Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Group Lab 3

Analysis on U.S. traffic fatalities: 1980-2004

Salman Bashir | Shu Ying Chen | YoungKoung Kim

December 12, 2020

Question 1: EDA

The purpose of this lab is to answer the research question : **"Do changes in traffic laws affect traffic fatalities?"** using the *driving.Rdata*, which includes 25 years of data that cover changes in various state drunk driving, seat belt, and speed limit laws (Freeman, 2007). In this section, we conduct the exploratory data analysis to answer this research question. The EDA begins with the overall trend over time. Then, we examine the relationship between total fatality rate *totfatrte* and the explanatory variables based on univariate as well as multivariate analysis.

EDA: Overall Trend

Before we check for univariate impact of laws on fatalities, let's get a sense of how the laws in effect have changed over time. We first see count of states with different laws and how that has evolved over time (Figure 1). Regarding the overall trend of *totfatrte*, here is the list of our findings:

- We notice that there's a gradual decrease in the fatality rates over time.
- We notice that states have changed their laws significantly over time. For instance, no states had seat belt laws until 1986. Similarly, more states have a speed limit of >55mph starting 1988. States have also generally moved towards a lower BAC and imposing Administrative License Revocation (per se) laws.
- Interestingly, the first three years saw a marked decline in the fatality rates, while the major laws related to driving remained mostly unchanged across states. We do notice, however a rise in unemployment, which could affect the fatality rate per 100,000 of population if there was a reduction in the amounts people drive (say as part of their work commutes affected by job losses).

```
data_ <- load("driving.rdata")
data_pdf <- data.frame(data, index = c("year","state"))
```

```
# Create categorical variables based on speed limit law status
data_pdf$SL <-
  ifelse(data_pdf$sl55==1,'55',
         ifelse(data_pdf$sl65==1,'65',
                ifelse(data_pdf$sl70==1,'70',
                       ifelse(data_pdf$sl75==1,"75","Transition"))))
data_pdf$BAC <-
  ifelse(data_pdf$bac08==1,'08',ifelse(data_pdf$bac10==1,'10','trans'))
data_pdf$PERIOD <- ifelse(data_pdf$year<=1987,"pre 88","post 88")
data_pdf$ALR = ifelse(data_pdf$perse==0,"0",
                      ifelse(data_pdf$perse==1,"1","Trans"))
```

```
dfsb <- as.data.frame(table(data_pdf$year, data_pdf$seatbelt))
dfsl <- data.frame(table(data_pdf$year, data_pdf$SL))
dfbac <- data.frame(table(data_pdf$year, data_pdf$BAC))
dfalr <- data.frame(table(data_pdf$year, data_pdf$ALR))

colnames(dfsb) <- c('year','seat_belt_rule','state_count')
colnames(dfsl) <- c('year','Sp_limit_rule','state_count')
colnames(dfbac) <- c('year','BAC','state_count')
colnames(dfalr) <- c('year','ALR','state_count')

plot_bar <- function(df, x, title) {
  p <- ggplot(dfsb, aes(year, state_count, fill=x)) +
    geom_bar(stat="identity")+
    scale_x_discrete(breaks = scales::pretty_breaks(n = 10))+
    ggtitle(title)
  return(p)
}

p2<-ggplot(dfsl, aes(year, state_count, fill=dfsl$Sp_limit_rule)) +
  geom_bar(stat="identity")+
  scale_x_discrete(breaks = scales::pretty_breaks(n = 10))+
  ggtitle("State counts for speed limit laws")

# Mean trend of total fatality rates across all states over time
fatrte_df <- aggregate(totfatrte ~ year,data=data_pdf,mean,na.rm=TRUE)
p<-ggplot(fatrte_df, aes(x=year, y = totfatrte)) + geom_bar(stat="identity") +
  ggtitle("Total Fatality Rate")

# Mean trend of unemployment rate across all states over time
unem_df <- aggregate(unem ~ year,data=data_pdf,mean,na.rm=TRUE)
p6<-ggplot(unem_df, aes(x=year, y = unem)) + geom_bar(stat="identity") +
  ggtitle("National Unemployment Rate")

lay <- rbind(c(1,1,6,6),
             c(2,2,3,3),
```

```
            c(4,4,5,5))

grid.arrange(p,plot_bar(dfsb, dfsb$seat_belt_rule,
                        "State counts for seat belt laws"),
             p2, plot_bar(dfbac, dfbac$BAC, "State counts for BAC laws"),
             plot_bar(dfalr, dfalr$ALR, "State counts for ALR laws"),
             p6, layout_matrix = lay)
```

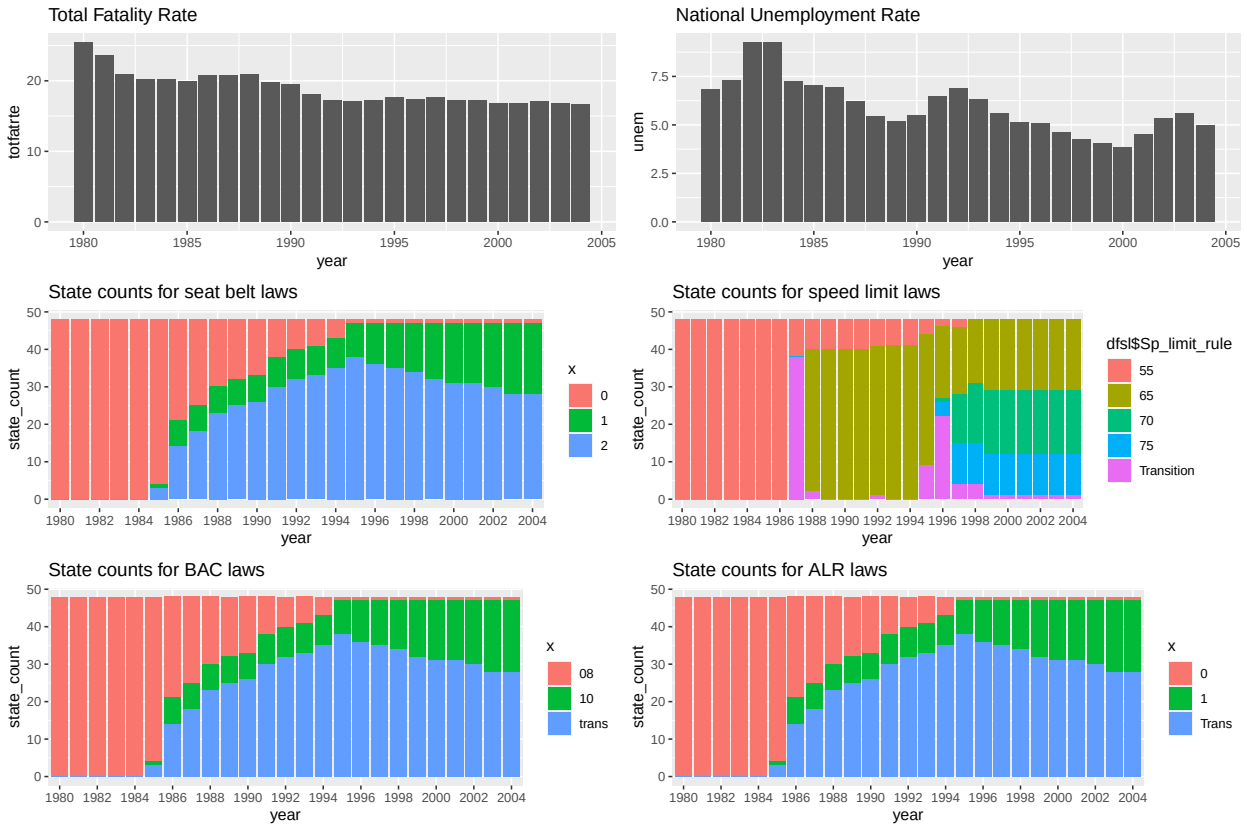

Figure 1: EDA Plots: Overall Trend

EDA: Univariate and Bivariate Analysis

Next we look at standalone distributions of some of the continuous / ordinal variables of interest (*perc14_24, unem, vehicmilespc bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*) to see if any transformations might be needed.

