Linear Statistical Models

SYS 6021

Project 1

Analysis of Train Accidents in the U.S. During 2001-2013

Debajyoti Datta
dd3ar@virginia.edu

Summary

The train accidents occur due to a variety of reasons and in this case severe train accidents have been analyzed in terms of the accident damage and the total number of casualties in such accidents. It can be concluded with 99% confidence that the trainspeed is positively correlated with the accident damage in sever train accidents and the confidence interval for the 99% confidence is [11626.663, 15639.337]. Head on collisions kill more people in train accidents than derailment even though derailment is the most common cause of train accidents in the FRA data of severe train accidents. The 99% confidence interval for the coefficient of head on collisons in measuring the total number of people killed is [0.4739, 0.7721]. Also the third observation is that rack, road bed structures cause more severe accident damage than human factors. So if the FRA's regulations related to trainspeed should be lowered conditioned on the number of cars and the tons of the train. More steps need to be taken to prevent head on collisions since that causes more casualties, even though derailment happens to be the major cause of train accidents. Track maintenance or prevention of problems arrising due to rack, road bed structures can lead to significant reduction in the economic loss through accident damage. So some of the actionable steps that need to be taken are more regular maintenance of tracks to prevent accidents caused due to rack, road bed and structures and lowering of trainspeed in the vicinity of another train or proper track management to prevent head on collisions.

Honor Pledge: On my honor, I pledge that I am the sole author of this paper and I have accurately cited all help and references used in its completion.

1. Problem Description

1.1 Situation

US railroad accidents [3] occur due to a variety of reasons from mechanical and electrical failures to human factors. These are not only responsible for the loss of human lives, but are also responsible for severe economic damage. This project [1] uses the template [2] for analyzing these rail accidents in terms of the economic loss and loss of human lives.

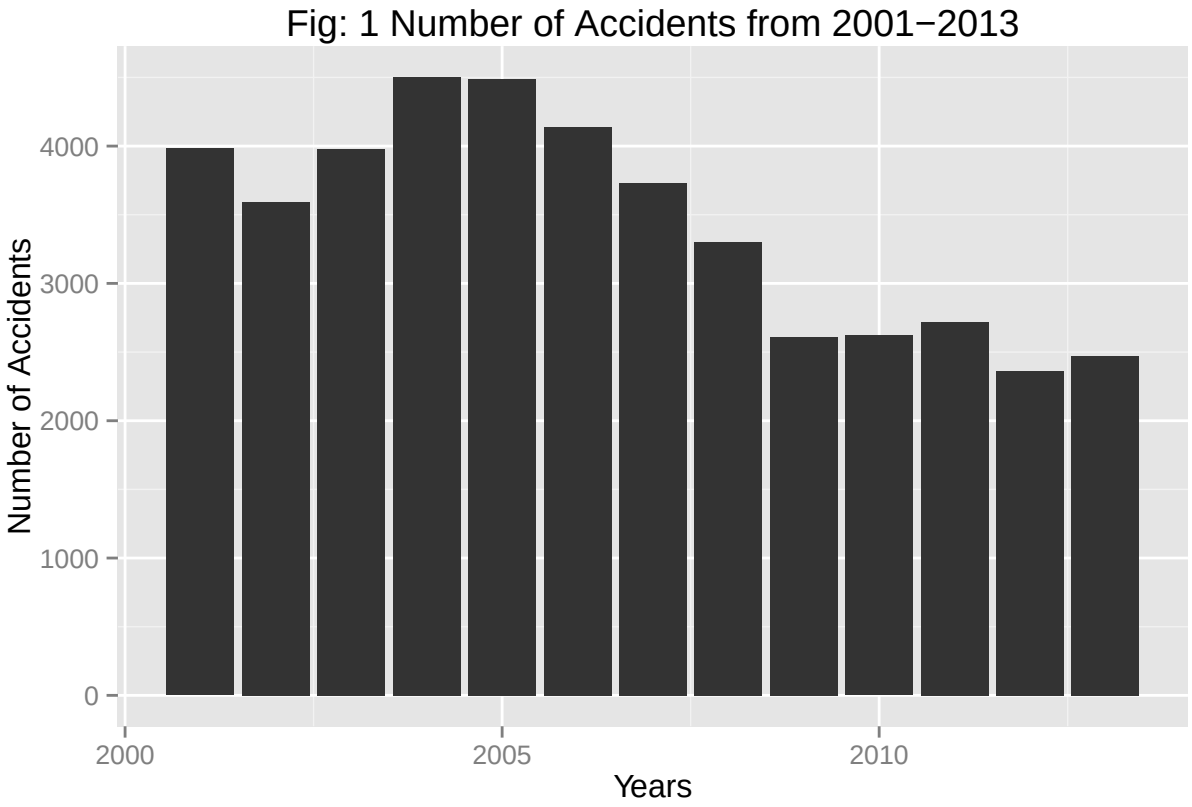
For that purpose of exploring the severity of rail accidents two datasets were created. One for exploring the severity based on the accident damage that was incurred by the FRA and the other based on the total number of casualties (total number of injuries).

In this case, we are looking at the really sever rail accidents or the extreme points of the datasets in terms of the economic damage and the other dataset in terms of the total number of injuries.

The current safety conditions and the present variation of accidents over the number of years and the various causes of accidents have been shown below.

The total number number of accidents that have happened is 44506 from the year 2001 to 2013.

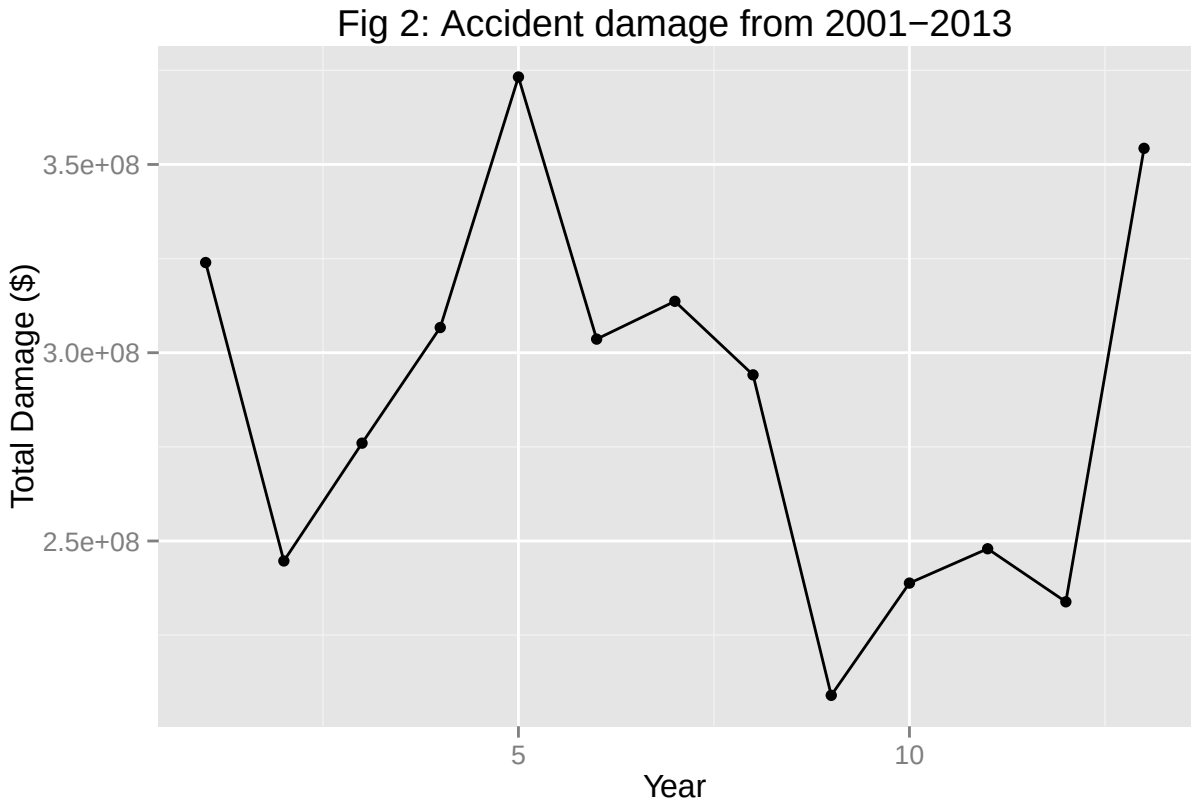
The number of accidents across the last 10 years is as follows:



Also as can be seen that total number of accidents have decreased since 2005. Even though this is a huge number, most of the accidents were not serious or there were very few casualties.

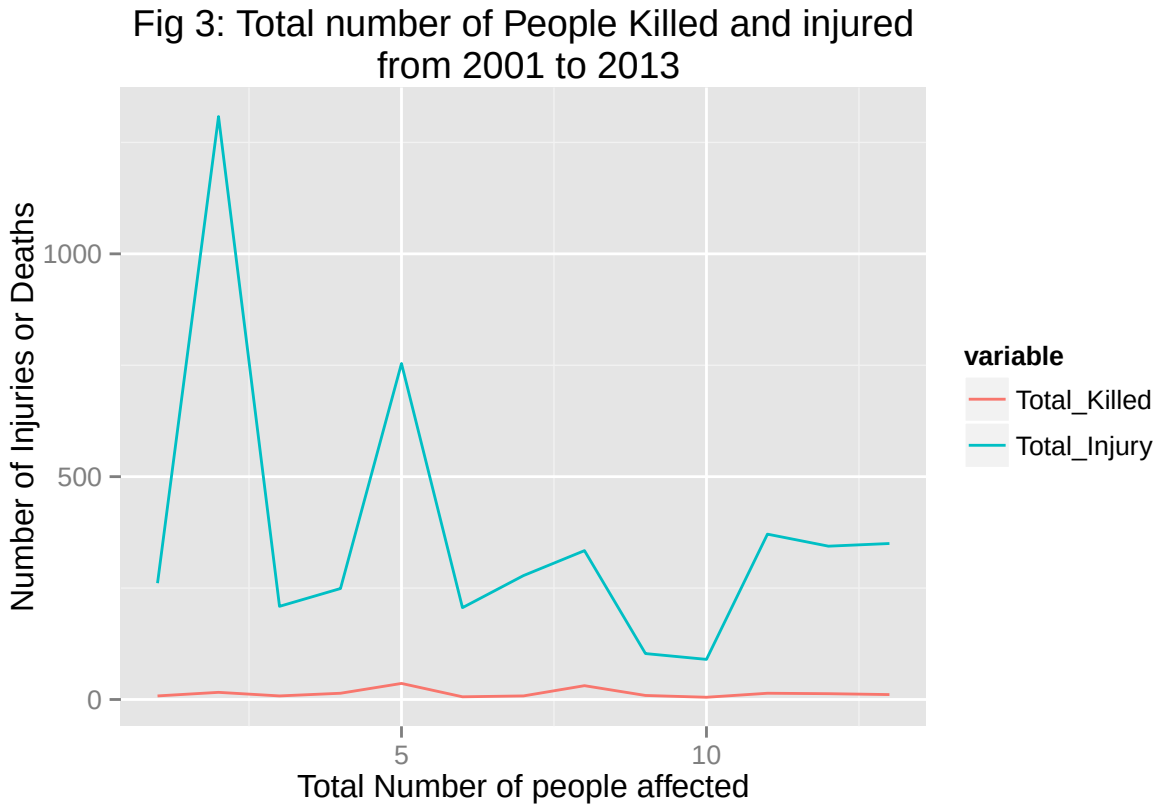
The most important factor that is of concern to the FRA is the loss of human lives and the economic damage.

The amount of economic damage in this accident in the years 2001-2013 is as follows:



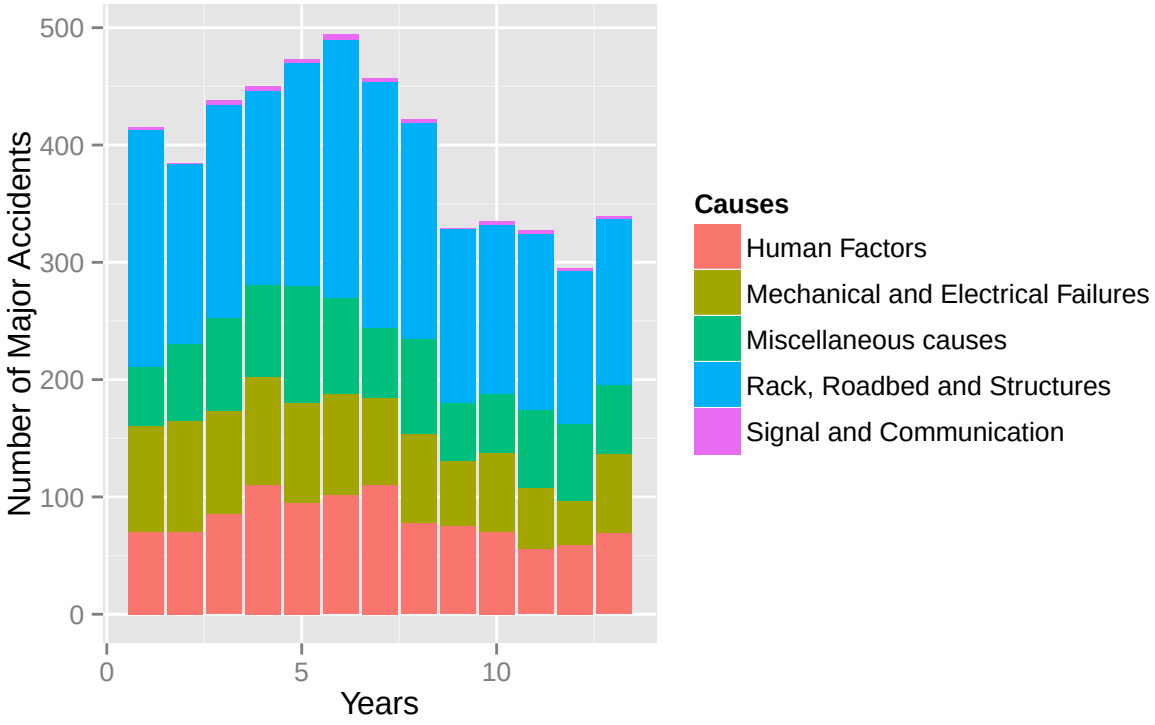
The amount of economic damage has been a cause of concern for the FRA and even though total number of accidents have decreased over the years, the amount of economic loss through accident damages have widely varied over the years. 2009 had the least amount of money lost in accident damages, but other than that there are not any significant patterns about the costs incurred through accident damages.

Other than the economic damage, another metric that we are looking at is the number of people who were killed and injured. The following graph highlights the number of people who were killed and injured.



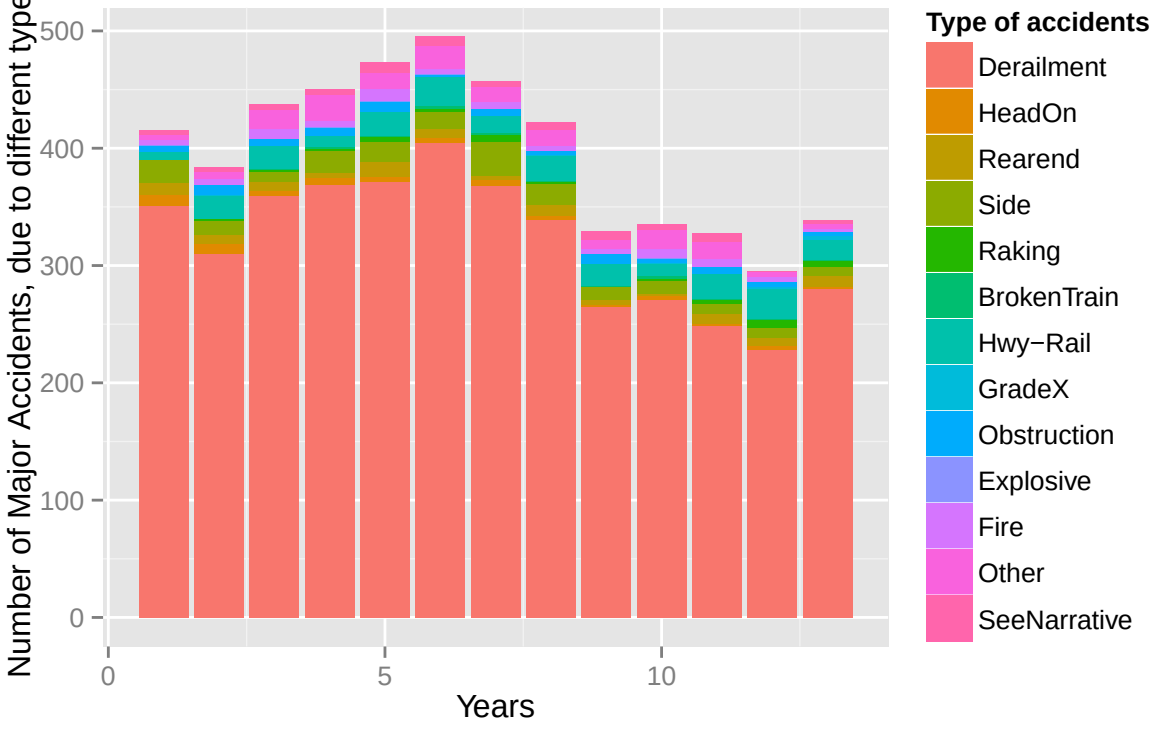
The total number of people injured have decreased on an average over the years, with some varitions in the years 2005, 2007, 2011 when there were peaks in the graph, but the total number of deaths have been more or less constant. It is worth noting however that the total number of deaths have been much lesser than the total number of injuries.

Fig 4: Number of major accidents due to different causes



There have been various factors responsible for the different causes of the accidents and from the stacked bar chart above it can be seen that rack, road bed structures cause more accident damage per year than the other factros like human factors.

Fig 5: Number of Accidents based on the type of the accident



As it is clearly evident from the graph that most of these major accidents occured due to Derailment. In fact derailment is the main source of accidents even when all accidents are observed across the last 14 years.

