SASEG 9* – Model Building; An Introduction

This SASEG is the beginning of SASEG 9C and was originally 9C but 9D had this content with additional content and so was combined. R. Freeze

(Fall 2015)

Sources (adapted with permission)-

T. P. Cronan, Jeff Mullins, Ron Freeze, and David E. Douglas Course and Classroom Notes
Enterprise Systems, Sam M. Walton College of Business, University of Arkansas, Fayetteville
Microsoft Enterprise Consortium
IBM Academic Initiative
SAS® Multivariate Statistics Course Notes & Workshop, 2010
SAS® Advanced Business Analytics Course Notes & Workshop, 2010
Microsoft® Notes
Teradata® University Network

Model Building and Interpretation

Model Selection

Eliminating one variable at a time manually for

- small data sets is a reasonable approach
- large data sets can take an extreme amount of time.

71

A process for selecting models might be to start with all the variables in the **`Fitness`** data set and eliminate the least significant terms, based on *p*-values.

For a small data set, a final model can be developed in a reasonable amount of time. If you start with a large model, however, eliminating one variable at a time can take an extreme amount of time. You would have to continue this process until only terms with *p*-values lower than some threshold value, such as 0.10 or 0.05, remain.

Model Selection Options

The Linear Regression task supports these model selection techniques:

All-possible regressions ranked using

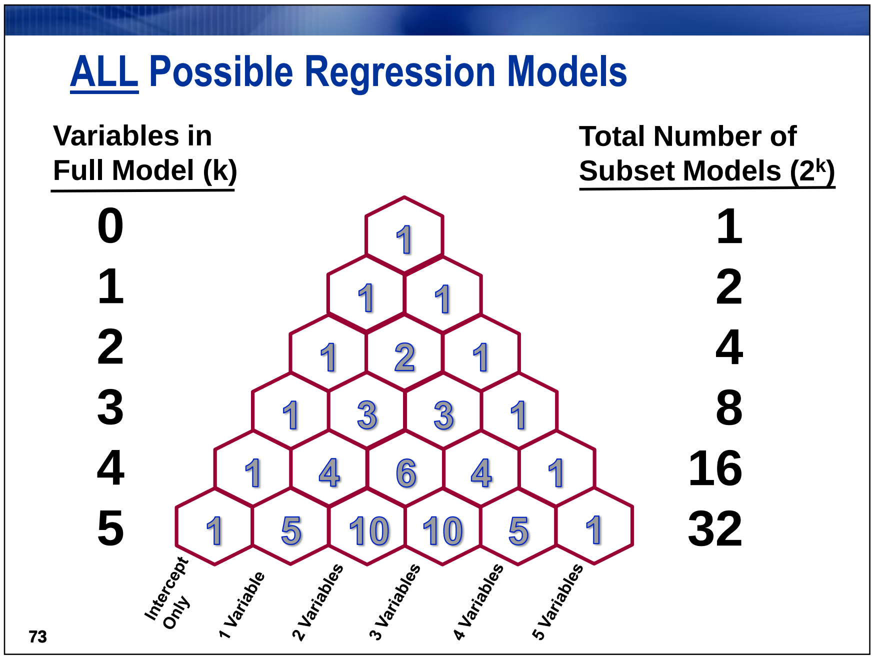
- R-squared, Adjusted R-Squared, or Mallows' Cp

Stepwise selection methods

- Stepwise, Forward, or Backward

✎ Full model fitted (no selection) is the default.

72



All-Possible Regression Techniques have in common that they literally assess each possible subset model of a given set of predictor variables in a regression model. The assessment is based on some overall model statistic value (such as R-Squared, Adjusted R-Square and Mallows' C_P). For a model with 2 predictor variables, X1 and X2, in the MODEL statement, there are 4 possible subset models: one intercept-only model (which is always a subset model); the X1 model; the X2 model; and the X1 X2 model. The intercept-only model is typically disregarded. The number of subset models for a set of *k* variables is 2^k or 2^k-1, ignoring the intercept-only model.

In the **`Fitness`** data set, there are 7 possible independent variables. Therefore, there are $2^7 - 1 = 127$ possible regression models. There are 7 possible one-variable models, 21 possible two-variable models, 35 possible three-variable models, and so on.

If there were 20 possible independent variables, there would be over 1,000,000 models. The number of calculations needed increases exponentially with the number of variables in the full model, so one must be cautious in judging when to use these techniques.

In a later demonstration, you will see another set of model selection techniques that do not have to examine all the models to help you choose a set of candidate "best subset" models.

Mallows' C_p

- Mallows' C_p is a simple indicator of model bias. Models with a large C_p are biased.
- Look for models with $C_p \le p$, where p equals the number of parameters in the model, including the intercept.
- Mallows recommends choosing the first (fewest variables) model where C_p approaches p.

$$C_p = p + \frac{(\mathrm{MSE}_p - \mathrm{MSE}_{\mathrm{full}})(n - p)}{\mathrm{MSE}_{\mathrm{full}}}$$

74

Mallows' C_p (1973) is estimated by $C_p = p + \frac{(\mathrm{MSE}_p - \mathrm{MSE}_{\mathrm{full}})(n - p)}{\mathrm{MSE}_{\mathrm{full}}}$

where

MSE_p	is the mean squared error for the model with p parameters.
$\mathrm{MSE}_{\mathrm{full}}$	is the mean squared error for the full model used to estimate the true residual variance.
n	is the number of observations.
p	is the number of parameters, including an intercept parameter, if estimated.

Bias in this context refers to the model underfitting or overfitting the data. In other words, important variables are left out of the model or there are redundant predictor variables in the model.

The choice of the best model based on C_p is up for some debate, as will be shown in the slide about Hocking's criterion. Many choose the model with the smallest C_p value. However, Mallows recommended that the best model will have a C_p value approximating p. The most parsimonious model that fits that criterion is generally considered to be a good choice, although subject-matter knowledge should also be a guide in the selection from among competing models. A *parsimonious* model is one with as few parameters as possible for a given degree of quality (predictive or explanatory ability).

Hocking's Criterion

Hocking (1976) suggests selecting a model based on the following:

- $C_p \le p$ for prediction
- $C_p \le 2p - p_{full} + 1$ for parameter estimation

75

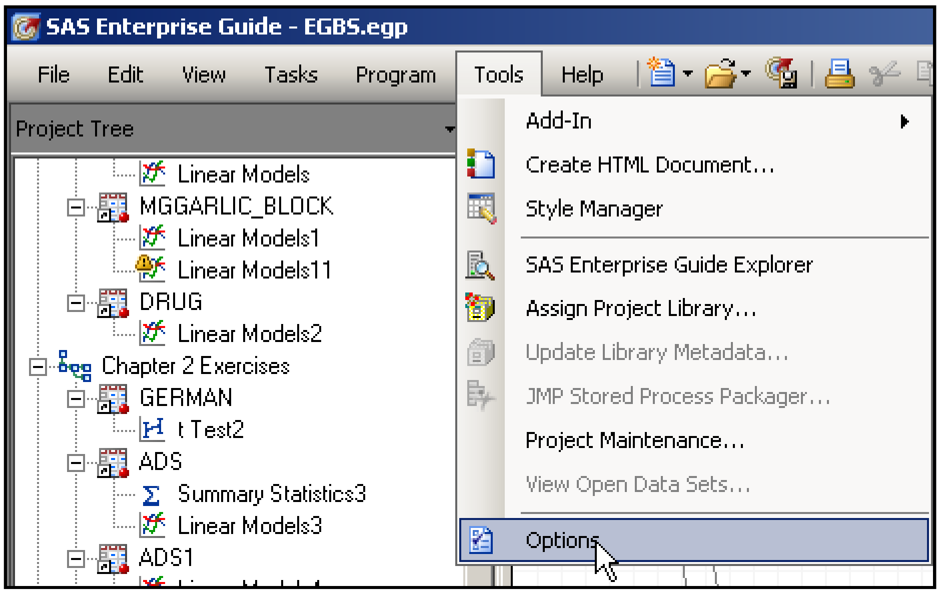
Hocking suggested the use of the C_p statistic, but with alternative criteria, depending on the purpose of the analysis. His suggestion of ($C_p \le 2p - p_{full} + 1$) is included in the REG procedure's calculations of criteria reference plots for best models.

Automatic Model Selection

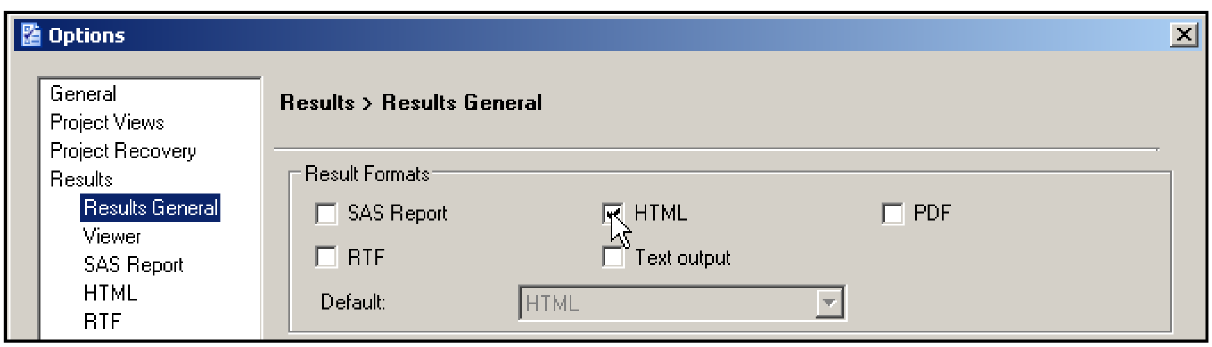
Invoke the Linear Regression task to produce a regression of **`Oxygen_Consumption`** on all the other variables in the **`Fitness`** data set and produce plots with tool (data) tips to aid in exploration of the results.

