

# Stat 480, Project 1, Spring 2003

## Using Regression to Predict the Weight of Rocks

Project 1 is a variation on the data collection and analysis activity described in Chapter 5 of *The Practice of Statistics: Putting the Pieces Together* by John Spurrier. The object of the activity is to collect, clean, analyze a set of data and then write a report on what you have found.

### 1 Part I: Data Collection, Thursday, March 6, 2003

- The class will be split into two groups of 5 and 6 members, respectively. Each group will turn in a single written report at the end of the project.
- Each group will use a digital (gram) scale and a caliper (100<sup>th</sup>'s of an inch). Each group will be given 10 rocks to measure and weigh. Once those 10 rocks are measured and weighed, the groups will swap rocks. At the end each group will measure all 20 rocks and we will have two version of the data set.
- Before measuring the rocks we must have operational definitions for what constitutes the Length, Width and Height of a rock.
- Once we have our operational definitions. Spend this class period measuring the rocks and entering the data into a JMP worksheet. At the end of class email me the JMP file that contains your group's data. My email address is [wrstephe@iastate.edu](mailto:wrstephe@iastate.edu)
- On Tuesday, March 11, 2003 we will meet in Snedecor 321 to work on cleaning the data.

### 2 Part II: Data Cleaning, Tuesday, March 11, 2003

- You will work in your two groups.
- The data you collected last time is in a single JMP file that is available from the course webpage. If you use the Internet Explorer Browser you can simply double click on the JMP file and JMP should open up the file. Note that there are two sets of measurements, one for Group A and a second for Group B.
- Investigate and resolve the large differences in the two sets of measurements.
  1. Let DiffWeight be Weight for Group A minus Weight for Group B. Large differences are those greater than 2 grams.
  2. Let DiffLength, DiffWidth, and DiffHeight be the corresponding Group A measurement minus the Group B measurement. Large differences are those greater than 0.1 inches.

3. Use the Table→Split command to get the group measurements split into a total of 8 columns in JMP. Then, create the Diff-variables described above. Using Graph→Overlay Plot with each Diff-variable specified as a Y-variable, identify each large difference (record the two measurements and the rock number for each large difference). Use the Overlay Plot options: uncheck Overly Plot option, check Separate Axes options. Also, right-click on the numerical values running up the vertical scale in each of the four plots and add Reference Lines at  $\pm 2$  for DiffWeight and  $\pm 0.1$  for the other Diff-variables. This makes it easy to visually identify large differences. Moving the cursor near a point will display the rock number of the point. Hold the Shift key down and click on any points that show large differences, these points will also be highlighted in the worksheet. Print out the four plots.
4. Go back to the rocks and remeasure/reweigh as needed to resolve discrepancies in the data. Remember, do this only for rocks that show large differences as defined. Create a new “correct” data set named **project1.clean.jmp** and email it to me. Also, email a table of the discrepancies containing the two measured values (indicate the variable measured), the rock number, and the “correct” value as found in **project1.clean.jmp**.

### 3 Part III: Data Analysis - Tuesday, March 25, 2003

- First investigate the distribution of each variable; Weight, Length, Width, Height and Volume=Length\*Width\*Height. Are there any outliers? Turn in JMP output.
- Next investigate several models.
  - Simple linear regression of Weight on Volume=Length\*Width\*Height.
  - Simple linear regression of Weight on Width.
  - Simple linear regression of the logarithm of Weight,  $\log(\text{Weight})$  on the logarithm of Width,  $\log(\text{Width})$ .

For each model be sure to include an analysis of the residuals from each model. A residual is defined as the Weight minus the Predicted Weight. Turn in JMP output.

- Pick one of the models as providing the “best fit” to the data and support your choice by referring to the analysis.
- For the model you selected as “best,” obtain output from JMP, Minitab and Excel. Write a brief description of any differences between the 3 sets of output. Also, make a brief statement about which you prefer to use to analyze data like these and why you have this preference. For this part, turn in hard copy of each application’s output along with the description of indicated differences in output and your usage preference statement.

**Final written report is due by 5 pm Friday, March 28, 2003.**