Nathan Groom

621 Homework 5

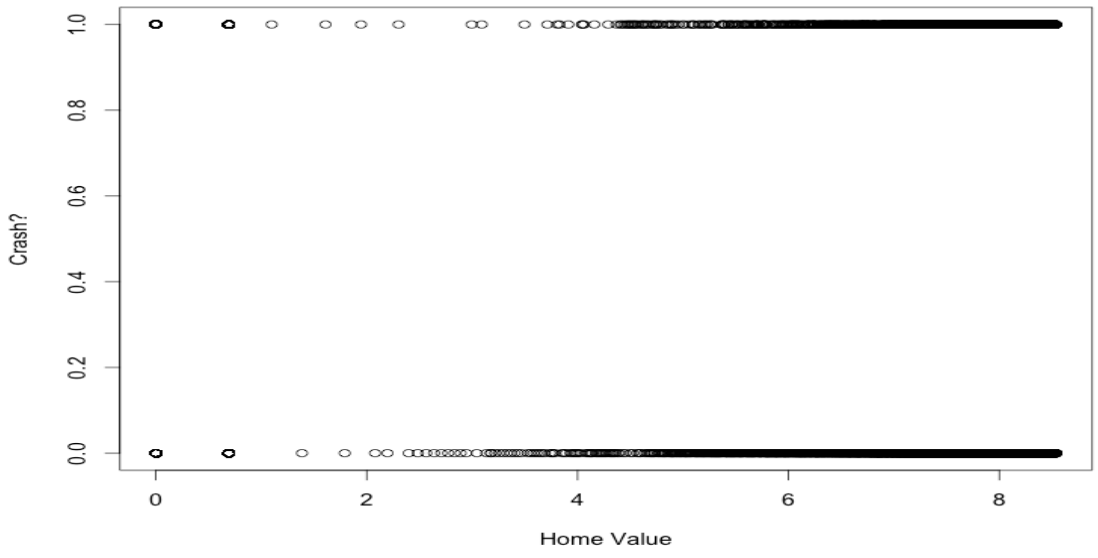
The purpose of this analysis is twofold. Firstly, through use of a dataset with approximately 8000 rows, we seek to determine which features of automobile drivers, or of the automobiles themselves, are most effective at predicting an accident. Secondly, if the car has indeed been in an accident, we also seek to predict which variables are most effective at predicting the repair costs. The analysis will be conducted using the R programming language to build multiple linear and binary logistic regression models.

**Data Exploration**

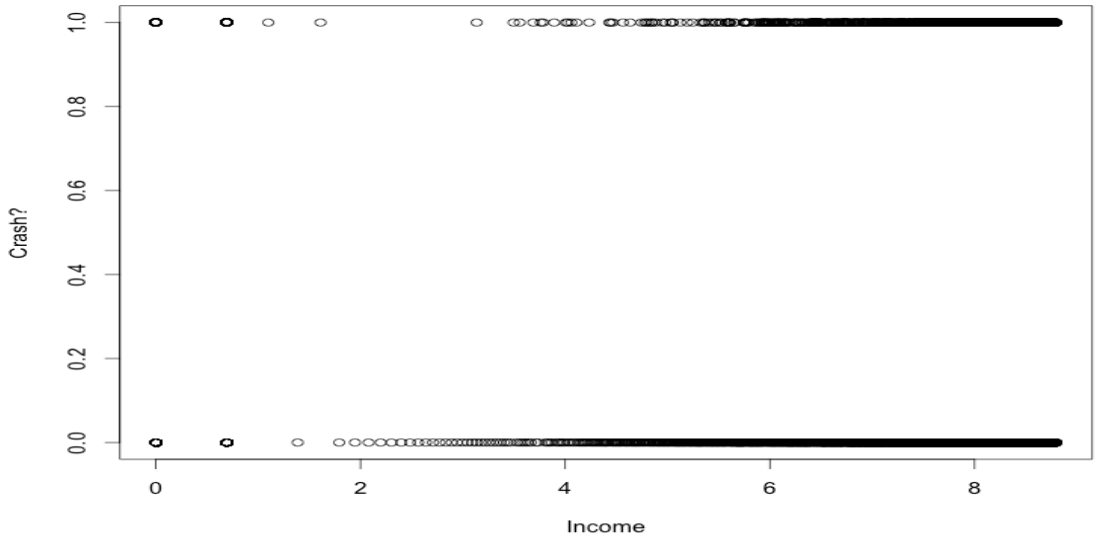
As noted above, the dataset we will use to conduct our analysis consists of approximately 8000 rows. Including the dependent variables, this dataset contains 26 columns, one of which is the index. Therefore, the number of predictor variables in our dataset is 23. This dataset is the training dataset. The evaluation dataset lacks target variables but contains the same predictors, and consists of approximately 2100 rows.

