

# Lab 3: Panel Models

## US Traffic Fatalities: 1980 - 2004

### U.S. traffic fatalities: 1980-2004

In this lab, we are asking you to answer the following **causal** question:

**“Do changes in traffic laws affect traffic fatalities?”**

To answer this question, please complete the tasks specified below using the data provided in `data/driving.Rdata`. This data includes 25 years of data that cover changes in various state drunk driving, seat belt, and speed limit laws.

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

```
load(file="./data/driving.RData")

## please comment these calls in your work
head(data)
```

##	year	state	sl55	sl65	sl70	sl75	slnone	seatbelt	minage	zerotol	gdl	bac10	bac08					
## 1	1980	1	1	0	0	0	0	0	18	0	0	1	0					
## 2	1981	1	1	0	0	0	0	0	18	0	0	1	0					
## 3	1982	1	1	0	0	0	0	0	18	0	0	1	0					
## 4	1983	1	1	0	0	0	0	0	18	0	0	1	0					
## 5	1984	1	1	0	0	0	0	0	18	0	0	1	0					
## 6	1985	1	1	0	0	0	0	0	20	0	0	1	0					
##	perse	totfat	nghtfat	wkndfat	totfatpvm	nghtfatpvm	wkndfatpvm	statepop										
## 1	0	940	422	236	3.20	1.437	0.803	3893888										
## 2	0	933	434	248	3.35	1.558	0.890	3918520										
## 3	0	839	376	224	2.81	1.259	0.750	3925218										
## 4	0	930	397	223	3.00	1.281	0.719	3934109										
## 5	0	932	421	237	2.83	1.278	0.720	3951834										
## 6	0	882	358	224	2.51	1.019	0.637	3972527										
##	totfatrte	nghtfatrte	wkndfatrte	vehicmiles	unem	perc14_24	sl70plus	sbprim										
## 1	24.14	10.84	6.06	29.37500	8.8	18.9	0	0										
## 2	24.07	11.08	6.33	27.85200	10.7	18.7	0	0										
## 3	21.37	9.58	5.71	29.85765	14.4	18.4	0	0										
## 4	23.64	10.09	5.67	31.00000	13.7	18.0	0	0										
## 5	23.58	10.65	6.00	32.93286	11.1	17.6	0	0										
## 6	22.20	9.01	5.64	35.13944	8.9	17.3	0	0										
##	sbsecon	d80	d81	d82	d83	d84	d85	d86	d87	d88	d89	d90	d91	d92	d93	d94	d95	d96
## 1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
## 2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
## 3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
## 4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

```

## 5      0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
## 6      0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
##      d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc
## 1      0  0  0  0  0  0  0  0      7543.874
## 2      0  0  0  0  0  0  0  0      7107.785
## 3      0  0  0  0  0  0  0  0      7606.622
## 4      0  0  0  0  0  0  0  0      7879.802
## 5      0  0  0  0  0  0  0  0      8333.562
## 6      0  0  0  0  0  0  0  0      8845.614

```

desc

```

##      variable                                label
## 1      year                                1980 through 2004
## 2      state                                48 continental states, alphabetical
## 3      sl55                                speed limit == 55
## 4      sl65                                speed limit == 65
## 5      sl70                                speed limit == 70
## 6      sl75                                speed limit == 75
## 7      slnone                              no speed limit
## 8      seatbelt      =0 if none, =1 if primary, =2 if secondary
## 9      minage                                minimum drinking age
## 10     zerotol                                zero tolerance law
## 11     gdl                                graduated drivers license law
## 12     bac10                                blood alcohol limit .10
## 13     bac08                                blood alcohol limit .08
## 14     perse administrative license revocation (per se law)
## 15     totfat                                total traffic fatalities
## 16     nghtfat                                total nighttime fatalities
## 17     wkndfat                                total weekend fatalities
## 18     totfatpvm      total fatalities per 100 million miles
## 19     nghtfatpvm      nighttime fatalities per 100 million miles
## 20     wkndfatpvm      weekend fatalities per 100 million miles
## 21     statepop                                state population
## 22     totfatrte      total fatalities per 100,000 population
## 23     nghtfatrte      nighttime fatalities per 100,000 population
## 24     wkndfatrte      weekend accidents per 100,000 population
## 25     vehicmiles      vehicle miles traveled, billions
## 26     unem                                unemployment rate, percent
## 27     perc14_24      percent population aged 14 through 24
## 28     sl70plus                                sl70 + sl75 + slnone
## 29     sbprim                                =1 if primary seatbelt law
## 30     sbsecon                                =1 if secondary seatbelt law
## 31     d80                                =1 if year == 1980
## 32     d81
## 33     d82
## 34     d83
## 35     d84
## 36     d85
## 37     d86
## 38     d87
## 39     d88
## 40     d89
## 41     d90
## 42     d91

```

```
## 43          d92
## 44          d93
## 45          d94
## 46          d95
## 47          d96
## 48          d97
## 49          d98
## 50          d99
## 51          d00
## 52          d01
## 53          d02
## 54          d03
## 55          d04                      =1 if year == 2004
## 56 vehicmilespc
```

