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

Due Monday 6 December 2021 11:59pm

Robert Hosbach, Charlie Boatwright, Wanyu Li

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.

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 data
load("./driving.RData")
# Look at first and last rows of data
head(data)
tail(data)
# Look at structure of data
str(data)
# Summarize columns
summary(data)
```

After loading the data, we looked at the first and last rows, the structure of the data frame, as well as a summary of every column. To save space, these items were omitted from the report. However, below we provide a high-level summary of important columns for this analysis, a check for missing values, as well as the panels summary below.

```r
# Summarize numeric columns
data %>%
  select(year, totfatrte, unem, perc14_24) %>%
  summary()
```

```
##       year        totfatrte          unem          perc14_24
##  Min.   :1980   Min.   : 6.20   Min.   : 2.200   Min.   :11.70
##  1st Qu.:1986   1st Qu.:14.38   1st Qu.: 4.500   1st Qu.:13.90
##  Median :1992   Median :18.43   Median : 5.600   Median :14.90
##  Mean   :1992   Mean   :18.92   Mean   : 5.951   Mean   :15.33
##  3rd Qu.:1998   3rd Qu.:22.77   3rd Qu.: 7.000   3rd Qu.:16.60
##  Max.   :2004   Max.   :53.32   Max.   :18.000   Max.   :20.30
```

```r
# Summarize key categorical columns
data %>%
  select(sbprim, sbsecon, gdl, bac10, bac08, perse, sl70plus) %>%
  describe()
```

```
## .
##
##  7  Variables      1200  Observations
## --------------------------------------------------------------------------------
## sbprim
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1200        0        2    0.441      215   0.1792   0.2944
##
## --------------------------------------------------------------------------------
## sbsecon
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1200        0        2    0.747      562   0.4683   0.4984
##
## --------------------------------------------------------------------------------
## gdl
##        n  missing distinct     Info     Mean      Gmd
##     1200        0        8    0.449   0.1741   0.2877
##
## lowest : 0.000 0.167 0.250 0.500 0.670, highest: 0.500 0.670 0.750 0.833 1.000
##
## Value      0.000 0.167 0.250 0.500 0.670 0.750 0.833 1.000
## Frequency    981     1     2    14     1     1     1   199
## Proportion 0.818 0.001 0.002 0.012 0.001 0.001 0.001 0.166
## --------------------------------------------------------------------------------
## bac10
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0       10    0.748   0.6231   0.4691        0        0
##      .25      .50      .75      .90      .95
##        0        1        1        1        1
##
```

```
## lowest : 0.000 0.250 0.333 0.417 0.500, highest: 0.583 0.667 0.750 0.833 1.000
##
## Value      0.000 0.250 0.333 0.417 0.500 0.583 0.667 0.750 0.833 1.000
## Frequency    424     4     4     1    28     4     8    13     3   711
## Proportion 0.353 0.003 0.003 0.001 0.023 0.003 0.007 0.011 0.002 0.593
## ---------------------------------------------------------------------------
## bac08
##        n  missing distinct     Info     Mean      Gmd
##     1200        0        8     0.54   0.2135   0.3358
##
## lowest : 0.000 0.250 0.333 0.417 0.500, highest: 0.417 0.500 0.667 0.750 1.000
##
## Value      0.000 0.250 0.333 0.417 0.500 0.667 0.750 1.000
## Frequency    921     9     5     4    19     1     2   239
## Proportion 0.767 0.007 0.004 0.003 0.016 0.001 0.002 0.199
## ---------------------------------------------------------------------------
## perse
##        n  missing distinct     Info     Mean      Gmd
##     1200        0        9     0.76   0.5471   0.4958
##
## lowest : 0.000 0.083 0.167 0.250 0.333, highest: 0.333 0.417 0.500 0.750 1.000
##
## Value      0.000 0.083 0.167 0.250 0.333 0.417 0.500 0.750 1.000
## Frequency    528     1     1     4     2     2    16     1   645
## Proportion 0.440 0.001 0.001 0.003 0.002 0.002 0.013 0.001 0.537
## ---------------------------------------------------------------------------
## sl70plus
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1200        0       15    0.515   0.2068   0.3283        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        1        1
##
## lowest : 0.000 0.042 0.083 0.333 0.375, highest: 0.750 0.792 0.833 0.984 1.000
##
## Value      0.000 0.042 0.083 0.333 0.375 0.417 0.500 0.583 0.625 0.667 0.750
## Frequency    938     1     5     1     1     3     3     2     1     3     3
## Proportion 0.782 0.001 0.004 0.001 0.001 0.002 0.002 0.002 0.001 0.002 0.002
##
## Value      0.792 0.833 0.984 1.000
## Frequency      2     2     1   234
## Proportion 0.002 0.002 0.001 0.195
## ---------------------------------------------------------------------------
```

```
# Check missing values
sapply(data, function(x) sum(is.na(x)))
```

```
##         year        state         sl55         sl65         sl70         sl75
##            0            0            0            0            0            0
```

```
##       slnone     seatbelt       minage      zerotol          gdl        bac10
##            0            0            0            0            0            0
##        bac08        perse       totfat      nghtfat      wkndfat    totfatpvm
##            0            0            0            0            0            0
##   nghtfatpvm   wkndfatpvm     statepop    totfatrte   nghtfatrte   wkndfatrte
##            0            0            0            0            0            0
##   vehicmiles         unem    perc14_24     sl70plus       sbprim      sbsecon
##            0            0            0            0            0            0
##          d80          d81          d82          d83          d84          d85
##            0            0            0            0            0            0
##          d86          d87          d88          d89          d90          d91
##            0            0            0            0            0            0
##          d92          d93          d94          d95          d96          d97
##            0            0            0            0            0            0
##          d98          d99          d00          d01          d02          d03
##            0            0            0            0            0            0
##          d04 vehicmilespc
##            0            0
```

```
# Check observations per year
table(data$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
```

