W271 Lab 4

Shan He, Joanna Huang, Tiffany Jaya, Robert Deng

```
# Load libraries
library(knitr)
library(car)
library(dplyr)
library(Hmisc)
library(gplots)
library(ggplot2)
library(lattice)
library(plm)
library(grid)
library(gridExtra)
# prevent source code from running off the page
opts_chunk$set(tidy.opts=list(width.cutoff=70), tidy=TRUE)
# remove all objects from current workspace
rm(list = ls())
# set seed number to reproduce results
set.seed(1)
# load data
load("data/driving.RData")
```

Introduction

According to the National Safety Council, data fatalities claimed 40,100 lives in 2017, 6% higher than the number of deaths in 2015. Alongside these death numbers are more than 4 million injuries that were a result of motor vehicle accidents. Furthermore, a fourth of those fatalities can be attributed to deaths from alcohol impaired driving. In this lab, we look into the general and alcohol-related data laws and demographic statistics of 48 continental U.S. states from 1980 through 2004 to answer the question **"Do changes in data laws affect data fatalities?"**
Source: https://www.nsc.org/road-safety/safety-topics/fatality-estimates, https://www.nhtsa.gov/risky-driving/drunk-driving

About the data

```
load("data/driving.RData")
str(data)
```

```
## 'data.frame':    1200 obs. of  56 variables:
##  $ year        : int  1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
##  $ state       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sl55        : num  1 1 1 1 1 ...
```

```
##  $ sl65        : num  0 0 0 0 0 ...
##  $ sl70        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ sl75        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ slnone      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ seatbelt    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ minage      : num  18 18 18 18 18 20 21 21 21 21 ...
##  $ zerotol     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ gdl         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ bac10       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ bac08       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ perse       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ totfat      : int  940 933 839 930 932 882 1080 1111 1024 1029 ...
##  $ nghtfat     : int  422 434 376 397 421 358 500 499 423 418 ...
##  $ wkndfat     : int  236 248 224 223 237 224 279 300 226 247 ...
##  $ totfatpvm   : num  3.2 3.35 2.81 3 2.83 ...
##  $ nghtfatpvm  : num  1.44 1.56 1.26 1.28 1.28 ...
##  $ wkndfatpvm  : num  0.803 0.89 0.75 0.719 0.72 ...
##  $ statepop    : int  3893888 3918520 3925218 3934109 3951834 3972527 3991569 4015261 402385
##  $ totfatrte   : num  24.1 24.1 21.4 23.6 23.6 ...
##  $ nghtfatrte  : num  10.84 11.08 9.58 10.09 10.65 ...
##  $ wkndfatrte  : num  6.06 6.33 5.71 5.67 6 ...
##  $ vehicmiles  : num  29.4 27.9 29.9 31 32.9 ...
##  $ unem        : num  8.8 10.7 14.4 13.7 11.1 ...
##  $ perc14_24   : num  18.9 18.7 18.4 18 17.6 ...
##  $ sl70plus    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ sbprim      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ sbsecon     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d80         : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ d81         : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ d82         : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ d83         : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ d84         : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ d85         : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ d86         : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ d87         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ d88         : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ d89         : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ d90         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d91         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d92         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d93         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d94         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d95         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d96         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d97         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d98         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d99         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d00         : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ d01         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d02         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d03         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d04         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ vehicmilespc: num  7544 7108 7607 7880 8334 ...
##  - attr(*, "datalabel")= chr ""
##  - attr(*, "time.stamp")= chr "22 Jan 2013 14:09"
##  - attr(*, "formats")= chr  "%8.0g" "%8.0g" "%9.0g" "%9.0g" ...
##  - attr(*, "types")= int  252 251 254 254 254 254 254 251 254 254 ...
##  - attr(*, "val.labels")= chr  "" "" "" "" ...
##  - attr(*, "var.labels")= chr  "1980 through 2004" "48 continental states, alphabetical" "s
##  - attr(*, "version")= int 12
```

```
unique(data$state)
```