## (30 points, total) Build and Describe the Data

- (5 points) Load the data and produce useful features. Specifically:
  - Produce a new variable, called `speed_limit` that re-encodes the data that is in `sl55`, `sl65`, `sl70`, `sl75`, and `slnone`;
  - Produce a new variable, called `year_of_observation` that re-encodes the data that is in `d80`, `d81`, `...`, `d04`.
  - Produce a new variable for each of the other variables that are one-hot encoded (i.e. `bac*` variable series).
  - Rename these variables to sensible names that are legible to a reader of your analysis. For example, the dependent variable as provided is called, `totfatrte`. Pick something more sensible, like, `total_fatalities_rate`. There are few enough of these variables to change, that you should change them for all the variables in the data. (You will thank yourself later.)
- (5 points) Provide a description of the basic structure of the dataset. What is this data? How, where, and when is it collected? Is the data generated through a survey or some other method? Is the data that is presented a sample from the population, or is it a *census* that represents the entire population? Minimally, this should include:
  - How is the our dependent variable of interest `total_fatalities_rate` defined?
- (20 points) Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable `total_fatalities_rate` and the potential explanatory variables. Minimally, this should include:
  - How is the our dependent variable of interest `total_fatalities_rate` defined?
  - What is the average of `total_fatalities_rate` in each of the years in the time period covered in this dataset?

As with every EDA this semester, the goal of this EDA is not to document your own process of discovery – save that for an exploration notebook – but instead it is to bring a reader that is new to the data to a full understanding of the important features of your data as quickly as possible. In order to do this, your EDA should include a detailed, orderly narrative description of what you want your reader to know. Do not include any output – tables, plots, or statistics – that you do not intend to write about.

```
#head(data)

#speed limit
data$speed_limit <- with(data,
  ifelse(as.integer(sl55) >= 0.5), 55,
  ifelse(as.integer(sl65) >= 0.5), 65,
  ifelse(as.integer(sl70) >= 0.5), 70,
  ifelse(as.integer(sl75) >= 0.5), 75,
  ifelse(as.integer(slnone) >= 0.5), NA, NA))))
```

```

#year of observation
year_vars <- grep("~d\\d{2}$", names(data), value = TRUE)
data$year_of_observation <- apply(data[, year_vars], 1, function(x) 1980 + which(x == 1) - 1)

#data <- data[, !(names(data) %in% c(year_vars, "sl55", "sl65", "sl70", "sl75", "slnone"))]
data <- data[, !(names(data) %in% c("sl55", "sl65", "sl70", "sl75", "slnone"))]

#make sure they are indicating 0 and 1 respectively
data$bac10 <- round(data$bac10)
data$bac08 <- round(data$bac08)
data$sbprim <- round(data$sbprim)
data$sbsecon <- round(data$sbsecon)

data$minage <- round(data$minage)
data$perse <- round(data$perse)
data$zerotol <- round(data$zerotol)
data$gdl <- round(data$gdl)
data$sl70plus <- round(data$sl70plus)

#rename for clarity
names(data)[names(data) == "totfatpvm"] <- "total_fatalities_pvm"
names(data)[names(data) == "nghtfatpvm"] <- "night_fatalities_pvm"
names(data)[names(data) == "wkndfatpvm"] <- "weekend_fatalities_pvm"
names(data)[names(data) == "totfatrte"] <- "total_fatalities_rate"
names(data)[names(data) == "nghtfatrte"] <- "night_fatalities_rate"
names(data)[names(data) == "wkndfatrte"] <- "weekend_fatalities_rate"
names(data)[names(data) == "unem"] <- "unemployment_rate"

names(data)[names(data) == "gdl"] <- "graduated_drivers_license_law"
names(data)[names(data) == "zerotol"] <- "zero_tolerance_law"
names(data)[names(data) == "totfat"] <- "total_fatalities"
names(data)[names(data) == "nghtfat"] <- "total_nighttime_fatalities"
names(data)[names(data) == "wkndfat"] <- "total_weekend_fatalities"
names(data)[names(data) == "sbprim"] <- "primary_seatbelt_law"
names(data)[names(data) == "sbsecon"] <- "secondary_seatbelt_law"

data$state[data$state == 1] <- 'al'
data$state[data$state == 3] <- 'az'
data$state[data$state == 4] <- 'ar'
data$state[data$state == 5] <- 'ca'
data$state[data$state == 6] <- 'co'
data$state[data$state == 7] <- 'ct'
data$state[data$state == 8] <- 'de'
data$state[data$state == 10] <- 'fl'
data$state[data$state == 11] <- 'ga'
data$state[data$state == 13] <- 'id'
data$state[data$state == 14] <- 'il'
data$state[data$state == 15] <- 'in'
data$state[data$state == 16] <- 'ia'
data$state[data$state == 17] <- 'ks'
data$state[data$state == 18] <- 'ky'

data$state[data$state == 19] <- 'la'

