Project 1

Reese Quillian, Sofia Zajec, Cecilia Smith, TJ Gwilliam

2022-10-24

## Load data / packages

# Libraries and files  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(lattice)  
library(GGally)

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

library(devtools)

## Loading required package: usethis

library(ggfortify)  
library(MASS)

##   
## Attaching package: 'MASS'

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

traindir <- "C:/Users/Student/OneDrive - University of Virginia/Documents/SYS4021/In Class/Data/Train Data"  
sourcedir <-"C:/Users/Student/OneDrive - University of Virginia/Documents/SYS4021/Project"  
  
setwd(sourcedir)  
  
# files for analysis  
source("AccidentInput.R")  
source("SPM\_Panel.R")  
source("PCAplots.R")

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

##   
## Attaching package: 'scales'

## The following objects are masked from 'package:psych':  
##   
## alpha, rescale

##   
## Attaching package: 'ggpubr'

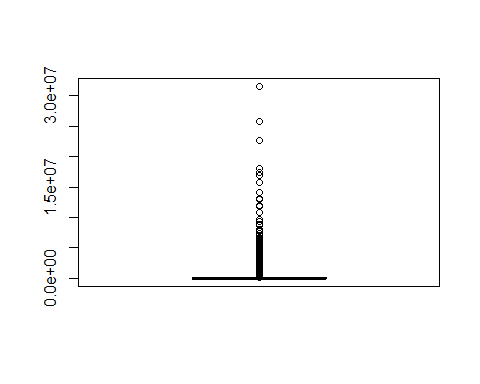
## The following object is masked from 'package:plyr':  
##   
## mutate

# load data  
acts <- file.inputl(traindir)  
  
totacts <- combine.data(acts)

## Cleaning

For this analysis, we only consider extreme accidents, as we are not concerned with accidents that do not lead to significant damages/casualties. For the ACCDMG metric, extreme accidents are those whose ACCDMG are above the upper whisker. For the Casualties metric, extreme accidents are those with at least 1 casualty.

# extreme accidents  
# Build a data frame with only extreme accidents for ACCDMG  
  
dmgbox <-boxplot(totacts$ACCDMG)



# accidents above upper whisker  
xdmg <- totacts[totacts$ACCDMG > dmgbox$stats[5],]  
  
#remove 9/11  
xdmg <- xdmg[-183,]  
  
## Remove duplicates from xdmg and call new data frame xdmgnd  
xdmgnd <- xdmg[!(duplicated(xdmg[, c("INCDTNO", "YEAR", "MONTH", "DAY", "TIMEHR", "TIMEMIN")])),]  
# xdmgnd = dataframe to use  
  
xdmgnd$Type <- factor(xdmgnd$TYPE, labels = c("Derailment", "HeadOn", "Rearend", "Side", "Raking", "BrokenTrain", "Hwy-Rail", "GradeX", "Obstruction", "Explosive", "Fire","Other","SeeNarrative"))  
  
# casualties = TOTINJ + TOTKLD  
xdmgnd <- xdmgnd %>% mutate(casualties = TOTKLD + TOTINJ)

# Setup categorical variables  
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)  
  
# Speed variable  
xdmgnd$Speed <- cut(xdmgnd$TRNSPD, c(min(xdmgnd$TRNSPD),median(xdmgnd$TRNSPD),max(xdmgnd$TRNSPD)), include.lowest = T, labels = c("low speed", "high speed"))  
  
# human factors variable  
xdmgnd$human\_factors <- rep(0, nrow(xdmgnd))  
xdmgnd$human\_factors[which(xdmgnd$Cause == "H")] <- 1   
xdmgnd$human\_factors <- factor(xdmgnd$human\_factors)

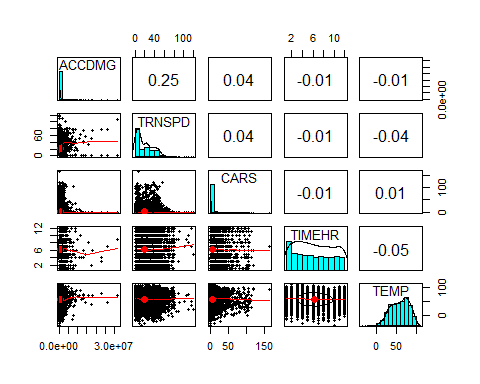
#Create a new dataframe with only 1 or more casualties  
xdmgnd\_cas <- xdmgnd %>% filter(casualties > 0)  
  
