BSAN 450 Assignment 7

1) For this problem consider the data in the file crash.csv. This is the data referenced on page 298 of the text in problem 45. The file contains the following information.

MAKE: The make of the car

MODEL: The model of the car.

HEAD\_INJ: the head injury severity

CHEST\_IN: the chest injury severity

LLEG\_INJ: the severity of the injury to the left leg

RLEG\_INJ: the severity of the injury to the right leg

DRIV\_PAS: a categorical variable = Driver if the driver was injured and Passen if the passenger was injured.

PROTECT: a categorical variable indicating the mode of protection in the crash: d airbag, d&p airbags, manual belts, Motorized belts, passive belts

DOORS: a categorical variable with the following categories: 2, 4 and -. The – indicates some type of truck.

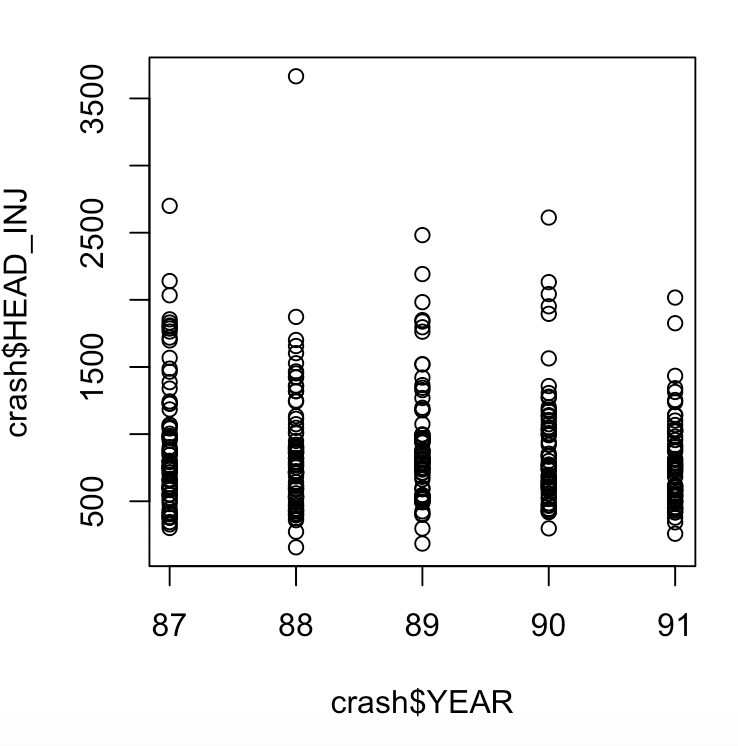
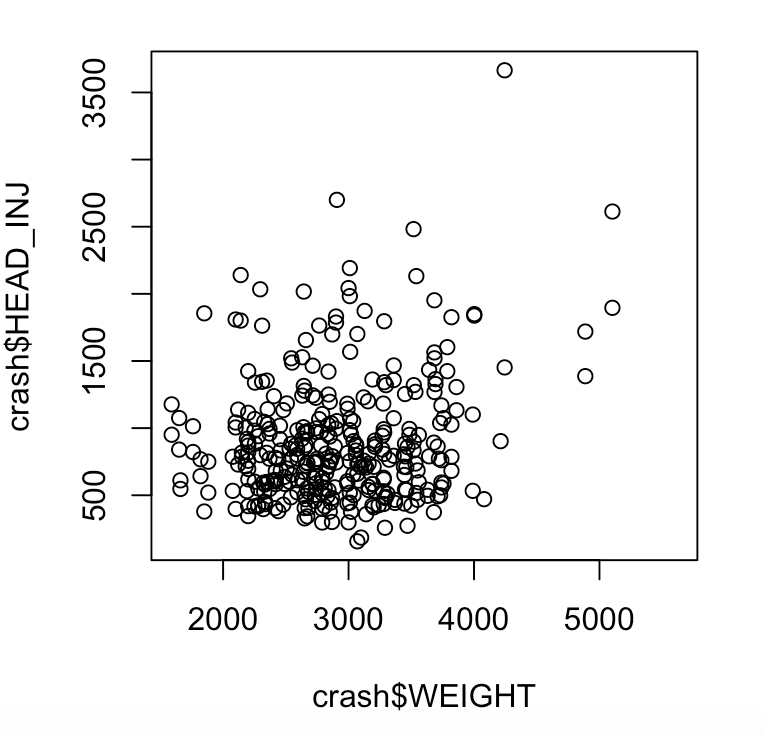
YEAR: The model year of the car.

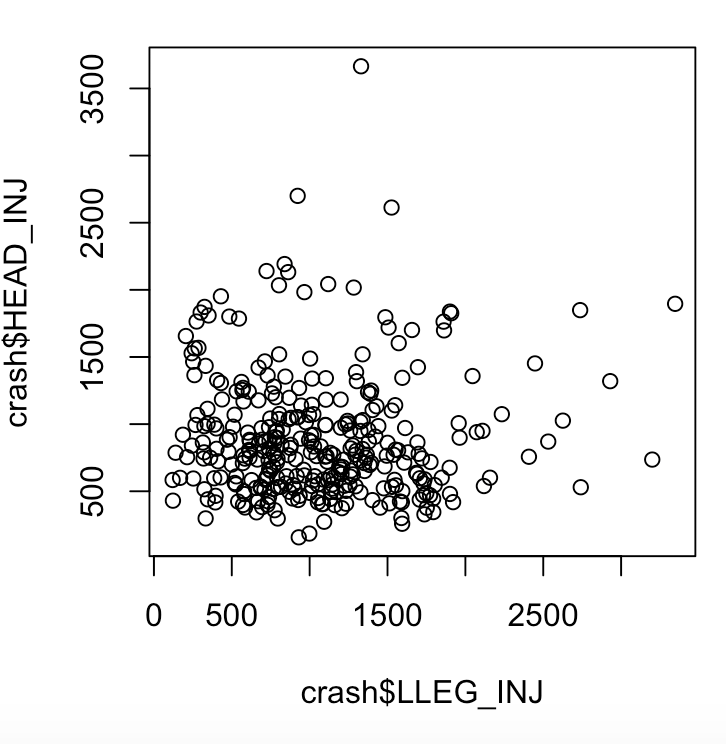
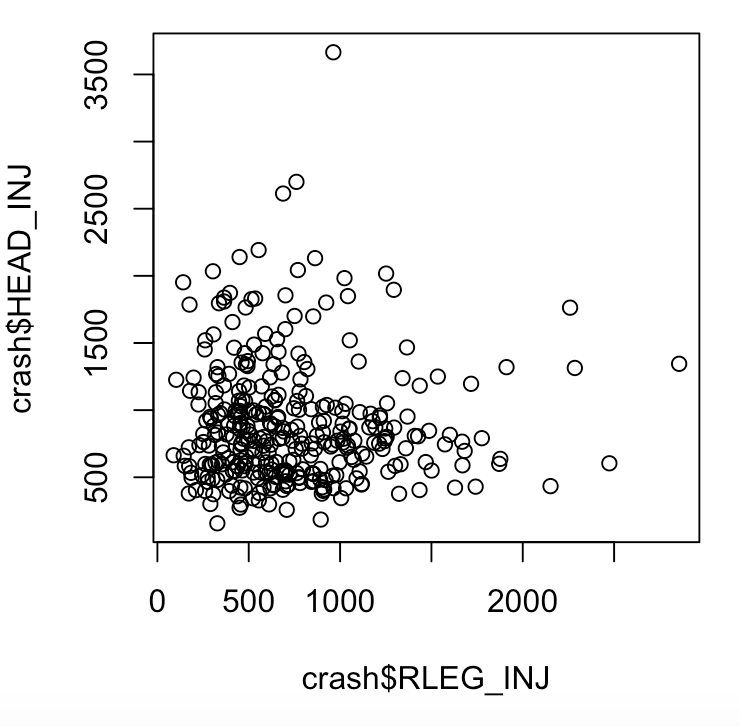
WEIGHT: The weight of the car.

