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

U.S. traffic fatalities: 1980-2004

In this lab, you are asked to answer the question **"Do changes in traffic laws affect traffic fatalities?"** To do so, you will conduct the tasks specified below using the data set *driving.Rdata*, which includes 25 years of data that cover changes in various state drunk driving, seat belt, and speed limit laws.

Specifically, this data set contains data for the 48 continental U.S. states from 1980 through 2004. Various driving laws are indicated in the data set, such as the alcohol level at which drivers are considered legally intoxicated. There are also indicators for "per se" laws—where licenses can be revoked without a trial—and seat belt laws. A few economics and demographic variables are also included. The description of the each of the variables in the dataset is come with the dataste.

```
library(foreign)
library(gplots)
library(ggplot2)
library(dplyr)
library(corrplot)
library(lattice)
library(plm)
library(viridis)
library(tsibble)
library(forecast)
library(tidyverse)
library(gridExtra)
```

Exercises:

1. (30%) Load the data. Provide a description of the basic structure of the dataset, as we have done throughout the semester. Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable *totfatrte* and the potential explanatory variables. You need to write a detailed narrative of your observations of your EDA. *Reminder: giving an "output dump" (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive a zero in this exercise.*

```
#driving <- miceadds::load.Rdata2( filename="driving.Rdata")

#load data
load('driving.RData')
driving.df <- data

#
dim(driving.df)
```

```
## [1] 1200   56
```

```r
#describe the variables
#desc

#examine the dat
rbind(head(data,2),tail(data,2))
```

```
##      year state sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl
## 1    1980     1    1    0    0    0      0        0     18       0   0
## 2    1981     1    1    0    0    0      0        0     18       0   0
## 1199 2003    51    0    0    0    1      0        2     21       1   0
## 1200 2004    51    0    0    0    1      0        2     21       1   0
##      bac10 bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm
## 1        1     0     0    940     422     236      3.20      1.437
## 2        1     0     0    933     434     248      3.35      1.558
## 1199     0     1     1    165      62      32      1.79      0.673
## 1200     0     1     1    164      67      31      1.77      0.723
##      wkndfatpvm statepop totfatrte nghtfatrte wkndfatrte vehicmiles unem
## 1         0.803  3893888     24.14      10.84       6.06   29.37500  8.8
## 2         0.890  3918520     24.07      11.08       6.33   27.85200 10.7
## 1199      0.347   501242     32.92      12.37       6.38    9.21788  4.4
## 1200      0.335   507000     32.35      13.21       6.11    9.26600  3.7
##      perc14_24 sl70plus sbprim sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88
## 1         18.9        0      0       0   1   0   0   0   0   0   0   0   0
## 2         18.7        0      0       0   0   1   0   0   0   0   0   0   0
## 1199      15.1        1      0       1   0   0   0   0   0   0   0   0   0
## 1200      14.9        1      0       1   0   0   0   0   0   0   0   0   0
##      d89 d90 d91 d92 d93 d94 d95 d96 d97 d98 d99 d00 d01 d02 d03 d04
## 1      0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 2      0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1199   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
## 1200   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1
##      vehicmilespc
## 1        7543.874
## 2        7107.785
## 1199    18390.080
## 1200    18276.135
```

```r
#check for nulls
#apply(data, 2, function(x) any(is.na(x)))

#check for gaps in panel
table(data$state)
```

```
## 
##  1  3  4  5  6  7  8 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 
## 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
```

```
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

```r
table(data$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
```

The Dataset is panel data, that contains observations about different US states from year 1980 to 2004. There are 1200 observations in total, with 56 columns . The data has 25 observations each, one per year, from 48 continental states except state ids 2,9 and 12 (which we will later identify as Alaska, District Of Columbia and Hawaii that are not part of continental US States). All variables are observed for all states and over all time periods, hence the panel is balanced. Important variables are:

Panel Index
year: 1980 through 2004
state: numeric id of 48 continental states, ordered alphabetically, ranging from 1 to 51.

Dependent Variable
totfatrte: total fatalities per 100,000 population by year by state. Values range from 6.2 to 53.32

Speed Limit Variables
sl55: 1 if speed limit == 55 for the whole year. If the law was in effect only during part of the year, it is set to fractions of 12. This applies for all indicator variables.
sl65: 1 if speed limit == 65
sl70: 1 if speed limit == 70
sl75: 1 if speed limit == 75
slnone: 1 if no speed limit
sl70plus: sl70 + sl75 + slnone

Drinking Laws
minage: minimum drinking age, ranges from 18 years to 21 years.
zerotol: 1 if zero tolerance law was in effect, and 0 if not. If the law was in effect only during part of the year, it is set to fractions of 12.
bac10: 1 if blood alcohol limit .10 in effect, and 0 if not. Fractions used to denote partial years, as above.
bac08: 1 if blood alcohol limit .08 in effect, and 0 if not. Fractions used to denote partial years, as above.
per se: 1 if administrative license revocation (per se law) in effect, and 0 if not. Fractions used to denote partial years, as above.

Seatbelt Laws
sbprim: 1 if primary seatbelt law was in effect, 0 otherswise. There are no fractions in this variable.
sbsecon: 1 if secondary seatbelt law was in effect, 0 otherswise. There are no fractions in this variable.
seatbelt: 0 if none, =1 if primary, =2 if secondary. There are no fractions in this variable.

Age iimit Laws
gdl: 1 if graduated drivers license law was in effect, and 0 if not. Fractions used to denote partial

years, similar to speed limit.

Demographic variables
statepop: state population by year by state. Values range from 453,401 to 35,894,000
vehicmiles: vehicle miles traveled, billions. Values range from 3.7027 to 329.6
unem: unemployment rate, percent. Values range from 3.2 to 18
perc14_24: percent population aged 14 through 24. Values range from 11.7 to 20.3

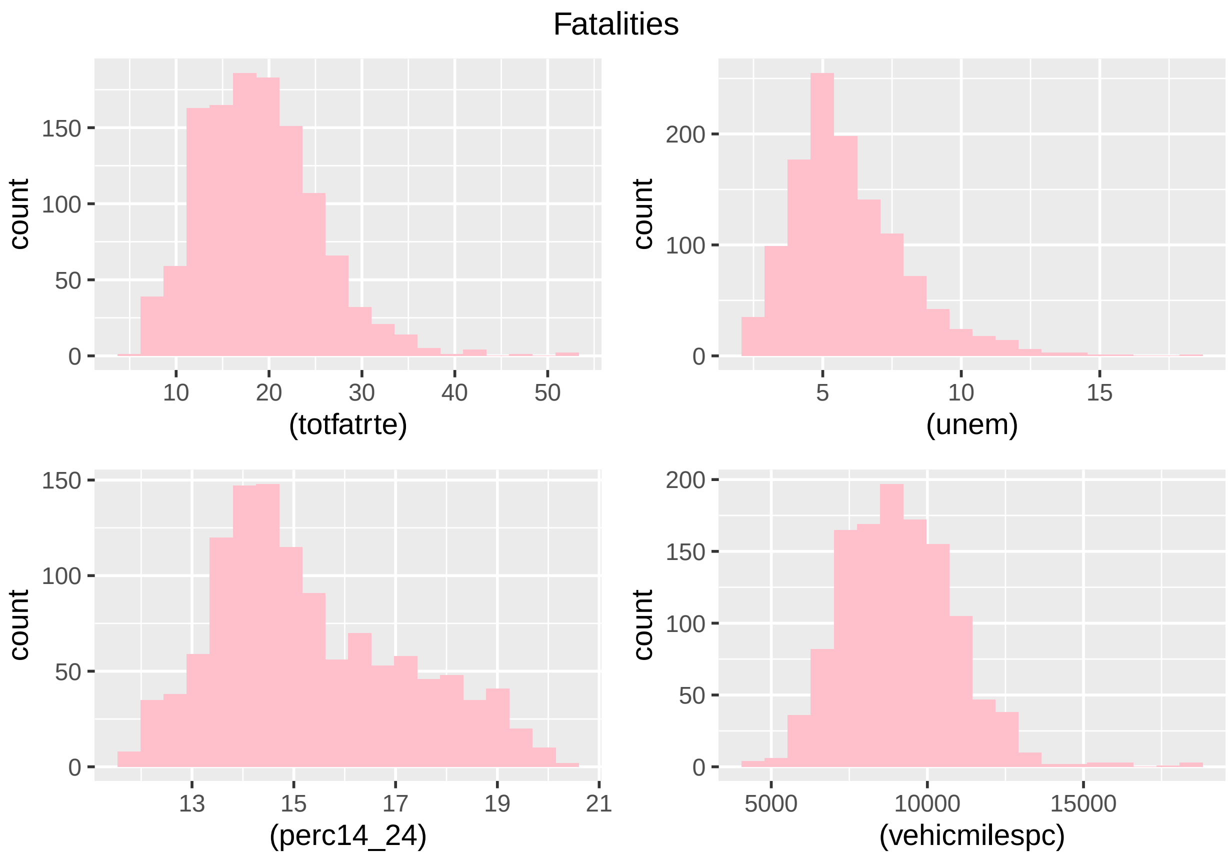
Year Dummy
Dummy variables $d80$ - $d04$ indicating years

It will be useful to add more context around the state information, in addition to the state id. Since we know the id is alphabetical, we get the aphabetical list of US states with two letter abbreviated code, and match with the state variable in fatality data.