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

This dataset contains 1200 observations of 56 variables that spans across 25 years from 1980 to 2004. There is a year column, as well as 25 dummy variables corresponding to each year. The dataset contains data for 48 continental U.S. states and each state is numbered from 1 to 51 with 2, 9 and 12 missing. Remaining columns hold variables related to state's data laws, data fatalities, and population demographics. Below we list the variables, their definitions and their values:

data Laws - "sl*" variables correspond to the speed limit mandated by the state in a given year. For example, "sl55" relates to whether or not a state had a 55 mph speed limit. Values range from 0 to 1, with fractions representing the amount of time for which the limit was enforced by law that year given a change in state law. - "seatbelt" consists of values 0, 1 and 2 with 0 being no seatbelt law, 1 being primary (no other violation required to give a ticket), and 2 being secondary (additional violated required for a ticket). Dummy variables, "sbprim" and "sbsecon", also signify the the same thing. - "gdl" represents graduated driver licensing, which is a three-stage approach to granting young drivers full license privileges. Values range from 0 to 1, with fractions representing the amount of time for which the limit was enforced by law that year given a change in state law.

Alcohol-Related data Laws - "minage" is the minimum drinking age, with values from 18-21 and the majority being 21 years old. - "zerotol" indiciates whether or not a state enacted a Zero Tolerance policy for drinking that makes it a criminal offense when drivers under the age of 21 drive with even a small amount of alcohol in their system. Values under this column consist mostly of 0's and 1's with occasional fractions. - "bac" columns denote the blood alcohol content allowed by each state, with "bac08" representing 8% and "bac10" representing 10%. - "perse" relates to "Per se" laws in DUI cases. This law establishes that once an individual is shown to have a BAC at or above the state's allowed percent, that person is considered intoxicated by law. Such laws allow a person to be established as impaired without on-scene evaluation such as sobriety testing. Values range from 0 to 1, with fractions representing the amount of time for which the limit was enforced by law that year given a change in state law.

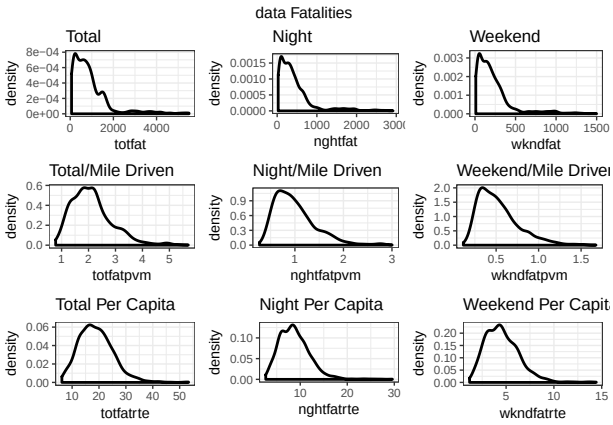
data Fatalities Fatality statistics are recorded for each state over the 25-year period. The statistics are provided in three ways: total fatalities, nighttime fatalities and weekend fatalities. Below are the exact definitions: - totfat, nghtfat, and wkndfat represent the total data/nighttime/weekend

fatalities with values ranging from ranging from 63 to 5504, 26-2918 and 10 to 1499 respectively - totfatpvm, nghtfatpvm, wkndfatpvm: total/nighttime/weekend fatalities per 100 million miles ranging from 0.78 to 5.70, 0.27 to 3.00, and 0.11 to 1.67 respectively - totfatrte,nghtfatrte, wkndfatrte: total/nighttime/weekend fatalities per 100,000 population ranging from 6.2 to 53.3, 2.7 to 29.6 and 1.2 to 14.4 respectively

Demographics In addition to the data laws and fatalities, this dataset contains demographic information including state population, unemployment rate, and percentage of population between the age 14 and 24. Also included are total vehicle miles driven (in billions) and per capita.

