SASEG 7 - Introduction to Regression Analysis

Datasets – Fitness; NewFitness; BodyFat2; AbdomenPred

(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

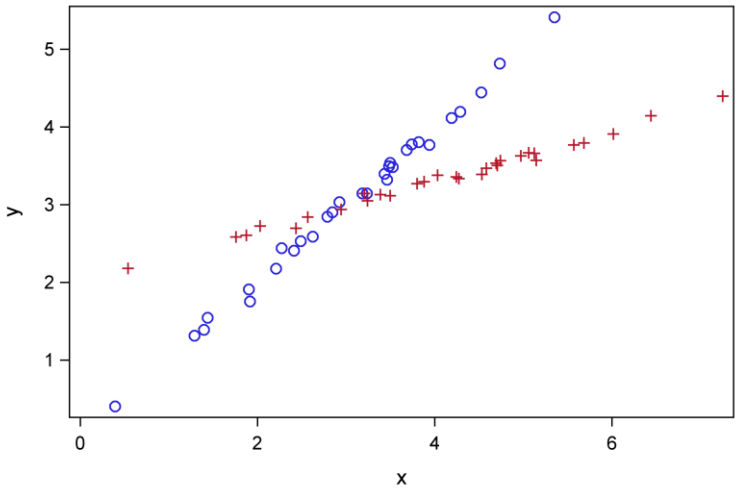
Simple Linear Regression

Objectives

- Explain the concepts of simple linear regression.
- Fit a simple linear regression using the Linear Regression task.
- Produce predicted values and confidence intervals.

23

Overview

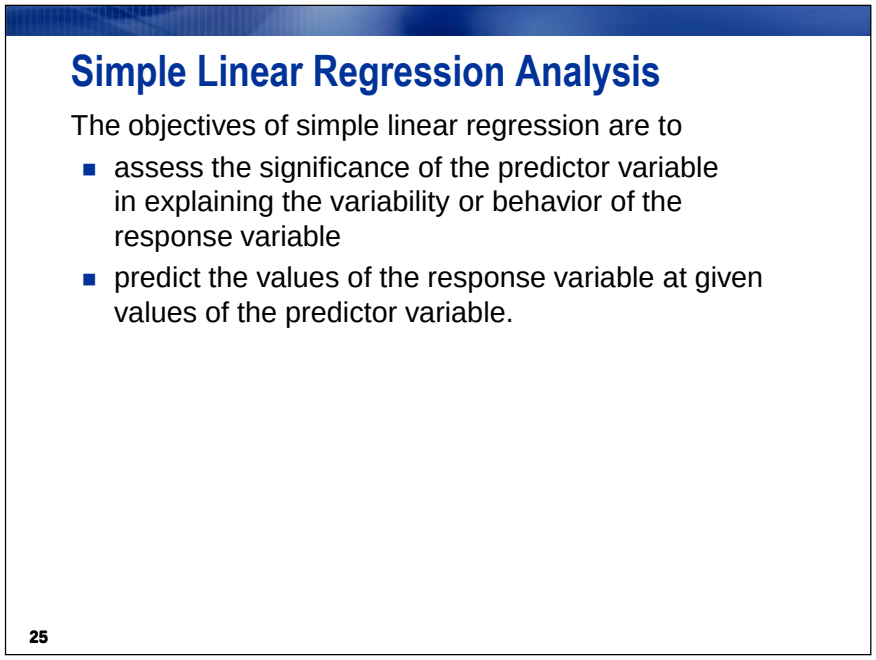


24

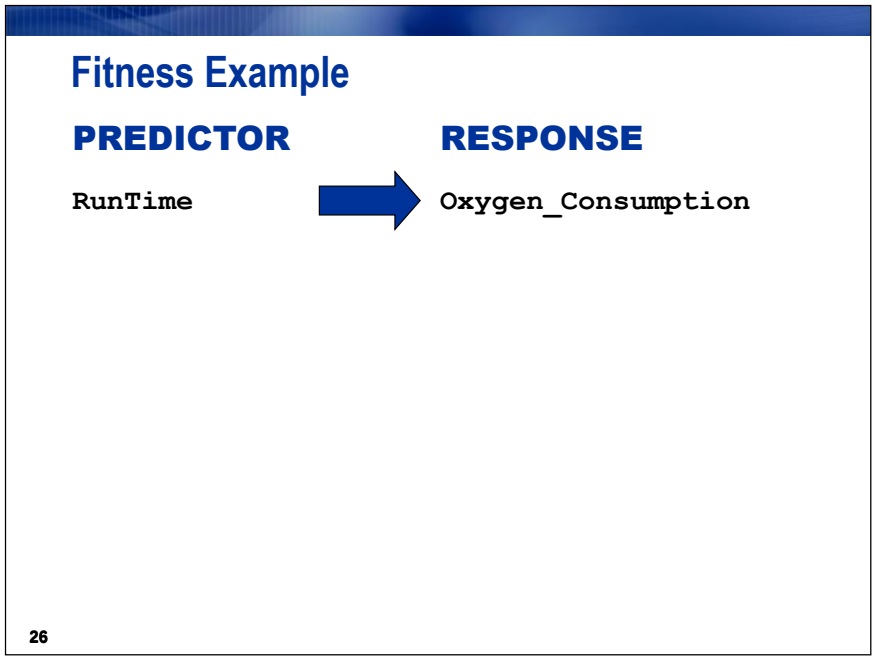
In the previous section, you used correlation analysis to quantify the linear relationships between continuous response variables. Two pairs of variables can have the same correlation, but very different linear relationships. In this section, you use simple linear regression to define the linear relationship between a response variable and a predictor variable.

The *response variable* is the variable of primary interest.

The *predictor variable* is used to explain the variability in the response variable.

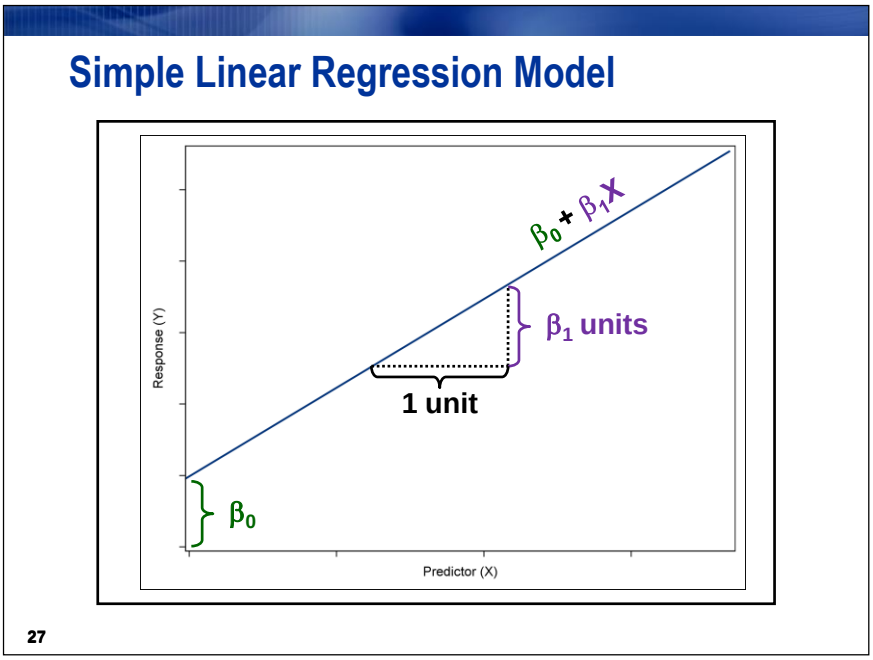


In simple linear regression, the values of the predictor variable are assumed fixed. Thus, you try to explain the variability of the response variable given the values of the predictor variable.



The analyst noted that the running time measure has the highest correlation with the oxygen consumption capacity of the club members. Consequently, she wants to further explore the relationship between **Oxygen_Consumption** and **RunTime**.

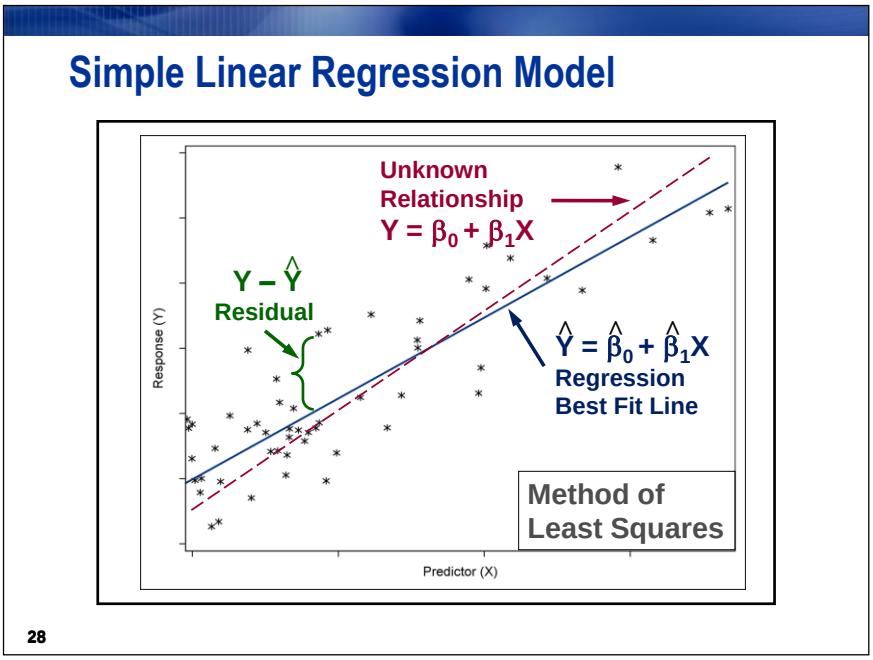
She decides to run a simple linear regression of **Oxygen_Consumption** versus **RunTime**.



The relationship between the response variable and the predictor variable can be characterized by the equation $Y = \beta_0 + \beta_1 X + \varepsilon$

where

Y	response variable
X	predictor variable
β_0	intercept parameter, which corresponds to the value of the response variable when the predictor is 0
β_1	slope parameter, which corresponds to the magnitude of change in the response variable given a one unit change in the predictor variable
ε	error term representing deviations of Y about $\beta_0 + \beta_1 X$.



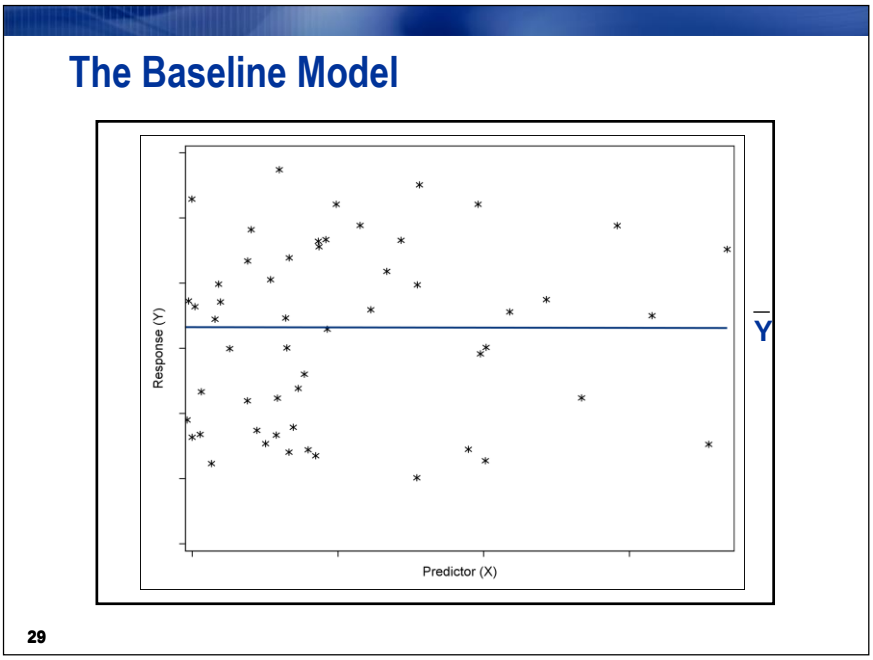
Because your goal in simple linear regression is usually to characterize the relationship between the response and predictor variables in your population, you begin with a sample of data. From this sample, you estimate the unknown population parameters (β_0, β_1) that define the assumed relationship between your response and predictor variables.

