****

EXST 7014, Lab 1: Review of SAS Programming Basics and Simple Linear Regression

****

OBJECTIVES

1. Prepare a scatter plot of the dependent variable on the independent variable
2. Do a simple linear regression in PROC REG
3. Get confidence intervals on the regression coefficients

Simple linear regression (SLR) is a common analysis procedure, used to describe the significant relationship a researcher presumes to exist between two variables: the dependent (or response) variable, and the independent (or explanatory) variable. This lab will familiarize you with how to perform SLR using the PROC REG procedure.

We will first look at the data graphically using a scatter plot to assess the nature of the relationship between the variables. Based on this assessment, we will use SLR to fit a straight line model relating two variables. The line will be fir using least-squares.

LABORATORY INSTRUCTIONS

Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78 ps=53;
Title1 'EXST7014 lab 1, Name, Section#';
```

Type your name and section number that will appear on the top of the output pages so that we know who performed the analysis.
Alternatively, you can use footnote as the following:

```
%let NAME=;
Footnote ‘EXST 7014 Lab1 Section#’;
Footnote2 ‘&NAME’;
```

Dataset

The data is from your textbook, chapter 7, problem 6and you can attain it through the link: http://www.stat.lsu.edu/exstweb/statlab/datasets/fwdata97/FW07P06.txt. The latitude (LAT) and the mean monthly range (RANGE), which is the difference between mean monthly maximum and minimum temperatures, are given for a selected set of US cities. The following program performs a SLR using RANGE as the dependent variable and LAT as the independent variable.

Copy the entire contents of the dataset (FW07P06) and paste into the SAS Editor. We will learn how to import external data in next lab. Be sure to type a semicolon after the last line in the dataset. It is also necessary to PROC PRINT the data to see whether the data is successfully obtained.

```
Data fw07p06;
Title2 'Latitudes and Temperature Ranges';
input CITY $ STATE $   LAT    RANGE;
cards;
 Montgome    AL      32.3     18.6
 Tuscon      AZ      32.1     19.7
 Bishop      CA      37.4     21.9
.
.
.
;
Proc print data=fw07p06;
Run;
```

Creating a Scatter Plot

When performing a regression analysis, it is always advisable to look at scatter plots of the data in order to get an idea of the type of relationship that exists between the response variable and the explanatory variables.

```
Proc plot data=fw07p06;
Title2 'Scatter plot of Temperature versus Latitude';
Plot RANGE*LAT;
Run;
```

The PROC PLOT statement above will create a scatter plot of RANGE vs. LAT. The graph is character based, so it is not fancy, but is sufficient for getting an idea of how RANGE and LAT are related.

To create more professional graphics, you will want to use procedures in SAS GRAPGICS package. You may refer to the various statements and options in the online SAS documentation if needed.

Fitting the Least-Squares Regression line Using SAS

Based on the scatter plot produced above, we will assume that an appropriate regression model relating RANGE and LAT is the liner model given by

$$y = \beta_0 + \beta_1\chi + \varepsilon$$

where Y is the RANGE, X is the LAT, and ε is a random error term that is normally distributed with mean 0 and unknown variances σ^2. β_0 is the estimate of Y-intercept, and β_1 is the estimate of the slope coefficient.

SAS has several procedures that will do this for us. In this lab we will use PROC REG.

```
Proc reg data=fw07p06;
Title3 'Simple Linear Regression between Temperature and latitude';
Model RANGE=LAT / CLB;
Run;
```

The output of PROC REG is very detailed and provides more information than we will be using in this lab. For this lab, we want to focus on the table of parameter estimates, and the coefficient of determination, denoted by R^2, that is the measure of how well the Least-Squares line fits the data. In particular, R^2 gives the proportion of the variability in the dependent variable that is accounted by the Least-Squares line. The coefficient of determination is labeled r-square in the output of PROC REG. The option CLB in the MODEL statement requests the 100(1-α)% upper and lower confidence limits for the parameter estimates. By default, the 95% limits are computed.

LAB ASSIGNMENT

1. Produce a scatter plot to show the relationship between RANGE and LAT. What is your observation?

2. Use PROC REG to fit the linear model: $y = \beta_0 + \beta_1\chi + \varepsilon$. Explain briefly your findings, which should include the parameter estimates, the interpretation of the parameters and appropriate hypothesis test.

3. Write the estimated regression function.

4. Fine the confidence interval for the intercept and the slope coefficients.

5. What proportion of the variability in the dependent variable RANGE is accounted for by LAT through the regression line?

*Remember to attach your SAS log for the lab report.