Chapter 740

Simple Linear Regression

Introduction

Simple linear regression is a commonly used procedure in statistical analysis to model a linear relationship between a dependent variable Y and an independent variable X. One of the main objectives in simple 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 different from zero. The conditional power calculation method is used.

Difference between Simple Linear Regression and Correlation

The correlation coefficient is used when *X* and *Y* are from a bivariate normal distribution. That is, *X* is assumed to be a random variable whose distribution is normal. The values of X will not be known until the study is completed. In the simple linear regression context, no statement is made about the distribution of *X*. In fact, *X* does not have to be a random variable. In this procedure the distribution of Y is conditioned on X.

Fixed or Random X

Gatsonis and Sampson (1989) present power analysis results for two approaches: *unconditional* and *conditional*. This procedure provides a calculation for the *conditional* (fixed X) approach.

The *unconditional* approach assumes that X is normally distributed and is based on the correlation coefficient. The normality assumption might occasionally be met, but not frequently. Our impression is that usually, the values of X will not be known at the planning stage and they will not follow (even approximately) the normal distribution. Hence, the only option available is to proceed with the sample size calculation using the *conditional* approach to power calculation and estimate the standard deviation of the X's as best you can.

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 a T-test to test the null hypothesis that the B is 0 versus the alternative hypothesis that the slope is something else.

Linear Regression Slope T-Test Statistic

It is anticipated that a *t*-test of a regression coefficient will be used to conduct the test. Hence, the formula of the test statistic is

$$t_{N-2} = \frac{b - 0}{s_b}$$

where *N* is the sample size, *b* is the estimate of *B*, and s_b is the standard error of *b*.

Power Calculation of the Test of the Regression Coefficient, B

The following presentation is based on the standard results for a t-test as shown by Neter, Wasserman, and Kutner (1983) pages 71 and 72.

The power is calculated as follows for a directional alternative (one-tailed test) in which $H1: B > 0$.

1. Find $t_{1-\alpha}$ such that $T_{df}(t_{1-\alpha}) = 1 - \alpha$, where $T_{df}(x)$ is the area under a central-*t* curve to the left of *x* and $df = N - 2$.
2. Calculate: $X_0 = (t_{1-\alpha})\sigma_e/\sqrt{N}$.
3. Calculate the noncentrality parameter: $\lambda = \sqrt{N}(B1 - 0)\sigma_X/\sigma_e$, where σ_X is the standard deviation of the X values in the regression and *B1* is the slope at which the power is to be calculated.
4. Calculate: $t_1 = \left(X_0 - (B1 - 0)\sigma_X\sqrt{N}/\sigma_e\right) + \lambda$.
5. Power = $1 - T'_{df,\lambda}(t_1)$, where $T'_{df,\lambda}(x)$ is the area to the left of *x* under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter λ.

The sample size can be easily found using a binary search with this power formula.

Calculation of σ_X

The above calculation requires the value of σ_X, the (population) standard deviation of the X values in the regression analysis. Except for the occasional experimental design that includes them (e.g., doses), the specific X values are unknown in the planning phase. Hence, a reasonable estimate must be found. PASS includes a special tool called the *Standard Deviation Estimator* that will aid in your search for accurate estimates of this parameter.

The following table provides examples of typical data configurations and their corresponding standard deviations.

σ_X	X Values	σ_X	X Values	σ_X	X Values	σ_X	X Values
0.500	1, 2	0.816	1, 2, 3	1.118	1, 2, 3, 4	1.414	1, 2, 3, 4, 5
1.000	1, 3	1.633	1, 3, 5	2.236	1, 3, 5, 7	2.828	1, 3, 5, 7, 9
1.500	1, 4	2.449	1, 4, 7	3.354	1, 4, 7, 10	4.243	1, 4, 7, 10, 13
2.000	1, 5	3.266	1, 5, 9	4.472	1, 5, 9, 13	5.657	1, 5, 9, 13, 17
4.000	1, 9	6.532	1, 9, 17	8.944	1, 9, 17, 25	11.314	1, 9, 17, 25, 33

Because of the direct impact on the power and sample size, it will be important to spend some time determining appropriate values for this parameter.

One final note: when a basic pattern is repeated, its population standard deviation remains the same. For example, the standard deviation of the values 1, 2, 1, 2, 1, 2, 1, 2 is 0.5. This is also the standard deviation of 1, 2 or 1, 2, 1, 2.

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 the alternative hypothesis assessed by the test.

