# 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

Copyright © 2013 ISYS 5503 Decision Support and Analytics, Information Systems; Timothy Paul Cronan. *For educational uses only - adapted from sources with permission. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission from the author/presenter.*

**Model Building and Interpretation**



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.





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’ CP). 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 2k or 2k-1, ignoring the intercept-only model.

In the **Fitness** data set, there are 7 possible independent variables. Therefore, there are 27 – 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’ Cp (1973) is estimated by 

where

MSE*p* is the mean squared error for the model with *p* parameters.

MSEfullis 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 Cp is up for some debate, as will be shown in the slide about Hocking’s criterion. Many choose the model with the smallest Cp value. However, Mallows recommended that the best model will have a Cp 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).



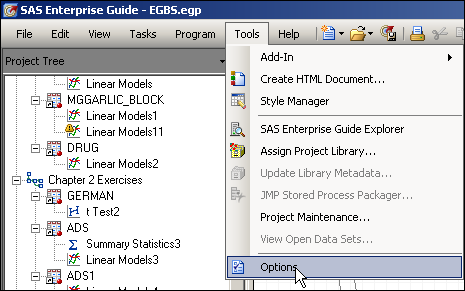
Hocking suggested the use of the Cp statistic, but with alternative criteria, depending on the purpose of the analysis. His suggestion of (Cp ≤ 2*p* − *p*full + 1) is included in the REG procedure’s calculations of criteria reference plots for best models.

## C:\Program Files\PowerServ\CourseGraphics\demo_eye.jpgAutomatic Model Selection

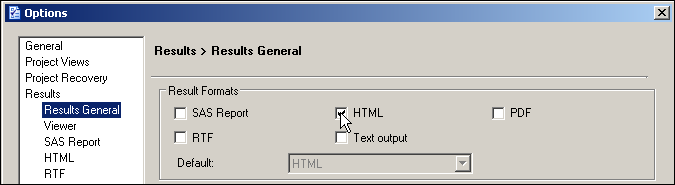
Invoke the Linear Regression task to produce a regression of **Oxygen\_Consumption** on all the other variables in the **Fitness** data set and produce plots with tool (data) tips to aid in exploration of the results.

**🖉** 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**.



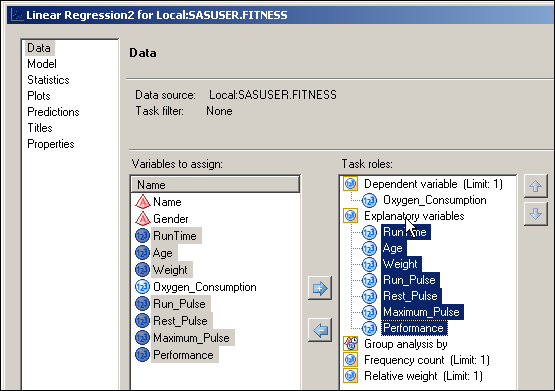
1. In the window that opens, select **Results General** under Results at the left and then uncheck the box for **SAS Report** and check the box for **HTML**.



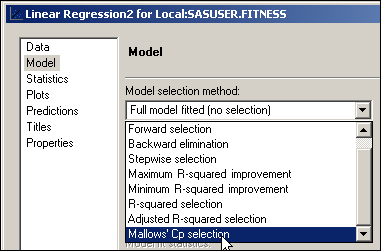
1. Click  and then click .

Now you are ready to run the Linear Regression task.

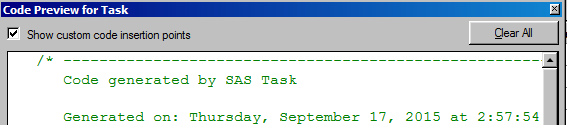
1. With the **Fitness** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
2. Drag **Oxygen\_Consumption** to the dependent variable task role and all other numeric variables   
   to the explanatory variables task role.



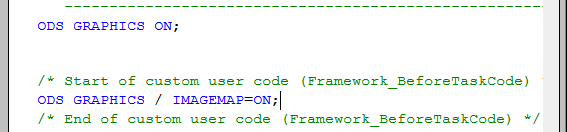
1. With **Model** selected at the left, find the pull-down menu for Model selection method   
   and click  to find **Mallows’ Cp selection** at the bottom.



1. Click .

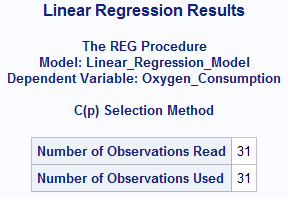


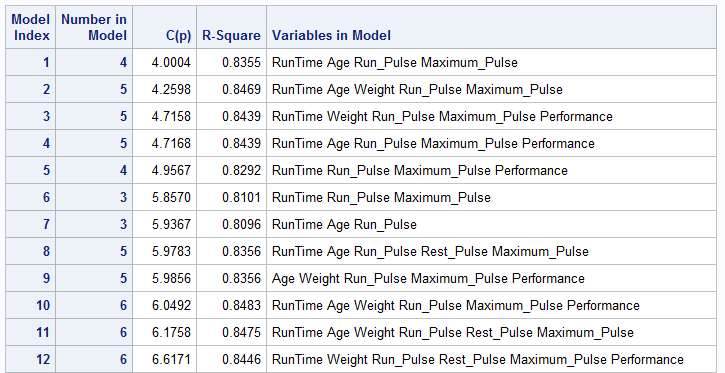
1. Enable the Show custom code insertion points box
2. Type **ODS GRAPHICS / IMAGEMAP=ON;** under the ODS GRAPHICS ON; statement in the <insert custom code here> area



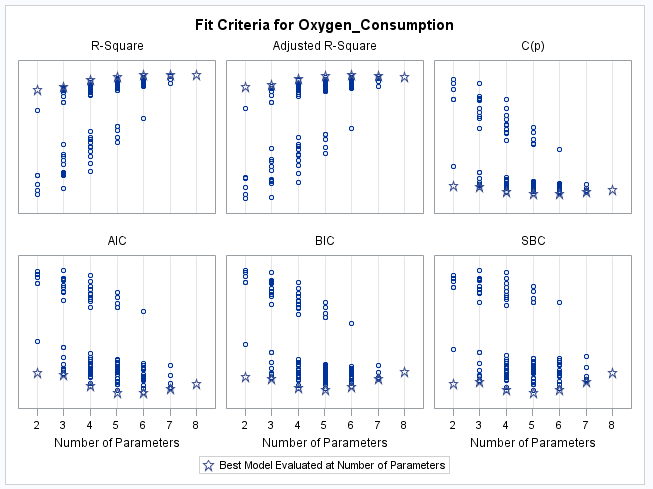
1. Click  in the Code Preview for Task window.
2. Click .

