Project 2 - Testing Hypotheses, Modeling, Producing Evidence & Recommendations

Will Adorno, Samarth Singh, Mehrdad Fazli

Introduction

Insert short intro here

Current Hypotheses

Total Accident Damage Cost:

1. Accidents caused by Human Factors at high train speeds significantly increase total accident damage cost

- Null Hypothesis: Human Factors combined with train speed do not significantly affect total accident damage cost
- Alternate Hypothesis: Human factors at high train speeds signficantly increases total accident damage cost

2. Derailment accidents that occur at high train speeds significantly increase total accident damage cost

- Null Hypothesis: Derailment accidents combined with train speed do not significantly affect total accident damage cost
- Alternate Hypothesis: Derailment accidents at high train speeds significantly increases total accident damage cost

Number of Casualties:

1. Higher train speeds and accidents caused by human factors cause a significant increase in the number of casualties

- Null Hypothesis: Train speed combined with the human factors accident type does not significantly affect the number of casualties
- Alternate Hypothesis: Accidents caused by human factors at high train speeds significantly increase the number of casualties

2. Derailment accidents on trains with a high number of cars containing HAZMAT will cause a significant increase in the number of casualties

- Null Hypothesis: Derailment accident types combined with the number of cars containing

HAZMAT has no significant effect on the number of casualties

- Alternate Hypothesis: Derailment accidents on trains with a high number of cars containing HAZMAT significantly increases the number of casualties.

Variable Selection

Before creating linear models, it important to screen variables to avoid multicollinearity and limit the number of parameters when including interactions. From Project 1, we know that train speed, number of cars carrying HAZMAT, weight tonnage, cause of accident, and type of accident all appeared to have a strong relationship with one of the severity metrics. There were several other predictors that we also think could be useful such as visibility, weather, train methods, head end train derailments and more. Below is a list of quantative and qualitative variables that we considered

Quantitative Variables

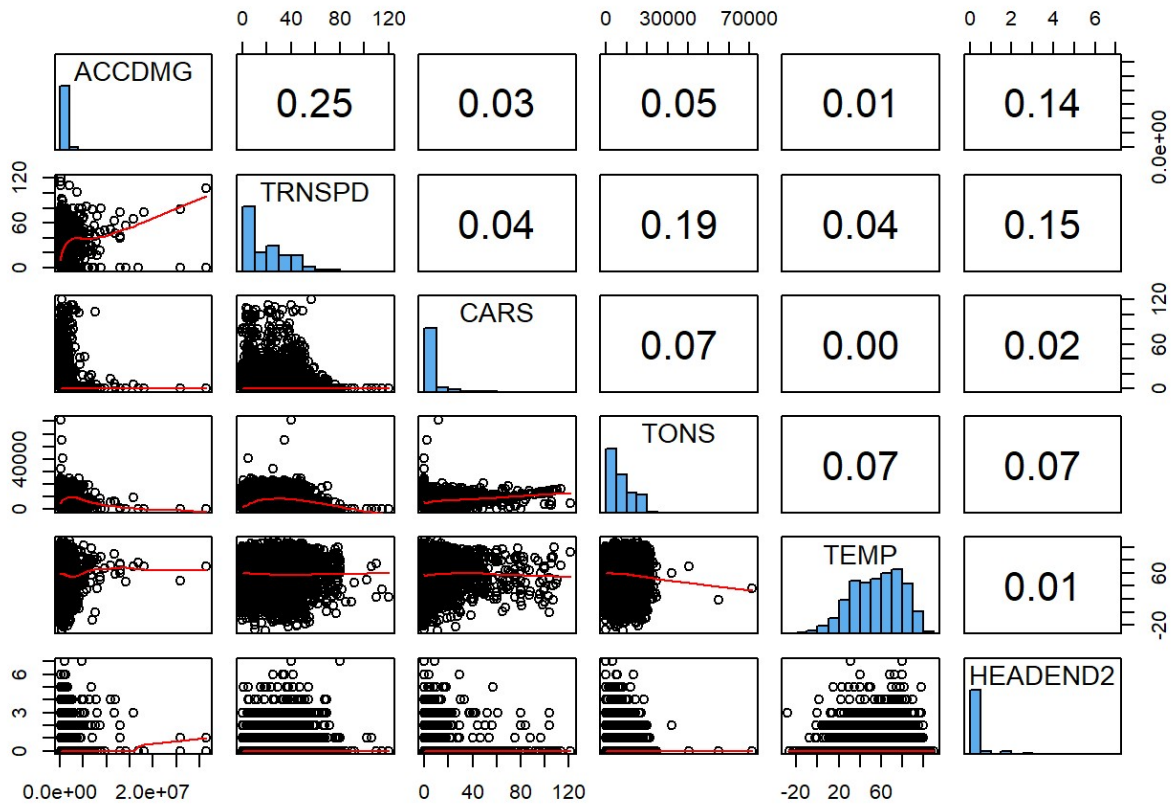
- CARS - number of cars carrying HAZMAT
- TRNSPD - train speed
- TONS - train weight in tonnage
- HEADEND2 - number of head end locomotives, derailed
- TEMP - temperature in Fahrenheit

Qualitative Variables

- TYPE - type of train accident. Derailments stood out in Project 1
- TYPEQ - type of Train
- Cause - cause of accident. Human factors stood out in Project 1
- METHOD - method of operation
- VISIBLTY - daylight period and specifically darkness
- WEATHER - weather conditions

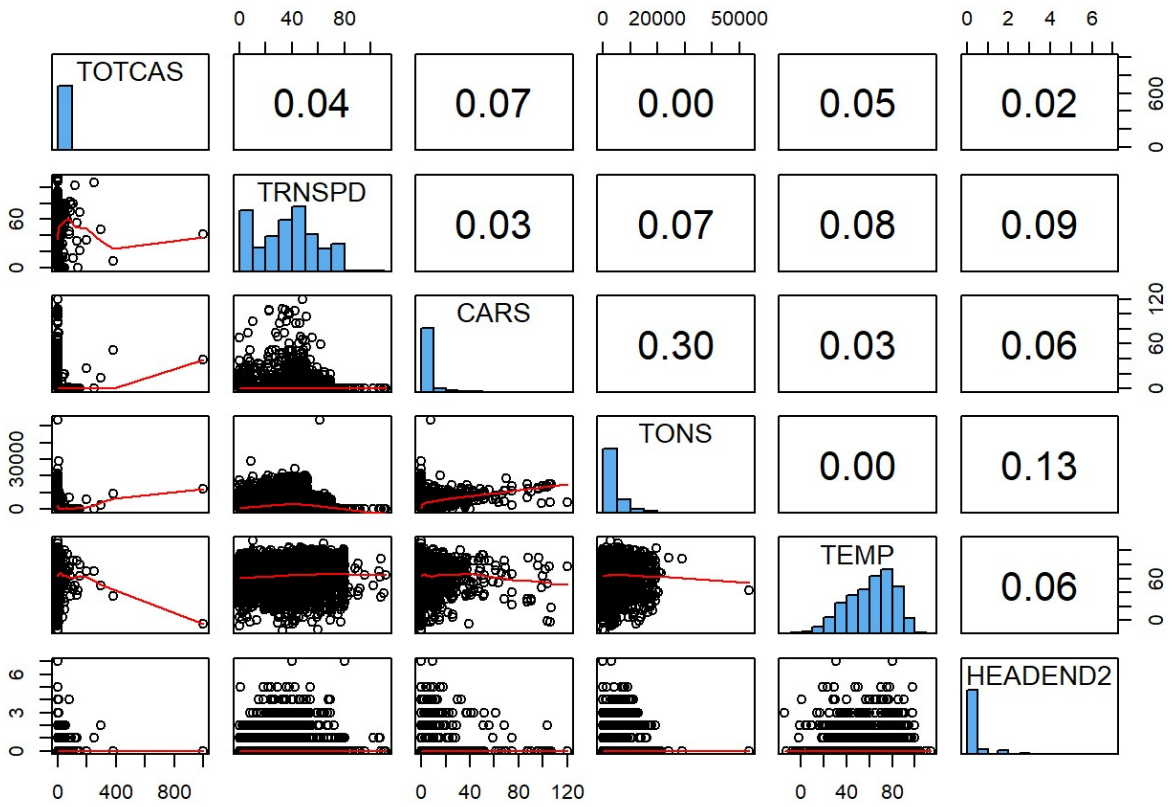
For the quantitative variables, we were concerned with multicollinearity and variables that lack significance. All quantitative variables will be centered (mean subtracted) to reduce multicollinearity if higher-order terms are later added to the model. Like in Project 1, we can look at scatterplot matrix to get an idea of correlation issues or that variables should be screened away. First, on the extreme ACCDMG dataset, TEMP appears to not have a significant relationship and we know it is unlikely to make accident more costly. Therefore, TEMP will not be included. Also, there doesn't appear to be any major correlation issues when looking at the pairwise comparisons.

```
uva.pairs(xdmg[,c("ACCDMG", "TRNSPD", "CARS", "TONS", "TEMP", "HEADEND2")])
```



