ES218 Project: US Vehicular Accidents

Curtis Zhuang

5/4/2020

Table of Contents

[Introduction 2](#_Toc40540497)

[Methods 2](#_Toc40540498)

[Results 3](#_Toc40540499)

[Discussion 15](#_Toc40540500)

[Reference: 16](#_Toc40540501)

# Install packages  
library(gridExtra)  
library(forcats)  
library(tidyr)  
library(ggplot2)  
library(dplyr)

# Want to test if sex, car type,   
dat <- readRDS('../Data/farsp.RDS')  
  
# group the states into four main regions in U.S  
  
# Regroup and create region variable  
west <- c(2,4,6,8,15,16,32,35,41,49,53,56 )  
midwe <- c(17,18,10,20,26,27,30,31,38,39,46)  
south <- c(1,5,11,12,13,21,22,24,28,29,37,40,45,47,48,51,54)  
northea <- c(9,10,23,25,33,34,36,42,44,50,55 )  
oth <- c(43,52)  
  
# Regroup car type  
  
dat <- dat %>%   
 mutate('Region' = case\_when(state %in% west ~ 'West',  
 state %in% midwe ~ 'MidWest',  
 state %in% south ~ 'South',  
 state %in% northea ~ 'NorthEast',  
 state %in% oth ~ 'Other'))  
  
dat\_full <- dat %>%   
 select(-c(day, hour, minute, county)) %>%   
 # filter out severity  
 filter(inj\_sev %in% c(1,2,3,4,5,6))%>%   
 # filter out Virgin Islands and Puerto Rico  
 drop\_na(Region)

### Introduction

This study is conducted in light of the dataset recorded by the NHTSA for the years 1996 through 2016 on vehicular accidents. This is a large dataset as it contains many dimensions of an accident from the time the accident happens, the age of the driver caught in the accident, the location the accident, type of vehicle, etc. However, for this study, I only focused on the following:

1. Are there regional differences in traffic accidents?
2. Is the severity of the accidents affected by some of the variables that are included in this dataset?

U.S. is a large country with diverse cultural backgrounds. I suspect that people living in different parts of the U.S. will have differetn driving habits. Thus, in order to study regional differences, I grouped the states into four main categories: West, Midwest, Northeast and South (which is the most common way of grouping states). Among all four regions, the South account for 38.3% of the whole population, followed by the West, Midwest and Northeast which accounts for 23.9%, 20.8%, 17.1% respectively (census.gov). From there I created graphs to compare some of the selected variables between the four groups.

To study the second question, I used bivariate models to analyze if the severity of the accidents is related to the variables of my choice, namely: age, region, sex, collision type and people involved in the accident. From my study, I find that all the variables I choose are significant at 5% level when it comes to severity.

The paper is structured as follows: the Methods section will touch briefly on what kind of graphs I made and what models I used; the Results section will focus on explaining the results shown by the graphs as well as interpreting the results from the models; the Discussion section will dig deeper into the findings of the results section.

### Methods

As discussed above, I came up with the most common way of grouping states and come out with four groups. I excluded Puerto Rico and Virgin Island due to the scarcity of samples they have compared with other groups and their nature of being small islands which may cause issues. The original dataset has 1906184 entries.

I performed piping operations and removed the samples from Puerto Rican and Virgin Island after I finished grouping. Another group that I discarded in this study is the data entries reported as “Unknown”/ “Not reported” which can be found for both the sex variable and the collision type variable. In total, they do not account for a large number and getting rid of these variables can help reduce the chance of having outliers. There are 1467635 samples left after these operations.

To study how the vehicular accidents can be affected by its location, I used univariate methods for some of the categorical variables and used bivariate methods for the quantitative variables.

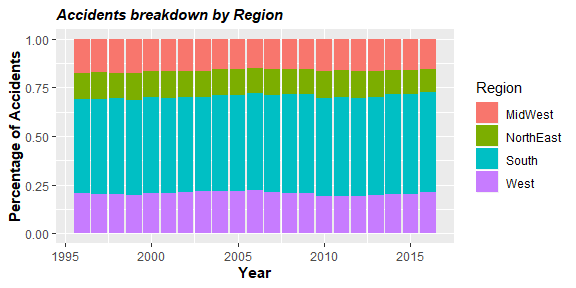
Using graph as guidance, I created two models. The first model has severity as the dependent variable, region and age as the independent variables. This model captures mainly how region can affect severity of the accident when age is controlled. The next model expands the scope of variables and studies more generally what variables can affect severity.