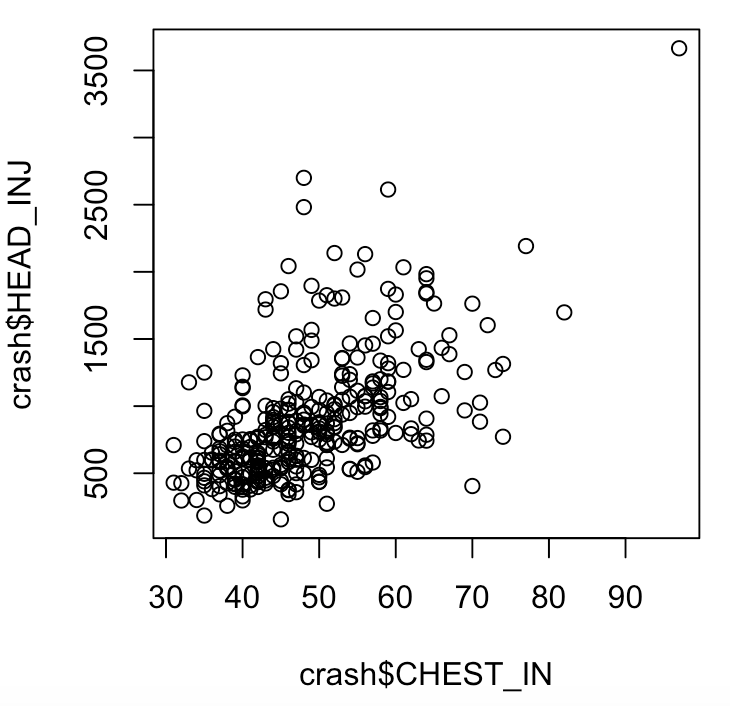
SIZE: a categorical variable indicating the size of the vehicle: comp, hev, lt, med, mini, mpv, pu, and van.

The goal is to develop a model to predict HEAD\_INJ, the severity of the head injury in these crashes.

a) There are 5 continuous variables that are available: CHEST-INJ, LLEG\_INJ, RLEG\_ING, YEAR, AND WEIGHT. Plot scatter plots of the variable HEAD\_INJ versus these 5 continuous variables. Comment on the plots, does it appear from these plots that there is a relationship between these variables and HEAD\_INJ?