```
#get state name
us.states = read.csv("usstates.csv", header = TRUE, sep = ",", dec = ".")

data.with.name <- merge(data, us.states, by=c("state","state"))
```

To start EDA, we perform univariate analyses of important variables fatality rate, unemployment, % of younger population, and vehiclemilespc to examine the distribution.

```
hist1 <- ggplot(driving.df, aes((totfatrte))) + geom_histogram(bins = 20, fill = "pink")

hist2 <- ggplot(driving.df, aes((unem))) + geom_histogram(bins = 20, fill = "pink")

hist3 <- ggplot(driving.df, aes((perc14_24))) + geom_histogram(bins = 20, fill = "pink")

hist4 <- ggplot(driving.df, aes((vehicmilespc))) + geom_histogram(bins = 20, fill = "pink")

grid.arrange(hist1, hist2, hist3, hist4, nrow = 2, ncol = 2, top="Fatalities")
```



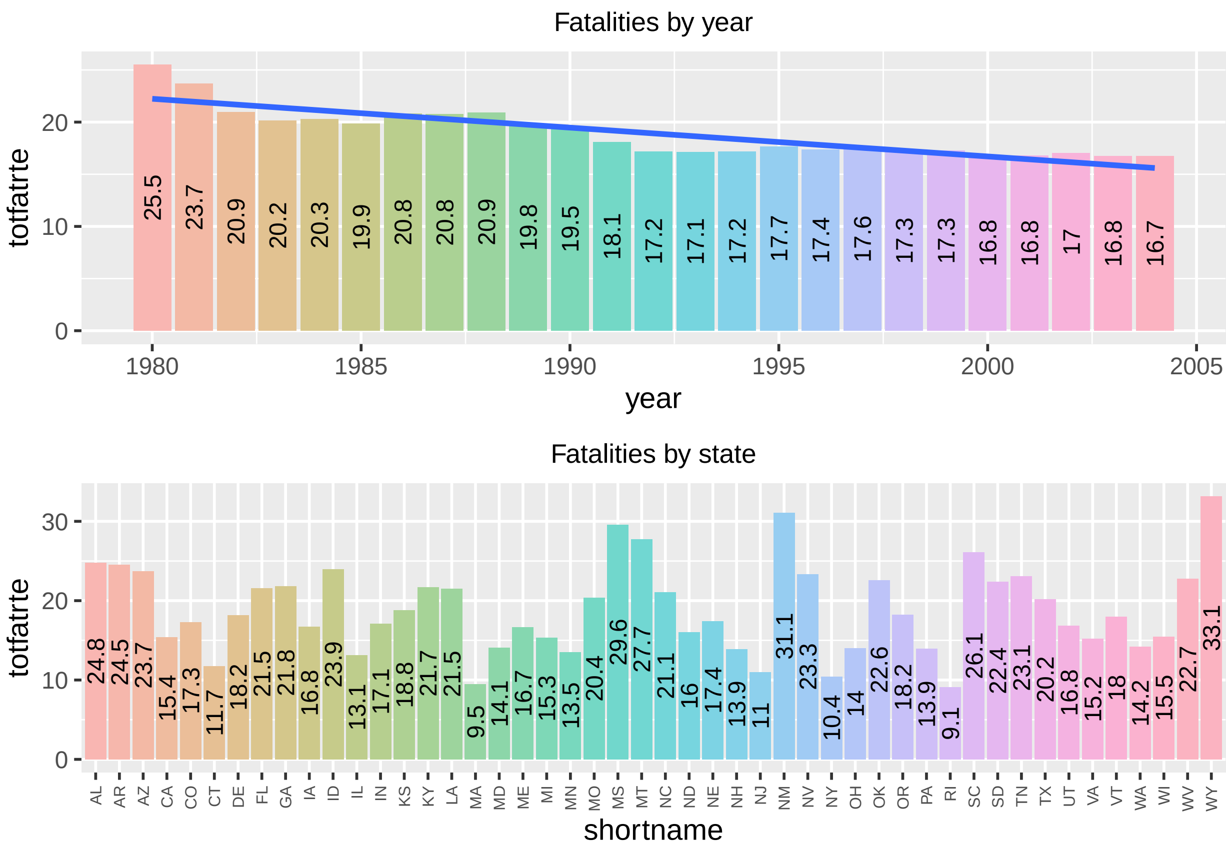
The distribution looks approximately normal with some tail for $totfatrte$, $unem$, and $vehiclemilespc$. It looks normal with higher slope at the top and lower slope at the bottom for $perc14-24$.

Then, determine both the common and individual driving behaviors of US States across time, we'll analyze the aggregate of traffic laws in US across time and across states. Then we'll focus on the fatality progression of top and bottom ranked US states across years. We'll also evaluate how the fatality pattern is different between years 1980 and 2004.

Below we analyze the fatality rate change by year and overall change by state .

```
#fatality change by year
traffic.yearly.aggr <- data %>%   group_by(year) %>%  summarise_at(vars(totfatrte, nghtfatrte,w

#fatality change by state
traffic.state.perc.aggr <- data.with.name %>%
  group_by(shortname) %>%
  summarise_at(vars(totfatrte,nghtfatrte,wkndfatrte), funs(mean))

year.plot <- ggplot(traffic.yearly.aggr, aes(year, totfatrte)) +
  geom_bar(aes(fill = factor(year)), position = "dodge", stat="identity") + ggtitle("Fatalitie
  #geom_abline(intercept, slope, linetype, color, size) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) + geom_text(data = traffic.yearly.agg
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")

state.plot <- ggplot(traffic.state.perc.aggr, aes(shortname, totfatrte)) +
  geom_bar(aes(fill = factor(shortname)), position = "dodge", stat="identity") + ggtitle("Fatal
```

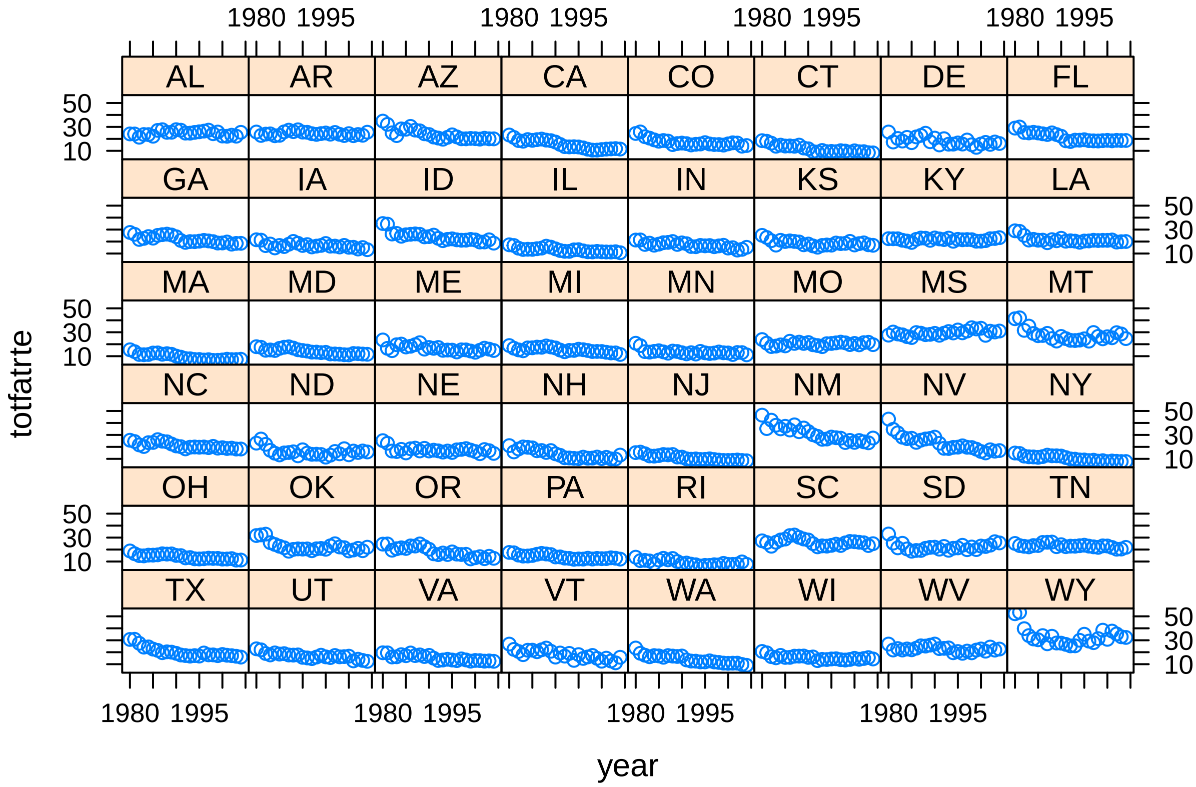
```
  scale_fill_hue(c=45,l=80) + theme(plot.title = element_text(size = 8, hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5, hjust=1)) +
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none") +

grid.arrange(year.plot, state.plot, nrow = 2,  ncol = 1)
```



We can see that the fatality rate is largely decreasing from 1980 to 2004. The fatality rates range from ~9 to ~34. Wyoming, New Mexico, Mississippi, Montana, and South Carolina are the states with highest fatality rates while New York, New Jersey and Rhode Island are the states with lowest fatality rates. The pattern shows that the states with more rural roads have higher fatality rates - the geography and road conditions are thus important omitted variables in the dataset. In addition, the fataility split by cause (drunk driving, speeding) by state by year could be an important predictor. Another omitted variable could be the a measure of compliance to the traffic laws - speed limit, seat belt - at the state level.

Next we analyze how the fatality rates varied over the years, in individual states.

