Linear Regression Analysis using SPSS Statistics

Introduction

Linear regression is the next step up after correlation. It is used when we want to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable (or sometimes, the outcome variable). The variable we are using to predict the other variable's value is called the independent variable (or sometimes, the predictor variable). For example, you could use linear regression to understand whether exam performance can be predicted based on revision time; whether cigarette consumption can be predicted based on smoking duration; and so forth. If you have two or more independent variables, rather than just one, you need to use multiple regression.

This "quick start" guide shows you how to carry out linear regression using SPSS Statistics, 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 linear regression to give you a valid result. We discuss these assumptions next.

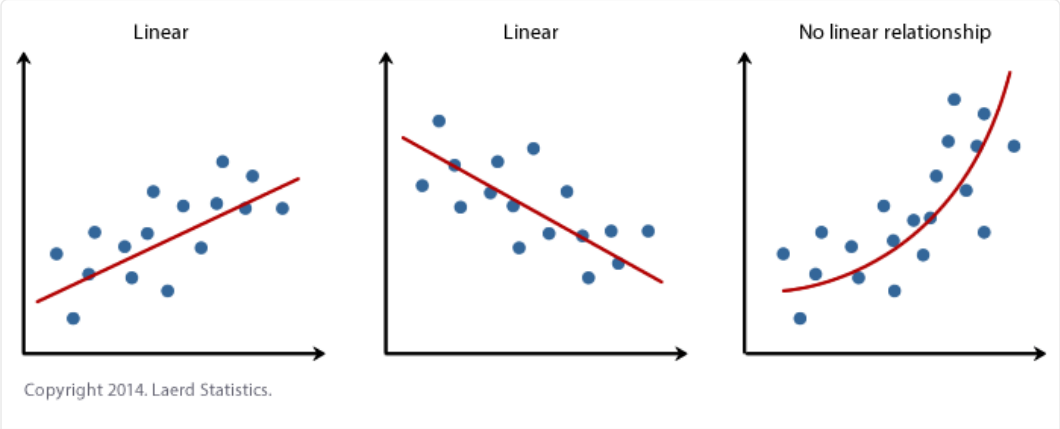
SPSS Statistics

Assumptions

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

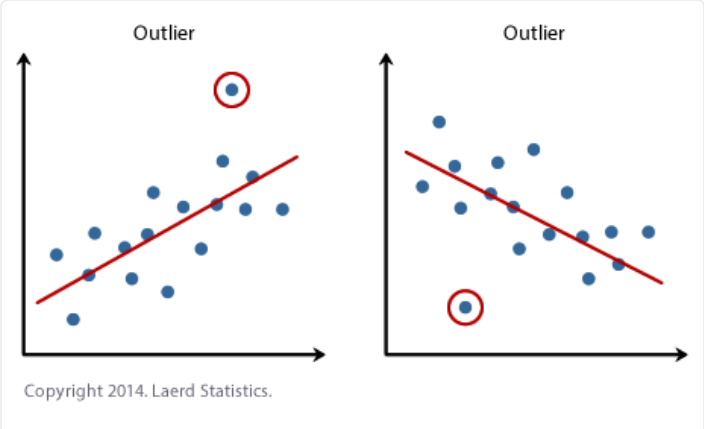
Before we introduce you to these six assumptions, do not be surprised if, when analysing your own data using SPSS Statistics, one or more of these assumptions is violated (i.e., not met). This is not uncommon when working with real-world data rather than textbook examples, which often only show you how to carry out linear 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 six assumptions:

- **Assumption #1:** Your two variables should be measured at the **continuous** level (i.e., they are either **interval** or **ratio** variables). Examples of **continuous variables** 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.
- **Assumption #2:** There needs to be a **linear relationship** between the two variables. Whilst there are a number of ways to check whether a linear relationship exists between your two variables, we suggest creating a scatterplot using SPSS Statistics where you can plot the dependent variable against your independent variable and then visually inspect the scatterplot to check for linearity. Your scatterplot may look something like one of the following:



If the relationship displayed in your scatterplot is not linear, you will have to either run a non-linear regression analysis, perform a polynomial regression or "transform" your data, which you can do using SPSS Statistics. In our enhanced guides, we show you how to: (a) create a scatterplot to check for linearity when carrying out linear regression using SPSS Statistics; (b) interpret different scatterplot results; and (c) transform your data using SPSS Statistics if there is not a linear relationship between your two variables.

