Modeling Project

Brooke Beanland uteid: btb949

Data Introduction

The dataset used for this project is an R dataset on factors influencing fatalties in fatal car accidents. Of the variables included in the dataset, main variables include estimated impact speeds, the sex of the occupants, the role of the occupant (i.e driver), whether the occupant was seat belted, whether the car had airbags and whether the airbag was deployed. Furthermore, information such as the year of the accident, year of the car, type of accident (i.e front on conclision), and whether or not the occupant died or not are all variables that may shed light on what factors are most influential on fatal car accident outcomes. The total number of observations in the dataset is 26217, and the total number of variables in the dataset is 15.

```
library(tidyverse)
library(readxl)
library(dplyr)
fatalities <- read.csv("~/Downloads/caraccident.csv")
head(fatalities)
```

```
##   X  dvcat  weight  dead airbag seatbelt frontal sex ageOFocc yearacc yearVeh   abcat occRole
## 1 1  25-39  25.069 alive   none   belted       1   f       26    1997    1990 unavail  driver
## 2 2 24-Oct  25.069 alive airbag   belted       1   f       72    1997    1995  deploy  driver
## 3 3 24-Oct  32.379 alive   none     none       1   f       69    1997    1988 unavail  driver
## 4 4  25-39 495.444 alive airbag   belted       1   f       53    1997    1995  deploy  driver
## 5 5  25-39  25.069 alive   none   belted       1   f       32    1997    1988 unavail  driver
## 6 6  40-54  25.069 alive   none   belted       1   f       22    1997    1985 unavail  driver
##   deploy injSeverity  caseid
## 1      0           3 2:03:01
## 2      1           1 2:03:02
## 3      0           4 2:05:01
## 4      1           1 2:10:01
## 5      0           3 2:11:01
## 6      0           3 2:11:02
```

MANOVA

A MANOVA test was conducted to determine if the numeric variables weight, age of occupant, year of accident, and year of vehicle, displayed mean differences between the categorical variable of level of injury severity. The levels of injury severityy were 0 (no injury), 1 (possible injury), 2(no incapacity), 3(incapacity), 4(death), 5(unknown), and 6(prior death).

```
library(dplyr)
fatalities<-fatalities%>%na.omit
man1<-manova(cbind(weight, ageOFocc, yearacc, yearVeh)~injSeverity, data=fatalities)
summary(man1)
```

```
##                Df   Pillai approx F num Df den Df    Pr(>F)
## injSeverity     1 0.055179   380.46      4  26058 < 2.2e-16 ***
## Residuals   26061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary.aov(man1)
```

```
##  Response weight :
##                Df     Sum Sq    Mean Sq F value    Pr(>F)
## injSeverity     1 2.5041e+09 2504098710  1118.8 < 2.2e-16 ***
## Residuals   26061 5.8328e+10    2238119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response ageOFocc :
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## injSeverity     1   69645   69645  219.11 < 2.2e-16 ***
## Residuals   26061 8283569     318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response yearacc :
##                Df Sum Sq Mean Sq F value    Pr(>F)
## injSeverity     1    133 133.295  46.088 1.155e-11 ***
## Residuals   26061  75373   2.892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response yearVeh :
##                Df Sum Sq Mean Sq F value    Pr(>F)
## injSeverity     1   4826  4825.5  155.26 < 2.2e-16 ***
## Residuals   26061 809958    31.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fatalities%>%group_by(injSeverity)%>%summarise(mean(weight),mean(ageOFocc),mean(yearacc),mean(yearVeh))
```

```
## # A tibble: 7 x 5
##   injSeverity `mean(weight)` `mean(ageOFocc)` `mean(yearacc)` `mean(yearVeh)`
##         <int>          <dbl>            <dbl>           <dbl>           <dbl>
## 1           0          970.              35.1           2000.           1993.
## 2           1          490.              37.3           2000.           1993.
## 3           2          414.              36.0           2000.           1993.
## 4           3          137.              38.5           1999.           1992.
## 5           4           51.2             43.8           1999.           1991.
## 6           5          386.              41.5           2000.           1993.
## 7           6           28.2             62.5           1998            1997
```

```r
pairwise.t.test(fatalities$weight, fatalities$injSeverity, p.adj="none")
```

```
##
```

```
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fatalities$weight and fatalities$injSeverity
##
##   0       1       2       3     4     5
## 1 < 2e-16 -       -       -     -     -
## 2 < 2e-16 0.013   -       -     -     -
## 3 < 2e-16 < 2e-16 < 2e-16 -     -     -
## 4 < 2e-16 < 2e-16 5.2e-13 0.070 -     -
## 5 8.0e-06 0.430   0.835   0.056 0.014 -
## 6 0.372   0.662   0.715   0.918 0.983 0.736
##
## P value adjustment method: none
```

```
pairwise.t.test(fatalities$ageOFocc, fatalities$injSeverity, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fatalities$ageOFocc and fatalities$injSeverity
##
##   0       1       2       3       4       5
## 1 1.2e-11 -       -       -       -       -
## 2 0.00759 0.00048 -       -       -       -
## 3 < 2e-16 6.8e-05 1.1e-13 -       -       -
## 4 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -       -
## 5 3.8e-05 0.00694 0.00047 0.05416 0.14974 -
## 6 0.02930 0.04503 0.03531 0.05645 0.13844 0.09744
##
## P value adjustment method: none
```

```
pairwise.t.test(fatalities$yearacc, fatalities$injSeverity, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fatalities$yearacc and fatalities$injSeverity
##
##   0       1       2       3       4       5
## 1 0.02267 -       -       -       -       -
## 2 0.00105 0.25534 -       -       -       -
## 3 1.4e-10 0.00019 0.02872 -       -       -
## 4 8.3e-07 0.00031 0.00476 0.09100 -       -
## 5 0.08850 0.02965 0.01511 0.00351 0.00076 -
## 6 0.16846 0.18739 0.19861 0.21963 0.24990 0.11489
##
## P value adjustment method: none
```

```
pairwise.t.test(fatalities$yearVeh, fatalities$injSeverity, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
```

```
##
## data:  fatalities$yearVeh and fatalities$injSeverity
##
##   0       1       2       3       4       5
## 1 0.27078 -       -       -       -       -
## 2 0.00371 0.00014 -       -       -       -
## 3 < 2e-16 < 2e-16 5.2e-07 -       -       -
## 4 < 2e-16 < 2e-16 6.8e-15 1.4e-07 -       -
## 5 0.78514 0.96562 0.35649 0.04453 0.00018 -
## 6 0.33254 0.34692 0.29373 0.23659 0.15577 0.35318
##
## P value adjustment method: none
```

