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

Poonam Parhar, Dean Wang, Dili Wang

12/8/2019

Introduction

Driving is one of the key technologies of modern life; it has allowed people to become much more mobile, changed the ways cities develop, and drastically altered the landscape of much of the world. However, the benefits of automobiles come with costs: there is an environmental impact from vehicle emissions, space must be made for roads and freeways, and people may even die in traffic accidents. It is the last of these costs and the factors that affect it which we will examine in this investigation.

There are a variety of variables that may affect how likely a person is to be killed in a car accident. Some of these include the amount of driving done (presumably driving more increases the chance of an accident), the speed limit, whether or not a seatbelt is used, the nature of training for new drivers, and whether the driver is intoxicated with alcohol. There are important policy implications of each of these factors: a state may choose to set speed limits, enforce seat belt use, change the type of mandatory drivers' training, or prohibit driving with a blood alcohol content (BAC) above a certain level.

This study examines a panel dataset that describes how the total traffic fatality rate changes over time for different US states. The dataset also includes many explanatory variables describing policies that were enacted in each state, including speed limits, seat belt laws, driver training laws, and BAC limits. If the most effective policy changes can be identified, this could have a great impact on future legislation, reducing traffic fatality rates and saving lives.

Functions

Here, we define a couple functions that will be used in our report below. The first, `summary.lm.adj()`, is used to ensure that the results of an OLS regression are reported with the correct standard error. The function will automatically apply the Breusch–Pagan test, and upon the return of a significant p-value that signifies the existence of heterosketasticity, it will report OLS regression results using adjusted heteroskedasticity-robust standard errors.

```
summary.lm.adj = function(model) {
    # test for heteroskedasticity
    if (bptest(model)$p.value >= 0.05) {
        return(summary(model))
    } else {
        return(coeftest(model, vcov = vcovHC(model, type = "HC3")))
    }
}
```

The second function, `summary.plm.adj()`, is used to ensure that linear model results for panel data are returned with properly adjusted standard errors. In this function, we first check for heteroskedasticity, then we check if there is serial correlation of idiosyncratic errors. In the case of no heteroskedasticity or serial correlation, the results are returned using the standard `summary()` function. In the event of heteroskedasticity but no serial correlation, we use `vcovHC` with method "white1". If there is both heteroskedasticity and serial correlation, we use "arellano" method with cluster by the group, which is defined as the state in section IV.

```
summary.plm.adj = function(model) {
    # test for heteroskedasticity
    if (bptest(model)$p.value >= 0.05 & pbgtest(model)$p.value >=
        0.5) {
        return(summary(model))
    } else {
        if (bptest(model)$p.value < 0.05 & pbgtest(model)$p.value >=
            0.5) {
            return(coeftest(model, vcov = vcovHC(model, type = "HC3",
                method = "white1")))
        } else {
            if (bptest(model)$p.value < 0.05 & pbgtest(model)$p.value <
                0.5) {
                return(coeftest(model, vcov = vcovHC(model, type = "HC3",
                  method = "arellano", cluster = "group")))
            }
        }
    }
}
```

I. Exploratory Data Analysis

Below, we load the data and examine the desc dataframe to gain an understanding of what each variable represents. While we examine the entirety of the data in our study, we only disply a summary of the key explanatory variables and the dependent variable, total fatality rate, in a table below.

```
# load datasets from driving.Rdata and look at the names of
# dataframes
driving.df <- load("driving.Rdata")
str(driving.df)
```

```
##  chr [1:4] ".Random.seed" "data" "desc" "self"
```

```
# examine the structure of data
desc
```

```
# examine the variables
str(data)
```

```
# summary of variables
summary(data)
```

Summary Table of Important Variables:

```
data.subset = subset(data, select = c(totfatrte, bac08, bac10,
    perse, sbprim, sbsecon, sl70plus, slnone, gdl, perc14_24,
    unem, vehicmilespc, year))
desc.summary = data.frame(variable = character(), n = character(),
    missing = character(), distinct = character(), mean = character(),
    min = double(), q1 = character(), median = character(), q3 = character(),
    max = double())
for (i in 1:ncol(data.subset)) {
    column = data.subset[[colnames(data.subset)[i]]]
    desc = describe(column)
    dr = desc$counts
    row = data.frame(variable = colnames(data.subset)[i], n = dr[1],
```

```
        missing = dr[2], distinct = dr[3], mean = dr[5], min = min(column),
        q1 = quantile(column)[2], median = quantile(column)[3],
        q3 = quantile(column)[4], max = max(column))
    desc.summary <- rbind(desc.summary, row)
}
kable(desc.summary)
```

	variable	n	missing	distinct	mean	min	q1	median	q3	max
n	totfatrte	1200	0	916	18.92	6.200	14.3775	18.435	22.7725	53.32
n1	bac08	1200	0	8	0.2135	0.000	0.0000	0.000	0.0000	1.00
n2	bac10	1200	0	10	0.6231	0.000	0.0000	1.000	1.0000	1.00
n3	perse	1200	0	9	0.5471	0.000	0.0000	1.000	1.0000	1.00
n4	sbprim	1200	0	2	215	0.000	0.0000	0.000	0.0000	1.00
n5	sbsecon	1200	0	2	562	0.000	0.0000	0.000	1.0000	1.00
n6	sl70plus	1200	0	15	0.2068	0.000	0.0000	0.000	0.0000	1.00
n7	slnone	1200	0	3	0.007569	0.000	0.0000	0.000	0.0000	1.00
n8	gdl	1200	0	8	0.1741	0.000	0.0000	0.000	0.0000	1.00
n9	perc14_24	1200	0	87	15.33	11.700	13.9000	14.900	16.6000	20.30
n10	unem	1200	0	112	5.951	2.200	4.5000	5.600	7.0000	18.00
n11	vehicmilespc	1200	0	1200	9129	4372.046	7788.0964	9012.670	10327.2732	18390.08
n12	year	1200	0	25	1992	1980.000	1986.0000	1992.000	1998.0000	2004.00

There are 56 variables, each with a total 1200 observations. The dataset has variables related to the driving laws such as alcohol limt, speed limit, and various traffic fatalities measurements. The dataset also includes a few economic and demographic variables. The data is collected for 48 different US states over 25 years.

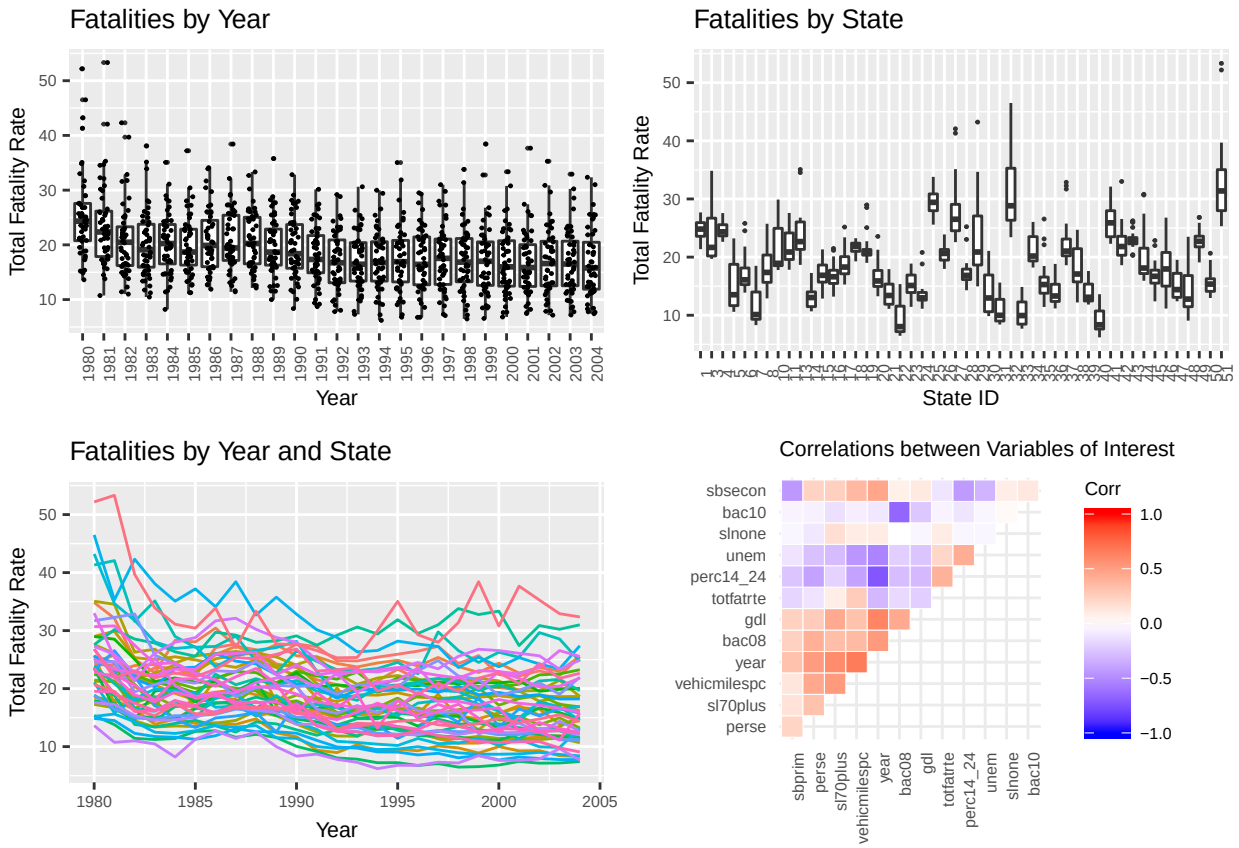
We can make a few observations in terms of specific variables:

- There are state and year variables, which serve as the `group` and `time` indices, respectively, for this panel dataset.
- `sl55` to `slnone`, `zerotol`, `gdl`, `bac08`, `bac10`, `perse`, `sbprim`, and `sbsecon` variables are indicator variables that show whether or not a given state in a given year has enacted specific legislation or not. `seatbelt` contains the same information that `sbprim` and `sbsecon` contain, but in a "factor" form. `sl70plus` binarizes information from the `sl` variables with a cutoff at 70 mph.
- The `totfat`, `nghtfat`, and `wkndfat` variables show the total, night, and weekend number of fatalities in a state. These variables provide data which our dependent variable can be derived from.
- The `totfatpvm`, `nghtfatpvm`, and `wkndfatpvm` variables show the fatalities in certain categories per number of miles driven. These variables will not be examined in this study
- The `totfatrte`, `nghtfatrte`, and `wkndfatrte` variables show the fatalities per 100,000 population - the population of the state, `statepop` is used to calculate these. It is the first of these variables, `totfatrte`, which we will examine as our dependent variable in our study.
- `vehicmiles`, `unem`, and `perc14_24` are continuous explanatory variables.
- There are indicator variables `d80` through `d04` for each of the years.

Using the `describe()` function on the entirety of the data, we ascertained that none of the variables had any missing values.

Next, we start to examine how the total fatality rate varies by year and state. We also examine correlations between the total fatality rate and variables of interest. The variables of interest include all the variables describing what laws exist in a state, as well as the level of unemployment and percentage of the population between 14 and 24. We include the law variables since finding the effect a law has on the fatality rate has important policy implications; we include the unemployment rate and percent of the population between 14 and 24 since we intuitively believe that these variables can serve as control covariates and explain some of the variation in total fatality rate as well, even if a state government cannot control these directly.

```
# Fatalities by year
p1 = ggplot(data, aes(as.factor(year), totfatrte)) + geom_boxplot(outlier.size = 0.2) +
    geom_jitter(width = 0.2, size = 0.2) + ggtitle("Fatalities by Year") +
    ylab("Total Fatality Rate") + xlab("Year") + theme(text = element_text(size = 8),
    axis.text.x = element_text(angle = 90, size = 6), axis.text.y = element_text(size = 6))
# Fatalities by state
p2 = ggplot(data, aes(as.factor(state), totfatrte)) + geom_boxplot(outlier.size = 0.2) +
    ggtitle("Fatalities by State") + ylab("Total Fatality Rate") +
    xlab("State ID") + theme(text = element_text(size = 8), axis.text.x = element_text(angle = 90,
    size = 6), axis.text.y = element_text(size = 6))
# Fatalities by year and state
p3 = ggplot(data, aes(year, totfatrte)) + geom_line(aes(col = as.factor(state))) +
    theme(legend.position = "none") + ggtitle("Fatalities by Year and State") +
    ylab("Total Fatality Rate") + xlab("Year") + labs(color = "State ID") +
    theme(text = element_text(size = 8))
# Correlation Plot
p4 = ggcorrplot(cor(data.subset), type = "upper", hc.order = TRUE,
    outline.col = "white", lab_size = 4) + theme(text = element_text(size = 7),
    axis.text.x = element_text(angle = 90, size = 6), axis.text.y = element_text(size = 6)) +
    ggtitle("Correlations between Variables of Interest")
grid.arrange(p1, p2, p3, p4, ncol = 2, nrow = 2)
```



In the first of the above plots, we can see the distributions of the state observations for total fatality rate varying by year. There seems to be a general decrease in the total fatality rate median, with a large decrease in the early 80s, a slight increase in the late 80s, and another decrease in the early 90s with the rate holding steady after that. However, the distributions of observations are overlapping so it is hard to say whether the decrease was truly significant at this point in this investigation, but we will examine this more closely in Part

II.

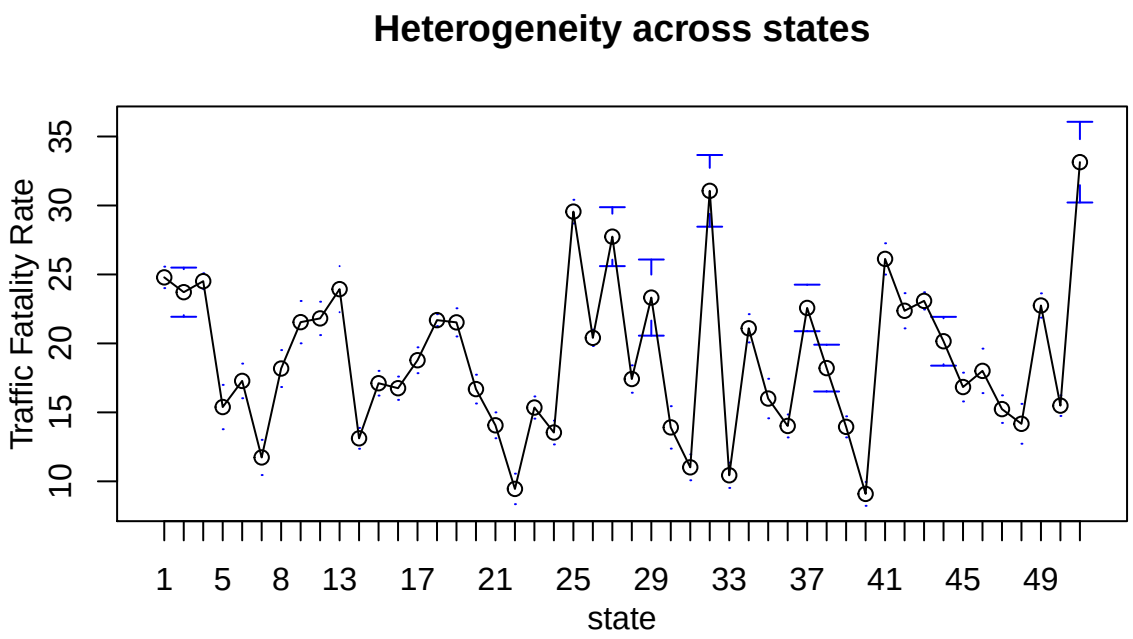
In the second of the graphs above, we see the distribution of total fatality rate observations by state, there is significant variation depending on the state. Some states have very high values like state 32, while others, like state 40, have very low values. Also, the variance appears to vary greatly: state 32 has a larger spread of values of total fatality rate, while state 40 has a much narrower distribution. The fact that the values of total fatality rate can vary so much by state makes a good argument that it may make more sense to use a fixed effects model by state than a pooled regression model.

In the third plot, we look at how the total fatality rate changes over time by state. There seems to be a general decrease in the rate, with most of the lines sloping downward. Almost all states see a decrease or slight decrease in total fatality rate.

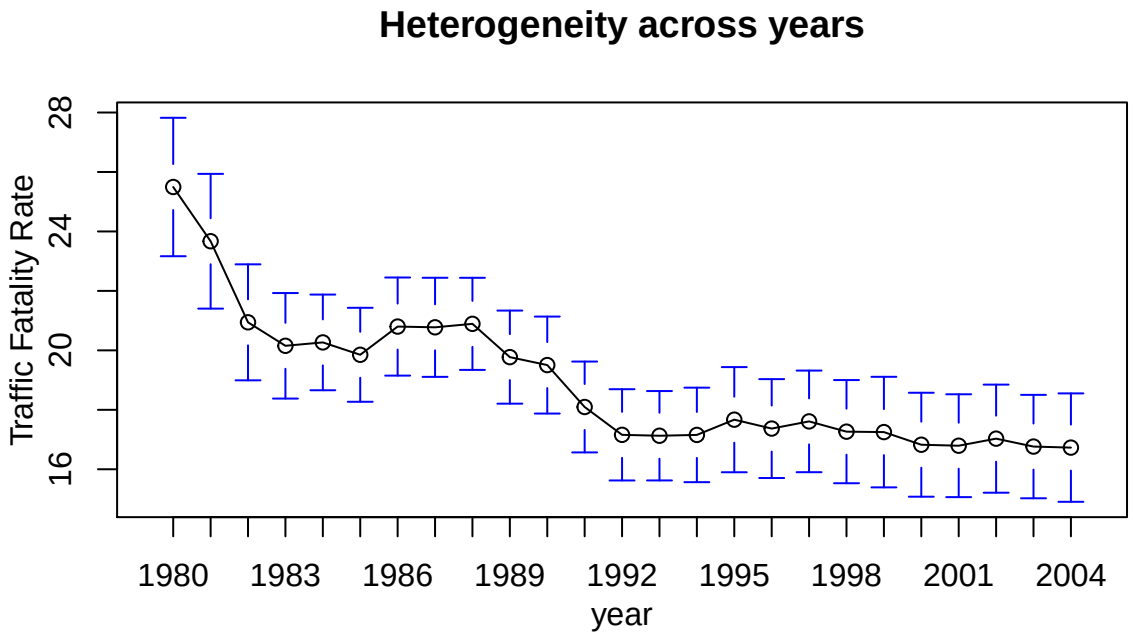
The forth plot shows correlations betweeen explanatory variables and covariates of interest. The dependent variable, `totfatrte`, has significant positive correlation with `perc14_24`, `vehicmilespc` and `unemp`. It is negatively correlated with `sbprim`, `gdl`, `bac08` and `bac10`. There also appears to strong correlations between `bac08` and `bac10`, `sl70plus` and `vehicmilespc`, `gdl` and `year`, and `vehicmilespc` and `year`.

