Lab 5

Ram Balasubramanian, John Kenney

December 14, 2017

Read and Examine the data: Load the data.

```
load("driving.RData")
traffic = data
# View(desc) View(traffic) table(with(data,state,year))

# hist(traffic$minage) length(unique(traffic$minage))

# hist(traffic$zerotol) length(unique(traffic$zerotol))
```

Question 1

Provide a description of the basic structure of the dataset, as we have done throughout the semester. Be sure your explanation includes what the variable means and how it is coded. **********

1.1 Description of Dataset

This data is structured as a long panel set, where each of the 48 continental states are numbered alphabetically from 1 to 51, with 2, 9, and 12 missing (Alaska, Hawaii, District of Columbia). Each state has associated with it 25 observations ranging from 1980 to 2004. The year is indicated both as its own variable and represented as one of 25 dummy variables. Within each Year-by-State observation there are observations that describe the state's traffic laws, traffic Fatalities, and population demographics.

Variables that are coded as dummy (1 or 0) will often show a number between these two dichotomous options. This indicates that the state's law was changed during this year and the fraction indicates the portion of the year for which the variable was active. For example, a value of 0.75 indicates that for three quarters of the year, the variable was $True$ and for the other quarter, the variable was $False$.

1.1.1 Traffic Laws

1.1.1.1 Speed Limit

There are six dummy variables starting with *sl*, which indicate the speed limit mandated by the state for the year. The first four variables code speed limits in 5 mph increments from 55 to 75 mph. The fifth variable *slnone* indicates there was no speed limit for the state that year. The sixth variable *sl*70*plus* indicates that either the speed limit was 70 mph or greater, or that there was no speed limit that year.

1.1.1.2 Seatbelts

The next variable *seatbelt* is categorical and describes the type of seat belt law that exists, "0" if no law, '1' if primary (no other violation required to give a ticket), "2" if secondary (another violation must have occurred for the officer to issue a seatbelt ticket). There also exist two dummy variables starting with *sb*, one for primary and the other for secondary.

1.1.1.3 Drinking

The variables *minage*, *zerotol*, and *bac* describe the state's approach to drinking laws. The *minage* variable is the state's legal drinking age for the year, taking on 12 distinct values from ranging from 18 to 21, with 21 making up the great majority of observations. Non-integer observations indicate a year of

The *zerotol* variable indicates if the state enacted a Zero Tolerance law for drinking, which makes it a criminal DUI offense for drivers under the age of 21 to drive with even a small amount of alcohol in their system. This variable takes on 11 distinct values, ranging from 0 to 1, with zero and 1 making up the vast majority of observations

The next two *bac* dummy variables indicate if the state's acceptable BAC is 0.08% or 0.10%.

Several states have adopted *perse* laws which allows for suspension or revocation of driver's license for DUI or DWI cases. The *perse* dummy variable indicates to the proportion of the year the state had this type of law enacted.

1.1.2 Fatality Statistics

Fatality statistics are given for each State-by-Year observation and include gross totals as well as totals normalized by vehicle miles driven by the state and per capita. These statistics are reported in terms of the time of occurrence, which include Fatalities over all times, Fatalities at nighttime, and Fatalities during the weekend.

1.1.3 Demographics

Observations also report a number of demographics specific to each state for each year reported. These include the number of vehicle miles in billions by the state's population for the year (total and per capita), the state's percent unemployment, and the percentage of the state's population between 14 and 24, inclusive.

1.2 Exploratory Data Analysis

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.

Remember, graphs must be well-labeled.

You need to write a detailed narrative of your observations of your EDA.

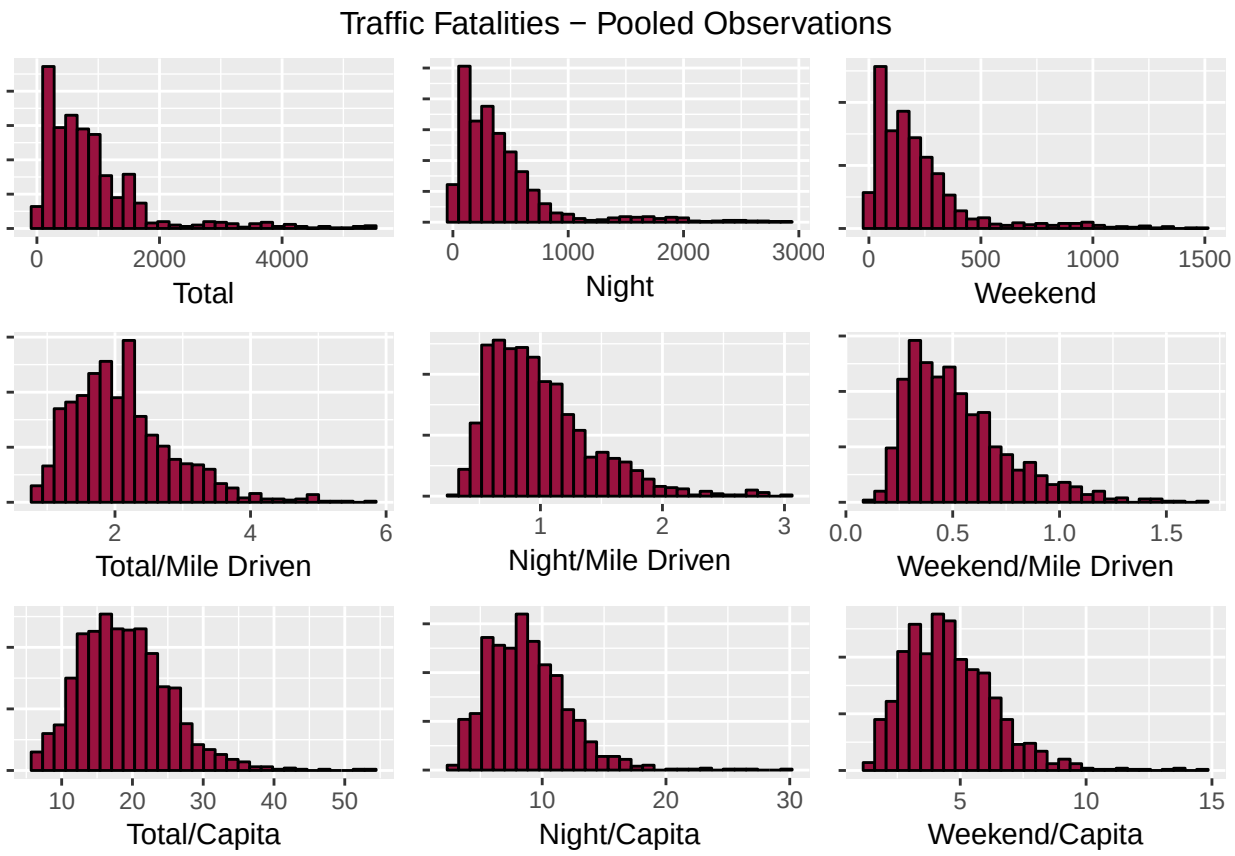
Reminder: giving an "output dump" (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive zero point in this exercise. **********

1.2.1 Univariate EDA

1.2.1.1 Dependent Variables - Fatalities

```
h = geom_histogram(aes(y = ..count..), bins = 30, fill = "#99123F", colour = "black")
t = theme(plot.title = element_text(lineheight = 1, face = "bold"), axis.text.y = element_blank(),
    axis.title.y = element_blank())

plot.hist1 = ggplot(traffic, aes(x = totfat)) + scale_x_continuous(name = "Total") +
    h + t
plot.hist2 = ggplot(traffic, aes(x = nghtfat)) + scale_x_continuous(name = "Night") +
    h + t
```

```
plot.hist3 = ggplot(traffic, aes(x = wkndfat)) + scale_x_continuous(name = "Weekend") +
    h + t + t
plot.hist4 = ggplot(traffic, aes(x = totfatpvm)) + scale_x_continuous(name = "Total/Mile Driven") +
    h + t
plot.hist5 = ggplot(traffic, aes(x = nghtfatpvm)) + scale_x_continuous(name = "Night/Mile Driven") +
    h + t + t
plot.hist6 = ggplot(traffic, aes(x = wkndfatpvm)) + scale_x_continuous(name = "Weekend/Mile Driven") +
    h + t + t
plot.hist7 = ggplot(traffic, aes(x = totfatrte)) + scale_x_continuous(name = "Total/Capita") +
    h + t + t
plot.hist8 = ggplot(traffic, aes(x = nghtfatrte)) + scale_x_continuous(name = "Night/Capita") +
    h + t + t
plot.hist9 = ggplot(traffic, aes(x = wkndfatrte)) + scale_x_continuous(name = "Weekend/Capita") +
    h + t
grid.arrange(plot.hist1, plot.hist2, plot.hist3, plot.hist4, plot.hist5,
    plot.hist6, plot.hist7, plot.hist8, plot.hist9, nrow = 3, ncol = 3,
    top = quote("Traffic Fatalities - Pooled Observations"))
```




```
shapiro.test(traffic$totfat)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  traffic$totfat
## W = 0.75, p-value <2e-16
```

```r
shapiro.test(traffic$totfatpvm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  traffic$totfatpvm
## W = 0.94, p-value <2e-16
```

```r
shapiro.test(traffic$totfatrte)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  traffic$totfatrte
## W = 0.97, p-value = 2e-15
```

```r
shapiro.test(log(traffic$totfat))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(traffic$totfat)
## W = 0.98, p-value = 1e-11
```

```r
shapiro.test(log(traffic$totfatpvm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(traffic$totfatpvm)
## W = 1, p-value = 0.04
```

```r
shapiro.test(log(traffic$totfatrte))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(traffic$totfatrte)
## W = 0.99, p-value = 9e-07
```

