SASEG 9C – Model Building – Advanced

(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

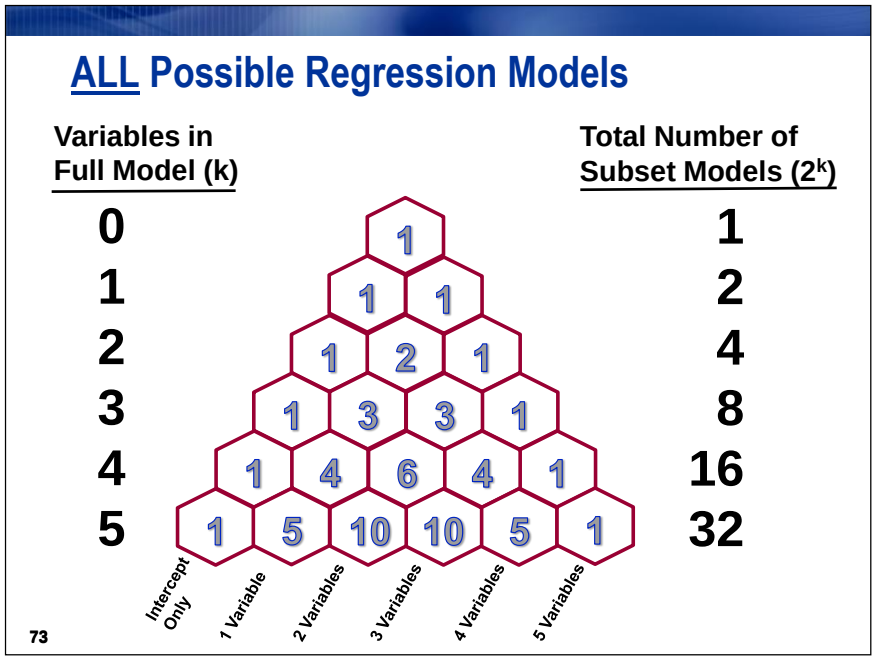
- R-squared, Adjusted R-Squared, or Mallows' Cp

Stepwise selection methods

- Stepwise, Forward, or Backward

Full model fitted (no selection) is the default.

72

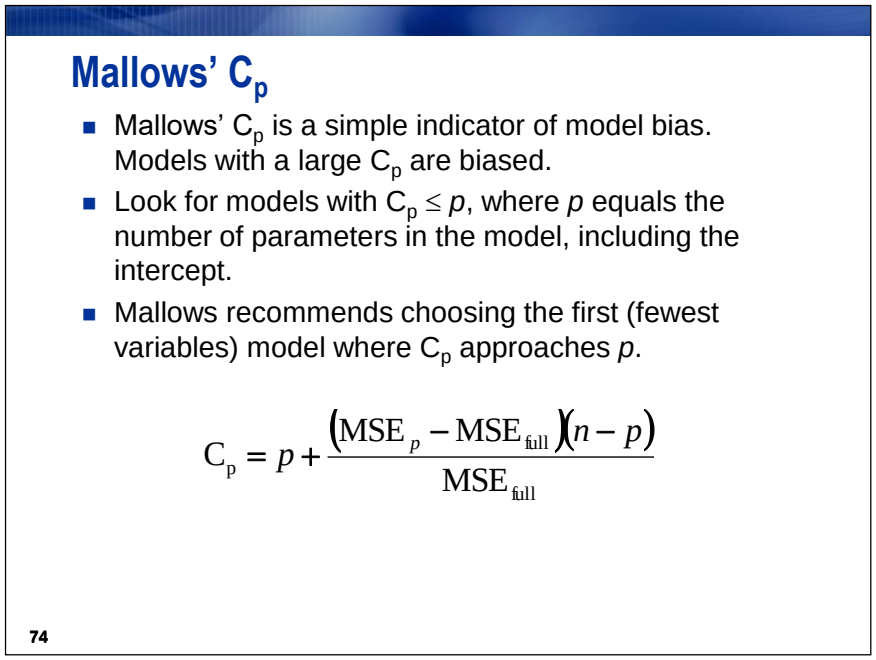


All-Possible Regression Techniques have in common that they literally assess each possible subset model of a given set of predictor variables in a regression model. The assessment is based on some overall model statistic value (such as R-Squared, Adjusted R-Square and Mallows' C_P). For a model with 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 (1973) is estimated by $C_p = p + \frac{(MSE_p - MSE_{full})(n-p)}{MSE_{full}}$

where

MSE_p	is the mean squared error for the model with p parameters.
MSE_{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:

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

75

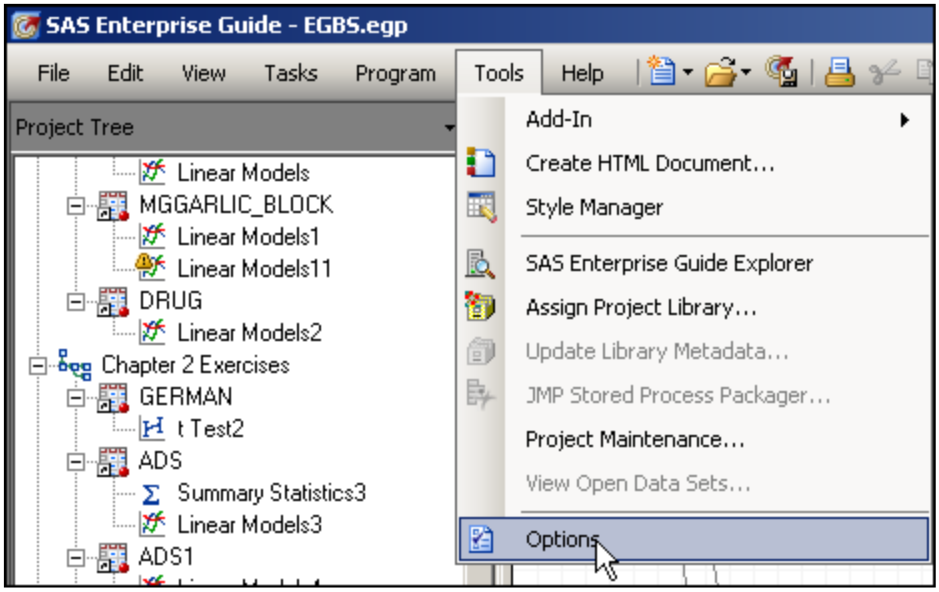
Hocking suggested the use of the C_p statistic, but with alternative criteria, depending on the purpose of the analysis. His suggestion of ($C_p \leq 2p - p_{full} + 1$) is included in the REG procedure's calculations of criteria reference plots for best models.

Automatic Model Selection

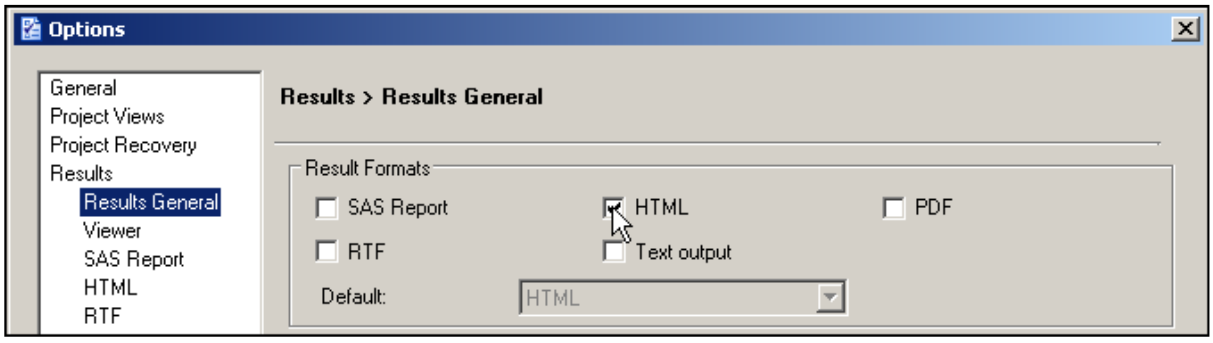
Invoke the Linear Regression task to produce a regression of **Oxygen_Consumption** on all the other variables in the **Fitness** data set and produce plots with tool (data) tips to aid in exploration of the results.

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**.



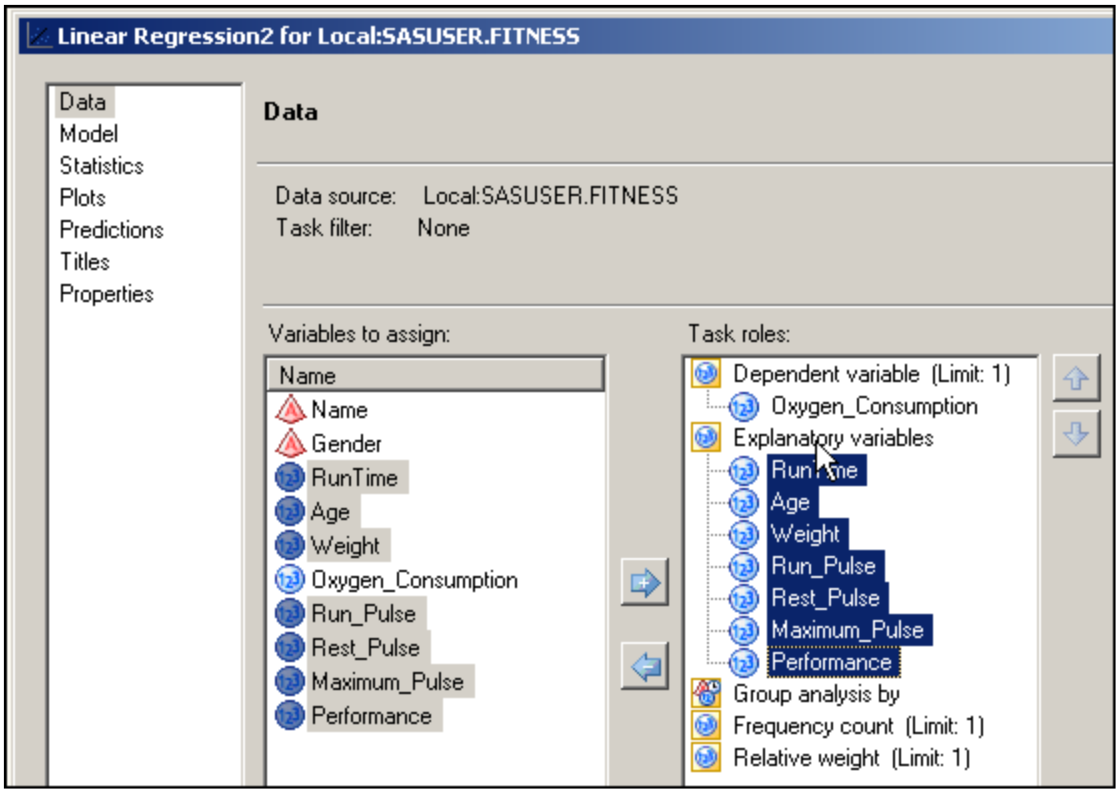
2. In the window that opens, select **Results General** under `Results` at the left and then uncheck the box for **SAS Report** and check the box for **HTML**.



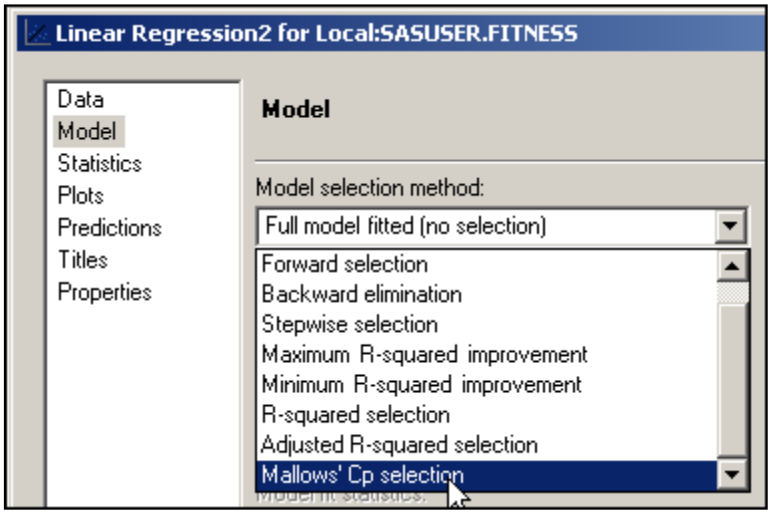
3. Click Apply and then click OK.

Now you are ready to run the Linear Regression task.

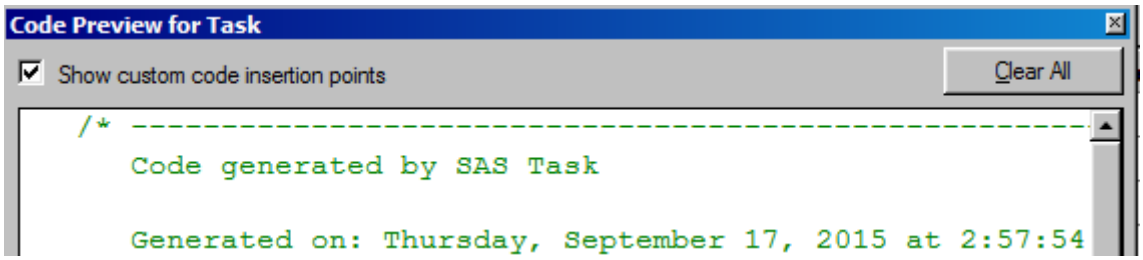
4. With the **Fitness** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
5. Drag **Oxygen_Consumption** to the dependent variable task role and all other numeric variables to the explanatory variables task role.



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



7. Click Preview code.



8. Enable the Show custom code insertion points box
9. Type **`ODS GRAPHICS / IMAGEMAP=ON;`** under the ODS GRAPHICS ON; statement in the <insert custom code here> area

```
ODS GRAPHICS ON;


/* Start of custom user code (Framework_BeforeTaskCode)
ODS GRAPHICS / IMAGEMAP=ON;
/* End of custom user code (Framework_BeforeTaskCode) */
```

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

Partial HTML Output

Linear Regression Results

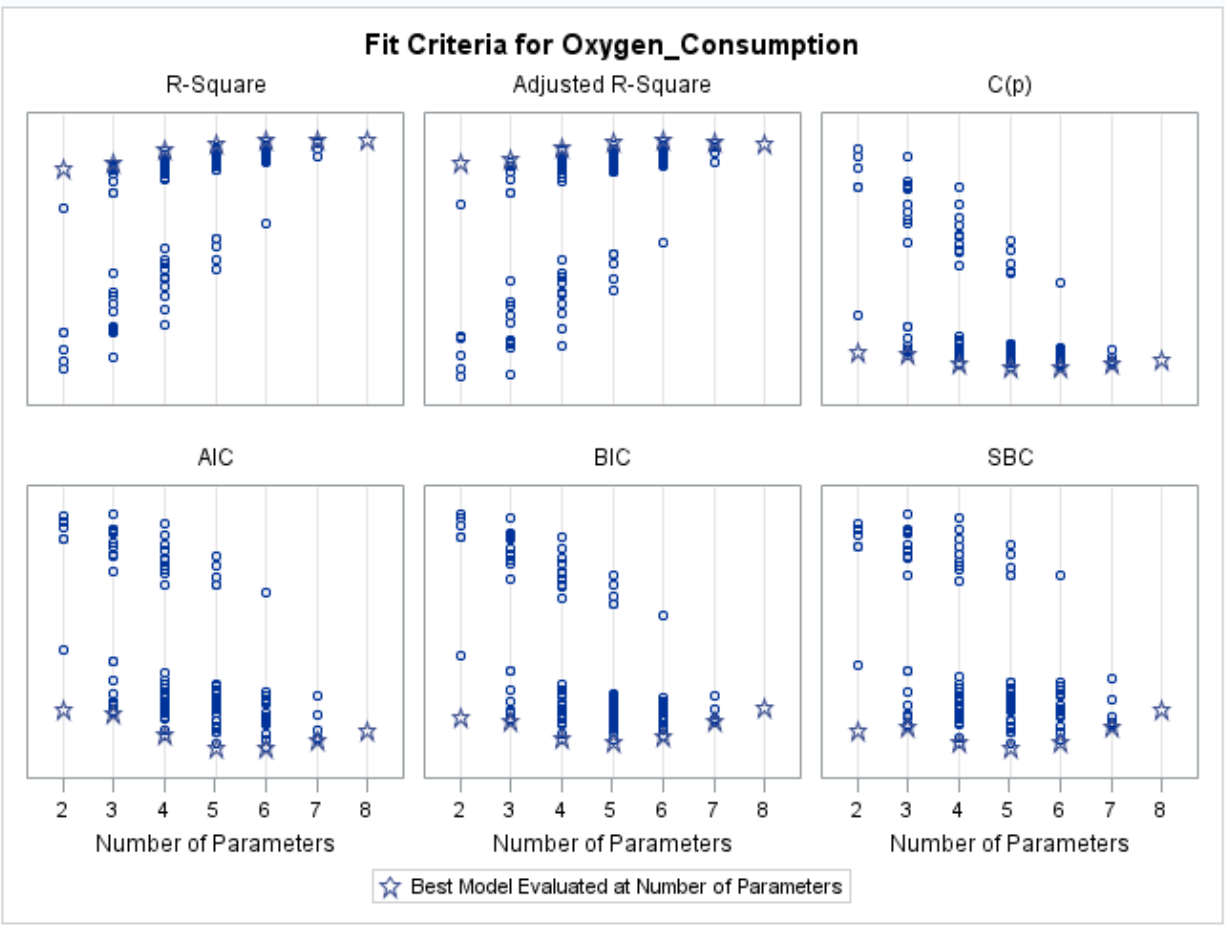
The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

