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

Sue Yang, Michelle Kim, Legg Yeung

August 20, 2017

Intro

This exercise intends to answer the question **"Do changes in traffic laws affect traffic fatalities?"**. To do so, we use data for the 48 continental U.S. states from 1980 through 2004 that cover changes in various traffic laws and a few economics and demographic variables.

We apply pooled OLS, fixed effect estimators (both first differenced and within) and random effect estimator. Through comparison between the theoretical foundation, mechanism and results between these options, we provide statistical insights to the research question using results from the within model.

Overall, our approach is to treat the dataset as a panel dataset which should be examined both as cross-sections and panels. Through techniques such as first-differencing, time-demeaning and GLS transformations, the preferred panel models eliminate omitted variable bias generated by state-fixed, time invariant information and improves efficiency by adjusting serial correlation.

We follow a set of procedures, including data exploration, model construction, assumptions evaluation and coefficient interpretation to conclude that some traffic laws are supported by statistical evidence to be associated with lower total fatality rate.

Exercises:

1. Load the data. Provide a description of the basic structure of the dataset, as we have done in 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. Remember, graphs must be well-labeled. You need to write a detailed narrative of your observations of your EDA.

```
rm(list=ls())
setwd("~/Desktop/W271/Lab4_2")
#setwd("D:/UCB/w271/lab4/")
load("driving.Rdata")
df<-data
library(car)
```

```
## Warning: package 'car' was built under R version 3.3.2
```

```
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
library(plm)
```

```
## Warning: package 'plm' was built under R version 3.3.2
## Loading required package: Formula
## Warning: package 'Formula' was built under R version 3.3.2
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.3.2
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(reshape)
```

```
## Warning: package 'reshape' was built under R version 3.3.2
```

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 3.3.2
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.3.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.3.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

```
# year dummies omitted
str(df[, c(1:30, 56)])
```

```
## 'data.frame':    1200 obs. of  31 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 ...
##  $ zerotol     : 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 403022
##  $ 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 ...
##  $ vehicmilespc: num  7544 7108 7607 7880 8334 ...
```

```
summary(df[, c(1:30, 56)])
```

```
##       year          state            sl55             sl65
##  Min.   :1980   Min.   : 1.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1986   1st Qu.:15.75   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1992   Median :27.50   Median :0.0000   Median :0.0000
##  Mean   :1992   Mean   :27.15   Mean   :0.3533   Mean   :0.4399
##  3rd Qu.:1998   3rd Qu.:39.25   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :2004   Max.   :51.00   Max.   :1.0000   Max.   :1.0000
##       sl70            sl75             slnone            seatbelt
##  Min.   :0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000
##  Median :0.000   Median :0.00000   Median :0.000000   Median :1.000
##  Mean   :0.119   Mean   :0.08024   Mean   :0.007569   Mean   :1.116
##  3rd Qu.:0.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:2.000
##  Max.   :1.000   Max.   :1.00000   Max.   :1.000000   Max.   :2.000
##      minage        zerotol            gdl              bac10
##  Min.   :18.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:21.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :21.0   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :20.6   Mean   :0.4519   Mean   :0.1741   Mean   :0.6231
##  3rd Qu.:21.0   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :21.0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      bac08            perse            totfat          nghtfat
##  Min.   :0.0000   Min.   :0.0000   Min.   :  63.0   Min.   :  26.0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 310.0   1st Qu.: 139.8
##  Median :0.0000   Median :1.0000   Median : 676.0   Median : 316.0
##  Mean   :0.2135   Mean   :0.5471   Mean   : 900.7   Mean   : 427.3
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1099.5   3rd Qu.: 518.2
##  Max.   :1.0000   Max.   :1.0000   Max.   :5504.0   Max.   :2918.0
##     wkndfat         totfatpvm       nghtfatpvm       wkndfatpvm
##  Min.   :  10.0   Min.   :0.780   Min.   :0.2700   Min.   :0.1140
##  1st Qu.:  70.0   1st Qu.:1.577   1st Qu.:0.6847   1st Qu.:0.3410
##  Median : 163.0   Median :2.020   Median :0.9130   Median :0.4770
##  Mean   : 222.3   Mean   :2.122   Mean   :0.9990   Mean   :0.5255
##  3rd Qu.: 277.0   3rd Qu.:2.500   3rd Qu.:1.2110   3rd Qu.:0.6420
##  Max.   :1499.0   Max.   :5.700   Max.   :3.0030   Max.   :1.6750
##     statepop          totfatrte       nghtfatrte       wkndfatrte
##  Min.   :  453401   Min.   : 6.20   Min.   : 2.660   Min.   : 1.180
##  1st Qu.: 1641938   1st Qu.:14.38   1st Qu.: 6.338   1st Qu.: 3.240
##  Median : 3700425   Median :18.43   Median : 8.420   Median : 4.390
##  Mean   : 5329896   Mean   :18.92   Mean   : 8.796   Mean   : 4.606
##  3rd Qu.: 6069563   3rd Qu.:22.77   3rd Qu.:10.650   3rd Qu.: 5.680
##  Max.   :35894000   Max.   :53.32   Max.   :29.600   Max.   :14.430
##    vehicmiles           unem          perc14_24        sl70plus
```

```
##  Min.   :  3.703   Min.   : 2.200   Min.   :11.70   Min.   :0.0000
##  1st Qu.: 14.574   1st Qu.: 4.500   1st Qu.:13.90   1st Qu.:0.0000
##  Median : 33.863   Median : 5.600   Median :14.90   Median :0.0000
##  Mean   : 46.323   Mean   : 5.951   Mean   :15.33   Mean   :0.2068
##  3rd Qu.: 58.639   3rd Qu.: 7.000   3rd Qu.:16.60   3rd Qu.:0.0000
##  Max.   :329.600   Max.   :18.000   Max.   :20.30   Max.   :1.0000
##      sbprim          sbsecon        vehicmilespc
##  Min.   :0.0000   Min.   :0.0000   Min.   : 4372
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 7788
##  Median :0.0000   Median :0.0000   Median : 9013
##  Mean   :0.1792   Mean   :0.4683   Mean   : 9129
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:10327
##  Max.   :1.0000   Max.   :1.0000   Max.   :18390
```

Examining the structure of the data as well as a summary, we note the following:

- The *state* variable is currently coded as an integer. We will want to transform this into a factor variable, and ideally match each level up with the appropriate state name.
- This data ranges from 1980 to 2004.
- The speed limit variables add up to 100% across each category.
- None of the variables have any missing values.

In addition to 25 indicator variables for year, there are 31 variables to this dataset. Our key variables of interest are as follows:

```
- totfatrte : total fatalities per 100,000 population
- bac08, bac10 : blood alcohol limit (.08 and .10 respectively)
- perse : administrative license revocation (per se law)
- sbprim : =1 if primary seatbelt law is enforced
- minage : minimum drinking age
- sbsecon : =1 if secondary seatbelt law is enforced
- sl70plus : speed limit >= 70
- gdl : graduated drivers license law
- perc14_24 : percent population aged 14 through 24
- unem : unemployment rate, percent
- vehicmilespc : vehicle miles per capita
```

Next, we match up the state indicators with the appropriate state name and calculate each state's average population over the years as a sanity check that our matching worked correctly.

```
states <- c(state.name, "District of Columbia")
states <- data.frame(state.name = stringr::str_sort(states),
    id = c(1:51))
df2 = merge(df, states, by.x = "state", by.y = "id")
statepops <- aggregate(df2[, 21], list(df2$state.name), mean)
head(statepops)
```

```
##       Group.1        x
## 1     Alabama  4187578
## 2     Arizona  4084063
## 3    Arkansas  2473339
## 4  California 30187070
## 5    Colorado  3659807
## 6 Connecticut  3300454
```

```
tail(statepops)
```

```
##          Group.1        x
```

```
## 43       Vermont  569551.4
## 44      Virginia 6388063.9
## 45    Washington 5137124.7
## 46 West Virginia 1847114.4
## 47     Wisconsin 5047634.7
## 48       Wyoming  486035.3
```

The average populations look reasonable, so we will proceed with our analysis.

Next we will create a new factor variable denoting the BAC Limit for each state.

```
df2$baclim <- as.factor(ifelse(df2$bac10 > 0, "0.1", ifelse(df2$bac08 >
    0, "0.08", "None")))
```

```
table(df2$year)
```

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

```
table(df2$state.name)
```

```
##
##              Alabama               Alaska              Arizona
##                   25                    0                   25
##             Arkansas           California             Colorado
##                   25                   25                   25
##          Connecticut             Delaware District of Columbia
##                   25                   25                    0
##              Florida              Georgia               Hawaii
##                   25                   25                    0
##                Idaho             Illinois              Indiana
##                   25                   25                   25
##                 Iowa               Kansas             Kentucky
##                   25                   25                   25
##            Louisiana                Maine             Maryland
##                   25                   25                   25
##        Massachusetts             Michigan            Minnesota
##                   25                   25                   25
##          Mississippi             Missouri              Montana
##                   25                   25                   25
##             Nebraska               Nevada        New Hampshire
##                   25                   25                   25
##           New Jersey           New Mexico             New York
##                   25                   25                   25
##       North Carolina         North Dakota                 Ohio
##                   25                   25                   25
##             Oklahoma               Oregon         Pennsylvania
##                   25                   25                   25
##         Rhode Island       South Carolina         South Dakota
##                   25                   25                   25
##            Tennessee                Texas                 Utah
##                   25                   25                   25
##              Vermont             Virginia           Washington
##                   25                   25                   25
```

```
##        West Virginia            Wisconsin              Wyoming
##                   25                   25                   25
```

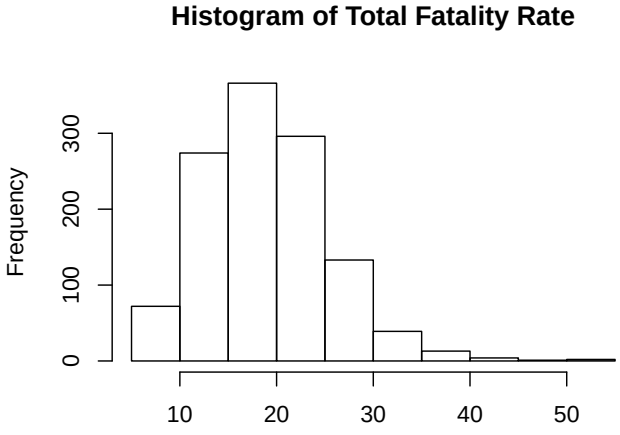
We note that we have a balanced panel, with 48 observations for each year (one for each of the 48 continental states), and 25 observations for each state (one for each year). As expected, we have no observations for Alaska, D.C., or Hawaii.

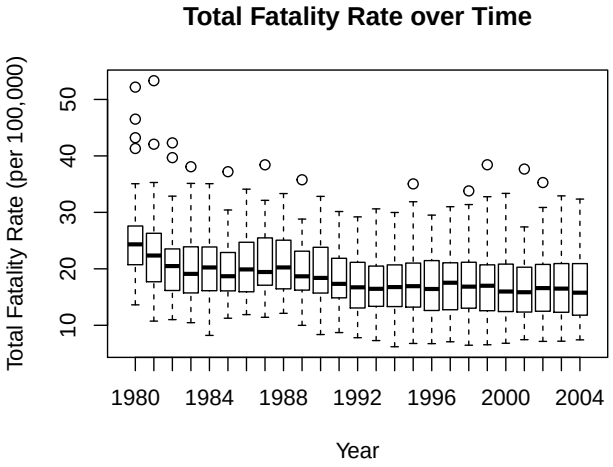
Univariate Investigation

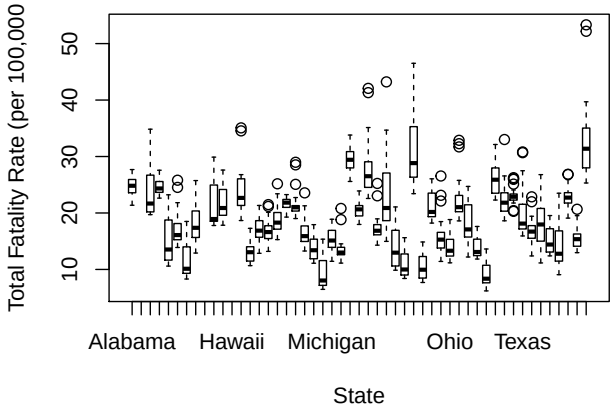
Next we will examine the distributions of each variable, as well as how or if the distribution changes over time and from state to state.

Key Dependent Variable : Total Fatality Rate

```
hist(df2$totfatrte, main = "Histogram of Total Fatality Rate",
    xlab = "Total Fatality Rate (per 100,000)")
boxplot(df2$totfatrte ~ df2$year, main = "Total Fatality Rate over Time",
    xlab = "Year", ylab = "Total Fatality Rate (per 100,000)")
boxplot(df2$totfatrte ~ df2$state.name, main = "Total Fatality Rate by State",
    xlab = "State", ylab = "Total Fatality Rate (per 100,000")
```







Mean total fatality rate in 1980: 25.4945837

Mean total fatality rate in 2004: 16.7289584

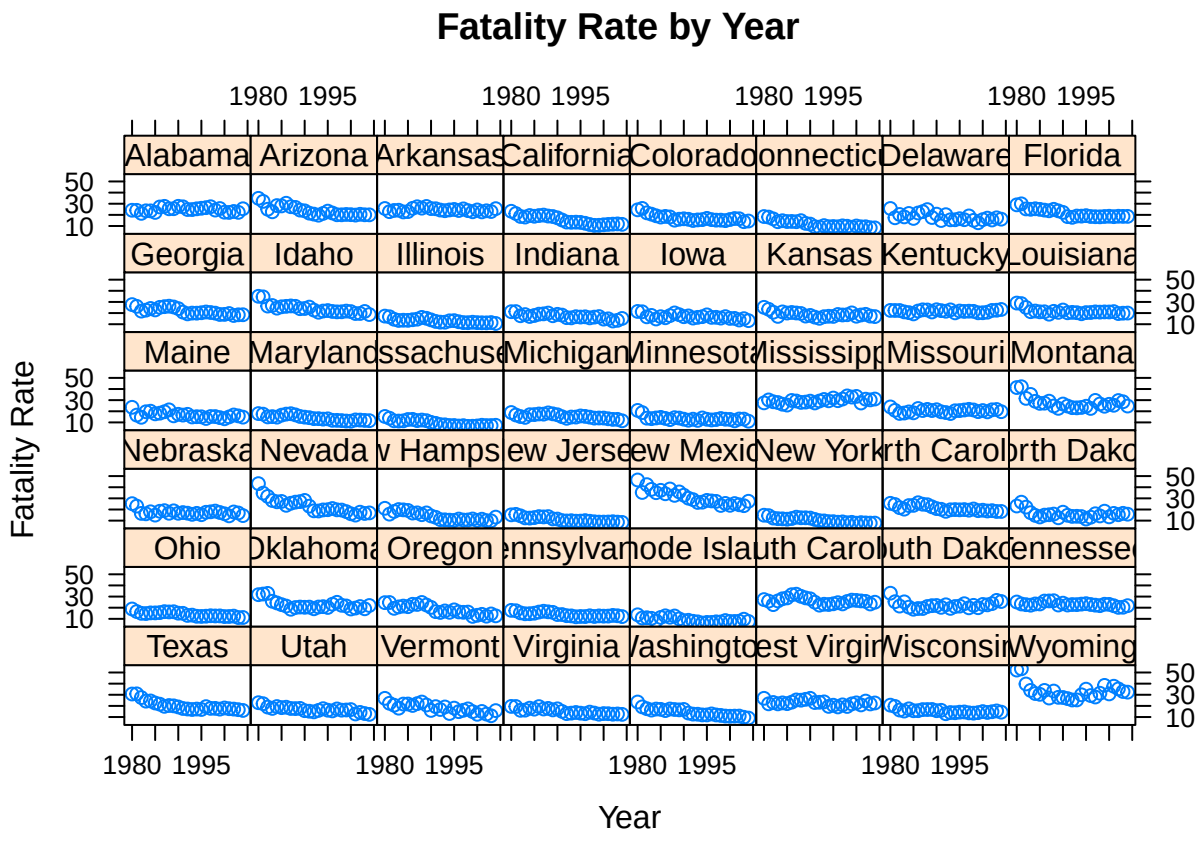
Mean total fatality rate of Wyoming: 33.1408

Mean total fatality rate of Massachusetts: 9.4512

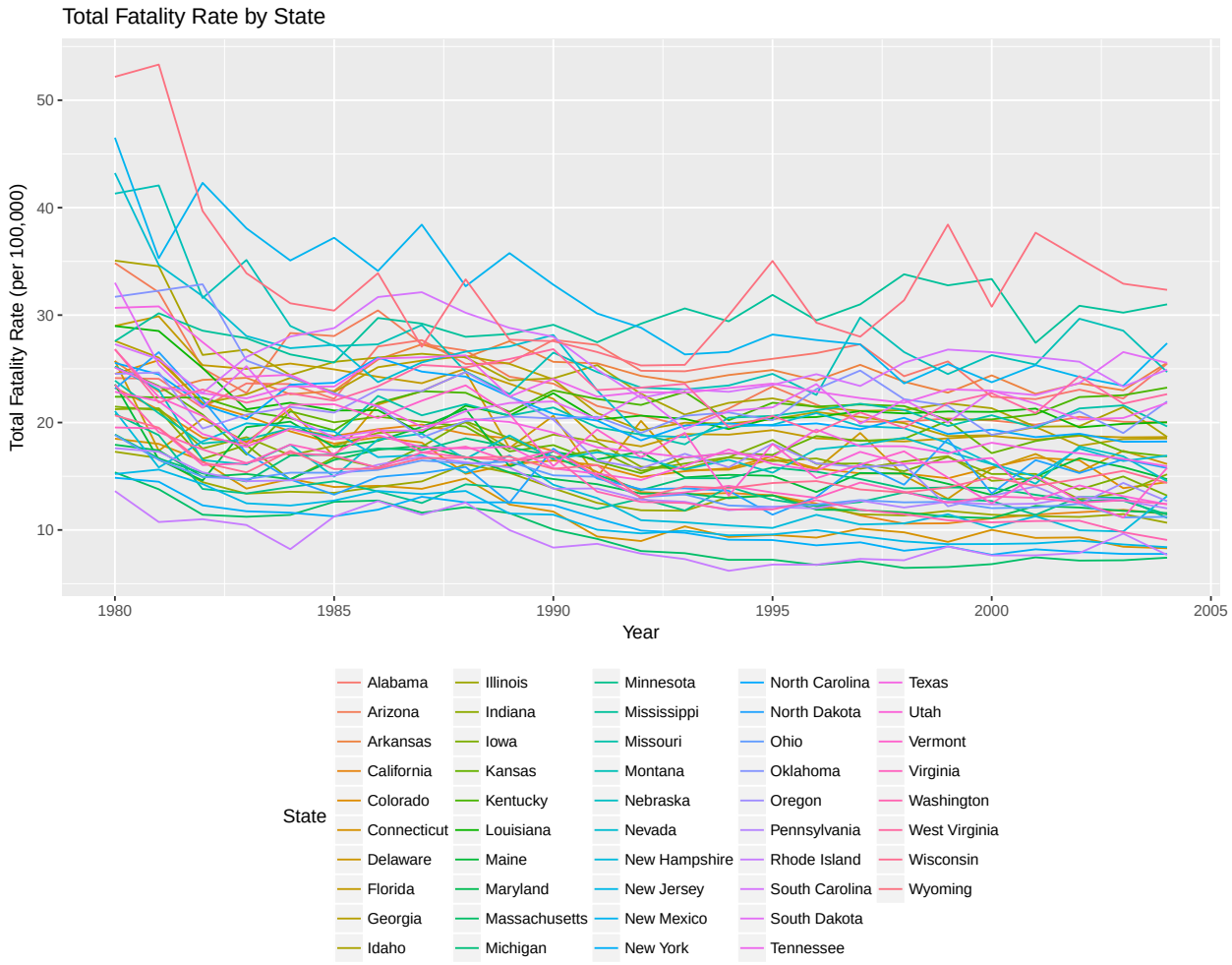
Mean total fatality rate of Rhode Island: 9.09

The distribution of *totfatrte* seems right skewed but not too severely. Across cross sections (time) from 1980 to 2004, we observe a gradual decrease in the mean of total fatality rate. The distributions also appear to shift (slightly more right skewed). After 1990, the variance of total fatality rate seem stable and there are fewer outliers. Across panels (states) we see much more variation in both mean and variance. For instance, Wyoming (far right) has the highest mean at 33.14, a lower bound that exceeds the higher bound of many other states and two outliers higher than 50 fatalities per 100,000 population. Other states, such as Massachusetts and Rhode Island, have the lowest means less than 10, upper bounds that clear the lower bounds of many other states and no outliers.

```
lattice::xyplot(totfatrte ~ year | state.name, data = df2, as.table = T,
    xlab = "Year", ylab = "Fatality Rate", main = "Fatality Rate by Year")
```



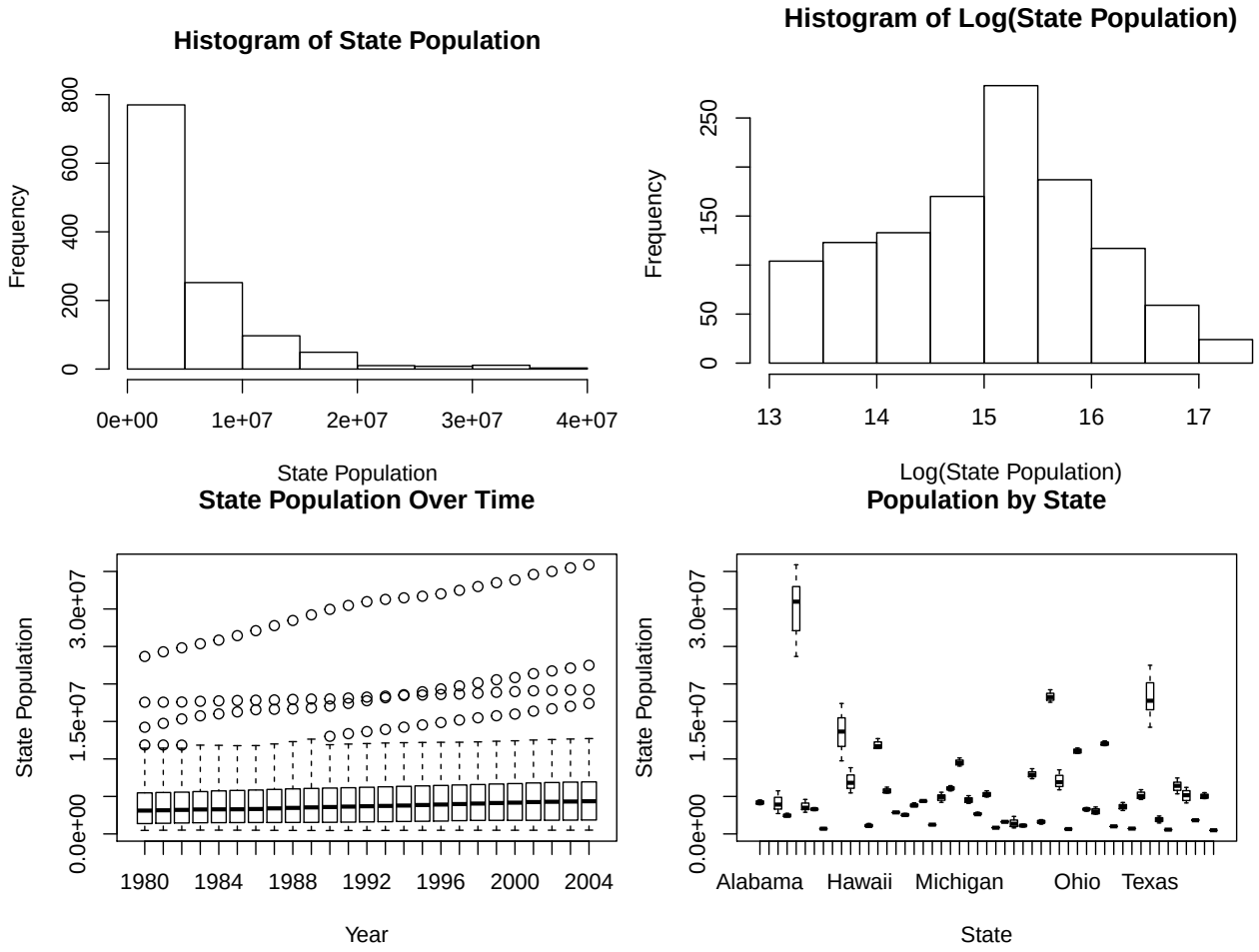

```
ggplot(data = df2, aes(x = year, y = totfatrte, group = state.name,
    colour = state.name)) + geom_line() + ggtitle("Total Fatality Rate by State") +
    xlab("Year") + ylab("Total Fatality Rate (per 100,000)") +
    labs(color = "State") + theme(legend.position = "bottom")
```



The time series plot by state suggests that the decreasing trend in total fatality rate doesn't manifest in all states. Montana, Nevada, New Mexico, and Wyoming exhibit a clear decreasing trend while other states such as Kentucky and Pennsylvania exhibit flat trends.

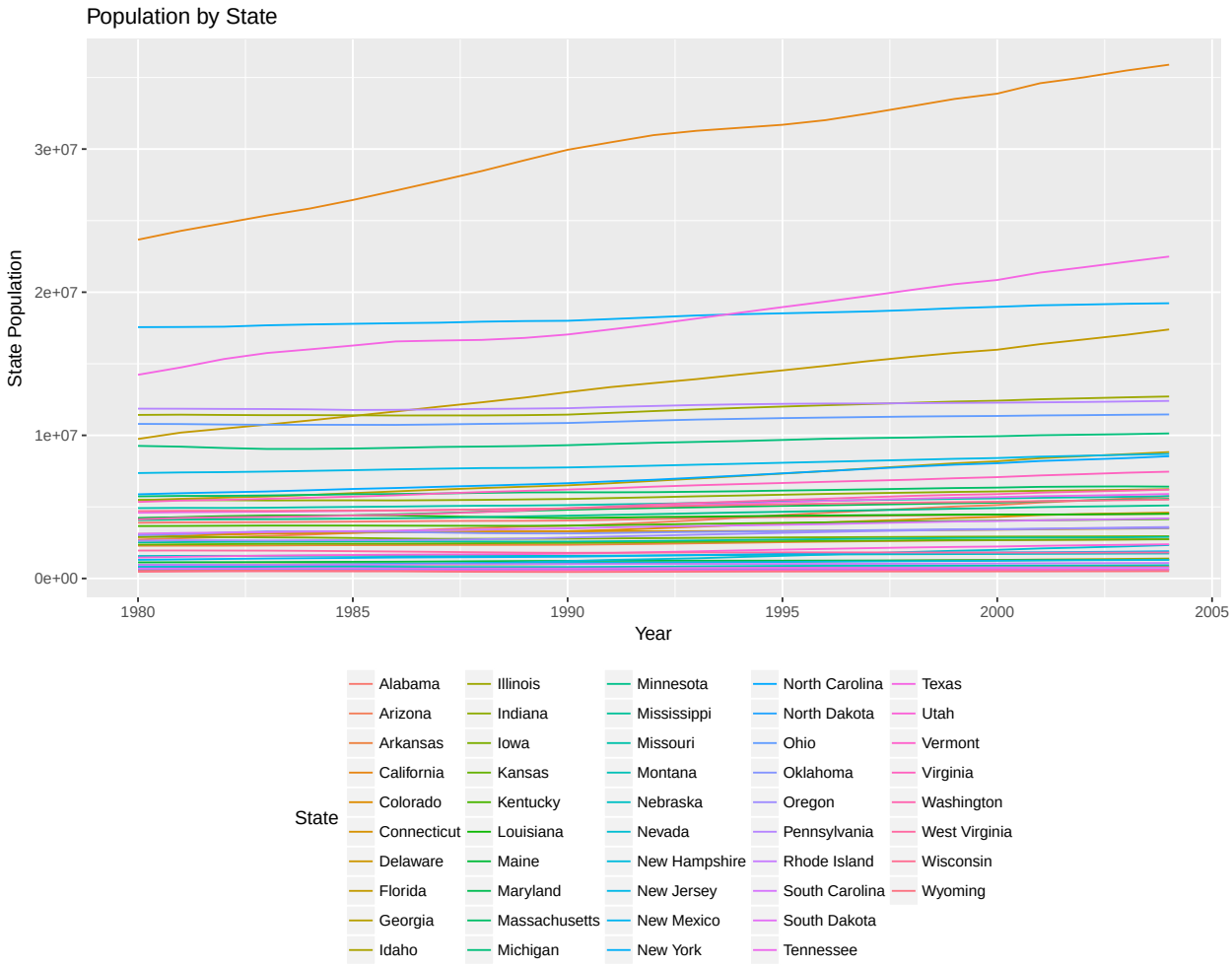
Independent Variable : State Population

```
hist(df2$statepop, main = "Histogram of State Population", xlab = "State Population")
hist(log(df2$statepop), main = "Histogram of Log(State Population)",
    xlab = "Log(State Population)")
boxplot(df2$statepop ~ df2$year, main = "State Population Over Time",
    xlab = "Year", ylab = "State Population")
boxplot(df2$statepop ~ df2$state.name, main = "Population by State",
    xlab = "State", ylab = "State Population")
```



State population is clearly right skewed; a log transformation can adjust the distribution closer to normal. Across cross sections (time), we observe a relatively stable range but 3 clearly trending outliers. Across panels (states) we observe various ranges of population. California, Texas and Florida exhibit far wider ranges than the rest of the states.

