From Laerd statistics

<https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php>

Multiple Regression Analysis using SPSS

Introduction

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

For example, you could use multiple regression to understand whether exam performance can be predicted based on revision time, test anxiety, lecture attendence, and gender. Alternately, you could use multiple regression to understand whether daily cigarette consumption can be predicted based on smoking duration, age when started smoking, smoker type, income, and gender.

Multiple regression 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. For example, you might want to know how much of the variation in exam performance can be explained by revision time, test anxiety, lecture attendence and gender "as a whole", but also the "relative contribution" of each independent variable in explaining the variance.

This "quick start" guide shows you how to carry out multiple regression using SPSS, as well as interpret and report the results from this test. However, before we introduce you to this procedure, you need to understand the different assumptions that your data must meet in order for multiple regression to give you a valid result. We discuss these assumptions next.

**SPSS**[**top ^**](https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php#top)

Assumptions

When you choose to analyse your data using multiple regression, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using multiple regression. You need to do this because it is only appropriate to use multiple regression if your data "passes" eight assumptions that are required for multiple regression to give you a valid result. In practice, checking for these eight assumptions just adds a little bit more time to your analysis, requiring you to click a few more buttons in SPSS when performing your analysis, as well as think a little bit more about your data, but it is not a difficult task.

Before we introduce you to these eight assumptions, do not be surprised if, when analysing your own data using SPSS, one or more of these assumptions is violated (i.e., not met). This is not uncommon when working with real-world data rather than textbook examples, which often only show you how to carry out multiple regression when everything goes well! However, don’t worry. Even when your data fails certain assumptions, there is often a solution to overcome this. First, let's take a look at these eight assumptions:

* **Assumption #1:** Your **dependent variable** should be measured on a continuous scale (i.e., it is either an **interval** or **ratio** variable). Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. You can learn more about interval and ratio variables in our article: [Types of Variable](https://statistics.laerd.com/statistical-guides/types-of-variable.php). If your dependent variable was measured on an **ordinal** scale, you will need to carry out [ordinal regression](https://statistics.laerd.com/spss-tutorials/ordinal-regression-using-spss-statistics.php) rather than multiple regression. Examples of **ordinal variables** include Likert items (e.g., a 7-point scale from "strongly agree" through to "strongly disagree"), amongst other ways of ranking categories (e.g., a 3-point scale explaining how much a customer liked a product, ranging from "Not very much", to "It is OK", to "Yes, a lot"). You can access our SPSS guide on ordinal regression [here](https://statistics.laerd.com/spss-tutorials/ordinal-regression-using-spss-statistics.php).
* **Assumption #2:** You have **two or more independent variables**, which can be either **continuous** (i.e., an **interval** or **ratio** variable) or categorical (i.e., an **ordinal** or **nominal** variable). For examples of **continuous variables**, see the bullet above. Examples of **ordinal variables** include Likert items (e.g., a 7-point scale from "strongly agree" through to "strongly disagree"), amongst other ways of ranking categories (e.g., a 3-point scale explaining how much a customer liked a product, ranging from "Not very much", to "It is OK", to "Yes, a lot"). Examples of **nominal variables** include gender (e.g., 2 groups: male and female), ethnicity (e.g., 3 groups: Caucasian, African American and Hispanic), physical activity level (e.g., 4 groups: sedentary, low, moderate and high), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth. Again, you can learn more about variables in our article: [Types of Variable](https://statistics.laerd.com/statistical-guides/types-of-variable.php).
* **Assumption #3:** You should have **independence of observations** (i.e., **independence of residuals**), which you can easily check using the Durbin-Watson statistic, which is a simple test to run using SPSS. We explain how to interpret the result of the Durbin-Watson statistic, as well as showing you the SPSS procedure required, in our enhanced multiple regression guide.
* **Assumption #4:** There needs to be a **linear relationship** between (a) the dependent variable and **each** of your independent variables, and (b) the dependent variable and the independent variables **collectively**. Whilst there are a number of ways to check for these linear relationships, we suggest creating **scatterplots** and **partial regression plots** using SPSS, and then visually inspecting these scatterplots and partial regession plots to check for linearity. If the relationship displayed in your scatterplots and partial regression plots are not linear, you will have to either run a non-linear regression analysis or "transform" your data, which you can do using SPSS. In our enhanced multiple regression guide, we show you how to: (a) create scatterplots and partial regression plots to check for linearity when carrying out multiple regression using SPSS; (b) interpret different scatterplot and partial regression plot results; and (c) transform your data using SPSS if you do not have linear relationships between your variables.
* **Assumption #5:** Your data needs to show **homoscedasticity**, which is where the variances along the line of best fit remain similar as you move along the line. We explain more about what this means and how to assess the homoscedasticity of your data in our enhanced multiple regression guide. When you analyse your own data, you will need to plot the studentized residuals against the unstandardized predicted values. In our enhanced multiple regression guide, we explain: (a) how to test for homoscedasticity using SPSS; (b) some of the things you will need to consider when interpreting your data; and (c) possible ways to continue with your analysis if your data fails to meet this assumption.
* **Assumption #6:** Your data must not show **multicollinearity**, which occurs when you have two or more independent variables that are highly correlated with each other. This leads to problems with understanding which independent variable contributes to the variance explained in the dependent variable, as well as technical issues in calculating a multiple regression model. Therefore, in our enhanced multiple regression guide, we show you: (a) how to use SPSS to detect for multicollinearity through an inspection of correlation coefficients and Tolerance/VIF values; and (b) how to interpret these correlation coefficients and Tolerance/VIF values so that you can determine whether your data meets or violates this assumption.
* **Assumption #7:** There should be **no significant outliers**, **high leverage points** or **highly influential points**. Outliers, leverage and influential points are different terms used to represent observations in your data set that are in some way unusual when you wish to perform a multiple regression analysis. These different classifications of unusual points reflect the different impact they have on the regression line. An observation can be classified as more than one type of usual point. However, all these points can have a very negative effect on the regression equation that is used to predict the value of the dependent variable based on the independent variables. This can change the output that SPSS produces and reduce the predictive accuracy of your results as well as the statistical significance. Fortunately, when using SPSS to run multiple regression on your data, you can detect possible outliers, high leverage points, and highly influential points. In our enhanced multiple regression guide, we: (1) show you how to detect outliers using "casewise diagnostics" and "studentized deleted residuals", which you can do using SPSS, and discuss some of the options you have in order to deal with outliers; (2) check for leverage points using SPSS, and discuss what you should do if you have any; and (3) check for influential points in SPSS using a measure of influence known as Cook's Distance, before presenting some practical approaches in SPSS to deal with any influential points you might have.
* **Assumption #8:** Finally, you need to check that the **residuals (errors)** are **approximately normally distributed** (we explain these terms in our enhanced multiple regression guide). Two common methods to check this assumption include using: (a) a histogram (with a superimposed normal curve) and a Normal P-P Plot; or (b) a Normal Q-Q Plot of the studentized residuals. Again, in our enhanced multiple regression guide, we: (a) show you how to check this assumption using SPSS, whether you use a histogram (with superimposed normal curve) and Normal P-P Plot, or Normal Q-Q Plot; (b) explain how to interpret these diagrams; and (c) provide a possible solution if your data fails to meet this assumption.

You can check assumptions #3, #4, #5, #6, #7 and #8 using SPSS. Assumptions #1 and #2 should be checked first, before moving onto assumptions #3, #4, #5, #6, #7 and #8. We suggest testing these assumptions in this order because it represents an order where, if a violation to the assumption is not correctable, you will no longer be able to use multiple regression (although you may be able to run another statistical test on your data instead). Just remember that if you do not run the statistical tests on these assumptions correctly, the results you get when running multiple regression might not be valid. This is why we dedicate a number of sections of our enhanced multiple regression guide to help you get this right. You can find out about our enhanced content as a whole [here](https://statistics.laerd.com/features-overview.php), or more specifically, learn how we help with testing assumptions [here](https://statistics.laerd.com/features-assumptions.php).

