SAS – Homework.

On your CD you will find a comma separated variable file called MBAHomework.csv.

This file contains 21 observations from a study undertaken by a large city bank of average account size (ACCTSIZE) in each of its branches to per capital income (INCOME) in the corresponding zip code area, number of business accounts (BUSIN), and number of competitive bank branches (COMPET).

Your task, should you chose to accept it, is:

1. Write a SAS program to read the data from the comma-separated file into a SAS Dataset.

   Cut and paste your program into this document here.

2. Write a SAS program to print the data within SAS.

   Cut and paste your program into this document here.

3. Write a SAS program to perform Regression Analysis of the relationship between average account size and the other variables. Use the SAS/STAT Procedure “Reg” and a forward selection method.

   (I have cut and pasted relevant parts of the SAS Online Documentation into the appendix to this document.)

   Cut and paste your program into this document here.

4. Look at the results in the results window and expand them using the + tabs until you can see the results of Step 3. From the Parameter Estimates page, type the regression equation parameters into the following equation.

   ```
   ACCTSIZE=          +/-          INCOME +/-          BUSINESS +/-          COMPET
   ```

5. Write a SAS program to provide Correlation and Covariance analysis of the data.

   Cut and paste your program into this document here.

6. Print out the pages of this document up to and including this line and hand it in.

The following is cut out of a couple of the pages of Online Documentation.

Example 55.1: Aerobic Fitness Prediction

The example assumes a SAS Dataset in the WORK library, called “fitness”.

Aerobic fitness (measured by the ability to consume oxygen) is fit to some simple exercise tests. The goal is to develop an equation to predict fitness based on the exercise tests rather than on expensive and cumbersome oxygen consumption measurements.

Three model-selection methods are used: forward selection, backward selection, and MAXR selection.

```
proc reg data=fitness;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
         / selection=forward;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
         / selection=backward;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
         / selection=maxr;
run;
```

Forward Selection (FORWARD)

The forward-selection technique begins with no variables in the model. For each of the independent variables, the FORWARD method calculates *F* statistics that reflect the variable's contribution to the model if it is included. The *p*-values for these *F* statistics are compared to the SLENTRY= value that is specified in the MODEL statement (or to 0.50 if the SLENTRY= option is omitted). If no *F* statistic has a significance level greater than the SLENTRY= value, the FORWARD selection stops. Otherwise, the FORWARD method adds the variable that has the largest *F* statistic to the model. The FORWARD method then calculates *F* statistics again for the variables still remaining outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant *F* statistic. Once a variable is in the model, it stays.

Backward Elimination (BACKWARD)

The backward elimination technique begins by calculating *F* statistics for a model, including all of the independent variables. Then the variables are deleted from the model one by one until all the variables remaining in the model produce *F* statistics significant at the SLSTAY= level specified in the MODEL statement (or at the 0.10 level if the SLSTAY= option is omitted). At each step, the variable showing the smallest contribution to the model is deleted.

Maximum R^2 Improvement (MAXR)

The maximum R^2 improvement technique does not settle on a single model. Instead, it tries to find the "best" one-variable model, the "best" two-variable model, and so forth, although it is not guaranteed to find the model with the largest R^2 for each size.

The MAXR method begins by finding the one-variable model producing the highest R^2. Then another variable, the one that yields the greatest increase in R^2, is added. Once the two-variable model is obtained, each of the variables in the model is compared to each variable not in the model. For each comparison, the MAXR method determines if removing one variable and replacing it with the other variable increases R^2. After comparing all possible switches, the MAXR method makes the switch that produces the largest increase in R^2. Comparisons begin again, and the process continues until the MAXR method finds that no switch could increase R^2. Thus, the two-variable model achieved is considered the "best" two-variable model the technique can find. Another variable is then added to the model, and the comparing-and-switching process is repeated to find the "best" three-variable model, and so forth.

The difference between the STEPWISE method and the MAXR method is that all switches are evaluated before any switch is made in the MAXR method . In the STEPWISE method, the "worst" variable may be removed without considering what adding the "best" remaining variable might accomplish. The MAXR method may require much more computer time than the STEPWISE method.