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

Issac Law, Mayukh Dutta, Mike King

Instructions (Please Read Carefully):

- **Due Date: Sunday 04/19/20 11:59pm**
- 20 page limit
- Do not modify fontsize, margin or line-spacing settings
- One student from each group should submit the lab to their student github repo by the deadline; submission and revisions made after the deadline will not be graded
- Answers should clearly explain your reasoning; do not simply 'output dump' the results of code without explanation
- Submit two files:

  1. A pdf file that details your answers. Include all R code used to produce the answers. Do not suppress the codes in your pdf file
  2. The R markdown (Rmd) file used to produce the pdf file

  The assignment will not be graded unless **both** files are submitted
- Name your files to include all group members names. For example the students' names are Stan Cartman and Kenny Kyle, name your files as follows:
  - `StanCartman_KennyKyle_Lab3.Rmd`
  - `StanCartman_KennyKyle_Lab3.pdf`
- Although it sounds obvious, please write your names on page 1 of your pdf and Rmd files
- For statistical methods that we cover in this course, use the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you must provide an explanation of why such libraries and functions are used and reference the library documentation. For data wrangling and data visualization, you are free to use other libraries, such as `dplyr`, `ggplot2`, etc.
- Your report needs to include:
  - A thorough analysis of the given dataset, which includ examiniation of anomalies, missing values, potential of top and/or bottom code, and other potential anomalies, in each of the variables.

  - A comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score. Be selective when choosing visuals and tables to illustrate your key points and concise with your explanations (please do not ramble).
  - A proper narrative for each question answered. Make sure that your audience can easily follow the logic of your analysis and the rationale of decisions made in your modeling, supported by empirical evidence. Use the insights generated from your EDA step to guide your modeling approach.
  - Clear explanations of all steps used to arrive at a final model, with conclusions that summarize results with respect to the question(s) being asked and key takeaways from the analysis.
- For mathematical formulae, type them in your R markdown file. Do not e.g. write them on a piece of paper, snap a photo, and use the image file.
- Incorrectly following submission instructions results in deduction of grades
- Students are expected to act with regard to UC Berkeley Academic Integrity

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.

```r
load("driving.Rdata")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
```

Exercises:

1. (40%) Load the data. Provide a description of the basic structure of the dataset, as we have done throughout the semester. Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable *totfatrte* and the potential explanatory variables. You need to write a detailed narrative of your observations of your EDA. *Reminder: giving an "output dump" (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive a zero in this exercise.* We have one row per year per state. No data points are missing - we have an observation for each of the lower 48 states for each year from 1980-2004. The data types are mostly appropriate, although the seatbelt and speed limit laws could be thought of as factors.

