Using Best Subsets for Regression

Data mining is the subject of Chapter 10 and includes a wide variety of statistical procedures for exploring data, including regression analysis. The *Data Mining Ribbon* in *Analytic Solver Basic* provides some advanced options not available in Excel's *Descriptive Statistics* tool, which we discuss in this section.

Analytic Solver offers five different procedures for selecting the best subsets of variables. *Backward Elimination* begins with all independent variables in the model and deletes one at a time until the best model is identified. *Forward Selection* begins with a model having no independent variables and successively adds one at a time until no additional variable makes a significant contribution. *Stepwise Selection* is similar to *Forward Selection* except that at each step, the procedure considers dropping variables that are not statistically significant. *Sequential Replacement* replaces variables sequentially, retaining those that improve performance. The fifth procedure is *Best Subsets*. These options might terminate with a different model.

Best-subsets regression evaluates either all possible regression models for a set of independent variables or the best subsets of models for a fixed number of independent variables. It helps you to find the best model based on the Adjusted R^2. Best-subsets regression evaluates models using a statistic called *Cp*. *Cp* estimates the bias introduced in the estimates of the responses by having an *underspecified model* (a model with important predictors missing). If *Cp* is much greater than $k + 1$ (the number of independent variables plus 1), there is substantial bias. The full model always has $Cp = k + 1$. If all models except the full model have large *Cp*s, it suggests that important predictor variables are missing. Models with a minimum value or having *Cp* less than or at least close to $k + 1$ are good models to consider.

Example: Using Best Subsets for the *Banking Data* Example

We will use the *Banking Data* example in Chapter 8. First, click on the *Data Mining Ribbon*. To use the linear regression tool, click the *Predict* button in the *Data Mining* group and choose *Linear Regression*. The dialog shown in Figure 1 will be displayed.

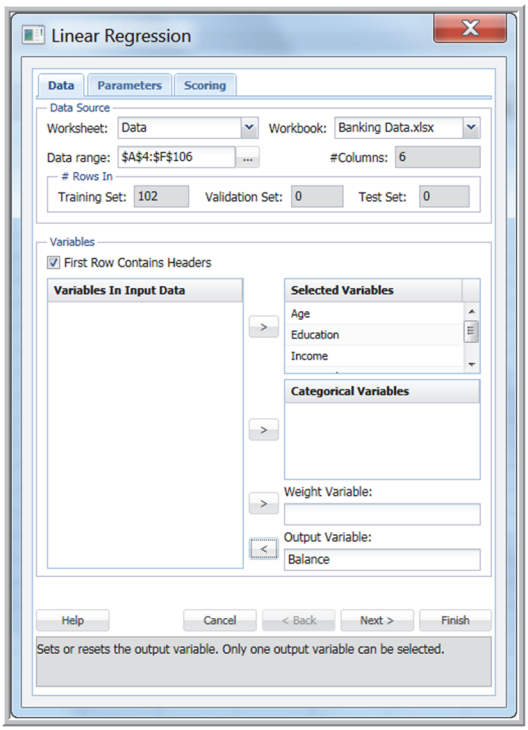


Figure 1 Linear Regression Dialog – Step 1

First, enter the data range (including headers) in the box near the top. All the variables will be listed in the left pane (*Variables in input data*). Select the independent variables and move them using the arrow button to the *Selected Variables* pane; then select the dependent variable and move it to the *Output Variable* pane as shown in the figure. Click *Next*. The second dialog shown in Figure 2 will appear.

