SYS 4021 Project 1: Multiple Linear Regression

Aatmika Deshpande, Harish Karumuri, Khin Kyaw, Grace Parzych

10/21/2020

Problem Background

Below, we load in the initial dataset provided, which contains longitudinal data on historical train accidents from the Federal Railroad Administration between 2001-2019.

```
acts = file.inputl(traindir)
totacts = combine.data(acts)
```

Casualties is not a variable in the initially loaded dataset; this is a variable that we created, summing up the total killed and total injured columns to create an aggregate ‘Casualty’ column.

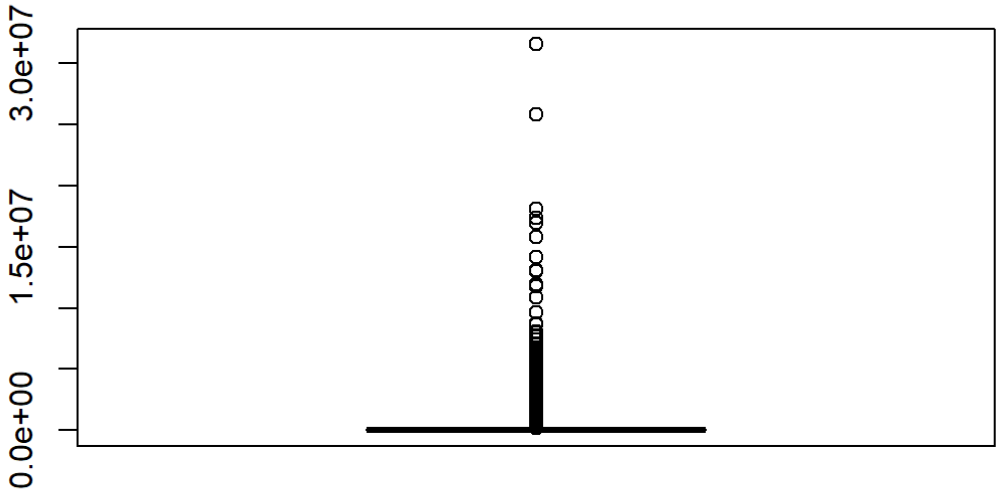
```
totacts$Casualty = totacts$TOTKLD + totacts$TOTINJ
```

Accident Damage

Exploratory Analysis

Cleaning up the Data

To begin our initial exploratory analysis and data visualization so that optimal variables may be picked for linear regression, we must first subset our larger dataframe to consist of just extreme accidents, as well as clean the data by recognizing and removing outlier points and duplicate entries. To do this, a boxplot was first generated of totacts, the train dataset, in order to visually see all the outlier accidents that we will be subsetting for and considering ‘extreme’.



Boxplot of ACCDMG for all Accidents

Next, the data was subsetted. Recognizing that this new extreme accidents (xdmg) dataset has an entry for 9/11, which has values that are much larger in magnitude than the rest of the entries in this dataset, this entry, at index 186, was removed. Finally, all duplicate entries were removed.

```
xdmg = totacts[totacts$ACCDMG > dmgbox$stats[5],]
xdmg = xdmg[-186,]
xdmg_nd = xdmg[!(duplicated(xdmg[, c("INCDTNO", "YEAR", "MONTH", "DAY",
                                     "TIMEHR", "TIMEMIN")])),]
```

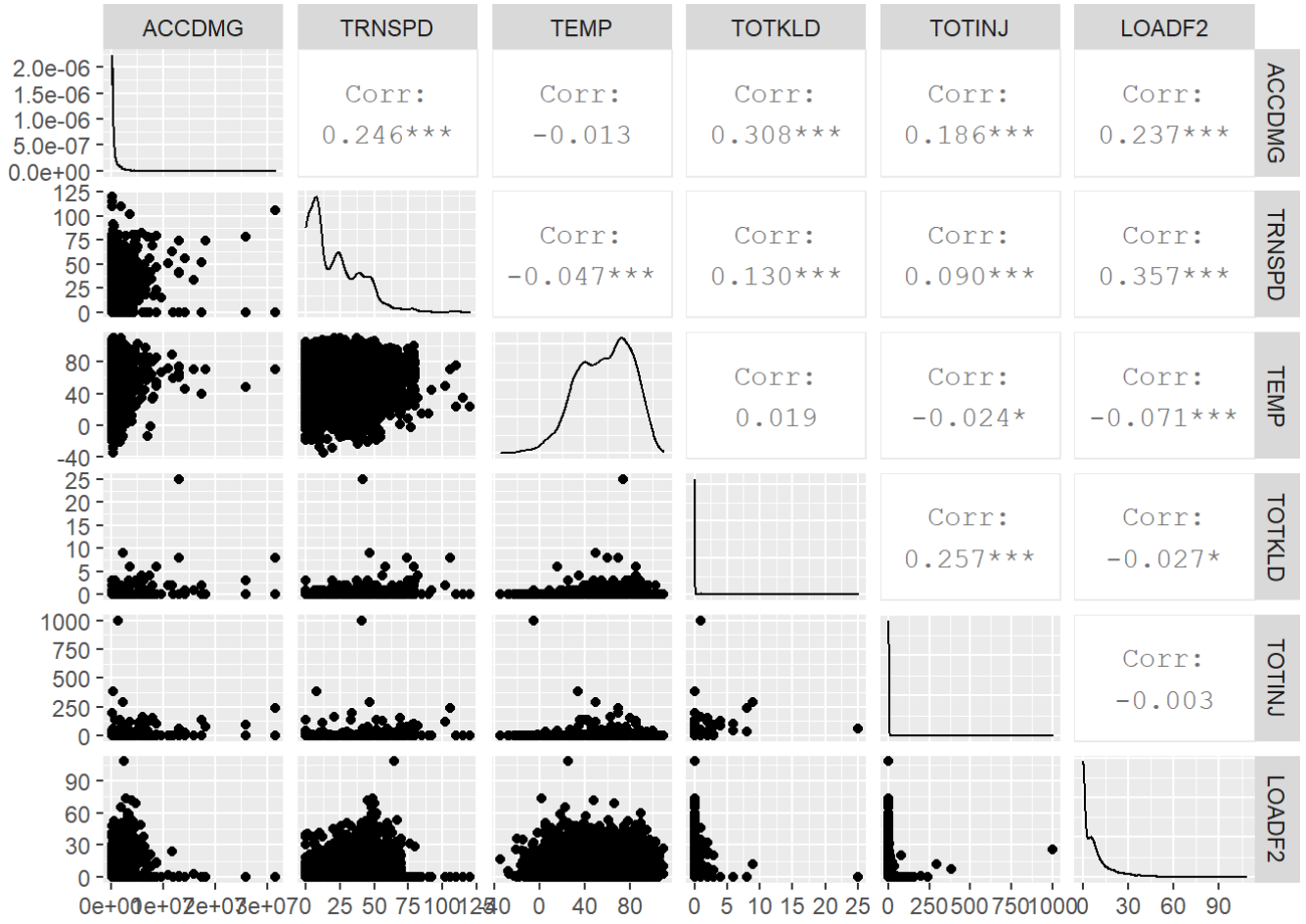
A Cause variable was created to better understand the root cause of all train accident entries in the dataset. M represents miscellaneous error; T represents track/rack, roadbed, and communications error; S represents signal and communication error; H represents human factors error; and E represents mechanical and electrical error.

```
xdmg_nd$Cause <- rep(NA, nrow(xdmg_nd))
xdmg_nd$Cause[which(substr(xdmg_nd$CAUSE, 1, 1) == "M")] <- "M"
xdmg_nd$Cause[which(substr(xdmg_nd$CAUSE, 1, 1) == "T")] <- "T"
xdmg_nd$Cause[which(substr(xdmg_nd$CAUSE, 1, 1) == "S")] <- "S"
xdmg_nd$Cause[which(substr(xdmg_nd$CAUSE, 1, 1) == "H")] <- "H"
xdmg_nd$Cause[which(substr(xdmg_nd$CAUSE, 1, 1) == "E")] <- "E"

xdmg_nd$Cause = as.factor(xdmg_nd$Cause)
```

Quantitative Variables

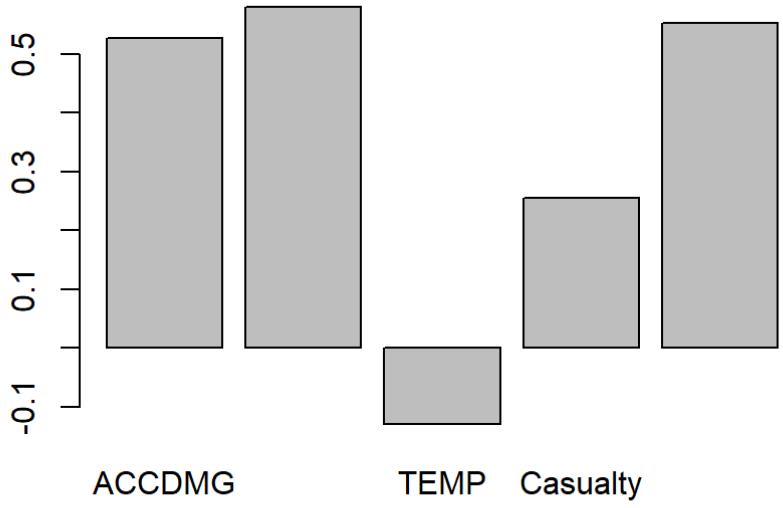
Variable analysis was split up into quantitative and qualitative variables. First, a scatter plot matrix was generated with some quantitative variables from the dataset that we believed could be important, along with ACCDMG, to see which were the most correlated with our response variable, ACCDMG. Of the 5 predictors that were chosen (train speed, temperature, total killed, total injured, and the number of loaded freight trains), total killed had the highest correlation of 0.308, followed by train speed at 0.246.



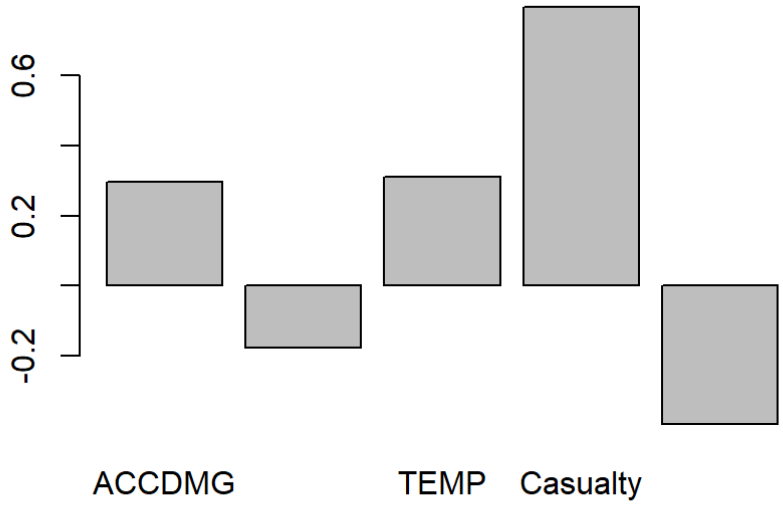
Scatterplot Matrix of ACCDMG with Quantitative Variables

PCA loading plots using the correlation matrix were also generated for the first 2 PCs to observe which variables contributed largest in the first and second directions of greatest variability. Train speed and loaded freight trains had the highest loading magnitudes, and varied in the same direction as ACCDMG in the first PC. The second PCs loadings plot was a lot less telling.

```
pca.xdmg.corr = princomp(xdmg_nd[,c("ACCDMG", "TRNSPD", "TEMP",
                                    "Casualty", "LOADF2")], cor=T)
```



First PC Loadings for xdmg_nd

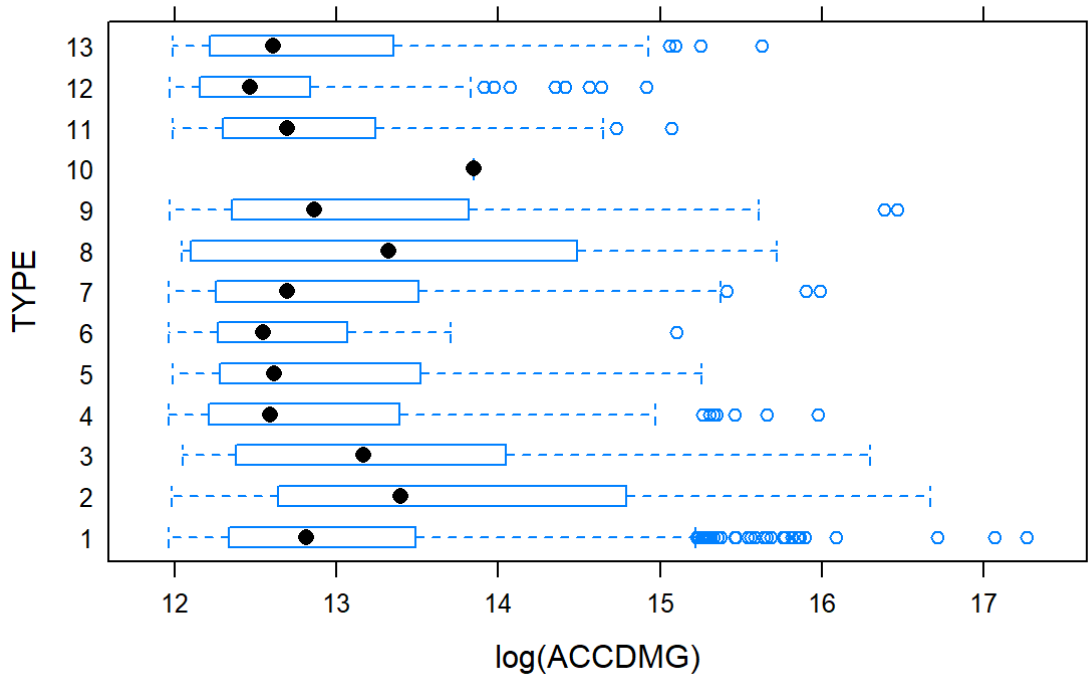


Second PC Loadings for xdmg_nd

Given the two visualizations made, TRNSPD was selected as the quantitative variable of to be used in the ACCDMG linear model because, while it did not have the highest correlation value with ACCDMG, it was the most actionable. In other words, knowing that the number of killed is highly correlated with ACCDMG is good insight, however there is not much that can be said recommendation wise in terms of reducing the number of killed, as we are still not able to pinpoint that cause. Thus, TRNSPD was selected.

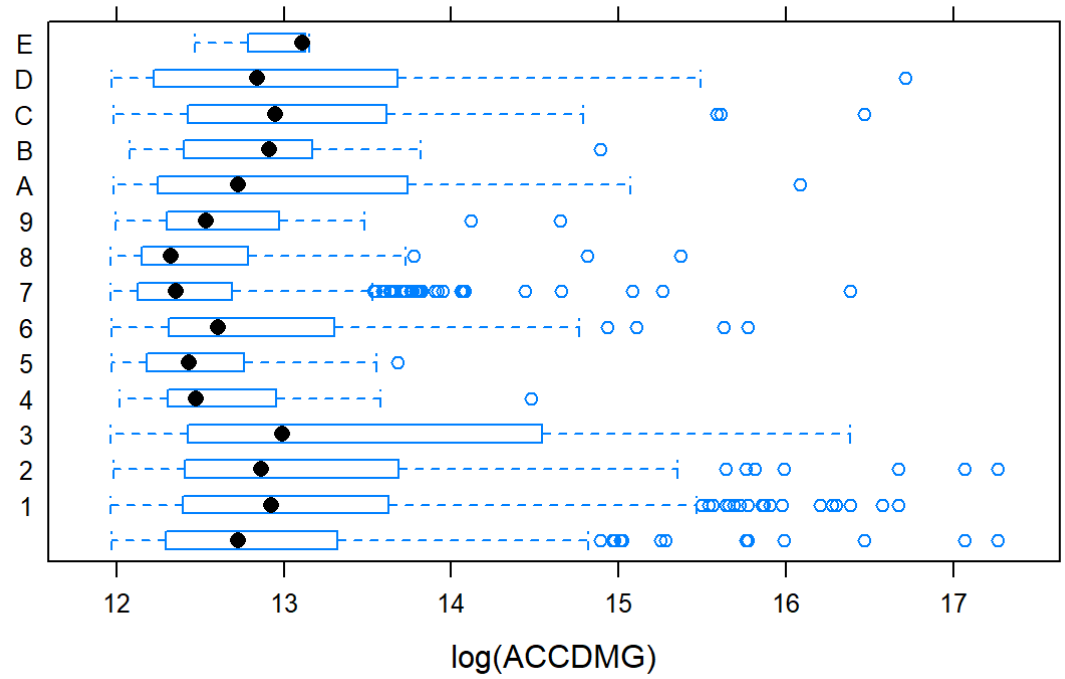
Qualitative Variables

Various qualitative variables were observed and considered for use in this model and for generating hypotheses. TYPE (type of accident), TYPEQ (type of train), WEATHER (weather at the time of accident), VISIBLTY (visibility at the time of accident), and Cause (cause of accident) were considered.



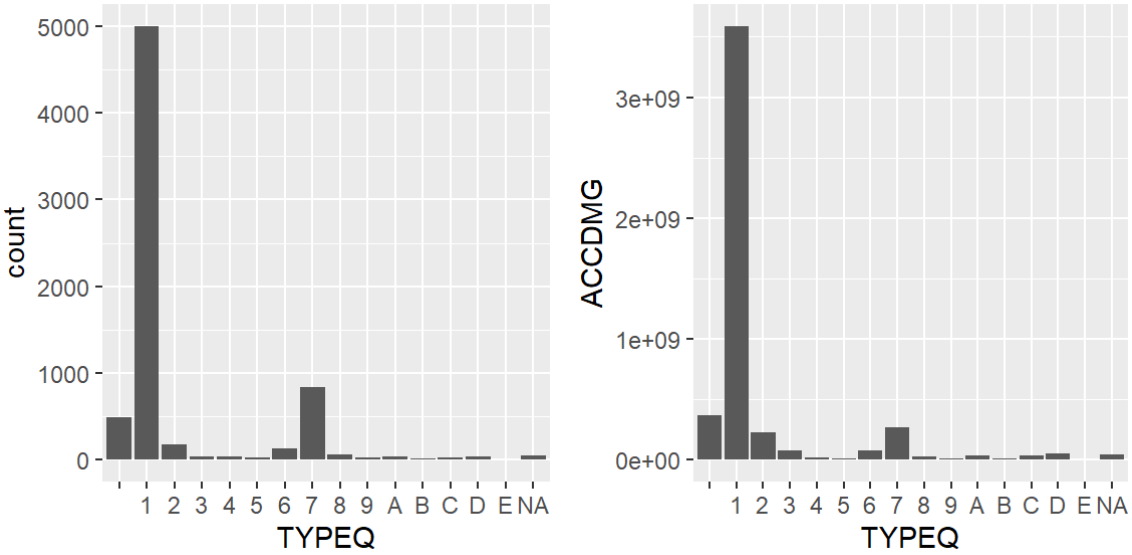
Factor Boxplot of TYPE with log(ACCDMG)

A factor boxplot with log(ACCDMG) on the x-axis and type of accident on the y-axis was plotted. Based on the visualization, one can see that Type 1, derailment, had the largest number of extreme/outlier accidents.



Factor Boxplot of TYPEQ with log(ACCDMG)

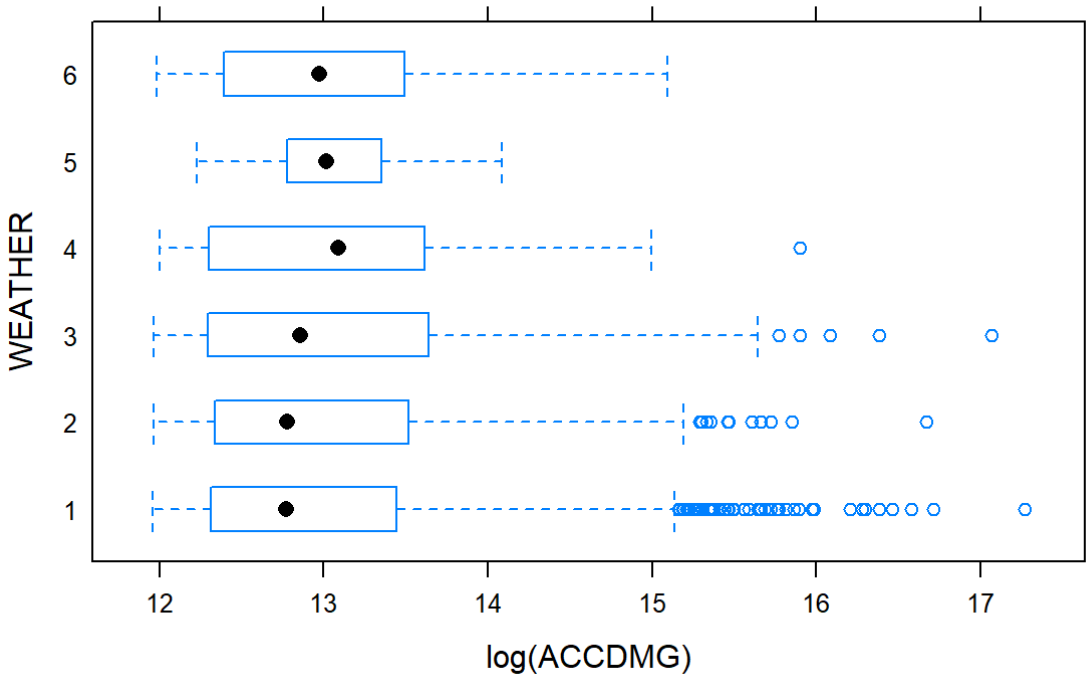
A similar plot was generated for TYPEQ; from here it can be observed that 1 and 7 had the largest number of extreme/outlier points in respect to ACCDMG, even if their median ACCDMG was consistent with the rest of the train types. This led us to generate bar plots with counts and ACCDMG on the y-axis. It can be seen that type 1, freight, has both the highest count and highest total accident damage, meaning this accident is very common, and as a result, although it may not cost much per accident, aggregately has high damage costs. Thus, it was decided to focus on Freight trains as a trigger for accident damage. In order to adhere towards Ocam's razor, the 9 factors of TYPEQ were recoded to Freight being 1 and all other levels the base case of 0.



