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

Rajiv Nair, Atit Wongnophadol, Julia Ying

U.S. traffic fatalities: 1980-2004

Part 1

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.*

There are 1,200 observations and 56 variables in this dataset. The panel dataset has two indices, `year` and `state`. The 48 continental states, numbered based on the alphabetical order, each have one row of data per year, spanning 25 years from 1980-2004, with no missing data in the dataset. There are nine traffic fatality rate measures, measuring total, weekend, and nighttime fatality count, fatality per 100,000 population, and fatality per 100 million miles. The fatality rate per 100,000 population, `totfatrte`, is the outcome variable of interest in this study. Figure 1 shows the univariate EDA on `totfatrte`. `totfatrte` is asymmetrically distributed with a positive skew, ranging from 6.2 to 53.32 with a median of 18.92. On a state level, New Mexico, Wyoming and Mississippi have the highest total fatality rate averaged over 25 years, while New York, New Jersey and Massachusetts have the lowest averaged total fatality rate. The year to year fatality rate have small fluctuations in variance. In general, fatality rate decreases over time, with the exception of a small increase around 1987.

Figure 2 displays the total fatality rate over time for each of the 48 states. Most of the states have declining or steady fatality rate, with New Mexico, Nevada and Montana having the largest declines. Exceptions to this pattern were Wyoming, which shows a u-shaped pattern for fatality rate, and Mississippi, which has a slight increasing trend. Both of these states are among the states with highest average fatality rate.

```
load("driving.RData")

paste(length(unique(data$year)), min(data$year), max(data$year), unique(table(data$year)))
```

```
## [1] "25 1980 2004 48"
```

```r
paste(length(unique(data$state)), unique(table(data$state)))
```

```
## [1] "48 25"
```

```r
unique(table(data$year, data$state))
```

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

```r
summary(data$totfatrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.20   14.38   18.43   18.92   22.77   53.32
```

```r
# merge with state codes
alpha_map <-read.csv("statecodes.csv")
data <- merge(x=data, y=alpha_map, by="state",all.x = TRUE) %>% dplyr::select(-state)
data <- dplyr::rename(data, c("state"="code"))

state_avg <- data %>% group_by(state) %>% summarise(avg_totfatrte=mean(totfatrte), .groups = '

# format options for combined plots
formatting <- theme(plot.title = element_text(hjust = 0.5, size=10),
               axis.title.x = element_text(size = 10),
               axis.title.y = element_text(size = 10),
               legend.title = element_text(size = 10))

tickformat <- theme(legend.position = "bottom", axis.text.y = element_text(size = 8),
                    axis.text.x = element_text(angle=90, vjust = 0.5, hjust=1, size = 7))

# functions for making histograms, boxplots and scatterplots
plotbox <- function (df, ycol, ylab){
          ggplot(data, aes(factor(year), get(ycol))) +
                          geom_boxplot() +
                          labs(y=ylab, x = "Year") +
                          formatting + tickformat +
                          scale_x_discrete(breaks=seq(1980, 2004, 2))}

plothist <- function(df, xcol, xlab){
  ggplot(df, aes(x=get(xcol)))+
      geom_histogram(color="black", fill = "white")+
      labs(y='Count', x =xlab)+
      formatting +
      geom_vline(aes(xintercept=mean(get(xcol))),
            color="blue", linetype="dashed", size=1)
}

plotscatter <- function(df, xcol, xlab){
  ggplot(df, aes(x=get(xcol), y=totfatrte)) + geom_point()+
  geom_smooth(method=lm, se=FALSE) + ggtitle(paste(xcol,' vs tolfatrte'))+
```

```
  xlab(xlab) + ylab('Fatality Per 100,000') + formatting
}

p1 <- plot_usmap(data = state_avg,  values="avg_totfatrte", color = "red") +
  scale_fill_continuous(name="Per 100,000", low="white", high="red") +
  theme(legend.position = "right") +ggtitle("Average Fatality Rate from 1980 to 2004 by State")

p2 <- plothist(data, "totfatrte", "Fatality per 100,000") +  ggtitle("Total Fatality Rate Distri
p3 <- plotbox(data, "totfatrte", "Fatality per 100,000") +  ggtitle("Boxplots of Total Fatality

year_avg <- data %>% group_by(year) %>% summarise(avg_totfatrte=mean(totfatrte))
p4 <- ggplot(data=year_avg, aes(x=year, y=avg_totfatrte)) +
  geom_line()+ ggtitle("Average Fatality Rate by Year") +
  xlab('Year') + ylab('Fatality per 100,000')+ formatting


grid.arrange(p1,p2,p3,p4, nrow = 2)
```

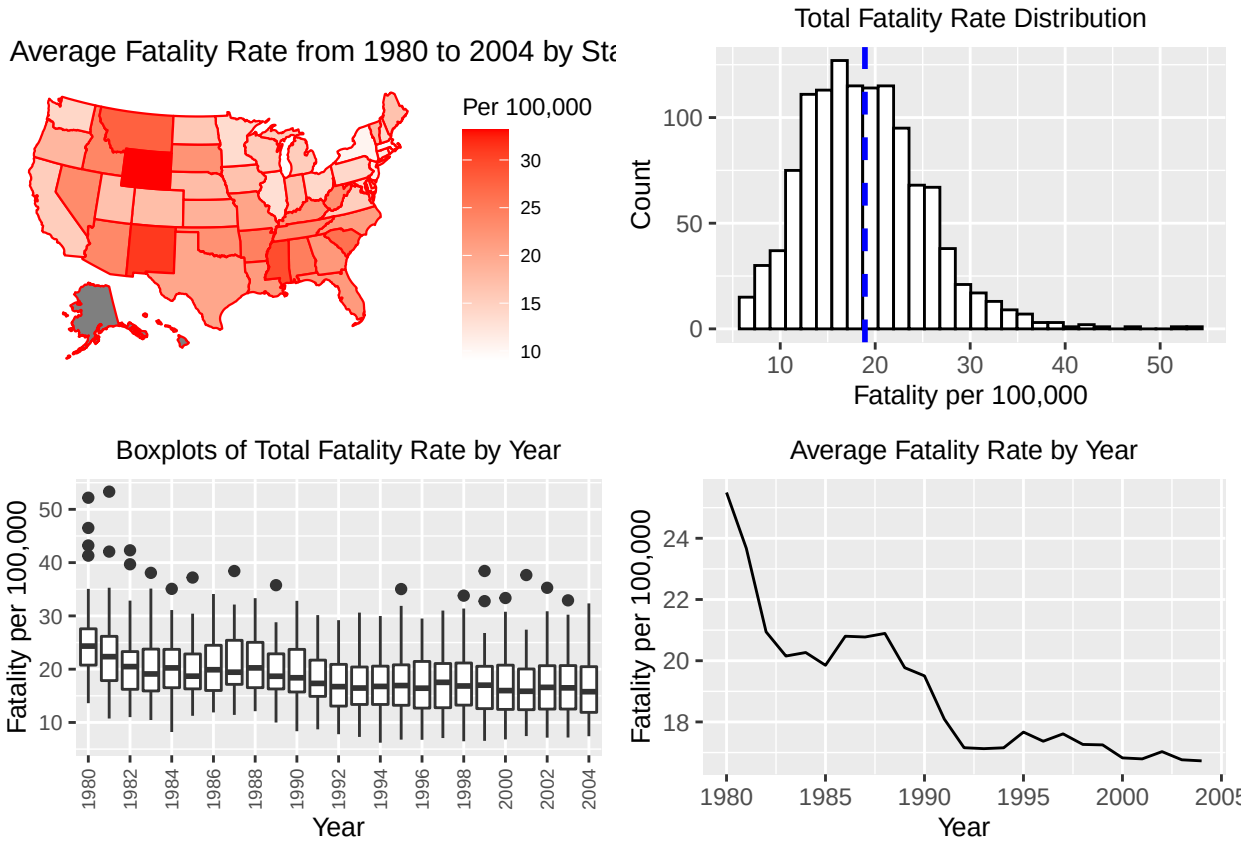

Figure 1: Univariate Analysis of Total Fatality Rate

```
xyplot(totfatrte~year | state, data=data,
       prepanel = function(x, y) prepanel.loess(x, y, family="gaussian"),
```

```
      xlab = "Year", ylab = "Fatality Rate per 100,000 Population",
      panel = function(x, y) {
         panel.xyplot(x, y)
         panel.loess(x,y, family="gaussian") },
      as.table=T)
```

