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

Instructions (Please Read Carefully):

- **Due 4pm Tuesday August 11 2020**
- 20 page limit (strict)
- Do not modify fontsize, margin or line-spacing settings
- One student from each group should submit the lab to their student github repo by the deadline; submission and revisions made after the deadline will not be graded
- Answers should clearly explain your reasoning; do not simply 'output dump' the results of code without explanation
- Submit two files:
  1. A pdf file that details your answers. Include all R code used to produce the answers. Do not suppress the codes in your pdf file
  2. The R markdown (Rmd) file used to produce the pdf file

  The assignment will not be graded unless **both** files are submitted
- Name your files to include all group members names. For example the students' names are Stan Cartman and Kenny Kyle, name your files as follows:
  - `StanCartman_KennyKyle_Lab3.Rmd`
  - `StanCartman_KennyKyle_Lab3.pdf`
- Although it sounds obvious, please write your names on page 1 of your pdf and Rmd files
- For statistical methods that we cover in this course, use the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you must provide an explanation of why such libraries and functions are used and reference the library documentation. For data wrangling and data visualization, you are free to use other libraries, such as `dplyr`, `ggplot2`, etc.
- Your report needs to include:
  - A thorough analysis of the given dataset, which includ examiniation of anomalies, missing values, potential of top and/or bottom code, and other potential anomalies, in each of the variables.
  - A comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score. Be

selective when choosing visuals and tables to illustrate your key points and concise with your explanations (please do not ramble).

  - A proper narrative for each question answered. Make sure that your audience can easily follow the logic of your analysis and the rationale of decisions made in your modeling, supported by empirical evidence. Use the insights generated from your EDA step to guide your modeling approach.
  - Clear explanations of all steps used to arrive at a final model, with conclusions that summarize results with respect to the question(s) being asked and key takeaways from the analysis.

- For mathematical formulae, type them in your R markdown file. Do not e.g. write them on a piece of paper, snap a photo, and use the image file.
- Incorrectly following submission instructions results in deduction of grades
- Students are expected to act with regard to UC Berkeley Academic Integrity

```
library(foreign)
library(gplots)
```

```
## Warning: package 'gplots' was built under R version 4.0.2
```

```
library(ggplot2)
library(stats)
library(Hmisc)
library(car)
library(usmap)
```

```
## Warning: package 'usmap' was built under R version 4.0.2
```

```
library(dplyr)
library(gridExtra)
library(stargazer)
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.0.2
```

```
library(data.table)
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
library(grid)
library(plm)
```

```
## Warning: package 'plm' was built under R version 4.0.2
```

```
#library(Rmisc) # for arranging plots
```

U.S. traffic fatalities: 1980-2004

In this lab, you are asked to answer the question **"Do changes in traffic laws affect traffic fatalities?"** To do so, you will conduct the tasks specified below using the data set *driving.Rdata*, which 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 come with the dataste.

Exercises:

Part 1

1. (30%) Load the data. Provide a description of the basic structure of the dataset, as we have done throughout the semester. Conduct a very thorough EDA, which should include

both graphical and tabular techniques, on the dataset, including both the dependent variable *totfatrte* and the potential explanatory variables. You need to write a detailed narrative of your observations of your EDA. *Reminder: giving an "output dump" (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive a zero in this exercise.*

This dataset is at an annual level - one row per state, per year for the 48 contiiguous states from 1980 to 2004. There are no missing values in the dataset.

```
load("driving.RData")

# one row per year per state
head(table(data$year, data$state))
```

```
##
##        1 3 4 5 6 7 8 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
##   1980 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1981 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1982 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1983 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1984 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1985 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##
##        30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
##   1980  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1981  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1982  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1983  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1984  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1985  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
```

```
min(data$year)
```

```
## [1] 1980
```

```
max(data$year)
```

```
## [1] 2004
```

```
# merge with state codes
fips_map <-read.csv("statecodes.csv")
d_data <- merge(x=data, y=fips_map, by="state",all.x = TRUE) %>% dplyr::select(-state)
d_data <- dplyr::rename(d_data, c("state"="code"))

d_data <- data.table(d_data)

# Zero missing values across all columns
d_data[, lapply(.SD, function(x) sum(is.na(x))), .SDcols = 1:56]
```

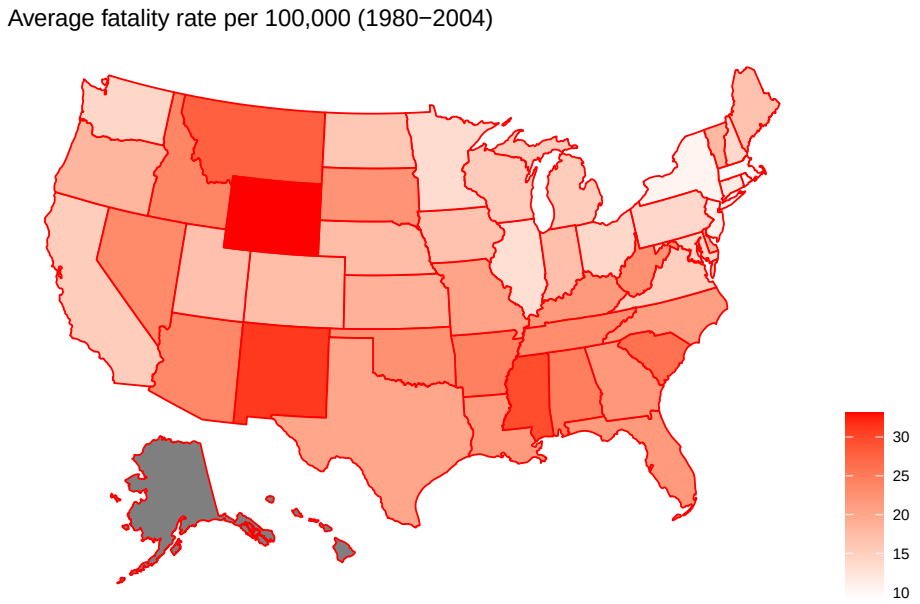
```
##    year sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl bac10 bac08
## 1:    0    0    0    0    0      0        0      0       0   0     0     0
```