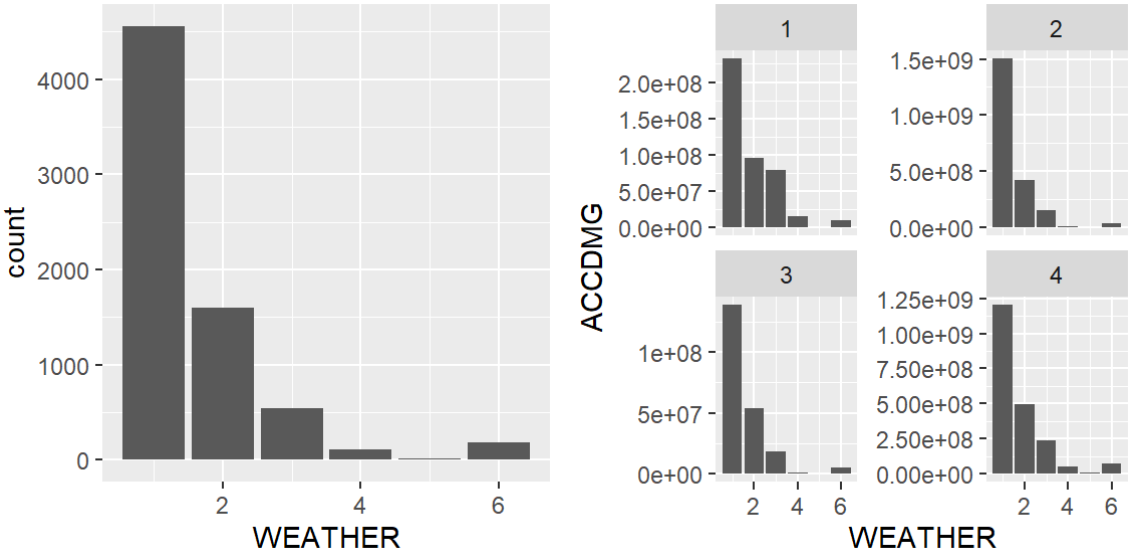
Histogram and Barplot of TYPEQ

```
xdmg_nd$Freight = ifelse(xdmg_nd$TYPEQ==1, 1, 0)
```

Weather was the next variable considered. Initially a boxplot was created, leading to further interest and barplots generated, one with counts and the other with ACCDMG faceted by visibility.



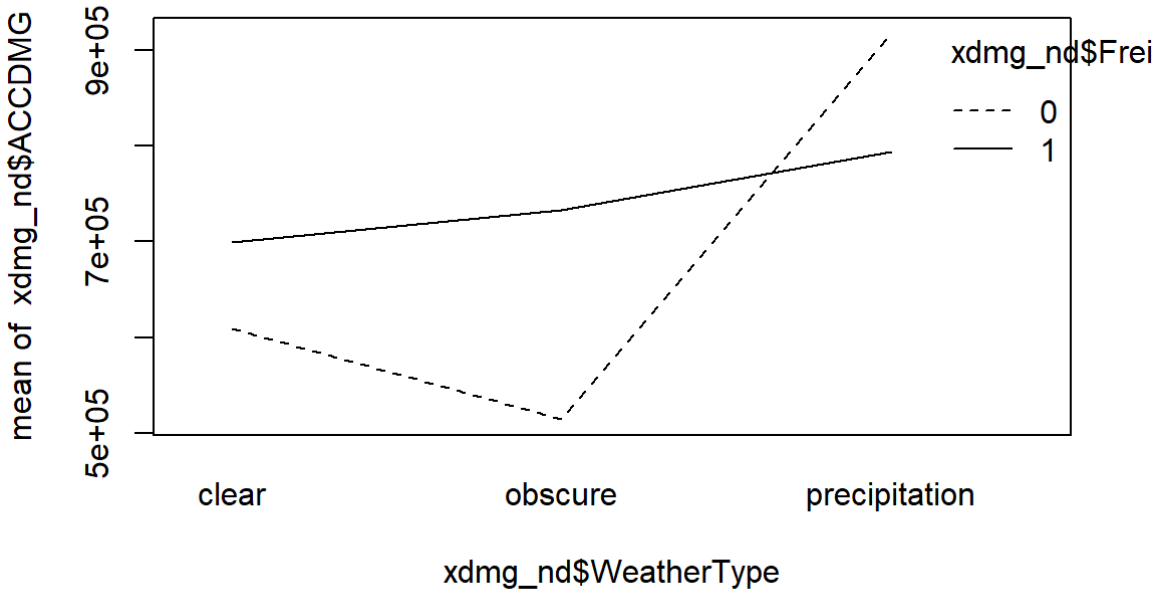
Factor Boxplot of WEATHER with log(ACCDMG)



Histogram and Barplot of WEATHER

After observing a clear impact of weather on ACCDMG, it was concluded that using the WEATHER variable as is had a large possibility to generate a lot of noise in the model and have larger numbers of parameters unnecessarily. A more concise weather variable, WeatherType, was thus created. It was factored to be simplified from 6 groups to 3 groups based on the characteristics of different weather conditions. Clear was used as its own level, fog or cloudy was merged into a level due to obscured light levels, and rain, sleet, and snow was merged to as they are all forms of precipitation.

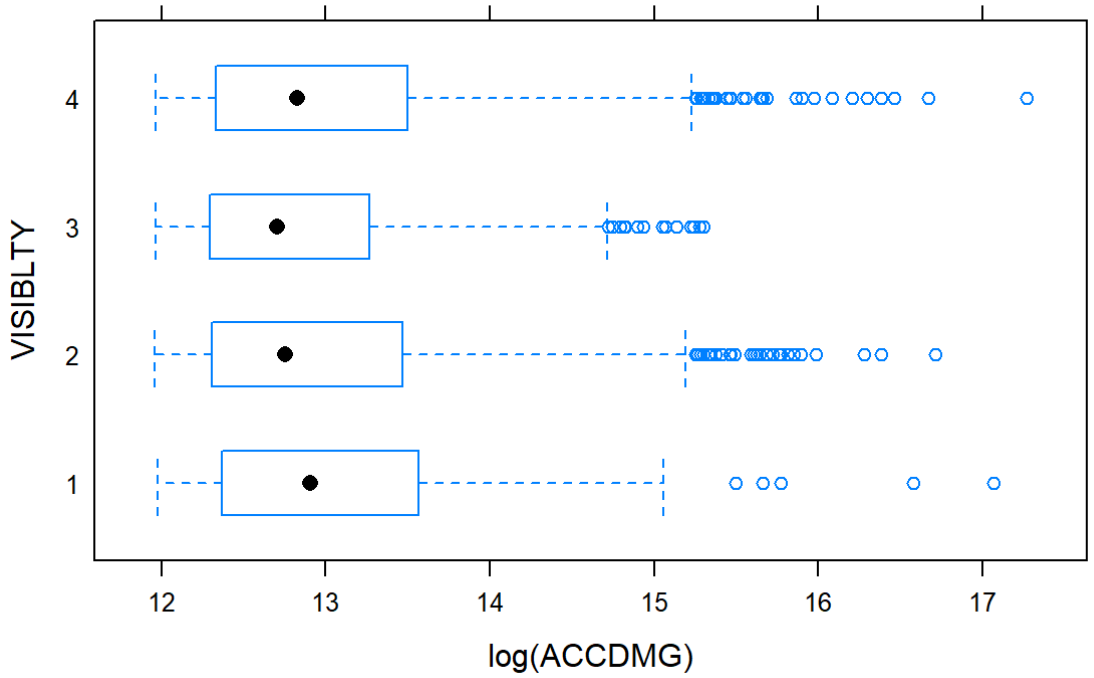
```
xdmg_nd = xdmg_nd %>% mutate(WeatherType = case_when(
  WEATHER == 1 ~ "clear",
  WEATHER %in% c(2, 4) ~ "obscure",
  WEATHER %in% c(3, 5, 6) ~ "precipitation"))
xdmg_nd$WeatherType = as.factor(xdmg_nd$WeatherType)
```



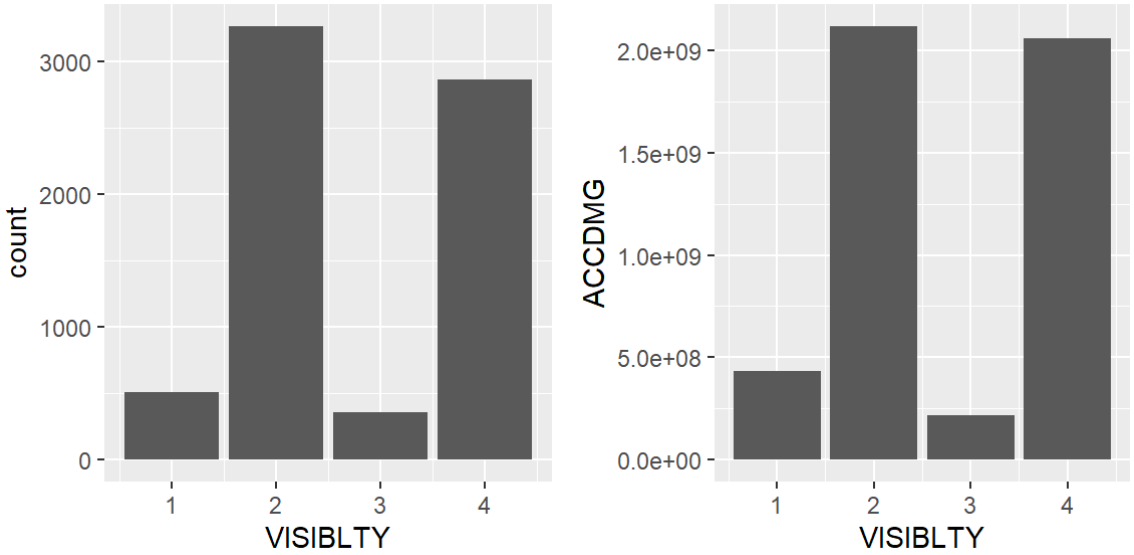
Interaction Plot of WeatherType and Freight with ACCDMG

This interaction plot shows a clear need for at least one interaction term between WeatherType and Freight given the stark slope differences between each of the 3 weather types for whether a train is Freight or not, and the intersection of these lines between obscure and precipitation.

Visibility corresponds to day, night, dusk and dawn for when the accident occurred. Similar to above, a factor boxplot was made, and then barplots were generated.



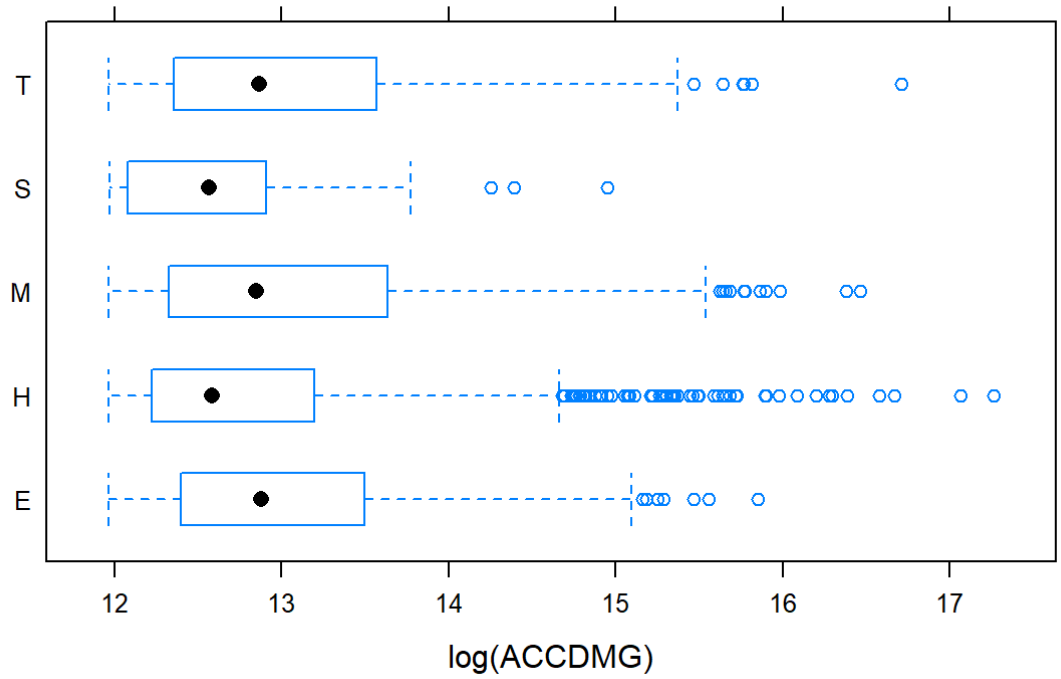
Factor Boxplot of VISIBLTY and log(ACCDMG)



Histogram and Barplot of VISIBLTY

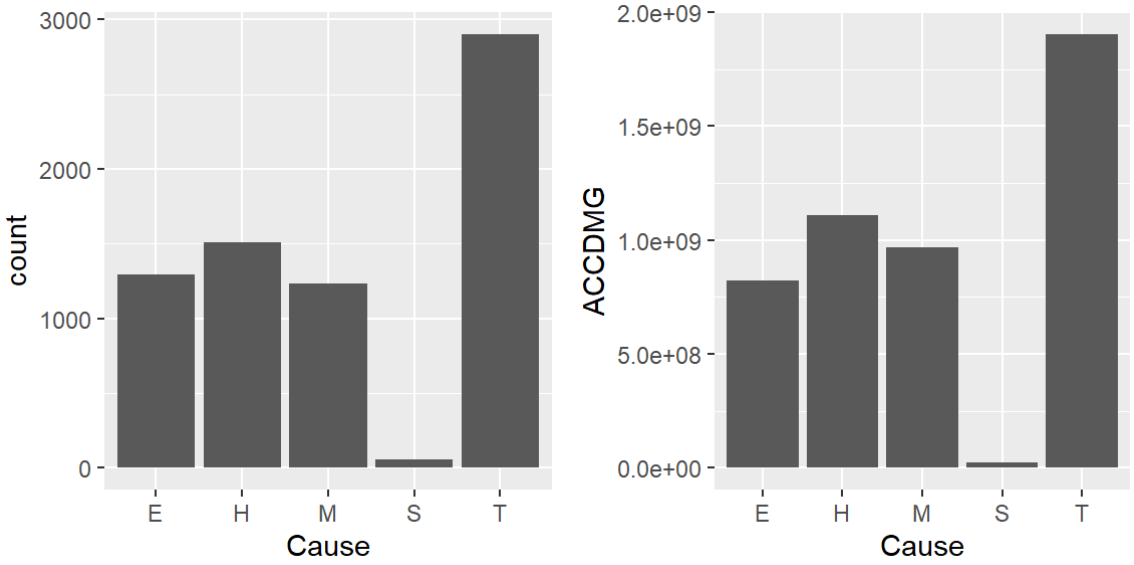
It seems the highest counts and damages occurred during visibility 2 and 4, which was day and night. This wasn't necessarily something we believed was unexpected, given that it makes sense most trains would run at these larger time spans versus the smaller and more obscure durations of dusk and dawn.

Cause was the last qualitative variable considered for ACCDMG.



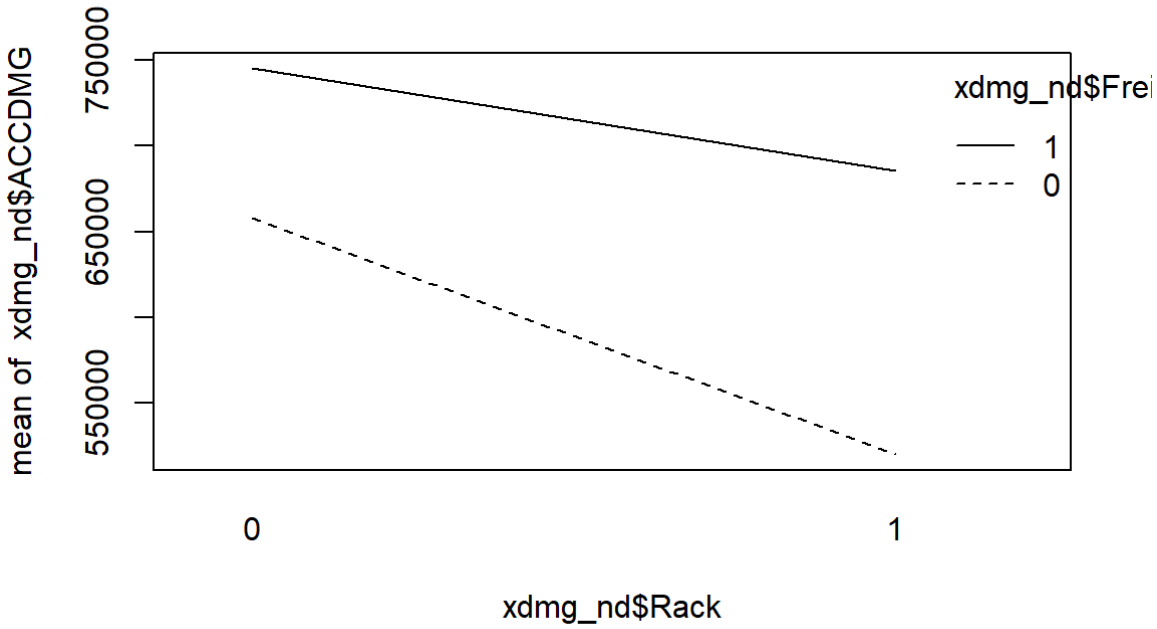
Factor Barplot of Cause with log(ACCDMG)

Similar to the instance with Freight, we observe that Cause T has both the highest count/frequency of accidents, as well as the highest accident damage, meaning it occurs often and is a cause we should look into addressing. Similar to Freight, the 5 factors of Cause were recoded to Rack, Roadbed, and Structure (T) as 1, and all other levels as the base case of 0.

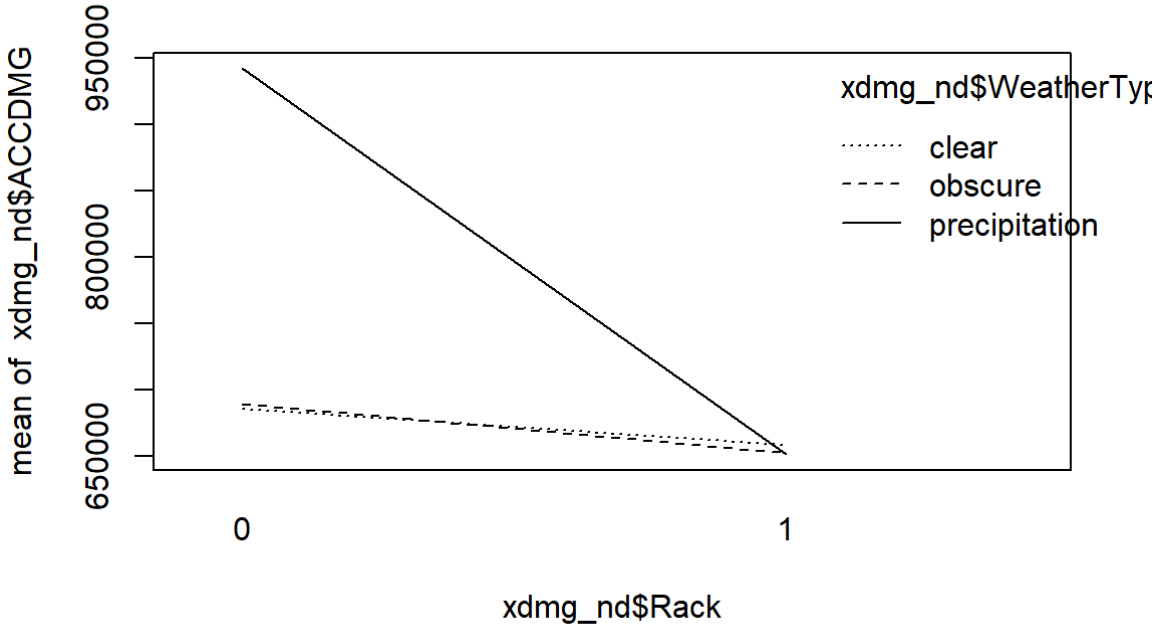


Histogram and Barplot of Cause

```
xdmg_nd$Rack = ifelse(xdmg_nd$Cause == "T", 1, 0)
```



Interaction Plot of Rack and Freight with ACCDMG



Interaction Plot of Rack and WeatherType with ACCDMG

Both interaction plots between Rack and Freight and Rack and WeatherType showed pretty consistent trends, and suggested the lack of need for an interaction between them. One's result in ACCDMG is not affected by the value of another.

