How do drunk driving laws affect traffic deaths?

BUAN 6312.004 – Applied Econometrics and Time-series Analysis

Final Project:

7th May 2020

Report by:

Revati Rajane

Project Mentor:

Dr. Moran Blueshtein

Table of Contents

1. Objective**:** 3

2. Approach Followed: 4

3. Expectations of Variable interpretation (Based on economic theory): 5

Data Manipulation**:** 5

4. Overall Trends of dependent variables: 9

5. Exploratory Data Analysis: 12

6. Multiple linear regression: 14

7. Model Selection: 16

1. Alcohol-Involved VFR [mraidall] 16
2. Night-time Vehicle Fatality [mralln] 19
3. All Vehicle Fatality [VFRate] 22

8. Conclusion: 27

9. R Code: 28

1. Objective:

The objective of this project is to understand how do drunk driving laws affects traffic death by considering various influencing factors that were recorded annually over a period of seven years across 48 states in the US.

2. Approach Followed:

1. Conduct a basic data exploratory analysis to gain general overview of each variable and the trend followed by them. It should give us a better understanding of the explanatory variables to assess which are the important independent variables.

2. Consider the economic theory and try to set the expectations about the sign and behavior of explanatory variables.

3. Carried out the exploratory data analysis to understand the nature and behavior of the data and decide dependent variables.

4. Model Selection:

   i. Run the pooling model without fixed effect considering all the variables based on the assumptions in Step 2 and call this as **model1**

   ii. Run the model with same variables as model1 with entities "fixed" effect and named this as **model2**.

   iii. Formulate and test the hypothesis (pFtest) to check which model is better out of above two. (model1 and model2)

   iv. Run the model with only significant variables from model2 with entities "fixed" effect. (**model3**)

   v. Run the model with only significant variables from model2 with entities "fixed" effect and time fixed effect. (**model4**)

   vi. Formulate and test the hypothesis (pFtest) to check which model is better out of above two. (model3 and model4)

5. In the process of selecting the model, I have eliminated the heteroskedasticity problem by carrying white test on each of the model. Also, endogeneity is handled in the process of model selection.

Note: *To handle this problem in detailed manner I have identified 3 dependent variables and followed the above approach for all 3 dependent variables*

3. Expectations of Variable interpretation (Based on economic theory):

Data Manipulation:

The traffic fatality rate is the number of traffic deaths in a given state in a given year, per 10,000 people living in that state in that year. Also, all the variables have different scales, therefore I have scaled some of them

- car$VFRate <- with(car, allmort/pop * 10000) – Vehicle fatality rate per 10000
- car$NVFRate <- with(car,mralln/pop * 10000)- Vehicle fatality rate at night per 10000
- car$AVFRate < - with(car,mraidall/pop *10000)- Vehicle fatality rate due to alcohol per 10000
- as.factor(year) - Converted year into a categorical variable to see yearly effect

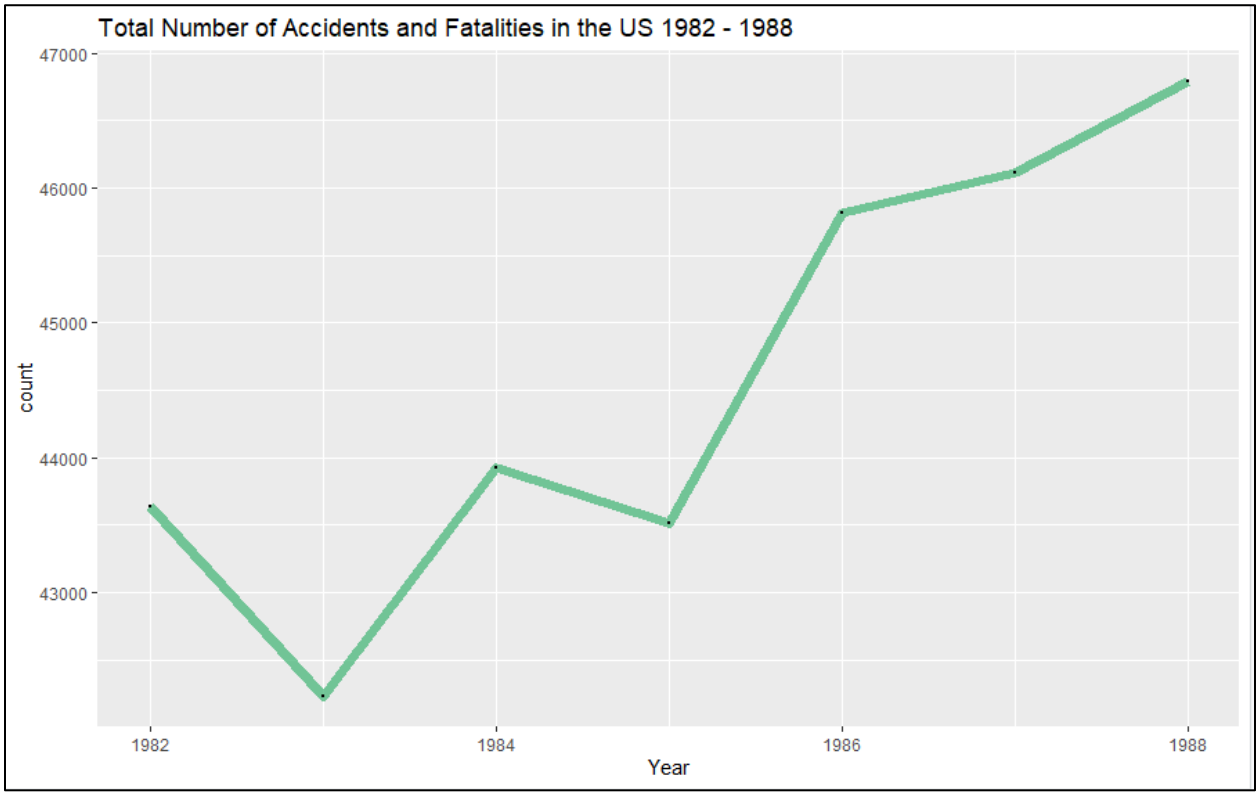
Variable	Descriptions	Expected Behaviour of Explanatory Variable
state	State ID (FIPS) Code	
year	Year	to capture annual effect
spircons	Per Capita Pure Alcohol Consumption (Annual, Gallons)	It is expected that the drunk driver increases the chance of getting involved in an accident. Hence, the spircons is expected to have positive correlation with vehicle fatality rates.
unrate	State Unemployment Rate (%)	Unemployment can cause decrease in fatality rate as less people will travel hence less fatalities.
perinc	Per Capita Personal Income ($)	This can be positively correlated as per capita income increases, more people will buy the cars. Thus, increase the chance of vehicle fatality rates.
beertax	Tax on Case of Beer ($)	It is expected to have inverse effect. Alcohol taxes are expected to reduce alcohol consumption which in turn will reduce VFR
sobapt	% Southern Baptist	This will tend to have less effect as this is a religious factor. I will ignore this.

