PSTAT 126

Regression Analysis

Spring 2017

Homework #2 – Due in Section Wednesday, April 26, 2017

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.

*Use R to perform data analyses, and provide annotated code and output.*

1. Consider a dataset with five pairs of data. The values of X are {1,2,3,4,5}. Create a set of Y values that have the following properties. Use R to confirm.
   1. Create a set of Y values that results in an intercept of 0 and a slope of 1.
   2. Create a set of Y values that results in an intercept of 10 and a slope of 1.
   3. Create a set of Y values that results in an intercept of 5 and a slope of 3.
   4. Create a set of Y values that results in a p-value between 0.1 and 0.15 for the hypothesis test of H0: 1 = 0. (Hint: you may need to try several different sets of Y values to obtain this).
2. Consider a hypothesis test for  1 = 0.
   1. Write the null and alternative hypotheses for a two-tailed test.
   2. Write the null and alternative hypotheses for a one-tailed test.
3. Refer to Problem 7 of Homework #1 where we used the **pima** dataset to analyze the relationship between **age** and **glucose**.
   1. Use the summary of the linear model that you produced for Homework #1 to test the hypothesis that the slope = 0.
      1. State the null and alternative hypotheses
      2. State the value of t, the standard error and the p-value
      3. Can you reject the null hypothesis? State why or why not.
      4. What does your statistical conclusion tell you about the population slope?
      5. What is the value for R2? What does this tell you about the amount of variance explained?
   2. Repeat the hypothesis test using the ANOVA function in R.
      1. Use the Anova output from R to complete the following ANOVA source table. Note: you may need to calculate some of the numbers in the table using the numbers in the R output.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **SS** | **df** | **MS** | **F** | **p** |
| Regression |  |  |  |  |  |
| Error |  |  |  | --- | --- |
| Total |  |  | --- |  |  |

* + 1. What is the p-value for the F-test?
    2. Can you reject the null hypothesis based on the ANOVA output from R? State why or why not?
  1. Calculate a 90% confidence interval for the slope. What are the lower and upper limits of this interval? Make a statement about the value of the population slope based on the 90% CI.

1. Refer to problem 9 from Homework #1 where we predicted Son’s height from Father’s height using the **GaltonFamilies** dataset in the **HistData** package. Repeat **all parts** of Problem 3 above using these data. Be sure to restrict the dataset to the male children.