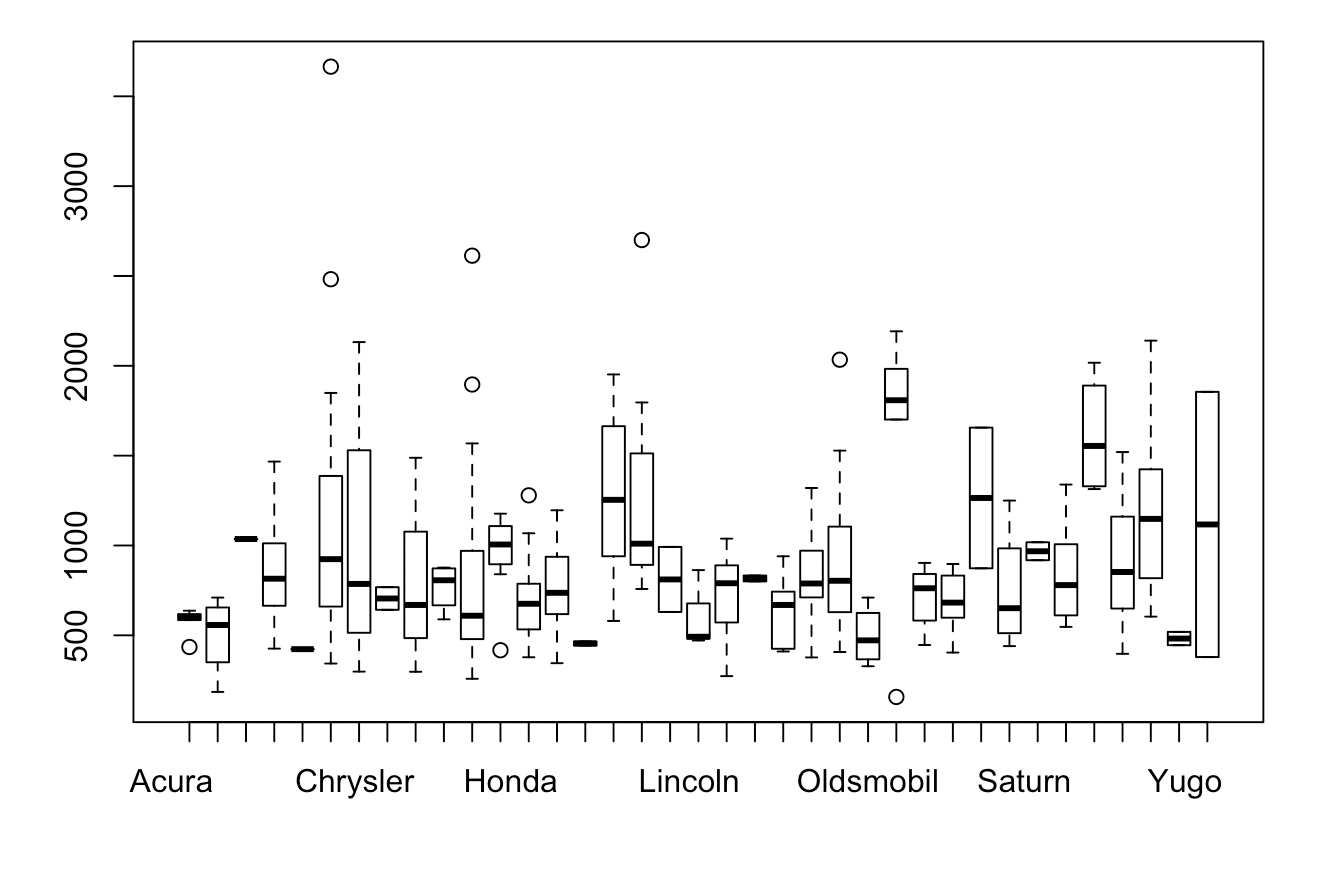
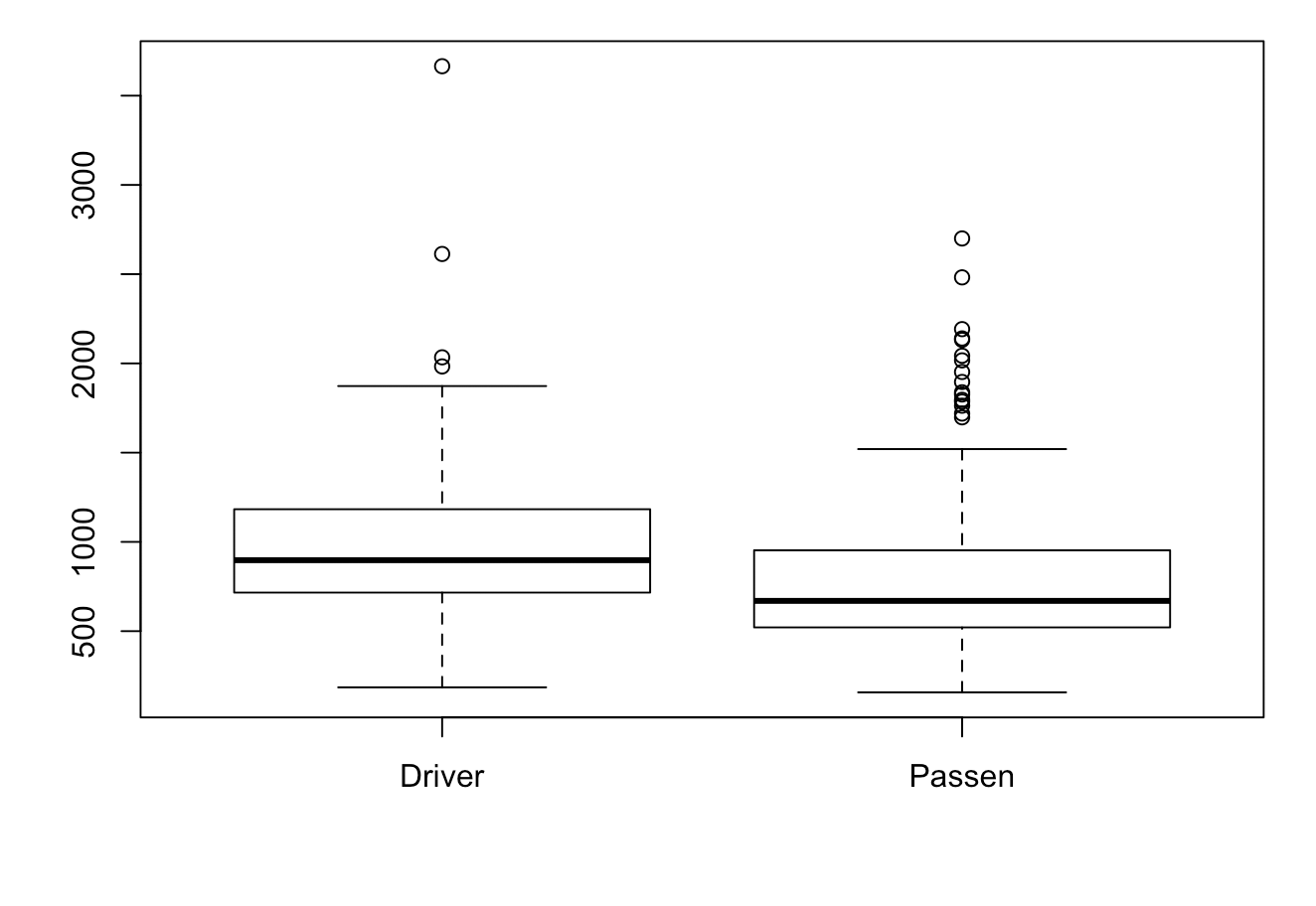
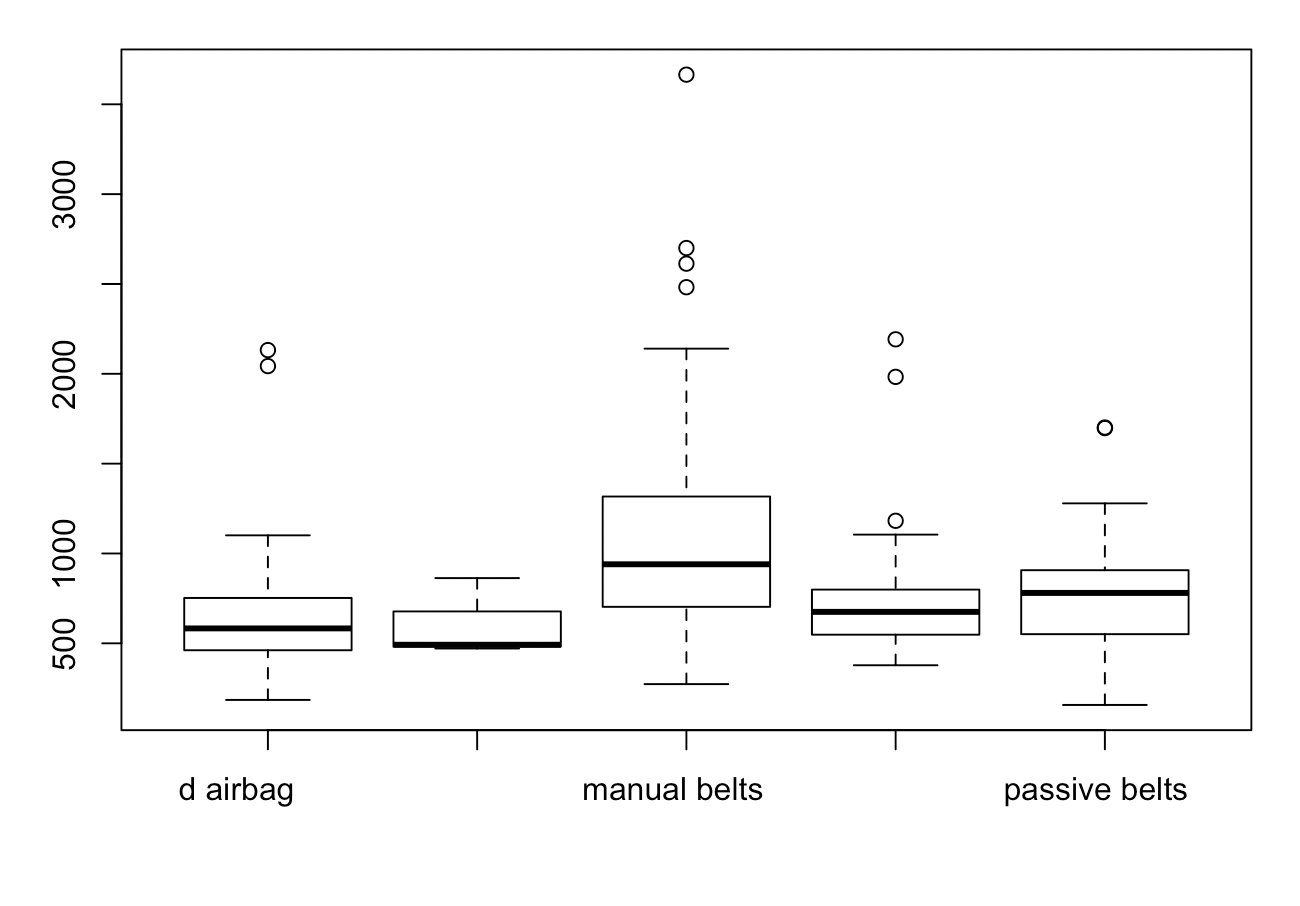


