Unit2 HW Solution

Jacob Turner

January 4, 2019

## MLR Conceptual Questions

1. The errors of the regression model must have constant variance, be normally distributed, and be statistically independent from each other.
2. FALSE. From above, the assumption is that the errors must be normally distributed, the response (without any additional stratification by predictor combinations) does not have to be.
3. Feature selection provides multiple helpful insights to the analysts. It can help the analyst filter through a large number of predictors and it is reasonable to assume that a large number of them are not important. It saves the analyst time, as more complex models can be easily implemented and then determined if they have any gain in explaining variability in the response. What a feature selection method can not do is make calls like transformations to deal with constant variance or make relationships with the response more linear. It also can’t automatically add model complexity terms. This all must be done by the person building the model.
4. The test ASE will be higher. The scenario being described is an example of overfitting the model. By overfitting the data set, there is more variability in the predictions from data set to data set and thus this model will suffer from high variance not high bias. An example of high bias for this case would be if the model being run was simply a linear model or an intercept model.

## HW Question #5

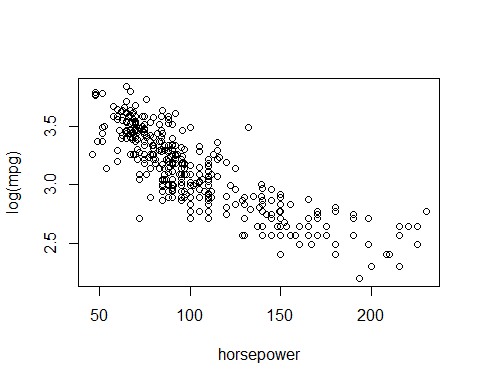
From the pairs plot, we can see that there are some pretty strong correlations between displacement, horsepower, and weight. Acceleration a little bit as well. We can also see this from noticing that the scatterplots of each of the predictors versus mpg have the same negative, slightly curved trend. I suspect that removing one or two of these predictors will be fine as models including both are just going to fight over the same variability to explain.

We can also see that the VIFs for the three variables are above 10 which corresponds well with my visual observations. The color coded pairs plot also suggests that cylinders is probably somewhat redundant with the 3 predictors previously discussed. Its VIF is low but that is hard to interpret since Cylinders is a categorical variable. Clearly from the graph when cylinders are high, displacement, horsepower, and weight all go up.

## HW Question #6

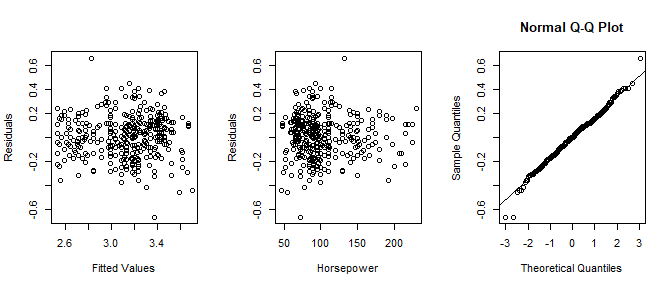
Before moving forward with the actual question, lets pause and just plot the data under the log scale to make some sense of things. The scatter plot below of log(mpg) vs horsepower suggest that the constant variance is indeed fised. There does appear to be a little curvature in the trend, but it is definitely more linear than on the original scale.

## Warning: package 'ISLR' was built under R version 3.5.1



Per the question. The following code produces the residual diagnostic plots of horsepower and a quadratic term regressed on log(mpg). We can see that constant variance is looks reasonable (left) and the qqplot of the residuals is as good as you could ever see in practice.

my.model<-lm(log(mpg)~horsepower+I(horsepower^2))  
par(mfrow=c(1,3))  
plot(my.model$fitted.values,my.model$residuals,xlab="Fitted Values",ylab="Residuals")  
plot(horsepower,my.model$residuals,xlab="Horsepower",ylab="Residuals")  
qqnorm(my.model$residuals)  
qqline(my.model$residuals)



With assumptions met we can statistically test whether or not the quadratic term is significant and should be kept in the model, thus determining if the curvature we see is more than just random variation in data.

summary(my.model)

##   
## Call:  
## lm(formula = log(mpg) ~ horsepower + I(horsepower^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.66460 -0.12041 0.00316 0.11349 0.66376   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.402e+00 7.260e-02 60.639 < 2e-16 \*\*\*  
## horsepower -1.711e-02 1.255e-03 -13.632 < 2e-16 \*\*\*  
## I(horsepower^2) 3.901e-05 4.922e-06 7.925 2.44e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1764 on 389 degrees of freedom  
## Multiple R-squared: 0.7324, Adjusted R-squared: 0.731   
## F-statistic: 532.2 on 2 and 389 DF, p-value: < 2.2e-16

The estimate for the quadratic term, although is quite small 3.9\*e-05, it is still statistically significant (t-stat 7.925, p-value 2.44e-14). So the curvature that we see is still relevant.