```
# Check observations per state
table(data$state)
```

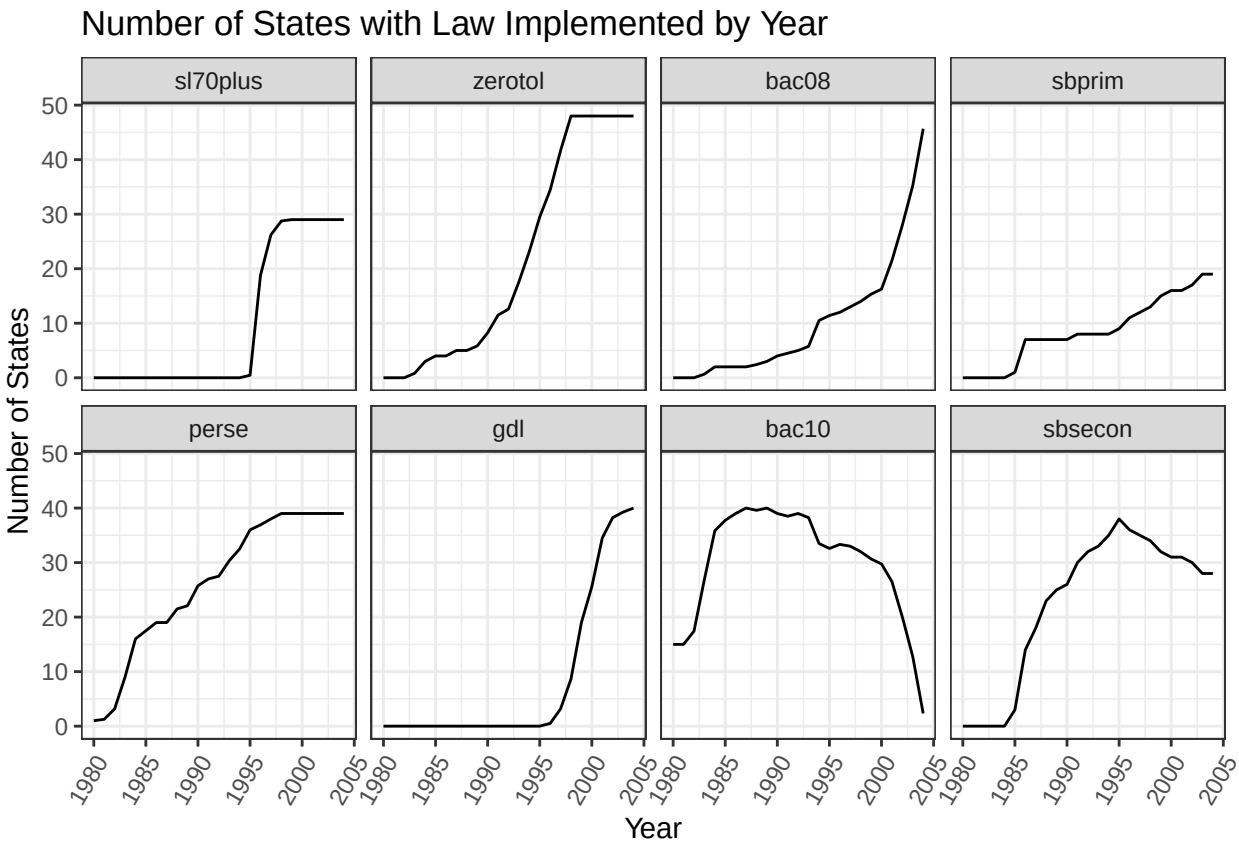
```
##
##  1  3  4  5  6  7  8 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
## 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

This data set has 1,200 observations of 56 variables, with no missing values. The data include 48 observations (corresponding to each state) for each year from 1980-2004, and 25 observations (corresponding each year from 1980-2004) for each state (numbered 1-51, excluding 2, 9, and 12). The data is arranged in order of increasing year by state (*e.g.*, the first 25 rows correspond to years 1980-2004 for state 1).

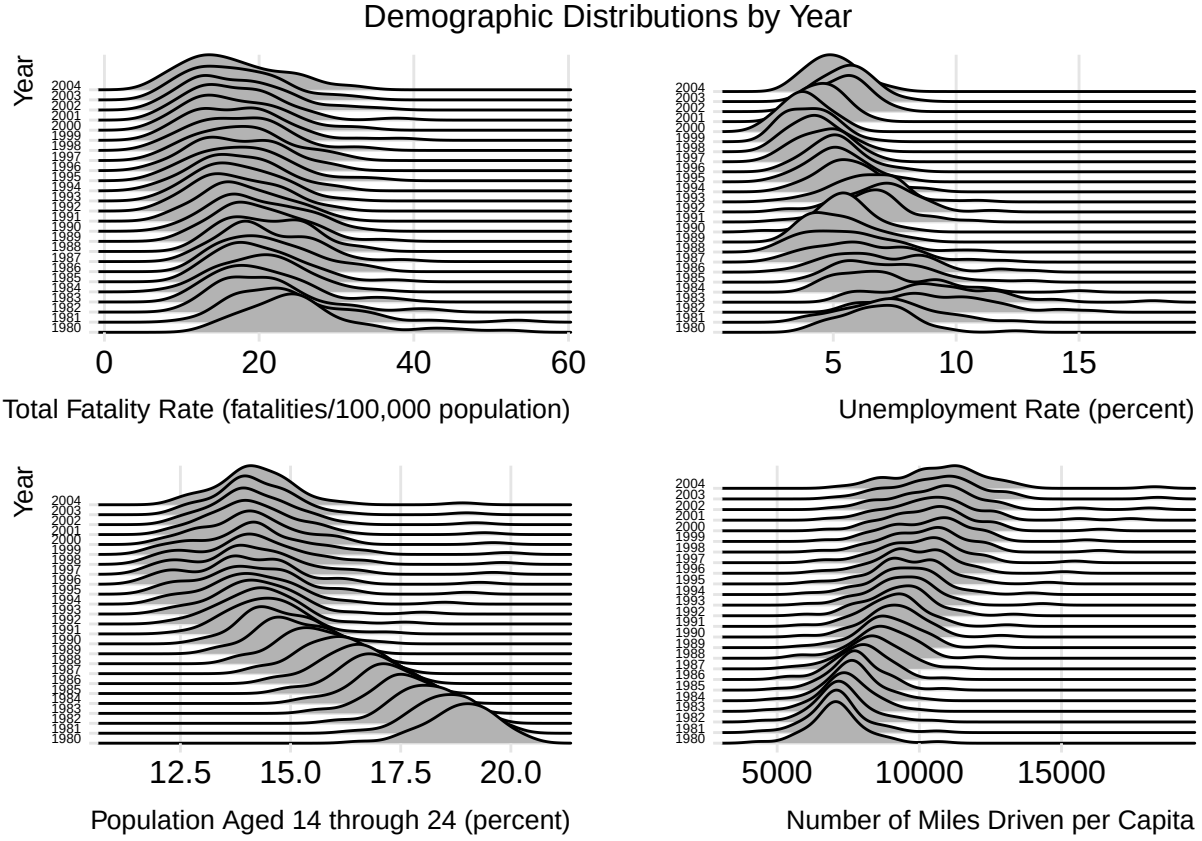
The following facet plot shows that there are many driving regulations implemented across the states over time. The facets show a variety of trends over time. For instance, while most of the states enacted the `sl70plus` law between 1995-1998, other laws were adopted by states at a more gradual rate. The number of states with `bac10` laws decreased after 1992 as states started to implement the stricter `bac08` law. We will use the `bac10` variable as it is throughout this study, arguing that the variables represent the law implementation effect (*i.e.* laws are implemented or not) rather than the behavior under the law as people may argue that driver under `bac08` laws

automatically follow the `bac10` laws hence the value for `bac10` should be 1 instead of 0 in states where the `bac08` law is implemented.