Estimates of the unknown population parameters β_0 and β_1 are obtained by the *method of least squares*. This method provides the estimates by determining the line that minimizes the sum of the squared vertical distances between the observations and the fitted line. In other words, the fitted or regression line is as close as possible to all the data points.

The method of least squares produces parameter estimates with certain optimum properties. If the assumptions of simple linear regression are valid, the least squares estimates are unbiased estimates of the population parameters and have minimum variance (efficiency). The least squares estimators are often called BLUE (Best Linear Unbiased Estimators). The term *best* is used because of the minimum variance property.

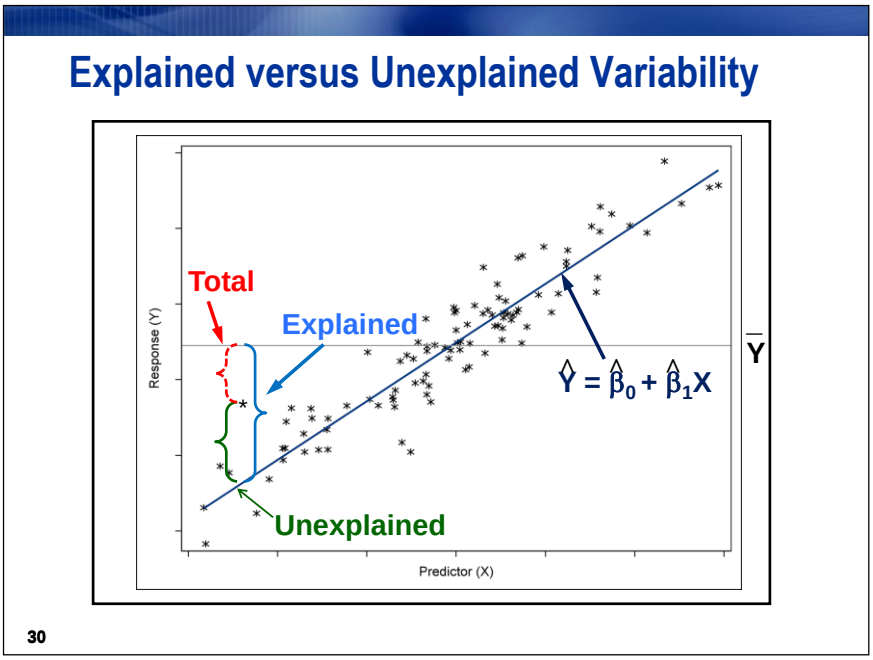
Because of these optimum properties, the method of least squares is used by many data analysts to investigate the relationship between continuous predictor and response variables.

With a large and representative sample, the fitted regression line should be a good approximation of the relationship between the response and predictor variables in the population. The estimated parameters obtained using the method of least squares should be good approximations of the true population parameters.



To determine whether the predictor variable explains a significant amount of variability in the response variable, the simple linear regression model is compared to the baseline model. The fitted regression line in a baseline model is a horizontal line across all values of the predictor variable. The slope of the regression line is 0 and the intercept is the sample mean of the response variable, ($\overline{Y}$).

In a baseline model, there is no association between the response variable and the predictor variable. Therefore, knowing the value of the predictor variable does not improve predictions of the response over simply using the mean of the response variable for everyone.



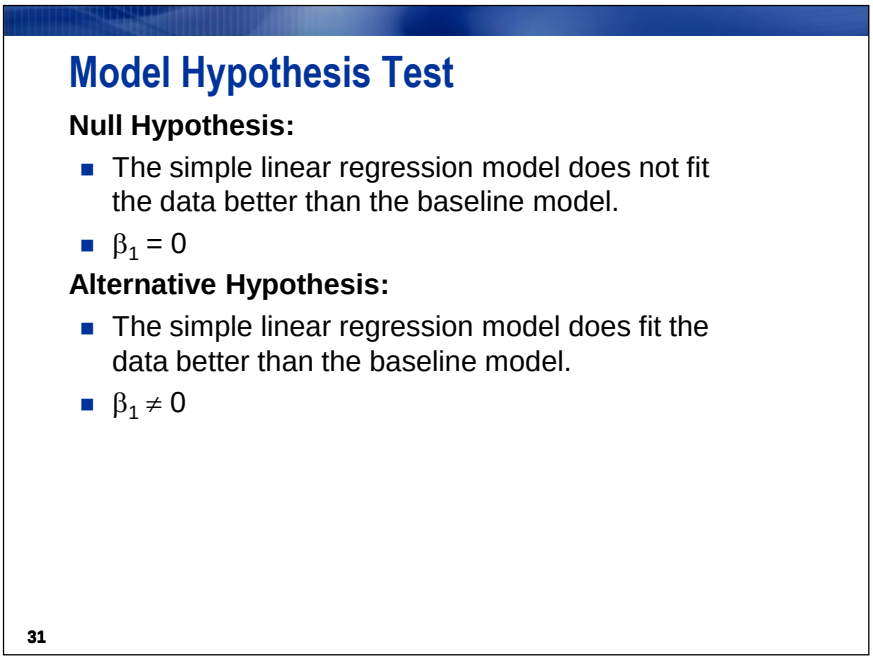
To determine whether a simple linear regression model is better than the baseline model, compare the explained variability to the unexplained variability.

Explained variability is related to the difference between the regression line and the mean of the response variable. The model sum of squares (SSM) is the amount of variability explained by your model. The model sum of squares is equal to $\sum\left(\hat{Y}_i - \bar{Y}\right)^2$.

Unexplained variability is related to the difference between the observed values and the regression line. The error sum of squares (SSE) is the amount of variability unexplained by your model. The error sum of squares is equal to $\sum\left(Y_i - \hat{Y}_i\right)^2$.

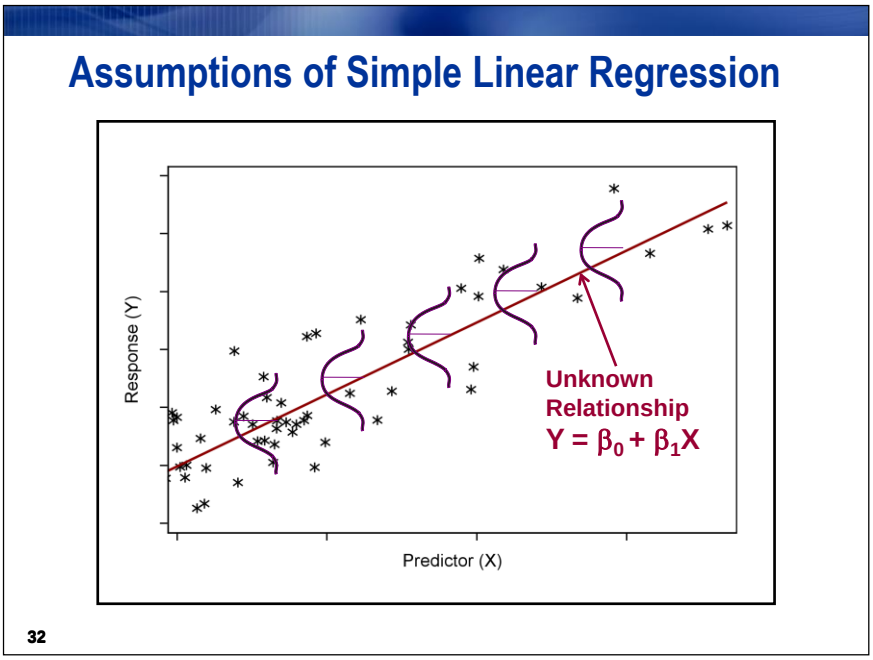
Total variability is related to the difference between the observed values and the mean of the response variable. The corrected total sum of squares is the sum of the explained and unexplained variability. The corrected total sum of squares is equal to $\sum\left(Y_i - \bar{Y}\right)^2$.

The plot shows a seemingly contradictory relationship between explained, unexplained and total variability. Contribution to total variability for the data point is smaller than contribution to explained and unexplained variability. Remember that the relationship of total=unexplained + explained holds for sums of squares over all observations and not necessarily for any individual observation.



If the estimated simple linear regression model **does not** fit the data better than the baseline model, you fail to reject the null hypothesis. Thus, you **do not** have enough evidence to say that the slope of the regression line in the population is **not** 0 and that the predictor variable explains a significant amount of variability in the response variable.

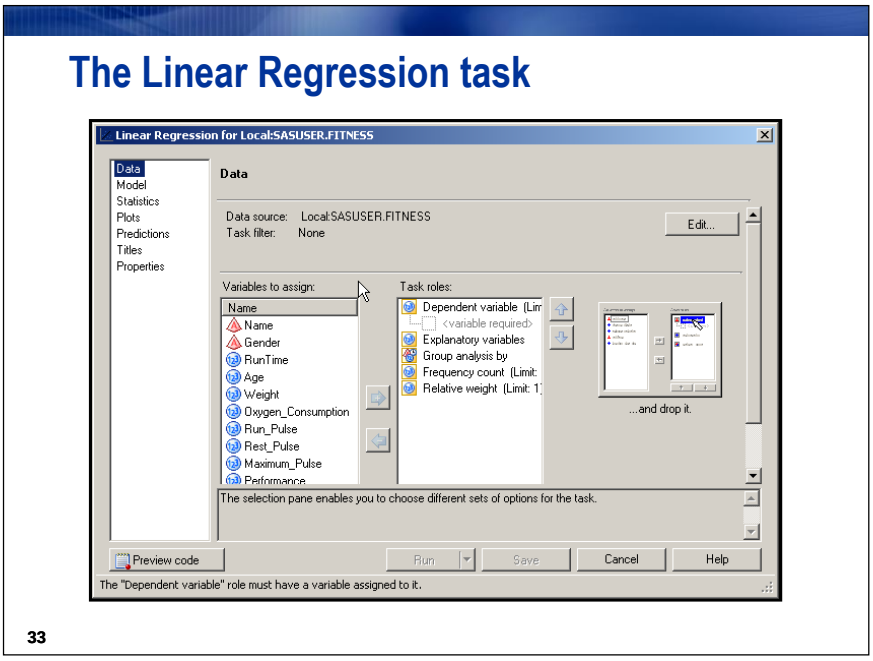
If the estimated simple linear regression model **does** fit the data better than the baseline model, you reject the null hypothesis. Thus, you **do** have enough evidence to say that the slope of the regression line in the population is not 0 and that the predictor variable explains a significant amount of variability in the response variable.



One of the assumptions of simple linear regression is that the mean of the response variable is linearly related to the value of the predictor variable. In other words, a straight line connects the means of the response variable at each value of the predictor variable.

The other assumptions are the same as the assumptions for ANOVA: the error terms are normally distributed, have equal variances, and are independent.