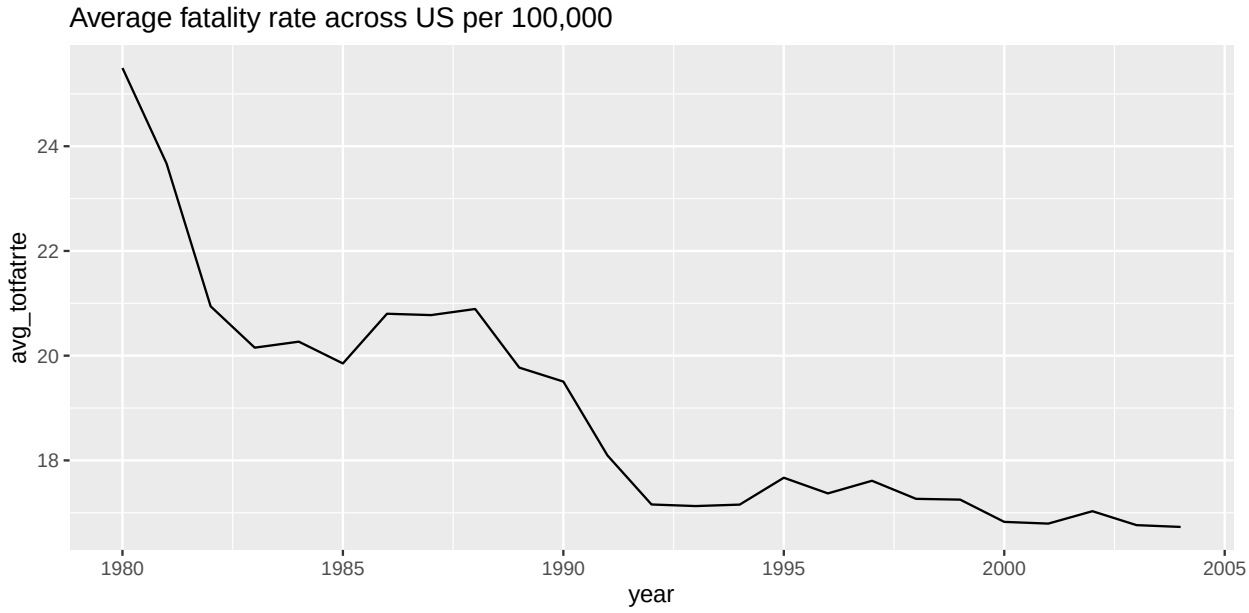
```
##    perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm statepop
## 1:     0      0       0       0         0          0          0        0
##    totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24 sl70plus sbprim
## 1:         0          0          0          0    0         0        0      0
##    sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91 d92 d93 d94 d95 d96
## 1:       0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##    d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc state
## 1:   0   0   0   0   0   0   0   0            0     0
```

```
# average fatality rate per 100,000 across states
state_avg <- d_data %>% group_by(state) %>% summarise(avg_totfatrte=mean(totfatrte), .groups =

state_avg <- dplyr::rename(state_avg, c("value"="avg_totfatrte"))

p1.1 <- plot_usmap(data = state_avg,  values="value", color = "red") +
  scale_fill_continuous(name="", low="white", high="red") +
  theme(legend.position = "right") +ggtitle("Average fatality rate per 100,000 (1980-2004)")

# average fatality rate per 100,000 across years
year_avg <- d_data %>% group_by(year) %>% summarise(avg_totfatrte=mean(totfatrte))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
p1.2 <- ggplot(data=year_avg, aes(x=year, y=avg_totfatrte)) +
  geom_line()+ ggtitle("Average fatality rate across US per 100,000")

grid.arrange(p1.1, p1.2, nrow=2)
```

