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

Over the years, individual states have enacted many laws to to help prevent car accident deaths ranging from restricting driver alcohol intake to mandating seatbelts for passengers. These laws have been effective to varying degrees and have often corresponded with demographic changes that make effects more difficult to parse. Determing the effectiveness of driving laws has far reaching policy implications. This lab will explore how different traffic laws among US states and the changes in these laws over times have affected traffic fatalities. We'll explore a dataset of traffic laws and fatalities for the 48 contiguous US states, and apply panel regression techniques to this dataset to find relationships between traffic laws and fatalities.

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 count NA
head(data)
```

```
##   year state sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl bac10
## 1 1980     1    1    0    0    0      0        0     18       0   0     1
## 2 1981     1    1    0    0    0      0        0     18       0   0     1
## 3 1982     1    1    0    0    0      0        0     18       0   0     1
## 4 1983     1    1    0    0    0      0        0     18       0   0     1
## 5 1984     1    1    0    0    0      0        0     18       0   0     1
## 6 1985     1    1    0    0    0      0        0     20       0   0     1
##   bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm
## 1     0     0    940     422     236      3.20      1.437      0.803
## 2     0     0    933     434     248      3.35      1.558      0.890
## 3     0     0    839     376     224      2.81      1.259      0.750
## 4     0     0    930     397     223      3.00      1.281      0.719
```

```
## 5     0     0    932     421     237      2.83      1.278      0.720
## 6     0     0    882     358     224      2.51      1.019      0.637
##   statepop totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24
## 1  3893888     24.14      10.84       6.06   29.37500  8.8      18.9
## 2  3918520     24.07      11.08       6.33   27.85200 10.7      18.7
## 3  3925218     21.37       9.58       5.71   29.85765 14.4      18.4
## 4  3934109     23.64      10.09       5.67   31.00000 13.7      18.0
## 5  3951834     23.58      10.65       6.00   32.93286 11.1      17.6
## 6  3972527     22.20       9.01       5.64   35.13944  8.9      17.3
##   sl70plus sbprim sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91
## 1        0      0       0   1   0   0   0   0   0   0   0   0   0   0   0
## 2        0      0       0   0   1   0   0   0   0   0   0   0   0   0   0
## 3        0      0       0   0   0   1   0   0   0   0   0   0   0   0   0
## 4        0      0       0   0   0   0   1   0   0   0   0   0   0   0   0
## 5        0      0       0   0   0   0   0   1   0   0   0   0   0   0   0
## 6        0      0       0   0   0   0   0   0   1   0   0   0   0   0   0
##   d92 d93 d94 d95 d96 d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc
## 1   0   0   0   0   0   0   0   0   0   0   0   0   0     7543.874
## 2   0   0   0   0   0   0   0   0   0   0   0   0   0     7107.785
## 3   0   0   0   0   0   0   0   0   0   0   0   0   0     7606.622
## 4   0   0   0   0   0   0   0   0   0   0   0   0   0     7879.802
## 5   0   0   0   0   0   0   0   0   0   0   0   0   0     8333.562
## 6   0   0   0   0   0   0   0   0   0   0   0   0   0     8845.614
```

```
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 be divided into several categories:
  1) Dummy variables that indicate what laws are implemented in a particular state for a specific
  2) Dummy variables for each year in the panel.
  3) Continuous variables that measure fatality counts, and ratios of fatalities to the population. This includes our outcome variable of interest, `totfatrte`
  4) Continuous demographic variables like `unem` (state unemployment rate), `perc14_24` (percentage of population aged 14 to 24).
  5) "Index" variables that indicate which year and state the data corresponds to (aptly named `year` and `state`).

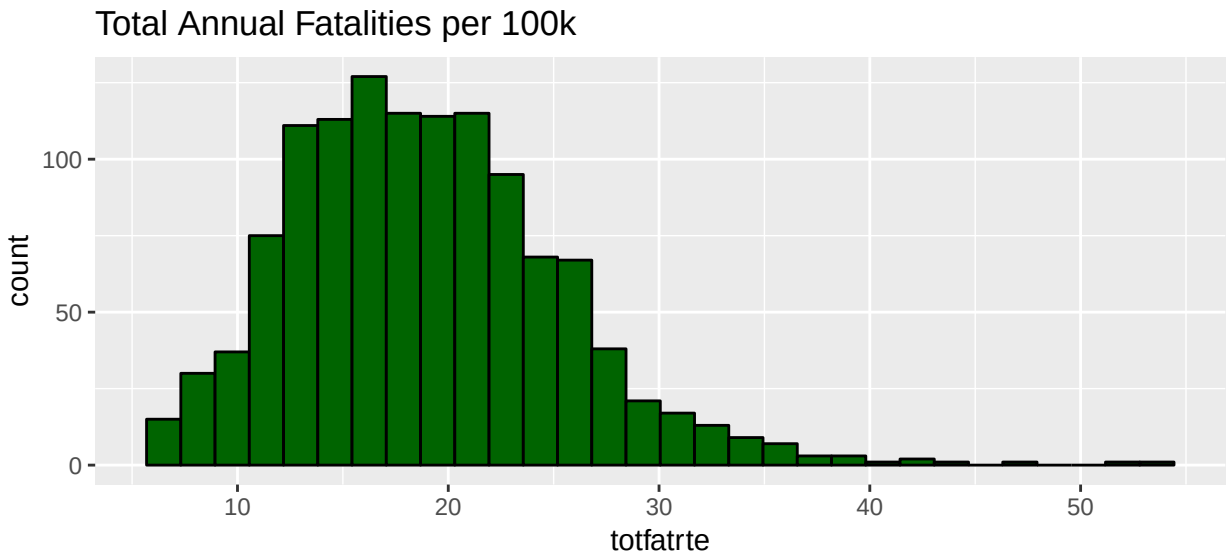
We'll examine these different variable types separately in our Exploratory Data Analysis.

Exploratory Data Analysis

Let's start by looking at our outcome variable, `totfatrte`. This variable measures the Total Number of Fatalities per 100 thousand individuals in the state/year. We'll first plot a histogram of this variable to understand its distribution:

```
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 variable is skewed right. This is typical of zero-bounded variables.
- There are a few outliers where the fatalities are above 50 people per 100k. Let's examine the values of the other variables for those observations 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
```