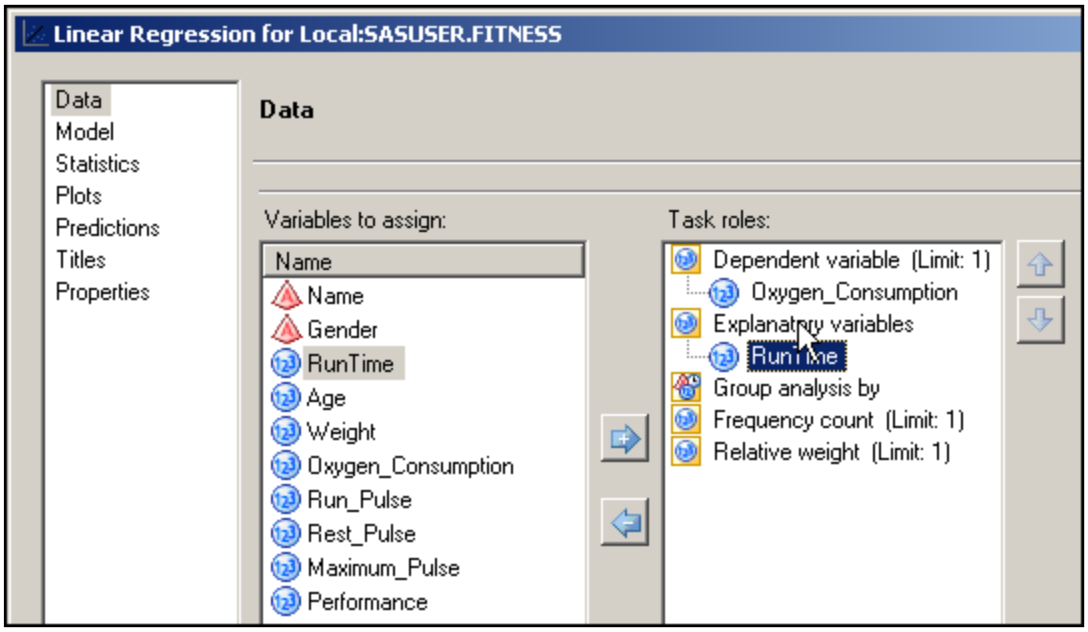
The verification of these assumptions is discussed in a later chapter.



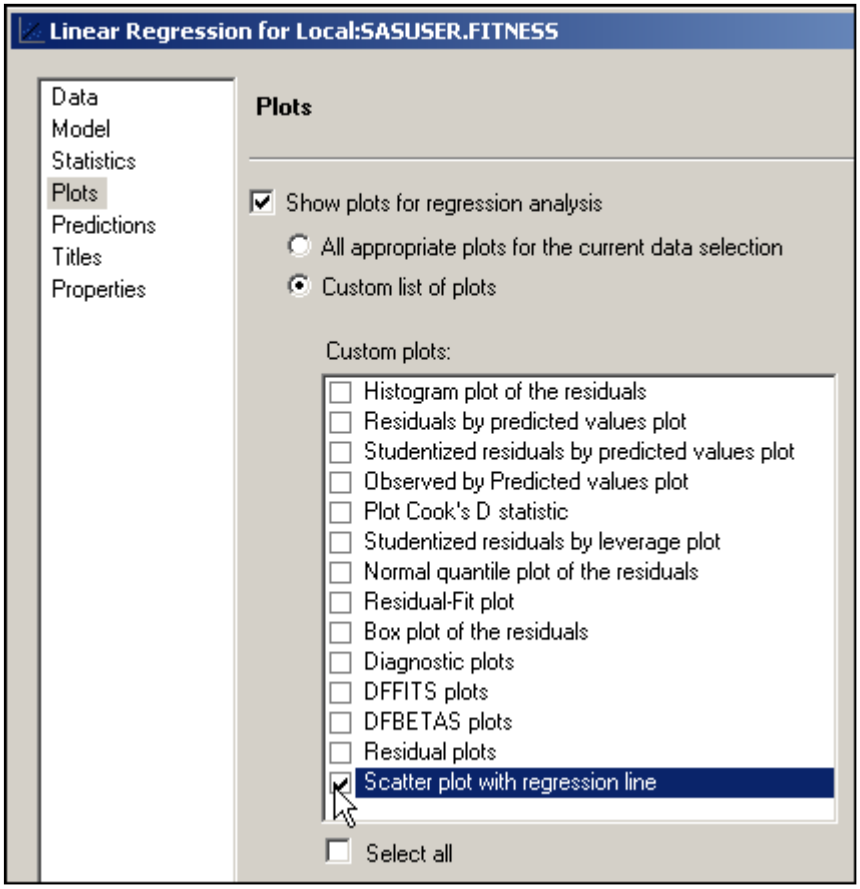
Performing Simple Linear Regression with SAS EG

Because there is an apparent linear relationship between **`Oxygen_Consumption`** and **`RunTime`**, perform a simple linear regression analysis with **`Oxygen_Consumption`** as the response variable.

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



3. With **Plots** selected at the left, select **Custom list of plots** under `Show plots for regression analysis`. In the menu that appears, uncheck the box for **Diagnostic plots** and check the box for **Scatter plot with regression line**.



4. Change the title, if desired.

5. Click Run.

The Number of Observations Read and the Number of Observations Used are the same, indicating that no missing values were detected for **`Oxygen_Consumption`** and **`RunTime`**.

The Analysis of Variance (ANOVA) table provides an analysis of the variability observed in the data and the variability explained by the regression line.

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			

The ANOVA table for simple linear regression is divided into six columns.

Source — labels the source of variability.

- Model — is the variability explained by your model (Between Group).
- Error — is the variability unexplained by your model (Within Group).
- Corrected Total — is the total variability in the data (Total).

DF — is the degrees of freedom associated with each source of variability.

Sum of Squares — is the amount of variability associated with each source of variability.

Mean Square — is the ratio of the sum of squares and the degrees of freedom. This value corresponds to the amount of variability associated with each degree of freedom for each source of variation.

F Value — is the ratio of the mean square for the model and the mean square for the error. This ratio compares the variability explained by the regression line to the variability unexplained by the regression line.

Pr > F — is the *p*-value associated with the *F* value.

The *F* value tests whether the slope of the predictor variable is equal to 0. The *p*-value is small (less than .05), so you have enough evidence at the .05 significance level to reject the null hypothesis. Thus, you can conclude that the simple linear regression model fits the data better than the baseline model. In other words, **`RunTime`** explains a significant amount of variability of **`Oxygen_Consumption`**.

The third part of the output provides summary measures of fit for the model.

Root MSE	2.74515	R-Square	0.7434
Dependent Mean	47.37581	Adj R-Sq	0.7345
Coeff Var	5.79442		

R-Square — the coefficient of determination also referred to as the R^2 value. This value is

- between 0 and 1.
- the proportion of variability observed in the data explained by the regression line. In this example, the value is 0.7434, which means that the regression line explains 74% of the total variation in the response values.
- the square of the multiple correlation between y and the x's.

Notice that the R-square is the squared value of the correlation you saw earlier between **`RunTime`** and **`Oxygen_Consumption`** (0.86219). This is no

coincidence. For simple regression, the R-square value will be the square of the value of the bivariate Pearson correlation coefficient.

Root MSE — the root mean square error is an estimate of the standard deviation of the response variable at each value of the predictor variable. It is the square root of the MSE.

Dependent Mean — the overall mean of the response variable, $\overline{Y}$.

Coeff Var — the coefficient of variation is the size of the standard deviation relative to the mean. The coefficient of variation is

- calculated as $\left(\frac{Root\,MSE}{\overline{Y}}\right) * 100$
- a unitless measure, so it can be used to compare data that has different units of measurement or different magnitudes of measurement.

Adj R-Sq — the adjusted R^2 is adjusted for the number of parameters in the model. This statistic is useful in multiple regression and is discussed in a later section.

The Parameter Estimates table defines the model for your data.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	82.42494	3.85582	21.38	<.0001
RunTime	1	-3.31085	0.36124	-9.17	<.0001

DF — represents the degrees of freedom associated with each term in the model.

Parameter Estimate — is the estimated value of the parameters associated with each term in the model.

Standard Error — is the standard error of each parameter estimate.

t Value — is the *t* statistic, which is calculated by dividing the parameter estimates by their corresponding standard error estimates.

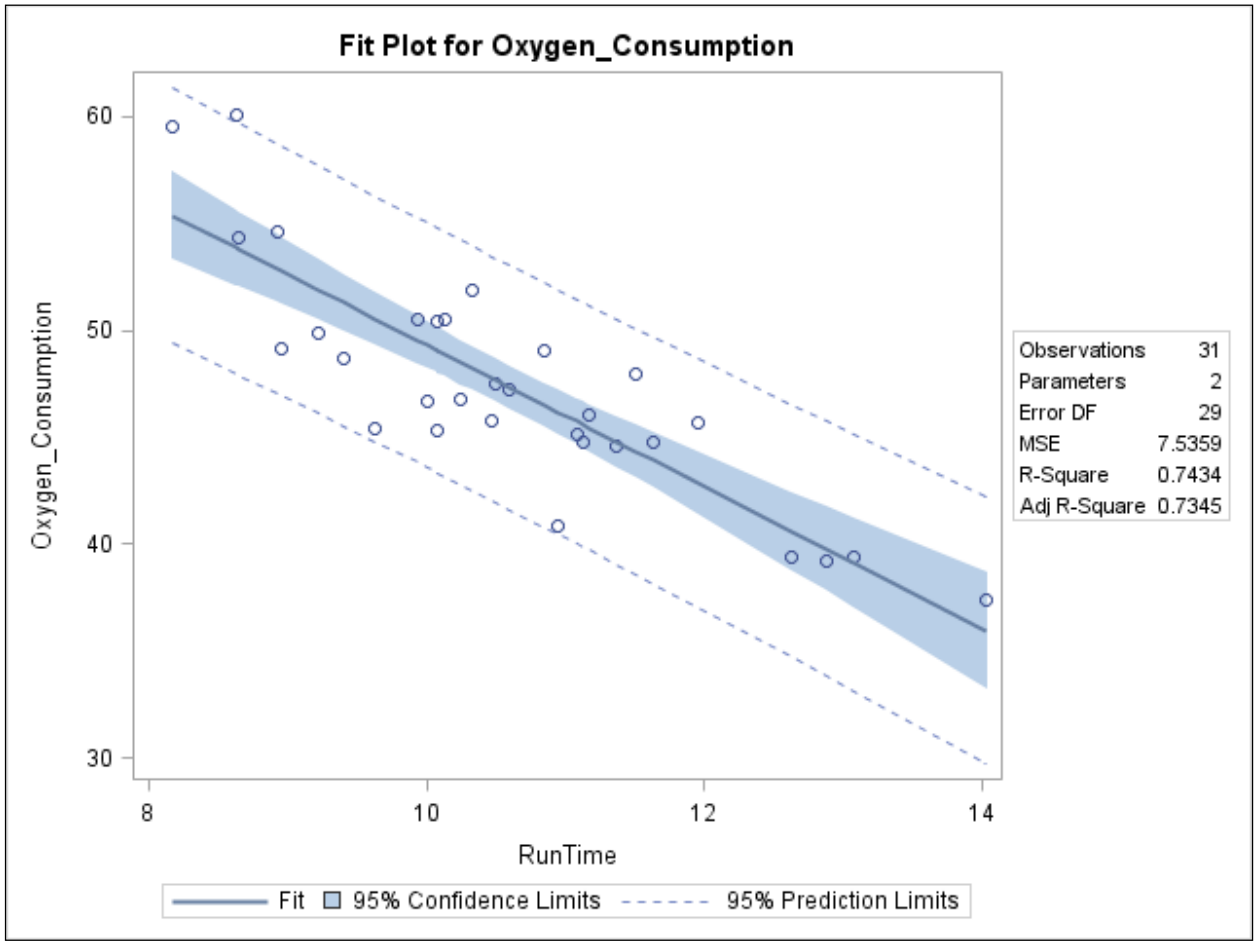
Pr > |t| — is the *p*-value associated with the *t* statistic. It tests whether the parameter associated with each term in the model is different from 0. For this example, the slope for the predictor variable is statistically different from 0. Thus, you can conclude that the predictor variable explains a significant portion of variability in the response variable.

Because the estimate of β_0=**82.42494** and β_1=**-3.31085**, the estimated regression equation is given by Predicted **`Oxygen_Consumption`** = 82.42494 - 3.31085 *(**`RunTime`**).