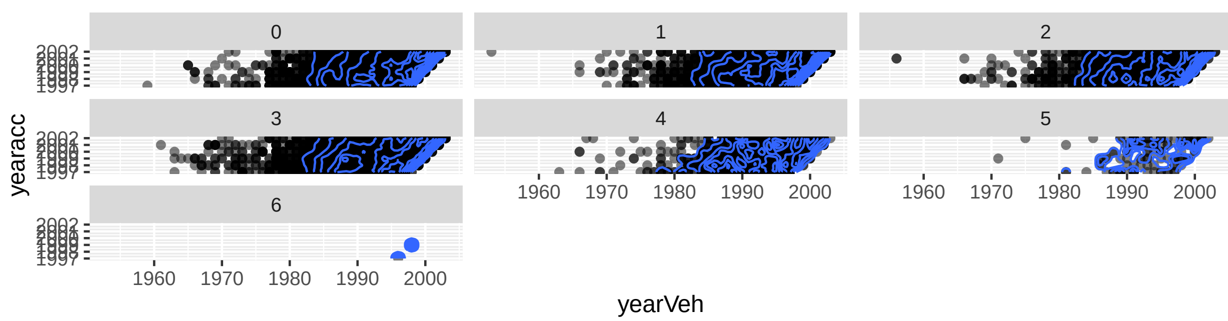
```r
1-0.95^29
```

```
## [1] 0.7740645
```

```r
0.05/29
```

```
## [1] 0.001724138
```

```r
#Assumptions
library(ggExtra)
ggplot(fatalities, aes(x =yearVeh, y = yearacc)) +
geom_point(alpha = .5) + geom_density_2d(h=2) + coord_fixed() + facet_wrap(~injSeverity)
```



The assumptions for conducting a MANOVA were assessed. The random sample with independent observations assumption was likely met due to the nature of the data collected. A DV plot was created to assess DV assumption of normality, and based on the plot shape the assumption of normality failed. The assumption of DV linear relationships may not have been met for the dependent variable of year of accident. Lastly, there is likely univariate and multivariate outliers as well. Multicolineraity was likely not met. Though these assumptions were analyzed theoretically by eye-balling, statistical analysis using specific ggplots and more tests would concretely determine if assumptions were met.

After conducting the MANOVA a significant p-value of < 2.2e-16 was obtained indicating that there was variation in at least one numeric variable across levels of injury severity. Single ANOVA tests were conducted to see which variables displayed between level variation.

With 1 MANOVA, 4 ANOVA, and 4 post hoc test (each with 6 levels), the number of hypothesis tests conduct in total was 29. The likelihood that a type I error occured was calculated to be an 77.41% chance. The adjusted p-value was determined to be 0.0017, and the bonferonni adjustment allowed for appropriate conclusions to be made. The mean weight was significantly different for the injury severities of no injury and no incapacity. The mean age of the occupant was signifcantly different for no injury, possible injury, no incapacity, and incapcity injury severities. The mean year of the accident happening had no significant differences based on severity of injuries. The mean year of the vechicle driven during the car accident was significantly different for possible injury, no incapacity, and incapacity.

Randomization Testing

A randomization test was conducted to determine if there was a significant mean difference in age based on whether or not the occupant died or lived in the crash. The randomization was conducted 5000 times, and p-values were analyzed.

```r
library(dplyr)
#conducting the t-test
fatalities%>%group_by(dead)%>%summarise(mean(ageOFocc))
```

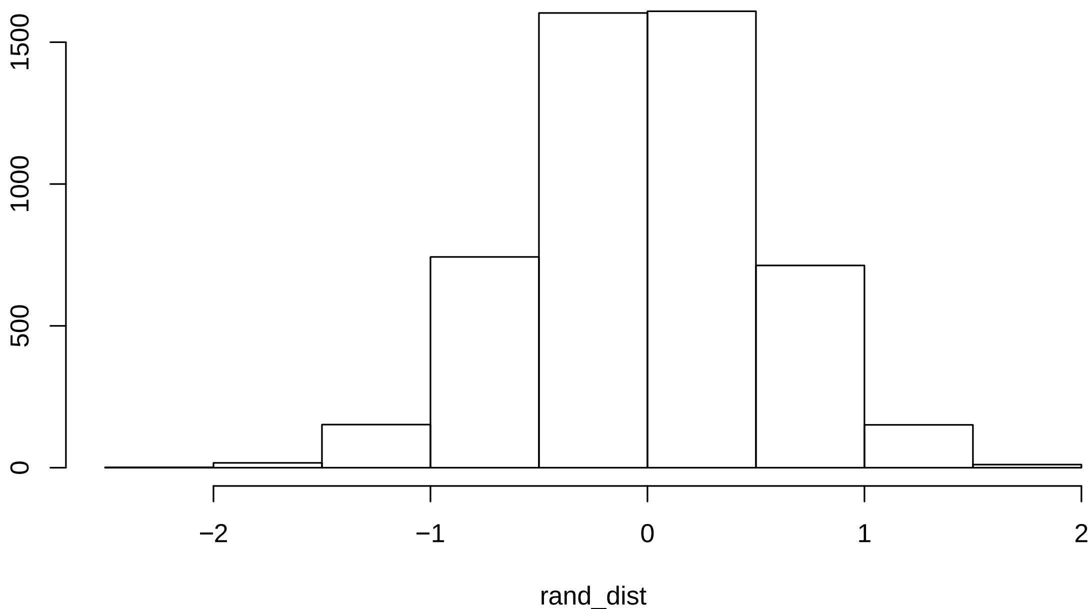
```
## # A tibble: 2 x 2
##   dead  `mean(ageOFocc)`
##   <fct>            <dbl>
## 1 alive             36.9
## 2 dead              44.6
```

```r
t.test(data=fatalities, ageOFocc~dead)
```

```
##
##  Welch Two Sample t-test
##
## data:  ageOFocc by dead
## t = -12.276, df = 1256.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.979871 -6.505109
## sample estimates:
## mean in group alive  mean in group dead
##            36.87276            44.61525
```

```r
#Randomization
rand_dist<-vector()
for(i in 1:5000){
new<-data.frame(age=sample(fatalities$ageOFocc),condition=fatalities$dead)
rand_dist[i]<-mean(new[new$condition=="dead",]$age)-
mean(new[new$condition=="alive",]$age)}


hist(rand_dist,main="",ylab=""); abline(v = -7.75824,col="red")
```




```r
mean(rand_dist>  7.75824| rand_dist< - 7.75824)
```

```
## [1] 0
```

