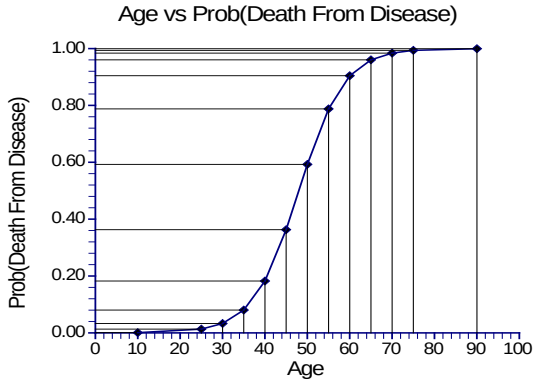
Chapter 860

Logistic Regression (Retired)

Introduction

Logistic regression expresses the relationship between a binary response variable and one or more independent variables called *covariates*. A covariate can be discrete or continuous.

Consider a study of death from disease at various ages. This can be put in a logistic regression format as follows. Let a binary response variable Y be one if death has occurred and zero if not. Let X be the individual's age. Suppose a large group of various ages is followed for ten years and then both Y and X are recorded for each person. In order to study the pattern of death versus age, the age values are grouped into intervals and the proportions that have died in each age group are calculated. The results are displayed in the following plot.



As you would expect, as age increases, the proportion dying of disease increases. However, since the proportion dying is bounded below by zero and above by one, the relationship is approximated by an "S" shaped curve. Although a straight-line might be used to summarize the relationship between ages 40 and 60, it certainly could not be used for the young or the elderly.

Under the logistic model, the proportion dying, P, at a given age can be calculated using the formula

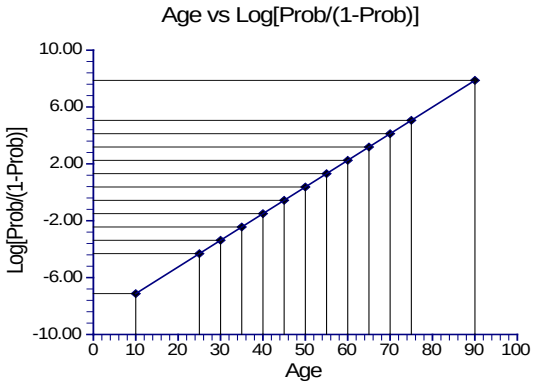
$$P = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

This formula can be rearranged so that it is linear in X as follows

$$Log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X$$

Note that the left side is the logarithm of the odds of death versus non-death and the right side is a linear equation for X. This is sometimes called the *logit* transformation of P. When the scale of the vertical axis of the plot is modified using the logit transformation, the following straight-line plot results.

Logistic Regression (Retired)



In the logistic regression model, the influence of *X* on *Y* is measured by the value of the slope of *X* which we have called β_1. The hypothesis that $\beta_1 = 0$ versus the alternative that $\beta_1 = B \neq 0$ is of interest since if $\beta_1 = 0$, *X* is not related to *Y*.

Under the alternative hypothesis that $\beta_1 = B$, the logistic model becomes

$$\log\left(\frac{P_1}{1 - P_1}\right) = \beta_0 + BX$$

Under the null hypothesis, this reduces to

$$\log\left(\frac{P_0}{1 - P_0}\right) = \beta_0$$

To test whether the slope is zero at a given value of *X*, the difference between these to quantities is formed giving

$$\beta_0 + BX - \beta_0 = \log\left(\frac{P_1}{1 - P_1}\right) - \log\left(\frac{P_0}{1 - P_0}\right)$$

which reduces to

$$\begin{aligned} BX &= \log\left(\frac{P_1}{1 - P_1}\right) - \log\left(\frac{P_0}{1 - P_0}\right) \\ &= \log\left(\frac{P_1 / (1 - P_1)}{P_0 / (1 - P_0)}\right) \\ &= \log(OR) \end{aligned}$$

where *OR* is odds ratio of P_1 and P_0. This relationship may be solved for *OR* giving

$$OR = e^{BX}$$

This shows that the odds ratio of P_1 and P_0 is directly related to the slope of the logistic regression equation. It also shows that the value of the odds ratio depends on the value of *X*. For a given value of *X*, testing that *B* is zero is equivalent to testing *OR* is one. Since *OR* is commonly used and well understood, it is used as a measure of effect size in power analysis and sample size calculations.

Power Calculations

Suppose you want to test the null hypothesis that $\beta_1 = 0$ versus the alternative that $\beta_1 = B$. Hsieh, Block, and Larsen (1998) have presented formulae relating sample size, α, power, and *B* for two situations: when X_1 is normally distributed and when X_1 is binomially distributed.

When X_1 is normally distributed, the sample size formula is

$$N = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{P*\left(1-P*\right)B^2}$$

where $P*$ is the event rate (probability that *Y* = 1) at the mean of X_1. Note that *B* is defined in terms of an increase of one standard deviation of X_1 above the mean.

