Lab 4

Section 2 | Beau Kramer, Kathryn Papandrew

12/11/2018

Lab 4

W271 | Section 2 | Kramer, Papandrew

1. Exercise 1

1.1 Load the Data

```
# Read in dataframe

raw.df <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vSK9DBvw_760rJTO-26YZtbmLjht6K5Q1zc
```

1.2 Initial Observations

Provide a description of basic structure of dataset.

The dataset includes a mixture of variable types: binary, continuous, and categorical.

There are a handful of variables describing the laws in each state for each year, such as speed limit (`sl55`, `sl65`, `sl70`, `sl75`, `slnone`, `sl70plus`), seat belt laws (`seatbelt`, `sbprim`, `sbsecon`), and drinking and driving laws (`perse`, `minage`, `zerotol`, `bac10`, `bac08`). These laws are expressed in proportion of the year where the state had that law. For example, if the value for `sl65` is 0.5, it means that the state had a speed limit of 65 for half the year. Though much of the data for these fields looks binary, they're actually continuous observations. Additionally, there is ancillary data about each state over time expressed in continuous variables: the population (`statepop`), unemployment rate (`unem`), percent of population between 14 and 24 (`perc14_24`), vehicle miles traveled in billions (`vehicmiles`), and miles driven per 100,000 capita (`vehicmilespc`).

There are also several continuous features describing fatalities in each state both overall (`totfat, ngthtfat, wkndfat`) and broken out into categories such as per 100,000 capita (`totfatrte, nghtfatrte, wkndfatrte`), and per 100 million miles (`totfatpvm, ngthfatpvm, wkndfatpvm`).

Finally, there are binary features for each year in the dataset indicating a row of data belongs to a certain year (nomenclature is `dYY`, such as `d80` for 1980). This data is likely present to be used for interaction terms.

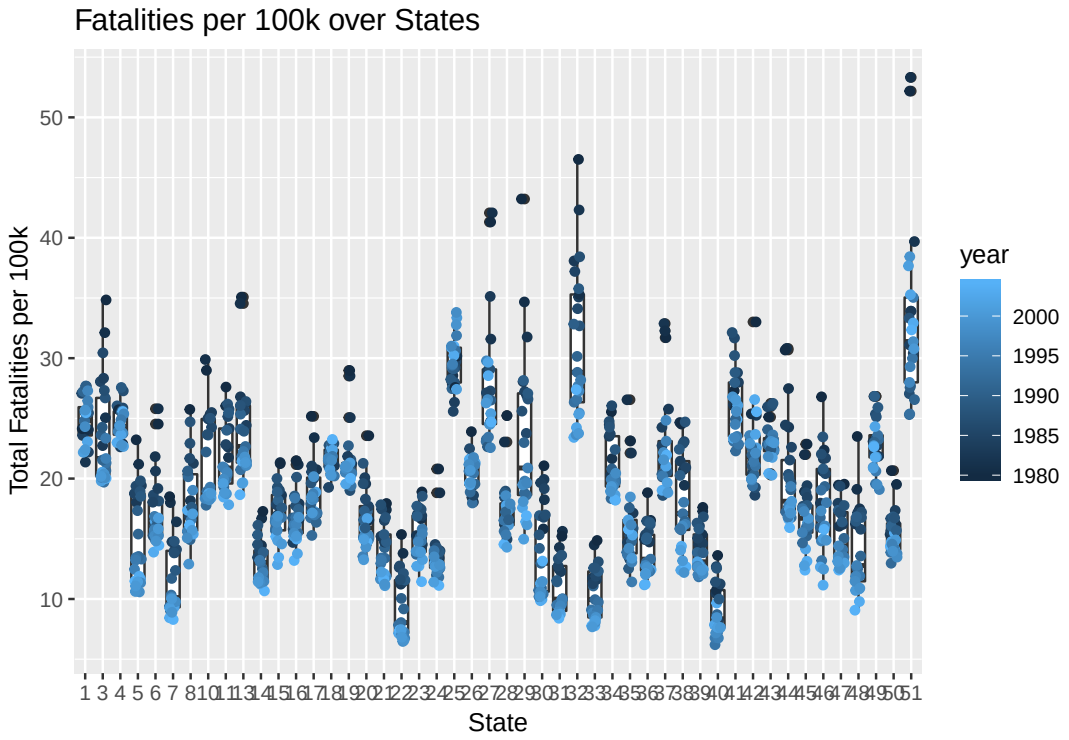
1.3 Exploratory Data Analysis

In this section, we will explore the data using both graphical and tabular techniques to get a better sense of the structure of the data and relationships between key model variables to use this information to determine if any transformations are needed for the model in the modeling sections.

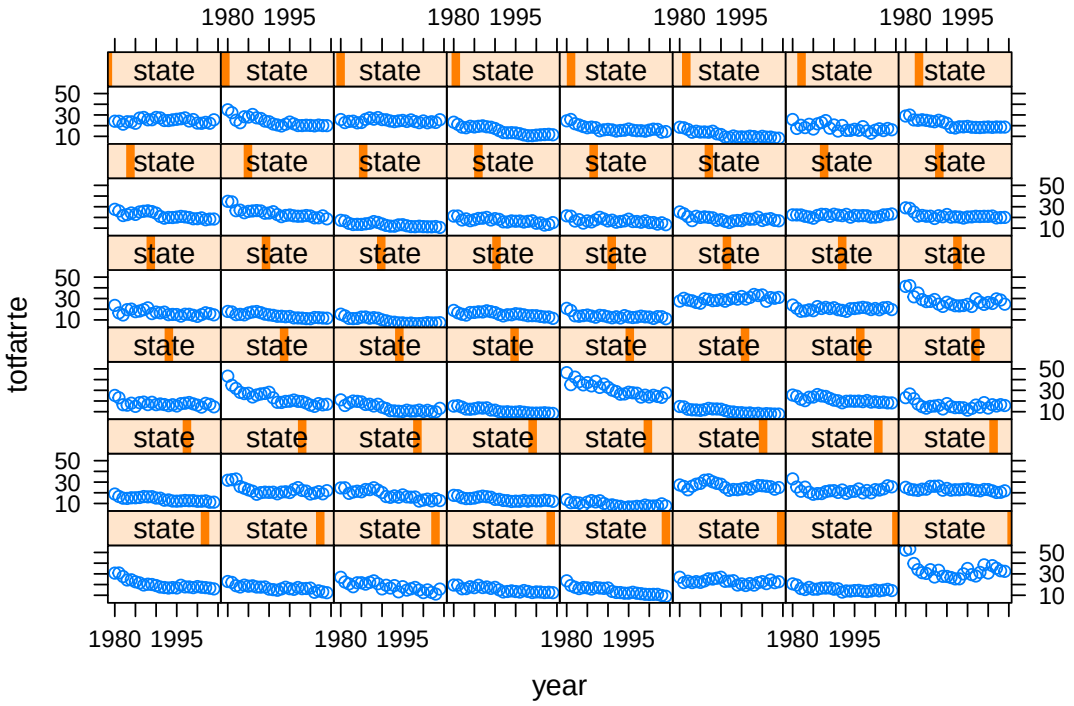
1.3.1 Dependent Variable - `totfatrte`

We start by analyzing our dependent variable, `totfatrte`, graphically below.

```
g2 <- ggplot(raw.df, aes(as.factor(state), totfatrte, colour=year)) +  geom_boxplot() + geom_jitter(wid
g3 <- xyplot(totfatrte ~ year | state, data=raw.df, as.table=T)
g2
```




```
g3
```




```
describe(raw.df$totfatrte)
```

```
## raw.df$totfatrte
```

```
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0      916        1    18.92    7.032    9.578   11.458
##      .25      .50      .75      .90      .95
##   14.377   18.435   22.773   26.790   29.895
##
## lowest :  6.20  6.47  6.55  6.75  6.76, highest: 42.31 43.22 46.51 52.18 53.32
```

The graphs above give us a few insights. First, we see that total fatalities are dropping as years progress (lighter colors in first graph) and both graphs show that that effects vary by state. We see some states varying a lot year over year while others stay fairly steady with a very slight slope downward.

