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?"**

2 Build and Describe the Data

2.0.1 Load the data and produce useful features.

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

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

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

We renamed the variables to more legible names.

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.

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 Mexico 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.

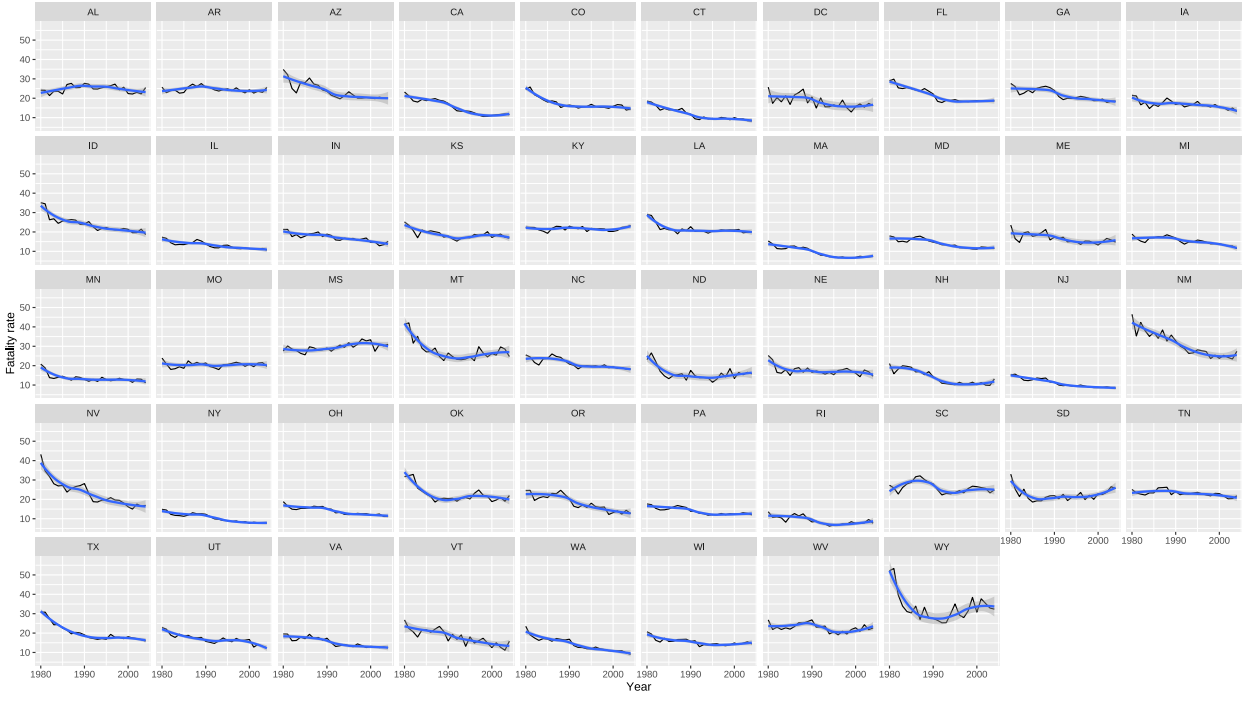

Figure 1: Fatality Rate by State (where data recorded) and Year Since 1980

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.