When X_1 is binomially distributed and X_1 = 0 or 1, the sample size formula is

$$N = \frac{\left(z_{1-\alpha/2}\sqrt{\frac{\overline{P}\left(1-\overline{P}\right)}{R}} + z_{1-\beta}\sqrt{P_0\left(1-P_0\right) + \frac{P_1\left(1-P_1\right)\left(1-R\right)}{R}} \right)^2}{\left(P_0 - P_1\right)^2\left(1-R\right)}$$

where P_0 is the event rate at $X_1 = 0$ and P_1 is the event rate at $X_1 = 1$, *R* is the proportion of the sample with $X_1 = 1$, and $\overline{P}$ is the overall event rate given by

$$\overline{P} = \left(1-R\right)P_0 + R\left(P_1\right).$$

Multiple Logistic Regression

The multiple logistic regression model relates the probability distribution of *Y* to two or more covariates $X_1, X_2, \cdots, X_k$ by the formula

$$\log\left(\frac{P}{1\text{-}P}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

where *P* is the probability that *Y* = 1 given the values of the covariates. It is a simple extension of the simple logistic regression model that was just presented. In power analysis and sample size work, attention is placed on a single covariate while the influence of the other covariates is statistically removed by placing them at their mean values.

When there are multiple covariates, the following adjustment was given by Hsieh (1998) to give the total sample size, N_m

$$N_m = \frac{N}{1\text{-}\rho^2}$$

where ρ is the multiple correlation coefficient between X_1 (the variable of interest) and the remaining covariates. Notice that the number of extra covariates does not matter in this approximation.

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. The parameters that may be selected are *P1*, *Sample Size*, *Alpha*, and *Power*. 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.

Commonly, 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. A type-II error occurs when you fail to reject the null hypothesis of equal probabilities of the event of interest when in fact they are different.

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.

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 of equal probabilities when in fact they are equal.

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 option specifies the total number of observations in the sample. You may enter a single value or a list of values.

Effect Size – Baseline Probability

P0 (Baseline Probability that Y=1)

This option specifies one or more P_0 values. The interpretation of P_0 depends on whether X_1 is binary or continuous.

Binomial Covariate

When X_1 is binary, P_0 is the probability that $Y = 1$ when $X_1 = 0$. All other covariates are assumed to be equal to their mean values. In this case, the logistic equation reduces to

$$\log\left(\frac{P_0}{1 - P_0}\right) = \beta_0$$

so that

$$P_0 = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

Normal Covariate

When X_1 is normally distributed, P_0 is the probability that $Y = 1$ when $X_1 = \mu_{X_1}$, where μ_{X_1} is the mean of X_1. That is, P_0 is the baseline probability that $Y = 1$ when X_1 is ignored. All other covariates are assumed to be equal to their mean values. In this case, the logistic equation reduces to

$$\log\left(\frac{P_0}{1 - P_0}\right) = \beta_0 + \beta_1\mu_{X_1}$$

so that

$$P_0 = \frac{e^{\beta_0 + \beta_1\mu_{X_1}}}{1 + e^{\beta_0 + \beta_1\mu_{X_1}}}$$

Effect Size – Alternative Probability

Use P1 or Odds Ratio

This option specifies the whether to specify *P1* directly or to specify it by specifying the odds ratio. Since the relationship between the odds ration, *P1*, and *P0* is given by

$$OR = \frac{P_1 / (1 - P_1)}{P_0 / (1 - P_0)}$$

specifying *OR* and *P0* implicitly specifies *P1*.

This options lets you specify whether you want to state the alternative hypothesis in terms of *P1* or the odds ratio.

P1 (Alternative Probability that Y=1)

This option specifies the effect size to be detected by specifying P_1. As was shown earlier, the slope of the logistic regression can be expressed in terms of P_0 and P_1. Hence, by specifying P_1, you are also specifying the slope.

This option is only used when the User P1 or Odds Ratio option is set to *P1*. Its interpretation depends on whether X_1 is binomial or normal.

Binomial Covariate

When X_1 is binary, *P1* is the probability that $Y = 1$ when $X_1 = 1$. All other covariates are assumed to be equal to their mean values. In this case, the logistic equation reduces to

$$\log\left(\frac{P_1}{1-P_1}\right) = \beta_0 + \beta_1$$

since $X_1 = 1$.

Normal Covariate

When X_1 is normally distributed, *P1* is the probability that $Y = 1$ when $X_1 = \mu_{x_1} + \sigma_{x_1}$. That is, when x_1 is one standard deviation above the mean. All other covariates are assumed to be equal to their mean values. In this case, the logistic equation reduces to

$$\log\left(\frac{P_1}{1-P_1}\right) = \beta_0 + \beta_1 x_1$$

so that

$$P_1 = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$$

Odds Ratio (Odds1/Odds0)

This option specifies the odds ratio to be detected by the study. As was shown earlier, the slope of the logistic regression can be expressed in terms of P_0 and the odds ratio. Hence, by specifying *OR*, you are also specifying the slope. Using the formula

$$P_1 = \frac{OR(P_0)}{1 - P_0 + OR(P_0)}$$

specifying *OR* and P_0 implicitly specifies P_1.

