Lab 3: Panel Models

US Traffic Fatalities: 1980 - 2004

Contents

1 U.S. traffic fatalities: 1980-2004

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

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

2.1 (5 points) Load the data and produce useful features.

```
# For the fractions, we are taking the majority as a speed limit
# We skipped year_of_observation since there a year column which aligns with dx
df <- data %>%
  mutate(speed_limit = ifelse(sl55 >= 0.5, '55',
                        ifelse(sl65 >= 0.5, '65',
                        ifelse(sl70 >= 0.5, '70',
                        ifelse(sl75 >= 0.5, '75',
                        ifelse(slnone >= 0.5, 'none', '0')
                        ))))) %>%
           mutate(speed_limit=factor(speed_limit,
                  levels=c('55', '65', '70', '75', 'none')),
         blood_alcohol_limit_10 = ifelse(bac10 >= 0.5, 1, 0),
         blood_alcohol_limit_08 = ifelse(bac08 >= 0.5, 1, 0)) %>%
  mutate(bac=ifelse(blood_alcohol_limit_10==1, '10',
             ifelse(blood_alcohol_limit_08==1, '8', 'none'))) %>%
  mutate(bac=factor(bac, levels=c('none', '10', '8'))) %>%
  select(!c((sl55:slnone), (d80:d04), bac10, bac08)) %>% # Excluding
  rename(minimum_drinking_age = minage, zero_tolerance_law = zerotol,
         graduated_drivers_license_law = gdl, per_se_law = perse,
         total_fatalities = totfat, nighttime_fatalities = nghtfat,
         weekend_fatalities = wkndfat, total_fatalities_per_100M_miles = totfatpvm,
         nighttime_fatalities_per_100M_miles = nghtfatpvm,
         weekend_fatalities_per_100M_miles = wkndfatpvm,
         state_population =  statepop, total_fatalities_rate = totfatrte,
         nighttime_fatalities_rate = nghtfatrte,
         weekend_fatalities_rate = wkndfatrte,
         vehicle_miles_traveled = vehicmiles, unemployment_rate = unem,
         population_aged_14_to_24_rate = perc14_24,
         speed_limit_70_plus = sl70plus,
         seat_belt = seatbelt,
         primary_seatbelt_law = sbprim, secondary_seatbelt_law = sbsecon,
         miles_driven_per_capita = vehicmilespc) %>%
  mutate(speed_limit_70_plus =
           ifelse(speed_limit_70_plus>0.5, 1, 0)
           ) %>%
  mutate(seat_belt_law =
           ifelse(seat_belt==0, 'none',
           ifelse(seat_belt==2, 'secondary',
           ifelse(seat_belt==1, 'primary', 'na')))) %>%
  mutate(seat_belt_law=factor(seat_belt_law,
                              levels=c('none', 'secondary', 'primary')),
         ) %>%
```

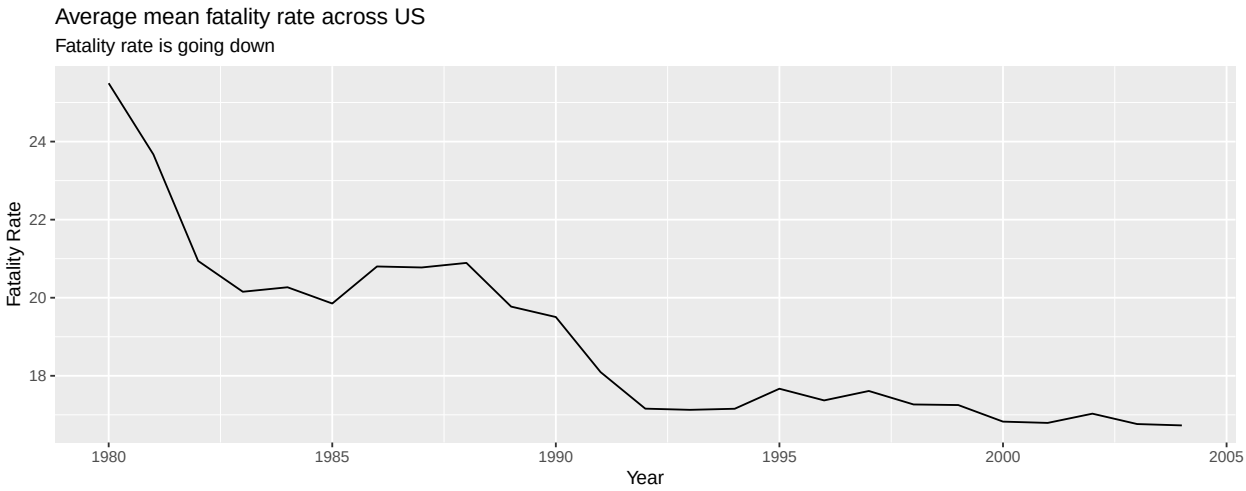
```
  mutate(per_se_law=round(per_se_law, 0)) %>%
  mutate(per_se_law=factor(per_se_law, levels=c(0, 1))) %>%
  mutate(log_total_fatalities_rate = log10(total_fatalities_rate))


# Adding states to the dataframe
state_df <- data.frame("index" = 1:51,
                       "state_name" = sort(c(state.name, "District of Columbia")))
main_df <- merge(df, state_df, by.x = 'state', by.y ='index')

pdata <- pdata.frame(main_df, index=c("state", "year"))
```

2.2 (5 points) Provide a description of the basic structure of the dataset.

The data contains state-level information on the total traffic fatalities rate in a specific year. The data also contains state law information that could have an impact on the traffic fatalities rate, like the speed limit, blood alcohol legal limit, seat belt laws, etc. 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 kindly 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.