Generating Hypotheses

Based on all the data visualization that was conducted above, our first hypothesis aims to identify if the inherent type of train poses a higher risk in terms of accident damage – based on the results of the experiment, we can recommend a process to further understand the specifics of why such a train type poses a capital risk. This hypothesis was reached given that the boxplot of TYPEQ with ACCDMG showed that although the median value of this type of train was generally around that of the other types of trains, there were a good number of outlier accidents that are resulting in high accident damages. This is indicative of potentially being a predictor variable and correlated with the response. Additionally, the histogram and barplots showed both high counts and high summations of accident damage, meaning that while each of the individual accidents may not account for much damage, the high number of accidents causes high aggregations. Thus, it would be important to look further into this variable. The null hypothesis is **Freight Trains do not increase the severity of ACCDMG relative to other types of trains**. the alternative hypothesis is **Freight**

Trains do increase the severity of ACCDMG relative to other types of trains. This hypothesis is actionable because once we are aware whether or not having a freight train is truly correlated to causing higher accident damage, we can look into potentially considering solutions to this issue. For example, considering alternative modes of transportation such as truck, or looking further into what makes freight trains explicitly different from other types and thus why these train types cause accidents more often.

The second hypothesis aims to see if the Rack, Roadbed, and Structures related cause further capital damages as well – a rejection of the null could help guide whether a maintenance process is needed on this underlying infrastructure, while a failure to reject would indicate a necessity to study other causes of accidents. Rack was picked as the other variable for our second hypothesis due to similar reasoning as the Freight train type. When looking at the histogram and barplots of Cause with ACCDMG, there were both high counts and high accident damage values. The null hypothesis is **Rack, roadbed, and structure causes do not increase the severity of ACCDMG relative to other kinds of track types**. The alternative hypothesis is **Rack, roadbed, and structure causes do increase the severity of ACCDMG relative to other kinds of track types**. This hypothesis is actionable because the rack, roadbed, and structure issues are something that are easily targetable. Things that go into this include track geometry, frog switches, etc. Looking into more regular maintenance or more heavy monitoring of the tracks in which this accident can occur could be looked into.

Initial Linear Model

Our model selection process consisted of an iterative process that involved creating full and smaller second order models, stepwise models, and determining which model provided us with more information. We approached the model selection process by creating a full second order model with our selected variables: trainspeed, freight, weather, and rack. Then, we decided to create a stepwise model to determine if this model was a better representation of our data and determined that the stepwise model was a better fit. Additionally, we created a smaller model that contained freight train, speed, weather, rack, and an interaction term between weather and freight. Upon comparing the small model to its stepwise model, the stepwise model result was better. Finally, we tested the full second order model against the smaller stepwise model and determined that the full second order stepwise model was overall the best model.

For every model we used to test our hypotheses, we used t-tests to determine if the specific terms (either freight or rack or interaction terms with them) were significant. If the resulting p-values led us to believe the term was significant, we would reject the null hypothesis if the coefficient was positive. We also used the global utility test on each model to determine if the model itself was significant.

```
xdmg.lm1 = lm(ACCDMG~(Freight + TRNSPD + WeatherType + Rack)^2,
              data=xdmg_nd)
xdmg.lm1.step = step(xdmg.lm1, trace=F)

#lm1.step is better than the lm1
anova(xdmg.lm1, xdmg.lm1.step)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	6923	9.360891e+15	*NA*	*NA*	*NA*	*NA*
2	6924	9.363584e+15	-1	-2.693183e+12	1.991788	0.158199

2 rows

```
summary(xdmg.lm1.step)
```

```
##
## Call:
## lm(formula = ACCDMG ~ Freight + TRNSPD + WeatherType + Rack +
##     Freight:TRNSPD + Freight:WeatherType + TRNSPD:WeatherType +
##     TRNSPD:Rack + WeatherType:Rack, data = xdmg_nd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3010540  -365646  -189965    56745 31105534
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    433220.23   38836.45  11.155  < 2e-16 ***
## Freight                        -74245.33   50563.49  -1.468 0.142052
## TRNSPD                          17837.88    1492.64  11.951  < 2e-16 ***
## WeatherTypeobscure             -27654.85   71021.19  -0.389 0.697001
## WeatherTypeprecipitation       323395.36   97561.85   3.315 0.000922 ***
## Rack                          -156086.08   49855.84  -3.131 0.001751 **
## Freight:TRNSPD                  -4000.47    1799.80  -2.223 0.026266 *
## Freight:WeatherTypeobscure      87237.03   82662.80   1.055 0.291309
## Freight:WeatherTypeprecipitation -292183.67  112332.03  -2.601 0.009313 **
## TRNSPD:WeatherTypeobscure       -1194.90    2067.54  -0.578 0.563329
## TRNSPD:WeatherTypeprecipitation  7177.40    2854.91   2.514 0.011958 *
## TRNSPD:Rack                      6952.18    1697.03   4.097 4.24e-05 ***
## WeatherTypeobscure:Rack           -52.48   68167.10  -0.001 0.999386
```

```
## WeatherTypeprecipitation:Rack    -226933.93   96567.06  -2.350 0.018801 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1163000 on 6924 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.07182,    Adjusted R-squared:  0.07008
## F-statistic: 41.21 on 13 and 6924 DF,  p-value: < 2.2e-16
```

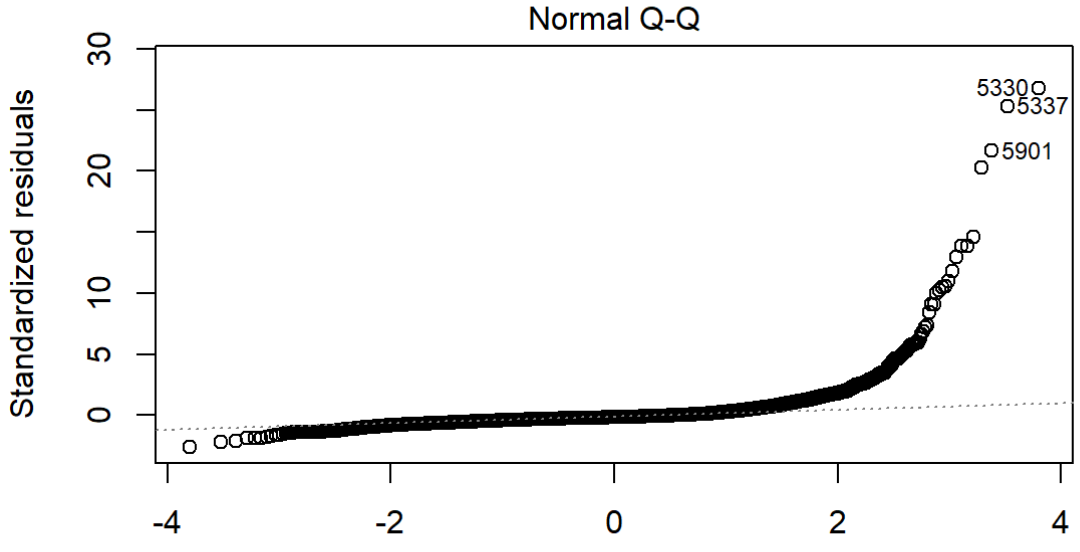
```
AIC(xdmg.lm1.step)
```

```
## [1] 213503.3
```

As described above, we created and tested multiple models before selecting the best one. The first transformations and testing we did was performing stepwise regression on two different sized models. For the full second order model, the step model generated from it was better given the p-value larger than .05 from the anova F test, which means that all extra coefficient values are equal to zero and we accepted the null for the Partial F test.

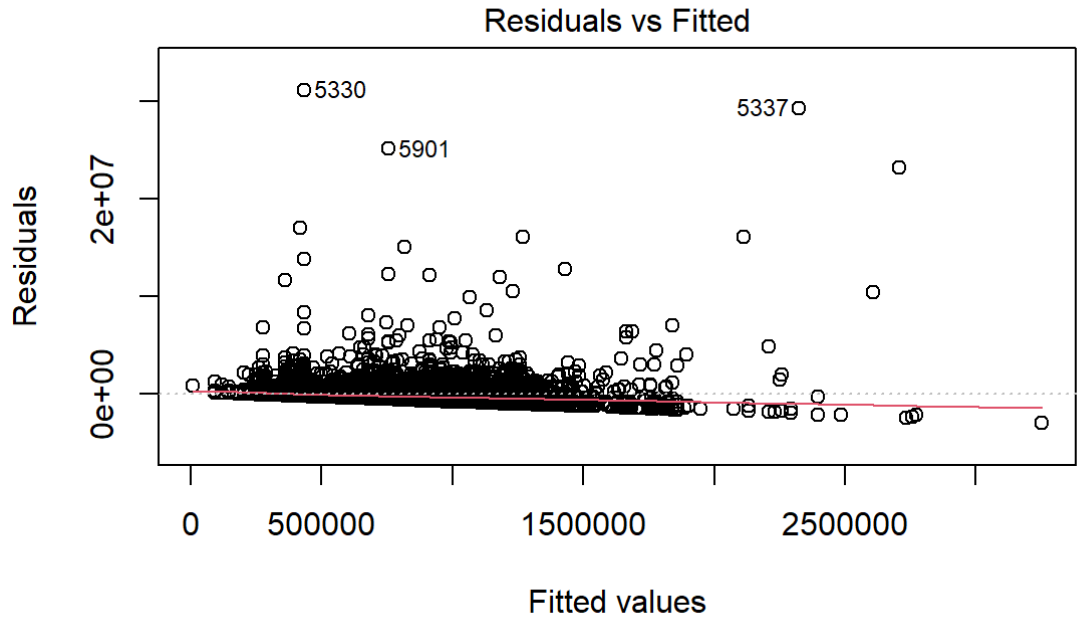
The first model (xdmg.lm1.step) we fully tested for these hypotheses was a full second order stepwise regression model using the exploratory variables: Freight, TRNSPD, WeatherTypeobscure, WeatherTypeprecipitation, Rack, Freight:TRNSPD, Freight:WeatherTypeobscure, Freight:WeatherTypeprecipitation, TRNSPD:WeatherTypeObscure, TRNSPD:WeatherTypeprecipiation, TRNSPD:Rack, WeatherTypeobscure:Rack, and WeatherTypeprecipiation:Rack. This model was significant based on the global utility test. After running t-tests, we found that Freight wasn't significant. The interaction term between freight and trainspeed was significant but had a negative relationship with accident damage. The interaction between freight and weather type precipitation was significant but also has a negative relationship with accident damage. Rack and the interaction term between rack and precipitation were significant and both had a negative relationship with accident damage. The interaction term between rack and trainspeed was significant and had a positive relationship with accident damage. To further test the adequacy of xdmg.lm1.step, we found the adjusted R2 and AIC, 7.008% and 213503.3 respectively. The adjusted R2 shows that our model isn't very accurate at predicting the actual accident damage of an event. The AIC is also very large which causes concern for this model. After this analysis, we used diagnostic plots to more adequately check our model. This led us to make transformations leading to our second model.

Diagnosing Model Problems

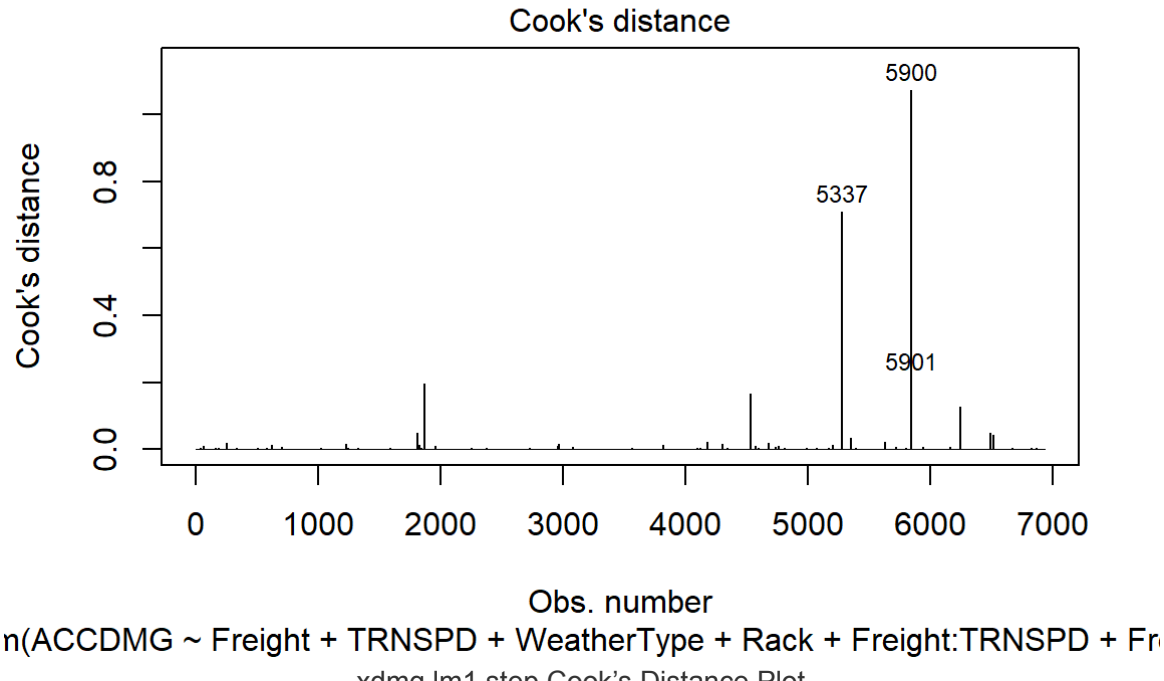


n(ACCDMG ~ Freight + TRNSPD + WeatherType + Rack + Freight:TRNSPD + Fr

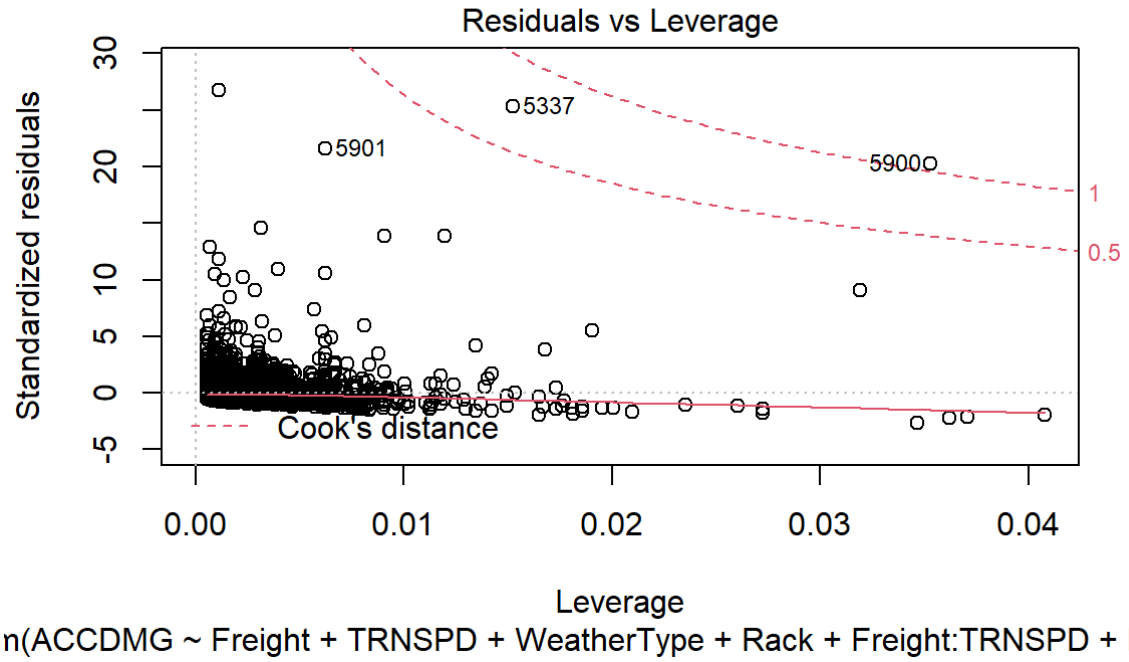
xdmg.lm1.step QQ Plot



xdmg.lm1.step Residuals vs. Fitted Plot



xdmg.lm1.step Cook's Distance Plot



xdmg.lm1.step Residuals vs. Leverage Plot

First, we made a QQ plot for xdmg.lm1.step. This model showed a violation of the normality assumption. We also ran a Residual vs Fitted graph; this graph showed that, while the model barely met the lack of fit assumption, it violated the constant variance assumption. We also ran leverage plots, such as cook's distance, which showed that the observations 5900 and 5337 were influential points.

To account for these leverage points, our next step in diagnosing problems was to remove observations 5900 and 5337 from the original data. We then created a new full second order model with the new data and a stepwise regression of that model. The stepwise model (xdmg.lm1_2.step) was a better model than the original, so we did initial analysis on the model and ran diagnostic plots.

```
#remove leverage points 5900, 5337
xdmg.lm1_2 = lm(ACCDMG~(Freight + TRNSPD + WeatherType + Rack)^2,
                data=xdmg_nd[-c(5900,5337),])
xdmg.lm1_2.step = step(xdmg.lm1_2, trace=F)
#lm1_2.step is better than the lm1
anova(xdmg.lm1_2, xdmg.lm1_2.step)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	6921	7.927665e+15	*NA*	*NA*	*NA*	*NA*
2	6926	7.930873e+15	-5	-3.207182e+12	0.559986	0.730791

2 rows

```
summary(xdmg.lm1_2.step)
```

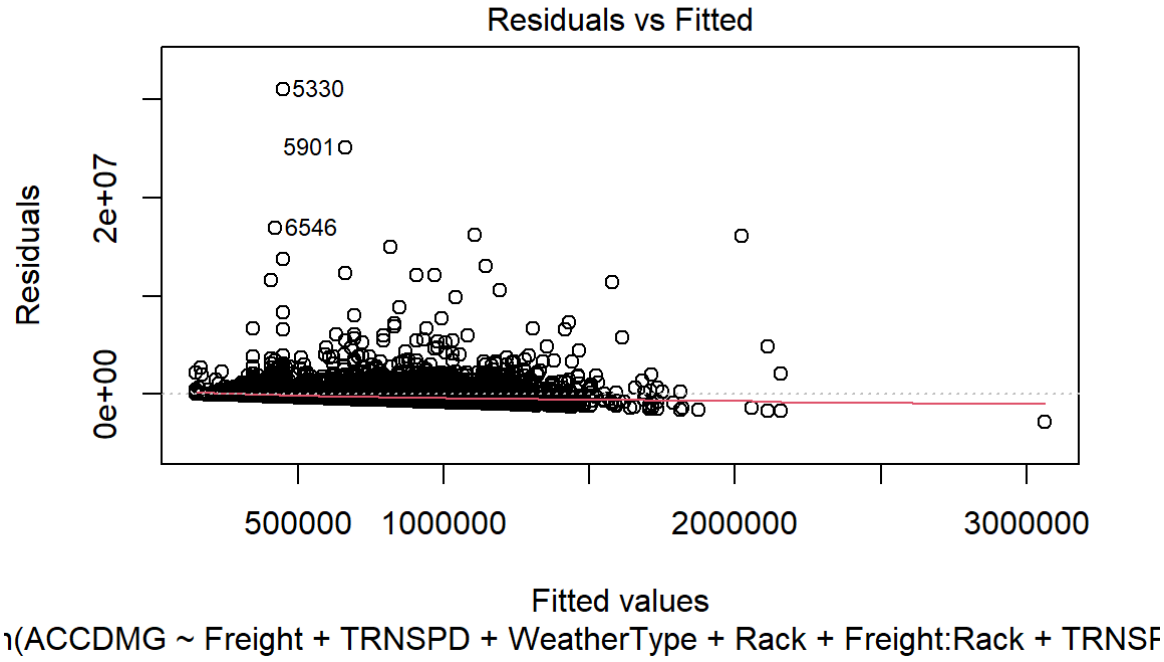
```
##
## Call:
## lm(formula = ACCDMG ~ Freight + TRNSPD + WeatherType + Rack +
##     Freight:Rack + TRNSPD:Rack + WeatherType:Rack, data = xdmg_nd[-c(5900,
##     5337), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2821846  -364023  -194678    55465 31091666
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    447088.1    32723.4  13.663  < 2e-16 ***
## Freight                        -39269.3    37559.5  -1.046   0.2958
## TRNSPD                          12421.6      942.1  13.186  < 2e-16 ***
## WeatherTypeobscure              12419.3    40257.1   0.308   0.7577
## WeatherTypeprecipitation       212780.3    54534.0   3.902 9.64e-05 ***
## Rack                          -101210.2    59403.4  -1.704   0.0885 .
## Freight:Rack                  -159432.8    66969.6  -2.381   0.0173 *
## TRNSPD:Rack                     10223.2     1636.1   6.248 4.39e-10 ***
## WeatherTypeobscure:Rack          4715.8    61899.7   0.076   0.9393
## WeatherTypeprecipitation:Rack -201514.2    88084.9  -2.288   0.0222 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1070000 on 6926 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.06739,    Adjusted R-squared:  0.06618
## F-statistic: 55.61 on 9 and 6926 DF,  p-value: < 2.2e-16
```

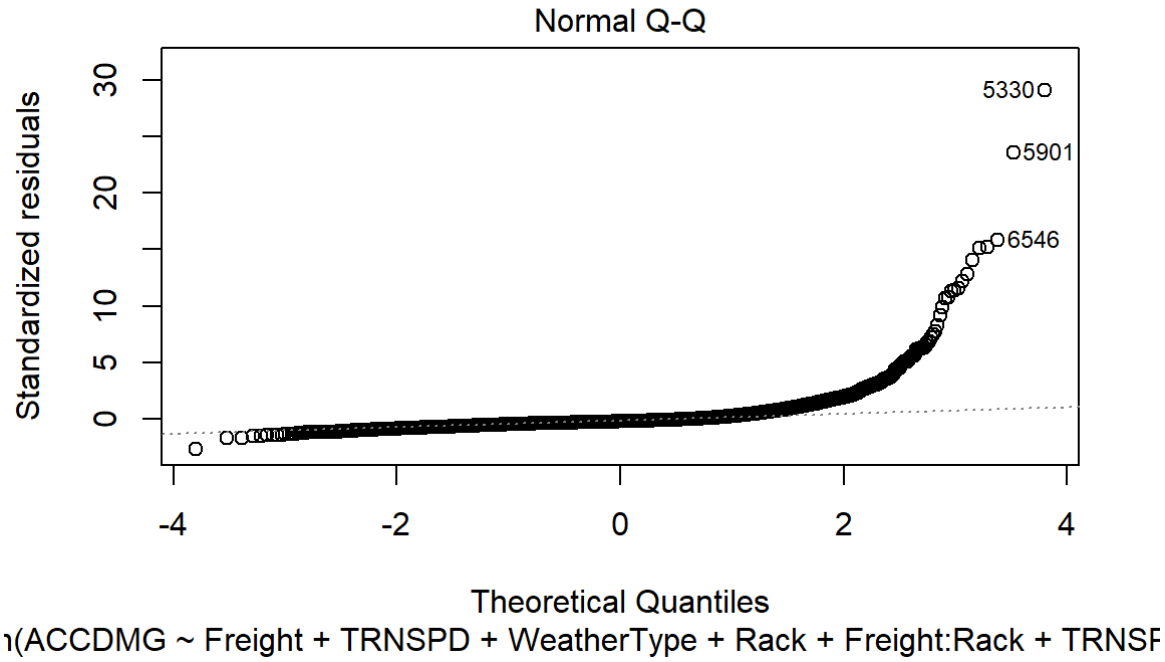
```
AIC(xdmg.lm1_2.step)
```

```
## [1] 212284
```

The cook's distance was much better and showed no leverage points. The Residual vs Fitted graph showed that the lack of fit assumption was met while the constant variance assumption was violated. The QQ plot showed that the model was even farther from a normality distribution than before. However, shown in the section above, this new model without influential points had some redeeming qualities such as a low AIC.