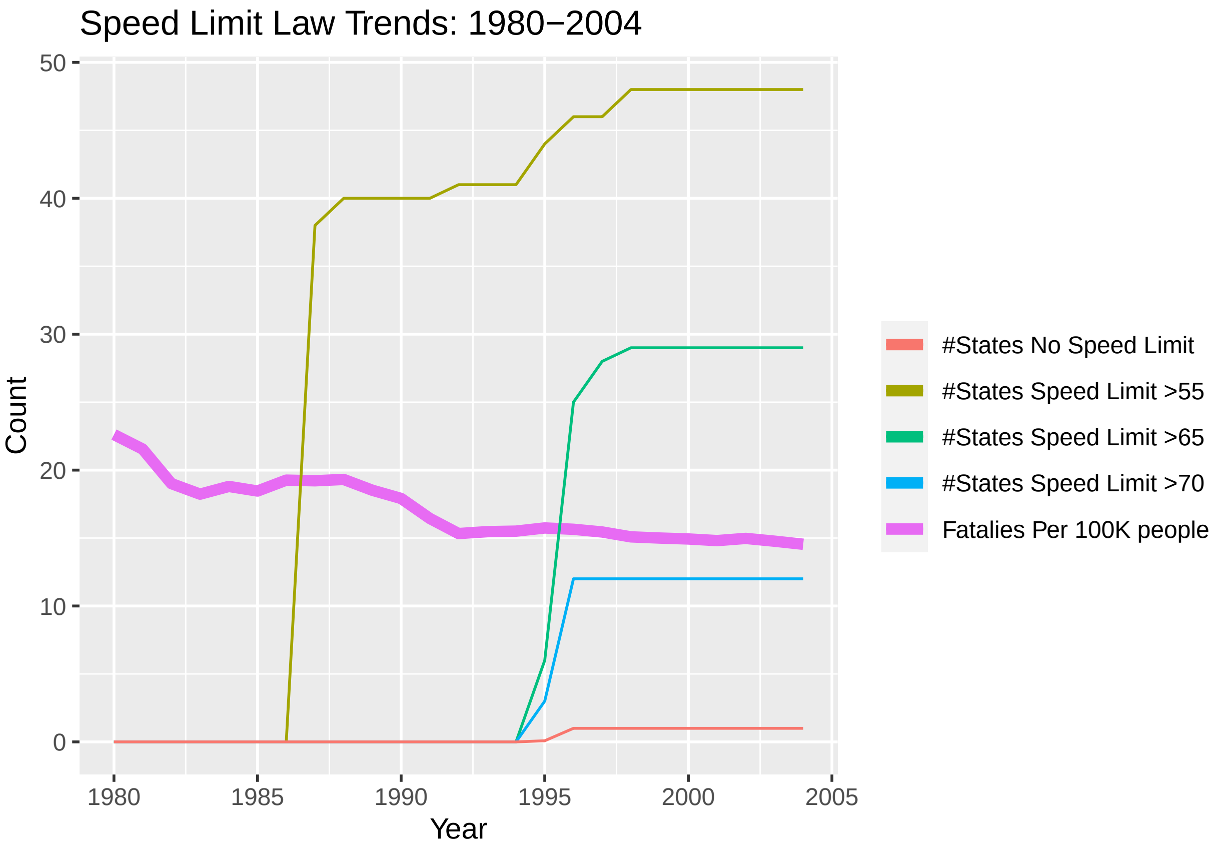
We must be careful when thinking about our primary outcome variable as explained by these explanatory variables. Many other changes occurred over this time that could be explanatory for fatality rates. Automotive technology changed (airbags, anti-lock brakes, influx of Japanese-built cars), national infrastructure changed (more multi-lane highways, nighttime lighting), and many other technology and cultural changes occurred as well (cell phones, urbanization, life expectancy.) There are potentially many unobserved explanatory variables.

Now aggregate the data. Combine the separate values by state into appropriate sums/averages for the nation. The intention here is to see how states have changed their laws over time.

```
aggregatedData = data %>%
mutate(
  sbany = as.integer(seatbelt > 0),
  slGt55 = as.integer((sl65 + sl70 +  sl75  + slnone) > 0),
  slGt65 = as.integer((sl70 +  sl75  + slnone) > 0),
  slGt70 = as.integer((sl75  + slnone) > 0),
  minageGt18 = as.integer(minage > 18),
  minageGt19 = as.integer(minage > 19),
  minageGt20 = as.integer(minage > 20)
  ) %>%
group_by(year) %>%
summarise(
  totfat = sum(totfat),
  population = sum(statepop),
  slGt55 = sum(slGt55),
  slGt65 = sum(slGt65),
  slGt70 = sum(slGt70),
  slnone = sum(slnone),
  minageGt18 = sum(minageGt18),
  minageGt19 = sum(minageGt19),
  minageGt20 = sum(minageGt20),
  bac08 = sum(bac08),
  gdl = sum(gdl),
  perse = sum(perse),
  zerotol = sum(zerotol),
  sbprim = sum(sbprim),
  sbsecon = sum(sbsecon),
  sbany = sum(sbany)
) %>%
mutate(
  totfatrte = 100000 * totfat / population
)
```

First look at how speed limit laws have changed:

```
aggregatedData %>% ggplot(aes(year)) +
geom_line(aes(y = totfatrte, col = "Fatalies Per 100K people"), size = 2) +
geom_line(aes(y = slGt55, col = "#States Speed Limit >55")) +
geom_line(aes(y = slGt65, col = "#States Speed Limit >65")) +
geom_line(aes(y = slGt70, col = "#States Speed Limit >70")) +
geom_line(aes(y = slnone, col = "#States No Speed Limit")) +
ggtitle("Speed Limit Law Trends: 1980-2004") + ylab("Count") + xlab("Year") + labs(col = "")
```