Mean difference test was conducted to determine if the mean age of occupants that died during the car accident is different than the mean of occupants that lived. Null Hypothesis: Mean age of occupant is the same for those classified as dead or alive. Alternative Hypothesis: Mean age of occupant is different for those classified as dead versus alive. The difference in mean age of dead or alive was calculated to be 7.75824. Based on the results of the randomization test, the p-value calculated using a two-tail calculation was 0. This would cause a failure to reject the null hypothesis because the p-value is greater than 0.05. This indicates that the randomization concluded the mean differences in age between dead and alive were the same. When conducting the actual welch t-test, the p-value is very small < 2.2e-16 causing a rejection of the null hypothesis which indicates the means are different.

Linear Regression

A linear regression model was created to see if age of the occupant and sex were predictive of injury severity sustained in the car accdient.

```
library(lmtest)
library(dplyr)
fatalities<-fatalities%>%na.omit()
head(fatalities)
```

```
##   X  dvcat  weight  dead airbag seatbelt frontal sex ageOFocc yearacc yearVeh   abcat occRole
## 1 1  25-39  25.069 alive   none   belted       1   f       26    1997    1990 unavail  driver
## 2 2 24-Oct  25.069 alive airbag   belted       1   f       72    1997    1995  deploy  driver
## 3 3 24-Oct  32.379 alive   none     none       1   f       69    1997    1988 unavail  driver
## 4 4  25-39 495.444 alive airbag   belted       1   f       53    1997    1995  deploy  driver
## 5 5  25-39  25.069 alive   none   belted       1   f       32    1997    1988 unavail  driver
## 6 6  40-54  25.069 alive   none   belted       1   f       22    1997    1985 unavail  driver
##   deploy injSeverity  caseid
## 1      0           3 2:03:01
## 2      1           1 2:03:02
## 3      0           4 2:05:01
## 4      1           1 2:10:01
## 5      0           3 2:11:01
## 6      0           3 2:11:02
```

```
fatalities$age_c <- fatalities$ageOFocc - mean(fatalities$ageOFocc)
any(is.na(fatalities))
```

```
## [1] FALSE
```

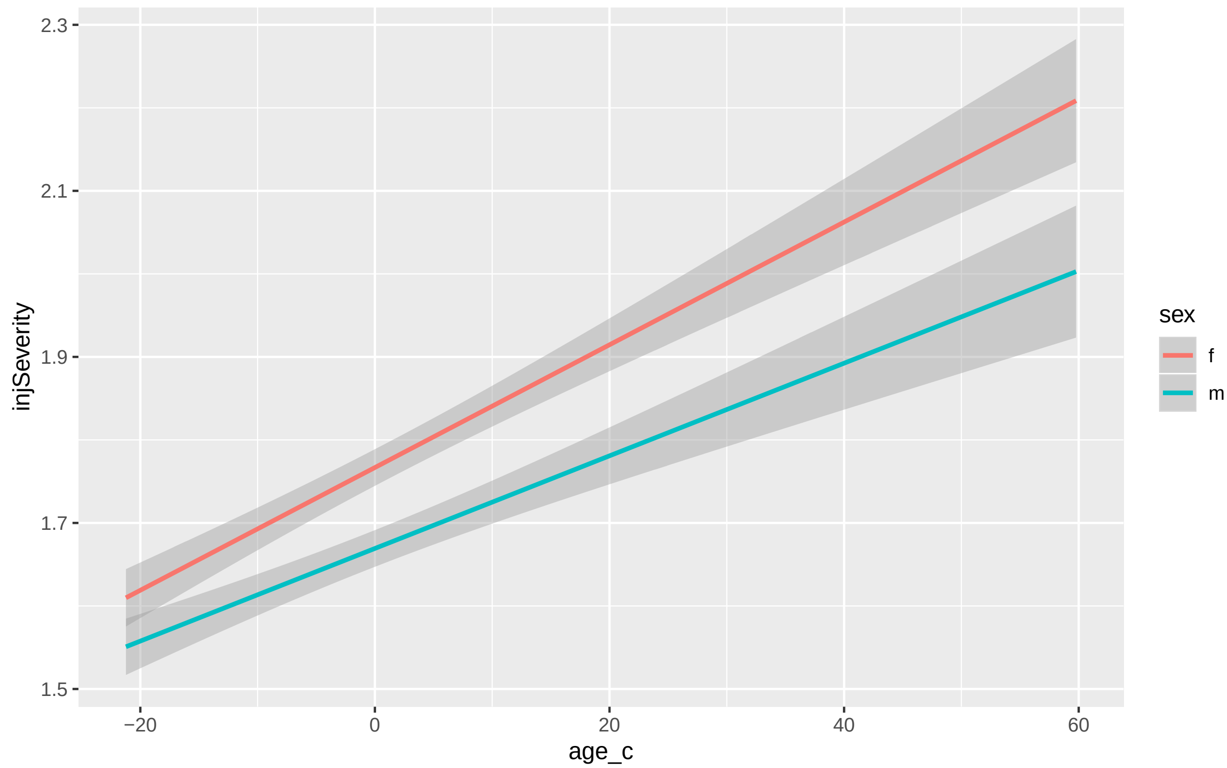
```
fatalities$injSeverity<-as.numeric(fatalities$injSeverity)
fatalities<-fatalities%>%na.omit

fit<-lm(injSeverity~age_c*sex, data = fatalities)
summary(fit)
```