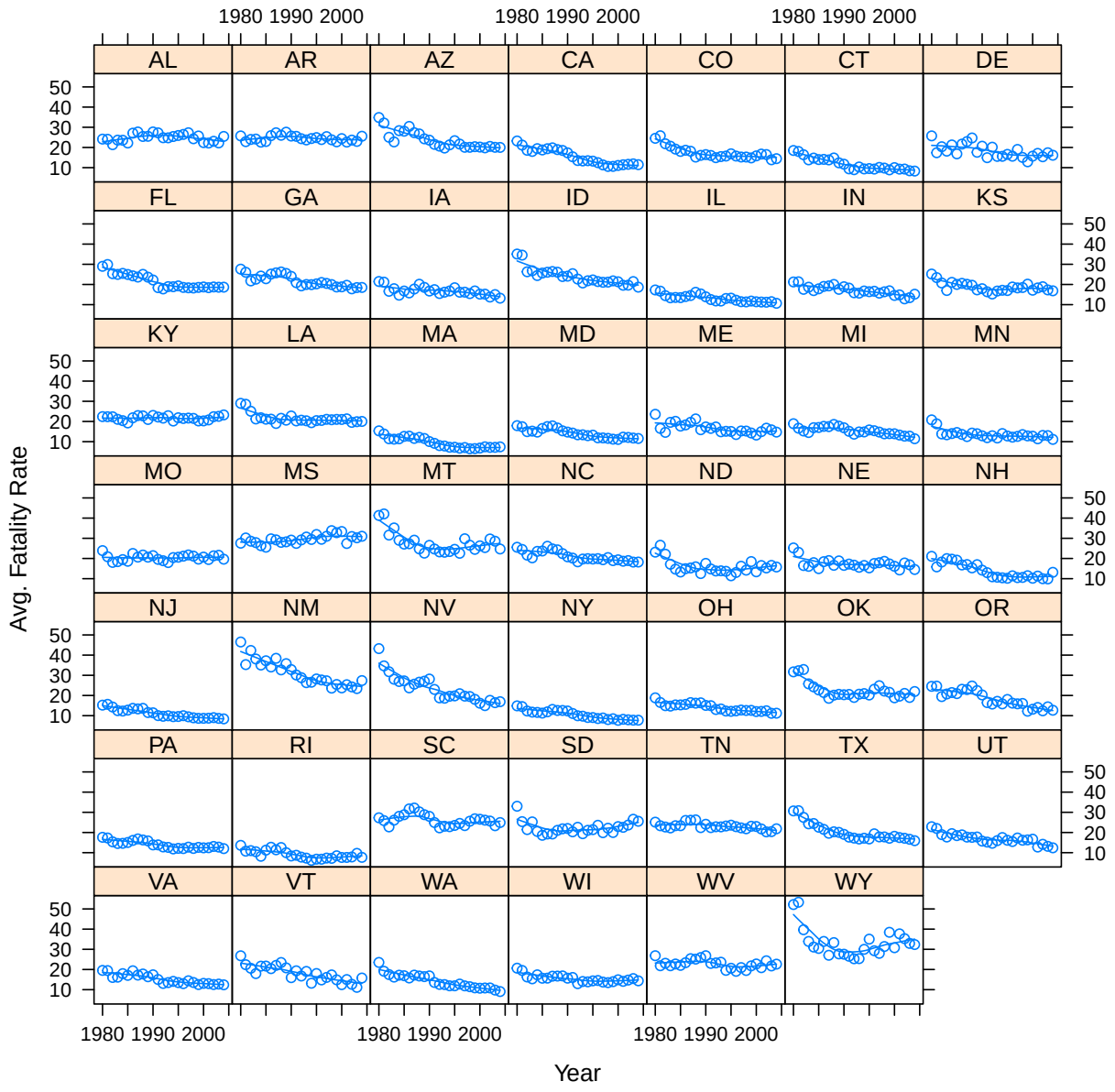




Growth Gurve Analysis

- Note general flat to downward trend with exception of Mississippi
- Nevada and New Mexico drop looks steep

```
xyplot(totfatrte~year | state, data=d_data,
       prepanel = function(x, y) prepanel.loess(x, y, family="gaussian"),
       xlab = "Year", ylab = "Avg. Fatality Rate",
       panel = function(x, y) {
          panel.xyplot(x, y)
          panel.loess(x,y, family="gaussian") },
       as.table=T)
```




```
# this is hard to read!
#g <- ggplot(data_state, aes(year, totfatrte, colour = as.factor(code)))
#g + geom_line() + ggtitle("Growth Curve by state")
```

