

Chapter 855

Linear Regression

Introduction

Linear regression is a commonly used procedure in statistical analysis. One of the main objectives in linear regression analysis is to test hypotheses about the slope (sometimes called the regression coefficient) of the regression equation. This module calculates power and sample size for testing whether the slope is a value other than the value specified by the null hypothesis.

Difference between Linear Regression and Correlation

The correlation coefficient is used when both X and Y are from the normal distribution (in fact, the assumption actually is that X and Y follow a bivariate normal distribution). That is, X is assumed to be a random variable whose distribution is normal. In the linear regression context, no statement is made about the distribution of X . In fact, X is not even a random variable. Instead, it is a set of fixed values such as 10, 20, 30 or -1, 0, 1. Because of this difference in definition, we have included both Linear Regression and Correlation algorithms. They gave different results. This module deals with the Linear Regression (fixed X) case.

Technical Details

Suppose that the dependence of a variable Y on another variable X can be modeled using the simple linear equation

$$Y = A + BX$$

In this equation, A is the Y -intercept, B is the slope, Y is the dependent variable, and X is the independent variable.

The nature of the relationship between Y and X is studied using a sample of N observations. Each observation consists of a data pair: the X value and the Y value. The values of A and B are estimated from these observations. Since the linear equation will not fit the observations exactly, estimated values of A and B must be used. These estimates are found using the method of least squares. Using these estimated values, each data pair may be modeled using the equation

$$Y_i = a + bX_i + e_i$$

Note that a and b are the estimates of the population parameters A and B . The e values represent the discrepancies between the estimated values ($a + bX$) and the actual values Y . They are called the errors or residuals.

If it is assumed that these e values are normally distributed, tests of hypotheses about A and B can be constructed. Specifically, we can employ an F ratio to test the null hypothesis that the slope is B_0 versus the alternative hypothesis that the slope is B (B not B_0). The power function of this F test can be written

$$Power = \Pr(F > F_\alpha)$$

where F_α is the critical value based on the central- F distribution with 1 and $N - 2$ degrees of freedom and the significant level α and F is distributed as a non-central F with degrees of freedom 1 and $N - 2$ and non-centrality parameter λ . The value of λ is

Linear Regression

$$\lambda = N \left(\frac{SX(B - B_0)}{\sigma} \right)^2$$

where σ^2 is the variance of the e 's and

$$SX = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}}$$

The values of each of the parameters α , β , σ^2 , N , SX , and B can be determined from the others using the above formulation.

Note that the power for a one-sided test may be found by using 2α for α in the above formulation.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design tab. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains most of the parameters and options that you will be concerned with.

Solve For

Solve For

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power* for a power analysis or *Sample Size* for sample size determination.

Select *Sample Size* when you want to calculate the sample size needed to achieve a given power and alpha level.

Select *Power* when you want to calculate the power of an experiment.

Test

Alternative Hypothesis

Specify whether the test is one-sided or two-sided. When a two-sided hypothesis is selected, the value of alpha is halved by *PASS*. Everything else remains the same.

Note that the accepted procedure is to use the Two Sided option unless you can justify using a one-sided test.

Power and Alpha

Power

This option specifies one or more values for power. Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Linear Regression

Alpha

This option specifies one or more values for the probability of a type-I error (alpha). A type-I error occurs when you reject the null hypothesis when in fact it is true.

Values of alpha must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

This is the number of observations in the study.

Effect Size – Slope

B0 (Slope|H0)

This is the value of the slope assumed by the null hypothesis. Often this value is set to zero, but this is not necessary. The alternative hypothesis is that the slope is not equal to this value.

B1 (Slope|H1)

This is the value of the slope at which the power is computed. The hypothesis being tested is that the slope is some value other than B0.

Effect Size – Standard Deviation of X's

SX (Standard Deviation of X's)

This is the standard deviation of the X values in the sample. It is not necessarily the standard deviation of X in the population. For example, suppose the X values are five -1's and five 1's. The computed standard deviation of these values (dividing by N rather than $N - 1$) is 1.0.

You will often want to compute the power or sample size for a specific set of X values. Instead of computing SX by hand, you can use the keyword XS (short for X 's) followed by the list of X values. For example, the phrase

`XS -1,1`

is translated into a 1.0 (which is the standard deviation of these two values). This calculation assumes that the sample is allocated equally to the two values. Hence, an N of 10 implies that five are assigned to -1 and five to 1.

