Lab 3: Panel Models

US Traffic Fatalities: 1980 - 2004

Contents

1 U.S. traffic fatalities: 1980-2004

We are answering the following **causal** question:

> **"Do changes in traffic laws affect traffic fatalities?"**

1.0.1 Preliminary Data Look

We have added state abbreviation in the data along with DC.

```
# Adding state abbriviation

# sorting stat.abb from R isn't correct because it puts Alaska and Alabama
# out of order

states.list <- c("AL","AK","AZ","AR","CA","CO","CT","DC","DE","FL","GA","HI",
                 "ID",
                 "IL","IN","IA","KS","KY","LA","ME","MD","MA","MI","MN","MS",
                 "MO","MT","NE","NV","NH","NJ","NM","NY","NC","ND","OH","OK",
                 "OR","PA","RI","SC","SD","TN","TX","UT","VT","VA","WA","WV",
                 "WI","WY")

states <- data.frame("index"=1:51,"abbr"=states.list)
data <- merge(data, states, by.x = "state", by.y = "index")
```

We then added year of observation in the data set.

```r
# year_of_observation
data <- data %>%
  mutate(year_of_observation = factor(year))
```

2 Build and Describe the Data

2.0.1 Load the data and produce useful features.

Producing a new variable, called `speed_limit` that re-encodes the data that is in `sl55`, `sl65`, `sl70`, `sl75`, and `slnone`;.

```r
data <- data %>%
  mutate(speed_limit =
           case_when(
             data$sl55 > .5 ~ 55,
             data$sl65 > .5 ~ 65,
             data$sl70 > .5 ~ 70,
             data$sl75 > .5 ~ 75,
             data$slnone > .5 ~ 0,
             data$year == 1996 & data$state == 6 ~ 75,
             data$year == 1996 & data$state == 11 ~ 70,
             data$year == 1995 & data$state == 21 ~ 65,
             data$year == 1997 & data$state == 24 ~ 70,
             data$year == 1988 & data$state == 47 ~ 65
           ))
```

Producing a new variable, called `year_of_observation` that re-encodes the data that is in `d80`, `d81`, ... , `d04`.

```r
data <- data %>%
  mutate(year_of_observation =
           case_when(
             data$d80 == 1 ~ 1980,
             data$d81 == 1 ~ 1981,
             data$d82 == 1 ~ 1982,
             data$d83 == 1 ~ 1983,
             data$d84 == 1 ~ 1984,
             data$d85 == 1 ~ 1985,
             data$d86 == 1 ~ 1986,
             data$d87 == 1 ~ 1987,
             data$d88 == 1 ~ 1988,
             data$d89 == 1 ~ 1989,
             data$d90 == 1 ~ 1990,
             data$d91 == 1 ~ 1991,
             data$d92 == 1 ~ 1992,
             data$d93 == 1 ~ 1993,
             data$d94 == 1 ~ 1994,
             data$d95 == 1 ~ 1995,
             data$d96 == 1 ~ 1996,
             data$d97 == 1 ~ 1997,
             data$d98 == 1 ~ 1998,
             data$d99 == 1 ~ 1999,
             data$d00 == 1 ~ 2000,
             data$d01 == 1 ~ 2001,
```

```
            data$d02 == 1 ~ 2002,
            data$d03 == 1 ~ 2003,
            data$d04 == 1 ~ 2004,
            TRUE ~ 00000
          )
)
# data %>% filter(year_of_observation == 00000)
```

Producing a new variable for each of the other variables that are one-hot encoded (i.e. `bac*` variable series).

```
data <- data %>%
  mutate(
    blood_alc =
      case_when(
        bac10 > .5 ~ .1,
        bac08 > .5 ~ .08,
        bac08 == 0 & bac10 == 0 ~ 0,
        bac10 > bac08 ~ .1,
        TRUE ~ .08
      )
  )
```

Renaming the variables to more legible names.

```
# rename the variables to sensible names
data <- data %>%
  dplyr::rename(
    "total_fatalities_rate"                     = "totfatrte",
    "minimum_drinking_age"                      = "minage",
    "zero_tolerance_law"                        = "zerotol",
    "state_population"                          = "statepop",
    "graduated_drivers_license_law"             = "gdl",
    "per_se_laws"                               = "perse",
    "total_traffic_fatalities"                  = "totfat",
    "total_nighttime_fatalities"                = "nghtfat",
    "total_weekend_fatalities"                  = "wkndfat",
    "total_fatalities_per_100_million_miles"    = "totfatpvm",
    "nighttime_fatalities_per_100_million_miles" = "nghtfatpvm",
    "weekend_fatalities_per_100_million_miles"  = "wkndfatpvm",
    "nighttime_fatalities_rate"                 = "nghtfatrte",
    "weekend_fatalities_rate"                   = "wkndfatrte",
    "vehicle_miles"                             = "vehicmiles",
    "unemployment_rate"                         = "unem",
    "pct_population_14_to_24"                   = "perc14_24",
    "vehicle_miles_per_capita"                  = "vehicmilespc",
    "primary_seatbelt_law"                      = "sbprim",
    "secondary_seatbelt_law"                    = "sbsecon",
    "Abbreviation"                              = "abbr"
  )

# Check data
#data %>% glimpse()
```

2.0.2 Description of the basic structure of the dataset.

The long-panel data being reviewed here has 60 columns and 1200 row, where each row is returning metrics having to do with traffic incidents in each state (excluding Alaska and Hawaii) from the years 1980 through 2004, at an annual granularity. Specifically, we will be focusing on traffic fatalities in each state to see if regulations like blood alcohol limits, speed limits, and so on appear to impact the rate of fatalities per state. This dataset also includes columns that show how many of these fatalities happened while driving at night or on the weekend. The data was compiled by Donald G Freedman for the paper "Drunk living legislation and traffic fatalities:New evidence on BAC 08 laws" - Contemporary Economic Policy 2007. In the paper it is noted that "Fatality data are from the Fatality Analysis Reporting System (FARS) compiled by NHTSA. Data on traffic legislation for the years 1982—1999 were provided by Thomas Dee. Earlier data on legislation were taken from Zador et al. (1989) and later data on legislation from the National Center for Statistics and Analysis at the NHTSA Web site at http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/. Data on graduated drivers' licenses are taken from Dee, Grabowski, and Morrisey (2005). State unemployment rates are from Dee and the Bureau of Labor Statistics; age data are from the Bureau of the Census". **The outcome of interest, `total_fatalities_rate` is defined as the number of fatalities per 100,000 people.**

2.0.3 EDA

A thorough EDA is conducted on the dataset to explore the relationship of certain variables at the state and aggregate level. First, data validity checks were conducted in order to determine that the observations across the states in the dataset were consistent and repeated across all years without large gaps (code is commented out to reduce output). This showed that the dataset did in fact have consistency and all observations were accounted for. We also verified that state numbers 2 and 13, which corresponded to Alaska and Hawaii, were indeed missing from the dataset.

```
#data %>%
#  dplyr::select(year_of_observation, state) %>%
#  table()
```

```
#data %>%
#  is.pconsecutive()
#
#pdim(data)
```

Next, we interrogated the total fatalities rate available in the data. A state by state view of each is shown in **Figure 1**. An initial observation can be made by studying this figure - states like Wyoming and New Jersey appear to stand out from others as having high fatality rates. States like Connecticut and Rhode Island appear to stand out as having low fatality rate. All states show a downward or flat trend in fatality over time.