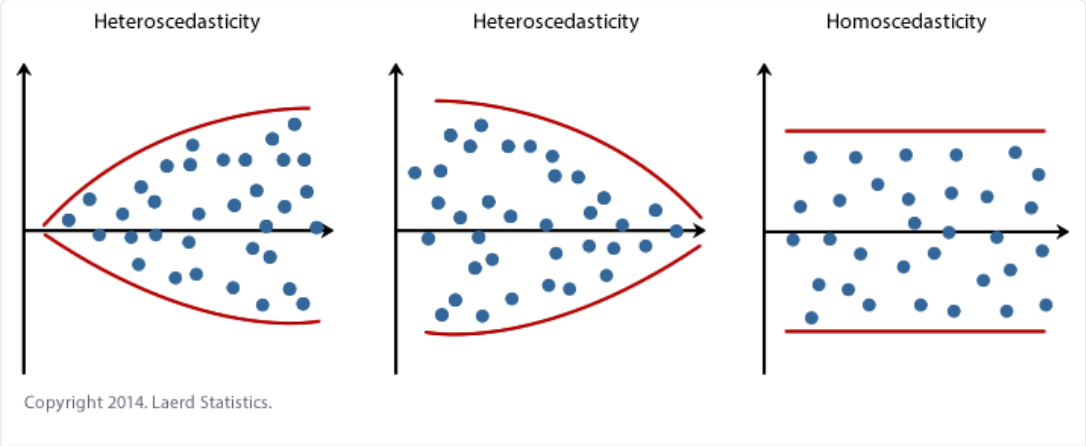
- **Assumption #3:** There should be **no significant outliers**. An outlier is an observed data point that has a dependent variable value that is very different to the value predicted by the regression equation. As such, an outlier will be a point on a scatterplot that is (vertically) far away from the regression line indicating that it has a large residual, as highlighted below:



The problem with outliers is that they can have a negative effect on the regression analysis (e.g., reduce the fit of the regression equation) that is used to predict the value of the dependent (outcome) variable based on the independent (predictor) variable. This will change the output that SPSS Statistics produces and reduce the predictive accuracy of your results. Fortunately, when using SPSS Statistics to run a linear regression on your data, you can easily include criteria to help you detect possible outliers. In our enhanced linear regression guide, we: (a) show you how to detect outliers using "casewise diagnostics", which is a simple process when using SPSS Statistics; and (b) discuss some of the options you have in order to deal with outliers.

- **Assumption #4:** You should have **independence of observations**, which you can easily check using the Durbin-Watson statistic, which is a simple test to run using SPSS Statistics. We explain how to interpret the result of the Durbin-Watson statistic in our enhanced linear regression guide.

- **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. Whilst we explain more about what this means and how to assess the homoscedasticity of your data in our enhanced linear regression guide, take a look at the three scatterplots below, which provide three simple examples: two of data that fail the assumption (called heteroscedasticity) and one of data that meets this assumption (called homoscedasticity):



Whilst these help to illustrate the differences in data that meets or violates the assumption of homoscedasticity, real-world data can be a lot more messy and illustrate different patterns of heteroscedasticity. Therefore, in our enhanced linear regression guide, we explain: (a) some of the things you will need to consider when interpreting your data; and (b) possible ways to continue with your analysis if your data fails to meet this assumption.