Plots with tool tips can only be created in HTML file, so before the task is created, the option to create HTML output must be selected in SAS Enterprise Guide.

1. Click **Tools** ⇨ **Options**.



2. In the window that opens, select **Results General** under `Results` at the left and then uncheck the box for **SAS Report** and check the box for **HTML**.

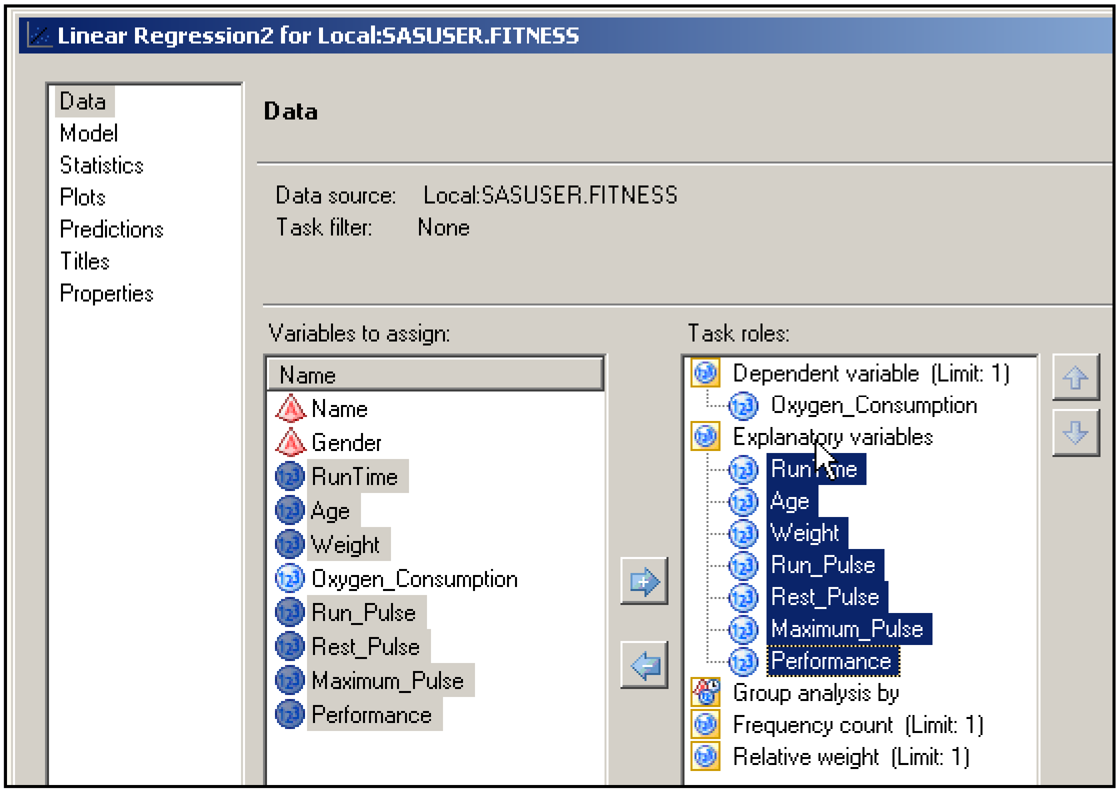


3. Click OK.

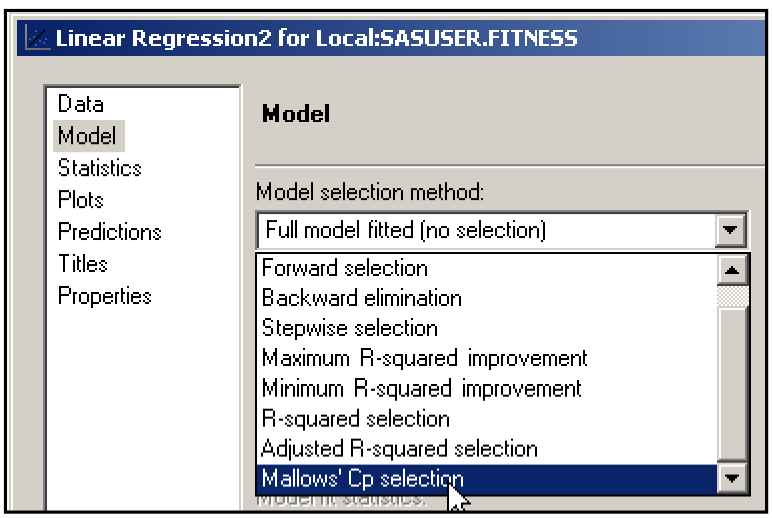
Now you are ready to run the Linear Regression task.

4. With the **Fitness** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.

5. Drag **`Oxygen_Consumption`** to the dependent variable task role and all other numeric variables to the explanatory variables task role.

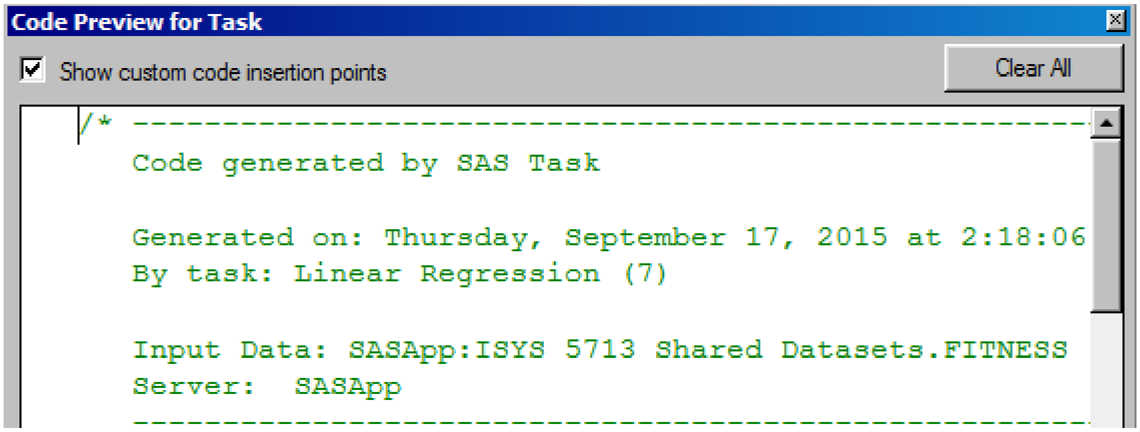


6. With **<u>Model</u>** selected at the left, find the pull-down menu for `Model selection method` and click to find **<u>Mallows' Cp selection</u>** at the bottom.



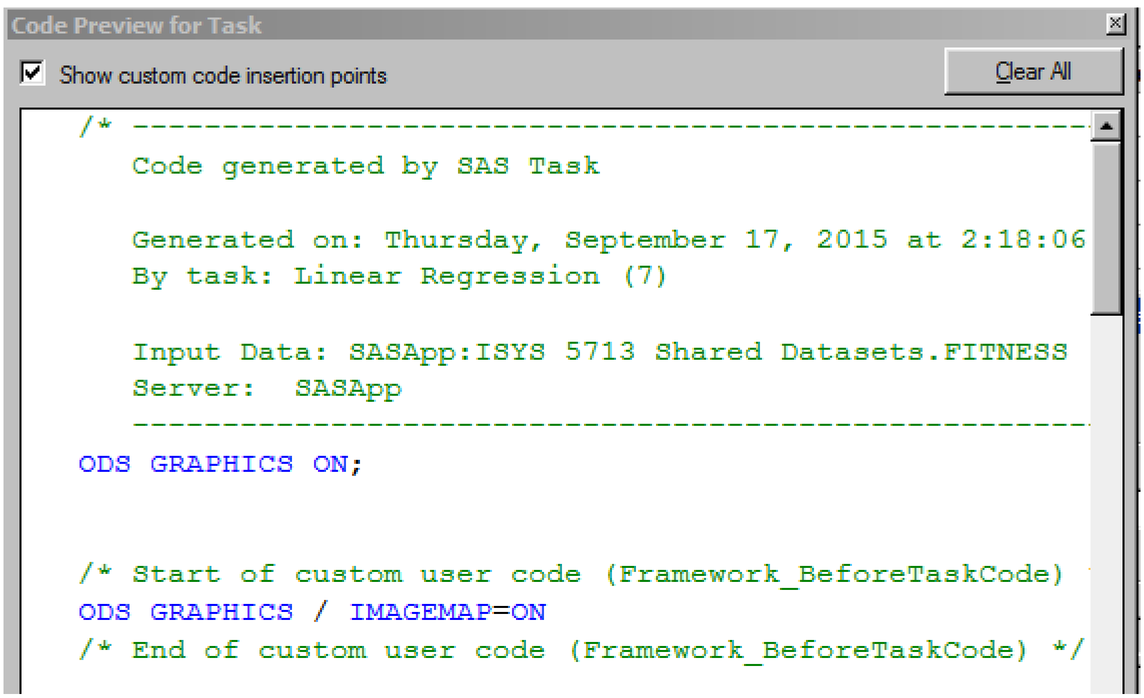
Note – under Plots, leave the defaults checked – Show plots for regression analysis > all appropriate plots

7. Click Preview code.



8. Enable the Show custom code insertion points box

9. Scroll down and under the ODS GRAPHICS ON statement, type **`ODS GRAPHICS / IMAGEMAP=ON;`** in the <insert custom code here> area