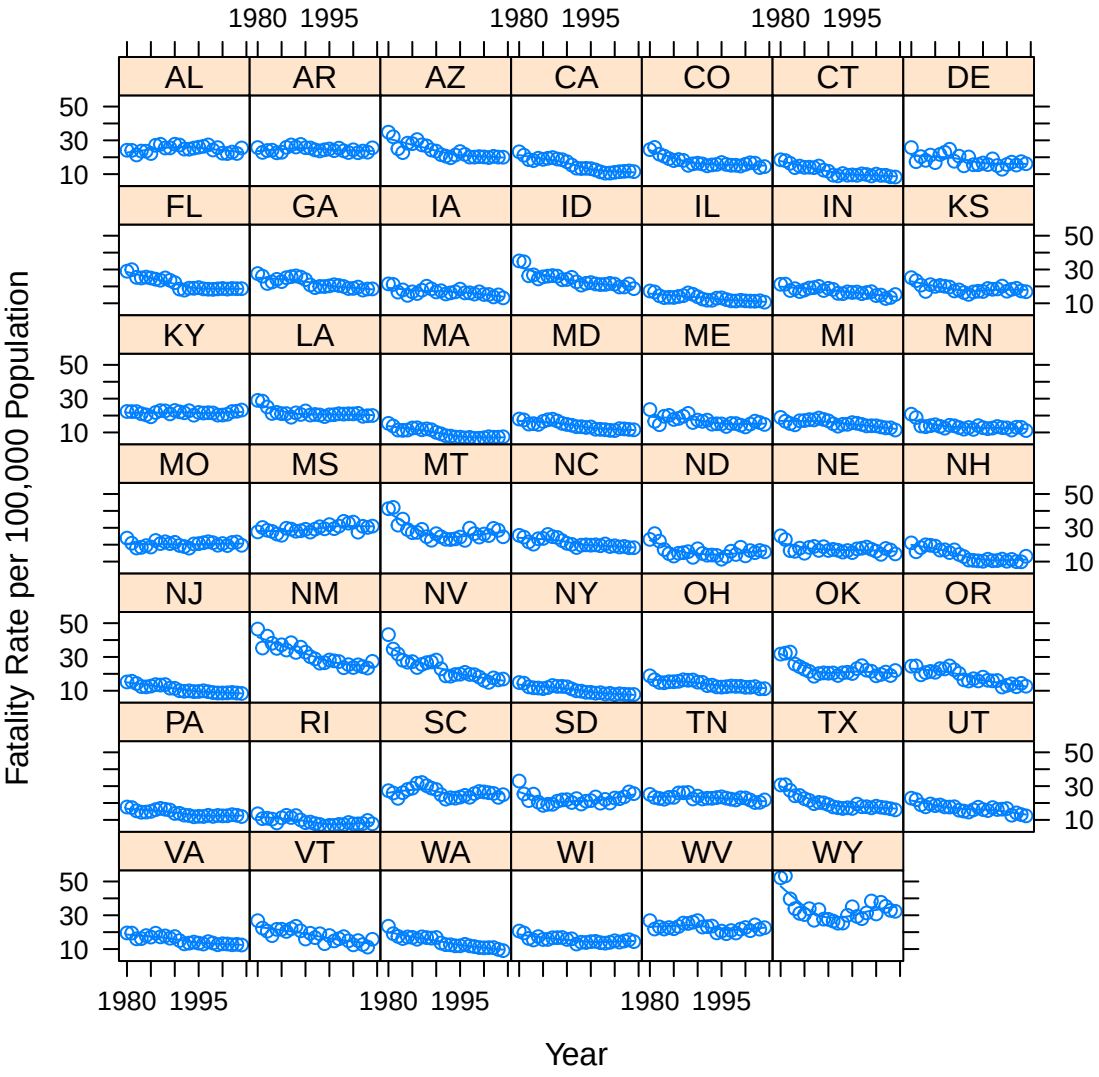

Figure 2: Total Fatality Rate by State

In addition to the indices and fatality measures, the panel data set contains 25 dummies variables representing each year in the dataset. The remaining columns are potential explanatory variables. State population size `statepop` is not pertinent to this study since fatality rate is per 100,000 people. There are four variables not directly related to traffic laws. Minimum drinking age (`minage`) ranges from 18 to 21 in the panel dataset. Starting in 1989, all states require minimal age of 21 to drink, so for most of the data set, `minage` is time invariant, and it is not included in the regression.

Figure 3 shows the univariate and bivariate EDA of the other three variables, `perc14_24`, `unem` and `vehicmilespc`. Percent population aged 14 through 24 (`perc14_24`) ranges from 11.7 to 20.30 with median of 14.9. The percentage decreases steadily until 1990, and then remains steady. Its variance is higher in the 90s, and every years since 1990 has at least one high outlier. Unemployment Rate Percentage (`unem`) ranges from 2.2 to 18% with median of 5.6%. `unem` fluctuates over time with a slight decreasing trend, and its variance reduces over time. Vehicle Miles per Capita (`vehicmilespc`) range from 4372 to 18390 miles with a median of 9013 miles, and it is increasing over time with increasing variance. Summary statistic and histogram suggest all three are asymmetrically distributed with a positive skew, and scatter plots show all three have weakly positive contemporaneous correlation to total fatality, with `perc_14_24` having the strongest correlation among the three.

```
# get class for columns besides year dummy and fatality
data %>% select(matches('^[^d]')) %>% select(-contains("fat")) %>% sapply(class)
```

```
##         year         sl55         sl65         sl70         sl75       slnone
##    "integer"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"
##     seatbelt       minage      zerotol          gdl        bac10        bac08
##    "integer"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"
##        perse     statepop   vehicmiles         unem    perc14_24     sl70plus
##    "numeric"    "integer"    "numeric"    "numeric"    "numeric"    "numeric"
##       sbprim      sbsecon vehicmilespc        state
##    "integer"    "integer"    "numeric"  "character"
```

```
summary(data[, c("minage", "perc14_24", "unem", "vehicmilespc")])
```

```
##      minage       perc14_24          unem         vehicmilespc
##  Min.   :18.0   Min.   :11.70   Min.   : 2.200   Min.   : 4372
##  1st Qu.:21.0   1st Qu.:13.90   1st Qu.: 4.500   1st Qu.: 7788
##  Median :21.0   Median :14.90   Median : 5.600   Median : 9013
##  Mean   :20.6   Mean   :15.33   Mean   : 5.951   Mean   : 9129
##  3rd Qu.:21.0   3rd Qu.:16.60   3rd Qu.: 7.000   3rd Qu.:10327
##  Max.   :21.0   Max.   :20.30   Max.   :18.000   Max.   :18390
```

```
data %>% group_by(year) %>% summarise(avg_min_age=mean(minage))
```

```
## # A tibble: 25 x 2
##     year avg_min_age
##    <int>       <dbl>
##  1  1980        19.2
##  2  1981        19.2
##  3  1982        19.3
##  4  1983        19.4
##  5  1984        19.6
##  6  1985        20.0
##  7  1986        20.5
##  8  1987        20.8
##  9  1988        21.0
## 10  1989        21
## # ... with 15 more rows
```

```
p1a <- plotbox(data, "perc14_24", "Population Age 14-24(%)")
p2a <- plotbox(data, "unem", "Unemployment (%)")
p3a <- plotbox(data, "vehicmilespc", "Miles Per Capita")

p1b <- plothist(data, "perc14_24", "Population Age 14-24(%)")
p2b <- plothist(data, "unem", "Unemployment (%)")
p3b <- plothist(data, "vehicmilespc", "Miles Per Capita")

p1c <- plotscatter(data, "perc14_24", "Population Age 14-24(%)")
p2c <- plotscatter(data, "unem", "Unemployment (%)")
p3c <- plotscatter(data, "vehicmilespc", "Miles Per Capita")

grid.arrange(p1a, p1b, p1c, p2a, p2b, p2c, p3a, p3b, p3c, ncol = 3)
```