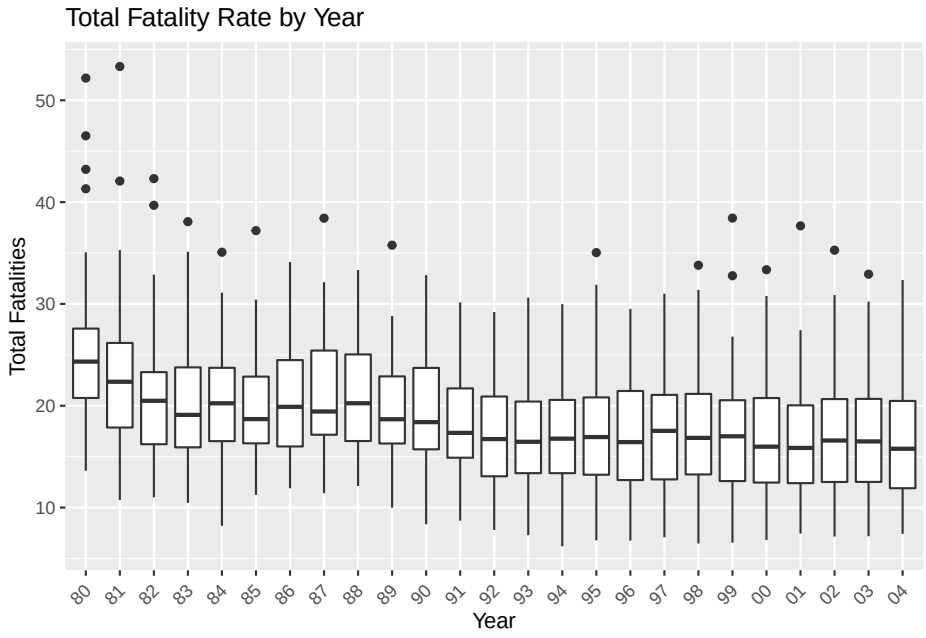
```
## statepop   469557.00 491713.000
## totfatrte      52.18     53.320
## nghtfatrte     29.60     27.250
## wkndfatrte     12.35     13.830
## vehicmiles      5.00      5.222
## unem            4.00      4.100
## perc14_24      19.60     18.500
## sl70plus        0.00      0.000
## sbprim          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 look particularly noteworthy.

Now let's look at how `totfatrte` changes over time:

```
# Get a list of the years w/o the millenium/century so the x axis looks cleaner.
years.nocent <- data %>% distinct(as.character(year)) %>% lapply(substr, 3, 4) %>% unlist(use.names=FALS

ggplot(data, aes(as.factor(year), totfatrte)) +
  geom_boxplot() +
  ggtitle('Total Fatality Rate by Year') +
  xlab('Year') + ylab('Total Fatalities') +
  scale_x_discrete(labels = years.nocent) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1.2))
```



The Total Fatality Rate has declined gradually but not steadily since the 80s. The variance across states remains fairly steady over time. It will be interesting to learn the drivers for this decline later in our analysis.

Let's now look at histograms for the continuous variables in the dataset, `perc14_24`, `unem`, and `vehicmilespc`. Variable definitions:

- **perc14_24** - The percentage of the population between the ages of 14 and 24.
- **unem** - The state-level unemployment rate.

- **vehicmilespc** - The number of vehicle miles per capita. This measures how much driving each person does, on average.

Like before, we'll plot histogramas of these variables to learn their distributions:

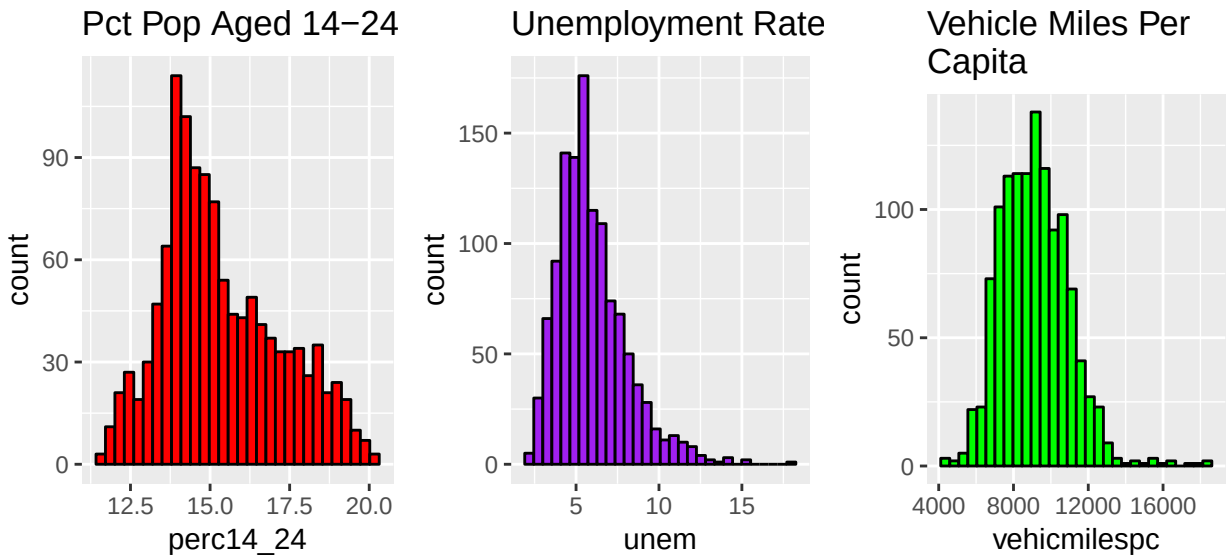
```
vehicpc.hist <- ggplot(data, aes(x=vehicmilespc)) +
  geom_histogram(color='black', fill='green') +
  ggtitle('Vehicle Miles Per \nCapita')

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

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

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 do not appear normally distributed, but nothing sticks out as unusual in them. Similar to the `totfatrte` histogram, we see right-ward skewness in the `unem` and `vehicmilespc`.

How do these variables change over time?

