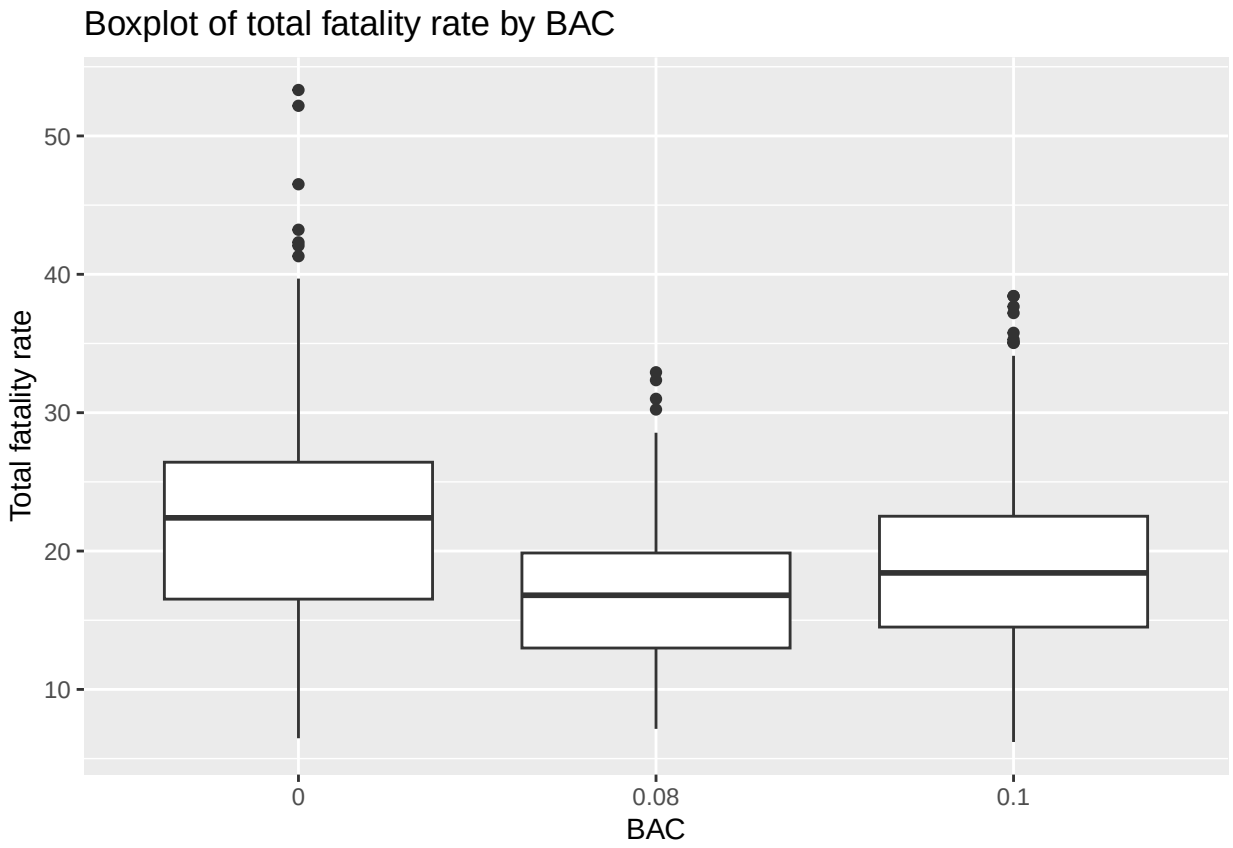
```
data %>%
    group_by(state, state_str) %>%
    summarise(total_fatality_rate = sum(total_fatality_rate)) %>%
    ggplot(aes(x = state_str, y = total_fatality_rate)) +
    geom_col() +
    labs(
        title = "Total fatality rate per state",
        x = "State",
        y = "Total fatality rate"
    ) +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

```
## `summarise()` has grouped output by 'state'. You can override using the
## `.groups` argument.
```



The figure below is a boxplot of the total fatality rate by BAC. it is clear that there is a slightly higher distribution of the total fatality rate in people that had a BAC of 0.1 compared to 0.8. Although, 0 (i.e. NA) is higher than both other BACs. NA could contain people that had a BAC higher than 0.1 or were not driving under the influence to incite the need for a BAC test.

