PSTAT 126

Regression Analysis

Spring 2017

Homework #1 – Due in Section Wednesday, April 19

Points total = 33. Note: ½ point deductions may be taken.

1. (4 pts) Write out the formal regression equation, and list the assumptions of the model.

Y’ = *0 + 1*X + 

Assumptions (from lecture slide ) are:

1. Y is a random variable with E[Y] = 0 + 1X
2. X is known and measured without error
3. The values of Y are independent, Cov(Yi,Yj) = 0, or Cov(i,j) = 0
4. (4 pts) 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 = 88 – 5.6X

* 1. What is the direction of the relationship between exam score and classes missed?

Negative, because the value of the slope is negative

* 1. What exam score would we predict for a student who misses 4 classes?

88 – 5.6(6) = 54.4

* 1. Describe the values of the slope (-5.6) and intercept (88) that the instructor obtained, in words?

Slope: We predict that exam score will decrease 8.2 points for each additional class missed.

Intercept: We predict that a student who misses zero classes will score 93 on the exam.

1. (1 pt) What is the difference between the formal regression model and the equation for the predicted value of Y?

The formal regression model includes an error term, or residual, 

1. (1 pt) What is minimized in the method of least squares?

The method of least squares minimizes the squared difference between Y and Y’

1. (1 pt) Evaluate the following statement: “For the least squares method to be fully valid, it is required that the distribution of Y be normal.”

The statement is false because the least squares method does not assume (or require) that Y be normally distributed in order to be valid.

1. (2 pts) The members of a yoga studio pay annual membership dues of $250 plus a charge of $5 for each visit to the spa. Let Y denote the dollar cost for the year for 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?

Cost = $250 + $5(number of visits). This is a functional relationship.

Use R to complete the following problems. **Include your R code (commands only) at the end of your answers. If R Code is NOT provided, deduct 5 points.**

1. (10 pts total) The **pima** dataset is described in the Faraway package in R. We are interested in the relationship between **age** and **glucose** (Note: you need to remove zero values before completing the analysis).

Note: The glucose variable has a few values of zero. Students should remove these before analysis. If they don’t, **take a 3 point deduction**.

* 1. For each variable:
     1. (2 pts) Create histogramwith appropriate labels for the X and Y axis

 

Note: The glucose variable has a few values of zero. Students should remove these before further analysis, but there is no deduction if they don’t.

* + 1. (2 pts) Calculate the mean and standard deviation.

With zeroes removed

Age: mean= 33.27, SD=11.77

Glucose: 121.69, sd=30.54

With zeroes included

Age: mean= 33.24, SD=11.76

Glucose (with zeroes included): 120.89, sd=31.97

* 1. Fit a simple linear regression model predicting **glucose** from **age**.  
     1. (2 pts) Give the values for the slope and intercept. Interpret these values in words.

Zeroes removed

Y’ = 98.63 + 0.693(X) – zeroes removed.

We predict that glucose will increase 0.693 points for each additional year of age. The intercept has a value of 98.63, but does not have a meaningful interpretation because Age cannot practically equal zero.

Zeroes included

Y’ = 97.08 + 0.716(X)

We predict that glucose will increase 0.716 points for each additional year of age. The intercept has a value of 97.08, but does not have a meaningful interpretation because Age cannot practically equal zero.

* + 1. (2 pts) Give the value for R2. Based on this value, how does knowing age effect the variance of glucose?

Zeroes included

R2 = 0.07136; 7% of the variance in glucose can be explained by knowing an individual’s age.

Zeroes removed

R2 = 0.069; 7% of the variance in glucose can be explained by knowing an individual’s age.

* 1. (2 pts) Plot glucose as a function of age using appropriate titles for the X and Y axis. Add the regression line to the plot.



d. Provide the R commands you used.

R Code:

newglucose<-subset(pima,pima$glucose>0)

attach(newglucose)

hist(age)

hist(glucose)

plot(age,glucose)

lm(glucose~age)

abline(lm(glucose~age))

summary(lm(glucose~age))

1. (10 pts total, Same points breakdown as problem 7) We are interested in predicting Son’s height from Father’s height using the **GaltonFamilies** dataset in the **HistData** package. Repeat each of the items from Problem 7 for these data. Note: you will need to select the male subjects from the data set using one of the following R commands:

>menheight<-GaltonFamilies[which(GaltonFamilies$gender=="male"),]

OR

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

* 1. Descriptive statistics
     1. Histograms Note the correct use of labels.

 

* + 1. Mean, sd

Father: mean=69.14, SD=2.31; Son: mean=69.23, SD=2.62

* 1. Linear model
     1. Slope and intercept

Y’ = 38.36 + 0.447(X).

We predict that Son’s height will increase 0.447 inches for each additional inch in Father’s height. The intercept does not have a meaningful interpretation because Father’s height cannot practically equal zero.

* + 1. R2

R2 = 0.154; 15.4% of the variance in Son’s height can be explained by knowing Father’s height.

* 1. Scatterplot Note the correct use of labels.



d. Provide the R commands you used.

R Code

install.packages("HistData")

data(GaltonFamilies,package='HistData')

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

attach(menheight)

hist(childHeight,xlab="Son's Height (inches)")

hist(father,xlab="Father's Height (inches)")

lm(childHeight~father)

plot(father,childHeight,xlab="Father's Height (inches)",ylab="Son's Height (inches)")

abline(lm(childHeight~father))

summary(lm(childHeight~father))