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

U.S. traffic fatalities: 1980-2004

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

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

Exercises:

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`. There are 48 states in the panel data spanning 25 years from 1980-2004. The 48 continental states, represented by their FIPS code, have 25 data points and there is one row of data per state per year. There is no missing data in the dataset. There are nine fatality rate measures, measuring total, weekend, and nighttime fatality count, fatality per 100,000 population, and fatality per 100 million miles. 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 Missipissipi, 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"
```

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

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

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

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

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

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

```
# merge with state codes
fips_map <-read.csv("statecodes.csv")
data <- merge(x=data, y=fips_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 = '
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")


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

p2 <- ggplot(data, aes(x=totfatrte))+ formatting +
      geom_histogram(color="black", fill = "white")+
      ggtitle("Total Fatality Rate Distribution") +
      labs(y='Count', x ='Fatality Per 100,000')+
      theme(plot.title = element_text(hjust = 0.5)) +
      geom_vline(aes(xintercept=mean(totfatrte)),
            color="blue", linetype="dashed", size=1)

p3 <- ggplot(data, aes(factor(year), totfatrte)) +
                  geom_boxplot() +
```

```
                  ggtitle("Boxplots of Total Fatality Rate by Year") +
                  labs(y='Fatality per 100,000', x ='Year') + formatting + tickformat

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)
```

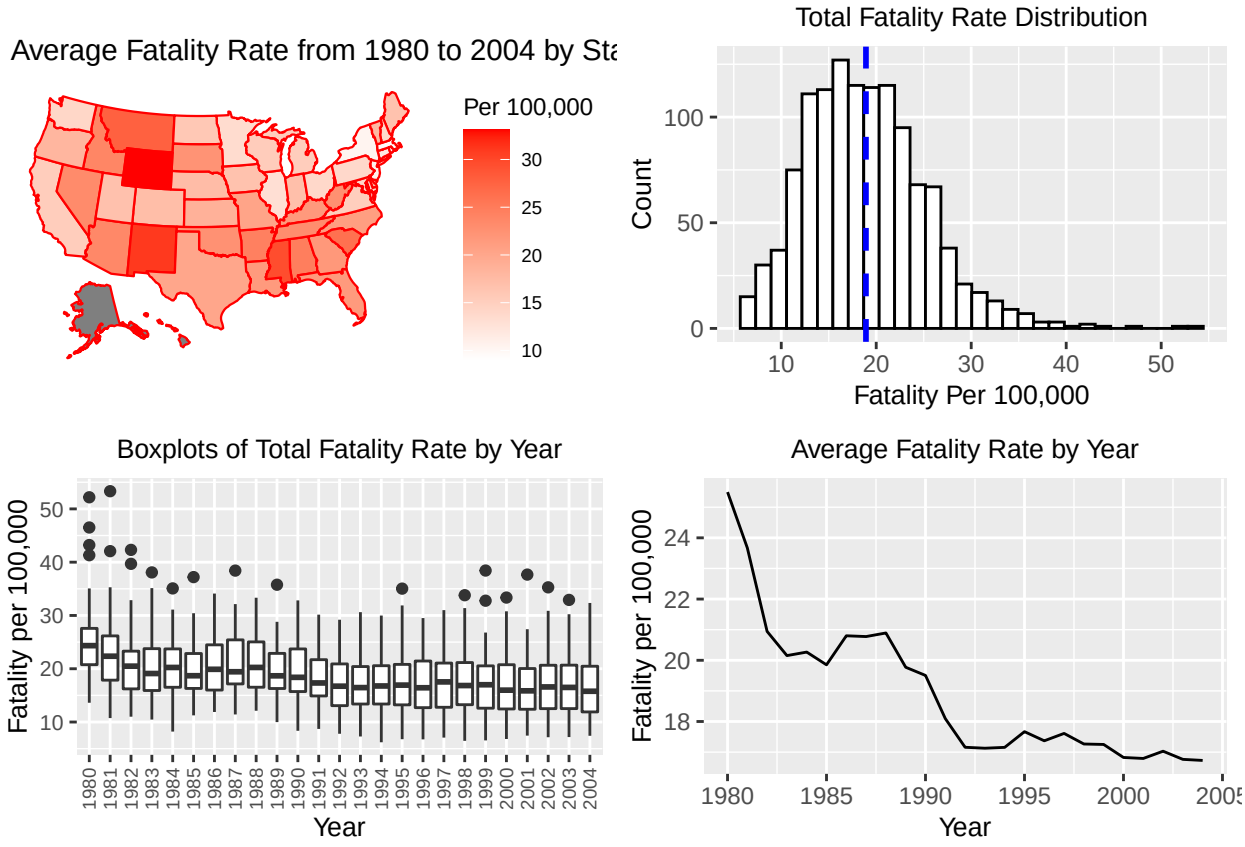


Figure 1: Univariate Analysis of Total Fatality Rate

