A standard multiple regression allows you to predict a dependent variable based on multiple independent variables and is an extension to simple linear regression. The dependent variable is also referred to as the outcome, target or criterion variable and the independent variables as predictor, explanatory or regressor variables. This method also allows you to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained.

## Y = b0 + b1X1 + b2X2 + e

## Where β0 is the intercept (also known as the constant), β1 is the slope parameter (also known as the slope coefficient) for X1, and so forth, and ε represents the errors. This type of statistical test relies on the initial assumption that there is, in fact, a linear relationship between each independent variable and the dependent variable and a linear relationship between the "composite" of the independent variables and the dependent variable. This assumption can be examined, as you will do in this guide. Confidence intervals can be calculated for the sample intercept and slope parameters to estimate the likely range of values that these parameters might take in the population. Furthermore, predictions can be made based on the regression equation calculated. You will calculate all these statistical measures in this guide.

## What is required

In order to run a multiple regression, you require the following:

1. Two or more independent variables that can be either **continuous** or **categorical** (e.g., height, exam performance, gender, etc.).
2. One dependent variable that is **continuous** (e.g., height, weight, etc.).

## Problems solved using multiple regression

Multiple regression can be used to answer the following problems:

### 1. Predict new values for the dependent variable given the independent variables

You can use multiple regression to predict the value of one variable when you know the value of other variables. For example, you might have individuals' heights, weights, age and gender, and you want to predict running performance. Using this data you construct a multiple regression equation, which you then use to predict new individuals' running performance based on their measured physical properties (i.e., their height, weight, age and gender).

### 2. Determine how much of the variation in the dependent variable is explained by the independent variables

Often, your goal is not to make predictions, but to determine how much of the variation in the dependent variable can be explained by all the independent variables. In addition, you can use multiple regression to understand the relative, unique contribution of each independent variable towards this total. For example, you might have individuals' heights, weights, age and gender, and you want to predict running performance. You want to know how much of the variation in running performance can be explained by the predictor variables. Additionally, you want to know the relative contribution of each predictor to the explanation of variance.

## Assumptions & order of testing

For a multiple regression to be a valid test to use (e.g., provide valid predictions), the following assumptions must hold:

1. One dependent variable measured at the interval or ratio level.
2. Two or more independent variables that are measured at either the continuous (ratio or interval) or nominal level.
3. Independence of errors (residuals).
4. A linear relationship between the predictor variables (and composite) and the dependent variable.
5. Homoscedasticity of residuals (equal error variances).
6. No multicollinearity.
7. No significant outliers or influential points.
8. Errors (residuals) are normally distributed.

These assumptions will allow you to (1) provide information on the accuracy of your predictions, (2) test how well the regression model fits your data, (3) determine the variation in your dependent variable explained by your independent variables, and (4) test hypotheses on your regression equation. If these assumptions are violated, you need to make corrections and re-test these assumptions. If they still do not pass, you must find alternative statistical tests.

With real-world data, it is not uncommon for one or more of these six assumptions to be violated. This guide will, therefore, provide explanations of how to implement techniques to overcome these violations and move forward with your analysis, if this is indeed possible with your data. The assumptions will be tackled in the order they have been addressed above. This order has been chosen because it represents an order whereby if a violation is not correctable, you cannot proceed with the analysis (if you want valid results). For example, if you have violated assumption (1), it is pointless testing assumptions (2) through (6) because the regression analysis will already have been rendered invalid by failure of assumption (1).

## EXAMPLE

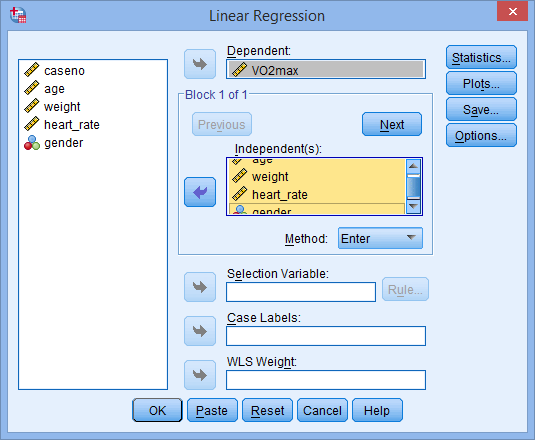
A health researcher wants to be able to predict maximal aerobic capacity (VO2max), an indicator of fitness and health. Normally, to perform this procedure requires expensive laboratory equipment and necessitates that an individual exercise to their maximum (i.e., until they can longer continue exercising due to physical exhaustion). This can put off those individuals that are not very active/fit and those individuals that might be at higher risk of ill health (e.g., older unfit subjects). For these reasons, it has been desirable to find a way of predicting an individual's VO2max based on more easily and cheaply measured attributes. To this end, the researcher recruits 100 participants to perform a maximum VO2max test, but also records their age, weight, heart rate and gender. Heart rate is the average of the last 5 mins of a 20 mins much easier, lower workload cycling test. The researcher's goal is to be able to predict VO2max based on age, weight, heart rate and gender.