Examining this variable in detail, we see that the mean total fatalities per 100,000 population is 18.92 and the median is 18.435. The lowest rates are in the sixes while the highest are in the mid-forties and low-fifties.

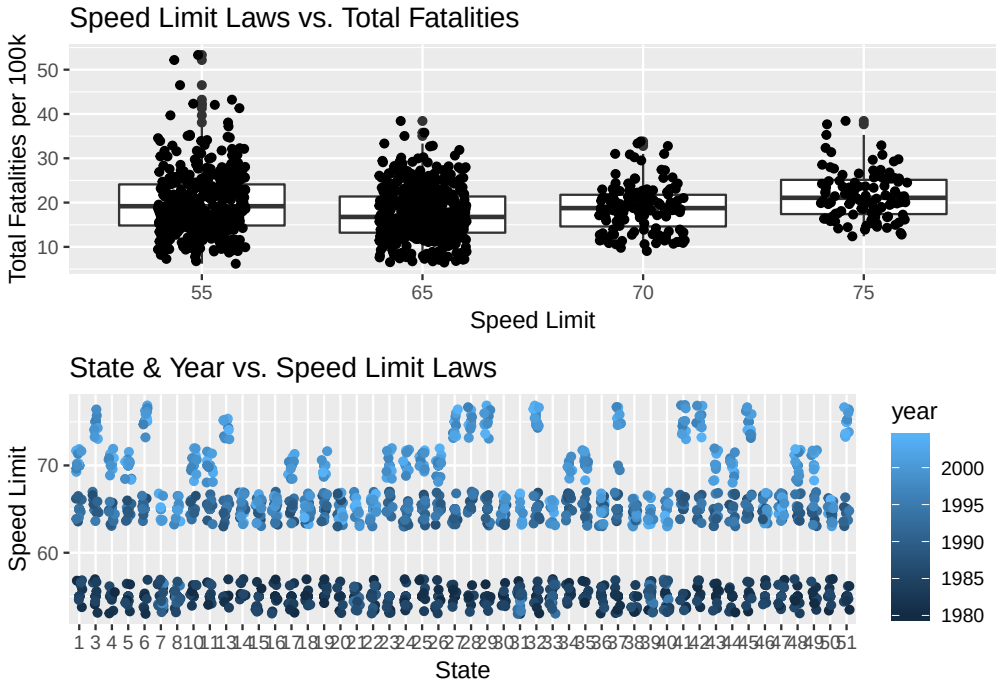
Overall, we don't see any gaping outliers or suspicious data so we will proceed with this data untransformed.

1.3.2 Explanatory Variables

1.3.2.1 Speed Limit Laws

To best visualize the speed limit for only EDA purposes, we decided to create a column giving the speed limit in that state where that speed limit was in place for the highest proportion of time for a given year. Please note that it was decided if there is no speed limit (i.e. `slnone` > 0), then the speed limit was set to `none`. The code to derive the speed limit can be seen in the .Rmd file but for the sake of space, this code is hidden in the PDF.

```
g1 <- ggplot(df.slmax, aes(as.factor(slwa), totfatrte)) + geom_boxplot() + geom_jitter(width = 0.2) + g
xlab("Speed Limit") + ylab("Total Fatalities per 100k")
g2 <- ggplot(df.slmax, aes(as.factor(state), slwa, colour=year)) + geom_point() + geom_jitter(width = 0
xlab("State") + ylab("Speed Limit")
grid.arrange(g1,g2,nrow=2)
```



The graph above displays the relationship between the speed limit in a state total fatalities per 100,000. There seems to be a trend for the highest variance to be in lower speed limit states in addition to having an

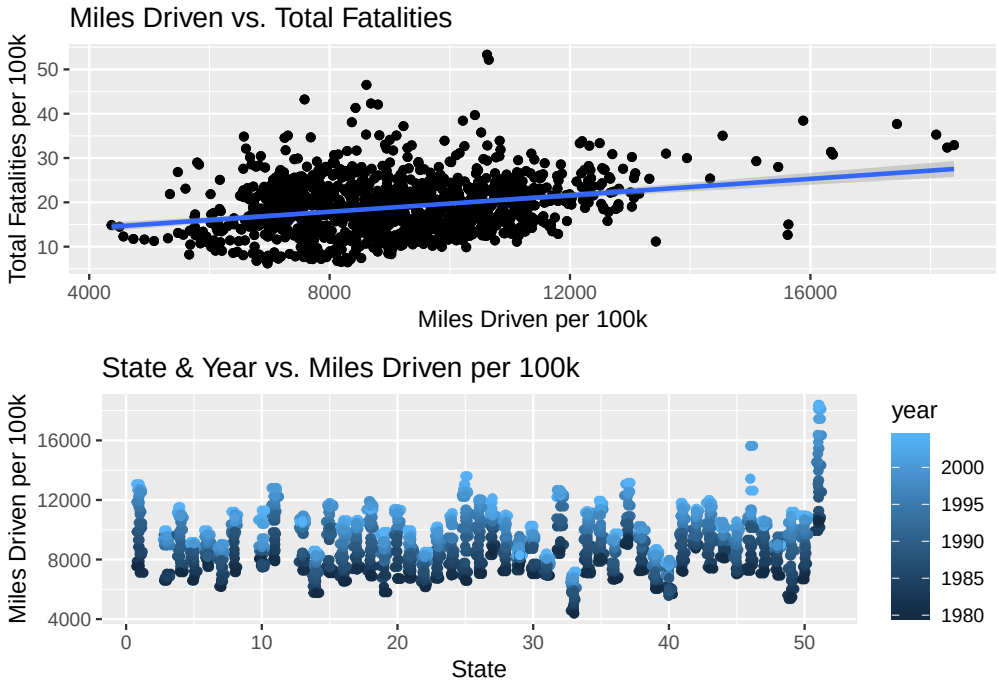
very slightly inverse relationship between the speed limit and fatalities.

We clearly see that there is a variation between state and year of the speed limit. As the years progressed, the speed limits generally rose - probably due to maturing roadways that were more like interstates rather than two-lane highways - and each state has significant variation in the later years. (We note that for the first few years, there is very little effect between states.)

1.3.2.2 Miles Driven

```
g1 <- ggplot(raw.df, aes(x = vehicmilespc, y = totfatrte)) + ylab('Total Fatalities per 100k') + xlab('
g2 <- ggplot(raw.df, aes(x = state, y = vehicmilespc, colour=year)) + ylab('Miles Driven per 100k') + x

grid.arrange(g1,g2,nrow=2)
```