Individual Explanatory Variables From the correlation matrix (Figure 2), we notice a high positive correlation between the fatality rate and percentage of population between the ages 14 and 24. With unemployment the fatality rate exhibits a positive correlation of about 60%. With vehicle miles traveled there's a negative correlation of 90%.The 3rd correlation figure is a bit unintuitive, as we'd expect more driving should in general correspond to more fatalities all things equal.

From the trend plots, we see that there are strong trends across the three factors as well as the fatality rate. Given that the data from the previous year tend to be related to the following year, the upward patterns are not surprising. Due to the nature of the dependency across years, however, we expect that the linear model assumptions including no serial error correlations and homoscedasticitywill be questionable.

Second observation is that taking logs may not be necessary. Since the variation profile across the factors is similar whether or not we take the logs. The only possible advantage is for the variable *unem*, for which the histogram is slightly more close to normal. So we will consider taking logs of unemployment only.

```
# a function to take differencing on the second variable in a dataframe
diff_df <- function(df) {
  n = nrow(df)
  fst_col = (1:n-1)
  sec_col = diff(df[,2])
  ret_df = data.frame(cbind(fst_col,sec_col))
  colnames(ret_df) <- c('year_count','value')

  return(ret_df)
}

# aggreate the response and the continuous explanatory variables by taking
# the mean of all states each year
agg_df <- aggregate(cbind(totfatrte,perc14_24,unem,vehicmilespc)~year,
                    data=data_pdf,mean,rm.na = TRUE)

# Histogram plot function
plot_hist <- function(df, x) {
  p <- ggplot(agg_df,aes(x)) + geom_histogram(fill="white", color = "black") +
    theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank())
}

# Histogram plot function for differenced data
plot_diff <- function(df) {
  p <- ggplot(diff_df(df), aes(year_count,value)) + geom_line() +
    theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank())
}

# Trend plot function
plot_trend <- function(x) {
  p <- ggplot(agg_df, aes(year,x)) + geom_line() +
    scale_x_continuous(breaks = scales::pretty_breaks(n = 3))+
    theme(axis.title.x=element_blank())
}

# for variables in the agg_df, plot trend, diff, diff of log, histogram of the original
# variable and the historgram of the logged variable
```

```
p1 <- plot_trend(agg_df$perc14_24)+ggtitle("trend") + ylab("perc14_24")
p2 <- plot_diff(data.frame(cbind(agg_df$year,agg_df$perc14_24)))+ggtitle("diff")
p3 <- plot_diff(data.frame(cbind(agg_df$year,log(agg_df$perc14_24))))+
  ggtitle("diff of log")
p11 <- plot_hist(agg_df,agg_df$perc14_24) + ggtitle("histogram")
p12 <- plot_hist(agg_df,log(agg_df$perc14_24)) + ggtitle("histogram of log")
p4 <- plot_trend(agg_df$unem) + ylab("unem")
p5 <- plot_diff(data.frame(cbind(agg_df$year,agg_df$unem)))
p6 <- plot_diff(data.frame(cbind(agg_df$year,log(agg_df$unem))))
p13 <- plot_hist(agg_df,agg_df$unem)
p14 <- plot_hist(agg_df,log(agg_df$unem))
p7 <- plot_trend(agg_df$vehicmilespc) + ylab("vehicmilespc")
p8 <- plot_diff(data.frame(cbind(agg_df$year,agg_df$vehicmilespc)))
p9 <- plot_diff(data.frame(cbind(agg_df$year,log(agg_df$vehicmilespc))))
p15 <- plot_hist(agg_df,agg_df$vehicmilespc)
p16 <- plot_hist(agg_df,log(agg_df$vehicmilespc))

# Correlation plot of state-average of totfatrte, perc14_24,
# umem and vehicmilesp
p10<- ggcorr(agg_df, palette = "RdBu", label = TRUE)

lay_ <- rbind(c(10,10,1,2,3,11,12),
              c(10,10,4,5,6,13,14),
              c(NA,NA,7,8,9,15,16))
grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,p13,p14,p15,p16,
             layout_matrix =lay_)
```

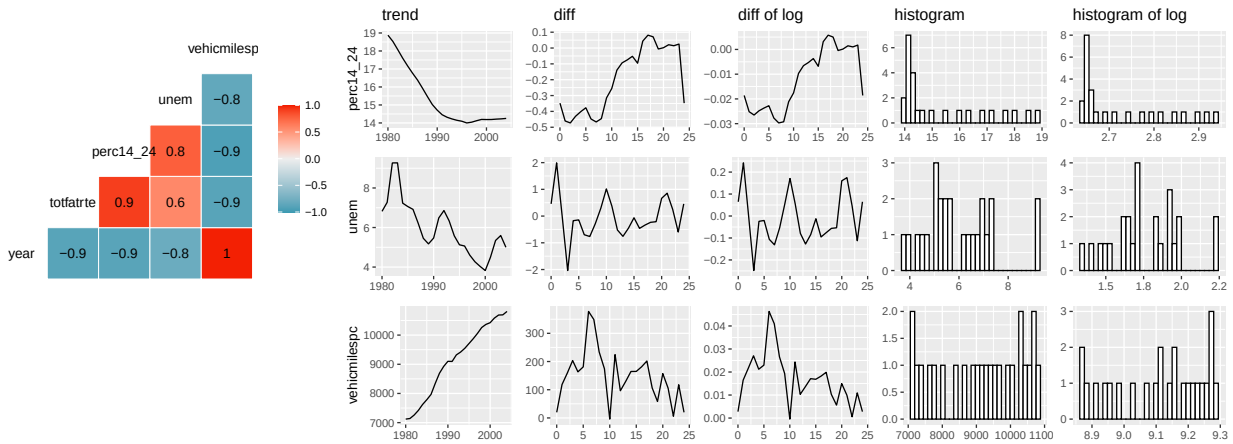

Figure 2: EDA Plots: Univariate and Bivariate Analysis

Speed Limits and Fatality Rates We first check through a univarate EDA if speed limit laws have any visible relationship with fatalities. We note that the dataset distinguishes fatalities at night, weekend and total fatalities. In addition, fatalities are measured in absolute terms, as a ratio of mileage and as a ratio of population.

It is important to account for these different measures. In particular, we expect the normalized fatalities (per mileage and population) to be particularly insightful, to control of different population sizes and driving trends across states.

Table 1 shows that until 1987 all states had 55mph as their speed limit. Afterwards, states have gone on to generally transition to a speed limit higher than 55mph. We added a "Transition" category, to account for records where states changed speed limit laws within a year. We notice that for those records the median is generally higher and inline with 55mph (presumably because the state was transitioning from 55mph to a higher speed limit).

Figure 3 presents the distributions of speed limit law status for total, night time and weekend fatality rates. We notice the following:

- The normalized fatality rates both show a similar pattern i.e. higher median for fatalities corresponding to 55mph speed. The rate falls for 65mph and then trends upwards.
- The same trend doesn't appear for non-normalized fatality rate - highlighting the need to normalize the fatality rate.
- It's not intuitive why 55mph would have a higher median. More states typically have 55mph as the speed limit. Also, states have changed the speed limits over time.