- **Assumption #6:** Finally, you need to check that the **residuals (errors)** of the regression line are **approximately normally distributed** (we explain these terms in our enhanced linear regression guide). Two common methods to check this assumption include using either a histogram (with a superimposed normal curve) or a Normal P-P Plot. Again, in our enhanced linear regression guide, we: (a) show you how to check this assumption using SPSS Statistics, whether you use a histogram (with superimposed normal curve) or Normal P-P 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 #2, #3, #4, #5 and #6 using SPSS Statistics. Assumptions #2 should be checked first, before moving onto assumptions #3, #4, #5 and #6. We suggest testing the assumptions in this order because assumptions #3, #4, #5 and #6 require you to run the linear regression procedure in SPSS Statistics first, so it is easier to deal with these after checking assumption #2. Just remember that if you do not run the statistical tests on these assumptions correctly, the results you get when running a linear regression might not be valid. This is why we dedicate a number of sections of our enhanced linear regression guide to help you get this right. You can find out more about our enhanced content as a whole on our Features: **Overview** page, or more specifically, learn how we help with testing assumptions on our Features: **Assumptions** page.

In the section, Procedure, we illustrate the SPSS Statistics procedure to perform a linear regression assuming that no assumptions have been violated. First, we introduce the example that is used in this guide.

SPSS Statistics

Example

A salesperson for a large car brand wants to determine whether there is a relationship between an individual's income and the price they pay for a car. As such, the individual's "income" is the independent variable and the "price" they pay for a car is the dependent variable. The salesperson wants to use this information to determine which cars to offer potential customers in new areas where average income is known.

SPSS Statistics

Setup in SPSS Statistics

In SPSS Statistics, we created two variables so that we could enter our data: Income (the independent variable), and Price (the dependent variable). It can also be useful to create a third variable, caseno, to act as a chronological case number. This third variable is used to make it easy for you to eliminate cases (e.g., significant outliers) that you have identified when checking for assumptions. However, we do not include it in the SPSS Statistics procedure that follows because we assume that you have already checked these assumptions. In our enhanced linear regression guide, we show you how to correctly enter data in SPSS Statistics to run a linear

regression when you are also checking for assumptions. You can learn about our enhanced data setup content on our Features: **Data Setup** page. Alternately, see our generic, "quick start" guide: Entering Data in SPSS Statistics.

