PSTAT 126 - Regression Analysis – Spring 2017

Lab 9 Handout

Model Selection Techniques

**Note:** These course materials are the sole property of Dr. Todd M. Gross. They are strictly for use by students enrolled in a course taught by Dr. Gross. They may not be altered, excerpted, distributed, or posted to any website or other document-sharing service.

**Note**: Please complete this lab with your project partner, if they are available.

1. For this exercise you will perform model selection using the **pima** dataset in the **faraway** package. Use **glucose**  as the outcome variable, and **age, bmi, diabetes, diastolic,** and **insulin** as the possible predictors. Before performing the analysis, restrict the dataset to non-zero values on all variables. Do NOT include interaction terms in the model fitting.  
     
   **> data(pima,package="faraway")**

**> newpima = pima[,c(2,8,6,7,3,5)]**Note: This command selects only the columns (variables) needed for the exercise, and drops the rest

**> newpima=subset(newpima,newpima$glucose>0 & newpima$age>0 & newpima$bmi>0 & newpima$diabetes>0 & newpima$diastolic>0 & newpima$insulin>0)**

* 1. Generate a summary of the descriptive statistics for each variable, yielding the mean, median, 1st and 3rd quartile, and the minimum and maximum value (Hint: you only need one R command to create this output).  
       
     **> summary(newpima)**
  2. Generate a scatterplot matrix for all variables named above. Do not include other variables.  
       
     **> pairs(newpima,pch=19)**
  3. Fit the following full regression model to these data, using the order of predictors in the problem description above.   
      *Y = β0 + β1X1 + β2X2 + β3X3 + + β4X4 + β5X5 ε*   
     Does this model fit these data? What is the overall p-value? Which predictors that have significant **partial** SS in this model?   
        
     **> pimamodel1=lm(glucose~.,data=newpima)**Note: The model specification of ~. Includes all predictors in the 1st order model (no interactions) **> summary(pimamodel1)**
  4. Produce the anova table that gives the sequential SS using the order provided in the problem description above. Which predictors have significant **sequential** SS?  
       
     **> anova(pimamodel1)**
  5. Perform a stepwise regression using forward addition. What is the resulting regression model?

First, define two models, null and full. The ~1 is the null model; the full model includes all two-way interactions with ~.^2  
  
**> null=lm(glucose~1,data=newpima)**

**> full=lm(glucose~.^2,data=newpima)**Now, perform the stepwise regression with forward addition.

**>step(null,scope=list(lower=null,upper=full),direction='forward')**Note: **scope** gives the range of models to consider in the forward addition of variables - **lower** is the smallest model, and **upper** is the largest model

* 1. Now repeat a stepwise regression using backward elimination. What is resulting regression model? Is it different than the model obtained with forward addition?  
       
     **> step(full,direction='backward')**
  2. Finally, repeat a stepwise regression using **direction=”both”**. What is resulting regression model? Does it match the results from **backward elimination** or **forward addition**? What model do the three methods of stepwise regression converge on?  
       
     **> step(null,scope=list(upper=full),direction='both')**

R Commands

**data(pima,package="faraway")**

**newpima = pima[,c(2,8,6,7,3,5)]**

**newpima=subset(newpima,newpima$glucose>0 & newpima$age>0 & newpima$bmi>0 & newpima$diabetes>0 & newpima$diastolic>0 & newpima$insulin>0)**

**summary(newpima)**

**pairs(newpima,pch=19)**

**pimamodel1=lm(glucose~.,data=newpima)**

**summary(pimamodel1)**

**anova(pimamodel1)**

**plot(fitted.values(pimamodel1),residuals(pimamodel1),pch=19)**

**null=lm(glucose~1,data=newpima)**

**full=lm(glucose~.^2,data=newpima)**

**step(null,scope=list(lower=null,upper=full),direction='forward')**

**step(full,direction='backward')**

**step(null,scope=list(upper=full),direction='both')**