Investigation of laws that changed over time

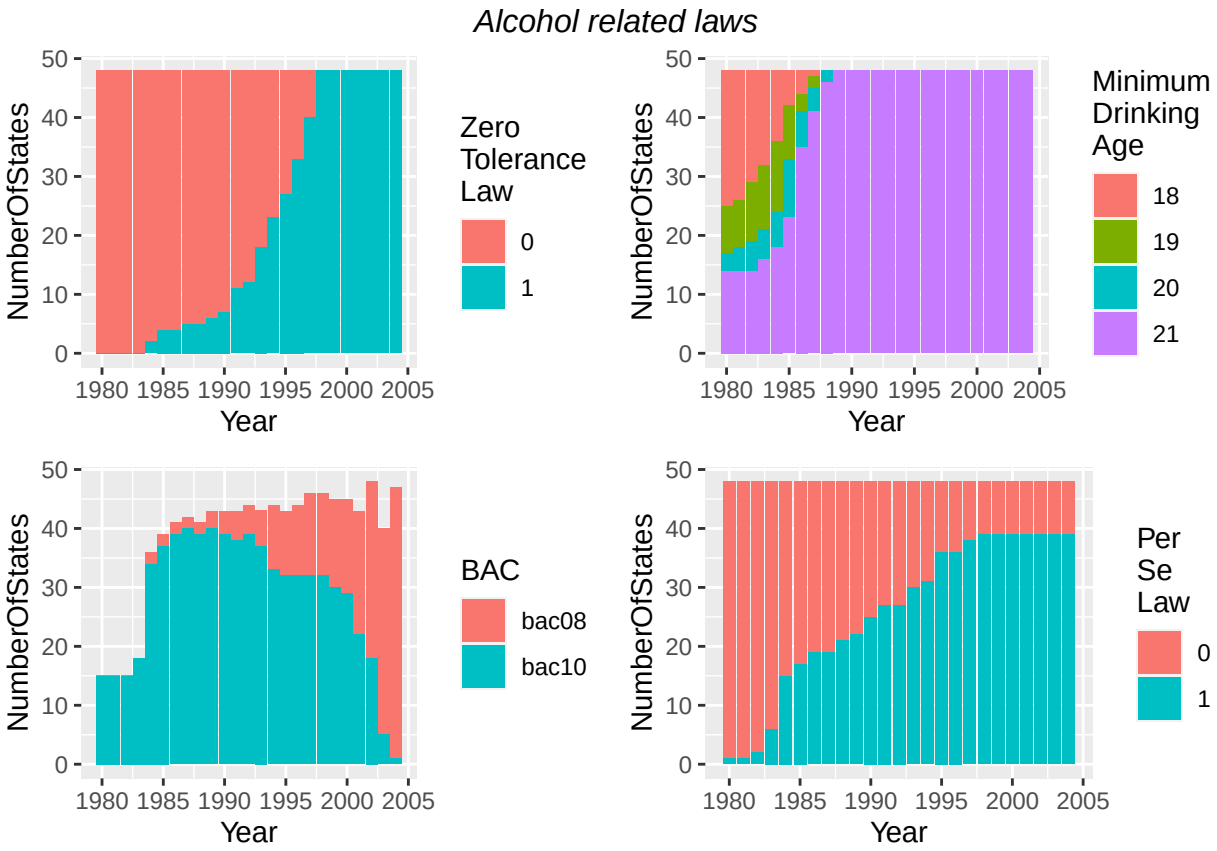
We note that the laws are not binary Variables. *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl and perc14_24* are yes-no indicator dummies, with the caveat that they can be fractional as noted in the problem statement. The fractional values will ideally need to be changed to 0 or 1 for *plotting purposes only* based on whether the variable is < 0.5 or $>= 0.5$. Note that we need to do special handling of the edge case of 0.5 in two categories. For the regression models, we use the original data as is since we do not see any value in dropping the information contained.

```
## generic plot function for stacked charts
genplot <- function(df, col, legend){
  df2 <- table(df[[col]], df$year) %>% reshape2::melt()
  colnames(df2) <- c(col, 'Year', 'NumberOfStates')
  df2.plot <- ggplot(df2, aes(fill=factor(get(col)), y=NumberOfStates, x=Year))+
                geom_bar(position="stack", stat = "identity") +
              guides(fill=guide_legend(title=legend))
#               scale_fill_discrete(palette=scales::hue_pal())


  return (df2.plot)
}
```