So some of the factors that do seem important from the above is the amount of economic damage incurred by the FRA and the total number of casualties. Derailment happens to be one of the major causes of accidents and it needs to be explored in the context of how much damage it is causing or the number of casualties caused due to derailment. Rack, road bed structures also seemed to cause the most amount of accident damage and it would be worth exploring the amount of accident damage incurred by the FRA due to the various causes of train accident.

1.2 Goal

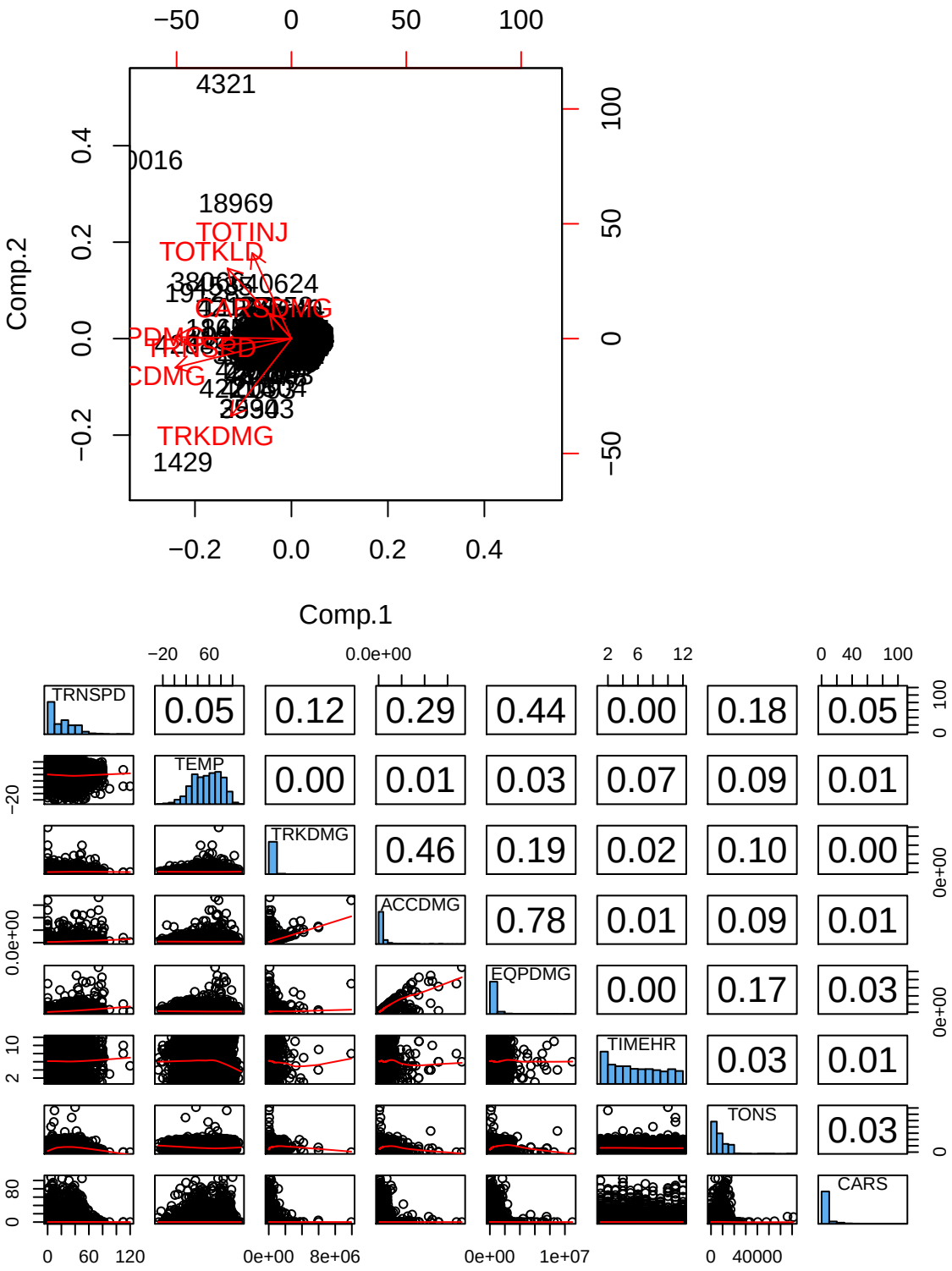
Thus the goal is to reduce the severity of Rail accidents and to recommend the actionable steps that needs to be taken by the Federal Railroad Administion to decrease the amount of accident damage and the number of casualties in rail accidents.

1.3 Metrics

- Reduce the amount of economic loss for FRA
- Reduce the amount of loss of human lives and the number of people injured.

1.4 Hypothesis

After doing a principal component analysis on the following variables, “CARSDMG”,“TRNSPD”,“EQPDMG”, “TRKDMG”,“ACCDMG”, “TOTKLD”, “TOTINJ”, we see that according to the principal component 1, equipment damage, accident damage and trainspeed (in this order) accounts for the majority of the variance in the data set that we are using. This data set essentially is analyzing the most severe accidents in terms of economic damage. These essentially involve the points that lie beyond the 1.5 times the interquartile range of the train data set. This data set has 5449 points out of the 44506 data points and we will concentrate on these accidents only.



From the scatter plot matricies for quantitative predictors, the above fact is further verified that the accident damage and trainspeed have the highest correlation among all the other variables that are being measured here. (Even though equipment damage has a higher correlation with TRNSPD, that has already been accounted for in ACCDMG).

Thus an hypothesis in this case based on quantitative variable Trainspeed is:

H0: In the FRA data of Train accidents, in the top 5449 accidents in terms of the economic damage, higher

train speeds do not lead to higher severity of accident damage.

H1: In the FRA data of Train accidents, in the top 5449 accidents in terms of the economic damage, higher train speeds do lead to higher severity of accident damage.

Another hypothesis based on a qualitative variables of type of accidents are as follows:

H0: Head-On Collisions of Trains do not kill more people than Derailments in severe Train Accidents, even though Derailments account for majority of the accidents under severe accident damages.

H1: Head-On Collisions of Trains do not kill more people than Derailments in severe Train Accidents, even though Derailments account for majority of the accidents under severe accident damages.

Another hypothesis based on the cause of the accident.

H0: Rack, Roadbed and structure related accidents do not cause more severe accidents than human factors in terms of high economic damage.

H1: Rack, Roadbed and structure related accidents cause more severe accidents than human factors in terms of high economic damage.

2. Approach

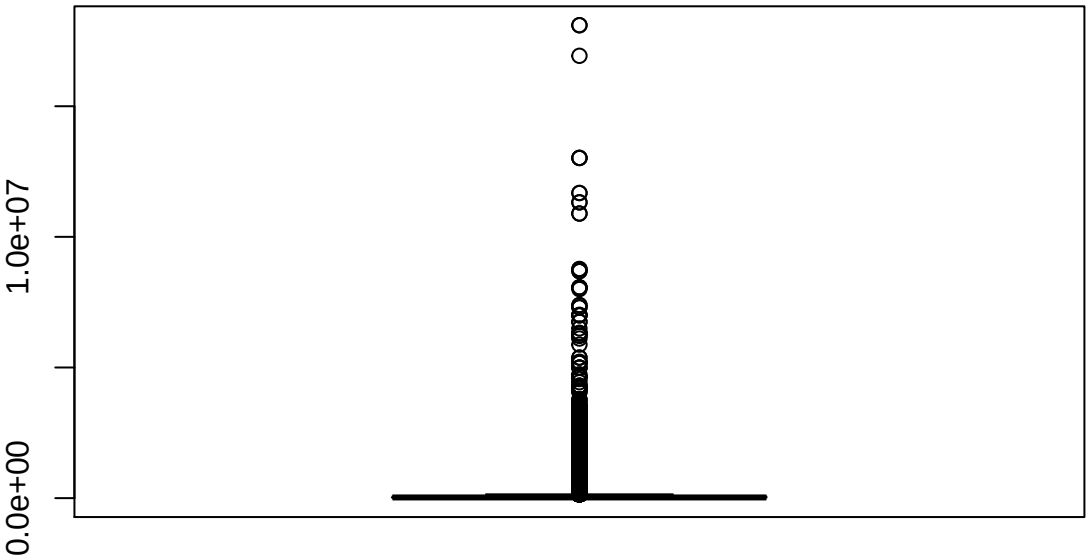
2.1 Data Description

The data [3] used in this study is from the year 2001-2013. There are a total of 44506 data points. As mentioned previously these accidents lead to a huge amount of economic damage and also lead to a loss of injuries and deaths.

About the top 13% of the accidents cause most amount of the damage. In this case we are trying to find out the points above the upper whisker in the box plot. One data point particularly the one that was part of the 9/11 attack was removed, since we are only interested in this dataset to make recommendations for the FRA.

Also in the extreme damage dataset that we are using, the 13% of the data set that account for the maximum amount of economic damage does not have missing values in the columns we are interested in. Trainspeed, accident damage cost, equipment damage cost are some of the columns that do not have any missing values.

Total number of such severe accidents is:




```
## [1] 5449
```

The proportion of accidents that are severe:

```
## [1] 0.1303
```

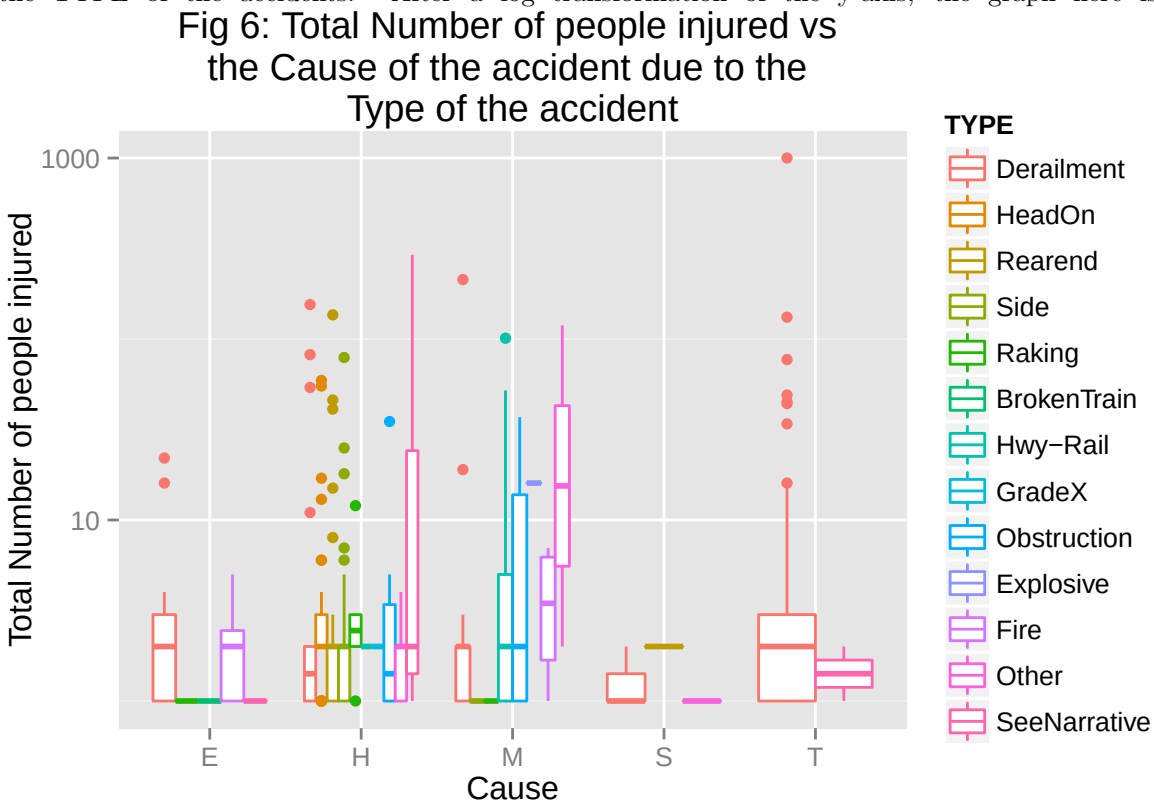
The proportional cost of these accidents:

```
## [1] 0.74
```

As we can see that 13% of the accidents cause the most amount of economic damage, lets concentrate on these 13% of the accidents since they are economically significant.

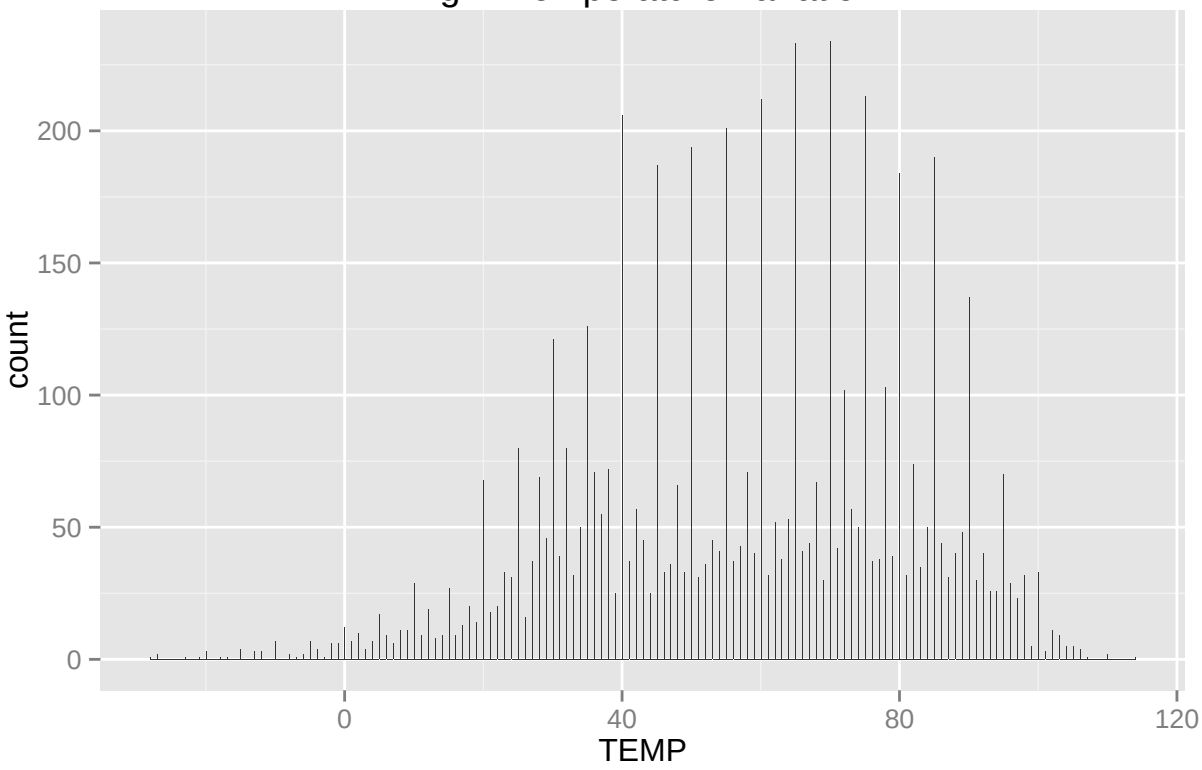
Lets look at these accidents more closely.The following discusses the distribution of these extreme accidents based on causes over the years.

Now in order to study the variations of accidents for the major accidents, two transformations were done for more information. Firstly quite a few of these accidents did not have any injuries. Thus excluding such accidents using the subset command, we get 556 such incidents where at least one person was injured. We then use the seggregate the color of the boxes with the TYPE of the accidents. After a log transformation of the y-axis, the graph here is below.



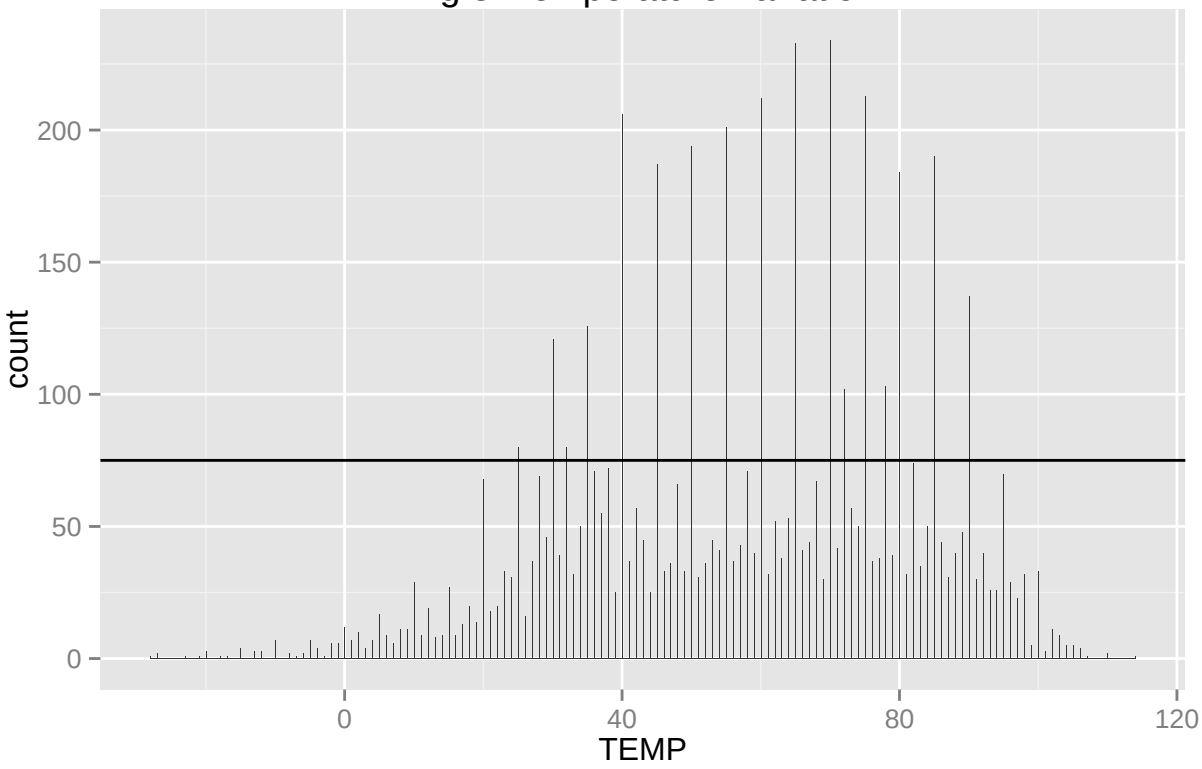
There are quite a few instances of bias in this data set, especially the ones that were reported by human-beings like the temperature and the speed of the train.