```

```

data$state[data$state == 20] <- 'me'
data$state[data$state == 21] <- 'md'
data$state[data$state == 22] <- 'ma'
data$state[data$state == 23] <- 'mi'
data$state[data$state == 24] <- 'mn'
data$state[data$state == 25] <- 'ms'
data$state[data$state == 26] <- 'mo'
data$state[data$state == 27] <- 'mt'
data$state[data$state == 28] <- 'ne'
data$state[data$state == 29] <- 'nv'
data$state[data$state == 30] <- 'nh'
data$state[data$state == 31] <- 'nj'

data$state[data$state == 32] <- 'nm'
data$state[data$state == 33] <- 'ny'
data$state[data$state == 34] <- 'nc'
data$state[data$state == 35] <- 'nd'
data$state[data$state == 36] <- 'oh'
data$state[data$state == 37] <- 'ok'
data$state[data$state == 38] <- 'or'
data$state[data$state == 39] <- 'pa'
data$state[data$state == 40] <- 'ri'
data$state[data$state == 41] <- 'sc'

data$state[data$state == 42] <- 'sd'
data$state[data$state == 43] <- 'tn'
data$state[data$state == 44] <- 'tx'
data$state[data$state == 45] <- 'ut'
data$state[data$state == 46] <- 'vt'
data$state[data$state == 47] <- 'va'
data$state[data$state == 48] <- 'wa'
data$state[data$state == 49] <- 'wv'
data$state[data$state == 50] <- 'wi'
data$state[data$state == 51] <- 'wy'

```

The traffic fatalities data originates from the Fatality Analysis Reporting System (FARS), managed by the National Highway Traffic Safety Administration (NHTSA). This system gathers data on every traffic crash across the 48 contiguous United States that results in the death of a vehicle occupant or a nonmotorist. The data collection process, conducted by state employees, employs a standardized format to ensure consistency and comparability across different states.

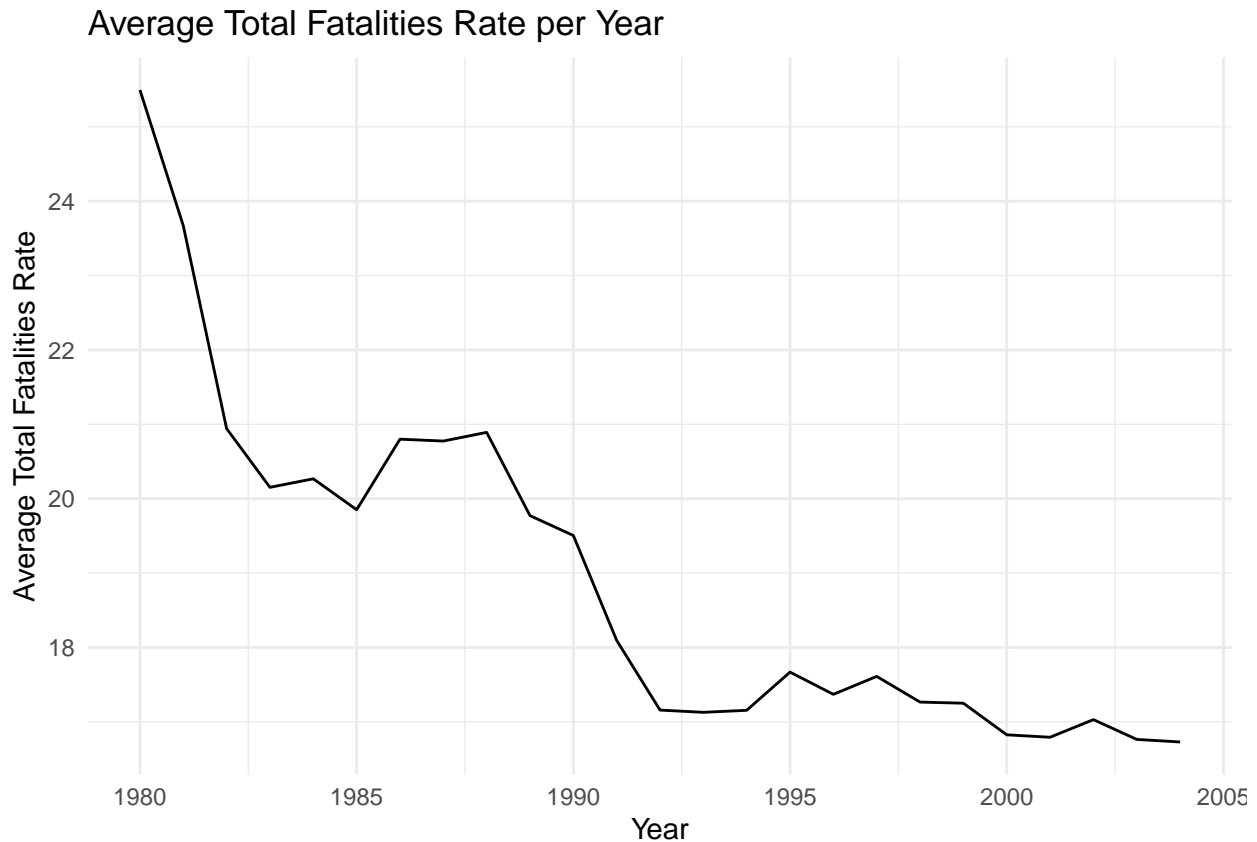
The dependent variable of interest, `total_fatalities_rate`, is defined as the number of traffic fatalities per 100,000 population at the state level over over the years 1980-2004. The dataset consists of approximately 1200 records, reflecting 48 records for each of the 25 years covered.

