

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

Due Sunday 8 August 2021 11:59pm

Dominik Graf, Mikayla Pugel, Prasad Valavade

Instructions (Please Read Carefully):

- Submit by the due date. **Late submissions will not be accepted**
- 20 page limit (strict)
- 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
- Submit two files:
 1. A pdf file that details your answers. Include all R code used to produce the answers
 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, if 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 name on page 1 of your pdf and Rmd files
- All answers should include a detailed narrative; make sure that your audience can easily follow the logic of your analysis. All steps used in modelling must be clearly shown and explained; do not simply 'output dump' the results of code without explanation
- If you use libraries and functions for statistical modeling that we have not covered in this course, you must provide an explanation of why such libraries and functions are used and reference the library documentation
- 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 economic and demographic variables are also included. The description of the each of the variables in the dataset is come with the dataset.

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

Response to Q1: The EDA below displays many different graphs and tables of the data. The first table is the correlation plot of all the variables. The dependent variable of interest is the *totfatrte*, which is the total fatality rate of driving incidents. From the correlation plot, the variables that seem to best correlate with this variable are, the year, seatbelt, minage, gdl, zerotol, vehiclemiles. All these variables are negatively correlated with the dependent variable. There are some variables that are strongly correlated with the dependent variable in a positive way and these include unem, perc14_24, and vehicmilespc. Since some of the speedlimit correlated go from being positive and negative, it seems that the data is inconclusive here on how changing the speed limit may affect driving incidents. Also, it is important to note that all the other measurements of fatality are highly correlated with the dependent variable, however these variables are not independent of one another and are essentially part of the same thing, therefore we will not use them in the model.

It is also important to note that there is no missing data from the data, and the data goes from 1980 through 2004 and includes 48 different states. The data is set up to have all the data in the time frame of 1980 to 2004 for each state, therefore we can see the changes over time for each individual state and model all the changes together in one model. Only the variables from the data that were discussed above for having a strong correlation to the dependent variable will be analyzed from here on out.

Since the year was negatively correlated I was interested to model the change in *totfatrte* over time. All the states were modeled on plots, however for sake of conducting a concise analysis only one plot of 12 different states is shown below. This plot shows the different states over time. We are not necessarily interested on how an individual model’s data looks, but rather the overall change, and overall the data looks like it does slightly decrease over the 25 year span. The other data plots that are not shown showed similar plots.

Lastly, the histograms of all the variable in consideration for the model are shown below. It is difficult to assess what some of the histograms mean, since we do not have information on what

some of these variable acronym's meaning. However, the point of plotting the histograms of these variables is to see any extreme distribution in the data. Most of the data is generally normal, mainly a lot of the data has right sided tails which may lead to a transformation of that variable in order to obtain a more normal distribution, however these transformations will be discussed in the model formulation part of the lab.

```
library(plm)
library(funModeling)
library(tidyverse)
library(Hmisc)
library(ggplot2)
library(forecast)
library(tseries)
library(corr)

```

```
#load("driving.RData", ex <- new.env())
#ls.str(ex)
load("driving.RData")
head(data,5)

```

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

```

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

```
df <- pdata.frame(data, index=c("state", 'year'))
#head(df, 20)
```

```
correlate(data)
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'

## # A tibble: 56 x 57
##   term      year      state  sl55  sl65  sl70  sl75  slnone seatbelt
##   <chr>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1 year      NA      -4.46e-22 -0.778 0.269 0.427 0.336 9.67e-2 0.649
## 2 state -4.46e-22 NA      -0.0192 0.0375 -0.0746 0.0548 -8.89e-4 0.0264
## 3 sl55 -7.78e- 1 -1.92e- 2 NA      -0.649 -0.280 -0.224 -6.60e-2 -0.639
## 4 sl65 2.69e- 1 3.75e- 2 -0.649 NA      -0.321 -0.261 -7.82e-2 0.339
## 5 sl70 4.27e- 1 -7.46e- 2 -0.280 -0.321 NA      -0.109 -3.27e-2 0.203
## 6 sl75 3.36e- 1 5.48e- 2 -0.224 -0.261 -0.109 NA      -2.62e-2 0.236
## 7 slnone 9.67e- 2 -8.89e- 4 -0.0660 -0.0782 -0.0327 -0.0262 NA      0.0863
## 8 seatbe~ 6.49e- 1 2.64e- 2 -0.639 0.339 0.203 0.236 8.63e-2 NA
## 9 minage 5.71e- 1 -1.66e- 2 -0.579 0.376 0.162 0.129 3.81e-2 0.509
## 10 zerotol 7.89e- 1 2.76e- 2 -0.570 0.124 0.400 0.279 5.74e-2 0.456
## # ... with 46 more rows, and 48 more variables: minage <dbl>, zerotol <dbl>,
## #   gdl <dbl>, bac10 <dbl>, bac08 <dbl>, perse <dbl>, totfat <dbl>,
## #   nghtfat <dbl>, wkndfat <dbl>, totfatpvm <dbl>, nghtfatpvm <dbl>,
## #   wkndfatpvm <dbl>, statepop <dbl>, totfatrte <dbl>, nghtfatrte <dbl>,
## #   wkndfatrte <dbl>, vehicmiles <dbl>, unem <dbl>, perc14_24 <dbl>,
## #   sl70plus <dbl>, sbprim <dbl>, sbsecon <dbl>, d80 <dbl>, d81 <dbl>,
## #   d82 <dbl>, d83 <dbl>, d84 <dbl>, d85 <dbl>, d86 <dbl>, d87 <dbl>,
## #   d88 <dbl>, d89 <dbl>, d90 <dbl>, d91 <dbl>, d92 <dbl>, d93 <dbl>,
## #   d94 <dbl>, d95 <dbl>, d96 <dbl>, d97 <dbl>, d98 <dbl>, d99 <dbl>,
## #   d00 <dbl>, d01 <dbl>, d02 <dbl>, d03 <dbl>, d04 <dbl>, vehicmilesperc <dbl>
```