A similar scatter plot can be done on the TOTCAS dataset. This time TEMP does have a slight trend, but TONS does not. For TOTCAS, we will screen away tons, but keep TEMP. After removing TONS, there are no major issues for pairwise correlations.

```
uva.pairs(xcas[,c("TOTCAS", "TRNSPD", "CARS", "TONS", "TEMP", "HEADEND2")])
```



For the qualitative variables, we need to reduce the number of bins per variable to focus the analysis on our hypotheses and to limit the number of parameters when using interaction terms. For the TYPE variable, we created a new variable to represent only derailments versus all other types. For Cause variable, we created a new variable to represent only accidents cause by human factors. We also tested other variables such as VISIBLTY, WEATHER, METHOD, and TYPEQ, but these variables either lack significance or their impact on the response was hard to explain.

```
# Centering Quantitative Variables to reduce multicollinearity with higher-order terms
# This subtracts each quantitative column by its mean. This is typically only done fo
r
# the higher order terms, but it was too difficult to separate. One downside is that
# it makes the predictors hard to transform, but that is a rare occurrence anyway.
xdmg$TRNSPD <- scale(xdmg$TRNSPD, scale = FALSE)
xdmg$TONS <- scale(xdmg$TONS, scale = FALSE)
xdmg$CARS <- scale(xdmg$CARS, scale = FALSE)
xdmg$HEADEND2 <- scale(xdmg$HEADEND2, scale = FALSE)

xcas$TRNSPD <- scale(xcas$TRNSPD, scale = FALSE)
xcas$TEMP <- scale(xcas$TEMP, scale = FALSE)
xcas$CARS <- scale(xcas$CARS, scale = FALSE)
xcas$HEADEND2 <- scale(xcas$HEADEND2, scale = FALSE)

# Create derailment categorical variable
xdmg_Derail <- rep(0, nrow(xdmg))
xdmg_Derail[which(xdmg$TYPE == "Derailment")] <- 1
xdmg_Derail <- as.factor(xdmg_Derail)

xcas_Derail <- rep(0, nrow(xcas))
xcas_Derail[which(xcas$TYPE == "Derailment")] <- 1
xcas_Derail <- as.factor(xcas_Derail)

# Create human factors categorical variable
xdmg_Human <- rep(0, nrow(xdmg))
xdmg_Human[which(xdmg$Cause == "Human Factors")] <- 1
xdmg_Human <- as.factor(xdmg_Human)

xcas_Human <- rep(0, nrow(xcas))
xcas_Human[which(xcas$Cause == "Human Factors")] <- 1
xcas_Human <- as.factor(xcas_Human)
```

Part 1: ACCDMG Analysis

First Model

There are an infinite number of ways to create a linear model. Sometimes you can start with just main effects and work up towards higher-order terms. The problem with this is hard to identify significant interactions if they’re not in the model. It’s possible the the main effect is insignificant, but the interaction is not. Therefore, as long as there are no multicollinearity issues we can model all main effects and interactions at first and then determine was parameters can be removed. To assess multicollinearity we calculated Variance Inflation Factors. A VIF of 1 means that variable is perfectly orthognal. VIFs greater than 10 or even 5 are typically problematic.

```
xdmg.lm1 <- lm(ACCDMG~(TRNSPD + CARS + TONS +  HEADEND2 + xdmg_Derail + xdmg_Human) ^
2, data=xdmg)
print(vif(xdmg.lm1))
```

```
##                 TRNSPD                   CARS                   TONS
##               5.336650              13.253336              14.070942
##               HEADEND2            xdmg_Derail             xdmg_Human
##               6.211638               2.893378               5.160050
##            TRNSPD:CARS            TRNSPD:TONS        TRNSPD:HEADEND2
##               1.130752               1.391577               1.465202
##     TRNSPD:xdmg_Derail      TRNSPD:xdmg_Human              CARS:TONS
##               4.614460               1.704459               1.636421
##          CARS:HEADEND2       CARS:xdmg_Derail        CARS:xdmg_Human
##               1.084139              12.427981               1.673090
##          TONS:HEADEND2       TONS:xdmg_Derail        TONS:xdmg_Human
##               1.184546              12.045003               1.825387
##   HEADEND2:xdmg_Derail    HEADEND2:xdmg_Human xdmg_Derail:xdmg_Human
##               3.827328               2.027332               4.050128
```

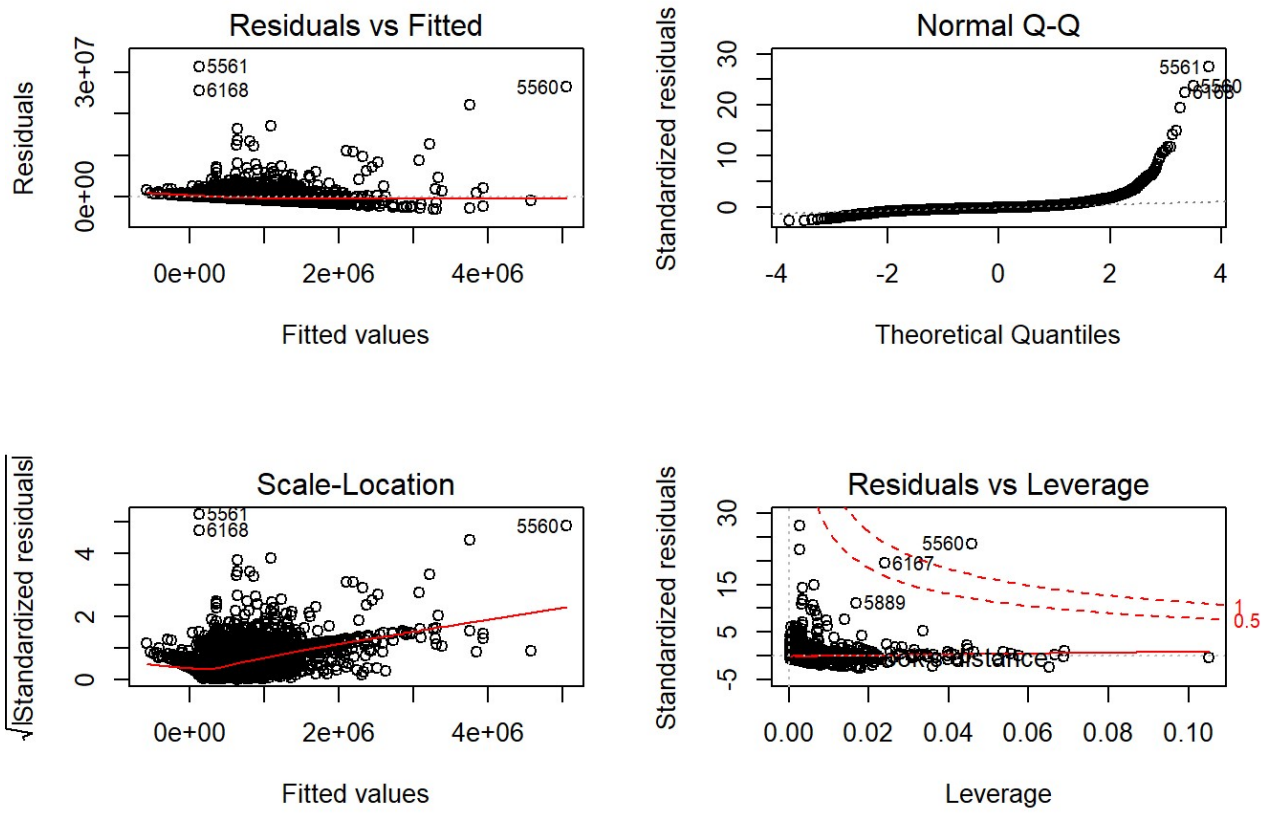
From this VIF report, there are four parameters with VIFs higher than 10. To alleviate this problem, the interaction terms can be removed first. If necessary, an entire effect can be removed. To alleviate some of the extreme multicollinearity, we removed HEADEND2, CARS:xmd_Derail, and TONS:xdmg_Derail. As you can see below, the model without these terms have much improved VIFs with the highest being less than 5.

```
## A couple of interactions and HEADEND2 have high VIFs, remove then and rerun
xdmg.lm2 <- lm(ACCDMG~(TRNSPD + CARS + TONS + xdmg_Derail + xdmg_Human) ^ 2
               - CARS:xdmg_Derail - TONS:xdmg_Derail, data=xdmg)
print(vif(xdmg.lm2))
```

```
##                 TRNSPD                   CARS                   TONS
##               4.894019               1.784800               1.453439
##            xdmg_Derail             xdmg_Human            TRNSPD:CARS
##               2.070941               4.540843               1.100552
##            TRNSPD:TONS     TRNSPD:xdmg_Derail      TRNSPD:xdmg_Human
##               1.342771               4.187501               1.509235
##              CARS:TONS        CARS:xdmg_Human        TONS:xdmg_Human
##               1.495730               1.401812               1.603937
## xdmg_Derail:xdmg_Human
##               3.782592
```

