Project 1 Example

Load the data

source("AccidentInput.R")  
  
# load the data  
acts <- file.inputl(traindir)  
  
# combine all the data into one data frame  
totacts <- combine.data(acts)

Source files and load libraries needed for analysis

source("SPM\_Panel.R")  
source("PCAplots.R")  
library(lattice)  
library(ggplot2)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(devtools) # for ggbiplot

## Loading required package: usethis

library(car)

## Loading required package: carData

library(here)

## here() starts at /Volumes/GoogleDrive/My Drive/UVA/Courses/LSM/Fall2021/Source

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggpubr)  
library(ggfortify)

##   
## Attaching package: 'ggfortify'

## The following object is masked \_by\_ '.GlobalEnv':  
##   
## ggbiplot

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(lindia)  
library(olsrr)

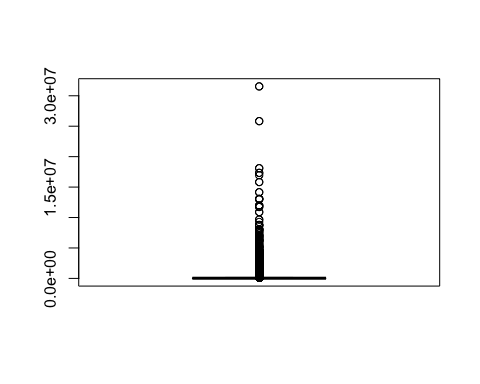
##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:MASS':  
##   
## cement

## The following object is masked from 'package:datasets':  
##   
## rivers

Extreme accident damage and remove duplicates

# Create a data frame with just the extreme ACCDMG accidents and remove duplicates  
dmgbox <-boxplot(totacts$ACCDMG)



xdmg <- totacts[totacts$ACCDMG > dmgbox$stats[5],]  
  
## Remove 9/11  
xdmg <- xdmg[-185,]  
  
#remove the duplicates  
xdmgnd <- xdmg[!(duplicated(xdmg[, c("INCDTNO", "YEAR", "MONTH", "DAY", "TIMEHR", "TIMEMIN")])),]

Obtain an understanding of the contributors to accident severity in your extreme accidents data using the multivariate visualization techniques discussed in class. This analysis should explore both quantitative and categorical variables to describe the severe accidents.

First we will setup our categorical variables to be able to explore our data better.

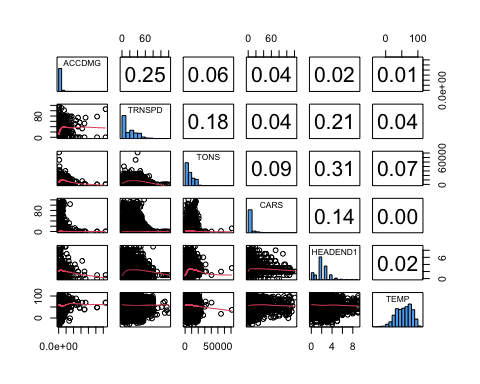
# rename TYPE and TYPEQ from a number to a character description  
xdmgnd$Type <- factor(xdmgnd$TYPE, labels = c("Derailment", "HeadOn", "Rearend", "Side", "Raking", "BrokenTrain", "Hwy-Rail", "GradeX", "Obstruction", "Explosive", "Fire","Other","SeeNarrative"))  
  
xdmgnd$TYPEQ <- as.numeric(xdmgnd$TYPEQ)

## Warning: NAs introduced by coercion

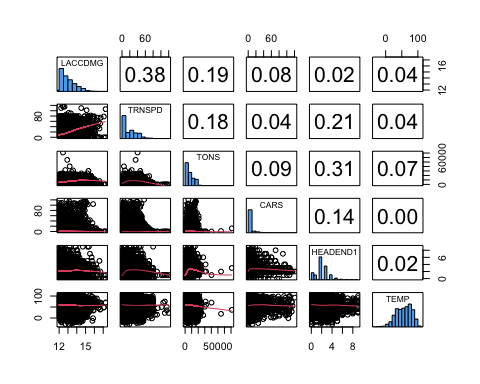
xdmgnd$TYPEQ <- factor(xdmgnd$TYPEQ, labels = c("Freight", "Passenger", "Commuter", "Work", "Single", "CutofCars", "Yard", "Light", "Maint"))#, "Spec"))  
  
