

Contents

1 Linear Regression Analysis

1.1 Introduction

Linear regression is used when you 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 consumptions 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*.

SPSS can be used to carry out linear regression, 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.

1.2 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 is appropriate for 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 when performing your analysis, as well as think a little bit more about your data, but it is not a difficult task.

Often 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 linear regression when everything goes well. However, 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 interval or ratio level (i.e., they are continuous). 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](#).
- **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 scatter-plot using SPSS, where you can plot the dependent variable against your independent variable, and then visually inspect the scatter-plot to check for linearity. Your scatter-plot 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 or *transform* your data, which you can do using SPSS. It is important to learn how to: (a) create a scatterplot to check for linearity when

carrying out linear regression using SPSS; (b) interpret different scatterplot results; and (c) transform your data using SPSS if there is not a linear relationship between your two variables.

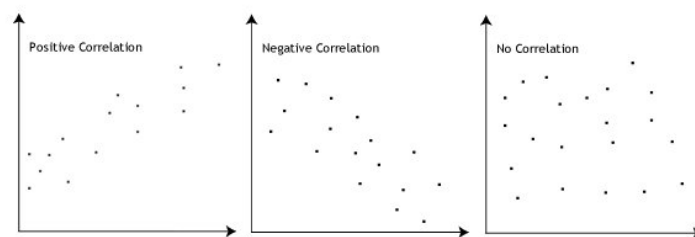


Figure 1: Types of Linear Relationship

- **Assumption 3:** There should be no significant outliers. Outliers are simply single data points within your data that do not follow the usual pattern (e.g., in a study of 100 students IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1% of IQ scores globally). The following scatterplots highlight the potential impact of outliers:

The problem with outliers is that they can have a negative effect on 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 produces and reduce the predictive accuracy of your results. Fortunately, when using SPSS to run linear regression on your data, you can easily include criteria to help you detect possible outliers.

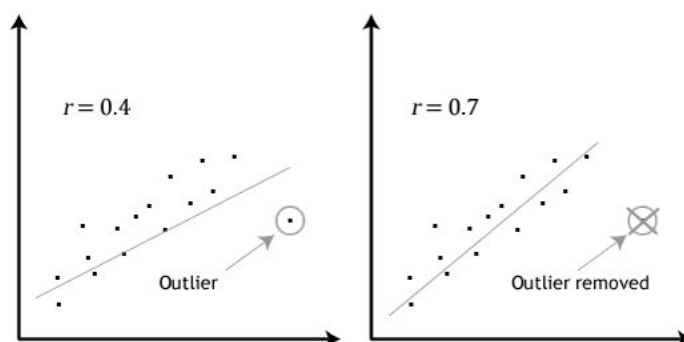


Figure 2: Effect of an Outlier

- **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. We explain how to interpret the result of the Durbin-Watson statistic later.
- **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 the linear regression line, take a look at the two scatter-plots below, which provide two simple examples: one of data that meets this assumption and one that fails the assumption:

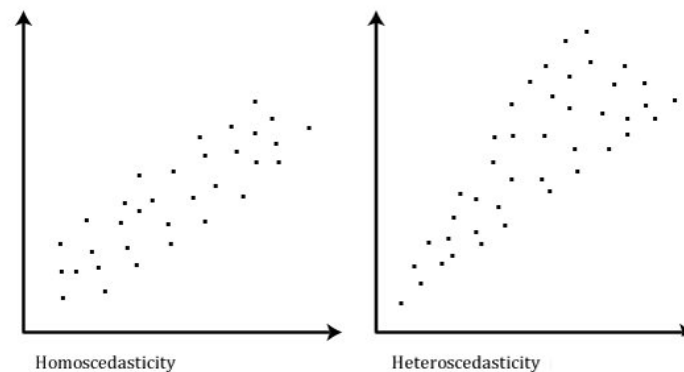


Figure 3: Constant Variance

When you analyse your own data, you will be lucky if your scatterplot looks like either of the two above. Whilst these help to illustrate the differences in data that meets or violates the assumption of homoscedasticity, real-world data is often a lot more messy. 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 your two variables 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 by using a Normal P-P Plot.

You can check assumptions all assumptions except no.1 using SPSS. It is recommended to test 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 a single linear 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 a linear regression might not be valid.

2 Output of Linear Regression Analysis

Linear regression is used when you 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).

$$\hat{y} = b_0 + b_1x$$

- b_0 is the intercept estimate
- b_1 is the slope estimate
- \hat{y} is the **fitted** y-value, given a specified value of the independent variable x .
- The fitted value and the observed value of y often differ. This difference is known as the **residual**

For example, you could use linear regression to understand whether exam performance can be predicted based on revision time; whether cigarette consumptions 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**.

SPSS will generate quite a few tables of output for a linear regression procedure. Only the three main tables required to understand your results from the linear regression procedure, assuming that no assumptions have been violated.

This includes relevant scatterplots, histogram (with superimposed normal curve) and Normal P-P Plot, and case-wise diagnostics and Durbin-Watson statistic tables. Below, we focus on the results for the linear regression analysis only.

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

Figure 4: Model Summary table

The first table of interest is the **Model Summary** table. This table provides the R and R^2 value. The R value is 0.873, which represents the simple correlation. It indicates a high degree of correlation. The R^2 value indicates how much of the dependent variable, **price** (Not evident on output), 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. This table indicates that the regression model predicts the outcome 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 applied. Here, the p-value is $p < 0.0005$, which is less than 0.05, and indicates that, overall, the model applied can statistically significantly predict the outcome variable.

The next table again, **Coefficients**, provides us with information on each predictor variable. This gives us the information we need to predict price from income. We can see that both the

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.418E7	1	4.418E7	57.737	.000 ^a
	Residual	1.377E7	18	765238.393		
	Total	5.796E7	19			

a. Predictors: (Constant), Income

b. Dependent Variable: Price

Figure 5: ANOVA Table

constant and income contribute significantly to the model (by looking at the Sig. column). By looking at the B column under the Unstandardized Coefficients column, we can present the regression equation as:

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

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

a. Dependent Variable: Price

Figure 6: Coefficients Table

3 Multiple Linear Regression

3.1 What is Multiple Linear Regression

Multiple regression is a statistical technique that allows us to predict a numeric value on the response variable on the basis of the observed values on several other independent variables.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots$$

- \hat{y} is the **fitted value** for the dependent variable Y , given a linear combination of values for the independent variables.
- x_i is the value for independent variable X_i . (For Example, x_1 is the value for independent variable X_1 .)
- b_0 is the constant regression estimate (commonly known as the **Intercept Estimate** in the case of simple linear regression).
- b_i is the regression estimate for Independent Variable X_i (commonly known as the **Slope Estimate** in the case of simple linear regression).

3.1.1 Simple Example

Suppose we were interested in predicting how much an individual enjoys their job. Independent Variables such as salary, extent of academic qualifications, age, sex, number of years in full-time employment and socioeconomic status might all contribute towards **job satisfaction**.

If we collected data on all of these variables, perhaps by surveying a few hundred members of the public, we would be able to see how many and which of these variables gave rise to the most accurate prediction of job satisfaction. We might find that job satisfaction is most accurately predicted by type of occupation, salary and years in full-time employment, with the other variables not helping us to predict job satisfaction.