Jared Yu
Module 10 Discussion

In building a regression model for the response variable y with a set of k regressors x_i, $i = 1, \cdots, k$, let r_{y,x_i} denote the observed correlation coefficient between y and x_i, $i = 1, \cdots, k$. Can the values of r_{y,x_i} suggest what subset of the regressors is likely to be selected when any of the subset selection techniques are applied to a set of orthogonal regressors? If so, explain how. [Note: the regressors x_i and x_j are orthogonal if $\sum_{h=1}^{n}(x_{ih} - \bar{x}_i)(x_{jh} - \bar{x}_j) = 0$], where n is the number of observations].

Ans:

The interpretation of the question is that all the potential regressors being considered are *orthogonal*. In other words, for x_i, $i = 1, \cdots, k$, $\sum_{h=1}^{n}(x_{ih} - \bar{x}_i)(x_{jh} - \bar{x}_j) = 0$, for $i \neq j$. If all regressors are orthogonal, then that seems to apply that they are all *uncorrelated*, since $r_{X,Y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2(Y-\bar{Y})^2}}$, where if they are orthogonal then the numerator will zero out. So, the question seems to be asking whether the subset selection methods in Chapter 10 Section 10.2.2 are impacted by this aspect of the dataset. Or more specifically, whether r_{y,x_i} will still function as intended, given that all the regressors are orthogonal to each other. The subset selection techniques that are going to be considered are *forward selection*, *backward elimination*, and *stepwise regression*.

The forward selection procedure works with the initial assumption that no regressors are included, except the intercept term. It will first add a regressor depending on which has the largest *simple correlation*, $r_{x_i,y}$. Furthermore, the resulting F statistic must be larger than a preselected F_{IN} value. From this stage, it is irrelevant (relatively speaking) whether the regressors are orthogonal to each other. The following regressors to be added are added depending on their *partial correlations*. For example, if x_1 is chosen in the first stage, then for the rest of the regressors, the following must be calculated,

$$\hat{x}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j}x_1, \qquad j = 2, 3, \cdots, K.$$

However, if the regressors are all orthogonal, then this will evaluate to $\bar{x}_j$ for all of the regressors (based on the SLR formula). Given this context, it is strange to think that the forward selection process still works as intended. That is, it is odd to think that it will choose a single regressor, then add the following regressors based on the size of their corresponding sample means. It seems then that $r_{x_i,y}$ works fine in the first stage, but the following stage is a bit problematic when analyzing the calculations that will be involved.

Backward elimination works differently in that initially all regressors are included into the potential first model. Then, the partial F statistic is calculated for all the regressors. For the regressor with the smallest F statistic, if it is lower than a predefined F_{OUT}, then it is dropped. This repeats until the last regressor is not dropped based on F_{OUT}. We can think of this calculation as

$$F = \frac{SS_R(x_K | x_1, x_2, \cdots, x_{K-1})}{MS_{Res}(x_1, x_2, \cdots, x_K)},$$

when looking at the $K'th$ regressor from a set of K regressors. This calculation seems to not be negatively impacted by the fact that the columns are orthogonal to each other. In fact, it seems to be ideal for the columns to be orthogonal and hence uncorrelated.

With stepwise regression, it adapts forward selection so that even after regressors are added to the model, they must be reassessed at each step by checking their partial F statistics. Therefore, it involves an F_{IN} and an F_{OUT}. The former determines whether to accept a new regressor, and the latter determines whether regressors are retained in the model. Like forward selection, in terms of the first regressor being added there is no issue. However, adding new regressors are possibly problematic for the same reason as before. The additional stage of checking whether the regressors should be retained however is not impacted.