mormon	% Mormon	This will tend to have less effect as this is a religious factor. I will ignore this.
mlda	Minimum Legal Drinking Age (years)	If more young population are allowed to drink the it can cause VFR to increase
dry	% Residing in Dry Counties A dry county is a county whose government forbids the sale of any kind of alcoholic beverages. Some prohibit off-premises sale, some prohibit on-premises sale, and some prohibit both.	Residing in dry counties might have positive or negative effect on VFR because it is related to drinking habbits of the population as they can avail alcohol from any other county easily
yngdrv	% of Drivers Aged 15-24	Young drivers can cause VFR to increase, hence positively correlated.
vmiles	Ave. Mile per Driver	As average number of miles per driver increases chances of VFR increases hence positively correlated.
jaild	Mandatory Jail Sentence	This punishment is expected to reduce VFR hence negatively correlated.
comserd	Mandatory Community Service	This punishment is expected to reduce VFR hence negatively correlated.
allmort	# of Vehicle Fatalities (#VF)	**This will be considered as dependent variable**
mrall	Vehicle Fatality Rate (VFR)	**This will be considered as dependent variable**
allnite	# of Night-time VF (#NVF)	This is a sub part of mralln , hence I will not include this in model
mralln	Night-time VFR (NFVR)	**This will be considered as dependent variable**
allsvn	# of Single VF (#SVN)	This is a sub part of mrall, hence I will not include this in model

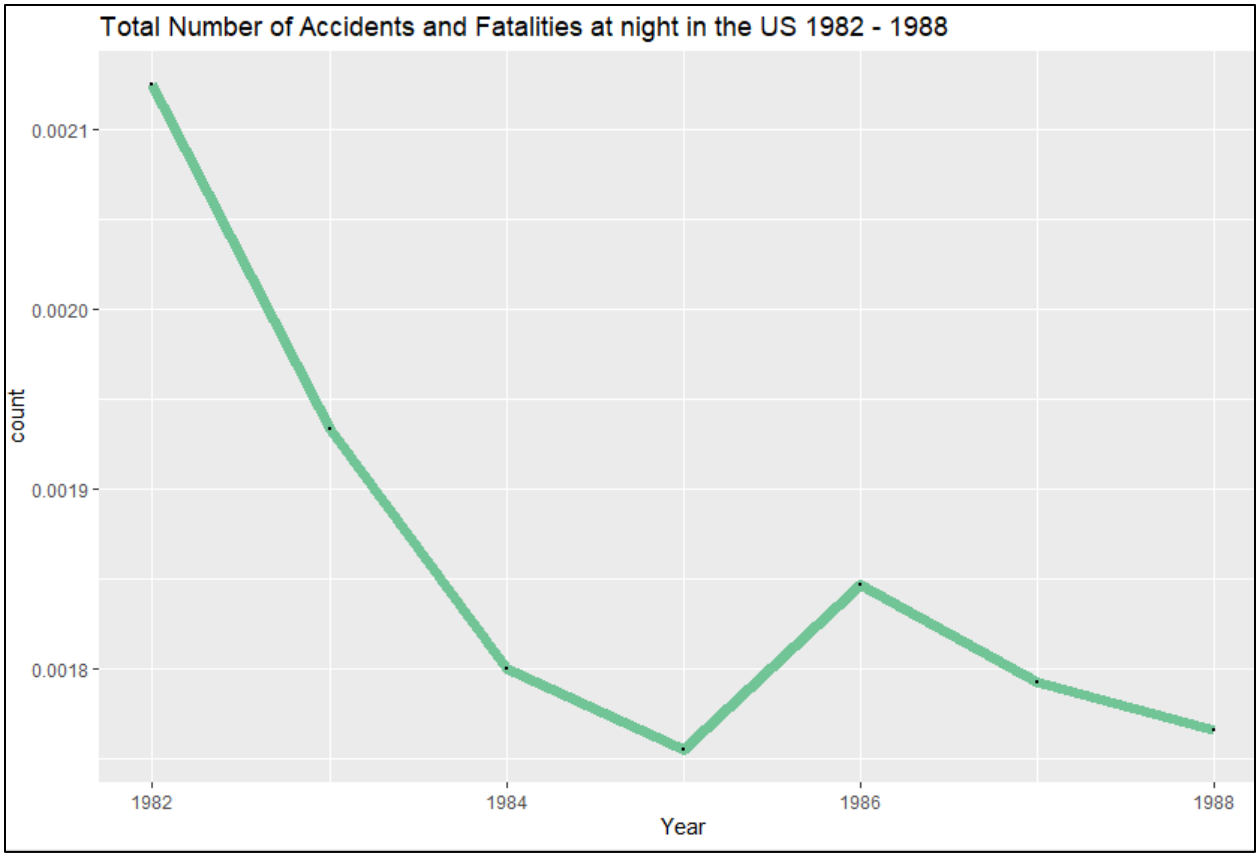
a1517	#VF, 15-17 year olds	This is a sub part of mrall, hence I will not include this in model
mra1517	VFR, 15-17 year olds	This is a sub part of mrall, hence I will not include this in model
a1517n	#NVF, 15-17 year olds	This is a sub part of mralln , hence I will not include this in model
mra1517n	NVFR, 15-17 year olds	This is a sub part of mralln , hence I will not include this in model
a1820	#VF, 18-20 year olds	This is a sub part of mrall, hence I will not include this in model
a1820n	#NVF, 18-20 year olds	This is a sub part of mralln , hence I will not include this in model
mra1820	VFR, 18-20 year olds	This is a sub part of mrall, hence I will not include this in model
mra1820n	NVFR, 18-20 year olds	This is a sub part of mralln , hence I will not include this in model
a2124	#VF, 21-24 year olds	This is a sub part of mrall, hence I will not include this in model
mra2124	VFR, 21-24 year olds	This is a sub part of mrall, hence I will not include this in model
a2124n	#NVF, 21-24 year olds	This is a sub part of mralln , hence I will not include this in model
mra2124n	NVFR, 21-24 year olds	This is a sub part of mrall, hence I will not include this in model
aidall	# of alcohol-involved VF	This is a sub part of mraidall, hence I will not include this in model
mraidall	Alcohol-Involved VFR	**This will be considered as dependent variable**
pop	Population	As the population increases, transportation increases. Hence, positively corelated.

pop1517	Population, 15-17 year olds	Either pop or the given variable is considered, otherwise it will lead to multicollinearity
pop1820	Population, 18-20 year olds	Either pop or the given variable is considered, otherwise it will lead to multicollinearity
pop2124	Population, 21-24 year olds	Either pop or the given variable is considered, otherwise it will lead to multicollinearity
miles	total vehicle miles (millions)	Since vmiles is considered, this can be ignored otherwise it will lead to multicollinearity
gspch	GSP Rate of Change This is a measure of economic growth	This can have positive correlation similar to that of unrate.