Fig 7: Temperature Variation

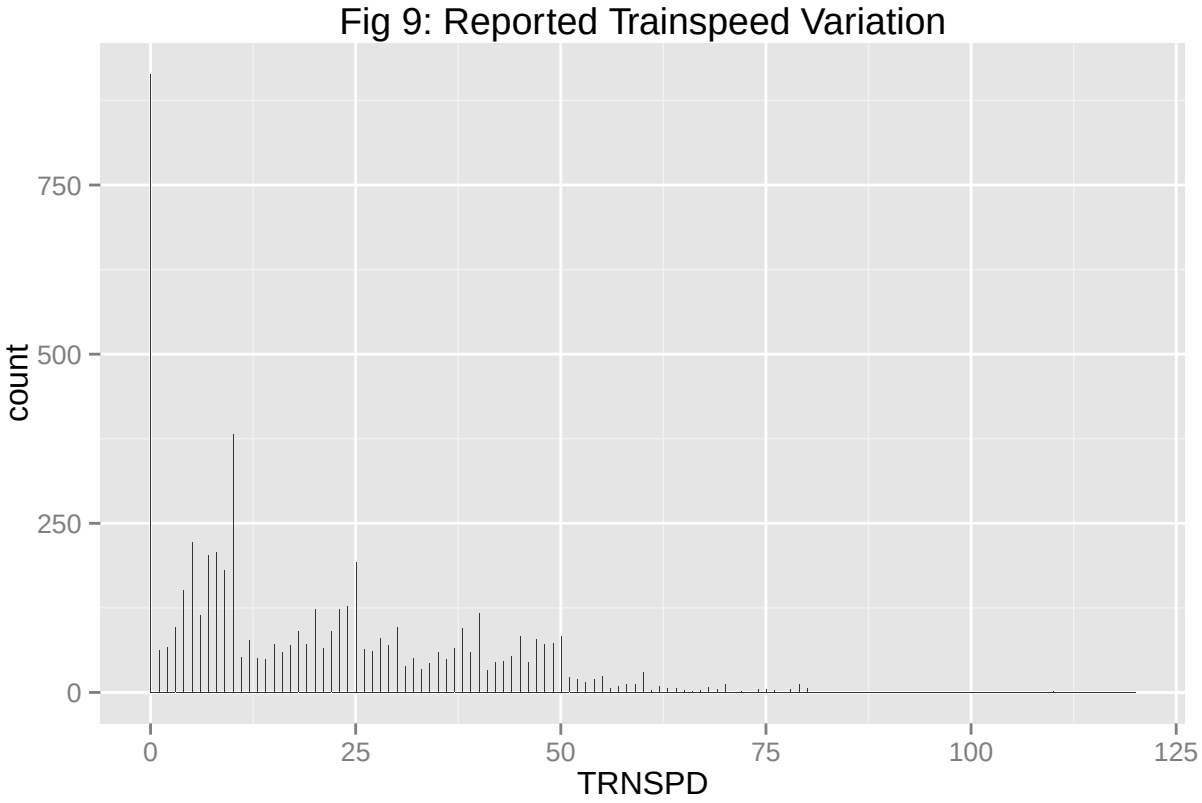


If we know look at a subset of the temperature whose count is above 75, we observe, that most of those temperature values are such that they are multiples of 0s and 5s. This is because of human inclination to such numbers.

Fig 8: Temperature Variation



A similar trend can be observed with Trainspeed.



2.2 Analysis

Analysis for the first Hypothesis:

Since we are exploring the factors that lead to high amount of accident damage, we are now exploring the various factors that lead to the severity of the accident damage.

Here accident damage which is the response variable is, as we had seen before is a continuous variable, and we are exploring only the continuos and discrete predictor variables. The categorical variables will be explored in a later section.

Now in order to find the accident damage lets explore the variables Temperature, Train Speed, Tons, Cars and the number of head end locomotives.

Here the R^2 is 0.09289 and the adjusted R^2 is around 0.9201. Since the adjusted R^2 is very close to R^2 the penalty for having many terms to fit the linear model for accident damage is hardly significant. However the correlation value is itself very small.

There can be various factors that can lead to such a small correlation value and heteroscedasticity is one of them. It is the phenomenon in which some observations are less reliable than others and should be downweighted in a fitting procedure.

The important point to remember is that heteroscedasticity does not cause ordinary least square (OLS) to be biased, but it does make the OLS inefficient. [7]

Since the model has a low R^2, it makes sense to see which combination of these variables gives the best model. This can be achieved through stepwise regression.

The results of the stepwise regression do not lead to a significant difference. According to the stepwise regression the final model should have TRNSPD, TONS, and HEADEND1 (No of head end locomotives) as

the predictor variables for the response variable ACCDMG or Accident Damage. The AIC difference between the initial model and the final model is hardly significant.

Initial Model:

$$\begin{aligned} ACCDMG = & \beta_0 \\ & + \beta_1 * TRNSPD \\ & + \beta_2 * TONS \\ & + \beta_3 * CARS \\ & + \beta_4 * HEADEND1 \end{aligned} \tag{1}$$

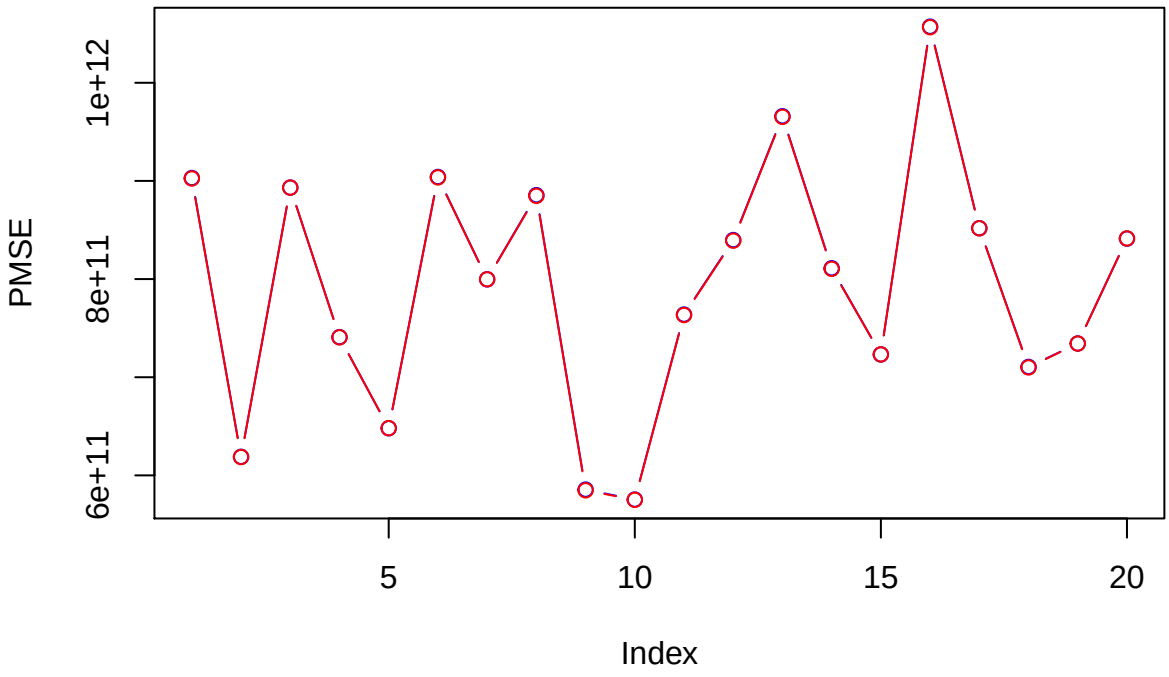
Final Model:

$$\begin{aligned} ACCDMG = & \beta_0 \\ & + \beta_1 * TRNSPD \\ & + \beta_2 * TONS \\ & + \beta_3 * CARS \end{aligned} \tag{2}$$

But in order to verify if these actually play any role two further tests can give us more data.

Anova test of these two models and or comparing the RMSE over a certain number of iterations can provide us more information.

Fig 10: Model Comparison Based on PMSE



Even with 20 runs of the difference is hardly noticeable.

Lets now observe another model and compare it with the exisiting Initial Model to see if this performance is better.

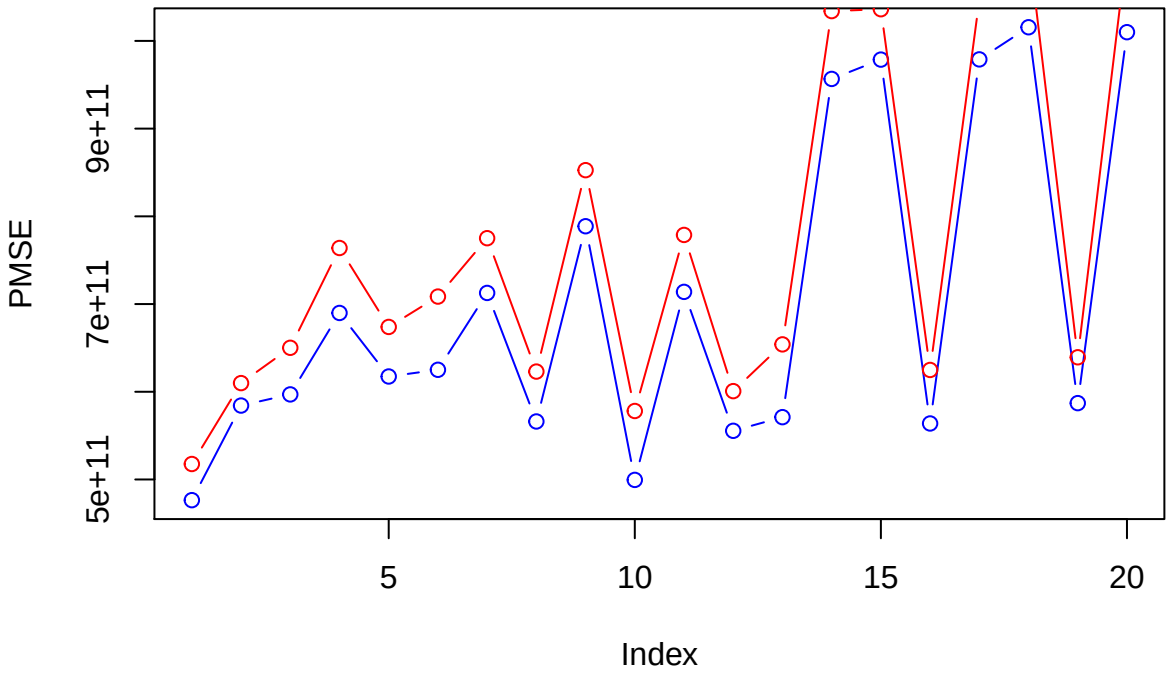
New Model:

$$\begin{aligned} ACCDMG =& \beta_0 \\ &+ \beta_1 * TEMP \\ &+ \beta_2 * TONS \\ &+ \beta_3 * CARS \end{aligned} \tag{3}$$

Compared to Initial Model:

$$\begin{aligned} ACCDMG =& \beta_0 \\ &+ \beta_1 * TRNSPD \\ &+ \beta_2 * TONS \\ &+ \beta_3 * CARS \\ &+ \beta_4 * HEADEND1 \\ &+ \beta_5 * TEMP \end{aligned} \tag{4}$$

Fig 11: Model Comparison Based on PMSE



From the visual inspection it appears that model 1 (in blue) is better than model 2 (in red).

In order to verify it we can do the t-test and the p-value happens to be 0.005651 signifying that the null hypothesis can be rejected and there is a significant difference between the two model.

Now that we have the model, it would be interesting to look for heteroscedasticity, normality, and influential observerations. Eliminating these can help in generating better fits and it can prevent the distortion of standard errors. [7]

