Lab 5: Panel Data

Eric Yang, Samir Datta, Carlos Castro

December 17, 2017

```
library(ggplot2)
library(reshape2)
library(plm)
```

```
## Warning: package 'plm' was built under R version 3.4.3
```

```
## Loading required package: Formula
```

```
load("driving.Rdata")
drivedata <- data
```

Structure of the data

```
head(drivedata)
```

```
##   year state sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl bac10
## 1 1980     1    1    0    0    0      0        0     18       0   0     1
## 2 1981     1    1    0    0    0      0        0     18       0   0     1
## 3 1982     1    1    0    0    0      0        0     18       0   0     1
## 4 1983     1    1    0    0    0      0        0     18       0   0     1
## 5 1984     1    1    0    0    0      0        0     18       0   0     1
## 6 1985     1    1    0    0    0      0        0     20       0   0     1
##   bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm
## 1     0     0    940     422     236      3.20      1.437      0.803
## 2     0     0    933     434     248      3.35      1.558      0.890
## 3     0     0    839     376     224      2.81      1.259      0.750
## 4     0     0    930     397     223      3.00      1.281      0.719
## 5     0     0    932     421     237      2.83      1.278      0.720
## 6     0     0    882     358     224      2.51      1.019      0.637
##   statepop totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24
## 1  3893888     24.14      10.84       6.06   29.37500  8.8      18.9
## 2  3918520     24.07      11.08       6.33   27.85200 10.7      18.7
## 3  3925218     21.37       9.58       5.71   29.85765 14.4      18.4
## 4  3934109     23.64      10.09       5.67   31.00000 13.7      18.0
## 5  3951834     23.58      10.65       6.00   32.93286 11.1      17.6
## 6  3972527     22.20       9.01       5.64   35.13944  8.9      17.3
##   sl70plus sbprim sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91
## 1        0      0       0   1   0   0   0   0   0   0   0   0   0   0   0
## 2        0      0       0   0   1   0   0   0   0   0   0   0   0   0   0
## 3        0      0       0   0   0   1   0   0   0   0   0   0   0   0   0
## 4        0      0       0   0   0   0   1   0   0   0   0   0   0   0   0
## 5        0      0       0   0   0   0   0   1   0   0   0   0   0   0   0
## 6        0      0       0   0   0   0   0   0   1   0   0   0   0   0   0
##   d92 d93 d94 d95 d96 d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc
## 1   0   0   0   0   0   0   0   0   0   0   0   0   0     7543.874
## 2   0   0   0   0   0   0   0   0   0   0   0   0   0     7107.785
```

```
## 3   0   0   0   0   0   0   0   0   0   0   0   0   0     7606.622
## 4   0   0   0   0   0   0   0   0   0   0   0   0   0     7879.802
## 5   0   0   0   0   0   0   0   0   0   0   0   0   0     8333.562
## 6   0   0   0   0   0   0   0   0   0   0   0   0   0     8845.614
```

```
unique(drivedata$year)
```

```
##  [1] 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993
## [15] 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004
```

```
unique(drivedata$state)
```

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

The data is in long form. There are 25 rows per state, each representing a different year from 1980 to 2005. The data description data frame tells us that the "state" variable is a number that simply represents the 50 states in alphabetical order.

Some variables like the BAC and seatbelt law variable have more than two levels represented by multiple dummy coded variables. For example, BAC has three levels - no BAC law, BAC limit = .08, and BAC limit = .10. These three levels are coded in the two variables bac08 and bac10. Each of the years is also dummy coded in its own column.

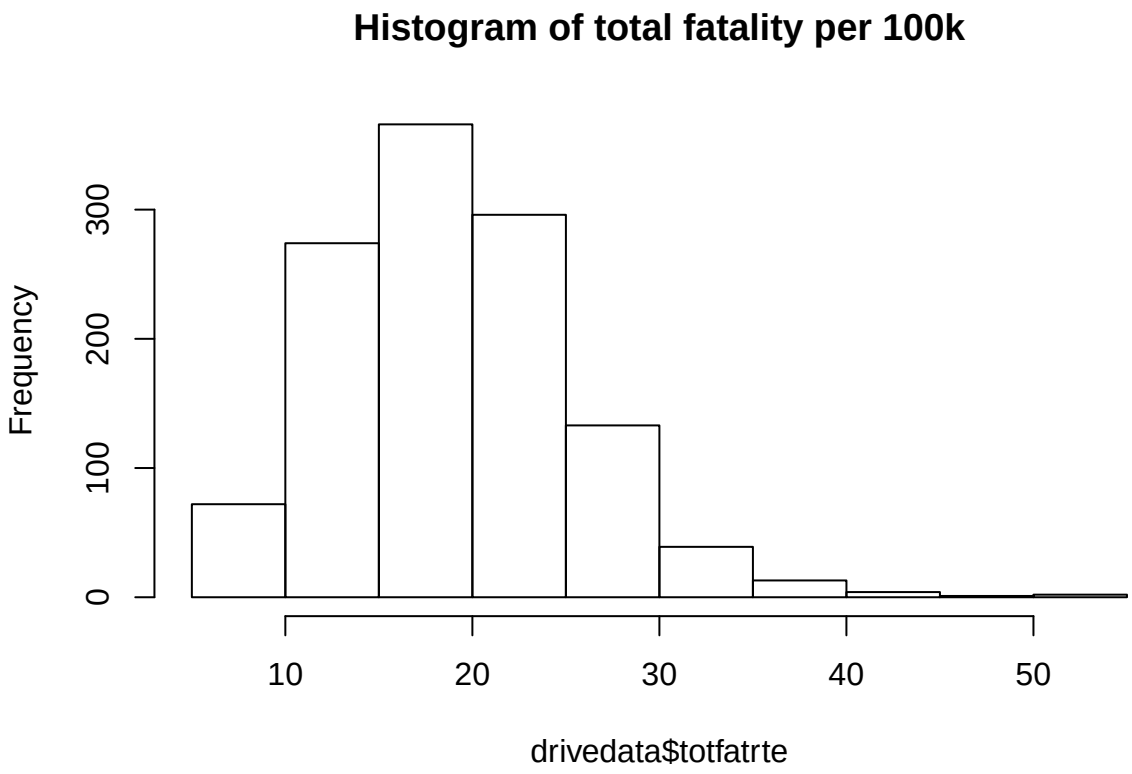
An odd fluke in the data noticed above that there is no state number 2. The number is suppose to represent the states in alphabetical order, but 2 is skipped and the maximum number is 51 instead of 50. This will not affect our modeling in any way since the number is effectively a categorical variable and the values don't matter.

EDA

Total fatality rate

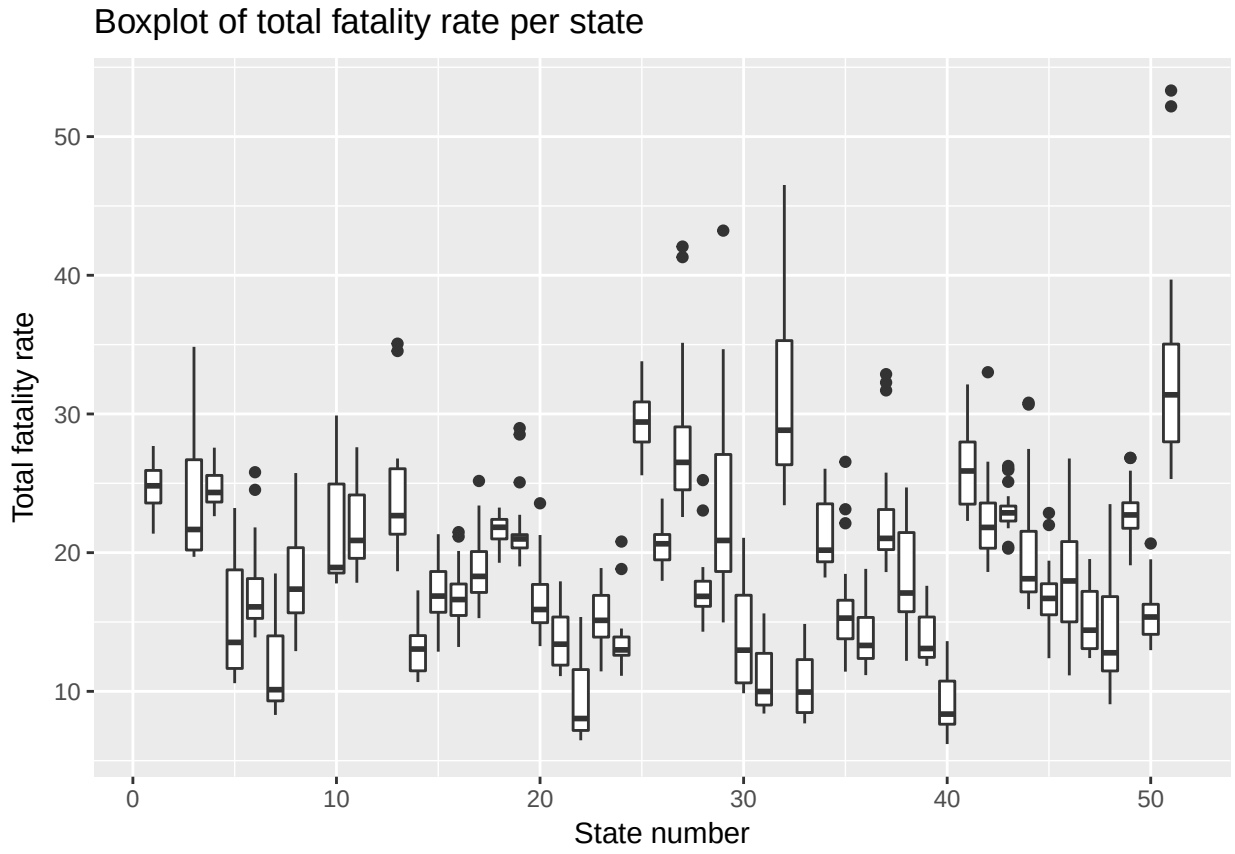
The primary variable of interest is "totfatrte", the total fatalities per 100,000 population.

```
hist(drivedata$totfatrte, main="Histogram of total fatality per 100k")
```



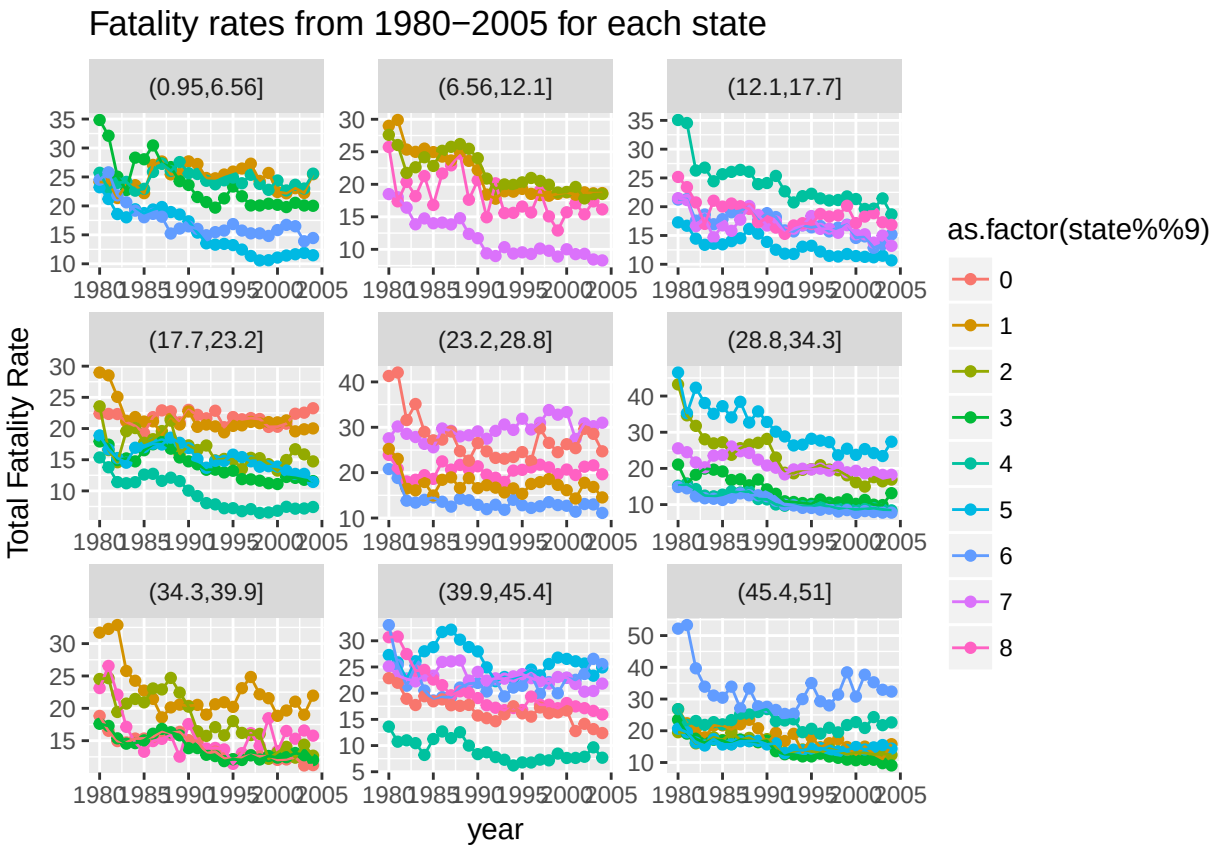
The data has a slight positive skew. Note that this histogram is of all values of total fatality rate, across all states and years.

```
ggplot(drivedata, aes(x=state, y=totfatrte, group=state)) + geom_boxplot()+
ylab("Total fatality rate")+xlab("State number")+ggtitle("Boxplot of total fatality rate per state")
```



