Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Group Lab 3

U.S. traffic fatalities: 1980-2004

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.

Exercises:

1. (40%) 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.*

```
# Import libraries
library(foreign)
library(gplots)
library(ggplot2)
library(stats)
library(Hmisc)
library(car)
library(dplyr)
library(corrplot)
library(corrgram)
library(lattice)
library(plm)

# clearing workspace and loading data
rm(list = ls())
#setwd('/Users/jamesdarmody/Documents/w271/labs/lab3')
driving <- get(load('driving.RData'))
```

```
# examining what loaded
ls()
```

```
## [1] "data"    "desc"    "driving" "self"
```

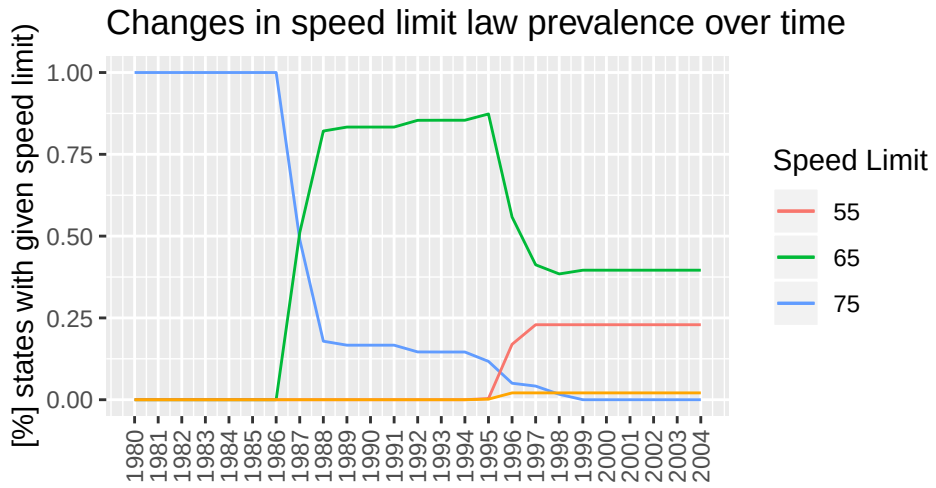
```
# looking at description of fields in the data
desc
```

```
#
str(data)
```

```r
# we first check the dataframe for missing values
apply(data, 2, function(x) any(is.na(x)))
# Dimension of the data
# summary(factor(data$year))
# summary(factor(data$state))
```

- We can see that there are 2004-1980+1 = 25 years and 48 states (missing state ID 2 and 9) in total, therefore, the datafram has 25*48 = 1200 observations.
- There do not appear to be missing values in this data, so we transition to examining features.
- We begin by examining the prevalence of various laws among states over time, starting with speed limit laws.

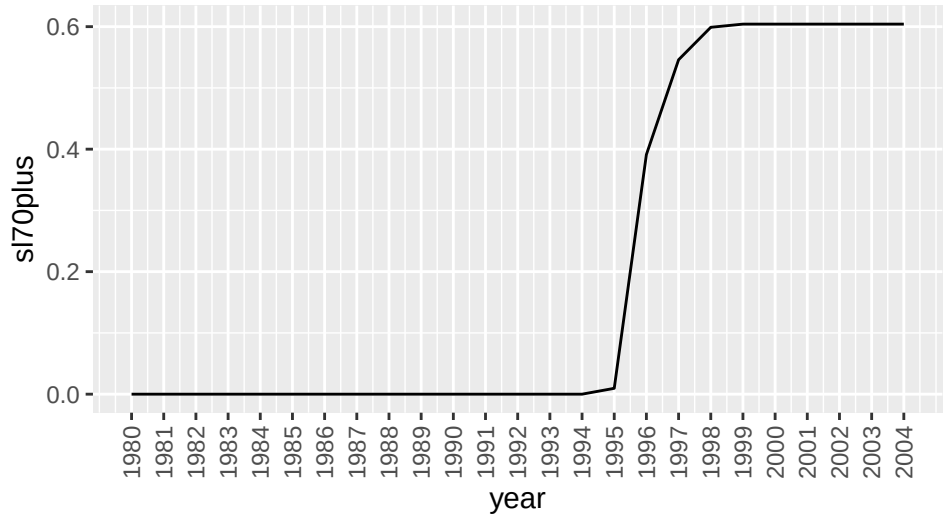
```r
# summarize the average statistics for speed limit in a data frame
sl.df <- data %>% group_by(year) %>%
  summarise(sl55 = mean(sl55), sl65 = mean(sl65),sl75 = mean(sl75), slnone = mean(slnone))
# plot the summary stats in a growth chart
sl.plot <- ggplot(sl.df, aes(x = year)) +
  geom_line(aes(y = sl55, color='purple')) + geom_line(aes(y=sl65, color='green')) +
  geom_line(aes(y=sl75, color='blue')) + geom_line(aes(y=slnone), color='orange') +
  scale_x_continuous(breaks = seq(min(sl.df$year), max(sl.df$year), 1)) +
  theme(axis.text.x = element_text(angle = 90, vjust=0.5)) +
  scale_color_discrete(name="Speed Limit", labels=c("55", "65", "75", "none")) +
  labs(title = "Changes in speed limit law prevalence over time",
       y = "[%] states with given speed limit)",
       x = "")
sl.plot
```



As can be seen from the chart above, speed limit laws have actually become more lax over time. At the start of the dataset, all states had a speed limit of 55mph, but in the late 1980s, 65mph became the norm, while in the late 1990s, states were split between 65mps, 75mps, and no speed limit; a situation which persisted fairly constantly until the end of the dataset in 2004.

Below, we look specifically at states that have speed limits over 70 or none in particular.

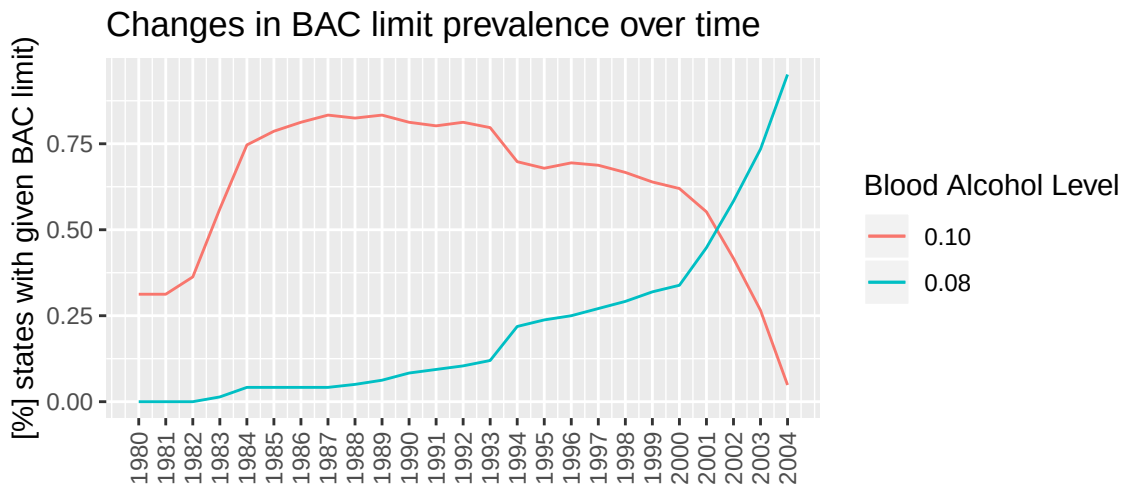
```r
# summarize the average statistics for high speed limit in a data frame
sl.high.df <- data %>% group_by(year) %>%
  summarise(sl70plus = mean(sl70plus))
# plot the summary stats in a growth chart
sl.high.plot <- ggplot(sl.high.df, aes(x = year)) +
  geom_line(aes(y = sl70plus)) +
  scale_x_continuous(breaks = seq(min(sl.high.df$year), max(sl.high.df$year), 1)) +
  theme(axis.text.x = element_text(angle = 90, vjust=0.5))
sl.high.plot
```