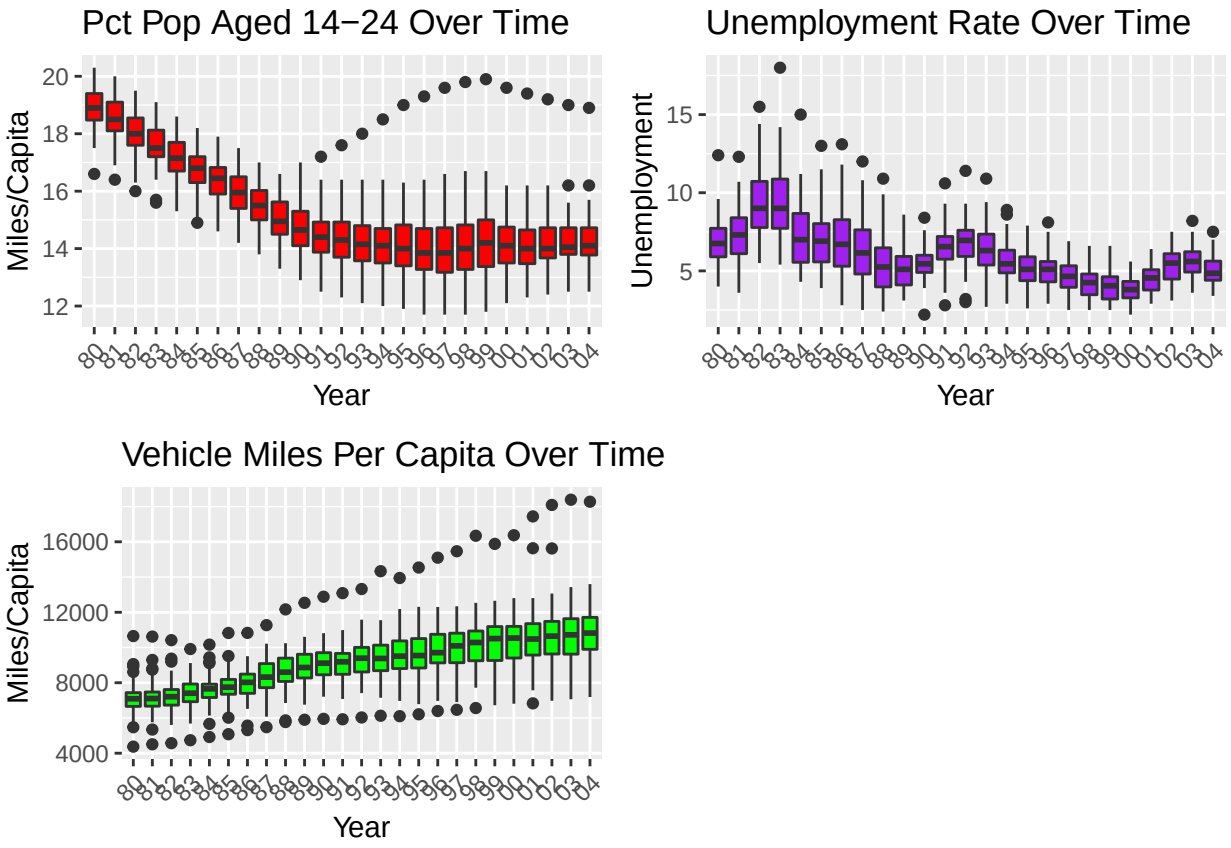
```
vehicpc.ot <- ggplot(data, aes(x=as.factor(year), y=vehicmilespc)) +
 geom_boxplot(fill='green') +
 ggtitle('Vehicle Miles Per Capita Over Time') +
 xlab('Year') + ylab('Miles/Capita') +
 scale_x_discrete(labels = years.nocent) +
 theme(axis.text.x = element_text(angle = 45, hjust = 1.0))

unem.ot <- ggplot(data, aes(x=as.factor(year), y=unem)) +
 geom_boxplot(fill='purple') +
 ggtitle('Unemployment Rate Over Time') +
```

```
 xlab('Year') + ylab('Unemployment') +
  scale_x_discrete(labels = years.nocent) +
 theme(axis.text.x = element_text(angle = 45, hjust = 1.0))

perc.ot <- ggplot(data, aes(x=as.factor(year), y=perc14_24)) +
 geom_boxplot(fill='red') +
 ggtitle('Pct Pop Aged 14-24 Over Time') +
 xlab('Year') + ylab('Miles/Capita') +
  scale_x_discrete(labels = years.nocent) +
 theme(axis.text.x = element_text(angle = 45, hjust = 1.0))

grid.arrange(perc.ot, unem.ot, vehicpc.ot, ncol=2, nrow=2)
```



perc14_24 and vehicmilespc both show a strong trend over time, with perc14_24 trending downward until stabilizing in the early 90s and vehicmilespc steadily trending upward. unem appears to show a very weak downward trend over the time period with cylical behavior (i.e. it's higher in times of recession). Both perc14_24 and vehicmilespc have a state that appears to be a persistent outlier. In perc14_24, the outlier appears to diverge from the stabilizing trend in the early 90s. The outliers in vehicmilespc trend upward with the prevailing trend, so that's not particularly interesting.

How do these variables correlate with totfatrte? We'll explore this by visualizing the absolute relationships between totfatrte and the relationships between the yearly first-differences. Let's look at scatter plots of the absolute variables first:

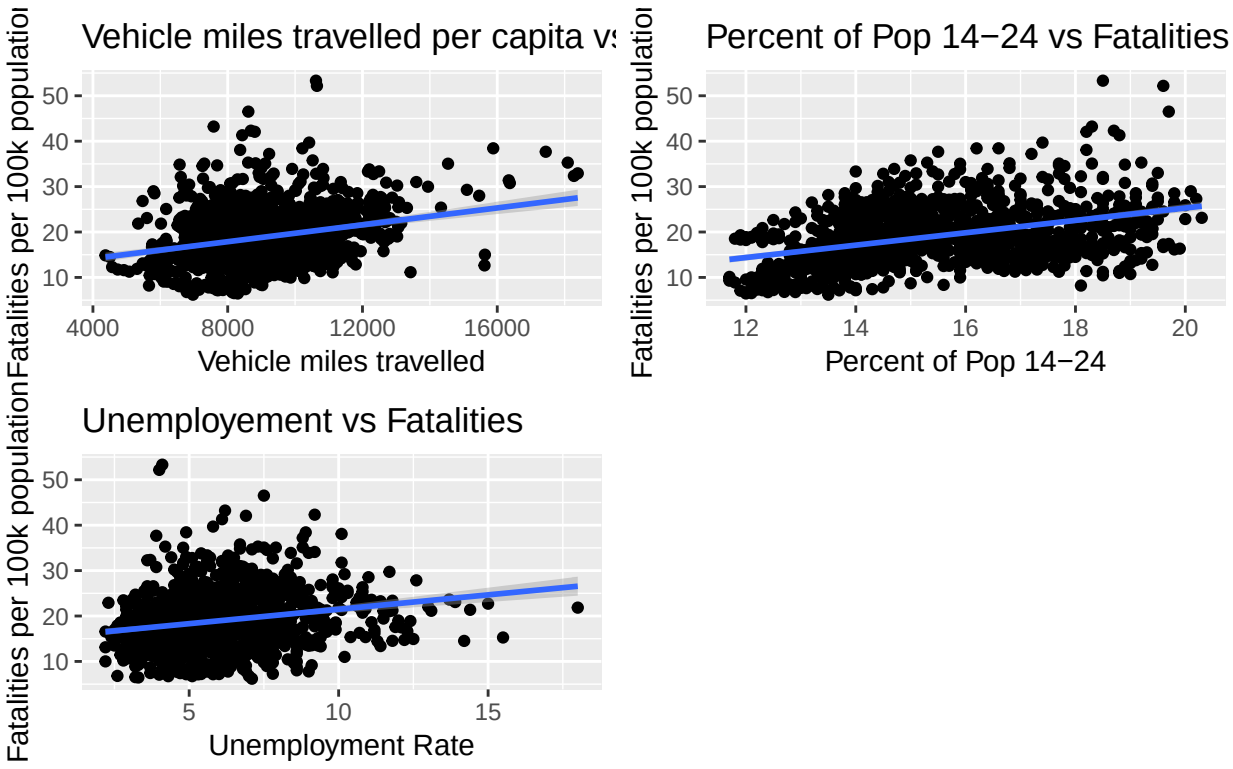
```
vmpc.vs.tot <- ggplot(data, aes(x=vehicmilespc, y=totfatrte)) +
  geom_point() +
  geom_smooth(method=lm) +
```

```
  ggtitle('Vehicle miles travelled per capita vs Fatalities') +
  xlab('Vehicle miles travelled') + ylab('Fatalities per 100k population')

perc.vs.tot <- 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')

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

grid.arrange(vmpc.vs.tot, perc.vs.tot, unem.vs.tot, ncol=2, nrow=2)
```



All of these variables appear to have positive correlations with `totfatrte`.