```
xyplot(totfatrte ~ year | shortname, data=data.with.name, as.table=T)
```

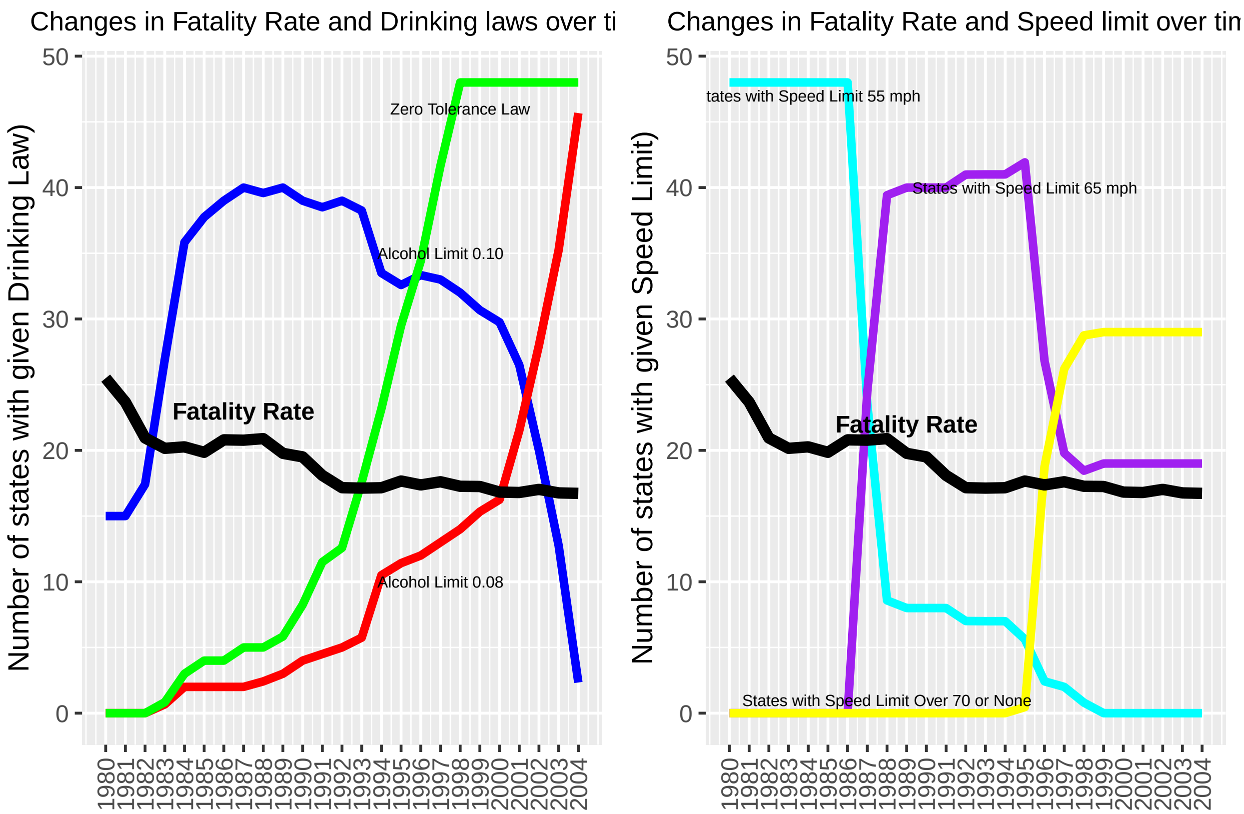


The above xyplot confirms that most of the States shows an overall decrease in the traffic fatality rate, except states like *Mississippi* . One interesting point is that the traffic fatality rate is not dependent on the state area or population - the top 2 states with size and population, Texas and California, are not amoing the top states in traffic fatality rate.

Below, we explore the prevalence of traffic laws over the years. We hypothesize that the fatality rate is influeced the most by drinking and overspeeding and proceed to examine the applicable laws.

```
# summarize the average statistics for blood alcohol in a data frame
bac.df <- data %>% group_by(year) %>%
summarise(bac10 = sum(bac10), bac08 = sum(bac08), zerotol = sum(zerotol), totfatrte = mean(tot

bac.plot <- ggplot(bac.df, aes(x = year)) +
geom_line(aes(y = bac10, color='bac10'), size = 1.5, group = 1) +
  geom_line(aes(y=bac08, color='bac08'), size = 1.5, group = 1) +
  geom_line(aes(y=zerotol, color='zerotol'), size = 1.5, group = 1) +
  geom_line(aes(y=totfatrte, color='totfatrte'), size = 2, group = 1) +
  scale_x_continuous(breaks = seq(min(bac.df$year), max(bac.df$year), 1)) + theme(axis.text.x =
      values = c(
           bac08="red",
           bac10="blue",
           zerotol="green",
           totfatrte="black")) + labs(title = "Changes in Fatality Rate and Drinking laws over
y = "Number of states with given Drinking Law)",
x = "") +
  annotate("text", x = 1987, y = 23, label = "Fatality Rate", size = 3, fontface = "bold") +
  annotate("text", x = 1997, y = 10, label = "Alcohol Limit 0.08", size = 2) +
```

```
  annotate("text", x = 1997, y = 35, label = "Alcohol Limit 0.10", size = 2) +
  annotate("text", x = 1998, y = 46, label = "Zero Tolerance Law", size = 2) + theme(plot.title

sl.df <- data %>% group_by(year) %>%
summarise(sl55 = sum(sl55), sl65 = sum(sl65),sl70plus = sum(sl70plus), totfatrte = mean(totfat

sl.plot <- ggplot(sl.df, aes(x = year)) +
geom_line(aes(y = sl55, color='sl55'), size = 1.5, group = 1) +
  geom_line(aes(y = sl65, color='sl65'), size = 1.5, group = 1) +
  geom_line(aes(y = sl70plus, color='sl70plus'), size = 1.5, group = 1) +
  geom_line(aes(y=totfatrte, color='totfatrte'), size = 2, group = 1) +
  scale_x_continuous(breaks = seq(min(sl.df$year), max(sl.df$year), 1)) + theme(axis.text.x = 
      values = c(
            sl55 = "cyan",
            sl65 = "purple",
           sl70plus="yellow",
           totfatrte="black")) + labs(title = "Changes in Fatality Rate and Speed limit over t
y = "Number of states with given Speed Limit)",
x = "") +
  annotate("text", x = 1989, y = 22, label = "Fatality Rate", size = 3, fontface = "bold") +
  annotate("text", x = 1984, y = 47, label = "States with Speed Limit 55 mph", size = 2) +
   annotate("text", x = 1995, y = 40, label = "States with Speed Limit 65 mph", size = 2) +
   annotate("text", x = 1988, y = 1, label = "States with Speed Limit Over 70 or None", size =
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")

grid.arrange(bac.plot, sl.plot, nrow = 1,  ncol = 2)
```

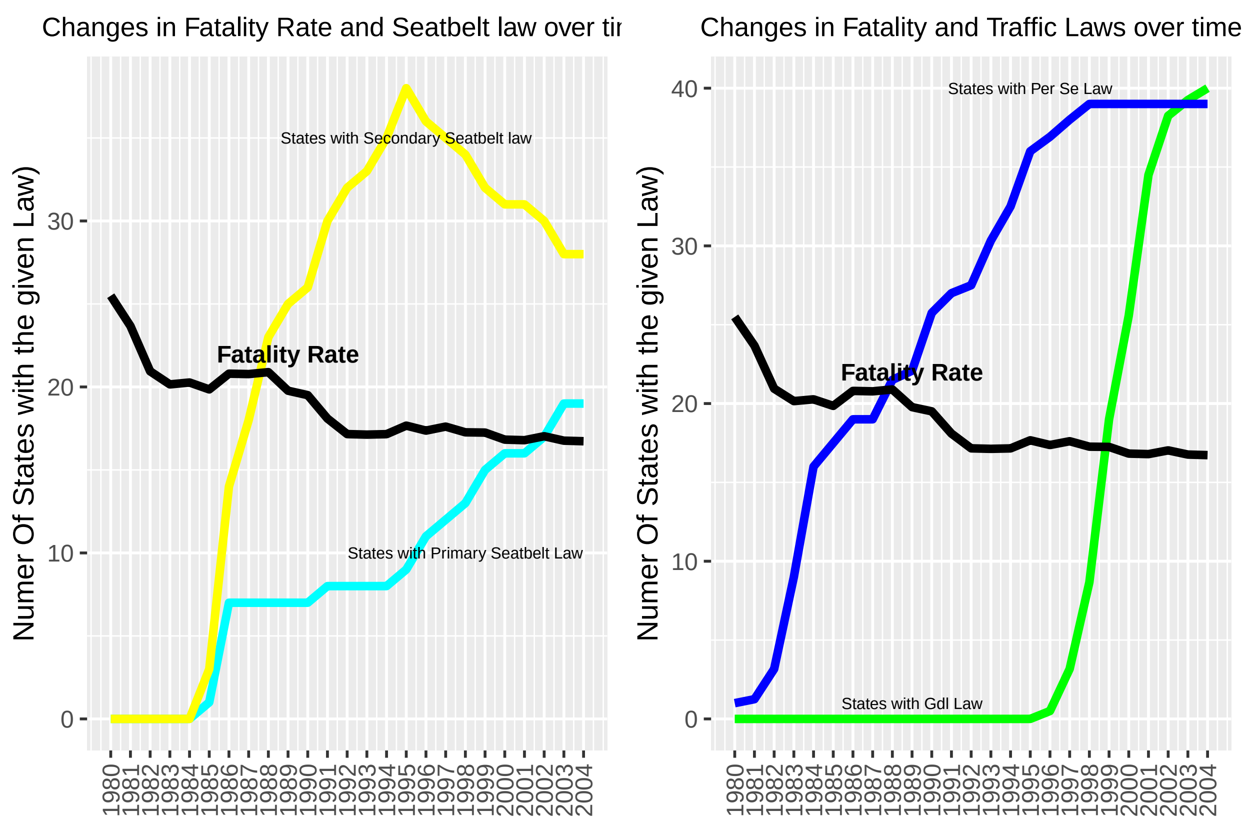


Over the years, more states are adopting stricter alcohol limits. In 2004, over 45 states have a stricter bac limit of 0.08, compared to 0 in 1980. Similarly, all states have adopted the zero tolerance law in 2004 compared to 0 states in 1980. This is consistent with the decrease in fatality rates over time that observed before. Regarding speed limit, states had lower speed limit in 1980 - however, the speed limits were more relaxed in the later years as can be seen by the increase in the number of states with speed limit 70 or above, as seen in the above graph.

Next, we explore how fatality rates vary with the stricter seatbelt requirements, and adoption of “Per Se” and “Graduate DL” license laws.