```
sum(is.na(data))
```

```
## [1] 0
```

```
summary(data)
```

```
##      year      state      sl55      sl65
## Min.   :1980   Min.    : 1.00   Min.    :0.0000   Min.    :0.0000
```

##	1st Qu.:1986	1st Qu.:15.75	1st Qu.:0.0000	1st Qu.:0.0000	
##	Median :1992	Median :27.50	Median :0.0000	Median :0.0000	
##	Mean :1992	Mean :27.15	Mean :0.3533	Mean :0.4399	
##	3rd Qu.:1998	3rd Qu.:39.25	3rd Qu.:1.0000	3rd Qu.:1.0000	
##	Max. :2004	Max. :51.00	Max. :1.0000	Max. :1.0000	
##	sl70	sl75	slnone	seatbelt	
##	Min. :0.000	Min. :0.00000	Min. :0.000000	Min. :0.000	
##	1st Qu.:0.000	1st Qu.:0.00000	1st Qu.:0.000000	1st Qu.:0.000	
##	Median :0.000	Median :0.00000	Median :0.000000	Median :1.000	
##	Mean :0.119	Mean :0.08024	Mean :0.007569	Mean :1.116	
##	3rd Qu.:0.000	3rd Qu.:0.00000	3rd Qu.:0.000000	3rd Qu.:2.000	
##	Max. :1.000	Max. :1.00000	Max. :1.000000	Max. :2.000	
##	minage	zerotol	gdl	bac10	
##	Min. :18.0	Min. :0.0000	Min. :0.0000	Min. :0.0000	
##	1st Qu.:21.0	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	
##	Median :21.0	Median :0.0000	Median :0.0000	Median :1.0000	
##	Mean :20.6	Mean :0.4519	Mean :0.1741	Mean :0.6231	
##	3rd Qu.:21.0	3rd Qu.:1.0000	3rd Qu.:0.0000	3rd Qu.:1.0000	
##	Max. :21.0	Max. :1.0000	Max. :1.0000	Max. :1.0000	
##	bac08	perse	totfat	nghtfat	
##	Min. :0.0000	Min. :0.0000	Min. : 63.0	Min. : 26.0	
##	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.: 310.0	1st Qu.: 139.8	
##	Median :0.0000	Median :1.0000	Median : 676.0	Median : 316.0	
##	Mean :0.2135	Mean :0.5471	Mean : 900.7	Mean : 427.3	
##	3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:1099.5	3rd Qu.: 518.2	
##	Max. :1.0000	Max. :1.0000	Max. :5504.0	Max. :2918.0	
##	wkndfat	totfatpvm	nghtfatpvm	wkndfatpvm	
##	Min. : 10.0	Min. :0.780	Min. :0.2700	Min. :0.1140	
##	1st Qu.: 70.0	1st Qu.:1.577	1st Qu.:0.6847	1st Qu.:0.3410	
##	Median : 163.0	Median :2.020	Median :0.9130	Median :0.4770	
##	Mean : 222.3	Mean :2.122	Mean :0.9990	Mean :0.5255	
##	3rd Qu.: 277.0	3rd Qu.:2.500	3rd Qu.:1.2110	3rd Qu.:0.6420	
##	Max. :1499.0	Max. :5.700	Max. :3.0030	Max. :1.6750	
##	statepop	totfatrte	nghtfatrte	wkndfatrte	
##	Min. : 453401	Min. : 6.20	Min. : 2.660	Min. : 1.180	
##	1st Qu.: 1641938	1st Qu.:14.38	1st Qu.: 6.338	1st Qu.: 3.240	
##	Median : 3700425	Median :18.43	Median : 8.420	Median : 4.390	
##	Mean : 5329896	Mean :18.92	Mean : 8.796	Mean : 4.606	
##	3rd Qu.: 6069563	3rd Qu.:22.77	3rd Qu.:10.650	3rd Qu.: 5.680	
##	Max. :35894000	Max. :53.32	Max. :29.600	Max. :14.430	
##	vehicmiles	unem	perc14_24	sl70plus	
##	Min. : 3.703	Min. : 2.200	Min. :11.70	Min. :0.0000	
##	1st Qu.: 14.574	1st Qu.: 4.500	1st Qu.:13.90	1st Qu.:0.0000	
##	Median : 33.863	Median : 5.600	Median :14.90	Median :0.0000	
##	Mean : 46.323	Mean : 5.951	Mean :15.33	Mean :0.2068	
##	3rd Qu.: 58.639	3rd Qu.: 7.000	3rd Qu.:16.60	3rd Qu.:0.0000	
##	Max. :329.600	Max. :18.000	Max. :20.30	Max. :1.0000	
##	sbprim	sbsecon	d80	d81	d82

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	:0.0000	:0.0000	:0.0000	:0.1792	:0.0000	:1.0000
	:0.0000	:0.0000	:0.0000	:0.4683	:1.0000	:1.0000
	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d83	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d84	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d85	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d86	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d87	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d88	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d89	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d90	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d91	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d92	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d93	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d94	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d95	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d96	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d97	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d98	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d99	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d00	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d01	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d02	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d03	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
d04	:0.00	:0.00	:0.00	:0.04	:0.00	:1.00
vehicmilespc	: 4372	: 7788	: 9013	: 9129	:10327	:18390

```

data1 <- subset(data, state == 1)
data2 <- subset(data, state == 2)
data3 <- subset(data, state == 3)
data4 <- subset(data, state == 4)
data5 <- subset(data, state == 5)

```

```

data6 <- subset(data, state == 6)
data7 <- subset(data, state == 7)
data8 <- subset(data, state == 8)
data9 <- subset(data, state == 9)
data10 <- subset(data, state == 10)
data11 <- subset(data, state == 11)
data12 <- subset(data, state == 12)

#plot the first data series using plot()
plot(data1$year, data1$totfatrte, type="o", col="blue", pch="o", ylab="y", lty=1, ylim=c(0,50))