To understand this, we next look at the time series plots (Figure 4). The timeseries plots shed some light on why 55mph seemed associated with higher median of fatality rates. It appears that the higher fatality rates in areas with 55mph speed limit in effect almost always occurred for older dates (1987 and before). After 1990s, 55mph was actually associated with a lower fatality rate vs a higher speed limit. Post 1990s, fatality rates generally increase with speed limits.

```
df_sl <- as.data.frame.matrix(table(data_pdf$year, data_pdf$SL))
df_bac <- as.data.frame.matrix(table(data_pdf$year, data_pdf$BAC))
df_sl_bac <- merge(df_sl, df_bac, by = "row.names", all = TRUE)
colnames(df_sl_bac) <- c("year", "sl 55", "sl 65", "sl 70", "sl 75",
                         "sl trans", "bac 08", "bac 10", "bac trans")
df_sl_bac
```

Table 1: Speed Limit and Blood Acohol Limit Laws by Year

year	sl 55	sl 65	sl 70	sl 75	sl trans	bac 08	bac 10	bac trans
1980	48	0	0	0	0	0	15	33
1981	48	0	0	0	0	0	15	33
1982	48	0	0	0	0	0	15	33
1983	48	0	0	0	0	0	18	30
1984	48	0	0	0	0	2	34	12
1985	48	0	0	0	0	2	37	9
1986	48	0	0	0	0	2	39	7
1987	10	0	0	0	38	2	40	6
1988	8	38	0	0	2	2	39	7
1989	8	40	0	0	0	3	40	5
1990	8	40	0	0	0	4	39	5

year	sl 55	sl 65	sl 70	sl 75	sl trans	bac 08	bac 10	bac trans
1991	8	40	0	0	0	4	38	6
1992	7	40	0	0	1	5	39	4
1993	7	41	0	0	0	5	37	6
1994	7	41	0	0	0	10	33	5
1995	4	35	0	0	9	11	32	5
1996	2	19	1	4	22	12	32	4
1997	2	18	13	11	4	12	32	4
1998	0	17	16	11	4	14	32	2
1999	0	19	17	11	1	15	30	3
2000	0	19	17	11	1	16	29	3
2001	0	19	17	11	1	18	22	8
2002	0	19	17	11	1	26	18	4
2003	0	19	17	11	1	30	5	13
2004	0	19	17	11	1	44	1	3

```
# Plot fatality variables by speed limit laws
p1<- ggplot(data_pdf,aes(x=SL, y =totfat ))+geom_boxplot()
p2<- ggplot(data_pdf,aes(x=SL, y =nghtfat ))+geom_boxplot()
p3<- ggplot(data_pdf,aes(x=SL,y  =wkndfat))+geom_boxplot()

p4<- ggplot(data_pdf,aes(x=SL, y =totfatpvm ))+geom_boxplot()
p5<- ggplot(data_pdf,aes(x=SL, y =nghtfatpvm ))+geom_boxplot()
p6<- ggplot(data_pdf,aes(x=SL,y  =wkndfatpvm))+geom_boxplot()

p7<- ggplot(data_pdf,aes(x=SL, y =totfatrte ))+geom_boxplot()
p8<- ggplot(data_pdf,aes(x=SL, y =nghtfatrte ))+geom_boxplot()
p9<- ggplot(data_pdf,aes(x=SL,y  =wkndfatrte))+geom_boxplot()

grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,nrow=3, ncol = 3)
```

```
# Plot state-average fatality variables over time with speed limit laws
by_sl <- data.frame()
agg_df <- aggregate(cbind(totfat,nghtfat,wkndfat,totfatpvm,nghtfatpvm,
                          wkndfatpvm,totfatrte,nghtfatrte,wkndfatrte)
                    ~ year + SL, data = data_pdf, mean, na.rm = TRUE)

p1<-ggplot(agg_df,aes(x=year,y=totfatpvm, group = SL, col=SL))+geom_line()
p2<-ggplot(agg_df,aes(x=year,y=nghtfatpvm, group = SL, col=SL))+geom_line()
p3<-ggplot(agg_df,aes(x=year,y=wkndfatpvm, group = SL, col=SL))+geom_line()

p4<-ggplot(agg_df,aes(x=year,y=totfatrte, group = SL, col=SL))+geom_line()
p5<-ggplot(agg_df,aes(x=year,y=nghtfatrte, group = SL, col=SL))+geom_line()
p6<-ggplot(agg_df,aes(x=year,y=wkndfatrte, group = SL, col=SL))+geom_line()

grid.arrange(p1,p2,p3,p4,p5,p6,nrow=2,ncol = 3)
```

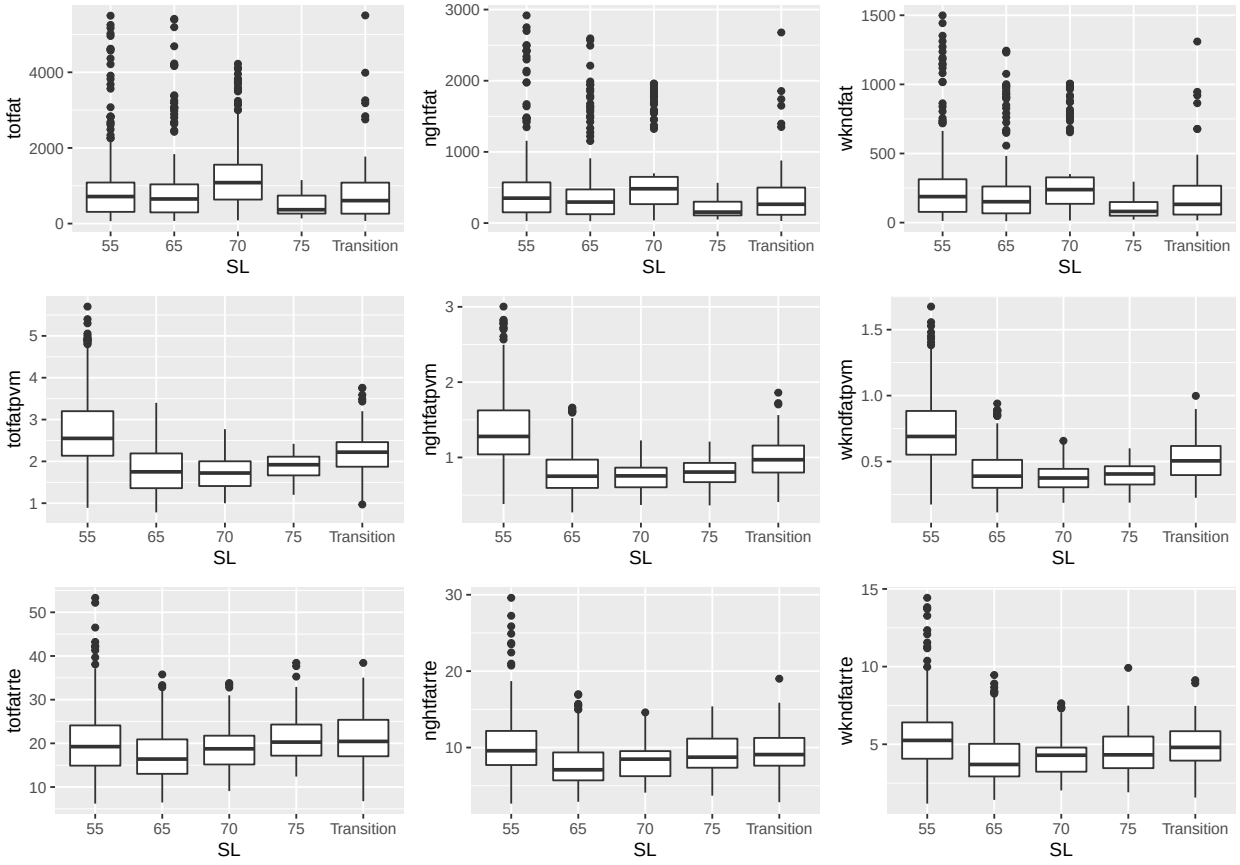

Figure 3: Speed Limits and Total/Night Time/Weekend Fatality Rates

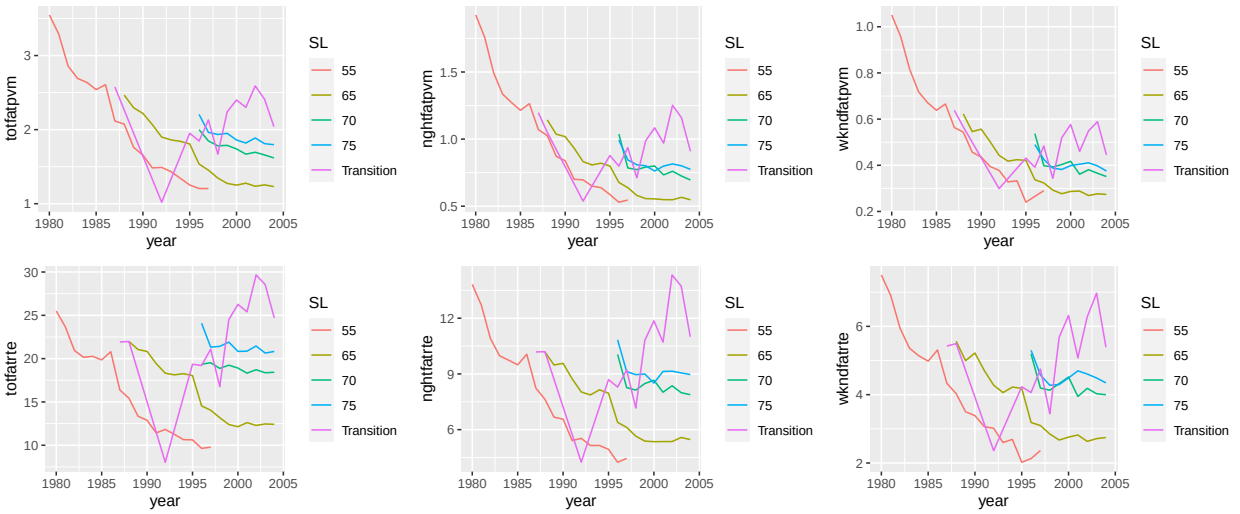

Figure 4: Speed Limits and Fatality Rates by Year

BAC and Fatality Rates Let's conduct a similar univarate EDA for BAC related laws. The last three columns in Table 1 show the changes of BAC limit related laws over time. Figure 5 shows the fatality rates by each BAC law status. We observed that:

- Fatality rates are generally higher with a higher BAC limit or if there's no limit at all.
- The trend is the same whether we use use the rate of the pvm measures.
- The data shows a higher fatality rate at night time than during weekends across BAC.

```
# Plot fatality variables by blood alcohol limit laws
p1<- ggplot(data_pdf,aes(x=BAC, y=totfatpvm ))+geom_boxplot()
p2<- ggplot(data_pdf,aes(x=BAC, y=nghtfatpvm ))+geom_boxplot()
p3<- ggplot(data_pdf,aes(x=BAC,y=wkndfatpvm))+geom_boxplot()

p4<- ggplot(data_pdf,aes(x=BAC, y=totfatrte ))+geom_boxplot()
p5<- ggplot(data_pdf,aes(x=BAC, y=nghtfatrte ))+geom_boxplot()
p6<- ggplot(data_pdf,aes(x=BAC,y=wkndfatrte))+geom_boxplot()

grid.arrange(p1,p2,p3,p4,p5,p6,nrow=2,ncol=3)
```

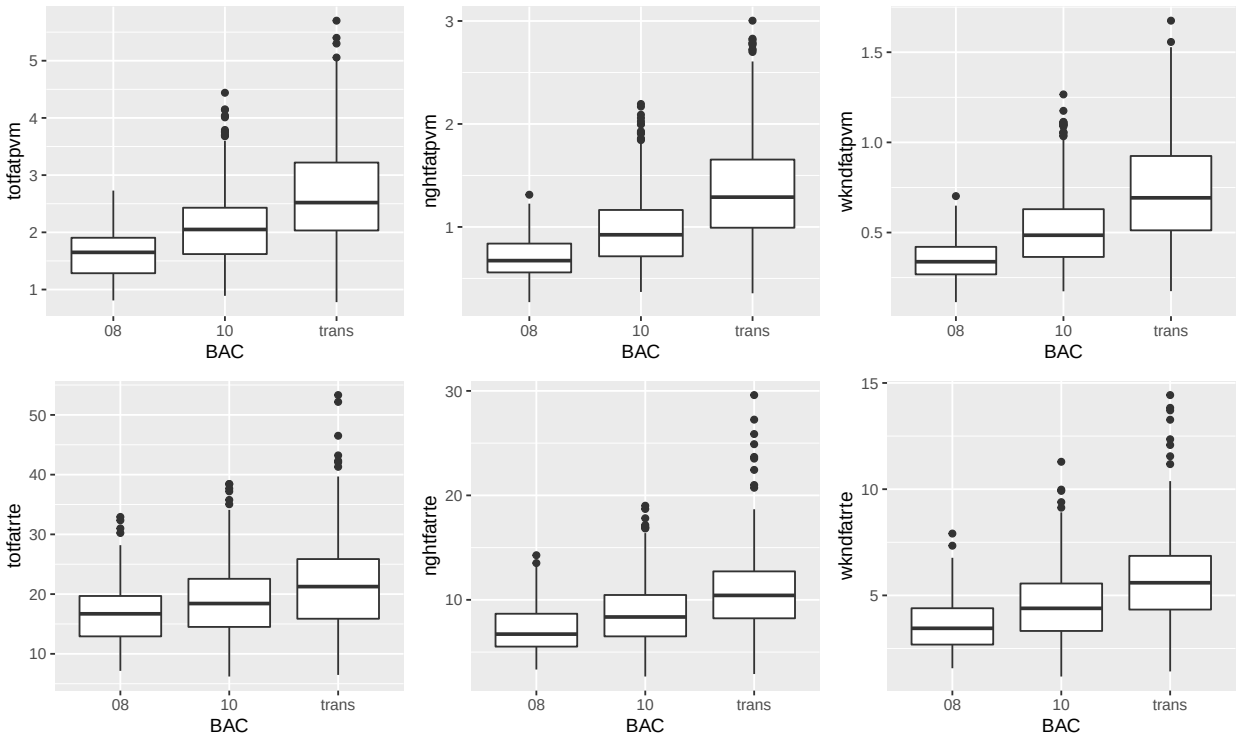

Figure 5: Blood Alcohol Limit Laws and Fatality Rates

Seatbelt Laws and Fatality Rates Figure 6 shows the fatality rates by each seatbelt law status. Here is the list of findings:

- The median fatality rates are higher for states where there are no seatbelt laws in effect.

- States with primary seatbelt laws appear to have the smallest median fatality rates.
- Again we notice the rate of fatalities at night are higher than weekends across seatbelt law types.

```
# Plot fatality variables by seatbelt laws
p1 <- ggplot(data_pdf, aes(x=factor(seatbelt),y=totfatpvm)) + geom_boxplot()
p2<- ggplot(data_pdf,aes(x=factor(seatbelt), y =nghtfatpvm ))+geom_boxplot()
p3<- ggplot(data_pdf,aes(x=factor(seatbelt),y  =wkndfatpvm))+geom_boxplot()

p4<- ggplot(data_pdf,aes(x=factor(seatbelt), y =totfatrte ))+geom_boxplot()
p5<- ggplot(data_pdf,aes(x=factor(seatbelt), y =nghtfatrte ))+geom_boxplot()
p6<- ggplot(data_pdf,aes(x=factor(seatbelt),y  =wkndfatrte))+geom_boxplot()

grid.arrange(p1,p2,p3,p4,p5,p6,nrow=2,ncol=3)
```