Furthermore, use group means plots below to observe that 'totfatrte' has heterogeneity across 'states' as well as across 'years'. The graphs below represents a summary of the information from the first 2 subplots shown above.

```
plotmeans(totfatrte ~ state, main = "Heterogeneity across states",
    n.label = "FALSE", ylab = "Traffic Fatality Rate", data = data,
    mgp = c(2, 1, 0))
```

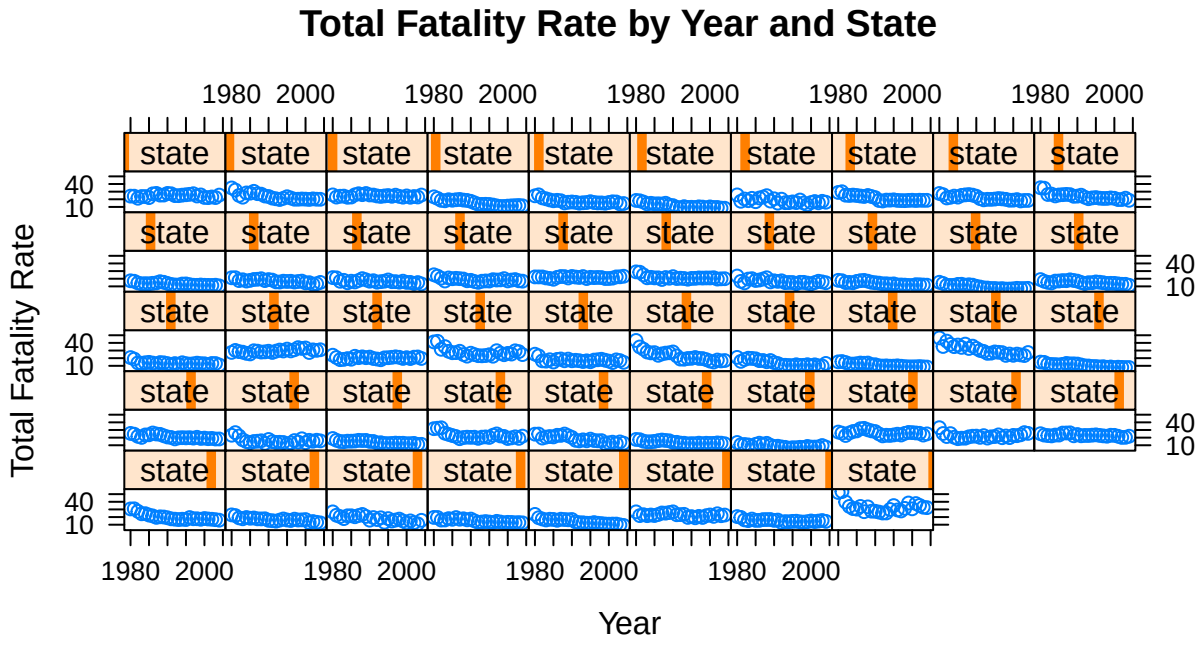



```
plotmeans(totfatrte ~ year, main = "Heterogeneity across years",
    n.label = "FALSE", ylab = "Traffic Fatality Rate", data = data,
    mgp = c(2, 1, 0))
```



To accentuate the relationship between fatality rate and time within each state over the 25 years, we also display the relationship between `totfatrte` and `year` per state in the xy-plot below. As seen here, the decrease in fatality rate over time is obvious for most states, although a few states either show a constant relationship or a slight increase. Nevertheless, the variations in the relationship between fatality rate and time for each state is further evidence that state fixed effects model is likely more trustworthy than an pooled OLS model.

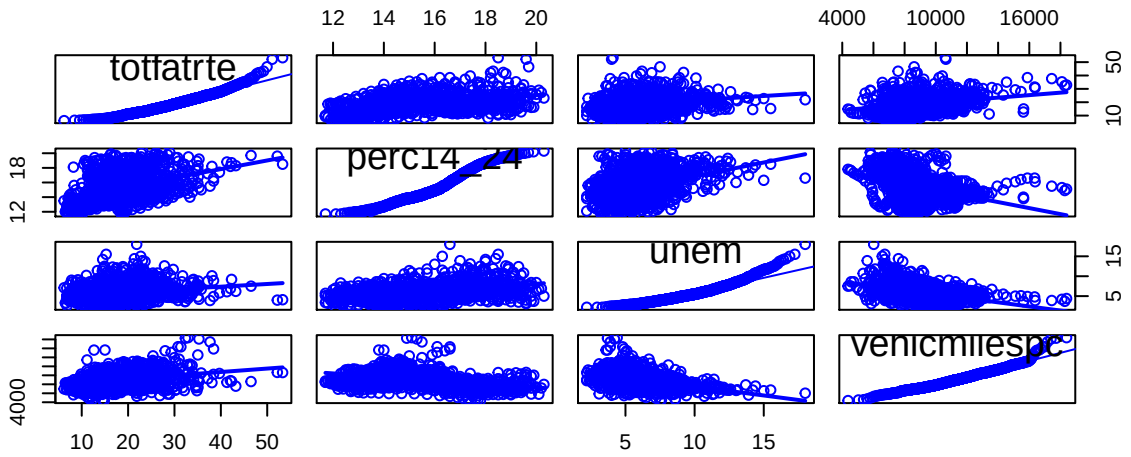
```
xyplot(totfatrte ~ year | state, data = data, as.table = T, ylab = "Total Fatality Rate",
    xlab = "Year", main = "Total Fatality Rate by Year and State")
```



Next, we examine the distributions of the continuous-valued variables as well as their scatterplots with each other.

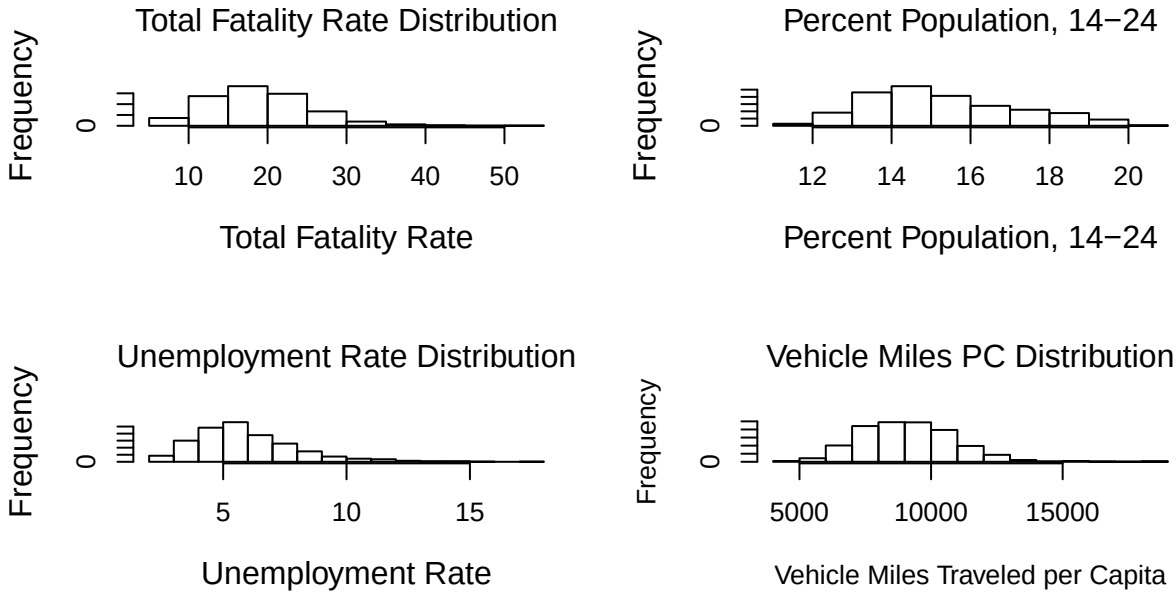
```
options(repr.plot.width = 12, repr.plot.height = 10, repr.plot.pointsize = 12)
```

```
scatterplotMatrix(~totfatrte + perc14_24 + unem + vehicmilespc,
    data = data, diagonal = list(method = "qqplot"), smooth = F)
```



