Project 1

Shayne Cassidy (sc5za) Sarah Winston Nathan (swn2bf) Sarah Nelson (skn5mq)

10/23/2019

Load data and combine data from 2001-2018 into totacts.

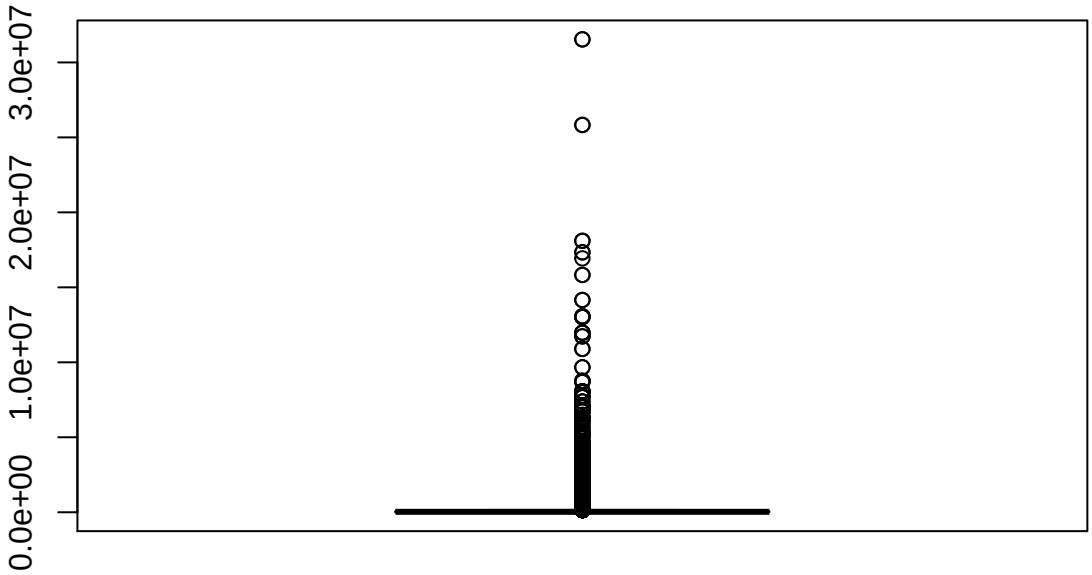
```
# Source AccidentInput
#setwd("/Users/shaynecassidy/Desktop/4021/RCode")
source("AccidentInput.R")

# list of data frames for each year of accident data
acts <- file.inputl(traindir)

# data frame with all accidents from all years from 2001 - 2018

# Get a common set the variables
comvar <- intersect(colnames(acts[[1]]), colnames(acts[[8]]))

# the combined data frame
totacts <- combine.data(acts, comvar)
```

Create a new variable, Casualty, where Casualty = TOTINJ + TOTKLD

```
totacts$Casualty <- totacts$TOTKLD + totacts$TOTINJ
```

Build a data frame with only extreme accidents for ACCDMG

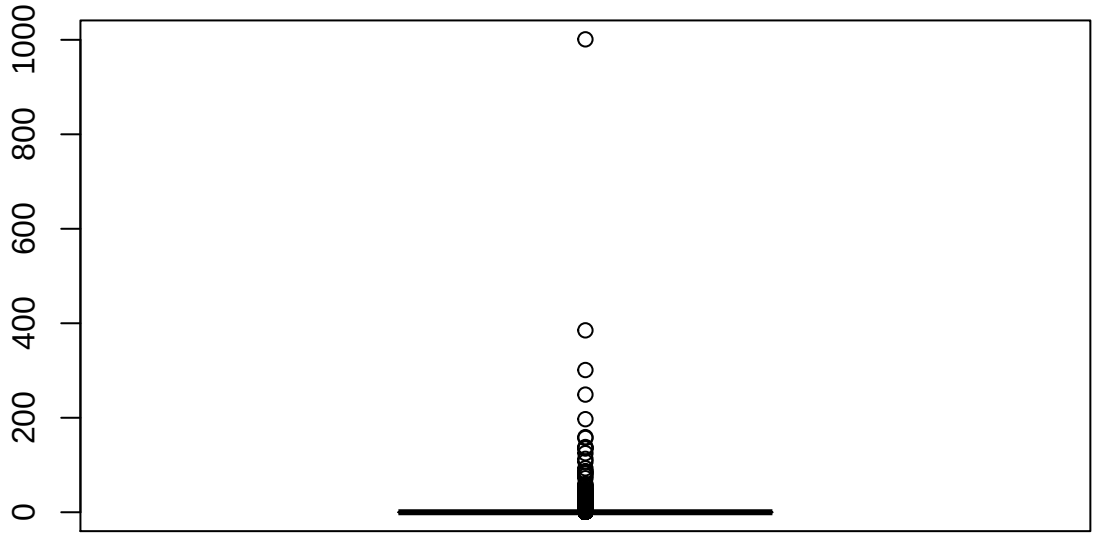
```
dmgbox <- boxplot(totacts$ACCDMG)
```




```
xdmg <- totacts[totacts$ACCDMG > dmgbox$stats[5],]
# Remove duplicates from xdmg and call new data frame xdmgnd
xdmgnd <- xdmg[!(duplicated(xdmg[, c("INCDTNO", "YEAR", "MONTH", "DAY", "TIMEHR",
                                     "TIMEMIN")])),]
```

Build a data frame with only extreme accidents for Casualty

```
casbox <- boxplot(totacts$Casualty)
```




```
cas <- totacts[totacts$Casualty > casbox$stats[5],]
# Remove duplicates from cdmg and call new data frame casnd
casnd <- cas[!(duplicated(cas[, c("INCDTNO", "YEAR", "MONTH", "DAY", "TIMEHR",
                                  "TIMEMIN")])),]
```

Treatment of categorical variables

These variables will need to be tested in our hypotheses. So, in order to appropriately test the hypotheses relating to these categorical variables, these categorical variables are recoded into two levels each.

```
# Create a freight train variable
Freight <- rep(0, nrow(xdmgnd))
Freight[which(xdmgnd$TYPEQ == 1)] <- 1
Freight <- as.factor(Freight)

# Create a human factors variable
HF <- rep(0, nrow(xdmgnd))
HF[which(substr(xdmgnd$CAUSE, 1, 1) == "H")] <- 1
HF <- as.factor(HF)
```

```
# Create a derailment variable
Derail <- rep(0, nrow(casnd))
Derail[which(casnd$TYPE == 1)] <- 1
Derail <- as.factor(Derail)

# Create a highway-rail crossing variable
```

```
HRX <- rep(0,nrow(casnd))
HRX[which(casnd$TYPE == 7)] <- 1
HRX <- as.factor(HRX)
```

Part 1: Generating Hypotheses

ACCDMG: We used the following visualizations, summary statistics, and external supporting evidence to arrive at our hypotheses for ACCDMG.

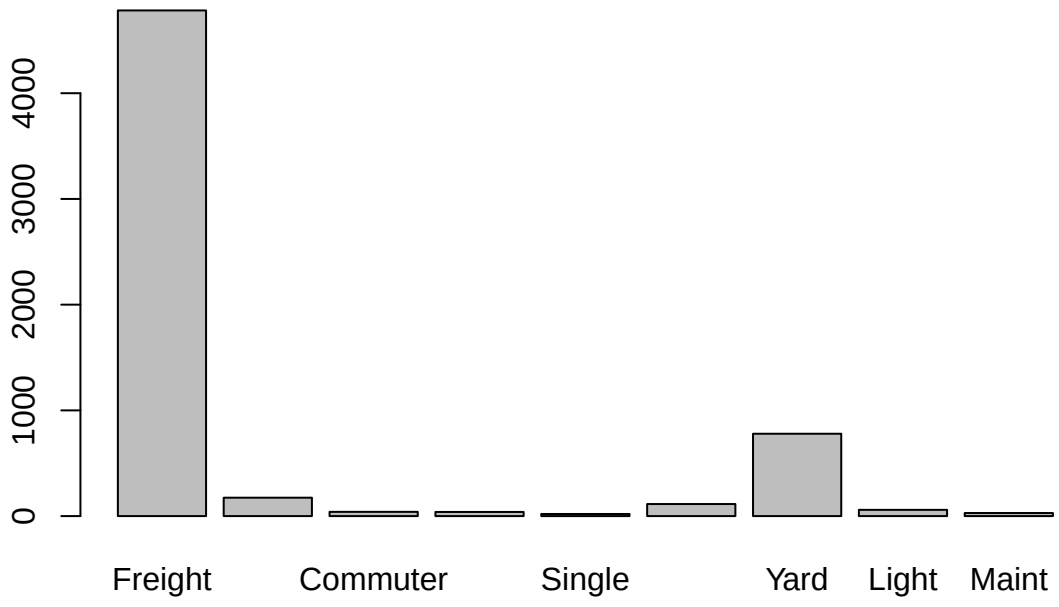
```
# Freight, Human Factors
library(lattice)
mean(xdmgnd$ACCDMG)
```

```
## [1] 689258.8
```

```
max(xdmgnd$ACCDMG)
```

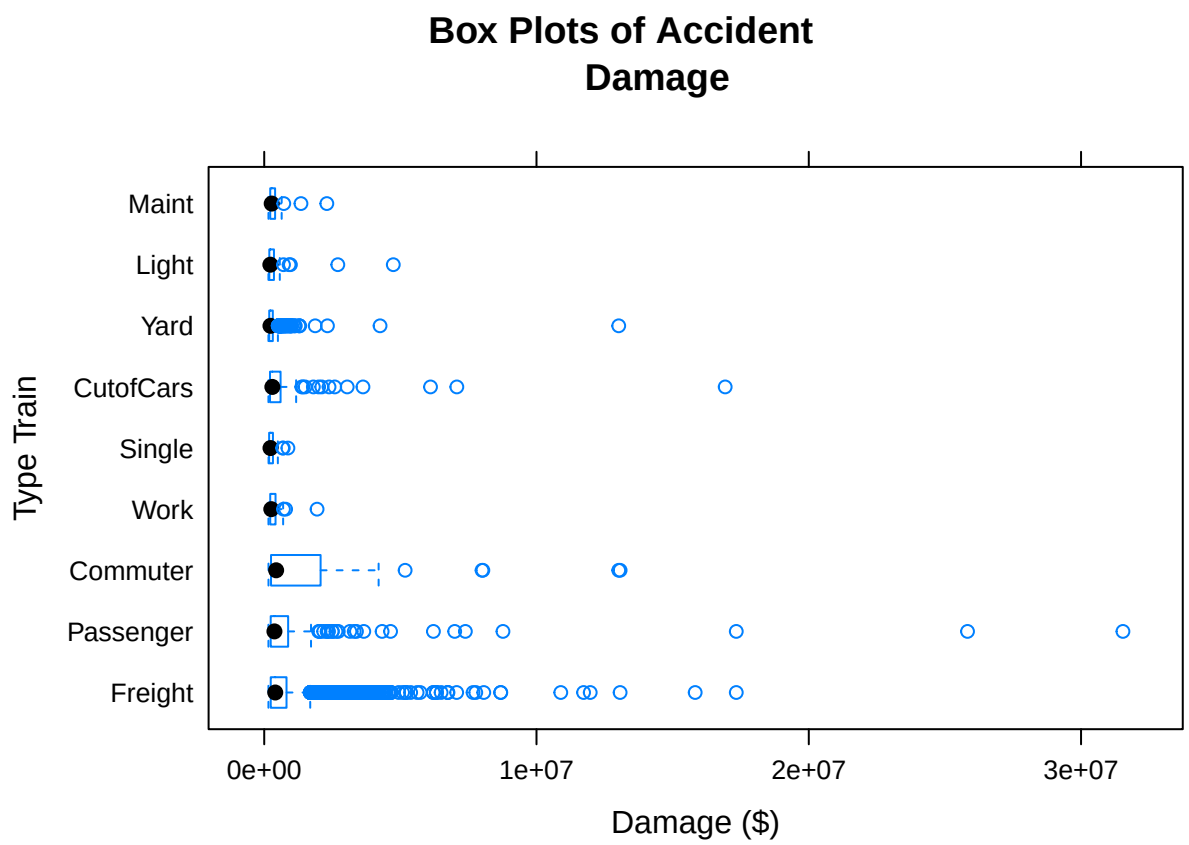
```
## [1] 31538754
```

```
xdmgnd$TYPEQ <- as.numeric(xdmgnd$TYPEQ)
xdmgnd$TYPEQ <- factor(xdmgnd$TYPEQ, labels = c("Freight", "Passenger", "Commuter",
                                                "Work",  "Single", "CutofCars",
                                                "Yard", "Light", "Maint"))
barplot(table(xdmgnd$TYPEQ))
```