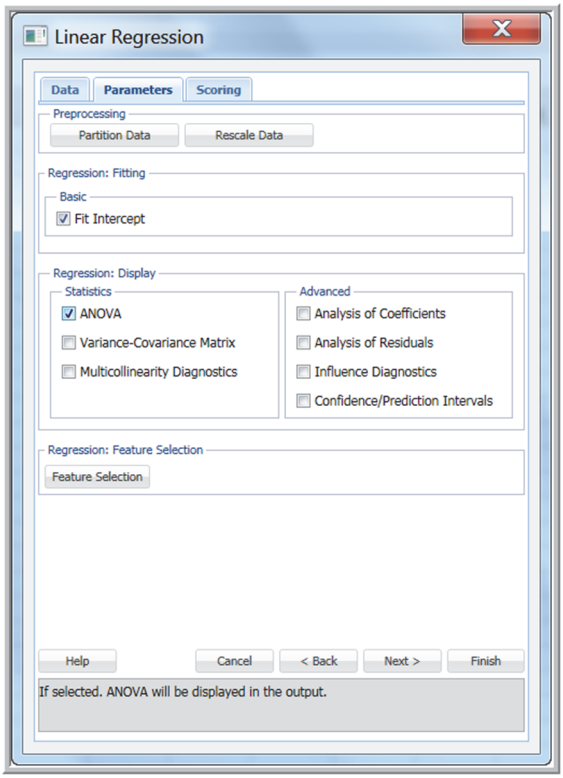


Figure 2 Linear Regression Dialog – Step 2

Select the output options (only ANOVA need be selected). However, before clicking *Finish*, click on the *Feature Selection* button. In the dialog shown in Figure 3, check the *Perform Feature Selection* box at the top and choose Best Subsets. Click *Done* and then click *Finish* in the Step 2 dialog.

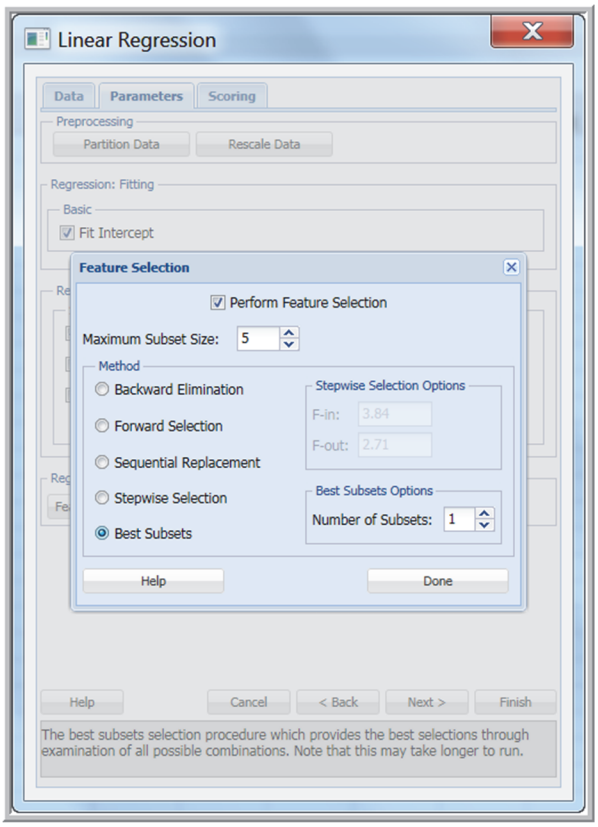
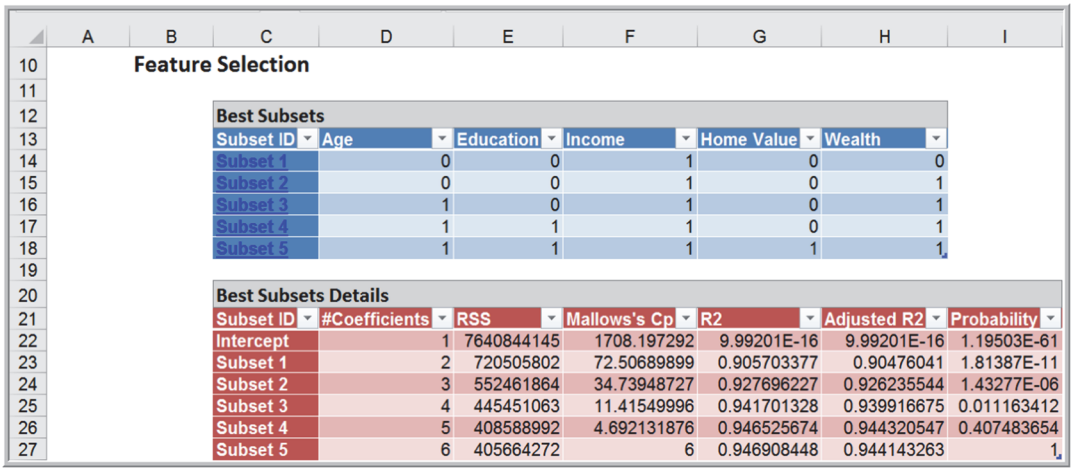