### Results

The results section looks at each variable following the order of: Region, Sex, Collision, Age, Number of ppl involved, and Severity.

#### Region

# Region by year  
dat\_reg <- dat\_full %>%  
 group\_by(year, Region) %>%   
 summarise(n = n()) %>%   
 ungroup() %>%   
 group\_by(year) %>%   
 mutate(percent = n/sum(n))%>%   
 drop\_na(Region)  
  
ggplot(dat\_reg, aes(year, percent, fill = Region)) +   
 geom\_bar(position = 'stack', stat = 'identity') +  
 xlab('Year') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by Region') +  
 theme(  
 plot.title = element\_text(color="black", size=11, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=11, face="bold"),  
 axis.title.y = element\_text(color="black", size=11, face="bold"))

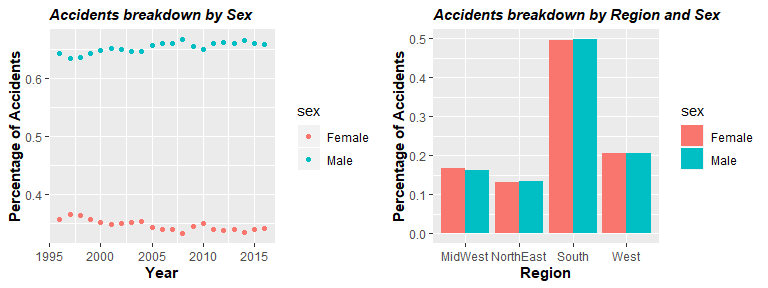


Based on the bar chart showing distribution of accidents across the four regions, we can observe that the percentage of accidents for each region did not fluctuate a lot over the course of 22 years. Comparing the four regions, we see that the South has the most accidents which accounted for nearly 50%.

This makes intuitive sense as geographically, we can see that the South covers lots of lands and includes states like Texas, Maryland, Florida which are all populated states. The large portion of accidents happening in the south will affect a lot of the results we will get from the other graphs.

#### Sex

# Make a graph of how accidents happen related to sex  
# male vs female by year  
dat\_sex <- dat\_full %>%  
 group\_by(year, sex) %>%   
 summarise(n = n()) %>%   
 filter(sex %in% c(1,2)) %>%   
 mutate(sex = ifelse(sex == 1, "Male", "Female")) %>%   
 ungroup() %>%   
 group\_by(year) %>%   
 mutate(percent = n/sum(n))  
  
# plot the accident by sex  
ggplot(dat\_sex, aes(year, percent, col = sex)) + geom\_point() +  
 xlab('Year') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by Sex') +  
 theme(  
 plot.title = element\_text(color="black", size=11, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=11, face="bold"),  
 axis.title.y = element\_text(color="black", size=11, face="bold")) ->p1  
  
# male vs female by region  
dat\_sex\_loc <- dat\_full %>%   
 group\_by(Region, sex) %>%   
 summarise(n = n()) %>%   
 filter(sex %in% c(1,2)) %>%   
 mutate(sex = ifelse(sex == 1, "Male", "Female")) %>%   
 ungroup() %>%   
 group\_by(sex) %>%   
 mutate(percent = n/sum(n))  
  
# univariate  
ggplot(dat\_sex\_loc, aes(Region, percent, fill = sex)) +   
 geom\_bar(position = 'dodge', stat = 'identity') +  
 xlab('Region') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by Region and Sex') +  
 theme(  
 plot.title = element\_text(color="black", size=11, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=11, face="bold"),  
 axis.title.y = element\_text(color="black", size=11, face="bold")) -> p2  
  
grid.arrange(p1, p2, nrow = 1)