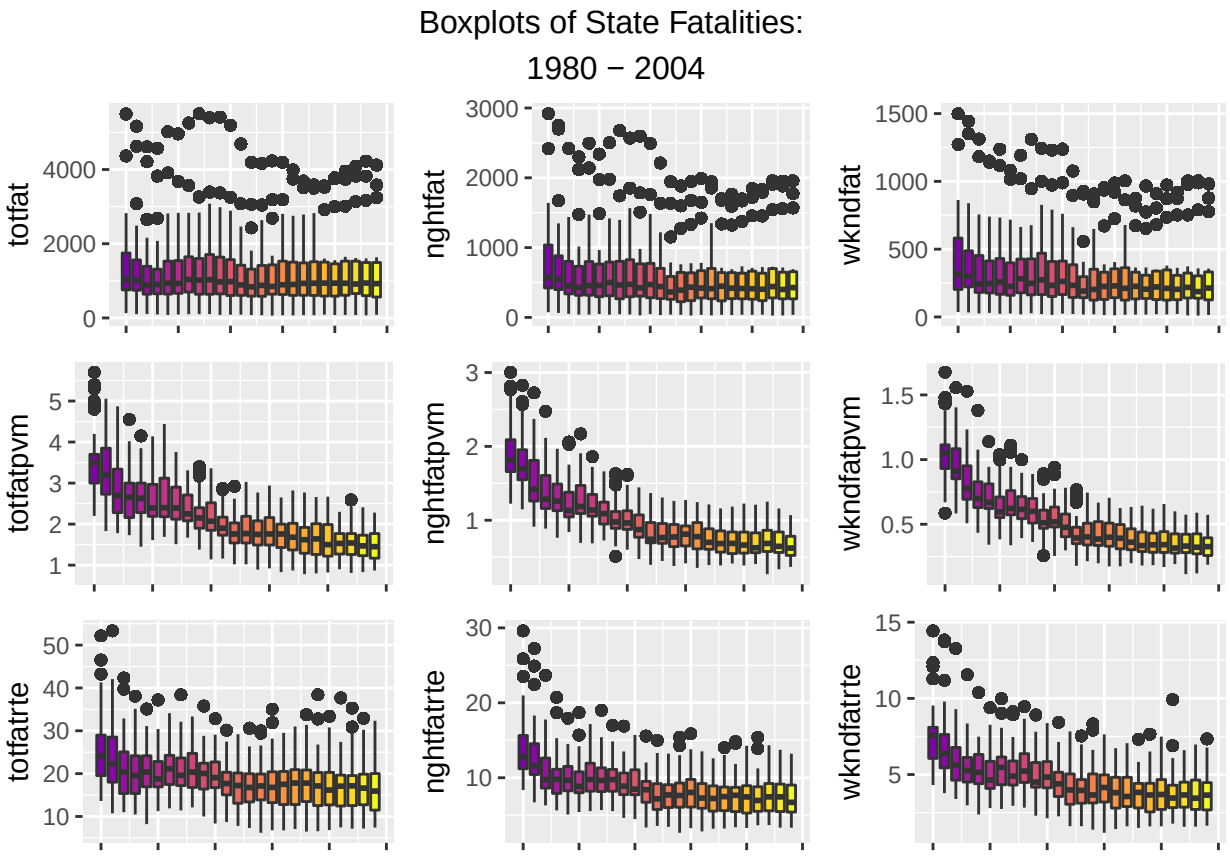
Observations: We see that pooling the State-Year observations undercovers meaningful information about the underlying trends of the fatality variables. Skew is most evident for the unnormalized fatality rates (Total, Night, and Weekend) and least evident when the data is normalized by the state's population.

We see that virtually every variable is positively skewed, which means a log transformation may improve normality. When considering normalization techniques (none, per vehicle mile, per capita), it appears that $totfatpvm$ responds most favorably to the log transformation.

Univariate Analysis of DVs by Year

```r
# Shape files for each state
states = map_data("state", projection = "albers", parameters = c(39, 45))

# Associate State Index with State Name
statenames = unique(states$region[])
statenames = c(statenames[1], "alaska", statenames[2:10], "hawaii", statenames[11:length(statenames)])
```

```r
states = states[states$region != "district of columbia", ]
states$state = match(states$region, statenames)
# Associate the state shapes with associated observations
traffic.map = merge(states, traffic, by = "state")


b = geom_boxplot(aes(fill = year, group = year))
t = theme(plot.title = element_text(lineheight = 1, face = "bold"), legend.position = "none",
    axis.text.x = element_blank(), axis.title.x = element_blank())

library(viridis)
c = scale_fill_viridis(option = "plasma", begin = 0.25)

plot.bp1 = ggplot(traffic.map, aes(year, totfat)) + b + t + c
plot.bp2 = ggplot(traffic.map, aes(year, nghtfat)) + b + t + c
plot.bp3 = ggplot(traffic.map, aes(year, wkndfat)) + b + t + c
plot.bp4 = ggplot(traffic.map, aes(year, totfatpvm)) + b + t + c
plot.bp5 = ggplot(traffic.map, aes(year, nghtfatpvm)) + b + t + c
plot.bp6 = ggplot(traffic.map, aes(year, wkndfatpvm)) + b + t + c
plot.bp7 = ggplot(traffic.map, aes(year, totfatrte)) + b + t + c
plot.bp8 = ggplot(traffic.map, aes(year, nghtfatrte)) + b + t + c
plot.bp9 = ggplot(traffic.map, aes(year, wkndfatrte)) + b + t + c
grid.arrange(plot.bp1, plot.bp2, plot.bp3, plot.bp4, plot.bp5, plot.bp6,
    plot.bp7, plot.bp8, plot.bp9, nrow = 3, ncol = 3, top = quote("Boxplots of State Fatalities: \n1980
```



Observations:

Total: Top outliers look to have decreased over time, but the national average of gross fatalities looks almost unchanged over time.

PVM: Significant decline over time in average rate per vehicle mile driven. Variance looks to be proportional to mean, also decreasing over time.

RTE: Significant decline over time in average fatalities per capita. Variance does not look to be proportional to mean.

Univariate Analysis of DVs by State

```
b = geom_boxplot(aes(fill = state, group = state))
t = theme(plot.title = element_text(lineheight = 1, face = "bold"), legend.position = "none",
    axis.text.x = element_blank(), axis.title.x = element_blank())
library(viridis)
c = scale_fill_viridis(option = "plasma", begin = 0.25)

plot.bp1 = ggplot(traffic.map, aes(state, totfat)) + b + t + c
plot.bp2 = ggplot(traffic.map, aes(state, nghtfat)) + b + t + c
plot.bp3 = ggplot(traffic.map, aes(state, wkndfat)) + b + t + c
plot.bp4 = ggplot(traffic.map, aes(state, totfatpvm)) + b + t + c
plot.bp5 = ggplot(traffic.map, aes(state, nghtfatpvm)) + b + t + c
plot.bp6 = ggplot(traffic.map, aes(state, wkndfatpvm)) + b + t + c
plot.bp7 = ggplot(traffic.map, aes(state, totfatrte)) + b + t + c
plot.bp8 = ggplot(traffic.map, aes(state, nghtfatrte)) + b + t + c
plot.bp9 = ggplot(traffic.map, aes(state, wkndfatrte)) + b + t + c
grid.arrange(plot.bp1, plot.bp2, plot.bp3, plot.bp4, plot.bp5, plot.bp6,
    plot.bp7, plot.bp8, plot.bp9, nrow = 3, ncol = 3, top = "Boxplots of Annual Fatalities by State")
```

Observations: There is good agreement between state and rates when considering timing of the incident (overall, at night, and during the weekend).

Total Fatalities: Variance is proportional to the mean of the state, which implies using these variables could result in significant heteroskedasticity. There are also a few outliers that would significantly affect the results (likely California and Texas).

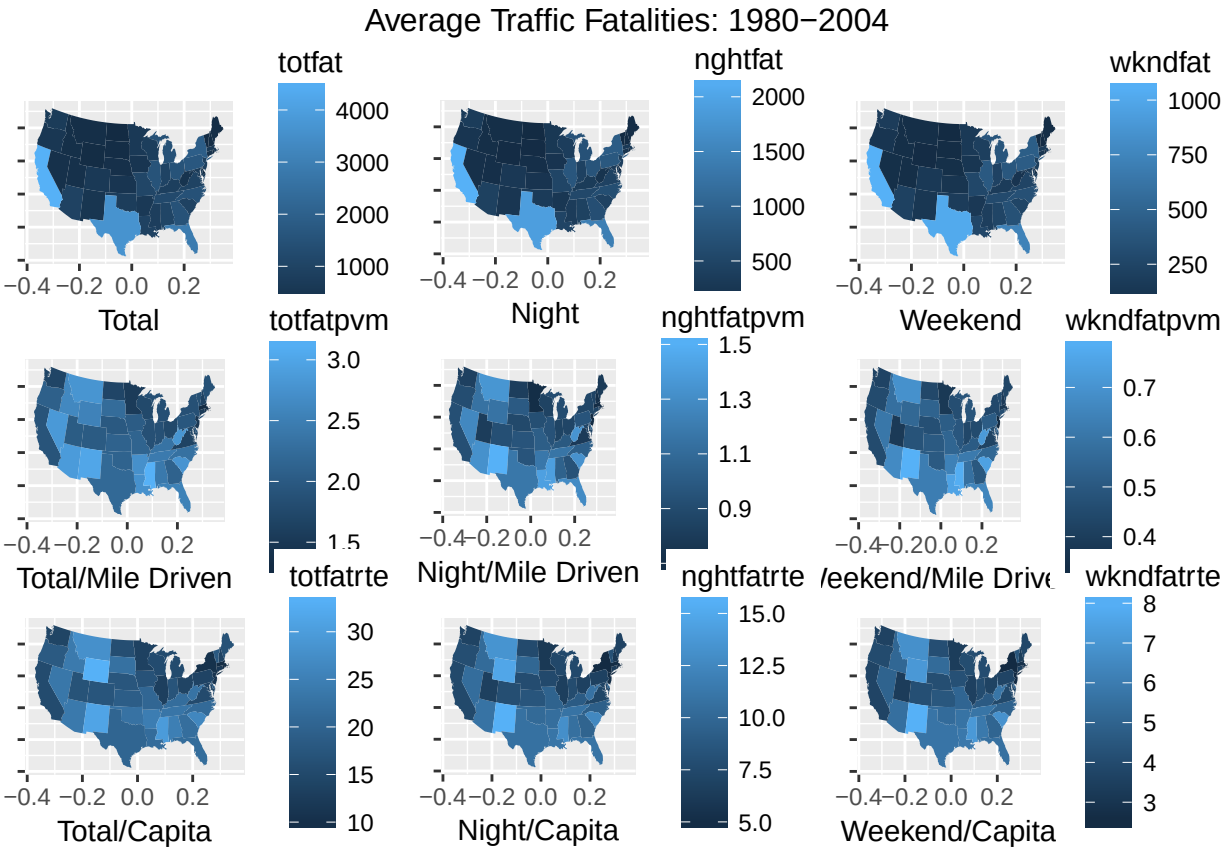
Per Vehicle Mile: Observations look constant with similar variance and no major outliers that could influence linear modeling.

Per Capita: No states that seem to be outliers, variance is not constant between states, nor a function of average.

Univariate Choropleths of DVs

```
states = map_data("state", projection = "albers", parameters = c(39, 45))
statenames = unique(states$region[])
statenames = c(statenames[1], "alaska", statenames[2:10], "hawaii", statenames[11:length(statenames)])
states = states[states$region != "district of columbia", ]
states$state = match(states$region, statenames)
traffic.map = merge(states, traffic, by = "state")

no_var = !names(traffic.map) %in% c("year", "region", "subregion", "lat",
    "long", "order", "group")
traffic.state.agg = aggregate(traffic.map[, no_var], list(traffic.map$state),
    mean)
traffic.map.agg = merge(states, traffic.state.agg, by = "state")
```

```
t = theme(plot.title = element_text(lineheight = 1, face = "bold"), axis.text.y = element_blank(),
    axis.title.y = element_blank())

plot.map1 = qplot(long, lat, data = traffic.map.agg, geom = "polygon",
    fill = totfat, group = group) + labs(x = "Total") + t
plot.map2 = qplot(long, lat, data = traffic.map.agg, geom = "polygon",
    fill = nghtfat, group = group) + labs(x = "Night") + t
plot.map3 = qplot(long, lat, data = traffic.map.agg, geom = "polygon",
    fill = wkndfat, group = group) + labs(x = "Weekend") + t
plot.map4 = qplot(long, lat, data = traffic.map.agg, geom = "polygon",
    fill = totfatpvm, group = group) + labs(x = "Total/Mile Driven") +
    t
plot.map5 = qplot(long, lat, data = traffic.map.agg, geom = "polygon",
    fill = nghtfatpvm, group = group) + labs(x = "Night/Mile Driven") +
    t
plot.map6 = qplot(long, lat, data = traffic.map.agg, geom = "polygon",
    fill = wkndfatpvm, group = group) + labs(x = "Weekend/Mile Driven") +
    t
plot.map7 = qplot(long, lat, data = traffic.map.agg, geom = "polygon",
    fill = totfatrte, group = group) + labs(x = "Total/Capita") + t
plot.map8 = qplot(long, lat, data = traffic.map.agg, geom = "polygon",
    fill = nghtfatrte, group = group) + labs(x = "Night/Capita") + t
plot.map9 = qplot(long, lat, data = traffic.map.agg, geom = "polygon",
    fill = wkndfatrte, group = group) + labs(x = "Weekend/Capita") + t
grid.arrange(plot.map1, plot.map2, plot.map3, plot.map4, plot.map5, plot.map6,
    plot.map7, plot.map8, plot.map9, nrow = 3, ncol = 3, top = quote("Average Traffic Fatalities: 1980-
```



Observations:

TOTAL: California, Texas, and Florida have the highest overall total fatalities averaged over all years observed. It's apparent that these observations are outliers and will skew our modeling.