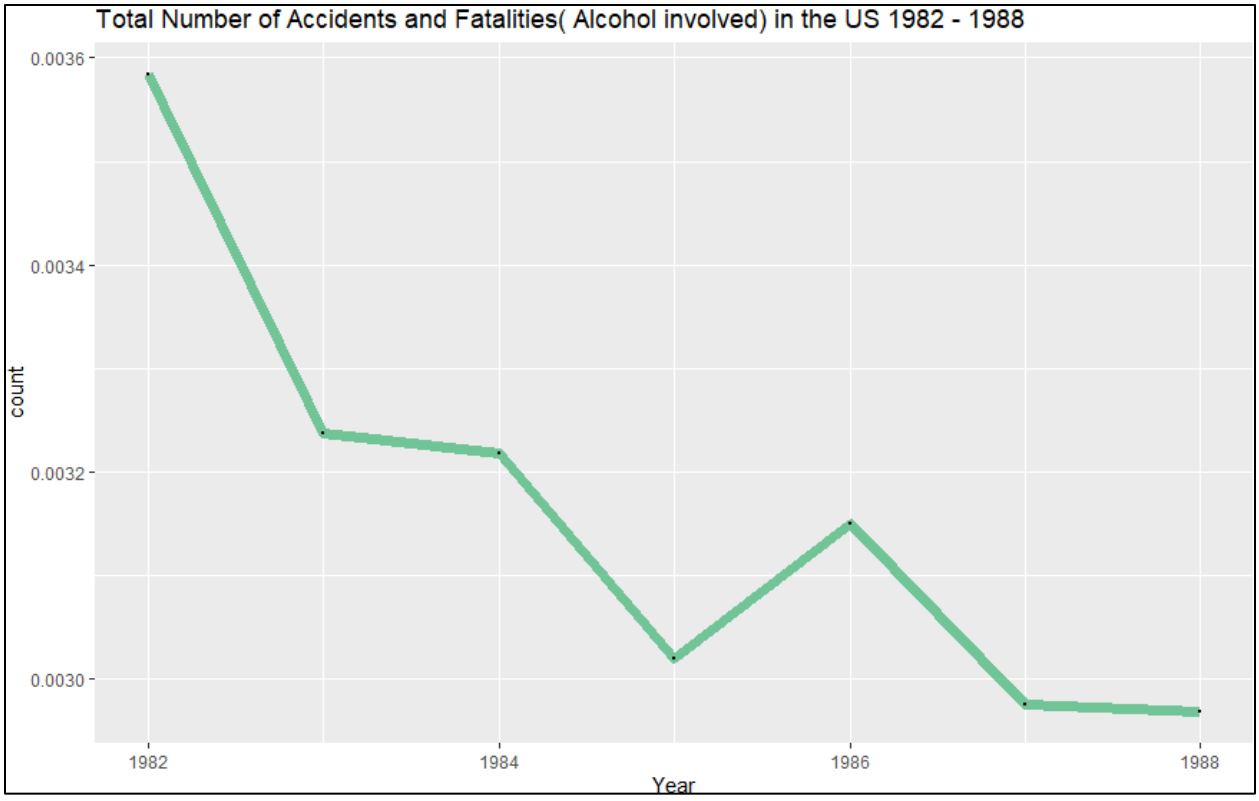
4. Overall Trends of dependent variables:



This graph shows VFR in each year from 1982 to 1988 across all the states. There was a decline in the VFR from 1982 to 1983 followed by a big spike from 1983 to 1986. After 1986, there was no significant trend in the average fatality rate.



This graph shows night VFR in each year from 1982 to 1988 across all the states. There was a decline in the NVFR from 1982 to 1985 followed by a big spike from 1985 to 1986. After 1986, again it is decreasing.



This graph shows VFR involving alcohol in each year from 1982 to 1988 across all the states. There was a decline in the AVFR from 1982 to 1985 followed by a big spike from 1985 to 1986. After 1986, there was a decline.

5. Exploratory Data Analysis:

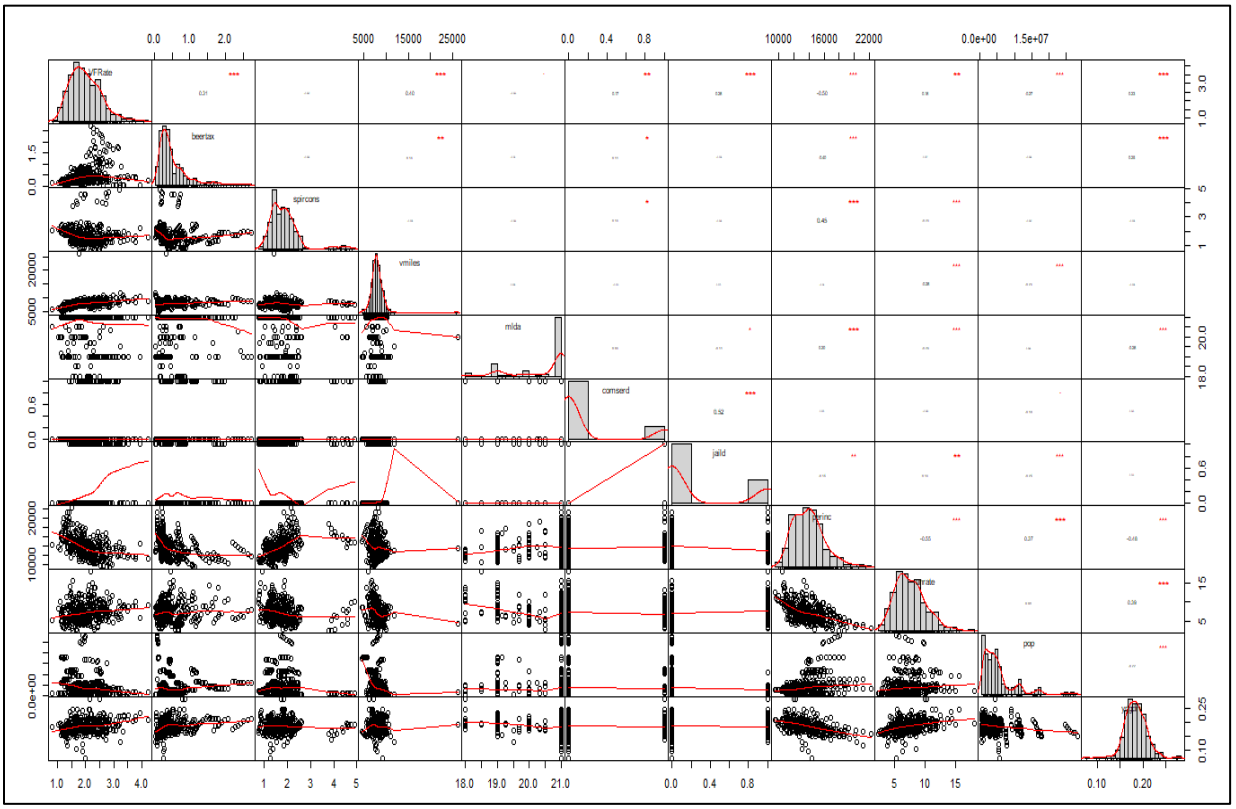
• The data is an observation of 48 states across U.S. annually from 1982 through 1988.

• As shown in screenshot, total number of observations is 336 across 48 states for 7 years. 33 = 48 * 7. So, it is a **balanced panel data**.