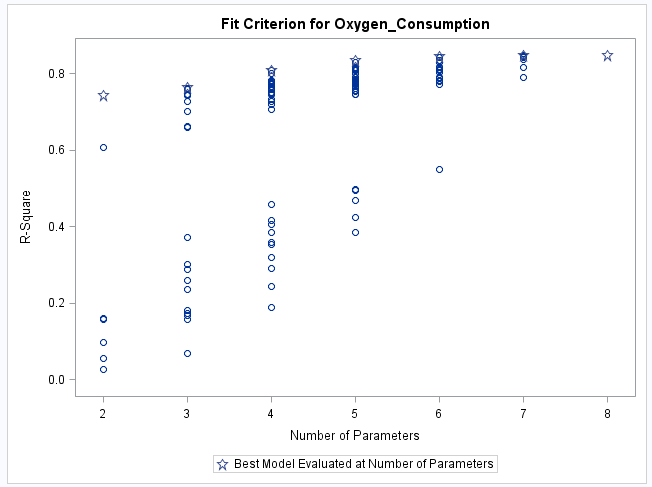
Partial HTML Output





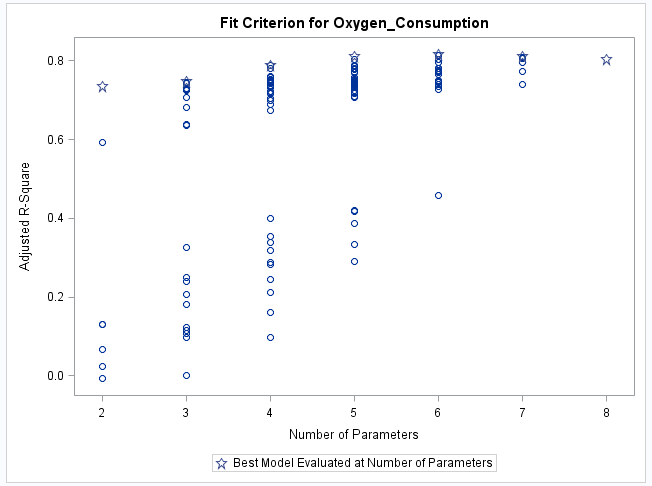
There are many models to compare. It would be unwieldy to try to determine the best model by viewing the output tables. Therefore, it is advisable to look at the plots.

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.

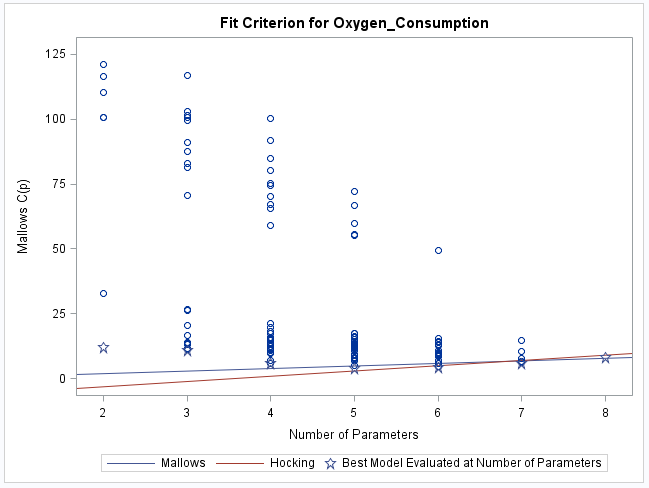


The R-Square plot compares all models based on their R2 values. As noted earlier, adding variables to a model will always increase R2 and therefore the full model will always be best. Therefore, one can only use the R2 value to compare models of equal numbers of parameters.

**🖉** The model with the greatest R2 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 Cp = *p* is plotted to help you identify models that satisfy the criterion Cp ≤ *p* for prediction. The lower line is plotted to help identify which models satisfy Hocking's criterion Cp ≤ 2*p* − *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 Cp = 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 Cp value is closest to *p*.

The smallest model that shows under the Hocking line has *p*=6.The model with the smaller Cp 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 Cp 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**.



*Some models might be essentially equivalent based on their Cp, R2 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.



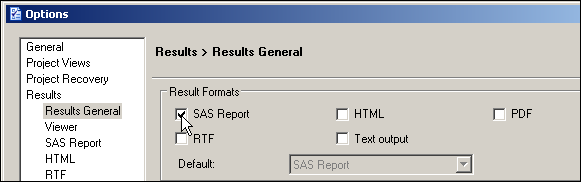
The variables **RunTime**, **Run\_Pulse**, and **Maximum\_Pulse** once again appear in all candidate models. The choice of models here depends on selection of pairs from **Performance**, **Age** and **Weight**. Here you again choose a model with objective measures, **Age** and **Weight**. That is the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

## Estimating and Testing the Coefficients for the SelectedC:\Program Files\PowerServ\CourseGraphics\demo_eye.jpg Models

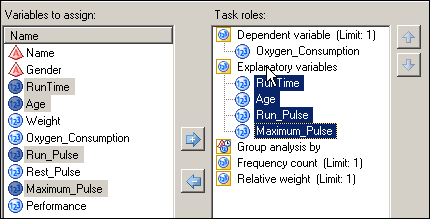
Invoke the Linear Regression task to compare the ANOVA tables and parameter estimates for the   
two-candidate models in the **Fitness** data set.

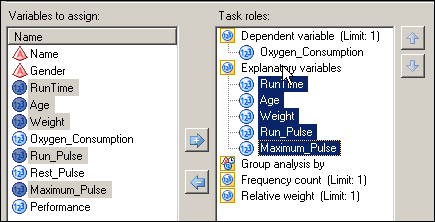
First, return reporting in SAS Enterprise Guide to SAS Report from HTML.

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**.



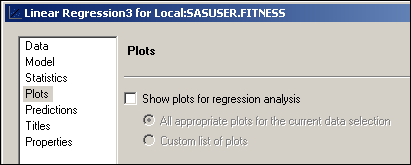
1. Click  and then click .
2. Run the Linear Regression task twice, once using the variables **Runtime**, **Age**, **Run\_Pulse**,   
   and **Maximum\_Pulse** as the explanatory variables and once using **Runtime**, **Age**, **Weight**, **Run\_Pulse**, and **Maximum\_Pulse** as the explanatory variables.