Clearly, each state has different fatality rates, and some states simply have higher or lower rates regardless of the year or other variables.

```
drivedata$statebucket <- cut(drivedata$state, 9)

ggplot(drivedata, aes(x=year, y=totfatrte, group=as.factor(state), color=as.factor(state%%9)))+geom_poi
  geom_line()+facet_wrap(~as.factor(statebucket), scales="free")+
  ylab("Total Fatality Rate")+guides(fill=F)+
  ggtitle("Fatality rates from 1980-2005 for each state")
```

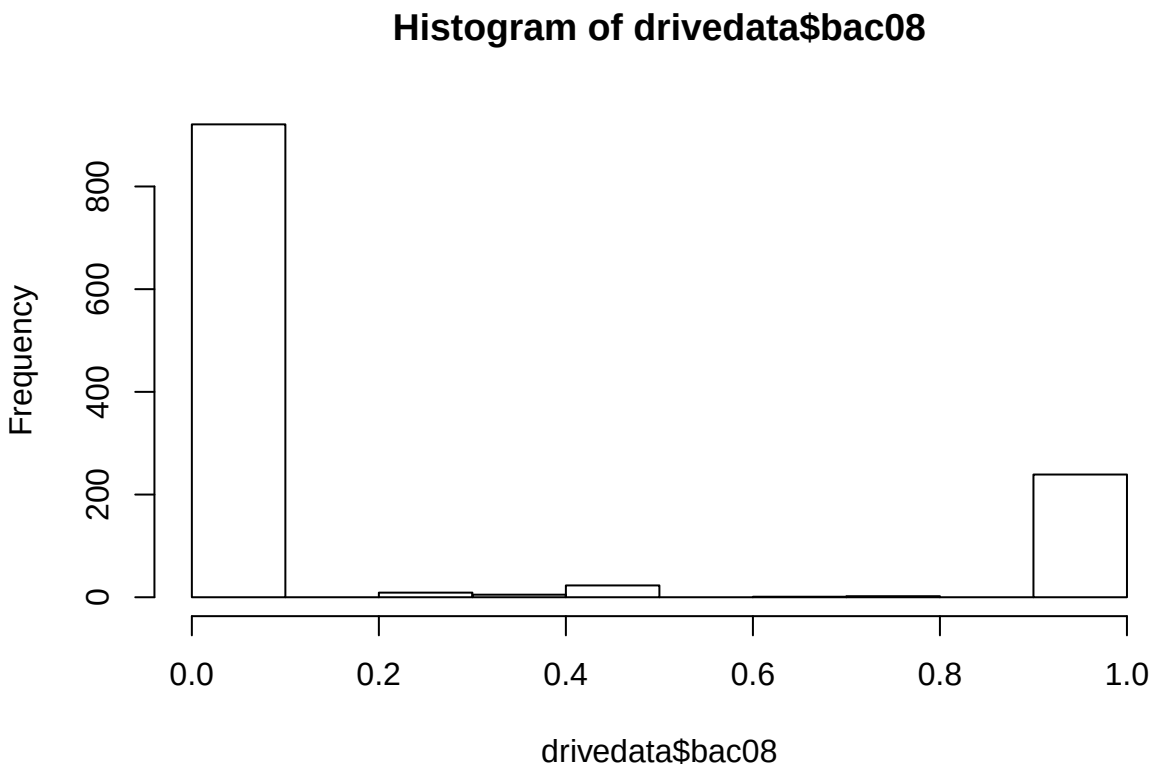


The above plot shows the fatality rates from 1980-205. Each set of dots connected with lines represents the data for a different state. 9 separate graphs are shown simply for the sake of avoiding clutter - the states have been binned by the number in the dataset (their alphabetical order).

This graph is simply to get an idea of the nature of year-to-year changes in fatlity rate. It seems for most states the rates steadily go down, but there are some interesting exceptions. It also seems that for a lot of states, the rates for the first one or two years is much higher, after which the rates drastically drop. The overall decrease in fatality rates over time is likely due to factors such as laws enacted.

Laws

```
hist(drivedata$bac08)
```



Many of the explanatory variables of interest (bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl) appear to be binary - 0 if the law was not in place, 1 if it was. As shown in the example histogram above, there are a very small number of fractional values.

```
head(drivedata[drivedata$year>1990,c("year", "bac08")], 10)
```

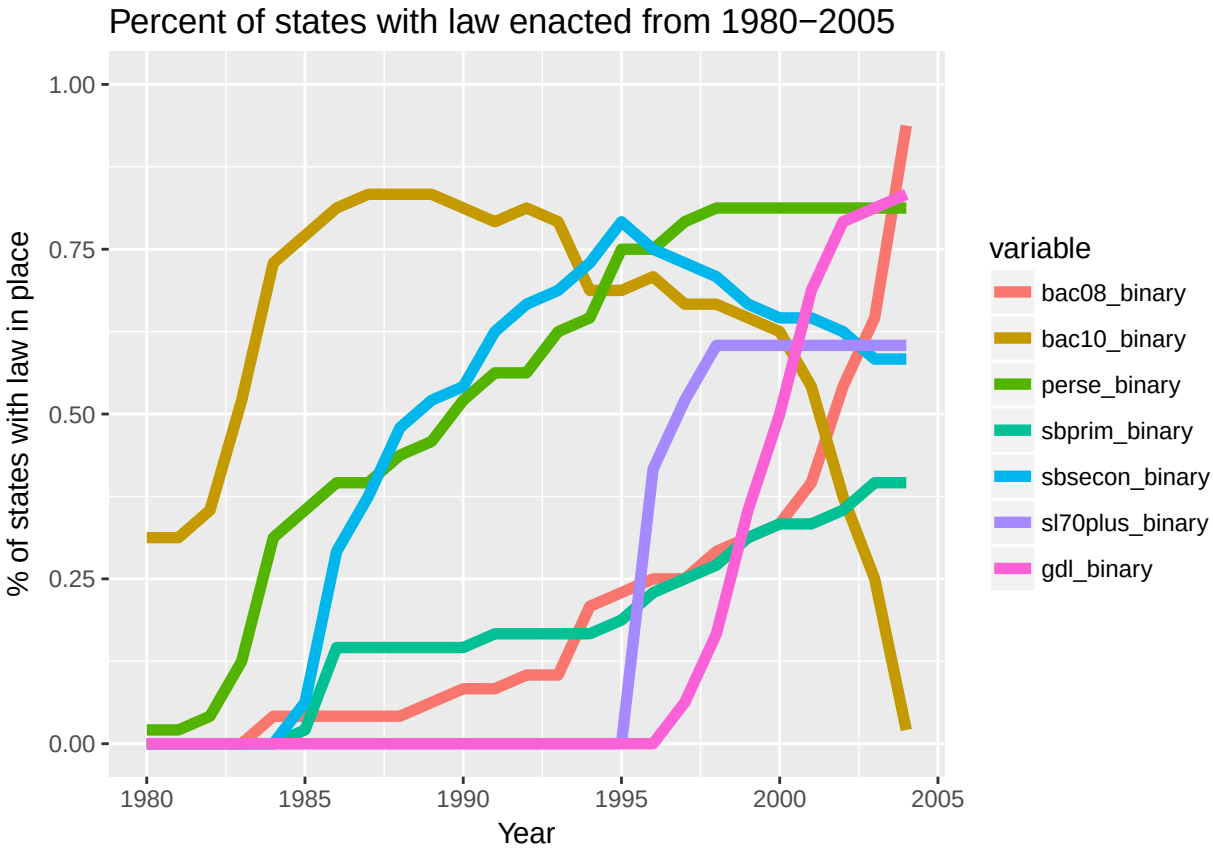
```
##    year bac08
## 12 1991 0.000
## 13 1992 0.000
## 14 1993 0.000
## 15 1994 0.000
## 16 1995 0.417
## 17 1996 1.000
## 18 1997 1.000
## 19 1998 1.000
## 20 1999 1.000
## 21 2000 1.000
```

As shown in the example section of the data above, these fractional values occur between stretches of 0 and 1. There is no additional clarifying information in the dataset or data descriptions. As such, we are moving forward with the assumption that fractional values occur when the law was enacted part way through the year, and the value represents how much of the year the law was in place for.

Keeping these fractional values in our data will force the model to treat them as continuous variables, which is very problematic. We are not interested in the effect of a law if it is present for part of the year; rather, we want to know the binary effect of a law being in place. It is also very problematic to use a continuous variable in a linear model with such an odd and non-normal distribution, given the sparsity of the fractional values. As such, we are moving forward using the rounded values which will turn the fractional values into 0 or 1.

```r
#code to binarize all law variables
drivedata$bac08_binary <- ifelse(round(drivedata$bac08)==1, 1, 0)
drivedata$bac10_binary <- ifelse(round(drivedata$bac10)==1, 1, 0)
drivedata$perse_binary <- ifelse(round(drivedata$perse)==1, 1, 0)
drivedata$sbprim_binary <- ifelse(round(drivedata$sbprim)==1, 1, 0)
drivedata$sbsecon_binary <- ifelse(round(drivedata$sbsecon)==1, 1, 0)
drivedata$sl70plus_binary <- ifelse(round(drivedata$sl70plus)==1, 1, 0)
drivedata$gdl_binary <- ifelse(round(drivedata$gdl)==1, 1, 0)
drivedata$sl55_binary <- ifelse(round(drivedata$sl55)==1, 1, 0)
drivedata$sl65_binary <- ifelse(round(drivedata$sl65)==1, 1, 0)
drivedata$sl70_binary <- ifelse(round(drivedata$sl70)==1, 1, 0)
drivedata$sl75_binary <- ifelse(round(drivedata$sl75)==1, 1, 0)
drivedata$slnone_binary <- ifelse(round(drivedata$slnone)==1, 1, 0)
```

```r
year_law_agg <- with(drivedata, aggregate(cbind(bac08_binary, bac10_binary, perse_binary, sbprim_binary

year_law_melt <- melt(year_law_agg, id.vars="year")

ggplot(year_law_melt, aes(x=year, y=value, color=variable, group=variable))+
  geom_line(size=2)+ylab("% of states with law in place")+xlab("Year")+
  ggtitle("Percent of states with law enacted from 1980-2005")+
  ylim(c(0,1))
```



