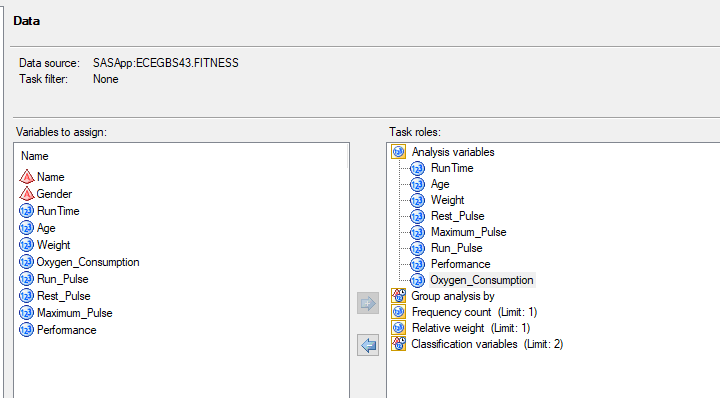
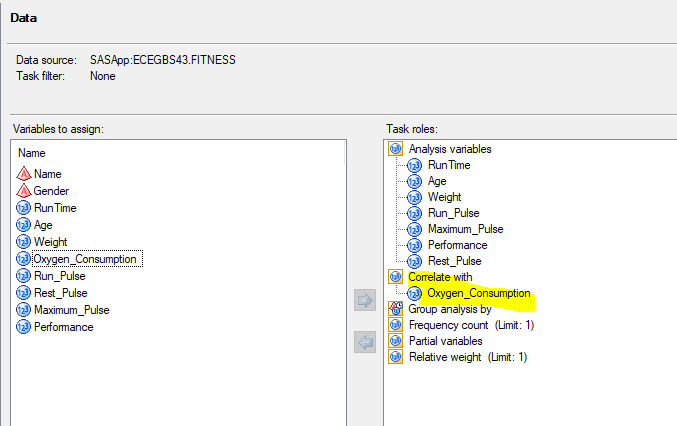
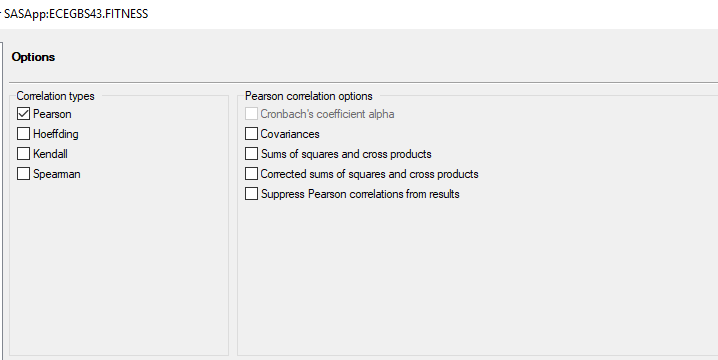
**Portions Copyright (c) 2015 SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC, USA. SAS Institute Inc, makes no warranties with respect to these materials and disclaims all liability therefor.**

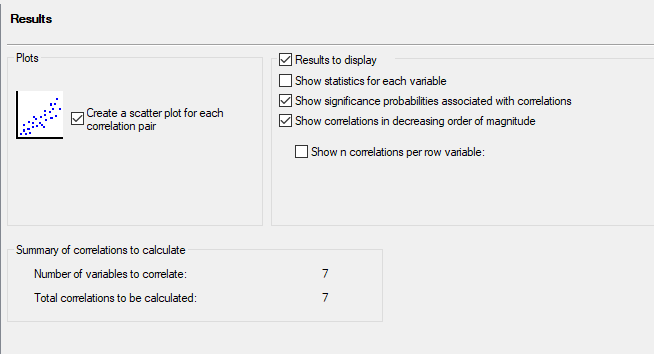
**Exercise 10: Regression and Correlation**