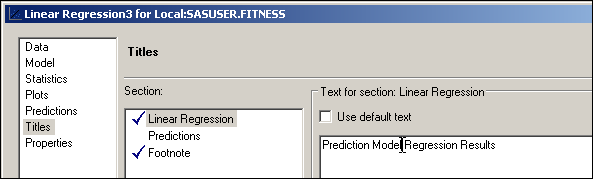


1. In each case, with **Plots** selected at the left, uncheck the box for **Show plots for regression analysis**.

**🖉** You will learn more about plots in a later chapter.

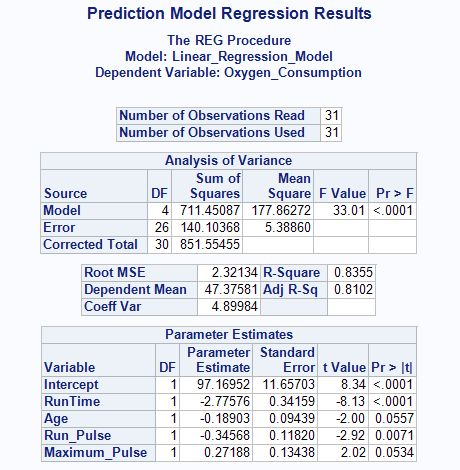


1. With **Titles** selected at the left, uncheck the box for **Use default text** and then type **Prediction Model Regression Results** in the text area for the first model and **Explanatory Model Regression Results** in the text area for the second model.



1. Click .

Output for the Prediction Model:

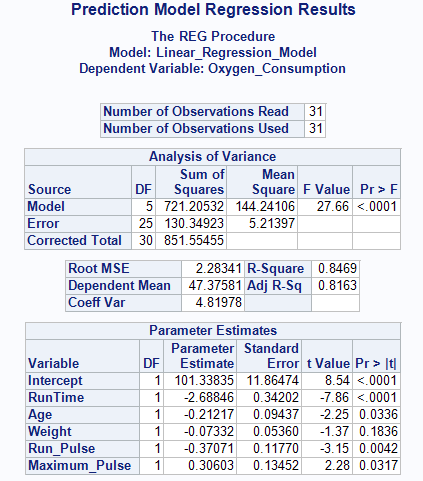


The R2 and adjusted R2 are the same as calculated during the model selection program. If there are missing values in the data set, however, this might not be true.

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 R2 is close to the R2, which suggests that there are not too many variables in the model.

Output for the Explanatory Model:



The adjusted R2 is slightly larger than in the Prediction model and very close to the R2.

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).*



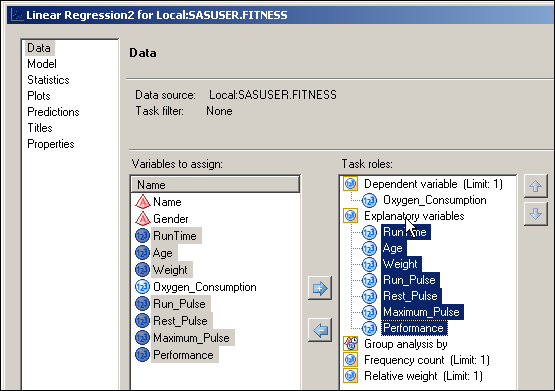
Statisticians give warnings and cautions about the over-interpretation of *p*-values from models chosen using any automated variable selection technique. Refitting many submodels in terms of an optimum fit to the data distorts the significance levels of conventional statistical tests. However, many researchers 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.

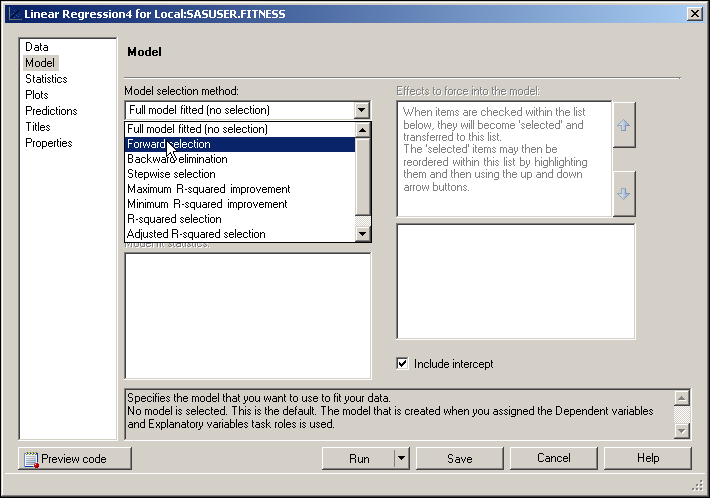
## Forward – Stepwise RegreC:\Program Files\PowerServ\CourseGraphics\demo_eye.jpgssion

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

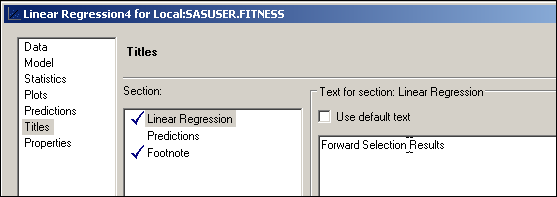
1. With the **Fitness** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
2. Drag **Oxygen\_Consumption** to the dependent variable task role and all other numeric variables to the explanatory variables task role.



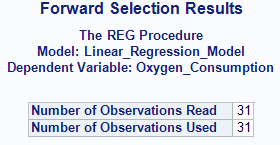
1. With **Model** selected at the left, find the pull-down menu for Model selection method and click  to find **Forward selection** at the bottom.

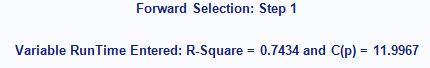


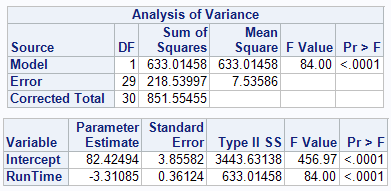
1. With **Titles** selected at the left, deselect the box for **Use default text** and then   
   type **Forward Selection Results** in the text area.



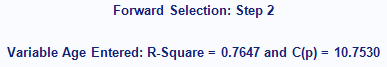
1. Click .