```
#summarize general laws
glaws.df <- data %>% group_by(year) %>%
summarise(gdl = sum(gdl), perse = sum(perse), sbprim = sum(sbprim), sbsecon = sum(sbsecon), to

glaws.plot <- ggplot(glaws.df, aes(x = year)) +
  geom_line(aes(y=gdl, color='gdl'), size = 1.5, group = 1) +
  geom_line(aes(y=perse, color='perse'), size = 1.5, group = 1) +
  geom_line(aes(y=totfatrte, color='totfatrte'), size = 1.5, group = 1)+
  scale_x_continuous(breaks = seq(min(glaws.df$year), max(glaws.df$year), 1)) + theme(axis.tex
      values = c(
           gdl="green",
           perse="blue",
           totfatrte="black")) + labs(title = "Changes in Fatality and Traffic Laws over time"
y = "Numer Of States with the given Law)",
x = "") +
  annotate("text", x = 1989, y = 22, label = "Fatality Rate", size = 3, fontface = "bold") +
```

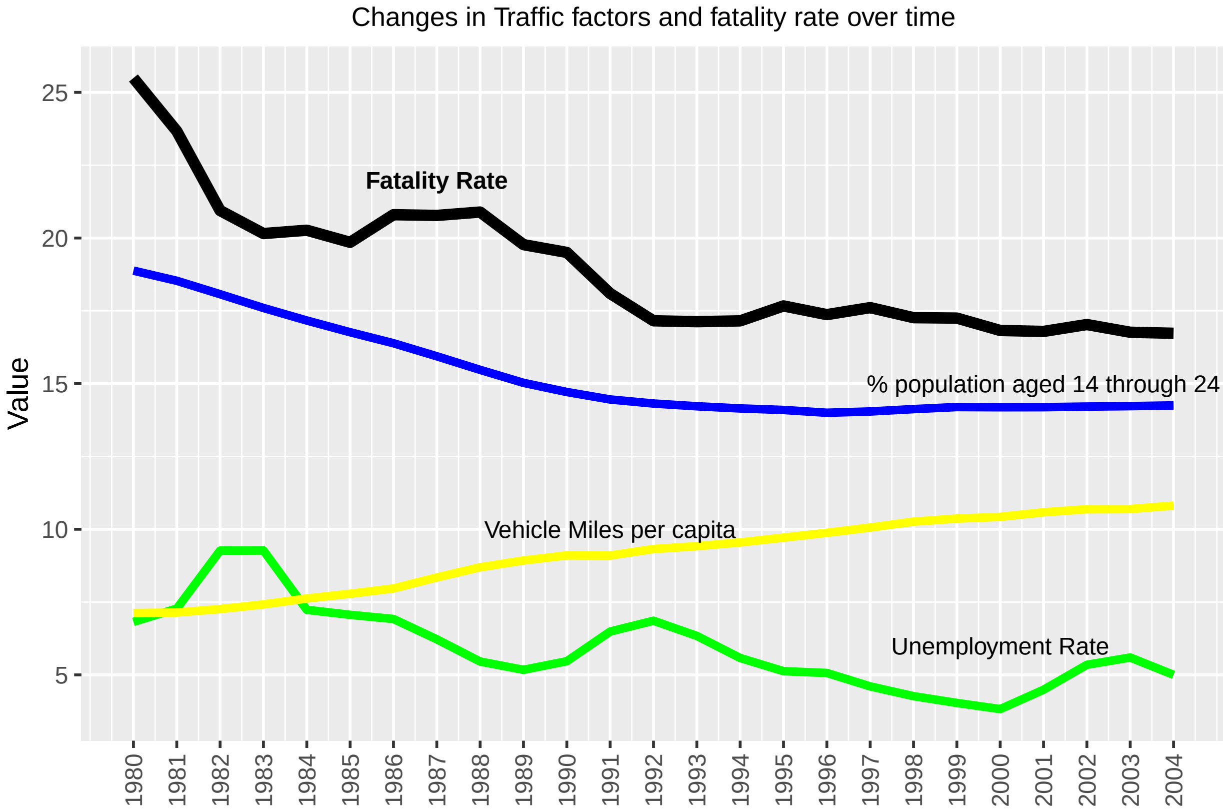
```
  annotate("text", x = 1989, y = 1, label = "States with Gdl Law", size = 2) +
   annotate("text", x = 1995, y = 40, label = "States with Per Se Law", size = 2) +
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")

sb.plot <- ggplot(glaws.df, aes(x = year)) +
  geom_line(aes(y=sbprim, color='sbprim'), size = 1.5, group = 1)+
  geom_line(aes(y=sbsecon, color='sbsecon'), size = 1.5, group = 1)+
  geom_line(aes(y=totfatrte, color='totfatrte'), size = 1.5, group = 1)+
  scale_x_continuous(breaks = seq(min(glaws.df$year), max(glaws.df$year), 1)) + theme(axis.tex
      values = c(
           sbprim="cyan",
           sbsecon="yellow",
           totfatrte="black")) + labs(title = "Changes in Fatality Rate and Seatbelt law over t
y = "Numer Of States with the given Law)",
x = "") +
  annotate("text", x = 1989, y = 22, label = "Fatality Rate", size = 3, fontface = "bold") +
  annotate("text", x = 1998, y = 10, label = "States with Primary Seatbelt Law", size = 2) +
   annotate("text", x = 1995, y = 35, label = "States with Secondary Seatbelt law", size = 2) +
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")

grid.arrange(sb.plot, glaws.plot, nrow = 1,  ncol = 2)
```



We find that states adopt stricter seatbelt laws over the years. We find that more states adopted the *PerSelaw* that mandates the "on the spot" suspension of an accused person's DL based upon a Law Enforcement Officer's belief that the person was operating a motor vehicle with an blood alcohol concentration of .08% or greater. We also see more adoption for *Gdl* law that mandates a multi-step process for new teen drivers to obtain the DL. The above patterns are consitent with the decrease in traffic fatality rates over years.

Then we look at how the pattern of demographic changes over years compares to fatality rates change. We examine the $percentofpopulationaged14-24$, $unemploymentrate$ and $vehicmilespc$.