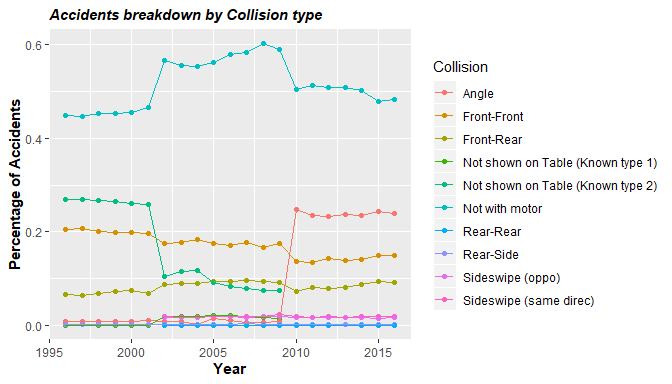


Here we consider how sex of the driver can contribute to the occurence of vehicular accident. From the graph on the left, we do not observe much fluctuation over the 20 years. About 65% of the accidents are caused by male drivers and about 35% of the accidents are caused by female drivers.

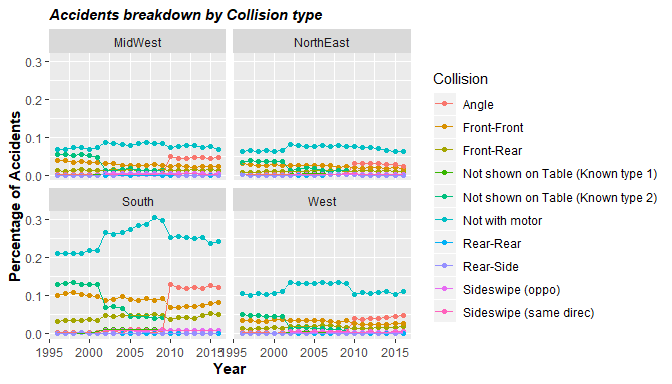
I then break down the accidents by sex across four regions to study regional difference on sex in terms of traffic accident. Here, when we observe the graph on the right, we see that for both male and female, the spread across four regions are almost identical: around 50% of the accidents happens in the South for both sex, followed by 21% in the West. Thus, sex of driver does not seem to be affected by where the accident happens and this pattern is also confirmed by the overall accidents distribution.

#### Collision

# collision by year  
  
dat\_col <- dat\_full %>%  
 group\_by(year, man\_coll) %>%   
 summarise(n = n()) %>%   
 mutate(Collision = case\_when(man\_coll == "0" ~ 'Not with motor',  
 man\_coll == "1" ~ 'Front-Rear',  
 man\_coll == "2" ~ 'Front-Front',  
 man\_coll == "3" ~ 'Not shown on Table (Known type 1)',  
 man\_coll == "4" ~ 'Not shown on Table (Known type 2)',  
 man\_coll == "6" ~ 'Angle',  
 man\_coll == "7" ~ 'Sideswipe (same direc)',  
 man\_coll == "8" ~ 'Sideswipe (oppo)',  
 man\_coll == "9" ~ 'Rear-Side',  
 man\_coll == "10" ~ 'Rear-Rear',  
 TRUE ~ 'Other')) %>%   
 ungroup() %>%   
 group\_by(year) %>%   
 filter(Collision != 'Other') %>%   
 mutate(percent = n/sum(n))  
  
ggplot(dat\_col, aes(year, percent, col = Collision)) + geom\_point() +  
 geom\_line() + xlab('Year') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by Collision type') +  
 theme(  
 plot.title = element\_text(color="black", size=11, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=11, face="bold"),  
 axis.title.y = element\_text(color="black", size=11, face="bold"))



# keep the Region variable  
dat\_col\_reg <- dat\_full %>%  
 group\_by(year, man\_coll, Region) %>%   
 summarise(n = n()) %>%   
 mutate(Collision = case\_when(man\_coll == "0" ~ 'Not with motor',  
 man\_coll == "1" ~ 'Front-Rear',  
 man\_coll == "2" ~ 'Front-Front',  
 man\_coll == "3" ~ 'Not shown on Table (Known type 1)',  
 man\_coll == "4" ~ 'Not shown on Table (Known type 2)',  
 man\_coll == "6" ~ 'Angle',  
 man\_coll == "7" ~ 'Sideswipe (same direc)',  
 man\_coll == "8" ~ 'Sideswipe (oppo)',  
 man\_coll == "9" ~ 'Rear-Side',  
 man\_coll == "10" ~ 'Rear-Rear',  
 TRUE ~ 'Other')) %>%   
 ungroup() %>%   
 group\_by(year) %>%   
 filter(Collision != 'Other') %>%   
 mutate(percent = n/sum(n))  
  
ggplot(dat\_col\_reg, aes(year, percent, col = Collision)) +   
 geom\_point() + geom\_line() + facet\_wrap( ~ Region) +  
 xlab('Year') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by Collision type') +  
 theme(  
 plot.title = element\_text(color="black", size=11, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=11, face="bold"),  
 axis.title.y = element\_text(color="black", size=11, face="bold"))