```
data %>%
    ggplot(aes(x = bac, y = total_fatality_rate)) +
    geom_boxplot() +
    labs(
        title = "Boxplot of total fatality rate by BAC",
        x = "BAC",
        y = "Total fatality rate"
    )
```



(15 points) Preliminary Model

In this section, we estimate a linear regression model using total fatalities as our outcome variable and the years 1981 - 2004 as our explanatory variables. Fitting a linear model gives us a baseline of the shape of the data before we move to more advanced modeling techniques.

```
# Fit linear model
lm_model <- lm(log_fatility_rate ~ year_of_observation, data = data)

lm_model %>% summary()
```

```
##
## Call:
## lm(formula = log_fatility_rate ~ year_of_observation, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96324 -0.22134  0.01005  0.23221  0.86830
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.19577    0.04697  68.035  < 2e-16 ***
## year_of_observation1981 -0.07878    0.06643  -1.186 0.235904
## year_of_observation1982 -0.19957    0.06643  -3.004 0.002719 **
```

```
## year_of_observation1983 -0.23523    0.06643  -3.541 0.000414 ***
## year_of_observation1984 -0.22585    0.06643  -3.400 0.000697 ***
## year_of_observation1985 -0.24301    0.06643  -3.658 0.000265 ***
## year_of_observation1986 -0.19681    0.06643  -2.963 0.003111 **
## year_of_observation1987 -0.19871    0.06643  -2.991 0.002836 **
## year_of_observation1988 -0.18885    0.06643  -2.843 0.004547 **
## year_of_observation1989 -0.24815    0.06643  -3.735 0.000196 ***
## year_of_observation1990 -0.26785    0.06643  -4.032 5.89e-05 ***
## year_of_observation1991 -0.34372    0.06643  -5.174 2.69e-07 ***
## year_of_observation1992 -0.40229    0.06643  -6.056 1.88e-09 ***
## year_of_observation1993 -0.40257    0.06643  -6.060 1.83e-09 ***
## year_of_observation1994 -0.40798    0.06643  -6.142 1.12e-09 ***
## year_of_observation1995 -0.38492    0.06643  -5.794 8.79e-09 ***
## year_of_observation1996 -0.39949    0.06643  -6.014 2.42e-09 ***
## year_of_observation1997 -0.38596    0.06643  -5.810 8.03e-09 ***
## year_of_observation1998 -0.40954    0.06643  -6.165 9.67e-10 ***
## year_of_observation1999 -0.41450    0.06643  -6.240 6.11e-10 ***
## year_of_observation2000 -0.43694    0.06643  -6.578 7.18e-11 ***
## year_of_observation2001 -0.43521    0.06643  -6.552 8.50e-11 ***
## year_of_observation2002 -0.42672    0.06643  -6.424 1.93e-10 ***
## year_of_observation2003 -0.43978    0.06643  -6.620 5.44e-11 ***
## year_of_observation2004 -0.44853    0.06643  -6.752 2.29e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3254 on 1175 degrees of freedom
## Multiple R-squared:  0.126,  Adjusted R-squared:  0.1081
## F-statistic: 7.057 on 24 and 1175 DF,  p-value: < 2.2e-16
```

As can be seen in the model, the coefficients of many years show statistically significant effects on the logged fatality rate, with all years having a p-value at least under 0.01 other than 1981. There is a general decrease in the year coefficients as time progresses, indicating that logged total fatalities goes down over time. However, this decline is not completely consistent, indicating some years have atemporal unobserved effect that this initial model is attempting to account for.

While a good place to start, there are a number of limitations to this initial model. For instance, it only takes into account the year as an explanatory variable. Also, pooled OLS ignores the structure of panel data. It only works if there is no unobserved/fixed effect in the individual states, which does not appear to be the case from our EDA. This can lead to omitted variable bias and make our estimates inconsistent. For example, if the explanatory variable(s) are positively correlated with the unobserved effect, we will get an upward bias.

(15 points) Expanded Model

In this next section, we add a number of explanatory variables. Many of the variables represent changes in state laws that occurred over the course of the data collection, including blood alcohol limits, seat belt laws, per se DUI laws, and highway speed limits. These variables were transformed into indicator variables, to show whether or not the law was in effect for a given observation.

