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

W271 Instructional Team

Fall 2018

page limit: 20 pages

Description of the Lab

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

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

```r
load('driving.RData')
#str(data)
desc
```

```
##       variable                                           label
## 1         year                               1980 through 2004
## 2        state             48 continental states, alphabetical
## 3         sl55                                speed limit == 55
## 4         sl65                                speed limit == 65
## 5         sl70                                speed limit == 70
## 6         sl75                                speed limit == 75
## 7       slnone                                  no speed limit
## 8     seatbelt      =0 if none, =1 if primary, =2 if secondary
## 9       minage                             minimum drinking age
## 10     zerotol                             zero tolerance law
## 11         gdl                  graduated drivers license law
## 12       bac10                         blood alcohol limit .10
## 13       bac08                         blood alcohol limit .08
## 14       perse administrative license revocation (per se law)
## 15      totfat                         total traffic fatalities
## 16     nghtfat                       total nighttime fatalities
## 17     wkndfat                         total weekend fatalities
## 18   totfatpvm          total fatalities per 100 million miles
## 19  nghtfatpvm      nighttime fatalities per 100 million miles
## 20  wkndfatpvm        weekend fatalities per 100 million miles
## 21    statepop                                 state population
## 22   totfatrte          total fatalities per 100,000 population
## 23  nghtfatrte      nighttime fatalities per 100,000 population
## 24  wkndfatrte        weekend accidents per 100,000 population
## 25  vehicmiles                 vehicle miles traveled, billions
## 26        unem                        unemployment rate, percent
```

```
## 27    perc14_24         percent population aged 14 through 24
## 28     sl70plus                        sl70 + sl75 + slnone
## 29       sbprim                  =1 if primary seatbelt law
## 30      sbsecon                =1 if secondary seatbelt law
## 31          d80                           =1 if year == 1980
## 32          d81
## 33          d82
## 34          d83
## 35          d84
## 36          d85
## 37          d86
## 38          d87
## 39          d88
## 40          d89
## 41          d90
## 42          d91
## 43          d92
## 44          d93
## 45          d94
## 46          d95
## 47          d96
## 48          d97
## 49          d98
## 50          d99
## 51          d00
## 52          d01
## 53          d02
## 54          d03
## 55          d04                           =1 if year == 2004
## 56 vehicmilespc
```

```r
#head(data)
#describe(data)
```

The dataset contains about 1200 observations ranging from 1980 to 2004 for the 48 contentential states. The observed variables include: Speed limits (`slXX`), seat belt and zero tolerance laws, graduated driver, blood alcohol level (`bacXX`), per se are in percent of year by months. `Sbl70plus`, `sbprim`, `sbsecon` and `dXX` variables are simply derivatives or dummy variables of the other variables in the data set.

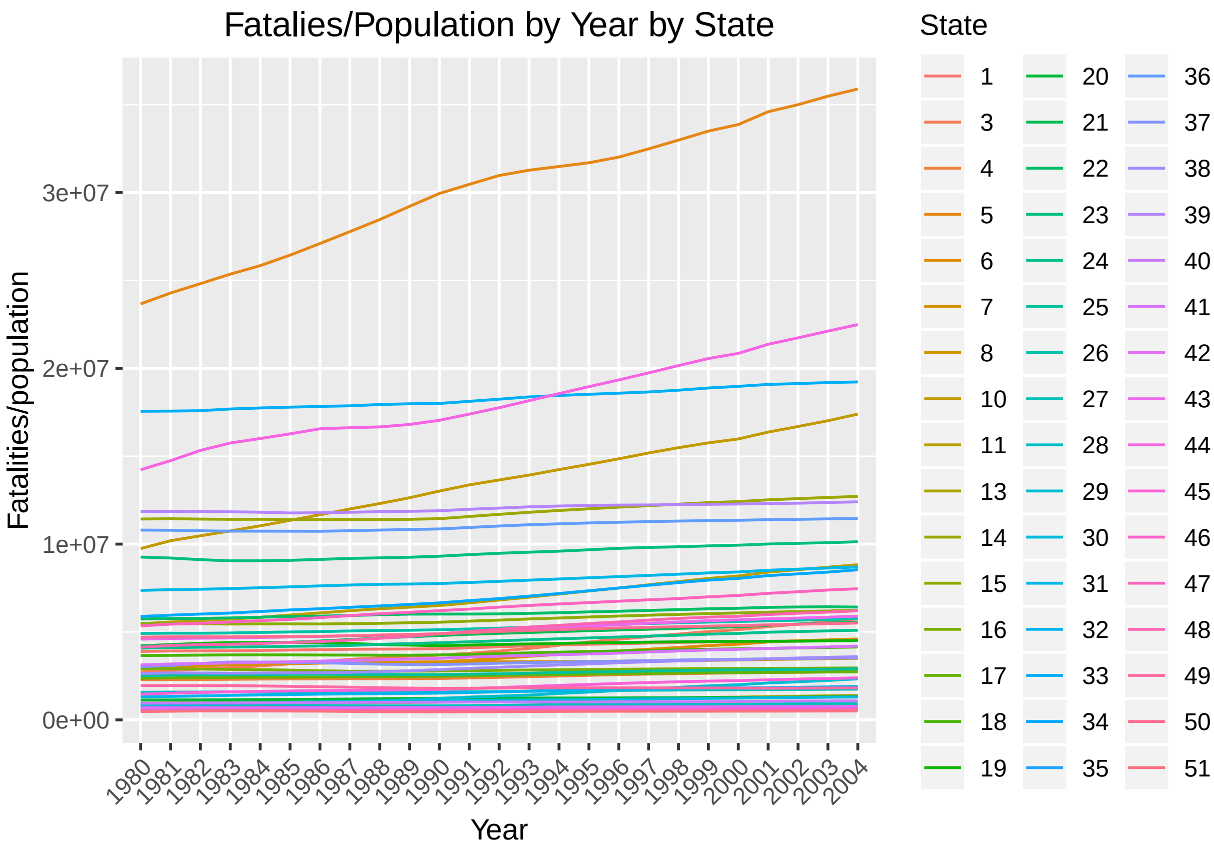
Our research question is whether or not traffic laws can affect total fatalities. Total fatalities is a function of population, vehicle miles, traffic laws and unobservable variables. Our dataset contains 9 fatality-related variables, some normalized in various ways. We will not consider the weekend and night fatality variables as we are focusing on total fatalities and not when they occurred.

```r
#df=data[,c('year','state','totfatrte','sl55','sl65','sl70','sl75','slnone','seatbelt','minage','zeroto
```

```r
#table(df$year)
#table(df$state)
```

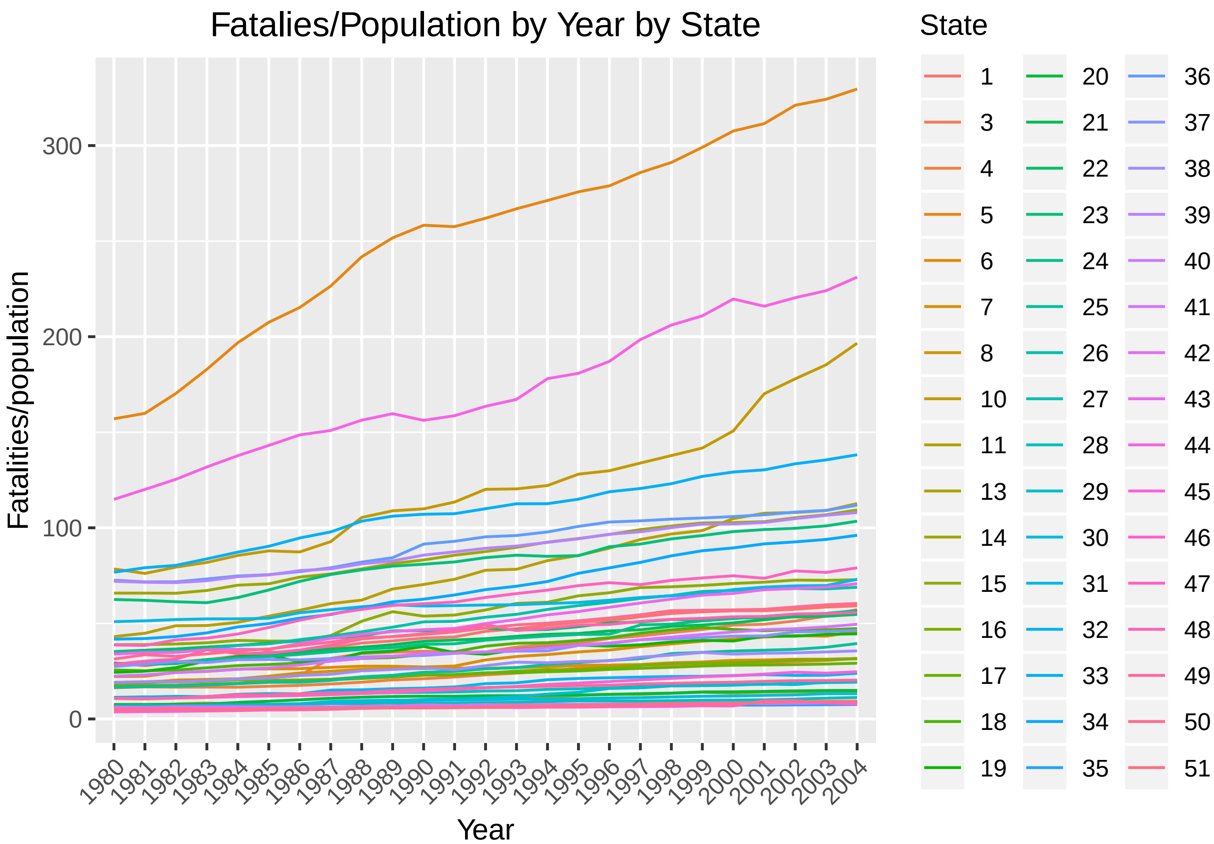
```r
#pairs(~totfatrte+zerotol+gdl+perse+vehicmilespc, data=tmp, lower.panel=panel.smooth)

#ggplot(tmp, aes(x=totfatrte, y=zerotol)) + geom_point()
```

```r
ggplot(data,aes(y=statepop,x=factor(year),group=factor(state),color=factor(state)))+geom_line()+theme(a
```



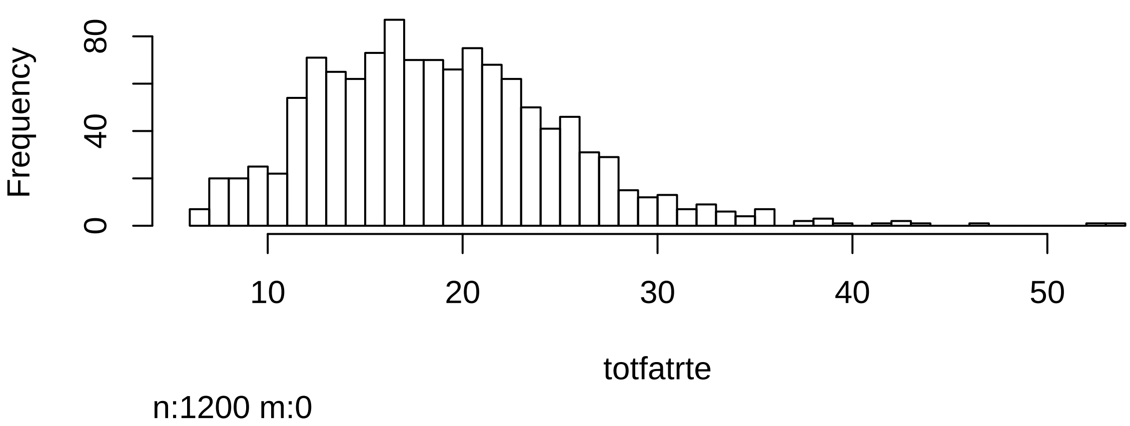
Three states, 5, 44 and 10, significantly increased in population while other states were relatively flat. This suggests regressing against a population normalized fatality measure, such as totfatrte, would be most useful for examining causing inferences of state driving laws.

```
ggplot(data,aes(y=vehicmiles,x=factor(year),group=factor(state),color=factor(state)))+geom_line()+theme
```