## MULTIPLE REGRESSION PROCEDURE

In order to check the assumptions of this test, you will first need to run the multiple regression procedure. This is mostly due to the fact that many of the assumptions are checked by inspection of the residuals, which can only be calculated once a regression line has been fitted/generated. To run a multiple regression you need to use the **Linear Regression** dialogue box. The instructions that follow will show you how to build the regression model in SPSS and which options to select in order to test the assumptions of the regression model.

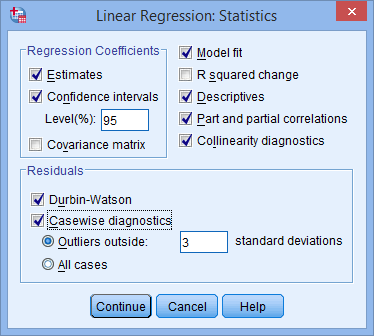
1. Click **Analyze > Regression > Linear...** on the main menu. You will be presented with the **Linear Regression** dialogue box

2. Transfer the dependent variable, VO2max, into the Dependent: box, and the independent variables, age, weight, heart\_rate and gender into the Independent(s): box, using the https://statistics.laerd.com/premium/mr/img/right-arrow-button.pngbuttons, as shown below (all other boxes can be ignored):



3. Click the https://statistics.laerd.com/premium/mr/img/statistics-button.pngbutton. You will be presented with the **Linear Regression: Statistics** dialogue box.

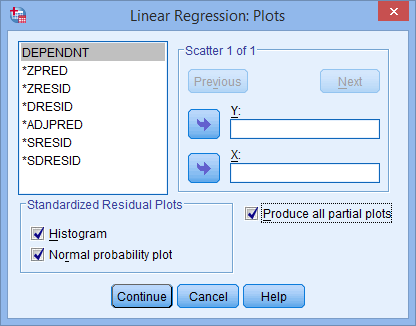
4. In addition to the options that are already selected, select Confidence intervals from the –Regression Coefficients– area and leave the Level(%): at 95. Also, select Casewise diagnostics from the –Residuals– area and leave the option value at 3 standard deviations, and select Durbin-Watson from the –Residuals– area. Then select Model Fit, Descriptives, Part and partial correlations and Collinearity diagnostics. You will end up with the following screen:

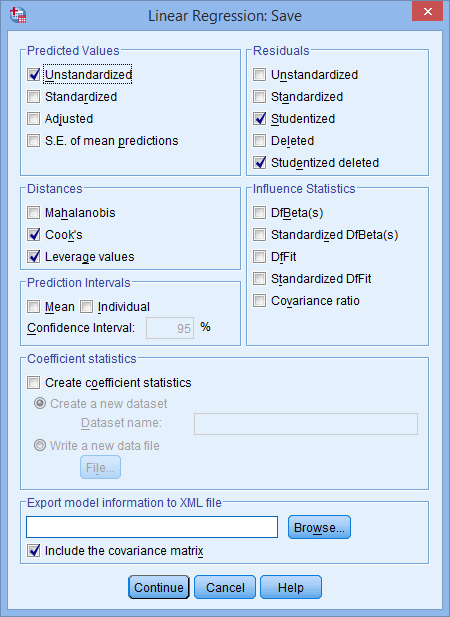


5. Click the https://statistics.laerd.com/premium/mr/img/continue-button.pngbutton. You will be returned to the **Linear Regression** dialogue box.

6. Click the https://statistics.laerd.com/premium/mr/img/plots-button.pngbutton and you will be presented with the **Linear Regression: Plots** dialogue box.

7. Select Histogram and Normal probability plot from the –Standardized Residual Plots– area, and also select Produce all partial plots, as shown below:



8. Click the https://statistics.laerd.com/premium/mr/img/continue-button.pngbutton. You will be returned to the **Linear Regression** dialogue box.

9. Click the https://statistics.laerd.com/premium/mr/img/save-button.pngbutton. You will be presented with the **Linear Regression: Save** dialogue box.

10. Check Unstandardized in the –Predicted Values– area, Cook's and Leverage values in the –Distances– area, and Studentized and Studentized deleted in the –Residuals– area, as shown to the right.

11. Click the https://statistics.laerd.com/premium/mr/img/continue-button.pngbutton. You will be returned to the **Linear Regression** dialogue box.

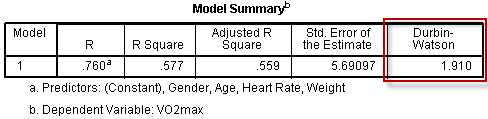
12. Click the https://statistics.laerd.com/premium/mr/img/ok-button.pngbutton. This will generate the output.

## TESTING OF ASSUMPTIONS - PART I

The majority of your time can often be spent analyzing the assumptions of multiple regression for any violations and making any necessary corrections to the data. Due to the number of assumptions, testing for assumptions will be split into three major parts: part 1 will deal with independence of cases, linearity, homoscedasticity and multicollinearity; part 2 will deal with the various ways to detect unusual points; and part 3 will deal with normality of the residuals.