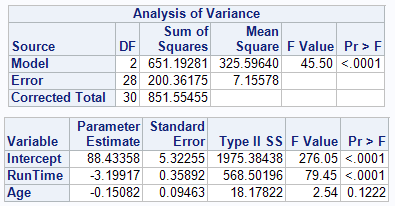






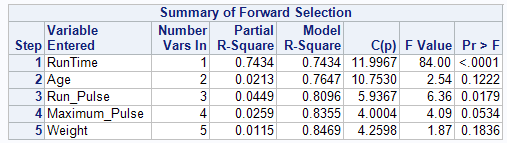
After the first step, one variable, **RunTime**, is in the model. If there are any variables that contribute significantly (*p*-value < 0.50, when controlling for **RunTime**) then the variable with the smallest *p*-value will be added to the model at the next step.





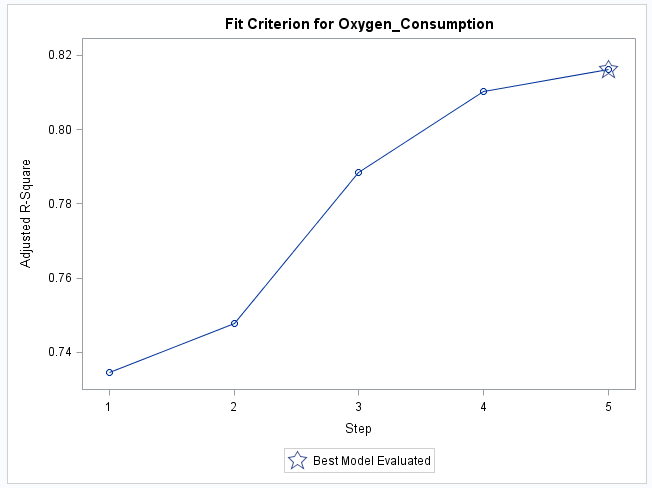
At step 2, **Age** is added to the model. The *p*-value associated with **Age** is 0.1222, which meets the significance level requirement set in the task.

Several steps are not displayed.



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 R2 shows the increase in the model R2 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.

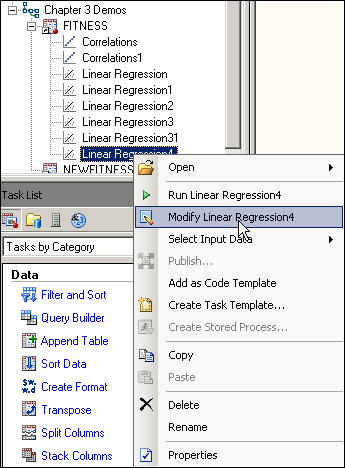


The Adjusted R-Square plot shows the progression of that statistic at each step. The star denotes the best model of the 5 tested. This is not necessarily the highest Adjusted R-Square value of all possible subsets, but is the best of the five tested in the forward selection model.

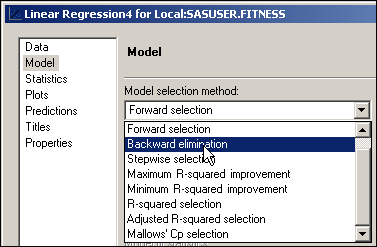
## Backward – Stepwise RegreC:\Program Files\PowerServ\CourseGraphics\demo_eye.jpgssion

Next, rerun the task using backward elimination.

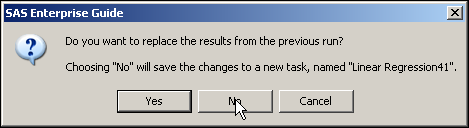
1. Reopen the previous task by right clicking the icon in the Project Tree and selecting   
   **Modify Linear Regression4** from the drop-down menu.



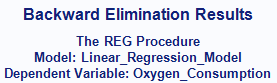
1. With **Model** selected, change the model selection method in the drop-down menu to   
   **Backward elimination**.

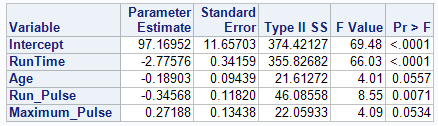


1. Change the title to **Backward Elimination Results** in the text area.
2. Click .
3. Do not replace the results of the previous run.

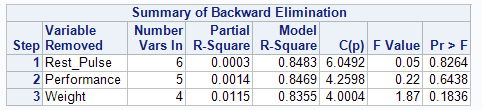


Partial Output









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 Cp, using the Mallows criterion.



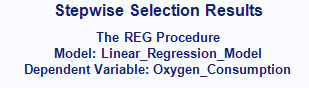
The Adjusted R-Square for the model at step 2 (before **Weight** was removed) was greatest of the three tested. Note the scale of the Y-axis for Adjusted R-Square. The differences in value among the three values is minimal. A [0-1] scale for the access would have shown how small the differences truly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

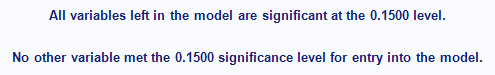
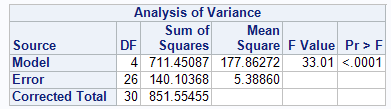
## Stepwise RegreC:\Program Files\PowerServ\CourseGraphics\demo_eye.jpgssion

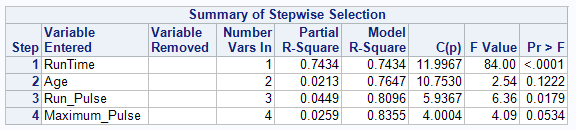
Finally, run the stepwise selection model.

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

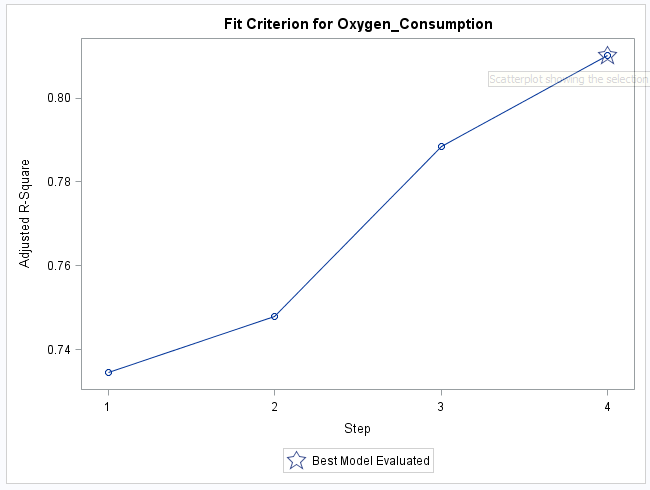
Partial Output







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 ExercisesC:\Program Files\PowerServ\CourseGraphics\exer_arrow.jpg – Above & Beyond