10. Click ☒ in the Code Preview for Task window.

11. Click Run.

Partial HTML Output

Linear Regression Results

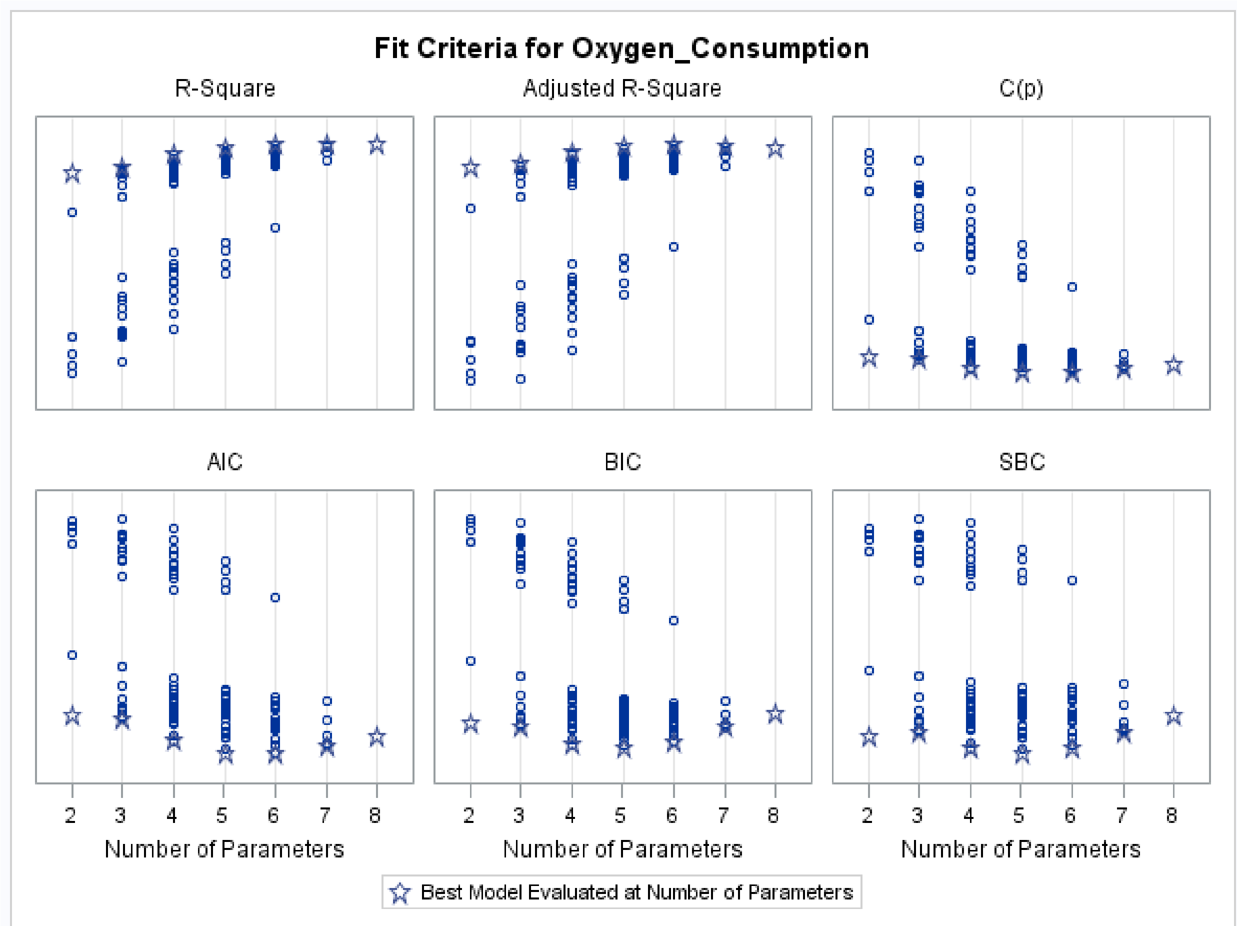
The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

C(p) Selection Method

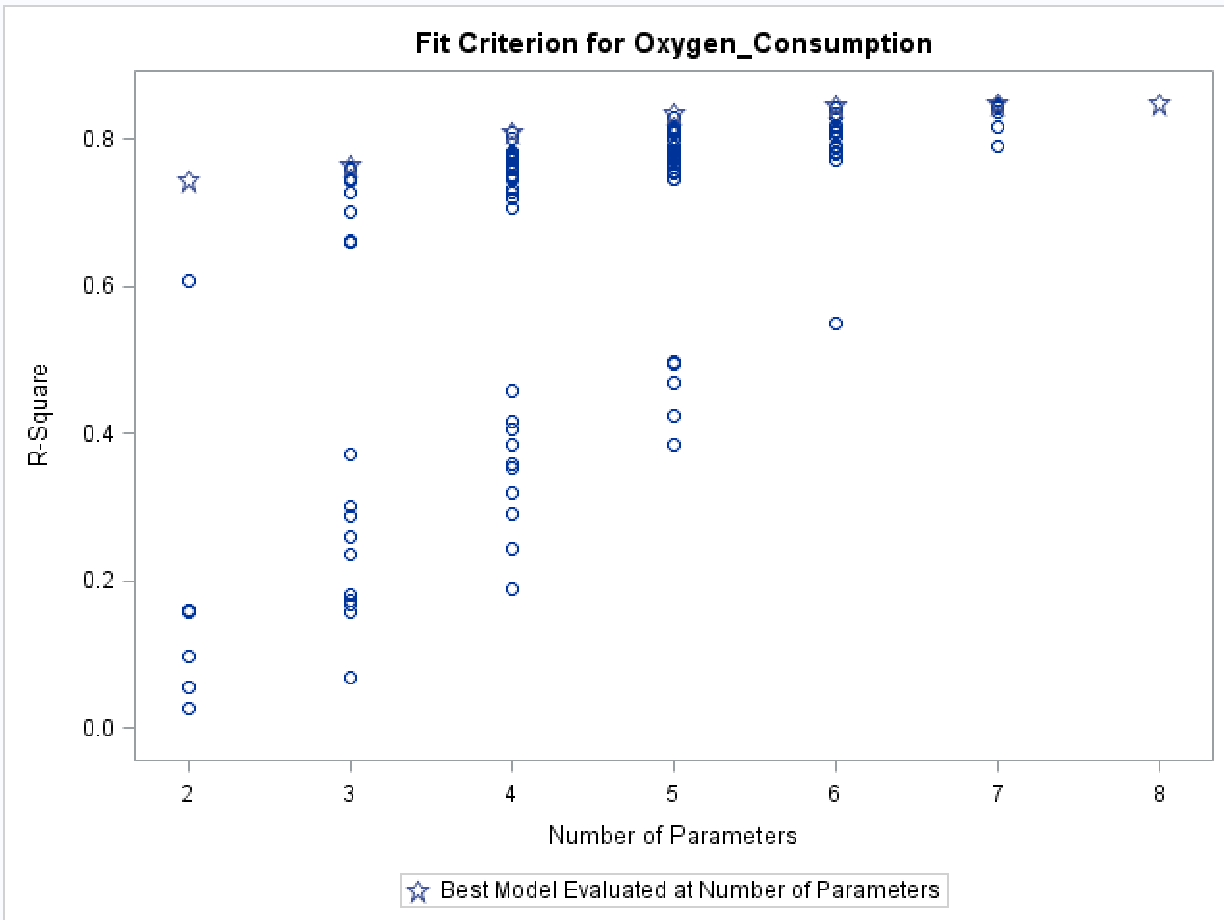
Number of Observations Read	31
Number of Observations Used	31

Model Index	Number in Model	C(p)	R-Square	Variables in Model
1	4	4.0004	0.8355	RunTime Age Run_Pulse Maximum_Pulse
2	5	4.2598	0.8469	RunTime Age Weight Run_Pulse Maximum_Pulse
3	5	4.7158	0.8439	RunTime Weight Run_Pulse Maximum_Pulse Performance
4	5	4.7168	0.8439	RunTime Age Run_Pulse Maximum_Pulse Performance
5	4	4.9567	0.8292	RunTime Run_Pulse Maximum_Pulse Performance
6	3	5.8570	0.8101	RunTime Run_Pulse Maximum_Pulse
7	3	5.9367	0.8096	RunTime Age Run_Pulse
8	5	5.9783	0.8356	RunTime Age Run_Pulse Rest_Pulse Maximum_Pulse
9	5	5.9856	0.8356	Age Weight Run_Pulse Maximum_Pulse Performance
10	6	6.0492	0.8483	RunTime Age Weight Run_Pulse Maximum_Pulse Performance
11	6	6.1758	0.8475	RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse
12	6	6.6171	0.8446	RunTime Weight Run_Pulse Rest_Pulse Maximum_Pulse Performance

There are many models to compare. It would be unwieldy to try to determine the best model by viewing the output tables. Therefore, it is advisable to look at the plots.