EDA

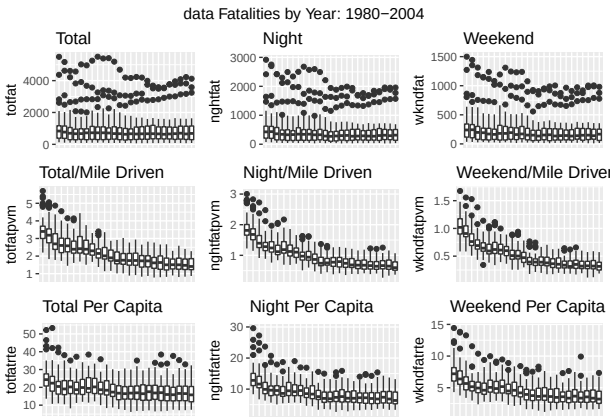
```
d1 <- data %>% ggplot(aes(totfat)) + geom_density(kernel = "gaussian",
    size = 1) + theme_bw() + ggtitle("Total")
d2 <- data %>% ggplot(aes(nghtfat)) + geom_density(kernel = "gaussian",
    size = 1) + theme_bw() + ggtitle("Night")
d3 <- data %>% ggplot(aes(wkndfat)) + geom_density(kernel = "gaussian",
    size = 1) + theme_bw() + ggtitle("Weekend")
d4 <- data %>% ggplot(aes(totfatpvm)) + geom_density(kernel = "gaussian",
    size = 1) + theme_bw() + ggtitle("Total/Mile Driven")
d5 <- data %>% ggplot(aes(nghtfatpvm)) + geom_density(kernel = "gaussian",
    size = 1) + theme_bw() + ggtitle("Night/Mile Driven")
d6 <- data %>% ggplot(aes(wkndfatpvm)) + geom_density(kernel = "gaussian",
    size = 1) + theme_bw() + ggtitle("Weekend/Mile Driven")
d7 <- data %>% ggplot(aes(totfatrte)) + geom_density(kernel = "gaussian",
    size = 1) + theme_bw() + ggtitle("Total Per Capita")
d8 <- data %>% ggplot(aes(nghtfatrte)) + geom_density(kernel = "gaussian",
    size = 1) + theme_bw() + ggtitle("Night Per Capita")
d9 <- data %>% ggplot(aes(wkndfatrte)) + geom_density(kernel = "gaussian",
    size = 1) + theme_bw() + ggtitle("Weekend Per Capita")
grid.arrange(d1, d2, d3, d4, d5, d6, d7, d8, d9, nrow = 3, ncol = 3, top = quote("data Fatalit
```



All fatality variables skew right, with the skew most apparent when not normalized by population or by mileage. We may consider taking the log of the fatalities given the values are always positive

and have a meaningful zero point. When comparing how a log transformation affects these variables, $totfatpvm$,$nghtfatpvm$ and $wkndfatpvm$ respond most favorably to the log transformation