• Dependent variable identified - mrall (Vehicle Fatality Rate) and then manipulated to VFR

```
car$VFRate <- with(car, allmort/pop * 10000)
```

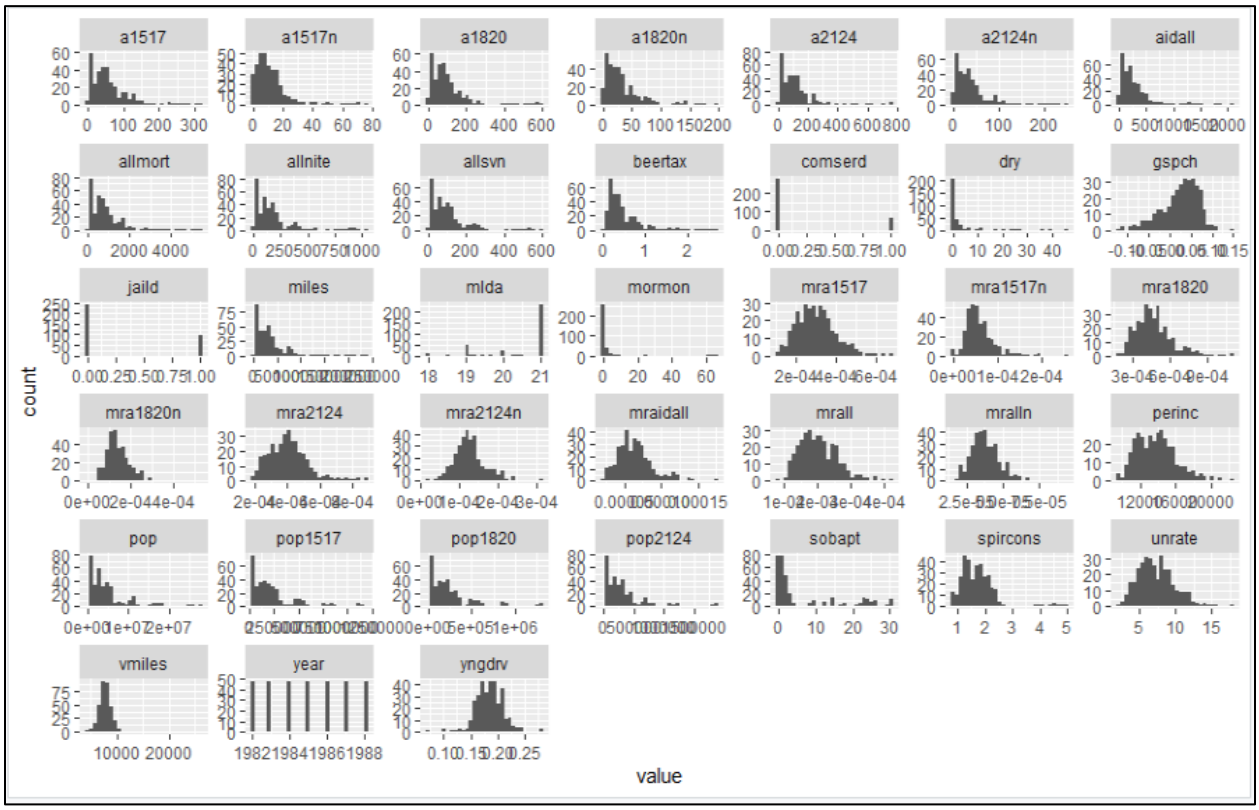
• Expected explanatory variables beertax, spircons, unrate, mlda,dry, yngdrv, vmiles, jaild, comserd, pop1517,gspch



- Young driving population and alcohol consumption are weakly correlated.
- Perinc and unrate are negatively correlated with each other, which is expected as unemployment rate increases per capita income decreases. This might lead to multicollinearity.
- Variables with high correlation are not considered to avoid multicollinearity.

Note: Here values are not visible clearly but re * can be seen.

```
*      p < 0.05
**     p < 0.01
***    p < 0.001
```



Histograms of all explanatory variables are plotted to get an idea of the distribution (left skewed or right skewed or normally distributed).
To start with I have plot a scatter plot between the beer tax variable and the alcohol involved vehicle fatality rate along with the linear regression.

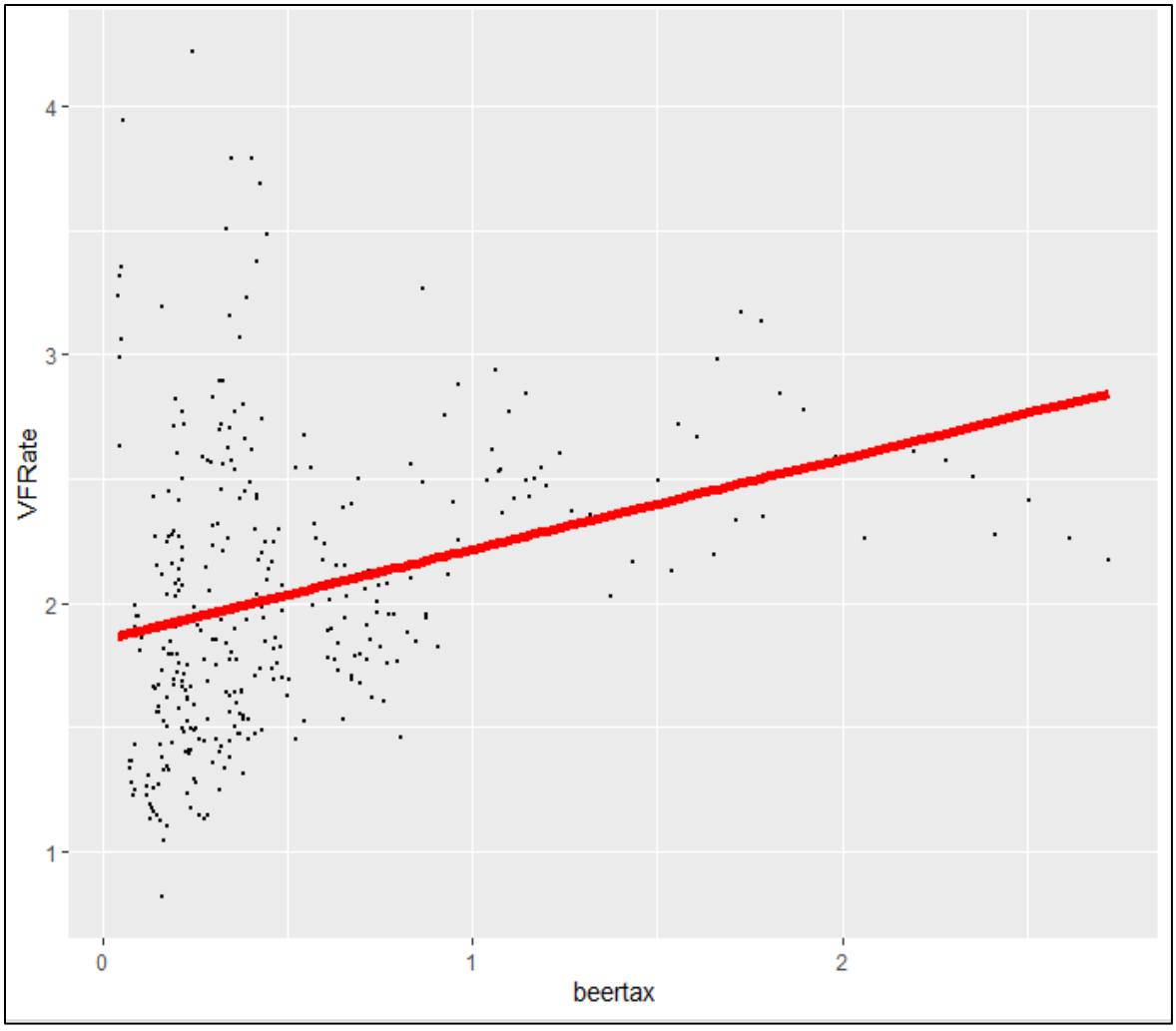
6. Linear regression:

```
#linear model

model_lm <- lm(VFRate~beertax ,data=car)
summary(model_lm)
tidy(model_lm)

plot1 <- car %>%
    ggplot(aes(x =beertax , y = VFRate, group = 1)) +
    geom_point(size = 0.5)
plot1<- plot1 + geom_line(aes(y=predict(lm(VFRate~beertax , data=car))), color="red",
plot1
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.85331    0.04357  42.539  < 2e-16 ***
beertax      0.36461    0.06217   5.865 1.08e-08 ***
```