Around the mid-90s, these states jump from being non-existent, to being nearly 60% of the dataset, which would reasonably lead us to expect higher rates of fatalities.

We then extend this analysis next to blood alcohol laws below.

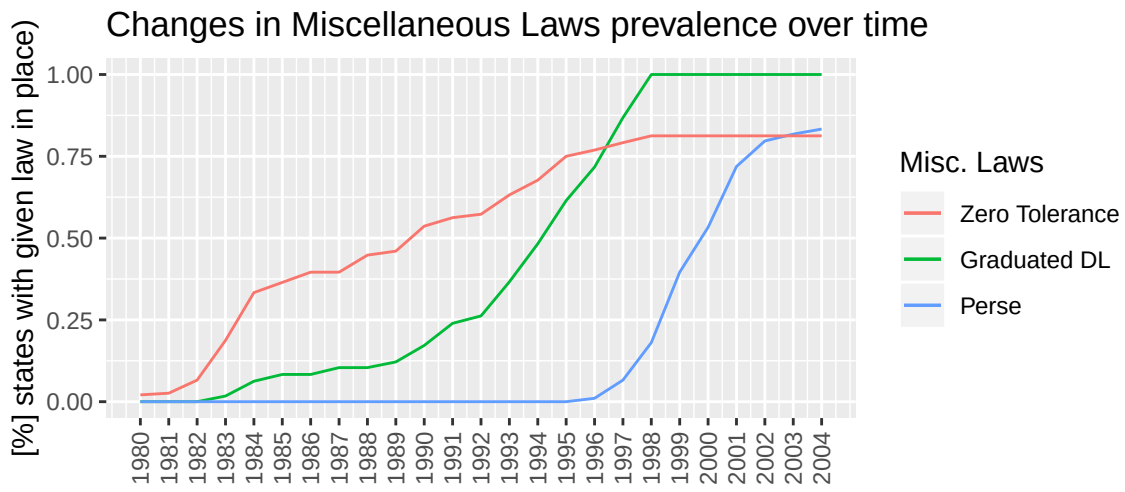
```r
# summarize the average statistics for speed limit in a data frame
bac.df <- data %>% group_by(year) %>%
  summarise(bac10 = mean(bac10), bac08 = mean(bac08))
# plot the summary stats in a growth chart
bac.plot <- ggplot(bac.df, aes(x = year)) +
  geom_line(aes(y = bac10, color='blue')) + geom_line(aes(y=bac08, color='orange')) +
  scale_x_continuous(breaks = seq(min(sl.df$year), max(sl.df$year), 1)) +
  theme(axis.text.x = element_text(angle = 90, vjust=0.5)) +
  scale_color_discrete(name="Blood Alcohol Level", labels=c("0.10", "0.08")) +
  labs(title = "Changes in BAC limit prevalence over time",
       y = "[%] states with given BAC limit)",
       x = "")
bac.plot
```



Examining the data, it seems that while speed limit laws may have gotten more lax over time, Blood Alcohol Limitations became stricter over time. At the beginning of the dataset, only about 30% of states had BAC limit laws, and all were set at 0.1%, but over time, 0.08% increased exponentially, until it was the near universal standard at the end of the dataset in 2004.

We then extend this analysis to a group of miscellaneous laws

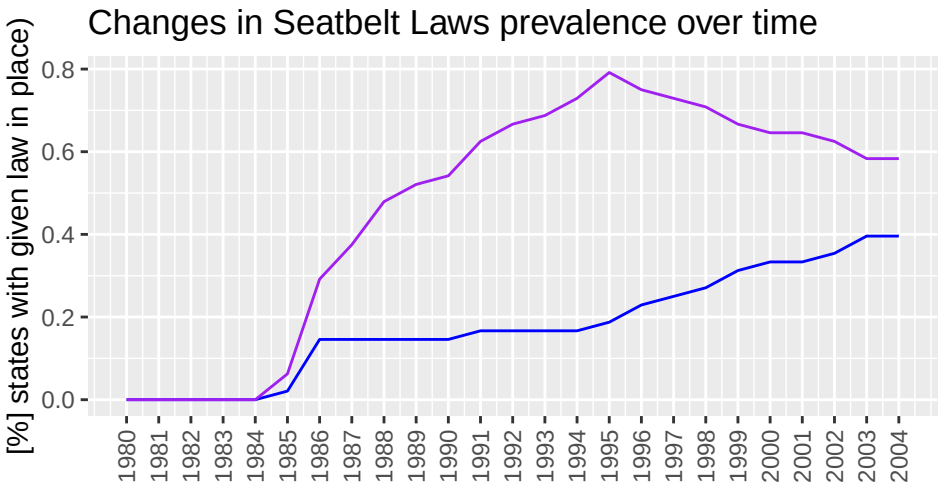
```
# summarize the average statistics for speed limit in a data frame
misc.df <- data %>% group_by(year) %>%
  summarise(zerotol = mean(zerotol), gdl = mean(gdl), perse = mean(perse))
# plot the summary stats in a growth chart
misc.plot <- ggplot(misc.df, aes(x = year)) +
  geom_line(aes(y = zerotol, color='green')) + geom_line(aes(y=gdl, color='purple')) +
  geom_line(aes(y = perse, color='gold')) +
  scale_x_continuous(breaks = seq(min(sl.df$year), max(sl.df$year), 1)) +
  theme(axis.text.x = element_text(angle = 90, vjust=0.5)) +
  scale_color_discrete(name="Misc. Laws", labels=c("Zero Tolerance", "Graduated DL", "Perse")) +
  labs(title = "Changes in Miscellaneous Laws prevalence over time",
       y = "[%] states with given law in place)",
       x = "")
misc.plot
```



The dataset shows a similar phenomenon to what occurred with BAC levels. Starting at the beginning of the dataset, most of these laws were close to non-existent, but zero tolerance laws grew exponentially in the early 1980s, while graduated DL laws grew exponentially in the early 1990s, and perse laws gres exponentially in the late 1990s.

