Traffic Fatality Analysis

Saisha Jain, Otto Miller, Cindy Wang

2023-04-08

Section 1: Introduction

Description: In this paper we will be analyzing a dataset containing traffic fatality data from 1982 to 1988 in all US states except for Alaska and Hawaii. Our data set includes 336 cross-sectional observations of 4 categorical and 30 numerical variables. Each entry contains traffic fatality information from a particular year in a particular state. We hope our analysis will help to describe the population of all US traffic fatalities. We chose this dataset because it represents a consistent and severe problem for everybody including ourselves. We will use RMarkdown in RStudio(2021.09.0) to answer three research questions with data visualization and appropriate statistical methods at a 5% level of significance. In the online description for our dataset it states that: "Traffic fatality data is from the US Department of Transportation Fatal Accident Reporting System. Total vehicle miles traveled annually by state was obtained from the Department of Transportation. Personal income was obtained from the US Bureau of Economic Analysis, and the unemployment rate was obtained from the US Bureau of Labor Statistics". The download link for our data set is **https://vincentarelbundock.github.io/Rdatasets/csv/AER/Fatalities.csv**.

Missing Values: Only one of the 336 observations in our data contains missing values so we omitted the row in which they occur for the questions that the missing values would affect (2 and 3).

Importance: According to Christopher J. Ruhm in their paper "Alcohol policies and highway vehicle fatalities", traffic fatalities are the leading cause of death for people under the age of 40(1). Understanding the factors correlated with traffic fatalities may allow the public and policy makers to make changes that reduce the amount of death and suffering they cause.

Research Questions

1: Is there an association between US regions and number of traffic fatalities?

The different regions of the US can often have very different cultural and political attitudes, by understanding which parts of the country suffer more traffic fatalities we enable future research to investigate those differences in the hopes of finding treatments for the problem.

We will make use of the state and net fatality variables in our data (state and fatal respectively) in an ANOVA model to answer this question. We decided to split the country into 5 regions: West, Midwest, Northeast, Southwest and Southeast (The way the states have been split into regions is described below). We then will verify the assumptions of the ANOVA model, making any transformations necessary, before generating the model with US region as our independent variable and total regional fatalities as our dependent.

West: WA, MT, OR, ID, WY, NV, UT, CO, CA

Midwest: ND, MN, SD, WI, NE, IA, MI, KS, MO, IL, IN, OH

Northeast: PA, NY, VT, ME, NH, MA, RI, CT, NJ, DE, MD

Southwest: AZ, NM, TX, OK

Southeast: AK, LA, MS, TN, AL, KY, GA, WV, VA, NC, SC, FL

2: Which factors significantly predict the number of traffic fatalities?

This question is taking a more specific look at what variables affect fatality rates, while ignoring explicit geographical factors. Our starting model includes most of the variables in the dataset that are not subsets of fatalities themselves or national statistics, specifically our unsimplified model will contain: year, spirits, unemp, income, emppop, beertax, baptist, mormon, drinkage, dry, youngdrivers, miles, breath, jail, service, pop, and milestot. We will build a multiple linear regression model with these variables as independent and fatalities as dependent. We will then simplify the model by removing multicollinearity and variables with lower significance. We will then compare all of our models and pick the best one that allows our assumptions to hold.

3: Is there an association between mandatory jail time and mandatory community service by state?

People may behave differently based on whether there is community service or jail time associated with reckless driving practices so understanding the relationship between their existences in different states in different years might shed more light on the problem. Especially when combined with our previous question which looked at their affects on traffic fatalities directly. We will construct a frequency table between the two variables and then if it fits our assumptions we will run a chi-square test on the table, and a further difference in proportion test if it is significant.

Section 2: EDA

Variable Table and Data Type

Variable	**state**	**year**	**fatal**	**spirits**	**unemp**
Type	Categorical	Categorical	Numeric	Numeric	Numeric
Subtype	Nominal	Ordinal	Discrete	Continuous	Continuous
Units	–	–	–	–	–

Variable	**income**	**emppop**	**beertax**	**baptist**	**mormon**
Type	Numeric	Numeric	Numeric	Numeric	Numeric
Subtype	Continuous	Continuous	Continuous	Continuous	Continuous
Units	Dollars	–	Dollars	–	–

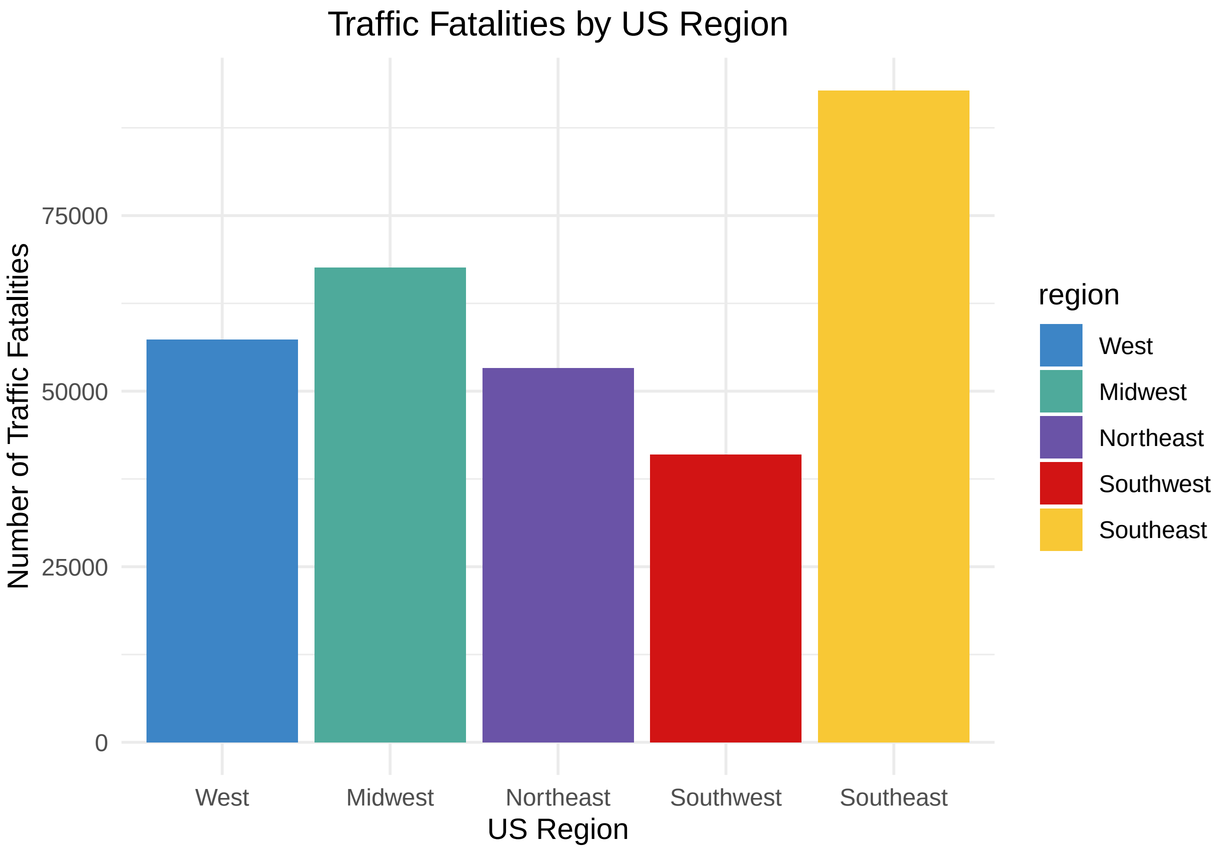
Variable	**drinkage**	**dry**	**youngdrivers**	**miles**	**breath**
Type	Numeric	Numeric	Numeric	Numeric	Categorical
Subtype	Continuous	Continuous	Continuous	Continuous	Nominal
Units	–	–	–	Miles	–

Variable	**jail**	**service**	**pop**	**milestot**	**unempus**
Type	Categorical	Categorical	Numeric	Numeric	Numeric
Subtype	Nominal	Nominal	Discrete	Discrete	Continuous
Units	–	–	–	Miles(mill)	–