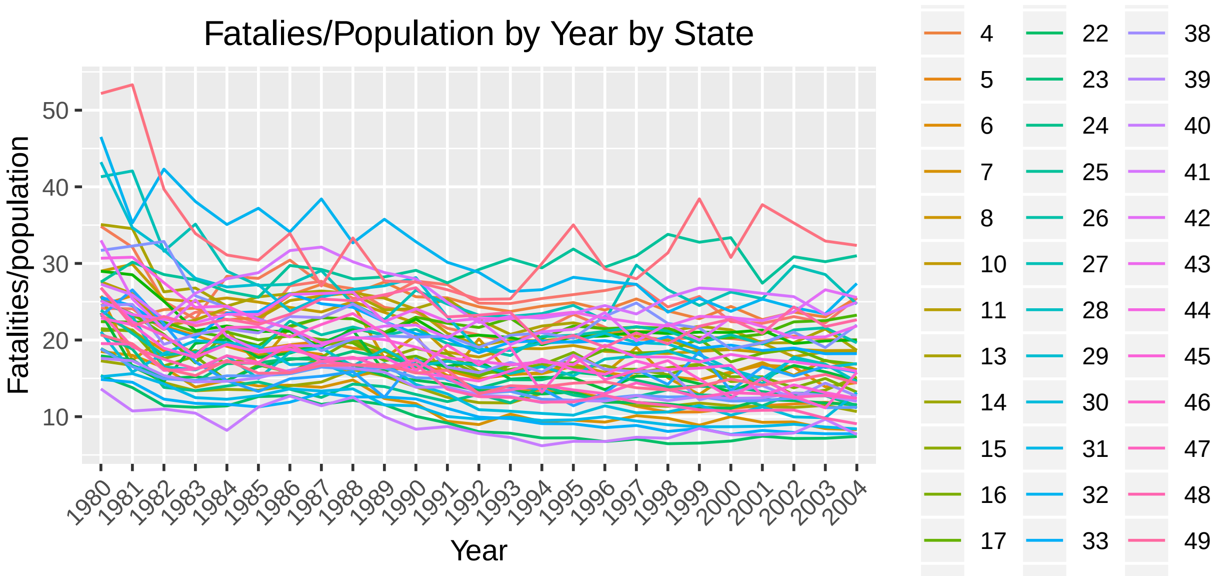


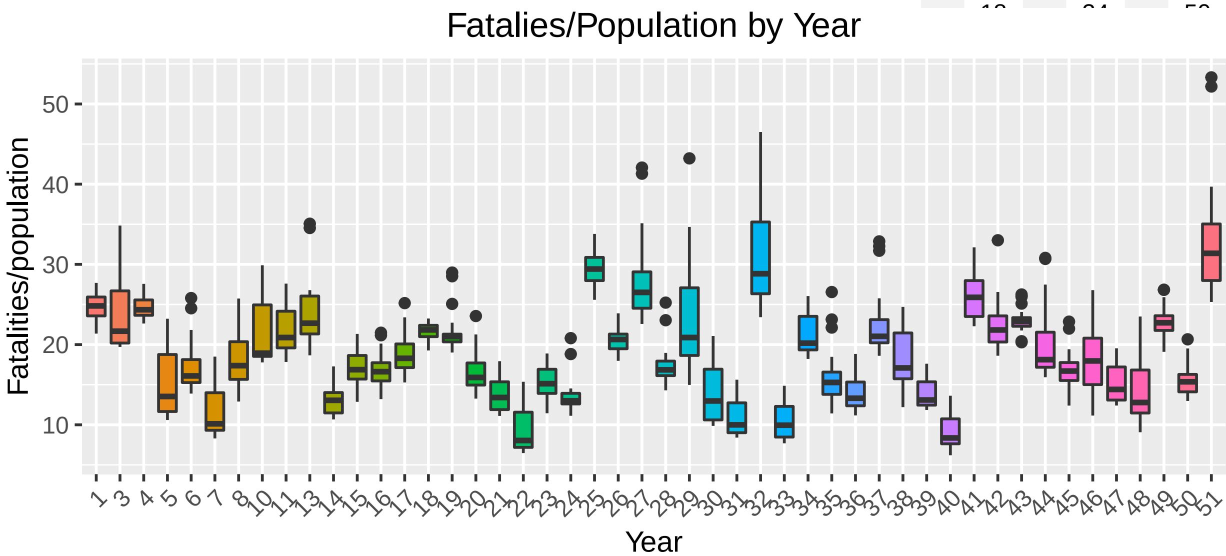
Vehicle miles has roughly similar trends for almost all states. This further confirms totfatrte as a dependent variable for traffic law causal inference as vehicle miles is more "stable" among the states than population through out time.

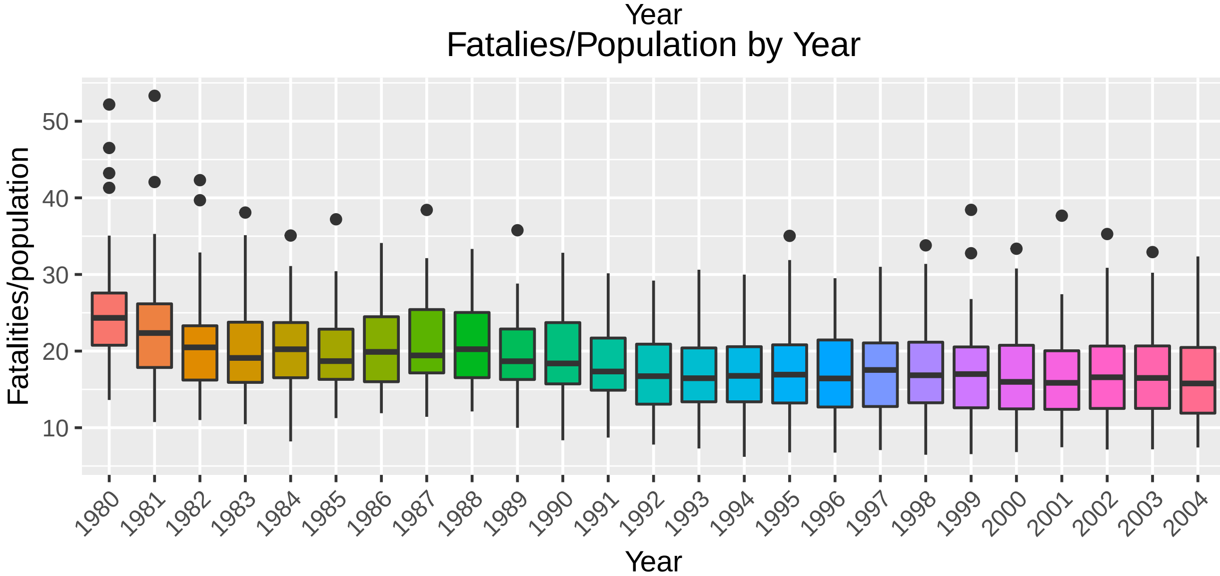
```
plottot('totfatrte')
```




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


Fatalies/Population by Year by State
Fatalities/population
Year


Fatalies/Population by Year
Fatalities/population
Year


Fatalies/Population by Year
Fatalities/population
Year

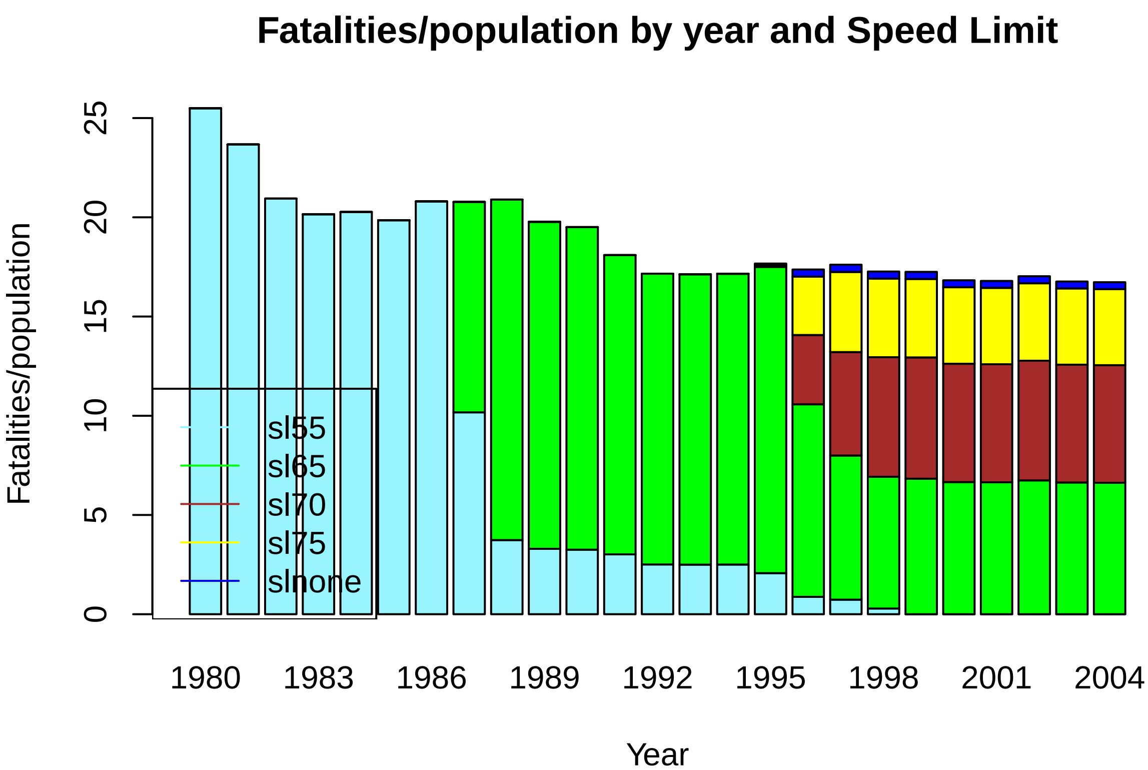
HISTOGRAM ISN'T DISPLAYING TITLES

Mean fatalities/population has dropped by about 10% from 1980 to 2004. State 51 has persistently stayed near the top of the fatalies while State 38 seems to stay near the bottom. The range of fatalities changes throughout time, with the highest fatality rate dropping from over 50% in 1980 to under 45% in 2004. The minimum values for fatalities drop at well, but less overall, from about just over 10 in 1980 to just under 10 in 2004."'{}

```
# I merged this back into the totfatrte with ggplot
#conditional_plot(data, data$totfatrte, data$state, "Total Fatality Rate per 100,000 population by Stat
```

There are several states that have higher overall fatality rates than the others. States 25, 32 and 51 may be further explored to see the relationships with the observed variables.

```
tmp=data %>% select(year,totfatrte,sl55,sl65,sl70,sl75,slnone) %>% group_by(year) %>% summarize_all(fun
tmp=tmp$totfatrte*tmp
tmp$year=tmp$year/sqrt(tmp$totfatrte)
tmp$totfatrte=NULL
rownames(tmp)=tmp$year
tmp$year=NULL
tmp=t(tmp)
barplot(as.matrix(tmp),col=c('cadetblue1','green','brown','yellow','blue'),xlab='Year',ylab='Fatalities
legend('bottomleft',legend=rownames(tmp),col=c('cadetblue1','green','brown','yellow','blue'),lty=c(1,1,
```

