MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Civil and Environmental Engineering

1.017 Computing and Data Analysis for Environmental Applications

Quiz 3
Tuesday, December 10, 2002

Please answer all questions on a separate piece(s) of paper with your name clearly identified:

Problem 1 (15 points)

A field survey attempts to relate nitrogen runoff (in 1000 kg/year) in a small watershed to 2 factors: A) pesticide use (none, light, moderate, or heavy) and B) soil type (sandy, sand-silt, silty clay, or clay). Answer the following questions about the two-way ANOVA table given below:

Source	SS	df	MS=SS/df	$\mathcal{F}$	p
Factor A	0.0403	3	0.0134	0.6992	0.5661
Factor B	0.3044	3	0.1015	5.2856	0.0100
Interaction AB	0.3265	9	0.0363	1.8896	0.1278
Error	0.3072	16	0.0192		
Total	0.9783	31			

1. Which, if any, of the two factors contributes significantly to nitrogen runoff? Why?
2. Are interactions between factors A and B significant? Why?

Problem 2 (15 points)

Species diversity indices are frequently used to measure the health of natural ecosystems. Suppose that you are given the following two sets of unitless diversity indices for a control site and a site affected by human activity:

Control: 2.72 1.98 8.13 0.42 4.17 6.66

Affected: 2.78 3.02 0.93 3.21

Use the normal probability plot provided at the end of this quiz to determine the *p* value for a two-sided large-sample test of the null hypothesis that the two means are the same. Do you think the control and affected populations are significantly different? Why? Is the large-sample assumption valid here? Why?

Problem 3 (25 points)

Provide specific one-sentence answers to the following questions:

a) Why did we use the transformation $C_T = ln(C+1)$ when carrying out an ANOVA of Boston Harbor coliform data?

b) Explain the difference between independent and dependent variables in a regression analysis.

c) Why is the sum-of-squared differences between the observed and modeled dependent variables a reasonable measure of "goodness-of-fit"? Suggest at least one other measure that could also be reasonable in some applications.

d) Suppose that there is a significant linear relationship between the area of a temperate watershed and the average annual runoff. Is a set of average annual runoff values selected at random from temperate watersheds of different sizes a random sample? Why?

e) Suppose that the sample mean of a set of 5 data points $x_1,\ldots,x_5$ is 6.0 and the sample standard deviation is 2.5. Compare the *p* values obtained from small and large sample tests of the hypothesis H0: $E[x]=0$ (i.e. which test will give a larger p value?). You do not need to compute the actual p values …. just rank the two possibilities.

Problem 4 (20 points)

Consider the EPA NOx (nitrous oxide) emissions data set attached to this quiz. Organize the data into groups appropriate for a one-way analysis of variance that tests the influence of fuel type on NOx emissions. Use only three replicates for each treatment level. Indicate directly on the data sheet (by writing two numbers at the end of the appropriate row) the treatment level and replicate for each of the sample you select for your ANOVA.

Next use your annotated data sheet to read off the values of the input data array required by the MATLAB function `anova1` (documentation attached). Please read the documentation carefully to make sure that you define the array properly.

Finally, construct a one-way ANOVA table with the degree of freedom values filled in for all table rows. Identify the quantity (e.g. sum-of-squares, etc.) that goes in each of the remaining table entries but do not compute numerical values for any quantities other than the number of degrees of freedom.

Problem 5 (25 points)

Suppose that you are given the following data describing the concentration (in mg/L) of decaying organic carbon remaining in a treatment tank after the indicated time has passed:

time	10	12	15	17	22	24
concentration	1.77	0.88	1.08	0.23	0.43	0.34

Also, suppose that you model the decaying organic carbon by the following exponential function:

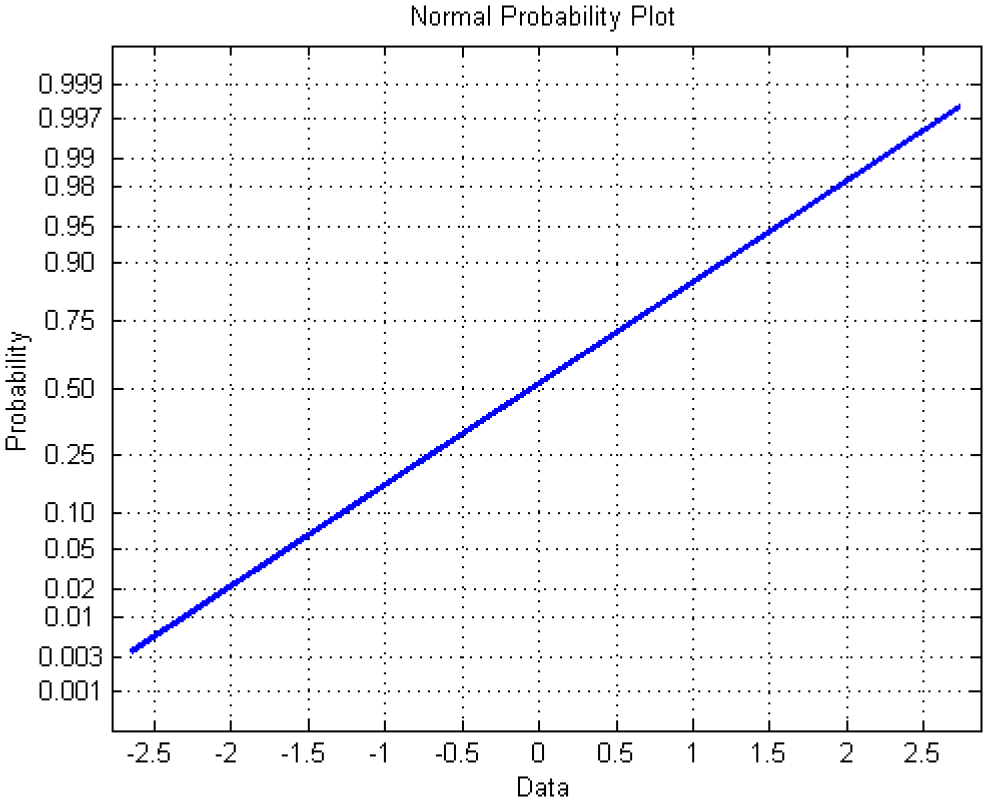
$$C(t) = C(0)e^{-rt}$$

This equation can be expressed as a linear regression model (with dependent variable $y(t) = log_e C(t)$ and unknown regression parameters $a_1 = log_e C(0)$ and $a_2 = -r$) if you take the log_e of each side and add a random measurement error to the result. You can use a regression approach to estimate the unknown regression parameters from the tabulated data and then to check the model's ability to explain observed temporal changes in concentration. The results of this regression are summarized in the ANOVA table provided below.

Source	SS	df	MS=SS/df	$\mathcal{F}$	p
Regression	1.09	1	1.09	7.15	0.056
Error	0.61	4	0.15		
Total	1.70	5			

Carry out the following tasks:

1) Define the arrays needed to carry out the regression analysis with the MATLAB function `regress` (documentation attached). Provide specific numerical values inferred from the data.
2) Use the $\mathcal{F}$ and p values in the table to determine whether the regression for this problem is significant (i.e. does the exponential model provide a good explanation of observed temporal variability).
3) Calculate the R^2 value for this regression. Does it suggest that the exponential fit is good or bad?


Normal Probability Plot
Probability
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
0.003
0.001
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
Data