## This section is on linear regression and will use the insurance.csv dataset

## Here we can look at the histograms of our numeric variables. We can also check out some scatterplots for how they relate to total insurance charges

hist(insurance$age)

hist(insurance$children)

hist(insurance$bmi)

hist(insurance$charges)

plot(insurance$bmi,insurance$charges)

plot(insurance$children,insurance$charges)

plot(insurance$age,insurance$charges)

## This command gives us scatter plots for all of our numeric variables. We need to look at these to check for multicollinearity. Looks fine

pairs(insurance[,c(1,3,4,7)])

## Next we want to code in dummy variables for sex, smoker, and regions

## This creates one for sex setting the value equal to 1 if they are a female

insurance$isFemale[sex=="female"]=1

insurance$isFemale[insurance$sex=="female"]=1

## This creates one for smoking setting the value equal to 1 if they are a smoker

insurance$isSmoker=0

insurance$isSmoker[insurance$smoker=="yes"]=1

## Here we see there are 4 regions so we create 3 dummy variables for SW, NW, and NE. SE will be the default

table(insurance$region)

insurance$southwest=0

insurance$southwest[insurance$region=="southwest"]=1

insurance$northwest=0

insurance$northwest[insurance$region=="northwest"]=1

insurance$northeast=0

insurance$northeast[insurance$region=="northeast"]=1

## This is how we fit the regression model and look at the results

fit<-lm(charges~age+bmi+children+isFemale+isSmoker+southwest+northwest+northeast,data=insurance)

summary(fit)

## We have a pretty good fit with an R^2 of 0.75. We need to check the assumptions about normality of the residuals

plot(fit)

## The first graph shows up if we have a linear relationship. There is a pretty big kink in there so this assumption looks to be violated

## The Second graph is our normality assumption. The dots should follow the line which they do not. Normality of the errors seems to be violated

## The third graph is for homoscedasticity or equal variance across the error terms. From left to right, it looks like the residuals are very spread out around the line but then get tighter. This would seem to violate our assumption

## The last graph is for Cook’s Distance. We want all of our observations to be within the area and not outside of the dashed red line. We don’t even see the dashed red line area so we aren’t even close. This one passes

## So from our graphs, we failed 3 of the 4 tests. There was some weird stuff going on with the age groups and charges, so lets drop age. Also, we will do a natural log transformation on bmi and charges

insurance$lbmi <- log(insurance$bmi)  
insurance$lcharges <- log(insurance$charges)

fit<-lm(lcharges~lbmi+children+isFemale+isSmoker+southwest+northwest+northeast,data=insurance)

summary(fit)

plot(fit)

## With these changes, we have fairly normal graphs 1,2, and 4 but the homoskedasticity assumption is still violated. We also lost quite a bit of fit as we are now at an R^2 of .4955