

# Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 4

*Heather Feinstein, Ryan Delgado, April Kim*

*Fall 2018*

## Instructions:

- **Due Date: 12/11/2018 (11:59 p.m. Pacific Time)**
- **Page limit of the pdf report: 20 (not include title and the table of content page)**
- Use the margin, linespace, and font size specification below:
  - fontsize=11pt
  - margin=1in
  - line\_spacing=single
- Submission:
  - Each group makes one submission to Github; please have one of your team members made the submission
  - Submit 2 files:
    1. A pdf file including the details of your analysis and all the R codes used to produce the analysis. Please do not suppress the codes in your pdf file.
    2. R markdown file used to produce the pdf file
  - Use the following file-naming convensation; fail to do so will receive 10% reduction in the grade:
    - \* FirstNameLastName1\_FirstNameLastName2\_FirstNameLastName3\_LabNumber.fileExtension
    - \* For example, if you have three students in the group for Lab Z, and their names are Gerard Kelley, Steve Yang, and Jeffrey Yau, then you should name your file the following
      - GerardKelley\_SteveYang\_JeffreyYau\_LabZ.Rmd
      - GerardKelley\_SteveYang\_JeffreyYau\_LabZ.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your pdf and Rmd files.
- This lab can be completed in a group of up to 3 students in your session. Students are encouraged to work in a group for the lab.
- For statistical methods that we cover in this course, use only the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you have to provide (1) explanation of why such libraries and functions are used instead and (2) reference to the suppressWarnings(suppressMessages(library documentation. Lacking the explanation and reference to the documentation will result in a score of zero for the corresponding question.
- Students are expected to act with regards to UC Berkeley Academic Integrity.

# Description of the Lab

In this lab, you are asked to answer the question “**Do changes in traffic laws affect traffic fatalities?**” To do so, you will conduct the tasks specified below using the data set *driving.Rdata*, which 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 come with the dataste.

## Introduction

\* Car accidents are bad \* There many be changes to laws we can make to prevent them \* This lab will examine how changes in traffic laws affect fatalities \* The findings will have policy implications

### Exercises:

1. Load the data. Provide a description of the basic structure of the dataset, as we have done throughout the semester. Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable *totfatrte* and the potential explanatory variables. You need to write a detailed narrative of your observations of your EDA. *Reminder: giving an “output dump” (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive a zero in this exercise.*

### Initial Examination

```
load("driving.RData")

# view data and descriptions
str(data)
```

```
## 'data.frame':    1200 obs. of  56 variables:
## $ year          : int  1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
## $ state         : int   1  1  1  1  1  1  1  1  1  1 ...
## $ sl55          : num   1  1  1  1  1 ...
## $ sl65          : num   0  0  0  0  0 ...
## $ sl70          : num   0  0  0  0  0  0  0  0  0  0 ...
## $ sl75          : num   0  0  0  0  0  0  0  0  0  0 ...
## $ slnone        : num   0  0  0  0  0  0  0  0  0  0 ...
## $ seatbelt      : int   0  0  0  0  0  0  0  0  0  0 ...
## $ minage        : num  18 18 18 18 18 20 21 21 21 21 ...
## $ zeroto1       : num   0  0  0  0  0  0  0  0  0  0 ...
## $ gdl           : num   0  0  0  0  0  0  0  0  0  0 ...
## $ bac10         : num   1  1  1  1  1  1  1  1  1  1 ...
## $ bac08         : num   0  0  0  0  0  0  0  0  0  0 ...
## $ perse         : num   0  0  0  0  0  0  0  0  0  0 ...
## $ totfat        : int  940 933 839 930 932 882 1080 1111 1024 1029 ...
## $ nghtfat       : int  422 434 376 397 421 358 500 499 423 418 ...
## $ wkndfat       : int  236 248 224 223 237 224 279 300 226 247 ...
```

```

## $ totfatpvm : num 3.2 3.35 2.81 3 2.83 ...
## $ nghtfatpvm : num 1.44 1.56 1.26 1.28 1.28 ...
## $ wkndfatpvm : num 0.803 0.89 0.75 0.719 0.72 ...
## $ statepop : int 3893888 3918520 3925218 3934109 3951834 3972527 3991569 4015261 4023858 4030222
## $ totfatrte : num 24.1 24.1 21.4 23.6 23.6 ...
## $ nghtfatrte : num 10.84 11.08 9.58 10.09 10.65 ...
## $ wkndfatrte : num 6.06 6.33 5.71 5.67 6 ...
## $ vehicmiles : num 29.4 27.9 29.9 31 32.9 ...
## $ unem : num 8.8 10.7 14.4 13.7 11.1 ...
## $ perc14_24 : num 18.9 18.7 18.4 18 17.6 ...
## $ sl70plus : num 0 0 0 0 0 0 0 0 0 0 ...
## $ sbprim : int 0 0 0 0 0 0 0 0 0 0 ...
## $ sbsecon : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d80 : int 1 0 0 0 0 0 0 0 0 0 ...
## $ d81 : int 0 1 0 0 0 0 0 0 0 0 ...
## $ d82 : int 0 0 1 0 0 0 0 0 0 0 ...
## $ d83 : int 0 0 0 1 0 0 0 0 0 0 ...
## $ d84 : int 0 0 0 0 1 0 0 0 0 0 ...
## $ d85 : int 0 0 0 0 0 1 0 0 0 0 ...
## $ d86 : int 0 0 0 0 0 0 1 0 0 0 ...
## $ d87 : int 0 0 0 0 0 0 0 1 0 0 ...
## $ d88 : int 0 0 0 0 0 0 0 0 1 0 ...
## $ d89 : int 0 0 0 0 0 0 0 0 0 1 ...
## $ d90 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d91 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d92 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d93 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d94 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d95 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d96 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d97 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d98 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d99 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d00 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d01 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d02 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d03 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ d04 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ vehicmilesperc : num 7544 7108 7607 7880 8334 ...
## - attr(*, "datalabel")= chr ""
## - attr(*, "time.stamp")= chr "22 Jan 2013 14:09"
## - attr(*, "formats")= chr "%8.0g" "%8.0g" "%9.0g" "%9.0g" ...
## - attr(*, "types")= int 252 251 254 254 254 254 254 251 254 254 ...
## - attr(*, "val.labels")= chr "" "" "" "" ...
## - attr(*, "var.labels")= chr "1980 through 2004" "48 continental states, alphabetical" "speed limit"
## - attr(*, "version")= int 12

```

```
print(paste("The number of NA values in the dataset is ", sum(is.na(data))))
```

```
## [1] "The number of NA values in the dataset is 0"
```

**\*\*Notes:\*\***

- There are 1200 pooled observations in the dataset, with each observation having 56 variables. There are no instances of NA values in the data, indicating that we have a balanced panel dataset.
- The variables can roughly be divided into several categories
- Dummy variables that indicate what laws are implemented in a particular state for a specific year, e.g. sl55-sl75, seatbelt.
- Non-discrete variables that measure fatality counts, and ratios of fatalities to the population. This includes our outcome variable of interest, `totfatrte`
- “Demographic” variables like `unem` (state unemployment rate), `perc14_24` (percentage of population aged 14 to 24).
- Dummy variables for each year in the panel.
- “Index” variables that indicate which year and state the data corresponds to (aptly named `year` and `state`).