Figure 3 Linear Regression Dialog – Best Subsets Selection

Analytic Solver creates a new workbook with several worksheets. On the worksheet LinReg_FS, an "Output Navigator" allows you to click on hyperlinks to see various portions of the output (see Figure 4. Click on *Feature Selection* to see a summary of the models created during the Best Subsets selection process (Figure 5). *RSS* is the residual sum of squares, or the sum of squared deviations between the predicted probability of success and the actual value (1 or 0). *Probability* is a quasi-hypothesis test that a given subset is acceptable; if this is less than 0.05, you can rule out that subset. Although subsets 4 and 5 meet the *Cp* criterion, Subset 4 can be ruled out because of its *Probability* value. Thus, subset 5, which includes all variables, is the best. If you click on Regression Summary in the Output Navigator, the regression statistics will be displayed (Figure 6). You can ignore the Predictor Screening section. We see that these results match those of Figure 8.21. However, this model differs from that found in Example 8.13, because that approach used the backward elimination process and a different criterion (p-values) for evaluating model acceptability.

	A	B	C	D	E	F	G	H	I	J	K
1		**Data Mining: Linear Regression**									
2											
3		**Output Navigator**									
4		Feature Selection		Inputs		Regression Summary		Predictor Screening		Coefficients	
5		ANOVA		PMML Model		Training: Prediction Sum					