## Independence of observations

If you examine the SPSS output generated by this test you will find a table called **Model Summary**, which contains the Durbin-Watson statistic, as highlighted below:



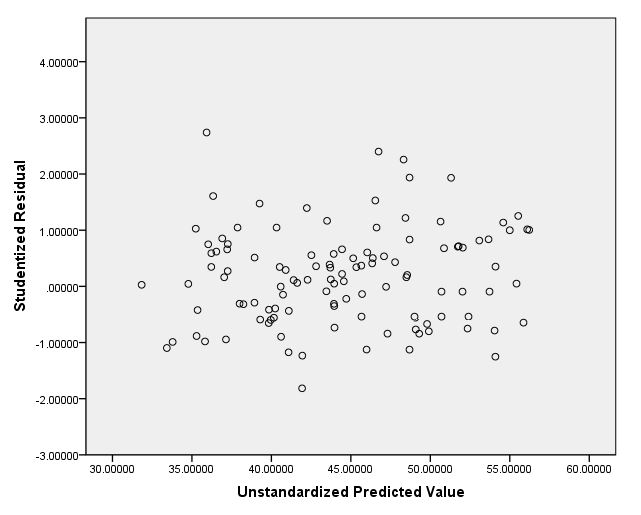
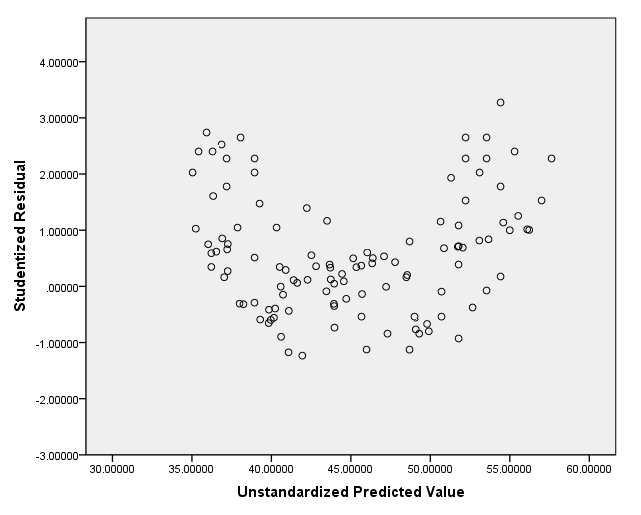
The Durbin-Watson statistic for this analysis is 1.910. The Durbin-Watson statistic can range from 0 to 4, but you are looking for a value of approximately 2 to indicate that there is no correlation between residuals. You can see that our value is very close to 2, so it can be accepted that there is independence of errors (residuals).

A large part of this assumption is also based on study design, which is not tested for statistically. Indeed, in situations where it is highly unlikely that observations will be related, this assumption might not be tested for statistically. If you do have correlated errors, multiple regression is not a suitable method of analysis and you will need to consider another type of analysis, such as time-series methods.

## Checking for a linear relationship

An assumption of multiple linear regression is that the independent variables collectively are linearly related to the dependent variable and also that each independent variable is linearly related to the dependent variable. This is tested for below:

You can check for this assumption by plotting the studentized residuals (SRE\_1) against the (unstandardized) predicted values (PRE\_1), the variables SPSS created when you requested these measures in the **Linear Regression: Save** dialogue box. If your residuals form a horizontal band, as shown in the scatterplot below, the relationship between your dependent variable and independent variables is likely to be linear:



Example of a nonlinear relationship via scatterplot

### Partial regression plots

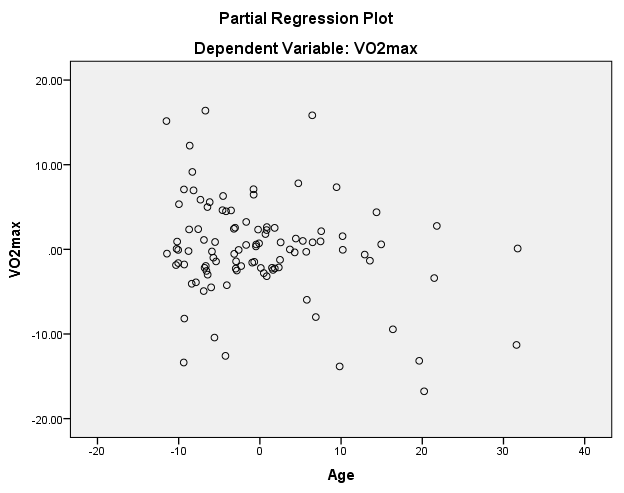
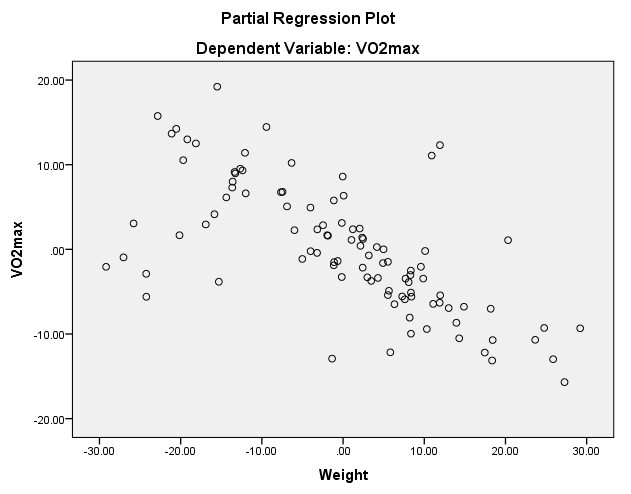
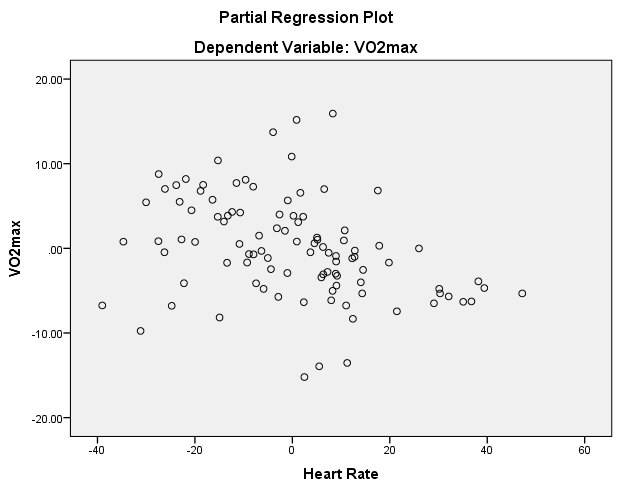
Partial regression plots will have been generated if you selected the Produce all partial plots in the **Linear Regression: Plots** dialogue box. There will be a partial plot between each independent variable and the dependent variable although you can ignore any categorical independent variables (e.g., gender). The partial regression plots should show a linear relationship.

### Partial regression plots

Partial regression plots will have been generated if you selected the Produce all partial plots in the **Linear Regression: Plots** dialogue box. There will be a partial plot between each independent variable and the dependent variable although you can ignore any categorical independent variables (e.g., gender). The partial regression plots should show a linear relationship.

**Age, Weight, and Heart Rate**

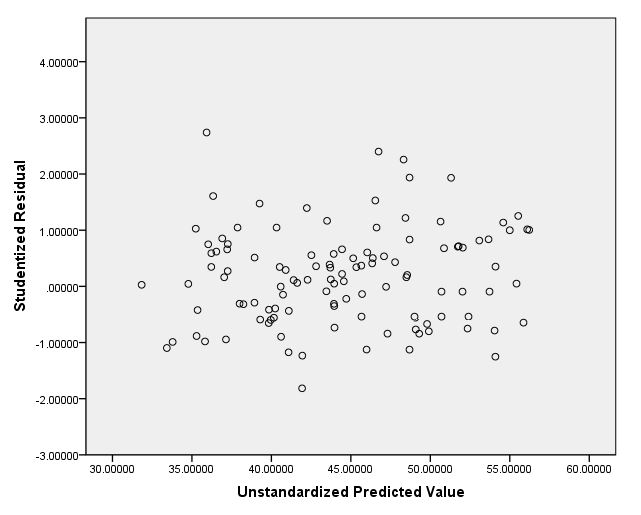
The partial regression plots below shows a somewhat linear relationship between VO2max and age; VO2max and weight; and VO2max and heart\_rate.



Note: If some or all of your relationships are non-linear, you will need to transform the variables involved in order to coax the variables into linear relationships. You will then have to re-run all analyses conducted so far, but with the newly transformed data.

## Checking for homoscedasticity

The assumption of homoscedasticity is that the residuals are equal for all values of the predicted dependent variable. To check for heteroscedasticity, you can use the plot you created to check linearity in the previous section, namely plotting the studentized residuals (SRE\_1) against the unstandardized predicted values (PRE\_1). This scatterplot is reproduced below:



If the residuals are not equally spread over the predicted values of the dependent variable, you have violated the assumption of homogeneity of variance. If there is homoscedasticity, the spread of the residuals will not increase or decrease as you move across the predicted values. In this example, you can see that there is homoscedasticity (i.e., the assumption has not been violated).

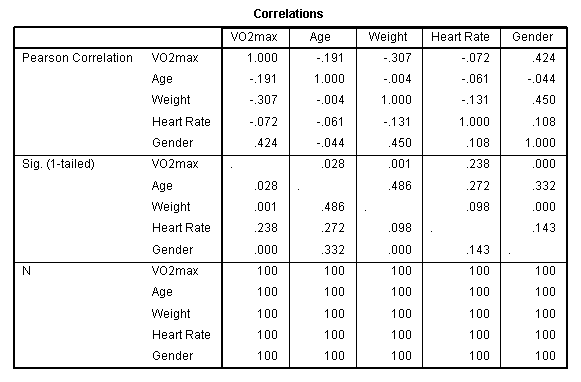
Note: If you have violated this assumption (i.e., you have heteroscedasticity), you will need to transform the dependent variable (or independent variables) in order to correct for this. You will then have to re-run all analyses conducted so far, but with the newly transformed data. An alternative, but much more advanced approach, is to use weight least squares (WLS) regression.