Let B represent the slope variable and 0 represent the specific value of the slope assumed by the null hypothesis.

Possible choices are:

- **H1: B ≠ 0**
  This is a two-sided hypothesis. The hypothesis set is H0: B = 0 vs. H1: B ≠ 0.
- **H1: B < 0**
  This is a lower, one-sided hypothesis. The hypothesis set is H0: B ≥ 0 vs. H1: B < 0.
- **H1: B > 0**
  This is an upper, one-sided hypothesis. The hypothesis set is H0: B ≤ 0 vs. H1: B > 0.

Note that this parameter impacts the value of alpha.

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 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.

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)

Enter one or more values for the number of observations (e.g., subjects) in the study.

Effect Size – Slope

B1 (Slope|H1)

Enter one or more values of the slope assumed by the alternative hypothesis, B1. This represents the actual value of B at which the power is computed.

You can enter a single value such as ‘1’ or a series of values such as ‘1 2 3’ or ‘1 to 10 by 1”.

The only restriction is that B1 ≠ 0.

Effect Size – Standard Deviation of X

σx Input Type

Select the method you want to use to enter the value(s) of σx. Your choices are

- **σx (Std Dev of X)**
  Enter one or more values for σx directly.
- **List of X Values**
  Enter a list of two or more numbers from which the standard deviation is to be calculated.

σx (Standard Deviation of X)

Enter one or more values for σx, the *population* standard deviation of the X values that will occur in a sample.

Usually, the actual X values are not known at the planning stage. When they are not known, you will have to estimate this value. You can press the *Standard Deviation Estimator* button at the right to obtain help in determining appropriate values for this parameter. Just be sure to use the *population*, not the *sample*, formula. That is, divide the sum of squares by N, not N-1.

The individual numbers can be any numeric value: positive, negative, or zero.

Fixed Xs

Determine the standard deviation of a typical set of fixed Xs. For example, suppose the X values will be five -1's and five 1's. The population standard deviation of these values (dividing by N, not N - 1) is 1.0. This is the value of σx. Note that '1 2' will result in the same σx as '1 2 1 2', '1 1 1 2 2 2', '-1 -2', or '11 12'.

Random Xs

Estimate one or more values of σx, the standard deviation of X, from your knowledge of X. If nothing else is available, you can use the likely range divided by 4, 5, or 6.

List of X Values

Enter a list of values from which the value of σx will be calculated. For example, entering "1, 3" results in σx = 1.0. Note that this calculation assumes that the N observations are allocated equally among the X's.

Effect Size – σe (Standard Deviation of Residuals)

σe Input Type

Select the method to use to enter σe (the standard deviation of the residuals).

- **σe (Std Dev of Residuals)**
  Specify σe directly.
- **σY (Std Dev of Y)**
  Specify σY. Calculate: $\sigma e^2 = \sigma Y^2 - B1^2 (\sigma x^2)$

σe (Std Dev of Residuals)

Enter one or more values of the standard deviation of the residuals from the regression of Y on X. The possible range is $0 < \sigma e$.

σY (Std Dev of Y)

Enter one or more values for the standard deviation of Y, ignoring the independent variable X. The value of σY is converted to σe using the formula: $\sigma e^2 = \sigma Y^2 - B1^2 (\sigma x^2)$. The allowable range is $0 < |B1(\sigma x)| < \sigma Y$.

Example 1 – Calculating the Power

Suppose a power analysis is required for a simple linear regression study that will test the relationship between two variables, *Y* and *X*. The analysis will look at the power of several sample sizes between 5 and 20. A one-sided test will be used with a significance level of 0.025. Based on previous studies, σe will be assumed to be 0.6. σx will assume that X is binary with equally-likely values of -1 and 1. The experimenter wants to test whether the slope is greater than zero. The power will be computed at B1 = 0.3, 0.4, and 0.5.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Simple Linear Regression** procedure window. 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**.

