STAT 2050: Exploratory Data Analysis
Winter Semester 2007

Instructor: Dr. Ron Hilburn
Office: Chase 105
Telephone: 494-1981
E-mail: rhilburn@mathstat.dal.ca

Lecture: Tu., Th. 11:35-12:55 pm
Class Room: Chase 319
Office Hours: Tu.,Th. 3:00-4:30 pm
Website: WebCT

Course Text: Statistical Computing: An Introduction to Data Analysis using S-Plus by Michael J. Crawley

Course Description: Introduction to exploratory data analysis and graphical techniques using S-plus. Methods of data analysis and inference commonly used in the experimental sciences are reviewed. Students will be introduced to and make use of the S-Plus data analysis and graphics system especially designed for data analysis of an exploratory nature. Programming in the S language will be introduced with emphasis placed on statistical problem solving.

Method of Evaluation:	Assignments	25%
	Midterm exam	25%
	Final exam	50%

Grading Scheme:	A+	93 and above
	A	85 up to but not including 93
	A-	80 up to but not including 85
	B+	75 up to but not including 80
	B	70 up to but not including 75
	B-	65 up to but not including 70
	C+	60 up to but not including 65
	C	57 up to but not including 60
	C-	53 up to but not including 57
	D	50 up to but not including 53
	F	below 50

Note: There will be no supplemental examination.

Assignments: There will be eight weekly assignments that provide an opportunity to use the techniques discussed in class to analyze data sets. Assignments will require both hand computations, using a calculator, and computations made with the S-Plus data analysis and graphics software package.

S-Plus Software: S-Plus (Windows Version) is available on computers in the Chase 007 computer lab and in LSC 200. You will need to have a Novell account to access these machines.

Students with Disabilities: Students with permanent or temporary disabilities who would like to discuss classroom or exam accommodations are asked to come and see me as soon as possible. You can meet me after class or privately during office hours. For your information, the phone number for Student Accessibility Services is 494 2836. Please call them and register.

STAT 2050: Exploratory Data Analysis

Winter, 2007

Week	Assignment	Topics	Chapters
1		Introduction to S-Plus GUI	
2		Command Line Operation in S-Plus	Cha 2
3	1	Probability Calculations	Chas 5, 7, 26
4	2	Graphical/Tabular Data Displays	Chas 9, 10, 21
5	3	Tests of Significance	Cha 11
6	4	Tests of Significance Resampling and Bootstrap Methods	Cha 11 Cha 12
7		Midterm Exam Permutation and Randomization Tests	 Cha 12
8		Study Break	
9		Statistical Models in S-Plus Regression Analysis	Cha 13 Cha 14
10	5	Multiple Regression Model Critique	Cha 24 Cha 17
11	6	Analysis of Variance	Cha 15
12	7	Analysis of Covariance	Cha 16
13	8	Regression Approach to ANOVA	
14		Review	

DALHOUSIE UNIVERSITY FACULTY OF SCIENCE

Department of Mathematics and Statistics

STATISTICS 2050B(01) Final Examination

Date and Time: 9:00 AM - 12:00 PM, Monday 23 April, 2007

Print Name _____ **Student ID** _____

Signature _____

The entire exam is worth 100 points. It is closed book, but you may refer to two sheets (8.5 x 14 both sides) of notes. The exam consists of 7 problems, the number of points allocated to each part of a problem is shown in the left hand column. You may use a calculator and straight edge.

1. An experiment was conducted to investigate the effects of various cold water temperatures on the pulse rates (in beats/minute)of small children. Data for 10 children participating in this investigation are shown below.

Child	1	2	3	4	5	6	7	8	9	10
Water Temp.(^0F)	68	65	70	62	60	55	58	57	52	50
Pulse Rate Decrease	2	5	1	10	9	13	10	12	16	17

Consider Pulse Rate Decrease to be the response (Y) and Water Temperature to be the explanatory (X1) variable.

(3) (a) Use S-Plus commands to form the Y vector of response values and the standard regression X matrix of explanatory variable values needed to perform the regression of Y on X1 using matrix methods.

(3) (b) Show the series of S-Plus commands needed to compute values of the regression coefficients from the Y vector and X matrix developed above. Do not use the lm() command.

(3) (c) Show the series of S-Plus commands needed to compute the value of the error sum of squares, SSE. Do not use the lm() command.

(2) (d) From the above results, show the series of S-Plus commands needed estimate the value of the error variance, σ^2.

(2) (e) Show the series of S-Plus commands needed compute values for elements of the regression coefficients variance-covariance matrix.

(2) (f) Show the series of S-Plus commands needed to compute standard errors for both regression coefficients, $SE_{\hat{\beta}_i}$.

(2) (g) Show the series of S-Plus commands needed to compute a 90% confidence interval for the regression line slope, β_1.

(3) (h) Show the series of S-Plus commands needed to compute the regression sum of squares, SSR.

(2) (i) Show the series of S-Plus commands needed to compute the value of the total sum of squares, SST.

(3) (j) Show the series of S-Plus commands needed to compute the value of the coefficient of determination, R^2.

2. Provide the sequence of S-Plus commands needed to accomplish the following:

(3) (a) Show the series of S-Plus commands needed to plot a smooth curve of x versus $4(x^2-1)^2$ over the range of x values 0. to 2.5, scale the horizontal (x) axis to range from 0 to 2.5 and the vertical axis (y) to range from 0 to 25. Label the horizontal axis "Input" and the vertical axis "Response".

