EXST 7014 - Statistical Inference II, Spring 2020

Lab 3: Simple Linear Regression: Regression Diagnostics

Due date: Feb 5th, 2020

OBJECTIVES

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. In previous labs, SLR was performed to fit a straight line model relating two variables. We learned how to interpret parameter estimates and R^2, and how to test specific hypothesis of SLR such as slope test and joint tests. We are also getting familiar with how to evaluate the assumptions of SLR by using residual and normality test.

You might realize 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. It is, however, very important to emphasize that simply discarding observations that appears to be outliers is not good statistical practice.

LABORATORY INSTRUCTIONS

Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78 ps=53;
title1 'EXST7015 lab 2, Name, Section#';
ods rtf  file = ‘c:/temp/lab2.rtf’;
ods html file = ‘c:/temp/lab2.html’;
```

Data set

The dataset is from Chapter 8, Problem 10 in your textbook. We are trying to estimate the survival of liver transplant patients using information on the patients collected before the operation. The variables are:
CLOT: a measure of the clotting potential of the patient’s blood;
PROG: a subjective index of the patient’s prospect of recovery;
ENZ: a measure of a protein present in the body;
LIV: a measure relating to white blood cell count and the response;
TIME: a measure of the survival time of the patient.
In this lab we will use TIME as the dependent and ENZ as the independent variable. The data is available at

http://statweb.lsu.edu/EXSTWeb/StatLab/DataSets/EXST7015/FW&M%20Data%202010/TEXT/DATATAB_8_31.TXT

```
data survival;
title2 'Survival of liver transplant patient';
input obs clot prog enz liv time;
cards;
1 3.7 51 41 1.55 34
2 8.7 45 23 2.52 58
3 6.7 51 43 1.86 65
4 6.7 26 68 2.1 70
.
;
Proc Print data= survival;
Run;
```

```
Proc plot data=survival;
title2 'Scatter plot of ENZ versus TIME';
plot time*enz;
run;
```

Fitting Simple Linear Regression

MODEL statement with options of INFLUENCE

OUTPUT statement (Predicted values of Y, Residual)

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

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where Y is the TIME, X is the ENZ, 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=survival;
title2 'Simple Linear Regression between TIME and ENZ';
Model time=enz/p 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 TIME predicted resid student rstudent;
run;
```

Model dependent=independent / influence clb cli clm;

The option "**influence**" to display the following diagnostic statistics: standard residuals (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.

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) Is the normality assumption violated? State the name and the value of the statistic that you use to reach your conclusion.

2) Does there appear to be any possible influential observations? State the name and the value of the statistics that you use to reach your conclusion.

3) What is the confidence interval for the mean RANGE for all cities with latitude 32.3?

4) What is the confidence interval for RANGE for a randomly selected city with latitude 32.3?