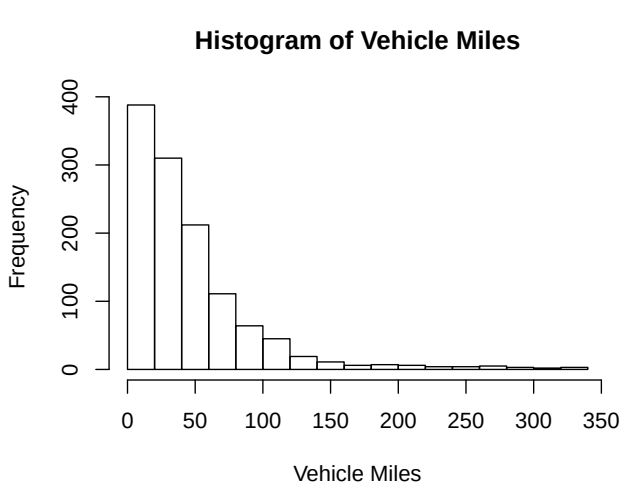
```
ggplot(data = df2, aes(x = year, y = statepop, group = state.name,
    colour = state.name)) + geom_line() + ggtitle("Population by State") +
    xlab("Year") + ylab("State Population") + labs(color = "State") +
    theme(legend.position = "bottom")
```

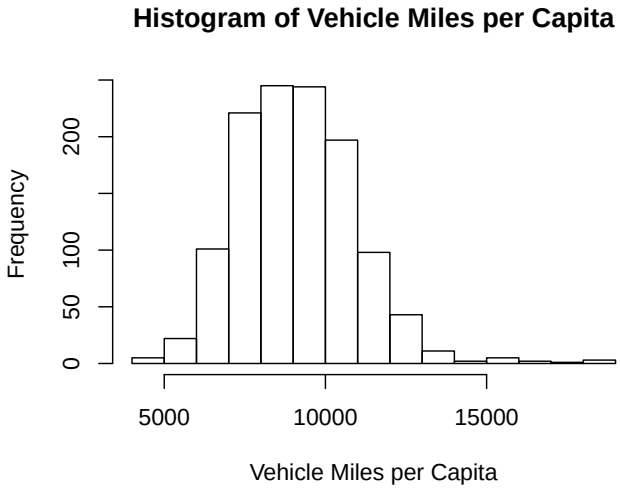


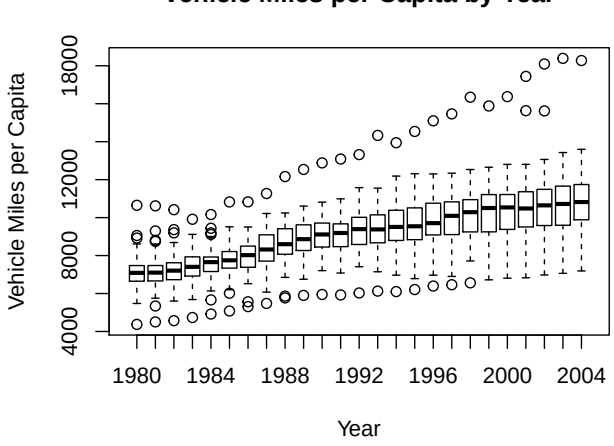
The time series plot by state echoes the panel box plot. Clearly California, Texas and Florida have been trending up. This agrees with existing population studies.

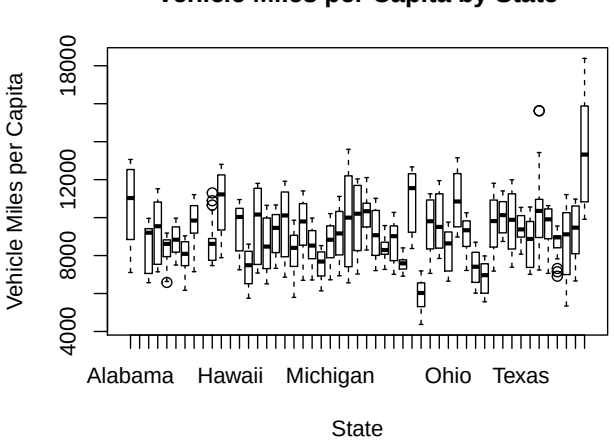
Independent Variable : Vehicle Miles per Capita

```
hist(df2$vehicmiles, main = "Histogram of Vehicle Miles", xlab = "Vehicle Miles")
hist(df2$vehicmilespc, main = "Histogram of Vehicle Miles per Capita",
    xlab = "Vehicle Miles per Capita")
boxplot(df2$vehicmilespc ~ df2$year, main = "Vehicle Miles per Capita by Year",
    xlab = "Year", ylab = "Vehicle Miles per Capita")
boxplot(df2$vehicmilespc ~ df2$state.name, main = "Vehicle Miles per Capita by State",
    xlab = "State", ylab = "Vehicle Miles per Capita")
```

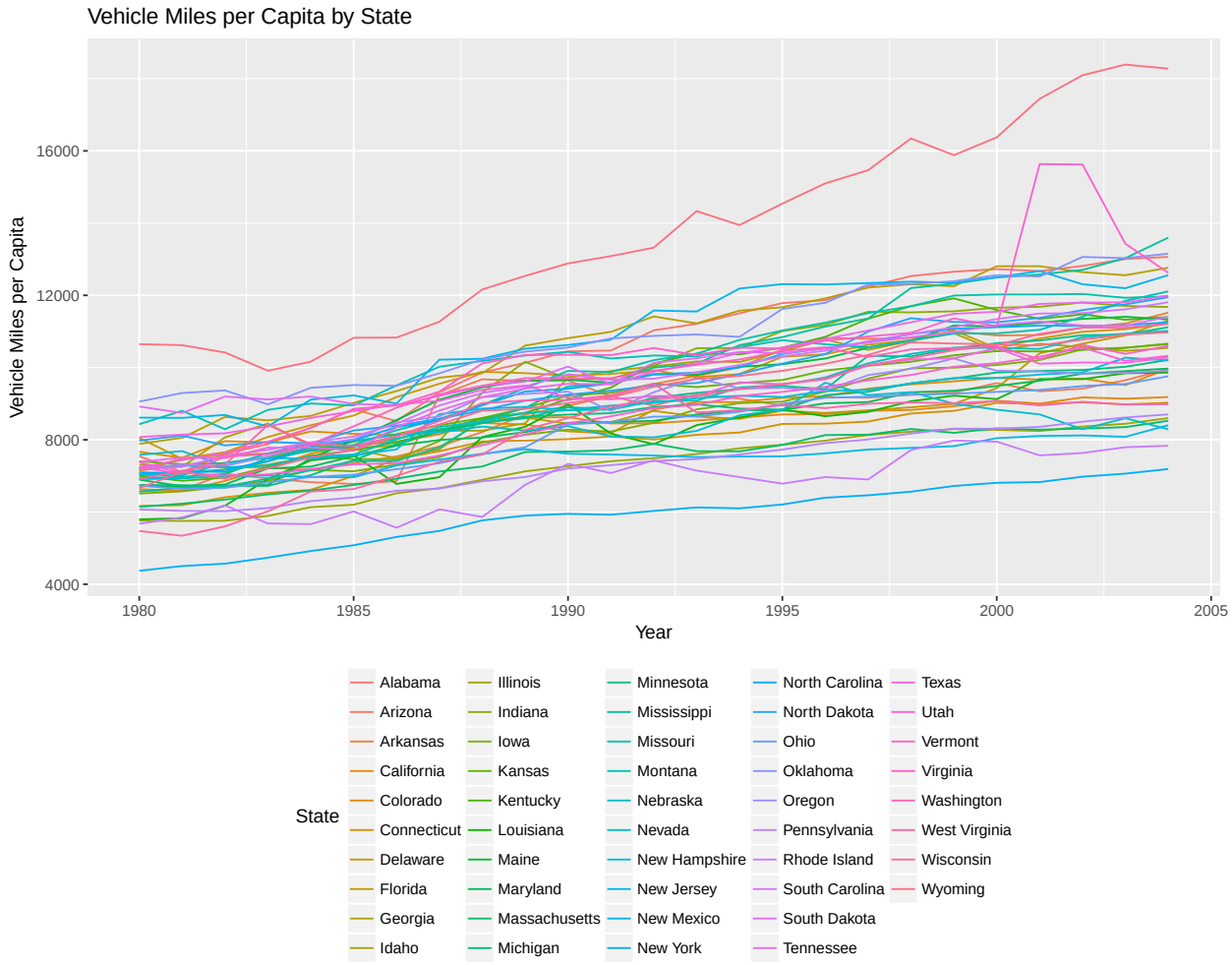









```
ggplot(data = df2, aes(x = year, y = vehicmilespc, group = state.name,
    colour = state.name)) + geom_line() + ggtitle("Vehicle Miles per Capita by State") +
    xlab("Year") + ylab("Vehicle Miles per Capita") + labs(color = "State") +
    theme(legend.position = "bottom")
```

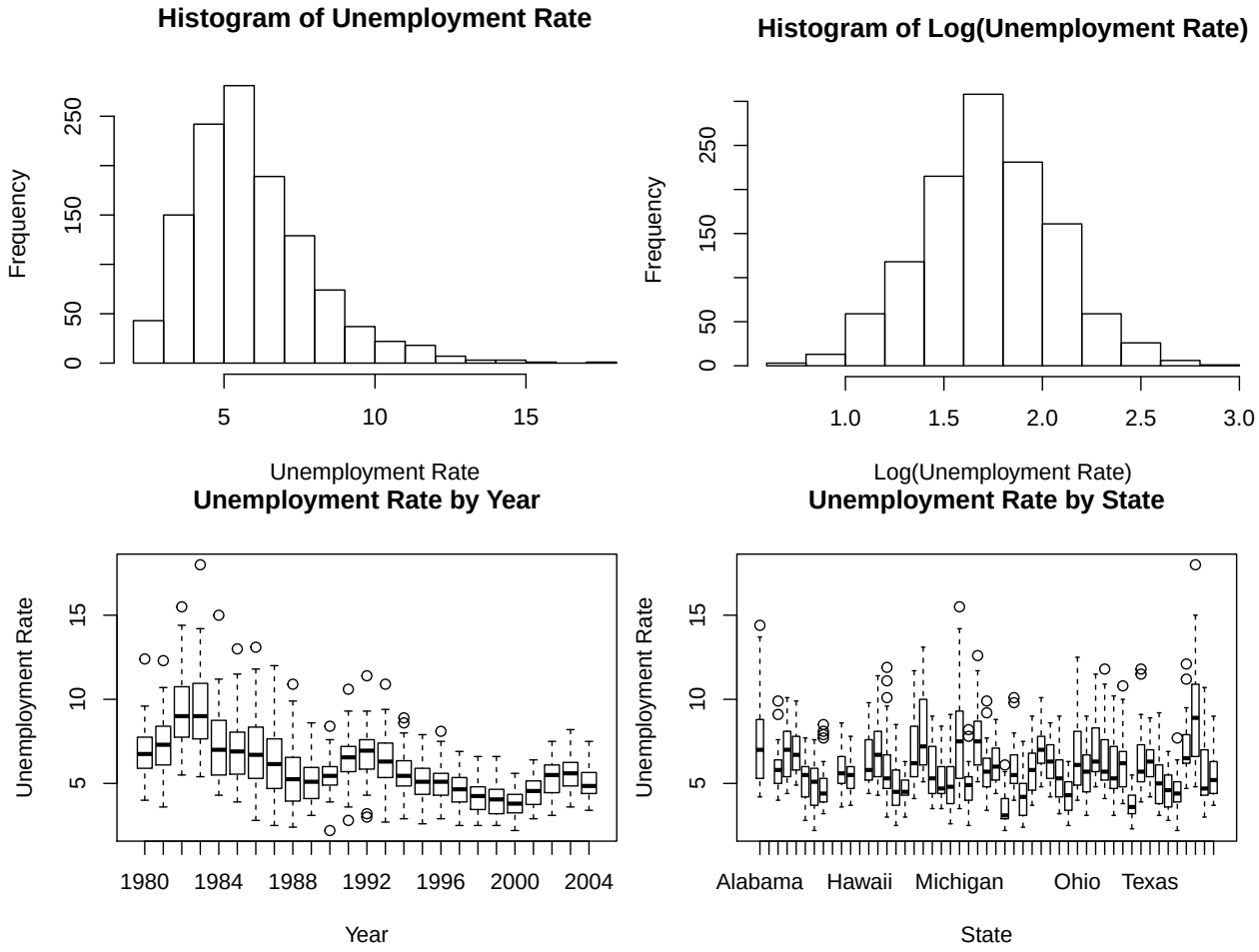


Examining the histogram of *Vehicle Miles*, we note that like state population it is highly skewed. We suspect that it is highly correlated with state population, so we switch our focus to the *Vehicle Miles per Capita* variable. We note an increasing trend in the number of vehicle miles over time, and very different distributions and levels across different states.

Vermont has a very pronounced step up in 2001-2002. Wyoming is consistently higher than all other states.

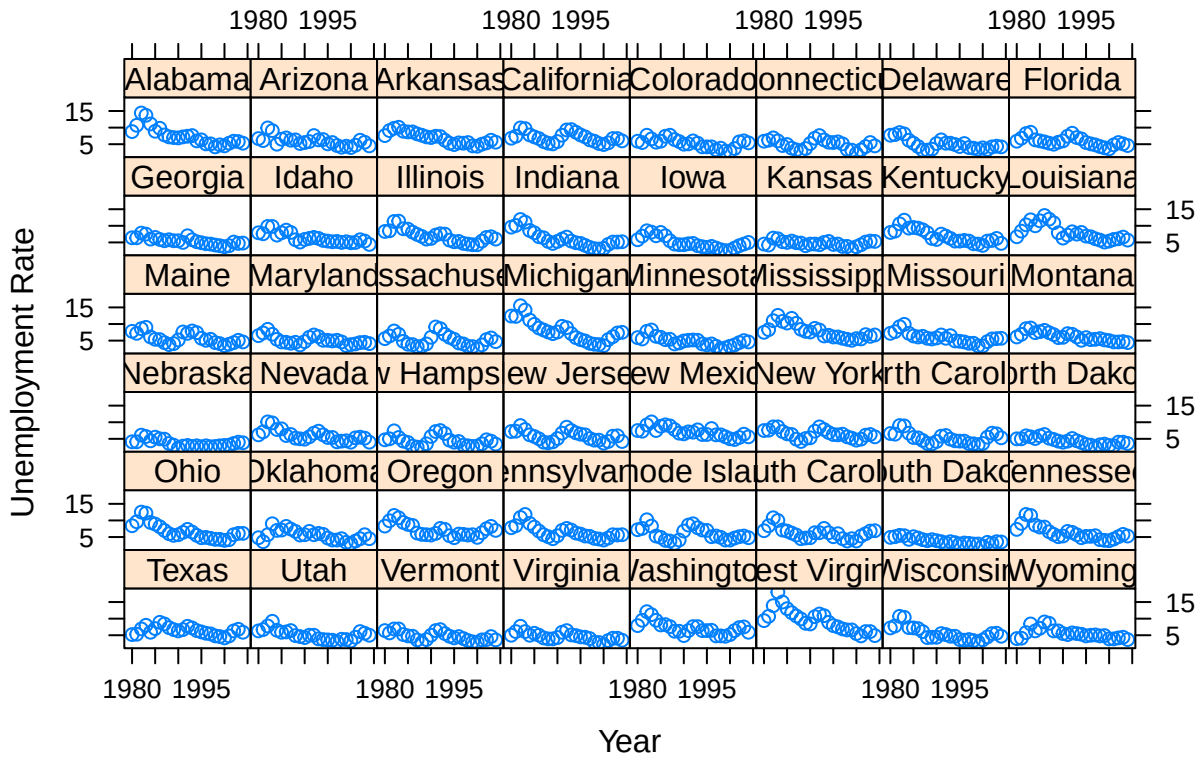
Independent Variable : Unemployment Rate

```
hist(df2$unem, main = "Histogram of Unemployment Rate", xlab = "Unemployment Rate")
hist(log(df2$unem), main = "Histogram of Log(Unemployment Rate)",
    xlab = "Log(Unemployment Rate)")
boxplot(df2$unem ~ df2$year, main = "Unemployment Rate by Year",
    xlab = "Year", ylab = "Unemployment Rate")
boxplot(df2$unem ~ df2$state.name, main = "Unemployment Rate by State",
    xlab = "State", ylab = "Unemployment Rate")
```

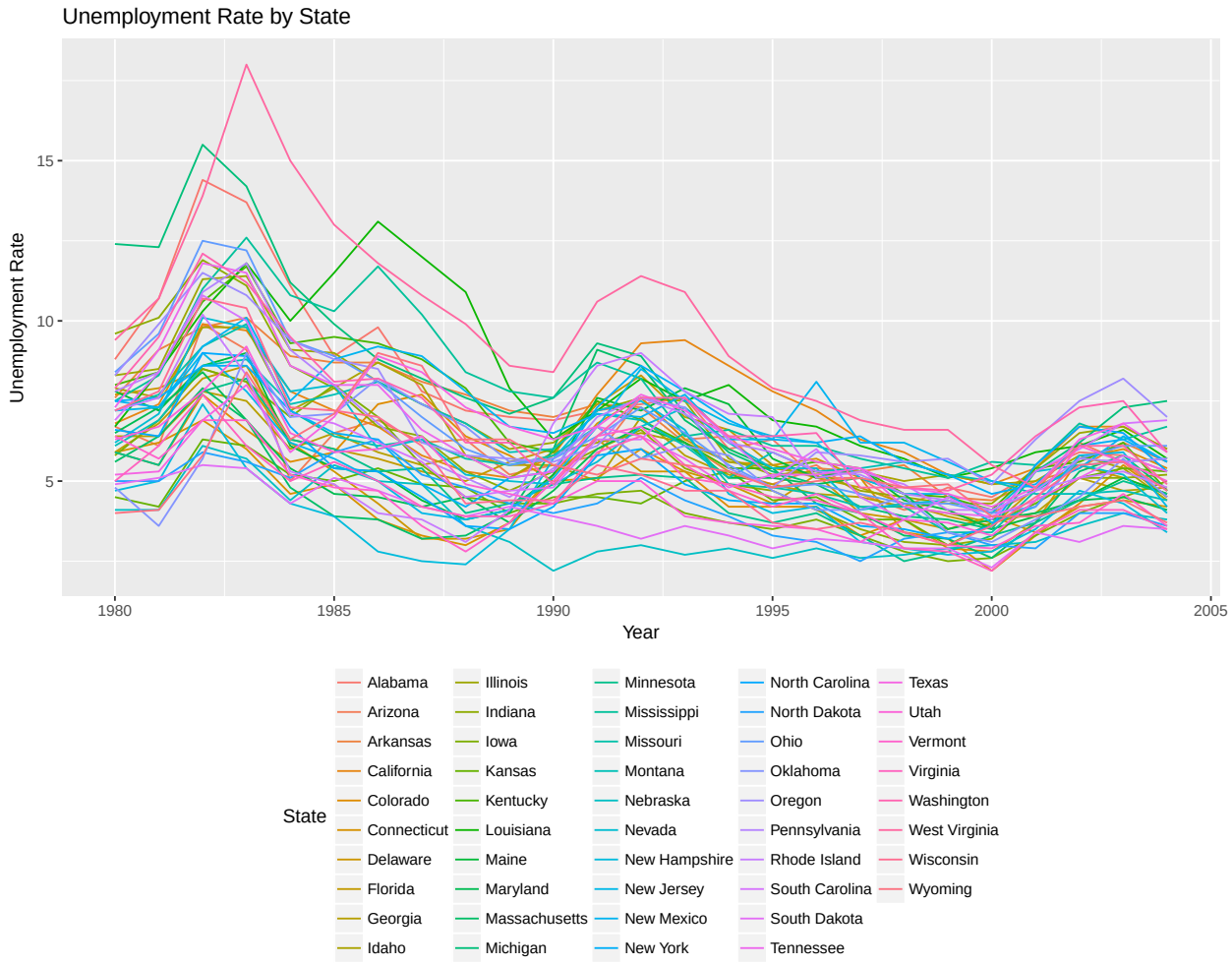


Unemployment rate is noticeably right skewed and we can consider a log transformation if a normal distribution is desired. Across cross-sections (time) from 1980 to 2004, we observe a gradual decrease. The local peaks correspond to the early 1980s, early 1990s and late 2000s recessions. Across panels variance is unequal across states. For instance, West Virginia (Bio-tech), Ohio (Energy and Technologies), Michigan (Automotive) and Alabama (Automotive) shows wider variance while South Dakota (Agriculture), Kansas (Agriculture), Arizona (Manufacturing, Mining) and Nebraska (Agriculture) show much narrower variance and lower means. It appears that unemployment rates in states dependent on agricultural industries are more resilient to economic downturn.

```
lattice::xyplot(unem ~ year | state.name, data = df2, as.table = T,
    xlab = "Year", ylab = "Unemployment Rate", main = "Unemployment Rate by Year")
```



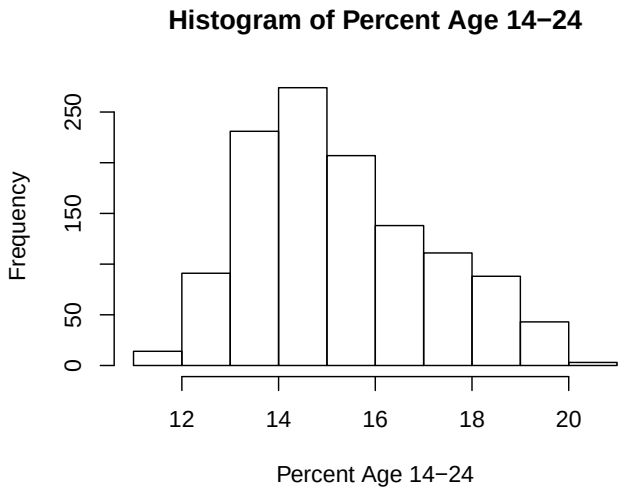

```
ggplot(data = df2, aes(x = year, y = unem, group = state.name,
    colour = state.name)) + geom_line() + ggtitle("Unemployment Rate by State") +
    xlab("Year") + ylab("Unemployment Rate") + labs(color = "State") +
    theme(legend.position = "bottom")
```

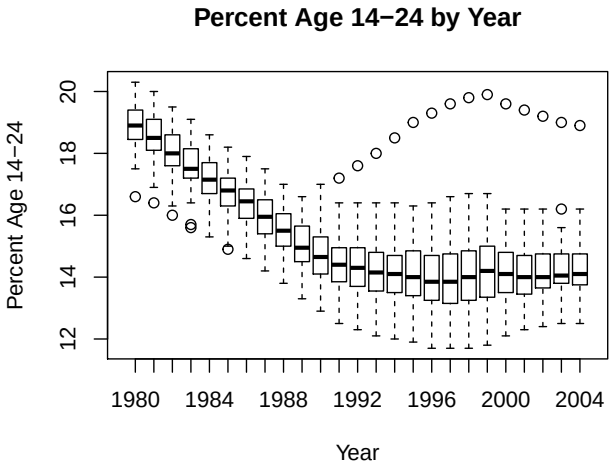


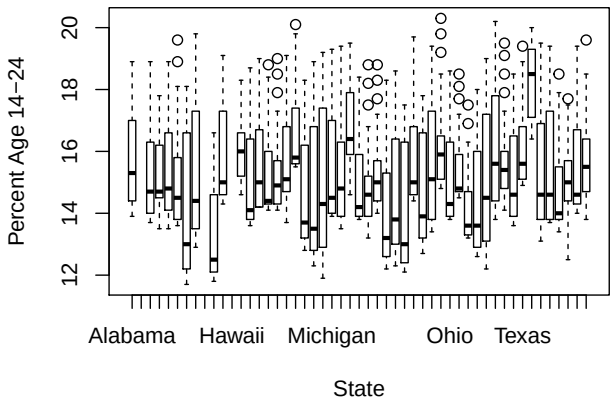
The time series plot shows that most states follow similar trends, with West Virginia clearly on top of most other states with the worst unemployment rates.