(2) (b) Show commands necessary to add the following $(x1, y1)$ data values to the above plot.

```
x1=c(0.0,0.25,0.5,0.75,1.0,1.3,1.5,1.75,1.8,2.0)
y1=c(3.8,3.0,4.0,0.5,0.1,0.7,8.0,14.0,19.0,23.8)
```

3. Consider the following lines of S-Plus code.

```
a1 = 2 ; t7 = 0
for(x in 3:7){
a1 = a1 + x
t7 = a1 + t7 - x
}
print(a1)
print(t7)
```

(4) (a) What values will variables "a1" and "t7" have when printed above?

(b) Using the sample() command, show the sequence of S-plus commands necessary to draw 10 random samples of size 5 from the data vector below. Sample with
(3) replacement.

```
> x3 = c(0.2, 3.1, 5.4, 2.6, 1.8, 3.4, 7.5, 2.8)
```

4. Using S-Plus commands other than stdev() and var(), create a function, called SDsamp,
(5) to compute the standard deviation of the following sample.

```
x = c(13.7, 14.1, 15.3, 14.2, 14.8, 14.4, 14.2)
```

Recall that:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

5. Three assembly machines are being considered by Sears Canada for use in their warehouse to assemble electronic components. Tests were conducted in which the length of time (in seconds) it takes a machine operator to assemble an electronic component using a particular machine was recorded.

Machine		
1	2	3
36.1	38.9	38.4
35.3	38.0	38.6
36.6	40.0	40.2
37.2	42.3	40.4

(3) (a) Show all commands necessary to perform nonparametric oneway ANOVA, using the Splus kruskal.test() command, to determine if mean assembly times are different for the three machines.

(3) (b) Now, using a sequence of Splus commands, compute the value of the Kruskal-Wallis test statistic. Do not use the kruskal.test() command.

(3) (c) Show the Splus commands needed to compute the pValue for the above test.

(2) (d) Perform standard Analysis of Variance to test the equality of mean assembly times for the 3 machines. Show the sequence of Splus commands necessary to perform ANOVA using the aov() command.

(3) (e) Show the sequence of S-Plus commands (don't use the aov() function) needed for computation of SST (the total sum of squares)

(3) (f) Show the sequence of S-Plus commands (don't use the aov() function) needed for computation of SSA (the treatment sum of squares)

(3) (g) Show the sequence of S-Plus commands (don't use the aov() function) needed for computation of SSE (the error sum of squares).

(2) (h) Using the above results, show the S-Plus commands needed to compute the factor F statistic.

(2) (i) Show the S-Plus commands needed to compute the pValue for the test.

6. A randomized block design was employed using 4 real golfers as blocks. Each golfer hit a random sample of 3 golf balls, one of each brand in a random sequence. The test compared the top 3 brands of golf balls for mean distance traveled when struck by a driver. Distance traveled in yards for each hit is presented in the table below.

Golfer	**Brand A**	**Brand B**	**Brand C**
1	202.4	203.2	223.7
2	242.0	248.7	259.8
3	220.4	227.3	240.0
4	230.0	243.1	247.7

(2) (a) Show all commands necessary to perform nonparametric randomized block ANOVA, using the Splus friedman.test() command, to determine if mean distances driven by the 3 brands of golfballs are different.

(3) (b) Now, using a sequence of Splus commands, compute the value of the Friedman test statistic. Do not use the friedman.test() command.

(2) (c) Show the Splus commands needed to compute the pValue for the above test.

(3) (d) Perform standard Analysis of Variance to test the equality of mean distances driven by the 3 brands of golfballs. Show the sequence of Splus commands necessary to perform ANOVA using the aov() command.

(3) (e) Show the sequence of S-Plus commands (don't use the aov() function) necessary for computation of SSA (treatment sum of squares).

(3) (f) Show the sequence of S-Plus commands (don't use the aov() function) necessary for computation of SSB (block sum of squares).

(2) (g) Show the S-Plus commands needed to compute the main factor F statistic.

(2) (h) Show the S-Plus commands needed to compute the pValue for the test.

(2) (i) Show the S-Plus commands needed to compute the block F statistic.

7. Use linear regression to perform the main factor F-test for the randomized block design outlined in problem 6.

(2) (a) Show the form of the regression model appropriate for solving this problem. This is the full model.

(3) (b) Make tables to show the correspondence between factor levels (golf ball brands), blocks (golfers), and indicator variable values.

(2) (c) Show the form of the reduced model (regression model) needed to evaluate the main factor.

(2) (d) Show the form of the reduced model (regression model) needed to evaluate the block effect.

(3) (e) Given the S-Plus commands shown below, provide all additional commands needed to produce the standard regression X matrix for the full model regression.

```
y = c(202.4,242.0,220.4,230.0,203.2,248.7,227.3
      ,243.1,223.7,259.8,240.0,247.7)
x0 = c(1,1,1,1,1,1,1,1,1,1,1,1)
x1 = c(0,0,0,0,1,1,1,1,0,0,0,0)
x2 = c(0,0,0,0,0,0,0,0,1,1,1,1)
```