****

Lab 07: <u>Multiple Linear Regression: Variable selection</u>

****

OBJECTIVES

1. Use PROC REG to fit multiple regression models.
2. Learn how to find the best reduced model.
3. Variable diagnostics and influential statistics

In multiple regression, 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. In the lab 4 and lab 5, we have used several statistics such as F-test, t-test of regression coefficient, standardized regression coefficients and partial R^2 to measure the relative importance of independent variables, which tell us which independent variables are more important than the others in predicating the values of the dependent variable.

Then the question is how to choose the 'best' model of multiple regression for the current data, i.e. which variables should remain in the model, to guide its application and future studies. Theoretically, the ideal model provides the best possible fit while using the fewest possible parameters, that is, a good model is easier both to fit and to interpret. In this lab, we will introduce common variable selection methods based on F-statistics or t-test of parameter estimates (the best criteria to measure the relative importance of independent variables) including Forward selection, Backward elimination, Stepwise selection that are widely used for multiple liner regression by different statistical analysis software like SAS.

Forward selection fits all possible simple linear models, and chooses the best one (largest F-statistics for type II SS or t-value for test of parameter estimate). Then all possible 2-varible models that include the first chosen variable are compared, and so on. The process continues until no remaining variable generates a significant F-statistics or t-test of parameter estimate. With this process, once a variable enters the model it remains in the model. The significant level of α, "alpha to enter".

Backward elimination starts with the full model including all the independent variables, and removes one variable at a time based on a user-defined selection criterion. The default in SAS is to remove the variable with the least significant F-test for type II SS or t-test for parameter estimate. Then the model is refitted and the process is repeated. When all of the statistical tests are significant (i.e. none of the parameter estimates are zero), the reduced model has been chosen. With this method, once a variable is dropped from the model it does not reenter. The preset significant level is called the "alpha to drop".

Stepwise selection works in much the same way as forward selection, with the exception that the significance of each variable is rechecked at each step along the process and removed if it falls below the significant threshold. Virtually this method combines forward selection and backward elimination. In this method, a variable may enter and leave the model several times during the procedure. The procedure depends on two preset significant levels, “alpha to enter” and “alpha to drop”.

LABORATORY INSTRUCTIONS

Part I.

Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78
ps=53; title1 'EXST7014 lab 7, Name, Section#';
ods rtf  file = ‘c:\temp\lab7.rtf’;
ods html file = ‘c:\temp\lab7.html’;
```

Data set

The data set is from Chapter 8, Problem 13 in “Statistical Methods” by Freund, Wilson and Mohr @ 2010 Elsevier Inc. This data set came from a study from an apartment owner to investigate what improvements or changes in her complex may bring in more rental income. From a sample of 34 complexes she obtains the monthly rent on single-bed room units and the following characteristics:

AGE: the age of the property,
SQFT: square footage of unit,
SD: amount of security deposit,
UNTS: number of units in complex,
GAR: present of a garage (0-no, 1-yes),
CP: presence of a carpet (0-no, 1-yes),
SS: Security system (0-no, 1-yes),
FIT: fitness facilities (0-no, 1-yes),
RENT: monthly rental.

We will perform a multiple linear regression using RENT as dependent variable and the others as independent variables. The data is available at:
http://www.stat.lsu.edu/exstweb/statlab/datasets/fwdata97/FW08P13.txt

```
DATA RENTS;
TITLE 'PREDICTING APARTMENT RENT';
INPUT AGE SQFT SD UNTS GAR CP SS FIT RENT;
CARDS;
7 692 150 408 0 0 1 0 508
7 765 100 334 0 0 1 1 553
8 764 150 170 0 0 1 1 488
```

```
13 808 100 533 0 1 1 1 558
7 685 100 264 0 0 0 0 471
.
.
.
;

PROC PRINT DATA= RENTS;
RUN;
```

Part II.

Multiple Linear Regression by using PROC REG

```
* Variable selection;
Proc reg data=rents;
Title2 'Multiple Linear Regression_Variable Selection';
Backward: model rent = AGE SQFT SD UNTS GAR CP SS FIT / selection=backward;
Forward: model rent = AGE SQFT SD UNTS GAR CP SS FIT / selection=forward;
Stepwise: model rent = AGE SQFT SD UNTS GAR CP SS FIT / selection=stepwise;
Run;

* Reduced model;
Proc reg data=rents;
Title2 'Reduced Model';
Reduced: model rent = /*insert your variables here'/ influence vif collin;
OUTPUT out=outdata1 p=Predicted r=resid lclm=lclm uclm=uclm lcl=lcl ucl=ucl;
Run;

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

* Residual analysis;
Proc univariate data=outdata1 normal plot;
Title2 'Normality test';
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 (Please do not print all the output. Only print the graphs and tables that you think are relevant to your answers).

1. Run all three of the selection methods discussed above. Report the final result of each method.

2. Do you get the same reduced model from three methods? Make brief comments.

3. Which model do you think is the “best” reduced model? Discuss why you choose this model.

4. Use PROC REG to fit the best reduced model. Report the usual results (Hints: hypothesis test results, parameter estimates, validity of assumptions, multicollinearity, outliers, and influential statistics).

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