# remove max (outlier)  
xdmgnd\_cas <- xdmgnd\_cas %>% filter(casualties != max(casualties))

## 1. Hypotheses

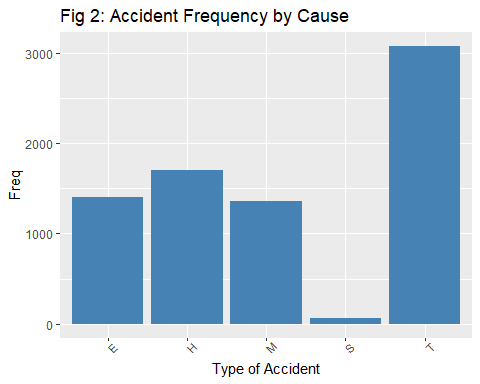
### 1.1 ACCDMG

We first began with a correlation matrix and saw that train speed (TRNSPD) was the most highly correlated with accident damage (ACCDMG). In exploring a categorical variable not included in the correlation matrix, we decided to make bar charts Accident Frequency by Cause (Fig. 2) and Mean Accident Damage by Cause (Fig. 3). These two visualizations showed that H, which corresponds to an accident attributed to human factors, is the second most common accident cause as well as the second highest mean accident cost, leading to human factors being a variable of interest because it is both common and costly per accident. We then recoded the train speed into a categorical variable with two levels, high and low speeds and created an interaction plot as shown below:

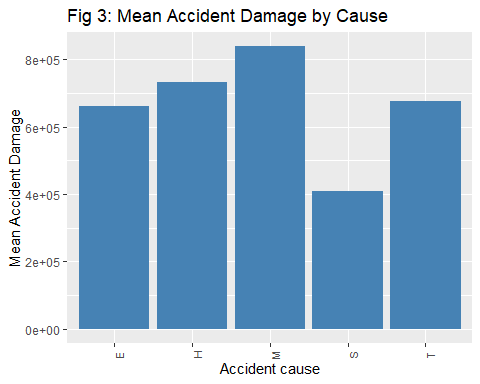
pairs.panels(xdmgnd[,c("ACCDMG", "TRNSPD", "CARS", "TIMEHR", "TEMP")])



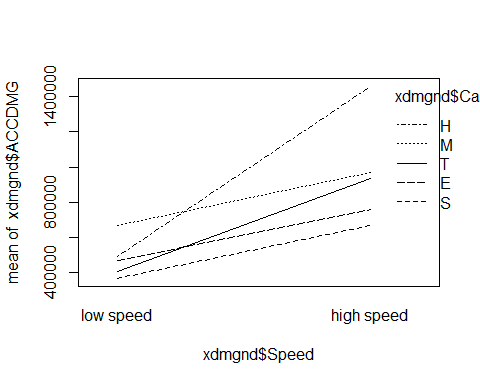
ggplot(as.data.frame(table(xdmgnd$Cause)), aes(x = Var1, y= Freq)) +  
 geom\_bar(stat="identity",fill= "steelblue")+   
 ggtitle("Fig 2: Accident Frequency by Cause") +  
 labs(x = "Type of Accident")+  
 theme(axis.text.x = element\_text(size = 8, angle = 45))



df\_causes<- xdmgnd %>% group\_by(Cause) %>% dplyr::summarise(Damage=mean(ACCDMG),n=n())  
  
ggplot(df\_causes, aes(x = Cause, y=Damage)) +  
 geom\_col(fill= "steelblue")+   
 ggtitle("Fig 3: Mean Accident Damage by Cause") +  
 labs(x = "Accident cause", y = "Mean Accident Damage")+  
 theme(axis.text.x = element\_text(size = 8, angle = 90))



interaction.plot(xdmgnd$Speed, xdmgnd$Cause, xdmgnd$ACCDMG)



The interaction plots showed that human factors have the greatest discrepancy of accident cost between low and high speeds than any other cause, leading to the generation of our first hypothesis: At high train speeds, human error is the most costly cause of accident.

#### Hypothesis 1: At high train speeds, human errors is the most costly cause of accident.

H0: ACCDMG for all causes are equal at high train speeds.

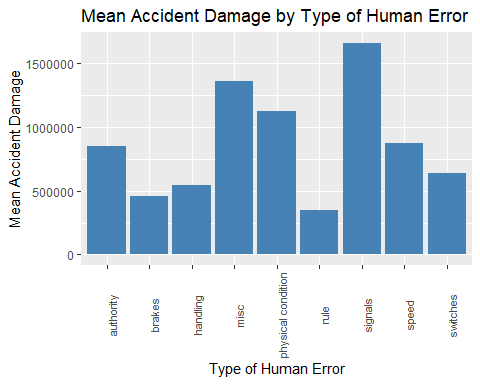
HA: ACCDMG for human factors is not equal to other causes at high speeds.

This hypothesis is actionable because trains known to go at higher speeds can be paid more attention to. For example, additional or more senior staff can be assigned to high speed trains. To arrive at this hypothesis, we looked at the interaction plot between cause and speed and saw that human factors had the steepest slope between low and high speeds. We then looked at the interaction between only human errors and speed and it appears that at higher speeds, human errors causes a disproportionate amount of damage.

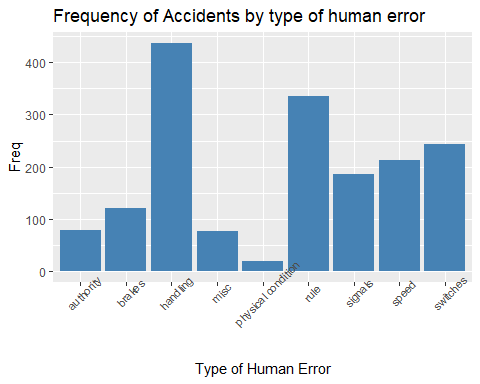
We then turned to exploring the more specific cause (CAUSE) of the accident. Exploration of our first hypothesis showed human factors were overrepresented in the accident damage severity metric, however in creating a bar chart Mean Accident Damage by Type of Human Error, Flagging, Fixed, Hand and Radio Signals (signals) were the most costly type of human error. The Frequency of Accidents by type of human error bar chart showed that signals were not the most common, but still made up ~25% of the cost of extreme accidents caused by human factors.