Key points that we observe from the above plots are that the fatality rate has positive correlation with `perc14_24`, `unem` and `vehicmilespc`. 3 of the four continuous varibles here, `totfatrte`, `unem` and `vehicmilespc`, have self Q-Q plots that show slight deviations from normality. However, before we consider any necessary logarithmic transformations, we examine the histograms of these 4 countinous variables below:

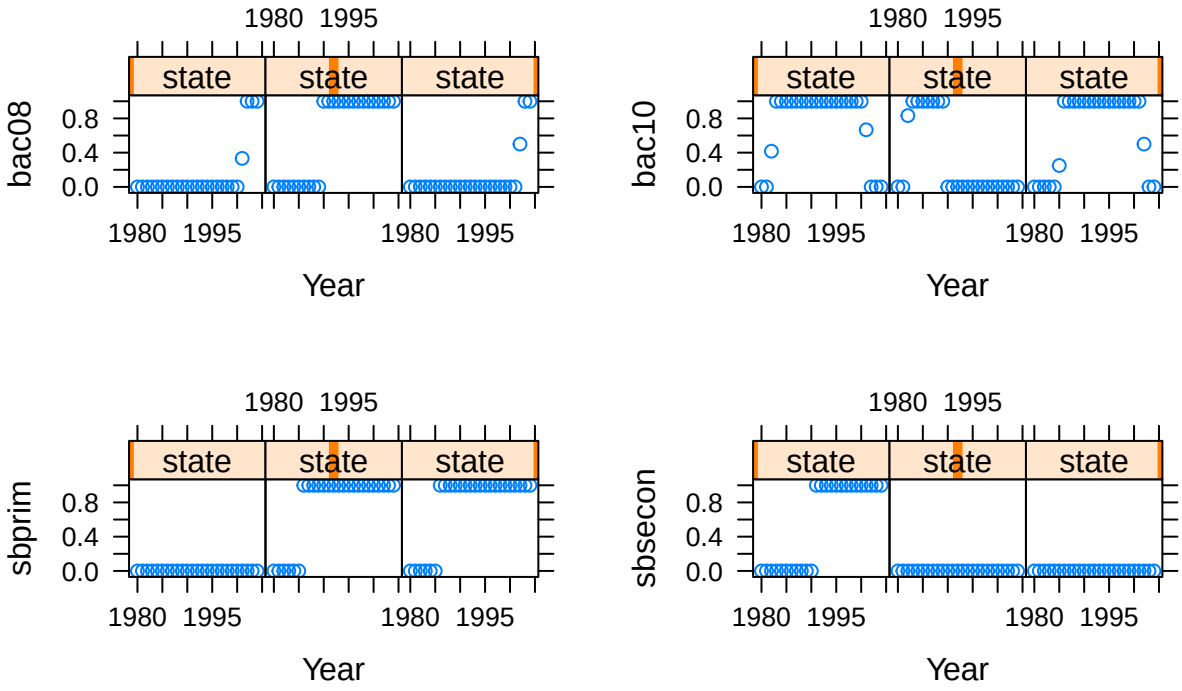
```
par(mfrow = c(2, 2))
hist(data$totfatrte, ylab = "Frequency", xlab = "Total Fatality Rate",
    main = "Total Fatality Rate Distribution", cex = 0.7, font.main = 6,
    cex.lab = 1.2)
hist(data$perc14_24, ylab = "Frequency", xlab = "Percent Population, 14-24",
    main = "Percent Population, 14-24", cex = 0.7, font.main = 6,
    cex.lab = 1.2)
hist(data$unem, ylab = "Frequency", xlab = "Unemployment Rate",
    main = "Unemployment Rate Distribution", cex = 0.7, font.main = 6,
    cex.lab = 1.2)
hist(data$vehicmilespc, ylab = "Frequency", xlab = "Vehicle Miles Traveled per Capita",
    main = "Vehicle Miles PC Distribution", cex = 0.7, font.main = 6)
```



Upon further examination of their histogram distributions, we see that these 4 aforementioned variables does not have extreme departures from normality. The observations in Q-Q plot that seemed to suggest departure from normality may be likley due to outliers.

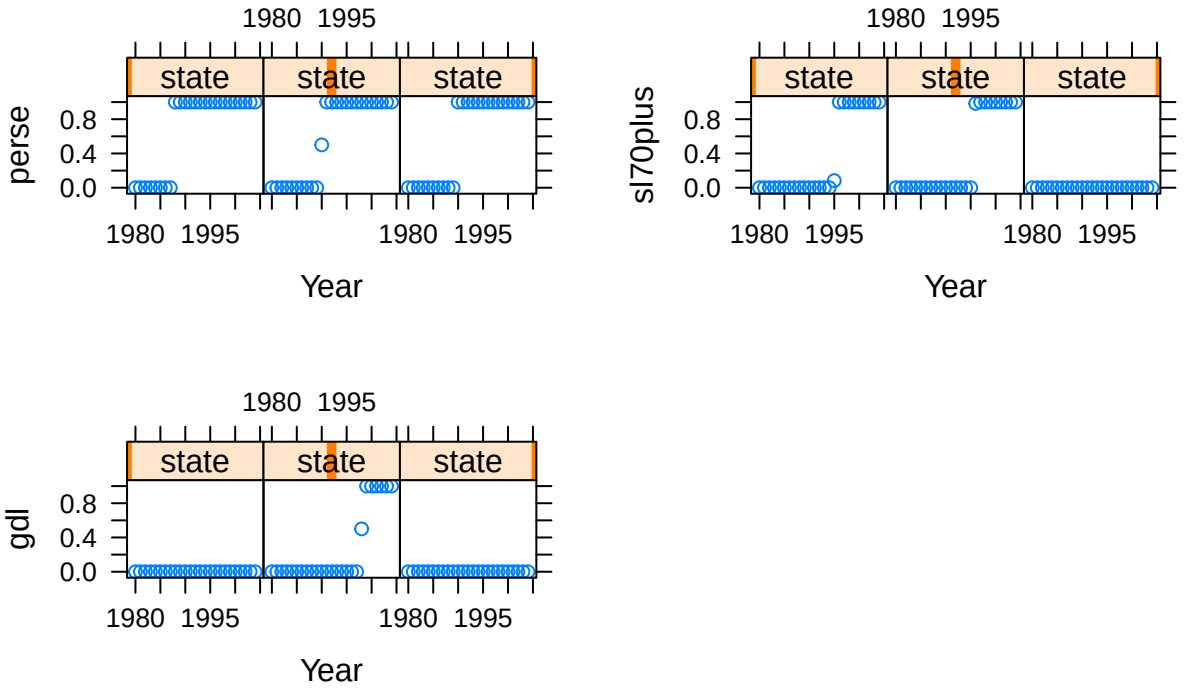
Based on earlier observations of the data, including summary table we created above, we noticed some of the variables that represent whether or not a law has been enacted appear as if they should be indicator variables, with minimum value of 0, maximum of 1 and very few unique values. However, some of these variables, in particular `perse`, `gdl`, `sl70plus`, `bac08`, `bac10`, have values that are neither 0 or 1. We found this information to be quite useful - `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`, and decided to plot these variables for a subset of states below to examine their relationship with `year`. These plots also will serve to inform us on decisions regarding the transformation of these variables for our models later on. Given the large number of states and the size of x-y plots, we only plot these variables with respect to `year` for a subset of states, as the relationships pictured below is similar for all states.

```
sample.states = subset(data, state == 3 | state == 5 | state ==
    7)
p1 = xyplot(bac08 ~ year | state, data = sample.states, as.table = T,
    ylab = "bac08", xlab = "Year")
p2 = xyplot(bac10 ~ year | state, data = sample.states, as.table = T,
    ylab = "bac10", xlab = "Year")
p3 = xyplot(sbprim ~ year | state, data = sample.states, as.table = T,
    ylab = "sbprim", xlab = "Year")
p4 = xyplot(sbsecon ~ year | state, data = sample.states, as.table = T,
    ylab = "sbsecon", xlab = "Year")
grid.arrange(p1, p2, p3, p4, ncol = 2, nrow = 2)
```




```
p5 = xyplot(perse ~ year | state, data = sample.states, as.table = T,
    ylab = "perse", xlab = "Year")
p6 = xyplot(sl70plus ~ year | state, data = sample.states, as.table = T,
    ylab = "sl70plus", xlab = "Year")
p7 = xyplot(gdl ~ year | state, data = sample.states, as.table = T,
    ylab = "gdl", xlab = "Year")
```

```
grid.arrange(p5, p6, p7, ncol = 2, nrow = 2)
```



Had we attempted to plot any of these variables with respect to `year`, without separations by `state`, we would have seen some rather unscrutable relationships between indicator variables and `year`. However, with x-y plots that are partitioned by state, the relationships between each indicator variable and `year` appear to be quite simplistic and clean. We can see that for the most part, all 7 variables that represent enactment of laws take on the value of 0 or 1 within each state for almost all years. Every state either has 0, 1, or at most 2 years where the variable takes on a non-binary value. The existence of a non-binary decimal value always represents the transition from 0 to 1 or 1 to 0. It appears that only the `bac10` variable has both transition from 0 to 1 and 1 to 0 in the 25 year time period, whereas most other variables makes the transition once.