Next let's examine some of our discrete variables. The variables we'll focus on are:

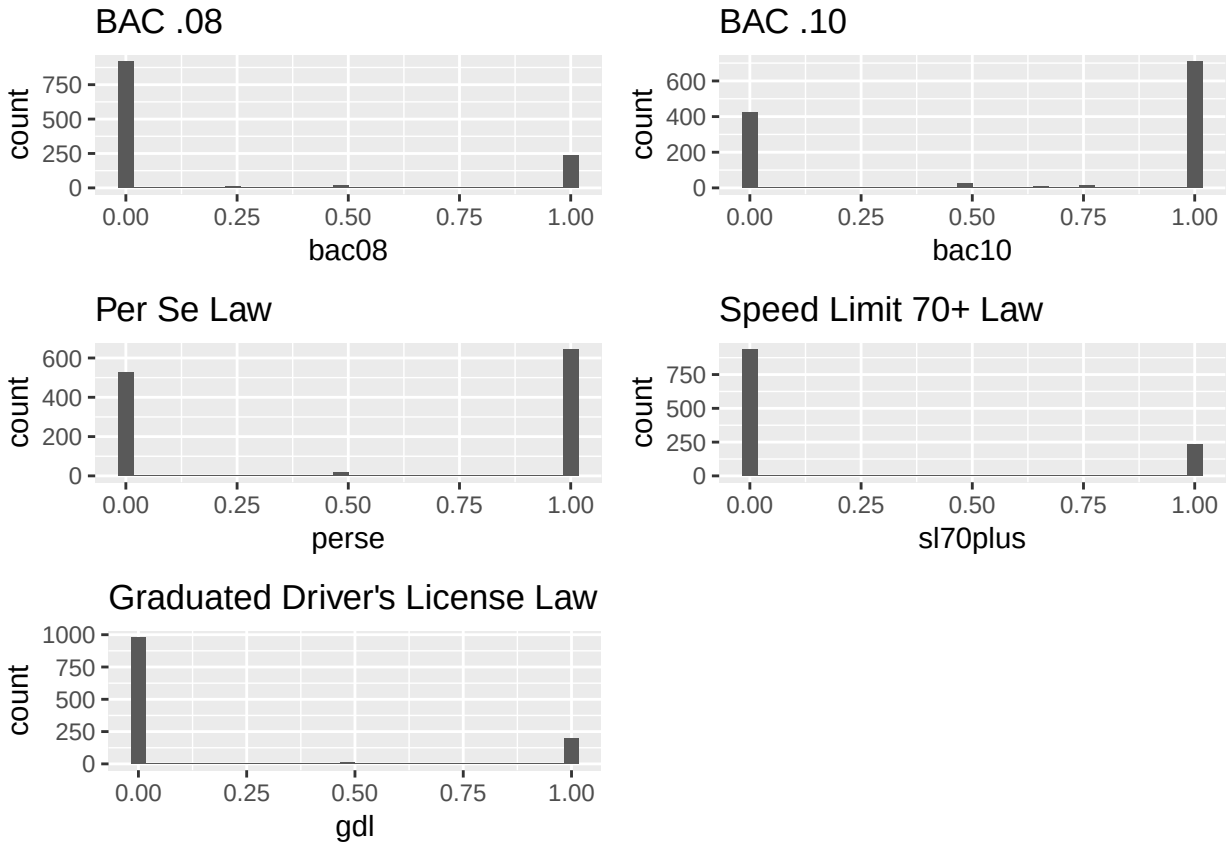
- `bac10` and `bac08`, indicator variables for the legal blood-alcohol (BAC) driving limit
- `perse`, an indicator variable for the implementation of Per Se Laws.
- `sbprim`, an indicator variable for Primary Seat belt laws (which allows law enforcement officers to ticket drivers for solely not wearing a seatbelt)
- `sbsecon`, an indicator variable for Secondary Seat belt laws (which means law enforcement officers cannot ticket drivers for solely not wearing a seatbelt, but can increase the ticket if stopped for another offense.)
- `sl70plus`, an indicator variable for whether the state allows for 70+ speed limits on its highways.
- `gdl`, an indicator variable for Graduated Driver's Licensing laws. These laws mean that drivers must first drive in a supervised learning period, then progress to an intermediate license before being granted

a full driver's license.

Let's first plot the distributions of these categorical variables:

```
bac08.hist <- ggplot(data, aes(x=bac08)) +
  geom_histogram() +
  ggtitle('BAC .08')
bac10.hist <- ggplot(data, aes(x=bac10)) +
  geom_histogram() +
  ggtitle('BAC .10')
perse.hist <- ggplot(data, aes(x=perse)) +
  geom_histogram() +
  ggtitle('Per Se Law')
sl.hist <- ggplot(data, aes(x=sl70plus)) +
  geom_histogram() +
  ggtitle('Speed Limit 70+ Law')
gdl.hist <- ggplot(data, aes(x=gdl)) +
  geom_histogram() +
  ggtitle('Graduated Driver\'s License Law')
grid.arrange(bac08.hist, bac10.hist, perse.hist, sl.hist, gdl.hist, ncol=2, nrow=3)
```

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



In each of the categorical variables we see values between 0 and 1. This is due to states implementing laws

intra year. We'll simplify these variables by rounding them to the nearest whole number.

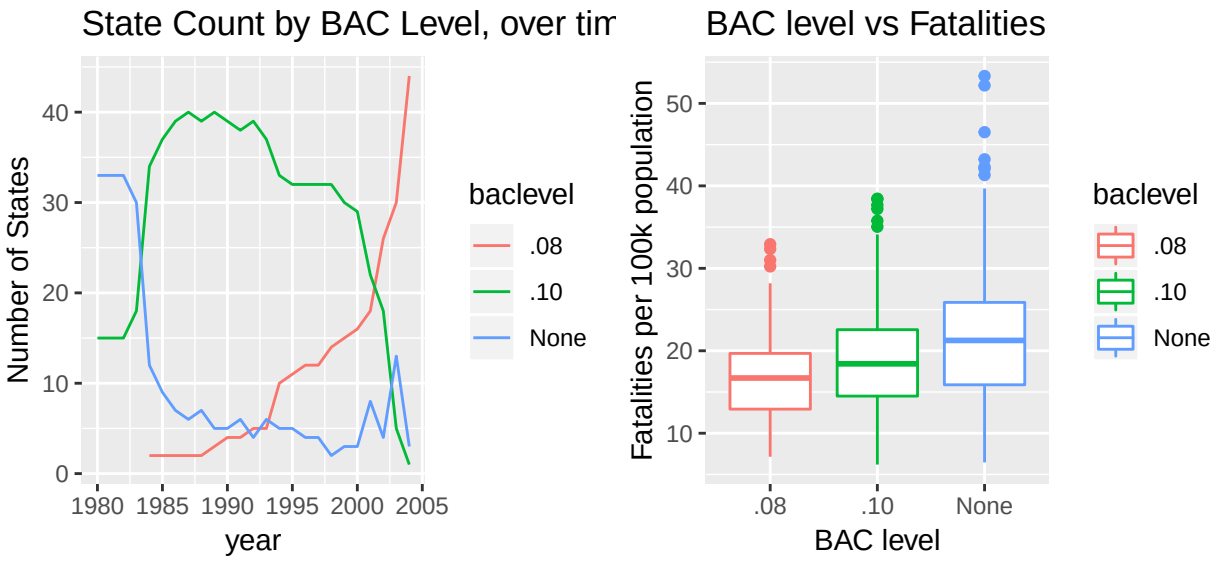
We'll visualize each class of categorical variables in two ways:

- With a time series plot of the number of states implementing them over time
- With a box & whisker plot, with `totfatrte` on the y-axis. This will give a clearer depiction of their impact on `totfatrte`.