Graphs

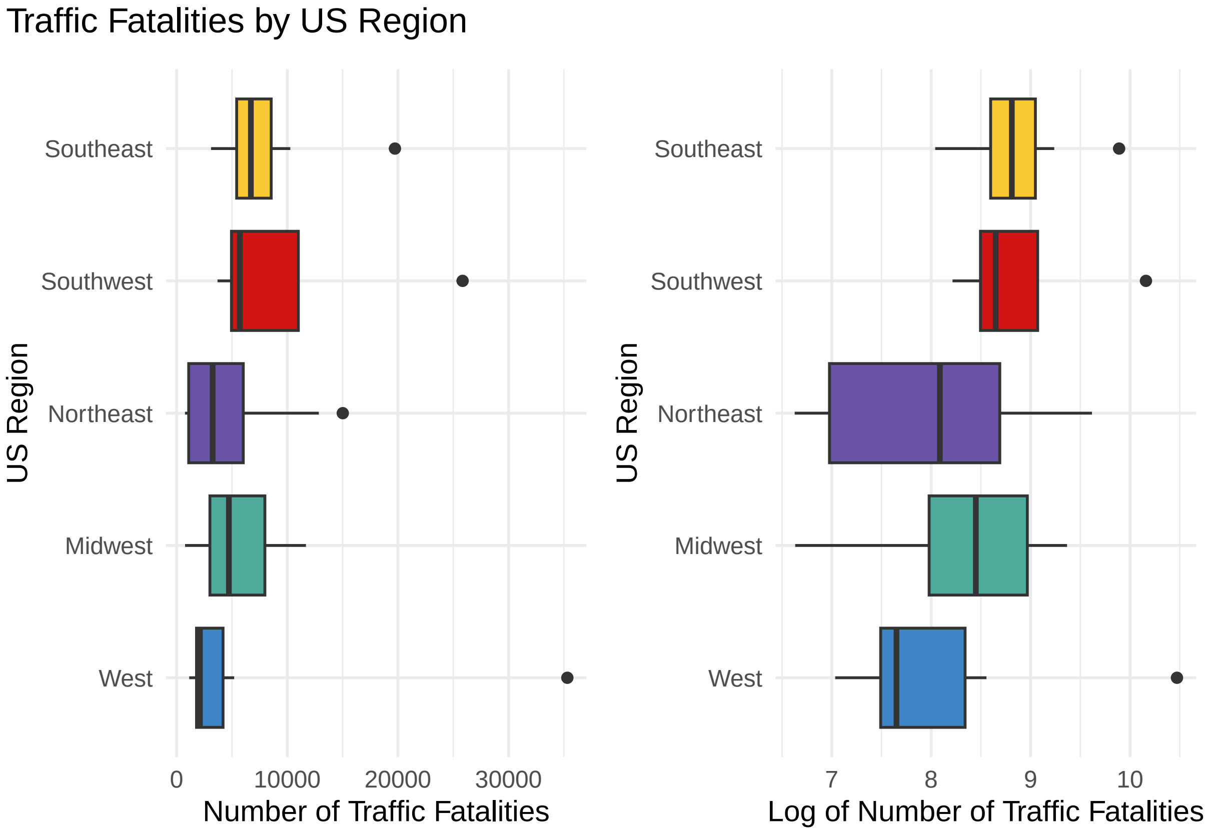
```
state_group <- c(5, 5, 4, 1, 1, 3, 3, 5, 5, 2, 1, 2, 2, 2, 5,
    5, 3, 3, 3, 2, 2, 2, 5, 1, 5, 2, 2, 3, 3, 4, 1, 3, 2, 4,
    1, 3, 3, 5, 2, 5, 4, 1, 5, 3, 1, 2, 5, 1)

tb <- aggregate(fatals$fatal, list(fatals$state), FUN = sum)
rownames(tb) <- paste(tb$Group.1, "_totals")
state_fatals <- tb[, 2]

total <- tibble(region = c("West", "Midwest", "Northeast", "Southwest",
    "Southeast"), fatalities = aggregate(state_fatals, list(state_group),
    FUN = sum)[, 2])
```

```
total$region <- factor(total$region, levels = c("West", "Midwest",
    "Northeast", "Southwest", "Southeast"))

p1 <- ggplot(data = total, aes(x = region, y = fatalities, fill = region)) +
    geom_bar(stat = "identity") + labs(x = "US Region", y = "Number of Traffic Fatalities") +
    ggtitle("Traffic Fatalities by US Region") + theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5))
p1 + scale_fill_manual(values = c("#3d85c6", "#4eaa9b", "#6a53a7",
    "#d21414", "#f8c835"))
```



Comments: From the bar plot of traffic fatalities by US regions, we can see that the southeast region has the most traffic fatalities and the southwest region has the least fatalities. We hypothesize that one of the biggest reason for this distribution is that the southeast region is more densely populated and includes more states than the southwest region which is less populated. For regions West, Midwest and Northeast, all these regions have states that are more populated and have have states that are not as populated and that they include roughly the same number of states so their total number of traffic fatalities are similar to each other.

```
total <- tibble(region = as.factor(state_group), fatalities = state_fatals)

total_log <- tibble(region = as.factor(state_group), fatalities_log = log(state_fatals))

p2 <- ggplot(total, aes(x = region, y = fatalities, fill = region)) +
    geom_boxplot() + labs(x = "US Region", y = "Number of Traffic Fatalities") +
    scale_x_discrete(labels = c("West", "Midwest", "Northeast",
        "Southwest", "Southeast")) + theme_minimal()

p3 <- ggplot(total_log, aes(x = region, y = fatalities_log, fill = region)) +
    geom_boxplot() + labs(x = "US Region", y = "Log of Number of Traffic Fatalities") +
    scale_x_discrete(labels = c("West", "Midwest", "Northeast",
        "Southwest", "Southeast")) + theme_minimal()

p2 + theme(legend.position = "none") + coord_flip() + scale_fill_manual(values = c("#3d85c6",
    "#4eaa9b", "#6a53a7", "#d21414", "#f8c835")) + p3 + theme(legend.position = "none") +
    coord_flip() + scale_fill_manual(values = c("#3d85c6", "#4eaa9b",
    "#6a53a7", "#d21414", "#f8c835")) + plot_annotation(title = "Traffic Fatalities by US Region")
```



Comments: We constructed box plots for each US region and their number of traffic fatalities and also the logged number of traffic fatalities. From the untransformed plot we see that there are 4 outliers and they are mostly not normally distributed. The West region appears to have the lowest median and Southeast has the highest median, they all look sightly right skewed and have less spread compared to after transforming the data. In the logged transformed plot, we can see that there are three outliers but the individual plots look more normally distributed compared to untransformed.

```
fivenum(fatals$fatal)
```

```
## [1]   79.0  292.5  701.0 1066.0 5504.0
```

```
mean(fatals$fatal)
```

```
## [1] 928.6637
```

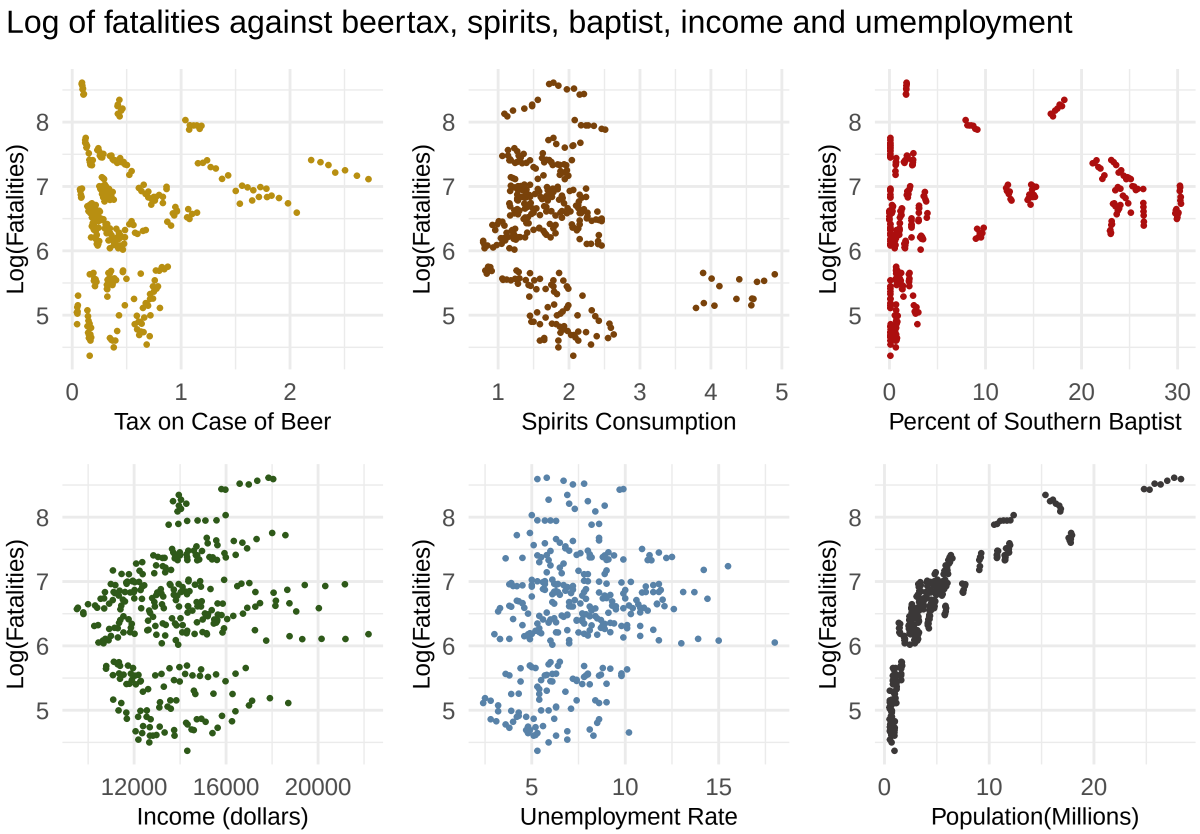
```
sd(fatals$fatal)
```

```
## [1] 934.0515
```