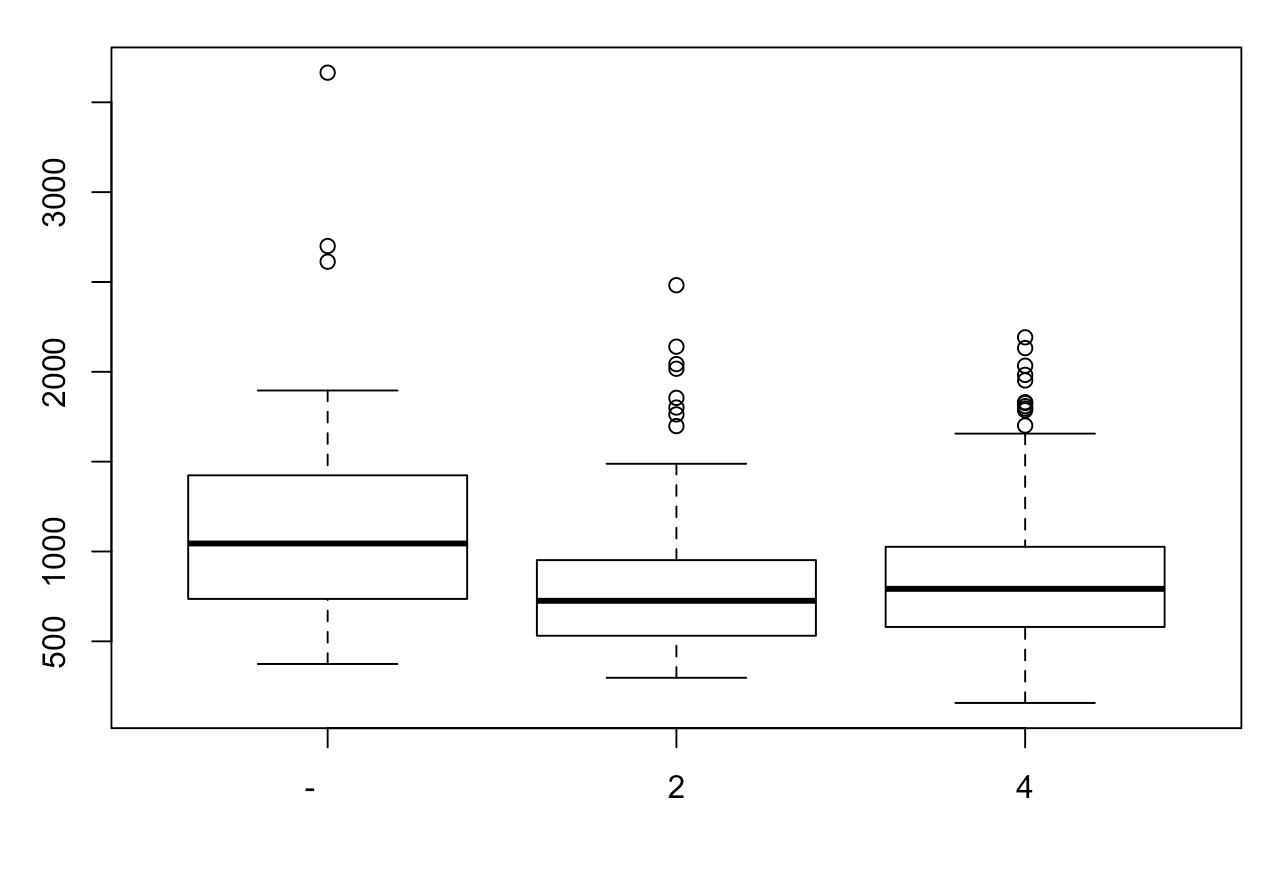
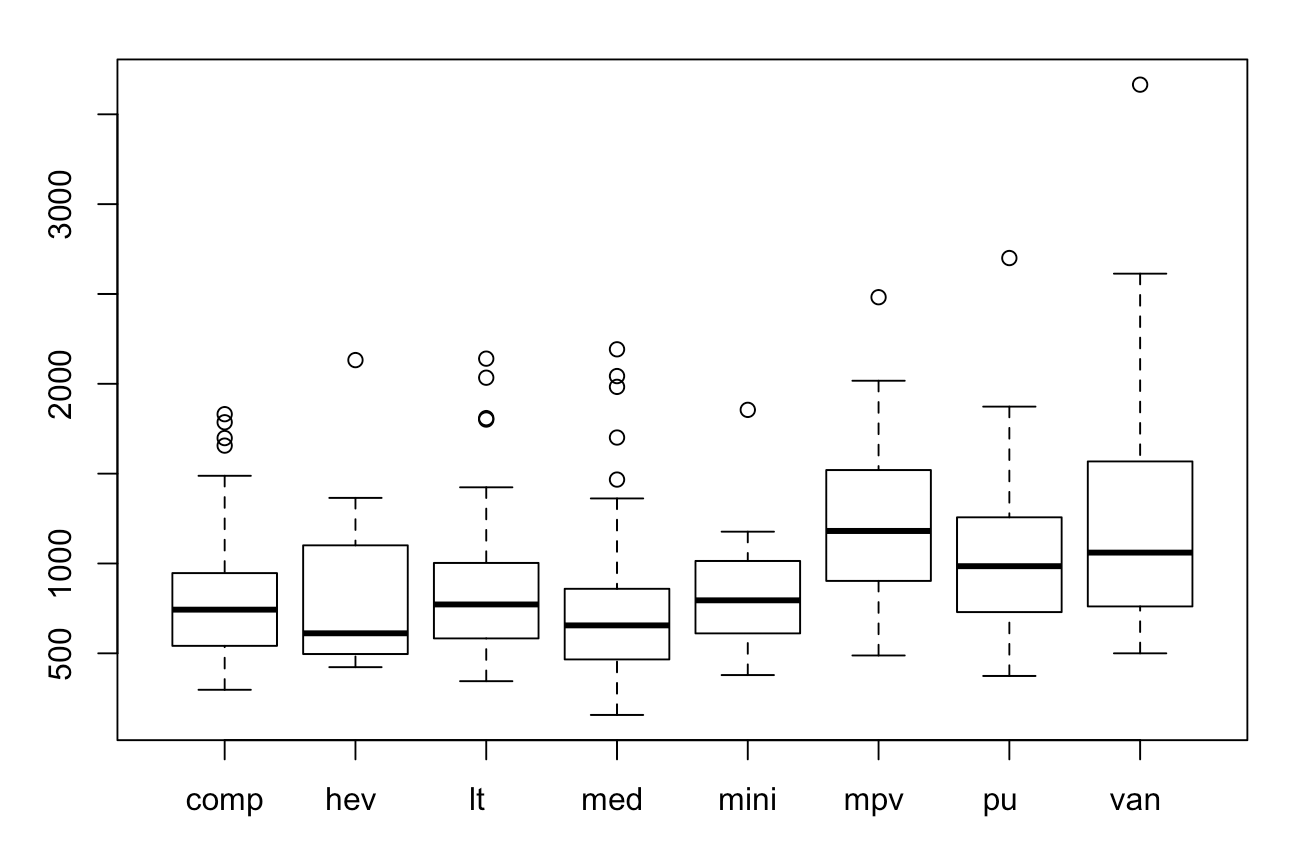




There appears to only be a noticeable relationship between head\_inj and chest\_in.

b) The following variables are categorical variables: MAKE, DRIV\_PAS, PROTECT, DOORS, and SIZE. Plot boxplots of HEAD\_INJ for the different categories in each of these categorical variables. Does it appear that any of these categorical variables have an impact upon HEAD\_INJ?

It appears that size and make of car might have an impact on head\_inj.

c) The number of groups for some of the categorical variables is large so that we would need to create a large number of indicator variables to include those categorical variables in a regression model. This can be an issue when using a variable selection method like best subsets regression because these methods add one variable at a time to the model. For example consider the categorical variable SIZE which has 8 groups. In order to capture all the information in this categorical variable, 7 indicator variables would need to be created. In order to judge the impact of the variable SIZE, one should consider the impact of adding all of the 7 indicator variables to the model; however all of the variable selection methods add one of these indicator variables at a time to the model. In addition, an indicator variable with a large number of categories adds many possible variables to those under consideration.

In order to deal with this issue for this problem, we will use the best subsets regression method using only the continuous variables to determine a model for these variables. After this is done, we will consider adding categorical variables to the model.