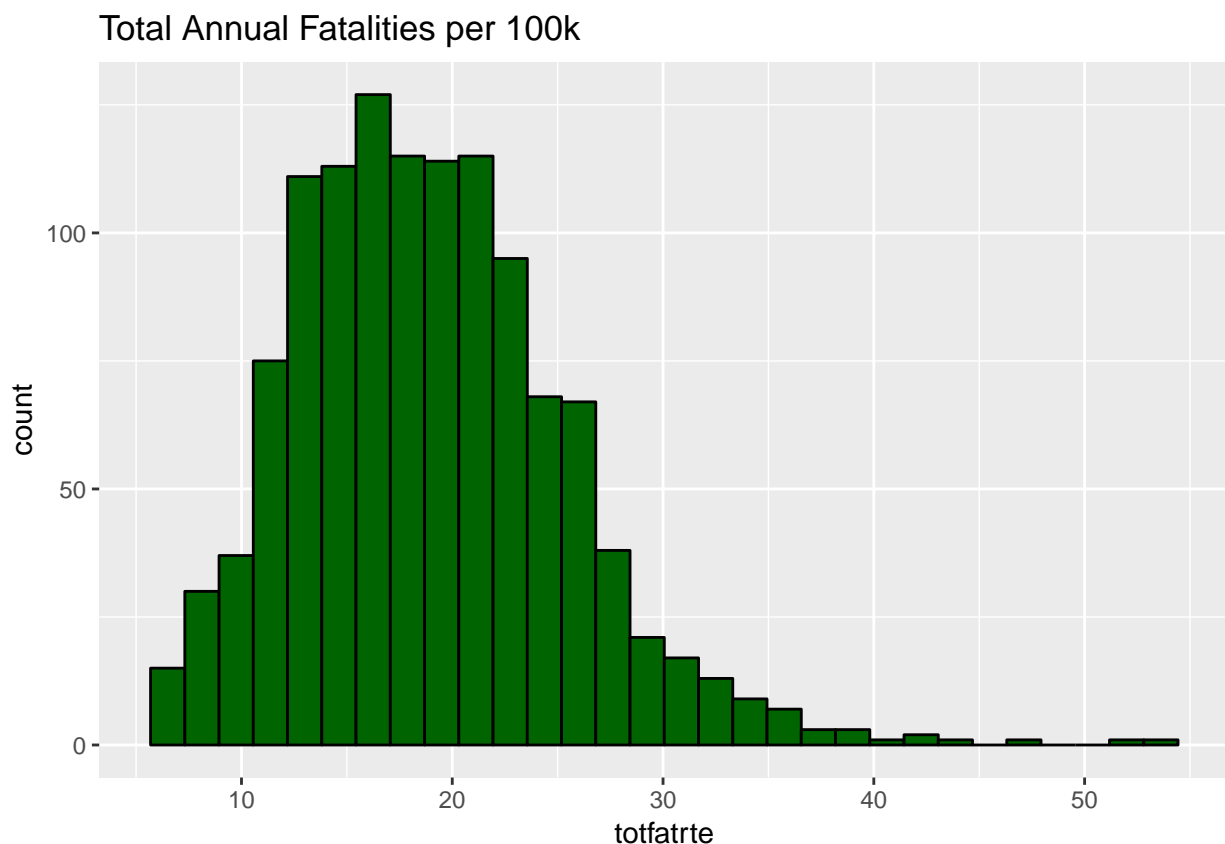
We’ll examine these different variable types separately in our EDA.

## Exploratory Data Analysis

Let’s start by looking at our outcome variable, `totfatrte`. We’ll plot a histogram of this variable:

```
ggplot(data, aes(x = totfatrte)) + geom_histogram(fill = "darkgreen",
  color = "black") + ggtitle("Total Annual Fatalities per 100k")
```

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



Observations:

- We see that the dataset is skewed right. This is typical of zero-bounded variables.
- There are a few cases where the fatalities are above 50 people per 100k. We will examine those in more detail to see if there's an explanation

Let's look at the outlier observations:

```
cols.wo.yrdummies <- c("year", "state", "sl55", "sl65", "sl70",
  "sl75", "slnone", "seatbelt", "minage", "zerotol", "gdl",
  "bac10", "bac08", "perse", "totfat", "nghtfat", "wkndfat",
  "totfatpvm", "nghtfatpvm", "wkndfatpvm", "statepop", "totfatrte",
  "nghtfatrte", "wkndfatrte", "vehicmiles", "unem", "perc14_24",
  "sl70plus", "sbprim", "sbsecon")

t(data[data$totfatrte > 50, cols.wo.yrdummies])
```

##	1176	1177
## year	1980.00	1981.000
## state	51.00	51.000
## sl55	1.00	1.000
## sl65	0.00	0.000
## sl70	0.00	0.000
## sl75	0.00	0.000
## slnone	0.00	0.000
## seatbelt	0.00	0.000
## minage	18.00	18.000
## zerotol	0.00	0.000
## gdl	0.00	0.000
## bac10	0.00	0.000
## bac08	0.00	0.000
## perse	0.00	0.000
## totfat	245.00	264.000
## nghtfat	139.00	134.000
## wkndfat	58.00	68.000
## totfatpvm	4.90	5.056
## nghtfatpvm	2.78	2.566
## wkndfatpvm	1.16	1.302
## statepop	469557.00	491713.000
## totfatrte	52.18	53.320
## nghtfatrte	29.60	27.250
## wkndfatrte	12.35	13.830
## vehicmiles	5.00	5.222
## unem	4.00	4.100
## perc14_24	19.60	18.500
## sl70plus	0.00	0.000
## sbprim	0.00	0.000
## sbsecon	0.00	0.000

Both observations are in state 51 and occur in the early 1980s. It appears this state's 1980 population (~470k) is around 10% of the average state population that year (~4.67 million). This low state population could explain the large fatality ratios in those years. The other variables for this state in these years don't see particularly extreme. **reword?**

**need to find out what vehicmilespc is...** A: vehicle miles per capita; decimal

Let's now look at histograms for `perc14_24`, `unem`, and `vehicmilespc`:

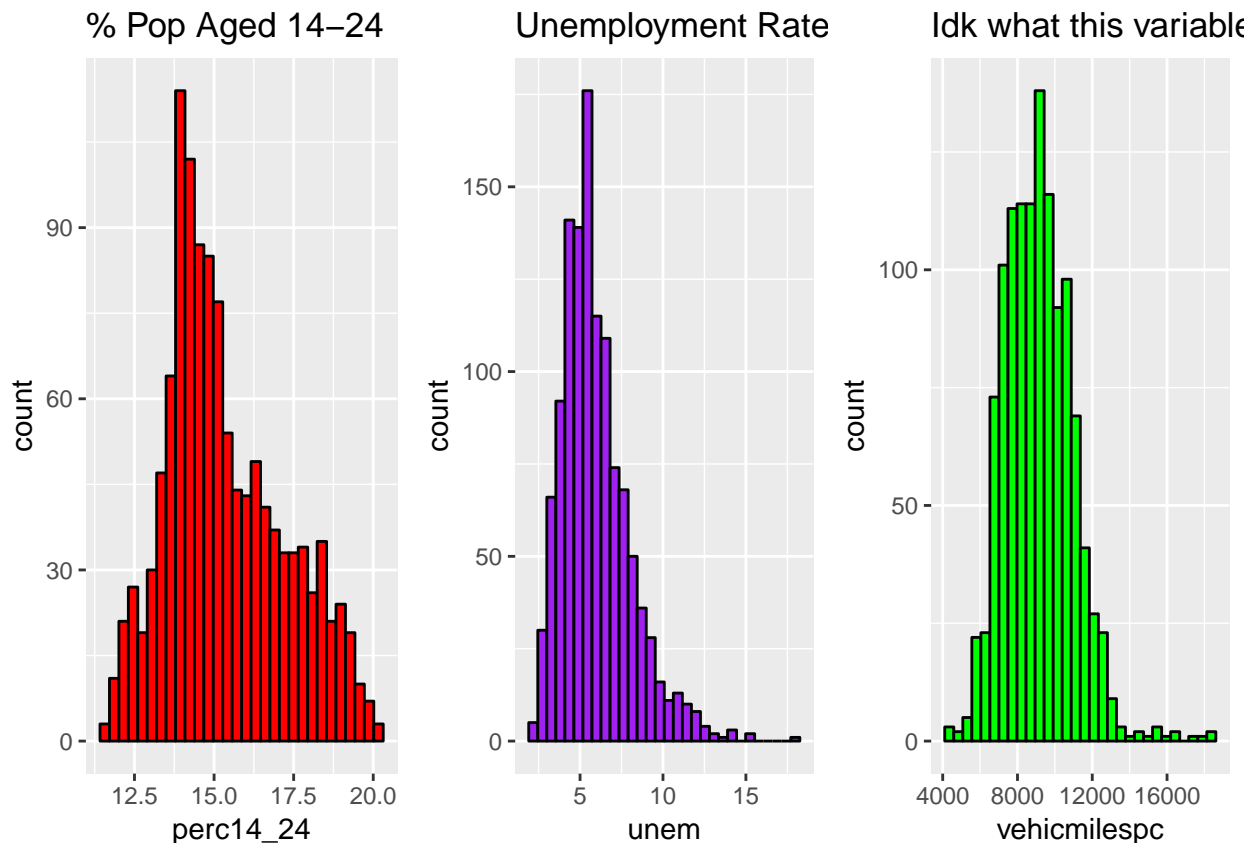
```
perc.hist <- ggplot(data, aes(x = perc14_24)) + geom_histogram(color = "black",
  fill = "red") + ggtitle("% Pop Aged 14-24")

unem.hist <- ggplot(data, aes(x = unem)) + geom_histogram(color = "black",
  fill = "purple") + ggtitle("Unemployment Rate")

vehicpc.hist <- ggplot(data, aes(x = vehicmilespc)) + geom_histogram(color = "black",
  fill = "green") + ggtitle("Idk what this variable means")

grid.arrange(perc.hist, unem.hist, vehicpc.hist, ncol = 3, nrow = 1)
```

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



These distributions look fairly ordinary. Similar to the `totfatrtc` histogram, we see right-ward skewness in the `unem` and `vehicmilespc`. Any other commentary. Maybe we can talk about log transforming these variables to deal with the skewness? Is this a good enough reason to log transform?

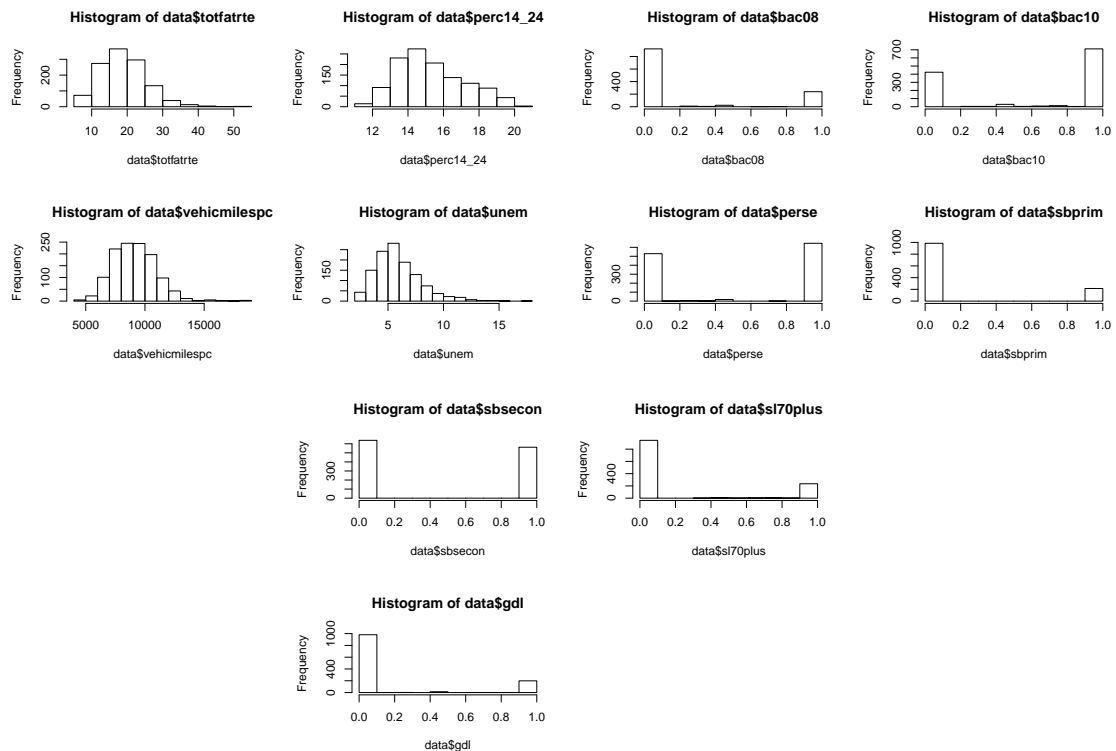
Let's plot histograms of the discrete variables now: gotta do this.

Questions I want to answer in the EDA: -Do these variables change over time? if not, comment how they'd be eliminated in the panel models unless we interact them with year. -Is there enough variance in them to be good candidate EVs in the model? -Does anything stick out as needing a transformation? e.g. rounding. Explain why this is a good idea.

```

par(mfrow = c(2, 2))
hist(data$totfatrte)
hist(data$perc14_24)
hist(data$vehicmilespc)
hist(data$unem)
hist(data$bac08)
hist(data$bac10)
hist(data$perse)
hist(data$sbprim)
hist(data$sbsecon)
hist(data$sl70plus)
hist(data$gdl)

```

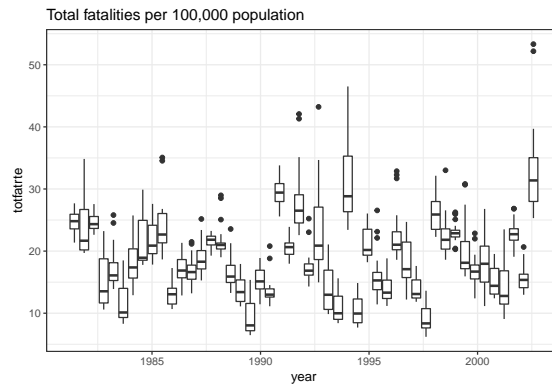


Variables displays some degree of skewness and may need to transform. Additionally, some variables that appear to be binary have values between 0 and 1 to indicate mid year changes. We will round these variables.

```

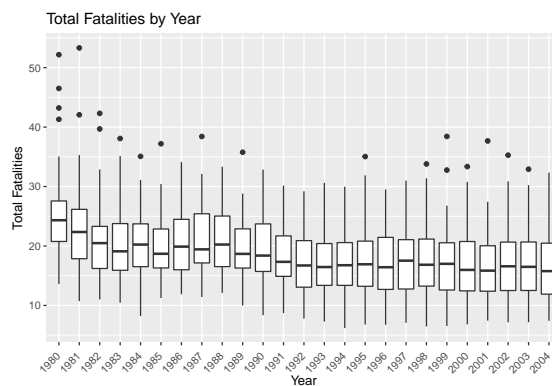
ggplot(data, aes(x = year, y = totfatrte, group = as.factor(state))) +
  geom_boxplot() + ggtitle("Total fatalities per 100,000 population") +
  theme_bw()