The independent variables include indicator variables for control legislation, including blood alcohol limit, graduated drivers license law, seat belt, and speed limit laws. Other controls include continuous variables for mileage traveled and demographic characteristics.

This dataset represents a census of traffic fatalities, rather than a sample. It documents every recorded instance of traffic-related fatalities within its scope and time frame, making it a complete record for the study period and not a subset of the total incidents.

```
average_fatalities_per_year <- aggregate(total_fatalities_rate ~ year_of_observation, data, mean)

ggplot(average_fatalities_per_year, aes(x = year_of_observation, y = total_fatalities_rate)) +
  geom_line() +
  labs(title = "Average Total Fatalities Rate per Year",
       x = "Year",
       y = "Average Total Fatalities Rate") +
  theme_minimal()
```



The graph shows a significant downward trend in the average total fatalities rate per year from 1980 to 2004. From the early 1990s onwards, the rate shows some fluctuations but generally remains at a lower rate than at the start of the period observed. This suggests that road safety may have improved, possibly due to policy changes, improved vehicle safety, or other factors.

```
average_by_age <- aggregate(total_fatalities_rate ~ minage, data = data, FUN = mean)
#print(average_by_age)
```

Table 1: Average Total Fatalities by Minimum Drinking Age

Minimum Drinking Age	18	19	20	21
Average Total Fatalities Rate	23.96	21.33	19.76	18.20

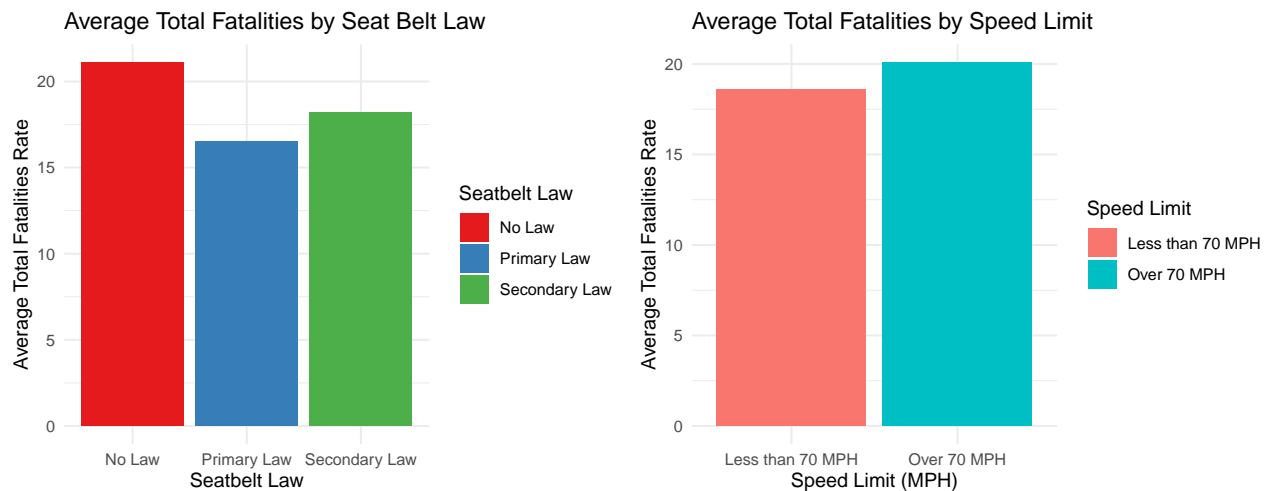
This table suggests that as the minimum legal drinking age increases, the average total fatalities rate decreases. This trend may imply that higher drinking age laws could be contributing to improved road safety, possibly by reducing alcohol consumption among younger drivers.

```
p1_bar <- ggplot(data, aes(x=factor(seatbelt), y=total_fatalities_rate, fill=factor(seatbelt))) +
  geom_bar(stat="summary", fun=mean) +
  labs(title="Average Total Fatalities by Seat Belt Law", x="Seatbelt Law", y="Average Total Fatalities Rate") +
  theme_minimal() +
  scale_fill_brewer(palette="Set1", labels=c("No Law", "Primary Law", "Secondary Law")) +
  scale_x_discrete(labels=c("No Law", "Primary Law", "Secondary Law"))

p2_bar <- ggplot(data, aes(x=factor(sl70plus), y=total_fatalities_rate, fill=factor(sl70plus))) +
  geom_bar(stat="summary", fun=mean) +
  labs(title="Average Total Fatalities by Speed Limit", x="Speed Limit (MPH)", y="Average Total Fatalities Rate") +
  theme_minimal() +
  scale_fill_brewer(palette="Set2", labels=c("0"="Less than 70 MPH", "1"="Over 70 MPH")) +
  scale_x_discrete(labels=c("0"="Less than 70 MPH", "1"="Over 70 MPH")) +
  scale_fill_discrete(labels=c("0"="Less than 70 MPH", "1"="Over 70 MPH"))
```