Finally, we examine one last type of law - seatbelt laws

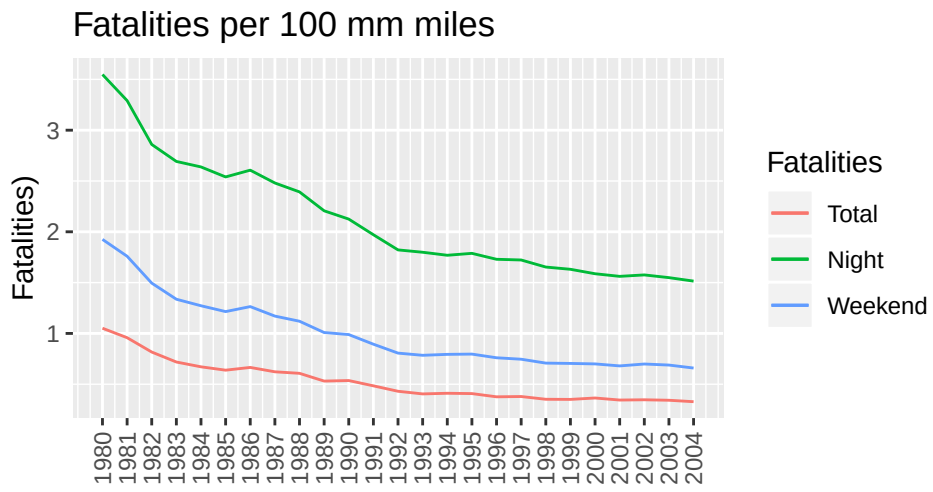
```
# summarize the average statistics for speed limit in a data frame
seatbelt.df <- data %>% group_by(year) %>%
  summarise(seatbelt = mean(seatbelt), sbprim = mean(sbprim), sbsecon = mean(sbsecon))
# plot the summary stats in a growth chart
seatbelt.plot <- ggplot(seatbelt.df, aes(x = year)) +
  geom_line(aes(y = sbprim), color='blue') +
  geom_line(aes(y = sbsecon), color='purple') +
  scale_x_continuous(breaks = seq(min(seatbelt.df$year), max(seatbelt.df$year), 1)) +
  scale_color_discrete(name="Seatbelt. Laws", labels=c("Primary", "Secondary")) +
  theme(axis.text.x = element_text(angle = 90, vjust=0.5)) +
  labs(title = "Changes in Seatbelt Laws prevalence over time",
       y = "[%] states with given law in place)",
       x = "")
seatbelt.plot
```



According to the data, seatbelt laws did not take effect at all until 1984, when they rose precipitously, and approached a relatively even balance between primary and secondary laws by the end of the dataset.

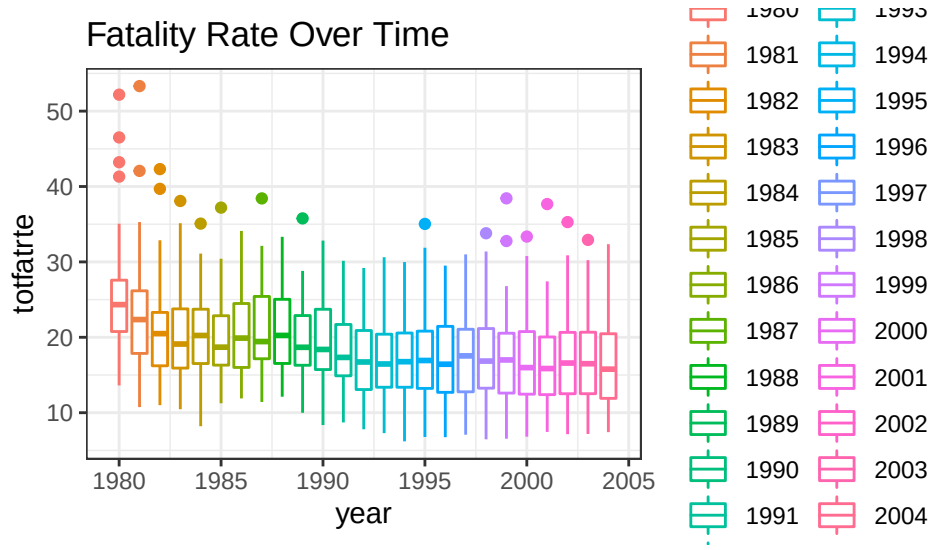
We now transition from examining the legal side of things to looking at some statistics on fatalities across time.

```
# summarize the average statistics for speed limit in a data frame
fat.df <- data %>% group_by(year) %>%
  summarise(totfatpvm = mean(totfatpvm), nghtfatpvm = mean(nghtfatpvm),
            wkndfatpvm = mean(wkndfatpvm))
# plot the summary stats in a growth chart
fat.plot <- ggplot(fat.df, aes(x = year)) +
  geom_line(aes(y = totfatpvm, color='green')) +
  geom_line(aes(y=nghtfatpvm, color='purple')) +
  geom_line(aes(y = wkndfatpvm, color='gold')) +
  scale_x_continuous(breaks = seq(min(sl.df$year), max(sl.df$year), 1)) +
  theme(axis.text.x = element_text(angle = 90, vjust=0.5)) +
  scale_color_discrete(name="Fatalities", labels=c("Total", "Night", "Weekend")) +
  labs(title = "Fatalities per 100 mm miles",
       y = "Fatalities)",
       x = "")
fat.plot
```



The data shows that all manner of fatalities have come down significantly since the dataset began. We expect this, therefore, to hold true for the dependent variable in the dataset, which we examine in the following graph.

```
# examining the the total fatality rate over time
qplot(year, totfatrte, colour = factor(year), geom = "boxplot", data = data) +
  theme_bw()+ ggtitle("Fatality Rate Over Time")
```

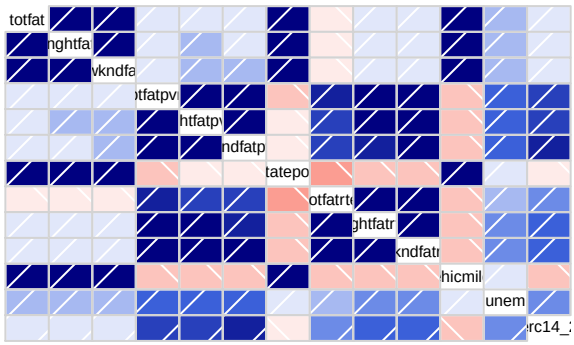


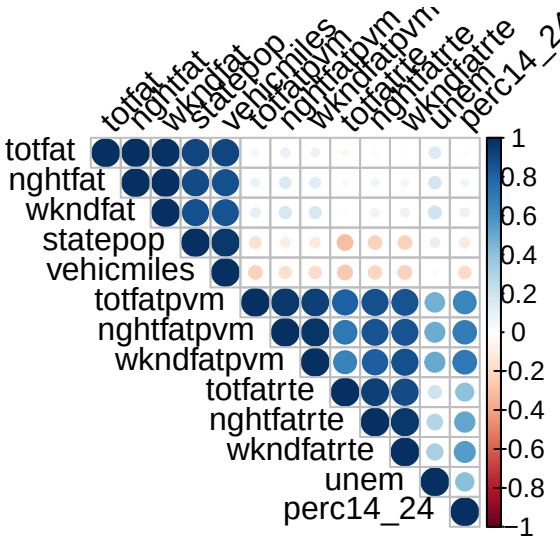
This boxplot, which shows the distribution of the total fatallities of different states by each year, shows a general decline trend in traffic fatalities over time for both average, majority and outliners, which is expected from the first visual on fatalites.