Using the information that a fraction indicates the part of the year that the law is enacted, we decided on a rather simple transformation for these indicator variables, which is discussed in detail in Part III.

II. Total Fatality Rate and Time Dependency

Our response variable 'totfatrte' is defined as 'total fatalities per 100,000 population for a given state in a particular year'. In the dataset, it is a numerical variable with values ranging from 6.20 to 53.32. Below, we estimate a linear model that regresses the total fatality rate on the year indicator variables, excluding the indicator variable d80 to avoid perfect multicollinearity. We note here that the function used to estimate the model below is equivalent `lm(totfatrte ~ factor(year), data = data)`, but the factor(year) term would not be able to be used in panel data linear models later on. Therefore, we will continue to use the year indicator variables moving forward.

```
# build linear regression model select totfatrte and d81 to
# d04 columns leave d80 out, as that will be the base year
# for our model
model.lm <- lm(totfatrte ~ ., data = data[, c(22, 32:55)])
summary.lm.adj(model.lm)
```

Here, we provide an equation that reports the coefficients of the model along with standard errors (The full summary of the model is present in our .rmd document). No heteroskedasticity was present in this model,

based on the Breusch–Pagan test.

$$
\begin{aligned}
totfatrte = & \underset{(0.8671)^{***}}{25.4946} - \underset{(1.2263)}{1.8244} \cdot d81 - \underset{(1.2263)^{***}}{4.5521} \cdot d82 - \underset{(1.2263)^{***}}{5.3417} \cdot d83 - \underset{(1.2263)^{***}}{5.2271} \cdot d84 - \underset{(1.2263)^{***}}{5.6431} \cdot d85 \\
& - \underset{(1.2263)^{***}}{4.6942} \cdot d86 - \underset{(1.2263)^{***}}{4.7198} \cdot d87 - \underset{(1.2263)^{***}}{4.6029} \cdot d88 - \underset{(1.2263)^{***}}{5.7223} \cdot d89 - \underset{(1.2263)^{***}}{5.9894} \cdot d90 \\
& - \underset{(1.2263)^{***}}{7.3998} \cdot d91 - \underset{(1.2263)^{***}}{8.3367} \cdot d92 - \underset{(1.2263)^{***}}{8.3669} \cdot d93 - \underset{(1.2263)^{***}}{8.3394} \cdot d94 - \underset{(1.2263)^{***}}{7.8260} \cdot d95 \\
& - \underset{(1.2263)^{***}}{8.1252} \cdot d96 - \underset{(1.2263)^{***}}{7.8840} \cdot d97 - \underset{(1.2263)^{***}}{8.2292} \cdot d98 - \underset{(1.2263)^{***}}{8.2442} \cdot d99 - \underset{(1.2263)^{***}}{8.6690} \cdot d00 \\
& - \underset{(1.2263)^{***}}{8.7019} \cdot d01 - \underset{(1.2263)^{***}}{8.4650} \cdot d02 - \underset{(1.2263)^{***}}{8.7310} \cdot d03 - \underset{(1.2263)^{***}}{8.7656} \cdot d04
\end{aligned}
$$

From the summary of the model, we note that all of the variable coefficients, including the intercept that represents the average totfatrte in the base year 1980, are statistically significant for the linear model. Multiple R-squared for the model is 0.1276 which means that only 12.76% of the variation in fatality rate is explained by the year indicator variables. The F-statistic is 7.164 and is highly statistically significant (p-value: $<$ 2.2e-16), so all the estimators of this linear model have joint statistical significance.

We also note that the intercept is positive and a large value (25.4946), and that the coefficients for all the indicator variables representing years are negative. These negative coefficients are generally decreasing except for a few exceptions from years 1986 to 1989, 1995 and 1997 when they increased slightly as compared to the coefficients of the previous year.

Here, we provide the steps necessary to interpret the results from the above model. Using the y-intercept, we see that the average totfatrte for the year 1980 is $totfatrte = 25.4946$. For any other year, we would use the coefficient associated with that year's indicator variable in the linear model to the calculate the average total fatality rate for that year with respect 1980. Here, the average is calculated over the 48 states for a given year. For instance, the total fataility rate for the year 2004 is $totfatrte = 25.4946 + (-8.7656) = 16.729$.

While we can see that there is general decrease in total fatality rate from year to year based on the mostly decreasing coefficients associated with the year indicators, we can also demostrate that the decrease yearly is significant. Since the year variable has ordinality and equal spacing between the values, there is validity to treating it as a numeric variable if we only want to model the relationship between `totfatrte` and `year`:

```
model.lm.2 <- lm(totfatrte ~ year, data = data)
summary.lm.adj(model.lm.2)
```

```
##
## Call:
## lm(formula = totfatrte ~ year, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9201  -4.3576  -0.7668   3.6596  31.3606
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 569.58781   48.24189   11.81   <2e-16 ***
## year         -0.27644    0.02422  -11.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.05 on 1198 degrees of freedom
## Multiple R-squared:  0.09809,    Adjusted R-squared:  0.09734
## F-statistic: 130.3 on 1 and 1198 DF,  p-value: < 2.2e-16
```

Here, we see that the coefficient associated with the numeric year variable is negative and highly statistically significant. This would mean that the average total fatality rate decreases with by 0.28 per 100,000 population with increasing year. From both of these linear models, we can conclude that the traffic fatality rate has decreased over the years.

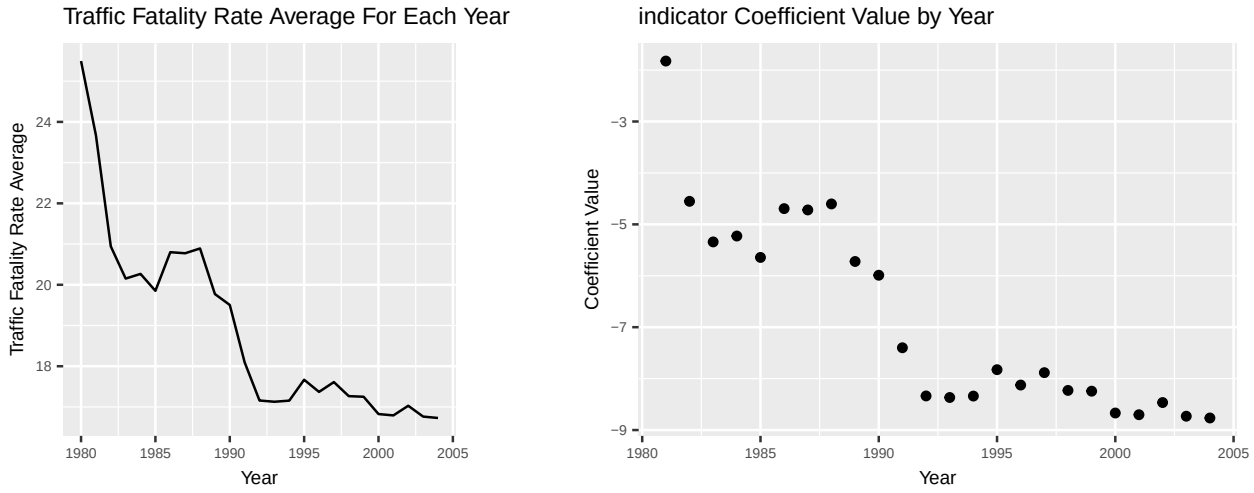
Below, we calculate the average of 'totfatrte' for each of the years in the dataset. we then compare these averages to the coefficient value for each year indicator variable.

```
# calculate fatality rate average for each year
totfatrte_means <- list()
j <- 1
# d80 to d04 indicators correspond to columns 31 to 55
for (i in c(31:55)) {
    totfatrte_subset <- subset(data, data[, i] == 1, select = totfatrte)
    totfatrte_means[[j]] <- mean(totfatrte_subset$totfatrte)
    j <- j + 1
}

p1 = ggplot() + geom_line(aes(x = seq(1980, 1980 + j - 2), y = as.vector(unlist(totfatrte_means)))) +
    ylab("Traffic Fatality Rate Average") + xlab("Year") + ggtitle("Traffic Fatality Rate Average For Ea
    theme(aspect.ratio = 1) + theme(text = element_text(size = 8))

p2 = qplot(seq(1981, 2004), model.lm$coefficients[-1], ylab = "Coefficient Value",
    xlab = "Year", main = "indicator Coefficient Value by Year") +
    theme(text = element_text(size = 8))

grid.arrange(p1, p2, ncol = 2, nrow = 1)
```



We can see from the plots above that the trend is nearly identical. The graph on the left represents true total fatality rates averages calculated over the states per year. Since we used indicator variables for years in the linear model, the coefficients for each year indicator represents the average difference in fatality rate with the respect to the base year of 1980. This means that the points in the graph on the right should follow the identical trend, but with a downward translation, since all averages are with respect to 1980. Again from these plots, we see that traffic fatality rates decrease over time.