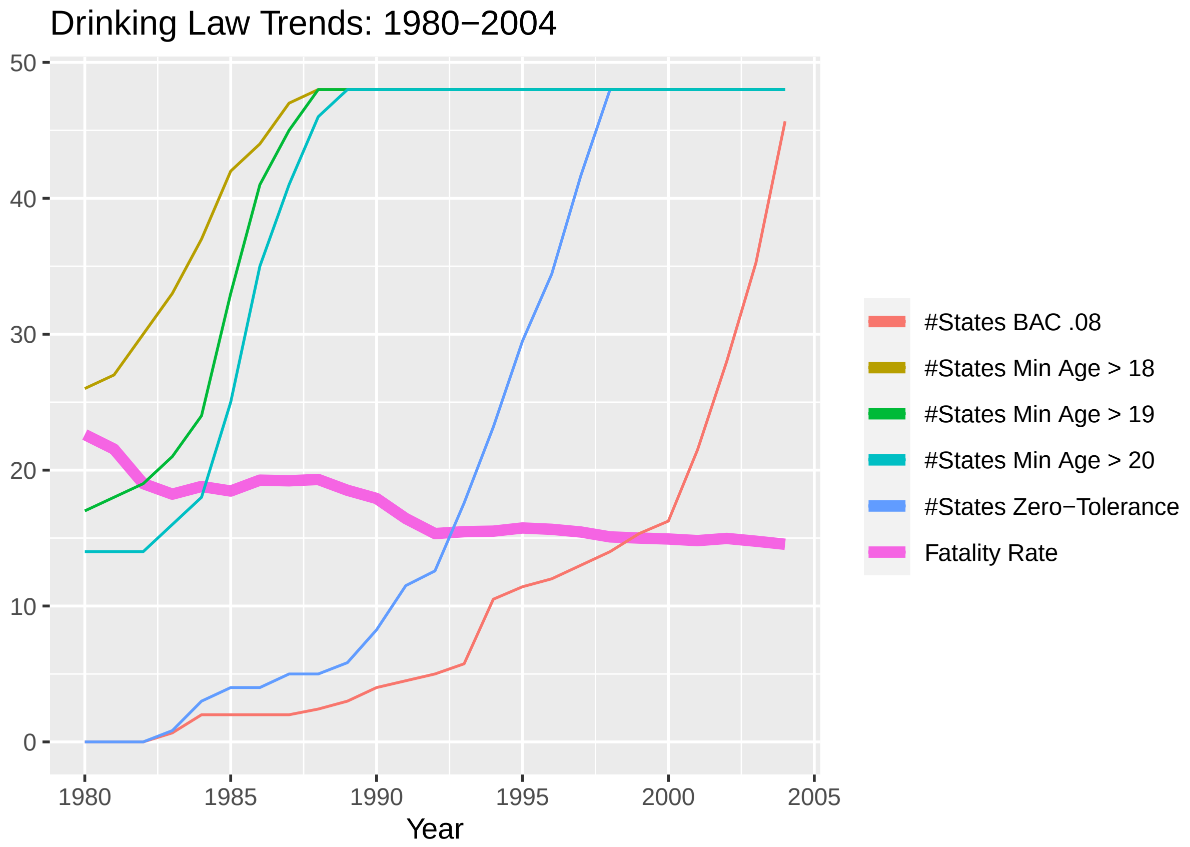
The first thing to notice about our primary outcome variable is that there are 2 periods over which the fatality rates changed more than any other time period in the data set. The first is 1980 through 1983 and the second is 1988 through 1992. The other timespans (1983-1988 and 1992 through 2004) show nearly stationary fatality rates.

We see that over 1987 through 1996 states increased their speed limits in an always-increasing pattern (for no year did the number states with a minimum speed limit decrease from the prior year.) Higher speed limits do seem to correlate with lower fatality rates although not particularly strongly. The drop in fatality rate from 1980-1983 occurred with zero changes in speed limit laws.

The speed limit laws are different from the other laws in this data set because the speed limit laws have been relaxed over time whereas the drinking and seatbelt laws have become more strict. One might expect that relaxed speed limit laws could lead to an increase in fatality rates. This data is observational (not experimental) however no evidenciary support for such a claim about speed limits is apparent.