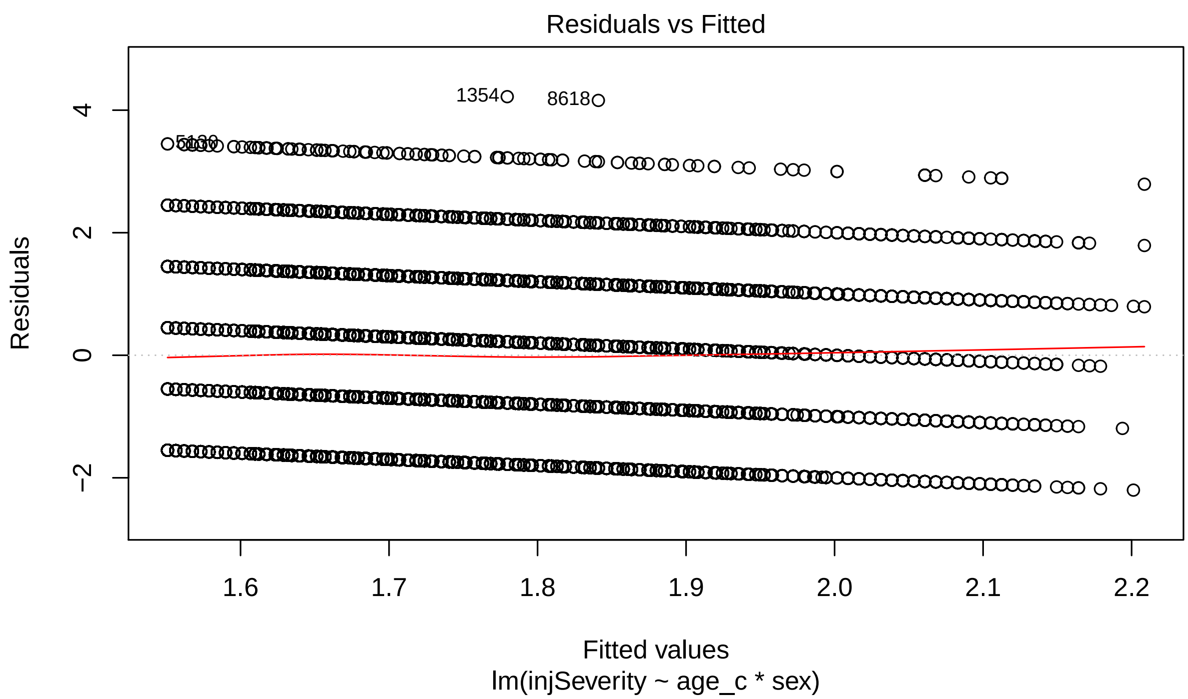
```
##
## Call:
## lm(formula = injSeverity ~ age_c * sex, data = fatalities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2012 -1.0977  0.2127  1.2645  4.2204
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7667397  0.0116767 151.305  < 2e-16 ***
## age_c        0.0073922  0.0006368  11.608  < 2e-16 ***
## sexm        -0.0974841  0.0159963  -6.094 1.12e-09 ***
## age_c:sexm  -0.0018135  0.0008918  -2.034    0.042 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.287 on 26059 degrees of freedom
## Multiple R-squared:  0.00991,    Adjusted R-squared:  0.009796
## F-statistic: 86.94 on 3 and 26059 DF,  p-value: < 2.2e-16
```

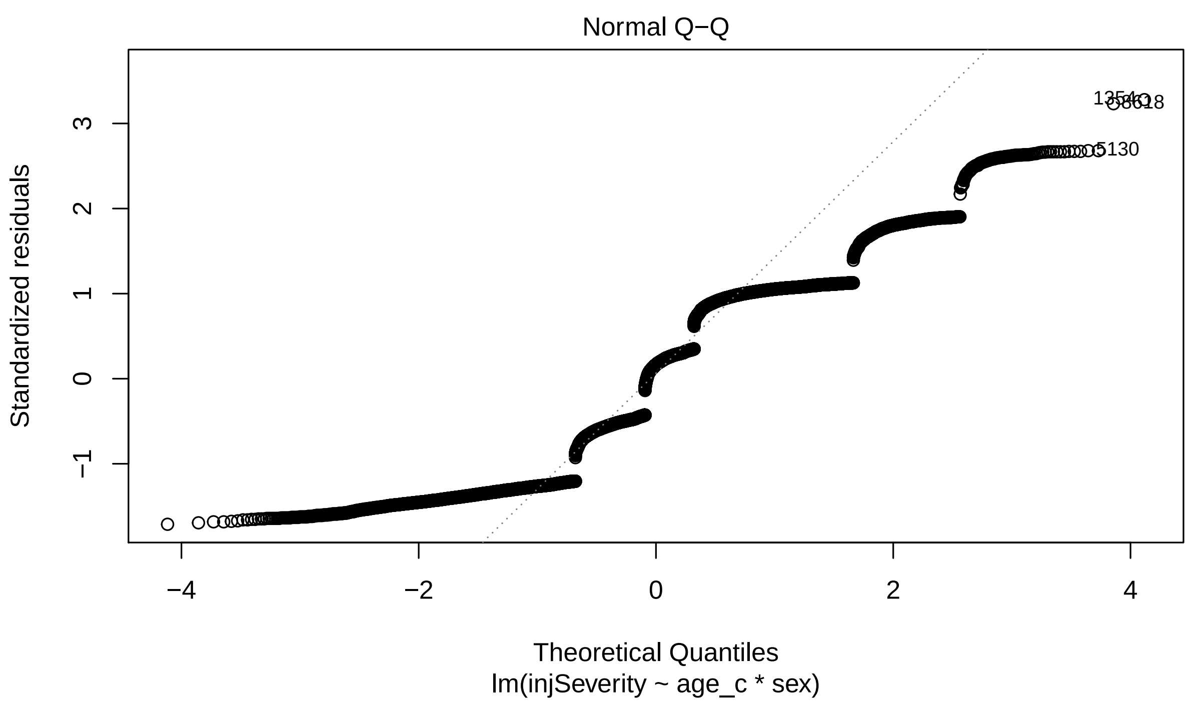
```
#graphical representation of regression model
library(ggplot2)
ggplot(fatalities,aes(y=injSeverity,x=age_c,color=sex))+geom_smooth(method="lm")
```