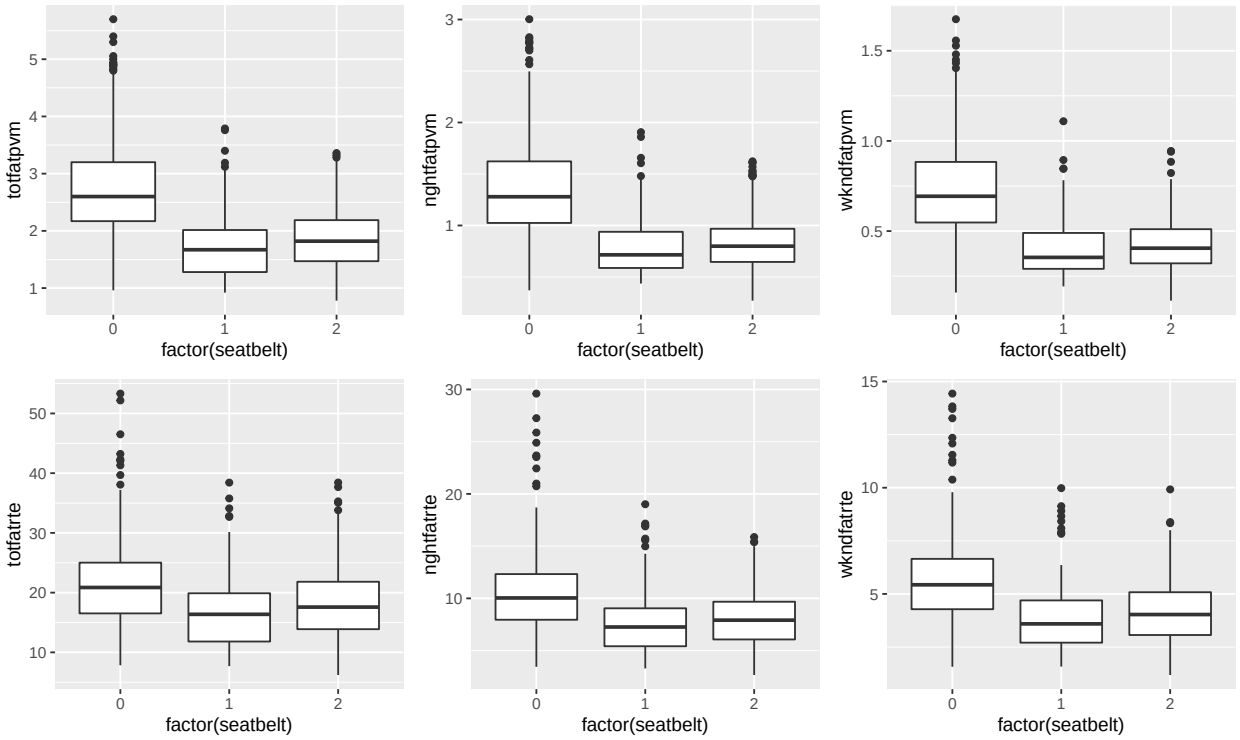

Figure 6: Seatbelt Laws and Fatality Rates

Multivariate EDA

We now look at the a few multivariate relationships (this is not exhaustive as there are too many variables, so we pick a few combinations that we think would be most illustrative). To simplify, we look only at *totfatrte*.

Speed Limits Laws and BAC Laws Before and After 1988 Combined effect of Speed Limits and BAC laws over different periods. Since we noticed a market change in laws related to speed

limit since 1987, we split the dataset into pre 1988 and post 1988 subsets. Based on Figure 8, here are our findings:

- We notice an upward trend in the median fatality rates in speed limits, even after we control for BAC and the time period.
- We notice generally a positive trend of median fatality rates in BAC even after we control for speed limits and time period.

```
# Separate timeseries to pre-1988 and post-1988 and boxplot totfatrte by
# speed limit laws in a matrix of BAC laws
ggplot(data_pdf,aes(x=SL,y=totfatrte)) + geom_boxplot() +
  facet_grid(PERIOD ~ BAC)
```

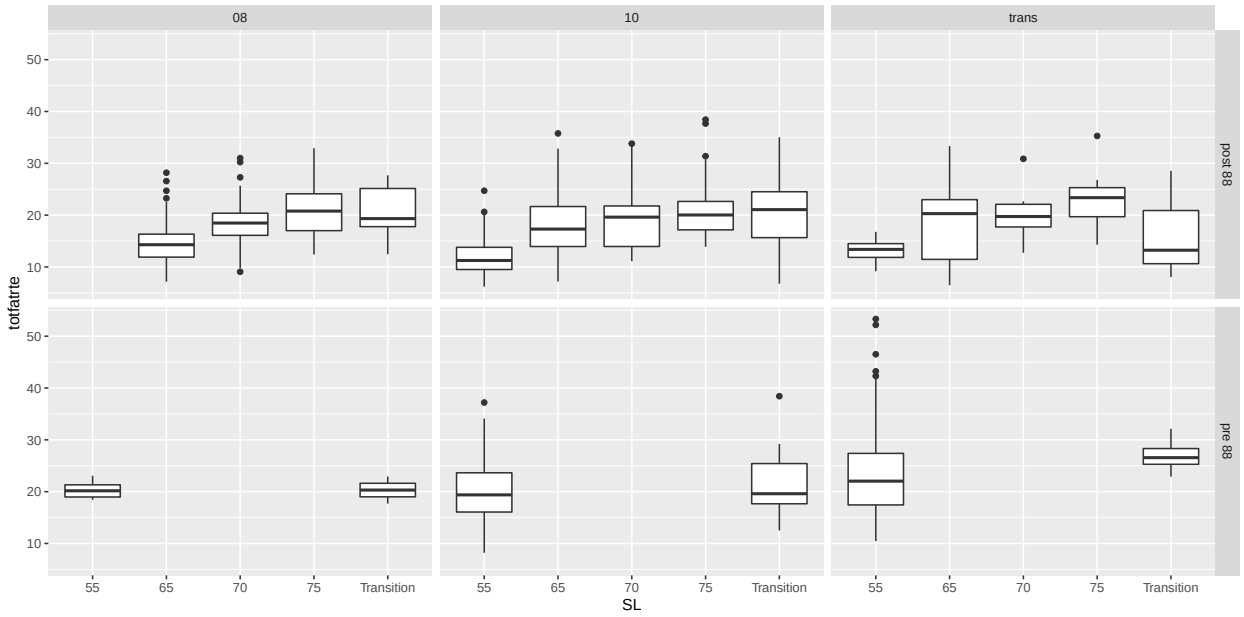

Figure 7: Speed Limits Laws and BAC Laws Before/After 1988

Speed Limits Laws, BAC Laws and Per Se Laws We also look at the effect of Administrative License Revocation (Per se law) Laws, controlling for BAC and SL (Figure 9):

- It appears that perse laws have a significant impact on the median of fatalities even after controlling for Speed Limit Laws and BAC Laws.
- In fact, median fatality rates seem lower in presence of perse laws, controlling for speed limit laws and BAC laws. This would appear to concur that drivers are more cautious in general when Administrative License Revocation laws are in place.

```
# boxplot totfatrte by speed limit laws in a matrix of BAC law by Per Se laws
ggplot(data_pdf,aes(x=SL,y=totfatrte)) + geom_boxplot() + facet_grid(BAC ~ ALR)
```

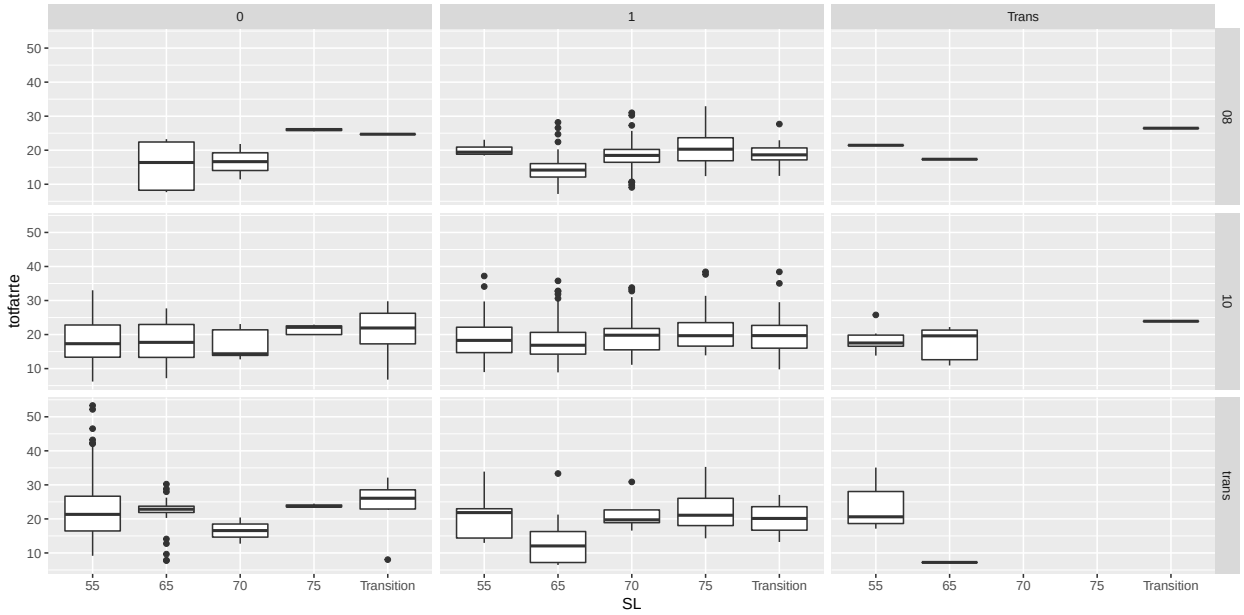

Figure 8: Speed Limits Laws, BAC Laws and Per Se Laws

Question 2 : Linear Regression Model for *totfatrte*

(15%) How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this model explain? Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.