There are a number of economic statistics in the dataset as well. We utilize a correlation matrix to examine whether there exists a relationship between the dependent variable and any of these

```
# Subset the economic stats
numerical.data <- data[,15:27]
# Plot the correlations
numerical.data.rcorr = (as.matrix(numerical.data))
```

```
corrplot(corrgram(numerical.data.rcorr), type="upper", order="hclust",
         tl.col="black", tl.srt=45)
```

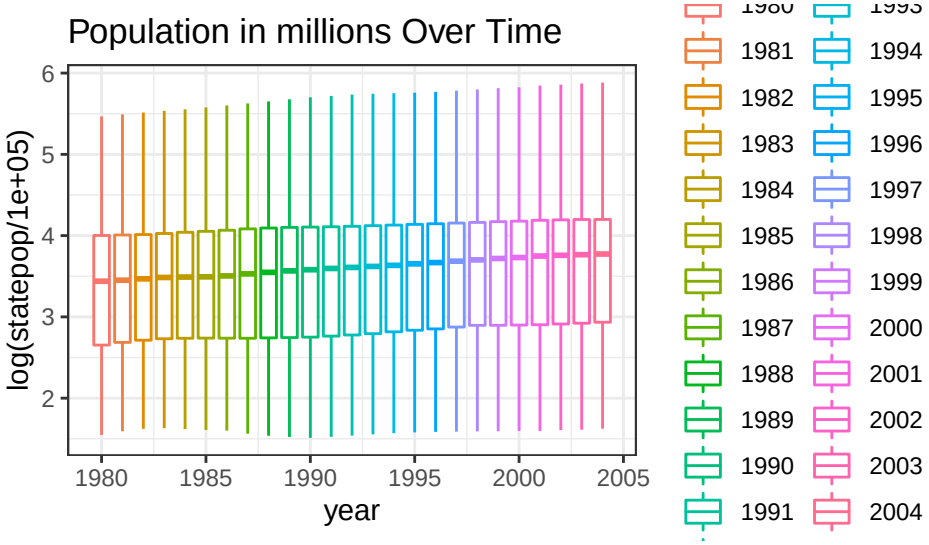




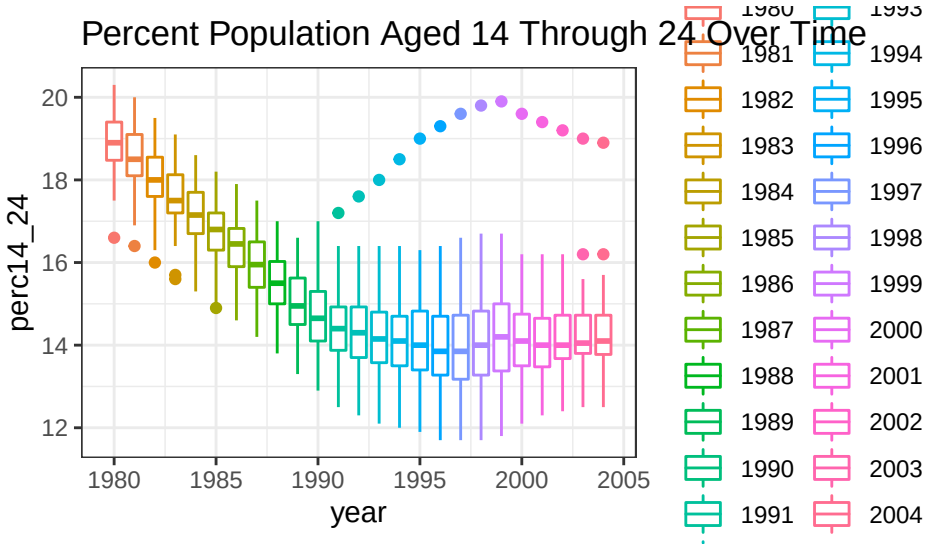
From the correlation matrix, it doesn't look as if there are any major correlations between the dependent variable, and economic statistics like unemployment. There are large correlations between totfatrte and the other fatality rates (i.e. nghtfatrte), but this is to be expected, as they are subsets of the dependent variable.

For completeness' sake, we then examine some of the economic statistics

```
# Log of state population boxplot
qplot( year, log(statepop/100000), colour = factor(year), geom = "boxplot", data = data) +
  theme_bw()+ ggtitle("Population in millions Over Time")
```




```
# Young population percentage
qplot( year, perc14_24, colour = factor(year), geom = "boxplot", data = data) +
  theme_bw()+ ggtitle("Percent Population Aged 14 Through 24 Over Time")
```

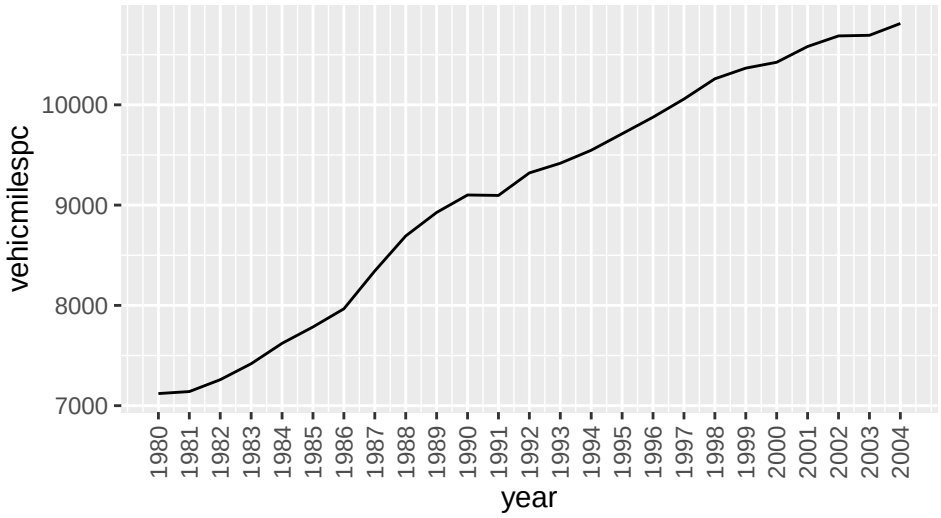


The state population has grown over the course of the data at what appears to be an exponential rate. With more people, you would expect more drivers on the road, and therefore more accident-related fatalities, but as we have seen prior, fatalities are coming down over the course of the data.