# recode human factor CAUSE  
  
xdmgnd$human\_factor\_level <- rep(NA, nrow(xdmgnd))  
  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H0")] <- "brakes"  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H1")] <- "physical condition"  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H2")] <- "signals"  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H3")] <- "rule"  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H4")] <- "authority"  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H5")] <- "handling"  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H6")] <- "speed"  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H7")] <- "switches"  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H8")] <- "cab"  
xdmgnd$human\_factor\_level[which(substr(xdmgnd$CAUSE,1,2)=="H9")] <- "misc"  
  
xdmgnd$human\_factor\_level <- factor(xdmgnd$human\_factor\_level)

df<- xdmgnd %>% filter(Cause == "H") %>% group\_by(human\_factor\_level) %>% dplyr::summarise(Damage=mean(ACCDMG),n=n())  
  
ggplot(df, aes(x = human\_factor\_level, y=Damage)) +  
 geom\_col(fill= "steelblue")+   
 ggtitle("Mean Accident Damage by Type of Human Error") +  
 labs(x = "Type of Human Error", y = "Mean Accident Damage")+  
 theme(axis.text.x = element\_text(size = 8, angle = 90))



df\_hf<- xdmgnd %>% filter(Cause == "H")  
  
ggplot(as.data.frame(table(df\_hf$human\_factor\_level)), aes(x = Var1, y= Freq)) +  
 geom\_bar(stat="identity",fill= "steelblue")+   
 ggtitle("Frequency of Accidents by type of human error") +  
 labs(x = "Type of Human Error")+  
 theme(axis.text.x = element\_text(size = 8, angle = 45))



# Total cost of human error accidents by type as a proportion of total accident damage  
sumbytype<- as.numeric(tapply(as.numeric(df\_hf$ACCDMG), as.factor(df\_hf$human\_factor\_level), sum))  
proptype <- sumbytype / sum(as.numeric(df\_hf$ACCDMG))  
  
proptype

## [1] 0.05305501 0.04376580 0.18849434 0.08280430 0.01708031 0.09407907 0.24698521  
## [8] 0.14849948 0.12523647

The high cost and proportion of total accident cost that signals make up of human factors led to the development of our second hypothesis: Signaling errors (a type of human errors) lead to disproportionately more costly accidents.

#### Hypothesis 2: Signaling errors lead to disproportionately more costly accidents.

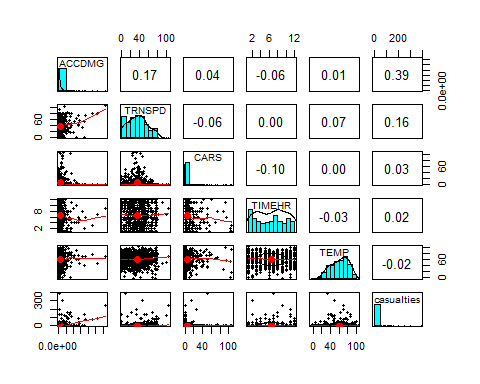
H0 = ACCDMG for signaling errors is the same as other errors.

HA = ACCDMG caused by signaling errors is higher than other errors.

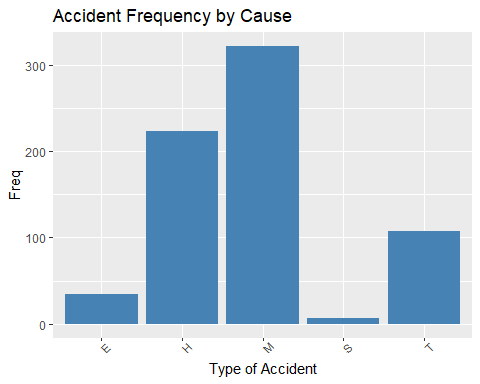
This hypothesis is actionable because training of conductors can be updated or improved if the evidence supports rejection of H0. We arrived at this hypothesis by first looking into the overall frequency of accidents by cause and found that human factors was the second most common cause of train accidents. This lead us to look into specific types of human errors, and found that signaling errors incur the most damage despite being 5/10 in terms of frequency.