```

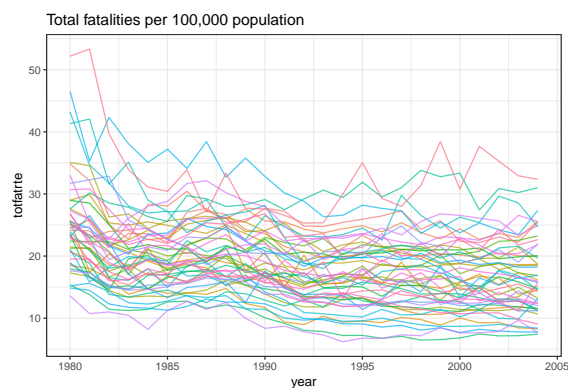


States display a large difference in distribution of total fatality rate.

```
# ryan's plot
ggplot(data, aes(as.factor(year), totfatrte)) + geom_boxplot() +
  ggtitle("Total Fatalities by Year") + xlab("Year") + ylab("Total Fatalities") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1.2))
```



```
# totfatrte: total fatalities per 100,000 population
ggplot(data, aes(x = year, y = totfatrte, colour = as.factor(state))) +
  geom_line(alpha = 0.7, show.legend = F) + ggtitle("Total fatalities per 100,000 population") +
  theme_bw()
```

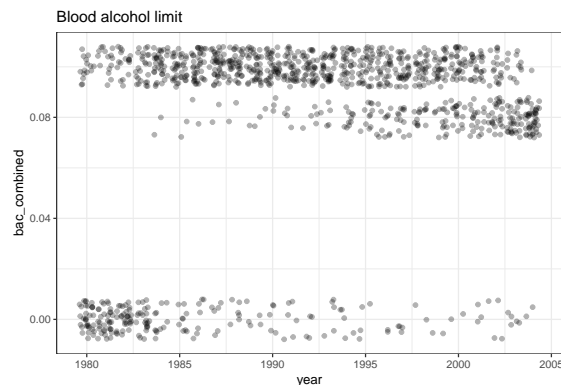


Total fatalities decreased from 1980 to 1985 and remained roughly constant post 1985.



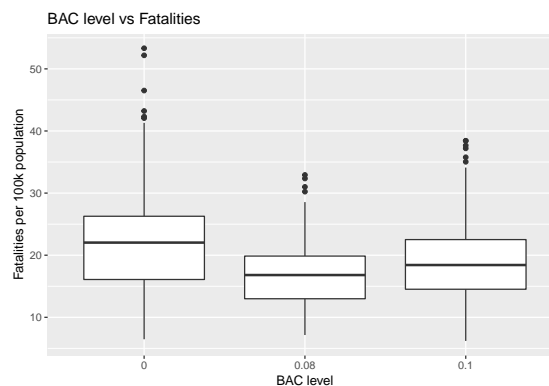
```
# if '1' in bac10, set as 0.1; else set as '0' or '1' in
# bac08
data$bac_combined <- ifelse(round(data$bac10) > 0, 0.1, 0.08 *
  round(data$bac08))
data$sl_combined <- ifelse(round(data$sl55) > 0, 55, ifelse(round(data$sl65) >
  0, 65, ifelse(round(data$sl70) > 0, 70, ifelse(round(data$sl75) >
  0, 75, 80))))
```

```
ggplot(data, aes(x = year, y = bac_combined, group = state)) +
  geom_jitter(alpha = 0.3) + theme_bw() + ggtitle("Blood alcohol limit")
```



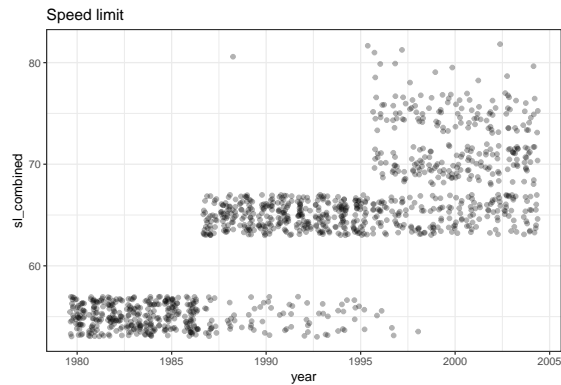
We noticed over time that fewer states have no BAC limit or 0.1 limit and more have a 0.08 limit.

```
ggplot(data, aes(x = as.factor(bac_combined), y = totfatrte)) +
  geom_boxplot() + ggtitle("BAC level vs Fatalities") + xlab("BAC level") +
  ylab("Fatalities per 100k population")
```



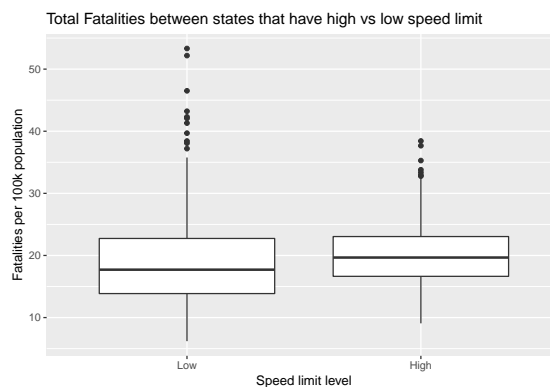
Fatalities appear to be lowest in states with a 0.08 BAC limit and highest in states with no BAC limit.

```
ggplot(data, aes(x = year, y = sl_combined, group = state)) +
  geom_jitter(alpha = 0.3) + theme_bw() + ggtitle("Speed limit")
```

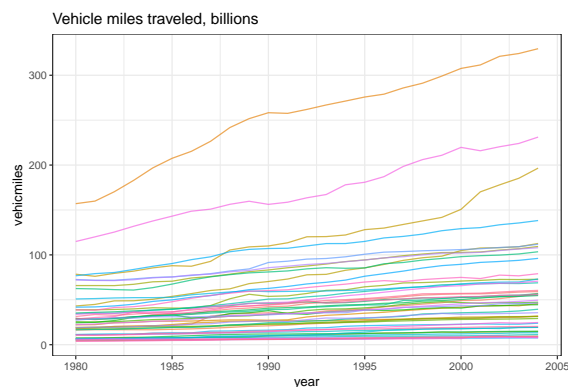


Over time, speed limits have increased.

```
data$slhigh <- (data$sl70plus == 1)
ggplot(data, aes(x = as.factor(slhigh), y = totfatrte)) + geom_boxplot() +
  ggtitle("Total Fatalities between states that have high vs low speed limit") +
  scale_x_discrete(labels = c("Low", "High")) + xlab("Speed limit level") +
  ylab("Fatalities per 100k population")
```

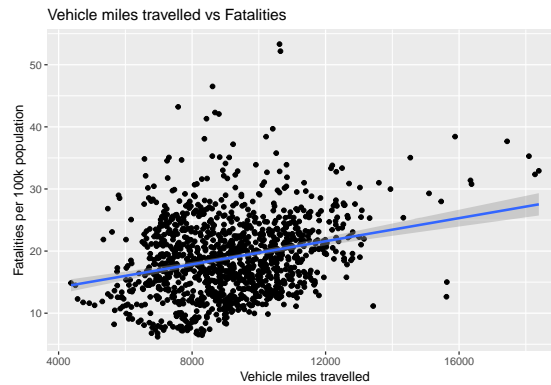


```
# vehicmiles vehicle miles traveled, billions
ggplot(data, aes(x = year, y = vehicmiles, colour = as.factor(state))) +
  geom_line(alpha = 0.7, show.legend = F) + ggtitle("Vehicle miles traveled, billions") +
  theme_bw()
```