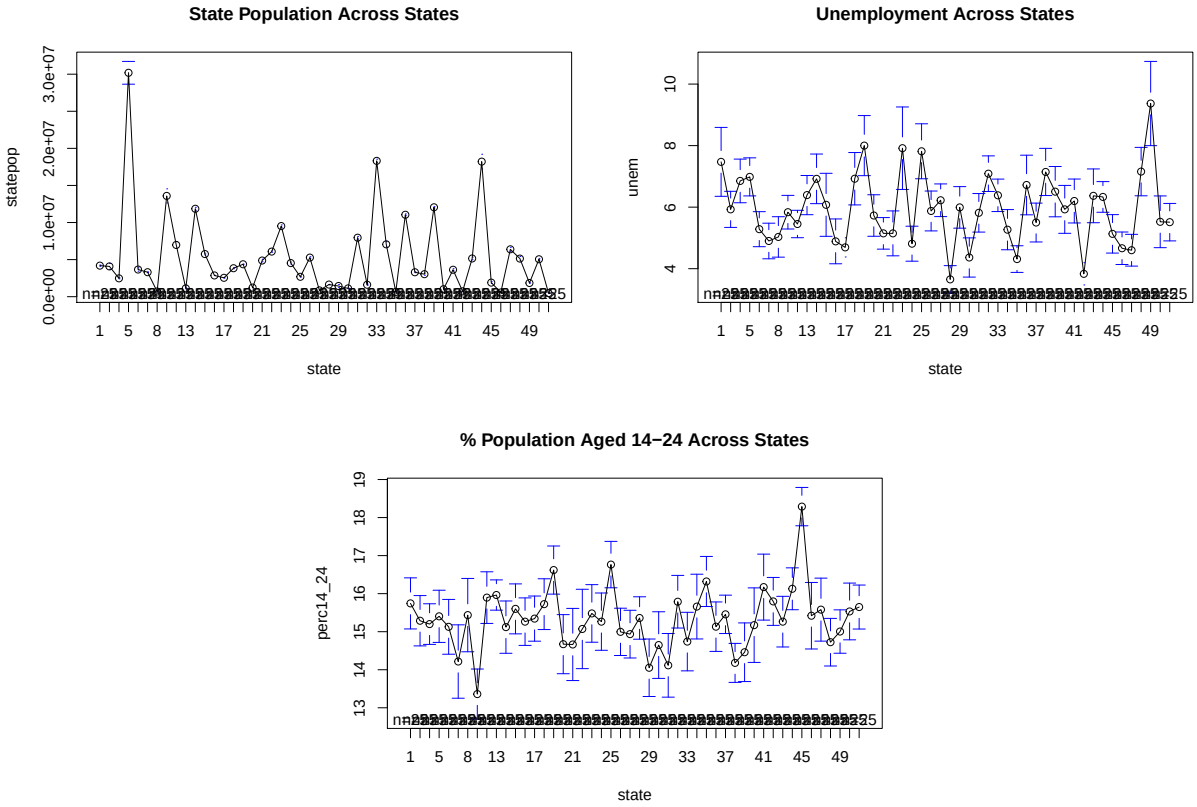
```
t <- theme(axis.title.x = element_blank(), axis.text.x = element_blank(),
    axis.ticks.x = element_blank())
b1 <- data %>% ggplot(aes(factor(year), totfat)) + geom_boxplot() + ggtitle("Total") +
    t
b2 <- data %>% ggplot(aes(factor(year), nghtfat)) + geom_boxplot() + ggtitle("Night") +
    t
b3 <- data %>% ggplot(aes(factor(year), wkndfat)) + geom_boxplot() + ggtitle("Weekend") +
    t
b4 <- data %>% ggplot(aes(factor(year), totfatpvm)) + geom_boxplot() +
    ggtitle("Total/Mile Driven") + t
b5 <- data %>% ggplot(aes(factor(year), nghtfatpvm)) + geom_boxplot() +
    ggtitle("Night/Mile Driven") + t
b6 <- data %>% ggplot(aes(factor(year), wkndfatpvm)) + geom_boxplot() +
    ggtitle("Weekend/Mile Driven") + t
b7 <- data %>% ggplot(aes(factor(year), totfatrte)) + geom_boxplot() +
    ggtitle("Total Per Capita") + t
b8 <- data %>% ggplot(aes(factor(year), nghtfatrte)) + geom_boxplot() +
    ggtitle("Night Per Capita") + t
b9 <- data %>% ggplot(aes(factor(year), wkndfatrte)) + geom_boxplot() +
    ggtitle("Weekend Per Capita") + t
grid.arrange(b1, b2, b3, b4, b5, b6, b7, b8, b9, nrow = 3, ncol = 3, top = quote("data Fatalit
```



The number of fatalities across total, night and weekend look mostly unchanged over the years. However, when looking at noramlized fatality rates, there are obvious downward trends at the national level.

Next, we look at how demographics differ between states.

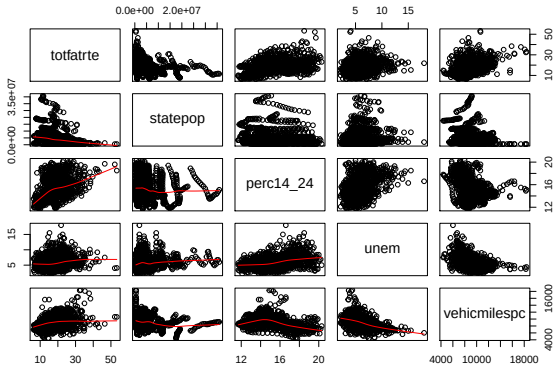
```
data.panel <- plm.data(data, c("state", "year"))
plotmeans(statepop ~ state, main = "State Population Across States", data = data.panel)
plotmeans(unem ~ state, main = "Unemployment Across States", data = data.panel)
plotmeans(perc14_24 ~ state, main = "% Population Aged 14-24 Across States",
    data = data.panel)
```



There are obvious differences in population across states with state 5, 33 and 44 standing out with 15 million+ population sizes compared to the average 5,329,896. Looking at unemployment rates, which average around 6%, state 28 and 42 demonstrate the lowest unemployment rates at less than 4% while state 49 demonstrates the highest with a median of 9. Percentage of population that is between the age of 14-24 also vary across state, with state 9 showing a low of 13.4% and state 45 showing a high of 18+%. Such variability in the demographics information shows how each state differs considerably from the other and how additional unobserved differences may be present.