C(p) Selection Method

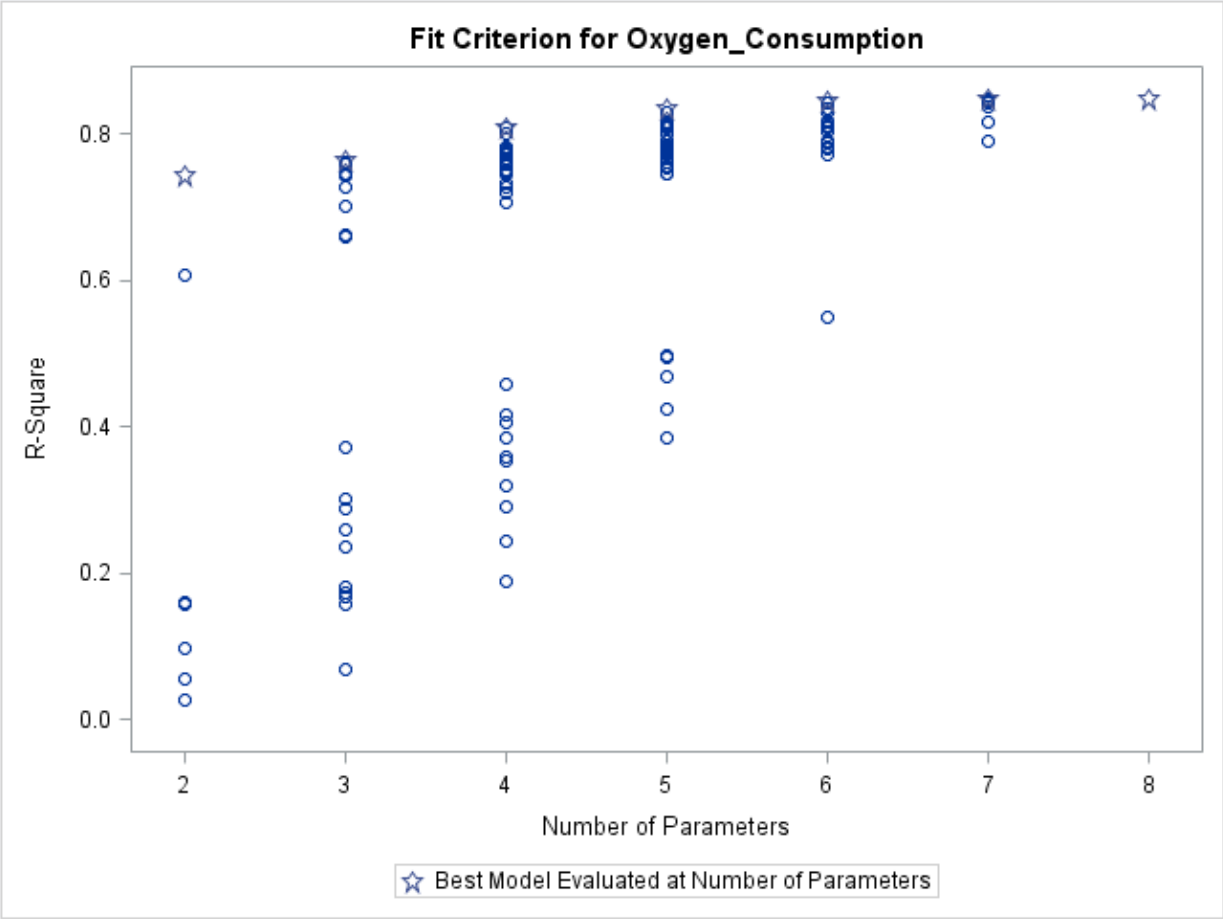
Number of Observations Read	31
Number of Observations Used	31

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.

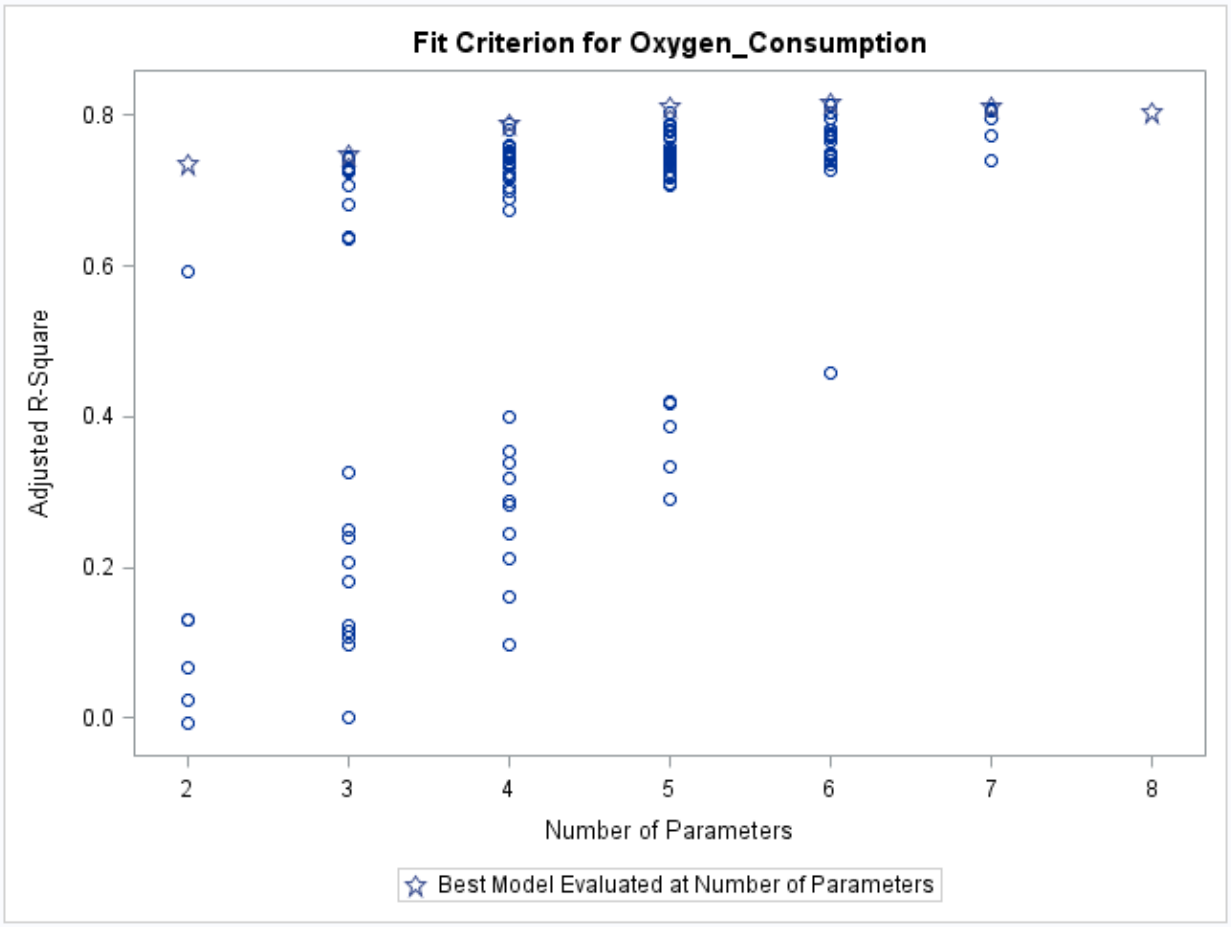


The first plot is a panel plot of several plots assessing each of the 127 possible subset models. Three of them will be further described below.

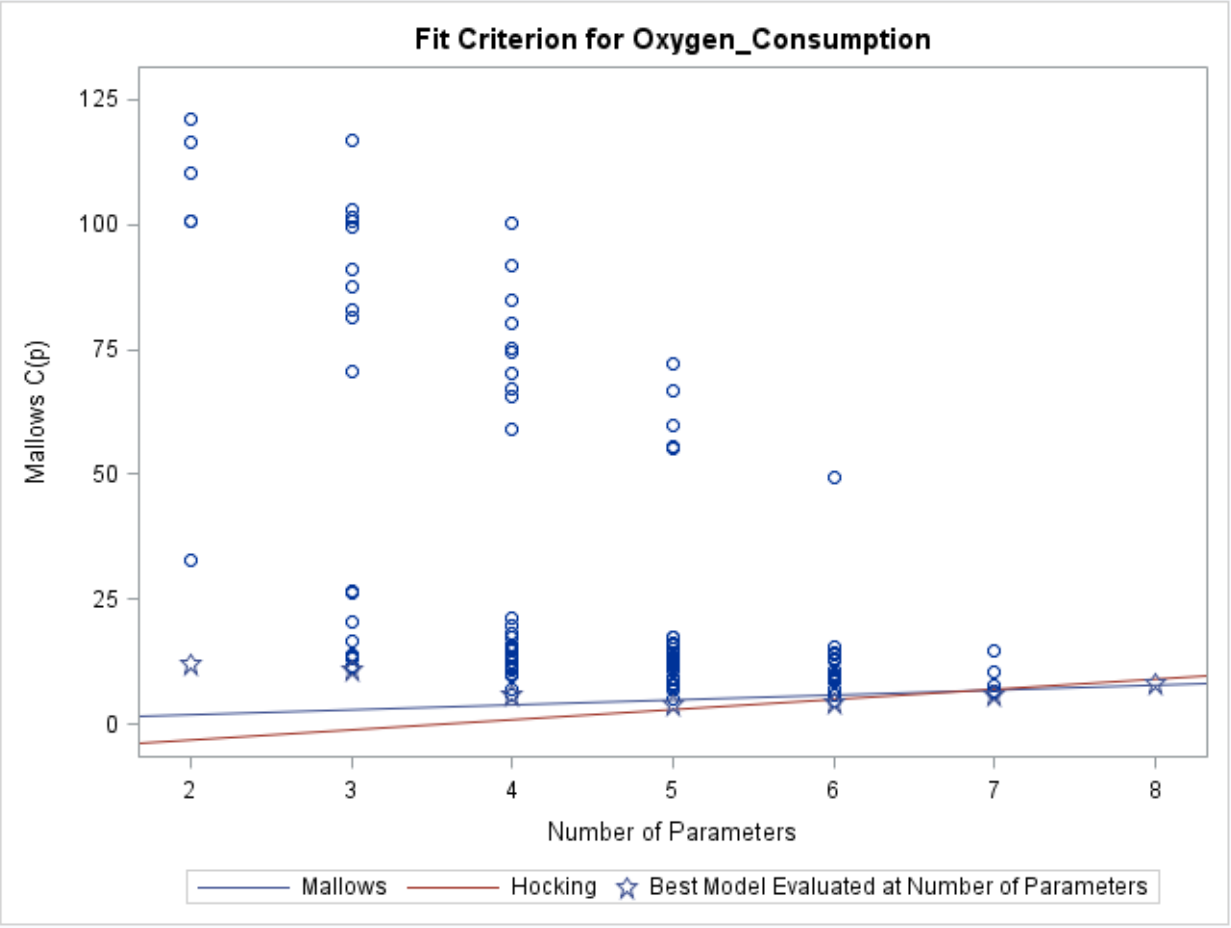

Fit Criterion for Oxygen_Consumption
R-Square
0.0
0.2
0.4
0.6
0.8
2
3
4
5
6
7
8
Number of Parameters
Best Model Evaluated at Number of Parameters

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 $C_p \leq p$ for prediction. The lower line is plotted to help identify which models satisfy Hocking's criterion $C_p \leq 2p - p_{full} + 1$ 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:

C(p) = 4.0004
Number of Parameters = 5
Model = RunTime Age Run_Pulse Maximum_Pulse

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.

"Best" Models – Parameter Estimation

The three best candidate models for Analytic purposes, according to Hocking, include:

p = 6	C_p = 4.2598 R^2=0.8469 Adj. R^2=0.8163	RunTime, Age, Weight, Run_Pulse, Maximum_Pulse
p = 6	C_p = 4.7158 R^2=0.8439 Adj. R^2=0.8127	Performance, RunTime, Weight, Run_Pulse, Maximum_Pulse
p = 6	C_p = 4.7168 R^2=0.8439 Adj. R^2=0.8127	Performance, RunTime, Age, Run_Pulse, Maximum_Pulse

78

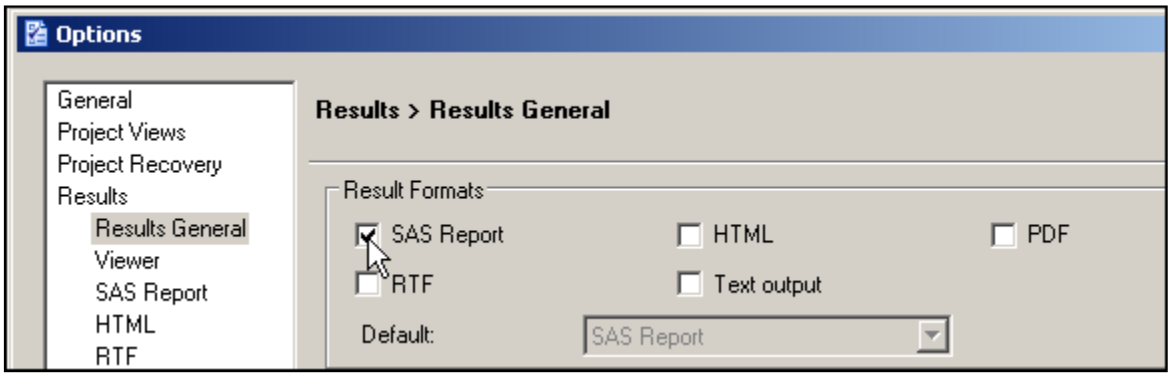
The variables **RunTime**, **Run_Pulse**, and **Maximum_Pulse** once again appear in all candidate models. The choice of models here depends on selection of pairs from **Performance**, **Age** and **Weight**. Here you again choose a model with objective measures, **Age** and **Weight**. That is the top model in the list. Your choice might differ.

Estimating and Testing the Coefficients for the Selected Models

Invoke the Linear Regression task to compare the ANOVA tables and parameter estimates for the two-candidate models in the **`Fitness`** data set.

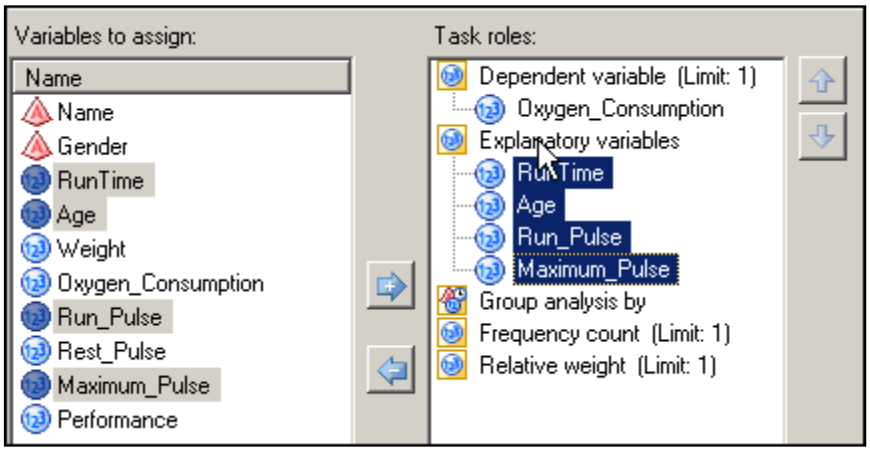
First, return reporting in SAS Enterprise Guide to SAS Report from HTML.

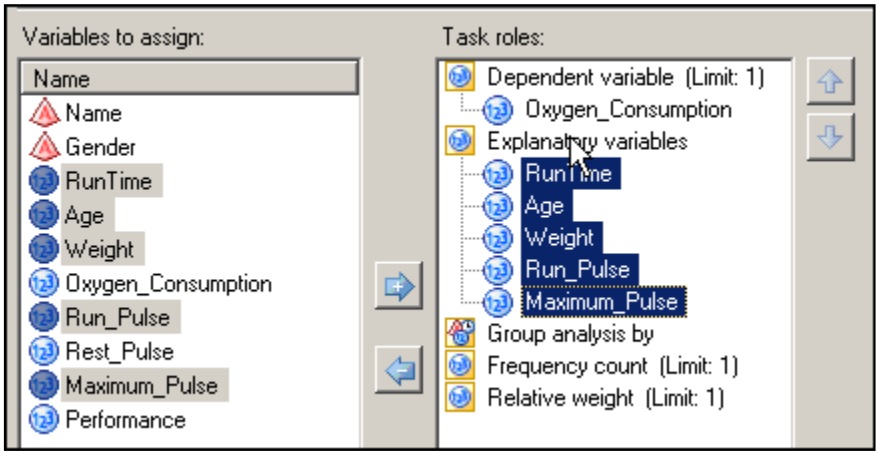
1. Select **Tools** ⇨ **Options**.
2. In the window that opens, select **Results General** under Results at the left and then uncheck the box for **HTML** and check the box for **SAS Report**.