In the section, [Procedure](https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php#procedure), we illustrate the SPSS procedure to perform a multiple regression assuming that no assumptions have been violated. First, we introduce the example that is used in this guide.

**SPSS**[**top ^**](https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php#top)

Example

A health researcher wants to be able to predict "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 attributes that can be measured more easily and cheaply. To this end, a researcher recruited 100 participants to perform a maximum VO2max test, but also recorded their "age", "weight", "heart rate" and "gender". Heart rate is the average of the last 5 minutes of a 20 minute, much easier, lower workload cycling test. The researcher's goal is to be able to predict VO2max based on these four attributes: age, weight, heart rate and gender.

**SPSS**[**top ^**](https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php#top)

Setup in SPSS

In SPSS, we created seven variables: (1) VO2max, which is the maximal aerobic capacity; (2) age, which is the participant's age; (3)weight, which is the participant's weight (technically, it is their 'mass'); (4) heart\_rate, which is the participant's heart rate; (5)gender, which is the participant's gender; and (6) caseno, which is the case number. The caseno variable is used to make it easy for you to eliminate cases (e.g., "significant outliers", "high leverage points" and "highly influential points") that you have identified when checking for assumptions. In our enhanced multiple regression guide, we show you how to correctly enter data in SPSS to run a multiple regression when you are also checking for assumptions. You can learn about our enhanced data setup content [here](https://statistics.laerd.com/features-data-setup.php). Alternately, we have a generic, "quick start" guide to show you how to enter data into SPSS, available [here](https://statistics.laerd.com/spss-tutorials/entering-data-in-spss-statistics.php).

**SPSS**[**top ^**](https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php#top)

Test Procedure in SPSS

The six steps below show you how to analyse your data using multiple regression in SPSS when none of the eight assumptions in the previous section, [Assumptions](https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php#assumptions), have been violated. At the end of these six steps, we show you how to interpret the results from your multiple regression. If you are looking for help to make sure your data meets assumptions #3, #4, #5, #6, #7 and #8, which are required when using multiple regression, and can be tested using SPSS, you can learn more in our enhanced guide [here](https://statistics.laerd.com/features-overview.php).

* Click **Analyze > Regression > Linear...** on the main menu, as shown below:



**Note:** Don't worry that you're selecting **Analyze > Regression > "Linear...**" on the main menu, or that the dialogue boxes in the steps that follow have the title, **Linear Regression**. You have not made a mistake. You are in the correct place to carry out the multiple regression procedure. This is just the title that SPSS gives, even when running a multiple regression procedure.

* You will be presented with the **Linear Regression** dialogue box below:



* 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  buttons, as shown below (all other boxes can be ignored):



**Note:** For a standard multiple regression you should ignore  and  buttons as they are for sequential (hierarchical) multiple regression. The Method: option needs to be kept at the default value, which is "Enter". If, for whatever reason, "Enter" is not selected, you need to change Method: back to "Enter". The "Enter" method is the name given by SPSS to standard regression analysis.

* Click the  button. You will be presented with the **Linear Regression: Statistics** dialogue box, as shown below. In the -Regression Coefficients- area, leave Estimates ticked, as well as Model Fit in the top right hand corner.



* In addition to the options that are already selected, select Confidence intervals from the -Regression Coefficients- area and leave theLevel(%): at 95. Then select Descriptives. You will end up with the following screen:



* Click the  button. You will be returned to the **Linear Regression** dialogue box.
* Click the  button. This will generate the output.