```
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)
```

In addition to the indicies and fatality measures, the panel data set contains 25 dummies variables representing each year in the dataset. The remaining columns are potential explanatory variables.

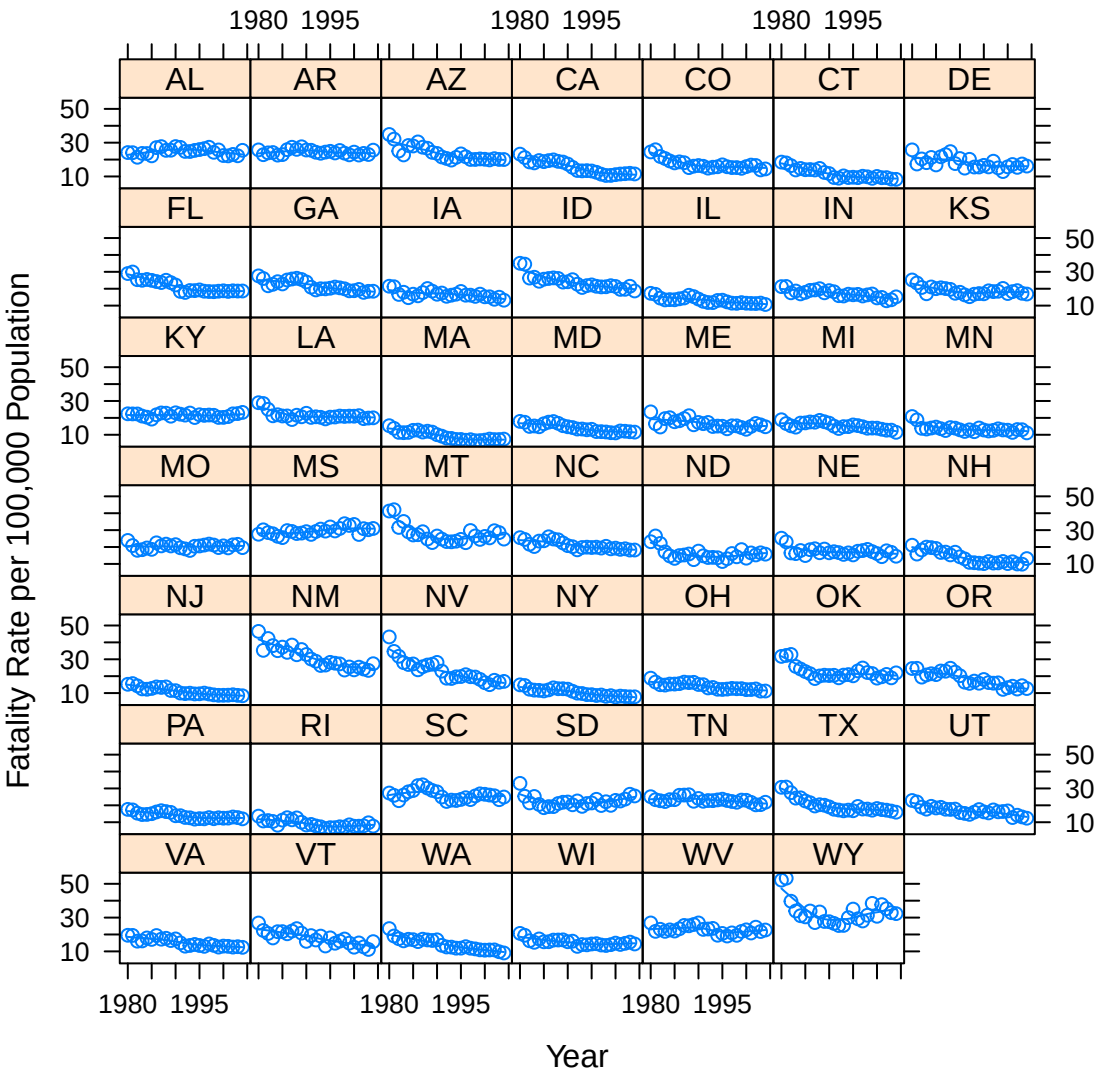


Figure 2: Total Fatality Rate by State

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. It is steadily decreasing until 1990, at which point it 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%. It's mostly fluctuating with a small decreasing trend. Variance of unemployment reduces over time. Vehicle Miles per Capita (`vehicmilespc`) range from 4372 to 18390 miles with median of 9013 miles, and it is increasing over time with steadily increasing variance. Summary statistic suggests all three are asymmetrically distributed with a positive skew, and scatter plot shows all three have weakly positive contemporaneous correlation to total fatality, with `perc_14_24` having the strongly 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
```

```
p1 <- ggplot(data, aes(factor(year), perc14_24)) +
                  geom_boxplot() +
                  labs(y='Population Age 14-24(%)', x ='Year') +
                  formatting + tickformat +
                  scale_x_discrete(breaks=seq(1980, 2004, 2))

p2<- ggplot(data, aes(x=perc14_24, y=totfatrte)) + geom_point()+
  geom_smooth(method=lm, se=FALSE) + ggtitle('perc14_24 vs tolfatrte')+
  xlab('Percent Population Aged 14-24') + ylab('Fatality Per 100,000') + formatting

p3 <- ggplot(data, aes(factor(year), unem)) +
                  geom_boxplot() +
                  labs(y='Unemployment (%)', x ='Year') +
                  formatting + tickformat+
                  scale_x_discrete(breaks=seq(1980, 2004, 2))

p4<- ggplot(data, aes(x=unem, y=totfatrte)) + geom_point()+
  geom_smooth(method=lm, se=FALSE) + ggtitle('unem vs tolfatrte')+
  xlab('Unemployment (%)') + ylab('Fatality Per 100,000') + formatting

p5 <- ggplot(data, aes(factor(year), vehicmilespc)) +
                  geom_boxplot() +
                  labs(y='Miles Per Capita', x ='Year') +
                  formatting + tickformat+
                  scale_x_discrete(breaks=seq(1980, 2004, 2))

p6<- ggplot(data, aes(x=vehicmilespc, y=totfatrte)) + geom_point()+
  geom_smooth(method=lm, se=FALSE) + ggtitle('vehicmilespc vs tolfatrte')+
  xlab('Miles Per Capita') + ylab('Fatality Per 100,000') + formatting

grid.arrange(p1,p3, p5, p2, p4, p6, ncol = 3)
```