The mean fatality rate is reported in **Figure 2**. It shows that, over time and across all the 48 contiguious states (plus DC), the fatality rate decreased between 1980 and 2004. Notably, there was a substantial decrease between 1980 and approximately 1993, with leveling off afterwards, from a high of >25 deaths per 100K to a low of ~17 per 100K.

An additional study across states was done to see how states performed on an equal axis, which is shown in **Figure 3**. The major takeaway is a repeat of **Figure 1**, which is that Wyoming and New Jersey are states that appear to have higher rates of fatality, while Connecticut, Rhode Island, New Mexico, and Idaho appear to be states with lower fatality rates.

To explore the impact of other variables of interest at the aggregagate and state level, we first used a scatterplot matrix to find baseline correlations between the variables when averaging across states. **Figure 4** shows the scatterplot matrix of the variables, averaged across the states and DC.

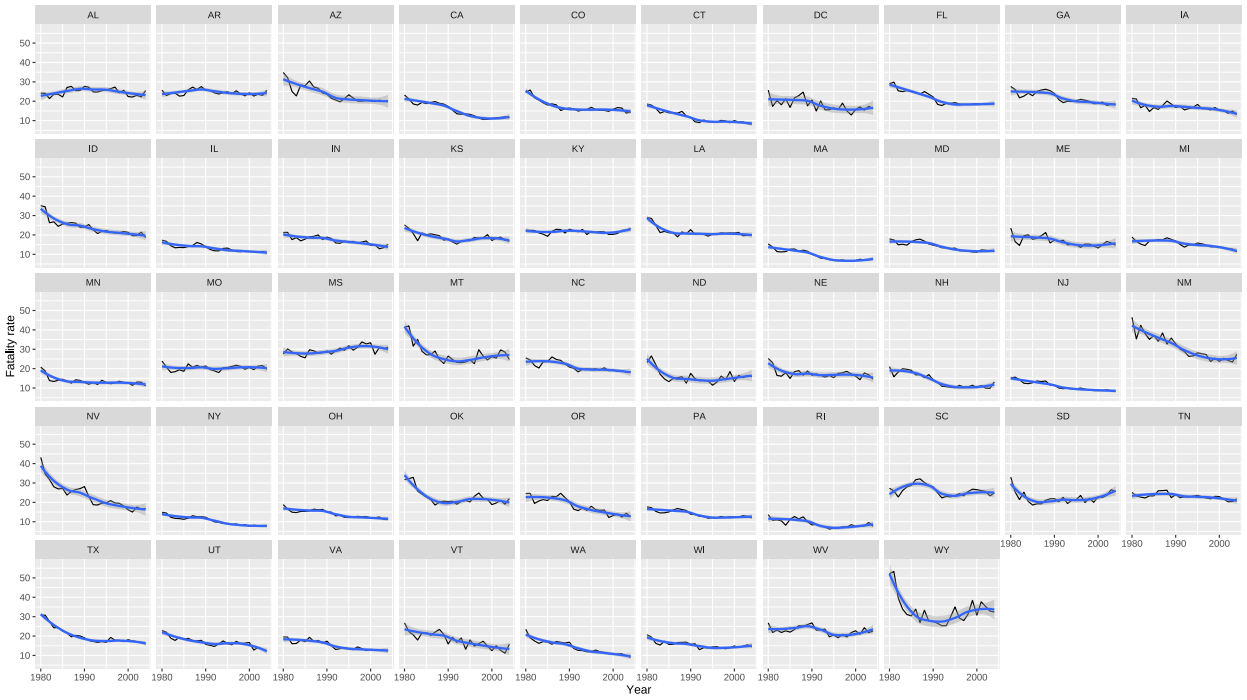

Figure 1: Fatality Rate by State (where data recorded) and Year Since 1980

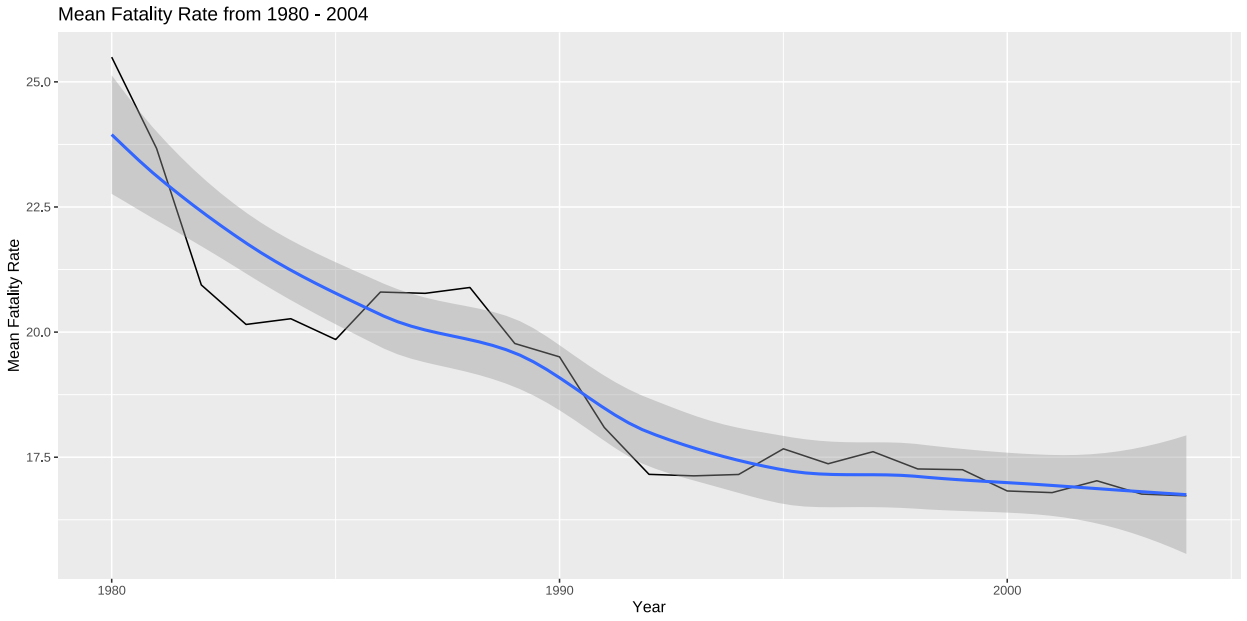

Figure 2: Time Series of Average Fatality

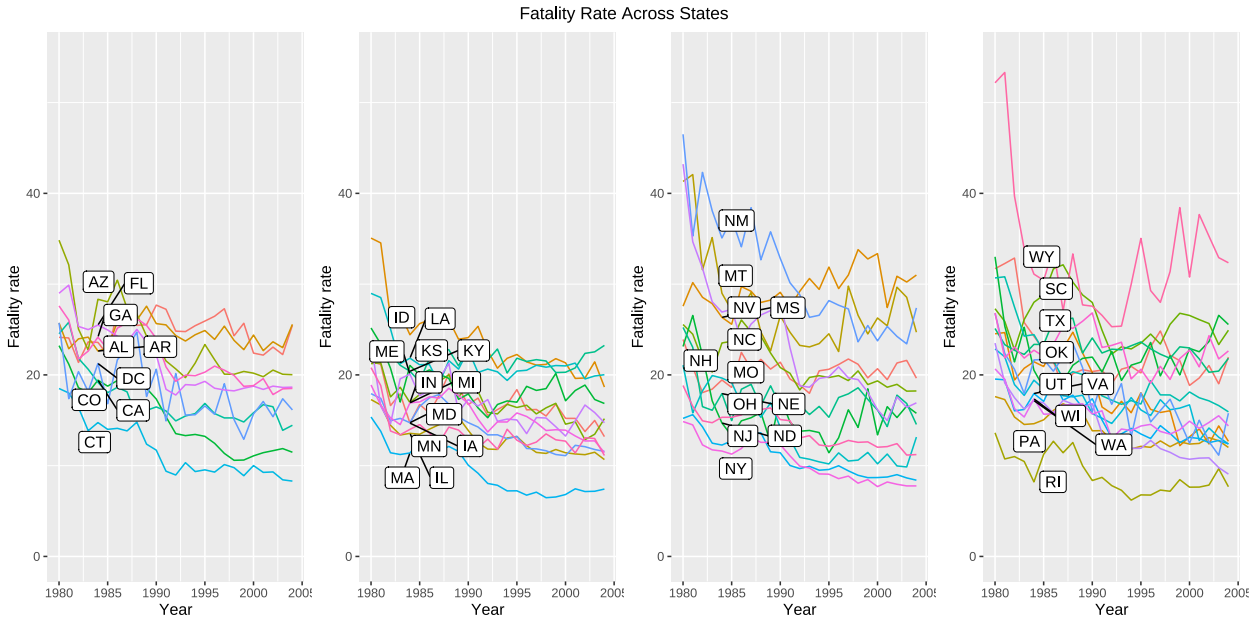

Figure 3: Fatality Rate by State on Common Axis