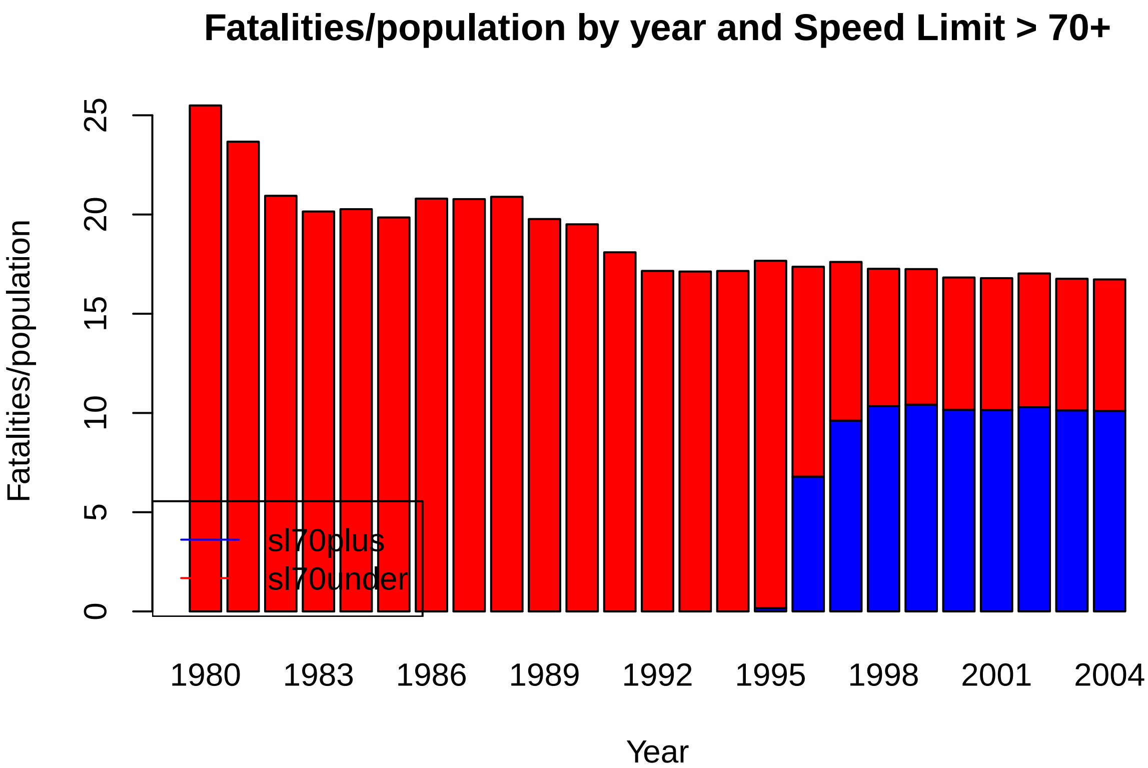


SS I'm a little confused as to who these plot are put together. same question on all stacked-like bar charts. Could we view stacked bar charts of the Speed limits and impose and average line for `totfatrte` on top

Initially, the speed limit (`slXX`) increase does not seem to affect the average total fatality rate across states immediately. In 1986-1987, speed limits increased in many states and in 1988-1991, fatalities fell by about

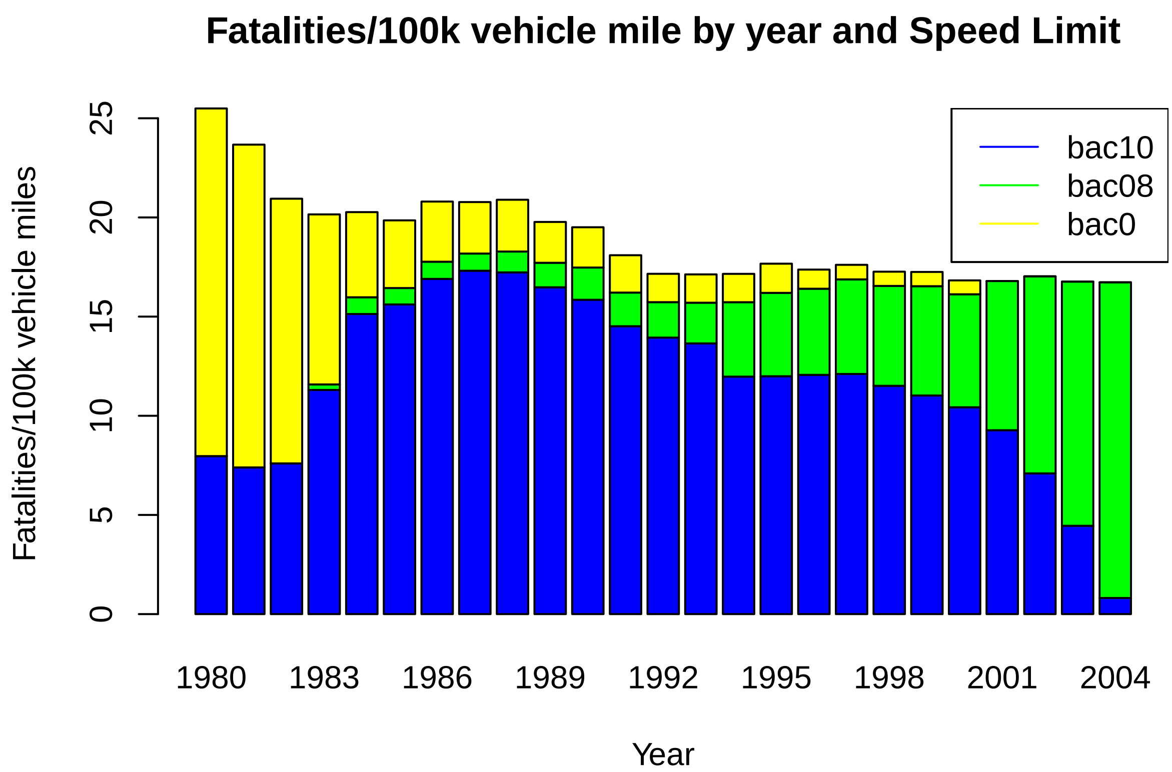
10%. The variable may be a candidate as an interaction variable with time.

```
tmp = data %>% select(year,totfatrte,sl70plus) %>% group_by(year) %>% summarize_all(funs(mean)) %>% muta
tmp=tmp$totfatrte*tmp
tmp$year=tmp$year/sqrt(tmp$totfatrte)
tmp$totfatrte=NULL
rownames(tmp)=tmp$year
tmp$year=NULL
tmp=t(tmp)
barplot(as.matrix(tmp),col=c('blue','red'),xlab='Year',ylab='Fatalities/population',main='Fatalities/pop
legend('bottomleft',legend=rownames(tmp),col=c('blue','red'),lty=c(1,1,1,1))
```



To further examine speed limits, we split the variables into two groups - states with speed limits under 70 and states with speed limits over 70. From the chart, it appears that speed limit over 70+ does not impact fatalities in a significant way.

```
tmp = data %>% select(year,totfatrte,bac10,bac08) %>% group_by(year) %>% summarize_all(funs(mean)) %>% m

tmp=tmp$totfatrte*tmp
tmp$year=tmp$year/sqrt(tmp$totfatrte)
tmp$totfatrte=NULL
rownames(tmp)=tmp$year
tmp$year=NULL
tmp=t(tmp)
barplot(as.matrix(tmp),col=c('blue','green','yellow'),xlab='Year',ylab='Fatalities/100k vehicle miles',m
legend('topright',legend=rownames(tmp),col=c('blue','green','yellow'),lty=c(1,1,1,1))
```



`bac0` represents where states did not have laws relating to blood alcohol content.

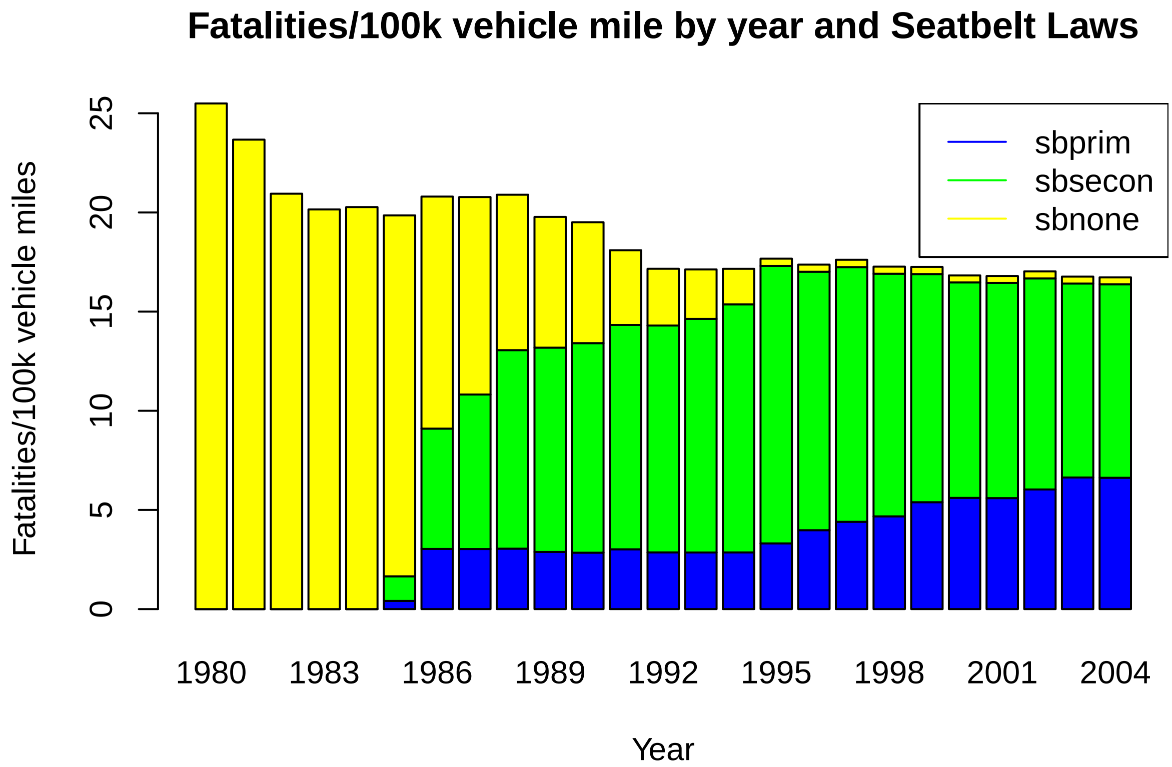
Blood Alcohol Content laws appear to have an immediate impact on fatalities. More interestingly, the graph suggests transformig the BAC laws into a binary variabe, as a `bac08` and `bac10` does not appear to affect fatalities, but the initial implementation of drinking-related laws has an impact.