You will use 10 fold cross validation along with the best subsets regression method to help determine which continuous variables should be used in the regression model. Note that since this involves R several complex R commands the R code is given below. In order for these commands to make sense you will need to read the data into a data frame named crash.

Notice that the leaps package must be called. In addition if there are any missing values in the data, is will cause problems, thus the second command removes missing data from the data frame named crash and stores these in the data frame named crash1.

library(leaps)

crash1=na.omit(crash)

# use 10 fold cross validation

set.seed(11)

folds=sample(rep(1:10,length=nrow(crash1)))

predict.regsubsets=function(object,newdata,id,...){

form=as.formula(object$call[[2]])

mat=model.matrix(form,newdata)

coefi=coef(object,id=id)

mat[,names(coefi)]%\*%coefi

}

cv.errors=matrix(NA,10,5)

for(k in 1:10){

best.fit=regsubsets(HEAD\_INJ~CHEST\_IN+LLEG\_INJ+RLEG\_INJ+WEIGHT+YEAR,data=crash1[folds!=k,],nvmax=5)

for(i in 1:5){

pred=predict(best.fit,crash1[folds==k,],id=i)

cv.errors[k,i]=mean( (crash1$HEAD\_INJ[folds==k]-pred)^2)

}

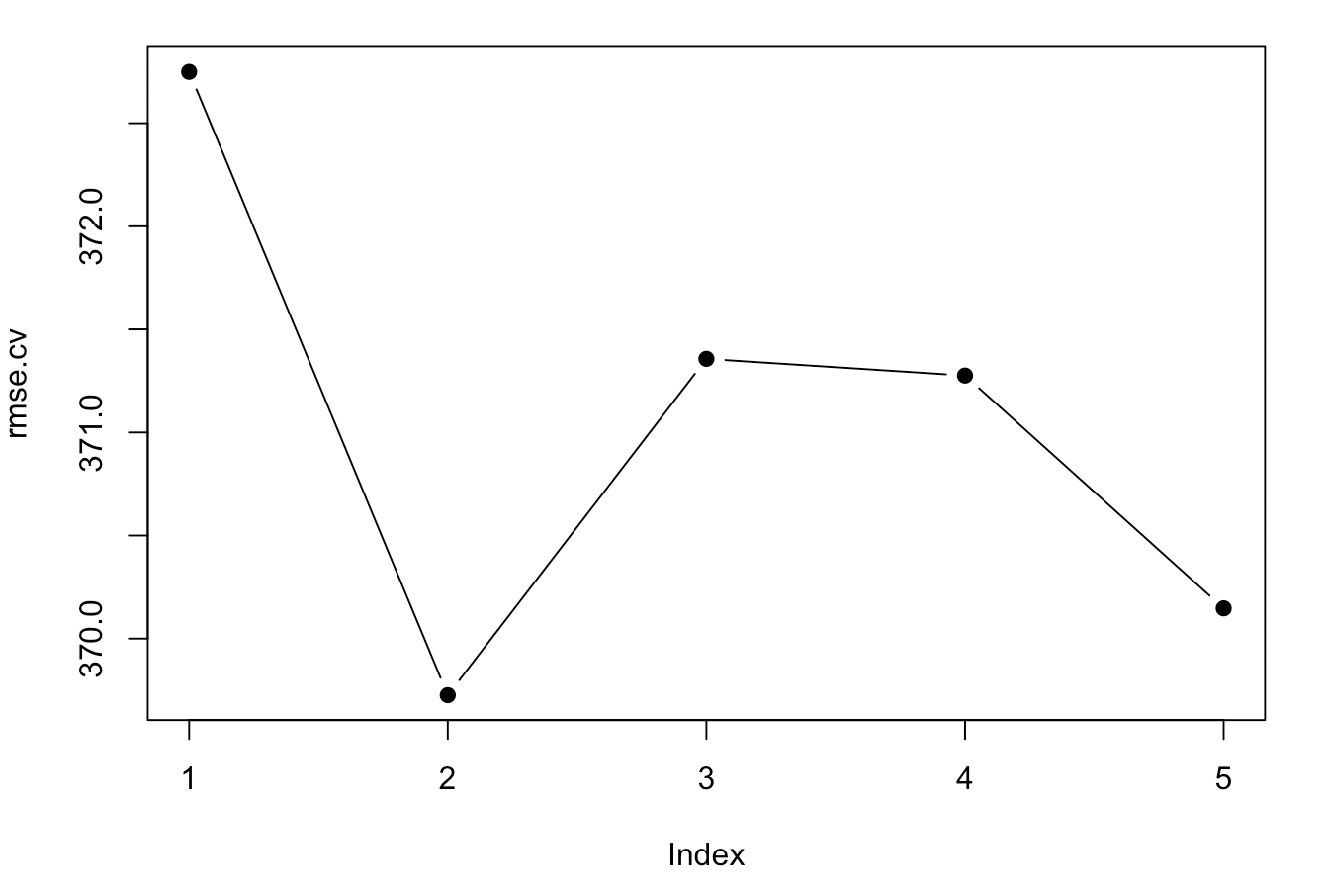
}

rmse.cv=sqrt(apply(cv.errors,2,mean))

plot(rmse.cv,pch=19,type="b")

rmse.cv

Execute the commands above and based upon the results determine an appropriate number of variables to include in the model.



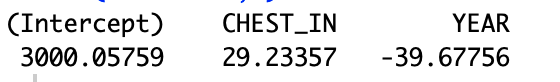
The RSME indicates that 2 variables would have the best fit for this model.

To determine the variables you will need to execute the following commands. In the second command put in the number of variables that you have determined in place of **numb**. Notice that in this command and the subsequent commands the data frame crash is being used because missing data does not cause any problems.

fit1.full=regsubsets(HEAD\_INJ~CHEST\_IN+LLEG\_INJ+RLEG\_INJ+WEIGHT+YEAR,data=crash, nvmax = 5)

coef(fit1.full,**numb**)

What is the model that you found?



d) We now want to consider adding the categorical variables to the model that you found in part c). The strategy is to use what is called a partial F test in which we fit two models: a full model and a reduced model. This will be illustrated with the following example.

Suppose we are considering whether or not the variable SIZE should be added to the model you found in part c? To include this variable would require that 7 indicator variables would be created and we would have the following two models. In the reduced model it is assumed that there are 3 continuous variables. In the full model the variables with the labels I are the 7 indicator variables.

Reduced model:

Full model:

We wish to test the null hypothesis: = 0 versus the alternative hypothesis: at least one of to is not equal to 0.

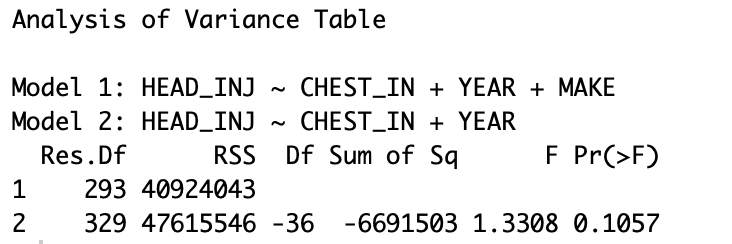
If the null hypothesis is true then the full model becomes the reduced model. Thus, if the null hypothesis cannot be rejected, then there is little evidence that the full model is needed to explain the data.