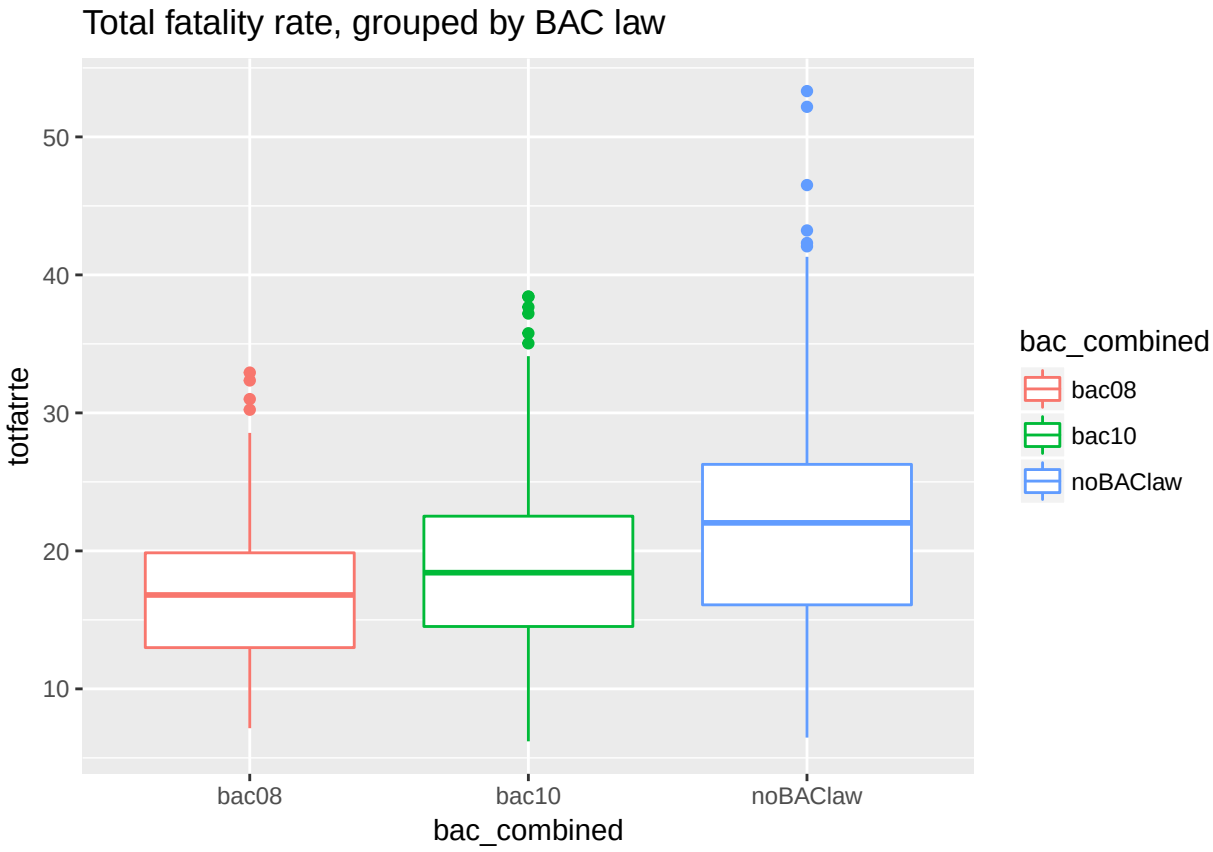
This graph shows the % of states that had a certain law in place in every year.

The blood alcohol content laws are interesting. Clearly in the early 2000s there was a big movement to decrease the legal BAC from .10 to .08 given the opposite directions of those lines. A similar, but weaker,

pattern is shown with the seatbelt law being primary or secondary around 1995.

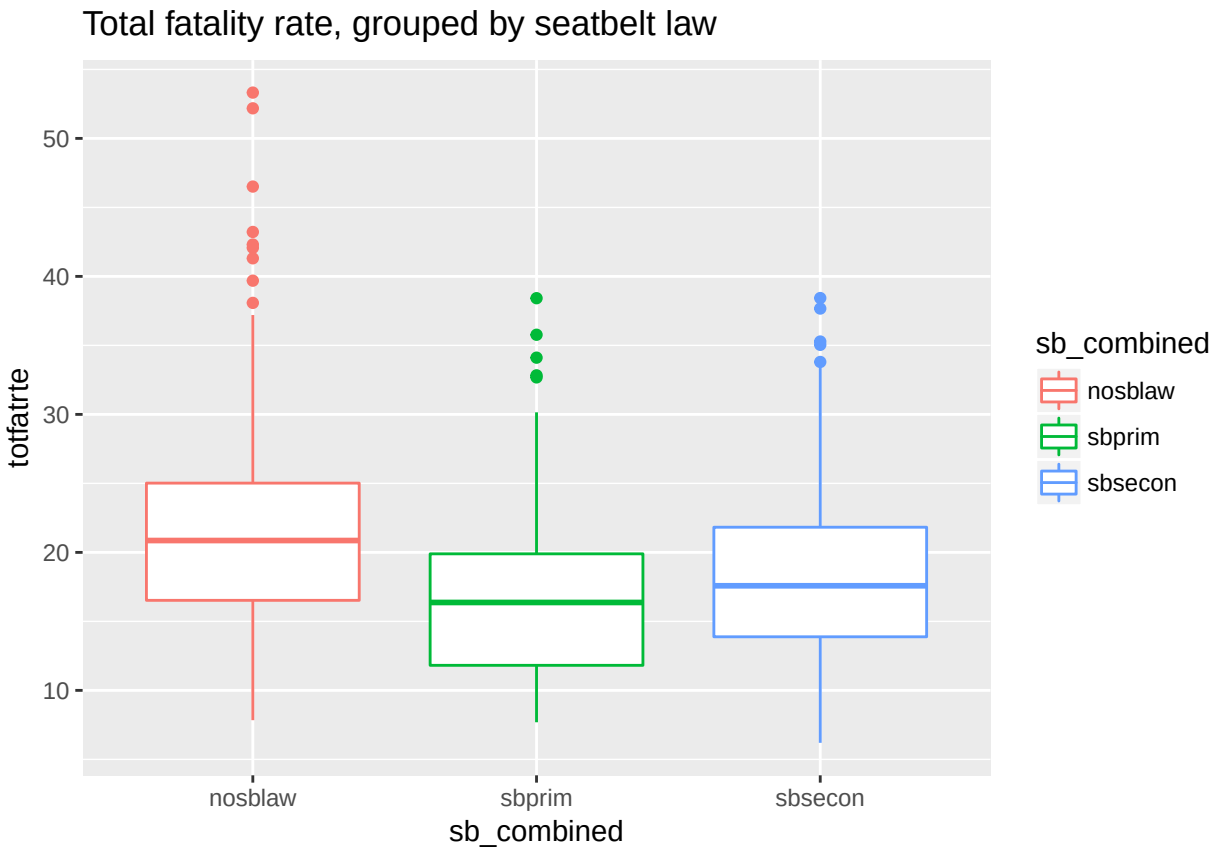
It is also interesting to note that both the graduated drivers license law and the speed limit being 70 or more changes did not happen at all until around 1995, and then fairly rapidly became more of a standard. Still, the speed limit change plateaus around 60%. It seems that about 60% of states adopted the speed limit change very quickly and the remaining states did not.

```
drivedata$bac_combined <- ifelse(drivedata$bac08_binary==1, "bac08",
                                 ifelse(drivedata$bac10_binary==1, "bac10", "noBAClaw"))
ggplot(drivedata, aes(x=bac_combined, y=totfatrte, color=bac_combined, group=bac_combined))+geom_boxplo
  ggtitle("Total fatality rate, grouped by BAC law")
```



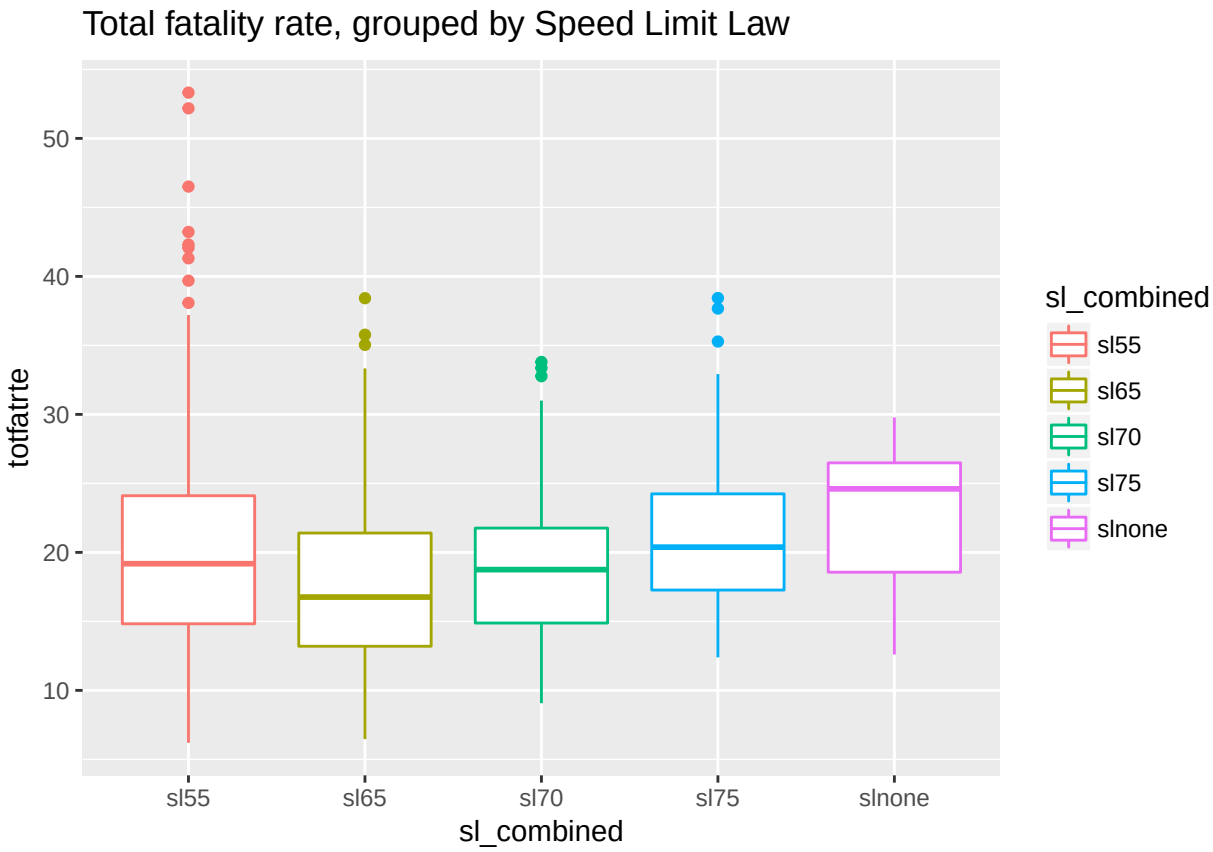
Predictably, in states/years where the BAC limit was .08, there were lower fatality rates than when the limit was .10. States/years where there was no BAC law in place had the highest rates on average.

```
drivedata$sb_combined <- ifelse(drivedata$sbprim_binary==1, "sbprim",
                                ifelse(drivedata$sbsecon_binary==1, "sbsecon", "nosblaw"))
ggplot(drivedata, aes(x=sb_combined, y=totfatrte, color=sb_combined, group=sb_combined))+geom_boxplot()+
  ggtitle("Total fatality rate, grouped by seatbelt law")
```