```r
# Fit expanded linear model
exp_lm_model <- lm(log_fatility_rate ~ year_of_observation + bac + perSe + sbprim + sbsecon +
                     sl70plus + gdl + perc14_24 + log_unem + log_vehicmilespc, data = data)

exp_lm_model %>% summary()
```

```
##
## Call:
## lm(formula = log_fatility_rate ~ year_of_observation + bac +
##     perSe + sbprim + sbsecon + sl70plus + gdl + perc14_24 + log_unem +
##     log_vehicmilespc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58173 -0.12555  0.00035  0.13886  0.62244
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.124e+01  4.016e-01 -27.993  < 2e-16 ***
## year_of_observation1981 -9.191e-02  4.111e-02  -2.236  0.02556 *
## year_of_observation1982 -2.938e-01  4.198e-02  -6.999 4.33e-12 ***
## year_of_observation1983 -3.480e-01  4.287e-02  -8.118 1.20e-15 ***
## year_of_observation1984 -2.982e-01  4.367e-02  -6.828 1.38e-11 ***
## year_of_observation1985 -3.364e-01  4.456e-02  -7.550 8.77e-14 ***
## year_of_observation1986 -3.134e-01  4.641e-02  -6.753 2.28e-11 ***
## year_of_observation1987 -3.497e-01  4.839e-02  -7.226 8.96e-13 ***
## year_of_observation1988 -3.606e-01  5.092e-02  -7.081 2.47e-12 ***
## year_of_observation1989 -4.461e-01  5.288e-02  -8.435  < 2e-16 ***
## year_of_observation1990 -5.056e-01  5.405e-02  -9.355  < 2e-16 ***
## year_of_observation1991 -6.207e-01  5.518e-02 -11.249  < 2e-16 ***
## year_of_observation1992 -7.264e-01  5.626e-02 -12.912  < 2e-16 ***
## year_of_observation1993 -7.184e-01  5.695e-02 -12.616  < 2e-16 ***
## year_of_observation1994 -7.047e-01  5.809e-02 -12.132  < 2e-16 ***
## year_of_observation1995 -6.849e-01  5.945e-02 -11.521  < 2e-16 ***
## year_of_observation1996 -8.066e-01  6.152e-02 -13.111  < 2e-16 ***
## year_of_observation1997 -8.282e-01  6.293e-02 -13.160  < 2e-16 ***
## year_of_observation1998 -8.704e-01  6.387e-02 -13.629  < 2e-16 ***
## year_of_observation1999 -8.718e-01  6.487e-02 -13.438  < 2e-16 ***
## year_of_observation2000 -8.837e-01  6.602e-02 -13.386  < 2e-16 ***
## year_of_observation2001 -9.370e-01  6.675e-02 -14.037  < 2e-16 ***
## year_of_observation2002 -9.809e-01  6.712e-02 -14.616  < 2e-16 ***
## year_of_observation2003 -1.004e+00  6.744e-02 -14.884  < 2e-16 ***
## year_of_observation2004 -9.868e-01  6.904e-02 -14.293  < 2e-16 ***
## bac0.08                 -6.002e-02  2.617e-02  -2.294  0.02198 *
## bac0.1                  -1.675e-02  1.937e-02  -0.865  0.38737
## perSe                   -2.049e-02  1.455e-02  -1.408  0.15932
## sbprim                   1.226e-04  2.454e-02   0.005  0.99601
```