Independent Variable : Percent Population Age 14 through 24

```
hist(df2$perc14_24, main = "Histogram of Percent Age 14-24",
    xlab = "Percent Age 14-24")
# hist(log(df2$perc14_24), main='Histogram of Percent Age
# 14-24', xlab='Percent Age 14-24')
boxplot(df2$perc14_24 ~ df2$year, main = "Percent Age 14-24 by Year",
    xlab = "Year", ylab = "Percent Age 14-24")
boxplot(df2$perc14_24 ~ df2$state.name, main = "Percent Age 14-24 by State",
    xlab = "State", ylab = "Percent Age 14-24")
```

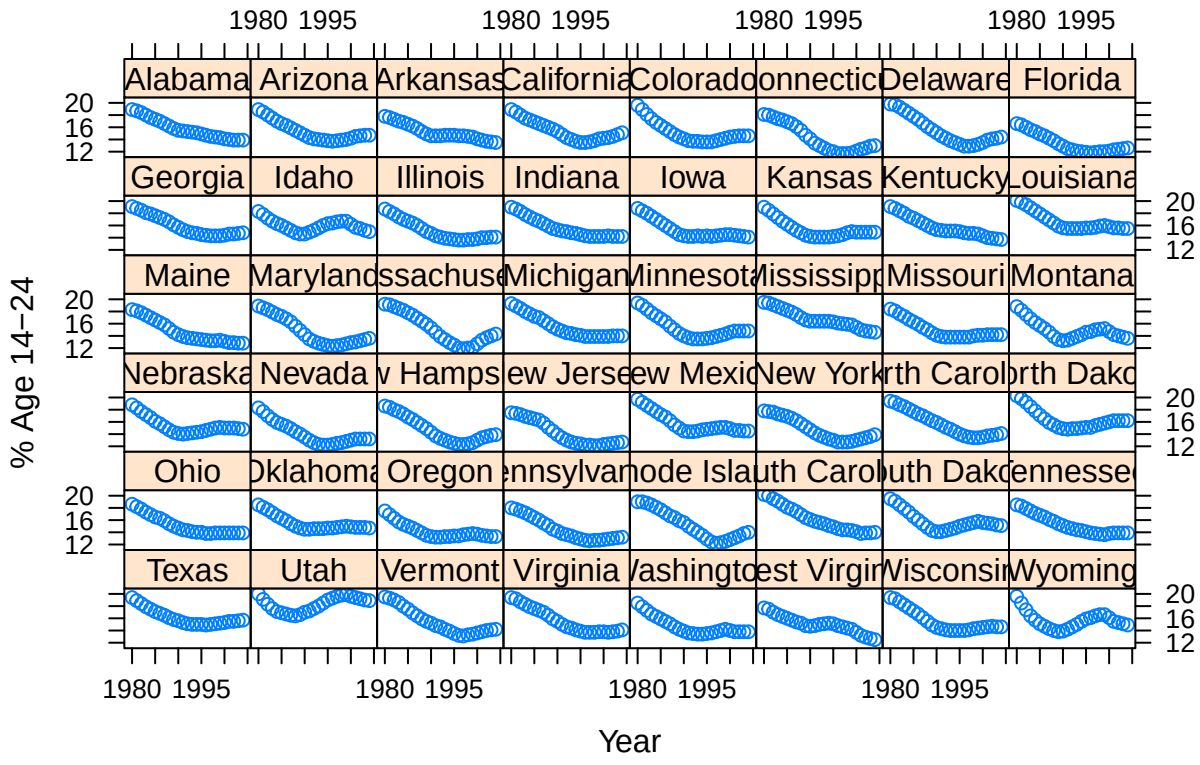




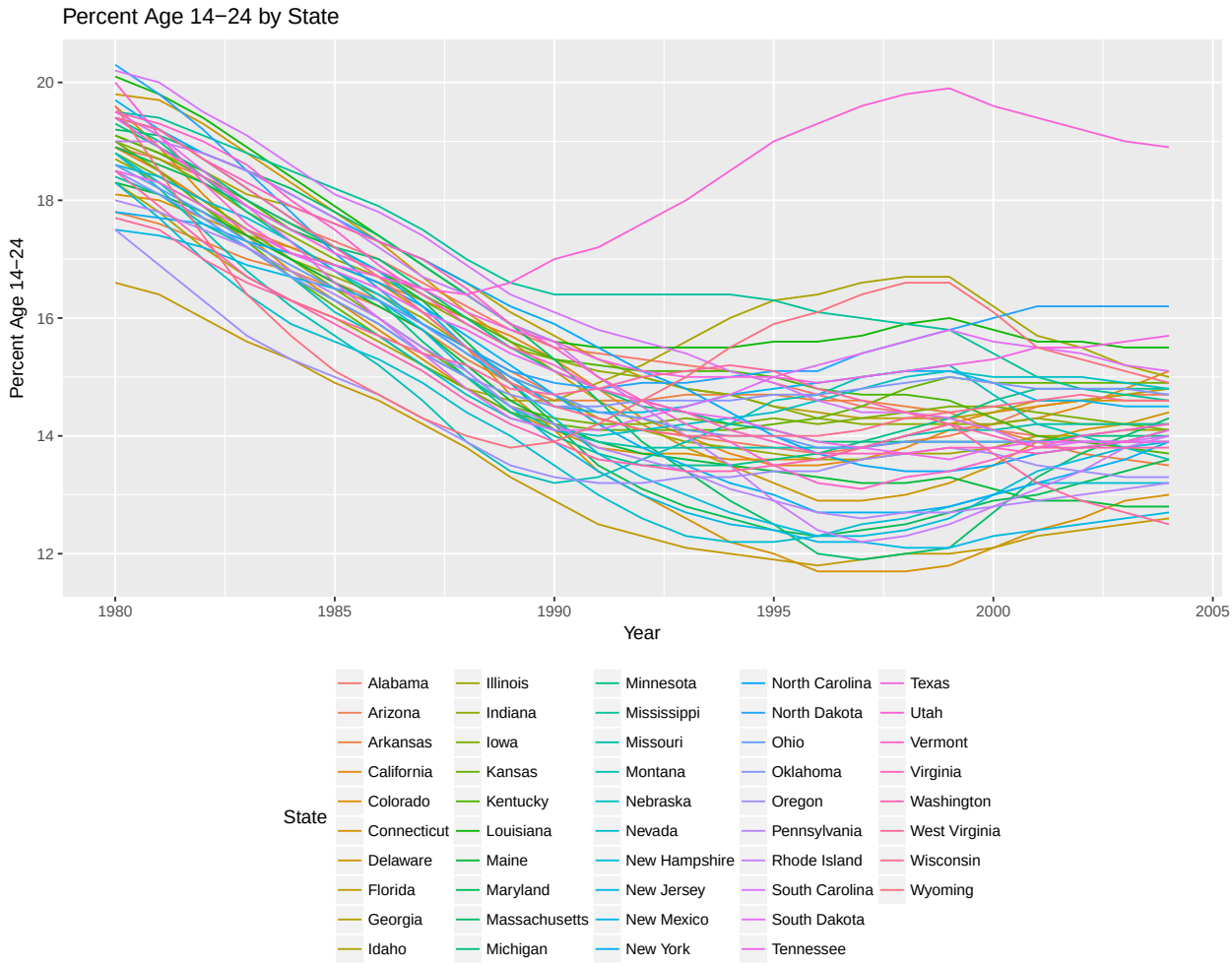


The Percent of the population age 14-24 is mildly right skewed. Across cross-sections from 1980 to 2004, we observe a clear decreasing trend that can be associated with the lowering birth rate since the 70s. Variance scales up in the 1990s. There is a clear outlier with noticeable trend. Across panels variance seems consistent and overlap greatly. Connecticut and Florida has lower means than most states.

```
lattice::xyplot(perc14_24 ~ year | state.name, data = df2, as.table = T,
    xlab = "Year", ylab = "% Age 14-24", main = "% Age 14-24 by Year")
```



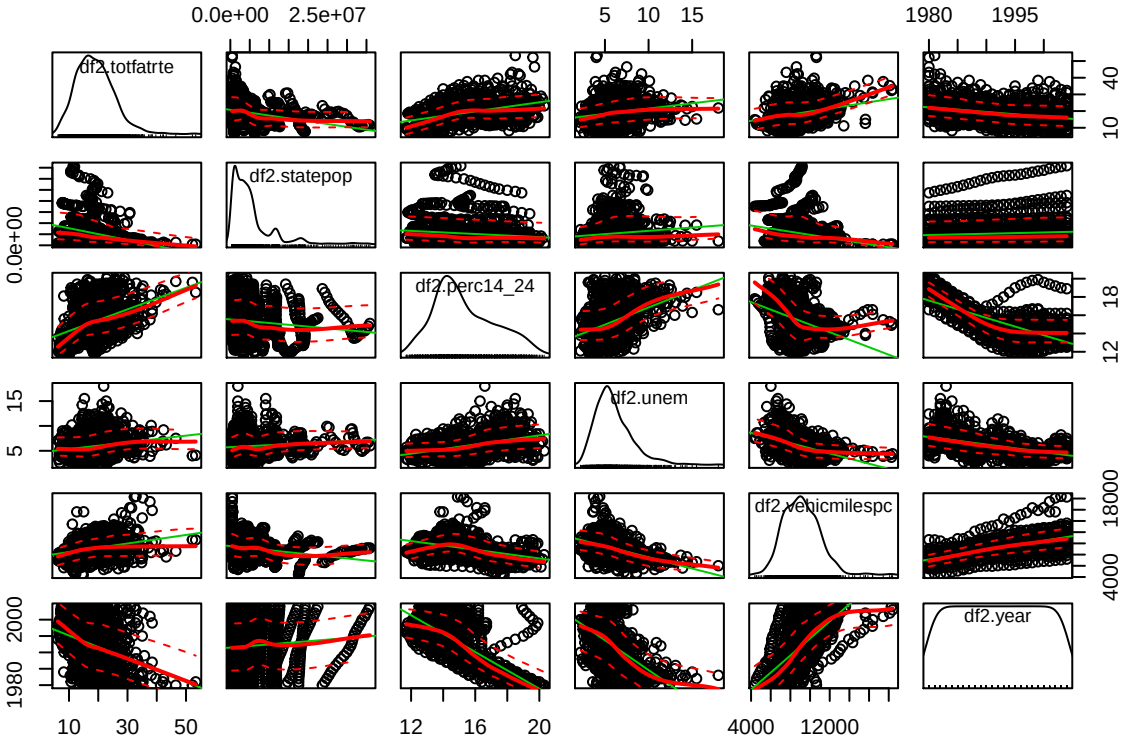

```
ggplot(data = df2, aes(x = year, y = perc14_24, group = state.name,
    colour = state.name)) + geom_line() + ggtitle("Percent Age 14-24 by State") +
    xlab("Year") + ylab("Percent Age 14-24") + labs(color = "State") +
    theme(legend.position = "bottom")
```



The time-series plot above reveals some trend differences across states starting in the late 1980s. States such as Utah and Idaho clearly pick up and peak around the late 1990s, while Florida, Texas and California dipped. This may contradict with the early population trends we saw about Florida, Texas and California, but external sources shows that population increase in these three states mostly come from migration.

Bivariate Correlation Matrix : Correlations between dependent and continuous independent variables

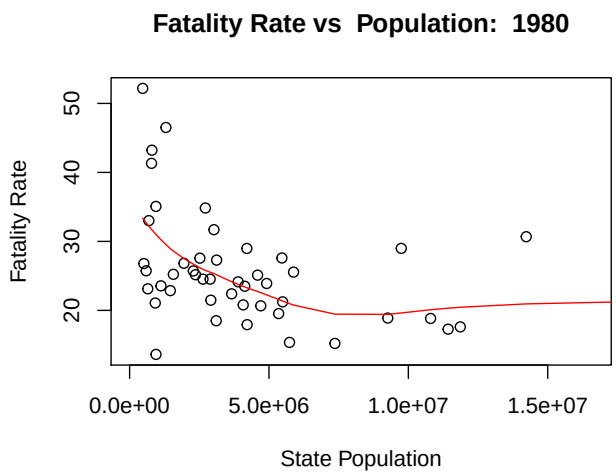
```
scatterplotMatrix(~df2$totfatrte + df2$statepop + df2$perc14_24 +
    df2$unem + df2$vehicmilespc + df2$year)
```

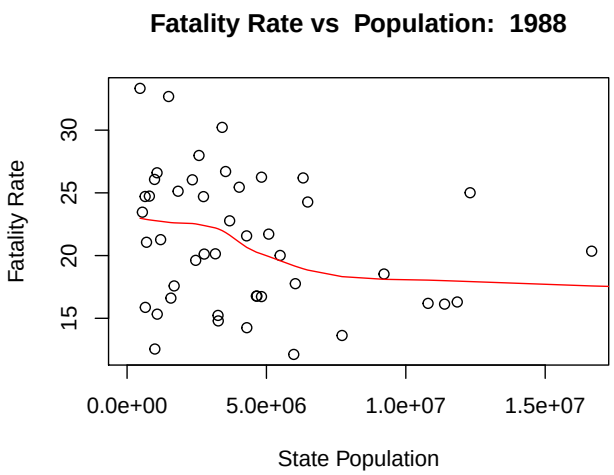


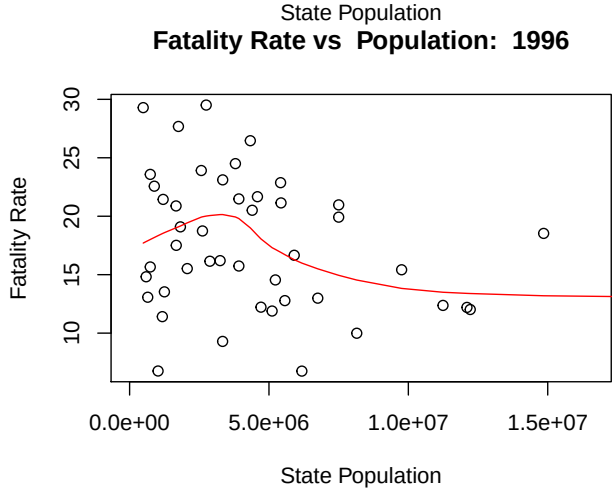
If we treat the dataset as one big cross-section (loses information more precise in panels or time series), *totfatre* has noticeable uphill or downhill relationships with the continuous independent variables in general. However, this view is not entirely appropriate to conclude bivariate relationships. Obvious time trends in both our dependent and continuous independent variables will cause errors to be serially correlated using pooled OLS thus it is inappropriate.

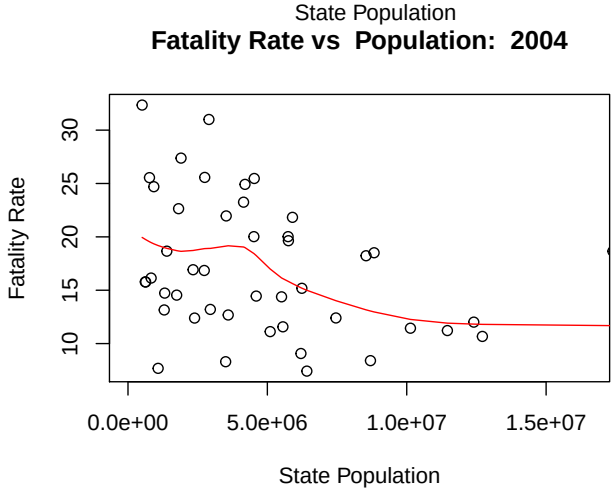
Bivariate Scatter plots by cross-sections: Correlations between dependent and continuous independent variables

```
# scatterplot by year : totfatrte vs statepop

for (i in c(1980, 1988, 1996, 2004)) {
    df.tmp = df2[df2$year == i, ]
    plot(df.tmp$totfatrte ~ df.tmp$statepop, xlab = "State Population",
        ylab = "Fatality Rate", xlim = c(0, mean(df2$statepop) +
            2 * sd(df2$statepop)))
    # Loess curve
    order.pred <- order(df.tmp$statepop)
    smooth.stand <- loess(formula = df.tmp$totfatrte ~ df.tmp$statepop,
        weights = rep(1, 48))
    lines(x = df.tmp$statepop[order.pred], y = predict(smooth.stand)[order.pred],
        lty = "solid", col = "red")
    title(paste("Fatality Rate vs  Population: ", toString(i),
        sep = " "))
}
```

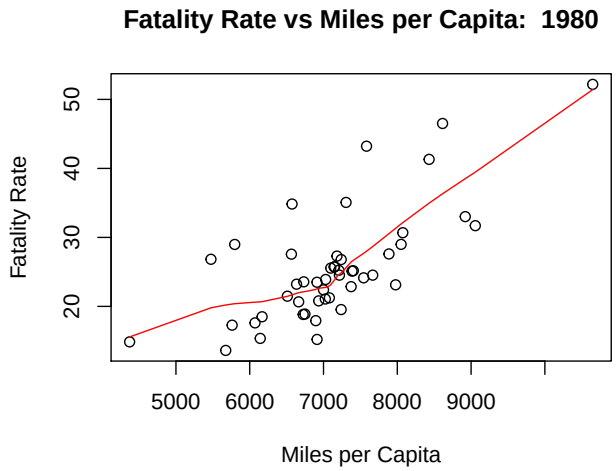


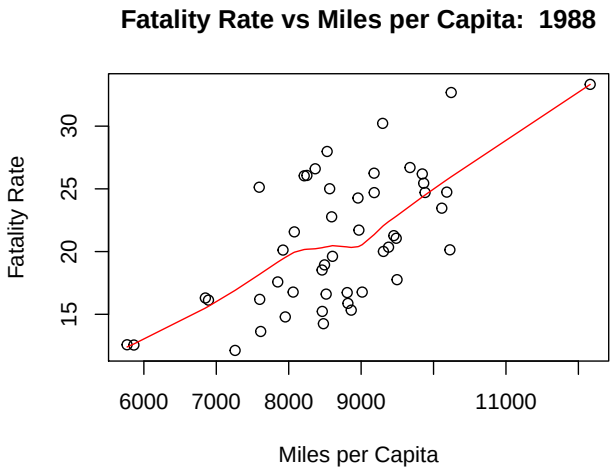


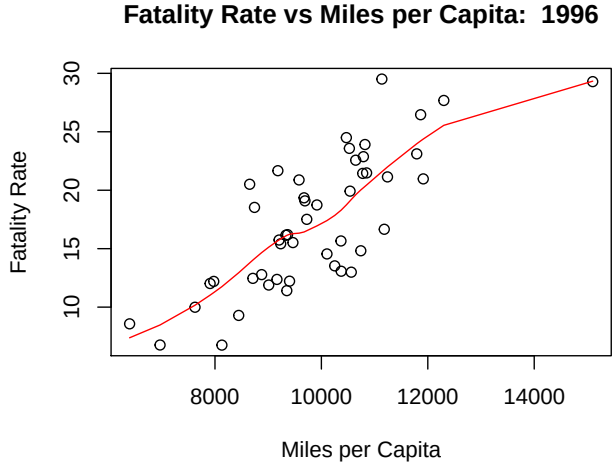


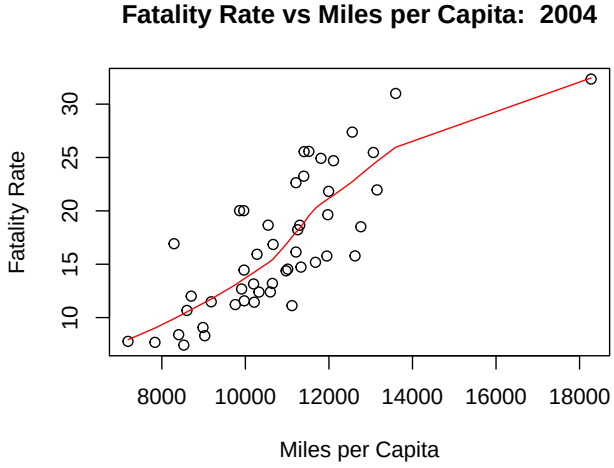



```
# scatterplot by year : totfatrte vs unem

for (i in c(1980, 1988, 1996, 2004)) {
    df.tmp = df2[df2$year == i, ]
    plot(df.tmp$totfatrte ~ df.tmp$vehicmilespc, xlab = "Miles per Capita",
        ylab = "Fatality Rate")
    # Loess curve
    order.pred <- order(df.tmp$vehicmilespc)
    smooth.stand <- loess(formula = df.tmp$totfatrte ~ df.tmp$vehicmilespc,
        weights = rep(1, 48))
    lines(x = df.tmp$vehicmilespc[order.pred], y = predict(smooth.stand)[order.pred],
        lty = "solid", col = "red")
    title(paste("Fatality Rate vs Miles per Capita: ", toString(i),
        sep = " "))
}
```

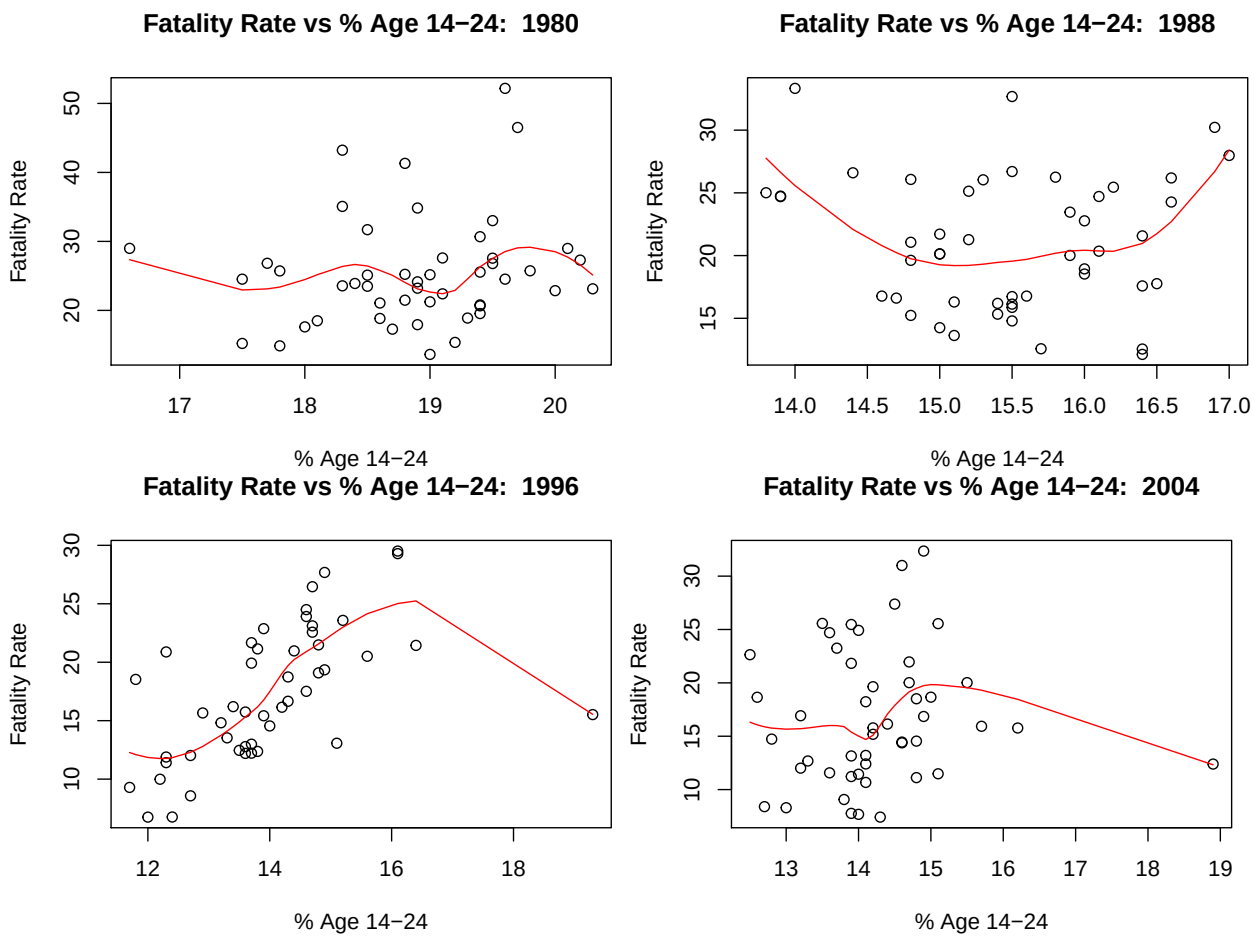









```
# scatterplot by year : totfatrte vs perc14_24

for (i in c(1980, 1988, 1996, 2004)) {
    df.tmp = df2[df2$year == i, ]
    plot(df.tmp$totfatrte ~ df.tmp$perc14_24, xlab = "% Age 14-24",
        ylab = "Fatality Rate")
    # Loess curve
    order.pred <- order(df.tmp$perc14_24)
    smooth.stand <- loess(formula = df.tmp$totfatrte ~ df.tmp$perc14_24,
        weights = rep(1, 48))
    lines(x = df.tmp$perc14_24[order.pred], y = predict(smooth.stand)[order.pred],
        lty = "solid", col = "red")
    title(paste("Fatality Rate vs % Age 14-24: ", toString(i),
        sep = " "))
}
```

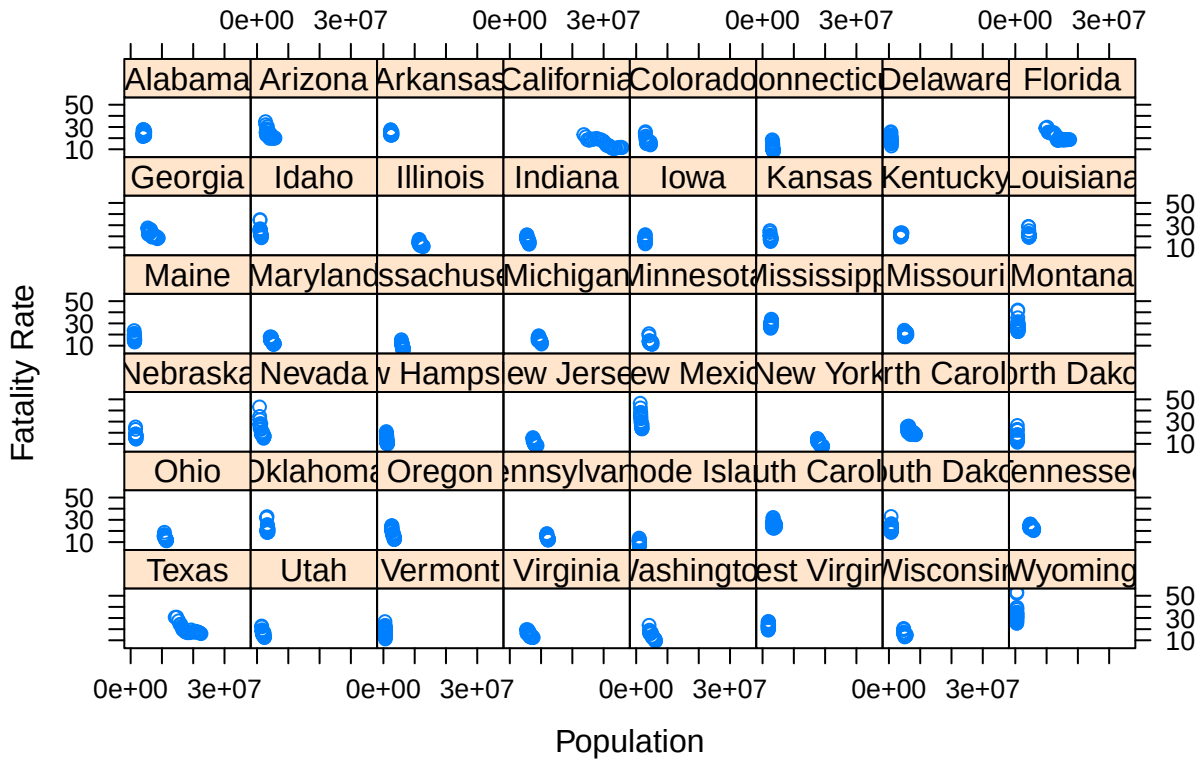



```
# scatterplot by year : totfatrte vs unem

for (i in c(1980, 1988, 1996, 2004)) {
    df.tmp = df2[df2$year == i, ]
    plot(df.tmp$totfatrte ~ df.tmp$unem, xlab = "Unemployment Rate",
        ylab = "Fatality Rate")
    # Loess curve
    order.pred <- order(df.tmp$unem)
    smooth.stand <- loess(formula = df.tmp$totfatrte ~ df.tmp$unem,
        weights = rep(1, 48))
    lines(x = df.tmp$unem[order.pred], y = predict(smooth.stand)[order.pred],
        lty = "solid", col = "red")
    title(paste("Fatality Rate vs Unemployment: ", toString(i),
        sep = " "))
}
```

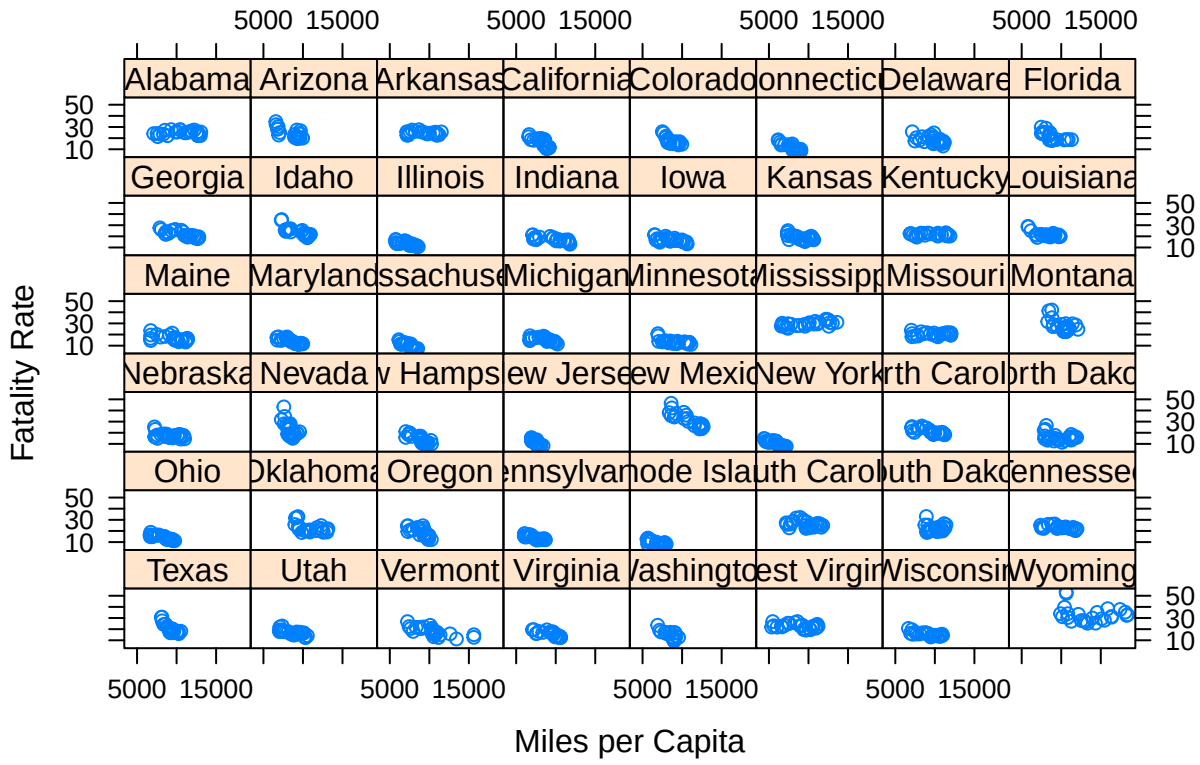
Examining the above plots, it is clear that bivariate relationships of *perc14_24* and *unem* against *totfatrte* are quite different at different time periods. Some periods show noticeable uphill or downhill relationships while the others are rather flat. For instance, *perc14_24* is only weakly correlated with *totfatrte* in 1980 but strongly so in 1996. We can consider interacting the variables to allow for different slopes for the variables across time. On the other hand, relationship between *vehicmilespc* and *unem* against *totfatrte* seem relatively stable across time.