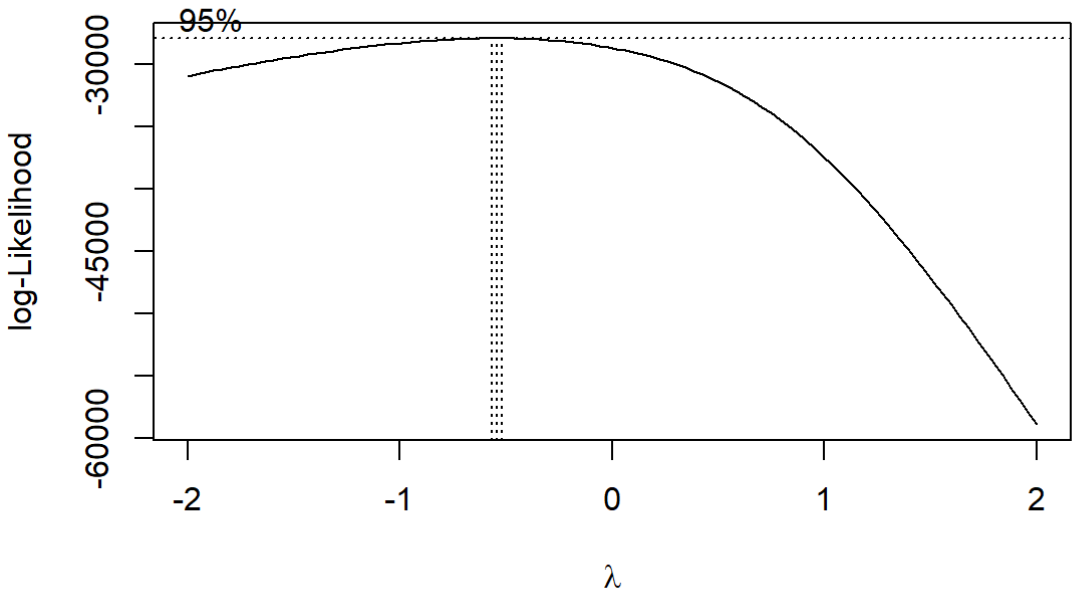
xdmg.lm1_2.step Residuals vs. Fitted Plot



xdmg.lm1_2.step QQ Plot

Adjusted Linear Models

To try and meet the normality assumption, we decided to run a boxcox transformation on the model xdmg.lm1 giving us xdmg.lm1.boxcox. After doing initial analysis on this model, we ran diagnostic plots. The residual vs fitted plot showed that the constant variance assumption was met, but lack of fit was no longer met. The QQ plot showed that, while the model is closer to being normally distributed, the assumption still isn't quite met. The graph showing cook's distance is better than the model without the boxcox transformation, but, overall, there are still some influential points that are worsening the model.



Boxcox Transformation for xdmg.lm1.step

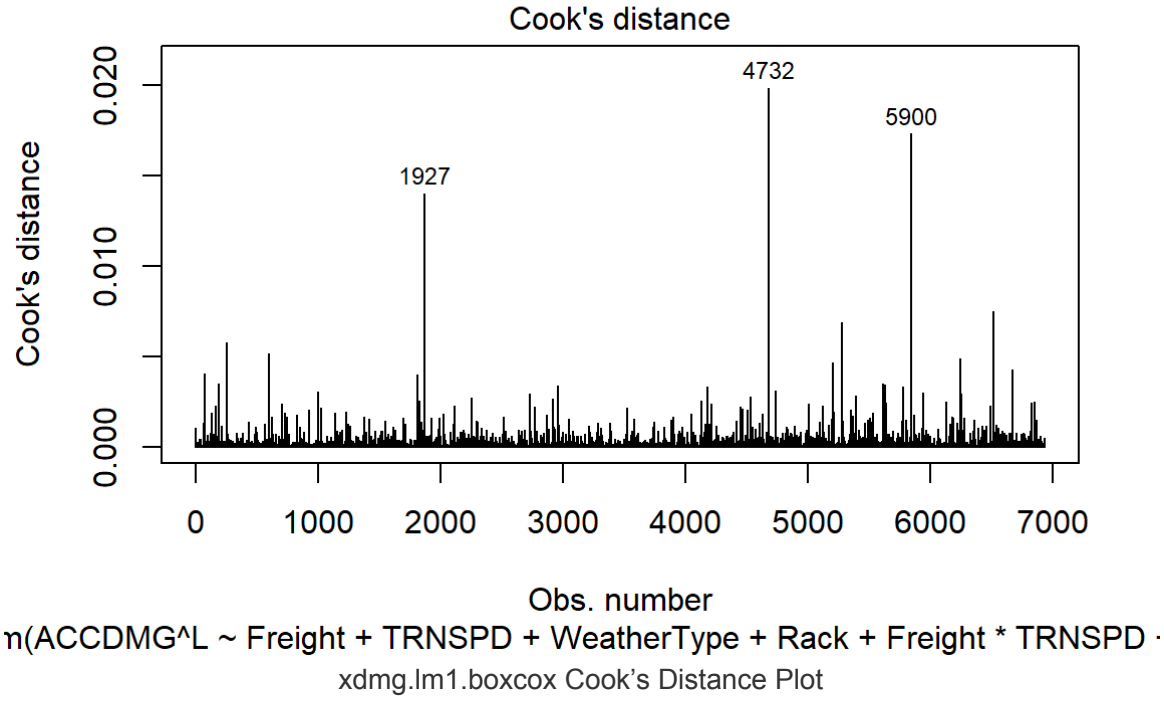
```
#lambda
boxcox(xdmg.lm1.step, plotit = F)$x[which.max(boxcox(xdmg.lm1.step,
                                                     plotit = F)$y)]
```

```
## [1] -0.5
```

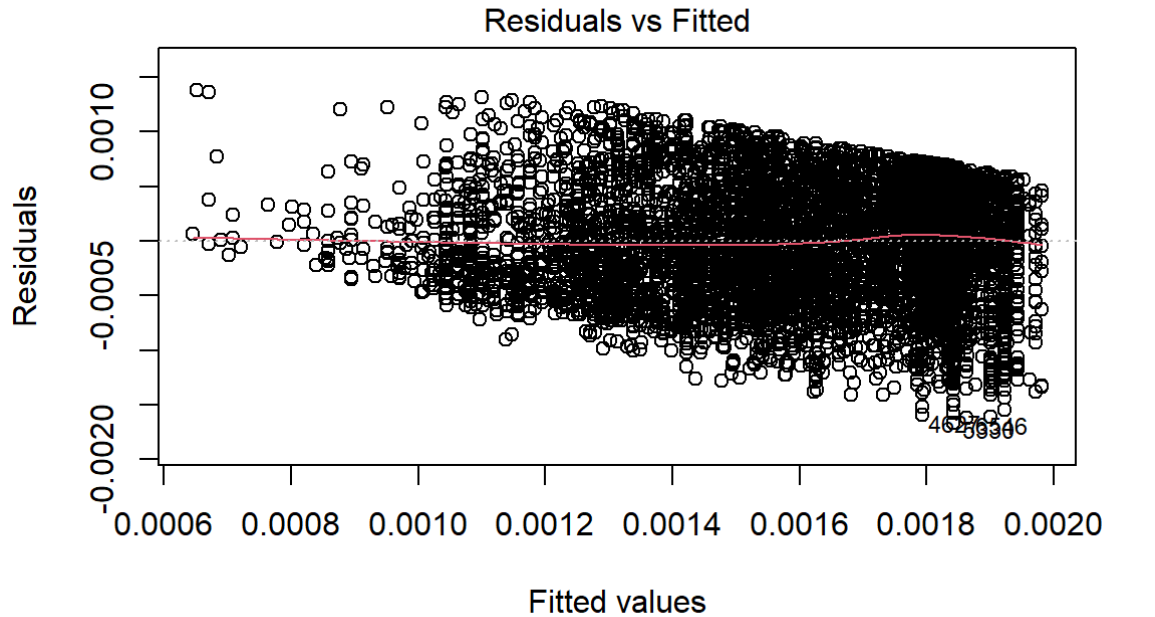
```
#max y value
max(boxcox(xdmg.lm1.step, plotit = F)$y)
```

```
## [1] -27905.89
```

```
#store best lambda
L<-boxcox(xdmg.lm1.step, plotit = F)$x[which.max(boxcox(xdmg.lm1.step,
                                                        plotit = F)$y)]
#model with best lambda
xdmg.lm1.boxcox<-lm(ACCDMG^L ~ Freight + TRNSPD + WeatherType + Rack +
                      Freight*TRNSPD + Freight*WeatherType +
                      TRNSPD*WeatherType +
                      TRNSPD*Rack + WeatherType*Rack,data=xdmg_nd)
```

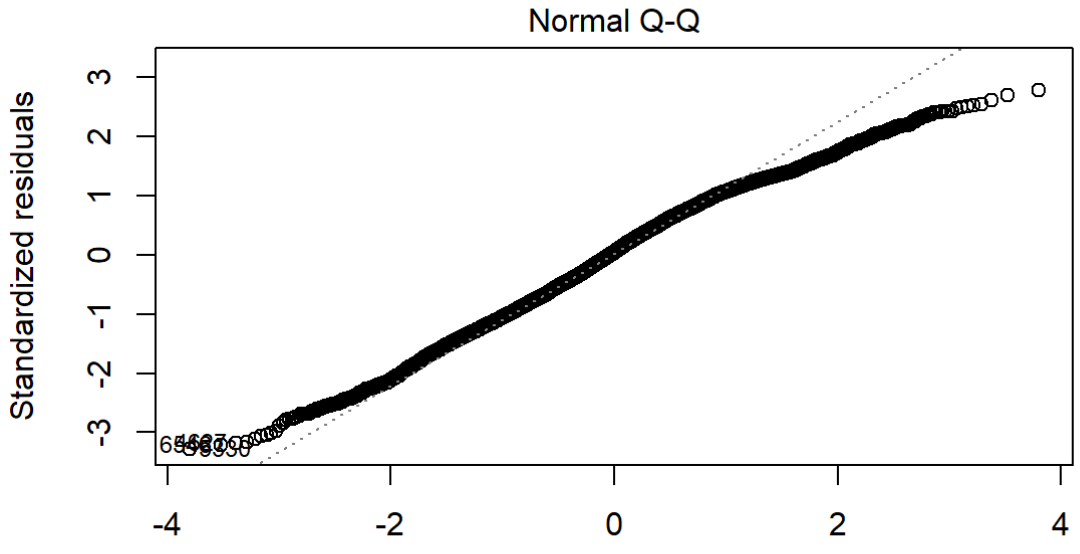


xdmg.lm1.boxcox Cook's Distance Plot



ɳ(ACCDMG^L ~ Freight + TRNSPD + WeatherType + Rack + Freight * TRNSPD ·

xdmg.lm1.boxcox Residuals vs. Fitted Plot



ɳ(ACCDMG^L ~ Freight + TRNSPD + WeatherType + Rack + Freight * TRNSPD ·

xdmg.lm1.boxcox QQ Plot

```
summary(xdmg.lm1.boxcox)
```

```
##
## Call:
## lm(formula = ACCDMG^L ~ Freight + TRNSPD + WeatherType + Rack +
##     Freight * TRNSPD + Freight * WeatherType + TRNSPD * WeatherType +
##     TRNSPD * Rack + WeatherType * Rack, data = xdmg_nd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.664e-03 -3.669e-04  2.319e-05  3.943e-04  1.384e-03
##
## Coefficients:
```

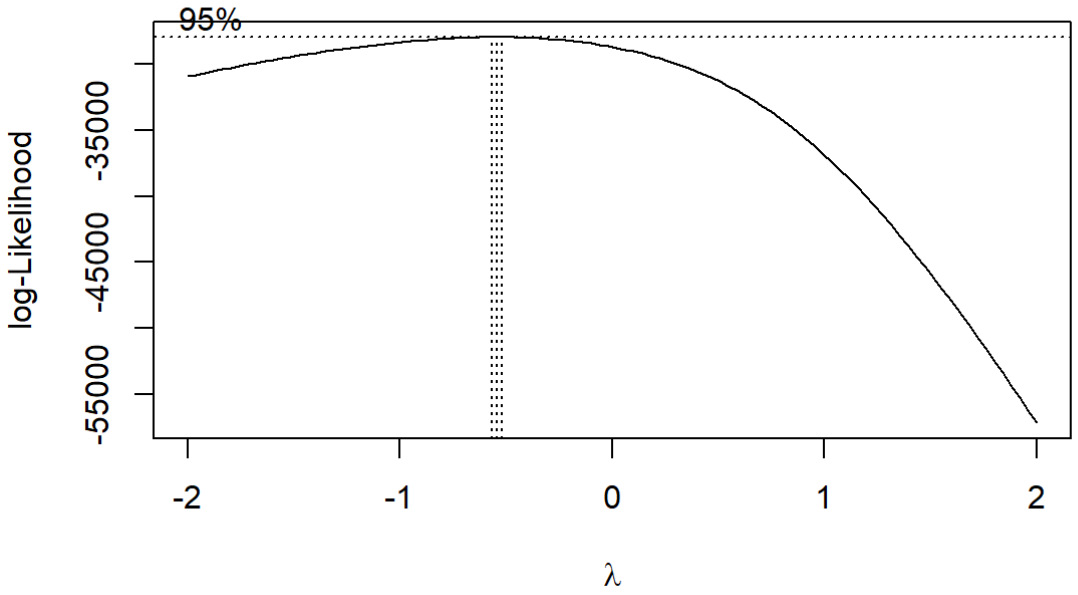
```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.842e-03  1.692e-05 108.860  < 2e-16 ***
## Freight                         5.905e-05  2.203e-05   2.681 0.007363 **
## TRNSPD                         -3.828e-06  6.502e-07  -5.887 4.11e-09 ***
## WeatherTypeobscure             -7.409e-06  3.094e-05  -0.240 0.810725
## WeatherTypeprecipitation       -4.884e-05  4.250e-05  -1.149 0.250466
## Rack                            8.058e-05  2.172e-05   3.710 0.000209 ***
## Freight:TRNSPD                 -8.130e-06  7.840e-07 -10.370  < 2e-16 ***
## Freight:WeatherTypeobscure     -3.102e-05  3.601e-05  -0.862 0.388944
## Freight:WeatherTypeprecipitation -6.675e-05  4.893e-05  -1.364 0.172547
## TRNSPD:WeatherTypeobscure      -2.349e-07  9.006e-07  -0.261 0.794220
## TRNSPD:WeatherTypeprecipitation  -2.156e-07  1.244e-06  -0.173 0.862399
## TRNSPD:Rack                    -6.762e-06  7.392e-07  -9.147  < 2e-16 ***
## WeatherTypeobscure:Rack         2.895e-05  2.969e-05   0.975 0.329694
## WeatherTypeprecipitation:Rack   9.792e-05  4.207e-05   2.328 0.019950 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0005066 on 6924 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.1871, Adjusted R-squared:  0.1856
## F-statistic: 122.6 on 13 and 6924 DF,  p-value: < 2.2e-16
```

```
AIC(xdmg.lm1.boxcox)
```

```
## [1] -85583.83
```

The second model we used to test our hypothesis was a boxcox transformation on xdmg.lm1.step giving us xdmg.lm1.boxcox. This model was significant based on the global utility test. After running t-tests, we found that Freight was significant, and the coefficient is positive showing a positive relationship with accident damage. The interaction term between freight and trainspeed was significant but had a negative relationship with accident damage. Rack and the interaction term between rack and precipitation were both significant and had positive relationships with accident damage. The interaction between rack and train speed was also significant and had a negative relationship with accident damage. To further test the adequacy of the boxcox transformation, we found the adjusted R2 and AIC, 18.56% and -85583.83 respectively. The adjusted R2 is much higher than the model without a boxcox transformation which shows that this model is more accurate at predicting the actual accident damage of an event. The AIC is larger, however, which causes more concern for this model. After this analysis, we decided to test a smaller initial model.

We decided to try and solve the normality problem on xdmg.lm1_2.step by running a boxcox transformation on xdmg.lm_2.step giving us xdmg.lm1_2.boxcox. The QQ plot showed that the normality assumption was now met. The Residual vs Fitted plot showed that constant variance was also met. The only assumption not fully met was lack of fit, but the residual vs fitted graph showed that it wasn't a large violation.



Boxcox Transformation of xdmg.lm1_2.step

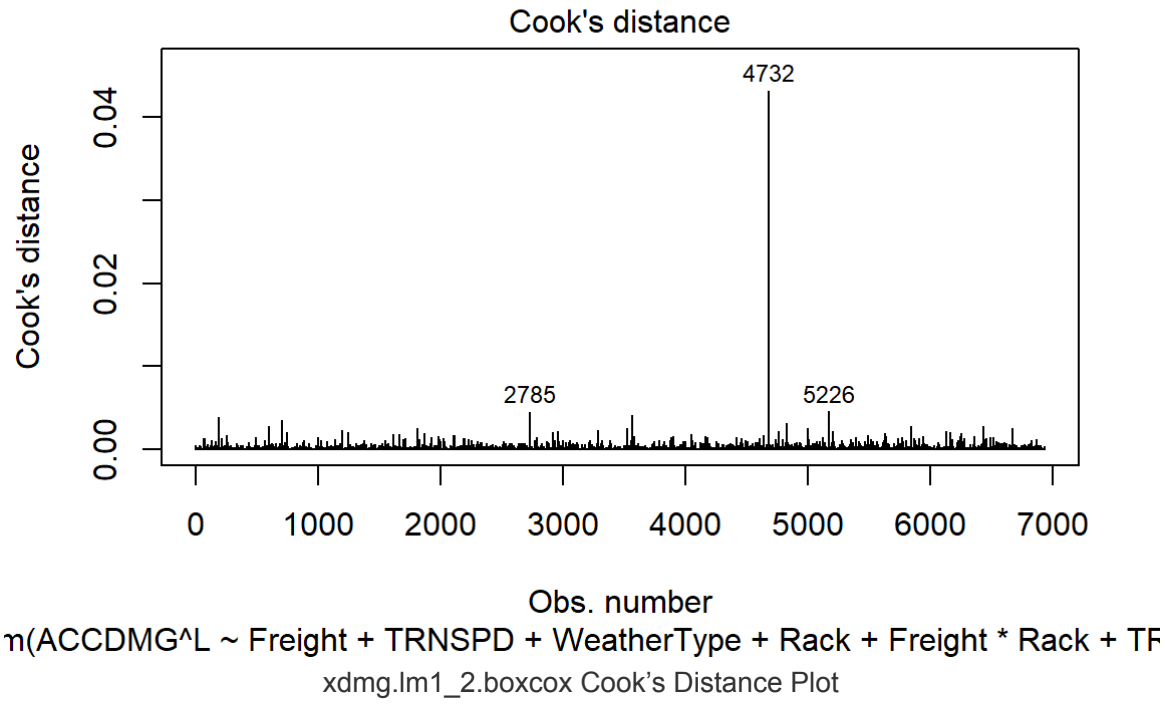
```
#lambda
boxcox(xdmg.lm1_2.step, plotit = F)$x[which.max(boxcox(xdmg.lm1_2.step,
                                                       plotit = F)$y)]
```

```
## [1] -0.5
```

