Karthik Srinivasan, Kineret Stanley, Kevin Fu: W271 Group Lab 3

Due Sunday 8 August 2021 11:59pm

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 economic and demographic variables are also included. The description of the each of the variables in the dataset is come with the dataset.

Q1

1. (30%) Load the data. Provide a description of the basic structure of the dataset, as we have done throughout the semester. Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable *totfatrte* and the potential explanatory variables. You need to write a detailed narrative of your observations of your EDA. *Reminder: giving an "output dump" (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive a zero in this exercise.*

```
load("driving.RData")
df <- data

# check to see if it's balanced
table(df$year)
```

```
##
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
##   48   48   48   48   48   48   48   48   48   48   48   48   48   48   48   48
## 1996 1997 1998 1999 2000 2001 2002 2003 2004
##   48   48   48   48   48   48   48   48   48
```

```r
colnames(df)
```

```
##  [1] "year"         "state"        "sl55"         "sl65"         "sl70"
##  [6] "sl75"         "slnone"       "seatbelt"     "minage"       "zerotol"
## [11] "gdl"          "bac10"        "bac08"        "perse"        "totfat"
## [16] "nghtfat"      "wkndfat"      "totfatpvm"    "nghtfatpvm"   "wkndfatpvm"
## [21] "statepop"     "totfatrte"    "nghtfatrte"   "wkndfatrte"   "vehicmiles"
## [26] "unem"         "perc14_24"    "sl70plus"     "sbprim"       "sbsecon"
## [31] "d80"          "d81"          "d82"          "d83"          "d84"
## [36] "d85"          "d86"          "d87"          "d88"          "d89"
## [41] "d90"          "d91"          "d92"          "d93"          "d94"
## [46] "d95"          "d96"          "d97"          "d98"          "d99"
## [51] "d00"          "d01"          "d02"          "d03"          "d04"
## [56] "vehicmilespc"
```

```r
robust_se <- function(mod, type = 'HC3') {
  sqrt(diag(vcovHC(mod, method = "arellano", type)))
}
#arellano recommended for both heteroskedasticity and serial correlation.
```

Description of Dataset

The dataset comprises of fatalities recorded from the continental 48 states between 1980 and 2004 with no missing or NA values in any of the columns. In particular, the total fatality rate is calculated per 100,000 in each and every US state. The dataset consists of 1200 observations with 56 variables, defined in detail in Appendix A and Appendix B.

The data is constructed in a panel format, with each panel consisting of data from the 48 states for each year between 1980 and 2004. In all, this results in 25 year panels, which result in the 1200 data points (25*48). The panel is balanced, and the 56 variables can be divided into the following groups:

Laws per state

Various traffic laws such as blood alcohol concentrations (bac08, bac10), Per Se laws (perse), zero tolerance laws (zerotol), speed limit laws (sl60, sl65, ...) and seat-belt laws (sbprim, sbsecon) are recorded for each state. These variables take up a value of 0 or 1, depending on whether the law was enforced in any given year. The years in which the laws were introduced results in a fractional value being recorded between 0 and 1. For example, a bac08 law (Blood Alcohol Concentration limit of 0.08%) that changed in March will get assigned 0.25 for that year while receiving 0 (for previous years) and 1 (for future years). The laws present in the dataset are defined in Appendix A.

Fatality Statistics per state

The dataset also records the fatalities as a result of drunk driving. The total fatality rate is measured as the number of fatalities per 100,000 of the population in each state. In addition to the

total fatality rate, the dataset also contains information on the night and weekend fatality rates, presumably due to a higher likelihood of drunk driving during this periods. The aforementioned fatality rates are also measured per 100 million miles of driving. The variables associated with the fatality rates are given by totfatrte, totfatpvm, wkndfatrte, nghtfatrte...

Demographics

The rest of the variables record the population statistics and the driving behavior (per capita mileage). In addition to the raw population statistics, the dataset also records the unemployment rate over the years alongwith the percentage of population between 14 and 24 years.

Dummy Variables

The dataset also contains a on-hot encoded dummy variable for each of years from 1980 to 2004. This is provided as a form of convenience and does not add any value to the overall dataset.

Please refer to Appendix C for some of the lengthier code blocks that drive the analyses and charts below.

```
# drop unused features and edit / reformat others
columns_to_keep <- c('totfatrte', 'year', 'state', 'bac08', 'bac10',
'perse', 'sbprim', 'sbsecon', 'sl70plus', 'gdl', 'perc14_24', 'unem', 'vehicmilespc')
numeric_columns <- c('totfatrte', 'perc14_24', 'unem', 'vehicmilespc')
round_up_columns <- c( 'bac08', 'bac10', 'perse', 'sl70plus', 'gdl')
factor_columns <- c(c('state','sbprim', 'sbsecon', 'us_state'), round_up_columns)
df <- df %>%
  # keep subset of columns
  dplyr::select(columns_to_keep) %>%
  # add in US state names
  merge(y = us_states, by.x = "state", by.y = "state") %>%
  # for partial years when law changed round up
  mutate(across(round_up_columns, round)) %>%
  mutate_each(funs(as.factor), factor_columns) %>%
  mutate_each(funs(as.numeric), numeric_columns)
```

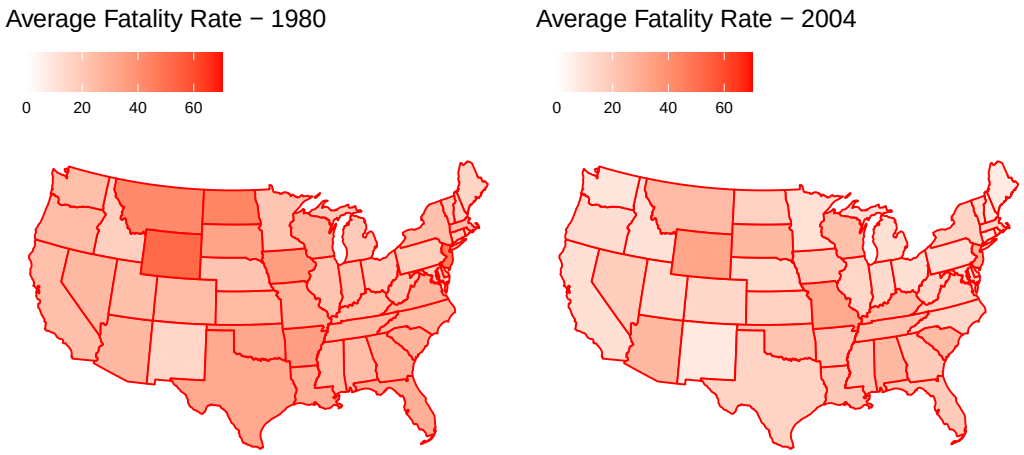
After all the data manipulations above, we get the following cleansed and reduced form of our dataset.

```
knitr::kable(head(df[,1:13],5), caption = "Cleaned and Reduced Dataset", digits = 1) %>%
  kable_styling(latex_options="scale_down")
```

We also visualize the total fatality rates at the end of 1981 and compare them at the endof 2004. We observe that the overall fatality rates across the states have reduced, and that the states that had high fatality rates in 1981 continue to be so at the end of 2004.

Table 1: Cleaned and Reduced Dataset

state	totfatrte	year	bac08	bac10	perse	sbprim	sbsecon	sl70plus	gdl	perc14_24	unem	vehicmilespc
1	24.1	1980	0	1	0	0	0	0	0	18.9	8.8	7544
1	24.1	1981	0	1	0	0	0	0	0	18.7	10.7	7108
1	21.4	1982	0	1	0	0	0	0	0	18.4	14.4	7607
1	23.6	1983	0	1	0	0	0	0	0	18.0	13.7	7880
1	23.6	1984	0	1	0	0	0	0	0	17.6	11.1	8334



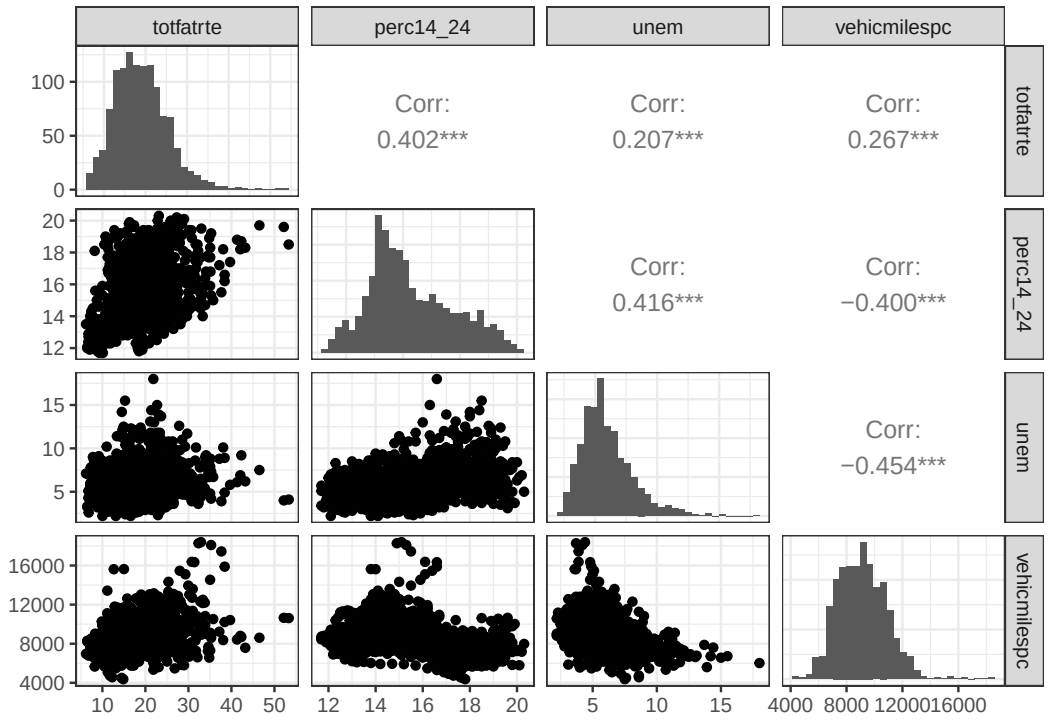
Dependent and Independent Variables

We would like to predict the total fatality rate (*totfatrte*) based on the various state laws (speed limits, drinking laws, and seat belt laws), driving behavior (vehicmilespc), and state demographics.

Continuous Variables

- We start by examining the distributions of the continuous variables (totfatrte, unem, perc14_24, vehicmilespc). From the distributions and skewness test below, we find that the variables `unem` variable is strongly skewed and after reviewing the Box-Cox lambda for this variable (0.1), we determined that a log transformation was sufficient to reduce the variance of the distribution.
- Correlation plot reveals positive correlations between the dependent variable totfatrte and a) perc14_24, b) unemployment rate and c) vehicmilespc.

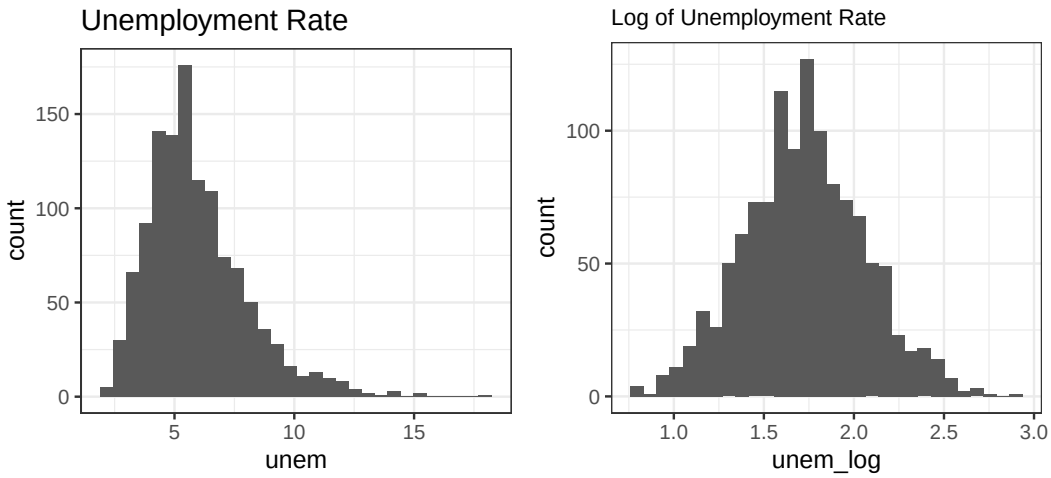
```
df %>% dplyr::select(all_of(numeric_columns)) %>% ggpairs(diag=list(continuous="barDiag"))
```




```
# check for skewness
apply(df[,c("totfatrte","perc14_24", "unem", "vehicmilespc")], 2, skewness)
```