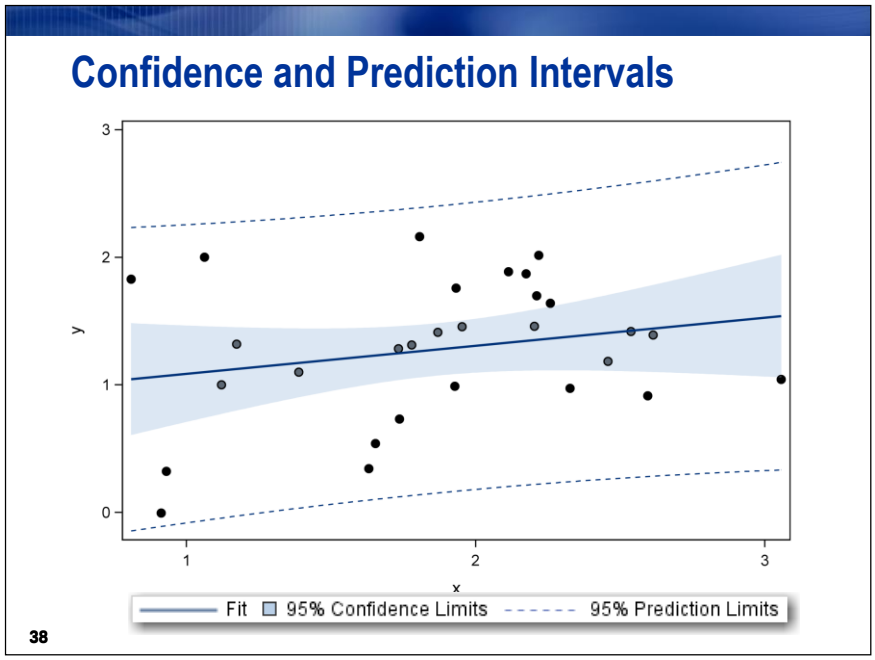
Interpretation

The model indicates that an increase of one unit for **Runtime** amounts to a 3.31085 decrease in **Oxygen_Consumption**. However, this equation is appropriate only in the range of values you observed for the variable **RunTime**.

The parameter estimates table also shows that the intercept parameter is not equal to 0. However, the test for the intercept parameter only has practical significance when the range of values for the predictor variable includes 0. In this example, the test could not have practical significance because **RunTime**=0 (running at the speed of light) is not inside the range of observed values.



The Fit Plot produced by ODS Graphics shows the predicted regression line superimposed over a scatter plot of the data. You will learn more about this plot next.



To assess the level of precision around the mean estimates of **`Oxygen_Consumption`**, you can produce **confidence intervals around the means**. This is represented in the shaded area in the plot.

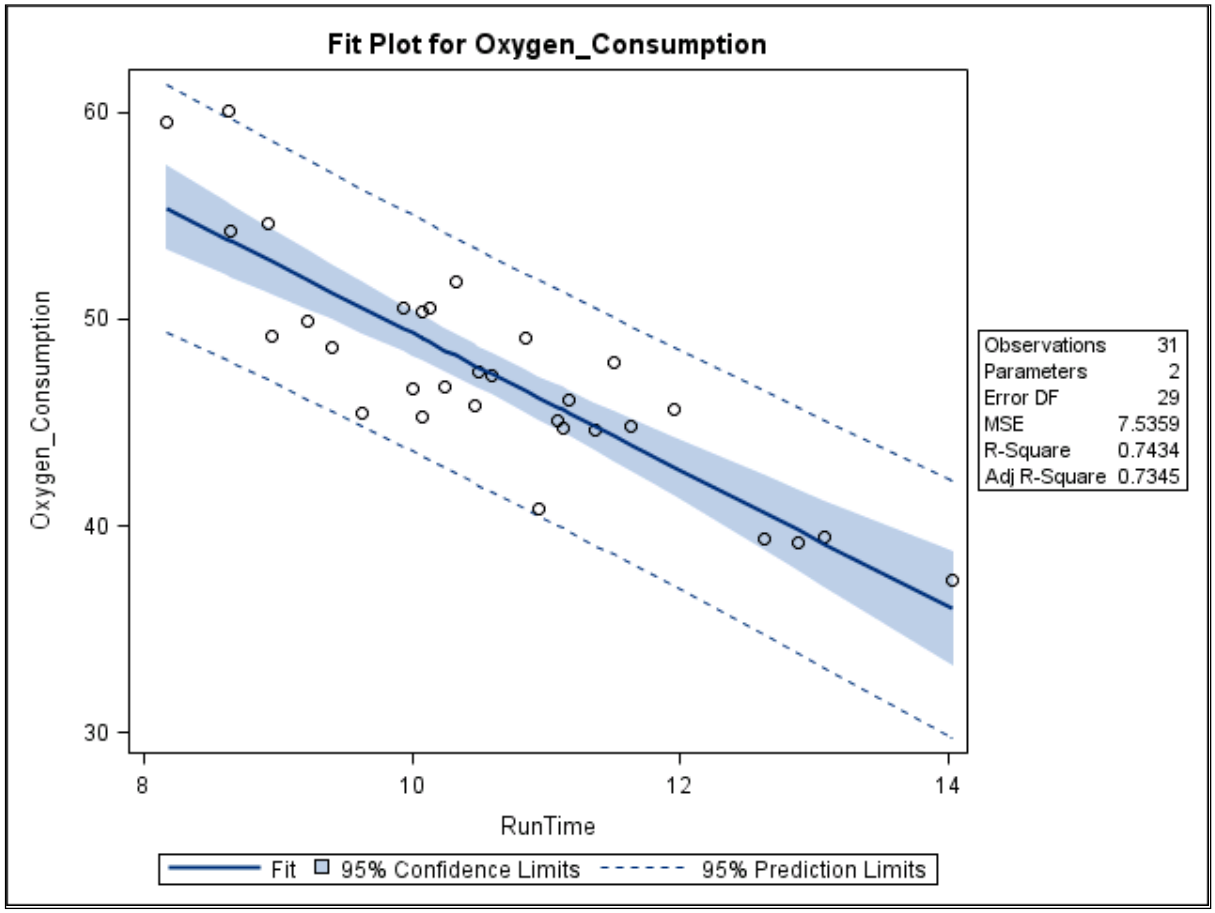
- A 95% confidence interval for the mean says that you are 95% confident your interval contains the population mean of Y for a particular X.
- Confidence intervals become wider as you move away from the mean of the independent variable. This reflects the fact that your estimates become more variable as you move away from the means of X and Y.

Suppose that the mean **`Oxygen_Consumption`** at a fixed value of **`Performance`** is not the focus. If you are interested in establishing an inference on a future single observation, you need a **prediction interval around the individual observations**. This is represented by the area between the broken lines in the plot.

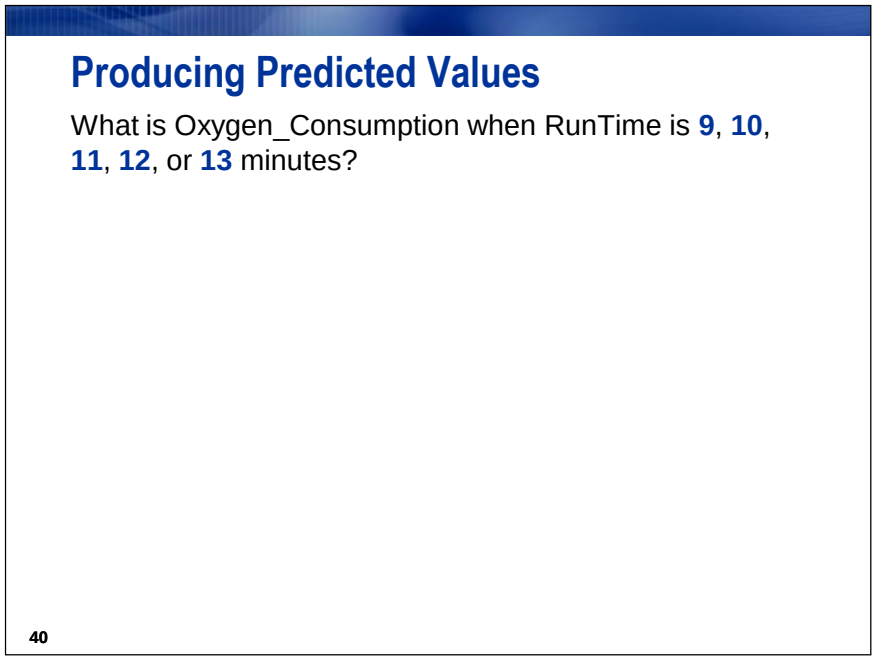
- A 95% prediction interval is one that you are 95% confident will contain a new observation.
- Prediction intervals are wider than confidence intervals because single observations have more variability than sample means.

Regression Lines with Confidence Intervals

Return to the output from the last demonstration and open the Fit Plot.



The Confidence Interval for the mean is represented by the shaded region. The Prediction Interval for observations is the area between the dotted lines. Model statistics are reported in the inset by default.

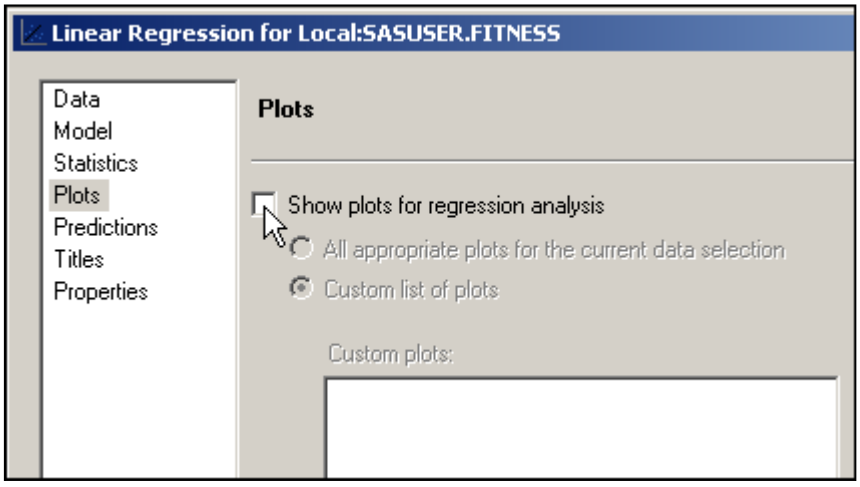


One objective in regression analysis is to predict values of the response variable given values of the predictor variables. You can obviously use the estimated regression equation to produce predicted values, but if you want a large number of predictions, this can be cumbersome.

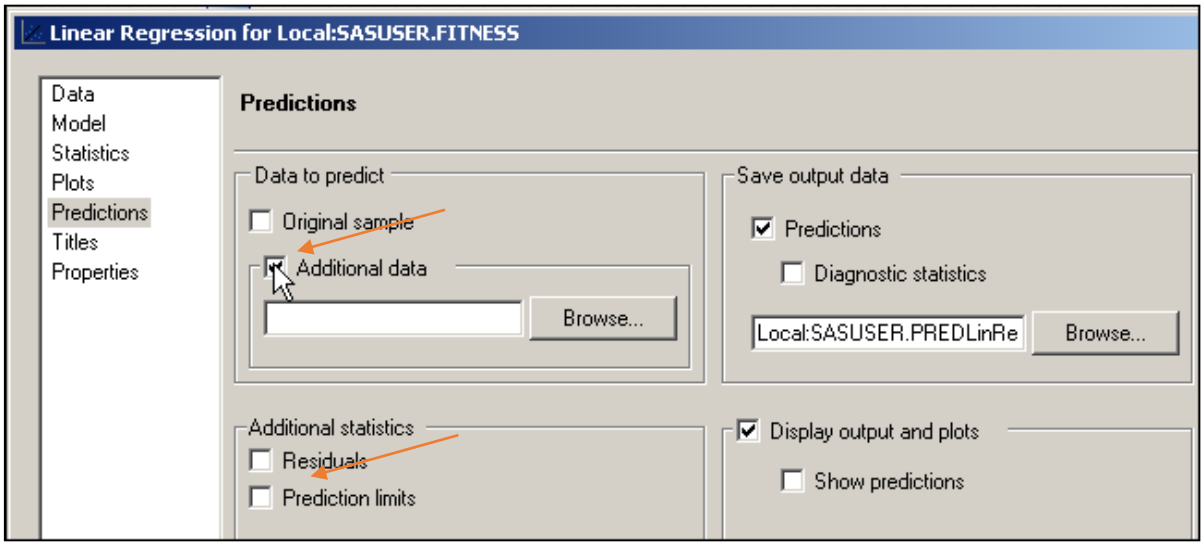
Producing Predicted Values

Produce predicted values of **`Oxygen_Consumption`** when **`Performance`** is 9, 10, 11, 12, or 13.