As is customary in the exploration of data, one of the first steps to take is to create plots. Since there are two dependent variables, and one of those depends on the value of the other, it might be wise to plot some of the variables twice: once against the binary variable representing whether or not that car has been in an accident, and then again against the target variable representing the repair costs (omitting rows representing cars that have not been in crashes).

We will first look at plots of the numeric variables against the binary dependent variable.

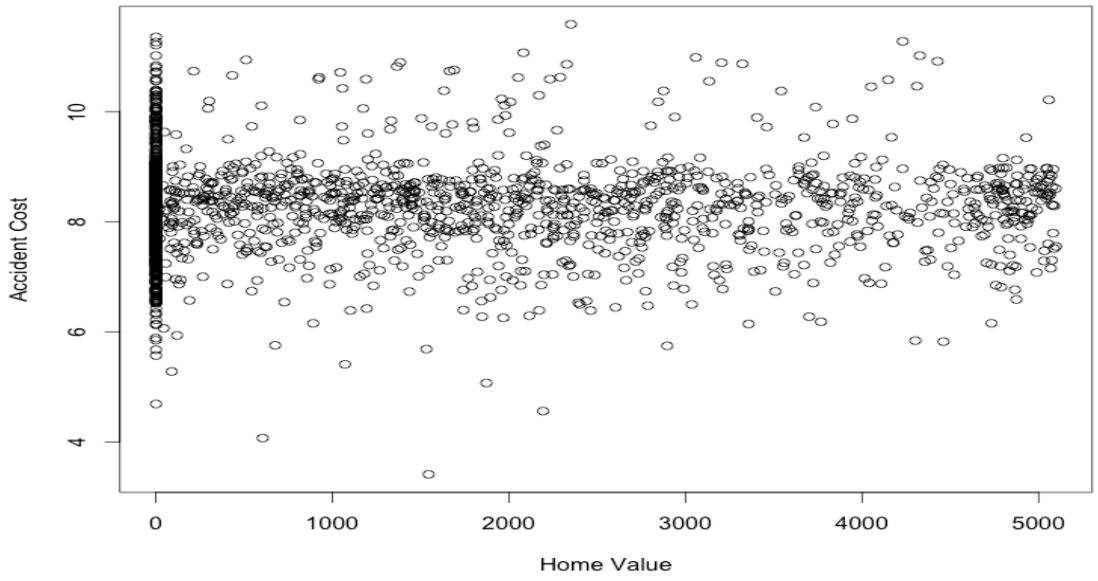


The above graph seems to suggest a slight positive correlation between home value and whether or not the car was in an accident (the x-values are the logarithm of the actual home values).

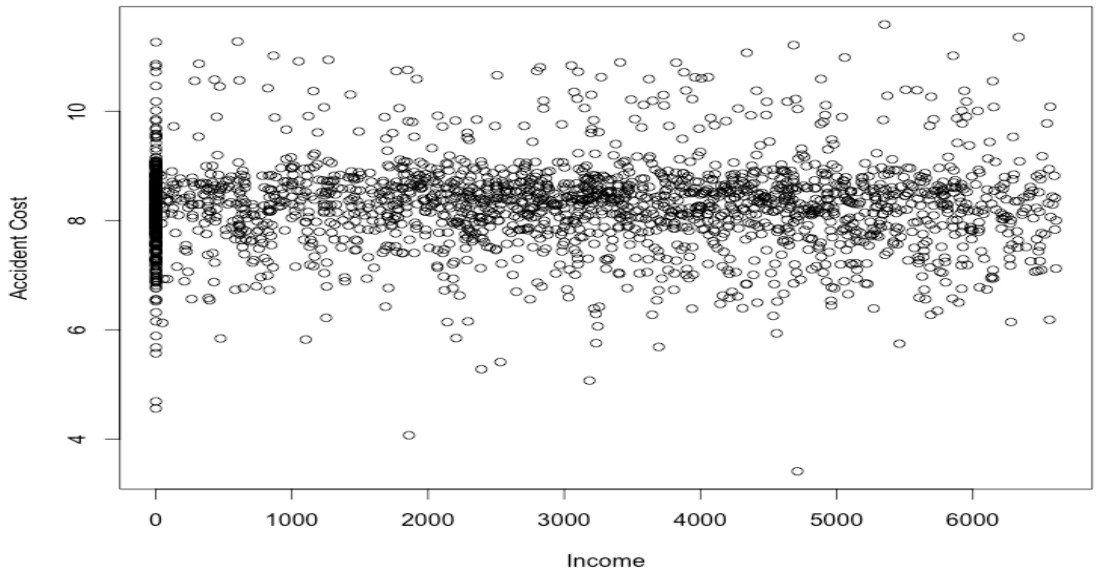



This plot suggests another positive correlation, this time between accidents and income.

