Stat 998, Fall 2013 (Larget)
Diagnostic Data Analysis Assignment

Assignment purpose and motivation.— The purpose of this assignment is to give you practice with simple data analysis and report writing. Your performance should be viewed in a diagnostic manner, helping you to determine areas in which you may be weak and may require extra effort. The data in these problems are not entirely "real" and have been edited and manipulated into the form shown. Your emphasis in this assignment should be on "statistical problem solving" and report writing (as opposed to solving a real client's problem).

Turning in the reports.— This assignment contains three separate data analysis problems, each of which asks you to do an analysis and write a brief report to a fictional client. Parts of the assignment will be due on different days so that you may incorporate feedback from my grading of early problems on later reports.

Due Date	**Part**
Tuesday, September 10	Problem 1
Thursday, September 12	Problem 2
Tuesday, September 17	Problem 3

You should turn in a paper copy that I will comment on and return to you as well as sending an electronic version in PDF format to `larget@stat.wisc.edu`. Each report file should be named following this template: `yourFamilyName-diagnostic-X.pdf` where the X is one of 1, 2, or 3. For example, a student named John Dingleberry would send an e-mail to me with the files `dingleberry-diagnostic-1.pdf`, `dingleberry-diagnostic-2.pdf`, and `dingleberry-diagnostic-3.pdf` attached.

You may use any statistical package you wish to for the analysis and any software you desire to write the report. However, you must convert your report to PDF format to send to me.

Report format.— The intended audience for each report is the "client", here, a fictional scientist who provided the data and whose scientific questions you aim to address. The reports should be a mixture of narrative, equations, graphs, tables, and results, written in such a way that communicates clearly to the client. You need to make it clear exactly what analyses you performed, but you do not need to provide excessive detail. In particular, your report should not contain the specific R or SAS code you used for the analysis and for graphs. A brief motivation for your choice of method is important. Keep in mind the importance of addressing the assumptions underlying the methods that you use. For your equations, graphs, and other computer output — as well as the results in general — make sure that you always provide interpretation. The vast majority of your report should be text in which you explain the analysis and interpret the results. Each figure or graph that you include in the report should be accompanied by a written description that stands alone, is descriptive, and includes interpretation. You should not include every graph you examined as part of your analysis. You should include key graphs that address each question asked.

For each problem, the **maximum length of the report is six pages**, which includes a one page summary and up to five pages of text and figures. Longer reports receive no credit and will receive no feedback. Essential figures and tables should be fully integrated into the text. Of course your reports can be shorter than six pages. Please hand in a separate report for each problem.

Note that there will not necessarily be a single correct "answer" to each problem (there can be, however, many incorrect answers!). In general, you should try to use the simplest appropriate statistical approach that will answer the question.

Additional tips/guidelines: A good summary provides a brief description of the scientific problem, a brief explanation of the design and the way the data were generated, and a brief statement of the main results. The body of the report may be broken up into similar sections, although the "results" portion may be divided into a section on exploratory data analysis, a section that presents the models you use, and a section that presents your results, along with any diagnostic analyses you conducted, and interpretations of the results. I am providing these as a guideline only, ultimately you will generate your own style that captures all of these.

Honesty and ethics.— Each problem in this assignment has been used in this course in previous semesters. Your work for this assignment must be completely independent. In particular, you are *not allowed* to examine the reports of any other students in this or any previous semester. For this assignment, you may not discuss approaches to the problem withanyone else. I want to have a fair assessment of your ability to find and carry out an appropriate statistical analysis and to communicate the results in writing at the beginning of the semester.

Data.— Data sets for this assignment are small and included on paper. You can also find electronic versions of the data at the course web page.

Three diagnostic problems

1. An experiment is conducted to compare the effects of four different soil additives on the assimilation of a particular complex molecule containing phosphorous in the roots of corn plants. The amount of this molecule is determined by a laborious chemical analysis on a 5 milligram portion of ground-up root material. The results of the analysis are given as concentrations in parts per million (ppm).