3. Click Apply and then click OK.

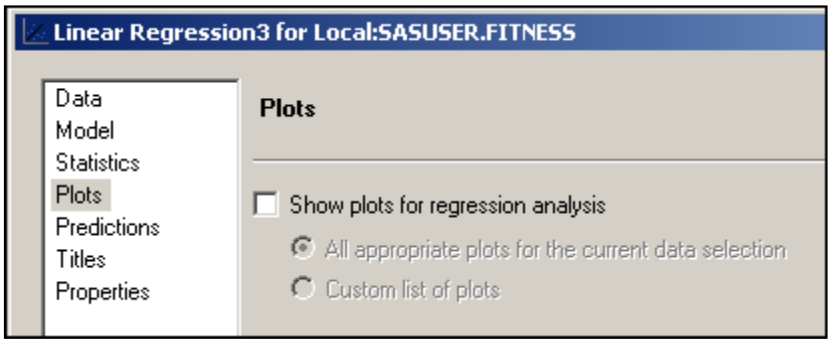
4. Run the Linear Regression task twice, once using the variables **`Runtime`**, **`Age`**, **`Run_Pulse`**, and **`Maximum_Pulse`** as the explanatory variables and once using **`Runtime`**, **`Age`**, **`Weight`**, **`Run_Pulse`**, and **`Maximum_Pulse`** as the explanatory variables.



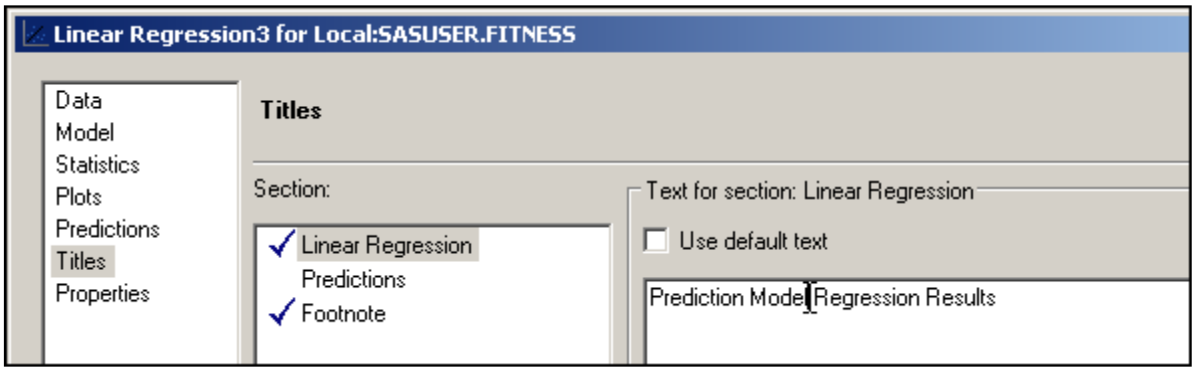


5. In each case, with **Plots** selected at the left, uncheck the box for **Show plots for regression analysis**.

You will learn more about plots in a later chapter.



6. With **Titles** selected at the left, uncheck the box for **Use default text** and then type **`Prediction Model Regression Results`** in the text area for the first model and **`Explanatory Model Regression Results`** in the text area for the second model.



7. Click Run.

Output for the Prediction Model:

Prediction Model Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

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			

Root MSE	2.32134	R-Square	0.8355
Dependent Mean	47.37581	Adj R-Sq	0.8102
Coeff Var	4.89984		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	97.16952	11.65703	8.34	<.0001
RunTime	1	-2.77576	0.34159	-8.13	<.0001
Age	1	-0.18903	0.09439	-2.00	0.0557
Run_Pulse	1	-0.34568	0.11820	-2.92	0.0071
Maximum_Pulse	1	0.27188	0.13438	2.02	0.0534

The R^2 and adjusted R^2 are the same as calculated during the model selection program. If there are missing values in the data set, however, this might not be true.

The model *F* is large and highly significant. **`Age`** and **`Maximum_Pulse`** are not significant at the 0.05 level of significance. However, all terms have *p*-values below 0.10.

The adjusted R^2 is close to the R^2, which suggests that there are not too many variables in the model.

Output for the Explanatory Model:

Prediction Model Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	721.20532	144.24106	27.66	<.0001
Error	25	130.34923	5.21397		
Corrected Total	30	851.55455			

Root MSE	2.28341	R-Square	0.8469
Dependent Mean	47.37581	Adj R-Sq	0.8163
Coeff Var	4.81978		

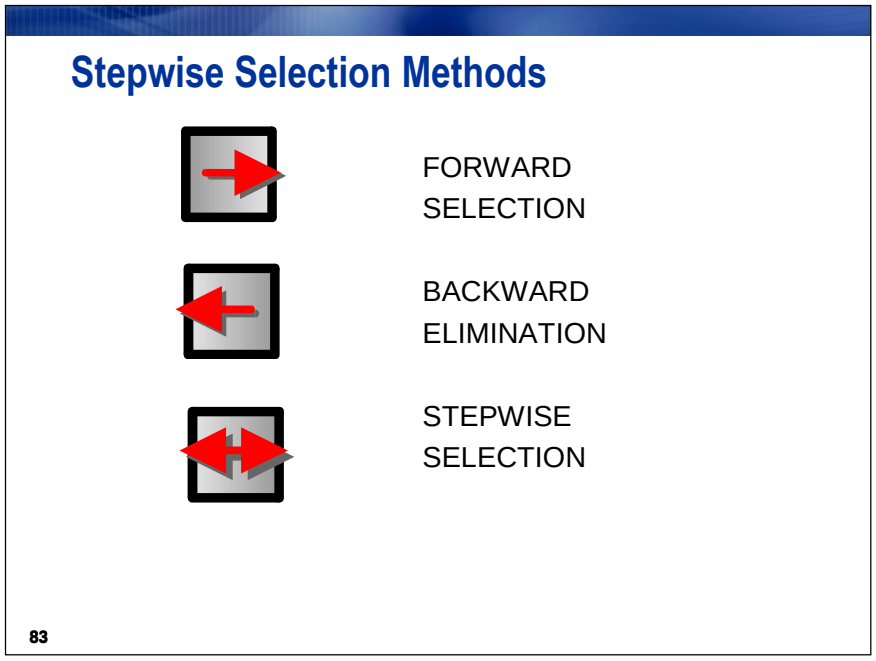
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	101.33835	11.86474	8.54	<.0001
RunTime	1	-2.68846	0.34202	-7.86	<.0001
Age	1	-0.21217	0.09437	-2.25	0.0336
Weight	1	-0.07332	0.05360	-1.37	0.1836
Run_Pulse	1	-0.37071	0.11770	-3.15	0.0042
Maximum_Pulse	1	0.30603	0.13452	2.28	0.0317

The adjusted R^2 is slightly larger than in the Prediction model and very close to the R^2.

The model *F* is large, but smaller than in the Prediction model. However, it is still highly significant. All terms included in the model are significant except **`Weight`**. Note that the *p*-values for **`Age`**, **`Run_Pulse`**, and **`Maximum_Pulse`** are smaller in this model than they were in the Prediction model.

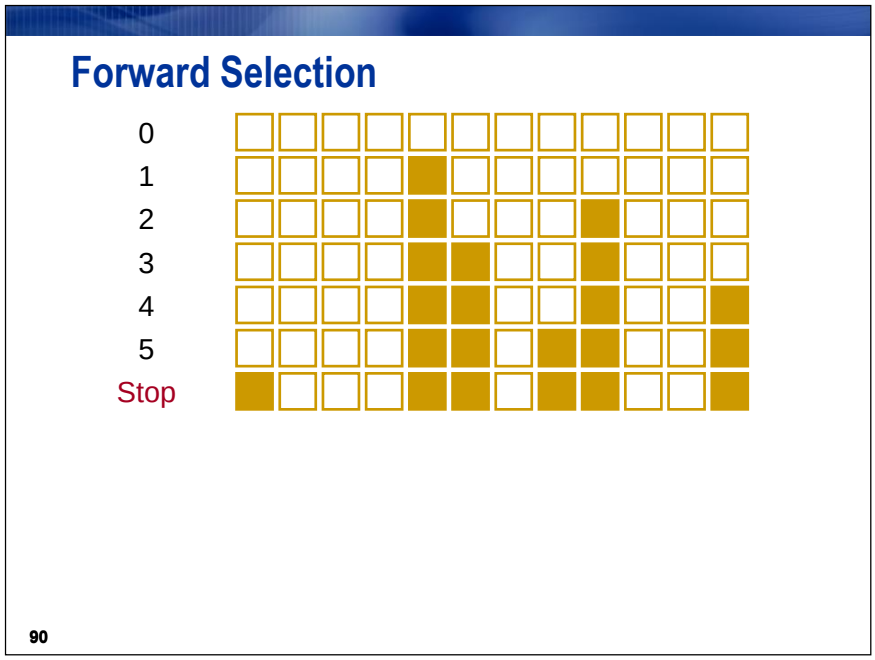
Including the additional variable in the model changes the coefficients of the other terms and changes the *t* Values for all.



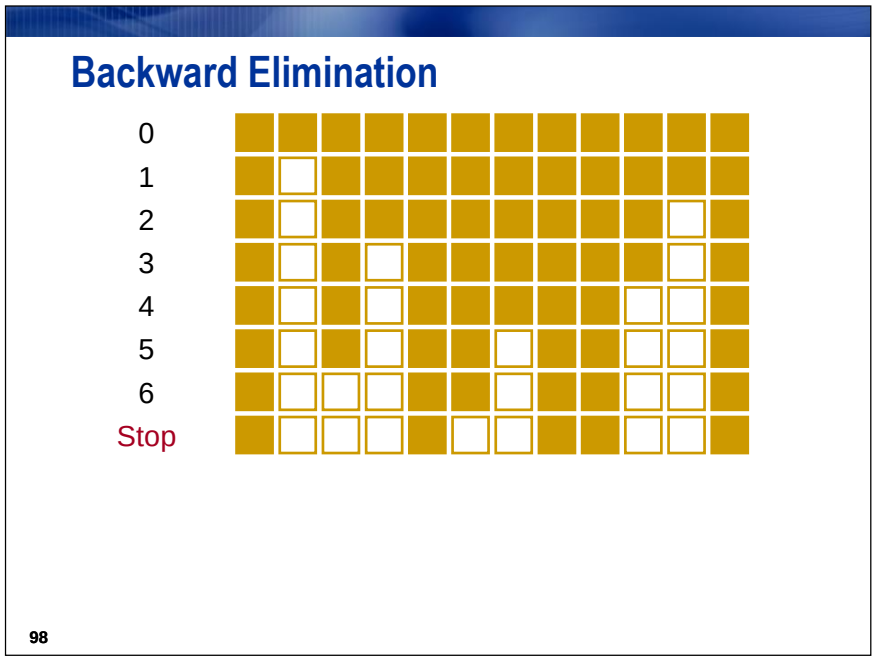
The all-possible regression technique that was discussed can be computer intensive, especially if there are a large number of potential independent variables.

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

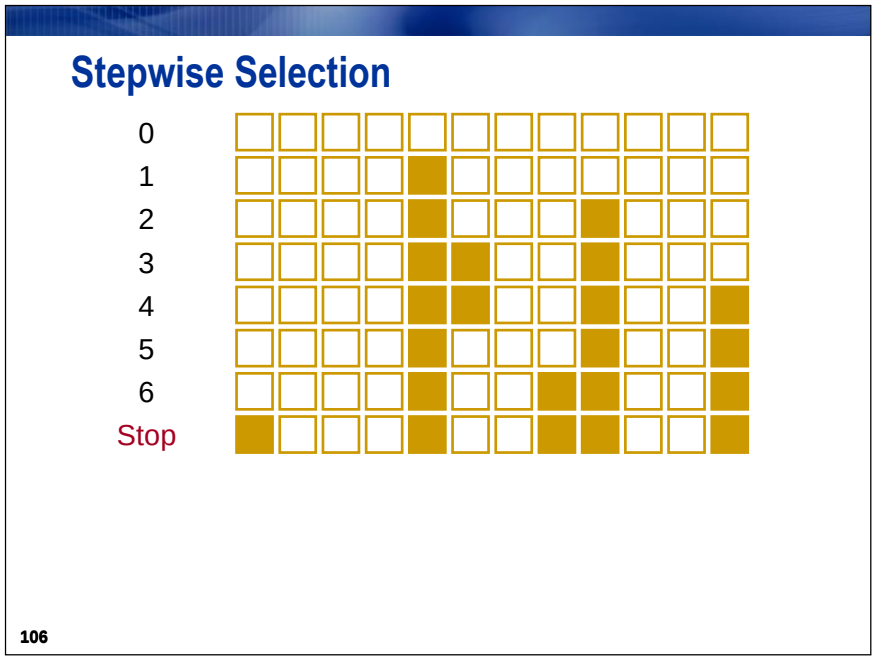
Forward selection	first selects the best one-variable model. Then it selects the best two variables among those that contain the first selected variable. Forward selection continues this process, but stops when it reaches the point where no additional variables have a *p*-value below some threshold (by 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.



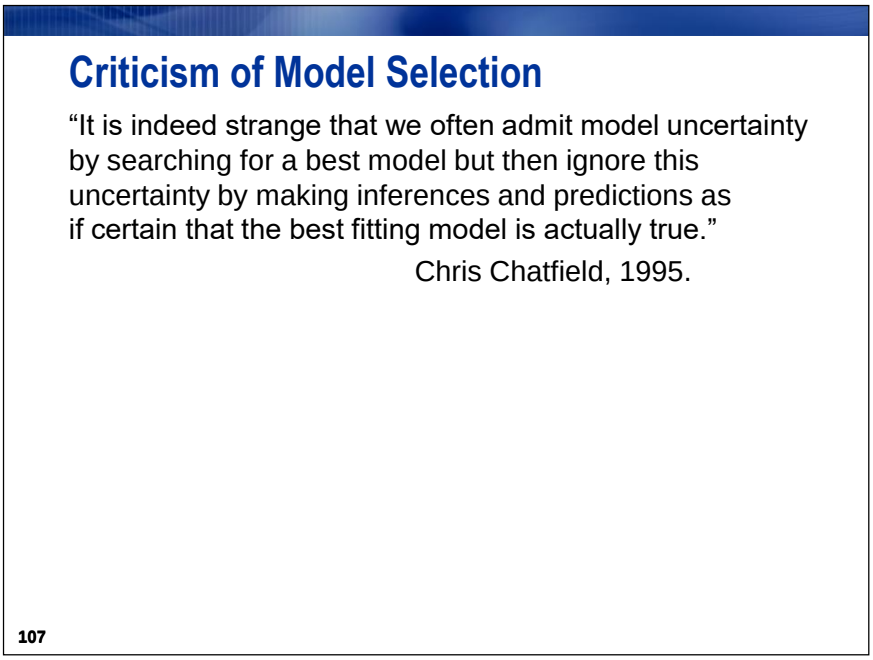
Backward elimination starts off with the full model. Results of the *F* test for individual parameter estimates are examined, and the least significant variable that falls above the specified significance level is removed. After a variable is removed from the model, it remains excluded. The process is 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).



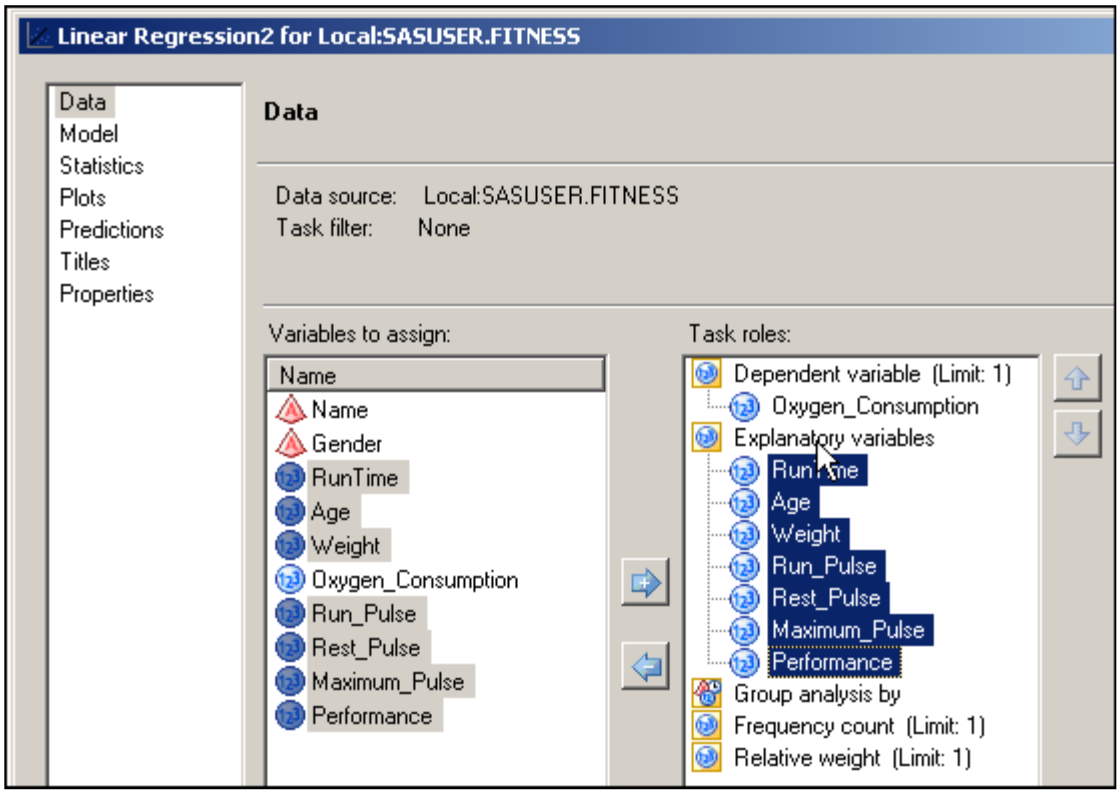
Statisticians give warnings and cautions about the over-interpretation of *p*-values from models chosen using any automated variable selection technique. Refitting many submodels in terms of an optimum fit to the data distorts the significance levels of conventional statistical tests. However, many researchers and users of statistical software neglect to report that the models they selected were chosen using automated methods. They report statistical quantities such as standard errors, confidence limits, *p*-values, and R-squared as if the resulting model were entirely pre-specified. These inferences are inaccurate, tending to err on the side of overstating the significance of predictors and making predictions with overly optimistic confidence. This problem is very evident when there are many iterative stages in model building. When there are many variables and you use stepwise selection to find a small subset of variables, inferences become less accurate (Chatfield 1995, Raftery 1994, Freedman 1983).