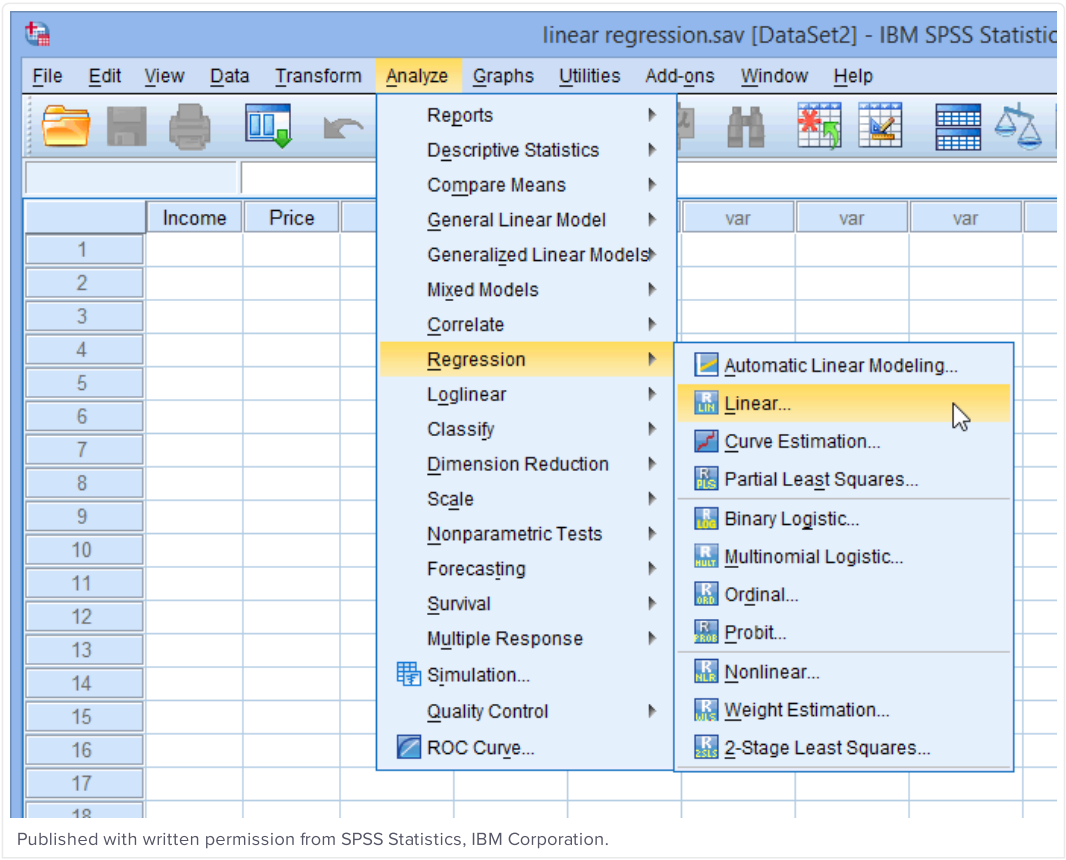
SPSS Statistics

Test Procedure in SPSS Statistics

The five steps below show you how to analyse your data using linear regression in SPSS Statistics when none of the six assumptions in the previous section, Assumptions, have been violated. At the end of these four steps, we show you how to interpret the results from your linear regression. If you are looking for help to make sure your data meets assumptions #2, #3, #4, #5 and #6, which are required when using linear regression and can be tested using SPSS Statistics, you can learn more about our enhanced guides on our Features: **Overview** page.

