Chapter 8 - Variable Selection Procedures

There are many ways to select the variables which are used in a multiple regression model. There is no one “best” way, although most people would agree that one wants the simplest model possible that explains the response variable adequately. The difficulty is in determining what is “adequate” and in deciding what trade-off to make in terms of model complexity for model fit.

Unlike the case with examining transformations of the variables, neither $R^2$ nor RMSE (\(\sqrt{MSE}\)) will be a particularly good criterion for selecting a best model from the $2^{p-1}$ possible multiple regression models which could be fit. $R^2$ will always increase as the number of variables increases, and RMSE will almost always decrease (until one gets too few degrees of freedom and it increases again). By such criteria, one would almost always pick the full $p$-variable model, or one close to it which leads to a needlessly complex model.

SAS offers 9 different selection procedures that are called in the model statement within PROC REG. One need only add the /selection option to the model statement followed by one or more of the 9 different selection procedures. Some of the procedures will actually pick a best model (stepwise procedure) while others may only list models which have the most optimal value of the statistic under consideration for a given number of explanatory variables (such as $R^2$). A brief review of these follows:

- /selection=none; This is the SAS default. If this is used, the requested model is fit, but no attempt to compare it to other models is made. This is used when one is sure that one wants to use a certain model, or after one has used some of the other procedures shown below to narrow the class of models being examined to a small group.

- /selection=FSLE = \(\alpha\); This is the FORWARD selection option. SAS starts with the null model (containing only the intercept parameter \(\beta_0\)), and then adds the most significant variable. After that, it adds the next most significant variable (with the first already entered into the equation). This process continues until none of the variables left unentered meets the Entry Level selection value (\(\alpha\)) specified by the SLE statement. If no SLE value is specified, SAS uses a default of \(\alpha = .50\). That is, any variable which is significant at .5 (not .05) will be entered into the model. This gives a reasonable idea of what variables might be important, but tends to keep too many unneeded ones.

- /selection=BSLS = \(\alpha\); This is the BACKWARD selection option. SAS starts with the full $p$-variable model, and deletes the least significant variable. After that, it deletes the next least significant variable remaining, etc., until all variables remaining are significant at the Stay Selection Level (SLS). If no SLS value is specified, SAS uses SLS=.10 For a small number of possible predictors, FORWARD and BACKWARD Regression (with SLE = SLS) tend to give the same final models, but as the number of predictors increases, there is a good chance that they disagree. This is why the next procedure shown was invented.

- /selection=stepwise SLS = \(\alpha_1\) SLE = \(\alpha_2\); This is the STEPWISE selection option. It is a combination of Forward and Backward Regression. For those options,
once a variable was entered (for FORWARD) or deleted (for BACKWARD), that variable was never re-examined. In STEPWISE, a variable can be added or deleted from the model several times before the final model is attained, dependent upon the other variables in the model. This is quite important when collinearity is present, because a variable which might initially have appeared insignificant (in the presence of some variables) might become very significant in the presence of others (and vice versa). The final model is achieved when no variables outside of the model meet the SLE criteria, and all in the model pass the SLS criteria. The default values for both SLE and SLS are .15 under STEPWISE.

- `/selection=maxr;` Too complex, most analysts do not use it (or `/selection=minr;`). Do not use these.

- `/selection=rsquare best=m;` Unlike the F, B, and STEPWISE options, this option does not yield a best model. It yields the $m$ models with the highest $R^2$ values for $k = 1, 2, \ldots, p-1$ predictors. If the BEST = $m$ option is not specified, this is done for all $2^{p-1}$ models, which could take an enormous amount of time and memory. A value of BEST = 5 works well in most applications. Once these models are printed out, one might want to look more closely at some of the models which were “best in the class of $p$-variable models” and use some other criteria to pick which of those are best.

- `/selection=adjrsq best=m;` This is very similar to RSQUARE, except that the $R^2$ statistic is adjusted for degrees of freedom. Unlike RSQUARE, ADJRSQ does not necessarily increase to 1.0 as more variables are added; an actual maximum can be attained. In this respect it does behave somewhat like RMSE (square root of mean square error). However, like RMSE, it is fairly useless for determining the best model, tending to over-parameterize by a lot. This method of variable selection is not used very frequently.