We will performed t-tests for 2001 and 2002 when the `bac08` and `bac10` are closest to 50% between the states. The two t-tests are performed to avoid the general downward trend of fatalities impacting the analysis. In both t-tests, the H_0: differences in means = 0 are not rejected. From the graph above, there aren't likely to be lagged effects for `bac10` and `bac08` as `bac10` decreased from 1997-2004, but fatalities appear roughly the same. The lack of lagged effect makes sense as bac laws will immediately impact drunk driver and remove theme from the roads.

```
#t.test(data[data$year==2001 & data$bac10==1,"totfatrte"],data[data$year==2001 & data$bac08==1,'totfatr
#t.test(data[data$year==2002 & data$bac10==1,"totfatrte"],data[data$year==2002 & data$bac08==1,'totfatr
```

We may later test this with a F-test of `bac08` and `bac10`. If both show insignificance in a multivariate regression but rejects the H_0 in a f-test, we should convert it into a binary variable of BAC laws or none

```
tmp = data %>% select(year,totfatrte,sbprim,sbsecon) %>% group_by(year) %>% summarize_all(funs(mean)) %

tmp=tmp$totfatrte*tmp
tmp$year=tmp$year/sqrt(tmp$totfatrte)
tmp$totfatrte=NULL
rownames(tmp)=tmp$year
tmp$year=NULL
tmp=t(tmp)
barplot(as.matrix(tmp),col=c('blue','green','yellow'),xlab='Year',ylab='Fatalities/100k vehicle miles',m
legend('topright',legend=rownames(tmp),col=c('blue','green','yellow'),lty=c(1,1,1,1))
```

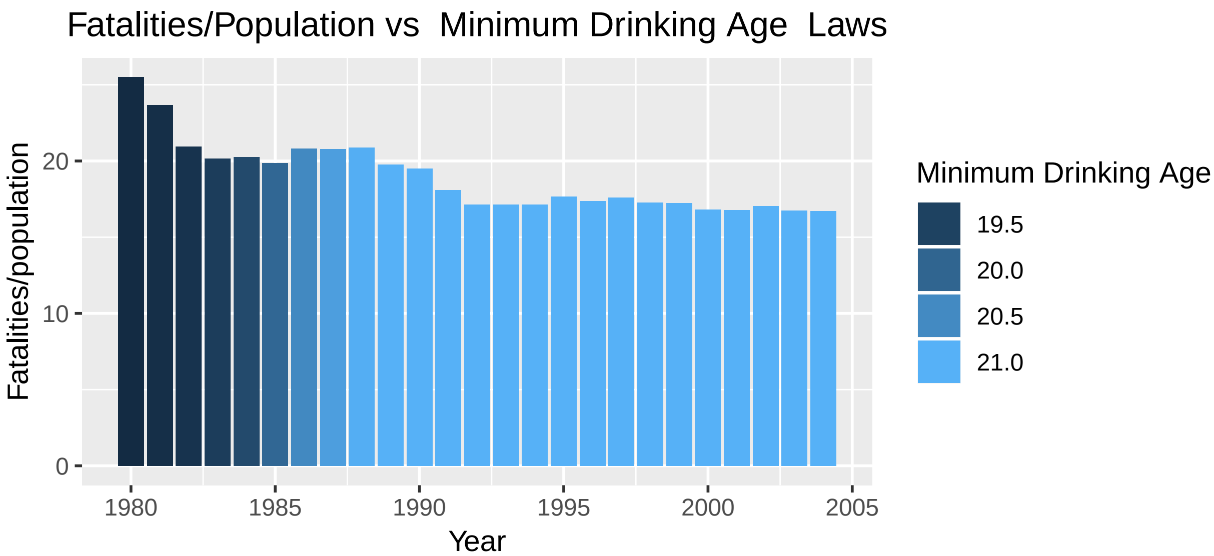


Seatbelts appear to have a simulatenous decrease with fatalities. It should be included as an independent variable. Much like BAC levels, the mix of seatbelt laws does not appear to affect fatalities. There appears to be slightly lagged effect on the binary seatbelt law - possibly the population is getting in the habit of putting on seatbelts. After most states have implemented at least primary wseatbelt laws (1992-1995), though, the average fatality rate evens off.

We performed t-test for 1999 and 2000 where the percentages are closer to even for sbprim and sbsecon. For both t-tests, H_0 is not rejected and there does not appear to be any contemporaneous impact difference between seatbelt laws. Despite the increase in sbprim mix from 1995 to 2004, the fatalities from 1995 to 2004 is similar. There is unlikely to be a lagged impact of seatbelt law differences. Finally, we also note that state 30 does not have seatbelt laws throughout the period.

```
#t.test(data[data$year==1999 & data$sbprim==1,"totfatrte"],data[data$year==1999 & data$sbsecon==1,'totf
#t.test(data[data$year==2000 & data$sbprim==1,"totfatrte"],data[data$year==2000 & data$sbsecon==1,'totf
```

```
tmp = data %>% group_by(year) %>% summarize_all(funs(mean))
plotmix(tmp,'minage','Minimum Drinking Age')
```




```
tmp2=data %>% group_by(state) %>% summarize_all(funs(mean)) %>% as.data.frame
tmp2=tmp2[tmp2$minage %in% unique(data$minage),c('state','minage')]
tmp2=tmp2[!(tmp2$state %in% c(47,51)),c('state','minage')]
tmp2
```

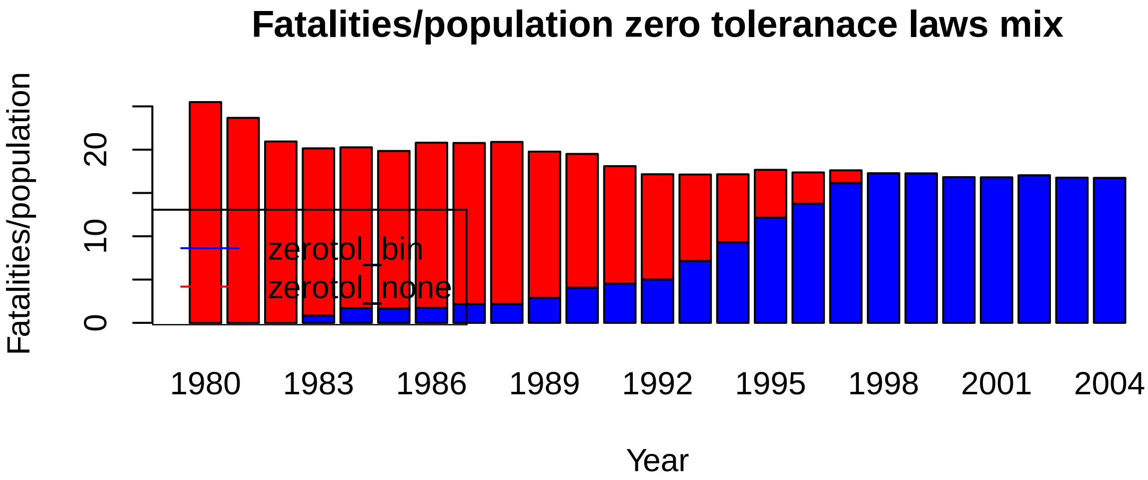
```
##    state minage
## 3      4     21
## 4      5     21
## 11    14     21
## 12    15     21
## 15    18     21
## 20    23     21
## 23    26     21
## 26    29     21
## 29    32     21
## 32    35     21
## 35    38     21
## 36    39     21
## 42    45     21
## 45    48     21
```

These might be better as distinct colors since we only have 4 categories - there are actually quite a few minage - like 5-6 - not sure what's the deal with these - I'll let it as above for now? Suggestions welcome!

Minmum drinking age appears to have some effect on fatalies. Interestingly, all states listed did not have minimum age laws changed during the period. The mean `minage` from 1980-1990 trended higher to 21 in 1990. If we were to focus on `minage`, we can split the data into 2 sets and run separate analysis, detrend and analyze the impact of `minage` on fatalities. We should also interact this variable with BAC variables as raising minimum drinking age may potentially offset some effects of BAC laws. The interaction term is expected to have a negative coefficient.

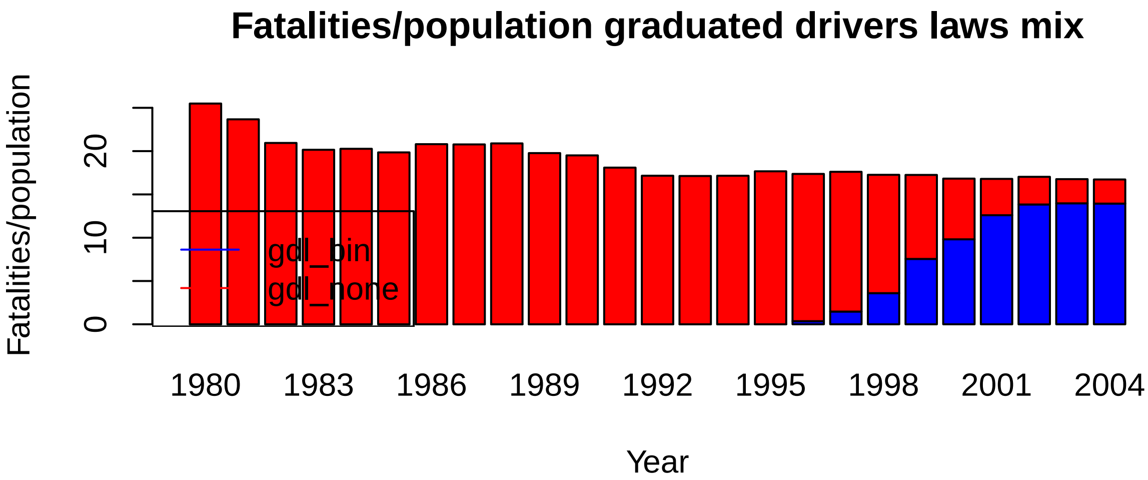
```
#plotmix(tmp,'zerotol','Zero Tolerance')
tmp=data
tmp$zerotol_bin=ifelse(tmp$zerotol>0,1,0)
```

```r
tmp = tmp %>% select(year,totfatrte,zerotol_bin) %>% group_by(year) %>% summarize_all(funs(mean)) %>% mu
tmp=tmp$totfatrte*tmp
tmp$year=tmp$year/sqrt(tmp$totfatrte)
tmp$totfatrte=NULL
rownames(tmp)=tmp$year
tmp$year=NULL
tmp=t(tmp)
barplot(as.matrix(tmp),col=c('blue','red'),xlab='Year',ylab='Fatalities/population',main='Fatalities/po
legend('bottomleft',legend=rownames(tmp),col=c('blue','red'),lty=c(1,1,1,1))
```