Bivariate Scatter plots by panels: Correlations between dependent and continuous independent variables

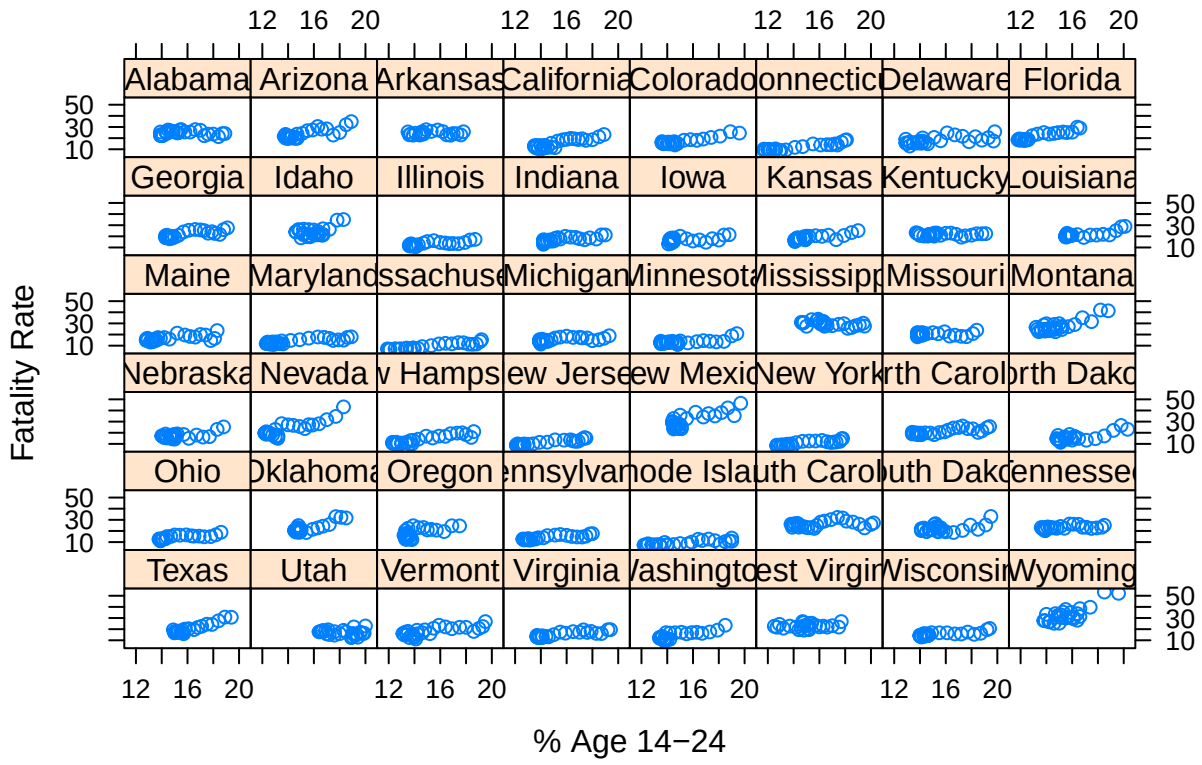
```
lattice::xyplot(totfatrte ~ statepop | state.name, data = df2,
    as.table = T, xlab = "Population", ylab = "Fatality Rate",
    main = "Population vs Fatality Rate by State")
```



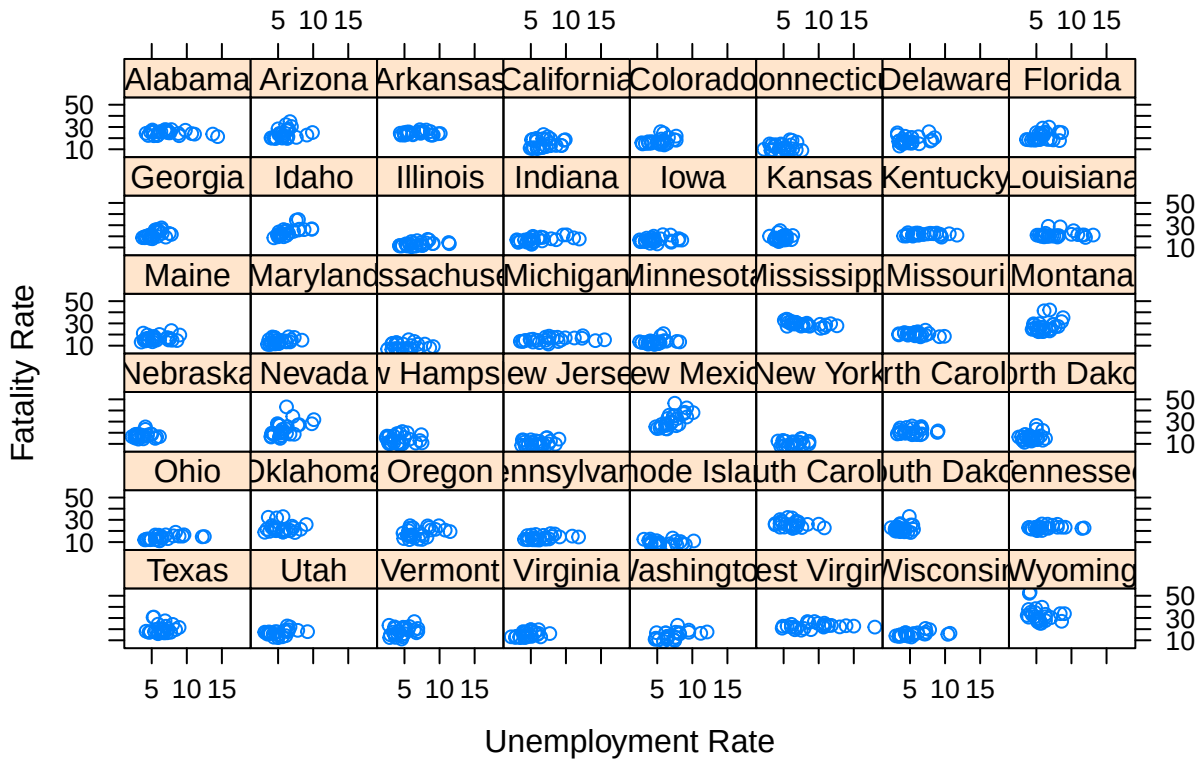

```
lattice::xyplot(totfatrte ~ vehicmilespc | state.name, data = df2,
    as.table = T, xlab = "Miles per Capita", ylab = "Fatality Rate",
    main = "Miles per Capita vs Fatality Rate by State")
```




```
lattice::xyplot(totfatrte ~ perc14_24 | state.name, data = df2,
    as.table = T, xlab = "% Age 14-24", ylab = "Fatality Rate",
    main = "% Age 14-24 vs Fatality Rate by State")
```



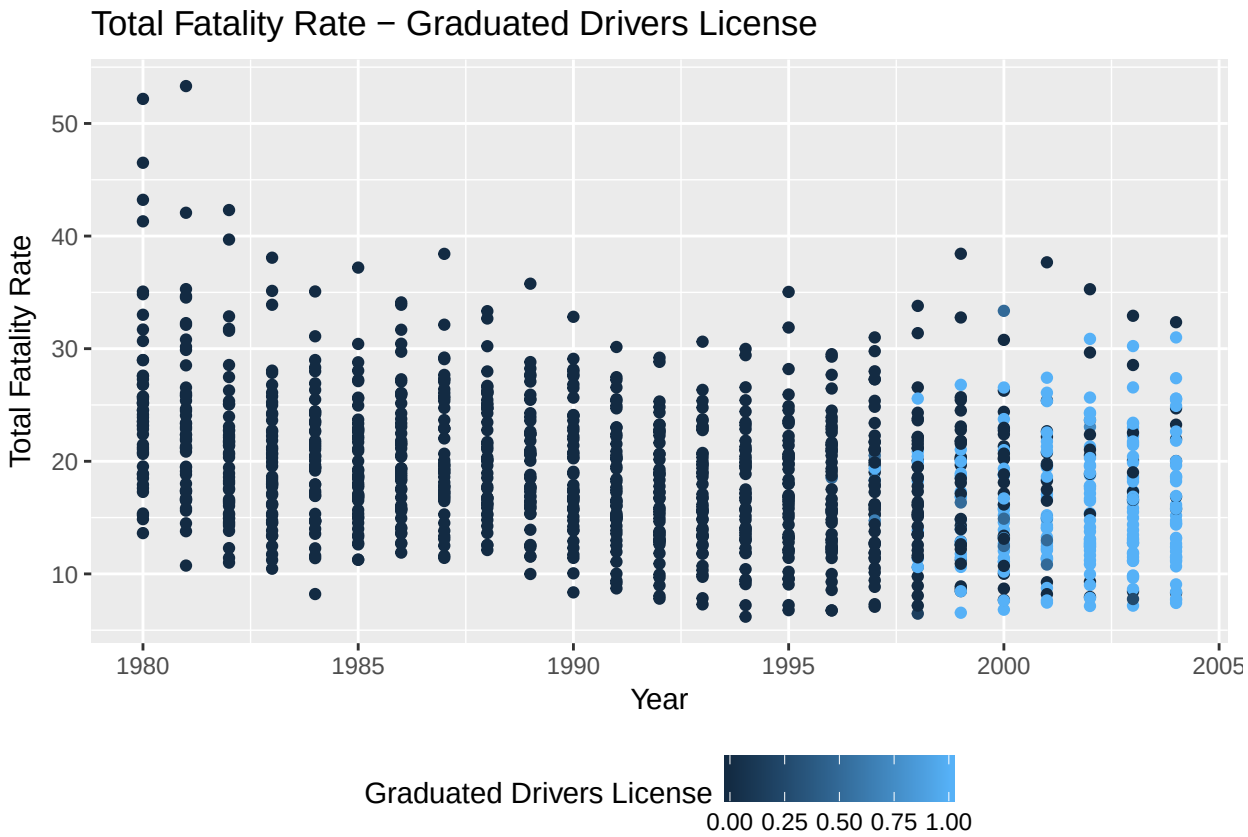

```
lattice::xyplot(totfatrte ~ unem | state.name, data = df2, as.table = T,
    xlab = "Unemployment Rate", ylab = "Fatality Rate", main = "Unemployment Rate vs Fatality Rate by S
```



Examining the above plots by panel, the noticeable variations of *totfatrte* corresponding to *perc12_24* says that there exists state fixed effects which affects how the independent variable correlates with the dependent variable. Because such fixed/unobserved effects exist, a random effects model may not be appropriate.

Bivariate Time Plot : Graduated Drivers License Law and Total Fatality Rate (Indicator variable)

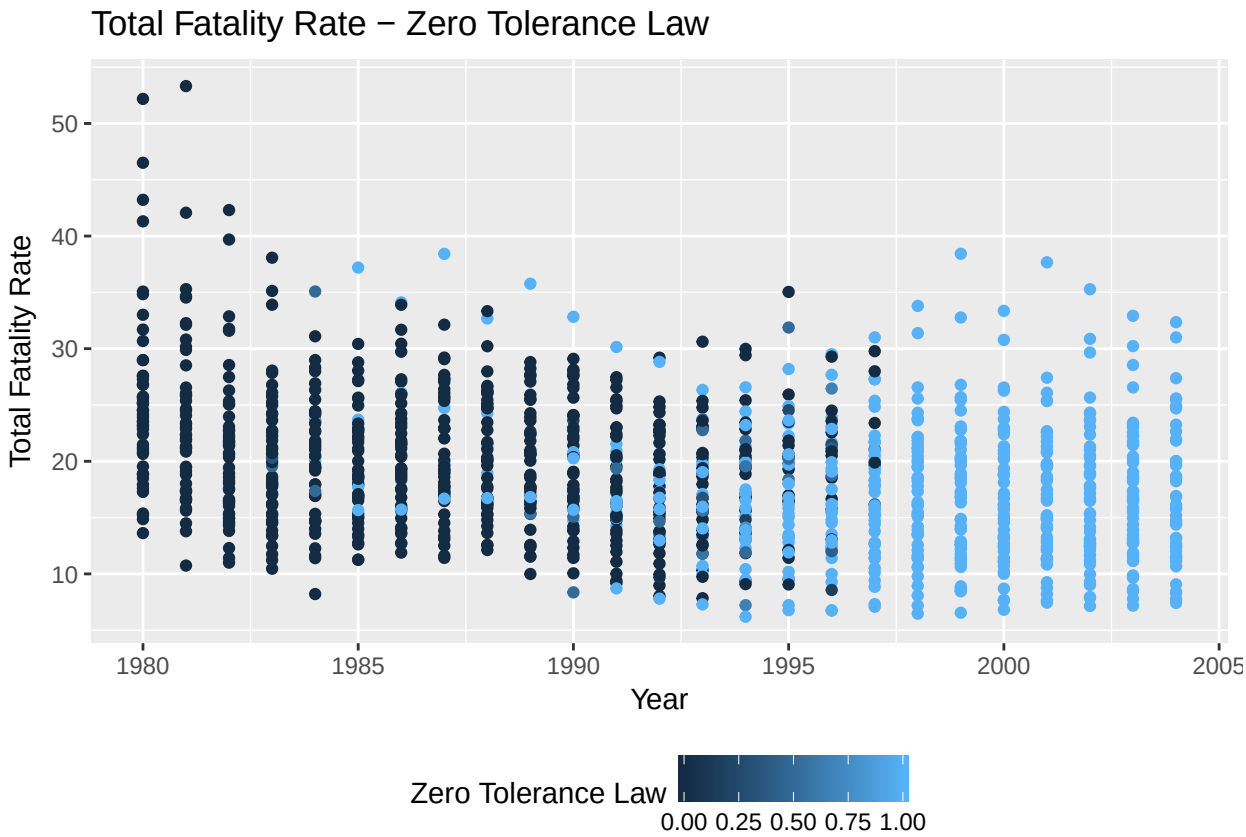
```
ggplot(data = df2, aes(x = year, y = totfatrte, group = gdl,
    colour = gdl)) + geom_point() + ggtitle("Total Fatality Rate - Graduated Drivers License") +
    xlab("Year") + ylab("Total Fatality Rate") + labs(color = "Graduated Drivers License") +
    theme(legend.position = "bottom")
```



It appears that the graduated driver's license law has been increasingly adopted since 1998. However, the variance and distribution of total fatality rate remained consistent since 1995.

Bivariate Time Plot : Zero Tolerance Law and Total Fatality Rate (Indicator variable)

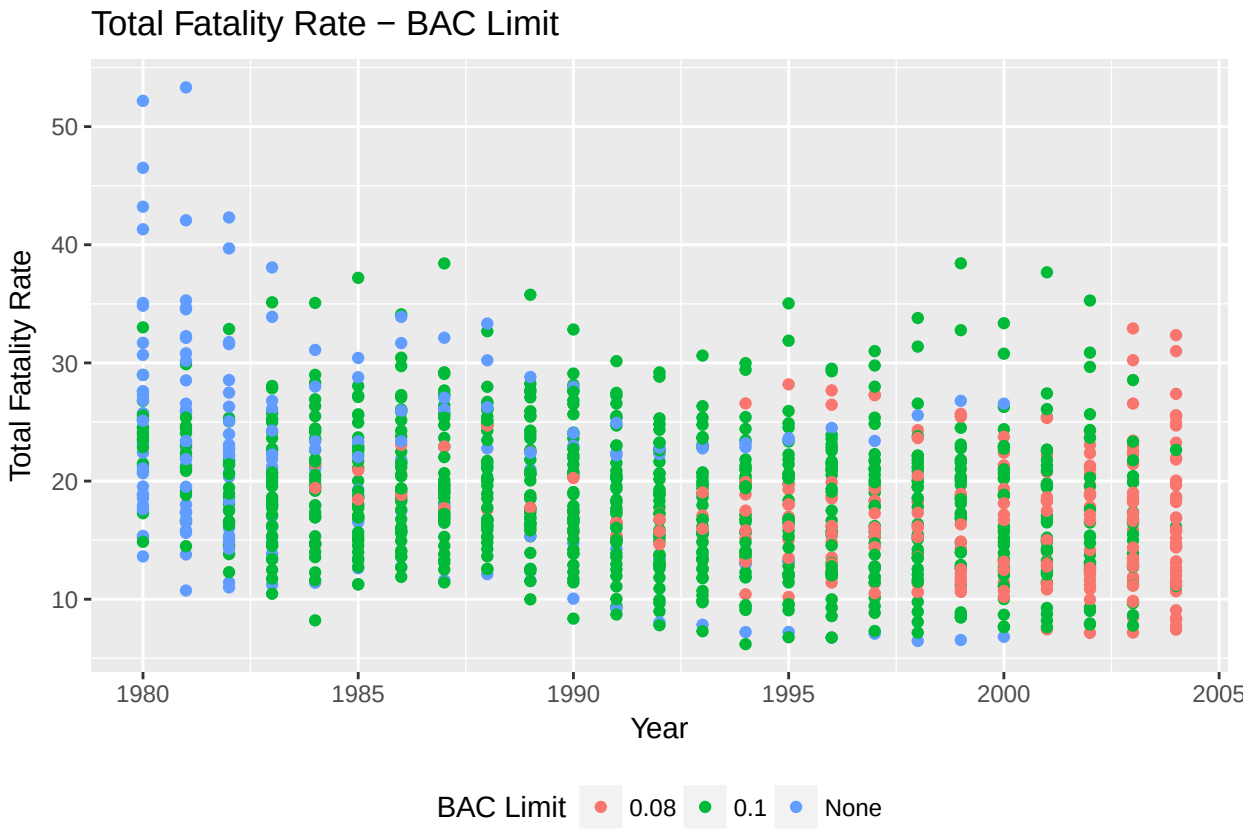
```
ggplot(data = df2, aes(x = year, y = totfatrte, group = zerotol,
    colour = zerotol)) + geom_point() + ggtitle("Total Fatality Rate - Zero Tolerance Law") +
    xlab("Year") + ylab("Total Fatality Rate") + labs(color = "Zero Tolerance Law") +
    theme(legend.position = "bottom")
```



Similar to graduated drivers license law, the zero tolerance law was widely adopted at a time when the variance, distribution and mean of total fatality rates have already stabilized.

Bivariate Time Plot : BAC limit and Total Fatality Rate (Indicator variable)

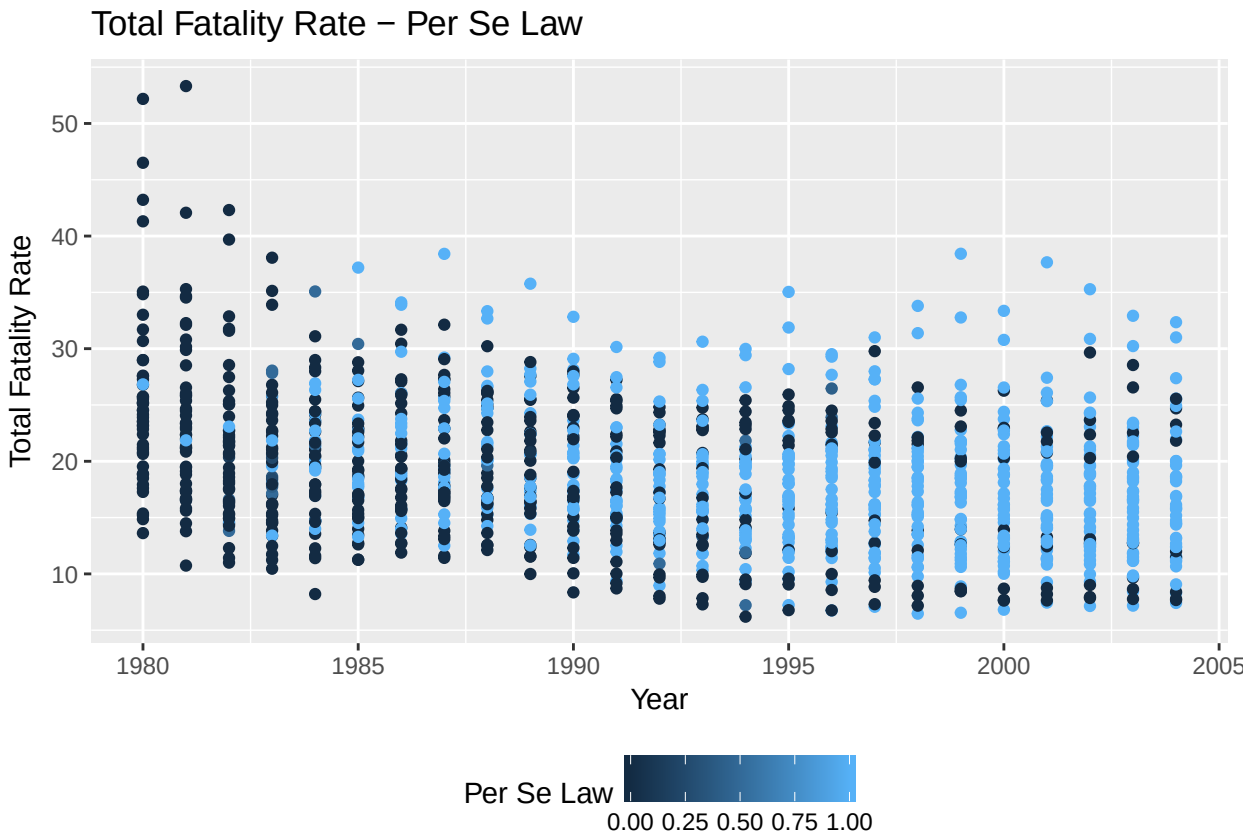
```
ggplot(data = df2, aes(x = year, y = totfatrte, group = baclim,
    colour = baclim)) + geom_point() + ggtitle("Total Fatality Rate - BAC Limit") +
    xlab("Year") + ylab("Total Fatality Rate") + labs(color = "BAC Limit") +
    theme(legend.position = "bottom")
```



We see that the blood alcohol limit of 0.10 has been enforced by most state since 1982, when the decreasing trend in total fatality rate started. This variable may have strong explanatory power on our dependent variable.

Bivariate Time Plot : Per Se Law and Total Fatality Rate (Indicator variable)

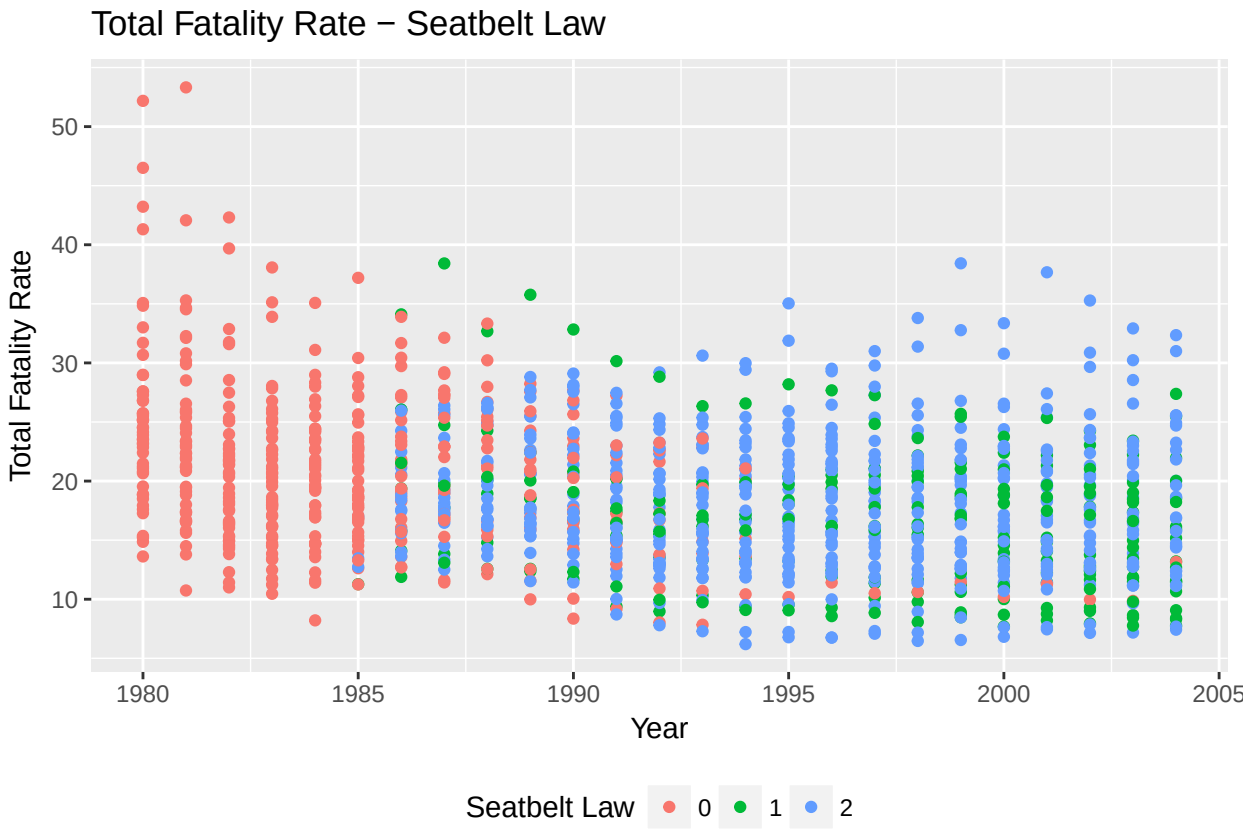
```
ggplot(data = df2, aes(x = year, y = totfatrte, group = perse,
    colour = perse)) + geom_point() + ggtitle("Total Fatality Rate - Per Se Law") +
    xlab("Year") + ylab("Total Fatality Rate") + labs(color = "Per Se Law") +
    theme(legend.position = "bottom")
```



Per se laws have seen increasing adoption at a steady rate since the early 1980s, compared with the overall gradual decreasing trend of *totfatrte*, this variable may have some explanatory power.

