STA 303 H1F / 1002 HF – Winter 2011 – Assignment 1

Due: Monday, February 14 at 14:00 a.m.
No late assignments will be accepted without a valid reason.
Presentation of solutions is important. It particular, it is inappropriate to hand in pages of SAS output without explanation or interpretation. The only SAS output you need to submit with assignments is relevant plots. Quote relevant numbers from your SAS output as part of your solutions. You don't need to hand in your SAS code.

The Data:
The source of these data is *Analyzing Medical Data Using S-PLUS* by Everitt and Rabe-Hesketh. The data for this assignment are in a text file on the course web site and at `www.utstat.utoronto.ca/alisong/Teaching/1011/Sta303/assignments.html`.

The data are from an investigation of girls suffering from anorexia. Patients were randomly assigned to receive one of three treatments: cognitive behavioural treatment, family therapy, or the standard therapy. The weight of each patient was recorded both before the treatment began, and then after a period of time receiving therapy. Therapies are considered successful if girls gain weight on the therapy.

The variables in the dataset are:
• `subject` – There are 72 girls in the study.
• `therapy` – The treatment the subject received; b = cognitive behavioural, f = family therapy, or c = control (the standard therapy).
• `before` – The patient's weight (in pounds) before treatment.
• `after` –The patient's weight (in pounds) after receiving treatment.
The column of 1's at the end of the data file can be ignored.

Use SAS to do the analysis for the following questions.

1. *ANALYSIS 1: One-way Analysis of Variance on the Change in Weight*

   (a) (1 mark) Use side-by-side boxplots to compare the change in weight among the three therapies? Do there appear to be differences?

   (b) (3 marks) Carry out an analysis of variance to examine whether there are differences in the means of the changes in weight among the three therapies. If there is evidence of differences, carry out an appropriate analysis to see which therapies differ. Can you rank the therapies in terms of effectiveness? Quote relevant numbers from the SAS output to support your answer.

   (c) (4 marks) Do the assumptions of the model appear to hold?

2. *ANALYSIS 2: Two-way Analysis of Variance on Weight*
   At the end of the assignment is some SAS code for reading in the data as two lines for each subject, one for the before (baseline) and one for the after (end) weight. If you're not sure what the code is doing, add `proc print;` to see the data set that is created.

   (a) (1 mark) From an interaction plot of the means, do there appear to be differences among the therapies? Do the therapies seem to be effective?

(b) (4 marks) Fit a linear model with weight as the response variable and predictor variables therapy and when (before versus after) the measurement was taken. What do you conclude about the therapies? Follow-up any significant results with an appropriate analysis to see where any differences in the means lie. Quote relevant numbers from the SAS output to support your answer.

(c) (4 marks) Do the assumptions of the model appear to hold?

3. (1 mark) Compare your conclusions for ANALYSIS 1 and ANALYSIS 2. Do they agree?

4. (2 marks) Which analysis do you prefer? Why? Consider both ease of interpretation and whether the model is valid.

Some potentially useful SAS code:

```
/* Calculate the changes in weight for ANALYSIS 1 */
/* Include this line in the data step when reading in the data file in its original form */
change=after-before;

/* Read in the data for ANALYSIS 2 */
data analysis2;
  infile 'a1data.txt';
  input subject therapy $ before @;
    weight=before;
    when='baseline';
    output;
  input after;
     weight=after;
     when='end     ';
     output;
  drop before after;
```

Something else to think about (that doesn't have to be handed in):
When there are differences at baseline among treatment groups (particularly a problem it situations when randomization is not used to assign subjects to treatments), doing analysis on change from baseline can be misleading because of the effect of "regression to the mean". In these situations, analysis of covariance is preferred. In STA 302/1001 you learned the analysis of covariance procedure for comparing regression lines for different groups. For these data, you could fit the model:

$$\texttt{after} = \beta_0 + \beta_1 \texttt{before} + \beta_2 I_{[therapy=b]} + \beta_3 I_{[therapy=f]} + e$$

How would you interpret the output from this model if fit to these data?