Panel Data Analysis of Traffic Fatalities

Stone Jiang

The purpose of this report is to apply and compare panel data regression techniques to modeling the importance of various numerical and legal factors in driving fatality rates across all 50 states. The procedure will be as follows:

1. We perform a very thorough exploratory panel data analysis in order to understand every possible caveat of the dataset.
2. We then move to an OLS model, pooling across states but controlling for year via indicator variables.
3. We then fit a fixed effects (FE) model, and discuss its advantages to pooled OLS such as heterogeneity bias, and assess all model assumptions through residual diagnostics of both models.
4. We compare a fixed effects regression model to a random effects regression model, and interpret the coefficients of the FE model, taking into account statistical significance.
5. We conclude with a discussion of standard errors associated with heteroskedasticity and serial correlation.

The data sets is provided by the textbook "Introductory Econometrics: A Modern Approach, 6e" by Jeffrey M. Wooldridge and cited below:

https://rdrr.io/cran/wooldridge/man/driving.html

U.S. traffic fatalities: 1980-2004

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

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

Exploratory Data Analysis

```
load('/Users/siduojiang/Desktop/Stone/Berkeley_MIDS/Time_Series_Panel/2020-spring-siduojiang/la
x <- c('knitr', 'ggplot2', 'GGally', 'tidyr', 'dplyr', 'patchwork', 'plm', 'car', 'sandwich',
o <- lapply(x, require, character.only = TRUE)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Our dependent variable is totfatrte, and the explanatory variables of interest include both numeric and indicator variables. First, we note that fatality is presented in multiple manners in this dataset, such as night fatality rates (nghtfat), weekend fatality rates (wkndfatrte), and as a unit of miles driven rather than population (such as totfatpvm). Since we will be interested in totfatrte, we will not look at these other measures, which are simply other ways to measure fatality, and are not independent explanatory variables that might be helpful for our regression models. In addition, for speed limit, we have indicators for whether speed limit laws of 55, 65, 70 and 75 are in effect. The variable we will focus on is sl70plus, which already aggregates these variables into a single indicator whether the speed limit was above 70, so we will focus on this variable. With that in mind, we now look at the tabular view of the dataset and interested variables.

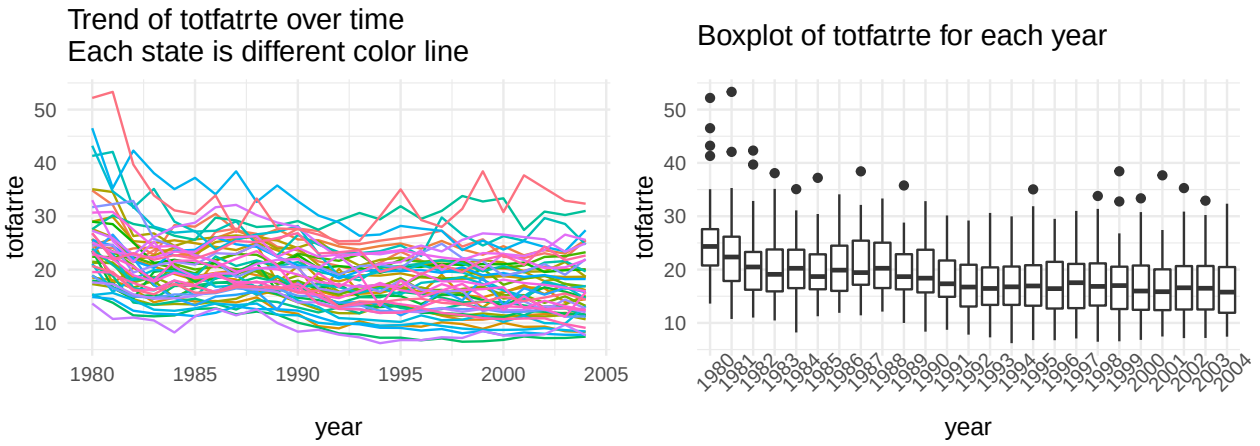
```
data.str <- data %>% dplyr::select(totfatrte, statepop, year,
    state, seatbelt, minage, zerotol, gdl, bac10, bac08, perse,
    vehicmilespc, unem, perc14_24, sl70plus, sbprim, sbsecon)
str(data.str, nchar.max = 50, give.attr = F)
```

```
## 'data.frame':    1200 obs. of  17 variables:
##  $ totfatrte   : num  24.1 24.1 21.4 23.6 23.6 ...
##  $ statepop    : int  3893888 3918520 3925218 3934109 395| __truncated__ ...
##  $ year        : int  1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
##  $ state       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ seatbelt    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ minage      : num  18 18 18 18 18 20 21 21 21 21 ...
##  $ zerotol     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ gdl         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ bac10       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ bac08       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ perse       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ vehicmilespc: num  7544 7108 7607 7880 8334 ...
```

```
##  $ unem        : num  8.8 10.7 14.4 13.7 11.1 ...
##  $ perc14_24   : num  18.9 18.7 18.4 18 17.6 ...
##  $ sl70plus    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ sbprim      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ sbsecon     : int  0 0 0 0 0 0 0 0 0 0 ...
```

Our dependent variable is numeric taking on decimal values, while our all of our explanatory variables are either integers or numeric. The integral variables are either categorical (such as state), or indicators (all other "law" variables) that represent whether or not a particular law or rule was in effect for that particular row of observation. Some "law" variables are currently marked as numeric, such as bac10, which is something we will deal with momentarily. First, we will look at the distribution of the dependent variable totfatrte across each of the years for each state in both a time series plot. We also generate a boxplot aggregated over each state for each year.

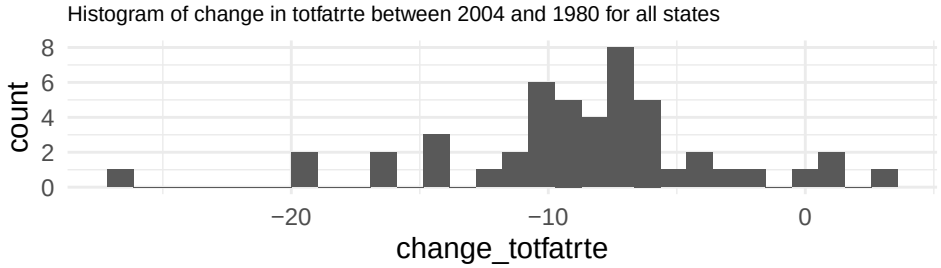
```
tsp <- ggplot(data = data, aes(x = year, y = totfatrte, colour = factor(state))) +
    geom_line() + ggtitle("Trend of totfatrte over time\nEach state is different color line") +
    theme_minimal() + theme(legend.position = "none")
hsp <- ggplot(data = data, aes(x = factor(year), y = totfatrte)) +
    geom_boxplot() + theme_minimal() + labs(x = "year", title = "Boxplot of totfatrte for each
    theme(legend.position = "none", axis.text.x = element_text(angle = 45))
tsp + hsp
```



Based on this time series plot above, we can see that most states have fatality rates that generally decline over time. However, there is quite a bit of fluctutation especially for states with overall high fatality rates (the blue, red and green lines at the top of the chart), but no line appears to be significant outliers. The boxplot reveals that in addition to the totfatrte decreasing overall and within-state fluctuation, the variations within each year also fluctuates, as indicated by the different sizes of the boxes for each year. In addition, in the earlier years before 1990, there were outlier(s) for almost every year, and outliers began appearing again after 1995. However, the proportion of outliers for each year is small. We can look the change in totfatrte for each state by generating a tabular form of the change between 2004 and 1980, and displaying this as a histogram.

```
change_totfatrte <- data %>% group_by(state) %>% dplyr::summarise(change_totfatrte = tail(totfa
    1) - head(totfatrte, 1))
ggplot(data = change_totfatrte, aes(x = change_totfatrte)) +
    geom_histogram() + theme_minimal() + ggtitle("Histogram of change in totfatrte between 200
```

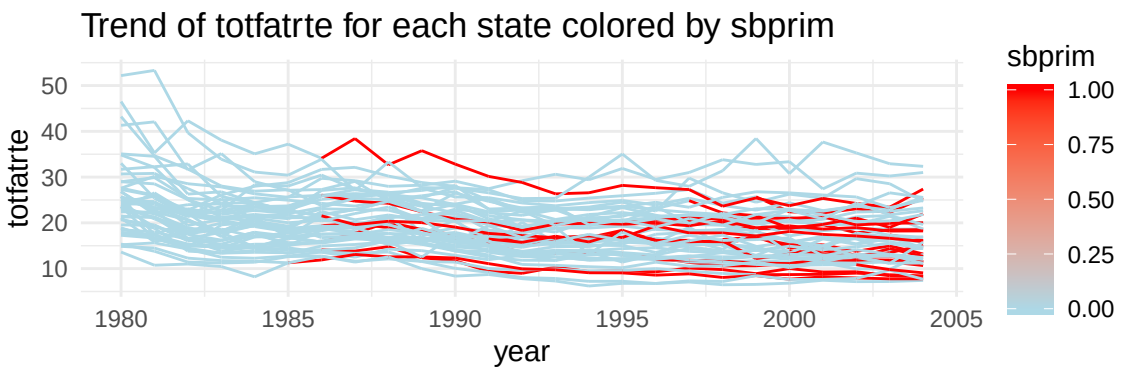
```
    theme(plot.title = element_text(size = 8))
```



We see that the change in totfatrte is indeed mostly negative, meaning rates for most states have dropped. There is 1 state for which the change between 2004 and 1980 was 0, and 3 for which the change is positive. The distribution of change is centered around -8, but is somewhat left skewed. There is one large outlier corresponding to state 29 which dropped by more than 25. This state started with a high value of 43.22, and dropped to around the average by 2004, at 16.92.

For the explanatory variables, we will first look at the indicator variables, where 1 represents the law is in effect, and 0 represents the law was not. The variable sbprim represents whether the state has a primary seatbelt law. We would expect that fatalities should decrease with implementation of seatbelt laws. We will look at a time plot of this variable with totfatrte.

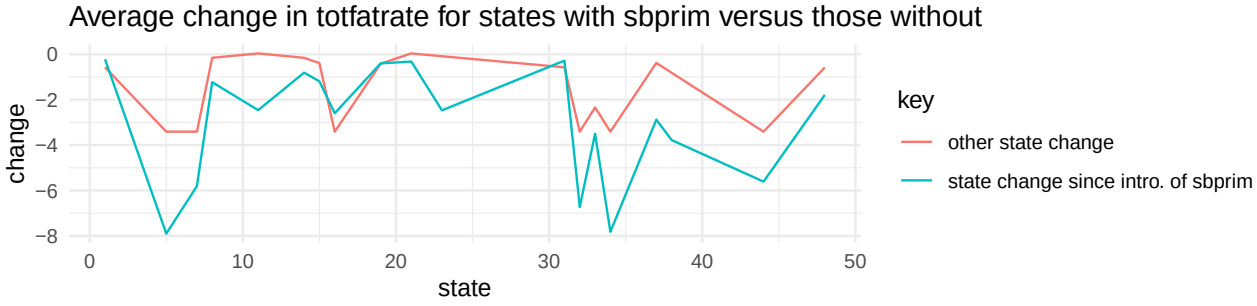
```
ggplot(data = data, aes(x = year, y = totfatrte, diff_lines = factor(state),
    color = sbprim)) + geom_line() + ggtitle("Trend of totfatrte for each state colored by sbpr
    scale_color_gradient(low = "lightblue", high = "red") + theme_minimal()
```



Based on the plot above, we see that for all states, once sbprim law was introduced, it was never banned. In addition, totfatrte for all states decreased since the time of introduction. For states that introduced this particular law, it would be interesting to see if the average rate of decrease for that state since the introduction of the law was similar to or different than states that never introduced this law for the same time period. For example, if a state introduced sbprim in 1999, we would like to compare the change of totfatrte in that state between 1999 and 2004 to the change in totfatrte for all states that never introduced sbprim between 1999 and 2004.