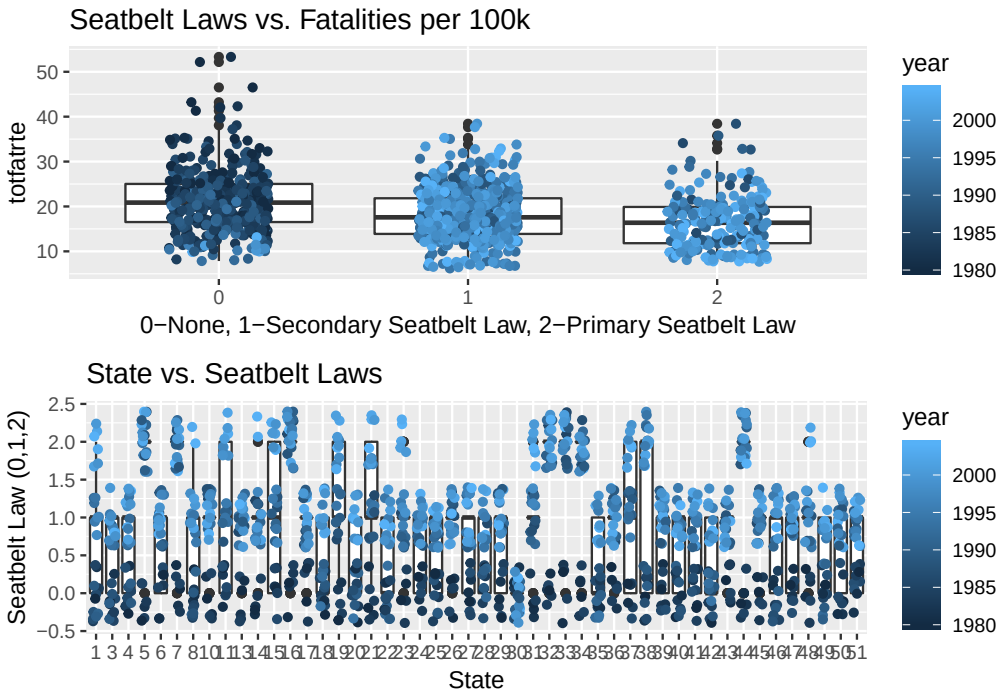


This top visual shows the relationship between miles driven per 100k and the total fatalities per 100k. There is a clear directly proporational relationship between the miles driven and the total fatalities. The bottom graph shows the Miles Driven per 100k over states and years. As years progress, the total miles driven per 100k tend to increase in each state, in addition to the fact that there are clear differences between each state on miles driven per 100k.

1.3.2.3 Seatbelt Laws

The raw `seatbelt` feature has the value 0 representing no seatbelt law, 1 representing primary seatbelt law, and 2 representing secondary seatbelt law. Because primary seatbelt law is more restrictive than secondary seatbelt law - a police officer can pull a driver over for lack of seatbelt and impose a greater fine - we decided to recode these laws to be in order of least restrictive to most restrictive. Therefore, a new variable `seatbelt_asc` was created with the value 0 representing no seatbelt law, 1 representing secondary seatbelt law, and 2 representing a primary seatbelt law. The code to do this transformation is below. Please note this recoding was *just* done for EDA purposes to see how the different restrictions related to total fatalities, state and year in a more intuitive order. The code that handled this recoding is hidden in the PDF for sake of space but is included in the .Rmd file for reference.

```
g1 <- ggplot(df.seatbelt, aes(as.factor(seatbelt_asc), totfatrte, colour=year)) + geom_boxplot() + geom_
xlab("0-None, 1-Secondary Seatbelt Law, 2-Primary Seatbelt Law")
g3 <- ggplot(df.seatbelt, aes(as.factor(state), seatbelt_asc, colour=year)) + geom_boxplot() + geom_jit
ylab("Seatbelt Law (0,1,2)") + xlab('State')
grid.arrange(g1, g3, nrow=2)
```



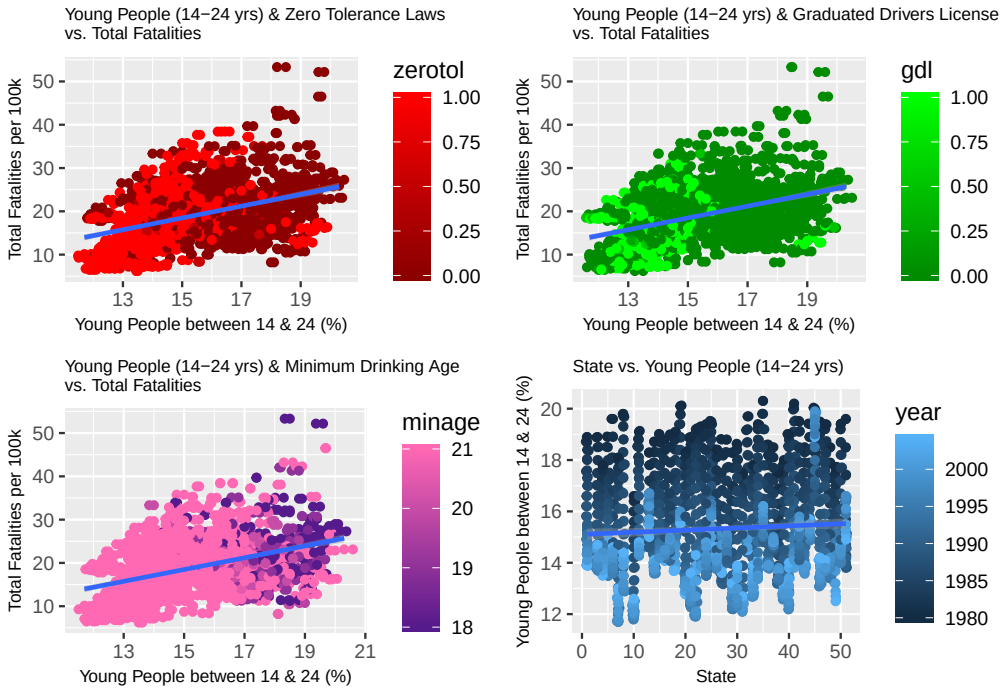
This relationship between seatbelt laws and total fatalities shows that when there is no seatbelt law, the median total fatalities per 100,000 is higher, along with the maximum total fatalities. The secondary seatbelt law has a higher median for total fatalities. Lastly, the primary seatbelt law, where police can pull a driver over for not having seatbelts fastened and carries the highest fine, has the lowest median total fatalities and smallest variation in total fatalities.

We also assess the effect of time and location here by looking at the seatbelt laws change over time and in each state. We see that there is a trend of greater restriction imposed in seatbelt laws over time. In the beginning of the 1980s there were almost no seatbelt laws and by the late 1990s and early 2000s, there were many more primary seatbelt laws. We also see clear differences in seat belt laws over states. Some states kept no seatbelt law throughout the timeframe captured in the dataset, while others had no seatbelt law all the way to the strictest seatbelt law during timeframe captured.

1.3.2.4 Young People, Zero Tolerance Laws, Graduated Driver's License

```
g1 <- ggplot(raw.df, aes(x = perc14_24, y = totfatrte, colour = zerotol)) + ylab('Total Fatalities per
g2 <- ggplot(raw.df, aes(x = perc14_24, y = totfatrte, colour = gdl)) + ylab('Total Fatalities per 100k
g4 <- ggplot(raw.df, aes(x = perc14_24, y = totfatrte, colour = minage)) + ylab('Total Fatalities per 1
g3 <- ggplot(raw.df, aes(x = state, y = perc14_24, colour = year)) + ylab('Young People between 14 & 24

grid.arrange(g1, g2, g4, g3, nrow=2)
```



There are four plots above: three (3) of which are showing the relationship between proportion of young people & total fatalities per 100k, the fourth showing the relationship between location & time on the proportion of young people. The top left shows the relationship between young people & total fatalities colored by the presence of a zero tolerance law (i.e. no BAC allowed for anyone under legal drinking age). The top right shows the the same relationship colored by the presence of a graduated driver's license law and the bottom left shows the relationship colored by the minimum drinking age.