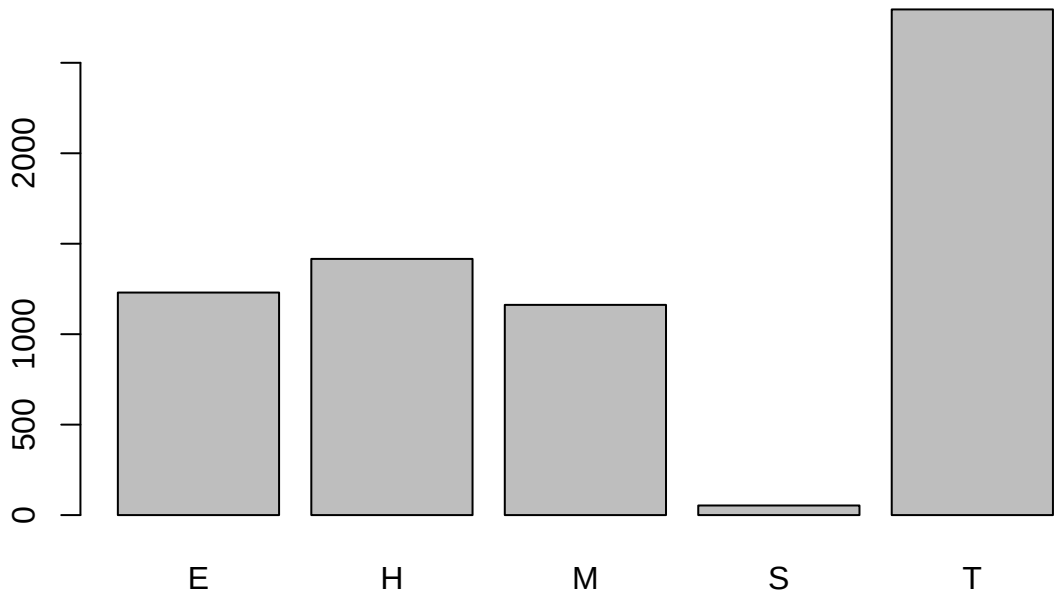
Freight trains have the highest number of accidents with extreme damage compared to the other types. Thus, we will further investigate freight trains by including this variable as a part of our hypotheses.

```
bwplot(as.factor(TYPEQ)~ACCDMG, data = xdmgnd, main = "Box Plots of Accident
       Damage", xlab = "Damage ($)", ylab = "Type Train")
```



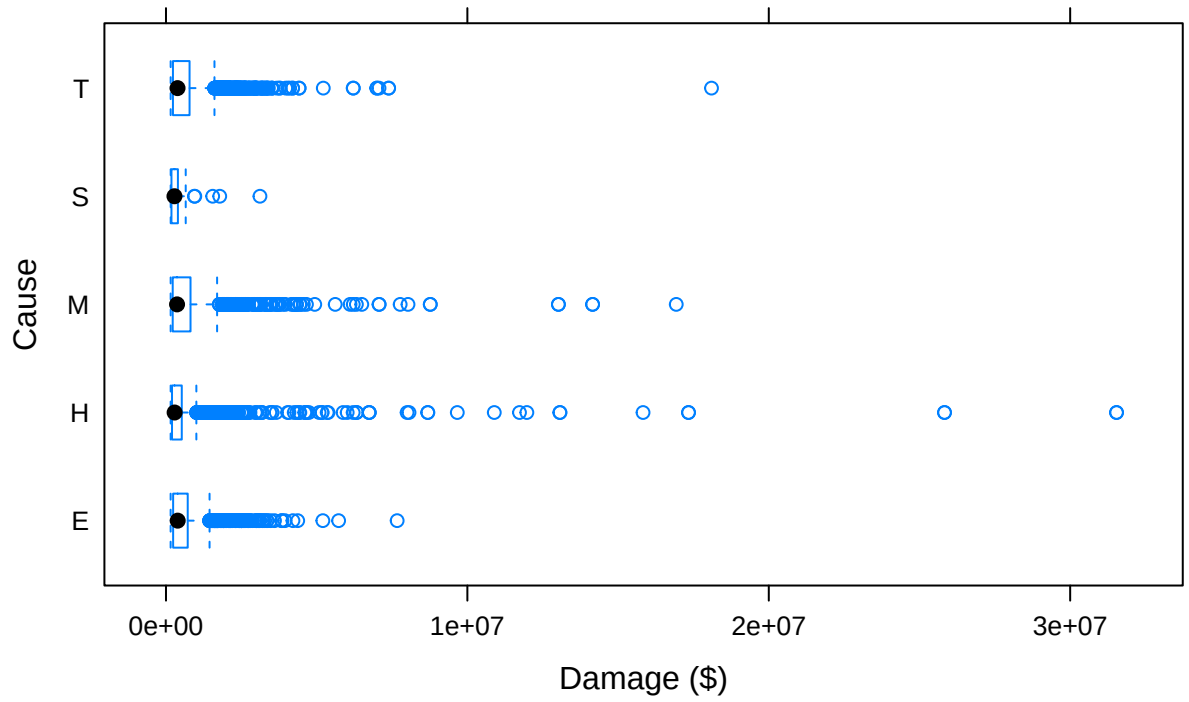
Although passenger trains have the accidents with the highest amount of accident damage, freight trains have a higher frequency of accidents with substantial accident damage.

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



Although Cause T (Rack, Roadbed and Structures) has the largest number of accidents with extreme accident damage, we thought that investigating Human Factors (the second largest number of accidents with extreme damage) would be a more actionable hypotheses.

```
bwplot(as.factor(Cause)~ACCDMG, data = xdmgnd, main = "Box Plots of Accident
       Damage", xlab = "Damage ($)", ylab = "Cause")
```



Additionally, the most costly accidents in this data set are accidents caused by human factors, thus human factors is an important factor to evaluate with our hypotheses.

Casualties: We used the following visualizations, summary statistics, and external supporting evidence to arrive at our hypotheses for Casualties.

```
# Derailment, Highway-Rail Crossing
casnd$TYPE <- factor(casnd$TYPE, labels = c("Derailment", "HeadOn", "Rearend",
                                            "Side", "Raking", "BrokenTrain",
                                            "Hwy-Rail", "GradeX", "Obstruction",
                                            "Explosive",
                                            "Fire","Other","SeeNarrative" ))
barplot(table(casnd$TYPE))
```


```
mean(casnd$Casualty)
```

```
## [1] 3.460352
```

```
max(casnd$Casualty)
```

```
## [1] 1001
```

As seen in the above histogram, Highway-Rail accidents and Derailments have the most number of accidents that result in casualties, so these are important factors to consider when developing hypotheses about casualties.

Based on the above analysis, we have been able to form some well-informed hypotheses from the visualizations. They are as follows, listed as null and alternative hypotheses and followed by how we arrived at them:

..As for accident severity in regards to **Accident Damage (ACCDMG)**: *1) H0: Accidents involving freight trains do not significantly increase accident damage. Ha: Accidents involving freight trains do significantly increase accident damage.*

This is an actionable hypothesis because more regulations can be put in place to promote safe driving with freight trains in dangerous areas.

2) H0: Accidents caused by Human Factors do not significantly increase accident damage. Ha Accidents caused by Human Factors do significantly increase accident damage.

This is an actionable hypothesis because better recurrent training programs can be implemented to improve conductor and engineer abilities to safely manage tasks when in control of a train.

..As for accident severity in regards to **Casualties**:

1) H0: Derailment caused accidents do not significantly increase the number of casualties.
Ha: Derailment caused accidents do significantly increase the number of casualties.

This is an actionable hypothesis because train wheels and tracks can be manufactured in order to prevent derailments.

2) H0: Highway-rail crossing accidents do not significantly increase number of casualties. Ha: Highway-rail crossing accidents are do significantly increase the number of casualties.

This is actionable because if the null hypothesis is accepted, then trains can be required to follow more safety precautions when traveling near highways, and other traffic laws can be put in place to encourage safe driving around highway-railroad crossings.

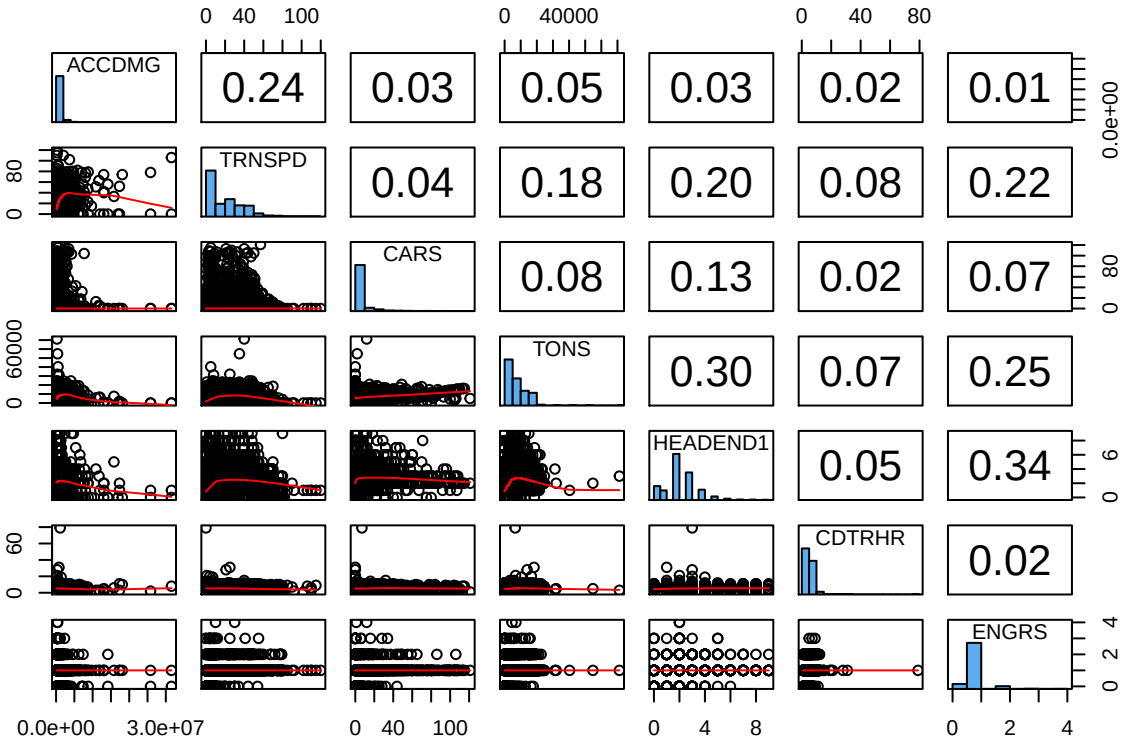
Part 2: ACCDMG Analysis

Hypotheses 1 and 2

1) H0: Accidents involving freight trains do not significantly increase accident damage. Ha: Accidents involving freight trains do significantly increase accident damage.

2) H0: Accidents caused by Human Factors do not significantly increase accident damage. Ha Accidents caused by Human Factors do significantly increase accident damage.

```
source("SPM_Panel.R")
uva.pairs(xdmgnd[,c("ACCDMG", "TRNSPD", "CARS", "TONS", "HEADEND1", "CDTRHR",
                    "ENGRS")])
```



Based on the visualizations and correlation values from the scatter plot matices above, we have decided to include TRNSPD, CARS, TONS, and HEADEND1 as quantitative variables in our model.

To test our hypotheses, we will include the two treated categorical variables, Freight and HF (Human Factors), with the selected quantitative variables in our linear model.