1. **Using Automated Model Selection Techniques**

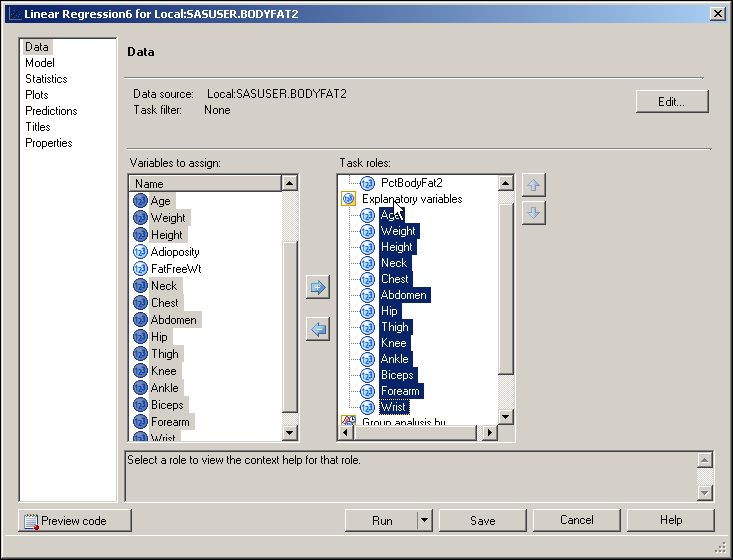
Use the **BodyFat2** data set to identify a set of “best” models.

* 1. 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?
  2. 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?
  3. 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?

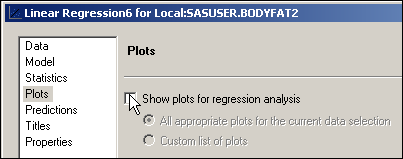
## Solutions

1. **Performing a Multiple Regression**
   1. 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**.

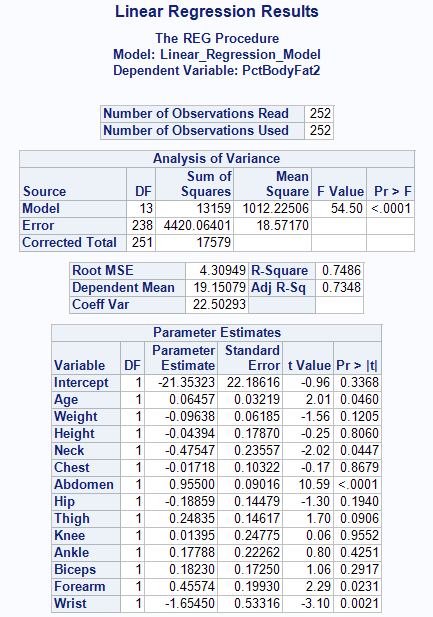
* With the **BodyFat2** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
* Drag **PctBodyFat2** to the dependent variable task role and **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and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

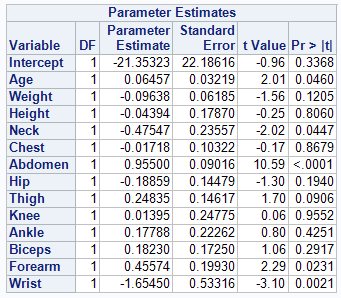


* With **Plots** selected at the right, deselect **Show plots for regression analysis**.



* Change the title, if desired.
* Click .





* + 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.
  + 1. How do the R2 and the adjusted R2 compare with these statistics for the **Abdomen** regression demonstration?

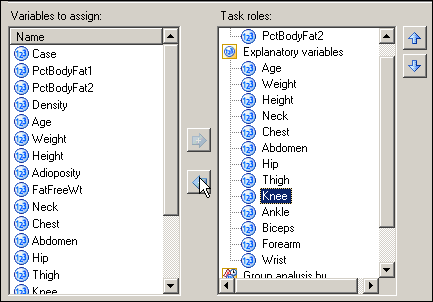
Both the R2 and adjusted R2 for the full models are larger than the simple linear regression. The multiple regression model explains almost 75 percent of the variation in the **PctBodyFat2** variable versus only about 66 percent explained by the simple linear regression model.

* + 1. Did the estimate for the intercept change? Did the estimate for the coefficient of **Abdomen** change?

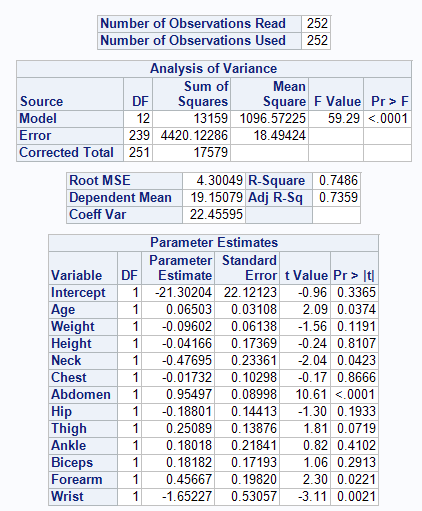
Yes, including the other variables in the model changed both the estimate of the intercept and the slope for **Abdomen**. Also, the *p*‑values for both changed dramatically. The slope and standard error of **Abdomen** are now greater.

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

* 1. 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 .
* Do not replace the results from the previous run.



* + 1. Did the *p*-value for the model change?

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

* + 1. Did the R2 and adjusted R2 change?

The R2 showed essentially no change. The adjusted R2 increased from .7348 to .7359. When an adjusted R2 increases by removing a variable from the models, it strongly implies that the removed variable was not necessary.

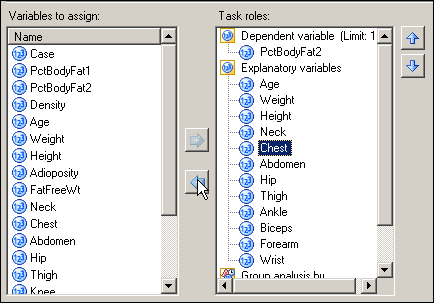
* + 1. Did the parameter estimates and their *p*‑values change?

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

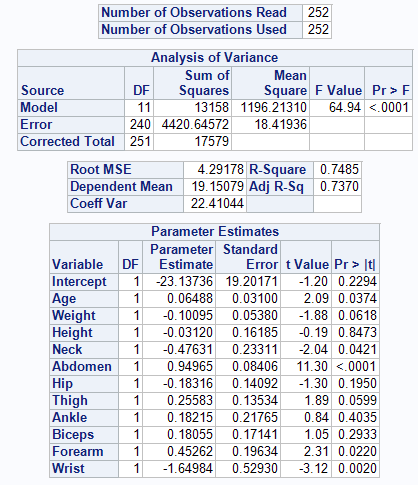
* 1. More Simplifying of the Model
     1. Rerun the model in Exercise **b**, but drop the variable with the highest *p*‑value.