**Part 1: Correlation**

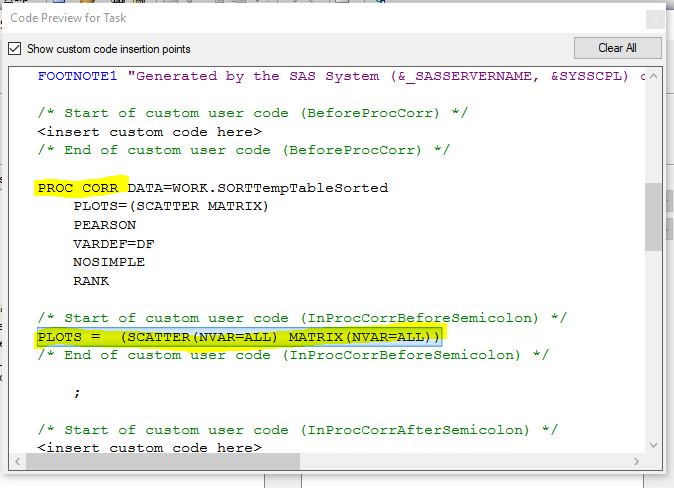
1. Start a new project and select the **Fitness** data set.
2. Run the **Distribution Analysis** to investigate the normality of the distribution of the numeric variables.  
   
3. Select **Analyze->Multivariate->Correlations.** 
   1. Assign **Runtime, Age, Weight, Run\_Pulse, Rest-Pulse, Maximum\_Pulse** and **Performance** as the analysis variables and **Oxygen\_Consumption** as the role to correlate with.



* 1. In **Options**, note that the type of correlation coefficient generated is **Pearson:**
  2. In **Results,** 
     1. check the boxes for **Create a scatter plot for each correlation pair** and **Show correlations in decreasing order of magnitude**
     2. uncheck **Show statistics for each variable**.



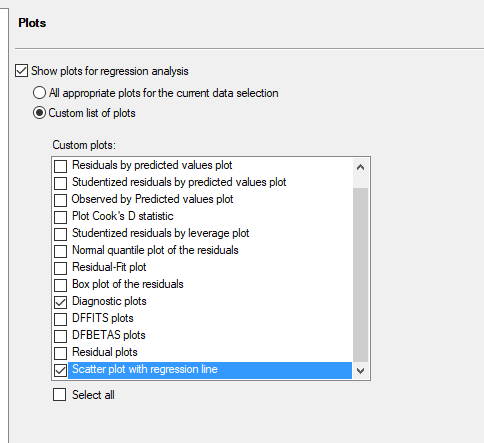
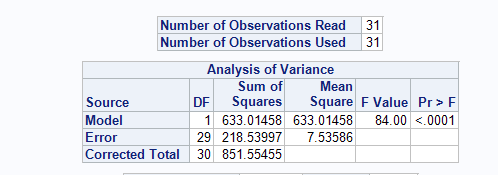
* 1. By default, there will be a maximum of five scatter plots, but since we have seven variables, we will have to modify the code:
     1. Click **Preview Code** and check **Show custom code insertion points**
     2. Add the statement shown below:  
        PLOTS = (SCATTER(NVAR=ALL) MATRIX(NVAR=ALL))



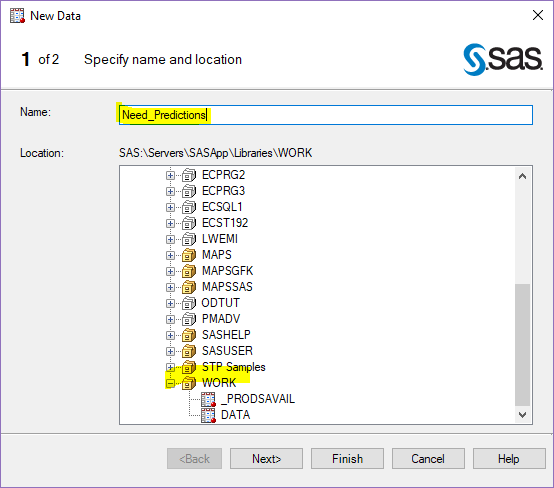
* 1. Close the code window and click **Run** to run the correlations task and examine the results.