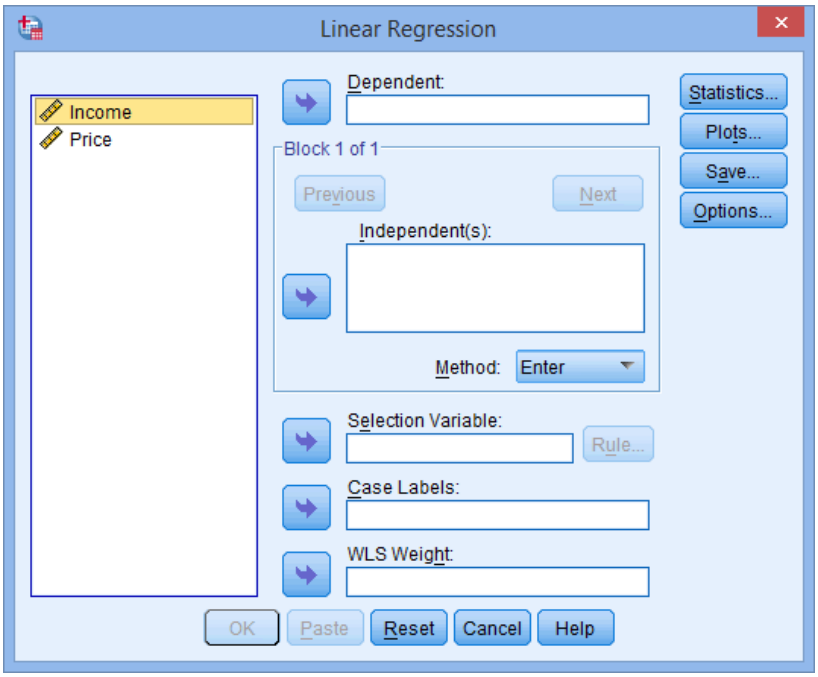
---

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



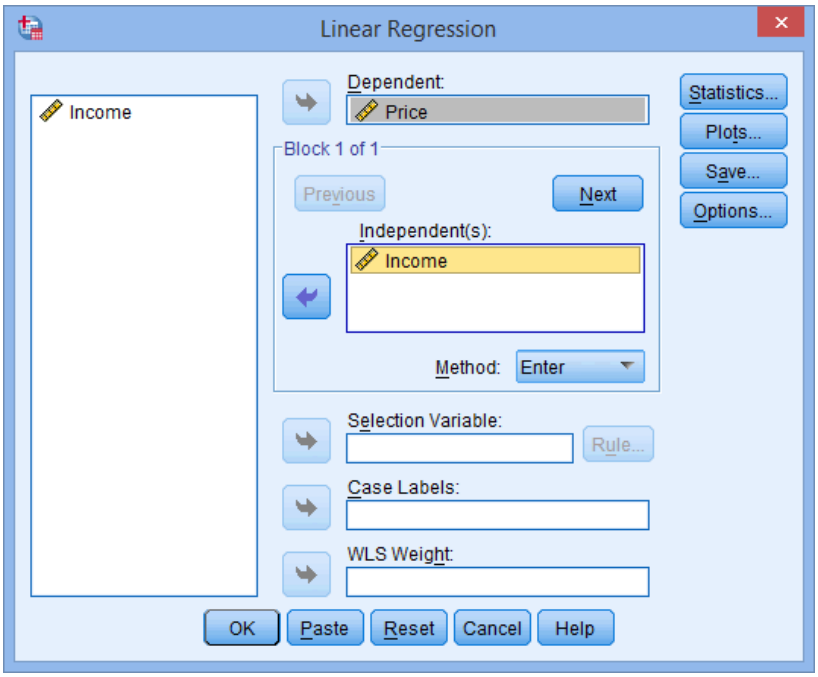
Published with written permission from SPSS Statistics, IBM Corporation.

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



2

Transfer the independent variable, Income, into the Independent(s): box and the dependent variable, Price, into the Dependent: box. You can do this by either drag-and-dropping the variables or by using the appropriate buttons. You will end up with the following screen:



3

You now need to check four of the assumptions discussed in the Assumptions section above: no significant outliers (assumption #3); independence of observations (assumption #4); homoscedasticity (assumption #5); and normal distribution of errors/residuals (assumptions #6). You can do this by using the Statistics... and Plots... features, and then selecting the appropriate options

within these two dialogue boxes. In our enhanced linear regression guide, we show you which options to select in order to test whether your data meets these four assumptions.

4 Click on the OK button. This will generate the results.

SPSS Statistics

Output of Linear Regression Analysis

SPSS Statistics will generate quite a few tables of output for a linear regression. In this section, we show you only the three main tables required to understand your results from the linear 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 six assumptions required to carry out linear regression is provided in our enhanced guide. This includes relevant scatterplots, histogram (with superimposed normal curve), Normal P-P Plot, casewise diagnostics and the Durbin-Watson statistic. Below, we focus on the results for the linear regression analysis only.

The first table of interest is the **Model Summary** table, as shown below:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.873[a]	.762	.749	874.779

a. Predictors: (Constant), Income

This table provides the *R* and R^2 values. The *R* value represents the simple correlation and is 0.873 (the "**R**" Column), which indicates a high degree of correlation. The R^2 value (the "**R Square**" column) indicates how much of the total variation in the dependent variable, Price, can be explained by the independent variable, Income. In this case, 76.2% can be explained, which is very large.

The next table is the **ANOVA** table, which reports how well the regression equation fits the data (i.e., predicts the dependent variable) and is shown below:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	44182633.37	1	44182633.37	57.737	.000[b]
	Residual	13774291.07	18	765238.393		
	Total	57956924.44	19			

a. Dependent Variable: Price

b. Predictors: (Constant), Income

This table indicates that the regression model predicts the dependent variable significantly well. How do we know this? Look at the "**Regression**" row and go to the "**Sig.**" column. This indicates the statistical significance of the regression model that was run. Here, $p < 0.0005$, which is less than 0.05, and indicates that, overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data).

The **Coefficients** table provides us with the necessary information to predict price from income, as well as determine whether income contributes statistically significantly to the model (by looking at the "**Sig.**" column). Furthermore, we can use the values in the "**B**" column under the "**Unstandardized Coefficients**" column, as shown below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8286.786	1852.256		4.474	.000
	Income	.564	.074	.873	7.598	.000

a. Dependent Variable: Price

to present the regression equation as:

$$\text{Price} = 8287 + 0.564(\text{Income})$$

If you are unsure how to interpret regression equations or how to use them to make predictions, we discuss this in our enhanced linear regression guide. We also show you how to write up the results from your assumptions tests and linear 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 on our Features: **Overview** page.

We also have a "quick start" guide on how to perform a linear regression analysis in Stata.