# Create a new factor variable called Cause that uses labels for cause.  
xdmgnd$Cause <- rep(NA, nrow(xdmgnd))  
xdmgnd$Cause[which(substr(xdmgnd$CAUSE, 1, 1) == "M")] <- "M"  
xdmgnd$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)

To understand the possible correlations between quantitative variables and accident damage and casualties for the extreme accidents, we use scatter plot matrices. A scatter plot matrix with the variables, train speed (TRNSPD), weight (TONS), the number of cars in the train carrying hazmat (CARS), and the number of hours a conductor is on duty versus each other and total accident damage (ACCDMG).

uva.pairs(xdmgnd[,c("ACCDMG", "TRNSPD", "TONS", "CARS", "HEADEND1", "TEMP")])

 The skewness observed in SPM with ACCDMG and quantitative variables suggests the log transformation should be used in order to better assess linear correlations and view relationships in scatter plots. Thus, we replot the scatter plots with the log transform of ACCDMG.

xdmgnd$LACCDMG <- log(xdmgnd$ACCDMG)  
uva.pairs(xdmgnd[,c("LACCDMG", "TRNSPD", "TONS", "CARS","HEADEND1", "TEMP")])

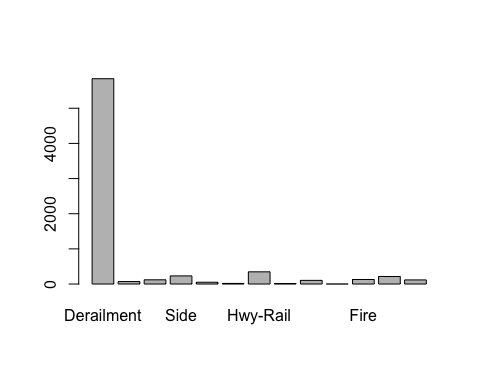
 1. Quantitative / Qualitative variable selection for linear models.

The scatter plots show that the variable most correlated with accident damage is train speed followed by tons. There is also some correlation between speed and the weight of the train. The number of head end locomotives and Temperature have the lowest correlation. Since we have a relatively low number of quantitative variables and lower correlations are relatively similar, we will add all of these quantitative predictors to our model and see if they have any value to an ANCOVA analysis.

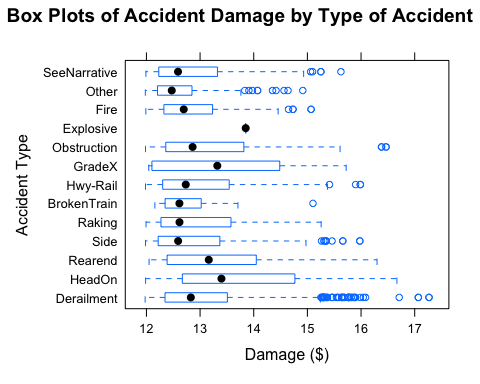
Now that we have identified potential quantitative predictors for our models, we will investigate the categorical variables that contribute to accident severity.

We will start by looking at how the Type of accident contributes to ACCDMG

#Frequency of different types of accidents  
barplot(table(xdmgnd$Type))



#boxplots of accident damage conditioned on accident typw  
bwplot(Type~ log(ACCDMG), main = "Box Plots of Accident Damage by Type of Accident", xlab = "Damage ($)", ylab = "Accident Type", data = xdmgnd)