```
# summarize the average statistics for speed limit in a data frame
demo.df <- data %>% group_by(year) %>%
#summarise(sl55 = sum(sl55), sl65 = sum(sl65),sl70 = sum(sl70),sl75 = sum(sl75), slnone = sum(
summarise(totfatrte = mean(totfatrte), perc14_24 = mean(perc14_24), unem = mean(unem), vehicmil

#perc14_24 + unem + vehicmilespc

sl.plot <- ggplot(demo.df, aes(x = year)) +
   geom_line(aes(x = year, y = perc14_24, color = "perc14_24"), size = 1.5, group = 1) +
  geom_line(aes(x = year, y = unem, color = "unem"), size = 1.5, group = 1) +
  geom_line(aes(x = year, y = vehicmilespc, color = "vehicmilespc"), size = 1.5, group = 1) +
   geom_line(aes(x = year, y = totfatrte, color = "totfatrte"), size = 2, group = 1) +
  scale_x_continuous(breaks = seq(min(sl.df$year), max(sl.df$year), 1)) + theme(axis.text.x = e
      values = c(
           perc14_24="blue",
           unem="green",
           vehicmilespc="yellow",
           totfatrte = "black")) + labs(title = "Changes in Traffic factors and fatality rate o
y = "Value",
x = "")  +
  annotate("text", x = 1987, y = 22, label = "Fatality Rate", size = 3, fontface = "bold") +
  annotate("text", x = 2000, y = 6, label = "Unemployment Rate", size = 3) +
  annotate("text", x = 1991, y = 10, label = "Vehicle Miles per capita", size = 3) +
  annotate("text", x = 2001, y = 15, label = "% population aged 14 through 24", size = 3) + the

sl.plot
```



Vehicle Miles shows an increase over years while % of youngest population and unemployment rate reduced ever so sightly.

Thus we see that some of the traffic and demographic variables show similar patterns as fatality rate and could be important predictors for the fatality rate.

Lets proceed to examine the individual behavior in the panel. First, we'll examine how the traffic fatality rates changed over years for the first 3 States from the top and bottom of the fatality rate.

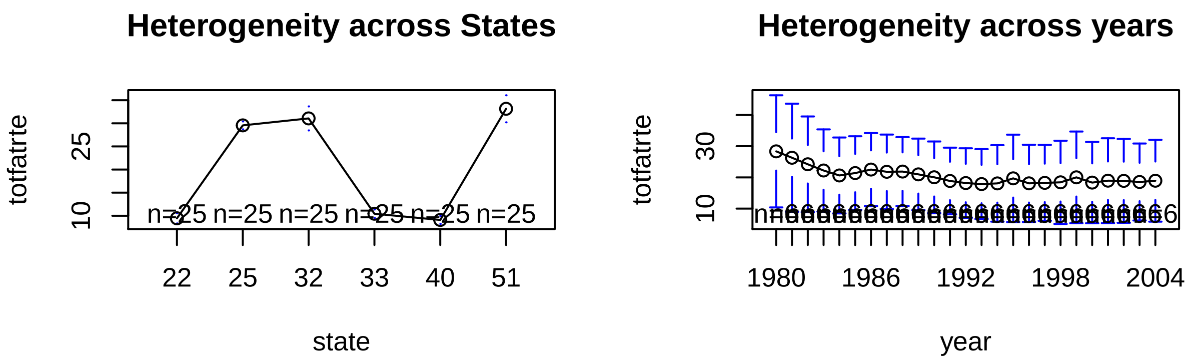
```
#fatality change by state
traffic.state.aggr <- data.with.name %>%
  group_by(shortname) %>%
  summarise_at(vars(totfat, nghtfat,wkndfat), funs(sum))

top.3.fatalities <- traffic.state.perc.aggr %>%
         filter(rank(desc(totfatrte))<=3) %>% arrange(desc(totfatrte))

bottom.3.fatalities <- traffic.state.perc.aggr %>%
         filter(rank((totfatrte))<=3) %>% arrange((totfatrte))

cbind(top.3.fatalities[,1:2],bottom.3.fatalities[,1:2])
```

```
##   shortname totfatrte shortname totfatrte
## 1        WY   33.1408        RI    9.0900
## 2        NM   31.0608        MA    9.4512
## 3        MS   29.5548        NY   10.4380
```

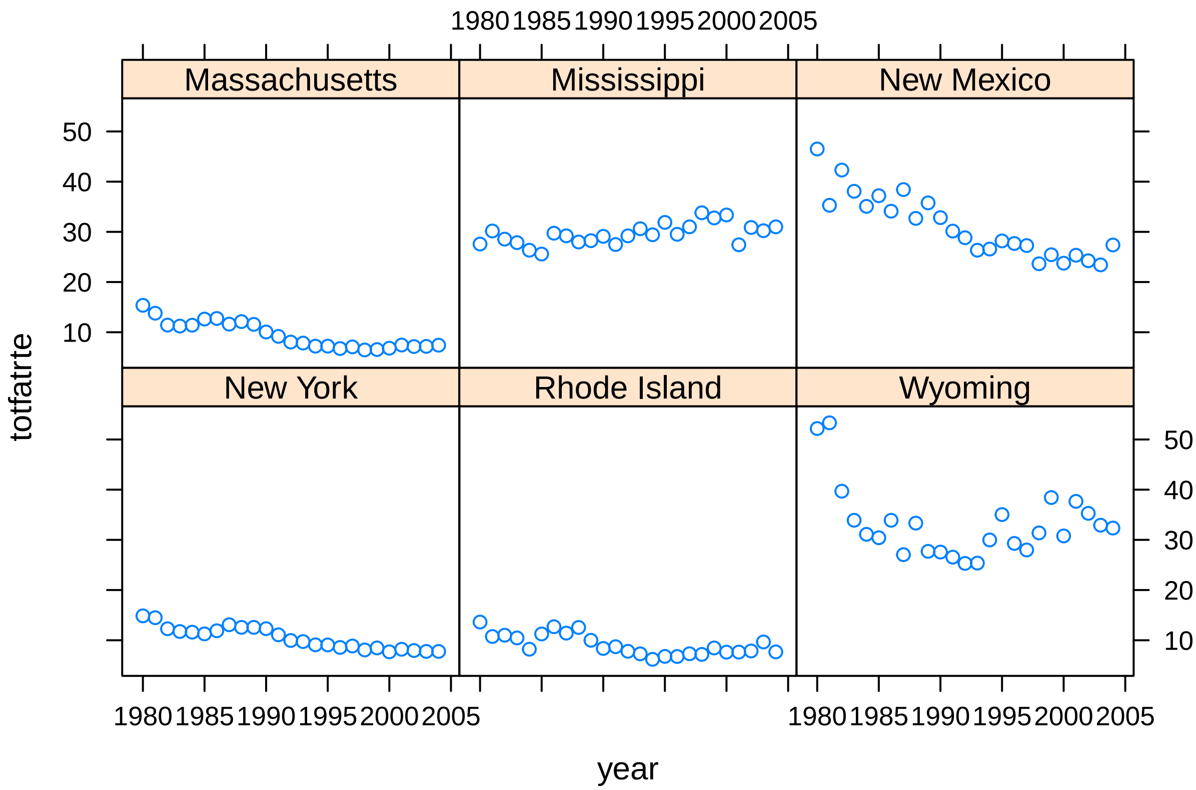
```r
data.top.filtered <- data.with.name %>% filter(shortname %in% c("MS","NM","WY"))
data.bottom.filtered <- data.with.name %>% filter(shortname %in% c("MA","NY","RI"))

#data.with.name %>% filter(shortname %in% c("MS","NM","WY","MA","NY","RI") & year == '2004') %
#data.with.name %>% filter(year == '2004') %>% dplyr::select(year,name,totfatrte) %>% arrange(

data.merged <- union(data.top.filtered,data.bottom.filtered)
```

The top 3 are Wyoming, New Mexico and Mississippi. The bottom 3 are Rhode Island, New York and Massachussets. To put this in context, in 2004, in Wyoming, the probability of dying in a motor vehicle accident is nearly 5 times as high as in Rhode Island, the state with the lowest death rate. Below, we see the average fatality rate for each state across years.

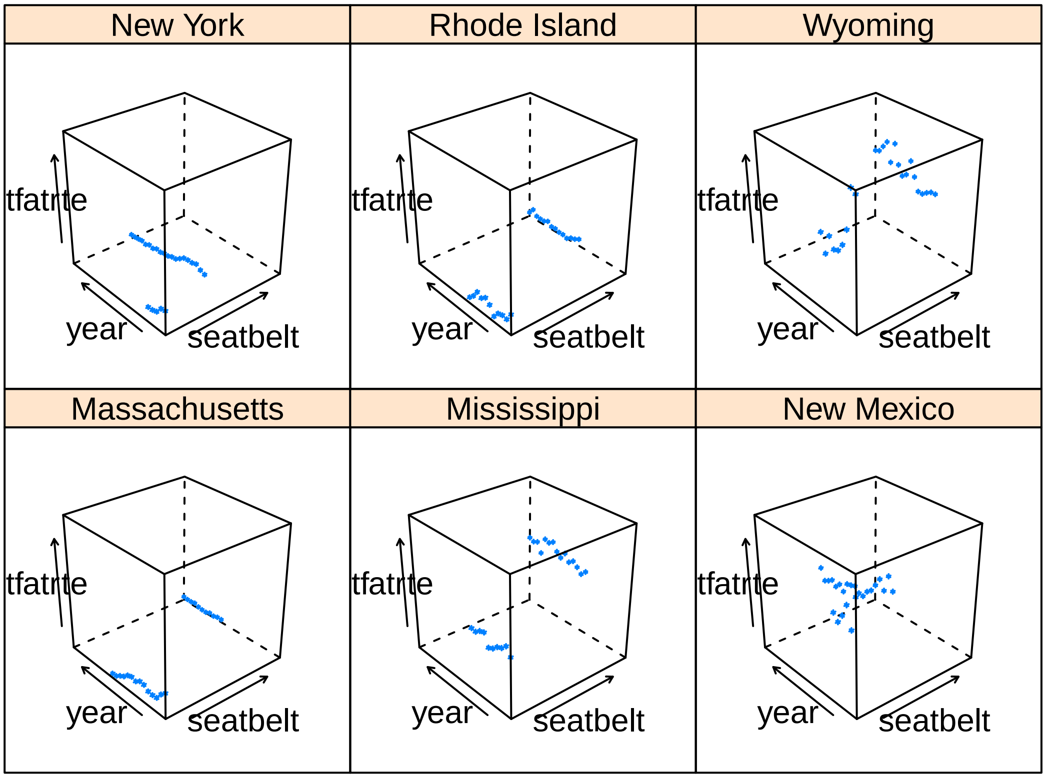
```r
# Conditional Box-plot
conditional_plot = function(data, plotvar, condvar, title) {
  g <- ggplot(data, aes(as.factor(condvar), plotvar, color = as.factor(condvar)))
  g + geom_boxplot() + geom_jitter(width = 0.2) + ggtitle(title) + theme(axis.text.x = element_
}