 The next variable that I considered in the study is the type of collision. The metadata for this variable is very incomplete as there is no explanation for type 3 and 4 which I think represents the collision between the front of one car and the side of the other and the collision between the front of one car and angle of anther. But without additional information, we cannot dig deeper into these two categories.

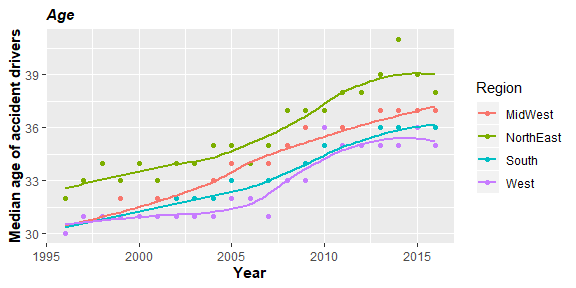
From the first graph, it is obvious that a majority of accidents happen between a vehicle and a non-vehicle (constantly over 40%) which is shown by the cyan line.

Another kind of collision that fluctuates as time increases is the growing number of accidents that happens at angle of the road (probably as a result of more roads) which is almost 0% in 1995 and experienced a sudden peak to over 20% starting in 2010. We noticed that the year it increased is also the time when Type 2 accidents vanished. So one explanation can be that collision type “Angle” replaces the original accident and that is why we can observe such an increase.

When we break down the collision by region, we observe that the accidents in the South are nearly identical to the overall accident patterns. As we have discussed previously, a large proportion of the accidents happens in the South and this makes a lot of sense here. For the other three regions, though we see accidents with Non-motor also appears to be them most frequent type, it does not diverge that much from other types of collision. Therefore, for collision type, we do observe a regional difference.

#### Age

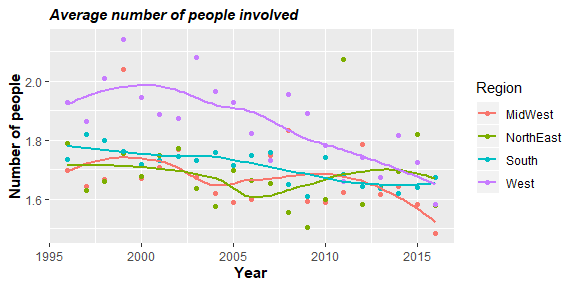
## AGE  
dat\_age <- dat\_full %>%  
 group\_by(year, Region) %>%   
 summarise(med = median(age))%>%   
 drop\_na(Region)  
  
ggplot(dat\_age, aes(year, med, col = Region)) + geom\_point() +  
 geom\_smooth(se = FALSE, method = 'loess') +  
 xlab('Year') + ylab('Median age of accident drivers') +   
 ggtitle('Age') +  
 theme(  
 plot.title = element\_text(color="black", size=11, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=11, face="bold"),  
 axis.title.y = element\_text(color="black", size=11, face="bold"))

 We then shift our focus to numerical variables. The first of which is the median age of people involved. By using a loess graph, we find that, interestingly, as time goes on, the median age of accident victims also increases. Over the 20 years, all four regions observe a increase in median age of victims. This is not surprising considering the increasing accessibility of cars and longer life expectancy.

Midwest shows the largest increase of 7 years from 30.5 to 37.5, while the West shows the smallest increase from 30.5 to 35. But across all four regions, Northeast remains the region with highest median across all regions. This might be the result of occupation as people in the Northeast may need to work in the office more compared with the other regions where people can work at home. It can also be a result of bad weather which may increase the need of Northwesterners to drive or it can be due to the demography. Still, we can conclude that there is regional difference when it comes to age of the drivers involved in accidents.