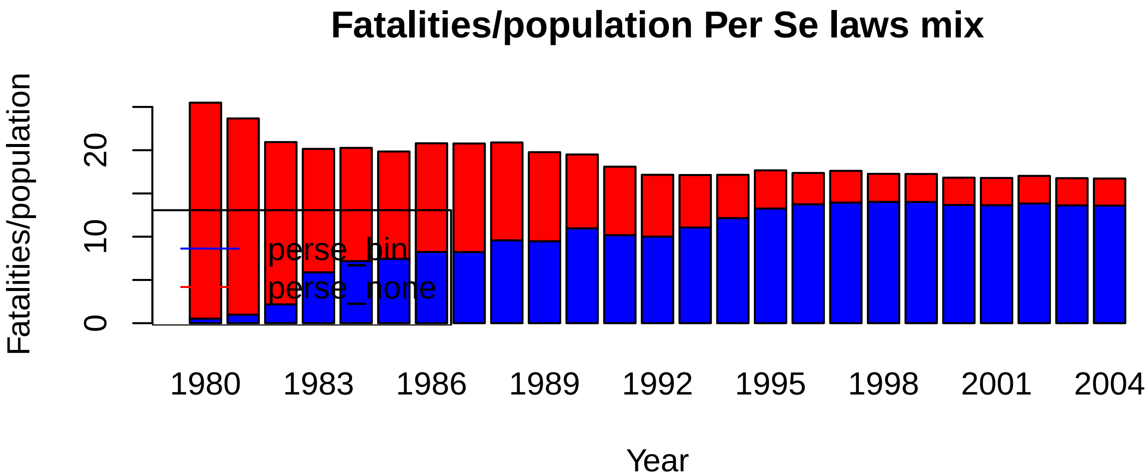
Zero tolerance laws do not appear to have a contemperous impact on fatalities based on the changes in laws from 1992-1997. It may potentally have a long-tailed effect.

```r
#tmp = data %>% group_by(year) %>% summarize_all(funs(mean))
#plotmix(tmp,'gdl','Graduated Driver License')
tmp=data
tmp$gdl_bin=ifelse(tmp$gdl>0,1,0)

tmp = tmp %>% select(year,totfatrte,gdl_bin) %>% group_by(year) %>% summarize_all(funs(mean)) %>% mutat
tmp=tmp$totfatrte*tmp
tmp$year=tmp$year/sqrt(tmp$totfatrte)
tmp$totfatrte=NULL
rownames(tmp)=tmp$year
tmp$year=NULL
tmp=t(tmp)
barplot(as.matrix(tmp),col=c('blue','red'),xlab='Year',ylab='Fatalities/population',main='Fatalities/po
legend('bottomleft',legend=rownames(tmp),col=c('blue','red'),lty=c(1,1,1,1))
```



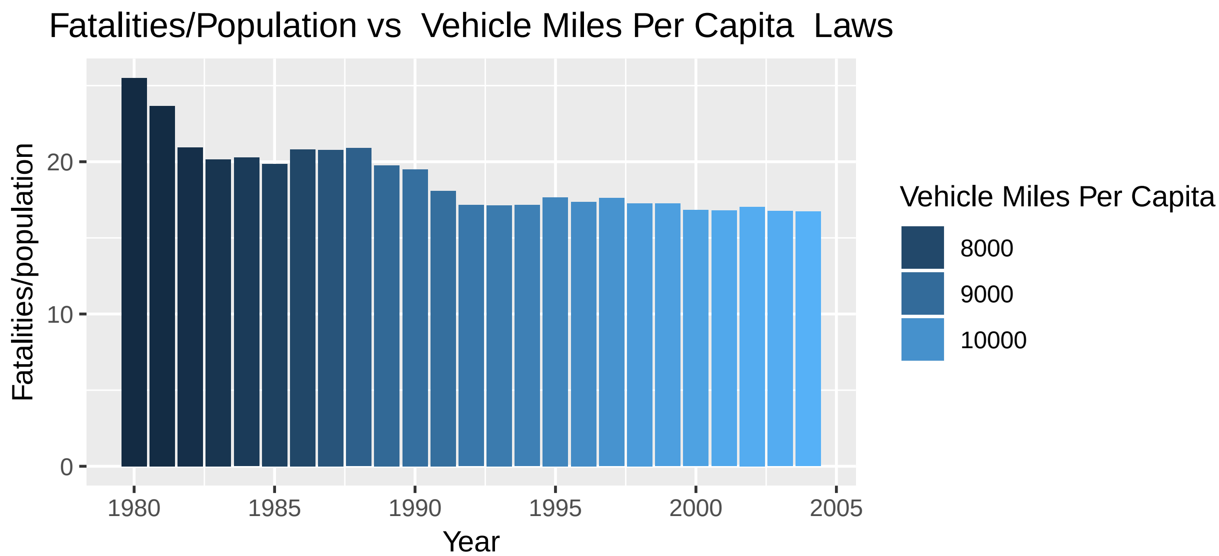
Graduated driver license laws do not appear to impact fatalies very much. Most likely, it will not impact fatalities even accounting for time lags. The changes from graduated license laws from 1999 to 2004 barely impacted fatalities with or without lag effects.

```
#tmp = data %>% group_by(year) %>% summarize_all(funs(mean))
#plotmix(tmp,'perse','Per Se Law')

tmp=data
tmp$perse_bin=ifelse(tmp$perse>0,1,0)

tmp = tmp %>% select(year,totfatrte,perse_bin) %>% group_by(year) %>% summarize_all(funs(mean)) %>% muta
tmp=tmp$totfatrte*tmp
tmp$year=tmp$year/sqrt(tmp$totfatrte)
tmp$totfatrte=NULL
rownames(tmp)=tmp$year
tmp$year=NULL
tmp=t(tmp)
barplot(as.matrix(tmp),col=c('blue','red'),xlab='Year',ylab='Fatalities/population',main='Fatalities/po
legend('bottomleft',legend=rownames(tmp),col=c('blue','red'),lty=c(1,1,1,1))
```



Per se law does not appear to impact fatalities. The increase from 1982 to 1983 did not appear to have a contemporaneous or lagged impact on fatalities. Per se laws may have interactions with BAC laws as it increasese the “harshness” of bac laws.

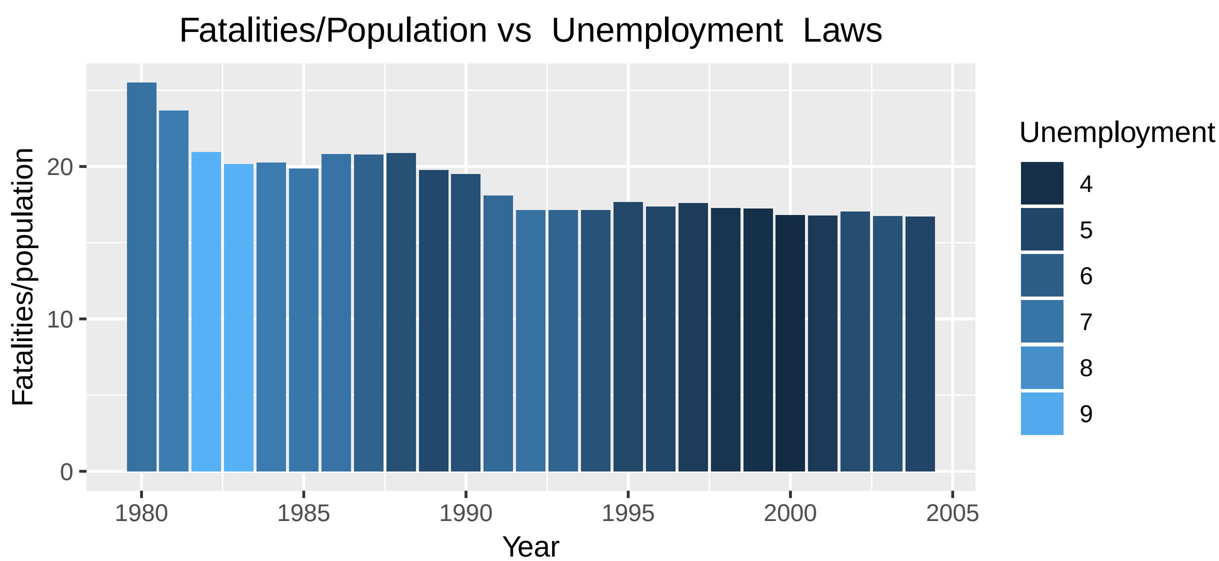
```
tmp = data %>% group_by(year) %>% summarize_all(funs(mean))
plotmix(tmp,'vehicmilespc','Vehicle Miles Per Capita')
```



Vehicle miles may be secondarily affected by traffic laws such as graduated license laws. Preliminarly, fatalities appear to decrease as it increases. This makes no sense and is likely to be trending effect through time.

While all states increased vehicmilespc, state 46 was interesting in that there was a large increase in 2001-2001 followed by a large drop back down to historical trend by 2003-2004. Closer, state-specific reasoning may be required. Given that we do not have that information, we will not examine further into it. **SHOW GRAPH**

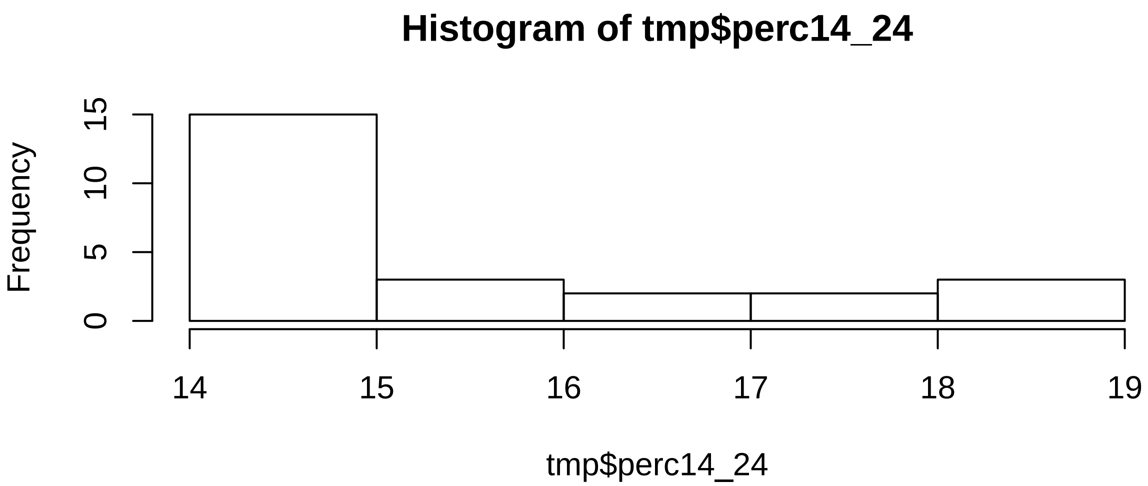
```
plotmix(tmp,'unem','Unemployment')
```



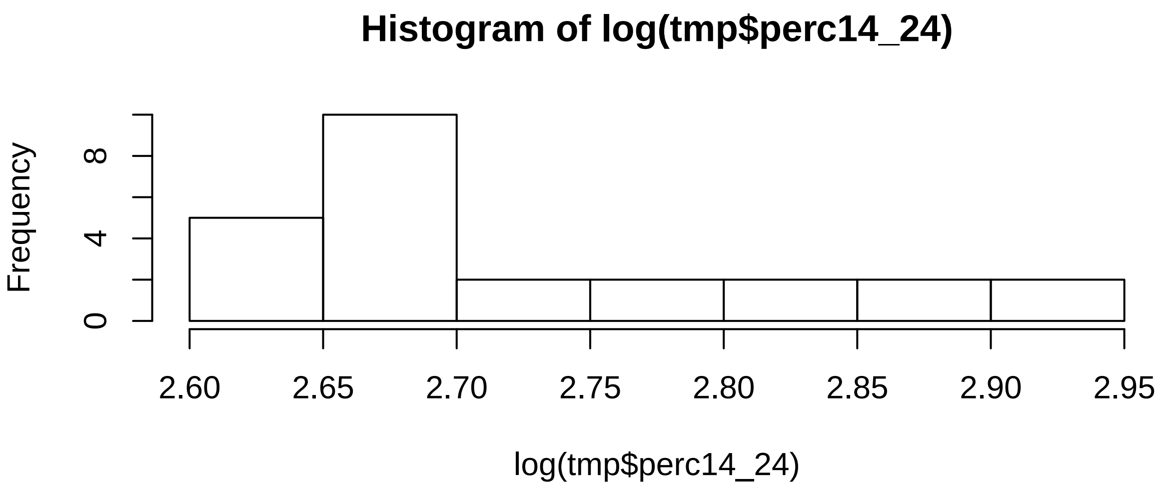
##Here I think using unique colors would look good. esp since the pattern is more random