Let's also look at the scatterplot matrix between key continuous variable and our dependent variable *totfatrte*

```
pairs(~totfatrte + statepop + perc14_24 + unem + vehicmilespc, data = data,
    lower.panel = panel.smooth)
```



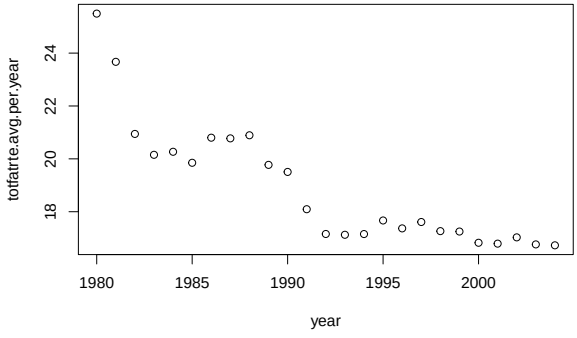
It doesn't seem like these variables need transformation since we didn't observe any nonlinear correlation.

Modeling

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

Our dependent variable *totfatrte* is defined as total fatalities per 100,000 population. Its average per year are listed below.

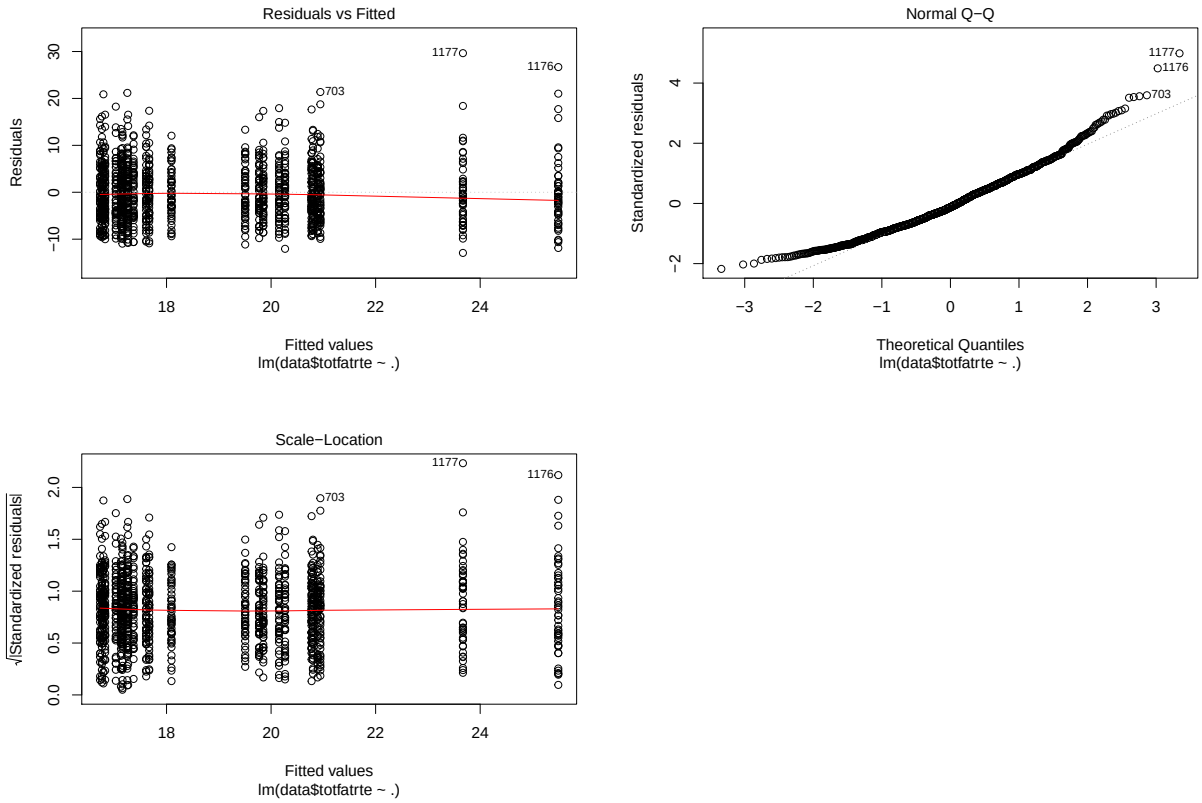
```
totfatrte.avg.per.year <- aggregate(list(totfatrte.avg.per.year = data$totfatrte),
    by = list(year = data$year), FUN = mean)
knitr::kable(totfatrte.avg.per.year)
plot(totfatrte.avg.per.year)
```