2.3 (20 points) EDA

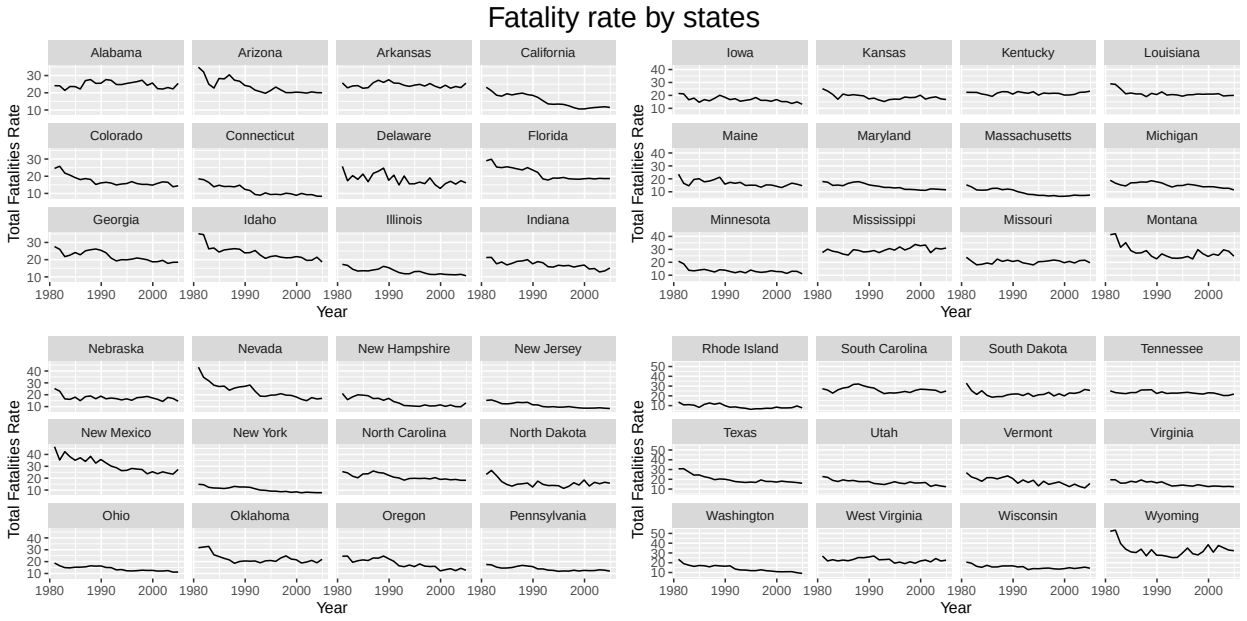
2.3.1 Average and state wide trends



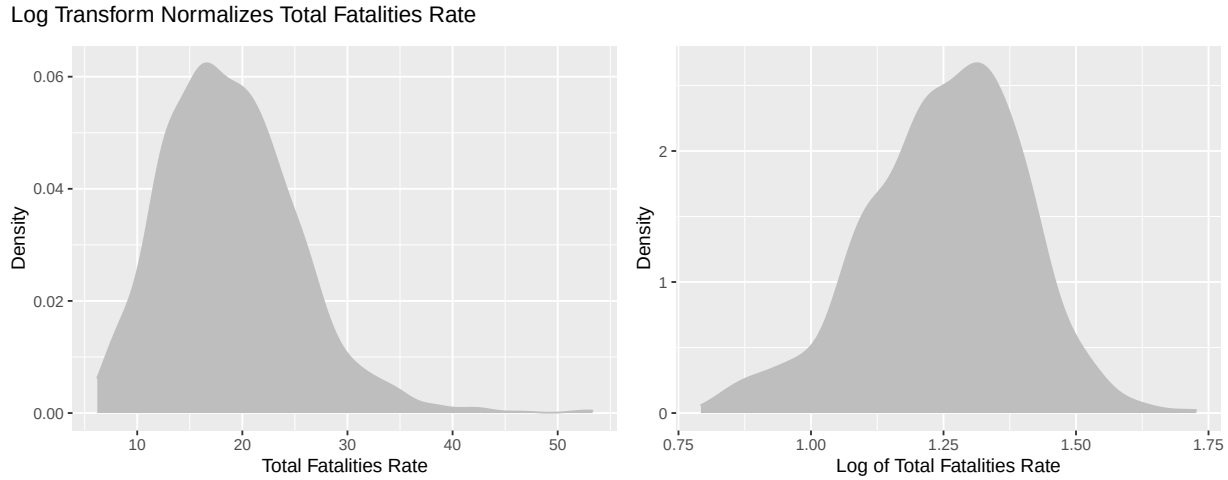
In the plot shown above, the fatality rates across states have been aggregated and the average fatality rate is plotted against time. The plot shows that, over time, average fatality rate has decreased in the United

States, beginning at an average of approximately 25 fatalities in 1980 to an average of about 17 fatalities in 2004, per 100,000 people.

However, given that the data is aggregated across states and the fact that the driving conditions and laws vary significantly across states is common knowledge, the graph shown above enables only a limited understanding of the trend in traffic fatalities across the United States over time.



Splitting the data by state provides the additional context needed to understand the trend in traffic fatalities. For most states, as is expected, the fatality rates decreased over the years. However, for some states, like Alabama and Arkansas, the trend does not show much change over time. Mississipi, on the other hand, shows an increase in fatality rate over time.



The `total_fatalities_rate` variable follows a right-skewed distribution. In the plot shown above, we see that the log-transformation helps make the distribution more normal