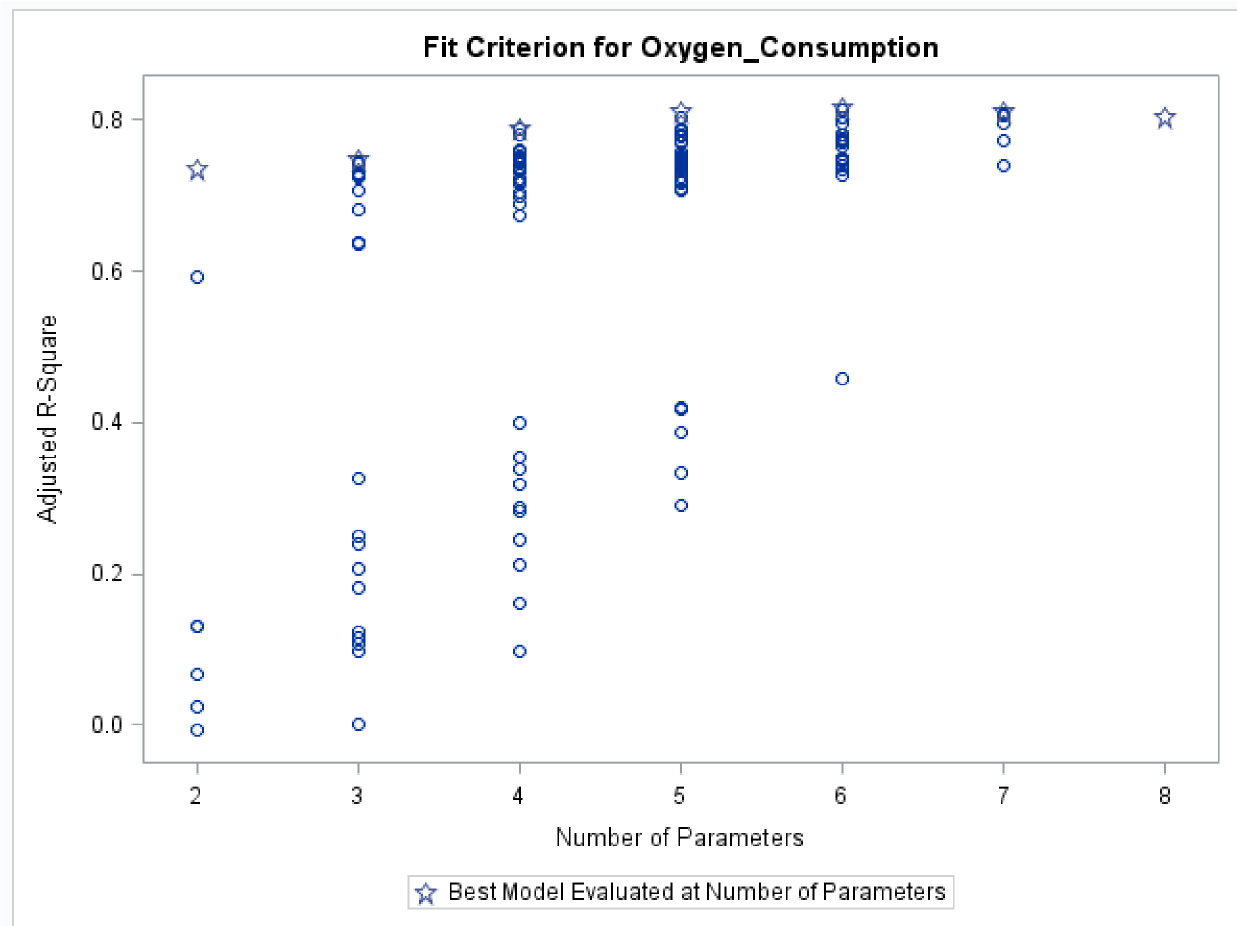

Fit Criteria for Oxygen_Consumption
R-Square
Adjusted R-Square
C(p)
AIC
BIC
SBC
2 3 4 5 6 7 8
Number of Parameters
Best Model Evaluated at Number of Parameters

The first plot is a panel plot of several plots assessing each of the 127 possible subset models. Three of them will be further described below.

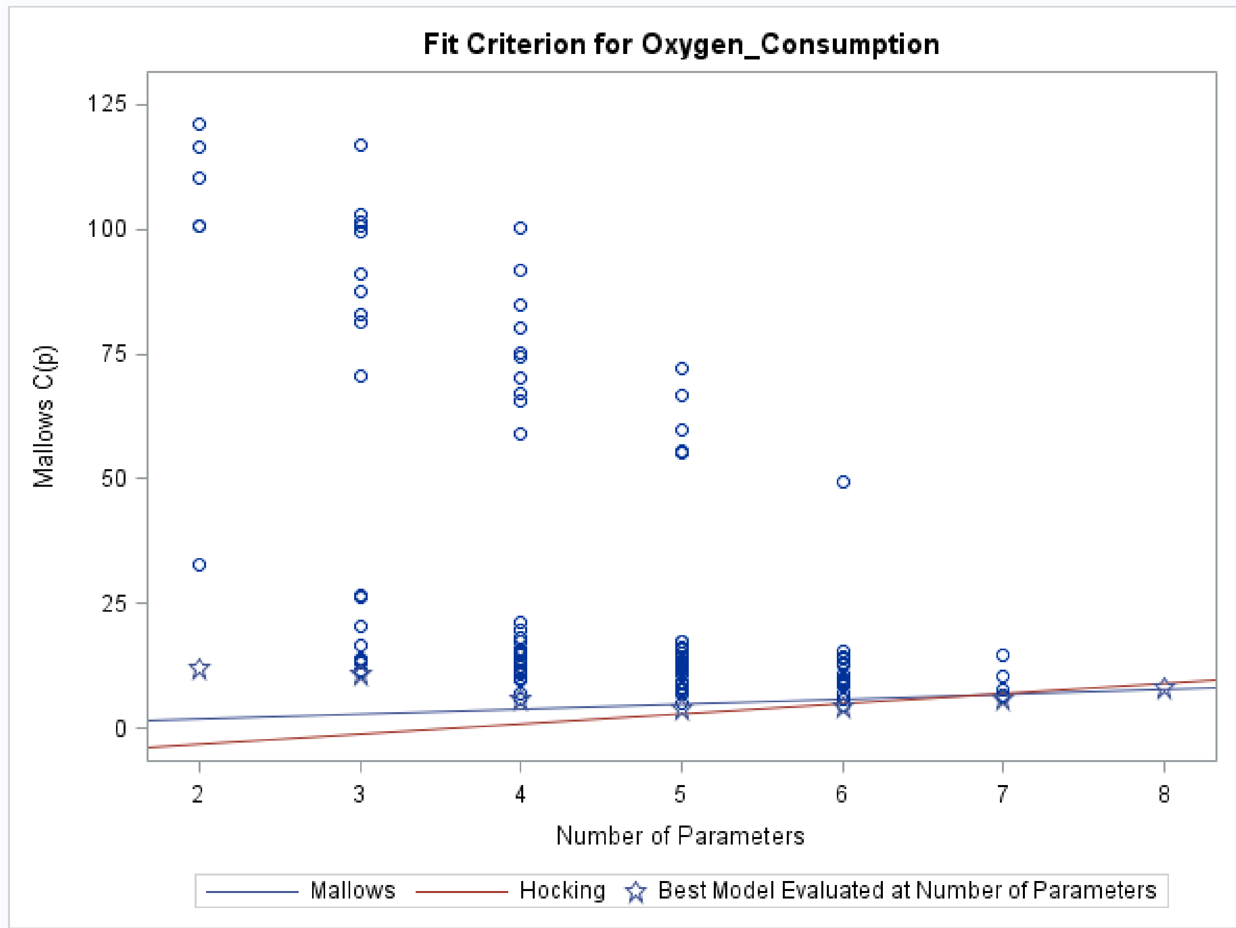


The R-Square plot compares all models based on their R^2 values. As noted earlier, adding variables to a model will always increase R^2 and therefore the full model will always be best. Therefore, one can only use the R^2 value to compare models of equal numbers of parameters.

The model with the greatest R^2 values are represented by stars within each category of "Number of Parameters".



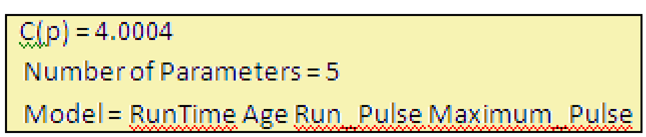
The Adjusted R-Square does not have the problem that the R-Square has. One can compare models of differing sizes. In this case, it is difficult to see which model has the higher Adjusted R-Square, the starred model for 6 parameters or 7 parameters.



The line $C_p = p$ is plotted to help you identify models that satisfy the criterion $C_p \le p$ for prediction. The lower line is plotted to help identify which models satisfy Hocking's criterion $C_p \le 2p - p_{full} + 1$ for parameter estimation.

Use the graph and review the output to select a relatively short list of models that satisfy the criterion appropriate for your objective. The first model to fall below the line for Mallows' criterion has five parameters. The first model to fall below Hocking's criterion has 6 parameters.

With tool tips activated using the IMAGEMAP=ON option, scrolling your mouse over an observation will cause a data box to hover over your mouse containing data values represented by that observation. In this case, the expanded data box shows that the first model that has a Cp value below the green threshold (where Cp=p) is:



In this example the number of variables in the full model, p_{full}, equals 8 (7 variables plus the intercept).

The smallest model with an observation below the Mallows line has $p = 5$ (Number in Model = 4). The model with the star at 5 parameters and the model just above it are considered "best", based on Mallows' original criterion. The starred model has a $C_p = 4.004$, satisfying Mallows' criterion (**Oxygen_Consumption** = **Runtime Age Run_Pulse Maximum_Pulse**) and the one above has a value of 4.9567 (**Oxygen_Consumption** = **Performance Runtime Run_Pulse Maximum_Pulse**). The only difference between the two models is that the first includes **Age** and the second includes **Performance**. By the strictest definition, the second model should be selected, because its C_p value is closest to p.

The smallest model that shows under the Hocking line has p=6. The model with the smaller C_p value will be considered the "best" explanatory model. The table shows the first model with p=6 is **Oxygen_Consumption = Runtime Age Weight Run_Pulse Maximum_Pulse**, with a C_p value of 4.2598. Two other models that are also below the Hocking line (they are nearly on top of one another in the plot) are **Oxygen_Consumption = Performance Runtime Weight Run_Pulse Maximum_Pulse** and **Oxygen_Consumption = Performance Runtime Age Run_Pulse Maximum_Pulse**.

