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)
#desc
```

(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.)
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?
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.

```
#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'
```

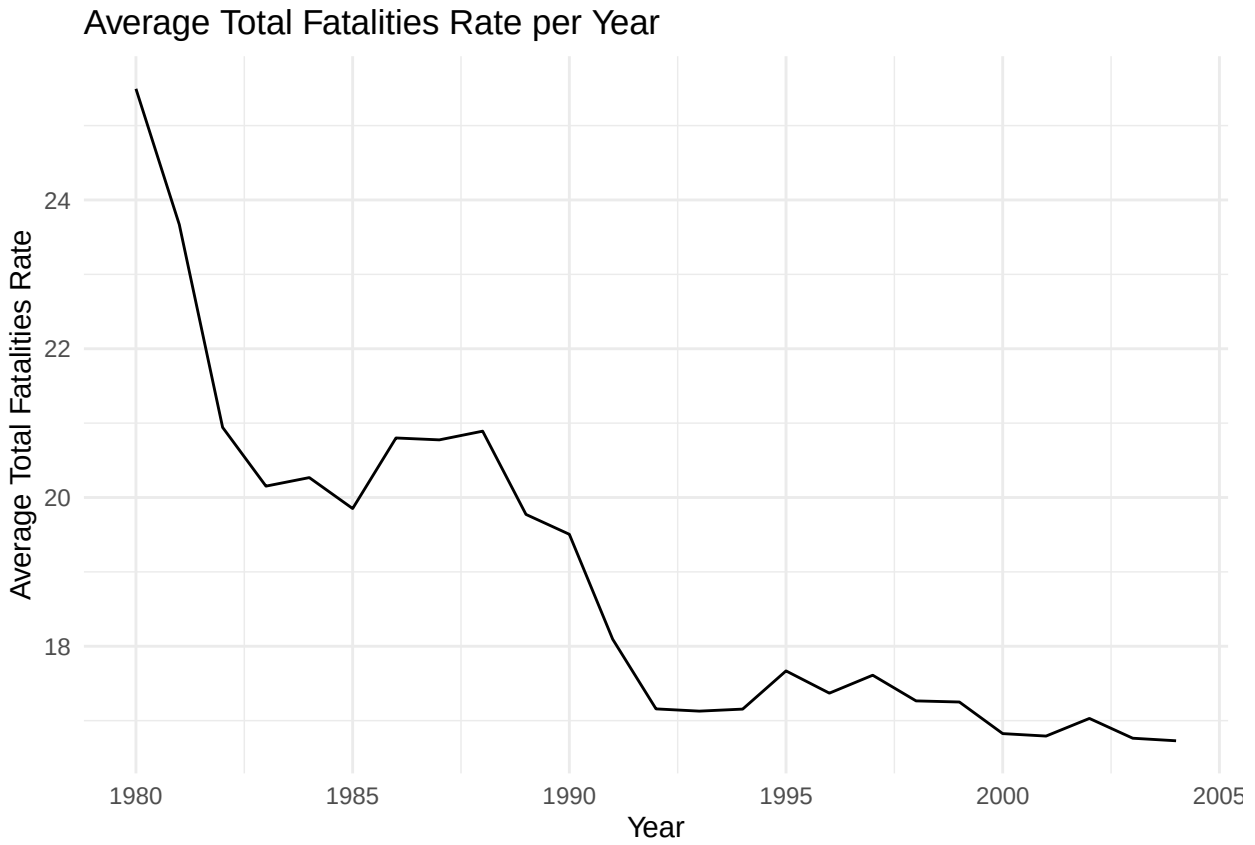
The traffic fatalities data originates from the Fatality Analysis Reporting System (FARS), managed by the National Highway Traffic Safety Administration (NHTSA). This system gathers data on every traffic crash across the 48 contiguous United States that results in the death of a vehicle occupant or a non-motorist. 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 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.

```
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

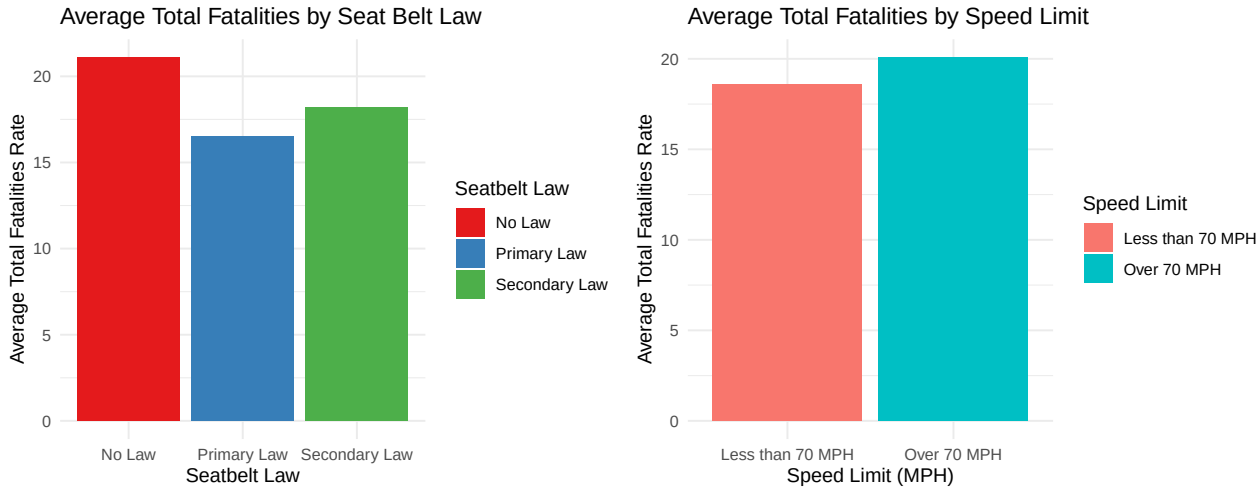
Minimum Drinking Age	18	19	20	21
Average Total Fatalities Rate	23.96	21.33	19.76	18.20

