MULTIPLE REGRESSION (Cont’d)

1. **Criterion based procedure with exhausted search:**

You will need a package called leaps.

Then, use these commands to do exhausted search

b = regsubsets(Native~Area+Elev+DistNear+DistSC+AreaNear,data=turtleData)

rs = summary(b)

rs$which

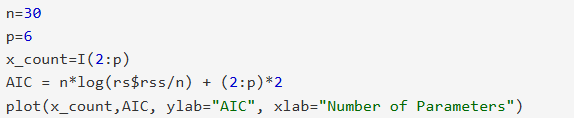
Look in rs variable to see what you have in there.

rs$which tell you the predictors (for each model size) to include such that the SSE is minimum.

Example 1: Use the data ex1220 from Sleuth2 to do model selection that fits the species of Native plant using Area, Elevation, Nearest Distance, Distance to Santa Cruz, and Area of the Nearest Island.

Then, plot the AIC, BIC, adjusted R2, and Cp values.

To plot AIC, use the following command where n is the number of data points, p is the number of parameters, AIC is the given formula for AIC, and x\_count is the vector from 2 to p (acts of number of parameters)



For other measures, check the rs variable above because the rs variable has all the other measures. Then, state what model you choose from here (how many predictors and what they are).

1. **Criterion based procedure with cheaper alternative:**

Sometimes, if you have many predictors to choose, it will be really time consuming doing exhausted search (2^(predictors)). Thus, we can use a cheaper alternative. It bears some comparison to the step wise methods, but only in the method of search — there is no hypothesis testing. The command is

allNumericalDataModel=lm(Native~Area+Elev+DistNear+DistSC+AreaNear,data=turtleData)

step(allNumericalDataModel,direction='forward')

Here, you can change direction from “forward” to “backward” to “both.”

Example 2: Do a model selection using forward, backward, and stepwise (both) using *step(*). What is your choice of the model in this case?

1. **Sensitivity**

The model selection procedures can be sensitive to outliers and influential points. Thus, we should check its sensitivity as well. First you need to figure out which point have a potential to be an outliers and influential points (there are 2-3 of them). You can do this by plotting out the diagnostic plot and look at the Cook statistic plot or you can compute the hat matrix as below

h <- lm.influence(allNumericalDataModel)$hat

names(h) =turtleData$Island

rev(sort(h))

Example 3: Figure out two potential influential/outlier points. Then do model selection again (both exhausted search and step () search without each of the point). Do you notice any difference before and after? To omit the point from the search, use

b\_12<-regsubsets(Native~Area+Elev+DistNear+DistSC+AreaNear,data=turtleData, subset=(Island!="Fernandina"))

or you can use the command similar to what we did before in the Simple Regression handout. Note that now you will only have n=29 data points.

1. **Partial plots and effect plots**

Example 4: Fits the species of Native plant using Area, Elevation, Nearest Distance, Distance to Santa Cruz, and Area of the Nearest Island. Then, fit the species of Native plant using Elevation only. Plot both regression lines on the same plot (Native plant species vs. Elev) to see the difference between the two lines. This is called effect plot.

Note, to plot the regression line from multiple regression model, you will need to first do a prediction in which other predictors are kept at constant (their mean values) and predict the response when elevation is changed.

Example 5: Do a partial regression plot following the steps listed in the lecture note to check the relationship between Native Species and Elevation

Example 6: Do a partial residual plot to check the relationship between Native Species and Elevation. In R, you can do it easily by where i is the index of the predictor.

termplot(Name of your Model, partial.resid=TRUE, terms=i)