BIOS663 Homework 3
Due noon on Tuesday, March 5 to my mailbox.

1. A group of subjects was recruited to a weight loss study in a medical center. The data consist of their weights (y = WGHT), average daily exercise times (x1 = TIME) and average daily running mileages (x2 = RUN). One of the objectives in this study is to investigate the effect of x1 and x2 on weight loss.

   (a) A partial ANOVA table for estimating WGHT from TIME is given below. Complete the table.

```
Dependent Variable: WGHT

                           Sum of
Source          DF        Squares     Mean Square    F Value     Pr > F

Model           1      2624.670184

Error           96     1827.099916

Corrected Total 97     4451.7701
```

   (b) State the model assumptions based on which the ANOVA table was computed.

   (c) Is average daily exercise time a significant predictor for predicting weight loss? State your answers in terms of the model from the previous questions and the statistical test of hypothesis.

   (d) To further explore the unexplained variation in the data, and in order to improve the predictive power of the model, the average daily running mileage (RUN) is also considered. The result is summarized in the following table. Does the analysis suggest that neither variable is significant? Why and why not?

```
Source    DF  Type III SS   Mean Square   F Value  Pr > F

TIME      1    88.173        88.173         3.10    0.08
RUN       1    70.339        70.339         2.47    0.12
```

2. A data set was collected by the Environmental Protection Agency (EPA) at the Health Effects Research Laboratory at UNC: Chapel Hill. One hundred seventy-two young adult males received a battery of pulmonary function tests. (The data are described in more detail in Muller and Fetterman on page 536).

For this homework, fit a model with average forced vital capacity (FVC) (in ml) as the outcome and height, weight, body mass index (BMI=$\frac{weight(kg)}{(height(m))^2}$), age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity as predictors. For the purpose of added-in-order tests, assume that this order is the preferred order for testing. The data are available on the course website in FILEN.DAT with associated SAS file hw3.SAS.

To report a test, provide H_0, the test statistic, the degrees of freedom, the p-value, the decision (accept/reject H_0), and an interpretation of the result in terms of the subject matter.

(a) Use Proc GLM to produce a table like Table 4.8.1 in Muller and Fetterman (pg56) (details about the table can be found in Lecture7.pdf, pg29) with the following predictors: height, weight and age. The table should contain six df values, six SS values, four MS values, three F values, and three p-values.

(b) Report the test of whether the group of predictors (height, weight, body mass index, age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity) is important.

(c) Report the corrected R^2 for these data.

(d) Give the two models being compared in testing the following hypotheses for these data, and report each test.

i. H_1: The entire group of predictors provides no useful information about FVC.

ii. H_2: Height provides no information about FVC, not adjusting for effects of other predictors (i.e., in a simple regression model).

iii. H_3: Adjusting for weight and BMI, height does not provide any additional information about FVC.

iv. H_4: After adjusting for weight, BMI, age, elevation, speed, temperature, barometric pressure, and humidity, height does not provide any additional information about FVC.

v. H_5: The group of body size variables (height, weight, BMI) provides no additional information about FVC compared to a model for only the mean level of FVC.

vi. H_6: The group of body size variables (height, weight, BMI) provides no additional information about FVC after adjusting for age, elevation, speed, temperature, barometric pressure, and humidity.

(e) Report a test of the hypothesis that humidity has no affect on FVC after adjusting for all the other variables in the model.

(f) Describe the relationship between the body size variables and FVC in these data.

(g) Based on the original model, which characteristics are associated with the best (largest) FVC?

3. For the same data in Q2, consider the following model

$$\begin{aligned} FVC_i &= \beta_0 + \beta_1 HEIGHT_i + \beta_2 WEIGHT_i + \beta_3 BMI_i + \beta_4 AREA_i \\ &\quad + \beta_5 AGE_i + \beta_6 AVTREL_i + \beta_7 AVTRSP_i + \beta_8 AVTREL_i AVTRSP_i \\ &\quad + \beta_9 TEMP_i + \beta_{10} BARM_i + \beta_{11} HUM_i + \varepsilon_i, \end{aligned}$$

$i = 1, \ldots, n$.

(a) Compute the following correlations, giving the interpretation of each, between FVC and age, and report tests of the hypotheses that each correlation equals zero.

i. the correlation between age and FVC, controlling both for all the other variables in the model

ii. the correlation between age and FVC, controlling only age for all the other variables in the model

iii. the simple correlation between age and FVC (not controlling for any other variables)

(b) Provide and interpret the following diagnostics (include subject ID when appropriate) for the regression model.

i. Largest 5 studentized residuals (in absolute value)

ii. Results of a test of the Gaussian distribution for the studentized residuals

iii. Histogram of the studentized residuals

iv. Plot of studentized residuals versus predicted values