"Best" Models – Prediction

The two best candidate models based on Mallows' original criterion includes these regressor variables:

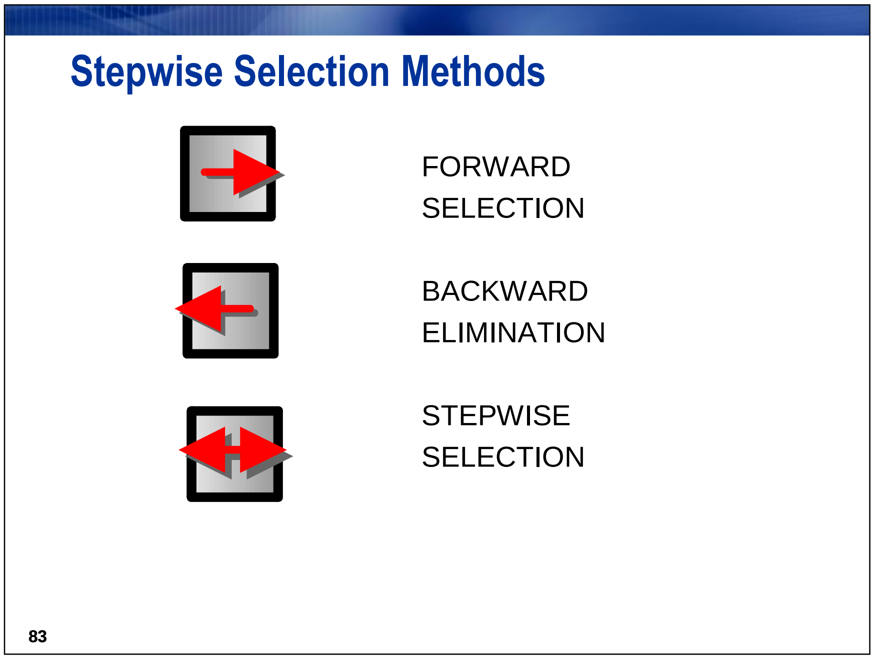
p = 5	C_p = 4.0004 R^2=0.8355 Adj. R^2=0.8102	RunTime, Age, Run_Pulse, Maximum_Pulse
p = 5	C_p = 4.9567 R^2=0.8292 Adj. R^2=0.8029	Performance, RunTime, Run_Pulse, Maximum_Pulse

77

Some models might be essentially equivalent based on their C_p, R^2 or other measures. When, as in this case, there are several candidate "best" models, it is up to the investigator to determine which model makes most sense based on theory and experience. The choice between these two models is essentially the choice between ***Age*** *and* ***Performance****. Because age is much easier to measure than the subjective measure of fitness, the first model is selected here.*

A limitation of the evaluation you have done thus far is that you do not know the magnitude and signs of the coefficients of the candidate models or their statistical significance.

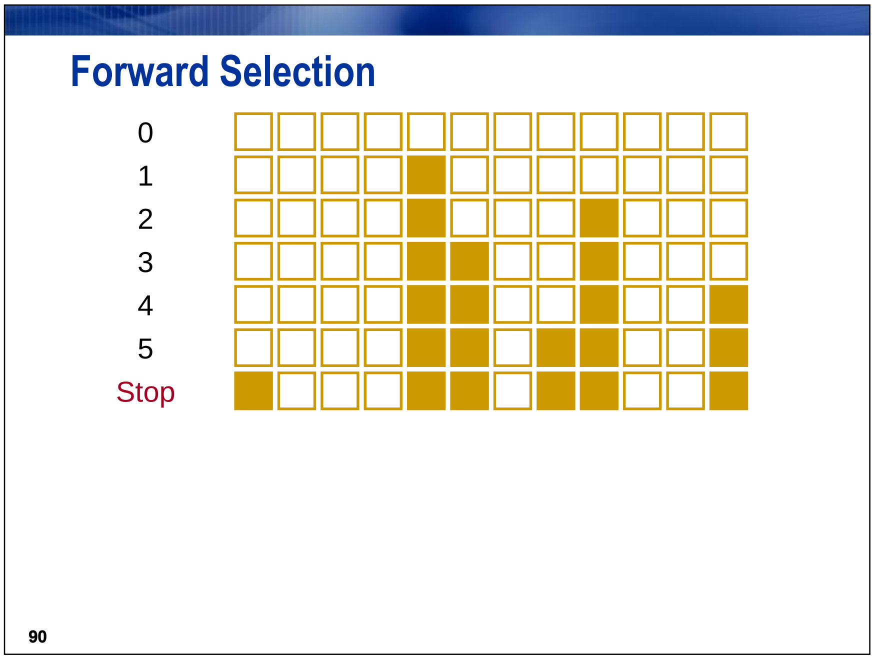
Using Stepwise Methods



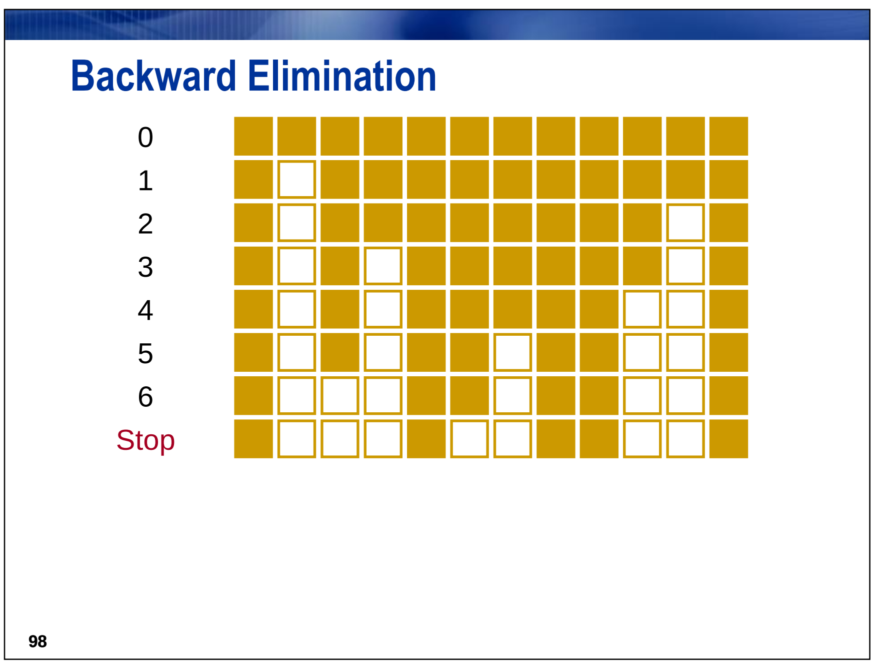
The all-possible regression technique that was discussed can be computer intensive, especially if there are a large number of potential independent variables.

The Linear Regression task also offers the following model selection options:

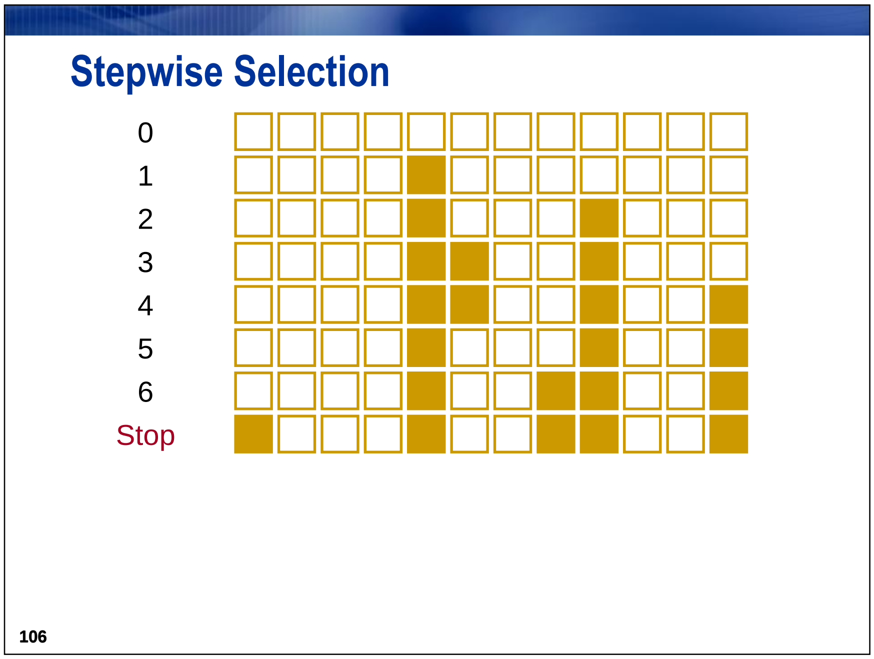
Forward selection	first selects the best one-variable model. Then it selects the best two variables among those that contain the first selected variable. Forward selection continues this process, but stops when it reaches the point where no additional variables have a *p*-value below some threshold (by default 0.50).
Backward elimination	starts with the full model. Next, the variable that is least significant, given the other variables, is removed from the model. Backward elimination continues this process until all of the remaining variables have a *p*-value below some threshold (by default 0.10).
Stepwise selection	works like a combination of the two previous methods. The default *p*-value threshold for entry is 0.15 and the default *p*-value threshold for removal is also 0.15.



Forward selection starts with an empty model. The method computes an *F* statistic for each predictor variable not in the model and examines the largest of these statistics. If it is significant at a specified significance level, the corresponding variable is added to the model. After a variable is entered in the model, it is never removed from the model. The process is repeated until none of the remaining variables meet the specified level for entry.



Backward elimination starts off with the full model. Results of the *F* test for individual parameter estimates are examined, and the least significant variable that falls above the specified significance level is removed. After a variable is removed from the model, it remains excluded. The process is repeated until no other variable in the model meets the specified significance level for removal.



Stepwise selection is similar to forward selection in that it starts with an empty model and incrementally builds a model one variable at a time. However, the method differs from forward selection in that variables already in the model do not necessarily remain. The backward component of the method removes variables from the model that do not meet the significance specified selection criterion. The stepwise selection process terminates if no further variable can be added to the model or if the variable just entered into the model is the only variable removed in the subsequent backward elimination.

Stepwise selection (forward, backward, and stepwise) has some serious shortcomings and is not the final answer. Simulation studies (Derksen and Keselman 1992) evaluating variable selection techniques found the following – collinearity (correlation among explanatory variables) and entry of noise variables.