## Checking for multicollinearity

Multicollinearity occurs when you have two or more independent variables that are highly correlated with each other. This leads to problems with understanding which variable contributes to the variance explained and technical issues in calculating a multiple regression model. There are two stages to identifying multicollinearity: inspection of correlation coefficients and Tolerance/VIF values, as discussed below:

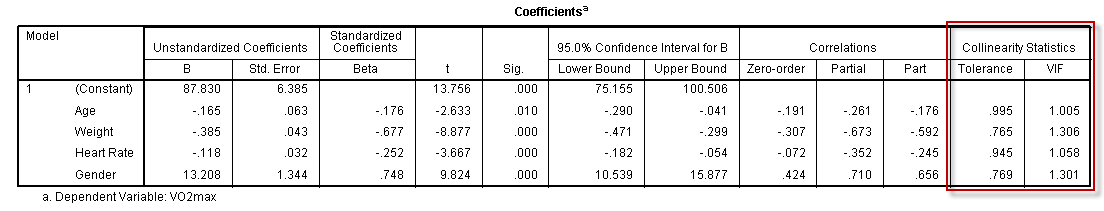
### Correlations

If you selected the Descriptives option in the **Linear Regression: Statistics** dialogue box, the **Correlations** table will be presented in the results, as shown below:



You need to check that none of the independent variables have correlations greater than 0.7. You can see from the **Correlations** table that there are no correlations larger than 0.7 in this example.

### Tolerance and VIF

Most importantly, you need to consult the "**Tolerance**" and "**VIF**" values in the **Coefficients** table, as highlighted below:

In reality, as VIF is simply the reciprocal of Tolerance (i.e., 1 divided by Tolerance), you need only consult one of these measures. If the Tolerance value is less than 0.1 – which is a VIF of greater than 10 – you might have a collinearity problem. In this example, all the Tolerance values are greater than 0.1 (the lowest is 0.765), so you can be fairly confident that you do not have a problem with collinearity in this particular data set.

Note: If you do have multicollinearity problems, these are very difficult to deal with. There are complicated methods that can be used, but the simplest solution is to simply drop one of the offending variables from the analysis. Selection of which variable to drop can be made on theoretical grounds.

There are three main objectives that you can achieve with the output from a multiple regression: (1) determine the proportion of the variation in the dependent variable explained by the independent variables; (2) predict dependent variable values based on new values of the independent variables; and (3) determine how much the dependent variable changes for a one unit change in the independent variables. All of these objectives will be answered in the following sections.

When interpreting and reporting your results from a multiple regression, we suggest working through two stages: (a) determine whether the multiple regression model is a good fit for the data and (b) understand the coefficients of the regression model. To recap:

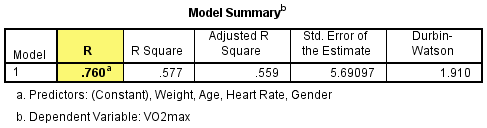
* First, you need to determine whether the multiple regression model is a good fit for the data: There are a number of statistics you can use to determine whether the multiple regression model is a good fit for the data. These are: (a) the multiple correlation coefficient, (b) the percentage (or proportion) of variance explained; (c) the statistical significance of the overall model; and (d) the precision of the predictions from the regression model. Therefore, on [page 16](https://statistics.laerd.com/premium/spss/mr/multiple-regression-in-spss-16.php) we take you through the **Model Summary** and **ANOVA** tables, which contain all the information you need to evaluate (a), (b) and (c), whilst (d) is addressed in the third bullet below.
* Second, you need to understand the coefficients of the regression model: Now that you have interpreted the overall model fit you can interpret and report the coefficients of the regression model on [page 17](https://statistics.laerd.com/premium/spss/mr/multiple-regression-in-spss-17.php). These coefficients are useful in order to understand whether there is a linear relationship between the dependent variable and the independent variables. In addition, you can use this regression equation to calculate predicted values of VO2max for a given set of values for age, weight, heart rate and gender.

### Determining how well the model fits

There are a number of measures you can use to determine whether the multiple regression model is a good fit for the data. These are: (a) the multiple correlation coefficient, (b) the percentage (or proportion) of variance explained; and (c) the statistical significance of the overall model.