A look at how drinking laws have changed over this time:

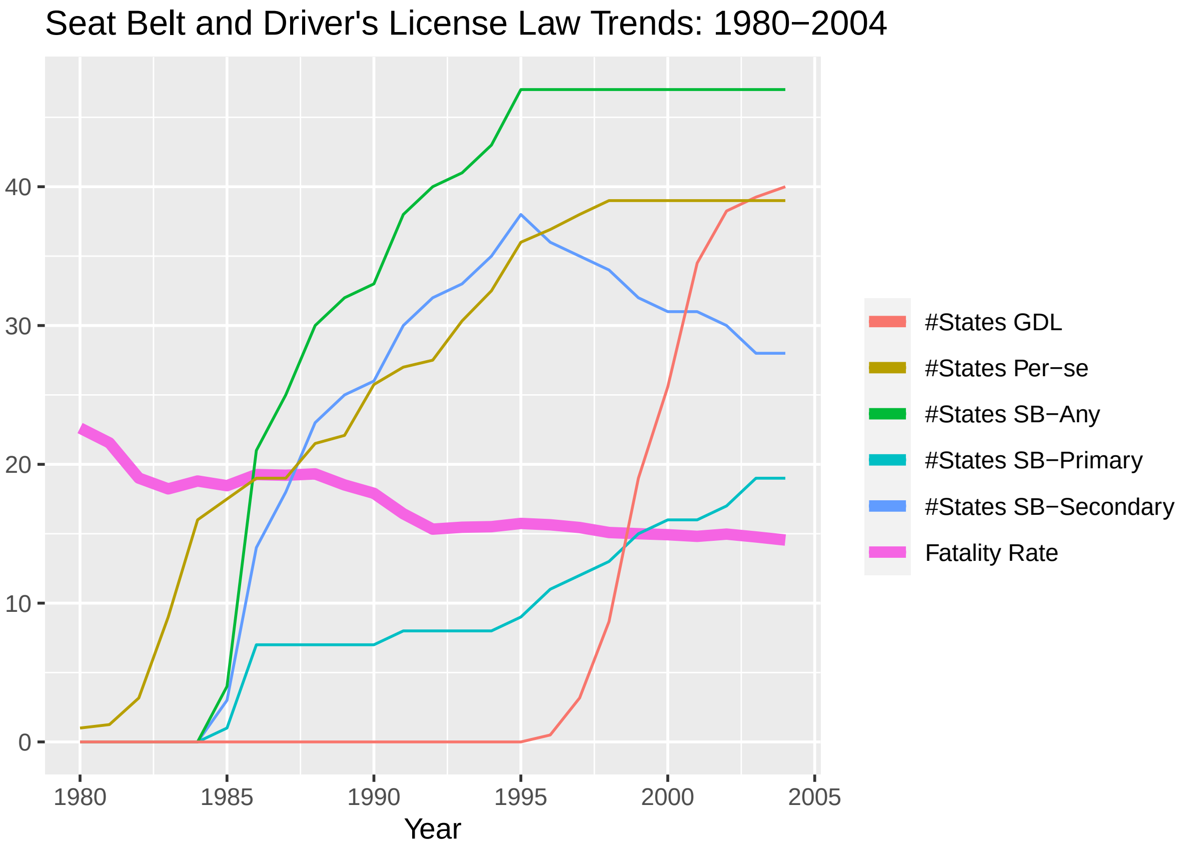
```
aggregatedData %>% ggplot(aes(year)) +
geom_line(aes(y = totfatrte, col = "Fatality Rate"), size = 2) +
geom_line(aes(y = bac08, col = "#States BAC .08")) +
geom_line(aes(y = zerotol, col = "#States Zero-Tolerance"))  +
geom_line(aes(y = minageGt18, col = "#States Min Age > 18"))  +
geom_line(aes(y = minageGt19, col = "#States Min Age > 19"))  +
geom_line(aes(y = minageGt20, col = "#States Min Age > 20"))  +
ggtitle("Drinking Law Trends: 1980-2004") + ylab("") + xlab("Year") + labs(col = "")
```



The drop in fatality rates from 1980-1983 correlates with the increased minimum drinking age in several states. The drop in fatality rates from 1988-1992 coincide with the increase in zero-tolerance states and the increase in BAC08 states. As with the speed limit laws, there are no years with reversions to prior values. The minimum drinking age either increases or remains the same, and the same for zero-tolerance and BAC08.