However, when we distill our data to only cars that have been in accidents, we find that the same variables are not significant in terms of predicting the cost of auto repairs.



In the above plot showing accident cost plotted against home value, it seems that there exists no correlation whatsoever.




The same can be said for income. In fact, based on such plots (most of which are not included in this report), not a single predictor variable appears to be significant for predicting these cost values. This is all the more reason to turn to model building to find answers.

**Data Transformation**

One look at the data and it is evident that there needs to be some significant transformation before any meaningful analysis can take place. First, all the variables with dollar amounts need to be transformed into simple numeric values. These include “bluebook”, “income” and “home value”. Next, I want to transform any variables with words as levels into numeric levels, or dummy variables. For example, “sex” will be transformed from “male” or “female” to 0 (male) or 1 (female). Each job will get a numeric level, as will each car type, and level of education. The rest, like male/female are binary. “Red”, “single parent”, “revoked”, “urban/rural”, and “marital status” are essentially questions with a yes or no answer. We will change these to binary levels, with 0 being “no” and 1 being “yes”. This makes it much easier to model the transition from yes to no or vice versa.

**Build Models**

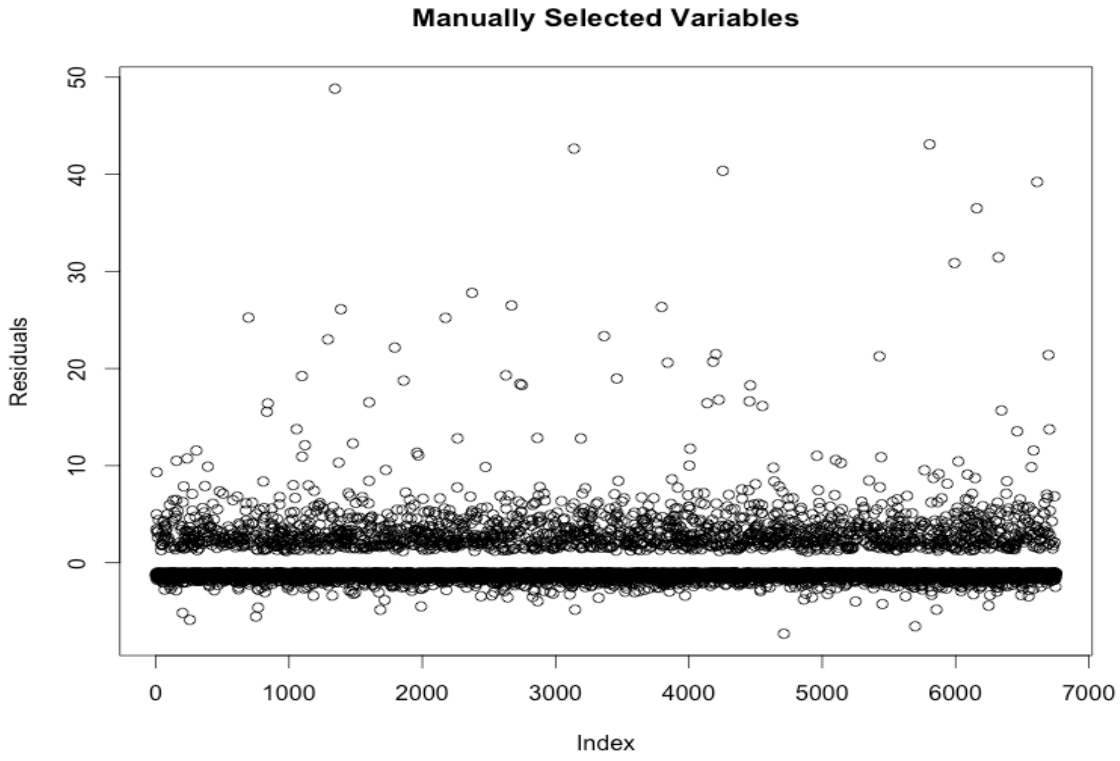
As one of the two target variables is contingent on the other, it is likely best to run what are essentially two separate analyses. First a binary logistic regression to predict whether or not the car had been in an accident in the first place, followed by a multiple linear regression to predict repair costs for cars that have indeed been in an accident.