```
## sbsecon                   2.040e-02  2.139e-02   0.954  0.34029
## sl70plus                  2.325e-01  2.211e-02  10.516  < 2e-16 ***
## gdl                      -2.621e-02  2.609e-02  -1.005  0.31533
## perc14_24                 1.734e-02  6.095e-03   2.845  0.00452 **
## log_unem                  2.645e-01  2.412e-02  10.966  < 2e-16 ***
## log_vehicmilespc          1.537e+00  4.433e-02  34.665  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.201 on 1165 degrees of freedom
## Multiple R-squared:  0.6693, Adjusted R-squared:  0.6597
## F-statistic: 69.36 on 34 and 1165 DF,  p-value: < 2.2e-16
```

The model results show that year continues to have a significant effect, but we also get significant effects from bac0.08, speed limit, percent of population between 14 and 24, log unemployment, and log miles driven per capita.

In our dataframe, we defined the blood alcohol as the level that was prevalent for the majority of the year for the given year-state-level observation. For example, if the BAC was 0.1 for 40% of 1994 in Michigan and 0.08 for 60%, we encoded it to be 0.08. The coefficient for bac0.08 in the expanded linear model is about -0.06. This means that a BAC of 0.08 decreases fatality rate by exp(-0.06) = 0.94, or a decrease of 1 - 0.94 = 5.82%, when compared to no BAC regulation. This estimate is statistically significant. However, bac0.1 is not significant.

The per se law coefficient is also negative, but statistically insignificant. Primary seatbelt laws are positive, which is counterintuitive, but also statistically insignificant.

```
##  bac0.08
## 5.825546
```

```
##   bac0.1
## 1.660879
```

```
##    perSe
## 2.027866
```

(15 points) State-Level Fixed Effects

Below, we've used the expanded model from the previous section and have added a fixed effect at the state level.

```r
# Convert data to a plm dataframe
data_plm <- pdata.frame(data, index = c("state", "year_of_observation"))

# Fit fixed effects model
fixed_model <- plm(log_fatility_rate ~ year_of_observation + bac + perSe + sbprim + sbsecon +
                     sl70plus + gdl + perc14_24 + log_unem + log_vehicmilespc,
                   data = data_plm,
                   model = "within",
                   effect = "individual")
```

```
fixed_model %>% summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_fatility_rate ~ year_of_observation + bac +
##     perSe + sbprim + sbsecon + sl70plus + gdl + perc14_24 + log_unem +
##     log_vehicmilespc, data = data_plm, effect = "individual",
##     model = "within")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.3781276 -0.0518472  0.0042803  0.0533520  0.2895206
##
## Coefficients:
##                           Estimate Std. Error  t-value  Pr(>|t|)
## year_of_observation1981 -0.0632829  0.0180429  -3.5074 0.0004706 ***
## year_of_observation1982 -0.1350743  0.0189613  -7.1237 1.879e-12 ***
## year_of_observation1983 -0.1680215  0.0197305  -8.5158 < 2.2e-16 ***
## year_of_observation1984 -0.2069996  0.0205690 -10.0636 < 2.2e-16 ***
## year_of_observation1985 -0.2324012  0.0215216 -10.7985 < 2.2e-16 ***
## year_of_observation1986 -0.1958681  0.0230678  -8.4910 < 2.2e-16 ***
## year_of_observation1987 -0.2420504  0.0250549  -9.6608 < 2.2e-16 ***
## year_of_observation1988 -0.2727972  0.0274213  -9.9484 < 2.2e-16 ***
## year_of_observation1989 -0.3472699  0.0292452 -11.8744 < 2.2e-16 ***
## year_of_observation1990 -0.3571833  0.0303953 -11.7513 < 2.2e-16 ***
## year_of_observation1991 -0.3941686  0.0310859 -12.6800 < 2.2e-16 ***
## year_of_observation1992 -0.4545399  0.0321384 -14.1432 < 2.2e-16 ***
## year_of_observation1993 -0.4725584  0.0327401 -14.4336 < 2.2e-16 ***
## year_of_observation1994 -0.5045791  0.0336728 -14.9848 < 2.2e-16 ***
## year_of_observation1995 -0.5056554  0.0347207 -14.5635 < 2.2e-16 ***
## year_of_observation1996 -0.5570297  0.0366968 -15.1793 < 2.2e-16 ***
## year_of_observation1997 -0.5838161  0.0379082 -15.4008 < 2.2e-16 ***
## year_of_observation1998 -0.6361488  0.0387249 -16.4274 < 2.2e-16 ***
## year_of_observation1999 -0.6544343  0.0392683 -16.6657 < 2.2e-16 ***
## year_of_observation2000 -0.6866487  0.0398755 -17.2198 < 2.2e-16 ***
## year_of_observation2001 -0.6557711  0.0401510 -16.3326 < 2.2e-16 ***
## year_of_observation2002 -0.6172334  0.0404163 -15.2719 < 2.2e-16 ***
## year_of_observation2003 -0.6196464  0.0406447 -15.2454 < 2.2e-16 ***
## year_of_observation2004 -0.6569959  0.0417471 -15.7375 < 2.2e-16 ***
## bac0.08                 -0.0170452  0.0164076  -1.0389 0.2990939
## bac0.1                  -0.0121941  0.0113511  -1.0743 0.2829376
## perSe                   -0.0552572  0.0098322  -5.6200 2.409e-08 ***
## sbprim                  -0.0404837  0.0149762  -2.7032 0.0069717 **
## sbsecon                  0.0059170  0.0109924   0.5383 0.5904882
```