The following are the Diagnostic Graphs

Fig 12: Residual vs. Fitted

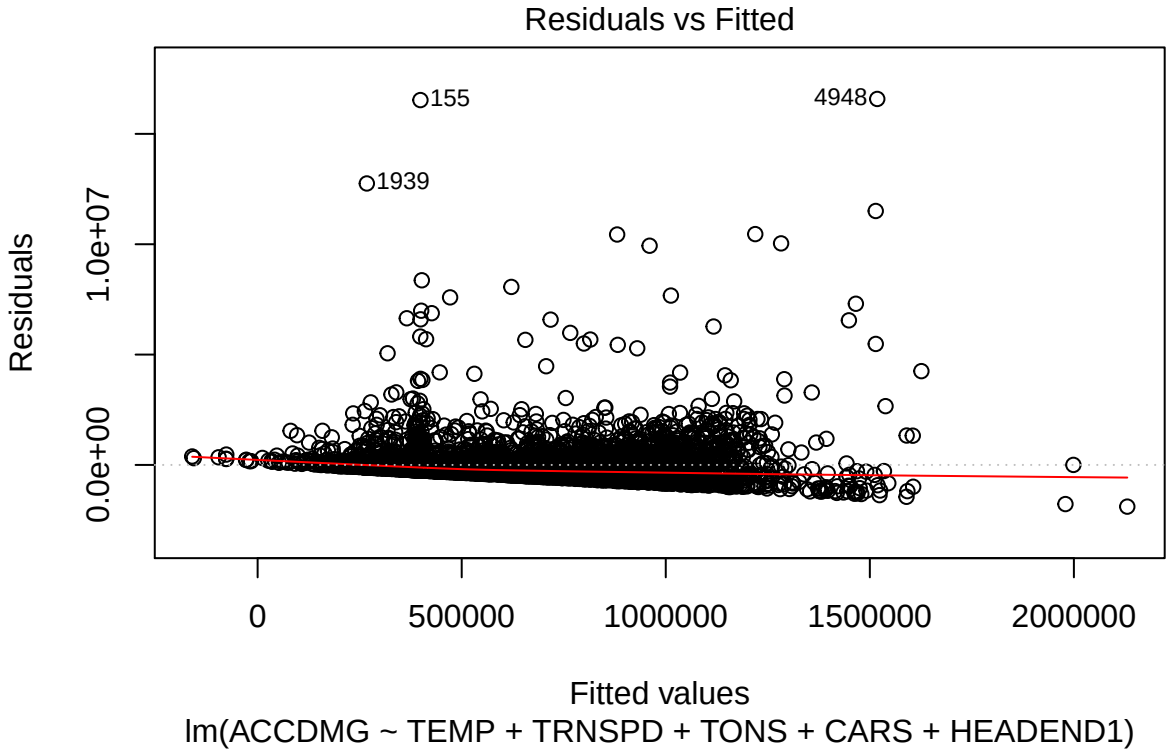


Fig 13: QQ Plot

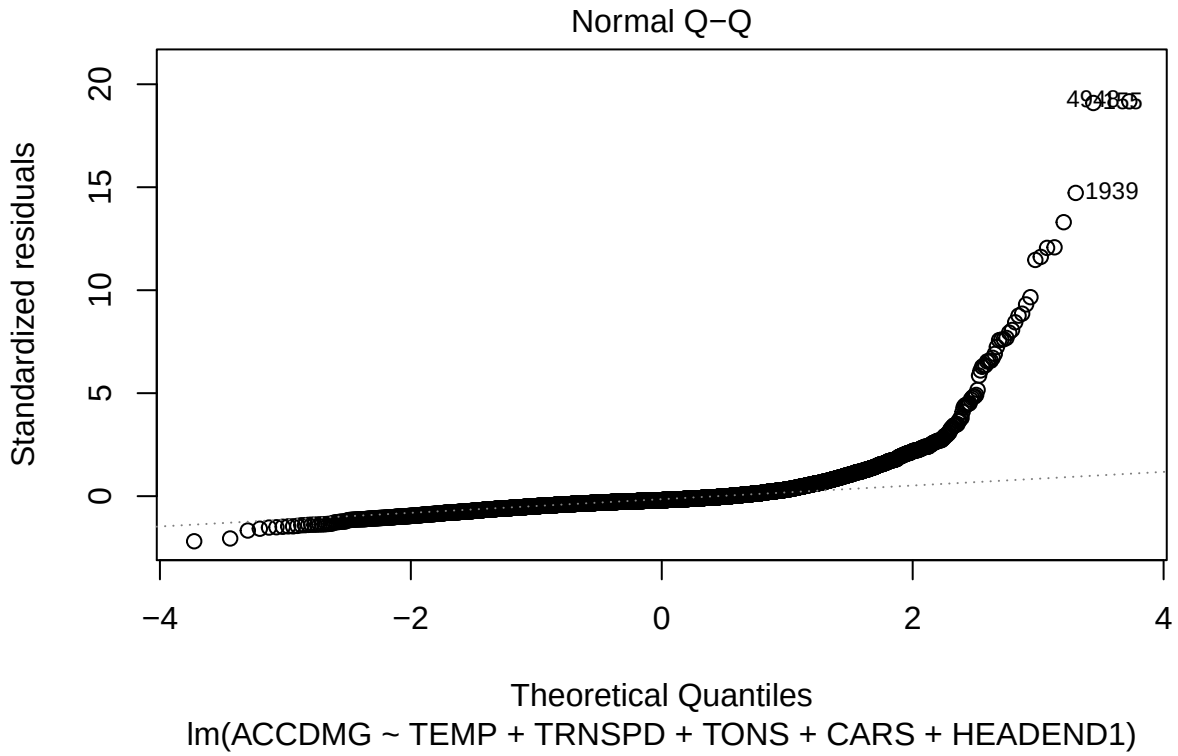


Fig 14: Scale Location Plot

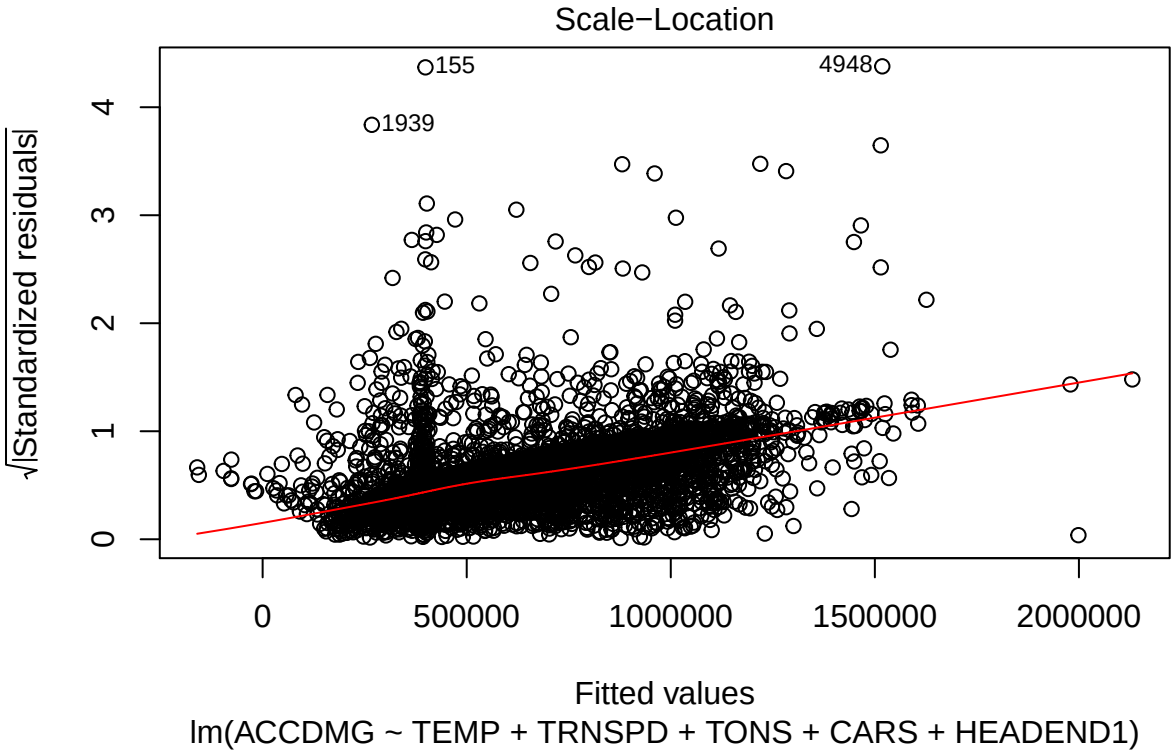


Fig 15: Cook's distance plot

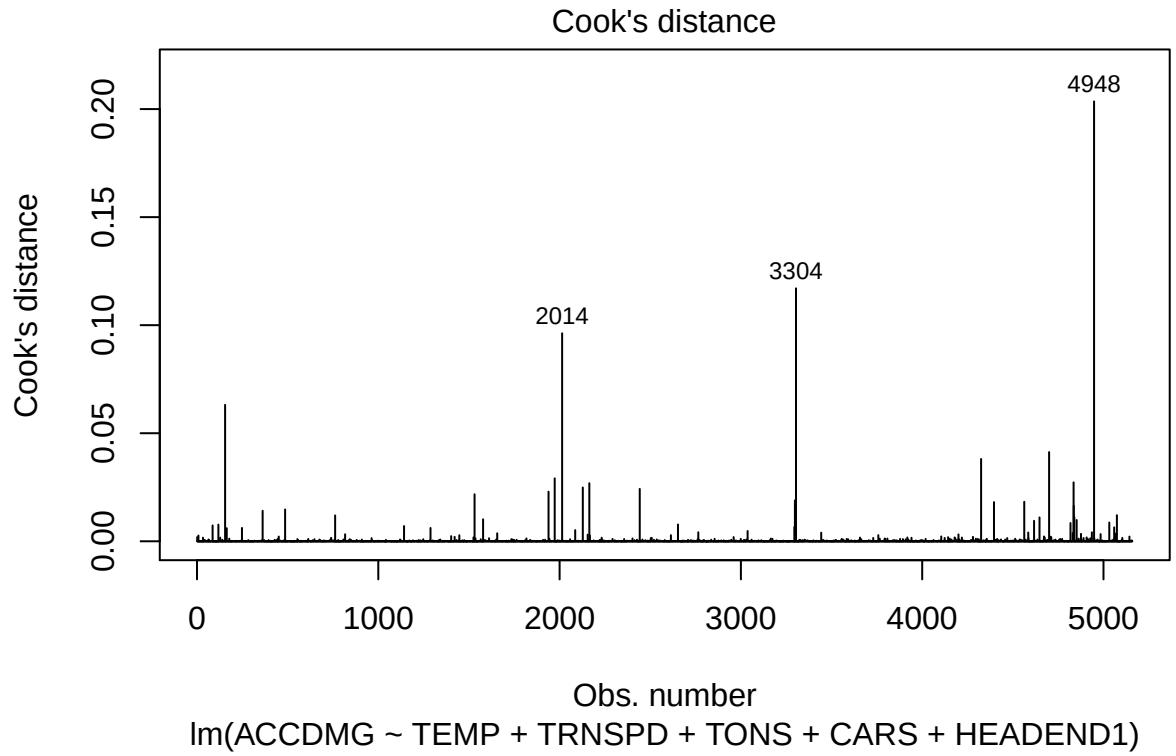


Fig 16: Residuals vs Leverage

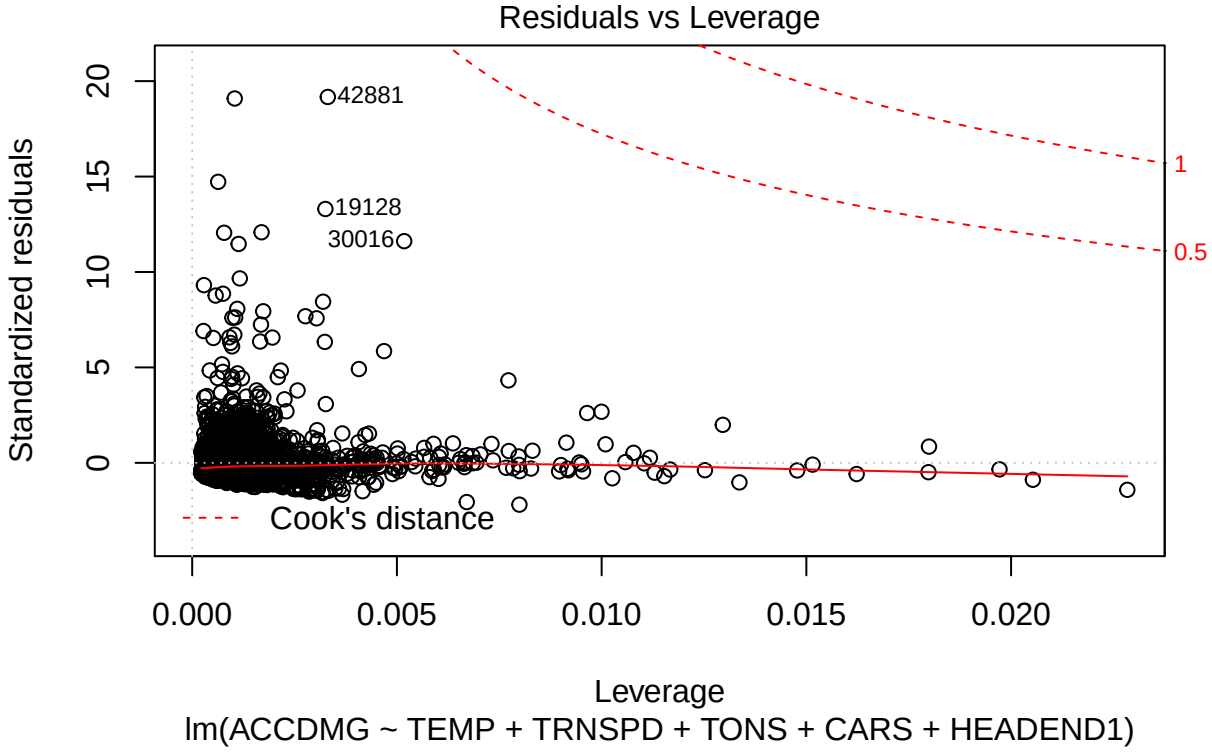
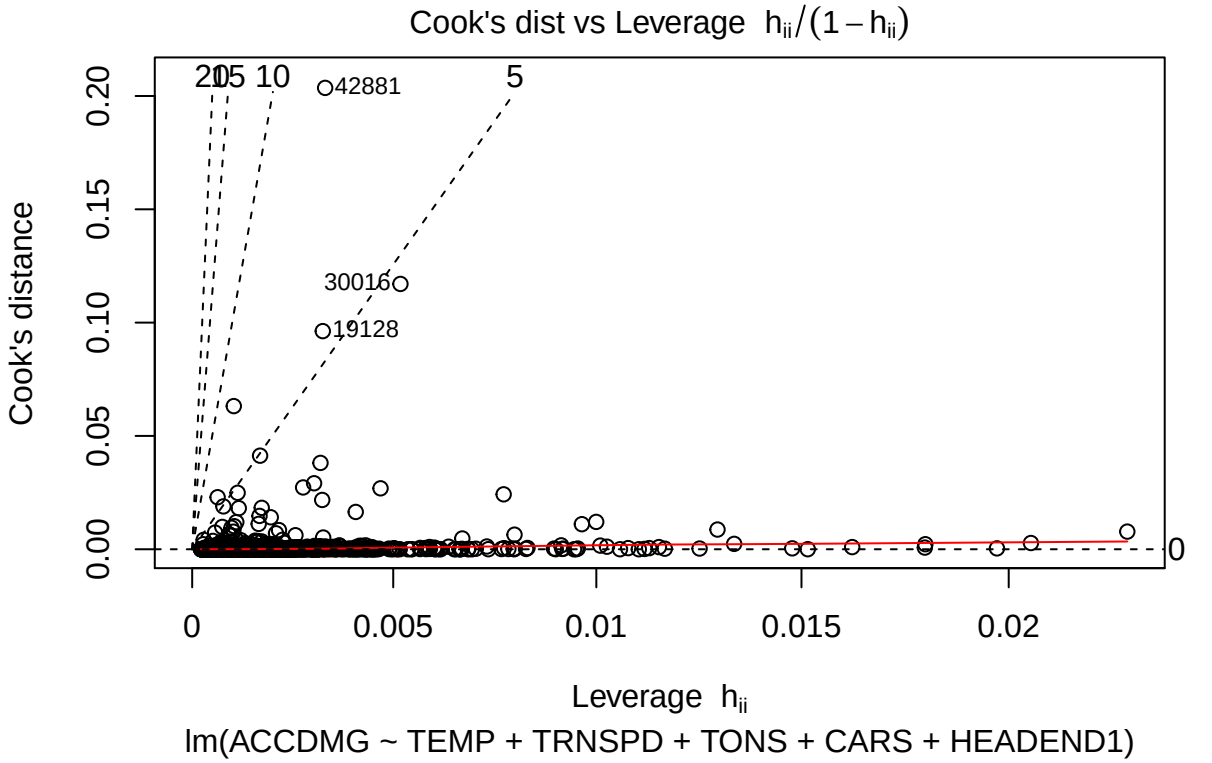


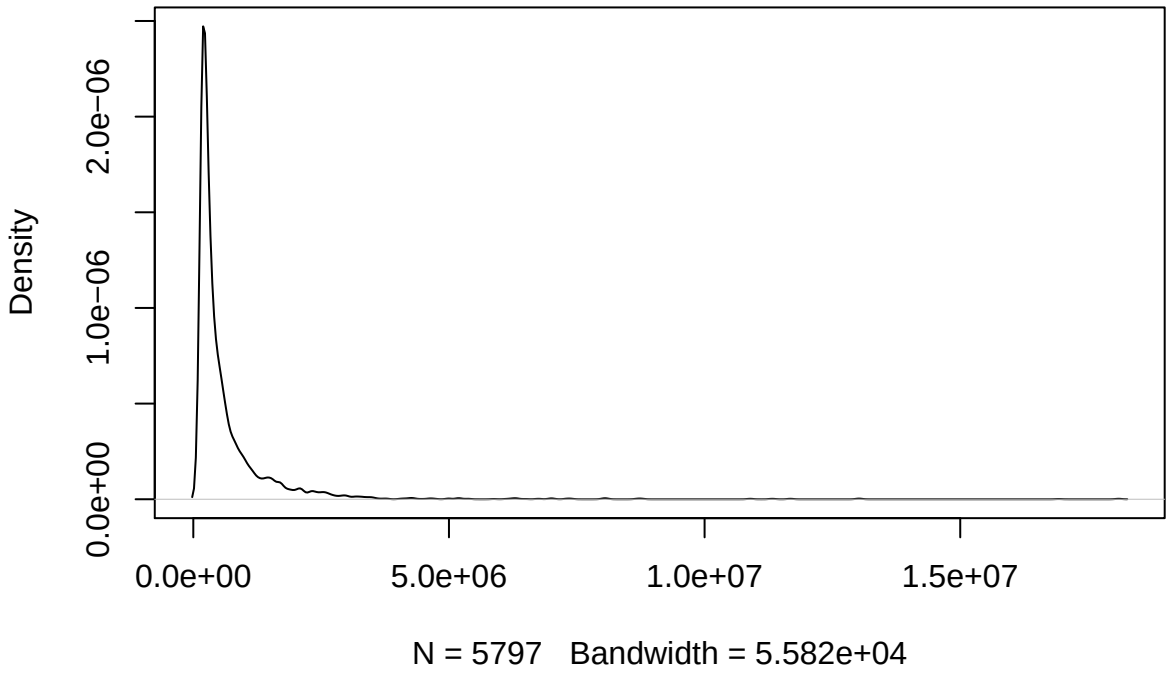
Fig 17: Cook's dist vs. Leverage



Now as it seems there are quite a few points that are influencing the data. We will investigate these points later in detail.

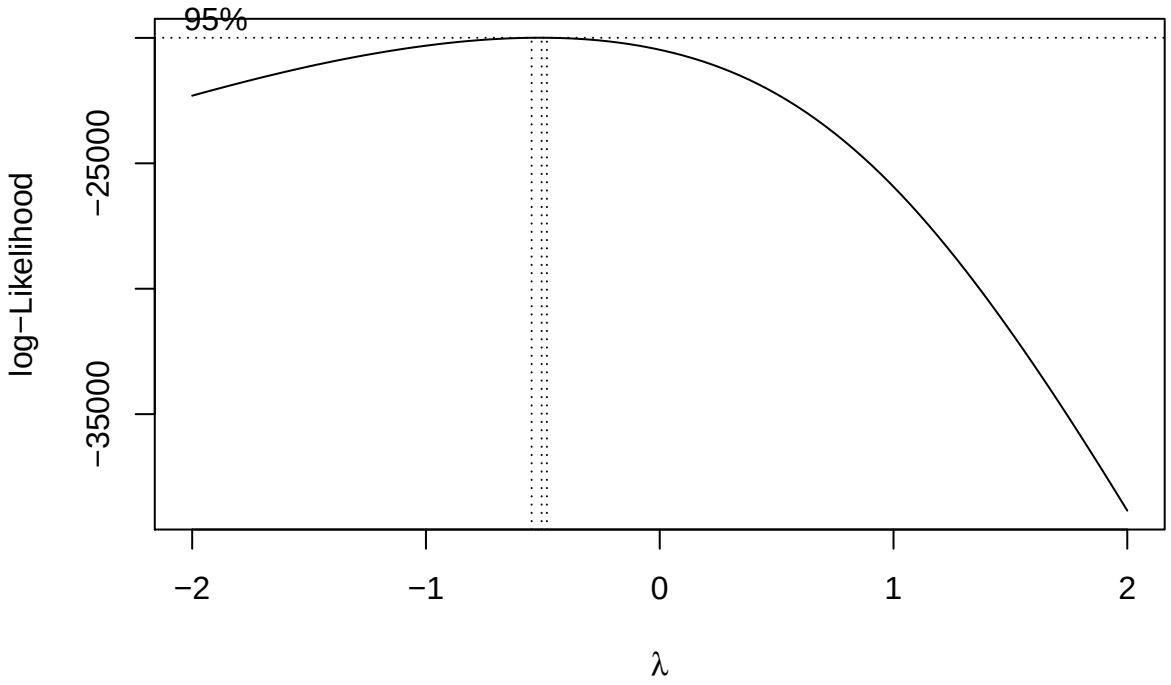
Let's take a look at the response variable ACCDMG.

Fig 18: Density plot of Accident Damage



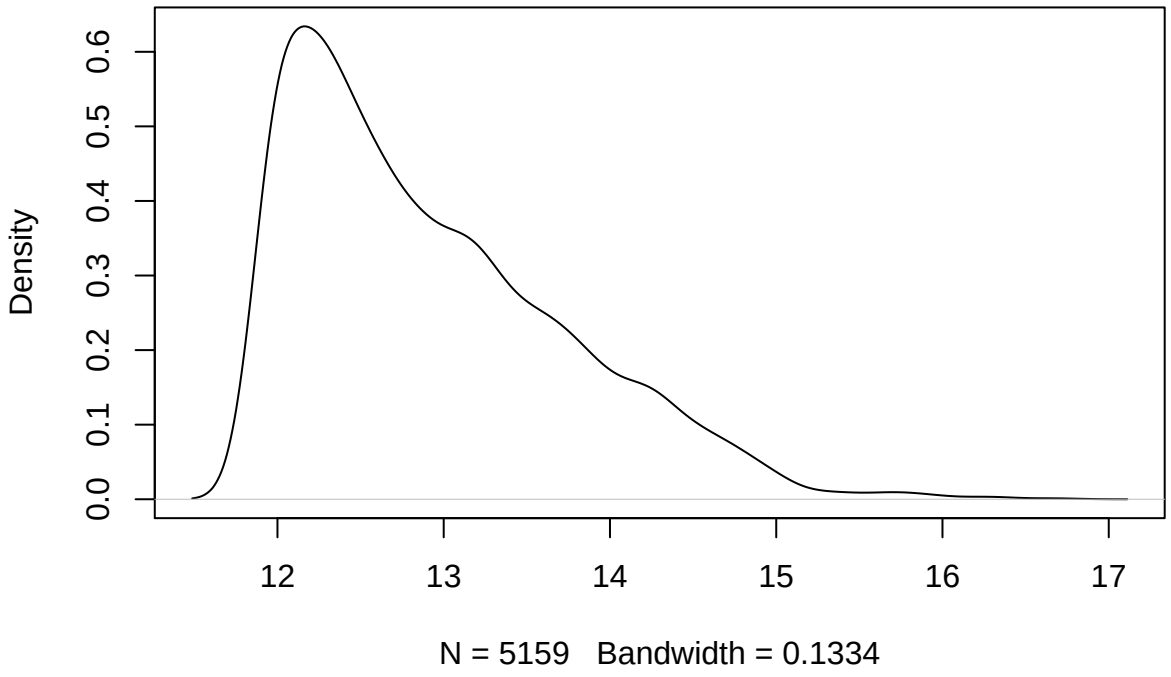
The density plot of the accident damage is not normal. However a lot of statistical techniques perform better when the data distribution is normal. Box-cox transformation is one such transformation that helps in finding a transformation that appropriately normalizes the data.