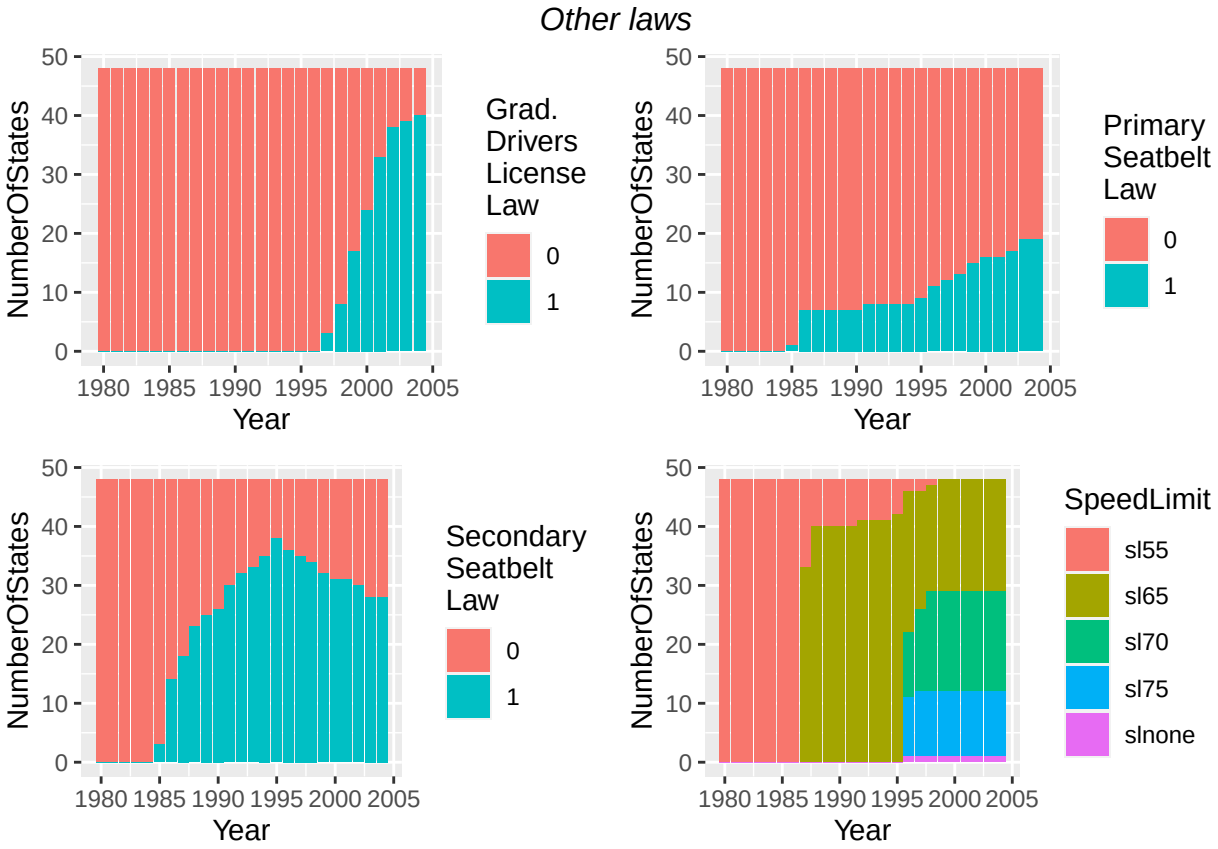
The stacked bar charts below show how the laws changed over time across the states. * We note that the alcohol related laws get more conservative over the years. Only about 15 states had *blood alcohol limit* laws in 1980. We see that proportion increase to about 40 states in the mid 80's with some states beginning to adopt the stricter limit of BAC08. More states also adopt the *Per Se law* that makes violating the BAC limit while driving on its own an offence. * We note the variation in the *minimum drinking age* among states in the early 80s with different states having different age limits. However by the late 80s all states conform to the minimum drinking age of 21. The *Zero tolerance law* that specifically targets youth drinking starts to get introduced in the mid 80s and we see a sharp increase in the number of states adopting the law over the next decade to full compliance. * The *Graduated driver license* law that alows young drivers to gain safe driving experience gets introduced in the early 90s with more states adopting it over the next decade. * We see a < 10 states beginning to adopt the *primary seatbelt law* in the mid 80s and there seems to be a gradual increase to about 20 states in the mid 2000's. The stricter secondary seatbelt law has a much more steep increase and then a gradual decline. Since the interpretation of the law in different states are different its a bit hard to determine the drop in adoption in the mid 90s. * We see a general trend of *Speed limits* increasing over time across the US starting in the late 80s. This is possibly due to the cars in general getting faster and roads getting better.

```
# Speed Limits

d_data[sl55==0.5 & sl65==0.5, c("sl55", "sl65"):=list(0,1) ]
d_data[sl65==0.5 & sl70==0.5, c("sl65", "sl70"):=list(0,1) ]
d_data[sl65==0.5 & sl75==0.5, c("sl65", "sl75"):=list(0,1) ]

for (sp in c("sl55","sl65","sl70","sl75", "slnone")){
  d_data[get(sp) >0.5, eval(quote(sp)):= 1]
  d_data[get(sp) <0.5, eval(quote(sp)):= 0]
}

d_data.speed <- d_data %>%
  gather(key="SpeedLimit", value="Value", "sl55","sl65","sl70","sl75", "slnone") %>%
  dplyr::filter(Value==1) %>%
  dplyr::select(-Value) %>% data.table()

d_data.speed[, SpeedLimit:=factor(SpeedLimit)]
```

```r
speed.plot <- genplot(d_data.speed, "SpeedLimit", "SpeedLimit")

# merge bac10 and bac08 to one
d_data[bac10==0.5 & bac08==0.5, c("bac10", "bac08"):=list(0,1) ]

d_data.bac <- d_data %>%
  gather(key="BAC", value="Value", "bac10","bac08") %>%
  dplyr::filter(Value==1) %>%
  dplyr::select(-Value) %>% data.table()

d_data.bac$BAC <- factor(d_data.bac$BAC)


bac.plot <- genplot(d_data.bac, "BAC", "BAC")


for (sp in c("zerotol","gdl", "perse", "sbprim", "sbsecon", "sl70plus")){
  d_data[get(sp) >0.5, eval(quote(sp)):= as.integer(1)]
  d_data[get(sp) <=0.5, eval(quote(sp)):= as.integer(0)]

}

# minage needs to be rounded
d_data[,minage:=round(minage)]

minage.plot <- genplot(d_data, "minage", "Minimum\nDrinking\nAge")
zerotol.plot <- genplot(d_data, "zerotol", "Zero\nTolerance\nLaw")
bac.plot <- genplot(d_data.bac, "BAC", "BAC")
perse.plot <- genplot(d_data, "perse", "Per\nSe\nLaw")

grid.arrange(zerotol.plot, minage.plot, bac.plot,perse.plot, top=textGrob("Alcohol related laws
```




```
# License laws
gdl.plot <- genplot(d_data, "gdl", "Grad.\nDrivers\nLicense\nLaw")

# seatbelt laws
sbprim.plot <- genplot(d_data, "sbprim", "Primary\nSeatbelt\nLaw")
sbsecon.plot <- genplot(d_data, "sbsecon",  "Secondary\nSeatbelt\nLaw")

#grid.arrange(gdl.plot, sbprim.plot, sbsecon.plot)
grid.arrange(gdl.plot, sbprim.plot, sbsecon.plot,speed.plot, top=textGrob("Other laws",gp=gpar
```



Part 2

2. (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.