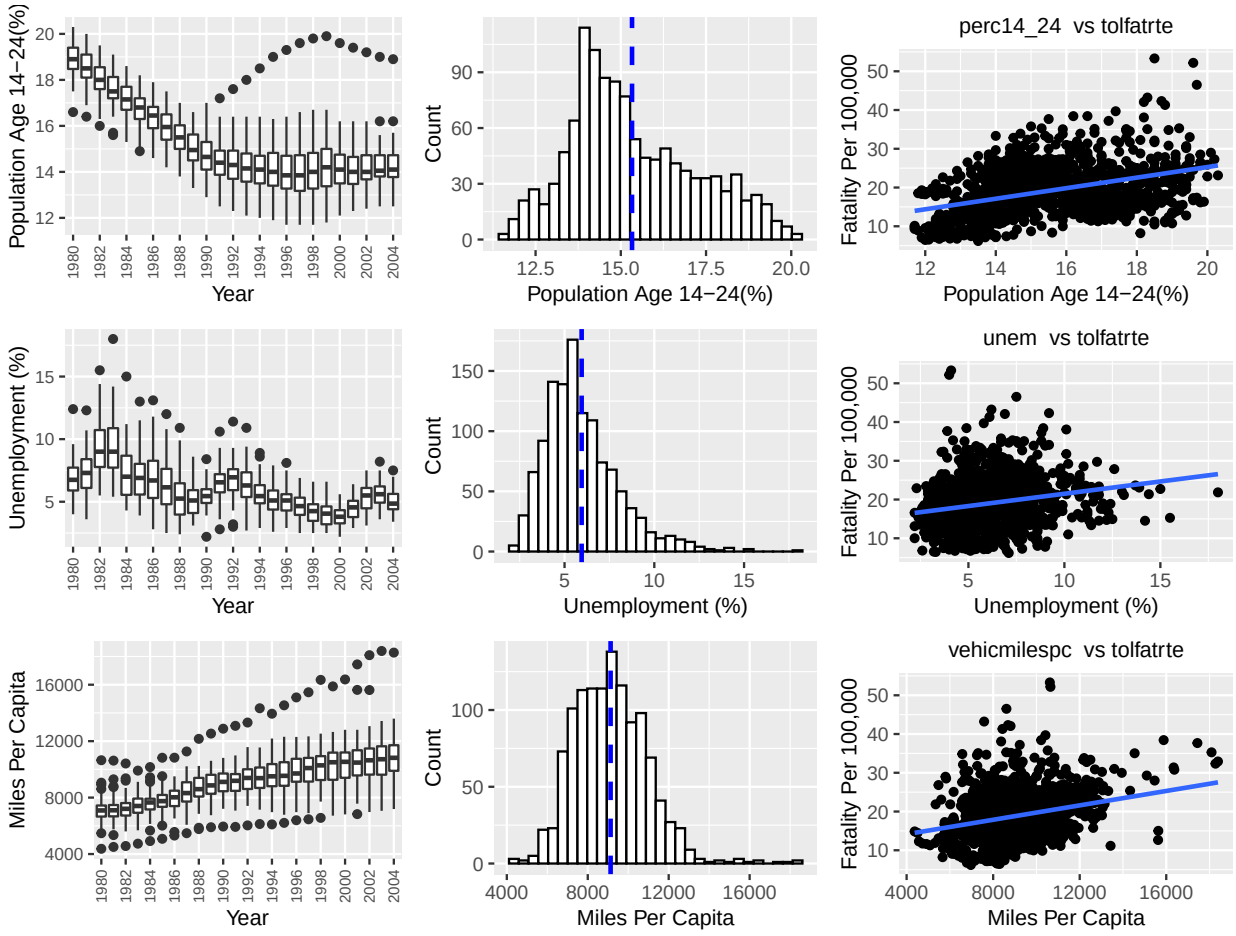

Figure 3: Univariate and Bivariate EDA of Factors Unrelated to Traffic Laws

There are six variables corresponding to speed limit laws, one each for 55, 65, 70, 75 mph limit, one for no speed limit, and one for speed limit 70 and over or no limit. These are not binary indicator variables. They contain decimal values to indicate proportion of the year when a law is in effect in the event a law is enabled in the middle of a year. Figure 4 shows the number of states enforcing each speed limit over time, with yearly averaged fatality rate superimposed. When there are two

different speed limits in a year, the speed limit in-effect for majority of the year is considered the speed limit for the purpose of the plot. In the special case where it's a 50-50 split, the higher speed limit is used. In general, speed limit is increasing over time across the US starting in the late 80s. This is possibly due to the cars in generally getting faster and roads getting better. It's worth noting that the small peak in fatality rate in the late 80s coincides with the first two years of speed limit increase. It's possible that driving was more dangerous in the initial years while people adapted to the faster speed limit. For the regression, speed limit over 70 mph or no limit (`sl70plus`) is used as the explanatory variable.

```
## generic plot function for stacked charts
genplot <- function(grouped_df, legend){
  names(grouped_df) <- c("Year", "condition", "n")
  bar.plot <- ggplot()+ geom_bar(data = grouped_df, aes(fill=condition, y=n, x=Year),
              position="stack", stat = "identity") + geom_line(data = year_avg,
              aes(x = year, y = avg_totfatrte * 1.5), colour = "blue") + formatting +
              scale_y_continuous(sec.axis = sec_axis(trans = ~ . / 1.5, name = "Average Fatali
              guides(fill=guide_legend(title=legend)) +ylab("Number of States")
  return (bar.plot)
}

data <- data %>%
    mutate(sl = case_when(
      sl55 > 0.5 ~ "sl55",
      sl65 > 0.5 ~ "sl65",
      sl70 > 0.5 ~ "sl70",
      sl75 > 0.5 ~ "sl75",
      slnone > 0.5 ~ "slnone",
      sl55 == 0.5 & sl65 == 0.5 ~ "sl65",
      sl65 == 0.5 & sl70== 0.5 ~ "sl70",
      sl65 == 0.5 & sl75== 0.5 ~ "sl75"
    ))


data %>% group_by(year, sl) %>% tally() %>% genplot("Speed Limit")
```

Laws pertaining alcohol are blood alcohol limit at .10 (`bac10`), blood alcohol limit at .08 (`bac8`), and zero tolerance (`zerotol`). `bac10` and `bac08` are rounded off to binary values depending on which law is in effect majority of the year and combined for the barchart. In case of a tie, the data point is considered having BAC limit of 0.1 in effect. Initially, majority of the states did not have a blood alcohol restriction. As seen in Figure **??**, starting in mid 80s, most states adopted a blood alcohol limit of 0.1 or lower, and in the later years, increasing number of states adopted the more restrictive limit of 0.08. The initial increase in BAC restriction correlates to the initial sharp drop in fatality rate in the early 80s. The increase in imposing BAC limit of 0.08 loosely corresponds to the second drop in fatality rate in the late 80s and early 90s. The zero-tolerance law was non-existent in the beginning of the dataset. It was first introduced in 1983, and the number of states implementing the law dramatically increased in the 90s. The increase of zero-tolerance is correlated to the decrease of fatality rate.