# yIndex by year (Heterogeineity across year)
cplot.1 <- conditional_plot(data.merged, data.merged$totfatrte, data.merged$year, "Totalfatrte

# yIndex by country (Heterogeineity across countries)
cplot.2 <- conditional_plot(data.merged, data.merged$totfatrte, data.merged$name, "Totalfatrte

#scatterplot(totfatrte ~  shortname|year, boxplots=FALSE, smooth=TRUE, data=data.merged)
#scatterplot(totfatrte ~ df$xIndex | df$country, boxplots=FALSE, xlab="xIndex", ylab="yhat",sm
#abline(lm(data.merged$totfatrte ~ data.merged$shortname),lwd=3, col="blue")

# Heterogeneity across countries
par(mfrow = c(2,2))
plotmeans(totfatrte ~ state, main="Heterogeneity across States", data=data.merged)
plotmeans(totfatrte ~ year, main="Heterogeneity across years", data=data.merged)

#
#coplot(totfatrte ~ year|state, type="l", data=data.merged)
#coplot(totfatrte ~ year|name, type="b", data=data.merged)

par(mfrow = c(2,2))
```



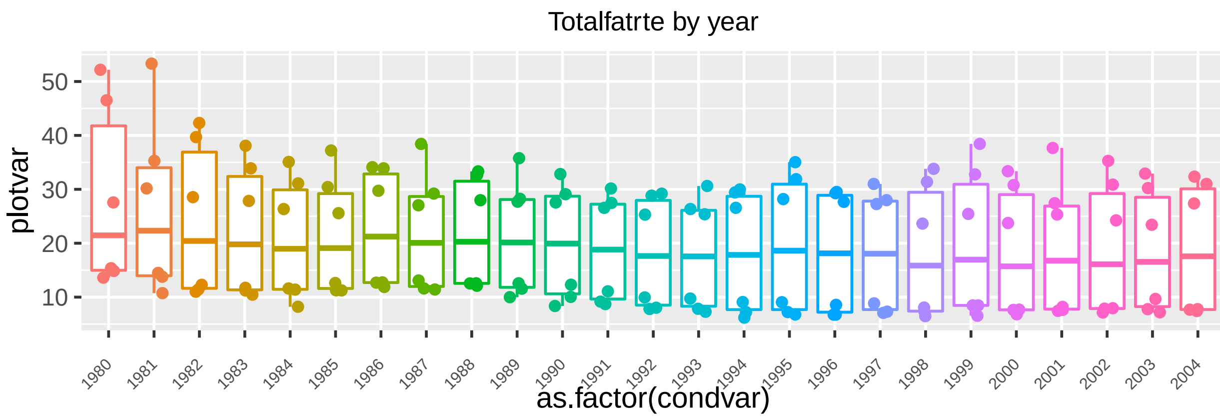

```
xyplot(totfatrte ~ year | name, data=data.merged, as.table=T)
```

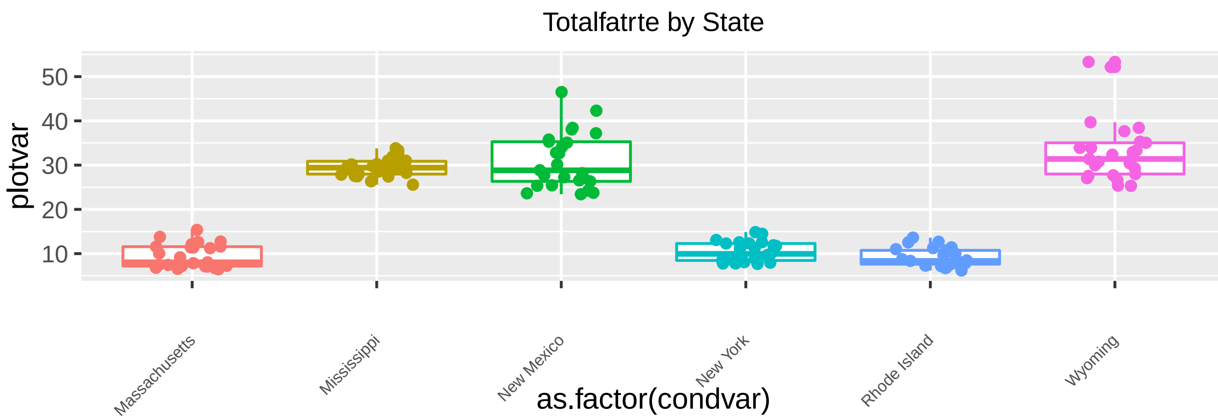



```
# check if we need this 3D plot
library(lattice)
cloud(totfatrte ~ seatbelt + year|name, data = data.merged, auto.key = TRUE)
```




```
# g <- ggplot(data.merged, aes(year, totfatrte, colour = as.factor(shortname)))
# g + geom_line() + ggtitle("Fatality Rate Change for Top and Bottom States") + facet_wrap(~sho

grid.arrange(cplot.1, cplot.2, nrow = 2,  ncol = 1)
```



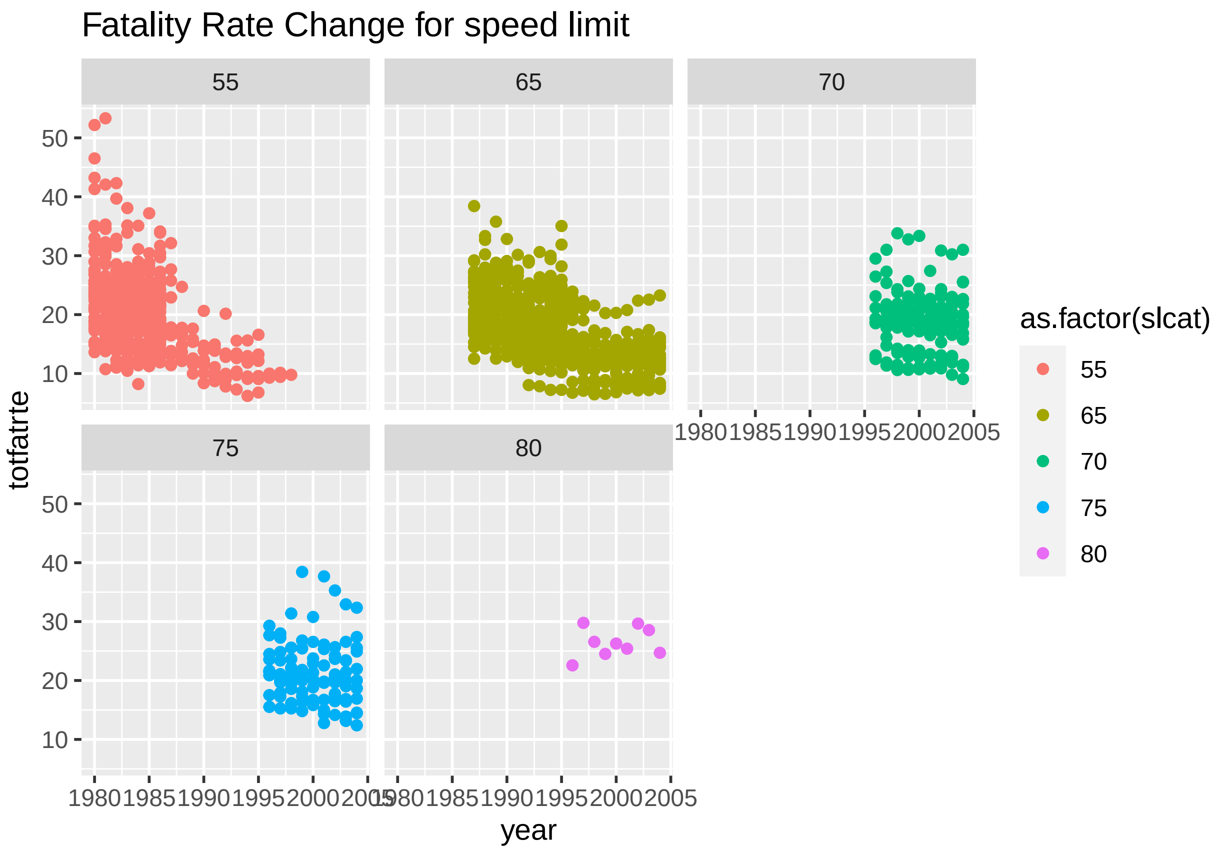


#TODO : The above graphs collectively provide the below information. Some of them are different ways of representing the same info, we need to pick and choose.

We can see that New Mexico and Wyoming has high variance in the data, with NM consistently reducing the traffic fatality rate over years. However, WY reduced the fatality rate from 80's to mid 90's and had a gradual increase after. The bottom 3 states have very low variance across years.