Now a look at seatbelt and driver's license law changes:

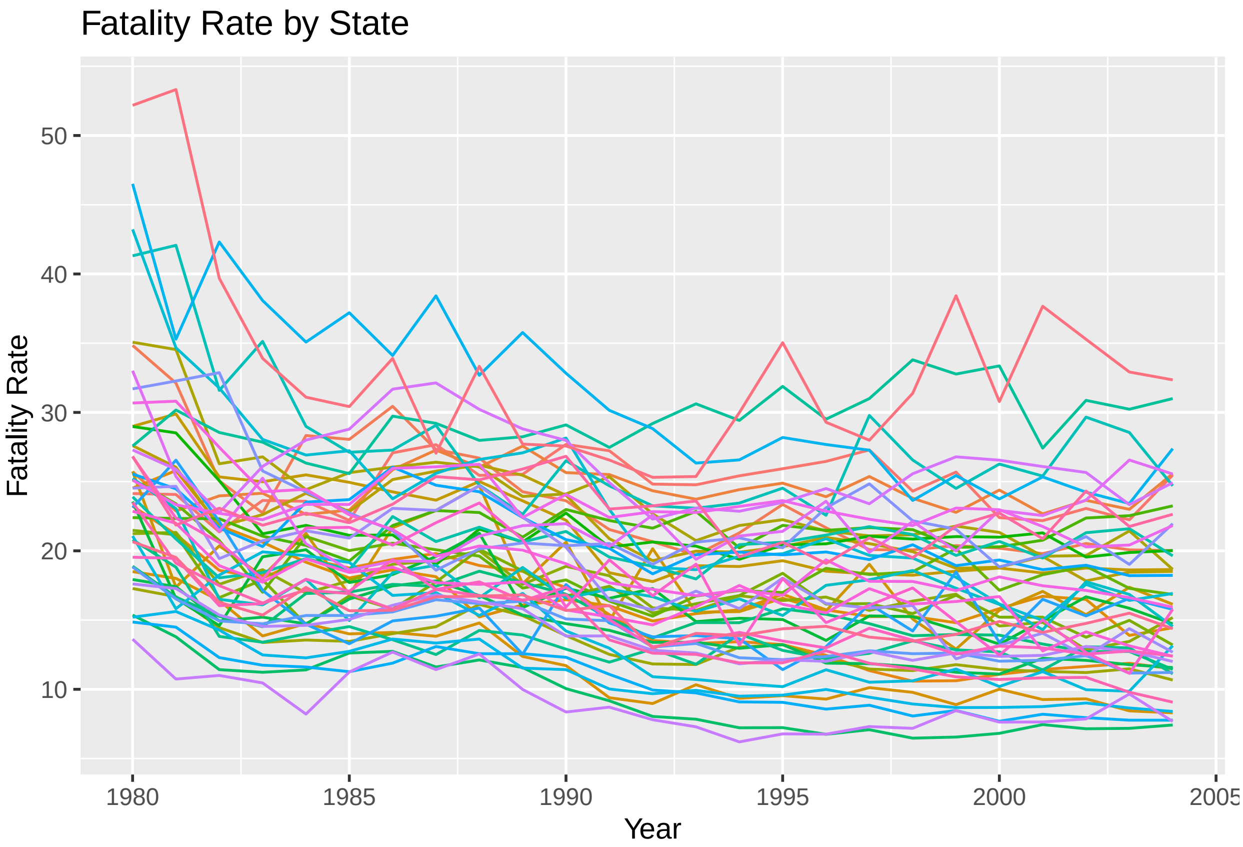
```
aggregatedData %>% ggplot(aes(year)) +
geom_line(aes(y = totfatrte, col = "Fatality Rate"), size = 2) +
geom_line(aes(y = sbsecon, col = "#States SB-Secondary")) +
geom_line(aes(y = sbprim, col = "#States SB-Primary"))  +
geom_line(aes(y = sbany, col = "#States SB-Any"))  +
geom_line(aes(y = gdl, col = "#States GDL"))  +
geom_line(aes(y = perse, col = "#States Per-se"))  +
ggtitle("Seat Belt and Driver's License Law Trends: 1980-2004") + ylab("") + xlab("Year") + la
```