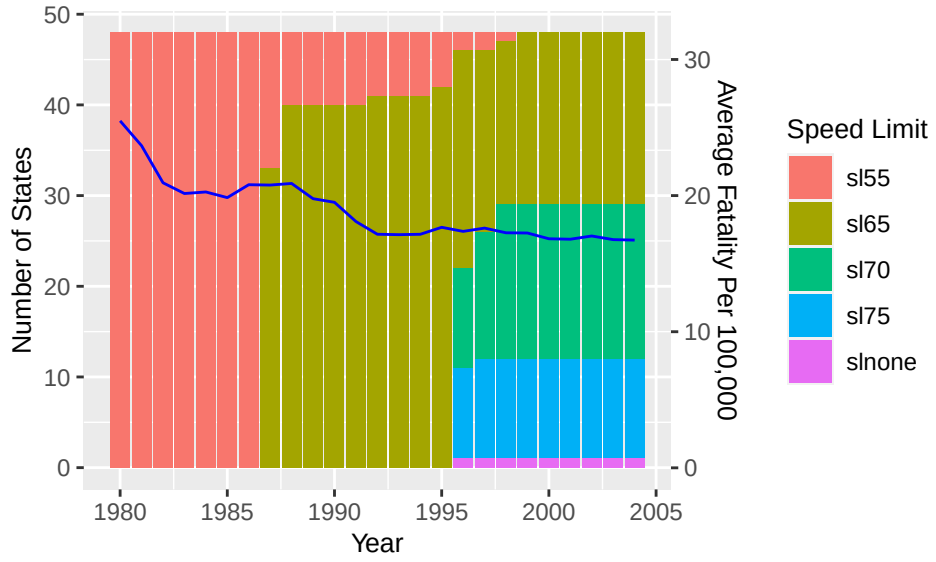

Figure 4: Speed Limit Distribution Over Time with Total Fatality Rate Superimposed

```
data <- data %>%
    mutate(bac = case_when(
      bac10 > 0.5 ~ "BAC 0.1",
      bac08 > 0.5 ~ "BAC 0.08",
      bac08 == 0.5 & bac10== 0.5 ~ "BAC 0.1",
    ), zerotol_bin = case_when(zerotol > 0.5 ~ "Enforced")
    )

p1 <- data[!is.na(data$bac),] %>% group_by(year, bac) %>% tally() %>% genplot("Blood Alcohol")
p2 <- data[!is.na(data$zerotol_bin),]%>% group_by(year, zerotol_bin) %>% tally() %>% genplot("

grid.arrange(p1,p2, nrow = 1)
```

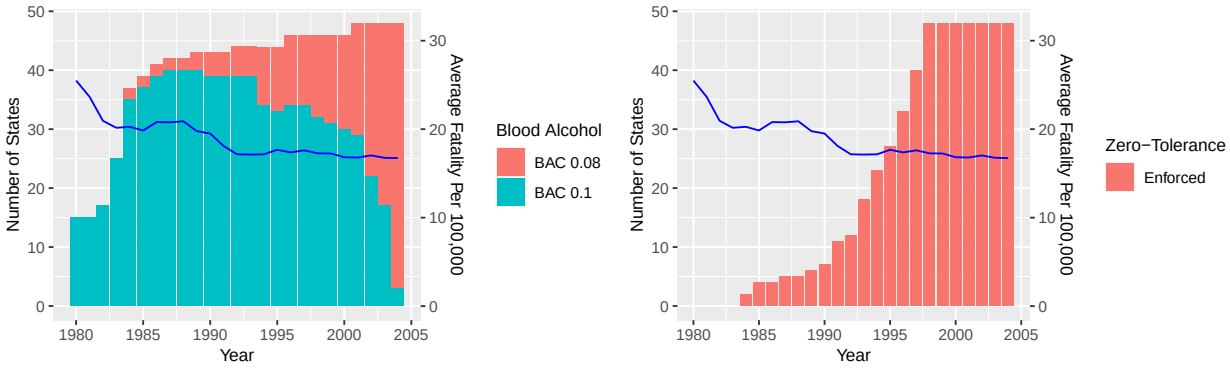

Figure 5: Enforcement of Laws Regarding Alcohol with Total Fatality Rate Superimposed

`seatbelt` indicates seatbelt laws, possible values are `primary`, `secondary`, or `none`. `sbprim` and `sbsecond` are binary indicator variables representing the same information. Other variables concern-

ing traffic laws are graduated drivers license law (`gdl`), administrative license revocation (`perse`). Similar to the speed limit variables, `gdl` and `perse` also contain decimals to represent laws in effect for parts of a year. They were also binarized for the bargraph visualizations (see Figure 6). Seat belt laws were first implemented in 1985 and saw near total adoption by 1995. Increasing number of states adopted primary seatbelt laws starting mid 90s. Overall, seatbelt laws is inversely correlated to fatality rate, though due to its absence in early 80s, it did not contribute to the initial decrease in fatality. Graduated Driver License Law first began in 1996 and dramatically increased in enforcement over the next decade. By mid 90s, the fatality rate was already steady, so `gdl` does not have an obvious impact on fatality. Per Se Law became increasingly common starting early 80s, and is inversely correlated to fatality rate.

```
data <- data %>%
    mutate(seatbelt_bin = case_when(seatbelt == 1 ~ "Primary",seatbelt == 2 ~ "Secondary")
      , gdl_bin =  case_when(gdl > 0.5 ~ "Enforced"),
    perse_bin = case_when(perse > 0.5 ~ "Enforced"))

p1 <- data[!is.na(data$seatbelt_bin),] %>% group_by(year, as.factor(seatbelt_bin)) %>% tally()
p2 <- data[!is.na(data$gdl_bin),] %>% group_by(year, gdl_bin) %>% tally() %>% genplot("Graduate
p3 <- data[!is.na(data$perse_bin),] %>% group_by(year, perse_bin) %>% tally() %>% genplot("Pers

grid.arrange(p1,p2,p3, nrow = 1)
```

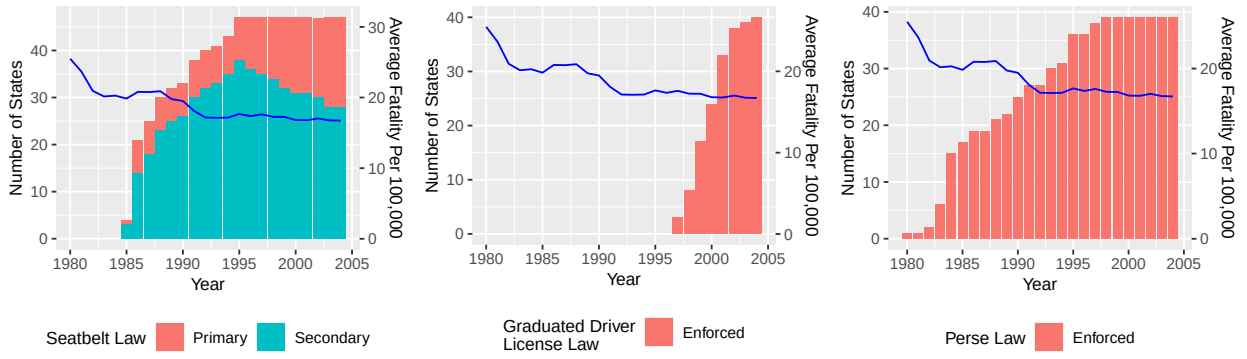

Figure 6: Enforcement of Traffic Laws with Total Fatality Rate Superimposed

Finally, correlation plot (Figure 7) provides a cursory look at the relationship between the predictors and the outcome variable `totfatrte`. None of the predictors have exceptionally high correlation with fatality rate. `sl70plus`, `vehicmilespc`, `perc14_24`, `unem` have positive contemporaneous correlation to the outcome variable, and the rest have negative correlation. As previously noted in the bivariate scatterplot (Figure 3), `perc14_24` has the strongest correlation to `totfatrte`. Among the predictors, in general the traffic law variables correlate positively to each other. `bac08` and `bac10` have strong negative correlation, which is expected, since a state can have only one of these two laws in effect at a time. `unem` has a notable negative correlation to `vehicmillespc`, which also makes intuitive sense, since higher unemployment rate would mean less commuting for work. Another well-correlated pair of predictors are `vehicmilespc` and `sl70plus`, implying that people travel more for places with less restrictive speed limit. Although `perse` is well correlated to `vehicmilespc`, this may be a coincidence since perse laws were increasingly enforced, while per capita miles travels were also increasing in time. None of the predictor variables of interest have perfect correlation so the

lack of perfect correlation assumption for linear regression is satisfied.

```
library(corrplot)
res2 <- cor(data[, c('totfatrte', 'bac08', 'bac10', 'perse', 'sbprim', 'sbsecon', 'sl70plus',
col<- colorRampPalette(c("blue", "white", "red"))(20)
corrplot(res2, type = 'lower', order = "hclust",  addCoef.col = "black",
         tl.col = "black", tl.srt = 45, tl.cex=0.8, number.cex = 0.8)
```