One solution to this problem is to split your data. One part could be used for finding the regression model and the other part could be used for inference. Another solution is to use bootstrapping methods to obtain the correct standard errors and *p*-values. Bootstrapping is a resampling method that tries to approximate the distribution of the parameter estimates to estimate the standard error. Unfortunately, bootstrapping is not part of the Linear Regression task and the computer programming is beyond the scope of this course.

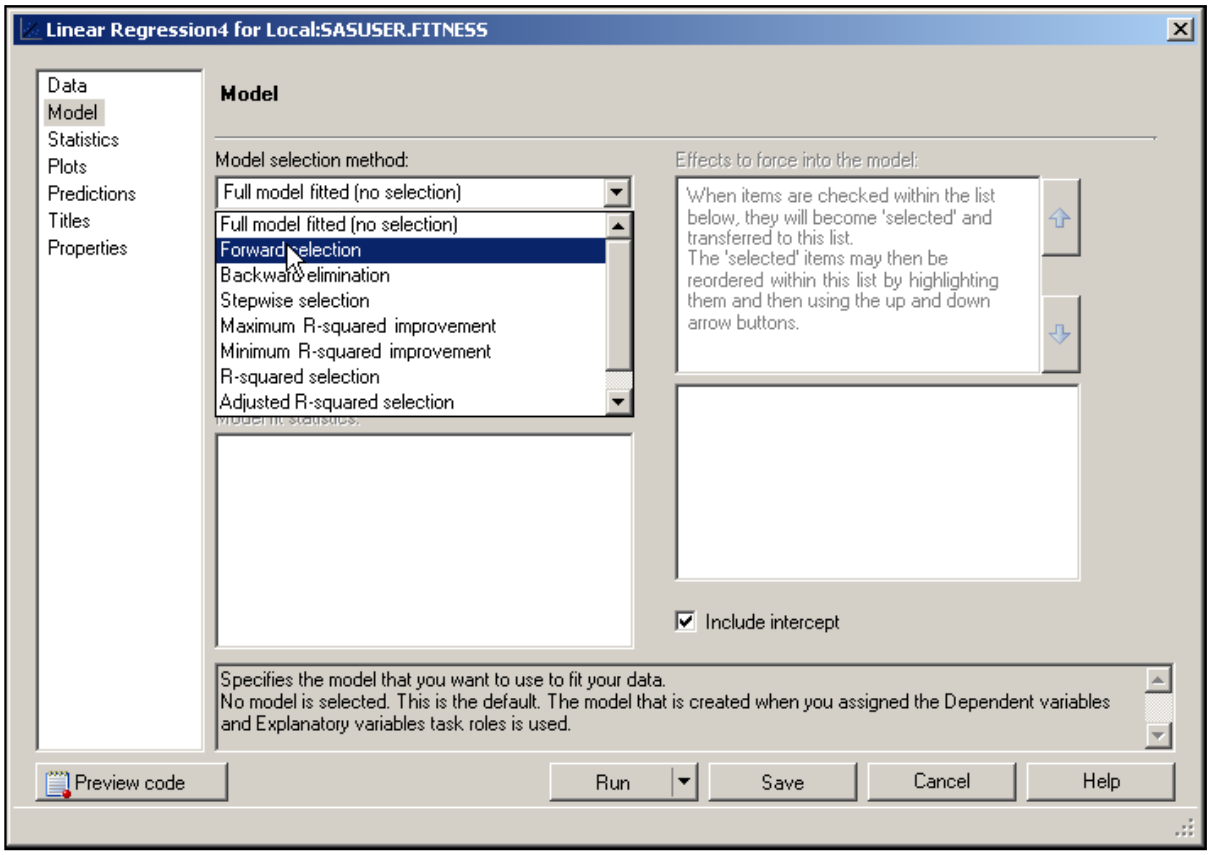
Forward – Stepwise Regression

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

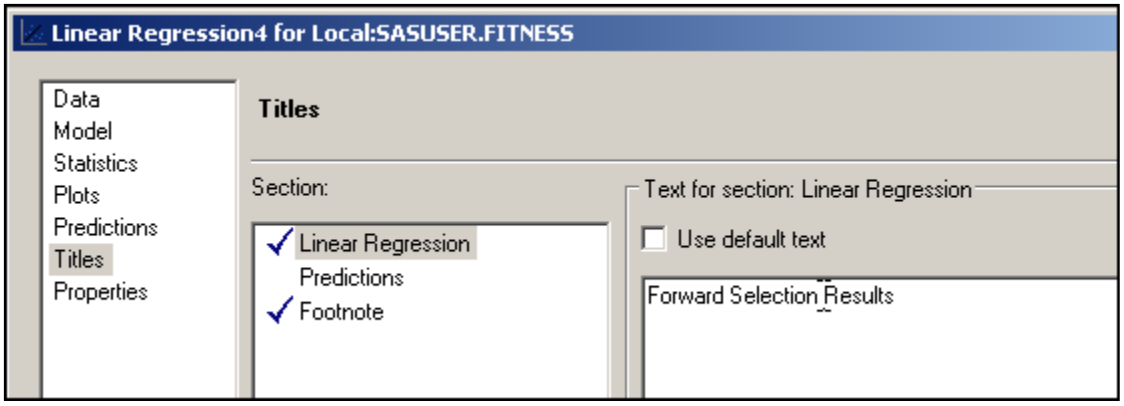
1. With the **Fitness** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
2. Drag **`Oxygen_Consumption`** to the dependent variable task role and all other numeric variables to the explanatory variables task role.



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

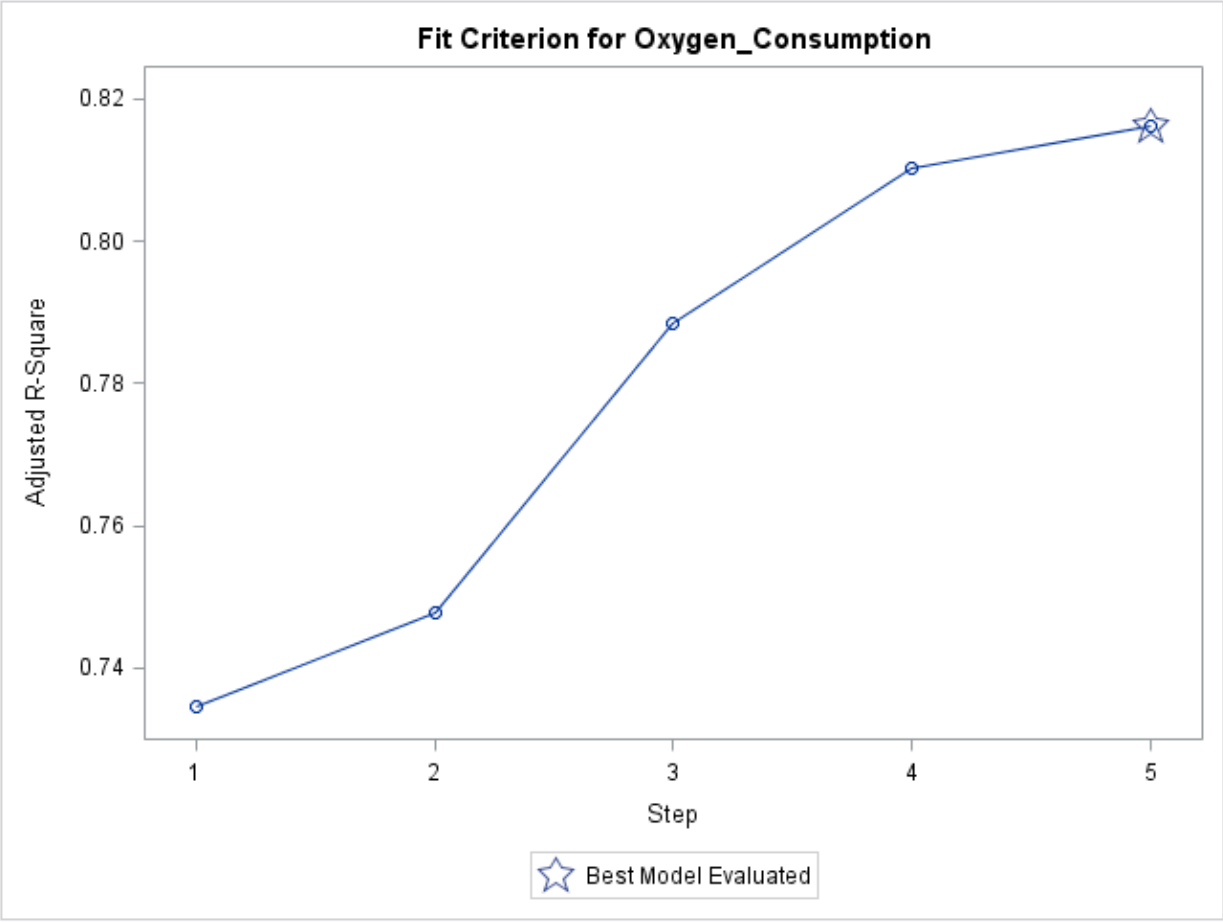
At step 2, **`Age`** is added to the model. The p-value associated with **`Age`** is 0.1222, which meets the significance level requirement set in the task.

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.

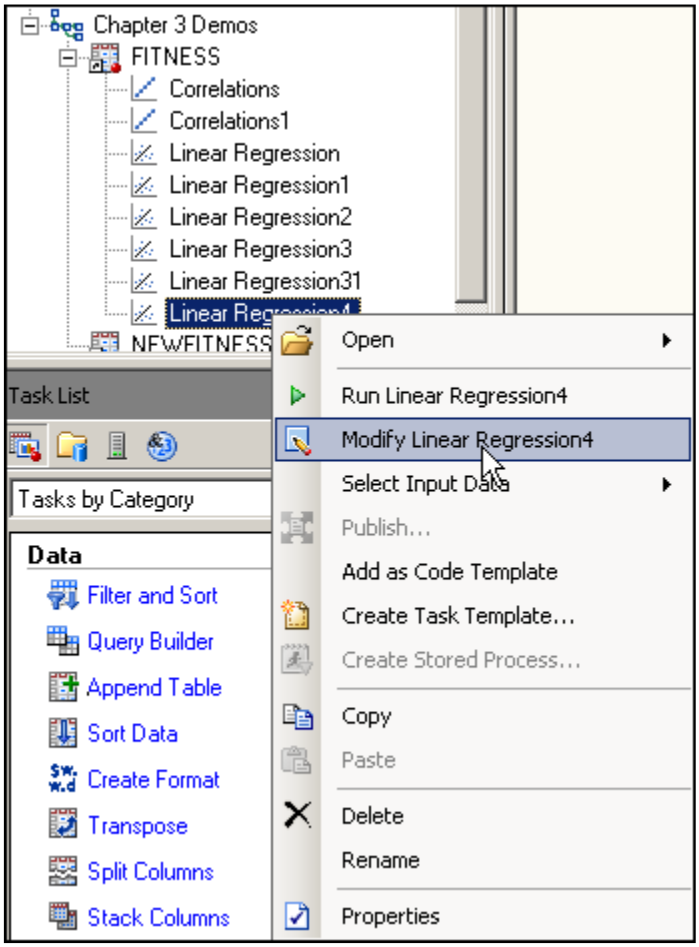


The Adjusted R-Square plot shows the progression of that statistic at each step. The star denotes the best model of the 5 tested. This is not necessarily the highest Adjusted R-Square value of all possible subsets, but is the best of the five tested in the forward selection model.

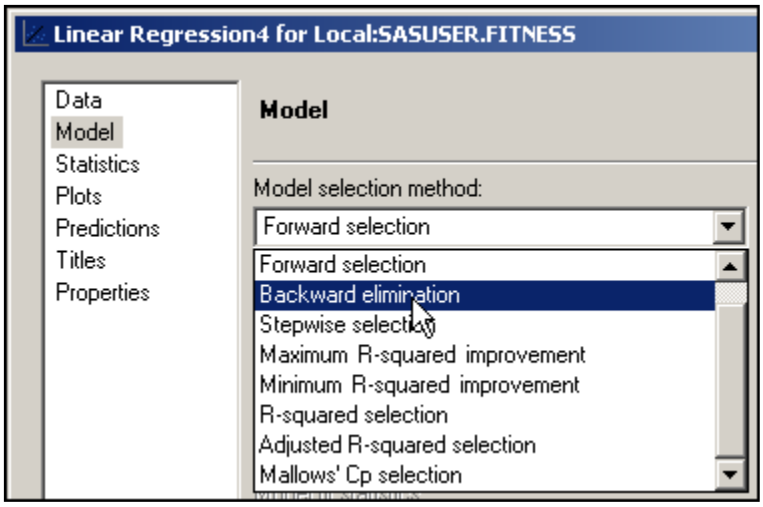
Backward – Stepwise Regression

Next, rerun the task using backward elimination.

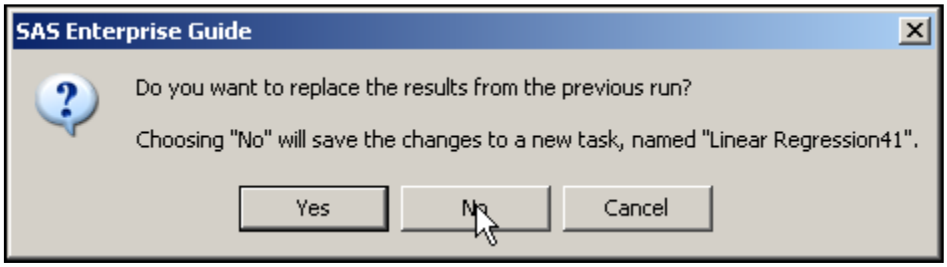
1. Reopen the previous task by right clicking the icon in the Project Tree and selecting **Modify Linear Regression4** from the drop-down menu.



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

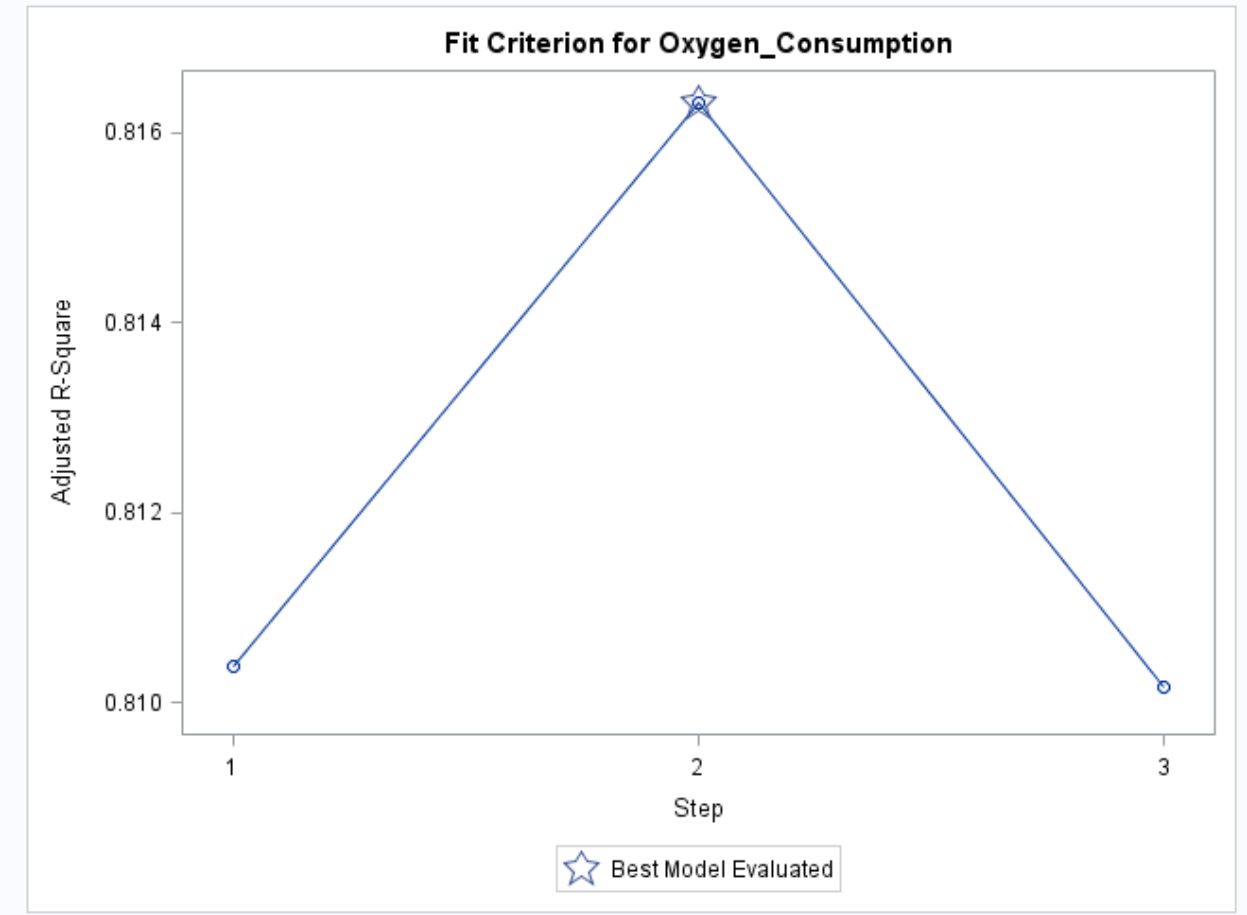
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	97.16952	11.65703	374.42127	69.48	<.0001
RunTime	-2.77576	0.34159	355.82682	66.03	<.0001
Age	-0.18903	0.09439	21.61272	4.01	0.0557
Run_Pulse	-0.34568	0.11820	46.08558	8.55	0.0071
Maximum_Pulse	0.27188	0.13438	22.05933	4.09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

