Comp 528 Computer Systems Performance Analysis
Spring 2005

Homework Assignment 2
Due: February 24

Analyzing Sample Data

This assignment focuses on data analysis with an emphasis on regression models. I encourage you to use Matlab for this assignment as it's interactive interface is well-suited to the analysis and visualization tasks that the problems require. Each problem part will have equal weight.

1. **Index of central tendency and dispersion.** The file `http://www.cs.rice.edu/~johnmc/comp528/homework/hw2p1` contains a set of round trip network latencies measured between Rice University and an off-campus site connected to the internet by cable modem.
   (a) Which index of central tendency would you choose to summarize the latencies and why?
   (b) Which index of dispersion would you choose to summarize the latencies and why?

2. **Confidence intervals, sample size and sample distribution.** On a 900MHz Itanium2 processor, 35 trials were performed using gzip to compress a file of 3,195,080 bytes, yielding the following set of running times (in seconds): {0.96, 0.90, 0.91, 0.94, 0.90, 0.91, 0.91, 0.92, 1.45, 0.92, 1.03, 1.04, 0.92, 0.96, 0.90, 0.92, 0.93, 0.93, 0.91, 1.23, 0.91, 0.90, 0.91, 0.90, 0.95, 0.92, 0.97, 0.90, 1.16, 0.92, 1.09, 0.92, 0.92, 1.40, 1.13}. A data file containing the set of times is available at `http://www.cs.rice.edu/~johnmc/comp528/homework/hw2p2`.
   (a) Compute a 90% confidence interval for the compression times.
   (b) How many trials would be necessary to estimate the mean within .01 seconds with 95% confidence?
   (c) A quantile-quantile plot is useful for understanding the distribution of measured values. Plot the measured times on a quantile-quantile plot against N(0,1). Interpret the plot (i.e., are the measurements normally distributed, short tailed, long-tailed, etc.). What conclusions can you draw about the distribution of compression execution times?

3. **Comparing file compression utilities.** Different file compression algorithms are effective to varying degrees, depending upon the size of the file and the characteristics of the data. File `http://www.cs.rice.edu/~johnmc/comp528/homework/hw2p3` contains three columns of data. Column 1 contains the size of an original file. Column 2 contains the size of the file after compression by gzip. Column 3 contains the size of the file size after compression with bzip2.
   (a) What is an appropriate mean to compare the effectiveness of the two compression utilities? Compute the mean compression ratio for gzip and bzip2. Interpret your results.
   (b) Compare bzip2 and gzip compression using paired sampling. Is one compression algorithm better at the 90% confidence level? Is paired sampling appropriate for analyzing this data? What do we know and what don't we know from analyzing this data using paired sampling?
   (c) For each compression algorithm, use linear regression to model the size of a compressed file in terms of the original file size. On a single graph, for each compression algorithm plot the compressed file size vs. the original file size and a line representing the regression model.
   (d) Use the visual tests presented in class and described in Section 14.7 of Jain to verify the regression assumptions of the model in part 3c. Show your plots. Is the behavior of each of the compression algorithms adequately modeled by a linear regression? If not, why not?

(e) Select and apply an appropriate transformation to compute a better regression model for each compression algorithm. (See Sections 15.3–15.4 in Jain for inspiration.) Report the transformation you picked, why you picked it, and the resulting regression models.

(f) Use visual tests to verify regression assumptions of the models computed in part 3e and report your findings. How do the results of these tests differ from those in part 3c?

(g) Apply the inverse of the transformation to the regression models of part 3e and plot the resulting models with the original data as before.

(h) Compute the coefficient of determination for the models you developed in part 3c and 3e. How much of the variation is accounted for by each regression?

(i) Compute the 90% confidence interval for b_1 for each compression algorithm. Is the compression ratio significantly different according to the regression? What are your thoughts about the accuracy and utility of the regression models?

4. **Modeling communication bandwidth** On an an HP Itanium2 cluster with a Quadrics Elan4 interconnect, the table below shows communication bandwidths measured using the MPI message passing library and ARMCI, a multiplatform library for one-sided communication for different data transfer sizes.

KB	MB/s	MB/s
0.5	56.31	49.57
1	108.11	94.56
2	187.68	168.8
4	342.58	280.03
8	483.32	418.38
16	606.58	555.07
32	732.87	666.41
64	778.61	738.43
128	804.54	781.78

Communication time can be modeled as $c_0 + c_1 x$, where c_0 reflects the fixed communication overhead, c_1 reflects the time per byte transferred, and x reflects the transfer size.

(a) Write an equational model for communication bandwidth.

(b) Transform the variables to yield a linear equational model. (See Section 15.3 in Jain for inspiration.)

(c) Apply linear regression analysis to estimate c_0 and c_1 for both MPI and ARMCI.

(d) Calculate the percentage of the variation explained by each of the regressions.

(e) What are the 90% confidence intervals for c_0 and c_1 for MPI and ARMCI?

(f) Are the estimates for c_0 and c_1 significantly different for MPI and ARMCI? If so, at what significance level?