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.



* + 1. How did the output change from the previous model?

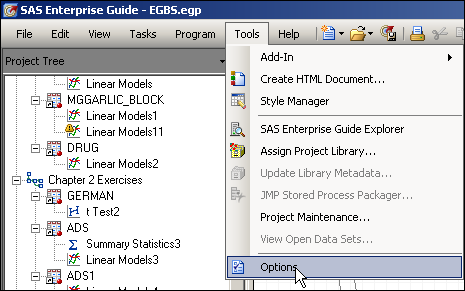
The ANOVA table did not change significantly. The R2 remained essentially unchanged. The adjusted R2 increased again, confirming that the variable **Chest** did not contribute much to explaining the variation in **PctBodyFat2** when the other variables are in the model.

* + 1. Did the number of parameters with *p*‑values less than 0.05 change?

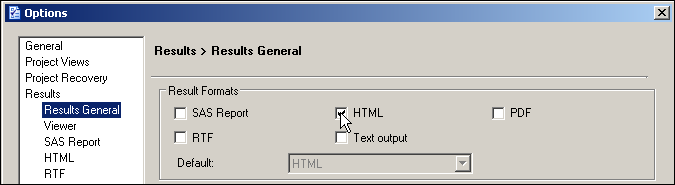
The *p*-value for **Weight** changed more than any other and is now just above 0.05. The *p*‑values and parameter estimates for other variables changed much less. There are no more variables in this model with *p*-values below 0.05, compared with the previous one.

1. **Using Automated Model Selection Techniques**
   1. 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**.



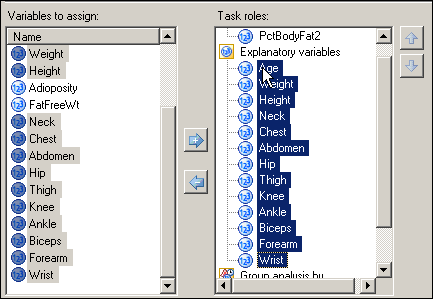
* In the window that opens, select **Results General** under Results at the left and then uncheck the box for **SAS Report** and check the box for **HTML**.



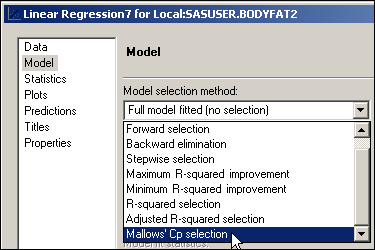
* Click  and then click .

Now you are ready to run the Linear Regression task.

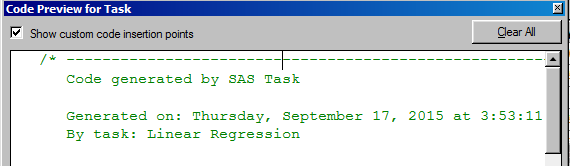
* With the **BodyFat2** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
* Drag **PctBodyFat2** to the dependent variable task role and **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and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



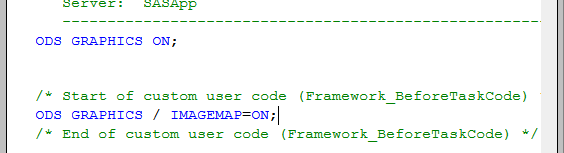
* With **Model** selected at the left, find the pull-down menu for Model selection method and click  to find **Mallows’ Cp selection** at the bottom.



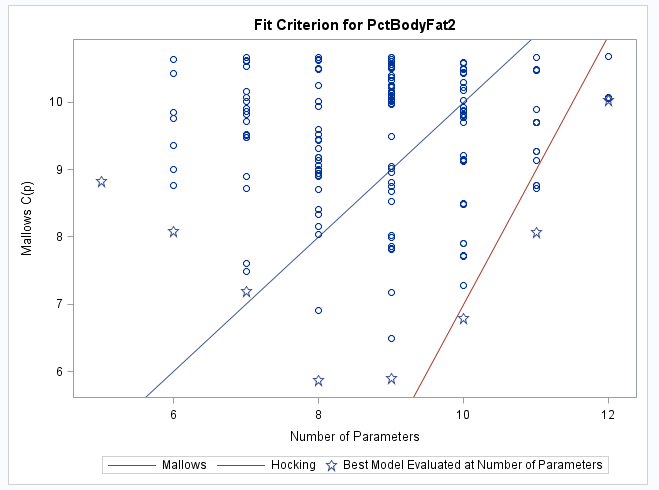
* Click .



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

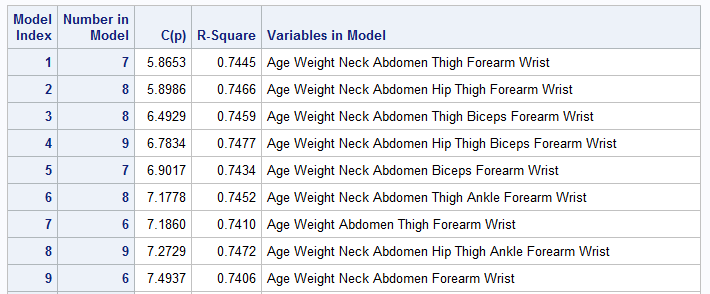


* Click  in the Code Preview for Task window.
* Click .



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.

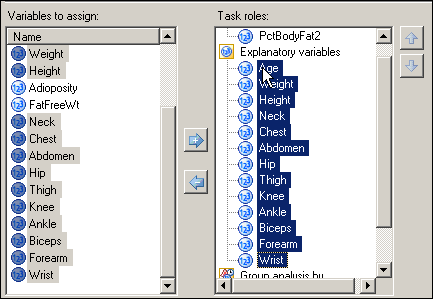


* + 1. Which set of variables was included in the best models according to each of the criteria published by Mallows and Hocking?

