MULTIPLE LINEAR REGRESSION
AND
SUBSET SELECTION

Prof. Tom Fomby
Department of Economics
Southern Methodist University
Dallas, TX 75275
February, 2008

Multiple Regression and Least Squares

Consider the following stochastic relationship between the output variable y_i for an i-th individual and the individual's K input variables $x_{i1}, x_{i2}, \cdots, x_{iK}$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \ , \quad i = 1,2,\cdots,N \qquad (1)$$

where $\beta_0, \beta_1, \cdots, \beta_K$ are the unknown regression coefficients (parameters) of the model and its is assumed that the stochastic errors ε_i are independent and identically distributed with zero mean and constant variance. In this setting it is optimal (ala the Gauss-Markov theorem) to estimate these parameters by the method of **least squares** which minimizes the following sum of squares errors function

$$S = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_K x_{iK})^2 \qquad (2)$$

over choices of $\beta_0, \beta_1, \cdots, \beta_K$. The solution for this problem comes in the form of the following normal equations:

$$\sum_{i=1}^{N} 2(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_K x_{iK})(-1) = 0$$
$$\sum_{i=1}^{N} 2(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_K x_{iK})(-x_{i1}) = 0$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$\sum_{i=1}^{N} 2(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_K x_{iK})(-x_{iK}) = 0$$

The "normal" equations represent a system of K non-homogeneous, linear equations in the K unknowns $\beta_0, \beta_1, \cdots, \beta_K$ and can easily be solved by means of matrix algebra assuming that none of the input variables can be written as a linear combination of the

other input variables (i.e. the absence of perfect multicollinearity). Let us denote these solutions, called the **ordinary least squares** estimates, $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_K$.

The fit of equation (1) is represented by the sum of squared errors (SSE)

$$SSE = \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK})^2 \qquad (2)$$

where $\hat{\varepsilon}_i$ denotes the residual of the regression fit of the observation on the output variable for the i-th individual and the fitted value of the observation of the output of the i-th individual is represented by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_K x_{iK} \quad . \qquad (3)$$

In this setting, the ordinary least squares estimates are optimal with respect to being the best of all unbiased, linear estimators of the parameters $\beta_0, \beta_1, \cdots, \beta_K$.

A measure of the goodness-of-fit of equation (1) is represented by the so-called R-square (R^2) or coefficient of determination:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \qquad (4)$$

where $SST = \sum_{i=1}^{N} (y_i - \bar{y})^2$ is the total sum of squares of the deviations of the observations from their mean and $SSR = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$ is the so-called sum of squares due to the regression. R^2 is defined in such a way as to satisfy the inequality $0 \le R^2 \le 1$. In words R^2 is the percentage of the variation in y_i that is explained by the input variables $x_{i1}, x_{i2}, \cdots, x_{iK}$.

As a measure of goodness-of-fit, R^2 is **inappropriate** for making a choice between competing regression models with differing numbers of input variables because R^2 can be arbitrarily grown to be as close to 1 (a perfect fit) as one would like by simply adding more and more input variables to the regression (1). An alternative measure of goodness-of-fit that contains a "penalty" for growing regression models unnecessarily large is the so-called **adjusted R^2 criterion**:

$$\bar{R}^2 = 1 - \frac{(N-1)}{(N-K-1)}(1 - R^2) .$$

This measure is not monotonically increasing in the number of input variables but instead reaches a point of decline after some large number of input variables has been introduced to the regression. Therefore, when comparing competing regression models by means of the $\overline{R}^2$ criterion, the preferred regression is the one that possesses the largest $\overline{R}^2$.

Exhaustive Search

One approach for choosing the "best" single regression equation for describing the relationship between the output y_i and the input variables $x_{i1}, x_{i2}, \cdots, x_{iK}$ is to search over **all** possible regressions to find the regression that provides the largest $\overline{R}^2$ goodness-of-fit measure. Of course this "exhaustive" (all subsets) search can be a very extensive one, especially when K becomes larger than 20. (The total number of possible subsets is given by the combinatorial expression $\binom{K}{K}+\binom{K}{K-1}+\cdots+\binom{K}{2}+\binom{K}{1}$, where $\binom{K}{M}=\frac{K!}{(K-M)!M!}$, and where, for any integer M, $M!=M\cdot(M-1)\cdot\cdots\cdot 2\cdot 1$. However, given the fast computational capabilities offered by current day computers, the "all subsets" is quite feasible in most cases.

Other Popular Subset Selection Algorithms

Three popular iterative search algorithms for choosing a "best subset" regression are **forward selection**, **backward elimination**, and **stepwise regression**. In contrast to all subset searches based on a goodness-of-fit criterion, these algorithms are called "directed search" algorithms because they avoid all subset searches by following certain rules in conducting the search. In **forward selection** we start with no input variables and then add one input variable at a time. The input variable chosen to be added to the regression equation is the one which makes the largest contribution to R^2 on top of the input variables already in the equation. Equivalently, the input variable chosen is the remaining input variable that has the **smallest** p-value for its t-statistic. Of course the smaller the p-value, the more significant the input variable is, statistically speaking. Moreover, in the forward selection method, once an input variable enters an equation, it remains in the equation thereafter. Of course, a decision must be made as to when the contribution to the regression of the best next input variable is so small as to warrant stopping the selection process. This is, of course, a subjective decision in many regards. In some software programs (e.g. SAS) the user can specify a "cut off" probability (frequently chosen to be p = 0.01, 0.05, or 0.10) to stop the forward selection process. (In SAS this cut off probability is labeled "SLENTRY" which stands for "significance level entry." Therefore, to implement the forward selection process, one must choose the value of the "tuning parameter" p that determines the stopping point in the selection process. Obviously, the choice of this tuning parameter is quite instrumental in determining the final model chosen in the selection process. The smaller the cutoff probability is chosen to be, other things held constant, the simpler the final model will be. Conversely, the larger the cutoff probability, the more complex (extensive) the final model will be.