1. Modify the previous Linear Regression task in the project.
2. Uncheck the box for **Show plots for regression analysis**.



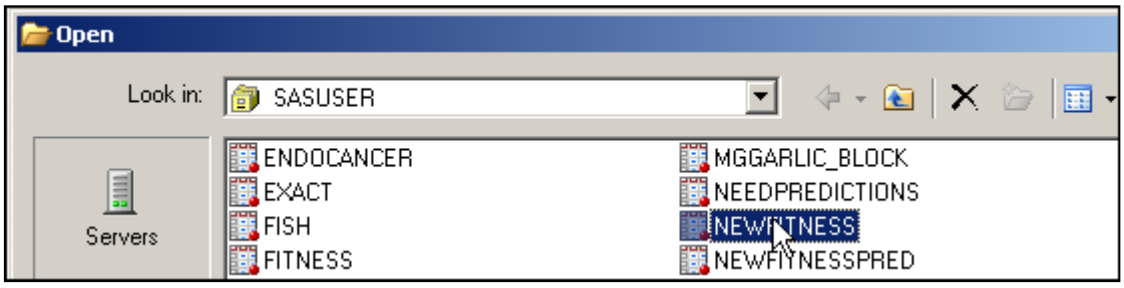
3. With **Predictions** selected at the left, check the box for **Additional data** *and* ***Prediction limits*** (to generate prediction Confidence Intervals.



4. Click 

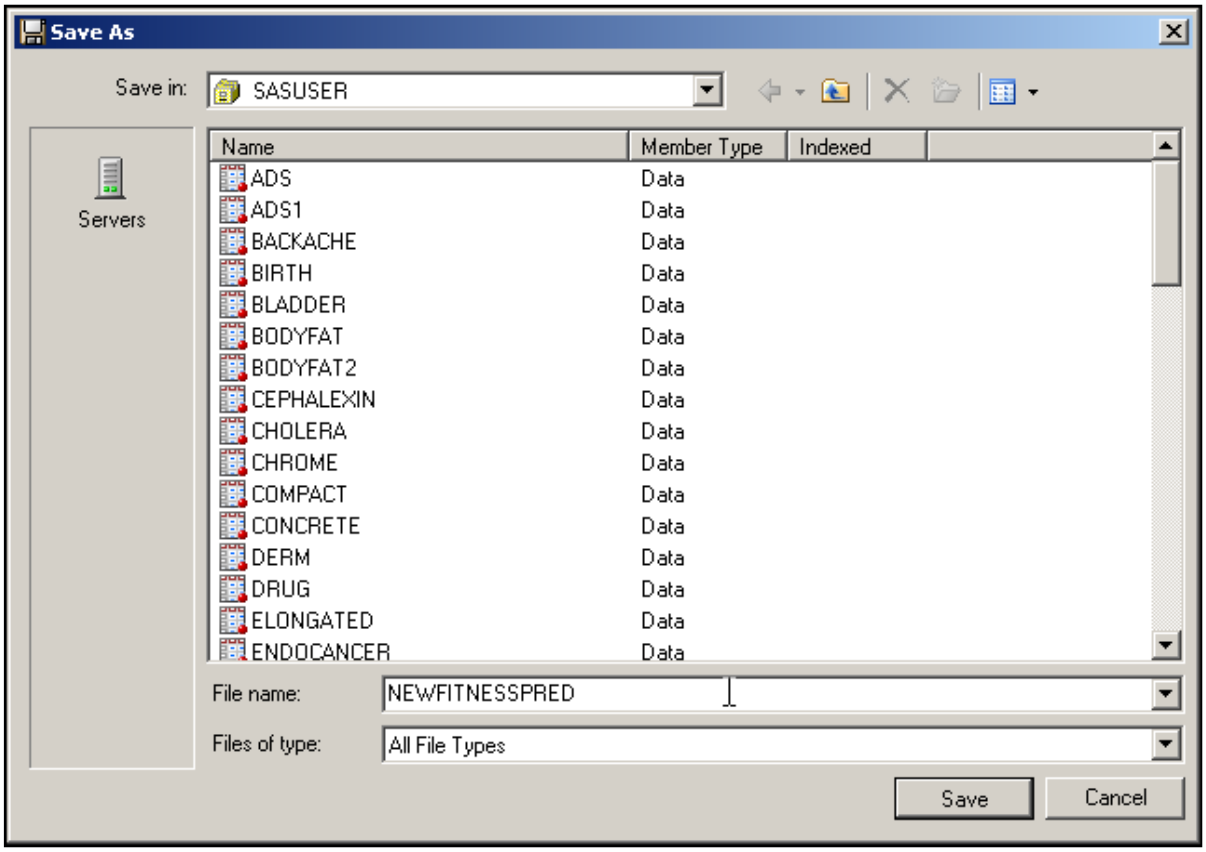
.

5. When the next window opens, select the **NewFitness** data set from the SASUSER library. Either double-click the name of the data set or highlight it and click Open.



6. Under `Save output data`, with **Predictions** checked, click Browse....

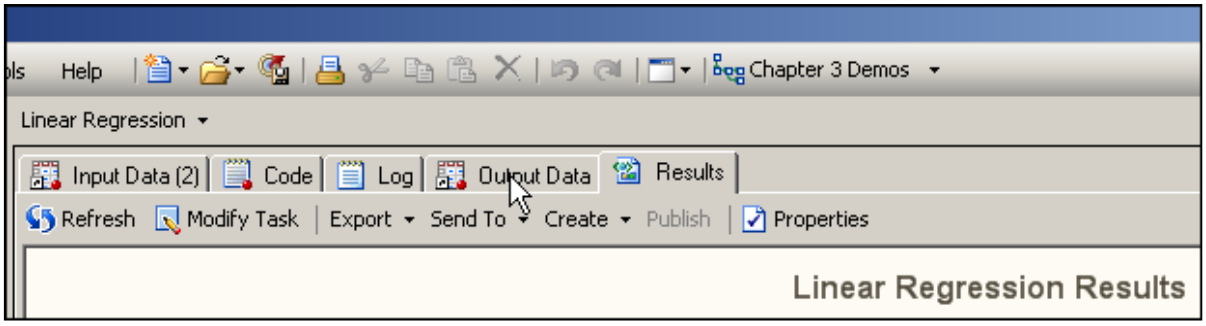
7. With that window open, overtype the default file name with **`NEWFITNESSPRED`**.



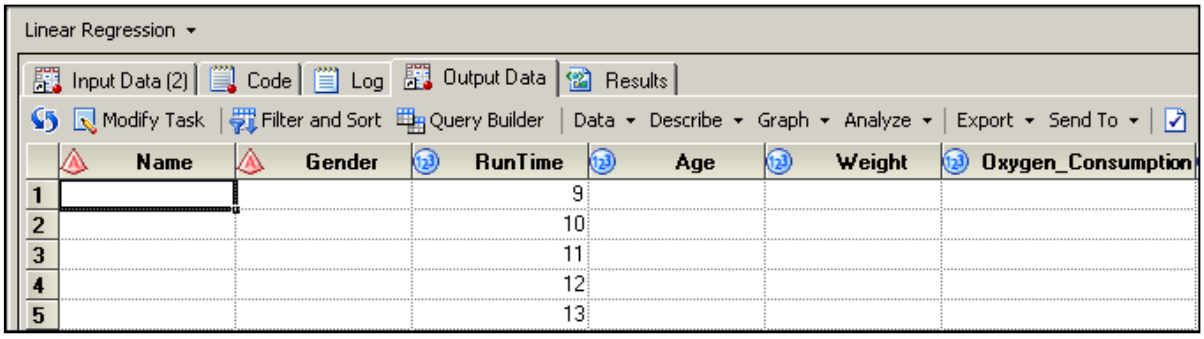
8. Click Save to close that window.

9. Click Run.

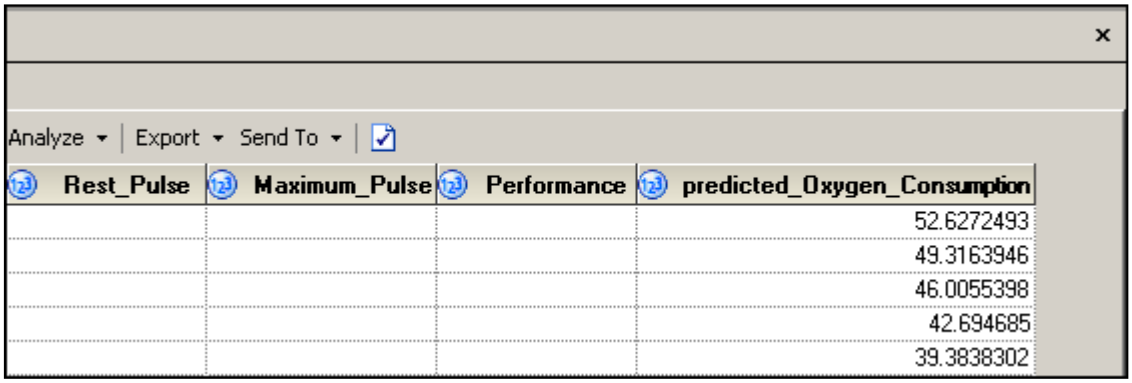
In the workspace, you will now see a tab for the newly created data set, **NEWFITNESSPRED**.



10. Click that tab to reveal the data.



11. Scroll all the way to the right to see the predictions column.



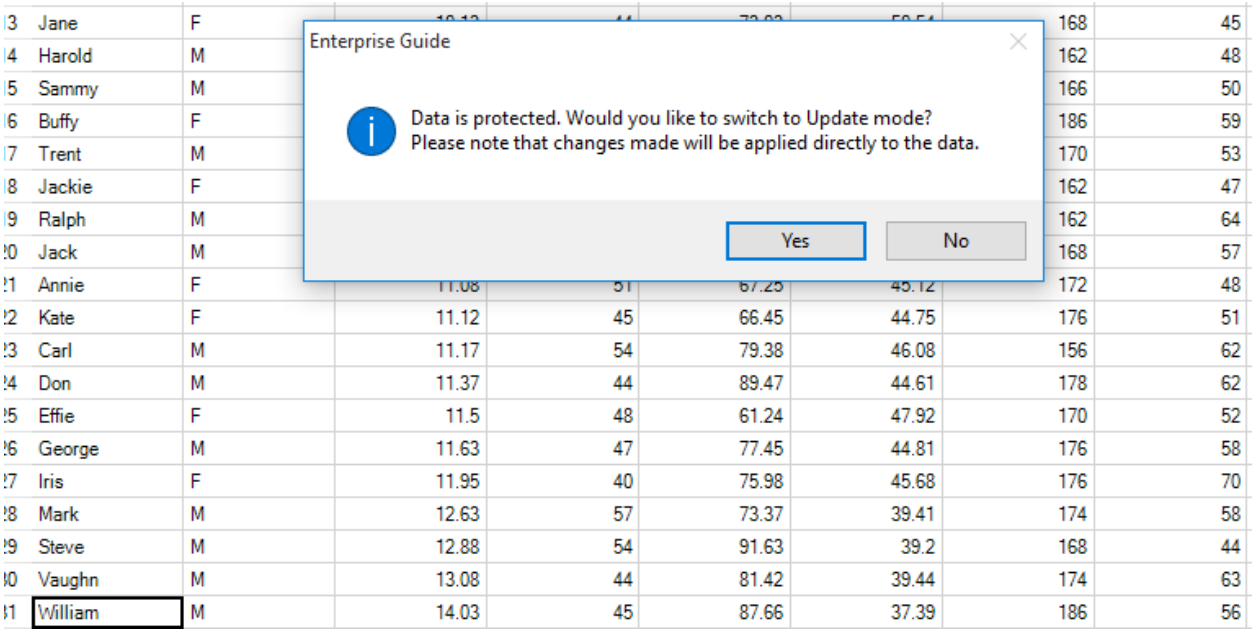
The new data set contains columns for all variables in the analysis data set, but the values of each record are set to missing for those variables that either were missing or did not exist in the scoring data set. *Also, note the 95% confidence limit for the prediction.*

Choose only values within or near the range of the predictor variable when you are predicting new values for the response variable. For this example, the values of the variable **RunTime** range from 8.17 to 14.03 minutes. Therefore, it is unwise to predict the value of **Oxygen_Consumption** for a **RunTime** of 18. The reason is that the relationship between the predictor variable and the response variable might be different beyond the range of your data.

