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.

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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**

Output

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

Numeric Reports

Numeric Results

Solve For: Power
Hypotheses: H0: B ≤ 0 vs. H1: B > 0

			Standard Deviation				
Power	Sample Size N	Slope | H1 B1	X σx	Y σY	Residuals σe	R-Squared R^2	Alpha
0.1242	5	0.3	1	0.671	0.6	0.200	0.025
0.1845	5	0.4	1	0.721	0.6	0.308	0.025
0.2585	5	0.5	1	0.781	0.6	0.410	0.025
0.2859	10	0.3	1	0.671	0.6	0.200	0.025
0.4581	10	0.4	1	0.721	0.6	0.308	0.025
0.6378	10	0.5	1	0.781	0.6	0.410	0.025
0.4338	15	0.3	1	0.671	0.6	0.200	0.025
0.6658	15	0.4	1	0.721	0.6	0.308	0.025
0.8466	15	0.5	1	0.781	0.6	0.410	0.025
0.5620	20	0.3	1	0.671	0.6	0.200	0.025
0.8049	20	0.4	1	0.721	0.6	0.308	0.025
0.9408	20	0.5	1	0.781	0.6	0.410	0.025

Power	The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N	The size of the sample drawn from the population.
B1	The slope at which the power is calculated. This is the slope under H1. The slope under H0 is assumed to be 0.
σx	The standard deviation of the X values.
σY	The standard deviation of Y (ignoring X).
σe	The standard deviation of the residuals.
R^2	The R-squared value when Y is regressed on X.
Alpha	The probability of rejecting a true null hypothesis.

Summary Statements

A simple linear regression (single group, Y versus X) design will be used to test whether the slope (B) is greater than 0 (H0: B ≤ 0 versus H1: B > 0). The comparison will be made using a one-sided simple linear regression slope t-test with a Type I error rate (α) of 0.025. The standard deviation of X is assumed to be 1 (calculated from the given list of X values), and the standard deviation of residuals is assumed to be 0.6 (corresponding to a standard deviation of Y of 0.671 and an R^2 of 0.2). To detect a slope of 0.3 with a sample size of 5, the power is 0.1242.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	5	7	2
20%	10	13	3
20%	15	19	4
20%	20	25	5

Dropout Rate	The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.
N	The evaluable sample size at which power is computed (as entered by the user). If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated power.
N'	The total number of subjects that should be enrolled in the study in order to obtain N evaluable subjects, based on the assumed dropout rate. N' is calculated by inflating N using the formula N' = N / (1 - DR), with N' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)
D	The expected number of dropouts. D = N' - N.

Dropout Summary Statements

Anticipating a 20% dropout rate, 7 subjects should be enrolled to obtain a final sample size of 5 subjects.

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.

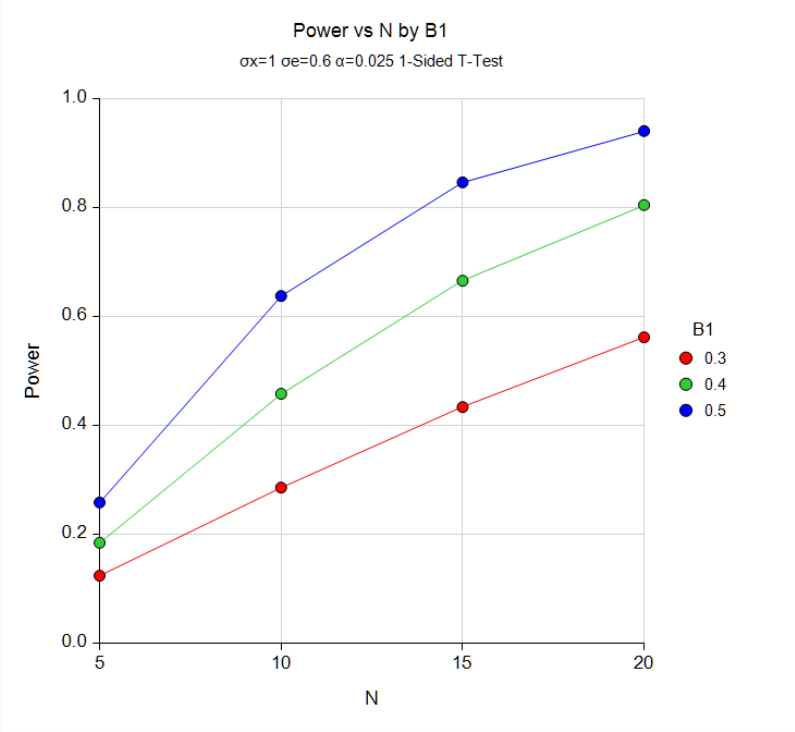
Sampson, Allan R. 1974. 'A Tale of Two Regressions'. JASA, Vol. 69, No. 347, Pages 682-689.