For the first model in the logistic regression analysis, I will employ the technique that worked best in R for the previous report, which happened to be the backwards stepwise selection. After running backwards stepwise selection with the every variable in the dataset, several variables are removed. Two factor variables, "education" and "job" only appear to be significant for some factors and not others. Therefore it is best to remove these variables as well. That leaves us with a model containing the following variables: "driving children", "years on job", "single parent", "marital status", "sex", "travel time", "car use", "time in force", "car type", "old claim", "claim freq.", "points", and "urban/rural". Variables "points", "urban/rural", and "car use" are the most significant. When we build a stepwise model using forward selection, we see the same results—the same variables and the same AIC of 6271.1.

I also want to build one model that is purely done by manual selection. There are many variables to choose from and the best approach here is to simply go through the variables and decide which are likely to have the greatest effect on whether or not that vehicle has been in a crash. I think that age probably should be included. Bluebook should not—I do not believe the value of the car makes it more likely to be involved in an accident, nor do I believe the age or type of car would have an effect. Claim frequency definitely should be included. Max education level should not be included, but I do think that number of children at home should be included because if the driver has children, and those children are in the car, maybe the driver will be more cautious. I will include income because I think this one variable will essentially have the same effect as bluebook, home value, job category, years on job and education. All of these variables directly affect or are affected by income. Even if income does not have any effect on whether or not a car has been in an accident, there are so many variables in this dataset that influence of are a result of the driver's socioeconomic status, that I believe it is best to include one of these variables to represent them all. Income is the best choice because it has the most direct influence on socioeconomic status. Next, I believe number of kids driving should be included for the same reason as "age" – younger drivers have been proven to get into more accidents. I do not believe marital status has an effect, but I do believe motor vehicle record points has an effect, as well as total claims in the past five years. Whether or not the driver is a single parent probably does not matter but whether the driver is a parent might. Still, I will leave this variable out as its effect is almost definitely the same as number of children at home. I do not believe whether the car is red will have a real effect. I do believe that if a driver has had his or her license revoked, that driver is more likely to get into a car crash in the future (perhaps that person is a compulsive drunk driver). Finally, I will include distance to work (more driving means higher likelihood of accidents), and also include "urban/rural" because I think driving is more inherently dangerous in cities and there are more opportunities to get into accidents. This brings the grand total of variables included, based purely on manual selection, to the following: "age", "claim frequency", "kids at home" , "income", "driving children", "motor vehicle record points", "total claims" , "revoked" , "travel time" and "urban/rural."

We build this model and look at the results. The good news is that every one of these variables is significant. In fact, the least significant variable is “total claims” with a significance level of 0.002723. It is also encouraging to see “income” with a significance level of 0.000183 because I do believe that there is no need to include more than one variable that measures a driver’s wealth and/or socioeconomic status. The underlying assumption for all such variables is most likely that if a driver is responsible enough to have a well paying job and accumulated wealth in his or her lifetime, perhaps this translates to responsibility behind the wheel as well. There is no need, in my opinion, to include more than one variable that measure this sort of potential correlation. The AIC is higher than the model built by stepwise selection, but this is probably to be expected with fewer variables. Furthermore, the manually selected model has a log likelihood of -3369 compared with -3115 for the stepwise models.

The residuals look random too:




I am happy with this manually selected model. It is almost as good as the stepwise models in terms of AIC and log likelihood and has far fewer variables. I think the best thing about this model is the fact that we could remove so many variables that measure, in essence, the same thing, or at least the same underlying cause that would perhaps contribute to an automobile accident. If there are so many predictor variables that have the same effect on the target variable, why include more than one of those variables?

A word on coefficients: there is not anything majorly counterintuitive here. The negative coefficients are: age, income, and urban/rural. And all of those make sense. As age goes up, accidents go down. As income goes up, accidents also go down. However, as we move from rural (0) to urban (1), accidents also go down. This is the opposite of what I thought in terms of more congested roads leading to more accidents but perhaps, like the "travel time" variable, the fact that a rural driver will drive longer distances at a time leads to more accidents. Despite the fact that this was not what I expected, it can still be reasonably explained. In terms of the positive coefficients, "clm. freq" is positive, "kids driv" is positive, as are "travtime", "oldclaim" , "revoked", "mvr. pts" and "home kids". The fact that kids at home lead to more accidents is also at odds with what I had assumed. However, if they include teen drivers as the kids at home, maybe those kids contribute to more accidents by driving themselves. All in all, nothing terribly counterintuitive.