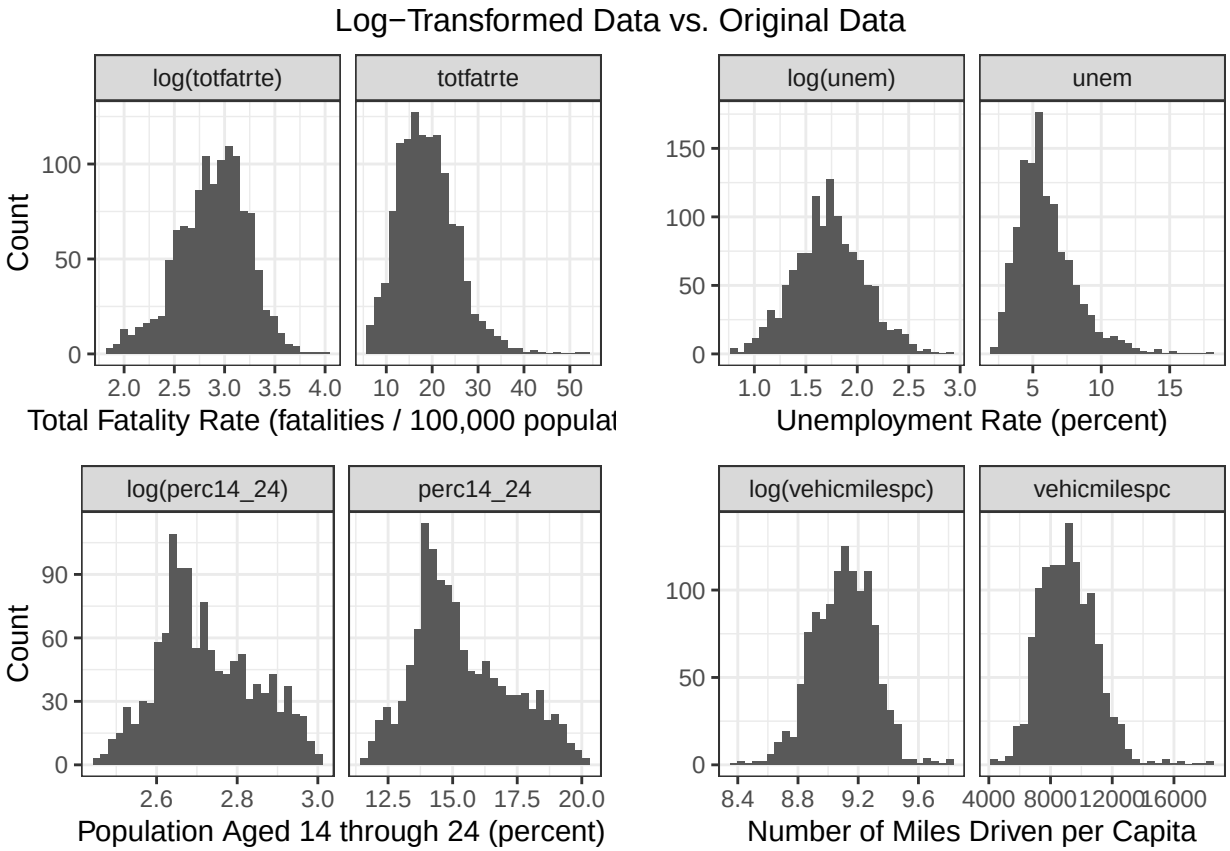
Another observation regarding the implementation of the laws is that they are highly correlated with time, once the law is implemented it is rarely removed. The observation of this relationship indicates that pooled OLS may not be a good solution for our study as serial correlation over time violates the assumptions in pooled OLS.



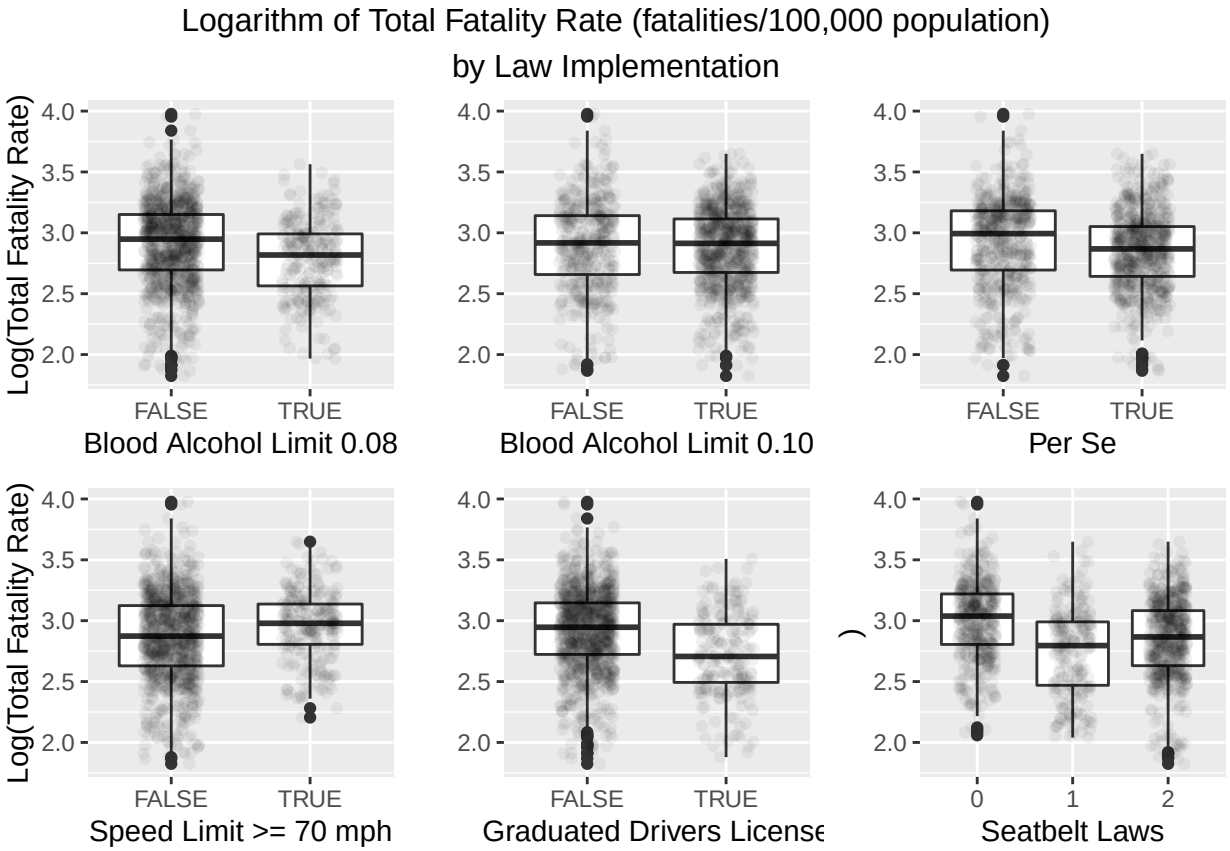
Below we show ridge plots characterizing how key continuous and demographic variables changed over time. In these plots, the annual distributions represent the range of values observed across the 48 contiguous states for that year. At a macro level, the stacked density plots of total fatality rate (`totfatrte`) show slightly decreasing fatality rates over time. Similarly, the percent of the population aged 14-24 (`perc14_24`) decreases rapidly until about 1997, but then stays relatively constant. On the other hand, the number of miles driven per capita (`vehicmilespc`) appears to have its mean and variance increasing over time, while the unemployment rate (`unem`) fluctuates from year to year.