What is the average of this variable in each of the years in the time period covered in this dataset

```
# yearly average nationwide
year_avg
```

```
## # A tibble: 25 x 2
##     year avg_totfatrte
##    <int>         <dbl>
##  1  1980          25.5
##  2  1981          23.7
##  3  1982          20.9
##  4  1983          20.2
##  5  1984          20.3
```

```
##  6  1985          19.9
##  7  1986          20.8
##  8  1987          20.8
##  9  1988          20.9
## 10  1989          19.8
## # ... with 15 more rows
```

Regression model and explanation

This model gives us the time effect on fatality rate. The intercept in this case is the average *totfatrte* across all states in the omitted year 2004. Each of the coefficients `d80`, `d81`…`d04` is the average increase in *totfatrte* relative to the base year 2004.

```
#stargazer(lm(totfatrte~d80+d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+d91+d92+d93+d94+d95+d96+d9

stargazer(lm(totfatrte~factor(year), data=d_data), type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                               totfatrte
## ------------------------------------------------
## factor(year)1981               -1.824
##                                (1.226)
##
## factor(year)1982              -4.552***
##                                (1.226)
##
## factor(year)1983              -5.342***
##                                (1.226)
##
## factor(year)1984              -5.227***
##                                (1.226)
##
## factor(year)1985              -5.643***
##                                (1.226)
##
## factor(year)1986              -4.694***
##                                (1.226)
##
## factor(year)1987              -4.720***
##                                (1.226)
##
## factor(year)1988              -4.603***
##                                (1.226)
##
## factor(year)1989              -5.722***
##                                (1.226)
```

```
##
## factor(year)1990            -5.989***
##                              (1.226)
##
## factor(year)1991            -7.400***
##                              (1.226)
##
## factor(year)1992            -8.337***
##                              (1.226)
##
## factor(year)1993            -8.367***
##                              (1.226)
##
## factor(year)1994            -8.339***
##                              (1.226)
##
## factor(year)1995            -7.826***
##                              (1.226)
##
## factor(year)1996            -8.125***
##                              (1.226)
##
## factor(year)1997            -7.884***
##                              (1.226)
##
## factor(year)1998            -8.229***
##                              (1.226)
##
## factor(year)1999            -8.244***
##                              (1.226)
##
## factor(year)2000            -8.669***
##                              (1.226)
##
## factor(year)2001            -8.702***
##                              (1.226)
##
## factor(year)2002            -8.465***
##                              (1.226)
##
## factor(year)2003            -8.731***
##                              (1.226)
##
## factor(year)2004            -8.766***
##                              (1.226)
##
## Constant                    25.495***
##                              (0.867)
```

```
##
## ------------------------------------------------
## Observations                      1,200
## R2                                0.128
## Adjusted R2                       0.110
## Residual Std. Error         6.008 (df = 1175)
## F Statistic             7.164*** (df = 24; 1175)
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

Did driving become safer

This model highlights that driving has become safer between the years 1982 through 2004 with respect to the base year 1980 since the differences are significant at the < 0.05.

Part 3

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.)

As seen earlier, variables `bac08`, `bac10`, `perse`, `sbprim`, `sbsecon`, `sl70plus`, `gdl` are fractional values between 0 and 1. We use the original values as is so as not to lose any information. Variables `unem`, `perc14_24` and `vehiclesmilespc` are continuous. We could log transform `vehiclesmilespc` only if we want to interpret the coefficients in terms of percentage changes.

```
model.3 <- lm(totfatrte~bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+unem+vehicmiles

stargazer(model.3, type="text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                              totfatrte
## ------------------------------------------------
## bac08                        -2.498***
##                               (0.538)
##
## bac10                        -1.418***
##                               (0.396)
##
## perse                        -0.620**
##                               (0.298)
```