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```

```
p1_bar | p2_bar
```



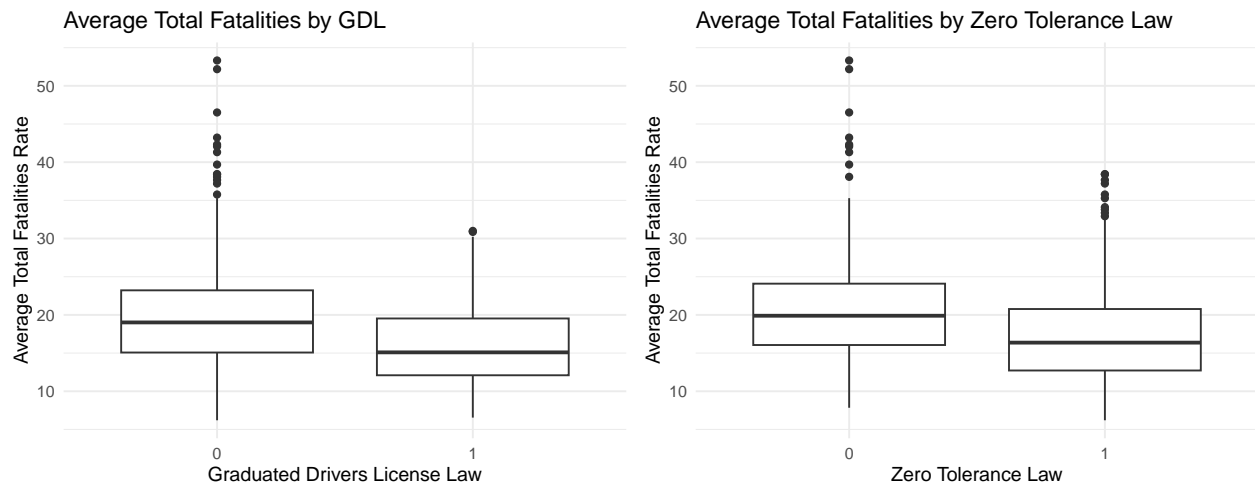
The seat belt law graph shows that regions with no seatbelt law (0) have the highest average fatality rate, while those with a primary enforcement law (1) show a lower rate, and secondary enforcement (2) has the lowest. This indicates that stricter seatbelt laws could be associated with reduced fatalities. The speed limit graph shows that when the speed limit is over 70 the average total fatalities rate is a bit higher, but interestingly not by much.

```
p1_box <- ggplot(data, aes(x=factor(graduated_drivers_license_law), y=total_fatalities_rate)) +
  geom_boxplot() +
  labs(title="Average Total Fatalities by GDL",
       x="Graduated Drivers License Law",
       y="Average Total Fatalities Rate") +
  theme_minimal()

p2_box <- ggplot(data, aes(x=factor(zero_tolerance_law), y=total_fatalities_rate)) +
  geom_boxplot() +
  labs(title="Average Total Fatalities by Zero Tolerance Law",
       x="Zero Tolerance Law",
       y="Average Total Fatalities Rate") +
```

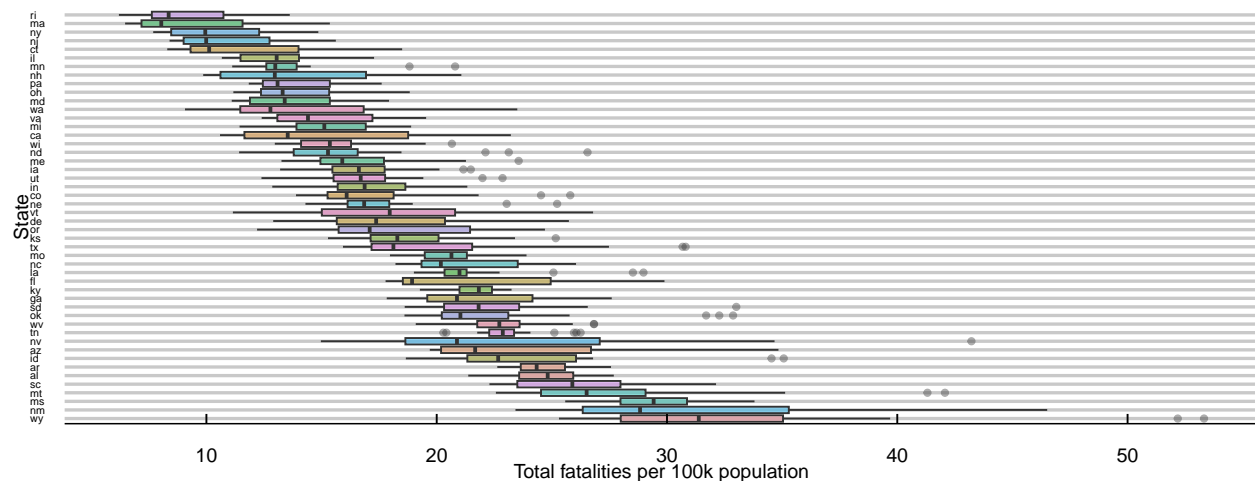
```
theme_minimal()
```

```
p1_box | p2_box
```