The young population(aged from 14 to 24) is rapidly decreased through the year 1980 to 1990 and became statble till year 2005, except for very few outliner states.

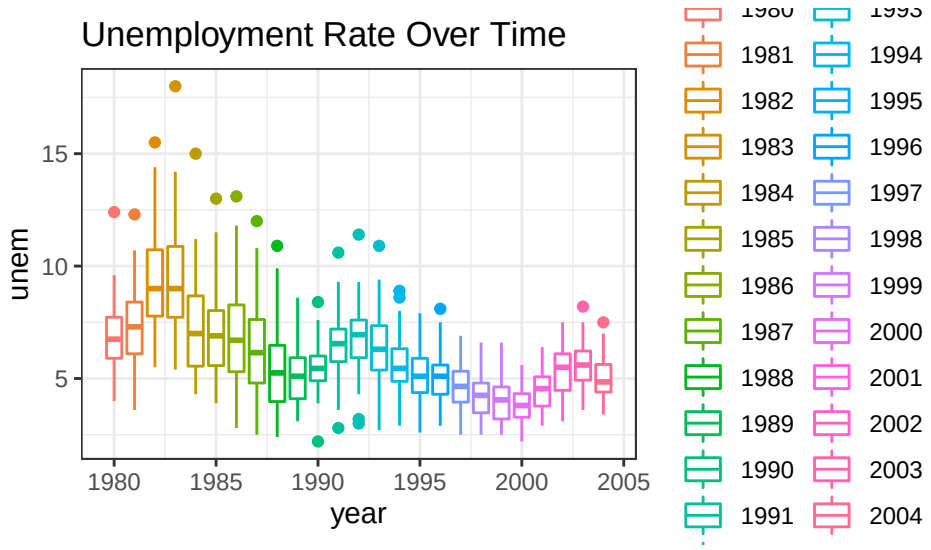
We now turn to the number of miles driven

```
# summarize the average statistics for miles driven in a data frame
miles.df <- data %>% group_by(year) %>%
  summarise(vehicmilespc = mean(vehicmilespc))
# plot the summary stats in a growth chart
miles.plot <- ggplot(miles.df, aes(x = year)) +
  geom_line(aes(y = vehicmilespc)) +
  scale_x_continuous(breaks = seq(min(miles.df$year), max(miles.df$year), 1)) +
  theme(axis.text.x = element_text(angle = 90, vjust=0.5))
miles.plot
```



Similar to the population, the number of miles driven per vehicle has risen rapidly over the course of the data set. Against this backdrop it would be reasonable to expect the number of fatalities to increase, but the opposite is the case. We also include a visual on how unemployment has changed over the past 25 years.

```
# Boxplot of unemployment rate over time
qplot( year, unem, colour = factor(year), geom = "boxplot", data = data) +
  theme_bw()+ ggtitle("Unemployment Rate Over Time")
```



The unemployment rate has been fluctuated through out the time, but in general, the mean and variance of the unemployment rate is getting smaller through out the time.

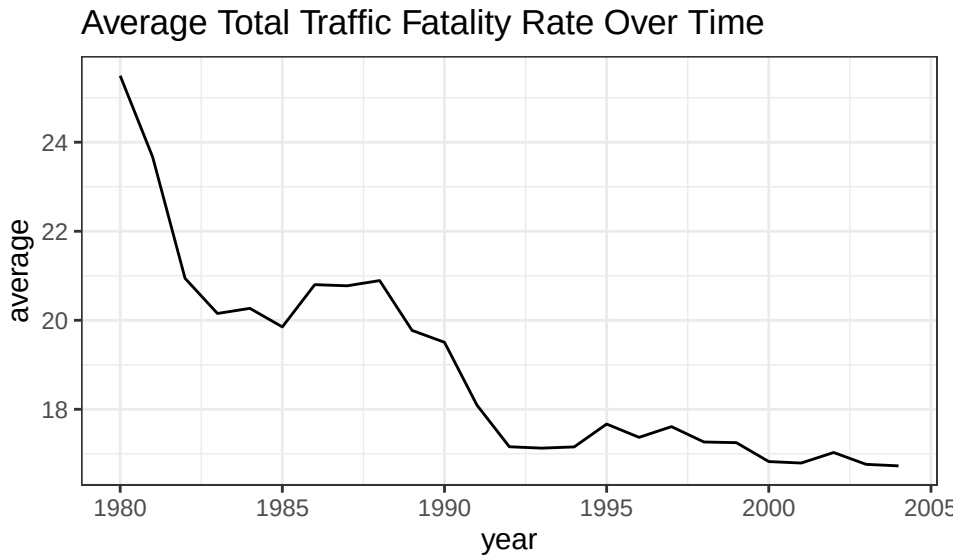
2. (15%) How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this model explain? Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.

Answer Question 2:

The dependent variable of interest, totfatrte, is defined in terms of traffic fatalities per 100,000 of population.

```
# grouping data by year and averaging totfatrte
df.avg.titfatrte <- data %>%
  group_by(year) %>%
  summarise(average = mean(totfatrte))

qplot( year, average, geom = "line", data = df.avg.titfatrte) +
  theme_bw()+ ggtitle("Average Total Traffic Fatality Rate Over Time")
```



Averaging by year appears to show a trend of average fatalities per 100,000 people coming down rapidly from ~25 deaths per 100,000 in 1980 to ~17 deaths per 100,000 in 1992, then stablized around ~17 death per 100,000 people from year 1992 to 2004.

```
# specifying dummy regression by year
dummy.lm <- lm(totfatrte ~ as.factor(year), data=data)
summary(dummy.lm)
```

```
##
## Call:
## lm(formula = totfatrte ~ as.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 ***
## as.factor(year)1981   -1.8244     1.2263  -1.488 0.137094
## as.factor(year)1982   -4.5521     1.2263  -3.712 0.000215 ***
## as.factor(year)1983   -5.3417     1.2263  -4.356 1.44e-05 ***
```