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", fill="Seatbelt Law") +
  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", fill="Speed Limit") +
  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.
```

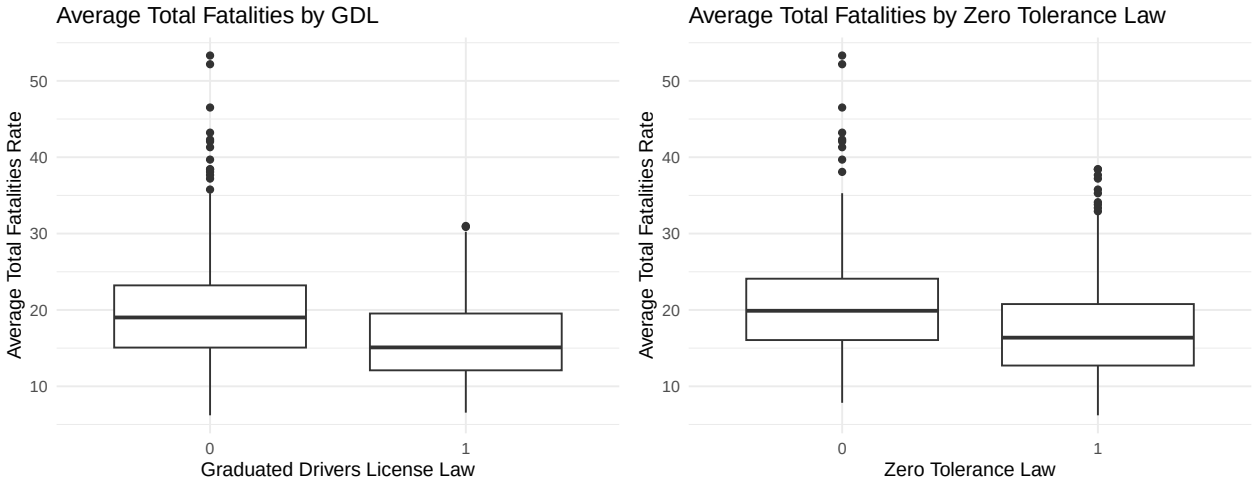
```
p1_bar | p2_bar
```



The seat belt law graph shows that regions with no seat belt law have the highest average fatality rate, while those with a primary enforcement law have the lowest average fatality rate. This indicates that stricter seat belt 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.

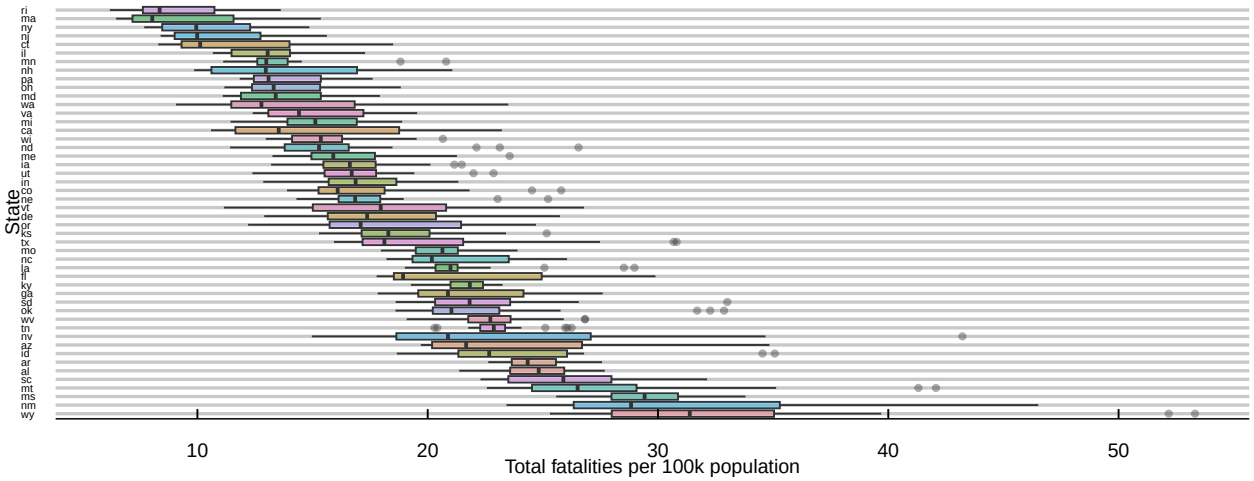
```
p1_box <- ggplot(data, aes(x=factor(graduated_drivers_license_law), y=total_fatalities_rate)) +
  geom_boxplot() +
  labs(title="Average Total Fatalities by GDL",
       x="Graduated Drivers License Law",
```

```
       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').