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 used 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 general 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

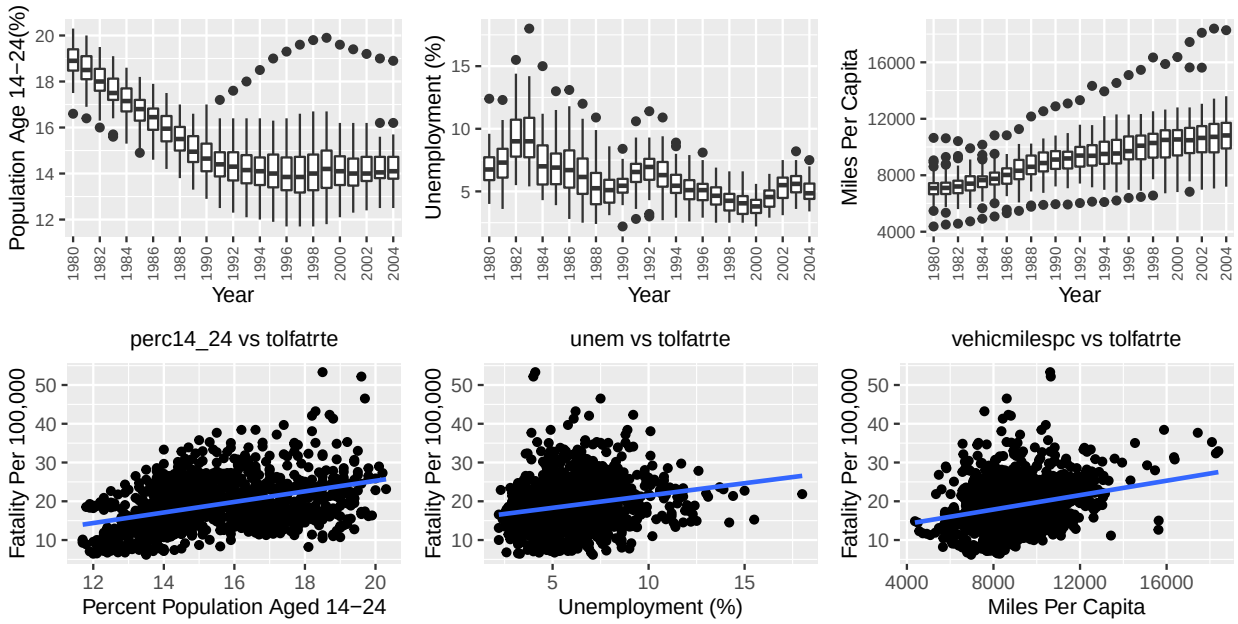


Figure 3: Univariate and Bivariate EDA of Factors Unrelated to Traffic Laws

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")
```

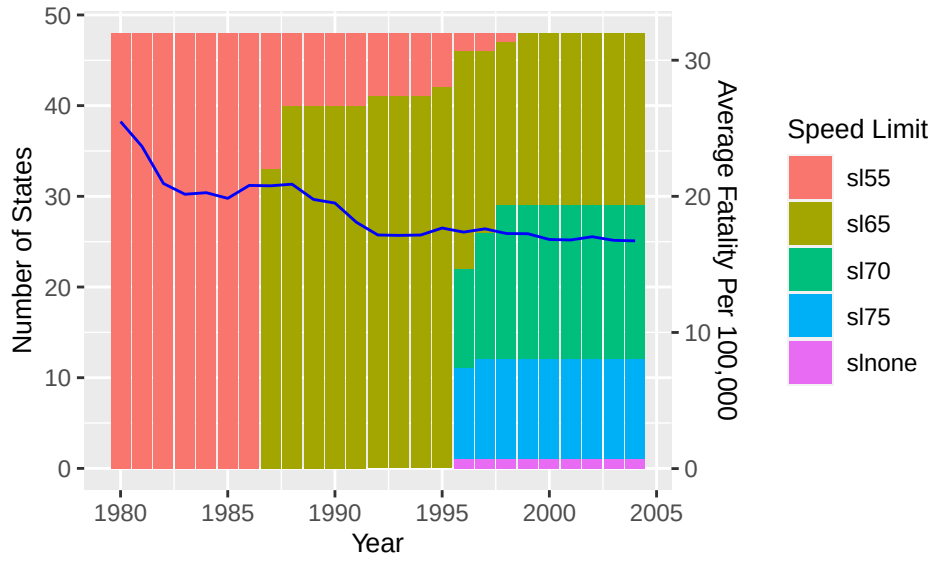


Figure 4: Speed Limit Distribution Over Time with Total Fatality Rate Superimposed

Laws pertaining alcohol are blood alcohol limit .10 (`bac10`), blood alcohol limit .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. 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 alchol 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, there are increasing number of states adopting 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 adopting 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. Its first occurance was 1983, and starting the 90s, there is a dramtic increase in number of states implementing the law. The increase of zero-tolerance is correlated to the decrease of fatality rate.

```
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 ~ "Zero-Tolerance")
    )

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

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

`seatbelt` indicates seatbelt laws, possible values are primary, secondary, or none. `sbprim` and

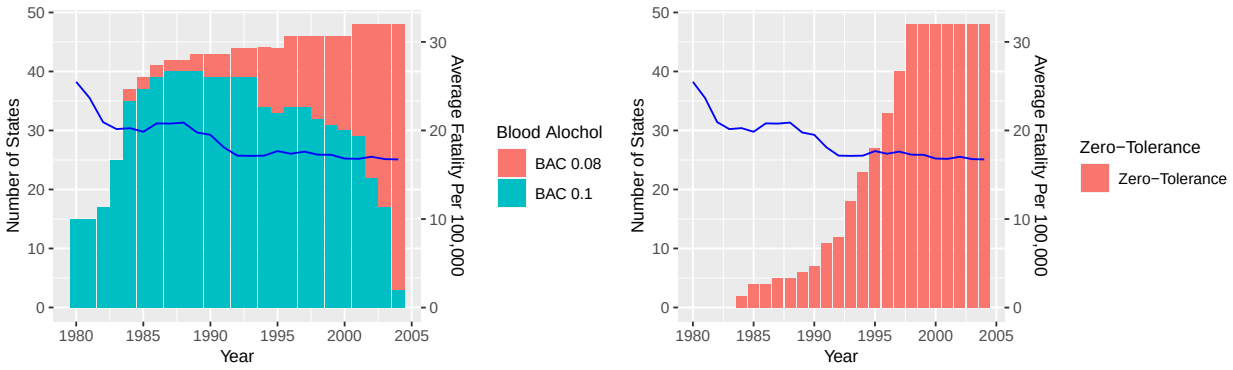


Figure 5: Enforcement of Laws Regarding Alcohol with Total Fatality Rate Superimposed