The variables with strong correlation to the average fatality rate include vehicle miles driven per capita (-0.88), average population (-0.857), and average percentage of the population consisting of individuals aged 14-24 (0.909). Lessor variables included average unemployment and the average BAC. The population 14-24 is interesting because it's known that young adults tend to engage in riskier activties.

```
all_means <- data %>%
  dplyr::group_by(year) %>%
  dplyr::summarise(
    avg_total_fatality_rate = mean(total_fatalities_rate),
    avg_drinking_age = mean(minimum_drinking_age),
    avg_pop = mean(state_population),
    avg_unemployment = mean(unemployment_rate),
    avg_perc_pop = mean(`pct_population_14_to_24`),
    avg_speed_limit = mean(speed_limit),
    avg_blood_alc_limit = mean(blood_alc),
    avg_vmd_percap = mean(vehicle_miles_per_capita)
  )

all_means %>% ggpairs() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

To see how each state compares to the other, **Figure 5** and **Figure 6** are displayed with shows the time series, for each state, of the vehicle miles driven per capita and the percent of the population aged 14-24. Interesting, again Wyoming stands out as a state where more vehicle miles are driven per capita than anywhere else. New York is an outlier in the oppostie direction. Wyoming does not appear to have a high number of adolescant adults, as a percentage of its population, shown in **Figure 6**.