Fatalities do not appear to be affected by unemployment rates at all. In fact, the pattern appears random. We expect this variable to have a β close to 0 in regressions.

```
hist(tmp$perc14_24)
```

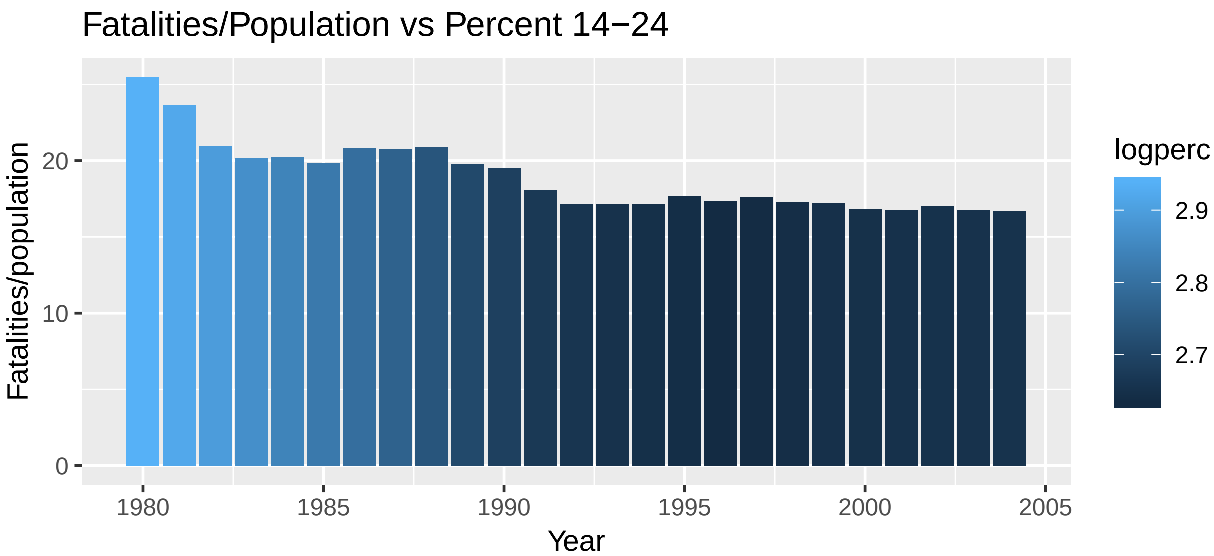



```r
hist(log(tmp$perc14_24))
```



The variable is extremely left skewed and a log transformation will be performed. It appears that percent 14-24 in is negatively correlated with fatalities. This makes no sense is it may simply be an artifact of general trend.

```r
#plotmix(tmp,'pe`rc14_24','Percent 14-24')
tmp=data
tmp$logperc14_24=log(data$perc14_24)
tmp = data %>% group_by(year) %>% mutate(logperc=log(perc14_24)) %>% summarize_all(funs(mean))
ggplot(tmp,aes(y=totfatrte,x=year,fill=logperc))+geom_bar(stat='identity')+labs(x='Year',y='Fatalities/
```



While all states decreased in their population of 14-24 year olds, a few states increased in the ratio. Most significantly, state 45's increase stood out amongst all the states. Again, without more state specific infomration, it's difficult to further examine the increase. ## PUT IN A GRAPH SHOWING THIS

From the EDA, there are variables that appear to be "incorrectly" correlated with fatalities, such as vehicmilespc. Others such as BAC may be better transformed into a binary on-off variable. Per se, zero tolerance, graduated license, minimum drinking age laws appears to have no effect while seatbelts appear to have a contemporaneous impact on fatalites. Speed Limit laws appear to have a lagged effect.

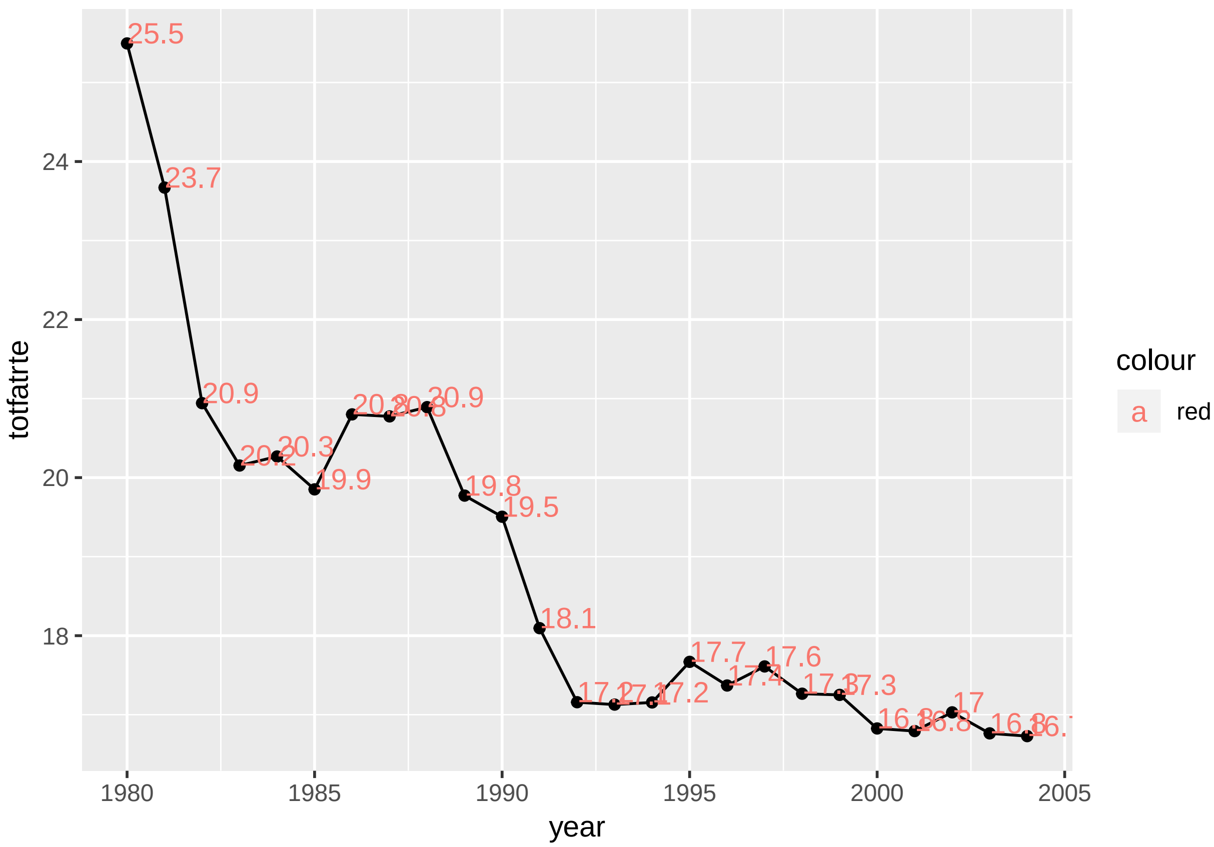
We will first examine the general time trend of fatalities. Recall that our fatalities variable, `totfatrte`, is Fatalities per 100,000 population is already normalized by population, so proper analysis of impact of traffic laws on fatalities can be analyzed. Note that traffic laws are most likely uncorrelated with state population as shown below and it (??) has a lower correlation than vehicle miles suggesting that normalizing fatalities on vehicle miles may be better since it's more likely to be independent.

SS we can include a chart with state total fatality rate v popultion? that would confirm there's not relationship.

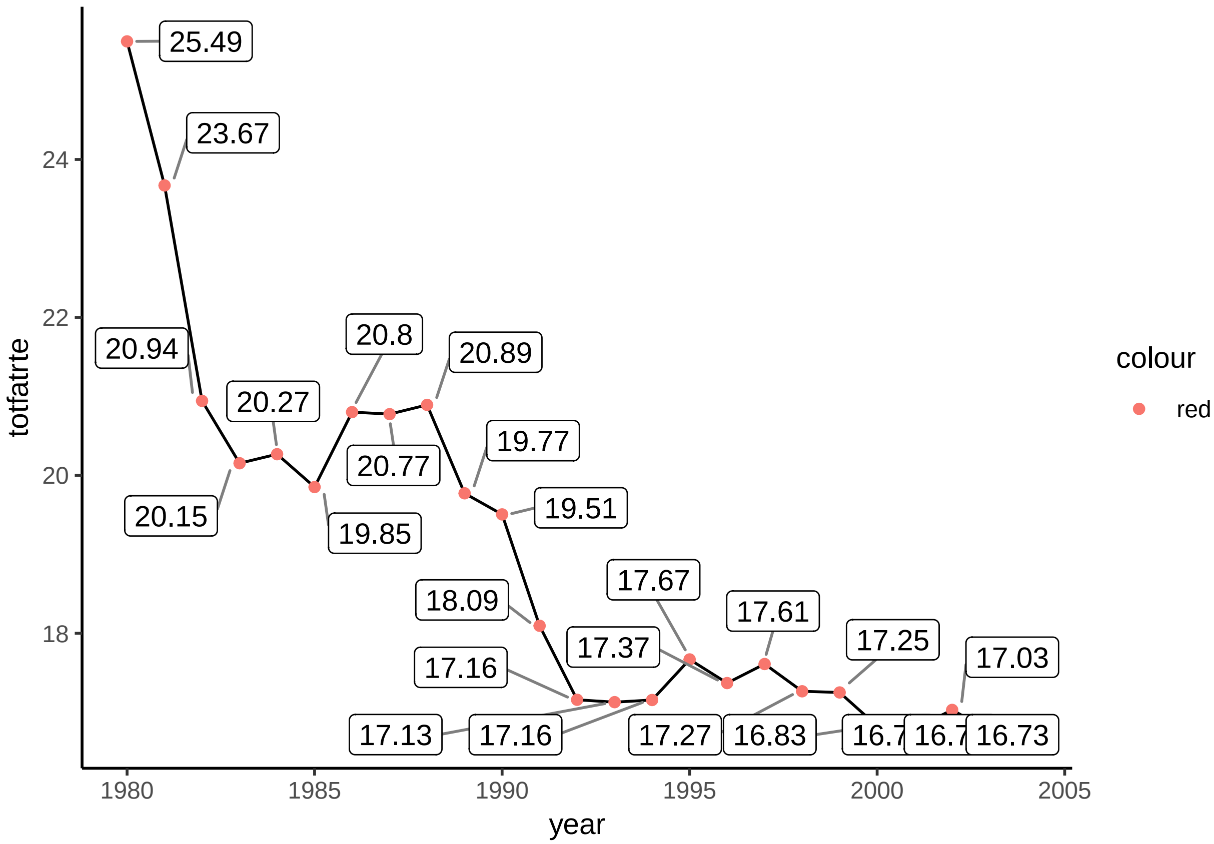
```
cor(data[,c('statepop','vehicmiles','minage','zerotol','gdl','seatbelt','vehicmilespc')])
```

```
##                 statepop  vehicmiles     minage   zerotol       gdl
## statepop      1.00000000  0.96989936 0.09719381 0.1026568 0.1124631
## vehicmiles    0.96989936  1.00000000 0.16131193 0.2065905 0.1950130
## minage        0.09719381  0.16131193 1.00000000 0.3784467 0.2019896
## zerotol       0.10265680  0.20659047 0.37844667 1.0000000 0.5178594
## gdl           0.11246313  0.19501297 0.20198956 0.5178594 1.0000000
## seatbelt      0.03990577  0.11427056 0.50901551 0.4560320 0.2265218
## vehicmilespc -0.22670325 -0.06233152 0.37605192 0.5111795 0.3178586
##                seatbelt vehicmilespc
## statepop     0.03990577  -0.22670325
## vehicmiles   0.11427056  -0.06233152
## minage       0.50901551   0.37605192
## zerotol      0.45603196   0.51117951
## gdl          0.22652184   0.31785864
## seatbelt     1.00000000   0.46796969
## vehicmilespc 0.46796969   1.00000000
```

```r
ggplot(data %>% select(year,totfatrte) %>% group_by(year) %>% summarize_all(funs(mean)),aes(year,totfat
```




```r
tmp = data %>% select(year,totfatrte) %>% group_by(year) %>% summarise_all(funs(mean))
ggplot(tmp,aes(year,totfatrte,label=totfatrte))+geom_line()+geom_point(aes(col='red'))+geom_label_repel
```