### 1.2 Casualties

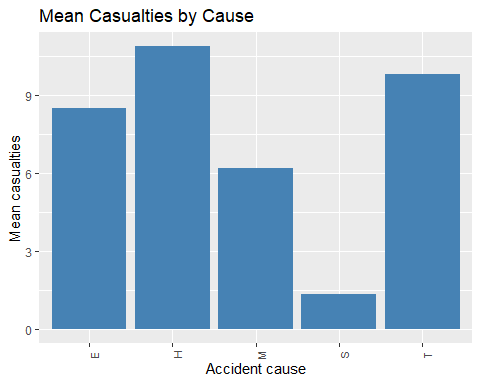
pairs.panels(xdmgnd\_cas[,c("ACCDMG", "TRNSPD", "CARS", "TIMEHR", "TEMP", "casualties")])



ggplot(as.data.frame(table(xdmgnd\_cas$Cause)), aes(x = Var1, y= Freq)) +  
 geom\_bar(stat="identity",fill= "steelblue")+   
 ggtitle("Accident Frequency by Cause") +  
 labs(x = "Type of Accident")+  
 theme(axis.text.x = element\_text(size = 8, angle = 45))



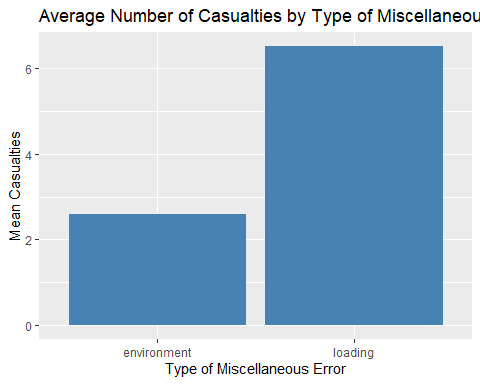
df\_causes\_cas<- xdmgnd\_cas %>% group\_by(Cause) %>% dplyr::summarise(average\_casualties=mean(casualties),n=n())  
  
ggplot(df\_causes\_cas, aes(x = Cause, y=average\_casualties)) +  
 geom\_col(fill= "steelblue")+   
 ggtitle("Mean Casualties by Cause") +  
 labs(x = "Accident cause", y = "Mean casualties")+  
 theme(axis.text.x = element\_text(size = 8, angle = 90))



Casualties are much more commonly cited as due to miscellaneous causes. Now we want to break down by general type within miscellaneous category:

xdmgnd\_cas$misc\_type <- rep(NA, nrow(xdmgnd\_cas))  
  
xdmgnd\_cas$misc\_type[which(substr(xdmgnd\_cas$CAUSE,1,2)=="M1")] <- "environment"  
xdmgnd\_cas$misc\_type[which(substr(xdmgnd\_cas$CAUSE,1,2)=="M2")] <- "loading"  
xdmgnd\_cas$misc\_type[which(substr(xdmgnd\_cas$CAUSE,1,2)=="M3")] <- "loading"  
xdmgnd\_cas$misc\_type[which(substr(xdmgnd\_cas$CAUSE,1,2)=="M4")] <- "loading"  
xdmgnd\_cas$misc\_type[which(substr(xdmgnd\_cas$CAUSE,1,2)=="M5")] <- "loading"  
  
xdmgnd\_cas$misc\_type <- factor(xdmgnd\_cas$misc\_type)

df2<- xdmgnd\_cas %>% filter(Cause == "M") %>% group\_by(misc\_type) %>% dplyr::summarize(average\_casualties = mean(casualties),n=n())  
  
ggplot(df2, aes(x = misc\_type, y=average\_casualties)) +  
 geom\_col(fill= "steelblue")+   
 ggtitle("Average Number of Casualties by Type of Miscellaneous Error") +  
 labs(x = "Type of Miscellaneous Error", y = "Mean Casualties")



# loading procedures variable (will use in model)  
xdmgnd\_cas$loading <- rep(0, nrow(xdmgnd\_cas))  
xdmgnd\_cas$loading[which(xdmgnd\_cas$misc\_type == "loading")] <- 1   
xdmgnd\_cas$loading <- factor(xdmgnd\_cas$loading)

Now we want to break down by general type within the miscellaneous category to understand what they might entail. Breakdowns of miscellaneous discussed in the data dictionary were applied through the recoding of the CAUSE variable and visualized to understand the number of fatal or injury-inducing accidents by cause, as seen in the average number of casualties by type of miscellaneous error bar chart.

#### Hypothesis 3: Errors in loading procedures lead to disproportionately more casualties.

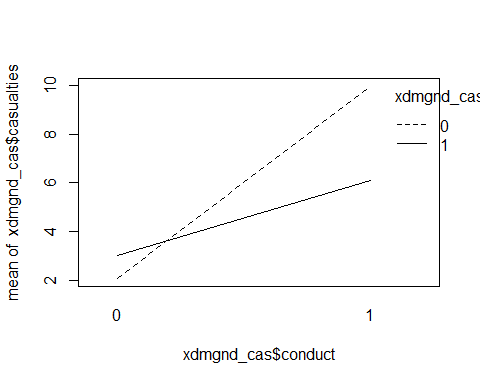
H0 = casualties in accidents caused by loading errors are the same as all other causes

HA = casualties in accidents caused by loading errors are higher than other human errors.

According to the Association of American Railroads, collisions at grade crossings and incidents involving trespassers on railroad property account for well over 95% of rail-related fatalities [1]. Based on this statistic, we decided to look into highway-rail accidents (TYPE) as being a predictor for casualties. Furthermore, we believed that having no conductors on board the train would likely result in more casualties as there would be no supervision on the train itself. Given this relationship, we created two binary variables to represent if an accident was a highway-rail and if there was a conductor on board and made an interaction plot with the two variables alongside casualties. The interaction plot showed that for highway-rail accidents, having a conductor seems to reduce the average number of casualties.

