****

Lab 06: Multiple Linear Regression: Partial correlations and Variable diagnostics

****

OBJECTIVES

1. Use PROC REG to fit multiple regression models.
2. Familiar with standardized regression coefficients and partial correlations.
3. Variable diagnostics and multicollinearity detection.

In multiple regression however, a number of variables can be involved and regressed on one another (model: $Y = \beta_0 + \beta_1X_1+\beta_2X_2+ \cdot \cdot \cdot + \beta_pX_P + \varepsilon$) .The overall test of hypothesis of multiple linear regression is H_0: $\beta_1=\beta_2 = \cdot \cdot \cdot = \beta_p =0$ v.s. H_1: at least one $\beta \neq 0$. Rejection of H_0 implies that at least one of the regressors, $X_1, X_2, \ldots, Xp$, contributes significantly to the model. In the lab05, we know that the problem of multicollinearity caused by highly correlated variables was introduced by identify the diagnostic statistics of sequential parameter estimates, simple correlation, variance inflation factor (VIF) and condition index. In this lab, an extreme case of muticolinearlity 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.

Some of you might realize that the absolute value of regression coefficient is not a good predictor of relative importance of the variables. Why it happens is that, most often, the variables are not on the same scale or are of arbitrary scale, which leads to un-meaningful regression coefficient (Y units per X units). In such cases, the variables could be standardized with a mean=0 and variance=1. Then the standardized regression coefficients are obtained, which is the relative measurements of the importance of the variable.

In multiple linear regression, 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.

LABORATORY INSTRUCTIONS

Part I.
Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78
ps=53; title1 'EXST7014 lab 06, Name,
Section#';
ods rtf  file = 'c:/temp/lab06.rtf';
ods html file = 'c:/temp/lab06.rtf';
```

Data set

The data set is from Chapter 6, Problem 18 in "Introduction to Regession 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. Nineteen 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;
```

Part II.
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=lcl
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;
quit;
```

All: Specify this option in your model is equivalent to requesting all the following options: ACOV, CLB, **CLI, CLM**, **CORRB**, COVB, I, P, **PCORR2**, R, **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.

PCORR2: requests partial R^2 type II

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

VIF: the value of VIF is expected to be 1 if the regressors are not correlated.
If the value is very large, 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 (Hints: Hypothesis test results, Parameter estimates, regression function, and the assumptions for homogenous and normality.)

2. Is there any multicollinearilty? Why?

3. Use RSTUDENT and Hat diag to check the outliers. And also use Cook's D, DFFITS, and DFFBetas to do influence diagnostics.

4. In the output there are two columns called "95% CL mean" and "95% CL Predicted". Explain what their difference is.

5. What is the partial R^2 type II for the variable DOSE. Can it be used to evaluate the importance of DOSE?

6. 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.

*Remember to attach your SAS log with your lab report.