#### Number of person involved per accident

# Number of person vs year  
dat\_per <- dat\_full %>%  
 group\_by(year, Region) %>%   
 summarise(avg = mean(per\_no))  
  
ggplot(dat\_per, aes(year, avg, col = Region)) + geom\_point() +  
 geom\_smooth(se = FALSE, method = 'loess') +  
 xlab('Year') + ylab('Number of people') +   
 ggtitle('Average number of people involved') +  
 theme(  
 plot.title = element\_text(color="black", size=11, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=11, face="bold"),  
 axis.title.y = element\_text(color="black", size=11, face="bold"))

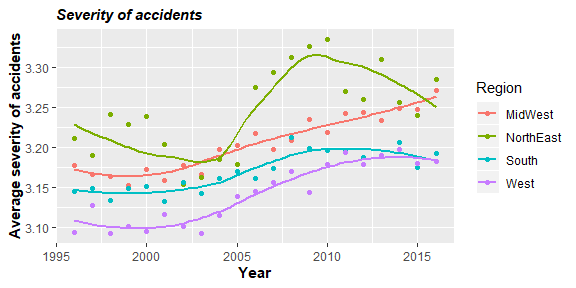


# # Create the bisquare function  
# wt.bisquare <- function(u, c=6) {  
# ifelse( abs(u/c) < 1, (1-(u/c)^2)^2, 0)  
# }  
#   
# lo <- loess(avg ~ year,dat\_per, span=1/3)  
#   
# wt <- rep(1,length(dat\_per$year))  
#   
# for(i in 1:10){  
# lo2 <- loess(avg ~ year,dat\_per, weights = wt, span=1/3)  
# wt <- wt.bisquare( lo2$res/ median(abs(lo2$res)), c=6 )  
# }  
#   
# # Plot the data  
# plot(avg ~ year,dat\_per, pch=16, col=rgb(0,0,0,0.2))  
#   
# # Add the robust loess  
# lines(dat\_per$year, predict(lo2), col="red")  
#   
# # Add the default loess  
# lines(dat\_per$year, predict(lo), col="grey50", lty=2)

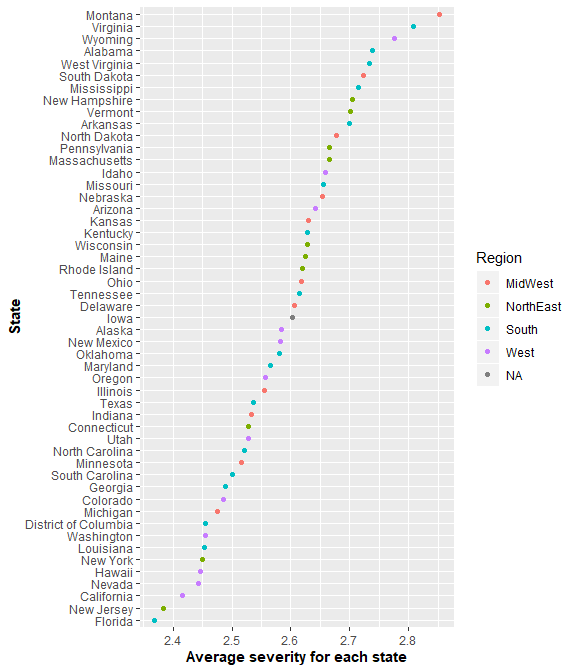
Contrary to the average age which shows a positive trend, we see a downturn of average number of people involved in accidents which is an exciting finding. This again can be explained by the accessibility of cars grows over time. However, when we look at the data for 2016, three regions(West, South, NorthEast) have very close results around 1.65 compared to the Midwest that averages around 1.45 people per accident. I have no findings regarding why Midwest has less people involved in the accident but it can be something to dig further into.