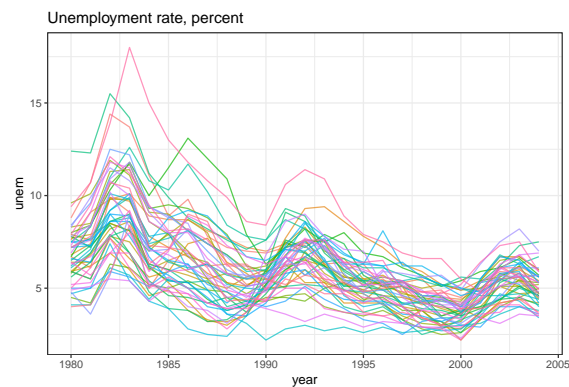
Miles traveled have increased over time.

```
ggplot(data, aes(x = vehicmilespc, y = totfatrte)) + geom_point() +
  geom_smooth(method = lm) + ggtitle("Vehicle miles travelled vs Fatalities") +
  xlab("Vehicle miles travelled") + ylab("Fatalities per 100k population")
```



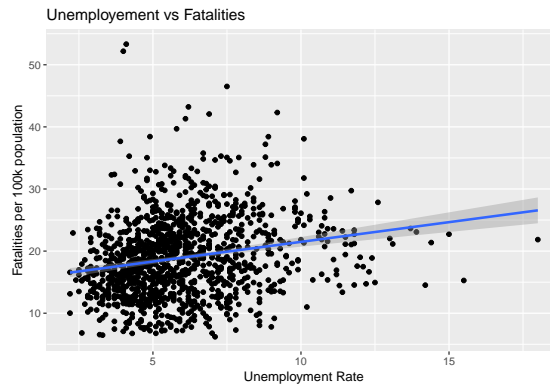
More miles travelled correlates with more fatalities.

```
# unem unemployment rate, percent
ggplot(data, aes(x = year, y = unem, colour = as.factor(state))) +
  geom_line(alpha = 0.7, show.legend = F) + ggtitle("Unemployment rate, percent") +
  theme_bw()
```



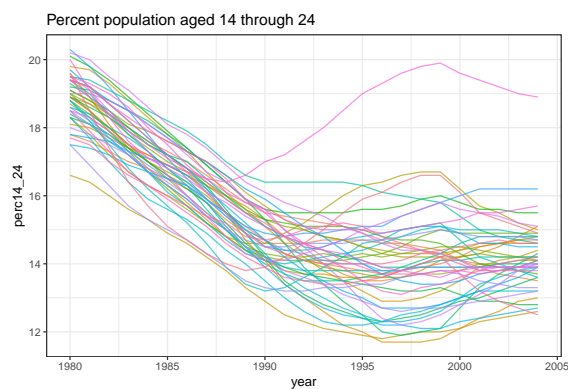
Increase in unemployment rate in early 80s followed by decrease until early 90s; steady decrease from mid 90s.

```
ggplot(data, aes(x = unem, y = totfatrte)) + geom_point() + geom_smooth(method = lm) +
  ggtitle("Unemployment vs Fatalities") + xlab("Unemployment Rate") +
  ylab("Fatalities per 100k population")
```



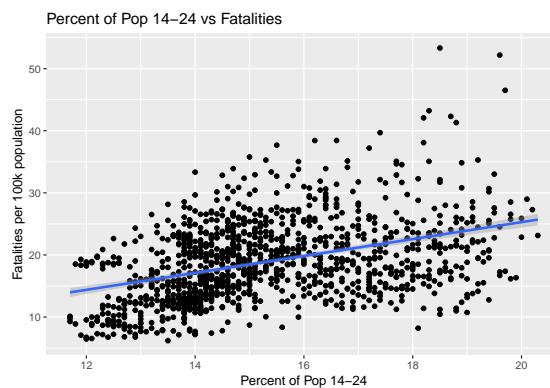
Higher unemployment correlates with more fatalities.

```
# perc14_24 percent population aged 14 through 24
ggplot(data, aes(x = year, y = perc14_24, colour = as.factor(state))) +
  geom_line(alpha = 0.7, show.legend = F) + ggtitle("Percent population aged 14 through 24") +
  theme_bw()
```



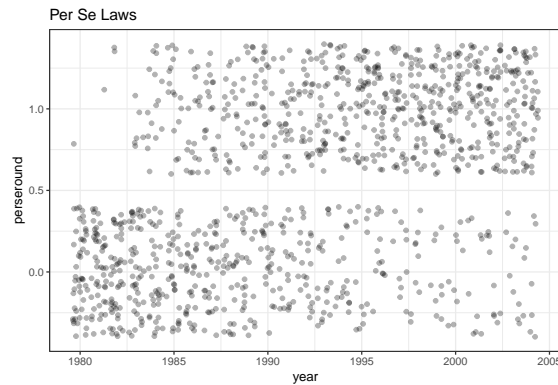
Rapid decrease in percent population aged 14 through 24 until early 90s; slight increase afterwards.

```
ggplot(data, aes(x = perc14_24, y = totfatrte)) + geom_point() +
  geom_smooth(method = lm) + ggtitle("Percent of Pop 14-24 vs Fatalities") +
  xlab("Percent of Pop 14-24") + ylab("Fatalities per 100k population")
```



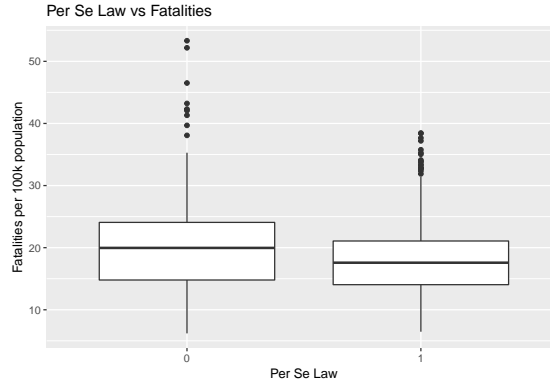
Higher percent of population 14-24 correlates with more fatalities.

```
data$perseround = round(data$perse)
ggplot(data, aes(x = year, y = perseround, group = state)) +
  geom_jitter(alpha = 0.3) + theme_bw() + ggtitle("Per Se Laws")
```



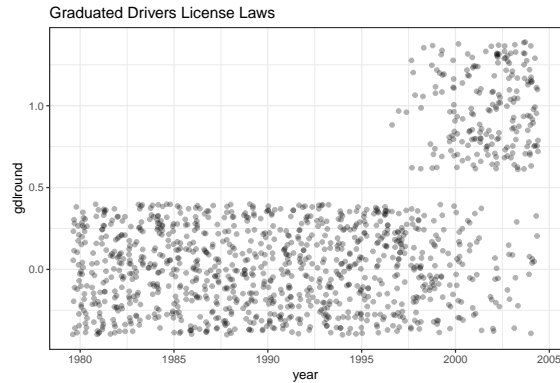
States have increasingly implemented per se laws over time.

```
ggplot(data, aes(x = as.factor(perseround), y = totfatrte)) +
  geom_boxplot() + ggtitle("Per Se Law vs Fatalities") + xlab("Per Se Law") +
  ylab("Fatalities per 100k population")
```



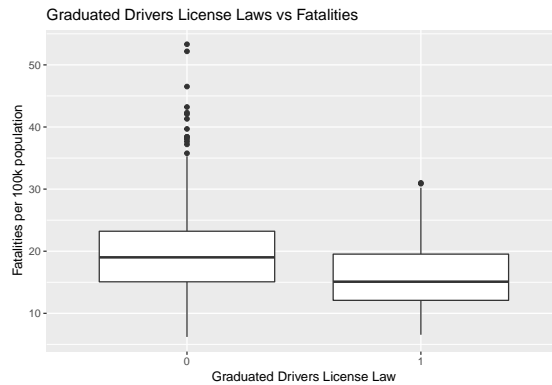
States with no per se laws have higher fatalities

```
data$gdlround = round(data$gdl)
ggplot(data, aes(x = year, y = gdlround, group = state)) + geom_jitter(alpha = 0.3) +
  theme_bw() + ggtitle("Graduated Drivers License Laws")
```



