PSTAT 126 - Regression Analysis – Fall 2017

Lab 1 Handout

**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.

Goals for this Lab

Learn how to use R to:

* Load a dataset
* Use descriptive statistics to summarize a variable
* Create a new dataset containing a subset of values
* Perform a linear regression and display a summary of the regression statistics
* Create a scatterplot that includes the regression line

Lab Exercise #1

We will use an existing data package, called **faraway**. We are going to use a dataset called **pima** that is in the **faraway** package.

We want to look at the relationship between two variables: body mass index, or **bmi,** and **glucose. BMI** is a measure of the amount of body fat you have. In general, the higher your body fat, the higher your risk of diabetes, which causes your fasting **glucose** to rise.

A description of the **pima** dataset is available in R. All of the R commands, output, and answers to the questions are provided for this example.

1. Install the **faraway** package and open the **pima**  dataset.

> install.packages('faraway')

> data(pima,package="faraway")

1. Display the names of the variables in the **pima** dataset

> head(pima)

1. Display summary statistics for each of the variables

> summary(pima$bmi)

> summary(pima$glucose)

1. Note that there are zero values on the **glucose** and **bmi** variables. There are data errors (or missing data). We need to remove those in order to perform our analysis. We will create a new dataset, called **newpima** that excludes the zero values.

> newpima<-subset(pima,pima$bmi>0 & pima$glucose>0)

1. Confirm that the zero values have been removed. Note that we have changed to the **newpima** dataset.

> summary(newpima$bmi)

> summary(newpima$glucose)

1. Create a histogram for **bmi** with appropriate labels

> hist(newpima$bmi,xlab='BMI',main='Histogram of BMI')

1. Now calculate the regression coefficients for predicting **glucose** from **age**. We do this by fitting a model, which we will call **fit1**. Note that we identify the dataset using **data=newpima** option. Then we summarize the **fit1** model to see the results.

> fit1=lm(glucose~bmi, data=newpima)

> summary(fit1)

1. Create a scatterplot with **age** on the X axis, and **glucose** on the Y axis. Use appropriate labels. Then add the regression line using the **abline** command.

> plot(newpima$bmi,newpima$glucose,xlab='BMI',ylab='Glucose', main='Scatterplot of BMI and Glucose')

> abline(fit1)

R Commands Used in this Lab

install.packages('faraway')

data(pima, package='faraway')

head(pima)

summary(pima$bmi)

summary(pima$glucose)

newpima<-subset(pima,pima$bmi>0 & pima$glucose>0)

summary(newpima$bmi)

summary(newpima$glucose)

hist(newpima$bmi,xlab='BMI',main='Histogram of BMI')

fit1=lm(glucose~bmi, data=newpima)

summary(fit1)

plot(newpima$bmi,newpima$glucose,xlab='BMI',ylab='Glucose', main='Scatterplot of BMI and Glucose')

abline(fit1)