The last binary model I will build will be similar to the above model but I am replacing "income" with a couple of the other variables that, as pointed out above, measure socioeconomic status, either directly or indirectly. Instead of "income", we will have "years on job" , "bluebook", and "home value". These variables are clearly tied to income in most cases. However, I do want to see how much of a difference it makes in terms of being a better fit.

This second manual model has a slightly lower AIC of 6608 compared to the first manual model's 6761 rating, and a slightly higher log likelihood of -3291 as opposed to the first manual model's -3369 rating. The coefficients are negative for both "years on job" and "home value", which is expected. However, "bluebook", has a slightly positive coefficient of 0.00014, meaning that as car value goes up, the odds of accidents happening are greater. This is very slight, even though we are considering the effects of single dollar changes. The coefficient of "home value" is -0.00019, meaning one dollar has a greater effect in terms of home prices, despite the fact that homes generally cost much more than automobiles. Nevertheless, the bluebook coefficient is positive, slight as it is. The only explanation I can think of is that perhaps more expensive cars are faster, which leads to more reckless driving and therefore more accidents. However, overall I think the best option is to stick with the first manual model. It has fewer variables and the AIC and log likelihood are not much worse than the second manual model. Plus I have concerns about multicollinearity when it comes to the variables I added to the second manual model.

The next step is to focus on the multiple linear regression to predict accident costs. Like the binary regression, the first thing we will try is a stepwise model, but this time fitting OLS instead of binary regression.

Unfortunately, the stepwise model proves to be less than a good fit. This model has a multiple R-squared value of 0.009 and an adjusted R-squared of 0.005. Only six variables have been included in the model, and of these, only one, marital status, is significant at the 0.05 level. The others, "yoj", "car.use", "oldclaim" , "revoked" and "mvr.pts" are not significant. Therefore, we will try to come up with a better model.

Next, we can repeat what we did for the binary regression by choosing variables manually. Which variables will have the biggest influence on how much an accident will cost? I think that most will have no effect. Number of kids driving, age, kids at home, and years on job may have an influence on whether a crash happens, but not necessarily the dollar amount of the repairs. The most obvious variable (to me) with an effect on repair costs is "bluebook". How expensive the car is will clearly affect how much it will cost to repair in the event of an accident. Note that the variable description simply says, "if the car was in a crash, what was the cost"—implying the general cost of repairs, not necessarily the amount paid out of pocket by the driver after taking into account what the insurance covers. For that reason, I will assume that a more wealthy person's better insurance coverage is irrelevant. We can also include car age and car type, but like "income" previously, for our purposes in building this model, car age and car type only serve to influence what the bluebook value will be. Therefore, intuitively, I do not feel there is a reason to include these. But a quick build of a model with only "bluebook" as a predictor shows a p-value of 0.732 and an R-squared value of nearly zero. We can ask ourselves here: what causes the cost of repairs to increase? Intuitively, it would be the value of the car as well as the extent of the damage. The value of the car is measured by "bluebook" but we can add "car age" and "car type" to see if it improves the model. The extent of the damage is not represented in one variable but I am inclined to believe that if the driver has had his or her license revoked and/or filed a higher number of claims in the past five years, the driver may be predisposed for more serious accidents. Therefore, the next model will include "bluebook", "car type", "car age", "revoked" and "oldclaim." However, after building this model, the only significant variables are car types 1 and 5, meaning panel trucks and vans. Somehow, these kinds of vehicles incur higher costs when in accidents. Other than that, it seems like none of these predictor variables are effective in predicting accident costs.

At this point it seems obvious that it is going to be quite difficult to come up with a model that has any predictive power for this particular variable. None of the models seem to have any significant variables.

**Select Models**

For the binary logistic model, I am going to stick with the first manually built model. As mentioned above, this model does have a lower AIC and log likelihood than the stepwise models, but it uses fewer variables, particularly due to the fact that we are eliminating what are essentially duplicate variables measuring socioeconomic status. In this way it makes more sense than a model chosen strictly through AIC or another summary statistic. However, when making predictions against the true binary variable in our test data, the first manual model proved itself to be very poor with about a 33% accuracy rate. Much better was the second manual model, with an accuracy rating of about 73%. This still is not ideal (it would be great if it were in the 90s) but not terrible either. This means the error rate is about 27%. Precision rate is at 49%, sensitivity is at 31%, specificity is at 88%, F1 score is 0.3819, AUC is

0.5981. The confusion matrix shows 5306 true positives, 674 true negatives, 702 false positives and 1479 false negatives—clearly the biggest error rate is coming from the false negatives. We see that the sensitivity is at 31%, meaning the proportion of positives that was correctly identified is quite low. When we predict values in the evaluation set, we get about 12% of the responses as crashes and 88% as non crashes, which also seems low.