After the box-cox transformation for the appropriate lambda, we do another fit and we can see that the p-value in this case has improved significantly.



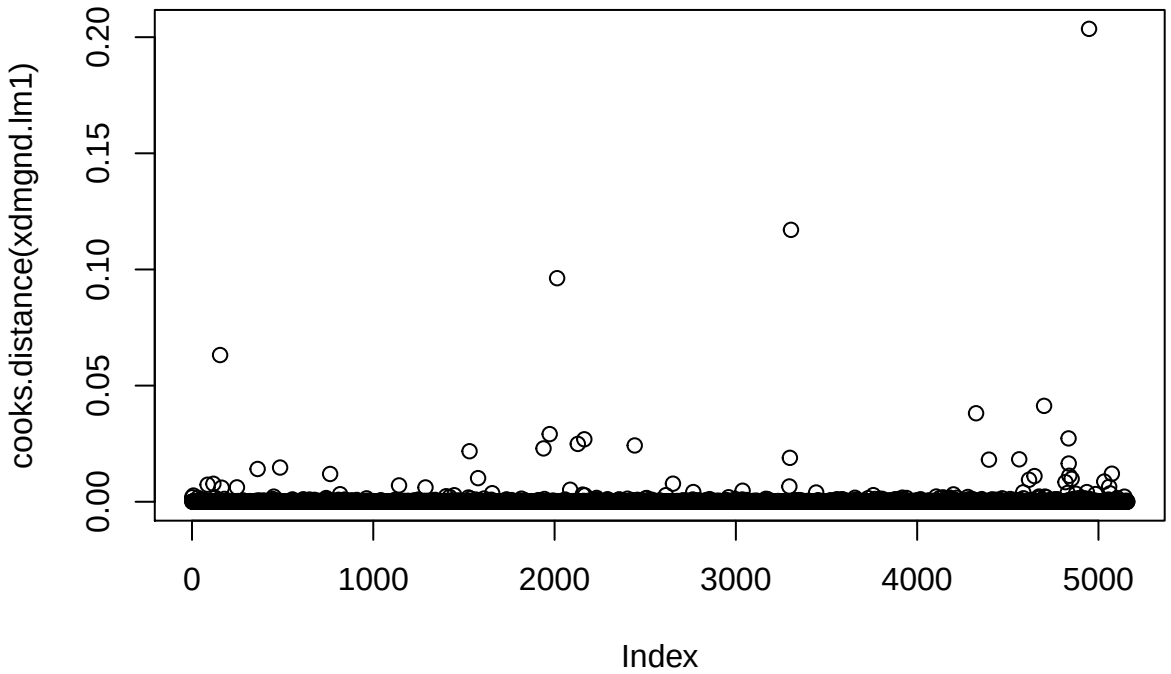
Now, let's try a logarithm transform of the response variable ACCDMG. This looks slightly closer to the normal distribution.

Fig 18: Log of Accident Damage



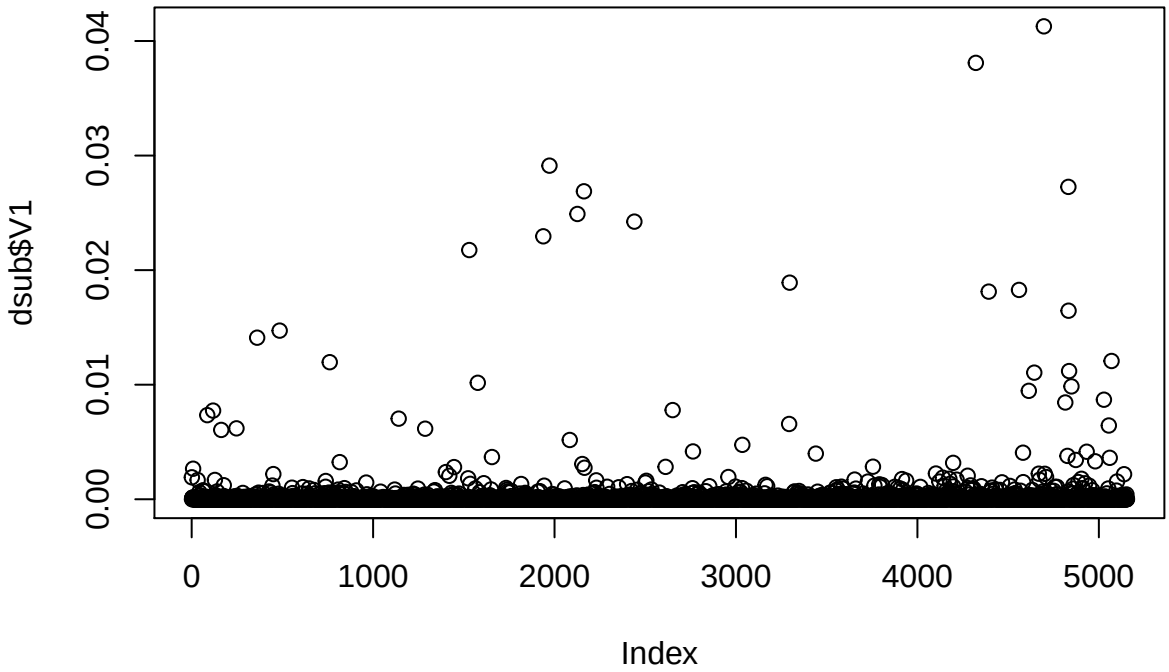
The Cook's Distance gives details about points that have a higher influence in the regression model. To see which points are those, plotting Cook's Distance on the graph we see that there are certain points with high influence on the data set.

Fig 19: Cook's Distance Plot



Eliminating the four points that have a Cook's distance of greater than 0.05, it is shown in Fig 19.

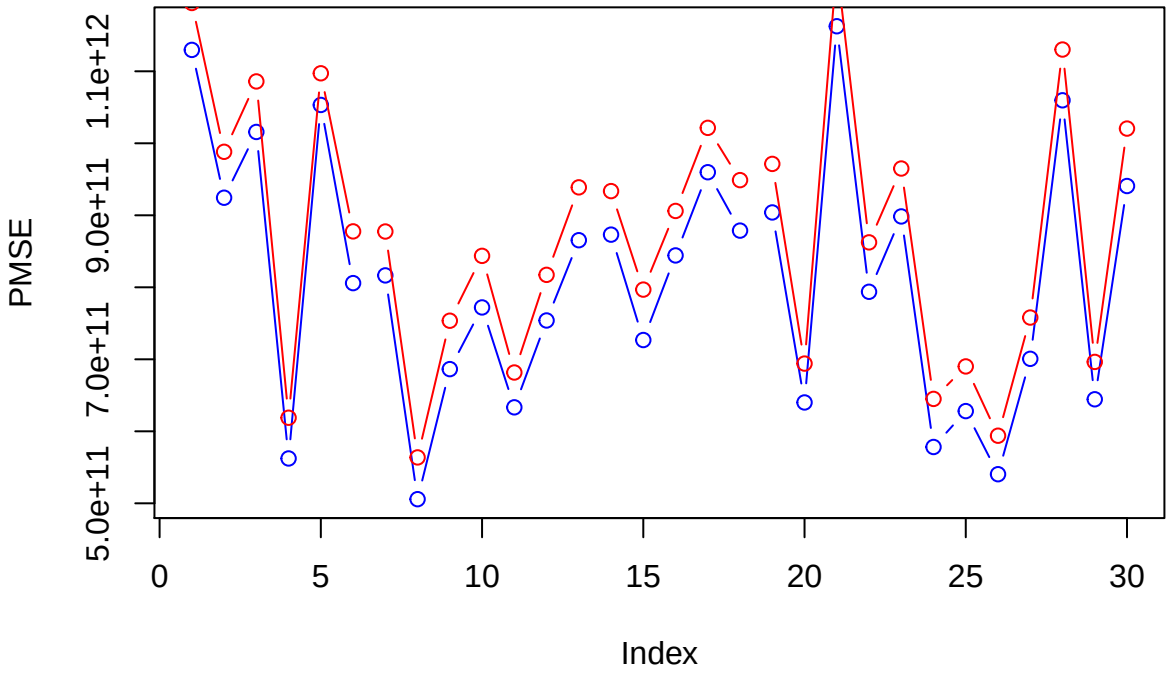
Fig 20: Modified Cook's Distance plot.



The NARRATIVE of the points did not give enough information to figure out if the points with high Cook's Distance should be eliminated. So instead of going ahead with eliminating the points another modern regression technique os being used. This is known as the robust regression. [5, 6]

Now for prediction purposes lets compare the two models and see how they perform on the test set.

Fig 21: Comparison of PMSE on the two new models on test set



And the t-statistics say that model 2, represented in blue has significanly less mean squared error than

model1.

So the final model is using a robust linear regression model, [5, 6] and it downweights some of the properties of data points with high Cook's distance and ensures a more stable model.

Analysis for the second Hypothesis:

Now exploring a categorical variable Derailment over other causes of accidents and seeing the effect of derailment and head-on collisions on the number of people killed.

In this case we want to condition on the number of CARS and the TEMPERATURE so that we can make sure that people killed in the train accidents died only because of the head-on Collison compared against Derailment and not because of more CARS or TEMPERATURE.

The following are the Diagnostic Graphs for the analysis of Total number of people killed in head-on collisions against Derailments.

