Lab 3: Panel Models

US Traffic Fatalities: 1980 - 2004

Contents

1 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.

2 (30 points, total) Build and Describe the Data

1. (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.)

```r
my_panel <- 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, "nolim", "NAN")))))) %>%
  select(-sl55, -sl65, -sl70, -sl75, -slnone) %>% # Drop unused columns
  mutate(speed_limit = as.factor(speed_limit))    # Convert to factor
```

```r
# Create an empty vector to store the names of non-zero columns for each row
result_vector <- character(nrow(my_panel))
year_names <- names(my_panel[,26:50])
# Loop through each row and find the non-zero column name (if any)
for (i in 1:nrow(my_panel)) {
  year_level <-  sub(".", "", year_names[my_panel[i,26:50] != 0])
  year_level <- paste(ifelse(as.numeric(year_level) < 80, "20", "19"), year_level, sep = "")
  result_vector[i] <- year_level
}
# Add the new column to the dataframe and relevel it so that 1980 was level 0
my_panel$year_of_observation <- relevel(as.factor(result_vector), ref = "1980")
# Drop unneeded columns
my_panel <- cbind(my_panel[,1:25], my_panel[,51:53])
```

```r
#Make BAC and seatbelt factor variables
my_panel <- my_panel %>%
  mutate(blood_alc_lmt = ifelse(bac10 >=0.5, "1.0",
                         ifelse(bac08 >=0.5, "0.8", "None"))) %>%
  select(-bac10, -bac08, -sbprim, -sbsecon) %>%
  mutate(blood_alc_lmt = as.factor(blood_alc_lmt),
         seatbelt = as.factor(seatbelt),
         zerotol = as.logical(ifelse(zerotol >=0.5, T, F)),
         gdl = as.logical(ifelse(gdl >=0.5, T, F)),
         perse = as.logical(ifelse(perse >=0.5, T, F)))
```

```r
# Add state names and abbreviations to the dataframe
# Create vectors for US states and their abbreviations
us_states <- c(
  "Alabama", "Alaska", "Arizona", "Arkansas", "California",
  "Colorado", "Connecticut", "Delaware", "District Columbia", "Florida", "Georgia",
  "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa",
  "Kansas", "Kentucky", "Louisiana", "Maine", "Maryland",
  "Massachusetts", "Michigan", "Minnesota", "Mississippi", "Missouri",
  "Montana", "Nebraska", "Nevada", "New Hampshire", "New Jersey",
  "New Mexico", "New York", "North Carolina", "North Dakota", "Ohio",
  "Oklahoma", "Oregon", "Pennsylvania", "Rhode Island", "South Carolina",
  "South Dakota", "Tennessee", "Texas", "Utah", "Vermont",
  "Virginia", "Washington", "West Virginia", "Wisconsin", "Wyoming")

state_abbreviations <- c(
  "AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DE", "DC", "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")

# Create a dataframe with US states and their abbreviations
us_states_df <- data.frame(State_name = us_states, Abbreviation = state_abbreviations, state = seq(1:51
# Merge the names to the main panel
my_panel <- merge(my_panel, us_states_df, by.x = "state", by.y = "state", all.x = TRUE)
```

```
names(my_panel) <- c("state", "year", "seat_belt", "min_age", "zero_tolerance",
   "gradual_dl", "no_proof_guilt",
   "total_fatal", "night_fatal", "weekend_fatal",
   "total_fatal_per_miles", "night_fatal_per_miles", "weekend_fatal_per_miles",
   "state_pop",
   "total_fatal_per_pop", "night_fatal_per_pop", "weekend_fatal_per_pop",
   "tot_miles_driven", "unemployment", "percent_young", "speed_limit_above_70",
   "miles_per_capita", "dum_speed_limit", "dum_year", "dum_blood_alc",
   "state_name", "state_abbr")
```

2.1 Description of Dataset

2. (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?

The data originates from the study by Freeman, D.G. (2007), "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws" published in Contemporary Economic Policy 25, 293–308. Author of the study made the data publicly accessible. It appears to describe an entire population of the US, although the details of data collection are behind a pay wall.

This data set is organized as a long table, with a set of observations for each US state (except Alaska, Hawaii, District of Columbia that are missing) for each year from 1980 to 2004. The years are coded as a numeric values as well as categorical variables. For each state-year category there are observations related to the state's traffic laws, traffic fatalities, and population/driving statistics. The data is balanced. ' To make the data set suitable for our analysis, we applied a set of transformations (see above) and the rest of the description will discuss the transformed data set.

Some of the variable that can only be binary (zero tolerance that can either be implemented or not) sometimes have fractional value. Presumably, that indicates the years when the variable changed its value. To simplify interpretations, we rounded these values to the nearest integer.

2.1.1 Traffic Laws and Drinking Regulations

Speed limit is represented by two categorical variables:

- Max speed limit in the state: dum_speed_limit with categories $sl55$, $sl65$, $sl70$, $sl75$, $slnone$
- Whether max speed limit in the state exceeds 70 mph: $speed\_limit\_above_70$

Seat belt rules are represented by a categorical variable $seat\_belt$ with categories "0" if there is no law, '1' if not wearing a seat belt is a sufficient reason for a ticket, "2" if seat belt ticket can only be issued in conjunction with another violation.

Drivers licensing process is represented by $gradual\_dl$ indicating if there is a step-wise process to obtaining a full licence.

Drinking regulations are represented by the following variables:

- *min_age* is a number representing the legal drinking age
- *zero_tolerance* is a logical variable showing if zero alcohol tolerance is in force
- *blood_alc_lmt* is a categorical with three levels: 'bac08' for 0.8‰ 'bac10' for 1.0‰ and 'none' for no upper limit.
- no_proof_guilt is a logical variable that indicates if drivers licence can be suspended without positive medical proof of driving under influence.

2.1.2 Fatalities

Fatality is given as $total\_fatal$, total fatalities at this year and state and additional $night\_fatal$ and $weekend\_fatal$ variables giving temporal details. Data normalized to 100,000 people and to 100 million miles driven is also presented in two groups of three variable.