Now that we've addressed assumptions associated with the predictor variable we must do the same for the response variable. The diagnostic plots below reveal some issues with constant variance and normality of the residuals. The normal quantile plot shows that the response has very heavy-tail in the high ACCDMG direction.

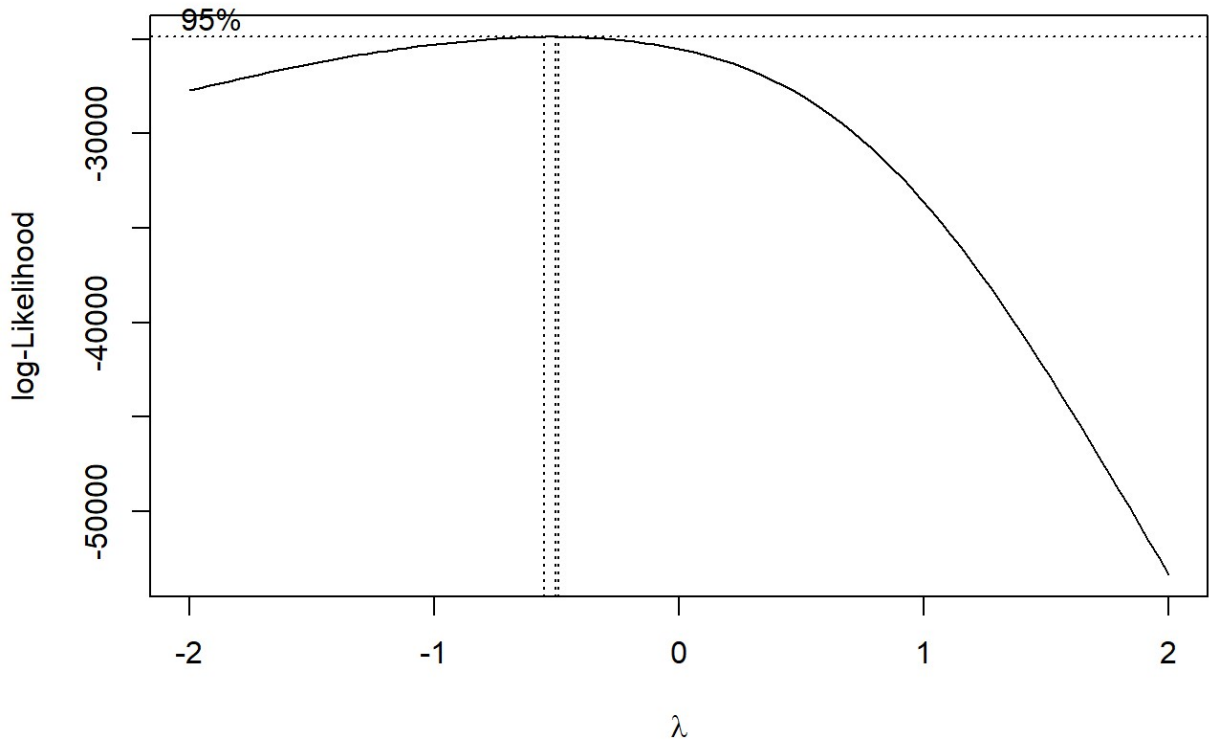
```
par(mfrow=c(2,2))
plot(xdmg.lm2, labels.id = NULL)
```




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

Transformation of the response can assist in achieving the residual normality assumption. The Box-Cox test can be applied to find an optimal lambda value. The optimal lambda in this case was -0.5 which is applied as an exponent to transform ACCDMG. However, this transformation will completely invert the response and make it very difficult to understand the model's output. Therefore, we selected a log transformation to improve normality, while also preserving most of the model's interpretability.

```
boxcox(xdmg.lm2, plotit=T, lambda=seq(-2,2,by=0.5))
```



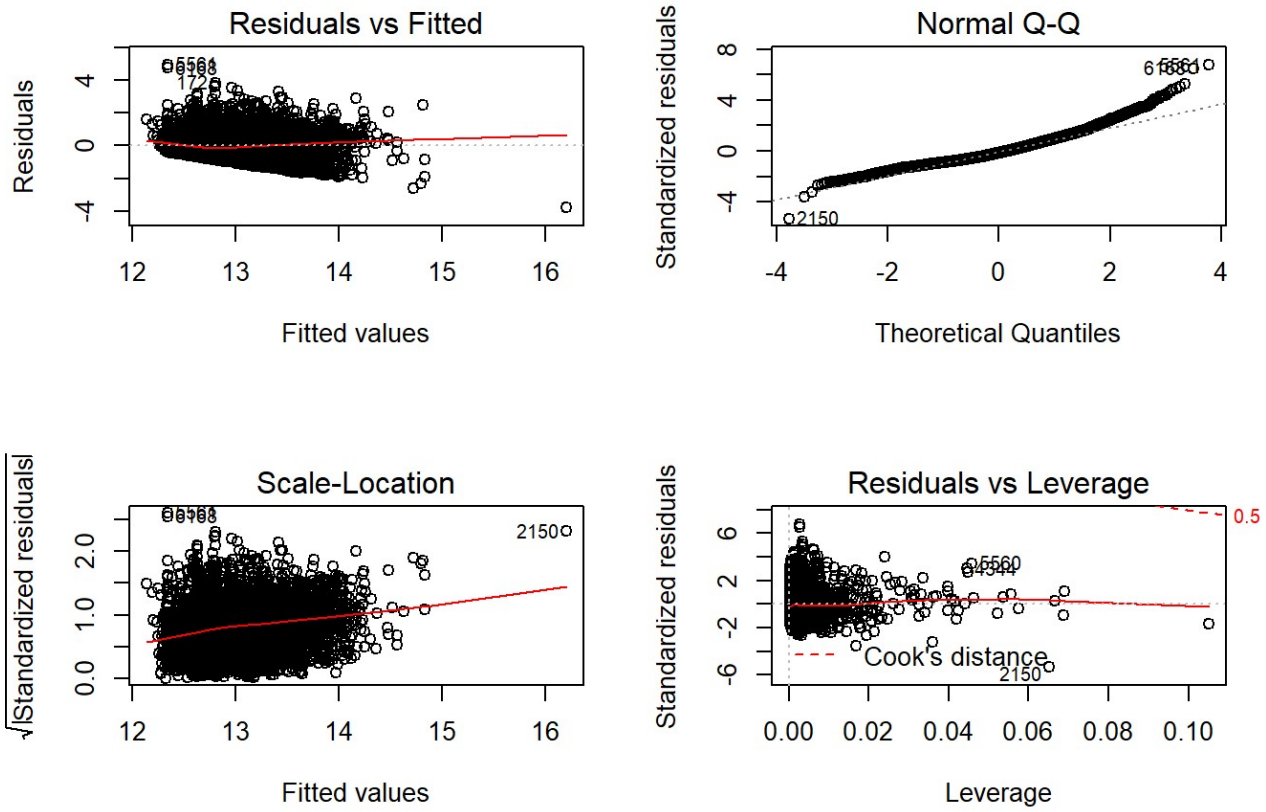

```
L_dmg<-boxcox(xdmg.lm2, plotit = F)$x[which.max(boxcox(xdmg.lm1, plotit = F)$y)]
print(L_dmg)
```