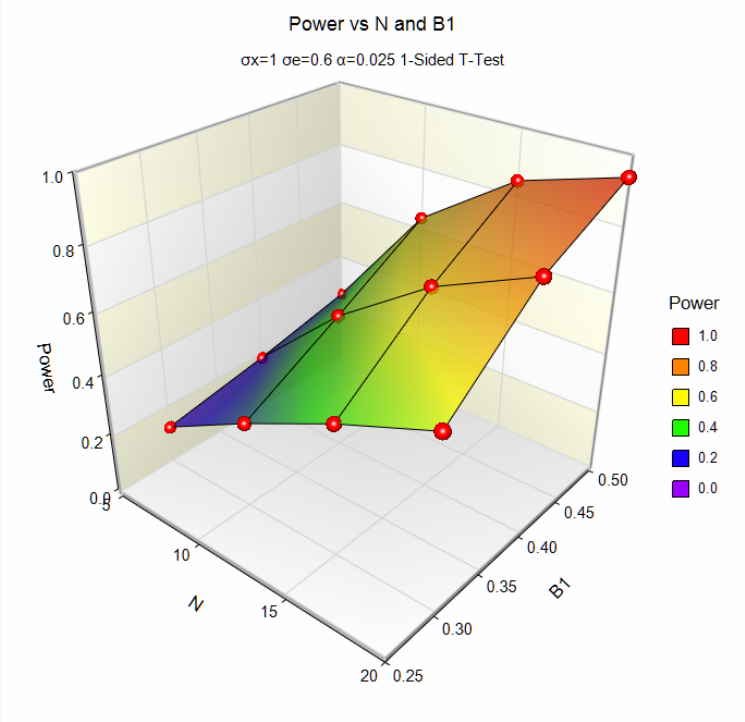
Neter, J., Wasserman, W., and Kutner, M. 1983. Applied Linear Regression Models. Richard D. Irwin, Inc. Chicago, Illinois.

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

Plots Section

Plots





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

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

They found the power to be approximately 0.97.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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

Solve For: Power
Hypotheses: H0: B = 0 vs. H1: B ≠ 0

			Standard Deviation				
Power	Sample Size N	Slope | H1 B1	X σx	Y σY	Residuals σe	R-Squared R²	Alpha
0.9797	10	0.25	18.439	5.59	3.162	0.68	0.05

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,

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

They found the power to be approximately 0.24.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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

Solve For: Power
Hypotheses: H0: B = 0 vs. H1: B ≠ 0

			Standard Deviation				
Power	Sample Size N	Slope | H1 B1	X σx	Y σY	Residuals σe	R-Squared R^2	Alpha
0.239	100	-0.067	7.5	4	3.969	0.016	0.05

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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

Solve For: Sample Size
Hypotheses: H0: B = 0 vs. H1: B ≠ 0

Power	Sample Size N	Slope | H1 B1	Standard Deviation X σx	Standard Deviation Y σY	Standard Deviation Residuals σe	R-Squared R^2	Alpha
0.9044	58	0.01	16.33	0.408	0.374	0.16	0.05

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.