Math 216
Problem Set 5

Part One

The Tennessee Valley Authority (TVA) receives data from rain gauges all over Tennessee, some operated by the TVA, others operated by other organizations. Data for the months of February and March 2012 for two of these gauges, one located in Manchester, Tennessee, and one located at the Nickajack Dam, have been compiled in a CSV file available here:

`http://derekbruff.org/blogs/math216/files/2012/04/Rain-Gauge-Data.csv`

Given their proximity, it is reasonable to expect that rainfall amounts in these two locations would be correlated. Use the above file and R/R Studio to answer the following questions about the relationship between rainfall in Manchester and at the Nickajack Dam over this time period.[1]

1. What is the equation of the least-squares line for these data? Your answer should be in the form of $y = \beta_0 + \beta_1 x$, where y is the rainfall at the Nickajack Dam and x is the rainfall in Manchester. (Reminder: Several R commands relevant to linear regression are illustrated on the course blog, `http://derekbruff.org/blogs/math216/?p=498`.)
2. Construct a scatterplot for these data (again, with x = Manchester rainfall and y = Nickajack Dam rainfall) that also shows the least-squares line you found in Question 1.
3. Determine the correlation coefficient for these data. What does this coefficient say about the possible linear relationship between these two variables?
4. Use the R function `residuals`, which works a bit like the function `summary`, to generate a residual plot for these data, one based on the least-squares line you found in Question 1.
5. Approximately what percent of the variance in the rainfall at the Nickajack Dam can be explained by a linear relationship between rainfall there and rainfall in Manchester?
6. Do these data provide strong evidence that there is a positive linear relationship between rainfall in Manchester and at Nickajack Dam? State the null and alternate hypotheses, report the p-value, and state your conclusion.
7. Given your linear regression model, if Manchester receives an additional inch of rain beyond what's forecasted on a given day, how much additional rain is the Nickajack Dam likely to receive?
8. Given the residual plot you constructed in Question 4, what concerns should you have about applying least squares to these data? (See §7.2.2 for a list of possible concerns.) Use your residual plot to justify each concern you raise.

Part Two

The website Information Is Beautiful compiled data on every Hollywood film released in the last five years.[2] Included in these data are the average critic scores (on a scale of 0 to 100) and average audience scores (also on a scale of 0 to 100) for each movie, drawn from the website Rotten Tomatoes. Critic and audience scores for 2011 releases can be found in the CSV file available here:

`http://derekbruff.org/blogs/math216/files/2012/04/Hollywood.csv`

Use the above file and R/R Studio to analyze the relationship between critic scores and audience scores for Hollywood films.

[1] Hat tip to Hayden Kelly for finding these data and bookmarking them in our Diigo group.

[2] See `http://www.informationisbeautifulawards.com/2012/01/challenge-of-the-stars/`.

1. What is the equation of the least-squares line for these data? Your answer should be in the form $y = \beta_0 + \beta_1 x$, where x is the average critic score and y is the average audience score.

2. Construct a scatterplot for these data (again with x = average critic score and y = average audience score) that also shows the least-squares line you found in Question 1.

3. Determine the correlation coefficient for these data. What does this coefficient say about the possible linear relationship between these two variables?

4. Construct a residual plot for these data based on your least-squares lines. Do you have any concerns about applying least squares to these data? Justify your answer using your residual plot.

5. Approximately what percent of the variance in average audience scores can be explained by a linear relationship between average audience scores and average critic scores?

6. Do these data provide strong evidence that there is a positive linear relationship between average critic scores and average audience scores? State the null and alternate hypotheses, report the p-value, and state your conclusion.

7. Do these data provide strong evidence that the linear relationship between average critic score and average audience score for all Hollywood movies has a slope greater than 0.50? State the null and alternate hypotheses, report the p-value, and state your conclusion.

8. If the average audience score for a given movie is 20, what does your model predict for the average critic score for this movie? Why is this prediction problematic?