```
df.transformed <- data.with.name %>%
  mutate(
    slcat = case_when(
      sl55 >= 0.5  ~ 55,
      sl65 >= 0.5  ~ 65,
      sl70 >= 0.5  ~ 70,
      sl75 >= 0.5  ~ 75,
      slnone >= 0.5  ~ 80,
      TRUE ~ 0
    ),
    bacat = case_when(
      bac10 >= 0.5  ~ 1,
      bac08 >= 0.5  ~ 2,
      TRUE ~ 0
    )
  )

g <- ggplot(df.transformed, aes(year, totfatrte, colour = as.factor(slcat)))
```

```
g + geom_point() + ggtitle("Fatality Rate Change for speed limit") + facet_wrap(~slcat)
```



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

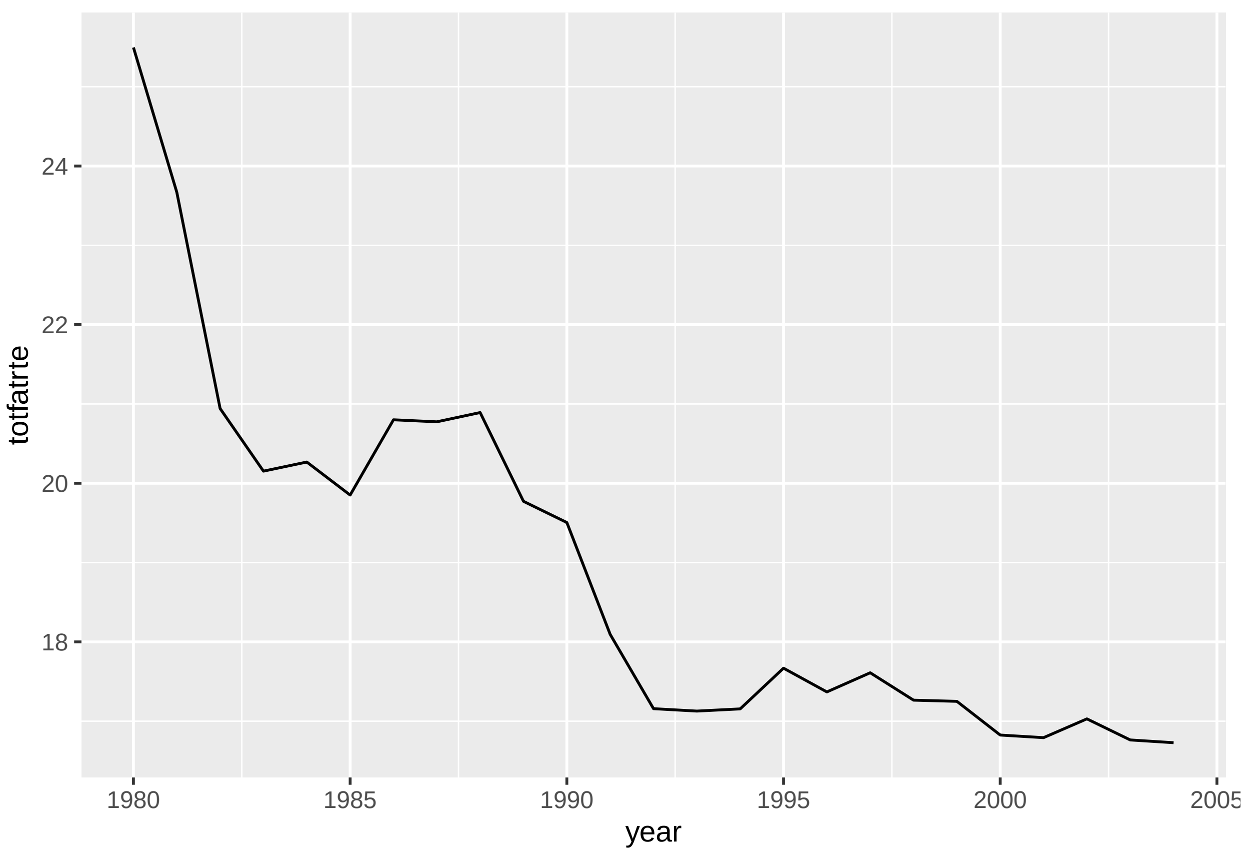
totfatrte defines the total fatalities per 100,000 population.<>

```
yearlyavg <- aggregate(totfatrte~year, driving.df, mean)

# Printing the yearly average for total fatality rate
yearlyavg
```

```
##    year totfatrte
## 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
```

```
# Plotting the yearly total fatality rate
ggplot(yearlyavg) +
  geom_line(
    mapping = aes(x = year, y = totfatrte)
)
```




```
lm.fit1 <- lm(totfatrte ~ d81+d82+d83+d84+d85+d86+d87+d88+d89+
                          d90+d91+d92+d93+d94+d95+d96+d97+d98+d99+
```

```r
                      d00+d01+d02+d03+d04, data=driving.df)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04, data = driving.df)
##
## 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
```

3. (15%) Expand your model in *Exercise 2* by adding variables *bac08, bac10, perse, sbprim,*

sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc, and perhaps *transformations of some or all of these variables*. Please explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed. How are the variables *bac8* and *bac10* defined? Interpret the coefficients on *bac8* and *bac10*. Do *per se laws* have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)

```r
lm.fit2 <- lm(totfatrte ~ d81+d82+d83+d84+d85+d86+d87+d88+d89+
                          d90+d91+d92+d93+d94+d95+d96+d97+d98+d99+
                          d00+d01+d02+d03+d04 + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
##     perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##     vehicmilespc, data = driving.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9160  -2.7384  -0.2778   2.2859  21.4203
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.716e+00  2.476e+00  -1.097 0.272847
## d81          -2.175e+00  8.276e-01  -2.629 0.008686 **
## d82          -6.596e+00  8.534e-01  -7.729 2.33e-14 ***
## d83          -7.397e+00  8.690e-01  -8.512  < 2e-16 ***
## d84          -5.850e+00  8.763e-01  -6.676 3.79e-11 ***
## d85          -6.483e+00  8.948e-01  -7.245 7.82e-13 ***
## d86          -5.853e+00  9.307e-01  -6.289 4.52e-10 ***
## d87          -6.367e+00  9.670e-01  -6.585 6.87e-11 ***
## d88          -6.592e+00  1.014e+00  -6.502 1.17e-10 ***
## d89          -8.071e+00  1.053e+00  -7.667 3.68e-14 ***
## d90          -8.959e+00  1.077e+00  -8.319 2.46e-16 ***
## d91          -1.107e+01  1.101e+00 -10.052  < 2e-16 ***
## d92          -1.288e+01  1.123e+00 -11.473  < 2e-16 ***
## d93          -1.273e+01  1.136e+00 -11.204  < 2e-16 ***
## d94          -1.236e+01  1.157e+00 -10.685  < 2e-16 ***
## d95          -1.195e+01  1.184e+00 -10.098  < 2e-16 ***
## d96          -1.388e+01  1.223e+00 -11.343  < 2e-16 ***
## d97          -1.426e+01  1.250e+00 -11.408  < 2e-16 ***
## d98          -1.504e+01  1.265e+00 -11.886  < 2e-16 ***
## d99          -1.509e+01  1.284e+00 -11.750  < 2e-16 ***
## d00          -1.544e+01  1.305e+00 -11.831  < 2e-16 ***
```

```
## d01          -1.618e+01  1.334e+00 -12.131  < 2e-16 ***
## d02          -1.672e+01  1.348e+00 -12.406  < 2e-16 ***
## d03          -1.702e+01  1.359e+00 -12.521  < 2e-16 ***
## d04          -1.671e+01  1.387e+00 -12.049  < 2e-16 ***
## bac08        -2.498e+00  5.375e-01  -4.648 3.73e-06 ***
## bac10        -1.418e+00  3.963e-01  -3.577 0.000362 ***
## perse        -6.201e-01  2.982e-01  -2.079 0.037791 *
## sbprim       -7.533e-02  4.908e-01  -0.153 0.878032
## sbsecon       6.728e-02  4.293e-01   0.157 0.875492
## sl70plus      3.348e+00  4.452e-01   7.521 1.09e-13 ***
## gdl          -4.269e-01  5.269e-01  -0.810 0.417978
## perc14_24     1.416e-01  1.227e-01   1.154 0.248675
## unem          7.571e-01  7.791e-02   9.718  < 2e-16 ***
## vehicmilespc  2.925e-03  9.497e-05  30.804  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.046 on 1165 degrees of freedom
## Multiple R-squared:  0.6078, Adjusted R-squared:  0.5963
## F-statistic:  53.1 on 34 and 1165 DF,  p-value: < 2.2e-16
```

bac10 is defined as the blood alcohol limit of .10 *bac08* is defined as the blood alcohol limit of .08

Both the variables *bac08* and *bac10* have the negative coefficients of -2.498 and -1.418 respectively. They are statistically significant and it implies that they have a strong negative correlation to the total fatality rate.If we come up with a stricter law and decrease the blood alcohol limit to .10 then the fatalities rate decreases more.

Yes. *perse* variable has a statistically significant negative correlation with the total fatality rate. The coefficient value is -0.6201 which implies a small change in the rate.

TODO write up about primary seatbelt law

4. (15%) Reestimate the model from *Exercise 3* using a fixed effects (at the state level) model. How do the coefficients on *bac08, bac10, perse, and sbprim* compare with the pooled OLS estimates? Which set of estimates do you think is more reliable? What assumptions are needed in each of these models? Are these assumptions reasonable in the current context?