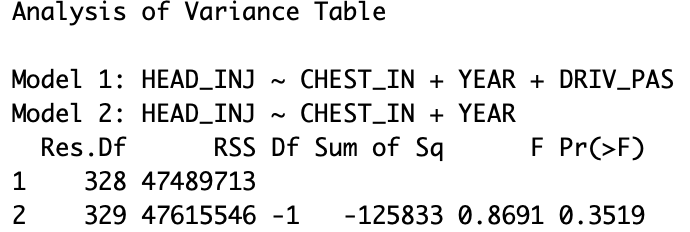
The following set of R commands compute the p-value associated with testing the above hypotheses. In this example it is assumed that the 3 continuous variables CHEST\_IN, WEIGHT, and YEAR are in the reduced model. These are not necessarily the variables that you would include in that model.

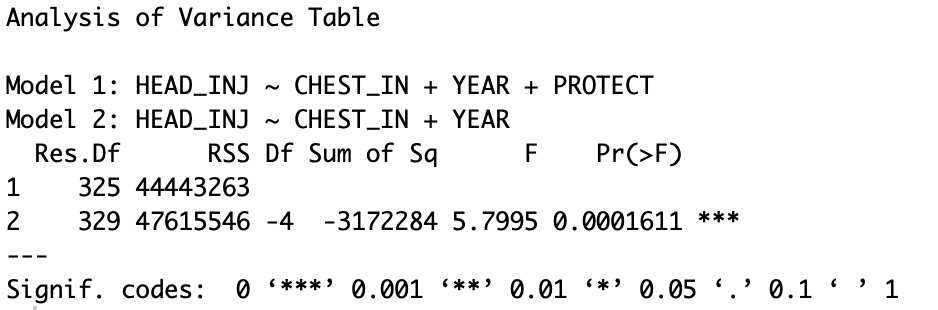
redmod = lm(HEAD\_INJ~CHEST\_IN+WEIGHT+YEAR, data=crash)

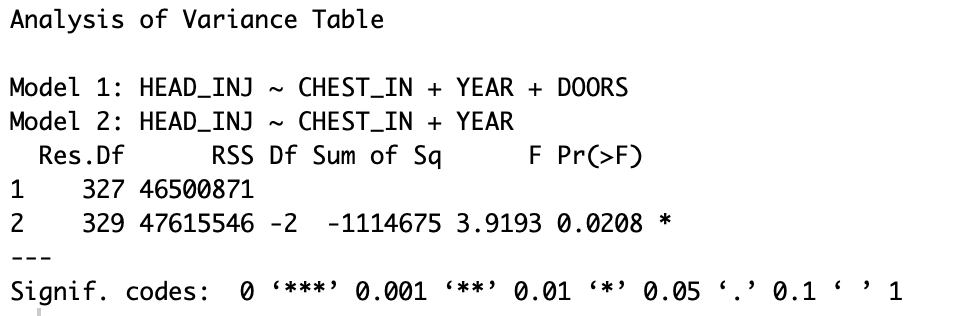
fullmod = lm(HEAD\_INJ~CHEST\_IN+WEIGHT+YEAR+PROTECT,data=crash)

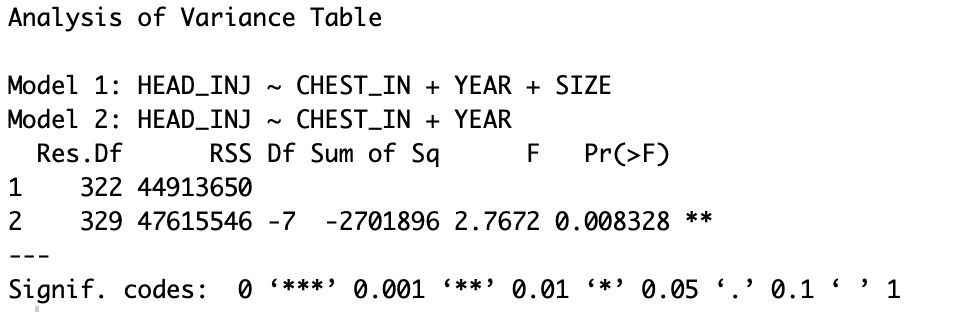
anova(fullmod,redmod)











SIZE, DOORS, PROTECT have p-values < 0.05.

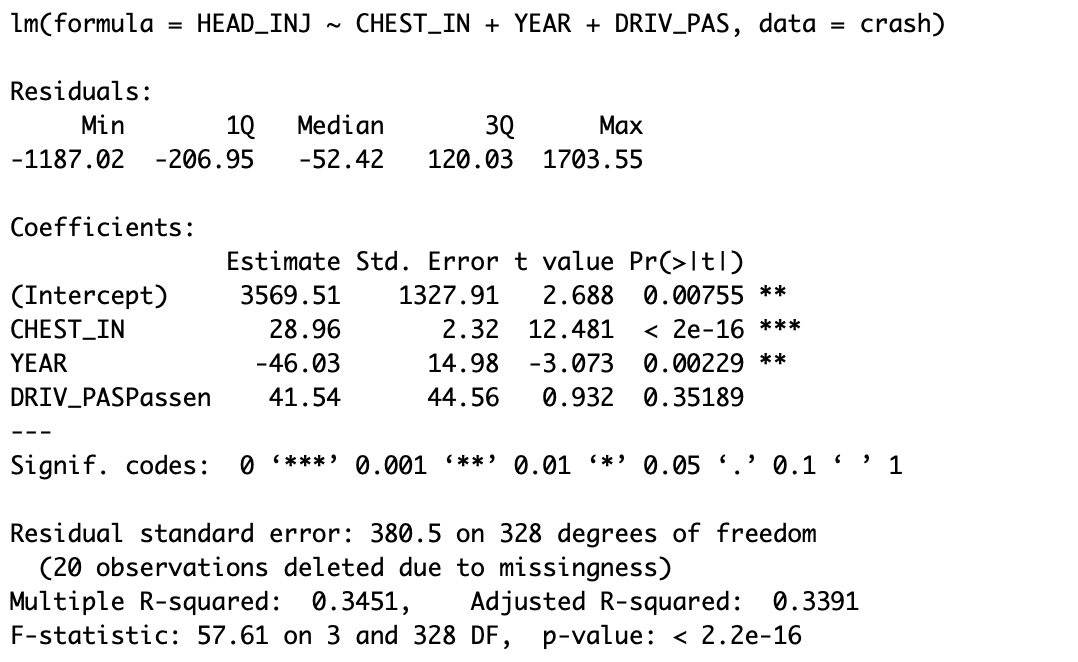
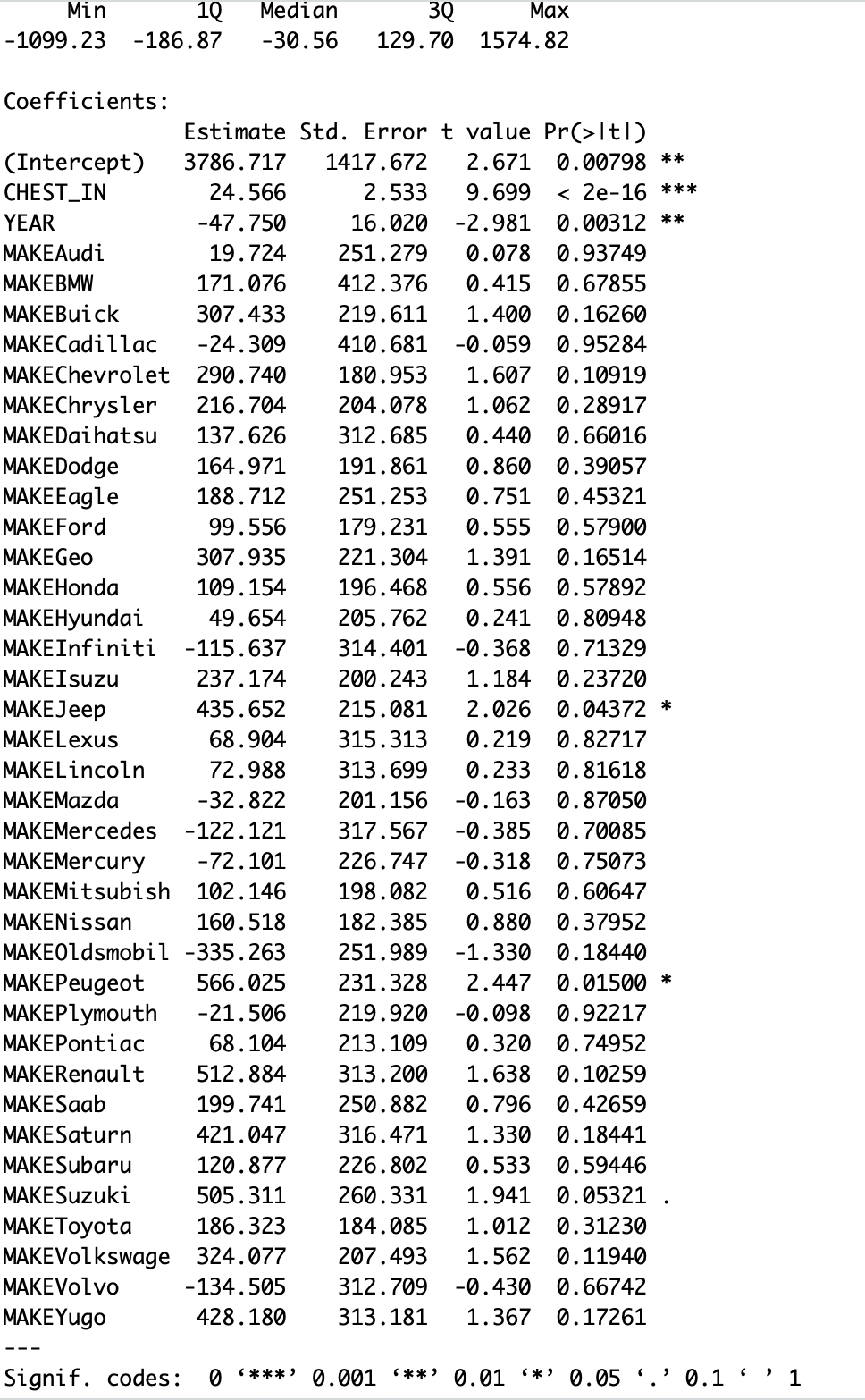
For all of the categorical variables MAKE, DRIV\_PAS, PROTECT, DOORS, and SIZE start with the model you came up with in part c) as the reduced model and then add the categorical variable to get the full model. Compute the p-value for testing the hypothesis that the coefficients for the categorical variable are all equal to 0.