#### Severity

# Number of person vs year  
# severity by region  
dat\_sev <- dat\_full %>%   
 group\_by(year, Region) %>%   
 summarise(avg = mean(inj\_sev)) %>%   
 drop\_na(Region)  
  
ggplot(dat\_sev, aes(year, avg, col = Region)) + geom\_point() +  
 geom\_smooth(se = FALSE, method = 'loess') +  
 xlab('Year') + ylab('Average severity of accidents') +   
 ggtitle('Severity of accidents') +  
 theme(  
 plot.title = element\_text(color="black", size=11, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=11, face="bold"),  
 axis.title.y = element\_text(color="black", size=11, face="bold"))



# digging deeper into state  
  
state <- read.csv("../Data/state\_code.csv") %>%  
 rename(state = state\_code)  
  
# Create the names for states  
dat\_state <- left\_join(dat, state, by = 'state')  
  
dat\_states <- dat\_state %>%   
 group\_by(state\_name, Region) %>%   
 summarise(avrg\_sev = mean(inj\_sev)) %>%   
 ungroup()  
  
# create plot  
ggplot(dat\_states, aes(fct\_reorder(state\_name, avrg\_sev, .fun = median), avrg\_sev, col = Region)) +  
 geom\_point(stat = 'identity') + coord\_flip() +   
 xlab("State") + ylab("Average severity for each state") +  
 theme(  
 plot.title = element\_text(color="black", size=11, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=11, face="bold"),  
 axis.title.y = element\_text(color="black", size=11, face="bold"))



When it comes to severity, I treat the original categorical variable, which has 6 levels (after dropping unknowns), as a numeric variable. 1 stands for almost no injury and 6 stands for death for at least one party.

Thus, we can treat it as a quantitative variable without losing its original function. From our graph, we can observe an increase in severity for all Regions except NorthEast which shows a ‘S’ shape. I find no policy in the Northeast between 2005 and 2010 that can be held accountable for this trend , but there is certainly some factors that affect the severity of accidents. Overall, the severity of Northeast accidents also increases. And is an alarming result for us to consider.

I also breakdown the severity into the state level, we can see that the average severity, between states does not differ a lot even when we compare Montana with Florida, and there is sign of clustering of states from the same region and are evenly spread. Thus, it is not clear if regions impact severity. And to come to a clear conclusion, I study the possible relationship by using models.

lm\_sev\_age <- lm(inj\_sev ~ Region, dat = dat\_full)  
summary(lm\_sev\_age)

##   
## Call:  
## lm(formula = inj\_sev ~ Region, data = dat\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2392 -1.1362 0.7608 0.8328 2.8638   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.198786 0.002169 1474.51 <2e-16 \*\*\*  
## RegionNorthEast 0.040402 0.003238 12.48 <2e-16 \*\*\*  
## RegionSouth -0.031570 0.002500 -12.63 <2e-16 \*\*\*  
## RegionWest -0.062592 0.002903 -21.56 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.062 on 1467631 degrees of freedom  
## Multiple R-squared: 0.0008694, Adjusted R-squared: 0.0008673   
## F-statistic: 425.7 on 3 and 1467631 DF, p-value: < 2.2e-16

I used a simple linear regression between severity and regions to see if there is regional diffence for severity. I used simple linear regression as each state can only appear in one region, thus with the three indicator variables, I can observe the difference in severity between the regions and also see if it is significant.

The result above suggets that there is a significant difference between severity of the four regions as the p value for the three variables are significant at the 0.05 level. This is in favor of my hypothesis that there is regional difference. On average, Midwest’s severity is 3.2 over the whole period. Compared with Midwest, we see that the Northeast has more severity while the other two regions have less severe accidents on average.

Thus, we see that there is a regional difference in severity when I apply the single linear regression.

#### Extended model

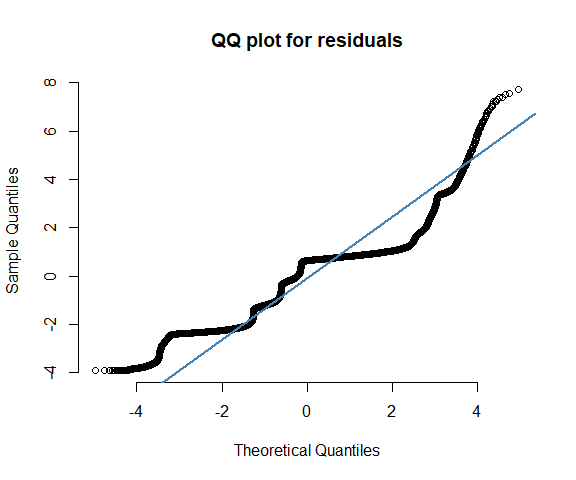
lm\_all <- lm(inj\_sev ~ age + Region + sex + man\_coll + per\_no, dat = dat\_full)  
summary(lm\_all)