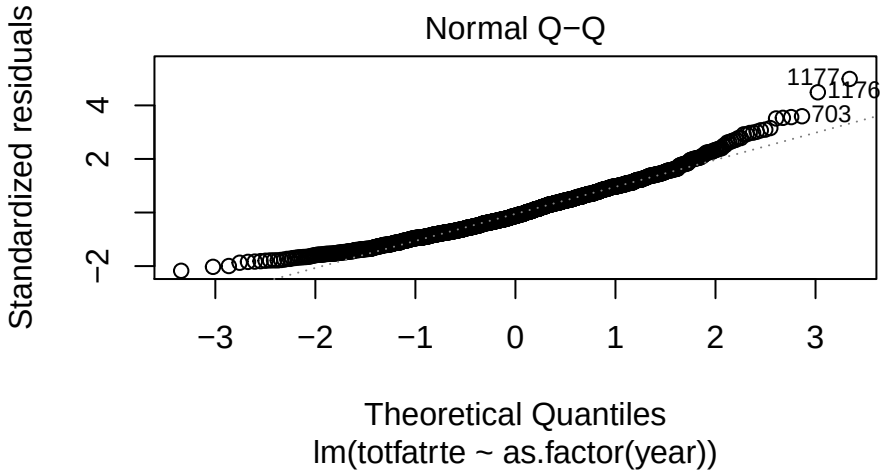
```
## as.factor(year)1984  -5.2271     1.2263  -4.263 2.18e-05 ***
## as.factor(year)1985  -5.6431     1.2263  -4.602 4.64e-06 ***
## as.factor(year)1986  -4.6942     1.2263  -3.828 0.000136 ***
## as.factor(year)1987  -4.7198     1.2263  -3.849 0.000125 ***
## as.factor(year)1988  -4.6029     1.2263  -3.754 0.000183 ***
## as.factor(year)1989  -5.7223     1.2263  -4.666 3.42e-06 ***
## as.factor(year)1990  -5.9894     1.2263  -4.884 1.18e-06 ***
## as.factor(year)1991  -7.3998     1.2263  -6.034 2.14e-09 ***
## as.factor(year)1992  -8.3367     1.2263  -6.798 1.68e-11 ***
## as.factor(year)1993  -8.3669     1.2263  -6.823 1.43e-11 ***
## as.factor(year)1994  -8.3394     1.2263  -6.800 1.66e-11 ***
## as.factor(year)1995  -7.8260     1.2263  -6.382 2.51e-10 ***
## as.factor(year)1996  -8.1252     1.2263  -6.626 5.25e-11 ***
## as.factor(year)1997  -7.8840     1.2263  -6.429 1.86e-10 ***
## as.factor(year)1998  -8.2292     1.2263  -6.711 3.01e-11 ***
## as.factor(year)1999  -8.2442     1.2263  -6.723 2.77e-11 ***
## as.factor(year)2000  -8.6690     1.2263  -7.069 2.67e-12 ***
## as.factor(year)2001  -8.7019     1.2263  -7.096 2.21e-12 ***
## as.factor(year)2002  -8.4650     1.2263  -6.903 8.32e-12 ***
## as.factor(year)2003  -8.7310     1.2263  -7.120 1.88e-12 ***
## as.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
```

As we can see from the model output, almost all the parameters are statistically significant. This model explains that in the absence of any legislative policies, there was a secular trend of declining traffic fatalities occuring from 1980 to 2004. For instance, the coefficient on the dummy variable for 2004 indicates that the fatality rate was ~9 deaths per 100,000 people lower than it was in 1980. Therefore, we can conclude that driving became safer over this time period. It should also be noted, according to the Woolridge text:

"When we include a full set of year dummies…we cannot estimate the effect of any variable whose change across time is constant."

Hypothetically, if we were only following a certain subset of drivers across time, we would not be able to specify how much 'experience' they had driving because it would be indistinguishable from the linear time trend.

```
# Model Diagnose - residual QQ plot
plot(dummy.lm, 2)
```



Since the model only includes the factor dummy varirbales, we will only take a look at the normaly of the residuals. As we can see from the QQ plot of dummy.lm model, we can see that the residual is generally normal distributed except for the head and tail is slightly off.

3. (15%) 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.)

EDA showed that bac08, gdl, sl70plus, and perse were exponential in nature and so we use logarithmic transformations for all. We add 1 to each logarithmic operation to account for negative values (we don't want to work with imaginary numbers.)

(Other variables?)

```
# specifying expanded OLS model
dummy.lm.expanded <- lm(totfatrte ~ as.factor(year) + log(bac08+1) + bac10 +
                          log(perse+1) + sbprim + sbsecon + log(sl70plus+1) +
                          log(gdl+1) + perc14_24 + unem + vehicmilespc, data=data)
summary(dummy.lm.expanded)
```

```
##
## Call:
## lm(formula = totfatrte ~ as.factor(year) + log(bac08 + 1) + bac10 +
##     log(perse + 1) + sbprim + sbsecon + log(sl70plus + 1) + log(gdl +
##     1) + perc14_24 + unem + vehicmilespc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9269  -2.7490  -0.2762   2.2880  21.4247
```

```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.649e+00  2.477e+00  -1.069  0.28518
## as.factor(year)1981 -2.175e+00  8.275e-01  -2.629  0.00868 **
## as.factor(year)1982 -6.595e+00  8.533e-01  -7.728 2.34e-14 ***
## as.factor(year)1983 -7.385e+00  8.692e-01  -8.497  < 2e-16 ***
## as.factor(year)1984 -5.860e+00  8.763e-01  -6.687 3.53e-11 ***
## as.factor(year)1985 -6.495e+00  8.947e-01  -7.260 7.07e-13 ***
## as.factor(year)1986 -5.867e+00  9.305e-01  -6.305 4.07e-10 ***
## as.factor(year)1987 -6.383e+00  9.668e-01  -6.603 6.12e-11 ***
## as.factor(year)1988 -6.604e+00  1.014e+00  -6.515 1.08e-10 ***
## as.factor(year)1989 -8.090e+00  1.052e+00  -7.687 3.19e-14 ***
## as.factor(year)1990 -8.977e+00  1.077e+00  -8.336  < 2e-16 ***
## as.factor(year)1991 -1.108e+01  1.101e+00 -10.067  < 2e-16 ***
## as.factor(year)1992 -1.290e+01  1.122e+00 -11.492  < 2e-16 ***
## as.factor(year)1993 -1.274e+01  1.136e+00 -11.214  < 2e-16 ***
## as.factor(year)1994 -1.238e+01  1.157e+00 -10.698  < 2e-16 ***
## as.factor(year)1995 -1.198e+01  1.184e+00 -10.125  < 2e-16 ***
## as.factor(year)1996 -1.398e+01  1.226e+00 -11.408  < 2e-16 ***
## as.factor(year)1997 -1.429e+01  1.251e+00 -11.430  < 2e-16 ***
## as.factor(year)1998 -1.507e+01  1.265e+00 -11.909  < 2e-16 ***
## as.factor(year)1999 -1.511e+01  1.285e+00 -11.757  < 2e-16 ***
## as.factor(year)2000 -1.546e+01  1.306e+00 -11.835  < 2e-16 ***
## as.factor(year)2001 -1.617e+01  1.336e+00 -12.102  < 2e-16 ***
## as.factor(year)2002 -1.672e+01  1.349e+00 -12.397  < 2e-16 ***
## as.factor(year)2003 -1.699e+01  1.362e+00 -12.470  < 2e-16 ***
## as.factor(year)2004 -1.672e+01  1.387e+00 -12.051  < 2e-16 ***
## log(bac08 + 1)      -3.571e+00  7.746e-01  -4.611 4.46e-06 ***
## bac10               -1.404e+00  3.955e-01  -3.551  0.00040 ***
## log(perse + 1)      -8.966e-01  4.298e-01  -2.086  0.03718 *
## sbprim              -7.722e-02  4.908e-01  -0.157  0.87500
## sbsecon              6.724e-02  4.293e-01   0.157  0.87555
## log(sl70plus + 1)    4.844e+00  6.409e-01   7.557 8.29e-14 ***
## log(gdl + 1)        -6.442e-01  7.591e-01  -0.849  0.39625
## perc14_24            1.383e-01  1.227e-01   1.127  0.26003
## unem                 7.565e-01  7.790e-02   9.712  < 2e-16 ***
## vehicmilespc         2.925e-03  9.499e-05  30.788  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.045 on 1165 degrees of freedom
## Multiple R-squared:  0.6078, Adjusted R-squared:  0.5964
## F-statistic: 53.11 on 34 and 1165 DF,  p-value: < 2.2e-16
```