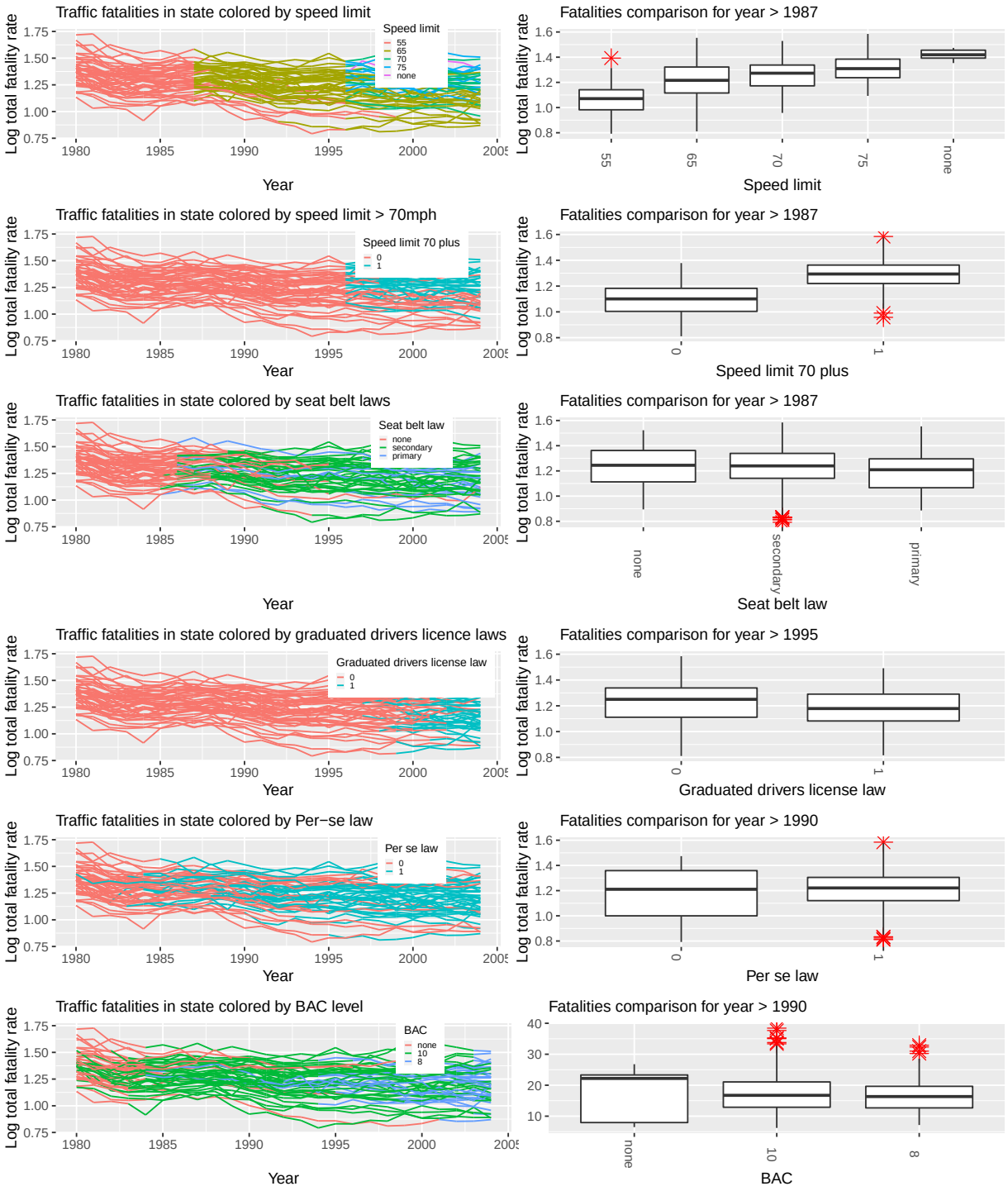
2.3.2 Description of variables

2.3.2.1 Factor variables

Referring to the graph on the following page

- The top-left panel shows the relationship between the highway speed limit and the logged `total_fatalities_rate` variable. Before 1987, the highway speed limit was uniformly set to 55mph across all states. Since then, different states have adopted different speed limits. In 1997, there was a significant increase in highway speeds across multiple states. The box plot in the top-right panel compares the fatality rate across different speed limits filtered for years greater than 1987. We see that increasing speed limits are associated with increased fatality rates. As there are states with no speed limit, this variable has been treated as a factor.
- Thresholding speed limits for greater than or lower than 70mph (second panel from the top) shows a similar pattern of higher speeds being associated with higher fatality rates.
- Seat belts started to become mandatory, across states, beginning mid-to-late 80s, and today there is only one state which does not mandate seat belts. Primary laws are the strictest and allow police to ticket drivers and passengers who are not wearing a proper safety restraint, even if that is the only traffic violation they are committing. Secondary seat belt laws, on the other hand, do not grant law enforcement officials the right to ticket drivers or passengers for failing to wear a safety restraint unless another traffic violation has occurred. There are 15 states with secondary seat belt laws. Source: https://www.cooper-law-firm.com/what-is-the-difference-between-primary-and-secondary-seat-belt-laws/.
- The graduated drivers licence law was beginning to be introduced in the late 90's. The box plot, which has been filtered for years greater than 1995, suggests that even for that time frame, there is a reduction in fatality rate between the two groups.
- Some states had Per-Se laws before the start of the data in 1980 and some still did not have Per-Se laws in 2004. There is a gradual increase in the adoption of the law from 1980 to about the 2000s. Surprisingly, there is an increase in fatality rates in comparison of data with the Per-Se law as compared to without it.
- Lastly, most states had adopted a blood alcohol limit by the mid 80s, with two states choosing a limit only in 2002.

Time series and distribution comparisons of factor variables

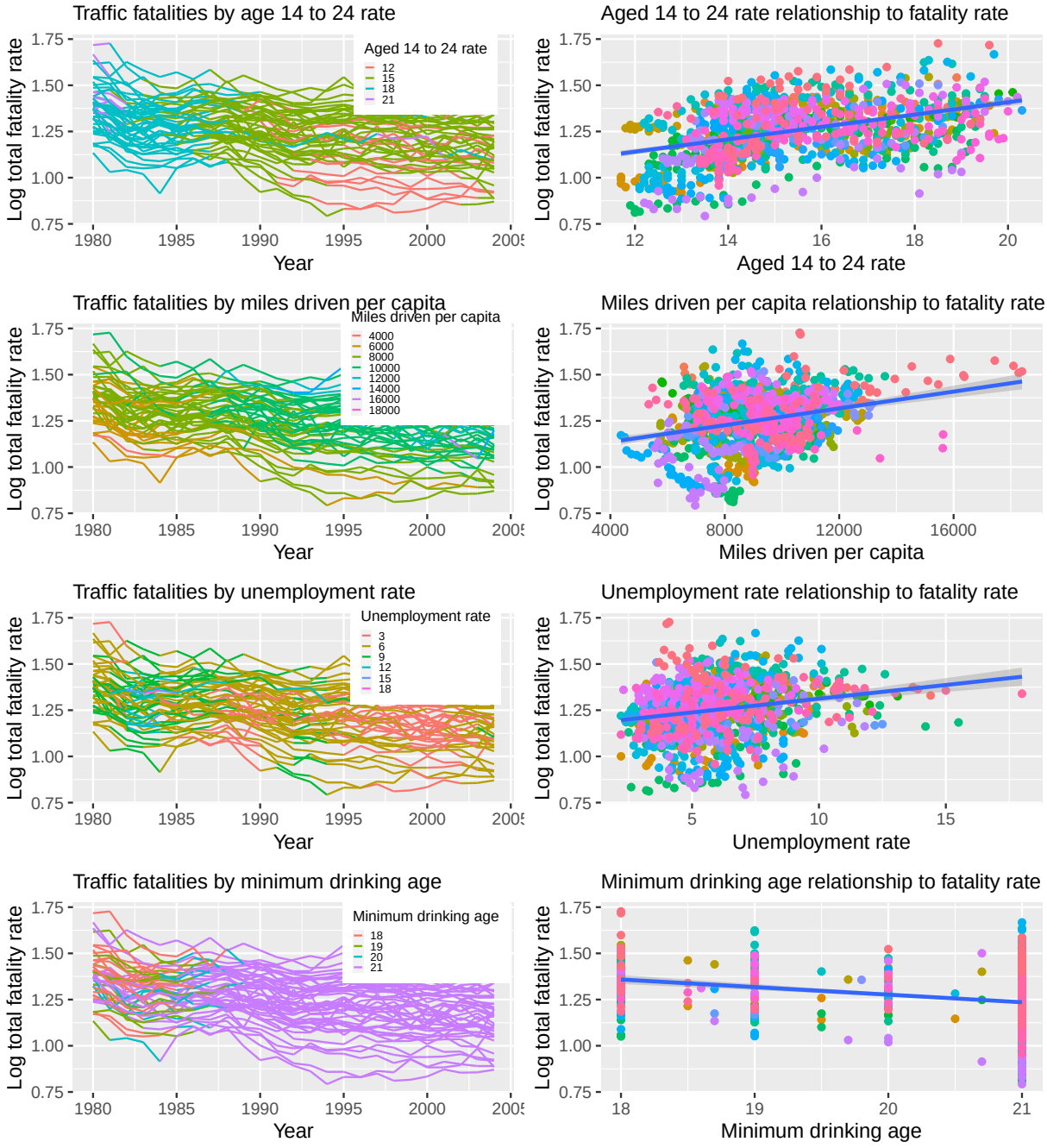


2.3.2.2 Description of Continous Variables

- There is a decrease in the percentage of 14 to 24 year-olds in the population over time. This is correlated to the decrease in the fatalities during that time period.