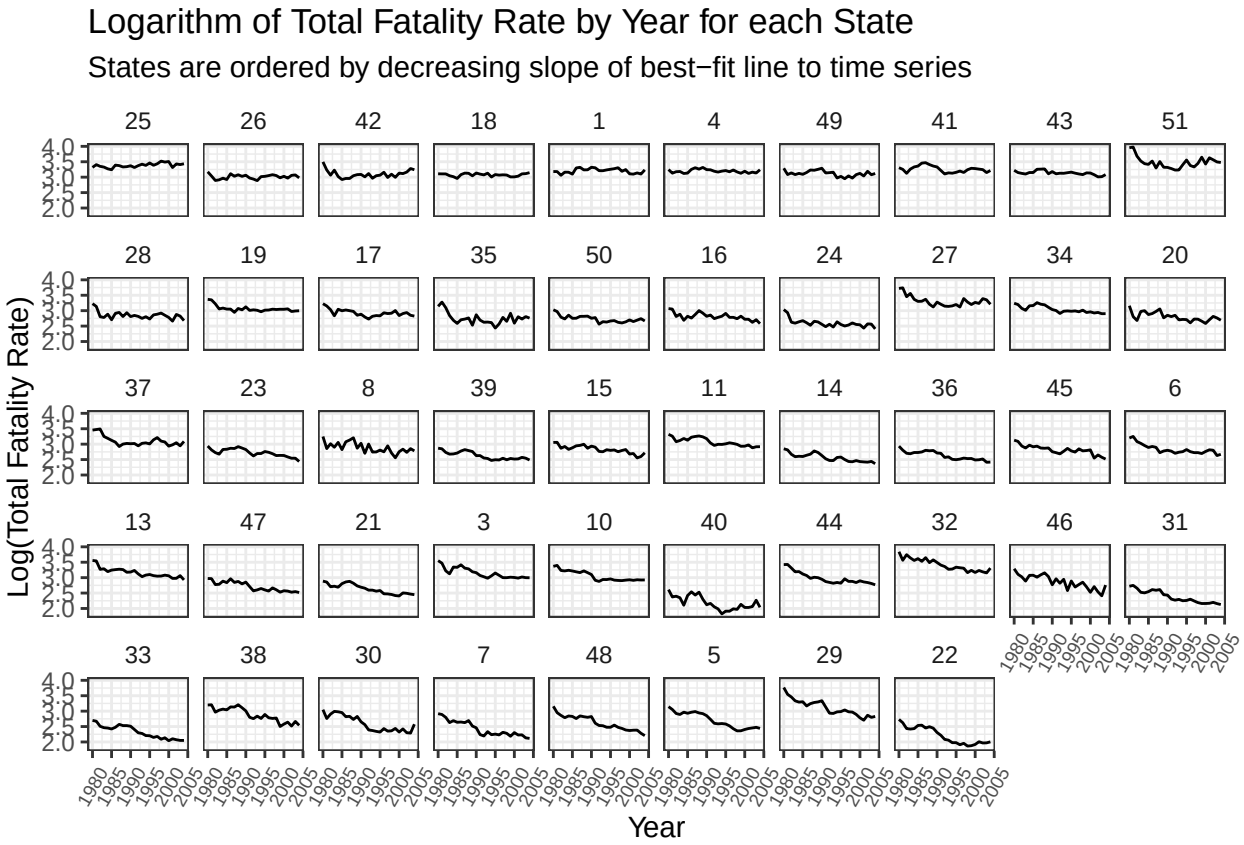
Below we provide side-by-side histograms of key analysis variables to explore if log-transformation improves in making the distribution more symmetric. While `totfatrte` and `unem` are right-skewed, a log-transformation makes them more symmetric (which will prove useful for modeling). Conversely, a log-transformation does not seem to improve the symmetry of `perc14_24` and `vehicmilespc` variables therefore we will use the original variables in the modeling.



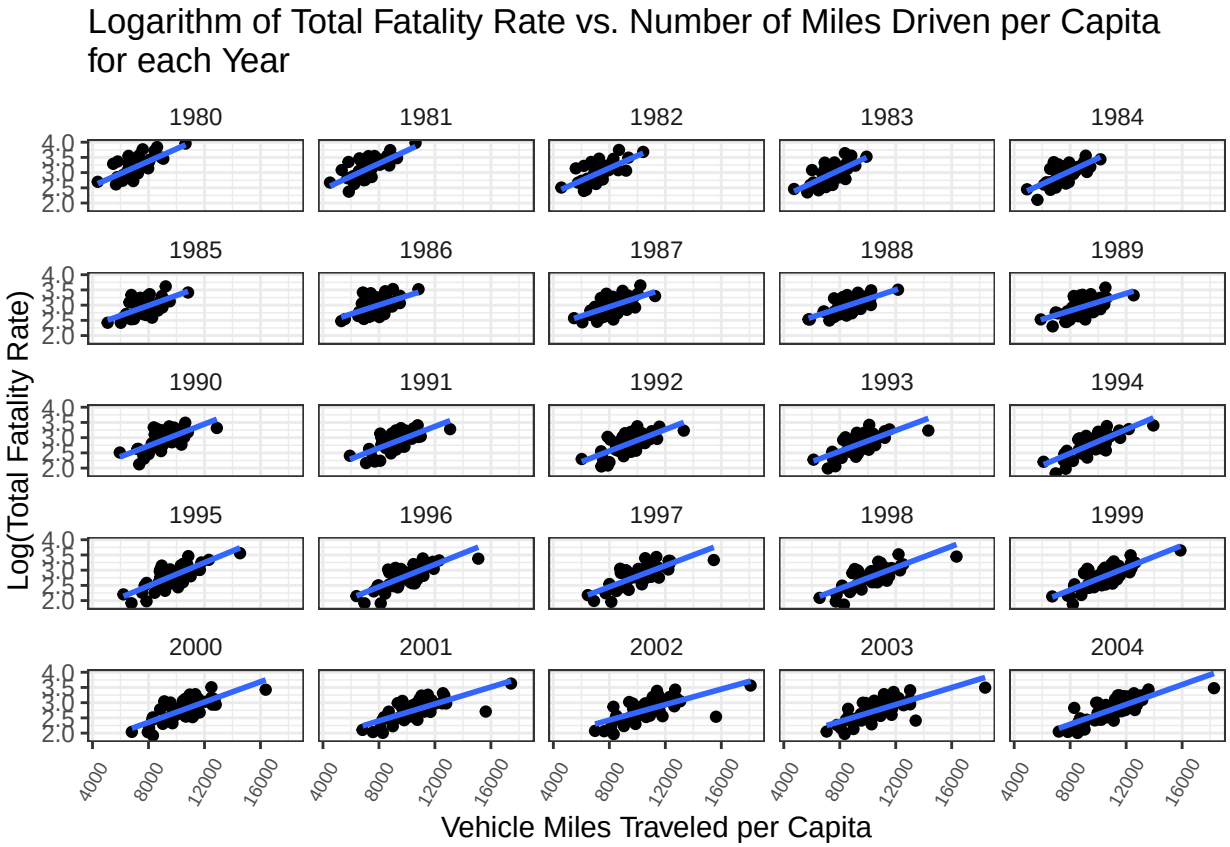
Checking on the interaction between implementation of laws and our key variable of interest, `totfatrte`, we observe that `totfatrte` tends to decrease when laws are enacted. `sl70plus` is the only exception in terms of median total fatality rate, but this is expected given that `sl70plus` indicates a very high (or nonexistent) speed limit, which one would expect to lead to a higher fatality rate. Note that we treat the law as true when it is implemented equal or more than half year, but we only do that on purpose of visualization.