#add third data series to the same chart using points() and lines()
points(data3$year, data3$totfatrte, col="dark red", pch="+")
lines(data3$year, data3$totfatrte, col="dark red", lty=3)

points(data4$year, data4$totfatrte, col="red", pch="*")
lines(data4$year, data4$totfatrte, col="red", lty=2)

points(data5$year, data5$totfatrte, col="purple", pch="-")
lines(data5$year, data5$totfatrte, col="purple", lty=1)

points(data6$year, data6$totfatrte, col="orange", pch="-")
lines(data6$year, data6$totfatrte, col="orange", lty=4)

points(data7$year, data7$totfatrte, col="pink", pch="-")
lines(data7$year, data7$totfatrte, col="pink", lty=2)

points(data8$year, data8$totfatrte, col="green", pch="-")
lines(data8$year, data8$totfatrte, col="green", lty=4)

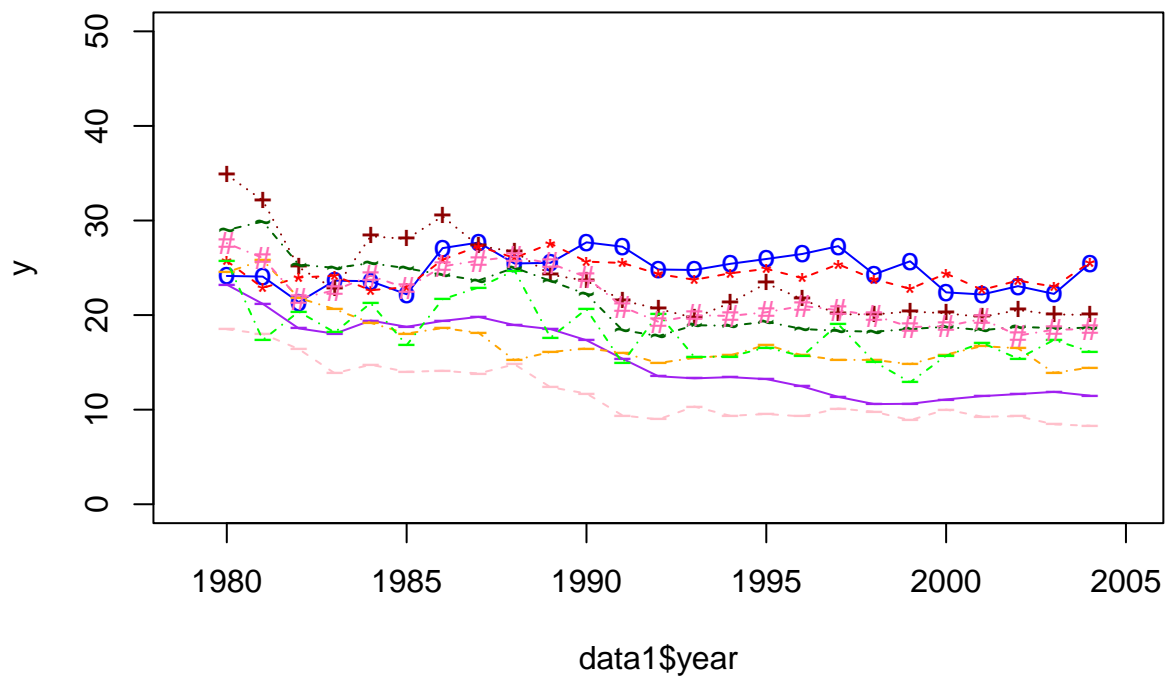
points(data9$year, data9$totfatrte, col="dark blue", pch="-")
lines(data9$year, data9$totfatrte, col="dark blue", lty=4)

points(data10$year, data10$totfatrte, col="dark green", pch="~")
lines(data10$year, data10$totfatrte, col="dark green", lty=4)

points(data11$year, data11$totfatrte, col="hot pink", pch="#")
lines(data11$year, data11$totfatrte, col="hot pink", lty=4)

points(data12$year, data12$totfatrte, col="sea green", pch="+")
lines(data12$year, data12$totfatrte, col="sea green", lty=4)