PVM: Evidence of regional influence. Significantly higher rates just inland from West Coast and in Lousisiana/Mississippi. Perhaps State GDP per Capita might be an unexplained factor?

RTE: Wyoming and Montana show higher per capita rates. There is some regional influence but probably less spatial correlation as compared to pvm.

Overall Univariate Conclusions: The $totfatrte$ variable is not a bad choice to use as our dependent variable. Of the variables available, it looks most normally distributed when pooled, doesn't have many influential outliers, nor does there seem to be regional correlation, which could contribute to omitted variable bias. There also doesn't look to be much serial correlation or heteroskedasticity over time, which means we may be able to get reasonable results from a pooled OLS.

1.2.2 Bivariate EDA

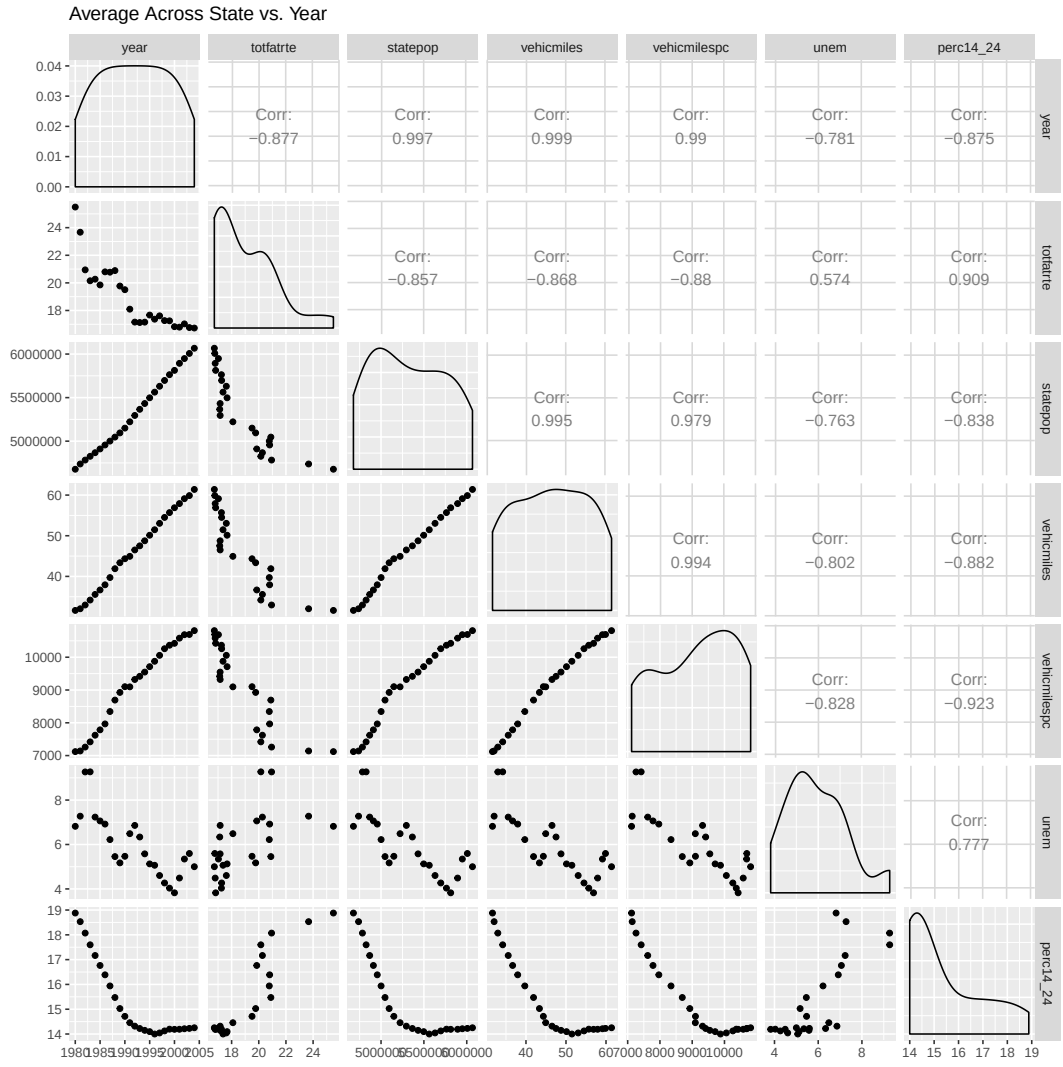
Average of State Observations Over Time

Regressor Interactions: Total Fatality Rate and Demographics

State-Aggregates Over Time

```
# ```{r}
mean_by_year = aggregate(traffic[, c("totfatrte", "statepop", "vehicmiles",
```

```
    "vehicmilespc", "unem", "perc14_24")], traffic["year"], FUN = mean)
ggpairs(mean_by_year, size = 12) + ggtitle(label = "Average Across State vs. Year")
```

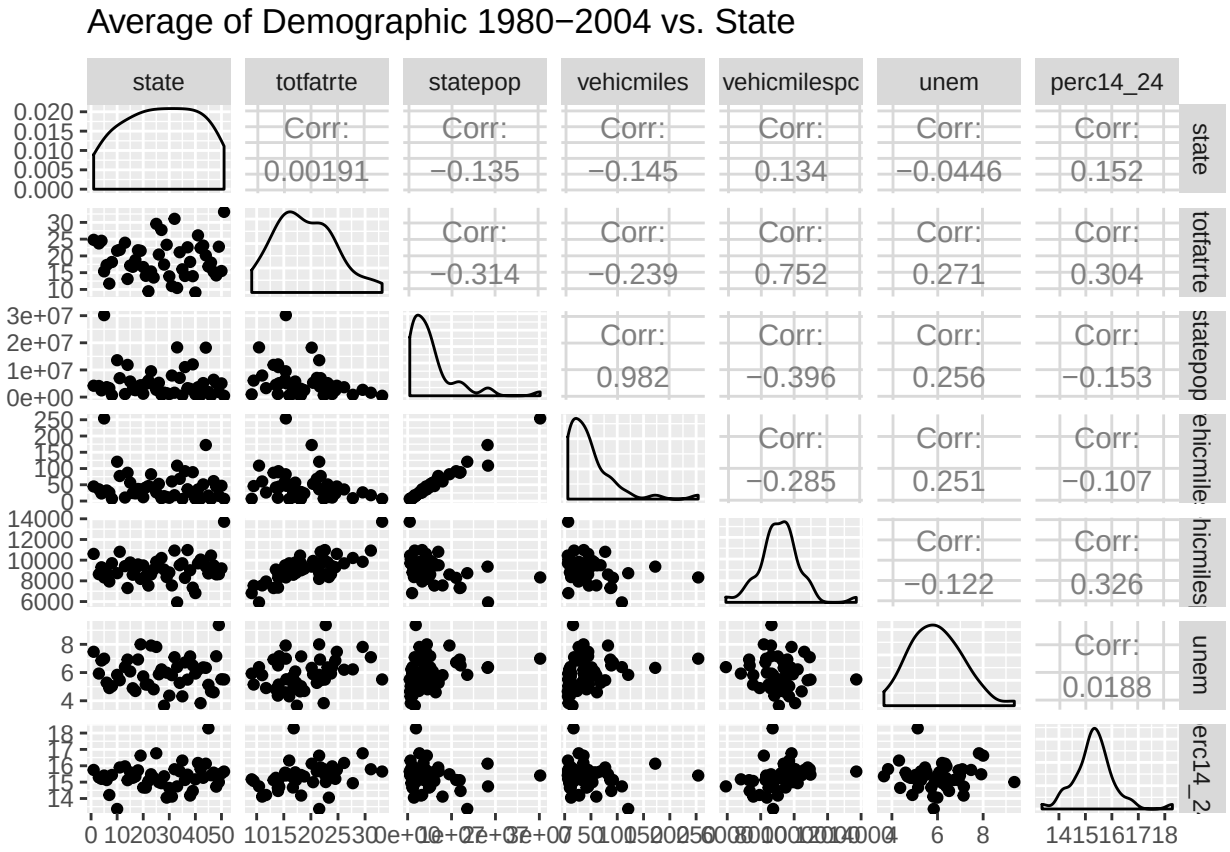


We see significant time dependence between each variable and time when we average all state observations. The fatality rate goes down, population increases and gross vehicle miles driven increase. Unemployment is cyclical, but shows a steady decline overall.

Miles driven per capita increases, which is not necessarily expected. This could be attributed to the fact that cars have become more affordable over time and are therefore more accessible. Percent 14_24 decreases, which indicate an aging population and might mean there are more experienced drivers and therefore fewer accidents. For this reason, this might be a good variable to include in the final model. Beyond this, when comparing scatters of any two variables, we cannot consider their covariance as significant due to spurious correlation.

Year-Aggregates By State

```
mean_by_state = aggregate(traffic[, c("totfatrte", "statepop", "vehicmiles",
    "vehicmilespc", "unem", "perc14_24")], traffic["state"], FUN = mean)
ggpairs(mean_by_state, size = 12) + ggtitle(label = "Average of Demographic 1980-2004 vs. State")
```



Observations When considering average of demographic variables from 1980-2004 vs. State, we see the most significant correlation between vehicle miles driven and population. This is not surprising, since more people to drive means more miles driven.

A significant observation is the correlation between vehicle miles per capita and fatality rate. This makes sense because more time on the road means greater chance of accident.