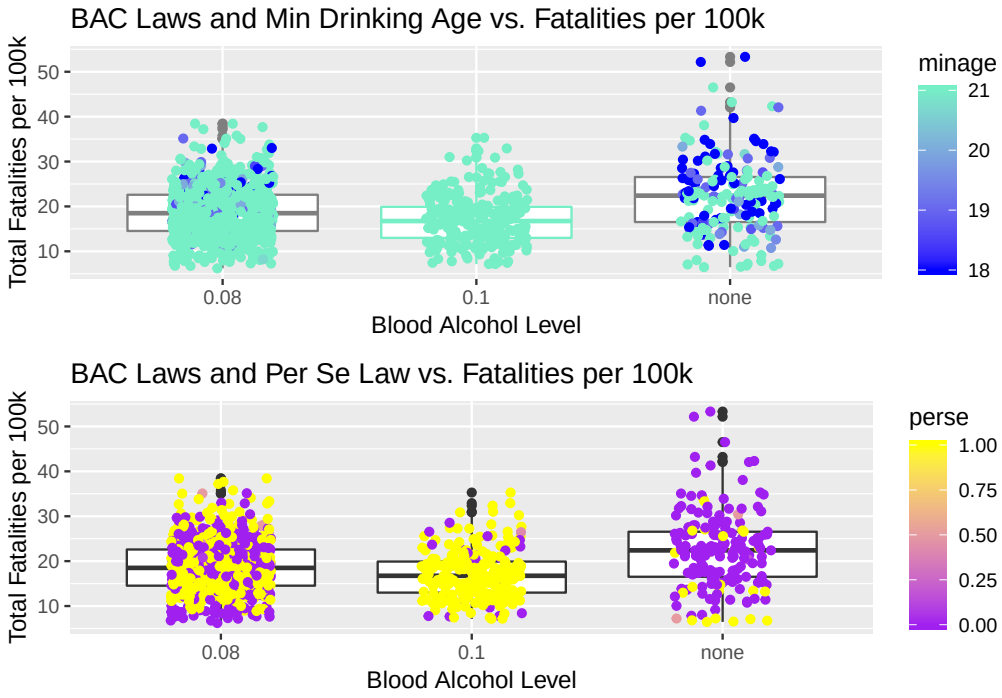
The relationship between young people & the total fatalities clearly shows that the total fatalities per 100k increases with the higher presence of young people. We see that zero tolerance law are more prevalent in places where the percentage of young people is lower, along with the minimum age to consume alcohol being higher in places where the percent of young people is lower. This relationship, howver, is not what could be taken for at face value. We want to confirm if this is truly a relationship or if there is another factor at play. So, we plot the bottom right graph to see if there was a demographic change during the same time laws were changing. We can confirm that there was a demographic shift (i.e. less people between 14 and 24) happening when these laws were being instituted.

1.3.2.5 Blood Alcohol Laws

In order to derive the Blood Alcohol Content (BAC) Law categories below, each row of the dataset was assessed to pick the max value of `bac08` and `bac10` and choosing the BAC law with the max proportion of effect in the state for that year to get a clearer picture of the effect of the BAC laws on total fatalities. The code for this is hidden below due to amount of space captured by the code but is included in .Rmd file for reference.

```
g1 <- ggplot(df.bacmax, aes(as.factor(bac), totfatrte,  colour = minage)) + geom_boxplot() +
geom_jitter(width = 0.2) + ggtitle("BAC Laws and Min Drinking Age vs. Fatalities per 100k") +
xlab("Blood Alcohol Level") + ylab("Total Fatalities per 100k") + scale_color_gradient(low='blue',high=
g2 <- ggplot(df.bacmax, aes(as.factor(bac), totfatrte,  colour = perse)) + geom_boxplot() +
geom_jitter(width = 0.2) + ggtitle("BAC Laws and Per Se Law vs. Fatalities per 100k") +
xlab("Blood Alcohol Level") + ylab("Total Fatalities per 100k") + scale_color_gradient(low='purple',high

grid.arrange(g1,g2,nrow=2)
```



The top graph shows that the total fatalities does not vary much between the states with a BAC law of 0.10 and 0.08, but the mean is noticeably higher for states without a 0.08 or 0.1 max BAC law. Layered into this display is the minimum drinking age denoted in color. For states with a higher minimum drinking age, the max total fatalities is lower. We also notice that with the stricter BAC laws comes stricter drinking age laws.

The graph on the bottom shows the same relationship between BAC and total fatalities per 100k, but instead of having the color hue related to the minimum drinking age, it's the presence of a *per se* law in the state for a given year. There doesn't seem to be a clear relational indication of *per se* vs. total fatalities here because states with *per se* and without *per se* both have fatalities ranging from the lowest end of the spectrum to highest end. We do see that states without a BAC law of 0.08 or 0.10 have a higher prevalance of no *per se* law vs. having a *per se* law.

1.3.2.6 Unemployment

Finally, we assess unemployment. In order to do so, we create two plots: unemployment rate and total fatalities and unemployment rates across states colored by year.

```
g1 <- ggplot(raw.df, aes(x = unem, y = totfatrte)) +
        ylab('Total Fatalities per 100k') +
        xlab('Unemployment (%)') +
        ggtitle('Unemployment vs. Total Fatalities') +
        geom_jitter(width=0.3)  +
        geom_point() +
        geom_smooth(method = "lm")
g2 <- ggplot(raw.df, aes(x = state, y = unem, colour=year)) + ylab('Unemployment (%)') + xlab('State') +
        ggtitle('State & Year vs. Unemployment') +
        geom_jitter(width=0.3)  +
        geom_point()

grid.arrange(g1,g2,nrow=2)
```

In the top graph, we see that as the unemployment rate rises, the total fatalities per 100,000 also rises. To dig further into this relationship, we then plot the unemployment rate over states hued by year. In the bottom plot, we see that over time, the unemployment rate is generally decreasing in all states. This makes sense because there was an economic upturn in the 1990s through to the early 2000s where we'd expect to see the unemployment drop. In the modeling, it will be interesting to explore if unemployment has a statistically significant relationship to total fatalities per 100,000 because here it seems that they have a notable relationship.

2. Exercise 2

2.1 How is `totfatrte` defined?

The variable `totfatrte` is total fatalities per 100,000 population.

2.2 What is average of this variable each year within time period covered in data?

```
aggregate(x = list(TotalFatalitiesPer100k = raw.df$totfatrte), list(Year = raw.df$year), mean)
```

```
##    Year TotalFatalitiesPer100k
## 1  1980               25.49458
## 2  1981               23.67021
## 3  1982               20.94250
## 4  1983               20.15292
## 5  1984               20.26750
## 6  1985               19.85146
## 7  1986               20.80042
## 8  1987               20.77479
## 9  1988               20.89167
## 10 1989               19.77229
## 11 1990               19.50521
```

```
## 12 1991               18.09479
## 13 1992               17.15792
## 14 1993               17.12771
## 15 1994               17.15521
## 16 1995               17.66854
## 17 1996               17.36938
## 18 1997               17.61062
## 19 1998               17.26542
## 20 1999               17.25042
## 21 2000               16.82562
## 22 2001               16.79271
## 23 2002               17.02958
## 24 2003               16.76354
## 25 2004               16.72896
```

