SASEG 9A - Multiple Regression

(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

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.

Concepts of Multiple Regression

Multiple Linear Regression with Two Variables

Consider the two-variable model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where

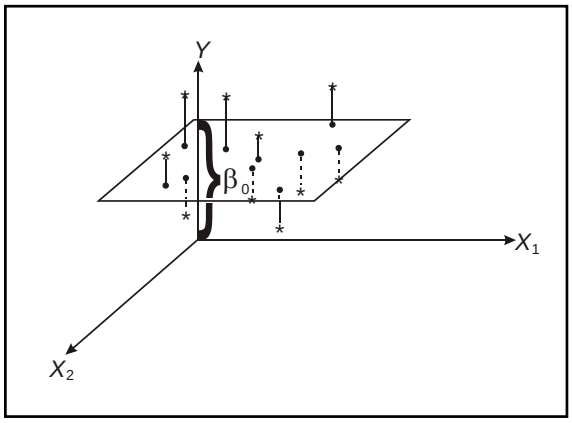
Y	is the dependent variable.
X_1 and X_2	are the independent or predictor variables.
ε	is the error term.
β_0, β_1, and β_2	are unknown parameters.

49

In simple linear regression, you can model the relationship between the two variables (two dimensions) with a line (one dimension).

For the two-variable model, you can model the relationship of three variables (three dimensions) with a plane (two dimensions).

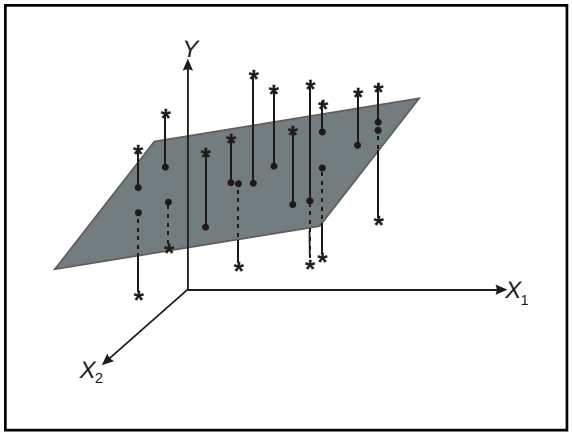
Picturing the Model: No Relationship



50

If there is no relationship among Y and X_1 and X_2, the model is a horizontal plane passing through the point ($Y = \beta_0$, $X_1 = 0$, $X_2 = 0$).

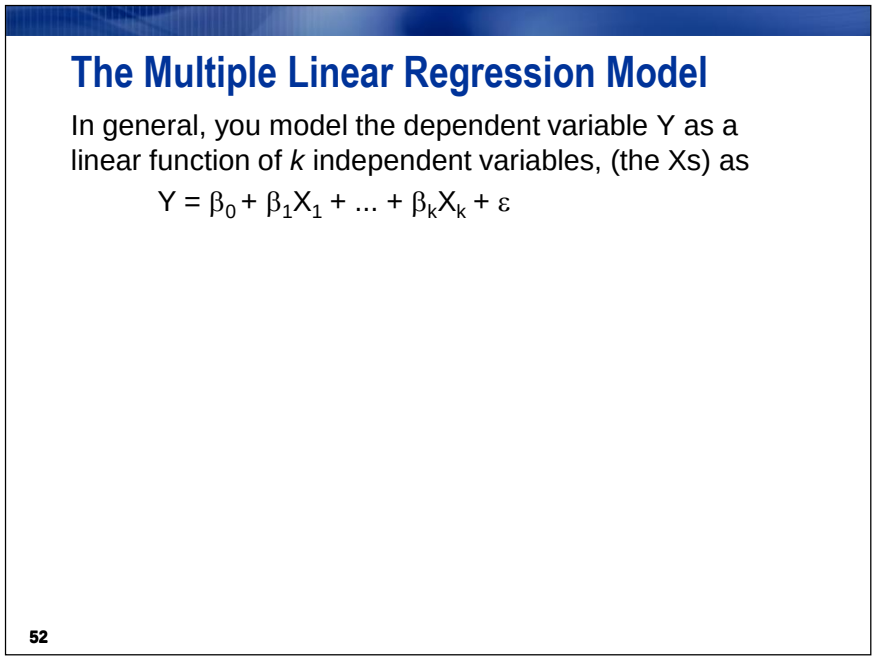
Picturing the Model: A Relationship



51

If there is a relationship among Y and X_1 and X_2, the model is a sloping plane passing through three points:

- ($Y = \beta_0$, $X_1 = 0$, $X_2 = 0$)
- ($Y = \beta_0 + \beta_1$, $X_1 = 1$, $X_2 = 0$)
- ($Y = \beta_0 + \beta_2$, $X_1 = 0$, $X_2 = 1$)

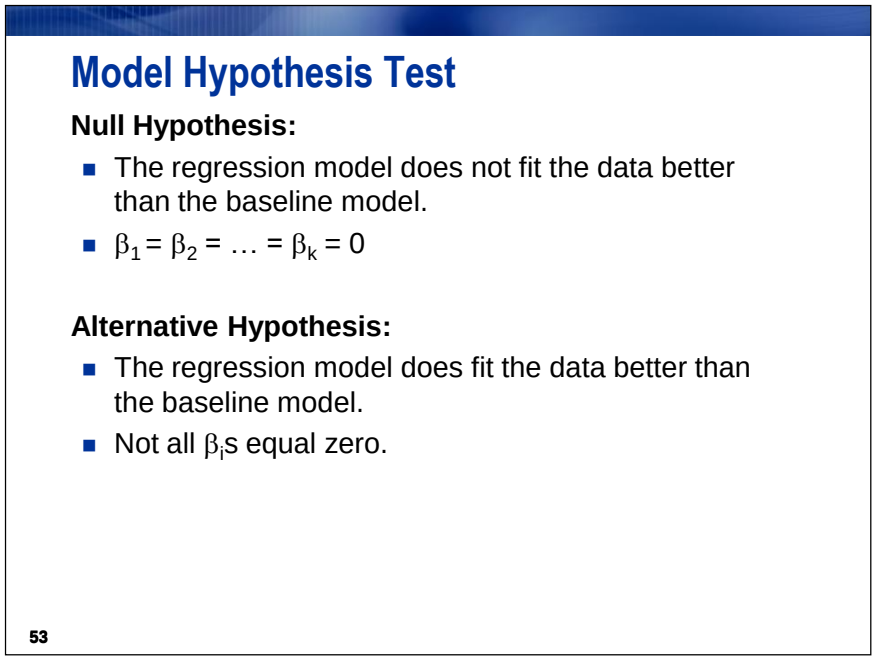


You investigate the relationship among $k + 1$ variables (k predictors + 1 response) using a k-dimensional surface for prediction.

The multiple general linear model is not restricted to modeling only planar relationships. By using higher order terms, such as quadratic or cubic powers of the Xs or cross products of one X with another, surfaces more complex than planes can be modeled.

In the examples, the models are limited to relatively simple surfaces.

The model has $p = k + 1$ parameters (the βs), including the intercept, β_0.



If the estimated 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 all of the slopes of the regression in the population are not 0 and that the predictor variables explain a significant amount of variability in the response variable.

If the estimated 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 at least one slope of the regression in the population is not 0 and that at least one predictor variable explains a significant amount of variability in the response variable.

Assumptions for Linear Regression

- The mean of the Ys is accurately modeled by a linear function of the Xs.
- The random error term, ε, is assumed to have a normal distribution with a mean of zero.
- The random error term, ε, is assumed to have a constant variance, σ^2.
- The errors are independent.

57

Techniques to evaluate the validity of these assumptions are discussed in a later chapter.

Multiple Linear Regression versus Simple Linear Regression

Main Advantage

Multiple linear regression enables you to investigate the relationship among Y and several independent variables simultaneously.