This plot shows that as beertax increases the VFR increases. This is not logical with the economic theory because as beertax increases, there should be decrease in VFR. But there could be a positive variable bias. Many of the factors have not been included.
Since, we have a panel data, there is a possibility that many of the factors remains constant over the time since we are considering 7 years of data.

7. Model Selection:

Here I would like to consider 3 different models with 3 different dependent variables namely,

1. **mraidall: Alcohol-Involved VFR**
2. **mralln: Night-time VFR (NFVR)**
3. **mrall: Vehicle Fatality Rate (VFR)**

1. Alcohol-Involved VFR [mraidall]

Initially I have considered ***pooled model*** for vehicle fatalities rate involving alcohol. After that ***white test*** is applied in order to remove heteroskedasticity.

```
model1 <- plm(mraidall  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
          I(comserd * jaild)+pop1517+gspch,model="pooling", index = c("state","year"), data=car)
summary(model1)
coeftest(model1,method=vcovHC)
```

```
                       Estimate  Std. Error t value  Pr(>|t|)
(Intercept)          2.3410e-05  3.5022e-05  0.6685 0.5043224
beertax              1.1877e-05  2.7181e-06  4.3697 1.681e-05 ***
spircons             2.5629e-06  1.9248e-06  1.3315 0.1839601
unrate               1.0800e-06  6.3813e-07  1.6925 0.0915213 .
mlda                -1.4068e-06  1.4345e-06 -0.9807 0.3274754
dry                  6.6400e-07  1.4207e-07  4.6737 4.356e-06 ***
yngdrv               1.2836e-04  5.7235e-05  2.2427 0.0255939 *
vmiles               3.4322e-09  8.8602e-10  3.8738 0.0001298 ***
jaild                1.6973e-05  3.9234e-06  4.3261 2.028e-05 ***
comserd              4.9180e-06  5.9600e-06  0.8252 0.4098888
I(comserd * jaild) -1.1833e-05  7.4838e-06 -1.5812 0.1148238
pop1517            -9.5279e-12  5.9364e-12 -1.6050 0.1094739
gspch              -1.1051e-04  3.3346e-05 -3.3142 0.0010240 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here most of the variables are highly insignificant. Also, the coefficient of beetax is positive which was expected as negative. This might be the case of ***endogeneity*** mostly because of omitted variable bias. So, using the fixed effect model might make more sense.

Also consider this hypothesis for this,

H0: OLS Model is better than fixed

H1: Fixed is better than OLS

```
model2 <- plm(mraidall  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
          I(comserd * jaild)+pop1517+gspch,model="within", index = c("state","year"), data=car)
summary(model2)
coeftest(model2,method=vcovHC)
```

```
            Estimate  Std. Error t value  Pr(>|t|)
beertax   -2.2787e-05  1.3442e-05 -1.6952  0.091164 .
spircons   3.1730e-05  7.5787e-06  4.1868 3.809e-05 ***
unrate    -1.8473e-06  6.5911e-07 -2.8027  0.005427 **
mlda      -2.1805e-09  1.4212e-06 -0.0015  0.998777
dry        4.6173e-07  1.0491e-06  0.4401  0.660185
yngdrv     1.0752e-04  6.0231e-05  1.7851  0.075336 .
vmiles    -3.5582e-10  7.1645e-10 -0.4966  0.619837
jaild      2.1400e-05  9.7725e-06  2.1898  0.029377 *
comserd   -2.0374e-05  1.1244e-05 -1.8119  0.071090 .
pop1517   -1.3850e-11  6.1522e-11 -0.2251  0.822050
gspch     -1.5255e-05  2.6345e-05 -0.5790  0.563028
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now Beertax has negative effect and it is significant at 10%. Minimum drinking age is a vital factor which is insignificant here. Surprisingly jaild has positive effect that was not expected and it is significant at 5% level. Community service has a negative effect as expected and the value is significant at 10%

Now we run the test to check our hypothesis.

```
pFtest(model2,model1)
```

```
        F test for individual effects

data:  mraidall ~ beertax + spircons + unrate + mlda + dry + yngdrv +  ...
F = 13.675, df1 = 46, df2 = 276, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Here If the p-value is < 0.05 then the fixed effects model is a better choice. So we reject the null. Hence the **Enitity fixed model is better.**

Now, I have removed all the insignificant variables and next model is generated.

```
model3 <- plm(mraidall  ~ beertax+spircons+unrate+yngdrv+jaild+comserd
              ,model="within", index = c("state","year"), data=car)
summary(model3)
coeftest(model3,method=vcovHC)
```

```
t test of coefficients:

             Estimate  Std. Error t value  Pr(>|t|)
beertax  -2.2418e-05  1.3196e-05 -1.6988  0.090457 .
spircons  3.1792e-05  7.1374e-06  4.4543 1.216e-05 ***
unrate   -1.6753e-06  5.5088e-07 -3.0412  0.002579 **
yngdrv    1.1038e-04  5.9482e-05  1.8557  0.064549 .
jaild     2.0792e-05  9.6376e-06  2.1573  0.031828 *
comserd  -2.0204e-05  1.1126e-05 -1.8160  0.070438 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After this, I have considered a model with time and entity fixed effect. Also, the hypothesis is formulated in order to decide whether time and entity fixed model is better to have or not.