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

```
## Rerun model with all main effects and interactions besides ones already removed
xdmg.lm2.trans <- lm(log(ACCDMG)~(TRNSPD + CARS + TONS + xdmg_Derail + xdmg_Human) ^
2 - CARS:xdmg_Derail - TONS:xdmg_Derail, data=xdmg)
```

We can re-examine the diagnostic plots now after rerunning the model with the log transformed response. The normal quantile plot now has a much straigter line. The other three plots do not reveal any major violations of assumptions either.

```
par(mfrow=c(2,2))
plot(xdmg.lm2.trans, labels.id = NULL)
```




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

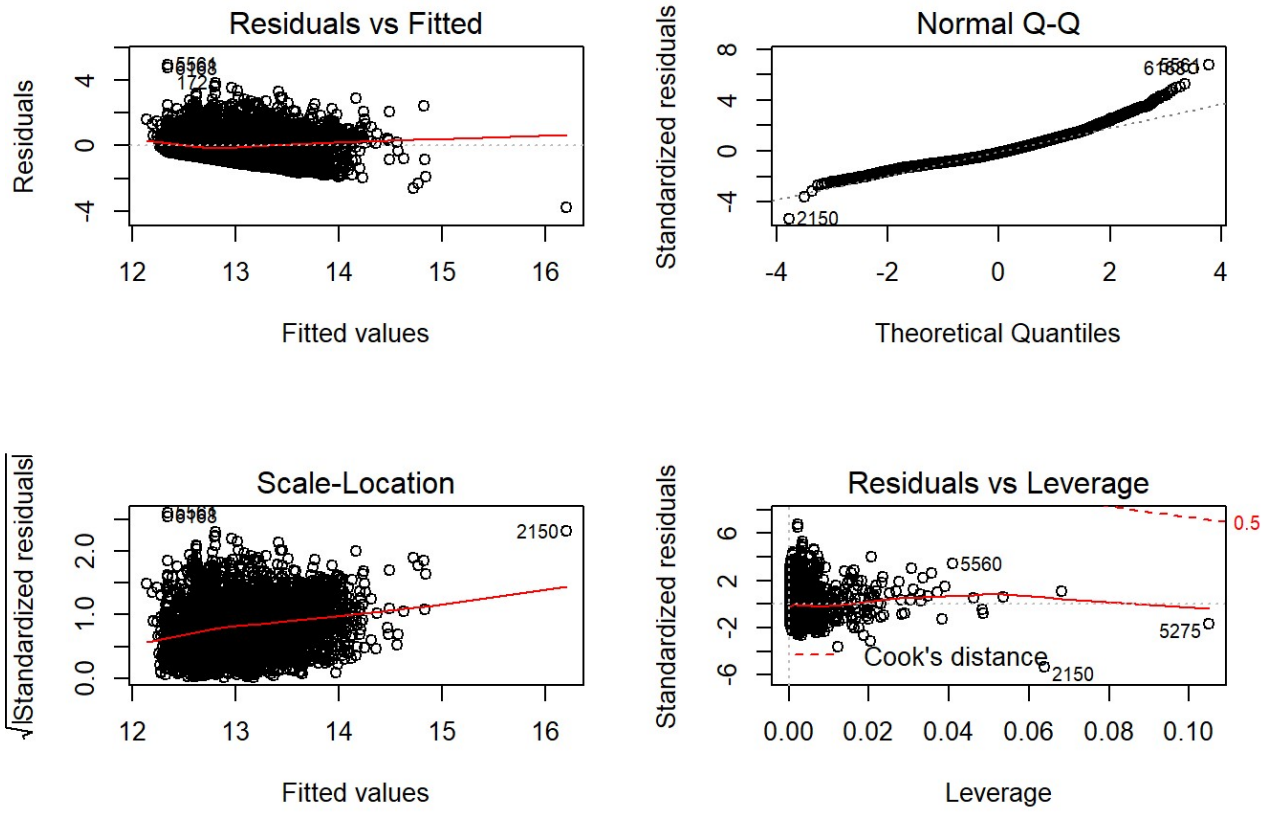
Now that we're comfortable with the model's assumptions, we can now begin to assess the impact of the model's predictors with the response. The summary below shows that this model currently exlains over 22% of the total variance of ACCDMG. There are a 7 terms that are significant at p-value of less than 0.001. There are also a number of terms that do not have a strong significance with the response. A stepwise regression can execute both backward and forward to subtract or add terms until it reaches a local minima.

```
summary(xdmg.lm2.trans)
```

```
##
## Call:
## lm(formula = log(ACCDMG) ~ (TRNSPD + CARS + TONS + xdmg_Derail +
##     xdmg_Human)^2 - CARS:xdmg_Derail - TONS:xdmg_Derail, data = xdmg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7701 -0.5150 -0.1158  0.4156  4.9240
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.297e+01  3.144e-02 412.355  < 2e-16 ***
## TRNSPD                      1.230e-02  1.138e-03  10.809  < 2e-16 ***
## CARS                        1.187e-03  9.783e-04   1.213   0.2251
## TONS                        1.601e-05  1.817e-06   8.809  < 2e-16 ***
## xdmg_Derail1               -1.688e-02  3.378e-02  -0.500   0.6174
## xdmg_Human1                 2.455e-01  4.798e-02   5.116 3.22e-07 ***
## TRNSPD:CARS                 9.847e-05  4.840e-05   2.034   0.0420 *
## TRNSPD:TONS                 1.306e-06  9.936e-08  13.145  < 2e-16 ***
## TRNSPD:xdmg_Derail1         7.810e-03  1.270e-03   6.151 8.18e-10 ***
## TRNSPD:xdmg_Human1          1.265e-02  1.514e-03   8.352  < 2e-16 ***
## CARS:TONS                   4.132e-07  1.689e-07   2.446   0.0145 *
## CARS:xdmg_Human1            8.273e-04  2.018e-03   0.410   0.6819
## TONS:xdmg_Human1           -1.062e-07  4.770e-06  -0.022   0.9822
## xdmg_Derail1:xdmg_Human1 -2.626e-01  5.505e-02  -4.771 1.88e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7296 on 6277 degrees of freedom
## Multiple R-squared:  0.2263, Adjusted R-squared:  0.2247
## F-statistic: 141.2 on 13 and 6277 DF,  p-value: < 2.2e-16
```

Before this reduced model can be fully accepted, the diagnostic plots should be reviewed again. There appears to not be much of a change from before which is expected since only insignificant terms were removed.

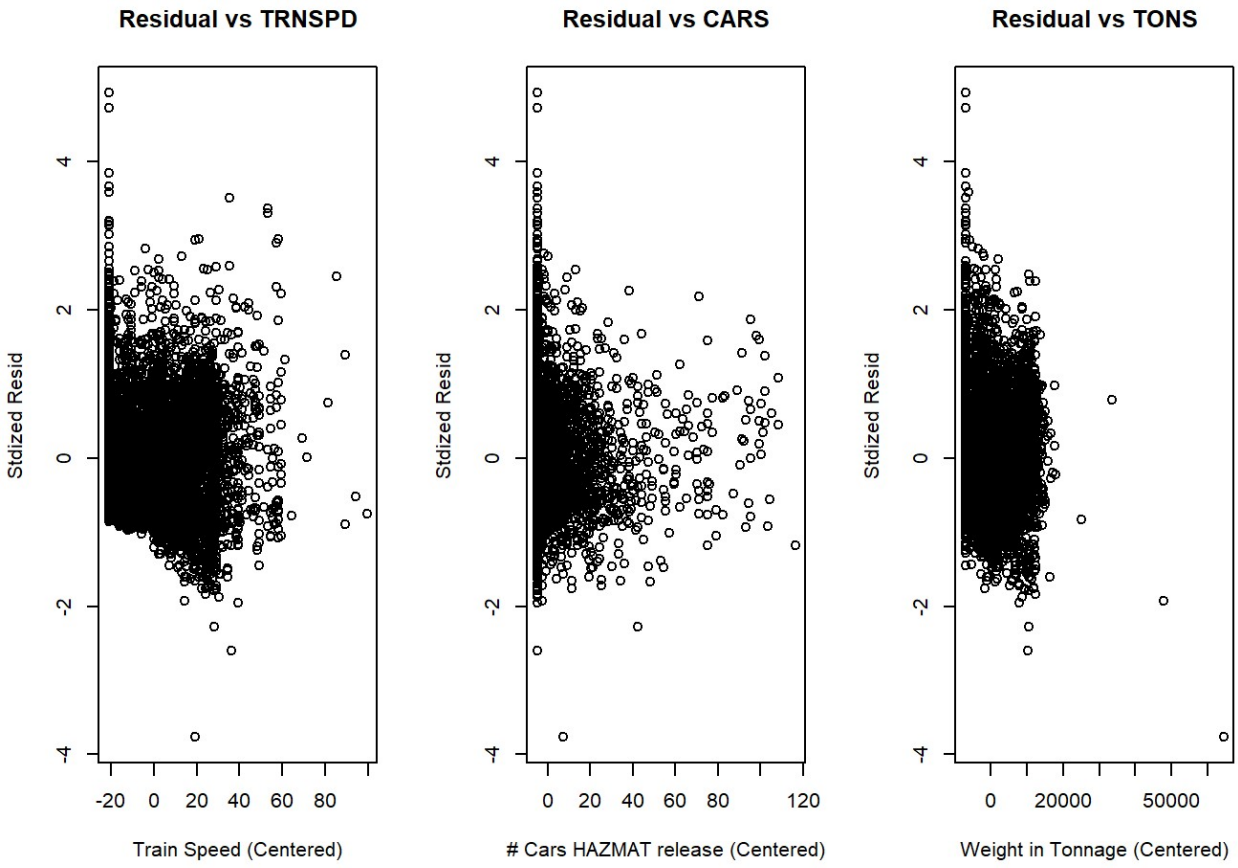
```
par(mfrow=c(2,2))
plot(xdmg.lm2.step, labels.id = NULL)
```