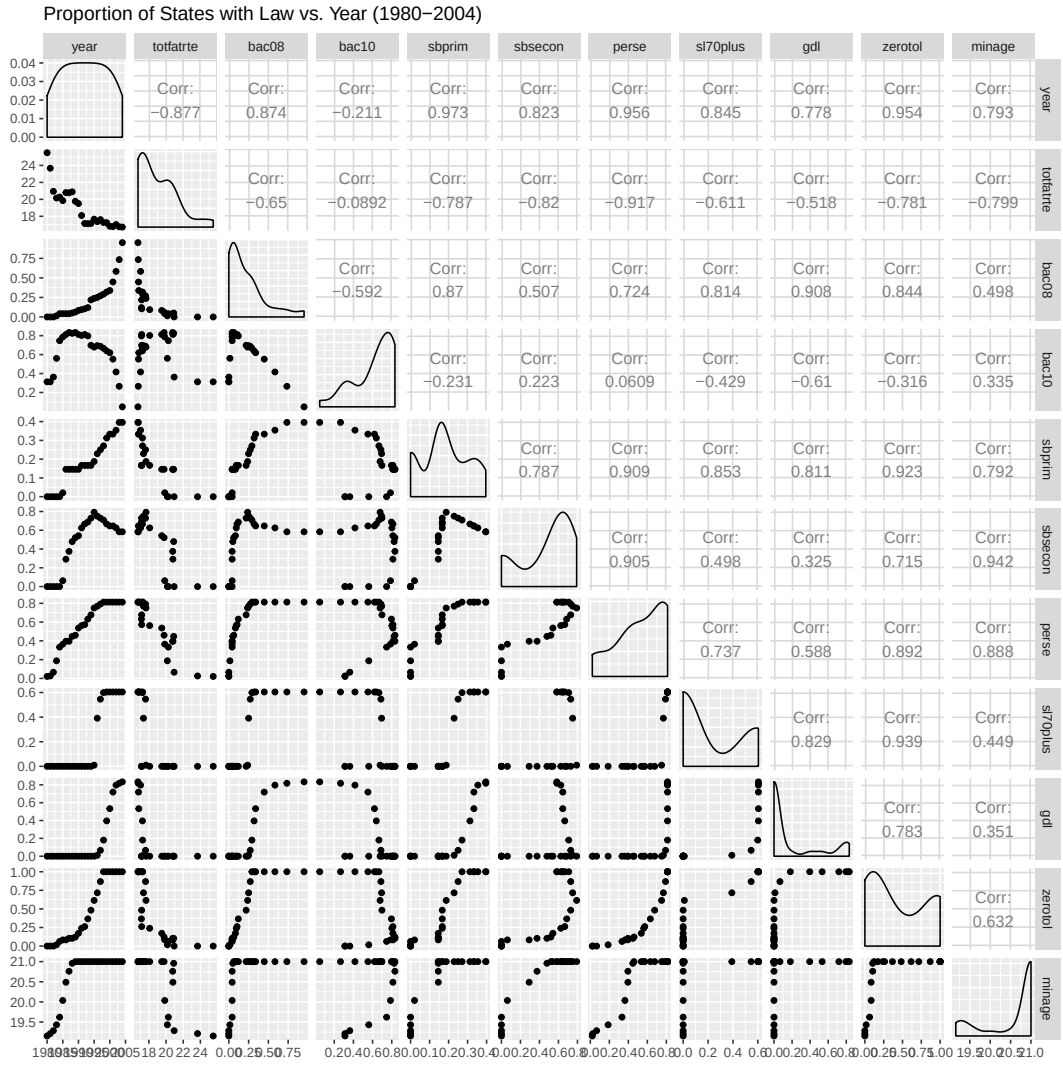
Negative correlation between state population and totfatrte. This is unexpected, but might be explained by more people in a state, which means more tax revenue, more investment in safe driving infrastructure. We see that vehicle miles pc go down with state population as well, meaning that more populous states tend to travel on the road less, thereby putting themselves at lower risk of fatality.

We see a medium correlation with youth and totfatrte. Certainly, the higher the proportion of youth, the more likely we are to have youth accidents, but there is also a medium effect for youth and miles pc. Younger states drive more, which could explain higher rate of accidents, not necessarily that younger drivers are worse at driving (although they probably are).

Regressor Interactions: Fatality Rate and Traffic Laws

State-Aggregates Over Time

```
# ```{r}
mean_by_year = aggregate(traffic[, c("totfatrte", "bac08", "bac10", "sbprim",
    "sbsecon", "perse", "sl70plus", "gdl", "zerotol", "minage")], traffic["year"],
    FUN = mean)
ggpairs(mean_by_year, size = 3) + ggtitle(label = "Proportion of States with Law vs. Year (1980-2004)")
```

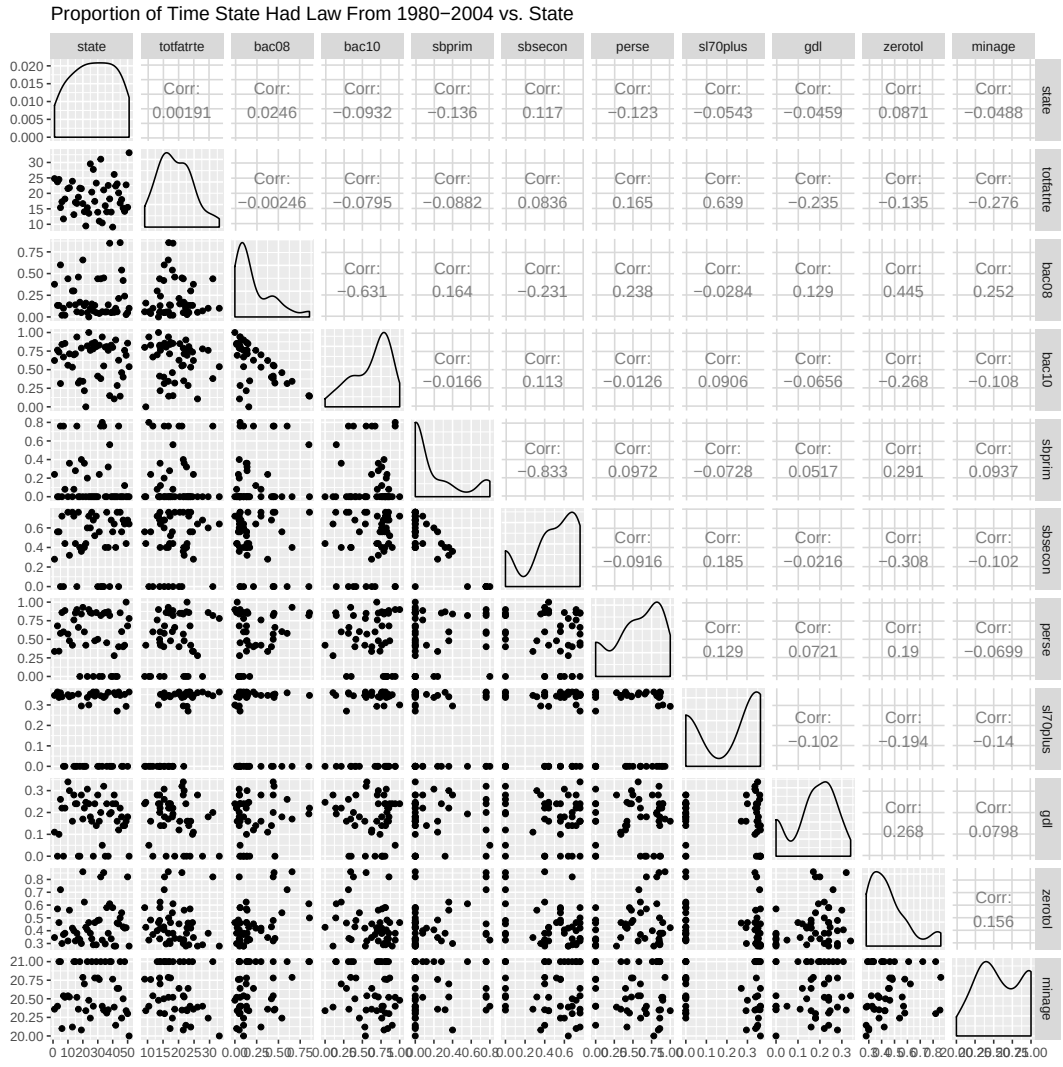


These data points indicate trends for the entire country and give us a sense of the adoption of each law over time. We see a strong time dependence between each of the law variables and time. Although the correlation between time and *bac*10 is low, it's evident that a quadratic term would do an excellent job of predicting the number of states with the law at any given year. When comparing scatters of any two variables, we cannot consider their covariance as significant due to spurious correlation. However, there

There seems to be strong covariance between the *gdl* and *sbprim* covariates. That is to say, for a given year, if we see a relatively high number of states with *gdl* laws, we would also expect there to be a relatively high number of states with *sbprim* laws for that year.

Year-Aggregates By State

```
mean_by_state = aggregate(traffic[, c("totfatrte", "bac08", "bac10", "sbprim",
    "sbsecon", "perse", "sl70plus", "gdl", "zerotol", "minage")], traffic["state"],
    FUN = mean)
ggpairs(mean_by_state, size = 3) + ggtitle(label = "Proportion of Time State Had Law From 1980-2004 vs.
```



There is strong correlation between *bac*08 and *bac*10 variables and *sbprim* and *sbsecon* variables. This makes sense, because these sets of laws are mutually exclusive; a state can only have one law or the other. Therefore, we would expect to see that the more years a state has one variable, the less years we would expect to see of its converse law. There are a few cases of significant interaction between covariates. For example, the correlation between *bac*08 and *zerotol* is 0.445. This means that the longer a state has had a *bac*08 law passed, the longer we would expect that state to have had a *zerotol* law passed as well.

Question 2

---

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

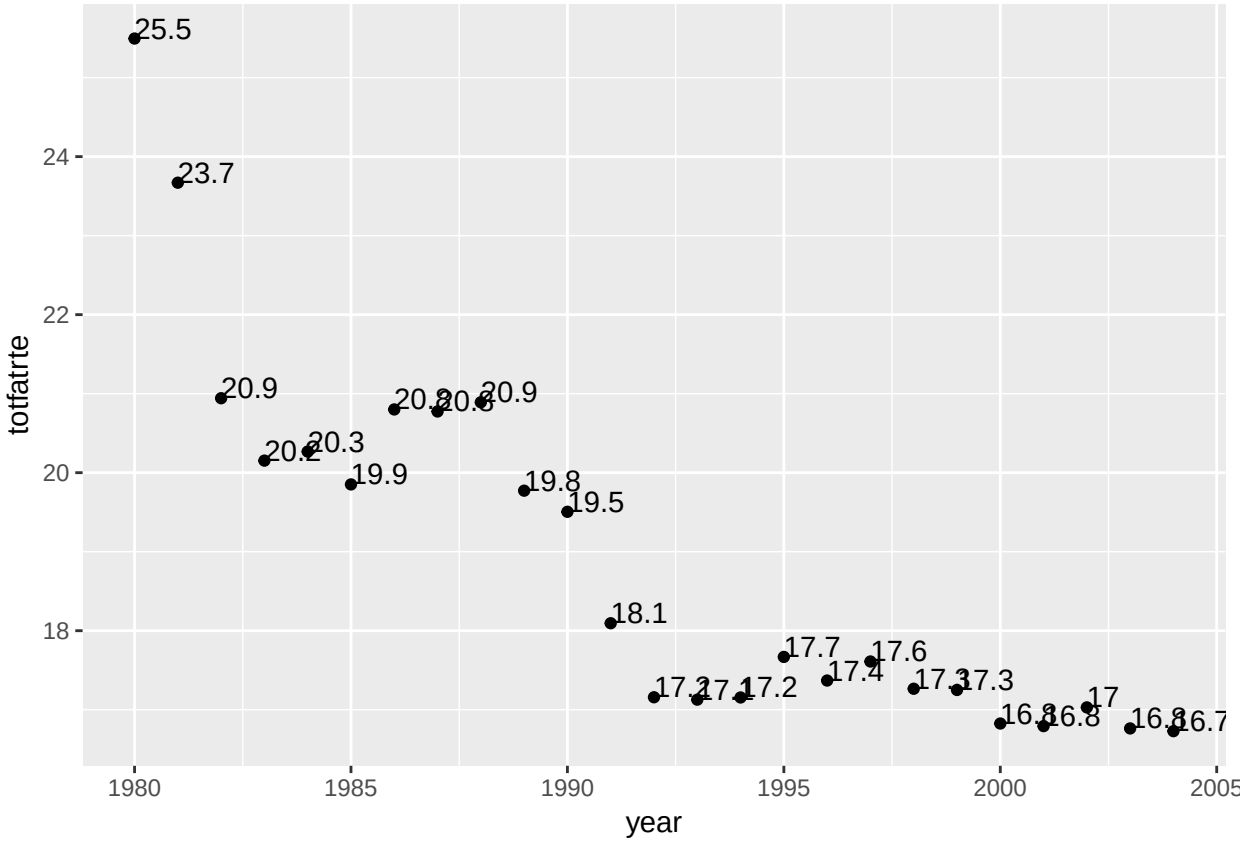
2.1 Totfatrte Definition

How is the our dependent variable of interest totfatrte defined? The *totfatrte* is defined as Fatalities per 100K population.