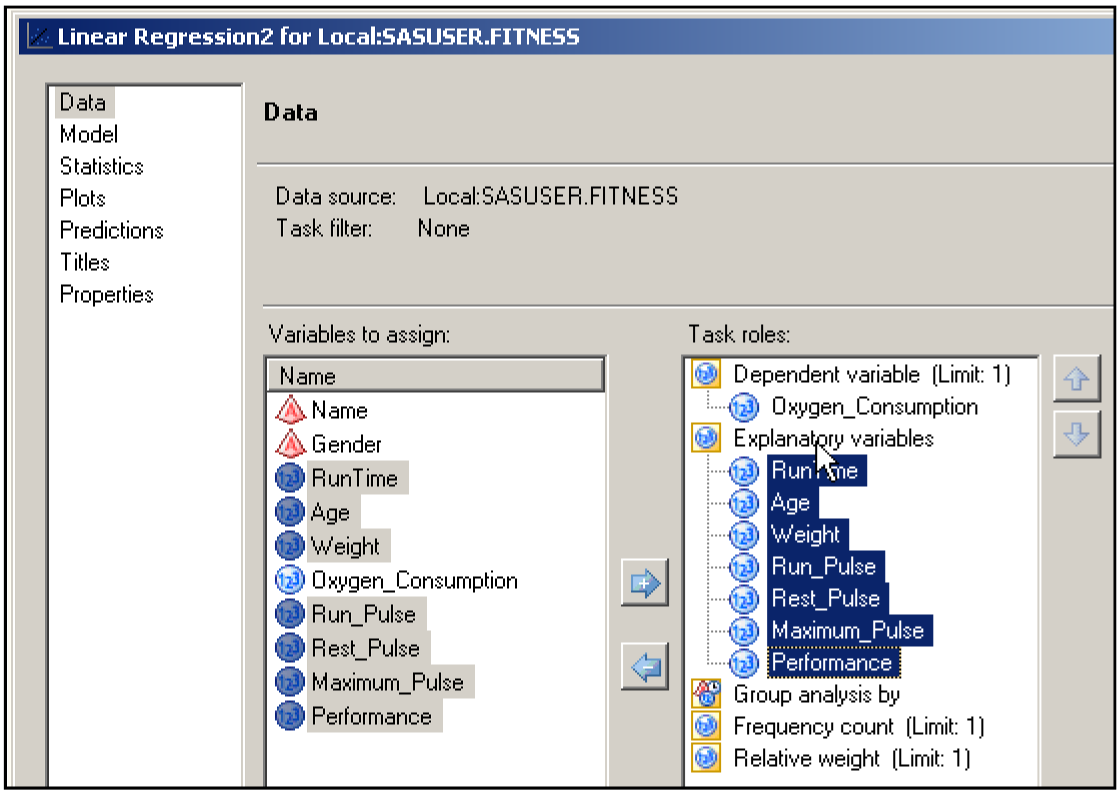
One recommendation is to use the variable selection methods to create several candidate models, and then use subject-matter knowledge to select the variables that result in the best model within the scientific or business context of the problem. Therefore, you are simply using these methods as a useful tool in the model-building process (Hosmer and Lemeshow 2000).

Stepwise Regression

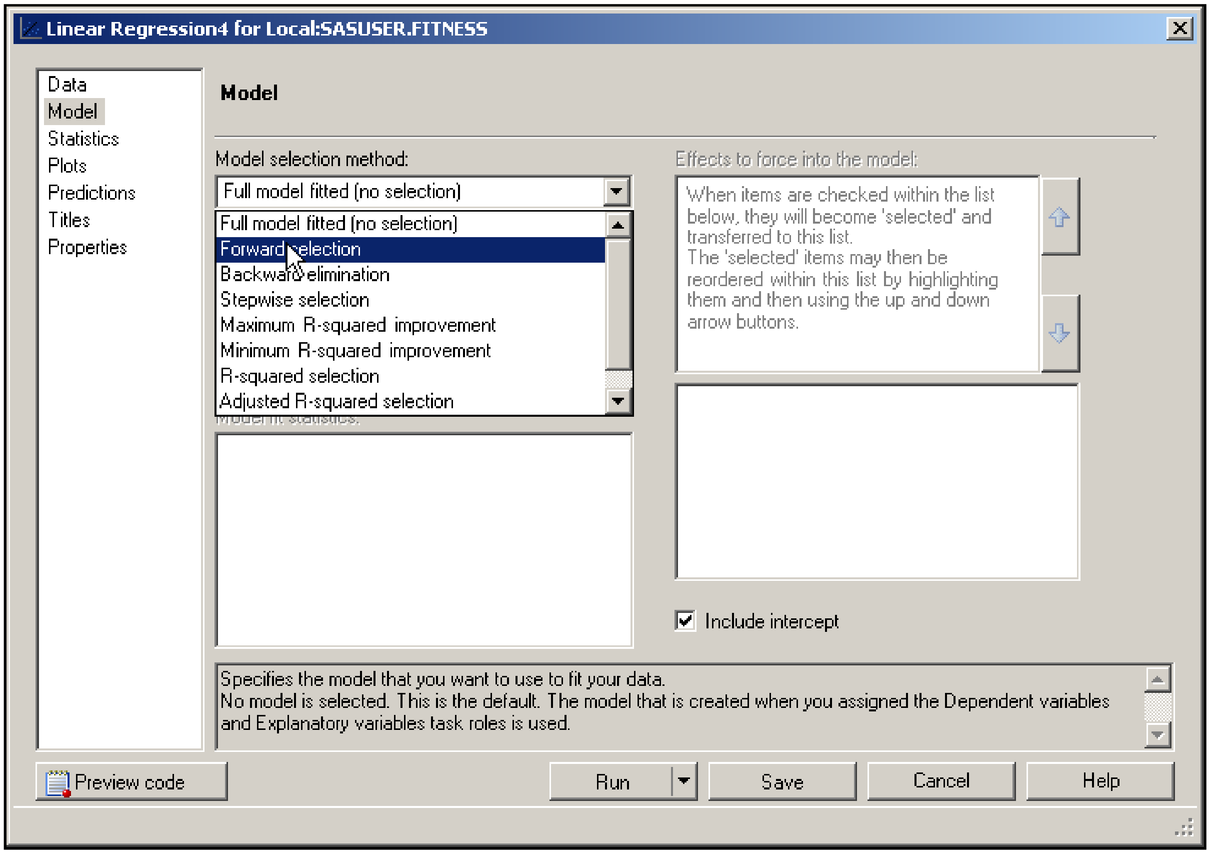
Select a model for predicting **`Oxygen_Consumption`** in the **`Fitness`** data set by using the forward, backward and stepwise methods.

Let's Begin with Forward Selection

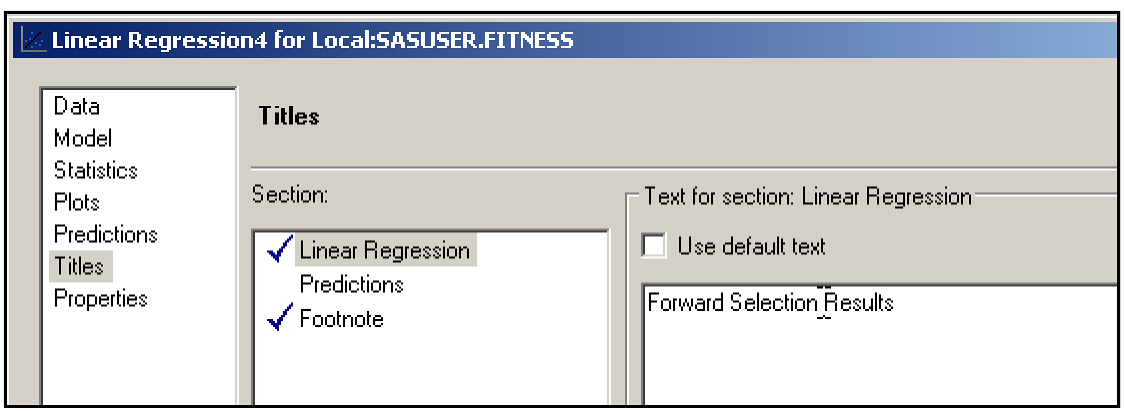
1. With the **Fitness** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression….**
2. Drag **`Oxygen_Consumption`** to the dependent variable task role and all other numeric variables to the explanatory variables task role.



3. With **Model** selected at the left, find the pull-down menu for `Model selection method` and click ▼ to find **Forward selection** at the bottom.



4. With **Titles** selected at the left, deselect the box for **Use default text** and then type **`Forward Selection Results`** in the text area.



5. Click `Run`.

Forward Selection Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

Forward Selection: Step 1

Variable RunTime Entered: R-Square = 0.7434 and C(p) = 11.9967

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	633.01458	633.01458	84.00	<.0001
Error	29	218.53997	7.53586		
Corrected Total	30	851.55455			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	82.42494	3.85582	3443.63138	456.97	<.0001
RunTime	-3.31085	0.36124	633.01458	84.00	<.0001

After the first step, one variable, **`RunTime`**, is in the model. If there are any variables that contribute significantly (*p*-value < 0.50, when controlling for **`RunTime`**) then the variable with the smallest *p*-value will be added to the model at the next step.