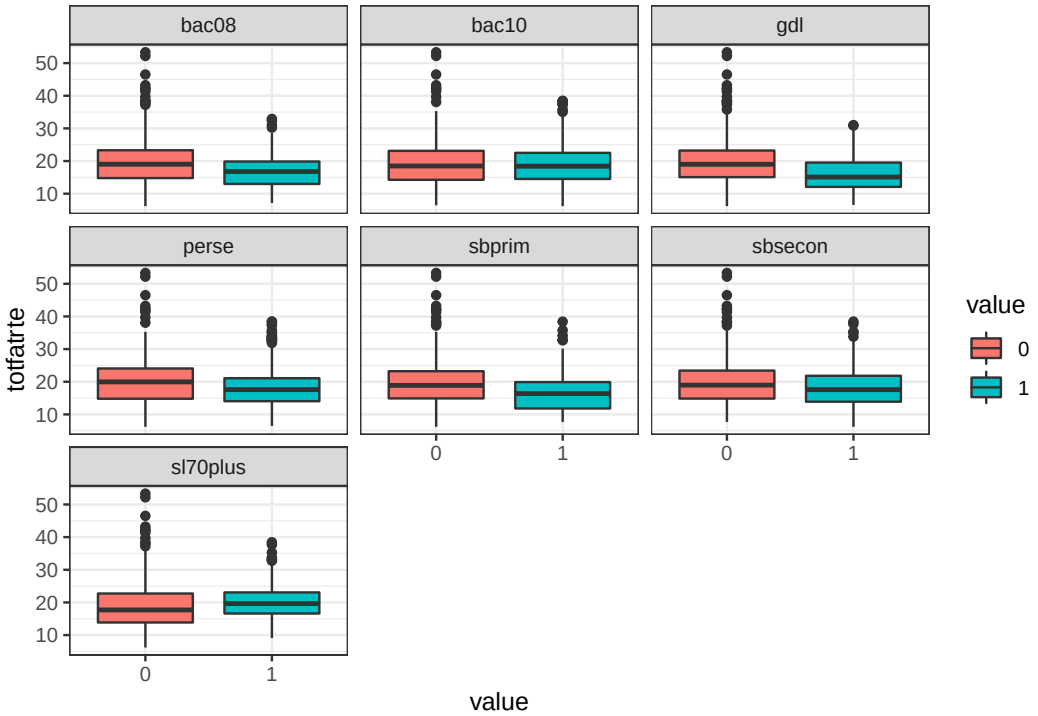
```
##    totfatrte    perc14_24         unem vehicmilespc
##       0.7900       0.5183       1.1755       0.7257
```

```
## save transformed variables
df$unem_log <- log(df$unem)

#plot
unem_raw_plot <- ggplot(df) + geom_histogram(aes(x = unem)) +
  ggtitle('Unemployment Rate')
unem_log_plot <- ggplot(df) + geom_histogram(aes(x = unem_log)) +
  ggtitle('Log of Unemployment Rate') +
  theme(plot.title = element_text(size = 10, lineheight=1))
grid.arrange(unem_raw_plot, unem_log_plot, ncol = 2)
```



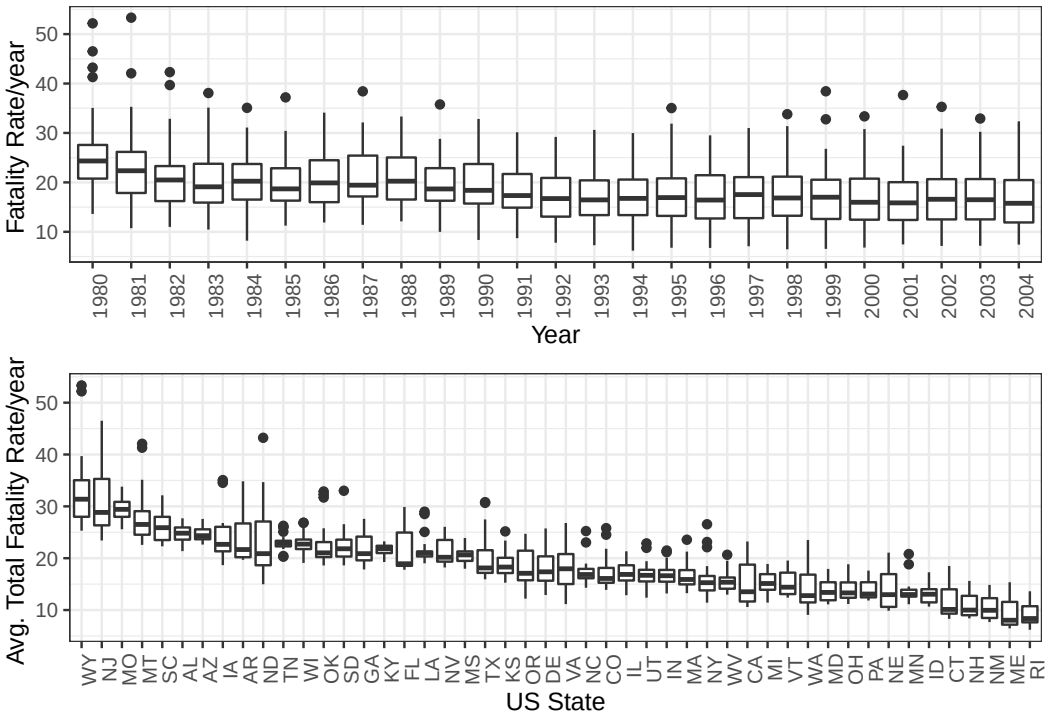
Categorical/Binary Variables

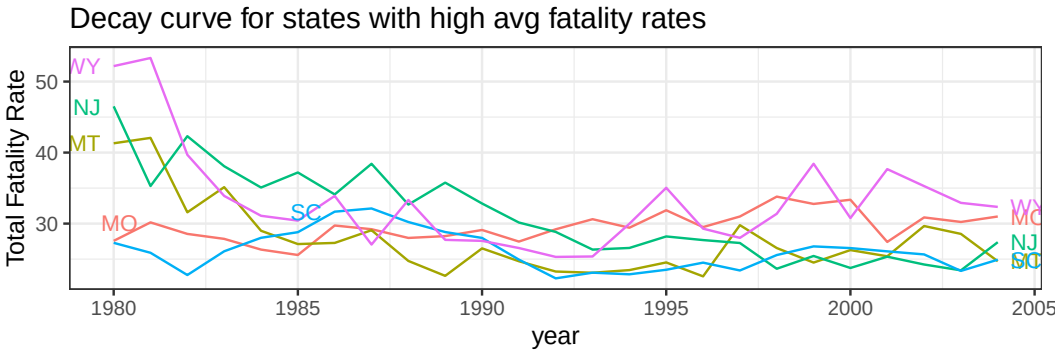
The categorical/binary variable of interest are a) bac08, b) bac10, c) spl70plus, d) perse, e) gdl, f) sbprim, g) sbsecon and h) the US states. We examine the effect of the categorical variables using box plots. We observe that the introduction of state seatbelt and drinking laws appear to have a positive effect in reducing the fatalities. We also observe that the BAC10 and sl70plus have no/adverse impact on the fatality rates. This is misleading in the current context since the removal of BAC10 implies introduction of BAC08, and we have no data pre-dating the introduction of BAC10 law. Similarly the non-existence of a sl70plus law does not imply the non-existence of a speed law. In fact, the laws may have become stricter (sl65, sl60 etc).

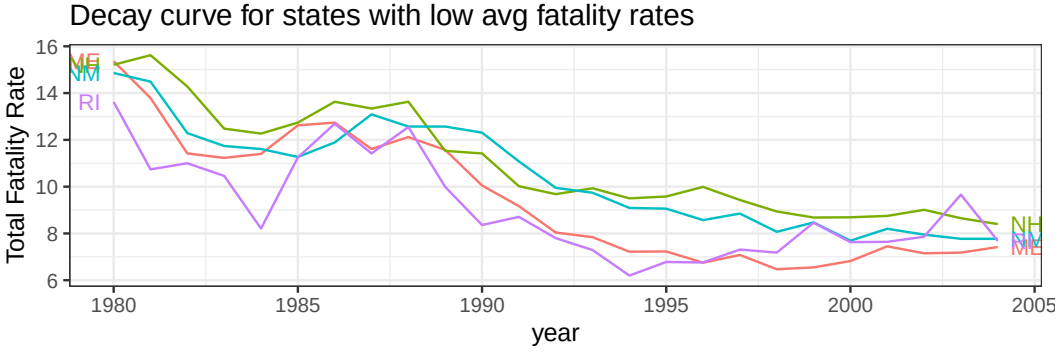


We also examine the decay curve of average total fatality rates over the years and across the states.

We find that the average fatality rates have declined but at a slow rate. However, certain states (WY, NJ, MO, MT, SC) have had high fatality rates over the years, while NH, NM, ME and RI have ~3x lower fatality rates.



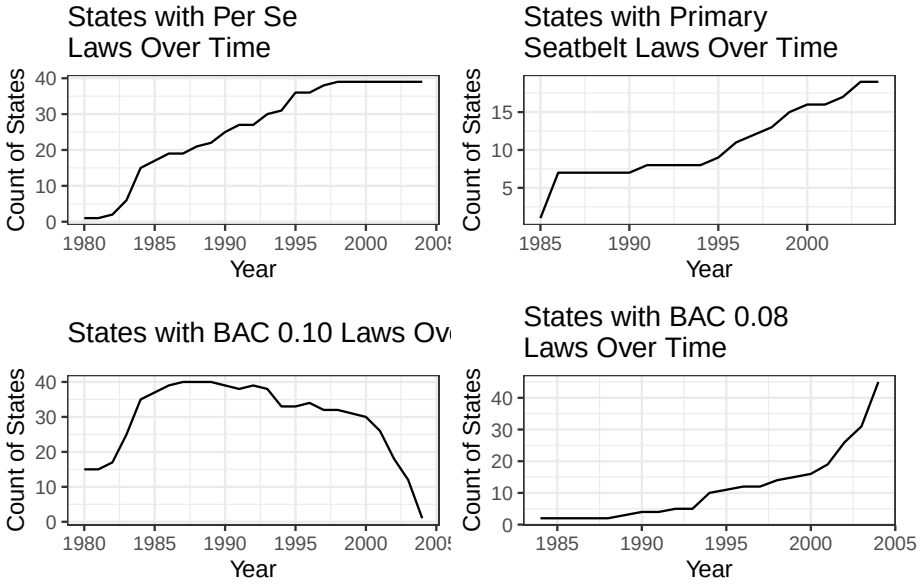




BAC08, BAC10 and Per Se laws

The blood alcohol limit in most states was first introduced at the 0.1% level and then reduced to 0.08%. The years when this change was introduced are represented as fractions, i.e, a 0.333 bac08 and a 0.667 bac10 would imply that the law had changed during the end of April. This also leads to the situation where bac08 and bac10 are strongly negatively correlated. From the EDA, we find that all states had either bac08 or bac10 enforced by the end of 2004. 44 of the 48 states had fully enforced bac08 by the end of 2004, while 3 of them were getting them enforced in 2004. On the other hand, bac10 was enforced in all the states. SC, TN and MD were the last three states to implement the bac10 law. On the other side of the spectrum, OR and UT introduced the much more stricter bac08 law in 1984.

At the end of 2004, 9 of states had no perse laws, 3 had no bac 08 limits enforced, 0 had no (either 0.08 or 0.10) bac limits enforced. The last state to enact a bac10 law was SC in 2001.



Q2

2. (15%) How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this model explain? Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.

Answer

The *totfatrte* variable is a continuous variable that denotes the total fatalities per 100,000 population, i.e., the ratio of number of fatalities occurring over a given year to the total population of the state, multiplied with 100,000. The average fatality rate per year is given by

Table 2: Average Total Fatality Rate

year	Avg	year	Avg	year	Avg	year	Avg	year	Avg
1980	25.49	1985	19.85	1990	19.51	1995	17.67	2000	16.83
1981	23.67	1986	20.80	1991	18.09	1996	17.37	2001	16.79
1982	20.94	1987	20.77	1992	17.16	1997	17.61	2002	17.03
1983	20.15	1988	20.89	1993	17.13	1998	17.27	2003	16.76
1984	20.27	1989	19.77	1994	17.16	1999	17.25	2004	16.73

We now fit a linear regression model using year as the dummy variable. This model examines the change in total fatality rates since 1980. We find almost all of the coefficients to be statistically significant, except for year 1981.

However, results from this model should be used with caution, because the repeated observations violate the independent and identically distributed assumption for an OLS model. The Durbin-Watson test confirms the violation.

The residuals are normally distributed according to a visual observation of the qq-plots and the Breusch-Pagan test p-value is sufficiently large that we fail to reject the null hypothesis (that the residuals are homoskedastically distributed).