2.2 Totfatrte Average

What is the average of this variable in each of the years in the time period covered in this dataset?

```
avg_by_year = aggregate(totfatrte ~ year, traffic, mean)
qplot(x = year, y = totfatrte, data = avg_by_year) + geom_text(aes(label = round(totfatrte,
    1)), hjust = 0, vjust = 0)
```



Over the years, we see a gradual decrease in the average *totfatrte* variable - from a high of 25.49 in 1980 to a low of 16.73 in 2004. We do see two steep declines in the rate of Fatalities (in the early 80's and then again between late 80's and early 90's), since then we have only seen a gradual decrease. It also appears there are periods when we see an increase in the average fatality rate (albeit small) mid-80's, '92-'95 and 2002.

2.3 Totfatrte Regression

2.3.1 Modeling

Estimate a very simple regression model of totfatrte on dummy variables for the years 1981 through 2004.

```
lm.mod = lm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +
    d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 +
```

```
    d01 + d02 + d03 + d04, data = traffic)
summary(lm.mod)
```

```
##
## Call:
## lm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04, data = traffic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.930  -4.347  -0.731   3.749  29.650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   25.495      0.867   29.40  < 2e-16 ***
## d81           -1.824      1.226   -1.49  0.13709
## d82           -4.552      1.226   -3.71  0.00022 ***
## d83           -5.342      1.226   -4.36  1.4e-05 ***
## d84           -5.227      1.226   -4.26  2.2e-05 ***
## d85           -5.643      1.226   -4.60  4.6e-06 ***
## d86           -4.694      1.226   -3.83  0.00014 ***
## d87           -4.720      1.226   -3.85  0.00013 ***
## d88           -4.603      1.226   -3.75  0.00018 ***
## d89           -5.722      1.226   -4.67  3.4e-06 ***
## d90           -5.989      1.226   -4.88  1.2e-06 ***
## d91           -7.400      1.226   -6.03  2.1e-09 ***
## d92           -8.337      1.226   -6.80  1.7e-11 ***
## d93           -8.367      1.226   -6.82  1.4e-11 ***
## d94           -8.339      1.226   -6.80  1.7e-11 ***
## d95           -7.826      1.226   -6.38  2.5e-10 ***
## d96           -8.125      1.226   -6.63  5.2e-11 ***
## d97           -7.884      1.226   -6.43  1.9e-10 ***
## d98           -8.229      1.226   -6.71  3.0e-11 ***
## d99           -8.244      1.226   -6.72  2.8e-11 ***
## d00           -8.669      1.226   -7.07  2.7e-12 ***
## d01           -8.702      1.226   -7.10  2.2e-12 ***
## d02           -8.465      1.226   -6.90  8.3e-12 ***
## d03           -8.731      1.226   -7.12  1.9e-12 ***
## d04           -8.766      1.226   -7.15  1.5e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.01 on 1175 degrees of freedom
## Multiple R-squared:  0.128,  Adjusted R-squared:  0.11
## F-statistic: 7.16 on 24 and 1175 DF,  p-value: <2e-16
```

```
mean(traffic$totfatrte[traffic$d80 == 1])
```

```
## [1] 25.49
```

2.3.2 Model Explanation

What does this model explain?

The regression explains how the fatality rate has changed (average across the 48 states) as compared to the year 1980. For example, d04 has a coefficient of -8.766. This implies that the average of $totfatrte$ was 8.8 less than 25.5 (average of $totfatrte$ in 1980) or 16.8, precisely what is reported in the graph above.

2.3.3 Model Findings

Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.

The fitted regression equation is $avg_fatality = 25.49 + (-1.824) \cdot d81 + (-4.552) \cdot d82 + \cdots + (-8.766) \cdot d04$ All the estimated coefficients are statistically significant at 5%, except for the one for 1981, which suggests that there was a lot of variability in fatality rates across the 48 states in 1981.

The intercept of the regression model is the average fatality rate for 1980. All other coefficients measure how the average fatality compares for the year (represented by the dummy variable) versus 1980. Each coefficient is negative, which means that the average fatality rate each year decreased relative to 1980. The coefficients are also mostly increasingly negative, represting the negative trend that we see in the graph above. *Based on this, we would say that driving (as measured by fatality rate per 100K population) has gotten safer over the years 1980-2004 on average in the United States.* Note, we have not accounted for any other factors in this assessment. The target variable $totfatrte$ is a simple average of the fatality rates of each state, which is not weighted by the population of each state, nor for the cumulative miles driven.

Question 3

---

3. Expand your model in Exercise 2 by adding variables bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc, and perhaps transformations of some or all of these variables. Please explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed. How are the variables bac8 and bac10 defined? Interpret the coefficients on bac8 and bac10. Do per se laws have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.) **********

3.1 Model Expansion

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.

```
traffic$bac08bin = ifelse(traffic$bac08 < 0.5, 0, 1)
traffic$bac10bin = ifelse(traffic$bac10 <= 0.5, 0, 1)
traffic$persebin = ifelse(traffic$perse < 0.5, 0, 1)
traffic$sbprimbin = ifelse(traffic$sbprim < 0.5, 0, 1)
traffic$sbseconbin = ifelse(traffic$sbsecon < 0.5, 0, 1)
traffic$sl70plusbin = ifelse(traffic$sl70plus < 0.5, 0, 1)
traffic$gdlbin = ifelse(traffic$gdl < 0.5, 0, 1)

traffic$bac0810bin = traffic$bac08bin + traffic$bac10bin
describe(traffic$bac0810bin)
```

```
## traffic$bac0810bin
##        n  missing distinct     Info      Sum     Mean      Gmd
```

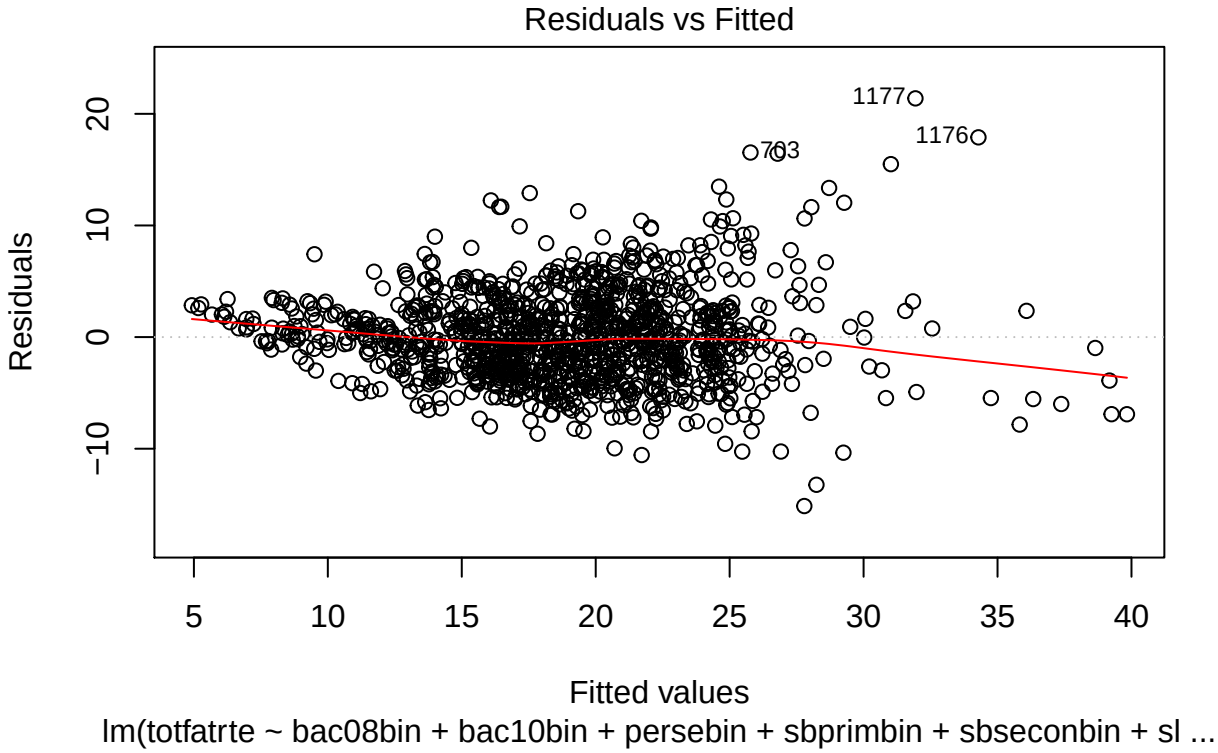
```
##     1200        0        2    0.417     1000   0.8333    0.278
```

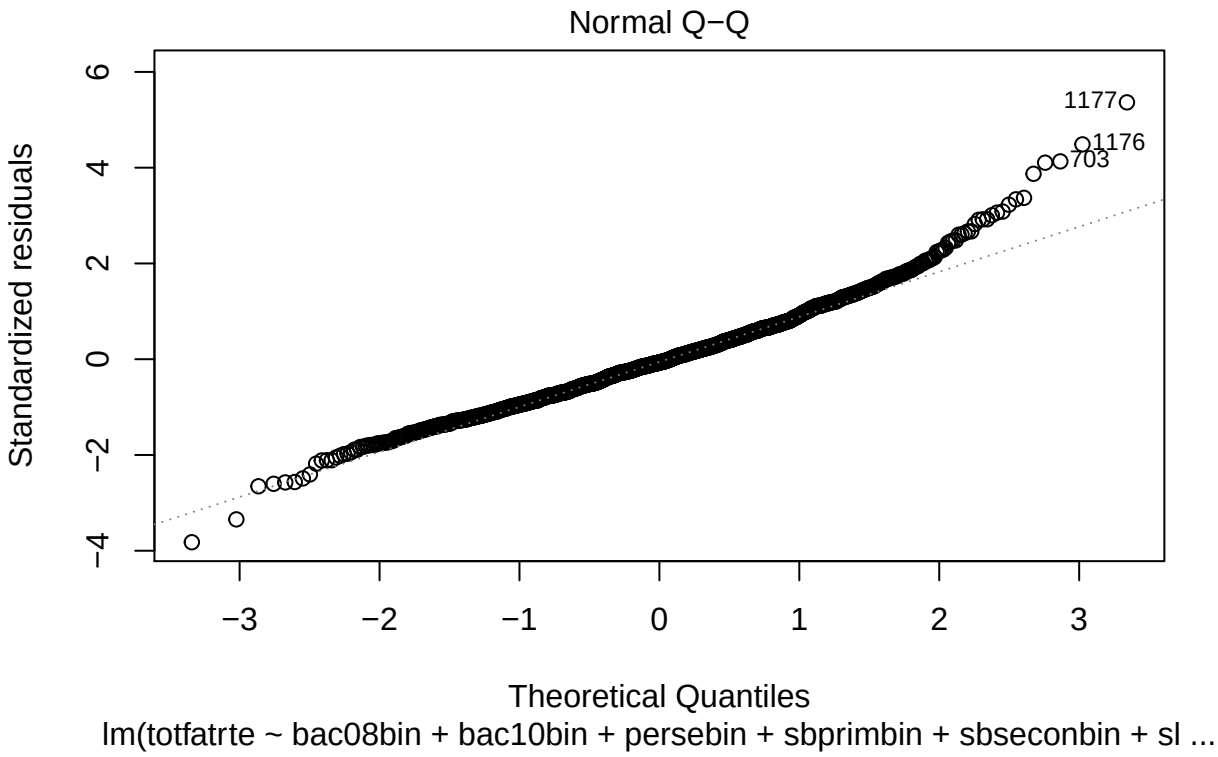
```
lm.mod2 = lm(totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin + sbseconbin +
    sl70plusbin + gdlbin + perc14_24 + unem + vehicmilespc + d81 + d82 +
    d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 +
    d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04, data = traffic)
summary(lm.mod2)
```