Stepwise Regression

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			

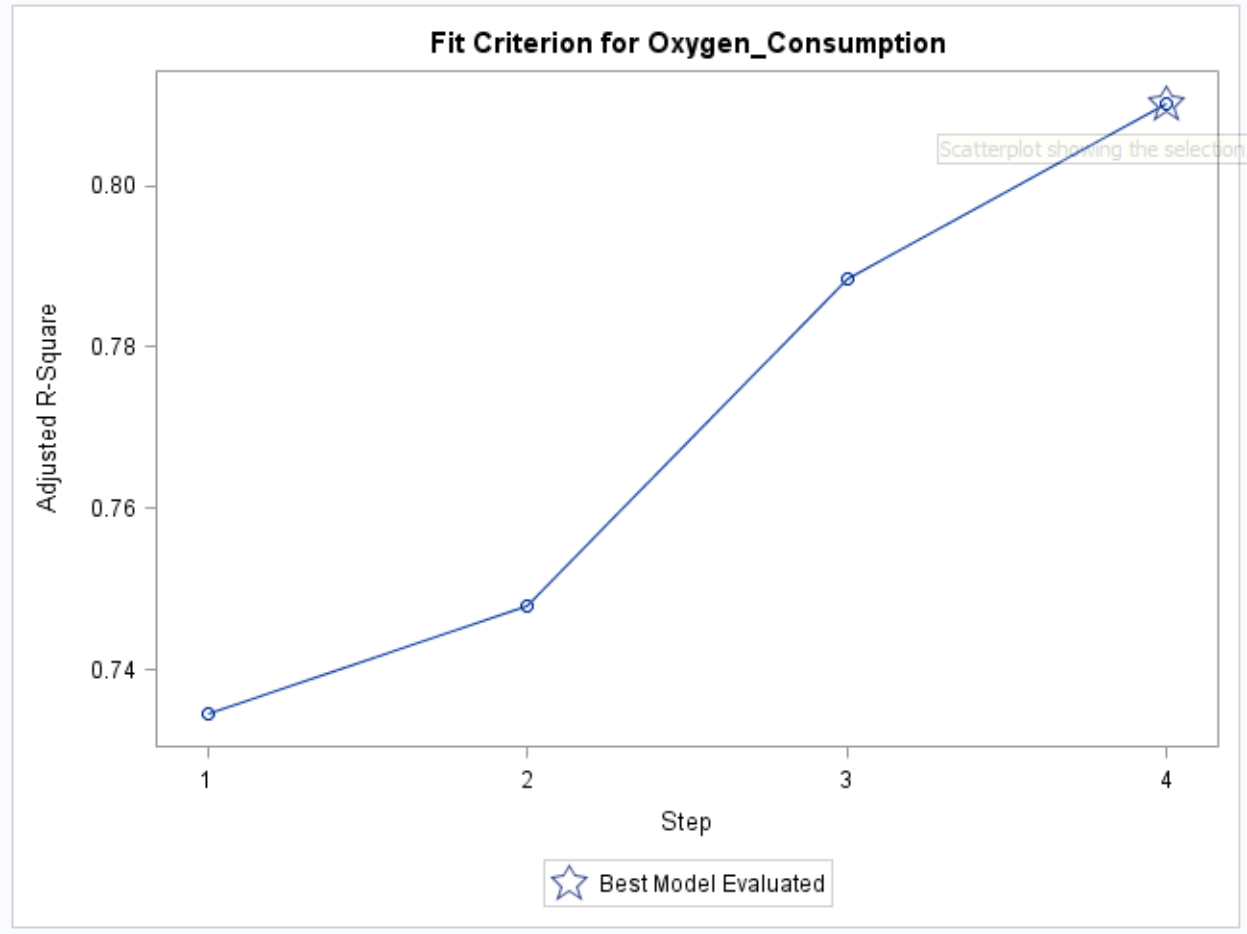
All variables left in the model are significant at the 0.1500 level.

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

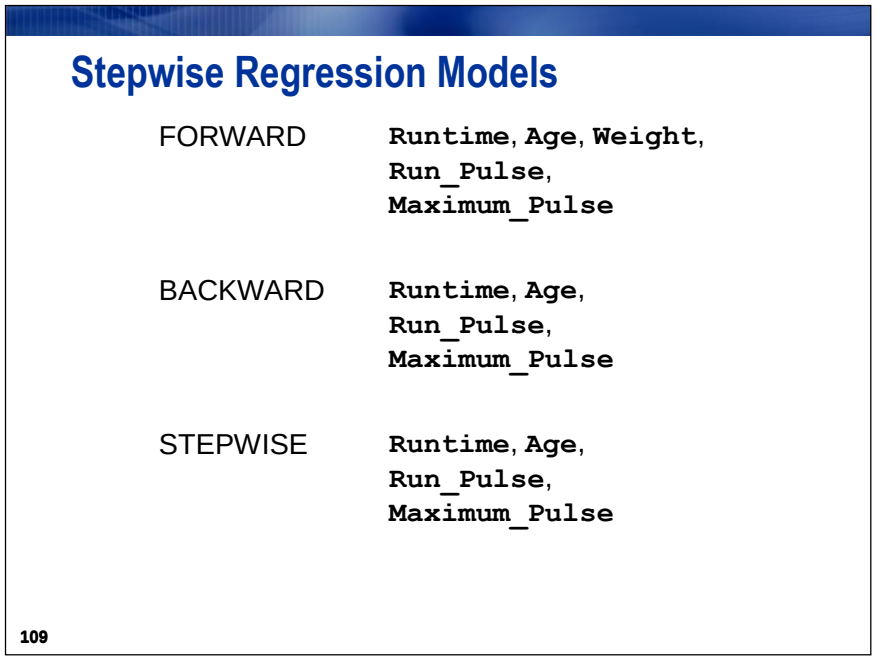
Summary of Stepwise Selection

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

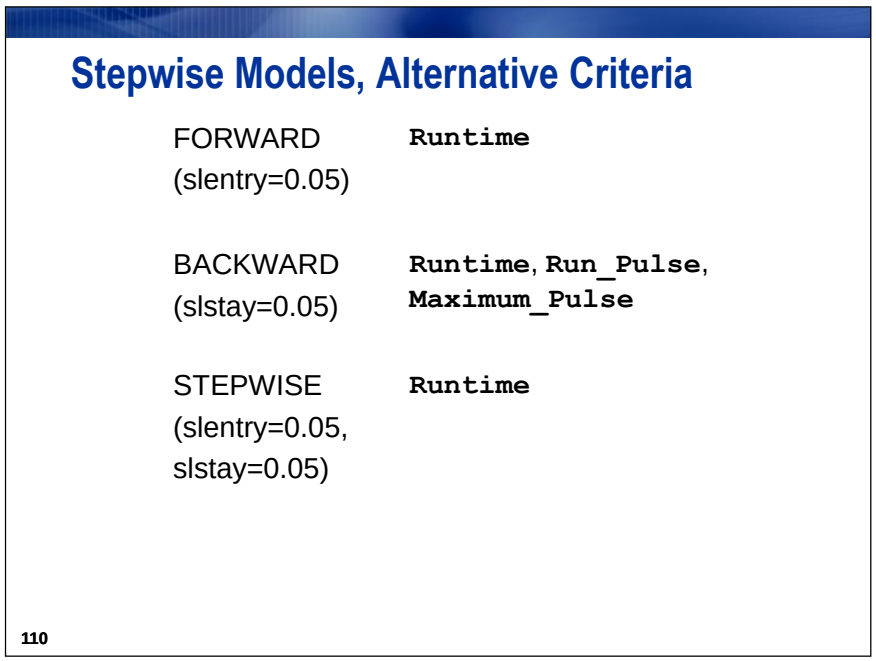
Using stepwise selection and the default *p*-value, the same subset resulted as that using backward elimination. However, it is not the same model as that resulting from forward selection.



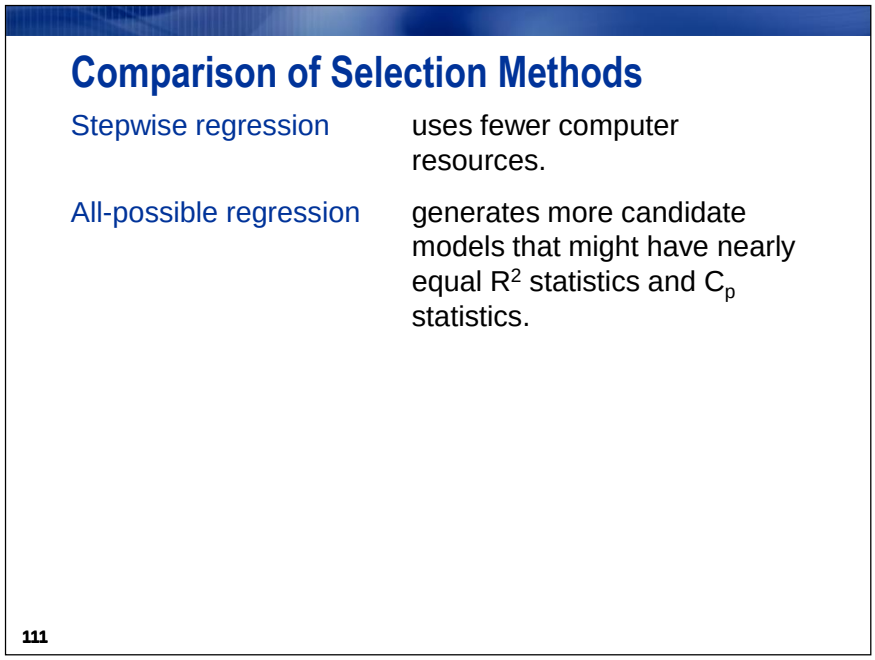
The default entry criterion is $p<.50$ for the forward selection method and $p<.15$ for the stepwise selection method. After **`RunTime`** was entered into the model, **`Age`** was entered at step 2 with a p-value of 0.1222. If the criterion were set to something less than 0.10, the final model would 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.



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.



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.



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.

Comprehensive Exercises – Above & Beyond

1. Using Automated Model Selection Techniques

Use the **BodyFat2** data set to identify a set of "best" models.

a. Use the Cp selection method to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.

1) Which set of variables was included in the best models according to each of the criteria published by Mallows and Hocking?

b. Use a stepwise regression method to select a candidate model; try forward and stepwise selection, and backward elimination. Use a significance level of 0.05 in each case.

1) Which variables were included in the final model produced with forward selection?

2) Which variables were included in the final model produced with backward elimination?

3) Which variables were included in the final model produced with stepwise selection?

c. Change the selection criterion for forward selection back to its default of 0.50.

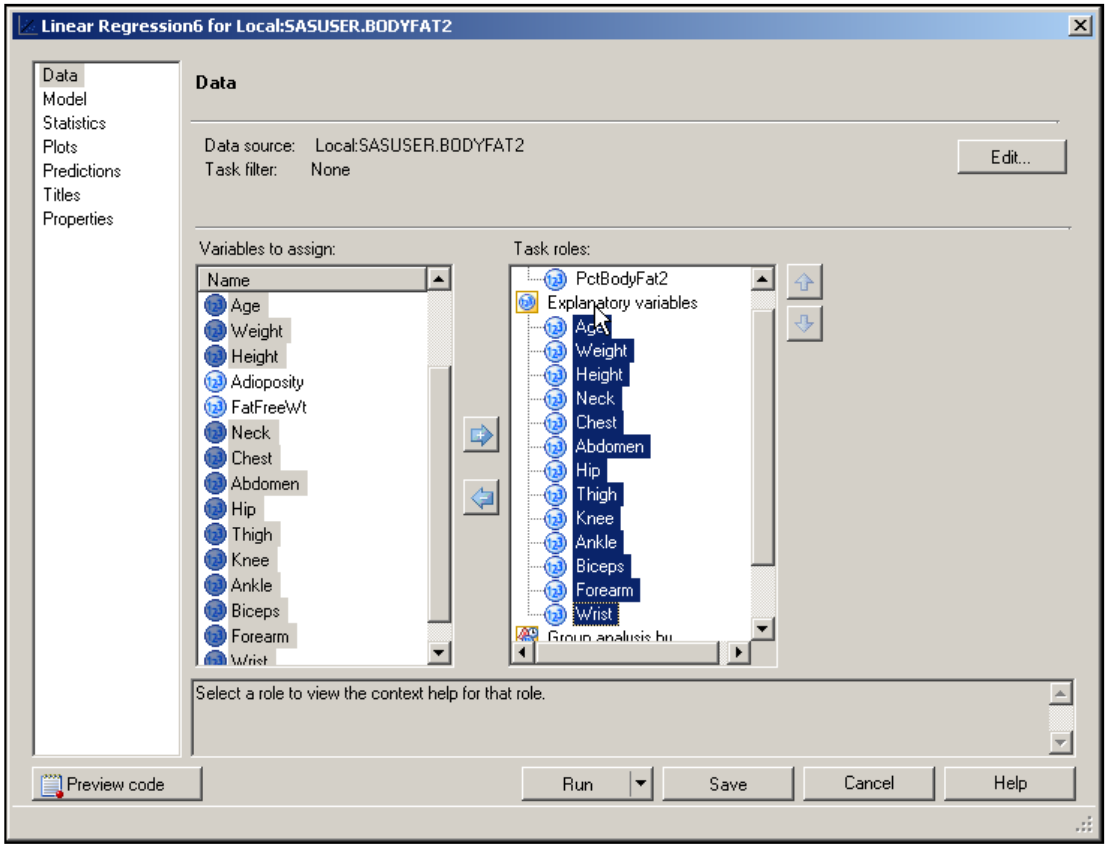
1) How many variables would have resulted from a model using forward selection and a significance level for entry criterion of 0.50 (the default), instead of 0.05?

Solutions

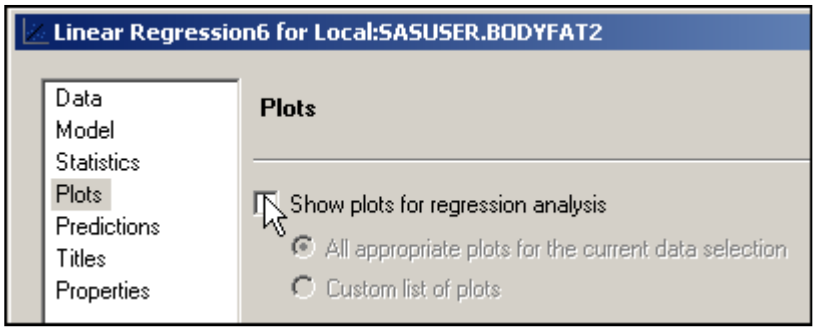
1. Performing a Multiple Regression

a. Using the **BodyFat2** data set, run a regression of **PctBodyFat2** on the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.

- With the **BodyFat2** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
- Drag **PctBodyFat2** to the dependent variable task role and **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist** to the explanatory variables task role.



- With **Plots** selected at the right, deselect **Show plots for regression analysis**.



- Change the title, if desired.

- Click Run.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	13159	1012.22506	54.50	<.0001
Error	238	4420.06401	18.57170		
Corrected Total	251	17579			

Root MSE	4.30949	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7348
Coeff Var	22.50293		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.35323	22.18616	-0.96	0.3368
Age	1	0.06457	0.03219	2.01	0.0460
Weight	1	-0.09638	0.06185	-1.56	0.1205
Height	1	-0.04394	0.17870	-0.25	0.8060
Neck	1	-0.47547	0.23557	-2.02	0.0447
Chest	1	-0.01718	0.10322	-0.17	0.8679
Abdomen	1	0.95500	0.09016	10.59	<.0001
Hip	1	-0.18859	0.14479	-1.30	0.1940
Thigh	1	0.24835	0.14617	1.70	0.0906
Knee	1	0.01395	0.24775	0.06	0.9552
Ankle	1	0.17788	0.22262	0.80	0.4251
Biceps	1	0.18230	0.17250	1.06	0.2917
Forearm	1	0.45574	0.19930	2.29	0.0231
Wrist	1	-1.65450	0.53316	-3.10	0.0021

1) Compare the ANOVA table with that from the model with only **`Abdomen`** in the previous exercise. What is different?

   There are key differences between the ANOVA table for this model and the Simple Linear Regression model.

   - The degrees of freedom for the model are much higher, 13 versus 1.
   - The Mean Square Error and the *F* Value are much smaller.

- The R-Square is higher.

2) How do the R^2 and the adjusted R^2 compare with these statistics for the **Abdomen** regression demonstration?

   Both the R^2 and adjusted R^2 for the full models are larger than the simple linear regression. The multiple regression model explains almost 75 percent of the variation in the **PctBodyFat2** variable versus only about 66 percent explained by the simple linear regression model.

3) Did the estimate for the intercept change? Did the estimate for the coefficient of **Abdomen** change?