```
##
## sbprim                          -0.075
##                                 (0.491)
##
## sbsecon                          0.067
##                                 (0.429)
##
## sl70plus                        3.348***
##                                 (0.445)
##
## gdl                             -0.427
##                                 (0.527)
##
## perc14_24                        0.142
##                                 (0.123)
##
## unem                            0.757***
##                                 (0.078)
##
## vehicmilespc                    0.003***
##                                 (0.0001)
##
## factor(year)1981                -2.175***
##                                 (0.828)
##
## factor(year)1982                -6.596***
##                                 (0.853)
##
## factor(year)1983                -7.397***
##                                 (0.869)
##
## factor(year)1984                -5.850***
##                                 (0.876)
##
## factor(year)1985                -6.483***
##                                 (0.895)
##
## factor(year)1986                -5.853***
##                                 (0.931)
##
## factor(year)1987                -6.367***
##                                 (0.967)
##
## factor(year)1988                -6.592***
##                                 (1.014)
##
## factor(year)1989                -8.071***
##                                 (1.053)
```

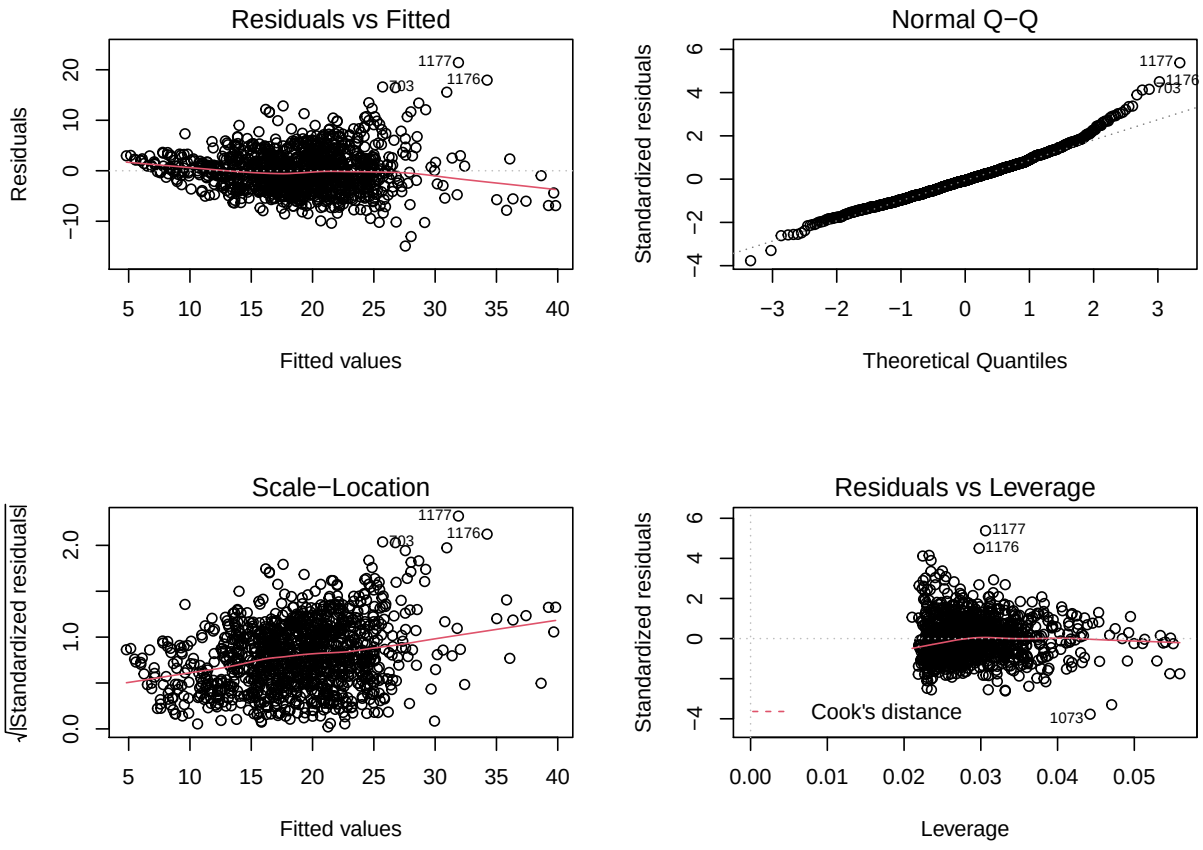
```
##
## factor(year)1990             -8.959***
##                               (1.077)
##
## factor(year)1991            -11.069***
##                               (1.101)
##
## factor(year)1992            -12.878***
##                               (1.123)
##
## factor(year)1993            -12.731***
##                               (1.136)
##
## factor(year)1994            -12.365***
##                               (1.157)
##
## factor(year)1995            -11.953***
##                               (1.184)
##
## factor(year)1996            -13.876***
##                               (1.223)
##
## factor(year)1997            -14.258***
##                               (1.250)
##
## factor(year)1998            -15.042***
##                               (1.265)
##
## factor(year)1999            -15.091***
##                               (1.284)
##
## factor(year)2000            -15.444***
##                               (1.305)
##
## factor(year)2001            -16.184***
##                               (1.334)
##
## factor(year)2002            -16.724***
##                               (1.348)
##
## factor(year)2003            -17.021***
##                               (1.359)
##
## factor(year)2004            -16.711***
##                               (1.387)
##
## Constant                      -2.716
##                               (2.476)
```

```
##
## ------------------------------------------------
## Observations                      1,200
## R2                                0.608
## Adjusted R2                       0.596
## Residual Std. Error          4.046 (df = 1165)
## F Statistic             53.096*** (df = 34; 1165)
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

Check model assumptions

We see some evidence of heteroskedasticity pointing to omitted variables.

```
par(mfrow=c(2,2))
plot(model.3)
```



Do *per se laws* have a negative effect on the fatality rate? From the model output above, it is evident that per se laws have had a negative effect on the fatality rate, as seen from the coefficient of -0.756 and p value < 0.05

What about having a primary seat belt law?

We do not see evidence of the primary seat belt law having any effect on the fatality rate as seen from the p value. This seems suspicous and points to the limitations inherent in pooled OLS models.

