EXST 7014

****

Lab 4: Simple Linear Regression and Curvilinear Regression

****

OBJECTIVES

1. Use PROC REG to fit a simple linear model;
2. Check the assumption of homogeneous variances and normality test;
3. Use PROC REG to fit curvilinear models.

Simple linear regression (SLR) is a common analysis procedure used to describe the significant relationship between two variables in such a manner that one variable can be predicted or explained by using information on the other. By using PROC REG and PROC UNIVARIATE, we learned how to evaluate the SLR model comprehensively through interpreting ANOVA table, R^2, parameter estimates, residual plot, normality test and diagnostic statistics.

However, many systems encountered in research are curvilinear relationships instead of simple linear relationships. Luckily, many curvilinear relationships may be expressed in linear relationships.

During the last lab, you might be aware that heterogeneity of variance is a common violation of one of the assumptions of linear regression, which assumes a constant variability about the regression line. If the variability increases as the values of the predicted value increases then certain transformations are applied. Among the choices are the log, square root, and reciprocal transformations. Usually the need for one of these transformations is determined by examining the residual plot. If the residual plot is fan shaped then heterogeneity of variance is assumed. Log transformation is the most commonly used to alleviate a problem with heterogeneity of variance. Using log transformation implies underlying relationship is exponential. If the transformation works and the underlying relationship is exponential then the regression model should improve, and the residual plot should be more oval than fan shaped.

LABORATORY INSTRUCTIONS

Part I.
Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78 ps=53;
title1 'EXST7014 lab 4, Name, Section#';
ods rtf  file = ‘c:/temp/lab4.rtf’;
ods html file = ‘c:/temp/lab4.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://www.stat.lsu.edu/EXSTWeb/StatLab/DataSets/FW&M%20Data%202010/TEXT/DATATAB_8_31.TXT

```
Data survival;
Title2 'Survival of liver transplant patient';
Input obs clot prog enz liv time;
logtime=log(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 between time and enz ';
Plot TIME*ENZ;
Run;

Proc plot data=survival;
Title2 'Scatter plot between log-time and enz';
Plot LOGTIME*ENZ;
Run;
```

Part II.

Fitting Simple Linear Regression model

$\text{TIME} = \beta_0 + \beta_1\text{ENZ} + \varepsilon$ 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 $\sigma2$. 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=outdata1 p=Predicted r=resid cookd=cooksd dffits=diffits H=hat
    student=student rstudent=rstudent lclm=lclm uclm=uclm lcl=lcl ucl=ucl;
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 of variance, 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.

```
Proc plot data=outdata1;
Title2 'Residual Plot (Simple Linear Regression)';
Plot resid*predicted;
Run;
```

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.

```
Proc univariate data=outdata1  normal plot;
Title2 'Residual Analysis (Simple Linear Regression)';
Var Resid;
Run;
```

Part III.

Fitting Exponential Model

$\text{logTIME} = \beta_0 + \beta_1\text{ENZ} + \varepsilon$ where Y is the logTIME, X is the ENZ, and ε is a random error term that is normally distributed with mean 0 and unknown variances σ2. 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 on logarithms transformation';
Model logtime=enz/p clb cli clm influence;
OUTPUT out=outdata2 p=Predicted r=resid cookd=cooksd dffits=diffits H=hat
    student=student rstudent=rstudent lclm=lclm uclm=uclm lcl=lcl ucl=ucl;
Run;

Proc plot data=outdata2;
Title2 'Residual plot (Log Transformation)';
Plot resid*predicted;
Run;
```

```
Proc Univariate data=outdata2  normal plot;
Title3 'Residual Analysis (Log Transformation)';
Var Resid;
Run;

ods rtf close;
ods html close;
```

LAB ASSIGNMENT

Your assignment is to perform necessary analysis using SAS and answer the following questions (Do not print all the output. Only print the graphs and tables that you think are relevant to your answers.)

1. Make a scatter plot to show the relationship between TIME and ENZ. What is your observation? How about the scatter plot to show the relationship between Log-TIME and ENZ?

2. Fit the simple linear regression model TIME = $\beta_0 + \beta_1$ENZ + ε. Write down the estimated regression function and exam the residual plot and normality test. Describe what you observed and make brief comments (Hints: you may need to check the ANOVA table, Parameter Estimates table, R-square, residual plot, and normality test).

3. Fit the exponential mdoel logTIME = $\beta_0 + \beta_1$ENZ + ε. Write down the regression equation. Does the model fit data well? Why? (Hints: you may need to check the ANOVA table, Parameter Estimates table, R-square, residual plot and normality test).

4. Compare the simple linear model in Question 1 and the exponential model in Question 2, do you see any improvement in the exponential model? Please give the details (such as R-square, residual plot and normality test).

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