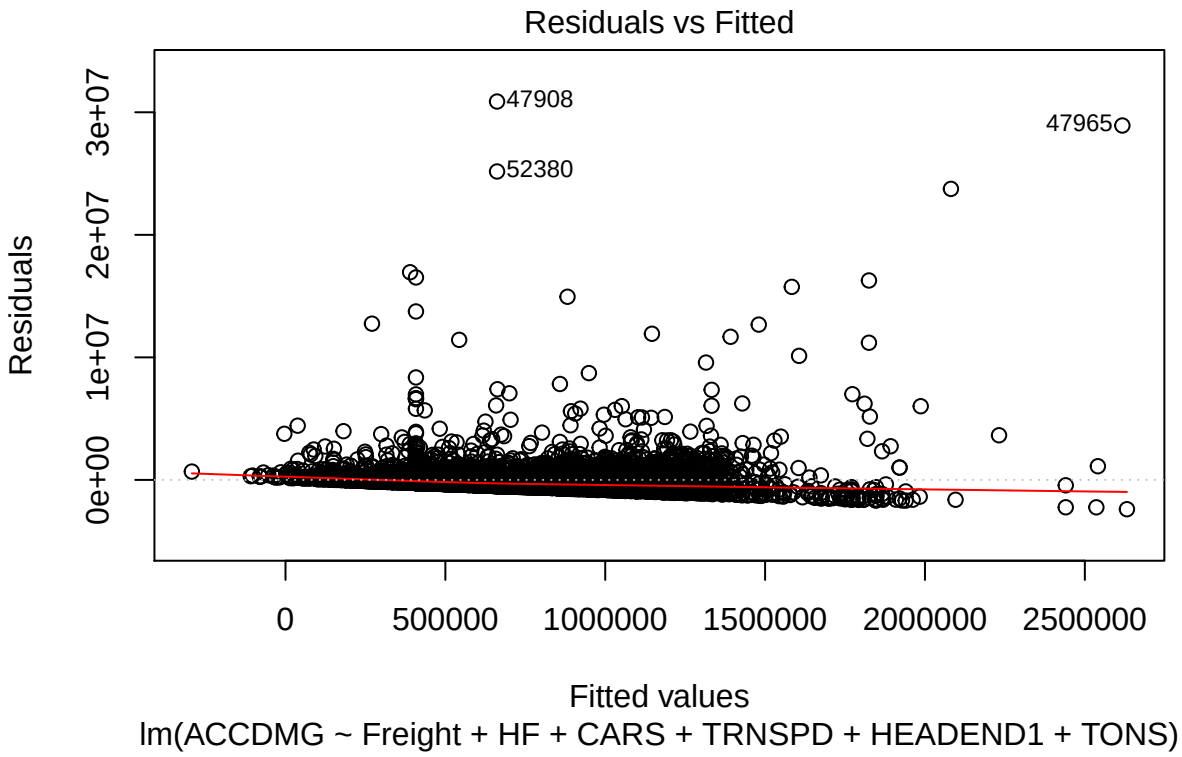
```
xdmgnd.lm1 <- lm(ACCDMG ~ Freight + HF + CARS + TRNSPD + HEADEND1 + TONS,
                 data=xdmgnd)
summary(xdmgnd.lm1)
```

```
##
## Call:
## lm(formula = ACCDMG ~ Freight + HF + CARS + TRNSPD + HEADEND1 +
##     TONS, data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2390342  -395614  -184295    78828 30877114
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.080e+05  3.530e+04  11.558  < 2e-16 ***
## Freight1    -1.503e+05  4.580e+04  -3.281  0.00104 **
## HF1          2.537e+05  3.749e+04   6.766 1.44e-11 ***
## CARS         3.091e+03  1.162e+03   2.659  0.00786 **
## TRNSPD       1.915e+04  8.867e+02  21.593  < 2e-16 ***
## HEADEND1    -7.380e+04  1.246e+04  -5.921 3.35e-09 ***
```

```
## TONS         1.446e+01  2.921e+00   4.951 7.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1193000 on 6649 degrees of freedom
## Multiple R-squared:  0.07275,    Adjusted R-squared:  0.07192
## F-statistic: 86.95 on 6 and 6649 DF,  p-value: < 2.2e-16
```

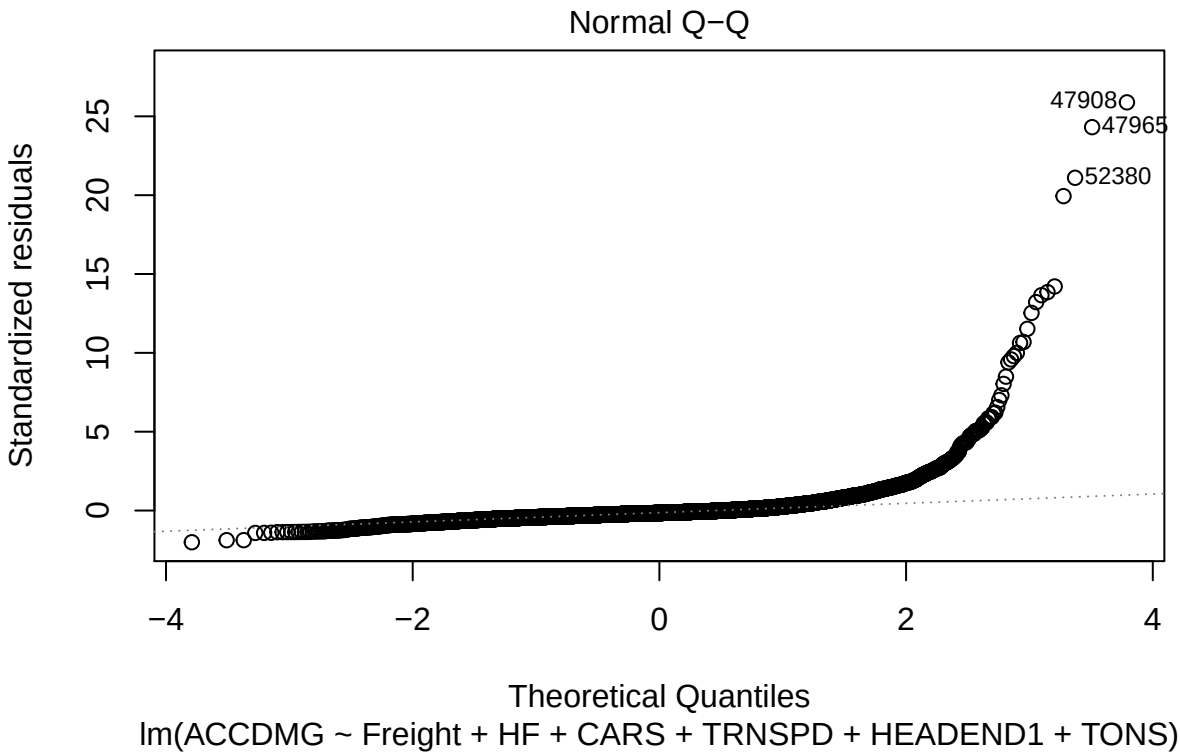
#Plot diagnostic graphs individually

```
plot(xdmgnd.lm1,which=1) #Residual vs. Fitted
```



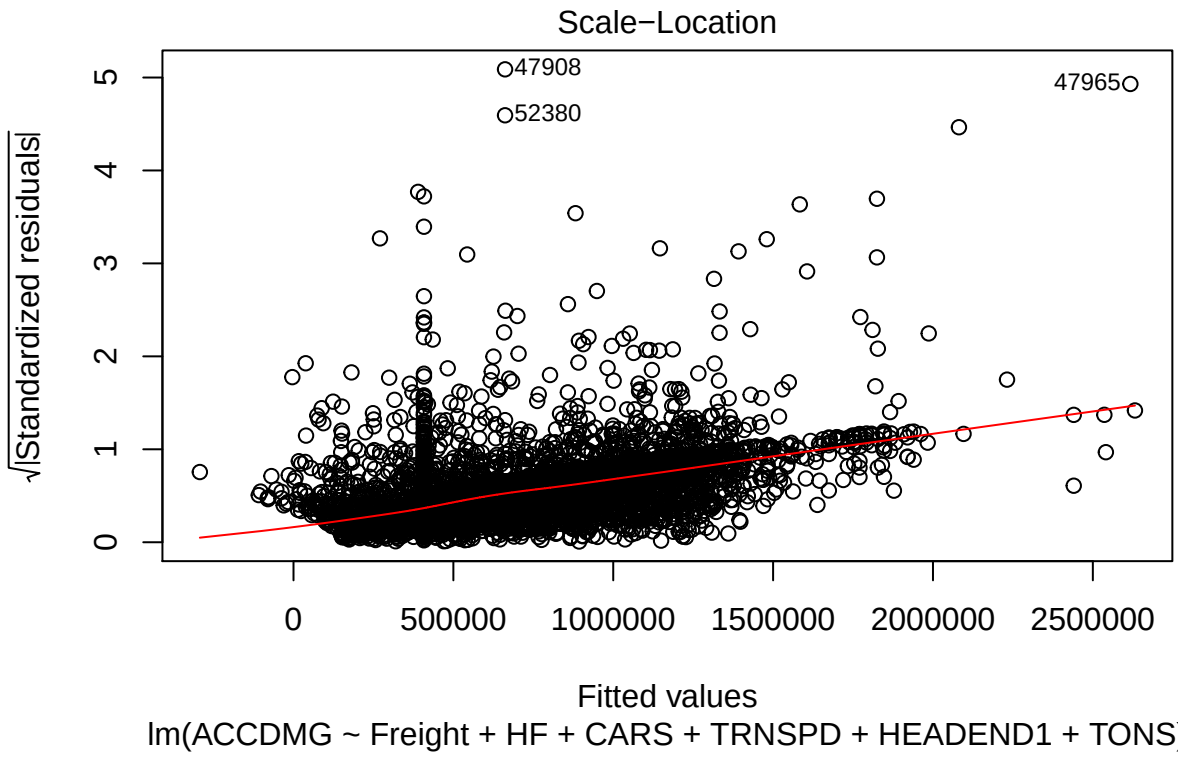
This plot violates the assumptions for the residual vs fitted diagnostic plot because there is not a mean of zero and there is not a constant variance.

```
plot(xdmgnd.lm1,which=2) #QQ
```



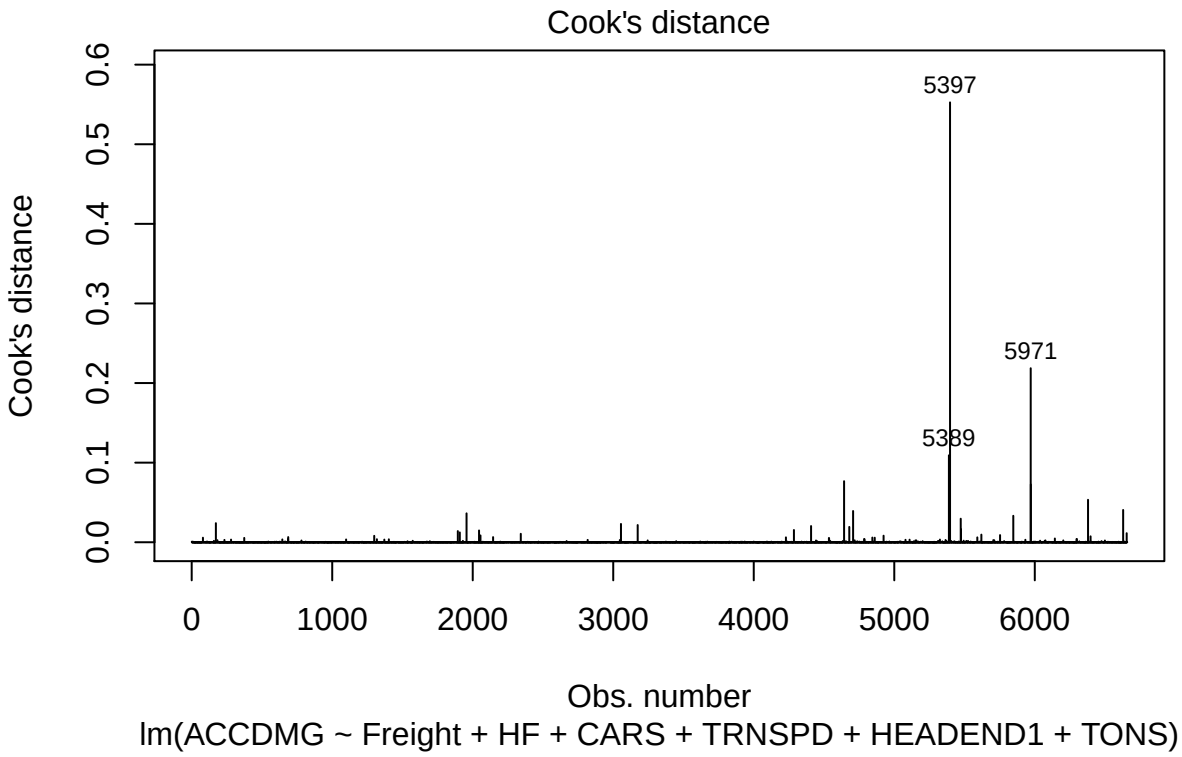
This graph shows a violation of assumptions because the points clearly do not follow the line. This represents a failure of the Gaussian assumption for the error term.

```
plot(xdmgnd.lm1,which=3) #Scale-Location
```