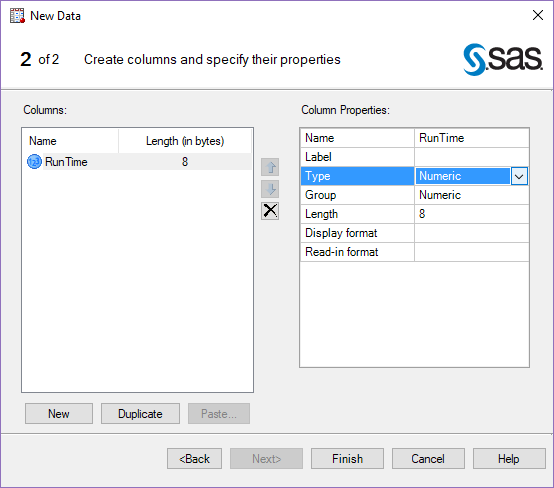
1. To check if there are any correlations between the predictor (x-) variables: reopen the previous task by right-clicking and selecting **Modify**
   1. Remove **Oxygen­\_Consumption** from the Task roles list.
   2. Click **Run**, but do not replace the results from the previous run.
   3. Identify the pairs of variables with have the strongest correlations.

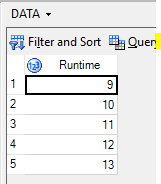
**Part 2: Performing Simple Linear Regression**

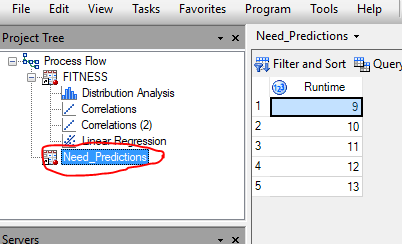
1. With the **Fitness** data set selected, select **Analyze-> Regression-> Linear Regression.**
   1. Drag **Oxygen\_Consumption** as the dependent variable and **RunTime** as the explanatory variable.
   2. With **Plots** selected, select **Custom Lists of Plots** and select **Scatter Plot for regression line.** 
   3. Click **Run**
2. Look at the results:  
     
   The analysis of variance provides an analysis of the variability observed in the data and the variability explained by the regression line.

