Multiple Regression in Practical Applications

In Chapters 11–12 we defined and described the computations and statistical inference procedures associated with linear regression analysis. In this chapter suggestions are given about how to proceed to formulate regression models and progress through an analysis.

Certainly the multiple linear regression tool is used to help accomplish one or more objectives:

- Often a major objective is to explain as much of the variation in $Y$ as possible
- Certainly, the better the model the more of the variance that will be explained, and that translates into better precision in prediction.
- Thus the objective in multiple regression analysis is determining which independent variables are most strongly related to $Y$, i.e., which ones in combination do the best job of predicting.
- Unfortunately, as we have seen, multicollinearity comes into play and models that are determined by the above criterion alone may contain variables that cause some coefficients to be inaccurately or incorrectly estimated and thus result in a less than useful model.
Variable Subset Selection

• Identifying potentially useful independent variables is not always so easy. Usually, people who are experts in the substantive area of the study will be called on to provide a list of possible $x$ variables. Then data ($y$) will be collected at selected levels of these $x$ variables.

• Usually many more $x$ variables are identified than are actually needed to form a good model. Some subsets of the $x$’s are often so closely related to one another that each provides about the same information about $y$, hence only one variable from the subset is needed in the model. Identifying such redundancies is the major challenge in variable selection.

• One obvious selection procedure which may be used is to do all possible regressions. This means fit every possible submodel of the full model involving all independent variables. Then a “good” submodel is selected based on some criterion related to the model fits. Some criteria often used are:

  1. Small $s^2_{\epsilon}$
  2. Large $R^2$
• Unfortunately, the full model always gives the smallest $s_\varepsilon^2$ and the largest $R^2$. What is a “good” submodel will be one that is a compromise between the desires to optimize $s_\varepsilon^2$ and $R^2$, and use as few $x$ variables as possible. Usually more than one “good” submodel will emerge, and a decision to select one must somehow be made.

• A statistics often used to determine “good” models when all possible regressions are considered is the $C_p$ statistic. It is defined as

$$C_p = \frac{SSE_p}{s_\varepsilon^2} - (n - 2p)$$

where $SSE_p$ is the error sum of squares due to fitting the submodel containing $p$ parameters (including $\beta_0$), and $s_\varepsilon^2$ is the error SS from the full model. “Good” submodels have $C_p$ which is nearly equal to $p$, i.e., $C_p \approx p$. This technique is especially good at helping to avoid selecting overspecified model having redundant $x$ variables.

• There are some classical algorithms available which automate the subset selection process. Some people call them “best subset selection algorithms”, but they cannot possibly select the “best” because it isn’t defined. Nevertheless these algorithms are used and they often select reasonably good subsets. Some of these are:
1. Backward Elimination

- This algorithm begins with the full model

\[ y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon \]

and attempts to delete a variable \( x_j \) to obtain a model with one less independent variable.

- The variable selected for deletion is the one that causes least increase in Error SS of the full model. It can be shown that this is equivalent to finding the smallest \textit{F-to-delete} statistic

\[ F_j = \frac{SS_{\text{DELETE}_j}}{MSE} \quad j = 1, 2, \ldots \]

where MSE is the mean square error from the full model fit, and the numerator is the difference in Error SS of the two models.

- If this F-statistic is not significant at a a preset value \( \alpha \) (usually called the \textit{significance level for deletion}) the corresponding \( x \) variable is deleted from the model.

- The process continued until a variable fails to be deleted at which time the \( x \) variables remaining in the model comprise the selected subset.
2. Forward Selection

- This algorithm begins with the model $y = \beta_0 + \epsilon$, and extends model by attempting to add one variable at a time. The variable selected at each stage is the one that causes largest decrease in Error SS of the full model. This is equivalent to finding the largest F-to-enter statistic

$$F_j = \frac{SS_{ADD_j}}{MSE} \quad j = 1, 2, \ldots$$

where MSE is the mean square error from the full model fit, and the numerator is the difference in Error SS of the two models.

- If this F-statistic is significant at a preset value cut-off $\alpha$ (usually called the *significance level for entry*), the corresponding $x$ variable is added to the model.

- When a variable fails to enter the current model, the process stops and the current submodel is declared the model selected by the procedure.
3. Stepwise Regression

- This algorithm combines a forward selection step and backward elimination step in each cycle. It begins with the model \( y = \beta_0 + \epsilon \) and does a forward selection step. If a variable is added then the current model is (if \( x_j \) is added) \( y = \beta_0 + \beta_1 x_j + \epsilon \).

Now a backward elimination step is made.

- If an \( x \) variable is deleted then the current model has one fewer terms. Otherwise it is unchanged and a forward selection step is made.

- Thus alternating forward and backward steps are made until no variable is added to the current subset in a forward step and no variabale is deleted in the following backward step. Then the algorithm stops and the current subset at that time comprise the selected model.

- Any or all of the methods mentioned can be used to help decide which set(s) of independent variables will be used. It isn’t necessary to decide absolutely on only one. We can go forward with more than one, and make a final decision at a later stage in the overall analysis.
Model Formulation

- Variable subset selection yields a subset $x_1, x_2, \ldots, x_k$ of the original full set of independent variables (sometimes we will proceed to this step with the full model).

- This second step is to decide exactly which terms will be in the model. For example, if we have $x_1, x_2, x_3$, do we want the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

or

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \epsilon$$

or

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3 + \epsilon$$

or

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \epsilon$$

and there are many other possibilities.

- A reasonable way to proceed (most people do this) is to fit the model which involves only first order terms, i.e.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

and do a complete residual and case diagnostics analysis.
• For example plotting residuals versus each \( x \) may often suggest which higher-degree terms may be appropriate. Then trying various higher order term model fits may lead to desirable extensions of the first order model.

• Plotting will not, however, suggest interaction between more than two variables. If such are suspected, a model having many higher order terms, some of them higher order interactions, can be subjected to the stepwise regression method to see which terms are retained. Any of the other algorithms (forward or backward) might be used instead.

• Also, if repeat \( y \) observations have been made at one or more combination of independent variable levels, the test for lack of fit can be used to identify model which may be improved by adding higher order terms. This is not an absolute because lack of fit may be due to the need for nonlinear functions of the parameters. We won’t consider nonlinear terms and models in Stat401.

• Remember, there is usually no “best” model because there is usually no exact functional relationship among \( Y \) population means. Thus different researchers may well find different very good models for the same data.