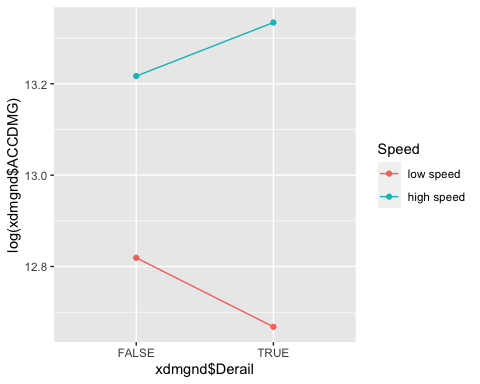
#Total cost of accidents by type as a proportion of total accident damage  
sumbytype<- as.numeric(tapply(as.numeric(xdmgnd$ACCDMG), as.factor(xdmgnd$Type), sum))  
proptype <- sumbytype / sum(as.numeric(xdmgnd$ACCDMG))

The figures and statistics show both the frequency of accidents by type in our extreme accident damage data frame and the cost of accident by type, respectively. Notably, derailments are occuring considerably more than other accident types and account for ~77% of the cost of extreme accidents.This suggests that derailment is a signifcant contributor to extreme accident damage.

1. Treatment of categorical variables for your linear model.

Now let’s look at the relationship between derailments and speed. We will create a new categorical variable Derail so we can compare derailments to all other types of accidents.

# Create the Derail variable to look explicitly at accidents of type derailment  
xdmgnd$Derail <- (xdmgnd$Type == "Derailment")  
  
# Get interaction plot for Derail and TRNSPD  
  
#Define cut point for high and low speed  
Speed <- cut(xdmgnd$TRNSPD, c(min(xdmgnd$TRNSPD), median(xdmgnd$TRNSPD),max(xdmgnd$TRNSPD)), include.lowest = T, labels = c("low speed", "high speed"))  
  
# Plot interaction between Derailments and Speed  
#interaction.plot(xdmgnd$Derail, Speed, log(xdmgnd$ACCDMG))  
ggplot() +  
 aes(x = xdmgnd$Derail, y = log(xdmgnd$ACCDMG), group = Speed, color = Speed) +  
 stat\_summary(fun = mean, geom = "point") +  
 stat\_summary(fun = mean, geom = "line")



Looking at the interaction plot, we can see that the slope of the lines differ for high speed and low speed derailments. The intercept and slopes vary, suggesting there should be both an interaction and a main effects term in the model.

1. Generating Hypotheses

From your analysis, generate at least 2 well-formed, actionable hypotheses for each severity metric. Explain why the hypotheses are actionable and demonstrate how you arrived at each hypothesis. Write out your null and alternative hypotheses.

Based on our observations above, we have formulated 2 hypotheses that relate factors that we can control to the severity of accidents. Our exploratory graphical analysis showed that derailment and speed appear to associate with extreme damage accidents.

1. Main effects term for derailment based on frequency and proportion of accident damage caused by derailments

H0: Derailments do not increase the severity of ACCDMG relative to other types of accidents.

HA: Derailments do increase the severity of ACCDMG relative to other types of accidents.

1. Interaction plot for derailment and train speed with accident damage

H0: The interaction between derailment and speed does not significantly affect the reported accident costs for extreme damage accidents.

HA: The interaction between derailment and speed does significantly affect the reported accident costs for extreme damage accidents.

Build a model to test the above hypotheses (skipping the transformation and jumping straight to model)

1. ACCDMG Analysis
2. We start with multiple models inclduing a main effects and full main effects interaction model with our qualitative and quantitative predictors. We will start with log transformed response because we know our response, ACCDMG, is skewed. We will use stepwise regression with AIC as the metric for evaluation to perform feature selection.
3. Our only categorical variable Derail (type of accident) has been recoded a 0 (accidents that are not derailments) and 1 (for accidents of type derailments) to test our hypotheses.

xdmgnd.lm1.main<-lm(log(ACCDMG)~Derail+TEMP + TRNSPD + CARS + HEADEND1,data=xdmgnd)  
  
xdmgnd.lm1.inter<-lm(log(ACCDMG)~(Derail+TEMP + TRNSPD + CARS + HEADEND1)^2,data=xdmgnd)  
  
xdmgnd.lm1.main.step<- step(xdmgnd.lm1.main, trace = F)  
summary(xdmgnd.lm1.main.step)