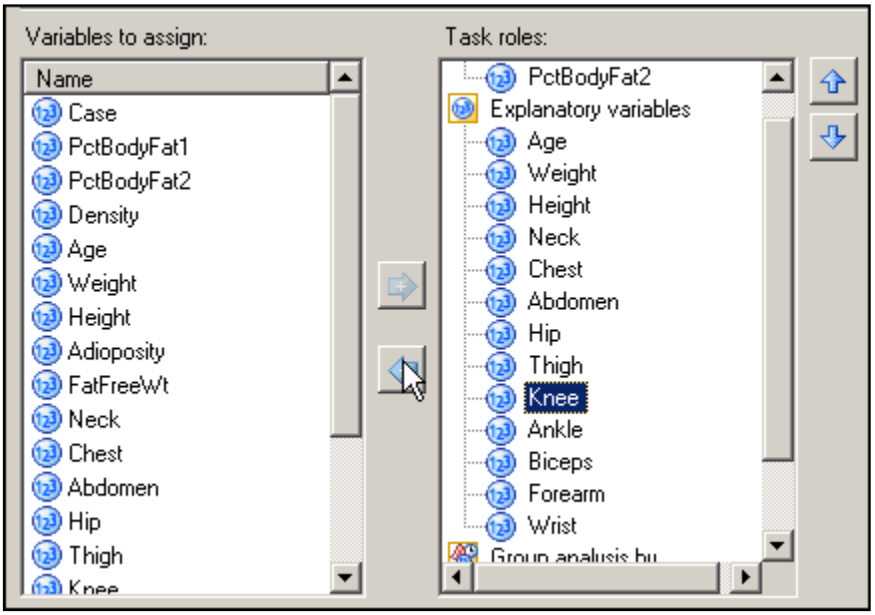
   Yes, including the other variables in the model changed both the estimate of the intercept and the slope for **Abdomen**. Also, the *p*-values for both changed dramatically. The slope and standard error of **Abdomen** are now greater.

Variable	DF	Parameter Estimate	Standard Error
Model1 Abdomen	1	0.63130	0.02855
Model2 Abdomen	1	0.95500	0.09016

b. Simplifying the Model

1) Rerun the model in **a.**, but eliminate the variable with the highest *p*-value. Compare the output with the Exercise **a.** model.

- This next step reruns the regression with **Knee** removed because it has the largest *p*-value (0.9552).
- Modify the previous Linear Regression task by right-clicking it and choosing **Modify…** from the drop-down menu.
- Remove **Knee** from task roles by selecting it and clicking  .



- Click Run.
- Do not replace the results from the previous run.

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	13159	1096.57225	59.29	<.0001
Error	239	4420.12286	18.49424		
Corrected Total	251	17579			

Root MSE	4.30049	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7359
Coeff Var	22.45595		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.30204	22.12123	-0.96	0.3365
Age	1	0.06503	0.03108	2.09	0.0374
Weight	1	-0.09602	0.06138	-1.56	0.1191
Height	1	-0.04166	0.17369	-0.24	0.8107
Neck	1	-0.47695	0.23361	-2.04	0.0423
Chest	1	-0.01732	0.10298	-0.17	0.8666
Abdomen	1	0.95497	0.08998	10.61	<.0001
Hip	1	-0.18801	0.14413	-1.30	0.1933
Thigh	1	0.25089	0.13876	1.81	0.0719
Ankle	1	0.18018	0.21841	0.82	0.4102
Biceps	1	0.18182	0.17193	1.06	0.2913
Forearm	1	0.45667	0.19820	2.30	0.0221
Wrist	1	-1.65227	0.53057	-3.11	0.0021

2) Did the *p*-value for the model change?

   No, the *p*-value for the model did not change out to four decimal places.

3) Did the R^2 and adjusted R^2 change?

   The R^2 showed essentially no change. The adjusted R^2 increased from .7348 to .7359. When an adjusted R^2 increases by removing a variable from the models, it strongly implies that the removed variable was not necessary.

4) Did the parameter estimates and their *p*-values change?

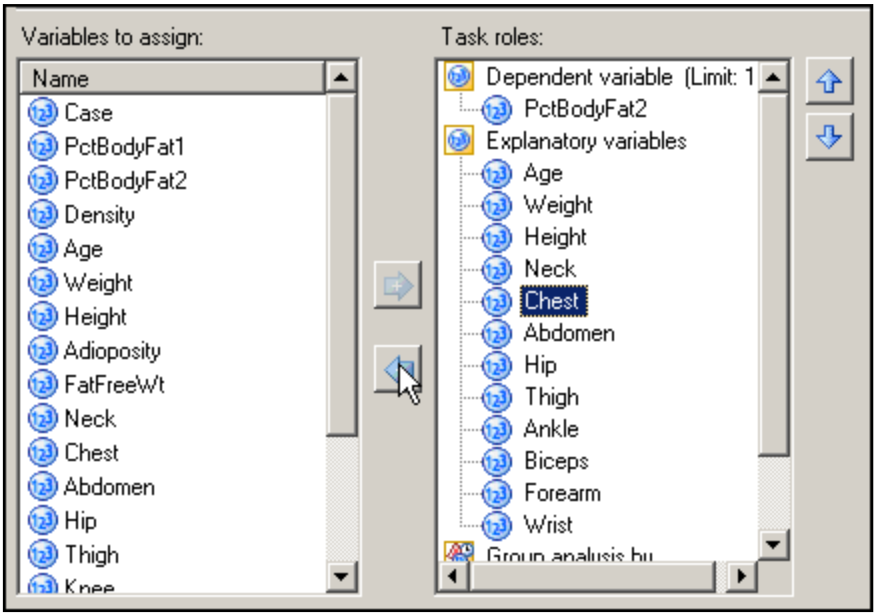
   The parameter estimates and their *p*-values changed slightly, none to any large degree.

c. More Simplifying of the Model

1) Rerun the model in Exercise **b**, but drop the variable with the highest *p*-value.

This next step reruns the regression, but with **`Chest`** removed because it has the largest *p*-value (0.8666).

- Modify the previous Linear Regression task by right-clicking it and choosing **Modify…** from the drop-down menu.
- Remove **`Chest`** from task roles by selecting it and clicking  .



- Click  .
- Do not replace the results from the previous run.

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	13158	1196.21310	64.94	<.0001
Error	240	4420.64572	18.41936		
Corrected Total	251	17579			

Root MSE	4.29178	R-Square	0.7485
Dependent Mean	19.15079	Adj R-Sq	0.7370
Coeff Var	22.41044		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-23.13736	19.20171	-1.20	0.2294
Age	1	0.06488	0.03100	2.09	0.0374
Weight	1	-0.10095	0.05380	-1.88	0.0618
Height	1	-0.03120	0.16185	-0.19	0.8473
Neck	1	-0.47631	0.23311	-2.04	0.0421
Abdomen	1	0.94965	0.08406	11.30	<.0001
Hip	1	-0.18316	0.14092	-1.30	0.1950
Thigh	1	0.25583	0.13534	1.89	0.0599
Ankle	1	0.18215	0.21765	0.84	0.4035
Biceps	1	0.18055	0.17141	1.05	0.2933
Forearm	1	0.45262	0.19634	2.31	0.0220
Wrist	1	-1.64984	0.52930	-3.12	0.0020

2) How did the output change from the previous model?

The ANOVA table did not change significantly. The R^2 remained essentially unchanged. The adjusted R^2 increased again, confirming that the variable **`Chest`** did not contribute much to explaining the variation in **`PctBodyFat2`** when the other variables are in the model.

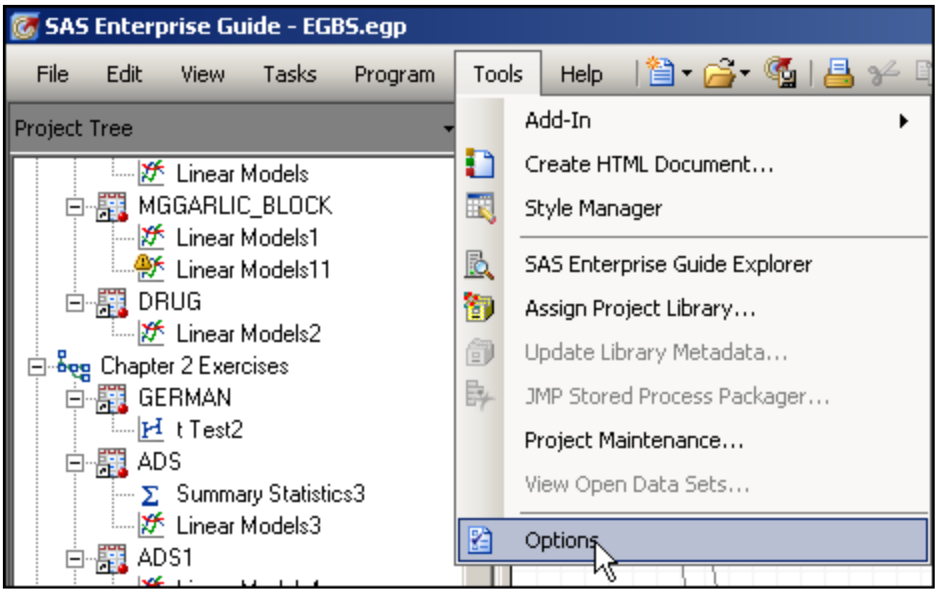
3) Did the number of parameters with *p*-values less than 0.05 change?

The *p*-value for **`Weight`** changed more than any other and is now just above 0.05. The *p*-values and parameter estimates for other variables changed much less. There are no more variables in this model with *p*-values below 0.05, compared with the previous one.

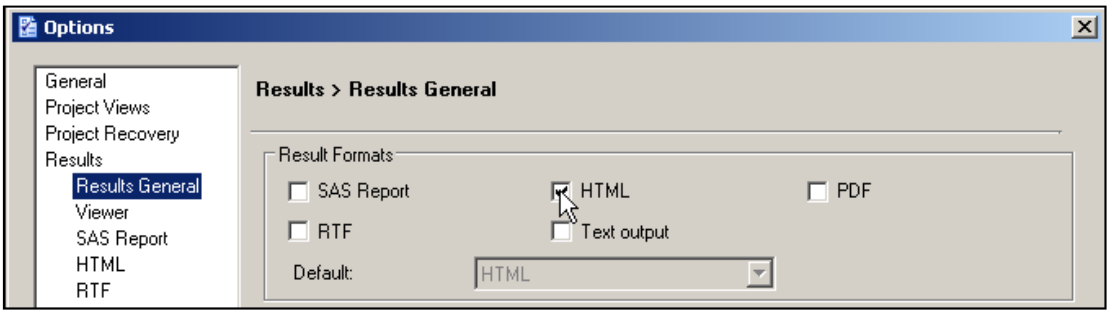
2. Using Automated Model Selection Techniques

a. Use an all-regressions technique to identify a set of candidate models, using the SELECTION=CP option, that predict **`PctBodyFat2`** as a function of the variables **`Age`**, **`Weight`**, **`Height`**, **`Neck`**, **`Chest`**, **`Abdomen`**, **`Hip`**, **`Thigh`**, **`Knee`**, **`Ankle`**, **`Biceps`**, **`Forearm`**, and **`Wrist`**.

- Click **Tools** ⇨ **Options**.



- 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**.

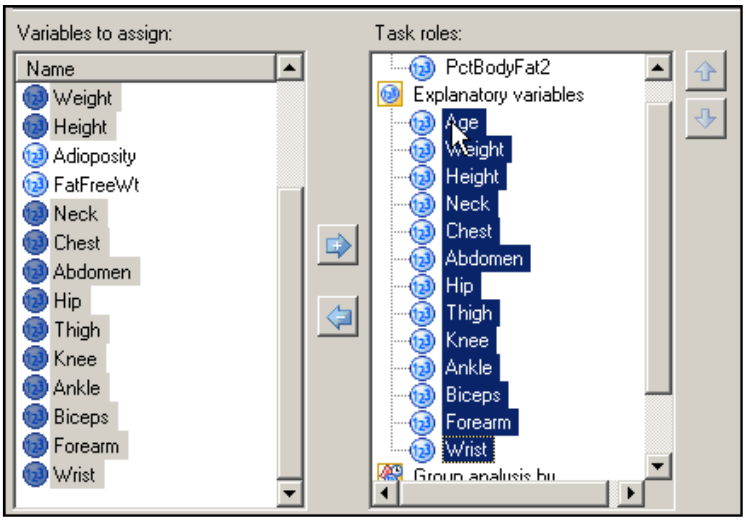


- Click Apply and then click OK.

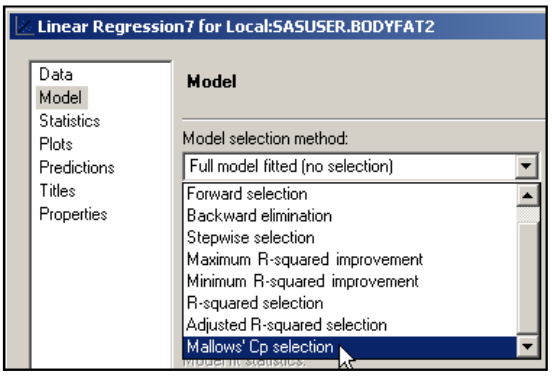
Now you are ready to run the Linear Regression task.

- With the **BodyFat2** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.

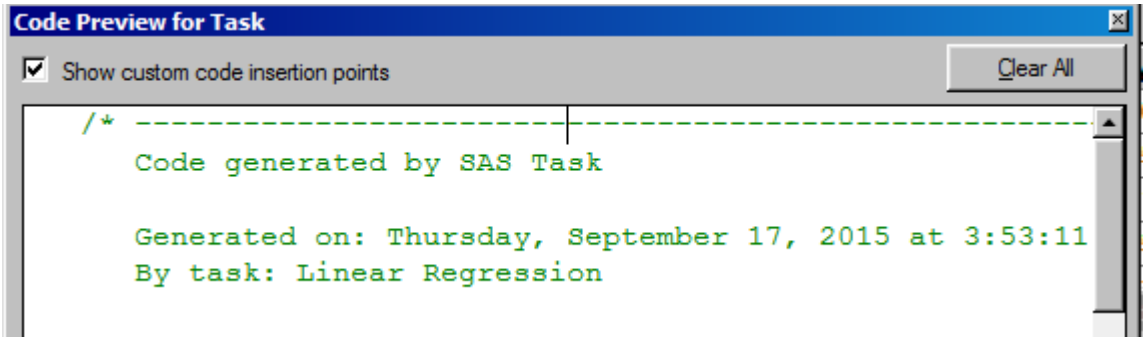
- Drag **PctBodyFat2** to the dependent variable task role and **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist** to the explanatory variables task role.



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



- Click Preview code.

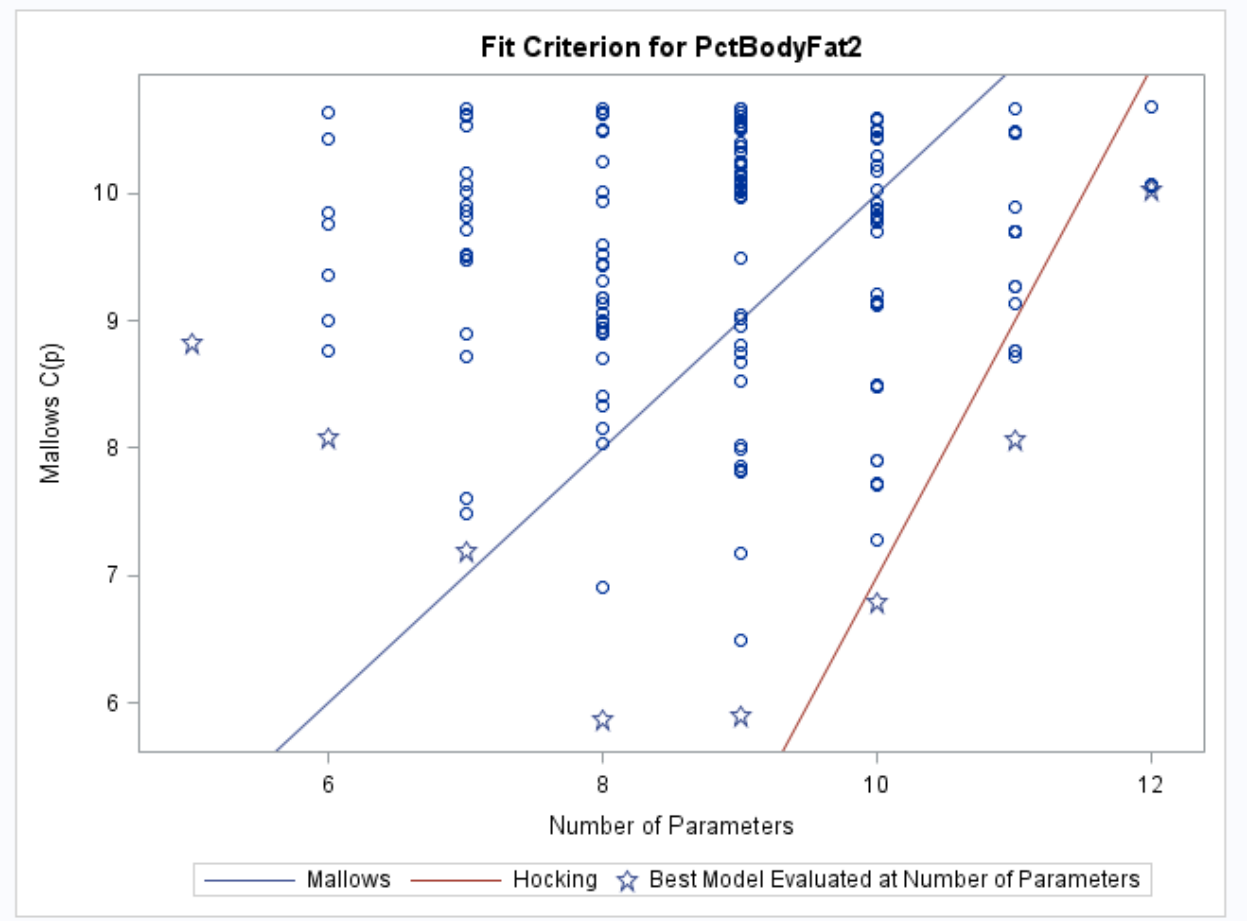


- Check the Show custom code insertion points box
- Type **GRAPHICS / IMAGEMAP=ON;** under ODS GRAPHICS ON; statement, in the <insert custom code here> area

```
   Server:  SASApp
   ------------------------------------------------------------
ODS GRAPHICS ON;


/* Start of custom user code (Framework_BeforeTaskCode)
ODS GRAPHICS / IMAGEMAP=ON;
/* End of custom user code (Framework_BeforeTaskCode) */
```

- Click ☒ in the Code Preview for Task window.
- Click Run.