Bivariate Time Plot : Seatbelt Law and Total Fatality Rate (Indicator variable)

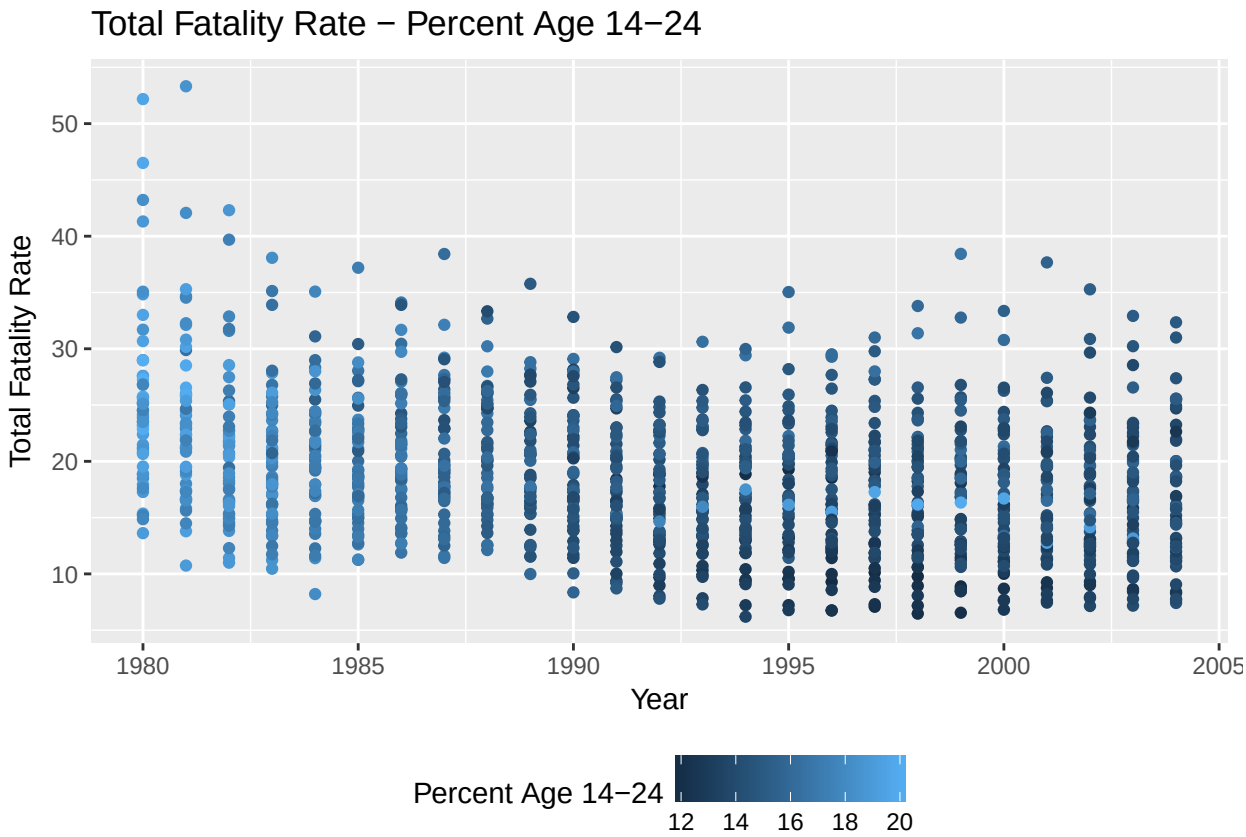
```
ggplot(data = df2, aes(x = year, y = totfatrte, group = as.factor(seatbelt),
    colour = as.factor(seatbelt))) + geom_point() + ggtitle("Total Fatality Rate - Seatbelt Law") +
    xlab("Year") + ylab("Total Fatality Rate") + labs(color = "Seatbelt Law") +
    theme(legend.position = "bottom")
```



The seatbelt laws went from almost no adoption in 1985 to almost full adoption in 1995. This period coincides with most of the decreasing trend in *totfatrte*. Judging from the pattern, *sbprim* (primary seatbelt law) should have strong explanatory power.

Bivariate Time Plot : % Age 14-24 and Total Fatality Rate

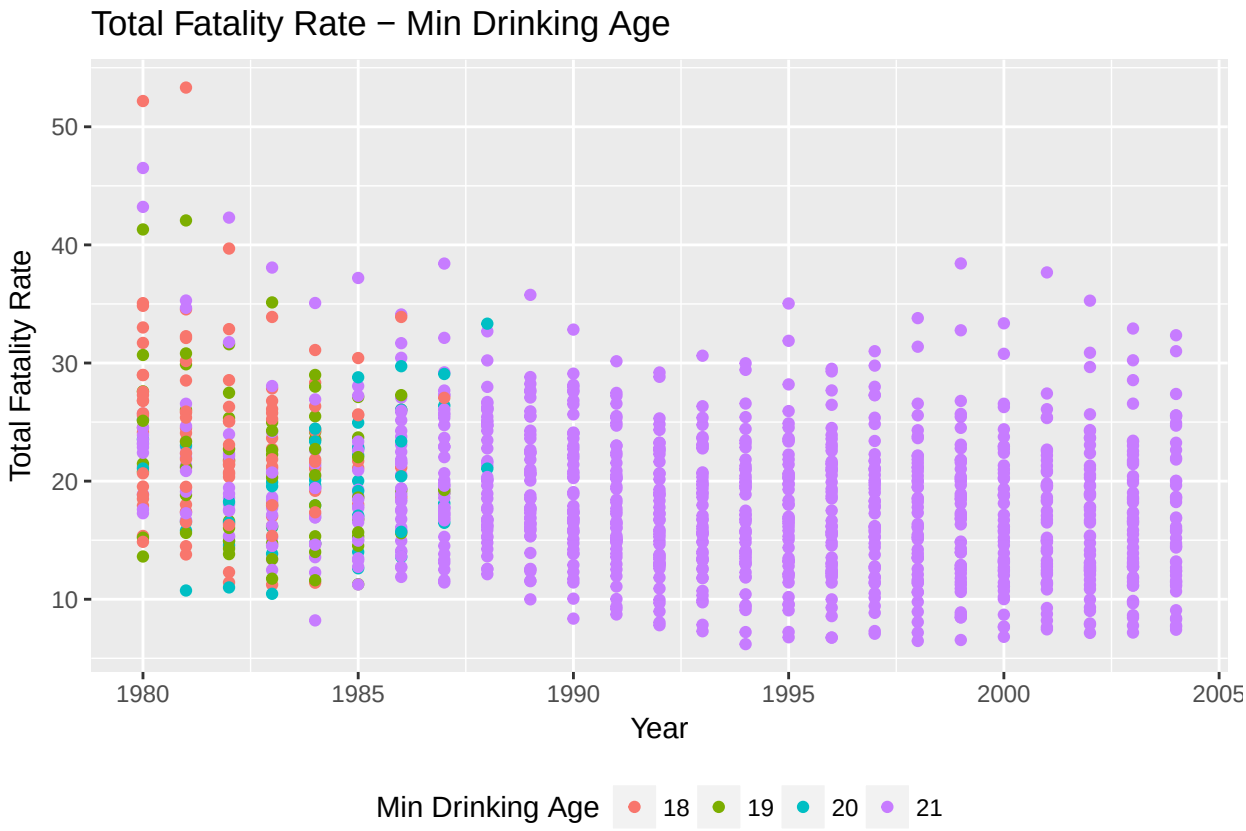
```
ggplot(data = df2, aes(x = year, y = totfatrte, group = perc14_24,
    colour = perc14_24)) + geom_point() + ggtitle("Total Fatality Rate - Percent Age 14-24") +
    xlab("Year") + ylab("Total Fatality Rate") + labs(color = "Percent Age 14-24") +
    theme(legend.position = "bottom")
```



The percent of the population aged 14 through 24 decreased with a steady trend for most states up until 1990. This period coincides with most of the decreasing trend in *totfatrte*. This variable should have strong explanatory power.

Bivariate Time Plot : Minimum Drinking Age and Total Fatality Rate (Indicator variable)

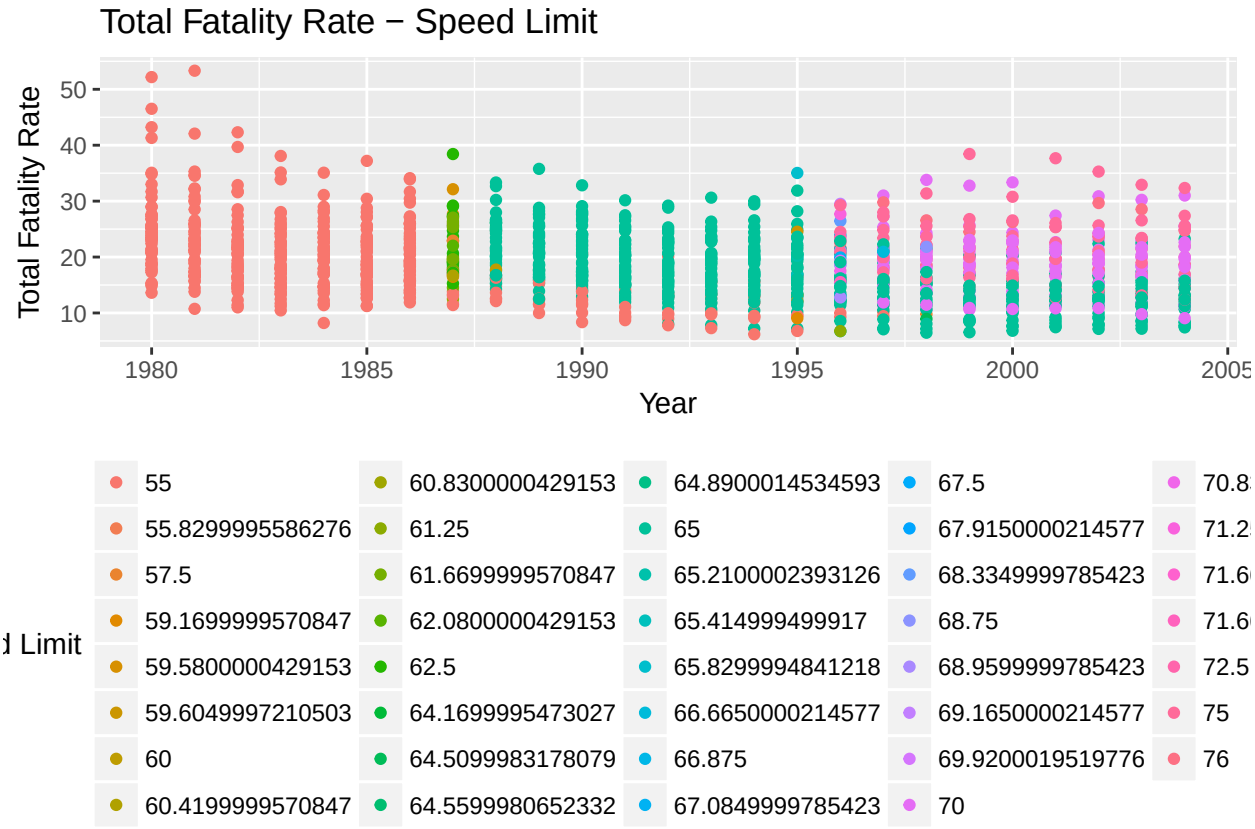
```
ggplot(data = df2, aes(x = year, y = totfatrte, group = as.factor(round(minage)),
    colour = as.factor(round(minage)))) + geom_point() + ggtitle("Total Fatality Rate - Min Drinking Ag
    xlab("Year") + ylab("Total Fatality Rate") + labs(color = "Min Drinking Age") +
    theme(legend.position = "bottom")
```



It appears that most states pushed their minimum drinking age from 18 to 21 between 1980 and 1985 when the trend of *totfatrte* was mildly decreasing but also mildly concave. This happened before the trend of *totfatrte* showed a clear decline between 1985 and 1990, therefore *minage* is expected to have some explanatory power but not too strong.

Bivariate Time Plot : Speed limit (Indicator variable)

```
speedlimit = df2$slnone * 0 + df2$sl55 * 55 + df2$sl65 * 65 +
    df2$sl70 * 70 + df2$sl75 * 75 + as.numeric(df2$sl70plus ==
    1 & (df2$sl70 == 0 & df2$sl75 == 0)) * 76

spl = as.factor(speedlimit)

ggplot(data = df2, aes(x = year, y = totfatrte, group = spl,
    colour = spl)) + geom_point() + ggtitle("Total Fatality Rate - Speed Limit") +
    xlab("Year") + ylab("Total Fatality Rate") + labs(color = "Speed Limit") +
    theme(legend.position = "bottom")
```

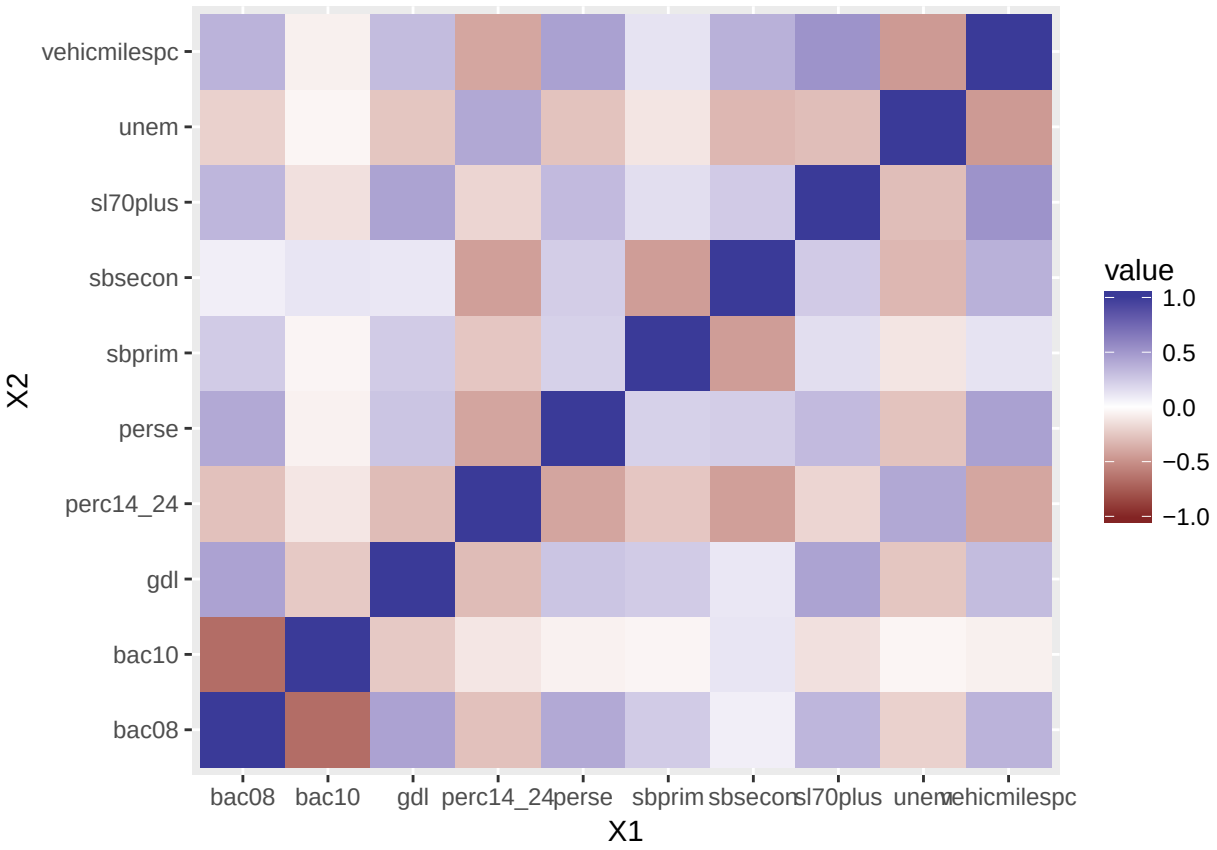


(* legend of 76 refers to speed limit higher than 75mph)

It appears that up until 1986, all states have the same speed limit of 55mph, then the limits gradually diverged. Later in the time spectrum, speed limit generally increased for most states. It is unintuitive that higher speed limit causes lower total fatality rate. We shall revisit this if the coefficient estimates for *sl70plus* for question 4 is significant.

Bivariate Heat map : Correlations between independent variables

```
data_filtered = df2[complete.cases(df2), c("bac08", "bac10",
    "perse", "sbprim", "sbsecon", "sl70plus", "gdl", "perc14_24",
    "unem", "vehicmilespc")]
g = qplot(x = X1, y = X2, data = melt(cor(data_filtered)), fill = value,
    geom = "tile") + scale_fill_gradient2(limits = c(-1, 1))
print(g)
```



Again, if we treat the dataset as one big cross-section, it appears that the following correlations among independent variables can raise standard errors of our estimates:

- bac08 and bac10 : moderate negative correlation, a state cannot adopt both but can adopt neither.
- sbprim and sbsecon : moderate negative correlation, a state cannot adopt both but can adopt neither.
- Within each pair the variables are not strictly mutually exclusive but we can consider combining each pair into one variable if it's necessary to get more precise estimates. This combination would indicate whether any BAC limit or any seatbelt law was enforced.

In summary of the EDA, we observe that:

For continuous variables:

- Relationships between our dependent variable and certain continuous variables (*unem* and *perc14_24*) vary across cross-sections and panels (time and individual independent).
- *statepop* doesn't seem to have a strong effect on the dependent variable
- *vehicmilespc* has a clear positive correlation with *totfatrte*.
- Certain variables are noticeably right skewed enough for transformation; further discussion followed in question 3.

For indicator variables: - *perc14_24*, *sbprim*, *bac08*, and *bac10* may have strong explanatory power

- *perse* and *minage* may have some explanatory power
- *gdl* and *zerotol* may have little or no explanatory power
- We can consider combining *bac08* and *bac10*, and/or *sbprim* and *sbsecon* into one variable based on their moderate correlations.

Since many of the variables exhibit time dependent trends, pooled OLS is clearly inappropriate. Since some bivariate relationships are state dependent, fixed effect models are likely more appropriate. Since bivariate relationships vary across time, we may consider interacting some variables with time.

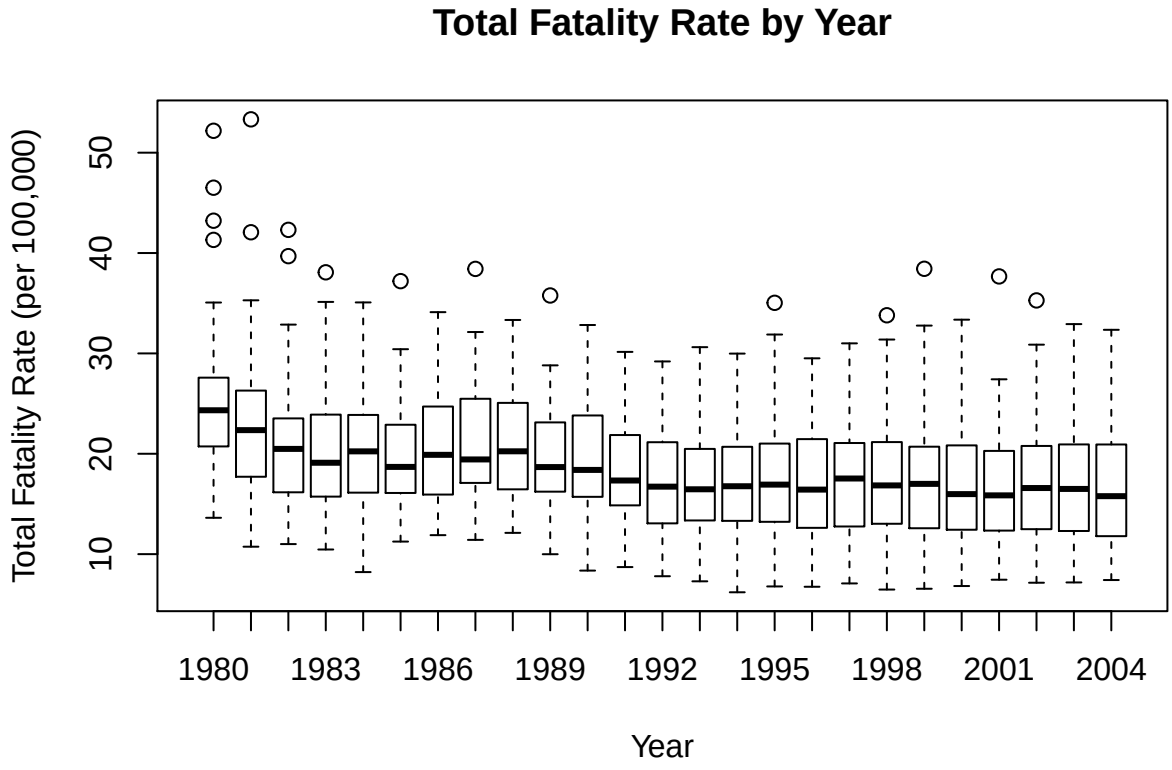
2. How is the dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a very simple regression model of totfatrte on dummy variables for the years 1981 through 2004. What does this model explain?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

The variable *totfatrte* is defined for each state as the total traffic fatalities per 100,000 population. The average for each year is as follows:

```
yearfatrte <- aggregate(df2[, 22], list(df2$year), mean)
yearfatrte
```

```
##    Group.1        x
## 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
```

```
boxplot(totfatrte ~ year, data = df2, main = "Total Fatality Rate by Year",
    xlab = "Year", ylab = "Total Fatality Rate (per 100,000)")
```




```
# 1980 is base year, omitted
yeardum1 <- paste("d", 81:99, sep = "")
yeardum2 <- c("d00", "d01", "d02", "d03", "d04")
yeardum <- c(yeardum1, yeardum2)
yeardum
```

```
##  [1] "d81" "d82" "d83" "d84" "d85" "d86" "d87" "d88" "d89" "d90" "d91"
## [12] "d92" "d93" "d94" "d95" "d96" "d97" "d98" "d99" "d00" "d01" "d02"
## [23] "d03" "d04"
```

```
fmla <- as.formula(paste("totfatrte ~ ", paste(yeardum, collapse = "+")))
lm1 <- lm(fmla, data = df2)
summary(lm1)
```

```
##
## Call:
## lm(formula = fmla, data = df2)
##
## 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 ***
## d81          -1.8244     1.2263  -1.488 0.137094
## d82          -4.5521     1.2263  -3.712 0.000215 ***
## d83          -5.3417     1.2263  -4.356 1.44e-05 ***
## d84          -5.2271     1.2263  -4.263 2.18e-05 ***
## d85          -5.6431     1.2263  -4.602 4.64e-06 ***
## d86          -4.6942     1.2263  -3.828 0.000136 ***
```

```
## d87          -4.7198     1.2263  -3.849 0.000125 ***
## d88          -4.6029     1.2263  -3.754 0.000183 ***
## d89          -5.7223     1.2263  -4.666 3.42e-06 ***
## d90          -5.9894     1.2263  -4.884 1.18e-06 ***
## d91          -7.3998     1.2263  -6.034 2.14e-09 ***
## d92          -8.3367     1.2263  -6.798 1.68e-11 ***
## d93          -8.3669     1.2263  -6.823 1.43e-11 ***
## d94          -8.3394     1.2263  -6.800 1.66e-11 ***
## d95          -7.8260     1.2263  -6.382 2.51e-10 ***
## d96          -8.1252     1.2263  -6.626 5.25e-11 ***
## d97          -7.8840     1.2263  -6.429 1.86e-10 ***
## d98          -8.2292     1.2263  -6.711 3.01e-11 ***
## d99          -8.2442     1.2263  -6.723 2.77e-11 ***
## d00          -8.6690     1.2263  -7.069 2.67e-12 ***
## d01          -8.7019     1.2263  -7.096 2.21e-12 ***
## d02          -8.4650     1.2263  -6.903 8.32e-12 ***
## d03          -8.7310     1.2263  -7.120 1.88e-12 ***
## d04          -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
```

Essentially the model estimates a different intercept for each year. Each intercept is relative to the base level 1980; therefore, its value is the difference of mean in fatality rate compared to 1980.

We notice that the intercept estimates generally decrease as the trend of *totfatrte* also decreases. This is echoed by the time plot of *totfatrte*. Its mean decreased from the mid 20s in 1980 to the mid teens in 2004. Note that all intercept estimates except for 1981 are strongly significant. Statistically, there exists a time trend *totfatrte*. If we do not consider effects of other variables and road safety is only measured by fatality rate, then driving became safer over this period.

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.)

Judging from the EDA, the variables *totfatrte* and *unem* can use log transformations. It is not strictly necessary for *totfatrte* since its skew is quite mild. *bac08* and *bac10*, *sbprim* and *sbsecon* may each be combined into one variable if interpretation of their individual coefficients is not necessary. The one variable for *bac08* and *bac10* would be an indicator for any BAC limit, and the variable for *sbprim* and *sbsecon* would be an indicator for any seatbelt law being enforced. *perc14_24* and *vehicmilespc* have close to normal distribution and the rest are bounded between 0 and 1 with few values in between, therefore no further transformations are applied.