Finally, a boxplot of the distribution across all years is built for a subset of the variables. This is presented in **Figure 7**. As expected, the state with the highest average fatality rate is Wyoming, and the state with the smallest is Rhode Island. When viewing the vehicle miles per capita, the

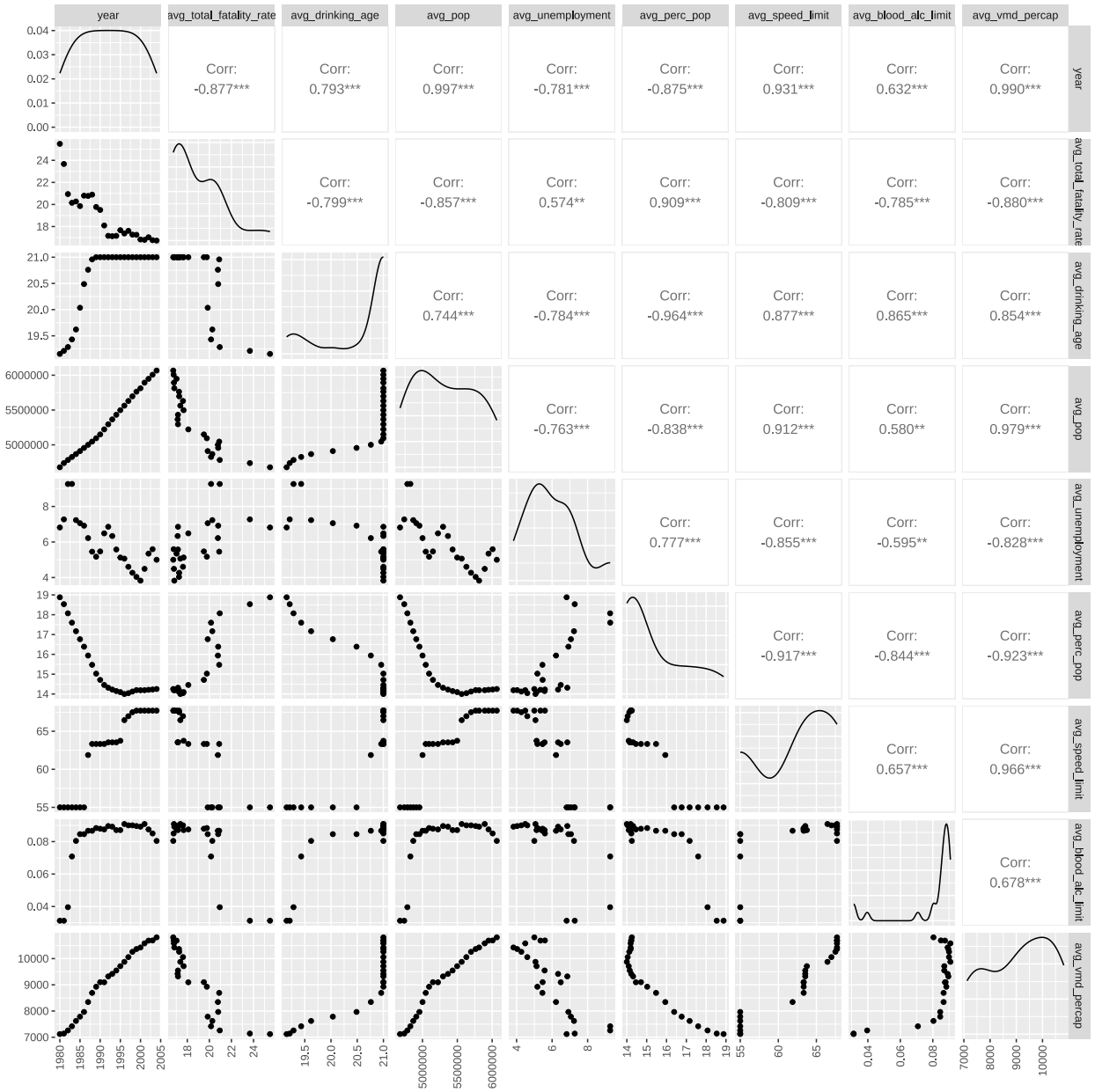

Figure 4: Scatterplot Matrix of Variables of Interest

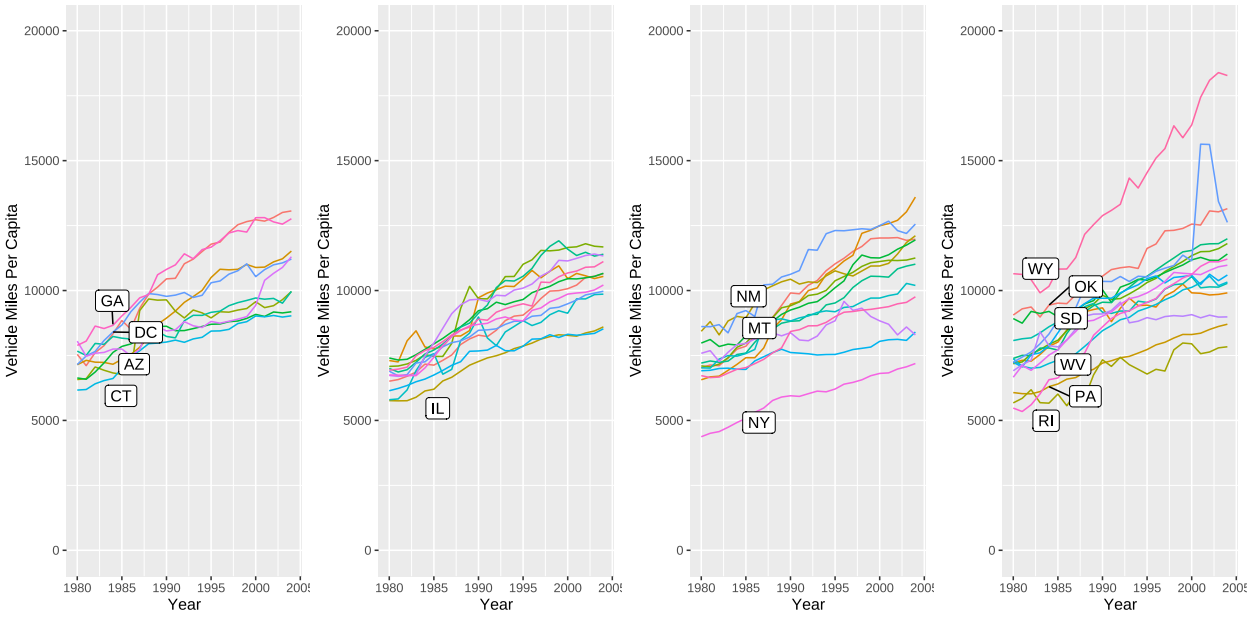

Figure 5: Vehicle Miles Driven Per Capita

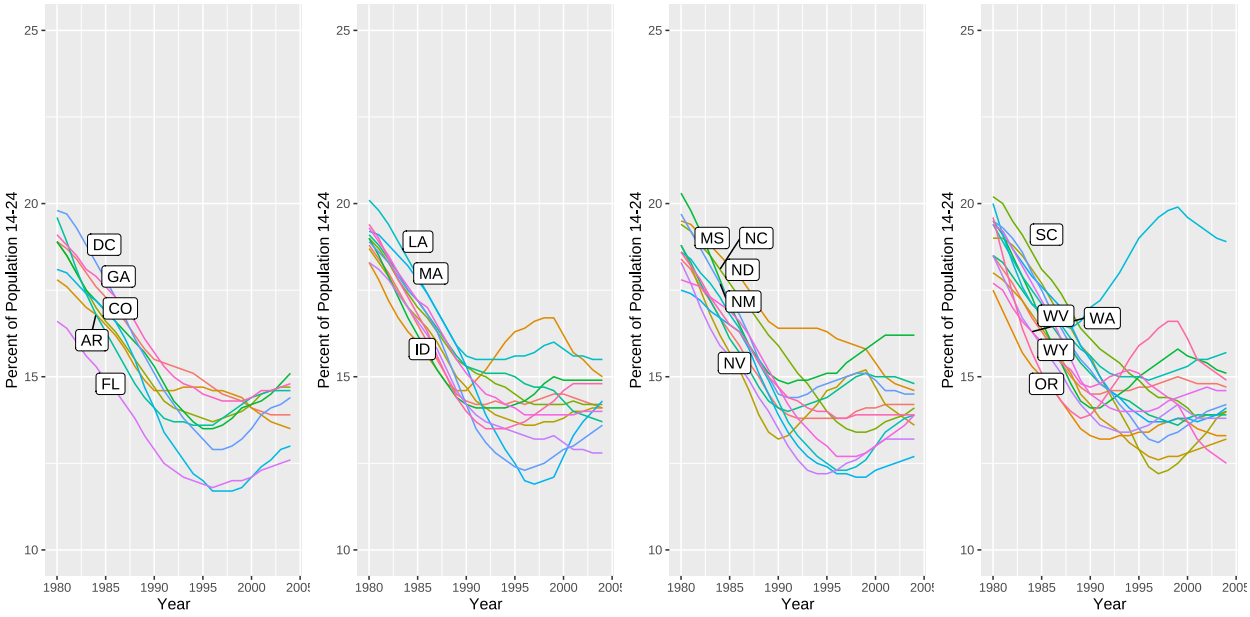

Figure 6: Percent of Population Aged 14-24

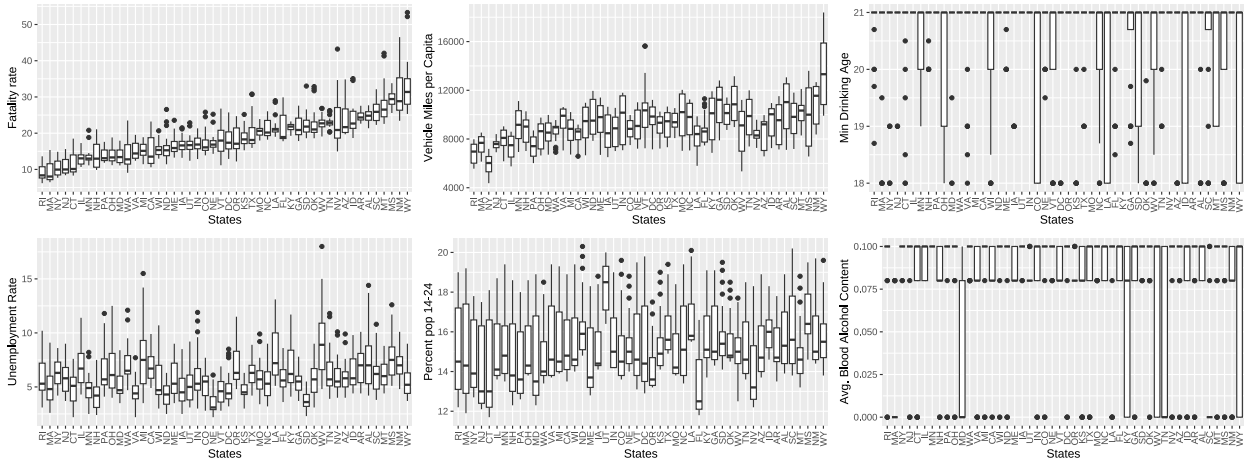

Figure 7: Box Plot of States - Sorted by Total Fatality Rate

relationship is striking - Wyoming is also the state with the highest average vehicle miles driven per capita across the years surveyed. The population aged 14-24 is also presented in this fashion but no major observations can be determined. We will the variables that we've identified in the EDA as important to build the causal models.

3 Preliminary Model

3.0.1 Linear Model

A linear model is a sensible starting place because we have non-independence in this data set. This dataset also contains some variability, since we have repeated measures taken over time, (i.e, finding the fatality rate of each state every year). A linear model can help us determine the relationships between different variables and the outcome and allow us to better understand the source of variability within the dataset.