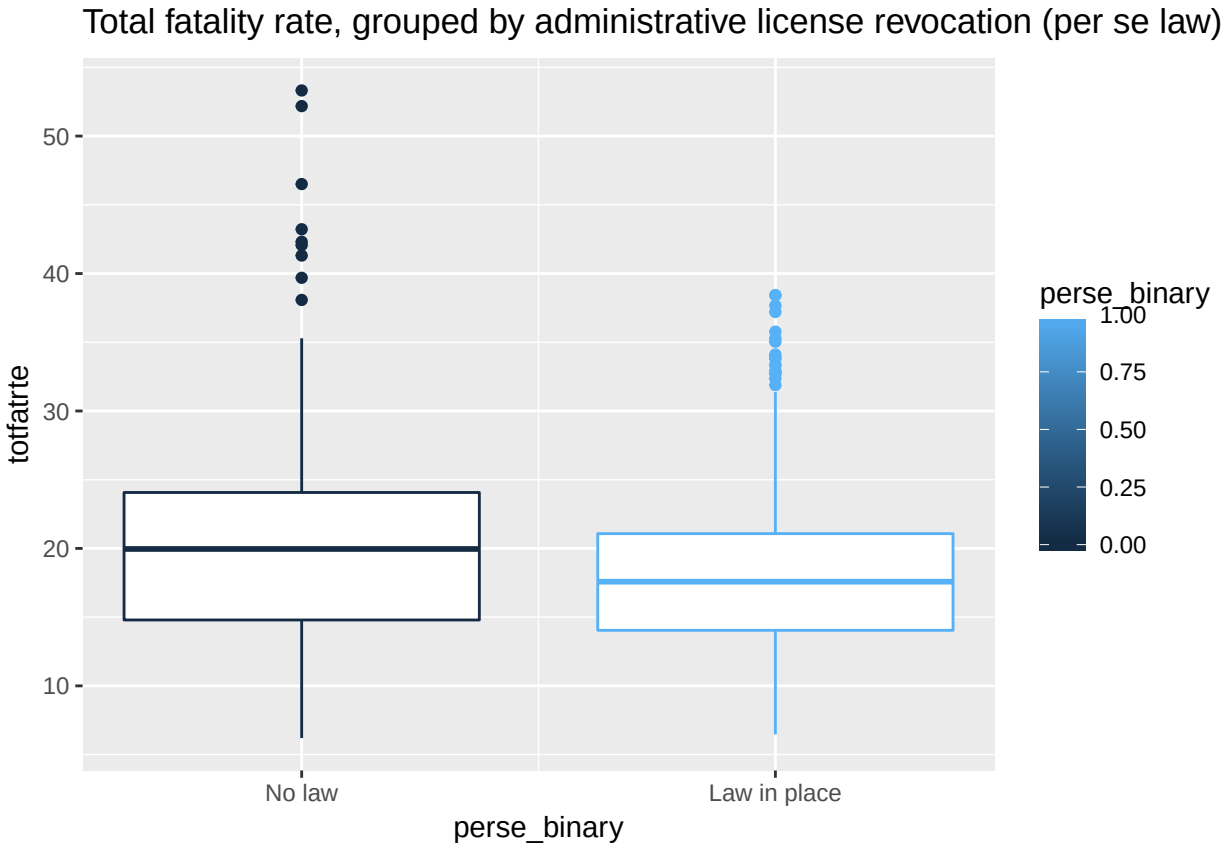
States/years without a seatbelt law have the highest fatality rate. States/years with the primary seatbelt law have lower rates than those with the secondary seatbelt law.

```
drivedata$sl_combined <- ifelse(drivedata$sl55_binary, "sl55",
                                ifelse(drivedata$sl65_binary==1, "sl65",
                                       ifelse(drivedata$sl70_binary==1, "sl70",
                                              ifelse(drivedata$sl75_binary==1, "sl75","slnone"))))
ggplot(drivedata, aes(x=sl_combined, y=totfatrte, color=sl_combined, group=sl_combined))+geom_boxplot()+
  ggtitle("Total fatality rate, grouped by Speed Limit Law")
```

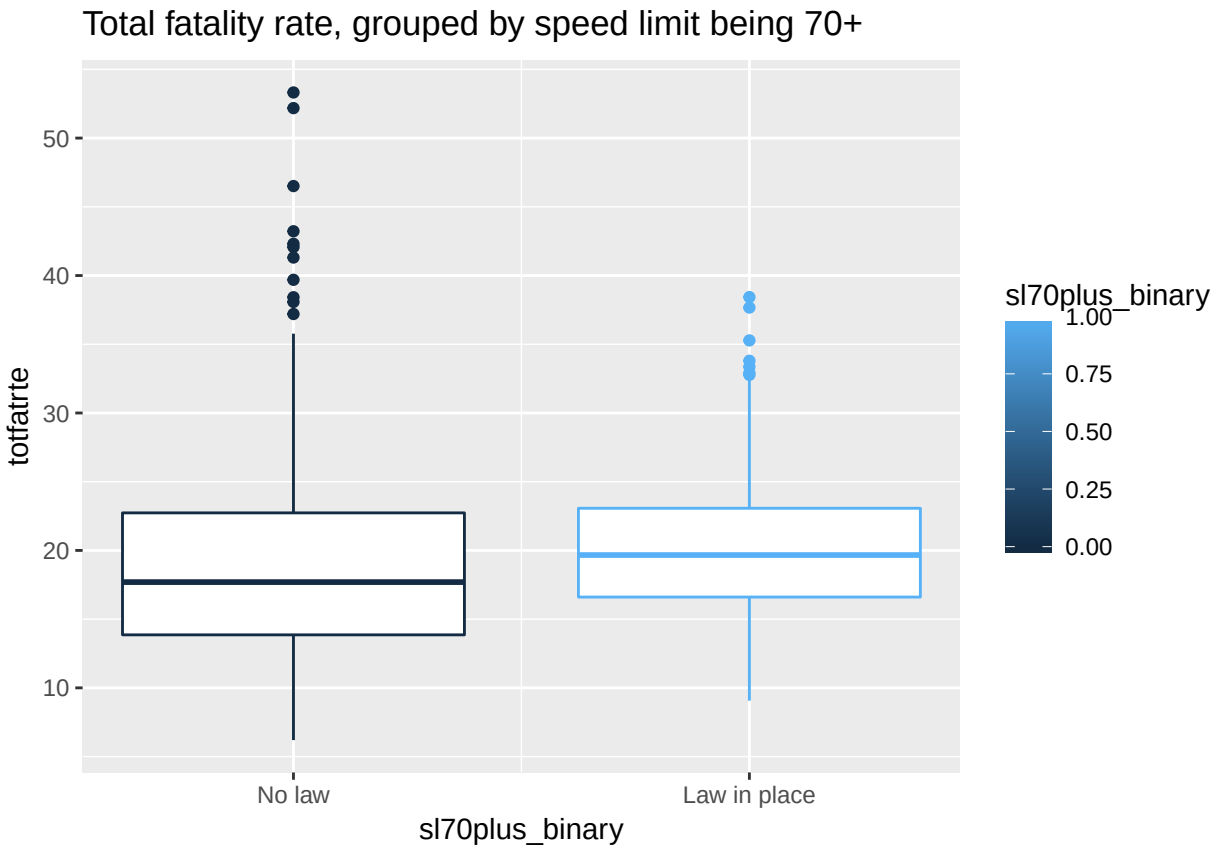


States/years with lower speed limits have lower fatalaty rates with the exception of speed limit of 55.

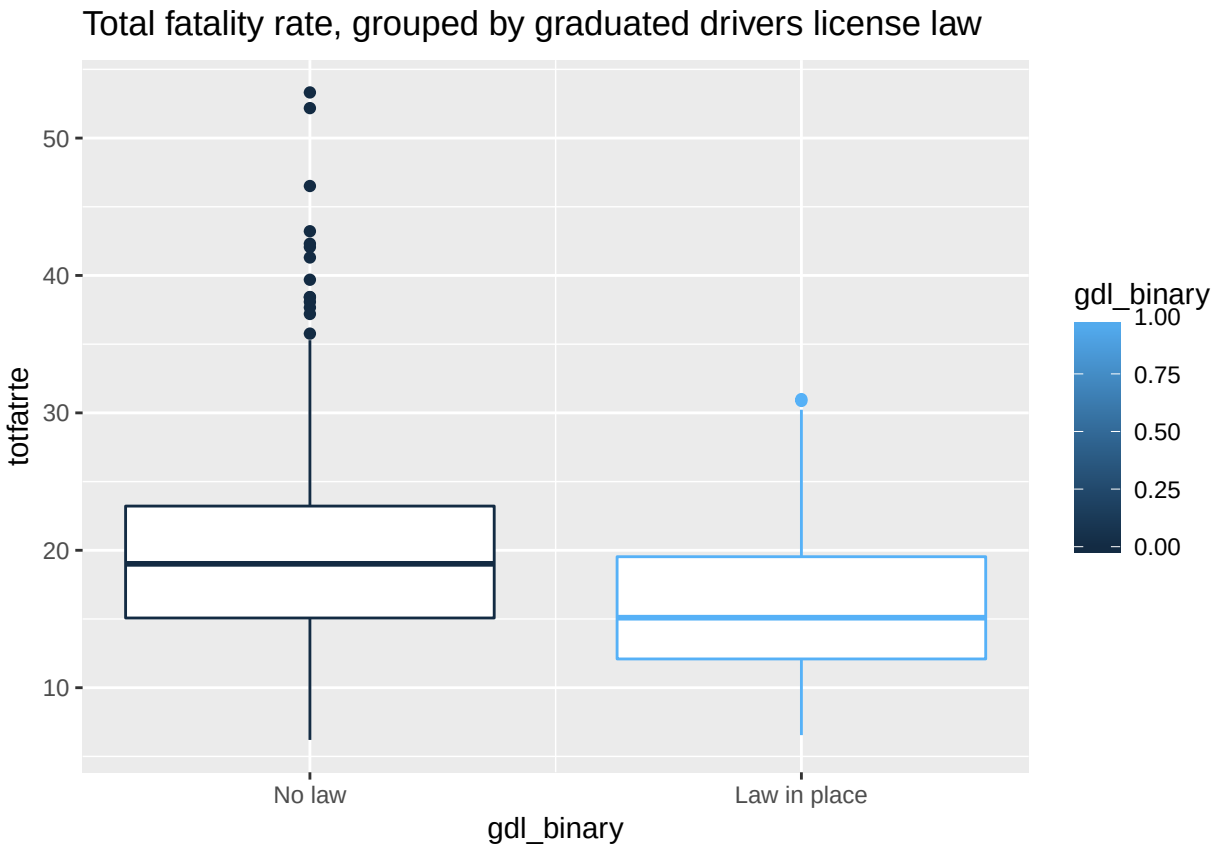
```
#bac, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc

ggplot(drivedata, aes(x=perse_binary, y=totfatrte, color=perse_binary, group=perse_binary))+geom_boxplo
  scale_x_continuous(breaks=c(0,1), labels=c("No law", "Law in place"))+
  ggtitle("Total fatality rate, grouped by administrative license revocation (per se law)")
```




```
ggplot(drivedata, aes(x=sl70plus_binary, y=totfatrte, color=sl70plus_binary, group=sl70plus_binary))+ge
  scale_x_continuous(breaks=c(0,1), labels=c("No law", "Law in place"))+
  ggtitle("Total fatality rate, grouped by speed limit being 70+")
```