Fig 22: Residual vs. Fitted

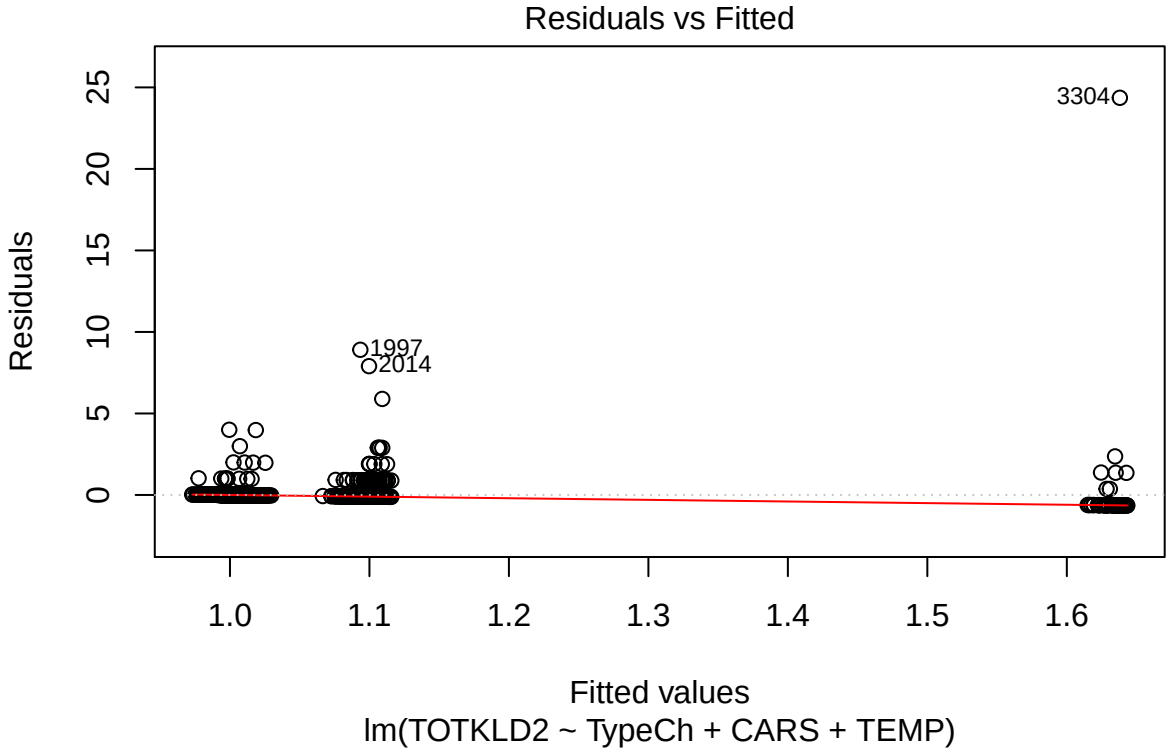


Fig 23: QQ Plot

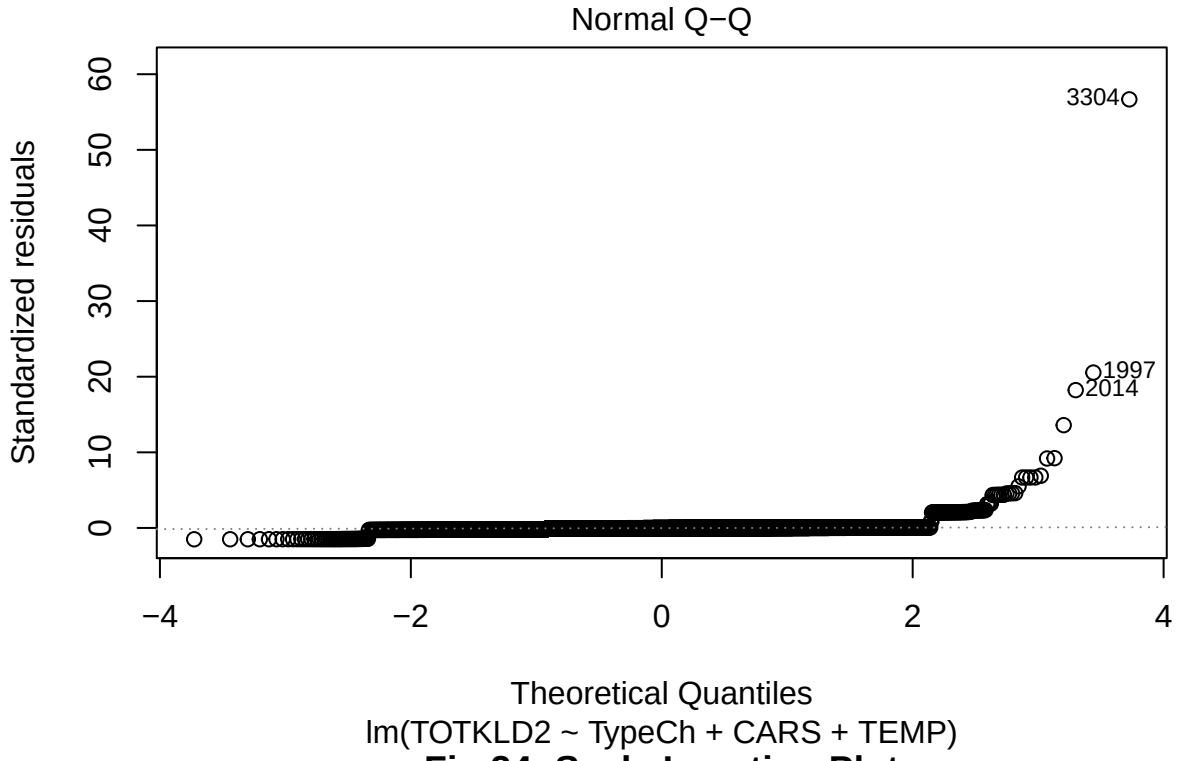


Fig 24: Scale Location Plot

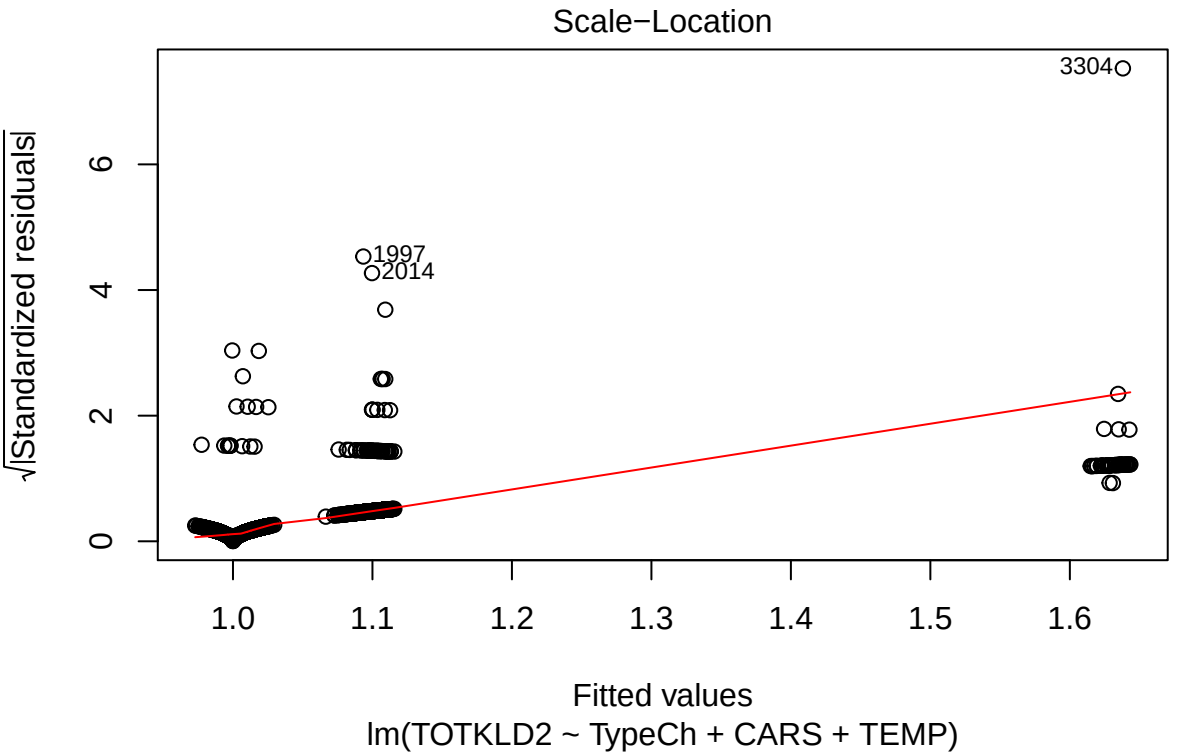


Fig 25: Cook's distance plot

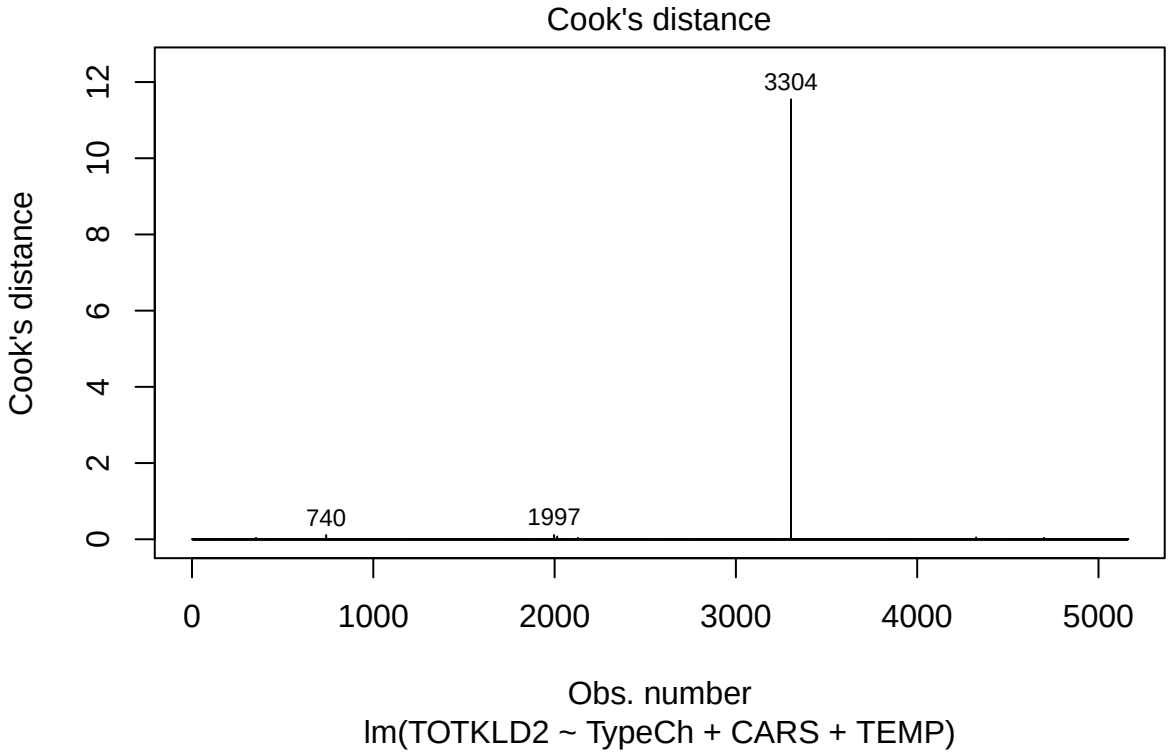


Fig 26: Redisuals vs. Leverage

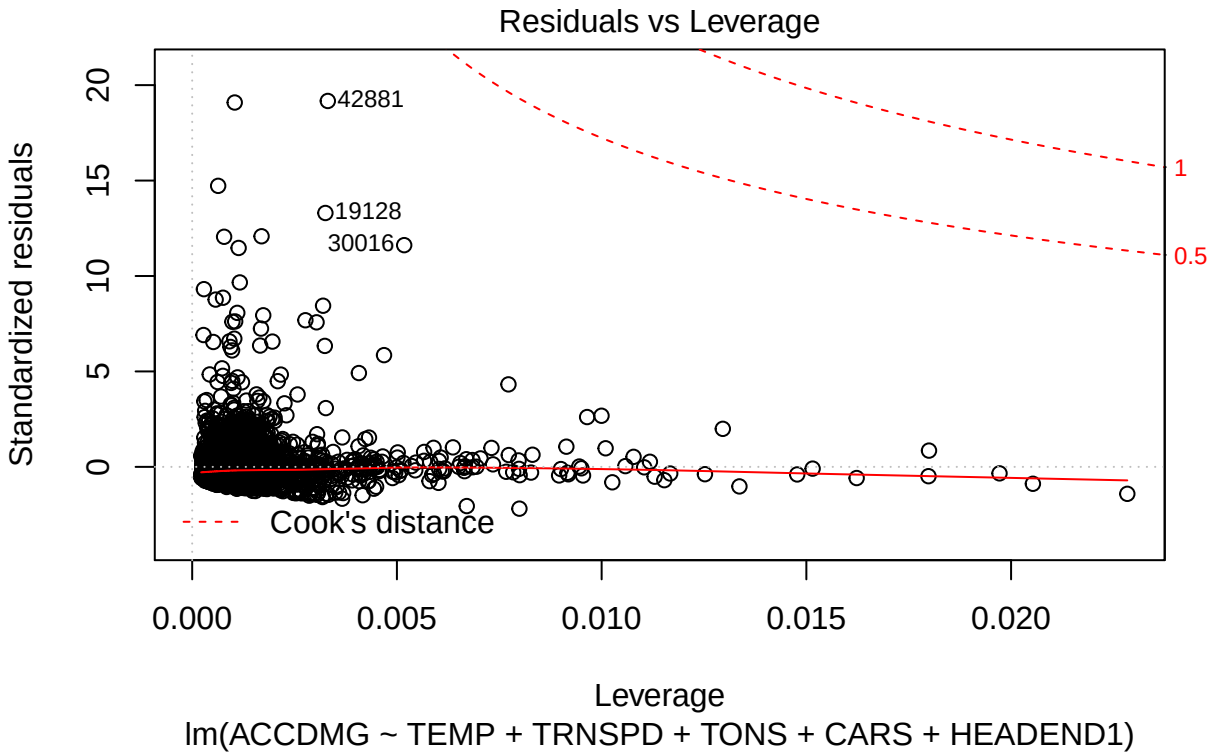
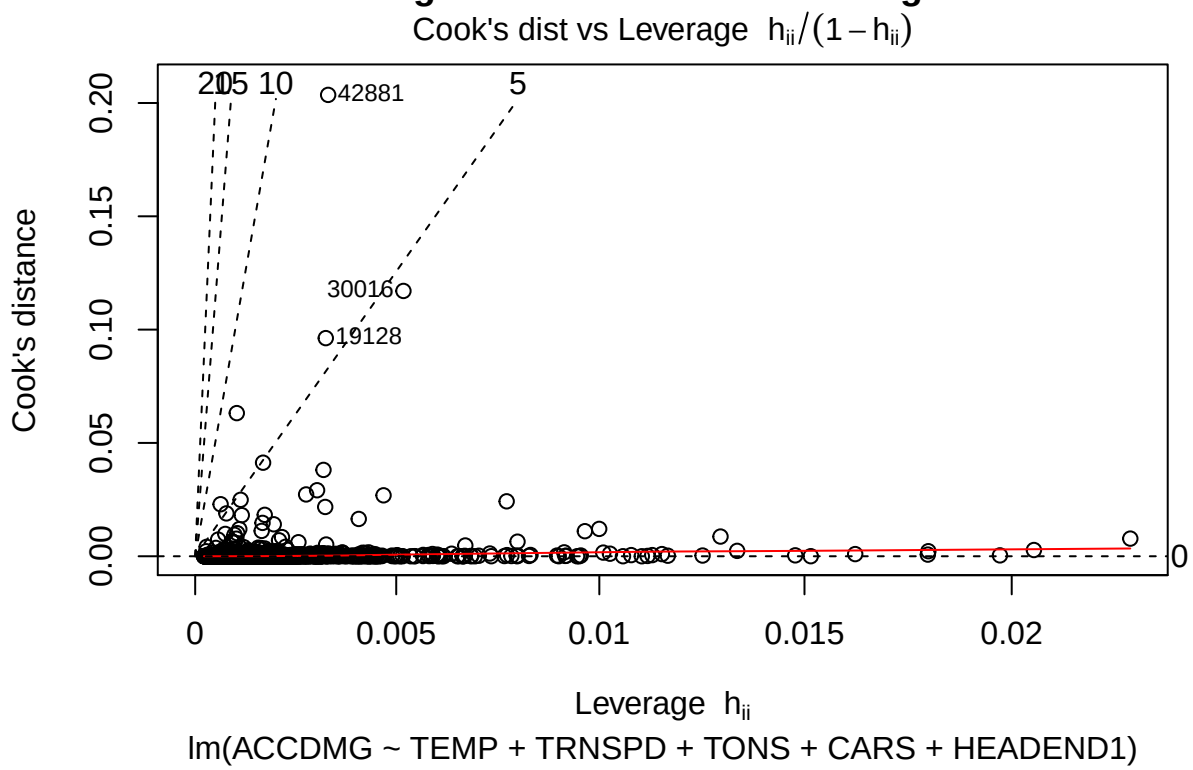


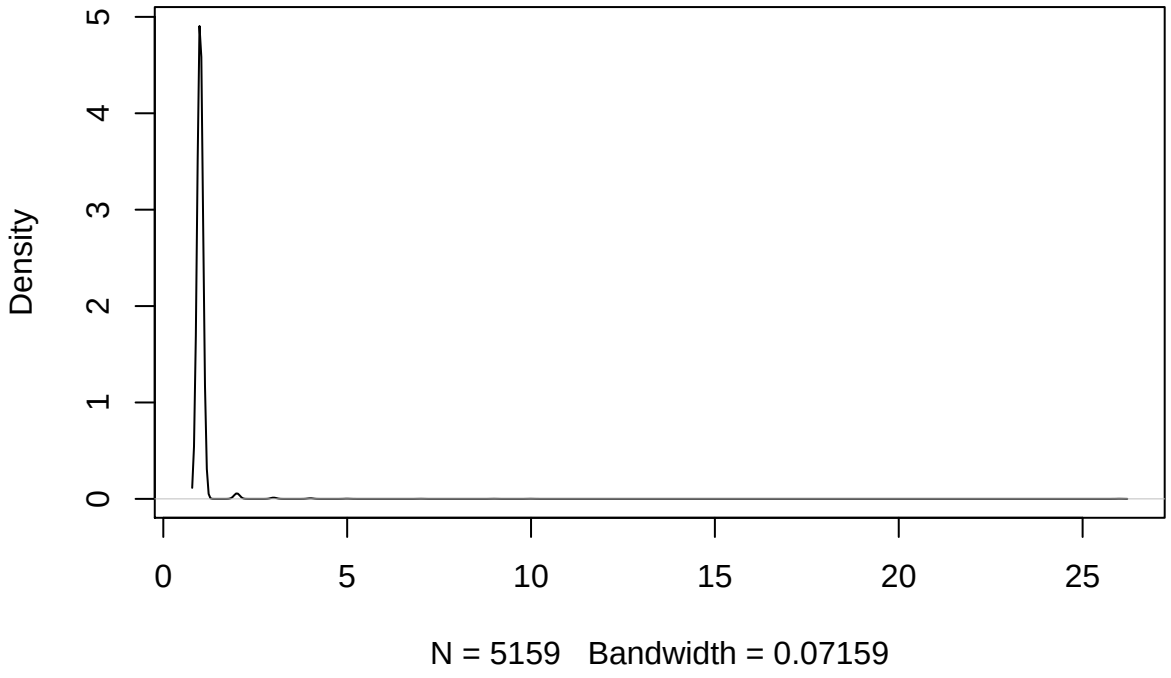
Fig 27: Cook's dist vs. Leverage



Now as it seems there are quite a few points that are influencing the data. We will investigate these points later in detail.

Let's take a look at the response variable TOTKLD.

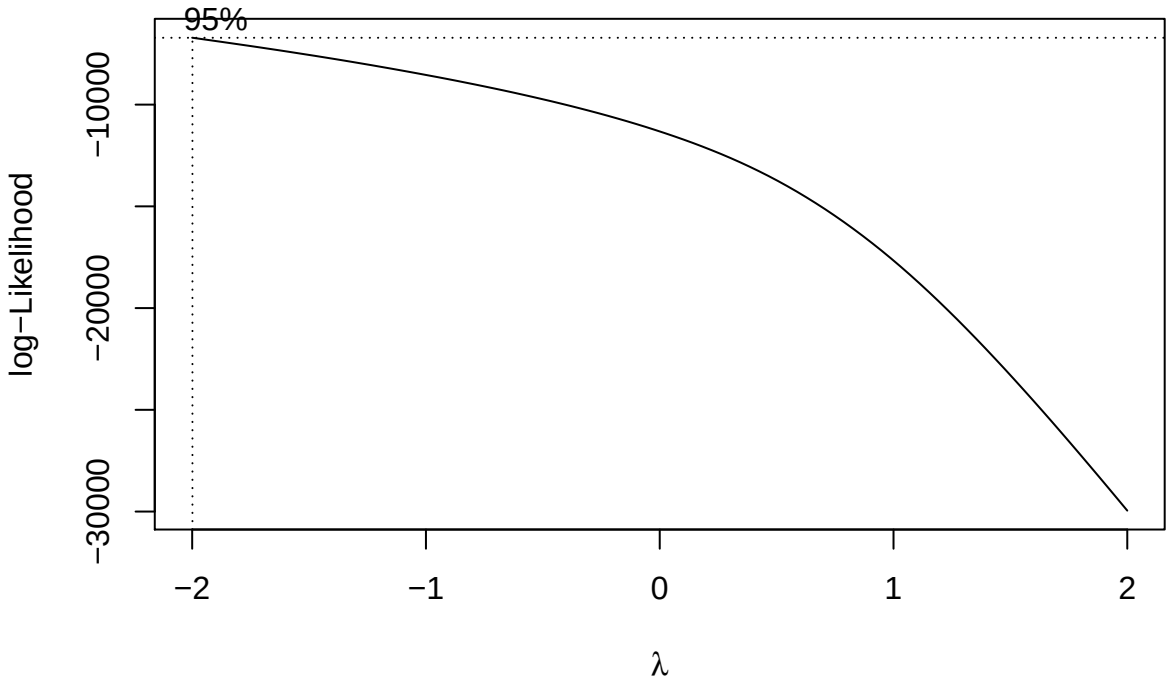
Fig 28: Density plot of Total Number of People Killed



The density plot of the Total number of people killed is not normal and extremely skewed. However a lot of statistical techniques perform better when the data distribution is normal. Box-cox transformation is one such transformation that helps in finding a transformation that appropriately normalizes the data.

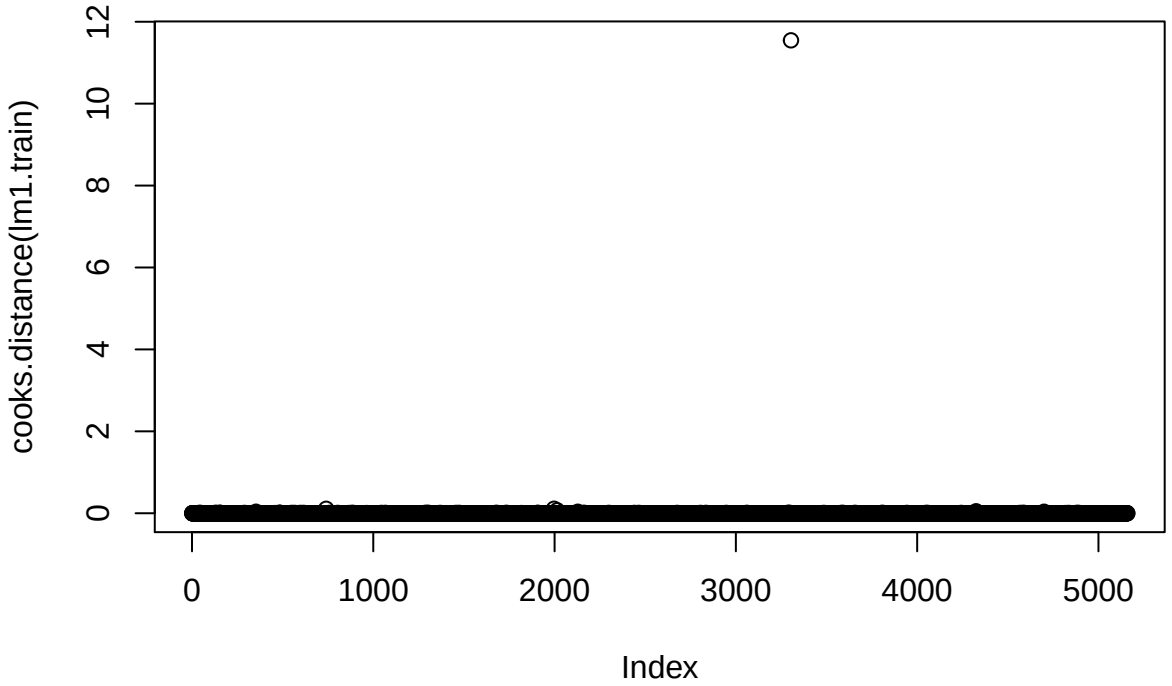
Also in some of the accidents the total number of people killed is zero. So box-cox transformations to those will lead to undefined values. Thus 1 is added uniformly to all the values of TOTKLD so as to take into account even those points.

After the box-cox transformation for the appropriate lambda, we do another fit and we can see that the p-value in this case has improved significantly.