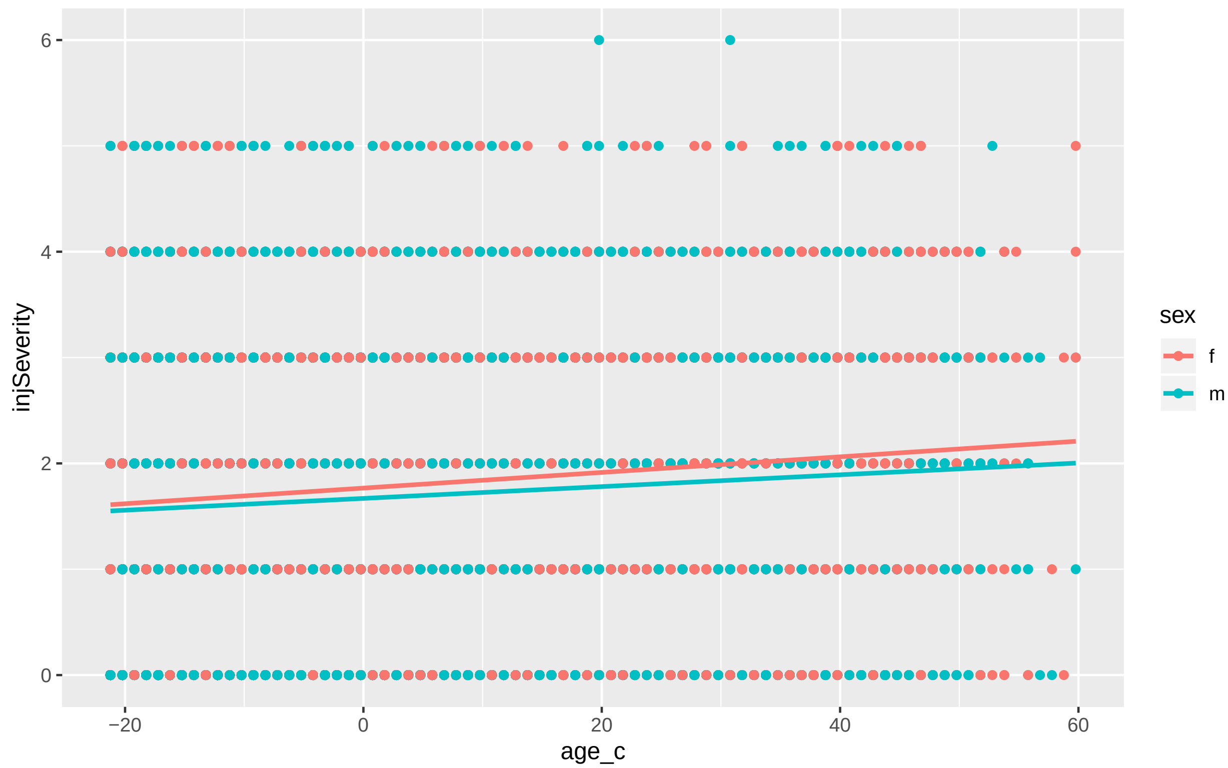

```
#Checking Assumptions
#linear- not met
plot(fit, 1)
```




```
#normality
plot(fit, 2)
```



```r
#homoskedastically- not met
ggplot(fatalities,aes(y=injSeverity,x=age_c,color=sex))+geom_point()+stat_smooth(method="lm",se=FALSE)
```




```r
#Robust standard errors
library(sandwich)
library(lmtest)
coeftest(fit, vcov=vcovHC(fit))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  1.76673971  0.01124763 157.0767 < 2.2e-16 ***
## age_c        0.00739220  0.00062400  11.8464 < 2.2e-16 ***
## sexm        -0.09748408  0.01592761  -6.1204 9.464e-10 ***
## age_c:sexm  -0.00181351  0.00090078  -2.0133    0.0441 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Assumptions were assessed graphically for homoskedasticity. By viewing the graph, it can be observed tha
After running the regression model to see the relationship of age of occupant and sex, as well as the in
By conducting a new regression with robust standard errors there were still 3 significant p-values. Spe
Overall, the regression model looking at sex and age as predictors of injury severity of the occupants i
```

Bootstrapped Linear Regression

The same linear regression model was completed using bootstrapped standard errors and differences were discussed.

```
fit1<-lm(injSeverity~age_c*sex, data = fatalities)
resids<-fit1$residuals
fitted<-fit1$fitted.values
resid_resamp<-replicate(5000,{
new_resids<-sample(resids,replace=TRUE)
fatalities$new_y<-fitted+new_resids
fit1<-lm(new_y~age_c*sex,data=fatalities)
coef(fit1)
})
coef(fit1)
```

```
##  (Intercept)        age_c         sexm   age_c:sexm
##  1.766739706  0.007392204 -0.097484080 -0.001813513
```

```
summary(fit1)
```

```
##
## Call:
## lm(formula = injSeverity ~ age_c * sex, data = fatalities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2012 -1.0977  0.2127  1.2645  4.2204
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7667397  0.0116767 151.305  < 2e-16 ***
## age_c        0.0073922  0.0006368  11.608  < 2e-16 ***
## sexm        -0.0974841  0.0159963  -6.094 1.12e-09 ***
## age_c:sexm  -0.0018135  0.0008918  -2.034    0.042 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.287 on 26059 degrees of freedom
## Multiple R-squared:  0.00991,    Adjusted R-squared:  0.009796
## F-statistic: 86.94 on 3 and 26059 DF,  p-value: < 2.2e-16
```

```
resid_resamp%>%t%>%as.data.frame%>%summarize_all(sd)
```

```
##   (Intercept)        age_c       sexm   age_c:sexm
## 1  0.01145936 0.0006332988 0.01594218 0.0008961404
```

```
coeftest(fit, vcov=vcovHC(fit))
```

```
##
## t test of coefficients:
```

```
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  1.76673971  0.01124763 157.0767 < 2.2e-16 ***
## age_c        0.00739220  0.00062400  11.8464 < 2.2e-16 ***
## sexm        -0.09748408  0.01592761  -6.1204 9.464e-10 ***
## age_c:sexm  -0.00181351  0.00090078  -2.0133    0.0441 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = injSeverity ~ age_c * sex, data = fatalities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2012 -1.0977  0.2127  1.2645  4.2204
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7667397  0.0116767 151.305  < 2e-16 ***
## age_c        0.0073922  0.0006368  11.608  < 2e-16 ***
## sexm        -0.0974841  0.0159963  -6.094 1.12e-09 ***
## age_c:sexm  -0.0018135  0.0008918  -2.034    0.042 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.287 on 26059 degrees of freedom
## Multiple R-squared:  0.00991,    Adjusted R-squared:  0.009796
## F-statistic: 86.94 on 3 and 26059 DF,  p-value: < 2.2e-16
```