H0: No time fixed effect is need

H1: Time fixed effect is needed.

```
model4 <- plm(mraidall  ~ as.factor(year)+beertax+spircons+unrate+yngdrv+jaild+
               comserd,model="within", index = c("state","year"), data=car)
summary(model4)
coeftest(model4,method=vcovHC)
coeftest(model4, vcov=vcovHC(model4, type="sss", cluster="time"))
```

This output is after applying **Cluster Standard Robust error**.

```
t test of coefficients:

                         Estimate  Std. Error  t value  Pr(>|t|)
as.factor(year)1983 -6.4390e-06  4.0142e-07 -16.0407 < 2.2e-16 ***
as.factor(year)1984 -1.1885e-05  1.2081e-06  -9.8376 < 2.2e-16 ***
as.factor(year)1985 -1.5114e-05  1.6015e-06  -9.4376 < 2.2e-16 ***
as.factor(year)1986 -9.4076e-06  2.6278e-06  -3.5801 0.0004059 ***
as.factor(year)1987 -1.3737e-05  3.0876e-06  -4.4490 1.254e-05 ***
as.factor(year)1988 -1.4969e-05  3.6620e-06  -4.0877 5.724e-05 ***
beertax             -2.1893e-05  8.0556e-06  -2.7178 0.0069897 **
spircons             2.6215e-05  8.5066e-06   3.0817 0.0022671 **
unrate              -3.0990e-06  4.1480e-07  -7.4709 1.062e-12 ***
yngdrv               4.1722e-05  4.5798e-05   0.9110 0.3630918
jaild                2.3856e-05  4.3604e-06   5.4711 1.006e-07 ***
comserd             -1.9865e-05  5.5752e-06  -3.5631 0.0004319 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pFtest(model4,model3)
```

```
        F test for individual effects

data:  mraidall ~ as.factor(year) + beertax + spircons + unrate + yngdrv +  ...
F = 3.8623, df1 = 6, df2 = 275, p-value = 0.00102
alternative hypothesis: significant effects
```

Here, if p value is < 0.05 then use time-fixed effects hence we reject the null and I concluded that **time and fixed effect model is better**.

2. Night-time Vehicle Fatality [mralln]

Initially I have considered ***pooled model*** for vehicle fatalities rate involving alcohol. After that ***white test*** is applied in order to remove heteroskedasticity.

```
model1 <- plm(mralln ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
              I(comserd * jaild)+pop1517+gspch,model="pooling", index = c("state","year"), data=car)
summary(model1)
coeftest(model1,method=vcovHC)
```

```
t test of coefficients:

                      Estimate  Std. Error t value  Pr(>|t|)
(Intercept)         1.7894e-05  1.5829e-05  1.1304  0.259148
beertax             1.6194e-06  1.2286e-06  1.3181  0.188410
spircons            1.1333e-06  8.7000e-07  1.3027  0.193620
unrate              7.0041e-07  2.8843e-07  2.4284  0.015714 *
mlda               -7.5337e-07  6.4839e-07 -1.1619  0.246131
dry                -1.9564e-09  6.4215e-08 -0.0305  0.975714
yngdrv              7.1277e-05  2.5870e-05  2.7552  0.006199 **
vmiles              1.9090e-09  4.0047e-10  4.7668 2.838e-06 ***
jaild               5.1686e-06  1.7733e-06  2.9146  0.003811 **
comserd             4.3830e-06  2.6939e-06  1.6270  0.104707
I(comserd * jaild) -7.6579e-06  3.3826e-06 -2.2639  0.024244 *
pop1517            -2.9874e-13  2.6832e-12 -0.1113  0.911419
gspch              -4.3273e-05  1.5072e-05 -2.8711  0.004362 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here most of the variables are highly insignificant. Also, the coefficient of beer tax is positive which was expected as negative. This might be the case of endogeneity mostly because of omitted variable bias. So, using the fixed effect model might make more sense.

Also consider this hypothesis for this,

H0: OLS Model is better than fixed

H1: Fixed is better than OLS

```
model2 <- plm(mralln  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
              I(comserd * jaild)+pop1517+gspch,model="within", index = c("state","year"), data=car)
summary(model2)
coeftest(model2,method=vcovHC)
```

```
t test of coefficients:

             Estimate  Std. Error t value  Pr(>|t|)
beertax  -1.0100e-05  6.5655e-06 -1.5383 0.1251180
spircons  1.3961e-05  3.7017e-06  3.7716 0.0001985 ***
unrate   -4.7327e-07  3.2193e-07 -1.4701 0.1426693
mlda      3.5818e-07  6.9417e-07  0.5160 0.6062838
dry       5.7373e-07  5.1239e-07  1.1197 0.2638119
yngdrv    1.6991e-05  2.9418e-05  0.5776 0.5640277
vmiles   -2.0814e-10  3.4993e-10 -0.5948 0.5524607
jaild     5.9754e-07  4.7732e-06  0.1252 0.9004661
comserd  -4.0027e-06  5.4921e-06 -0.7288 0.4667321
pop1517   2.0312e-11  3.0049e-11  0.6759 0.4996402
gspch    -2.0815e-05  1.2867e-05 -1.6177 0.1068794
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now Beertax has negative effect and it is insignificant. Minimum drinking age is a vital factor which is insignificant here. Surprisingly jaild has positive effect that was not expected. Community service has a negative effect as expected but both are insignificant.

Now we run the test to check our hypothesis.

```
pFtest(model2,model1)
```

```
        F test for individual effects

data:  NVFRate ~ beertax + spircons + unrate + mlda + dry + yngdrv +  ...
F = 39.289, df1 = 46, df2 = 276, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Here If the p-value is < 0.05 then the fixed effects model is a better choice. So we reject the null. Hence **Fixed effect mode is better**

Now, I have removed all the insignificant variables and next model is generated.

```
model3 <- plm(mralln  ~ spircons+gspch
              ,model="within", index = c("state","year"), data=car)
summary(model3)
coeftest(model3,method=vcovHC)
```

```
t test of coefficients:

             Estimate  Std. Error t value  Pr(>|t|)
spircons  1.2657e-05  2.3511e-06  5.3835 1.525e-07 ***
gspch    -1.8281e-05  1.1006e-05 -1.6610    0.0978 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After this, I have considered a model with time and entity fixed effect. Also, the hypothesis is formulated in order to decide whether time and entity fixed model is better to have or not.

H0: No time fixed effect is need

H1: Time fixed effect is needed.

```
model4 <- plm(mralln  ~ spircons+gspch+as.factor(year)
              ,model="within", data=car)
summary(model4)
coeftest(model4,method=vcovHC)
coeftest(model4, vcov=vcovHC(model4, type="sss", cluster="time"))
```

This output is after applying Cluster Standard Robust error.

```
t test of coefficients:

                        Estimate  Std. Error t value  Pr(>|t|)
spircons              1.3164e-05  5.9815e-06  2.2007  0.028571 *
gspch                 3.5114e-05  2.4858e-05  1.4126  0.158883
as.factor(year)1983 -5.1888e-06  1.1573e-06 -4.4835 1.072e-05 ***
as.factor(year)1984 -8.5862e-06  1.8903e-06 -4.5422 8.281e-06 ***
as.factor(year)1985 -7.4499e-06  1.0402e-06 -7.1620 7.033e-12 ***
as.factor(year)1986 -4.0340e-06  1.4362e-06 -2.8089  0.005322 **
as.factor(year)1987 -4.6574e-06  1.6022e-06 -2.9069  0.003942 **
as.factor(year)1988 -4.9750e-06  1.8497e-06 -2.6896  0.007584 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> pFtest(model4,model3)

        F test for individual effects

data:  mralln ~ spircons + gspch + as.factor(year)
F = 5.8661, df1 = 6, df2 = 280, p-value = 8.83e-06
alternative hypothesis: significant effects
```

Here, if p value is < 0.05 then use time-fixed effects hence we reject the null. Hence model4 is the better model.