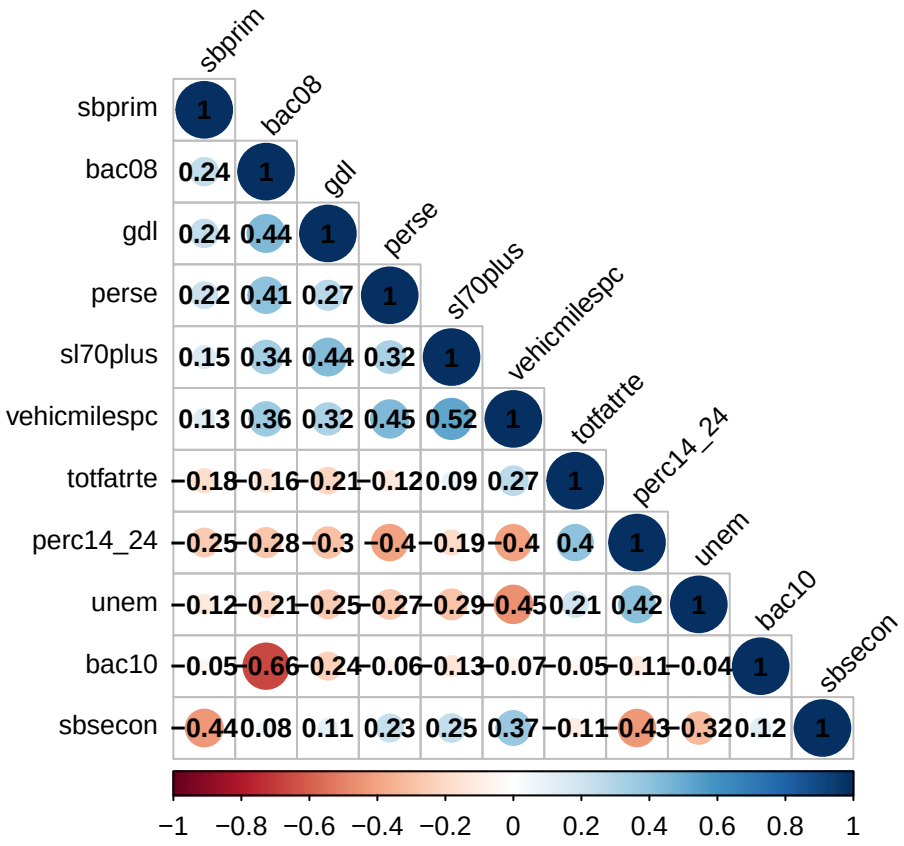

Figure 7: Correlation Matrix of Predictor Variables and Fatality Rate

Part 2

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 *totfatrte* is defined as the total fatality per 100,000 people. The average in each year can be calculated as a simple average or weighted average after controlling for each state's population size, as shown in Table 1. In 1980, the simple and weighted annual average of `totfatrte` is 25.49 and 22.61, respectively. In comparison, in 2004, the simple and weighted annual average of `totfatrte` is 16.73 and 14.54. In general, the fatality rates are decreasing over time, and the weighted average

is lower than simple average in the same year, because some of the more populated states have lower fatality rate.

```
# totfat averaged factoring population
weighted_avg <- data %>% group_by(year) %>% summarise(avg_totfatrte_weighted = sum(totfat)*1000
year_avg$avg_totfatrte_weighted <- weighted_avg$avg_totfatrte_weighted

kable(year_avg, caption = "\\label{tab:avg_fatality}Yearly Average of Total Fatality Rate")
```

Table 1: Yearly Average of Total Fatality Rate

year	avg_totfatrte	avg_totfatrte_weighted
1980	25.49458	22.61325
1981	23.67021	21.54580
1982	20.94250	19.00720
1983	20.15292	18.23027
1984	20.26750	18.79777
1985	19.85146	18.45940
1986	20.80042	19.26071
1987	20.77479	19.21391
1988	20.89167	19.31130
1989	19.77229	18.51417
1990	19.50521	17.91138
1991	18.09479	16.43934
1992	17.15792	15.32955
1993	17.12771	15.46917
1994	17.15521	15.50936
1995	17.66854	15.74177
1996	17.36938	15.64513
1997	17.61062	15.44655
1998	17.26542	15.08738
1999	17.25042	15.00136
2000	16.82562	14.93093
2001	16.79271	14.81142
2002	17.02958	14.97254
2003	16.76354	14.76937
2004	16.72896	14.54297

A pooled linear regression model was fitted using just the indicator variables for years. In the EDA we noted that `totfatrte` is strongly positively skewed, therefore we log transformed the variable. The follow is the truncated equation summarizing the result of the regression. See Table 2 in the appendix for complete table of coefficients.

$$\log(totfatrte) = 3.20 - 0.079d81 - 0.20d82 - 0.24d83 - 0.226d84 - 0.24d85 \cdots - 0.43d02 - 0.44d03 - 0.45d04$$

This model gives us the time effects on total fatality rate. The intercept in this case is the logged average `totfatrte` across all states in 1980, the baseline year. Each of the coefficients `d81`, `d82`...`d04` is the average change in logged `totfatrte` relative to the base year 1980. The coefficients for

the dummy variable for 1981 is not statistically significant at the 5% level, the rest are all highly significant. Using `d80` as the base level, all coefficients have a negative sign, implying the total fatality rate comparing to 1980 is lower for all years starting 1981. The magnitude of the coefficients are for most part increasing, meaning as time goes on, in general there's an increasingly larger negative difference in fatality rate comparing to 1980.

While total fatality rate is decreasing over the 25 year period, it doesn't necessarily mean driving has become safer. Firstly, driving safety encompasses both fatality rate in accidents, as well as accident rates in general. This dataset does not capture overall accident rates, so it's possible that vehicular accident rates remained the same or even increased over time, but because the newer car models have better safety features, drivers are much less likely to be injured or killed in accidents and hence the drop in fatality rate. Additionally, because the fatality rate per fixed population rate, changes in demographics or lifestyle could indirectly lead to what appears to be decreasing fatality rate. For example, most major metropolises are growing in population size over time, and people living in the city tend to travel using means other than private vehicles. Along the same veins, in recent years, due to combination of improvement in public transit and environment advocacy, more people are shifting to public transportation. These people would be included in the denominator for traffic fatality rate, while not contributing much to the numerator, which would lower the total fatality rate as defined in this dataset.

```
q2.lm <- lm(log(totfatrte)~d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+d91+d92+d93+d94+d95+d96+d97+
q2.lm.se = sqrt(diag(vcovHC(q2.lm)))
```

Part 3

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.)

For modeling purposes, we chose to not binarize the variables representing enactment of laws (`bac08`, `bac10`, `perse`, `sbprim`, `sbsecon`, `sl70plus`, `gdl`). From an intuitive perspective, if a law has effect on traffic fatality, implementing it middle of the year should result in that year's fatality rate averaging out to be somewhere in between if the law was in full effect the whole year and entirely not enacted the whole time. Binarizing the predictors would lose this meaningful relationship. By leaving the variables as decimals, the variables can be interpreted as the fraction of year when the law is effective, instead of an indicator variable representing simple presence or absence of a law.