Main Disadvantages

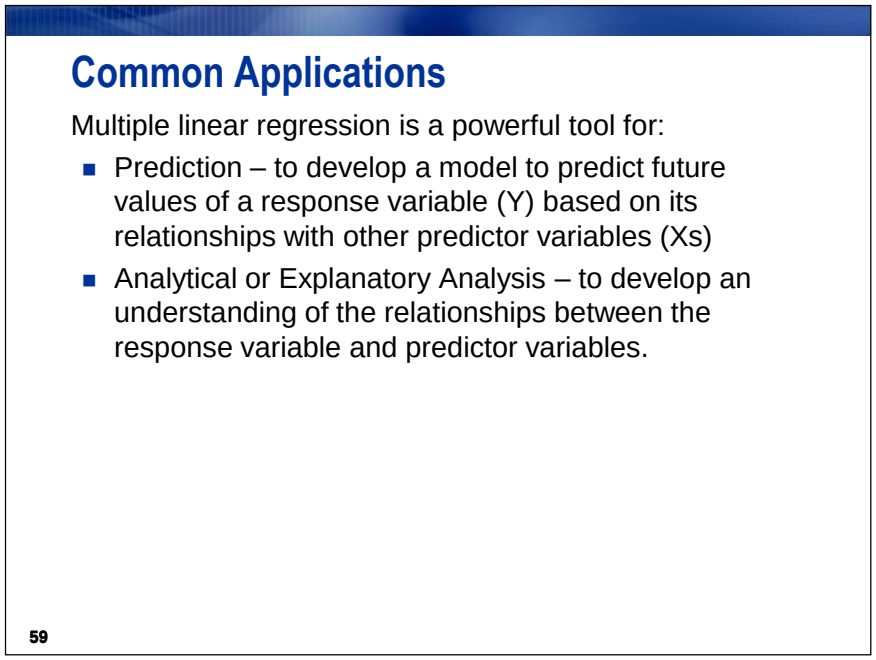
Increased complexity makes it more difficult to

- ascertain which model is “best”
- interpret the models.

58

The advantage of performing multiple linear regression over a series of simple linear regression models far outweighs the disadvantages. In practice, many responses depend on multiple factors that might interact in some way.

SAS tools help you decide upon a “best” model, a choice that might depend upon the purposes of the analysis, as well as subject-matter expertise.



Even though multiple linear regression enables you to analyze many different experimental designs, ranging from simple to complex, you will focus on applications for analytical studies and predictive modeling. Other SAS Enterprise tasks, such as the Linear Models or Mixed Models tasks, are better suited for analyzing experimental data.

The distinction between using multiple regression for an analytic analysis and prediction modeling is somewhat artificial. A model developed for prediction will probably be a good analytic model. Conversely, a model developed for an analytic study will probably be a good prediction model.

Myers (1999) actually refers to four applications of regression: prediction, variable screening, model specifications, and parameter estimation. The term *analytical analysis* is similar to Myers' parameter estimation application and variable screening.

Prediction vs. Explanation

Prediction

- The terms in the model, the values of their coefficients, and their statistical significance are of secondary importance.
- The focus is on producing a model that is the best at predicting future values of Y as a function of the Xs. The predicted value of Y is given by