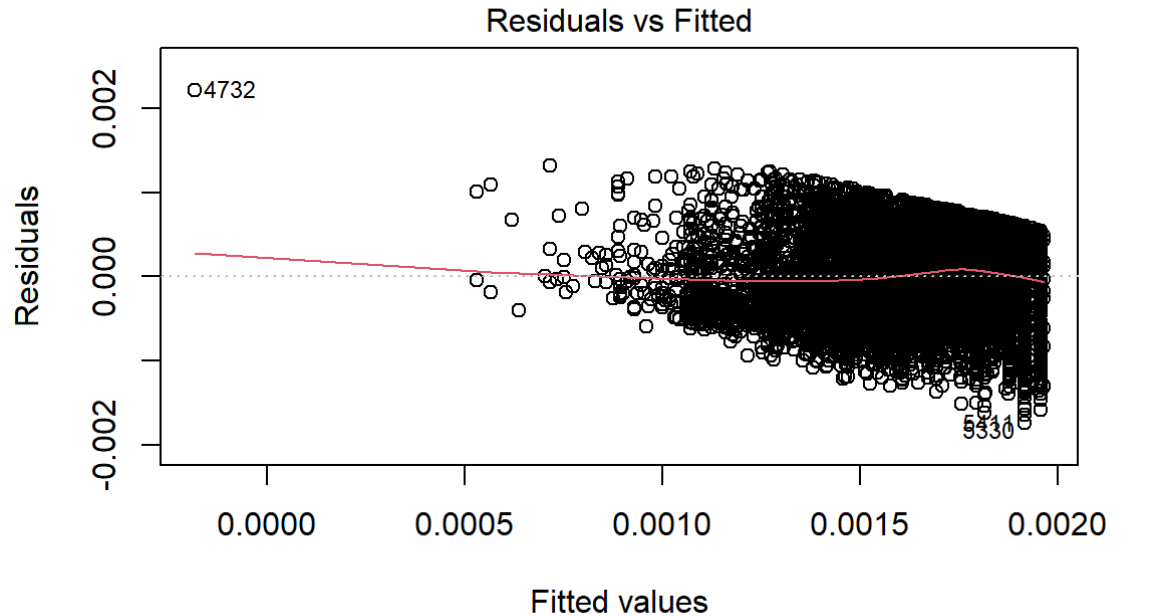
```
#max y value
max(boxcox(xdmg.lm1_2.step, plotit = F)$y)
```

```
## [1] -27938.47
```

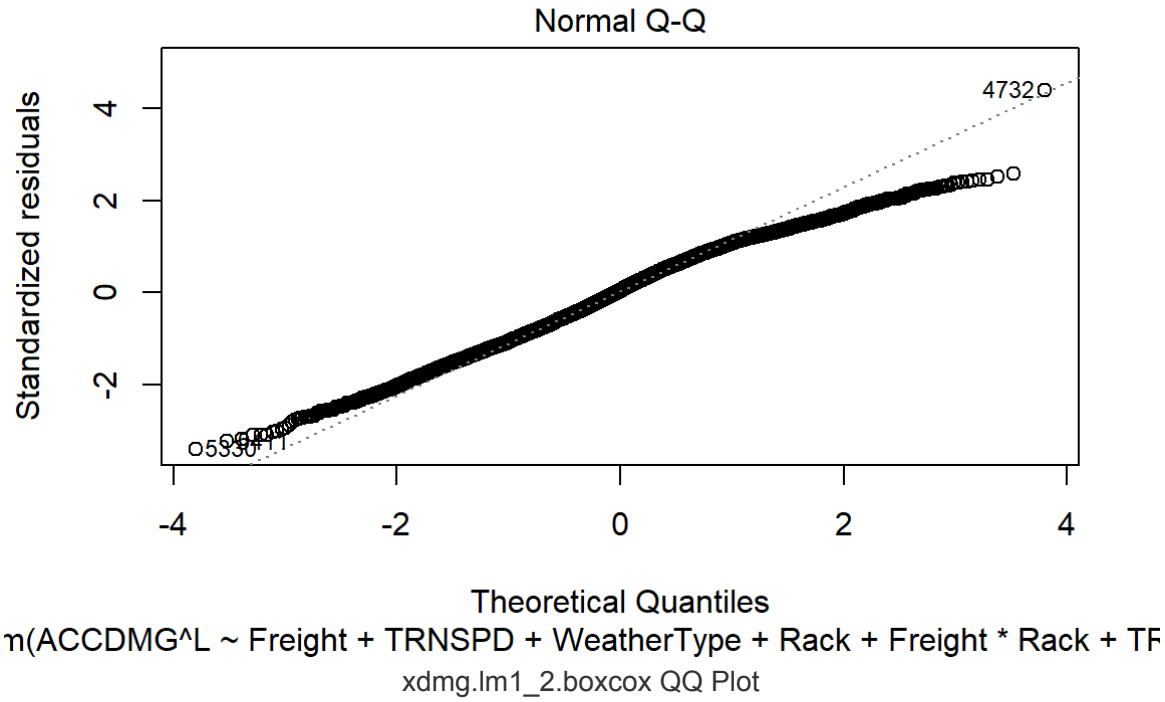
```
#store best lambda
L<-boxcox(xdmg.lm1_2.step, plotit = F)$x[which.max(boxcox(xdmg.lm1_2.step,
                                                         plotit = F)$y)]
#model with best lambda
xdmg.lm1_2.boxcox<-lm(ACCDMG^L ~ Freight + TRNSPD + WeatherType +
                        Rack + Freight*Rack + TRNSPD*Rack +
                        WeatherType*Rack, data = xdmg_nd[-c(5900,5337), ])
```



xdmg.lm1_2.boxcox Cook's Distance Plot



xdmg.lm1_2.boxcox Residuals vs. Fitted Plot



xdmg.lm1_2.boxcox QQ Plot

```
summary(xdmg.lm1_2.boxcox)
```

```
##
## Call:
## lm(formula = ACCDMG^L ~ Freight + TRNSPD + WeatherType + Rack +
##     Freight * Rack + TRNSPD * Rack + WeatherType * Rack, data = xdmg_nd[-c(5900,
##     5337), ])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.739e-03 -3.771e-04  2.046e-05  4.031e-04  2.219e-03
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.917e-03  1.559e-05 122.956  < 2e-16 ***
## Freight                      -1.222e-04  1.789e-05  -6.832 9.10e-12 ***
## TRNSPD                       -8.117e-06  4.488e-07 -18.088  < 2e-16 ***
## WeatherTypeobscure           -3.694e-05  1.918e-05  -1.926 0.054104 .
## WeatherTypeprecipitation     -1.011e-04  2.598e-05  -3.890 0.000101 ***
## Rack                          4.164e-05  2.830e-05   1.471 0.141234
## Freight:Rack                  1.287e-04  3.190e-05   4.033 5.56e-05 ***
## TRNSPD:Rack                  -9.737e-06  7.794e-07 -12.492  < 2e-16 ***
## WeatherTypeobscure:Rack       2.579e-05  2.949e-05   0.874 0.381919
## WeatherTypeprecipitation:Rack  1.017e-04  4.196e-05   2.424 0.015360 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0005098 on 6926 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.175,  Adjusted R-squared:  0.1739
## F-statistic: 163.2 on 9 and 6926 DF,  p-value: < 2.2e-16
```

```
AIC(xdmg.lm1_2.boxcox)
```

```
## [1] -85475.54
```

The third model (xdmg.lm1_2.boxcox) we used to test our hypothesis was a boxcox transformation of the regression model of thestep model generated from the full second order model with influential points, 5900 and 5337, from the original model. This model was significant based on the global utility test. After running t-tests, we found that Freight was significant, and the relationship between accident damage was negative. Rack was not significant using this model. The interaction between rack and trainspeed is significant and has a negative relationship with accident damage. The interaction between rack and precipitation was significant and had a positive relationship with accident damage. Lastly, the interaction between freight and rack was significant and had a positive relationship with accident damage. To further test the adequacy of this model we found the adjusted R2 and AIC, 17.38% and -85475.5 respectively. The adjusted R2 is relatively adequate, and the AIC value is rather large.

However, based on diagnostic plots that are described, this model, boxcox transformation on a second order model without influential points is the best model for predicting accident damage. It has a slightly lower R2 value than the 18.56% value from the model that conducted a boxcox transformation on the initial model's step function, however, this model of the boxcox tranformation on the step function from the full second order

made after leverage points were removed has a signicantly lower AIC value, which is what we are looking for. Thus, **the xdmg.lm1_2.boxcox is the ideal model to be selected for this hypothesis**.

Casualties

Exploratory Analysis

Cleaning up the Data

Similar to accident damage, we must subset the larger train dataset for the purpose of analyzing different variable impacts on Casualties. There are over 56,000 entries in the dataset that currently have a casualty of zero. These entries are not relevant for our purpose of modeling variable impacting casualty numbers, because there are no casualties at all. Thus, we created a new casualties dataset that contains only entries with a casualty value of 1 or more. We also removed any duplicate entries.

```
casualties = totacts %>% filter(Casualty>=1)
casualties_nd <- casualties[!(duplicated(casualties[, c("INCDTNO",
                     "YEAR", "MONTH", "DAY", "TIMEHR", "TIMEMIN")])),]
```

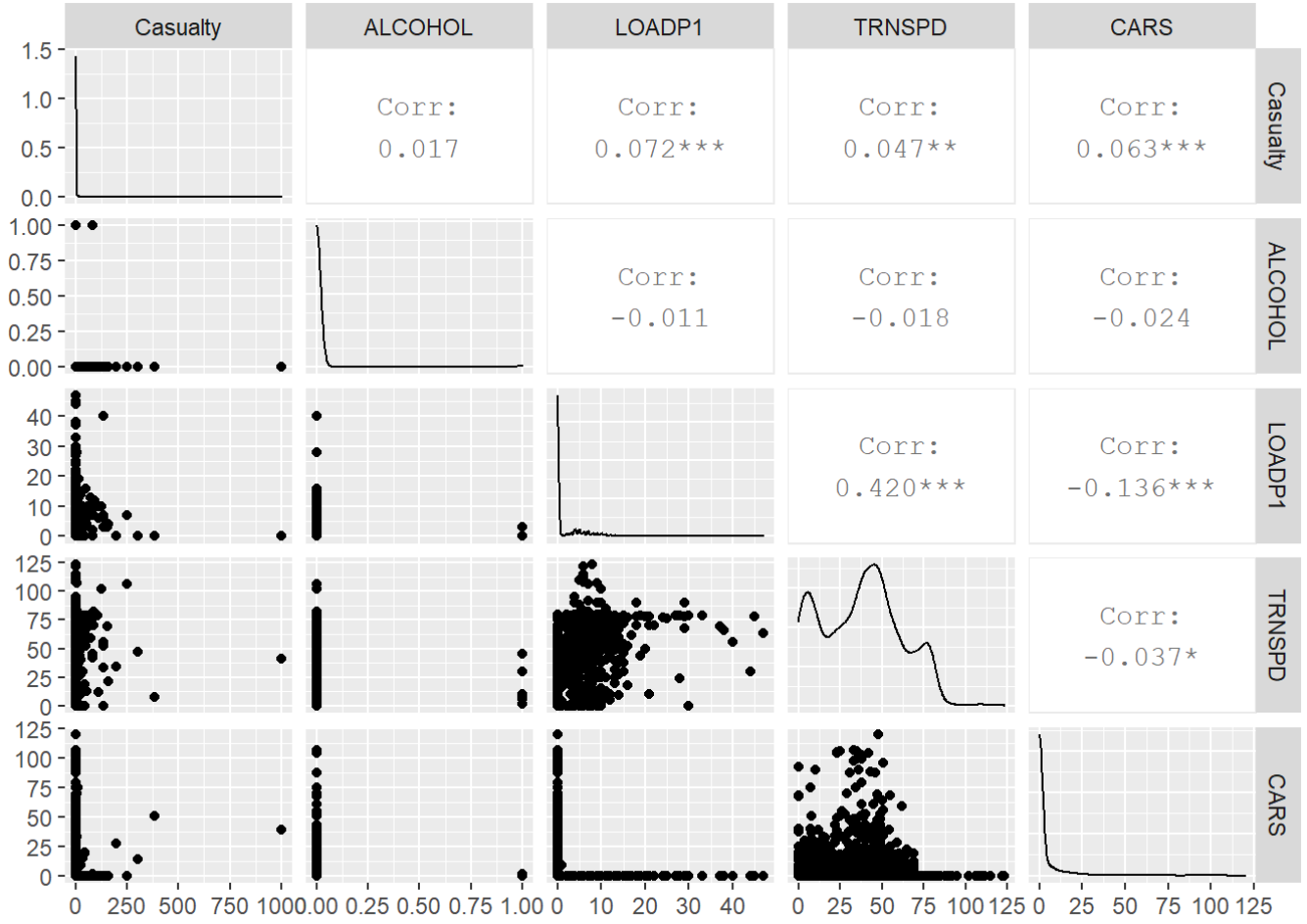
Just as was done with the xdmg_nd dataset, a Cause variable was created for the casualties dataset with the same letters (M, T, S, H, and E) and the same causes corresponded to each.

```
casualties_nd$Cause <- rep(NA, nrow(casualties_nd))
casualties_nd$Cause[which(substr(casualties_nd$CAUSE, 1, 1) == "M")] <- "M"
casualties_nd$Cause[which(substr(casualties_nd$CAUSE, 1, 1) == "T")] <- "T"
casualties_nd$Cause[which(substr(casualties_nd$CAUSE, 1, 1) == "S")] <- "S"
casualties_nd$Cause[which(substr(casualties_nd$CAUSE, 1, 1) == "H")] <- "H"
casualties_nd$Cause[which(substr(casualties_nd$CAUSE, 1, 1) == "E")] <- "E"

casualties_nd$Cause = as.factor(casualties_nd$Cause)
```

Quantitative Variables

Similar to the ACCDMG response variable process, in order to determine which predictor variables to be using in the Casualty linear model and figure out what our hypotheses are, first quantitative and then qualitative variables were tested.



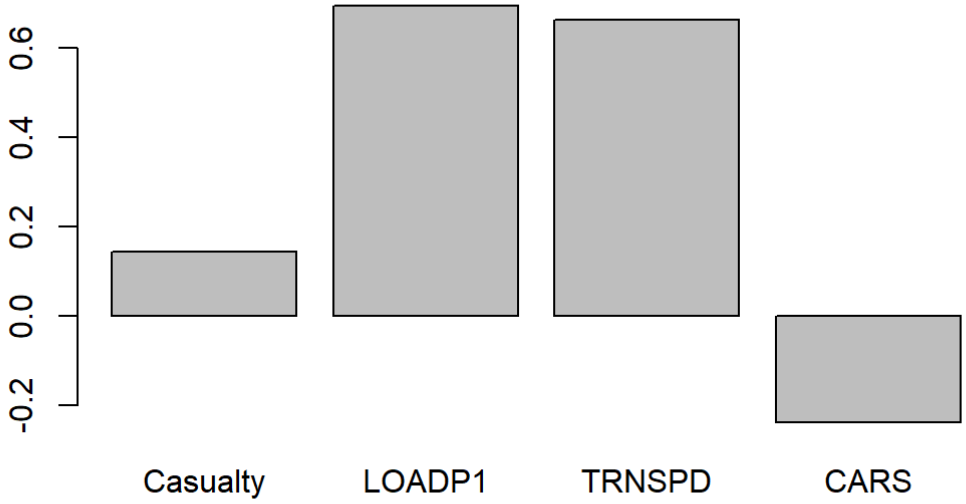
Scatterplot Matrix of Casualty with Quantitative Variables

After running a scatterplot matrix with Casualty along with ALCOHOL (number of people on train under the influence), LOADP1 (the number of loaded passenger cars), TRNSPD (train speed), and CARS (the number of cars with hazmat), we observed that LOADP1 had the highest correlation with Casualty, although it was quite low at 0.072.

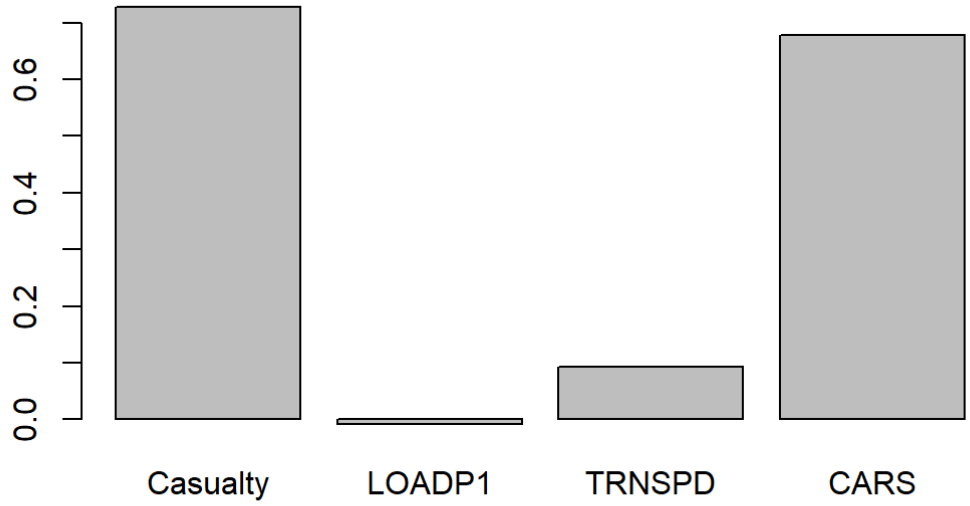
PCA loadings plots for the first 2 PCs were created too, using a correlation matrix. The first loadings plot shows the majority of the 1st level of greatest variability in the dataset being due to LOADP1 and TRNSPD. They both vary together with Casualty. The second level of greatest

variability is due largest in part to Casualty and Cars.

```
pca.casualty.corr = princomp(casualties_nd[,c("Casualty", "LOADP1",
                                              "TRNSPD", "CARS")], cor=T)
```



First PC Loadings for casualties_nd

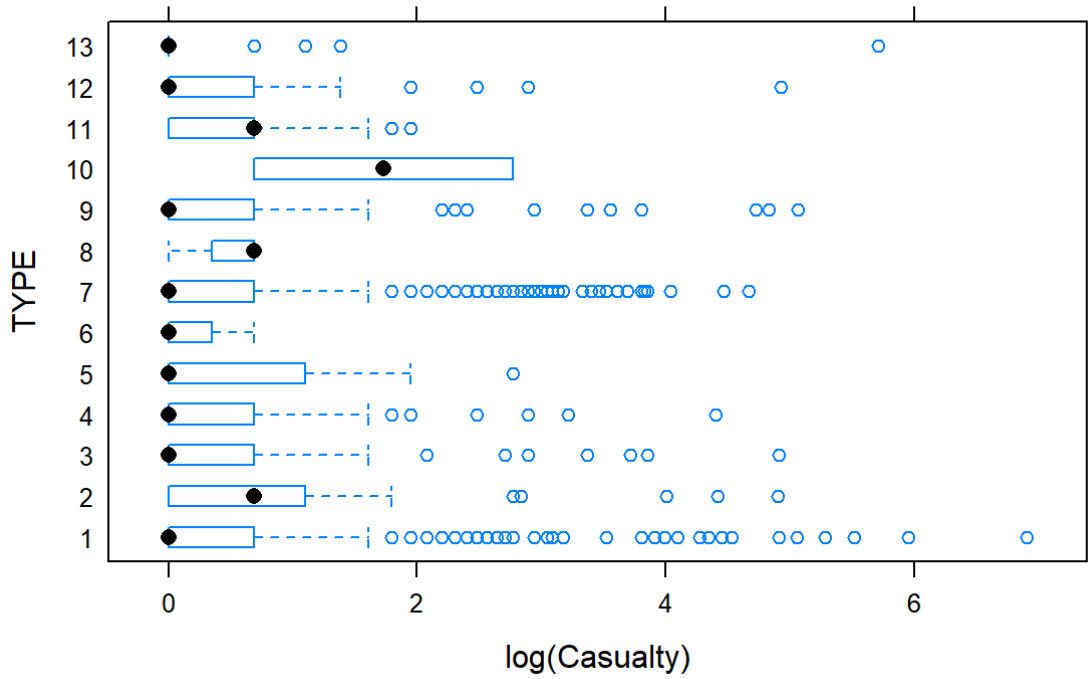


Second PC Loadings for casualties_nd

Qualitative Variables

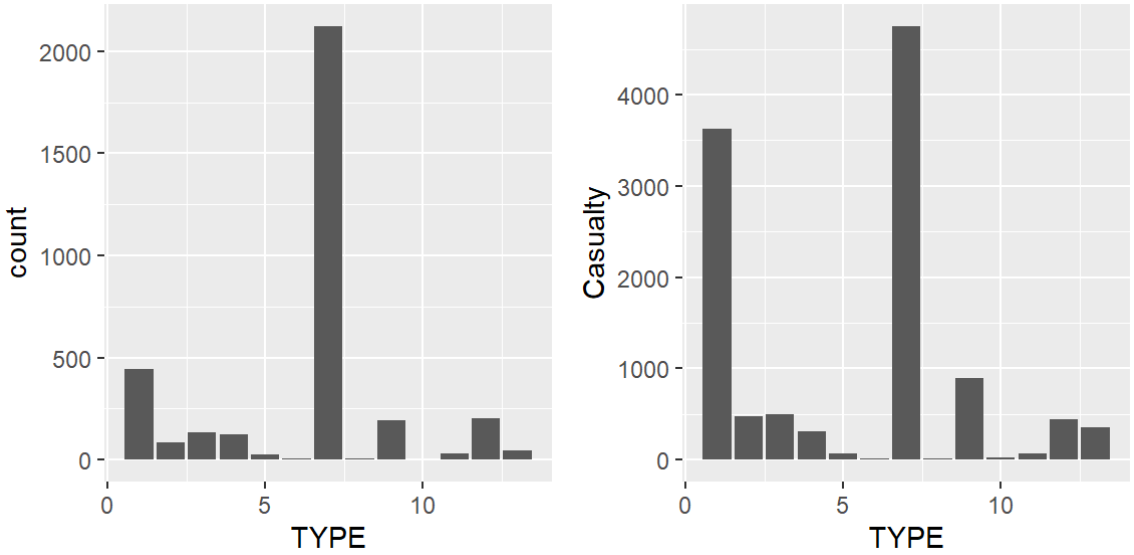
In terms of qualitative variables, TYPE, CAUSE, and PartofDay, a new variable we created, were considered.