There is a noticeable downward trend with steep declines occuring twice: once from 1980 to 1983 and another one from 1988 to 1992.

```
dummy.1981.to.2004 <- c(paste("d", seq(81, 99, by = 1), sep = ""), paste("d0",
    seq(0, 4, by = 1), sep = ""))
m1 <- lm(data$totfatrte ~ ., data = data[dummy.1981.to.2004])
summary(m1)
plot(m1)
```

```
## hat values (leverages) are all = 0.02083333
##  and there are no factor predictors; no plot no. 5
```



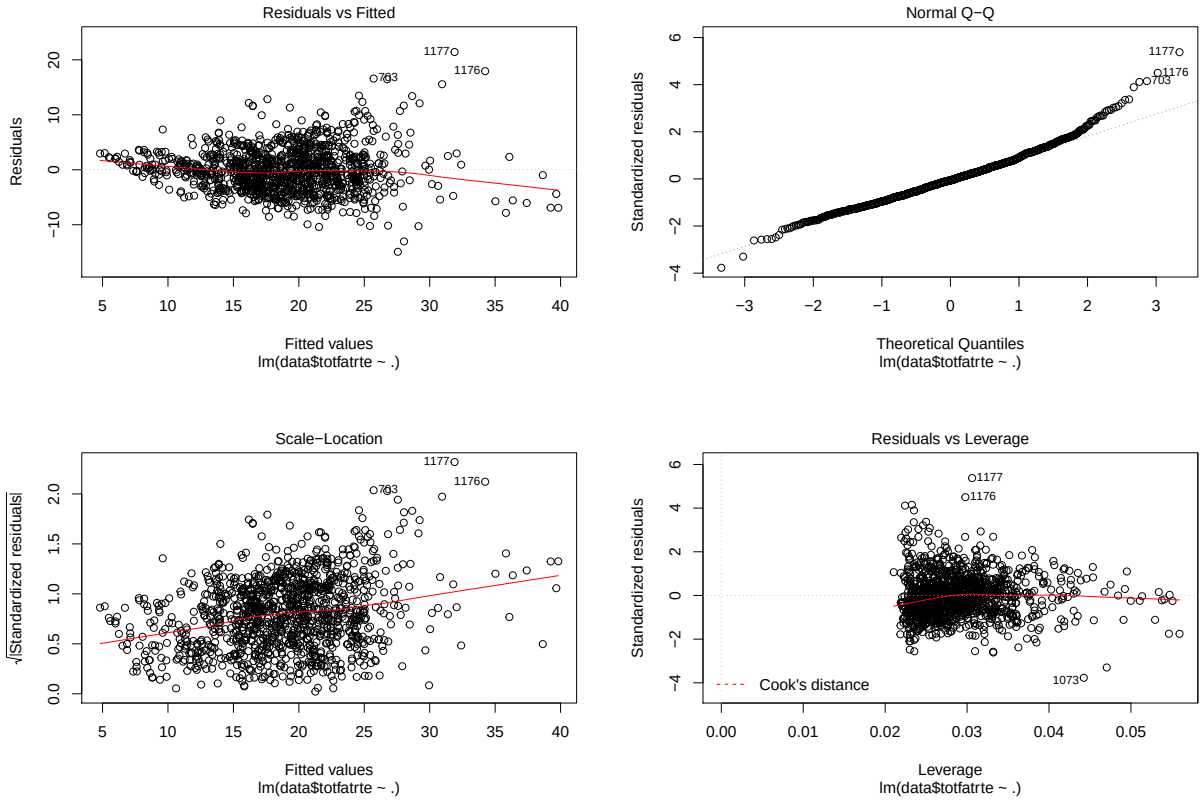
With 1980 as the base year, the intercept represents total fatality rate in that year. We can see that the intercept value 25.5 matches that of the average total fatality rate in 1980 (calculated in the first part of this question). In 1981, the estimated total fatality rate is 25.5 - 1.8(1), which equals 23.7 and so forth. Therefore, the fitted regression equation is totfatrte.avg.per.year $= 25.5 - 1.8 \cdot d81 - 4.6 \cdot d82 - 5.3 \cdot d83 - 5.2 \cdot d84 - 5.6 \cdot d85 - 4.7 \cdot d86 - 4.7 \cdot d87 - 4.6 \cdot d88 - 6.0 \cdot d90 - 7.4 \cdot d91 - 8.3 \cdot d92 - 8.4 \cdot d93 - 8.3 \cdot d94 - 7.8 \cdot d95 - 8.1 \cdot d96 - 7.9 \cdot d97 - 8.2 \cdot d98 - 8.2 \cdot d99 - 8.7 \cdot d00 - 8.7 \cdot d01 - 8.5 \cdot d02 - 8.7 \cdot d03 - 8.8 \cdot d04$. In other words, the model explains the average difference in the yearly fatalities as compared to the baseline year of 1980.