To summarize, we fit two different linear models: one with one indicator variable per year and another that treated `year` as a numeric variable. In both cases, a statistically significant decrease in total fatality rate from the 1980 level could be seen. We also compared the averages of total fatality rate over the years and observed that this also decreased, allowing us to conclude that driving has become safer, on average, during this time period, due to decrease in fatality rate.

III. Pooled OLS Modeling

In this section, we build a extended Pooled OLS models including other explanatory variables such as *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*. Prior to the inclusion of these additional explanatory variables, we considered possible transformations. As noted in the exploratory data analysis in part III, the deviations from normality in the Q-Q plots of the continuous variables `unem`, `vehicmilespc` and the dependent variable `totfatrte` did call into question whether or not logarithmic transformation might be needed. However, our histograms of these continuous variables showed that deviations from normality were only due to outliers. Therefore, we decided not to transform any of the continuous variables.

However, the variables `bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl` do require further consideration due to their current coding as numeric variables. One of our concerns is that treating these variables as numeric will confound interpretation and give imporoper statistical power to decimal values. One thing we considered is that if decimal values were left in the regression, this would lead to overfitting of data to a few observations that could bias or skew our linear model results. Therefore, we used the plots of these variables with respect to year, but separated by state, to determine the best transformation to binarize all 7 variables.

We considered a few possible tranformations as several arguments could be made for how the indicators should be treated. For instance, one could argue that given the incubation period of a law after it is enacted that could take up to a year, all of the decimals should be treated as 0s and therefore a `floor()` function could be used for tranformation. One the other hand, one can also make the argument that if a law exists in a year, and data is collected a the end of the year, then that law must have some influence on the dependent variable, so the simple existence of a law for even part of a year should have some impact on `totfatrte`. For this argument, we considered using the `ciel()` function for transformation. However, the `ciel()` function which would always round up has a flaw for the varialbe bac10, which uses decimals to represent the transition from 1 to 0, or the retraction of a the blood alcohol limit at 0.10 law. This would require us to examine the value in the previous year, which makes this transformation difficult to implemnent with panel data grouped by state.

Ultimately, we felt that rounding the decimals to the nearest integer would be the best option as it balances out both of scenarios stated above. We also thought it was reasonable to assume that if a law was enacted for more than 6 months in a year, then its effect on traffic incident rates would be measureable, while a law that was enacted for less than half a year may not have the same effect since it existed for a period that was too short. Therefore, we went with using the `round()` function for our indicator variables transformation.

Additionally, we took a look at indicator variables 'bac08' and 'bac10'. Variables 'bac08' and 'bac10' are defined as:

- bac10: blood alcohol limit .10 (.10g of alcohol for every 100 ml of blood) constitutes drunk driving
- bac08: blood alcohol limit .08 (.08g of alcohol for every 100 ml of blood) constitutes drunk driving

One of our major concerns here is the existence of near perfect multicollinearity between these two variables, since there is bound to be a relationship between the two. The reason for this is that if a state has the bac10 law enacted, it is highly unlikely that the state will also have the more strict bac08 law enacted simultaneously. The only way for perfect multicollinearity to be broken is if there is a number of states that have neither laws during the same year. We test to make sure that there is no perfect multicollinearity below, and see that binarization suprisingly decreases the correlation between these two variables.

```
cor(data$bac08, data$bac10)
```

```
## [1] -0.6637454
```

```
cor(round(data$bac08), round(data$bac10))
```

```
## [1] -0.6363509
```

Next, we fit an "extended" linear model. This model still contains year indicator variables, but also now has all the indicators for the state laws as well as the three continuous-valued explanatory variables.

```
model.extended.lm <- lm(totfatrte ~ factor(round(bac08)) + factor(round(bac10)) +
    factor(round(perse)) + factor(round(sbprim)) + factor(round(sbsecon)) +
    factor(round(sl70plus)) + factor(round(gdl)) + perc14_24 +
    unem + vehicmilespc + d81 + d82 + d83 + d84 + d85 + d86 +
    d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
    d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04, data = data)
summary.lm.adj(model.extended.lm)
```

The model did produce a statistically significant p-value in the Breusch–Pagan test, leading us to report the model results using heteroskedasticity-robust standard errors in the equation below. Please note that since we are primarily concerned with the coefficients of variables other than the year indicators in this section, those coefficients are not directly reported in the model equation shown here, but are instead abbreviated by $\vec{\beta}_{OLS} \cdot \vec{d}_{years}$. To see all coefficients associated with the all variables in the pooled OLS model, please refer to the table in section IV.

$$\begin{aligned} totfatrte = &-\underset{(2.6958)}{2.8262} - \underset{(0.4460)^{***}}{2.1944} \cdot bac08 - \underset{(0.3677)^{***}}{1.2379} \cdot bac10 - \underset{(0.2706)^{*}}{0.6499} \cdot perse - \underset{(0.4638)}{0.0942} \cdot sbprim + \underset{(0.4159)}{0.0643} \cdot sbsecon \\ &+ \underset{(0.4010)^{***}}{3.2389} \cdot sl70plus - \underset{(0.4762)}{0.3476} \cdot gdl + \underset{(0.1234)}{0.1401} \cdot perc14\_24 + \underset{(0.0853)^{***}}{0.7675} \cdot unem + \underset{(0.00013)^{***}}{0.00293} \cdot vehicmilespc \\ &+ \vec{\beta}_{OLS} \cdot \vec{d}_{years} \end{aligned}$$

For the extended regression model above, we can see that the coefficients for `bac08`, `bac10`, and `sl70plus` are strongly significant with p-values under 0.01, while the coefficient for `perse` is moderately significant with a p-value of 0.016.

The coefficient signs for `bac08` and `bac10` are negative, meaning that when a state has enacted blood alcohol content limiting laws, the total fatality rate has decreased. In contrast, the coefficient sign for `sl70plus` is positive, meaning that if a state has laws allowing a higher speed limit, then the total fatality rate increases. Neither of these effects are surprising, as drunk driving and high speed collisions are two factors that lead to traffic fatality. We see here that the coefficient associated with bac08 (-2.1944) is more negative than that associated with bac10 (-1.2379). This tells us that (1) the enactment of either law is associated with a decreased fatality rate when compared to a state that has no such drunk driving law, and (2) the stricter of the two laws, blood alcohol limit at 0.08, has a larger decrease in fatality rate than the loser `bac10` law. The postive coefficient for `sl70plus` strong suggests that states which allow higher speed limits also experience higher fatality rates, possibly due to the increase probablity of collisions.

The sign of the `perse` coefficient is negative, which is also not suprising. The suspension of administrative licenses for drivers who are caught violating laws would likely lead to more careful drivers. Therefore, we would expect the enactment of a perse law to be associated with a decrease of traffic incidents, and therefore traffic fatalities.

The coefficient for `sbprim` and `sbsecon` which is the indicator for primary seat belt and secondary seat belt laws, respectively, is not statistically significant, with a p-value of 0.84. Thus, we cannot reject the null hypothesis that primary seatbelt laws have no effect on total fatality rates. While seat belts are likely to protect passengers from fatal accidents, these laws may not have the same effect as bac laws or speed limit laws because violation of a seat belt law is harder to detect by law enforcement officials and therefore do not lead to driver punishment in the same way that drunk driving and speed limit laws can.

Other than the year indicators, the coefficients for `unem` and `vehicmilespc` are significant and have positive signs. This means that as unemployment and miles driven per capita increase, the total fatality rate also increases. We expect `vehicmilespc` to be positively associated with fatality rates as higher driving mileage means general increased likelihood of traffic accidents. The positive relationship between `unem` and `totfatrte` is likely not as direct, but due to some other causal factor, such as state economy, that could lead to both

increasing. However, we find out in Part IV, that the positive relationship observed here may not be very reliable.

Additionally, we run the `plmtest` to check for the presence of *individual* and *time* effects in the OLS Pooled model for the driving panel data.

```
# prepare panel data
driving.panel <- pdata.frame(data, index = c("state", "year"),
    drop.index = TRUE, row.names = TRUE)
```

```
g1 <- plm(totfatrte ~ factor(round(bac08)) + factor(round(bac10)) +
    factor(round(perse)) + factor(round(sbprim)) + factor(round(sbsecon)) +
    factor(round(sl70plus)) + factor(round(gdl)) + perc14_24 +
    unem + vehicmilespc, data = driving.panel, model = "pooling")
plmtest(g1, effect = "individual")
```

```
##
##  Lagrange Multiplier Test - (Honda) for balanced panels
##
## data:  totfatrte ~ factor(round(bac08)) + factor(round(bac10)) + factor(round(perse)) +  ...
## normal = 64.316, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
plmtest(g1, effect = "time")
```

```
##
##  Lagrange Multiplier Test - time effects (Honda) for balanced
##  panels
##
## data:  totfatrte ~ factor(round(bac08)) + factor(round(bac10)) + factor(round(perse)) +  ...
## normal = 14.268, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Without the year indicator variables, we can see that there are significant "individual" and "time" effects (p-value $<$ 2.2e-16) present in the Pooled OLS model. We run the same test after including year indicator variables that should absorb the time fixed effects.