##### Multiple correlation coefficient (R)

Therefore, let's first consider the value of the multiple correlation coefficient found in the "R" column of the **Model Summary** table, as highlighted below:



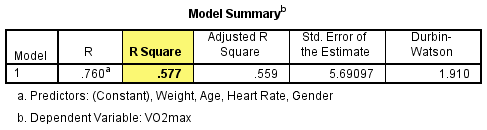
The multiple correlation coefficient, which can be abbreviated to just R, is simply the Pearson correlation coefficient between the scores predicted by the regression model (i.e., the predicted scores, PRE\_1) and the actual values of the dependent variable (i.e., the VO2max scores). As such, R is a measure of the strength of the linear association between these two variables and can give an indication as to the goodness of the model fit with a value that can range from 0 to 1, with higher values indicating a stronger linear association. A multiple correlation coefficient of 0 (zero) indicates no linear association between the dependent variable and the independent variables and a value of 1 a perfect linear association. A value of **0.760**, in this example, indicates a moderate to strong level of association. You should note, however, that the multiple correlation coefficient, R, is not a common measure used to assess goodness of fit. A much more popular method of assessing model fit is presented next.

##### Total variation explained (R2 and adjusted R2 )

The coefficient of determination – more commonly known as R2 – is a measure of the proportion of variance in the dependent variable that is explained by the independent variable. More specifically (and accurately), it is the proportion of variance in the dependent variable that is explained by the independent variables over and above the mean model. You might also hear this expressed as the proportion of variation accounted for by the regression model over and above the mean model. Let us explain this statement as it is very common for people to misinterpret what R2 measures.

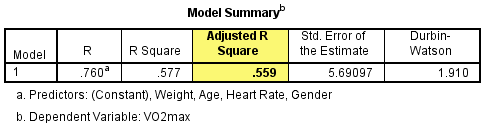
Given a desire to predict a dependent variable with multiple independent variables the simplest model we could choose is one without any independent variables at all. This is called the mean model and it is simply the mean of the dependent variable (VO2max in this example). In this situation, our best prediction of the dependent variable is its mean value. This is also the worst possible prediction (which makes sense when you think that we are not using any of our independent variables to help us). In this situation, you can assess the amount of variability in the model (i.e., as a measure of the error of prediction). Then, you run the multiple regression with all the independent variables added (which stands to reason will give you your best prediction as you are using all the available information) and measure the variability of this model (i.e., as a measure of the error of prediction). This model's variability will be lower than the mean model's variability because there has been a reduction in variability, which has been "caused" or "explained" by the addition of the independent variables. This is often expressed as a proportion or percentage and is what is referred to as R2. It assesses overall model fit.

This value of R2 is presented in the "**R Square**" column in the **Model Summary** table, as highlighted below:



You can see that R2 is equal to **0.577** in this example. This means that the addition of all our independent variables into a regression model explained **57.7%** (i.e., 0.577 x 100 = 57.7%) of the variability of our dependent variable, VO2max (compared to the mean model).

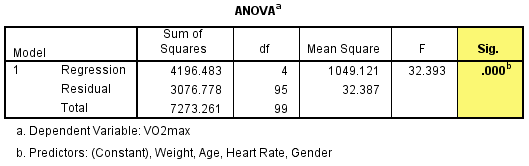
However, R2 is based on the sample and is considered a positively-biased estimate of the proportion of the variance of the dependent variable accounted for by the regression model (i.e., it is larger than it should be when generalizing to a larger population). Despite this criticism, it is still considered by some to be a good starting measure to understanding your results (Draper & Smith, 1998). That said, there is another measure called adjusted R2 which corrects for this positive bias in order to provide a value that would be expected in the population. The adjusted R2 value is found in the "**Adjusted R Square**" column of the **Model Summary** table, as highlighted below:



You can see that adjusted R2 is **0.559** in this example. Adjusted R2 will always be smaller than R2, but it is preferable that you use this value to report the proportion of variance explained (i.e., report 55.9% rather than 57.7%), although ideally you might be able to report both. Adjusted R2 is also an estimate of effect size, which at 0.559 (55.9%), is indicative of a large effect size according to Cohen's (1988) classification.

##### **Statistical significance of the model**

The statistical significance of the overall model (i.e., the model containing all independent variables) is presented in the "**Sig.**" column of the **ANOVA** table, as highlighted below:

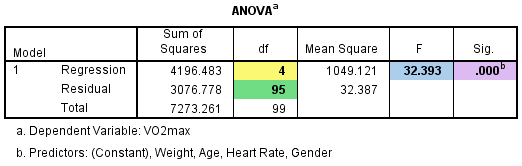


You can see that the "**Sig.**" value is **.000**, which actually means that p < .0005. If p < .05, you have a statistically significant result. On the other hand, if p > .05, you do not have a statistically significant result.

As p < .0005 satisfies p < .05, we have a statistically significant result. This means that the addition of all our independent variables (i.e., our overall model) leads to a model that: (a) is statistically significantly better at predicting the dependent variable than the mean model; and (b) is a statistically significantly better fit to the data than the mean model.

The null hypothesis of this test is that the multiple correlation coefficient, R, is equal to 0 (zero). You can also deduce from this result that at least one regression (slope) coefficient (i.e., except the intercept) is statistically significantly different to zero.

