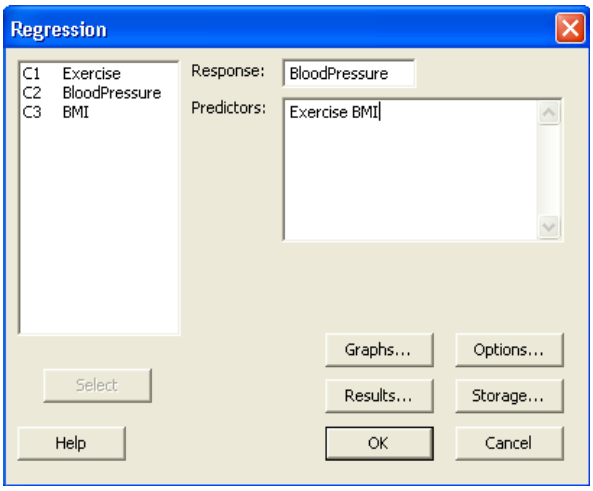
Suppose we are interested in how the exercise and body mass index affect the blood pressure and selecting the best model based on these variables. A random sample of 10 males 50 years of age is selected and their height, weight, number of hours of exercise and the blood pressure are measured. Body mass index is calculated by the following formula: $BMI\ (kg/m^2) = \frac{(Weight\ in\ pounds * 703)}{Height\ in\ Inches^2}$.

Worksheet 1 ***

	C1	C2	C3	C4	C5
	Exercise	BloodPressure	BMI		
1	4	120	18.4		
2	10	110	20.1		
3	2	120	22.4		
4	3	135	25.9		
5	3	140	26.5		
6	5	115	28.9		
7	1	150	30.4		
8	2	165	32.9		
9	2	160	33.0		
10	0	180	34.7		
11					
12					

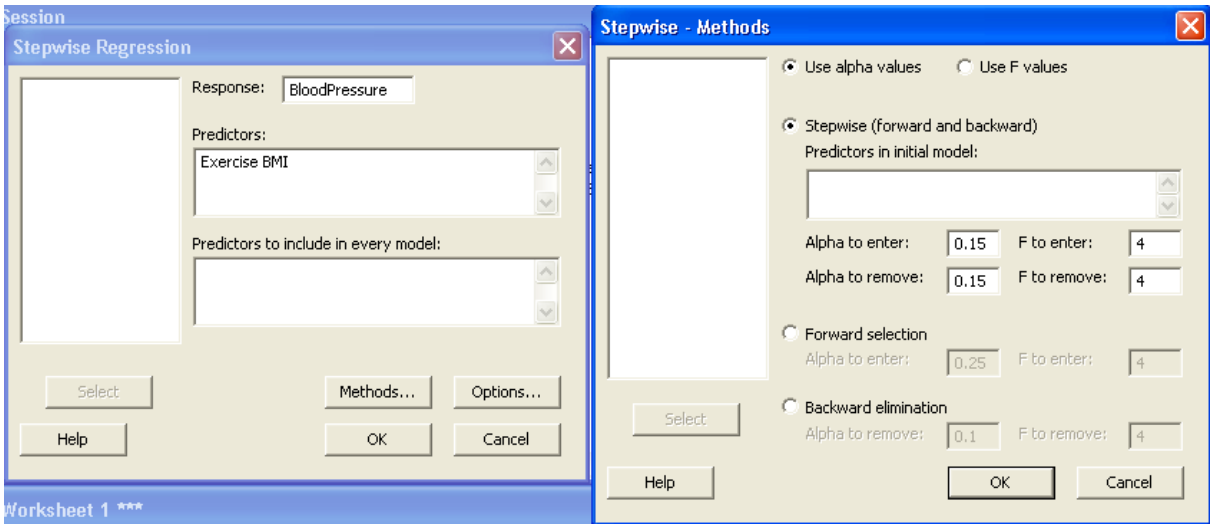


Model selection (Variable selection)

For variable selection procedure we can choose stepwise, remove, backward, and forward. This procedure automatically selects variables that are significantly important in the model using different procedure.

Stepwise Regression

To run a Stepwise, Forward Selection, or Backward Elimination model click **Stat-Regression-Stepwise…** from the drop-down menu. Then insert your variables as needed to run the appropriate test.



A Stepwise model will begin with forward selection, and it will find the most important variable to be selected. After the forward selection, the variables are then evaluated again using backward elimination to see if any of the variables should be removed. The last model presented in Minitab will be the best model found using Stepwise Regression. Using our data for a stepwise regression we find the following results:

```
Alpha-to-Enter: 0.05  Alpha-to-Remove: 0.1

Response is BloodPressure on 2 predictors, with N = 10

Step            1
Constant    40.98

BMI          3.61
T-Value      4.76
P-Value     0.001

S            12.9
R-Sq        73.94
R-Sq(adj)   70.68
Mallows Cp    3.3
```

This means our model would include BMI, but not exercise, since the alpha-to-enter=.05 and alpha-to-remove=.1 excluded exercise from being introduced to the model.