```
sbprim_sum <- data %>% group_by(state) %>% summarise(s = sum(sbprim))
states_with_sbprim <- unique(data$state)[sbprim_sum$s > 0]
# Get dataframe for states that never introduced sbprim
data_no_sb <- data %>% dplyr::filter(!state %in% states_with_sbprim)
sbprim_dat <- matrix(nrow = length(states_with_sbprim), ncol = 2)
ind <- 1
```

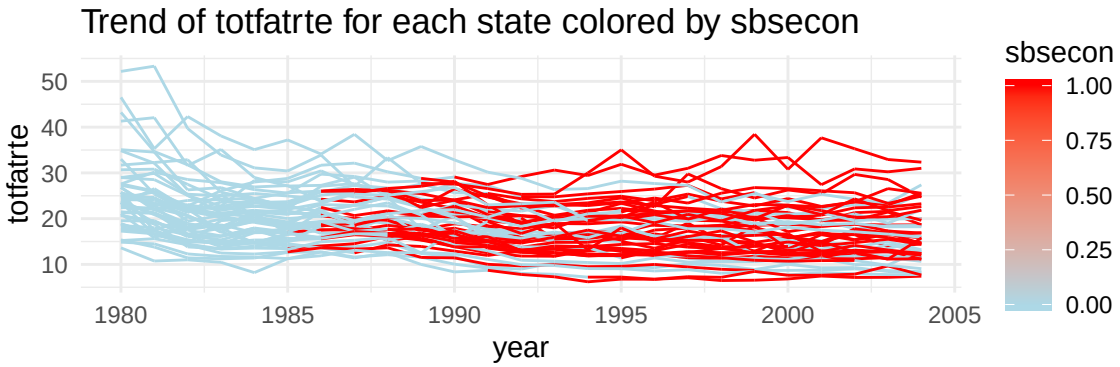
```
for (st in states_with_sbprim) {
    # Get first date for which sbprim was introduced
    year_introduced <- data %>% filter(state == st, sbprim ==
        1) %>% dplyr::select(year) %>% head(1) %>% as.integer()
    # Get average rate of change for the state since introduction
    # of the law
    state_change <- data %>% filter(state == st, year >= year_introduced) %>%
        dplyr::summarise(s = tail(totfatrte, 1) - head(totfatrte,
            1))
    sbprim_dat[ind, 1] <- state_change[1, ] %>% unname
    # Get average rate of change states that never introduced
    # sbprim during same time period
    states_change <- data_no_sb %>% filter(year >= year_introduced) %>%
        group_by(state) %>% dplyr::summarise(s = tail(totfatrte,
        1) - head(totfatrte, 1)) %>% dplyr::select(s) %>% dplyr::summarise(mean(s))
    sbprim_dat[ind, 2] <- states_change %>% as.numeric()
    ind <- ind + 1
}
sbprim_dat <- data.frame(sbprim_dat)
colnames(sbprim_dat) <- c("state change since intro. of sbprim",
    "other state change")
sbprim_dat$state <- states_with_sbprim
sbprimt_dat_plot <- gather(sbprim_dat, key, change, -state)
ggplot(sbprimt_dat_plot, aes(x = state, y = change, color = key)) +
    geom_line() + theme_minimal() + ggtitle("Average change in totfatrate for states with sbpri
```



This plot shows that for most states where sbprim was introduced, the average rate of change in totfatrat decreased more in that state since the introduction of that law, compared to the average rate of change for states that never introduced the law for the same time period. This indicates that this variable might be an important indicator in our regression.

Next, we look at the same plot for secondary seat belt law, sbsecon:

```
ggplot(data = data, aes(x = year, y = totfatrte, diff_lines = factor(state),
    color = sbsecon)) + geom_line() + ggtitle("Trend of totfatrte for each state colored by sbs
    scale_color_gradient(low = "lightblue", high = "red") + theme_minimal()
```



Here, we see that more states introduced sbsecon than states that introduced sbprim (39 versus 19). The decrease is less clear, since there are some states where the fatality rate seems to have increased since the introduction of the law (top of the chart). Also, there is one state where the secondary law was introduced, but removed at a later date (at the bottom of the plot). These all suggest that this variable may be less correlated than sbprim with fatality rate.

Next, we look at the impact of blood alcohol levels, bac08, bac10. These variables indicate the tolerated blood alcohol level, so they are never both 1 at the same time. They can both be 0, meaning there's no set level. In addition, some values are decimals, meaning the alcohol levels are implemented during the middle of the year. For the EDA, we will round decimals so that if the law went into effect during the middle of the year, if it is for half of the year or more, we consider the law to be in effect for the whole year. Some years, we have both bac08 and bac10 as 0.5 An example is shown below:

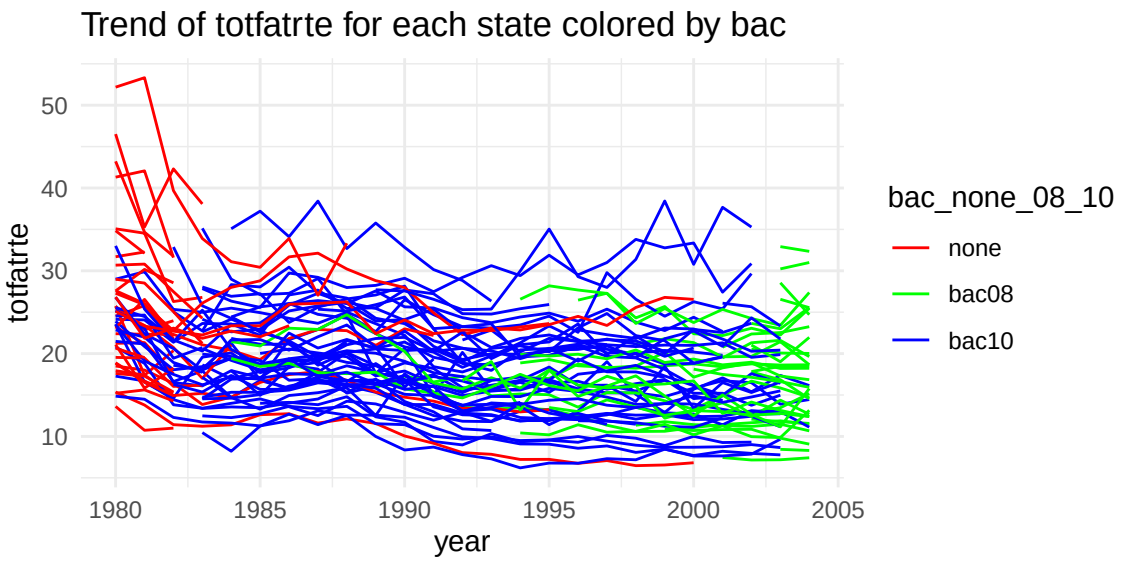
```
filter(data, bac10 == 0.5, bac08 == 0.5) %>% dplyr::select(year,
    state, bac10, bac08) %>% head(2)
```

```
##   year state bac10 bac08
## 1 2004     6   0.5   0.5
## 2 2002     7   0.5   0.5
```

In this case, for that particular year, we will take the previous year as rule (for example, if previous year was bac10, and current year transitioned to bac08 half way in between, we treat current year as bac10). This is because bac08 and bac10 cannot simultaneously be 1, and purely rounding both 0.5 up would result in this. We will generate a plot and color based on the level of tolerance (none, 0.08%, and 0.1%).

```
bac_none_08_10 <- matrix(ncol = 1, nrow = dim(data)[1])
prev_year <- 'none'
for (row in seq(1,dim(data)[1])){
  #Neither law was in effect this year
  if ((data[row,]$bac10 == 0) & (data[row,]$bac08 == 0)) {bac_none_08_10[row] <- 'none'}
  #Both equal 0.5
  else if (data[row,]$bac10 == data[row,]$bac08){bac_none_08_10[row] <- prev_year}
  #Majority of year was bac10
  else if (data[row,]$bac10 > data[row,]$bac08){bac_none_08_10[row] <- 'bac10'}
  else {bac_none_08_10[row] <- 'bac08'}
  prev_year <- bac_none_08_10[row]}
data$bac_none_08_10 <- factor(bac_none_08_10, c('none', 'bac08', 'bac10'))
```

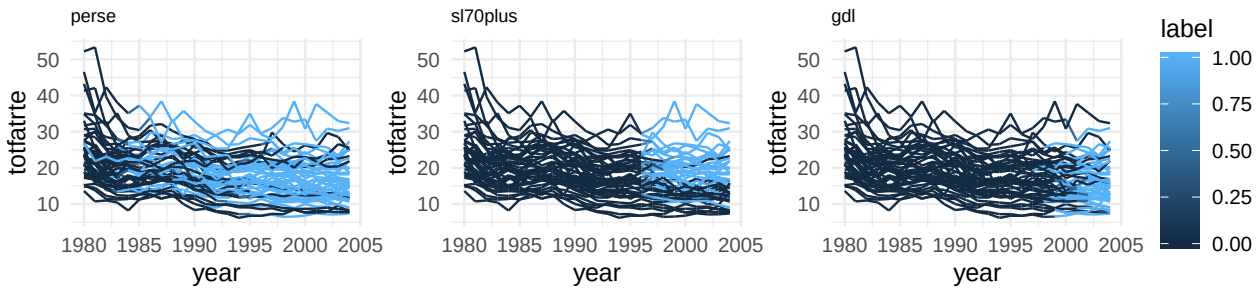
```
ggplot(data = data, aes(x = year, y = totfatrte, diff_lines = factor(state),color = bac_none_08
```



Based on this plot, most states started with bac10, and by the end of 2004, most have introduced bac08. For the few states that did not have bac laws until ~1995 or ~2000, their totfatrates are spread pretty evenly across the states for any given year. For states that already had bac10 in effect in 1980, their totfatrte were at around 30 or below already, indicating a lower starting baseline for those states. Overall, for observations where bac10 or bac08 was in effect, the totfatrte appears lower compared to, for example, the levels of states before 1983 without the law, indicating bac10 or bac08 could be an important explanatory variable.

We now look at 3 traffic laws, perse, sl70plus, gdl, which are per se law, whether the state has speed limit of 70 or above, and graduated drivers license law.