Option	Value
Design Tab	
Solve For	**Power**
Alternative Hypothesis	**One-Sided (H1: B > 0)**
Alpha	**0.025**
N (Sample Size)	**5 10 15 20**
B1 (Slope|H1)	**0.3 0.4 0.5**
σx Input Type	**List of X Values**
List of X Values	**-1 1**
σe Input Type	**σe (Std Dev of Residuals)**
σe (Std Dev of Residuals)	**0.6**

Annotated Output

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

Numeric Results

Hypotheses: H0: B ≤ 0 vs. H1: B > 0

Power	Sample Size N	Actual Slope B1	Std Dev of X σx	Std Dev of Y σY	Std Dev of Resids σe	R^2	Alpha
0.1242	5	0.300	1.000	0.671	0.600	0.200	0.025
0.1845	5	0.400	1.000	0.721	0.600	0.308	0.025
0.2585	5	0.500	1.000	0.781	0.600	0.410	0.025
0.2859	10	0.300	1.000	0.671	0.600	0.200	0.025
0.4581	10	0.400	1.000	0.721	0.600	0.308	0.025
0.6378	10	0.500	1.000	0.781	0.600	0.410	0.025
0.4338	15	0.300	1.000	0.671	0.600	0.200	0.025
0.6658	15	0.400	1.000	0.721	0.600	0.308	0.025
0.8466	15	0.500	1.000	0.781	0.600	0.410	0.025
0.5620	20	0.300	1.000	0.671	0.600	0.200	0.025
0.8049	20	0.400	1.000	0.721	0.600	0.308	0.025
0.9408	20	0.500	1.000	0.781	0.600	0.410	0.025

References

Dupont, W.D. and Plummer, W.D. Jr. 1998. 'Power and Sample Size Calculations for Studies Involving Linear Regression'. Controlled Clinical Trials, Vol. 19, Pages 589-601.
Sampson, Allan R. 1974. 'A Tale of Two Regressions'. JASA, Vol. 69, No. 347, Pages 682-689.
Neter, J., Wasserman, W., and Kutner, M. 1983. Applied Linear Regression Models. Richard D. Irwin, Inc. Chicago, Illinois.

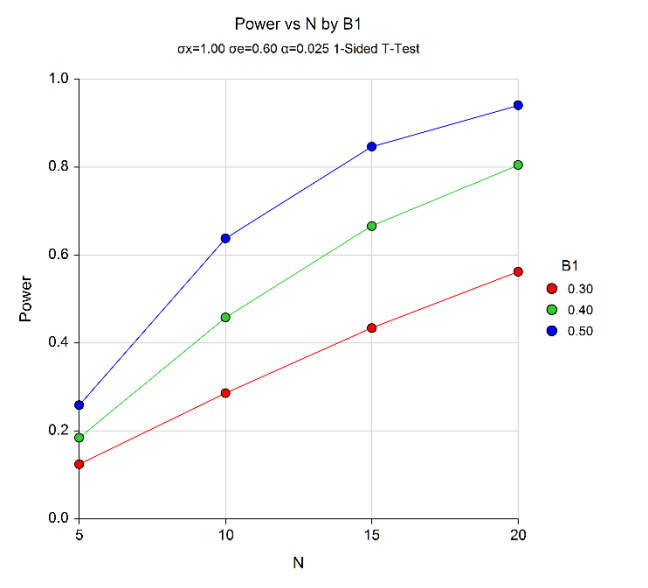
Simple Linear Regression

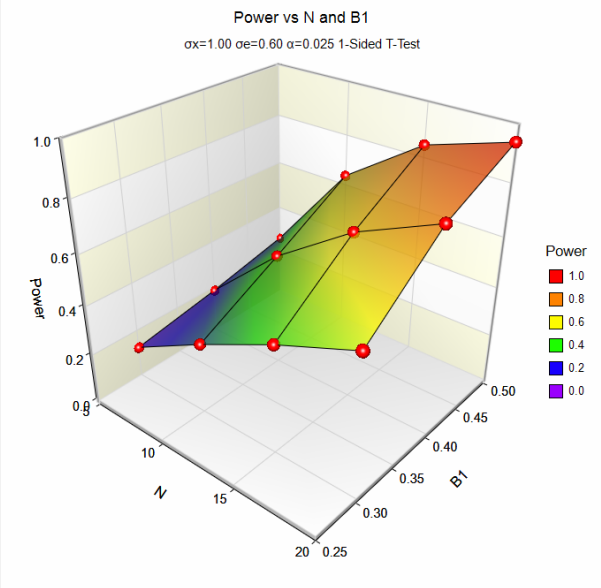
Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population. To conserve resources, it should be small.
The slope under the null hypothesis is assumed to be 0.
B1 is the slope at which the power is calculated.
σx is the standard deviation of the X values.
σY is the standard deviation of Y (ignoring X).
σe is the standard deviation of the residuals.
R^2 is the R-squared when Y is regressed on X.
Alpha is the probability of rejecting a true null hypothesis.