2.3 Linear Regression Model

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

2.3.1 Estimate the model

```
mod.dum <- 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=raw.df)
summary(mod.dum)
```

```
##
## 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 = raw.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.4946     0.8671  29.401  < 2e-16 ***
## d81          -1.8244     1.2263  -1.488 0.137094
## d82          -4.5521     1.2263  -3.712 0.000215 ***
## d83          -5.3417     1.2263  -4.356 1.44e-05 ***
## d84          -5.2271     1.2263  -4.263 2.18e-05 ***
## d85          -5.6431     1.2263  -4.602 4.64e-06 ***
## d86          -4.6942     1.2263  -3.828 0.000136 ***
## d87          -4.7198     1.2263  -3.849 0.000125 ***
## d88          -4.6029     1.2263  -3.754 0.000183 ***
## d89          -5.7223     1.2263  -4.666 3.42e-06 ***
## d90          -5.9894     1.2263  -4.884 1.18e-06 ***
## d91          -7.3998     1.2263  -6.034 2.14e-09 ***
## d92          -8.3367     1.2263  -6.798 1.68e-11 ***
## d93          -8.3669     1.2263  -6.823 1.43e-11 ***
## d94          -8.3394     1.2263  -6.800 1.66e-11 ***
```

```
## d95          -7.8260     1.2263  -6.382 2.51e-10 ***
## d96          -8.1252     1.2263  -6.626 5.25e-11 ***
## d97          -7.8840     1.2263  -6.429 1.86e-10 ***
## d98          -8.2292     1.2263  -6.711 3.01e-11 ***
## d99          -8.2442     1.2263  -6.723 2.77e-11 ***
## d00          -8.6690     1.2263  -7.069 2.67e-12 ***
## d01          -8.7019     1.2263  -7.096 2.21e-12 ***
## d02          -8.4650     1.2263  -6.903 8.32e-12 ***
## d03          -8.7310     1.2263  -7.120 1.88e-12 ***
## d04          -8.7656     1.2263  -7.148 1.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.008 on 1175 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1098
## F-statistic: 7.164 on 24 and 1175 DF,  p-value: < 2.2e-16
```

2.3.2 What does model explain?

This model with the dummy time variables explains change in the total fatality rate over time. The base year is 1980 so the coefficients represent the change in the total fatality rate relative to 1980.

2.3.3 Model Findings

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

The coefficients for the year dummy variables are all negative. Aditionally, the coefficients for the dummy variables are all significant at the 0.1 % level, except for 1981. This means that, relative to the reference year of 1980, the total fatality rate fell in subsequent years. That the absolute value of the coefficient is larger in later years is unsurprising because as the rate falls further over time it becomes farther away from the initial level in 1980. However, the decline is not monotonic. We can see in several places, especially the 1990s and early 2000s, that the total fatality rate increased in some years relative to the prior year. However, there could be other omitted variables that explain this decline in the fatality rate. For example, car safety systems could have been improved. So there could be the same number of potentially fatal accidents but people simply survive now due to better safety equipment. In summary, driving did become safer over time according to this model however there could be other factors at work.

3. Exercise 3

3.1 Expanding Model

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*

Having conducted our EDA, we concluded that no transformations were required. First, for any binary or quasi binary variable, transformations seem unnecessary given the structure (a state either has a law or does not) and the ease it creates in model interpretation. For example, in the case of BAC laws, if a state has a 0.08 BAC law for half the year (`bac08` = 0.5) and a 0.10 BAC law for the other half of the year (`bac10` = 0.5), there is no such thing as a 0.09 BAC law, which is what would happen if we normalized these two variables into a weighted average. Therefore, we decided to keep these as-is.

Second, variables provided as percentages are close to neither zero nor one and so do not need to be stretched or compressed respectively. Finally, variables provided on a per capita basis are on the same basis (per 100,000) as the dependent variable. If they were not we would consider transforming them so they shared the same basis.

```
model.ols <- plm(totfatrte ~  d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +
   d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 +
  d02 + d03 + d04 +bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem + vehic
summary(model.ols)
```

```
## Pooling Model
##
## Call:
## plm(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 + bac08 + bac10 +
##     perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##     vehicmilespc, data = raw.df, model = "pooling")
##
## Balanced Panel: n = 25, T = 48, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -14.91602  -2.73839  -0.27779   2.28591  21.42027
##
## Coefficients:
##                 Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)  -2.7161e+00  2.4758e+00  -1.0970 0.2728472
## d81          -2.1755e+00  8.2761e-01  -2.6286 0.0086859 **
## d82          -6.5960e+00  8.5340e-01  -7.7290 2.330e-14 ***
## d83          -7.3967e+00  8.6902e-01  -8.5115 < 2.2e-16 ***
## d84          -5.8504e+00  8.7634e-01  -6.6760 3.792e-11 ***
## d85          -6.4833e+00  8.9480e-01  -7.2455 7.820e-13 ***
## d86          -5.8528e+00  9.3067e-01  -6.2888 4.516e-10 ***
## d87          -6.3674e+00  9.6696e-01  -6.5850 6.869e-11 ***
## d88          -6.5916e+00  1.0137e+00  -6.5024 1.170e-10 ***
## d89          -8.0710e+00  1.0526e+00  -7.6675 3.684e-14 ***
## d90          -8.9587e+00  1.0770e+00  -8.3185 2.463e-16 ***
## d91          -1.1069e+01  1.1012e+00 -10.0517 < 2.2e-16 ***
## d92          -1.2878e+01  1.1225e+00 -11.4728 < 2.2e-16 ***
## d93          -1.2731e+01  1.1363e+00 -11.2038 < 2.2e-16 ***
## d94          -1.2365e+01  1.1572e+00 -10.6849 < 2.2e-16 ***
## d95          -1.1953e+01  1.1836e+00 -10.0985 < 2.2e-16 ***
## d96          -1.3876e+01  1.2233e+00 -11.3430 < 2.2e-16 ***
## d97          -1.4258e+01  1.2498e+00 -11.4085 < 2.2e-16 ***
## d98          -1.5042e+01  1.2655e+00 -11.8861 < 2.2e-16 ***
## d99          -1.5091e+01  1.2843e+00 -11.7499 < 2.2e-16 ***
## d00          -1.5444e+01  1.3053e+00 -11.8314 < 2.2e-16 ***
## d01          -1.6184e+01  1.3340e+00 -12.1314 < 2.2e-16 ***
## d02          -1.6724e+01  1.3480e+00 -12.4065 < 2.2e-16 ***
## d03          -1.7021e+01  1.3595e+00 -12.5206 < 2.2e-16 ***
## d04          -1.6711e+01  1.3870e+00 -12.0488 < 2.2e-16 ***
## bac08        -2.4985e+00  5.3751e-01  -4.6483 3.729e-06 ***
## bac10        -1.4176e+00  3.9633e-01  -3.5768 0.0003622 ***
## perse        -6.2011e-01  2.9820e-01  -2.0795 0.0377907 *
```

```
## sbprim       -7.5335e-02  4.9078e-01  -0.1535 0.8780318
## sbsecon       6.7280e-02  4.2930e-01   0.1567 0.8754918
## sl70plus      3.3479e+00  4.4517e-01   7.5205 1.086e-13 ***
## gdl          -4.2691e-01  5.2691e-01  -0.8102 0.4179781
## perc14_24     1.4159e-01  1.2268e-01   1.1542 0.2486752
## unem          7.5705e-01  7.7906e-02   9.7176 < 2.2e-16 ***
## vehicmilespc  2.9254e-03  9.4968e-05  30.8042 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    48612
## Residual Sum of Squares: 19067
## R-Squared:      0.60778
## Adj. R-Squared: 0.59633
## F-statistic: 53.0957 on 34 and 1165 DF, p-value: < 2.22e-16
```

3.1.2 bac8 and bac10

How are the variables bac8 and bac10 defined? Interpret the coefficients on bac8 and bac10.

The variables `bac08` and `bac10` indicate the legal limit for blood acohol content in a state. The coefficient for `bac08` is -2.498. This implies that ceteris paribus, having a legal blood alcohol content limit of 0.08 reduces on average the total fatalities per 100,000 by 2.498 fatalities. The coefficient for `bac10` is -1.418. Similarly, this implies that ceteris paribus, having a legal blood alcohol content limit of 0.1 reduces on average the total fatalities per 100,000 by 1.418 fatalities. Both coefficients are statistically significant at the 0.1 % level.