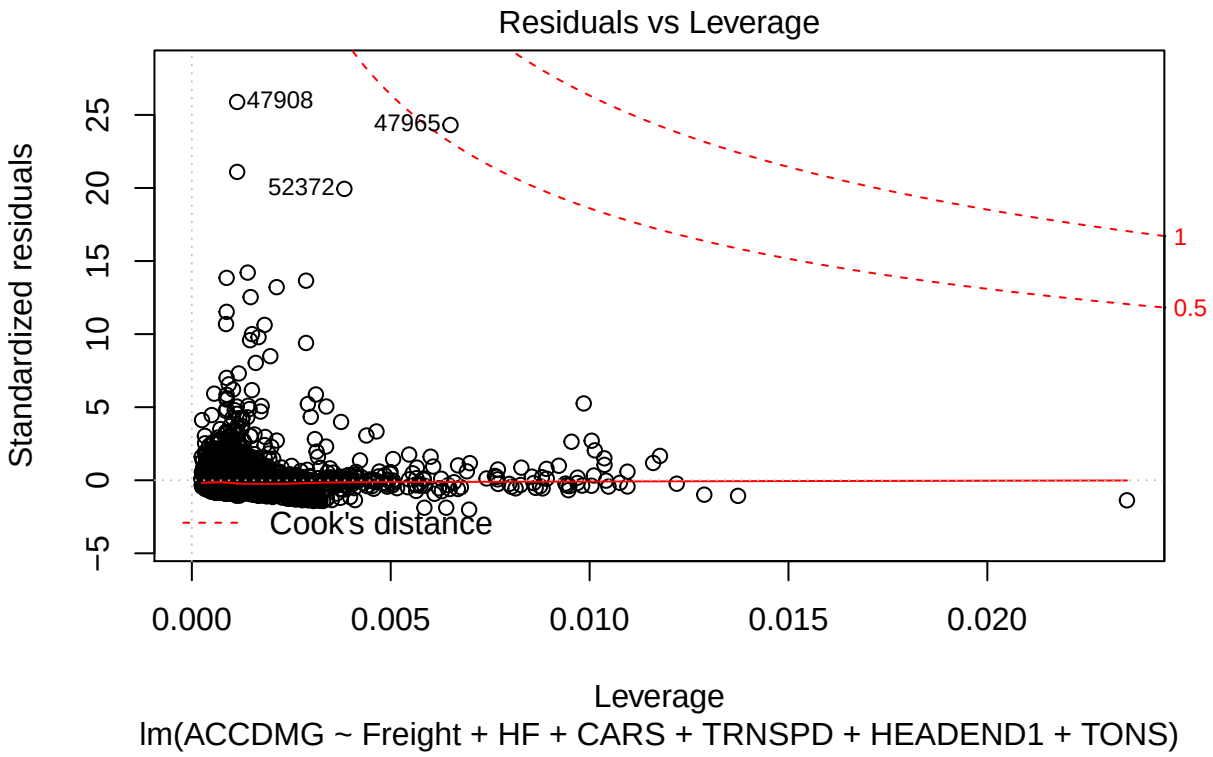


This graph shows a violation of assumptions because the points do not have a mean of zero and there is not a constant variance.

```
plot(xdmgnd.lm1,labels.id = NULL, which=4) #Cook's distance
```




```
plot(xdmgnd.lm1,which=5) #Redisuals vs. Leverage
```



These graphs shows a violation of assumptions because there are incidents with very large Cook's distances.

The analysis above shows that this model does not meet the assumptions required for assessment of this linear model. So, we can conclude that this is not a good model for assessing casualties from train accidents. Therefore, we next perform a log transormation of this model to assess if this creates a more effective model of casualties.

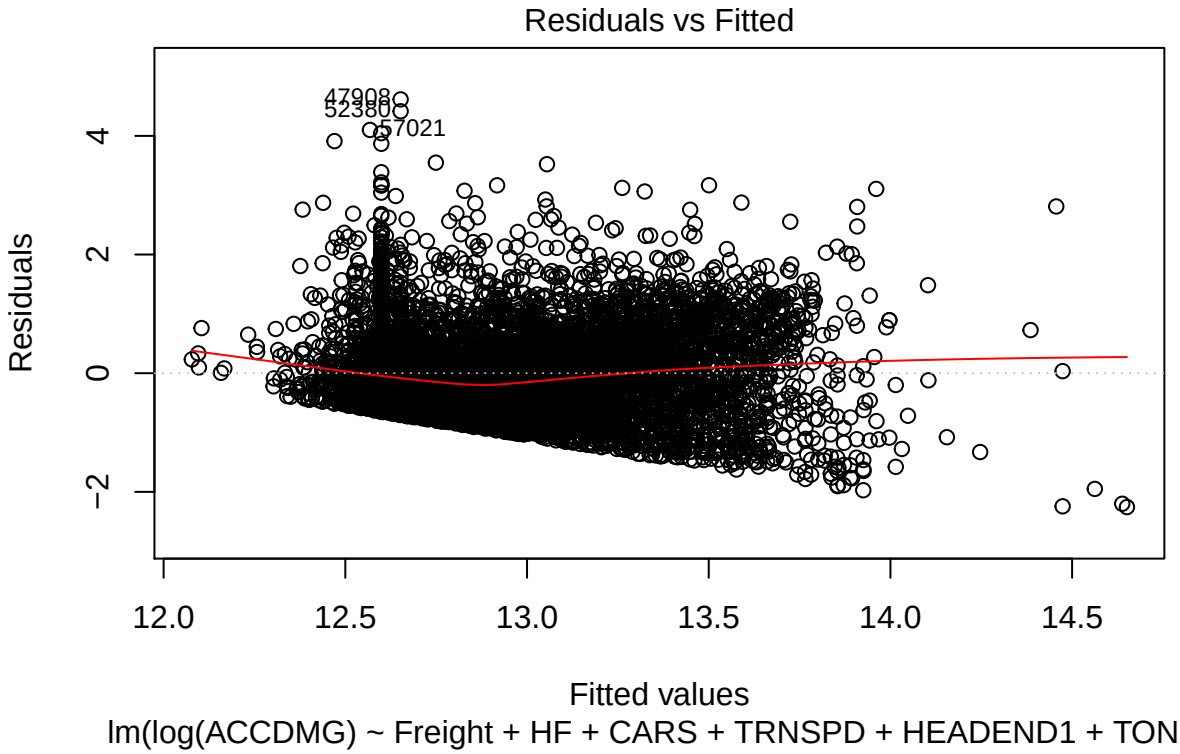
```
xdmgnd.lm2 <- lm(log(ACCDMG) ~ Freight + HF + CARS + TRNSPD + HEADEND1 + TONS,
                 data=xdmgnd)
summary(xdmgnd.lm2)
```

```
##
## Call:
## lm(formula = log(ACCDMG) ~ Freight + HF + CARS + TRNSPD + HEADEND1 +
##     TONS, data = xdmgnd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2567 -0.5296 -0.1374  0.4322  4.6145
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.260e+01  2.232e-02 564.376  < 2e-16 ***
## Freight1     2.157e-02  2.897e-02   0.745   0.4565
## HF1          5.323e-02  2.371e-02   2.245   0.0248 *
## CARS         3.113e-03  7.352e-04   4.234 2.33e-05 ***
## TRNSPD       1.770e-02  5.608e-04  31.556  < 2e-16 ***
```

```
## HEADEND1    -7.170e-02  7.882e-03  -9.097  < 2e-16 ***
## TONS         2.089e-05  1.847e-06  11.308  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7548 on 6649 degrees of freedom
## Multiple R-squared:  0.1765, Adjusted R-squared:  0.1758
## F-statistic: 237.6 on 6 and 6649 DF,  p-value: < 2.2e-16
```

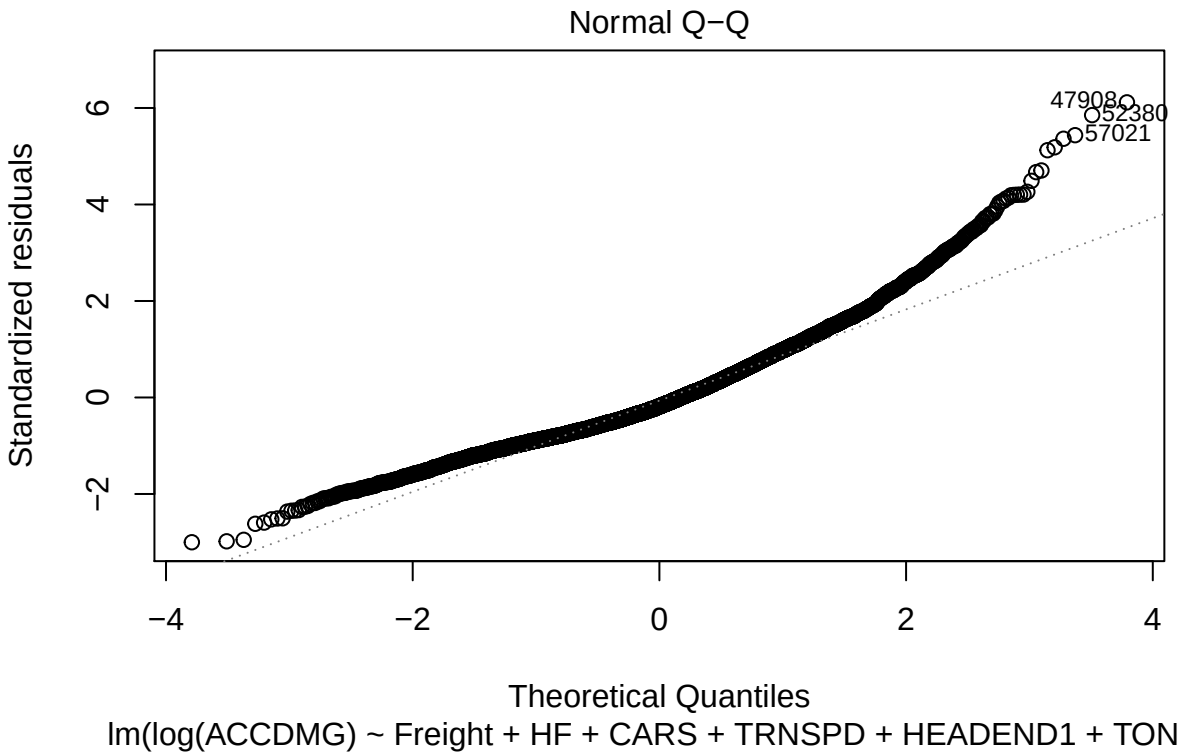
#Plot diagnostic graphs individually

```
plot(xdmgnd.lm2,which=1) #Residual vs. Fitted
```



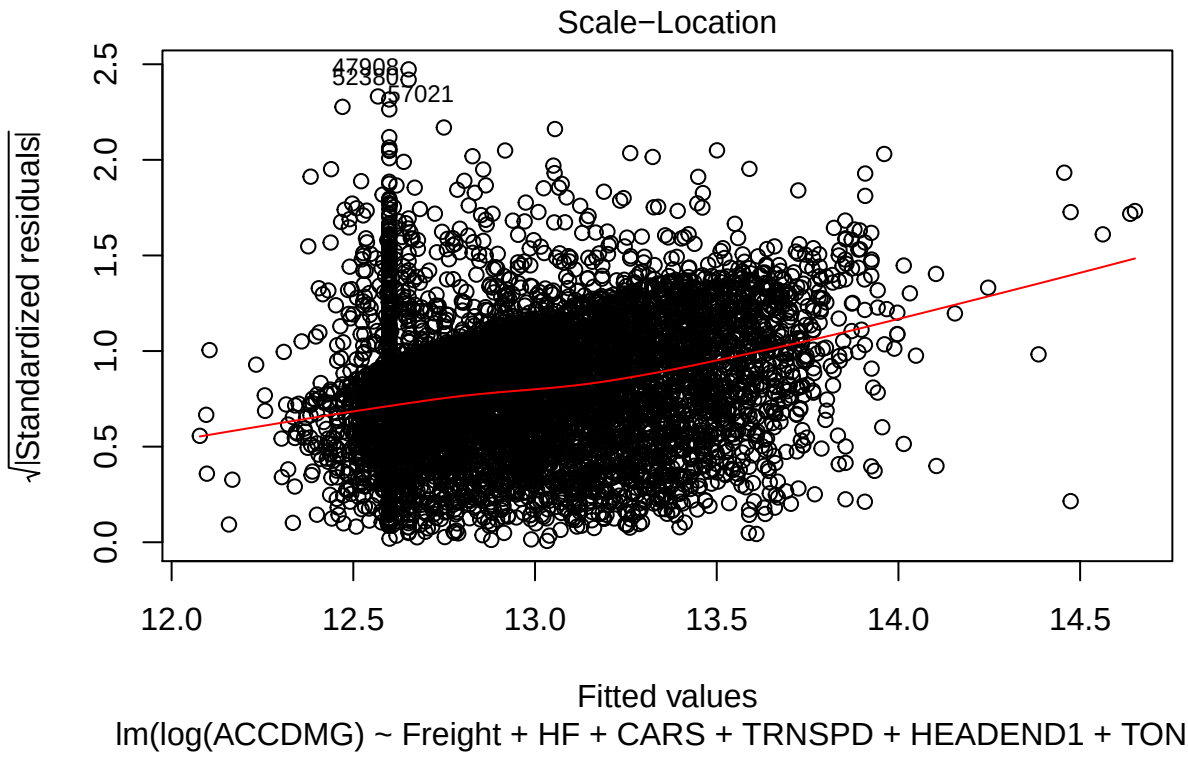
This graph shows significant improvement from the non-transformed model (xdmgnd.lm1), as there is more constant variance and the mean is closer to 0.

```
plot(xdmgnd.lm2,which=2) #QQ
```