Dependent Variable : *totfatrte*

totfatrte is defined as total fatalities per 100,000 population. This variable is measured for each year and for each state in the dataset. As shown Figure 1, the average *totfatrte* per year is declined. We see a marked decline in fatality rates earlier in the period (1980-1983). We discussed this point in the EDA, that it does coincide with a higher unemployment and may be due to changes in driving patterns as people commute less for work due to high unemployment. There is a smaller decrease between 1989 and 1992.

Linear Regression Model

The linear regression model essentially allows us to estimate the time fixed effects for the fatality rate. The coefficients can be interpreted as the magnitude of the fixed effects for each unit (units being the years). In other words, the coefficients represent the average fatality rate for the year, across all states.

Figure 11 shows the estimated coefficient for the time dummy variables based on the linear regression model. We can see that coefficients generally become more negative over time. For instance,

(relative to the reference year of 1980) the change in fatality rate for 1981 was -1.8. While it was close -8.8 for 2004. And there's no overlap between their 90% confidence intervals.

Thus the fatality rate across the states has been falling over this time period. The range of estimates for the decrease for start of the period 1981 and end of the period 2004 are different with 90% confidence.

Table 2 shows the estimated coefficients based on the linear regression model (Column (1)). The results show that all fixed effects (except for the one for 1981) are statistically different at p<0.01.

```
# Simple linear regression model with only dummy year variables
m1.ols <- lm(totfatrte ~ d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+
                         d91+d92+d93+d94+d95+d96+d97+d98+d99+d00+
                         d01+d02+d03+d04, data = data_pdf)
```

```
# Plot CIs of all coefficients in model #1
plot_summs(m1.ols, scale = TRUE, inner_ci_level = .9)
```

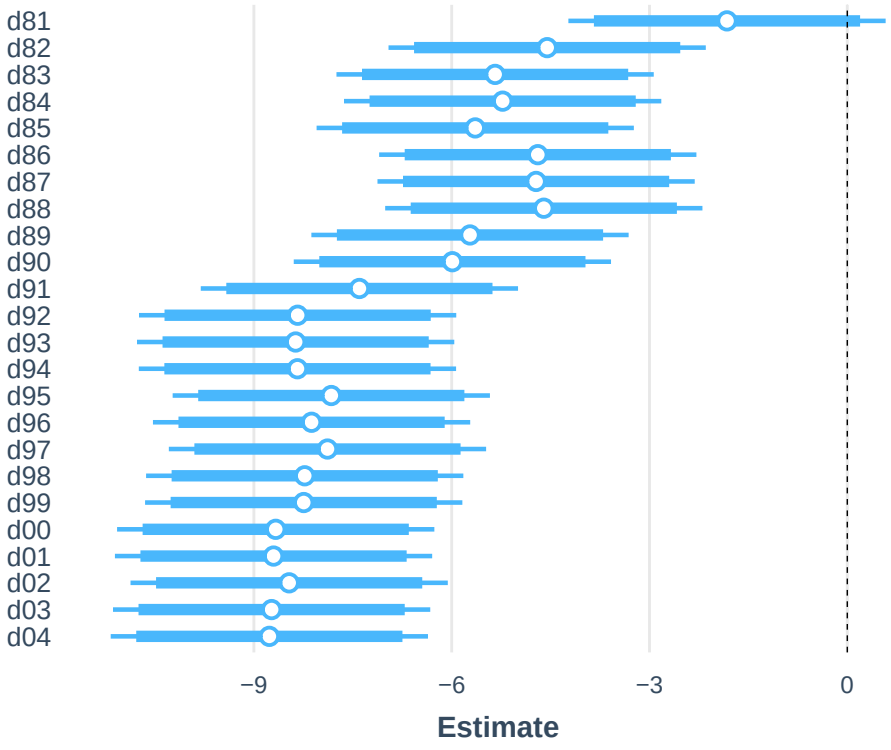

Figure 9: Coefficients for Time Dummy Variables based on Linear Regression Model

Question 3

(15%) Expand your model in *Exercise 2* by adding variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*, and perhaps *transformations of some or all of these*

variables. Please 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 variables *bac8* and *bac10* defined? Interpret the coefficients on *bac8* and *bac10*. Do *per se laws* have a negative effect on the fatality rate? What about having a primary seat belt law? (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.)

Transformation

Indicator Variables *bac8* and *bac10* are related to blood alcohol content (BAC) laws. They are indicator variables, which are coded 0 or 1 for states that have or have not been enacted the described traffic control law which mandates a lower blood alcohol limit threshold of 0.08 or 0.10, respectively. *perse*(administrative license revocation `per se law`), *sbprim*(primary seatbelt law), *sbsecon*(secondary seatbelt law), *sl70plus*(combination of laws related to speed limits of 70, 75 and none), *gdl* (graduated drivers license law) are also indicator variables when each corresponding law was enacted. They are coded 0 or 1 if the laws not or have not been enacted for states, respectively.

If states enacted a law within a year, these seven law indicator variables - *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl* are coded for a fraction of the year. Once the traffic related laws are in effect, they can start to affect traffic fatalities immediately. Thus, for this analysis, these seven law indicator variables are re-coded as dichotomous variables, i.e. the fractional values of the law indicator variables are re-coded as 1 if the laws have been enacted more than half of a year and as 0 otherwise.

Continuous Variables As we found in EDA, *unem*(unemployment rate in percent) is highly skewed. Thus, a logarithm transformation was applied to the variable. On the other hand, *perc14_24*(percent population aged 14 through 24) and *vehicmilespc*(vehicle miles traveled, billions) seem to be normally distributed, so we decided not to transform these two variables.

Model Results

The results of the expanded OLS model are presented under Column (2) in Table 2.

Blood alcohol content law : *bac8 and bac10* The recoded dichotomous variables *bac8* and *bac10* represents whether the blood alcohol content (BAC) laws are in effect.

The estimated coefficients for *bac8* and *bac10* are negative and statistically significant (-2.21 at $p<0.001$ and -1.13 at $p<0.01$, respectively). The results suggest that, on average, a state that has the BAC law in effect has a lower total traffic fatality rate (by 2.21 for *bac8* and 1.13 for *bac10*) than a state that does not have the BAC law holding all other variables equal. Given that the coefficient for *bac8* is smaller than the one for *bac10*, the BAC law limit threshold 0.08 provides a lower total traffic fatality rate than the BAC law limit threshold of 0.10.

Administrative license revocation law : *per se laws* There is marginal statistical evidence that *per se laws* have a negative effect on the fatality rate ($p<0.1$). Holding all other variables constant, a state that has *per se laws* has lower a total traffic fatality rate by 0.54 than a state that does not have the law.

Primary seat belt law: *sbprim* The coefficient for *sbprim* is -0.35 and so the primary seat belt law seems to have a negative effect on the fatality rate holding all other variables constant. However, the result is not statistically significant.

```
# transformation
selected_vars_list <- c("bac08", "bac10", "perse", "sbprim", "sbsecon",
                        "sl70plus", "gdl")
df_traffic_recoded <-
  data.frame(data_pdf,(data_pdf[c(selected_vars_list)] >= 0.5)*1,
             log.unem = log(data_pdf$unem), log.totfatrte = log(data_pdf$totfatrte))
```

```
# Pooled OLS
m2.ols <-plm(totfatrte ~ bac08.1+bac10.1+perse.1+sbprim.1+sbsecon.1+sl70plus.1+
               gdl.1+perc14_24+log.unem+vehicmilespc+
               d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+
               d91+d92+d93+d94+d95+d96+d97+d98+d99+d00+
               d01+d02+d03+d04, data=df_traffic_recoded,
             index=c('state','year'), model = "pooling")
```

Question 4

(15%) Reestimate the model from *Exercise 3* using a fixed effects (at the state level) model. How do the coefficients on *bac08, bac10, perse, and sbprim* compare with the pooled OLS estimates? Which set of estimates do you think is more reliable? What assumptions are needed in each of these models? Are these assumptions reasonable in the current context?

The results of the fixed effects at the state model are presented under Column (3) in Table 2.

Coefficients for *bac08, bac10, perse, and sbprim*

The estimated coefficients for *bac08, bac10, perse, and sbprim* from the fixed effects model are negative and statistically significant. The coefficients for *perse, and sbprim* with the fixed effects model are larger and significant compared to the pooled OLS model estimates. The coefficients for *bac08, bac10* with the fixed effects model are smaller than the pooled OLS model estimates.

