EXST 7015
Fall 2014
Lab 06: Multiple Linear Regression Variable diagnostics

OBJECTIVES

In multiple regression a dependent variable can be regressed on a number of other variables (e.g. $Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \cdots + \beta_p XP_i + \varepsilon_i$). The overall test of hypothesis of multiple linear regression is H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ v.s. H_1: at least one $\beta \neq 0$. Rejection of H_0 implies that at least one of the regressors, *X1, X2, . . . , Xp*, contributes significantly to the model. In the lab5, the problem of detecting multicolinearity, caused by highly correlated variables, introduced several diagnostic statistics including the sequential parameter estimates, simple correlations, the variance inflation factor (VIF) and the condition index. In this lab, an extreme case of multicolinearity will be presented to help you fully understand those statistics.

Since there are more than one independent variables in the model of multiple linear regression, many of you have raised the question that which variables are more important than the others. By using partial SS F-test (Type II, III, IV) and t-test of regression coefficients, the larger the F-value or t-value (the smaller the P-value), the more significant of the variable to the model as you might be aware in lab5. In addition, standardized regression coefficients and partial R^2 will be discussed to help you evaluate the relative importance of individual variables in the model in this lab.

You may have realized that the absolute value of regression coefficient is not a good predictor of relative importance of the variables. The coefficients may have meaningful interpretations (Y units per X unit), but are not directly comparable because they are on different scales. The importance of each variable can be assessed with the t-value or P value, or they can be compared by putting the variables on the same scale; standardized with a mean=0 and variance=1. The standardized regression coefficients are a relative measure of the importance of the variable.

In multiple linear regression, the R^2 for overall model is the proportion of variation in dependent variable explained by all independent variables included in the model (SSModel/SSTotal). Likewise, a partial R^2 could be calculated for each individual variable, which measures the marginal contribution of one independent variable when all the other variables are already included in model. However, such interpretation is not valid unless there is no problem of multi collinearity.

LABORATORY INSTRUCTIONS

Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78 ps=53;
title1 'EXST7015 lab 6, Name, Section#';
ods rtf  file = ‘c:/temp/lab6.rtf’;
ods html file = ‘c:/temp/lab6.html’;
```

Data set

The data set is from Chapter 6, Problem 18 in "Introduction to Regression Analysis" by Abraham and Ledolter @ 2006 Thomson Brook. This data set came from an experiment to investigate the amount of drug retained in the liver of a rat. 19 rats were weighted and dosed. The dose was approximately 40mg/kg of body weight. It can be expected that the liver is strongly correlated with body weight. After a fixed length of time the rat was sacrificed, the liver weighted and the percentage of dose in the liver was determined.

The variables are: bodyWT (body weight), liverWT (liver weight), DOSE and Y (Dose remained in liver).We will perform a multiple regression using Y as dependent variable and bodyWT, liverWT and DOSE as independent variables.

```
data Liver;
title2 'Multilinear regression_Variable Diagnostics';
input bodyWT liverWT dose Y;
datalines;
176   6.5   0.88  0.42
176   9.5   0.88  0.25
190   9     1     0.56
176   8.9   0.88  0.23
200   7.2   1     0.23
167   8.9   0.83  0.32
188   8     0.94  0.37
195   10    0.98  0.41
176   8     0.88  0.33
165   7.9   0.84  0.38
158   6.9   0.8   0.27
148   7.3   0.74  0.36
149   5.2   0.75  0.21
163   8.4   0.81  0.28
170   7.2   0.85  0.34
186   6.8   0.94  0.28
146   7.3   0.73  0.3
181   9     0.9   0.37
149   6.4   0.75  0.46
;

Proc print data=liver;
run;
;
```

Multiple Linear Regression by using PROC REG

```
Proc reg data=liver;
title2 'Multiple Linear Regression_Variable diagonostics';
model Y=bodyWT liverWT dose/all influence collin;
OUTPUT out=outdata1 p=Predicted r=resid 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 'Normality test';
var resid;
run;
```

All: Specify this option in your model is equivalent to requesting all the following options: ACOV, **CLB**, CLI, CLM, **CORRB**, COVB, I, P, **PCORR1**, **PCORR2**, R, **SCORR1**, **SCORR2**, **SEQB**, SPEC, SS1, SS2, **STB**, TOL, **VIF**, and XPX.

In this lab, we are particularly interested in the analysis performance by those in bold letters. Note that, while it is nice not having to memorize and type a lot of options, pages of possible irrelevant information are generated and you need to be able to navigate through to find what you need.

STB: prints standardized regression coefficients.

CORRB: prints the correlation matrix of estimates.

PCORR1: requests partial R^2 type I {SeqSSX/(SeqSSXi+SSError)**}**

PCORR2: requests partial R^2 type II{PartialSSXi/(PartialSSXi+SSError)}

SCORR1: requests semi-partial R^2 type I (SeqSSX/SSTotal)

SCORR2: requests semi-partial R^2 type II (PartialSSX/SSTotal)

Carefully exam the output, you will find that Type I semi-partial R^2 sums to overall R^2 since the type I SS sums to the SSReg. In contrast, partial R^2 type II is not predictable since the type II SS may sum to more or less than SSReg. The values of partial R^2 type II follows similar trends of the t-value in t-test of regression coefficients.

CLM: prints the 95% upper and lower confidence limits for the expected value of the dependent variable (mean) for each observation.

CLI: requests the 95% upper and lower confidence limits for an individual predicted value.

Collin: generates a number of collinearity diagnostics include condition indicies. If condition index exceeds 30, multi collinearity might be a problem.

VIF: the value of VIF is expected to be 1 if the regressors are not correlated. If the value is much greater than 2, serious problems are suggested.

SEQB: generate the sequential parameter estimates to exam whether there are large fluctuations as variable enters.

LAB ASSIGNMENT

Use PROC REG with appropriate options to fit the multiple linear model $Y = \beta_0 + \beta_1 bodyWT + \beta_2 liverWT + \beta_3 DOSE + \varepsilon$, and answer the following questions.

1. Report the usual results of multiple linear regression (Hypothesis test results, Parameter estimates, regression function, validity of assumptions, multi collinearity, influential statistics, etc.)

2. In the output there are two columns called “95% CL mean” and “95% CL Predicted”. Explain their difference.

3. What are the semi-Partial R^2 type I and partial R^2 type II for the variable DOSE, respectively, what is their difference?

4. Carefully exam the values of standardized regression coefficients and partial R^2 type II for individual independent variables, do you see the similar trends that you see in t-values in the t-test of regression coefficient? Make brief comments.