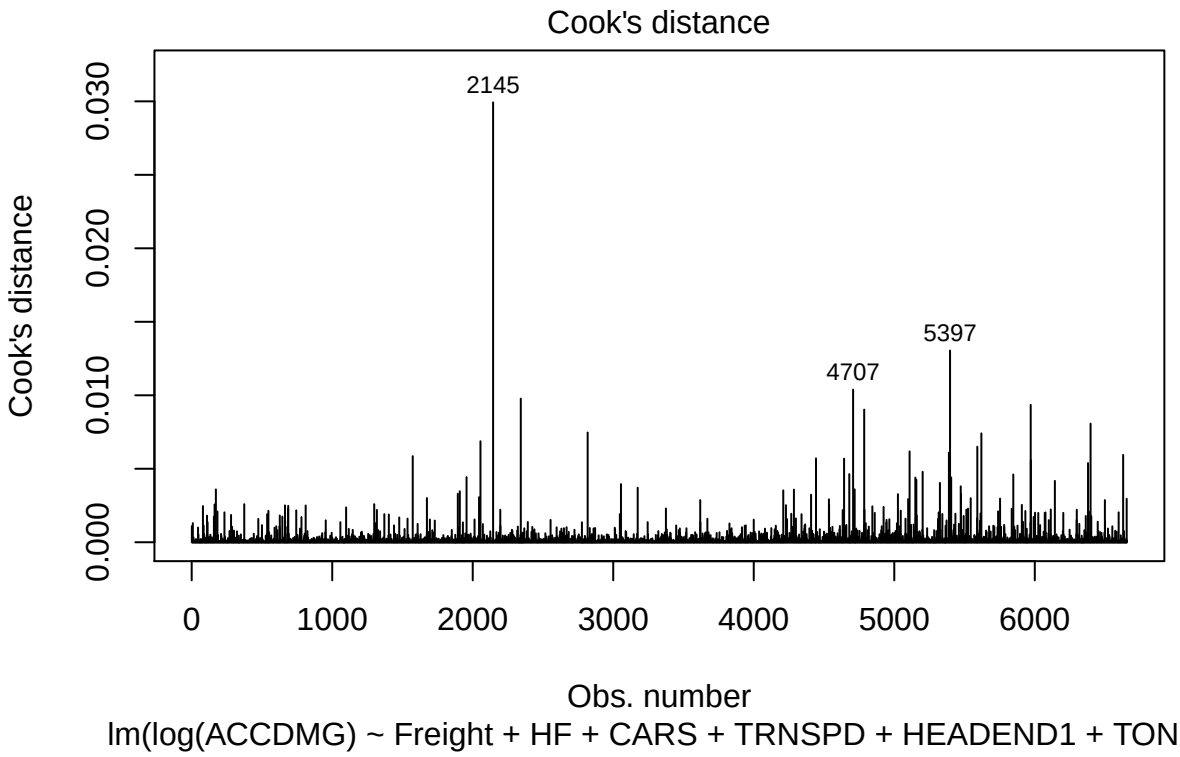
This graph shows significant improvement from the non-transformed model (xdmgnd.lm1), as the points are more linear. However, there is still some deviation from the line which represents a failure of the Gaussian assumption for the error term.

```
plot(xdmgnd.lm2,which=3) #Scale-Location
```

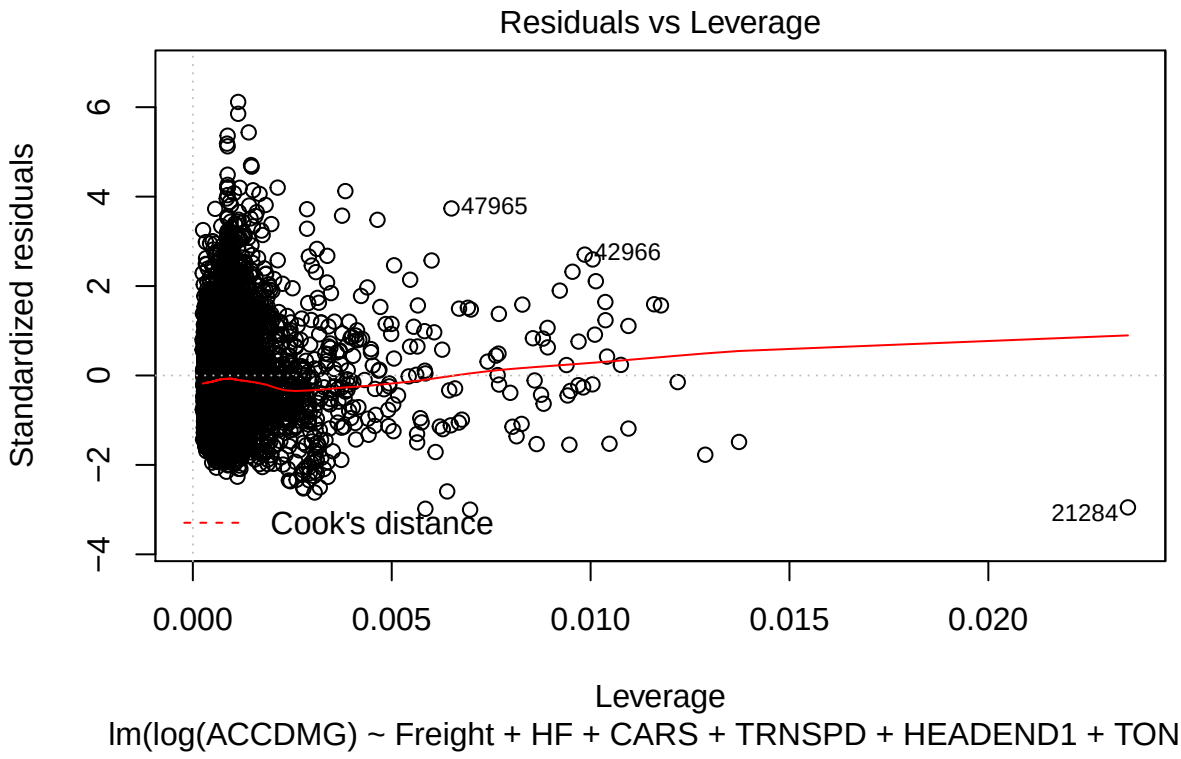


Although the points still do not have a mean of zero after the transformation, there is more constant variance.

```
plot(xdmgnd.lm2,labels.id = NULL, which=4) #Cook's distance
```




```
plot(xdmgnd.lm2,which=5) #Redisuals vs. Leverage
```



Clearly, the log transform helped reduce the impact of influential point as the Cook's distances are smaller than in the previous model.

Based on the diagnostic plots, we feel confident with this transformed model and will run a stepwise regression to determine if we can further improve it.

```
xdmgnd.lm2.step<-step(xdmgnd.lm2, trace=T)
```

```
## Start:  AIC=-3737.46
## log(ACCDMG) ~ Freight + HF + CARS + TRNSPD + HEADEND1 + TONS
##
##             Df Sum of Sq    RSS     AIC
## - Freight    1      0.32 3788.5 -3738.9
## <none>                   3788.2 -3737.5
## - HF         1      2.87 3791.1 -3734.4
## - CARS       1     10.21 3798.4 -3721.5
## - HEADEND1   1     47.14 3835.4 -3657.1
## - TONS       1     72.85 3861.1 -3612.7
## - TRNSPD     1    567.34 4355.6 -2810.6
##
## Step:  AIC=-3738.9
## log(ACCDMG) ~ HF + CARS + TRNSPD + HEADEND1 + TONS
##
##             Df Sum of Sq    RSS     AIC
## <none>                   3788.5 -3738.9
## - HF         1      2.72 3791.3 -3736.1
## - CARS       1     10.54 3799.1 -3722.4
```

```
## - HEADEND1  1     53.34 3841.9 -3647.9
## - TONS      1     99.10 3887.6 -3569.0
## - TRNSPD    1    609.89 4398.4 -2747.4
```

```
summary(xdmgnd.lm2.step)
```

```
##
## Call:
## lm(formula = log(ACCDMG) ~ HF + CARS + TRNSPD + HEADEND1 + TONS,
##     data = xdmgnd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2748 -0.5290 -0.1390  0.4332  4.6126
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.260e+01  2.181e-02 577.702  < 2e-16 ***
## HF1          5.160e-02  2.361e-02   2.186   0.0289 *
## CARS         3.153e-03  7.331e-04   4.301 1.72e-05 ***
## TRNSPD       1.780e-02  5.440e-04  32.719  < 2e-16 ***
## HEADEND1    -6.924e-02  7.156e-03  -9.676  < 2e-16 ***
## TONS         2.153e-05  1.632e-06  13.189  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7548 on 6650 degrees of freedom
## Multiple R-squared:  0.1765, Adjusted R-squared:  0.1758
## F-statistic:   285 on 5 and 6650 DF,  p-value: < 2.2e-16
```

```
anova(xdmgnd.lm2,xdmgnd.lm2.step)
```

```
## Analysis of Variance Table
##
## Model 1: log(ACCDMG) ~ Freight + HF + CARS + TRNSPD + HEADEND1 + TONS
## Model 2: log(ACCDMG) ~ HF + CARS + TRNSPD + HEADEND1 + TONS
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   6649 3788.2
## 2   6650 3788.5 -1   -0.3159 0.5545 0.4565
```

The stepwise function removed the Freight variable from our model, indicating that this may not be as important of a factor as we thought when forming our hypothesis. Since the p-value from the partial F-test is greater than 0.05, we fail to reject the null hypothesis and choose the smaller model due to Okham's Razor.

```
summary(xdmgnd.lm1)$adj.r.squared
```

```
## [1] 0.07191597
```

```
summary(xdmgnd.lm2)$adj.r.squared
```

```
## [1] 0.1757921
```

```
summary(xdmgnd.lm2.step)$adj.r.squared
```

```
## [1] 0.1758473
```

```
#AIC
AIC(xdmgnd.lm1)
```

```
## [1] 205164.3
```

```
AIC(xdmgnd.lm2)
```

```
## [1] 15153.45
```

```
AIC(xdmgnd.lm2.step)
```

```
## [1] 15152.01
```

```
#BIC
AIC(xdmgnd.lm1,k=log(nrow(xdmgnd)))
```

```
## [1] 205218.7
```

```
AIC(xdmgnd.lm2,k=log(nrow(xdmgnd)))
```

```
## [1] 15207.88
```

```
AIC(xdmgnd.lm2.step,k=log(nrow(casnd)))
```

```
## [1] 15194.59
```

The adjusted R2 value for the first model is 0.07191597 and for the transformed model it is 0.1757921, which is clearly a significant improvement. The stepwise regression, which removed the freight variable has an R2 value of 0.1758473, which is a slight improvement from the original transformed model. Additionally, which each new model, the AIC and BIC decreased, particularly after doing the log transformation.

The p-values for both Freight (0.00104, **0.4565**, NA) and HF (1.4e-11, 0.0248, **0.0289**) increased when we performed the transformation. However, Human Factors remains significant in each model, while Freight does not. Thus, we can reject the null hypothesis that Human Factors does not significantly increase accident damage and we fail to reject the null hypothesis that Freight does not significantly increase accident damage.

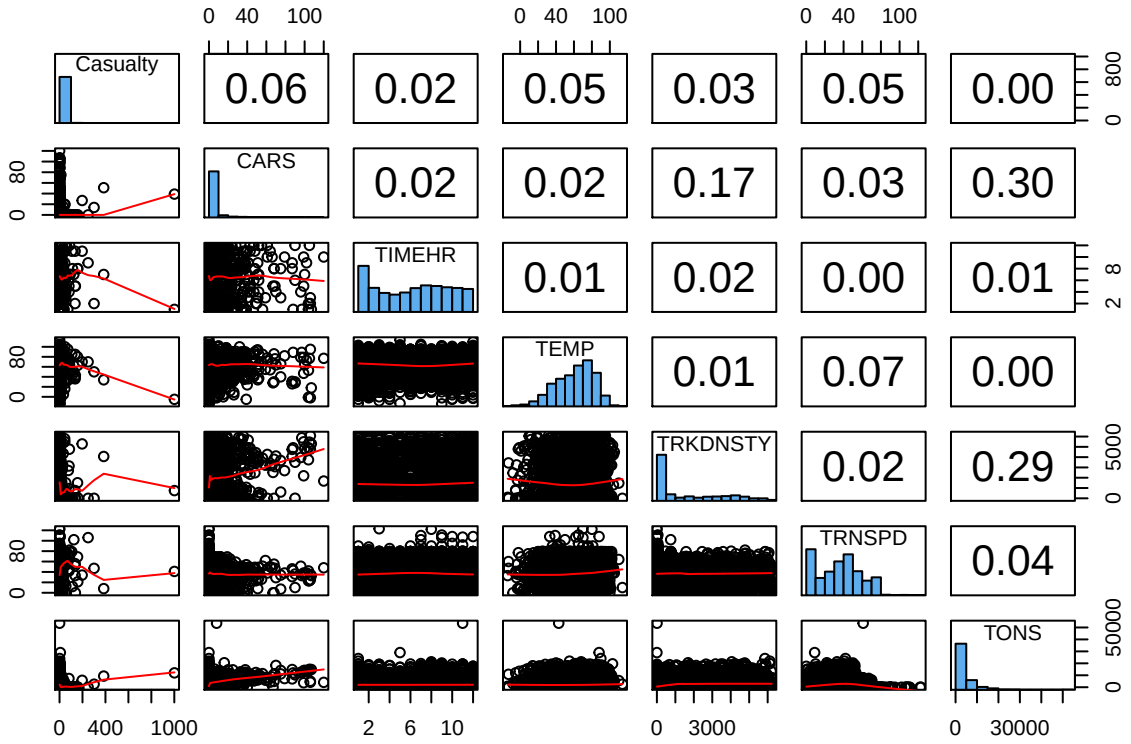
Part 3: Casualties

Hypotheses 1 and 2

1) H0: Derailment caused accidents do not significantly increase the number of casualties.
Ha: Derailment caused accidents do significantly increase the number of casualties.