As for the numerical variables, as seen in the EDA histogram in (Figure 3), `perc14_24` is asymmetric but not strongly skewed, therefore a log transformation is not necessary. In contrast, `unem`, `vehicmilespc` as well as the outcome variable `totfatrte` were shown in the EDA to be obviously skewed in the positive direction, and `vehicmilespc` has increasing variance over time. This makes them good candidates for log transformation. Additionally, Shapiro-Wilk normality test shows that the linear model with untransformed variable would result in non-normal residuals (P = 9.96e-13), while residuals after transforming `vehicmilespc` and `totfatrte` is normally distributed (p = 0.14).

Figure 8 displays the diagnostic plots of the untransformed linear model. Comparing to that of the transformed model (See Figure 9 for diagnostic models), the residuals of the untransformed linear model non-normal, heteroskedastic, and violates zero conditional mean. For these reasons, log-transformation of `totfatre`, `unem`, and `vehicmilespc` is preferable.

The following is the truncated equation summarizing the result of the regression. See Table 2 in the appendix for complete table of coefficients.

$$\begin{aligned}\log(totfatrte) = &-11.246 - 0.092d81 - 0.29d82 \cdots - 1.00d03 - 0.98d04 - 0.063bac08 - 0.018bac10 \\ &- 0.020perse + 0.0004sbprim + 0.020sbsecon + 0.232sl70plus - 0.027gdl + 0.017perc14\_24 \\ &+ 0.26\log(unem) + 1.54\log(vehicmilespc)\end{aligned}$$

Breusch-Pagan test suggests the model using transformed variables still violates heteroskedascity. The coefficients for `bac10` , `perse`, `sbprim`, `sbsecon` and `gdl` are not significant when using heteroskedasticity robust standard error. The rest of the coefficients are significant at the 5% level. `bac8` and `bac10` indicate the proportion of the year when blood alcohol limit is at 0.08 and 0.1, respectively. Holding all other factors constant, in any given year, enforcing the legal BAC limit at 0.08 for the entire year is associated with about 6.3% decrease in fatality rate; enforcing the BAC limit at 0.1 is associated with 1.8% decrease in fatality rate, though this decrease is not statistically significantly different from zero.

The signs for the coefficients of `perse` and `sbprim` are both negative, but neither is significant at 5% level, so even though the regression shows per se laws and primary seat belt law both have a negative effect on fatality rate, the effect may not be significantly different from zero.

```
q3.untransformed <- lm(totfatrte~factor(year)+ bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+pe

shapiro.test(residuals(q3.untransformed))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(q3.untransformed)
## W = 0.97747, p-value = 9.96e-13
```

```
par(mfrow=c(2,2))

plot(q3.untransformed)
```

```
q3.lm <- lm(log(totfatrte)~d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+d91+d92+d93+d94+d95+d96+d97+
              bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+log(unem)+log(vehicmiles

q3.lm.se = sqrt(diag(vcovHC(q3.lm)))
shapiro.test(residuals(q3.lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(q3.lm)
```

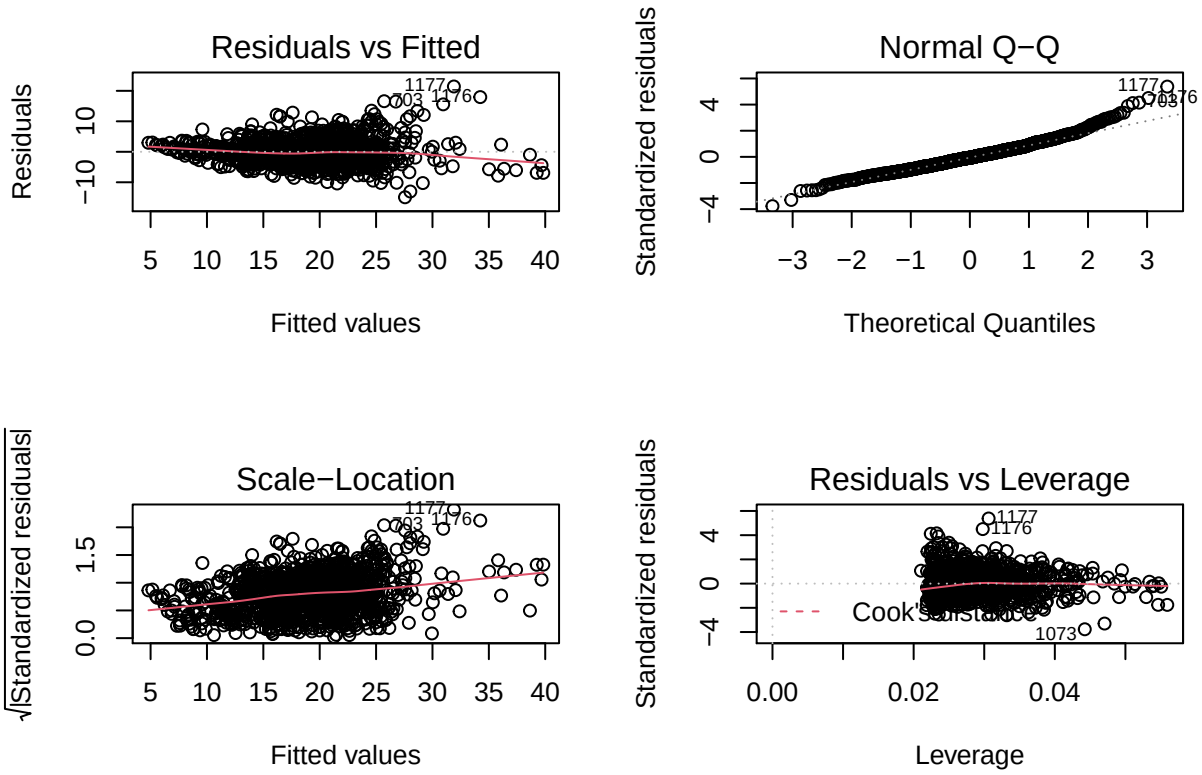

Figure 8: Diagnostic Plots of the Pooled OLS Model Without Transformation

```
## W = 0.99793, p-value = 0.141
```

```
bptest(q3.lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  q3.lm
## BP = 98.34, df = 34, p-value = 3.61e-08
```

```
par(mfrow=c(2,2))
plot(q3.lm)
```