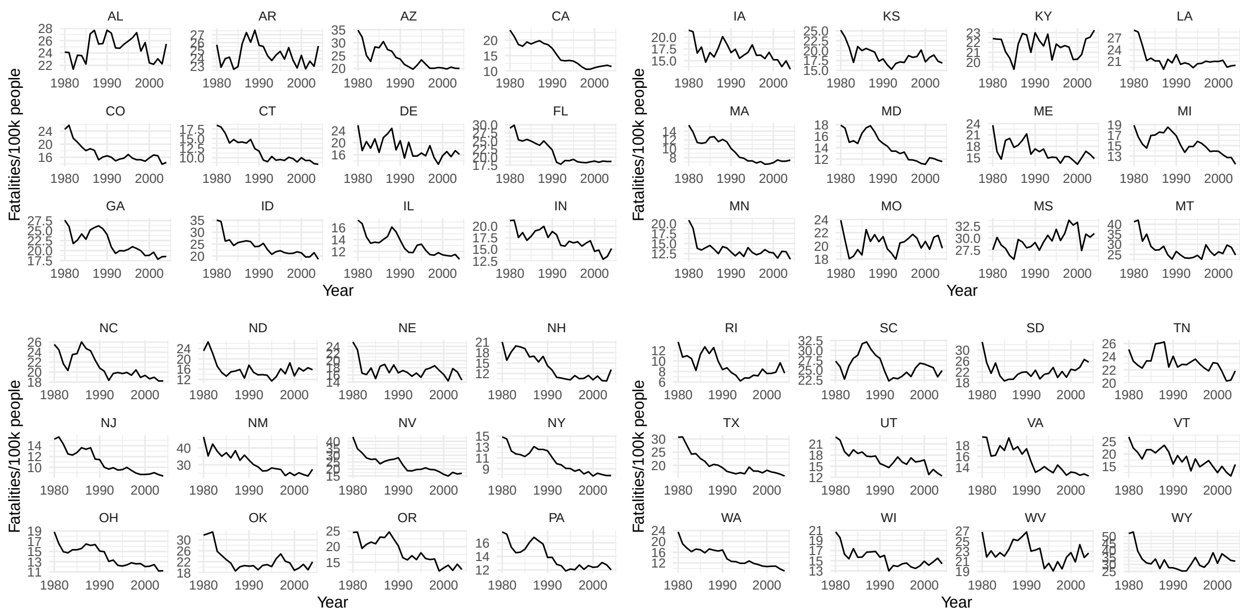
2.1.2.1 State demographics Additional information for each year and each state is presented in numeric variables:

- *state_pop* is the state population
- *unemployment* is the percent of unemployment
- *percent_young* is the percent of population 18 to 24 years old
- *miles_per_capita* is the number of miles driven per person

2.2 EDA

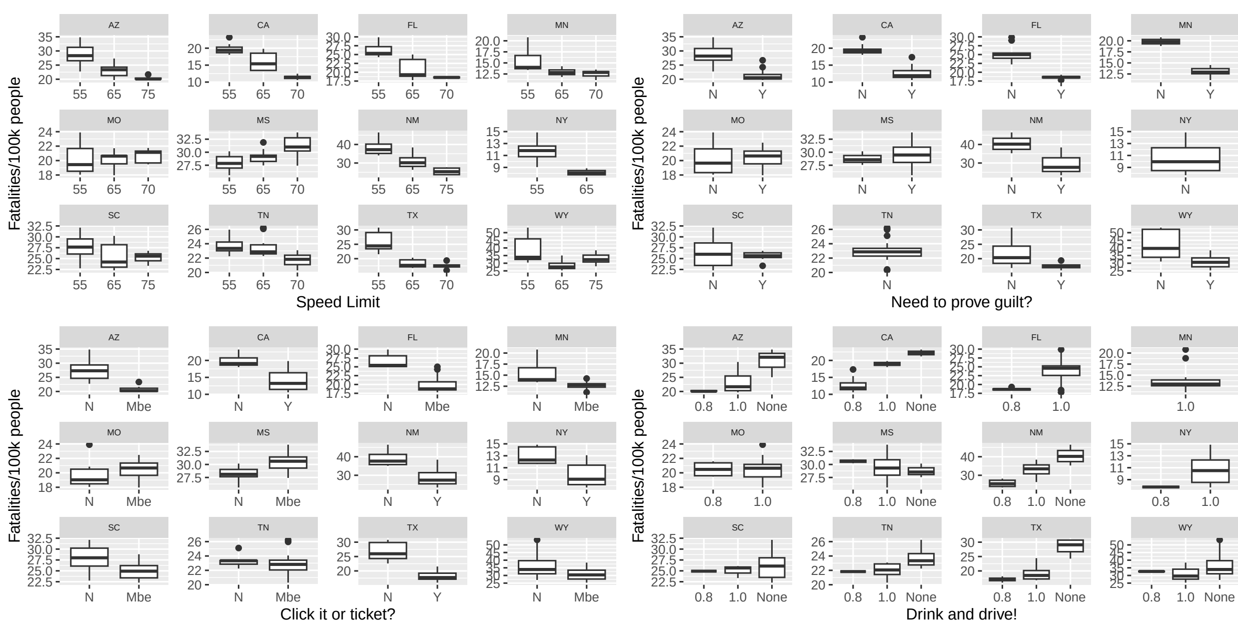
3. (20 points) Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable $total\_fatalities\_rate$ and the potential explanatory variables. Minimally, this should include:
   - How is the our dependent variable of interest $total\_fatalities\_rate$ 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.



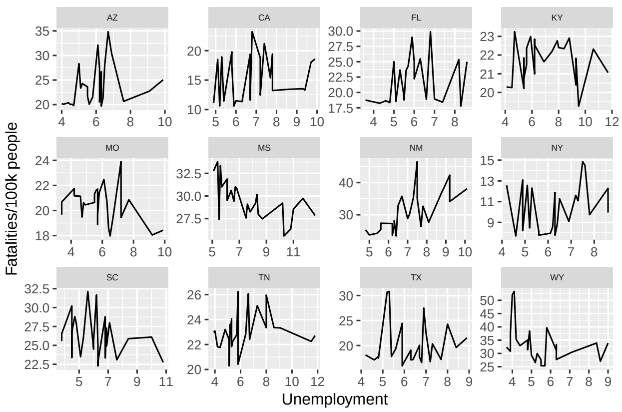
Plots above show how fatality per person changes over time for each state represented in the data set. There is a general downward trend that is particularly visible in case of the more populous states such as Texas, California and Florida. Only South Dakota is trending up consistently, but that might be a statistical artifact caused by small population. Some states, like Texas, Utah and Colorado, exhibit a very consistent downward trend, that would be hard to explain by the influence of any one-time decision. A large group of states (Oregon, Pennsylvania and Maryland, to name a few) exhibit strong increase in the fatality rate in the late 80's early 90's before sudden drop to almost modern levels. This patters would be consistent with the influence of some policy change that became popular in the late 80's.