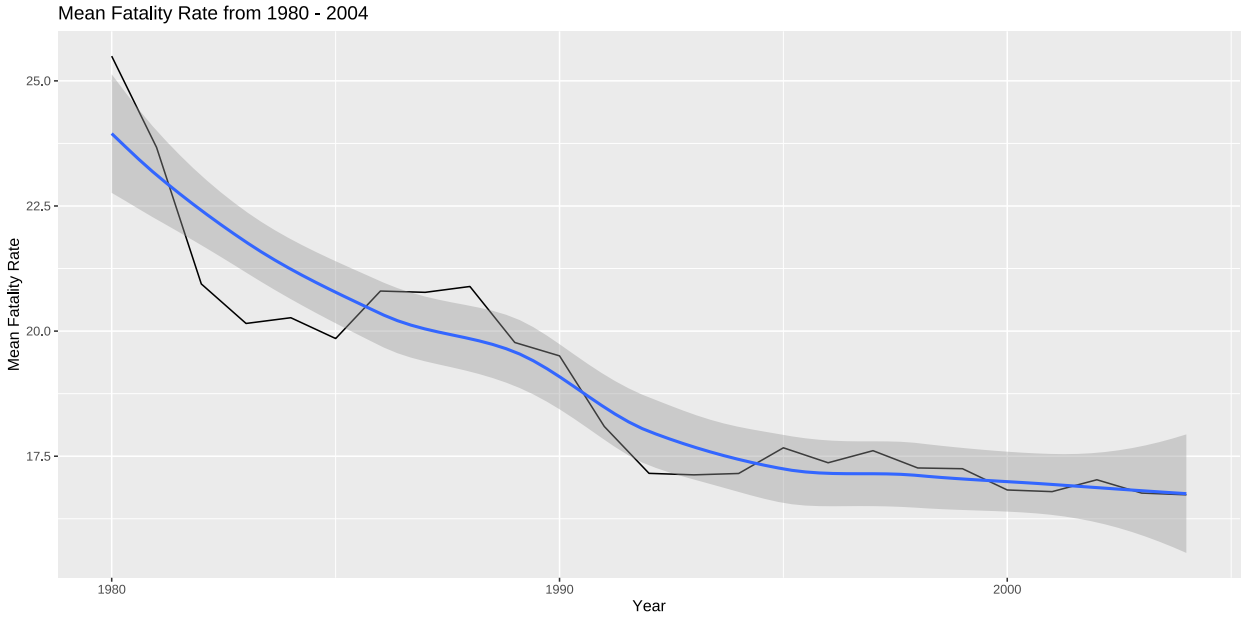

Figure 2: Time Series of Average Fatality

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 western states like Wyoming and New Mexico are states that appear to have higher rates of fatality, while NE states Connecticut, Massachussets, and Rhode Island appear to be states with lower fatality rates.

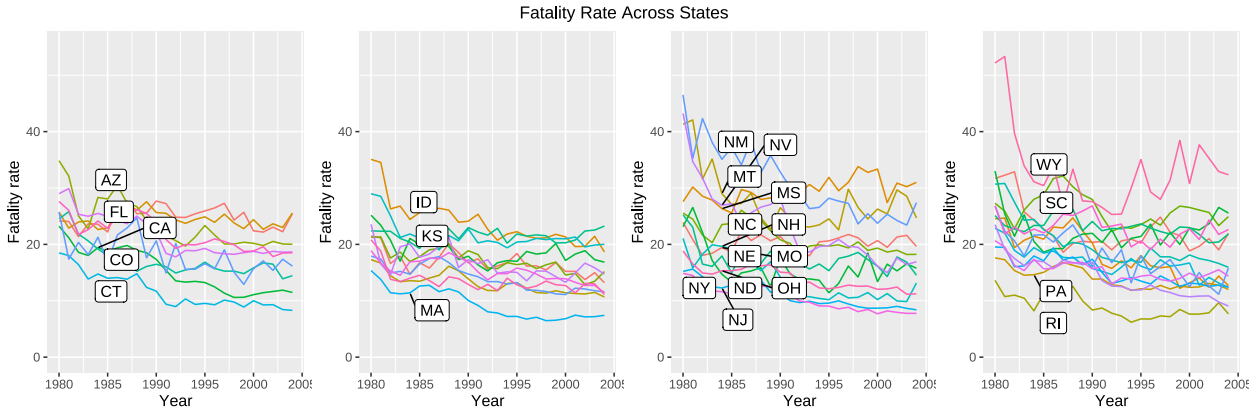

Figure 3: Fatality Rate by State on Common Axis

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. 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.