3.1.3 Effects of Laws

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

The coefficient on `perse` is -0.6201. This implies that all else equal, *per se laws* do have a negative effect on the fatality rate, reducing it on average by 0.6201 fatalities per 100,000. This effect is significant at the 5% level.

3.1.3.2 Do primary seatbelt laws have negative effect on fatality rate?

The coefficient on `sbprim` is -0.07533. This means that ceteris paribus, seatbelt laws have negative effect on the fatality rate, reducing it on average by 0.07533 fatalities per 100,000. This effect however is not statistically significant at any level so this conclusion is not valid.

4. Exercise 4

4.1 Fixed Effects Model

Reestimate Exercise 3 model using fixed effects (at state level) model.

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

```
## Oneway (individual) effect Within Model
##
## Call:
```

```
## plm(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 + bac08 + bac10 +
##     perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##     vehicmilespc, data = raw.df, model = "within", index = "state")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -8.4273592 -1.0258600 -0.0029547  0.9572345 14.8109310
##
## Coefficients:
##                  Estimate  Std. Error  t-value  Pr(>|t|)
## d81          -1.51107133  0.41321486  -3.6569 0.0002672 ***
## d82          -3.02549578  0.44243119  -6.8383 1.316e-11 ***
## d83          -3.50360069  0.45657705  -7.6736 3.628e-14 ***
## d84          -4.25936110  0.46494255  -9.1610 < 2.2e-16 ***
## d85          -4.72679311  0.48547032  -9.7365 < 2.2e-16 ***
## d86          -3.66118540  0.51769787  -7.0721 2.686e-12 ***
## d87          -4.30578838  0.55532856  -7.7536 2.001e-14 ***
## d88          -4.76712131  0.60155650  -7.9246 5.501e-15 ***
## d89          -6.12997263  0.64019069  -9.5752 < 2.2e-16 ***
## d90          -6.22973766  0.66485076  -9.3701 < 2.2e-16 ***
## d91          -6.91714040  0.68195432 -10.1431 < 2.2e-16 ***
## d92          -7.77417239  0.70288580 -11.0604 < 2.2e-16 ***
## d93          -8.09410864  0.71594741 -11.3055 < 2.2e-16 ***
## d94          -8.50421668  0.73410866 -11.5844 < 2.2e-16 ***
## d95          -8.25540198  0.75623634 -10.9164 < 2.2e-16 ***
## d96          -8.60661913  0.79594975 -10.8130 < 2.2e-16 ***
## d97          -8.70781739  0.81975686 -10.6224 < 2.2e-16 ***
## d98          -9.34924025  0.83373487 -11.2137 < 2.2e-16 ***
## d99          -9.47489124  0.84399083 -11.2263 < 2.2e-16 ***
## d00          -9.99185979  0.85606370 -11.6719 < 2.2e-16 ***
## d01          -9.63121721  0.87255395 -11.0380 < 2.2e-16 ***
## d02          -8.90673015  0.88205263 -10.0977 < 2.2e-16 ***
## d03          -8.93650263  0.88994687 -10.0416 < 2.2e-16 ***
## d04          -9.33936115  0.91107045 -10.2510 < 2.2e-16 ***
## bac08        -1.43722116  0.39421213  -3.6458 0.0002788 ***
## bac10        -1.06266776  0.26883763  -3.9528 8.208e-05 ***
## perse        -1.15161719  0.23398721  -4.9217 9.867e-07 ***
## sbprim       -1.22739974  0.34271485  -3.5814 0.0003564 ***
## sbsecon      -0.34970784  0.25217091  -1.3868 0.1657826
## sl70plus     -0.06253283  0.26931063  -0.2322 0.8164283
## gdl          -0.41177619  0.29257391  -1.4074 0.1595790
## perc14_24     0.18712169  0.09509969   1.9676 0.0493567 *
## unem         -0.57183997  0.06057851  -9.4397 < 2.2e-16 ***
## vehicmilespc  0.00094005  0.00011104   8.4656 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4535.3
## R-Squared:      0.62624
```

```
## Adj. R-Squared: 0.59916
## F-statistic: 55.0943 on 34 and 1118 DF, p-value: < 2.22e-16
```

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

```
print("Pooled OLS Estimates")
```

```
## [1] "Pooled OLS Estimates"
```

```
coef(summary(model.ols))[c('bac08','bac10','perse','sbprim'), ]
```

```
##           Estimate Std. Error    t-value     Pr(>|t|)
## bac08  -2.49848306  0.5375054 -4.6482941 3.728596e-06
## bac10  -1.41756515  0.3963277 -3.5767503 3.621614e-04
## perse  -0.62010810  0.2982021 -2.0794896 3.779065e-02
## sbprim -0.07533472  0.4907848 -0.1534985 8.780318e-01
```

```
print("State Fixed Effects Estimates")
```

```
## [1] "State Fixed Effects Estimates"
```

```
coef(summary(model.fe))[c('bac08','bac10','perse','sbprim'), ]
```

```
##         Estimate Std. Error   t-value     Pr(>|t|)
## bac08  -1.437221  0.3942121 -3.645807 2.788147e-04
## bac10  -1.062668  0.2688376 -3.952824 8.208262e-05
## perse  -1.151617  0.2339872 -4.921710 9.867302e-07
## sbprim -1.227400  0.3427149 -3.581402 3.564423e-04
```

The coefficients estimated by the state fixed effect model differ from the pooled OLS estimates. For the two blood acohol content measures, `bac08` and `bac10`, the coefficients are smaller than the pooled OLS estimates. In the case of `bac08` the difference is 1.06, or about 1 fatality per 100,000 people. According to the state fixed effects model these blood alcohol concentration laws did not reduce the fatality rate by as much as the pooled OLS model indciates. The coefficients for `perse` and `sbprim` are larger than the pooled OLS estimates, implying that these measures reduce the fatality rate by more than the pooled OLS model states. In the case of a primary seatbelt laws this difference was 1.15 fatalities per 100,000 people.

4.3 Which set of estimates is more reliable?

The estimates from the State Fixed Effects model are more reliable. The Pooled OLS model ignores the heterogeneity of the individual states . As we saw in `1.3 Exploratory Data Analysis`, there is considerable variation across the states. Pooling them together throws this information away resulting in less precise measures. We can observe this by reviewing the standard errors for each coefficient. The standard errors are smaller for all coefficients in the state fixed effects model so these estimates are more precise. However, we still have to review the model assumptions to confirm that the standard errors do not require correction.

4.4 What assumptions are needed in each of the models?

4.4.1 Pooled OLS

1) *Model Specification* - The model can be written as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$ where β_0, $\beta_1 \ldots \beta_k$ are the parameters of interest and u is the unobserved random error.

2) *Random Sample* - We assume we have a random sample of n observations that follows the model specified above.

3) *No Perfect Collinearity* - None of the explanatory variables are constant and there is no exact linear relationship among any of them.

4) *Zero Conditional Mean* - The error term has an expected value of zero given any values of the explanatory variables. $E(u|x_1, x_2, \dots x_k) = 0$

5) *Homoskedasticity* - The error u has the same variance given any values of the explanatory variables. $\text{Var}(u|x_1, x_2, \dots x_k) = \sigma^2$

4.4.2 Fixed Effects

1) *Model Specification* - For each i the model is $y_{it} = \beta_1 x_{it1} + \dots + \beta_k x_{itk} + a_i + u_{it}, \quad t = 1, \dots, T$ where β_j are the parameter estimates and a_i is the unobserved effect.

2) *Random Sample* - We assume we have a random sample from the cross section.

3) *No Perfect Collinearity* - Each explanatory variable changes over time for at least some i and no perfect linear relationships exist among the explanatory variables.