We see that seatbelt laws have become more strict (or remained the same) each year. The downward trend in seatbelt-secondary is explained by the increase in seatbelt-primary and the contant seatbelt-any. Clearly the states abadoning their seatbelt-secondary laws were simply switching to the more strict seatbelt-primary laws. No obvious correlation is seen with the 1980-1983 drop in fatality rate, but the 1988-1992 drop occurs over a time when seatbelt laws and per-se laws were increasing quickly. The graduated license law does not seem correlated with any meaningful drop in fatality rate, especially compared to seatbelt-primary over the time when the GDL laws were adopted.

Now a look at the data by state:

```
ggplot(data, aes(year, totfatrte)) + geom_line(aes(col = as.factor(state))) + ggtitle("Fatality
```



```
#varData = data %>% group_by(state) %>% summarise(Variance = var(totfatrte))
#plot(varData$state, varData$Variance)
```

The downward trend can be seen, although the image is very busy. There are 4 states that had much higher fatality rates in 1980 than the others, but none would qualify as an outlier. One state appears to have the lowest fatality rate for many of the years in the timespan.

So far the most strongly correlated explanatory variables seem to be a drinking age > 18, a BAC08 law, any seatbelt law, and a zero-tolerance law. Changes to speed limit laws seem to have no correlation with changes in fatality rates.

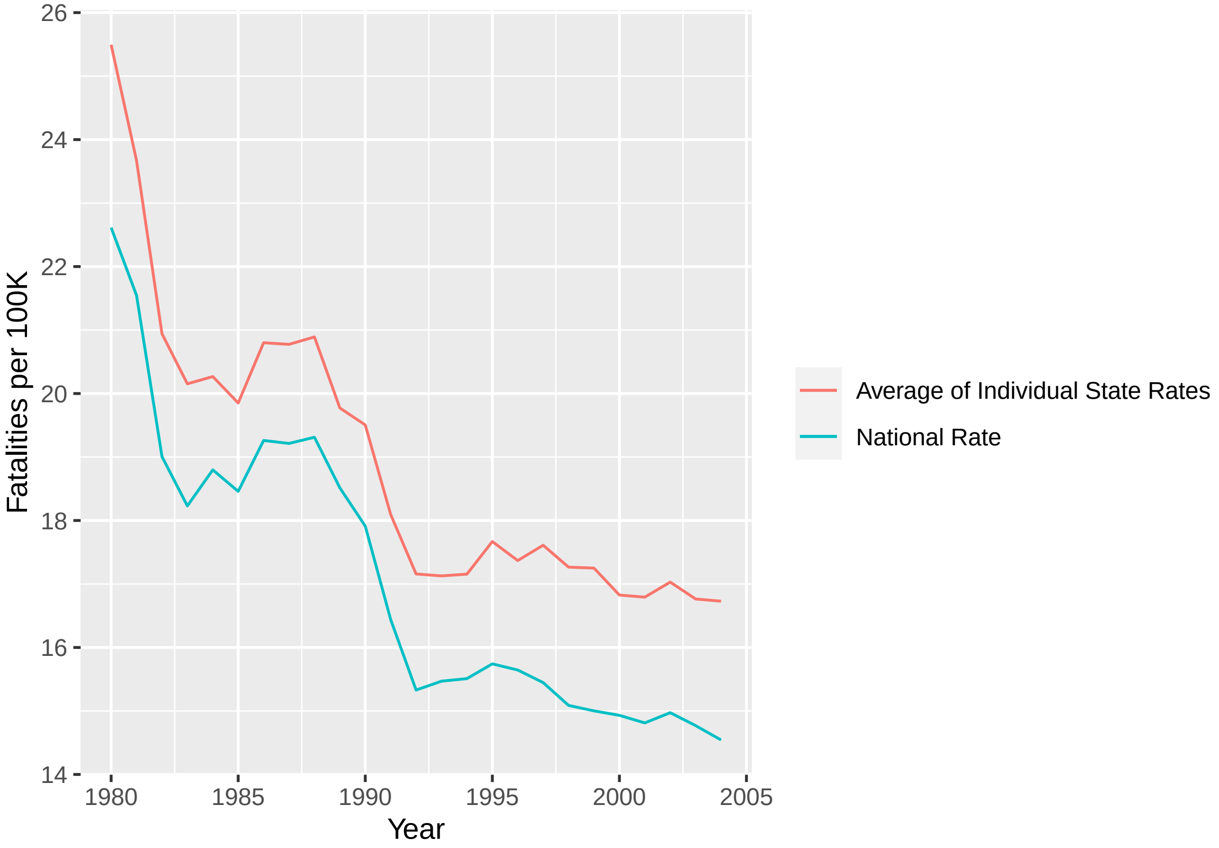
2. (15%) How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this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