```
p1 <- ggplot(data = data, aes(x = year, y = totfatrte, diff_lines = factor(state),
    color = perse)) + guides(color = FALSE) + geom_line() + ggtitle("perse") +
    theme_minimal() + theme(plot.title = element_text(size = 8))
p2 <- ggplot(data = data, aes(x = year, y = totfatrte, diff_lines = factor(state),
    color = sl70plus)) + ggtitle("sl70plus") + geom_line() +
    guides(color = FALSE) + theme_minimal() + theme(plot.title = element_text(size = 8))
p3 <- ggplot(data = data, aes(x = year, y = totfatrte, diff_lines = factor(state),
    color = gdl)) + ggtitle("gdl") + labs(color = "label") +
    geom_line() + theme_minimal() + theme(plot.title = element_text(size = 8))
p1 + p2 + p3
```



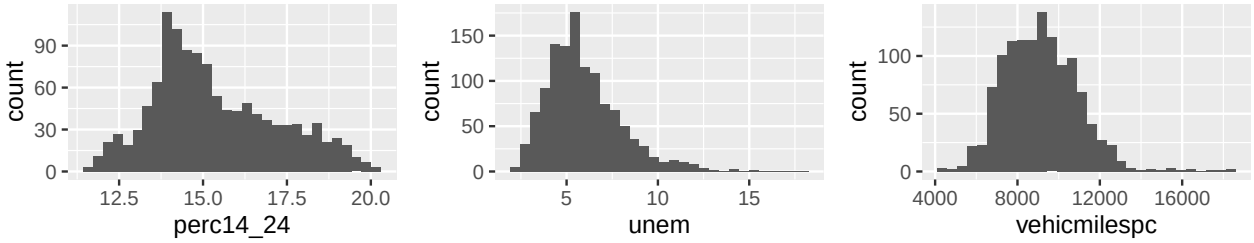
First, we can see that for sl70plus and gdl, the laws were implemented in the later years when

fatality rates have already decreased. For example, even if observations with gdl has significantly less totfatrte, this effect is clearly not causal. Instead, it could be that including gdl laws helped to keep totfatrte down. For the perse laws, most states that introduced the law did so early on this panel. However, most states did not introduce perse until after 1985, after the steepest period of totfatrte decline. As a result, perse might suffer the same issue: the law was implemented after totfatrte has already dropped significantly. In addition, for any given year, the blue lines of sl70plus were part of states with overall higher fatality rates, while gdl part of middle range to lower. This indicates that for a given year after 1995, states with speed limits of 70+ has an overall higher fatalities compared to states with speed limits < 70, during the same period. Within each year, sl70plus could have a significant role in explaining the variation in totfatrte.

We have completed our analysis of indicator variables, and will now move to numeric variables perc14_24, unem and vehicmilespc. First, we perform univariate analysis by looking at the distribution of these variables pooled across state and time.

```
phist <- ggplot(data = data, aes(perc14_24)) + geom_histogram(bins = 30)
uhist <- ggplot(data = data, aes(unem)) + geom_histogram(bins = 30)
vhist <- ggplot(data = data, aes(vehicmilespc)) + geom_histogram(bins = 30)
phist + uhist + vhist + plot_annotation(title = "Pooled histograms of numeric variables of inte
```

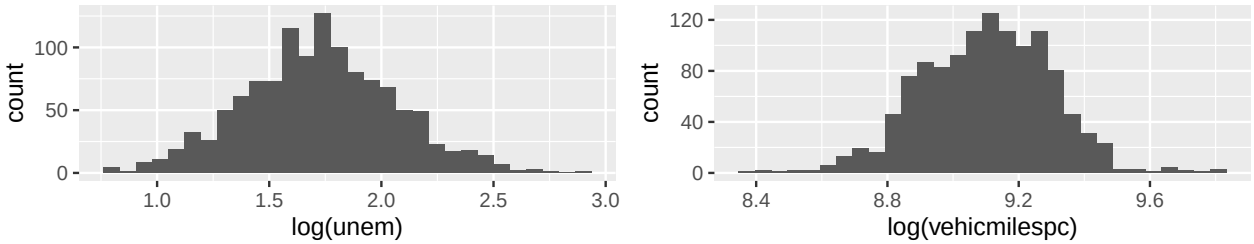
Pooled histograms of numeric variables of interest



Based on the histogram, both unem and vehicmilespc has a highly skewed distribution. We will perform a log transformation to see if we can eliminate some of the large skews. perc14_25 does not have extreme values and does not require a transformation.

```
uhist.log <- ggplot(data = data, aes(log(unem))) + geom_histogram(bins = 30)
vhist.log <- ggplot(data = data, aes(log(vehicmilespc))) + geom_histogram(bins = 30)
uhist.log + vhist.log + plot_annotation(title = "Histograms of log numeric variables of interes
```

Histograms of log numeric variables of interest



We see that taking the log fixes the skew. Since we would like the variables to be roughly normal, we can test to see whether the resulting distributions are normal using the Shapiro-Wilk Normality Test, where the null hypothesis is that the distribution is normal.

```
shapiro.test(log(data$unem))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(data$unem)
## W = 0.99834, p-value = 0.2991
```

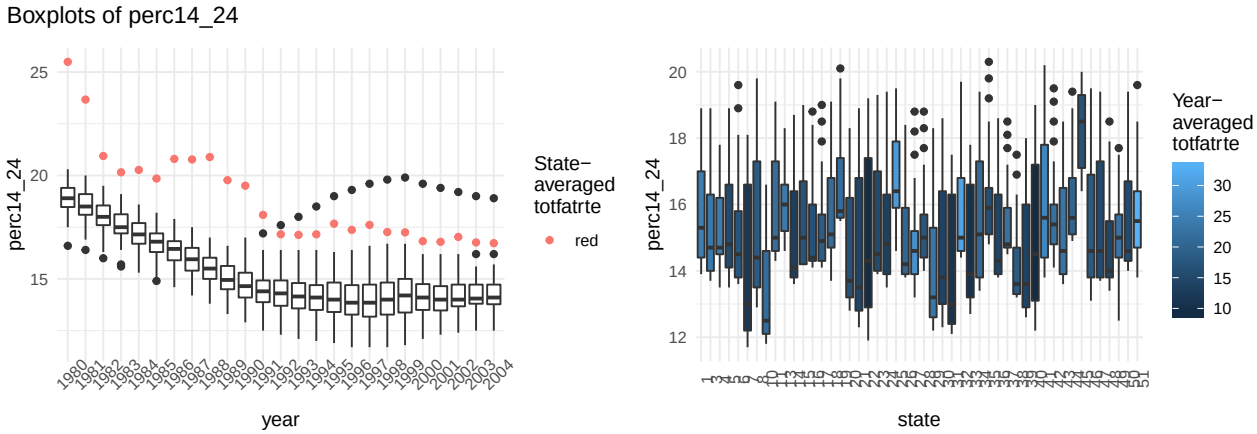
```
shapiro.test(log(data$vehicmilespc))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(data$vehicmilespc)
## W = 0.99476, p-value = 0.0003426
```

Based on the test, a transformation turns unem into a roughly normal distribution (failing to reject H0 since p-value > 0.05), but a log transformation does not transform vehicmilespc into a roughly normal distribution. We will log transform unem in our regression analysis, and use residual analysis to determine whether vehicmilespc requires a transformation.

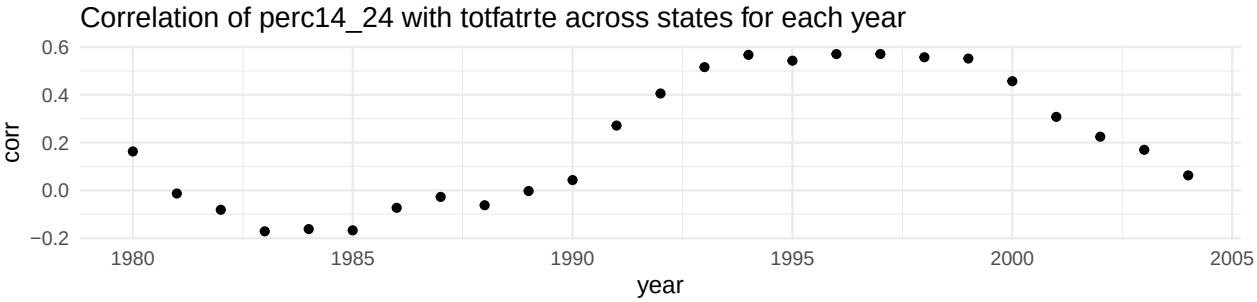
Next, we look at the percentage of population between 14 and 24 in relation to totfatrte. In the plots below, the left plot will show boxplots grouped by year, where each boxplot shows the spread of the explanatory variables across each state. Red dots will indicated average totfatrte for each year, averaged across state. The plot on the right will show boxplots grouped by state and colored by the year-averaged totfatrte for that state.

```
totfatrte_ave_state <- data %>% group_by(year) %>% summarise(totfatrte_mean_across_state = mea
totfatrte_ave_year <- data %>% group_by(state) %>% summarise(totfatrte_mean_across_year = mean
d <- totfatrte_ave_year[rep(seq_len(nrow(totfatrte_ave_year)),
    each = 25), ][, "totfatrte_mean_across_year"]
data$totfatrte_mean_across_year <- d$totfatrte_mean_across_year
yr <- ggplot(data = data, aes(x = factor(year), y = perc14_24)) +
    geom_boxplot() + geom_point(data = totfatrte_ave_state, aes(x = factor(year),
    y = totfatrte_mean_across_state, color = "red")) + theme_minimal() +
    labs(x = "year", color = "State-\naveraged\ntotfatrte") +
    theme(axis.text.x = element_text(angle = 45))
st <- ggplot(data = data, aes(x = factor(state), y = perc14_24)) +
    geom_boxplot(aes(fill = totfatrte_mean_across_year)) + theme_minimal() +
    labs(x = "state", fill = "Year-\naveraged\ntotfatrte") +
    theme(axis.text.x = element_text(angle = 90))
yr + st + plot_annotation(title = "Boxplots of perc14_24")
```



On the left plot, we see that the percentage of drivers between ages 14-24 is generally decreasing between 1980 and 2005, although some large single outliers crop up for years 1990-2004 (black points). Compared to the red dots which represent state averaged totfatrte for that year, the two variables appear correlated. Since young drivers are inexperienced, this correlation would make sense. On the right plot, we see that some states have much higher variability in perc14 than others. The year-averaged totfatrte is also highly variable relative to the median of perc14_24. For example, for state 1, the year-averaged totfatrte is around 25, and the median for perc14_24 is around 16. On the contrary, for state 22, the average totfatrte is below 10 for a similar median. This variable is clearly not enough to fully explain the variation within a state. To further explore the explanatory power of this variable, we can look at correlation of perc14_24 across the states for each year.