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

One last possibilty for model inadequacy is due to lack of fit because there are missing parameters. This could be due to missing variables all together or missing higher-order terms. A lack of fit test is one of way of doing so, but we haven't figured out how to do that in R yet. To determine if any current parameters require a higher-order term we can plot each quantitative variable by the model's residual. Those three plots are shown below. There did not appear to be any issues with constant variance amongst the predictors, so no higher-order terms or transformations are required.

```
par(mfrow=c(1,3))
plot(xdmg$TRNSPD, resid(xdmg.lm2.step), main = "Residual vs TRNSPD", ylab = "Stdized R
esid", xlab = "Train Speed (Centered)")
plot(xdmg$CARS, resid(xdmg.lm2.step), main = "Residual vs CARS", ylab = "Stdized Resi
d", xlab = "# Cars HAZMAT release (Centered)")
plot(xdmg$TONS, resid(xdmg.lm2.step), main = "Residual vs TONS", ylab = "Stdized Resi
d", xlab = "Weight in Tonnage (Centered)")
```




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

Finally, the reduced model summary after stepwise regression is shown below. The model's R^2, remained virtually the same after the insignificant parameter reduction. The current takeaways are that TRNSPD, TONS, Human Factors main effects all significantly increase accident damage costs. For interactions, train speed combined with TONS, Derailments, or Human Factors all significantly increase accident damage.

```
summary(xdmg.lm2.step)
```

```
##
## Call:
## lm(formula = log(ACCDMG) ~ TRNSPD + CARS + TONS + xdmg_Derail +
##     xdmg_Human + TRNSPD:CARS + TRNSPD:TONS + TRNSPD:xdmg_Derail +
##     TRNSPD:xdmg_Human + CARS:TONS + xdmg_Derail:xdmg_Human, data = xdmg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7660 -0.5159 -0.1156  0.4152  4.9221
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.297e+01  3.126e-02 414.729  < 2e-16 ***
## TRNSPD                     1.229e-02  1.133e-03  10.854  < 2e-16 ***
## CARS                       1.375e-03  8.658e-04   1.588   0.1123
## TONS                       1.600e-05  1.682e-06   9.516  < 2e-16 ***
## xdmg_Derail1              -1.741e-02  3.357e-02  -0.519   0.6041
## xdmg_Human1                2.455e-01  4.566e-02   5.376 7.90e-08 ***
## TRNSPD:CARS                9.385e-05  4.707e-05   1.994   0.0462 *
## TRNSPD:TONS                1.307e-06  9.777e-08  13.371  < 2e-16 ***
## TRNSPD:xdmg_Derail1        7.805e-03  1.268e-03   6.156 7.94e-10 ***
## TRNSPD:xdmg_Human1         1.268e-02  1.489e-03   8.512  < 2e-16 ***
## CARS:TONS                  4.010e-07  1.662e-07   2.413   0.0159 *
## xdmg_Derail1:xdmg_Human1 -2.621e-01  5.406e-02  -4.848 1.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7295 on 6279 degrees of freedom
## Multiple R-squared:  0.2263, Adjusted R-squared:  0.2249
## F-statistic: 166.9 on 11 and 6279 DF,  p-value: < 2.2e-16
```

Second Model

Compare Adjusted R^2 of two models and discuss anything else like diagnostics.

Model Comparison

Make sure to address the following steps: a. Feature and model selection techniques b. Treatment of categorical variables c. Model assessment d. Model diagnostics e. Model adjustment based on analytical and graphical diagnostics

Part 2: Casualties Analysis

First Model

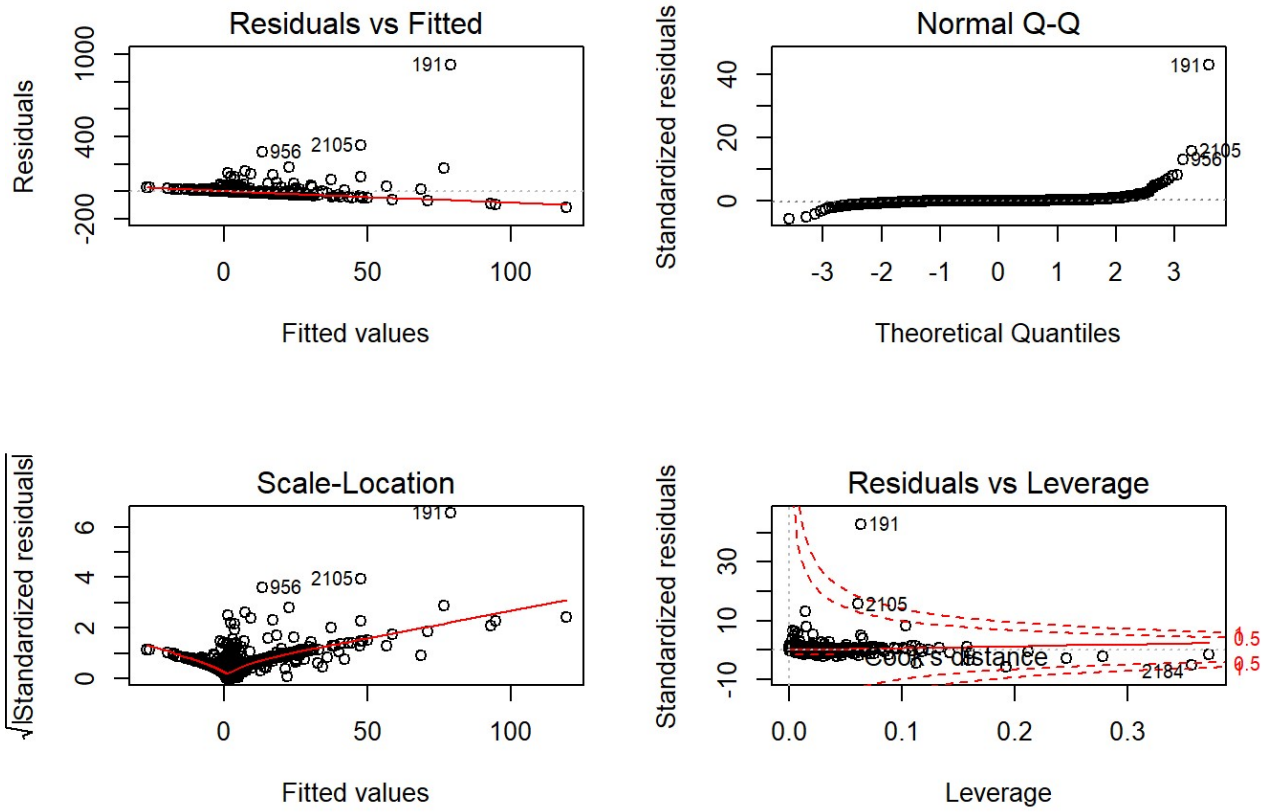
The first model for total casualties was obtained using a similar strategy to the firts model for ACCDMG. First, the VIFs of the model must be checked to ensure there is not problematic multicollinearity between the predictor variables. This time, there were not any major issues with multicollinearity, so no terms were removed.

```
xcas.lm1 <- lm((TOTCAS)~(TRNSPD + CARS + TEMP + HEADEND2 + xcas_Derail + xcas_Human)
^ 2, data=xcas)
print(vif(xcas.lm1))
```

```
##                 TRNSPD                   CARS                   TEMP
##               1.968432               1.459198               1.547652
##               HEADEND2            xcas_Derail             xcas_Human
##               2.895357               2.418597               3.924153
##            TRNSPD:CARS            TRNSPD:TEMP        TRNSPD:HEADEND2
##               1.364107               1.355689               1.285255
##     TRNSPD:xcas_Derail      TRNSPD:xcas_Human              CARS:TEMP
##               2.098119               3.624052               1.194726
##          CARS:HEADEND2       CARS:xcas_Derail        CARS:xcas_Human
##               1.386654               1.536010               1.628982
##          TEMP:HEADEND2       TEMP:xcas_Derail        TEMP:xcas_Human
##               1.220287               1.447603               1.711757
##   HEADEND2:xcas_Derail    HEADEND2:xcas_Human xcas_Derail:xcas_Human
##               2.032195               2.363313               1.932226
```