The negative coefficients represent a decrease in average fatality rate each year relative to 1980. Since the coefficients for the year dummy variables (except for 1981) are statistically significant at the 1% significance level, we see strong evidence that drivers become safer over the years.

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.)

```
explanatory_vars <- c("bac08", "bac10", "perse", "sbprim", "sbsecon", "sl70plus",
    "gdl", "perc14_24", "unem", "vehicmilespc", dummy.1981.to.2004)
m2 <- lm(data$totfatrte ~ ., data = data[explanatory_vars])
summary(m2)
plot(m2)
```



No transformation was needed for the continuous variables since 1) they are normalized and 2) they don't demonstrate obvious non-linear relationship with the dependent variable.

bac8 and *bac10* are defined to denote the blood alcohol content allowed by each state, with "bac08" representing 8% and "bac10" representing 10%. They are not strictly binary and the fraction denotes the fraction of time that the policy was imposed, just like the other "binary" variables" for the other traffic laws.

The coefficient for bac8 is -2.50 and bac10 -1.42. This means that with all other variables held constant, the fatality rate drops by 2.50 if 8% alchohol content is allowed and is enforced 100% of the time. On the other hand, the fatality rate only drops by 1.42 if 10% alchohol content is allowed and is enforced 100% of the time. Both coefficients are statistically significant but the model results suggest that all else held equal, a state with a blood alcohol content allowance of 8% would have a

lower $totfatrte$ than a state that has a threshold of 10%.

Per se law seemed to have a significant negative effect on fatality rate at the 5% significance level with a coefficient of -.62. Primary seat belt law, on the other hand, doesn't seem to have a significant impact on the fatality rate with a coefficient of -.08 that is not statistically significant.

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?

```
data.panel <- plm.data(data, c("state", "year"))
m3.fe <- plm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +
    d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 +
    d01 + d02 + d03 + d04 + bac08 + bac10 + perse + sbprim + sbsecon +
    sl70plus + gdl + perc14_24 + unem + vehicmilespc, data = data.panel,
    model = "within")
summary(m3.fe)
```

The coefficients for bac08 and bac10 decreased in size (from -2.5 to -1.4 and from -1.4 to -1.1 respectively) yet those for perse and sbprim became more pronounced (from -.62 to -1.2 and from -.08 to -1.2 respectively). The p-value for perse and sbprim also decreased, implying strong statistical significance (at 99.9% confidence level).

Pooled OLS regression assumes that the fixed effects and idiosyncratic errors are uncorrelated with any of the explanatory variables. A violation of this assumption leads to heterogeneity bias (or bias caused from omitting a time-constant variable) in the model estimates. In our context, such an assumption does not hold as we are observed the same states multiple times over a period of time. Thus there are bound to be constant state-specific effects, such as road conditions that are not included in the model and can affect our results.