5 num summary for the number of fatalities in each state for each year

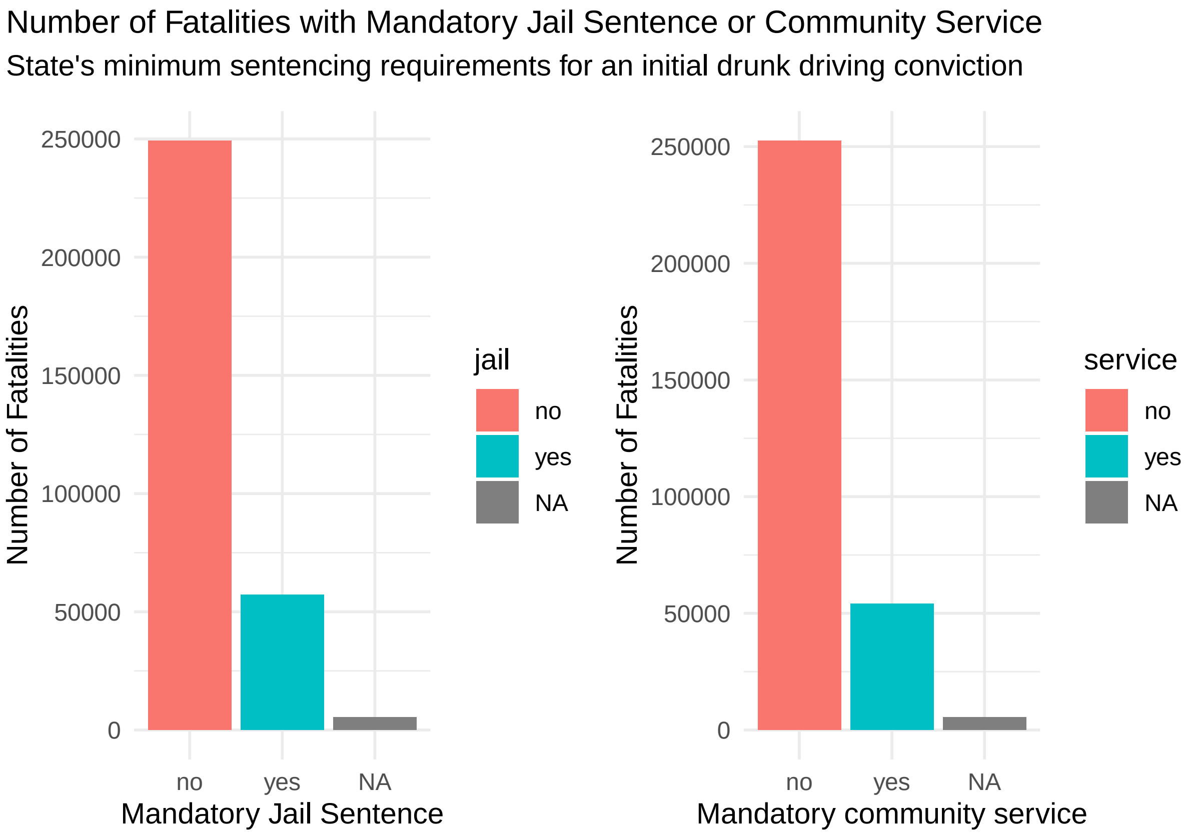
Min	1st Quartile	Med	Mean	SD	3rd Quartile	Max
79.0	292.5	701.0	928.6637	934.0515	1066.0	5504.0

```r
p4 <- ggplot(fatals, aes(x = beertax, y = log(fatal))) + geom_point(color = "#b88f11",
    size = 0.5) + labs(x = "Tax on Case of Beer", y = "Log(Fatalities)") +
    theme_minimal() + theme(axis.title = element_text(size = 9))

p5 <- ggplot(fatals, aes(x = spirits, y = log(fatal))) + geom_point(color = "#79420a",
    size = 0.5) + labs(x = "Spirits Consumption", y = "Log(Fatalities)") +
    theme_minimal() + theme(axis.title = element_text(size = 9))

p6 <- ggplot(fatals, aes(x = baptist, y = log(fatal))) + geom_point(color = "#ac0e0e",
    size = 0.5) + labs(x = "Percent of Southern Baptist", y = "Log(Fatalities)") +
    theme_minimal() + theme(axis.title = element_text(size = 9))

p7 <- ggplot(fatals, aes(x = income, y = log(fatal))) + geom_point(color = "#2e5919",
    size = 0.5) + labs(x = "Income (dollars)", y = "Log(Fatalities)") +
    theme_minimal() + theme(axis.title = element_text(size = 9))

p8 <- ggplot(fatals, aes(x = unemp, y = log(fatal))) + geom_point(color = "#5882a8",
    size = 0.5) + labs(x = "Unemployment Rate", y = "Log(Fatalities)") +
    theme_minimal() + theme(axis.title = element_text(size = 9))

p9 <- ggplot(fatals, aes(x = pop/1e+06, y = log(fatal))) + geom_point(color = "#3b3838",
    size = 0.5) + labs(x = "Population(Millions)", y = "Log(Fatalities)") +
    theme_minimal() + theme(axis.title = element_text(size = 9))

p4 + p5 + p6 + p7 + p8 + p9 + plot_annotation(title = "Log of fatalities against beertax, spirits, bapti
    theme = theme(plot.title = element_text(size = 12)))
```



Comments: From the beer tax scatter plot we can see that there doesn't appear to be any linearity and there is a very weak positive correlation between taxes and number of fatalities. In the spirits consumption plot, there also doesn't appear to be any linearity or correlation, there are some outliers where consumption is around 4 and 5 where the number of fatalities is small. The scatter plot for percentage of baptist shows a very weak positive correlation but no linearity between the two variables. Both the scatter plot for income and unemployment rate shows no linearity and both has an extremely weak positive correlation. The population plot shows a postive correlation but doesn't show linearity.

```
p10 <- ggplot(data = fatals, aes(x = jail, y = fatal, fill = jail)) +
    geom_col() + labs(x = "Mandatory Jail Sentence", y = "Number of Fatalities") +
    theme_minimal()

p11 <- ggplot(data = fatals, aes(x = service, y = fatal, fill = service)) +
    geom_col() + labs(x = "Mandatory community service", y = "Number of Fatalities") +
    theme_minimal()

p10 + p11 + plot_annotation(title = "Number of Fatalities with Mandatory Jail Sentence or Community Ser
    subtitle = "State's minimum sentencing requirements for an initial drunk driving conviction",
    theme = theme(plot.title = element_text(size = 12)))
```



Comments: From the two bar plots we can see that when states have no mandatory jail sentence or mandatory community service for first time drunk driving convictions, the number of fatalities is far greater than when states have mandatory jail sentences or community services.

Table for mandatory jail or community service for each state in each year

service jail	no	yes
no	227	14
yes	46	48

Section 3: Statistical Results

Question 1

ANOVA ASSUMPTIONS:

ANOVA requires a single categorical independent and a single numerical dependent variable, our independent categorical variable corresponds to the 5 US regions defined above, and our numerical dependent variable is the number of traffic fatalities. Our dependent variable is at an interval level and independent variable is categorical with 5 groups that are independent of each other, meaning the number of fatalities in each region does not affect the number of fatalities in another region. Each US region only maps to one value of our dependent variable. **EDITS: ANOVA also requires the dependent variable to be normally distributed for each category of the independent variable, Looking at the box plot above for "Traffic Fatalities by US Region", the non transformed data looked a bit skewed so we performed a log transformation. After transforming the data looks more normal within each region group, so we are going to assume normality and use the logged values for the anova test.** According to the Levene's test we get a p-value of 0.1545 which is greater than 0.05 so it's safe to assume that the variances for each group of US regions is equal. From the box plot above, we can see that there are

3 outliers, so we removed the outliers from the data set.

H_o : There is no statistically significant difference in mean traffic fatalities with respect to the five US region groups

H_A : There is a statistically significant difference in mean traffic fatalities with respect to the five US region groups

```
# new logged data we will be using
state_fatals_log = log(state_fatals)

# test for equal variance
leveneTest(state_fatals_log ~ as.factor(state_group))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  4    1.76 0.1545
##       43
```