# hwyrail variable  
xdmgnd\_cas$hwyrail <- rep(0, nrow(xdmgnd\_cas))  
xdmgnd\_cas$hwyrail[which(xdmgnd\_cas$Type == "Hwy-Rail")] <- 1   
xdmgnd\_cas$hwyrail <- factor(xdmgnd\_cas$hwyrail)  
  
# conductor variable  
xdmgnd\_cas$conduct <- rep(0, nrow(xdmgnd\_cas))  
xdmgnd\_cas$conduct[which(xdmgnd\_cas$CONDUCTR > 0)] <- 1  
xdmgnd\_cas$conduct <- factor(xdmgnd\_cas$conduct)

# interaction  
interaction.plot(xdmgnd\_cas$conduct, xdmgnd\_cas$hwyrail, xdmgnd\_cas$casualties)



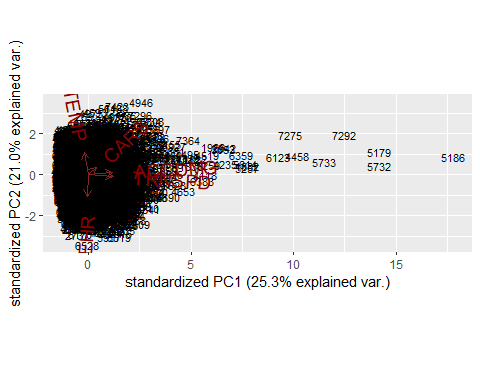
#### Hypothesis 4: Trains with conductors are more likely to reduce the number of casualties in highway-rail accidents.

H0 = For highway-rail type accidents, the number of casualties on trains with conductors is equal to the number of casualties on trains without conductors. HA = For highway-rail type accidents, the number of casualties on trains with conductors is less than the number of casualties on trains without conductors.

## 2. ACCDMG Analysis

### a) feature and model selection techniques

predictors.accdmg.pca <- princomp(xdmgnd[,c("ACCDMG", "TRNSPD", "CARS", "TIMEHR", "TEMP")], cor = T )  
  
ggbiplot(predictors.accdmg.pca, varname.size = 5, labels=row(xdmgnd)[,1])



TRNSPD and ACCDMG are very correlated, so we will add it as a predictor to our models.

### b) treatment of ordinal and categorical variables

* Cause
* human\_factors
* human\_factor\_level
* signals Y/N

### c) model assessment, d) diagnosing issues, e) adjustments

#### Hypothesis 1: At high train speeds, human errors is the most costly cause of accident.

H0: ACCDMG for all causes are equal at high train speeds. HA: ACCDMG for human factors is not equal to other causes at high speeds.

# interaction model because our hypothesis is about human errors AT high speeds  
accdmg.lm1 <- lm(ACCDMG~(TRNSPD+human\_factors)^2,data=xdmgnd)  
summary(accdmg.lm1)

##   
## Call:  
## lm(formula = ACCDMG ~ (TRNSPD + human\_factors)^2, data = xdmgnd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3770704 -366831 -201007 69903 31304687   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 388093.8 24669.8 15.732 < 2e-16 \*\*\*  
## TRNSPD 14212.7 851.7 16.686 < 2e-16 \*\*\*  
## human\_factors1 -154027.1 45250.0 -3.404 0.000668 \*\*\*  
## TRNSPD:human\_factors1 27363.2 2214.9 12.354 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1194000 on 7597 degrees of freedom  
## Multiple R-squared: 0.08355, Adjusted R-squared: 0.08319   
## F-statistic: 230.9 on 3 and 7597 DF, p-value: < 2.2e-16

#### Hypothesis 2: Signaling errors lead to disproportionately more costly accidents.

H0 = ACCDMG for signaling errors is the same as other errors. HA = ACCDMG caused by signaling errors is higher than human errors.

# signals variable  
xdmgnd$signals <- rep(0, nrow(xdmgnd))  
xdmgnd$signals[which(xdmgnd$human\_factor\_level == "signals")] <- 1   
xdmgnd$signals <- factor(xdmgnd$signals)  
  
accdmg.lm2 <- lm(ACCDMG~signals+TRNSPD,data=xdmgnd)  
summary(accdmg.lm2)

##   
## Call:  
## lm(formula = ACCDMG ~ signals + TRNSPD, data = xdmgnd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2168664 -353422 -182995 60797 31197312   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 341442.4 20813.0 16.41 <2e-16 \*\*\*  
## signals1 1018594.6 89040.2 11.44 <2e-16 \*\*\*  
## TRNSPD 17238.5 765.9 22.51 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1199000 on 7598 degrees of freedom  
## Multiple R-squared: 0.07609, Adjusted R-squared: 0.07585   
## F-statistic: 312.9 on 2 and 7598 DF, p-value: < 2.2e-16

signals significant when controlling for speed

Determine whether signaling errors or human factor errors as a whole is a better predictor: - human\_factors is more significant in lm4 than signals in lm3 - therefore we will move forward using only human\_factors as a predictor (we can’t include both because of multicollinearity)