The definition is the number of fatalities per 100K of the state's population. However it is per state per year, and not every state has the same population. So the national average per year is different from the average of the 48 individual states.

The national mean of the value can be computed by aggregating the state level data. The average-of-state-level can be computed by averaging totfatrte in the original data set.

```
nationalData = data %>%
group_by(year) %>%
summarise(
  totfat = sum(totfat),
  population = sum(statepop),
) %>%
mutate(
  totfatrte = 100000 * totfat / population
)

stateRateAverage = data %>% group_by(year) %>% summarise(totfatrte = mean(totfatrte))
ggplot() + geom_line(data = nationalData, aes(x = year, y = totfatrte, col = "National Rate"))
```



A model of the original data with dummy variables for the years:

```
mod2 = lm(totfatrte ~ as.factor(year), data = data)
summary(mod2)
```

```
##
## Call:
## lm(formula = totfatrte ~ as.factor(year), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           25.4946     0.8671  29.401  < 2e-16 ***
## as.factor(year)1981   -1.8244     1.2263  -1.488 0.137094    
## as.factor(year)1982   -4.5521     1.2263  -3.712 0.000215 ***
## as.factor(year)1983   -5.3417     1.2263  -4.356 1.44e-05 ***
## as.factor(year)1984   -5.2271     1.2263  -4.263 2.18e-05 ***
## as.factor(year)1985   -5.6431     1.2263  -4.602 4.64e-06 ***
## as.factor(year)1986   -4.6942     1.2263  -3.828 0.000136 ***
## as.factor(year)1987   -4.7198     1.2263  -3.849 0.000125 ***
## as.factor(year)1988   -4.6029     1.2263  -3.754 0.000183 ***
## as.factor(year)1989   -5.7223     1.2263  -4.666 3.42e-06 ***
## as.factor(year)1990   -5.9894     1.2263  -4.884 1.18e-06 ***
## as.factor(year)1991   -7.3998     1.2263  -6.034 2.14e-09 ***
## as.factor(year)1992   -8.3367     1.2263  -6.798 1.68e-11 ***
## as.factor(year)1993   -8.3669     1.2263  -6.823 1.43e-11 ***
## as.factor(year)1994   -8.3394     1.2263  -6.800 1.66e-11 ***
## as.factor(year)1995   -7.8260     1.2263  -6.382 2.51e-10 ***
## as.factor(year)1996   -8.1252     1.2263  -6.626 5.25e-11 ***
## as.factor(year)1997   -7.8840     1.2263  -6.429 1.86e-10 ***
## as.factor(year)1998   -8.2292     1.2263  -6.711 3.01e-11 ***
## as.factor(year)1999   -8.2442     1.2263  -6.723 2.77e-11 ***
## as.factor(year)2000   -8.6690     1.2263  -7.069 2.67e-12 ***
## as.factor(year)2001   -8.7019     1.2263  -7.096 2.21e-12 ***
## as.factor(year)2002   -8.4650     1.2263  -6.903 8.32e-12 ***
## as.factor(year)2003   -8.7310     1.2263  -7.120 1.88e-12 ***
## as.factor(year)2004   -8.7656     1.2263  -7.148 1.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.008 on 1175 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1098 
## F-statistic: 7.164 on 24 and 1175 DF,  p-value: < 2.2e-16
```