If you are planning a study involving two random variables, X and Y , that come from a bivariate normal population, you should enter the actual standard deviation of X here.

Effect Size – Residual Variance Calculation

Residual Variance Method

The standard deviation of the residuals is needed for the power and sample size calculations. These residuals are the e_i in the regression model

$$Y_i = a + bX_i + e_i$$

However, their standard deviation is not available until after the study is complete. *PASS* provides three methods for specifying the standard deviation of the residuals: *Std Deviation of Y*, *Correlation*, and specifying it directly.

Linear Regression

SY (Standard Deviation of Y)

Enter an estimate of the standard deviation of Y . This standard deviation ignores X . An estimate of this value must be found from previous studies, pilot studies, or using your best guess. This option is used when 'Residual Variance Method' is set to 'SY'.

When this value is used, the standard deviation of the residuals is computed using the relationship

$$\sigma = \sqrt{\sigma_Y^2 - B^2 SX^2}$$

This value must be greater than zero.

R (Correlation)

Enter an estimate of the correlation between Y and the X values. An estimate of this correlation must be found from previous studies, pilot studies, or using your best guess. This value must be greater than zero and less than one—negative values are allowed. This option is used when 'Residual Variance Method' is set to 'R'.

When this method is used, the standard deviation of the residuals is computed using the relationship

$$\sigma = B(SX)\sqrt{1/R^2 - 1}$$

where R is the correlation.

S (Standard Deviation of Residuals)

Enter an estimate for the value of the standard deviation of the residuals. This option is used when 'Residual Variance Method' is set to 'S'.

Linear Regression

Example 1 – Calculating the Power

Suppose a power analysis must be conducted for a linear regression study that will test the relationship between two variables, Y and X . The test will look at the power using two significance levels, 0.01 and 0.05 and several sample sizes between 5 and 85. Based on previous studies, the standard deviation of Y will be assumed to be 1.0. The standard deviation of the X 's in the sample will also be assumed as 1.0. The experimenter decides that unless the slope is at least 0.5, the relationship between X and Y is too weak to be considered.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Linear Regression** procedure window by clicking on **Regression**, and then clicking on **Linear Regression**. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
Design Tab	
Solve For	Power
Alternative Hypothesis	Two-Sided
Alpha.....	0.01 0.05
N (Sample Size).....	5 to 85 by 10
B0 (Slope H0)	0.0
B1 (Slope H1)	0.5
SX (Standard Deviation of X's).....	1
Residual Variance Method	SY (Std. Dev. of Y)
SY (Standard Deviation of Y)	1

Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Two-Sided Testing of $B=B_0$ where $B_0 = 0.00$

Power	Sample Size (N)	Slope (B)	Standard Deviation of X (SX)	Standard Deviation of Y (SY)	Standard Deviation of Residuals (S)	Alpha	Beta
0.03533	5	0.50	1.00	1.00	0.87	0.01000	0.96467
0.15194	5	0.50	1.00	1.00	0.87	0.05000	0.84806
0.26836	15	0.50	1.00	1.00	0.87	0.01000	0.73164
0.54369	15	0.50	1.00	1.00	0.87	0.05000	0.45631
0.54097	25	0.50	1.00	1.00	0.87	0.01000	0.45903
0.78944	25	0.50	1.00	1.00	0.87	0.05000	0.21056
0.74759	35	0.50	1.00	1.00	0.87	0.01000	0.25241
0.91225	35	0.50	1.00	1.00	0.87	0.05000	0.08775
0.87415	45	0.50	1.00	1.00	0.87	0.01000	0.12585
0.96601	45	0.50	1.00	1.00	0.87	0.05000	0.03399
0.94183	55	0.50	1.00	1.00	0.87	0.01000	0.05817
0.98755	55	0.50	1.00	1.00	0.87	0.05000	0.01245
0.97470	65	0.50	1.00	1.00	0.87	0.01000	0.02530
0.99564	65	0.50	1.00	1.00	0.87	0.05000	0.00436
0.98953	75	0.50	1.00	1.00	0.87	0.01000	0.01047
0.99853	75	0.50	1.00	1.00	0.87	0.05000	0.00147
0.99585	85	0.50	1.00	1.00	0.87	0.01000	0.00415
0.99952	85	0.50	1.00	1.00	0.87	0.05000	0.00048