- Miles driven per capita has a positive relationship with total fatality rate. An increase in miles driven is associated with an increase in fatality rate
- There is an increase in fatality rate with an increase in unemployment rate.
- The minimum drinking age has been 21 in most states since the late 80s. There is a general decrease in fatality rate with an increased minimum drinking age but there also has been a general decrease in fatality rates during the time period when the age limits were changed.

Time series and correlation comparison of continous variables colored by variable values



3 (15 points) Preliminary Model

3.0.1 Preliminary Model Creation

```
#  Pooled OLS model
pooled_ols <- plm(log_total_fatalities_rate ~ year, data = pdata,
                  index = c("state", "year"),
                  effect = "individual", model = "pooling")
summary(pooled_ols)
```

```
## Pooling Model
##
## Call:
## plm(formula = log_total_fatalities_rate ~ year, data = pdata,
##     effect = "individual", model = "pooling", index = c("state",
##         "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.4183307 -0.0961280  0.0043659  0.1008470  0.3770988
##
## Coefficients:
##              Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  1.387907   0.020400 68.0351 < 2.2e-16 ***
## year1981    -0.034213   0.028850 -1.1859 0.2359037
## year1982    -0.086671   0.028850 -3.0042 0.0027191 **
## year1983    -0.102159   0.028850 -3.5411 0.0004141 ***
## year1984    -0.098087   0.028850 -3.3999 0.0006966 ***
## year1985    -0.105539   0.028850 -3.6582 0.0002652 ***
## year1986    -0.085474   0.028850 -2.9627 0.0031107 **
## year1987    -0.086297   0.028850 -2.9912 0.0028363 **
## year1988    -0.082018   0.028850 -2.8429 0.0045473 **
## year1989    -0.107768   0.028850 -3.7355 0.0001963 ***
## year1990    -0.116324   0.028850 -4.0320 5.886e-05 ***
## year1991    -0.149274   0.028850 -5.1742 2.690e-07 ***
## year1992    -0.174714   0.028850 -6.0560 1.875e-09 ***
## year1993    -0.174832   0.028850 -6.0601 1.830e-09 ***
## year1994    -0.177185   0.028850 -6.1416 1.116e-09 ***
## year1995    -0.167169   0.028850 -5.7945 8.794e-09 ***
## year1996    -0.173498   0.028850 -6.0138 2.416e-09 ***
## year1997    -0.167622   0.028850 -5.8102 8.028e-09 ***
## year1998    -0.177863   0.028850 -6.1651 9.666e-10 ***
## year1999    -0.180014   0.028850 -6.2397 6.108e-10 ***
## year2000    -0.189762   0.028850 -6.5776 7.181e-11 ***
## year2001    -0.189010   0.028850 -6.5515 8.500e-11 ***
## year2002    -0.185324   0.028850 -6.4238 1.927e-10 ***
## year2003    -0.190996   0.028850 -6.6204 5.437e-11 ***
## year2004    -0.194794   0.028850 -6.7520 2.286e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Total Sum of Squares:    26.854
## Residual Sum of Squares: 23.471
## R-Squared:      0.12598
## Adj. R-Squared: 0.10813
## F-statistic: 7.05692 on 24 and 1175 DF, p-value: < 2.22e-16
```

Starting from a linear model gives us an easy and intuitive way to understand the overall trend pattern over the years via the coefficients on the year dummy variables. In later sections, the report details panel data analysis that can be compared against this benchmark.

This model explains the log of fatality rate over different years compared to the baseline year, which is 1980. All coefficients except for the 1981 year are statistically significant. From this model, we learn that in all years following the baseline up to 2004, the fatality rate goes down compared to the baseline year 1980. The decline accelerates as the year increases but not consistently, as seen, for example, between 1986 - 1988, where the fatality rate increases compared to previous years. However, it returns to the decline track/trend in 1989. The higher, in absolute value, coefficients over the years means that driving becomes safer as the fatality rate keeps decreasing. For example, the fatality rate, which was 24.43 as of 1980, became 0.77 in 1990, 0.65 in 2000, and 0.64 in 2004, showing that driving over the years has become safer. In other words, 8.83 fewer people are predicted to be involved in traffic fatalities out of 100,000 people in 2004 than in 1980 as per this model.

Note that we are ignoring unobserved Heterogeneity and the group structure by taking each entry as a separate observation. Because of that, residuals generally correlate across time and have heteroskedasticity across and/or within groups. Heteroscedastic residuals are a violation of the OLS Homoscedasticity assumption, which will make it difficult to trust the standard error. As a result, the confidence interval can not be trusted as it can be too wide or narrow. Also, the independence assumption (no autocorrelation) is violated since we did not accommodate the lag/trend component, which makes the OLS estimates to be unreliable; in other words, our OLS estimator is not the Best Linear Unbiased Estimator.

4 (15 points) Expanded Model

4.1 Transformation (treatment) of Variables

- As described in a previous section, a log transformation was performed on the response variable to make the distribution closer to normal.
- States where highway speeds were made over 70mph during the middle of the year contained a fractional value for that year. The fraction was cut off at the value of 0.5 to make this a binary variable. As there is no meaningful interpretation for this variable as a continuous value, this threshold was necessary to convert it into a factor.
- A state can have a primary, secondary or no seat belt laws. The seat_belt_law reflects these three factors combined into one variable.
- Graduated drivers licence law and per-se law parameters both have fractional value for years where the law was implemented mid-year. The fractions were cut off at the value of 0.5 to make this a binary variable. As there is no meaningful interpretation for these variables as a continuous value, this threshold was necessary to convert them into a factor.
- Although Blood alcohol content (BAC) levels of 0.08% or 0.1% lends itself to a numeric interpretation, there is no numeric value associated with no BAC limit. For this reason BAC has been treated as a factor variable with levels none, 10 and 8.
- None of the continuous variables, namely, unemployment rate, miles drive per-capita, proportion of 14 to 24 aged people in population required any transformation. This can be seen in the correlation plots in Figure y where their relationship to log fatality rates appear linear.

