Fundamentals of Data Analysis: Assignment #11

Deadline: 1/19/2004 (Monday)

Please post to the mailbox next to the IS management office (2nd floor of IS building)

1. Run a numerical experiment according to the following steps.

a. Determine arbitrary constants b_0, b_1, and b_2.

b. Determine 10 pairs of appropriate values for two variables x_1 and x_2.

c. Calculate $y = b_0 + b_1 x_1 + b_2 x_2 + c\varepsilon$ for every pair of x_1 and x_2 chosen in step b, where ε is a random number obeying a standard normal distribution. You can make this random number by subtracting 6 from sum of 12 samples of uniform random values in the range (0, 1), as you did in the previous assignment.

d. Estimate b_0, b_1, and b_2 from the data made in steps b and c, according to the linear regression method, and compare the estimates with the true values which you determined in step b.

e. Calculate the mean squared error (D^2), the unbiased estimate of the error, and the coefficient of determination (R^2).

f. Calculate R^2 for various values of c, and examine how R^2 changes dependent on c. You should choose appropriate values of c so as to obtain meaningful results. Note that it is also important to repeat the experiment for an identical c with different settings.

g. Calculate $z = b_0 + b_1 x_1 + b_2 x_1^2 + c\varepsilon$ for 10 values of x_1 which you determined in step b. Choose the constant c so that noise has some effect on z.

h. Estimate the coefficient a_i of the following four models, using the linear regression, and calculate the mean squared error (D^2), the coefficient of determination (R^2) and AIC($= n \log D^2 + 2p$). Compare the results among the models.

   1) $z = a_0 + a_1 x_1$,
   2) $z = a_0 + a_1 x_1 + a_2 x_1^2$,
   3) $z = a_0 + a_1 x_1 + a_2 x_1^2 + a_3 x_1^3$, and
   4) $z = a_0 + a_1 x_1 + a_2 x_1^2 + a_3 x_1^3 + a_4 x_1^4$

i. Run the experiments for different noise and x_1, and examine how well the minimum AIC model agrees with the true model (i.e., 2)).

2. Test the hypothesis that the mean value of the following three groups are the same, using a significance level $\alpha = 0.01$. Is it appropriate to run a HSD test (i.e., post hoc test) for these data? Do the test if appropriate. Please refer to the homepage for the table of q distribution (studentized range statistic).

Group 1	8	7	10	9	11
Group 2	6	8	9	10	7
Group 3	7	11	11	12	9

3. You want to know whether there is any difference in lifetime among three brands of light bulbs. You get 6 samples for each bland and measure their lifetime. The following table shows the data.

a. Test the hypothesis that average lifetimes of three brands are the same, using a significance level $\alpha = 0.05$.

b. Is it appropriate to perform a HSD test for these data? Run the test if so.

Brand 1	23	20	27	25	23	25
Brand 2	33	30	29	28	31	32
Brand 3	21	24	29	27	28	26

4. Write your comments and requests on this lecture (if any).

END.