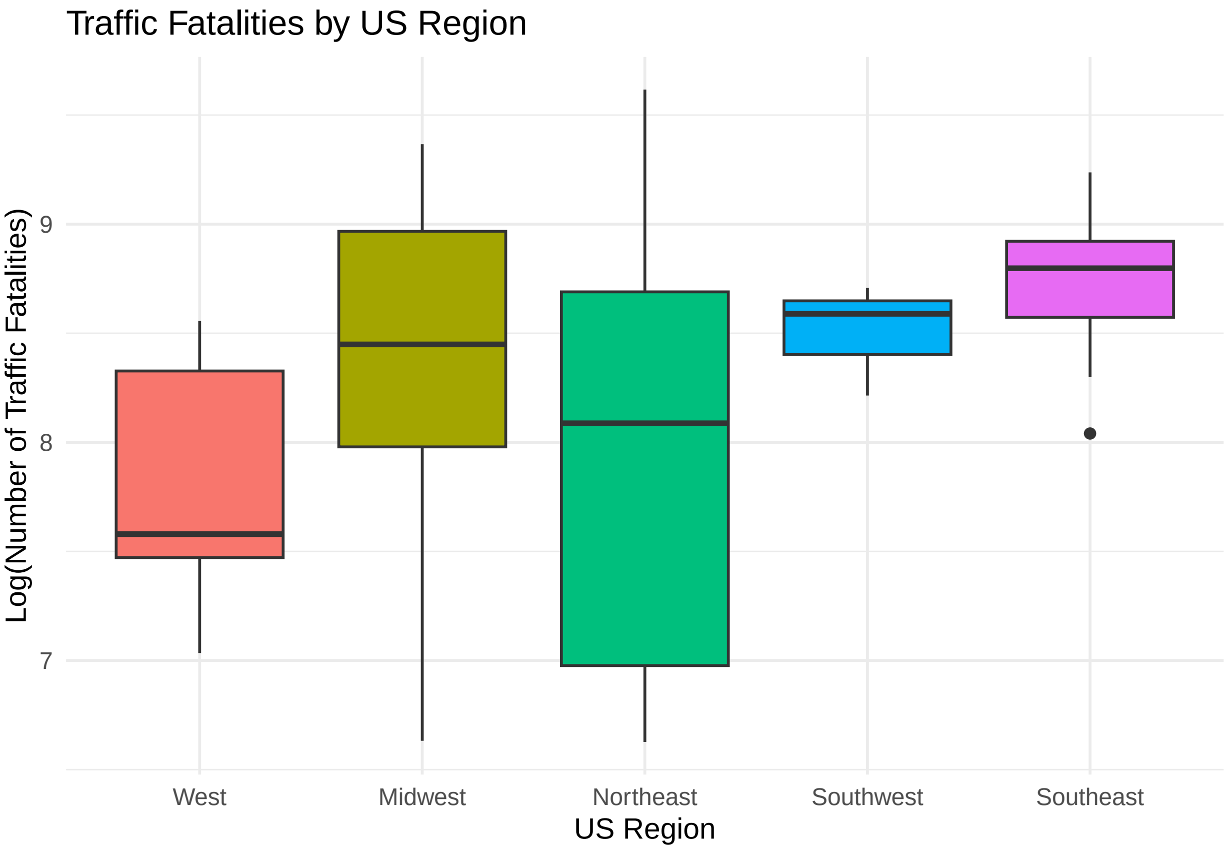
```
# 1st anova test on complete data
AOV_model <- aov(state_fatals_log ~ as.factor(state_group))
summary(AOV_model)
```

```
##                        Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(state_group)  4   6.40  1.5999   1.988  0.113
## Residuals              43  34.61  0.8049
```

```
# highest three values are outliers and their indexes in
# the original data is 4,8,41
sort(state_fatals_log)
```

```
##  [1]  6.626718  6.632002  6.668228  6.811244  6.918695  7.034388  7.142037
##  [8]  7.332369  7.411556  7.491645  7.507690  7.542213  7.650645  8.040447
## [15]  8.087333  8.124743  8.125335  8.214465  8.298540  8.301273  8.322394
## [22]  8.340933  8.493720  8.524169  8.526747  8.555837  8.589142  8.596374
## [29]  8.622094  8.707648  8.794825  8.797851  8.824237  8.850088  8.852093
## [36]  8.852522  8.863616  8.991189  9.218110  9.237274  9.276877  9.337061
## [43]  9.366404  9.461099  9.616805  9.889997 10.159950 10.472063
```

```
new_log <- state_fatals_log[-c(4, 8, 41)]
new_state <- state_group[-c(4, 8, 41)]

total_log <- tibble(region = as.factor(new_state), fatalities_log = new_log)
outliercheck <- ggplot(total_log, aes(x = region, y = fatalities_log,
    fill = region)) + geom_boxplot() + labs(x = "US Region",
    y = "Log(Number of Traffic Fatalities)", title = "Traffic Fatalities by US Region") +
    scale_x_discrete(labels = c("West", "Midwest", "Northeast",
        "Southwest", "Southeast")) + theme_minimal()
outliercheck + theme(legend.position = "none")
```



```
# from the box plot we removed another outlier(lowest
# value) in the southeast region
a <- data.frame(new_log, new_state)
a1 <- a %>%
    filter(new_state == 5)
sort(a1$new_log)
```

```
##  [1] 8.040447 8.298540 8.524169 8.622094 8.794825 8.797851 8.824237 8.852093
##  [9] 8.991189 9.218110 9.237274
```

```
# Anova test after removing outliers
new_log1 <- new_log[-c(44)]
new_state1 <- new_state[-c(44)]
AOV_model_2 <- aov(new_log1 ~ as.factor(new_state1))
summary(AOV_model_2)
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(new_state1)  4  6.052  1.5130   2.417 0.0649 .
## Residuals             39 24.408  0.6258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANALYSIS: We did a log transformation and performed an anova test and got a p-value of 0.113 which is again greater than 0.05. We then looked at the box plots and removed the outliers from the logged traffic fatalities data and checked for any remaining outliers, we ended up removing another outlier and we ran the anova test again and got a p-value of 0.0649 which is still greater than 0.05. From both the tests we can come to the conclusion that since all the p-values are greater than 5% of significance level, then we fail to reject our null hypothesis so that there is no statistically significant difference in mean traffic fatalities with respect to the five US region groups.

The p-value for the anova test after we removed the outlier is not statistically significant for a 5% significance level, but it is statistically significant for a 10% significance level, so for a post-hoc test, we've decided to run the Tukey test at the 10% significance level. We are using a Tukey test because it allows us to compare all possible combinations of any 2 US regions against each other, since we are comparing all pairs, we will not be using the Dunnett test because that one makes comparison with a reference group which we don't have, we will also not be using the Bonferroni test because it is best used when we have a small set of planned comparison but in this case we are comparing all the possible sets. We will not be using the Scheffe method because it compares more than two means at once however we only need to compare two means. We will also not be using the Fisher's LSD because it doesn't control the family-wise error rate and it is less conservative than the Tukey test (Lee & Lee, 2018).

```r
# Post-Hoc test
TukeyHSD(AOV_model_2, conf.level = 0.9)
```

```
##   Tukey multiple comparisons of means
##     90% family-wise confidence level
##
## Fit: aov(formula = new_log1 ~ as.factor(new_state1))
##
## $`as.factor(new_state1)`
##           diff         lwr      upr     p adj
## 2-1  0.5385040 -0.38269300 1.459701 0.5740943
## 3-1  0.1759614 -0.76183507 1.113758 0.9888693
## 4-1  0.7143656 -0.65199034 2.080722 0.6722206
## 5-1  1.0266520  0.06931594 1.983988 0.0665554
## 3-2 -0.3625426 -1.20500415 0.479919 0.8065490
## 4-2  0.1758616 -1.12690769 1.478631 0.9968469
## 5-2  0.4881479 -0.37601145 1.352307 0.6056176
## 4-3  0.5384042 -0.77615510 1.852964 0.8328686
## 5-3  0.8506905 -0.03114266 1.732524 0.1208195
## 5-4  0.3122863 -1.01628290 1.640856 0.9743123
```

From the Tukey test, we can see that the only p-value that is less than the 10% significance level is from group 5-1, which is the difference in means of the West and Southeast region. **Looking at the differences, we can see that the difference between the groups 5 and 1 is approximately 1.0266520, since this is a positive number, we can conclude that the mean traffic fatality number for Southeast is greater than the mean traffic fatality for West by 1.0266520.** Looking at the confidence interval, the only interval that doesn't include 0 is also the 5-1 group and it's positive. So we can say that there is a statistically significant difference in means between the US regions Southeast and West at a 10% level of significance where the mean for Southeast is greater than the mean for West.

Looking at the other p-values, Midwest-west has a p-value of 0.57. Northeast-West has a p-value of 0.99. Southwest-West has a p-value of 0.67. Northeast-Midwest has a p-value of 0.81. Southwest-Midwest has a p-value of 0.997. Southeast-Midwest has a p-value of 0.61. Southwest-Northeast has a p-value of 0.83. Southeast-Northeast has a p-value of 0.12. Southeast-Southwest has a p-value of 0.97. All these p-values are greater than 0.1 so there is no statistically significant difference between the means of these groups.

Question 2:

INITIAL MULTIPLE LINEAR REGRESSION ASSUMPTIONS:

Normality of Dependent Variable: We tried a series of transformations (logarithmic, squareroot and box-cox) and were unable to achieve a p-value below 0.05 on a Shapiro test. However the logarithmic transformation did significantly improve the qqnorm plot of fatalities, as well as reduce outliers, so we used it as the dependent of our model and will assume normality going forward.

```
# Removing missing data and attempting transformations:
fatals_clean <- na.omit(fatals)

shapiro.test(fatals_clean$fatal)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fatals_clean$fatal
## W = 0.7441, p-value < 2.2e-16
```

```
shapiro.test(log(fatals_clean$fatal))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(fatals_clean$fatal)
## W = 0.97727, p-value = 3.841e-05
```

```
shapiro.test(sqrt(fatals_clean$fatal))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt(fatals_clean$fatal)
## W = 0.92516, p-value = 6.632e-12
```

```
shapiro.test(1/fatals_clean$fatal)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  1/fatals_clean$fatal
## W = 0.76086, p-value < 2.2e-16
```

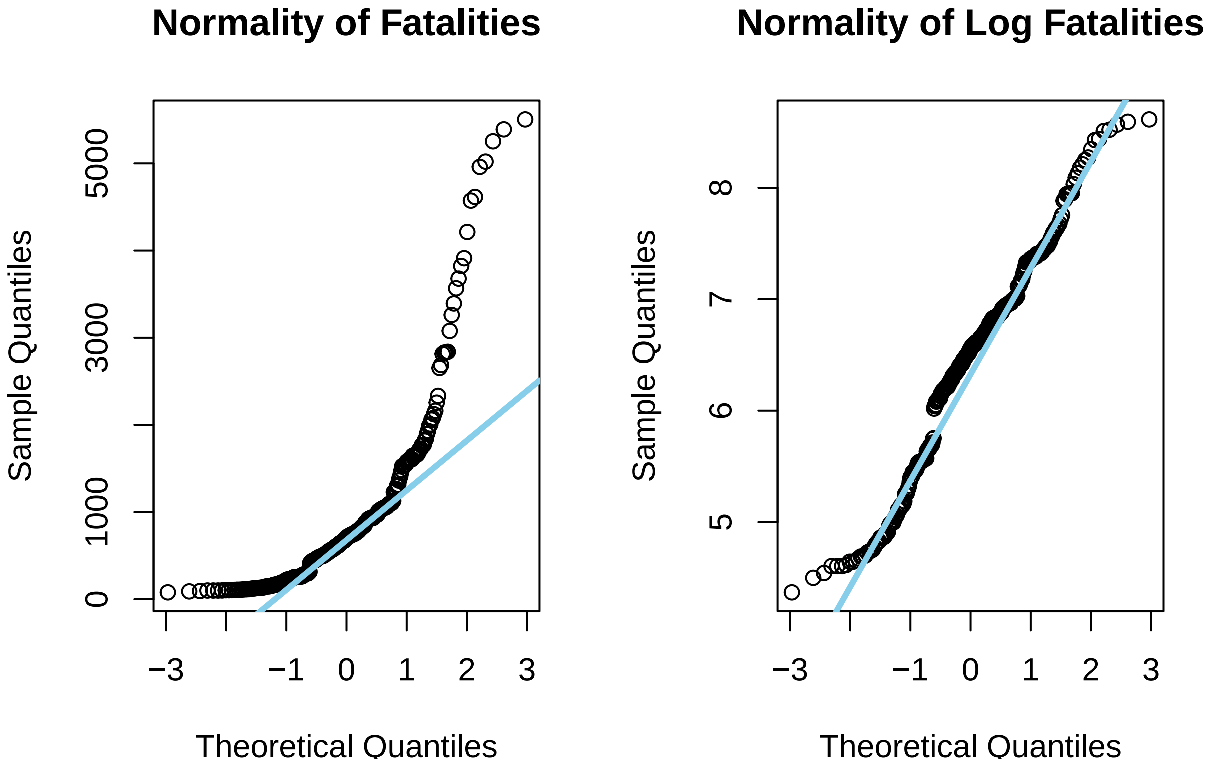
```
bc1 <- boxcox(fatals_clean$fatal ~ factor(fatals_clean$year) +
    fatals_clean$spirits + fatals_clean$unemp + fatals_clean$income +
    fatals_clean$emppop + fatals_clean$beertax + fatals_clean$baptist +
    fatals_clean$mormon + fatals_clean$drinkage + fatals_clean$dry +
    fatals_clean$youngdrivers + fatals_clean$miles + factor(fatals_clean$breath) +
    factor(fatals_clean$jail) + factor(fatals_clean$service) +
    fatals_clean$pop + fatals_clean$milestot)
```