4) *Strict Exogeneity* - For each t, the expected value of the error u_{it} given the explanatory variables in all time periods $\mathbf{X}_i$ and the unobserved effect a_i is zero. $E(u_{it}|\mathbf{X}_i, a_i) = 0$

5) *Homoskedasticity* - The variance of the idiosyncratic errors is constant across time and not conditional on either the explanatory variables or unobserved effect. $\text{Var}(u_{it}|\mathbf{X}_i, a_i) = \text{Var}(u_{it}) = \sigma_u^2$ for all $t = 1, \dots, T$.

6) *No Serial Correlation* - For all $t \neq s$ the idiosyncratic errors are uncorrelated, conditional on all explanatory variables and unobserved effect. $\text{Cov}(u_{it}, u_{is}|\mathbf{X}_i, a_i) = 0$

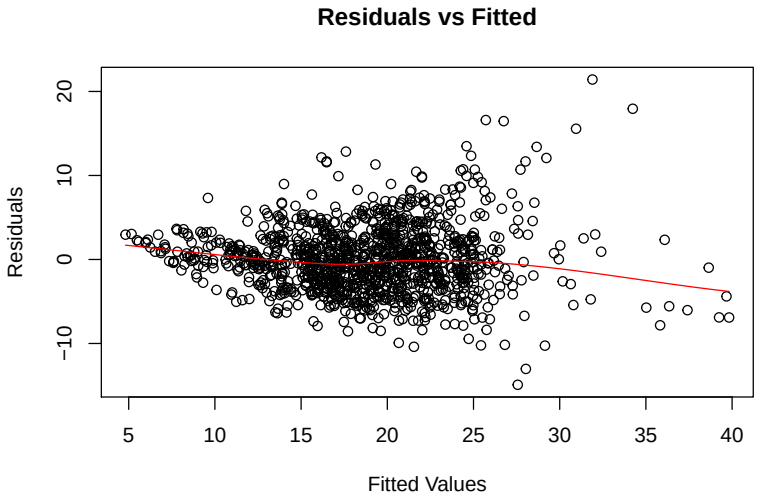
7) *Normality* - Conditional on all explanatory variables and the unobserved effect, the idiosyncratic errors are independent and identically distributed as $\text{Normal}(0, \sigma_u^2)$.

4.5 Are these assumptions reasonable in current context?

4.5.1 Reasonability of Pooled OLS Assumptions

Given the context, several of the assumptions made by the pooled OLS model seem unreasonable. First, Assumption 4 (Zero Conditional Mean) seems violated. The model tucks the unobserved effects on a state level into the error term. These unobserved state effects are conditional on the explanatory variables provided. So, the error term u has a nonzero expected value that *is* conditional on the explanatory variables $E(u|x_1, x_2, \dots, x_k) \neq 0$. We confirm this below with the Residuals vs Fitted plot which shows the residuals are not quite centered at zero. Additionally, the "cone" shape of the residuals suggests that Assumption 5 (Homoskedasticity) is similarly violated. We examine this next.

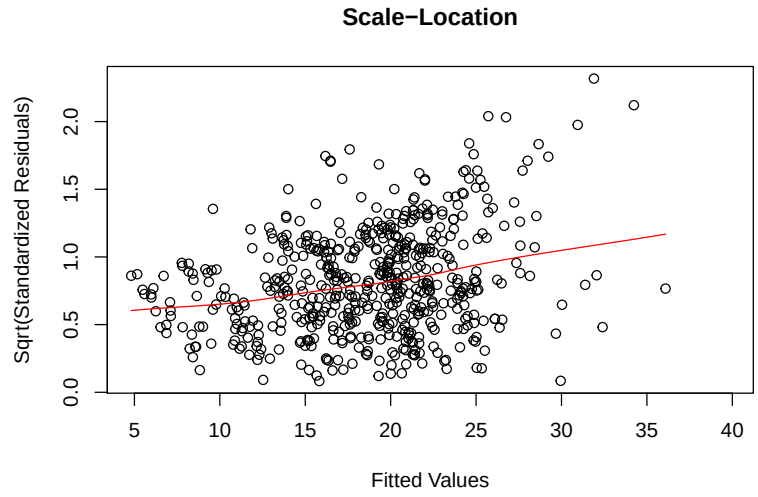
```
# zero conditional mean
scatter.smooth(fitted.values(model.ols), residuals(model.ols), xlab = "Fitted Values", ylab = "Residuals
```



We examine the possible presence of heteroskedasticity by reviewing the Scale-Location plot. The steep angle in the plot shows that the spread becomes wider. This is good evidence of heteroskedasticity. We confirm this with the Breusch–Pagan test below. The p-value is $2.2E-16$, so we reject the null hypothesis of homoskedasticity. Clustered standard errors may help correct these two violations, which we discuss more in Exercise 7.

```
# homoskedasticity
scatter.smooth(fitted.values(model.ols), sqrt(residuals(model.ols)/sd(residuals(model.ols))), main = "S
```

```
## Warning in sqrt(residuals(model.ols)/sd(residuals(model.ols))): NaNs
## produced
```




```
bptest(model.ols)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model.ols
## BP = 150.07, df = 34, p-value < 2.2e-16
```

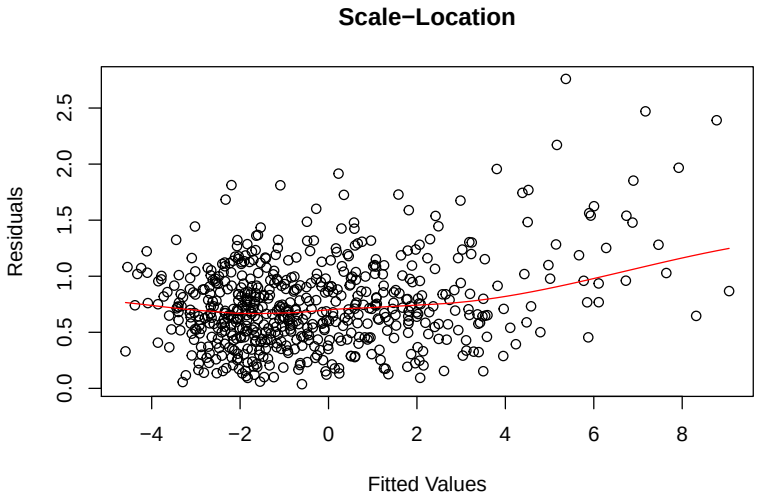
4.5.2 Reasonability of Fixed Effects Assumptions

On the surface, the fixed effect model assumptions are more reasonable given the heterogeneity in the

state level data. However we must still examine the assumptions in detail. In particular we want to see if Assumption 5 (Homoskedasticity), Assumption 6 (No Serial Correlation), and Assumption 7 (Normality), are violated. To observe any heteroskedasticity, we examine the Scale-Location plot below. The distribution of the residuals appears more evenly distributed than in the Pooled OLS model. However, there still does seem to be a change in the distribution of residuals. We confirm this potential heteroskedasticity with the Bresuch-Pagan test. The p-value is $2.2E-16$, so we reject the null hypothesis of homoskedasticity. Assumption 5 (Homoskedasticity) seems violated.

```
scatter.smooth(fitted.values(model.fe), sqrt(residuals(model.fe)/sd(residuals(model.fe))), main="Scale-
```

```
## Warning in sqrt(residuals(model.fe)/sd(residuals(model.fe))): NaNs produced
```




```
bptest(model.fe)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model.fe
## BP = 150.07, df = 34, p-value < 2.2e-16
```