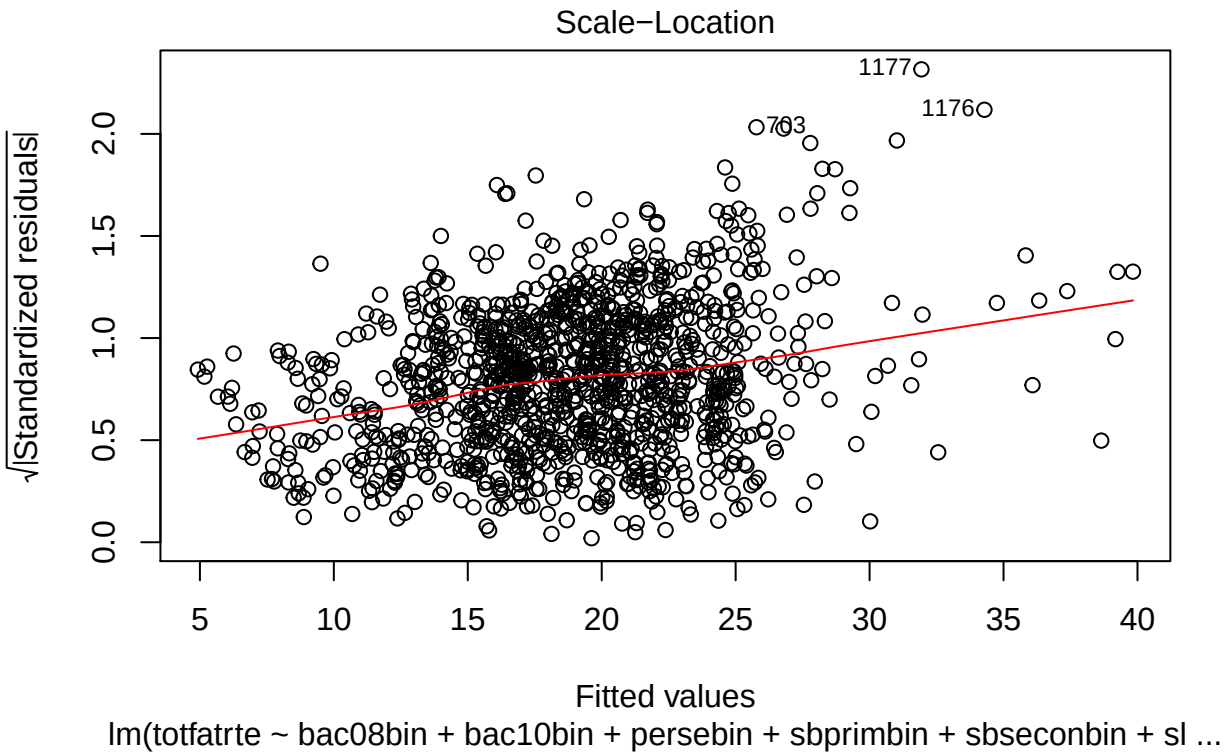
```
##
## Call:
## lm(formula = totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin +
##     sbseconbin + sl70plusbin + gdlbin + perc14_24 + unem + vehicmilespc +
##     d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 +
##     d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 +
##     d01 + d02 + d03 + d04, data = traffic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.137  -2.753  -0.272   2.305  21.385
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.95e+00   2.47e+00   -1.19   0.2330
## bac08bin     -2.63e+00   5.22e-01   -5.03  5.6e-07 ***
## bac10bin     -1.57e+00   3.85e-01   -4.07  5.1e-05 ***
## persebin     -5.74e-01   2.92e-01   -1.97   0.0496 *
## sbprimbin    -4.25e-02   4.91e-01   -0.09   0.9310
## sbseconbin    9.29e-02   4.29e-01    0.22   0.8287
## sl70plusbin   3.12e+00   4.34e-01    7.17  1.3e-12 ***
## gdlbin       -4.49e-01   5.06e-01   -0.89   0.3751
## perc14_24     1.50e-01   1.23e-01    1.22   0.2221
## unem          7.67e-01   7.78e-02    9.86  < 2e-16 ***
## vehicmilespc  2.93e-03   9.50e-05   30.87  < 2e-16 ***
## d81          -2.18e+00   8.28e-01   -2.63   0.0086 **
## d82          -6.61e+00   8.54e-01   -7.75  2.1e-14 ***
## d83          -7.47e+00   8.67e-01   -8.61  < 2e-16 ***
## d84          -5.80e+00   8.75e-01   -6.63  5.0e-11 ***
## d85          -6.43e+00   8.93e-01   -7.20  1.1e-12 ***
## d86          -5.79e+00   9.30e-01   -6.22  6.8e-10 ***
## d87          -6.30e+00   9.67e-01   -6.51  1.1e-10 ***
## d88          -6.52e+00   1.01e+00   -6.43  1.9e-10 ***
## d89          -7.99e+00   1.05e+00   -7.59  6.4e-14 ***
## d90          -8.88e+00   1.08e+00   -8.25  4.4e-16 ***
## d91          -1.10e+01   1.10e+00   -9.99  < 2e-16 ***
## d92          -1.28e+01   1.12e+00  -11.42  < 2e-16 ***
## d93          -1.27e+01   1.14e+00  -11.15  < 2e-16 ***
## d94          -1.23e+01   1.16e+00  -10.60  < 2e-16 ***
## d95          -1.19e+01   1.18e+00  -10.04  < 2e-16 ***
## d96          -1.39e+01   1.23e+00  -11.31  < 2e-16 ***
## d97          -1.40e+01   1.25e+00  -11.24  < 2e-16 ***
## d98          -1.48e+01   1.26e+00  -11.75  < 2e-16 ***
## d99          -1.49e+01   1.28e+00  -11.59  < 2e-16 ***
## d00          -1.52e+01   1.30e+00  -11.69  < 2e-16 ***
## d01          -1.59e+01   1.33e+00  -11.99  < 2e-16 ***
## d02          -1.65e+01   1.34e+00  -12.28  < 2e-16 ***
```

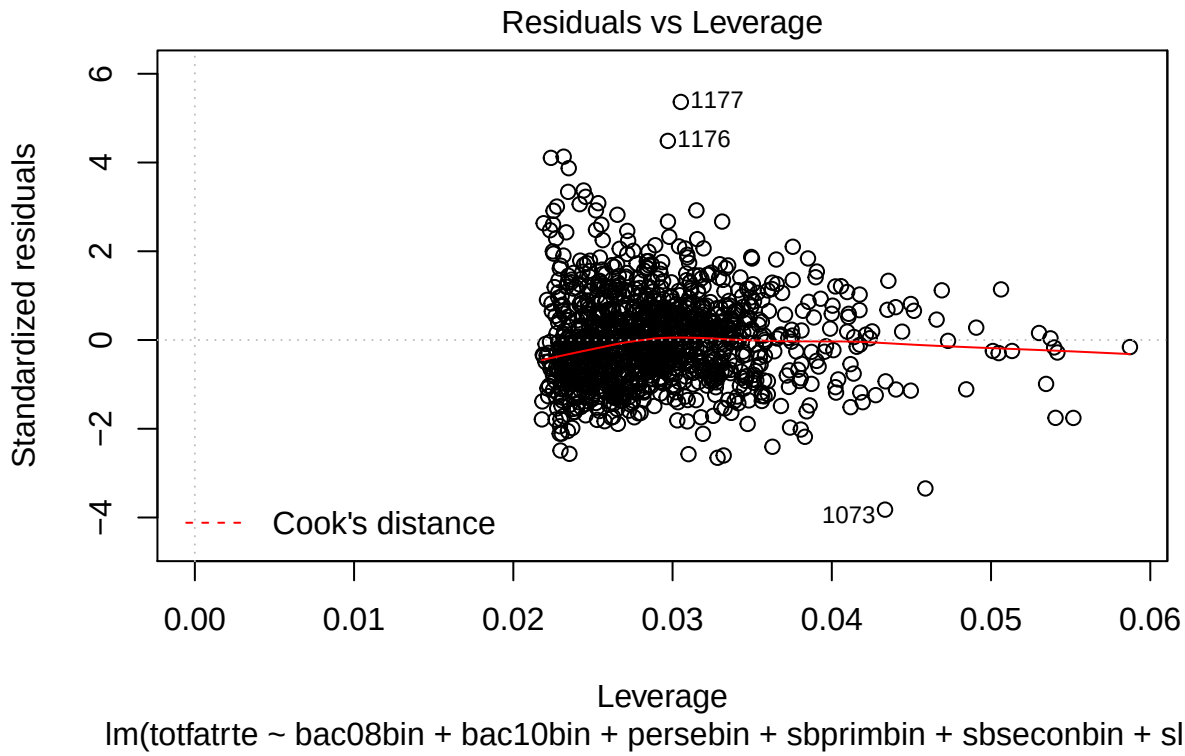
```
## d03          -1.68e+01   1.35e+00  -12.48  < 2e-16 ***
## d04          -1.65e+01   1.38e+00  -11.98  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.05 on 1165 degrees of freedom
## Multiple R-squared:  0.607,  Adjusted R-squared:  0.596
## F-statistic:   53 on 34 and 1165 DF,  p-value: <2e-16
```

```
plot(lm.mod2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin + sbseconbin + sl ...
1177
1176
703
−4
−2
0
2
4
6
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1177
1176
703
5
10
15
20
25
30
35
40
Fitted values
lm(totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin + sbseconbin + sl ...



##3.2 Rationale 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.

In most instances we can expect there to be a lag between when certain types of laws are enacted and when they impact change in behavior. Given that, we suggest transforming the variables that are essentially binary in nature (whether a law was in effect or not) to be strictly binary (so if it was in effect for less than 1/2 the year, we will code as 0 and 1 otherwise).

As for the other variables - $perc14_24$, $unem$, $vehicmilespc$, they are all already expressed as normalized metrics (where the denominator is some measure of the population, either as per 100 people or per person).

3.3 BAC

3.3.1 Variable Definition

How are the variables bac8 and bac10 defined?

The BAC (blood alcohol content) is a measure used to determine if someone is driving under the influence of alcohol. For $bac10$ the threshold is 0.10 and $bac8$ represents a threshold of 0.08.

3.3.2 Coefficient Interpretation

Interpret the coefficients on bac8 and bac10.

The coefficient on bac08 is -2.13 and the coefficient for bac10 is -1.12; The mathematical interpretation of the coefficients in the regression would suggest that all else being equal, a state that has a threshold of 0.08 would have a lower $totfatrte$ than a state that has a threshold of 0.1 because the coeff for bac08 is -2.13

and the coeff for bac10 is -1.12. This model structure also suggests that the effect of having neither of the thresholds implies an increase of the fatality rate by 3.25.

3.4 Per Se Laws

Do per se laws have a negative effect on the fatality rate?

According to the linear regression model - YES, perse laws have a negative effect on $totfatrte$. All else being equal, the existence of perse laws lowers $totfatrte$ by -0.57 compared to when these laws do not exist.

3.5 Primary Seat Belt

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.)

Per the model specification, inclusion of the primary seatbelt variable *sbprimbin* implies a reduction in Fatalities by 0.0425. However, because the coefficient is not significantly different from zero, we interpret it as having no effect on $totfatrte$. This does not imply that seatbelts do not affect fatality rates, just that compulsory laws may not change drivers' willingness to comply with the law.

Question 4

---

4. Reestimate the model from Exercise 3 using a fixed effects (at the state level) model. How do the coefficients on bac08, bac10, perse, and sbprim compare with the pooled OLS estimates? Which set of estimates do you think is more reliable? What assumptions are needed in each of these models? Are these assumptions reasonable in the current context? **********

4.1 Fixed Effect Modeling

Reestimate the model from Exercise 3 using a fixed effects (at the state level) model. ###

4.2 Coefficient Comparison

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

The coefficients for bac08, bac10, perse and sbprim are all statistically significant with the panel fixed-effects model. The estimate for the bac08 coeff is slightly lower, the coeff for bac10 is about the same as before. The perse coeff has changed significanty from -6.1 to -1.1; We also fnid that the coeff for sbprim is now statistically significant.

4.3 Assessment of Reliability

Which set of estimates do you think is more reliable?
Both models explain about the same amount of variation in the $totfatrte$ variable (comparable adjusted-R-sq values); however the fixed-effects model is more reliable because it models the variation over time in $totfatrte$ and all the independent variables *within* each state

4.4 Modeling Assumptions

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

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

The pooled effects assumption (made in a random effects model) is that the individual-specific effects are uncorrelated with the independent variables. This is a very strong assumption and a difficult one to meet. We are assuming there is no omitted variable bias and no correlation between same-state observations.

Question 5

---

5. Would you prefer to use a random effects model instead of the fixed effects model you build in Exercise 4? Why? Why not? **********

Random effects model would pool everything together and would rely on the assumption that same-state observations are independent. While we have quite a few variables we could add to the mix, this isn't a realistic assumption to make. We need to account for cultural/economic differences between states. We are able to remove this omitted variable bias using fixed effects modeling with dummy variables of each year to explain the unaccounted for time-variant error dependence.

Question 6

---

6. 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? Be sure to interpret the estimate as if explaining to a layperson. **********

From the fixed-effects regression, the coefficient for the *vehicmilespc* variable was 0.000942 Fatalities/100K people per mile-driven/capita. This implies that for all other things held equal, an increase of 1,000 miles driven per capita would result in an increase of 0.942 (approximately 1) fatalities per 100K people.

Question 7

---

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

If there existed serial correlation in the model errors, we would be too liberal with our confidence intervals for estimations of the model parameters. That is, we would commit Type I error, rejecting the null hypothesis too easily, saying that our variables are significant when they actually aren't.

Conversely, with the presence of heteroskedasticity in our error, we may commit Type II error, failing to reject the null hypothesis when it should be rejected, which means we do not detect the significance of a potentially valuable regressor.

Let's now compare the three possible models we could realistically use.