`sbsecond` are binary indicator variables representing the same information. Other variables concerning traffic laws are graduated drivers license law (`gdl`), administrative license revocation (`perse`). Similar to the speed limit variables, these also contain decimals to represent laws in effect for parts of a year. They were also binarized for the bargraph visualizations (see Figure 6). Seatbelts 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 = ifelse(gdl >0.5, "Enforced", "Not Enforced"),
    perse_bin = ifelse(perse >0.5, "Enforced", "Not Enforced"))

p1 <- data[!is.na(data$seatbelt_bin),] %>% group_by(year, as.factor(seatbelt_bin)) %>% tally()
p2 <- data %>% group_by(year, gdl_bin) %>% tally() %>% genplot("Graduated Driver\nLicense Law")
p3 <- data %>% group_by(year, perse_bin) %>% tally() %>% genplot("Perse Law")+ theme(legend.pos

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

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 positive to each other,`bac08` and `bac10` have strong negative correlation, which is expected, since states 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

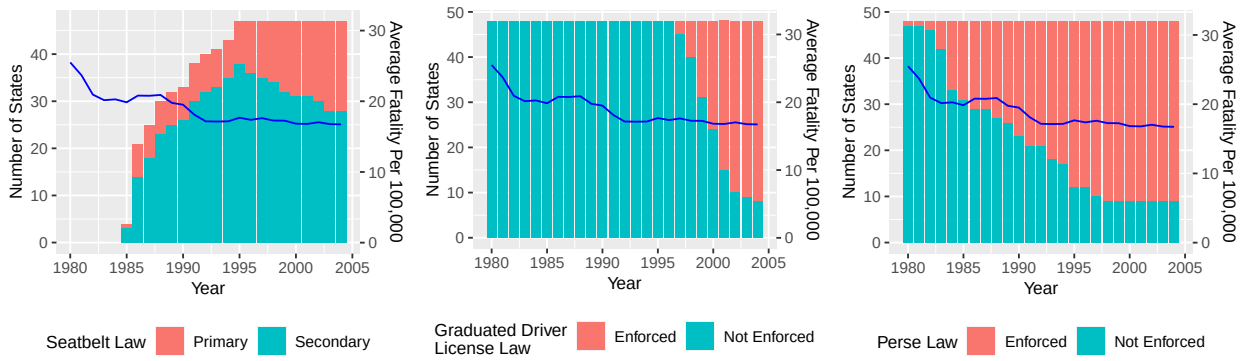


Figure 6: Enforcement of Traffic Laws with Total Fatality Rate Superimposed

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)
```

```
## Warning: package 'corrplot' was built under R version 4.0.2
```

```
## corrplot 0.84 loaded
```

```
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)
```

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 the table below. 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 = "Yearly Average of Total Fatality Rate")
```