Linear Regression

Report Definitions

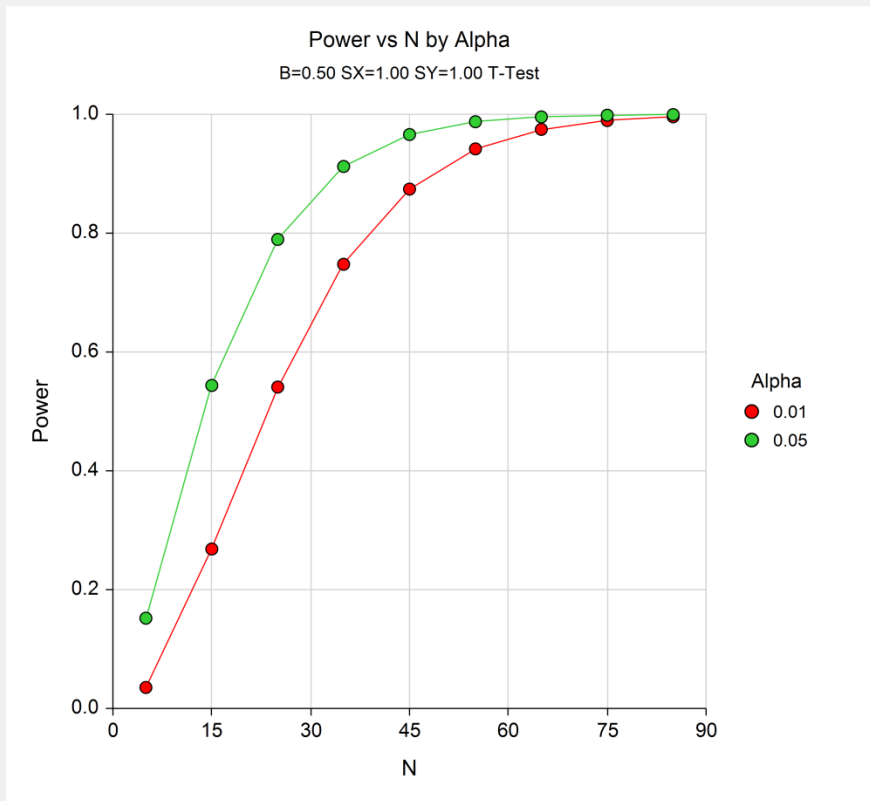
Power is the probability of rejecting a false null hypothesis. It should be close to one.
 N is the size of the sample drawn from the population. To conserve resources, it should be small.
 B0 is the slope under the null hypothesis.
 B is the slope at which the power is calculated.
 SX is the standard deviation of the X values.
 SY is the standard deviation of Y.
 Alpha is the probability of rejecting a true null hypothesis. It should be small.
 Beta is the probability of accepting a false null hypothesis. It should be small.