```
library(plm)
```

```
##
## Attaching package: 'plm'

## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

```
# Pooled
plm.mod.pl = plm(totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin +
    sbseconbin + sl70plusbin + gdlbin + 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 = traffic, model = c("pooling"), index = c("state", "year"))
summary(plm.mod.pl)
```

```
## Pooling Model
##
## Call:
## plm(formula = totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin +
##     sbseconbin + sl70plusbin + gdlbin + 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 = traffic, model = c("pooling"),
##     index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -15.137  -2.753  -0.272   2.305  21.385
##
## Coefficients:
##                Estimate Std. Error t-value Pr(>|t|)
## (Intercept)  -2.95e+00   2.47e+00   -1.19   0.2330
## bac08bin     -2.63e+00   5.22e-01   -5.03  5.6e-07 ***
## bac10bin     -1.57e+00   3.85e-01   -4.07  5.1e-05 ***
## persebin     -5.74e-01   2.92e-01   -1.97   0.0496 *
## sbprimbin    -4.25e-02   4.91e-01   -0.09   0.9310
## sbseconbin    9.29e-02   4.29e-01    0.22   0.8287
## sl70plusbin   3.12e+00   4.34e-01    7.17  1.3e-12 ***
## gdlbin       -4.49e-01   5.06e-01   -0.89   0.3751
## perc14_24     1.50e-01   1.23e-01    1.22   0.2221
## unem          7.67e-01   7.78e-02    9.86  < 2e-16 ***
## vehicmilespc  2.93e-03   9.50e-05   30.87  < 2e-16 ***
## d81          -2.18e+00   8.28e-01   -2.63   0.0086 **
## d82          -6.61e+00   8.54e-01   -7.75  2.1e-14 ***
## d83          -7.47e+00   8.67e-01   -8.61  < 2e-16 ***
## d84          -5.80e+00   8.75e-01   -6.63  5.0e-11 ***
## d85          -6.43e+00   8.93e-01   -7.20  1.1e-12 ***
## d86          -5.79e+00   9.30e-01   -6.22  6.8e-10 ***
## d87          -6.30e+00   9.67e-01   -6.51  1.1e-10 ***
## d88          -6.52e+00   1.01e+00   -6.43  1.9e-10 ***
## d89          -7.99e+00   1.05e+00   -7.59  6.4e-14 ***
## d90          -8.88e+00   1.08e+00   -8.25  4.4e-16 ***
## d91          -1.10e+01   1.10e+00   -9.99  < 2e-16 ***
```

```
## d92          -1.28e+01   1.12e+00  -11.42  < 2e-16 ***
## d93          -1.27e+01   1.14e+00  -11.15  < 2e-16 ***
## d94          -1.23e+01   1.16e+00  -10.60  < 2e-16 ***
## d95          -1.19e+01   1.18e+00  -10.04  < 2e-16 ***
## d96          -1.39e+01   1.23e+00  -11.31  < 2e-16 ***
## d97          -1.40e+01   1.25e+00  -11.24  < 2e-16 ***
## d98          -1.48e+01   1.26e+00  -11.75  < 2e-16 ***
## d99          -1.49e+01   1.28e+00  -11.59  < 2e-16 ***
## d00          -1.52e+01   1.30e+00  -11.69  < 2e-16 ***
## d01          -1.59e+01   1.33e+00  -11.99  < 2e-16 ***
## d02          -1.65e+01   1.34e+00  -12.28  < 2e-16 ***
## d03          -1.68e+01   1.35e+00  -12.48  < 2e-16 ***
## d04          -1.65e+01   1.38e+00  -11.98  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    48600
## Residual Sum of Squares: 19100
## R-Squared:      0.607
## Adj. R-Squared: 0.596
## F-statistic: 52.9708 on 34 and 1165 DF, p-value: <2e-16
```

```
# Fixed Effects
plm.mod.fe = plm(totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin +
    sbseconbin + sl70plusbin + gdlbin + 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 = traffic, model = c("within"), index = c("state", "year"), effect = "twoways")
summary(plm.mod.fe)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin +
##     sbseconbin + sl70plusbin + gdlbin + 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 = traffic, effect = "twoways",
##     model = c("within"), index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -8.43054 -1.03297 -0.00645  0.96011 14.80921
##
## Coefficients:
##               Estimate Std. Error t-value Pr(>|t|)
## bac08bin     -1.407134   0.369658   -3.81  0.00015 ***
## bac10bin     -1.088696   0.254904   -4.27  2.1e-05 ***
## persebin     -1.102371   0.224707   -4.91  1.1e-06 ***
## sbprimbin    -1.225085   0.342461   -3.58  0.00036 ***
## sbseconbin   -0.346177   0.251932   -1.37  0.16969
## sl70plusbin  -0.053950   0.260574   -0.21  0.83601
## gdlbin       -0.426583   0.279756   -1.52  0.12758
```

```
## perc14_24     0.186621   0.094926    1.97  0.04955 *
## unem         -0.569569   0.060607   -9.40  < 2e-16 ***
## vehicmilespc  0.000942   0.000111    8.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5930
## Residual Sum of Squares: 4530
## R-Squared:      0.236
## Adj. R-Squared: 0.18
## F-statistic: 34.4483 on 10 and 1118 DF, p-value: <2e-16
```

```
# Random Effects
plm.mod.re = plm(totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin +
    sbseconbin + sl70plusbin + gdlbin + 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 = traffic, model = c("random"), index = c("state", "year"))
summary(plm.mod.re)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin +
##     sbseconbin + sl70plusbin + gdlbin + 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 = traffic, model = c("random"),
##     index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Effects:
##                 var std.dev share
## idiosyncratic 4.05    2.01  0.33
## individual    8.35    2.89  0.67
## theta: 0.862
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##  -8.301  -1.148  -0.157   0.921  16.447
##
## Coefficients:
##               Estimate Std. Error t-value Pr(>|t|)
## (Intercept)   17.13010    2.09447    8.18  7.4e-16 ***
## bac08bin      -1.54100    0.37903   -4.07  5.1e-05 ***
## bac10bin      -1.16570    0.26187   -4.45  9.3e-06 ***
## persebin      -1.04806    0.22950   -4.57  5.5e-06 ***
## sbprimbin     -1.17405    0.35104   -3.34  0.00085 ***
## sbseconbin    -0.34410    0.25988   -1.32  0.18574
## sl70plusbin    0.02779    0.26867    0.10  0.91764
## gdlbin        -0.39653    0.28912   -1.37  0.17048
## perc14_24      0.19726    0.09701    2.03  0.04224 *
## unem          -0.49003    0.06184   -7.92  5.3e-15 ***
```

```
## vehicmilespc   0.00118    0.00011   10.70  < 2e-16 ***
## d81           -1.55554    0.42797   -3.63  0.00029 ***
## d82           -3.24149    0.45712   -7.09  2.3e-12 ***
## d83           -3.79746    0.46964   -8.09  1.5e-15 ***
## d84           -4.37537    0.47786   -9.16  < 2e-16 ***
## d85           -4.87247    0.49829   -9.78  < 2e-16 ***
## d86           -3.83521    0.53192   -7.21  1.0e-12 ***
## d87           -4.50527    0.57009   -7.90  6.3e-15 ***
## d88           -4.97966    0.61731   -8.07  1.8e-15 ***
## d89           -6.37867    0.65620   -9.72  < 2e-16 ***
## d90           -6.54115    0.68107   -9.60  < 2e-16 ***
## d91           -7.31414    0.69833  -10.47  < 2e-16 ***
## d92           -8.24499    0.71931  -11.46  < 2e-16 ***
## d93           -8.56647    0.73230  -11.70  < 2e-16 ***
## d94           -8.90705    0.75187  -11.85  < 2e-16 ***
## d95           -8.70897    0.77284  -11.27  < 2e-16 ***
## d96           -9.10536    0.81880  -11.12  < 2e-16 ***
## d97           -9.23412    0.83628  -11.04  < 2e-16 ***
## d98           -9.91823    0.85050  -11.66  < 2e-16 ***
## d99          -10.04491    0.86114  -11.66  < 2e-16 ***
## d00          -10.56558    0.87289  -12.10  < 2e-16 ***
## d01          -10.28819    0.88888  -11.57  < 2e-16 ***
## d02           -9.65536    0.89726  -10.76  < 2e-16 ***
## d03           -9.72738    0.90186  -10.79  < 2e-16 ***
## d04          -10.10285    0.92330  -10.94  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12800
## Residual Sum of Squares: 5070
## R-Squared:      0.605
## Adj. R-Squared: 0.593
## F-statistic: 52.4298 on 34 and 1165 DF, p-value: <2e-16
```

```
phtest(plm.mod.fe, plm.mod.re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ bac08bin + bac10bin + persebin + sbprimbin + sbseconbin +  ...
## chisq = 150, df = 10, p-value <2e-16
## alternative hypothesis: one model is inconsistent
```

```
names(plm.mod.fe)
```