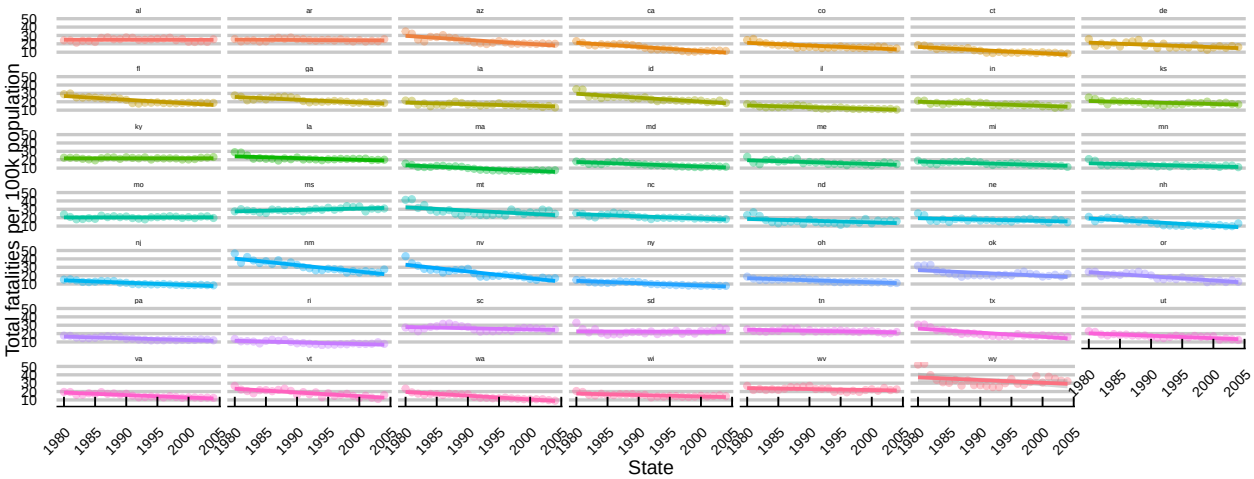
```
data |>
  ggplot(aes(reorder(state,desc(total_fatalities_rate)), total_fatalities_rate, fill=state))+
  geom_boxplot(alpha=0.4) +
  theme_economist_white(gray_bg=F)+
  theme(legend.position = "none", axis.text.y = element_text(size=6)) +
  scale_y_continuous(label=scales::number_format(accuracy = 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.

```
summary(data$perc14_24)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.70   13.90   14.90   15.33   16.60   20.30
```

```r
summary(data$unemployment_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.200   4.500   5.600   5.951   7.000  18.000
```

```r
summary(data$vehicmilespc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4372    7788    9013    9129   10327   18390
```

```r
data$year_of_observation <- as.factor(data$year_of_observation)
```

The scale of vehicle miles driven per capita is far larger than the percentage and rate variables.

(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
mod_pm <- lm(total_fatalities_rate ~ year_of_observation, data=data)

summary(mod_pm)
```

```
##
## Call:
## lm(formula = total_fatalities_rate ~ year_of_observation, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               25.4946     0.8671  29.401  < 2e-16 ***
## year_of_observation1981   -1.8244     1.2263  -1.488 0.137094
## year_of_observation1982   -4.5521     1.2263  -3.712 0.000215 ***
## year_of_observation1983   -5.3417     1.2263  -4.356 1.44e-05 ***
## year_of_observation1984   -5.2271     1.2263  -4.263 2.18e-05 ***
## year_of_observation1985   -5.6431     1.2263  -4.602 4.64e-06 ***
## year_of_observation1986   -4.6942     1.2263  -3.828 0.000136 ***
## year_of_observation1987   -4.7198     1.2263  -3.849 0.000125 ***
## year_of_observation1988   -4.6029     1.2263  -3.754 0.000183 ***
## year_of_observation1989   -5.7223     1.2263  -4.666 3.42e-06 ***
## year_of_observation1990   -5.9894     1.2263  -4.884 1.18e-06 ***
## year_of_observation1991   -7.3998     1.2263  -6.034 2.14e-09 ***
## year_of_observation1992   -8.3367     1.2263  -6.798 1.68e-11 ***
## year_of_observation1993   -8.3669     1.2263  -6.823 1.43e-11 ***
```

```
## year_of_observation1994  -8.3394     1.2263  -6.800 1.66e-11 ***
## year_of_observation1995  -7.8260     1.2263  -6.382 2.51e-10 ***
## year_of_observation1996  -8.1252     1.2263  -6.626 5.25e-11 ***
## year_of_observation1997  -7.8840     1.2263  -6.429 1.86e-10 ***
## year_of_observation1998  -8.2292     1.2263  -6.711 3.01e-11 ***
## year_of_observation1999  -8.2442     1.2263  -6.723 2.77e-11 ***
## year_of_observation2000  -8.6690     1.2263  -7.069 2.67e-12 ***
## year_of_observation2001  -8.7019     1.2263  -7.096 2.21e-12 ***
## year_of_observation2002  -8.4650     1.2263  -6.903 8.32e-12 ***
## year_of_observation2003  -8.7310     1.2263  -7.120 1.88e-12 ***
## year_of_observation2004  -8.7656     1.2263  -7.148 1.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.008 on 1175 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1098
## F-statistic: 7.164 on 24 and 1175 DF,  p-value: < 2.2e-16
```

Fitting a linear model is a good place to start given its simplicity and comprehension to illuminate first insights into the data. It may also signal if ignoring cross sectional units has detrimentally effected the model. Here the model explains that in general as the year increases the total fatality rate falls. This gives evidence that driving did in fact become safer overtime as the fatality rate was much lower in 2004 opposed to 1980.

This model however has considerable limitations due to its lack of independent observations, omitted variable bias and the accompanying ignorance to unobserved heterogeneity for these omitted variables. Omitted variables such as different policy measures and state are thus potentially problematic to not include and would be a leading cause to biased parameter estimations. Additionally, the presence of heteroskedasticity signals potential bias and unreliability in uncertainty estimates.

Taking the data structure for granted and applying a simple pooled regression while convenient can thus obscure the correct conclusions to be drawn from this panel data set.

(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?

```
# second option using binary year variables and add scaling
```

```r
#data$perc14_24 <- rescale(data$perc14_24)
#data$unemployment_rate <- rescale(data$unemployment_rate)
#data$vehicmilespc<- rescale(data$vehicmilespc)

cat_vars <- c("bac10", "bac08", "perse", "primary_seatbelt_law",
              "secondary_seatbelt_law", "sl70plus", "graduated_drivers_license_law")
data[cat_vars] <- lapply(data[cat_vars], factor)

num_vars <- c("perc14_24", "unemployment_rate", "vehicmilespc")
combined_vars <- c(cat_vars, num_vars)

fn_exp <- as.formula(paste('total_fatalities_rate ~ year_of_observation +',
                           paste(combined_vars, collapse='+')))
mod_exp <- lm(fn_exp, data=data)

summary(mod_exp)
```

```
##
## Call:
## lm(formula = fn_exp, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8962  -2.7265  -0.3033   2.3323  21.5064
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2.826e+00  2.478e+00  -1.141 0.254236
## year_of_observation1981       -2.184e+00  8.290e-01  -2.634 0.008539 **
## year_of_observation1982       -6.657e+00  8.547e-01  -7.789 1.49e-14 ***
## year_of_observation1983       -7.589e+00  8.671e-01  -8.752  < 2e-16 ***
## year_of_observation1984       -5.974e+00  8.730e-01  -6.843 1.25e-11 ***
## year_of_observation1985       -6.603e+00  8.915e-01  -7.407 2.47e-13 ***
## year_of_observation1986       -5.947e+00  9.290e-01  -6.401 2.23e-10 ***
## year_of_observation1987       -6.459e+00  9.656e-01  -6.689 3.48e-11 ***
## year_of_observation1988       -6.691e+00  1.013e+00  -6.607 5.97e-11 ***
## year_of_observation1989       -8.159e+00  1.052e+00  -7.757 1.89e-14 ***
## year_of_observation1990       -9.060e+00  1.076e+00  -8.421  < 2e-16 ***
## year_of_observation1991       -1.121e+01  1.099e+00 -10.194  < 2e-16 ***
## year_of_observation1992       -1.300e+01  1.121e+00 -11.591  < 2e-16 ***
## year_of_observation1993       -1.288e+01  1.134e+00 -11.358  < 2e-16 ***
## year_of_observation1994       -1.253e+01  1.154e+00 -10.855  < 2e-16 ***
## year_of_observation1995       -1.203e+01  1.183e+00 -10.176  < 2e-16 ***
## year_of_observation1996       -1.403e+01  1.224e+00 -11.459  < 2e-16 ***
## year_of_observation1997       -1.430e+01  1.242e+00 -11.517  < 2e-16 ***
## year_of_observation1998       -1.512e+01  1.262e+00 -11.978  < 2e-16 ***
## year_of_observation1999       -1.518e+01  1.276e+00 -11.900  < 2e-16 ***
## year_of_observation2000       -1.554e+01  1.296e+00 -11.996  < 2e-16 ***
## year_of_observation2001       -1.645e+01  1.316e+00 -12.500  < 2e-16 ***
## year_of_observation2002       -1.703e+01  1.331e+00 -12.798  < 2e-16 ***
## year_of_observation2003       -1.742e+01  1.336e+00 -13.033  < 2e-16 ***
## year_of_observation2004       -1.698e+01  1.369e+00 -12.399  < 2e-16 ***
## bac101                        -1.238e+00  3.616e-01  -3.423 0.000641 ***
## bac081                        -2.194e+00  4.891e-01  -4.487 7.94e-06 ***
```

```
## perse1                          -6.499e-01  2.943e-01  -2.208 0.027433 *
## primary_seatbelt_law1           -9.420e-02  4.910e-01  -0.192 0.847868
## secondary_seatbelt_law1          6.430e-02  4.299e-01   0.150 0.881124
## sl70plus1                        3.239e+00  4.352e-01   7.443 1.91e-13 ***
## graduated_drivers_license_law1 -3.476e-01  5.101e-01  -0.682 0.495682
## perc14_24                        1.401e-01  1.229e-01   1.140 0.254611
## unemployment_rate                7.675e-01  7.796e-02   9.844  < 2e-16 ***
## vehicmilespc                     2.927e-03  9.485e-05  30.860  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.052 on 1165 degrees of freedom
## Multiple R-squared:  0.6064, Adjusted R-squared:  0.595
## F-statistic:  52.8 on 34 and 1165 DF,  p-value: < 2.2e-16
```

Primary seat belt laws and Secondary seat belt laws are rounded to represent a factor 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 simlarly rounded to eliminate any data that may have not been in factor form. Speed limit faster than 70 mph was rounded as well to be a proper Boolean. Unemployment Rate, Vehicle miles driven per capita, Percent of the population between 14 and 24 years old are rescaled to the [0,1] range to improve interpretability and mitigate the effect of a large numerical variable from biasing results.

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 with the base case being without any legal restrictions on BAC level. The coefficient values stand at -1.2 and -2.2 respectfully and are both highly significant. Meaning that the presence of BAC laws lower fatalities by 1.2 to 2.2 per 100,000 people where stricter regulations are correlated with less fatalities. Per Se laws also seem to lower fatalities at a lower rate of 0.6 per 100,000 people, however the coefficient is not as significant, but does fall below the 0.05 p-value. Surprisingly, the presence of a primary seat belt law yields a marginally negative and non-significant parameter estimate. This finding raises questions regarding the validity of employing a pooled regression model for the dataset, given the widely acknowledged role of seat belts in saving lives during emergencies.

(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?

```
mod_fe <- plm(fn_exp, index="state", data=data, model="within")
summary(mod_fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = fn_exp, 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|)
## year_of_observation1981        -1.5124e+00  4.1379e-01  -3.6549 0.0002692 ***
## year_of_observation1982        -3.0540e+00  4.4318e-01  -6.8912 9.221e-12 ***
## year_of_observation1983        -3.6638e+00  4.5516e-01  -8.0495 2.111e-15 ***
## year_of_observation1984        -4.3998e+00  4.5966e-01  -9.5719 < 2.2e-16 ***
## year_of_observation1985        -4.8603e+00  4.8010e-01 -10.1237 < 2.2e-16 ***
## year_of_observation1986        -3.7692e+00  5.1357e-01  -7.3392 4.123e-13 ***
## year_of_observation1987        -4.4123e+00  5.5162e-01  -7.9989 3.118e-15 ***
## year_of_observation1988        -4.8877e+00  5.9837e-01  -8.1684 8.379e-16 ***
## year_of_observation1989        -6.2395e+00  6.3732e-01  -9.7901 < 2.2e-16 ***
## year_of_observation1990        -6.3564e+00  6.6196e-01  -9.6024 < 2.2e-16 ***
## year_of_observation1991        -7.0442e+00  6.7895e-01 -10.3752 < 2.2e-16 ***
## year_of_observation1992        -7.8905e+00  7.0039e-01 -11.2659 < 2.2e-16 ***
## year_of_observation1993        -8.2366e+00  7.1290e-01 -11.5536 < 2.2e-16 ***
## year_of_observation1994        -8.6823e+00  7.3004e-01 -11.8930 < 2.2e-16 ***
## year_of_observation1995        -8.3889e+00  7.5324e-01 -11.1370 < 2.2e-16 ***
## year_of_observation1996        -8.7648e+00  7.9400e-01 -11.0388 < 2.2e-16 ***
## year_of_observation1997        -8.9164e+00  8.1140e-01 -10.9889 < 2.2e-16 ***
## year_of_observation1998        -9.5333e+00  8.2867e-01 -11.5044 < 2.2e-16 ***
## year_of_observation1999        -9.6940e+00  8.3614e-01 -11.5938 < 2.2e-16 ***
## year_of_observation2000        -1.0223e+01  8.4713e-01 -12.0683 < 2.2e-16 ***
## year_of_observation2001        -9.9608e+00  8.5745e-01 -11.6168 < 2.2e-16 ***
## year_of_observation2002        -9.2546e+00  8.6613e-01 -10.6850 < 2.2e-16 ***
## year_of_observation2003        -9.3270e+00  8.6980e-01 -10.7232 < 2.2e-16 ***
## year_of_observation2004        -9.6676e+00  8.9310e-01 -10.8248 < 2.2e-16 ***
## bac101                         -8.6977e-01  2.2522e-01  -3.8619 0.0001190 ***
## bac081                         -1.1805e+00  3.2987e-01  -3.5786 0.0003603 ***
## perse1                         -1.0587e+00  2.2415e-01  -4.7230 2.619e-06 ***
## primary_seatbelt_law1          -1.2506e+00  3.4313e-01  -3.6447 0.0002800 ***
## secondary_seatbelt_law1        -3.5659e-01  2.5230e-01  -1.4133 0.1578360
## sl70plus1                      -3.2440e-02  2.6034e-01  -0.1246 0.9008577
## graduated_drivers_license_law1 -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
```

The blood alcohol variables (`bac08` and `bac10`) both remain statistically significant in the fixed effects model at the state level. The coefficient for `bac08` dropped slightly from -2.19 to -1.18 and the coefficient for `bac10` also dropped from -1.24 to -0.87. The coefficients for the blood alcohol variables are negative and significant

in both models. The interpretation is that when the bac08 law is in effect the traffic fatality rate is reduced by 1.18, whereas when the bac10 law is in effect the traffic fatality rate is only reduced by 0.87. This practically makes sense as we would expect a lower blood alcohol level to reduce the fatality rate. By controlling for any inherent fixed characteristics that vary across states, we that the magnitude of the effect of bac laws on fatality decreases, suggesting that there is a decent amount of variation in the characteristics from state to state.

The statistical significance associated with `perse` increases in the fixed effects model and the coefficient increases in magnitude from -0.65 to -1.06, indicating that when the perse laws are in effect, controlling for fixed effects across states, the laws appear to be more effective at reducing the fatality rate.

The `primary_seatbelt_law` variable becomes significant in the fixed effects model. The coefficient also changes from -0.09 to -1.25 suggesting that when controlling for state level fixed effects, primary seat belt laws do reduce the traffic fatality rate.

The fixed effects model produces a more reliable set of estimates because it relaxes the assumption of iid across data points. In the linear extended model, the assumption is that the data are iid, however due to the panel structure we know that there is dependency among observations. The fixed effects model accounts for this dependency by identifying that the data is linked by state. This model assumes that the data are iid within panels i.e., at the state level. This assumption is more reasonable than ignoring the panel structure of the data. It is reasonable to believe that there are fixed characteristics that vary across states i.e., no two states are identical.

(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?

In this case, a random effects model assumes the state specific effects are uncorrelated with all the predictors in the model, which means there is no omitted variable bias from omitting fixed effects. We think this is a very strong assumption that are not met in our data. If we were to inappropriately estimate a random effects model, the coefficient estimates and standard error estimates will be biased, and will not be consistent. Only fixed effect model is the solely consistent model for this case.

```
mod_re <- plm(fn_exp, index="state", data=data, model="random")
#summary(mod_re)
```

```
phtest(mod_fe, mod_re)
```

```
##
##  Hausman Test
##
## data:  fn_exp
## chisq = 164.12, df = 34, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

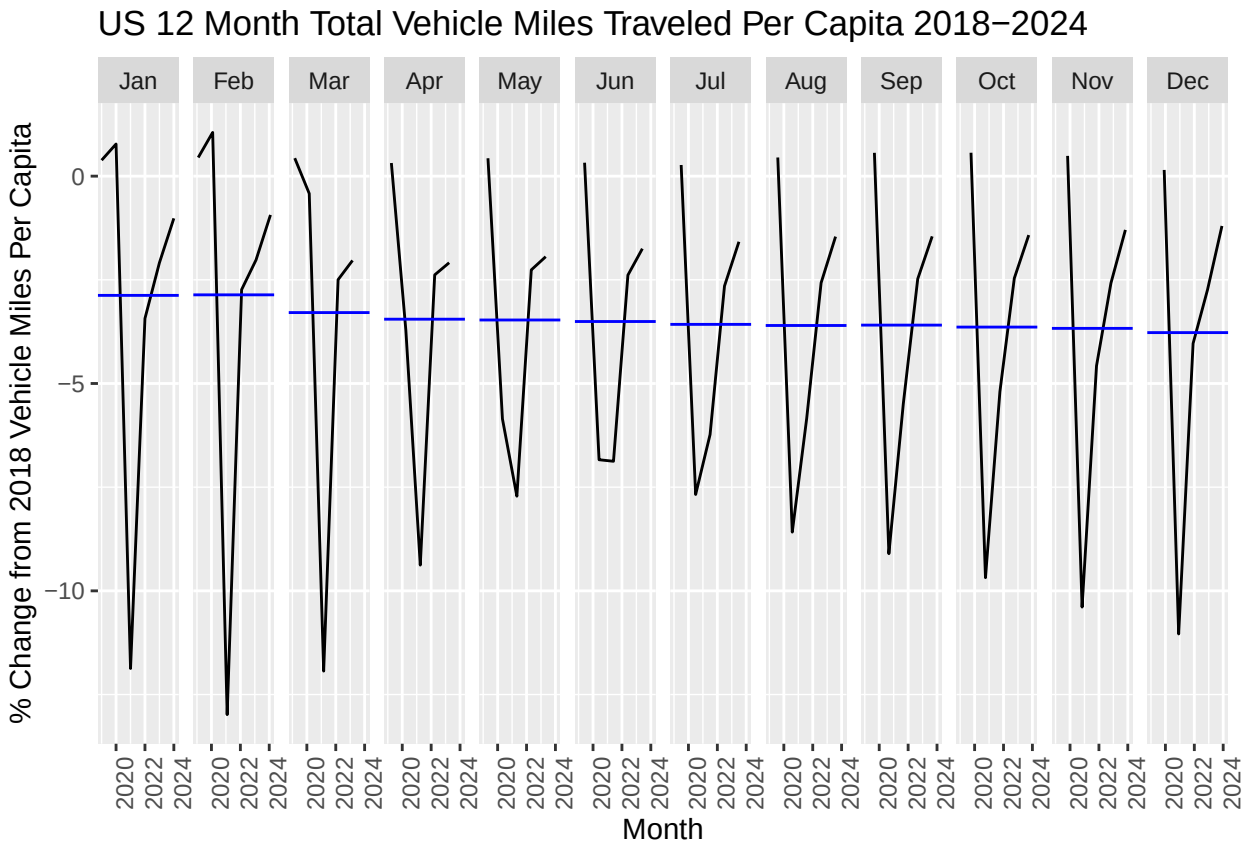
Very small p-value, significantly less than 5%, suggesting we should reject the null hypothesis that random effects model is appropriate. This confirms our original evaluation against a random effects model.

(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:

```
fredr_set_key("0ba6b0a40845d6abcf6b761a190609c7")
driving.df <- fredr(
  series_id = "M12MTVUSM227NFWA",
  observation_start = as.Date("2018-01-01")
) |>
  mutate(year = year(date)) |>
  mutate(month = month(date)) |>
  mutate(time_index = yearmonth(date)) |>
  mutate(value = value*1000000) |>
  select(c(time_index, year, month, value))

fredr_set_key("0ba6b0a40845d6abcf6b761a190609c7")
population.df <- fredr(
  series_id = "POPTHM",
  observation_start = as.Date("2018-01-01")
) |>
  mutate(year = year(date)) |>
  mutate(month = month(date)) |>
  mutate(time_index = yearmonth(date)) |>
  rename(pop_index = time_index) |>
  rename(pop = value) |>
  mutate(pop = 1000*pop) |>
  select(c(pop_index, pop))

driving.df <- driving.df |>
  left_join(population.df, by = c('time_index' = 'pop_index')) |>
  mutate(vehicmilespc = value/pop) |>
  select(c(time_index, year, month, vehicmilespc))

baseline_2018_driving <- driving.df |>
  filter(year == 2018) |>
  select(month, vehicmilespc) |>
  rename(baseline_month = month, baseline_vehicmilespc = vehicmilespc)

post_2018<-driving.df |>
  filter(year != 2018)

driving.change.ts <- post_2018 |>
  left_join(baseline_2018_driving, by = c('month' = 'baseline_month')) |>
  mutate(change = (vehicmilespc-baseline_vehicmilespc)) |>
  mutate(pct_change = (vehicmilespc-baseline_vehicmilespc)/baseline_vehicmilespc*100) |>
  select(time_index, vehicmilespc, baseline_vehicmilespc, change, pct_change) |>
  as_tsibble(index = time_index)
```

```
driving.change.ts |>
  gg_subseries(pct_change) +
  labs(y = "% Change from 2018 Vehicle Miles Per Capita", x = "Month",
       title = "US 12 Month Total Vehicle Miles Traveled Per Capita 2018-2024")
```




```
driving.change.ts |> filter(pct_change == max(pct_change))
```

```
## # A tsibble: 1 x 5 [1M]
##   time_index vehicmilespc baseline_vehicmilespc change pct_change
##        <mth>        <dbl>                 <dbl>  <dbl>      <dbl>
## 1   2020 Feb        9908.                 9804.   103.       1.05
```

```
driving.change.ts |> filter(pct_change == min(pct_change))
```

```
## # A tsibble: 1 x 5 [1M]
##   time_index vehicmilespc baseline_vehicmilespc change pct_change
##        <mth>        <dbl>                 <dbl>  <dbl>      <dbl>
## 1   2021 Feb        8531.                 9804. -1273.      -13.0
```

```
sd_increase <- sd(data$vehicmilespc)*.00092612

tail(driving.change.ts)
```

```
## # A tsibble: 6 x 5 [1M]
##   time_index vehicmilespc baseline_vehicmilespc change pct_change
##        <mth>        <dbl>                 <dbl>  <dbl>      <dbl>
## 1   2023 Sep        9678.                 9821. -143.      -1.45
## 2   2023 Oct        9684.                 9823. -140.      -1.42
## 3   2023 Nov        9699.                 9826. -127.      -1.30
## 4   2023 Dec        9711.                 9829. -118.      -1.20
## 5   2024 Jan        9701.                 9801.  -99.7     -1.02
## 6   2024 Feb        9712.                 9804.  -91.7     -0.935
```

- 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?

During the pandemic, Feb 2021 was the month that had experienced the largest decrease in driving when compared to Feb 2018 with -12.98% decrease.

- What month demonstrated the largest increase in driving? How much, in percentage terms, higher was this driving?

During the pandemic, Feb 2020 was the month that had experienced largest increase in driving when compared to Feb 2018 with 1.05% 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.

Using February 2024 as a reference point, when the US vehicle miles per capita stood at 9712.464, a 1.05% increase suggests the FE model would anticipate an additional 0.089 fatalities per 100,000 population. While any rise is regrettable, the observed effect appears to be minimal in practical terms.

- 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.

Using February 2024 as the benchmark, when US vehicle miles per capita stood at 9712.464, a 12.98% decrease implies the FE model would forecast 1.103 fewer fatalities per 100,000 population. This promising outcome indicates a reduced reliance on single occupancy vehicles, leading to fewer driving miles, could precipitate fewer fatalities in a meaningful way at this large a difference.

(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?

```
# Breusch-Godfrey serial correlation
pbgtest(mod_fe)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  fn_exp
## chisq = 306.43, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

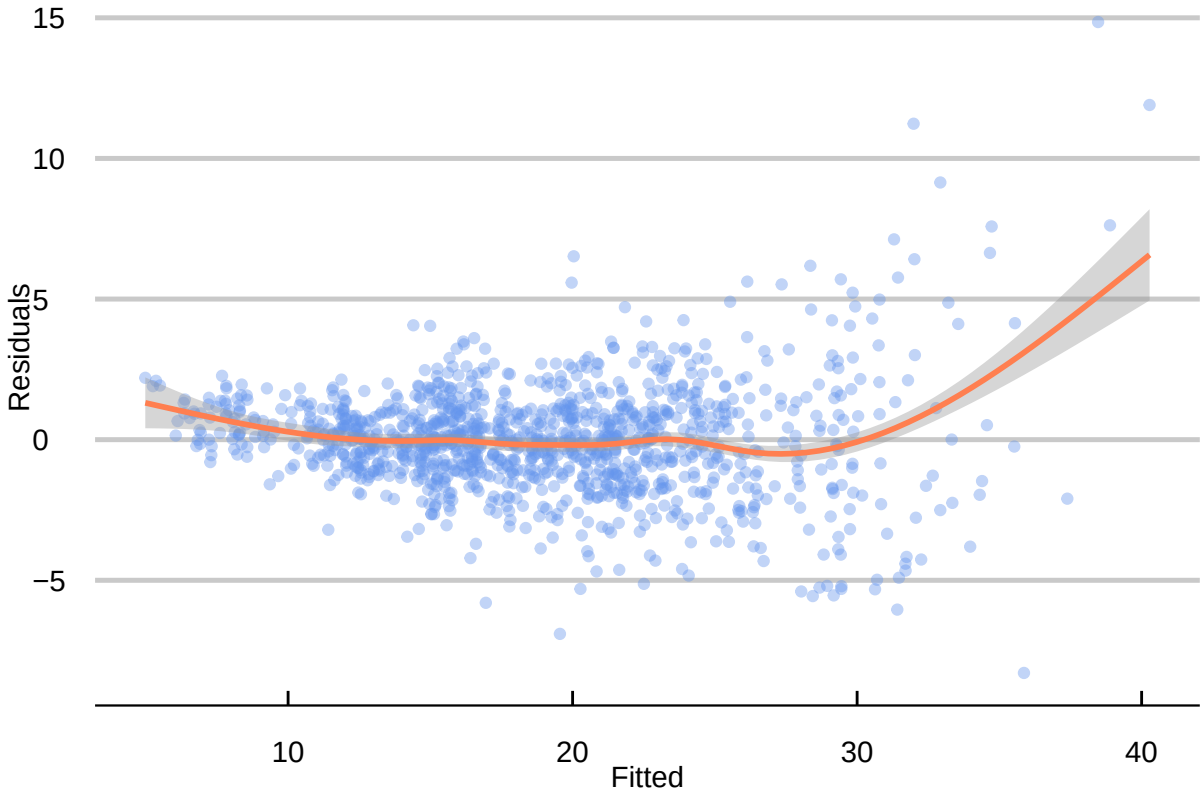
```
# breush pagan test for heteroskedasticty
pcdtest(mod_fe)
```

```
##
##  Pesaran CD test for cross-sectional dependence in panels
##
## data:  total_fatalities_rate ~ year_of_observation + bac10 + bac08 +     perse + primary_seatbelt_lau
## z = -0.4412, p-value = 0.6591
## alternative hypothesis: cross-sectional dependence
```

```
data.frame("resid"=resid(mod_fe),"fitted"=predict(mod_fe)) %>%
ggplot(aes(fitted,resid)) +
geom_point(alpha=0.4,color="cornflowerblue") +
```

```
geom_smooth(method="gam",color="coral") +
theme_economist_white(gray_bg=F) +
theme(legend.position="none") +
xlab("Fitted") +
ylab("Residuals")
```

```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```



If serial correlation or heteroskedasticity were present in the idiosyncratic errors of the model, the significance levels of the regression coefficients would be overestimated which results in a biased model. The Breusch-Godfrey test provides evidence to reject the null hypothesis of no serial correlation due to the small p-value, therefore serial correlation is present. The Breush-Pagan test provides evidence to not reject the null hypothesis of cross-sectional dependence. Therefore, heteroskedasticity is present in our data. This result is also supported by the residual plot which shows dispersion of the residuals as the fitted value increases, implying un-equal variance.