A pooled linear regression model was fitted using just the indicator variables for years. The follow

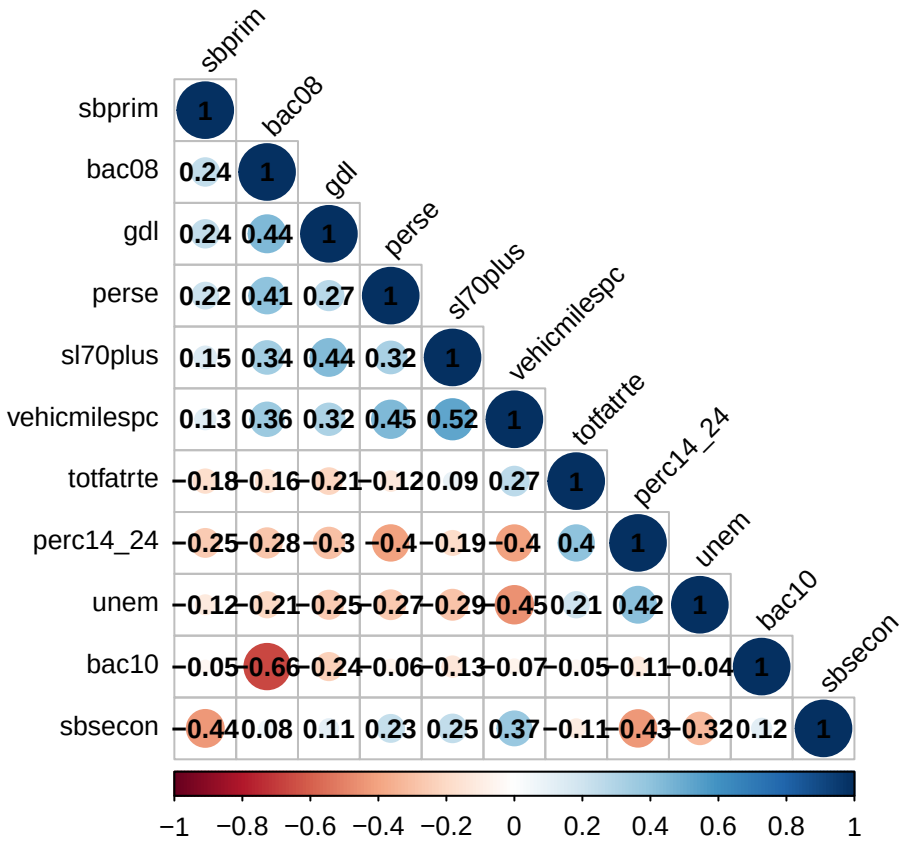


Figure 7: Correlation Matrix of Predictor Variables and Fatality Rate

is the truncated equation summarizing the result of the regression. See table in Appendix A for complete table of coefficients.

$$totfatrte = 25.49 - 1.82d81 - 4.55d82 - 5.34d83 - 5.23d84 - 4.69d85 \cdots - 8.73d02 - 8.73d03 - 8.77d04$$

This model gives us the time effect on total fatality rate. The intercept in this case is the average *totfatrte* across all states in 1980, the baseline year. Each of the coefficients `d80`, `d81…d04` is the average increase in *totfatrte* relative to the base year 2004. 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 injuried or killed in accidents and hence the drop in fatality rate. Additionally, because the fatalty 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 development of 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 as much to the numerator, which would lower the total fatality rate as defined in this dataset.

```
q2.lm <- lm(totfatrte~d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+d91+d92+d93+d94+d95+d96+d97+d98+d
```

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.)

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, implmenting 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, or 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 summary statistic and the boxplots in (Figure 3),

perc14_24 and unem are asymmetric but not strongly skewed. Therefore a log transformation is not necessary. In contrast, 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. 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 = 0.16). For this reason, log-transformation of these two variables are preferred.

