EXST 7015
Fall 2014
Lab 07: Multiple Linear Regression Variable Selection

OBJECTIVES

In multiple regression, a number of variables can be involved and regressed on one another (model: $Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \cdots + \beta_p XP + \varepsilon$). The overall test of hypothesis of multiple linear regression is H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs. 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 lab 5 and lab 6, 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. In practice, however, in addition to the expensive and time-consuming processes of data collection, problems of multi collinearity, poor combination of independent variables and influential observations make the model fitting quite challenging, as we have learned in previous labs. 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-variable 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", needs to be specified

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". Today, we will focus on Backward Elimination and Stepwise Selection.

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/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),
CARP: 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.

The data are presented in Table 8.34. We will perform a multiple linear regression using RENT as dependent variable and the others as independent variables.

```
Data rents;
title2 'Multilinear Regression_Variable Selection';
input obs age sqft sd unts gar carp ss fit rent;
cards;
1 7 692 150 408 0 0 1 0 508
2 7 765 100 334 0 0 1 1 553
3 8 764 150 170 0 0 1 1 488
4 13 808 100 533 0 1 1 1 558
5 7 685 100 264 0 0 0 0 471
6 7 710 100 296 0 0 0 0 481
7 5 718 100 240 0 1 1 1 577
8 6 672 100 420 0 1 0 1 556
9 4 746 100 410 1 1 1 1 636
10 4 792 100 404 1 0 1 1 737
11 8 797 150 252 0 0 1 1 546
12 7 708 100 276 0 0 1 0 445
13 8 797 150 252 0 0 0 1 533
14 6 813 100 416 0 1 0 0 617
15 7 708 100 536 0 0 1 1 475
16 16 658 100 188 1 1 1 1 525
;
```

```
Proc print data=rents;
run;
```

Multiple Linear Regression by using PROC REG

```
Proc reg data=rents;
title2 'Multiple Linear Regression_Variable Selection';
Backward: model rent=age sqft sd unts gar carp ss fit/selection=backward;
stepwise: model rent=age sqft sd unts gar carp ss fit/selection=stepwise;
run;
```

Note: the default level of significance is 0.1 rather than the 0.05 we usually use.

```
Proc reg data=rents;
title2 'reduced model';
Reduced: model rent = /*insert your variables here*/all vif 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;
ods rtf close;
ods html close;
```

LAB ASSIGNMENT

1. Using backward selection and stepwise selection to fit the model. Report the final results of each method. Do you get the same reduced model from two methods?

2. Use PROC REG to fit the best reduced model. Report the usual results (F-statistics, parameter estimates, validity of assumptions, multi collinearity, and influential statistics).

3. Compare the full model and reduced model, do you find any difference in VIF? Make brief comments.

Suggestion; Try testing between the full and reduced models using the GENERAL LINEAR HYPOTHESIS TEST to get an overall P value.