The plot indicates that the best model according to Mallows' criterion is an 8-parameter model (a 7-parameter model comes close and would be worth investigating). The best model according to Hocking's criterion has 10 parameters (including the intercept).

A partial table of the models, their C(p) values and the numbers of variables in the models is displayed.

Model Index	Number in Model	C(p)	R-Square	Variables in Model
1	7	5.8653	0.7445	Age Weight Neck Abdomen Thigh Forearm Wrist
2	8	5.8986	0.7466	Age Weight Neck Abdomen Hip Thigh Forearm Wrist
3	8	6.4929	0.7459	Age Weight Neck Abdomen Thigh Biceps Forearm Wrist
4	9	6.7834	0.7477	Age Weight Neck Abdomen Hip Thigh Biceps Forearm Wrist
5	7	6.9017	0.7434	Age Weight Neck Abdomen Biceps Forearm Wrist
6	8	7.1778	0.7452	Age Weight Neck Abdomen Thigh Ankle Forearm Wrist
7	6	7.1860	0.7410	Age Weight Abdomen Thigh Forearm Wrist
8	9	7.2729	0.7472	Age Weight Neck Abdomen Hip Thigh Ankle Forearm Wrist
9	6	7.4937	0.7406	Age Weight Neck Abdomen Forearm Wrist

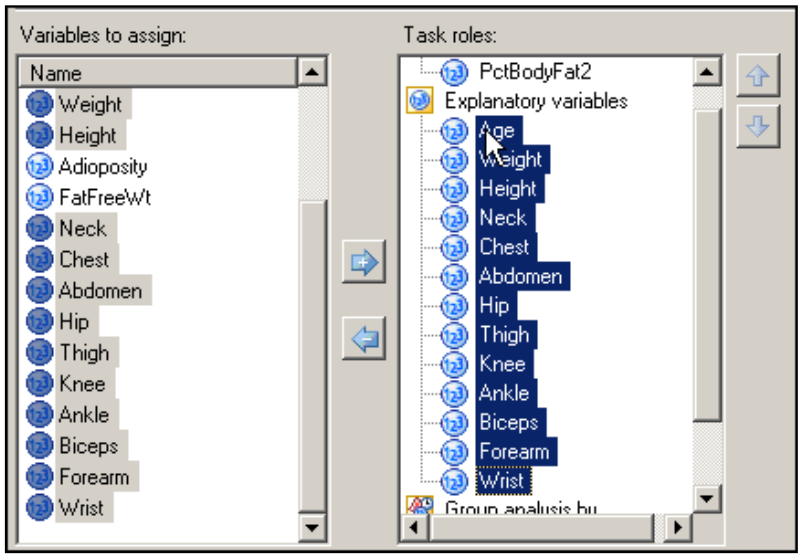
1) Which set of variables was included in the best models according to each of the criteria published by Mallows and Hocking?

   The best Mallows model is number 1 (7 variables in model plus an intercept equals 8 parameters). This model includes the variables **`Age`**, **`Weight`**, **`Neck`**, **`Abdomen`**, **`Thigh`**, **`Forearm`**, and **`Wrist`**.

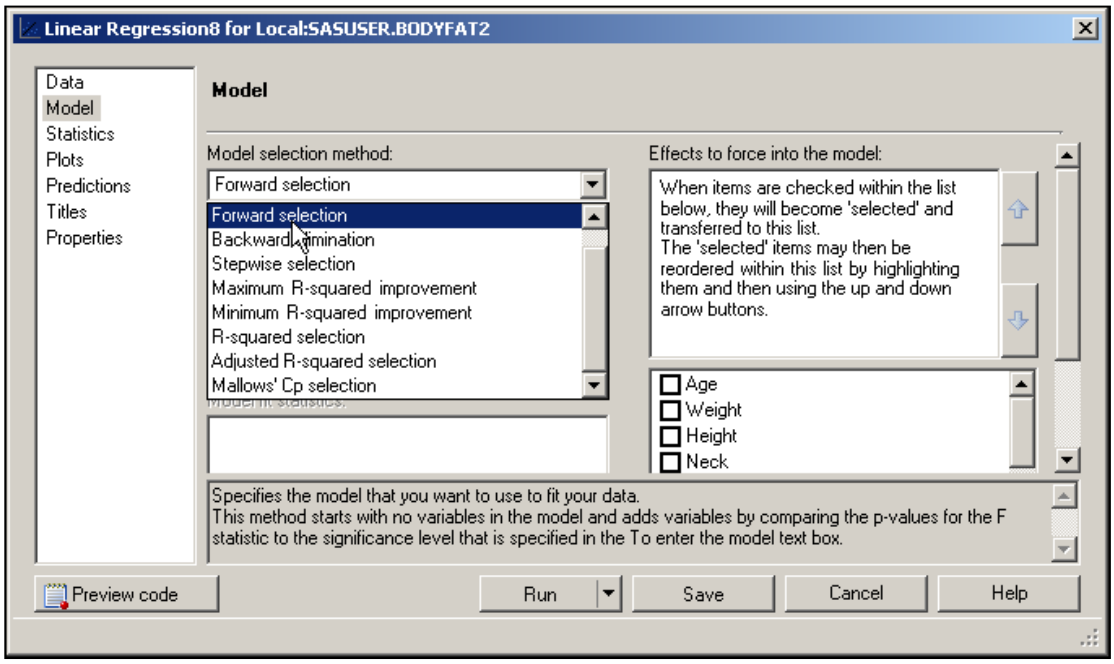
   The best Hocking model is number 4. It includes **`Hip`** and **`Biceps`**, along with the variables in the best Mallows model.

b. Use a stepwise regression method to select a candidate model; try forward and stepwise selection, and backward elimination. Use a significance level of 0.05 in each case.

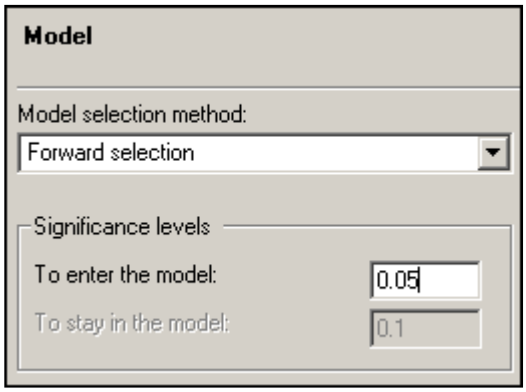
- With the **BodyFat2** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
- Drag **`PctBodyFat2`** to the dependent variable task role and **`Age`**, **`Weight`**, **`Height`**, **`Neck`**, **`Chest`**, **`Abdomen`**, **`Hip`**, **`Thigh`**, **`Knee`**, **`Ankle`**, **`Biceps`**, **`Forearm`**, and **`Wrist`** to the explanatory variables task role.



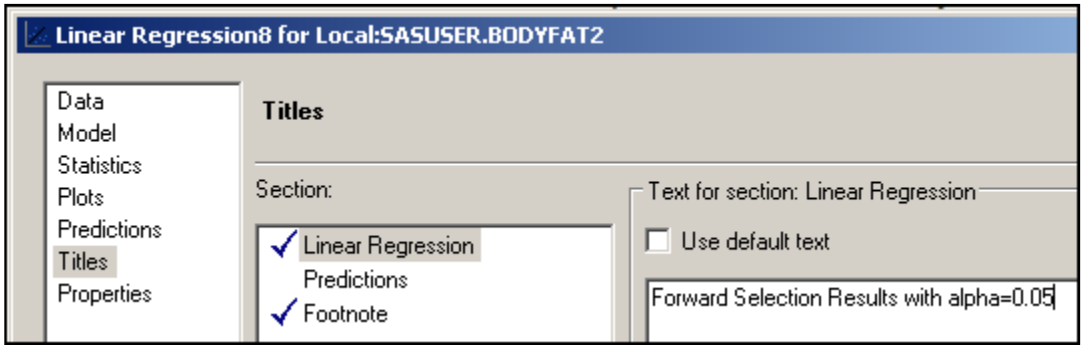
- With **Model** selected at the left, find the pull-down menu for `Model selection method` and click ▼ to find **Forward selection** at the bottom.



- Change the significance level to 0.05.



- With **Titles** selected at the left, uncheck the box for **Use default text** and then type **`Forward Selection Results with alpha=0.05`** in the text area.



- Click Run.

Forward Selection Results with alpha=0.05

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

Skip to the last step in the forward selection process:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12921	3230.18852	171.28	<.0001
Error	247	4658.23577	18.85925		
Corrected Total	251	17579			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-34.85407	7.24500	436.46987	23.14	<.0001
Weight	-0.13563	0.02475	566.43299	30.03	<.0001
Abdomen	0.99575	0.05607	5948.85562	315.43	<.0001
Forearm	0.47293	0.18166	127.81846	6.78	0.0098
Wrist	-1.50556	0.44267	218.15750	11.57	0.0008

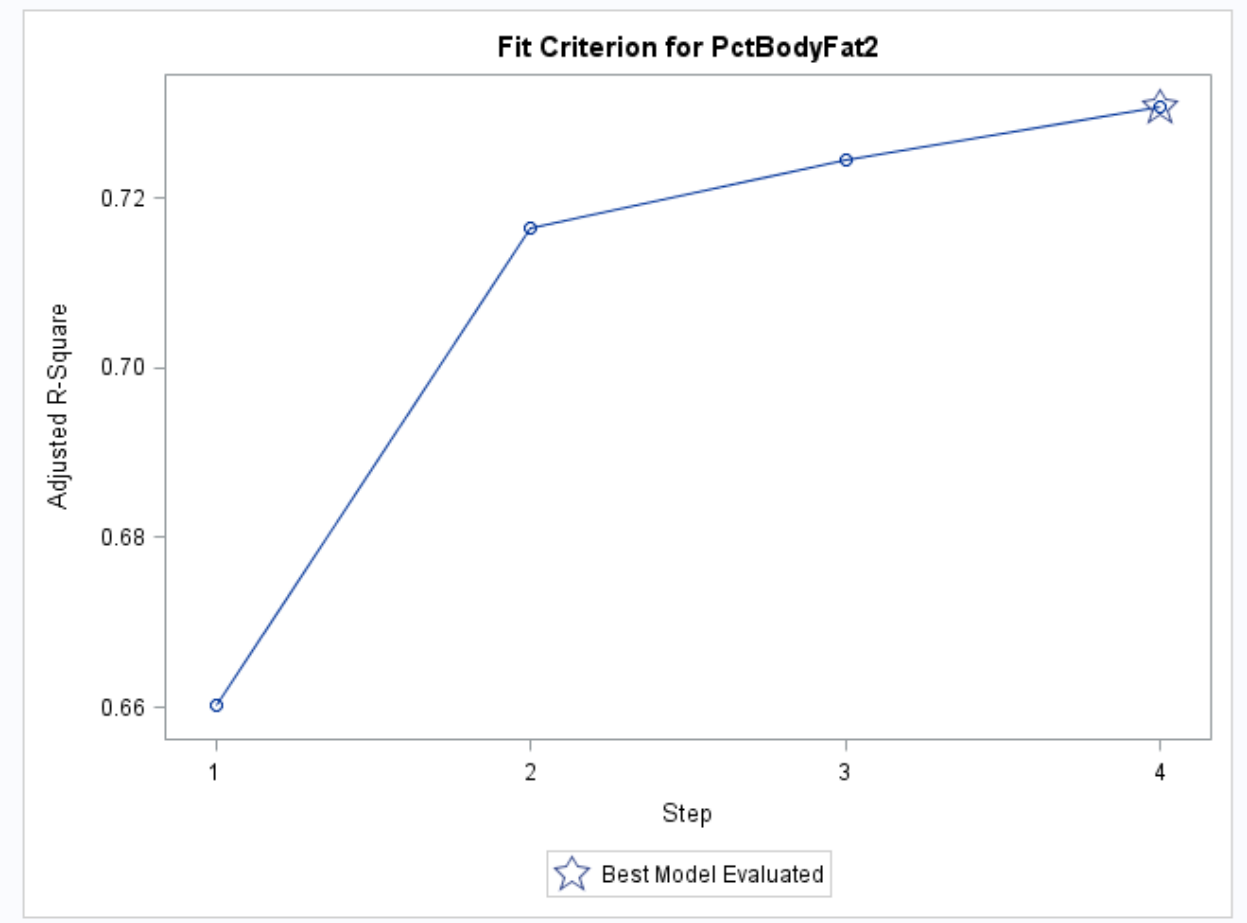
No other variable met the 0.0500 significance level for entry into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Abdomen	1	0.6617	0.6617	72.2434	488.93	<.0001
2	Weight	2	0.0571	0.7188	20.1709	50.58	<.0001
3	Wrist	3	0.0089	0.7277	13.7069	8.15	0.0047
4	Forearm	4	0.0073	0.7350	8.8244	6.78	0.0098

1) Which variables were included in the final model produced with forward selection?

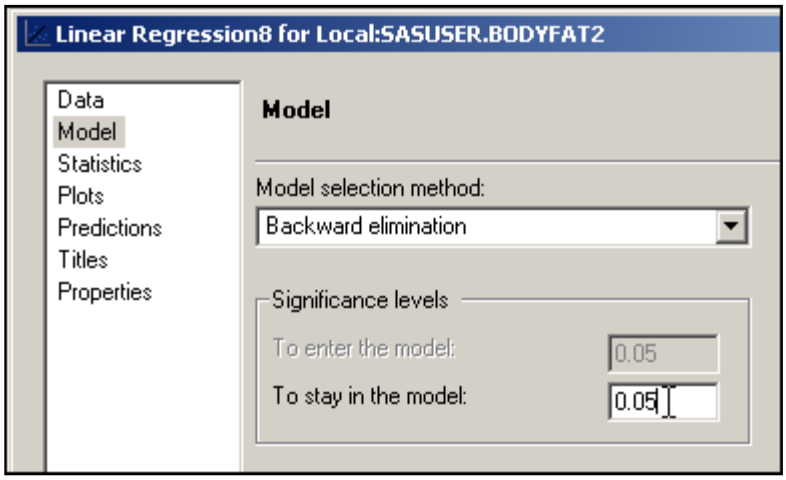
Abdomen, **Weight**, **Wrist**, and **Forearm** were included in the final model.

The Summary of Forward Selection shows that **Abdomen** alone contributed 0.6617 to the total R-square for the model. **Weight** adds 0.0571 to that total and **Wrist** and **Forearm** add less than 0.01 each. The total R-Square of this model (0.7350) is nearly as great as that for the full model (0.7485), which had 13 predictors.

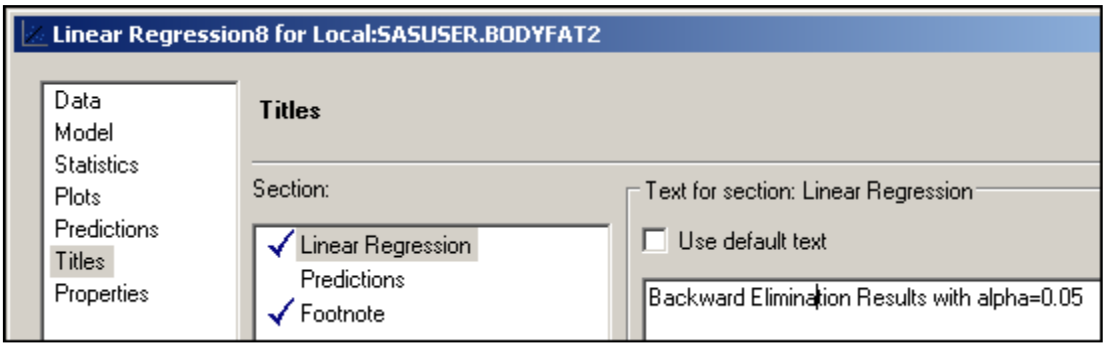


The adjusted R-Square plot shows how the adjusted R-square changes at each step. In this case, that value also increases monotonically.

- Modify the previous model by right-clicking it in the Project Tree and selecting **Modify** from the drop-down menu.
- With **Model** selected at the left, change the model selection method to **Backward elimination** and change the significance level to stay in the model to `0.05`.



- With **Titles** selected at the left, type **`Backward Elimination Results with alpha=0.05`** in the text area.



- Click Run.
- Do not replace the results from the previous run.