3. All Vehicle Fatality [VFRate]

Initially I have considered ***pooled model*** for vehicle fatalities rate. After that ***white test*** is applied in order to remove heteroskedasticity.

```
model1 <- plm(VFRate  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
              I(comserd * jaild)+pop1517+gspch,model="pooling", index = c("state","year"), data=car)
summary(model1)
coeftest(model1,method=vcovHC)
```

```
Coefficients:
                     Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)         1.5177e-02  7.3968e-01  0.0205 0.9836428
beertax             2.4773e-01  5.7408e-02  4.3153 2.124e-05 ***
spircons            6.9839e-02  4.0654e-02  1.7179 0.0867747 .
unrate              3.3985e-02  1.3478e-02  2.5215 0.0121668 *
mlda               -2.8730e-03  3.0298e-02 -0.0948 0.9245139
dry                 1.0178e-02  3.0007e-03  3.3918 0.0007809 ***
yngdrv              1.8867e+00  1.2088e+00  1.5607 0.1195711
vmiles              1.4991e-04  1.8713e-05  8.0108 2.108e-14 ***
jaild               3.0951e-01  8.2864e-02  3.7352 0.0002218 ***
comserd             2.7871e-01  1.2588e-01  2.2141 0.0275203 *
I(comserd * jaild) -2.5236e-01  1.5806e-01 -1.5966 0.1113445
pop1517            -2.5047e-07  1.2538e-07 -1.9977 0.0465892 *
gspch              -1.4929e+00  7.0429e-01 -2.1198 0.0347911 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here most of the variables are highly significant. Also the coefficient of beer tax is positive which was expected as negative and which is significant at 5%. This might be the case of endogeneity mostly because of omitted variable bias. So using the fixed effect model might make more sense.

Also consider this hypothesis for this,

H0: OLS Model is better than fixed

H1: Fixed is better than OLS

```
model2 <- plm(VFRate  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
              I(comserd * jaild)+pop1517+gspch,model="within", index = c("state","year"), data=car)
summary(model2)
coeftest(model2,method=vcovHC)
```

```
t test of coefficients:

            Estimate  Std. Error t value  Pr(>|t|)
beertax  -5.0823e-01  1.7150e-01 -2.9634  0.003308 **
spircons  6.6986e-01  9.6695e-02  6.9276 3.014e-11 ***
unrate   -6.7296e-02  8.4094e-03 -8.0025 3.400e-14 ***
mlda      3.2976e-02  1.8133e-02  1.8185  0.070067 .
dry       2.7901e-02  1.3385e-02  2.0846  0.038027 *
yngdrv    6.0426e-01  7.6847e-01  0.7863  0.432352
vmiles    1.3091e-05  9.1409e-06  1.4322  0.153228
jaild     2.5295e-02  1.2468e-01  0.2029  0.839381
comserd   7.8443e-03  1.4346e-01  0.0547  0.956435
pop1517   7.9348e-07  7.8494e-07  1.0109  0.312961
gspch    -6.0105e-01  3.3613e-01 -1.7882  0.074846 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, Beer tax has negative effect and it is highly significant at 1%. Minimum drinking age is a vital factor which is significant at 10% here. Surprisingly jaild and somserd has positive effect that was not expected, and they are highly insignificant. Also, the population in dry county is significant at 5% but it has positive effect on the vehicle fatality rate which was not expected.

Now we run the test to check our hypothesis.

```
> pFtest(model2,model1)

        F test for individual effects

data:  VFRate ~ beertax + spircons + unrate + mlda + dry + yngdrv +  ...
F = 47.915, df1 = 46, df2 = 276, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Here If the p-value is < 0.05 then the fixed effects model is a better choice. So, we reject the null. Hence the model with fixed effect is better to consider.

Now, I have removed all the insignificant variables and next model is generated.

```
model3 <- plm(VFRate  ~ beertax+spircons+unrate++mlda+dry+gspch
              ,model="within", index = c("state","year"), data=car)
summary(model3)
coeftest(model3,method=vcovHC)
```

```
t test of coefficients:

           Estimate Std. Error t value  Pr(>|t|)
beertax  -0.5209221  0.1675008 -3.1100  0.002063 **
spircons  0.6918432  0.0783229  8.8332 < 2.2e-16 ***
unrate   -0.0652170  0.0077299 -8.4370  1.72e-15 ***
mlda      0.0324146  0.0180065  1.8002  0.072905 .
dry       0.0274180  0.0133241  2.0578  0.040530 *
gspch    -0.5741620  0.3256458 -1.7631  0.078958 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, 4 explanatory variables are highly significant. But here time fixed effects have not considered.

After this, I have considered a model with time and entity fixed effect. Also, the hypothesis is formulated in order to decide whether time and entity fixed model is better to have or not.

H0: No time fixed effect is need

H1: Time fixed effect is needed.

```
model4 <- plm(VFRate  ~ as.factor(year)+beertax+spircons+unrate+mlda+dry+gspch
              ,model="within", index = c("state","year"), data=car)
summary(model4)
coeftest(model4,method=vcovHC)
coeftest(model4, vcov=vcovHC(model4, type="sss", cluster="time"))
```

This output is after Cluster Standard robust error.

```
t test of coefficients:

                      Estimate Std. Error t value  Pr(>|t|)
as.factor(year)1983 -0.0692081  0.0298838 -2.3159 0.0212965 *
as.factor(year)1984 -0.1914366  0.0660894 -2.8966 0.0040740 **
as.factor(year)1985 -0.1927064  0.0471683 -4.0855 5.769e-05 ***
as.factor(year)1986 -0.0283493  0.0529685 -0.5352 0.5929355
as.factor(year)1987 -0.0639028  0.0630034 -1.0143 0.3113396
as.factor(year)1988 -0.0883755  0.0776118 -1.1387 0.2558214
beertax             -0.4602320  0.1278468 -3.5999 0.0003773 ***
spircons             0.7747690  0.1415157  5.4748 9.847e-08 ***
unrate              -0.0724307  0.0127153 -5.6963 3.133e-08 ***
mlda                 0.0186440  0.0101058  1.8449 0.0661277 .
dry                  0.0203626  0.0048388  4.2082 3.486e-05 ***
gspch                0.4751061  0.6697429  0.7094 0.4786839
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> pFtest(model4,model3)

        F test for individual effects

data:  VFRate ~ as.factor(year) + beertax + spircons + unrate + +mlda +  ...
F = 8.4381, df1 = 6, df2 = 276, p-value = 2.036e-08
alternative hypothesis: significant effects
```

Here, if p value is < 0.05 then use time-fixed effects hence we reject the null.

8. Conclusion:

I encountered several issues such as heteroskedasticity, omitted variable bias, endogeneity, and collinearity while working on the given data. After detailed statistical analysis, and regression techniques, I got quite promising results from entity and time fixed effects model.

Based on which, I can infer the following:

1. Beer tax is negatively corelated to the vehicle fatality rate which makes sense and I assumed the same during the expectations. With the increasing tax, beer consumption would reduce leading to fewer fatalities.
2. However, other drunk driving laws such as jail or community service doesn't seem to be significant and effective in controlling the vehicle fatality rate.
3. Per capita alcohol consumption has a significant positive impact on the vehicle fatality rate and that is what I have expected before.
4. Unemployment Rate has negative correlation to the vehicle fatality rate as less people will travel due to work resulting less traffic.

from the above study, I conclude that drunk driving laws may have an impact on fatality rate to an extent but not all the laws seem effective. There are no improvements in the states having laws of mandatory jail and community service. Additionally, I did not find minimum legal drinking age having any significant impact on vehicle fatalities rate.

