EXST 7015
Fall 2014
Lab 04: Curvilinear Regression

OBJECTIVES

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 last a couple of labs, 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

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.
The data is available at:
http://www.stat.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;
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;
```

Fitting Simple Linear Regression model

$TIME = \beta_0 + \beta_1 ENZ + \varepsilon$ 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 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=outdata1 p=Predicted r=resid cookd=cooksd dffits=diffits H=hat
     STUDENT=student rstudent=rstudent lclm=lclm uclm=uclm lcl=ccl ucl=ucl;
run;

Proc plot data=outdata1;
Title2 'Residual plot';
Plot resid*predicted;
Run;

Proc Univariate data=outdata1  normal plot;
Title2 'Residual Analysis';
Var Resid;
Run;
```

Fitting Exponential Model

$logTIME = \beta_0 + \beta_1 ENZ + \varepsilon$ 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 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 'Exponetial Model';
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=ccl ucl=ucl;
run;

Proc Plot data=outdata2;
Title2 'Residual plot';
Plot resid*predicted;
Run;

Proc Univariate data=outdata2  normal plot;
Title3 'Residual Analysis';
Var Resid;
Run;
ods rtf close;
ods html close;
ods listing;
```

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.

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. Fit the simple linear regression model TIME = $\beta_0 + \beta_1$ENZ + ε. Write down the regression equation and exam the residual plot and normality test. Describe what you observed and make brief comments.

2. Fit the exponential model logTIME = $\beta_0 + \beta_1$ENZ + ε. Write down the regression equation. Does the model fit data well? Why? (check the ANOVA table, predicted values of parameters, R-square, residual plot and normality test).

3. 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 ANOVA table, R-square, residual plot and normality test).