**Part 3: Producing Predicted Values**

1. In the same project, select **New->Data**
   1. Enter the name as **Need\_Predictions** and select the WORK library  
      
   2. Create only one column and call it **Runtime**, of type Numeric.

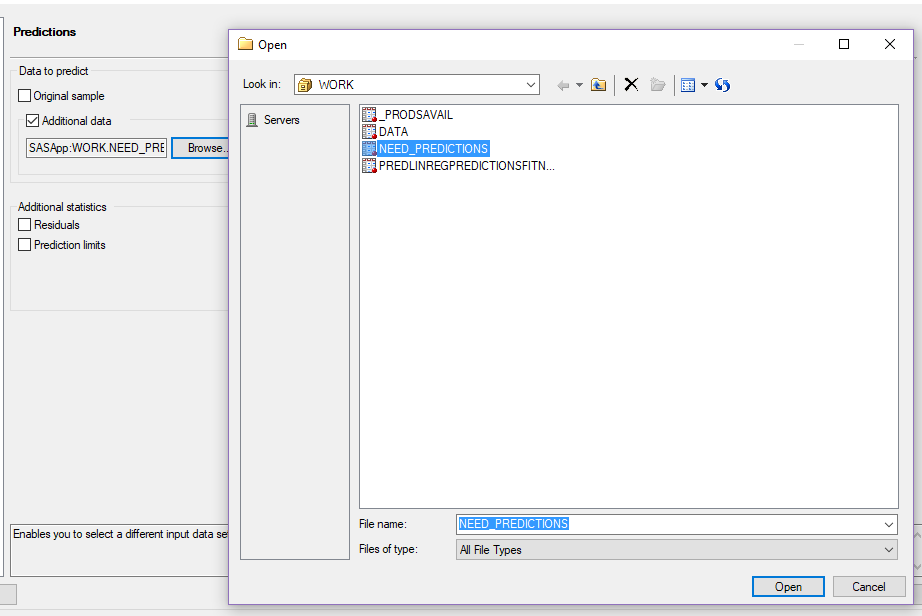


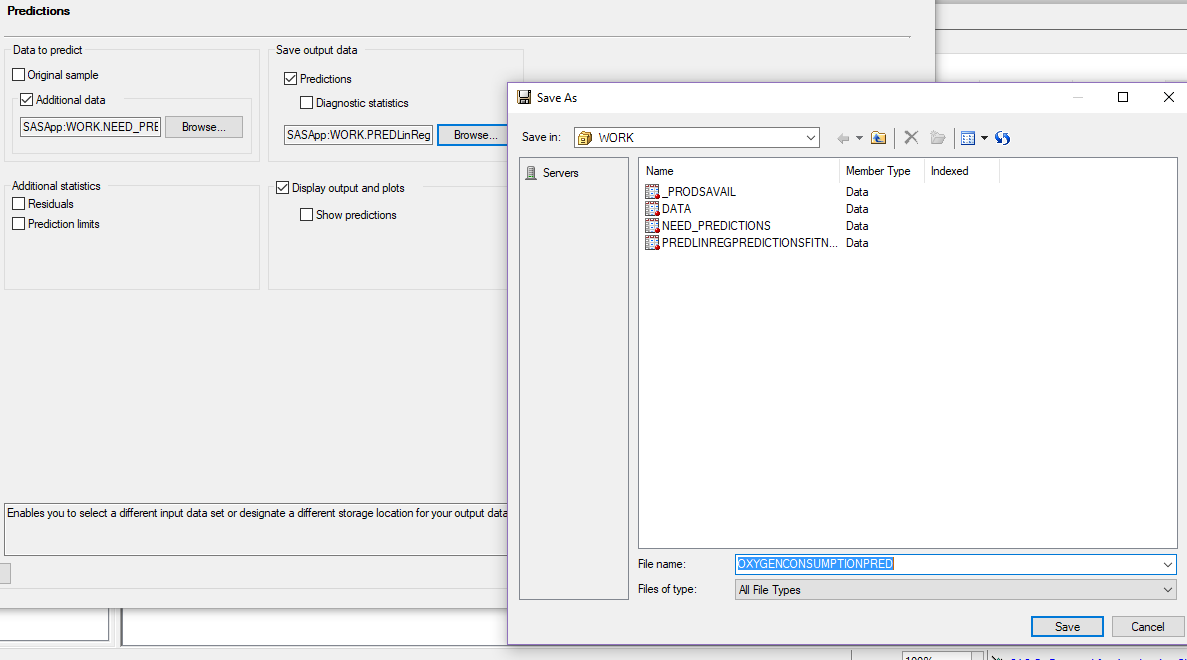
* 1. Click OK, then add the data values 9, 10, 11, 12, 13 and delete the extra rows.  
     



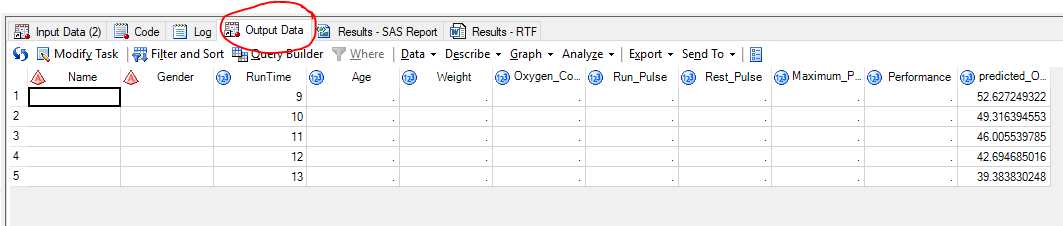
* 1. Select **Export** and make sure the data is saved in the WORK library.