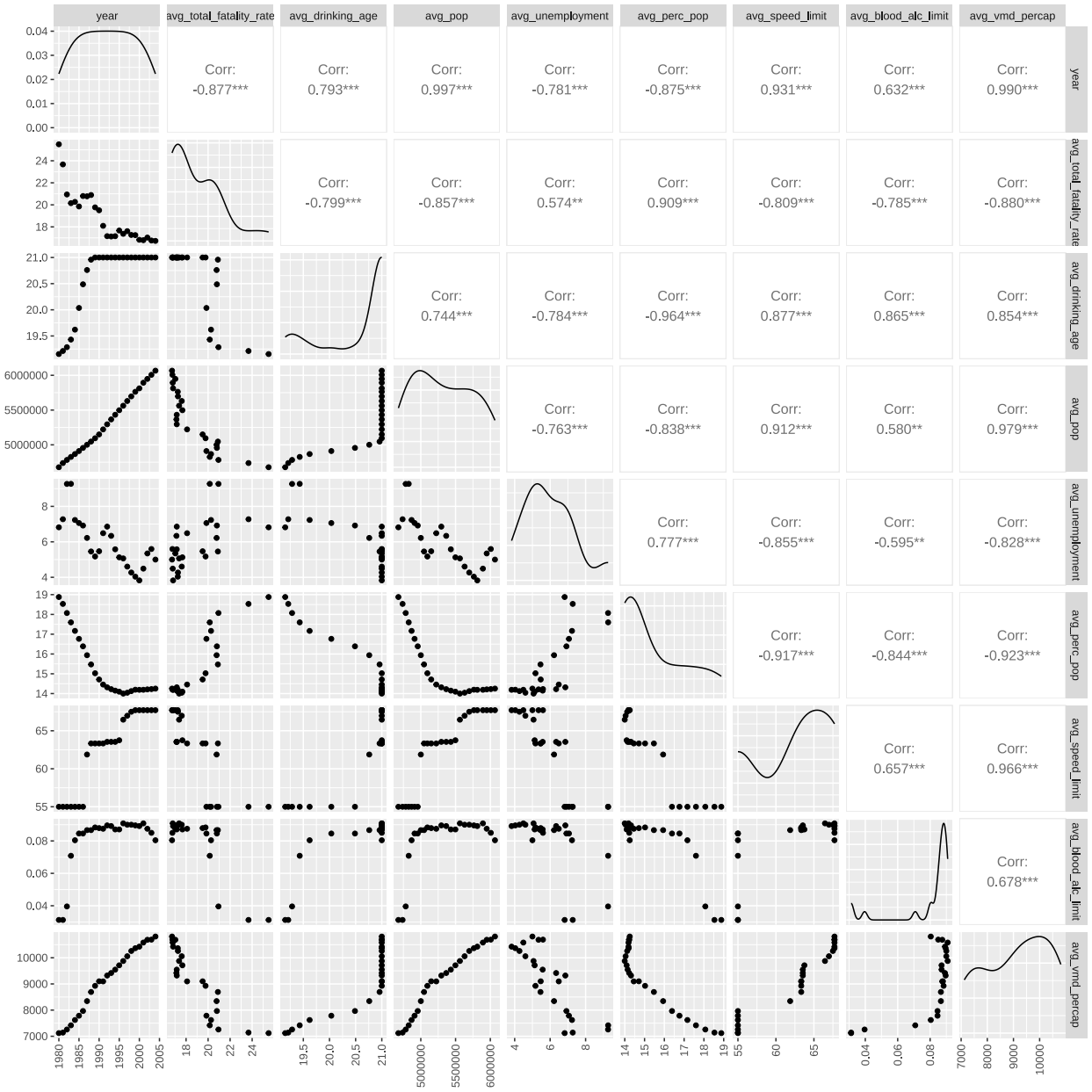

Figure 4: Scatterplot Matrix of Variables of Interest

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

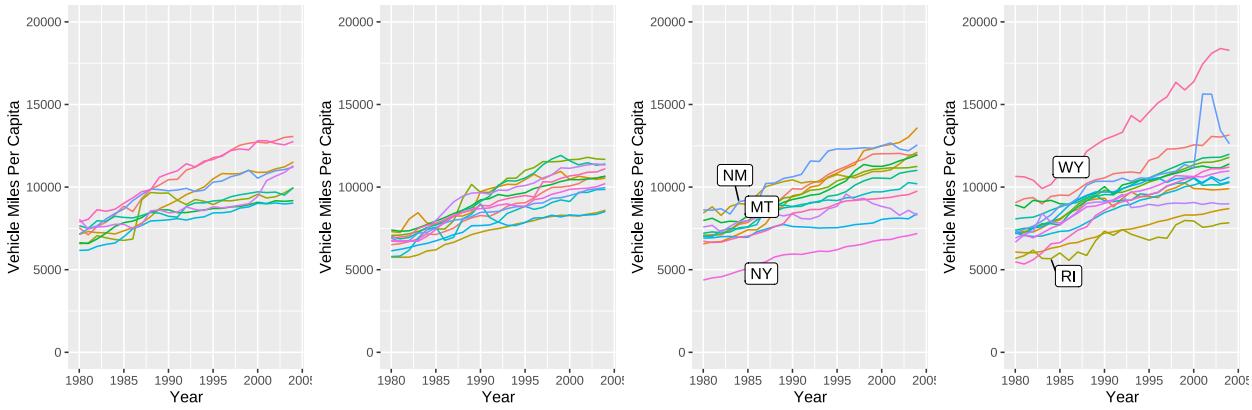

Figure 5: Vehicle Miles Driven Per Capita

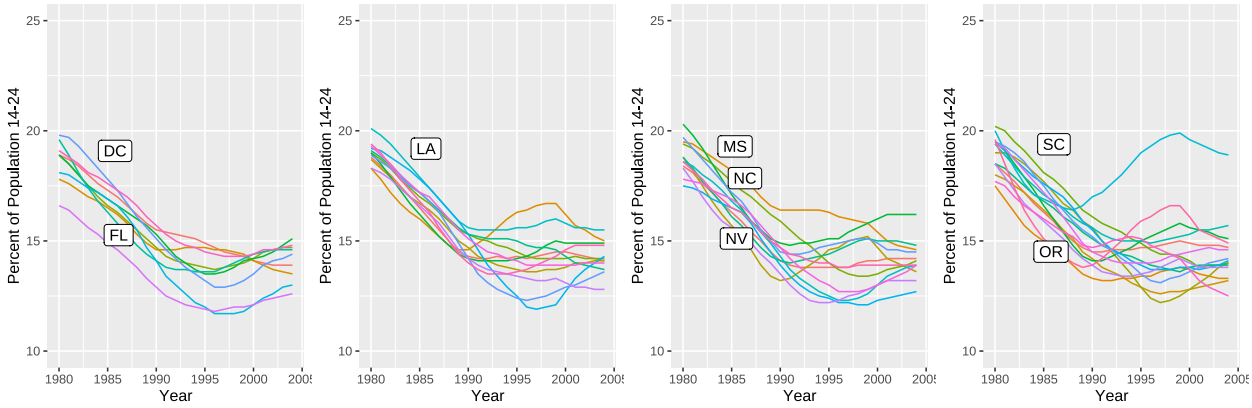

Figure 6: Percent of Population Aged 14-24

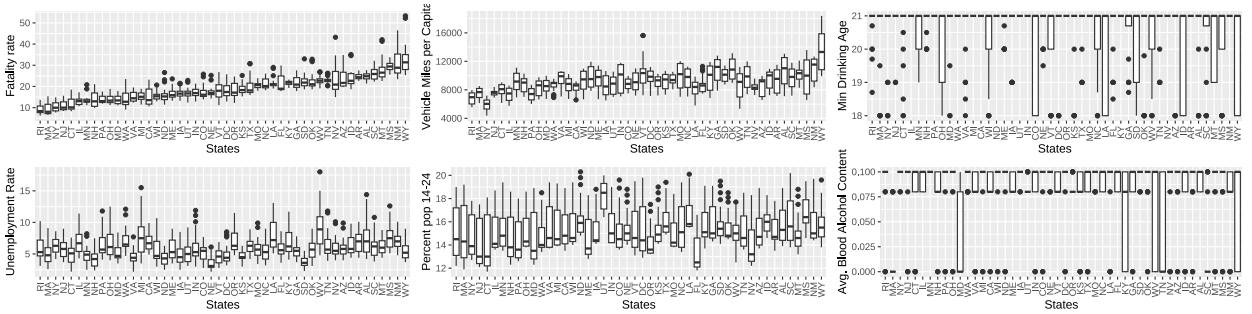

Figure 7: Box Plot of States - Sorted by Total Fatality Rate

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 use 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(
  log(total_fatalities_rate) ~ year,
  index = c("year","state"),
  model = "pooling",
  data=data
)

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

3.0.2 Discussion on Linear Model

We incorporated dummy variables for the year and across all states. The model doesn't include any other variable, so omitted variable bias could be present. However, the results imply the year is statistically significant and a negative coefficient is present across all years. This is consistent with our observations in the EDA, where the fatality rate was shown to decrease over time. Driving became safer over this period. The dummy variable results are shown in **Table 1**, excluding each of the dummy year variables