##   
## Call:  
## lm(formula = log(ACCDMG) ~ Derail + TEMP + TRNSPD + CARS + HEADEND1,   
## data = xdmgnd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.4547 -0.5419 -0.1461 0.4484 4.5500   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.7129665 0.0338112 375.998 < 2e-16 \*\*\*  
## DerailTRUE 0.0541573 0.0230489 2.350 0.0188 \*   
## TEMP -0.0007199 0.0003759 -1.915 0.0555 .   
## TRNSPD 0.0183131 0.0005108 35.851 < 2e-16 \*\*\*  
## CARS 0.0043535 0.0006816 6.387 1.80e-10 \*\*\*  
## HEADEND1 -0.0451128 0.0068187 -6.616 3.95e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7599 on 7247 degrees of freedom  
## Multiple R-squared: 0.157, Adjusted R-squared: 0.1564   
## F-statistic: 269.9 on 5 and 7247 DF, p-value: < 2.2e-16

AIC(xdmgnd.lm1.main.step)

## [1] 16608.89

xdmgnd.lm1.inter.step<- step(xdmgnd.lm1.inter, trace = F)  
summary(xdmgnd.lm1.inter.step)

##   
## Call:  
## lm(formula = log(ACCDMG) ~ Derail + TEMP + TRNSPD + CARS + HEADEND1 +   
## Derail:TRNSPD + Derail:HEADEND1 + TEMP:CARS + TRNSPD:CARS +   
## TRNSPD:HEADEND1 + CARS:HEADEND1, data = xdmgnd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.9312 -0.5279 -0.1391 0.4278 4.5260   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.290e+01 4.352e-02 296.472 < 2e-16 \*\*\*  
## DerailTRUE -1.494e-01 4.107e-02 -3.639 0.000276 \*\*\*  
## TEMP -1.630e-04 3.915e-04 -0.416 0.677136   
## TRNSPD 3.266e-03 1.039e-03 3.143 0.001677 \*\*   
## CARS 2.042e-03 2.358e-03 0.866 0.386413   
## HEADEND1 -3.830e-02 1.657e-02 -2.311 0.020845 \*   
## DerailTRUE:TRNSPD 1.561e-02 1.114e-03 14.013 < 2e-16 \*\*\*  
## DerailTRUE:HEADEND1 -6.495e-02 1.655e-02 -3.924 8.80e-05 \*\*\*  
## TEMP:CARS -6.907e-05 2.633e-05 -2.623 0.008728 \*\*   
## TRNSPD:CARS 1.139e-04 4.291e-05 2.654 0.007971 \*\*   
## TRNSPD:HEADEND1 2.069e-03 3.886e-04 5.323 1.05e-07 \*\*\*  
## CARS:HEADEND1 1.386e-03 5.899e-04 2.350 0.018804 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7432 on 7241 degrees of freedom  
## Multiple R-squared: 0.1944, Adjusted R-squared: 0.1931   
## F-statistic: 158.8 on 11 and 7241 DF, p-value: < 2.2e-16

AIC(xdmgnd.lm1.inter.step)

## [1] 16291.67

anova(xdmgnd.lm1.main.step, xdmgnd.lm1.inter.step)

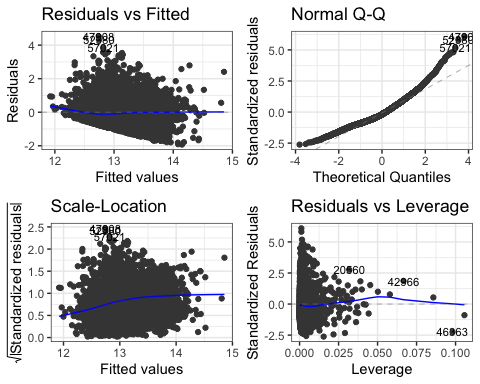
## Analysis of Variance Table  
##   
## Model 1: log(ACCDMG) ~ Derail + TEMP + TRNSPD + CARS + HEADEND1  
## Model 2: log(ACCDMG) ~ Derail + TEMP + TRNSPD + CARS + HEADEND1 + Derail:TRNSPD +   
## Derail:HEADEND1 + TEMP:CARS + TRNSPD:CARS + TRNSPD:HEADEND1 +   
## CARS:HEADEND1  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7247 4185.1   
## 2 7241 3999.4 6 185.72 56.042 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. Model assessment