AIC(accdmg.lm1)

## [1] 234299.4

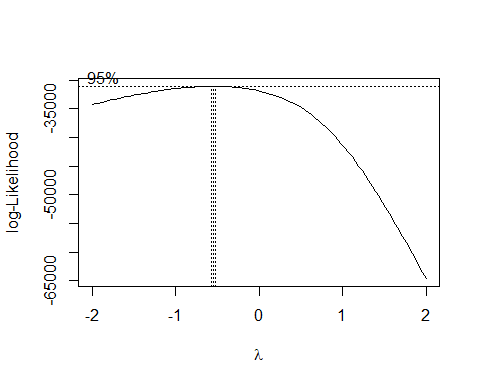
AIC(accdmg.lm2)

## [1] 234359

Based on the model assessment above, accdmg.lm1 is the better model. We will move forward with using human factors as a predictor and not signals specifically.

Now we can check if a transformation is needed.

#Box-Cox Transformation   
boxcox(accdmg.lm1) #box-cox plot



The plot above suggests that we need to use a boxcox transformation for ACCDMG. The optimal lambda value is:

L<-boxcox(accdmg.lm1, plotit = F)$x[which.max(boxcox(accdmg.lm1, plotit = F)$y)]   
L

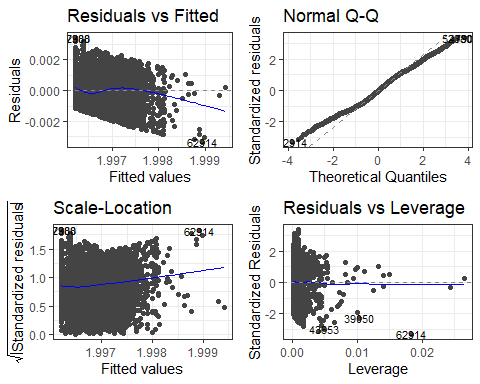
## [1] -0.5

# boxcox transformation  
accdmg.lm1.boxcox <- lm((ACCDMG^L-1)/L~(TRNSPD+human\_factors)^2,data=xdmgnd)  
summary(accdmg.lm1.boxcox)

##   
## Call:  
## lm(formula = (ACCDMG^L - 1)/L ~ (TRNSPD + human\_factors)^2, data = xdmgnd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0033451 -0.0008175 -0.0000368 0.0007836 0.0034365   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.996e+00 2.104e-05 94895.401 < 2e-16 \*\*\*  
## TRNSPD 2.160e-05 7.263e-07 29.739 < 2e-16 \*\*\*  
## human\_factors1 -1.724e-04 3.859e-05 -4.467 8.03e-06 \*\*\*  
## TRNSPD:human\_factors1 8.650e-06 1.889e-06 4.580 4.73e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.001018 on 7597 degrees of freedom  
## Multiple R-squared: 0.1458, Adjusted R-squared: 0.1454   
## F-statistic: 432.1 on 3 and 7597 DF, p-value: < 2.2e-16

Model diagnostics:

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

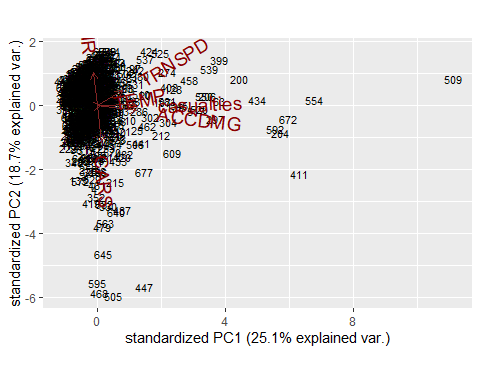


## 3. Casualties Analysis

### a) feature and model selection techniques

Our first step in feature selection was to set up the casualties variable by adding total killed and total injured for every accident. We then defined a new data frame with only one or more casualties to use in our analysis. To identify quantitative predictors that we could use in our model, we chose to do PCA analysis:

predictors.casualties.pca <- princomp(xdmgnd\_cas[,c("casualties", "TRNSPD", "CARS", "TIMEHR", "TEMP","ACCDMG")], cor = T )  
  
ggbiplot(predictors.casualties.pca, varname.size = 5, labels=row(xdmgnd\_cas)[,1])



Based on the biplot displayed above, we can see that TRNSPD and ACCDMG are correlated with casualties (though the temperature variable is pointing in the same direction, the arrow is very small). The correlation with ACCDMG suggests that the two severity metrics may have similar predictors. We can use TRNSPD in our casualties model and add predictors from our ACCDMG analysis if others found are not predictive enough. We discuss feature selection and treatment of ordinal and categorical variables in the following section.

### b) treatment of ordinal and categorical variables

The first, and main categorical variable we investigate in our analysis is the CAUSE variable. We recoded CAUSE to a factor with 5 levels that represent the 5 broad classes of accident causes (M: miscellaneous, T: rack, roadbed and structures, S: signal and communication, H: human factors, E: mechanical and electrical failures). A bar plot showed the relative frequencies of accidents with casualties among each cause, and demonstrated that accidents caused by “miscellaneous - other factors” were the most common in our casualties dataframe. This led us to break down the variable further to examine specific errors that occurred within accidents due to “miscellaneous” errors, using the first two indices of the cause code (M1XX, M2XX, etc.). Further breakdown is more useful so that we can actually understand what is leading to a higher number of casualties. We find that most of these accidents were cited as caused by loading procedures. Based on this, we created a loading procedures binary variable to use in our model for our third hypothesis: that errors in loading procedures lead to disproportionately more casualties.