bac08 and bac10 are defined as binary variables that represent states either having or not having a law in place which places a legal limit on blood alcohol content at either 0.08% or 0.10% respectively. The coefficients on these variables are both highly statistically significant. Specifically, holding

all other variables constant, for the bac08 variable, as can be seen below, every 1% increase of bac08 limit corresponding to 0.036% decrease of fatality rate; while the presence of of the law corresponding to every 1 unit bac10 limit increase corresponds with a decrease in fatality rates of about -1.4%.

```
# coefficient for bac08
bac08.ols <- dummy.lm.expanded$coefficients['log(bac08 + 1)']/100
bac08.ols
```

```
## log(bac08 + 1)
##    -0.03571129
```

```
# coefficient for bac10
bac10.ols <- dummy.lm.expanded$coefficients['bac10']
bac10.ols
```

```
##     bac10
## -1.404195
```

Perse laws do appear to have a small fraction of correspondence to a reduction in fatality rates as shown below. Every 1% increase of the perse, we expect the total fatality rate will decrease by 0.009%.

```
# coefficient for perse
perse.ols <- dummy.lm.expanded$coefficients['log(perse + 1)']/100
perse.ols
```

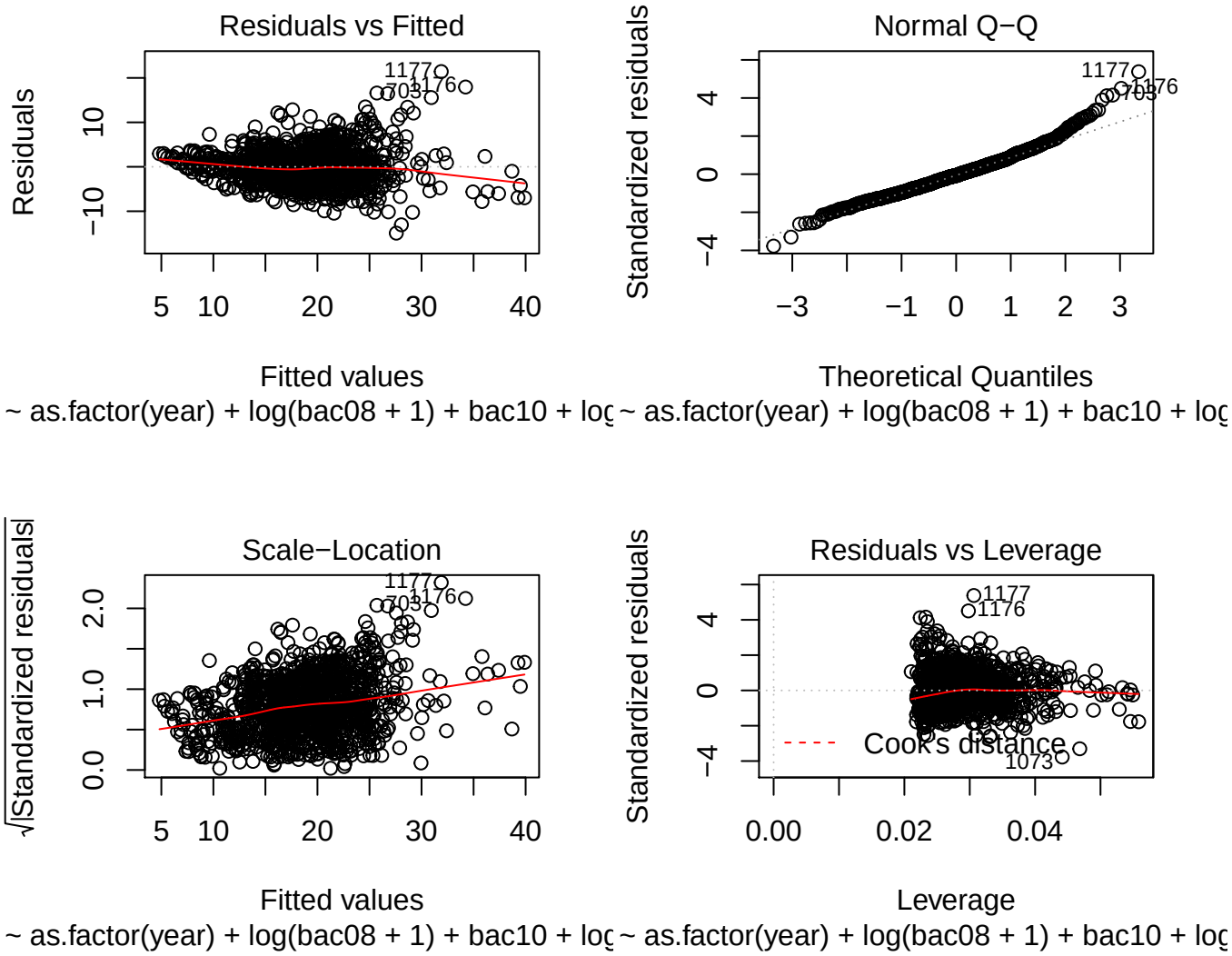
```
## log(perse + 1)
##    -0.00896565
```

Primary seatbelt laws do seem to correspond to a decrease in fatality rates as well.

```
sbprim.ols <- dummy.lm.expanded$coefficients['sbprim']
sbprim.ols
```

```
##      sbprim
## -0.07721643
```

```
# Model Diagnose
plot(dummy.lm.expanded)
```



Based on diagnostic plots of dummy.lm.expanded model, we can see that there is a pattern of the mean and variance of the residuals which could due to lack of data at the tail; the normality of the residuals is bascially the same to dummy.lm model.