```
cor_data <- data %>% group_by(year) %>% summarise(corr = cor(perc14_24,
    totfatrte))
ggplot(data = cor_data, aes(x = year, y = corr)) + geom_point() +
    theme_minimal() + ggtitle("Correlation of perc14_24 with totfatrte across states for each y
```

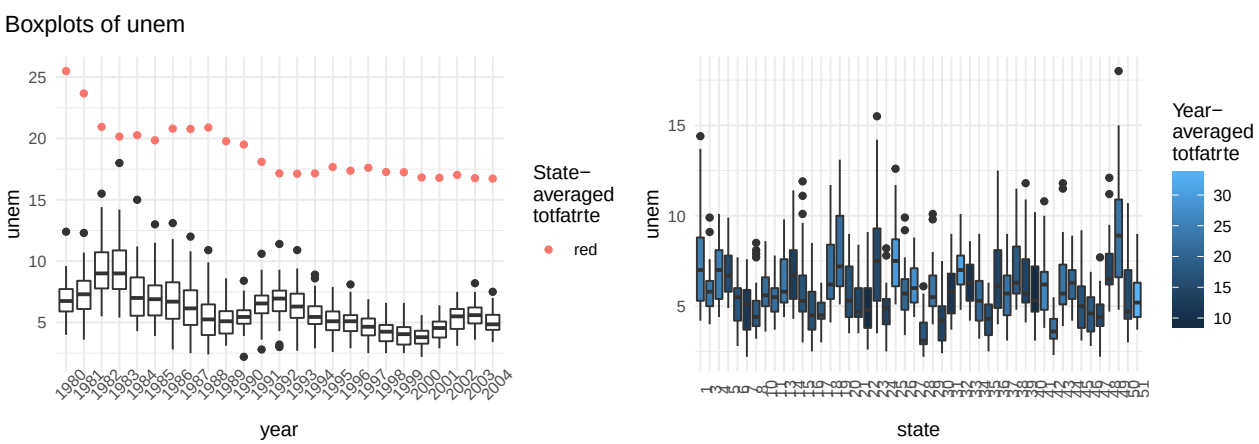


Interestingly, the correlation is low, even negative, between 1990, high between 1990 and 2000, but drops again after 2000. As a result, this variable may be useful, especially for explaining variation between 1990 and 2000.

Next, we will look at the unemployment rate.

```
yr <- ggplot(data = data, aes(x = factor(year), y = unem)) +
    geom_boxplot() + geom_point(data = totfatrte_ave_state, aes(x = factor(year),
    y = totfatrte_mean_across_state, color = "red")) + theme_minimal() +
    labs(x = "year", color = "State-\naveraged\ntotfatrte") +
    theme(axis.text.x = element_text(angle = 45))
st <- ggplot(data = data, aes(x = factor(state), y = unem)) +
```

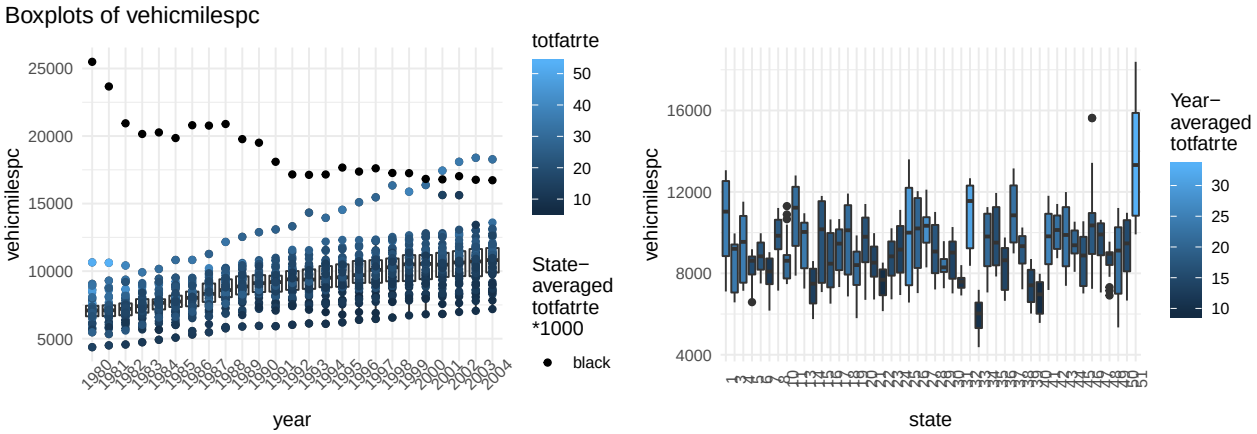
```
    geom_boxplot(aes(fill = totfatrte_mean_across_year)) + theme_minimal() +
    labs(x = "state", fill = "Year-\naveraged\ntotfatrte") +
    theme(axis.text.x = element_text(angle = 90))
yr + st + plot_annotation(title = "Boxplots of unem")
```



From the left plot, unemployment clearly fluctuates drastically, and there appears to be a weak inverse correlation between unemployment and state-averaged totfatrte. For example, as unem decreases from 1983 to 1987, totfatrte increased, and unem rose from 1988 to 1993, totfatrte decreases. This makes sense because the lower the unem rate, the more people who drive, the greater the fatality rate per unit population. There appears to be little relation from the plot on the right on whether the yearly-averaged totfatrte has anything to do with unem within a state, similar to perc14_24.

Finally, vehicmilespc is the vehicle miles per capita, a direct measure of miles driven. We would expect the higher the miles driven, the greater the fatality rate, both within states and within years. We will look at boxplot of vehicmilespc, and 1000 * totfatrte for visualization.

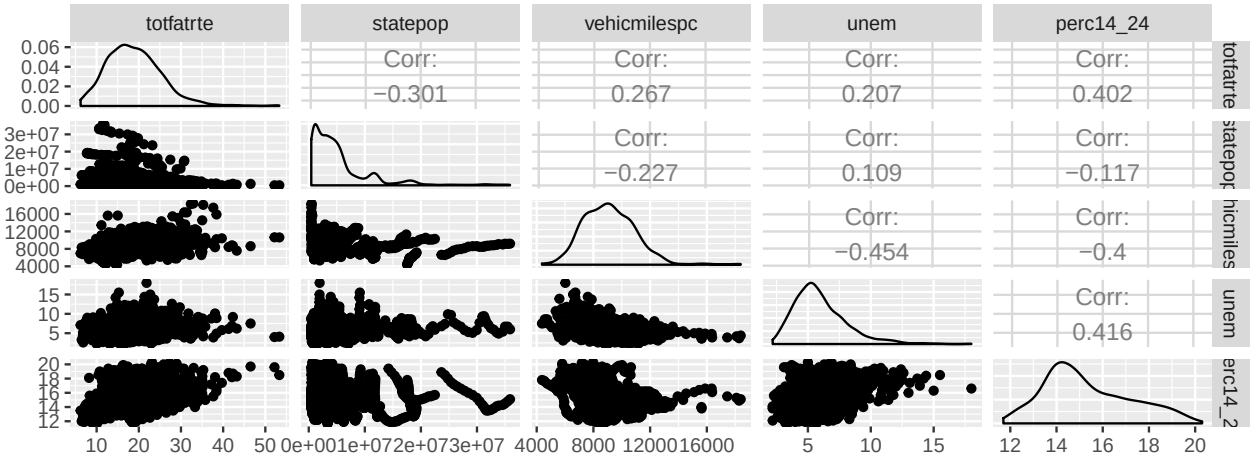
```
yr <- ggplot(data = data, aes(x = factor(year), y = vehicmilespc)) +
    geom_boxplot() + geom_point(data = data, aes(x = factor(year),
    y = vehicmilespc, color = totfatrte)) + geom_point(data = totfatrte_ave_state,
    aes(x = factor(year), y = totfatrte_mean_across_state * 1000,
        fill = "black")) + theme_minimal() + labs(x = "year",
    fill = "State-\naveraged\ntotfatrte\n*1000") + theme(axis.text.x = element_text(angle = 45)
st <- ggplot(data = data, aes(x = factor(state), y = vehicmilespc)) +
    geom_boxplot(aes(fill = totfatrte_mean_across_year)) + theme_minimal() +
    labs(x = "state", fill = "Year-\naveraged\ntotfatrte") +
    theme(axis.text.x = element_text(angle = 90))
yr + st + plot_annotation(title = "Boxplots of vehicmilespc")
```



Interestingly, the amount driven is increasing over the years, but the fatality rate is decreasing. This is likely not a direct causal relationship. Most likely, what is happening is that as the population becomes urbanized, more individuals can afford cars, and more mobility is required for work or day-to-day activities. A dotplot of vehicmilespc has been overlaid on top of the left boxplots, where each point represents a different state and is colored based on totfatrte for that state within the indicated year (x-axis). This shows that while the overall trend across time between the two variables are inversely related, within each year, higher levels of vehicmilespc (top of the dot plot) tends to be correlated with higher levels of totfatrte (light blue dots). In fact, all outliers are near the light blue end of the spectrum (high totfatrte). On the right plot, states with higher distributions of vehicmilespc tends to have higher year-averaged fatality rates (such as state 25, 32, 51, colored light blue). Opposite is true for low levels of vehicmilespc, such as state 33 and state 40. As a result, the vehicmilespc positively correlates with totfatrte within and averaged across each year.

We will conclude analysis of the numerical variables with a correlation plot analysis of all values pooled across time with totfatrte. While we lose panel information from doing this, this will give us a sense of how much information is contained in each numerical variable when pooling.