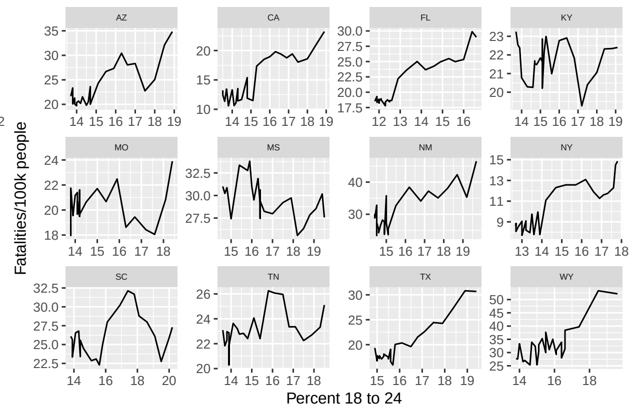
Visualizing all 50 states would be visually overwhelming and for the next set of graphs we decided to use the most populous states and the states with the most interesting fatality-time patterns as representatives of the entire country: TX, CA, NY, MN, FL, AZ, SC, NM, WY, MO, TN, MS

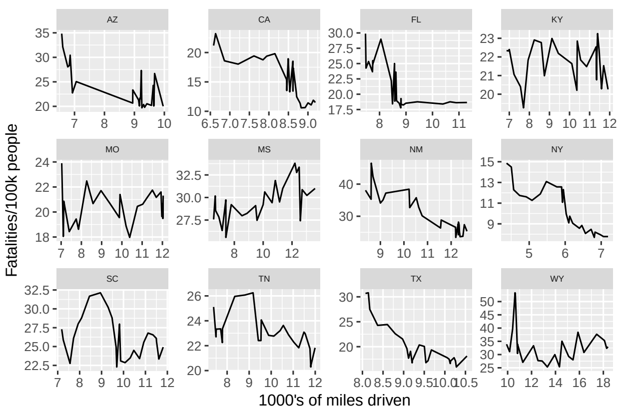


The group of plots above show correlation of four policies and the road fatality rates: changes in speed limit, changes in seat belt laws, institution of no-proof guilt when it comes to DUI and blood alcohol limits.

Unexpectedly, the speed limit seems to be inversely correlated with the fatality: the higher the limit the lower the fatality. This is likely the case of spurious correlation. More stringent seat belt laws are also related to the lower fatality rate more often than not. Introduction of *per se* laws is also associated with the reduced fatality. Unsurprisingly, there is strong negative association between allowed blood alcohol and fatalities, although the difference between BAC 0.8‰ and 1.0‰ is often negligible. Having a mixture of expected and unexpected results, these plots are good reminders that correlation does not equal causation.





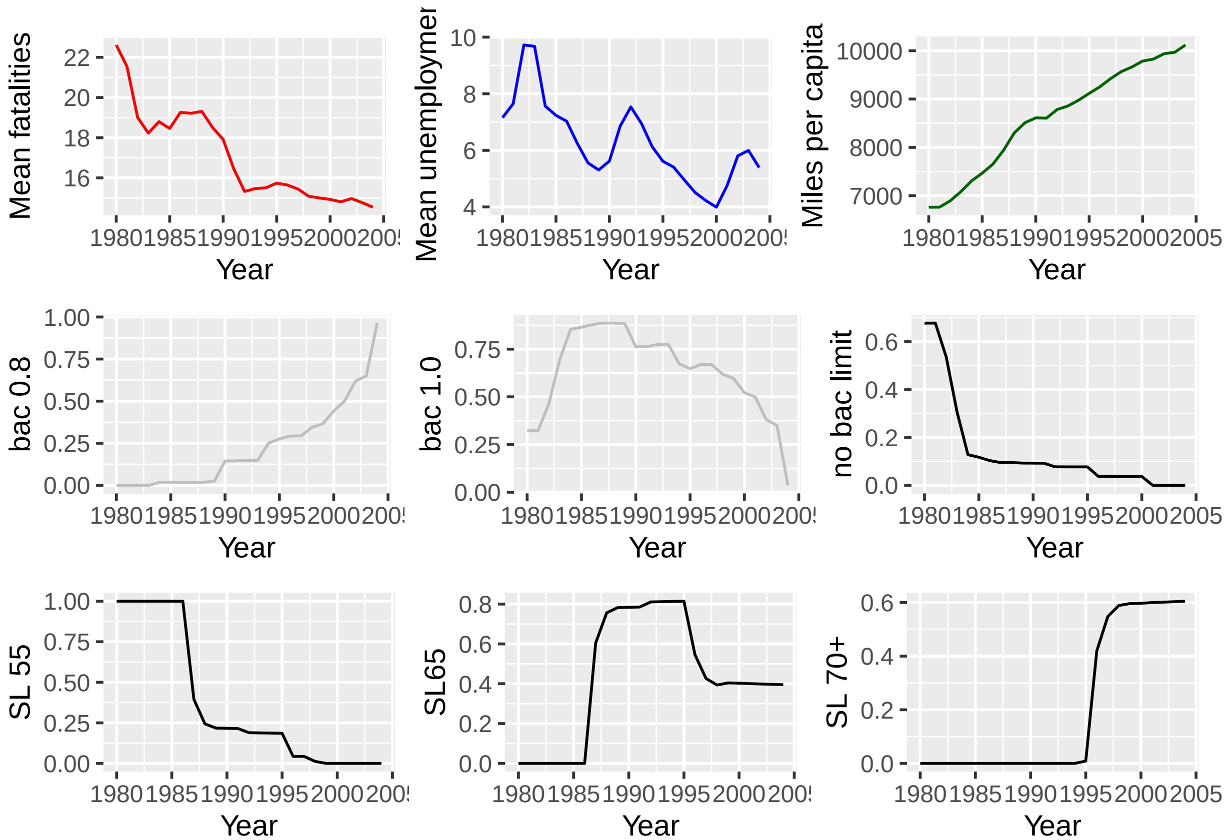


The graphs above demonstrate correlation of factors that are outside of the legislator's control. Thee factors change gradually and would be better suited to explain gradual change in the fatality that we observed above. Unemployment does not seem to have a clear correlation with the fatality, while percentage of the young population has clear positive correrlation. Given the car insurance rates for young driver, we suspect that this effect is well known in the insurance world. Unexpectedly, the number of miles driven per person has negative correlation with the fatality: as the millage goes up the fatality goes down in most populous states. This effect is hard to rationalize and we expect to shed some light on it with more sophisticated analysis.