```
lsdv_mod <- plm(total_fatalities_rate ~  minimum_drinking_age + state,
                index=c("state", "year"), model = "within", data=data)

stargazer(lsdv_mod, type = "latex", header=FALSE,
          omit.stat = c("ser","f","adj.rsq"), dep.var.labels = "",
          column.labels = c("Within"), title="Preliminary Model")
```

Table 1: Preliminary Model

	Dependent variable:
	Within
minimum_drinking_age	-1.450^{***}
	(0.099)
Observations	1,200
R^2	0.158
Note:	*p<0.1; **p<0.05; ***p<0.01

3.0.2 Discussion on Linear Model

We incorporate dummy variables for the year and for all states. The model doesn't include any other variable, so omitted variable bias could be present, however, the results imply that minimum drinking age does have an impact on the total fatality rate within states.

4 Expanded Model

4.0.1 Modeling

```
# create speed limit over 70 variable
data <- data %>%
  mutate(speed_lim_70 = ifelse(speed_limit == 55 | speed_limit == 65, 0, 1))


exp_mod <- plm(
  log(total_fatalities_rate) ~ minimum_drinking_age + state + blood_alc +
    per_se_laws + primary_seatbelt_law + secondary_seatbelt_law + speed_lim_70 +
    graduated_drivers_license_law + log(pct_population_14_to_24) +
    log(unemployment_rate) + log(vehicle_miles_per_capita),
  index=c("state", "year"),
  model = "within",
  data=data
)

stargazer(exp_mod, type = "latex", header=FALSE,
          omit.stat = c("ser","f","adj.rsq"), dep.var.labels = "",
          column.labels = c("Within"), title="Expanded Model")
```

Based on the expanded model, the vehicle miles driven per capita, unemployment rate, percent of population between 14 and 24, as well as per se laws and each year of observation were statistically significant.

5 State-Level Fixed Effects

5.0.1 Modeling

```
within.model <- plm(
  total_fatalities_rate ~ as.numeric(year) +
    blood_alc +
    per_se_laws +
    primary_seatbelt_law +
    speed_limit +
    graduated_drivers_license_law +
    log(`pct_population_14_to_24`) +
    log(unemployment_rate) +
    log(vehicle_miles_per_capita),
  data = data,
  index=c("state", "year"),
  model="within"
)
```

Table 2: Expanded Model

	Dependent variable:
	Within
minimum_drinking_age	0.011^{**}
	(0.005)
blood_alc	-0.280^{**}
	(0.119)
per_se_laws	-0.092^{***}
	(0.011)
primary_seatbelt_law	-0.090^{***}
	(0.016)
secondary_seatbelt_law	-0.030^{**}
	(0.012)
speed_lim_70	-0.029^{***}
	(0.011)
graduated_drivers_license_law	-0.067^{***}
	(0.011)
log(pct_population_14_to_24)	0.982^{***}
	(0.054)
log(unemployment_rate)	-0.158^{***}
	(0.015)
log(vehicle_miles_per_capita)	0.140^{***}
	(0.048)
Observations	1,200
R^2	0.614
Note:	*p<0.1; **p<0.05; ***p<0.01

```
pool.model <- plm(
  total_fatalities_rate ~ as.numeric(year) +
    blood_alc +
    per_se_laws +
    primary_seatbelt_law +
    speed_limit +
    graduated_drivers_license_law +
    log(`pct_population_14_to_24`) +
    log(unemployment_rate) +
    log(vehicle_miles_per_capita),
  data = data,
  index=c("state", "year"),
  model="pooling"
)

fd.model <- plm(total_fatalities_rate ~ as.numeric(year) +
    blood_alc +
    per_se_laws +
    primary_seatbelt_law +
    speed_limit +
    graduated_drivers_license_law +
    log(`pct_population_14_to_24`) +
    log(unemployment_rate) +
    log(vehicle_miles_per_capita),
  data = data,
  index=c("state", "year"),
  model="fd"
)

between.model <- plm(
  total_fatalities_rate ~ as.numeric(year) +
    blood_alc +
    per_se_laws +
    primary_seatbelt_law +
    speed_limit +
    graduated_drivers_license_law +
    log(`pct_population_14_to_24`) +
    log(unemployment_rate) +
    log(vehicle_miles_per_capita),
  data = data,
  index=c("state", "year"),
  model="between"
)

random.model <- plm(
  total_fatalities_rate ~ as.numeric(year) +
    blood_alc +
    per_se_laws +
    primary_seatbelt_law +
    speed_limit +
    graduated_drivers_license_law +
    log(`pct_population_14_to_24`) +
    log(unemployment_rate) +
```

```
    log(vehicle_miles_per_capita),
  data = data,
  index=c("state", "year"),
  model="random"
)


stargazer(pool.model, fd.model, between.model, within.model, random.model,
          type = "latex", header=FALSE,
          omit.stat = c("ser","f","adj.rsq"), dep.var.labels = "",
          column.labels = c("Pooled", "FD", "Between", "Within", "Random"),
          title="Comparison of Models")
```

State level fixed effect models are estimated using the `plm` function in R, which takes advantage of the panel data structure for model estimations. A *within*, *pooling*, *first difference*, *between*, and *random effects* model are estimated and compared in **Table 1**.

5.0.2 Model analysis

```
pFtest(within.model, pool.model)
```

```
    F test for individual effects

  data:  total_fatalities_rate ~ as.numeric(year) + blood_alc + per_se_laws +  ...
  F = 77.881, df1 = 47, df2 = 1143, p-value < 2.2e-16
  alternative hypothesis: significant effects
```

A pFtest is conducted to determine where state and time fixed effects should be included. Our pFtest returns a significant p value, meaning that we reject the null hypothesis. This means we should include the state and time fixed effects in our model.

```
pwfdtest(fd.model, data=data, index=c("state", "year"), h0="fe")
```

```
    Wooldridge's first-difference test for serial correlation in panels

  data:  fd.model
  F = 21.758, df1 = 1, df2 = 1102, p-value = 3.473e-06
  alternative hypothesis: serial correlation in original errors
```

```
pwfdtest(fd.model, data=data, index=c("state", "year"), h0="fd")
```

```
    Wooldridge's first-difference test for serial correlation in panels

  data:  fd.model
  F = 78.459, df1 = 1, df2 = 1102, p-value < 2.2e-16
  alternative hypothesis: serial correlation in differenced errors
```

```
phtest(within.model, random.model)
```

```
    Hausman Test

  data:  total_fatalities_rate ~ as.numeric(year) + blood_alc + per_se_laws +  ...
```

Table 3: Comparison of Models

	Dependent variable:				
	Pooled	FD	Between	Within	Random
	(1)	(2)	(3)	(4)	(5)
as.numeric(year)	-0.510^{***}			-0.374^{***}	-0.399^{***}
	(0.041)			(0.029)	(0.029)
blood_alc	-13.763^{***}	-5.409	-2.498	-12.184^{***}	-12.658^{***}
	(3.790)	(3.340)	(25.103)	(2.373)	(2.411)
per_se_laws	-0.907^{***}	-0.413	-0.428	-1.462^{***}	-1.389^{***}
	(0.305)	(0.394)	(1.691)	(0.228)	(0.230)
primary_seatbelt_law	-0.065	0.298	-0.491	-0.678^{***}	-0.638^{**}
	(0.341)	(0.445)	(1.779)	(0.259)	(0.262)
speed_limit	-0.008	0.024	0.064	-0.024^{**}	-0.025^{**}
	(0.018)	(0.019)	(0.131)	(0.011)	(0.011)
graduated_drivers_license_law	0.213	-0.035	-4.999	0.819^{***}	0.835^{***}
	(0.449)	(0.359)	(5.143)	(0.247)	(0.252)
log(pct_population_14_to_24)	11.831^{***}	17.039^{***}	5.098	9.394^{***}	9.853^{***}
	(1.586)	(3.416)	(10.094)	(1.068)	(1.081)
log(unemployment_rate)	3.926^{***}	-2.656^{***}	11.254^{***}	-3.318^{***}	-2.997^{***}
	(0.438)	(0.357)	(2.520)	(0.305)	(0.309)
log(vehicle_miles_per_capita)	28.265^{***}	4.226^{**}	30.930^{***}	10.908^{***}	12.875^{***}
	(0.906)	(1.937)	(4.107)	(1.161)	(1.128)
Constant	-268.585^{***}	-0.263^{***}	-298.433^{***}		-111.373^{***}
	(8.517)	(0.078)	(39.043)		(10.754)
Observations	1,200	1,152	48	1,200	1,200
R^2	0.562	0.082	0.713	0.582	0.568

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

```
  chisq = 32.73, df = 9, p-value = 0.0001488
  alternative hypothesis: one model is inconsistent
```

5.0.3 Blood Alcohol Affects

The fixed effect model estimates that blood alcohol has a coefficient of -6.163. In action, this means that for a unit increase in blood alcohol limit, the fatality rate will decrease by approximately 6 fatalities per 100,000 people. However in our case, blood alcohol is measured in hundreths and tenths rather than in unitary increments, so a full unit increase isn't likely to take place. In the Pooled model, the coefficient for blood alcohol is -8.736, which is the largest impact out of the five models. The coefficient for the First Difference model is -3.092, the smallest impact of the models. The between model has a similar coefficient of -3.313, and lastly the Random Effects model has a coefficient of -6.610, very simlilar to the Fixed Effect Model. All of these coefficients are significant.

5.0.4 Per se Laws

The Fixed Effect model estimates that the coefficient for the per-se parameter is -1.219. In our data, the per se parameter is a one-hot-encoded variable with some values being between zero and one, in cases where states took on the law mid-year. In the case where a state does have per se laws in a given year, our model estimates that the rate of fatalities will decrease by 1.2 per 100,000 people. The is the largest impact of the five models, with the coefficients of the Pooled Effect, First Difference, Between, and Random Effect being -0.881, -0.669, -0.467, and -1.178 respectively. All of these coefficients are significant.

5.0.5 Primary Seat Belt Laws

The primary seat belt law variable is also a Boolean, indicating whether a state uses primary seat belt laws or secondary. For the Pooled model, the coefficient is not significant and is -0.322, a fairly small impact. Similarly, the coefficient for the First Difference model is -0.201, and the coefficient for the Between model is -0.491. This variable is not significant in either model. However, the coefficients for both the Fixed Effect and Random Effect models are both significant, being -0.867 and -0.833 respectively.

5.0.6 Reliability of Estimates

The Wooldridge first-difference test for serial correlation in panels returns significant p values for both the Fixed Effect (within) and First Difference models, however the p value for the Fixed Effect model is far more significant. This leads us to believe that the Fixed Effect model is more reliable than the First Difference and Pooled models. Moving on, the outcome of the Hausmen test is not significant, meaning we do not reject the null hypothesis that random effects are appropriate, suggesting that we should not use the Fixed Effect model. Because of these outcomes, the estimates created by the Random Effects model will be the most reliable.

5.0.7 Model Assumptions

5.0.8 Are the assumptions reasonable?

6 Consider a Random Effects Model

Instead of estimating a fixed effects model, should you have estimated a random effects model?

```
data <- data %>%
  mutate(speed_limit = ifelse(sl55 >= 0.5, 55,
                       ifelse(sl65 >= 0.5, 65,
                       ifelse(sl70 >= 0.5, 70,
```

```
                        ifelse(sl75 >= 0.5, 75,
                        ifelse(slnone >= 0.5, 'none', 0)
                        )))))

data = data %>% mutate(
    seatbelt = factor(seatbelt), # 'seatbelt' categorizes primary or secondary
    speed_limit_70plus = ifelse(speed_limit == 55 | speed_limit == 65, 0, 1)
  )

data <- data %>% mutate(blood_alcohol_limit_10 = ifelse(bac10 >= 0.5, 1, 0),
                        blood_alcohol_limit_08 = ifelse(bac08 >= 0.5, 1, 0)
                        ) %>%
                 mutate(bac = ifelse(blood_alcohol_limit_10==1, '10',
                                     ifelse(blood_alcohol_limit_08==1, '8', 'none')))

random.effect.model <- plm(log(total_fatalities_rate) ~
                             year_of_observation +
                             factor(bac) +
                             per_se_laws +
                             primary_seatbelt_law +
                             secondary_seatbelt_law +
                             speed_limit_70plus +
                             graduated_drivers_license_law +
                             pct_population_14_to_24 +
                             unemployment_rate +
                             vehicle_miles_per_capita,
                             #log(pct_population_14_to_24) +
                             #log(unemployment_rate) +
                             #log(vehicle_miles_per_capita),
                           data = data,
                           index = c("state", "year"),
                           model = "random")

stargazer(random.effect.model, type='latex', header=FALSE,
          omit.stat = c("ser","f","adj.rsq"),
          dep.var.labels = "", title='Random Effects Model')
```

```
coeftest(random.effect.model, vcov. = vcovHC, type = "HC1")
```

```
  t test of coefficients:

                                 Estimate  Std. Error t value  Pr(>|t|)
  (Intercept)                   5.1292e+01  6.3787e+00  8.0412 2.136e-15 ***
  year_of_observation          -2.4764e-02  3.2031e-03 -7.7312 2.259e-14 ***
  factor(bac)8                  1.6779e-02  1.8205e-02  0.9217 0.3568809
  factor(bac)none               2.0124e-02  2.0031e-02  1.0047 0.3152635
  per_se_laws                  -6.9043e-02  1.8225e-02 -3.7884 0.0001592 ***
  primary_seatbelt_law         -2.8268e-02  2.7342e-02 -1.0338 0.3014188
  secondary_seatbelt_law        8.4095e-03  1.6847e-02  0.4992 0.6177549
  speed_limit_70plus            4.3623e-02  2.2466e-02  1.9417 0.0524050 .
  graduated_drivers_license_law 2.8001e-02  2.1487e-02  1.3032 0.1927615
  pct_population_14_to_24       3.1154e-02  8.5074e-03  3.6620 0.0002613 ***
  unemployment_rate            -2.7074e-02  2.9911e-03 -9.0515 < 2.2e-16 ***
```

Table 4: Random Effects Model

	Dependent variable:
year_of_observation	-0.025^{***}
	(0.002)
factor(bac)8	0.017
	(0.011)
factor(bac)none	0.020^{*}
	(0.011)
per_se_laws	-0.069^{***}
	(0.011)
primary_seatbelt_law	-0.028^{*}
	(0.016)
secondary_seatbelt_law	0.008
	(0.011)
speed_limit_70plus	0.044^{***}
	(0.011)
graduated_drivers_license_law	0.028^{**}
	(0.012)
pct_population_14_to_24	0.031^{***}
	(0.004)
unemployment_rate	-0.027^{***}
	(0.002)
vehicle_miles_per_capita	0.0001^{***}
	(0.00000)
Constant	51.292^{***}
	(2.997)
Observations	1,200
R^2	0.670
Note:	*p<0.1; **p<0.05; ***p<0.01

```
vehicle_miles_per_capita       6.8163e-05  1.4109e-05  4.8312 1.534e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.0.1 Assumptions of Random Effects

The first assumption of the random effect model is that there are no perfect linear relationships among the explanatory variables.

```
library(car)
car::vif(random.effect.model)
```

```
                                  GVIF Df GVIF^(1/(2*Df))
year_of_observation           15.198952  1        3.898583
factor(bac)                   2.634612  2        1.274029
per_se_laws                   2.089643  1        1.445560
primary_seatbelt_law          2.247484  1        1.499161
secondary_seatbelt_law        2.955369  1        1.719119
speed_limit_70plus            2.258779  1        1.502923
graduated_drivers_license_law 2.245655  1        1.498551
pct_population_14_to_24       4.829649  1        2.197646
unemployment_rate             1.898915  1        1.378011
vehicle_miles_per_capita      5.852089  1        2.419109
```

We see high values for percent_pop_aged_14_to_24, vehicle_miles_per_capita indicating the possible presence of multicollinearity in these variables.

The second assumption is that there is no correlation between the unobserved random and fixed effects and the explanatory variables. Using a random effects model imposes the error structure that the error term** v_{it} **is equal to the sum of variation between groups and variation within groups onto the model residuals, allowing to properly specify the residuals and more efficiently estimate the coefficients of interest. This requires the assumption of independence between random effects and the other predictors in the model. The assumptions for the fixed effect model are discussed above, the additional assumption of independence of random effects and other predictors in the model is evaluated below. The test we run is the Hausman Test for fixed versus random effects. The null hypothesis is that the random effects model is acceptable while the alternative hypothesis is that there is correlation between residuals and predictors, meaning that we should use the FE model.**

We conduct a Hausman test for random vs. fixed effects using `phtest`. We perform this test with an $\alpha = 0.05$

```
res <- phtest(within.model, random.effect.model)
```

With a p-value of $5.2125866 \times 10^{-12}$ less than α, we reject the null hypothesis that random effects are appropriate, suggesting that we should use the fixed models. The random effects model is not likely to be consistent in this case.

The third assumption is that of homoskedastic errors, which we can test below using the Breusch-Pagan Lagrange Multiplier for random effects. Null is no panel effect:

```
plmtest(random.effect.model)
```

```
    Lagrange Multiplier Test - (Honda)

data:  log(total_fatalities_rate) ~ year_of_observation + factor(bac) +  ...
normal = 73.795, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Here we failed to reject the null and conclude that random effects is not appropriate.

6.0.2 Note on Assumptions

As we have seen that the assumptions for random effect model are not met. If we were to inappropriately estimate a random effect model, we would be incorrectly assuming that the random effects and other predictors are independent of one another. This would lead to omitted variable bias as the correlation between the random effects and the explanatory variables of interest would not allow for accurate estimation of the coefficient. Standard errors will also be biased as we are assuming that the random effects, which are included in the error term, are incorrectly uncorrelated with the predictors - given that there is correlation, this will introduce bias into the standard errors.

7 Model Forecasts

7.0.1 Data on Vehicle Miles Traveled

We have downloaded population data from https://fred.stlouisfed.org/series/POPTHM and vehicle driven data from https://fred.stlouisfed.org/series/TRFVOLUSM227NFWA. Population includes resident population plus armed forces overseas. The monthly estimate is the average of estimates for the first of the month and the first of the following month. Vehicle Miles Traveled and the 12-Month Moving Vehicle Miles Traveled series are created by appending the recent monthly figures from the FHWA's Traffic Volume Trends to their Historic Monthly Vehicle Miles Traveled (VMT) data file. We have defined the pandemic period between March 2020 through March 2021 when the Covid vaccine became widely available.

```
library(fredr)

fredr_set_key("cd565a10e83d56f9f1150d5a2c067e2a")

data.vhcl <- fredr(
                   series_id = "TRFVOLUSM227NFWA",
                   observation_start = as.Date("2018-01-01"),
                   observation_end = as.Date("2023-08-01")
                  ) %>% dplyr::select(date,value) %>% as_tsibble(index = date)
data.pop <- fredr(
                  series_id = "POPTHM",
                  observation_start = as.Date("2018-01-01"),
                  observation_end = as.Date("2023-08-01")
                 )%>% dplyr::select(date, value) %>% as_tsibble(index = date)

# Merge vehicle miles driven and population data
data.temp <- merge(x = data.vhcl, y = data.pop, by = "date")

# Calculate vehicle miles per capita
data.temp <- data.temp %>%
  mutate(vehicle_miles_per_capita = 1000 * value.x / value.y)

data.vhcl.ml.per.capita <- data.temp[, c('date', 'vehicle_miles_per_capita')]
data.vhcl.ml.per.capita$year = year(data.vhcl.ml.per.capita$date)
data.vhcl.ml.per.capita$month = month(data.vhcl.ml.per.capita$date)

data.vhcl.ml.per.capita <- data.vhcl.ml.per.capita %>%
  mutate(group = ifelse(year < 2020, "pre-pandemic",
                ifelse(year == 2020 | year == 2021, "pandemic",
```

```
                      "post-pandemic")))


data.pre.pandemic <- data.vhcl.ml.per.capita %>%
  filter(year == 2018)
data.pandemic <- data.vhcl.ml.per.capita %>%
  filter(year == 2020 | year == 2021)
data.pandemic.arranged <- data.pandemic %>%
  arrange(month)

vehicle_miles_per_capita.diff <-
  data.pre.pandemic$vehicle_miles_per_capita -
  data.pandemic.arranged$vehicle_miles_per_capita

drive_pandemic <- data.pandemic %>% slice(3:15)
data.pandemic$group <- 'pandemic'
data.pre.pandemic$group <- 'pre_pandemic'
data.pandemic.pre.post.comparison <- rbind(data.pre.pandemic, data.pandemic)

plot.orig <- data.vhcl.ml.per.capita %>%
  ggplot(aes(x = date, y = vehicle_miles_per_capita, color=group)) +
  geom_line() + xlab("Date") + ylab("Veh. Miles per Capita") +
  ggtitle('Miles Driven during Pandemic')

plot.comparison <- ggplot(
  data.pandemic.pre.post.comparison,
         aes(x = month,
             y = vehicle_miles_per_capita,
             group = group)) +
  geom_point(aes(color=group)) +
  geom_smooth(aes(color=group)) +
  xlab('Month') +
  ylab('Miles Driven') +
  ggtitle('Pre/Post Pandemic Miles Driven')

(plot.orig | plot.comparison)
```

The COVID-19 pandemic dramatically changed patterns of driving. Find data (and include this data in your analysis, here) that includes some measure of vehicle miles driven in the US. Your data should at least cover the period from January 2018 to as current as possible. With this data, produce the following statements:

- Comparing monthly miles driven in 2018 to the same months during the pandemic:
  - What month demonstrated the largest decrease in driving? How much, in percentage terms, lower was this driving?
  - What month demonstrated the largest increase in driving? How much, in percentage terms, higher was this driving?

Now, use these changes in driving to make forecasts from your models.

- Suppose that the number of miles driven per capita, increased by as much as the COVID boom. Using the FE estimates, what would the consequences be on the number of traffic fatalities? Please interpret the estimate.

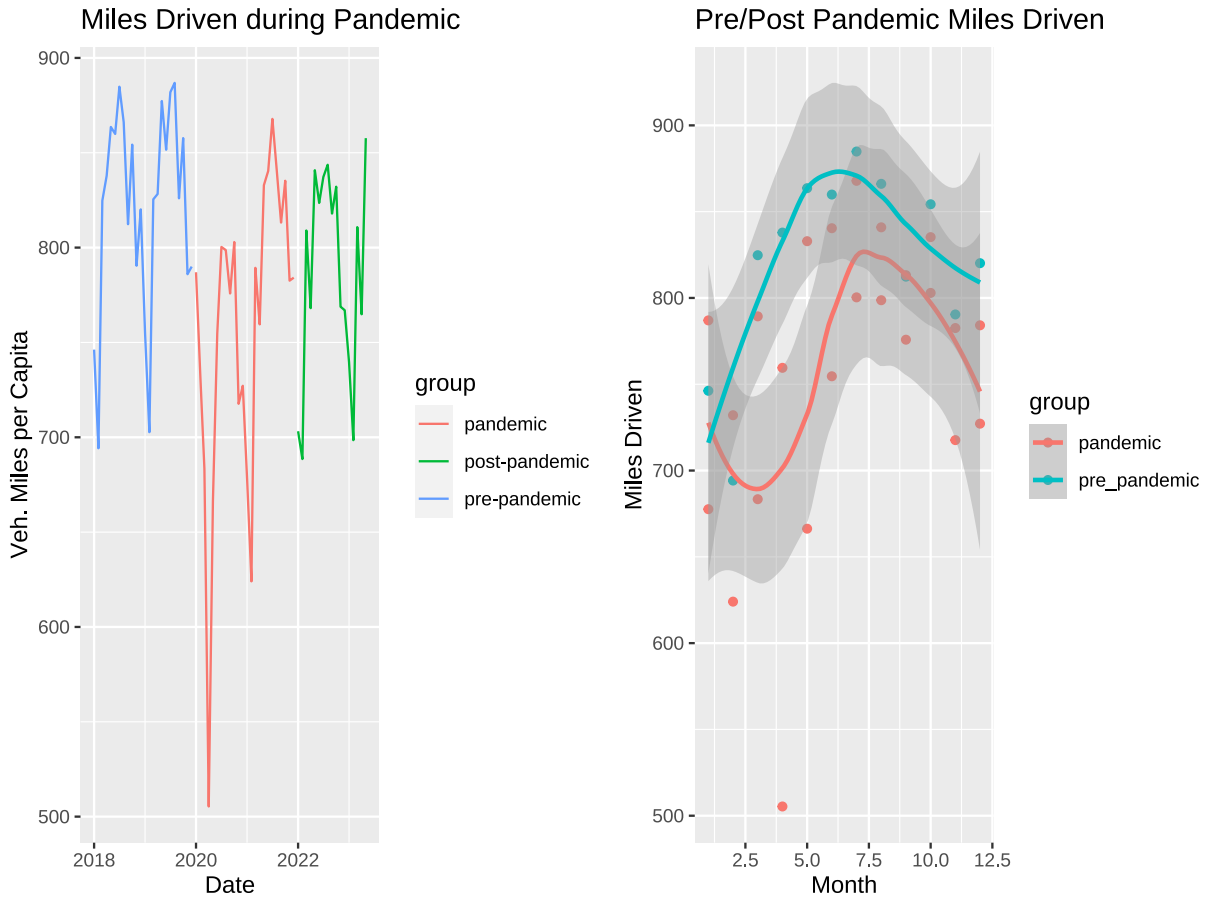

Figure 8: (#fig:download external data)Vehicle Miles Traveled Series from the St. Louis Fed

8 Evaluate Error

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