The time series plots by state shown below are ordered by decreasing slope of the best-fit line to each state's time series of `log(totfatrte)`. In other words, the states with the highest slopes are shown first, and the states with the lowest slopes are shown last. By ordering the facets in this way, we can easily see how the total fatality rate remained relatively constant for some states, whereas for most states the total fatality rate declined (sometimes substantially) over the time period.



We notice from the annual facet plots below that the logarithm of `totfatrte` appears to have a strong positive correlation with `vehicmilespc` in all years. This plot reinforces our earlier findings that there were fewer vehicle miles traveled per capita and a higher fatality rate in earlier years (*e.g.*, in 1980 the cloud of points is in the top left corner of the facet plot) compared to later years, where the cloud of points tends to be lower (lower fatality rate) and more to the right (more vehicle miles traveled).

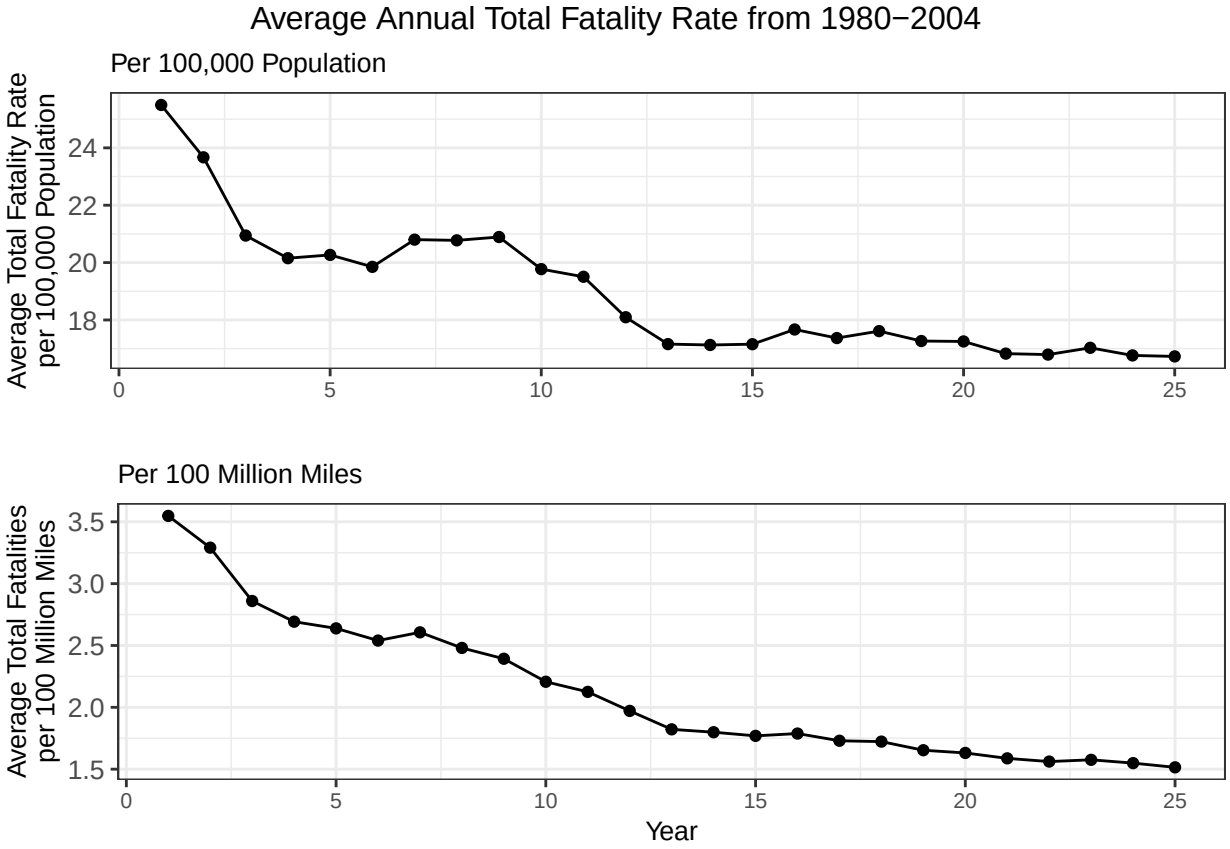


Due to the page limit, we have provided more EDA plots in the appendix for interested readers.

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.

`totfatrte`, or total fatality rate, is defined as the total number of traffic fatalities per 100,000 people in the population. The average `totfatrte` for each year in the data set is shown in the top plot below. We can see that the average in 1980 was more than 25 fatalities per 100,000 population, and the series experiences a decline over the time period until it reaches an average of under 17 fatalities per 100,000 population by 2004. Relative to the rest of the series, there is a notable increase in average fatality rate across the states from 1986-1988. The bottom plot below shows a similar plot for the average total fatalities per 100 million miles driven, which has a decreasing trend over the entire time period, with the only exception being 1986 (a slight increase relative to 1985). These plots provide initial evidence that driving has become more safe over this period.



Below we generated a pooled OLS linear regression model with log(`totfatrte`) as the dependent variable, using dummy variables for 1981-2004 as the only explanatory variables. It takes the form:

$$log(totfatrte) = \beta_0 + \beta_1 d_{81} + \beta_2 d_{82} + ... + \beta_{24} d_{04} + u$$

We observe slight heteroskedasticity in the model residuals (based on a plot of residuals vs. fitted values as well as a marginally significant Breusch-Pagan test), so we report the estimated coefficients with their associated robust standard errors. Note that due to space constraints we forego showing the model coefficients here, and instead provide the estimated coefficients and standard errors for all developed models in question 4.

This model captures the change in the U.S. average traffic fatality rate relative to 1980. For example, the estimated intercept of the model is 3.20, which works out to an estimated 24.4 fatalities / 100,000 population in 1980 (a slight underestimate relative to the average fatality rate across all 48 contiguous states in 1980 of 25.5 fatalities / 100,000 population). The estimated coefficient on `d81` (β_1) is -0.079, which indicates that the average fatality rate across all 48 contiguous states dropped by approximately 1.85 fatalities / 100,000 population relative to the average in 1980 ($exp(intercept) - exp(\beta_1 + intercept)$). Similar interpretations apply to the estimated coefficients on the `d82` to `d04` dummy variables. Because 1980 had the highest average fatality rate across the contiguous states, the estimated coefficients for all year dummy variables are negative. Additionally, using robust standard errors we see that the estimated coefficients for all of the year dummy variables are highly statistically significant. Comparing the magnitude of neighboring coefficients, because they are all interpreted relative to 1980, allows one to see how the average fatality rate

increased or decreased year-over-year. Importantly, the R-squared and adjusted R-squared values are 0.13 and 0.11, respectively, indicating that this model does not capture much of the variance in log(`totfatrte`).

As a final note, this pooled OLS model provides unreliable results due to failure to meet the independence assumption (*i.e.*, we are observing the exact same sample at different time periods).

```
q2lmmod <- plm(
  data = data.panel,
  formula = log.totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +
    d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 +
    d01 + d02 + d03 + d04,
  model = "pooling")
# Breusch-Pagan Test
bptest(q2lmmod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  q2lmmod
## BP = 36.089, df = 24, p-value = 0.05381
```

```
# p-value of 0.054; so, marginal evidence of heteroskedasticity.
# Use robust SEs to be on the safe side
# Get robust standard errors for model
se.q2lmmod = coeftest(q2lmmod, vcov = vcovHC)[ , "Std. Error"]
```

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.)