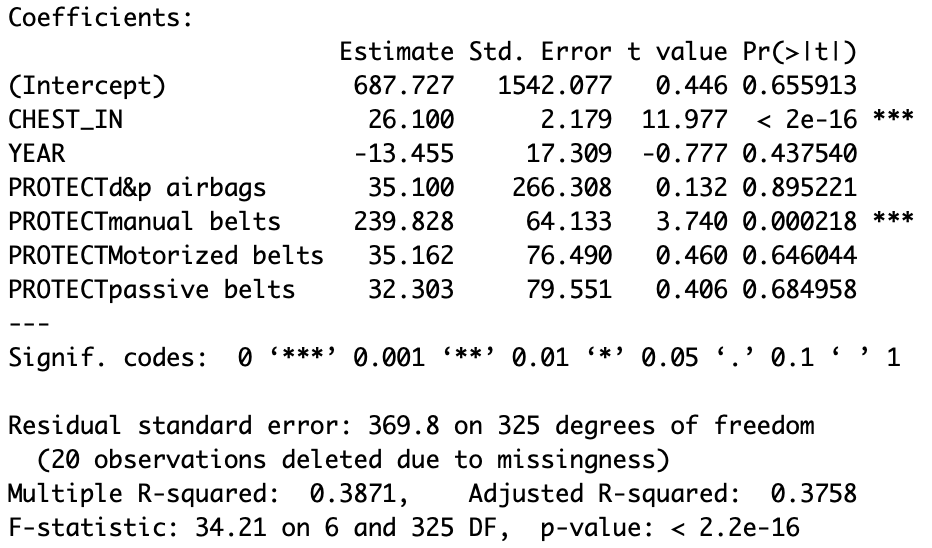
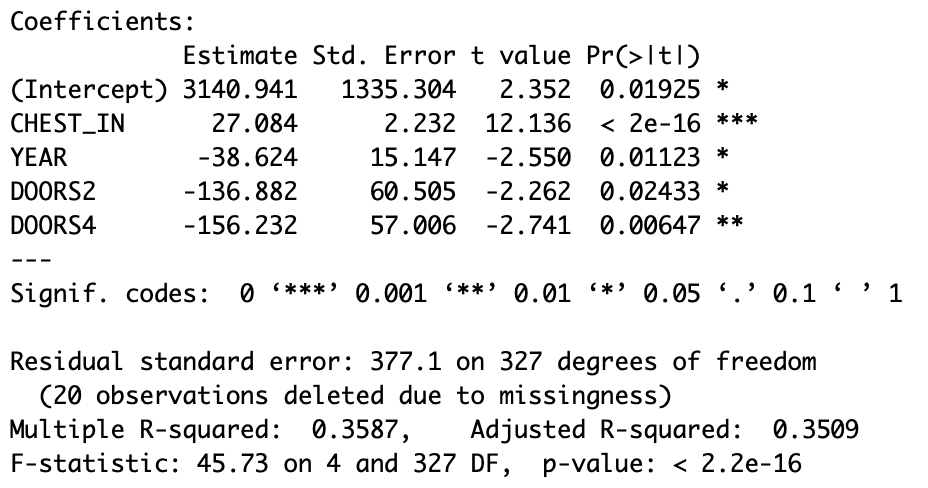
Do this for each of the 5 categorical variables. Report the p-values for each of these categorical variables and order these p-values from the smallest to the largest. Which categorical variables have p-values smaller than .05?

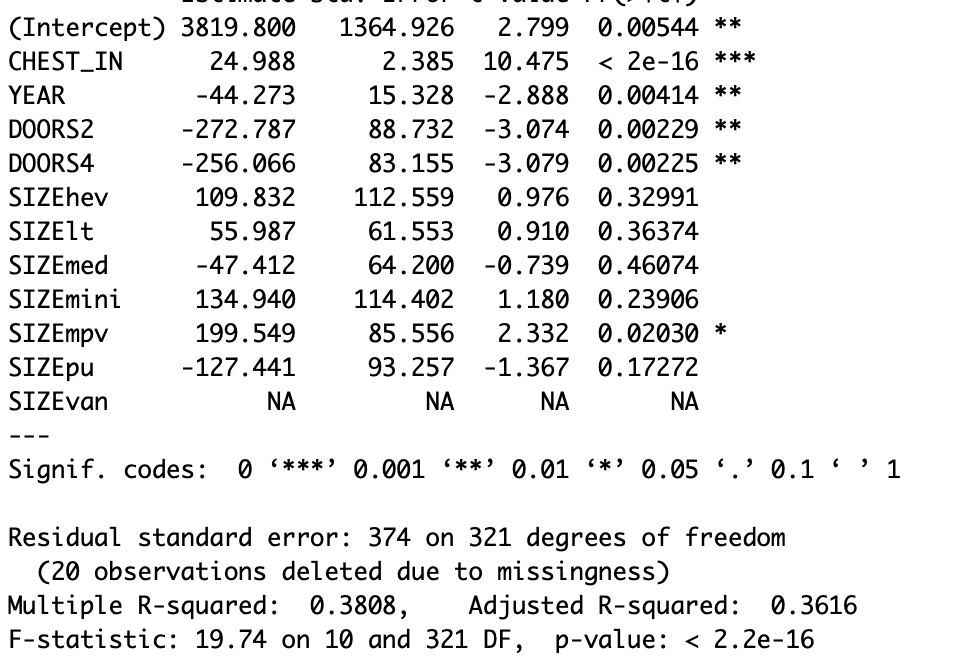
e) Next we will add one categorical variable at a time to the model that you came up with in part c). Start with this model and add the categorical variable which has the smallest p-value from part d) and add this categorical variable to the model. Based upon the result is this categorical variable statistically significant?