Let's start with the BAC law frequencies over time and the `totfatrte` box plot by BAC laws:

```
data$baclevel <- 'None'
data[(data$bac08 == 1), 'baclevel'] = '.08'
data[(data$bac10 == 1), 'baclevel'] = '.10'
data$baclevel <- as.factor(data$baclevel)

agged.fat.bybacyr <- data %>%
  group_by(baclevel, year) %>%
  dplyr::summarize(totfatrte = mean(totfatrte),  # calculate average total fatality rates
                   count = n())  # number of states in each bac
bac.cnt.t <- ggplot(agged.fat.bybacyr, aes(x=year, y=count, colour=baclevel)) +
  geom_line() + ggtitle('State Count by BAC Level, over time') +
  ylab('Number of States')
bac.boxplot <- ggplot(data, aes(x=baclevel, y=totfatrte, color=baclevel)) +
  geom_boxplot() +
  ggtitle('BAC level vs Fatalities') +
  xlab('BAC level') + ylab('Fatalities per 100k population')
grid.arrange(bac.cnt.t, bac.boxplot, ncol=2, nrow=1)
```

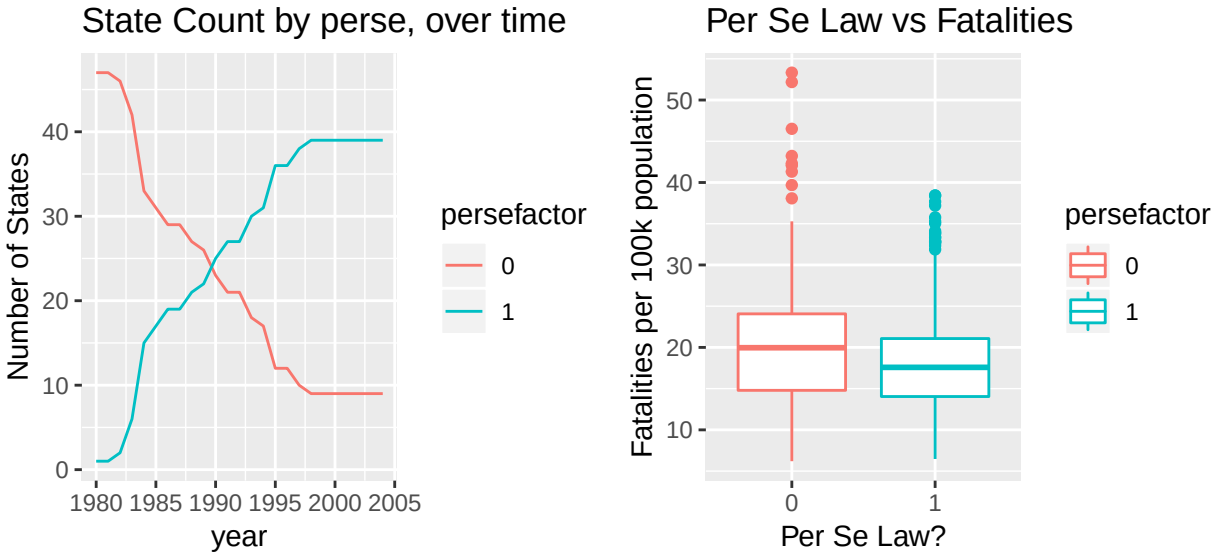


We see several things in these plots:

- There's a sharp adoption of BAC laws in the early 1980s, with many states adopting BAC .10 laws.
- The number of states with .08 BAC limit laws has risen steadily starting in the late 80s, with the vast majority of states adopting this limit by the end of the sample.
- States with more restrictive BAC laws (i.e. with BAC levels of .08) tend to have lower fatality rates than states with more lax BAC laws.

Let's now examine the same plots for `perse`:

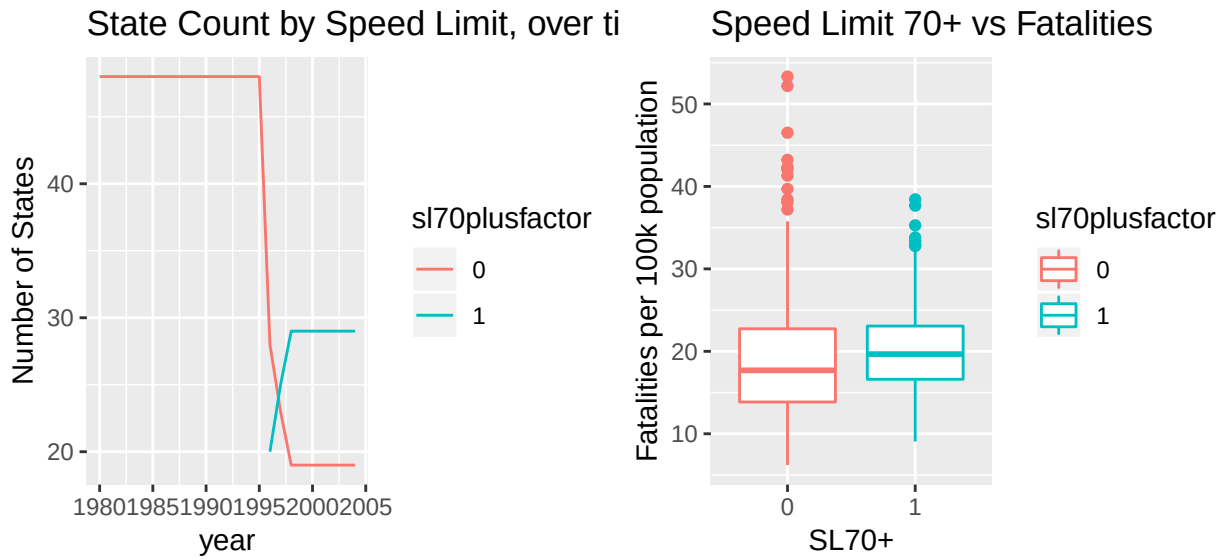
```
data$persefactor <- as.factor(round(data$perse))
agged.fat.byperseyr <- data %>%
  group_by(persefactor, year) %>%
  dplyr::summarize(totfatrte = mean(totfatrte),  # calculate average total fatality rates
                   count = n())
perse.cnt.t <- ggplot(agged.fat.byperseyr, aes(x=year, y=count, color=persefactor)) +
  geom_line() + ggtitle('State Count by perse, over time') +
  ylab('Number of States')
perse.boxplot <- ggplot(data, aes(x=persefactor, y=totfatrte, color=persefactor)) +
  geom_boxplot() +
  ggtitle('Per Se Law vs Fatalities') +
  xlab('Per Se Law?') + ylab('Fatalities per 100k population')
grid.arrange(perse.cnt.t, perse.boxplot, ncol=2, nrow=1)
```



Per se laws saw steady adoption among states until stabilizing in the late 90s, and we see that these states tended to average lower fatality rates.