Based on the lower AIC and the significance of our additional predictors based on the result of the partial F test, we choose our stepwise interaction model over our main effects stepwise model.

1. How you diagnosed problems with the models

autoplot(xdmgnd.lm1.inter.step, which = c(1,2,3,5), ncol = 2, label.size = 3) + theme\_bw()



These plots show a good fit of the residuals to a Gaussian distribution so our inferential analysis should be supported by these results. The residuals versus fitted plot shows mostly constant variance but it also shows a lack of fit. particular, there is a clear floor effect on the residual values and this is the result of the extreme damage threshold used to define these data. The lack of fit is also evident by the apparent slight curve in the scale-location plot. Not surprisingly this particular lack of fit suggests one or more missing variables to explain the variance in accident damage. So, the variables we used from the accident reports provided do not have sufficient data to fully model accident damage, but it is an adequate starting point.

1. How you adjusted the models based on these assessment

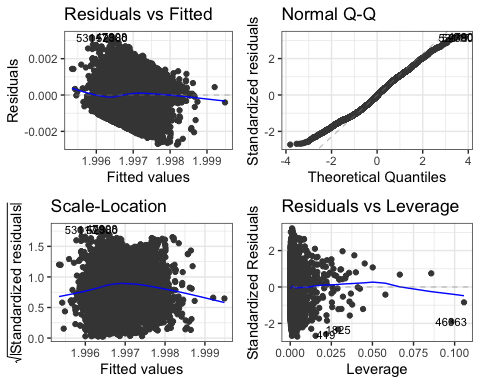
Let us also try a box cox transformation and choose between the log and box cox for our model.

xdmgnd.lm1.inter<-lm(ACCDMG~(Derail+TEMP + TRNSPD + CARS + HEADEND1)^2,data=xdmgnd)  
  
L<-boxcox(xdmgnd.lm1.inter.step, plotit = F)$x[which.max(boxcox(xdmgnd.lm1.inter, plotit = F)$y)]   
  
xdmgnd.lm1.inter.boxcox<-lm((ACCDMG^L-1)/L~(Derail+TEMP + TRNSPD + CARS + HEADEND1)^2,data=xdmgnd)  
  
xdmgnd.lm1.inter.boxcox.step<- step(xdmgnd.lm1.inter.boxcox, trace = F)  
summary(xdmgnd.lm1.inter.boxcox.step)