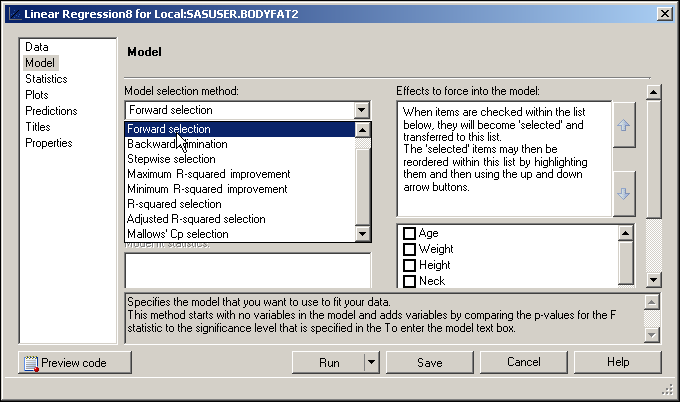
The best Mallows model is number 1 (7 variables in model plus an intercept equals 8 parameters). This model includes the variables **Age**, **Weight**, **Neck**, **Abdomen**, **Thigh**, **Forearm**, and **Wrist**.

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

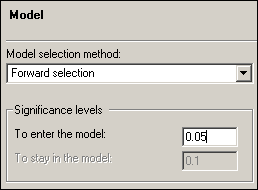
* 1. Use a stepwise regression method to select a candidate model; try forward and stepwise selection, and backward elimination. Use a significance level of 0.05 in each case.
* With the **BodyFat2** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
* Drag **PctBodyFat2** 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.



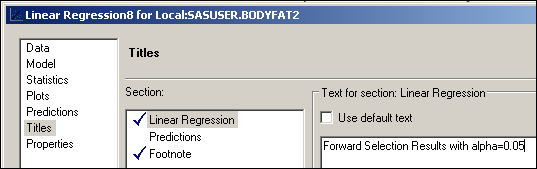
* With **Model** selected at the left, find the pull-down menu for Model selection method and click  to find **Forward selection** at the bottom.



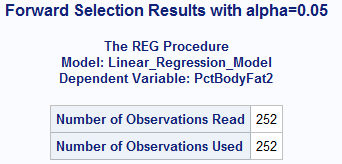
* Change the significance level to enter the model to 0.05.



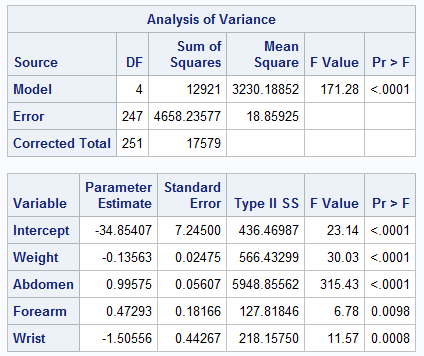
* With **Titles** selected at the left, uncheck the box for **Use default text** and then type **Forward Selection Results with alpha=0.05** in the text area.



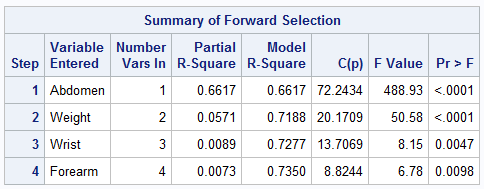
* Click .



Skip to the last step in the forward selection process:



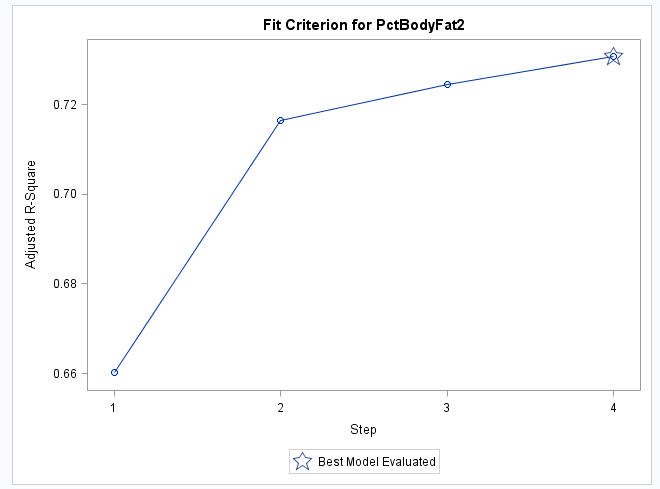




* + 1. Which variables were included in the final model produced with forward selection?

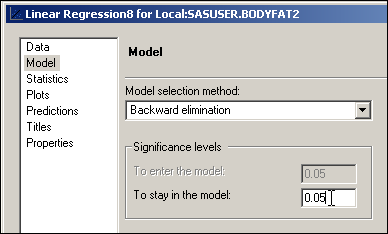
**Abdomen**, **Weight**, **Wrist**, and **Forearm** were included in the final model.

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.

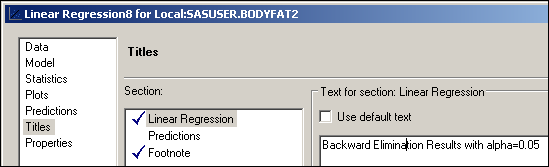


The adjusted R-Square plot shows how the adjusted R-square changes at each step. In this case, that value also increases monotonically.

* 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.

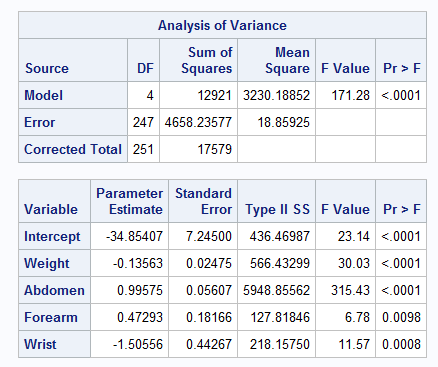


* Click .
* Do not replace the results from the previous run.

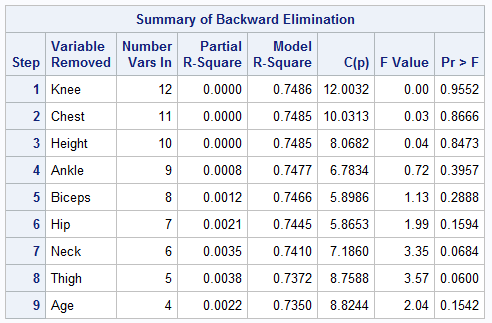
Partial Output



Skip to the last step in the backward elimination process:

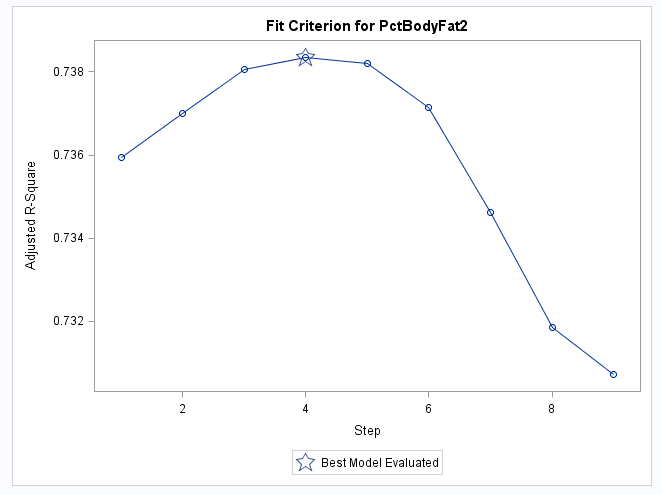






* + 1. Which variables were included in the final model produced with backward elimination?