Next, the model's diagnostic plots must be reviewed to ensure the assumptions of a linear regression model are met. Constant variance and normality do not look overly problematic. The major problem lies in the influence points contained in this model. Point #191 is over 40 standard deviations from the mean which for a normal distribution is unfathomably rare.

```
par(mfrow=c(2,2))
plot(xcas.lm1, labels.id = NULL)
```

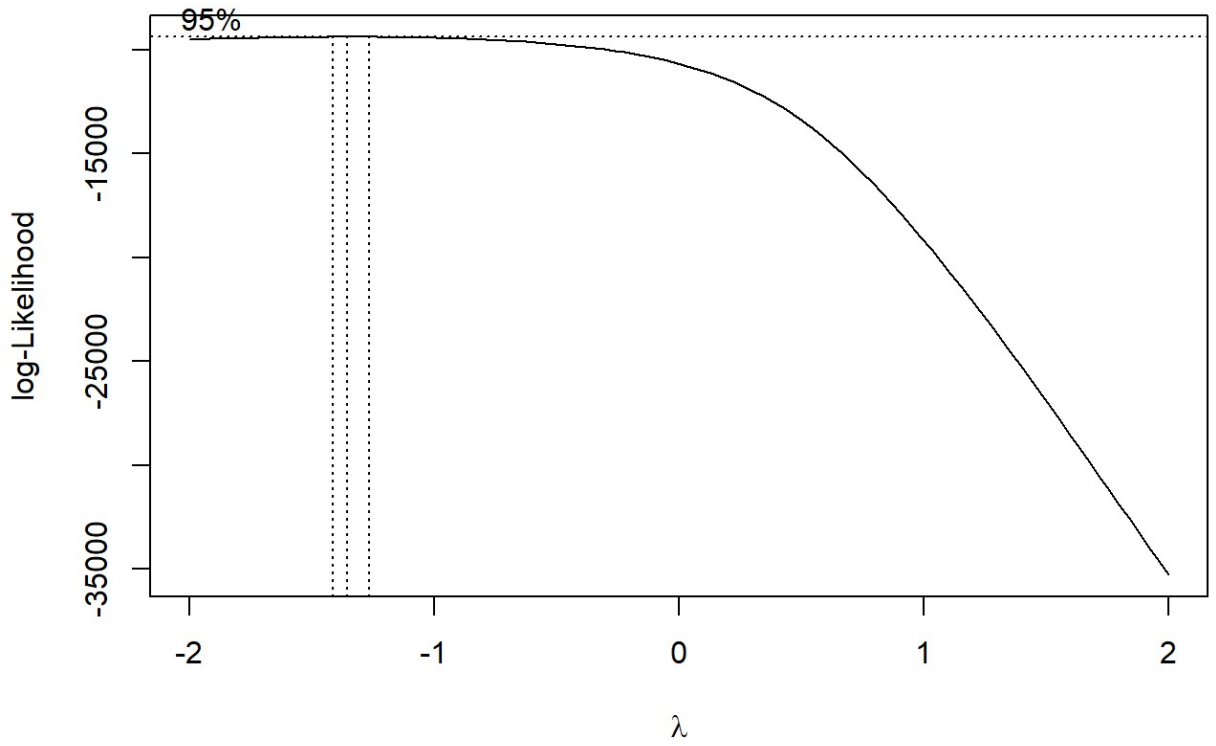



```
par(mfrow=c(1,1))
```

Sometimes influence points are present due to errors or data that is not appropriate. To check the validity of some of these potential points, we must read the narrative entries in these rows. After reading these narratives, it was clear that these data points are valid and should not be discarded for the sake of improving the model's adequacy.

Response transformations can be tried to reduce skew and hopefully eliminate the influence points. The Box-Cox test calculated an optimal lambda of -1.3.

```
boxcox(xcas.lm1, plotit=T, lambda=seq(-2,2,by=0.5))
```



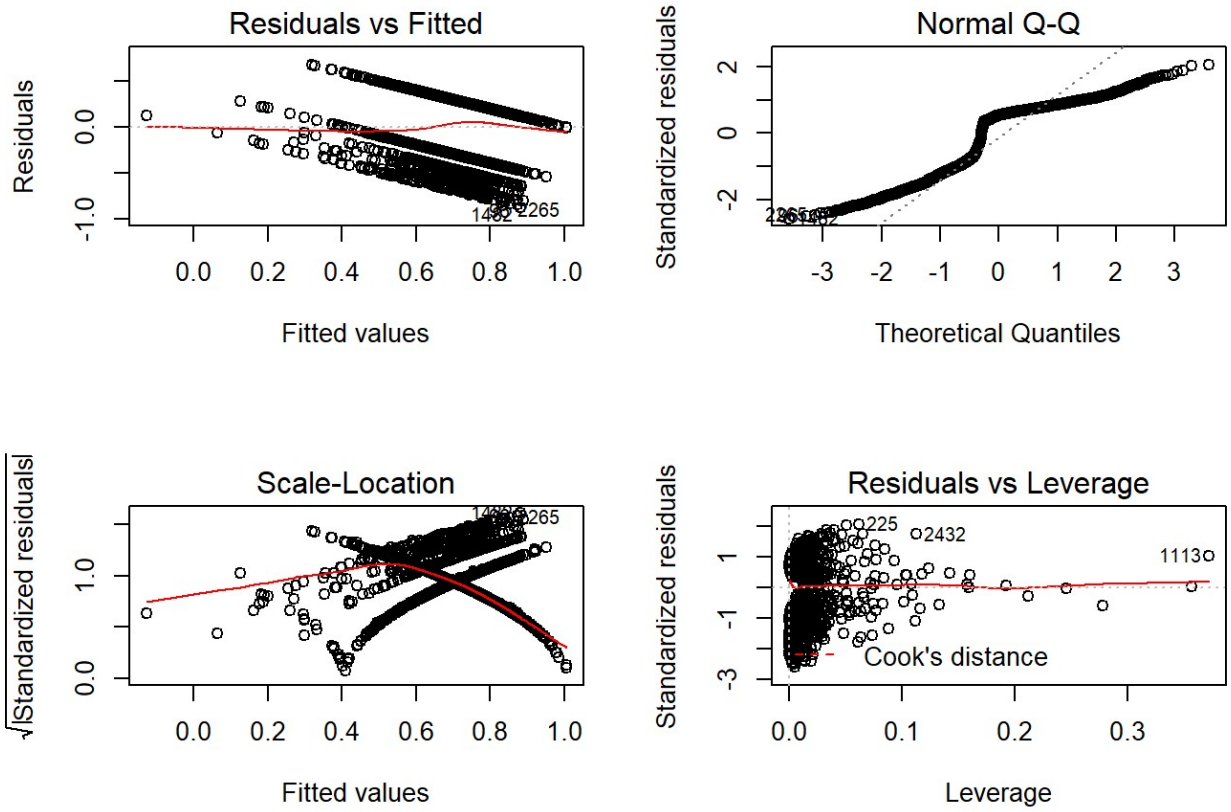

```
L_xcas <-boxcox(xcas.lm1, plotit = F)$x[which.max(boxcox(xcas.lm1, plotit = F)$y)]
print(L_xcas)
```

```
## [1] -1.3
```