2) H0: Highway-rail crossing accidents do not significantly increase number of casualties. Ha: Highway-rail crossing accidents are do significantly increase the number of casualties.

```
source("SPM_Panel.R")
uva.pairs(casnd[,c("Casualty", "CARS", "TIMEHR", "TEMP", "TRKDNSTY", "TRNSPD",
                   "TONS")])
```



Due to the low correlations between casualty and TIMEHR, TONS, and TRKDNSTY, we will not include these variables in our model.

To test our hypotheses, we will include the two treated categorical variables, Derail and HRX (Highway-Rail Crossing), with the selected quantitative variables in our linear model.

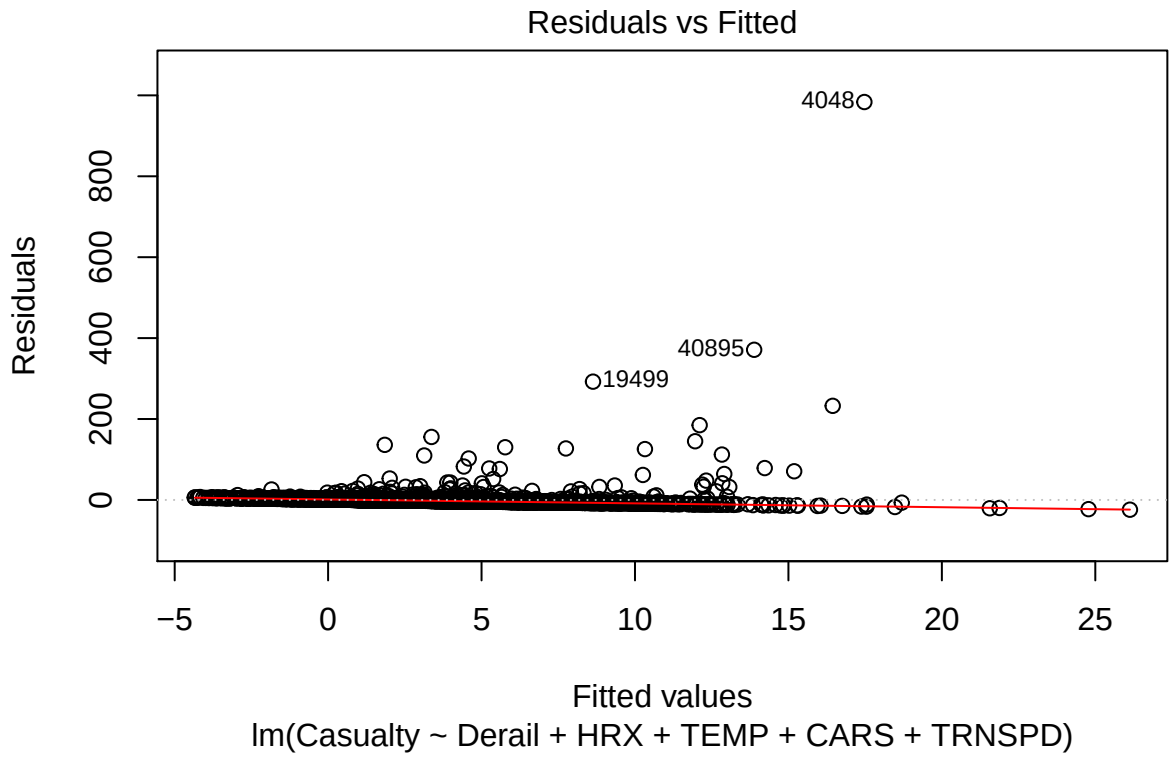
```
casnd.lm1 <- lm(Casualty ~ Derail + HRX + TEMP + CARS + TRNSPD, data=casnd)
summary(casnd.lm1)
```

```
##
## Call:
## lm(formula = Casualty ~ Derail + HRX + TEMP + CARS + TRNSPD,
##     data = casnd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24.13  -3.56  -1.11   0.63 983.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.55892    1.40450   2.534 0.011325 *
## Derail1      3.96653    1.31165   3.024 0.002514 **
```

```
## HRX1        -4.31697    1.04546  -4.129 3.73e-05 ***
## TEMP        -0.03983    0.01855  -2.147 0.031873 *
## CARS         0.13390    0.03534   3.788 0.000154 ***
## TRNSPD       0.11044    0.01929   5.727 1.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.94 on 3235 degrees of freedom
## Multiple R-squared:  0.02363,    Adjusted R-squared:  0.02212
## F-statistic: 15.66 on 5 and 3235 DF,  p-value: 2.938e-15
```

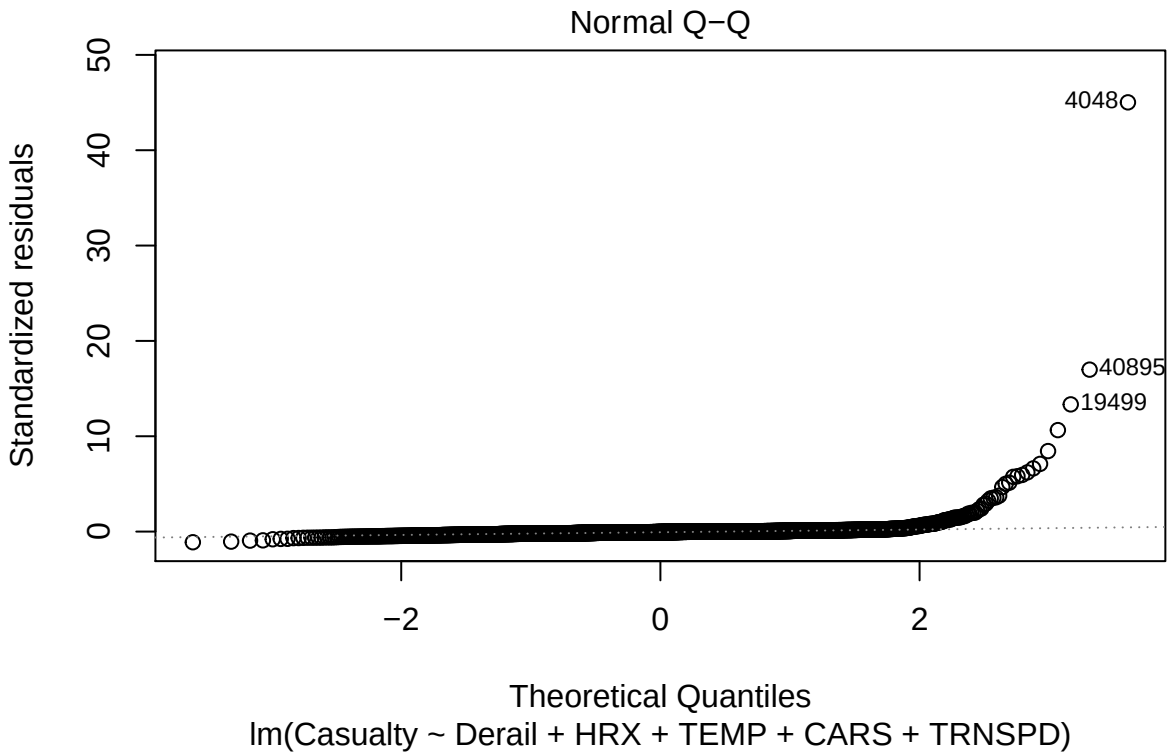
#Plot diagnostic graphs individually

```
plot(casnd.lm1,which=1) #Residual vs. Fitted
```



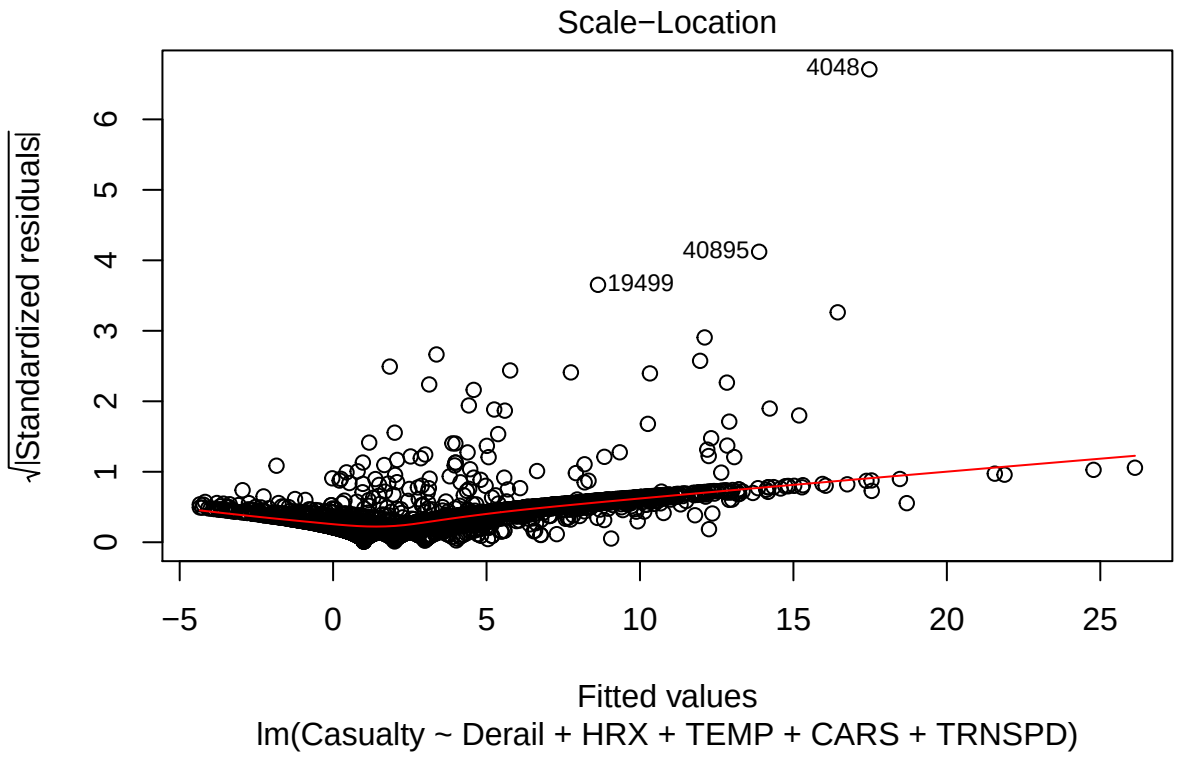
This plot violates the assumptions for the residual vs fitted diagnostic plot because there is not a mean of zero and there is not a constant variance.

```
plot(casnd.lm1,which=2) #QQ
```