```
data.numeric <- data %>% dplyr::select(totfatrte, statepop, vehicmilespc,
    unem, perc14_24)
ggpairs(data.numeric)
```



We see that population is inversely related to totfatrte. This could indicate for example that states with larger populations have stricter laws, among many other things. vehicmilespc, unem,

perc14_24 all have a positive correlation with totfatrte. The direction of correlation for perc14_24 is consistent with our panel analysis, which tends to be mostly positive for each year, but the sign of the correlation is different for unem and vehicmilespc. For example, we previously observed that vehicmilespc increased with time, which makes sense, and totfatrte decreases with time (the time dimension on the 2D scatterplot for these two variables would start on the bottom right, and go toward the upper right; see Appendix). Year information appears critical for these variables.

Regression model for the mean

The dependent variable totfatrte is the total fatalities per 100,000 population. Average of this variable in each of the years in the time period covered in this dataset:

```
m <- data %>% group_by(year) %>% summarize(mean = round(mean(totfatrte),
    3))
cbind(m[1:5, ], m[6:10, ], m[11:15, ], m[16:20, ], m[21:25, ])
```

```
##   year   mean year   mean year   mean year   mean year   mean year   mean
## 1 1980 25.495 1985 19.851 1990 19.505 1995 17.669 2000 16.826
## 2 1981 23.670 1986 20.800 1991 18.095 1996 17.369 2001 16.793
## 3 1982 20.942 1987 20.775 1992 17.158 1997 17.611 2002 17.030
## 4 1983 20.153 1988 20.892 1993 17.128 1998 17.265 2003 16.764
## 5 1984 20.267 1989 19.772 1994 17.155 1999 17.250 2004 16.729
```

Linear regression with only the years generates a model that explains the change in average totfatrte in a given year compared to the base year 1980:

```
mod.means <- lm(totfatrte ~ factor(year), data = data)
coeftest(mod.means, vcovHC)
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       25.4946     1.1695 21.7988 < 2.2e-16 ***
## factor(year)1981  -1.8244     1.6322 -1.1178 0.2639003
## factor(year)1982  -4.5521     1.5259 -2.9831 0.0029121 **
## factor(year)1983  -5.3417     1.4705 -3.6326 0.0002927 ***
## factor(year)1984  -5.2271     1.4217 -3.6766 0.0002470 ***
## factor(year)1985  -5.6431     1.4136 -3.9921 6.955e-05 ***
## factor(year)1986  -4.6942     1.4335 -3.2745 0.0010892 **
## factor(year)1987  -4.7198     1.4383 -3.2815 0.0010630 **
## factor(year)1988  -4.6029     1.4047 -3.2768 0.0010808 **
## factor(year)1989  -5.7223     1.4094 -4.0601 5.231e-05 ***
## factor(year)1990  -5.9894     1.4280 -4.1942 2.944e-05 ***
## factor(year)1991  -7.3998     1.3989 -5.2895 1.462e-07 ***
## factor(year)1992  -8.3367     1.4008 -5.9512 3.508e-09 ***
## factor(year)1993  -8.3669     1.3922 -6.0097 2.476e-09 ***
## factor(year)1994  -8.3394     1.4161 -5.8890 5.066e-09 ***
## factor(year)1995  -7.8260     1.4687 -5.3287 1.185e-07 ***
## factor(year)1996  -8.1252     1.4373 -5.6531 1.977e-08 ***
```

```
## factor(year)1997  -7.8840     1.4509 -5.4338 6.703e-08 ***
## factor(year)1998  -8.2292     1.4588 -5.6412 2.114e-08 ***
## factor(year)1999  -8.2442     1.4969 -5.5076 4.464e-08 ***
## factor(year)2000  -8.6690     1.4626 -5.9271 4.047e-09 ***
## factor(year)2001  -8.7019     1.4570 -5.9724 3.093e-09 ***
## factor(year)2002  -8.4650     1.4837 -5.7054 1.468e-08 ***
## factor(year)2003  -8.7310     1.4598 -5.9810 2.939e-09 ***
## factor(year)2004  -8.7656     1.4854 -5.9010 4.721e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model we fit is:

$$totfatrte = 25.4946 - 1.8244 * x_{1981} - 4.5521 * x_{1982} + ... - 8.7656 * x_{2004}$$

Since 1980 was the base year, the intercept of this model is the average value of totfatrte in 1980 (25.495). The coefficients of the other years represent the change in average totfatrte in that year compared to 1980. For example, in 1981, the change is -1.824, so the average totfatrte in 1981 is 25.495-1.824 = 23.671, equivalent to the dataframe of year averages from previous. This is because we are allowing each year to have a separate intercept. Linear regression coefficients are estimated to produce the expected value of y. Due to zero conditional mean assumption $E(u|year) = 0$, the expected value of y given a vector of x value is 0, so given that we want to estimate y for a particular year x after 1980, the equation boils down to:

$$E(\text{totfatrte} \mid \text{year}_{1980+x}) = \beta_0 + \beta_x \text{year}_{1980+x} \qquad \text{for x} > 0$$

For the year 1980, all "year" variables are 0, so the expected value of totfatrte in that year is β_0, whereas for any other year, the expected value is $\beta_0 + \beta_x$, making the other coefficients the change in expected value of totfatrte with respect to year 1980.

All coefficients are negative, meaning totfatrte is less compared to 1980 for any following year in the dataset. Furthermore, all coefficients except 1981 is statistically significant, meaning the change of totfatrte in any year except 1981 with respect to 1980 is statistically negative. In addition, the absolute value of the coeffients is generally becoming larger. This means that the magnitude of decline in totfatrte compared to 1980 is generally increasing, though there are fluctuations, such as the increase between 1985 and 1986. Based on this decline of coefficients, it is getting safer to drive. However, the rate of decrease is also slowing down, which is supported by the EDA time series plot flattening out. So while driving has gotten safer during this period between 1980 in 2004, further models are required to know whether the decline is significant for example between 2002 and 2004.

Expandeded pooled OLS with added variables bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc

Transformations As indicated by the EDA, 'bac08', 'bac10', 'perse', 'sl70plus', 'gdl' are all variables that are currently represented as decimals, indicating that a particular law went into/out of effect part way through the year. The question remains how we treat a particular year for which the transition happened. The most logical is if the law was in effect for less than half the time (value is less than 0.5), then we treat the year as not having the law. If the value is equal to or more than 0.5, then we treat the year as having the law. This rounding ensures that each year is indicated as having the law or not having the law, allowing us to treat the variable as a factor.

Note that for bac08 and bac10, we've created a single variable with 3 levels, no bac, bac08, and bac10 as done in the EDA. 'sbprim' and 'sbsecon' are currently indicator variables that are either 0 or 1, so these require no transformation.

We saw that the unem variable is highly skewed, and by taking the log transformation, the variable passed the normality test. As a result, we will take the log transform of this variable. This makes it less likely that single extreme points have too much leverage or influence on the regression model. perc14_24 did not have a skewed distribution and as a result we will not transform this variable. We show in the appendix that the residual plot against this variable shows no correlation, and zero mean. A log transformation helped remove extreme right skew outliers from vehicmilespc, but the distribution was not normal. A residual analysis in the appendix shows that this transformation introduced correlation of the residuals with the fitted values. As a result, we will not log transform vehicmilespc.

In the appendix, we show that log transforming the outcome variable totfatrte leads to correlation of the residuals with the fitted values. As a result, we will not transform this variable for our model.

```
# Transform the indicator variables
data.fit <- data %>% dplyr::select(totfatrte, year, bac_none_08_10)
data.fit$year <- factor(data.fit$year)
to_transform <- c("perse", "sl70plus", "gdl")
other_x <- c("sbprim", "sbsecon", "perc14_24", "unem", "vehicmilespc")
# Generate full data frame
data.fit <- cbind(data.fit, data %>% dplyr::select(other_x))
data.fit <- cbind(data.fit, data %>% dplyr::select(to_transform) %>%
    round())
# Generate the model with 1980 as base year
model.3 <- lm(totfatrte ~ factor(year) + sbprim + sbsecon + perc14_24 +
    log(unem) + vehicmilespc + bac_none_08_10 + perse + sl70plus +
    gdl, data = data.fit)
```

Based on the residual versus fitted values plot in the appendix, we satisfy zero conditional mean, but we have heteroskedasticity. Therefore, we use heteroskedastic-robust standard errors in order to test the coefficients.

```
coeftest(model.3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                         Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)         -8.0797e+00  2.8677e+00  -2.8175  0.004922 **
## factor(year)1981    -2.1029e+00  1.3225e+00  -1.5902  0.112069
## factor(year)1982    -6.2337e+00  1.2209e+00  -5.1058 3.844e-07 ***
## factor(year)1983    -6.9311e+00  1.1332e+00  -6.1163 1.306e-09 ***
## factor(year)1984    -5.7540e+00  1.1135e+00  -5.1677 2.787e-07 ***
## factor(year)1985    -6.3857e+00  1.1337e+00  -5.6326 2.224e-08 ***
## factor(year)1986    -5.6177e+00  1.2089e+00  -4.6471 3.750e-06 ***
## factor(year)1987    -6.0540e+00  1.2346e+00  -4.9036 1.074e-06 ***
## factor(year)1988    -6.1705e+00  1.2580e+00  -4.9051 1.066e-06 ***
```

```
## factor(year)1989    -7.6808e+00  1.3268e+00  -5.7888 9.108e-09 ***
## factor(year)1990    -8.6695e+00  1.3679e+00  -6.3376 3.328e-10 ***
## factor(year)1991    -1.0822e+01  1.3698e+00  -7.8999 6.418e-15 ***
## factor(year)1992    -1.2605e+01  1.4056e+00  -8.9678 < 2.2e-16 ***
## factor(year)1993    -1.2458e+01  1.4019e+00  -8.8861 < 2.2e-16 ***
## factor(year)1994    -1.2025e+01  1.4099e+00  -8.5292 < 2.2e-16 ***
## factor(year)1995    -1.1452e+01  1.4619e+00  -7.8337 1.061e-14 ***
## factor(year)1996    -1.3368e+01  1.4463e+00  -9.2433 < 2.2e-16 ***
## factor(year)1997    -1.3434e+01  1.4911e+00  -9.0096 < 2.2e-16 ***
## factor(year)1998    -1.4163e+01  1.5025e+00  -9.4264 < 2.2e-16 ***
## factor(year)1999    -1.4119e+01  1.5116e+00  -9.3407 < 2.2e-16 ***
## factor(year)2000    -1.4370e+01  1.5572e+00  -9.2277 < 2.2e-16 ***
## factor(year)2001    -1.5544e+01  1.5437e+00 -10.0689 < 2.2e-16 ***
## factor(year)2002    -1.6310e+01  1.5771e+00 -10.3421 < 2.2e-16 ***
## factor(year)2003    -1.6709e+01  1.5858e+00 -10.5368 < 2.2e-16 ***
## factor(year)2004    -1.6170e+01  1.6237e+00  -9.9587 < 2.2e-16 ***
## sbprim              -3.5695e-01  4.7008e-01  -0.7593  0.447810
## sbsecon             -1.5181e-01  4.2192e-01  -0.3598  0.719048
## perc14_24            1.8294e-01  1.2374e-01   1.4784  0.139563
## log(unem)            5.1082e+00  5.0175e-01  10.1809 < 2.2e-16 ***
## vehicmilespc         2.9284e-03  1.2367e-04  23.6803 < 2.2e-16 ***
## bac_none_08_10bac08 -2.3937e+00  4.9188e-01  -4.8663 1.293e-06 ***
## bac_none_08_10bac10 -1.2707e+00  4.0942e-01  -3.1037  0.001957 **
## perse               -5.6645e-01  2.7334e-01  -2.0723  0.038456 *
## sl70plus             3.0843e+00  3.9509e-01   7.8065 1.303e-14 ***
## gdl                 -3.1855e-01  4.6464e-01  -0.6856  0.493104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

bac8 and bac10 are blood alcohol content levels. bac8 means if blood alcohol level (concentration of alcohol in the blood) is at or above 0.08 percent, a person is considered intoxicated and it is illegal for them to drive. If such a person is found driving, they are violating a driving under the influence law. bac10 means if blood alcohol content at or above 0.1%, the individual is intoxicated and it is illegal for them to drive.

The base level for our variable is no bac laws in effect. For bac08, the coefficient -2.3937. Keeping all other conditions constant, having a bac08 law compared to no law at all decreases totfatrte by an estimated amount of 2.394. This is highly statistically significant (p-value $<<$ 0.01). Keeping all other variables constant, introducing a bac10 law compared to no law at all decreases totfatrte by an estimated amount of 1.271. This is also highly statistically significant (p-value $<$ 0.001).

Since the sign on both perse and sbprim are negative, the effect is that keeping all other variables constant, fatality rate is estimated to decrease by 0.566 when perse laws are in effect compared to no perse law, and by 0.356 with sbprim in effect compared to not having sbprim. At a level of 0.05, only perse variable is statistically significant with a p-value of 0.038 $<$ 0.05. Primary seatbelt was not significant with a p-value of 0.448. As a result, with 95% confidence, we can conclude that perse laws have a statistically signficant negative effect on fatality rate, but that sbprim does not.

Fixed effects regression model

Fixed effects regression model allows us to control for time-invarying factors across each of the states. We will assess the assumptions set out by this model and pooled OLS. First, we fit the model. Note that we will not use a "twoway" model since we are interested in eliminating fixed a_i effects at the state level, and not necessarily effects specific to each time period (typically denoted θ_i). Later, we note that the residuals display heteroskedasticity, so we will use robust standard errors to look at coefficient significance.