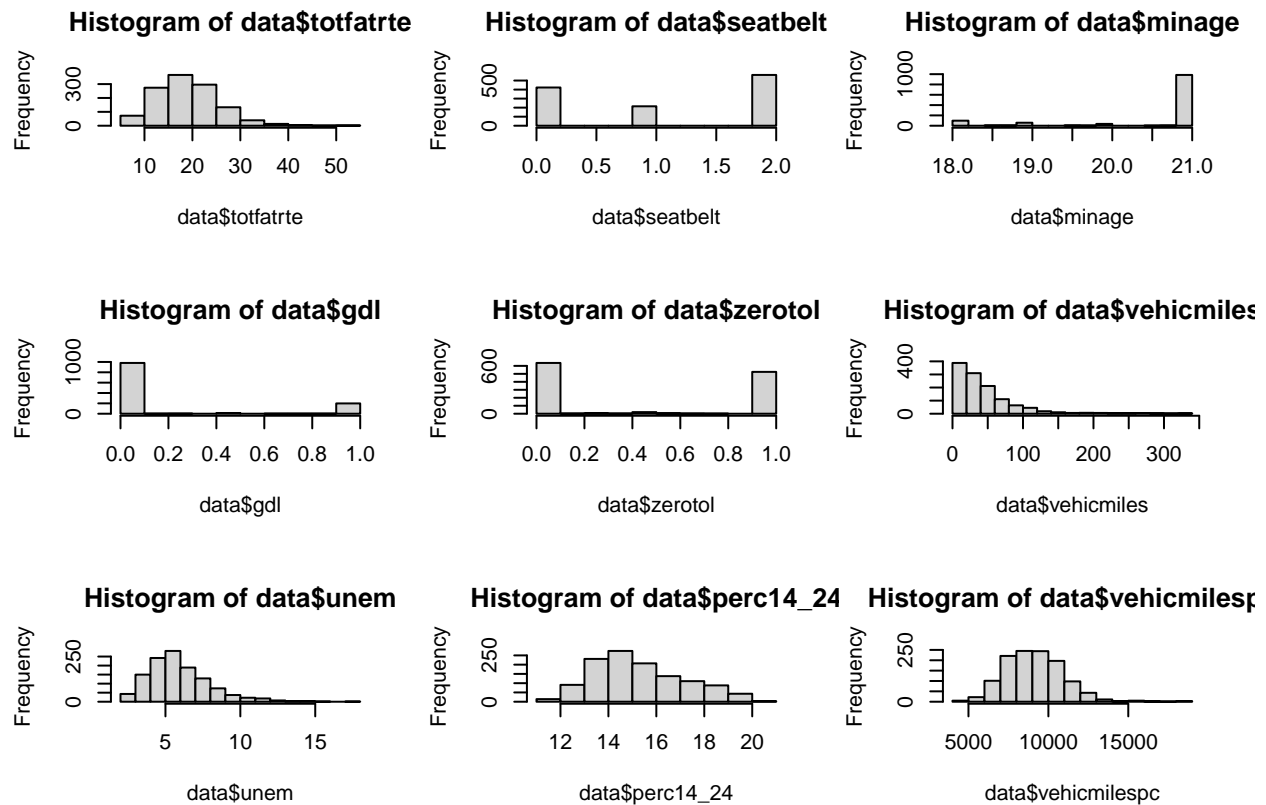
```



```

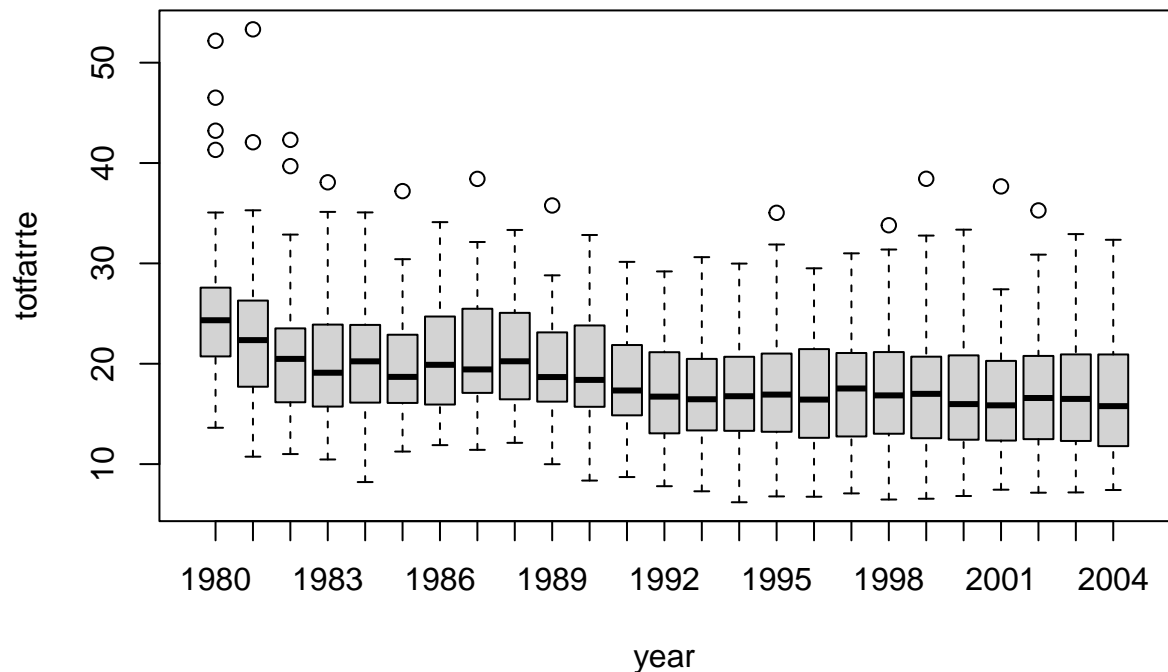
par(mfrow=c(3,3))
hist(data$totfatrte)
hist(data$seatbelt)
hist(data$minage)
hist(data$gdl)
hist(data$zerotol)
hist(data$vehicmiles)
hist(data$unem)
hist(data$perc14_24)
hist(data$vehicmilespc)

```

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.

```
boxplot(totfatrte ~ year, data=data)
```



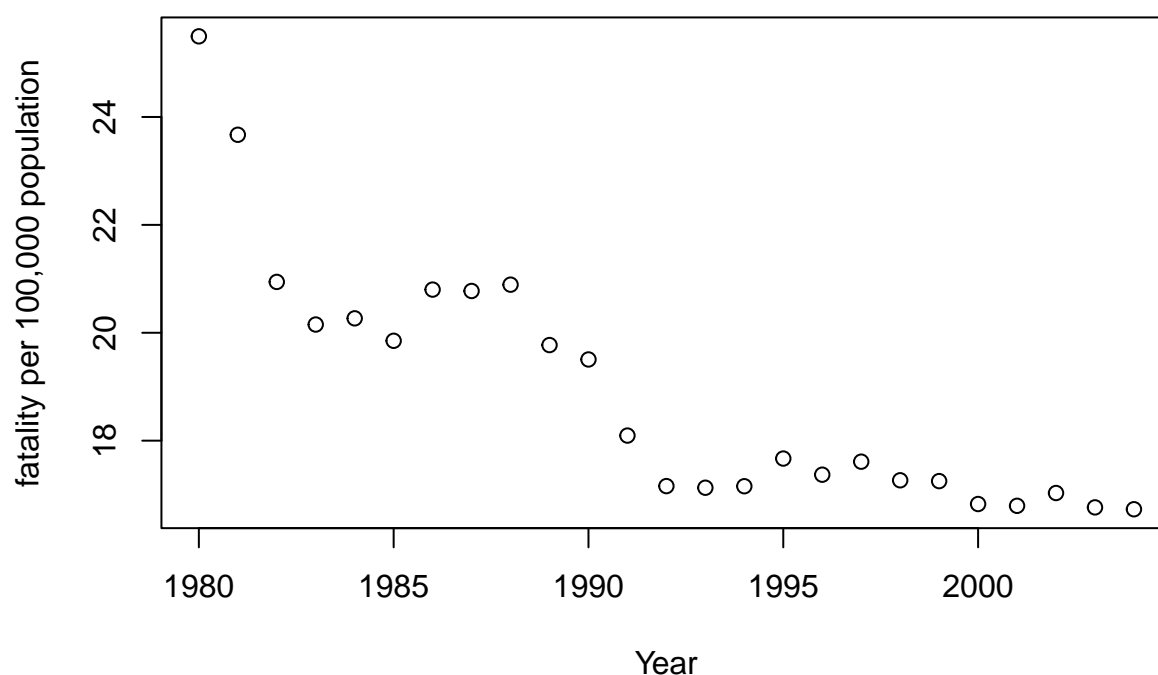
```
data_sub = data[, c('totfatrte', 'year')]
for (year in unique(data_sub$year)) {
  if (year == min(data_sub$year)) {
    next
  }
  data_sub[paste('year_', year, sep='')] = (data_sub$year == year) * 1
}
data_sub$year = NULL
mdl1 = lm('totfatrte ~ .', data=data_sub)
summary(mdl1)
```

