****

Lab 2: Simple Linear Regression: Regression Diagnostics and Assumption Tests

****

OBJECTIVES

Simple linear regression (SLR) is a common analysis procedure, used to describe the significant relationship between two variables: the dependent (or response) variable, and the independent (or explanatory) variable. In lab 1, SLR was performed to fit a straight line model relating two variables. We learned how to interpret parameter estimates and R^2, and understood the hypothesis test of SLR.

You might notice that a single observation that is substantially different from all other observations can make a large difference in the results of your regression analysis. If a single observation (or small group of observations) substantially changes your results, you would want to know about this and investigate further. In this lab exercise, we will use appropriate regression diagnostics to detect outliers (or unusual observations) besides evaluation of assumption.

In this lab exercise, you will get familiar with and understand as listed:

1) Use appropriate regression diagnostics to detect outliers (or unusual observations)
2) Evaluate the assumptions of SLR using Residual plot and Normality test

LABORATORY INSTRUCTIONS

Part I.
Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78 ps=53;
title1 'EXST7014 lab 2, Name, Section#';
ods rtf  file = 'c:/temp/lab2.rtf';
ods html file = 'c:/temp/lab2.html';
```

ods rtf file ='<path your output file>';
The output delivery system (ODS) improves the appearance of your output. The format of the output file can be html, rtf(rich text format which can be opened in MicroSoft Word) or PDF. It is necessary to add '**ods rtf close**' at the end of the program

Data set

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.

```
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;

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

Part II.

Fitting Simple Linear Regression

MODEL statement with options of INFLUENCE

OUTPUT statement (Predicted values of Y, Residual, rstudent, dffits, etc.)

Based on the scatter plot produced above, we 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.

```
Proc reg data=fw07p06;
Title2 'Simple Linear Regression between RANGE and LAT';
Model RANGE=LAT/ clb cli clm influence;
OUTPUT out=outdata p=Predicted r=resid cookd=cooksd dffits=diffits H=hat
     student=student rstudent=rstudent lclm=lclm uclm=uclm lcl=ccl ucl=ucl;
Run;
```

```
Proc print data=outdata;
Title2 'Listing of Observation Diagnostics';
Var RANGE predicted resid student rstudent;
Run;
```

Model dependent=independent / influence clb cli clm;

The option "**influence**" to display the following diagnostic statistics: standardized residual (RSTUDENT), hat diagonal values (Hat Diag H), and influence diagnostics (DFFITS and DFBETAS). These statistics are usually used to detect possible outliers. The **clb**, **cli** and **clm** options provide 95% confidence intervals for betas, prediction, and mean. The confidence level can be changed by the option **alpha = the desired level.**

Output out=Name of SAS data set < Varible1 …VaribleN > ;

The OUTPUT statement creates a new SAS data set containing diagnostic measures calculated after fitting the model. You can request a variety of diagnostic measures that are calculated for each observation in the data set. The new data set contains the variables specified in the MODEL statement in addition to the requested variables. If no variables are specified, the data set contains only the predicted values.

Part III.

Evaluate Assumption by Residual Analysis

```
Proc plot data=outdata;
Title3 'Residual plot';
Plot resid*predicted;
Run;

Proc univariate  data=outdata normal plot;
Title3 'Residual Analysis';
Var Resid;
Run;
```

Residual plot can be used to detect various problems such as non-linear pattern, non-homogeneous variances and outliers. If the data is of homogeneity, most of residual points of data randomly scatter around zero. If problems such as curvature or non-homogenous variance are detected in residual plot, we may need to consider fitting more complicated model.

The **UNIVARIATE** procedure on the residual is used to test normality. **Shapiro-Wilk Test** is a popular statistics to evaluate whether the data is normally distributed. It should be noticed that the null hypothesis test of Shapiro-Wilk is that the data is normally distributed. If P-value of this test is less than the significant level of 0.05, the null hypothesis is rejected and we conclude the data is not normally distributed. Otherwise, the null hypotheses could not be rejected and we conclude that the data is normally distributed.

LAB ASSIGNMENT

Your assignment is to perform necessary analysis using SAS and answer the following questions.

1. Use **Residual plot** to check the assumption of homogeneity of variance. Does the data set appear to be homogenous? Why?

2. Use **Proc univariate** to analyze RANGE and LAT. Does RANGE appear to be normally distributed? Why? Is this test relevant to the normality assumption? Why?

3. Use **Proc reg** to fit the model. Write down the regression equation and answer: Does the model fit data well? Why? (Hints: you may need to check the ANOVA table, predicted values of parameters, R-square, residual plot).

4. What is the predicted value of RANGE at LAT=42?

5. Does there appear to be any possible outlier(s)? State the name and value of the statistics that you use to reach your conclusion.