PSTAT 126 - Regression Analysis - Fall 2015

Lab 7 Handout

Extra Sum of Squares

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

Goals for this Lab

* Learn to perform some basic matrix multiplication
* Learn to read in data from a text file
* Perform and interpret sequential and partial Sum of Squares tests
* Test the contribution of a subset of predictors

Lab Exercise #1

In order to understand how matrix algebra works, here are a few basic matrix algebra problems for you to work by hand. The answers are provided so that you can check your work.

A = B = C= D = E =

A) A \* BT = B) BT \* A = C) D\*E =

Lab Exercise #2

For the course data project you will need to read in data from a text file that is stored on Gauchospace. We will practice this by reading in the Dwaine.txt file that was used in Lecture 8.

1. Click on the Dwaine.txt file on Gauchospace and save the text file to a location on your hard drive.
2. Read the data into R using the read.table command. The text file has variable names in the first row, so use the Header = TRUE option. You will also need to know the path of the location where you saved the text file. For example, if you save the file to the Downloads directory, you will use:  
     
   Dwaine=read.table("C:/Users/Gross/Downloads/Dwaine.txt",header=TRUE)
3. Confirm that you have read the data correctly by generating a summary of the variables using summary(Dwaine).

Lab Exercise #3

We will now fit several different linear models using the **pima** data set. Note that for these exercises you can assume that the assumptions of the linear model are valid. You do NOT need to evaluate diagnostics.

1. Open the **pima** dataset in the **faraway** package.
2. Create a new dataset (name it **newpima**)that uses the subset command to remove values of zero on any of five variables (**diastolic**, **glucose**, **insulin, bmi**,and **age**).
3. Fit the full linear model predicting **diastolic** from **glucose**, **insulin, bmi**,and **age**. Be sure to enter the predictors in that order.
4. Generate the summary output for the **fitnewpima** linear model. What is R2 for this model? Is the overall p-value significant for this model?

Partial Sum of Squares (SS)

The partial SS for any predictor reflects the contribution of that predictor given all other predictors in the model. The partial p-values can be obtained directly from the summary linear model output. Use the Summary output to answer the following questions:

* 1. Is glucose significant given insulin, BMI, and age? What is the p-value?
  2. Is insulin significant given glucose, BMI, and age? What is the p-value?
  3. Is age significant given glucose, insulin, and BMI? What is the p-value?

Sequential Sum of Squares (SS)

The sequential SS for any predictor reflects the contribution of that predictor given earlier predictors in the model. Therefore, sequential SS depend on the order of the predictors in the lm statement. The anova command provides sequential SS and p-values.

1. Generate ANOVA output for the **fit.full** model using the ANOVA command. Answer the following questions:
   1. Is glucose significant by itself? What is the p-value?
   2. Is insulin significant given glucose? What is the p-value?
   3. Is age significant given glucose, insulin, and BMI? What is the p-value? Is this p-value the same as its partial p-value? Why or why not?

Testing a subset of predictors

We can also test whether two or more predictors together make a contribution. We do this by fitting a reduced model that does not include the predictors to be tested, and then using ANOVA to compare it to the full model that does contain the additional predictors.  
  
We want to test whether **glucose** and **insulin** contribute to the model given that **BMI** and **age** are already in the model.

1. Fit a reduced model (name it **fit.reduced**) that predicts **diastolic** from **BMI** and **age**. What is R2 for this model? Is the overall p-value significant for this model?
2. Compare this R2 to the R2 obtained for the **fit.full** model above? What is the difference in R2?
3. Use the anova command to compare the two models. What is the SS(glucose,insulin|age,bmi)? What is the p-value? Do **glucose** and **insulin** make a significant contribution when added to **BMI** and **age**?

R Commands

Dwaine=read.table("C:/Users/Gross/Downloads/Dwaine.txt",header=TRUE)

summary(Dwaine)

newpima<-subset(pima,age>0 & bmi>0 & glucose>0 & insulin>0 & diastolic>0)

fit.full<-lm(diastolic~glucose+insulin+bmi+age,data=newpima)

summary(fit.full)

anova(fit.full)

fit.reduced<-lm(diastolic~bmi+age,data=newpima)

anova(fit.reduced,fit.full)