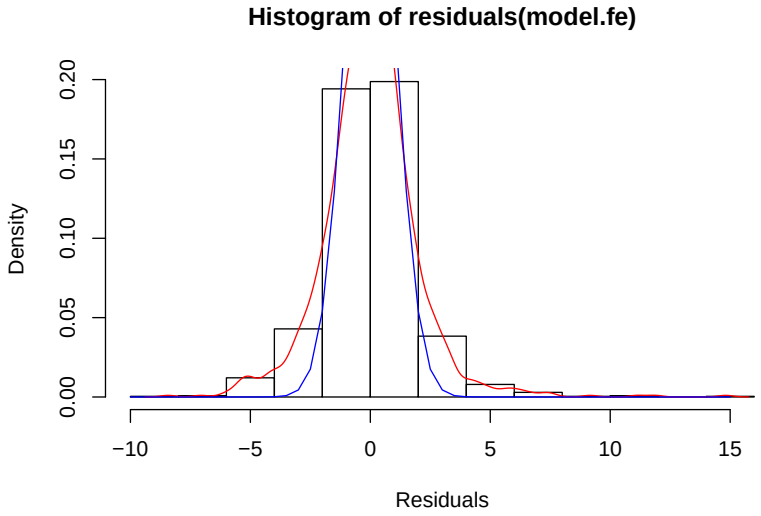
We next test Assumption 6 (No Serial Correlation). If the idiosyncraticc errors, conditional on all explanatory variables and unobserved effect, exhibit correlation then the model assumptions are violated and we cannot trust the errors. We can test for the presence of serial correlation in panel data models with the Bresuch-Godfrey test. The p-value is $2.2E-16$, so we reject the null hypothesis of no serial correlation in the idiosyncratic errors. Assumption 6 (No Serial Correlation) appears violated as well.

```
pbgtest(model.fe)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +     d90 + d91 + d92 + d93 +
## chisq = 87.862, df = 25, p-value = 6.369e-09
## alternative hypothesis: serial correlation in idiosyncratic errors
```

Finally, we examine the distribution of errors to test Assumption 7 (Normality). Given the prior two results, we should not anticipate the errors to be normally distributed. We overlay a normal distribution in blue and the distribution of the residuals in red. The residuals appear wider than approximately normal. We can confirm this with the Shaprio-Wilk test of normality. The p-value for the test is $2.2E-16$ so we reject the null hypothesis of normality. We cannot assume that the errors are normally distributed around zero.

```
# normality
hist(residuals(model.fe), freq = F, xlab = 'Residuals')
lines(density(residuals(model.fe)), col="red")
lines(seq(-10, 15, by=.5), dnorm(seq(-10, 15, by=.5), 0, 1), col="blue")
```




```
shapiro.test(residuals(model.fe))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.fe)
## W = 0.93739, p-value < 2.2e-16
```

We can correct for the violations of heteroskedasticiy and serial correlation using a heteroskedasticity consistent covariance estimator. We show the coefficients, standard errors, and p-values after making this correction below.

```
coeftest(model.fe, vcovHC(model.fe, method="arellano"))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## d81         -1.51107133  0.43628851  -3.4635 0.0005534 ***
## d82         -3.02549578  0.48661275  -6.2175 7.123e-10 ***
## d83         -3.50360069  0.50722346  -6.9074 8.267e-12 ***
## d84         -4.25936110  0.43860375  -9.7112 < 2.2e-16 ***
## d85         -4.72679311  0.45737157 -10.3347 < 2.2e-16 ***
## d86         -3.66118540  0.58416846  -6.2673 5.234e-10 ***
## d87         -4.30578838  0.67027727  -6.4239 1.961e-10 ***
## d88         -4.76712131  0.75197856  -6.3394 3.339e-10 ***
## d89         -6.12997263  0.85707933  -7.1522 1.541e-12 ***
## d90         -6.22973766  0.90514332  -6.8826 9.773e-12 ***
## d91         -6.91714040  0.97365427  -7.1043 2.149e-12 ***
## d92         -7.77417239  1.06203082  -7.3201 4.725e-13 ***
## d93         -8.09410864  1.08994968  -7.4261 2.212e-13 ***
## d94         -8.50421668  1.07290343  -7.9264 5.430e-15 ***
## d95         -8.25540198  1.15959440  -7.1192 1.938e-12 ***
## d96         -8.60661913  1.14818594  -7.4958 1.337e-13 ***
```

```
## d97          -8.70781739  1.19826853  -7.2670 6.884e-13 ***
## d98          -9.34924025  1.21240453  -7.7113 2.742e-14 ***
## d99          -9.47489124  1.33589783  -7.0925 2.331e-12 ***
## d00          -9.99185979  1.31554509  -7.5952 6.469e-14 ***
## d01          -9.63121721  1.43580699  -6.7079 3.130e-11 ***
## d02          -8.90673015  1.44602568  -6.1595 1.017e-09 ***
## d03          -8.93650263  1.48817832  -6.0050 2.584e-09 ***
## d04          -9.33936115  1.62460726  -5.7487 1.159e-08 ***
## bac08        -1.43722116  0.80266842  -1.7906 0.0736354 .
## bac10        -1.06266776  0.47955910  -2.2159 0.0268973 *
## perse        -1.15161719  0.43310154  -2.6590 0.0079493 **
## sbprim       -1.22739974  0.54500980  -2.2521 0.0245113 *
## sbsecon      -0.34970784  0.36181365  -0.9665 0.3339824
## sl70plus     -0.06253283  0.55279337  -0.1131 0.9099545
## gdl          -0.41177619  0.37386299  -1.1014 0.2709556
## perc14_24     0.18712169  0.16965909   1.1029 0.2702960
## unem         -0.57183997  0.11744736  -4.8689 1.283e-06 ***
## vehicmilespc  0.00094005  0.00033858   2.7765 0.0055868 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Exercise 5

Would you prefer to use random effects model instead of fixed effects model built in Exercise 4?

There are several factors to take into account when deciding which model to use. The goal of any model is to most accurately depict the relationships while still maintaining interpretability of the results. Fixed effects models are generally simpler to interpret but run the risk of providing inaccurately precise estimators. Random effects can account for study-level variability but typically at the cost of interpretability.

In order to determine which we'd prefer to use, we find it appropriate to assess primarily the following: is there variability beyond sampling error present from our data? To evaluate this question, we can use the Hausman test to test the null hypothesis that unique errors (u_i) are not correlated with the regressors. If we reach a p-value significance (here we use $\alpha = 0.05$) to reject the null hypothesis, then we are confident we have correlated errors (fixed effects).

In order to run the Hausman test, we will have to first define a random effects model comparable to the fixed effects model we ran in section `4.1 Fixed Effects Model`. The code below fits the model then runs the Hausman test.

```
# Define a random effects model
model.re <- plm(totfatrte ~  d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +
   d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 +
              bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl+ perc14_24 + unem + vehicmilespc

# Run a Hausman test
phtest(model.fe, model.re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +  ...
## chisq = 148.69, df = 34, p-value = 2.727e-16
## alternative hypothesis: one model is inconsistent
```

Because we have a p-value of $2.7e^{-}16$, we can safely reject the null hypothesis and therefore believe our errors are correlated. In this circumstance, we'd prefer to use the fixed effects model over the random effects model.

6. Exercise 6

Suppose that `vehicmilespc`*, the number of miles driven per capita, increases by 1,000. Using FE estimates, what is estimated effect on* `totfatrte`*? Please interpret estimate*

In the State fixed effects model in `4.1 Fixed Effects Model`, the coefficient of `vehicmilespc` is 0.00094. Therefore, when the number of miles driven per 100,000 increases by 1000, the total fatalities per 100,000 population (`totfatrte`) increases by 0.94, holding all else constant.

7. Exercise 7

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

When either serial correlation or heteroskedasticity are present in the model, the standard errors and p-values associated with coefficients are inaccurate. To correct for the presence of serial correlation or heteroskedasticity, clustered-robust standard errors should be used instead which will change the standard errors and usually affects the p-values for the estimators.