Stepwise Regression

Pravat Uprety

Stepwise Regression

- One option in regression analysis is to bring all possible independent variables into the model in one step. This is what we have done previously.
- We use the term full regression to describe this approach. Another option for developing a regression model is called stepwise regression. Stepwise regression, as the name implies, develops the least squares regression equation in steps, either through forward selection, backward elimination, or standard stepwise regression

Forward Selection

- The forward selection procedure begins (Step 1) by selecting a single independent variable from all those available. The independent variable selected at Step 1 is the variable that is most highly correlated with the dependent variable. A t-test is used to determine if this variable explains a significant amount of the variation in the dependent variable. At Step 1, if the variable is statistically significant, it is selected to be part of the final model used to predict the dependent variable. If it is not significant, the process is terminated. If no variables are found to be significant, the researcher will have to search for different independent variables than the ones already tested.

- In the next step (Step 2), a second independent variable is selected based on its ability to explain the remaining unexplained variation in the dependent variable. Recall that the coefficient of determination R^2 measures the proportion of variation explained by all of the independent variables in the model. Thus, after we select the first variable (say, x1), R^2 indicates the percentage of variation this variable explains. The forward selection routine then computes all possible two-variable regression models, with x1 included, and determines the R^2 for each model. The coefficient of partial determination at Step 2 is the proportion of the as yet unexplained variation (after x1 is in the model) that the additional variable explains. The independent variable that adds the most to R^2, given the variable(s) already in the model, is the one we select. Then, we conduct a t-test to determine if the newly added variable is significant. This process continues until either we have entered all available independent variables or the remaining independent variables do not add appreciably to R^2. For the forward selection procedure, the model begins with no variables. We enter variables one at a time, and after a variable is entered, it cannot be removed.

Backward Elimination

- Backward elimination is the reverse of the forward selection procedure. In the backward elimination procedure, all variables are forced into the model to begin the process. Then we remove the variables one insignificant variable at a time until no more insignificant variables are found. Once we have removed a variable from the model, it cannot be re-entered

Standard Stepwise Regression

- The standard stepwise procedure (sometimes referred to as forward stepwise regression—not to be confused with forward selection) combines attributes of both backward elimination and forward selection. The standard stepwise method serves one more important function. If two or more independent variables are correlated, a variable selected in an early step may become insignificant when other variables are added at later steps. The standard stepwise procedure will drop this insignificant variable from the model. Standard stepwise regression also offers a means of observing multicollinearity problems, because we can see how the regression model changes as each new variable is added to it.

Mallows' Cp

- **Mallows' Cp** is a metric that is used to pick the best regression model among several different models.
- It is calculated as:
- $Cp = RSS_p / S^2 \quad - n + 2(P+1)$

- **Where**

RSS_p: The residual sum of squares for a model with *p* predictor variables

S^2: The residual mean square for the model (estimated by MSE)

n: The sample size

P: The number of predictor variables

Mallows' Cp is used when we have several potential predictor variables that we'd like to use in a regression model and we'd like to identify the best model that uses a subset of these predictor variables.

- We can identify the "best" regression model by identifying the model with the lowest Cp value that is less than P+1, where P is the number of predictor variables in the model.