4 Expanded Model

4.0.1 Modeling

We used the Shapiro-Wilks test to test for normality with the total fatalities rate, percent of population 14-24, unemployment, and vehicle miles per capita variables. The test resulted in very small p-values for each, so we decided to log transform these variables in our expanded model.

```
# 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 + year +
    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("year","Abbreviation"),
  model = "pooling",
  data=data
)

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

Based on the expanded model, the vehicle miles driven per capita, unemployment rate, and percent of population between 14 and 24 were all highly statistically significant. Both per se laws and primary seat belt laws had negative effects on fatality rate. Additionally, per se laws and minimum drinking age laws had less significant but still negative effects on fatality rates.

The blood alcohol content (BAC) variable represents the alcohol level at which drivers are considered legally intoxicated. BAC is calculated by Alcohol consumed in grams divided by body weight in grams, then multiplied by a constant that represents biological sex. This number is then multiplied by 100 to give a final measurement of grams per 100mL of blood. The coefficient estimate for the blood alcohol content (BAC) variable in our expanded model is -0.0042982. The negative relationship suggests that stricter blood alcohol content laws may be associated with a decrease in the log-transformed total fatalities rate. However, this impact was not found to be statistically significant (p-value = 0.982275).

The results are shown alongside the preliminary model in **Table 1**, excluding each of the dummy year variables

5 State-Level Fixed Effects

5.0.1 Modeling

```
if(!"kableExtra"%in%rownames(installed.packages())) {install.packages("kableExtra")}
library(kableExtra)