```
data.plm.fit <- cbind(data.fit, data %>% dplyr::select(state,
    year, totfatrte))
model.fe <- plm(totfatrte ~ sbprim + sbsecon + perc14_24 + log(unem) +
    vehicmilespc + bac_none_08_10 + perse + sl70plus + gdl, data = data.plm.fit,
    index = c("state", "year"), model = "within")
coeftest(model.fe, vcovHC)
```

```
##
## t test of coefficients:
##
##                          Estimate  Std. Error t value  Pr(>|t|)
## sbprim               -1.77489457  0.74545208 -2.3810  0.017431 *
## sbsecon              -0.80760577  0.47855444 -1.6876  0.091762 .
## perc14_24             0.96207189  0.16992427  5.6618 1.894e-08 ***
## log(unem)            -3.28537828  0.52027302 -6.3147 3.868e-10 ***
## vehicmilespc          0.00033785  0.00027226  1.2409  0.214892
## bac_none_08_10bac08 -1.80729814  0.61849783 -2.9221  0.003546 **
## bac_none_08_10bac10 -1.31773351  0.47357982 -2.7825  0.005483 **
## perse                -1.51311090  0.38873433 -3.8924  0.000105 ***
## sl70plus             -1.17466017  0.51743338 -2.2702  0.023383 *
## gdl                  -0.53266737  0.31020274 -1.7172  0.086221 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimate for bac08 is now slightly smaller in magnitude, meaning slightly smaller effect (-1.81 vs -2.39), and the SE for the bac08 level has slightly increased. However, the variable is still highly statistically significant (p-value < 0.01). bac10 has a slightly larger in magnitude and approximately the same SE, maintaining high statistical significant. The variable perse has a significantly larger absolute value (almost 3x) and only a slightly higher SE, making this variable now highly statistically significant, compared to only marginally significant previously. For sbprim, the estimate is now about 5x larger in magnitude, and the SE has has only slightly increased. The variable is now statistically significant, compared to not being so in the pooled OLS model. FE model has the greatest effect on perse and sbprim out of these 4 variables, making the former highly statistically significant, and the latter statistically significant.

Reliability: The fixed effects model is likely more reliable because pooled OLS pools the data across the states, and is biased if time-invarying unobserved effects a_i is correlated with any of the explanatory variables. For example, factors such as general political leaning could be correlated with perc14_24 (younger populations are typically more liberal). FE eliminates these time-invarying effects, whereas pooled OLS will suffer from heterogeneity bias leading to biased estimates. The pFtest formalized this, and the null hypothesis is that pooled OLS is sufficient and FE is not

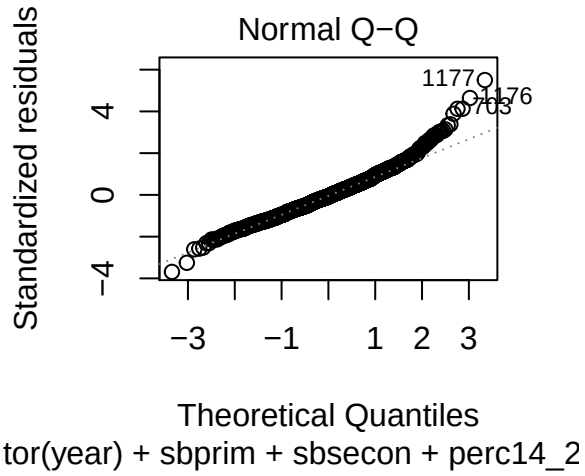
necessary:

```
pFtest(model.fe, model.3)
```

```
##
##  F test for individual effects
##
## data:  totfatrte ~ sbprim + sbsecon + perc14_24 + log(unem) + vehicmilespc +  ...
## F = 115.39, df1 = 23, df2 = 1142, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Since p-value << 0.01, we reject H0 and conclude that FE is a better model. Several OLS assumptions are also violated below. **CLM1: Linear in parameters** We define the model with an error term so the parameters are linear. **CLM2: Random Sampling** This assumes that all observations are independent, which is strongly violated by this model. Observations of the same state across multiple years is clearly not independent, and neither are different states within the same year. **violated**. **CLM 3: No perfect multi-collinearity** R would have warned us if this were the case (by evaluating whether the covariance matrix is singular), so we have fulfilled this requirement. **CLM4: Zero Conditional Mean** Based on the residual analysis of model 3 in the appendix, the residuals all average around 0 for all fitted values, so this is satisfied. **CLM 5: Homoskedasticity** Based on the residual analysis of model 3 in the appendix, there is clear homoskedasticity for larger fitted values, which is why we used robust errors. **CLM 6: Normality** Population error is independent of the explanatory variables. We check this with a QQ plot:

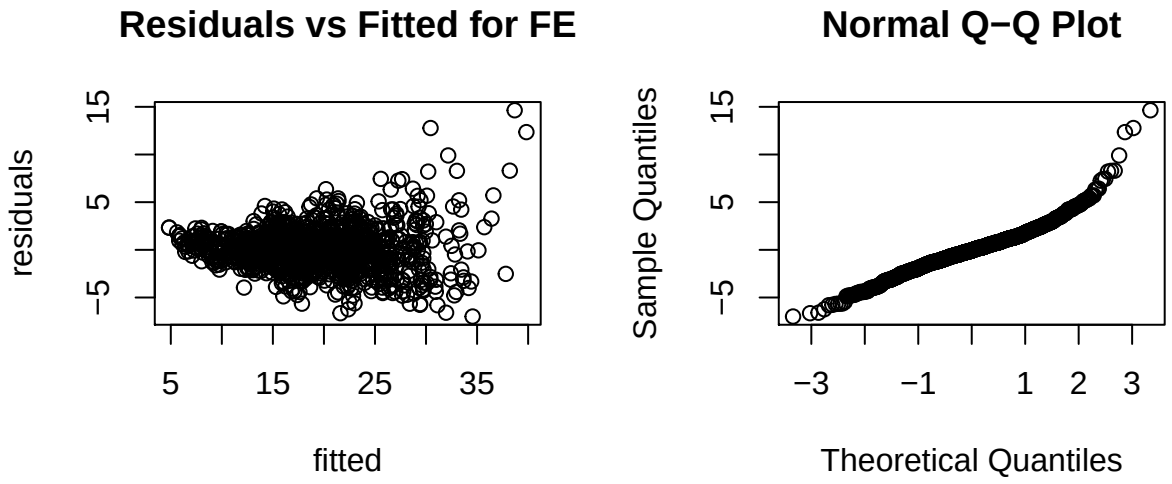
```
plot(model.3, which = 2)
```



We see that for most values, the quartiles fall on the theoretical line, except for extreme values on the upper end. This could indicate a skew. Since we do not have independent observations, CLT does not apply, so this assumption is likely also **violated**. The fixed effects assumptions and whether they are reasonable is listed below: **Assumption FE.1** For each i, the model is $y_{it} = \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it}, t = 1, , T$ where the $beta_j$ are the parameters to estimate and a_i is the time-invariant unobserved effect. This is the assumption and nothing to assess here. **Assumption FE.2** We have a random sample from the cross section. Since we have a consensus of all of the states for each cross section, this is completely reasonable. **Assumption FE.3** Each explanatory variable changes over time and no perfect linear relationships exist among the explanatory variables. This is true that there is some unique variation in all of the explanatory

variables. R would have warned us otherwise if this isn't the case. **Assumption FE.4** For each t, the expected value of the idiosyncratic error given the explanatory variables in all time periods and the unobserved effect is zero. The residual versus fitted values for all samples is displayed below, along with a loess smoother to the plot. The residuals here would represent the idiosyncratic error since FE eliminates unobserved effects.