This brings us to the model for the multiple linear regression. I have to conclude that none of these variables is a good predictor for how much an accident is going to cost. By "good" I of course mean statistically significant. Some might be better than others, but at the p=0.05 significance level, nothing seems to work. From what I have found, there are two factors in the car type variable that are significant, and the "revoked" variable is almost significant but not quite. Therefore, unfortunately I believe I have no choice other than to use a model that only includes car type and revoked as predictors. When we build a model with only these variables, "revoked" does become significant at the p=0.1 level. This is not as significant as I would like to see but probably as good as we will find in this data. The residuals look random as well—there are some outliers but not many. This is therefore the best model for the multiple linear regression, even as it has an adjusted R-squared value of only 0.00665, a terrible fit. But intuitively, this actually does make sense—none of these variables would be a good predictor for the repair costs. We would need to know far more about the accident itself: how fast the car was going, whether it was day or night, if the driver had been drinking, etc. Any prediction made with this data is going to be very indirect at best. Nevertheless, we can still make a prediction and see how it turns out.

Indeed, the summary statistics do not turn out well. The R-squared statistic is a mere 0.009419 and the F statistic is 3.401 on 6 and 2146 degrees of freedom. The MSE is a huge 59364539 (though we are measuring dollars here so a squared dollar amount might be expected to be high). As stated earlier, the residuals do look normal, though there are some outliers.

One interesting tidbit, however, is when I predict values with the multiple linear regression model using the training data, the values, while not so accurate that the model could realistically be used, are not as inaccurate as one might believe based on the summary statistics mentioned above. The mean is 2946 for the real costs and 4021 for the predicted costs—not highly accurate but only off by about $1000. So while I would not recommend using this model, the actual predicted results are better, at least, than I would have expected. Finally, the predicted values of costs using the evaluation data set are a mean of $5760.96, a median of $5590.54, and a standard deviation of $722.79, again not as bad a prediction as one might expect.