# Fixed effects
within.model <- pvcm(
  log(total_fatalities_rate) ~ minimum_drinking_age +
    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),
  data = data,
  index=c("Abbreviation", "year"),
  effect="individual",
  model="within"
```

Table 1: Comparison of Linear Models

	Dependent variable:	
	Prelim	Expanded
	(1)	(2)
blood_alc		−0.004
		(0.193)
per_se_laws		−0.032**
		(0.014)
primary_seatbelt_law		0.001
		(0.025)
secondary_seatbelt_law		0.026
		(0.021)
speed_lim_70		0.217***
		(0.022)
graduated_drivers_license_law		−0.026
		(0.026)
log(pct_population_14_to_24)		0.295***
		(0.093)
log(unemployment_rate)		0.269***
		(0.024)
log(vehicle_miles_per_capita)		1.528***
		(0.045)
Observations	1,200	1,200
R^2	0.126	0.668
Note:	*p<0.1; **p<0.05; ***p<0.01	

)

State level fixed effect models are estimated using the `pvcm` function in R, which takes advantage of the panel data structure for model estimations and allows for individual state effects to be estimated on the control variables found in the expanded model. These individual effects were too large to present in a table format so instead they are presented as distributions in **Figure 8**.

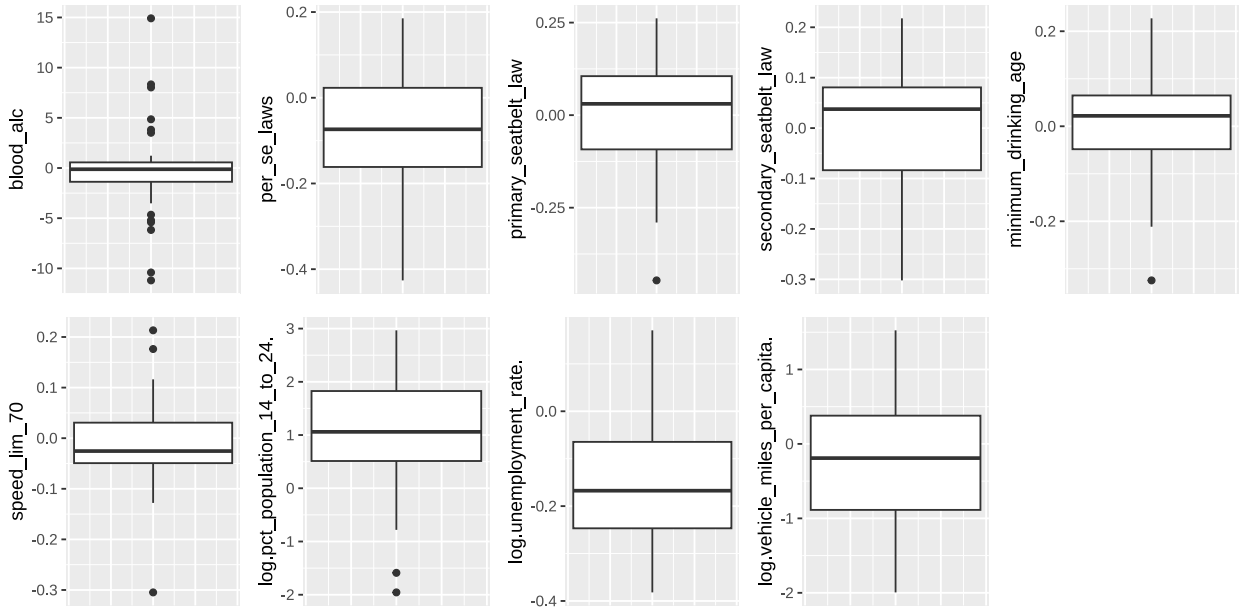

Figure 8: Estimated Coefficients Across State - Fixed Effect Model

5.0.2 Fixed Effects Discussion

The estimated coefficients for the model are presented in **Figure 8** across the states queried in this dataset (continental US). One can see the variability present in the estimates, but clearly some parameters, like the young population and the unemployment rate do not have an IQR which contains zero, indicating a significant effect.

5.0.2.1 Blood Alcohol Affects

The estimated effects of the blood alcohol content tend to be on the negative end, with an average of -0.3337, with 1Q being -1.37 and 3Q being 0.56 (greater than zero). The fact that the range across states encompasses zero shows this parameter may not be very impactful. This is consisent with the OLS results, which generally indicated a small impact from this variable and was ultimately not statisticaly significant.

5.0.2.2 Per se Laws

The estimated effects of per-se laws were much more biased negative than the BAC content just described. For the per-se laws, the mean effect was a coefficient of -0.07, with a 1Q being -0.16 and 3Q being 0.02. This meant that, on average, for a change from 0 to 1, it meant a 7.25% reduction in the fatality rate, per 100K. The pooled models found the per-se laws to also be significant, but with a higher impact.

5.0.2.3 Primary Seat Belt Laws

The estimated effects of primary seat belt laws were found to be on average -0.01, with a 1Q of -0.09 and a 3Q of 0.10. This does not appear to be impactful in the fixed effects model controlling for each state. This is consistent with the results of the expanded model from the previous section.

5.0.3 Reliability of Estimates

5.0.3.1 Fixed effect model assumptions

- For each 'i' the model is $y_{it} = \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it}, t = 1..T$
- We have a random sample from the cross section
- Each explanatory variable changes over time(for at least some time) and no perfect linear relationship exists between explanatory variables
- For each t, the expected value of the idiosyncratic error given the explanatory ariables in all time periods and the unobserved effect is zero: $E(u_{it}|X_i, a_i) = 0$
- The variance of the difference errors, conditional on all explanatory variables, is constant $Var(\triangle u_{it}|X_i) = \sigma_u^2, t = 2, ..., T$. This is required for homoskedastic errors
- For all $t \neq s$, the differences in the idiosyncratic errors are uncorrelated(conditional on all explanatory variables). This is for serially uncorrelated residuals

5.0.3.2 Conclusion on Reliability

The assumption in an pooled OLS model is that the data is IID. Here in the data set, a sample of a large population is collected on different years. It is unlikely that a particular individual sample data point is measured twice. In such a circumstance a pooled OLS model would be applicable. However in this data set, data granularity is at the state level and the same state is measured multiple times across years. This violates the assumption of IID in the pooled OLS. A fixed effect model is then expected to be a better model in this scenario.

6 Random Effects Model

6.0.1 Modeling