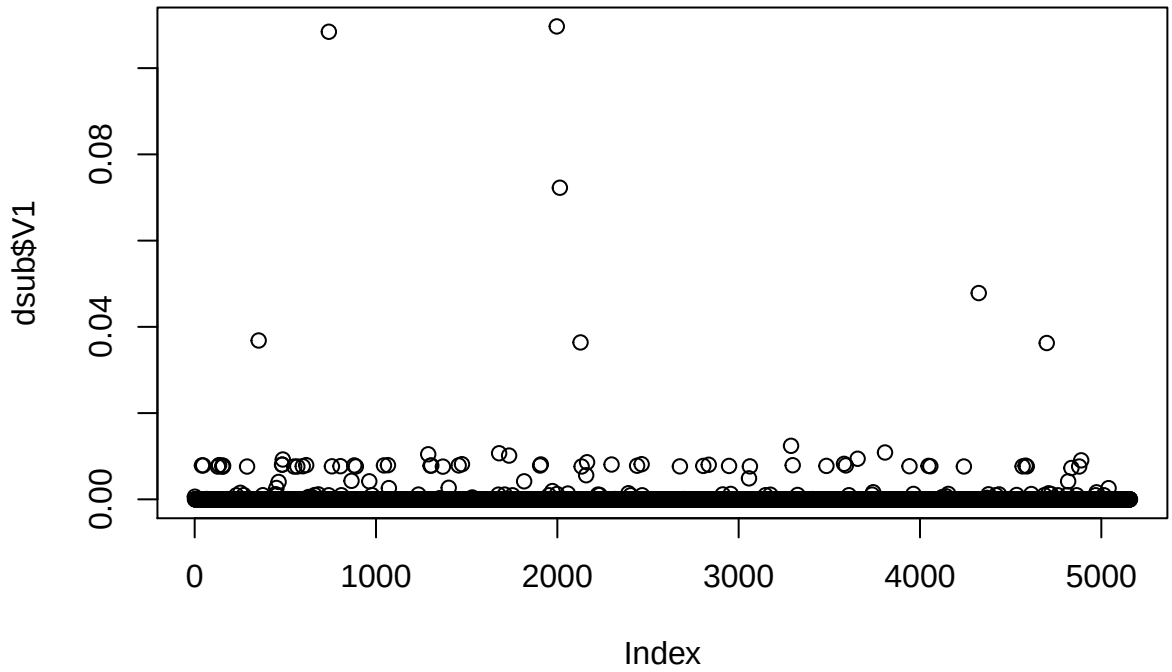
The Cook's Distance gives details about points that have a higher influence in the regression model. To see which points are those, plotting Cook's Distance on the graph we see that there are certain points with high influence on the data set.

Fig 29: Cook's Distance Plot



Eliminating the single points that have a Cook's distance of greater than 10, it is shown in Fig 18.

Fig 30: Modified Cook's Distance plot.



Thus after taking all the aforementioned processes into account, the final model analyzes the Total Number of People Killed against Derailment and Head-On Collisions conditioned on the number of cars and temperature.

This model has a p-value of 0.385 for Derailments and 2e-16 for Head On Collisions suggesting that head on collisions indeed result in more deaths than derailment even though derailments are the chief cause of severe

accidents.

Also the model's p value of 2e-16 is low suggesting that the model is highly significant.

Analysis for the third Hypothesis:

Now exploring a categorical variable Rack, Roadbed and Structure related problems over other causes of accidents and comparison of the effect of Rack, Roadbed and Structures against human factors in terms of the economic damage incurred in the accident.

In this case we want to condition on the number of CARS,TEMPERATURE and TONS so that we can make sure that accident damage in the train accidents is due to the CAUSE of the accident like Road Bed structures compared to the Human Factors in terms of the accident damage.

The following are the Diagnostic Graphs for the analysis of the accident damage due to the Cause of the accident being Rack, Road Bed and Structures against Human Factors.

Fig 31: Residual vs. Fitted

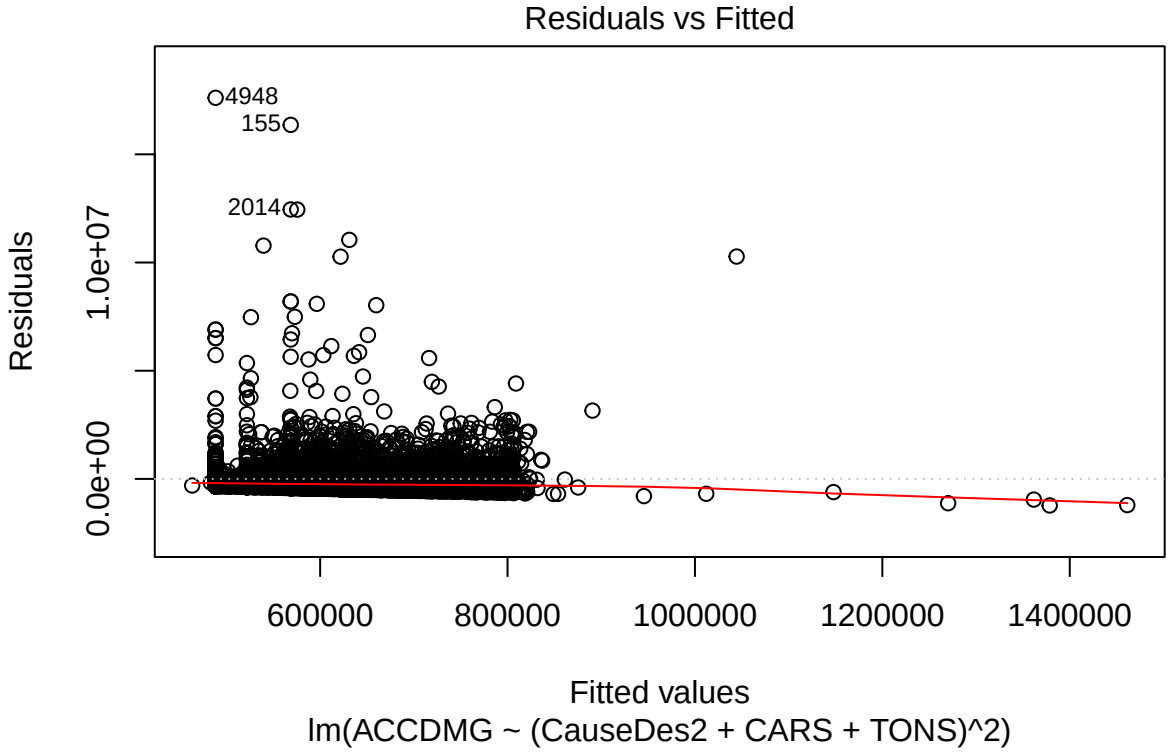


Fig 32: QQ Plot

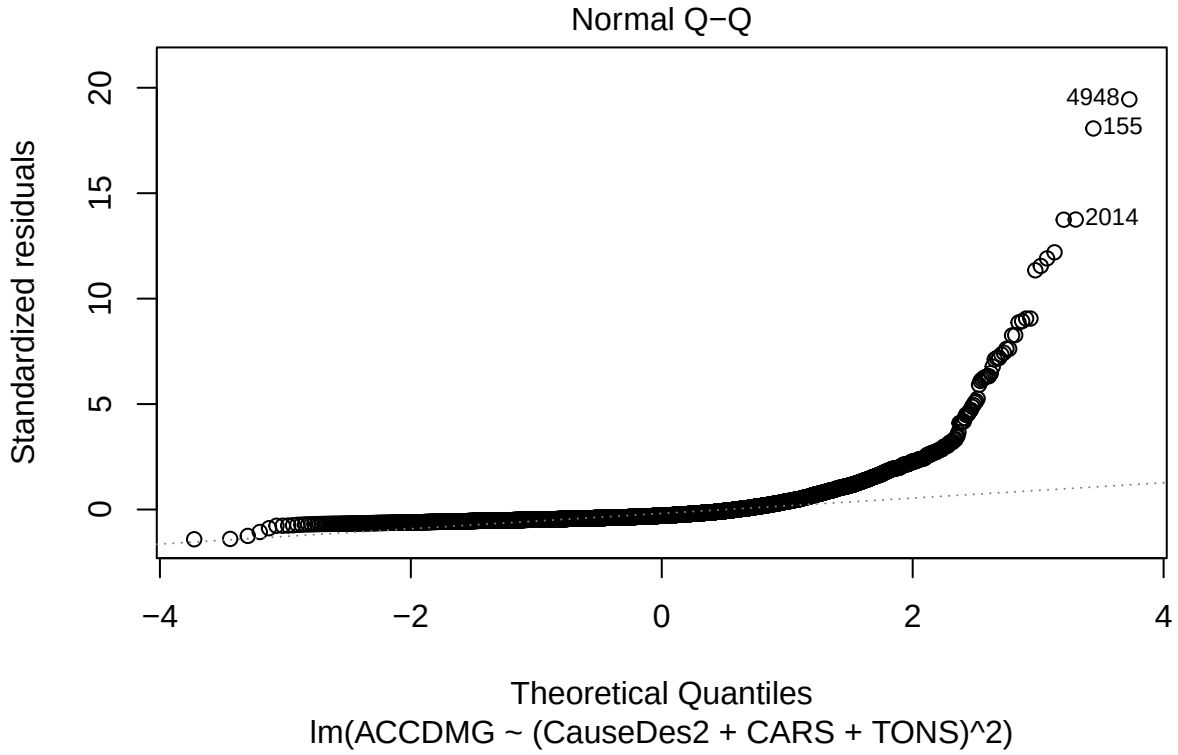


Fig 33: Scale Location Plot

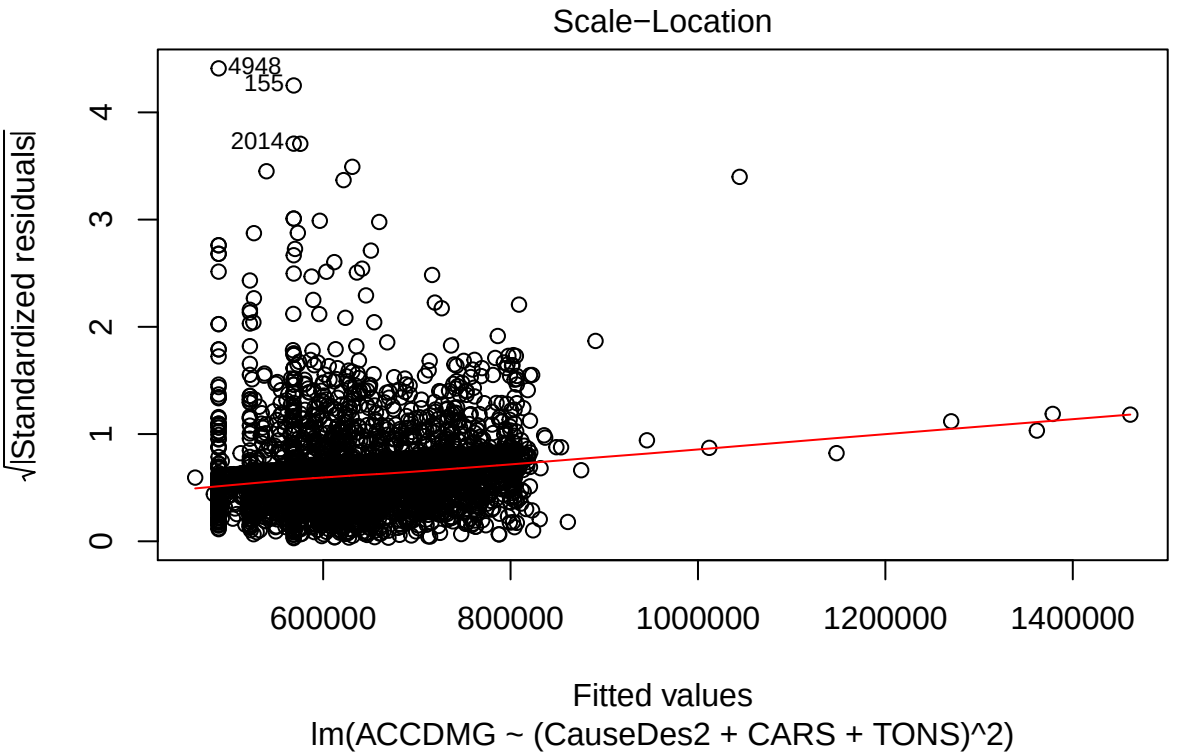


Fig 34: Cook's distance plot

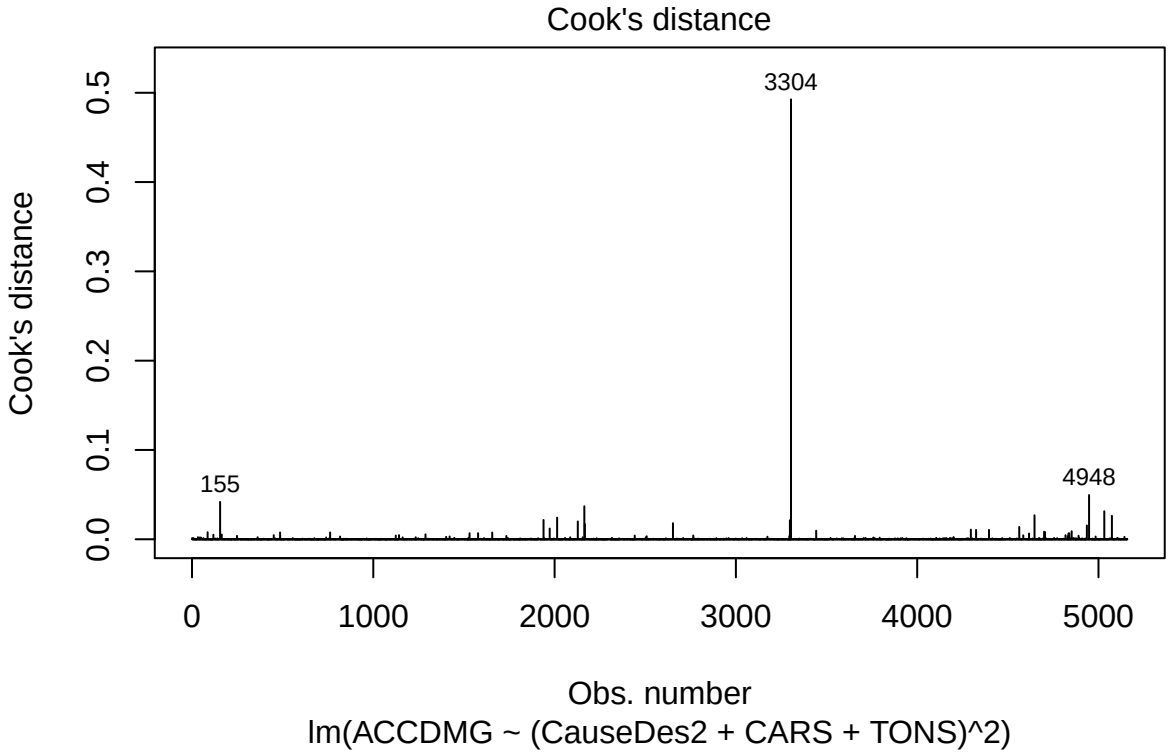


Fig 35: Redisuals vs. Leverage

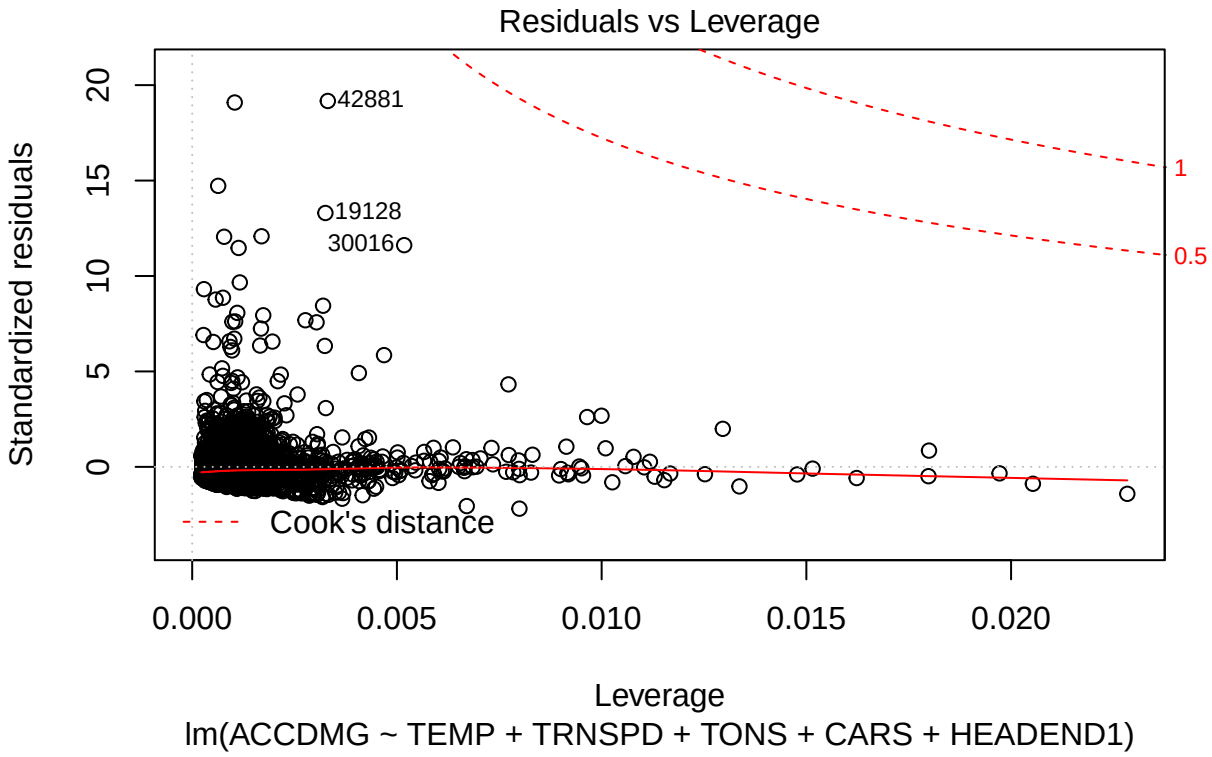
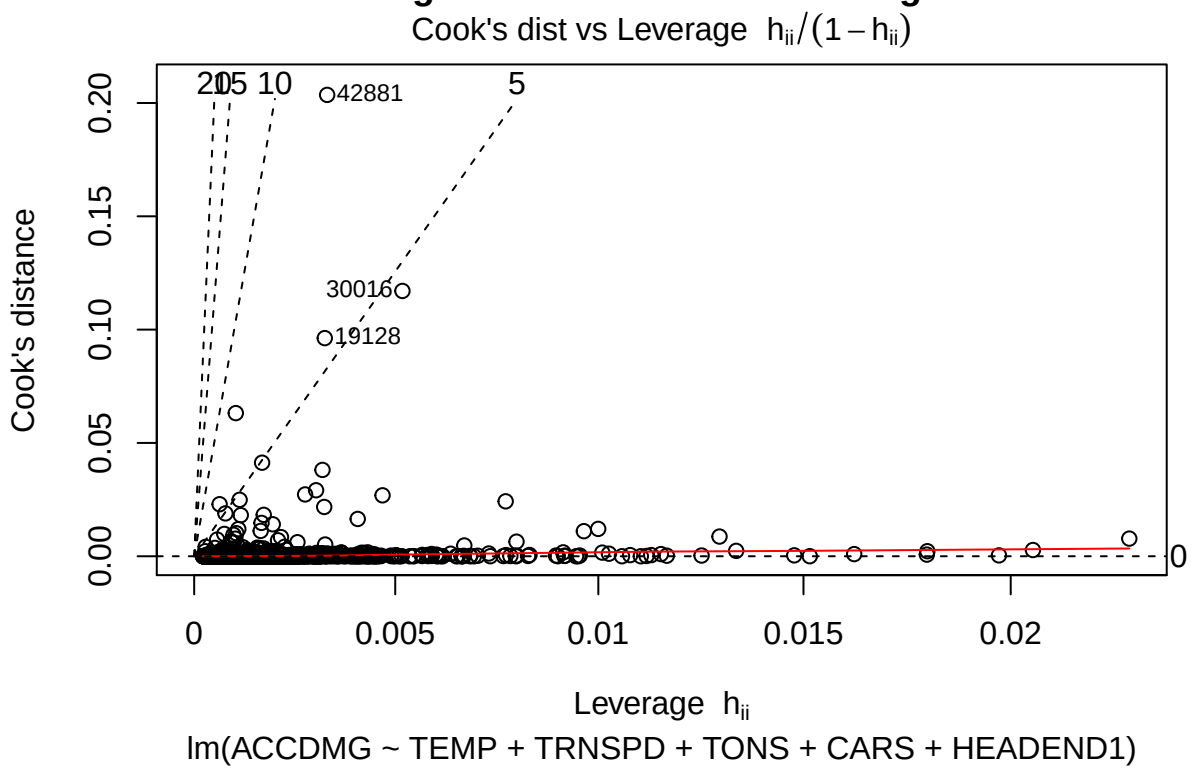


Fig 36: Cook's dist vs. Leverage



Now as it seems there are quite a few points that are influencing the data. We will investigate these points later in detail.