Residual plots of a linear model shows a trend in the residuals vs fitted graph.

## HW Question #7

1. The following code deletes the 3 and 5 cylinder observations then creates a random training and test set split. It also deletes variables that are not relevant like “name”.

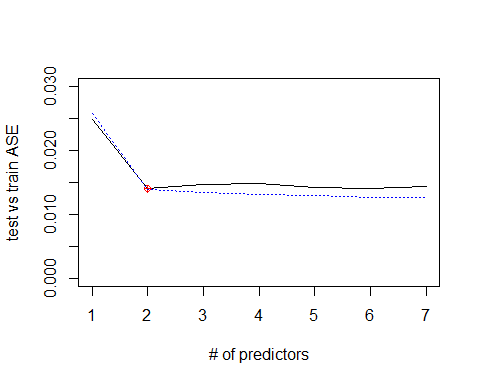
delete.index<-which(cylinders %in% c("3","5"))  
Auto.reduced<-Auto[-delete.index,]  
#Auto.reduced<-Auto.reduced[,-9] # removing names  
  
set.seed(1234)  
index<-sample(1:dim(Auto.reduced)[1],200,replace=F)  
train<-Auto.reduced[index,]  
test<-Auto.reduced[-index,]

1. By deleting the 3 and 5 cylinder observations, we have effectively redefined the population of which we could draw inference too or make claims that the model could predict for. Any model we build discussing effects of displacement, origin, etc would have to be done with a caveat that a resonable population definition would be all cars from these particular origins and standard 4/6/8 cylinder models. We couldn’t even predict the mpg for a new observation that was 3 cylinder even if we tried. There is no coefficient to allow for that to happen.
2. The following runs forward selection on the set of predictors specified in the problem. As we examine the ASE plot, we can see that improvement is made up until 3 predictors where the test ASE levels off suggesting only two predictors are needed.

library(leaps)

## Warning: package 'leaps' was built under R version 3.5.2

reg.fwd=regsubsets(log(mpg)~displacement+horsepower+weight+acceleration+year+origin,data=train,method="forward",nvmax=7)  
  
predict.regsubsets =function (object , newdata ,id ,...){  
 form=as.formula (object$call [[2]])  
 mat=model.matrix(form ,newdata )  
 coefi=coef(object ,id=id)  
 xvars=names(coefi)  
 mat[,xvars]%\*%coefi  
}  
  
testASE<-c()  
#note my i is 1 to 20 since that what I set it in regsubsets  
for (i in 1:7){  
 predictions<-predict.regsubsets(object=reg.fwd,newdata=test,id=i)   
 testASE[i]<-mean((log(test$mpg)-predictions)^2)  
}  
par(mfrow=c(1,1))  
plot(1:7,testASE,type="l",xlab="# of predictors",ylab="test vs train ASE",ylim=c(0,0.03))  
index<-which(testASE==min(testASE))  
points(index,testASE[index],col="red",pch=10)  
rss<-summary(reg.fwd)$rss  
lines(1:7,rss/200,lty=3,col="blue")



1. Below we run the forward selection on the entire data set and observe what the first two predictors that are included in the model. From that information, I simply used the lm function one last time so I could easily get the residuals, t-tests, etc. By the residual diagnostics, overall this model in terms of assumptions appears to be reasonable. Normality is fine. There is a little bit of potential curvature in the residuals vs fitted plot (top left). We could potentially try an interaction between the two predictors to see if that improves or consider another transformation other than log to maybe stablize the variance a little more. I think the latter is less of a concern.

fwd.full=regsubsets(log(mpg)~displacement+horsepower+weight+acceleration+year+origin,data=Auto.reduced,method="forward",nvmax=2)  
  
coef(fwd.full,2)

## (Intercept) weight year   
## 1.7125803256 -0.0003124427 0.0305032558

final.model<-lm(log(mpg)~weight+year)  
par(mfrow=c(2,2))  
plot(final.model)

