Northeastern University
College of Science

Module 9 – Homework

Problem 1 (25 points)

On the Golub et al. (1999) data set, find the expression values for the GRO2 GRO2 oncogene and the GRO3 GRO3 oncogene. (Hint: Use grep() to find the gene rows in golub.gnames. Review module 2, or page 12 of the textbook on how to do this. Be careful to search *only in the column with gene names*.)

(a) Find the correlation between the expression values of these two genes.

(b) Find the parametric 90% confident interval for the correlation with cor.test(). (Hint: use ?cor.test to learn how to set the confidence level different from the default value of 95%.)

(c) Find the bootstrap 90% confident interval for the correlation.

Problem 2 (25 points)

On the Golub et al. (1999) data set, we consider the correlation between the Zyxin gene expression values and each of the gene in the data set.

(a) How many of the genes have correlation values less than negative 0.5? (Those genes are highly negatively correlated with Zyxin gene).

(b) Find the gene names for the top five genes that are most negatively correlated with Zyxin gene.

(c) Using the t-test, how many genes are negatively correlated with the Zyxin gene? Use a false discovery rate of 0.05. (Hint: use cor.test() to get the p-values then adjust for FDR. Notice that we want a one-sided test here.)

Problem 3 (30 points)

On the Golub et al. (1999) data set, regress the expression values for the GRO3 GRO3 oncogene on the expression values of the GRO2 GRO2 oncogene.

(a) Is there a statistically significant linear relationship between the two genes' expression? Use appropriate statistical analysis to make the conclusion. What proportion of the GRO3 GRO3 oncogene expression's variation can be explained by the regression on GRO2 GRO2 oncogene expression?

(b) Test if the slope parameter is less than 0.5 at the $\alpha = 0.05$ level.

(c) Find an 80% prediction interval for the GRO3 GRO3 oncogene expression when GRO2 GRO2 oncogene is not expressed (zero expression value).

(d) Check the regression model assumptions. Can we trust the statistical inferences from the regression fit?

Problem 4 (20 points)

For this problem, work with the data set stackloss that comes with R. You can get help on the data set with ?stackloss command. That shows you the basic information and source reference of the data set. Note: it is a data frame with four variables. The variable stack.loss contains the ammonia loss in a manufacturing (oxidation of ammonia to nitric acid) plant measured on 21 consecutive days. We try to predict it using the other three variables: air flow (Air.Flow) to the plant, cooling water inlet temperature (C) (Water.Temp), and acid concentration (Acid.Conc.)

(a) Regress stack.loss on the other three variables. What is the fitted regression equation?

(b) Do all three variables have statistical significant effect on stack.loss? What proportion of variation in stack.loss is explained by the regression on the other three variables?

(c) Find a 90% confidence interval and 90% prediction interval for stack.loss when Air.Flow=60, Water.Temp=20 and Acid.Conc.=90.