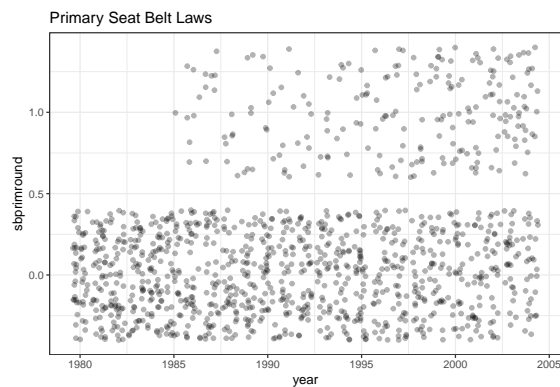
States have increasingly implemented graduated drivers license laws beginning in the late 90s.

```
ggplot(data, aes(x = as.factor(gdlround), y = totfatrte)) + geom_boxplot() +
  ggtitle("Graduated Drivers License Laws vs Fatalities") +
  xlab("Graduated Drivers License Law") + ylab("Fatalities per 100k population")
```



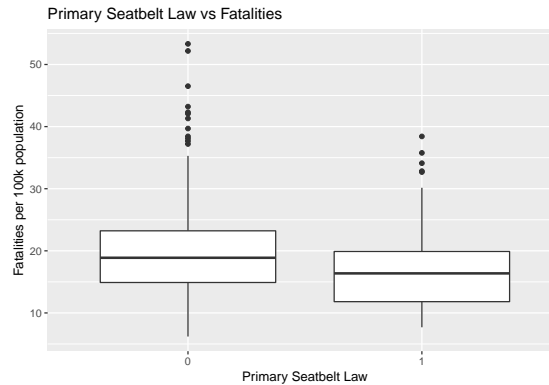
States with without graduated drivers license laws have higher fatalities.

```
data$sbprimround = round(data$sbprim)
ggplot(data, aes(x = year, y = sbprimround, group = state)) +
  geom_jitter(alpha = 0.3) + theme_bw() + ggtitle("Primary Seat Belt Laws")
```



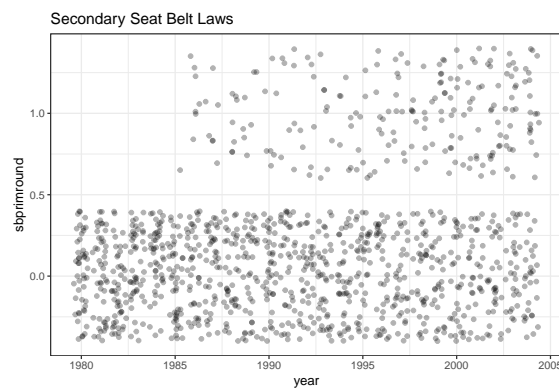
States have increasingly implemented primary seatbelt laws over time.

```
data$sbprimround = round(data$sbprim)
ggplot(data, aes(x = as.factor(sbprimround), y = totfatrte)) +
  geom_boxplot() + ggtitle("Primary Seatbelt Law vs Fatalities") +
  xlab("Primary Seatbelt Law") + ylab("Fatalities per 100k population")
```



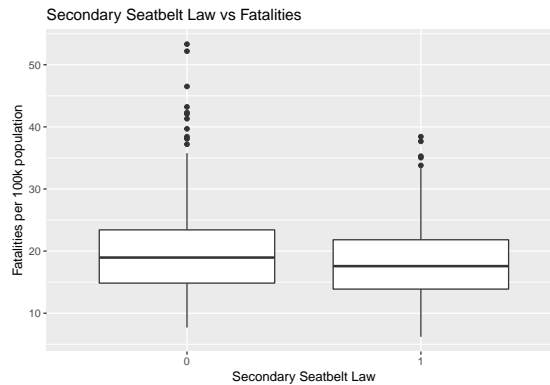
States with no primary seatbelt law have higher fatalities.

```
data$sbseconround = round(data$sbsecon)
ggplot(data, aes(x = year, y = sbprimround, group = state)) +
  geom_jitter(alpha = 0.3) + theme_bw() + ggtitle("Secondary Seat Belt Laws")
```



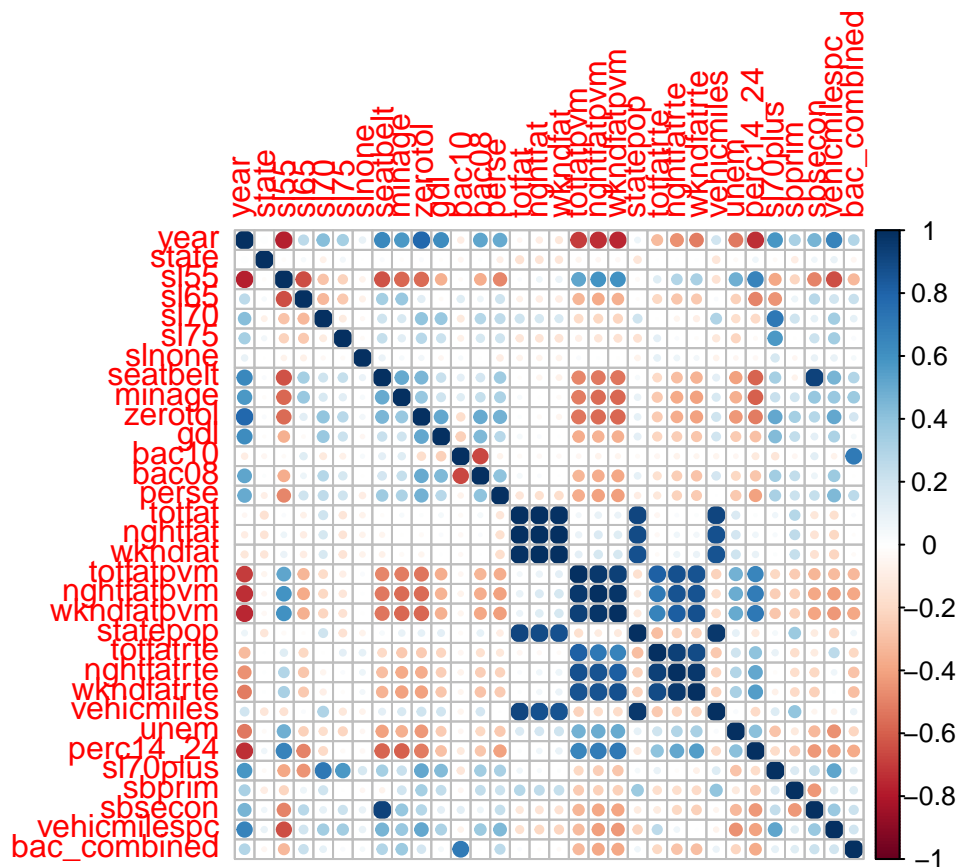
States have increasingly implemented secondary seatbelt laws over time.

```
data$sbseconround = round(data$sbsecon)
ggplot(data, aes(x = as.factor(sbseconround), y = totfatrte)) +
  geom_boxplot() + ggtitle("Secondary Seatbelt Law vs Fatalities") +
  xlab("Secondary Seatbelt Law") + ylab("Fatalities per 100k population")
```



States with no secondary seatbelt law have higher fatalities.

```
data1 <- data[, -(31:55)]
data1 <- data1[, -(33:39)]
m = cor(data1)
corrplot(m, method = "circle")
```



2. How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this model explain? Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.