```
## sl70plus                 0.0769801  0.0117235   6.5663 7.882e-11 ***
## gdl                     -0.0215866  0.0127401  -1.6944 0.0904699 .
## perc14_24                0.0194307  0.0041618   4.6688 3.396e-06 ***
## log_unem                -0.1925304  0.0171667 -11.2154 < 2.2e-16 ***
## log_vehicmilespc         0.6778581  0.0507376  13.3601 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    31.924
## Residual Sum of Squares: 8.6606
## R-Squared:      0.72871
## Adj. R-Squared: 0.70906
## F-statistic: 88.3264 on 34 and 1118 DF, p-value: < 2.22e-16
```

The coefficient for bac0.08 in this fixed effects model is about -0.02, which is less extreme than in the expanded linear model, but this time is insignificant. The coefficient of the per se variable is -0.06, which is three times as extreme as the linear model, and this time it is significant. This equates to a 5.38% decrease in fatality rate. Unlike in the linear model, primary seatbelt law is statistically significant. It also has a much more practical effect, -0.04, which equates to a 3.97% decrease in fatality rate.

The assumptions for the linear model are:

- Observations are i.i.d.
- Homoscedasticity
- Normality
- No multicollinearity

The assumptions for the fixed effects model are:

- Individuals are i.i.d.
- No serial correlation in the error term (i.e. expectation = 0)
- No perfect multicollinearity
- Homoscedasticity (i.e. error term has constant variance)
- There is a fixed effect that is correlated with at least one of the explanatory variables

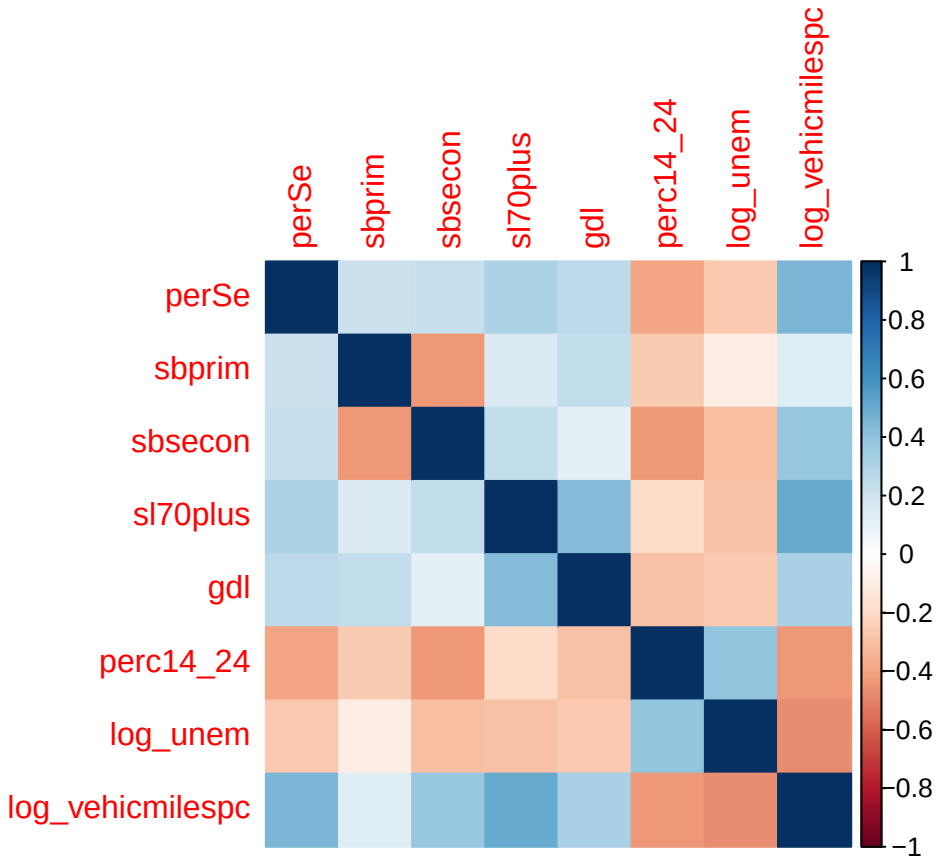
The assumptions of the linear model and fixed effects model are very similar. Because we're looking at panel data, we know the observations aren't i.i.d. in the linear assumption sense. A state observation from one year is not independent from an observation for the same state in the following year. However, we could argue that states are independent from other states, in which case the i.i.d. assumption holds well enough for fixed effects.