Producing Predicted Values – The quick and easy way

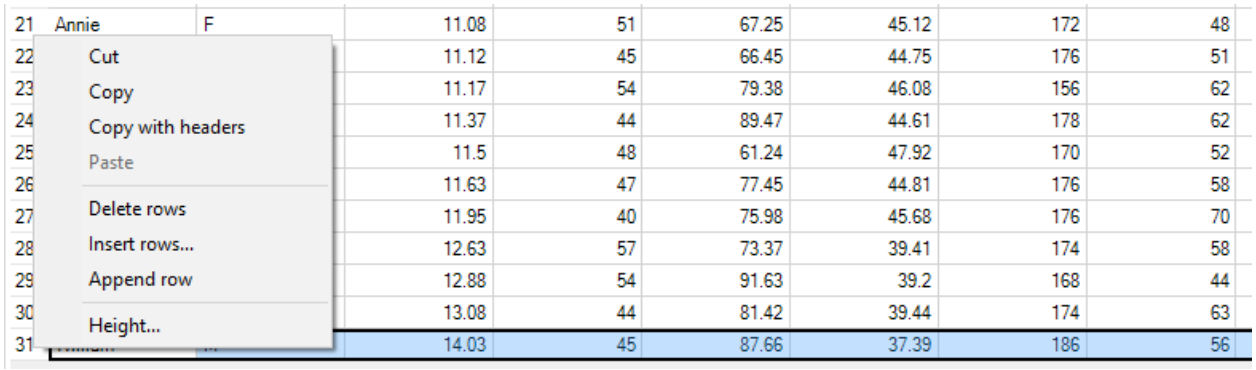
Similar to the previous method, we will produce predicted values of **`Oxygen_Consumption`** when **`run time`** is 9, 10, 11, 12, or 13. However, this time, we will be manipulating the original input data rather than adding another file to generate these predictions.

1. Click on the input data tab of the previous Linear Regression task in the project.
2. Scroll down to the last row of data and double click on the first column of the last row. You will see a message as seen in the screenshot below.



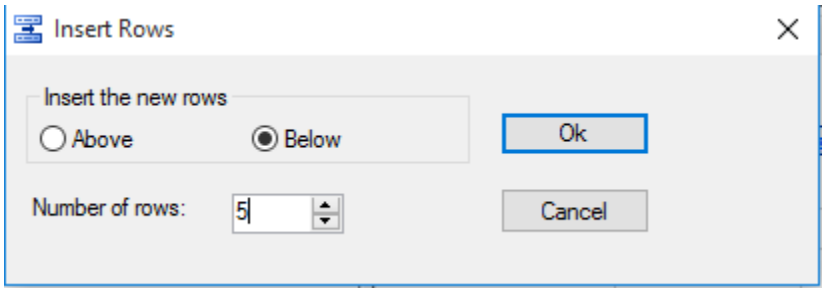
3	Jane	F					168	45
4	Harold	M					162	48
5	Sammy	M					166	50
6	Buffy	F					186	59
7	Trent	M					170	53
8	Jackie	F					162	47
9	Ralph	M					162	64
10	Jack	M					168	57
11	Annie	F	11.08	51	67.25	45.12	172	48
12	Kate	F	11.12	45	66.45	44.75	176	51
13	Carl	M	11.17	54	79.38	46.08	156	62
14	Don	M	11.37	44	89.47	44.61	178	62
15	Effie	F	11.5	48	61.24	47.92	170	52
16	George	M	11.63	47	77.45	44.81	176	58
17	Iris	F	11.95	40	75.98	45.68	176	70
18	Mark	M	12.63	57	73.37	39.41	174	58
19	Steve	M	12.88	54	91.63	39.2	168	44
20	Vaughn	M	13.08	44	81.42	39.44	174	63
21	William	M	14.03	45	87.66	37.39	186	56

3. Click on the Yes button and the data will switch to the update mode.
4. Now right click on the row marker for the last row and you will see a menu tab appear as seen below.



21	Annie	F	11.08	51	67.25	45.12	172	48
22			11.12	45	66.45	44.75	176	51
23			11.17	54	79.38	46.08	156	62
24			11.37	44	89.47	44.61	178	62
25			11.5	48	61.24	47.92	170	52
26			11.63	47	77.45	44.81	176	58
27			11.95	40	75.98	45.68	176	70
28			12.63	57	73.37	39.41	174	58
29			12.88	54	91.63	39.2	168	44
30			13.08	44	81.42	39.44	174	63
31			14.03	45	87.66	37.39	186	56

5. Click on the Insert Rows option. We will be inserting rows below the last row, so we will select that radio button for that option. We can insert any number of rows. In this case we will be inserting 5 rows i.e., based on the number of values that you are going to predict.



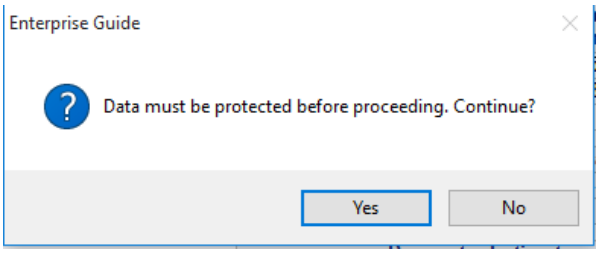
6. We are predicting the dependent variable that is oxygen consumption using the independent variable run time(for which the linear regression model was developed).

7. Double click on the first empty cell for RunTime and enter the value (s) for which the prediction will be made.

	Name	Gender	RunTime	Age	Weight	Oxygen_Consumption	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance
11	Bob	M	10.07	40	75.07	45.31	185	62	185	79
12	Harriett	F	10.08	49	73.37	50.39	168	67	168	57
13	Jane	F	10.13	44	73.03	50.54	168	45	168	67
14	Harold	M	10.25	48	91.63	46.77	162	48	164	61
15	Sammy	M	10.33	54	83.12	51.85	166	50	170	49
16	Buffy	F	10.47	52	73.71	45.79	186	59	188	47
17	Trent	M	10.5	52	82.78	47.47	170	53	172	51
18	Jackie	F	10.6	47	79.15	47.27	162	47	164	56
19	Ralph	M	10.85	43	81.19	49.09	162	64	170	65
20	Jack	M	10.95	51	69.63	40.84	168	57	172	48
21	Annie	F	11.08	51	67.25	45.12	172	48	172	43
22	Kate	F	11.12	45	66.45	44.75	176	51	176	55
23	Carl	M	11.17	54	79.38	46.08	156	62	165	40
24	Don	M	11.37	44	89.47	44.61	178	62	182	58
25	Effie	F	11.5	48	61.24	47.92	170	52	176	45
26	George	M	11.63	47	77.45	44.81	176	58	176	50
27	Iris	F	11.95	40	75.98	45.68	176	70	180	56
28	Mark	M	12.63	57	73.37	39.41	174	58	176	20
29	Steve	M	12.88	54	91.63	39.2	168	44	172	23
30	Vaughn	M	13.08	44	81.42	39.44	174	63	176	41
31	William	M	14.03	45	87.66	37.39	186	56	192	30
32			9	.	.	.	.	.	.	.
33			10	.	.	.	.	.	.	.
34			11	.	.	.	.	.	.	.
35			12	.	.	.	.	.	.	.
36			13	.	.	.	.	.	.	.

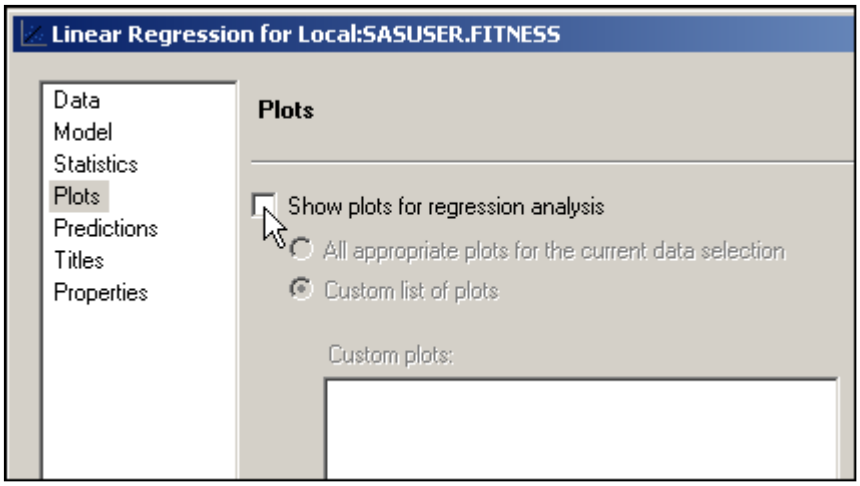
8. Select the results tab and then Modify Task.

9. This will throw up the dialogue box asking to protect the data. Click Yes to continue to the modify task dialogue box

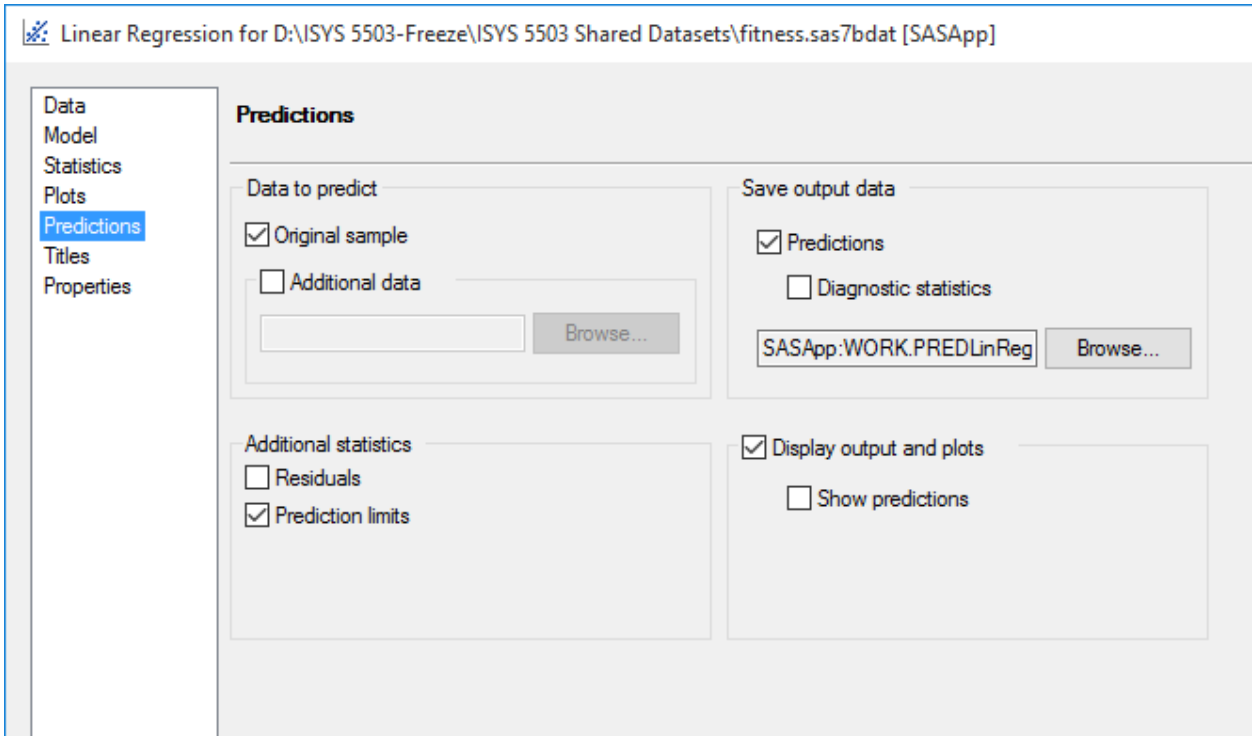


10. Go to Plots

11. Uncheck the box for **Show plots for regression analysis**.



12. Select **Predictions** at the left.
13. Check the box for **Original Sample** *and* ***Prediction limits*** (to generate prediction Confidence Intervals.
14. Run the task now.