Fixed effects models, on the other hand, allow for persisting influence, such as state economy or perspectives on alcohol, to be correlated with the explanatory variables, and also eliminates such within-state heterogeneity. Therefore, the estimates from this fixed effects model are more reliable since it controls for not just year-level but also state-level unobserved effect. To further test whether this is true, we can use the pFtest.

```
pFtest(m3.fe, m2)  # Testing for fixed effects, null: OLS better than fixed
```

```
##
##  F test for individual effects
##
## data:  totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +  ...
## F = 76.217, df1 = 47, df2 = 1118, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

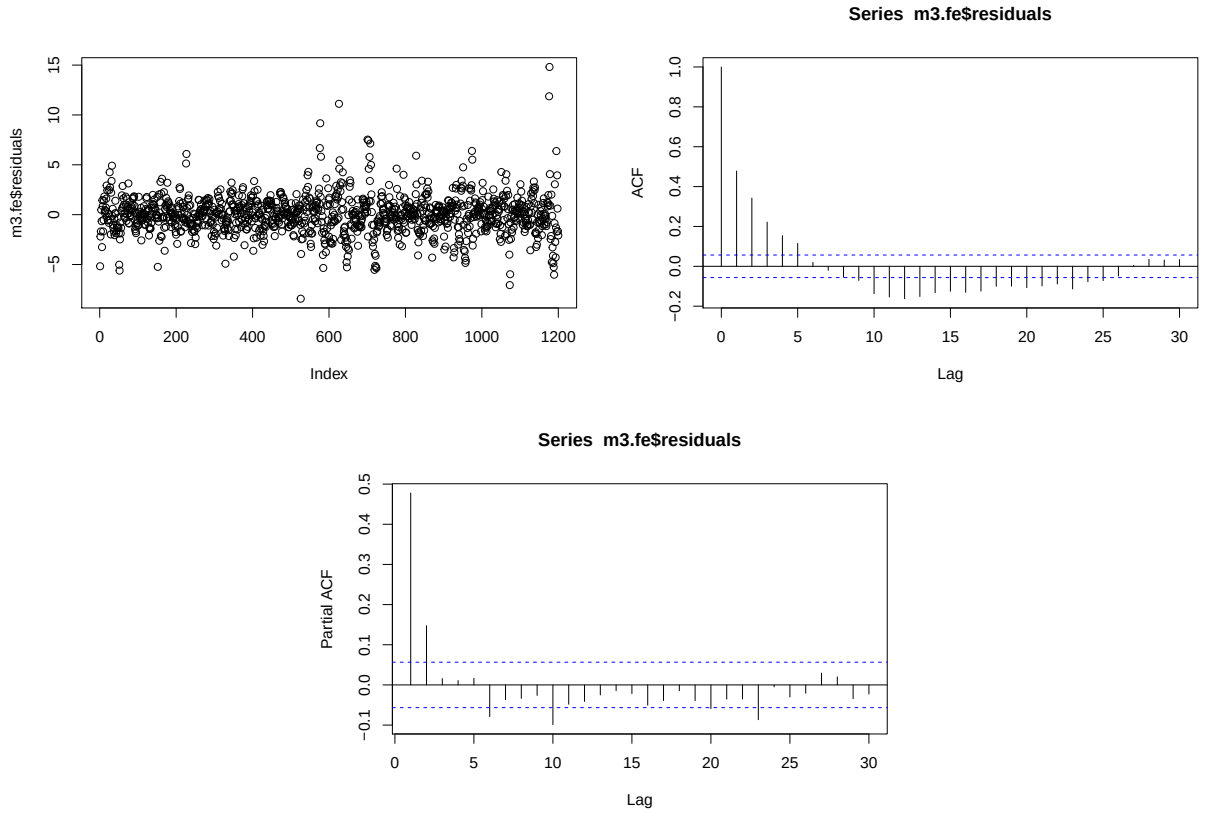
Since the p-value is less than .01 and statistically significant at the 1% level, we can reject the null that the pooled OLS regression is a better model than the fixed model.

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 as if explaining to a layperson.

Since the coefficient is 0.000940, an increase by 1,000 indicates that, with all other variables held constant, the fatality rate will go up by 0.94. In simple English, with all other variables being the same, a 1000 miles increase in the number of miles driven per captia will increase the fatality rate by 0.94.

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

```
plot(m3.fe$residuals)
acf(m3.fe$residuals)
pacf(m3.fe$residuals)
```



If there is existence of heteroskedasticity, then our estimates would not be consistent. From the plot of the residuals, our model seems robust against heteroskedasticity.

If there existed postive serial correlation in the model errors, we would have underestimated the p-values of our estimates. Potential p-value underestimation means that we could wrongfully reject the null hypothesis and commit a Type I error. From the acf and pacf plots, we see evidence of postive serial correlation between our residuals.