```
##
## Call:
## lm(formula = "totfatrte ~ .", data = data_sub)
##
## 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 ***
```

```
## year_1981    -1.8244      1.2263   -1.488  0.137094
## year_1982    -4.5521      1.2263   -3.712  0.000215 ***
## year_1983    -5.3417      1.2263   -4.356  1.44e-05 ***
## year_1984    -5.2271      1.2263   -4.263  2.18e-05 ***
## year_1985    -5.6431      1.2263   -4.602  4.64e-06 ***
## year_1986    -4.6942      1.2263   -3.828  0.000136 ***
## year_1987    -4.7198      1.2263   -3.849  0.000125 ***
## year_1988    -4.6029      1.2263   -3.754  0.000183 ***
## year_1989    -5.7223      1.2263   -4.666  3.42e-06 ***
## year_1990    -5.9894      1.2263   -4.884  1.18e-06 ***
## year_1991    -7.3998      1.2263   -6.034  2.14e-09 ***
## year_1992    -8.3367      1.2263   -6.798  1.68e-11 ***
## year_1993    -8.3669      1.2263   -6.823  1.43e-11 ***
## year_1994    -8.3394      1.2263   -6.800  1.66e-11 ***
## year_1995    -7.8260      1.2263   -6.382  2.51e-10 ***
## year_1996    -8.1252      1.2263   -6.626  5.25e-11 ***
## year_1997    -7.8840      1.2263   -6.429  1.86e-10 ***
## year_1998    -8.2292      1.2263   -6.711  3.01e-11 ***
## year_1999    -8.2442      1.2263   -6.723  2.77e-11 ***
## year_2000    -8.6690      1.2263   -7.069  2.67e-12 ***
## year_2001    -8.7019      1.2263   -7.096  2.21e-12 ***
## year_2002    -8.4650      1.2263   -6.903  8.32e-12 ***
## year_2003    -8.7310      1.2263   -7.120  1.88e-12 ***
## 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
```

```
year_avg = as.numeric mdl1$coefficients
year_avg[2:length(year_avg)] = year_avg[2:length(year_avg)] + year_avg[1]
plot(unique(data$year), year_avg, main='Total fatalities per 100,000 population through time',
```

Total fatalities per 100,000 population through time



Response to Q2

The average fatalities per 100,000 people (*totfatrte*) for each year can be computed through the regression, where the fatalities are regressed on year dummy variables, and by skipping the first year since the intercept of the regression will represent the average for 1980, and all further coefficients will represent the adjustment to the intercept term to arrive at the average for that given year. Looking at the chart, we can see that the average fatalities per 100,000 people have been trending down overall, with a big improvements in the early and late 1980's, as well as early 1990's, with only a minor improvement from then until 2004.

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.)

```
vars = c('bac08', 'bac10', 'perse', 'sbprim', 'sbsecon', 'sl70plus', 'gdl', 'perc14_24', 'unem', 'vehicmilespc')

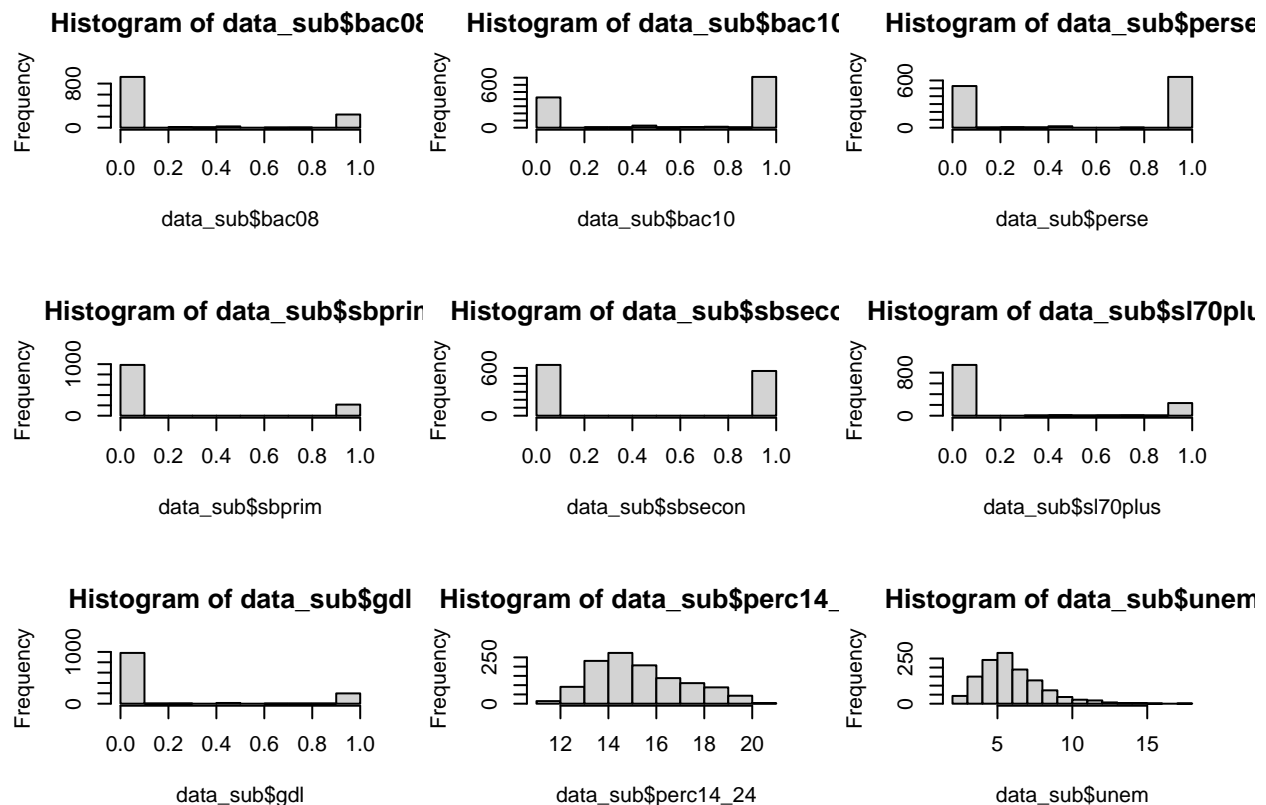
for (var in vars) {
  data_sub[var] = data[var]
```