Normalizing the legislations to population makes our observations even more clear. The graphs below present time-course of the following variables: nation-wide fatalities normalized to population, unemployment and miles per capita. Importantly, these values are not averages by state, but the average weighted byt the state population. The middle set of graphs presents the fraction of the US population living with the BAC 0.8, BAC 1.0 or no limit whatsoever. Similarily, the bottom row shows the fraction of the US population living with the speed limits 55, 65 or above 70 mph. These plots show, that mandating some BAC limit is anti-correlated with the fatalities very strongly, but the exact limit matters little. Miles driven per capita does not seem to have anything in common with the fatalities, nether does the speed limit.

```
av_explore <- my_panel %>%
  group_by(year) %>%
  summarise( pop = sum(state_pop),
                          mean_fat = 100000 * sum(total_fatal)/pop,
```

```r
                       unempl = sum(unemployment * state_pop)/pop,
                       miles = sum(tot_miles_driven*1000000000)/pop,
                       pop_08 = sum(state_pop * ifelse(dum_blood_alc == "0.8", 1, 0))/pop,
                       pop_1 = sum(state_pop * ifelse(dum_blood_alc == "1.0", 1, 0))/pop,
                       pop_non = sum(state_pop * ifelse(dum_blood_alc == "None", 1, 0))/pop,
                       pop_55 = sum(state_pop * ifelse(dum_speed_limit == "55", 1, 0))/pop,
                       pop_65 = sum(state_pop * ifelse(dum_speed_limit == "65", 1, 0))/pop,
                       pop_70 = sum(state_pop * speed_limit_above_70)/pop)
```



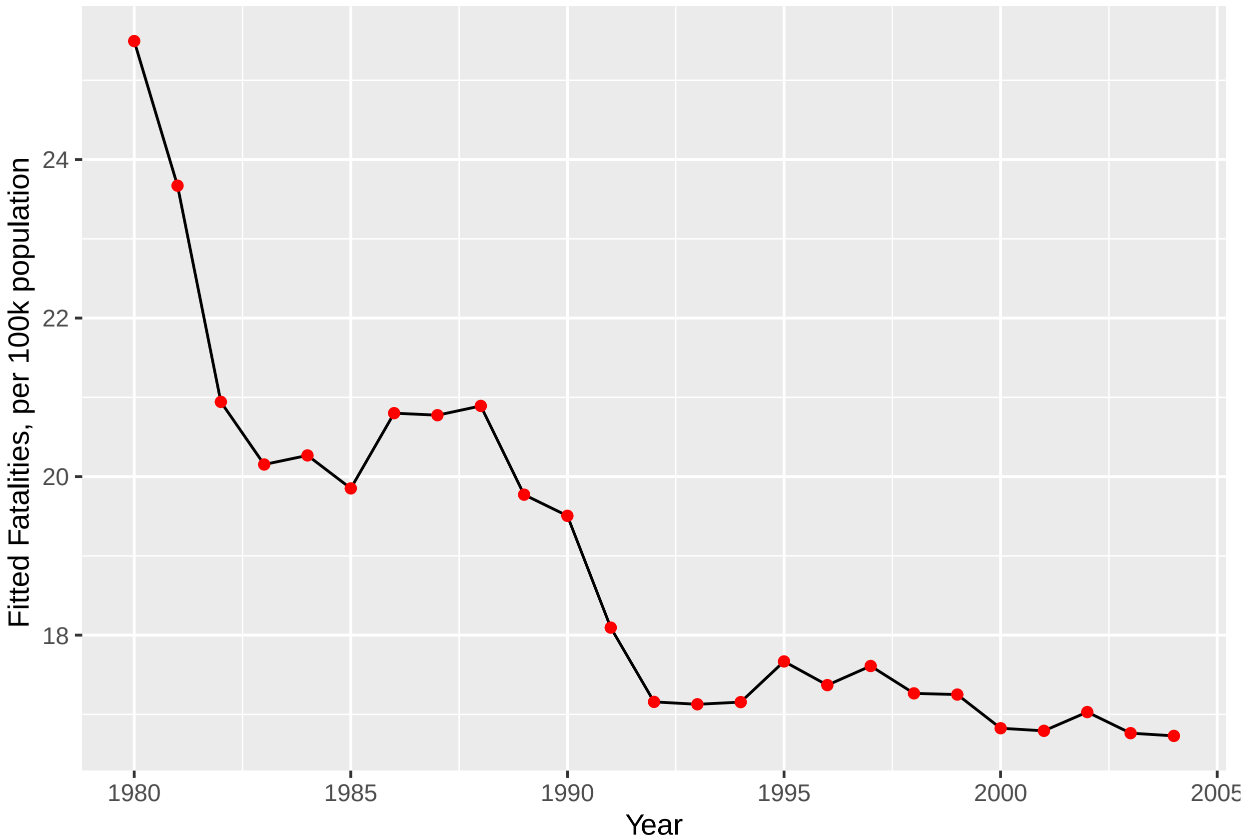
3 (15 points) Preliminary Model

Estimate a linear regression model of *totfatrte* 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?

```r
lsdv_model <- plm(total_fatal_per_pop ~ dum_year, data = my_panel,
                  effect = "individual", model = "pooling")
# summary(lsdv_model)
```

```r
av_fat <- my_panel %>% group_by(year) %>% summarise(mean_fat = mean(total_fatal_per_pop))

ggplot(data = broom::augment(lsdv_model), aes(x = 1979+as.numeric(dum_year), y = .fitted)) +
  geom_line() +
  geom_point(aes(x = year, y = mean_fat), color = "red", data = av_fat) +
  labs(x = "Year", y = "Fitted Fatalities, per 100k population ") +
  theme(legend.position = "none")
```



Fitting a linear model might be a good first step, because most of the assumptions of a linear regression are satisfied:

1. Linearity: each dummy variable has only two levels, therefore model is guaranteed linear
2. I.I.D. : Fatality rate is reasonably independent across states
3. No perfect colinearity in the data
4. No obvious correlation between the error term and individual-specific effects

This model shows how much safer driving becomes for each year, as compared to 1980. All coefficients are statistically significant and negative, showing that driving in 1980 was the most dangerous of all years in observation. Generally, fitted fatality rate decreases, with a slight increase in the late 80s. Driving, on average became safer in this period. Because the data set is a complete census of the population, and the model is descriptive, there is no uncertainty associated with these estimates. These conclusions are purely descriptive, the errors can be calculated down to a single fatality. So, while linear models of this kind do suffer from bias in the estimation of the coefficients, because of the potential correlation between state-level unobserved effects and the error terms, this would only matter if the data was a sample of the population.