*totfatrte* is defined as “nighttime fatalities per 100,000 population”

```
# avg per year covered in data set
ddply(data, .(year), summarize, Total = mean(totfatrte))
```

```
##   year   Total
## 1 1980 25.49458
## 2 1981 23.67021
## 3 1982 20.94250
## 4 1983 20.15292
## 5 1984 20.26750
## 6 1985 19.85146
## 7 1986 20.80042
## 8 1987 20.77479
## 9 1988 20.89167
## 10 1989 19.77229
## 11 1990 19.50521
## 12 1991 18.09479
## 13 1992 17.15792
## 14 1993 17.12771
## 15 1994 17.15521
## 16 1995 17.66854
## 17 1996 17.36938
## 18 1997 17.61062
## 19 1998 17.26542
## 20 1999 17.25042
## 21 2000 16.82562
## 22 2001 16.79271
## 23 2002 17.02958
## 24 2003 16.76354
## 25 2004 16.72896
```

We’ll estimate the linear regression model on the year dummies using the *totfatrte* column and the *year* column converted to a factor:

```
# linear model
mod1 <- lm(totfatrte ~ factor(year), data = data)
summary(mod1)
```

```
##
## Call:
## lm(formula = totfatrte ~ factor(year), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    25.4946     0.8671  29.401  < 2e-16 ***
## factor(year)1981  -1.8244     1.2263  -1.488  0.137094
## factor(year)1982  -4.5521     1.2263  -3.712  0.000215 ***
## factor(year)1983  -5.3417     1.2263  -4.356  1.44e-05 ***
```

```
## factor(year)1984 -5.2271      1.2263 -4.263 2.18e-05 ***
## factor(year)1985 -5.6431      1.2263 -4.602 4.64e-06 ***
## factor(year)1986 -4.6942      1.2263 -3.828 0.000136 ***
## factor(year)1987 -4.7198      1.2263 -3.849 0.000125 ***
## factor(year)1988 -4.6029      1.2263 -3.754 0.000183 ***
## factor(year)1989 -5.7223      1.2263 -4.666 3.42e-06 ***
## factor(year)1990 -5.9894      1.2263 -4.884 1.18e-06 ***
## factor(year)1991 -7.3998      1.2263 -6.034 2.14e-09 ***
## factor(year)1992 -8.3367      1.2263 -6.798 1.68e-11 ***
## factor(year)1993 -8.3669      1.2263 -6.823 1.43e-11 ***
## factor(year)1994 -8.3394      1.2263 -6.800 1.66e-11 ***
## factor(year)1995 -7.8260      1.2263 -6.382 2.51e-10 ***
## factor(year)1996 -8.1252      1.2263 -6.626 5.25e-11 ***
## factor(year)1997 -7.8840      1.2263 -6.429 1.86e-10 ***
## factor(year)1998 -8.2292      1.2263 -6.711 3.01e-11 ***
## factor(year)1999 -8.2442      1.2263 -6.723 2.77e-11 ***
## factor(year)2000 -8.6690      1.2263 -7.069 2.67e-12 ***
## factor(year)2001 -8.7019      1.2263 -7.096 2.21e-12 ***
## factor(year)2002 -8.4650      1.2263 -6.903 8.32e-12 ***
## factor(year)2003 -8.7310      1.2263 -7.120 1.88e-12 ***
## factor(year)2004 -8.7656      1.2263 -7.148 1.54e-12 ***
## ---
## 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:  0.1276, Adjusted R-squared:  0.1098
## F-statistic: 7.164 on 24 and 1175 DF, p-value: < 2.2e-16
```

The summary shows that the coefficients are estimated to be negative for each year after 1980, with the coefficients being statistically significant for all years except 1981. This corroborates the observations we made in the box plot of `totfatrte` over time. Does this mean that driving became safer over this time period? The answer to that question is multi-faceted - better driving habits, more/less speeding, frequency of drunk driving, car safety. **need to elaborate and reword this portion.**

- Expand your model in *Exercise 2* by adding variables `bac08`, `bac10`, `perse`, `sbprim`, `sbsecon`, `sl70plus`, `gdl`, `perc14_24`, `unem`, `vehicmiles`, and perhaps *transformations of some or all of these variables*. Please explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed. How are the variables `bac8` and `bac10` defined? Interpret the coefficients on `bac8` and `bac10`. Do *per se* laws have a negative effect on the fatality rate? What about having a primary seat belt law? (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.)

I think we should log transform `unem` Should we include a latex rendering of the model we'll estimate? We may score extra points by doing that... **yesssss we def should**

```
data$bac08round = round(data$bac08)
data$bac10round = round(data$bac10)
data$sl70plusround = round(data$sl70plus)

mod2 <- lm(totfatrte ~ factor(year) + bac08round + bac10round +
  perseround + sbprimround + sbseconround + sl70plusround +
  gdlround + perc14_24 + log(unem) + vehicmiles, data = data)
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = totfatrte ~ factor(year) + bac08round + bac10round +
##     perseround + sbprimround + sbseconround + sl70plusround +
##     gdlround + perc14_24 + log(unem) + vehicmilespc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4031  -2.6086  -0.3265   2.2414  21.8650
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.012e+00  2.620e+00  -3.058 0.002277 **
## factor(year)1981 -2.107e+00  8.229e-01  -2.560 0.010578 *
## factor(year)1982 -6.304e+00  8.397e-01  -7.508 1.19e-13 ***
## factor(year)1983 -7.190e+00  8.515e-01  -8.445 < 2e-16 ***
## factor(year)1984 -5.826e+00  8.666e-01  -6.723 2.78e-11 ***
## factor(year)1985 -6.458e+00  8.852e-01  -7.296 5.48e-13 ***
## factor(year)1986 -5.634e+00  9.231e-01  -6.103 1.42e-09 ***
## factor(year)1987 -6.065e+00  9.613e-01  -6.309 3.98e-10 ***
## factor(year)1988 -6.176e+00  1.011e+00  -6.109 1.36e-09 ***
## factor(year)1989 -7.688e+00  1.049e+00  -7.325 4.43e-13 ***
## factor(year)1990 -8.682e+00  1.072e+00  -8.102 1.36e-15 ***
## factor(year)1991 -1.087e+01  1.093e+00  -9.944 < 2e-16 ***
## factor(year)1992 -1.263e+01  1.114e+00 -11.335 < 2e-16 ***
## factor(year)1993 -1.250e+01  1.128e+00 -11.085 < 2e-16 ***
## factor(year)1994 -1.208e+01  1.150e+00 -10.500 < 2e-16 ***
## factor(year)1995 -1.147e+01  1.180e+00  -9.722 < 2e-16 ***
## factor(year)1996 -1.340e+01  1.223e+00 -10.962 < 2e-16 ***
## factor(year)1997 -1.352e+01  1.244e+00 -10.864 < 2e-16 ***
## factor(year)1998 -1.420e+01  1.268e+00 -11.197 < 2e-16 ***
## factor(year)1999 -1.415e+01  1.284e+00 -11.019 < 2e-16 ***
## factor(year)2000 -1.440e+01  1.307e+00 -11.021 < 2e-16 ***
## factor(year)2001 -1.567e+01  1.317e+00 -11.903 < 2e-16 ***
## factor(year)2002 -1.649e+01  1.326e+00 -12.434 < 2e-16 ***
## factor(year)2003 -1.692e+01  1.331e+00 -12.716 < 2e-16 ***
## factor(year)2004 -1.633e+01  1.367e+00 -11.947 < 2e-16 ***
## bac08round     -2.288e+00  4.858e-01  -4.709 2.79e-06 ***
## bac10round      -1.256e+00  3.591e-01  -3.497 0.000489 ***
## perseround      -5.625e-01  2.919e-01  -1.927 0.054231 .
## sbprimround     -3.795e-01  4.898e-01  -0.775 0.438515
## sbseconround    -1.535e-01  4.279e-01  -0.359 0.719911
## sl70plusround    3.112e+00  4.331e-01   7.186 1.19e-12 ***
## gdlround        -3.014e-01  5.066e-01  -0.595 0.552051
## perc14_24        1.776e-01  1.222e-01   1.453 0.146542
## log(unem)        5.152e+00  4.812e-01  10.707 < 2e-16 ***
## vehicmilespc     2.921e-03  9.393e-05  31.096 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.024 on 1165 degrees of freedom
```

```
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6006
## F-statistic: 54.02 on 34 and 1165 DF,  p-value: < 2.2e-16
```