9. R Code:

```
library(data.table)
library(sandwich)
library(lmtest)
library(DBI)
library(RSQLite)
library(tidyverse)
library(ggplot2)
library(plm)
library(margins)
library(broom)
```

```
library(tidyverse)
library(AER)
library(dplyr)
library(foreign)
library(thePackage)
library(plm)
car <- read.dta("F:/UTD/Spring 20/Econometrics/project/car_fatalities.dta")
summary(car)
```

```
# Data understanding
```

```
# Histograms of all the variables
library(ggplot2)
car %>%
  keep(is.numeric) %>%                  # Keep only numeric columns
  gather() %>%                          # Convert to key-value pairs
  ggplot(aes(value)) +                  # Plot the values
  facet_wrap(~ key, scales = "free") +   # In separate panels
  geom_histogram()                      # as density
#missing Value
sum(is.na(car$jaild))
sum(is.na(car$comserd))

# Data Manuplation
car$VFRate <- with(car, allmort/pop * 10000)
car$NVFRate <- with(car,mralln/pop * 10000)
car$AVFRate < - with(car,mraidall/pop *10000)



# # Number of accidents per year in USA
 car_year<- aggregate(x = car$allmort,              # Specify data column
          by = list(car$year),            # Specify group indicator
          FUN = sum)


 ggplot(aes(x=Group.1, y=x), data = car_year)  + geom_line(size = 2.5, alpha = 0.7, color =
"mediumseagreen", group=1) +
   geom_point(size = 0.5) +
   ggtitle('Total Number of Accidents and Fatalities in the US 1982 - 1988') +
```

```
  ylab('count') +
  xlab('Year')



# # Number of accidents per year in USA
car_year<- aggregate(x = car$mraidall,                # Specify data column
                     by = list(car$year),             # Specify group indicator
                     FUN = sum)


ggplot(aes(x=Group.1, y=x), data = car_year)  + geom_line(size = 2.5, alpha = 0.7, color =
"mediumseagreen", group=1) +
  geom_point(size = 0.5) +
  ggtitle('Total Number of Accidents and Fatalities( Alcohol involved) in the US 1982 - 1988') +
  ylab('count') +
  xlab('Year')



#Exploratory Data Analysis


count(car)
unique(car$state)
unique(car$year)


# Corelation between independent variables


 Data.num = car[c("VFRate" , "beertax", "spircons", "vmiles" ,"mlda" ,"comserd", "jaild",
"perinc", "unrate", "pop", "yngdrv")]
```

```
library(PerformanceAnalytics)

chart.Correlation(Data.num,
                  method="pearson",
                  histogram=TRUE,
                  pch=16)
library(psych)
corr.test(Data.num,
          use    = "pairwise",
          method = "pearson",
          adjust = "none")


ggcorr(Data.num,
       label = TRUE,
       label_alpha = TRUE)

library(ppcor)
pcor(Data.num, method = "pearson")


#linear model

model_lm <- lm(VFRate~beertax ,data=car)
summary(model_lm)
tidy(model_lm)
```

```
plot1 <- car %>%
   ggplot(aes(x =beertax , y = VFRate, group = 1)) +
   geom_point(size = 0.5)
plot1<- plot1 + geom_line(aes(y=predict(lm(VFRate~beertax , data=car))), color="red", size=2)
plot1



#Model Selection



############################## 1. Alcohol involved
##############################################

################################################################################
###############
model1 <- plm(mraidall  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
        I(comserd * jaild)+pop1517+gspch,model="pooling", index = c("state","year"), data=car)
summary(model1)
coeftest(model1,method=vcovHC)

model2 <- plm(mraidall  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
        I(comserd * jaild)+pop1517+gspch,model="within", index = c("state","year"), data=car)
summary(model2)
coeftest(model2,method=vcovHC)

pFtest(model2,model1)
```

```
#If the p-value is < 0.05 then the fixed effects model is a better choice

model3 <- plm(mraidall  ~ beertax+spircons+unrate+yngdrv+jaild+comserd
          ,model="within", index = c("state","year"), data=car)
summary(model3)
coeftest(model3,method=vcovHC)


model4 <- plm(mraidall  ~ as.factor(year)+beertax+spircons+unrate+yngdrv+jaild+
           comserd,model="within", index = c("state","year"), data=car)
summary(model4)
coeftest(model4,method=vcovHC)
coeftest(model4, vcov=vcovHC(model4, type="sss", cluster="time"))

pFtest(model4,model3)
#If this number is < 0.05 then use time-fixed effects. In this example, no need to use time-fixed
effects.

############################ 2. Night Vehicle fatality Rate
############################################

#############################################################################################
#################

model1 <- plm(mralln  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
           I(comserd * jaild)+pop1517+gspch,model="pooling", index = c("state","year"),
data=car)
summary(model1)
```

```
coeftest(model1,method=vcovHC)

model2 <- plm(mralln  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
            I(comserd * jaild)+pop1517+gspch,model="within", index = c("state","year"), data=car)
summary(model2)
coeftest(model2,method=vcovHC)

pFtest(model2,model1)
#If the p-value is < 0.05 then the fixed effects model is a better choice

model3 <- plm(mralln  ~ spircons+gspch
           ,model="within", index = c("state","year"), data=car)
summary(model3)
coeftest(model3,method=vcovHC)


model4 <- plm(mralln  ~ spircons+gspch+as.factor(year)
            ,model="within", data=car)
summary(model4)
coeftest(model4,method=vcovHC)
coeftest(model4, vcov=vcovHC(model4, type="sss", cluster="time"))

pFtest(model4,model3)
#If this number is < 0.05 then use time-fixed effects. In this example, no need to use time-fixed
effects.
```

```
######################################## 3.  All Vehicle fatality Rate
##############################################

#############################################################################################
##############
```

```
model1 <- plm(VFRate  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
          I(comserd * jaild)+pop1517+gspch,model="pooling", index = c("state","year"),
data=car)
summary(model1)
coeftest(model1,method=vcovHC)

model2 <- plm(VFRate  ~ beertax+spircons+unrate+mlda+dry+yngdrv+vmiles+jaild+comserd+
          I(comserd * jaild)+pop1517+gspch,model="within", index = c("state","year"), data=car)
summary(model2)
coeftest(model2,method=vcovHC)

pFtest(model2,model1)
#If the p-value is < 0.05 then the fixed effects model is a better choice

model3 <- plm(VFRate  ~ beertax+spircons+unrate++mlda+dry+gspch
         ,model="within", index = c("state","year"), data=car)
summary(model3)
coeftest(model3,method=vcovHC)

model4 <- plm(VFRate  ~ as.factor(year)+beertax+spircons+unrate+mlda+dry+gspch
```

```
        ,model="within", index = c("state","year"), data=car)
summary(model4)
coeftest(model4,method=vcovHC)
coeftest(model4, vcov=vcovHC(model4, type="sss", cluster="time"))



pFtest(model4,model3)
#If this number is < 0.05 then use time-fixed effects. In this example, no need to use time-fixed effects.
```