4 (15 points) Expanded Model

Expand the *PreliminaryModel* 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 *perselaws* have a negative effect on the fatality rate?
- Does having a primary seat belt law?

```
lsdv_model_expand <- plm(total_fatal_per_pop ~ dum_year +
                           dum_blood_alc +
                           no_proof_guilt +
                           seat_belt +
                           speed_limit_above_70 +
                           gradual_dl +
                           percent_young +
                           unemployment +
                           miles_per_capita,
                      data = my_panel,
                      effect = "individual", model = "pooling")
summary(lsdv_model_expand)
```

```
## Pooling Model
##
## Call:
## plm(formula = total_fatal_per_pop ~ dum_year + dum_blood_alc +
##     no_proof_guilt + seat_belt + speed_limit_above_70 + gradual_dl +
##     percent_young + unemployment + miles_per_capita, data = my_panel,
##     effect = "individual", model = "pooling")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -14.90250  -2.71443  -0.28619   2.27799  21.45800
##
## Coefficients:
##                         Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)          -5.0171e+00  2.5080e+00  -2.0004  0.045689 *
## dum_year1981         -2.1793e+00  8.2850e-01  -2.6305  0.008639 **
## dum_year1982         -6.5755e+00  8.5445e-01  -7.6956 2.990e-14 ***
## dum_year1983         -7.3602e+00  8.7246e-01  -8.4361 < 2.2e-16 ***
```

```
## dum_year1984        -5.8609e+00  8.7884e-01  -6.6689 3.972e-11 ***
## dum_year1985        -6.5303e+00  8.9559e-01  -7.2916 5.641e-13 ***
## dum_year1986        -5.9150e+00  9.3089e-01  -6.3541 3.001e-10 ***
## dum_year1987        -6.4319e+00  9.6719e-01  -6.6501 4.494e-11 ***
## dum_year1988        -6.6625e+00  1.0141e+00  -6.5699 7.575e-11 ***
## dum_year1989        -8.1436e+00  1.0528e+00  -7.7349 2.230e-14 ***
## dum_year1990        -9.0318e+00  1.0773e+00  -8.3840 < 2.2e-16 ***
## dum_year1991        -1.1157e+01  1.1013e+00 -10.1304 < 2.2e-16 ***
## dum_year1992        -1.2952e+01  1.1230e+00 -11.5339 < 2.2e-16 ***
## dum_year1993        -1.2834e+01  1.1361e+00 -11.2965 < 2.2e-16 ***
## dum_year1994        -1.2449e+01  1.1577e+00 -10.7529 < 2.2e-16 ***
## dum_year1995        -1.2057e+01  1.1833e+00 -10.1894 < 2.2e-16 ***
## dum_year1996        -1.3953e+01  1.2241e+00 -11.3990 < 2.2e-16 ***
## dum_year1997        -1.4378e+01  1.2492e+00 -11.5097 < 2.2e-16 ***
## dum_year1998        -1.5147e+01  1.2646e+00 -11.9782 < 2.2e-16 ***
## dum_year1999        -1.5193e+01  1.2847e+00 -11.8257 < 2.2e-16 ***
## dum_year2000        -1.5553e+01  1.3039e+00 -11.9283 < 2.2e-16 ***
## dum_year2001        -1.6351e+01  1.3289e+00 -12.3047 < 2.2e-16 ***
## dum_year2002        -1.6895e+01  1.3384e+00 -12.6233 < 2.2e-16 ***
## dum_year2003        -1.7257e+01  1.3417e+00 -12.8620 < 2.2e-16 ***
## dum_year2004        -1.6882e+01  1.3725e+00 -12.3001 < 2.2e-16 ***
## dum_blood_alc1.0     1.0221e+00  3.6906e-01   2.7696  0.005702 **
## dum_blood_alcNone    2.3624e+00  5.2702e-01   4.4826 8.102e-06 ***
## no_proof_guiltTRUE  -5.7895e-01  2.9303e-01  -1.9757  0.048425 *
## seat_belt1          -9.2094e-02  4.9132e-01  -0.1874  0.851348
## seat_belt2           6.8642e-02  4.2953e-01   0.1598  0.873061
## speed_limit_above_70 3.3548e+00  4.4554e-01   7.5298 1.015e-13 ***
## gradual_dlTRUE      -4.1051e-01  5.0600e-01  -0.8113  0.417373
## percent_young        1.3860e-01  1.2278e-01   1.1288  0.259201
## unemployment         7.5628e-01  7.7987e-02   9.6976 < 2.2e-16 ***
## miles_per_capita     2.9220e-03  9.5050e-05  30.7412 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    48612
## Residual Sum of Squares: 19107
## R-Squared:      0.60694
## Adj. R-Squared: 0.59547
## F-statistic: 52.9099 on 34 and 1165 DF, p-value: < 2.22e-16
```