The model diagnostic plots for the TOTCAS model with transformed response are shown below. While the transformation did take care of the influence point issue, the other assumption tests are much worse than before. The normality assumption is severely violated. Due to such as poor performance by the transformations, we decided to stick with the original modle even with the influence point issues. With data sets that include this many outliers or influence points, it is sometimes appropriate to use a Robust Regression technique instead of Ordinary Least Squares (OLS). Robust Regression enables another distribution to be fit to the response that can have much wider tails than the normal distribution. This outside of the current scope of the class, so we will continue to utilize OLS.

```
xcas.lm1.trans <- lm(TOTCAS^L_xcas~(TRNSPD + CARS + TEMP + HEADEND2 + xcas_Derail + xc
as_Human) ^ 2, data=xcas)

par(mfrow=c(2,2))
plot(xcas.lm1.trans, labels.id = NULL)
```




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

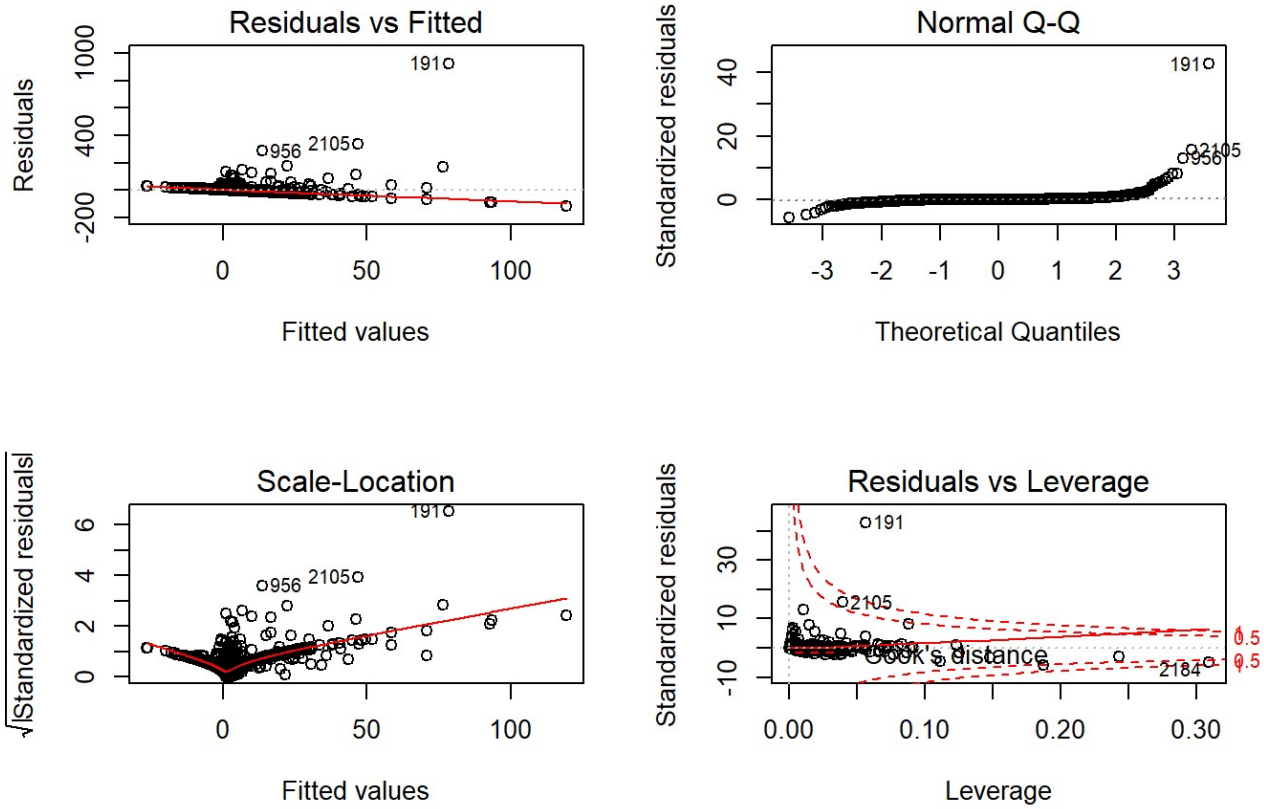
Below is the model summary for the original (non-transformed) TOTCAS model. This model is only able to explain just over 9% of the variance of total casualties which is much lower than the ACCDMG models.

```
summary(xcas.lm1)
```

```
##
## Call:
## lm(formula = (TOTCAS) ~ (TRNSPD + CARS + TEMP + HEADEND2 + xcas_Derail +
##     xcas_Human)^2, data = xcas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -117.34   -1.84   -0.87    0.36  922.10
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.285e+00  5.405e-01   4.227 2.44e-05 ***
## TRNSPD                     2.079e-02  2.532e-02   0.821  0.41165
## CARS                      -3.460e-02  4.550e-02  -0.760  0.44711
## TEMP                       1.144e-03  2.453e-02   0.047  0.96279
## HEADEND2                   6.527e-01  8.788e-01   0.743  0.45768
## xcas_Derail1               1.330e+01  1.874e+00   7.101 1.55e-12 ***
## xcas_Human1                1.192e+01  2.099e+00   5.680 1.48e-08 ***
## TRNSPD:CARS               -3.707e-05  3.004e-03  -0.012  0.99015
## TRNSPD:TEMP                6.049e-04  1.040e-03   0.581  0.56097
## TRNSPD:HEADEND2           -1.163e-02  3.162e-02  -0.368  0.71313
## TRNSPD:xcas_Derail1        4.574e-01  6.899e-02   6.631 3.98e-11 ***
## TRNSPD:xcas_Human1         3.711e-01  6.988e-02   5.310 1.18e-07 ***
## CARS:TEMP                 -8.165e-03  1.532e-03  -5.331 1.05e-07 ***
## CARS:HEADEND2             -1.558e-01  5.212e-02  -2.989  0.00283 **
## CARS:xcas_Derail1          4.853e-01  9.839e-02   4.933 8.58e-07 ***
## CARS:xcas_Human1          -4.896e-02  1.506e-01  -0.325  0.74518
## TEMP:HEADEND2              5.593e-02  2.498e-02   2.239  0.02523 *
## TEMP:xcas_Derail1         -3.272e-01  5.991e-02  -5.461 5.15e-08 ***
## TEMP:xcas_Human1           7.476e-02  6.003e-02   1.245  0.21310
## HEADEND2:xcas_Derail1     -5.448e+00  1.216e+00  -4.480 7.77e-06 ***
## HEADEND2:xcas_Human1      -1.478e-01  1.251e+00  -0.118  0.90596
## xcas_Derail1:xcas_Human1 -2.234e+00  2.925e+00  -0.764  0.44509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.2 on 2873 degrees of freedom
## Multiple R-squared:  0.1015, Adjusted R-squared:  0.09492
## F-statistic: 15.45 on 21 and 2873 DF,  p-value: < 2.2e-16
```

Like before, we can apply stepwise regression to subtract and add terms until a local minima for AIC is found. The diagnostic plots should be reassessed to affirm that the model's assumptions are met. The results are similar to before. The normality distribution is skewed and there is at least one data point that exceeds 1.0 for Cook's distance. Since the data points sre valid and transformations did not improve the situation, we will continue to utilize this model.