Forward Selection: Step 2

Variable Age Entered: R-Square = 0.7647 and C(p) = 10.7530

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	651.19281	325.59640	45.50	<.0001
Error	28	200.36175	7.15578		
Corrected Total	30	851.55455			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	88.43358	5.32255	1975.38438	276.05	<.0001
RunTime	-3.19917	0.35892	568.50196	79.45	<.0001
Age	-0.15082	0.09463	18.17822	2.54	0.1222

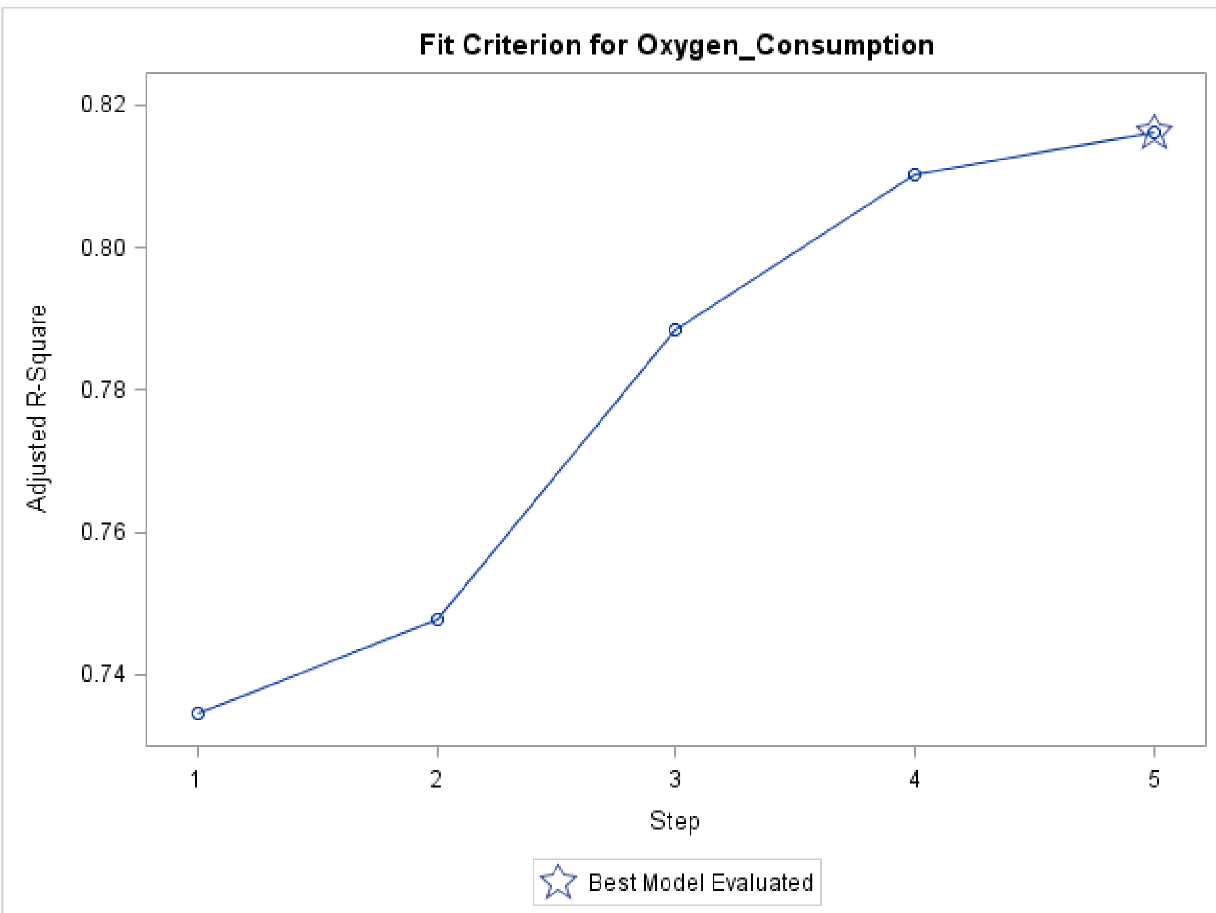
At step 2, **`Age`** is added to the model. The *p*-value associated with **`Age`** is 0.1222, which meets the significance level requirement set in the task.

Several steps are not displayed.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RunTime	1	0.7434	0.7434	11.9967	84.00	<.0001
2	Age	2	0.0213	0.7647	10.7530	2.54	0.1222
3	Run_Pulse	3	0.0449	0.8096	5.9367	6.36	0.0179
4	Maximum_Pulse	4	0.0259	0.8355	4.0004	4.09	0.0534
5	Weight	5	0.0115	0.8469	4.2598	1.87	0.1836

The model selected at each step is printed and a summary of the sequence of steps is given at the end of the output. In the summary, the variables are listed in the order in which they were selected. The partial R^2 shows the increase in the model R^2 as each term was added.

The model selected has the same variables as the model chosen using Mallows' Cp selection with the Hocking criterion. This will not always be the case.

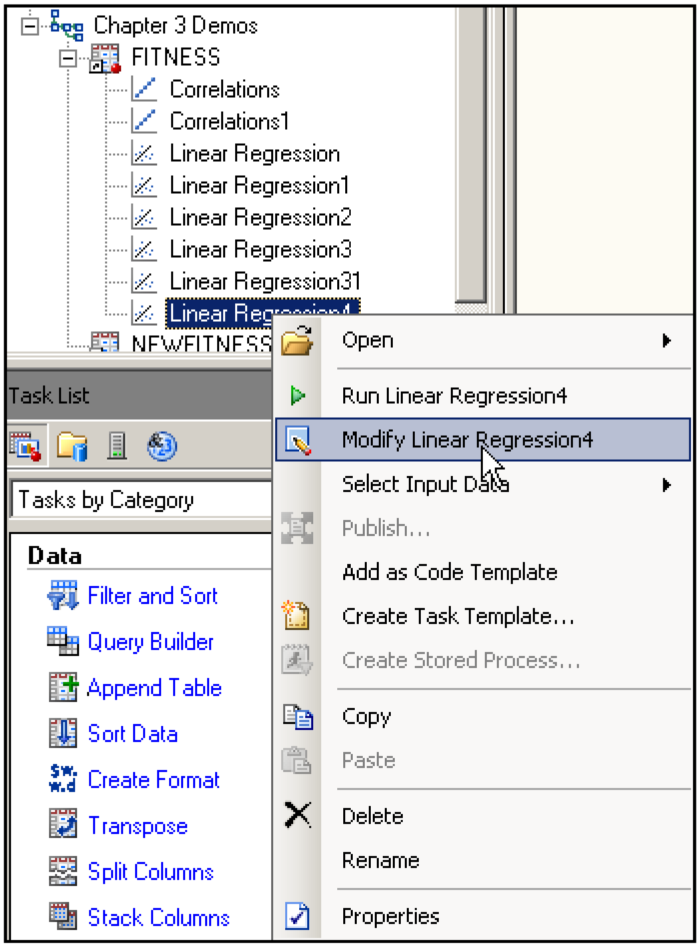


The Adjusted R-Square plot shows the progression of that statistic at each step. The star denotes the best model of the 5 tested. This is not necessarily the highest Adjusted R-Square value of all possible subsets, but is the best of the five tested in the forward selection model.

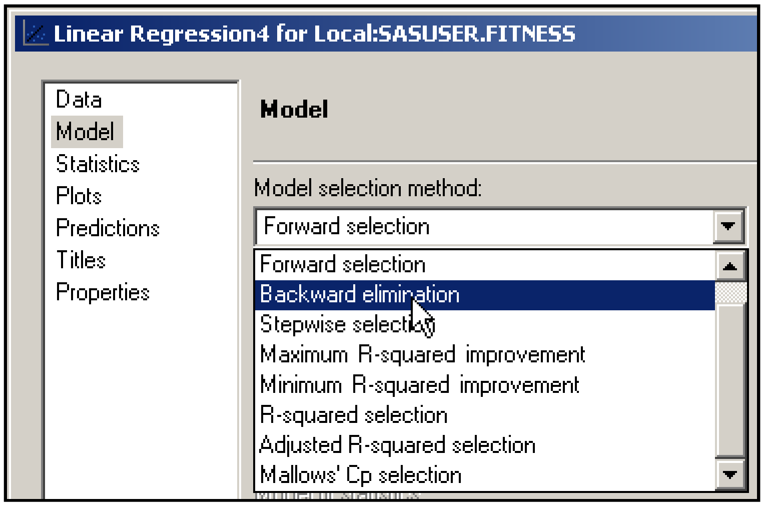
Backward Elimination

Next, rerun the task using **backward elimination**.

1. Reopen the previous task by right clicking the icon in the Project Tree and selecting **Modify Linear Regression4** from the drop-down menu.



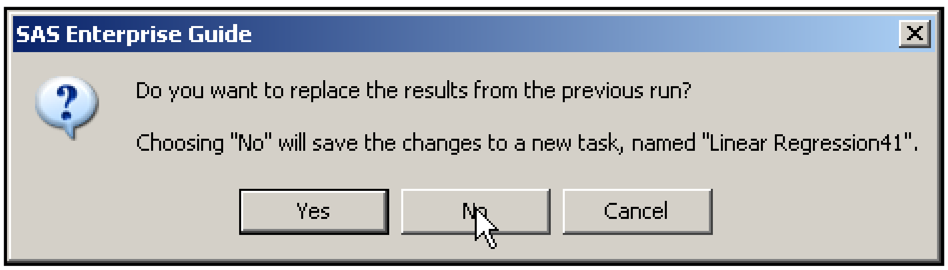
2. With **Model** selected, change the model selection method in the drop-down menu to **Backward elimination**.