15. Notice the Results tabs has additional information as seen below. SASEG now reads 36 observations but only uses the 31 observations to develop the model. It is able to identify the five values that have missing fields in the oxygen consumption column.

16. In order to retrieve the predicted values, Click on the output data tab and scroll to the bottom. You will notice that the output data gives the predicted values along with the control limits similar to the output of the previous method as seen below.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

Number of Observations Read	36
Number of Observations Used	31
Number of Observations with Missing Values	5

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			

Root MSE	2.74515	R-Square	0.7434
Dependent Mean	47.37581	Adj R-Sq	0.7345
Coeff Var	5.79442		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	82.42494	3.85582	21.38	<.0001
RunTime	1	-3.31085	0.36124	-9.17	<.0001

Generated by the SAS System ('SASApp', X64_ES08R2) on May 03, 2017 at 2:21:11 PM

Input Data | Code | Log | Output Data | Results - SAS Report | Results - HTML | Results - PDF | Results - RTF | Results - Listing

Modify Task | Filter and Sort | Query Builder | Where | Data | Describe | Graph | Analyze | Export | Send To

	RunTime	Age	Weight	Oxygen_Consumption	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance	predicted_Oxygen_Cons...	lclm_Oxygen_Consumpti...	uclm_Oxygen_Consumpti...	I C
13	10.13	44	73.03	50.54	168	45	168	67	48.885983433	47.822773589	49.949193277	4
14	10.25	48	91.63	46.77	162	48	164	61	48.488680861	47.450162947	49.527198775	4
15	10.33	54	83.12	51.85	166	50	170	49	48.223812479	47.197822167	49.249802792	4
16	10.47	52	73.71	45.79	186	59	188	47	47.760292812	46.748261262	48.772324362	4
17	10.5	52	82.78	47.47	170	53	172	51	47.660967169	46.650573333	48.671361005	4
18	10.6	47	79.15	47.27	162	47	164	56	47.329881692	46.321441608	48.338321776	4
19	10.85	43	81.19	49.09	162	64	170	65	46.502168	45.475107311	47.529228689	4
20	10.95	51	69.63	40.84	168	57	172	48	46.171082523	45.127473514	47.214691532	4
21	11.08	51	67.25	45.12	172	48	172	43	45.740671403	44.668296935	46.813045871	4
22	11.12	45	66.45	44.75	176	51	176	55	45.608237212	44.525450416	46.691024009	3
23	11.17	54	79.38	46.08	156	62	165	40	45.442694474	44.345911022	46.539477926	3
24	11.37	44	89.47	44.61	178	62	182	58	44.78052352	43.617659014	45.943388026	3
25	11.5	48	61.24	47.92	170	52	176	45	44.3501124	43.136551803	45.563672998	3
26	11.63	47	77.45	44.81	176	58	176	50	43.91970128	42.65019155	45.189211011	3
27	11.95	40	75.98	45.68	176	70	180	56	42.860227754	41.434664707	44.285790802	3
28	12.63	57	73.37	39.41	174	58	176	20	40.608846512	38.793044003	42.424649021	3
29	12.88	54	91.63	39.2	168	44	172	23	39.78113282	37.809054764	41.753210876	3
30	13.08	44	81.42	39.44	174	63	176	41	39.118961866	37.018537621	41.219386111	3
31	14.03	45	87.66	37.39	186	56	192	30	35.973649836	33.236696157	38.710603515	2
32	9	.	.	.	.	.	.	.	52.627249322	51.081244412	54.173254231	4
33	10	.	.	.	.	.	.	.	49.316394553	48.21895386	50.413835246	4
34	11	.	.	.	.	.	.	.	46.005539785	44.951809844	47.059269725	4
35	12	.	.	.	.	.	.	.	42.694685016	41.242774786	44.146595246	3
36	13	.	.	.	.	.	.	.	39.383830248	37.335058502	41.432601993	3

The output data set contains columns for all variables in the analysis data set, but the values of each record are set to missing for those variables that either were missing or did not exist in the scoring data set. *Also, note the 95% confidence limit for the prediction.*

Choose only values within or near the range of the predictor variable when you are predicting new values for the response variable. For this example, the values of the variable **`RunTime`** range from 8.17 to 14.03 minutes. Therefore, it is unwise to predict the value of **`Oxygen_Consumption`** for a **`RunTime`** of 18. The reason is that the relationship between the predictor variable and the response variable might be different beyond the range of your data.

An Additional Exercise

1. Fitting a Simple Linear Regression Model

Use the **BodyFat2** data set (the one created in the previous exercise) for this exercise.

a. Perform a simple linear regression model with **PctBodyFat2** as the response variable and **Abdomen** as the predictor. Produce regression plots with confidence and prediction intervals.

1) What is the value of the *F* statistic and the associated *p*-value? How would you interpret this with regards to the null hypothesis?

2) Write out the predicted regression equation.

3) What is the value of the R^2 statistic? How would you interpret this?

b. Produce predicted values for **PctBodyFat2** when **Abdomen** is 80, 100 and 120. The **AbdomenPred** data set in SASUSER contains the 3 observations needed.

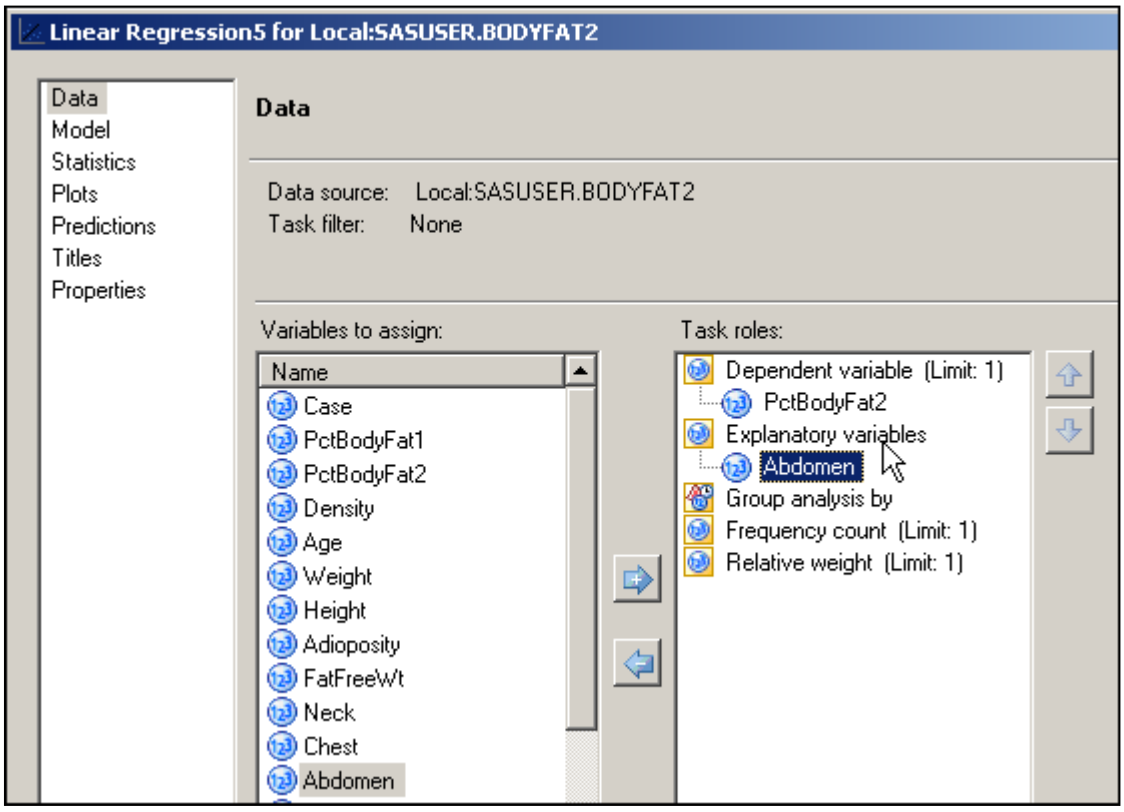
1) What are the predicted values?

2) Is it appropriate to predict **PctBodyFat2** when **Abdomen** is 140?

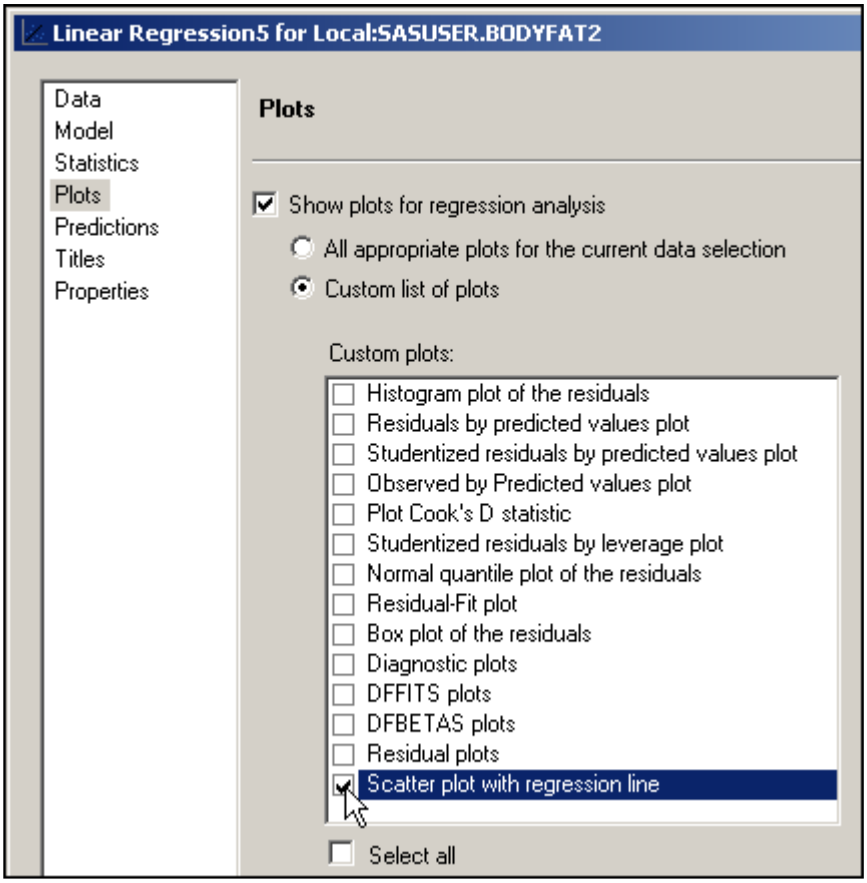
Fitting a Simple Linear Regression Model

a. Perform a simple linear regression model with **`PctBodyFat2`** as the response variable and **`Abdomen`** as the predictor. Produce regression plots with confidence and prediction intervals.

- With the **BodyFat2** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.
- Drag **`PctBodyFat2`** to the dependent variable task role and **`Abdomen`** to the explanatory variables task role.



- With **Plots** selected at the right, select **Custom list of plots** under `Show plots for regression analysis`. From the menu that appears, uncheck the box for **Diagnostic plots** and check the box for **Scatter plot with regression line**.