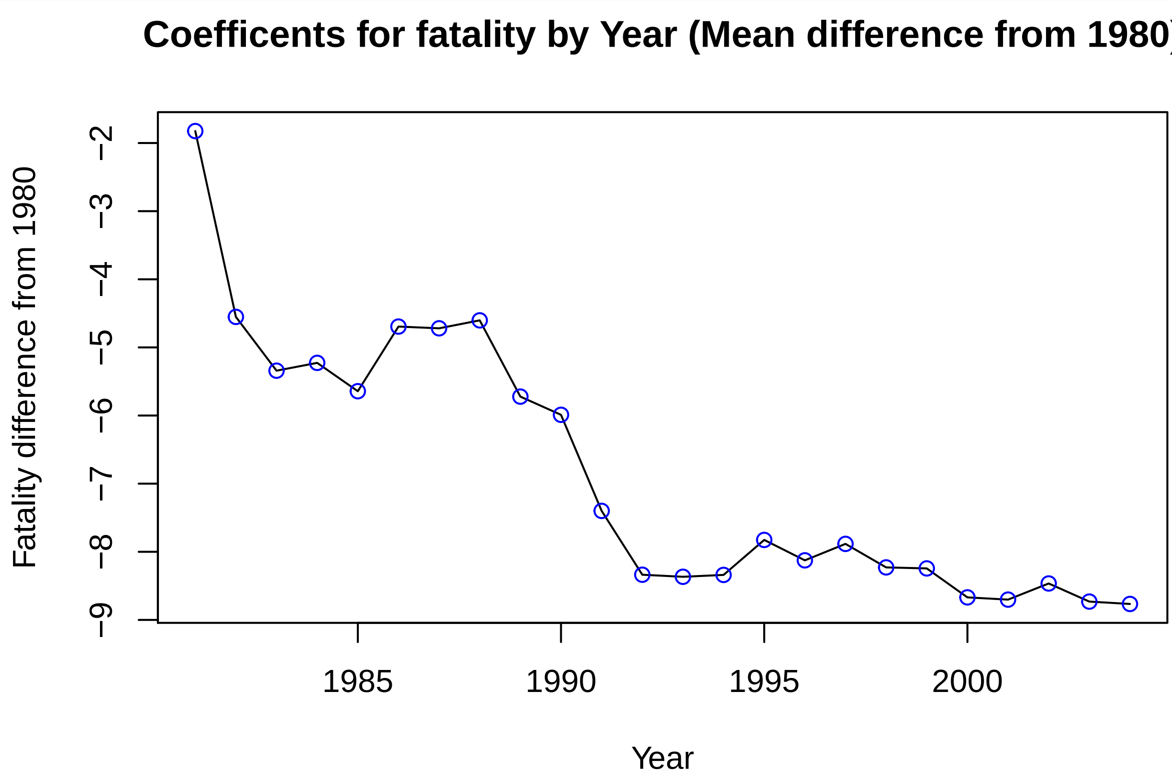

```r
m=lm(totfatrte~factor(year),data)
summary(m)
```

```
##
## Call:
## lm(formula = totfatrte ~ factor(year), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        25.4946     0.8671  29.401  < 2e-16 ***
## factor(year)1981  -1.8244     1.2263  -1.488 0.137094
## factor(year)1982  -4.5521     1.2263  -3.712 0.000215 ***
## factor(year)1983  -5.3417     1.2263  -4.356 1.44e-05 ***
## factor(year)1984  -5.2271     1.2263  -4.263 2.18e-05 ***
## factor(year)1985  -5.6431     1.2263  -4.602 4.64e-06 ***
## factor(year)1986  -4.6942     1.2263  -3.828 0.000136 ***
## factor(year)1987  -4.7198     1.2263  -3.849 0.000125 ***
## factor(year)1988  -4.6029     1.2263  -3.754 0.000183 ***
## factor(year)1989  -5.7223     1.2263  -4.666 3.42e-06 ***
## factor(year)1990  -5.9894     1.2263  -4.884 1.18e-06 ***
## factor(year)1991  -7.3998     1.2263  -6.034 2.14e-09 ***
## factor(year)1992  -8.3367     1.2263  -6.798 1.68e-11 ***
## factor(year)1993  -8.3669     1.2263  -6.823 1.43e-11 ***
```

```
## factor(year)1994  -8.3394     1.2263  -6.800 1.66e-11 ***
## factor(year)1995  -7.8260     1.2263  -6.382 2.51e-10 ***
## factor(year)1996  -8.1252     1.2263  -6.626 5.25e-11 ***
## factor(year)1997  -7.8840     1.2263  -6.429 1.86e-10 ***
## factor(year)1998  -8.2292     1.2263  -6.711 3.01e-11 ***
## factor(year)1999  -8.2442     1.2263  -6.723 2.77e-11 ***
## factor(year)2000  -8.6690     1.2263  -7.069 2.67e-12 ***
## factor(year)2001  -8.7019     1.2263  -7.096 2.21e-12 ***
## factor(year)2002  -8.4650     1.2263  -6.903 8.32e-12 ***
## factor(year)2003  -8.7310     1.2263  -7.120 1.88e-12 ***
## factor(year)2004  -8.7656     1.2263  -7.148 1.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.008 on 1175 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1098
## F-statistic: 7.164 on 24 and 1175 DF,  p-value: < 2.2e-16
```

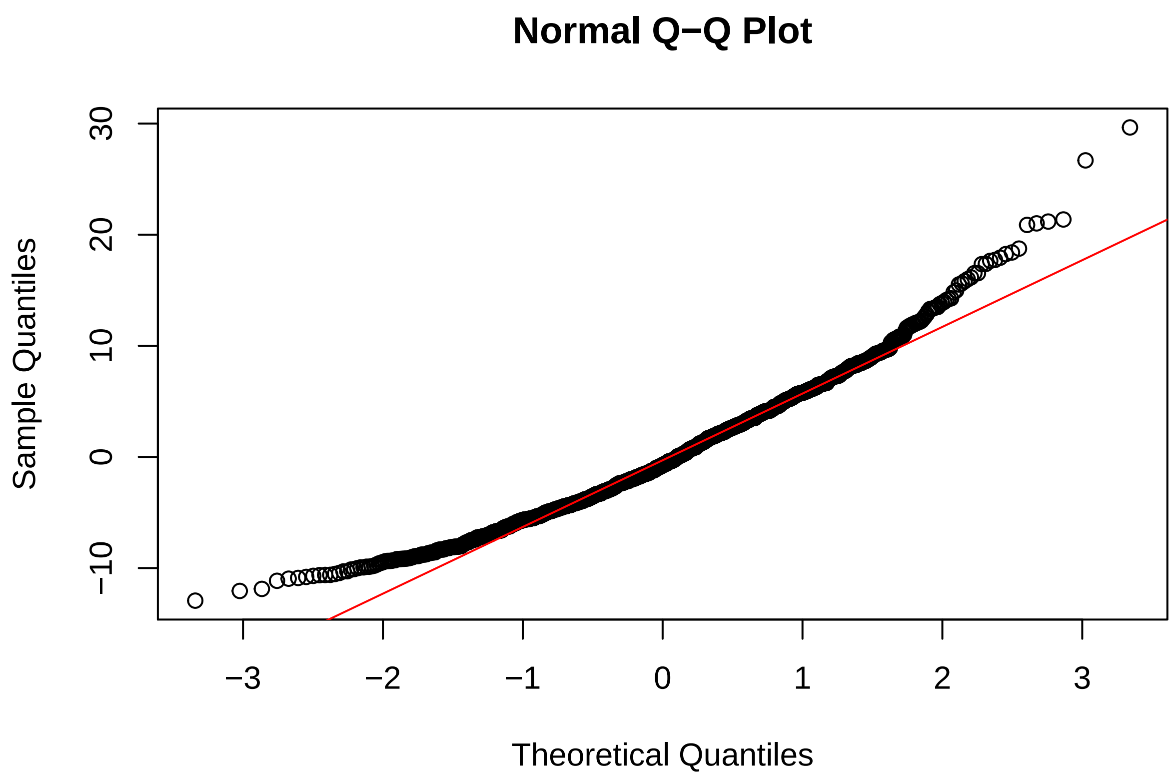
Regression of fatalities vs each year shows that there is a clear significant downward trend. The $F - test$ p-value of ~ 0 shows that the dummy variables for year is jointly significant. The regression suggests that fatalities have been decreasing through time and the βs show the mean differential between the year t and 1980. The intercept of the regression is the mean fatalities in 1980 and the coefficients is the mean differences from 1980 for each year respectively. Notice the 2 chart are exact same shape after the 1st year (1980).

```
plot(x=1981:2004,y=m$coefficients[2:length(m$coefficients)],type='l',main='Coefficents for fatality by Y
points(x=1981:2004,y=m$coefficients[2:length(m$coefficients)],col='blue')
```



Note that the residuals are not normally distributed and fails the Shapiro Wilks test.

```
qqnorm(m$residuals)
qqline(m$residuals,col='red')
```




```r
shapiro.test(m$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m$residuals
## W = 0.9703, p-value = 5.637e-15
```