```
##    perSe
## 5.375827
```

```
##   sbprim
## 3.967516
```

We turn to the heatmap below to check whether there is multicollinearity. There are no numeric variables that are dangerously close to 1 or -1.

```
# Calculate the correlation matrix
cor_matrix <- cor(data[, c("perSe", "sbprim", "sbsecon", "sl70plus",
                           "gdl", "perc14_24", "log_unem", "log_vehicmilespc")])

# Create a correlation heatmap
corrplot::corrplot(cor_matrix, method = "color")
```



We will test for serial correlation and homoscedasticity in the final section of this report.

As has already been noted, pooled OLS ignores the structure of the panel data and the possibility of a fixed effect. So, the fixed effect model likely has a more reliable result. This is especially true since there is likely a state-fixed effect, which we'll discuss more in the next section. To be certain, we run a pFtest(), which shows a highly significant p-value, indicating we should use the fixed effects model.

```
pFtest(fixed_model, exp_lm_model)
```

```
##
##  F test for individual effects
##
## data:  log_fatility_rate ~ year_of_observation + bac + perSe + sbprim +  ...
## F = 105.52, df1 = 47, df2 = 1118, p-value < 2.2e-16
```

```
## alternative hypothesis: significant effects
```

(10 points) Consider a Random Effects Model

The assumptions of a random effects model are: * All of the fixed effect assumptions * The unobserved individual effect is independent of all explanatory variables

We cannot be sure that the unobserved individual effect is independent of the explanatory variables. For example, it's certainly possible that unobserved attitudes about drinking and driving at the individual (i.e. state) level are tied to state-level policies about the legally accepted BAC. Because these assumptions aren't met with this data, a random effects model is not likely to be consistent, may give biased estimates, and may return incorrect standard errors.

To be sure, we conduct a Hausman test. The p-value in the results is very small, so we reject the null hypothesis that random effects are appropriate. In other words, we can stick with the fixed effects model.

```
# Test for random effects
phtest(fixed_model, random_model)
```

```
##
##  Hausman Test
##
## data:  log_fatility_rate ~ year_of_observation + bac + perSe + sbprim +  ...
## chisq = 78.868, df = 34, p-value = 2.006e-05
## alternative hypothesis: one model is inconsistent
```

(10 points) Model Forecasts

For this section, we found data for the amount of vehicle miles driven in the U.S. from January 2018 up to May 2023. This data was retrieved from the Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/series/TRFVOLUSM227NFWA), who compiled monthly millions of vehicle miles driven using data from the U.S. Federal Highway Administration (https://highways.dot.gov/). To convert to per capita data, we also retrieved US population data from the same source (https://fred.stlouisfed.org/series/POPTHM). We define the U.S. pandemic era as March 2020 through May 2023 (basically up until the data ends).