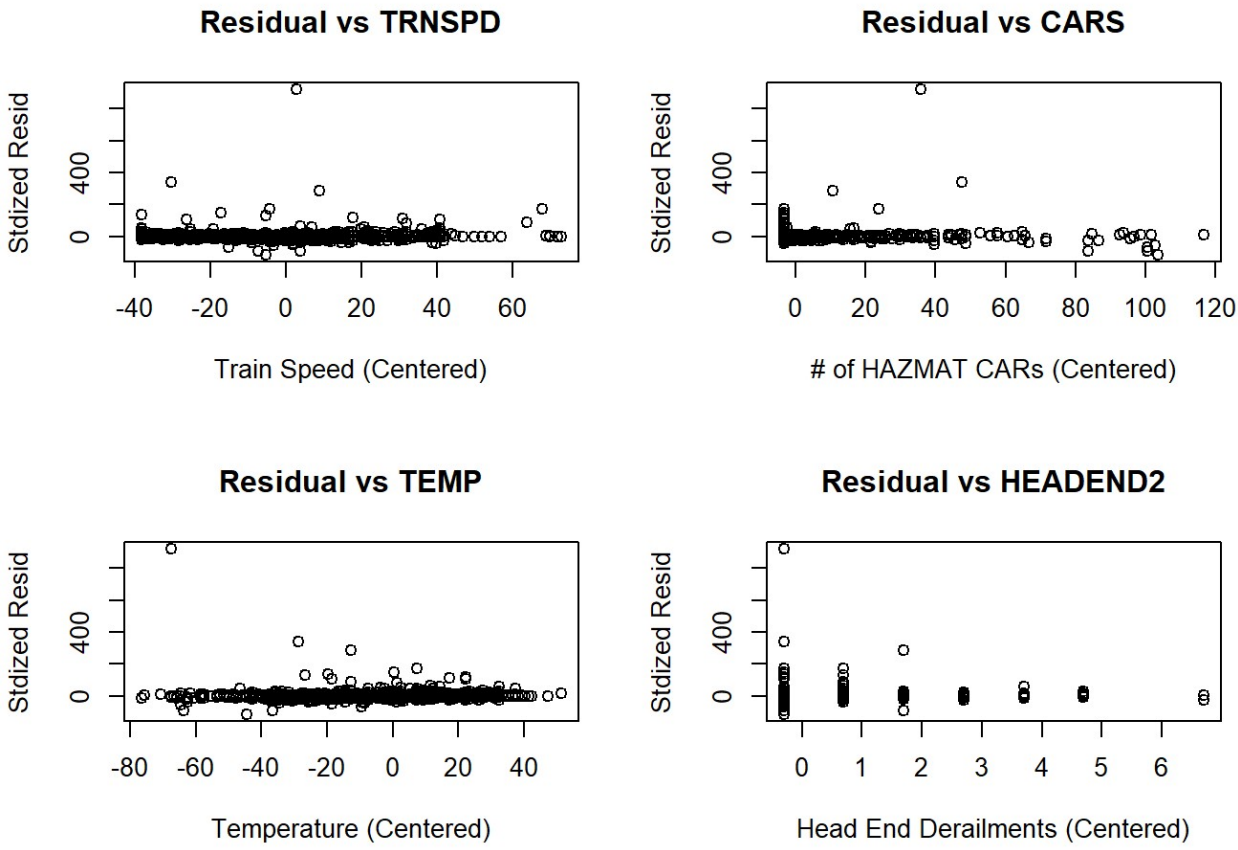
```
par(mfrow=c(2,2))
plot(xcas.lm1.step, labels.id = NULL)
```




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

Next, residual versus predictors plots can help determine if transformations or higher-order terms are required on the predictor variables. Beside the otuliers point, no alarming patterns appear in these charts.

```
par(mfrow=c(2,2))
plot(xcas$TRNSPD, resid(xcas.lm1), main = "Residual vs TRNSPD", ylab = "Stdized Resi
d", xlab = "Train Speed (Centered)")
plot(xcas$CARS, resid(xcas.lm1), main = "Residual vs CARS", ylab = "Stdized Resid", xl
ab = "# of HAZMAT CARs (Centered)")
plot(xcas$TEMP, resid(xcas.lm1), main = "Residual vs TEMP", ylab = "Stdized Resid", xl
ab = "Temperature (Centered)")
plot(xcas$HEADEND2, resid(xcas.lm1), main = "Residual vs HEADEND2", ylab = "Stdized Re
sid", xlab = "Head End Derailments (Centered)")
```




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

Finally, the model's summary is shown below. Interestingly, none of the quantitative variables main effects are significant. Categorical variables main effects for Human Factors and Derailments do show a significant relationship with TOTCAS. For the interaction terms, only 4 out of the 8 terms significantly increased TOTCAS. Like for ACCDMG, TRNSPD combined with accidents caused Human Factors or Derailments will significantly increase the TOTCAS severity metric. The other interaction of note is CARS:xcas_Derail. If there are high number of cars carrying HAZMAt and the train derails, there could be a greater chance HAZMAT spillage. The HAZMAT spillage will require that train crew and passangers receive medical attention to assess their exposure to the HAZMAT.

```
summary(xcas.lm1.step)
```

```
##
## Call:
## lm(formula = (TOTCAS) ~ TRNSPD + CARS + TEMP + HEADEND2 + xcas_Derail +
##     xcas_Human + TRNSPD:xcas_Derail + TRNSPD:xcas_Human + CARS:TEMP +
##     CARS:HEADEND2 + CARS:xcas_Derail + TEMP:HEADEND2 + TEMP:xcas_Derail +
##     HEADEND2:xcas_Derail, data = xcas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -117.41   -1.87   -0.96    0.28  922.37
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.391682   0.514543   4.648 3.50e-06 ***
## TRNSPD                  0.018648   0.023336   0.799 0.424290
## CARS                   -0.040973   0.043148  -0.950 0.342407
## TEMP                    0.014511   0.021653   0.670 0.502795
## HEADEND2                0.682202   0.715586   0.953 0.340494
## xcas_Derail1           12.824295   1.714982   7.478 9.97e-14 ***
## xcas_Human1            10.810575   1.780090   6.073 1.42e-09 ***
## TRNSPD:xcas_Derail1     0.465092   0.065603   7.089 1.69e-12 ***
## TRNSPD:xcas_Human1      0.351797   0.064031   5.494 4.27e-08 ***
## CARS:TEMP              -0.008329   0.001511  -5.512 3.87e-08 ***
## CARS:HEADEND2          -0.161013   0.045934  -3.505 0.000463 ***
## CARS:xcas_Derail1       0.491828   0.095845   5.131 3.07e-07 ***
## TEMP:HEADEND2           0.063424   0.023878   2.656 0.007948 **
## TEMP:xcas_Derail1      -0.331387   0.057739  -5.739 1.05e-08 ***
## HEADEND2:xcas_Derail1 -5.426858   1.189504  -4.562 5.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.19 on 2880 degrees of freedom
## Multiple R-squared:  0.1007, Adjusted R-squared:  0.09632
## F-statistic: 23.03 on 14 and 2880 DF,  p-value: < 2.2e-16
```

Second Model

Repeat analysis for TOTCAS second model

Model Comparison

Compare Adjusted R^2 of two models and discuss anything else like diagnostics.

Part 3: Evidence and Recommnedation to FRA (20 points)

Quick summary can go here

Evaluating the Hypotheses

ACCDMG

1. Accidents caused by Human Factors at high train speeds significantly increase total accident damage cost

- At over 99% confidence, we reject the null hypothesis that Human factors and train speed have no significant influence on ACCDMG. Human factors at high train speeds are shown to significantly increase total accident damage cost.

2. Derailment accidents that occur at high train speeds significantly increase total accident damage cost

- At over 99% confidence, we reject the null hypothesis that derailmetns and train speed have no significant influence on ACCDMG. Derailments at high train speeds are shown to significantly increase total accident damage cost.

TOTCAS

1. Higher train speeds and accidents caused by human factors cause a significant increase in the number of casualties

- At over 99% confidence, we reject the null hypothesis that Human factors and train speed have no significant influence on the TOTCAS. Human factors at high train speeds are shown to significantly increase total casualties.

2. Derailment accidents on trains with a high number of cars containing HAZMAT will cause a significant increase in the number of casualties

- At over 99% confidence, we reject the null hypothesis that cars carrying HAZMAT and derailments have no significant influence on the TOTCAS. Derailments with a high number of cars carry HAZMAT significantly increases total casualties.

Recommendations

For both ACCDMG and TOTCAS models, it was found that human factors can have a great impact to accident severity when they are operating at high speeds. The United States train system is already notoriously slow compared to the rest of the world, so we would not recommend reducing speed limits. Our recommendation would be to add cyber-physical elements to train operating systems that can autonomously control speeds or inform the operator when things seem awry. One example of these is positive train control which will slow down or stop the train if it is going at an excessive speed in a

certain location. Positive train control also alerts the operator when speed limit changes are incoming or there are poor track conditions. Much like back-up cameras now installed on every car, we recommend that every train be outfitted with positive train control.

It is also possible to one day remove train operators all together. Positive train control would fall under Automated Train Protection, but there are even higher stages of autonomy. Automated Train Operation can automate features like changing tracks, starting, and stopping. Driverless Train Operation means there are no drivers, but there is still humans available in case of emergency. Finally, trains could have full Unattened Train Operation. While train conductors and engineers may not want their current jobs automated, these automation possibilities could to lead to more efficient and safer railroad transportation.

References

1. “Positive Train Control.” UP: Positive Train Control, www.up.com/media/media_kit/ptc/about-ptc/.
2. Sankaran, Vishwam. “Fully Autonomous Trains Are Better Suited for Moving Ores than People.” The Next Web, 13 July 2018, thenextweb.com/artificial-intelligence/2018/07/13/fully-autonomous-trains-are-better-suited-for-moving-ores-than-people/.