```
random.effect.model <- plm(
  log(total_fatalities_rate) ~ minimum_drinking_age +
    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("year","state"),
  model = "random",
  data=data
)
```

6.0.2 Assumptions of Random Effects

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

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 $\alpha = 0.05$

With a p-value of $3.4618701 \times 10^{-50}$ 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 using the Breusch-Pagan Lagrange Multiplier for random effects. Null is no panel effect. In our test, we are able to reject the null hypothsis, again indicating the panel data structure.

6.0.3 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

Table 2: Random Effects Model

	Dependent variable:
blood_alc	−0.187
	(0.210)
per_se_laws	−0.052***
	(0.016)
primary_seatbelt_law	−0.115***
	(0.025)
secondary_seatbelt_law	−0.078***
	(0.021)
speed_lim_70	0.077***
	(0.021)
graduated_drivers_license_law	−0.201***
	(0.021)
log(pct_population_14_to_24)	1.183***
	(0.082)
log(unemployment_rate)	0.261***
	(0.024)
log(vehicle_miles_per_capita)	1.250***
	(0.046)
Observations	1,200
R^2	0.554
Note:	*p<0.1; **p<0.05; ***p<0.01

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.

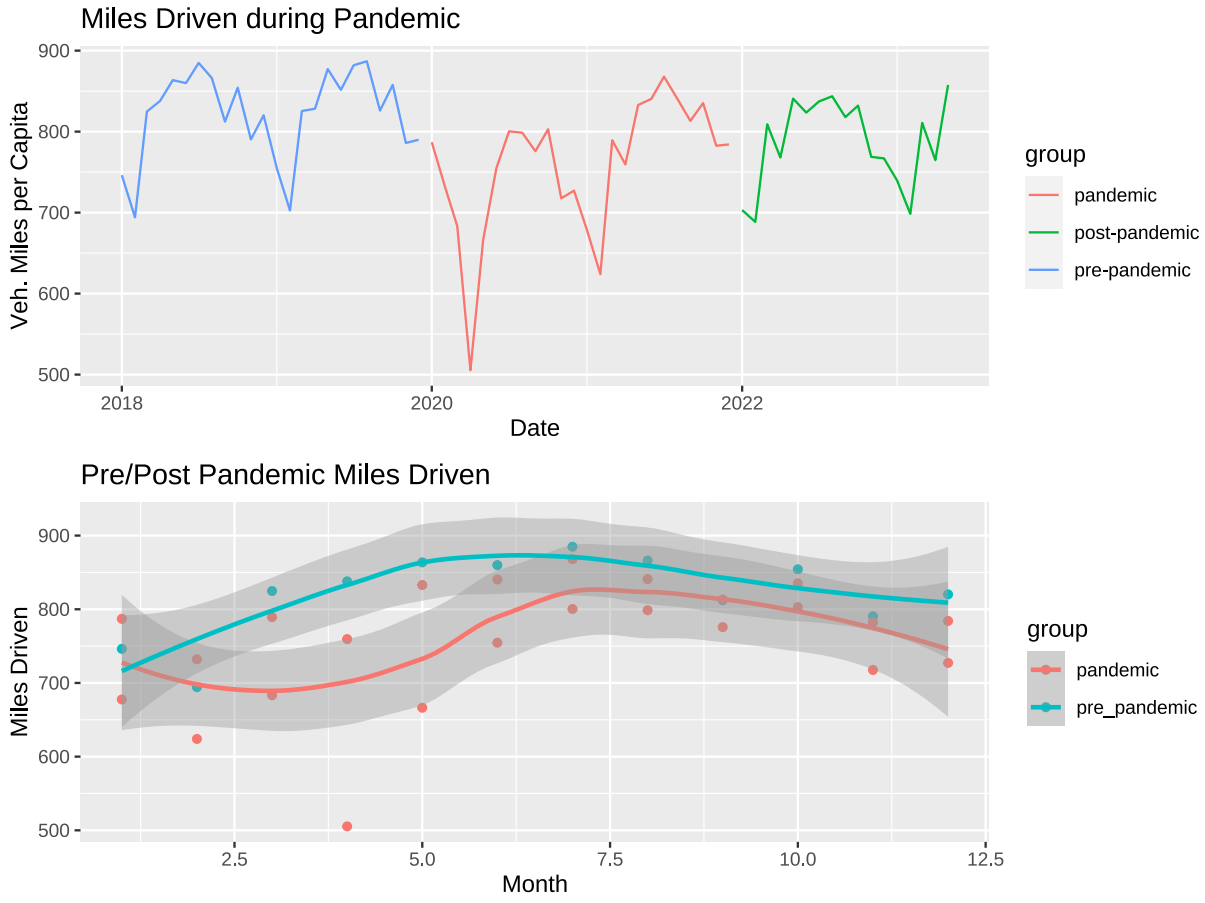

Figure 9: Vehicle Miles Traveled Series from the St. Louis Fed

7.0.2 Forecasting changes in driving

The pandemic caused a rapid decrease in the vehicle miles traveled per capita. To forecast the impact of this decrease, we assessed the pre-pandemic peak to the pandemic lull - which was a

decrease of around 0.4 percent. This is reflected in **Figure 9**, which shows the pre-pandemic data population peaking at 884.8 miles per capita in late 2019 decreaseing to 505.4 miles per capita in spring 2020.