Since, the longitudinal data presented here violate the fundamental assumptions of independence and homogeneity of variance, the estimates are not reliable and the statistics are invalid and therefore the statistical inference may be incorrect.

```
lm1 <- lm(formula=totfatrte~factor(year), data=df)
stargazer(lm1, type = "text", single.row = TRUE)
```

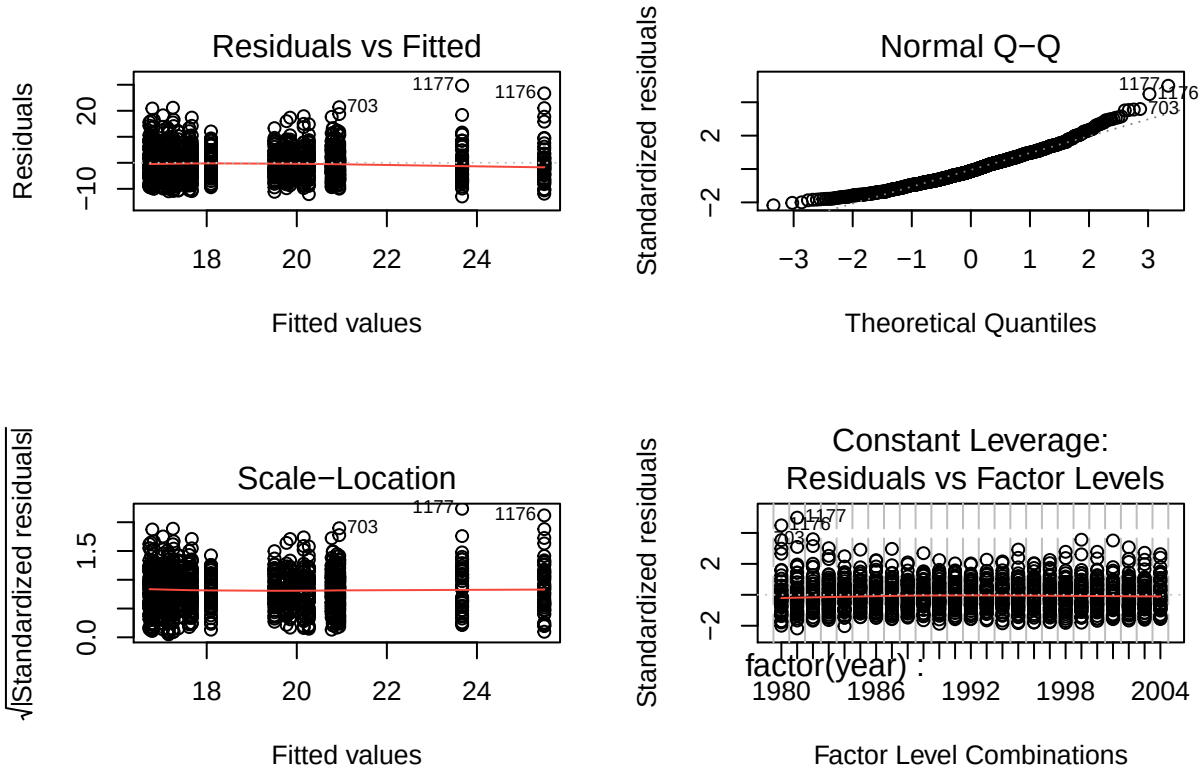
```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                totfatrte
## ------------------------------------------------
## factor(year)1981           -1.824 (1.226)
## factor(year)1982         -4.552*** (1.226)
## factor(year)1983         -5.342*** (1.226)
## factor(year)1984         -5.227*** (1.226)
## factor(year)1985         -5.643*** (1.226)
## factor(year)1986         -4.694*** (1.226)
## factor(year)1987         -4.720*** (1.226)
## factor(year)1988         -4.603*** (1.226)
## factor(year)1989         -5.722*** (1.226)
## factor(year)1990         -5.989*** (1.226)
## factor(year)1991         -7.400*** (1.226)
## factor(year)1992         -8.337*** (1.226)
## factor(year)1993         -8.367*** (1.226)
## factor(year)1994         -8.339*** (1.226)
```

```
## factor(year)1995         -7.826*** (1.226)
## factor(year)1996         -8.125*** (1.226)
## factor(year)1997         -7.884*** (1.226)
## factor(year)1998         -8.229*** (1.226)
## factor(year)1999         -8.244*** (1.226)
## factor(year)2000         -8.669*** (1.226)
## factor(year)2001         -8.702*** (1.226)
## factor(year)2002         -8.465*** (1.226)
## factor(year)2003         -8.731*** (1.226)
## factor(year)2004         -8.766*** (1.226)
## Constant                 25.500*** (0.867)
## ------------------------------------------------
## Observations                    1,200
## R2                              0.128
## Adjusted R2                     0.110
## Residual Std. Error       6.008 (df = 1175)
## F Statistic           7.164*** (df = 24; 1175)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
# Durbin Watson
durbinWatsonTest(lm1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1          0.8973        0.1997       0
##  Alternative hypothesis: rho != 0
```

```
# residual plots
par(mfrow=c(2,2))
plot(lm1)
```




```
# Breusch-Pagan test
bptest(lm1)
```

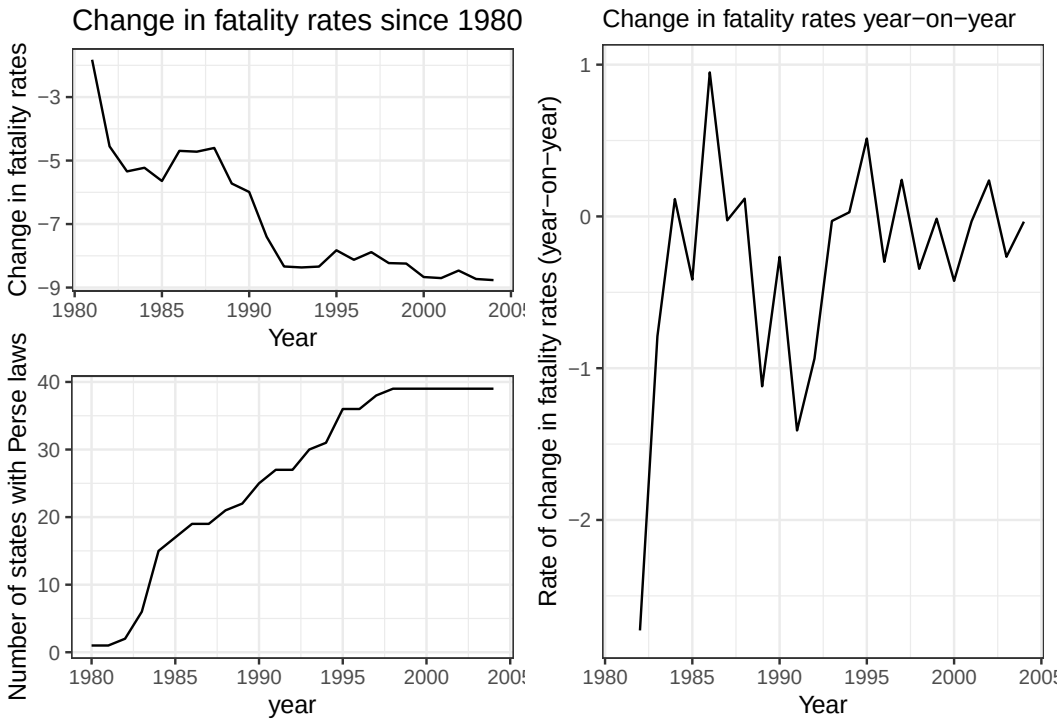
```
##
##  studentized Breusch-Pagan test
##
## data:  lm1
## BP = 25, df = 24, p-value = 0.4
```

The linear regression model is then given by,

$$totfatrte = 25.495 - 1.824\delta_{1981} - 4.552\delta_{1982}...$$

From the model coefficient plots (left figure), we find that the fatality rates are decreasing over the years. The rate of change of this decrease had stabilized after 1995. This is observed in the rate of change plot on the right. We see a sharp decline in fatality rates between 1985 and 1990. This coincides with the state drinking laws coming into effect in various states. In particular, the number of states that introduced the perse law over the years, shown in the plot below, reflect the decline in total fatality rates over the years. The rate of decline has saturated as the number of states with perse laws plateaued. In all 9 states had not yet introduced perse laws in 2004.

```
coef_model <- coef(lm1)
coef_df <- data.frame(year=sort(unique(df$year))[2:length(coef_model)],
                      coefs=coef_model[2:length(coef_model)])
g1 <- ggplot(coef_df, aes(year,coefs)) + geom_line()
g1 <- g1 + theme_bw() + labs(x='Year', y='Change in fatality rates')
g1 <- g1 + ggtitle('Change in fatality rates since 1980')
temp_df <- df %>% group_by(year) %>%
        summarise(num_states = sum(as.numeric(levels(perse))[perse]))
g11 <- ggplot() + geom_line(data=temp_df,aes(x=year, y=num_states))
g11 <- g11 + theme_bw() + labs(y='Number of states with Perse laws')
coef_df$diff_coefs <- difference(coef_df$coefs)
g2 <- ggplot(coef_df, aes(year,diff_coefs)) + geom_line()
g2 <- g2 + theme_bw() + labs(x='Year', y='Rate of change in fatality rates (year-on-year)')
g2 <- g2 + ggtitle('Change in fatality rates year-on-year') +
  theme(plot.title = element_text(size = 11, lineheight=1))

grid.arrange(grid.arrange(g1,g11,ncol=1),g2,ncol=2)
```



Q3

3. (15%) Expand your model in *Exercise 2* by adding variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*, and perhaps *transformations of some or all of these variables*. Please explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed. How are the variables *bac8* and *bac10* defined? Interpret the

coefficients on *bac8* and *bac10*. Do *per se laws* have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)

We find that the skewness of the continuous variables is most observed in the 'unem' variable. A Box-Cox transform of this variable results in a $\lambda = 0.1$, as described in the EDA. For low value of λ, we prefer to use a log transform for better explainability of model coefficients. Also, since the skewness of the other continuous variables are not significant, we do not transform them.

In addition to violating I.I.D. assumptions like the model in Q2, this model also has heteroskedasticity in its residuals and thus we are even more skeptical of the outcomes.

As a reminder, if the law changed halfway through the year or more we consider the law as implemented in a state; otherwise we do not consider it a law.

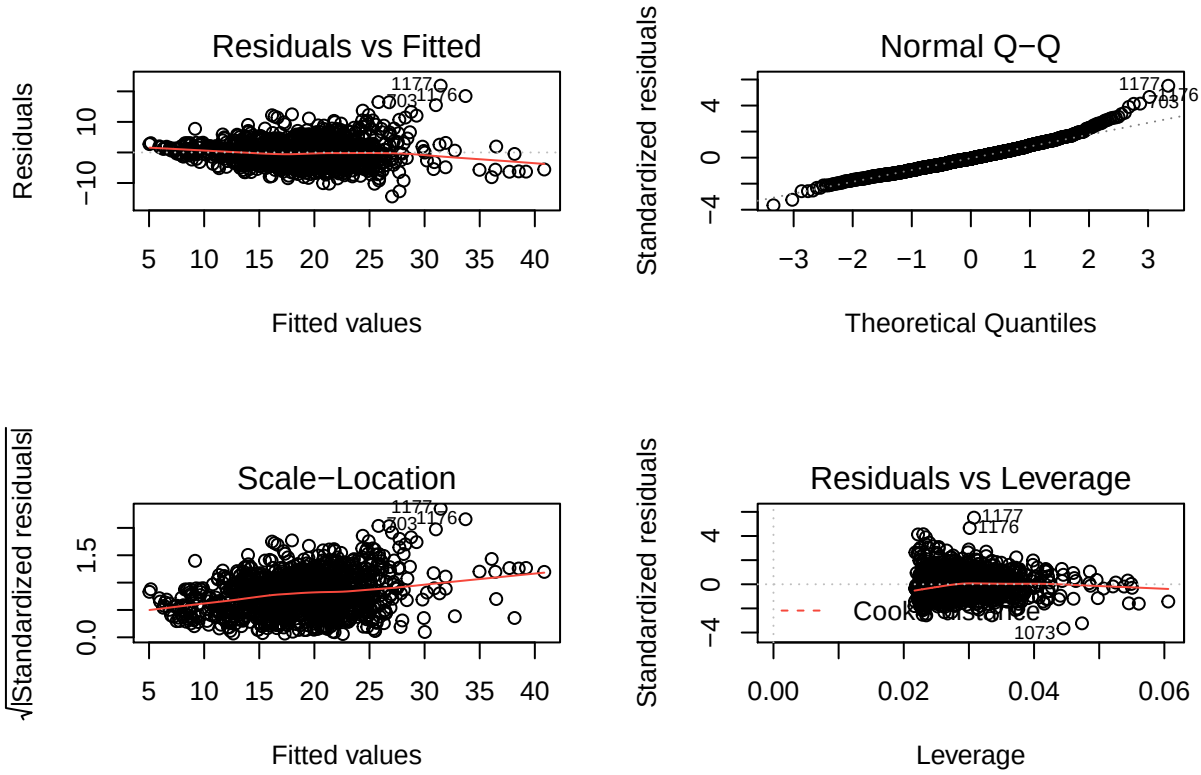
- The coefficient for `bac08` and `bac10` indicate whether a state had a law restricting blood alcohol content for drivers and the effect of having such a law on fatality rates. Specifically, implementing a `bac08` led to a reduction of 2.04 per 100,000 people in `totfatrte` and `bac10` had a smaller impact of 0.94 per 100,000. This aligns with our intuition that more stringent requirements lead to fewer fatalities. Both were significant at 95% confidence level.
- `perse` also was significant and led to a decrease of fatalities equal to 0.7 per 100,000 people.
- `sbprim` laws did not have a statistically significant effect on `totfatrte`.

Models are in Appendix D, and are reported with robust standard errors.

The new linear regression model is now given by,

$$\begin{aligned} totfatrte = & -8.01 - 2.29bac08 - 1.26bac10 - 0.562perse - 0.38sbprim \\ & -0.153sbsecon - 3.11sl70plus - 0.3gdl + 0.18perse + 5.15log(unem) + 0.0029vehicmilespc \\ & -2.11\delta_{1981} - 6.30\delta_{1982}... \end{aligned}$$

```
lm2 <- lm(formula=totfatrte ~ factor(year) + bac08 + bac10 +
            perse + sbprim + sbsecon + sl70plus + gdl +
            perc14_24 + unem_log + vehicmilespc, data=df)

mod_pooled <- plm(totfatrte ~ factor(year) + bac08 + bac10 +
            perse + sbprim + sbsecon + sl70plus + gdl +
            perc14_24 + unem_log + vehicmilespc,
            data=df, model='pooling', index = c("state", "year"))

# residuals
par(mfrow=c(2,2))
plot(lm2)
```




```
# Breusch-Pagan test
bptest(lm2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm2
## BP = 141, df = 34, p-value = 5e-15
```

Q4

4. (15%) Reestimate the model from *Exercise 3* using a fixed effects (at the state level) model. How do the coefficients on *bac08, bac10, perse, and sbprim* compare with the pooled OLS estimates? Which set of estimates do you think is more reliable? What assumptions are needed in each of these models? Are these assumptions reasonable in the current context?

