****

Lab 05: Multiple Linear Regression: matrix algebra and ExtraSS

****

OBJECTIVES

1. Use PROC REG to fit multiple regression models.
2. Familiar multiple regression with matix algebra and ExtraSS.
3. Detect multicollinearity in data.
4. Use PROC GLM to fit multiple regression models.

In SLR, only a single dependent variable can be regressed on a single independent variable. 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, . . . , Xp, contributes significantly to the model. As in SLR, the F-test is used to test this hypothesis. The assumptions for the multiple regression are the same for SLR. Thus the same sets of analysis, such as residual plot, normality test and diagnostic statistics are used to evaluate the assumptions.

In this lab, we will use PROC GLM and PROC REG to perform multiple linear regression. You are required to identify various types of sum-of-squares (TypeI, TypeII, TypeIII and Type IV) by using PROC GLM, and the components in X'X matrix (cross products X'X, X'Y, and Y'Y) and $(X'X)^{-1}$ matrix (X'X inverse, parameters and SSE) by using PROC REG; to understand that F-Test and T-test give the same results for parameter estimates test of hypothesis

In multiple regression, when two independent variables are highly correlated, the problem occurs because X'X matrix could not be inverted. This problem is called multicollinearity, which could cause large fluctuations of the regression coefficients and inflated variance estimates. Therefore, the regression coefficient estimates are not useful. In this lab, you will also get familiar with the statistics (sequential parameter estimates, variance inflation factor (VIF) and condition index), that evaluate the multicollinearity.

LABORATORY INSTRUCTIONS

Part I.

Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78
ps=53; title1 'EXST7014 lab 5, Name, Section#';
ods rtf  file = ‘c:\temp\lab5.rtf’;
ods html file = ‘c:\temp\lab5.html’;
```

Data set

The data set is from Chapter 8, Problem 3 in your textbook (Table 8.24). It's the results of a test for the strength of asphalt concrete mix. The test consisted of applying a compressive force on the top of different sample specimens. Two responses were collected: the stress and stain at which a sample specimen failed. The factors relate to mixture proportions, rates of speed at which the force was applied, and ambient temperature. Higher values of response variables indict stronger materials.

The variables are:
X_1: percent binder (the amount of asphalt in the mixture);
X_2: loading rate (the speed at which the force is applied);
X_3: the ambient temperature;
Y_1: the stress at which the sample specimen failed;
Y_2: the strain at which the specimen failed.
We will perform a multiple regression using Y_2 as dependent variable and X_1, X_2 and X_3 as independent variables.

The data is available at:
http://www.stat.lsu.edu/exstweb/statlab/datasets/fwdata97/FW08P03.txt

```
Data asphalt;
Title2 'The Strength of an asphalt concrete mix';
Input obs x1 x2 x3 y1 y2;
Cards;
1 5.3 0.02 77 42 3.2
2 5.3 0.02 32 481 0.73
3 5.3 0.02 0 543 0.16
4 6 2 77 609 1.44
5 7.8 0.2 77 444 3.68
6 8 2 104 194 3.11
.
.
.
;

Proc print data= asphalt;
Run;
```

Part II.

Multiple Linear Regression by using PROC REG

```
Proc reg data=asphalt;
Title2 'Full model of multi-linear regression fitted by using PROC REG';
Full: model y2=x1 x2 x3/xpx i ss1 ss2 p clb cli clm collin vif seqb;
OUTPUT out=outdata1 p=Predicted r=resid lclm=lclm uclm=uclm lcl=lcl
ucl=ucl; Reduced: model y2=x1 x3/xpx i ss1 ss2 clb collin vif;
Run;

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

Proc univariate data=outdata1 normal plot;
Title2 'Normality test: full model';
Var resid;
Run;
```

Now we have new options for the model statements:

XPX---display sums-of-squares and crossproducts matrix (X’X)
I---display inverse of sums-of-squares and crossproducts $(X'X)^{-1}$
SS1---displays the sequential sums of squares (Type I SS)
SS2---displays the partial sums of squares (Type II SS)

We can also find the extra SS by subtracting SSRegression of a reduced model from that of a full model. The full model includes all three X’s while the reduced model only have X1 and X3. Therefore, the difference of their regression gives the type II SS for X2.

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 much greater than 2, serious problems are suggested.

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

Note: the OUTPUT statement has to appear before the second MODEL statement, otherwise the residual in the output data will be for the reduced model rather than the full model.

Part III.
Multiple Linear Regression by using PROC GLM

```
Proc glm data=asphalt;
Title2 'Multi-linear regression fitted by using PROC GLM';
Model y2=x1 x2 x3/ss1 ss2 ss3 ss4;
Run;
```

GLM (General linear model) procedure works much like PROC REG except that we can combine regressor type variables with categorical (class) factors that we will learn later in the lab. In this lab, the purpose of using PROC GLM is to get all four types of sums-of-squares (Type I, Type II, Type III and Type VI), some of which (Type III and Type VI) we could not get from PROC REG.

Carefully examine the F-test (type II, type III and type VI) and T-test of the parameter estimates, and you will find the virtually identical results from two tests.

LAB ASSIGNMENT

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

1. Using PROC REG to fit the multiple linear regression model $Y2 = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$. Write down the estimated regression equation. What hypothesis does the F-test in ANOVA table test? What is the conclusion based on the ANOVA table?
2. Are the regression coefficients significant? Are they consistent with the F test in question 1?

3. Suppose that there is a specimen with $X_1 = 9$, $X_2 = 2.0$ and $X_3 = 83$, estimates the strain of the specimen as well as its 95% confidence interval.
4. What are the assumptions of the fitted model? Evaluate those assumptions by using proper SAS output.
5. Fit necessary models to find SS(X3|X1, X2) and SS(X3|X2).
6. Is there indication of any possible problem of multicollinearity? Support your answer with proper SAS output.
7. Fit the above model using PROC GLM. Examine each F-test of parameter estimates with different type of SS. Which F-test results are identical to the results of T-test for parameter estimates? What is relationship between F-value and T-value?