```
lamda1 <- bc1$x[which.max(bc1$y)]

shapiro.test(((fatals_clean$fatal^lamda1) - 1)/lamda1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ((fatals_clean$fatal^lamda1) - 1)/lamda1
## W = 0.82916, p-value < 2.2e-16
```

```
par(mfrow = c(1, 2))

qqnorm(fatals$fatal, main = "Normality of Fatalities")
qqline(fatals$fatal, lw = 3, col = "skyblue")
```

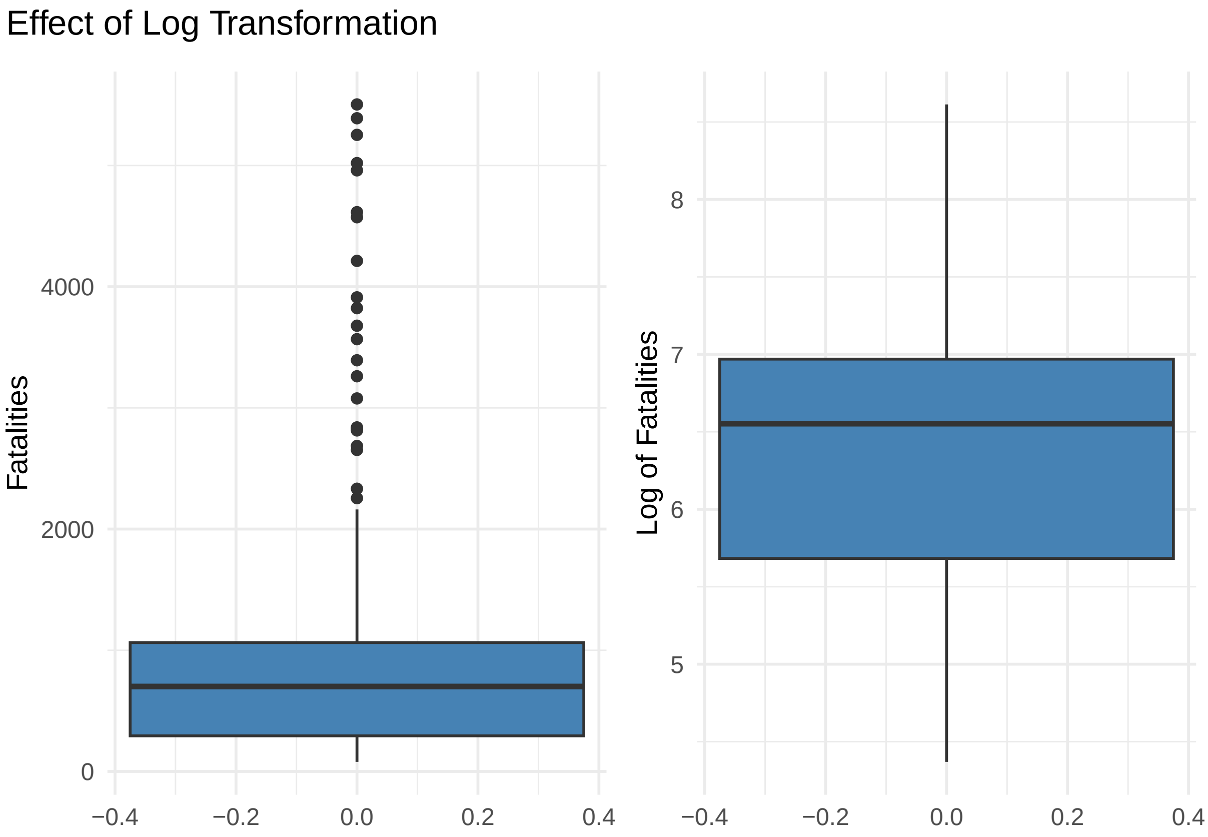
```
qqnorm(log(fatals$fatal), main = "Normality of Log Fatalities")
qqline(log(fatals$fatal), lw = 3, col = "skyblue")
```



Linear Relationship of Variables: Based on the EDA above the data does not show very clear linearity between the variables but they do not appear randomly scattered either, however we will assume linearity while building our model.

No significant outliers: Taking the log of the dependent variable serves the dual purpose of improving normality and reducing the number of outliers present to 0.

```
p12 <- ggplot(fatals, aes(y = fatal)) + labs(y = "Fatalities") +
    geom_boxplot(fill = "steelblue") + theme_minimal()
p13 <- ggplot(fatals, aes(y = log(fatal))) + labs(y = "Log of Fatalities") +
    geom_boxplot(fill = "steelblue") + theme_minimal()

p12 + p13 + plot_annotation(title = "Effect of Log Transformation")
```



With our initial assumptions made we can move on to building and refining a model. Our first model will contain almost all of the available independent variables.

CHOICE OF MODEL SELECTION CRITERION: According to a paper from Behavioral Ecology and Sociobiology titled "A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion" AIC is better suited than other model comparison metrics for cases "where no one model is strongly supported"(1), which as will be seen applies very easily to our model selection process, as the differences between the different models created are quite minor.

```
# Building the first model with all independent variables.
mlrmodel1 <- lm(log(fatals_clean$fatal) ~ factor(fatals_clean$year) +
    fatals_clean$spirits + fatals_clean$unemp + fatals_clean$income +
    fatals_clean$emppop + fatals_clean$beertax + fatals_clean$baptist +
    fatals_clean$mormon + fatals_clean$drinkage + fatals_clean$dry +
    fatals_clean$youngdrivers + fatals_clean$miles + factor(fatals_clean$breath) +
    factor(fatals_clean$jail) + factor(fatals_clean$service) +
    fatals_clean$pop + fatals_clean$milestot)
summary(mlrmodel1)
```

```
##
## Call:
## lm(formula = log(fatals_clean$fatal) ~ factor(fatals_clean$year) +
##     fatals_clean$spirits + fatals_clean$unemp + fatals_clean$income +
##     fatals_clean$emppop + fatals_clean$beertax + fatals_clean$baptist +
##     fatals_clean$mormon + fatals_clean$drinkage + fatals_clean$dry +
##     fatals_clean$youngdrivers + fatals_clean$miles + factor(fatals_clean$breath) +
##     factor(fatals_clean$jail) + factor(fatals_clean$service) +
##     fatals_clean$pop + fatals_clean$milestot)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.12614 -0.17467  0.04918  0.23348  0.74344 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      4.527e+00  8.363e-01   5.413 1.24e-07 ***
## factor(fatals_clean$year)1983   -8.886e-02  7.209e-02  -1.233  0.21864    
## factor(fatals_clean$year)1984   -7.257e-03  7.773e-02  -0.093  0.92568    
## factor(fatals_clean$year)1985   -7.303e-02  8.033e-02  -0.909  0.36398    
## factor(fatals_clean$year)1986   -1.042e-01  8.695e-02  -1.198  0.23187    
## factor(fatals_clean$year)1987   -1.217e-01  9.401e-02  -1.294  0.19660    
## factor(fatals_clean$year)1988   -9.899e-02  1.040e-01  -0.952  0.34177    
## fatals_clean$spirits            -2.566e-01  4.052e-02  -6.333 8.39e-10 ***
## fatals_clean$unemp               6.925e-02  1.632e-02   4.244 2.90e-05 ***
## fatals_clean$income              1.245e-04  1.731e-05   7.194 4.69e-12 ***
## fatals_clean$emppop             -1.359e-02  8.861e-03  -1.533  0.12619    
## fatals_clean$beertax             1.711e-01  6.033e-02   2.836  0.00487 ** 
## fatals_clean$baptist             2.725e-02  3.692e-03   7.380 1.44e-12 ***
## fatals_clean$mormon             -2.709e-03  2.483e-03  -1.091  0.27603    
## fatals_clean$drinkage            2.047e-02  2.416e-02   0.847  0.39768    
## fatals_clean$dry                -7.013e-04  2.771e-03  -0.253  0.80039    
## fatals_clean$youngdrivers       -1.360e+00  1.130e+00  -1.203  0.22972    
## fatals_clean$miles              -1.155e-05  1.778e-05  -0.650  0.51636    
## factor(fatals_clean$breath)yes  -4.137e-02  4.849e-02  -0.853  0.39420    
## factor(fatals_clean$jail)yes    -8.048e-02  6.278e-02  -1.282  0.20078    
## factor(fatals_clean$service)yes  4.123e-01  6.442e-02   6.400 5.69e-10 ***
## fatals_clean$pop                 7.034e-08  2.393e-08   2.940  0.00353 ** 
## fatals_clean$milestot            7.690e-06  3.182e-06   2.417  0.01624 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3479 on 312 degrees of freedom
## Multiple R-squared:  0.8678, Adjusted R-squared:  0.8585 
## F-statistic:  93.1 on 22 and 312 DF,  p-value: < 2.2e-16
```

We can begin refining our base model by using a backwards step function to create a second model:

```
ols_step_backward_p(mlrmodel1)
```

```
## 
## 
##                                     Elimination Summary                                      
## ---------------------------------------------------------------------------------------------
##         Variable                                   Adj.                                          
## Step             Removed             R-Square    R-Square     C(p)        AIC        RMSE    
## ---------------------------------------------------------------------------------------------
##    1    fatals_clean$dry                0.8678      0.8589    11.0640    265.4255    0.3473    
##    2    factor(fatals_clean$year)       0.8662      0.8599    12.7684    257.3788    0.3461    
##    3    fatals_clean$drinkage           0.8661      0.8602    11.0726    255.7013    0.3457    
##    4    fatals_clean$miles              0.8659      0.8604     9.6069    254.2671    0.3455    
##    5    fatals_clean$youngdrivers       0.8656      0.8606     8.1854    252.8786    0.3452    
## ---------------------------------------------------------------------------------------------
```

Removing the five variables recommended by the function gives the following model:

```
mlrmodel2 <- update(mlrmodel1, . ~ . - fatals_clean$dry - factor(fatals_clean$year) -
    fatals_clean$drinkage - fatals_clean$miles - fatals_clean$youngdrivers)

summary(mlrmodel2)
```

```
##
## Call:
## lm(formula = log(fatals_clean$fatal) ~ fatals_clean$spirits +
##     fatals_clean$unemp + fatals_clean$income + fatals_clean$emppop +
##     fatals_clean$beertax + fatals_clean$baptist + fatals_clean$mormon +
##     factor(fatals_clean$breath) + factor(fatals_clean$jail) +
##     factor(fatals_clean$service) + fatals_clean$pop + fatals_clean$milestot)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03778 -0.18294  0.03158  0.23301  0.74605
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.742e+00  6.194e-01   7.656 2.25e-13 ***
## fatals_clean$spirits            -2.419e-01  3.705e-02  -6.529 2.58e-10 ***
## fatals_clean$unemp               6.448e-02  1.452e-02   4.441 1.23e-05 ***
## fatals_clean$income              1.262e-04  1.613e-05   7.822 7.52e-14 ***
## fatals_clean$emppop             -1.706e-02  8.494e-03  -2.009 0.045384 *
## fatals_clean$beertax             1.552e-01  5.637e-02   2.753 0.006233 **
## fatals_clean$baptist             2.688e-02  3.072e-03   8.750  < 2e-16 ***
## fatals_clean$mormon             -2.922e-03  2.416e-03  -1.210 0.227317
## factor(fatals_clean$breath)yes  -5.600e-02  4.638e-02  -1.207 0.228188
## factor(fatals_clean$jail)yes    -9.353e-02  5.910e-02  -1.582 0.114539
## factor(fatals_clean$service)yes  4.169e-01  6.279e-02   6.640 1.33e-10 ***
## fatals_clean$pop                 7.968e-08  2.055e-08   3.878 0.000128 ***
## fatals_clean$milestot            6.573e-06  2.738e-06   2.401 0.016921 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3452 on 322 degrees of freedom
## Multiple R-squared:  0.8656, Adjusted R-squared:  0.8606
## F-statistic: 172.8 on 12 and 322 DF,  p-value: < 2.2e-16
```

```
compareLM(mlrmodel1, mlrmodel2)
```

```
## $Models
##   Formula
## 1 "log(fatals_clean$fatal) ~ factor(fatals_clean$year) + fatals_clean$spirits + fatals_clean$unemp +
## 2 "log(fatals_clean$fatal) ~ fatals_clean$spirits + fatals_clean$unemp + fatals_clean$income + fatals
##
## $Fit.criteria
##   Rank Df.res   AIC  AICc   BIC R.squared Adj.R.sq    p.value Shapiro.W
## 1   23    312 267.4 271.2 358.9    0.8678   0.8585 6.555e-123    0.9708
## 2   13    322 252.9 254.2 306.3    0.8656   0.8606 2.233e-132    0.9732
##   Shapiro.p
## 1 2.769e-06
## 2 6.953e-06
```

We can see that the backwards step did significantly improve the model based on AIC. We will now remove some of the less significant independent variables to see if we can create a simpler model that still performs well:

```
mlrmodel3 <- update(mlrmodel2, . ~ . - fatals_clean$mormon -
    factor(fatals_clean$breath) - factor(fatals_clean$jail))
summary(mlrmodel3)
```

```
##
## Call:
## lm(formula = log(fatals_clean$fatal) ~ fatals_clean$spirits +
##     fatals_clean$unemp + fatals_clean$income + fatals_clean$emppop +
##     fatals_clean$beertax + fatals_clean$baptist + factor(fatals_clean$service) +
##     fatals_clean$pop + fatals_clean$milestot)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02734 -0.17802  0.03306  0.23786  0.76194
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.570e+00  5.994e-01   7.625 2.71e-13 ***
## fatals_clean$spirits             -2.350e-01  3.573e-02  -6.577 1.91e-10 ***
## fatals_clean$unemp                6.494e-02  1.431e-02   4.538 8.01e-06 ***
## fatals_clean$income               1.342e-04  1.525e-05   8.799  < 2e-16 ***
## fatals_clean$emppop              -1.751e-02  8.099e-03  -2.162 0.031337 *
## fatals_clean$beertax              1.585e-01  5.611e-02   2.825 0.005020 **
## fatals_clean$baptist              2.892e-02  2.865e-03  10.095  < 2e-16 ***
## factor(fatals_clean$service)yes  3.482e-01  5.006e-02   6.955 1.94e-11 ***
## fatals_clean$pop                  7.809e-08  2.008e-08   3.888 0.000122 ***
## fatals_clean$milestot             7.009e-06  2.689e-06   2.607 0.009561 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3457 on 325 degrees of freedom
## Multiple R-squared:  0.864,  Adjusted R-squared:  0.8602
## F-statistic: 229.4 on 9 and 325 DF,  p-value: < 2.2e-16
```

```
compareLM(mlrmodel1, mlrmodel2, mlrmodel3)
```

```
## $Models
##   Formula
## 1 "log(fatals_clean$fatal) ~ factor(fatals_clean$year) + fatals_clean$spirits + fatals_clean$unemp +
## 2 "log(fatals_clean$fatal) ~ fatals_clean$spirits + fatals_clean$unemp + fatals_clean$income + fatal
## 3 "log(fatals_clean$fatal) ~ fatals_clean$spirits + fatals_clean$unemp + fatals_clean$income + fatal
##
## $Fit.criteria
##   Rank Df.res   AIC  AICc   BIC R.squared Adj.R.sq    p.value Shapiro.W
## 1   23    312 267.4 271.2 358.9    0.8678   0.8585 6.555e-123    0.9708
## 2   13    322 252.9 254.2 306.3    0.8656   0.8606 2.233e-132    0.9732
## 3   10    325 250.9 251.7 292.8    0.8640   0.8602 4.665e-135    0.9765
##   Shapiro.p
## 1 2.769e-06
## 2 6.953e-06
## 3 2.767e-05
```

The third model has three less variables and has slightly lower AIC than the other two models so we will use it going forward.

MULTIPLE LINEAR REGRESSION FINAL ASSUMPTIONS:

EDIT: The multicollinearity section has been moved to the end of the model creation process. Some variables are still removed during this step to fix the multicollinearity, as is done in the Multiple Linear Regression notes on our canvas page.

Multicollinearity: We will begin checking multicollinearity by computing the VIF of the variables in our model:

```r
vif(mlrmodel3)
```

```
##         fatals_clean$spirits           fatals_clean$unemp
##                     1.671845                     3.677427
##          fatals_clean$income          fatals_clean$emppop
##                     3.275888                     4.094202
##         fatals_clean$beertax         fatals_clean$baptist
##                     2.010095                     2.191415
## factor(fatals_clean$service)             fatals_clean$pop
##                     1.059607                    27.255574
##        fatals_clean$milestot
##                    25.890889
```

We can see that population and total miles both have very high VIF's because they are correlated, total miles is less significant so we will remove it from the model rather than population.

```r
mlrmodel4 <- update(mlrmodel3, . ~ . - fatals_clean$milestot)
```

All of our VIF's are now below 5, to finish checking multicollinearity we will compute the correlation matrix of the numerical variables and remove any that are highly correlated (above 0.8).