*Join the 1,000s of students, academics and professionals who rely on Laerd Statistics.*[**TAKE THE TOUR**](https://statistics.laerd.com/features-overview.php) [**PLANS & PRICING**](https://statistics.laerd.com/pricing.php)

**SPSS**[**top ^**](https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php#top)

Interpreting and Reporting the Output of Multiple Regression Analysis

SPSS will generate quite a few tables of output for a multiple regression analysis. In this section, we show you only the three main tables required to understand your results from the multiple regression procedure, assuming that no assumptions have been violated. A complete explanation of the output you have to interpret when checking your data for the nine assumptions required to carry out multiple regression is provided in our enhanced guide. This includes relevant scatterplots and partial regression plots, histogram (with superimposed normal curve), Normal P-P Plot and Normal Q-Q Plot, correlation coefficients and Tolerance/VIF values, casewise diagnostics and studentized deleted residuals.

However, in this "quick start" guide, we focus only on the three main tables you need to understand your multiple regression results, assuming that your data has already met the nine assumptions required for multiple regression to give you a valid result:

**Determining how well the model fits**

The first table of interest is the **Model Summary** table. This table provides the *R*, *R2*, adjusted *R2*, and the standard error of the estimate, which can be used to determine how well a regression model fits the data:



The "**R**" column represents the value of *R*, the ***multiple correlation coefficient***. *R* can be considered to be one measure of the quality of the prediction of the dependent variable; in this case, VO2max. A value of 0.760, in this example, indicates a good level of prediction. The "**R Square**" column represents the *R2* value (also called the coefficient of determination), which is the proportion of variance in the dependent variable that can be explained by the independent variables (technically, it is the proportion of variation accounted for by the regression model above and beyond the mean model). You can see from our value of 0.577 that our independent variables explain 57.7% of the variability of our dependent variable, VO2max. However, you also need to be able to interpret "**Adjusted R Square**" (*adj. R2*) to accurately report your data. We explain the reasons for this, as well as the output, in our enhanced multiple regression guide.

**Statistical significance**

The *F*-ratio in the **ANOVA** table (see below) tests whether the overall regression model is a good fit for the data. The table shows that the independent variables statistically significantly predict the dependent variable, *F*(4, 95) = 32.393, *p* < .0005 (i.e., the regression model is a good fit of the data).



**Estimated model coefficients**

The general form of the equation to predict VO2max from age, weight, heart\_rate, gender, is:

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

This is obtained from the **Coefficients** table, as shown below:



Unstandardized coefficients indicate how much the dependent variable varies with an independent variable, when all other independent variables are held constant. Consider the effect of age in this example. The unstandardized coefficient, B1, for age is equal to -0.165 (see **Coefficients** table). This means that for each 1 year increase in age, there is a decrease in VO2max of 0.165 ml/min/kg.

**Statistical significance of the independent variables**

You can test for the statistical significance of each of the independent variables. This tests whether the unstandardized (or standardized) coefficients are equal to 0 (zero) in the population. If *p* < .05, you can conclude that the coefficients are statistically significantly different to 0 (zero). The *t*-value and corresponding *p*-value are located in the "**t**" and "**Sig.**" columns, respectively, as highlighted below:



You can see from the "**Sig.**" column that all independent variable coefficients are statistically significantly different from 0 (zero). Although the intercept, B0, is tested for statistical significance, this is rarely an important or interesting finding.

**Putting it all together**

You could write up the results as follows:

* General

A multiple regression was run to predict VO2max from gender, age, weight and heart rate. These variables statistically significantly predicted VO2max, *F*(4, 95) = 32.393, *p* < .0005, *R2* = .577. All four variables added statistically significantly to the prediction, *p* < .05.

If you are unsure how to interpret regression equations, or how to use them to make predictions, we discussed this in our enhanced multiple regression guide. We also show you how to write up the results from your assumptions tests and multiple regression output if you need to report this in a dissertation/thesis, assignment or research report. We do this using the Harvard and APA styles. You can learn more about our enhanced content [here](https://statistics.laerd.com/features-overview.php).