You would normally report the result as follows: F(4, 95) = 32.393, p < .0005; rather than just a p-value. The breakdown of the last part is as follows, F(4, 95) = 32.393, p < .0005, as shown by the highlighted cells in the **ANOVA** table below:



The meaning of the information in the **ANOVA** table is explained below:

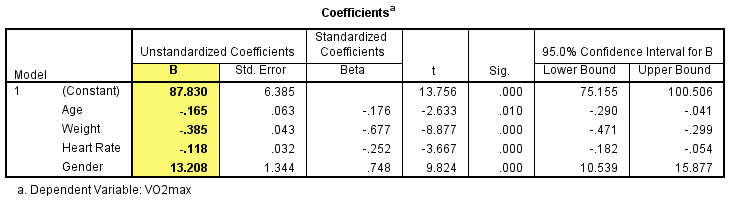
|  | Cell name | Cell meaning |
| --- | --- | --- |
|  | F | Indicates that we are comparing to an F-distribution (F-test). |
|  | 4 in (4, 95) | Indicates the regression (aka model) degrees of freedom ("df"). |
|  | 95 in (4, 95) | Indicates the residual (aka error) degrees of freedom ("df"). |
|  | 32.393 | Indicates the obtained value of the F-statistic (obtained F-value). |
|  | p < .0005 | Indicates the probability of obtaining the observed F-value if the null hypothesis is true. |

### Interpreting the coefficients

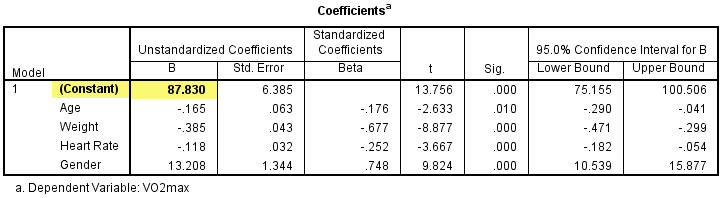
The regression equation for the current example can be expressed in the following form:

predicted VO2max = b0 + (b1 x age) + (b2 x weight) + (b3 x heart\_rate) + (b4 x gender)

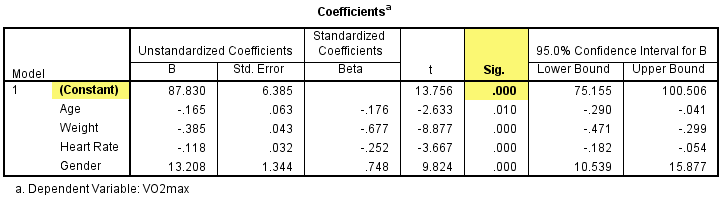
where b0 is the intercept (aka constant) and b1 through b4 are the slope coefficients (one for each variable). By substituting the values for b0through b4 you will be able to predict VO2max given any values you enter for age, weight, heart rate or gender. You can ascertain the value of these coefficients by inspecting the **Coefficients** table, as highlighted below:



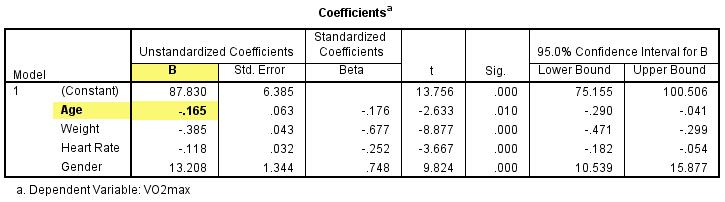
The intercept is called the constant in SPSS Statistics. The value of the intercept is found in the "**(Constant)**" column under the "**B**" column, as highlighted below:



The intercept is not usually of much interest. It is the value of the dependent variable when all the independent variables are zero. The intercept usually has no "real world" meaning, and we will not consider it in any more detail here. You can determine whether this intercept is statistically significant by consulting the "**(Constant)**" row under the "**Sig.**" column, as highlighted below:



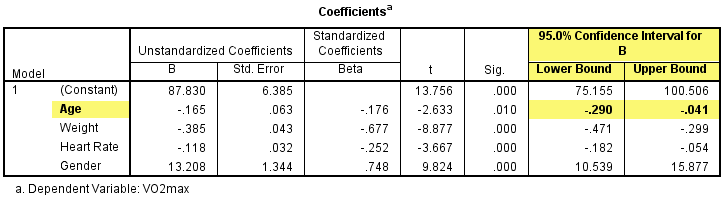
You can see that the intercept is statistically significant (i.e., p < **.0005**), meaning that it is different from 0 (zero). Again, this is of little interest. Much more importantly, and of much greater interest, are the slope coefficients. The first of these slope coefficients is for the variable, age, as reported in the "**Age**" row under the "**B**" column, as highlighted below:



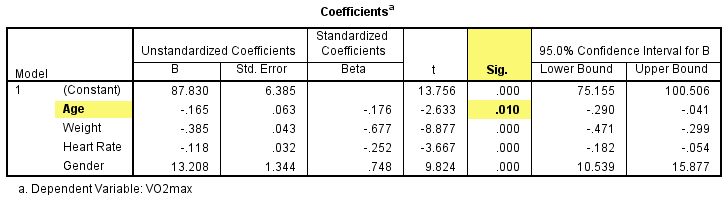
You can see that the coefficient for age is **-0.165**. The slope coefficient represents the change in the dependent variable for a one unit change in the independent variable. As such, an increase in age of one year is associated with a decrease in VO2max of 0.165 ml/min/kg. There is a decrease in VO2max because the slope coefficient is negative. If the slope coefficient had been positive then an increase in age would have been associated with an increase in VO2max. As it stands, our result makes sense. The multiple regression equation predicts that the older you are the lower your VO2max and this is known to be true – your aerobic capacity does in fact decrease with age (unfortunately!). It is important to note that this decrease in VO2max for each extra year in age is when all other independent variables are held constant. It does not matter what those values are, as long as they are kept constant.

If you prefer to consider differences in slope coefficients in units that are larger or smaller than the ones that the slope coefficient represents you can do this. For example, if you would prefer to report the decrease in VO2max that occurs every decade (i.e., 10 years) rather than every year, you can simply multiply your slope coefficient by 10 to get the decrease in VO2max every decade (i.e., every 10 years your VO2max is predicted to decrease by 1.651 ml/min/kg).

It is also possible to define a range of "plausible" values for the slope coefficient. The 95% confidence intervals are reported in the "**Age**" row in the "**Lower Bound**" and "**Upper Bound**" columns found under the "**95% Confidence Interval for B**" column, as highlighted below:



You can see that the 95% confidence intervals (CI) are between **-0.290** and **-0.041** ml/min/kg. That is, you can be 95% confident that the true value of the slope coefficient is between these lower and upper bounds. You can determine whether this slope coefficient is statistically significant by interpreting the value in the "**Sig.**" column, as highlighted below:

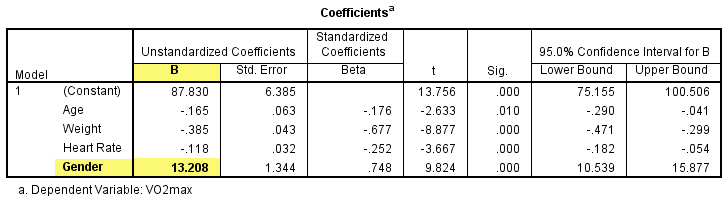


You can see that the p-value is **.010** (i.e., p = .010). If p <. 05, the slope coefficient is statistically significant. This means that the coefficient is statistically significantly different to 0 (zero). You can also interpret this as meaning that there is a linear relationship in the population.

If *p* > .05, you can declare that the slope coefficient is not statistically significant; that is, the slope coefficient is not different to 0 (zero) in the population (i.e., there is no linear relationship).

You can perform the same interpretations on the other continuous independent variables in your multiple regression. So, in our example, an increase in weight of 1 kg is associated with a decrease in VO2max of **0.385** ml/min/kg and an increase in heart rate of 1 bpm (beats per minute) is associated with a decrease in VO2max of **0.118** ml/min/kg.

However, a dichotomous independent variable such as gender has a different interpretation than that of continuous independent variables. In the dichotomous independent variable situation, the value of the slope coefficient represents the difference in the dependent variable between the two categories of the dichotomous independent variable. Remember that we coded the two categories of the gender variable as: 0 = females and 1 = males. The comparison between the two categories is with respect to the category with a value of 0. In this example, we are comparing males to females as females are coded as 0. That is, the coefficient represents the difference in predicted VO2max of males compared to females (i.e., males' VO2max minus females' VO2max). Another way of looking at this is that you are going from 0 to 1, which is going from females to males. As such, the coefficient represents the difference in VO2max for being male. The value of our slope coefficient is found in the "**Gender**" column under the "**B**" column, as highlighted below:



You can see that the value of our slope coefficient is **13.208**. This means that predicted VO2max for males is 13.208 ml/min/kg greater than that predicted for females (with all values of all other independent variables being held constant). So, all other things being equal, males have VO2max values that are 13.208 ml/min/kg (on average) greater than females. Again, this makes sense, as we know that males have higher VO2max values than females, on average. We can also evaluate the 95% confidence intervals and statistical significance of this difference in the same way as we did with our continuous independent variables.

You can now substitute the values of the coefficients into the regression equation, as shown below:

predicted VO2max = 87.83 – (0.165 x age) – (0.385 x weight) – (0.118 x heart\_rate) + (13.208 x gender)

A multiple regression was run to predict VO2max from gender, age, weight and heart rate. The multiple regression model statistically significantly predicted VO2max, F(4, 95) = 32.393, p < .001, adj. *r2* = .56. All four variables added statistically significantly to the prediction, p < .05. Regression coefficients and standard errors can be found in Table 1 (below).

### Summarizing your multiple regression analysis

You can present the results from the multiple regression analysis in a simple table, as shown below:

