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.

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

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

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`;

```r
data_cleaned <- data %>%
  mutate(speed_limit = case_when(sl55 > 0.5 ~ '55', #Higher than 0.5 as 55
                                 sl65 >= 0.5 ~ '60', #Higher than 0.5 as 60 overriding any prior value
                                 sl70 >= 0.5 ~ '65', #Higher than 0.5 as 65 overriding any prior value
                                 sl75 >= 0.5 ~ '70', #Higher than 0.5 as 70 overriding any prior value
                                 slnone > 0.5 ~ NA)) #Higher than 0.5 as NA

data_cleaned <- data_cleaned %>%
  select(-c("sl55", "sl65", "sl70", "sl75", "slnone")) #Drop columns

data_cleaned$speed_limit <- as.numeric(data_cleaned$speed_limit) #sets to numeric
```

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

```r
data_cleaned <- data_cleaned %>%
       rename("year_of_observation" = "year") #sets year as year_of_observation

data_cleaned <- data_cleaned[-c(which(colnames(data_cleaned)=="d80"):which(colnames(data_cleaned)=="d04

data_cleaned$year_of_observation <- as.factor(data_cleaned$year_of_observation) #sets as a factor
```

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

```r
data_cleaned <- data_cleaned %>%
  mutate(blood_alcohol_level = case_when(bac08 >= 0.5 ~ '8',
                                         bac10 >= 0.5 ~ '10',
                                         TRUE ~ 'None')) #Remainder are none

data_cleaned <- data_cleaned %>%
  select(-c("bac08", "bac10")) #drops columns

data_cleaned$blood_alcohol_level <- as.factor(data_cleaned$blood_alcohol_level) #sets as a factor
```

- 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.)

```
col_name_lookup <- c(
  "min_drinking_age" = "minage",
  "zero_tol_law" = "zerotol",
  "grad_driver_law" = "gdl",
  "per_se_law" = "perse",
  "total_fatalities" = "totfat",
  "night_fatalities" = "nghtfat",
  "weekend_fatalities" = "wkndfat",
  "total_fatalities_per_100mil_miles" = "totfatpvm",
  "night_fatalities_per_100mil_miles" = "nghtfatpvm",
  "weekend_fatalities_per_100mil_miles" = "wkndfatpvm",
  "state_population" = "statepop",
  "total_fatalities_rate" = "totfatrte",
  "night_fatalities_rate" = "nghtfatrte",
  "weekend_fatalities_rate" = "wkndfatrte",
  "vehicle_miles" = "vehicmiles",
  "unemployment_rate" = "unem",
  "percent_age_14_24" = "perc14_24",
  "speed_limit_70_or_higher" = "sl70plus",
  "primary_seatbelt" = "sbprim",
  "secondary_seatbelt" = "sbsecon",
  "vehicle_miles_per_capita"="vehicmilespc"
  )

data_cleaned <- data_cleaned %>%
  rename(any_of(col_name_lookup))
```

```
#Adjusting factor variables to binary
data_cleaned <- data_cleaned %>%
  mutate(zero_tol_law = case_when(zero_tol_law >= 0.5 ~ '1',
                                  zero_tol_law < 0.5 ~ '0'))

data_cleaned <- data_cleaned %>%
  mutate(grad_driver_law = case_when(grad_driver_law >= 0.5 ~ '1',
                                     grad_driver_law < 0.5 ~ '0'))

data_cleaned <- data_cleaned %>%
  mutate(per_se_law = case_when(per_se_law >= 0.5 ~ '1',
                                per_se_law < 0.5 ~ '0'))

data_cleaned <- data_cleaned %>%
  mutate(speed_limit_70_or_higher = case_when(speed_limit_70_or_higher >= 0.5 ~ '1',
                                speed_limit_70_or_higher < 0.5 ~ '0'))
```

```
pdriving <- pdata.frame(
  data_cleaned,
  index=c("state", "year_of_observation")
  )
```

```r
pdim(pdriving)
```

```
## Balanced Panel: n = 48, T = 25, N = 1200
```

```r
# Renaming States
replacement_state <- c("AL", "AZ", "AR", "CA", "CO", "CT", "DE", "FL",
                       "GA", "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")

levels(pdriving$state) <- replacement_state

replacement_region <- c(
  "south", "southwest", "south", "west", "west", "northeast", "northeast", "south", "south",
  "west", "midwest", "midwest", "midwest", "midwest", "south", "south", "northeast", "northeast",
  "northeast", "midwest", "midwest", "south", "midwest", "west", "midwest", "southwest", "northeast",
  "northeast", "southwest", "northeast", "south", "midwest", "midwest", "southwest", "west", "northeast
  "northeast", "south", "midwest", "south", "southwest", "west", "northeast", "south", "west", "south",
  "midwest", "west")

pdriving$region <- pdriving$state
levels(pdriving$region) <- replacement_region
```

```r
#renameing state column rows
data_cleaned <- data_cleaned %>%
  mutate(state = case_when(
    state == '1' ~ "AL", state == '3' ~ "AZ", state == '4' ~ "AR", state == '5' ~ "CA",
    state == '6' ~ "CO", state == '7' ~ "CT", state == '8' ~ "DE", state == '10' ~ "FL",
    state == '11' ~ "GA", state == '13' ~ "ID", state == '14' ~ "IL", state == '15' ~ "IN",
    state == '16' ~ "IA", state == '17' ~ "KS", state == '18' ~ "KY", state == '19' ~ "LA",
    state == '20' ~ "ME", state == '21' ~ "MD", state == '22' ~ "MA", state == '23' ~ "MI",
    state == '24' ~ "MN", state == '25' ~ "MS", state == '26' ~ "MO", state == '27' ~ "MT",
    state == '28' ~ "NE", state == '29' ~ "NV", state == '30' ~ "NH", state == '31' ~ "NJ",
    state == '32' ~ "NM", state == '33' ~ "NY", state == '34' ~ "NC", state == '35' ~ "ND",
    state == '36' ~ "OH", state == '37' ~ "OK", state == '38' ~ "OR", state == '39' ~ "PA",
    state == '40' ~ "RI", state == '41' ~ "SC", state == '42' ~ "SD", state == '43' ~ "TN",
    state == '44' ~ "TX", state == '45' ~ "UT", state == '46' ~ "VT", state == '47' ~ "VA",
    state == '48' ~ "WA", state == '49' ~ "WV", state == '50' ~ "WI", state == '51' ~ "WY"))
```

```r
#Convert variables to numeric or factor
data_cleaned$year_of_observation <- as.factor(data_cleaned$year_of_observation)
data_cleaned$state <- as.factor(data_cleaned$state)
data_cleaned$seatbelt <- as.factor(data_cleaned$seatbelt)
data_cleaned$min_drinking_age <- as.numeric(data_cleaned$min_drinking_age)
data_cleaned$zero_tol_law <- as.factor(data_cleaned$zero_tol_law)
data_cleaned$grad_driver_law <- as.factor(data_cleaned$grad_driver_law)
data_cleaned$per_se_law <- as.factor(data_cleaned$per_se_law)
data_cleaned$total_fatalities <- as.numeric(data_cleaned$total_fatalities)
data_cleaned$night_fatalities <- as.numeric(data_cleaned$night_fatalities)
data_cleaned$weekend_fatalities <- as.numeric(data_cleaned$weekend_fatalities)
```

```r
data_cleaned$total_fatalities_per_100mil_miles <- as.numeric(data_cleaned$total_fatalities_per_100mil_mi
data_cleaned$night_fatalities_per_100mil_miles <- as.numeric(data_cleaned$night_fatalities_per_100mil_mi
data_cleaned$weekend_fatalities_per_100mil_miles <- as.numeric(data_cleaned$weekend_fatalities_per_100mi
data_cleaned$state_population <- as.numeric(data_cleaned$state_population)
data_cleaned$total_fatalities_rate <- as.numeric(data_cleaned$total_fatalities_rate)
data_cleaned$night_fatalities_rate <- as.numeric(data_cleaned$night_fatalities_rate)
data_cleaned$weekend_fatalities_rate <- as.numeric(data_cleaned$weekend_fatalities_rate)
data_cleaned$vehicle_miles <- as.numeric(data_cleaned$vehicle_miles)
data_cleaned$unemployment_rate <- as.numeric(data_cleaned$unemployment_rate)
data_cleaned$percent_age_14_24 <- as.numeric(data_cleaned$percent_age_14_24)
data_cleaned$speed_limit_70_or_higher <- as.factor(data_cleaned$speed_limit_70_or_higher)
data_cleaned$primary_seatbelt <- as.factor(data_cleaned$primary_seatbelt)
data_cleaned$secondary_seatbelt <- as.factor(data_cleaned$secondary_seatbelt)
data_cleaned$vehicle_miles_per_capita <- as.numeric(data_cleaned$vehicle_miles_per_capita)
data_cleaned$speed_limit <- as.numeric(data_cleaned$speed_limit)
data_cleaned$blood_alcohol_level <- as.factor(data_cleaned$blood_alcohol_level)
```

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?

   Provide a description of the basic structure of the dataset

   After cleaning the data, the dataset has 27 columns and 1,200 rows. With each year from 1980 to 2004 having 48 rows for each of the lower 48 states (i.e., excluding Alaska and Hawaii). For each year and for each state information is given about the total number of fatalities per 100,000 people under several different criteria (e.g., number that occured at night or on the weekend), as well as certain demographic data such as unemployment rate, and various laws for each state in each year about blood alcohol level and whether or not a seatbelt is required.

   What is the data?

   The data is collecting the total fatalities per 100,000 in each state, as well as several other potentially significant features.

   How, where, and when is it collected?

   The data is collected for each of the continental 48 states for each year from 1980 to 2004. The dependent variable of interest is the total fatalities per 100,000 population in each state. It's unclear how population is determined, if it is for the population as a whole, population of drivers only, another metric.

   Is the data generated through a survey or some other method?

   In the article "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws" (Donald G. Freeman, 2007), the data was collected as follows:

"Fatality data are compiled from the Fatality Analysis Reporting System (FARS) administered by NHTSA. FARS compiles data on all traffic crashes that result in the death of a vehicle occupant or a nonmotorist. The data are gathered by state employees using a standard format for comparability across jurisdictions. Each record contains information on date, time, day of week, road conditions, age of victim, age of driver(s), number of vehicles, vehicle speed, and many other crash attributes."

Is the data that is presented a sample from the population, or is it a census that represents the entire population?

The data was collected by the Fatality Analysis Reporting System (FARS) which is administered by the U.S. Department of Transportation, National Highway Traffic Safety Administrtaion. We therefore believe that the figures represent the entire population of the state, to the extent that the fatalities were entered.

How is the dependent variable of interest defined

"Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws" (Donald G. Freeman, 2007) defines the dependent variable (total_fatalities_rate) as:

"The dependent variable in the empirical analyses to follow is the rate of traffic fatalities per 100,000 population at the state level over the years 1980-2004 for the 48 contiguous staes. The FARS is used to generate fatality rates for total, weekend night, and multiple daytime crashes. Ideally, alcohol related crashes would be used for a test of alcohol laws but only since 1982 has a consistent methodology been established for counting alcohol-related traffic deaths, and even now there are wide variations across states in teh proportion of drivers tested, alive or dead."

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?

     The dependent variable of interest is defined above, but it is the total number of fatalities per 100,000 of the population.

- What is the average of `total_fatalities_rate` in each of the years in the time period covered in this dataset?

```
summary_by_year <- pdriving %>%
  select(year_of_observation, total_fatalities_rate) %>%
  group_by(year_of_observation) %>%
  summarize(avg = mean(total_fatalities_rate))

summary_by_year %>%
  kbl(caption = "Average of Total Fatalities Rate by Year",
      col.names = c("Year", "Average Fatalities Rate"),
      longtable = TRUE) %>%
  kable_styling()
```

Table 1: Average of Total Fatalities Rate by Year

Year	Average Fatalities Rate
1980	25.49458
1981	23.67021
1982	20.94250
1983	20.15292
1984	20.26750
1985	19.85146
1986	20.80042
1987	20.77479
1988	20.89167
1989	19.77229
1990	19.50521
1991	18.09479
1992	17.15792
1993	17.12771
1994	17.15521
1995	17.66854
1996	17.36938
1997	17.61062
1998	17.26542
1999	17.25042
2000	16.82562
2001	16.79271
2002	17.02958
2003	16.76354
2004	16.72896

The graph above shows the average falities per 100,000 from 1980 to 2004. From the graph we can see that the fatalities decline from 1980 to 1985 before having a few years of increase from 1985 to 1988. The rates than decline from 1988 to 1995 where they appear to stabalize.

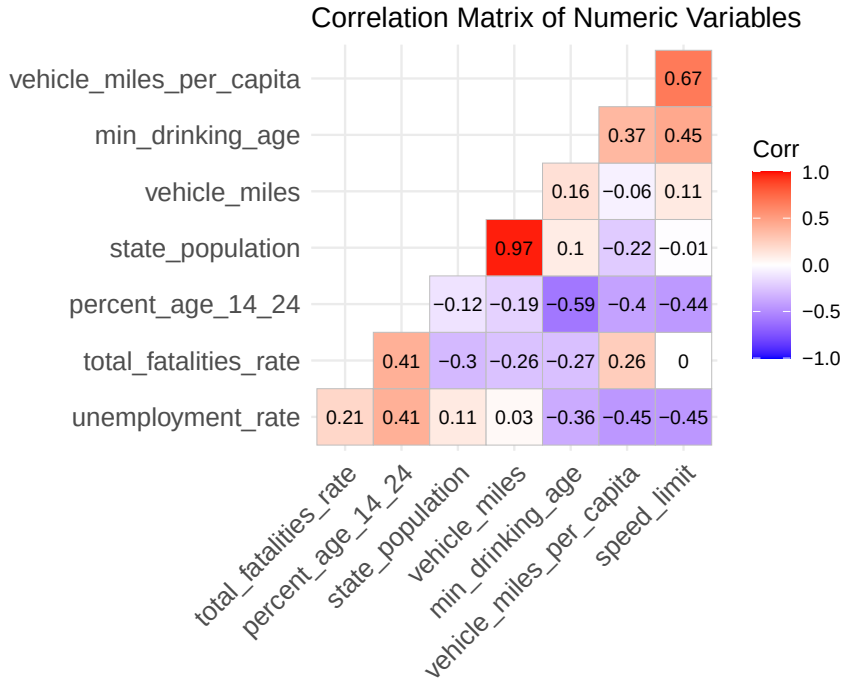
- Full Exploratory Data Analysis

2.1 Numeric Value Analysis

To simplify the analysis, we've removed the weekend and night variables from the EDA and focused only on the total_fatalieis_rate rather than total fatalities or fatalities per 100mil miles. The correlation matrix is presented below:

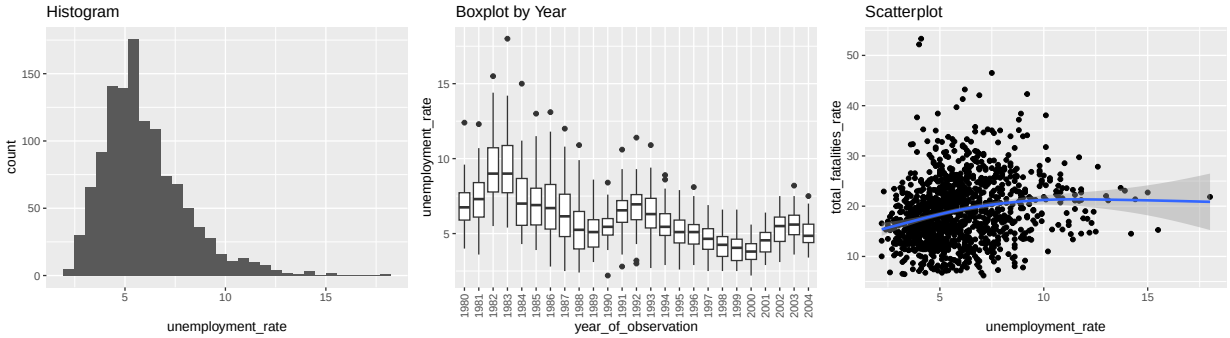
```
data_analysis <- data_cleaned %>%
  select(-c("total_fatalities", "night_fatalities", "weekend_fatalities",
            "total_fatalities_per_100mil_miles", "night_fatalities_per_100mil_miles", "weekend_fatalitie
            "night_fatalities_rate", "weekend_fatalities_rate"))

correlation_matrix <- round(cor(data_analysis[sapply(data_analysis, is.numeric)], use = "complete.obs")
ggcorrplot(correlation_matrix, hc.order = TRUE, type="lower", lab=TRUE, lab_size = 3) +
  labs(title = "Correlation Matrix of Numeric Variables")
```



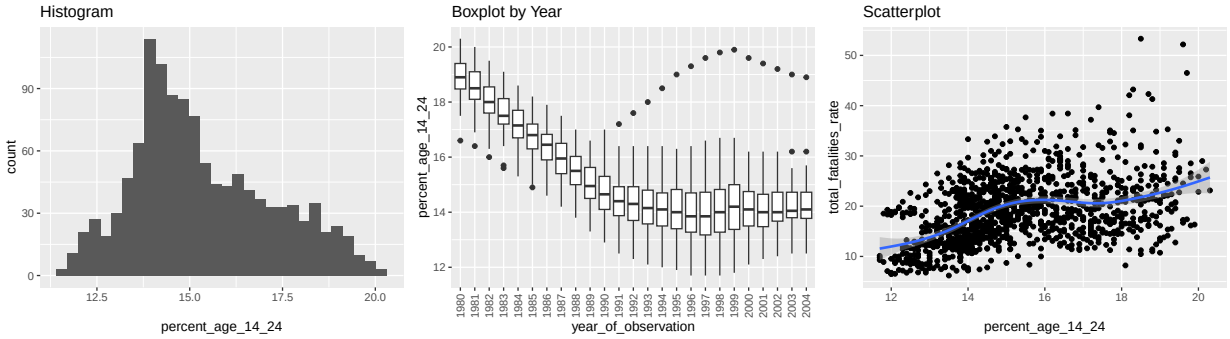
With the total fatalities rate, we can see that there is a positive correlation with unemployment_rate, percent_age_14_24 and vehicle_miles_per_capita, while all of the other variables have a negative correlation. Each of the variables are analyzed in more detail below.

2.1.1 Unemployment Rate



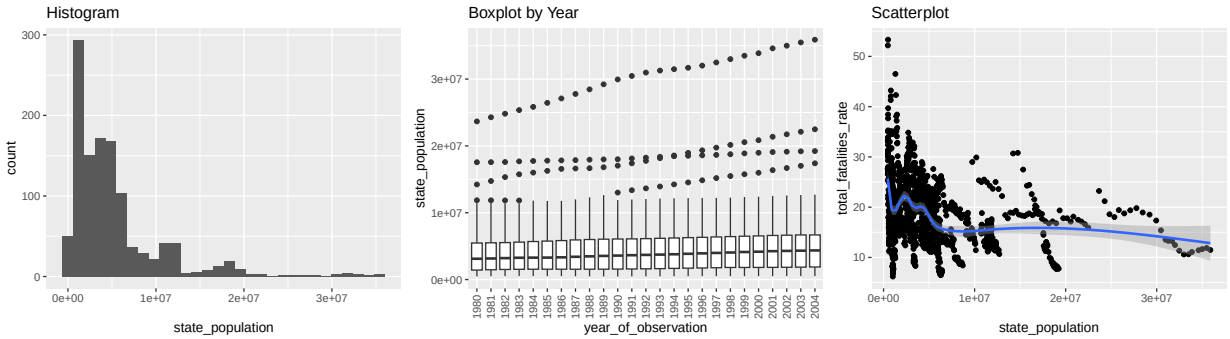
From the histogram, the unemployment rate appears to be slightly skewed, and would benefit from a log transfrmation. The boxplot shows that there are some distinct differences in unemployment rate by year, and the scatter plot illustrates the upward trend, with fewer observations at higher unemployment rates.

2.1.2 percent population aged 14 through 24



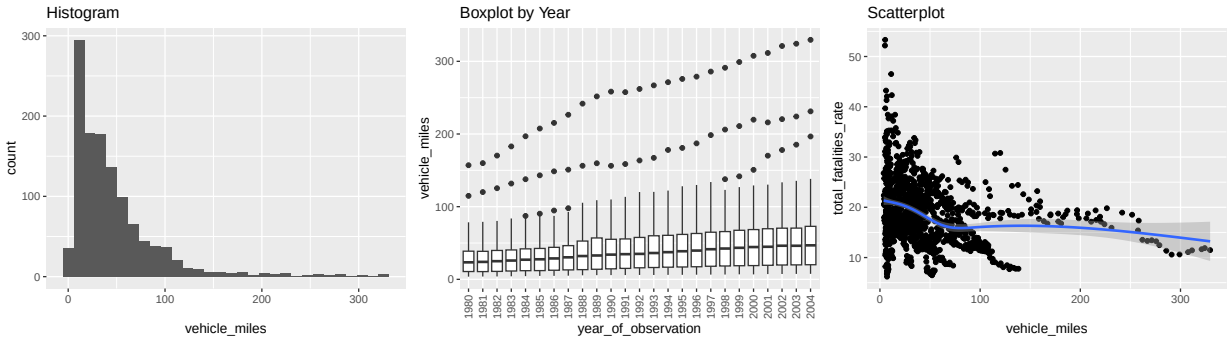
From the histogram, the variable seems to be relatively normally distributed and no transformation is suggested. There are distinct differences by year, with an evident downward trend. Related to total fatalities, there is a higher fatality rate the higher percentage of age 14 to 24 year olds.

2.1.3 State population



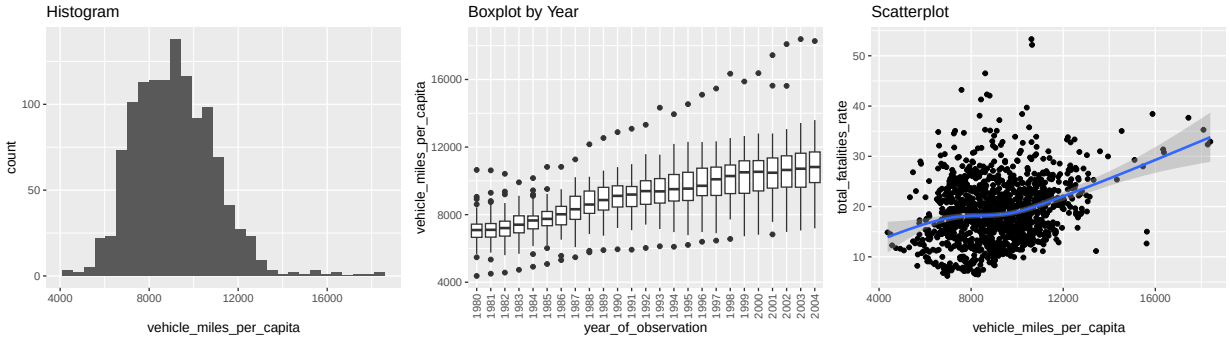
The histogram is rather skewed and would benefit from a log transformation. From the boxplot, the the median amount is rather flat, but the outliers are increasing. The scatterplot shows higher fatalities for lower populations.

2.1.4 Vehicle miles



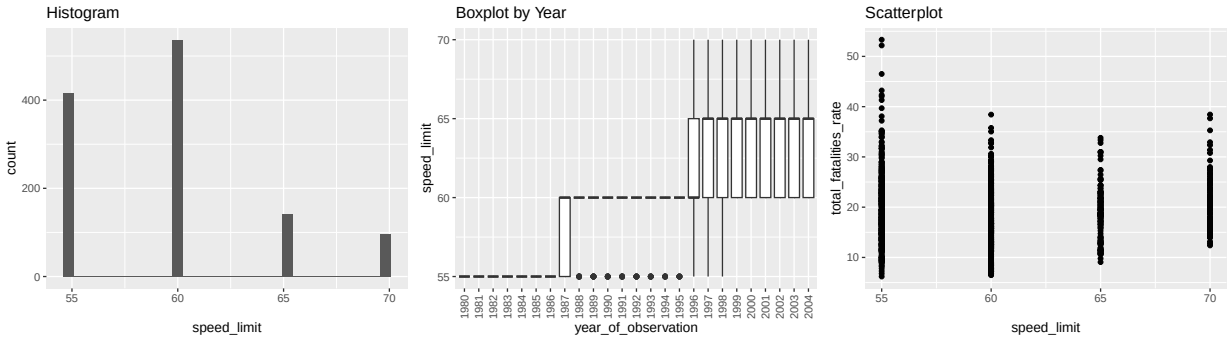
The histogram is rather skewed and would benefit from a log transformation. From the boxplot, the the median amount is rather flat, but the outliers are increasing. The scatterplot shows higher fatalities for lower vehicle populations.

2.1.5 Vehicle miles per capita



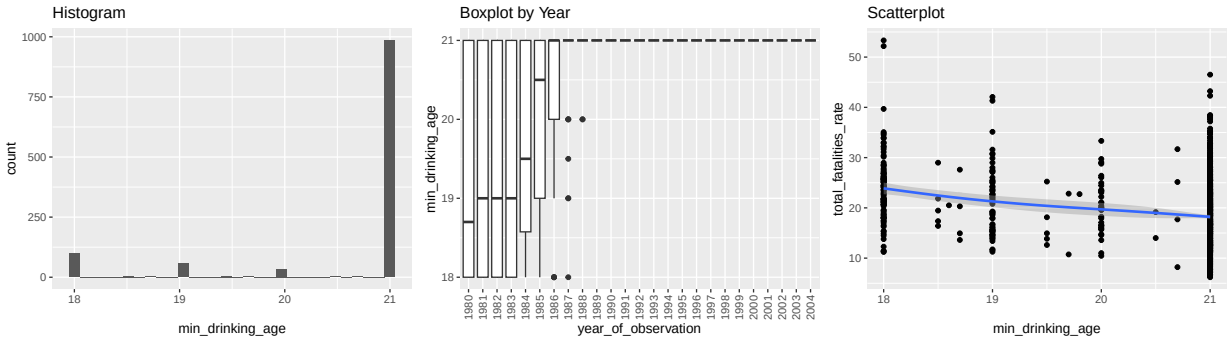
> The histogram is rather normally distributed, doesn't necessarily need a tranformation. From the boxplot, the the median amount is increasing per year of population, and the scatterplot shows a positive correlation between fatalities rate and vehicle miles per capita.

2.1.6 Speed limit



We're treating speed limit as a numeric variable, but it could otherwise be a factor variable. The most common speed limit is 60mpg, and from the boxplot we can see an increase in median speed limit as the years progress. From the scatterplot R wasn't able to fit a smoothed line, but there does appear to be a slight upward trend.

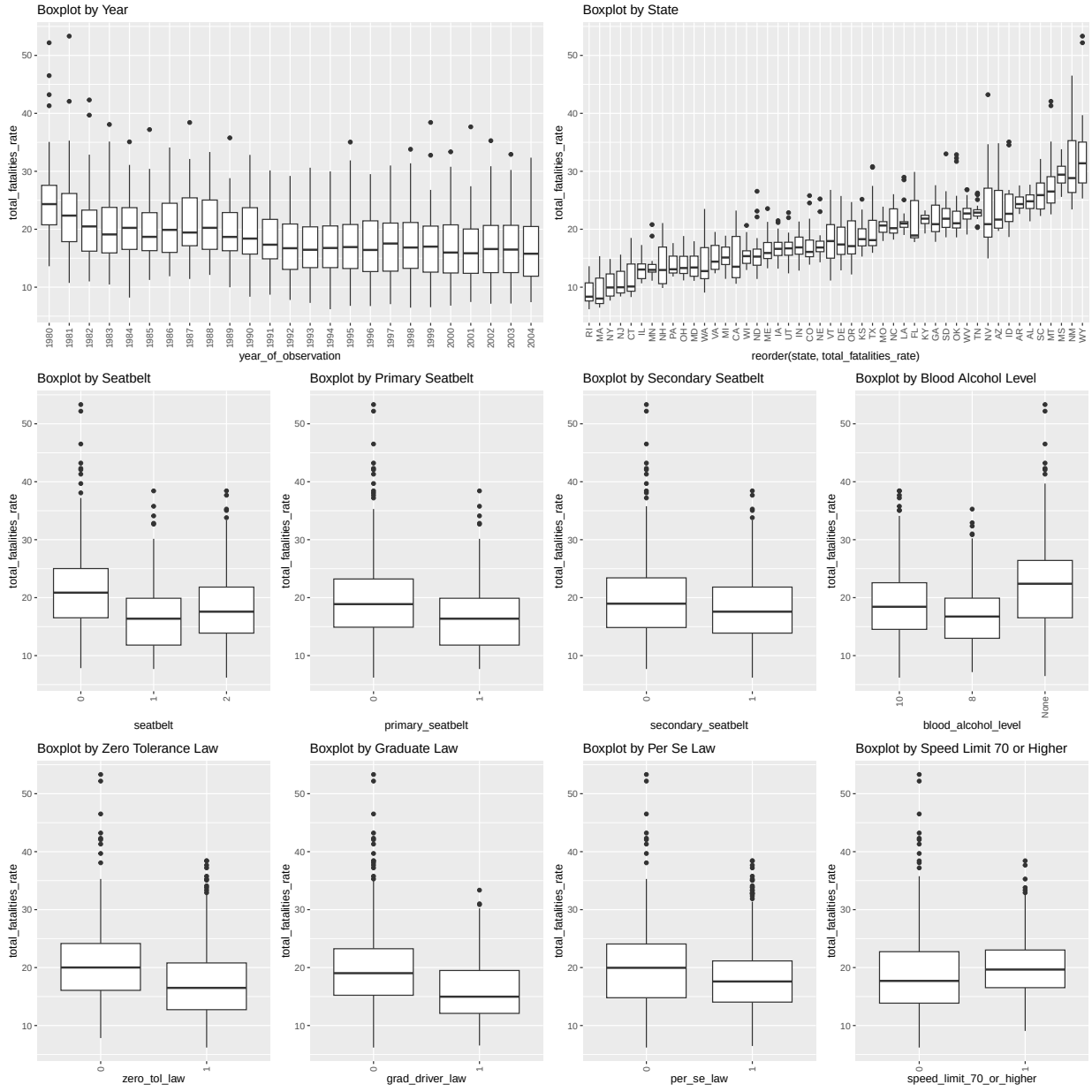
2.1.7 Minimum Drinking Age



From the histogram, the most common minimum drinking age is 21, and that we can see the drinking age increasing over time until it is 21 across all the states. From the scatterplot we can see that as the minimum drinking age increases the total fatalities rate decreases.

2.2 Factor Variable Analysis

2.2.1 Year of Observation



The boxplots above show the total fatality rate mapped against the different variabkes. From the graphs, in order, we see:

1. *Year*: There is a subtle decline in fatalities by year.
2. *State*: The fatality rate by state varies drastically, and we expect this to be a significant variable in our model.
3. *Seatbelt*: Fatality rates are the highest when ther are no seatbelts.
4. *Primary Seatbelt Law*: There appears to be a slight deline in the fatality rate when there are primary seatbelt laws.

5. *Secondary Seatbelt Law*: The different in total fatality rate is relatively minor between both factors.
6. *Blood Alcohol Level*: There appears to be higher fatalities when there are no blood alcohol laws, and also when the laws are higher (0.10 vs 0.08).
7.
   - 9. *Zero tolerance, Graduate driver and Per Se Laws*: All of the laws show a decline in total fatality rate
8. *Speed Limit 70 or Higher*: There appears to be a slightly lower fatality rate when the speed limit is 70 or lower.

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:

```
mod.prel <- plm(
  formula = total_fatalities_rate ~ year_of_observation,
  data = pdriving,
  model = "pooling" #same as linear model
  )
```

Each of the dummay variables are statistically significant.

- Why is fitting a linear model a sensible starting place?

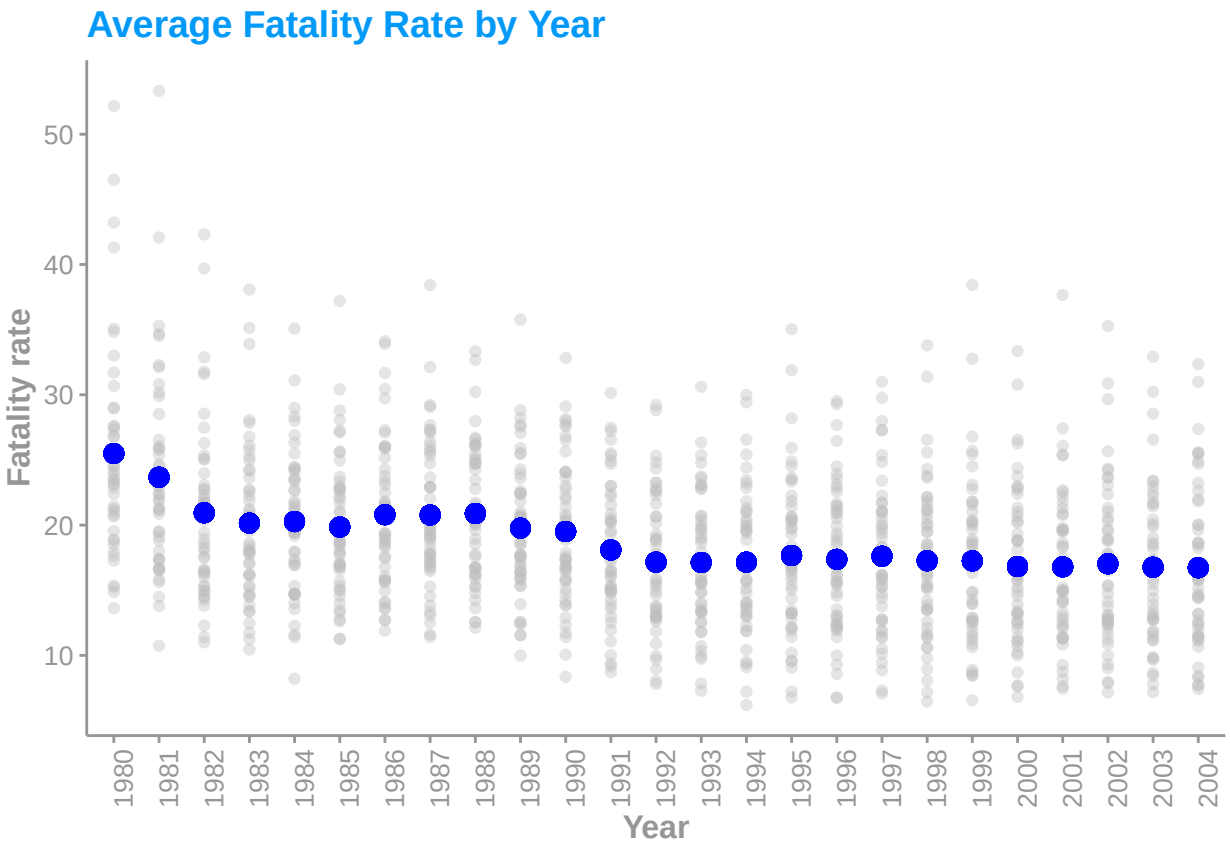
  A linear model is always a sensible place to start as they are generally the most parsimoneous and easily explainable models. For this dataset in particular, in the EDA we have already identified some areas in which there appear to be linear relationships, therefore, it's entirely appropriate to start with linear methods in the modelling process.

- What does this model explain, and what do you find in this model?

  The linear model is above is only a function of year. Where each year is predicting the average in that particularly. As we can see, that as year increases there is a slight deline in the total fatalities rate, and eventually begins to stabalize. This can be visualized in the graph below.

Table 2: Preliminary Model

	Dependent variable:
	Total Fatality Rate
Dummy Variable 1981	−1.824 (1.226)
Dummy Variable 1982	−4.552*** (1.226)
Dummy Variable 1983	−5.342*** (1.226)
Dummy Variable 1984	−5.227*** (1.226)
Dummy Variable 1985	−5.643*** (1.226)
Dummy Variable 1986	−4.694*** (1.226)
Dummy Variable 1987	−4.720*** (1.226)
Dummy Variable 1988	−4.603*** (1.226)
Dummy Variable 1989	−5.722*** (1.226)
Dummy Variable 1990	−5.989*** (1.226)
Dummy Variable 1991	−7.400*** (1.226)
Dummy Variable 1992	−8.337*** (1.226)
Dummy Variable 1993	−8.367*** (1.226)
Dummy Variable 1994	−8.339*** (1.226)
Dummy Variable 1995	−7.826*** (1.226)
Dummy Variable 1996	−8.125*** (1.226)
Dummy Variable 1997	−7.884*** (1.226)
Dummy Variable 1998	−8.229*** (1.226)
Dummy Variable 1999	−8.244*** (1.226)
Dummy Variable 2000	−8.669*** (1.226)
Dummy Variable 2001	−8.702*** (1.226)
Dummy Variable 2002	−8.465*** (1.226)
Dummy Variable 2003	−8.731*** (1.226)
Dummy Variable 2004	−8.766*** (1.226)
Constant	25.495*** (0.867)
Observations	1,200
R^2	0.128
Adjusted R^2	0.110
F Statistic	7.164*** (df = 24; 1175)
Note:	*p<0.1; **p<0.05; ***p<0.01



We also run the Breusch-Pagam LM test for heteroskedasticity, where we reject the null hypothesis of homoskedasticity win the model as the p-value is below 0.05.

```
pcdtest(mod.prel, test="lm")
```

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  total_fatalities_rate ~ year_of_observation
## chisq = 5033.8, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

We also run the Durbin-Watson test for serial correlation in the model, and we reject the null hypothesis of no serial correlation in the model.

```
pdwtest(mod.prel)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  total_fatalities_rate ~ year_of_observation
## DW = 0.19967, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

- Did driving become safer over this period? Please provide a detailed explanation.

  While the overall fatalities rate has generally declined, we can't say that driving has become safer over the period, as we can see see a significant variance in the results versus the projection. There may be missing variables that are causing the decrease in fatalities.

- 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?

    The only variable is for year, which is not necessarily unbiased estimator of the truth. There is omitted variable risk, and we need to explore alternative models which may be better predictors of total fatalities than just time alone.
  - Are the uncertainty estimate reliable, unbiased estimates of sampling based variability? Or, are they biased due to the way that the data is structured?

    Since the total fatalities can potentially differ by the states due to their laws and safety requirements, there is a potential bias in the data if these features aren't considered within the final model. There is also uncertainty within our model as there is potential serial correlation.

4 (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.

```
mod.exp <- plm(
  formula = total_fatalities_rate ~ year_of_observation +
    blood_alcohol_level + per_se_law + primary_seatbelt +
    secondary_seatbelt + speed_limit_70_or_higher + grad_driver_law +
    percent_age_14_24 + log(unemployment_rate) + vehicle_miles_per_capita,
  data = pdriving,
  model = "pooling" #same as linear model
  )
```

- Transformation of variables

Table 3: Expanded Model

	Dependent variable:
	Total Fatality Rate
Dummy Variable 1981	-2.102^{**} (0.824)
Dummy Variable 1982	-6.237^{***} (0.841)
Dummy Variable 1983	-6.971^{***} (0.858)
Dummy Variable 1984	-5.692^{***} (0.873)
Dummy Variable 1985	-6.359^{***} (0.890)
Dummy Variable 1986	-5.576^{***} (0.926)
Dummy Variable 1987	-6.004^{***} (0.964)
Dummy Variable 1988	-6.107^{***} (1.014)
Dummy Variable 1989	-7.623^{***} (1.052)
Dummy Variable 1990	-8.604^{***} (1.074)
Dummy Variable 1991	-10.750^{***} (1.097)
Dummy Variable 1992	-12.546^{***} (1.117)
Dummy Variable 1993	-12.386^{***} (1.132)
Dummy Variable 1994	-11.918^{***} (1.155)
Dummy Variable 1995	-11.405^{***} (1.182)
Dummy Variable 1996	-13.381^{***} (1.230)
Dummy Variable 1997	-13.332^{***} (1.252)
Dummy Variable 1998	-14.030^{***} (1.270)
Dummy Variable 1999	-13.945^{***} (1.292)
Dummy Variable 2000	-14.190^{***} (1.314)
Dummy Variable 2001	-15.278^{***} (1.332)
Dummy Variable 2002	-16.030^{***} (1.339)
Dummy Variable 2003	-16.411^{***} (1.344)
Dummy Variable 2004	-15.932^{***} (1.374)
Factor Variable for Blood Alcohol of .08	-1.146^{***} (0.368)
Factor Variable for Blood Alcohol of None	1.364^{***} (0.388)
Factor Variable Per Se Law in Effect	-0.498^{*} (0.291)
Factor Variable Primary Seatbelt Law in Effect	-0.348 (0.491)
Factor Variable Secondary Seatbelt Law in Effect	-0.137 (0.428)
Factor Variable Speed Limit 70 or Higher in Effect	2.988^{***} (0.433)
Factor Variable of Graduate Driver Laws in Effect	-0.396 (0.503)
Numeric Variable Percentage of Population aged 14-24	0.192 (0.122)
Numeric Variable Log of Unemployment Rate	5.137^{***} (0.481)
Numeric Variable Vehicles per Capita	0.003^{***} (0.0001)
Constant	-9.637^{***} (2.613)
Observations	1,200
R^2	0.611
Adjusted R^2	0.600
F Statistic	53.887^{***} (df = 34; 1165)
Note:	*p<0.1; **p<0.05; ***p<0.01

For the variables within this model, we have made the following transformations:

- Due to the skew on the results for unemployment rate, we've applied a log tranformation to bring the model more towards normal disrtibution.
- For each factor variable we've set values below 0.5 as zero and above 0.5 as one. This only impacts years when there were changes in laws, which were relately few. This change was made to increase interpretability and to have more parsimoneous models.

While note included in our model, we would also suggest logging state population and vehicle miles (note: different than vehicle miles per capita) as there was also a skew in their data.

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

  "bac10" means there is a blood alcohol limit of 0.10, and "bac08" means there is a blood alcohol limit of 0.08 to be considered legally intoxicated. When neither bac10 or bac08 was flagged, we created a new indicator for "none".

  The baseline factor is a blood alcohol limit of 0.10, and we can see that when the the limit decreases to .08, the total fatality rate decreases by 1.146 which is statistically significant.

  The change between a blood alcohol limit of 0.10 and none, increases the total fatalities rate by 1.364 and this increase is also significant.

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

  When per se laws (i.e. administrative license revocation) are present the total fatality rate decreases by 0.498, and this is statistically significant at 0.1 only.

- Does having a primary seat belt law?

  Primary seatbelt laws are shown to decrease the total number of fatalities by 0.348 and secondary seatbelt laws are shown decrease the fatality rate by 0.137, however, neither variables are statistically significant even at the 0.1 level.

- Impact of Other Variables

  Year: Each year has a negative coefficient and are statistically significant. Relatively to the preliminary model, the impacts are lessened as presumably some of the effect was transferred to other variables. *Speed Limit of 70 or Higher*: When the speed limit is 70 or higher, the fatality rate increases by 2.988 which is statistically significant. *Percentage of Population aged 14-24*: For each percentage of the population aged 14-24 the fatality rate increases by 0.192, however, this is not statistically significant. *Log of Unemployment Rate*: As the log of unemployment increases, the fatality rate increases by 5.137 which is statistically significant. *Vehicles per Capita*: As the vehicles per capita increases, the fatality rate increases by 0.003 which is statistically significant.

- Tests

  We also run the Breusch-Pagam LM test for heteroskedasticity, where we reject the null hypothesis of homoskedasticity win the model as the p-value is below 0.05.

```
pcdtest(mod.exp, test="lm")
```

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  total_fatalities_rate ~ year_of_observation + blood_alcohol_level +     per_se_law + primary_
## chisq = 4082.8, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

We also run the Durbin-Watson test for serial correlation in the model, and we reject the null hypothesis of no serial correlation in the model.

```
pdwtest(mod.exp)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  total_fatalities_rate ~ year_of_observation + blood_alcohol_level +  ...
## DW = 0.41629, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

5 (15 points) State-Level Fixed Effects

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

```
mod.fix <- plm(
  formula = total_fatalities_rate ~ state + year_of_observation + #Do we even need to put state in here
    blood_alcohol_level + per_se_law + primary_seatbelt +
    secondary_seatbelt + speed_limit_70_or_higher + grad_driver_law +
    percent_age_14_24 + unemployment_rate + vehicle_miles_per_capita,
  data = pdriving, #indexed on state and year
  model = "within" #adjusting to a fixed effect model
  )
```

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

  The variables for blood alcohol change as follows:

Model	Level 0.08	Level None
Pooled	-1.146	1.364
Fixed Effect	-0.308	0.944

  As we can see that when removing the effect of state mathematically with the within model the coefficients for blood alcohol levels dampens the impact of blood alcohol laws.

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

Table 4: Fixed Effect Model

	Dependent variable:
	Total Fatality Rate
Dummy Variable 1981	-1.513^{***} (0.414)
Dummy Variable 1982	-2.983^{***} (0.443)
Dummy Variable 1983	-3.458^{***} (0.459)
Dummy Variable 1984	-4.285^{***} (0.466)
Dummy Variable 1985	-4.788^{***} (0.485)
Dummy Variable 1986	-3.742^{***} (0.517)
Dummy Variable 1987	-4.392^{***} (0.555)
Dummy Variable 1988	-4.852^{***} (0.601)
Dummy Variable 1989	-6.230^{***} (0.640)
Dummy Variable 1990	-6.325^{***} (0.665)
Dummy Variable 1991	-7.008^{***} (0.682)
Dummy Variable 1992	-7.857^{***} (0.703)
Dummy Variable 1993	-8.202^{***} (0.716)
Dummy Variable 1994	-8.584^{***} (0.735)
Dummy Variable 1995	-8.384^{***} (0.755)
Dummy Variable 1996	-8.715^{***} (0.801)
Dummy Variable 1997	-8.834^{***} (0.818)
Dummy Variable 1998	-9.491^{***} (0.832)
Dummy Variable 1999	-9.608^{***} (0.843)
Dummy Variable 2000	-10.133^{***} (0.854)
Dummy Variable 2001	-9.773^{***} (0.870)
Dummy Variable 2002	-9.048^{***} (0.879)
Dummy Variable 2003	-9.102^{***} (0.883)
Dummy Variable 2004	-9.518^{***} (0.904)
Factor Variable for Blood Alcohol of .08	-0.308 (0.243)
Factor Variable for Blood Alcohol of None	0.944^{***} (0.261)
Factor Variable Per Se Law in Effect	-1.086^{***} (0.226)
Factor Variable Primary Seatbelt Law in Effect	-1.228^{***} (0.343)
Factor Variable Secondary Seatbelt Law in Effect	-0.355 (0.252)
Factor Variable Speed Limit 70 or Higher in Effect	-0.059 (0.261)
Factor Variable of Graduate Driver Laws in Effect	-0.411 (0.280)
Numeric Variable Percentage of Population aged 14-24	0.189^{**} (0.095)
Numeric Variable Log of Unemployment Rate	-0.580^{***} (0.061)
Numeric Variable Vehicles per Capita	0.001^{***} (0.0001)
Observations	1,200
R^2	0.625
Adjusted R^2	0.598
F Statistic	54.755^{***} (df = 34; 1118)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient on per se laws changes as follows:

Model	Per Se Law
Pooled	-0.498
Fixed Effect	-1.086

Here we can see a stronger impact of implementing per se laws, with the reduction in fatalities going down by 1.1 vs. 0.5 in the expanded pooling model.

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

  The coefficient on seatbelt laws changes as follows:

Model	Primary	Secondary
Pooled	-0.348	-0.137
Fixed Effect	-1.228	-0.355

  Note the secondary seatbelt is not statistically significant for either model, and the primary seatbelt is only significant in the fixed effect model.

  Here we can see that by implementing seatbelt laws there is a decrease in the total fatality rate, which is more consistent with our natural understanding.

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

```
pFtest(mod.fix, mod.exp)
```

```
##
##  F test for individual effects
##
## data:  total_fatalities_rate ~ state + year_of_observation + blood_alcohol_level +  ...
## F = 74.936, df1 = 47, df2 = 1118, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

The pFtest has the null hypothesis that the Pooled OLS model is preferable and the alternative hypothesis that the fixed effect model is preferrable. Since we have a significant p-value the fixed effect model (where we mathematically remove the state effect) is preferred.

- What assumptions are needed in each of these models?

  For a fixed effect model, the necessary assumptions are as follows (taken directly from the live session):

  1. *Linearity*: The model is linear in parameters
  2. *i.i.d.*: The observations are independent across individuals but not necessarily across time. This is guaranteed by random sampling of individuals.

3. *Indentifiability*: The regressors, including a constant, are not perfectly collinear, and all regressors (but the constant) have non-zero variance and not too many extreme values.
4. *Zero conditional means (strict exogeneity)*: $E(x_{it}, u_{is}) = 0$ for $s = 1, 2, 3, ..., T$

For a pooled effect model, the necessary assumptions are as follows (taken directly from the live session):

1. *Linearity*: The model is linear in parameters
2. *i.i.d.*: The observations are independent across individuals but not necessarily across time. This is guaranteed by random sampling of individuals.
3. *Indentifiability*: The regressors, including a constant, are not perfectly collinear, and all regressors (but the constant) have non-zero variance and not too many extreme values.
4. $x_i t$ is uncorrelated with idiosyncratic error term u_{it} and individual-specific effect γ_i

a)

$$E(u_{it} x_{it}) = 0$$

b)

$$E(x_{it}, \gamma_i) = 0$$

- Are these assumptions reasonable in the current context?

We believe that the model is linear within it's parameters, as we are able to estimate the model with a linear equation. We also believe that the data is i.i.d. in respect to the observations across individuals due to the robust data collection methodology. When we analyze the results of the models, we don't see any variables with zero variance which helps us to assume that the identifiability assumption is met.

We also run the Breusch-Pagam LM test for heteroskedasticity, where we reject the null hypothesis of homoskedasticity win the model as the p-value is below 0.05.

```
pcdtest(mod.fix, test="lm")
```

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  total_fatalities_rate ~ state + year_of_observation + blood_alcohol_level +     per_se_law + p
## chisq = 3431.8, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

We also run the Durbin-Watson test for serial correlation in the model, and we reject the null hypothesis of no serial correlation in the model.

```
pdwtest(mod.fix)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  total_fatalities_rate ~ state + year_of_observation + blood_alcohol_level +  ...
## DW = 1.0353, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

6 (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.

  In addition to the fixed effect assumptions, there are the three additional assumptions for a random effect model (taken from the textbook) are:

  1. There are no perfect linear relationships among the explanatory variables.
  2. The expected value of a_i given all explanatory variables is constant: $E(a_i|X_i) = \beta_0$
  3. The variance of a_i given all explanatory variables is constant: $Var(a_i|X_i) = \sigma_a^2$

  In order to test the assumptions we run a Hausman Test for Fixed vs. Random Effects, which determines if the residuals are uncorrelated with other predictors in the model:

```
mod.random <- plm(
  formula = total_fatalities_rate ~ year_of_observation +
    blood_alcohol_level + per_se_law + primary_seatbelt +
    secondary_seatbelt + speed_limit_70_or_higher + grad_driver_law +
    percent_age_14_24 + unemployment_rate + vehicle_miles_per_capita,
  data = pdriving, #indexed on state and year
  model = "random" #adjusting to a fixed effect model
  )

phtest(mod.fix, mod.random)
```

```
##
##  Hausman Test
##
## data:  total_fatalities_rate ~ state + year_of_observation + blood_alcohol_level +  ...
## chisq = 146.91, df = 34, p-value = 5.478e-16
## alternative hypothesis: one model is inconsistent
```

  Since the p-value is less than 0.05, we reject the null hypothesis that random effects are appropriate, suggesting that we should use a fixed effect model.

- 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.

  Since we reject the null hypothesis of the Hausman test, that means the assumptions are not valid and we should not run a random effect 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?

  Using a random effect model inappropriately could impact the validity of the model as the assumptions are not satisfied. The random effect model removes the time varying impact, and would only be showing the overall cofficient averages which doesn't tell us anything about how the variables change over time.

7 (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:

```
#Source: https://www.fhwa.dot.gov/policyinformation/travel_monitoring/tvt.cfm (September 2023 data tab)
vmd_data <-  read.csv("./data/23septvt_formatted.csv")
```

```
#Create a column for year and month
vmd_data$year <- str_sub(vmd_data$OBS_DATE, -2, -1)
vmd_data$month <- str_sub(vmd_data$OBS_DATE, 1, 3)

#Select 2018 and 2022 data
vmd_2018 <- vmd_data %>%
  filter(year == 18) %>%
  select("VMT", "month") %>%
  rename("vmd_2018" = "VMT")

vmd_2020 <- vmd_data %>%
  filter(year == 20) %>%
  select("VMT", "month")%>%
  rename("vmd_2020" = "VMT")


#Create new datatable
vmd_combined <- merge(vmd_2018, vmd_2020)
```

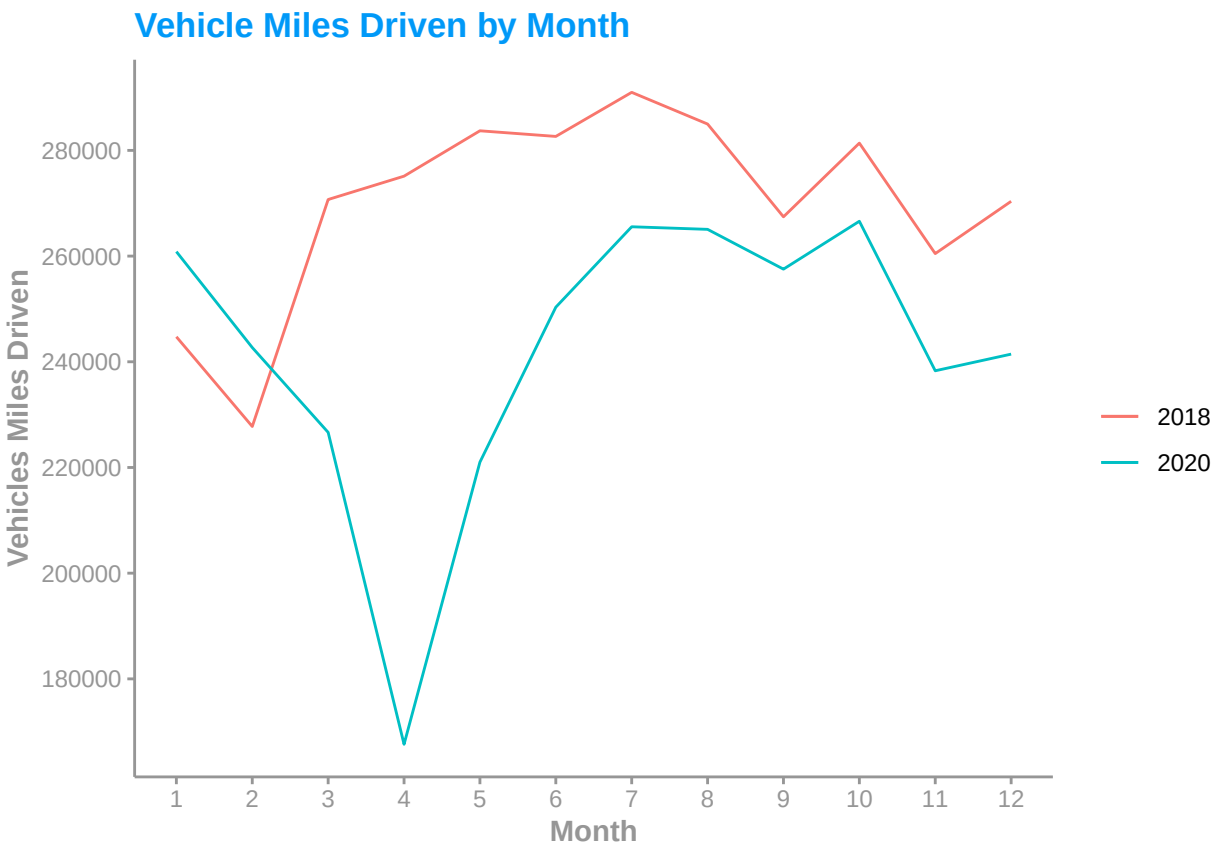
```
#Something went wrong, as I cant convert the variables to numeric, so I'm making it from scratch

my_list <- list(index = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12),
                month = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov",
                vmd_2018 = c(244736, 227759, 270705, 275127, 283713, 282648, 290989, 284989, 267434, 28
                vmd_2020 = c(260847, 242695, 226638, 167617, 221006, 250330, 265550, 265060, 257531, 26

df <- as.data.frame(my_list)

df$difference <- (df$vmd_2020 - df$vmd_2018)
df$percent <- df$difference / df$vmd_2018
```

- 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?

  The largest drop in driving was April, which was the first full month of COVID-19 lockdown. Decreasing from 275,127 in 2018 tp 167,617 in 2020. This is a 39.08% decrease.

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

  January had the highest increase in driving, increasing from 244,726 in 2018 to 260,847 in 2020. This is a 6.58% 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.
- 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.

8 (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-Pagan LM test for cross-sectional dependence in panels
##
## data:  total_fatalities_rate ~ state + year_of_observation + blood_alcohol_level +     per_se_law + p
## chisq = 3431.8, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  total_fatalities_rate ~ state + year_of_observation + blood_alcohol_level +  ...
## DW = 1.0353, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatalities_rate ~ state + year_of_observation + blood_alcohol_level +  ...
## chisq = 298.78, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

From the tests above, reject the null hypothesis of homoskedasticity in the fixed effect model from the Breusch-Pagan test, and we reject the null hypothesis of the Dubin-Watson test which implies there is serial correlation in the fixed effect model. We get a similar result when we run the Breusch-Godfrey/Woolridge test.

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

```
##             Type        SE
## 1    Regular OLS 0.4140226
## 2         Robust 0.7606427
## 3 Cluster Robust 0.3991404
## 4     Newey West 0.6229521
## 5      Arrellano 0.4379227
```

Our model is heteroskedastic and has serial correlation, we would suggest the Arrellano standard errors as the most robust.