**Index: R Programming Language Code**

```
library(plyr)

library(dplyr)

library(leaps)

library(pROC)

it=read.csv("/users/nathangroom/desktop/621csvs/insurance_training_data.csv")


ie=read.csv("/users/nathangroom/desktop/621csvs/insurance-evaluation-
data.csv")

#names(it)<-
c("index","targetcrash","targetcost","driver.age","bluebook","car.age","car.type","car
.use","claims","edu","kids","home.value","income","job","kids.driv","mstatus","points
","oldclaim","parent1","red","revoked","sex","tif","travtime","urban","yoj")

names(ie)<-
c("index","targetcrash","targetcost","driver.age","bluebook","car.age","car.type","car
.use","claims","edu","kids","home.value","income","job","kids.driv","mstatus","points
","oldclaim","parent1","red","revoked","sex","tif","travtime","urban","yoj")

# 2 response variables. Best to analyze one at a time. First predict logistic regression

# for binary variable, then take only data that has a "yes" for crash yes/no and
perform

# generalized least squares on that data.

#get rid of 2nd target (For now)

it.1=it[,-3]

#get rid of index

it.1=it.1[,-1]

names(it.1)<-
c("targetcrash","kids.driv","age","home.kids","yoj","income","parent1","home.val","
mstatus","sex","education","job","travtime","car.use","bluebook","tif","car.type","red
","oldclaim","clm.freq","revoked","mvr.pts","car.age","urban")

#first build a model with all the numeric variables

model.1=glm(targetcrash~kids.driv+age+home.kids+yoj+travtime+tif+clm.freq+mv
r.pts+car.age,family="binomial",data=it.1)
```

```
#everything here is highly significant

#now transform the data to make it all numeric

it.1$income<-as.numeric(it.1$income)

it.1$home.val<-as.numeric(it.1$home.val)

#marital status: 0 = no, 1= yes

it.1$mstatus=revalue(it.1$mstatus,c("z_No"=0,"Yes" = 1))

#sex: 0=male, 1=female

it.1$sex=revalue(it.1$sex,c("M"=0,"z_F" = 1))

# education: less than HS = 0, HS = 1, bachelors = 2, masters = 3 , phd = 4

it.1$education=revalue(it.1$education,c("<High School"=0,"z_High School" =
1,"Bachelors"=2,"Masters"=3,"PhD"=4))

it.1$job[it.1$job==""]<-NA

it.1$job=revalue(it.1$job,c("Student" = 1,"Home Maker"=2,"z_Blue
Collar"=3,"Clerical"=4,"Manager"=5,"Professional"=6,"Lawyer"=7,"Doctor"=8))

it.1$car.use=revalue(it.1$car.use,c("Commercial"=0,"Private" = 1))

it.1$bluebook<-as.numeric(it.1$bluebook)

it.1$car.type=revalue(it.1$car.type,c("Minivan"=0,"Panel
Truck"=1,"Pickup"=2,"Sports Car"=3,"Van"=4,"z_SUV"=5))

it.1$red=revalue(it.1$red,c("no"=0,"yes" = 1))

it.1$parent1=revalue(it.1$parent1,c("No"=0,"Yes" = 1))

it.1$oldclaim<-as.numeric(it.1$oldclaim)

it.1$revoked=revalue(it.1$revoked,c("No"=0,"Yes" = 1))

it.1$urban=revalue(it.1$urban,c("Highly Urban/ Urban"=0,"z_Highly Rural/ Rural" =
1))


#finally we can make a model with all features involved

model.all<-glm(targetcrash~.,data=it.1,family="binomial")


#we can graph some of the numeric variables
```

```
plot(x=it.1$age,y=it.1$targetcrash)

plot(x=it.1$bluebook,y=it.1$targetcrash)

#both of the above variables don't seem significant when plotted

plot(x=log(it.1$home.val),y=it.1$targetcrash)

plot(x=log(it.1$income),y=it.1$targetcrash)

#both of the next two do seem significant when plotted

#create a new dataframe with only cars that have been in crashes.

targetcost<-it$TARGET_AMT

it.3=cbind(it.1,targetcost)

it.cost=filter(it.3,targetcrash==1)

plot(it.cost$home.val,log(it.cost$targetcost),xlab="Home Value",ylab="Accident
Cost")


#build models

#first try most effective from last homework, which was stepwise:

it.2<-na.omit(it.1)

m.all2<-glm(targetcrash~.,data=it.2,family='binomial')

back2=step(m.all2)

summary(back2)

#now we can observe the variables that are not significant

#and rebuild the models without those variables.

# it seems that only a couple factors in 'job' and 'education' are significant

# therefore we should take out those variables

# home.kids also is not significant enough

m.all3<-
glm(targetcrash~kids.driv+yoj+parent1+home.val+mstatus+sex+travtime+car.use+
tif+car.type+oldclaim+clm.freq+revoked+mvr.pts+urban,data=it.2,family='binomial'
)

summary(m.all3)
```

```
back3=step(m.all3)
summary(back3)
#forwards
nothing=glm(it.2$targetcrash~1,family='binomial')
kids.driv=it.2$kids.driv;
tif=it.2$tif
car.type=it.2$car.type
red=it.2$red
oldclaim=it.2$oldclaim