Forward Selection

Forward Selection begins by running a simple regression analysis on all candidate explanatory variables and examining the variables with the largest partial F or t-statistic and the smallest p-value. It then tests $H_0: \beta_k = 0, H_A: \beta_k \neq 0$. If H_0 is rejected, then the variable is selected for the model and the process continues. If H_0 is not rejected, the variable is not selected. When you are no longer adding new variables you have the best model created using forward selection. Using our data, the following regression equation is found using forward selection in Minitab:

```
Forward selection.  Alpha-to-Enter: 0.25


Response is BloodPressure on 2 predictors, with N = 10


Step            1      2
Constant    40.98  74.49

BMI          3.61   2.71
T-Value      4.76   2.97
P-Value     0.001  0.021

Exercise            -2.8
T-Value            -1.52
P-Value            0.171

S            12.9   11.9
R-Sq        73.94  80.43
R-Sq(adj)   70.68  74.84
Mallows Cp    3.3    3.0
```

This means our model, using forward selection with an alpha-to-enter=0.25, would include both BMI and Exercise. Our model equation would be
Blood Pressure=74.49+(2.71)*(BMI)+(-2.8)*(Exercise). This result is very similar to the multiple linear regression equation found earlier when the regression was completed. It should be noted that making the alpha-to-enter lower makes it harder to include variables. For example, if alpha-to-enter=0.15 the Exercise variable would have failed to be included since it has a p-value of 0.171, which is greater than 0.15.

Backward Elimination

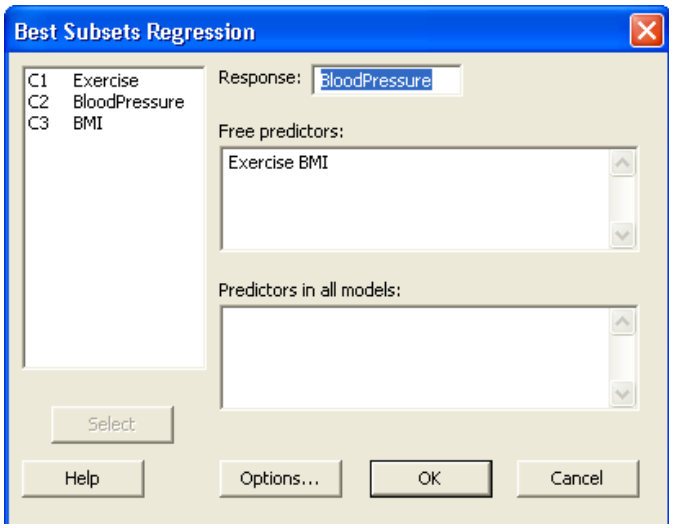
Backward Elimination is a variable selection technique that begins with all *K* candidate regressors. One of three values can then be used to decide if any and which variable(s) should be eliminated from the model: Partial *F* statistic, *t*-test statistics, *p*-value for *t*-test statistic. If the p-value for a variable is less than or equal to your chosen alpha-value you do not eliminate the variable from your model. If the p-value for your variable is greater than your chosen alpha-value you do eliminate the variable from your model. When no more variables are eliminated you have the best model created using backward elimination. Using our data, the following regression equation is found using forward selection in Minitab:

```
Backward elimination.  Alpha-to-Remove: 0.1


Response is BloodPressure on 2 predictors, with N = 10


Step            1      2
Constant    74.49  40.98

Exercise     -2.8
T-Value     -1.52
P-Value     0.171

BMI          2.71   3.61
T-Value      2.97   4.76
P-Value     0.021  0.001

S            11.9   12.9
R-Sq        80.43  73.94
R-Sq(adj)   74.84  70.68
Mallows Cp    3.0    3.3
```

The best equation found using backward elimination would be
Blood Pressure=40.98+(3.61)*(BMI), since Exercise has a p-value greater than the Alpha-to-Remove value of 0.1. It should be noted, however, that our Alpha-to-Remove value is completely subjective for what the tester wants. This means you should still looks at the S, R-Sq, R-Sq(adj), and Mallows Cp statistics to make a judgment call as to if the backward elimination model is actually more efficient or useful as explaining blood pressure. (For a brief explanation of S, R-Sq, R-Sq (adj), and Mallows Cp see further below).

Best Subset Regression

To run a Best Subsets variable selection click **Stat-Regression-Best Subsets…** from the drop-down menu. Then insert your variables as needed to run the test.