We will now expand the previous regression with additional regressors - bac08, bac10, perse, sbprim, sbsecon, sl70plus, gld, perc14_24, unem and vehicmilespc. perc14-24 is logged since it is very left skewed to expand out variance between the observations and assist the regression. unem and vehicmilepc do not appear to require transformations as they appear more normally/uniformly distributed. The rest of variables are binary variables and no transformations are done. BAC and speed limit variables are not binarized and no interactions are implemented as found through the EDA due to the scope of the analysis.

```r
m=lm(totfatrte~factor(year)+bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+log(perc14_24)+unem+vehicmile
summary(m)
```

```
##
## Call:
## lm(formula = totfatrte ~ factor(year) + bac08 + bac10 + perse +
##     sbprim + sbsecon + sl70plus + gdl + log(perc14_24) + unem +
##     vehicmilespc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9146  -2.7322  -0.2732   2.2793  21.4225
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -6.258e+00  5.479e+00  -1.142 0.253641
```

```
## factor(year)1981 -2.185e+00  8.273e-01  -2.641 0.008376 **
## factor(year)1982 -6.615e+00  8.519e-01  -7.765 1.78e-14 ***
## factor(year)1983 -7.425e+00  8.655e-01  -8.579  < 2e-16 ***
## factor(year)1984 -5.887e+00  8.699e-01  -6.767 2.07e-11 ***
## factor(year)1985 -6.526e+00  8.856e-01  -7.369 3.26e-13 ***
## factor(year)1986 -5.900e+00  9.191e-01  -6.419 1.99e-10 ***
## factor(year)1987 -6.418e+00  9.525e-01  -6.738 2.52e-11 ***
## factor(year)1988 -6.645e+00  9.969e-01  -6.666 4.06e-11 ***
## factor(year)1989 -8.124e+00  1.034e+00  -7.854 9.08e-15 ***
## factor(year)1990 -9.011e+00  1.058e+00  -8.513  < 2e-16 ***
## factor(year)1991 -1.112e+01  1.083e+00 -10.264  < 2e-16 ***
## factor(year)1992 -1.293e+01  1.106e+00 -11.692  < 2e-16 ***
## factor(year)1993 -1.278e+01  1.120e+00 -11.410  < 2e-16 ***
## factor(year)1994 -1.241e+01  1.141e+00 -10.873  < 2e-16 ***
## factor(year)1995 -1.200e+01  1.169e+00 -10.264  < 2e-16 ***
## factor(year)1996 -1.392e+01  1.210e+00 -11.500  < 2e-16 ***
## factor(year)1997 -1.430e+01  1.237e+00 -11.557  < 2e-16 ***
## factor(year)1998 -1.508e+01  1.253e+00 -12.038  < 2e-16 ***
## factor(year)1999 -1.513e+01  1.271e+00 -11.901  < 2e-16 ***
## factor(year)2000 -1.549e+01  1.291e+00 -11.991  < 2e-16 ***
## factor(year)2001 -1.623e+01  1.320e+00 -12.292  < 2e-16 ***
## factor(year)2002 -1.677e+01  1.334e+00 -12.570  < 2e-16 ***
## factor(year)2003 -1.707e+01  1.345e+00 -12.689  < 2e-16 ***
## factor(year)2004 -1.676e+01  1.372e+00 -12.214  < 2e-16 ***
## bac08            -2.499e+00  5.375e-01  -4.649 3.72e-06 ***
## bac10            -1.423e+00  3.962e-01  -3.592 0.000342 ***
## perse            -6.189e-01  2.982e-01  -2.075 0.038194 *
## sbprim           -7.731e-02  4.908e-01  -0.158 0.874867
## sbsecon           6.741e-02  4.293e-01   0.157 0.875256
## sl70plus          3.344e+00  4.468e-01   7.485 1.41e-13 ***
## gdl              -4.258e-01  5.269e-01  -0.808 0.419230
## log(perc14_24)    2.125e+00  1.869e+00   1.137 0.255868
## unem              7.563e-01  7.788e-02   9.710  < 2e-16 ***
## vehicmilespc      2.923e-03  9.546e-05  30.618  < 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: 53.09 on 34 and 1165 DF,  p-value: < 2.2e-16
```

WE SHOULD PROBABLY COMMENT ON RESIDUALS AND MODEL ASSUMPTIONS HERE

POOLED OLS IS NOT VALID BCS OF SERIAL CORRELATION CAUSED BY REPEATED OBSERVATIONS

```r
library(plm)
data.p=pdata.frame(data,index = c('state','year'))
#m.pool=plm(totfatrte~bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+log(perc14_24)+unem+vehicmilespc,da
#summary(m.pool)
```

bac08 and bac10 are coefficients of -2.4987093 and -1.4230058 with p-values of 0 and 0.038 respectively. The β_{bac08} and β_{bac10} represent the impact of having abac08 and bac10 laws in that year (regardless of the year) on the mean fatalities across the states. Per se laws also decrease the mean fatalities by -0.6188569 once it's enacted. Primay seat belt laws does not seem to have an impact on fatalities despite the $\beta_{sbprim} = -0.077$, the p-value is at 0.85 indicating insignificance.

```
m.fe=plm(totfatrte~bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+log(perc14_24)+unem+vehicmilespc,data =
summary(m.fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon +
##     sl70plus + gdl + log(perc14_24) + unem + vehicmilespc, data = data.p,
##     model = "within")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -7.253710 -1.171182 -0.056489  1.108649 14.505522
##
## Coefficients:
##                   Estimate  Std. Error  t-value  Pr(>|t|)
## bac08          -1.95294344  0.38218230  -5.1100 3.774e-07 ***
## bac10          -1.56051125  0.26549425  -5.8778 5.452e-09 ***
## perse          -1.56001405  0.24612845  -6.3382 3.340e-10 ***
## sbprim         -1.80490926  0.34433211  -5.2418 1.893e-07 ***
## sbsecon        -0.86479619  0.24746941  -3.4946  0.000493 ***
## sl70plus       -1.12371942  0.24356869  -4.6136 4.405e-06 ***
## gdl            -0.59430184  0.22813383  -2.6051  0.009305 **
## log(perc14_24) 14.66024332  1.08747617  13.4810 < 2.2e-16 ***
## unem           -0.58697467  0.05086431 -11.5400 < 2.2e-16 ***
## vehicmilespc    0.00028437  0.00010221   2.7822  0.005488 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 5496.2
## R-Squared:      0.54705
## Adj. R-Squared: 0.52444
## F-statistic: 137.922 on 10 and 1142 DF, p-value: < 2.22e-16
```

```
# shoudl add p-Value or SE in here
d=data.frame(pooled=c(
  data.frame(t(m$coefficients))$bac08,
  data.frame(t(m$coefficients))$bac10,
  data.frame(t(m$coefficients))$perse,
  data.frame(t(m$coefficients))$sbprim),
  FE=c(
  data.frame(t(m.fe$coefficients))$bac08,
  data.frame(t(m.fe$coefficients))$bac10,
  data.frame(t(m.fe$coefficients))$perse,
  data.frame(t(m.fe$coefficients))$sbprim))
d
```

```
##       pooled        FE
## 1 -2.4987093 -1.952943
## 2 -1.4230058 -1.560511
## 3 -0.6188569 -1.560014
## 4 -0.0773098 -1.804909
```

COMMENT ON THE DIFFERENCES

The coefficients are significantly different between the pooled OLS and Fixed Effects regression. FE model is better since it removes the fixed effects. Pooled OLS assumes that there is no correlation between unobserved variable and any of the regressors. If this assumption is broken, *heterogeniety bias* is introduced into the model. For example, dry laws, which are unobserved, may be correlated with bac08 laws and affect fatalities. Even if the assumption is not broken, the potential serial correlation in the composite error is not accounted for in pooled OLS. The standarded errors in a pooled OLS are incorrect as are statistical tests. For the FE models, the assumption is that the idiosyncratic errors are uncorrelated conditional on the indepdendent variables and time-invariant unobservable variables. Given the current context, it the FE assumptions are more reasonable as time-invariant error can be eliminated.

```r
pooltest(totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +  gdl + log(perc14_24) + unem
```

```
##
##  F statistic
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +  ...
## F = 3.066, df1 = 470, df2 = 672, p-value < 2.2e-16
## alternative hypothesis: unstability
```

```r
pbgtest(m.fe,order=2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +     gdl + log(perc14_24) + u
## chisq = 340.02, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

In comparing FE models with RE models, FE models is likely to be a better estimate in the current context. Like the Pooled OLS, RE model assumes no correlation between fixed effects and independent variables. The difference between the 2 models is that RE corrects the serial correlation within the composite error by estimating a correlation. The advantage of RE models over FE is the ability to estimate time-invariant variables. However, it also requires an extremely strong assumption on those variables and the indepdendent variables. Given the endogeniety issues, we believe fixed effects is a much better model than random effects.

```r
summary(m.fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon +
##     sl70plus + gdl + log(perc14_24) + unem + vehicmilespc, data = data.p,
##     model = "within")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
```

```
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -7.253710 -1.171182 -0.056489  1.108649 14.505522
##
## Coefficients:
##                   Estimate  Std. Error  t-value  Pr(>|t|)
## bac08          -1.95294344  0.38218230  -5.1100 3.774e-07 ***
## bac10          -1.56051125  0.26549425  -5.8778 5.452e-09 ***
## perse          -1.56001405  0.24612845  -6.3382 3.340e-10 ***
## sbprim         -1.80490926  0.34433211  -5.2418 1.893e-07 ***
## sbsecon        -0.86479619  0.24746941  -3.4946  0.000493 ***
## sl70plus       -1.12371942  0.24356869  -4.6136 4.405e-06 ***
## gdl            -0.59430184  0.22813383  -2.6051  0.009305 **
## log(perc14_24) 14.66024332  1.08747617  13.4810 < 2.2e-16 ***
## unem           -0.58697467  0.05086431 -11.5400 < 2.2e-16 ***
## vehicmilespc    0.00028437  0.00010221   2.7822  0.005488 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 5496.2
## R-Squared:      0.54705
## Adj. R-Squared: 0.52444
## F-statistic: 137.922 on 10 and 1142 DF, p-value: < 2.22e-16
```

```r
var1 <- data.frame(t(m.fe$coefficients))$vehicmilespc
var2 <- data.frame(t(confint(m.fe)))$vehicmilespc
```

WRITE THE FE MODEL

NOT SURE HOW THEY INTERPRET increase by 1000

If $vehicmilepc$ increase by 1,000 in a time period t assuming the $\bar{vehicmilepc}$ does not change form the FE model, $totfatrte$ is expected to increase r var1 with a 95% confidence interval of r var2. NEED MORE INTERPRETATION

When autocorrelation and heteroskedasticity exists in errors, it implies that the samples are not iid. This causes your estimates to be biased. This can be seen in the $sl70plus$ variable. In the pooled model, it suggests that increase speed limit increases total fataility. The fixed model, closer to the EDA expectations, indicated that it actually decreased fatalities. With the pooled model, there is autocorrelation and heteroskedasticity in the residuals. The estimates are unstable and inconsistent as shown by the poolability test that tests whether the coefficients are across time. Finally, the estimates are inefficient as the SE may be too low or too high depending on the value of independent variable.

Exercises:

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

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

```
lm1 <- lm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94

lm1
```

```
##
## 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 = data)
##
## Coefficients:
## (Intercept)          d81          d82          d83          d84
##      25.495       -1.824       -4.552       -5.342       -5.227
##         d85          d86          d87          d88          d89
##      -5.643       -4.694       -4.720       -4.603       -5.722
##         d90          d91          d92          d93          d94
##      -5.989       -7.400       -8.337       -8.367       -8.339
##         d95          d96          d97          d98          d99
##      -7.826       -8.125       -7.884       -8.229       -8.244
##         d00          d01          d02          d03          d04
##      -8.669       -8.702       -8.465       -8.731       -8.766
```

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.)

```
lm2 <- lm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94

lm2
```

```
##
## 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 = data)
##
## Coefficients:
##  (Intercept)            d81            d82            d83            d84
##    -2.716054      -2.175479      -6.595970      -7.396690      -5.850394
##          d85            d86            d87            d88            d89
##    -6.483252      -5.852796      -6.367393      -6.591578      -8.070967
##          d90            d91            d92            d93            d94
##    -8.958670     -11.068552     -12.878398     -12.730718     -12.364833
##          d95            d96            d97            d98            d99
```

```
##   -11.952549   -13.876377   -14.258378   -15.041676   -15.090547
##          d00          d01          d02          d03          d04
##   -15.443946   -16.183715   -16.724350   -17.021308   -16.711273
##        bac08        bac10        perse       sbprim      sbsecon
##    -2.498483    -1.417565    -0.620108    -0.075335     0.067280
##     sl70plus          gdl    perc14_24         unem  vehicmilespc
##     3.347914    -0.426911     0.141590     0.757053     0.002925
```

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?

```
plm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95
```

```
##
## Model 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
##
## Coefficients:
##        bac08        bac10        perse       sbprim      sbsecon
##   -2.4984831   -1.4175652   -0.6201081   -0.0753347    0.0672804
##     sl70plus          gdl    perc14_24         unem vehicmilespc
##    3.3479143   -0.4269107    0.1415903    0.7570529    0.0029254
```

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

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

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