The four soil additives are (1) a standard mixture of inorganic material including some phosphorous; (2) a new blend of inorganic material with no added phosphorous; (3) the same new blend as in (2) with 1% (by weight) phosphorous supplement; and (4) the same new blend as in (2) with a 2% (by weight) phosphorous supplement. (Note: The amount of phosphorous in (2) cannot be readily quantified).

Each of the treatments is applied to three randomly selected large pots that are otherwise identically prepared. Newly germinated corn plants are planted in the pots. The twelve pots are randomly located in an environmentally controlled growth chamber. The pots are all watered daily.

At the end of 15 days the roots from each plant are removed and ground up. Two 5 milligram portions from each plant are randomly taken from the ground-up root material and analysed for amount of the particular complex molecule containing phosphorous.

The data are presented in twelve rows as follow with each row representing one pot. Each row gives the concentration of the molecule in ppm for the two 5 mq portions of the root material.

	Molecule	
Additive	Concentrations	
(1)	1.9	2.1
(1)	2.4	2.8
(1)	1.4	1.6
(2)	2.0	1.8
(2)	1.2	1.2
(2)	1.9	1.6
(3)	2.9	3.0
(3)	3.7	3.2
(3)	2.2	2.2
(4)	5.1	4.5
(4)	3.3	3.0
(4)	3.0	3.5

Determine the effect of soil additive on the concentration of the particular complex molecule.

2. A brief study was undertaken to determine how well a small number of simple measures could predict senility in aging women. A random sample of 23 elderly women was selected from a registry of patients in a large city and four quantities were recorded for each women: (1) age (in years); (2) years of education; (3) score on a standardized test of cognitive function; and (d) a doctor's determination of senility with [senile = 1] and [not senile = 0].

The recorded data are listed below. Determine which measure or (measures) best predicts senility and interpret your findings.

	age	educ	score	senility
1	77	18	12	0
2	88	20	8	0
3	74	14	13	1
4	81	12	14	0
5	93	16	14	0
6	74	16	7	0
7	77	12	9	1
8	86	12	11	1
9	68	14	8	1
10	89	11	13	1
11	75	12	13	0
12	95	14	9	1
13	72	12	16	1
14	78	14	11	1
15	72	11	10	1
16	83	12	7	1
17	83	12	10	1
18	88	16	13	0
19	84	14	17	0
20	90	16	10	1
21	79	14	15	0
22	82	15	11	1
23	90	13	18	0

3. A group of bacterial pathogens is known to cause damage to soybeans. In mid-August, 25 different fields were scored for pathogen damage. Each field was assigned a number from 1 to 10 where 1 represents negligible damage and 10 represents severe damage. The scoring system is based on a visual examination from a small plane flying overhead. Each field had associated with it a weather station for obtaining climatological data. The objective of this problem is to find a useful model relating damage score to the factors of interest. These factors are described as follows:

Rainfall: Total precipitation in inches for the 30 days prior to date of scoring.
Wind: Average wind speed in miles per hour for the 30 days prior to date of scoring.
Temperature: Average high temperature (degrees Fahrenheit) for the 30 days prior to scoring date.
Crop History: Code for crop planted on each field during the previous growing season. Code 1 = soybeans, 2 = oats and 3 = snap beans.

Score	Rainfall	Wind	Temperature	Crop History
4	2.84	11.2	77.2	3
4	4.12	10.4	82.7	1
2	1.23	12.1	81.0	1
3	1.79	12.9	82.4	3
1	2.04	9.4	79.1	2
6	3.72	8.4	82.3	3
2	2.76	12.5	80.2	2
1	2.14	7.8	76.8	2
8	5.01	12.4	79.3	1
2	2.99	11.5	83.9	2
2	3.47	8.6	81.2	3
3	3.04	8.0	79.6	2
6	2.22	11.1	80.2	2
3	3.16	8.6	80.4	3
5	4.22	12.0	78.0	3
1	1.80	9.8	83.0	2
1	2.53	7.4	81.4	3
6	3.96	8.9	78.2	3
9	4.37	12.2	82.4	1
3	2.21	11.7	80.9	1
4	3.30	10.3	80.9	3
4	2.62	11.7	82.1	1
3	3.98	10.6	79.6	2
6	4.51	11.2	79.1	3
2	4.73	10.3	80.6	2