The box plots comparing average total fatalities rates against the presence of Graduated Drivers License Law and Zero Tolerance Law show a lower median fatality rate when these laws are in place (indicated by '1') as opposed to when they are not (indicated by '0').

```
data |>
  ggplot(aes(reorder(state, desc(total_fatalities_rate)), total_fatalities_rate, fill=state)) +
  geom_boxplot(alpha=0.4) +
  theme_economist_white(gray_bg=F) +
  theme(legend.position = "none", axis.text.y = element_text(size=6)) +
  scale_y_continuous(label=scales::number_format(accuracy = 1)) +
  xlab("State") +
  ylab("Total fatalities per 100k population") +
  coord_flip()
```



We see strong differences in fatality rate across different states over time, suggesting that fixed effects are important for controlling for unobserved differences.

```
data |>
  ggplot(aes(year_of_observation, total_fatalities_rate, color = state)) +
  geom_point(alpha=0.4) +
  geom_smooth(method="lm") +
```

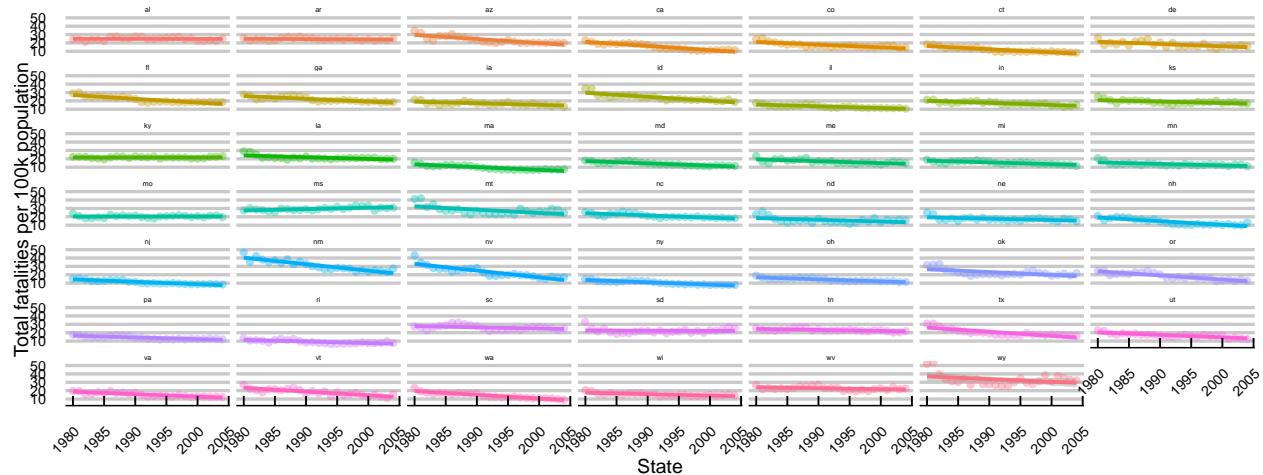


```

facet_wrap(~state) +
theme_economist_white(gray_bg=F)+
theme(legend.position = "none", axis.text.x=element_text(angle=45,hjust=1,vjust=1,size=8),
      axis.text.y = element_text(size=8)) +
theme(strip.text = element_text(size=4)) +
scale_y_continuous(label=scales::number_format(accuracy = 1))+
xlab("State")+
ylab("Total fatalities per 100k population")

```

```
## `geom_smooth()` using formula = 'y ~ x'
```



We see that fatality rate appear to have been trending down in most state over time, with some being more flat.

## (15 points) Preliminary Model

Estimate a linear regression model of *totfatrtte* on a set of dummy variables for the years 1981 through 2004 and interpret what you observe. In this section, you should address the following tasks:

- Why is fitting a linear model a sensible starting place?
- What does this model explain, and what do you find in this model?
- Did driving become safer over this period? Please provide a detailed explanation.
- What, if any, are the limitation of this model. In answering this, please consider **at least**:
  - Are the parameter estimates reliable, unbiased estimates of the truth? Or, are they biased due to the way that the data is structured?
  - Are the uncertainty estimate reliable, unbiased estimates of sampling based variability? Or, are they biased due to the way that the data is structured?

```

prelim_lm <- lm(total_fatalities_rate ~ year_of_observation + bac10 + bac08, data=data)
summary(prelim_lm)

```

```

##
## Call:
## lm(formula = total_fatalities_rate ~ year_of_observation + bac10 +
##     bac08, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8520  -4.3317  -0.6817   3.7333  30.1607