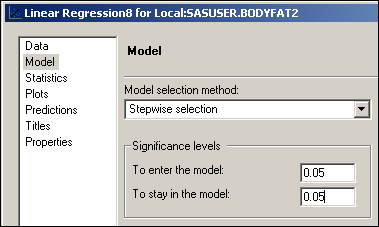
The backward elimination method using alpha=0.05 resulted in the same model as the one that resulted from the forward selection method.



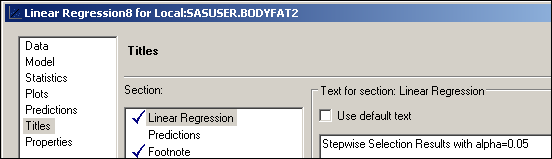
The Adjusted R-Square plot shows that, even though the backward elimination process continued to step 9, the adjusted R-square actually stopped improving at step 4, when 9 variables remained in the model. In fact, the adjusted R-square continued to get worse after step 4.

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.

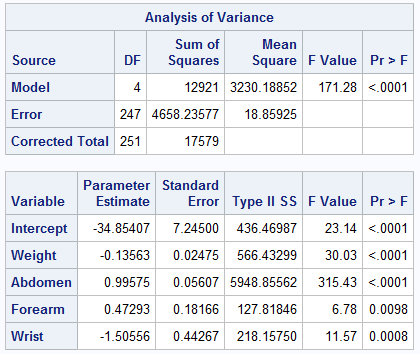


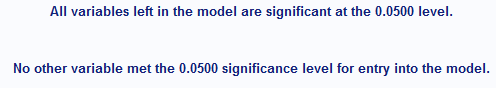
* Click .
* Do not replace the results from the previous run.

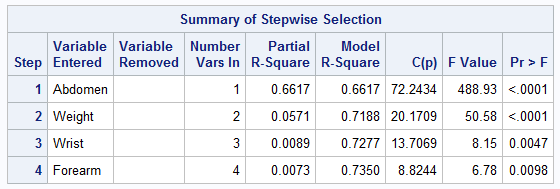
Partial Output



Skip to the last step in the stepwise selection process:



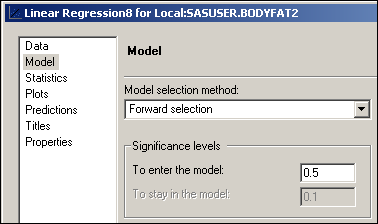




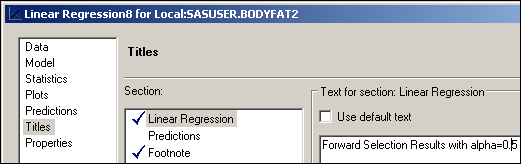
* + 1. Which variables were included in the final model produced with stepwise selection?

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

* 1. 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**.

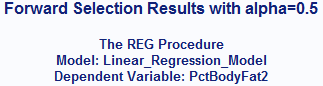


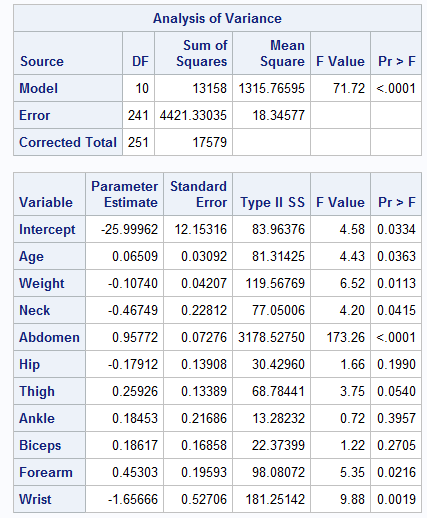
* With **Titles** selected at the left, type **Forward Selection Results with alpha=0.5** in the text area.



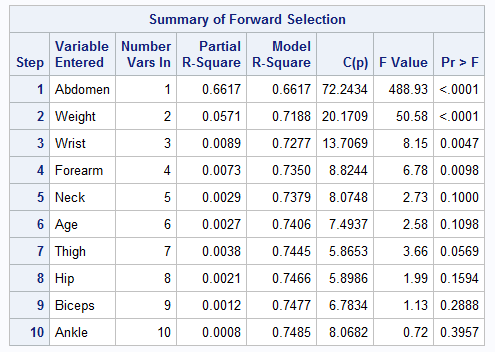
* Click .
* Do not replace the results from the previous run.

Partial Output



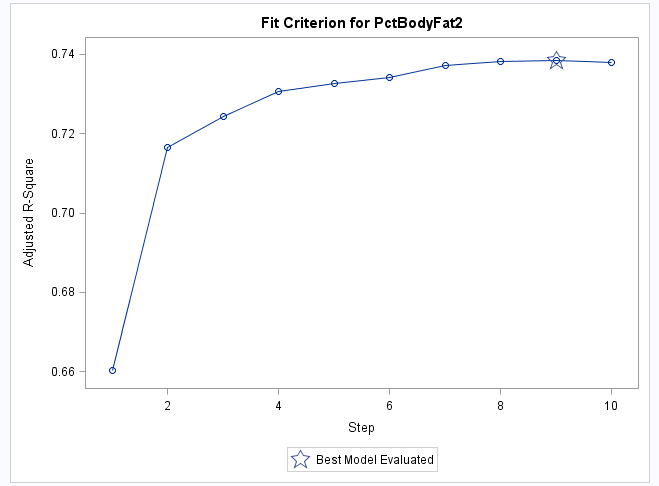






* + 1. How many variables would have resulted from a model using forward selection and a significance level for entry criterion of 0.50 (the default), instead of 0.05?

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.