```
fitted <- data$totfatrte - model.fe$residuals %>% as.vector()
residuals <- model.fe$residuals %>% as.vector()
par(mfrow = c(1, 2))
plot(fitted, residuals, main = "Residuals vs Fitted for FE")
qqnorm(model.fe$residuals)
```



As can be seen from the smoother plot, the average residual for all values of fitted is approximately 0. The slight deviance toward the positive end is likely due to small amounts of data with extremely large fitted totfatrte. Given FE1-FE4, we have unbiased estimators. **Assumption FE.5** $Var(u_{it}|X_i, a_i) = Var(u_{it}) = \sigma_u^2 \quad$ for all $\quad t = 1, , T$. Based on the residual plot above, as fitted values increase, so does the variance of the residuals, meaning we likely do not have homoskedasticity. This assumption is therefore not valid. We resolved this issue by using heteroskedastic robust standard errors. **Assumption FE.6** For all $t \neq s$, the idiosyncratic errors are uncorrelated (conditional on all explanatory variables and a_i). In other words, the errors for each state at different time periods are uncorrelated given the model fit. We test this with a BG test:

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

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  totfatrte ~ sbprim + sbsecon + perc14_24 + log(unem) + vehicmilespc +     bac_none_08
## chisq = 402.61, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

Based on the BG test, we do have serial correlation in the idiosyncratic errors, so this assumption is likely violated. Under Assumptions FE.1 through FE.6, the fixed effects estimator is the best linear unbiased estimator (BLUE). **Assumption FE.7** Conditional on X_i and a_i, the u_{it} are independent and identically distributed as Normal(0,σ_u^2). Based on the qqplot above, the sample quantiles and

theoretical quantiles do not match up too well. We have a large cross section (N == 50) , and reasonably large number of year (25 panels) so CLT may or may not kick in. Since this is the case, the standard errors may not be fully accurate (see final question).

Comparison to a random effects model

In this case, since we do not have any explicit explanatory variables that are constant in time in our model, there is no advantage to using the random effect model, whose primary advantage is that variables that are constant in time can be included in the estimation. FE is usually considered a more convincing tool for estimating the true effects of time-varying variables for panel data regression. When using random effects, usually one wants to use as many time-constant variables as possible in order to control time-constant effects. While we have some of these variables in the dataset, we have many reasons to believe that other time-invarying unobserved effects not in the dataset, call them a_i, are correlated with the existing explanatory variables. We therefore cannot use the random effects model and prefer fixed effects as this would violate the random effects model assumption:

$$Cov(x_{itj}, a_i) = 0; t = 1, 2, , T; j = 1, 2, , k$$

For example, one a_i effect that might be fixed in time but different for each state is the number of interstate highway that goes through a state. We would expect this variable to be correlated with vehicmilespc, as individuals living in highways-dense states should drive more. Another example is general political leaning of the state. This variable might also be correlated with perc14_24, as generally the larger the youth population the more blue the state. In these cases, since we expect these a_i variable that are constant in time to be correlated with our explanatory variables, using an RE model would violate the assumption of a random effects model. Finally, we generate RE model, and then perform the Hausman test. The null hypothesis is that the preferred model is random effects vs. the alternative the fixed effects. It tests whether the unique errors u_i are correlated with the explanatory variables; the null hypothesis is they are not.

```
model.re <- plm(totfatrte ~ sbprim + sbsecon + perc14_24 + log(unem) +
    vehicmilespc + bac_none_08_10 + perse + sl70plus + gdl, data = data.plm.fit,
    index = c("state", "year"), model = "random")
phtest(model.fe, model.re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ sbprim + sbsecon + perc14_24 + log(unem) + vehicmilespc +  ...
## chisq = 1006.6, df = 10, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

We reject H0 that the preferred model is RE and go with FE, which is consistent with our analysis above.

FE model interpretation

Keeping all variables constant including the year, a one-unit increase in vehicmilespc results in an estimated 0.00033785 increase in totfatrte, so a 1000 increase in miles driven per capita results in an estimated 0.33785 increase in total fatalities per 100000 population. The standard error for this estimate is large and the coefficient is not statistically significant. The 95% CI is below:

```
se <- sqrt(diag(vcovHC(model.fe))[["vehicmilespc"]])
ci <- c(lower = model.fe$coefficients[["vehicmilespc"]] + qnorm(0.05/2) *
    se, upper = model.fe$coefficients[["vehicmilespc"]] - qnorm(0.05/2) *
    se)
ci
```

```
##         lower         upper
## -0.0001957691  0.0008714736
```

With 95% confidence, a 1000 increase in miles driven per capita results in a -0.1957691 to 0.8714736 change in fatalities per 100000 population, keeping all other variables constant. We can convert this to an absolute number for any state in any year since we are given the state population in the data. For example, row 500, in year 2004, state 23, given all other variables are constant, an increase in 1000 miles driven per capita would have resulted in the following 95%CI (-19.8 to 88.3) and estimated (34.2) change in estimated fatalities:

```
c(ci = data[500, ]$statepop * ci[1] * 1000/1e+05, estimate = data[500,
    ]$statepop * 0.00033785 * 1000/1e+05, ci = data[500, ]$statepop *
    ci[2] * 1000/1e+05)
```

```
##  ci.lower  estimate  ci.upper
## -19.83729  34.23434  88.30642
```

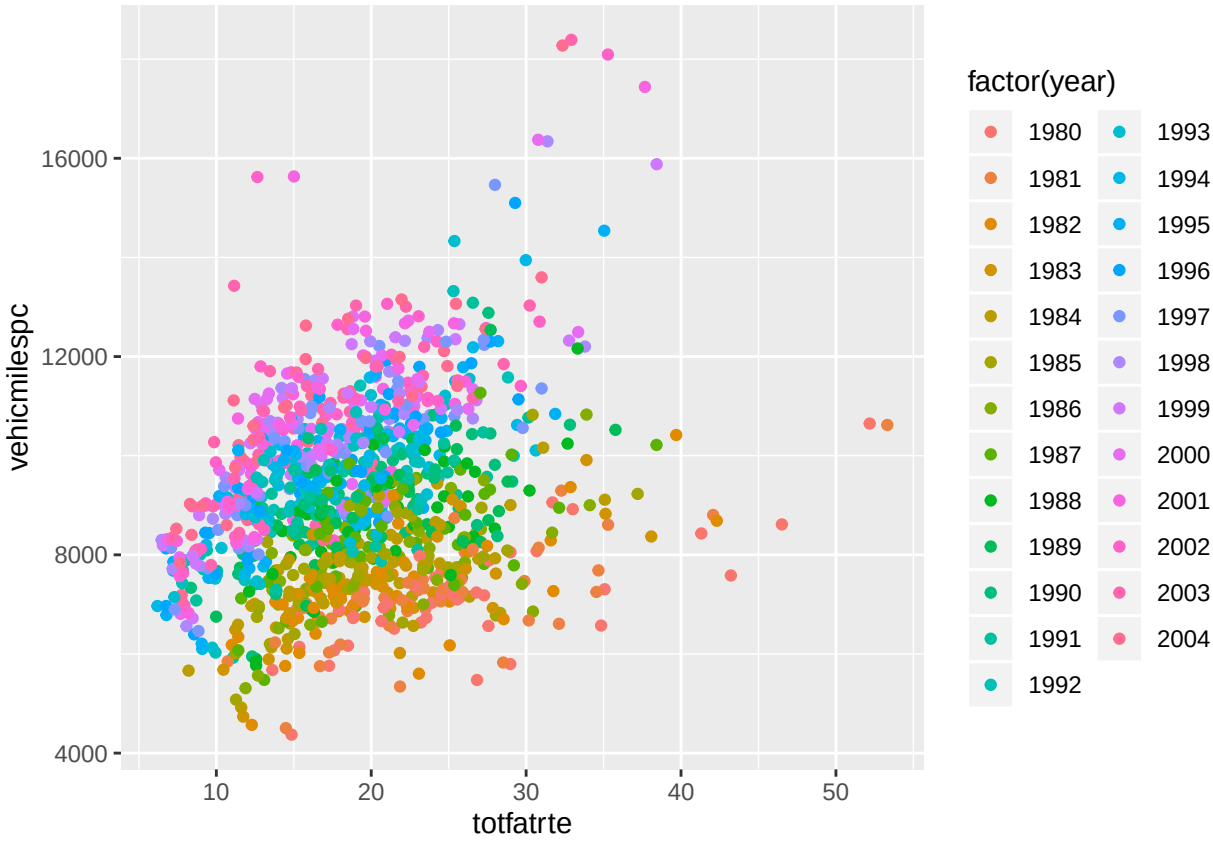
Discussion of serial correlation and heteroskedasticity

The standard errors are only accurate if there is no serial correlation and the errors are homoskedastic based on FE5 and FE6 described above. There is no consequence on the estimated coefficients values. The estimators are unbiased if they fulfill Assumptions FE1-4, regardless of serial correlation or heteroskedasticity. If there's serial correlation, the standard errors will be too small, because serial correlation means that consecutive values do not vary as much as if all points are independent. Therefore the SE estimated from the data will be smaller than an SE estimated from random data, leading to potential Type I error. If we have heteroskedasticity in the errors, then the estimated standard errors can be either too small or too large. In this case, since our variance increases as we get to larger values, we would most likely have larger SE than if we didn't have heteroskedasticity. This is because the larger spread for larger fitted values potentially increases the estimated SE compared to constant homoskedasticity, which would have overall lower variance. This leads to a greater chance of Type II error. Heteroskedasticity was resolved by using robust standard errors; however, serial correlation was still an issue in our model.

Appendix

totfatrte vs vehicmilespc with year dimension added in

```
ggplot(data = data, aes(x = totfatrte, y = vehicmilespc, color = factor(year))) +
    geom_point()
```

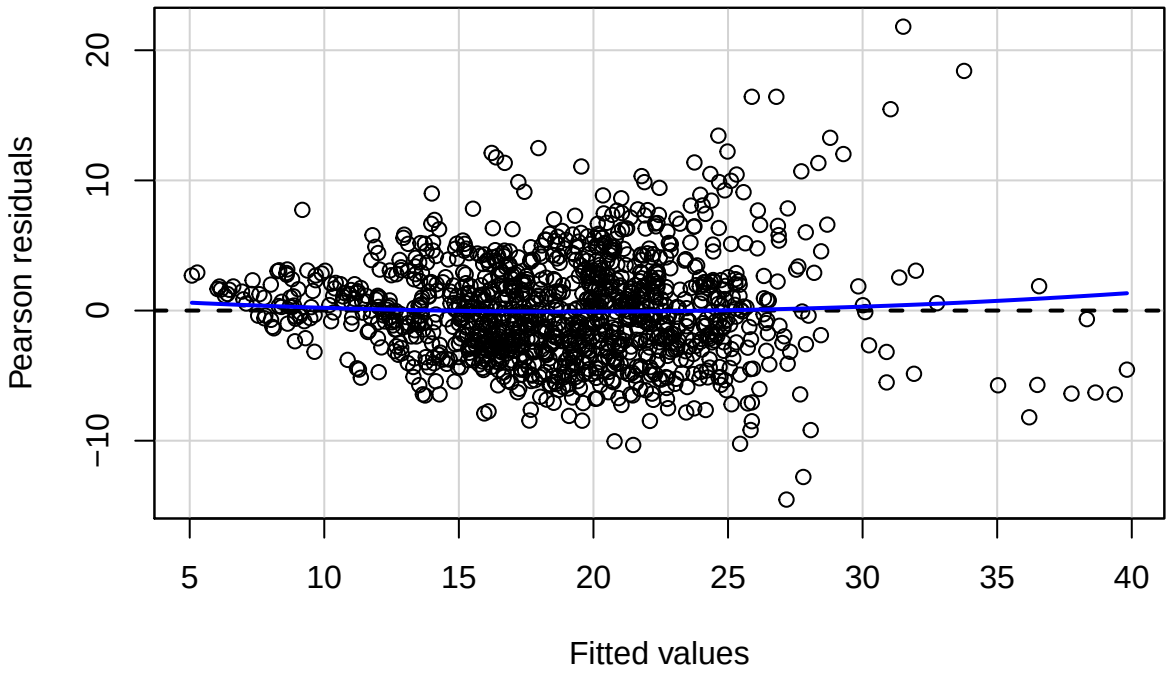


We can clearly see that with a time dimension, totfatrte decreases with time while vehicmilespc increases, even though the pooled correlation of the two variables are positive.

