PSTAT 126 - Regression Analysis – Spring 2017

Lab 6 Handout

Multiple Linear Regression

**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.

Goals for this Lab

* Fit a multiple linear regression model, produce summary output and interpret the resulting model.
* Test the hypothesis that all slopes are zero, then test hypotheses that each slope equals zero.
* Produce a scatterplot matrix containing all variables in the linear model.
* Produce residuals plots appropriate for diagnosing possible violations of the assumptions of the general linear model.

Lab Exercise

We will fit a linear model using the **pima** data set. We will use three variables (**age**, **bmi**, and **insulin**) to predict **diagnostic** blood pressure. The R commands for this lab appear below, but you should attempt to program the linear model yourself, before looking at the code provided.

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 the four variables (**diastolic**, **age**, **bmi**, and **insulin**).
3. Fit a linear model (name it **fitnewpima**) that predicts diastolic from age, bmi, and insulin.
4. Generate summary output for the **fitnewpima** linear model.
   1. Write out the linear regression equation using the estimates of intercept and slopes you obtained. How would you interpret the slope for bmi?
   2. Test the null hypothesis (H0: **1 = **2 = **3 = 0) against the alternative (H1: not all **j = 0). State the value of *F*, the corresponding p-value, and the statistical conclusion. What can you conclude about the linear model?
   3. Interpret the value for multiple R2 (not adjusted R2).
   4. Test the hypothesis that the slope for age equals zero. State the value of *t*, the corresponding p-value, and the statistical conclusion. Repeat for BMI and insulin. Do all of the predictors have non-zero slopes?
5. Generate a scatterplot matrix for the four variables (diastolic, age, bmi, and insulin). What do you see in the plots of diastolic against each of the three predictors?
6. Generate the following data displays:
   1. Residuals plot with fitted values on the X axis, residuals on the Y axis and a horizontal line at zero
   2. QQ normal plot, with QQ line
   3. Histogram of residuals
   4. Use these three plots to evaluate non-linearity, non-constant variance, non-normality, and outliers. Do you see any issues?

R Commands

newpima=subset(pima,diastolic>0 & age>0 & bmi>0 & insulin>0,package="faraway")

fitnewpima=lm(diastolic~age+bmi+insulin,data=newpima)

summary(fitnewpima)

pairs(diastolic~age+bmi+insulin,data=newpima)

plot(fitted(fitnewpima),residuals(fitnewpima))

abline(h=0)

qqnorm(residuals(fitnewpima))

qqline(residuals(fitnewpima))

hist(residuals(fitnewpima))