```r
# Creating a version of fatals containing only the numeric
# variables we are using
fatals_numeric_only <- fatals_clean[, -c(1:3, 15:17, 19:27, 29:31,
    32:35)]

# Generating the correlation matrix of the numeric values
cor(fatals_numeric_only)
```

```
##                  spirits        unemp      income     emppop     beertax
## spirits       1.00000000 -0.239529558  0.45475142  0.4108859 -0.08970862
## unemp        -0.23952956  1.000000000 -0.55251636 -0.8004410  0.05473103
## income        0.45475142 -0.552516362  1.00000000  0.5218040 -0.39506992
## emppop        0.41088589 -0.800440977  0.52180404  1.0000000 -0.16051830
## beertax      -0.08970862  0.054731029 -0.39506992 -0.1605183  1.00000000
## baptist      -0.29449906  0.263027640 -0.47413254 -0.3578730  0.63262502
## mormon       -0.17884609 -0.008138022 -0.21815266  0.1114799  0.00487962
## drinkage     -0.08376709 -0.258533795  0.20087463  0.1637044 -0.05853347
## dry          -0.26820761  0.255657203 -0.34303933 -0.3393203  0.17668206
## youngdrivers -0.05893215  0.384462745 -0.47413361 -0.2065180  0.24519522
## miles        -0.05643949 -0.277448172 -0.08305713  0.3203545  0.14332441
## fatal        -0.09570603  0.103377100  0.22200623 -0.1917515  0.06526388
## pop          -0.06661884  0.093829152  0.35323140 -0.1810682 -0.07865120
##                   baptist        mormon     drinkage         dry youngdrivers
## spirits      -0.294499060 -0.1788460919 -0.083767094 -0.26820761  -0.05893215
## unemp         0.263027640 -0.0081380215 -0.258533795  0.25565720   0.38446275
## income       -0.474132541 -0.2181526641  0.200874630 -0.34303933  -0.47413361
## emppop       -0.357872959  0.1114799497  0.163704353 -0.33932029  -0.20651797
## beertax       0.632625022  0.0048796199 -0.058533472  0.17668206   0.24519522
## baptist       1.000000000 -0.1482848944  0.057159292  0.57136604   0.17009628
## mormon       -0.148284894  1.0000000000  0.009930955 -0.09096457   0.20591120
## drinkage      0.057159292  0.0099309546  1.000000000  0.14018210  -0.27804654
## dry           0.571366036 -0.0909645708  0.140182103  1.00000000   0.06205779
## youngdrivers  0.170096279  0.2059112007 -0.278046541  0.06205779   1.00000000
## miles         0.136178065  0.0006018359  0.058539616 -0.08218782  -0.05429201
## fatal         0.174184635 -0.1538592729  0.038674132  0.08004450  -0.15101732
## pop           0.008483297 -0.1617292375  0.059827975  0.03350139  -0.20866906
##                      miles       fatal          pop
## spirits      -0.0564394864 -0.09570603 -0.066618840
## unemp        -0.2774481723  0.10337710  0.093829152
## income       -0.0830571315  0.22200623  0.353231395
## emppop        0.3203545367 -0.19175149 -0.181068220
## beertax       0.1433244084  0.06526388 -0.078651201
## baptist       0.1361780649  0.17418463  0.008483297
## mormon        0.0006018359 -0.15385927 -0.161729238
## drinkage      0.0585396156  0.03867413  0.059827975
## dry          -0.0821878244  0.08004450  0.033501394
## youngdrivers -0.0542920148 -0.15101732 -0.208669060
## miles         1.0000000000 -0.13232041 -0.255741134
## fatal        -0.1323204059  1.00000000  0.945098207
## pop          -0.2557411338  0.94509821  1.000000000
```

The only numerical values that appear highly correlated are employment population and unemployment percentage, this is expected and we can remove employment population as it is less significant in our models than unemployment percentage:

```
mlrmodel5 <- update(mlrmodel4, . ~ . - fatals_clean$emppop)
summary(mlrmodel5)
```

```
##
## Call:
## lm(formula = log(fatals_clean$fatal) ~ fatals_clean$spirits +
##     fatals_clean$unemp + fatals_clean$income + fatals_clean$beertax +
##     fatals_clean$baptist + factor(fatals_clean$service) + fatals_clean$pop)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00523 -0.19895  0.03296  0.24911  0.81061
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.423e+00  2.483e-01  13.784  < 2e-16 ***
## fatals_clean$spirits           -2.662e-01  3.384e-02  -7.865 5.44e-14 ***
## fatals_clean$unemp              8.208e-02  1.029e-02   7.975 2.58e-14 ***
## fatals_clean$income             1.333e-04  1.539e-05   8.662  < 2e-16 ***
## fatals_clean$beertax            1.766e-01  5.625e-02   3.140  0.00184 **
## fatals_clean$baptist            3.128e-02  2.769e-03  11.294  < 2e-16 ***
## factor(fatals_clean$service)yes 3.528e-01  5.050e-02   6.986 1.59e-11 ***
## fatals_clean$pop                1.299e-07  4.854e-09  26.766  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3491 on 327 degrees of freedom
## Multiple R-squared:  0.8605, Adjusted R-squared:  0.8575
## F-statistic: 288.1 on 7 and 327 DF,  p-value: < 2.2e-16
```

Autocorrelation: We will employ a Durbin-Watson test to check for autocorrelation:

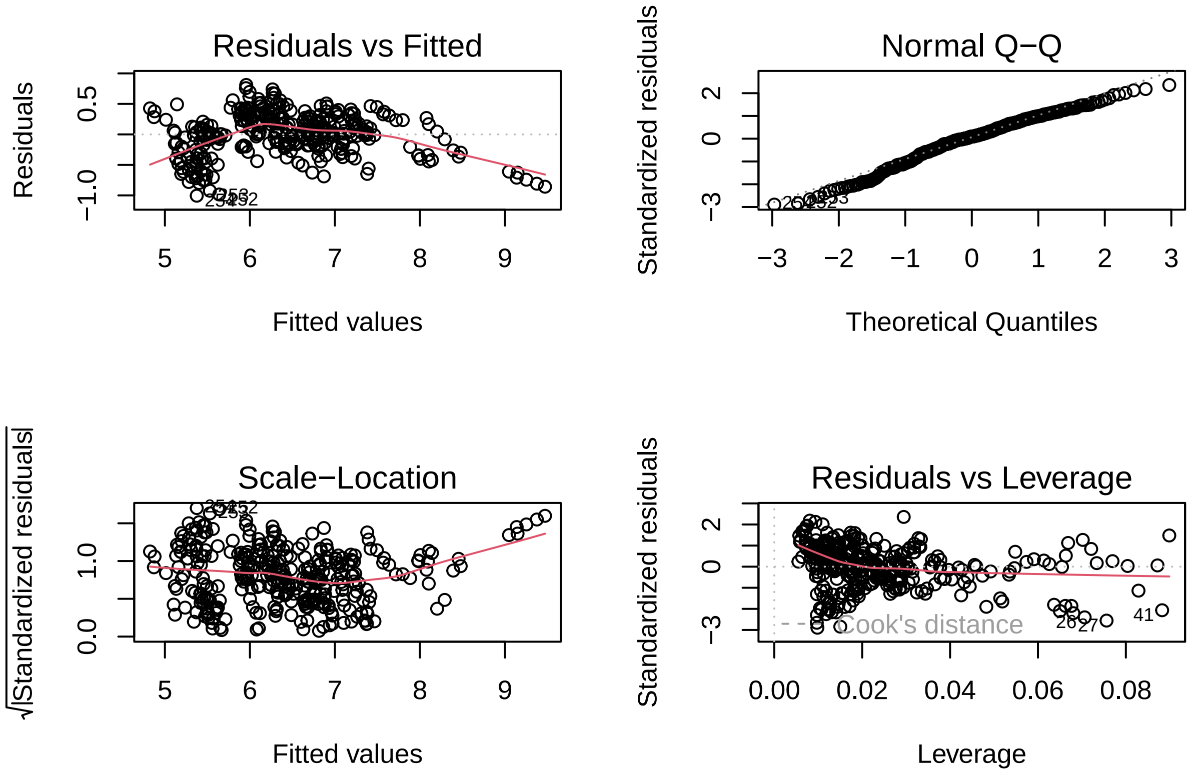
```
dwtest(mlrmodel5)
```

```
##
##  Durbin-Watson test
##
## data:  mlrmodel5
## DW = 0.40037, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

The test was significant so the model does not fit the no or little autocorrelation assumption.

Residuals: We will check residual assumptions by plotting our chosen model:

```
par(mfrow = c(2, 2))
plot(mlrmodel5)
```



The residual graphs show that the residuals are relatively normally distributed and that there aren't any extreme values. However, they also show that there is low linearity between the independent and dependent variables and that that the variances are relatively unequal. Overall the residual assumptions do not fully hold.