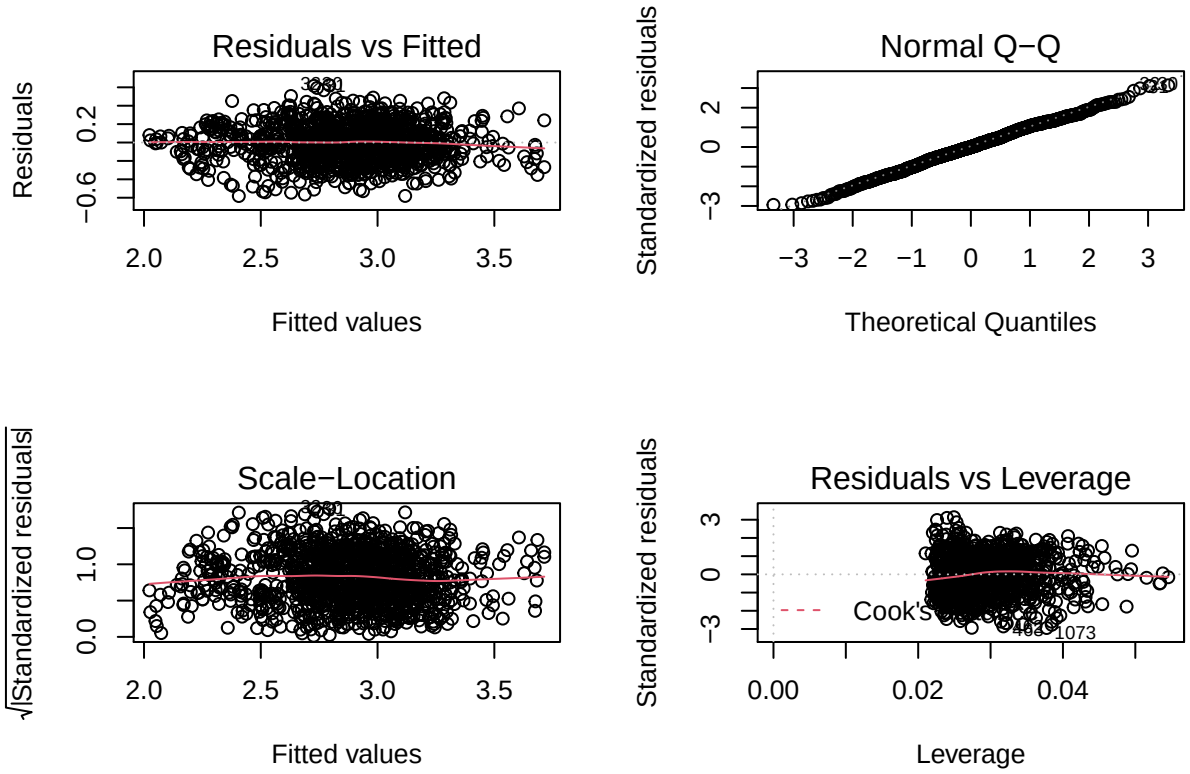

Figure 9: Diagnostic Plots of the Pooled OLS Model

Part 4

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?

The fixed effect model takes the form

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{it2} \cdots + \beta_k \ddot{x}_{itk} + \ddot{u}_{it}, t = 1, 2, \cdots, T$$

See table in Table 2 in the appendix for complete table of coefficients with comparison to the model in Exercise 3.

In the fixed effects model, the coefficients for `bac8`, `bac10`, `sbprim`, `sbsecon` and `gdl` are not significant when using heteroskedasticity robust SE. The coefficient for `perc14_24` is marginally significant at the 10% level. The rest of the coefficients are significant at the 5% level.

Comparing to the pooled OLS model, in the fixed effect model, the coefficient for `bac08` became smaller and non-significant; `perse` became highly significant with coefficient more than twice as large; `bac10` and `sbprim` remain non-significant as they are in the pooled OLS model.

According to the fixed effects model, the traffic laws that has the largest effects on traffic fatalities are the Graduated Driver License Law (Per Se Law) and higher speed limits. Hold all other factors constant, by enforcing the Per Se Law, traffic fatality rate is expected to decrease by approximate 5.9%. Conversely, by allowing a high speed limit (70mph or higher) or not enforcing speed limit at all, traffic fatality is expected to increase by 7.8%.

The most important drawback to using pooled OLS is that it assumes the unobserved effect a_i is uncorrelated with any of the predictor variables at all times. If this assumption is not true or if the idiosyncratic error is correlated to the predictors, then pooled OLS is biased and inconsistent. Additionally, the pooled OLS estimates requires the six classical linear regression assumptions. CLM1 (linearity of parameters) is met by definition. It's unknown whether CLM2 (random sampling) is true due to the unknown sampling methods. CLM3 (no perfect linear relationship) was checked using the correlation matrix. CLM4 (Zero Conditional Mean or Exogeneity) was checked using diagnostic plots. The residual vs. fitted plot for the linear model after transformation shows the residuals to be roughly symmetrical around 0, with no strong patterns throughout the fitted values. CLM5 (homoskedascity) was checked using the Breusch–Pagan test, which suggests significant heteroskedascity. This was addressed by using heteroskedascity-robust standard errors. CLM6 (normality of residuals) was checked using the Shapiro-Wilk test and qq-plot, which suggests the model with transformation upholds the assumption.

Unlike the pooled OLS model, the fixed effect model allows for arbitrary correlation between the unobserved effect and the explanatory variables in any time period, and only requires the idiosyncratic error to be uncorrelated to the predictors. The other assumptions are similar to that of the pooled OLS model. QQ-plot of the residual shows rough normality (Figure 10). The residual vs fitted plot shows the idiosyncratic error to have approximately mean of zero, so exogeneity assumption is likely met. Absence of perfect linear relationship was confirmed through correlation matrix, and none of the explanatory variables included in the model are time invariant. The residuals for the fixed effect model is heteroskedastic according to Breusch–Pagan test (p = 1.81e-07). This issue is addressed by using heteroskedastic robust coefficients. Lastly, the model assumes no serial correlation between the idiosyncratic errors conditional on all explanatory variables and unobserved effects. This is the most problematic assumption this specific model. The Breusch-Godfrey test suggested there is serial correlation in idiosyncratic errors (p <2.2e-16). In conjunction to the violation of heteroskedasticity, we could potentially address this by using heteroskedasticity and autocorrelation consistent standard error, or include lag terms in the regression to attempt to remove the serial correlation in the error.

Taking in account of the shortcomings of both models, the estimates from the Fixed-Effect model are more reliable. The assumption of no correlation between composite error $(a_i + u_{it})$ is unlikely to be true for the pool OLS to give unbiased estimates. Intuitive example of such violation is a state's geographical features and city layouts. The layout of cities in each state is mostly time invariant, and it would have a tangible impact on one of the predictors, the amount of traveling by vehicle per capita. Simultaneously, it would have a direct effect on vehicle accidental fatality rate. States with closely packed cities and narrow streets are likely to end up with higher accidental rate and higher fatalities.

```
panel_data <- pdata.frame(data, c("state", "year"))
q4.fe <- plm(log(totfatrte) ~ d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+d91+d92+d93+d94+d95+d96+d
               bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + log(uner
q4.fe.se = sqrt(diag(vcovHC(q4.fe)))

bptest(q4.fe, studentize = F)
```

```
##
##  Breusch-Pagan test
##
## data:  q4.fe
## BP = 93.568, df = 34, p-value = 1.811e-07
```

```
pbgtest(q4.fe)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  log(totfatrte) ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +     d89 + d90 + d91
## chisq = 219.61, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
par(mfrow=c(1,2 ))

Fitted.Values <- as.numeric(q4.fe$model[[1]] - q4.fe$residuals)
Residual <- as.numeric(q4.fe$residuals)

scatter.smooth(Fitted.Values, Residual, lpars = list(col = "red", lwd = 1, lty = 1))

qqPlot(Residual)
```

```
## [1] 1098 1099
```

Part 5

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

The random effects model is reported in Table 2 in the appendix.

The fixed effects model is more preferred over the random effects model. Hausman test also rejects the null-hypothesis that the random effect model is consistent (p = 1.68e-05). The main issue with