5 (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?

```r
lsdv_model_fixed_effects <- plm(total_fatal_per_pop ~ dum_year +
                           dum_blood_alc +
                           no_proof_guilt +
                           seat_belt +
                           speed_limit_above_70 +
                           gradual_dl +
                           percent_young +
                           unemployment +
                           miles_per_capita,
                      data = my_panel,
                      effect = "individual", model = "within")
summary(lsdv_model_fixed_effects)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = total_fatal_per_pop ~ dum_year + dum_blood_alc +
##     no_proof_guilt + seat_belt + speed_limit_above_70 + gradual_dl +
##     percent_young + unemployment + miles_per_capita, data = my_panel,
##     effect = "individual", model = "within")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -8.4044751 -1.0440737 -0.0032183  0.9573453 14.8233136
##
## Coefficients:
##                         Estimate  Std. Error  t-value  Pr(>|t|)
## dum_year1981         -1.5127e+00  4.1381e-01  -3.6556 0.0002685 ***
## dum_year1982         -2.9849e+00  4.4287e-01  -6.7398 2.535e-11 ***
## dum_year1983         -3.4598e+00  4.5889e-01  -7.5395 9.727e-14 ***
## dum_year1984         -4.2826e+00  4.6626e-01  -9.1849 < 2.2e-16 ***
## dum_year1985         -4.7846e+00  4.8523e-01  -9.8605 < 2.2e-16 ***
## dum_year1986         -3.7351e+00  5.1705e-01  -7.2238 9.332e-13 ***
## dum_year1987         -4.3825e+00  5.5469e-01  -7.9009 6.592e-15 ***
## dum_year1988         -4.8418e+00  6.0111e-01  -8.0548 2.027e-15 ***
## dum_year1989         -6.2159e+00  6.3957e-01  -9.7188 < 2.2e-16 ***
## dum_year1990         -6.3107e+00  6.6441e-01  -9.4982 < 2.2e-16 ***
## dum_year1991         -7.0007e+00  6.8167e-01 -10.2700 < 2.2e-16 ***
## dum_year1992         -7.8407e+00  7.0290e-01 -11.1548 < 2.2e-16 ***
## dum_year1993         -8.1930e+00  7.1542e-01 -11.4521 < 2.2e-16 ***
## dum_year1994         -8.5658e+00  7.3422e-01 -11.6665 < 2.2e-16 ***
## dum_year1995         -8.3566e+00  7.5545e-01 -11.0618 < 2.2e-16 ***
## dum_year1996         -8.6835e+00  7.9610e-01 -10.9076 < 2.2e-16 ***
## dum_year1997         -8.8090e+00  8.1898e-01 -10.7560 < 2.2e-16 ***
## dum_year1998         -9.4488e+00  8.3300e-01 -11.3431 < 2.2e-16 ***
## dum_year1999         -9.5643e+00  8.4371e-01 -11.3360 < 2.2e-16 ***
## dum_year2000         -1.0087e+01  8.5503e-01 -11.7968 < 2.2e-16 ***
## dum_year2001         -9.7406e+00  8.6952e-01 -11.2023 < 2.2e-16 ***
## dum_year2002         -9.0080e+00  8.7708e-01 -10.2705 < 2.2e-16 ***
## dum_year2003         -9.0591e+00  8.8023e-01 -10.2917 < 2.2e-16 ***
## dum_year2004         -9.4269e+00  9.0281e-01 -10.4417 < 2.2e-16 ***
```

```
## dum_blood_alc1.0      4.0447e-01  2.4481e-01   1.6522 0.0987846 .
## dum_blood_alcNone     1.3602e+00  3.7610e-01   3.6165 0.0003120 ***
## no_proof_guiltTRUE   -1.0577e+00  2.2605e-01  -4.6791 3.233e-06 ***
## seat_belt1           -1.2396e+00  3.4323e-01  -3.6115 0.0003179 ***
## seat_belt2           -3.6313e-01  2.5234e-01  -1.4390 0.1504177
## speed_limit_above_70 -7.6225e-02  2.6958e-01  -0.2828 0.7774172
## gradual_dlTRUE       -4.0829e-01  2.7999e-01  -1.4582 0.1450606
## percent_young         1.9005e-01  9.5181e-02   1.9968 0.0460930 *
## unemployment         -5.7887e-01  6.0567e-02  -9.5575 < 2.2e-16 ***
## miles_per_capita      9.3498e-04  1.1127e-04   8.4031 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4548.1
## R-Squared:      0.62518
## Adj. R-Squared: 0.59803
## F-statistic: 54.8466 on 34 and 1118 DF, p-value: < 2.22e-16
```

When using a within estimator as opposed to a pooled estimator the estimates are significantly different:

- BAC Limit of 1.0: The estimate changes from 1.02 (p = 0.006) to 4.04 (p = 0.099) deaths per 100k
- No BAC Limit: The estimate changes from 2.36 (p = 0) to 1.36 (p = 0) deaths per 100k
- Per Se Law: The estimate changes from -5.79 (p = 0.048) to -1.06 (p = 0) deaths per 100k
- Primary Seat Belt Law: The estimate changes from -9.21 (p = 0.85) to -1.24 (p = 0) deaths per 100k

```r
pFtest(lsdv_model_fixed_effects, lsdv_model_expand)
```

```
##
##  F test for individual effects
##
## data:  total_fatal_per_pop ~ dum_year + dum_blood_alc + no_proof_guilt +  ...
## F = 76.148, df1 = 47, df2 = 1118, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Both the Pooled and Fixed Effects Models assume the following:

1. Linearity in the parameters
2. Samples are independent and identically distributed
3. No multi-collinearity

However, the pooled OLS model also assumes that individual unobserved heterogeneity is uncorrelated with the independent variable. This is not a necessary assumption in the fixed effects model which makes it easier to satisfy the assumptions. In this case, the assumption of uncorrelated, unobserved heterogeneity is not valid. This is a difficult assumption to satisfy and is further supported by the `pFtest` which rejected the null hypothesis that there are no significant fixed effects in the data. Because of this, the results from the fixed effects model are more trustworthy as the pooled model does not sufficiently satisfy the assumptions required.

```
# (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
- If the assumptions are, in fact, met in the data, then estimate a random effects model and interpret t
- If the assumptions are **not** met, then do not estimate the data. But, also comment on what the cons
```

```
- Independence across Units and Time: The error term should be uncorrelated with the explanatory variabl
- No Endogeneity: The individual-specific effect (unobserved heterogeneity) should be uncorrelated with
- Homoskedasticity: The error term should have constant variance.
- No Serial Correlation: There should be no correlation between the error term of different time periods

The random effects model assumes that the individual-specific effects are uncorrelated with the explana

The Hausman test compares the fixed effects and random effects models to determine if the individual ef

Based on the Hausman test, this assumption is not met in our data. The p-value is virtually zero, leadi
```

```r
library(plm)

# Define the formula
formula <- total_fatal_per_pop ~ dum_year + dum_blood_alc + no_proof_guilt + seat_belt +
                                speed_limit_above_70 + gradual_dl + percent_young +
                                unemployment + miles_per_capita

# Fit the random effects model
random_model <- plm(formula, data = my_panel, model = "random")

# Fit the fixed effects model
fixed_model <- plm(formula, data = my_panel, model = "within")

# Conduct the Hausman test
hausman_test <- phtest(random_model, fixed_model)
print(hausman_test)
```

```
##
##  Hausman Test
##
## data:  formula
## chisq = 150.67, df = 34, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

If you were to inappropriately estimate a random effects model when the assumptions are not met, the coefficient estimates would be biased and inconsistent. They would not accurately capture the relationships between the variables in the model. The standard error estimates would also be biased, leading to incorrect inference. This could result in wrong conclusions about statistical significance. The random effects model also assumes that the variance components are constant across entities and time, and a failure of these assumptions could also lead to misleading results.