Analyzing the new SEs from the bootstrapped model, the intercept SE has been reduced slightly from the original model to 0.01497 as it was previously 0.0150. The other standard errors stayed essentially the same between the models compared to the bootstrap model. In addition, the p-values stayed the same as well as the the significance cutoffs obtained from both the original model and robust errors model when compared to the bootstrapped model. In other words, there was no change in significance.

Logistic Regression

A logistic regression was conducted to explore the relationship of occupant role and frontal crashes on whether the occupant lived or died.

```r
library(tidyverse)
library(dplyr)
library(lmtest)
fatalities<-fatalities%>%mutate(y=ifelse(dead=="dead",1,0))
head(fatalities)
```

```
##   X  dvcat  weight  dead airbag seatbelt frontal sex ageOFocc yearacc yearVeh   abcat occRole
## 1 1  25-39  25.069 alive   none   belted       1   f       26    1997    1990 unavail  driver
## 2 2 24-Oct  25.069 alive airbag   belted       1   f       72    1997    1995  deploy  driver
```

```
## 3 3 24-Oct  32.379 alive   none     none       1    f       69    1997    1988 unavail  driver
## 4 4  25-39 495.444 alive airbag   belted       1    f       53    1997    1995  deploy  driver
## 5 5  25-39  25.069 alive   none   belted       1    f       32    1997    1988 unavail  driver
## 6 6  40-54  25.069 alive   none   belted       1    f       22    1997    1985 unavail  driver
##   deploy injSeverity  caseid      age_c y
## 1      0           3 2:03:01 -11.223305 0
## 2      1           1 2:03:02  34.776695 0
## 3      0           4 2:05:01  31.776695 0
## 4      1           1 2:10:01  15.776695 0
## 5      0           3 2:11:01  -5.223305 0
## 6      0           3 2:11:02 -15.223305 0
```

```
fatalities<-fatalities%>%na.omit()
fitl<-glm(y~occRole+frontal, data=fatalities, family=binomial(link="logit"))
coeftest(fitl)
```

```
##
## z test of coefficients:
##
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.718107   0.045939 -59.1676  < 2e-16 ***
## occRolepass  0.171240   0.069676   2.4577  0.01398 *
## frontal     -0.644540   0.059769 -10.7838  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(coef(fitl))
```

```
## (Intercept) occRolepass     frontal
##  0.06599956  1.18677528  0.52490372
```

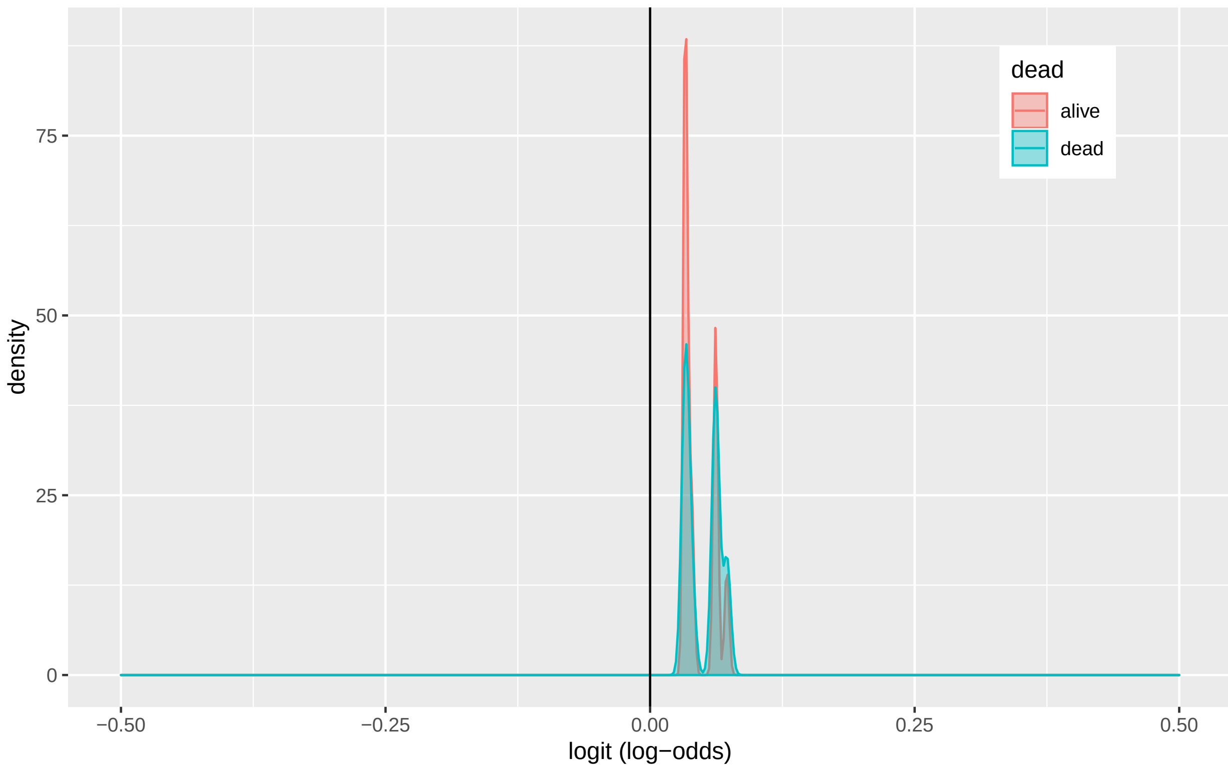
```
#confusion matrix
probs<-predict(fitl,type="response")
table(truth=fatalities$dead,predict=as.numeric(probs>.5))%>%addmargins
```

```
##        predict
## truth       0   Sum
##   alive 24883 24883
##   dead   1180  1180
##   Sum   26063 26063
```