$$\underline{\hat{Y}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$$

60

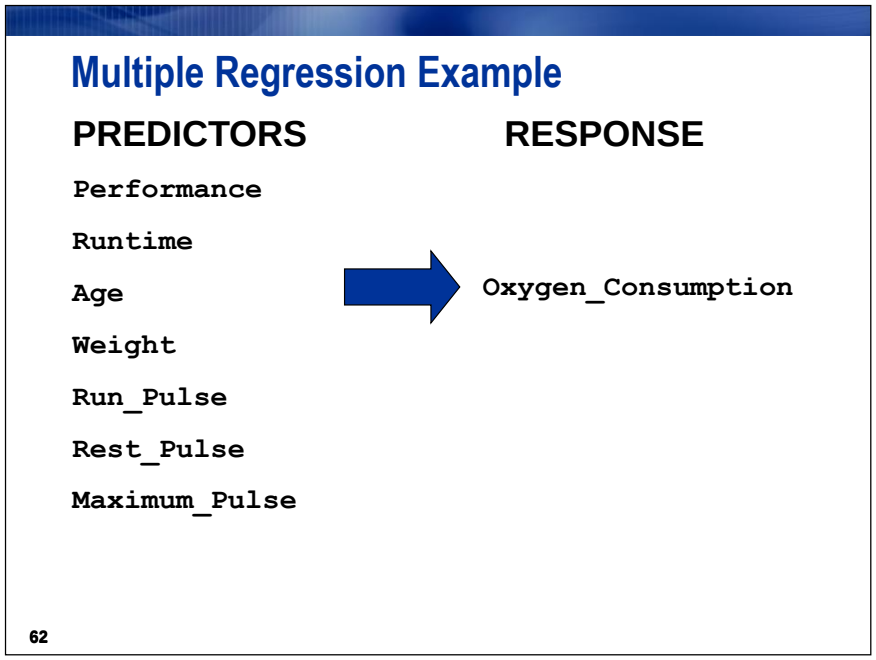
Most investigators do not ignore the terms in the model (the Xs), the values of their coefficients (the βs), or their statistical significance (the *p*-values). They use these statistics to help choose among models with different numbers of terms and predictive capabilities.

Analytical or Explanatory Analysis

- The focus is on understanding the relationship between the dependent variable and the independent variables.
- Consequently, the statistical significance of the coefficients is important as well as the magnitudes and signs of the coefficients.

$$\hat{Y} = \underline{\hat{\beta}_0} + \underline{\hat{\beta}_1} X_1 + \ldots + \underline{\hat{\beta}_k} X_k$$

61



An analyst knows from doing a simple linear regression that the measure of performance is an important variable in explaining the oxygen consumption capability of a club member.

The analyst is interested in investigating other information to ascertain whether other variables are important in explaining the oxygen consumption capability.

Recall that you did a simple linear regression on **Oxygen_Consumption** with **RunTime** as the predictor variable.

The R^2 for this model was 0.7434, which suggests that 25.64% of the variation in **Oxygen_Consumption** is still unexplained.

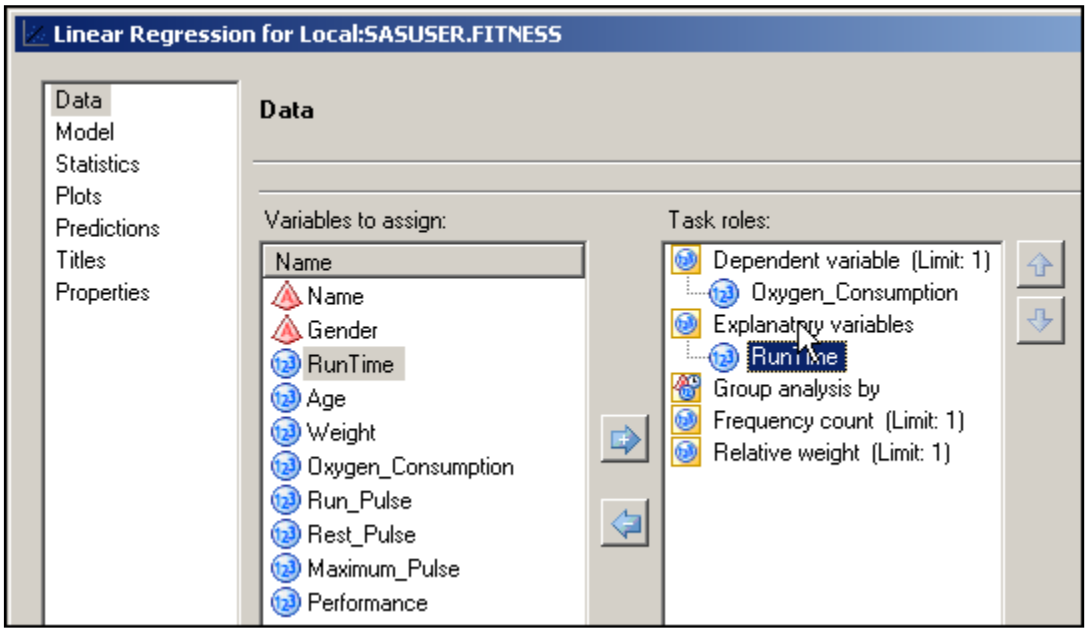
Consequently, adding other variables to the model, such as **Performance** or **Age**, might provide a significantly better model.