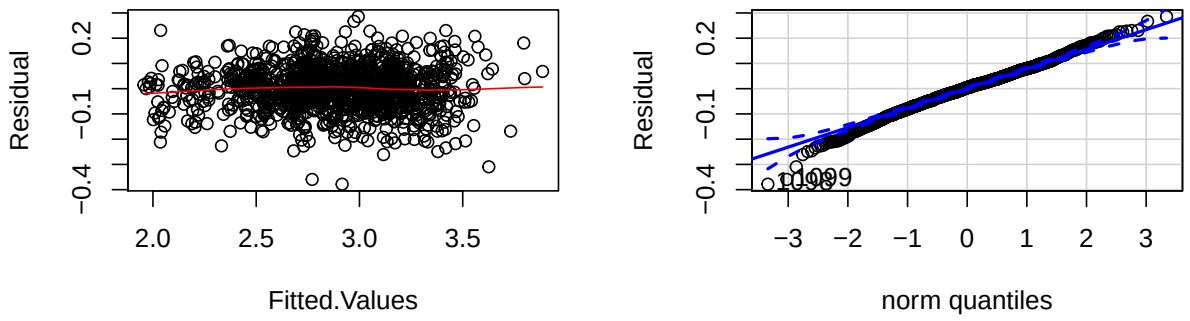

Figure 10: Residual vs. Fitted and QQ-plot for Fixed Effects Model

the random effect model is that it assumes the unobserved effect given all explanatory variables is constant, that is, there is no correlation between the unobserved effect and the explanatory variables. As discussed previously, this assumption is unlikely to hold. On the other hand, because we are not using any time-invariant variables as predictors, fixed effect model can estimate the effects of all predictors on `totfatrte`. It's worth noting that all coefficients significant in the fixed effect model are also significant in the random effect model, all coefficients not significant in the FE model are also non-significant in RE model, and the signs of the coefficients remain the same between the two models.

```
q5.re <-plm(log(totfatrte) ~ d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+d91+d92+d93+d94+d95+d96+d9
               bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + log(uner
q5.re.se = sqrt(diag(vcovHC(q5.re)))

phtest(q4.fe, q5.re)
```

```
##
##  Hausman Test
##
## data:  log(totfatrte) ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +  ...
## chisq = 79.459, df = 34, p-value = 1.675e-05
## alternative hypothesis: one model is inconsistent
```

Part 6

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.

Due to the log transformation of `totfatrte` and `vehicmilespc`, there isn't a constant estimated effects from a raw increase of miles driven per capita. Holding all other factors constant, if `vehicmilespc` is originally 10,000 miles, increasing it by 1,000 miles is a 10% increase, then according to the Fixed Effect model's coefficient for `log(vehicmilespc)`, it would result in a 6.76% increase in total fatality rate. In contrast, if `vehicmilespc` was originally 5,000 or 20,000, increasing

it by 1,000 miles would lead to 13.52% and 3.38% increase in `totfatrte`, respectively.

Part 7

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?

If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, the errors would be closer together and the standard errors are smaller than they should be. Consequently, the p-values obtained would be smaller than it should be and it would be easier to obtain a significant coefficient comparing to how significant it is in reality. One way to address this is to use clustering to obtain fully robust standard errors and test statistics. As discussed in part 4, the fixed effect model in this analysis has both serial correlation and heteroskedasticity in the residuals according to Breusch-Godfrey and Breusch-Pagan test, respectively. Without using heteroskedasticity and autocorrelation consistent standard errors, we should consider the possibility that the marginally significant coefficient (`perc14_24`) is in reality not significant.

Appendix

Regression Model Results from Questions 2-5, using heteroskedastic robust standard error.

```
stargazer(q2.lm, q3.lm, q4.fe, q5.re, type = "latex", single.row = TRUE, font.size="small",
          se = list(q2.lm.se, q3.lm.se, q4.fe.se, q5.re.se), column.sep.width = "-15pt",
          column.labels = c("Pooled OLS 1", "Pooled OLS 2", "Fixed Effect", "Random Effect"),
          star.char = c("+", "*", "**", "***"), star.cutoffs = c(0.1, 0.05, 0.01, 0.001),
          notes = c("+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001"), header = FALSE)
```

Table 2:

	Dependent variable:			
	log(totfatrte)			
	OLS		*panel linear*	
	Pooled OLS 1	Pooled OLS 2	Fixed Effect	Random Effect
	(1)	(2)	(3)	(4)
d81	−0.079 (0.061)	−0.092* (0.046)	−0.063*** (0.017)	−0.064*** (0.017)
d82	−0.200** (0.061)	−0.294*** (0.046)	−0.136*** (0.018)	−0.143*** (0.018)
d83	−0.235*** (0.060)	−0.348*** (0.043)	−0.169*** (0.022)	−0.177*** (0.022)
d84	−0.226*** (0.060)	−0.298*** (0.044)	−0.206*** (0.023)	−0.212*** (0.023)
d85	−0.243*** (0.058)	−0.336*** (0.045)	−0.231*** (0.028)	−0.238*** (0.027)
d86	−0.197*** (0.058)	−0.312*** (0.049)	−0.194*** (0.035)	−0.202*** (0.035)
d87	−0.199*** (0.058)	−0.349*** (0.050)	−0.240*** (0.040)	−0.250*** (0.040)
d88	−0.189*** (0.057)	−0.359*** (0.052)	−0.271*** (0.049)	−0.282*** (0.048)
d89	−0.248*** (0.058)	−0.445*** (0.056)	−0.345*** (0.054)	−0.357*** (0.054)
d90	−0.268*** (0.061)	−0.504*** (0.061)	−0.355*** (0.060)	−0.370*** (0.059)
d91	−0.344*** (0.061)	−0.619*** (0.062)	−0.391*** (0.064)	−0.410*** (0.063)
d92	−0.402*** (0.063)	−0.725*** (0.065)	−0.452*** (0.067)	−0.474*** (0.067)
d93	−0.403*** (0.063)	−0.716*** (0.064)	−0.469*** (0.068)	−0.490*** (0.068)
d94	−0.408*** (0.066)	−0.703*** (0.064)	−0.503*** (0.068)	−0.523*** (0.068)
d95	−0.385*** (0.068)	−0.683*** (0.066)	−0.503*** (0.073)	−0.522*** (0.073)
d96	−0.399*** (0.067)	−0.805*** (0.066)	−0.554*** (0.076)	−0.577*** (0.075)
d97	−0.386*** (0.067)	−0.825*** (0.067)	−0.581*** (0.078)	−0.603*** (0.077)
d98	−0.410*** (0.068)	−0.868*** (0.068)	−0.633*** (0.078)	−0.656*** (0.078)
d99	−0.414*** (0.069)	−0.869*** (0.067)	−0.651*** (0.081)	−0.674*** (0.080)
d00	−0.437*** (0.068)	−0.881*** (0.070)	−0.684*** (0.081)	−0.706*** (0.080)
d01	−0.435*** (0.067)	−0.932*** (0.071)	−0.653*** (0.085)	−0.679*** (0.083)
d02	−0.427*** (0.069)	−0.976*** (0.073)	−0.615*** (0.083)	−0.645*** (0.082)
d03	−0.440*** (0.068)	−0.996*** (0.074)	−0.618*** (0.086)	−0.649*** (0.084)
d04	−0.449*** (0.070)	−0.982*** (0.076)	−0.656*** (0.089)	−0.686*** (0.088)
bac08		−0.063* (0.028)	−0.015 (0.032)	−0.019 (0.033)
bac10		−0.018 (0.021)	−0.014 (0.020)	−0.016 (0.020)
perse		−0.020 (0.016)	−0.059*** (0.017)	−0.057** (0.018)
sbprim		0.0004 (0.025)	−0.040 (0.025)	−0.038 (0.025)
sbsecon		0.020 (0.023)	0.006 (0.016)	0.007 (0.016)
sl70plus		0.232*** (0.022)	0.078** (0.024)	0.082*** (0.023)
gdl		−0.027 (0.026)	−0.022 (0.021)	−0.022 (0.021)
perc14_24		0.017** (0.007)	0.019+ (0.011)	0.020+ (0.011)
log(unem)		0.264*** (0.024)	−0.192*** (0.023)	−0.173*** (0.024)
log(vehicmilespc)		1.537*** (0.049)	0.676*** (0.136)	0.759*** (0.128)
Constant	3.196*** (0.042)	−11.246*** (0.451)		−3.567** (1.155)
Observations	1,200	1,200	1,200	1,200
R^2	0.126	0.669	0.729	0.714
Adjusted R^2	0.108	0.660	0.710	0.705
Residual Std. Error	0.325 (df = 1175)	0.201 (df = 1165)		
F Statistic	7.057*** (df = 24; 1175)	69.380*** (df = 34; 1165)	88.512*** (df = 34; 1118)	2,901.701***

Note: *p<0.1; **p<0.05; ***p<0.01
+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001