Biostat 510
Homework 5
Due Thursday, February 16, 2006

Before beginning this homework, submit a **libname** statement so you can use the permanent data sets created previously (b510.allgroups and b510.afifi). Remember, the libname statement points to a *folder*, not to a data file.

The first questions apply to the data set (b510.allgroups) we created from our class data.

1. Create a **Scatterplot** using the b510.allgroups data set, with PULSE1 as the X variable and PULSE2 as the Y variable.
   a) Include two linear regression lines in your scatterplot, one for those who ran and one for those who did not run.
   b) ***Make sure you exclude those who are missing on the variable RAN** in your scatterplot, by using a where statement, as shown below:

```
where ran not = .;
```

   c) Include the scatterplot with the regression lines in your homework.
2. Create centered versions of some variables and create interaction terms.
   a) Get descriptive statistics for PULSE1 for all cases in the data set.
   b) Create PULSE1_RAN, by multiplying PULSE1 by RAN.
   c) Create CENTPULSE1 by subtracting the overall mean of PULSE1 from each value (use 72 as the overall mean of PULSE1).
   d) Create CENTPULSE1_RAN by multiplying CENTPULSE1 by RAN.

   Your SAS commands to create the new temporary data set will look something like those shown below:

```
data new;
  set b510.allgroups;

  pulse1_ran =                              ;
  centpulse1  =                             ;
  centpulse1_ran =                          ;
run;
```

3. **Run an ANCOVA model** with an interaction term on the **uncentered variables.**
   a) Run an ANCOVA model using Proc Reg, with PULSE2 as the dependent variable and PULSE1, RAN and PULSE1_RAN as the predictors.
   b) Include the output from this ANCOVA model in your homework.
4. **Rerun the ANCOVA model** with the **centered variables.**

a) Rerun the ANCOVA model using Proc Reg, with PULSE2 as the dependent variable and CENTPULSE1, RAN, and CENTPULSE1_RAN as the predictors.
b) Include the output from this ANCOVA model in your homework.

5. **Run a separate regression** for those who **ran and those who did not run** to predict PULSE2 as a function of PULSE1.
   a) First, sort the data set by RAN.
   b) Run the regression model by RAN.
   c) Use the centered version of PULSE1 as your predictor.
   d) Include the Parameter Estimates output for these regressions in your output (you do not need to include the ANOVA portion of the output for these regressions).
   e) Note: you will probably want to exclude RAN=. from your analysis.
6. **Run an ANCOVA model excluding the interaction** term. Run this model using the centered version of PULSE1.
   a) The dependent variable in this model should be PULSE2. The predictors should be CENTPULSE1, and RAN.
   b) Include the output from this regression model in your homework.

The last 4 questions apply to the data set b510.afifi.

7. **Run a stepwise regression model.**
   a) Run a stepwise regression to predict SBP2 using SBP1, DBP1, BSA1, CI1, HGB1, and MAP1 as predictors.
   b) You do not need to use the **details** option for this model.
   c) Include the final model and the summary of the stepwise selection in your homework.
8. **Run a backward selection regression model.**
   a) Run a backward selection regression to predict SBP2 using SBP1, DBP1, BSA1, CI1, HGB1, and MAP1 as predictors.
   b) You do not need to use the details option for this model.
   c) Include the final model and the summary of backward selection in your homework.
9. **Run a regression model using all possible regressions** and **adjusted R^2 selection methods.**
   a) Request that SAS prints the best 10 models for each of these criteria.
   b) Include the output from the all possible regressions and the adjusted R^2 selection methods in your homework.
10. **Answer the following questions** about your analysis.
    a) (Scatterplot for those who ran and did not run)
       i. Does there appear to be a linear relationship between PULSE1 and PULSE2 for those who ran? For those who did not run?
       ii. Do the slopes of these regression lines for those who ran and those who did not run appear to be the same or different? Which group (those who ran or those who did not run) has the steeper slope?

iii. Can you tell if there is a significant difference in the slopes by the graph alone?

b) (ANCOVA model with uncentered variables as predictors)
   i. What are the sample size and the Model R^2 for this model?
   ii. What is the overall significance of this model? Use the F-test in the ANOVA output to answer this question. Write out the null and alternative hypotheses for this test, and report the F-test statistic, the numerator and denominator degrees of freedom, and the p-value for the test. What do you conclude?
   iii. Interpret each coefficient in this model, including the intercept.
      a. What is the estimated slope for those who ran?
      b. What is the estimated slope for those who did not run?
   iv. Is there a significant interaction between PULSE1 and RAN?
      a. Report the test statistic, the degrees of freedom and the p-value for this interaction test.
      b. Does this result surprise you, based on what you saw in the scatterplot earlier?
      c. Comment on why you think the interaction is or is not significant.

c) (ANCOVA model with centered variables as predictors)
   i. What are the sample size, Model R^2 and overall significance for this model? How do they compare to the values for the previous model?
   ii. Interpret each coefficient in this model, including the intercept.
   iii. What is the slope for those who ran? For those who did not run?
   iv. Is there a significant interaction in this model? Compare the test for the interaction in this model to the test for the interaction in the previous model.

d) (Separate Regressions for those who ran and did not run)
   i. How many observations are included in each regression?
   ii. What are the intercept and slope for those who ran and those who did not run?
   iii. How do these values compare to the values you calculated in the previous ANCOVA model?

e) (ANCOVA model without interaction)
   i. What are the sample size and Model R^2 for this model? What is the overall test of significance for this model?

f) (Stepwise regression model)
   i. What are the variables included in the final model when using stepwise selection?
   ii. What is the model R^2 for the final model?

g) (Backward selection model)
   i. What are the variables included in the final model when using backward selection?
   ii. What is the model R^2 for the final model?