If you have a statistically significant categorical variable, take that model as the new reduced model and add the categorical variable with the next smallest p-value to the new model. Is that categorical variable statistically significant?

Continue until it no longer is helpful to add the next categorical variable.

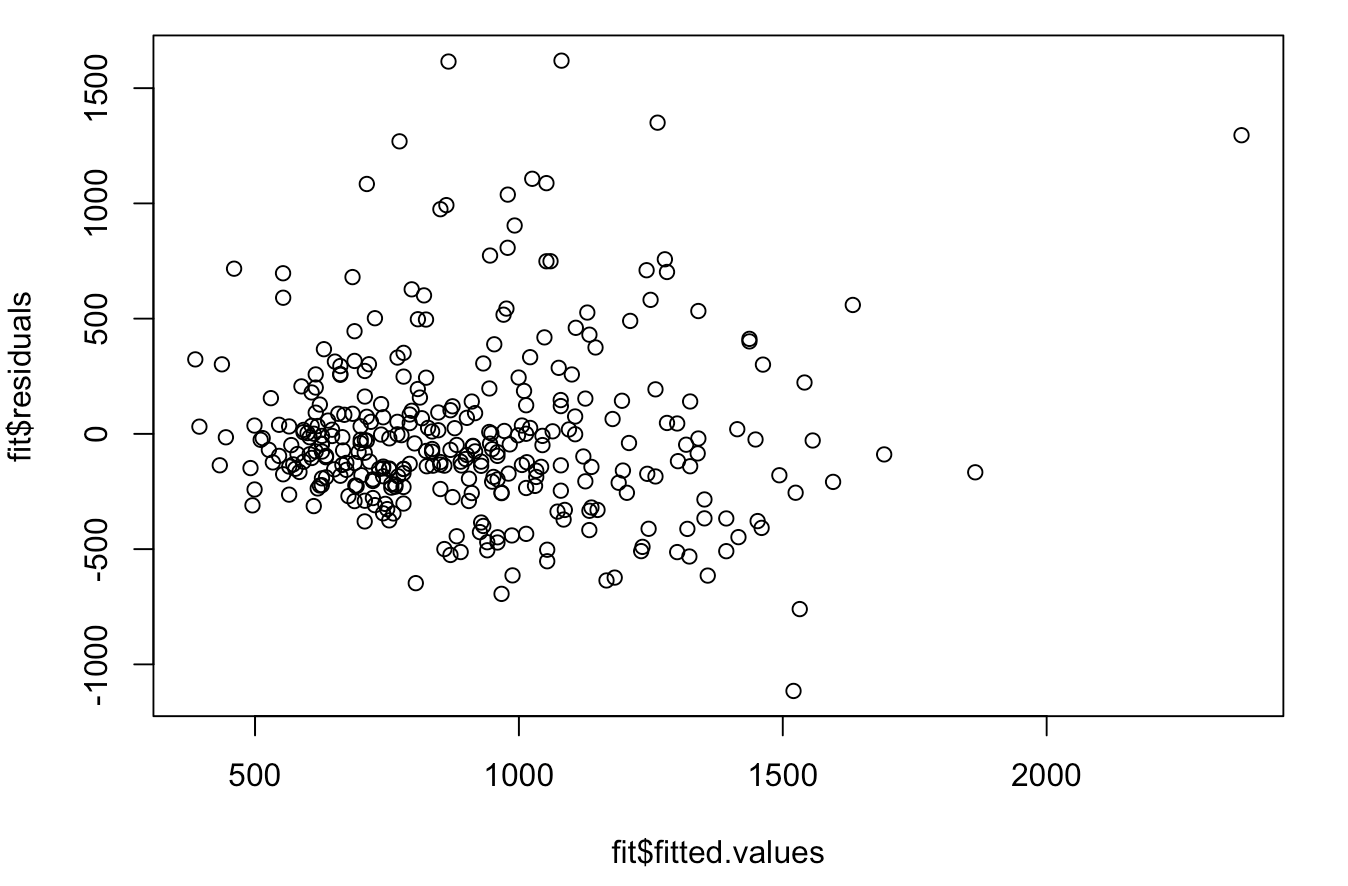
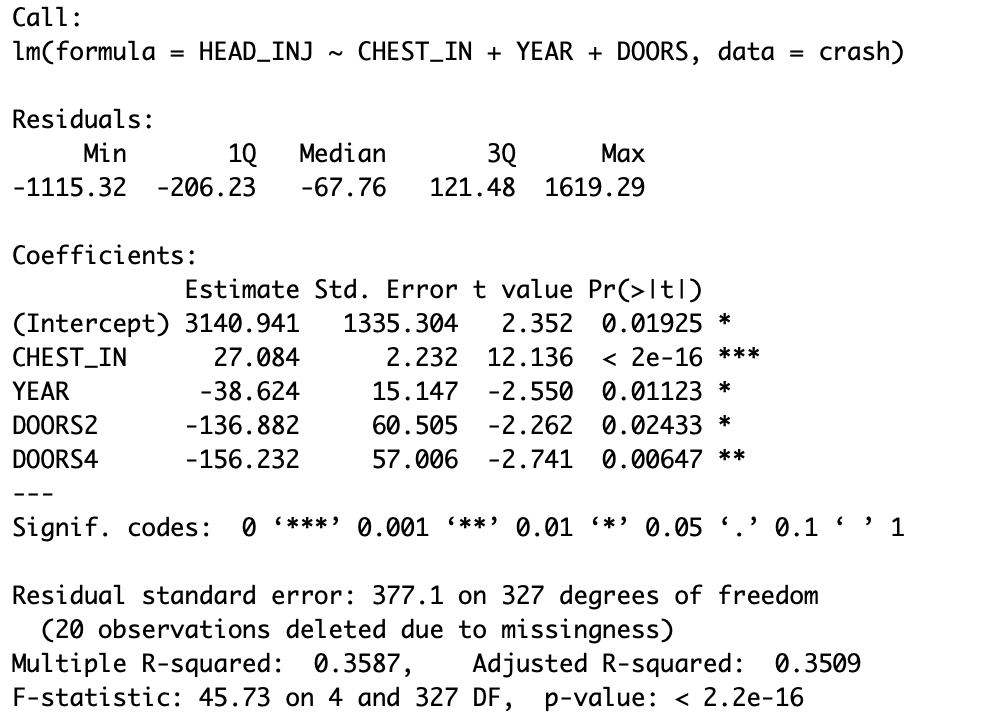
 



I found only Doors to be statistically significant.

f) Once you have determined the continuous and categorical variables that should be in the model. Estimate this model and check the residuals.

The residual shows a slight increase in variance but overall looks good.