4.1.1 Pooled OLS Model Creation

```
expanded.ols.data <- main_df %>% select(
  c(log_total_fatalities_rate, bac, per_se_law, seat_belt_law,
    graduated_drivers_license_law, population_aged_14_to_24_rate,
    minimum_drinking_age, unemployment_rate, speed_limit_70_plus,
    miles_driven_per_capita, year, state))

main_p <- pdata.frame(expanded.ols.data, index=c("state", "year"))
expanded.ols <- plm(log_total_fatalities_rate ~  year + bac +
                      population_aged_14_to_24_rate + miles_driven_per_capita +
                      unemployment_rate + speed_limit_70_plus + per_se_law +
                      seat_belt_law + graduated_drivers_license_law,
                    data = main_p,
                  index = c("state", "year"),
                  effect = "individual", model = "pooling")

summary(expanded.ols)
```

```
## Pooling Model
##
## Call:
## plm(formula = log_total_fatalities_rate ~ year + bac + population_aged_14_to_24_rate +
##     miles_driven_per_capita + unemployment_rate + speed_limit_70_plus +
##     per_se_law + seat_belt_law + graduated_drivers_license_law,
##     data = main_p, effect = "individual", model = "pooling",
##     index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -0.34758836 -0.05686774  0.00052907  0.06279643  0.25888532
##
## Coefficients:
##                                  Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)                    6.5776e-01  5.5569e-02  11.8367 < 2.2e-16 ***
## year1981                      -4.0821e-02  1.8599e-02  -2.1948   0.02838 *
## year1982                      -1.3080e-01  1.9183e-02  -6.8182 1.477e-11 ***
## year1983                      -1.5099e-01  1.9581e-02  -7.7113 2.659e-14 ***
## year1984                      -1.2006e-01  1.9720e-02  -6.0885 1.545e-09 ***
## year1985                      -1.3316e-01  2.0105e-02  -6.6233 5.354e-11 ***
## year1986                      -1.2207e-01  2.0907e-02  -5.8386 6.819e-09 ***
## year1987                      -1.3467e-01  2.1723e-02  -6.1997 7.839e-10 ***
## year1988                      -1.3907e-01  2.2769e-02  -6.1076 1.376e-09 ***
## year1989                      -1.7282e-01  2.3648e-02  -7.3079 5.026e-13 ***
## year1990                      -1.9557e-01  2.4190e-02  -8.0846 1.550e-15 ***
## year1991                      -2.4396e-01  2.4741e-02  -9.8605 < 2.2e-16 ***
## year1992                      -2.8991e-01  2.5218e-02 -11.4964 < 2.2e-16 ***
## year1993                      -2.8707e-01  2.5525e-02 -11.2464 < 2.2e-16 ***
## year1994                      -2.8368e-01  2.5976e-02 -10.9207 < 2.2e-16 ***
## year1995                      -2.7658e-01  2.6580e-02 -10.4057 < 2.2e-16 ***
## year1996                      -3.3148e-01  2.7527e-02 -12.0418 < 2.2e-16 ***
```

```
## year1997                      -3.4014e-01  2.7968e-02 -12.1615 < 2.2e-16 ***
## year1998                      -3.6607e-01  2.8402e-02 -12.8887 < 2.2e-16 ***
## year1999                      -3.7110e-01  2.8800e-02 -12.8855 < 2.2e-16 ***
## year2000                      -3.8060e-01  2.9269e-02 -13.0035 < 2.2e-16 ***
## year2001                      -4.0014e-01  2.9859e-02 -13.4009 < 2.2e-16 ***
## year2002                      -4.1553e-01  3.0151e-02 -13.7815 < 2.2e-16 ***
## year2003                      -4.2378e-01  3.0308e-02 -13.9825 < 2.2e-16 ***
## year2004                      -4.2116e-01  3.0978e-02 -13.5954 < 2.2e-16 ***
## bac10                         -9.9381e-03  8.7474e-03  -1.1361   0.25614
## bac8                          -2.6054e-02  1.1804e-02  -2.2073   0.02749 *
## population_aged_14_to_24_rate  6.8653e-03  2.7564e-03   2.4906   0.01289 *
## miles_driven_per_capita        6.8781e-05  2.1301e-06  32.2903 < 2.2e-16 ***
## unemployment_rate              1.6720e-02  1.7501e-03   9.5533 < 2.2e-16 ***
## speed_limit_70_plus            9.6254e-02  9.7589e-03   9.8632 < 2.2e-16 ***
## per_se_law1                   -6.8721e-03  6.6160e-03  -1.0387   0.29915
## seat_belt_lawsecondary         9.1036e-03  9.6452e-03   0.9438   0.34545
## seat_belt_lawprimary          -8.6245e-04  1.1029e-02  -0.0782   0.93768
## graduated_drivers_license_law -1.2209e-03  1.1840e-02  -0.1031   0.91789
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    26.854
## Residual Sum of Squares: 9.6296
## R-Squared:      0.64141
## Adj. R-Squared: 0.63095
## F-statistic: 61.29 on 34 and 1165 DF, p-value: < 2.22e-16
```

4.2 Interpretation of Results

4.2.1 How are the blood alcohol variables defined? Interpret the coefficients that you estimate for this concept.

In an earlier section we have defined our treatment of the BAC variable as a factor with levels 'none', '10' and '8'. In this model we note that the base level is no blood alcohol limit. We note that bac value of 0.08% is statistically significant and 0.1% is not statistically significant. The model suggests that setting a blood alcohol limit of 0.1% is associated with, ceteris paribus, a 0.98 times decrease in fatality rate as compared to no BAC limit. The model suggests that setting a blood alcohol limit of 0.08% is associated with, ceteris paribus, a 0.94 times decrease in fatality rate as compared to no BAC limit.

4.2.2 Do *per se laws* have a negative effect on the fatality rate?

We note that per-se law is not a statistically significant parameter. The model suggests that having a per-se law is associated with, ceteris paribus, a 0.98 times decrease in fatality rate as compared to not having per-se law.

4.2.3 Does having a primary seat belt law reduce fatality rates?

We note that the seat belt law factors are not statistically significant. We also note that the implementation of the primary law leads to, ceteris paribus, a 0.998 times decrease in fatality rate which is practically insignificant.