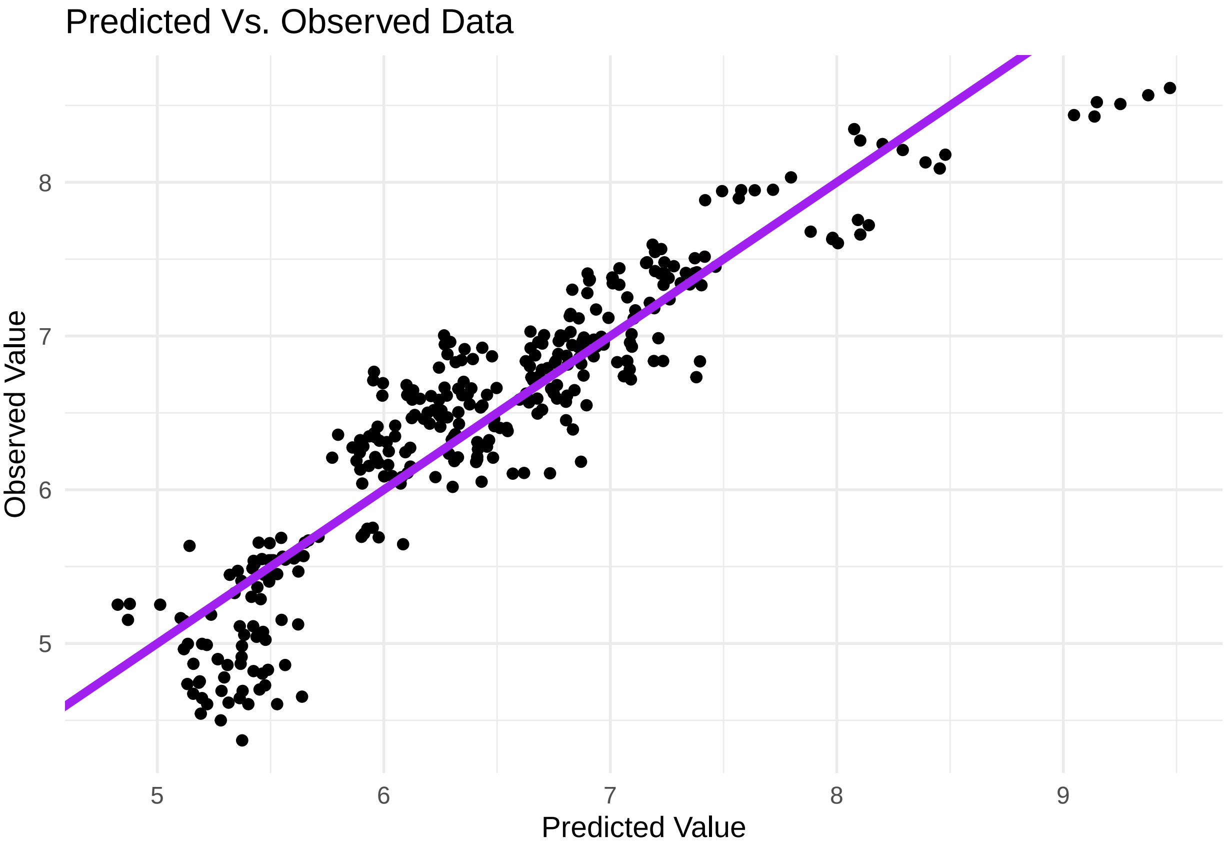
MULTIPLE LINEAR REGRESSION INTERPRETATIONS:
Due to the nature of the independent variables involved, the intercept of the model is of no pragmatic interest however the slopes of the model are. All of the slopes of model the are statistically significant at a 5% level of significance. The slopes of the model are all very small, meaning no one factor has a very large amount of predictive capability on its own. Unemployment, Income, Beertax, Percent of Southern Baptist, Mandatory Community Service, and Population are all predicted to be positively correlated with traffic

fatalities according to our model, in fact the only factor that is negatively correlated is the amount of Spirit Consumption. The majority of these associations might be surprising to the average person, the only ones we hypothesized would be positively correlated from the start were Population and Unemployment.

MULTIPLE LINEAR REGRESSION PREDICTIONS:

Next we use the model to make predictions of fatalities based on the observations of the independent variables in the dataset and compare them to the observed fatalities:

```
plot_data <- data.frame(predicted_value = predict(mlrmodel5),
    observed_value = log(fatals_clean$fatal))

ggplot(plot_data, aes(x = predicted_value, y = observed_value)) +
    geom_point() + labs(title = "Predicted Vs. Observed Data",
    x = "Predicted Value", y = "Observed Value") + geom_abline(intercept = 0,
    slope = 1, color = "purple", lwd = 1.5) + theme_minimal()
```



We see that the model does a relatively good job of predicting the data within the dataset despite several of our assumptions for multiple linear regression not holding.

Question 3:

CHI-SQUARED AND DIFFERENCE IN PROPORTIONS ASSUMPTIONS:

EDITS: The chi-square test that was on its own before has been removed

Chi-Square and Difference in Proportions tests both require two categorical variables that should be measured at an ordinal or nominal level. The two variables should consist of two or more categorical, independent groups. Our independent categorical variable corresponds to mandatory jail sentence laws and mandatory community service sentence laws. Both the variables are nominal that take on yes or no values. Our sample size is 335 observations and the expected frequencies for each cell is above 5 so it meets the assumptions for the chi-square section of the test.

H_o : There is no statistically significant correlation between mandatory jail sentence laws and mandatory community service sentence laws.

H_A : There is a statistically significant correlation between mandatory jail sentence laws and mandatory community service sentence laws.

```
tb2 <- table(fatals_clean$jail, fatals_clean$service)
prop.test(tb2, correct = FALSE, conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  tb2
## X-squared = 91.825, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.3472655 0.5578285
## sample estimates:
##    prop 1    prop 2
## 0.9419087 0.4893617
```

Because the p-value of the chi-square was well below our 5% level of significance, we can say that there is a statistically significant relationship between mandatory jail sentencing and mandatory community service sentencing.

According to the results of the difference in proportion: out of all of the states that do not have mandatory jail sentences ~94% also do not have mandatory community service, and out of all states that do have mandatory jail sentences ~49% do not have mandatory community service. So based on this data it appears that in a given state in a given year, it is very likely that if there is not mandatory jail sentences there are also not mandatory community service sentences, however if there are mandatory jail sentences there is only a slightly above 50% chance that there will also be mandatory community service sentences.

Section 4:

SUMMARY OF RESULTS: Our ANOVA model did not show a statistically significant difference in mean fatalities between US regions at our originally stated 5% significance level. After the data had been cleaned however, it did show a difference at a 10% level of significance, so we ran a Tukey post-hoc test at that 10% level of significance. The post-hoc test revealed that the only statistically significant difference in mean traffic fatalities was between the West and Southeast regions of the US, where Southeast had a greater mean. This result is also consistent with what can be seen visually in the boxplots and barplots of our EDA. Our multiple linear regression model appears to predict the values within our dataset relatively well and has slope values that are very interesting from cultural and political perspectives. However, it fails to fulfill several major assumptions such as normality of dependent variable, low autocorrelation, and equal variances, so it would be very risky to use the model to make real world predictions and any interpretations gained from the formula of the model are suspect. Our chi-square test indicated a statistically significant correlation between mandatory jail sentences and mandatory community service sentences, and after performing a post-hoc difference in proportion test we learned that a lack of mandatory jail sentences and a lack mandatory community service sentences are very highly correlated, and that the prescence of mandatory jail sentences and the prescence of mandatory community service sentences are only slightly correlated.

FINDINGS RELATIVE TO EXISTING LITERATURE: The findings of the paper our dataset has been derived from and the findings from our analysis are somewhat consistent. The paper concluded that the impact from beer tax and “administrative per se laws”, mandatory jail and mandatory community service, have a positive correlation with vehicle fatalities. In the paper, legal drinking ages were found to be strongly negatively related to the fatalities of 18 to 20 year olds. This was different from our findings as our analysis concluded that legal drinking age was statistically insignificant. Additionally, some of the variables compared in the paper and in our analysis were different. In the paper, the dram shop laws and Mothers Against Drunk Driving were compared as well. Dram shop laws were found to have a statistically significant negative impact on traffic mortality. In our analysis, we found that Percent of Southern Baptists and Population were positively correlated with traffic fatalities.

The paper from National Center for Statistics and Analysis on alcohol-impaired driving talks about motor vehicle crashes that involve an alcohol-impaired driver. This is relevant to our paper as some of the variables that are statistically significant are alcohol related like Beertax. This shows that alcohol and fatalities are positively correlated. The paper concludes that about 30 percent of all traffic fatalities are caused because of drunk driving. It also found that young drivers and night time accidents are positively correlated with fatal crashes. In our paper, we didn't include night fatalities (nfatal) in our model because fatalities and night fatalities are correlated.

SUGGESTIONS FOR FUTURE RESEARCH: The majority of our uncertainty in regards to this project is from the multiple linear regression model. Having a larger number of observations with more independent variables would allow more precise analysis of the problems we are interested in. It would also allow us to narrow the range of data we examine at once which might lead to more normal and/or more consistent data distributions. Our dependent variable for the first two questions was not normally distributed in its unaltered state which is likely to have been the cause of many of our issues. A statistical test more suited to the natural distribution of the data may give more reliable results.

A paper from the journal Accident Analysis & Prevention titles "Traffic fatalities and economic growth" addresses two of these concerns by having a much larger dataset and narrowing their tests, and as a result found that income was a negatively correlated to traffic fatalities after a certain point, which is opposite of what we found. This is because "Traffic fatalities and economic growth" compared multiple countries including low-income (developing) countries whereas in our paper, we compare only the data from the United States.

WORKS CITED:

Lee, S., & Lee, D. K. (2018). What is the proper way to apply the multiple comparison test?. Korean journal of anesthesiology, 71(5), 353–360. https://doi.org/10.4097/kja.d.18.00242

Ruhm, C. J. (1996). Alcohol policies and highway vehicle fatalities. Journal of Health Economics, 15(4), 435–454. https://doi.org/10.1016/s0167-6296(96)00490-0

Symonds, M. R. E., & Moussalli, A. (2010, August 25). A brief guide to model selection, Multimodel inference and model averaging in behavioural ecology using Akaike's information criterion - behavioral ecology and sociobiology. SpringerLink. Retrieved November 14, 2022, from https://link.springer.com/article/10.1007/s00265-010-1037-6

Kopits, E., & Cropper, M. (January 2005). Traffic Fatalities and Economic Growth. Accident Analysis & Prevention. Retrieved November 15, 2022 from https://www.sciencedirect.com/science/article/pii/S0001457504000685

Traffic Safety Facts. National Center for Statistics and Analysis. (2022, April). Retrieved November 15, 2022, from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813294