A factor boxplot, similar to with ACCDMG, was generated.



Factor Boxplot of TYPE and log(Casualty)

Types 1 and 7, while their medians were fairly consistent with the rest of the accident types, had very large numbers of outlier points that accounted for large Casualty values. Theses are the types we decided to focus on in order to type to reduce those outlier accidents.

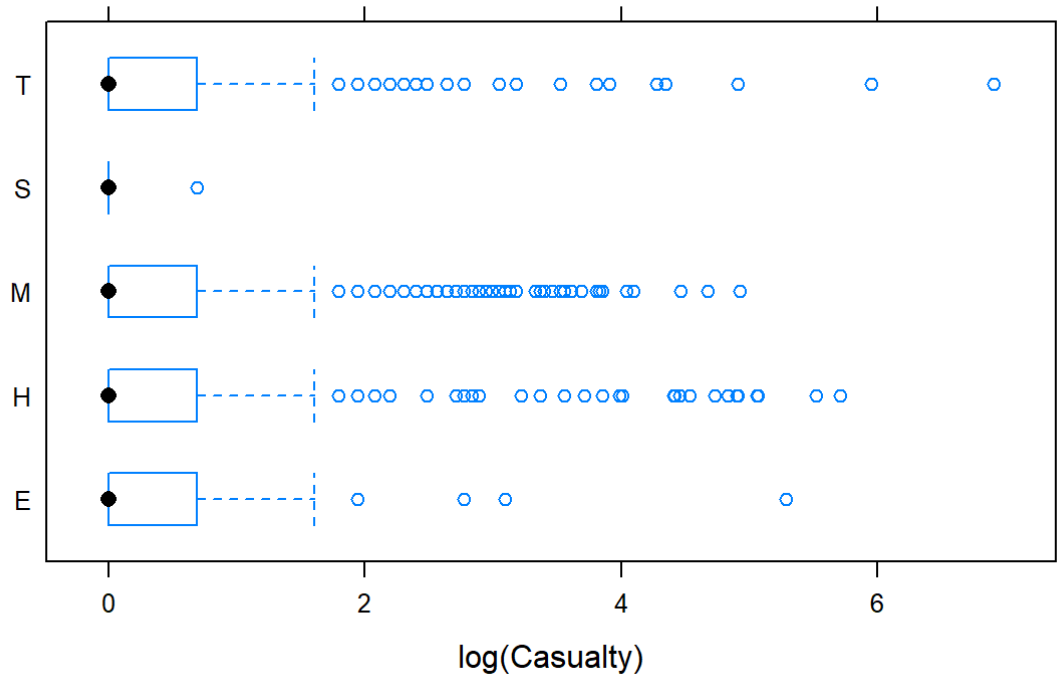


Histogram and Barplot of TYPE

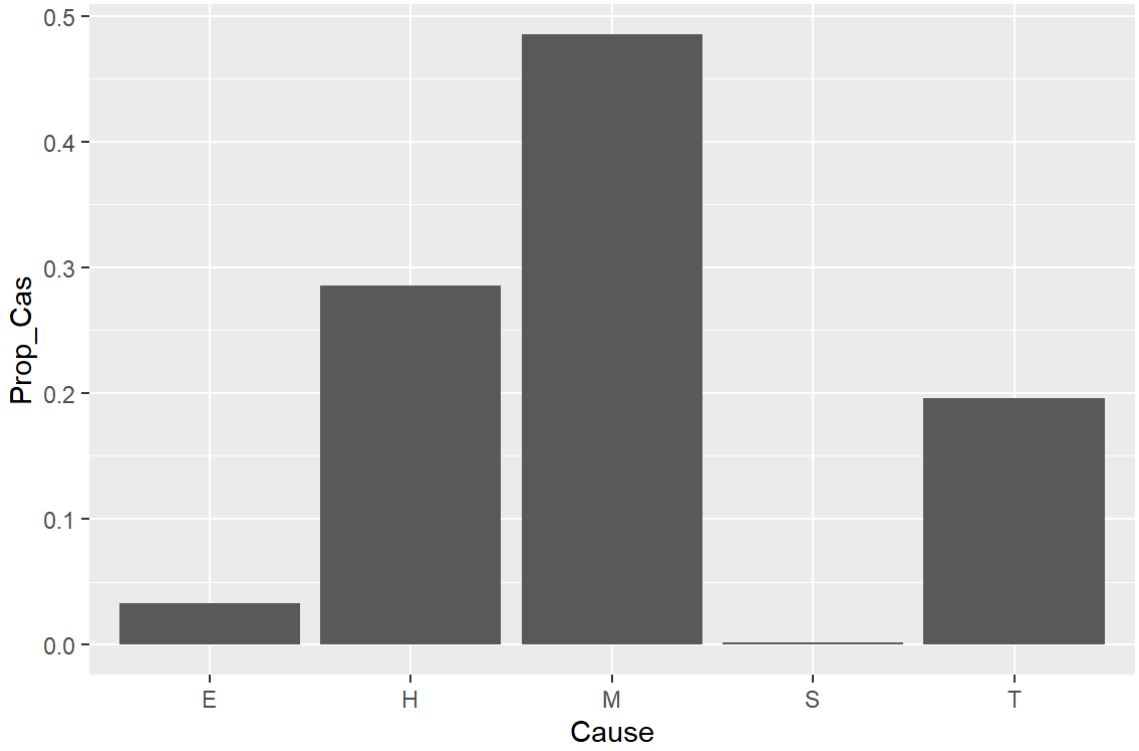
```
casualties_nd$HRCrossing = ifelse(casualties_nd$TYPE==7,1,0)
```

Type 7 was Highway-Railroad crossing. This type had both high counts as well as high casualty values. We interpreted this as although each individual accident may cause small casualty numbers, the sheer number of instances of this accident type results in vary large aggregate casualty numbers, and should be a accident type that should be addressed. A dummy variable for HRCrossing, which is a 1 for all Highway-Crossing accidents and 0 otherwise, was made.

Observing the factor boxplot and histogram, miscellaneous causes of train accidents is the largest proportion of casualties as well as has the highest number of outlier values. However, this cause is very ambigious. It is not very actionable, as we ourselves do not know the exact cause of the accidents that have occurred. All we know is what they are not. Thus, the next largest category, Human Factors errors, was selected to be focused on. A dummy variable, Human, which is a 1 for all human factors errors and a 0 otherwise, was made.



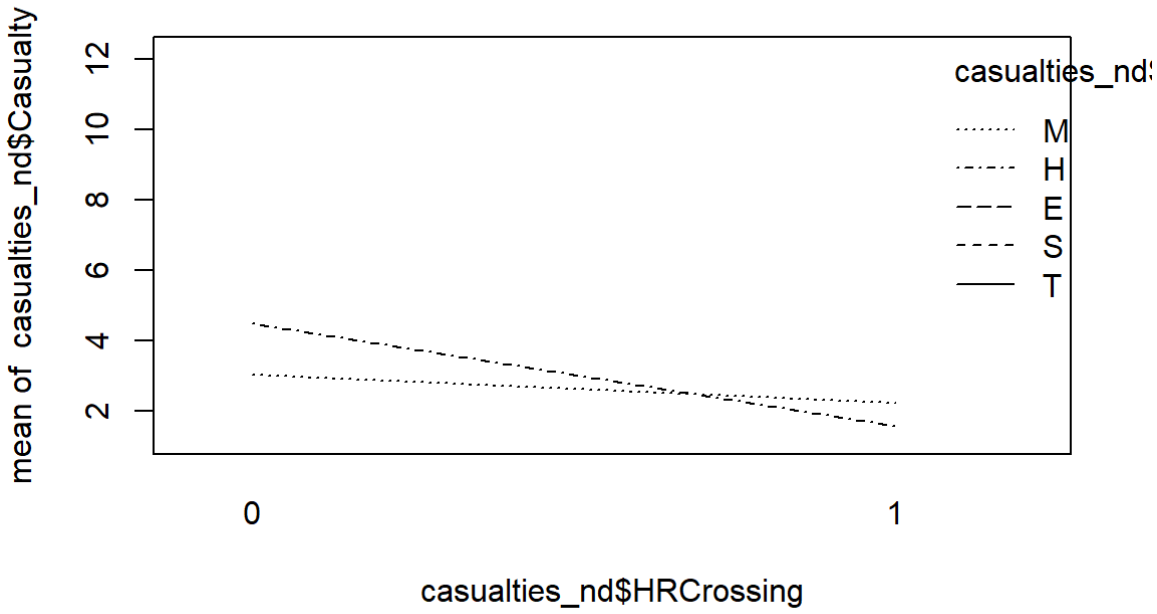
Factor Boxplot of Cause and log(Casualty)



Proportion of Casualties by Cause

```
casualties_nd$Human = ifelse(casualties_nd$Cause == "H", 1,0)
```

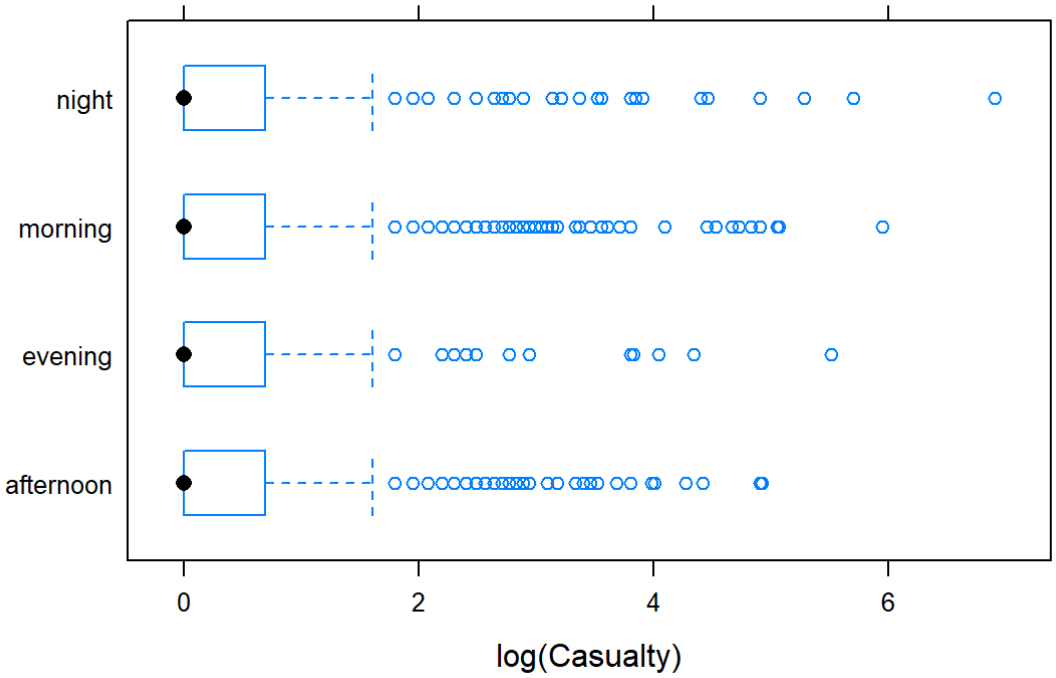
Human factors and miscellaneous errors were the only causes that occurred for Highway-Rail Crossing accidents. However, due to the difference in slope and the intersection between these two lines in the interaction plot below, it was concluded that at least this one interaction is needed in the linear model being used.



Interaction Plot between HRCrossing and Cause with Casualty

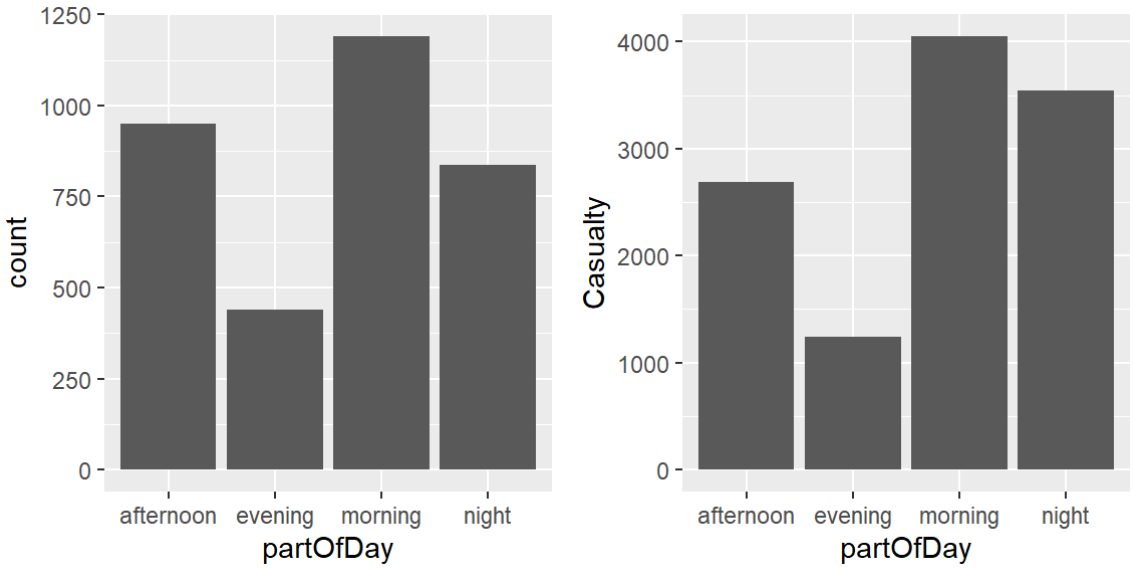
Part of Day is a variable that we created to equate to a different time of day: morning, afternoon, evening, and night. While this is not something that can be explicitly controlled for, we thought this aspect would be important to investigate to see if it had any relevance to casualty numbers. 7am to 11am was morning, 12pm to 5pm was afternoon, 6pm to 9pm was evening, and 10pm to 6am was night.

```
    #PART OF DAY - Night or not (10pm - 6am)
casualties_nd = casualties_nd %>% mutate(partOfDay = case_when(
  TIMEHR >= 7 & AMPM =="AM" ~ "morning",
  TIMEHR <= 9 & TIMEHR >5 & AMPM =="PM" ~ "evening",
  TIMEHR <=5 & AMPM =="PM" ~ "afternoon",
  TIMEHR > 9 & AMPM == "PM" ~ "night",
  TRUE ~ "night"
))
casualties_nd$partOfDay = as.factor(casualties_nd$partOfDay)
```



Factor Boxplot of partOfDay and log(Casualty)

The factor barplot shows a pretty even distribution with casualty values across all parts of the day.

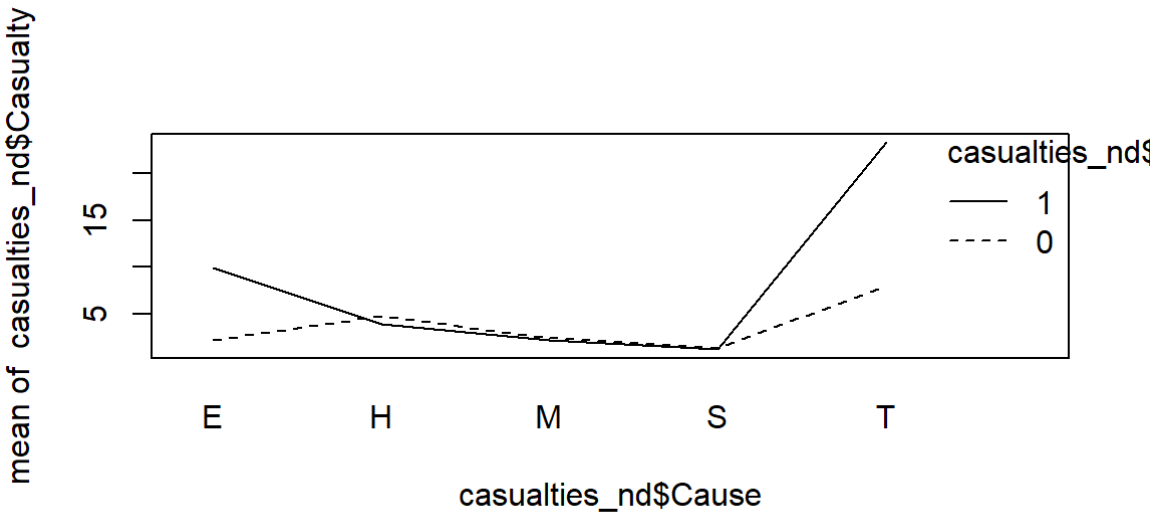


Histogram and Barplot of parOfDay

```
casualties_nd$Night = ifelse(casualties_nd$partOfDay == "night", 1, 0)
```

The histograms above show that while the count for night accidents with casualties was the second lowest, it was the second highest in terms of aggregate casualties, meaning that intuitively we can assume that each of those night accidents had higher casualties per accident. As a result, we decided to look further into night time accidents, making a dummy variable Night from the PartOfDay variable, 1 when it was Night, 0 otherwise.

The interaction plot below between Cause and Night suggests the need for an interaction term given the slope differences and the number of casualties for each cause being higher or lower at night changing.



Interaction Plot of Cause and Night with Casualty

Generating Hypotheses

Our first hypothesis identifies the highway rail crossing as a situation where more casualties can occur. We reached this conclusion based on the fact that, when looking at the factor boxplots the median of type 7 is around the same as those of the other types; however, this type of accident had many outlier values, which is indicative of it being a potential issue when accounting for the reasoning behind high casualties. Additionally, when looking at the histograms and barplots comparing the count and the casualties of various types of passenger trains the same thing occurred in the idea that there were both high counts and high casualty numbers. Reducing the number of this type of accident will then, in effect, reduce aggregate casualty numbers. Choosing this situation is actionable, as extra monitoring and care can be placed at highway rail crossing. In addition, there is the ability to change elements of a passenger train and highway rail crossing in order to improve safety. The null hypothesis is **Highway rail crossing does not have any effect on the number of Casualties relative to other types of accidents**. The alternative is **Highway rail crossing increases the number of Casualties relative to the other types of accidents**.

Our second hypothesis considers the involvement of human factors and its effect on the number of casualties in an accident. This hypothesis was formulated in recognizing that it accounted for the second highest proportion of casualties in the dataset. It also followed similar trends with the factor boxplot as mentioned with prior hypothesis. The cause “M” had the highest proportion of casualties and the highest number/value of outliers, however this represents ‘miscellaneous’ error, and would not be actionable given that we are not sure what exactly falls into this category apart from the fact that the accidents did not fit into the other 4 options for cause. This is actionable because knowing this provides the advantage of

knowing whether or not human error is a big factor in casualty numbers. Once this is confirmed or denied, further analysis can be made to look into what exact the failure was, such as with breaks. Targeted training protocols or safety procedures can be formulated, and such changes can be made as necessary. The null hypothesis is **Human factors does not have any effect on the severity of Casualties relative to other kinds of Causes**. The alternative hypothesis is **Human Factors does increase the severity of Casualties relative to other kinds of Causes**.

In the case of failing to reject a null hypothesis for either, we identify a situation that does not need extreme concern and research can be prioritized in other situations.

Initial Linear Model

In the Causality analysis, our first model was a full second order model and stepwise model with the variables highwayRailCrossing, loadp1, human factors, and night. By looking at the results of the F-test, we concluded that the stepwise model was a better fit for the dataset, but noticed that there were a few points that were extreme outliers. We executed a repetitive process of removing the outlier with the highest leverage and highest residuals and visualizing how it impacted the full second order model and its stepwise function to optimize our overall model. Once it was determined the best model for our hypothesis was the third stepwise model, we created a box cox plot and compared it to the stepwise model. Ultimately, the third stepwise model was the best fit because the box cox plot failed to meet the normality assumption and did not improve the model.