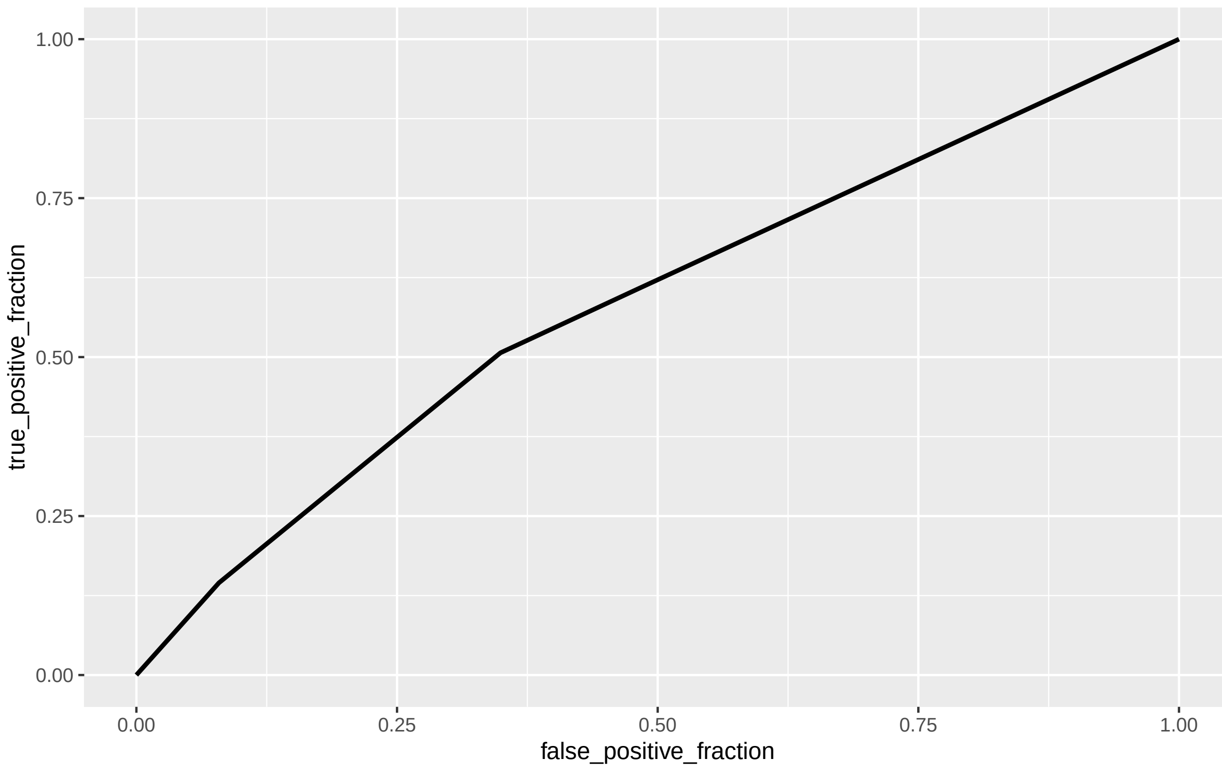
```
14969/15677
```

```
## [1] 0.9548383
```

```
#Density Plot
fatalities$logit<-predict(fitl, type = "link")
fatalities%>%ggplot()+geom_density(aes(probs,color=dead,fill=dead), alpha=.4)+theme(legend.position=c(.8
```




```
#ROC
library(plotROC)
probs<-predict(fitl,type="response")
ROCplot<-ggplot(fatalities)+geom_roc(aes(d=y,m=probs), n.cuts=0)
ROCplot
```




```
calc_auc(ROCplot)
```

```
##   PANEL group      AUC
## 1     1    -1 0.584216
```

```
#CV
class_diag <- function(probs,truth){
tab<-table(factor(probs>.5,levels=c("FALSE","TRUE")),truth)
acc=sum(diag(tab))/sum(tab)
sens=tab[2,2]/colSums(tab)[2]
spec=tab[1,1]/colSums(tab)[1]
ppv=tab[2,2]/rowSums(tab)[2]
if(is.numeric(truth)==FALSE & is.logical(truth)==FALSE) truth<-as.numeric(truth)-1

ord<-order(probs, decreasing=TRUE)
probs <- probs[ord]; truth <- truth[ord]
TPR=cumsum(truth)/max(1,sum(truth))
FPR=cumsum(!truth)/max(1,sum(!truth))
dup<-c(probs[-1]>=probs[-length(probs)], FALSE)
TPR<-c(0,TPR[!dup],1); FPR<-c(0,FPR[!dup],1)
n <- length(TPR)
auc<- sum( ((TPR[-1]+TPR[-n])/2) * (FPR[-1]-FPR[-n]) )
data.frame(acc,sens,spec,ppv,auc)
}

set.seed(1234)
k=10
fatalities<-fatalities[sample(nrow(fatalities)),]
```

```r
folds<-cut(seq(1:nrow(fatalities)),breaks=k,labels=F)
diags<-NULL
for(i in 1:k){
train<-fatalities[folds!=i,]
test<-fatalities[folds==i,]
truth<-test$y
fit<-glm(y~occRole+frontal,data=fatalities,family="binomial")
probs<-predict(fit,newdata = test,type="response")
diags<-rbind(diags,class_diag(probs,truth))
}
diags%>%summarize_all(mean)
```

```
##         acc sens spec ppv       auc
## 1 0.9547249    0    1 NaN 0.5842944
```