Model assumptions: Fixed Effects Model vs. Pooled OLS

The fixed-effects model estimates are more reliable because the fixed effects model takes into account the variability *within* each state. The fixed effects model assumes that the idiosyncratic errors $u_i$are serially uncorrelated as well as constant, i.e. homoscedastic. The fixed effects model allows for the correlation between the unobserved effect α_i and the explanatory variables. On the other hand, the Pooled OLS assumes no correlations between the unobserved effect and the explanatory variables. Therefore, if there is unobserved heterogeneity (i.e. some unobserved factor that affects the dependent variable), and this is correlated with some observed explanatory variables, then the Pooled OLS is inconsistent, whereas FE is consistent.

For the current analysis, it is highly likely to have unobserved effects on the total fatality rate, which is also related to the explanatory variables. For example, the changes in laws are related to unobserved social/historical background of states. Thus, the Pooled OLS model assumption, i.e. no correlations between the unobserved effect and the explanatory variables, is not reasonable and the fixed effect model is more appropriate.

Given that explanatory variables are time variant, it is reasonable that the assumptions of no serial error correlations as well as homoscedasticity are more likely to be violated. In fact, the violations of these assumptions are show in the model diagnostic plots. Although the Pooled OLS seems to satify the normality assumption, the residual plots show that errors seem to have a increasing pattern indicating that no serial error and homoscedasticity assumptions are violated. Compared to the Pooled OLS model, this pattern in the residual plots for the FE model is weaker but still exist (Figure 12). For the further study, we can explore two-way fixed effect regress model to adjust for unobserved unit-specific and time-specific confounders at the same time (Hanck, Arnold, Gerber & Schmelzer, 2020).

```
# Fixed-effect model
m3.fe <-plm(totfatrte ~ bac08.1+bac10.1+perse.1+sbprim.1+sbsecon.1+sl70plus.1+
               gdl.1+perc14_24+log.unem+vehicmilespc+
               d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+
               d91+d92+d93+d94+d95+d96+d97+d98+d99+d00+
               d01+d02+d03+d04,
            data=df_traffic_recoded,
            index=c('state','year'), model = "within")
```

```
# Random-effect model
m4.re <-plm(totfatrte ~ bac08.1+bac10.1+perse.1+sbprim.1+sbsecon.1+sl70plus.1+
               gdl.1+perc14_24+log.unem+vehicmilespc +
               d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+
               d91+d92+d93+d94+d95+d96+d97+d98+d99+d00+
               d01+d02+d03+d04, data=df_traffic_recoded,
            index=c('state','year'), model = "random")
```

```
se <- list(sqrt(diag(vcov(m1.ols, type = "HC1"))),
               sqrt(diag(vcov(m2.ols, type = "HC1"))),
               sqrt(diag(vcov(m3.fe, type = "HC1"))),
               sqrt(diag(vcov(m4.re, type = "HC1"))))

stargazer(m1.ols, m2.ols, m3.fe, m4.re,
          digits = 3, header = FALSE,
          type = "latex", se = se,
          single.row = TRUE, no.space = TRUE,
          font.size = "small",
          column.sep.width = "1pt",
          title = "Regression Models for Total Traffic Fatalities Panel Data",
          model.numbers = FALSE,
          column.labels = c("(1)", "(2)", "(3)", "(4)"))
```

Table 2: Regression Models for Total Traffic Fatalities Panel Data

	Dependent variable:			
	totfatrte			
	OLS		*panel linear*	
	(1)	(2)	(3)	(4)
bac08.1		−2.200*** (0.490)	−0.750** (0.330)	−0.880*** (0.340)
bac10.1		−1.100*** (0.360)	−0.550** (0.230)	−0.630*** (0.240)
perse.1		−0.540* (0.290)	−1.200*** (0.220)	−1.100*** (0.230)
sbprim.1		−0.350 (0.490)	−1.200*** (0.340)	−1.100*** (0.350)
sbsecon.1		−0.130 (0.430)	−0.300 (0.250)	−0.300 (0.260)
sl70plus.1		3.000*** (0.430)	0.021 (0.260)	0.092 (0.270)
gdl.1		−0.400 (0.500)	−0.380 (0.280)	−0.350 (0.290)
perc14_24		0.200 (0.120)	0.170* (0.096)	0.180* (0.098)
log.unem		5.100*** (0.480)	−3.700*** (0.390)	−3.200*** (0.400)
vehicmilespc		0.003*** (0.0001)	0.001*** (0.0001)	0.001*** (0.0001)
d81	−1.800 (1.200)	−2.100** (0.820)	−1.600*** (0.410)	−1.600*** (0.430)
d82	−4.600*** (1.200)	−6.200*** (0.840)	−3.300*** (0.430)	−3.500*** (0.450)
d83	−5.300*** (1.200)	−7.000*** (0.860)	−3.900*** (0.450)	−4.100*** (0.470)
d84	−5.200*** (1.200)	−5.800*** (0.870)	−4.600*** (0.460)	−4.700*** (0.480)
d85	−5.600*** (1.200)	−6.500*** (0.890)	−5.100*** (0.480)	−5.200*** (0.500)
d86	−4.700*** (1.200)	−5.700*** (0.920)	−4.100*** (0.520)	−4.300*** (0.530)
d87	−4.700*** (1.200)	−6.100*** (0.960)	−4.800*** (0.560)	−5.000*** (0.570)
d88	−4.600*** (1.200)	−6.200*** (1.000)	−5.300*** (0.610)	−5.500*** (0.620)
d89	−5.700*** (1.200)	−7.700*** (1.000)	−6.700*** (0.640)	−6.900*** (0.660)
d90	−6.000*** (1.200)	−8.700*** (1.100)	−6.700*** (0.670)	−7.000*** (0.690)
d91	−7.400*** (1.200)	−11.000*** (1.100)	−7.400*** (0.680)	−7.700*** (0.700)
d92	−8.300*** (1.200)	−13.000*** (1.100)	−8.200*** (0.700)	−8.700*** (0.720)
d93	−8.400*** (1.200)	−12.000*** (1.100)	−8.600*** (0.720)	−9.000*** (0.740)
d94	−8.300*** (1.200)	−12.000*** (1.200)	−9.000*** (0.740)	−9.400*** (0.760)
d95	−7.800*** (1.200)	−12.000*** (1.200)	−8.900*** (0.760)	−9.200*** (0.780)
d96	−8.100*** (1.200)	−13.000*** (1.200)	−9.300*** (0.810)	−9.700*** (0.830)
d97	−7.900*** (1.200)	−13.000*** (1.300)	−9.500*** (0.830)	−9.900*** (0.850)
d98	−8.200*** (1.200)	−14.000*** (1.300)	−10.000*** (0.840)	−11.000*** (0.870)
d99	−8.200*** (1.200)	−14.000*** (1.300)	−10.000*** (0.860)	−11.000*** (0.880)
d00	−8.700*** (1.200)	−14.000*** (1.300)	−11.000*** (0.870)	−12.000*** (0.890)
d01	−8.700*** (1.200)	−15.000*** (1.300)	−10.000*** (0.880)	−11.000*** (0.900)
d02	−8.500*** (1.200)	−16.000*** (1.300)	−9.600*** (0.880)	−10.000*** (0.900)
d03	−8.700*** (1.200)	−16.000*** (1.300)	−9.600*** (0.890)	−10.000*** (0.910)
d04	−8.800*** (1.200)	−16.000*** (1.400)	−10.000*** (0.910)	−11.000*** (0.930)
Constant	25.000*** (0.870)	−8.500*** (2.600)		20.000*** (2.300)
Observations	1,200	1,200	1,200	1,200
R^2	0.130	0.610	0.620	0.600
Adjusted R^2	0.110	0.600	0.590	0.590
Residual Std. Error	6.000 (df = 1175)			
F Statistic	7.200*** (df = 24; 1175)	54.000*** (df = 34; 1165)	54.000*** (df = 34; 1118)	1,743.000***

