EXST 7015

Fall 2014

Lab1: Simple Linear Regression Analysis

Simple linear regression (SLR) is a common analysis procedure, used to describe potential relationship between two 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 help assess the nature of the relationship between the variables. Based on this assessment, we will use SLR to fit a straight line model for two variables. The line will be fit using least-squares.

Objectives

1. Prepare a scatter plot of the dependent variable (Y_i) on the independent variable (X_i)
2. Print the data set
3. Do a regression analysis in **PROC REG** including:
    - A scatter plot of data
    - Confidence intervals on the regression coefficients (b_0 and b_1)
    - Output of the predicted values, residuals and standardized residuals
4. Create a residual plot using the standardized residuals (RSTUDENT)
5. Use **PROC UNIVARIATE** to test the residuals for normality

Housekeeping Statements

The following statements include some useful “housekeeping ” functions that may facilitate the execution and interpretation of your results.

```
dm 'log;clear;output;clear';
ods html close;
ods html style=minimal body='EXST7010 Assign01.html';
ods graphics off;
options nodate nocenter pageno=1 ls=111 ps=53;
options FORMCHAR="|----|+|---+=|-/\<>*";;
ODS listing;
```

The following statements will serve to identify any log printouts as your own.

```
*********************************************;
*** James P Geaghan                         ***;
*** EXST 7015 Regression Example            ***;
*** Assignment 01                           ***;
*********************************************;
```

The following statements will be printed at the top of each page and serve to identify the owner of the results.

```
title1 "James P Geaghan";
title2 "Assignment 1 - Regression example";
```

Dataset

The data is from your textbook, chapter 7, problem 6 and can be accessed by 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. The code provided is minimal. Feel free to embellish the program and make it your own.

Copy the entire contents of the dataset (FW07P06) and paste into the SAS Editor. We will learn how to import external data in a future lab. Be sure to type a semicolon after the last line in the dataset. It is expedient to print the data to insure that it was successfully obtained.

```
data fw07p06;
   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=data fw07p06;
   title3 'LAT and TEMP ranges for selected US cities';
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;
   Title3 'Scatter plot of Temperature versus Latitude';
   plot range*lat;
run;
```

The PROC PLOT statement above will create a scatter plot of RANGE on 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 GRAPHICS package. You may refer to the various statements and options in the online SAS documentation if needed. The SAS ODS option "graphics on" also produced some interesting graphics but takes longer to execute and may cause issues in execution and printing the program.

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 linear model given by

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

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 the Y-intercept, and β_1 is the estimate of the slope.

SAS has several procedures that would fit this model to the data. In this lab we will use PROC REG.

```
Proc reg data=fw07p06;
   title3 'SLR between Temperature and latitude';
   Model range = lat / CLB;
  output out=next01 r=resid p=predicted rstudent=rstudent;
run;
```

The output of PROC REG 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.

```
Proc plot data=next01;
   Title3 'Scatter plot of Temperature versus Latitude';
   plot rstudent * lat = state / vref = -3 -2 2 3;
options ls=111 ps=52; run; options ls=111 ps=512;

PROC UNIVARIATE data=next01 normal plots;
  TITLE3 'Test of residuals from REG';
  VAR resid;
  ods exclude BasicMeasures ExtremeObs ExtremeValues Modes
       MissingValues Quantiles TestsForLocation;
RUN;

ods html close;
```

Previously stated objectives

1. Prepare a scatter plot of the dependent variable (Y_i) on the independent variable (X_i)
2. Print the data set
3. Do a regression analysis in **PROC REG** including:
    - A scatter plot of data
    - Confidence intervals on the regression coefficients (b_0 and b_1)
    - Output of the predicted values, residuals and standardized residuals
4. Create a residual plot using the standardized residuals (RSTUDENT)
5. Use **PROC UNIVARIATE** to test the residuals for normality

LAB ASSIGNMENT

1. Run the program and print the data set. **(1 point)**

2. Make a scatter plot to show the relationship between RANGE and LAT. How does it look? Does there appear to be a good linear association? **(1 point)**

3. Use PROC REG to fit the linear model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Explain briefly your findings. In particular; what are the parameter estimates, do the estimated differ significantly from zero (give a P value) and do the confidence intervals include zero? **(3 point)**

4. Write the estimated regression function. **(1 point)**

5. What proportion of the variability in the dependent variable RANGE is accounted for by LAT through the regression line? **(1 point)**

6. Does the residual plot appear to represent just a random scattering of points? If there are outliers do they appear to be explicable? **(2 point)**

7. Does the residual analysis indicate that the assumption of normality has been met? **(1 point)**