In the following we provide a regression table to compare coefficient estimates across four models:

- Model 1 : year dummies as the only covariates (Question 2 Model)
- Model 2 : Suggested covariates added to model 1, log transform *unem*

- Model 3 : Based on Model 2, with =*bac08* and *bac10* combined into one Bernoulli variable , *sbprim* and *sbsecon* combined into one variable
- Model 4 : Based on Model 3, with *totfatrte* log transformed.

```
fmla2 <- as.formula(paste("totfatrte ~ ", paste(yeardum, collapse = "+"),
    "+ bac08 + bac10 + perse + sbprim +
                          sbsecon + sl70plus + gdl + perc14_24 +
                          log(unem) + vehicmilespc"))
lm2 <- lm(fmla2, data = df2)
# summary(lm2)
```

```
# combine bac08 bac10 and sbprim sbsecon
df2$bac.none = as.numeric(df2$bac08 == df2$bac10)
df2$sb.none = as.numeric(df2$sbprim == df2$sbsecon)

fmla2.1 <- as.formula(paste("totfatrte ~ ", paste(yeardum, collapse = "+"),
    "+ bac.none + perse + sb.none +
                            sl70plus + gdl + perc14_24 +
                            log(unem) + vehicmilespc"))
lm2.1 <- lm(fmla2.1, data = df2)
# summary(lm2.2)
```

```
# combine bac08 bac10 and sbprim sbsecon
df2$bac.none = as.numeric(df2$bac08 == df2$bac10)
df2$sb.none = as.numeric(df2$sbprim == df2$sbsecon)

fmla2.2 <- as.formula(paste("log(totfatrte) ~ ", paste(yeardum,
    collapse = "+"), "+ bac.none + perse + sb.none +
                            sl70plus + gdl + perc14_24 +
                            log(unem) + vehicmilespc"))
lm2.2 <- lm(fmla2.2, data = df2)
# summary(lm2.2)
```

```
stargazer(lm1, lm2, lm2.1, type = "text", column.labels = c("yrs only",
    "added covariates", "merged(bac), merged(sb)"))
```

```
##
## ================================================================================================
##                                                Dependent variable:
##                     ---------------------------------------------------------------------------
##                                                     totfatrte
##                              yrs only            added covariates       merged(bac), merged(sb)
##                                (1)                       (2)                      (3)
## ------------------------------------------------------------------------------------------------
## d81                           -1.824                  -2.100**                 -2.098**
##                              (1.226)                   (0.822)                  (0.826)
##
## d82                          -4.552***                -6.250***                -6.200***
##                              (1.226)                   (0.838)                  (0.843)
##
## d83                          -5.342***                -7.004***                -6.899***
##                              (1.226)                   (0.853)                  (0.863)
##
## d84                          -5.227***                -5.698***                -5.802***
##                              (1.226)                   (0.870)                  (0.874)
```

```
##
## d85                    -5.643***             -6.332***             -6.453***
##                         (1.226)               (0.888)               (0.892)
##
## d86                    -4.694***             -5.535***             -5.732***
##                         (1.226)               (0.925)               (0.927)
##
## d87                    -4.720***             -5.969***             -6.190***
##                         (1.226)               (0.963)               (0.966)
##
## d88                    -4.603***             -6.073***             -6.309***
##                         (1.226)               (1.012)               (1.015)
##
## d89                    -5.722***             -7.594***             -7.857***
##                         (1.226)               (1.050)               (1.054)
##
## d90                    -5.989***             -8.575***             -8.840***
##                         (1.226)               (1.073)               (1.076)
##
## d91                    -7.400***            -10.722***            -11.013***
##                         (1.226)               (1.095)               (1.097)
##
## d92                    -8.337***            -12.506***            -12.787***
##                         (1.226)               (1.115)               (1.119)
##
## d93                    -8.367***            -12.345***            -12.660***
##                         (1.226)               (1.130)               (1.133)
##
## d94                    -8.339***            -11.905***            -12.335***
##                         (1.226)               (1.153)               (1.154)
##
## d95                    -7.826***            -11.384***            -11.808***
##                         (1.226)               (1.181)               (1.183)
##
## d96                    -8.125***            -13.245***            -13.710***
##                         (1.226)               (1.222)               (1.224)
##
## d97                    -7.884***            -13.452***            -14.026***
##                         (1.226)               (1.253)               (1.251)
##
## d98                    -8.229***            -14.106***            -14.667***
##                         (1.226)               (1.271)               (1.271)
##
## d99                    -8.244***            -14.048***            -14.658***
##                         (1.226)               (1.293)               (1.291)
##
## d00                    -8.669***            -14.295***            -14.938***
##                         (1.226)               (1.317)               (1.314)
##
## d01                    -8.702***            -15.389***            -16.232***
##                         (1.226)               (1.335)               (1.326)
##
## d02                    -8.465***            -16.162***            -17.175***
##                         (1.226)               (1.344)               (1.327)
```

```
##
## d03                          -8.731***              -16.500***              -17.717***
##                               (1.226)                (1.354)                 (1.326)
##
## d04                          -8.766***              -16.039***              -17.447***
##                               (1.226)                (1.384)                 (1.341)
##
## bac08                                               -2.605***
##                                                      (0.534)
##
## bac10                                               -1.446***
##                                                      (0.393)
##
## bac.none                                                                     1.120***
##                                                                              (0.363)
##
## perse                                                -0.537*                -0.815***
##                                                      (0.296)                 (0.287)
##
## sbprim                                               -0.354
##                                                      (0.490)
##
## sbsecon                                              -0.147
##                                                      (0.427)
##
## sb.none                                                                      0.107
##                                                                             (0.420)
##
## sl70plus                                            3.204***                3.277***
##                                                      (0.443)                 (0.445)
##
## gdl                                                  -0.380                  -0.361
##                                                      (0.523)                 (0.526)
##
## perc14_24                                            0.180                   0.169
##                                                      (0.122)                 (0.123)
##
## log(unem)                                           5.088***                5.003***
##                                                      (0.481)                 (0.482)
##
## vehicmilespc                                        0.003***                0.003***
##                                                     (0.0001)                (0.0001)
##
## Constant                     25.495***              -7.878***               -8.815***
##                               (0.867)                (2.621)                 (2.667)
##
## ------------------------------------------------------------------------------------------
## Observations                   1,200                  1,200                   1,200
## R2                             0.128                  0.613                   0.608
## Adjusted R2                    0.110                  0.602                   0.597
## Residual Std. Error     6.008 (df = 1175)      4.018 (df = 1165)       4.041 (df = 1167)
## F Statistic         7.164*** (df = 24; 1175) 54.310*** (df = 34; 1165) 56.552*** (df = 32; 1167)
## ==========================================================================================
## Note:                                                        *p<0.1; **p<0.05; ***p<0.01
```

```
stargazer(lm2.2, type = "latex", omit = "d[0-9][0-9]", omit.labels = "Year Dummies Included?",
    omit.yes.no = c("Yes", "No"))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 20, 2017 - 22:13:42

Table 1:

	Dependent variable:
	log(totfatrte)
bac.none	0.017
	(0.019)
perse	−0.023
	(0.015)
sb.none	−0.008
	(0.022)
sl70plus	0.230***
	(0.023)
gdl	−0.002
	(0.027)
perc14_24	0.017***
	(0.006)
log(unem)	0.243***
	(0.025)
vehicmilespc	0.0002***
	(0.00000)
Constant	1.288***
	(0.138)
Year Dummies Included?	Yes
Observations	1,200
R^2	0.643
Adjusted R^2	0.633
Residual Std. Error	0.209 (df = 1167)
F Statistic	65.734*** (df = 32; 1167)
Note:	*p<0.1; **p<0.05; ***p<0.01

Across the three models using the added set of covariates, the seatbelt law variables, whether separate or combined, are insignificant. The standard errors themselves are larger than the estimates. This is somewhat surprising from the time plot we saw in the EDA. If we disregard the state fixed(observed) variables, other covariates in the model, such as speed limit, unemployment rate and vehicle miles per capita, seem to have explained away most of the variations in *totfatrte*.

The coefficient for per se law is significant in both model 2 and 3. Its signs in both models are negative,

suggesting that the adoption of a per se law is associated with lower total fatality rate. In Model 2, adoption of a per se law is estimated to lower total fatality rate by 0.537 fatalities per 100,000 population, which is less than one-tenth of its standard deviation.

bac8 and *bac10* are defined as the legal upper limits (0.08 and 0.10) of blood alcohol level for driving. Their coefficient estimates, whether separate or combined, are consistent in both Model 2 and 3. In Model 2, setting legal blood alcohol limit to 0.08 is estimated to lower total fatality rate by 2.605 fatalities per 100,000 population, and setting it to 0.10 is estimated to do so by 1.446 fatalities. Comparison of these two estimates makes sense, since setting a higher limit is more lenient to drunk driving.

We acknowledge that with panel data, pooled OLS are generally invalid due to omitted variable biases and serial correlations. To use OLS we must assume that the errors are uncorrelated with our explanatory variables in order to consistently estimate our parameters. However, this is likely not the case here. There are several factors such as topography (e.g., steep mountains, winding roads) and climate (e.g., icy roads, precipitation) that affect the fatality rate and may be correlated with our explanatory variables, but that are omitted from this model. These factors just mentioned are time-invariant, but may be correlated with our explanatory variables. For example, regions with winding or frequently icy roads may be more likely to enact seatbelt laws. Since we have this heterogeneity bias, OLS is not appropriate and will not produce unbiased estimates of our parameters of interest. Out of curiosity, we still conducted brief residual studies to highlight how the models differ from each other.

Assumption CLM.1 (Linear Specification) :

- Weak assumption, we have specified linear relationship between coefficients and the dependent variable.

$$totfatrte = \beta_0 + \beta_1 bac08 + \beta_2 bac10 + \beta_3 perse + \beta_4 sbprim + \beta_5 sbsecon$$
$$+\beta_6 sl70plus + \beta_7 gdl + \beta_8 perc14_24 + \beta_9 log(unem) + \beta_{10} vehicmilespc + \sum_{yr=81}^{yr=04} \beta_{yr} yr + u$$

Assumption CLM.2 (Random Sampling) :

- Clearly violated due to serial correlation across time and dropping out of Alaska, Hawaii and DC. These three states are either geographically disconnected or administratively distinct from other states.

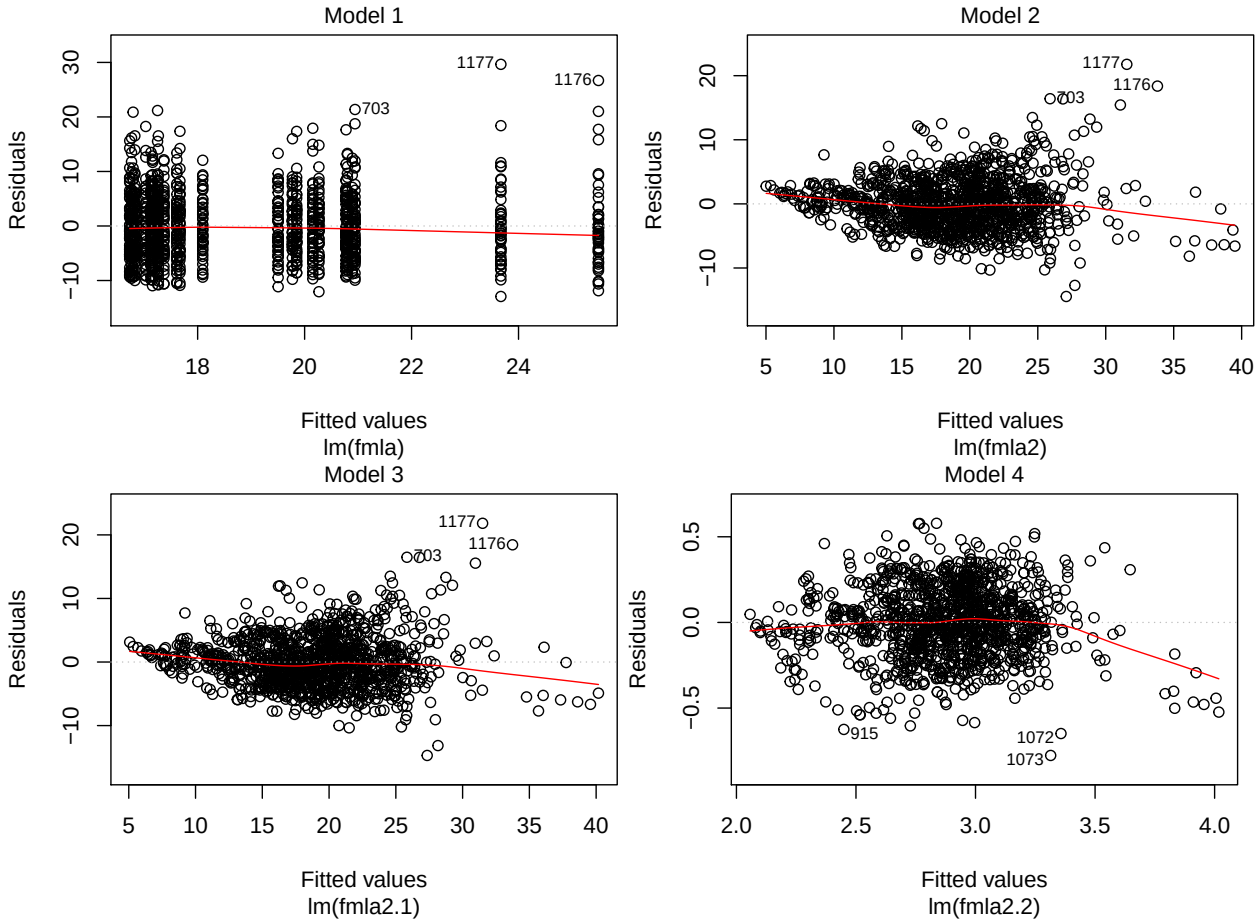
Assumption CLM.3 (no perfect collinearity) :

- We have seen from the correlation heat map in the EDA that this assumption holds.

Assumption CLM.4 (Zero-conditional mean $E(u|X) = 0$) :

- Model 1 : Loess curve shows gradual decline. Recall from the "Total Fatality Rate over Time" box plot in the EDA, if we dissect *totfatrte* by year its distribution gradual shifts from normal to slightly right skewed. Estimating an intercept for each year will fit the regression line through these means and produce residuals that mirror that data behavior.
- Model 2, 3, 4 : Loess curve shows a mild dip when fitted values is between 15-20. Further log transformation *totfatrte* in model 4 seem to flip the loess curve, which hasn't solved the problem.

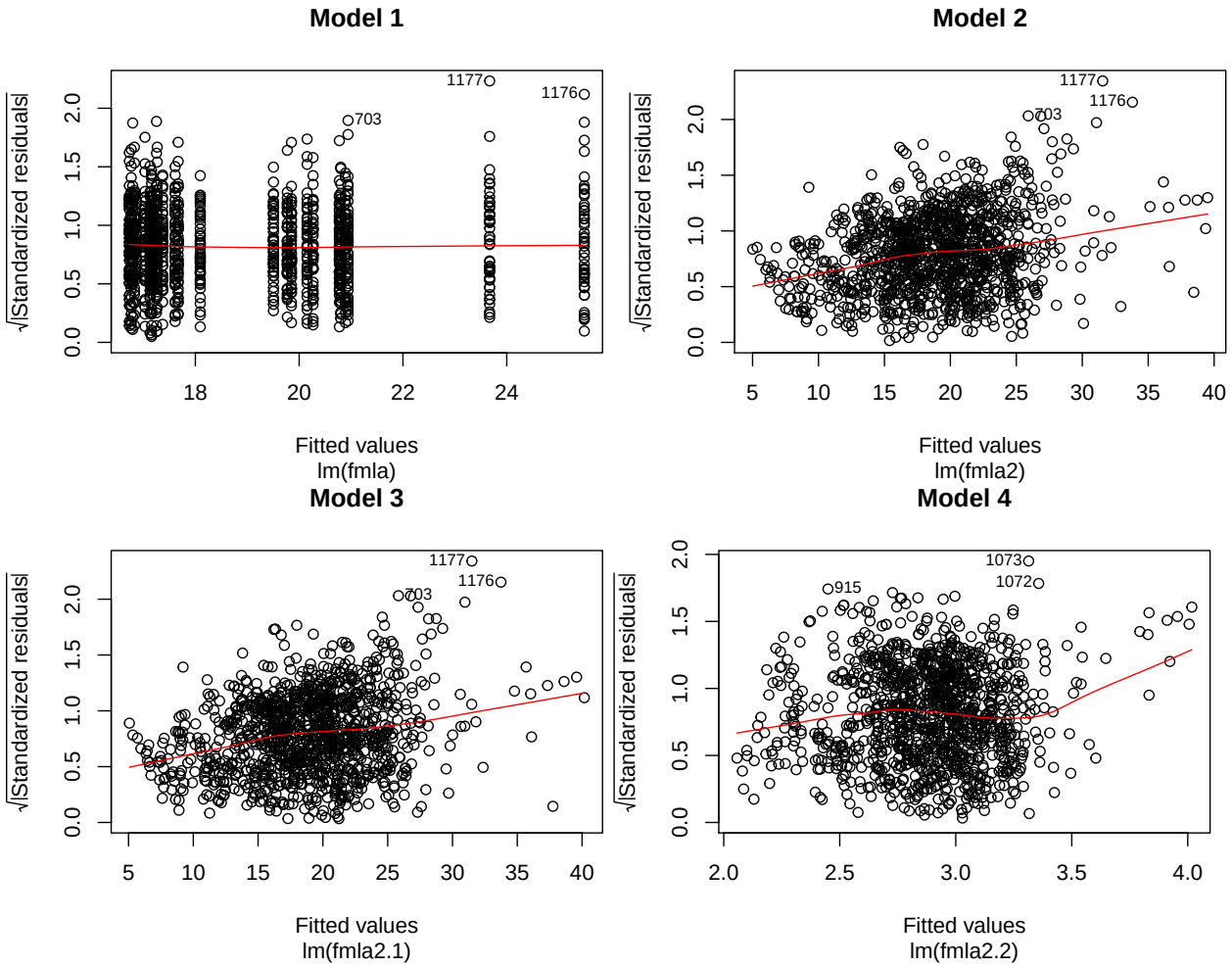
```
i = 1
plot(lm1, which = i, caption = "Model 1")
plot(lm2, which = i, caption = "Model 2")
plot(lm2.1, which = i, caption = "Model 3")
plot(lm2.2, which = i, caption = "Model 4")
```



Assumption CLM.5 (Homoskedasticity $var(u|X) = var(u) = \sigma_u^2$) :

- Model 2,3 : Without log transformation, there's clear heteroskedasticity as fitted values increases
- Model 4 : With log transformations applied on *totfatrte*, variance of residuals seems more stabilized.

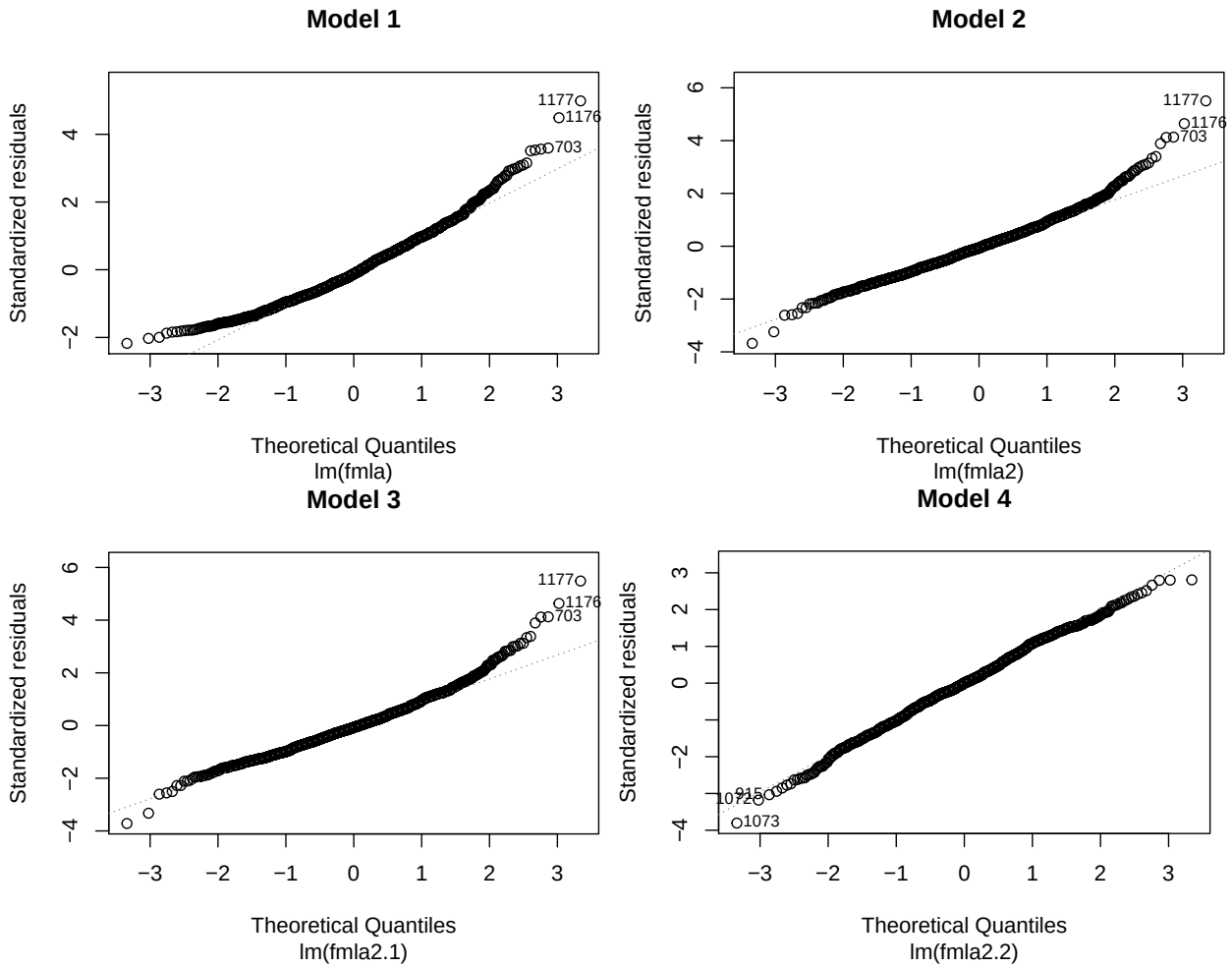
```
i = 3
plot(lm1, which = i, caption = "Model 1")
title("Model 1")
plot(lm2, which = i, caption = "Model 2")
title("Model 2")
plot(lm2.1, which = i, caption = "Model 3")
title("Model 3")
plot(lm2.2, which = i, caption = "Model 4")
title("Model 4")
```



Assumption CLM.6 (Normality of residuals $N(0, \sigma_u^2)$) :

Model 1,2,3 vs 4 : The log transformed *totfatrte* model produces closer to normal residuals than the other two.

```
i = 2
plot(lm1, which = i, caption = "Model 1")
title("Model 1")
plot(lm2, which = i, caption = "Model 2")
title("Model 2")
plot(lm2.1, which = i, caption = "Model 3")
title("Model 3")
plot(lm2.2, which = i, caption = "Model 4")
title("Model 4")
```



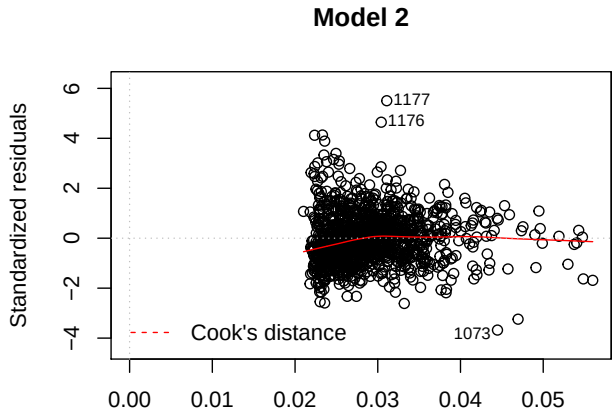
Outlier Analysis :

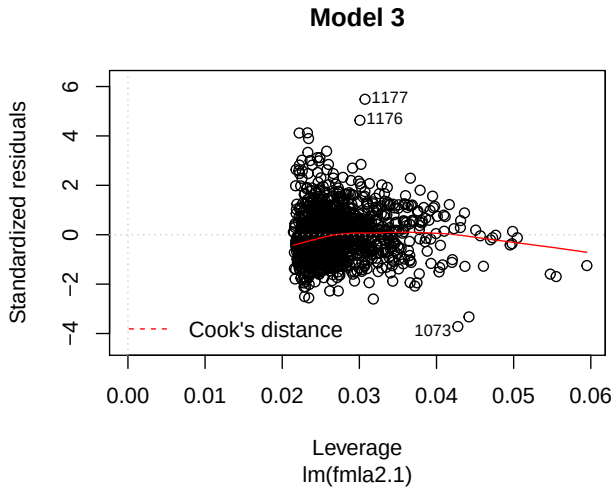
None of the observations are close to cook's distance of 0.5; it looks like the variation in the data is fairly captured.

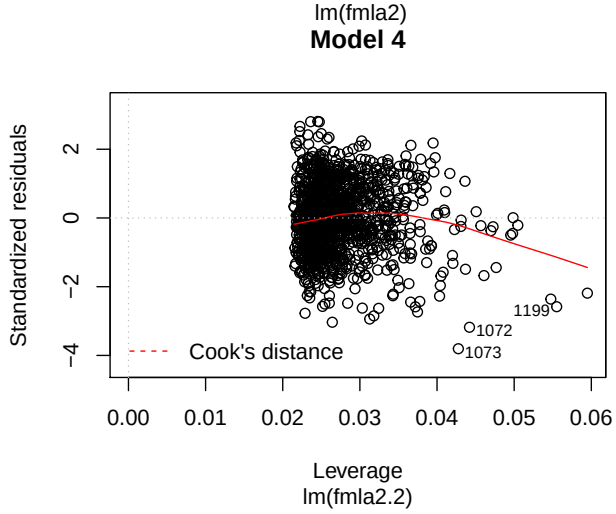
```
i = 5
plot(lm1, which = i, caption = "Model 1")
```

```
## hat values (leverages) are all = 0.02083333
##  and there are no factor predictors; no plot no. 5
```

```
plot(lm2, which = i, caption = "Model 2")
title("Model 2")
plot(lm2.1, which = i, caption = "Model 3")
title("Model 3")
plot(lm2.2, which = i, caption = "Model 4")
title("Model 4")
```







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?

```
df.panel <- plm.data(df2, c("state.name", "year"))
# summary(df.panel)


fmla3 <- as.formula(paste("totfatrte ~ ", paste(yeardum, collapse = "+"),
    "+ bac08 + bac10 + perse + sbprim +
                          sbsecon + sl70plus + gdl + perc14_24 +
                          log(unem) + vehicmilespc"))
plm1 <- plm(fmla3, data = df.panel, model = "fd")
# summary(plm1)

fmla3.1 <- as.formula(paste("totfatrte ~ ", paste(yeardum, collapse = "+"),
    "+ bac08 + bac10 + perse + sbprim +
                            sbsecon + sl70plus + gdl + perc14_24 +
                            log(unem) + vehicmilespc"))
plm1.1 <- plm(fmla3.1, data = df.panel, model = "within")
# summary(plm1.1)
```

```
stargazer(lm2, plm1.1, plm1, type = "latex", omit = "d[0-9][0-9]",
    omit.labels = "Year Dummies Included?", omit.yes.no = c("Yes",
        "No"), column.labels = c("OLS", "Within", "First Differenced"))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 20, 2017 - 22:13:52

From the bivariate plots in our EDA, we had the intuition of the existence of state fixed (unobserved effects). Therefore, we attempted two fixed effect estimators (first-differenced (FD) and within (FE)). Generally estimates from the fixed effect estimators are quite different compared to that of pooled OLS (see Table 2).

- For blood alcohol limit variables *bac08* and *bac10*, the estimates are less negative using the fixed effect estimators. They stayed significant in the within model but not in the first differenced model. The signs of coefficients are consistent but generally blood alcohol limit is estimated to have a lesser effect on total fatality rate. Using the within model, setting legal blood alcohol limit as 0.10 is estimated to lower total fatality rate by 0.959 fatalities per 100,000 population and that of 0.08 is 1.317.
- For per se law, the within and first differenced estimators yielded coefficients that diverge from the OLS model in different directions. On the one hand, the within model estimated a more significant and more negative coefficient, which suggests that enforcement of per se law should have a larger effect on total fatality rate. On the other, the first differenced model estimated a less significant and less negative coefficient, which suggests that the adoption of per se law should have a lesser, or no effect on total fatality rate. Using the within model, enforcement of per se law is estimated to lower total fatality rate by 1.218 fatalities per 100,000 population.
- For primary seat belt law, the OLS model and first difference model results are very similar, while the within model estimated a more negative and significant coefficient. Using the within model, enforcing the primary seatbelt law is estimated to lower total fatality rate by 1.168 fatalities per 100,000 population.

In the last question, we have already examined the CLM assumptions for the OLS models. In brief, the models violated random sampling because observations of each state are serially correlated; there are some mild violations of zero-conditional mean from the residual plot and signs of heteroskedasticity from the scale-location plot. From the above regression table and discussion, we saw how fixed effect estimators (both FD and FE) produced results that are noticeably different than the OLS model in terms of both magnitude and significance. This is a sign of heterogeneity bias in the OLS model. The omitted time invariant variables are effectively removed by taking differences between each observation and the panel mean, or its last record in the same panel.

Below, we evaluate if the fixed effect model assumptions hold. At the same time, we compare mechanisms between the within and first differenced estimators.

Assumption FE.1 (linear specification of parameters and unobserved effect)

- This is a weak assumption. For each state i, the general model is

$$\begin{aligned} totfatrte_{it} = \beta_0 + \beta_1 bac08_{it} + \beta_2 bac10_{it} + \beta_3 perse_{it} + \beta_4 sbprim_{it} \\ + \beta_5 sbsecon_{it} + \beta_6 sl70plus_{it} + \beta_7 gdl_{it} + \beta_8 perc14_24_{it} \\ + \beta_9 log(unem)_{it} + \beta_{10} vehicmilespc_{it} + \sum_{yr=81}^{yr=04} \delta_{yr} I(yr_t) + a_i + u_{it} \end{aligned}$$

where $\beta_i t$ are the parameters to estimate and a_i are the unobserved state-level effects. If a variable doesn't change over time it will be differenced or demeaned away.