Note: *p<0.1; **p<0.05; ***p<0.01

```r
# diagnostic function
diagnostic_plot = function(model) {
  df_plot <- data.frame(fitted = fitted(model), resid = residuals(model),
                        stdresid = residuals(model)/sd(residuals(model)))

  p1 <- ggplot(df_plot, aes(fitted, resid))+geom_point()+
    stat_smooth(method="loess")+
    geom_hline(yintercept=0, col="red", linetype="dashed")+
    xlab("Fitted values")+ylab("Residuals")+
    ggtitle("Residual vs Fitted Plot")+
    theme(plot.title = element_text(hjust = 0.5, lineheight=1, face="bold"))

  p2<-ggplot(df_plot, aes(sample = stdresid))+
    stat_qq() +
    stat_qq_line()+
    xlab("Theoretical Quantiles")+ylab("Standardized Residuals") +
    ggtitle("Normal Q-Q")+
    theme(plot.title = element_text(hjust = 0.5, lineheight=1, face="bold"))

  p3<-ggplot(df_plot, aes(fitted, sqrt(abs(stdresid))))+
    geom_point(na.rm=TRUE) +
    stat_smooth(method="loess", na.rm = TRUE)+
    xlab("Fitted Value") +ylab(expression(sqrt("|Standardized residuals|")))+
    ggtitle("Scale-Location")+
    theme(plot.title = element_text(hjust = 0.5, lineheight=1, face="bold"))

  return (grid.arrange(p1, p2, p3, ncol = 3))
}
```

```r
diagnostic_plot(m2.ols)
```

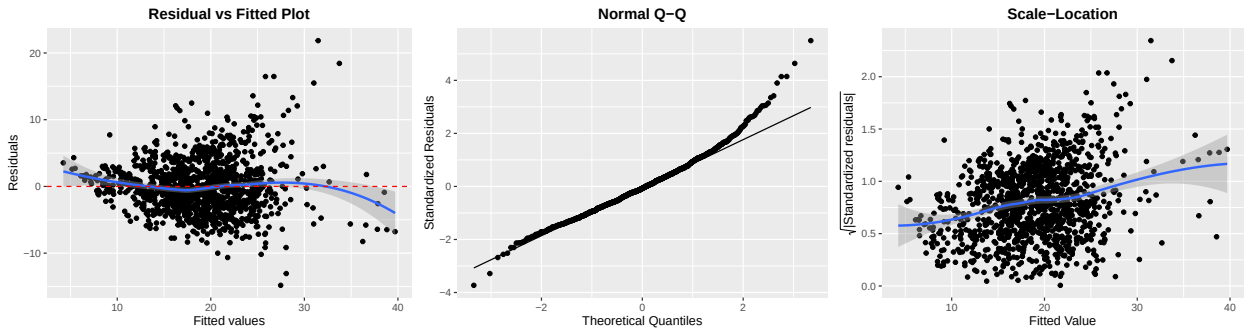

Figure 10: Diagnostics Plots: Pooled OLS

```r
diagnostic_plot(m3.fe)
```

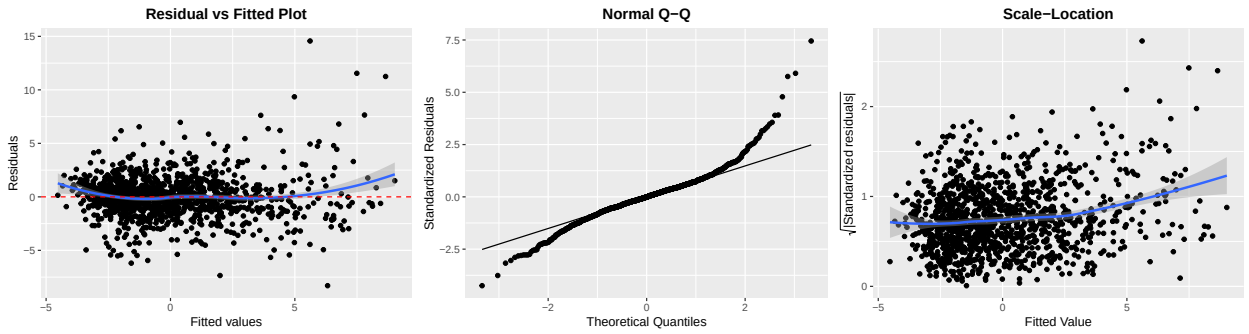

Figure 11: Diagnostics Plots: Fixed Effect Model

```
diagnostic_plot(m4.re)
```

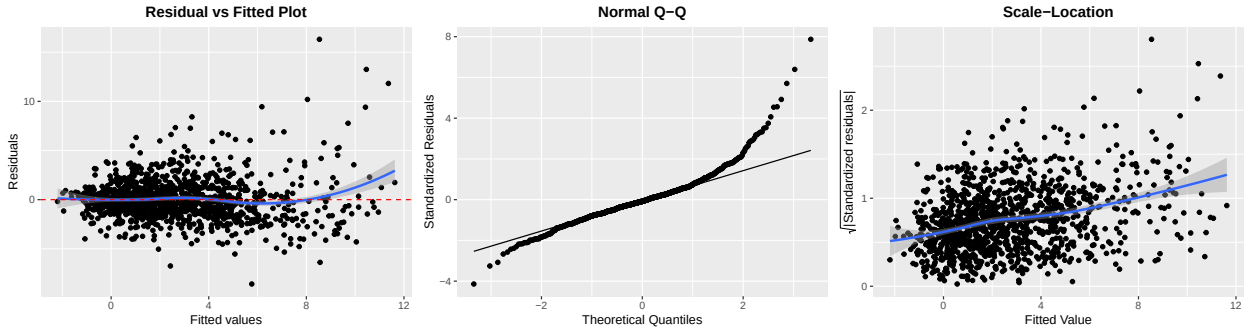

Figure 12: Diagnostics Plots: Random Effect Model

Question 5

(10%) Would you prefer to use a random effects model instead of the fixed effects model you built in *Exercise 4*? Please explain.

The result of Hausman test has a small p-value, which suggests to reject the null hypothesis of that the fixed-effects and random-effects models are indifferent. That is, the fixed-effects model is preferred. We also determined that this suggestion makes sense in a practical perspective. The random-effect model assumptions include all of the fixed effects assumptions plus the additional requirement that the individual state effect a_i is independent of all explanatory variables in all times periods. However, this additional assumption is not satisfied in this case because the explanatory variables are themselves outcomes of changes in driving laws over time and are correlated with individual state driving behaviors captured by a_i. Furthermore, we observed that the two models have similar coefficients and level of significance. Compared to the diagnostic plots of the FE model, those of the RE model (Figure 13) do not exhibit a significant improvement on the residual term.

```
# Hausman Test
phtest(m3.fe, m4.re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ bac08.1 + bac10.1 + perse.1 + sbprim.1 + sbsecon.1 +  ...
## chisq = 162, df = 34, p-value <2e-16
## alternative hypothesis: one model is inconsistent
```

Question 6

(10%) Suppose that *vehicmilespc*, the number of miles driven per capita, increases by 1,000. Using the FE estimates, what is the estimated effect on *totfatrte*? Please interpret the estimate.

```
m3.fe$coefficients['vehicmilespc']*1000
```

```
## vehicmilespc
##         0.95
```

Using the FE estimates, the estimated effect is 0.95 increase in total fatalities per 100,000 population for every 1000 increase in the number of miles driven per capital.

Question 7

(5%) If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors?

Based on the result of Breusch-Godfrey test in question 4, we reject the null hypothesis of of no serial correlation in residuals since p-value is extremely small. Hence, there is heteroskedasticity in the idiosyncractic errors of the model. The consequenses are:

1. The standard errors of the estimators are not valied as estimates of $sd(\hat{\beta}_j)$;
2. The confidence interval of the estimators computed using these standard errors will not truly be a 95% confidence intervals; and
3. The tests of hypothesis using the standard errors (t-test and F-test) are no longer valid, and thus we would commit type I or II error.

Reference

Freeman, D.G. (2007), Drunk driving legislation and traffic fatalities: new evidence on BAC 08 laws, *Contemporary Economic Policy 25*, 293-308.

Hanck, C., Arnold M., Gerber, A., & Schmelzer, M. (2020). *Introduction to Econometrics with R.* https://www.econometrics-with-r.org/