Prepared by Richard Rennehan

NSCC Institute of Technology

DBAS3075 Introduction to Statistical Learning

SubSet selection assignment

March 21, 2018

Questions and Answers

#1. Perform best subset regression on the Body Mass data set using the regsubsets() function in R, and then use the summary() function to look at and report which variables are included in the 2 and 6 variable models.

data = read.csv("http://www.statsci.org/data/oz/physical.txt", header=TRUE, sep="\t")

na.omit(data)

best\_subset = regsubsets(Mass~.,data=data,method="exhaustive")

best\_summary = summary(best\_subset)

best\_summary

* The fore and waist variables are included in the two variable model
* In the 6 variable model, the variables included are fore, waist, height, calf, thigh, and head

#2. Create a plot of the values of Adjusted-R^2 for each subset size (1 through 10). Between which two model sizes is the biggest change in Adjusted-R^2?

plot(best\_summary$adjr2,xlab="Number of variables",ylab="Adjusted R-squared",main="Adj-R^2 values of models of size 1-10 found by best subset selection")

* The biggest change in Adjusted-R^2 is between the one variable model and the two variable model. It jumps up from approximately 0.83 to approximately 0.93.

#3. Perform forward stepwise regression on the Body Mass data set. Do the model subsets of each size match those found when performing best subset regression? Which models, if either, are different?

fwd\_subset = regsubsets(Mass~.,data=data,method="forward")

fwd\_summary = summary(fwd\_subset)

* Both are the same. By performing a summary on each, we can see that each use the exact same variables used for each sized model. By plotting adjusted r-squared for each, the graphs look exactly the same.

#4. Perform backward stepwise regression on the Body Mass data set. Do the model subsets of each size match those found when performing best subset regression?

bwd\_subset = regsubsets(Mass~.,data=data,method="backward")

bwd\_summary = summary(bwd\_subset)

* Like with forward stepwise regression, neither are different. By performing a summary on each, we can see that each use the exact same variables used for each sized model. By plotting adjusted r-squared for each, the graphs look exactly the same.

#5. To do this question, you must include the ISLR libray to be able to get the Auto data we previously worked with. Assign the Auto data to a variable (I suggest "cars") and then remove the "name" and "origin" variables, since they are categorical. Once this has been done, perform best subset selection on the remaining data. Which two variables are included in the two variable model?

car\_data = Auto

car\_data = car\_data[-c(8,9)] #Remove name and origin variables

summary(car\_data)

car\_best\_subset = regsubsets(mpg~.,data=car\_data,nvmax=6,method="exhaustive")

car\_best\_summary = summary(car\_best\_subset)

car\_best\_summary

* The two variables included in the two variable model are weight and year.

#6. Create a plot of the values of Adjusted-R^2 for each subset size (1 through 6). Between which two model sizes is the biggest change in Adjusted-R^2?

plot(car\_best\_summary$adjr2,xlab="Number of variables",ylab="Adjusted R-squared",main="Adj-R^2 values of models of size 1-6 found by best subset selection")

* The biggest jump in Adjusted-R^2 is between the one variable model and the two variable model. It jumps up from approximately 0.69 to approximately 0.81

# References

Data Set: <http://www.statsci.org/data/oz/physical.txt>

## R Code:

data = read.csv("http://www.statsci.org/data/oz/physical.txt", header=TRUE, sep="\t")

na.omit(data)

best\_subset = regsubsets(Mass~.,data=data,nvmax=10,method="exhaustive")

best\_summary = summary(best\_subset)

fwd\_subset = regsubsets(Mass~.,data=data,nvmax=10,method="forward")

fwd\_summary = summary(fwd\_subset)

bwd\_subset = regsubsets(Mass~.,data=data,nvmax=10,method="backward")

bwd\_summary = summary(bwd\_subset)

best\_summary

fwd\_summary

bwd\_summary

plot(best\_summary$adjr2,xlab="Number of variables",ylab="Adjusted R-squared",main="Adj-R^2 values of models of size 1-10 found by best subset selection")

plot(fwd\_summary$adjr2,xlab="Number of variables",ylab="Adjusted R-squared",main="Adj-R^2 values of models of size 1-10 found by forward stepwise subset selection")

plot(bwd\_summary$adjr2,xlab="Number of variables",ylab="Adjusted R-squared",main="Adj-R^2 values of models of size 1-10 found by backward stepwise subset selection")

car\_data = Auto

car\_data = car\_data[-c(8,9)] #Remove name and origin variables

summary(car\_data)

car\_best\_subset = regsubsets(mpg~.,data=car\_data,nvmax=6,method="exhaustive")

car\_best\_summary = summary(car\_best\_subset)

car\_best\_summary

plot(car\_best\_summary$adjr2,xlab="Number of variables",ylab="Adjusted R-squared",main="Adj-R^2 values of models of size 1-6 found by best subset selection")