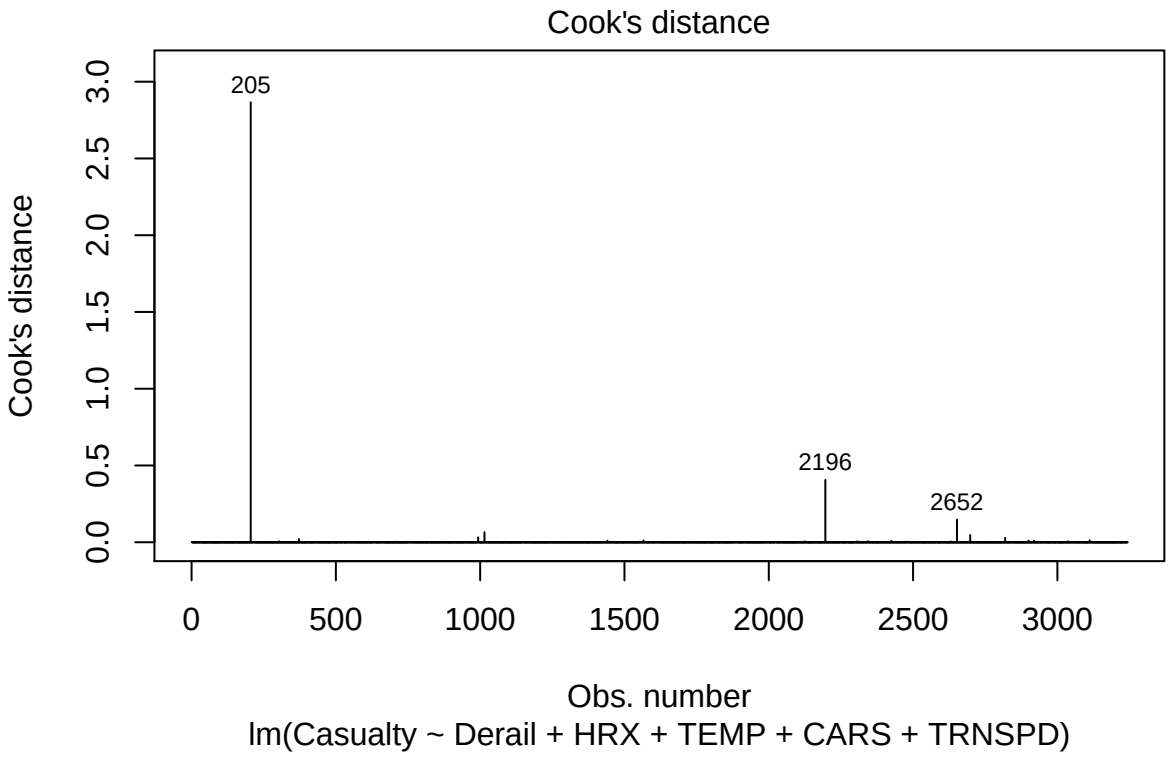
This graph shows a violation of assumptions because the points clearly do not follow the line. This represents a failure of the Gaussian assumption for the error term.

```
plot(casnd.lm1,which=3) #Scale-Location
```

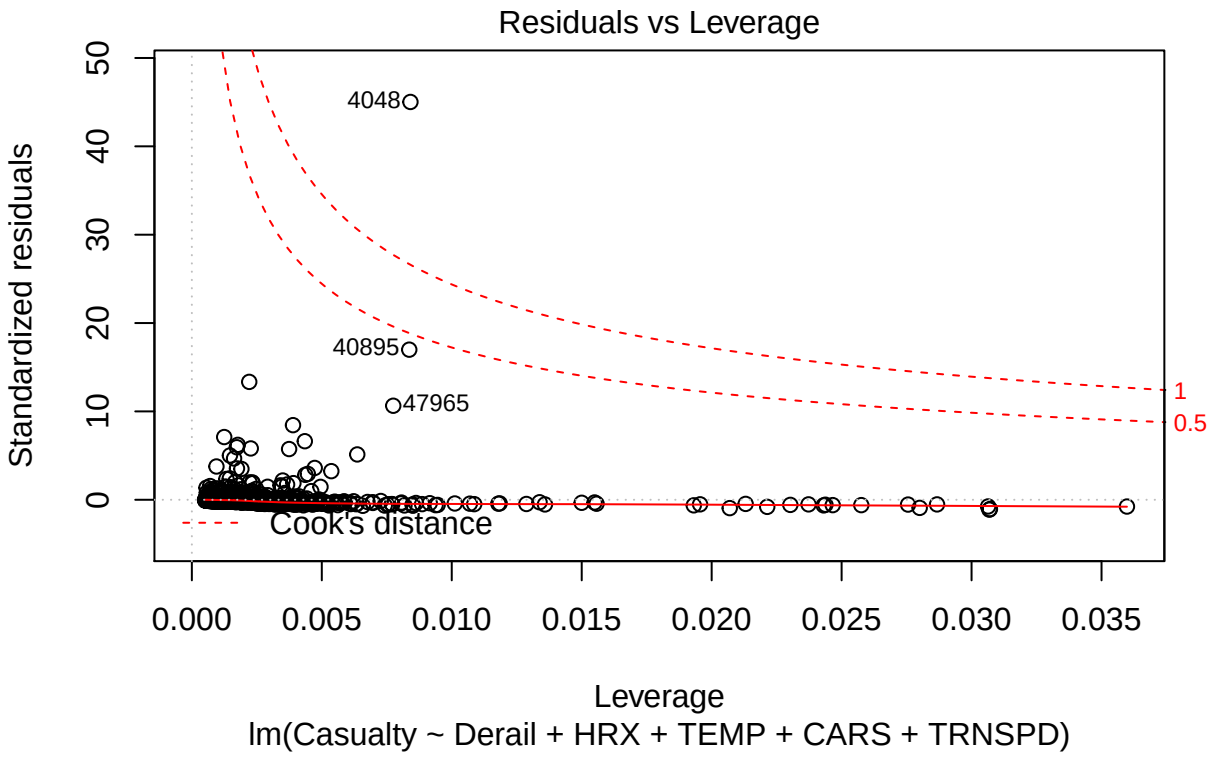


This graph shows a violation of assumptions because the points do not have a mean of zero and there is not a constant variance.

```
plot(casnd.lm1,labels.id = NULL, which=4) #Cook's distance
```




```
plot(casnd.lm1,which=5) #Redisuals vs. Leverage
```



These graphs shows a violation of assumptions because there are incidents with very large Cook's distances.

The analysis above shows that this model does not meet the assumptions required for assessment of this linear model. So, we can conclude that this is not a good model for assessing casualties from train accidents. Therefore, we next perform a log transormation of this model to assess if this creates a more effective model of casualties.

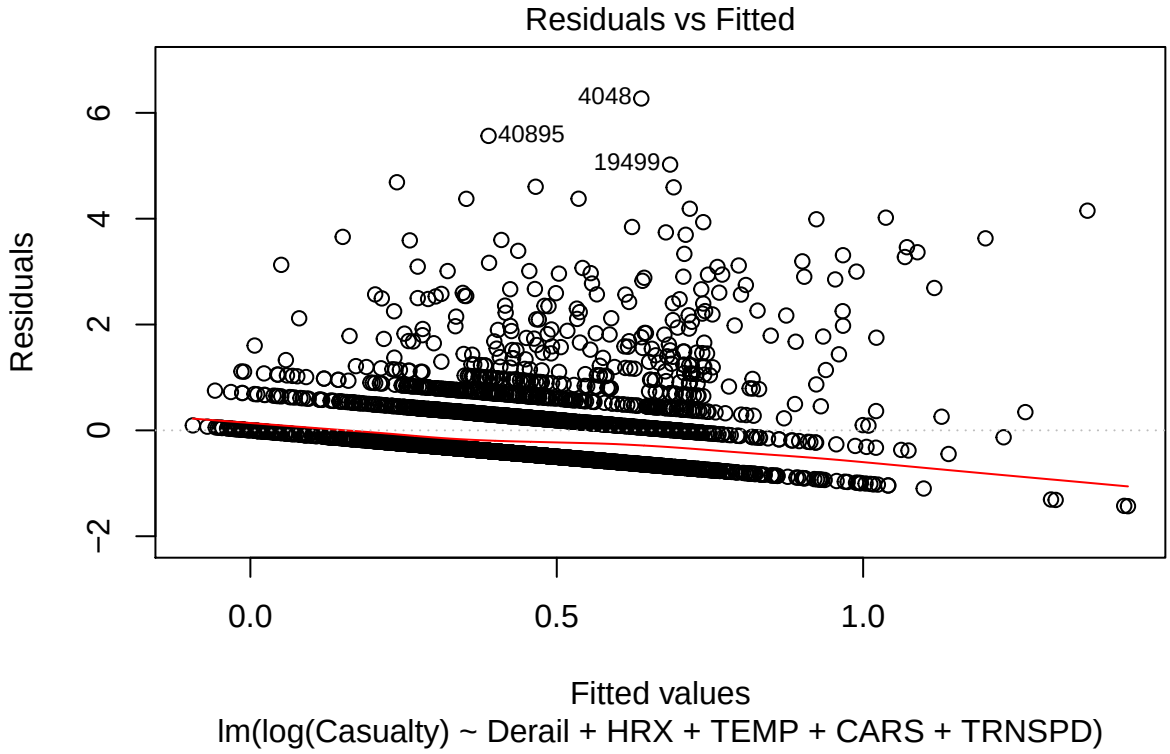
```
casnd.lm2 <- lm(log(Casualty) ~ Derail + HRX + TEMP + CARS + TRNSPD, data=casnd)
summary(casnd.lm2)
```

```
##
## Call:
## lm(formula = log(Casualty) ~ Derail + HRX + TEMP + CARS + TRNSPD,
##     data = casnd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4321 -0.4256 -0.2802  0.2707  6.2708
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1728245  0.0475055   3.638 0.000279 ***
## Derail1      0.0983370  0.0443650   2.217 0.026724 *
## HRX1        -0.3004348  0.0353615  -8.496  < 2e-16 ***
## TEMP         0.0015433  0.0006275   2.459 0.013977 *
## CARS        -0.0001847  0.0011955  -0.154 0.877235
## TRNSPD       0.0093101  0.0006523  14.273  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.742 on 3235 degrees of freedom
## Multiple R-squared:  0.06613,    Adjusted R-squared:  0.06468
## F-statistic: 45.81 on 5 and 3235 DF,  p-value: < 2.2e-16
```

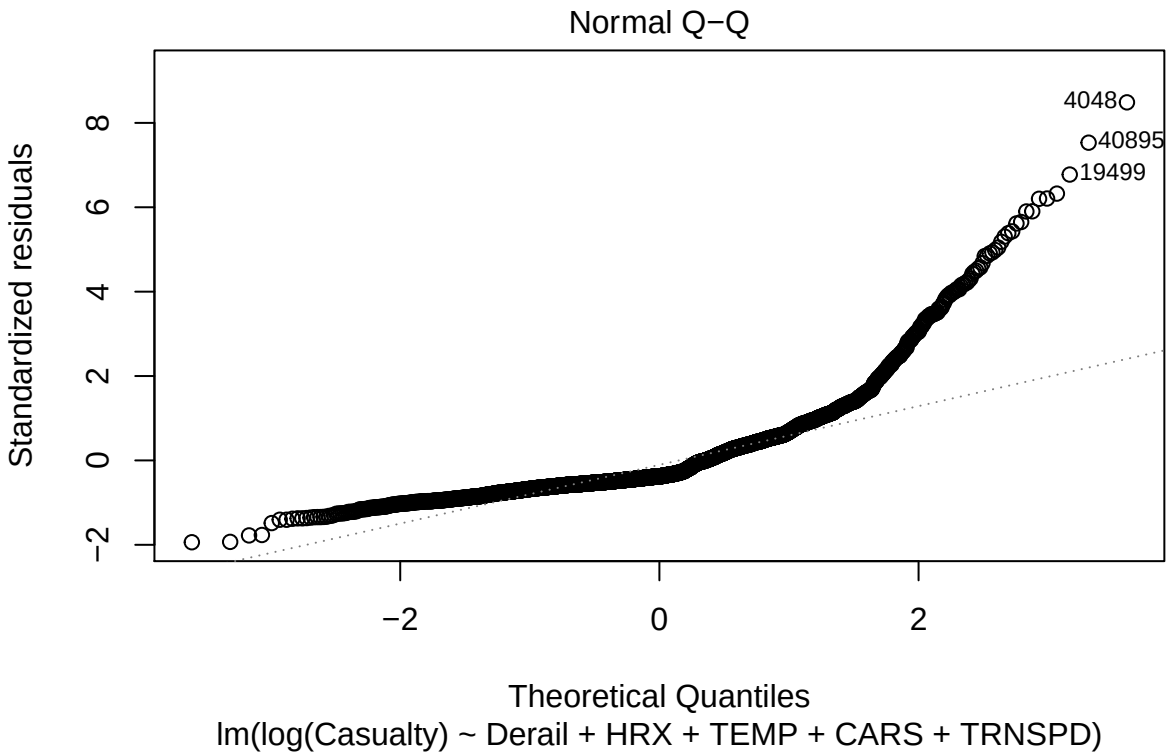
#Plot diagnostic graphs individually

```
plot(casnd.lm2,which=1) #Residual vs. Fitted
```