5 (15 points) State-Level Fixed Effects

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

Model estimation for a fixed effect (within) model.

```
expanded.within <- plm(log_total_fatalities_rate ~  bac + year  +
                       population_aged_14_to_24_rate + miles_driven_per_capita +
                       unemployment_rate + speed_limit_70_plus + per_se_law +
                       seat_belt_law + graduated_drivers_license_law,
                     data = main_p,
                   index = c("state", "year"),
                   effect = "individual", model = "within")

summary(expanded.within)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_total_fatalities_rate ~ bac + year + population_aged_14_to_24_rate +
##     miles_driven_per_capita + unemployment_rate + speed_limit_70_plus +
##     per_se_law + seat_belt_law + graduated_drivers_license_law,
##     data = main_p, effect = "individual", model = "within", index = c("state",
##         "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.1835098 -0.0223288  0.0020551  0.0227251  0.1279559
##
## Coefficients:
##                                Estimate  Std. Error  t-value  Pr(>|t|)
## bac10                       -1.1165e-02  4.9828e-03  -2.2407  0.025241 *
## bac8                        -1.6135e-02  7.1737e-03  -2.2492  0.024694 *
## year1981                    -2.5783e-02  7.9252e-03  -3.2533  0.001175 **
## year1982                    -5.0669e-02  8.4888e-03  -5.9689 3.204e-09 ***
## year1983                    -6.1384e-02  8.7911e-03  -6.9825 4.964e-12 ***
## year1984                    -7.8231e-02  8.9205e-03  -8.7698 < 2.2e-16 ***
## year1985                    -8.6982e-02  9.2924e-03  -9.3605 < 2.2e-16 ***
## year1986                    -6.6274e-02  9.9105e-03  -6.6872 3.585e-11 ***
## year1987                    -8.1819e-02  1.0632e-02  -7.6954 3.088e-14 ***
## year1988                    -9.1423e-02  1.1516e-02  -7.9387 4.941e-15 ***
## year1989                    -1.2221e-01  1.2261e-02  -9.9670 < 2.2e-16 ***
## year1990                    -1.2726e-01  1.2730e-02  -9.9971 < 2.2e-16 ***
## year1991                    -1.4334e-01  1.3067e-02 -10.9697 < 2.2e-16 ***
## year1992                    -1.6845e-01  1.3465e-02 -12.5101 < 2.2e-16 ***
## year1993                    -1.7549e-01  1.3715e-02 -12.7957 < 2.2e-16 ***
## year1994                    -1.8929e-01  1.4049e-02 -13.4739 < 2.2e-16 ***
## year1995                    -1.8602e-01  1.4479e-02 -12.8476 < 2.2e-16 ***
## year1996                    -2.0517e-01  1.5272e-02 -13.4345 < 2.2e-16 ***
## year1997                    -2.1113e-01  1.5630e-02 -13.5084 < 2.2e-16 ***
## year1998                    -2.3185e-01  1.5945e-02 -14.5404 < 2.2e-16 ***
```

```
## year1999                      -2.3797e-01  1.6125e-02 -14.7580 < 2.2e-16 ***
## year2000                      -2.5039e-01  1.6352e-02 -15.3122 < 2.2e-16 ***
## year2001                      -2.4326e-01  1.6634e-02 -14.6247 < 2.2e-16 ***
## year2002                      -2.2996e-01  1.6803e-02 -13.6862 < 2.2e-16 ***
## year2003                      -2.3124e-01  1.6894e-02 -13.6878 < 2.2e-16 ***
## year2004                      -2.4447e-01  1.7323e-02 -14.1121 < 2.2e-16 ***
## population_aged_14_to_24_rate  9.0309e-03  1.8219e-03   4.9569 8.270e-07 ***
## miles_driven_per_capita        2.6879e-05  2.1286e-06  12.6277 < 2.2e-16 ***
## unemployment_rate             -1.2761e-02  1.1601e-03 -11.0002 < 2.2e-16 ***
## speed_limit_70_plus            2.3830e-02  4.9909e-03   4.7748 2.038e-06 ***
## per_se_law1                   -2.3217e-02  4.2933e-03  -5.4078 7.801e-08 ***
## seat_belt_lawsecondary        -6.6982e-04  4.8314e-03  -0.1386  0.889762
## seat_belt_lawprimary          -2.0458e-02  6.5736e-03  -3.1122  0.001904 **
## graduated_drivers_license_law -7.3527e-03  5.6051e-03  -1.3118  0.189860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    6.0213
## Residual Sum of Squares: 1.6682
## R-Squared:      0.72294
## Adj. R-Squared: 0.70287
## F-statistic: 85.8017 on 34 and 1118 DF, p-value: < 2.22e-16
```

5.2 Interpretation of the Model

Table 1:

	log_total_fatalities_rate	
	Pooled OLS	Within
	(1)	(2)
bac10	−0.010	−0.011**
	(0.009)	(0.005)
bac8	−0.026**	−0.016**
	(0.012)	(0.007)
per_se_law1	−0.007	−0.023***
	(0.007)	(0.004)
seat_belt_lawprimary	−0.001	−0.020***
	(0.011)	(0.007)
N	1,200	1,200
R^2	0.641	0.723
Notes:	***Significant at the 1 percent level.	
	**Significant at the 5 percent level.	
	*Significant at the 10 percent level.	

5.2.1 What do you estimate for coefficients on the blood alcohol variables? How do the coefficients on the blood alcohol variables change, if at all?

We note that BAC value of 0.1% and 0.08% are both statistically significant parameters.

This model suggests that setting a blood alcohol limit of 0.1% is associated with, ceteris paribus, a 0.97 times decrease in log fatality rate as compared to no BAC limit. Compared to the OLS model which showed that setting a blood alcohol limit of 0.1% is associated with, ceteris paribus, a 0.98 times decrease in fatality rate as compared to no BAC limit.

Additionally, setting a blood alcohol limit of 0.08% is associated with, ceteris paribus, a 0.96 times decrease in fatality rate as compared to no BAC limit. Compared to the OLS model which showed that setting a blood alcohol limit of 0.08% is associated with, ceteris paribus, a 0.96 times decrease in fatality rate as compared to no BAC limit.

5.2.2 What do you estimate for coefficients on per se laws? How do the coefficients on per se laws change, if at all?

We note that per-se law is a statistically significant parameter. This model suggests that having a per-se law is associated with, ceteris paribus, a 0.95 times decrease in fatality rate as compared to not having per-se law. Compared to the OLS model which showed that having a per-se law is associated with, ceteris paribus, a 0.98 times decrease in fatality rate as compared to not having per-se law

5.2.3 What do you estimate for coefficients on primary seat-belt laws? How do the coefficients on primary seatbelt laws change, if at all?

In this model we note that seat belt primary law is a significant factor. This model suggests that having a primary seat belt law is associated with, ceteris paribus, a 0.95 unit decrease in fatality rate as compared to not having primary seat belt law. Compared to the OLS model for which the parameter was not statistically significant and had a close to 0 value of 1

5.3 Model Assumptions

5.3.1 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 variables 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

An assumption in an pooled OLS model is that the data is IID. Let's consider a data set where a sample of a large population was collected on different years. It is unlikely that a particular individual sample is measured twice. In such a circumstance a pooled OLS model would be applicable. However in this data set, the individual is the state 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. We perform a F-test between the pooled and the fixed effect model to check for fixed effects. The

null hypothesis is that there are no fixed effects and the alternate hypothesis is that there are fixed effects. We test against an alpha of 0.05

```
res <- pFtest(expanded.within, expanded.ols)
```

With a p-value of 0 less than 0.05, we reject the null hypothesis of no fixed effects. This means we should include state and/or time fixed effects in our model. Hence the fixed effect model is better for this scenario.

6 (10 points) Consider a Random Effects Model

6.1 Random Effect Model Assumptions

- There are no perfect linear relationships among the explanatory variables
- For each t, the expected value of the idiosyncratic error given the explanatory variables in all time periods and the unobserved effect is zero: $E(u_{it}|X_i, a_i) = 0$
- The expected value of a_i given all explanatory variables is a constant $E(a_i|X_i) = \beta_0$. This ensures that the fixed effects is uncorrelated to the independent variable at any time period, i.e. $cov(x_{it}, a_i) = 0$.