Residual plots for model.3, and identical model with log transformed totfatrte

Below is the residual plot for model.3

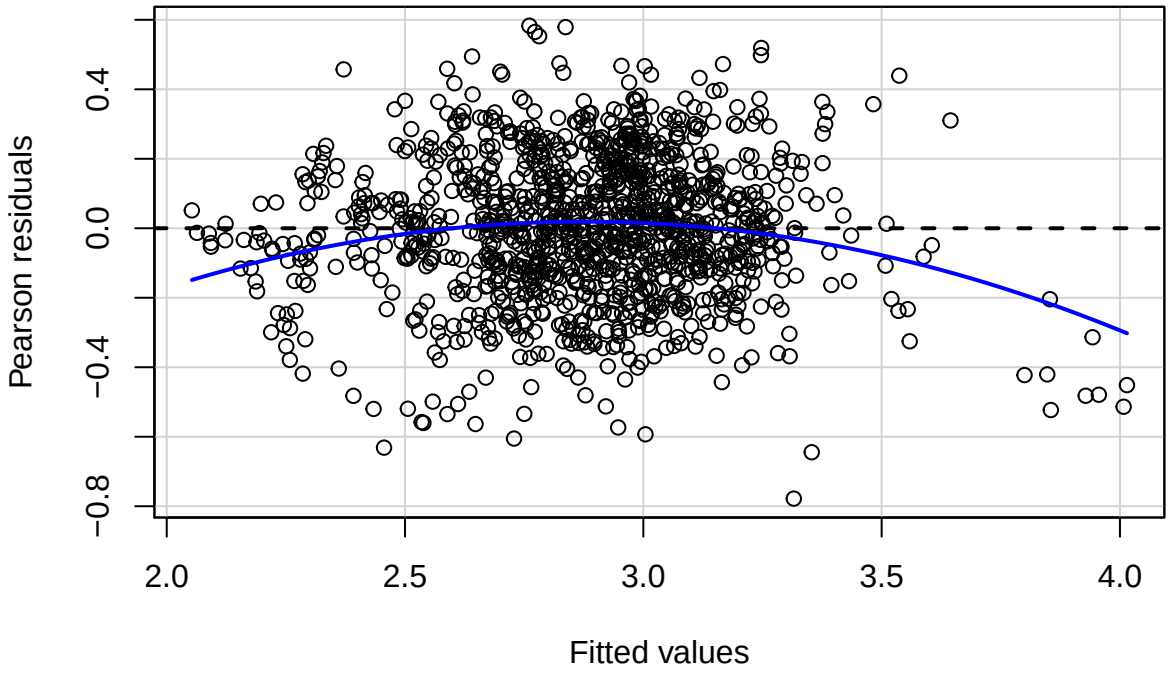
```
residualPlot(model.3)
```



We see that with the set of variables and transformations we were using, we satisfy zero conditional mean (no correlation of fitted values with the residuals). In addition, we appear to have heteroskedasticity, and therefore performed our standard error analysis with robust errors.

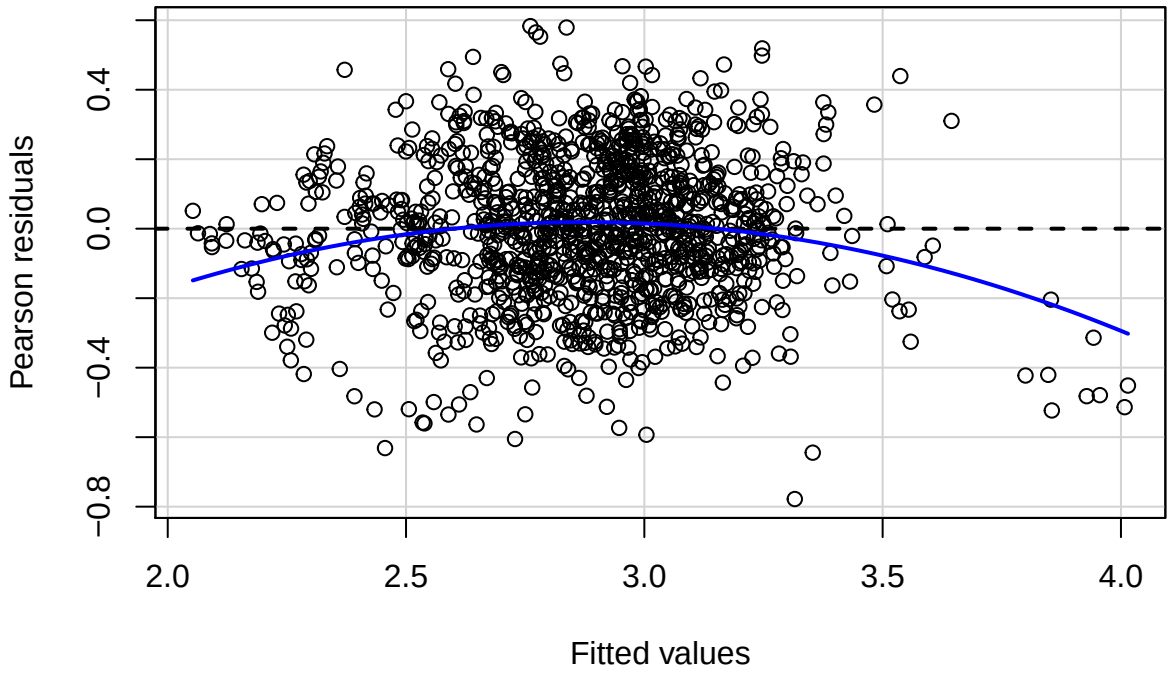
Model 3, but with log transformation on totfatrte, introduced a pattern in the residuals:

```
model.3.log <- lm(log(totfatrte) ~ factor(year) + sbprim + sbsecon +
    perc14_24 + log(unem) + vehicmilespc + bac_none_08_10 + perse +
    sl70plus + gdl, data = data.fit)
residualPlot(model.3.log)
```



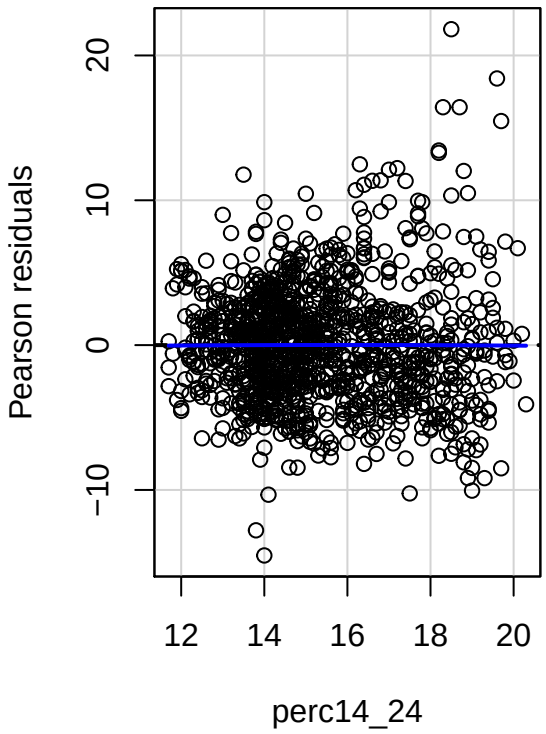
Model 3, but with log transformation on vehicmilespc. Even though log transformation of this variable removed extreme outliers, the residual plot now shows a pattern against this fitted values.

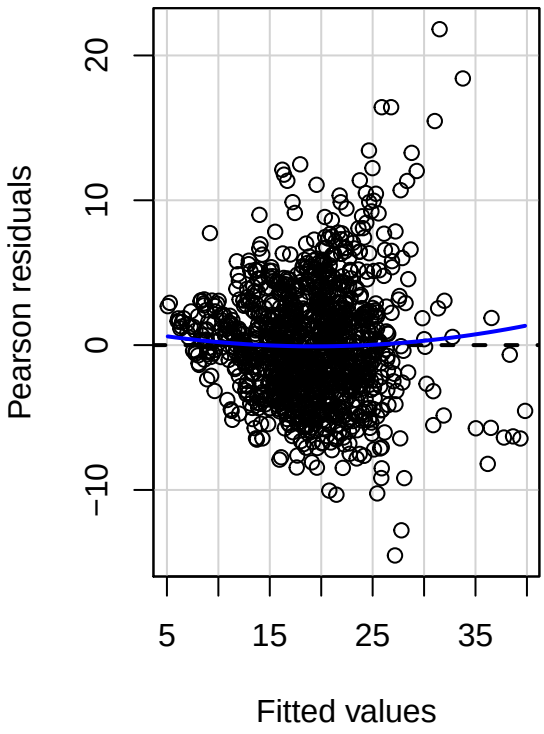
```
model.3.veh.log <- lm(totfatrte ~ factor(year) + sbprim + sbsecon +
    perc14_24 + log(unem) + log(vehicmilespc) + bac_none_08_10 +
    perse + sl70plus + gdl, data = data.fit)
residualPlot(model.3.log)
```



Residual plot for model.3 with non-transformed perc14_24. We see that there is no correlation of the residuals with this variable, making it suitable for use without transformation.

```
residualPlots(model.3, terms = ~perc14_24)
```






```
##            Test stat Pr(>|Test stat|)
## perc14_24    -0.1101           0.9123
## Tukey test    1.3903           0.1645
```

*This work was done as part of the W271 - Statistical Methods for Discrete Response, Time Series, and Panel Data course under the U.C. Berkeley Master of Information and Data Science program.