```

#hist(data_sub[var])
}

par(mfrow=c(3,3))
hist(data_sub$bac08)
hist(data_sub$bac10)
hist(data_sub$perse)
hist(data_sub$sbprim)
hist(data_sub$sbsecon)
hist(data_sub$sl70plus)
hist(data_sub$gdl)
hist(data_sub$perc14_24)
hist(data_sub$unem)

```



```

#hist(data_sub$vehicmilespc)

data_sub$unem = log(data_sub$unem)
for (i in 1:7) {
  data_sub[vars[i]] = (data_sub[vars[i]] > 0.5) * 1
}
mdl2 = lm('totfatrte ~ .', data=data_sub)

```

```
summary mdl2)
```

```
##
## Call:
## lm(formula = "totfatrte ~ .", data = data_sub)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4031  -2.6086  -0.3265   2.2414  21.8650
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.012e+00  2.620e+00  -3.058 0.002277 **
## year_1981    -2.107e+00  8.229e-01  -2.560 0.010578 *
## year_1982    -6.304e+00  8.397e-01  -7.508 1.19e-13 ***
## year_1983    -7.190e+00  8.515e-01  -8.445 < 2e-16 ***
## year_1984    -5.826e+00  8.666e-01  -6.723 2.78e-11 ***
## year_1985    -6.458e+00  8.852e-01  -7.296 5.48e-13 ***
## year_1986    -5.634e+00  9.231e-01  -6.103 1.42e-09 ***
## year_1987    -6.065e+00  9.613e-01  -6.309 3.98e-10 ***
## year_1988    -6.176e+00  1.011e+00  -6.109 1.36e-09 ***
## year_1989    -7.688e+00  1.049e+00  -7.325 4.43e-13 ***
## year_1990    -8.682e+00  1.072e+00  -8.102 1.36e-15 ***
## year_1991    -1.087e+01  1.093e+00  -9.944 < 2e-16 ***
## year_1992    -1.263e+01  1.114e+00 -11.335 < 2e-16 ***
## year_1993    -1.250e+01  1.128e+00 -11.085 < 2e-16 ***
## year_1994    -1.208e+01  1.150e+00 -10.500 < 2e-16 ***
## year_1995    -1.147e+01  1.180e+00  -9.722 < 2e-16 ***
## year_1996    -1.340e+01  1.223e+00 -10.962 < 2e-16 ***
## year_1997    -1.352e+01  1.244e+00 -10.864 < 2e-16 ***
## year_1998    -1.420e+01  1.268e+00 -11.197 < 2e-16 ***
## year_1999    -1.415e+01  1.284e+00 -11.019 < 2e-16 ***
## year_2000    -1.440e+01  1.307e+00 -11.021 < 2e-16 ***
## year_2001    -1.567e+01  1.317e+00 -11.903 < 2e-16 ***
## year_2002    -1.649e+01  1.326e+00 -12.434 < 2e-16 ***
## year_2003    -1.692e+01  1.331e+00 -12.716 < 2e-16 ***
## year_2004    -1.633e+01  1.367e+00 -11.947 < 2e-16 ***
## bac08        -2.288e+00  4.858e-01  -4.709 2.79e-06 ***
## bac10        -1.256e+00  3.591e-01  -3.497 0.000489 ***
## perse        -5.625e-01  2.919e-01  -1.927 0.054231 .
## sbprim       -3.795e-01  4.898e-01  -0.775 0.438515
## sbsecon      -1.535e-01  4.279e-01  -0.359 0.719911
## sl70plus      3.112e+00  4.331e-01   7.186 1.19e-12 ***
## gdl          -3.014e-01  5.066e-01  -0.595 0.552051
## perc14_24     1.776e-01  1.222e-01   1.453 0.146542
## unem          5.152e+00  4.812e-01  10.707 < 2e-16 ***
## vehicmilespc  2.921e-03  9.393e-05  31.096 < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.024 on 1165 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6006
## F-statistic: 54.02 on 34 and 1165 DF,  p-value: < 2.2e-16

anova(mdl1, mdl2, test="Chisq")