6.2 Estimating the Random Effect Model

```
expanded.re <- plm(log_total_fatalities_rate ~  bac + year  +
                     population_aged_14_to_24_rate + miles_driven_per_capita +
                     unemployment_rate + speed_limit_70_plus + per_se_law +
                     seat_belt_law + graduated_drivers_license_law,
                   data = main_p,
                 index = c("state", "year"),
                 model = "random")

summary(expanded.re)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = log_total_fatalities_rate ~ bac + year + population_aged_14_to_24_rate +
##     miles_driven_per_capita + unemployment_rate + speed_limit_70_plus +
##     per_se_law + seat_belt_law + graduated_drivers_license_law,
##     data = main_p, model = "random", index = c("state", "year"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Effects:
##                     var  std.dev share
## idiosyncratic 0.001492 0.038629 0.232
## individual    0.004933 0.070238 0.768
## theta: 0.8907
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
```

```
## -0.20038250 -0.02385632  0.00023109  0.02427459  0.13173156
##
## Coefficients:
##                                   Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept)                     1.0848e+00  4.0710e-02  26.6472 < 2.2e-16 ***
## bac10                          -1.1910e-02  5.1011e-03  -2.3348  0.019553 *
## bac8                           -1.7389e-02  7.3360e-03  -2.3703  0.017774 *
## year1981                       -2.6305e-02  8.1578e-03  -3.2245  0.001262 **
## year1982                       -5.3662e-02  8.7278e-03  -6.1484 7.825e-10 ***
## year1983                       -6.4716e-02  9.0356e-03  -7.1624 7.927e-13 ***
## year1984                       -7.9928e-02  9.1675e-03  -8.7187 < 2.2e-16 ***
## year1985                       -8.8998e-02  9.5438e-03  -9.3253 < 2.2e-16 ***
## year1986                       -6.8770e-02  1.0173e-02  -6.7603 1.377e-11 ***
## year1987                       -8.4696e-02  1.0903e-02  -7.7681 7.970e-15 ***
## year1988                       -9.4564e-02  1.1799e-02  -8.0147 1.104e-15 ***
## year1989                       -1.2573e-01  1.2554e-02 -10.0147 < 2.2e-16 ***
## year1990                       -1.3167e-01  1.3028e-02 -10.1066 < 2.2e-16 ***
## year1991                       -1.4886e-01  1.3372e-02 -11.1327 < 2.2e-16 ***
## year1992                       -1.7506e-01  1.3773e-02 -12.7102 < 2.2e-16 ***
## year1993                       -1.8188e-01  1.4026e-02 -12.9669 < 2.2e-16 ***
## year1994                       -1.9522e-01  1.4366e-02 -13.5889 < 2.2e-16 ***
## year1995                       -1.9206e-01  1.4802e-02 -12.9748 < 2.2e-16 ***
## year1996                       -2.1214e-01  1.5609e-02 -13.5910 < 2.2e-16 ***
## year1997                       -2.1838e-01  1.5970e-02 -13.6740 < 2.2e-16 ***
## year1998                       -2.3955e-01  1.6289e-02 -14.7063 < 2.2e-16 ***
## year1999                       -2.4588e-01  1.6471e-02 -14.9276 < 2.2e-16 ***
## year2000                       -2.5832e-01  1.6704e-02 -15.4644 < 2.2e-16 ***
## year2001                       -2.5245e-01  1.6989e-02 -14.8601 < 2.2e-16 ***
## year2002                       -2.4041e-01  1.7157e-02 -14.0120 < 2.2e-16 ***
## year2003                       -2.4196e-01  1.7250e-02 -14.0264 < 2.2e-16 ***
## year2004                       -2.5478e-01  1.7689e-02 -14.4027 < 2.2e-16 ***
## population_aged_14_to_24_rate   9.1466e-03  1.8592e-03   4.9196 8.673e-07 ***
## miles_driven_per_capita         3.0075e-05  2.1269e-06  14.1405 < 2.2e-16 ***
## unemployment_rate              -1.1668e-02  1.1828e-03  -9.8650 < 2.2e-16 ***
## speed_limit_70_plus             2.5018e-02  5.1220e-03   4.8844 1.037e-06 ***
## per_se_law1                    -2.2411e-02  4.3815e-03  -5.1149 3.139e-07 ***
## seat_belt_lawsecondary         -6.2907e-04  4.9599e-03  -0.1268  0.899073
## seat_belt_lawprimary           -1.9878e-02  6.7212e-03  -2.9575  0.003101 **
## graduated_drivers_license_law -6.8574e-03  5.7606e-03  -1.1904  0.233887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    6.2703
## Residual Sum of Squares: 1.8422
## R-Squared:      0.7062
## Adj. R-Squared: 0.69762
## Chisq: 2800.26 on 34 DF, p-value: < 2.22e-16
```

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

```
res <- phtest(expanded.within, expanded.re)
```

With a p-value of $2.5404639 \times 10^{-75}$ much 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.

6.2.1 Consequences of inapproriately using a random effect model

A fixed effect model allows for an intercept per individual that is correlated to the independent variables. The fixed effect model normalizes each individual to their means to estimate the coefficients. A random effect model does not require this normalization to the mean due to the assumption of the uncorrelated individual fixed effects. When the assumption is valid then the random effect model acts like a partial pooling, the groups effect estimate will be partially based on groups that have abundant data. If such a correlation exists then there are two consequences for it:

- The estimate of the individual coefficients would be incorrect. Especially low-sample groups will have poor estimates.
- The residuals will have a serial correlation which makes the standard errors unreliable

7 (10 points) Model Forecasts

```
# Source
# U.S. Federal Highway Administration, Vehicle Miles Traveled [TRFVOLUSM227NFWA],
#   retrieved from FRED, Federal Reserve Bank of St. Louis;
# https://fred.stlouisfed.org/series/TRFVOLUSM227NFWA, November 29, 2022.
# Driving mile units in millions
miles.driven <-
  readr::read_csv("./data/TRFVOLUSM227NFWA.csv", show_col_types = FALSE) %>%
  rename(miles = TRFVOLUSM227NFWA, date = DATE)

largest.decrease <- 0
largest.increase <- 0
max <- -Inf
min <- Inf

# Comparing each month miles driven
for (i in 1:12) {
  val.18 <- miles.driven %>%
    filter(date == mdy(paste(i, 1, 2018, sep="/")))

  # Comparing 2018 with 2020 and 2021
  for (j in 2020:2021) {
    val.covid <- miles.driven %>%
      filter(date == mdy(paste(i, 1, j, sep="/")))

    diff <- val.covid$miles - val.18$miles
    if(diff < min) {
      largest.decrease <- mdy(paste(i, 1, j, sep="/"))
      min = diff
    }
    if(diff > max) {
      largest.increase <- mdy(paste(i, 1, j, sep="/"))
      max = diff
    }
```

```
  }
}

# Calculating percentage
prev <- miles.driven %>%
    filter(date == largest.decrease)
lower.percentage  <- min/prev$miles

prev <- miles.driven %>%
    filter(date == largest.increase)
higher.percentage <- max/prev$miles
```