Fitting a Multiple Linear Regression Model (Two or More Independent Variables)

First, let's recall the Simple Linear Regression Model with

```
Oxygen_Consumption = f(RunTime)
```

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. Click Run.

The results were –

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance

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

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`**.

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

Note 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.

The Parameter Estimates table was as follows -

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

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`**).

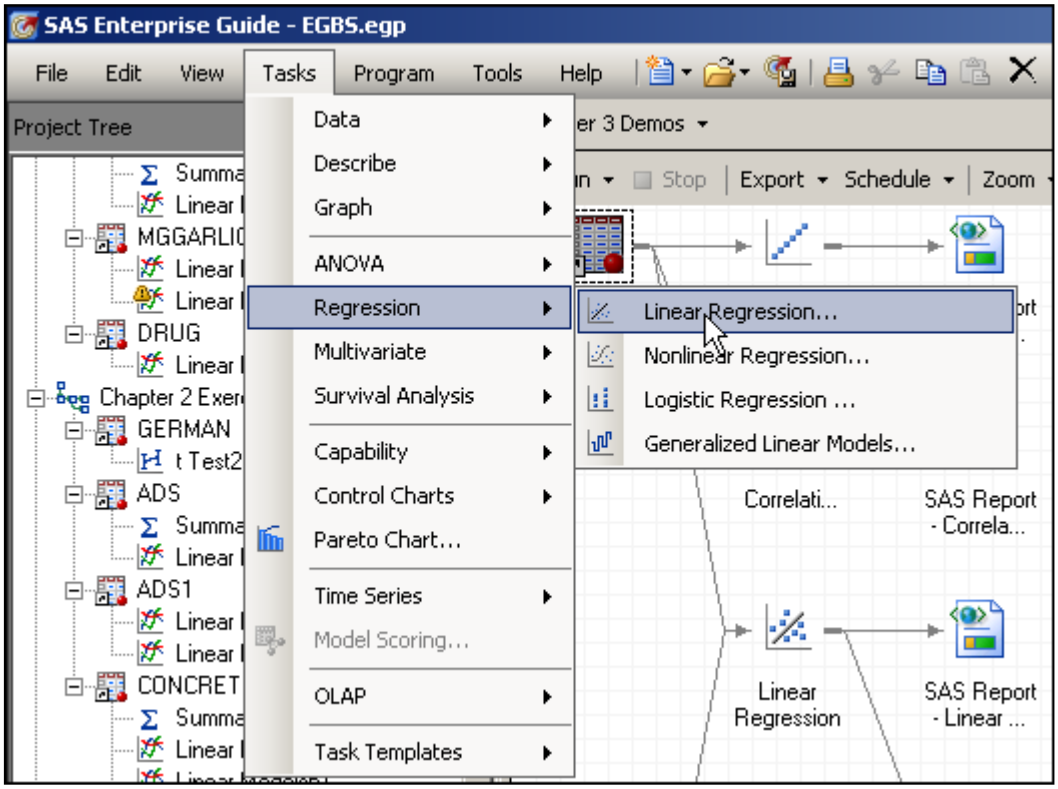
*The model indicates that an increase of one unit for **`Runtime`** amounts to a 3.31085 decrease in **`Oxygen_Consumption`**. However, this equation is appropriate only in the range of values you observed for the variable **`RunTime`**.*

Multiple Linear Regression – (Two Independent Variables)