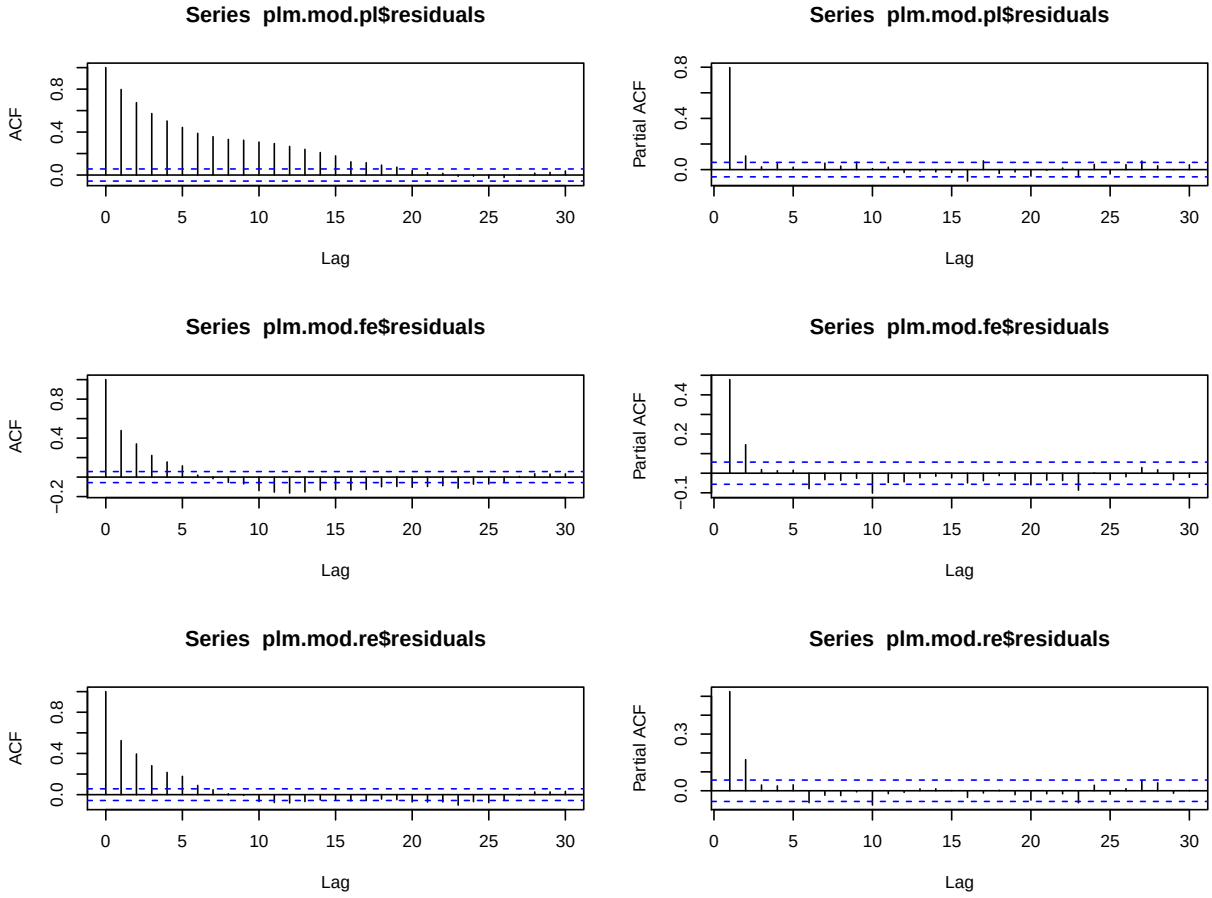
```
##  [1] "coefficients" "vcov"         "residuals"    "df.residual"
##  [5] "formula"      "model"        "assign"       "args"
##  [9] "aliased"      "call"
```

```
describe(plm.mod.fe$residuals)
```

```
## plm.mod.fe$residuals
##          n    missing   distinct       Info       Mean        Gmd
##       1200          0       1200          1 -7.969e-16      2.026
##        .05        .10        .25        .50        .75        .90
##  -2.998638  -2.130841  -1.032969  -0.006452   0.960110   1.994515
```

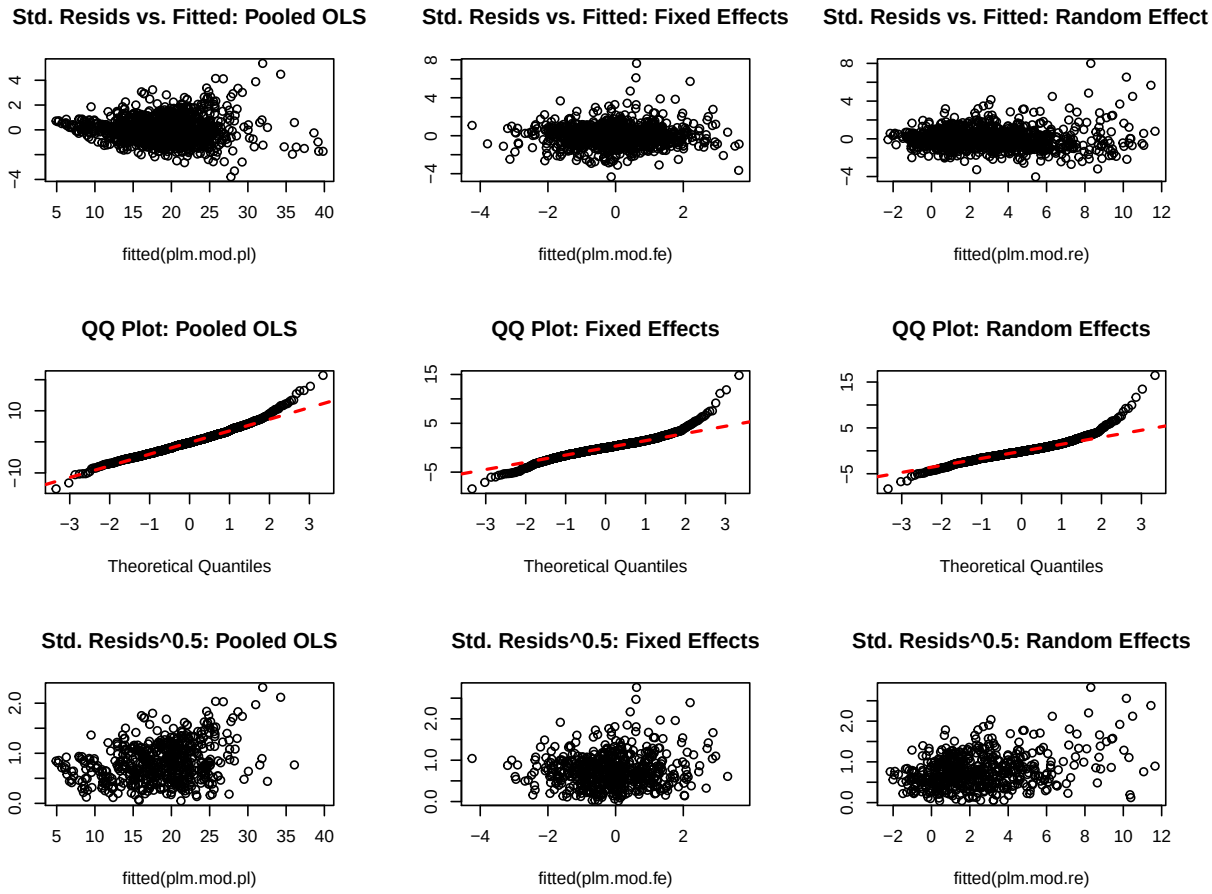
```
##        .95
##   2.881907
##
## lowest : -8.431 -7.105 -6.037 -5.989 -5.661, highest:  7.554  9.148 11.120 11.862 14.809
```

```r
# Recall that our data is chronologically ordered for each state
par(mfrow = c(3, 2))
acf(plm.mod.pl$residuals)
pacf(plm.mod.pl$residuals)
acf(plm.mod.fe$residuals)
pacf(plm.mod.fe$residuals)
acf(plm.mod.re$residuals)
pacf(plm.mod.re$residuals)
```




```r
par(mfrow = c(3, 3))
plot(fitted(plm.mod.pl), residuals(plm.mod.pl)/sd(residuals(plm.mod.pl)),
    main = "Std. Resids vs. Fitted: Pooled OLS", ylab = "")
plot(fitted(plm.mod.fe), residuals(plm.mod.fe)/sd(residuals(plm.mod.fe)),
    main = "Std. Resids vs. Fitted: Fixed Effects", ylab = "")
plot(fitted(plm.mod.re), residuals(plm.mod.re)/sd(residuals(plm.mod.re)),
    main = "Std. Resids vs. Fitted: Random Effects", ylab = "")

qqnorm(residuals(plm.mod.pl), main = "QQ Plot: Pooled OLS", ylab = "")
qqline(residuals(plm.mod.pl), col = 2, lwd = 2, lty = 2)
qqnorm(residuals(plm.mod.fe), main = "QQ Plot: Fixed Effects", ylab = "")
```

```
qqline(residuals(plm.mod.fe), col = 2, lwd = 2, lty = 2)
qqnorm(residuals(plm.mod.re), main = "QQ Plot: Random Effects", ylab = "")
qqline(residuals(plm.mod.re), col = 2, lwd = 2, lty = 2)

plot(fitted(plm.mod.pl), sqrt(residuals(plm.mod.pl)/sd(residuals(plm.mod.pl))),
    main = "Std. Resids^0.5: Pooled OLS", ylab = "")
plot(fitted(plm.mod.fe), sqrt(residuals(plm.mod.fe)/sd(residuals(plm.mod.fe))),
    main = "Std. Resids^0.5: Fixed Effects", ylab = "")
plot(fitted(plm.mod.re), sqrt(residuals(plm.mod.re)/sd(residuals(plm.mod.re))),
    main = "Std. Resids^0.5: Random Effects", ylab = "")
```



We see from these plots that we made the best modeling choice to go with Fixed Effects over Random Effects or Pooled OLS models. Our residuals are less heterskedastic and the Hausman test shows us that we have a great deal of correlation between our unobserved effect and dependent variable. We also see that we have the least amount of autocorrelation in the fixed-effects model than we do for any other model.