```
monthly_miles <- read_csv("../data/monthly_miles_driven.csv")

# Show first few rows
head(monthly_miles)
```

```
## # A tibble: 6 x 2
##   date     millions_of_miles
##   <chr>                <dbl>
## 1 1/1/2018            244736
## 2 2/1/2018            227759
## 3 3/1/2018            270705
## 4 4/1/2018            275127
## 5 5/1/2018            283713
```

```
## 6 6/1/2018            282648
```

```
monthly_population <- read_csv("../data/POPTHM.csv")

# Show first few rows
head(monthly_population)
```

```
## # A tibble: 6 x 2
##   DATE       POPTHM
##   <date>      <dbl>
## 1 2018-01-01 327969
## 2 2018-02-01 328085
## 3 2018-03-01 328219
## 4 2018-04-01 328364
## 5 2018-05-01 328521
## 6 2018-06-01 328692
```

When comparing 2018 to COVID-era monthly data, the largest decrease in per capita driving was in April 2020 with -39.68%. The largest increase in per capita driving was in September 2022 with +0.69%.

```
# What month demonstrated the largest decrease in driving?
covid_bust_date <- comparison_data$date[which.min(comparison_data$perc_change)]
print(covid_bust_date)
```

```
## [1] "2020-04-01"
```

```
covid_bust_perc <- min(comparison_data$perc_change)
print(covid_bust_perc)
```

```
## [1] -39.68631
```

```
covid_bust_tot <- min(comparison_data$total_log_change)
print(covid_bust_tot)
```

```
## [1] -0.5056112
```

```
#What month demonstrated the largest increase in driving?
covid_boom_date <- comparison_data$date[which.max(comparison_data$perc_change)]
print(covid_boom_date)
```

```
## [1] "2022-09-01"
```

```
covid_boom_perc <- max(comparison_data$perc_change)
print(covid_boom_perc)
```

```
## [1] 0.6897441
```

```
covid_boom_tot <- max(comparison_data$total_log_change)
print(covid_boom_tot)
```

```
## [1] 0.006873763
```

For these two months, we can estimate the percentage change in total fatality rate that we'd expect

to see by utilizing the coefficient for miles driven per capita from our fixed model. Taking advantage of the log transformations in our variables, we estimate a 34% decrease in traffic fatalities for April 2020, and a modest 0.5% increase in fatalities in September 2022.

```
forecast_ci
```

```
##   scenario      month   lower estimate   upper
## 1     Boom 2022-09-01  0.0040   0.0047  0.0053
## 2     Bust 2020-04-01 -0.2925  -0.3427 -0.3930
```

(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?

When there's no omitted variable bias, even in the presence of serial correlation or heteroskedasticity, the OLS estimators remain unbiased. This means that on average, the OLS estimator will be correct. However, OLS estimators will no longer be the Best Linear Unbiased Estimators, meaning there might be other linear estimators that have smaller variances. Heteroskedasticity: When heteroskedasticity is present, the usual OLS standard errors are generally inconsistent. This can lead to incorrect inference, such as invalid t-statistics and confidence intervals. Serial Correlation: In the presence of serial correlation in a time series context, the standard OLS standard errors are not valid. This can again lead to misleading t-statistics and confidence intervals. Hypothesis tests rely on valid standard errors. If standard errors are incorrect due to serial correlation or heteroskedasticity, these test statistics can be misleading. This can lead to incorrect rejections or failures to reject the null hypothesis.

In this section, we test for serial correlation and heteroskedasticity.

```
pbgtest(fixed_model)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  log_fatility_rate ~ year_of_observation + bac + perSe + sbprim +  ...
## chisq = 243.19, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pdwtest(fixed_model)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  log_fatility_rate ~ year_of_observation + bac + perSe + sbprim +  ...
## DW = 1.2136, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

To test for serial correlation, we look at two tests: The Durbin-Watson test and the Breusch-Pagan test.

When using the Durbin-Watson test, we get a p-value of .496, which means we fail to reject the null hypothesis that there is no serial correlation. However, the Durbin-Watson test is only limited

to singular lag, which makes it a less robust test than the Breusch-Godfrey test. In order to be confident in our assessment, we used the Breusch-Godfrey test, which resulted in a p-value significantly less than .05, which means that we reject the null hypothesis and conclude that there is serial correlation.

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

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  log_fatility_rate ~ year_of_observation + bac + perSe + sbprim +     sbsecon + sl70pl
## chisq = 2745.2, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

To test for heteroskedasticity, we employed the Breusch-Pagan test. Our p-value was well below .05, which means that we reject the null hypothesis and conclude that there is heteroskedasticity.