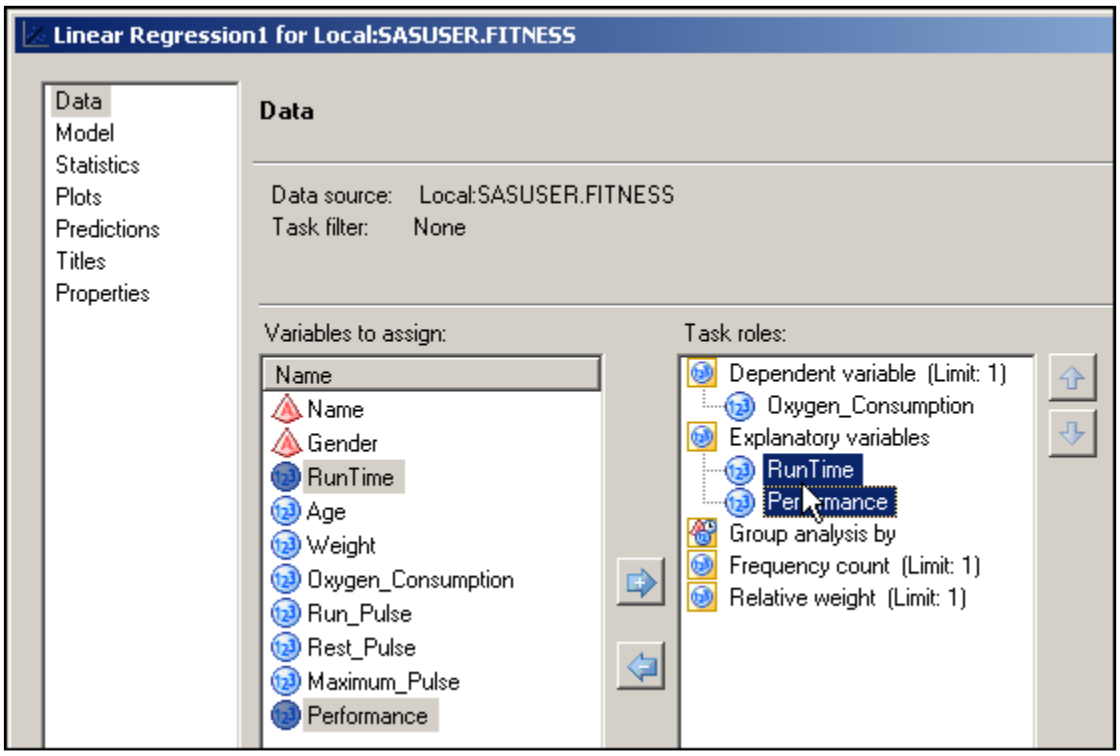
`Oxygen_Consumption = f(RunTime, Performance)`

Invoke the Linear Regression task and perform multiple linear regression analysis of **`Oxygen_Consumption`** with **`Performance`** and **`Runtime`** as explanatory variables. Interpret the output for the two-variable model.

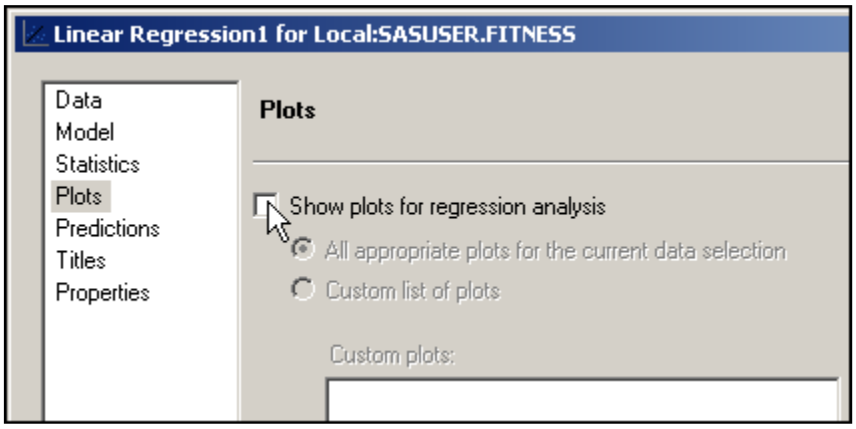
1. With the **Fitness** data set selected, click **Tasks** ⇨ **Regression** ⇨ **Linear Regression…**.



2. Drag **`Oxygen_Consumption`** to the dependent variable task role and **`RunTime`** and **`Performance`** to the explanatory variables task role.