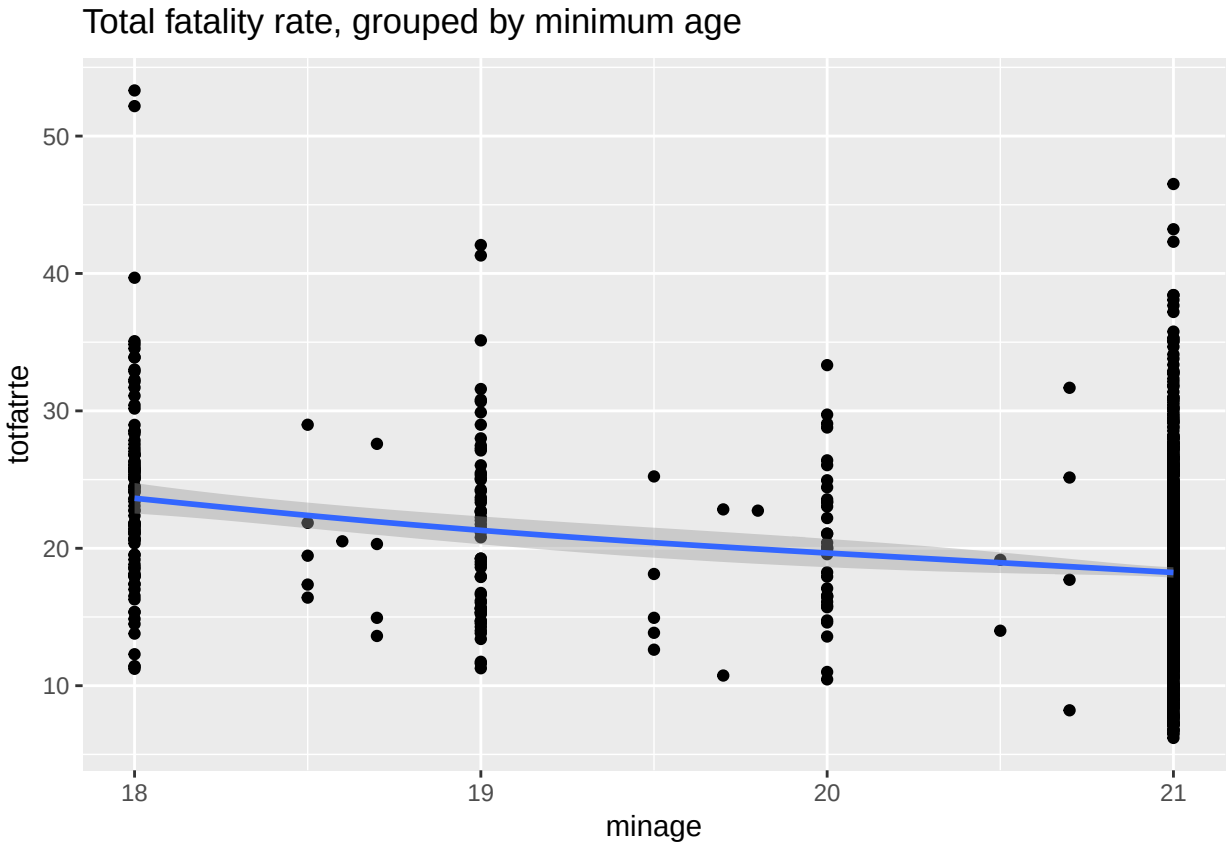

```
ggplot(drivedata, aes(x=gdl_binary, y=totfatrte, color=gdl_binary, group=gdl_binary))+geom_boxplot()+
  scale_x_continuous(breaks=c(0,1), labels=c("No law", "Law in place"))+
  ggtitle("Total fatality rate, grouped by graduated drivers license law")
```



The above boxplots show that fatality rates are lower when the "per se" law and the graduated drivers license law are in place, and higher when the speed limit is above 70. At least based on this plot, it seems the speed limit has the weakest effect. Raising the speed limit would intuitively increase fatality rates due to faster and more reckless driving. However, based on the previous time plot, the speed limit change appeared to happen around the same time as other safe laws were being enacted more often. Including all of these variables into a model should clarify their individual, ceteris paribus effects.

```
ggplot(data, aes(x=minage, y=totfatrte))+
  geom_point()+geom_smooth(method="auto")+ggtitle("Total fatality rate, grouped by minimum age")
```

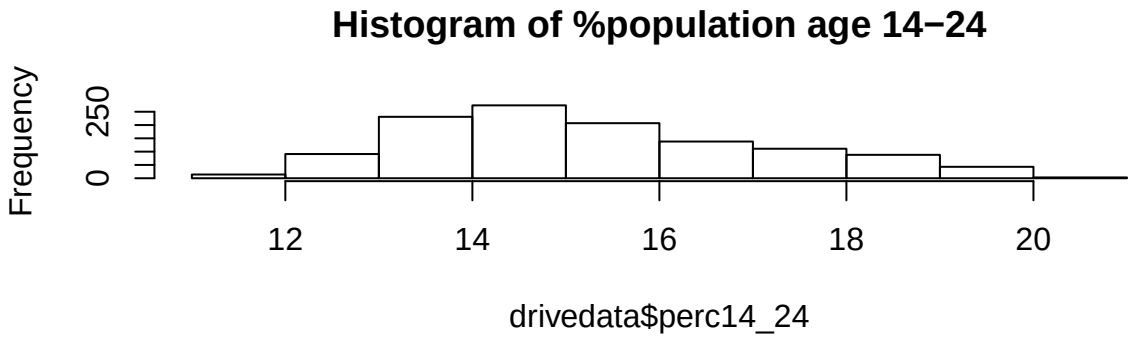
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

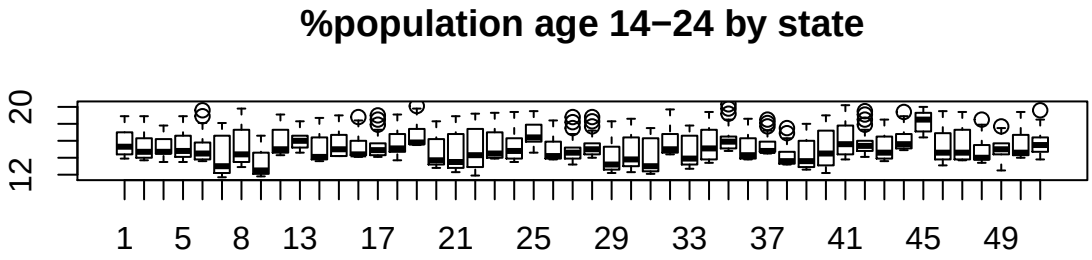


States/years with higher minimum driving age show lower fatality rates. The data seems to be either 18, 19, 20, or 21. A very small number of observations have a fractional value. Similar to the fractional values for the binary law variables, this may represent when the minimum driving age was changed part of the way through the year.

Other explanatory variables

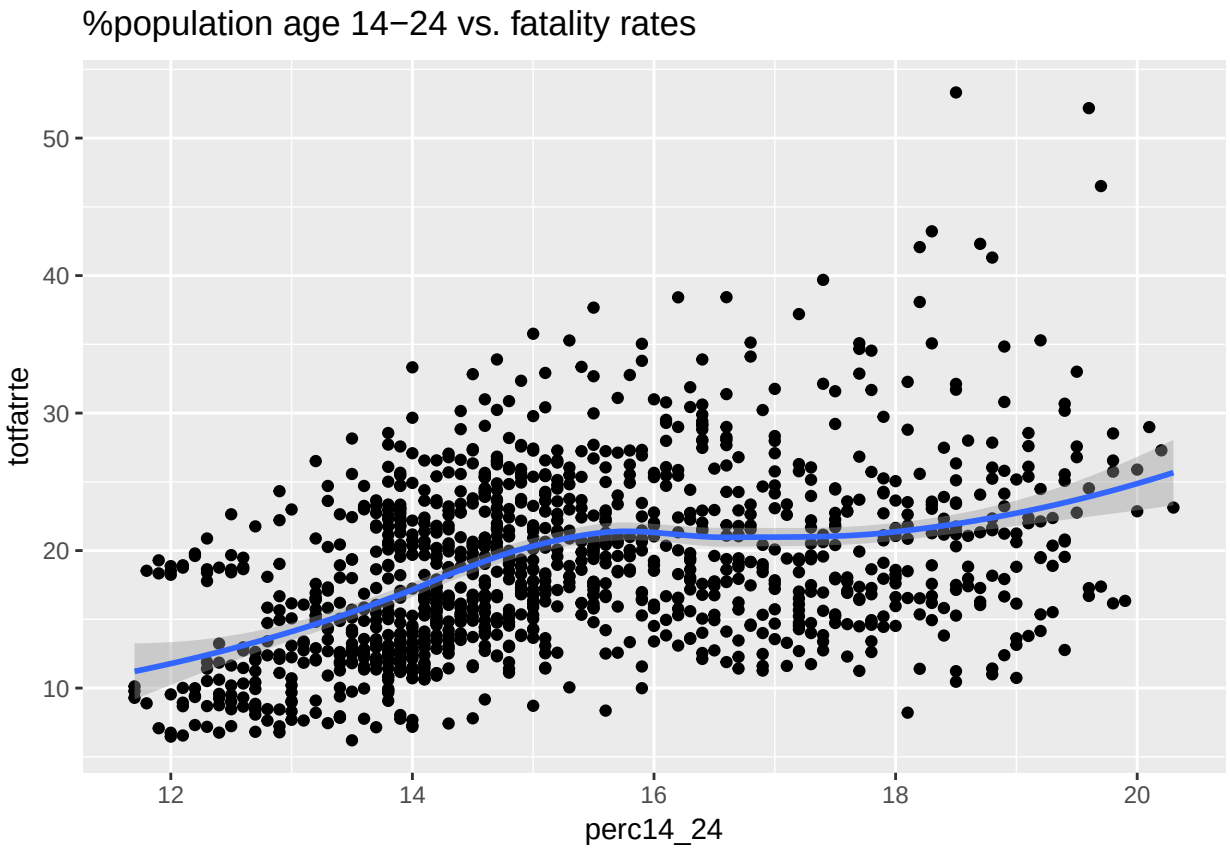
```
par(mfrow=c(2,1))
hist(drivedata$perc14_24, main="Histogram of %population age 14-24")
boxplot(perc14_24~state, data=drivedata, main="%population age 14-24 by state")
```





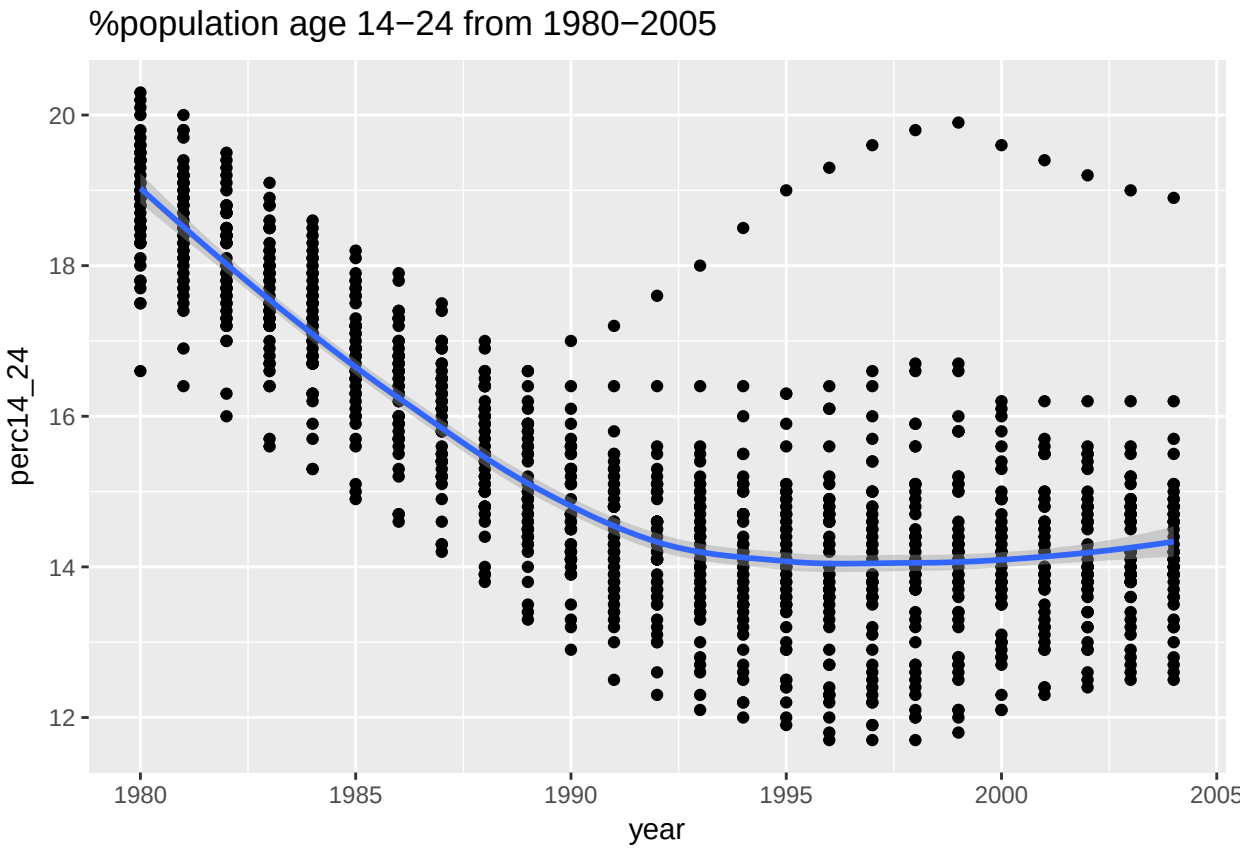
The percent of population across states/years appears to be a normally distributed variable. This variable doesn't appear to be especially different from state to state - there are no states that tend to have an especially high or low % of ages 14-24.

```
ggplot(drivedata, aes(x=perc14_24, y=totfatrte))+
  geom_point()+geom_smooth(method="loess")+ggtitle("%population age 14-24 vs. fatality rates")
```