Figure 4 Output Navigator



Feature Selection

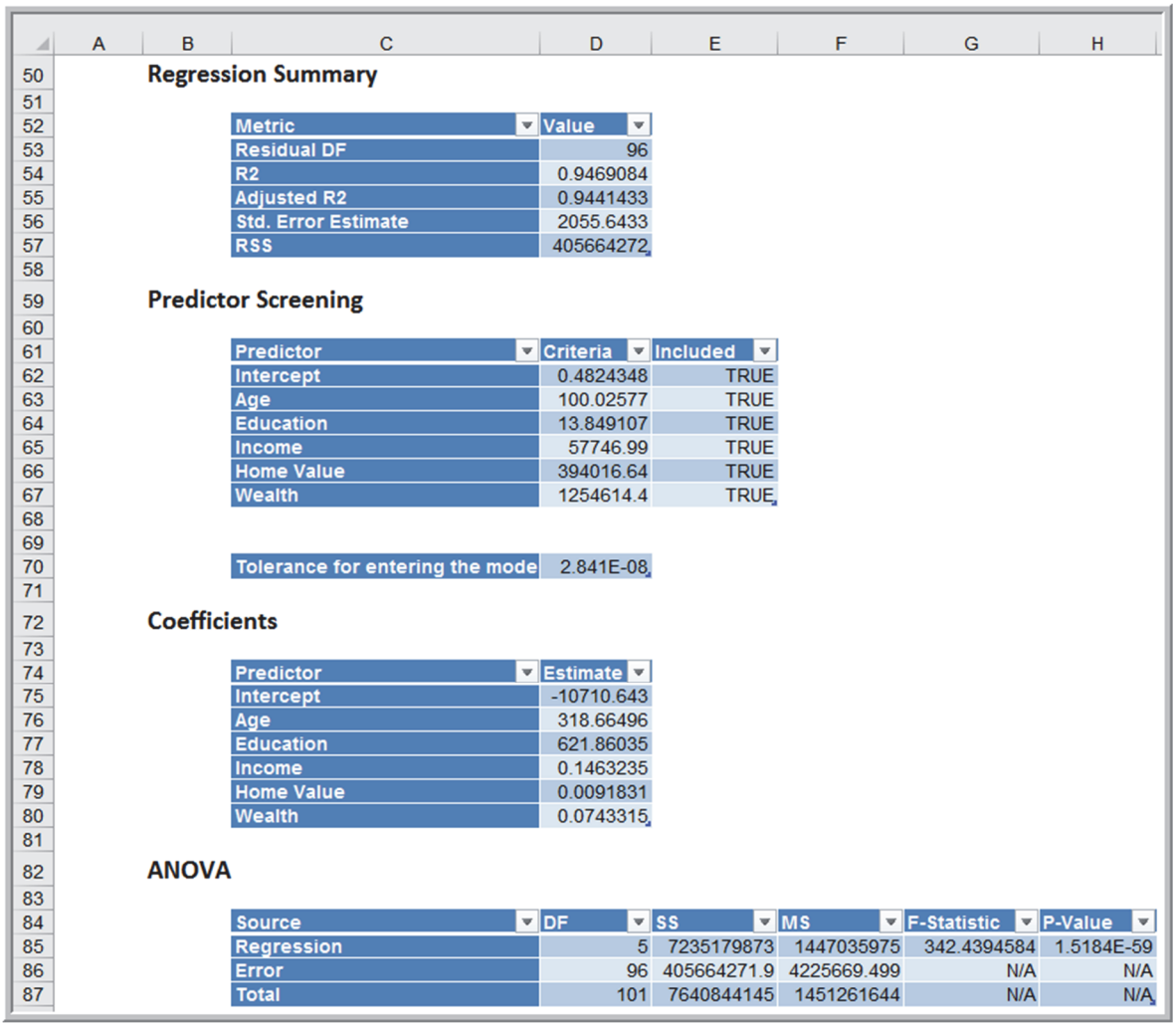
Best Subsets

Subset ID	Age	Education	Income	Home Value	Wealth
Subset 1	0	0	1	0	0
Subset 2	0	0	1	0	1
Subset 3	1	0	1	0	1
Subset 4	1	1	1	0	1
Subset 5	1	1	1	1	1

Best Subsets Details

Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2	Probability
Intercept	1	7640844145	1708.197292	9.99201E-16	9.99201E-16	1.19503E-61
Subset 1	2	720505802	72.50689899	0.905703377	0.90476041	1.81387E-11
Subset 2	3	552461864	34.73948727	0.927696227	0.926235544	1.43277E-06
Subset 3	4	445451063	11.41549996	0.941701328	0.939916675	0.011163412
Subset 4	5	408588992	4.692131876	0.946525674	0.944320547	0.407483654
Subset 5	6	405664272	6	0.946908448	0.944143263	1

Figure 5 Best Subsets Summary



Regression Summary

Metric	Value
Residual DF	96
R2	0.9469084
Adjusted R2	0.9441433
Std. Error Estimate	2055.6433
RSS	405664272

Predictor Screening

Predictor	Criteria	Included
Intercept	0.4824348	TRUE
Age	100.02577	TRUE
Education	13.849107	TRUE
Income	57746.99	TRUE
Home Value	394016.64	TRUE
Wealth	1254614.4	TRUE

Tolerance for entering the mode	2.841E-08

Coefficients

Predictor	Estimate
Intercept	-10710.643
Age	318.66496
Education	621.86035
Income	0.1463235
Home Value	0.0091831
Wealth	0.0743315

ANOVA

Source	DF	SS	MS	F-Statistic	P-Value
Regression	5	7235179873	1447035975	342.4394584	1.5184E-59
Error	96	405664271.9	4225669.499	N/A	N/A
Total	101	7640844145	1451261644	N/A	N/A

Figure 6 Regression Summary and Statistics