[add discussion of hwyrail and conductor variables]

### c) model assessment, d) diagnosing issues, e) adjustments

#### Hypothesis 3: Errors in loading procedures lead to disproportionately more casualties.

H0 = casualties in accidents caused by loading errors are the same as all other causes. HA = casualties in accidents caused by loading errors are higher than other human errors.

casualties.lm1 <- lm(casualties~loading+TRNSPD,data=xdmgnd\_cas)  
summary(casualties.lm1)

##   
## Call:  
## lm(formula = casualties ~ loading + TRNSPD, data = xdmgnd\_cas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.46 -9.04 -4.79 -0.50 381.14   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.50453 2.03411 0.740 0.459763   
## loading1 -8.96036 2.38124 -3.763 0.000182 \*\*\*  
## TRNSPD 0.29430 0.05309 5.543 4.23e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27.85 on 689 degrees of freedom  
## Multiple R-squared: 0.04562, Adjusted R-squared: 0.04285   
## F-statistic: 16.47 on 2 and 689 DF, p-value: 1.032e-07

#### Hypothesis 4: Trains with conductors are more likely to reduce the number of casualties in highway-rail accidents.

H0 = For highway-rail type accidents, the number of casualties on trains with conductors is equal to the number of casualties on trains without conductors.

HA = For highway-rail type accidents, the number of casualties on trains with conductors is less than the number of casualties on trains without conductors.

casualties.lm2 <- lm(casualties~(conduct+hwyrail+TRNSPD)^2,data=xdmgnd\_cas)  
summary(casualties.lm2)

##   
## Call:  
## lm(formula = casualties ~ (conduct + hwyrail + TRNSPD)^2, data = xdmgnd\_cas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.52 -8.70 -3.89 0.19 383.05   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.07174 6.35350 0.169 0.866  
## conduct1 -2.02791 6.66918 -0.304 0.761  
## hwyrail1 4.64990 13.39353 0.347 0.729  
## TRNSPD 0.09035 0.27465 0.329 0.742  
## conduct1:hwyrail1 -7.99409 13.60418 -0.588 0.557  
## conduct1:TRNSPD 0.27352 0.27381 0.999 0.318  
## hwyrail1:TRNSPD -0.16013 0.11866 -1.349 0.178  
##   
## Residual standard error: 27.83 on 685 degrees of freedom  
## Multiple R-squared: 0.05303, Adjusted R-squared: 0.04473   
## F-statistic: 6.393 on 6 and 685 DF, p-value: 1.452e-06

Both casualties.lm1 and casualties.lm2 have adjusted R-squared of ~.04. We can use AIC to further assess these models:

AIC(casualties.lm1)

## [1] 6573.322

AIC(casualties.lm2)

## [1] 6575.934

# combine predictors from casualties.lm1 and casualties.lm2 and check AIC  
casualties.lm2.step <- step(casualties.lm2, trace=T)

## Start: AIC=4610.12  
## casualties ~ (conduct + hwyrail + TRNSPD)^2  
##   
## Df Sum of Sq RSS AIC  
## - conduct:hwyrail 1 267.37 530665 4608.5  
## - conduct:TRNSPD 1 772.65 531170 4609.1  
## - hwyrail:TRNSPD 1 1410.12 531808 4610.0  
## <none> 530398 4610.1  
##   
## Step: AIC=4608.47  
## casualties ~ conduct + hwyrail + TRNSPD + conduct:TRNSPD + hwyrail:TRNSPD  
##   
## Df Sum of Sq RSS AIC  
## - conduct:TRNSPD 1 505.71 531171 4607.1  
## <none> 530665 4608.5  
## - hwyrail:TRNSPD 1 1711.45 532376 4608.7  
##   
## Step: AIC=4607.13  
## casualties ~ conduct + hwyrail + TRNSPD + hwyrail:TRNSPD  
##   
## Df Sum of Sq RSS AIC  
## - conduct 1 36.8 531207 4605.2  
## <none> 531171 4607.1  
## - hwyrail:TRNSPD 1 1735.8 532906 4607.4  
##   
## Step: AIC=4605.18  
## casualties ~ hwyrail + TRNSPD + hwyrail:TRNSPD  
##   
## Df Sum of Sq RSS AIC  
## <none> 531207 4605.2  
## - hwyrail:TRNSPD 1 1764.5 532972 4605.5

Comparing the step model to casualties.lm1:

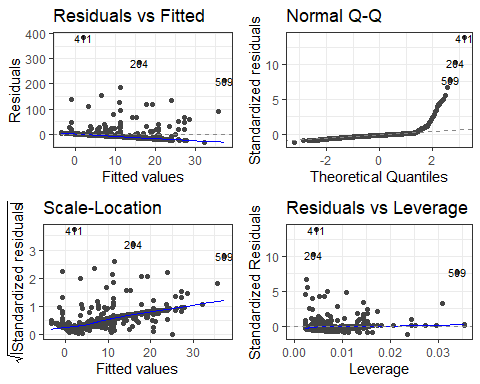
AIC(casualties.lm1)