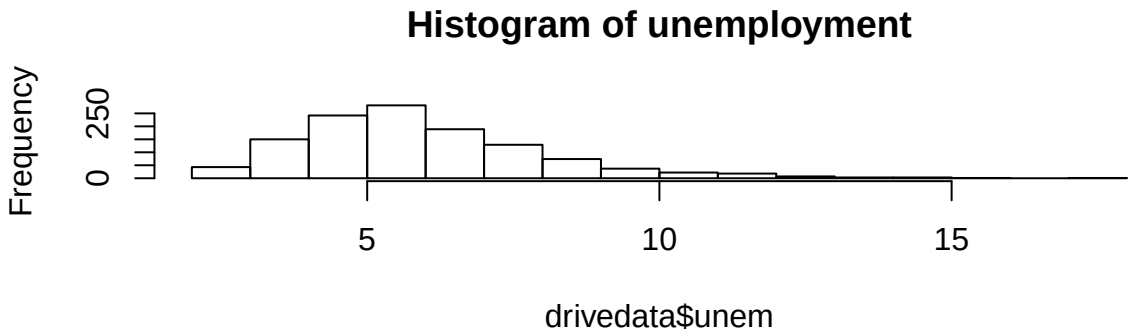
There appears to be a linear relationship where states/years with a higher % of the population between ages 14-24 have higher fatality rates. Perhaps younger drivers are less safe.

```
ggplot(drivedata, aes(x=year, y=perc14_24))+
  geom_point()+geom_smooth(method="loess")+ggtitle("%population age 14-24 from 1980-2005")
```



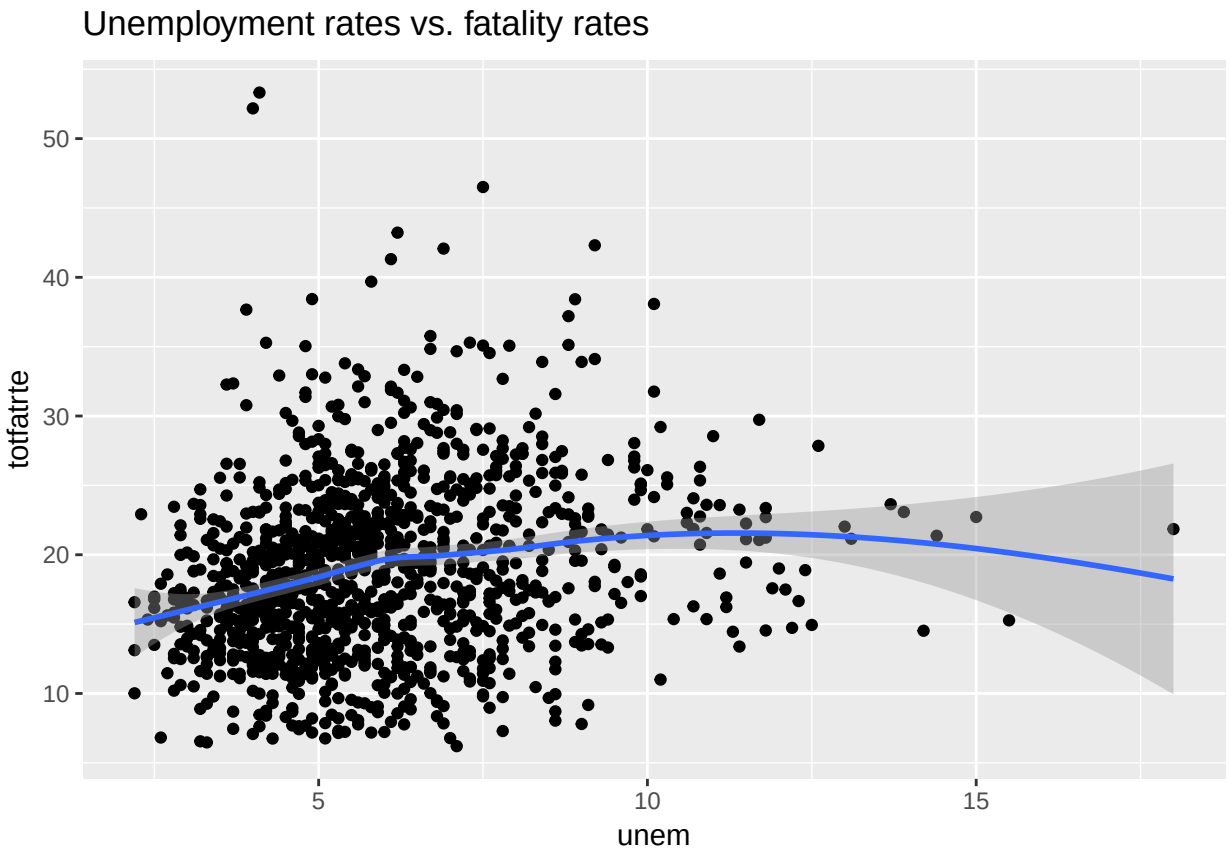
The % of the population between 14 and 24 very distinctly goes down from 1980 to about 1990, after which it flattens out to around 14%.

```
par(mfrow=c(2,1))
hist(drivedata$unem, main="Histogram of unemployment")
boxplot(unem~state, data=drivedata, main="unemployment rate by state")
```



The unemployment rate, which is a percent, has a positive skew when looking at all the values in the dataset. We did not believe the skew to be problematic enough to warrant any transformation. Some states appear to have distinctly higher or lower unemployment rates.

```
ggplot(drivedata, aes(x=unem, y=totfatrte))+
  geom_point()+geom_smooth(method="loess")+ggtitle("Unemployment rates vs. fatality rates")
```

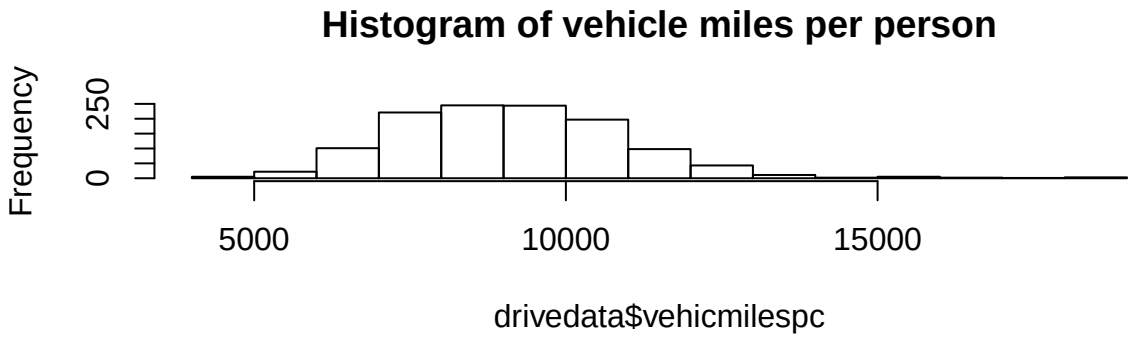


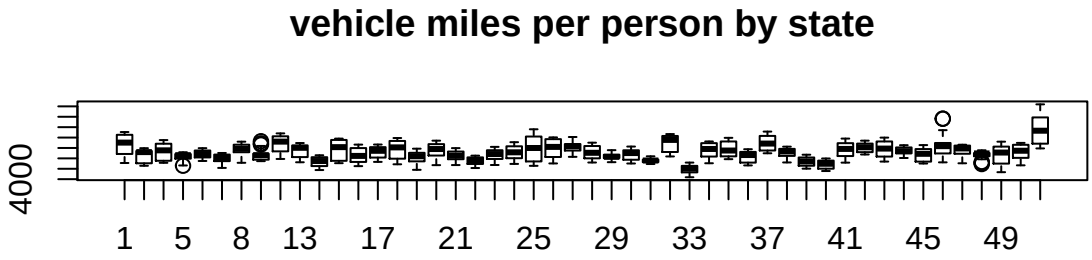
There does not appear to be any discernible relationship between unemployment rates and driving fatality rates.

```
ggplot(drivedata, aes(x=year, y=log(unem)))+
  geom_point()+geom_smooth(method="loess")+ggtitle("Unemployment rates per person vs. fatality rates")
```

Unemployment apears to have gone down over time, with a slight uptick in 1990.

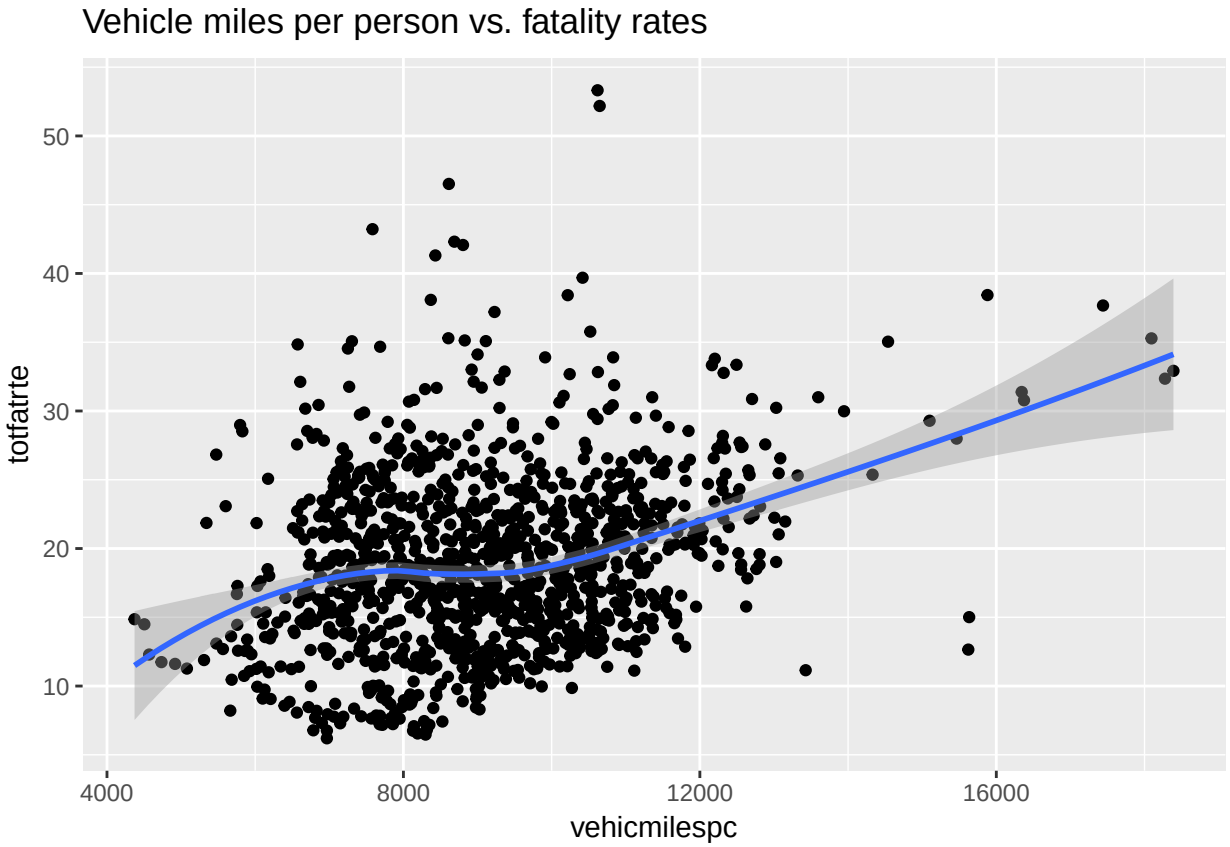
```
par(mfrow=c(2,1))
hist(drivedata$vehicmilespc, main="Histogram of vehicle miles per person")
boxplot(vehicmilespc~state, data=drivedata, main="vehicle miles per person by state")
```





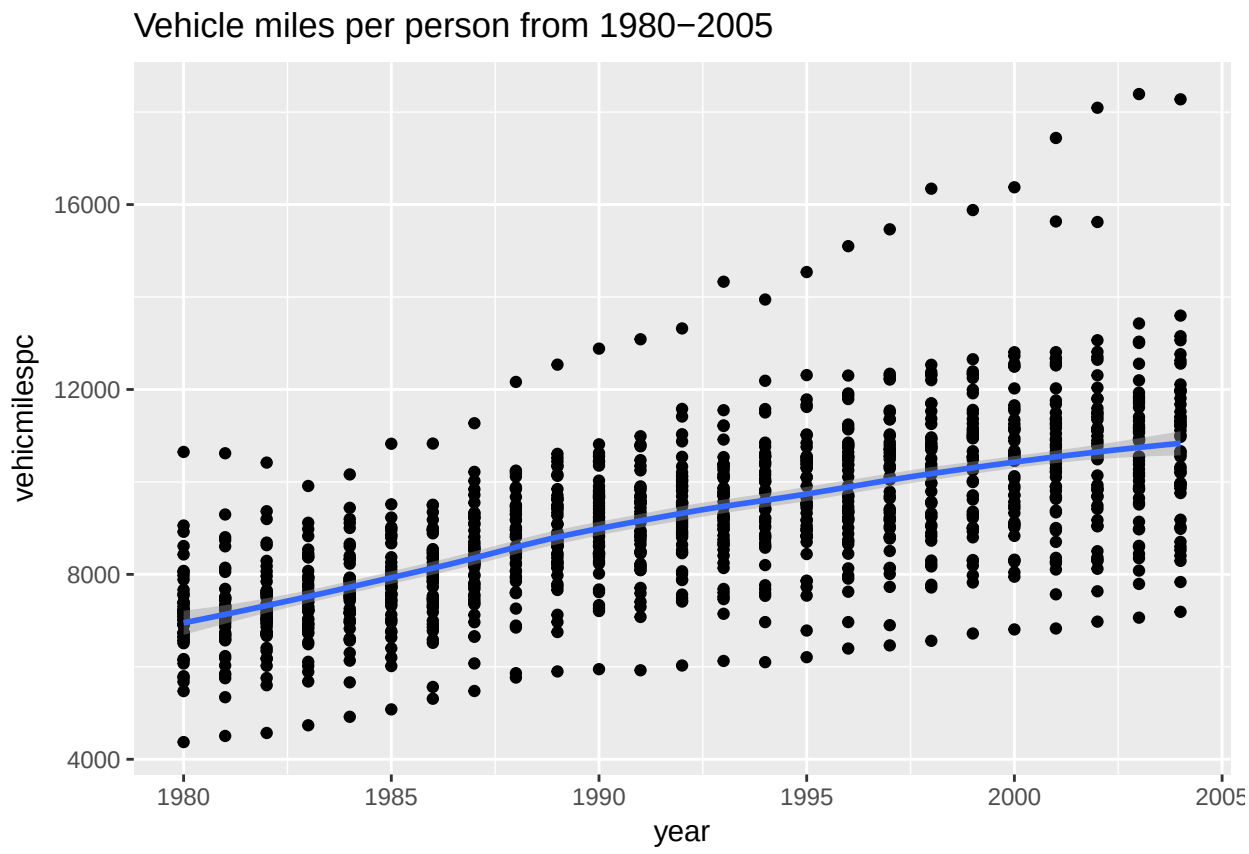
"vehicmilespc" is the vehicle miles per capita, which was calculated as the the number of vehicle miles travelled (in billions) divided by the state population. Despite this attempt at correction, it is clear from the boxplots that some states tend to have a higher or lower value for this variable.

```
ggplot(drivedata, aes(x=vehicmilespc, y=totfatrte))+
  geom_point()+geom_smooth(method="loess")+ggtitle("Vehicle miles per person vs. fatality rates")
```



Fatality rates appears to be positively correlated with vehicle miles, although these relationship may be driven by a relatively small number of data points with large influence.

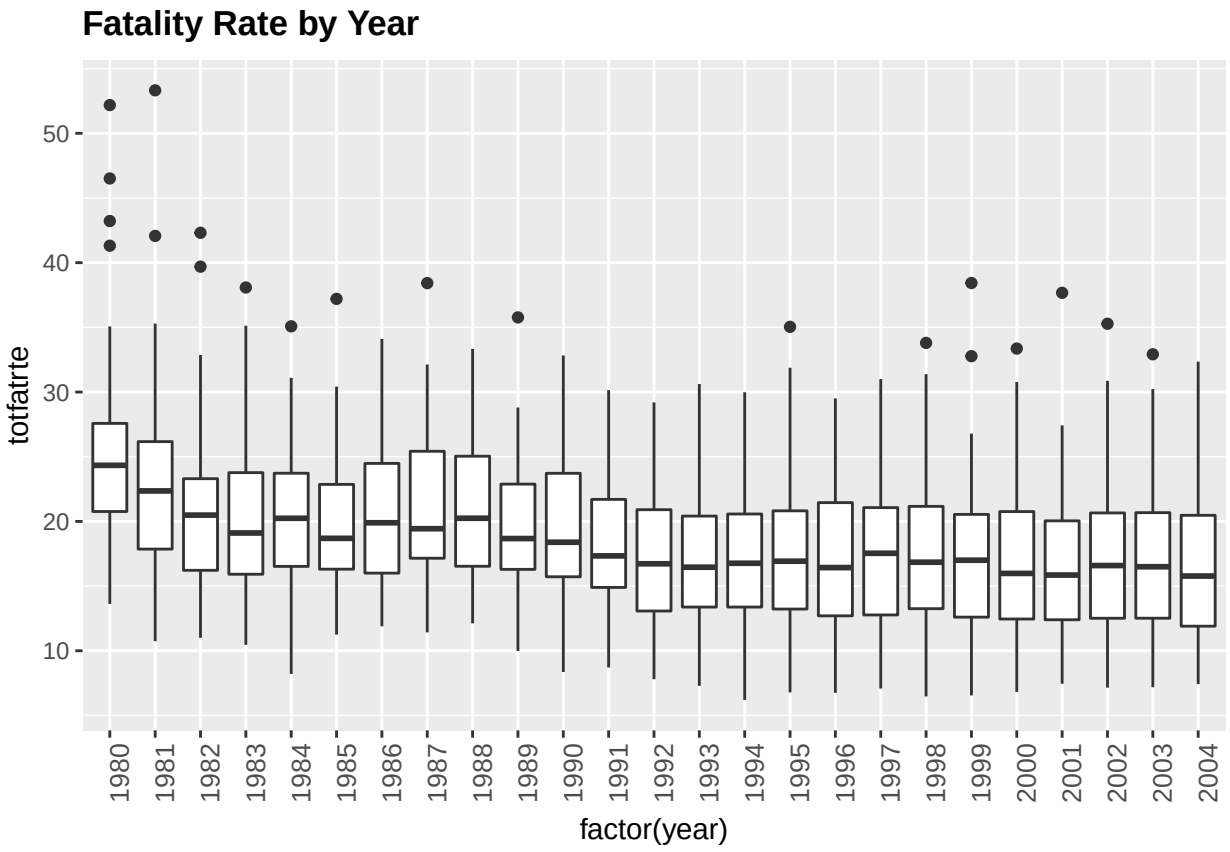
```
ggplot(drivedata, aes(x=year, y=vehicmilespc))+
  geom_point()+geom_smooth(method="loess")+ggtitle("Vehicle miles per person from 1980-2005")
```



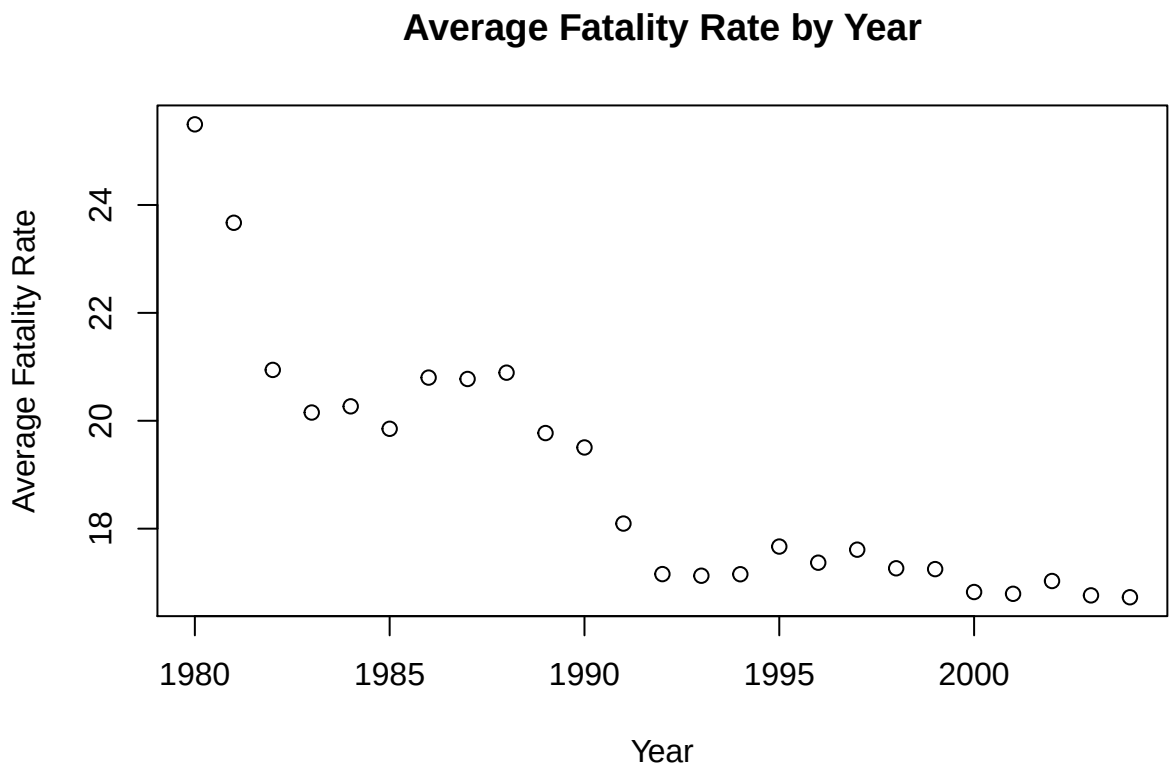
The number of vehicle miles per person appears to steadily increase over time across states in a linear fashion.

2. How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a 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.

```
ggplot(data, aes(factor(year), totfatrte)) +
  geom_boxplot() +
  ggtitle("Fatality Rate by Year") +
  theme(plot.title = element_text(lineheight=1, face="bold"),
        axis.text.x = element_text(size=10,angle=90))
```




```
plot(aggregate(drivedata$totfatrte, list(drivedata$year), mean), ylab = 'Average Fatality Rate', xlab =
```



totfatrte is defined as total fatalities per 100,000 population. Fatalaty Rate trends down from 80s to the early 90s and is somewhat flat from later 90s to 2000s. Steepest decline in the 80s.

```
glm.mod <- glm(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 = drivedata)

summary(glm.mod)
```

```
##
## Call:
## glm(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 = drivedata)
##
## Deviance 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
##
## (Dispersion parameter for gaussian family taken to be 36.09097)
##
##     Null deviance: 48612  on 1199  degrees of freedom
## Residual deviance: 42407  on 1175  degrees of freedom
## AIC: 7735.4
##
## Number of Fisher Scoring iterations: 2
```

The model shows the average change is fatality rate by each year compared to 1980. All years show a decrease in fatalaty rates when compared to 1980. The only year where the change is not significant is 1981. From the years 1982 to 2004 we see significant drops in fatalaty rates ranging from ~4.5% to ~8.8%. It appears when comparing to 1980, from 1982 to 2004 driving did become safer when measured by fatalaty rates.

