PSTAT 126

Regression Analysis

Fall 2017

Homework #1 – Due in Your Enrolled Section Oct 10-12

**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 without express permission.

1. Write out the formal regression equation, and list the assumptions of the model.
2. An instructor measures the number of classes missed and the final exam score for each student in his course. She then performs a regression analysis predicting final exam score from number of classes missed. She finds the following regression equation:

Final exam score = 85 – 5.3X

* 1. What is the direction of the relationship between exam score and classes missed?
  2. What exam score would we predict for a student who misses 6 classes?
  3. Describe the values of the slope (-5.3) and intercept (85) that the instructor obtained, in words?

1. What is the difference between the formal regression model and the equation for the predicted value of Y (the fitted regression equation)?
2. What is minimized in the method of least squares?
3. Evaluate the following statement: “For the least squares method to be fully valid, it is required that the distribution of Y be normal.”
4. The members of a yoga studio pay annual membership dues of $250 plus a charge of $5 for each visit to the studio. Let Y denote the yearly dollar cost a member and X the number of visits by the member during the year. Express the relation between X and Y mathematically. Is it a functional relation or a statistical relation?

Use R to complete the following problems. **Include your R code (commands only) at the end of your answers.**

1. The **pima** dataset is described in the Faraway package in R. We are interested in the relationship between **age** and **glucose** (Note: you will need to remove zero values before completing the analysis).
   1. For each variable:
      1. Create a histogramwith appropriate labels for the X and Y axis
      2. Calculate the mean and standard deviation.
   2. Fit a simple linear regression model predicting **glucose** from **age**.  
      1. Give the values for the slope and intercept. Interpret these values in words.
      2. Give the value for R2. Based on this value, how does knowing age effect the variance of glucose?
   3. Plot glucose as a function of age using appropriate titles for the X and Y axis. Add the regression line to the plot.
   4. Provide the R commands you used.
2. We are interested in predicting the Son’s adult height (**childheight)** from the Father’s height (**father)** using the **GaltonFamilies** dataset in the **HistData** package. Note: you will need to select the male subjects from the data set using the following R command:

>menheight<-subset(GaltonFamilies,GaltonFamilies$gender=="male")

* 1. For each variable:
     1. Create a histogramwith appropriate labels for the X and Y axis
     2. Calculate the mean and standard deviation.
  2. Fit a simple linear regression model predicting **childheight** from **father**.  
     1. Give the values for the slope and intercept. Interpret these values in words.
     2. Give the value for R2. Based on this value, how does knowing the father’s height affect the variance of predicting the son’s adult height?
  3. Plot the sons’ adult heights as a function of the fathers’ heights using appropriate titles for the X and Y axis. Add the regression line to the plot.
  4. Provide the R commands you used.