```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    493.61361    58.12676   8.492 < 2e-16 ***
## year_of_observation -0.23748    0.02927  -8.114 1.20e-15 ***
## bac10           -2.02391    0.48917  -4.137 3.76e-05 ***
## bac08           -1.89414    0.67873  -2.791 0.00534 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.012 on 1196 degrees of freedom
## Multiple R-squared:  0.1108, Adjusted R-squared:  0.1086
## F-statistic: 49.69 on 3 and 1196 DF,  p-value: < 2.2e-16

std_residuals <- rstandard(prelim_lm)
```

Fitting a linear model is a good place to start given its simplicity and comprehension to give first insights into the data. It also may immediately signal if ignoring cross sectional units has detrimentally effected the model. Here the model explains that year has a negative relationship with total fatality rate. Additionally, that establishing legality levels of BAC have negative effects on fatality rates as well.

Driving did become safer over time as the trend in years clearly shows, but the BAC legality had a strong influence as well. At the beginning of the time period many states had no BAC limit, but as regulations changed thresholds were established. The model indicates these thresholds helped in the decrease in fatality rates.

The limitations of this model are that the different policy journeys of each state are mixed together in a potentially problematic way that fails to respect the true timing of state decisions to adjust policy and make driving safer.

Additionally states could have started at very distinct levels of driving fatality rates that are obscured when data is pooled together. This is a leading reason that parameter estimates are not reliable and unbiased. The omitted variable of state not being considered leads to this. The influence of heteroscedasticity also gives evidence that the uncertainty estimates are not reliable and biased due to the data structure problems previously mentioned.

## (15 points) Expanded Model

Expand the **Preliminary Model** by adding variables related to the following concepts:

- Blood alcohol levels
- Per se laws
- Primary seat belt laws (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.)
- Secondary seat belt laws
- Speed limits faster than 70
- Graduated drivers licenses
- Percent of the population between 14 and 24 years old
- Unemployment rate
- Vehicle miles driven per capita.

If it is appropriate, include transformations of these variables. Please carefully 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 blood alcohol variables defined? Interpret the coefficients that you estimate for this concept.
- Do *per se laws* have a negative effect on the fatality rate?

- Does having a primary seat belt law?

```
expand_lm <- lm(total_fatalities_rate ~ year_of_observation + graduated_drivers_license_law + bac10 + bac08 + primary_seatbelt_law + secondary_seatbelt_law + perse + unemployment_rate + perc14_24 + sl70plus + vehicmiles, data = data)

summary(expand_lm)

##
## Call:
## lm(formula = total_fatalities_rate ~ year_of_observation + graduated_drivers_license_law + bac10 + bac08 + primary_seatbelt_law + secondary_seatbelt_law + perse + unemployment_rate + perc14_24 + sl70plus + vehicmiles, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2850  -3.6877  -0.3546   3.3226  29.6470
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    540.279681  120.678631   4.477 8.30e-06 ***
## year_of_observation    -0.266993   0.060004  -4.450 9.41e-06 ***
## graduated_drivers_license_law  -1.404149   0.573180  -2.450  0.01444 *
## bac10            -1.289623   0.460641  -2.800  0.00520 **
## bac08            -1.283000   0.639323  -2.007  0.04500 *
## primary_seatbelt_law    1.485568   0.669913   2.218  0.02677 *
## secondary_seatbelt_law    1.256634   0.539304   2.330  0.01997 *
## perse            0.234827   0.387471   0.606  0.54460
## unemployment_rate    0.281398   0.090454   3.111  0.00191 **
## perc14_24         0.611647   0.145887   4.193 2.96e-05 ***
## sl70plus          5.732376   0.530907  10.797 < 2e-16 ***
## vehicmiles        -0.031032   0.003799  -8.169 7.91e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.4 on 1188 degrees of freedom
## Multiple R-squared:  0.2875, Adjusted R-squared:  0.2809
## F-statistic: 43.57 on 11 and 1188 DF,  p-value: < 2.2e-16
```

Primary seat belt laws and Secondary seat belt laws are rounded to represent a Boolean flag and from our chart above should be considered separately due to the different levels of fatality rates. Graduated driver license and Per se laws were similarly rounded to eliminate any data that may have not been a Boolean. Speed limit faster than 70 mph was kept as the original Boolean due to the EDA indicating a slight difference in the fatality rates associated. Unemployment Rate, Vehicle miles driven per capita, Percent of the population between 14 and 24 years old are also unaltered due to them being continuous variables.

For Blood Alcohol levels the BAC 10% and BAC 8% law columns are considered separately to help us evaluate the relative impact in the model. The coefficient values are nearly identical at roughly -1.3. Meaning that the presence of a BAC law lowers fatalities by -1.2 per 100,000 people. Per Se laws have a slightly positive effect on fatalities however the coefficient is not significant and will be removed from the model. Having a primary seatbelt law has a significant and positive impact on fatalities showing the clearest evidence that a pooled linear model is not the appropriate approach to model the trends of the this panel data.

## (15 points) State-Level Fixed Effects

Re-estimate the **Expanded Model** using fixed effects at the state level.

- What do you estimate for coefficients on the blood alcohol variables? How do the coefficients on the blood alcohol variables change, if at all?
- What do you estimate for coefficients on per se laws? How do the coefficients on per se laws change, if at all?
- What do you estimate for coefficients on primary seat-belt laws? How do the coefficients on primary seatbelt laws change, if at all?

Which set of estimates do you think is more reliable? Why do you think this?

- What assumptions are needed in each of these models?
- Are these assumptions reasonable in the current context?