The largest month-over-month decrease in driving was in April 2020 and it was lower by -64.14%. This was right as the pandemic was starting to make its way through the country and the response was to shut everything down. So a rapid decrease in driving would make sense. In contrast, the largest month-over-month increase in driving was in January 2020 and it was higher by 6.18%. This was right before the pandemic and the increase is not as severe as the pandemic decrease was.

```
# Source
# U.S. Bureau of Economic Analysis, Population [POPTHM],
# Retrieved from FRED, Federal Reserve Bank of St. Louis;
# https://fred.stlouisfed.org/series/POPTHM, December 3, 2022.
# Pop Units in thousands
us.pop.month <-
  readr::read_csv("./data/POPTHM.csv", show_col_types = FALSE) %>%
  rename(total_population = POPTHM, date = DATE) %>%
  mutate(date = as.Date(date, "%m/%d/%Y"))

# Filter to get population on date with largest increase
max_pop <- us.pop.month %>%
  filter(date == largest.increase)

# Filter to get population on date with largest decrease
min_pop <- us.pop.month %>%
  filter(date == largest.decrease)

# Calculate increase/decrease per capita based on min/max drive values and pop.
max_month_pop <- max_pop$total_population * 1000 # Units in thousands
min_month_pop <- min_pop$total_population * 1000
max_miles <- max * 1000000 # Units in millions
min_miles <- min * 1000000

max_miles_driven_pc <- max_miles / max_month_pop
min_miles_driven_pc <- min_miles / min_month_pop

# fatal rate - represents increases/decreases in the fatality rate
fatal_rate_covid_max <- round(coef(expanded.within)['miles_driven_per_capita']
                              * max_miles_driven_pc, 3)
fatal_rate_covid_min <- round(coef(expanded.within)['miles_driven_per_capita']
                              * min_miles_driven_pc, 3)
```

The FE model suggests that a one mile increase in the miles driven per capita is associated with, ceteris paribus, a 0.00003 log increase in the fatality rate. During the COVID boom, the largest increase in driving

was on January 2020 and it results, ceteris paribus, in an estimated increase of 48.61 miles driven per capita and this is expected to result in a 0.001 log increase in the fatality rate.

In contrast, the largest decrease in driving was on April 2020 and it results, ceteris paribus, in an estimated decrease of -324.14 miles driven per capita and this is expected to result in a -0.009 log decrease in the fatality rate.

8 (5 points) Evaluate Error

8.1 Analysis of Residuals

We test the residuals for serial correlation. The null hypothesis for the test is that the residuals are not serially correlated and the alternate hypothesis is that the residuals are serially correlated. We check against an α of 0.05.

```
pwd <- pdwtest(expanded.within)
```

With a p-value of $6.5724615 \times 10^{-46}$, we reject the null hypothesis of no serial correlation. This suggests that we must use robust standard errors for model parameters.

8.2 Hetroskedasticity Test

We perform a Breusch Pagan test for heteroskedasticity. The null hypothesis for the test is that the residuals are homoskedastic and the alternate hypothesis is that the residuals are hetroskedastic. We check against an α of 0.05.

```
pcd <- pcdtest(expanded.within, test = "lm")
```

With a p-value of $1.4895406 \times 10^{-146}$, we reject the null hypothesis of homoskedasticity. This suggests that heteroscedasticity is present and our residuals are not distributed with equal variance. We may be able to correct for this by using heteroskedasticity-consistent standard errors or robust standard errors.'

We compare the results of the Durbin Watson and Breusch-Godfrey test with order = 2. The null hypothesis is that there is no serial correlation of any order less than or equal to p.

```
pbg <- pbgtest(expanded.within, order = 2)
```

With a p-value of $3.5077171 \times 10^{-48}$, we reject the null hypothesis and conclude that autocorrelation exists among the residuals at some order less than or equal to p.

Using vcovHC, we calculate robust standard errors (white1), cluster robust standard errors (white2), arellano standard errors (arellano), and newey west standard errors (using vcovNW) for the coefficients within our FE model.

```
reg.se <- coef(summary(expanded.within))[1,2]
het.se <- sqrt(vcovHC(expanded.within, method = "white1", type = "HC0")[1,1])
cluster.se <- sqrt(vcovHC(expanded.within, method = "white2", type = "HC0")[1,1])
nw.se <- sqrt(vcovNW(expanded.within, type = "HC0", maxlag = 1)[1,1])
arrellano.se <- sqrt(vcovHC(expanded.within, method = "arellano", type = "HC0")[1,1])

data.frame(
```

```
  "Type" = c("Regular OLS", "Robust", "Cluster Robust", "Newey West", "Arellano"),
  "SE" = c(reg.se, het.se, cluster.se, nw.se, arrellano.se)
)
```

```
##              Type          SE
## 1    Regular OLS 0.004982780
## 2         Robust 0.005353132
## 3 Cluster Robust 0.004787973
## 4     Newey West 0.006195990
## 5       Arellano 0.008663517
```

Our analysis shows that the residuals in our model are both heterskedastic and have serial correlation. The results from the vcovHC analysis suggests that cluster robust standard errors are the most robust. Specifically, the standard errors for cluster robust standard errors are slightly smaller than the regular OLS errors and robust standard error, while quite a bit higher (e.g, greater than 0.003) compared to the other standard error types.

When using robust cluster standard errors, our analysis shows that the fixed effects model has the highest R-squared value among the pooled and expanded OLS counterparts, as well as the lowest standard errors. This is shown in the table in the following page.

Table 2:

	log_total_fatalities_rate		
	Pooled	Expanded OLS	Within
	(1)	(2)	(3)
bac10		−0.010	−0.011**
		(0.009)	(0.005)
bac8		−0.026**	−0.016**
		(0.012)	(0.007)
population_aged_14_to_24_rate		0.007**	0.009***
		(0.003)	(0.002)
miles_driven_per_capita		0.0001***	0.00003***
		(0.00000)	(0.00000)
unemployment_rate		0.017***	−0.013***
		(0.002)	(0.001)
speed_limit_70_plus		0.096***	0.024***
		(0.010)	(0.005)
per_se_law1		−0.007	−0.023***
		(0.007)	(0.004)
seat_belt_lawsecondary		0.009	−0.001
		(0.010)	(0.005)
seat_belt_lawprimary		−0.001	−0.020***
		(0.010)	(0.006)
graduated_drivers_license_law		−0.001	−0.007
		(0.012)	(0.005)
Constant	1.388***	0.658***	
	(0.020)	(0.058)	
N	1,200	1,200	1,200
R^2	0.126	0.641	0.723

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.