Partial Output

Backward Elimination Results with alpha=0.05

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: PctBodyFat2

Skip to the last step in the backward elimination process:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12921	3230.18852	171.28	<.0001
Error	247	4658.23577	18.85925		
Corrected Total	251	17579			

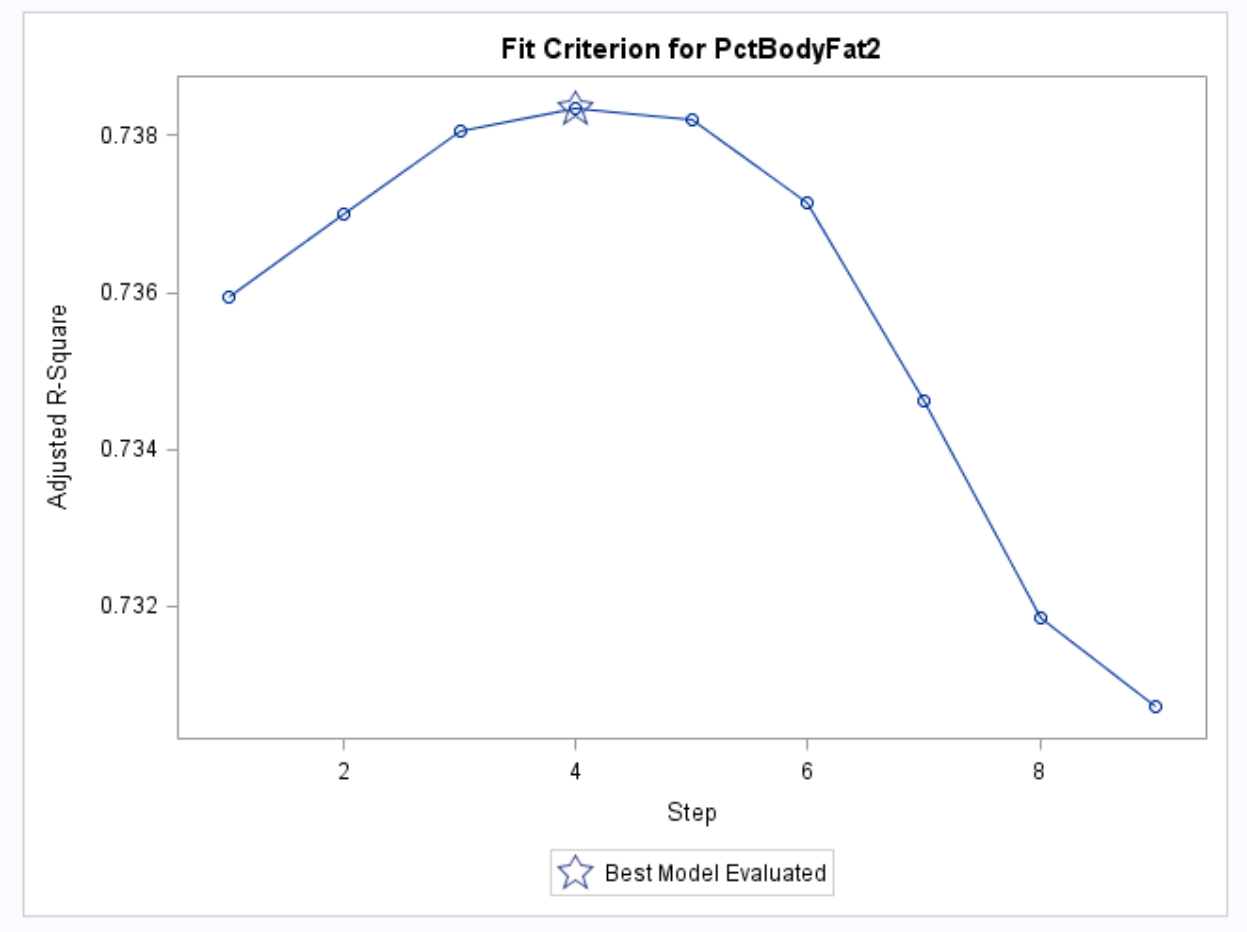
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-34.85407	7.24500	436.46987	23.14	<.0001
Weight	-0.13563	0.02475	566.43299	30.03	<.0001
Abdomen	0.99575	0.05607	5948.85562	315.43	<.0001
Forearm	0.47293	0.18166	127.81846	6.78	0.0098
Wrist	-1.50556	0.44267	218.15750	11.57	0.0008

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Knee	12	0.0000	0.7486	12.0032	0.00	0.9552
2	Chest	11	0.0000	0.7485	10.0313	0.03	0.8666
3	Height	10	0.0000	0.7485	8.0682	0.04	0.8473
4	Ankle	9	0.0008	0.7477	6.7834	0.72	0.3957
5	Biceps	8	0.0012	0.7466	5.8986	1.13	0.2888
6	Hip	7	0.0021	0.7445	5.8653	1.99	0.1594
7	Neck	6	0.0035	0.7410	7.1860	3.35	0.0684
8	Thigh	5	0.0038	0.7372	8.7588	3.57	0.0600
9	Age	4	0.0022	0.7350	8.8244	2.04	0.1542

2) Which variables were included in the final model produced with backward elimination?

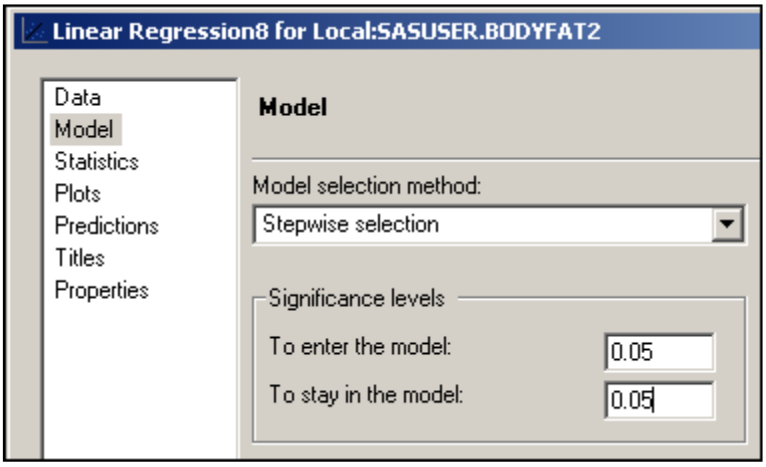
The backward elimination method using alpha=0.05 resulted in the same model as the one that resulted from the forward selection method.



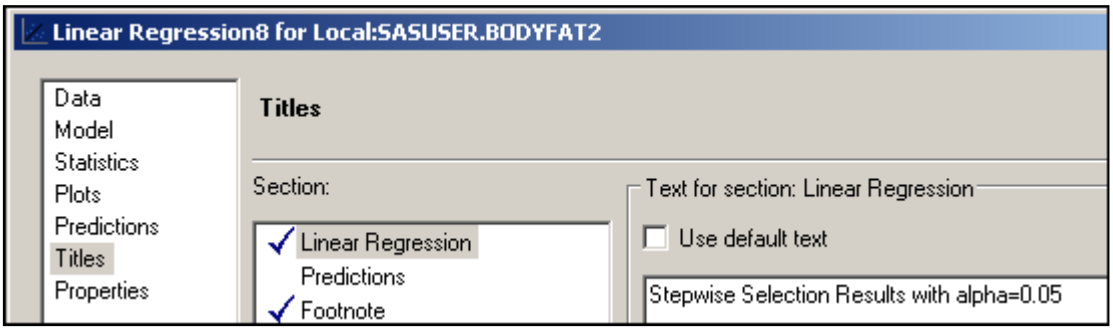
The Adjusted R-Square plot shows that, even though the backward elimination process continued to step 9, the adjusted R-square actually stopped improving at step 4, when 9 variables remained in the model. In fact, the adjusted R-square continued to get worse after step 4.

The reliance on stepwise *p*-values alone to reach a "best" model has many limitations. It is suggested that any model-building process be followed up by model validation on a separate set of data.

- Modify the forward selection model by right-clicking it in the Project Tree and selecting **Modify** from the drop-down menu.
- With **Model** selected at the left, change the model selection method to **Stepwise selection** and change both significance levels to `0.05`.



- With **Titles** selected at the left, type `Stepwise Selection Results with alpha=0.05` in the text area.



- Click Run.
- Do not replace the results from the previous run.

Partial Output

Stepwise Selection Results with alpha=0.05

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: PctBodyFat2

Skip to the last step in the stepwise selection process:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12921	3230.18852	171.28	<.0001
Error	247	4658.23577	18.85925		
Corrected Total	251	17579			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-34.85407	7.24500	436.46987	23.14	<.0001
Weight	-0.13563	0.02475	566.43299	30.03	<.0001
Abdomen	0.99575	0.05607	5948.85562	315.43	<.0001
Forearm	0.47293	0.18166	127.81846	6.78	0.0098
Wrist	-1.50556	0.44267	218.15750	11.57	0.0008

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 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	Abdomen		1	0.6617	0.6617	72.2434	488.93	<.0001
2	Weight		2	0.0571	0.7188	20.1709	50.58	<.0001
3	Wrist		3	0.0089	0.7277	13.7069	8.15	0.0047
4	Forearm		4	0.0073	0.7350	8.8244	6.78	0.0098

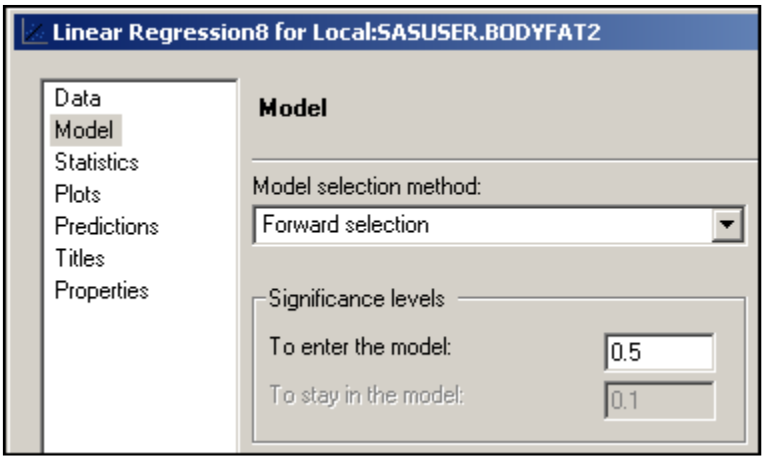
3) Which variables were included in the final model produced with stepwise selection?

The resulting model is identical to the one obtained using forward selection. This will not always be the case.

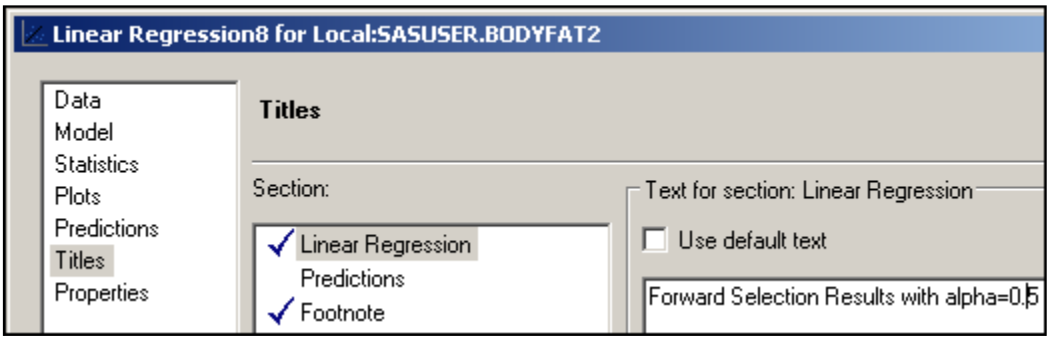
c. Change the selection criterion for forward selection back to its default of 0.50.

- Modify the forward selection model by right-clicking it in the Project Tree and selecting **Modify** from the drop-down menu.

- With **Model** selected at the left, change the significance level to `0.5`.



- With **Titles** selected at the left, type **`Forward Selection Results with alpha=0.5`** in the text area.



- Click Run.
- Do not replace the results from the previous run.

Partial Output

Forward Selection Results with alpha=0.5

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: PctBodyFat2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	13158	1315.76595	71.72	<.0001
Error	241	4421.33035	18.34577		
Corrected Total	251	17579			

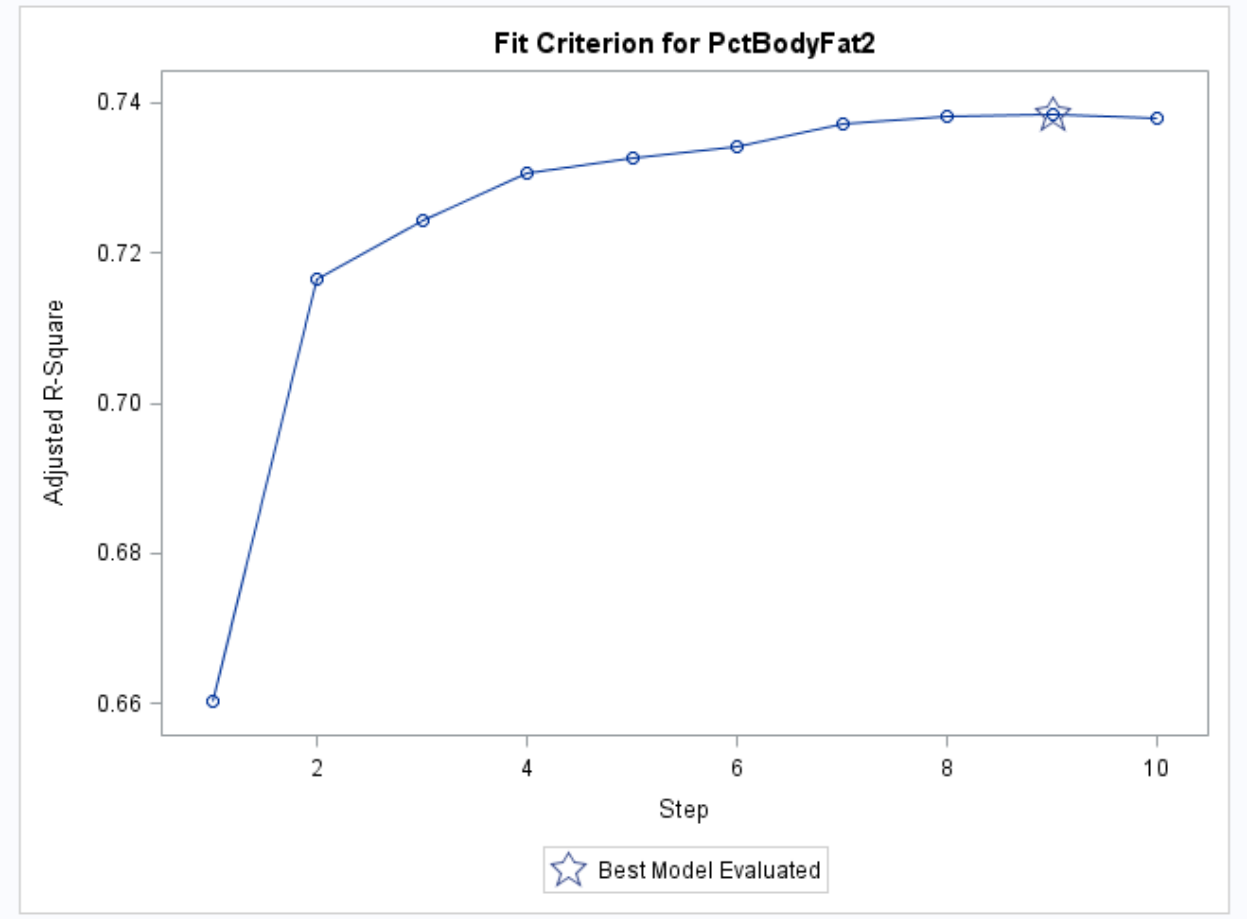
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-25.99962	12.15316	83.96376	4.58	0.0334
Age	0.06509	0.03092	81.31425	4.43	0.0363
Weight	-0.10740	0.04207	119.56769	6.52	0.0113
Neck	-0.46749	0.22812	77.05006	4.20	0.0415
Abdomen	0.95772	0.07276	3178.52750	173.26	<.0001
Hip	-0.17912	0.13908	30.42960	1.66	0.1990
Thigh	0.25926	0.13389	68.78441	3.75	0.0540
Ankle	0.18453	0.21686	13.28232	0.72	0.3957
Biceps	0.18617	0.16858	22.37399	1.22	0.2705
Forearm	0.45303	0.19593	98.08072	5.35	0.0216
Wrist	-1.65666	0.52706	181.25142	9.88	0.0019

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Abdomen	1	0.6617	0.6617	72.2434	488.93	<.0001
2	Weight	2	0.0571	0.7188	20.1709	50.58	<.0001
3	Wrist	3	0.0089	0.7277	13.7069	8.15	0.0047
4	Forearm	4	0.0073	0.7350	8.8244	6.78	0.0098
5	Neck	5	0.0029	0.7379	8.0748	2.73	0.1000
6	Age	6	0.0027	0.7406	7.4937	2.58	0.1098
7	Thigh	7	0.0038	0.7445	5.8653	3.66	0.0569
8	Hip	8	0.0021	0.7466	5.8986	1.99	0.1594
9	Biceps	9	0.0012	0.7477	6.7834	1.13	0.2888
10	Ankle	10	0.0008	0.7485	8.0682	0.72	0.3957

1) How many variables would have resulted from a model using forward selection and a significance level for entry criterion of 0.50 (the default), instead of 0.05?

   The final model contains 10 variables, rather than the 4 that resulted from using a significance level for entry value of 0.05. Variables added in the final steps contribute very little to the overall R-Square.



Adjusted R-square stops improving at step 9.