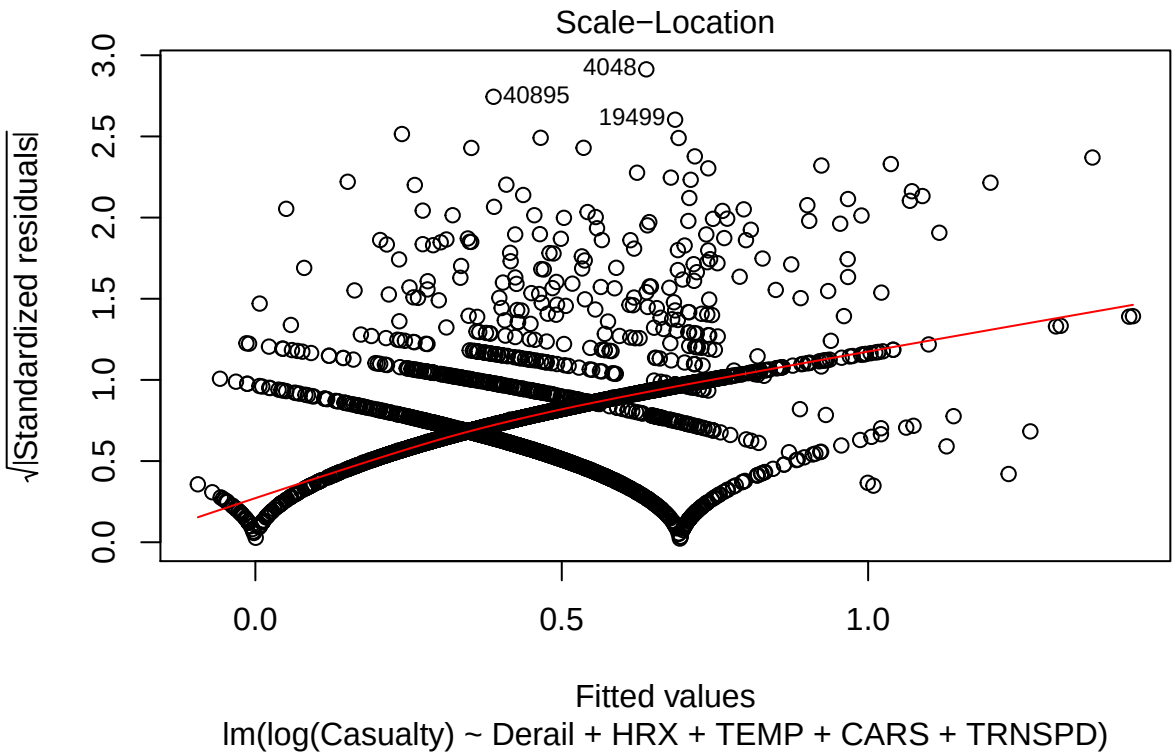
This plot violates the assumptions for the residual vs fitted diagnostic plot because there is not a mean of zero and there is not a constant variance.

```
plot(casnd.lm2,which=2) #QQ
```



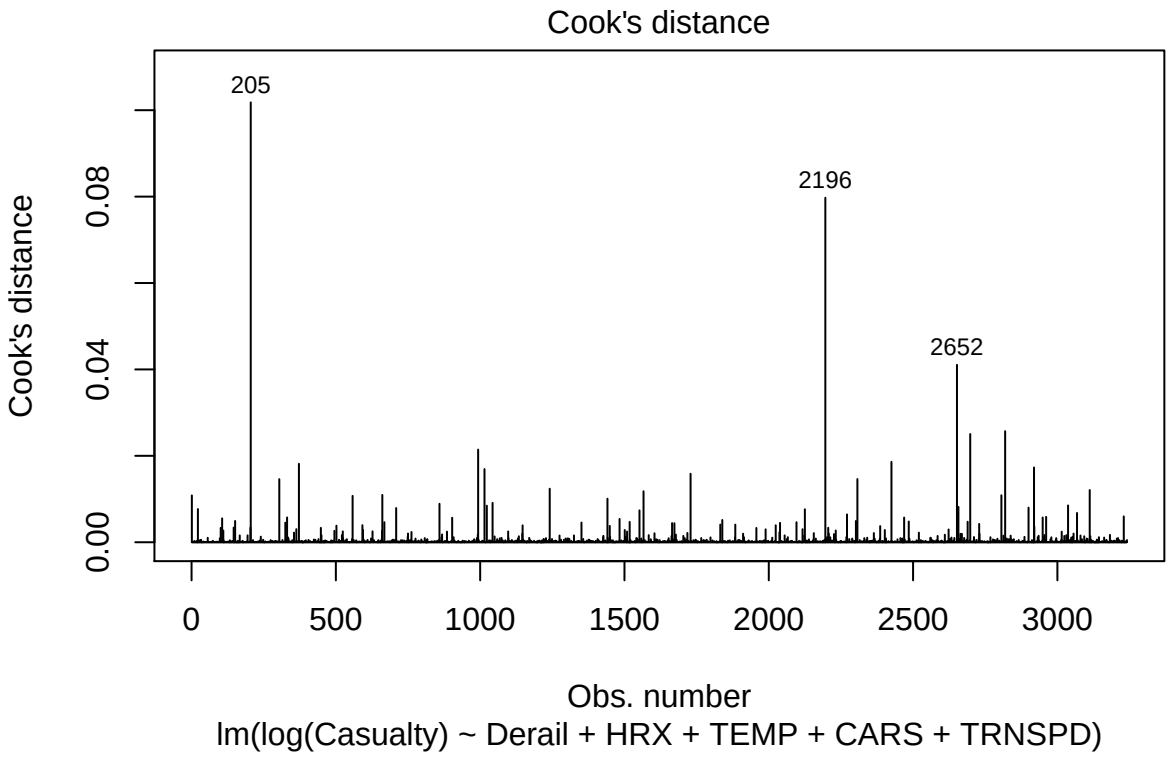
This graph shows a violation of assumptions because the points clearly do not follow the line. This represents a failure of the Gaussian assumption for the error term.

```
plot(casnd.lm2,which=3) #Scale-Location
```

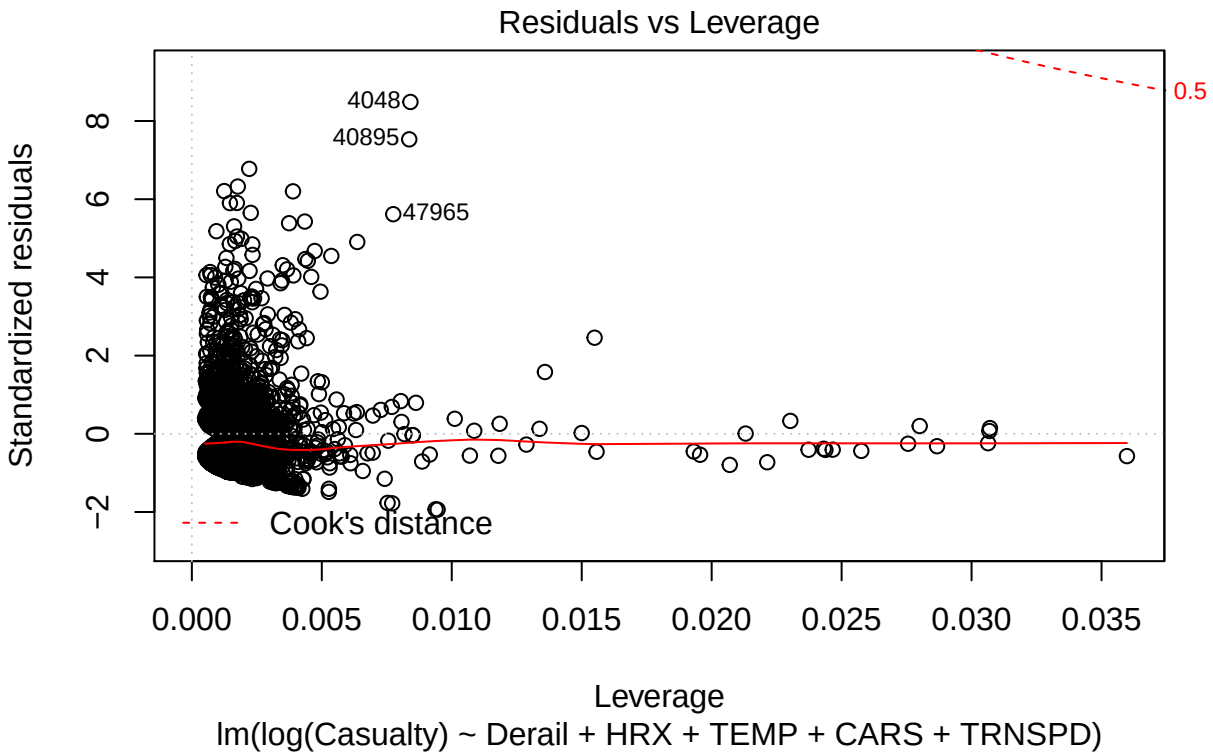


This graph shows a violation of assumptions because the points do not have a mean of zero and there is not a constant variance.

```
plot(casnd.lm2,labels.id = NULL, which=4) #Cook's distance
```




```
plot(casnd.lm2,which=5) #Redisuals vs. Leverage
```



Although these graphs still show a violation of assumptions due to large Cook's distances, there is a slight improvement from the previous model after taking the log transformation.

```
summary(casnd.lm1)
```

```
##
## Call:
## lm(formula = Casualty ~ Derail + HRX + TEMP + CARS + TRNSPD,
##     data = casnd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24.13  -3.56  -1.11   0.63 983.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.55892    1.40450   2.534 0.011325 *
## Derail1      3.96653    1.31165   3.024 0.002514 **
## HRX1        -4.31697    1.04546  -4.129 3.73e-05 ***
## TEMP        -0.03983    0.01855  -2.147 0.031873 *
## CARS         0.13390    0.03534   3.788 0.000154 ***
## TRNSPD       0.11044    0.01929   5.727 1.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.94 on 3235 degrees of freedom
```

```
## Multiple R-squared:  0.02363,    Adjusted R-squared:  0.02212
## F-statistic: 15.66 on 5 and 3235 DF,  p-value: 2.938e-15
```

```
summary(casnd.lm2)
```

```
##
## Call:
## lm(formula = log(Casualty) ~ Derail + HRX + TEMP + CARS + TRNSPD,
##     data = casnd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4321 -0.4256 -0.2802  0.2707  6.2708
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1728245  0.0475055   3.638 0.000279 ***
## Derail1      0.0983370  0.0443650   2.217 0.026724 *
## HRX1        -0.3004348  0.0353615  -8.496  < 2e-16 ***
## TEMP         0.0015433  0.0006275   2.459 0.013977 *
## CARS        -0.0001847  0.0011955  -0.154 0.877235
## TRNSPD       0.0093101  0.0006523  14.273  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.742 on 3235 degrees of freedom
## Multiple R-squared:  0.06613,    Adjusted R-squared:  0.06468
## F-statistic: 45.81 on 5 and 3235 DF,  p-value: < 2.2e-16
```

```
#AIC
AIC(casnd.lm1)
```

```
## [1] 29223.3
```

```
AIC(casnd.lm2)
```

```
## [1] 7271.419
```

```
#BIC
AIC(casnd.lm1,k=log(nrow(casnd)))
```

```
## [1] 29265.88
```

```
AIC(casnd.lm2,k=log(nrow(casnd)))
```

```
## [1] 7314.004
```

The adjusted R2 value for the first model is 0.02212 and for the transformed model it is 0.06468. While the transformed model shows a larger adjusted R2 value and lower AIC and BIC values than the first model, none of these values indicate a strong model. This indicates that neither model is appropriate to assess both

hypotheses related to casualties. Thus, in order to determine an effective model to predict casualty severity, we would like to investigate non-linear regression models.

Although the p-values for Derail (0.002514, 0.026724) and HRX (3.73e-05, $<$ 2e-16) technically indicate significance, the assumptions for the linear models were not met, thus we cannot draw conclusions about our hypotheses from these models.

Part 4: Evidence and Recommendations to FRA

Part a)

ACCDMG The p-values for both Freight (0.00104, **0.4565**, NA) and HF (1.4e-11, 0.0248, **0.0289**) increased when we performed the transformation. However, Human Factors remains significant in each model, while Freight does not. Thus, we can reject the null hypothesis that Human Factors does not significantly increase accident damage and we fail to reject the null hypothesis that Freight does not significantly increase accident damage.

Casualties Although the p-values for Derail (0.002514, 0.026724) and HRX (3.73e-05, $<$ 2e-16) technically indicate significance, the assumptions for the linear models were not met, thus we cannot draw conclusions about our hypotheses from these models. So, in order to reject or accept these hypotheses, more models would need to be explored to perform a statistically significant assessment.

Part b)

ACCDMG We recommend to not focus on freight trains when attempting to reduce accident damage. We do recommend further investigating human factors. They do impact accident damage, so measures could be put in place to reduce accidents caused by human factor errors. According to the US Department of Transportation, the main initiatives of the Action Plan are intended to reduce accidents caused by human factors by addressing fatigue, improving highway-rail-grade crossing safety, and enhancing emergency preparedness training (https://www.transportation.gov/testimony/role-human-factors-rail-accidents).

Casualties Because the diagnostic plot assumptions were not met in any of our models assessing casualties, we recommend looking into non-linear regression models to evaluate accidents with many casualties. Because many of the accidents with casualties only have a small number of casualties, it might be beneficial to look into a Poisson or Weibull models.