After running the logisitic regression, the coefficients were interpretted in context. The odds of death for passengers in the car accident, controling for type of crash, are 1.1867 times higher than that of the driver. Further, the odds of death when involved in a frontal car accident, controling for occupant role, are 0.5249 times higher than non-frontal crashes. The intercept is interpreted to communicate that the odds of dying in a car accident for the driver when frontal=0 (not a frontal crash), based on the data studied here, is 0.065. The confusion matrix produced informs the viewer on the Accuracy, Sensitivity (TPR), Specificity (TNR), and Recall (PPV) of the model. The Accuracy of the model = 95.48% which indicates that 95 percent of the cases were correctly classified. The Sensitivity (TPR) of the model = 0 which indicates the proportion of deaths correctly classified as death. The Specificity (TNR) of the model = 0 which indicates the proportion of living cases correctly classified as living. The PPV of this model would be 0 because that describes the proportion classified as dead that were actually dead, and there were no predicitions of dead (1), (no p>.5). An ROC curve was generated and the AUC value was calculated to be 0.5883. This is a bad AUC value because it communicates that the test is only slightly better at predicting the correct outcome than a completely uninformative test. A 50/50 chance at correct prediction would produce a straight line, and as seen in the ROC curve for this model the line only slightly deviates from the straight line. Ultimately, this is a bad ROC curve and bad AUC value. A 10-fold cross-validation test was conducted, and the AUC values stayed virtually the same whith the CV AUC coming out to 0.5863. This, again, is a bad AUC value indicating that the model is a poor predictor of the outcome of death. The sensitivity of the CV model was 0, indicating that there were zero deaths correctly classified as deaths by the CV model. The accuracy by the CV model was 95.45% which is similar to the values indicated in the original regression model's confusion matrix. The ppv was reported as NA by the cv model.

LASSO

```r
library(glmnet)
fatalities<-fatalities%>%select(!logit)
fatalities<-fatalities%>%select(!y)
fatalities<-fatalities%>%select(!caseid)

y<-as.matrix(fatalities$frontal)
x<-model.matrix(frontal~.,data=fatalities)[,-1]
head(x)
```

```
##           X dvcat24-Oct dvcat25-39 dvcat40-54 dvcat55+   weight deaddead airbagnone seatbeltnone
## 7452   7487           0          1          0        0   23.427        0          1            1
```

```
## 8016   8058           0          1          0        0    3.856          0          0             0
## 7162   7197           0          1          0        0   24.166          0          1             0
## 8086   8128           1          0          0        0   41.323          0          1             0
## 23653 23794           0          1          0        0  832.725          0          0             0
## 9196   9246           1          0          0        0 4826.845          0          0             0
##       sexm ageOFocc yearacc yearVeh abcatnodeploy abcatunavail occRolepass deploy injSeverity
## 7452     1       20    1998    1994             0            1           1      0           3
## 8016     1       20    1998    1993             1            0           0      0           1
## 7162     1       16    1998    1985             0            1           0      0           1
## 8086     0       18    1998    1985             0            1           1      0           1
## 23653    1       16    2002    2000             1            0           1      0           0
## 9196     0       82    1999    1994             0            0           0      1           2
##           age_c
## 7452  -17.22331
## 8016  -17.22331
## 7162  -21.22331
## 8086  -19.22331
## 23653 -21.22331
## 9196   44.77669
```

```
x<-scale(x)
cv<-cv.glmnet(x,y,family="binomial")
lasso<-glmnet(x,y,family="binomial",lambda=cv$lambda.1se)
coef(lasso)
```

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept)    0.6490732949
## X              0.0471184833
## dvcat24-Oct    .
## dvcat25-39     .
## dvcat40-54     0.0118958137
## dvcat55+       0.0801514071
## weight         .
## deaddead      -0.1366036806
## airbagnone     .
## seatbeltnone   0.1084867366
## sexm           0.0666511669
## ageOFocc      -0.0053210387
## yearacc        .
## yearVeh        0.0593401189
## abcatnodeploy -0.6368240757
## abcatunavail   .
## occRolepass   -0.0265909127
## deploy         0.2410075980
## injSeverity   -0.1548721358
## age_c         -0.0001612575
```

```
#cross-validating lasso model
set.seed(1234)
k=10
data <- fatalities %>% sample_frac
folds <- ntile(1:nrow(data),n=10)
```

```r
diags<-NULL
for(i in 1:k){
train <- data[folds!=i,]
test <- data[folds==i,]
truth <- test$frontal
fit <- glm(frontal~`dvcat`+`dvcat`+`weight`+`dead`+`seatbelt`+`sex`+`yearVeh`+`abcat`+`occRole`+`deploy
probs <- predict(fit, newdata=test, type="response")
diags<-rbind(diags,class_diag(probs,truth))
}
diags%>%summarize_all(mean)
```

```
##         acc      sens      spec       ppv       auc
## 1 0.7181448 0.8794158 0.4266079 0.7348388 0.7197367
```

Upon conducting the LASSO on predictors of frontal crashes, a binary response variable, there were a good amount of variables retained. For speed of impact, categories 25-39, 40-54, and 55+ mph were all retained as predictive variables of whether a crash was frontal or not. Further, weight, dead, seatbelt, sex, age of occupant, year of the vehicle, the airbag deploying or not, the role of the occupant, and injury severity 1 and 3 were all retained as predictive variables. Most of these variables are inuitive, for example whether the airbag deployed or not seems logical that is would be a predictor of whether the accident was a frontal crash. Note: X and afe_c were not included in the following CV as age_c would be a redundant predictor, and X is a variable indicating the observation number. To see how this model held, cross-validation was conducted. The cross validation out-of-sample accuracy was 0.7178 which is lower than the accuracy observed from the previous cross validation in question 5 of 0.9547. Interesting, the AUC of the lasso cross-validation increased measurably to 0.7197 classifying the model as fair. This is greatly different than the "bad"" model from question 5 that had an AUC value of 0.58.