##   
## Call:  
## lm(formula = inj\_sev ~ age + Region + sex + man\_coll + per\_no,   
## data = dat\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9192 -0.9583 0.6384 0.7587 7.7012   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.487e+00 3.130e-03 1113.944 <2e-16 \*\*\*  
## age 1.662e-03 1.953e-05 85.120 <2e-16 \*\*\*  
## RegionNorthEast 2.885e-02 3.153e-03 9.148 <2e-16 \*\*\*  
## RegionSouth -3.472e-02 2.435e-03 -14.263 <2e-16 \*\*\*  
## RegionWest -4.992e-02 2.828e-03 -17.649 <2e-16 \*\*\*  
## sex -1.018e-01 1.611e-03 -63.163 <2e-16 \*\*\*  
## man\_coll -2.721e-02 2.029e-04 -134.060 <2e-16 \*\*\*  
## per\_no -9.202e-02 4.253e-04 -216.373 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.034 on 1467627 degrees of freedom  
## Multiple R-squared: 0.05272, Adjusted R-squared: 0.05271   
## F-statistic: 1.167e+04 on 7 and 1467627 DF, p-value: < 2.2e-16

I then continue and included other variables (sex, age, collision type and number of people involved) into the model to see what other factors also contribute to the severity apart from the regions. The results are very satisfying as the variables that I included are all significant at 5% level.

Judging from the sign of the variables, we see that woman is less likely to involve in severe accidents, and the collisions with smaller index are more dangerous (e.g. traffic accidents involving a non-motor party). The number of person involved in accident is inversely related to severity, as we have discussed previously.

qqnorm(lm\_all$residuals, pch = 1, frame = FALSE, main = 'QQ plot for residuals')  
qqline(lm\_all$residuals, col = "steelblue", lwd = 2)



Though it may be important to check the residuals to see if the conditions for the linear model holds, especially even spread. From the qq-plot, we see that the shape of residuals does not follow a straight line. But as the severity is categorical in nature, we would still consider it to fit the line. More importantly, the fact that our model included all the data points, rather than samples of the vehicular accidents data, provide support for our model.

### Discussion

Overall, both the study on regional effect on traffic accident and the study on how other variables affect severity shows significant results.

Regional effect does show for some of the variables including collsion type, age and number of people involved in the accident. This effect can be boiled down to many aspects like weather, road condition, policies in effect and even cultural reasons.

The South, which makes up nearly half of the vehicular accidents, shows a strong impact on the overall data when studying the type of collisions. Considering the fact that the South only consists of around 38% of the total population, it is important for policy makers to consider why so many accidents happen in the South, especially accidents that involve one party that is not a motor.

In terms of severity, the result suggests that accidents in the Northeast are of the most severity followed by Midwest, South and the West. The second model suggests that severity is positively related to age and this can be a possible explanation of the trend we have observed as Northeast has the highest meidan age of drivers. While I consider age to be the most important factor, other factors including sex, number of people involved and the type of collision affect the severity as well.

I tried to identify what factors can lead to more detrimental accidents. Results show that an aged male driver who is passing a crossroad with only himself in the car driving in the Northeast will be the most likely to get involved in a severe vehicular accident than a light one. It is reasonable that fewer people in the car leads to more severe accident. As if there is one more person sitting in the car, that person is able to identify possible risks that other neglect. Thus, one more person can lead to likelihood of avoiding an accident.

The study finds female drivers will involve in lighter and less traffic accidents than men. Considering that fact that there are as many female drivers as male driver in the U.S. (statista), we find that compared with man, women drivers are less likely to get involved in a traffic accident from both the graph and our model. This is contradictary to our conventional wisdom that female drivers lead to accidents more (at least in China people tend to believe this).

Notice that even after including so many variables, our model only accounts for 5% of the total variability. It is important to understand what other factors leads to what outcome in order to prevent severe traffic accident, whether it is smarter regulation, stricter road tests or better road condition.

So this study provides insight of what affects severity of traffic accidents and how location can affect the accident. This is important for future decision making and accident prevention as we know under what circumstances the accidents may happen and how severe it can be. Still, this is just a very preliminary study and in order to diminish the accidents, more factors have to be included and the states need to learn from each other to find out what factors are contributing the road safety and what are not.

### Reference:

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.