3. Change the title to **`Backward Elimination Results`** in the text area.

4. Click Run.

5. Do not replace the results of the previous run.



Partial Output

Backward Elimination Results

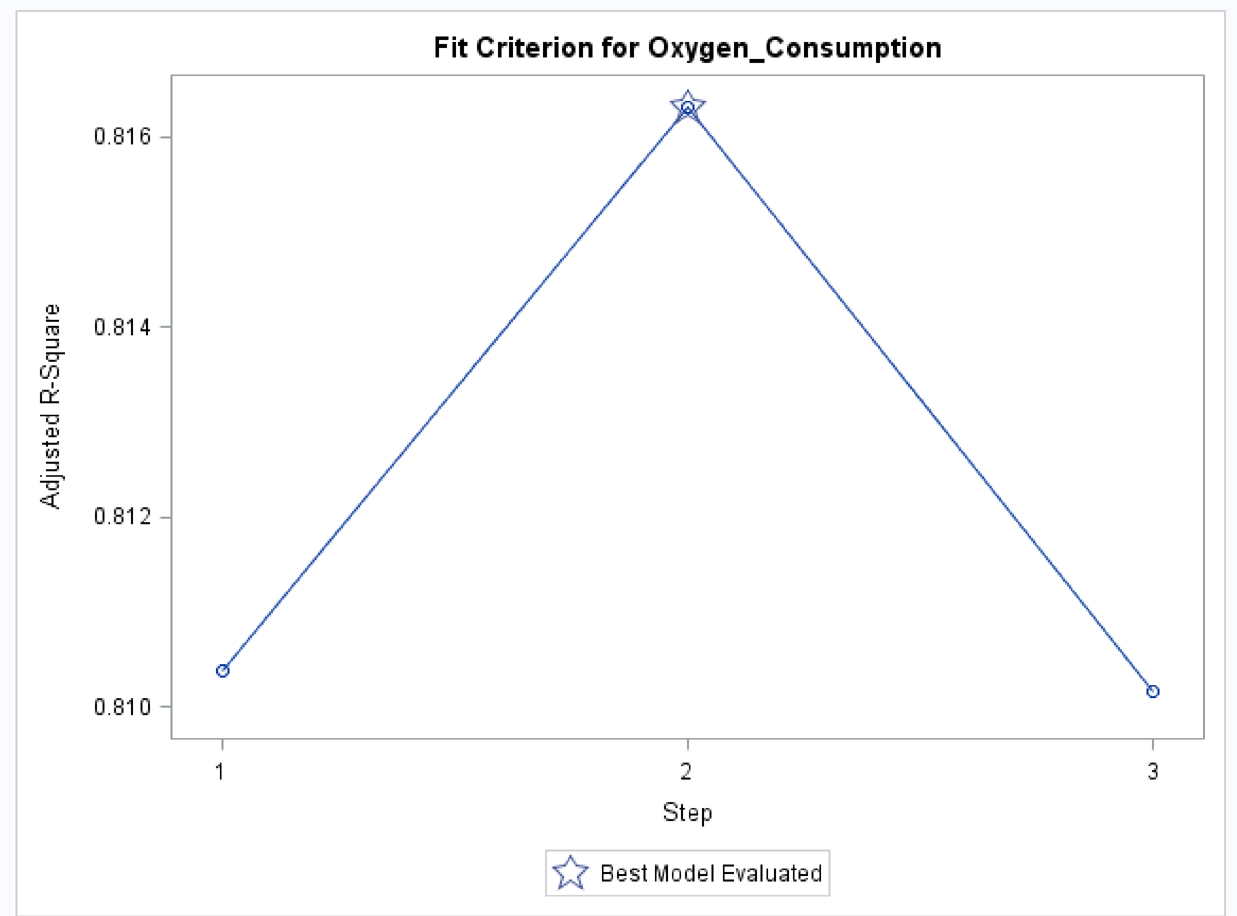
The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	97.16952	11.65703	374.42127	69.48	<.0001
RunTime	-2.77576	0.34159	355.82682	66.03	<.0001
Age	-0.18903	0.09439	21.61272	4.01	0.0557
Run_Pulse	-0.34568	0.11820	46.08558	8.55	0.0071
Maximum_Pulse	0.27188	0.13438	22.05933	4.09	0.0534

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Rest_Pulse	6	0.0003	0.8483	6.0492	0.05	0.8264
2	Performance	5	0.0014	0.8469	4.2598	0.22	0.6438
3	Weight	4	0.0115	0.8355	4.0004	1.87	0.1836

Using the backward elimination option and the default *p*-value criterion for staying in the model, three independent variables were eliminated. By coincidence the final model is the same as the one considered best base on C_p, using the Mallows criterion.



The Adjusted R-Square for the model at step 2 (before **`Weight`** was removed) was greatest of the three tested. Note the scale of the Y-axis for Adjusted R-Square. The differences in value among the three values is minimal. A [0-1] scale for the access would have shown how small the differences truly are.

Finally, run the Stepwise selection model.

1. Reopen the previous task by right clicking the icon in the Project Tree and selecting **Modify…** from the drop-down menu.
2. With **Model** selected, change the model selection method in the drop-down menu to **Stepwise selection**.
3. Change the title to **`Stepwise Selection Results`** in the text area.
4. Click Run.
5. Do not replace the results of the previous run.

Partial Output

Stepwise Selection Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

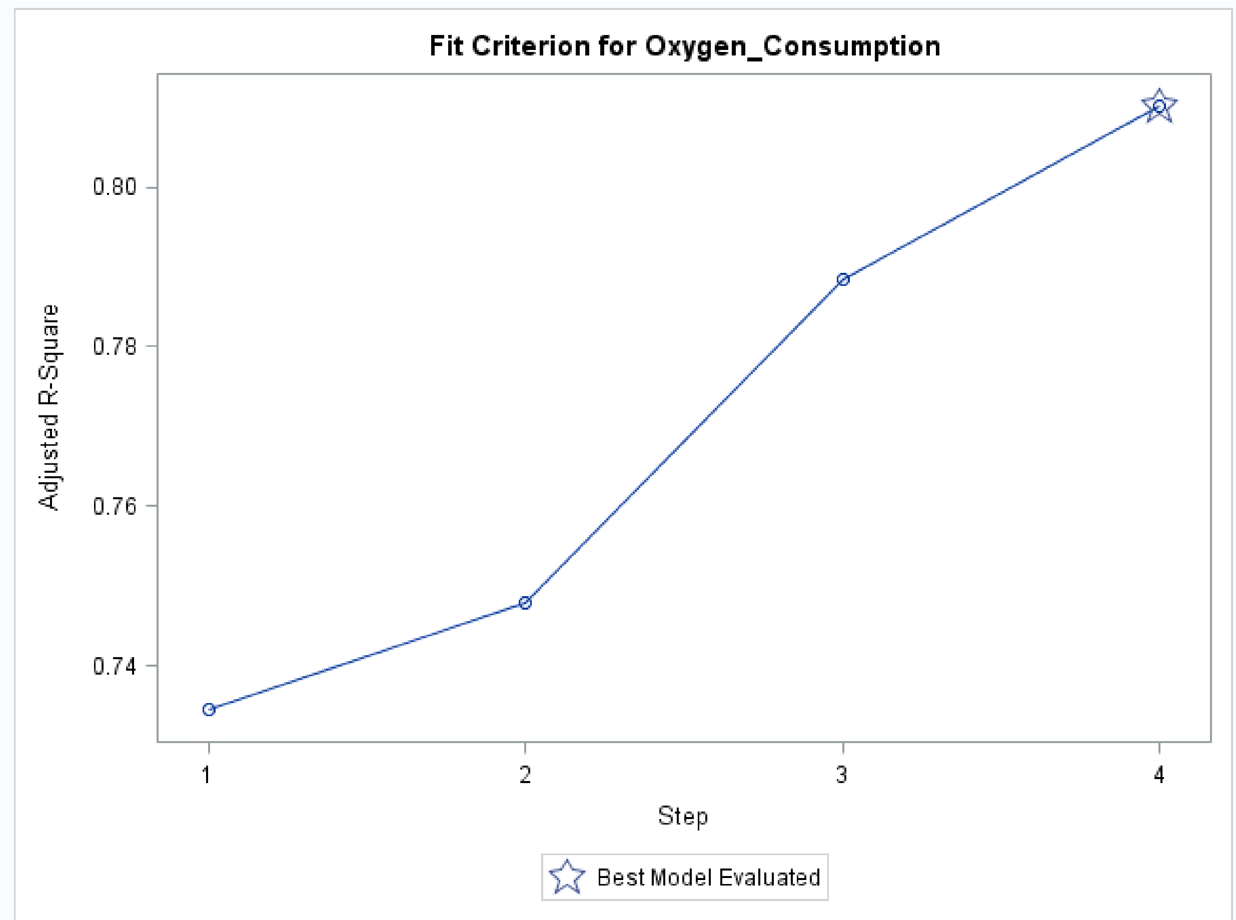
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	711.45087	177.86272	33.01	<.0001
Error	26	140.10368	5.38860		
Corrected Total	30	851.55455			

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RunTime		1	0.7434	0.7434	11.9967	84.00	<.0001
2	Age		2	0.0213	0.7647	10.7530	2.54	0.1222
3	Run_Pulse		3	0.0449	0.8096	5.9367	6.36	0.0179
4	Maximum_Pulse		4	0.0259	0.8355	4.0004	4.09	0.0534

Using stepwise selection and the default *p*-value, the same subset resulted as that using backward elimination. However, it is not the same model as that resulting from forward selection.



The default entry criterion is $p<.50$ for the forward selection method and $p<.15$ for the stepwise selection method. After **RunTime** was entered into the model, **Age** was entered at step 2 with a p-value of 0.1222. If the criterion were set to something less than 0.10, the final model would have been quite different. It would have included only one variable, **RunTime**. This underscores the precariousness of relying on one stepwise method for defining a "best" model.

Comparing Forward, Backward, & Stepwise Results

Stepwise Regression Models

FORWARD	**Runtime**, **Age**, **Weight**, **Run_Pulse**, **Maximum_Pulse**
BACKWARD	**Runtime**, **Age**, **Run_Pulse**, **Maximum_Pulse**
STEPWISE	**Runtime**, **Age**, **Run_Pulse**, **Maximum_Pulse**

109

The final models obtained using the default selection criteria are displayed. It is important to note that the choice of criterion levels can greatly affect the final models that are selected using stepwise methods.

Stepwise Models, Alternative Criteria

FORWARD (slentry=0.05)	**Runtime**
BACKWARD (slstay=0.05)	**Runtime**, **Run_Pulse**, **Maximum_Pulse**
STEPWISE (slentry=0.05, slstay=0.05)	**Runtime**

110

The final models using 0.05 as the forward and backward step criteria resulted in very different models than those chosen using the default criteria.

Comparison of Selection Methods

Stepwise regression	uses fewer computer resources.
All-possible regression	generates more candidate models that might have nearly equal R^2 statistics and C_p statistics.

111

The stepwise regression methods have an advantage when there are a large number of independent variables.

With the all-possible regressions techniques, you can compare essentially equivalent models and use your knowledge of the data set and subject area to select a model that is more easily interpreted.