## Analysis of Variance Table
##
## Model 1: totfatrte ~ year_1981 + year_1982 + year_1983 + year_1984 + year_1985 +
##   year_1986 + year_1987 + year_1988 + year_1989 + year_1990 +
##   year_1991 + year_1992 + year_1993 + year_1994 + year_1995 +
##   year_1996 + year_1997 + year_1998 + year_1999 + year_2000 +
##   year_2001 + year_2002 + year_2003 + year_2004
## Model 2: totfatrte ~ year_1981 + year_1982 + year_1983 + year_1984 + year_1985 +
##   year_1986 + year_1987 + year_1988 + year_1989 + year_1990 +
##   year_1991 + year_1992 + year_1993 + year_1994 + year_1995 +
##   year_1996 + year_1997 + year_1998 + year_1999 + year_2000 +
##   year_2001 + year_2002 + year_2003 + year_2004 + bac08 + bac10 +
##   perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##   vehicmilespc
##   Res.Df    RSS Df Sum of Sq  Pr(>Chi)
## 1    1175 42407
## 2    1165 18867 10      23540 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Response to Q3

Since bac08, bac10, perse, sbprim, sbsecon, sl70plus and gdl, are 0-1 variables that represent the fraction of a year that a specific speed limit / blood alcohol level / law was in place for (for a given year), these are transformed into 0-1 hot encodings (0 = 0.50 or less, 1 if > 0.50) for better interpretability. Furthermore, the unem variable, which is the unemployment rate, is log transformed to take out the skew of its distribution as mentioned in the EDA. The remaining variables (perc14_14 and vehicmilespc) are not transformed since their distributions are not skewed.

After running the regression with the added variables, we can see from the anova (chi-square) test that the added variables are adding significant explanatory power to the model. bac08 and bac10 variables represent the blood alcohol level limits that are prescribed by law, and these two variables can either be 0 for both if there are no limits, 1 for one and 0 for the other, or a fractions that can sum to 1 if one limit was used for part of one year and the other was used for the remaining. The coefficients for these two variables in the regression are -2.288 and -1.256, which implies that total fatalities per 100,000 people decreases by these amounts when these limits are in place. Interestingly and not surprisingly, going from no limit to 0.08 limit reduces the fatalities more than going for 0.08 to 0.1. Both coefficients are significant at the 0.01 confidence level.

Looking at per se laws (perse), the coefficient is negative as expected, and specifically, the model estimates that by introducing this per se law, total fatalities per 100,000 people is reduced by -0.5625. This variable is somewhat significant at the 0.1 level. Moving to the primary seat belt law, there is also an estimated negative effect with total fatalities per 100,000 people of -0.3795, however, this variable does not seem significant, and interestingly, the sbsecon (secondary seal belt law), is not significant either and has an even lower magnitude in estimated effect on totfatrte.

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?

```
#head(data)

# Transform the data (as per inputs from Dominik)
data$bac08bin = ifelse(data$bac08<=0.5,0,1)
data$bac10bin = ifelse(data$bac10<=0.5,0,1)
data$persebin = ifelse(data$perse<=0.5,0,1)
data$sbprimbin = ifelse(data$sbprim<=0.5,0,1)
data$sbseconbin = ifelse(data$sbsecon<=0.5,0,1)
data$sl70plusbin = ifelse(data$sl70plus<=0.5,0,1)
data$gdlbin = ifelse(data$gdl<0.5,0,1)
data$unemlog = log(data$unem)

#Fixed Effects Model
model.fe = plm(totfatrte~bac08bin+bac10bin+persebin+sbprimbin+sbseconbin+sl70plusbin+gdlbin+
               perc14_24+unemlog+vehicmiles pc + 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, model=c("within"), index=c('state','year'), effect = "twoways")
summary(model.fe)

## Twoways effects Within Model
##
## Call:
## plm(formula = totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin +
##       sbseconbin + sl70plusbin + gdlbin + perc14_24 + unemlog +
##       vehicmiles pc + 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, effect = "twoways",
##       model = c("within"), index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.      1st Qu.      Median      3rd Qu.      Max.
## -8.261550 -1.041583 -0.019056  0.984573 14.642734
##
```



```

## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## bac08bin      -1.10399614  0.33022452 -3.3432 0.0008557 ***
## bac10bin      -0.80560450  0.22564600 -3.5702 0.0003718 ***
## persebin      -1.12079848  0.22340455 -5.0169 6.104e-07 ***
## sbprimbin     -1.19280857  0.34308412 -3.4767 0.0005271 ***
## sbseconbin    -0.30877884  0.25213895 -1.2246 0.2209698
## sl70plusbin   0.05315694  0.26093278  0.2037 0.8386103
## gdlbin        -0.37500598  0.27964169 -1.3410 0.1801853
## perc14_24     0.16346044  0.09545138  1.7125 0.0870819 .
## unemlog       -3.71639544  0.39225223 -9.4745 < 2.2e-16 ***
## vehicmilespsc 0.00094479  0.00011034  8.5629 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5928.9
## Residual Sum of Squares: 4549.2
## R-Squared:              0.23271
## Adj. R-Squared: 0.17711
## F-statistic: 33.9069 on 10 and 1118 DF, p-value: < 2.22e-16

```

Response to Q4

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

The coefficients for bac08, bac10, perse and sbprim are all statistically significant with the panel fixed-effects model. The estimate for the bac08 coeff is slightly lower, the coeff for bac10 is about the same as before. The perse coeff has changed from -2.06 to -1.12; Also the coeff for sbprim is now statistically significant which was not the case earlier.

Which set of estimates do you think is more reliable?

1. The FE model models the variation over time in *totfatrte* and all the independent variables *within* each state. So it is more reliable

What assumptions are needed in each of these models? Are these assumptions reasonable in the current context?

FE Assumption - The fixed effect assumption is that the individual-specific effects are correlated with the independent variables.

Pooled OLS assumption - The Key assumption of Pooled OLS is that there are unique, time constant attributes of individuals that are not correlated with the individual regressors.

In the current context, it is very difficult to test assumption for Poole OLS model. In other words, FE model assumptions are more reasonable to proceed.