Below we add in all of the additional variables to the pooled OLS model, which takes the form:

$$\begin{aligned} log(totfatrte) = \beta_0 + \beta_1 d_{81} + \beta_2 d_{82} + ... + \beta_{24} d_{04} + \\ \beta_{25} bac08 + \beta_{26} bac10 + \beta_{27} perse + \beta_{28} sbprim + \\ \beta_{29} sbsecon + \beta_{30} sl70plus + \beta_{31} gdl + \beta_{32} perc14\_24 + \\ \beta_{33} log(unem) + \beta_{34} vehicmilespc + u \end{aligned}$$

As before, we log-transformed `totfatrte` to make it more symmetric, and we also log-transformed `unem` for the same reason. We did not transform the other two continuous variables—`perc14_24` and `vehicmilespc`—because while they are skewed, log-transformation does not provide much improvement to their distributions. The remaining variables are largely binary, but in some cases

a decimal value between 0 and 1 is provided as an indication that a law went into effect at some time (*month*/12) during the year. One might think to make these variables strictly binary by, for instance, doing a simple rounding of any decimal values. But, we did not transform any of these variables because we wanted to retain all of the information captured therein.

`bac08` and `bac10` correspond to laws specifying maximum blood alcohol content (BAC) limits of .08 and .10, respectively. The coefficients on `bac08` and `bac10`, neither of which are statistically significant, are -0.066 and -0.026, respectively. This means that, all else being equal, for each additional state that adds an .08 BAC law, there is a 6.6% lower fatality rate on average. Similarly, for each additional state that adds a .10 BAC law, there is a 2.6% lower fatality rate on average.

The coefficient on `perse` is -0.012 (also not statistically significant), which implies that for each additional state that adds an administrative license revocation (per se) law the average fatality rate decreases by 1.2% on average (holding all other variables constant).

It should be noted that the model results and these interpretations only hold under the assumption of independence, which we already discussed as being violated for the pooled OLS model in this case.

```
q3lmmod <- plm(
  data = data.panel,
  formula = log.totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +
    d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 +
    d01 + d02 + d03 + d04 + bac08 + bac10 + perse + sbprim + sbsecon +
    sl70plus + gdl + perc14_24 + log.unem + vehicmilespc,
  model = "pooling")
# Breusch-Pagan Test
bptest(q3lmmod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  q3lmmod
## BP = 70.247, df = 34, p-value = 0.0002529
```

```
# p-value of 0.00025; so, strong evidence of heteroskedasticity.
# Use robust SEs
# Get robust standard errors for model
se.q3lmmod = coeftest(q3lmmod, vcov = vcovHC)[ , "Std. Error"]
```

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?

We build the fixed effects model below, which takes the following form (where the double dots over a variable mean the time-demeaned data for that variable):

$$
\begin{aligned}
\ddot{log.totfatrte}_{it} = {} & \beta_1 \ddot{d81}_{it} + \beta_2 \ddot{d82}_{it} + ... + \beta_{24} \ddot{d04}_{it} + \beta_{25} \ddot{bac08}_{it} + \\
& \beta_{26} \ddot{bac10}_{it} + \beta_{27} \ddot{perse}_{it} + \beta_{28} \ddot{sbprim}_{it} + \beta_{29} \ddot{sbsecon}_{it} + \\
& \beta_{30} \ddot{sl70plus}_{it} + \beta_{31} \ddot{gdl}_{it} + \beta_{32} \ddot{perc14\_24}_{it} + \\
& \beta_{33} \ddot{log(unem)}_{it} + \beta_{34} \ddot{vehicmilespc}_{it} + \ddot{u}_{it}
\end{aligned}
$$

```
q4lmmod <- plm(
  data = data.panel,
  formula = log.totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
    d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 +
    d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 + perse + sbprim +
    sbsecon + sl70plus + gdl + perc14_24 + log.unem + vehicmilespc,
  model = "within")
# Breusch-Pagan Test
bptest(q4lmmod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  q4lmmod
## BP = 70.247, df = 34, p-value = 0.0002529
```

```
# p-value of 0.0002529; so, strong evidence of heteroskedasticity.
# Use robust SEs
# Get robust standard errors for model
se.q4lmmod = coeftest(q4lmmod, vcov = vcovHC)[ , "Std. Error"]
```

The table below shows a side-by-side comparison of the simple pooled OLS model, the expanded pooled OLS model, and the fixed effects model. When we run the fixed effects model, some of the coefficient values have increased compared to the expanded pooled OLS model (this is the case for all of the year dummy variables). We also see that there are more statistically significant variables compared with the expanded pooled OLS model: `perse` is very statistically significant with an estimated coefficient of -0.06, `sbprim` is marginally statistically significant with an estimated coefficient of -0.04, and `perc14_24` is marginally statistically significant with an estimated coefficient of 0.02.

Assumptions

OLS Assumptions

For the pooled OLS model all of the standard assumptions from a linear regression must be met (IID, no perfect collinearity, normally distributed and homoskedastic errors), in addition to these, we also have to account for the following since we are dealing with panel data:

- The relationship between the dependent variables and the independent variables remains constant over time.
- To test consistency, our independent variables (x) for a given moment (i) at a given time (t), x_{it}, must be uncorrelated with the fixed effects a_i.

Table 1:

	Log of Total Fatality Rate per 100,000 Population		
	Simple Pooled OLS	Expanded Pooled OLS	Fixed Effects
d81	−0.08*** (0.02)	−0.09*** (0.02)	−0.06*** (0.02)
d82	−0.20*** (0.02)	−0.28*** (0.03)	−0.13*** (0.02)
d83	−0.24*** (0.02)	−0.32*** (0.04)	−0.16*** (0.02)
d84	−0.23*** (0.02)	−0.27*** (0.06)	−0.19*** (0.02)
d85	−0.24*** (0.02)	−0.30*** (0.06)	−0.21*** (0.03)
d86	−0.20*** (0.02)	−0.26*** (0.08)	−0.16*** (0.03)
d87	−0.20*** (0.03)	−0.29*** (0.10)	−0.20*** (0.04)
d88	−0.19*** (0.02)	−0.30*** (0.11)	−0.23*** (0.04)
d89	−0.25*** (0.02)	−0.38*** (0.12)	−0.30*** (0.05)
d90	−0.27*** (0.02)	−0.43*** (0.13)	−0.31*** (0.05)
d91	−0.34*** (0.03)	−0.54*** (0.14)	−0.34*** (0.06)
d92	−0.40*** (0.03)	−0.65*** (0.15)	−0.40*** (0.06)
d93	−0.40*** (0.03)	−0.64*** (0.15)	−0.42*** (0.06)
d94	−0.41*** (0.03)	−0.63*** (0.15)	−0.45*** (0.06)
d95	−0.38*** (0.03)	−0.61*** (0.16)	−0.45*** (0.07)
d96	−0.40*** (0.03)	−0.73*** (0.17)	−0.49*** (0.07)
d97	−0.39*** (0.03)	−0.76*** (0.18)	−0.52*** (0.07)
d98	−0.41*** (0.03)	−0.81*** (0.18)	−0.57*** (0.07)
d99	−0.41*** (0.03)	−0.81*** (0.18)	−0.59*** (0.08)
d00	−0.44*** (0.03)	−0.83*** (0.19)	−0.62*** (0.07)
d01	−0.44*** (0.03)	−0.89*** (0.19)	−0.59*** (0.08)
d02	−0.43*** (0.03)	−0.94*** (0.19)	−0.55*** (0.08)
d03	−0.44*** (0.03)	−0.95*** (0.19)	−0.56*** (0.08)
d04	−0.45*** (0.03)	−0.94*** (0.20)	−0.59*** (0.08)
bac08		−0.07 (0.09)	−0.03 (0.03)
bac10		−0.03 (0.07)	−0.02 (0.02)
perse		−0.01 (0.05)	−0.06*** (0.02)
sbprim		−0.02 (0.06)	−0.04* (0.03)
sbsecon		0.01 (0.04)	0.002 (0.02)
sl70plus		0.23*** (0.06)	0.06** (0.03)
gdl		−0.001 (0.06)	−0.02 (0.02)
perc14_24		0.02 (0.02)	0.02* (0.01)
log.unem		0.25*** (0.07)	−0.20*** (0.02)
vehicmilespc		0.0002*** (0.0000)	0.0001*** (0.0000)
Constant	3.20*** (0.04)	1.29** (0.50)	
Observations	1,200	1,200	1,200
R^2	0.13	0.65	0.73
Adjusted R^2	0.11	0.64	0.71
F Statistic	7.06*** (df = 24; 1175)	62.47*** (df = 34; 1165)	87.28*** (df = 34; 1118)

Note: *p<0.1; **p<0.05; ***p<0.01

Are these assumptions met? In addition to the independence assumption violation (mentioned previously), the main issue with using a pooled OLS model on this particular data set is that individual states make decisions to legislate the laws intended to curb driving fatalities, so it is more likely that some of these laws come into effect *after* there has been a recent increase in driving related fatalities. This indicates that there is likely a relationship between fatalities and the laws coming into effect, which then may impact the rate of future fatalities. Such a scenario would break the exogeneity assumption when using a pooled OLS model and panel data analysis across time is better suited to handle this kind of situation. Additionally, as we will discuss in a moment, we also can see that the plots of the residuals of the OLS models do not pass the assumption that the errors are homoskedastic although both models have normally distributed errors.

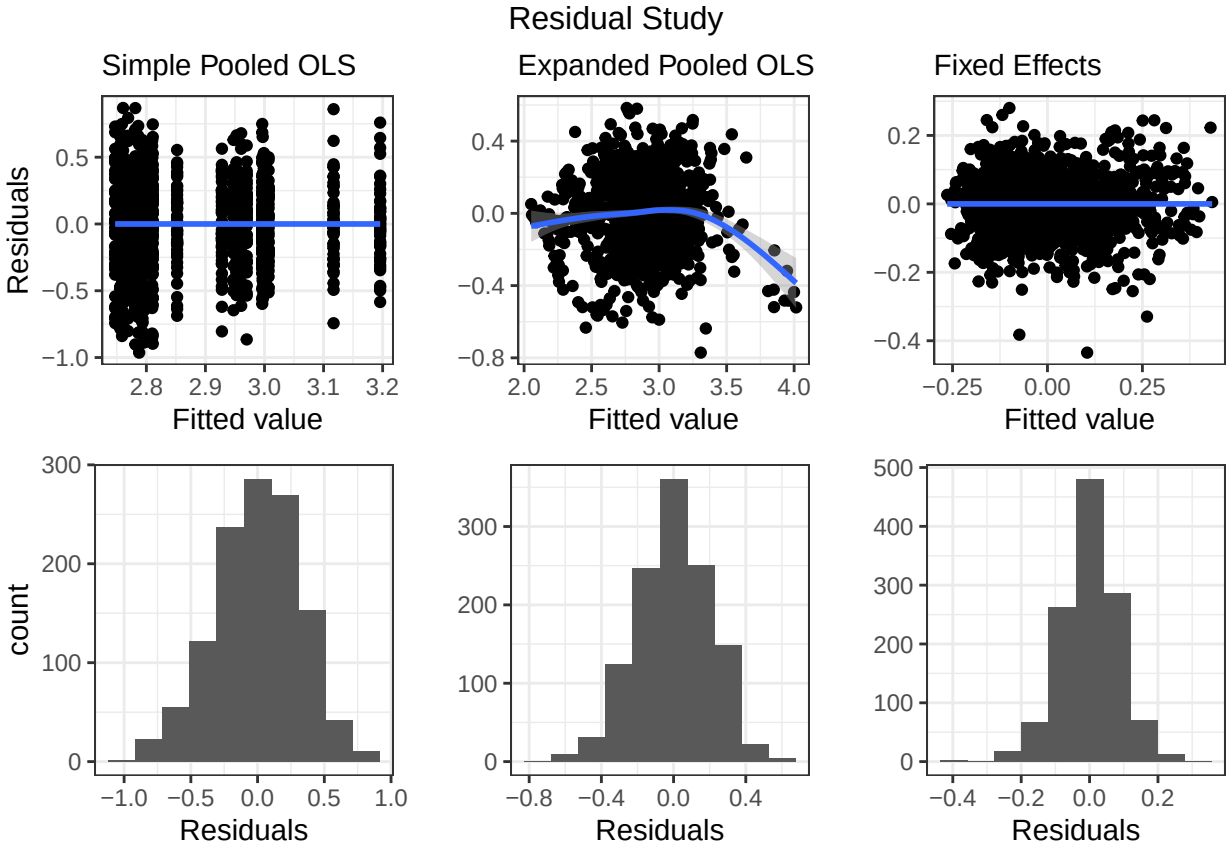
Fixed Effect Assumptions

For the Fixed Effects model, after the data is time-demeaned, we end up with a model that needs to pass all OLS model assumptions like above. The additional fixed effect assumptions we have to pass are (these come from Wooldridge p. 537):

- We have a random sample from the cross section.
- All the explanatory variables change over time and no perfect linear relationships exist among explanatory variables.
- The expected value of the idiosyncratic error given the explanatory variables in all time periods and the unobserved effect are zero.
- $Var(u_{it}|X_i, a_i) = Var(u_{it}) = \sigma_u^2$ for all t = 1,2, ..., T, which is to say that the variance of the idiosyncratic error, given the explanatory variables and unobserved effect remains constant across time.
- The model must have strict exogeneity and the time-varying error (u_{it}) must be uncorrelated with each explanatory variable across time. However, the time-invariant error (a_i) is allowed to be arbitrarily correlated with the independent variables.
- The time-varying errors (u_{it}) need to be homoskedastic and serially uncorrelated.
- Conditional on X_i and a_i, the idiosyncratic errors u_{it} are independent and identically distributed as normal errors.

Are these assumptions met? While we continue to have problems with the exogeneity assumption, as our time variant error is still correlated to our independent variables the fixed effect model mitigates this far better than the OLS model because we are looking at the relationship over time and able to better examine what happens *after* the law is in place. If our explanatory variables were just the measured variables, then a pooled OLS would *perhaps* be fine to use (assuming the other OLS assumptions were met) although panel data remains the better model since we're still examining a relationship over a long time period. Based on the residual plot discussed below, it does appear as if this model reasonably passes the assumptions of homoskedasticity among the time varying errors. All of our explanatory variables change over time, no perfect linear relationship exists among the explanatory variables and the errors are normally distributed. Overall, the fixed effect model does a better job at passing a majority of the assumptions listed above whereas the OLS models fail to pass some critical assumptions like homoskedastic error terms.

Comparing the plots of residuals against the fitted model values for all the models, it becomes apparent why the fixed effect model is the best model for conducting this analysis. The residuals of the simple pooled OLS show that the variance in our errors is inconsistent violating the assumption of homoskedasticty. The residuals in the expanded OLS model are similarly heteroskedastic as the variance of the errors change for different fitted values. Finally, the fixed effects model appears to have nearly perfectly homoskedastic residuals validating our prior discussion that the fixed effect model is likely the best model for our analysis. All three of the residuals are normally distributed with the fixed models residuals having the lowest variance.



5. (10%)

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

To run a random effects model, we have to assume that the unobserved effect (a_i) is uncorrelated with all explanatory variables in all time periods. This is not an assumption we can make because we are dealing with states over a 24-year period, and there could be a multitude of unobserved effects correlated to our input variables. For example, one unobserved effect has to do with the elected officials and what kind of policies are they enacting. This variable can change regularly in a democracy, and has a direct effect on when, and in response to what, legislation is passed. This relationship indicates the laws are highly correlated to the unobserved effect, so a fixed effect model is better for this analysis than a random effects model.

Wooldridge indicates in section 14-2a that researchers commonly apply both random effects and

fixed effects and then test for differences in the coefficients on the time-varying explanatory variables using the Hausman test. Moreover, Wooldridge states that a rejection of the Hausman test is taken to mean that the key random effects assumption is false, and so the fixed effects model should be used. Below, we develop a random effects model and use the Hausman test as suggested by Wooldridge. The resulting p-value is less than 2.2e-16, which indicates very strong evidence to reject the null hypothesis that the key random effects assumption is correct. This provides further evidence that the fixed effects model should be selected over the random effects model.

```
q4remod <- plm(
  data = data.panel,
  formula = log.totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
    d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 +
    d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 + perse + sbprim +
    sbsecon + sl70plus + gdl + perc14_24 + log.unem + vehicmilespc,
  model = "random")
phtest(q4lmmod, q4remod)
```

```
##
##  Hausman Test
##
## data:  log.totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +  ...
## chisq = 852.66, df = 34, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

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.

The effect on `totfatrte` due to a change in `vehicmilespc` can be calculated based on the following equation:

$$\ddot{log(totfatrte)}_{it} = \beta_{34}\ddot{vehicmilespc}_{it}$$

In this equation, $\ddot{vehicmilespc}_{it}$ is actually $vehicmilespc_{it} - \overline{vehicmilespc_i}$. By assuming that a nominal change of 1,000 miles does not impact the state's overall average vehicle miles traveled per capita, the effect on `log(totfatrte)` is simply interpreted in the same manner as would be done for a standard linear regression model with a log-transformed dependent variable. Therefore, the percent change in `totfatrte` based on a 1-unit change in `vehicmilespc` is simply $(exp(\beta_{34})-1)*100$. This equation is scaled by 1,000 to estimate the percent change in `totfatrte` based on a 1,000 mile increase in `vehicmilespc`. The results of this calculation are below (we omit the code from this report to save space):

```
# calculate the mean of the effect
mean <- (exp(coefficients(q4lmmod)['vehicmilespc']) - 1) * 100 * 1000
# calculate the confidence interval of the effects
CI <- (exp(confint(q4lmmod)["vehicmilespc", ]) - 1) * 100 * 1000
result<- round(data.frame(mean = mean, CI.lower = CI[1], CI.lower = CI[2]), 1)
knitr::kable(result)
```

	mean	CI.lower	CI.lower.1
vehicmilespc	6.2	5.3	7.2

Holding other explanatory variables constant, a $1,000$ unit increase in `vehicmilespc` will lead to an average 6.2% increase on `totfatrte` (with a 95% confidence interval from 5.3% to 7.2%). This result is sensible based on the intuition that more miles traveled should result in a higher likelihood of fatal accidents.

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?

Under the FE.1 through FE.6 provided by Wooldridge—with FE.6 being the assumption of idiosyncratic errors being uncorrelated—the fixed effects estimator is the best linear unbiased estimator. Therefore, if there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, the estimators will be biased and statistically inefficient. Moreover, due to the serial correlation, the reported standard errors on the estimators will be lower than they really are, which can lead to incorrect statistical inference.

We can check for these issues in our fixed effects model. To check the serial correlation, we use the Breusch-Godfrey/Wooldridge test. The code below runs this test and provides a p-value of less than 2.2e-16, which is (unfortunately) very strong evidence to reject the null hypothesis of no serial correlation in the idiosyncratic errors. Therefore, there is serial correlation in the idiosyncratic errors of our fixed effects model.

```
pbgtest(q4lmmod)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  log.totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +  ...
## chisq = 257.19, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

To check the heteroskedasticity, we use the Breusch-Pagan test. The code below runs this test and provides a p-value of less than 0.00025, which is (unfortunately) very strong evidence to reject the null hypothesis of homoskedasticity in the idiosyncratic errors. Therefore, there is heteroskedasticity in the idiosyncratic errors of our fixed effects model.

```
bptest(q4lmmod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  q4lmmod
## BP = 70.247, df = 34, p-value = 0.0002529
```

The idiosyncratic errors of the fixed effects model shows evidence of serial correlation and heteroskedasticity in this case, leading to a biased and inefficient model. We did use robust standard

errors in our reporting to mitigate the smaller than true standard error due to heteroskedasticity, but concerns still remain.

Conclusion

In this study we conducted a thorough analysis on a data set that contains data for the 48 continental U.S. states from 1980 through 2004, examining the core question of how various laws implemented over time impacted the total fatality rate. Ultimately, the fixed-effects model does a better job than a pooled OLS or a random-effects model in modeling our outcome variable. This is because it exceeds the pooled OLS and random effects models in terms of passing the required assumptions. Using robust standard errors, we conclude that "per se" laws (`perse`), speed limit laws greater than or equal to 70 mph (`sl70plus`), the logarithm of unemployment rate (`log(unem)`), and vehicle miles traveled per capita (`vehicmilespc`) have significant effects on the log of the total fatality rate per 100,000 population (`totfatrte`). We also found that blood alcohol limit laws at 10% (`bac10`) is marginally significant, and all of our year dummy variables are highly statistically significant, indicating there are potentially some unobserved effects being captured by the dummy variables that lead to decreases in fatality rate (*e.g.*, improvements in engineering over the 24 years leading to safer cars in cases of an accident).

One limitation of our final fixed-effects model is that our residuals did not pass the Breusch-Godfrey/Wooldridge test, which indicates that there is serial correlation in the idiosyncratic errors. Without correction, this serial correlation can bias our model results. A linear mixed-effects model may be able to mitigate such drawbacks, and could be an avenue for future study.