Part 4

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 Pooled OLS model from Part 3 does not consider heterogeniety across each state. Each state has state specific factors that influence the fatality rate over time and the Pooled OLS model does not differentiate between the state specific differences over time and therefore treats all the observations the same way. Therefore the fixed effects model is more reliable for this inference. The output of the model produces different estimates from the pooled OLS model. We note that the model correctly shows the statistically significant negative effect of `sbprim` on the dependent variable.

```
d_data2 <- pdata.frame(d_data, c("state", "year"))
model.4 <-
  plm(
    totfatrte ~ bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+unem+vehicmilespc+facto
    model = "within",
    data = d_data2
  )

stargazer(model.4, type="text")
```

```
##
## ===============================================
##                       Dependent variable:
##                   ---------------------------
##                              totfatrte
## -----------------------------------------------
## bac08                       -1.377***
##                              (0.387)
##
## bac10                       -1.053***
##                              (0.269)
##
## perse                       -1.077***
##                              (0.224)
##
## sbprim                      -1.225***
##                              (0.343)
##
```

```
## sbsecon                    -0.350
##                            (0.252)
##
## sl70plus                   -0.058
##                            (0.261)
##
## gdl                        -0.331
##                            (0.281)
##
## perc14_24                  0.197**
##                            (0.095)
##
## unem                      -0.576***
##                            (0.061)
##
## vehicmilespc              0.001***
##                           (0.0001)
##
## factor(year)1981          -1.512***
##                            (0.414)
##
## factor(year)1982          -3.038***
##                            (0.443)
##
## factor(year)1983          -3.566***
##                            (0.457)
##
## factor(year)1984          -4.296***
##                            (0.465)
##
## factor(year)1985          -4.753***
##                            (0.486)
##
## factor(year)1986          -3.677***
##                            (0.518)
##
## factor(year)1987          -4.322***
##                            (0.556)
##
## factor(year)1988          -4.799***
##                            (0.602)
##
## factor(year)1989          -6.152***
##                            (0.641)
##
## factor(year)1990          -6.271***
##                            (0.666)
##
```

```
## factor(year)1991          -6.934***
##                            (0.683)
##
## factor(year)1992          -7.805***
##                            (0.704)
##
## factor(year)1993          -8.125***
##                            (0.717)
##
## factor(year)1994          -8.572***
##                            (0.735)
##
## factor(year)1995          -8.302***
##                            (0.757)
##
## factor(year)1996          -8.681***
##                            (0.798)
##
## factor(year)1997          -8.766***
##                            (0.818)
##
## factor(year)1998          -9.431***
##                            (0.833)
##
## factor(year)1999          -9.588***
##                            (0.841)
##
## factor(year)2000         -10.115***
##                            (0.852)
##
## factor(year)2001          -9.762***
##                            (0.868)
##
## factor(year)2002          -9.036***
##                            (0.879)
##
## factor(year)2003          -9.071***
##                            (0.887)
##
## factor(year)2004          -9.498***
##                            (0.906)
##
## ------------------------------------------
## Observations                1,200
## R2                          0.625
## Adjusted R2                 0.598
## F Statistic       54.807*** (df = 34; 1118)
## ==========================================
```

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

Model assumtions

Our *Pooled OLS model* has the form. Here $y_{81}...y_{04}$ are dummy variables that represent each year. a_i and u_{it} together form the composite error term. a_i is the unobserved state effect. This represents all factors that affecting fatality rates that do not change over time, such as population demographics which are slow to change. u_{it} is the time varying error. One of the assumptions of the pooled OLS model is that the composite error term is uncorrelated with the explanatory variables, which does not make sense in this case since there could be state specific factors that effect the fatality rate. Pooled OLS is therefore biased and inconsistent if the explanatory variables bac_{08}, bac_{10}... $\beta_2 vehicmilespc$ are correlated to a_i

$$\begin{aligned} totfatrte = \beta_o + \beta_1 bac08 + \beta_2 bac10 + \beta_3 perse + \beta_4 sbprim + \\ \beta_5 sbsecon + \beta_6 sl70plus + \beta_7 gdl + \beta_2 8perc14\_24 + \\ \beta_9 unem + \beta_1 0vehicmilespc + \\ \delta_o y_{81} + \delta_1 y_{82} + ... + \delta_{23} y_{04} + a_i + u_{it}, \\ t = 82, 83...04 \end{aligned}$$

The *Fixed effect model* has the form seen below. Here we see that the state specific unobserved effect a_i has disappeared. Therefore we note that the fixed effects model allows for arbitrary correlation between a_i and all the explanatory variables since any explanatory variables that is constant over time gets swept away by the fixed effects transformation. The other assumptions for this fixed effects model is that errors u_{it} are homoskedastic and serially uncorrelated across t

$$\begin{aligned} totfatrte_{it} - \hat{totfatrte}_{it} = \beta_1(bac08 - \hat{bac08}) + ... + \beta_2(vehicmilespc - \hat{vehicmilespc}) + u_{it}, \\ t = 82, 83...04 \end{aligned}$$

Part 5

5. (10%) Would you perfer to use a random effects model instead of the fixed effects model you built in *Exercise 4*? Please explain.

The null hypothesis of the Hausman test is the Random effects is preferred, the alternate hypothesis is that the fixed effects model is preferred. The below result suggests that the fixed effect model is preferred.

```
model.re <-   plm(
    totfatrte ~ bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+unem+vehicmilespc+facto
    index = c("state"),
    model = "random",
    data = d_data
  )
phtest(model.4, model.re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +  ...
## chisq = 151.57, df = 34, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Part 6

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.

The estimated effect of 1 mile increase per capita is 0.001 increase in `totfatrte`. Therfore a 1,000 mile increase per capita would result in a corresponding 1% increase in fatality rate. The 95% confidence interval of the estimate is $1.0 \pm 1.96 * (1000 * 0.0001)$ which equals (0.804,1.196)

#Part 7

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?