```
year_vars_minus_one <- year_vars[2:length(year_vars)]
expanded_vars <- c("bac10", "bac08", "perse", "primary_seatbelt_law", "secondary_seatbelt_law", "sl70pl
combined_vars <- c(year_vars_minus_one, expanded_vars)

fn <- as.formula(paste('total_fatalities_rate ~', paste(combined_vars, collapse='+'))

fe_mod <- plm(fn, index="state", data=data, model="within")
summary(fe_mod)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = fn, data = data, model = "within", index = "state")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.      1st Qu.      Median      3rd Qu.      Max.
## -8.2942745 -1.0561099  0.0055578  0.9788361 14.8497791
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## d81             -1.5124e+00  4.1379e-01 -3.6549 0.0002692 ***
## d82             -3.0540e+00  4.4318e-01 -6.8912 9.221e-12 ***
## d83             -3.6638e+00  4.5516e-01 -8.0495 2.111e-15 ***
## d84             -4.3998e+00  4.5966e-01 -9.5719 < 2.2e-16 ***
## d85             -4.8603e+00  4.8010e-01 -10.1237 < 2.2e-16 ***
## d86             -3.7692e+00  5.1357e-01 -7.3392 4.123e-13 ***
## d87             -4.4123e+00  5.5162e-01 -7.9989 3.118e-15 ***
## d88             -4.8877e+00  5.9837e-01 -8.1684 8.379e-16 ***
## d89             -6.2395e+00  6.3732e-01 -9.7901 < 2.2e-16 ***
## d90             -6.3564e+00  6.6196e-01 -9.6024 < 2.2e-16 ***
## d91             -7.0442e+00  6.7895e-01 -10.3752 < 2.2e-16 ***
## d92             -7.8905e+00  7.0039e-01 -11.2659 < 2.2e-16 ***
## d93             -8.2366e+00  7.1290e-01 -11.5536 < 2.2e-16 ***
## d94             -8.6823e+00  7.3004e-01 -11.8930 < 2.2e-16 ***
## d95             -8.3889e+00  7.5324e-01 -11.1370 < 2.2e-16 ***
## d96             -8.7648e+00  7.9400e-01 -11.0388 < 2.2e-16 ***
## d97             -8.9164e+00  8.1140e-01 -10.9889 < 2.2e-16 ***
## d98             -9.5333e+00  8.2867e-01 -11.5044 < 2.2e-16 ***
## d99             -9.6940e+00  8.3614e-01 -11.5938 < 2.2e-16 ***
## d00             -1.0223e+01  8.4713e-01 -12.0683 < 2.2e-16 ***
## d01             -9.9608e+00  8.5745e-01 -11.6168 < 2.2e-16 ***
```

```
## d02                -9.2546e+00  8.6613e-01 -10.6850 < 2.2e-16 ***
## d03                -9.3270e+00  8.6980e-01 -10.7232 < 2.2e-16 ***
## d04                -9.6676e+00  8.9310e-01 -10.8248 < 2.2e-16 ***
## bac10              -8.6977e-01  2.2522e-01 -3.8619 0.0001190 ***
## bac08              -1.1805e+00  3.2987e-01 -3.5786 0.0003603 ***
## perse              -1.0587e+00  2.2415e-01 -4.7230 2.619e-06 ***
## primary_seatbelt_law -1.2506e+00  3.4313e-01 -3.6447 0.0002800 ***
## secondary_seatbelt_law -3.5659e-01  2.5230e-01 -1.4133 0.1578360
## sl70plus           -3.2440e-02  2.6034e-01 -0.1246 0.9008577
## graduated_drivers_license_law -3.0503e-01  2.8029e-01 -1.0883 0.2767100
## perc14_24           1.9367e-01  9.5068e-02  2.0372 0.0418646 *
## unemployment_rate   -5.7652e-01  6.0592e-02 -9.5147 < 2.2e-16 ***
## vehicmilespc        9.2612e-04  1.1066e-04  8.3691 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4547.9
## R-Squared:              0.6252
## Adj. R-Squared: 0.59804
## F-statistic: 54.8501 on 34 and 1118 DF, p-value: < 2.22e-16
```

*Need to add commentary*

## (10 points) Consider a Random Effects Model

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

- Please state the assumptions of a random effects model, and evaluate whether these assumptions are met in the data.
- If the assumptions are, in fact, met in the data, then estimate a random effects model and interpret the coefficients of this model. Comment on how, if at all, the estimates from this model have changed compared to the fixed effects model.
- If the assumptions are **not** met, then do not estimate the data. But, also comment on what the consequences would be if you were to *inappropriately* estimate a random effects model. Would your coefficient estimates be biased or not? Would your standard error estimates be biased or not? Or, would there be some other problem that might arise?

## (10 points) Model Forecasts

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.

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

### **(5 points) 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?