The following is the truncated equation summarizing the result of the regression. See table in Appendix A for complete table of coefficients. The coefficients for bac10 , sbprim, sbsecon and gdl are not significant, the coefficient for perse is marginally significant at the 10% level. The rest of the coefficients are significant at the 5% level. The untransformed bac8 and bac10 indicates 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 5.77% decrease in fatality rate; enforcing the BAC limit at 0.1 is associated with 1.62% decrease in fatality rate, though this decrease is not 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.

$$log(totfatrte) = -11.07 - 0.096d81 - 0.32d82 \cdots - 1.02d03 - 1.01d04 - 0.0577bac08 - 0.016bac10 - 0.025perse + 0.014sbprim + 0.031sbsecon + 0.24sl70plus - 0.029gdl + 0.016oerc14_24_0.041unem + 1.54vehic$$

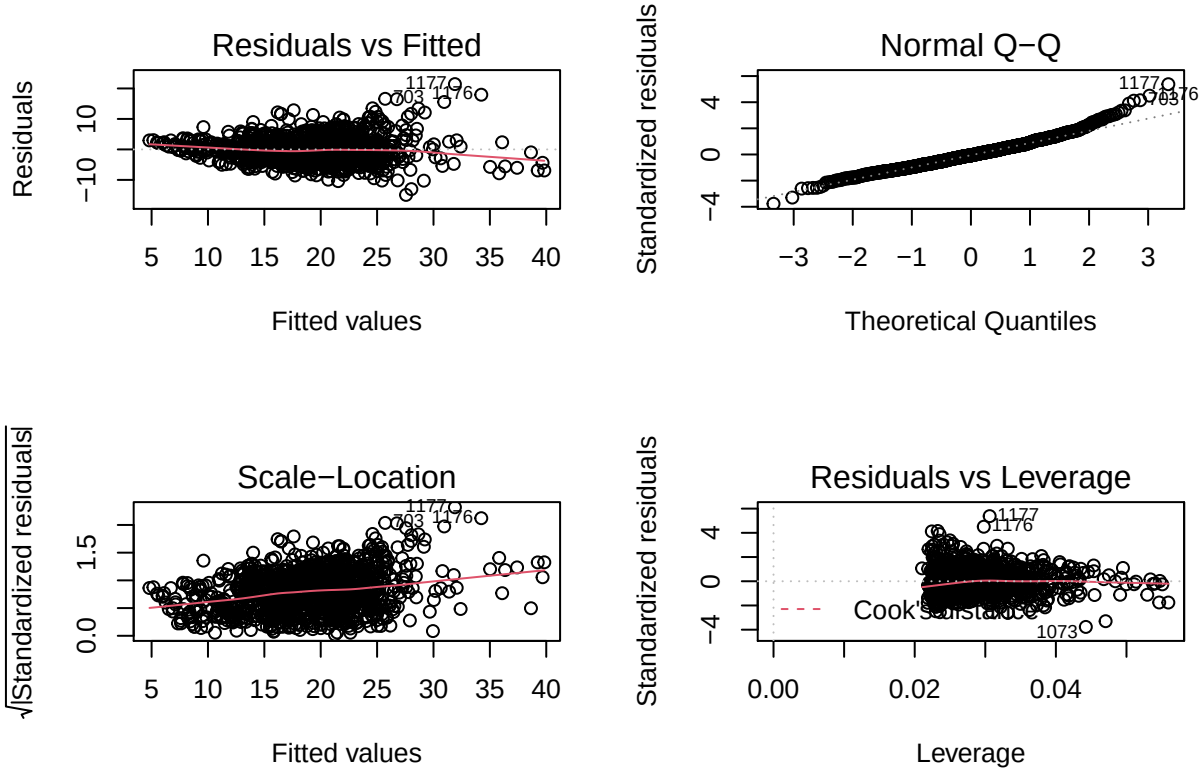
```
q3.untransformed <- lm(totfatrte~factor(year)+ bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+pe

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)
```