- The within model for OLS application is

Table 2:

	Dependent variable:		
	totfatrte		
	OLS	*panel linear*	
	OLS	Within	First Differenced
	(1)	(2)	(3)
Constant	-7.878^{***}		
	(2.621)		
bac08	-2.605^{***}	-1.317^{***}	-0.819
	(0.534)	(0.396)	(0.587)
bac10	-1.446^{***}	-0.959^{***}	-1.018^{**}
	(0.393)	(0.270)	(0.445)
perse	-0.537^{*}	-1.218^{***}	-0.616
	(0.296)	(0.233)	(0.390)
sbprim	-0.354	-1.168^{***}	-0.345
	(0.490)	(0.343)	(0.482)
sbsecon	-0.147	-0.297	-0.311
	(0.427)	(0.252)	(0.295)
sl70plus	3.204^{***}	0.033	0.335
	(0.443)	(0.270)	(0.564)
gdl	-0.380	-0.388	-0.203
	(0.523)	(0.293)	(0.370)
perc14_24	0.180	0.160^{*}	0.904^{***}
	(0.122)	(0.096)	(0.311)
log(unem)	5.088^{***}	-3.664^{***}	-1.549^{***}
	(0.481)	(0.393)	(0.484)
vehicmilespc	0.003^{***}	0.001^{***}	0.0004^{*}
	(0.0001)	(0.0001)	(0.0002)
Constant	-7.878^{***}		
	(2.621)		
Year Dummies Included?	Yes	Yes	Yes
Observations	1,200	1,200	1,152
R^2	0.613	0.626	0.178
Adjusted R^2	0.602	0.598	0.153
Residual Std. Error	4.018 (df = 1165)		
F Statistic	54.310^{***} (df = 34; 1165)	54.932^{***} (df = 34; 1118)	7.324^{***} (df = 33; 1118)
Note:			*p<0.1; **p<0.05; ***p<0.01

$$totfatrte_{it} - \bar{totfatrte_i} = \beta_1(bac08_{it} - \bar{bac08_i}) + \beta_2(bac10_{it} - \bar{bac10_i})$$
$$+\beta_3(perse_{it} - \bar{perse_i}) + \beta_4(sbprim_{it} - \bar{sbprim_i}) + \beta_5(sbsecon_{it} - \bar{sbsecon_i})$$
$$+\beta_6(sl70plus_{it} - \bar{sl70plus_i}) + \beta_7(gdl_{it} - \bar{gdl_i}) + \beta_8(perc14\_24_{it} - \bar{perc14\_24_i})$$
$$+\beta_9(log(unem)_{it} - \bar{log(unem)_i}) + \beta_{10}(vehicmilespc_{it} - \bar{vehicmilespc_i})$$
$$+ \sum_{yr=81}^{yr=04} \delta_{yr}(I(yr_t) - \bar{I(yr)}) + (u_{it} - \bar{u_i})$$

- The first differenced model for OLS application is

$$\Delta totfatrte_{it} = \beta_1 \Delta bac08_{it} + \beta_2 \Delta bac10_{it} + \beta_3 \Delta perse_{it} + \beta_4 \Delta sbprim_{it}$$
$$+\beta_5 \Delta sbsecon_{it} + \beta_6 \Delta sl70plus_{it} + \beta_7 \Delta gdl_{it} + \beta_8 \Delta perc14_24_{it}$$
$$+\beta_9 \Delta log(unem)_{it} + \beta_{10} \Delta vehicmilespc_{it} + \sum_{yr=81}^{yr=04} \delta_{yr} \Delta I(yr_t) + \Delta u_{it}$$

Assumption FE.2 (random sample in each cross section)

As mentioned in the CLM assumptions, we believe Alaska, Hawaii and DC were not randomly excluded from the population. Therefore this model cannot be generalized to geographically disconnected areas or the capital.

Assumption FE.3 (each explanatory variable changes over time, no perfect linear relationships between them)

From the EDA, we saw that most states enforced their blood alcohol limits, seatbelt laws and per se law early and there isn't much variation across time. First differencing causes many zero values in the differenced variables, while time demeaning better preserves the magnitudes. Therefore, the FD model estimated large standard errors and insignificant coefficients, while the within model more efficiently estimate the coefficients.

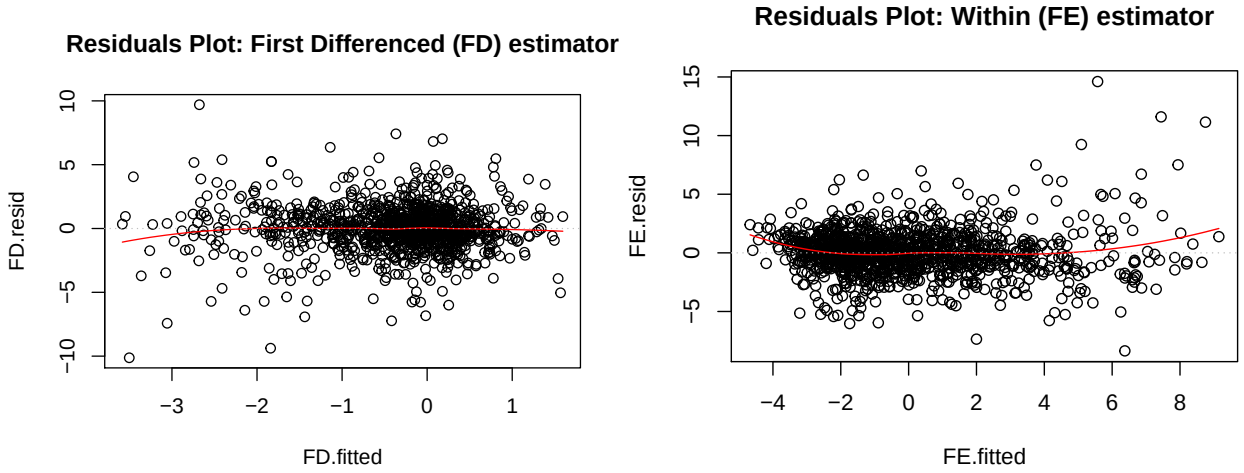
Assumption FE.4 (strict exogeneity assumption $E(u_{it}|X_i, a_i) = 0$)

Although time invariant unobserved variables are differenced or demeaned away. We don't have a good way to identify unobserved time variant variables. Therefore we cannot guarantee the expected value of the idiosyncratic error given the explanatory variables in all time periods and the unobserved time variant effect is zero.

Assuming no time variant variables are omitted. Since we can't estimate u_{it} directly, so we examine if $E(\Delta u_{it}|\Delta X_i) = 0$. The residuals plots for the FD and FE estimators below each show a mostly flat Loess curve where most of our observations lie. Within the FE and FD estimators, the assumption appear to hold. But we should be careful about generalizing this to the general, untransformed forumla.

```
FD.fitted = predict(plm1)
FD.resid = plm1$residuals
plot(FD.resid ~ FD.fitted)

# Loess curve
order.pred <- order(FD.fitted)
smooth.stand <- loess(formula = FD.resid ~ FD.fitted, weights = rep(1,
    1152))
lines(x = FD.fitted[order.pred], y = predict(smooth.stand)[order.pred],
    lty = "solid", col = "red")
abline(h = 0, col = "gray", lty = "dotted")
```

```
title("Residuals Plot: First Differenced (FD) estimator")

FE.fitted = predict(plm1.1)
FE.resid = plm1.1$residuals
plot(FE.resid ~ FE.fitted)

# Loess curve
order.pred <- order(FE.fitted)
smooth.stand <- loess(formula = FE.resid ~ FE.fitted, weights = rep(1,
    1200))
lines(x = FE.fitted[order.pred], y = predict(smooth.stand)[order.pred],
    lty = "solid", col = "red")
abline(h = 0, col = "gray", lty = "dotted")
title("Residuals Plot: Within (FE) estimator")
```

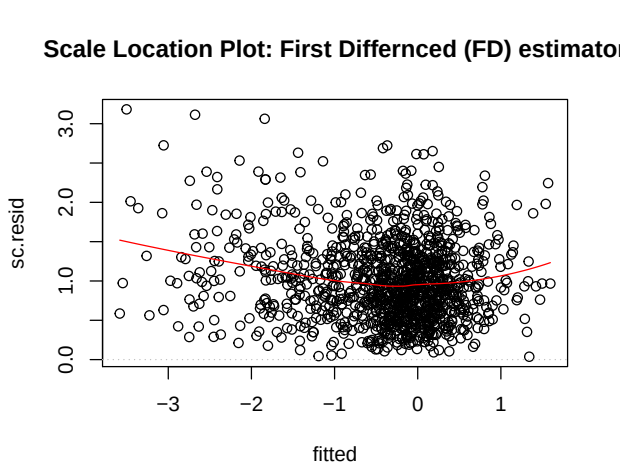


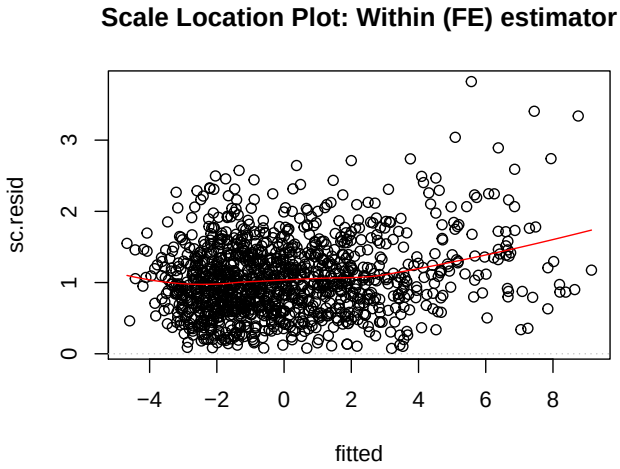
Assumption FE.5 (homoskedasticity $Var(u_{it}|X_i, a_i) = Var(u_{it}) = \sigma_u^2$ for all $t = 1, ..., T$)

- Scale location plots show clear signs of heteroskedasticity. Residuals tend to scatter more with lower fitted values in the FD model and with higher fitted values in the FE estimator. This indicates that the FD model doesn't capture variations in observations well when the change from one time period to the other is more negative, and the FE model doesn't capture variations in observations well when the observations are high above the panel means. The Breusch Pagan tests also strongly reject the null hypotheses of homoskedasticity for both model.
- We also provided the robust standard errors to correct for heteroskedasticity. The table below shows that after robust standard errors of larger magnitudes are applied and *bac08* becomes insignificant in the within (FE) model.

```
scale_location.plot = function(plm_model) {
    fitted = predict(plm_model)
    resid = plm_model$residuals
    sc.resid = sqrt(abs(resid))
    plot(sc.resid ~ fitted)

    # Loess curve
    order.pred <- order(fitted)
    smooth.stand <- loess(formula = sc.resid ~ fitted, weights = rep(1,
        length(fitted)))
    lines(x = fitted[order.pred], y = predict(smooth.stand)[order.pred],
        lty = "solid", col = "red")
```

```
    abline(h = 0, col = "gray", lty = "dotted")
}

scale_location.plot(plm1)
title("Scale Location Plot: First Differnced (FD) estimator")

scale_location.plot(plm1.1)
title("Scale Location Plot: Within (FE) estimator")
```






```
cat("First Differenced Estimator (FD) : \n")
```

```
## First Differenced Estimator (FD) :
```

```
lmtest::bptest(plm1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  plm1
## BP = 146.53, df = 34, p-value = 6.363e-16
```

```
cat("Within Estimator (FE) : \n")
```

```
## Within Estimator (FE) :
```

```
lmtest::bptest(plm1.1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  plm1.1
## BP = 146.53, df = 34, p-value = 6.363e-16
```

```
FD.se.robust = lmtest::coeftest(plm1, vcov = vcovHC)[, "Std. Error"]
FE.se.robust = lmtest::coeftest(plm1.1, vcov = vcovHC)[, "Std. Error"]
```

```
get_robust.ci.lower = function(se, pt.estimate, alpha, df) {
    pt.estimate + qt(p = alpha/2, df = df) * se
}

get_robust.ci.upper = function(se, pt.estimate, alpha, df) {
    pt.estimate + qt(p = 1 - alpha/2, df = df) * se
```

```
}
```

```
FD.coef = plm1$coefficients  #[25:34]
FE.coef = plm1.1$coefficients  #[25:34]

FD.lower = get_robust.ci.lower(FD.se.robust, FD.coef, 0.05, 1118)
FD.upper = get_robust.ci.upper(FD.se.robust, FD.coef, 0.05, 1118)
FD.sign = (FD.lower > 0 & FD.upper > 0) | (FD.lower < 0 & FD.upper <
    0)

FE.lower = get_robust.ci.lower(FE.se.robust, FE.coef, 0.05, 1118)
FE.upper = get_robust.ci.upper(FE.se.robust, FE.coef, 0.05, 1118)
FE.sign = (FE.lower > 0 & FE.upper > 0) | (FE.lower < 0 & FE.upper <
    0)
```

```
cbind(FD.coef, se.robust = FD.se.robust, FD.lower, FD.upper,
    significance = FD.sign)
```

```
##                    FD.coef    se.robust      FD.lower      FD.upper
## (intercept)  -0.2125634989 0.0867863780 -3.828460e-01 -0.0422809765
## d81          -1.2150566914 0.4254627686 -2.049852e+00 -0.3802612422
## d82          -2.8991708296 0.3812709526 -3.647258e+00 -2.1510836181
## d83          -2.8267251200 0.3268659472 -3.468065e+00 -2.1853853240
## d84          -2.2738101942 0.3989094303 -3.056506e+00 -1.4911147349
## d85          -2.1152833842 0.4042341631 -2.908426e+00 -1.3221403298
## d86          -0.5717120415 0.5237845682 -1.599424e+00  0.4559994439
## d87          -0.2432007186 0.5349147505 -1.292751e+00  0.8063491655
## d88           0.2436586209 0.5819790038 -8.982355e-01  1.3855527178
## d89          -0.3494620857 0.6274954236 -1.580663e+00  0.8817392384
## d90          -0.0327083238 0.6549590089 -1.317796e+00  1.2523789750
## d91          -0.6942381420 0.6194810650 -1.909715e+00  0.5212383042
## d92          -1.2525713448 0.5457789367 -2.323438e+00 -0.1817049699
## d93          -1.1031063932 0.5573843135 -2.196744e+00 -0.0094692462
## d94          -1.0134684043 0.6260430849 -2.241820e+00  0.2148833033
## d95          -0.3669436821 0.5928097078 -1.530089e+00  0.7962012107
## d96          -0.5277998008 0.5834558334 -1.672592e+00  0.6169919660
## d97          -0.3481991838 0.5653933710 -1.457551e+00  0.7611524398
## d98          -0.7285801925 0.5457462647 -1.799382e+00  0.3422220771
## d99          -0.6838677420 0.5475376857 -1.758185e+00  0.3904494534
## d00          -0.9686467179 0.4104496880 -1.773985e+00 -0.1633082560
## d01          -0.5275924536 0.3400177388 -1.194737e+00  0.1395523166
## d02           0.1247298261 0.3039273556 -4.716024e-01  0.7210620837
## d03           0.1057633136 0.2619890673 -4.082823e-01  0.6198089535
## bac08        -0.8185256429 0.5452897817 -1.888432e+00  0.2513809668
## bac10        -1.0177167419 0.3951583946 -1.793052e+00 -0.2423811453
## perse        -0.6161969651 0.3765135500 -1.354950e+00  0.1225558034
## sbprim       -0.3454437140 0.4462021571 -1.220932e+00  0.5300442433
## sbsecon      -0.3106623897 0.3311263044 -9.603614e-01  0.3390366026
## sl70plus      0.3346554035 0.5043929717 -6.550081e-01  1.3243188676
## gdl          -0.2032203819 0.2562878361 -7.060797e-01  0.2996389399
## perc14_24     0.9043104961 0.3213303124  2.738321e-01  1.5347888888
## log(unem)    -1.5490894134 0.4457594875 -2.423709e+00 -0.6744700128
## vehicmilespc  0.0003614197 0.0002300485 -8.995572e-05  0.0008127951
##              significance
```

```
## (intercept)             1
## d81                      1
## d82                      1
## d83                      1
## d84                      1
## d85                      1
## d86                      0
## d87                      0
## d88                      0
## d89                      0
## d90                      0
## d91                      0
## d92                      1
## d93                      1
## d94                      0
## d95                      0
## d96                      0
## d97                      0
## d98                      0
## d99                      0
## d00                      1
## d01                      0
## d02                      0
## d03                      0
## bac08                    0
## bac10                    1
## perse                    0
## sbprim                   0
## sbsecon                  0
## sl70plus                 0
## gdl                      0
## perc14_24                1
## log(unem)                1
## vehicmilespc             0
```

```
cbind(FE.coef, se.robust = FE.se.robust, FE.lower, FE.upper,
    significance = FE.sign)
```

```
##                    FE.coef    se.robust      FE.lower     FE.upper
## d81          -1.577868e+00 0.4253888601 -2.412519e+00 -0.743217980
## d82          -3.346397e+00 0.4287783621 -4.187698e+00 -2.505095991
## d83          -3.877137e+00 0.4447331422 -4.749742e+00 -3.004531186
## d84          -4.425062e+00 0.4324491645 -5.273565e+00 -3.576558721
## d85          -4.883527e+00 0.4540810375 -5.774474e+00 -3.992579924
## d86          -3.898229e+00 0.5811823144 -5.038560e+00 -2.757898392
## d87          -4.582660e+00 0.6682905794 -5.893905e+00 -3.271415405
## d88          -5.105365e+00 0.7369838612 -6.551393e+00 -3.659338177
## d89          -6.431512e+00 0.8343464061 -8.068573e+00 -4.794450661
## d90          -6.471918e+00 0.8876459030 -8.213558e+00 -4.730278540
## d91          -7.147248e+00 0.9574752675 -9.025899e+00 -5.268596820
## d92          -8.034183e+00 1.0415190965 -1.007774e+01 -5.990630550
## d93          -8.348390e+00 1.0756336736 -1.045888e+01 -6.237901501
## d94          -8.786037e+00 1.0507563816 -1.084771e+01 -6.724360680
## d95          -8.595692e+00 1.1384898265 -1.082951e+01 -6.361874422
## d96          -8.991145e+00 1.1242446379 -1.119701e+01 -6.785277913
```

```
## d97          -9.210849e+00 1.1806585264 -1.152741e+01 -6.894293024
## d98          -9.944602e+00 1.1905647611 -1.228060e+01 -7.608609377
## d99          -1.015240e+01 1.2991364513 -1.270142e+01 -7.603376596
## d00          -1.074782e+01 1.2871176501 -1.327326e+01 -8.222380311
## d01          -1.016207e+01 1.4079582153 -1.292460e+01 -7.399528189
## d02          -9.293965e+00 1.4207513934 -1.208160e+01 -6.506325521
## d03          -9.294010e+00 1.4691502071 -1.217661e+01 -6.411407316
## d04          -9.784410e+00 1.6044858941 -1.293255e+01 -6.636267296
## bac08        -1.316974e+00 0.7970670782 -2.880890e+00  0.246941498
## bac10        -9.590871e-01 0.4879294477 -1.916448e+00 -0.001726567
## perse        -1.217545e+00 0.4269612935 -2.055281e+00 -0.379809742
## sbprim       -1.168263e+00 0.5519034642 -2.251146e+00 -0.085380006
## sbsecon      -2.971611e-01 0.3703814625 -1.023882e+00  0.429559984
## sl70plus      3.280878e-02 0.5521608533 -1.050579e+00  1.116197036
## gdl          -3.884104e-01 0.3671798016 -1.108850e+00  0.332028743
## perc14_24     1.599316e-01 0.1671922871 -1.681144e-01  0.487977594
## log(unem)    -3.663879e+00 0.7333710799 -5.102818e+00 -2.224940276
## vehicmilespc  9.624022e-04 0.0003516965  2.723427e-04  0.001652462
##              significance
## d81                     1
## d82                     1
## d83                     1
## d84                     1
## d85                     1
## d86                     1
## d87                     1
## d88                     1
## d89                     1
## d90                     1
## d91                     1
## d92                     1
## d93                     1
## d94                     1
## d95                     1
## d96                     1
## d97                     1
## d98                     1
## d99                     1
## d00                     1
## d01                     1
## d02                     1
## d03                     1
## d04                     1
## bac08                   0
## bac10                   1
## perse                   1
## sbprim                  1
## sbsecon                 0
## sl70plus                0
## gdl                     0
## perc14_24               0
## log(unem)               1
## vehicmilespc            1
```

Assumption FE.6 (For all $t \neq s$, the idiosyncratic errors are uncorrelated, conditional on all explanatory variables and unobserved variables. $Cov(u_{it}, u_{is}|X_i, a_i) = 0$)

- The PBG tests for serial correlation
  - Ho : The idiosyncratic errors are uncorrelated
  - Ho : The idiosyncratic errors are correlated
- From the test results below, $\Delta\hat{u_{it}}$ from the first differenced estimator shows substantial, negative serial correlation, possibly an AR(1) itself. This is a sign that u_{it} is not a random walk and we have over-differenced the observations. This assumption is broken for the FD model.
- On the other hand $\ddot{\hat{u}}_{it}$ from the within estimator shows substantial positive serial correlation, possibly an AR(2) itself. This time-demeaned series is harder to directly relate to u_{it}, yet the serial correlation is a problem for accurate standard error estimates in the within model.

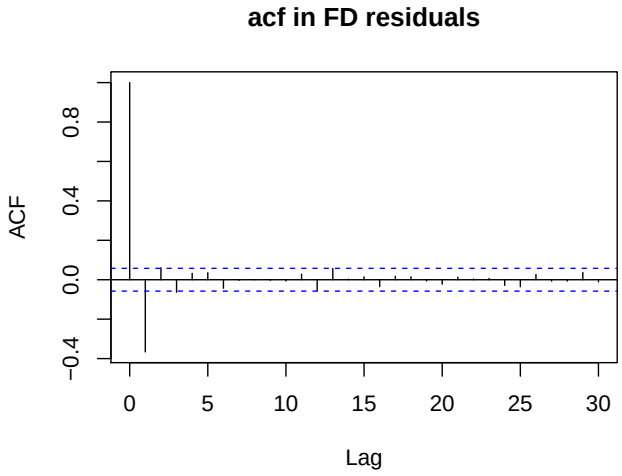
```
cat("First differenced (FD) model : \n")
```

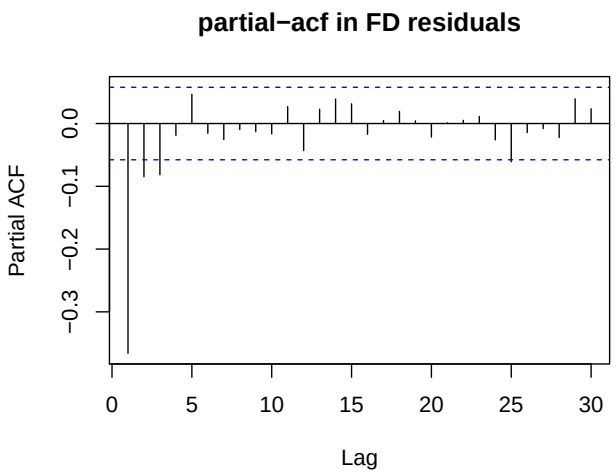
```
## First differenced (FD) model :
```

```
pbgtest(plm1)  # fd model
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  fmla3
## chisq = 187.23, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
acf(plm1$residuals, main = "")
title("acf in FD residuals")
pacf(plm1$residuals, main = "")
title("partial-acf in FD residuals")
```






```
cat("Within (FE) model : \n")
```

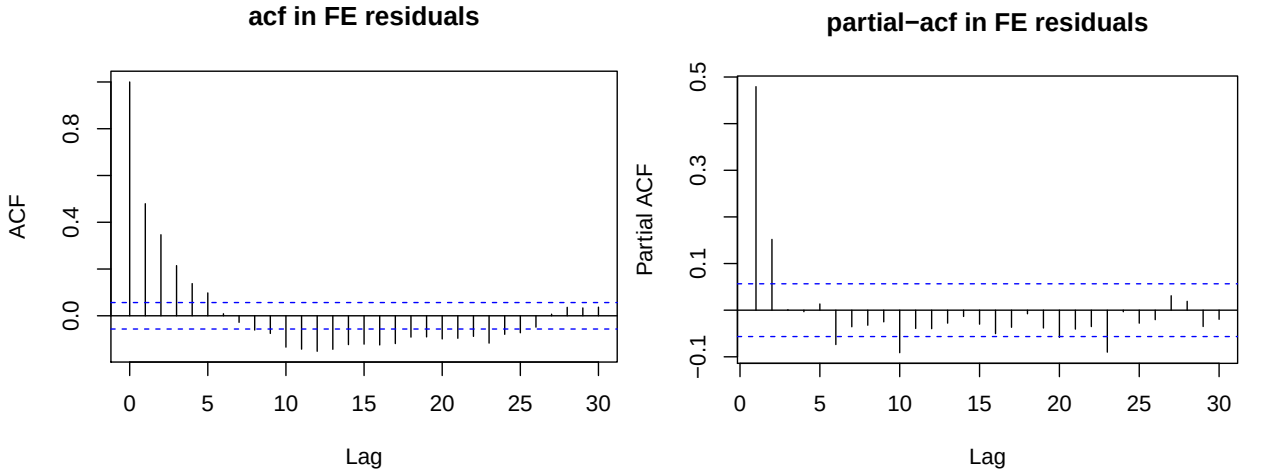
```
## Within (FE) model :
```

```
pbgtest(plm1.1)  # fd model
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
```

```
##
## data:  fmla3.1
## chisq = 341.1, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
acf(plm1.1$residuals, main = "")
title("acf in FE residuals")
pacf(plm1.1$residuals, main = "")
title("partial-acf in FE residuals")
```



Assumption FE.7 (normal distribution of idiosyncratic error conditional on time variant variables and observed variables, $Normal(0, \sigma_u^2)$)

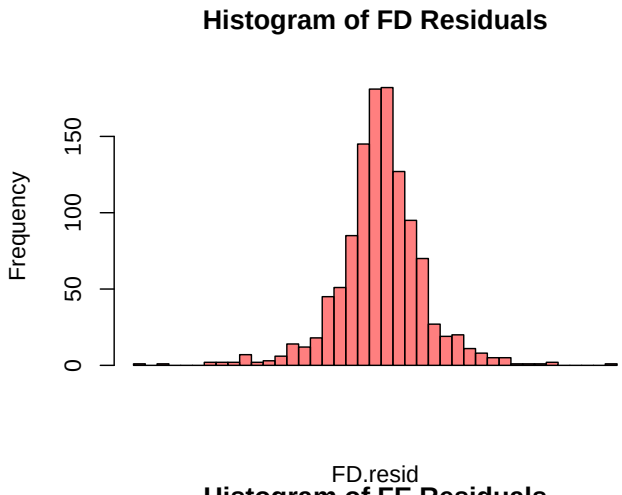
- For this assumption to hold, the last three must be true. From the above discussion we have reservations about each of them. Not being able to directly estimate u_{it} is also a problem. If we can assume that these three assumptions hold, we observe the following distribution properties of the residuals:
- The histograms and QQ plots of the overall residuals tell us that the FD residuals are mildly left skewed and the FE residuals are mildly right skewed.
  - The box plots per year and per state show only minor skews occur at some states and years. We should be able to rely on sample size of 48 in each cross-section and sample size of 25 in each panel to test significance of coefficients.