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.)

```
glm.mod <- glm(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_binary + bac10_

summary(glm.mod)
```

```
##
## Call:
## glm(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_binary +
##     bac10_binary + perse_binary + sb_combined + sl_combined +
##     gdl_binary + perc14_24 + unem + vehicmilespc, data = drivedata)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -14.399   -2.570   -0.345    2.253   21.808
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -8.840e-01  2.515e+00  -0.351 0.725319
## d81                -2.199e+00  8.226e-01  -2.673 0.007619 **
## d82                -6.665e+00  8.486e-01  -7.855 9.08e-15 ***
## d83                -7.591e+00  8.615e-01  -8.812  < 2e-16 ***
## d84                -5.966e+00  8.675e-01  -6.877 9.96e-12 ***
## d85                -6.594e+00  8.862e-01  -7.441 1.93e-13 ***
## d86                -5.936e+00  9.239e-01  -6.425 1.92e-10 ***
## d87                -8.096e+00  1.018e+00  -7.950 4.40e-15 ***
## d88                -8.762e+00  1.090e+00  -8.040 2.20e-15 ***
## d89                -1.014e+01  1.125e+00  -9.008  < 2e-16 ***
## d90                -1.103e+01  1.146e+00  -9.622  < 2e-16 ***
## d91                -1.317e+01  1.166e+00 -11.292  < 2e-16 ***
## d92                -1.499e+01  1.188e+00 -12.613  < 2e-16 ***
## d93                -1.487e+01  1.200e+00 -12.387  < 2e-16 ***
## d94                -1.451e+01  1.219e+00 -11.905  < 2e-16 ***
## d95                -1.413e+01  1.248e+00 -11.329  < 2e-16 ***
## d96                -1.643e+01  1.310e+00 -12.544  < 2e-16 ***
## d97                -1.659e+01  1.325e+00 -12.526  < 2e-16 ***
## d98                -1.730e+01  1.342e+00 -12.888  < 2e-16 ***
## d99                -1.743e+01  1.358e+00 -12.837  < 2e-16 ***
## d00                -1.780e+01  1.375e+00 -12.941  < 2e-16 ***
## d01                -1.870e+01  1.391e+00 -13.442  < 2e-16 ***
## d02                -1.928e+01  1.403e+00 -13.743  < 2e-16 ***
## d03                -1.968e+01  1.408e+00 -13.973  < 2e-16 ***
## d04                -1.923e+01  1.438e+00 -13.374  < 2e-16 ***
## bac08_binary       -2.227e+00  4.874e-01  -4.569 5.42e-06 ***
## bac10_binary       -1.288e+00  3.596e-01  -3.582 0.000355 ***
## perse_binary       -7.740e-01  2.963e-01  -2.612 0.009118 **
## sb_combinedsbprim   4.516e-02  4.883e-01   0.092 0.926334
## sb_combinedsbsecon -1.291e-02  4.269e-01  -0.030 0.975885
## sl_combinedsl65     2.414e+00  5.287e-01   4.567 5.47e-06 ***
## sl_combinedsl70     5.007e+00  7.346e-01   6.816 1.50e-11 ***
## sl_combinedsl75     6.307e+00  8.024e-01   7.860 8.69e-15 ***
## sl_combinedslnone   6.954e+00  1.250e+00   5.565 3.25e-08 ***
## gdl_binary         -2.631e-01  5.126e-01  -0.513 0.607791
## perc14_24           7.951e-02  1.242e-01   0.640 0.522275
## unem                7.583e-01  7.762e-02   9.769  < 2e-16 ***
## vehicmilespc        2.826e-03  9.698e-05  29.141  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 16.16562)
##
##     Null deviance: 48612  on 1199  degrees of freedom
## Residual deviance: 18784  on 1162  degrees of freedom
## AIC: 6784.3
##
## Number of Fisher Scoring iterations: 2
```

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.

The only transformation we made was to binarize the law variables. The justification was explained in more detail earlier in the EDA section. In short, a very small number of fractional numbers betwen 0 and 1 exit for the law variables which were clearly intended to be binary (law was in place or wasn't that year). Based on patterns in the data it seems as if the fracitonal numbers represent if the law was in place only for part of the year. It would be bad for our model to include variables with such an odd distribution, so we transformed those variabels into binary variables by rounding it. While we noticed slight positive skews for the dependent variable totfatrte and unem, we did not believe the skew was big enough to warrant any transformation.

How are the variables bac8 and bac10 defined?

BAC stands for Blood Alcohol Concentration, and BAC laws are legislations that punish driving with certain BAC. In this case, the BAC variables have more than two levels represented by multiple dummy coded variables. The variable three levels - no BAC law, BAC limit = .08, and BAC limit = .10. These three levels are coded in the two variables bac08 and bac10.

Interpret the coefficients on bac8 and bac10.

Both coefficients for bac08 and bac10 are highly statistically significant and the negative coefficients for both bac08 and bac10 suggest that BAC laws reduce the fatality rate under study. Curiously, the coefficient for bac08 for this model suggests that a limit for .08 has a more significant effect than .10, since the coefficient for .08 has a larger absolute value.

Do per se laws have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)

We can also observe that per se laws reduce the fatality rate (coefficient -7.740e-01, p-value < 0.001). Interestingly, our model doesn't support statistically significant support for the primary seat belt law coefficient.

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?

Let's first estimate a fixed effects model at the state level.

```
plm.mod <- 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_binary + bac10_
```

```
## Warning: use of 'plm.data' is discouraged, better use 'pdata.frame' instead
```

```
summary(plm.mod)
```

```
## 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_binary +
##     bac10_binary + perse_binary + sb_combined + sl_combined +
##     gdl_binary + perc14_24 + unem + vehicmilespc, data = plm.data(drivedata,
##     c("state", "year")))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
```

```
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -8.074294 -1.041043  0.045493  0.950932 14.614063
##
## Coefficients:
##                       Estimate  Std. Error t-value  Pr(>|t|)
## d81                -1.49220519  0.41040004 -3.6360 0.0002896 ***
## d82                -3.04893725  0.43948860 -6.9375 6.754e-12 ***
## d83                -3.61009565  0.45153363 -7.9952 3.215e-15 ***
## d84                -4.23574900  0.45831323 -9.2420 < 2.2e-16 ***
## d85                -4.65332276  0.47999307 -9.6946 < 2.2e-16 ***
## d86                -3.53106801  0.51467425 -6.8608 1.133e-11 ***
## d87                -3.53637675  0.59063755 -5.9874 2.872e-09 ***
## d88                -3.82709843  0.65335545 -5.8576 6.174e-09 ***
## d89                -5.12068899  0.68942249 -7.4275 2.195e-13 ***
## d90                -5.21497249  0.71298490 -7.3143 4.933e-13 ***
## d91                -5.91615518  0.72784376 -8.1283 1.148e-15 ***
## d92                -6.73958932  0.75010156 -8.9849 < 2.2e-16 ***
## d93                -7.06270896  0.76241329 -9.2636 < 2.2e-16 ***
## d94                -7.46928076  0.78018922 -9.5737 < 2.2e-16 ***
## d95                -7.14510470  0.80600431 -8.8648 < 2.2e-16 ***
## d96                -7.30104667  0.86605156 -8.4303 < 2.2e-16 ***
## d97                -7.40762067  0.88249305 -8.3940 < 2.2e-16 ***
## d98                -8.01017294  0.89860256 -8.9140 < 2.2e-16 ***
## d99                -8.13639562  0.90945947 -8.9464 < 2.2e-16 ***
## d00                -8.64637562  0.92118529 -9.3861 < 2.2e-16 ***
## d01                -8.38673169  0.93172431 -9.0013 < 2.2e-16 ***
## d02                -7.68725498  0.94059584 -8.1728 8.120e-16 ***
## d03                -7.75520358  0.94474733 -8.2088 6.124e-16 ***
## d04                -8.04874144  0.96965676 -8.3006 2.967e-16 ***
## bac08_binary       -1.29097375  0.32820028 -3.9335 8.889e-05 ***
## bac10_binary       -0.88790969  0.22343701 -3.9739 7.526e-05 ***
## perse_binary       -1.02347458  0.22325624 -4.5843 5.069e-06 ***
## sb_combinedsbprim  -1.26733407  0.34106435 -3.7158 0.0002126 ***
## sb_combinedsbsecon -0.29447937  0.25058871 -1.1752 0.2401855
## sl_combinedsl65    -0.83122073  0.31486369 -2.6399 0.0084079 **
## sl_combinedsl70    -0.59035476  0.46600114 -1.2669 0.2054726
## sl_combinedsl75    -1.90397421  0.50886006 -3.7416 0.0001921 ***
## sl_combinedslnone  -0.67592558  0.72910565 -0.9271 0.3540954
## gdl_binary         -0.35895778  0.28070452 -1.2788 0.2012426
## perc14_24           0.29159504  0.09780928  2.9813 0.0029329 **
## unem               -0.54541851  0.06088084 -8.9588 < 2.2e-16 ***
## vehicmilespc        0.00091135  0.00011043  8.2530 4.322e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4461
## R-Squared:      0.63235
## Adj. R-Squared: 0.60466
## F-statistic: 51.8329 on 37 and 1115 DF, p-value: < 2.22e-16
```

How do the coefficients on bac08, bac10, perse, and sbprim compare with the pooled OLS estimates?

The BAC coefficients for the fixed effects model are relatively similar to those obtained in the pooled OLS model. In both models, BAC laws suggest a reduced fatality rate, while a BAC limit of .08 seems to have a larger effect on fatality rate than a BAC limit of .10.

In the fixed effect model, the coefficient for per se laws is negative just as in the previous model, suggesting per se laws reduce the fatality rates. However, the coefficient for per se laws is one order of magnitude larger than in the pooled OLS model. Finally, while the pooled OLS model coefficient for seat belt primary laws was not statistically significant, the fixed effects model has a highly statistically significant model and a negative coefficient, suggesting that primary seat belt laws do in fact reduce fatality rates, which makes sense.

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?

The estimates from the fixed effects model should be more reliable. Simple regression suffers more from omitted variable bias, specifically we need to assume that onobserved effects a_i are not correlated with the explanatory variables. If this is not true we have heterogeneity bias. For the fixed effects estimator we difference away the time constant unobserved effects.

First Difference: $\Delta y_t = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$

The key assumption for the first difference equation is that Δu_i is not correlated with Δx_i. This is a version of strict exogeneity. If we violate this assumption we will have bias in the estimators. We also need to assume that Δx_i varies across states and time. We do see that there is variation across explanatory across states and time intervals.

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

We would prefer to use of fixed effects model for excercise 4. For our data our unobserved effects (a_i), are likely correlated with our explanatory variables. Things such as road and weather conditions are likely specific to states and are slow to change over time and should have an effect on fatality rate. In addition since we have data from states we might not be able to consider all our observations to be random draws from a large enough population. This also suggests that a fixed effect model would be more appropriate.

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.

Our coefficient for *vehicmilespc* is 0.0009. This means that there is a positive relationship between miles driven per capita and fatality rates, or the more miles driven the more likely you are the be in a fatal accident. Specifically our estimator suggests that for every 1000 mile increase in miles per capita we expect a 0.9% increase in the fatality rate. The result is significant so we are confident this is a true effect and not due to random variation.

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 we have heteroskedasticity in the error of the model we will have both bias in the coefficient estimates and the standard errors will be incorrect. If we have positive serial correlation we will tend to overestimate coefficients and underestimate standard errors. These lower standard errors would lead to artificially low p-values and results that overstate the statistical significance. The opposite would be true of negative serial correlation - with artificaily higher standard errors we would be more likely to commit type II errors.