clm.freq=it.2$clm.freq
revoked=it.2$revoked
mvr.pts=it.2$mvr.pts
car.age=it.2$car.age
urban=it.2$urban
yoj=it.2$yoj
income=it.2$income
parent1=it.2$parent1
home.val=it.2$home.val
mstatus=it.2$mstatus
sex=it.2$sex
education=it.2$education
job=it.2$job
travtime=it.2$travtime
car.use=it.2$car.use
bluebook=it.2$bluebook
forwards = step(nothing,
```

```
        scope=list(lower=formula(nothing),upper=formula(m.all3)),
direction="forward")


summary(forwards)


#manually selected

manual<-
glm(targetcrash~age+clm.freq+home.kids+income+kids.driv+mvr.pts+oldclaim+re
voked+travtime+urban,data=it.2,family='binomial')

logLik(forwards)

logLik(manual)

logLik(back3)

plot(manual$resid,main="Manually Selected
Variables",xlab="Index",ylab="Residuals")

manual$coef


#another model with some slight changes

# income taken out, now we add "home.val", "bluebook" and "yoj"

manual2<-
glm(targetcrash~age+clm.freq+home.kids+yoj+bluebook+home.val+kids.driv+mvr.
pts+oldclaim+revoked+travtime+urban,data=it.2,family='binomial')

coef(manual2)


##model to predict accident costs##


m.cost<-lm(targetcost~.,data=na.omit(it.cost1))

back4=step(m.cost)

summary(back4)
```

```
m.cost1<-lm(targetcost~bluebook,data=it.cost1)

summary(m.cost1)

#no good

m.cost2=lm(targetcost~bluebook+car.type+car.age,data=it.cost1)

summary(m.cost2)

#apparently panel trucks and vans are more likely to incur higher costs

# adding revoked and oldclaim

m.cost3=lm(targetcost~bluebook+car.type+car.age+oldclaim+revoked,data=it.cost1
)

#neither new variable is significant. Revoked is close

m.cost4<-lm(targetcost~car.type+revoked,data=it.cost1)

plot(m.cost4$resid)

###predictions###


names(ie)<-
c('index','targetcrash','targetcost','kids.driv','age','home.kids','yoj','income','parent1',
'home.value','mstatus','sex','education','job','travtime','car.use','bluebook','tif','car.ty
pe','red','oldclaim','clm.freq','revoked','mvr.pts','car.age','urban')

ie.1=ie[-1]

ie.1=ie.1[-2]

#re transform data for evaluation set

ie.1$income<-as.numeric(ie.1$income)

ie.1$home.value<-as.numeric(ie.1$home.value)

ie.1$mstatus=revalue(ie.1$mstatus,c("z_No"=0,"Yes" = 1))

ie.1$sex=revalue(ie.1$sex,c("M"=0,"z_F" = 1))

ie.1$job[ie.1$job==""]<-NA

ie.1$job=revalue(ie.1$job,c("Student" = 1,"Home Maker"=2,"z_Blue
Collar"=3,"Clerical"=4,"Manager"=5,"Professional"=6,"Lawyer"=7,"Doctor"=8))

ie.1$car.use=revalue(ie.1$car.use,c("Commercial"=0,"Private" = 1))
```

```
ie.1$bluebook<-as.numeric(ie.1$bluebook)

ie.1$car.type=revalue(ie.1$car.type,c("Minivan"=0,"Panel
Truck"=1,"Pickup"=2,"Sports Car"=3,"Van"=4,"z_SUV"=5))

ie.1$red=revalue(ie.1$red,c("no"=0,"yes" = 1))

ie.1$parent1=revalue(ie.1$parent1,c("No"=0,"Yes" = 1))

ie.1$oldclaim<-as.numeric(ie.1$oldclaim)

ie.1$revoked=revalue(ie.1$revoked,c("No"=0,"Yes" = 1))

ie.1$urban=revalue(ie.1$urban,c("Highly Urban/ Urban"=0,"z_Highly Rural/ Rural"
= 1))

ie.1$education=revalue(ie.1$education,c("<High School"=0,"z_High School" =
1,"Bachelors"=2,"Masters"=3,"PhD"=4))


it.1$predicted<-round(predict(manual,newdata=it.1,type="response"))

it.1$predicted[2971]<-1

it.1$predicted[1315]<-1

it.1$predicted[240]<-1

it.1$predicted[1043]<-1

it.1$predicted[3460]<-1

it.1$predicted[4156]<-1

#not that accurate


it.1$predicted<-round(predict(manual2,newdata=it.1,type="response"))

#just a couple NA values, replacing with 1

it.1$predicted[is.na(it.1$predicted)]<-1

#better

dat=cbind(it.1[1],it.1[25])

names(dat)<-c("class","scored.class")
```

```
ie.1<-ie.1[-1]
#predictions for evaluation set
ie.1$predicted<-round(predict(manual2,newdata=ie.1,type="response"))
predict<-ie.1$predicted[!is.na(ie.1$predicted)]

it.cost1$predictions<-predict(m.cost4,it.cost1)
p<-cbind(it.cost1$targetcost,it.cost1$predictions)
mean(predict[1])
mean(predict[2])
median(predict[1])
median(predict[2])
#for how poor the model supposedly is, the predictions are not wildly off
# mean and median are both 2946 for real costs and 4021 for predicted costs
# so it's off (and model probably not usable) but still not as  awful as
# some of the summary statistics may have us thinking
mse=mean(summary(m.cost4)$residuals^2)
plot(m.cost4$resid,ylab="Residuals",main="Residuals For Crash Costs")

ie.1<-ie.1[-25]
ie.1$predict.cost<-predict(m.cost4,ie.1)
p2=ie.1$predict.cost
mean(p2);median(p2);sd(p2);
```