In **backward elimination** we start with all of the input variables in the regression equation and then at each step we eliminate the input variable that reduces R^2 by the least amount when dropping the input variable from the regression. Equivalently, the input variable that is eliminated is the one remaining input variable that has the **largest** p-value for its t-statistic. (In SAS this probability value is labeled "SLSTAY.") Moreover, in the backward elimination method, once an input variable is eliminated from an equation, it remains eliminated thereafter. Again, the smaller the cutoff probability is chosen to be, other things held constant, the simpler the final model will be when using the backward elimination method. Conversely, the larger the cutoff probability, the more complex (extensive) the final model will be.

Finally, there is the **stepwise regression procedure** which is like the forward selection and backward selection procedures except the "once in, always in" and "once out, always out" rules of forward and backward selection are no longer invoked and variables can come in and drop out, and even can come back in again until there is no remaining variable in the equation that is not statistically significant at the required levels. To affect this procedure you need to choose two "tuning parameters," namely the probability level that determines when an input variable should be admitted to the equation (in SAS "SLENTRY") and when an input variable should be eliminated from the equation (in SAS "SLSTAY"). The entry probability is invariably set higher than the exit probability. Of course, the higher these two probabilities are set in the stepwise procedure, the simpler the final model will tend to be, while the lower these two probabilities are set, the more complex the final model will tend to be.

Of course, these four best subset selection procedures need not choose the same subset regression given a specific training data set. Even for a given "architecture" (i.e. subset selection method) the choice of the probability value (tuning parameter) is crucial in determining the final model chosen for use in prediction. **Following the data mining approach of using data partitioning and cross-validation to choose between competing models, one could generate several subset models generated by different selection methods and cut-off values and then choose the subset regression that performs the best (i.e. has the best predictive accuracy) in the validation data set.** Beyond, this choice, however, one could build **an ensemble model** made up of the "better"subset regression models and, hopefully in doing so, have a model that, **in a test data set**, performs even better in the test data set than the very "best" subset model determined by the validation data set.

The "Too Small" P-Values Derived by Subset Selection Algorithms

Another way for controlling the tendency to "over-train" regression models chosen by subset selection algorithms is to **adjust upward** the p-values that the computer reports for the chosen variables after the subset selection algorithm has completed its task. The problem with forward selection or backward selection algorithms is that, instead of having tested the significance of the variables in the final model only once, the

final model is derived by a sequence of tests that proceed one at a time from one model specification to the next.

To see what distortion this multiple individual significance testing causes in terms of the implied Type I error of the testing process, consider the following two-input regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \quad .$$

Suppose, in truth that $\beta_1 = \beta_2 = 0$ and that neither input is related to the output variable. Furthermore, for simplicity, let us assume that the two input variables, x_{i1} and x_{i2}, are independent (orthogonal) to each other and, thus, the t-statistics of the two inputs are independent of each other. What is the probability of finding one or both t-statistics to be statistically significant, that is, greater in absolute value than 1.96? It is

$$1 - \Pr(|t_1| \le 1.96) \cdot \Pr(|t_2| \le 1.96) = 1 - (0.95)^2 = 0.0975 \quad .$$

Thus the Type I error associated with the multiple separate t-tests is 0.0975. It is not 0.05 as implied by using the individual t-test critical values of ± 1.96. Likewise, when starting with C independent candidate input variables, the probability of finding one or more of the inputs to be statistically significant at the 5% level is $1 - (0.95)^C$. This probability of course approaches one as C approaches infinity indicating the likelihood of finding at least a few significant input variables when searching over a multitude of candidate input variables all of which may not be important at all.

This point of the likelihood of finding at least one significant input variable among many potential input variables even if none of them is, in the population, significant is demonstrated in the Monte Carlo program Cross Validation.sas that is posted on the website for this course. As a pragmatic approach, one can adjust the final p-values obtained from a subset selection search by using the rule-of-thumb suggested by Lovell (1983). (You will find Lovell's paper posted on the class website under the file named Data Mining_Lovell.pdf.) He suggests multiplying the p-values of the final subset model by the factor (C/K) where C represents the original number of candidate input variables (not including the intercept term) and K represents the number of input variables (not including the intercept term) that remain after the final subset model has been selected. One can see in the Cross Validation.sas program that starts with 20 spurious input variables if, say a subset of 2 variables is selected in the final model, all reported p-values should be multiplied by a factor of 20/2 = 10 which would, more often than not, reveal that the retained variables are more than likely spurious in their effects on the output variable.

The above comments equally apply to the "all subset" searches. Although it can be shown that, as the sample size goes to infinity, the model with the highest $\overline{R}^2$ will, with probability one, be the correct model. Of course, we rarely have an infinite sized sample therefore this "correct model in infinity" property (i.e. consistent model choice) is

not always that comforting. Moreover, even with the all subset selection methods, the reported p-values of the final model will need to be adjusted because the computer has computed them under the assumption that the final model is the only one that was considered in the process of selection.

All of this said **it seems a good practice to use a validation data set to test the predictive power of a subset selected regression model and to use the test data set to obtain unconditional p-values of its retained variables.**