4. (15%) 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?

```
traffic.fe <- plm(totfatrte ~ as.factor(year) + log(bac08+1) + bac10 + log(perse+1) +
                    sbprim + sbsecon + log(sl70plus+1) + log(gdl+1) + perc14_24 +
                    unem + vehicmilespc, data=data,
                  index=c("state","year"), model="within")
summary(traffic.fe)
```

```
## Oneway (individual) effect Within Model
##
```

```
## Call:
## plm(formula = totfatrte ~ as.factor(year) + log(bac08 + 1) +
##     bac10 + log(perse + 1) + sbprim + sbsecon + log(sl70plus +
##     1) + log(gdl + 1) + 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.
## -8.4320421 -1.0173616  0.0072081  0.9517076 14.8047879
##
## Coefficients:
##                          Estimate  Std. Error  t-value  Pr(>|t|)
## as.factor(year)1981 -1.5098e+00  4.1331e-01  -3.6529 0.0002713 ***
## as.factor(year)1982 -3.0206e+00  4.4251e-01  -6.8261 1.428e-11 ***
## as.factor(year)1983 -3.4880e+00  4.5678e-01  -7.6361 4.787e-14 ***
## as.factor(year)1984 -4.2687e+00  4.6496e-01  -9.1807 < 2.2e-16 ***
## as.factor(year)1985 -4.7400e+00  4.8543e-01  -9.7646 < 2.2e-16 ***
## as.factor(year)1986 -3.6781e+00  5.1760e-01  -7.1061 2.122e-12 ***
## as.factor(year)1987 -4.3246e+00  5.5520e-01  -7.7893 1.531e-14 ***
## as.factor(year)1988 -4.7827e+00  6.0158e-01  -7.9503 4.523e-15 ***
## as.factor(year)1989 -6.1512e+00  6.4008e-01  -9.6100 < 2.2e-16 ***
## as.factor(year)1990 -6.2480e+00  6.6484e-01  -9.3977 < 2.2e-16 ***
## as.factor(year)1991 -6.9366e+00  6.8201e-01 -10.1708 < 2.2e-16 ***
## as.factor(year)1992 -7.7944e+00  7.0289e-01 -11.0891 < 2.2e-16 ***
## as.factor(year)1993 -8.1102e+00  7.1610e-01 -11.3256 < 2.2e-16 ***
## as.factor(year)1994 -8.5204e+00  7.3422e-01 -11.6048 < 2.2e-16 ***
## as.factor(year)1995 -8.2777e+00  7.5638e-01 -10.9439 < 2.2e-16 ***
## as.factor(year)1996 -8.6284e+00  7.9795e-01 -10.8133 < 2.2e-16 ***
## as.factor(year)1997 -8.7298e+00  8.2037e-01 -10.6412 < 2.2e-16 ***
## as.factor(year)1998 -9.3746e+00  8.3380e-01 -11.2433 < 2.2e-16 ***
## as.factor(year)1999 -9.4924e+00  8.4425e-01 -11.2436 < 2.2e-16 ***
## as.factor(year)2000 -1.0007e+01  8.5642e-01 -11.6850 < 2.2e-16 ***
## as.factor(year)2001 -9.6301e+00  8.7380e-01 -11.0210 < 2.2e-16 ***
## as.factor(year)2002 -8.9150e+00  8.8271e-01 -10.0995 < 2.2e-16 ***
## as.factor(year)2003 -8.9280e+00  8.9196e-01 -10.0095 < 2.2e-16 ***
## as.factor(year)2004 -9.3578e+00  9.1124e-01 -10.2693 < 2.2e-16 ***
## log(bac08 + 1)      -2.0110e+00  5.6731e-01  -3.5449 0.0004091 ***
## bac10               -1.0377e+00  2.6788e-01  -3.8736 0.0001135 ***
## log(perse + 1)      -1.6741e+00  3.3802e-01  -4.9525 8.454e-07 ***
## sbprim              -1.2267e+00  3.4279e-01  -3.5785 0.0003604 ***
## sbsecon             -3.4816e-01  2.5223e-01  -1.3803 0.1677644
## log(sl70plus + 1)   -8.0903e-02  3.8867e-01  -0.2082 0.8351446
## log(gdl + 1)        -6.2162e-01  4.2198e-01  -1.4731 0.1410077
## perc14_24            1.8443e-01  9.5248e-02   1.9363 0.0530767 .
## unem                -5.7275e-01  6.0591e-02  -9.4528 < 2.2e-16 ***
## vehicmilespc         9.3856e-04  1.1112e-04   8.4462 < 2.2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4537.4
## R-Squared:      0.62606
## Adj. R-Squared: 0.59897
## F-statistic: 55.0528 on 34 and 1118 DF, p-value: < 2.22e-16
```

```
bac08.fe <- traffic.fe$coefficients['log(bac08 + 1)']/100
bac08.fe - bac08.ols
```

```
## log(bac08 + 1)
##     0.01560112
```

```
bac10.fe <- (traffic.fe$coefficients['bac10'])
bac10.fe - bac10.ols
```

```
##     bac10
## 0.3665331
```

```
perse.fe <- traffic.fe$coefficients['log(perse + 1)']/100
perse.fe - perse.ols
```

```
## log(perse + 1)
##   -0.007775114
```

```
sbprim.fe <- traffic.fe$coefficients['sbprim']
sbprim.fe - sbprim.ols
```

```
##    sbprim
## -1.149452
```

Data indicates that the fixed effect model calculates larger coeffcicients for bac08 and bac10, but smaller coefficients for perse and sbprim. The fixed effects model is most likely more appropriate in this instance because it gets rid of unobserved effects (i.e. things like geography that aren't captured in the model) and only looks at the time variation in y and x within each cross-sectional observation. This is valuable, because we are dealing with different states, and different states have different legal frameworks. This is the power of the fixed effect model over the random effect model.

The fixed effect model requires a strict exogeneity assumption (idiosyncratic errors should be uncorrelated with all explanatory variables over all time periods). In addition to this, an OLS model requires that the errors are homoskedastic and serially uncorrelated across all time periods.

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

To assess whether a fixed or random effects model is preferred, we perform a Hausman test, the null hypothesis of which is that the preferred model is random effects vs. the alternative the fixed effects. It tests whether the unique errors u_i are correlated with the explanatory variables; the null hypothesis is they are not.

```
traffic.re <- plm(totfatrte ~ as.factor(year) + log(bac08+1) + bac10 +
                    log(perse+1) + sbprim + sbsecon + log(sl70plus+1) +
                    log(gdl+1) + perc14_24 + unem + vehicmilespc,
                  data=data, index=c("state","year"), model="random")
phtest(traffic.fe, traffic.re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ as.factor(year) + log(bac08 + 1) + bac10 + log(perse +  ...
## chisq = 147.96, df = 34, p-value = 3.627e-16
## alternative hypothesis: one model is inconsistent
```

The Hausman test rejects the null hypothesis that unique errors are not correlated with the explanatory variables, and therefore the fixed effect model is preferred.

6. (5%) 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.

```
# increase in dependent variable given increase in miles driven per capita
traffic.fe$coefficients['vehicmilespc']*1000
```

```
## vehicmilespc
##    0.9385644
```

According to the data above, an increase in vehicle miles driven per capita would lead to an increase in the total fatality rate of about 0.94%($\pm$ 0.22%).

7. (5%) 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 presence of serial correlation and heteroskedasticity do not affect the coefficients themselves, but cause the model to underestimate the true standard errors, which can lead to larger t-statistics, and wrongful conclusions that the coefficients are significant, when they are in fact, not. This is an error with false positives, which can be more harmful than false negatives (taking a conservative approach.)