This model is just offering a coefficient for each year. Of course the coefficient is equal to the difference in the fatality rate for that year (averaged by state) from the fatality rate in 1980 (averaged by state). So this model provides no additional insight from the red line on the previous graph. They are all significant because they contain all the explanatory power needed to get the average for that year vs the average in 1980 - ALL of the necessary power. They have exactly the same values:

```
round(stateRateAverage$totfatrte - 25.49458, 4)[2:25]   #29.49458 is the base fatality rate fro
```

```
##  [1] -1.8244 -4.5521 -5.3417 -5.2271 -5.6431 -4.6942 -4.7198 -4.6029 -5.7223
## [10] -5.9894 -7.3998 -8.3367 -8.3669 -8.3394 -7.8260 -8.1252 -7.8840 -8.2292
## [19] -8.2442 -8.6690 -8.7019 -8.4650 -8.7310 -8.7656
```

```
round(as.vector(mod2$coefficients), 4)[2:25]                #coefficient#1
```

```
##  [1] -1.8244 -4.5521 -5.3417 -5.2271 -5.6431 -4.6942 -4.7198 -4.6029 -5.7223
## [10] -5.9894 -7.3998 -8.3367 -8.3669 -8.3394 -7.8260 -8.1252 -7.8840 -8.2292
## [19] -8.2442 -8.6690 -8.7019 -8.4650 -8.7310 -8.7656
```

Driving did become safer over this time period if you measure safety by fatalities per capita, giving equal weight to each state or looking at the national average as a whole. However putting a coefficient on each independent year offers no insight as to why (unless you accept the possibility that the calendar year literally changes fatality rates holding all other things constant, a preposterous conjecture.)

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.)

```
mod3 = lm(totfatrte ~ as.factor(year) + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + g
summary(mod3)
```

```
##
## Call:
## lm(formula = totfatrte ~ as.factor(year) + bac08 + bac10 + perse +
##     sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem + vehicmilespc,
##     data = data)
##
## 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
## as.factor(year)1981 -2.175e+00  8.276e-01  -2.629 0.008686 **
## as.factor(year)1982 -6.596e+00  8.534e-01  -7.729 2.33e-14 ***
## as.factor(year)1983 -7.397e+00  8.690e-01  -8.512  < 2e-16 ***
## as.factor(year)1984 -5.850e+00  8.763e-01  -6.676 3.79e-11 ***
## as.factor(year)1985 -6.483e+00  8.948e-01  -7.245 7.82e-13 ***
## as.factor(year)1986 -5.853e+00  9.307e-01  -6.289 4.52e-10 ***
## as.factor(year)1987 -6.367e+00  9.670e-01  -6.585 6.87e-11 ***
## as.factor(year)1988 -6.592e+00  1.014e+00  -6.502 1.17e-10 ***
## as.factor(year)1989 -8.071e+00  1.053e+00  -7.667 3.68e-14 ***
## as.factor(year)1990 -8.959e+00  1.077e+00  -8.319 2.46e-16 ***
## as.factor(year)1991 -1.107e+01  1.101e+00 -10.052  < 2e-16 ***
## as.factor(year)1992 -1.288e+01  1.123e+00 -11.473  < 2e-16 ***
## as.factor(year)1993 -1.273e+01  1.136e+00 -11.204  < 2e-16 ***
## as.factor(year)1994 -1.236e+01  1.157e+00 -10.685  < 2e-16 ***
## as.factor(year)1995 -1.195e+01  1.184e+00 -10.098  < 2e-16 ***
## as.factor(year)1996 -1.388e+01  1.223e+00 -11.343  < 2e-16 ***
## as.factor(year)1997 -1.426e+01  1.250e+00 -11.408  < 2e-16 ***
## as.factor(year)1998 -1.504e+01  1.265e+00 -11.886  < 2e-16 ***
## as.factor(year)1999 -1.509e+01  1.284e+00 -11.750  < 2e-16 ***
## as.factor(year)2000 -1.544e+01  1.305e+00 -11.831  < 2e-16 ***
## as.factor(year)2001 -1.618e+01  1.334e+00 -12.131  < 2e-16 ***
## as.factor(year)2002 -1.672e+01  1.348e+00 -12.406  < 2e-16 ***
## as.factor(year)2003 -1.702e+01  1.359e+00 -12.521  < 2e-16 ***
## as.factor(year)2004 -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
```

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?

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

6. (5%) Suppose that *vehicmilespc*, the number of miles driven per capita, increases by $1,000$. Using the FE estimates, what is the estimated effect on *totfatrte*? Please interpret the estimate.

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?