For every model we used to test our hypotheses against Casualty, we used t-tests to determine if the specific terms (either highway crossing or human factors or interaction terms with them) were significant. If the resulting p-values led us to believe the term was significant, we used the coefficient's sign to determine if we would reject or fail to reject the null hypothesis. We also used the global utility test on each model to determine if the model itself was significant.

```
casualty.lm1 = lm(Casualty~(HRCrossing + LOADP1 + Human + Night)^2,
                  casualties_nd)
casualty.lm1.step = step(casualty.lm1, trace=F)
#lm1.step better
anova(casualty.lm1, casualty.lm1.step)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	3406	1553426	*NA*	*NA*	*NA*	*NA*
2	3407	1553442	-1	-16.14895	0.03540776	0.850755

2 rows

```
summary(casualty.lm1.step)
```

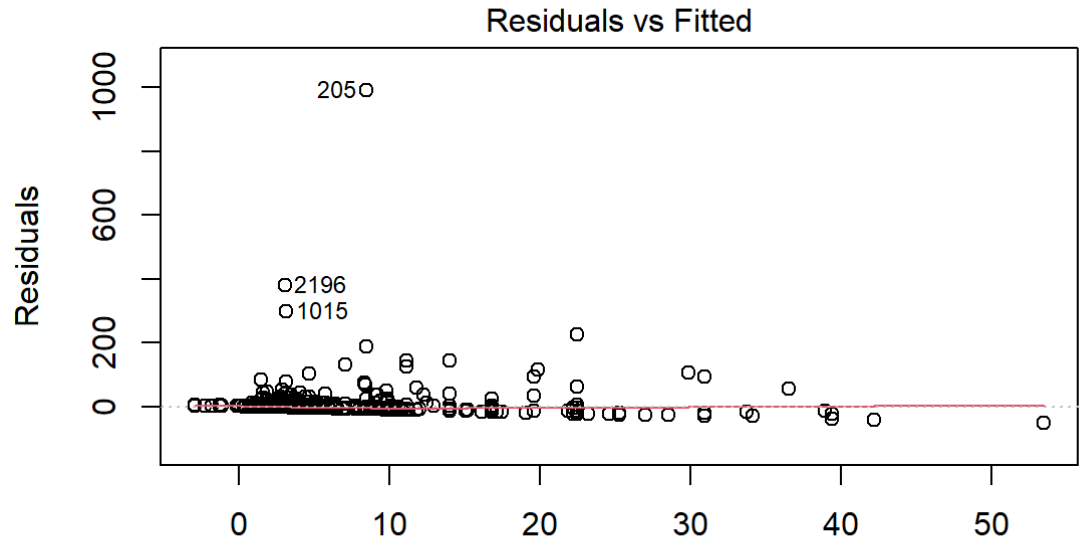
```
##
## Call:
## lm(formula = Casualty ~ HRCrossing + LOADP1 + Human + Night +
##     HRCrossing:LOADP1 + HRCrossing:Night + LOADP1:Human + LOADP1:Night +
##     Human:Night, data = casualties_nd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.47  -1.86  -0.73  -0.62 992.51
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.0839     1.1984   2.573 0.010111 *
## HRCrossing         -1.4595     1.3109  -1.113 0.265634
## LOADP1              0.6707     0.1966   3.412 0.000652 ***
## Human              -0.3519     1.5583  -0.226 0.821356
## Night               5.4109     1.9292   2.805 0.005063 **
## HRCrossing:LOADP1  -0.3617     0.2153  -1.680 0.093000 .
## HRCrossing:Night   -4.6284     2.2335  -2.072 0.038319 *
## LOADP1:Human        2.1481     0.4124   5.208 2.02e-07 ***
## LOADP1:Night       -0.4352     0.2019  -2.156 0.031182 *
## Human:Night        -4.9736     2.5166  -1.976 0.048204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.35 on 3407 degrees of freedom
## Multiple R-squared:  0.0266, Adjusted R-squared:  0.02403
## F-statistic: 10.35 on 9 and 3407 DF,  p-value: 6.588e-16
```

```
AIC(casualty.lm1.step)
```

```
## [1] 30629.24
```

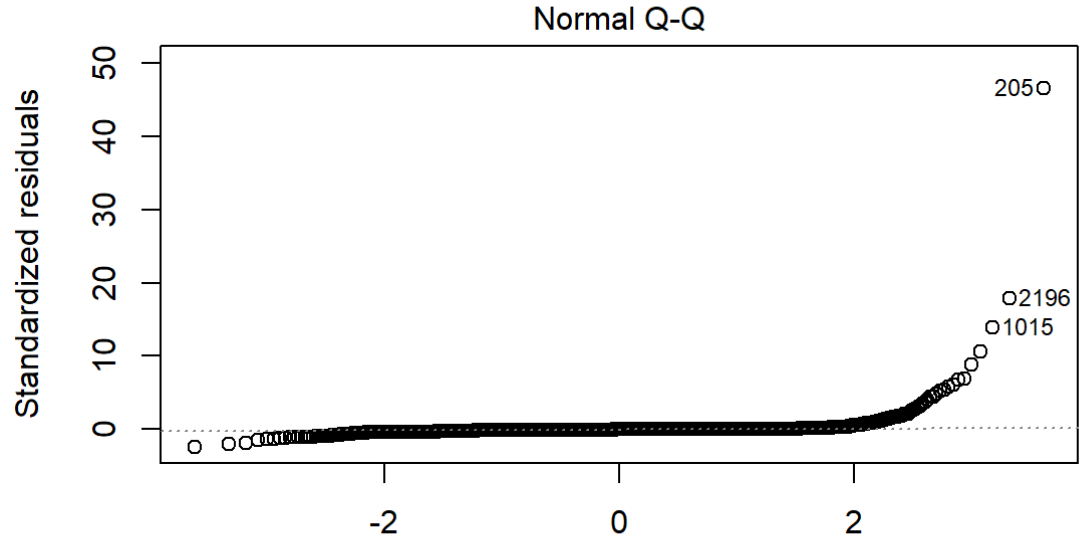
The first model we created for our Casualty hypotheses was a full second order model using the exploratory variables: highway rail crossing, loadP1, human factors, and Night. We ran stepwise regression on this model and compared the original model and the step model using an anova test. This test showed that the stepwise regression model (casualty.lm1.step) was better. We assessed this model in multiple ways. First, we ran a global utility test; it showed that the model was significant. Next, we ran t-tests on each term in the model. The highway rail crossing and the human factors term are not significant in the model. The interaction term between highway rail crossing and night was significant and has a negative relationship with the number of casualties. The interaction term between human factors and loadP1 was significant and has a positive relationship with the number of casualties. The interaction term between human factors and night was significant and has a negative relationship with the number of casualties. To further test the adequacy of casualty.lm1.step, we found the adjusted R2, 2.403%, and AIC, 30629.24. The adjusted R2 shows that our model isn't very accurate at predicting the actual accident damage of an event. The AIC is also relatively small which is a good sign. After this analysis, we used diagnostic plots to more adequately check our model. This led us to make changes to our model.

Diagnosing Model Problems



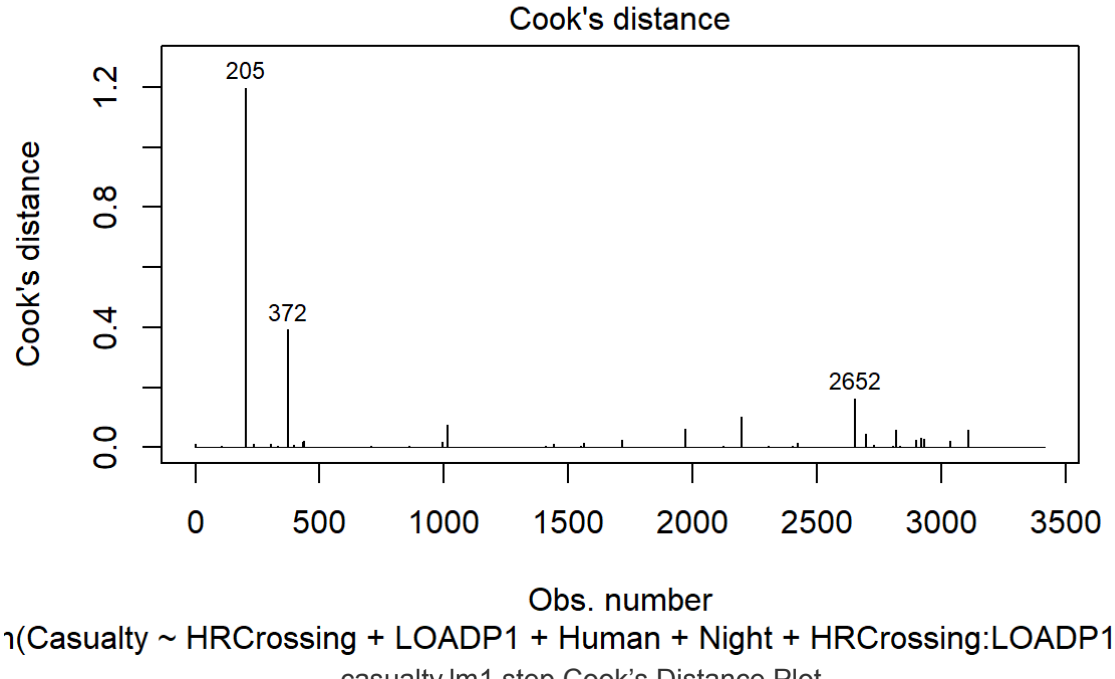
ı(Casualty ~ HRCrossing + LOADP1 + Human + Night + HRCrossing:LOADP1 + I

casualty.lm1.step Residuals vs. Fitted Plot



ı(Casualty ~ HRCrossing + LOADP1 + Human + Night + HRCrossing:LOADP1 + I

casualty.lm1.step QQ Plot



casualty.lm1.step Cook's Distance Plot

Based on the Residual vs Fitted model, we determined the lack of fit assumption was met but that the constant variance assumption was violated. We, also, determined that the normality assumption was violated based on the QQ plot. The cook's distance graph also was very abnormal. After further analysis on leverage, we decided to remove the leverage point 205 from the dataset to diagnose problems. This resulted in our second full second order model; the only difference from the first being removing the observation from the data. We, once again, performed stepwise regression and determined the stepwise model (casualty.lm1_2.step) was a better model. We performed initial analysis, described above, and created diagnostic plots. Based on the Residual vs Fitted model, we determined that constant variance was still violated. We determined that the lack of fit assumption was still met but barely. Based on the QQ plot, we determined the normality assumption still wasn't met. The cook's distance graph, while better, still was being skewed by leverage points. We decided our next step to diagnosing problems was to remove more leverage points.

```
#REMOVE 205 NEW MODEL
casualty.lm1_2 = lm(Casualty~(HRCrossing + LOADP1 + Human + Night)^2,
                    casualties_nd[-205,])
```

```
casualty.lm1_2.step = step(casualty.lm1_2, trace=F)
#lm1_2.step better
anova(casualty.lm1_2, casualty.lm1_2.step)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	3405	562948.7	*NA*	*NA*	*NA*	*NA*
2	3408	562958.3	-3	-9.645352	0.01944666	0.9963164

2 rows

```
summary(casualty.lm1_2.step)
```

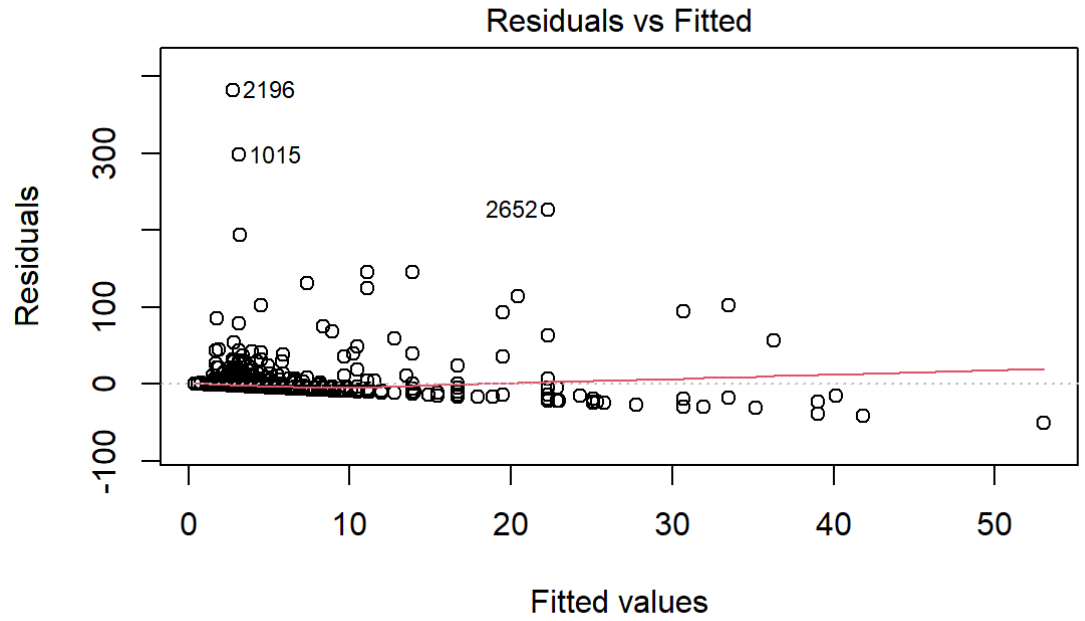
```
##
## Call:
## lm(formula = Casualty ~ HRCrossing + LOADP1 + Human + Night +
##     HRCrossing:LOADP1 + LOADP1:Human + LOADP1:Night, data = casualties_nd[-205,
##     ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.04  -1.74  -0.74  -0.70 382.22
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.77726    0.62471   4.446 9.04e-06 ***
## HRCrossing        -1.07427    0.67392  -1.594  0.11102
## LOADP1             0.76740    0.11784   6.512 8.49e-11 ***
## Human             -0.04198    0.76414  -0.055  0.95619
## Night              0.38575    0.56468   0.683  0.49458
## HRCrossing:LOADP1 -0.48706    0.12942  -3.764  0.00017 ***
## LOADP1:Human       2.02711    0.24740   8.194 3.54e-16 ***
## LOADP1:Night      -0.32674    0.11913  -2.743  0.00613 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.85 on 3408 degrees of freedom
## Multiple R-squared:  0.06228,    Adjusted R-squared:  0.06035
## F-statistic: 32.33 on 7 and 3408 DF,  p-value: < 2.2e-16
```

```
AIC(casualty.lm1_2.step)
```

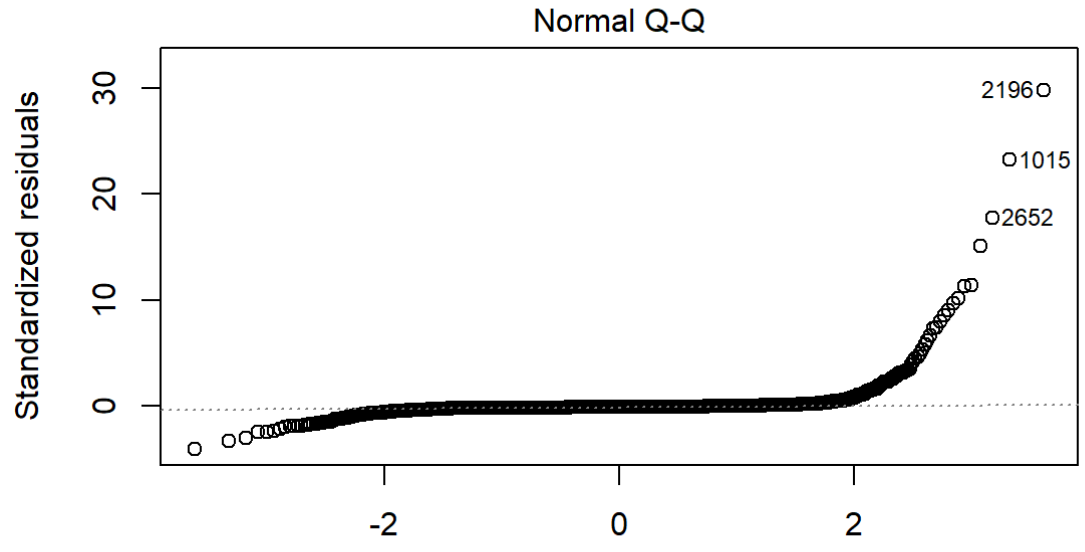
```
## [1] 27149.96
```

The second model we utilized was the same as the first full second order model, but we removed the observation 205, a leverage point, from the data. We ran stepwise regression on this model and compared the original model and the step model using an anova test. This test showed that the stepwise regression model (casualty.lm1_2.step) was better. This second step model was significant based on the global utility test. We, once again, ran t-tests on the terms of the model. In this model, highway rail crossing and human factors were both not significant. The interaction term between highway rail crossing and loadP1 was significant and had a negative relationship with the number of casualties. The interaction term between human factors and loadP1 was significant and had a positive relationship with the number of casualties. To further test the adequacy of this model, we found the adjusted R2 and AIC, 6.305% and 27149.96 respectively. The adjusted R2 for this new model is better than the first model we tested; however, it is still incredibly low showing that this model isn't great for predicting the number of casualties. The AIC is also smaller than the original model showing that this model is better than our first model. After running the diagnostic plots, we decided to remove more leverage points.



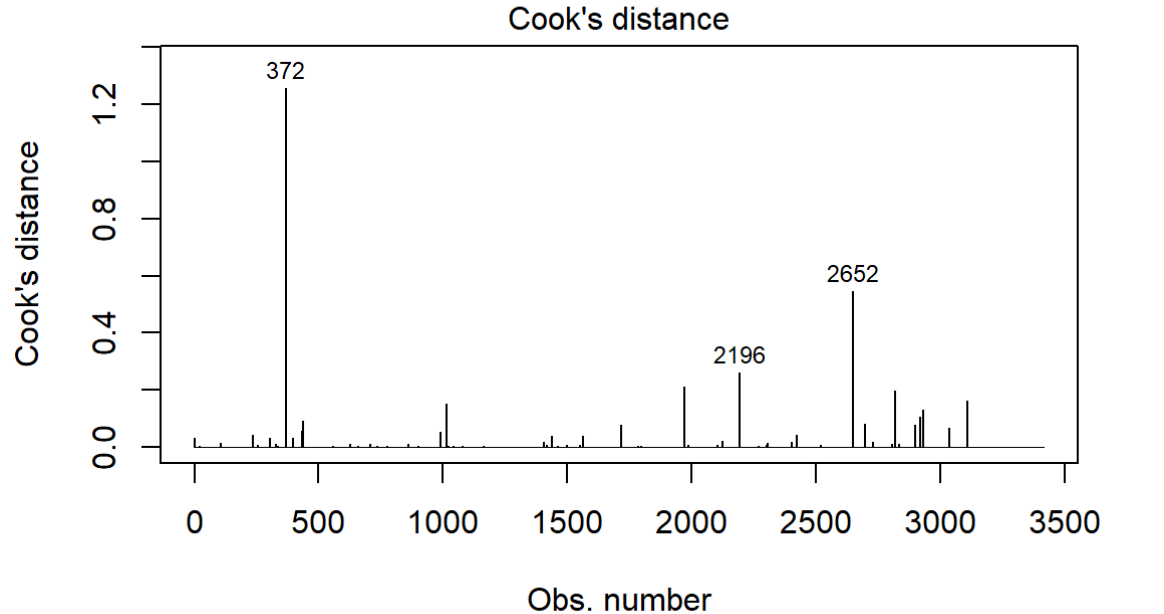
n(Casualty ~ HRCrossing + LOADP1 + Human + Night + HRCrossing:LOADP1 +

casualty.lm1_2.step Residuals vs. Fitted Plot



n(Casualty ~ HRCrossing + LOADP1 + Human + Night + HRCrossing:LOADP1 +

casualty.lm1_2.step QQ Plot



casualty.lm1_2.step Cook's Distance Plot

Adjusted Linear Models

```
#REMOVE 205, 372, 2652 NEW MODEL
casualty.lm1_3 = lm(Casualty~(HRCrossing + LOADP1 + Human + Night)^2, casualties_nd[-c(205,372,2652),])
casualty.lm1_3.step = step(casualty.lm1_3, trace=F)
#lm1_3.step better
anova(casualty.lm1_3, casualty.lm1_3.step)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	3403	498775.9	*NA*	*NA*	*NA*	*NA*

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
2	3407	499015.4	-4	-239.4509	0.4084256	0.8027065

2 rows

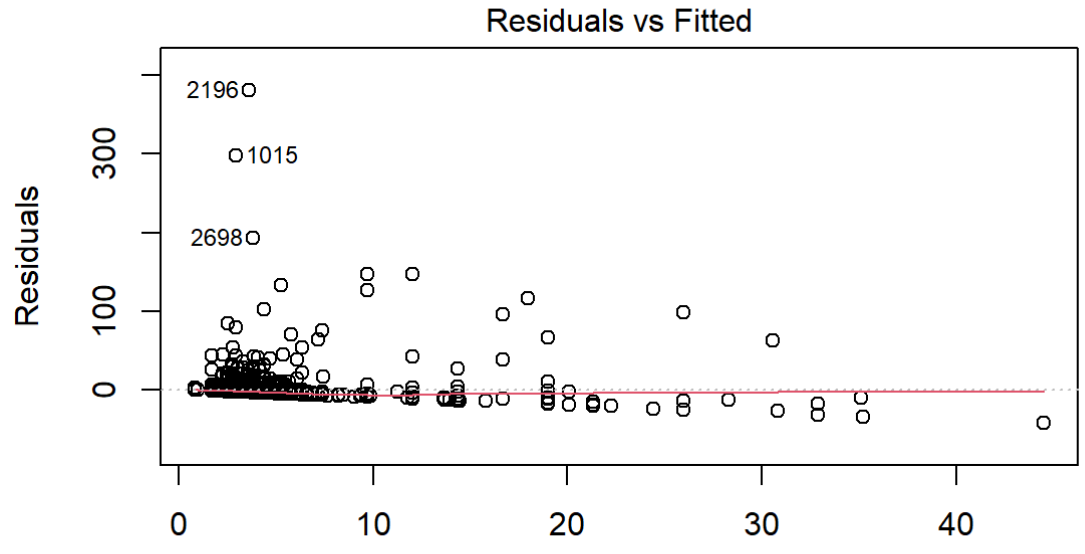
```
summary(casualty.lm1_3.step)
```