```r
pnldata <- pdata.frame(driving.df, c("state", "year"))

model.fe <- plm(totfatrte ~ d81+d82+d83+d84+d85+d86+d87+d88+d89+
                      d90+d91+d92+d93+d94+d95+d96+d97+d98+d99+
                      d00+d01+d02+d03+d04 + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus
summary(model.fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
```

```
##     perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##     vehicmilespc, data = pnldata, model = "within")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -8.4273592 -1.0258600 -0.0029547  0.9572345 14.8109310
##
## Coefficients:
##                Estimate  Std. Error  t-value  Pr(>|t|)
## d81          -1.51107133  0.41321486  -3.6569 0.0002672 ***
## d82          -3.02549578  0.44243119  -6.8383 1.316e-11 ***
## d83          -3.50360069  0.45657705  -7.6736 3.628e-14 ***
## d84          -4.25936110  0.46494255  -9.1610 < 2.2e-16 ***
## d85          -4.72679311  0.48547032  -9.7365 < 2.2e-16 ***
## d86          -3.66118539  0.51769787  -7.0721 2.686e-12 ***
## d87          -4.30578838  0.55532856  -7.7536 2.001e-14 ***
## d88          -4.76712131  0.60155650  -7.9246 5.501e-15 ***
## d89          -6.12997263  0.64019069  -9.5752 < 2.2e-16 ***
## d90          -6.22973766  0.66485076  -9.3701 < 2.2e-16 ***
## d91          -6.91714040  0.68195432 -10.1431 < 2.2e-16 ***
## d92          -7.77417239  0.70288580 -11.0604 < 2.2e-16 ***
## d93          -8.09410864  0.71594741 -11.3055 < 2.2e-16 ***
## d94          -8.50421668  0.73410866 -11.5844 < 2.2e-16 ***
## d95          -8.25540198  0.75623634 -10.9164 < 2.2e-16 ***
## d96          -8.60661913  0.79594975 -10.8130 < 2.2e-16 ***
## d97          -8.70781739  0.81975686 -10.6224 < 2.2e-16 ***
## d98          -9.34924025  0.83373487 -11.2137 < 2.2e-16 ***
## d99          -9.47489124  0.84399083 -11.2263 < 2.2e-16 ***
## d00          -9.99185979  0.85606370 -11.6719 < 2.2e-16 ***
## d01          -9.63121721  0.87255395 -11.0380 < 2.2e-16 ***
## d02          -8.90673015  0.88205263 -10.0977 < 2.2e-16 ***
## d03          -8.93650263  0.88994687 -10.0416 < 2.2e-16 ***
## d04          -9.33936116  0.91107045 -10.2510 < 2.2e-16 ***
## bac08        -1.43722116  0.39421213  -3.6458 0.0002788 ***
## bac10        -1.06266776  0.26883763  -3.9528 8.208e-05 ***
## perse        -1.15161719  0.23398721  -4.9217 9.867e-07 ***
## sbprim       -1.22739974  0.34271485  -3.5814 0.0003564 ***
## sbsecon      -0.34970784  0.25217091  -1.3868 0.1657826
## sl70plus     -0.06253283  0.26931063  -0.2322 0.8164283
## gdl          -0.41177619  0.29257391  -1.4074 0.1595790
## perc14_24     0.18712169  0.09509969   1.9676 0.0493567 *
## unem         -0.57183997  0.06057851  -9.4397 < 2.2e-16 ***
## vehicmilespc  0.00094005  0.00011104   8.4656 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4535.3
## R-Squared:      0.62624
## Adj. R-Squared: 0.59916
## F-statistic: 55.0943 on 34 and 1118 DF, p-value: < 2.22e-16
```

5. (10%) Would you perfer to use a random effects model instead of the fixed effects model you built in *Exercise 4*? Please explain.

```r
model.re <- plm(totfatrte ~ d81+d82+d83+d84+d85+d86+d87+d88+d89+
                      d90+d91+d92+d93+d94+d95+d96+d97+d98+d99+
                      d00+d01+d02+d03+d04 + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus
summary(model.re)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
##     perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##     vehicmilespc, data = pnldata, model = "random")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Effects:
##                 var std.dev share
## idiosyncratic 4.057   2.014 0.328
## individual    8.294   2.880 0.672
## theta: 0.8615
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -8.25582 -1.15221 -0.15787  0.93086 16.45691
##
## Coefficients:
##                 Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept)   1.7149e+01  2.0964e+00   8.1801 2.835e-16 ***
## d81          -1.5489e+00  4.2830e-01  -3.6164 0.0002988 ***
## d82          -3.2433e+00  4.5772e-01  -7.0858 1.383e-12 ***
## d83          -3.7447e+00  4.7212e-01  -7.9318 2.161e-15 ***
## d84          -4.3729e+00  4.8064e-01  -9.0981 < 2.2e-16 ***
## d85          -4.8609e+00  5.0136e-01  -9.6954 < 2.2e-16 ***
## d86          -3.8295e+00  5.3416e-01  -7.1693 7.539e-13 ***
## d87          -4.5014e+00  5.7213e-01  -7.8678 3.610e-15 ***
## d88          -4.9819e+00  6.1887e-01  -8.0500 8.279e-16 ***
## d89          -6.3713e+00  6.5797e-01  -9.6833 < 2.2e-16 ***
## d90          -6.5357e+00  6.8279e-01  -9.5720 < 2.2e-16 ***
```

```
## d91          -7.3027e+00  7.0030e-01 -10.4279 < 2.2e-16 ***
## d92          -8.2390e+00  7.2126e-01 -11.4230 < 2.2e-16 ***
## d93          -8.5418e+00  7.3449e-01 -11.6296 < 2.2e-16 ***
## d94          -8.9183e+00  7.5297e-01 -11.8442 < 2.2e-16 ***
## d95          -8.6769e+00  7.7541e-01 -11.1902 < 2.2e-16 ***
## d96          -9.0969e+00  8.1573e-01 -11.1518 < 2.2e-16 ***
## d97          -9.2203e+00  8.3984e-01 -10.9786 < 2.2e-16 ***
## d98          -9.8922e+00  8.5380e-01 -11.5860 < 2.2e-16 ***
## d99          -1.0032e+01  8.6426e-01 -11.6071 < 2.2e-16 ***
## d00          -1.0549e+01  8.7667e-01 -12.0330 < 2.2e-16 ***
## d01          -1.0274e+01  8.9336e-01 -11.5000 < 2.2e-16 ***
## d02          -9.6376e+00  9.0278e-01 -10.6755 < 2.2e-16 ***
## d03          -9.6828e+00  9.1090e-01 -10.6300 < 2.2e-16 ***
## d04          -1.0054e+01  9.3254e-01 -10.7816 < 2.2e-16 ***
## bac08        -1.5693e+00  4.0384e-01  -3.8860 0.0001019 ***
## bac10        -1.1380e+00  2.7604e-01  -4.1227 3.744e-05 ***
## perse        -1.0933e+00  2.3885e-01  -4.5772 4.712e-06 ***
## sbprim       -1.1761e+00  3.5144e-01  -3.3465 0.0008184 ***
## sbsecon      -3.4758e-01  2.6024e-01  -1.3356 0.1816862
## sl70plus      2.9969e-02  2.7772e-01   0.1079 0.9140655
## gdl          -3.8524e-01  3.0249e-01  -1.2736 0.2028095
## perc14_24     1.9695e-01  9.7213e-02   2.0259 0.0427722 *
## unem         -4.9238e-01  6.1839e-02  -7.9622 1.690e-15 ***
## vehicmilespc  1.1744e-03  1.0983e-04  10.6933 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12834
## Residual Sum of Squares: 5078.6
## R-Squared:      0.60429
## Adj. R-Squared: 0.59274
## Chisq: 1779.05 on 34 DF, p-value: < 2.22e-16
```

```r
phtest(model.fe, model.re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +  ...
## chisq = 148.69, df = 34, p-value = 2.727e-16
## alternative hypothesis: one model is inconsistent
```

Fixed effect model should be used

6. (10%) Suppose that *vehicmilespc*, the number of miles driven per capita, increases by $1,000$. Using the FE estimates, what is the estimated effect on *totfatrte*? Please interpret the estimate.

The coefficient for the *vehicmilespc* variable is 0.00094005 using the FE estimates and it is highly statistically significant. In other words, There will be an increase of 0.94 fatalities per 100k for an increase of 1000 vehicle miles driven per capita.

7. (5%) If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors?

There is no serial correlation in the idiosyncratic errors of our model as shown in the p-value below. However if there is Serial correlation then it will not affect the unbiasedness or consistency of OLS estimators, but it does affect their efficiency. With positive serial correlation, the OLS estimates of the standard errors will be smaller than the true standard errors. This will lead to the conclusion that the parameter estimates are more precise than they really are. There will be a tendency to reject the null hypothesis when it should not be rejected.

```r
pbgtest(model.fe)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +     d90 + d91 + d9
## chisq = 340.4, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```