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

```
model.re = plm(totfatrtte~bac08bin+bac10bin+persebin+sbprimbin+sbseconbin+sl70plusbin+gdlbin+
               perc14_24+unemlog+vehicmilespc + 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, model=c("random"), index=c('state','year'))
```

```
summary(model.re)
```

```
## Oneway (individual) effect Random Effect Model
##      (Swamy-Arora's transformation)
##
## Call:
## plm(formula = totfatrtte ~ bac08bin + bac10bin + persebin + sbprimbin +
##      sbseconbin + sl70plusbin + gdlbin + perc14_24 + unemlog +
##      vehicmilespc + 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, model = c("random"),
##      index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Effects:
##              var std.dev share
## idiosyncratic 4.069    2.017 0.342
## individual    7.819    2.796 0.658
## theta: 0.8572
##
## Residuals:
##      Min.   1st Qu.   Median   3rd Qu.    Max.
## -8.41735 -1.21591 -0.16209  0.94795 16.39919
##
## Coefficients:
##              Estimate Std. Error z-value Pr(>|z|)
## (Intercept)  1.9880e+01 2.2639e+00  8.7814 < 2.2e-16 ***
## bac08bin     -1.2065e+00 3.4049e-01 -3.5434 0.0003950 ***
## bac10bin     -8.6902e-01 2.3321e-01 -3.7264 0.0001943 ***
## persebin     -1.0683e+00 2.2905e-01 -4.6640 3.100e-06 ***
## sbprimbin    -1.1434e+00 3.5286e-01 -3.2404 0.0011937 **
## sbseconbin   -3.0957e-01 2.6110e-01 -1.1856 0.2357635
## sl70plusbin   1.3203e-01 2.7008e-01  0.4889 0.6249426
## gdlbin       -3.4065e-01 2.9016e-01 -1.1740 0.2403837
## perc14_24     1.7892e-01 9.7841e-02  1.8286 0.0674526 .
```

```

## unemlog      -3.1341e+00  4.0073e-01  -7.8209  5.244e-15 ***
## vehicmilespc  1.1891e-03  1.0929e-04  10.8797 < 2.2e-16 ***
## d81          -1.6116e+00  4.3015e-01  -3.7465  0.0001793 ***
## d82          -3.5672e+00  4.5056e-01  -7.9172  2.429e-15 ***
## d83          -4.2443e+00  4.6221e-01  -9.1826 < 2.2e-16 ***
## d84          -4.6567e+00  4.7702e-01  -9.7620 < 2.2e-16 ***
## d85          -5.1300e+00  4.9800e-01 -10.3012 < 2.2e-16 ***
## d86          -4.1427e+00  5.3277e-01  -7.7759  7.493e-15 ***
## d87          -4.8456e+00  5.7325e-01  -8.4528 < 2.2e-16 ***
## d88          -5.3904e+00  6.2372e-01  -8.6423 < 2.2e-16 ***
## d89          -6.7373e+00  6.6286e-01 -10.1639 < 2.2e-16 ***
## d90          -6.8719e+00  6.8557e-01 -10.0238 < 2.2e-16 ***
## d91          -7.6375e+00  7.0062e-01 -10.9011 < 2.2e-16 ***
## d92          -8.5935e+00  7.2137e-01 -11.9127 < 2.2e-16 ***
## d93          -8.9144e+00  7.3512e-01 -12.1264 < 2.2e-16 ***
## d94          -9.3465e+00  7.5459e-01 -12.3862 < 2.2e-16 ***
## d95          -9.1062e+00  7.8024e-01 -11.6710 < 2.2e-16 ***
## d96          -9.5842e+00  8.2246e-01 -11.6530 < 2.2e-16 ***
## d97          -9.8431e+00  8.4367e-01 -11.6669 < 2.2e-16 ***
## d98          -1.0564e+01  8.6460e-01 -12.2180 < 2.2e-16 ***
## d99          -1.0769e+01  8.7732e-01 -12.2754 < 2.2e-16 ***
## d00          -1.1361e+01  8.9168e-01 -12.7414 < 2.2e-16 ***
## d01          -1.0985e+01  8.9318e-01 -12.2989 < 2.2e-16 ***
## d02          -1.0261e+01  8.9371e-01 -11.4813 < 2.2e-16 ***
## d03          -1.0335e+01  8.9485e-01 -11.5491 < 2.2e-16 ***
## d04          -1.0715e+01  9.2133e-01 -11.6301 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12878
## Residual Sum of Squares: 5131.3
## R-Squared:      0.60155
## Adj. R-Squared: 0.58992
## Chisq: 1758.8 on 34 DF, p-value: < 2.22e-16

```

Response to Q5

1. In a random effects model, the unobserved variables are assumed to be uncorrelated with (or, more strongly, statistically independent of) all the observed variables.
2. In a fixed effects model, the unobserved variables are allowed to have any associations whatsoever with the observed variables.
3. For the given data, assumption for fixed effect model seems to be more practical as it is very difficult to test/prove that unobserved variables are not correlated with observed variables if we use random effect modeling. On other hand, it is possible to remove omitted/unobserved variable bias using fixed effects modeling

with dummy variables of each year to explain the unaccounted for time-variant error dependence.

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

Response to Q6

1. From the FE model, the coefficient for the *vehicmilespc* variable is 0.00094479
2. *totfatrte* is Fatalities/100K people per mile-driven/capita.
3. If *vehicmilespc* increases by 1,000, the value of *totfatrte* increases as follows

$$\text{Increase} - \text{In} - \text{totfatrte} - \text{value} = 1000 * 0.00094479 = 0.94479$$

4. In summary, estimated effect on *totfatrte* *vehicmilespc*, if the number of miles driven per capita, increases by 1,000, is an increase in its value by 0.94479 (provided all other things do not change)
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?

Response for Q7

1. With, positive serial correlation, the OLS estimates of the standard errors will be smaller than the true standard errors. This will lead to the conclusion that the parameter estimates are more precise than they really are. There will be a tendency to reject the null hypothesis when it should not be rejected. In other words, we will consider insignificant variables as significant.
2. On the contrary, with the presence of heteroskedasticity in our error, we will fail to reject null hypothesis when it should be rejected. In other words, we will not be in a position to detect the significance of potentially valuable regressor.