3. Uncheck the box for **<u>Show plots for regression analysis</u>** under Plots.



4. Click Run.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	646.33101	323.16550	44.09	<.0001
Error	28	205.22355	7.32941		
Corrected Total	30	851.55455			

This model is statistically significant at the alpha level of 0.05 (p < .0001).

Root MSE	2.70729	R-Square	0.7590
Dependent Mean	47.37581	Adj R-Sq	0.7418
Coeff Var	5.71450		

Comparing this two variable model (`RunTime, Performance`) to the one variable model (`RunTime`) - the R^2 for this model, 0.7590, is only slightly larger than the R^2 for the model in which `RunTime` is the only predictor variable, 0.7434.

The R^2 always increases as you include more terms in the model. However, choosing the "best" model is not as simple as just making the R^2 as large as possible.

The adjusted R^2 is a measure similar to R^2, but it takes into account the number of terms in the model.

The adjusted R^2 for this model is 0.7418, slightly higher than the adjusted R^2 of 0.7345 for the **`RunTime`** only model. This suggests, albeit mildly, that **`Performance`** does improve the model predicting **`Oxygen_Consumption`**.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71.52626	8.93520	8.00	<.0001
RunTime	1	-2.62163	0.62320	-4.21	0.0002
Performance	1	0.06360	0.04718	1.35	0.1885

Using the estimates for β_0, β_1, and β_2 above, this model can be written as:

`Oxygen_Consumption` = 71.52626 – 2.62163*__`Runtime`__ + 0.06360*__`Performance`__

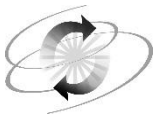
The p-value for ***`Performance`*** *is large (Pr > |t| = 0.1885), which suggests that the slope is not significantly different from 0.*

Note - the correlation between **`Performance`** and **`Oxygen_Consumption`** was large and statistically significant (r=.77890, p<.0001).

How did this seeming contradiction occur? The test for β_i=0 is conditioned on the other terms in the model. That is the reason that neither **`RunTime`** nor **`Performance`** have the same *p*-values (or parameter estimates) when used alone as when used in a model that includes both. The test for β_1=0 (for **`Runtime`**) is conditional on (or adjusted for) X_2 (**`Performance`**). Similarly, the test for β_2=0 is conditioned on X_1 (**`RunTime`**).

The significance level of the test does **not** depend on the order in which you list the independent variables in the Task roles, but it does depend upon which set of variables are included in the model.

In a later section, you will look more at the difficulties involved with analyzing and selecting the best models due to the relationships among predictor variables.



Exercises

1. Some practice (going above and beyond) -- Performing a Multiple Regression to discover the model

a. (One variable model) Using the **BodyFat2** data set, run a regression of **PctBodyFat2** on the variable **Abdomen**

PctBodyFat2 = f(Abdomen)

1) Note the results – R^2 (around 66%) as well as the coefficients.

b. (Multiple variable model) 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**.

PctBodyFat2 = f(Age, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen, Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist)**

1) Compare the ANOVA table of this model with the the ANOVA table of one variable model with only **Abdomen** in **a.** above. What is different?

2) How do the R^2 and the adjusted R^2 of this model compare with the statistics for single variable model (**Abdomen**) in **a.**?

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

c. Simplifying the Model

1) Rerun the model in **b.**, but eliminate all the variables with a *p*-value greater than .05 (those that are not significant).

Hint: You should be removing from the model - **Weight**, **Height**, **Chest**, **Hip**, **Thigh**, **Knee**, **Ankle**, and **Biceps**.

2) Compare the output from this new model with the output from the Exercise **b.** model (above).

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

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

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

Note – we should notice that the one variable model **(Abdomen)** has an R^2 of about 66%; the full model with all variables has an R^2 of about 74.8% (improved explanatory ability). However, the reduced model (**c.** above) has an R^2 of about 73%. A model with six (6) variables **(Age**, **Neck**, **Chest**, **Abdomen, Forearm**, and **Wrist)** is able to explain just as well as the full model – 73% compared to 74.8% (but with a smaller number of variables – *six variables compared to thirteen variables*).