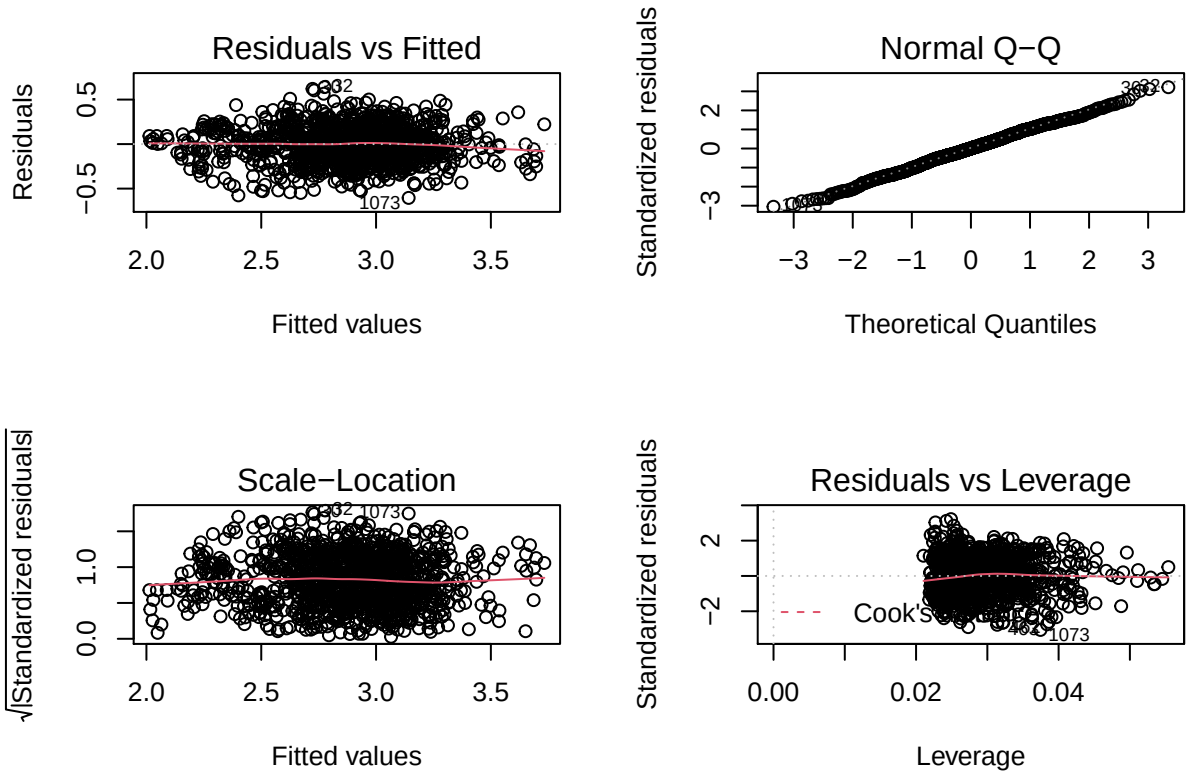



```
q3.lm <- lm(log(totfatrte)~d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+d91+d92+d93+d94+d95+d96+d97+
              bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+unem+log(vehicmilespc),

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

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(q3.lm)
## W = 0.99802, p-value = 0.1652
```

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



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?

See table in Appendix A for complete table of coefficients with comparison to the model in Exercise 3. In the fixed effects model, the coefficients for `bac8` and `sbsecon` are not significant, the coefficient for `bac10` and `gdl` are marginally significant at the 10% level. The rest of the coefficients are significant at the 5% level. In comparison to the pooled OLS, in the fixed effect model, the coefficient for `bac08` became smaller and non-significant; the coefficient for `bac10` increased, its SE decreased, so the p-value is closer to significance though its still not significant at 5%; `perse` became highly significant; `sbprim` became significant as well.

The Pooled OLS model has the following form, where $\delta_t y$ is the time effect on $totfatrte$, a_i and u_{it} form the composite error term, where a_i is the time invariant unobserved factors across the states and $u_i t$ is the idiosyncratic error.

$$totfatrte = \beta_o + \beta_1 bac08 + \beta_2 bac10 + \beta_3 perse + \beta_4 sbprim + \beta_5 sbsecon + \beta_6 sl70plus + \beta_7 gdl + \beta_2 8perc14\_24 + \beta_9 unem + \beta_1 0vehicmilespc + \delta_o y_{81} + \delta_1 y_{82} + ... + \delta_{23} y_{04} + a_i + u_{it}, \quad t = 81, 82...04$$

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 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 BP-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.

The *Fixed effect model* has the following form, where a transformation is applied so the outcome variable as well as each explanatory variable is differenced with the time demeaned value. We see that as a result, the time invariant unobserved effect a_i has disappeared.

$$totfatrte_{it} - \bar{totfatrte}_{it} = \beta_1(bac08 - \bar{bac08}) + ... + \beta_2(vehicmilespc - \bar{vehicmilespc}) + u_{it}, \quad t = 81, 82...04$$

As a result, 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. Perfect linear relationship was confirmed to be absent through correlation matrix, and none of the explanatory variables included in the model are time invariant. The residual vs fitted plot shows the idiosyntractic error to have roughly a mean of zero, so exogeneity assumption is likely met. The residuals for the fixed effect model is also heteroskedastic, shown using BP-test. This issue is addressed by using heteroskedastic robust coefficients. Last, 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. 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 composit 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 urban planning 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+
               bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +

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

```
##
##  Breusch-Pagan test
##
## data:  q4.fe
## BP = 90.939, df = 34, p-value = 4.333e-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 = 228.67, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
q4.fe.se = sqrt(diag(vcovHC(q4.fe)))
```

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 fixed effects model is preferred over the random effects model. Hausman test also rejects the null-hypothesis that the random effect model is consistent (p = 6.576e-05). The main issue with 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 perviously, this assumption is unlikely to hold. On the flip side, because we are not using any time-invariant variables as predictors, fixed effect model can estimate the effects of all predictors on `totfatrte`.

```
q5.re <-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 + unem +
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 = 74.906, df = 34, p-value = 6.586e-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 others 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, it would result in a 6.59% increase in total fatality rate. Compare to this, if `vehicmilespc` was originally 5,000 or 20,000, increasing it by 1,000 miles would lead to 13.18% and 3.30% increase in `totfatrte`, respectively.

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.

###Appendix A Regression Model Results from Question 3-6, using heteroskedastic robust standard error.

```
stargazer(q3.lm, q4.fe, q5.re, type = "text",
          se = list(q3.lm.se, q4.fe.se, q5.re.se),
          column.labels = c("Pooled Linear Model", "Fixed Effect Model", "Random Effect Model"),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## ================================================================================
##                                          Dependent variable:
##                     ---------------------------------------------------------
##                                          log(totfatrte)
##                              OLS                              panel
##                                                              linear
##                       Pooled Linear Model   Fixed Effect Model   Random Effect Model
##                              (1)                   (2)                  (3)
## --------------------------------------------------------------------------------
## d81                         -0.097*              -0.060***            -0.062***
##                             (0.047)               (0.017)              (0.017)
##
## d82                         -0.317***            -0.124***            -0.133***
```

```
##                              (0.047)              (0.021)              (0.020)
##
## d83                        -0.374***            -0.153***            -0.164***
##                              (0.044)              (0.025)              (0.024)
##
## d84                        -0.307***            -0.197***            -0.203***
##                              (0.044)              (0.023)              (0.024)
##
## d85                        -0.344***            -0.221***            -0.229***
##                              (0.045)              (0.028)              (0.028)
##
## d86                        -0.329***            -0.179***            -0.189***
##                              (0.049)              (0.036)              (0.036)
##
## d87                        -0.369***            -0.221***            -0.232***
##                              (0.050)              (0.041)              (0.041)
##
## d88                        -0.384***            -0.246***            -0.259***
##                              (0.052)              (0.051)              (0.050)
##
## d89                        -0.467***            -0.321***            -0.336***
##                              (0.056)              (0.056)              (0.055)
##
## d90                        -0.522***            -0.334***            -0.352***
##                              (0.060)              (0.061)              (0.060)
##
## d91                        -0.637***            -0.374***            -0.394***
##                              (0.062)              (0.065)              (0.064)
##
## d92                        -0.745***            -0.434***            -0.457***
##                              (0.065)              (0.069)              (0.068)
##
## d93                        -0.736***            -0.450***            -0.472***
##                              (0.063)              (0.070)              (0.069)
##
## d94                        -0.725***            -0.481***            -0.502***
##                              (0.064)              (0.070)              (0.069)
##
## d95                        -0.710***            -0.475***            -0.497***
##                              (0.066)              (0.075)              (0.074)
##
## d96                        -0.834***            -0.524***            -0.549***
##                              (0.065)              (0.077)              (0.077)
##
## d97                        -0.863***            -0.543***            -0.569***
##                              (0.066)              (0.080)              (0.079)
##
## d98                        -0.912***            -0.590***            -0.617***
```

```
##                              (0.067)             (0.080)             (0.080)
##
## d99                        -0.918***           -0.604***           -0.631***
##                              (0.067)             (0.084)             (0.082)
##
## d00                        -0.935***           -0.631***           -0.659***
##                              (0.069)             (0.084)             (0.082)
##
## d01                        -0.969***           -0.614***           -0.644***
##                              (0.071)             (0.087)             (0.085)
##
## d02                        -1.003***           -0.585***           -0.619***
##                              (0.073)             (0.086)             (0.084)
##
## d03                        -1.021***           -0.590***           -0.624***
##                              (0.074)             (0.088)             (0.086)
##
## d04                        -1.014***           -0.623***           -0.656***
##                              (0.076)             (0.092)             (0.090)
##
## bac08                       -0.058*             -0.022              -0.025
##                              (0.028)             (0.032)             (0.033)
##
## bac10                       -0.016              -0.020              -0.022
##                              (0.022)             (0.020)             (0.020)
##
## perse                       -0.025             -0.057**            -0.055**
##                              (0.016)             (0.018)             (0.019)
##
## sbprim                       0.014              -0.043              -0.041
##                              (0.025)             (0.024)             (0.024)
##
## sbsecon                      0.031               0.004               0.005
##                              (0.023)             (0.016)             (0.016)
##
## sl70plus                   0.238***             0.072**             0.077**
##                              (0.022)             (0.024)             (0.024)
##
## gdl                         -0.029              -0.022              -0.022
##                              (0.026)             (0.021)             (0.021)
##
## perc14_24                    0.016*              0.021*              0.021*
##                              (0.007)             (0.011)             (0.011)
##
## unem                       0.041***            -0.028***           -0.025***
##                              (0.004)             (0.005)             (0.005)
##
## log(vehicmilespc)          1.545***            0.659***            0.745***
```

```
##                               (0.049)                  (0.137)                (0.128)
##
## Constant                   -11.065***                                        -3.630**
##                               (0.453)                                         (1.152)
##
## ------------------------------------------------------------------------------------------
## Observations                  1,200                    1,200                  1,200
## R2                            0.668                    0.725                  0.710
## Adjusted R2                   0.658                    0.705                  0.702
## Residual Std. Error     0.202 (df = 1165)
## F Statistic         68.837*** (df = 34; 1165) 86.653*** (df = 34; 1118)    2,851.986***
## ==========================================================================================
## Note:                                                      *p<0.05; **p<0.01; ***p<0.001
```