1. Now select the previous **Linear Regression** task and select **Modify Task.**
   1. Select **Additional data** under **Data to predict** and select the **NEED\_PREDICTIONS** data.



* 1. Select **Save output data** and enter a new filename OXYGENCONSUMPTIONPRED in WORK   
     
  2. Under Properties, select Edit and change the label to **Predictions.**
  3. Click RUN, and do **not** replace the previous run.

1. The results will be shown under the **output data.** Scroll to the right to see the predicted oxygen consumption values.



Homework:

1. Describing the relationship between continuous variables. Percentage of body fat, age, weight, height and 10 body circumference measurements were recorded for 252 men. The data are stored in the **BodyFat2** data set. Body fat, one measure of health, was accurately estimated by an underwater weighing technique. There are two measures of percentage body fat in this data set. The variables are:

|  |  |
| --- | --- |
| **Case** | Case Number |
| **PctBodyFat1** | Percent body fat using Brozek’s equation, 457/Density – 414.2 |
| **PctBodyFat2** | Percent body fat using Siri’s equation, 459/Density – 450 |
| **Density** | Density (gm/cm3) |
| **Age** | Age (yrs) |
| **Weight** | Weight (lbs) |
| **Height** | Height (inches) |
| **Adiposity** | Adiposity index = Weight / Height2 (kg/m2) |
| **FatFreeWt** | Fat Free Weight (1 – fraction of body fat)\* Weight, using Brozek’s formula |
| **Neck** | Neck circumference (cm) |
| **Chest** | Chest circumference (cm) |
| **Abdomen** | Abdomen circumference (cm) |
| **Hip** | Hip circumference (cm) |
| **Thigh** | Thigh circumference (cm) |
| **Knee** | Ankle circumference (cm) |
| **Biceps** | Biceps circumference (cm) |
| **Forearm** | Forearm circumference (cm) |
| **Wrist** | Wrist circumference (cm) |

* 1. Generate scatter plots and correlations for the variables **Age, Weight, Height** versus the variable, **PctBodyFat2.**
  2. Generate scatter plots and correlations for the circumference measures versus the variable, **PctBodyFat2.**
  3. What variable has the highest correlation with **PctBodyFat2?**
  4. What is the value of the coefficient?
  5. Is the correlation statistically significant at the 0.05 level?
  6. Can straight lines adequately describe the relationships?
  7. Are there any outliers that you should investigate?
  8. Generate correlations among the variables (**Age, Weight, Height**). Are there any notable relationships?
  9. Generate correlations among the circumference measures. Are there any notable relationships?

1. Now perform a simple linear regression model using the **BodyFat2** data, using **PctBodyFat2** as the response variable and **Weight** as the predictor.
   1. For the ANOVA model:
      1. What is the null hypothesis?
      2. What is the value of the *F* statistic and the associated *p*-value?
      3. How would you interpret this with regard to the null hypothesis?
   2. Write the predicted regression equation as  
       Y = β0 + β1X
      1. What is the null hypothesis related to the parameter estimate β1?
      2. How would you interpret the parameter estimate β1?
      3. What is the t-value for the parameter estimate β1?
      4. How would you interpret the *p­-*value?
   3. What is the value of the R2 statistic? How would you interpret this?
   4. What is the predicted value for PctBodyFat2 when Weight is 130?
   5. Produce predicted values for PctBodyFat2 when Weight is 125, 150, 175, 200 and 225.