```
##
## Call:
## lm(formula = Casualty ~ HRCrossing + LOADP1 + Human + Night +
##     LOADP1:Human + LOADP1:Night, data = casualties_nd[-c(205,
##     372, 2652), ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41.51  -1.78  -0.76  -0.69 381.37
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.62708    0.55021   6.592 5.01e-11 ***
## HRCrossing   -1.93938    0.57459  -3.375 0.000746 ***
## LOADP1        0.27205    0.06331   4.297 1.78e-05 ***
## Human        -0.86426    0.69264  -1.248 0.212201
## Night         0.18980    0.53113   0.357 0.720846
## LOADP1:Human  2.04710    0.21622   9.467  < 2e-16 ***
## LOADP1:Night -0.17436    0.11270  -1.547 0.121933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.1 on 3407 degrees of freedom
## Multiple R-squared:  0.04432,    Adjusted R-squared:  0.04264
## F-statistic: 26.34 on 6 and 3407 DF,  p-value: < 2.2e-16
```

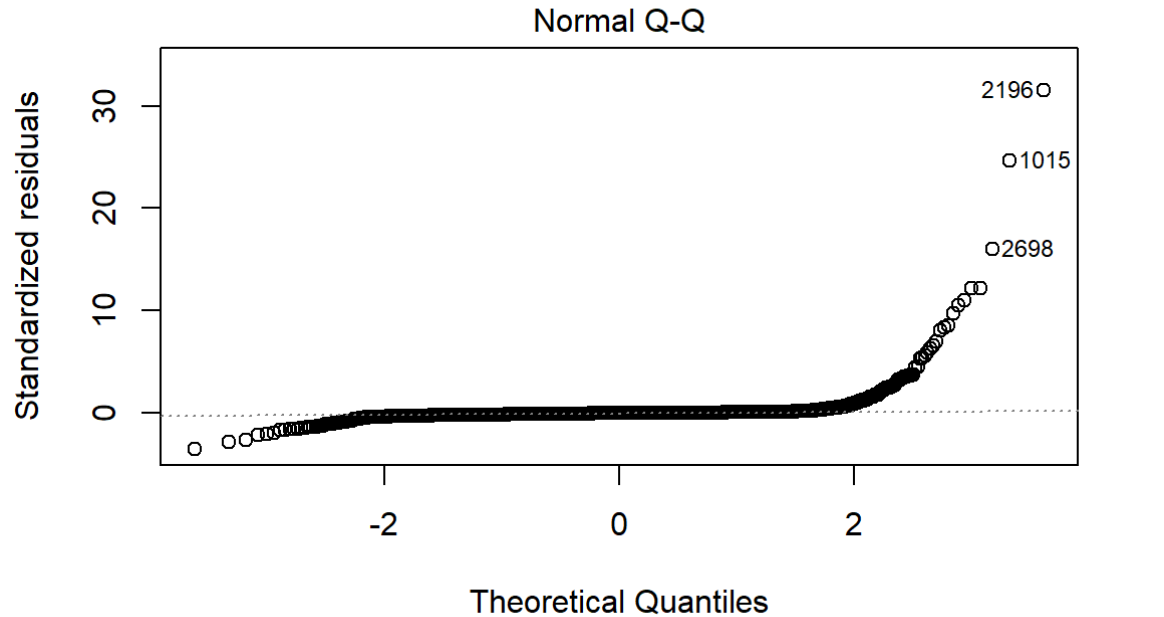
```
AIC(casualty.lm1_3.step)
```

```
## [1] 26722.46
```

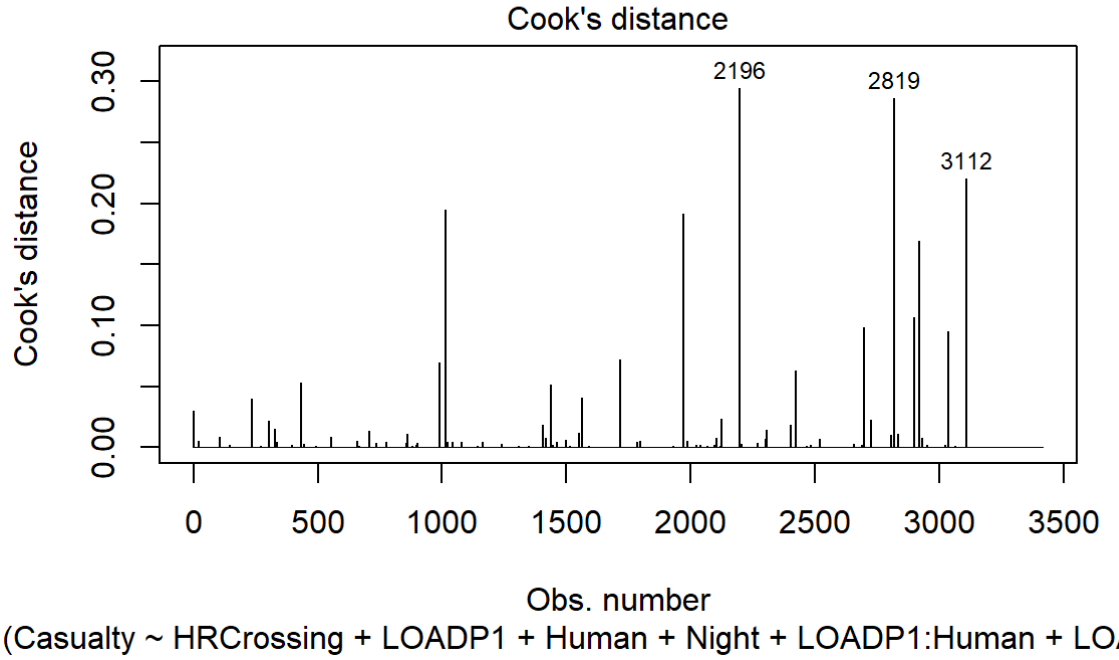
Next, we removed the leverage points 372, 2652, and the aforementioned 205 from the data set and remade our full second order model. Once again, we performed stepwise regression on the model and determined the step model (casualty.lm1_3.step) was best. We concluded that this model was significant based on the global utility test. To determine if individual terms in the model are significant and how they affect the number of casualties, we performed t-tests. In this model, highway rail crossing was significant and had a negative relationship with the number of casualties. The human factors term was not significant. The interaction term between human factors and loadP1 was significant and had a positive relationship with the number of casualties. To further test the adequacy of this model we found the adjusted R2 and AIC, 4.264% and 26722.46 respectively. The adjusted R2 is smaller than before meaning this model accounts for less variance than the second model we tested. The AIC, on the other hand, is the best for this model. Based on diagnostic plots and these assessments, we determined this step model with three leverage points taken out was the best. However, based on specific findings from the diagnostic plots, we decided to do a boxcox transformation on the model.



casualty.lm1_3.step Residuals vs. Fitted Plot



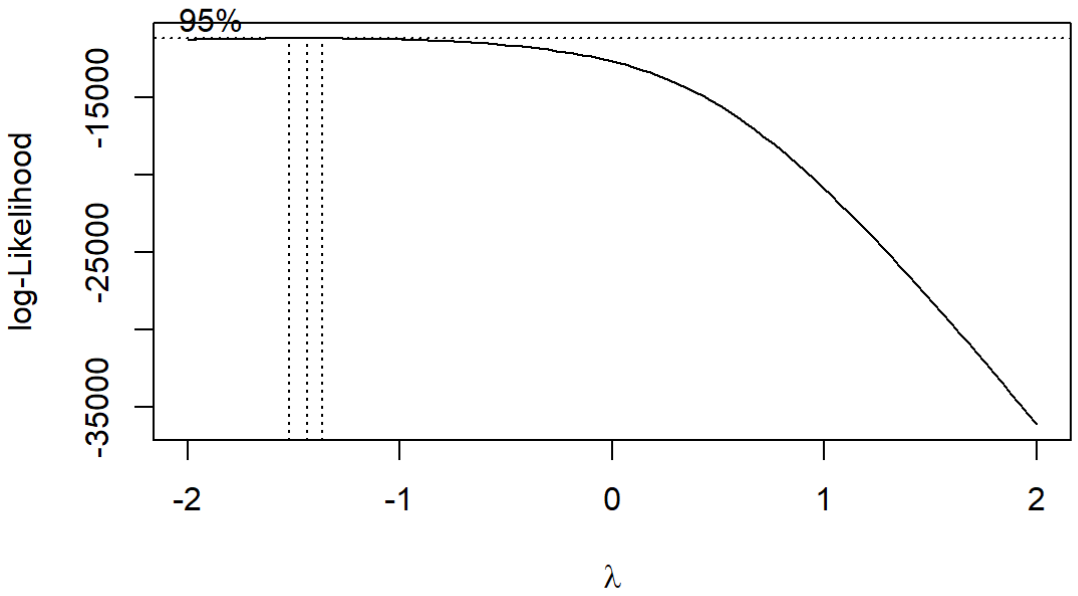
casualty.lm1_3.step QQ Plot



casualty.lm1_3.step Cook's Distance Plot

We removed observation 372, 2562, and 205 from the data set and created the same full second order model creating a third version of the first model. As for the first two versions of the model, we performed stepwise regression on this model and deemed it a better model using an anova test. Next, we did an initial assessment of the model, more details in the section above, and created diagnostic plots. The lack of fit assumption is met based on the Residual vs Fitted graph while the constant variance assumption was violated. The normality assumption still was not met based on the QQ plot. However, the cook's distance plot was much better meaning there weren't more leverage points for us to remove.

We decided to try and solve some of the assumption problems and do a boxcox transformation. This resulted in our final model, a boxcox transformation of casualty.lm1_3.step. We continued with the same process and did an initial assessment of the model as well as created diagnostic plots. The cook's distance was very similar to the model before. Based on the Residual vs Fitted graph, the model still violated the constant variance and the lack of fit assumption. The model still violated the normality function based on the QQ plot.



Boxcox Tranformation of casualty.lm1_3.step

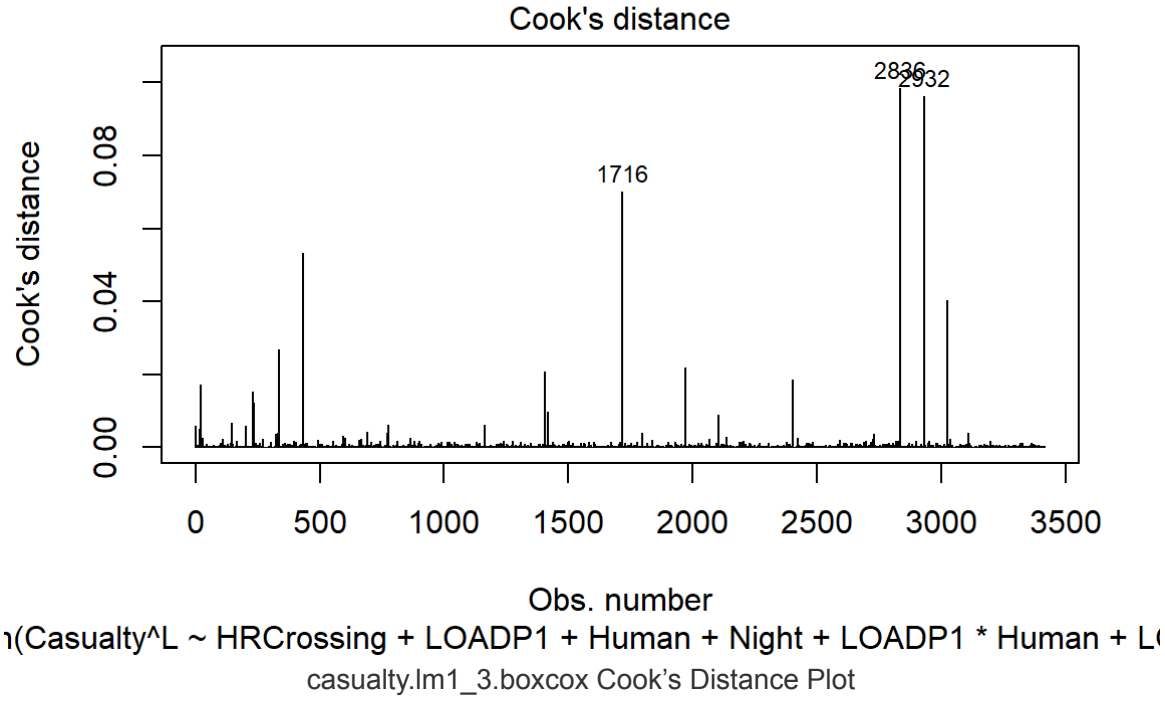
```
#lambda
boxcox(casualty.lm1_3.step, plotit = F)$x[which.max(boxcox(casualty.lm1_3.step,
                                                    plotit = F)$y)]
```

```
## [1] -1.4
```

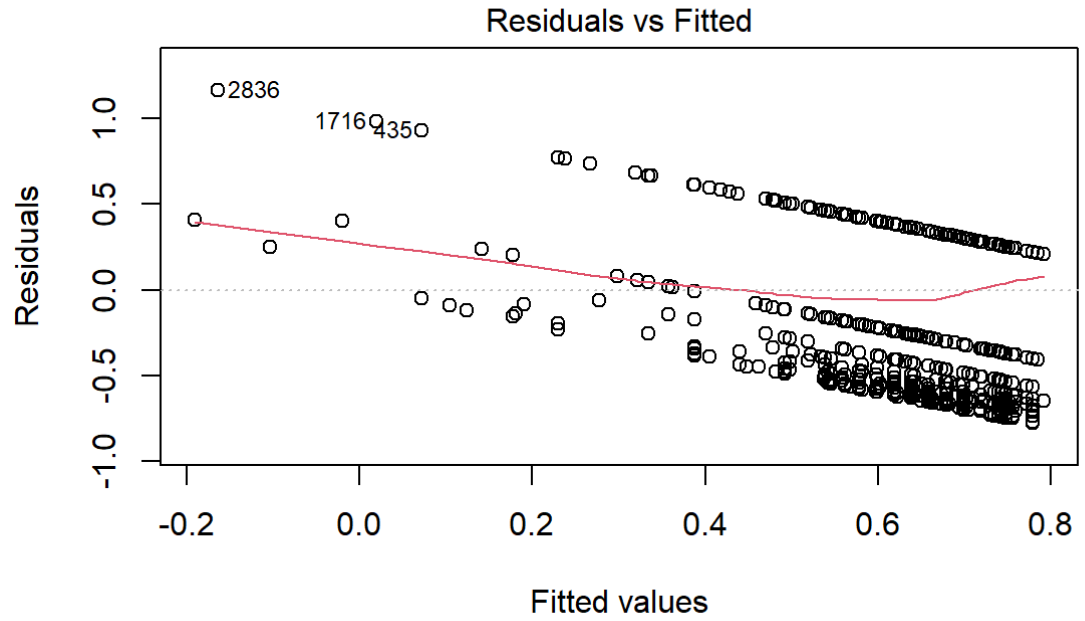
```
#max y value
max(boxcox(casualty.lm1_3.step, plotit = F)$y)
```

```
## [1] -11209.83
```

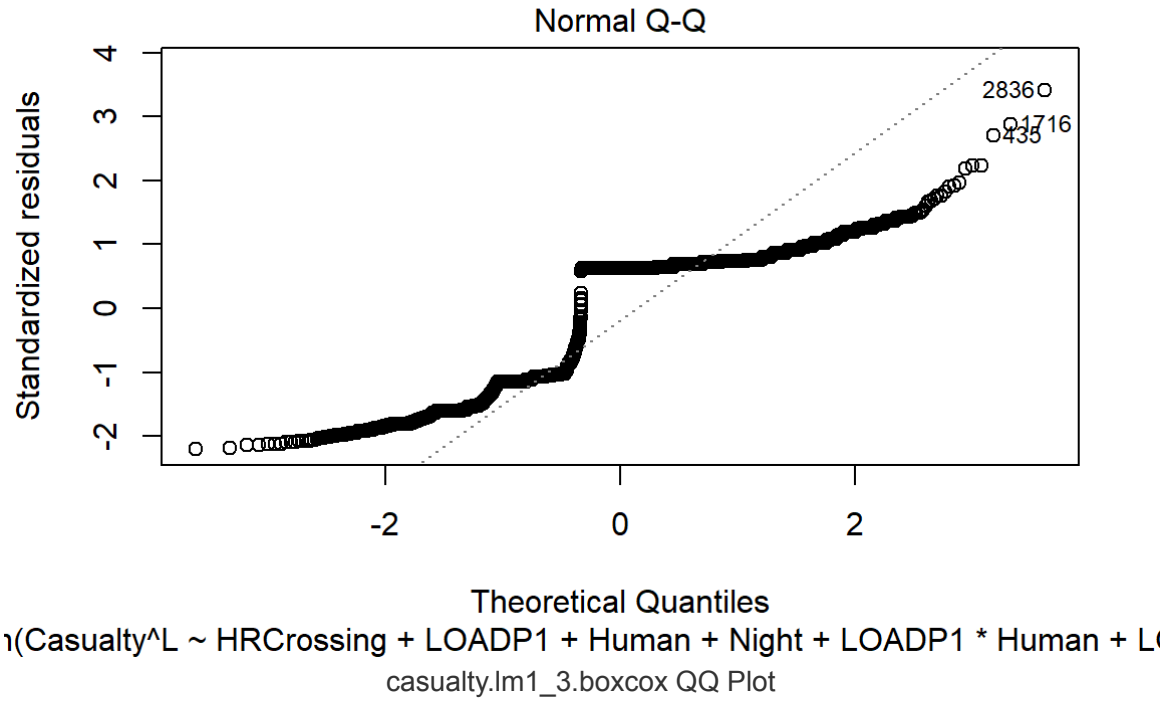
```
#store best lambda
L=
boxcox(casualty.lm1_3.step, plotit = F)$x[which.max(boxcox(casualty.lm1_3.step,
                                                    plotit = F)$y)]
#model with best lambda
casualty.lm1_3.boxcox<-lm(Casualty^L ~ HRCrossing + LOADP1 + Human +
                            Night + LOADP1*Human + LOADP1*Night, data =
                              casualties_nd[-c(205, 372, 2652), ])
```



casualty.lm1_3.boxcox Cook's Distance Plot



casualty.lm1_3.boxcox Residuals vs. Fitted Plot



casualty.lm1_3.boxcox QQ Plot

```
summary(casualty.lm1_3.boxcox)
```

```
##
## Call:
## lm(formula = Casualty^L ~ HRCrossing + LOADP1 + Human + Night +
##     LOADP1 * Human + LOADP1 * Night, data = casualties_nd[-c(205,
##     372, 2652), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7744 -0.3765  0.2207  0.2446  1.1638
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.742745   0.015945  46.581  < 2e-16 ***
## HRCrossing    0.036532   0.016652   2.194  0.02831 *
## LOADP1       -0.020065   0.001835 -10.935  < 2e-16 ***
## Human         0.012680   0.020073   0.632  0.52764
## Night        -0.007154   0.015392  -0.465  0.64212
## LOADP1:Human -0.032524   0.006266  -5.190 2.22e-07 ***
## LOADP1:Night  0.009647   0.003266   2.954  0.00316 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3507 on 3407 degrees of freedom
## Multiple R-squared:  0.05655,    Adjusted R-squared:  0.05488
## F-statistic: 34.03 on 6 and 3407 DF,  p-value: < 2.2e-16
```

```
AIC(casualty.lm1_3.boxcox)
```

```
## [1] 2543.553
```

The last model we tested for our hypotheses against the number of casualties was a boxcox transformation of the stepwise regression model of a full second order model with leverage points removed from the dataset. Based on the global utility test, we concluded that this model was significant. Next, we performed t-test to determine if individual terms in the model are significant and how they affect the number of casualties. In this final model, highway rail crossing was significant and had a positive relationship with the number of casualties; however, human factors term was not significant. The interaction term between human and loadP1 was significant and had a negative relationship with the number of casualties. To further test the adequacy of this model we found the adjusted R2, 5.488%, and AIC, 2543.553. The adjusted R2 is small meaning this model accounts for less variance than the second model we tested. The AIC, on the other hand, is the best for this model.

Based on diagnostic plots and these assessments, we determined **casualty.lm1_3.step was the best model for predicting the number of casualties**. This is because although the boxcox tranformation model from this step one had an extremely lower AIC value, it caused a lot of the assumptions for linear regression to no longer be met, included heteroscedasticity, lack of fit, and non-Gaussian distribution. However, the step function model did not display a lack of fit although there was non-constant variance. Even though the QQ plot showed a bit of a deviation from the Gaussian distribution line, this diagnostic plot appears much more similar to the line then that of the boxcox tranformation model.

Evidence-Based Recommendations

ACCDMG

For accident damage, we are able to reject the null hypothesis in both situations; for the first hypothesis despite freight showing a negative relationship with accident damage, it's interaction with rack showed an increase in severity of accident damages. Our final model selection was xdmg.lm1_2.boxcox, the boxcox transformation of the step function formed from the linear model with leverage points removed. The model was ACCDMG^L ~ Freight + TRNSPD + WeatherType + Rack + Freight * Rack + TRNSPD * Rack + WeatherType * Rack, where L was -0.5.

Based on this, our recommendation to the FRA is to verify safety conditions on railroads and pathways frequently taken by freight trains, alongside more frequent maintenance checks . Similarly, the interaction between rack and precipitation has a statistically significant likelihood of increasing the severity of accident damage. This statistically significant coefficient suffices to reject the second null hypothesis. The previous recommendation stands for this as well, in addition conducting further analysis and safety testing of tracks in situations of rain, sleet and snow.

Casualties

For casualties, we failed to reject the null hypothesis and conclude Human factors do not result in an increased number of casualties. Our final model section was casualty.lm1_3.step, which was the step function made from the full second order model with leverage points removed. The model was Casualty ~ HRCrossing + LOADP1 + Human + Night + LOADP1:Human + LOADP1:Night.

We recommend research and automation tools, reviewing safety protocols and implementing the necessary procedures, and doing further analysis for a more specific cause of human factors. However, we reject the null hypothesis for highway rail crossing and recommend conducting further analysis in these situations, and reviewing rail crossing procedures and safety measures. Since there were low summary statistics, we recommend conducting another analysis with other variables as the model showed a violation of regression assumptions and low R^2.