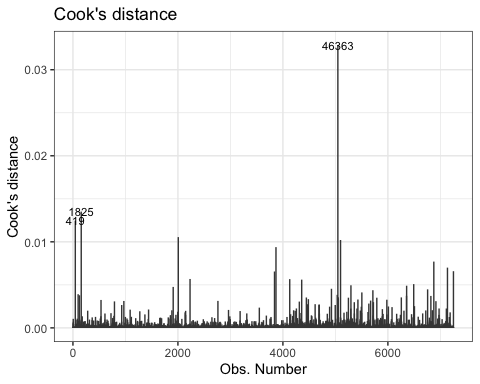
##   
## Call:  
## lm(formula = (ACCDMG^L - 1)/L ~ Derail + TEMP + TRNSPD + CARS +   
## HEADEND1 + Derail:TRNSPD + Derail:HEADEND1 + TEMP:CARS +   
## TRNSPD:CARS + TRNSPD:HEADEND1 + CARS:HEADEND1, data = xdmgnd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0027280 -0.0007892 -0.0000481 0.0007516 0.0032294   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.997e+00 5.872e-05 33999.799 < 2e-16 \*\*\*  
## DerailTRUE -1.841e-04 5.542e-05 -3.323 0.000896 \*\*\*  
## TEMP -3.891e-07 5.283e-07 -0.737 0.461381   
## TRNSPD 3.175e-06 1.402e-06 2.265 0.023554 \*   
## CARS 3.075e-06 3.181e-06 0.967 0.333776   
## HEADEND1 -6.812e-05 2.236e-05 -3.046 0.002328 \*\*   
## DerailTRUE:TRNSPD 2.119e-05 1.503e-06 14.099 < 2e-16 \*\*\*  
## DerailTRUE:HEADEND1 -6.805e-05 2.234e-05 -3.046 0.002325 \*\*   
## TEMP:CARS -7.698e-08 3.553e-08 -2.167 0.030305 \*   
## TRNSPD:CARS 1.213e-07 5.790e-08 2.096 0.036140 \*   
## TRNSPD:HEADEND1 3.271e-06 5.244e-07 6.237 4.72e-10 \*\*\*  
## CARS:HEADEND1 1.408e-06 7.960e-07 1.769 0.076939 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.001003 on 7241 degrees of freedom  
## Multiple R-squared: 0.193, Adjusted R-squared: 0.1917   
## F-statistic: 157.4 on 11 and 7241 DF, p-value: < 2.2e-16

Plot diagnostics again for your model and compare to log transformed

autoplot(xdmgnd.lm1.inter.boxcox.step, which = c(1,2,3,5), ncol = 2, label.size = 3) + theme\_bw()



autoplot(xdmgnd.lm1.inter.boxcox.step, which=4, ncol = 1, label.size = 3) + theme\_bw() #Cook's distance



The boxcox plot is slightly better visually than the log transformation, but still suffers from lack of fit suggesting more data is needed to appropriately explain accident damage. However, the model is a good starting point to exploring how to reduce accident severity.

Compare your best models. \*\*At this point, you could compare different models with same response using criterion-based metrics such as AIC, BIC or PMSE from training and test sets or CV. For example, a main effects vs. an interaction or models with transformed predictors.

1. Evidence Recommendation to FRA
2. Assess whether or not you can reject your null hypotheses.

Based on the results, we fail to reject the hypothesis that derailments do not increase the severity of rail accident damage. However, we can reject H0 for the hypothesis that the interaction of train speed and accidents of type derailment do not increase the severity of rail accident damage at p<0.05.

1. Summarize your findings and your recommendations for safety improvements based on the evidence you have discovered and include next steps the FRA should take to reduce the severity of rail accidents.

The results from this study suggest two types of recommendations: 1. Steps for immediate improvement and 2. Directions for future study. We base recommendations for immediate improvement on the evidence discussed in regards to hypotheses 1 and 2.

The major finding is the association between derailment and derailment at high speed to high cost accidents, while controlling for the values of the other predictor variables (p < 0.0001). Hence, we can improve accident safety by preventing derailment accidents and particularly those that occur at higher speeds. We make several recommendations for achieving this improvement. The first is to continue development, deployment, and use of Positive Train Control (PTC). PTC has been developed under Federal Railroad Administration funding and consists of a suite of technologies proposed to achieve automated or semi-automated control of train operations [1]. PTC uses and integrates technologies about the condition of the trains, the environment, and the track. When implemented PTC has the potential to prevent or reduce the incidents of high speed derailments.

Second, the model used for analyzing extreme damage accidents showed a lack of fit to the data. This can be improved with the inclusion of other relevant data in the accident report or collecting new data. We specifically recommend the inclusion of operational data. These data may give additional variables that are not in the accident reports. Data that are in the accident reports but were not used in this study are in the narratives or free text. These data should be mined for possible relevance to the contributors to extreme accident damage.

[1] Positive train control (ptc). Federal Railroad Administration. [Online]. Available: <http://safetydata.fra.dot.gov/>