```
g2 <- plm(totfatrte ~ factor(round(bac08)) + factor(round(bac10)) +
    factor(round(perse)) + factor(round(sbprim)) + factor(round(sbsecon)) +
    factor(round(sl70plus)) + factor(round(gdl)) + perc14_24 +
    unem + vehicmilespc + d81 + d82 + d83 + d84 + d85 + d86 +
    d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
    d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04, data = driving.panel,
    model = "pooling")
plmtest(g2, effect = "individual")
```

```
##
##  Lagrange Multiplier Test - (Honda) for balanced panels
##
## data:  totfatrte ~ factor(round(bac08)) + factor(round(bac10)) + factor(round(perse)) +  ...
## normal = 65.116, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
plmtest(g2, effect = "time")
```

```
##
##  Lagrange Multiplier Test - time effects (Honda) for balanced
##  panels
```

```
##
## data:  totfatrte ~ factor(round(bac08)) + factor(round(bac10)) + factor(round(perse)) +  ...
## normal = -3.5729, p-value = 0.9998
## alternative hypothesis: significant effects
```

The results indicate that after we include year indicators, there are no significant time fixed effects (p-value = 0.9998) left in the Pooled OLS model, but there are still statistically significant "individual effects" (state level effects) remaining in the model. This is a strong suggestion that a Fixed Effects model that can control for the state level effects would be more appropriate for the given panel data than an OLS model.

IV. State Level Fixed Effects Modeling

We build a Fixed Effects model to account for the 'state' level effects, including the year indicators to absorb the 'time' effects. Let's compare the coefficients from the Pooled OLS and the Fixed Effects models:

```
plm.fe <- plm(totfatrte ~ factor(round(bac08)) + factor(round(bac10)) +
    factor(round(perse)) + factor(round(sbprim)) + factor(round(sbsecon)) +
    factor(round(sl70plus)) + factor(round(gdl)) + perc14_24 +
    unem + vehicmilespc + d81 + d82 + d83 + d84 + d85 + d86 +
    d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
    d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04, data = driving.panel,
    model = "within")
# all results
pOLS.results = summary.lm.adj(model.extended.lm)
FE.results = summary.plm.adj(plm.fe)
# extracting p-values for significant markers
OLS.p = pOLS.results[2:35, 4]
FE.p = FE.results[, 4]
# include all coefficients from both models intercept will be
# excluded, since FE has no intercept
comparison.table = data.frame(FE.coef = round(FE.results[, 1],
    5), FE.rse = round(FE.results[, 2], 5), FE.p = round(FE.results[,
    4], 5), FE.sig = case_when(FE.p >= 0 & FE.p < 0.001 ~ "***",
    FE.p >= 0.001 & FE.p < 0.01 ~ "**", FE.p >= 0.01 & FE.p <
        0.05 ~ "*", FE.p >= 0.05 & FE.p < 0.1 ~ ".", TRUE ~ ""),
    OLS.coef = round(pOLS.results[2:35, 1], 5), OLS.rse = round(pOLS.results[2:35,
        2], 5), OLS.p = round(pOLS.results[2:35, 4], 5), OLS.sig = case_when(OLS.p >=
        0 & OLS.p < 0.001 ~ "***", OLS.p >= 0.001 & OLS.p < 0.01 ~
        "**", OLS.p >= 0.01 & OLS.p < 0.05 ~ "*", OLS.p >= 0.05 &
        OLS.p < 0.1 ~ ".", TRUE ~ ""))
# intercept row
OLS.intercept = data.frame(FE.coef = NA, FE.rse = NA, FE.p = NA,
    FE.sig = NA, OLS.coef = round(pOLS.results[1, 1], 5), OLS.rse = round(pOLS.results[1,
        2], 5), OLS.p = round(pOLS.results[1, 4], 5), OLS.sig = "")
# add intercept row
comparison.table = rbind(OLS.intercept, comparison.table)
rownames(comparison.table)[1] = "Intercept"
options(knitr.kable.NA = "")
kable(comparison.table)
```

	FE.coef	FE.rse	FE.p	FE.sig	OLS.coef	OLS.rse	OLS.p	OLS.sig
Intercept					-2.82621	2.69577	0.29468	
factor(round(bac08))1	-1.18045	0.60889	0.05279	.	-2.19437	0.44596	0.00000	***
factor(round(bac10))1	-0.86977	0.34176	0.01106	*	-1.23789	0.36773	0.00079	***

	FE.coef	FE.rse	FE.p	FE.sig	OLS.coef	OLS.rse	OLS.p	OLS.sig
factor(round(perse))1	-1.05865	0.39691	0.00776	**	-0.64989	0.27061	0.01648	*
factor(round(sbprim))1	-1.25061	0.55379	0.02412	*	-0.09420	0.46380	0.83908	
factor(round(sbsecon))1	-0.35659	0.36629	0.33051		0.06430	0.41594	0.87716	
factor(round(sl70plus))1	-0.03244	0.52364	0.95061		3.23891	0.40100	0.00000	***
factor(round(gdl))1	-0.30503	0.37490	0.41603		-0.34762	0.47619	0.46554	
perc14_24	0.19367	0.17466	0.26772		0.14010	0.12338	0.25641	
unem	-0.57652	0.12266	0.00000	***	0.76749	0.08530	0.00000	***
vehicmilespc	0.00093	0.00035	0.00738	**	0.00293	0.00013	0.00000	***
d81	-1.51238	0.45294	0.00087	***	-2.18402	1.32710	0.10009	
d82	-3.05403	0.50577	0.00000	***	-6.65721	1.23718	0.00000	***
d83	-3.66381	0.52593	0.00000	***	-7.58904	1.14875	0.00000	***
d84	-4.39985	0.46759	0.00000	***	-5.97447	1.12374	0.00000	***
d85	-4.86034	0.48448	0.00000	***	-6.60315	1.14077	0.00000	***
d86	-3.76923	0.61034	0.00000	***	-5.94667	1.20870	0.00000	***
d87	-4.41235	0.69948	0.00000	***	-6.45877	1.23512	0.00000	***
d88	-4.88774	0.79533	0.00000	***	-6.69054	1.25484	0.00000	***
d89	-6.23948	0.90618	0.00000	***	-8.15884	1.32504	0.00000	***
d90	-6.35637	0.94858	0.00000	***	-9.05967	1.35866	0.00000	***
d91	-7.04423	1.01658	0.00000	***	-11.20604	1.36374	0.00000	***
d92	-7.89053	1.11616	0.00000	***	-12.99595	1.39679	0.00000	***
d93	-8.23657	1.15146	0.00000	***	-12.88173	1.39360	0.00000	***
d94	-8.68229	1.13506	0.00000	***	-12.52992	1.40417	0.00000	***
d95	-8.38889	1.22501	0.00000	***	-12.03327	1.45293	0.00000	***
d96	-8.76480	1.19816	0.00000	***	-14.02527	1.43828	0.00000	***
d97	-8.91637	1.25870	0.00000	***	-14.30415	1.47964	0.00000	***
d98	-9.53329	1.27110	0.00000	***	-15.11958	1.48246	0.00000	***
d99	-9.69404	1.38621	0.00000	***	-15.18480	1.48877	0.00000	***
d00	-10.22347	1.36678	0.00000	***	-15.54436	1.53450	0.00000	***
d01	-9.96079	1.49052	0.00000	***	-16.44872	1.53812	0.00000	***
d02	-9.25456	1.49587	0.00000	***	-17.02795	1.57188	0.00000	***
d03	-9.32704	1.53521	0.00000	***	-17.41791	1.57822	0.00000	***
d04	-9.66760	1.69059	0.00000	***	-16.97948	1.61445	0.00000	***

Fixed Effects model does not have Intercept because the unobserved time-constant effects for States are removed by the Fixed Effects transformation. The Intercept in the Pooled OLS represents the base effect of all the States on the total fatality rate, which is not reliable because it assumes that the effect of all the states on traffic fatality rate is constant, and it does not capture the difference of effects among different States.

We can see that the coeffcients of `bac08`, `bac10`, `perse`, and `sbprim` are all negatively related to the traffic fatality rate which means that the fatality rate decreases with the enforcement of these laws.

Estimated effect of `bac08` in the Pooled OLS is -2.1944 and is statistically significant, whereas its estimator in the Fixed effects model is -1.1805 and is marginally statistically significant. Estimated coefficient of 'bac10' in the Pooled OLS is -1.2379 and is highly statistically significant, whereas its estimator in the Fixed effects model is -0.86977 and is moderately statistically significant. The effect of `bac08` and `bac10` are both greater in magnitude in the OLS model than the Fixed Effects Model. However, the trend we see in the Fixed Effects model is the same as that of the pooled OLS model, which is that the magnitude of the effect of the stricter `bac08` law is greater than that of the `bac10` law.

The estimated effect of the `perse` is greater in magnitude in the Fixed effects model, with a statistically significant value of -1.0587 compared to the smaller and moderately significant value of -0.6499 from the pooled OLS model. The difference in the estimated effect of `sbprim` between the OLS and Fixed effect models is alarmingly large, as the pooled OLS produces an estimate of not significant -0.094205 while the Fixed

Effects model produces a moderately significant estimate of -1.2506, which is two orders of magnitude larger.

Another outstanding difference between the coefficient of `unem` between the two models, as it has changed signs. In the OLS Pooled model, it is 0.767 whereas in the Fixed effects model, it is -0.577. In the OLS pooled model it is positively related to the fatality rate, meaning the estimated traffic fatality would increase with an increase in the unemployment rate. This relationship is opposite in the Fixed effects model; the estimated traffic fatality rate would decrease with an increase in the unemployment. We think that the estimator of the Fixed effects model makes more sense as one possible argument is that there would be more vehicles on road with more economic growth and lower unemployment, thereby increasing the chances of accidents and fatalities.

The estimators of `perc14_24` are positively related to the fatality rate, and are not statistically significant in both the models. The coefficients of `vehicmilespc` are statistically significant in both the models, and are also positively related to the traffic fatality rate. The magnitude of effect of this variable is greater in the OLS pooled model as compared to the Fixed Effects model, but both are very very small.

Finally, we consider the assumptions for both the OLS model and the Fixed Effects model and assess the validity of these assumptions below.

For the Pooled OLS:

1. It is assumed that all the observations are independent. However, in the current context, the observations collected over years for a particular State may not be independent as the unobserved fixed effects of that State would make the observations correlated to each other.
2. It is also assumed that all the observations across all time periods are identically distributed. In the current context, there might be year specific effects that could make the distribution of observations over years different. An example of that could be the development of safety features in vehicles in a particular time period affecting driving safety in all the States. By adding year indicator varibles in the pooled OLS, we overcame this problem and this assumption holds good.
3. Another assumption is that the composite errors (unobserved fixed effects + idiosyncratic errors) are uncorrelated with each of the explanatory variables across all time periods, and there is no serial correlation in the composite error. However, this exogeneity assumption does not hold good in the current context because there are some unobserved fixed effects specific to each State that might be correlated with explanatory variables. Due to these unobserved fixed effects, the composite errors are also serially correlated.

For the Fixed Effects model:

1. It is assumed that all the observations are independent. Fixed effects model removes the unobserved fixed effects making the observations independent.
2. It is also assumed that all the observations across all time periods are identically distributed. By adding the indicator variables for years, this assumption is upheld.
3. Strict exogeneity assumption is needed, that is the idiosyncratic errors are uncorrelated with each explanatory variable across all time periods. In the current context, since the unobserved fixed effects of each State which are time-constant are eliminated by the fixed effects transformation, the assumption of non-correlation between the idiosyncratic errors with explanatory variables holds well.

In conclusion, we believe that the coefficients estimated by the Fixed Effects model are more reliable, consistent and unbiased because that model controls for the unobserved fixed effects of states taking into account the heterogeneity across States. Furthermore, more of the Fixed Effects model assumptions are met by the panel data transformation and year indicator variable inclusion than the assumptions of the OLS models. Our assertion here is also supported by the second of the two plmtests ran in section III.

V. Random Effects vs Fixed Effects Model

Let's first run 'Hausman Test' to compare the Fixed effects and Random effects models, where the failure to reject the null hypothesis would support the Random Effects model. The output of the model coefficients are errors are suppressed here, but they can be seeing in our .rmd file. Errors have been adjusted according to the existence of heteroskedasticity and/or serial correlation.

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