6 (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:

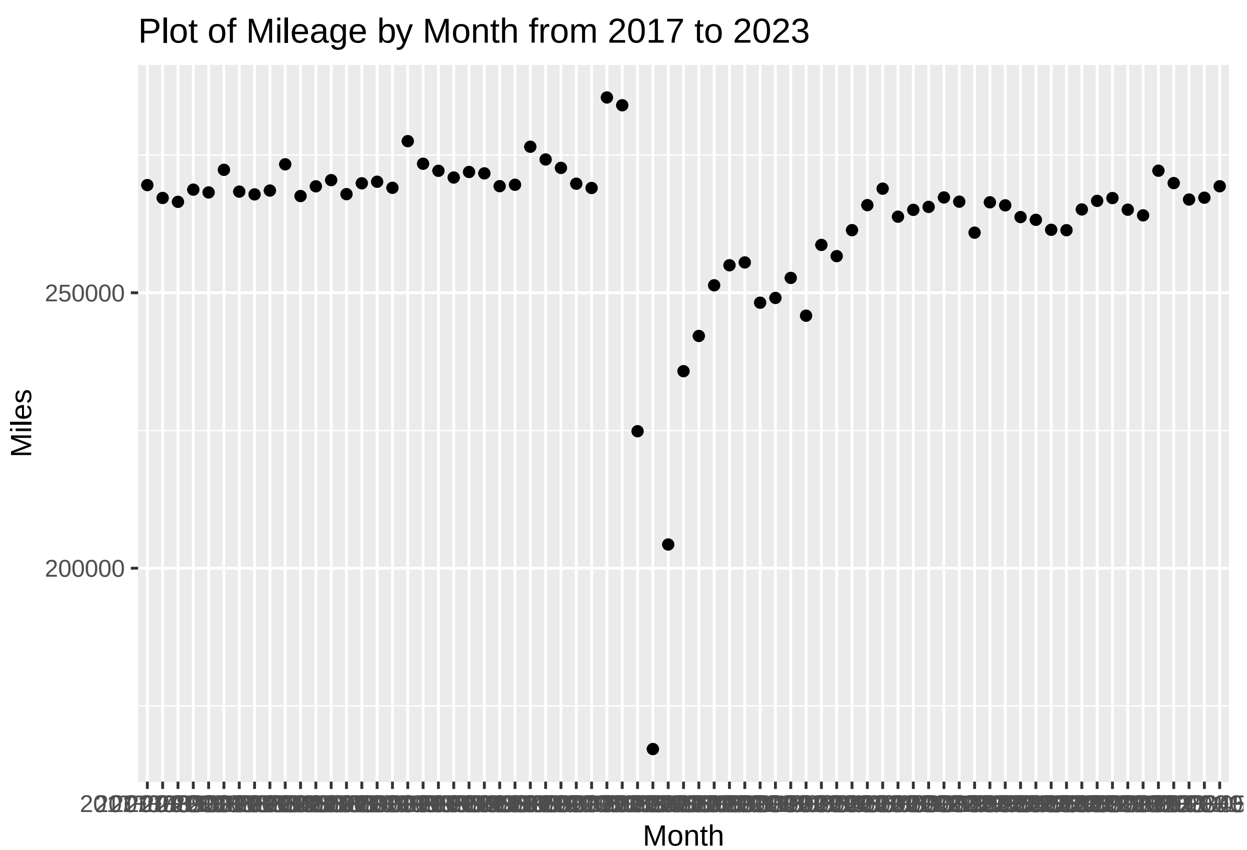
The data I am using is the Vehicle Miles Traveled dataset from Fred. The unit is in millions of miles. The dataset includes the monthly vehicle miles traveled from 2012 to 2023. For this experiment, I am only using the data from 2018 to 2023. Source: https://fred.stlouisfed.org/series/TRFVOLUSM227SFWA

```r
df = read.csv("travel_data.csv")
df = df[211:281,]
colnames(df)[1] = "date"
colnames(df)[2] = "miles"
summary(df$miles)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  167174  262337  267256  262980  269839  285452
```

```r
travel_plot <- ggplot(df,aes(x = df$date, y = df$miles))+geom_point()+ggtitle("Plot of Mileage by Month

travel_plot
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```



After doing a quick check, we see no anomaly in the data. The monthly miles ranges from 167174 (during covid) to 285452. The monthly plot shows a similar picture.

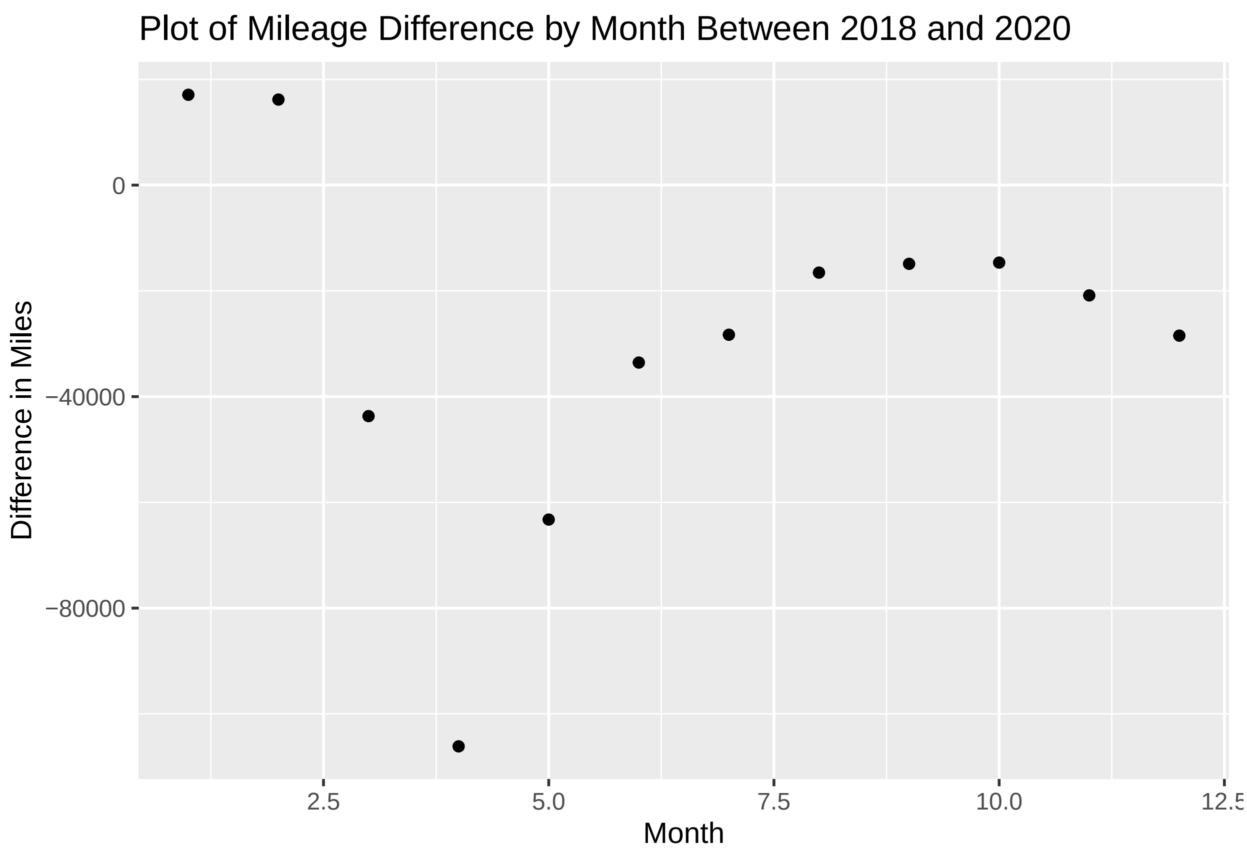
```r
jan_2018 = df[df$date == "2018-01-01",]$miles
feb_2018 = df[df$date == "2018-02-01",]$miles
mar_2018 = df[df$date == "2018-03-01",]$miles
apr_2018 = df[df$date == "2018-04-01",]$miles
may_2018 = df[df$date == "2018-05-01",]$miles
jun_2018 = df[df$date == "2018-06-01",]$miles
jul_2018 = df[df$date == "2018-07-01",]$miles
aug_2018 = df[df$date == "2018-08-01",]$miles
sep_2018 = df[df$date == "2018-09-01",]$miles
```