As we had seen before that the variable accident damage in Figure 18, did not follow a normal distribution and we did a box cox transformation for the variable accident damage so as to improve the model.

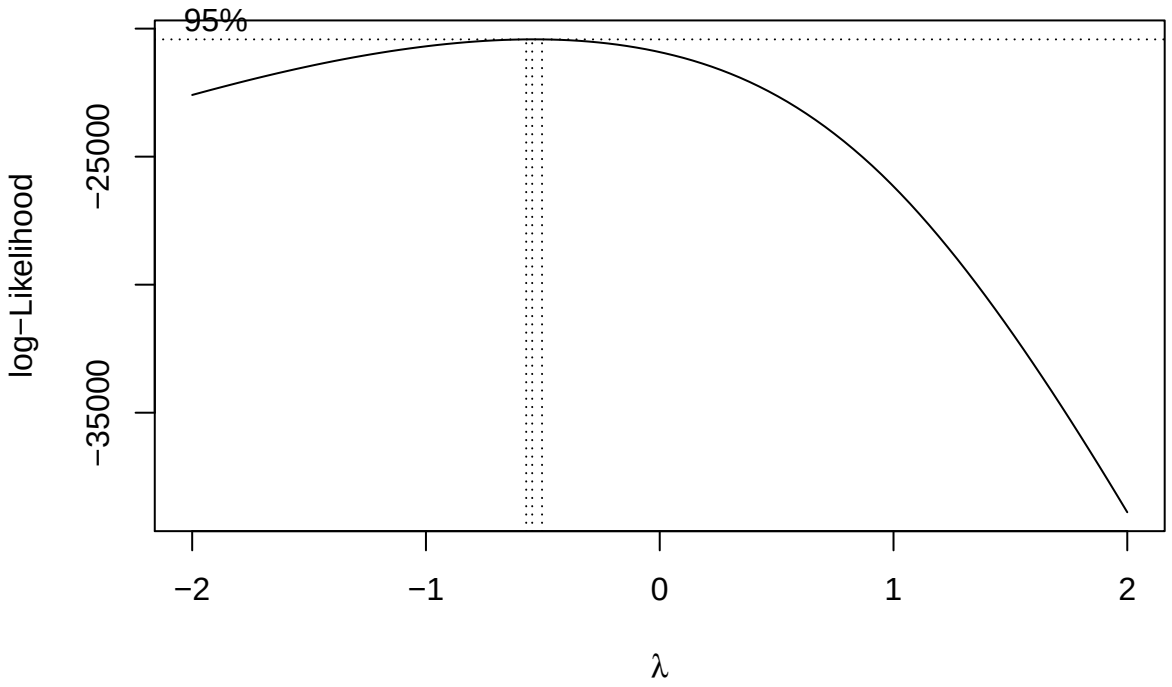
Here the R^2 is 0.08137 and the adjusted R^2 is around 0.07174. Since the adjusted R^2 is very close to R^2 the penalty for having many terms to fit the linear model for accident damage is hardly significant. However the correlation value is itself very small.

There can be various factors that can lead to such a small correlation value and heteroscedasticity is one of them. It is the phenomenon in which some observations are less reliable than others and should be downweighted in a fitting procedure.

The important point to remember is that heteroscedasticity does not cause ordinary least square (OLS) to be biased, but it does make the OLS inefficient. [7]

As we had seen in case of accident damage for the first hypothesis, the density plot of the density plot of the accident damage not normal and extremely skewed. However a lot of statistical techniques perform better when the data distribution is normal. Box-cox transformation is one such transformation that helps in finding a transformation that appropriately normalizes the data.

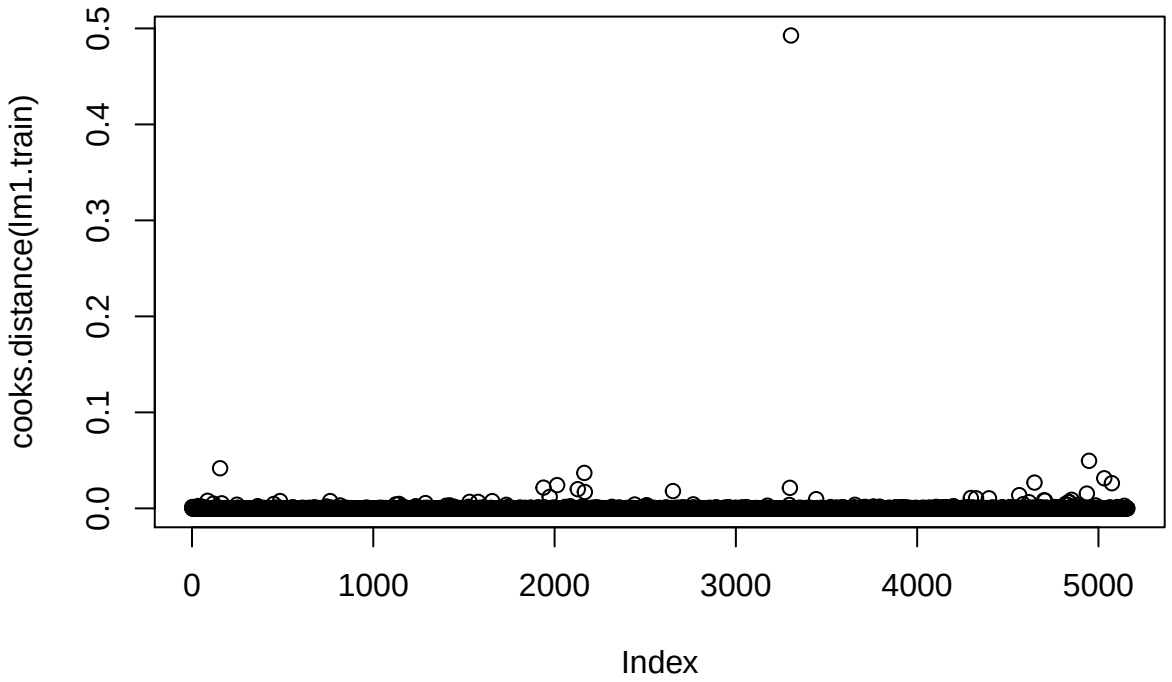
After the box-cox transformation for the appropriate lambda, we do another fit and we can see that the p-value in this case has improved significantly.




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

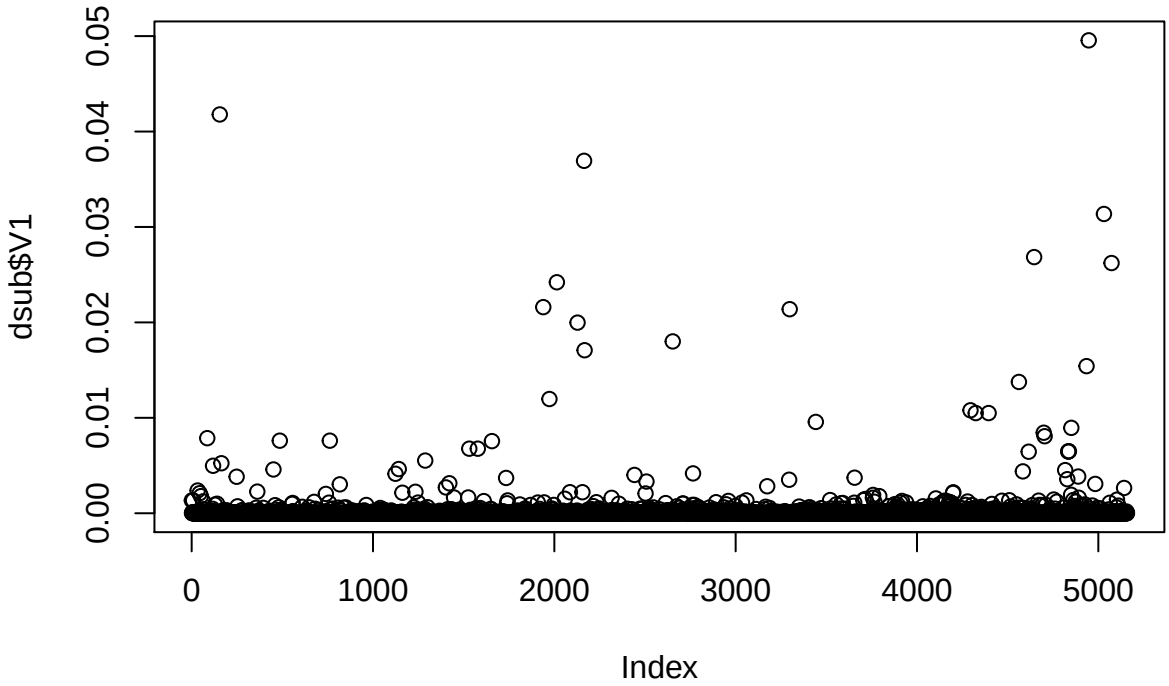
The Cook's Distance gives details about points that have a higher influence in the regression model. To see which points are those, plotting Cook's Distance on the graph we see that there are certain points with high influence on the data set.

Fig 37: Cook's Distance Plot



Eliminating the single points that have a Cook's distance of greater than 0.4, it is shown in Fig 38.

Fig 38: Modified Cook's Distance plot.



Thus after taking all the aforementioned processes into account, the total accident damage based on the cause of the accident, where the comparison is between accidents that are cause by human factors and accidents that are caused by rack, roadbed structures.

This model has a p-value of 0.641 for accidents related to human factors and 2e-16 for accidents due to rack, road bed and structures suggesting that rack, road bed and structures cause more accident damage than human factors in terms of the economic loss.

Also the model's p value of 1.582e-08 is low suggesting that the model is highly significant.

3. Evidence

Hypothesis 1: In the FRA data of Train accidents [3], in the top 5449 accidents in terms of the economic damage, higher train speeds do not lead to higher severity of accident damage.

There is significant amount of evidence to support that higher trainspeed conditional on temperature, the number of tons, the number of cars and the number of headend locomotives show that it does have a significant impact on the accident damage.

The pmse difference in Figure 19 is significant since the blue line, which was a linear regression model built with the conditional variables along with trainspeed performs better on the test set compared to the redline model which is also the accident damage conditional on the aforementioned variables without trainspeed.

This is further validated by the fact that the AIC of the model with Trainspeed is lower than the AIC of the model without trainspeed and thus Trainspeed do play a role in determining the amount of Accident damage conditional on the aforementioned variables.

Since these are Robust linear regresion models [5], these take into account the variations that are introduced by data points with high Cook's Distance and downweights them accordingly, we can say that one of the actionable recommendations is to lower the current speed limit for Trains so as to reduce severity of economic loss through accident damages.

Diagnostic plots are shown in Figures 12,13,14,15,16 and 17 and the and the density plot and the box cox transformation are shown in figure 18. The value of l = -0.5 results in a more normal distribution of the accident damage and that was used to build the model in the analysis section.

It can be concluded with 99% confidence that the trainspeed is positively correlated with the accident damage in sever train accidents and the confidence interval for the 99% confidence is [11626.663, 15639.337].

This model was a robust linear regression model that was developed using step wise regression and taking into account the influential data points that were observed from the residual measures as shown above in the analysis section.

Hypothesis 2: Head-On Collisions of Trains do not kill more people than Derailments in severe Train Accidents, even though Derailments account for majority of the accidents under severe accident damages.

Head on collisions kill more people in train accidents than derailment even though derailment is the most common cause of train accidents in the FRA data of severe train accidents. The 99% confidence interval for the coefficient of head on collisons in measuring the total number of people killed is [0.4739, 0.7721].

The diagnostic graphs were shown in Fig 22, 23, 24, 25, 26 and 27 and the density plot and box cox transformation were shown in Figure 28.

The p-value for the model is 2.2e-16 highlighting the fact that indeed head-on collisions kill more people than derailments and we can thus reject the hypothesis. Also when the p-value is observed, the p-value is 2e-16 for HeadOn collisions against a significant higher p-value of 0.385 for Derailments, proving further that head-on collisions lead to more number of deaths than accident damage. The adjusted R^2 of 0.0271 is quite close to the Multiple R^2 signifying that our model does not have excess predictor variables for predicting the Total Number of People Killed. The AIC of the model with the interaction terms seem to have done better than the one without the interaction terms and also validate the fact that head-on collisions are indeed responsble for more causalties compared to the derailement. The AIC value of the model with the interaction terms happen to be 5975.802 and that of the other model is 6027.613 and it highlights the fact that head-on collisions conditioned on the number of cars and temperature indeed result in more deaths than derailment.

Hypothesis 3: Rack, Roadbed and structure related accidents do not cause more severe accidents than human factors in terms of high economic damage.

The p-value of the model is 1.582e-08 highlingting the fact that accident damage due to rack, roadbed and structures is indeed more significant than human factors in terms of economic damage. The p-value of 2e-16, in case of the Rack,Road bed and structures is indeed significant compared to the p-value of 0.641 for human factors and thus the amount of economic damage due to rack, road bed structures is significant in comparison to Human Factors and other causes. The AIC of the model with the interaction terms are more significant than the one without it. The value of the AIC with the interaction terms in 156176 compared to the AIC of 158613.

The diagnostic plots for this hypothesis are in Figures 31-36

Track maintenance or prevention of problems arrising due to rack, road bed structures can lead to significant reduction in the economic loss through accident damage. An actionable step in this case that need to be taken is more regular maintenance of tracks to prevent accidents caused due to rack, road bed and structures.

4. Recommendations

As was observed in the situation that derailments were one of the major types of train accidents and the rack, road bed structures was one of the major types of accidents. It can be concluded with 99% confidence that the trainspeed is positively correlated with the accident damage in sever train accidents and the confidence interval for the 99% confidence is [11626.663, 15639.337]. Thus FRA should take steps to reduce the avergae trainspeed so as to reduce economic damage caused due to train accidents. Adjusted R^2 measurements and AIC values have been mentioned in the preceding section.

Head on collisions kill more people in train accidents than derailment even though derailment is the most common cause of train accidents in the FRA data of severe train accidents. The 99% confidence interval for the coefficient of head on collisons in measuring the total number of people killed is [0.4739, 0.7721]. Some of the possible steps in this case would be more efficient tracks management and efficient communication between trains so as to prevent head on collsions. This can also include extremely slow speeds when trains are in the near vicinity of each other and during bad weather conditions when visibility is low so as to prevent head on collisions.

Also the third observation is that rack, road bed structures cause more severe accident damage than human factors. So if the FRA's regulations related to trainspeed should be lowered conditioned on the number of cars and the tons of the train. More steps need to be taken to prevent head on collisions since that causes more casualties, even though derailment happens to be the major cause of train accidents. Track maintenance or prevention of problems arrising due to rack, road bed structures can lead to significant reduction in the economic loss through accident damage.

An actionable step in this case would be to have more regular track maintenance and increased track maintenance during bad weather conditions. Newer equipment of track failure detection automatically can also be installed to detect rack and road bed structure problems so as to prevent the severe economic loss because of accident damage due to this cause.

References

[1] L. E. Barnes and D. E. Brown, *Project 1: Train accidents,Class project in SYS 6021, 2014.*

[2] Project 1 template, Class template in SYS 4021, 2014.

[3] F. R. Administration. (2014) Federal railroad administration office of safety analysis. *http://safetydata.fra.dot.gov/*

[4] Code sample provided in Class, SYS 6021, L. E. Barnes and D. E. Brown, 2014, UVa

[5] Li, G. 1985. Robust regression. *In Exploring Data Tables, Trends, and Shapes, ed. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, Wiley.*

[6] John Fox, Applied regression analysis, linear models, and related models, Sage publications, Inc, 1997

[7] Lecture20partoftheStatistics511course,lecture20. *http://sites.stat.psu.edu/jls/stat511/lectures/lec20.pdf*

[8] @Book, author = Hadley Wickham, title = ggplot2: elegant graphics for data analysis, publisher = Springer New York, year = 2009, isbn = 978-0-387-98140-6, url = http://had.co.nz/ggplot2/book,

Optional Appendices