- 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

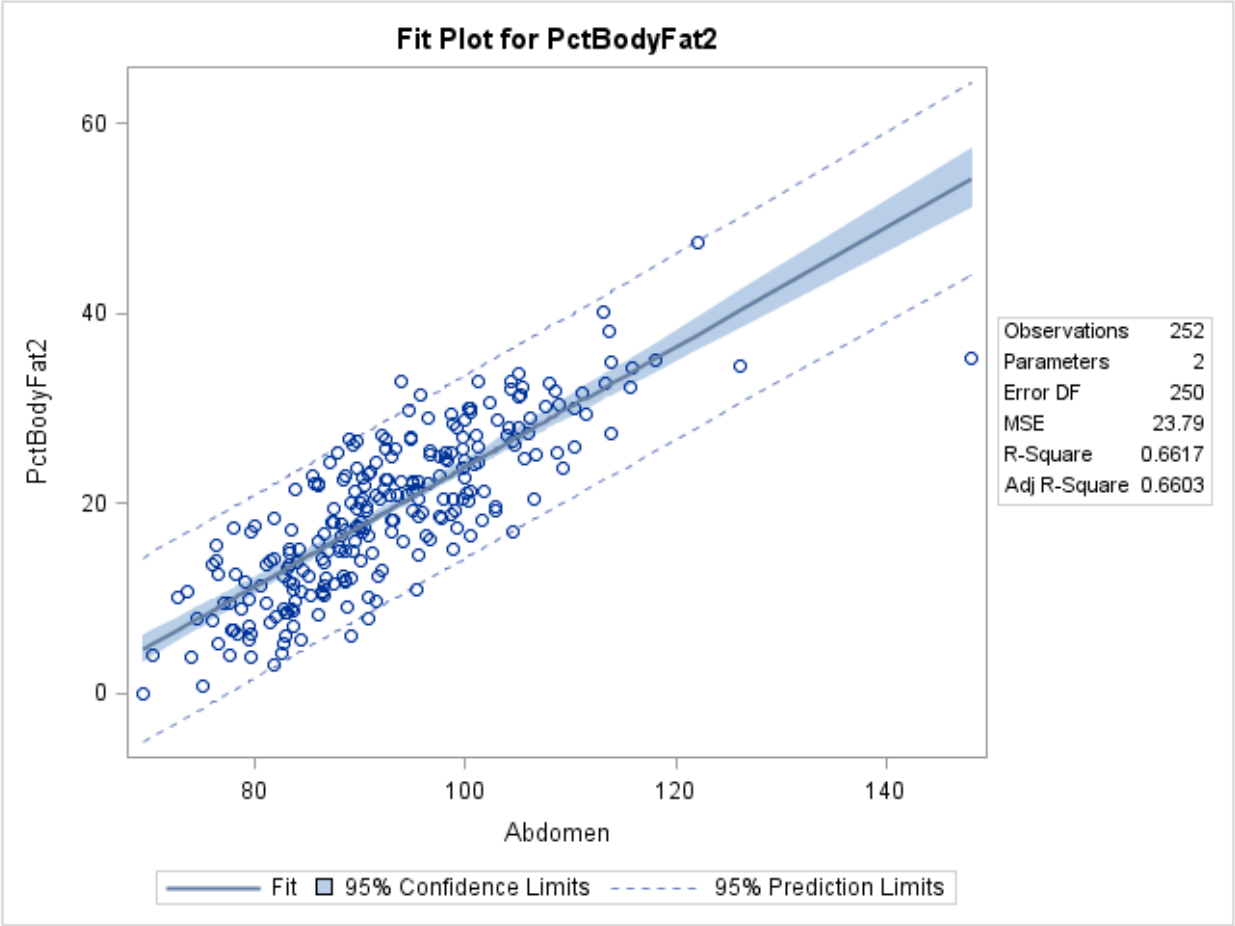
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11632	11632	488.93	<.0001
Error	250	5947.46303	23.78985		
Corrected Total	251	17579			

Root MSE	4.87748	R-Square	0.6617
Dependent Mean	19.15079	Adj R-Sq	0.6603
Coeff Var	25.46884		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-39.28018	2.66034	-14.77	<.0001
Abdomen	1	0.63130	0.02855	22.11	<.0001



1) What is the value of the *F* statistic and the associated *p*-value? How would you interpret this with regards to the null hypothesis?

   The *F* value is 488.93 and the *p*-value is <.0001. You would reject the null hypothesis of no relationship.

2) Write out the predicted regression equation.

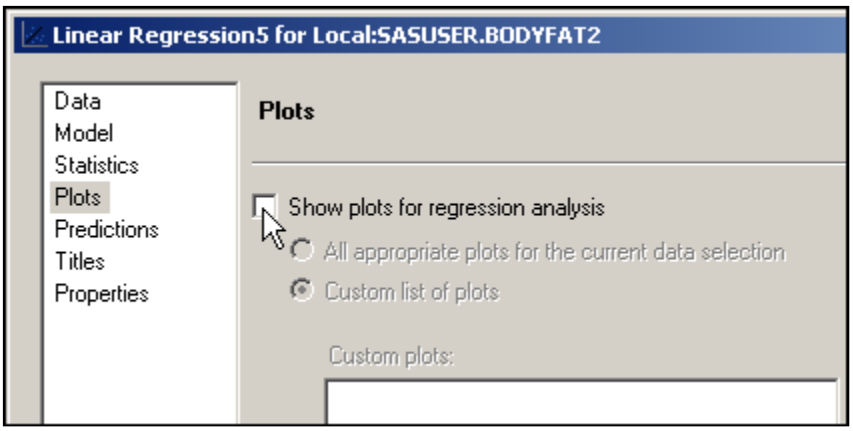
   From the parameter estimates table, the predicted value of **`PctBodyFat2`** = -39.28018 + 0.63130 * **`Abdomen`**.

3) What is the value of the R^2 statistic value? How would you interpret this?

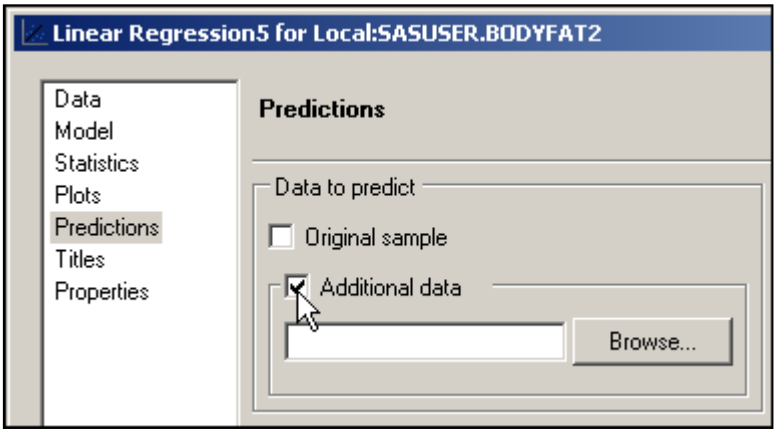
   The R^2 value of 0.6617 can be interpreted to mean that 66.17% of the variability in **`PctBodyFat2`** can be explained by **`Abdomen`**.

b. Produce predicted values for **`PctBodyFat2`** when **`Abdomen`** is 80, 100 and 120. The **`AbdomenPred`** data set in SASUSER contains the 3 observations needed.

- Modify the previous Linear Regression task in the project.
- Uncheck the box for **Show plots for regression analysis**.



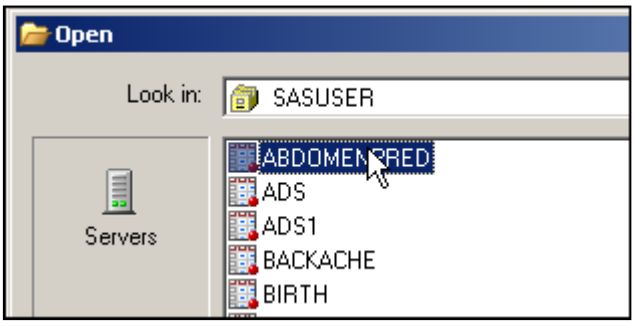
- With **Predictions** selected at the left, check the box for **Additional data** and **Prediction limits**.



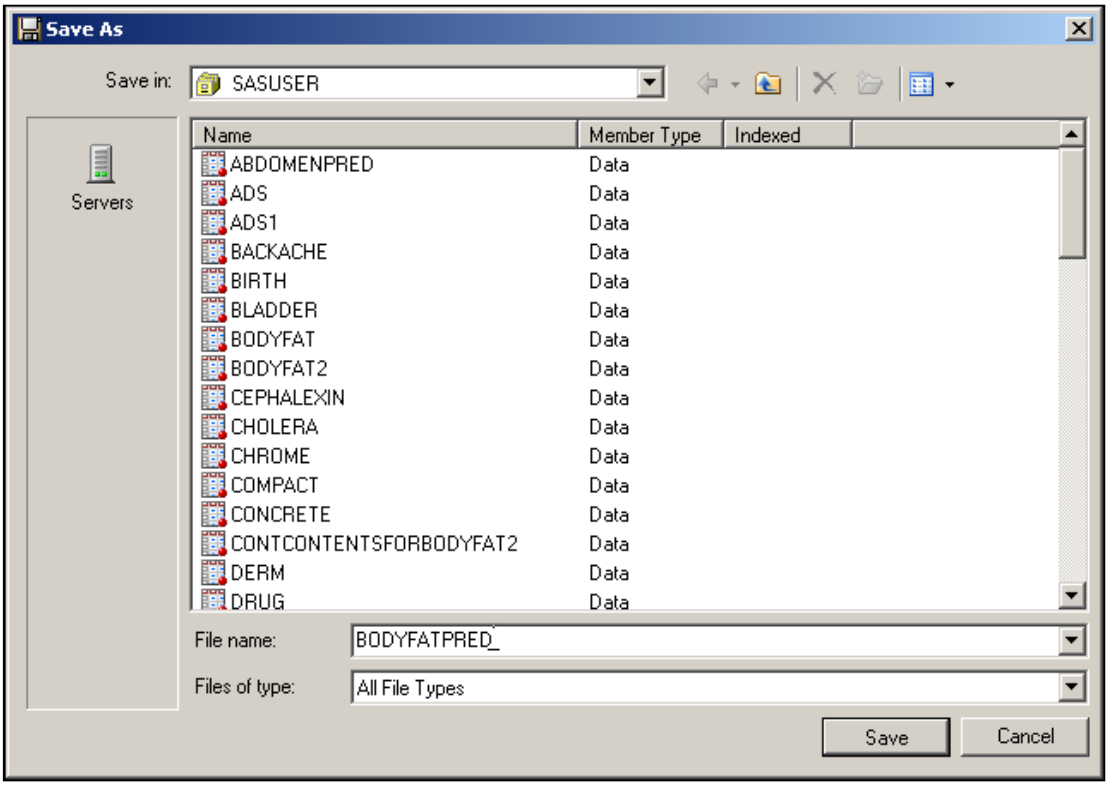
- Click 

.

- When the next window opens, select the **AbdomenPred** data set from the SASUSER library. Either double-click the name of the data set or highlight it and click Open.

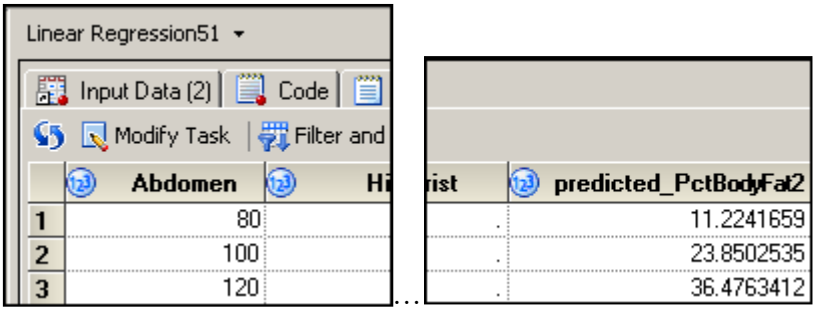


- Under `Save output data`, with **Predictions** checked, click Browse...
- With that window open, overtype the default **File name:** with **BODYFATPRED**.



- Click Save to close that window.
- Click Run.
- In the workspace, you will now see a tab for the newly created data set. Click the tab to open the data table.

- Scroll to the right to see the predicted values for **PctBodyFat2**.



1) What are the predicted values?

   The predicted values at **Abdomen**=80, 100 and 120, are 11.22, 23.85 and 36.48, respectively.

2) Is it appropriate to predict **PctBodyFat2** when **Abdomen** is 160?

   No, because there are no data in the model data set with **Abdomen** greater than 148.1. You should not predict beyond the range of your data.