## [1] 6573.322

AIC(casualties.lm2.step)

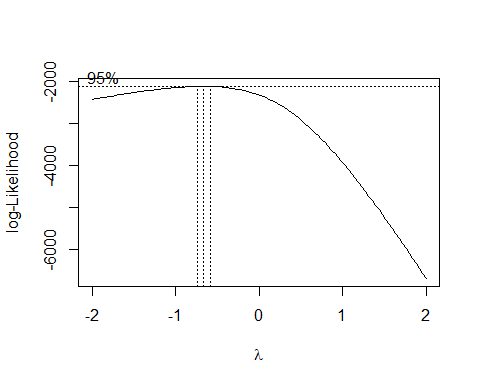
## [1] 6570.99

autoplot(casualties.lm2.step, which = c(1,2,3,5), ncol = 2, label.size = 3) + theme\_bw()



Diagnostics above show we need to transform the casualties variable:

#Box-Cox Transformation   
boxcox(casualties.lm2.step) #box-cox plot



# boxcox transformation  
L<-boxcox(accdmg.lm1, plotit = F)$x[which.max(boxcox(accdmg.lm1, plotit = F)$y)]  
  
casualties.lm2.boxcox <- lm((casualties^L-1)/L~(conduct+hwyrail+TRNSPD)^2,data=xdmgnd\_cas)  
summary(casualties.lm2.boxcox)

##   
## Call:  
## lm(formula = (casualties^L - 1)/L ~ (conduct + hwyrail + TRNSPD)^2,   
## data = xdmgnd\_cas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.07002 -0.42997 -0.00192 0.31091 1.52636   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4112589 0.1177482 3.493 0.000509 \*\*\*  
## conduct1 -0.1060546 0.1235985 -0.858 0.391161   
## hwyrail1 0.1151219 0.2482196 0.464 0.642945   
## TRNSPD 0.0005843 0.0050900 0.115 0.908636   
## conduct1:hwyrail1 -0.2474165 0.2521235 -0.981 0.326776   
## conduct1:TRNSPD 0.0077289 0.0050744 1.523 0.128194   
## hwyrail1:TRNSPD 0.0005574 0.0021991 0.253 0.799970   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5157 on 685 degrees of freedom  
## Multiple R-squared: 0.09559, Adjusted R-squared: 0.08767   
## F-statistic: 12.07 on 6 and 685 DF, p-value: 6.483e-13

Create a step model with casualties transformed:

casualties.lm2.boxcox.step <- step(casualties.lm2.boxcox, trace=T)

## Start: AIC=-909.56  
## (casualties^L - 1)/L ~ (conduct + hwyrail + TRNSPD)^2  
##   
## Df Sum of Sq RSS AIC  
## - hwyrail:TRNSPD 1 0.01709 182.19 -911.50  
## - conduct:hwyrail 1 0.25611 182.43 -910.59  
## <none> 182.17 -909.56  
## - conduct:TRNSPD 1 0.61695 182.79 -909.22  
##   
## Step: AIC=-911.5  
## (casualties^L - 1)/L ~ conduct + hwyrail + TRNSPD + conduct:hwyrail +   
## conduct:TRNSPD  
##   
## Df Sum of Sq RSS AIC  
## - conduct:hwyrail 1 0.24032 182.43 -912.59  
## <none> 182.19 -911.50  
## - conduct:TRNSPD 1 0.60136 182.79 -911.22  
##   
## Step: AIC=-912.59  
## (casualties^L - 1)/L ~ conduct + hwyrail + TRNSPD + conduct:TRNSPD  
##   
## Df Sum of Sq RSS AIC  
## - conduct:TRNSPD 1 0.36552 182.80 -913.20  
## <none> 182.43 -912.59  
## - hwyrail 1 1.18317 183.61 -910.11  
##   
## Step: AIC=-913.2  
## (casualties^L - 1)/L ~ conduct + hwyrail + TRNSPD  
##   
## Df Sum of Sq RSS AIC  
## - conduct 1 0.0182 182.81 -915.13  
## <none> 182.80 -913.20  
## - hwyrail 1 1.2075 184.00 -910.65  
## - TRNSPD 1 17.3269 200.12 -852.53  
##   
## Step: AIC=-915.13  
## (casualties^L - 1)/L ~ hwyrail + TRNSPD  
##   
## Df Sum of Sq RSS AIC  
## <none> 182.81 -915.13  
## - hwyrail 1 1.1947 184.01 -912.63  
## - TRNSPD 1 17.7338 200.55 -853.07

AIC(casualties.lm2.boxcox.step)

## [1] 1050.678

Best AIC yet!

## 4. Recommendation

human factor errors leading to accidents with disproportionately higher amounts of damage and casualties

## 5. References

[1] “Freight Rail Safety Record,” Association of American Railroads. <https://www.aar.org/issue/freight-rail-safety-record/>