```
phtest(plm.fe, plm.random)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ factor(round(bac08)) + factor(round(bac10)) + factor(round(perse)) +  ...
## chisq = 164.12, df = 34, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

From the above test results, p-value < 2.783e-06, is highly statistically significant, which means a Fixed Effects model is preferred in this case.

In general, a Random Effects model is considered more appropriate when there is no correlation between the unobserved fixed effects (of states) and the explanatory variables. We believe that that is not true for the case of our panel data. There are unobserved fixed effects of States that have correlation with other explanatory variables of interest. For instance, the amount of alcohol that people consume in a particular state might be correlated to its bac08/bac10 laws, or the infrastructure, population or the size of the state might be determining the number of miles driven per capita.

A Random effects model is more useful when key explanatory variables are constant over time. However, from our EDA, many of the key expanatory variables included in our model show high variation over time. In this case, the Fixed Effects model is best as it allows us to study the effects of these time-varying variables, as they are not swept away by the Fixed Effects transformation. Additionally, we would want to use random effects model when the size of the sample is very small as compared to the population size. In this case, we have the data on 48 states of the US, which is very close to the population size that is 50 US states. So, for this study, Fixed Effects model is ultimately more appropriate.

VI. Total Fatality Rate based on Vehicle Miles Driven per Capita

As we noted above that the Fixed Effects model is model appropriate for the given panel data, we will use the estimator from the Fixed Effects model to compute the estimated effect of *vehicmilespc* on *totfatrte*.

```
plm.fe$coefficients[10] * 1000
```

```
## vehicmilespc
##    0.9261162
```

We can interpret that with an increase of 1,000 in the number of miles driven per capita, *totfatrte* would increase by 0.9261162, or nearly 1 additional traffic fatality amongst 100,000 people, while holding all the other variables in the model constant.

VII. Serial Correlation or Heteroskedasticity Discussion

For fixed effects with serial correlation and heteroskedasticity in idiosyncratic errors, having all other Fixed Effects model assumptions hold (Section V), estimators would still be unbiased but would no longer be the best linear unbiased estimator. This means that unadjusted standard errors in these models are understated and we would not be able to trust the p-values and the confidence intervals of estimators, as the significance of the t-statistics calculated from these understated errors will become unreliable.

For random effects, it is expected that the composite errors ν_{it} will be serially correlated across time and is taken into account in the model (Wooldridge 2016, Chapter 14). This serial correlation in composite errors has no negative effects on the random effects model. In contrast, if heteroskedasticity is found, the random effects standard errors (and by extension the test statistics) are underestimated and invalid.

Below, we use the Breusch-Pagan test to detect heteroskedasticity in both the Fixed Effects and Random Effects plm models. We also use the Breusch-Godfrey/Wooldridge test to detect the existence of serial correlation in the idiosyncratic errors of both models.

```
# test for heteroskedasticity
bptest(plm.fe)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  plm.fe
## BP = 144.26, df = 34, p-value = 1.553e-15
```

```
bptest(plm.random)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  plm.random
## BP = 144.26, df = 34, p-value = 1.553e-15
```

```
# run serial correlation tests for both the models
pbgtest(plm.fe)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  totfatrte ~ factor(round(bac08)) + factor(round(bac10)) + factor(round(perse)) +     factor(r
## chisq = 336.34, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(plm.random)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  totfatrte ~ factor(round(bac08)) + factor(round(bac10)) + factor(round(perse)) +     factor(r
## chisq = 376.81, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

We note from the above test results that we can reject both null hypotheses and conclude that there is both heteroskedasticity and serial correlation in the idiosyncratic errors of both the Fixed effects and Random effects models. This means that had we used unadjusted standard errors to report the model results, we would

have had understated errors. Given these test results, the standard errors of our models must be recalculated to have adjustements that are robust to both the existence of heteroskedasticity and serial correlation.

Our results in sections IV and V for the Fixed Effects and Random Effects models, respectively, are reported with heteroskedasticity-robust and serial correlation robust standard errors. The function we used to report these results is defined in the "Functions" Section immediately following the Introduction. For both plm's the function that adjusts their errors checks for both heteroskedasticity and serial correlation. If only heteroskedasticity is detected without serial correlation, adjusted heterskedasticity-robust standard errors, using `vcovHC()` method "white1" are reported. If both heteroskedasticity and serial correlation are both dectect, then clustered standard errors for clusters defined by `group = state` are used. These adjusted errors are larger than those for heteroskedasticity alone, as they use the `vcovHC()` method "arellano".

Conclusion

In this study, we examined a panel dataset describing traffic fatality rate, traffic laws and a few other economic and demographic measures for 48 US states over the period of 25 years. We attempted to fit various models in order to determine how the total traffic fatality rate changes over time for different US states, and which other factors may significantly affect the fatality rate. It is important to note that the models built in this study is only used to detect certain effects but cannot be used to predict future fatality rates, as our models were not designed with assessment of best fit in mind.

We started out with building an OLS Pooled model, but found evidence that the panel data had both 'individual' as well as 'time' effects, and a Fixed Effects model would be more appropriate to fit the given panel data. We also built a Random Effects model, and the comparison of this model against the Fixed effects model by conducting a Hausman Test showed that the Fixed effects model was a better fit in this case.

From our analysis we learned that the average traffic fatality rate declined over the years. Coefficients of the Fixed Effects model suggested that the enforcement of driving laws such as 'blood alcohol content limit', 'per se' and 'seat belt laws' are important for the model, and significantly impacted traffic fatalities rate. We also detected that economic growth, lower unemployment rate and larger number of vehicle miles driven per capita contribute significantly toward the traffic fatalities.