```r
oct_2018 = df[df$date == "2018-10-01",]$miles
nov_2018 = df[df$date == "2018-11-01",]$miles
dec_2018 = df[df$date == "2018-12-01",]$miles

jan_2020 = df[df$date == "2020-01-01",]$miles
feb_2020 = df[df$date == "2020-02-01",]$miles
mar_2020 = df[df$date == "2020-03-01",]$miles
apr_2020 = df[df$date == "2020-04-01",]$miles
may_2020 = df[df$date == "2020-05-01",]$miles
jun_2020 = df[df$date == "2020-06-01",]$miles
jul_2020 = df[df$date == "2020-07-01",]$miles
aug_2020 = df[df$date == "2020-08-01",]$miles
sep_2020 = df[df$date == "2020-09-01",]$miles
oct_2020 = df[df$date == "2020-10-01",]$miles
nov_2020 = df[df$date == "2020-11-01",]$miles
dec_2020 = df[df$date == "2020-12-01",]$miles
jan_2021 = df[df$date == "2021-01-01",]$miles


jan_diff = jan_2020 - jan_2018
feb_diff = feb_2020 - feb_2018
mar_diff = mar_2020 - mar_2018
apr_diff = apr_2020 - apr_2018
may_diff = may_2020 - may_2018
jun_diff = jun_2020 - jun_2018
jul_diff = jul_2020 - jul_2018
aug_diff = aug_2020 - aug_2018
sep_diff = sep_2020 - sep_2018
oct_diff = oct_2020 - oct_2018
nov_diff = nov_2020 - nov_2018
dec_diff = dec_2020 - dec_2018
```

```r
month = c(1:12)
diff = c(jan_diff,feb_diff,mar_diff,apr_diff,may_diff,jun_diff,jul_diff,aug_diff,sep_diff,oct_diff,nov_

diff_df = data.frame(month,diff)
diff_df
```

```
##    month    diff
## 1      1   17091
## 2      2   16183
## 3      3  -43686
## 4      4 -106146
## 5      5  -63252
## 6      6  -33554
## 7      7  -28293
## 8      8  -16542
## 9      9  -14882
## 10    10  -14650
## 11    11  -20854
## 12    12  -28464
```

```r
ggplot(diff_df,aes(x = month, y = diff))+geom_point()+ggtitle("Plot of Mileage Difference by Month Betw
```



- Comparing monthly miles driven in 2018 to the same months during the pandemic:

6.0.1 Raw Data

```
apr_2020
```

```
## [1] 167174
```

```
apr_2018
```

```
## [1] 273320
```

```
apr_2020 - apr_2018
```

```
## [1] -106146
```

```
apr_2020 / apr_2018 - 1
```

```
## [1] -0.388358
```

```
jan_2020
```

```
## [1] 285452
```

```
jan_2018
```

```
## [1] 268361
```

```
jan_2020 - jan_2018
```

```
## [1] 17091
```

```r
jan_2020 / jan_2018 - 1
```

```
## [1] 0.0636866
```

###Normalizated to per 100000 capita: The population in 2018 was 326.8 million. The population in 2020 was 329.5 million.

```r
apr_2020
```

```
## [1] 167174
```

```r
apr_2020/329.5
```

```
## [1] 507.3566
```

```r
apr_2018
```

```
## [1] 273320
```

```r
apr_2018/326.8
```

```
## [1] 836.3525
```

```r
apr_2020/329.5 - apr_2018/326.8
```

```
## [1] -328.9959
```

```r
0.00093498*(apr_2020/329.5 - apr_2018/326.8)*100000
```

```
## [1] -30760.46
```

```r
(apr_2020/329.5 - apr_2018/326.8)/(apr_2018/326.8)
```

```
## [1] -0.3933699
```

```r
jan_2020
```

```
## [1] 285452
```

```r
jan_2020/329.5
```

```
## [1] 866.3187
```

```r
jan_2018
```

```
## [1] 268361
```

```r
jan_2018/326.8
```

```
## [1] 821.1781
```

```r
jan_2020/329.5-jan_2018/326.8
```

```
## [1] 45.14057
```

```r
0.00093498*(jan_2020/329.5-jan_2018/326.8)*100000
```

```
## [1] 4220.553
```

```r
(jan_2020/329.5-jan_2018/326.8)/(jan_2018/326.8)
```

```
## [1] 0.0549705
```

```r
summary(my_panel$total_fatal_per_pop)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     6.20   14.38   18.43   18.92   22.77   53.32
```

The month of April 2020 demonstrated the largest decrease in driving, from 273320 in April 2018 to 167174 in April 2020. This was a 38.8% decrease. The month of January 2020 demonstrated the largest increase in driving, from 268361 in January 2018 to 285462 in January 2020. This was a 6.4% increase.

If we normalize our results to per 100000 capita: The month of April 2020 demonstrated the largest decrease in miles per capita, from 836.35 in April 2018 to 507.36 in April 2020. This was a 39.3% decrease. The month of January 2020 demonstrated the largest increase in miles per capita, from 821.18 in January 2018 to 866.32 in January 2020. This was a 5.5% increase.

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.

The fixed effect model in part 6 shows that the slope of the variable “miles driven per capita” is 0.00093498 with a statistically significantly p-value. If the number of miles driven per capita were to increase by as much as 45.14, the difference between January 2018 and January 2020, the model predicts that the number of traffic fatalities would increase by about 4220 per 100000 capita.

- 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.

The fixed effect model in part 6 shows that the slope of the variable “miles driven per capita” is 0.00093498 with a statistically significantly p-value. If the number of miles driven per capita were to decrease by as much as 329, the difference between April 2018 and April 2020, the model predicts that the number of traffic fatalities would decrease by about 30760 per 100000 capita.

7 (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?