bac8 is blood alcohol limit .08 bac10 is blood alcohol limit .10

4. Reestimate the model from *Exercise 3* using a fixed effects (at the state level) model. How do the coefficients on *bac08*, *bac10*, *perse*, and *sbprim* compare with the pooled OLS estimates? Which set of estimates do you think is more reliable? What assumptions are needed in each of these models? Are these assumptions reasonable in the current context?

```
data$bac08round <- round(data$bac08)
data$bac10round <- round(data$bac10)
data$sl70plusround <- round(data$sl70plus)
data$perseround <- round(data$perse)
data$sbprimround <- round(data$sbprim)
data$sbseconround = round(data$sbsecon)
data$gdlround = round(data$gdl)

model.fe <- plm(totfatrtte ~ factor(year) + bac08round + bac10round +
  perseround + sbprimround + sbseconround + sl70plusround +
  gdlround + perc14_24 + log(unem) + vehicmiles, data = data,
  index = c("state", "year"), model = "within")

summary(model.fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrtte ~ factor(year) + bac08round + bac10round +
##      perseround + sbprimround + sbseconround + sl70plusround +
##      gdlround + perc14_24 + log(unem) + vehicmiles, data = data,
##      model = "within", index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.   Median   3rd Qu.    Max.
## -8.24190 -1.03559 -0.01383  0.97540 14.63249
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## factor(year)1981 -1.5789e+00  4.1364e-01 -3.8171 0.0001424 ***
## factor(year)1982 -3.3715e+00  4.3392e-01 -7.7699 1.771e-14 ***
## factor(year)1983 -4.0253e+00  4.4540e-01 -9.0376 < 2.2e-16 ***
## factor(year)1984 -4.5466e+00  4.5989e-01 -9.8863 < 2.2e-16 ***
## factor(year)1985 -4.9961e+00  4.8064e-01 -10.3947 < 2.2e-16 ***
## factor(year)1986 -3.9860e+00  5.1473e-01 -7.7438 2.152e-14 ***
## factor(year)1987 -4.6704e+00  5.5481e-01 -8.4180 < 2.2e-16 ***
## factor(year)1988 -5.2103e+00  6.0465e-01 -8.6170 < 2.2e-16 ***
## factor(year)1989 -6.5240e+00  6.4330e-01 -10.1416 < 2.2e-16 ***
## factor(year)1990 -6.5808e+00  6.6583e-01 -9.8835 < 2.2e-16 ***
## factor(year)1991 -7.2511e+00  6.8047e-01 -10.6560 < 2.2e-16 ***
## factor(year)1992 -8.1283e+00  7.0115e-01 -11.5928 < 2.2e-16 ***
```

```
## factor(year)1993 -8.4679e+00 7.1473e-01 -11.8476 < 2.2e-16 ***
## factor(year)1994 -8.9443e+00 7.3383e-01 -12.1885 < 2.2e-16 ***
## factor(year)1995 -8.7095e+00 7.5904e-01 -11.4743 < 2.2e-16 ***
## factor(year)1996 -9.1278e+00 8.0059e-01 -11.4013 < 2.2e-16 ***
## factor(year)1997 -9.3884e+00 8.2150e-01 -11.4283 < 2.2e-16 ***
## factor(year)1998 -1.0104e+01 8.4200e-01 -11.9999 < 2.2e-16 ***
## factor(year)1999 -1.0347e+01 8.5217e-01 -12.1422 < 2.2e-16 ***
## factor(year)2000 -1.0957e+01 8.6601e-01 -12.6518 < 2.2e-16 ***
## factor(year)2001 -1.0458e+01 8.6660e-01 -12.0674 < 2.2e-16 ***
## factor(year)2002 -9.6024e+00 8.6989e-01 -11.0387 < 2.2e-16 ***
## factor(year)2003 -9.6414e+00 8.7258e-01 -11.0493 < 2.2e-16 ***
## factor(year)2004 -1.0076e+01 8.9852e-01 -11.2135 < 2.2e-16 ***
## bac08round -1.1048e+00 3.3064e-01 -3.3413 0.0008616 ***
## bac10round -8.0362e-01 2.2577e-01 -3.5594 0.0003873 ***
## perseround -1.1270e+00 2.2337e-01 -5.0456 5.273e-07 ***
## sbprimround -1.1891e+00 3.4319e-01 -3.4649 0.0005505 ***
## sbseconround -3.0375e-01 2.5224e-01 -1.2042 0.2287712
## sl70plusround 4.7115e-02 2.6094e-01 0.1806 0.8567466
## gdlround -2.8250e-01 2.8028e-01 -1.0079 0.3137115
## perc14_24 1.6728e-01 9.5431e-02 1.7529 0.0798887 .
## log(unem) -3.7089e+00 3.9240e-01 -9.4519 < 2.2e-16 ***
## vehicmilespc 9.5062e-04 1.1018e-04 8.6281 < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares: 12134
## Residual Sum of Squares: 4552.4
## R-Squared: 0.62483
## Adj. R-Squared: 0.59765
## F-statistic: 54.7637 on 34 and 1118 DF, p-value: < 2.22e-16
```

5. Would you prefer to use a random effects model instead of the fixed effects model you built in *Exercise 4*? Please explain.

To determine whether random effects model should be used over the fixed effects model, we can conduct a Hausman test with the null hypothesis that the preferred model is random effects.

```
model.re <- plm(totfatrte ~ factor(year) + bac08round + bac10round +
  perseround + sbprimround + sbseconround + sl70plusround +
  gdlround + perc14_24 + log(unem) + vehicmilespc, data = data,
  index = c("state", "year"), model = "random")
phtest(model.fe, model.re)
```

```
##
## Hausman Test
##
## data: totfatrte ~ factor(year) + bac08round + bac10round + perseround + ...
## chisq = 180.54, df = 34, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

With the p-value < 2.2e-16, we can reject the null hypothesis that the random effects assumptions are correct and we would prefer to use the fixed effects model.

6. Suppose that *vehicmilespc*, the number of miles driven per capita, increases by 1,000. Using the FE estimates, what is the estimated effect on *totfatrte*? Please interpret the estimate.

According to our fixed effect model, the coefficient for *vehicmilespc* variable was 0.000951 fatalities/100k people per mile driven per capita. For all other things held equal, if, on average, there's an increase of 1,000 miles driven per capita, we would expect an increase of 0.951 (approximately 1) fatalities per 100k people.

7. If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors?

The fixed effects model assumes that the idiosyncratic errors are uncorrelated. If there is serial correlation in the model errors, the estimated variance will be biased which will result in underestimated standard errors and thus rendering most statistical tests invalid. This would most likely commit Type I error, and reject the null hypothesis too easily.

Heteroskedasticity in the idiosyncratic errors would result in overstated standard errors and may commit Type II error. We may fail to reject the null hypothesis since significance of potentially valuable regressor will not be detected.