The output from the Best Subsets test will look like this:

Best Subsets Regression: BloodPressure versus Exercise, BMI

```
Response is BloodPressure

                                          E
                                          x
                                          e
                                          r
                                          c
                                          i  B
                          Mallows         s  M
Vars   R-Sq  R-Sq(adj)       Cp       S   e  I
   1   73.9       70.7      3.3  12.854      X
   1   55.8       50.3      9.8  16.734   X
   2   80.4       74.8      3.0  11.909   X  X
```

Using the R-Sq, R-Sq (adj), Mallows Cp, and S (standard error of the regression, or $s_e = \sqrt{MS_E}$) you can choose which model would be the strongest and the most useful. You want R-Sq and R-Sq (adj) to be large (as close to 1 as possible). At the same time you want S to be the smallest possible. You also need to find a Cp that is as close to the number of coefficients used in the model. Be aware, though, that the full model (in our example, the one with 2 variables) will always have Cp=K+1=the number of coefficients. *The best model will be the one that you feel best meets all these criteria.*

Model Selection Criterion

Another technique that can be used to select variables within a model or to choose among various types of models is known as model selection criterion. Specifically this tutorial covers how to solve for the AIC (Akaike Information Criterion). Minitab does not contain a default means for calculating this statistic, so information will be taken from Minitab and plugged manually into a formula. Open-source software, such as R (the statistical programming language), has tools to

calculate this and many other values as well, but this is worthy of its own specific tutorial and will not be discussed here.

The formula for AIC that can be easy calculated from Minitab is

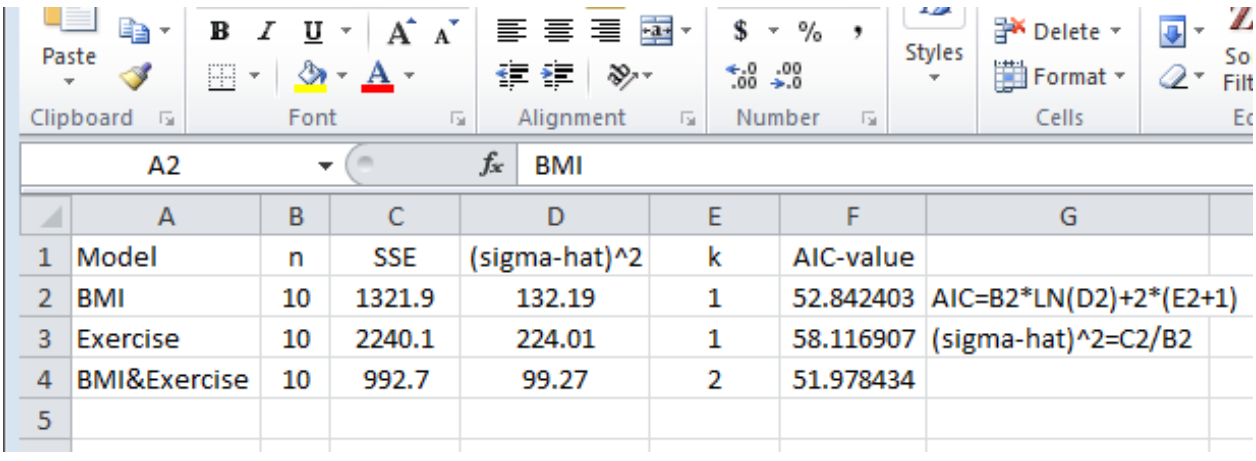
$$AIC = n * \ln(\hat{\sigma}^2) + 2(k+1) \text{ and } \hat{\sigma}^2 = \frac{SSE}{n}.$$

Where n=sample size and k=number of variables in model. Other variations of this formula also exist, but this tutorial uses this one. When comparing multiple AIC values to determine which model statement is best, the lowest AIC value is selected as the best fit model. We will use the BMI, exercise, and blood pressure set used earlier as our example here. We are attempting to determine which model is the best fit with our two variables. We will need to test the dependent variable blood pressure against the three possible combinations of our independent variables (BMI, exercise, and BMI & exercise).

First, run all the Minitab tests for regression (**Stat-Regression-Regression…**). The relevant ANOVA tables are listed below:

```
Analysis of Variance for Blood Pressure vs. BMI

Source           DF      SS      MS      F      P
Regression        1  3750.6  3750.6  22.70  0.001
Residual Error    8  1321.9   165.2
Total             9  5072.5

Analysis of Variance for Blood Pressure vs. Exercise

Source           DF      SS      MS      F      P
Regression        1  2832.4  2832.4  10.12  0.013
Residual Error    8  2240.1   280.0
Total             9  5072.5

Analysis of Variance for Blood Pressure vs. BMI and Exercise

Source           DF      SS      MS      F      P
Regression        2  4079.8  2039.9  14.38  0.003
Residual Error    7   992.7   141.8
Total             9  5072.5
```

Then this information can then be plugged into the formula. Here we use Excel:



	A	B	C	D	E	F	G
1	Model	n	SSE	(sigma-hat)^2	k	AIC-value	
2	BMI	10	1321.9	132.19	1	52.842403	AIC=B2*LN(D2)+2*(E2+1)
3	Exercise	10	2240.1	224.01	1	58.116907	(sigma-hat)^2=C2/B2
4	BMI&Exercise	10	992.7	99.27	2	51.978434	
5							

Note that the two formulas listed in column G are both for calculating the model Blood Pressure and BMI written in the notation needed for Excel.

Now we have the AIC values for each model:

	AIC
Blood Pressure and BMI	52.842
Blood Pressure and Exercise	58.117
Blood Pressure and BMI & Exercise	51.978

The final step is to determine which value is smallest. In this case the model with both variables has the smallest AIC value. Therefore we determine that this model is the best fit given our data.