- `/selection=cp best=m;` This option will rank the models exactly the same way as RSQUARE does, so most people run them at the same time, if at all. The $CP$ statistic (called Mallow’s $C(P)$) is more interesting than $R^2$, since it can be used to determine the best model. (Whether the model so determined is really the best is a matter of interpretation.) Mallow’s CP is calculated as

$$C(P) = (k - p) \times [F(p) - 1] + p,$$

where $k$ = number of possible original predictors, and $F(p)$ is the $F$-statistic calculated when there are “$p$-1” variables in the model (i.e. $p$ parameters which includes $\beta_0$). Pick the simplest model such that Mallow’s $C(P)$ is approximately equal to $p$, where $p$ in this context means the number of parameters including the intercept ($\beta_0$) in the model under consideration.

- `/selection=aic best=m;` AIC (Akaike Information Criteria) can also be used determine the best model. The model chosen using this method will have the smallest AIC among all models. Again, whether it does so is really a matter of interpretation.
Letting \( k \) be total number of possible predictors in the model, the Akaike Information Criteria is given by

\[
AIC = n \times \ln(SSE/n) + (2k).
\]

- \( /selection=sbc \) \( =m \); SBC (Sawa’s Bayesian Criteria) is another statistic that can be used to determine the best model. Again, the chosen model will be the one that minimizes SBC. Whether the chosen model is indeed the best is also a matter of interpretation as with AIC and Mallow’s CP. Letting \( k \) be total number of possible predictors in the model, Sawa’s Bayesian Criteria is given by

\[
SBC = n \times \ln(SSE/n) + (k \times \ln(n)).
\]

Once one has run these selection procedures, one usually has several candidate models which one could use as best. Frequently, it really doesn’t matter which one uses, and one familiar with the data may be better able than a statistician to pick a “best” model. The five most common methods are:

1. Pick best model by STEPWISE with SLE=.20 and SLS=.10.
2. Pick the simplest model using Mallow’s \( C(P) \).
3. Pick the model that minimizes Akaike Information Criteria.
4. Pick the model that minimizes Sawa’s Bayes Criteria.
5. Calculate the PRESS statistic, and then choose the model with the smallest PRESS(RMSE), (given by \( \sqrt{PRESS/n} \)).

Note: The PRESS statistic is something which is calculated each time the “R” option is given (displays residual information). It is similar to the SSE, but is usually a little larger because it tries to estimate what the SSE would be if the chosen model were applied to a new set of \( n \) data points for the same \( X \)-values as the original data. The SSE from the original data set tends to give an overly optimistic answer, since it is based on \( \beta \)’s which have been chosen to fit the original data well, but which may not be applicable in practice. The calculation of the PRESS statistic is as follows: We first select the 1st observation and fit the model using the remaining \( n - 1 \) observations. From that observation we have calculate \( e_{(1)} = Y_1 - \hat{Y}_{(1)} \). Thus, the notation of \( \hat{Y}_{(1)} \) denotes the estimated regression line with the 1st observation deleted. If we do this for each of the \( n \) observations we end up with \( n \) PRESS residuals denoted \( e_{(1)}, e_{(2)}, \ldots, e_{(n)} \). The PRESS statistic then is given by

\[
PRESS = \sum_{i=1}^{n} e_{(i)}^2.
\]

Note: Items 2, 3, and 4 above are all methods which try to get a good fit (low SSE), but with a penalty for over-parameterizing. They’re all a bit overrated, but 4 is best. If I really had to pick a best model in a problem with a large number of possible regressors, I would probably
1. First use Method 1

2. Run PROC REG with MODEL Y = $X_1 \ X_2 \ ... \ X_{p-1}$ / Selection=Rsquare CP AIC SBC Best=5;

3. Evaluate these models by criteria 2, 3 or 4, above. If they give the same answer as the model in Step 1 I would stop.

4. If I could not pick a winner, run PROC REG on each of the previous models with the /R option and pick the candidate model with the lowest $\sqrt{PRESS/n}$. 