Summary Statements
A sample size of 5 achieves 12% power to detect a change in slope from 0.000 under the null hypothesis to 0.300 under the alternative hypothesis when the statistical hypothesis is one-sided, the significance level is 0.025, the standard deviation of X is 1.000, the standard deviation of Y is 0.671, the standard deviation of residuals is 0.600, and R^2 is 0.200.

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

Plots Section





These plots show the power versus the sample size for the three values of B1.

Example 2 – Validation using Neter, Wasserman, and Kutner (1983)

Neter, Wasserman, and Kutner (1983) pages 71 and 72 present a power analysis when

$$N = 10, B1 = 0.25, \alpha = 0.05, \sigma_X = \sqrt{3400/10} = 18.439, \qquad \sigma_e = \sqrt{10} = 3.16228.$$

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 **Simple Linear Regression** procedure window. 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**.

Option	Value
Design Tab	
Solve For	**Power**
Alternative Hypothesis	**Two-Sided (H1: B ≠ 0)**
Alpha	**0.05**
N (Sample Size)	**10**
B1 (Slope|H1)	**0.25**
σx Input Type	**σx (Std Dev of X)**
σx (Std Dev of X)	**18.439**
σe Input Type	**σe (Std Dev of Residuals)**
σe (Std Dev of Residuals)	**3.16228**

Output

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

Numeric Results

Hypotheses: H0: B = 0 vs. H1: B ≠ 0

Power	Sample Size N	Actual Slope B1	Std Dev of X σx	Std Dev of Y σY	Std Dev of Resids σe	R^2	Alpha
0.9797	10	0.250	18.439	5.590	3.162	0.680	0.050

The power of 0.9797 matches their approximate result to two decimals. Note that they used interpolation from a table to obtain their answer.

Example 3 – Observational Study given in Dupont and Plummer (1998)

Dupont and Plummer (1998) page 593 present a power analysis example for an *observational* study in which the values of the X variable is not fixed. In this example,

$$N = 100, B1 = -0.0667, \alpha = 0.05, \sigma_X = 7.5, \qquad \sigma_Y = 4.$$

They found the power to be approximately 0.24.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Simple Linear Regression** procedure window. You may then make the appropriate entries as listed below, or open **Example 3** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Alternative Hypothesis	**Two-Sided (H1: B ≠ 0)**
Alpha	**0.05**
N (Sample Size)	**100**
B1 (Slope|H1)	**-0.0667**
σx Input Type	**σx (Std Dev of X)**
σx (Std Dev of X)	**7.5**
σe Input Type	**σY (Std Dev of Y)**
σY (Std Dev of Y)	**4**

Output

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

Numeric Results

Hypotheses: H0: B = 0 vs. H1: B ≠ 0

Power	Sample Size N	Actual Slope B1	Std Dev of X σx	Std Dev of Y σY	Std Dev of Resids σe	R^2	Alpha
0.2390	100	-0.067	7.500	4.000	3.969	0.016	0.050

The power of 0.2390 matches their result of 0.24 to two decimals.

Example 4 – Fixed-X Sample Size Study given in Dupont and Plummer (1998)

Dupont and Plummer (1998) pages 593-594 present a sample size example for a *fixed-X* study in which the values of the X variable are fixed at 10, 30, and 50. In this example,

$$Power = 0.90, B1 = 0.01, \alpha = 0.05, r = 0.4.$$ So, using the relationship $r = \frac{B\sigma_X}{\sigma_Y}$,

we obtain $\sigma_Y = \frac{0.01(16.330)}{0.4} = 0.4082$. They found the sample size to be 57.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Simple Linear Regression** procedure window. You may then make the appropriate entries as listed below, or open **Example 4** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Alternative Hypothesis	**Two-Sided (H1: B ≠ 0)**
Power	**0.9**
Alpha	**0.05**
B1 (Slope|H1)	**0.01**
σx Input Type	**List of X Values**
List of X Values	**10 30 50**
σe Input Type	**σY (Std Dev of Y)**
σY (Std Dev of Y)	**0.4082**

Output

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

Numeric Results

Hypotheses: H0: B = 0 vs. H1: B ≠ 0

Power	Sample Size N	Actual Slope B1	Std Dev of X σx	Std Dev of Y σY	Std Dev of Resids σe	R^2	Alpha
0.9044	58	0.010	16.330	0.408	0.374	0.160	0.050

PASS has computed the sample size as 58, one more than the 57 Dupont and Plummer found. We assume that this is a rounding problem. **PASS** computed the power for an N of 57 to be 0.8993 which does round to 0.90 but is actually slightly less than the desired 0.90.