We estimate the impact to the fatality rate by applying the percentage decrease to the coefficients estimated by the fixed effects model. Our modeling choice, which regressed the log of the fatality rate on the log of the vehicle miles leads to easy math.

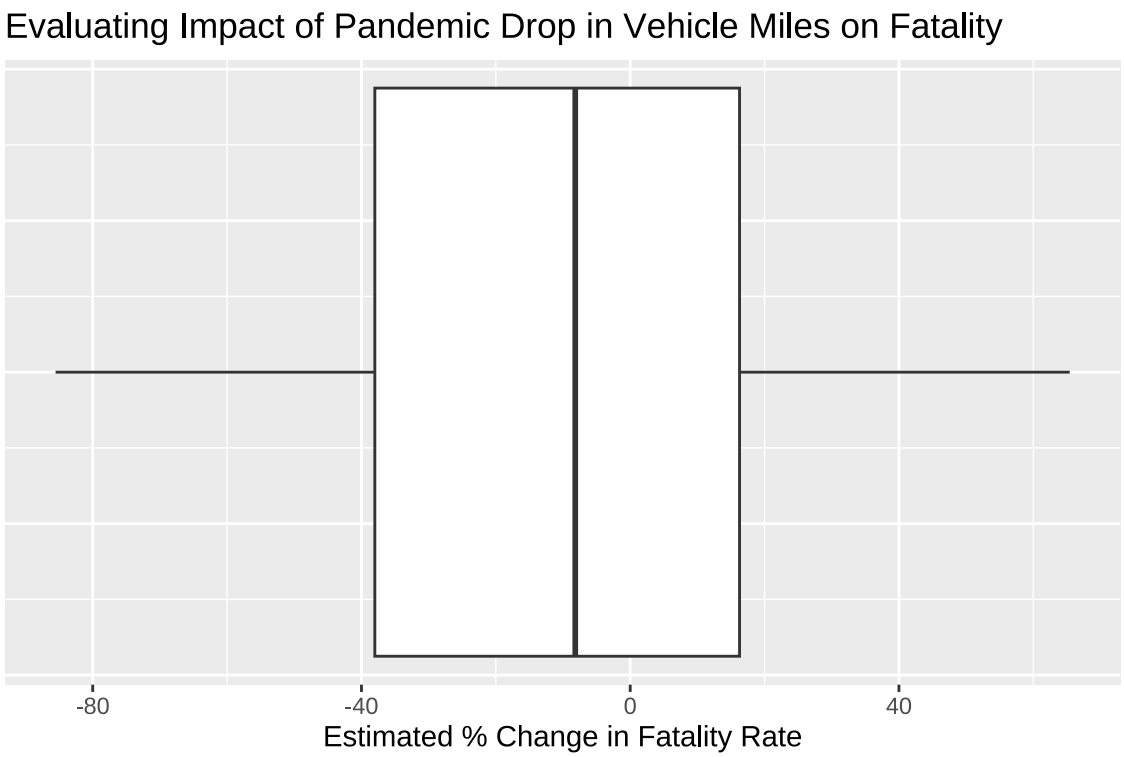

Figure 10: Estimating Impact of Miles Traveled Drop During Pandemic

The mean estimated % change in fatality rate from the drop in vehicle miles traveled is -7%, with a standard deviation of37%.

8 Evaluate Error

8.0.1 Consequences of Serial Correlation / Heteroskadicity

According to literature, the consequences of serial correlation and heteroskedacity in panel data models is a loss in efficiency of the estimates (Jianhong). This means that the true coefficients relative to the estimated coefficients likely have higher variance than what is currently estimated in the standard errors.

8.0.2 Are there serial correlations or heteroskedasticity?

Yes, while testing the random effects model, we conducted the Lagrange Multiplier test and rejected the null hypothesis. We also conducted a Hausman test and rejected the null hypothesis. By rejecting the null hypothesis in both cases, we can conclude that there exists both serial correlation and heteroskedasticity.

9 References

(1) Jianhong Wu, A joint test for serial correlation and heteroscedasticity in fixed-T panel regression models with interactive effects, Economics Letters, Volume 197, 2020, 109594, ISSN 0165-1765, https://doi.org/10.1016/j.econlet.2020.109594. (https://www.sciencedirect.com/science/article/pii/S0165176520303578)