Summary Statements

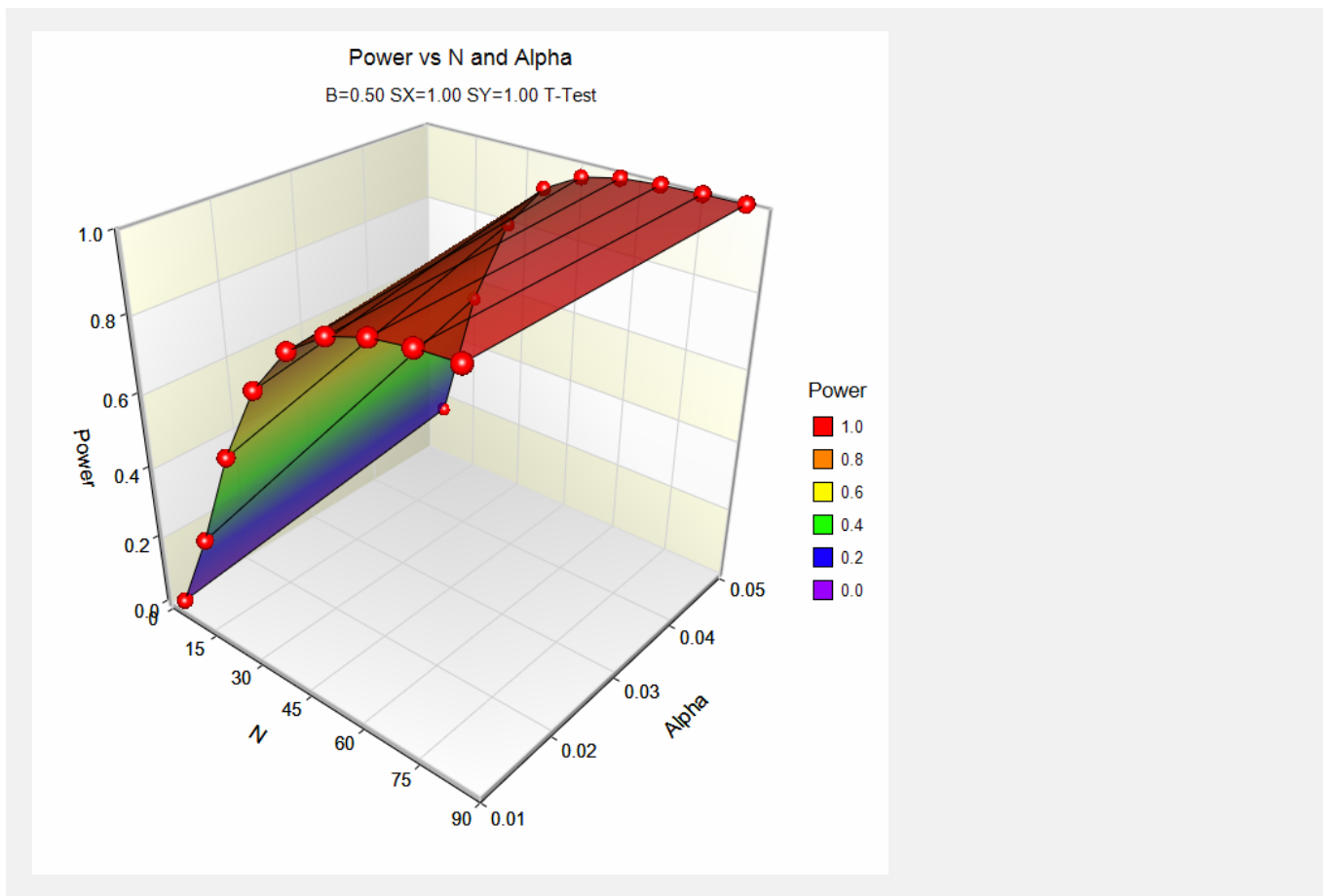
A sample size of 5 achieves 4% power to detect a change in slope from 0.00 under the null hypothesis to 0.50 under the alternative hypothesis when the standard deviation of the X's is 1.00, the standard deviation of Y is 1.00, and the two-sided significance level is 0.01000.

This report shows the calculated sample size for each of the scenarios.

Plots Section



Linear Regression



These plots show the power versus the sample size for the two values of alpha.

Linear Regression

Example 2 – Validation using Neter, Wasserman, and Kutner

Neter, Wasserman, and Kutner (1983) pages 71 and 72 present a power analysis when $N = 10$, $Slope = 0.25$, $\alpha = 0.05$, $SX = \sqrt{(3400 / 10)} = 18.439$, and $SY = \sqrt{10 + (0.25)^2(3400 / 10)} = 5.59015$. They found the power to be approximately 0.97.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Linear Regression** procedure window by clicking on **Regression**, and then clicking on **Linear Regression**. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
Design Tab	
Solve For	Power and Beta
Alternative Hypothesis	Two-Sided
Alpha.....	0.05
N (Sample Size).....	10
B0 (Slope H0)	0.00
B1 (Slope H1)	0.25
SX (Standard Deviation of X's).....	18.439
Residual Variance Method	SY (Std. Dev. of Y)
SY (Standard Deviation of Y)	5.59015

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Two-Sided Testing of $B=B_0$ where $B_0 = 0.00$							
Power	Sample Size (N)	Slope (B)	Standard Deviation of X (SX)	Standard Deviation of Y (SY)	Standard Deviation of Residuals (S)	Alpha	Beta
0.97975	10	0.25	18.44	5.59	3.16	0.05000	0.02025

The power of 0.97975 matches their result to two decimals. Note that they used interpolation from a table to obtain their answer.