The Pooled OLS model is equivalent to the Linear Model produced above, see Appendix D for model summary table.

- In the pooled OLS model, `bac08` is highly statistically significant at the 95% and even 99% confidence interval, while `bac10` and `perse` show statistical significance at the 95% confidence interval. `sbprim` was not statistically significant.

- In the fixed effect model, we see a difference set of results. All four variables (`bac08, bac10, perse, and sbprim`) are statistically significant at the 99% confidence interval.

Tests

- We ran F-tests (the Chow Test) and found the p-value is significant and so we reject the null hypothesis (that the same coefficients apply across all individuals). Indicating that the Fixed Effects model is appropriate. *We ran a Lagrange Multiplier Test below and the p-value is significant indicating that we should reject the null hypothesis (residuals across entities are correlated) and suspect that FE model may be more appropriate.
- We ran a Breusch-Godfrey test for serial correlation and found the p-value is significant so we reject the null hypothesis (that there is no serial correlation), and conclude the FE model is more appropriate.

FE model

The FE model is indicated based on the tests above.

Based on the findings in Q3, the unobserved effect is correlated with the explanatory variables and so the pooled OLS / linear models result in biased and inconsistent estimates. The pooled effect model assumes that the independent variables are uncorrelated with the error term. The FE model removes the unobserved effect and leads to more robust and reliable estimates, because it assumes that the independent variables are correlated with the error term.

The FE model is then defined by,

$$totfatrte_{it} = \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \alpha_i + u_{it}$$

where,

$$X \in \{bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14\_24, log(unem), vehicmilespc\}$$

and α_i are the time-invariant unobserved effects observed for each state i, $t \in [1980, 2004]$, u_{it} are the errors, and X are the explanatory variables.

Note: In our `robust_se` function, we use `arellano` in our `vcovHC` parameter because we have both heteroskedasticity and serial correlation. We also use `HC3` in our `type` parameter because it gives less weight to outliers.

```
# F-test / Chow Test
fixed_effects_mod.time <- pvcm(totfatrte ~  bac08 + bac10 +
  perse + sbprim + sbsecon + sl70plus + gdl +
  perc14_24 + unem_log + vehicmilespc, data=df, model="within",
            index = c("state", "year"))

fixed_effects_mod <- plm(totfatrte ~ bac08 + bac10 +
  perse + sbprim + sbsecon + sl70plus + gdl +
  perc14_24 + unem_log + vehicmilespc, data=df, model="within",
            index = c("state", "year"))

pooltest(fixed_effects_mod, fixed_effects_mod.time)
```

```
##
##  F statistic
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +  ...
## F = 3, df1 = 470, df2 = 672, p-value <2e-16
## alternative hypothesis: unstability
```

```
# Lagrange Multiplier Test
plmtest(fixed_effects_mod, c("time"), type=("bp"))
```

```
##
##  Lagrange Multiplier Test - time effects (Breusch-Pagan) for balanced
##  panels
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +  ...
## chisq = 213, df = 1, p-value <2e-16
## alternative hypothesis: significant effects
```

```
# Serial Correlation
pbgtest(fixed_effects_mod )
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +     gdl + perc14_24
## chisq = 402, df = 25, p-value <2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
# FE model
mod_fe <- plm(totfatrte ~ factor(year) + bac08 + bac10 +
            perse + sbprim + sbsecon + sl70plus + gdl +
            perc14_24 + unem_log + vehicmilespc,
                  data=df, model='within', index = c("state", "year"))

summary(mod_fe, vcov=vcovHC(mod_fe, method="arellano", type="HC3"))
```

```
## Oneway (individual) effect Within Model
##
## Note: Coefficient variance-covariance matrix supplied: vcovHC(mod_fe, method = "arellano", t
##
## Call:
## plm(formula = totfatrte ~ factor(year) + bac08 + bac10 + perse +
##     sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem_log +
##     vehicmilespc, data = df, model = "within", index = c("state",
##     "year"))
##
```

```
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -8.2419 -1.0356 -0.0138  0.9754 14.6325
##
## Coefficients:
##                    Estimate Std. Error t-value Pr(>|t|)
## factor(year)1981 -1.58e+00   4.42e-01   -3.58  0.00036 ***
## factor(year)1982 -3.37e+00   4.44e-01   -7.59  6.8e-14 ***
## factor(year)1983 -4.03e+00   4.62e-01   -8.72  < 2e-16 ***
## factor(year)1984 -4.55e+00   4.60e-01   -9.88  < 2e-16 ***
## factor(year)1985 -5.00e+00   4.81e-01  -10.38  < 2e-16 ***
## factor(year)1986 -3.99e+00   6.07e-01   -6.57  7.8e-11 ***
## factor(year)1987 -4.67e+00   7.00e-01   -6.68  3.8e-11 ***
## factor(year)1988 -5.21e+00   7.84e-01   -6.64  4.8e-11 ***
## factor(year)1989 -6.52e+00   8.89e-01   -7.34  4.2e-13 ***
## factor(year)1990 -6.58e+00   9.38e-01   -7.02  4.0e-12 ***
## factor(year)1991 -7.25e+00   1.01e+00   -7.19  1.2e-12 ***
## factor(year)1992 -8.13e+00   1.10e+00   -7.37  3.3e-13 ***
## factor(year)1993 -8.47e+00   1.15e+00   -7.39  2.8e-13 ***
## factor(year)1994 -8.94e+00   1.12e+00   -7.98  3.6e-15 ***
## factor(year)1995 -8.71e+00   1.21e+00   -7.18  1.2e-12 ***
## factor(year)1996 -9.13e+00   1.18e+00   -7.71  2.7e-14 ***
## factor(year)1997 -9.39e+00   1.25e+00   -7.51  1.2e-13 ***
## factor(year)1998 -1.01e+01   1.26e+00   -8.03  2.5e-15 ***
## factor(year)1999 -1.03e+01   1.36e+00   -7.62  5.5e-14 ***
## factor(year)2000 -1.10e+01   1.35e+00   -8.14  1.0e-15 ***
## factor(year)2001 -1.05e+01   1.47e+00   -7.11  2.1e-12 ***
## factor(year)2002 -9.60e+00   1.48e+00   -6.48  1.4e-10 ***
## factor(year)2003 -9.64e+00   1.53e+00   -6.32  3.8e-10 ***
## factor(year)2004 -1.01e+01   1.68e+00   -6.00  2.7e-09 ***
## bac081           -1.10e+00   6.01e-01   -1.84  0.06637 .
## bac101           -8.04e-01   3.44e-01   -2.34  0.01951 *
## perse1           -1.13e+00   3.92e-01   -2.87  0.00414 **
## sbprim1          -1.19e+00   5.61e-01   -2.12  0.03429 *
## sbsecon1         -3.04e-01   3.76e-01   -0.81  0.41955
## sl70plus1         4.71e-02   5.23e-01    0.09  0.92821
## gdl1             -2.83e-01   3.68e-01   -0.77  0.44324
## perc14_24         1.67e-01   1.72e-01    0.97  0.33178
## unem_log         -3.71e+00   7.61e-01   -4.87  1.3e-06 ***
## vehicmilespc      9.51e-04   3.58e-04    2.65  0.00804 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12100
## Residual Sum of Squares: 4550
## R-Squared:      0.625
```

```
## Adj. R-Squared: 0.598
## F-statistic: 59.7253 on 34 and 47 DF, p-value: <2e-16
```

Q5

5. (10%) Would you prefer to use a random effects model instead of the fixed effects model you built in *Exercise 4*? Please explain.

In theory the random effects model allows for time invariant explanatory variables to be included in our model. A fixed effect model eliminates all time invariant variables. However, RE assumes that the unobserved effect is uncorrelated with all explanatory variables in all time periods, and that's difficult to believe in a case where state-level differences are likely to exist.

We also run the Hausman test to make sure that the fixed effects model is preferable over the random effect model (below). Since we generate a small p-value, we reject the null hypothesis (that unique errors are correlated with regressors) and we should use the fixed effects model.

$$totfatrte_{it} = \beta_0 + \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + a_i + u_{it},$$

where i is the state, a_i is the unobserved effect uncorrelated with all explanatory variables X, and $t \in [1980, 2004]$.

```
# random effects model
mod_re <- plm(totfatrte ~ factor(year) + bac08 + bac10 +
            perse + sbprim + sbsecon + sl70plus + gdl +
            perc14_24 + unem_log + vehicmilespc, data=df, model="random",
            index = c("state", "year"))

summary(mod_re, vcov=vcovHC(mod_re, method="arellano", type="HC3"))
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Note: Coefficient variance-covariance matrix supplied: vcovHC(mod_re, method = "arellano", t
##
## Call:
## plm(formula = totfatrte ~ factor(year) + bac08 + bac10 + perse +
##     sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem_log +
##     vehicmilespc, data = df, model = "random", index = c("state",
##     "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Effects:
##                 var std.dev share
## idiosyncratic 4.07    2.02  0.34
## individual    7.82    2.80  0.66
```

```
## theta: 0.857
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##  -8.439  -1.205  -0.160   0.937  16.390
##
## Coefficients:
##                   Estimate Std. Error z-value Pr(>|z|)
## (Intercept)       1.98e+01   4.52e+00    4.38  1.2e-05 ***
## factor(year)1981 -1.61e+00   4.40e-01   -3.66  0.00025 ***
## factor(year)1982 -3.57e+00   4.39e-01   -8.12  4.6e-16 ***
## factor(year)1983 -4.24e+00   4.59e-01   -9.24  < 2e-16 ***
## factor(year)1984 -4.65e+00   4.81e-01   -9.67  < 2e-16 ***
## factor(year)1985 -5.13e+00   5.02e-01  -10.21  < 2e-16 ***
## factor(year)1986 -4.14e+00   6.34e-01   -6.52  6.8e-11 ***
## factor(year)1987 -4.84e+00   7.42e-01   -6.53  6.7e-11 ***
## factor(year)1988 -5.39e+00   8.11e-01   -6.64  3.1e-11 ***
## factor(year)1989 -6.73e+00   9.15e-01   -7.36  1.9e-13 ***
## factor(year)1990 -6.87e+00   9.59e-01   -7.16  8.2e-13 ***
## factor(year)1991 -7.63e+00   1.01e+00   -7.55  4.5e-14 ***
## factor(year)1992 -8.59e+00   1.10e+00   -7.78  7.1e-15 ***
## factor(year)1993 -8.91e+00   1.16e+00   -7.68  1.6e-14 ***
## factor(year)1994 -9.34e+00   1.12e+00   -8.33  < 2e-16 ***
## factor(year)1995 -9.10e+00   1.21e+00   -7.50  6.3e-14 ***
## factor(year)1996 -9.59e+00   1.18e+00   -8.12  4.7e-16 ***
## factor(year)1997 -9.85e+00   1.24e+00   -7.95  1.8e-15 ***
## factor(year)1998 -1.06e+01   1.25e+00   -8.45  < 2e-16 ***
## factor(year)1999 -1.08e+01   1.32e+00   -8.20  2.3e-16 ***
## factor(year)2000 -1.14e+01   1.34e+00   -8.55  < 2e-16 ***
## factor(year)2001 -1.11e+01   1.39e+00   -7.97  1.6e-15 ***
## factor(year)2002 -1.03e+01   1.39e+00   -7.43  1.1e-13 ***
## factor(year)2003 -1.04e+01   1.43e+00   -7.26  3.8e-13 ***
## factor(year)2004 -1.08e+01   1.60e+00   -6.74  1.5e-11 ***
## bac081           -1.21e+00   6.11e-01   -1.98  0.04799 *
## bac101           -8.67e-01   3.54e-01   -2.45  0.01440 *
## perse1           -1.07e+00   3.78e-01   -2.84  0.00452 **
## sbprim1          -1.14e+00   5.43e-01   -2.10  0.03584 *
## sbsecon1         -3.05e-01   3.82e-01   -0.80  0.42458
## sl70plus1         1.27e-01   5.11e-01    0.25  0.80396
## gdl1             -2.59e-01   3.61e-01   -0.72  0.47238
## perc14_24         1.82e-01   1.67e-01    1.09  0.27566
## unem_log         -3.13e+00   7.67e-01   -4.08  4.6e-05 ***
## vehicmilespc      1.19e-03   3.23e-04    3.70  0.00022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12900
## Residual Sum of Squares: 5130
```

```
## R-Squared:      0.601
## Adj. R-Squared: 0.59
## Chisq: 1937.74 on 34 DF, p-value: <2e-16
```

```
# Hausman
phtest(mod_fe, mod_re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ factor(year) + bac08 + bac10 + perse + sbprim + sbsecon +  ...
## chisq = 181, df = 34, p-value <2e-16
## alternative hypothesis: one model is inconsistent
```

As stated above, since the fixed effect model assumptions are more reasonable than the random effect model assumptions, we would opt to use the fixed effect model. Some examples of how the error term could be correlated with the independent variables include:

- Difference of income / cost of living between different states (e.g. higher cost of living in CA/NY vs. Midwest)
- Different political and cultural preferences between residents in different states

Q6

6. (10%) Suppose that *vehicmilespc*, the number of miles driven per capita, increases by 1,000. Using the FE estimates, what is the estimated effect on *totfatrte*? Please interpret the estimate.

```
fe_robust_coef <- coef(mod_fe, vcovHC(mod, method = "arellano", type= "HC3"))
```

The increase in fatalities per 100,000 people would be 0.951. We apply a robust standard error to estimate our coefficient of interest.

Q7

7. (5%) If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors?

The estimators would be incorrect likely due to an overly optimistic standard error. In our models, we applied robust standard errors in order to adjust our estimates.

Appendix

Appendix A: Glossary

Laws per state

Speed limit Laws

- sl55, sl65, sl70, sl75: *(Binary)* An imposed speed limit of 55, 65, 70 and 75mph, respectively.
- slnone: *(Binary)* No imposed speed limit
- sl70plus: *(Binary)* sl70 + sl75 + slnone

Seat Belt Laws

- seatbelt: *(Binary)* =0 if none, =1 if primary, =2 if secondary
- sbprim: *(Binary)* =1 if primary seatbelt law
- sbsecon: *(Binary)* =1 if secondary seatbelt law

Drinking Laws

- minage: *(Continuous)* State level imposition of minimum drinking age (years)
- zerotol: *(Binary)* Zero tolerance laws make it illegal for drivers under age 21 to drive with any measurable amount of alcohol in their system, regardless of the BAC limit for older drivers.
- gdl: *(Binary)* Graduated Driver Licensing (GDL) programs allow young drivers to safely gain driving experience before obtaining full driving privileges. Different states have differing versions of the law implemented based off minimum age requirements for graduating through the phases of the learner programs.
- bac10: *(Binary)* Blood Alcohol Concentration (BAC) legal limit set to .10%. This preceded the bac08 law.
- bac08: *(Binary)* Blood Alcohol Concentration (BAC) legal limit set to .08%, at which driving skills are proven to be compromised.
- perse: *(Binary)* administrative license revocation (ALR), An ALR law gives state officials the authority to suspend administratively the license of any driver who fails or refuses to take a BAC test.

Fatality Statistics per state

- totfat: *(Continuous)* total traffic fatalities
- nghtfat: *(Continuous)* total nighttime fatalities
- wkndfat: *(Continuous)* total weekend fatalities
- totfatpvm: *(Continuous)* total fatalities per 100 million miles
- nghtfatpvm: *(Continuous)* nighttime fatalities per 100 million miles
- wkndfatpvm: *(Continuous)* weekend fatalities per 100 million miles
- totfatrte: *(Continuous)* total fatalities per 100,000 population
- nghtfatrte: *(Continuous)* nighttime fatalities per 100,000 population
- wkndfatrte: *(Continuous)* weekend accidents per 100,000 population

Demographics

- year: *(Continuous)* 1980 through 2004
- state: *(Categorical*) 48 continental states, alphabetical
- statepop: *(Continuous)* state population
- vehicmiles: *(Continuous)* vehicle miles traveled, billions
- unem: *(Continuous)* unemployment rate, percent
- perc14_24: *(Continuous)* percent population aged 14 through 24
- vehicmilespc: *(Continuous)* the number of miles driven per capita

Dummy Variables

- d80 to d04: *(Binary)* One hot encoded features for years 1980 to 2004

Appendix B: More EDA

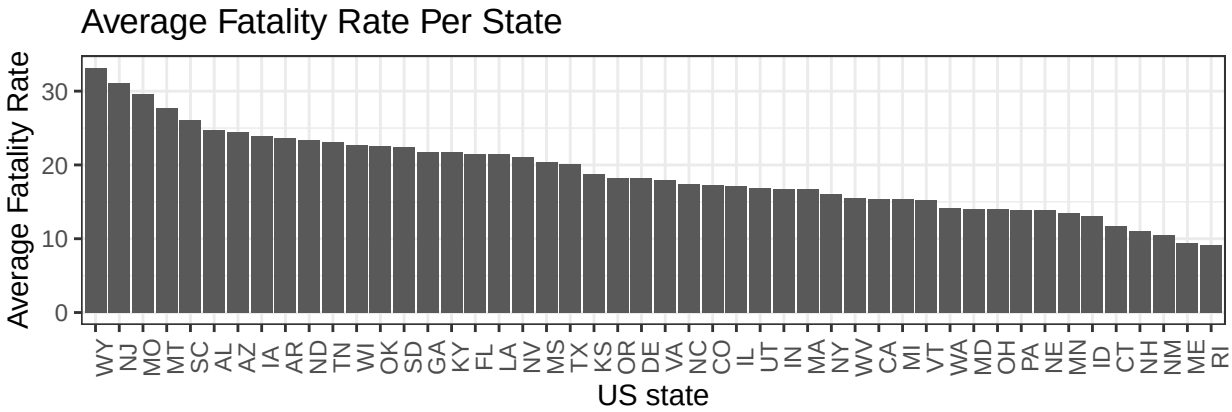
```
#skimr::skim(df)
```

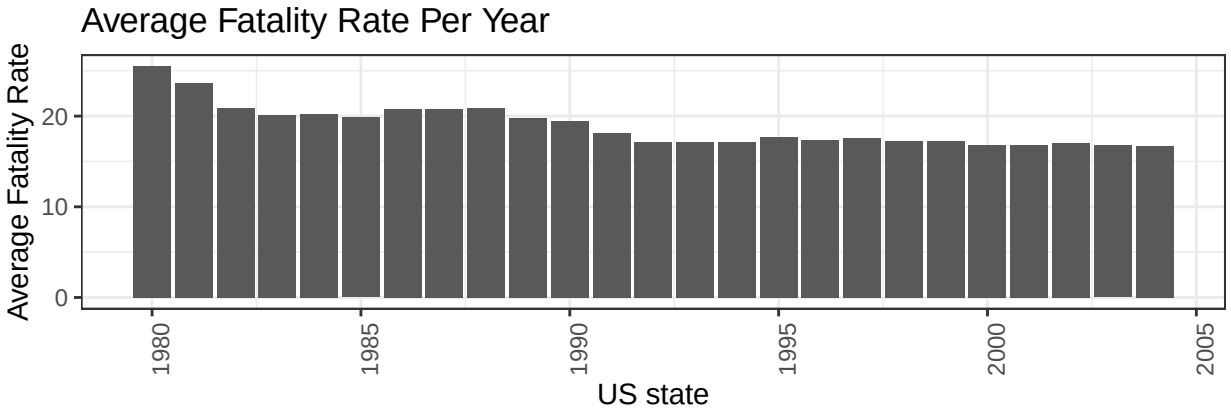
```r
# chart without and with transformations
plot_wo_transforms <- df %>%
  dplyr::select(totfatrte, perc14_24, unem, vehicmilespc) %>%
  pivot_longer(cols = c("totfatrte","perc14_24", "unem", "vehicmilespc"),
               names_to = "key", values_to = "values") %>%
  #necessary to keep order of items
  mutate(key = factor(key, levels = c("totfatrte","perc14_24", "unem", "vehicmilespc"))) %>%
  ggplot(aes(x = values)) +  geom_histogram()  +
  ggtitle('Variables without transforms') +
  facet_wrap(key ~ ., scales = "free_x", ncol = 4)

# identify whether Box-Cox or log transform are appropriate
apply(df[, (names(df) %in% c("totfatrte","perc14_24", "unem",
                             "vehicmilespc"))], 2, BoxCoxTrans)
```

```
## $totfatrte
## Box-Cox Transformation
##
## 1200 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     6.2    14.4    18.4    18.9    22.8    53.3
##
## Largest/Smallest: 8.6
## Sample Skewness: 0.788
##
## Estimated Lambda: 0.3
##
##
## $perc14_24
## Box-Cox Transformation
##
## 1200 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    11.7    13.9    14.9    15.3    16.6    20.3
##
## Largest/Smallest: 1.74
## Sample Skewness: 0.517
##
## Estimated Lambda: -1.1
##
##
## $unem
## Box-Cox Transformation
```

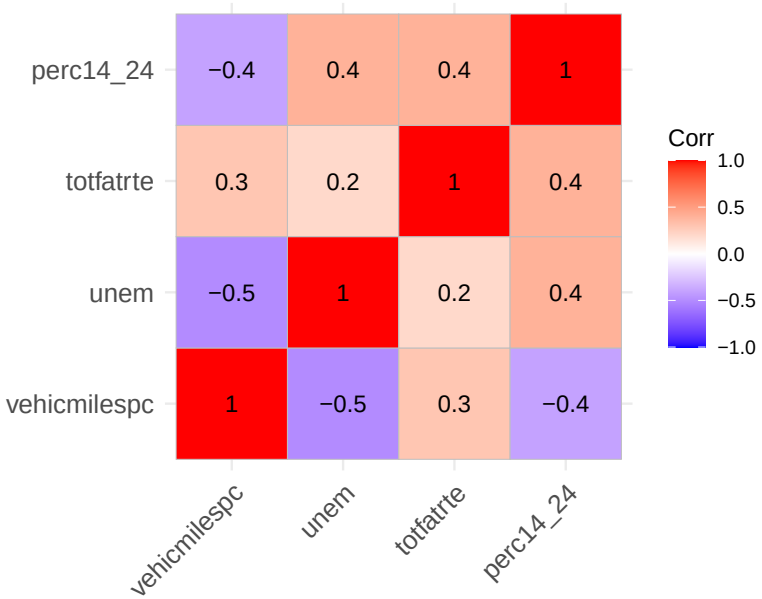
```
##
## 1200 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.20    4.50    5.60    5.95    7.00   18.00
##
## Largest/Smallest: 8.18
## Sample Skewness: 1.17
##
## Estimated Lambda: -0.1
## With fudge factor, Lambda = 0 will be used for transformations
##
##
## $vehicmilespc
## Box-Cox Transformation
##
## 1200 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4372    7788    9013    9129   10327   18390
##
## Largest/Smallest: 4.21
## Sample Skewness: 0.724
##
## Estimated Lambda: 0.1
## With fudge factor, Lambda = 0 will be used for transformations
```

```
g1 <- df %>%
        group_by(us_state) %>%
        summarise(totfatrte_avg = mean(totfatrte)) %>%
        ggplot() + geom_bar(aes(x=reorder(us_state,-totfatrte_avg), y=totfatrte_avg),
                            stat='identity') + theme_bw() +
        theme(axis.text.x = element_text(angle=90))  +
        ggtitle('Average Fatality Rate Per State') +
  labs(x='US state',y='Average Fatality Rate')

g2 <- df %>%
        group_by(year) %>%
        summarise(totfatrte_avg = mean(totfatrte)) %>%
        ggplot() + geom_bar(aes(x=year, y=totfatrte_avg), stat='identity') +
        theme_bw() + theme(axis.text.x = element_text(angle=90)) +
        ggtitle('Average Fatality Rate Per Year') +
  labs(x='US state',y='Average Fatality Rate')

grid.arrange(g1,g2, ncol=1)
```






```
#CORRELATION PLOT
temp_df <- df %>% dplyr::select(all_of(numeric_columns)) %>% as_tibble()
corr <- round(cor(temp_df),1)
ggcorrplot(corr, hc.order = TRUE, #type = "lower",
   lab = TRUE)
```



```r
plot_first_time <- function (df_plot, title) {
  g1 <- df_plot%>% mutate(dummy=1) %>%
  ggplot(aes(year,dummy)) + theme_bw() +
  geom_jitter(position = position_jitter(seed=1)) +
  ggtitle(title) +
  geom_text(position=position_jitter(seed=1), hjust=-0.1,
            aes(label=ifelse(year>1990,as.character(us_state),''))) +
  theme(axis.title.y=element_blank(), axis.text.y=element_blank(),
        axis.ticks.y=element_blank())
  return(g1)
}

perse_per_state <- df %>%
        group_by(us_state) %>%
        summarise(min_perse = min(as.numeric(levels(perse)[perse])),
                  min_bac_08 = min(as.numeric(levels(bac08)[bac08])),
                  min_bac_10 = min(as.numeric(levels(bac10)[bac10])),
                  max_perse = max(as.numeric(levels(perse)[perse])),
                  max_bac_08 = max(as.numeric(levels(bac08)[bac08])),
                  max_bac_10 = max(as.numeric(levels(bac10)[bac10])),
                  min_sbprim = min(as.numeric(levels(sbprim)[sbprim])),
                  min_sbsecon = min(as.numeric(levels(sbsecon)[sbsecon])),
                  max_sbprim = max(as.numeric(levels(sbprim)[sbprim])),
                  max_sbsecon = max(as.numeric(levels(sbsecon)[sbsecon])))#,
                  # range_perse = max(as.numeric(perse))-min(as.numeric(perse)),
                  # range_bac_08 = max(as.numeric(bac08))-min(as.numeric(bac08)),
                  # range_bac_10 = max(as.numeric(bac10))-min(as.numeric(bac10)))

states_with_no_bac <- perse_per_state %>% filter(max_bac_08<1 & max_bac_10<1)
states_with_bac08 <- perse_per_state %>% filter(max_bac_08==1)
states_with_no_bac08 <- perse_per_state %>% filter(max_bac_08==0)
states_with_no_perse <- perse_per_state %>% filter(max_perse<1)

print(paste('The number of states that had no bac limits enforced (1980-2004):',
            nrow(states_with_no_bac)))
```

```
## [1] "The number of states that had no bac limits enforced (1980-2004): 0"
```

```r
print(paste('The number of states that had bac 08 limits enforced (1980-2004):',
            nrow(states_with_bac08)))
```

```
## [1] "The number of states that had bac 08 limits enforced (1980-2004): 45"
```

```r
print(paste('The number of states that had no bac 08 limits enforced (1980-2004):',
            nrow(states_with_no_bac08)))
```

```
## [1] "The number of states that had no bac 08 limits enforced (1980-2004): 3"
```

```
print(paste('The number of states that had no perse laws (1980-2004):',
            nrow(states_with_no_perse)))
```

```
## [1] "The number of states that had no perse laws (1980-2004): 9"
```

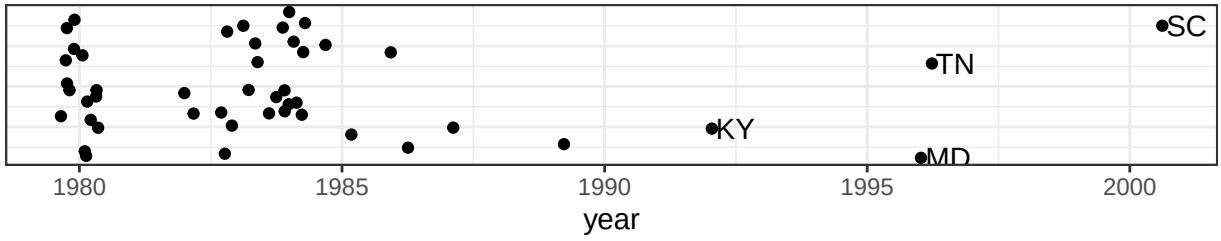
```
bac10_law_introduction <- df %>%
        group_by(us_state) %>% filter(bac10==1) %>% filter(row_number()==1) %>%
        arrange(year)
bac10_law_introduction_rev <- df %>%
        group_by(us_state) %>% filter(bac10==0) %>% filter(row_number()==1) %>%
        arrange(year, descending=TRUE)
bac08_law_introduction <- df %>%
        group_by(us_state) %>% filter(bac08==1) %>% filter(row_number()==1) %>%
        arrange(year, descending=FALSE)
perse_law_introduction <- df %>%
        group_by(us_state) %>% filter(perse==1) %>% filter(row_number()==1) %>%
        arrange(year)
sbprim_law_introduction <- df %>%
        group_by(us_state) %>% filter(sbprim==1) %>% filter(row_number()==1) %>%
        arrange(year)
sbsecon_law_introduction <- df %>%
        group_by(us_state) %>% filter(sbsecon==1) %>% filter(row_number()==1) %>%
        arrange(year)

last_state_to_bac10 <- tail(bac10_law_introduction,n=1)
print(paste('The last state to enact a bac10 law was: ',
            last_state_to_bac10$us_state, ' in the year: ',
            last_state_to_bac10$year), max.levels=0)
```

```
## [1] "The last state to enact a bac10 law was:  SC  in the year:  2001"
```

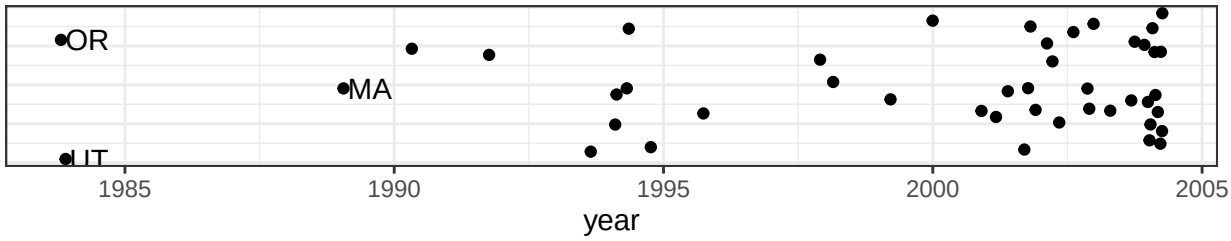
```
g1 <- bac10_law_introduction %>% mutate(dummy=1) %>%
  ggplot(aes(year,dummy)) + theme_bw() +
  geom_jitter(position = position_jitter(seed=1)) +
  ggtitle('Year of BAC10 introduction') +
  geom_text(position=position_jitter(seed=1), hjust=-0.1,
            aes(label=ifelse(year>1990,as.character(us_state),''))) +
  theme(axis.title.y=element_blank(), axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

g2 <- bac08_law_introduction %>% mutate(dummy=1) %>%
  ggplot(aes(year,dummy)) + theme_bw() +
  geom_jitter(position = position_jitter(seed=1)) +
  ggtitle('Year of BAC08 introduction') +
```

```
  geom_text(position=position_jitter(seed=1), hjust=-0.1,
            aes(label=ifelse(year<1990,as.character(us_state),''))) +
  theme(axis.title.y=element_blank(), axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

g3 <- perse_law_introduction %>% mutate(dummy=1) %>%
  ggplot(aes(year,dummy)) + theme_bw() +
  geom_jitter(position = position_jitter(seed=1)) +
  ggtitle('Year of PerSe introduction') +
  geom_text(position=position_jitter(seed=1), hjust=-0.1,
            aes(label=ifelse(year>1990,as.character(us_state),''))) +
  theme(axis.title.y=element_blank(), axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

g4 <- sbprim_law_introduction %>% mutate(dummy=1) %>%
  ggplot(aes(year,dummy)) + theme_bw() +
  geom_jitter(position = position_jitter(seed=1)) +
  ggtitle('Year of sbprim introduction') +
  geom_text(position=position_jitter(seed=1), hjust=-0.1,
            aes(label=ifelse(year>1990,as.character(us_state),''))) +
  theme(axis.title.y=element_blank(), axis.text.y=element_blank(),
        axis.ticks.y=element_blank())
g5 <- plot_first_time(sbsecon_law_introduction, 'Year of sbsecon introduction')
grid.arrange(g1,g2,g3,g4,g5,ncol=1)
```

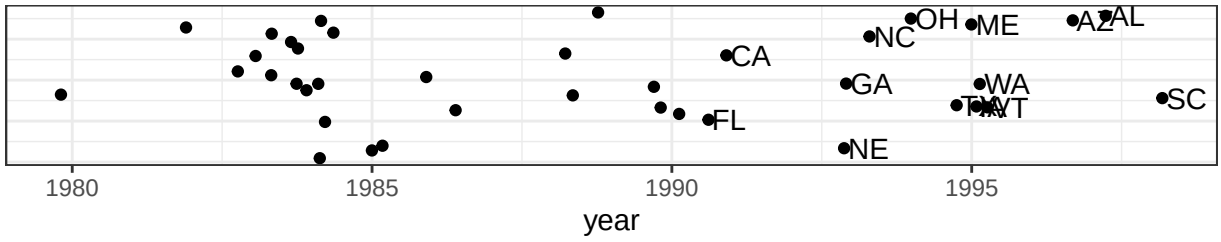
Year of BAC10 introduction



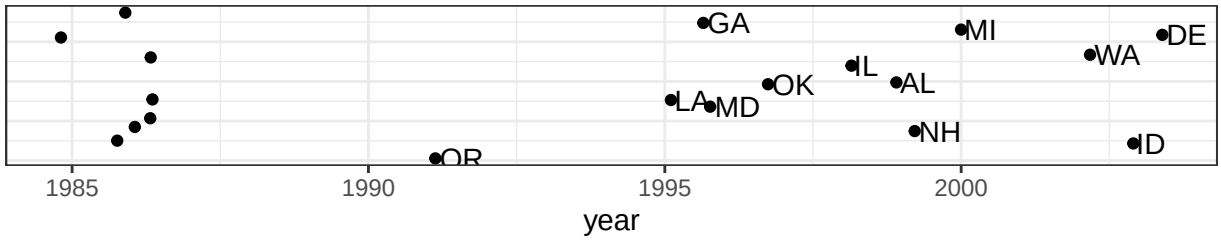
Year of BAC08 introduction



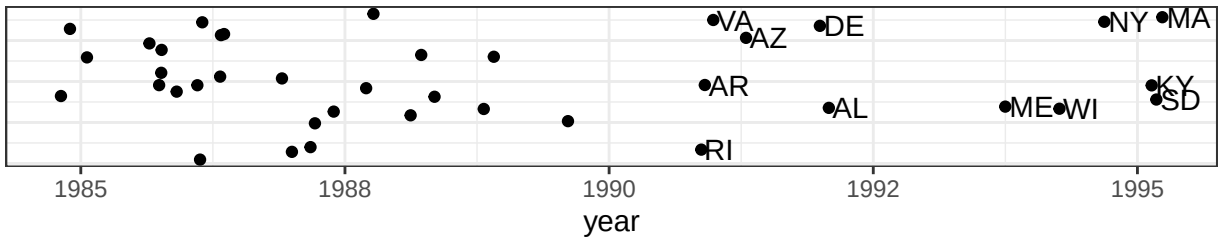
Year of PerSe introduction



Year of sbprim introduction

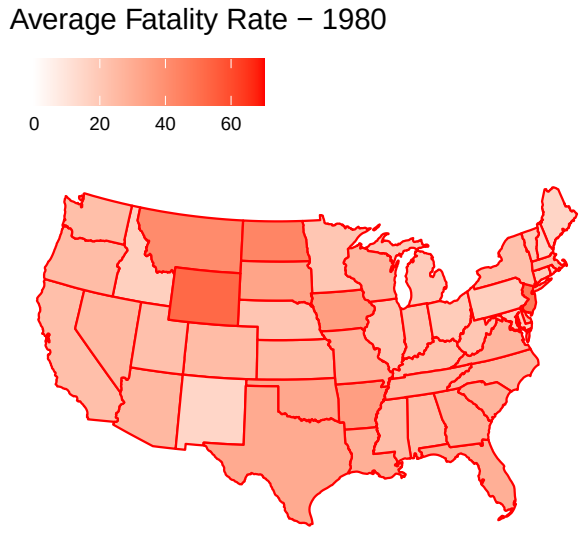


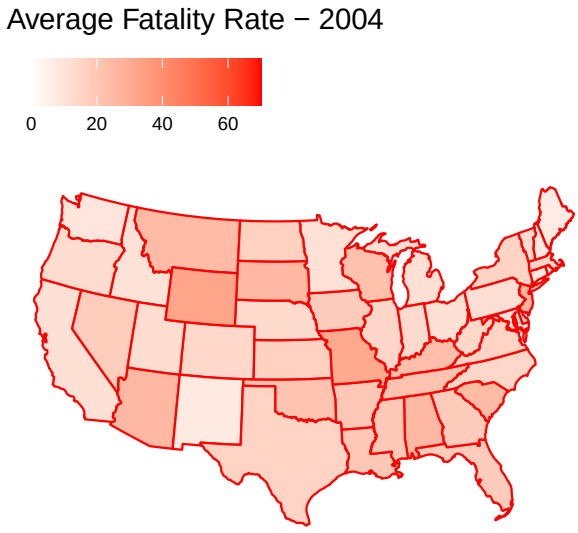
Year of sbsecon introduction



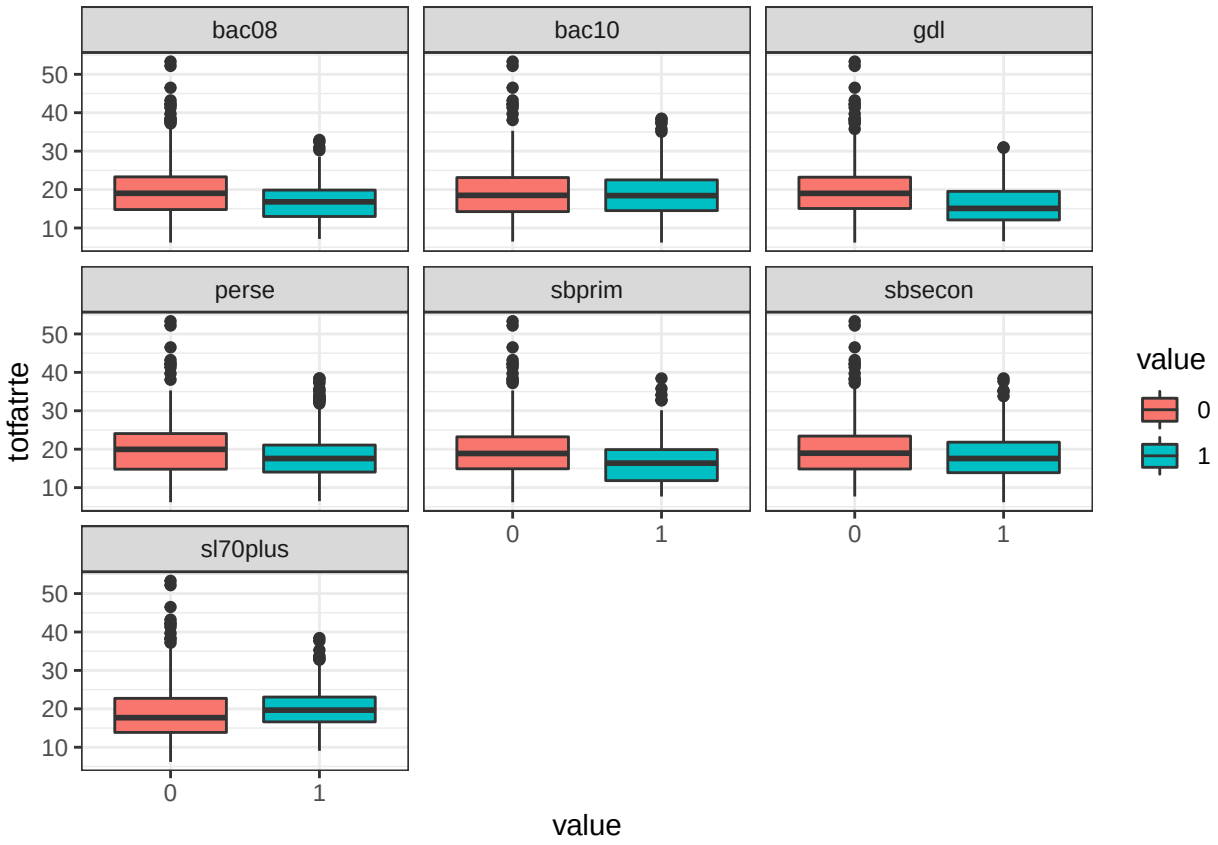
Appendix C: Code for this report

```
knitr::opts_chunk$set(cache = TRUE)
us_states <- data.frame(state = unique(df$state), us_state = c('AL','AR','AZ','CA','CO','CT',
               'ME','MI','MN','MO','MS','MT','NC','ND','NE',
               'NH','NJ','NM','NV','NY','OH','OK','OR','PA',
               'RI','SC','SD','TN','TX','UT','VA','VT','WA',
               'WI','WV','WY'))

#map of US with average fatality rates at t0 and tn

# average by year
t0_fatrte <- df%>%
        dplyr::select(us_state, totfatrte, year) %>%
        group_by(us_state) %>%
        rename(state = us_state) %>%
        filter(year == min(year)) %>%
        summarise(totfatrte_avg = mean(totfatrte))

tn_fatrte <- df%>%
        dplyr::select(us_state, totfatrte, year) %>%
        group_by(us_state) %>%
        rename(state = us_state) %>%
        filter(year == max(year)) %>%
        summarise(totfatrte_avg = mean(totfatrte))

#maps
map_t0 <- plot_usmap(data = t0_fatrte,
                     exclude = c("HI", "AK"), values =  "totfatrte_avg",
                     color = "red") +
  labs(title = "Average Fatality Rate - 1980") +
  scale_fill_continuous(low = "white", high = "red", limits =c(0,70)) +
              theme(legend.position = "top",
                    legend.title=element_blank())

map_tn <- plot_usmap(data = tn_fatrte, exclude = c("HI", "AK"),
            values = "totfatrte_avg", color = "red") +
            labs(title = "Average Fatality Rate - 2004") +
            scale_fill_continuous(low = "white",
                                  high = "red",
                                  limits =c(0,70)) +
              theme(legend.position = "top",
                    legend.title=element_blank())
grid.arrange(map_t0, map_tn, ncol = 2)
```

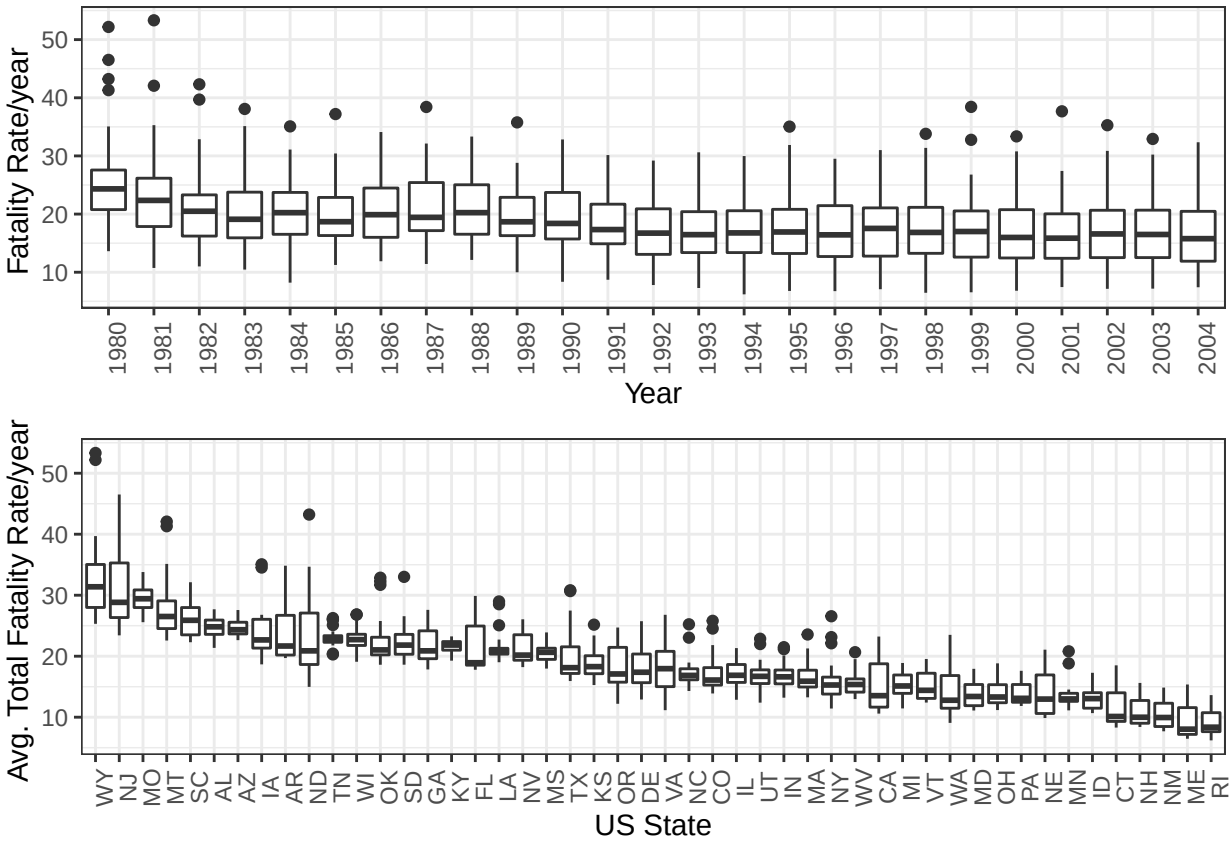





```
df %>% dplyr::select(all_of(c(factor_columns, c('totfatrte'))), -c('state','us_state')) %>%
  pivot_longer(cols = c("sbprim","sbsecon", "bac08","bac10","perse","sl70plus",
                        "gdl"), names_to = "key", values_to = "value")%>%
   ggplot(aes(x = value, y =totfatrte)) +
   geom_boxplot(aes(fill = value)) +
   facet_wrap(.~key)
```

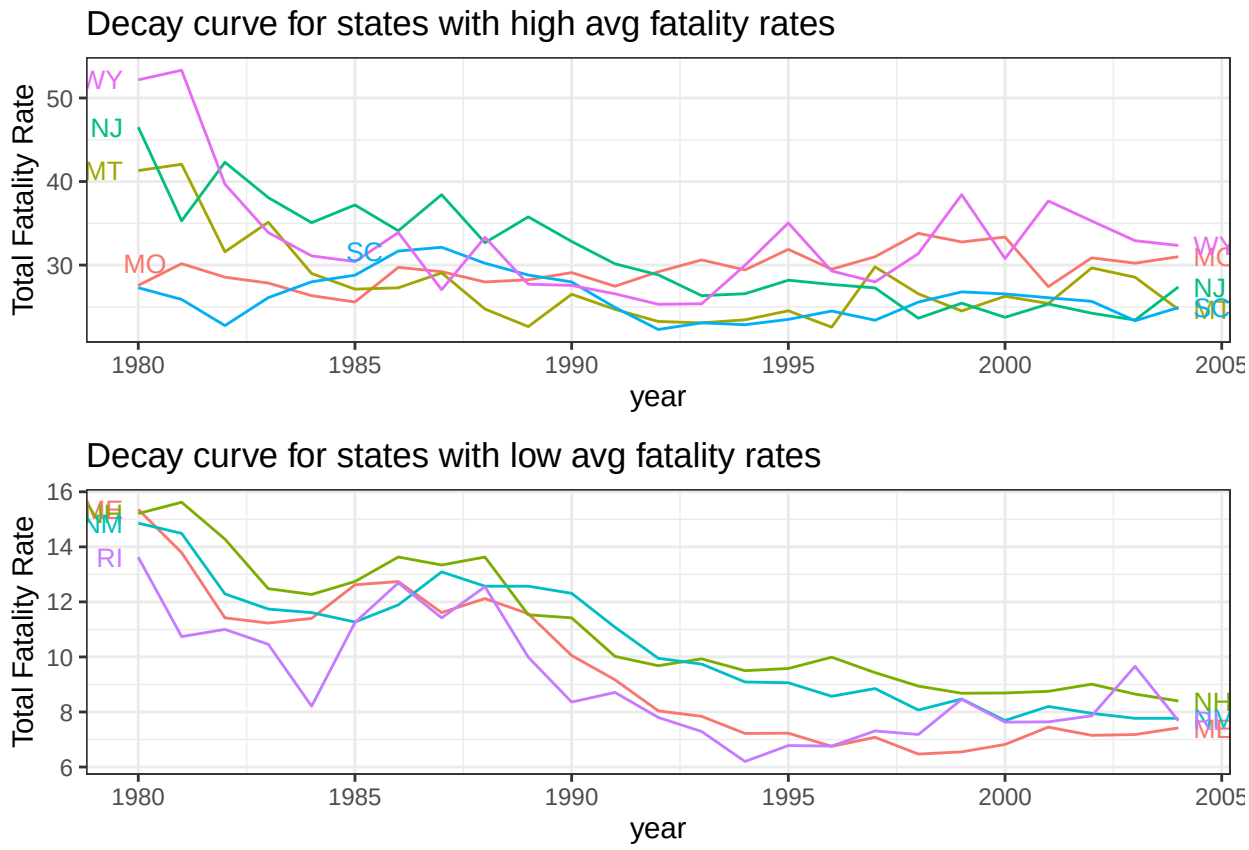



```
g2 <- ggplot(df, aes(x=reorder(us_state,-totfatrte), y=totfatrte)) +
    geom_boxplot() + theme_bw() + theme(axis.text.x = element_text(angle=90)) +
    scale_y_continuous(name = "Avg. Total Fatality Rate/year ") + labs(x='US State')

g3 <- ggplot(df, aes(x=factor(year), y=totfatrte)) +
    geom_boxplot() + theme_bw() + theme(axis.text.x = element_text(angle=90)) +
    scale_y_continuous(name = "Fatality Rate/year ") + labs(x='Year')
grid.arrange(g3,g2,ncol=1)
```




```
# Select States
scale <- 1

state_columns_hfr <- c("WY", "NJ", "MO", "MT", "SC")
state_columns_lfr <- c("NH", "NM", "ME", "RI")

df_hfr <- df %>% filter(us_state %in% state_columns_hfr)
df_lfr <- df %>% filter(us_state %in% state_columns_lfr)

g1 <- ggplot(df_hfr,aes(x=year, y=totfatrte, color=us_state)) +
    geom_line() + theme_bw() +
    scale_y_continuous(name = "Total Fatality Rate") +
    scale_colour_discrete(guide = 'none') +
    # scale_x_discrete(expand=c(0, 1)) +
    geom_dl(data = subset(df_hfr, totfatrte>30),
            aes(label=us_state), method = list(dl.trans(x = x - 0.2),
                                               "first.points", cex = 0.8)) +
    geom_dl(data = subset(df_hfr, totfatrte<40),
            aes(label=us_state), method = list(dl.trans(x = x + 0.2),
                                               "last.points", cex = 0.8)) +
    ggtitle('Decay curve for states with high avg fatality rates')
g2 <- ggplot(df_lfr,aes(x=year, y=totfatrte, color=us_state)) +
    geom_line() + theme_bw() +
```

```
    scale_y_continuous(name = "Total Fatality Rate") +
    scale_colour_discrete(guide = 'none') +
    # scale_x_discrete(expand=c(0, 1)) +
    geom_dl(data = subset(df_lfr, totfatrte>10),
            aes(label=us_state), method = list(dl.trans(x = x - 0.2),
                                               "first.points", cex = 0.8)) +
    geom_dl(data = subset(df_lfr, totfatrte<10),
            aes(label=us_state), method = list(dl.trans(x = x + 0.2),
                                               "last.points", cex = 0.8)) +
    ggtitle('Decay curve for states with low avg fatality rates')
grid.arrange(g1,g2,ncol=1)
```




```
plot_first_time <- function (df_plot, title) {
  g1 <- df_plot%>% mutate(dummy=1) %>%
  ggplot(aes(year,dummy)) + theme_bw() +
  geom_jitter(position = position_jitter(seed=1)) +
  ggtitle(title) +
  geom_text(position=position_jitter(seed=1), hjust=-0.1,
            aes(label=ifelse(year>1990,as.character(us_state),''))) +
  theme(axis.title.y=element_blank(), axis.text.y=element_blank(),
        axis.ticks.y=element_blank())
  return(g1)
```

```
}

perse_per_state <- df %>%
        group_by(us_state) %>%
        summarise(min_perse = min(as.numeric(levels(perse)[perse])),
                  min_bac_08 = min(as.numeric(levels(bac08)[bac08])),
                  min_bac_10 = min(as.numeric(levels(bac10)[bac10])),
                  max_perse = max(as.numeric(levels(perse)[perse])),
                  max_bac_08 = max(as.numeric(levels(bac08)[bac08])),
                  max_bac_10 = max(as.numeric(levels(bac10)[bac10])),
                  min_sbprim = min(as.numeric(levels(sbprim)[sbprim])),
                  min_sbsecon = min(as.numeric(levels(sbsecon)[sbsecon])),
                  max_sbprim = max(as.numeric(levels(sbprim)[sbprim])),
                  max_sbsecon = max(as.numeric(levels(sbsecon)[sbsecon])))

states_with_no_bac <- perse_per_state %>% filter(max_bac_08<1 & max_bac_10<1)
states_with_bac08 <- perse_per_state %>% filter(max_bac_08==1)
states_with_no_bac08 <- perse_per_state %>% filter(max_bac_08==0)
states_with_no_perse <- perse_per_state %>% filter(max_perse<1)

bac10_law_introduction <- df %>%
        group_by(us_state) %>% filter(bac10==1) %>% filter(row_number()==1) %>%
        arrange(year)
bac10_law_introduction_rev <- df %>%
        group_by(us_state) %>% filter(bac10==0) %>% filter(row_number()==1) %>%
        arrange(year, descending=TRUE)
bac08_law_introduction <- df %>%
        group_by(us_state) %>% filter(bac08==1) %>% filter(row_number()==1) %>%
        arrange(year, descending=FALSE)
perse_law_introduction <- df %>%
        group_by(us_state) %>% filter(perse==1) %>% filter(row_number()==1) %>%
        arrange(year)
sbprim_law_introduction <- df %>%
        group_by(us_state) %>% filter(sbprim==1) %>% filter(row_number()==1) %>%
        arrange(year)
sbsecon_law_introduction <- df %>%
        group_by(us_state) %>% filter(sbsecon==1) %>% filter(row_number()==1) %>%
        arrange(year)

last_state_to_bac10 <- tail(bac10_law_introduction,n=1)

# plot how state laws evolved over time (adoption counts)
perse <- df %>% group_by(perse) %>%
  count(year) %>% filter(perse == 1) %>%
  as_tsibble(index = year) %>% autoplot() +
   labs(y = "Count of States", x = "Year",
       title = "States with Per Se \nLaws Over Time")
```

```
## Plot variable not specified, automatically selected `.vars = n`
```

```
sbprim <- df %>% group_by(sbprim) %>%
  count(year) %>% filter(sbprim == 1) %>%
  as_tsibble(index = year) %>% autoplot() +
   labs(y = "Count of States", x = "Year",
       title = "States with Primary \nSeatbelt Laws Over Time")
```

```
## Plot variable not specified, automatically selected `.vars = n`
```

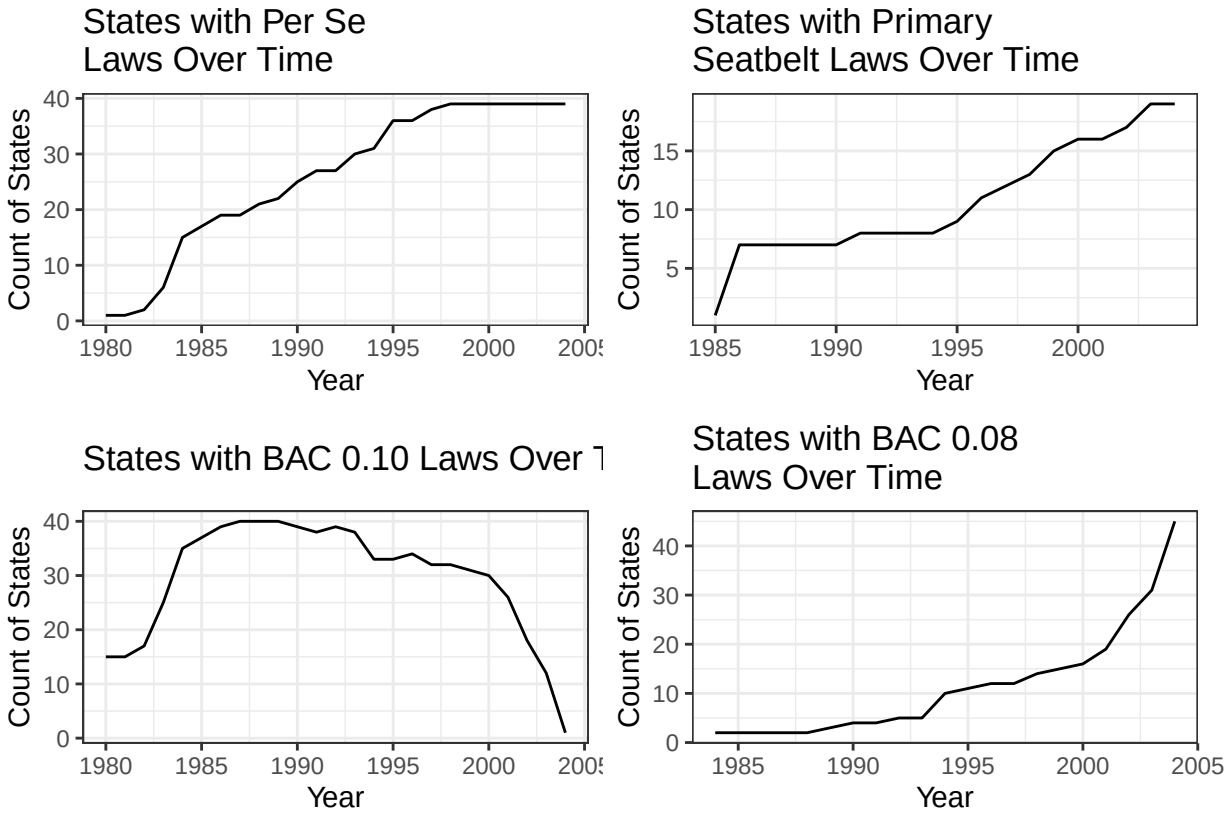
```
bac_10 <- df %>% group_by(bac10) %>%
  count(year) %>% filter(bac10 == 1) %>%
  as_tsibble(index = year) %>% autoplot() +
   labs(y = "Count of States", x = "Year",
       title = "States with BAC 0.10 Laws Over Time")
```

```
## Plot variable not specified, automatically selected `.vars = n`
```

```
bac_08 <- df %>% group_by(bac08) %>%
  count(year) %>% filter(bac08 == 1) %>%
  as_tsibble(index = year) %>% autoplot() +
   labs(y = "Count of States", x = "Year",
       title = "States with BAC 0.08 \nLaws Over Time")
```

```
## Plot variable not specified, automatically selected `.vars = n`
```

```
(perse +sbprim) / (bac_10 + bac_08)
```




```
# Compute average fatality by year

avg_totfatrte <- df %>% group_by(year) %>% summarise(Avg = round(mean(totfatrte),2))

knitr::kable(cbind(
  avg_totfatrte[1:5,],
  avg_totfatrte[6:10,],
  avg_totfatrte[11:15,],
  avg_totfatrte[16:20,],
  avg_totfatrte[21:25,]),col.names = NA,
caption = "Average Total Fatality Rate", "pipe")
```

Table 3: Average Total Fatality Rate

year	Avg	year	Avg	year	Avg	year	Avg	year	Avg
1980	25.49	1985	19.85	1990	19.51	1995	17.67	2000	16.83
1981	23.67	1986	20.80	1991	18.09	1996	17.37	2001	16.79
1982	20.94	1987	20.77	1992	17.16	1997	17.61	2002	17.03
1983	20.15	1988	20.89	1993	17.13	1998	17.27	2003	16.76
1984	20.27	1989	19.77	1994	17.16	1999	17.25	2004	16.73

Appendix D: Pooled OLS and Linear Regression

```
stargazer(lm2, mod_fe, mod_re,
          se.list = list(robust_se(lm2),
                         robust_se(mod_fe),
                         robust_se(mod_re)),
          notes = "Robust Standard Errors",
          column.labels = c("OLS", "Fixed Effects", "Random Effects"),
          type = "text",
          single.row = TRUE)
```

```
##
## ================================================================================
##                                          Dependent variable:
##                     ---------------------------------------------------------
##                                              totfatrte
##                                OLS                            panel
##                                                               linear
##                                OLS                Fixed Effects         Random Effects
##                                (1)                     (2)                   (3)
## --------------------------------------------------------------------------------
## factor(year)1981        -2.107** (0.823)        -1.579*** (0.414)     -1.611*** (0.430)
## factor(year)1982        -6.304*** (0.840)       -3.372*** (0.434)     -3.567*** (0.451)
## factor(year)1983        -7.190*** (0.851)       -4.025*** (0.445)     -4.243*** (0.462)
## factor(year)1984        -5.826*** (0.867)       -4.547*** (0.460)     -4.653*** (0.477)
## factor(year)1985        -6.458*** (0.885)       -4.996*** (0.481)     -5.126*** (0.498)
## factor(year)1986        -5.634*** (0.923)       -3.986*** (0.515)     -4.139*** (0.533)
## factor(year)1987        -6.065*** (0.961)       -4.670*** (0.555)     -4.842*** (0.573)
## factor(year)1988        -6.176*** (1.011)       -5.210*** (0.605)     -5.387*** (0.624)
## factor(year)1989        -7.688*** (1.049)       -6.524*** (0.643)     -6.733*** (0.663)
## factor(year)1990        -8.682*** (1.072)       -6.581*** (0.666)     -6.868*** (0.686)
## factor(year)1991       -10.870*** (1.093)       -7.251*** (0.680)     -7.634*** (0.701)
## factor(year)1992       -12.630*** (1.114)       -8.128*** (0.701)     -8.591*** (0.722)
## factor(year)1993       -12.500*** (1.128)       -8.468*** (0.715)     -8.911*** (0.735)
## factor(year)1994       -12.070*** (1.150)       -8.944*** (0.734)     -9.343*** (0.755)
## factor(year)1995       -11.470*** (1.180)       -8.709*** (0.759)     -9.102*** (0.780)
## factor(year)1996       -13.400*** (1.223)       -9.128*** (0.801)     -9.586*** (0.823)
## factor(year)1997       -13.520*** (1.244)       -9.388*** (0.821)     -9.849*** (0.844)
## factor(year)1998       -14.200*** (1.268)      -10.100*** (0.842)    -10.580*** (0.865)
## factor(year)1999       -14.150*** (1.284)      -10.350*** (0.852)    -10.820*** (0.875)
## factor(year)2000       -14.400*** (1.307)      -10.960*** (0.866)    -11.420*** (0.889)
## factor(year)2001       -15.670*** (1.317)      -10.460*** (0.867)    -11.060*** (0.890)
## factor(year)2002       -16.490*** (1.326)       -9.602*** (0.870)    -10.330*** (0.893)
## factor(year)2003       -16.920*** (1.331)       -9.641*** (0.873)    -10.400*** (0.895)
## factor(year)2004       -16.330*** (1.367)      -10.080*** (0.899)    -10.780*** (0.922)
## bac081                  -2.288*** (0.486)       -1.105*** (0.331)     -1.208*** (0.341)
```

```
## bac101                  -1.256*** (0.359)        -0.804*** (0.226)     -0.867*** (0.233)
## perse1                   -0.562* (0.292)         -1.127*** (0.223)     -1.074*** (0.229)
## sbprim1                  -0.380 (0.490)          -1.189*** (0.343)     -1.140*** (0.353)
## sbsecon1                 -0.153 (0.428)           -0.304 (0.252)        -0.305 (0.261)
## sl70plus1               3.112*** (0.433)          0.047 (0.261)         0.127 (0.270)
## gdl1                     -0.301 (0.507)          -0.283 (0.280)        -0.259 (0.291)
## perc14_24                 0.178 (0.122)          0.167* (0.095)        0.182* (0.098)
## unem_log                5.152*** (0.481)         -3.709*** (0.392)     -3.127*** (0.401)
## vehicmilespc            0.003*** (0.0001)        0.001*** (0.0001)     0.001*** (0.0001)
## Constant                -8.012*** (2.620)                              19.770*** (2.261)
## ------------------------------------------------------------------------------------------
## Observations                  1,200                    1,200                 1,200
## R2                            0.612                    0.625                 0.601
## Adjusted R2                   0.601                    0.598                 0.590
## Residual Std. Error     4.024 (df = 1165)
## F Statistic         54.020*** (df = 34; 1165) 54.760*** (df = 34; 1118)    1,757.000***
## ==========================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
##                                                                     Robust Standard Errors
##
## ==========================================================================================
## (Intercept) factor(year)1981 factor(year)1982 factor(year)1983 factor(year)1984 factor(year)
## ------------------------------------------------------------------------------------------
## 2.864            1.324            1.217            1.138            1.119            1.138
## ------------------------------------------------------------------------------------------
## Robust Standard Errors
##
## ==========================================================================================
## factor(year)1981 factor(year)1982 factor(year)1983 factor(year)1984 factor(year)1985 factor
## ------------------------------------------------------------------------------------------
## 0.442                0.444            0.462            0.460            0.481            0
## ------------------------------------------------------------------------------------------
## Robust Standard Errors
##
## ==========================================================================================
## (Intercept) factor(year)1981 factor(year)1982 factor(year)1983 factor(year)1984 factor(year)
## ------------------------------------------------------------------------------------------
## 4.517            0.440            0.439            0.459            0.481            0.502
## ------------------------------------------------------------------------------------------
## Robust Standard Errors
```