Let's look at how speed limits change over time and how they affect fatality rates:

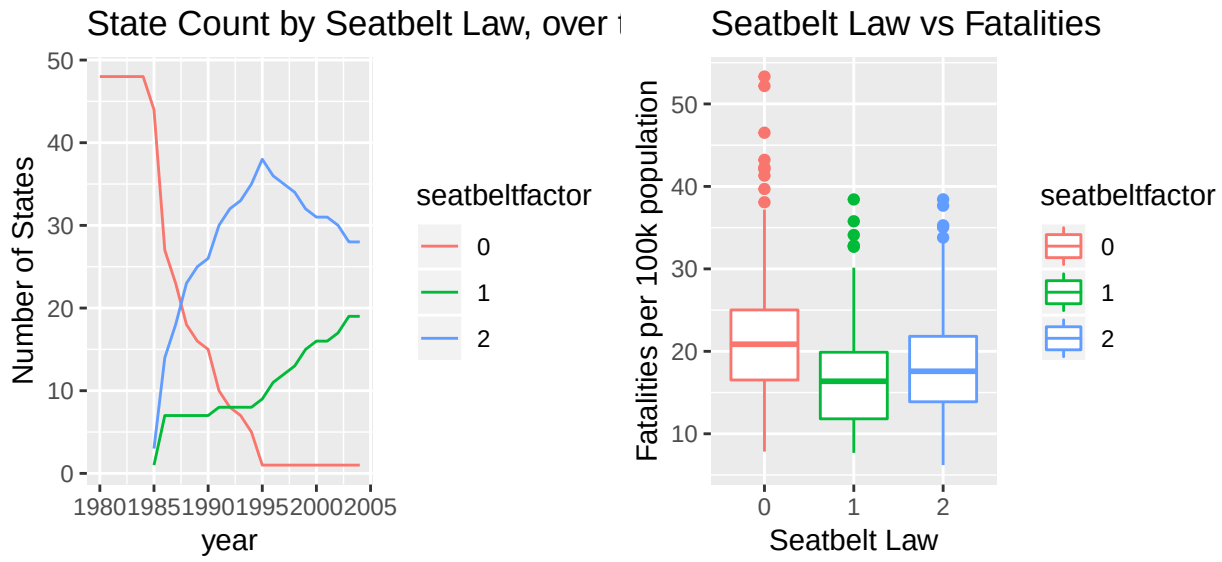
```
data$sl70plusfactor <- as.factor(round(data$sl70plus))
agged.fat.bysl <- data %>%
  group_by(sl70plusfactor, year) %>%
  dplyr::summarize(totfatrte = mean(totfatrte),
                   count = n())
sl.cnt.t <- ggplot(agged.fat.bysl, aes(x=year, y=count, color=sl70plusfactor)) +
  geom_line() + ggtitle('State Count by Speed Limit, over time') +
  ylab('Number of States')
sl.boxplot <- ggplot(data, aes(x=sl70plusfactor, y=totfatrte, color=sl70plusfactor)) +
  geom_boxplot() +
  ggtitle('Speed Limit 70+ vs Fatalities') +
  xlab('SL70+') + ylab('Fatalities per 100k population')
grid.arrange(sl.cnt.t, sl.boxplot, ncol=2, nrow=1)
```



It appears there was significant adoption of 70+ speed limit laws in the 90s, with the number of 70+ speed limit states stabilizing before 2000. Additionally, it appears that states with higher speed limits have higher fatality rates than states with lower speed limits. This makes sense, as higher speeds likely make drivers more at risk of injury in collisions.

What about seatbelt laws?:

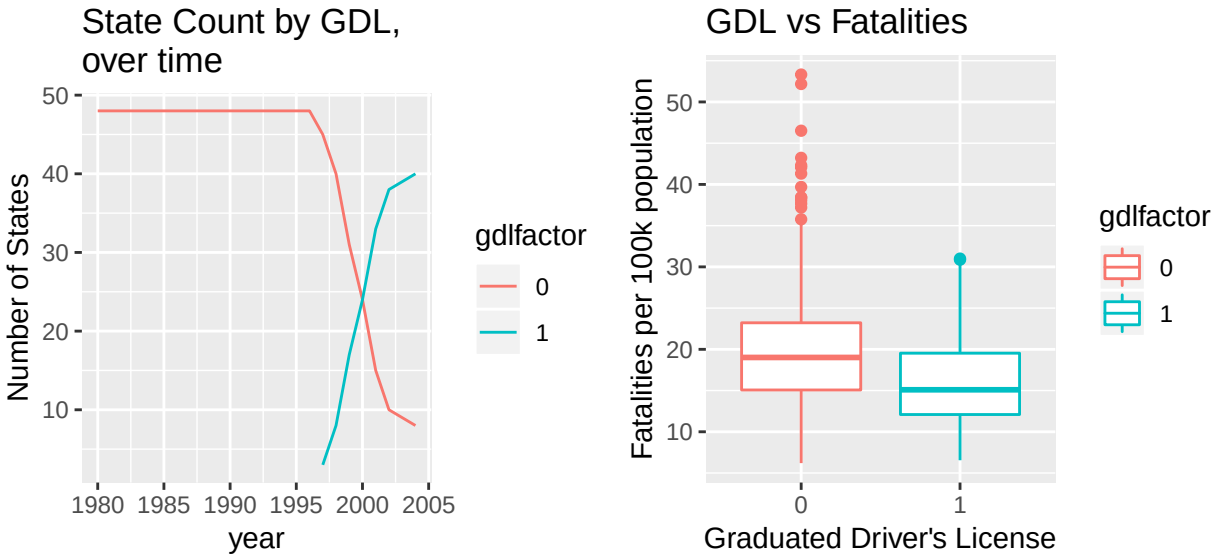
```
data$seatbeltfactor <- as.factor(data$seatbelt)
agged.fat.bysb <- data %>%
  group_by(seatbeltfactor, year) %>%
  dplyr::summarize(totfatrte = mean(totfatrte),
                   count = n())
sb.cnt.t <- ggplot(agged.fat.bysb, aes(x=year, y=count, color=seatbeltfactor)) +
  geom_line() + ggtitle('State Count by Seatbelt Law, over time') +
  ylab('Number of States')
sb.boxplot <- ggplot(data, aes(x=seatbeltfactor, y=totfatrte, color=seatbeltfactor)) +
  geom_boxplot() +
  ggtitle('Seatbelt Law vs Fatalities') +
  xlab('Seatbelt Law') + ylab('Fatalities per 100k population')
grid.arrange(sb.cnt.t, sb.boxplot, ncol=2, nrow=1)
```



We see that more states have adopted primary or secondary seatbelt laws over time, with the vast majority states having some seatbelt law by 1995. We also see that states with Primary seatbelt laws tend to average lower fatality rates compared to states with secondary seatbelt laws and states with no seatbelt laws.

Let's wrap-up by looking at Graduated Driver's License laws:

```
data$gdlfactor <- as.factor(round(data$gdl))
agged.fat.bygdl <- data %>%
  group_by(gdlfactor, year) %>%
  dplyr::summarize(totfatrte = mean(totfatrte),
                   count = n())
gdl.cnt.t <- ggplot(agged.fat.bygdl, aes(x=year, y=count, color=gdlfactor)) +
  geom_line() + ggtitle('State Count by GDL,\nover time') +
  ylab('Number of States')
gdl.boxplot <- ggplot(data, aes(x=gdlfactor, y=totfatrte, color=gdlfactor)) +
  geom_boxplot() +
  ggtitle('GDL vs Fatalities') +
  xlab('Graduated Driver\'s License') + ylab('Fatalities per 100k population')
grid.arrange(gdl.cnt.t, gdl.boxplot, ncol=2, nrow=1)
```



States began to enact Graduated Driver's License laws starting in the late 90s, and states that enact these laws tend to have lower fatality rates.

EDA Key Takeaways:

- Our continuous variables `vehicmilespc`, `unem`, and `perc14_24` are all positively correlated with `totfatrte`. These correlations also held when first-differencing and de-meaning the variables.
- States with more restrictive BAC laws tended to have lower fatality rates, with more states enacting these laws over time.
- States with Per Se laws tended to have lower fatality rates. These laws have become more common with time.
- States with higher speed limits tended to have higher fatality rates. Higher speed limits have become more common with time.
- States with primary and secondary seatbelt laws tended to have lower fatality rates, with primary seatbelt laws having the lowest fatality rates.
- States with Graduated Driver's Licensing laws tended to have lower fatality rates, and more states have adopted these laws over time.

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 "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:

$$totfatrte_{it} = \beta_0 + \sum_{k=1980}^{2004} \delta_k year_k + u_{it}$$

```
#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 statistically significant for all years except 1981. This aligns with what we saw in the box plot of `totfatrte` over time in the EDA.

3. Expand your model in *Exercise 2* by adding variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*, 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.)

We chose not to perform any transformations, as the EDA did not show nonlinear relationships between `totfatrte` and the continuous variables. We do see that these variables are skewed right which log transforming could "fix". However this will change the interpretations of the coefficients, and this will only cause problems if it skews the residuals. We'll examine the residuals of the model after estimating it to ensure they're not heavily skewed.

We'll estimate the pooled OLS model with the additional variables:

$$totfatrte_{it} = \beta_0 + \sum_{k=1980}^{2004} \delta_k year_k + \beta_1 bac08_{it} + \beta_2 bac10_{it} + \beta_3 perse_{it} + \beta_4 sbprim_{it} + \beta_5 sbsecon_{it} + \beta_6 sl70plus_{it} + \beta_7 gdl_{it} + \beta_8 perc14$$

```
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)
```

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

summary(mod2)
```

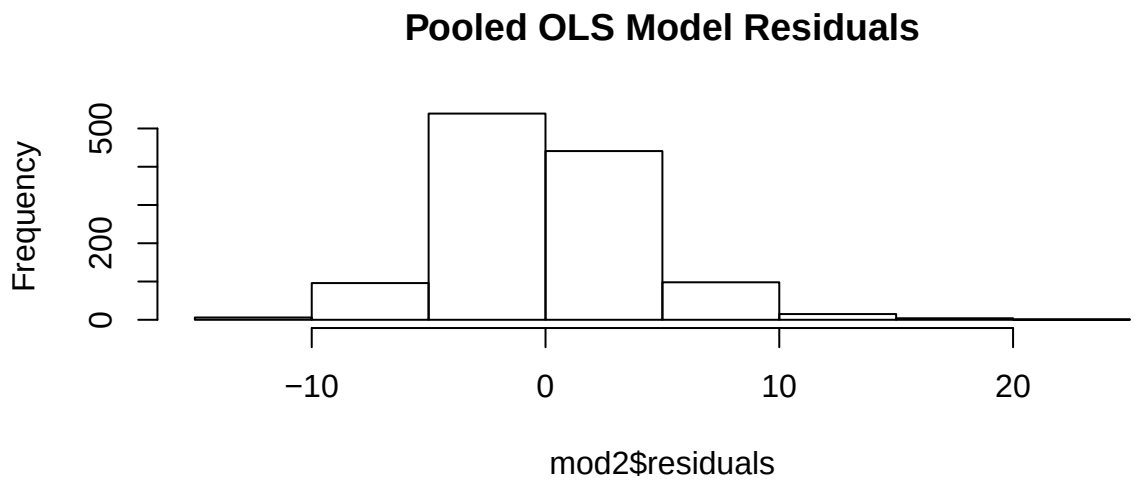
```
##
## Call:
## lm(formula = totfatrte ~ factor(year) + bac08round + bac10round +
##     perseround + sbprimround + sbseconround + sl70plusround +
##     gdlround + perc14_24 + unem + vehicmilespc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8962  -2.7265  -0.3033   2.3323  21.5064
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.826e+00  2.478e+00  -1.141 0.254236
## factor(year)1981 -2.184e+00  8.290e-01  -2.634 0.008539 **
## factor(year)1982 -6.657e+00  8.547e-01  -7.789 1.49e-14 ***
## factor(year)1983 -7.589e+00  8.671e-01  -8.752  < 2e-16 ***
## factor(year)1984 -5.974e+00  8.730e-01  -6.843 1.25e-11 ***
## factor(year)1985 -6.603e+00  8.915e-01  -7.407 2.47e-13 ***
## factor(year)1986 -5.947e+00  9.290e-01  -6.401 2.23e-10 ***
## factor(year)1987 -6.459e+00  9.656e-01  -6.689 3.48e-11 ***
## factor(year)1988 -6.691e+00  1.013e+00  -6.607 5.97e-11 ***
## factor(year)1989 -8.159e+00  1.052e+00  -7.757 1.89e-14 ***
## factor(year)1990 -9.060e+00  1.076e+00  -8.421  < 2e-16 ***
## factor(year)1991 -1.121e+01  1.099e+00 -10.194  < 2e-16 ***
## factor(year)1992 -1.300e+01  1.121e+00 -11.591  < 2e-16 ***
## factor(year)1993 -1.288e+01  1.134e+00 -11.358  < 2e-16 ***
## factor(year)1994 -1.253e+01  1.154e+00 -10.855  < 2e-16 ***
## factor(year)1995 -1.203e+01  1.183e+00 -10.176  < 2e-16 ***
## factor(year)1996 -1.403e+01  1.224e+00 -11.459  < 2e-16 ***
## factor(year)1997 -1.430e+01  1.242e+00 -11.517  < 2e-16 ***
## factor(year)1998 -1.512e+01  1.262e+00 -11.978  < 2e-16 ***
## factor(year)1999 -1.518e+01  1.276e+00 -11.900  < 2e-16 ***
## factor(year)2000 -1.554e+01  1.296e+00 -11.996  < 2e-16 ***
## factor(year)2001 -1.645e+01  1.316e+00 -12.500  < 2e-16 ***
## factor(year)2002 -1.703e+01  1.331e+00 -12.798  < 2e-16 ***
## factor(year)2003 -1.742e+01  1.336e+00 -13.033  < 2e-16 ***
## factor(year)2004 -1.698e+01  1.369e+00 -12.399  < 2e-16 ***
## bac08round       -2.194e+00  4.891e-01  -4.487 7.94e-06 ***
## bac10round       -1.238e+00  3.616e-01  -3.423 0.000641 ***
## perseround       -6.499e-01  2.943e-01  -2.208 0.027433 *
## sbprimround      -9.420e-02  4.910e-01  -0.192 0.847868
## sbseconround      6.430e-02  4.299e-01   0.150 0.881124
## sl70plusround     3.239e+00  4.352e-01   7.443 1.91e-13 ***
## gdlround         -3.476e-01  5.101e-01  -0.682 0.495682
## perc14_24         1.401e-01  1.229e-01   1.140 0.254611
## unem              7.675e-01  7.796e-02   9.844  < 2e-16 ***
## vehicmilespc      2.927e-03  9.485e-05  30.860  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.052 on 1165 degrees of freedom
## Multiple R-squared:  0.6064, Adjusted R-squared:  0.595
## F-statistic:  52.8 on 34 and 1165 DF,  p-value: < 2.2e-16
```

This model indicates that in addition to years, laws with blood alcohol limis of either .08 or .1 have a significant impact decreasing fatalities in car accidents. Further, it indicates that laws that allow speed limits over 70, high unemployment, and high vehicle miles per capita have significant impact increasing fatalities from car accidents.

Let's plot the model residuals:

```
hist(mod2$residuals, main='Pooled OLS Model Residuals')
```



The residuals of the model do not appear heavily skewed, so the skewness we saw in the continuous variables in the EDA will not negatively impact our inferences.

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?

We'll estimate this fixed effects model:

Let

$$\widetilde{x}_{it} = x_{it} - \overline{x}_i$$

$$\widetilde{totfatrte}_{it} = \beta_0 + \sum_{k=1980}^{2004} \delta_k year_k + \beta_1 \widetilde{bac08}_{it} + \beta_2 \widetilde{bac10}_{it} + \beta_3 \widetilde{perse}_{it} + \beta_4 \widetilde{sbprim}_{it} + \beta_5 \widetilde{sbsecon}_{it} + \beta_6 \widetilde{sl70plus}_{it} + \beta_7 \widetilde{gdl}_{it} + \beta_8 \widetilde{perc14}$$

```r
model.fe <- plm(totfatrte ~ factor(year) + bac08round + bac10round +
                perseround + sbprimround + sbseconround + sl70plusround +
                gdlround + perc14_24 + unem + vehicmilespc,
                data=data,
                index=c('state', 'year'), model='within')

summary(model.fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrte ~ factor(year) + bac08round + bac10round +
##     perseround + sbprimround + sbseconround + sl70plusround +
##     gdlround + perc14_24 + unem + vehicmilespc, data = data,
##     model = "within", index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -8.2942745 -1.0561099  0.0055578  0.9788361 14.8497791
##
## Coefficients:
##                      Estimate  Std. Error  t-value  Pr(>|t|)
## factor(year)1981 -1.5124e+00  4.1379e-01  -3.6549 0.0002692 ***
## factor(year)1982 -3.0540e+00  4.4318e-01  -6.8912 9.221e-12 ***
## factor(year)1983 -3.6638e+00  4.5516e-01  -8.0495 2.111e-15 ***
## factor(year)1984 -4.3998e+00  4.5966e-01  -9.5719 < 2.2e-16 ***
## factor(year)1985 -4.8603e+00  4.8010e-01 -10.1237 < 2.2e-16 ***
## factor(year)1986 -3.7692e+00  5.1357e-01  -7.3392 4.123e-13 ***
## factor(year)1987 -4.4123e+00  5.5162e-01  -7.9989 3.118e-15 ***
## factor(year)1988 -4.8877e+00  5.9837e-01  -8.1684 8.379e-16 ***
## factor(year)1989 -6.2395e+00  6.3732e-01  -9.7901 < 2.2e-16 ***
## factor(year)1990 -6.3564e+00  6.6196e-01  -9.6024 < 2.2e-16 ***
## factor(year)1991 -7.0442e+00  6.7895e-01 -10.3752 < 2.2e-16 ***
## factor(year)1992 -7.8905e+00  7.0039e-01 -11.2659 < 2.2e-16 ***
## factor(year)1993 -8.2366e+00  7.1290e-01 -11.5536 < 2.2e-16 ***
## factor(year)1994 -8.6823e+00  7.3004e-01 -11.8930 < 2.2e-16 ***
## factor(year)1995 -8.3889e+00  7.5324e-01 -11.1370 < 2.2e-16 ***
## factor(year)1996 -8.7648e+00  7.9400e-01 -11.0388 < 2.2e-16 ***
## factor(year)1997 -8.9164e+00  8.1140e-01 -10.9889 < 2.2e-16 ***
## factor(year)1998 -9.5333e+00  8.2867e-01 -11.5044 < 2.2e-16 ***
## factor(year)1999 -9.6940e+00  8.3614e-01 -11.5938 < 2.2e-16 ***
## factor(year)2000 -1.0223e+01  8.4713e-01 -12.0683 < 2.2e-16 ***
## factor(year)2001 -9.9608e+00  8.5745e-01 -11.6168 < 2.2e-16 ***
## factor(year)2002 -9.2546e+00  8.6613e-01 -10.6850 < 2.2e-16 ***
## factor(year)2003 -9.3270e+00  8.6980e-01 -10.7232 < 2.2e-16 ***
## factor(year)2004 -9.6676e+00  8.9310e-01 -10.8248 < 2.2e-16 ***
## bac08round       -1.1805e+00  3.2987e-01  -3.5786 0.0003603 ***
## bac10round       -8.6977e-01  2.2522e-01  -3.8619 0.0001190 ***
## perseround       -1.0587e+00  2.2415e-01  -4.7230 2.619e-06 ***
## sbprimround      -1.2506e+00  3.4313e-01  -3.6447 0.0002800 ***
## sbseconround     -3.5659e-01  2.5230e-01  -1.4133 0.1578360
## sl70plusround    -3.2440e-02  2.6034e-01  -0.1246 0.9008577
```

```
## gdlround          -3.0503e-01  2.8029e-01  -1.0883 0.2767100
## perc14_24          1.9367e-01  9.5068e-02   2.0372 0.0418646 *
## unem              -5.7652e-01  6.0592e-02  -9.5147 < 2.2e-16 ***
## vehicmilespc       9.2612e-04  1.1066e-04   8.3691 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4547.9
## R-Squared:      0.6252
## Adj. R-Squared: 0.59804
## F-statistic: 54.8501 on 34 and 1118 DF, p-value: < 2.22e-16
```

- All four estimates are directionally negative in both models and all four estimates are statistically significant at p<0.001 level in the fixed effect model. *bac08* and *bac10* coefficients from the pooled OLS has higher absolute estimates when compared to those from the fixed effect model, where they are statistically significant at p<0.001 from both models. On the other hand, *perse* and *sbprim* coefficients from the fixed effect model has higher absolute estimates when compared to those from the pooled OLS and are significant only in the fixed effect model.

- The fixed effect estimates are likely to be more reliable because the standard errors are uniformly lower when compared to those from the pooled OLS, indicating higher precision in the fixed effect model. In addition, the pooled OLS assumes there is independence between the observations and does not account for unobserved heterogeneity, which makes the fixed effect model more consistent compared to the pooled OLS.
- The fixed effect model assumes that the state fixed effects are time independent and the explanatory variables change over time with no perfect linear relationship between the variables. In comparison, the pooled OLS assumes that the response variable is normally distributed and errors are uncorrelated with the explanatory variables, which are valid assumptions based on the residual diagnostics.

5. Would you perfer 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 conducts 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 + 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 = 164.12, 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 rending 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.