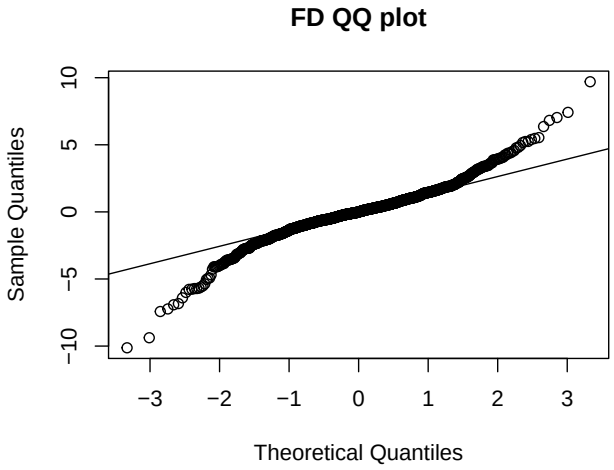
```
# FD
hist(FD.resid, breaks = 30, xaxt = "n", col = rgb(1, 0, 0, 0.5),
    main = "Histogram of FD Residuals")
qqnorm(FD.resid, main = "")
qqline(FD.resid)
title("FD QQ plot")
# FE within
hist(FE.resid, breaks = 30, xaxt = "n", col = rgb(0, 0, 1, 0.5),
    main = "Histogram of FE Residuals")
qqnorm(FE.resid, main = "")
qqline(FE.resid)
title("FE QQ plot")

moments::skewness(FD.resid)
```

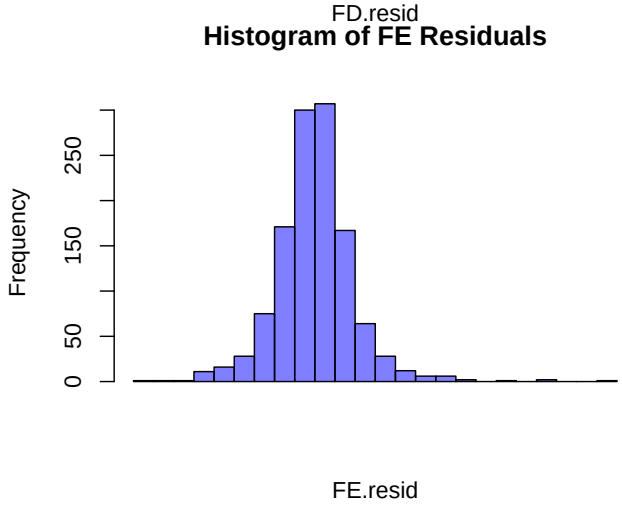
```
## [1] -0.2431281
```

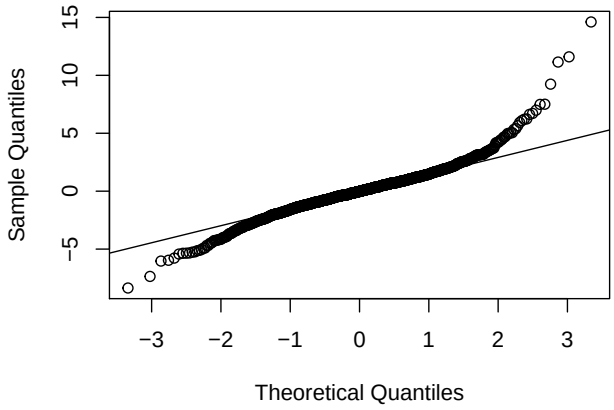
```
moments::skewness(FE.resid)
```

```
## [1] 0.765644
```

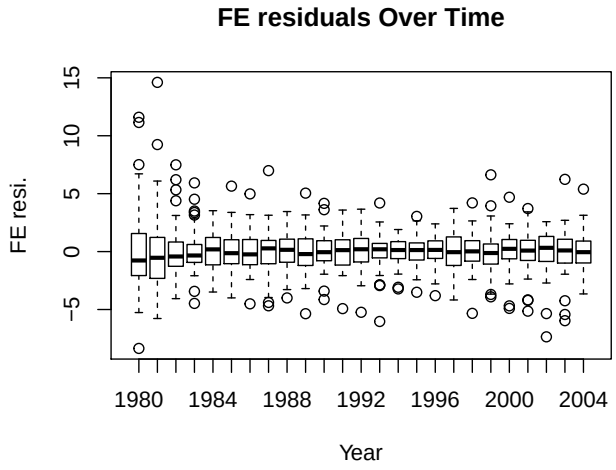


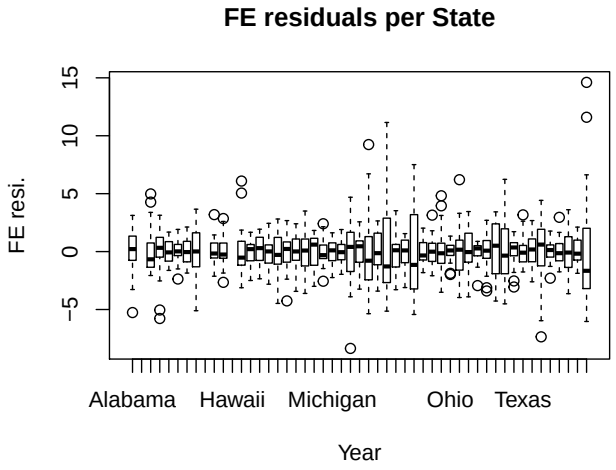




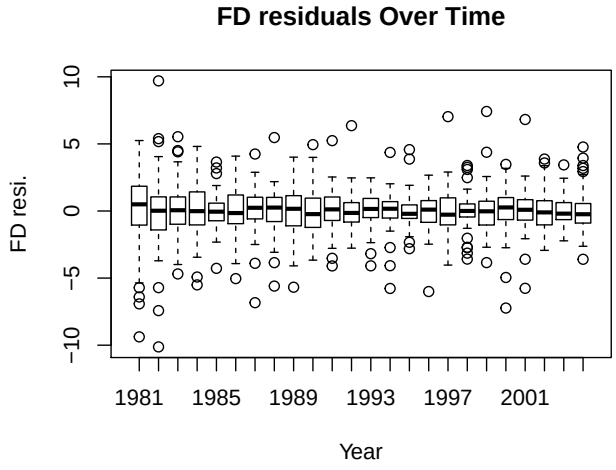


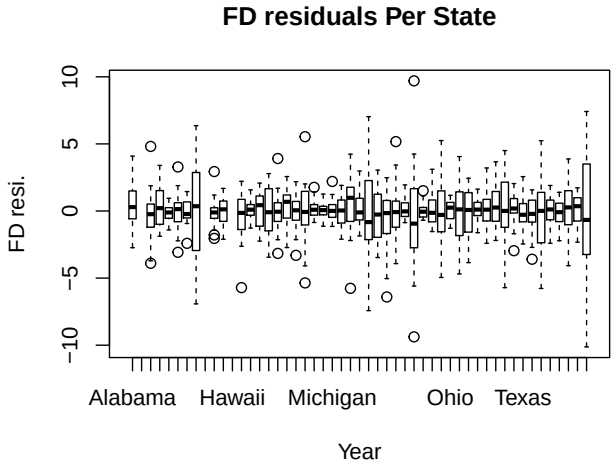

```
FE.resid.df = data.frame(state.name = df2$state.name, year = as.numeric(df2$year),
    FE.resid)

boxplot(FE.resid ~ df2$year, main = "FE residuals Over Time",
    xlab = "Year", ylab = "FE resi.")
boxplot(FE.resid ~ df2$state.name, main = "FE residuals per State",
    xlab = "Year", ylab = "FE resi.")
```






```
FD.resid.df = data.frame(state.name = df2$state.name[-seq(1,
    1176, by = 25)], year = df2$year[-seq(1, 1176, by = 25)],
    FD.resid)

boxplot(FD.resid ~ FD.resid.df$year, main = "FD residuals Over Time",
    xlab = "Year", ylab = "FD resi.")
boxplot(FD.resid ~ FD.resid.df$state.name, main = "FD residuals Per State",
    xlab = "Year", ylab = "FD resi.")
```





We have covered above that within (FE) estimator is more appropriate than first differenced (FD) estimator in our case because the variables of interest don't have much variation across time. Furthermore, the serial correlation test on residuals of the FD model showed negative serial correlation and suggests that u_{it} in the untransformed model is not a random walk. The FE model residuals is also an AR process, for a more efficient estimator we will have to apply GLS on top of the FE results. Lastly, we deem FE estimator more appropriate than Pooled OLS because:

- From the coefficient differences, we know that heterogeneity bias exists in the Pooled OLS model. Unobserved time-invariant variables need to be demeaned away.
- The Test for Poolability doesn't work for our model specification because it has a minimum requirement of $T > k + 1$, where T is the number of periods in each panel and k is the number of parameters to estimate. Since our models include the time intercepts it's not possible. Instead, we test poolability of the models without the year dummies. We acknowledge without the year dummies, the year effect would be absorbed into the time variant variables and idiosyncractic errors (not necessarily desirable). The test results strongly reject the null hypothesis that the dataset is poolable.

```
fmla.pooltest <- as.formula(paste("totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gd

FDnp = pvcm(fmla.pooltest, data = df2, model = "within")
FDp = plm(fmla.pooltest, data = df2)
pooltest(FDp, FDnp)
```

```
##
##  F statistic
##
## data:  fmla.pooltest
## F = 3.1498, df1 = 470, df2 = 672, p-value < 2.2e-16
## alternative hypothesis: unstability
```

```
fmla3 <- as.formula(paste("totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl + perc
```

5. Would you perfer to use a random effects model instead of the fixed effects model you build in *Exercise 4*? Why? Why not?

No, because we need stronger assumptions and we get biased estimates. We'd need to assume every unobserved state-level effect is uncorrelated with every explanatory variable, which is probably not true in this case. For example, certain states may have more aggressive drivers in general or more treacherous road conditions, which may affect the relevant laws that are enacted (some of which are included in this model).

Another key assumption for RE estimator is that we have large cross-section (large N) and short panel (small T). Our cross-section is decent but panel is rather long with 25 years. So we should remain skeptical about the RE results.

We estimate the random effect model below with the same variables. Notice that for the same variables are estimated to be significant and the standard errors estimates are similar. Further examination of the RE regression output tells us that $\theta = 0.8439$ which is close to 1 . We should understand that θ comes from the transformed RE equation:

$$y_{it} - \theta\bar{y_i} = \beta_0(1-\theta) + \beta_1(x_{it1} - \theta\bar{x_{i1}}) + \beta_2(x_{it2} - \theta\bar{x_{i2}}) + ... + \beta_k(x_{itk} - \theta\bar{x_{ik}}) + (v_{it} - \theta\bar{v_i})$$

where v_{it} is the composite form of unobserved effect a_i (not correlated with $x_i t$ for all t) and u_{it}. θ here refers to an adjustment applied to correct for serial correlation in the RE model. If θ is 1 then the RE estimator is essentially the same as FE estimator (the partial time demeaning in RE becomes full time demeaning in FE). In our case, since θ is close to one, the RE results are close to the FE results. What this suggests is that unobserved, time invariant effects are mild but still exist. Again, the FE model is still more appropriate.

```
plm.re <- plm(fmla3.1, data = df.panel, model = "random")

plm.re$ercomp
```

```
##                  var std.dev share
## idiosyncratic 4.064   2.016 0.385
## individual    6.505   2.551 0.615
## theta:  0.8439
```

```
se.lm2 = lmtest::coeftest(lm2)[, "Std. Error"]
se.plm1.1 = lmtest::coeftest(plm1.1)[, "Std. Error"]
se.plm.re = lmtest::coeftest(plm.re)[, "Std. Error"]
```

```
stargazer(lm2, plm1.1, plm.re, type = "latex", omit = "d[0-9][0-9]",
    se = list(se.lm2, se.plm1.1, se.plm.re), omit.labels = "Year Dummies Included?",
    omit.yes.no = c("Yes", "No"), column.labels = c("OLS", "Within",
```

```
        "Random Effects"), star.cutoffs = c(0.05, 0.01, 0.001),
    model.names = F)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 20, 2017 - 22:14:03

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? Be sure to interpret the estimate as if explaining to a layperson.

If *vehicmilespc* increases by 1,000, assuming all other factors stay the same, then the average fatality rate per state would increase by 1 fatality per year per 100,000 people. Interpretation and significance is same for both Fixed Effect and Random Effect models.

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

If there is serial correlation in the idiosyncratic errors and the series is mean stationary, our coefficient estimates would remain consistent and unbiased, given the FE.4 assumption 1-4 mentioned above. But without adjustment to this serial correlation, we are not able to take advantage of the serial correlation information. Therefore our estimator is either too progressive (gives standard errors that are too low), or inefficient (gives standard errors that are higher than necessary) enough. If there is heteroskedasticity, in most cases our estimators yield standard errors that are too low.

The size of standard errors affects t-test results, since we depend on it to reject the null hypotheses that variables have coefficients that are zero. Having standard errors higher than necessary will increase type II errors (false negatives), while lower than necessary will increase type I errors (false positive).

These two conditions are based on FE assumption 5 and 6. Together with assumption FE 7, or a large sample size for asymptotic properties, we can trust t-test results for our estimates.

Let's revisit how these assumptions apply to our models:

Serial correlation : In question 4, we have tested that the FD model errors are serially correlated and the autocorrelation is negative. This is a sign that our errors in the untransformed model are serially uncorrelated and we have over differenced the errors. But on the other hand, the FE model also produced residuals that have positive serial correlation. It says that both the FE and FD estimators are not efficient enough. The RE model, for the reasons mentioned above, gave similar test results and correlation plots as the FE model (See below). For a more efficient and unbiased estimator, we will have to try applying GLS on top of the FE estimator in the future.

```
cat("Random Effects (RE) model : \n")
```

```
## Random Effects (RE) model :
```

```
pbgtest(plm.re)  # fd model
```

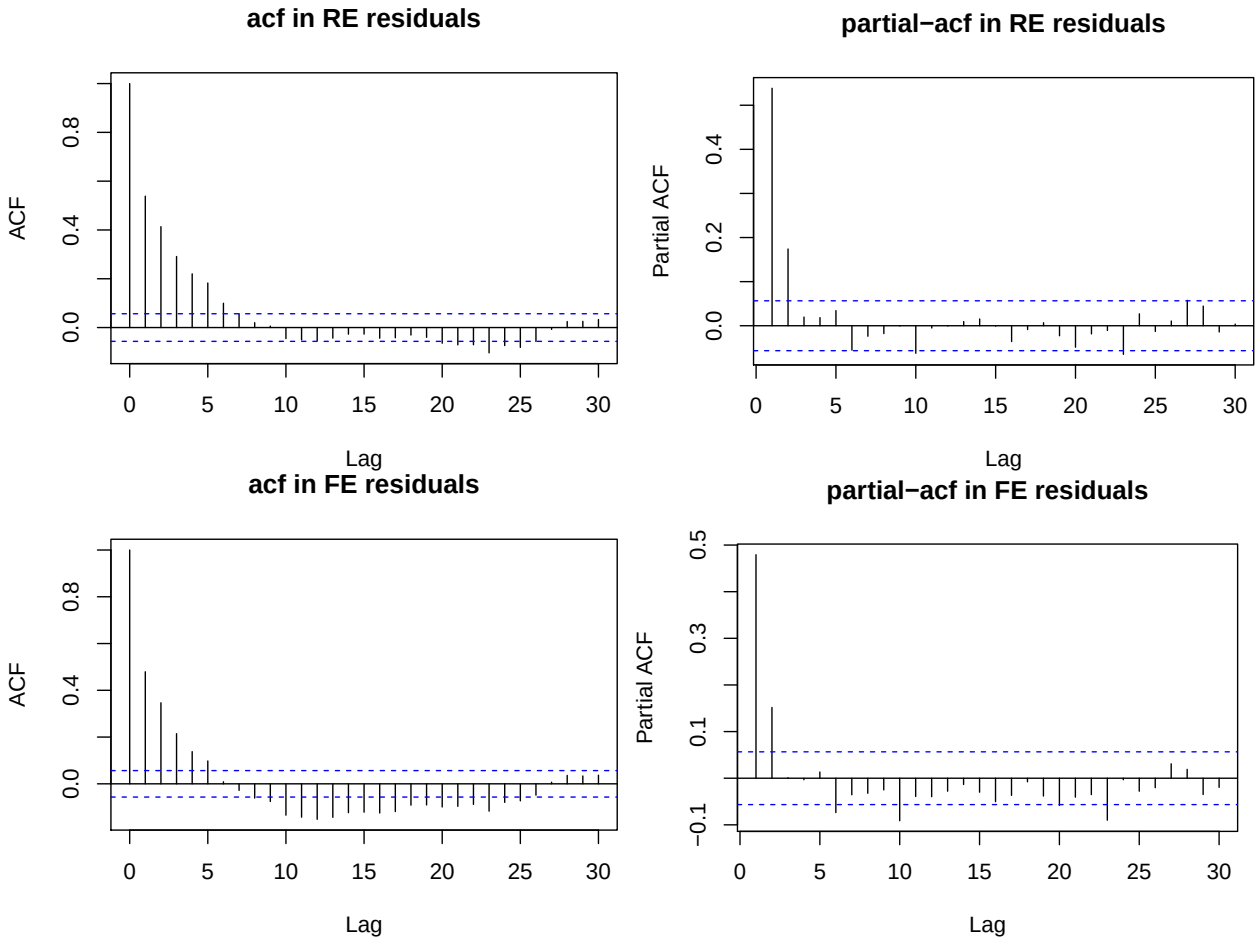
```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  fmla3.1
## chisq = 395.6, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
acf(plm.re$residuals, main = "")
title("acf in RE residuals")
pacf(plm.re$residuals, main = "")
title("partial-acf in RE residuals")
acf(plm1.1$residuals, main = "")
```

Table 3:

	Dependent variable:		
	totfatrte		
	OLS	Within	Random Effects
	(1)	(2)	(3)
bac08	−2.605***	−1.317***	−1.508***
	(0.534)	(0.396)	(0.410)
bac10	−1.446***	−0.959***	−1.075***
	(0.393)	(0.270)	(0.280)
perse	−0.537	−1.218***	−1.138***
	(0.296)	(0.233)	(0.241)
sbprim	−0.354	−1.168***	−1.108**
	(0.490)	(0.343)	(0.356)
sbsecon	−0.147	−0.297	−0.299
	(0.427)	(0.252)	(0.264)
sl70plus	3.204***	0.033	0.131
	(0.443)	(0.270)	(0.282)
gdl	−0.380	−0.388	−0.359
	(0.523)	(0.293)	(0.307)
perc14_24	0.180	0.160	0.178
	(0.122)	(0.096)	(0.099)
log(unem)	5.088***	−3.664***	−2.977***
	(0.481)	(0.393)	(0.404)
vehicmilespc	0.003***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)
Constant	−7.878**		19.228***
	(2.621)		(2.272)
Year Dummies Included?	Yes	Yes	Yes
Observations	1,200	1,200	1,200
R^2	0.613	0.626	0.598
Adjusted R^2	0.602	0.598	0.587
Residual Std. Error	4.018 (df = 1165)		
F Statistic	54.310*** (df = 34; 1165)	54.932*** (df = 34; 1118)	51.055*** (df = 34; 1165)
Note:			*p<0.05; **p<0.01; ***p<0.001

```
title("acf in FE residuals")
pacf(plm1.1$residuals, main = "")
title("partial-acf in FE residuals")
```



Heteroskedasticity : In question 4, we have also seen heteroskedasticity in both the FE and FD models. We have also provided clustered-robust(white) standard errors for corrections. Acknowleding that our errors are both serially correlated and heteroskedastic in both the FE and FD model, here we used the "arellano" method for standard error adjustment. As per the help document of "vcovHC" function, method parameter "white1" allows for general heteroskedasticity but no serial (cross-sectional) correlation;"white2" is "white1" restricted to a common variance inside every group (time period); "arellano" allows a fully general structure w.r.t. heteroskedasticity and serial (cross-sectional) correlation. So we used "arellano" method here.

The overall clustered-robust(white) standard errors of each variable has increased. Also note that when using robust standard error, *bac08* pvalue in the FE model becomes 0.099 so it's insignificant at 5% level of significance.

Also, if we set the cut-off level of p-value at 1%(0,01), all law related explanatory variables (*bac08*, *bac10*, *perse*, *sbprim*) in the FD model are insignificant. So after using robust standard error, the results become more conservative.

```
cat("Within (FE) model: \n")
```

```
## Within (FE) model:
```

```
options(width = 200)
robust <- round(coeftest(plm1.1, plm::vcovHC(plm1.1, method = "arellano",
```

```
    type = "HC0")), digits = 3)
# convert coeftest object to data frame
robust_df <- tidy(robust)
plm1.1_df <- tidy(round(coeftest(plm1.1), digits = 3))
# drop unecessary column
robust_df <- within(robust_df, rm("statistic"))
plm1.1_df <- within(plm1.1_df, rm("statistic"))
# merge data frame for easy comparison
merge_df <- merge(plm1.1_df, robust_df, by = "term")
colnames(merge_df) <- c(" ", "coef.FE", "se.FE", "pvalue.FE",
    "coef.Robust", "se.Robust", "pvalue.Robust")
stargazer(merge_df, type = "text", summary = FALSE)
```

```
##
## ================================================================================
##                 coef.FE se.FE pvalue.FE coef.Robust se.Robust pvalue.Robust
## --------------------------------------------------------------------------------
## 1     bac08     -1.317  0.396   0.001     -1.317      0.797       0.099
## 2     bac10     -0.959  0.270     0       -0.959      0.488       0.050
## 3      d00      -10.748 0.876     0       -10.748     1.287         0
## 4      d01      -10.162 0.883     0       -10.162     1.408         0
## 5      d02      -9.294  0.887     0       -9.294      1.421         0
## 6      d03      -9.294  0.894     0       -9.294      1.469         0
## 7      d04      -9.784  0.918     0       -9.784      1.604         0
## 8      d81      -1.578  0.413     0       -1.578      0.425         0
## 9      d82      -3.346  0.433     0       -3.346      0.429         0
## 10     d83      -3.877  0.447     0       -3.877      0.445         0
## 11     d84      -4.425  0.466     0       -4.425      0.432         0
## 12     d85      -4.884  0.486     0       -4.884      0.454         0
## 13     d86      -3.898  0.519     0       -3.898      0.581         0
## 14     d87      -4.583  0.559     0       -4.583      0.668         0
## 15     d88      -5.105  0.609     0       -5.105      0.737         0
## 16     d89      -6.432  0.647     0       -6.432      0.834         0
## 17     d90      -6.472  0.670     0       -6.472      0.888         0
## 18     d91      -7.147  0.684     0       -7.147      0.957         0
## 19     d92      -8.034  0.704     0       -8.034      1.042         0
## 20     d93      -8.348  0.718     0       -8.348      1.076         0
## 21     d94      -8.786  0.739     0       -8.786      1.051         0
## 22     d95      -8.596  0.763     0       -8.596      1.138         0
## 23     d96      -8.991  0.804     0       -8.991      1.124         0
## 24     d97      -9.211  0.831     0       -9.211      1.181         0
## 25     d98      -9.945  0.849     0       -9.945      1.191         0
## 26     d99      -10.152 0.861     0       -10.152     1.299         0
## 27     gdl      -0.388  0.293   0.185     -0.388      0.367       0.290
## 28  log(unem)   -3.664  0.393     0       -3.664      0.733         0
## 29  perc14_24    0.160  0.096   0.094      0.160      0.167       0.339
## 30    perse     -1.218  0.233     0       -1.218      0.427       0.004
## 31    sbprim    -1.168  0.343   0.001     -1.168      0.552       0.034
## 32   sbsecon    -0.297  0.252   0.239     -0.297      0.370       0.423
## 33   sl70plus    0.033  0.270   0.903      0.033      0.552       0.953
## 34 vehicmilespc  0.001    0       0        0.001        0         0.006
## --------------------------------------------------------------------------------
```

```r
cat("First Differenced (FD) model: \n")
```

```
## First Differenced (FD) model:
```

```r
options(width = 200)
robust <- round(coeftest(plm1, plm::vcovHC(plm1, method = "arellano",
    type = "HC0")), digits = 3)
# convert coeftest object to data frame
robust_df <- tidy(robust)
plm1_df <- tidy(round(coeftest(plm1), digits = 3))
# drop unecessary column
robust_df <- within(robust_df, rm("statistic"))
plm1_df <- within(plm1_df, rm("statistic"))
# merge data frame for easy comparison
merge_df <- merge(plm1_df, robust_df, by = "term")
colnames(merge_df) <- c(" ", "coef.FD", "se.FD", "pvalue.FD",
    "coef.Robust", "se.Robust", "pvalue.Robust")
stargazer(merge_df, type = "text", summary = FALSE)
```

```
##
## ================================================================================
##                coef.FD se.FD pvalue.FD coef.Robust se.Robust pvalue.Robust
## --------------------------------------------------------------------------------
## 1  (intercept)  -0.213  0.094   0.024     -0.213      0.087       0.014
## 2     bac08     -0.819  0.587   0.163     -0.819      0.545       0.134
## 3     bac10     -1.018  0.445   0.022     -1.018      0.395       0.010
## 4      d00      -0.969  0.619   0.118     -0.969      0.410       0.018
## 5      d01      -0.528  0.514   0.305     -0.528      0.340       0.121
## 6      d02       0.125  0.407   0.759      0.125      0.304       0.682
## 7      d03       0.106  0.283   0.709      0.106      0.262       0.687
## 8      d81      -1.215  0.269     0       -1.215      0.425       0.004
## 9      d82      -2.899  0.415     0       -2.899      0.381         0
## 10     d83      -2.827  0.515     0       -2.827      0.327         0
## 11     d84      -2.274  0.598     0       -2.274      0.399         0
## 12     d85      -2.115  0.669   0.002     -2.115      0.404         0
## 13     d86      -0.572  0.731   0.435     -0.572      0.524       0.275
## 14     d87      -0.243  0.807   0.763     -0.243      0.535       0.649
## 15     d88       0.244  0.890   0.784      0.244      0.582       0.676
## 16     d89      -0.349  0.960   0.716     -0.349      0.627       0.578
## 17     d90      -0.033  0.993   0.974     -0.033      0.655       0.960
## 18     d91      -0.694  1.008   0.491     -0.694      0.619       0.263
## 19     d92      -1.253  1.005   0.213     -1.253      0.546       0.022
## 20     d93      -1.103  0.984   0.262     -1.103      0.557       0.048
## 21     d94      -1.013  0.955   0.289     -1.013      0.626       0.106
## 22     d95      -0.367  0.928   0.693     -0.367      0.593       0.536
## 23     d96      -0.528  0.891   0.554     -0.528      0.583       0.366
## 24     d97      -0.348  0.840   0.678     -0.348      0.565       0.538
## 25     d98      -0.729  0.770   0.344     -0.729      0.546       0.182
## 26     d99      -0.684  0.686   0.319     -0.684      0.548       0.212
## 27     gdl      -0.203  0.370   0.583     -0.203      0.256       0.428
## 28  log(unem)   -1.549  0.484   0.001     -1.549      0.446       0.001
## 29  perc14_24    0.904  0.311   0.004      0.904      0.321       0.005
## 30    perse     -0.616  0.390   0.114     -0.616      0.377       0.102
## 31    sbprim    -0.345  0.482   0.474     -0.345      0.446       0.439
```

```
## 32   sbsecon    -0.311  0.295   0.293     -0.311      0.331       0.348
## 33   sl70plus    0.335  0.564   0.553      0.335      0.504       0.507
## 34 vehicmilespc    0      0     0.074        0          0         0.116
## ------------------------------------------------------------------------
```

Conclusion:

To answer the question **"Do changes in traffic laws affect traffic fatalities?"**, we decide to defer to the within (FE) estimator because it removes unobserved, state fixed, time invariant variables more effectively than the pooled OLS or RE estimators. Compared to the first differenced (FD) estimator it is more efficient because our variables of interest don't vary much over time. However, we acknowledge that the FE, FD and RE estimated errors all show serial correlations. For a more efficient estimator, we suggest applying GLS on the FE model with more advanced R packages.

From the FE results, coefficients for blood alcohol limits (both 0.08 and 0.10), per se law and primary seatbelt law were statistically and practically significant. Assuming that no other time-variant variables are missing, we have statistical evidence that enforcing these particular traffic laws can reduce traffic fatalities.

Secondly, we don't have strong enough statistical evidence to support the effect secondary seatbelt law, high speed limit and graduate driver's license laws on traffic fatalities. Thirdly, unemployment rate has statistically and practically significant coefficient while vehicle miles per capita has statistically but only moderately practically significant coefficient. It's interesting to see how these economics related variables also have effects on traffic fatalities.

Note that our conclusion cannot be generalized to Alaska, Hawaii and D.C. nor other traffic laws.