This option is only used when the User P1 or Odds Ratio option is set to *Odds Ratio*. Its interpretation depends on whether X_1 is binomial or normal.

Binomial Covariate

When X_1 is binary, this option gives the odds ratio of P_1 and P_0. All other covariates are assumed to be equal to their mean values. In this case, the logistic equation reduces to

$$\log\left(\frac{P_1}{1-P_1}\right) = \beta_0 + \beta_1$$

since $X_1 = 1$.

This odds ratio compares the odds of obtaining $Y = 1$ when $X_1 = 1$ to the odds of obtaining $Y = 1$ when $X_1 = 0$.

Normal Covariate

When X_1 is normally distributed, this option gives the odds ratio of P_1 and P_0, where P_1 is the probability that $Y = 1$ when $X_1 = x_1$, where x_1 is a value other than μ_{x_1}. All other covariates are assumed to be equal to their mean values. In this case, the logistic equation reduces to

$$\log\left(\frac{P_1}{1-P_1}\right) = \beta_0 + \beta_1 x_1$$

so that

$$P_1 = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

This odds ratio compares the odds of obtaining $Y = 1$ when $X_1 = x_1$ to the odds of obtaining Y when $X_1 = \mu_{x_1}$.

Effect Size – Covariates (X1 is the Variable of Interest)

R-Squared of X1 with Other X's

This is the R-Squared that is obtained when X_1 is regressed on the other X's (covariates) in the model. Use this to study the influence on power and sample size of adding other covariates. Note that the number of additional variables does not matter in this formulation. Only their overall relationship with X_1 through this R-Squared value is used.

Of course, this value is restricted to being greater than or equal to zero and less than one. Use zero when there are no other covariates.

X1 (Independent Variable of Interest)

This option specifies whether the covariate is binary (binomial) or continuous (normal). This is a very important distinction since the sample size required for a particular power level is much larger for a binary covariate than for a continuous covariate.

This selection also changes the meaning of *P0* and *P1*.

Percent of N with X1 = 1

When X_1 is binary, this option specifies the proportion, *R*, of the sample in which $X_1 = 1$. Note that the value is specified as a percentage.

Example 1 – Power for a Continuous Covariate

A study is to be undertaken to study the relationship between post-traumatic stress disorder and heart rate after viewing video tapes containing violent sequences. Heart rate is assumed to be normally distributed. The event rate is thought to be 7% among soldiers. The researchers want a sample size large enough to detect an odds ratios of 1.5 or 2.0 with 90% power at the 0.05 significance level with a two-sided test. They decide to calculate the power at level sample sizes between 20 and 1200.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Logistic Regression (Retired)** procedure. 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	**Two-Sided**
Alpha	**0.05**
N (Sample Size)	**20 50 100 200 300 500 700 1000 1200**
P0 (Baseline Probability that Y=1)	**0.07**
Use P1 or Odds Ratio	**Odds Ratio**
Odds Ratio (Odds1/Odds0)	**1.5 2.0**
R-Squared of X1 with Other X's	**0**
X1 (Independent Variable of Interest)	**Continuous (Normal)**

Annotated Output

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

Numeric Results

Power	N	P0	P1	Odds Ratio	R Squared	Alpha	Beta
0.06716	20	0.070	0.101	1.500	0.000	0.05000	0.93284
0.10964	50	0.070	0.101	1.500	0.000	0.05000	0.89036
0.17737	100	0.070	0.101	1.500	0.000	0.05000	0.82263
0.30962	200	0.070	0.101	1.500	0.000	0.05000	0.69038
0.43325	300	0.070	0.101	1.500	0.000	0.05000	0.56675
0.63808	500	0.070	0.101	1.500	0.000	0.05000	0.36192
0.78147	700	0.070	0.101	1.500	0.000	0.05000	0.21853
0.90516	1000	0.070	0.101	1.500	0.000	0.05000	0.09484
0.94779	1200	0.070	0.101	1.500	0.000	0.05000	0.05221
0.12119	20	0.070	0.131	2.000	0.000	0.05000	0.87881
0.23903	50	0.070	0.131	2.000	0.000	0.05000	0.76097
0.42410	100	0.070	0.131	2.000	0.000	0.05000	0.57590
0.70579	200	0.070	0.131	2.000	0.000	0.05000	0.29421
0.86504	300	0.070	0.131	2.000	0.000	0.05000	0.13496
0.97696	500	0.070	0.131	2.000	0.000	0.05000	0.02304
0.99673	700	0.070	0.131	2.000	0.000	0.05000	0.00327
0.99986	1000	0.070	0.131	2.000	0.000	0.05000	0.00014
0.99998	1200	0.070	0.131	2.000	0.000	0.05000	0.00002

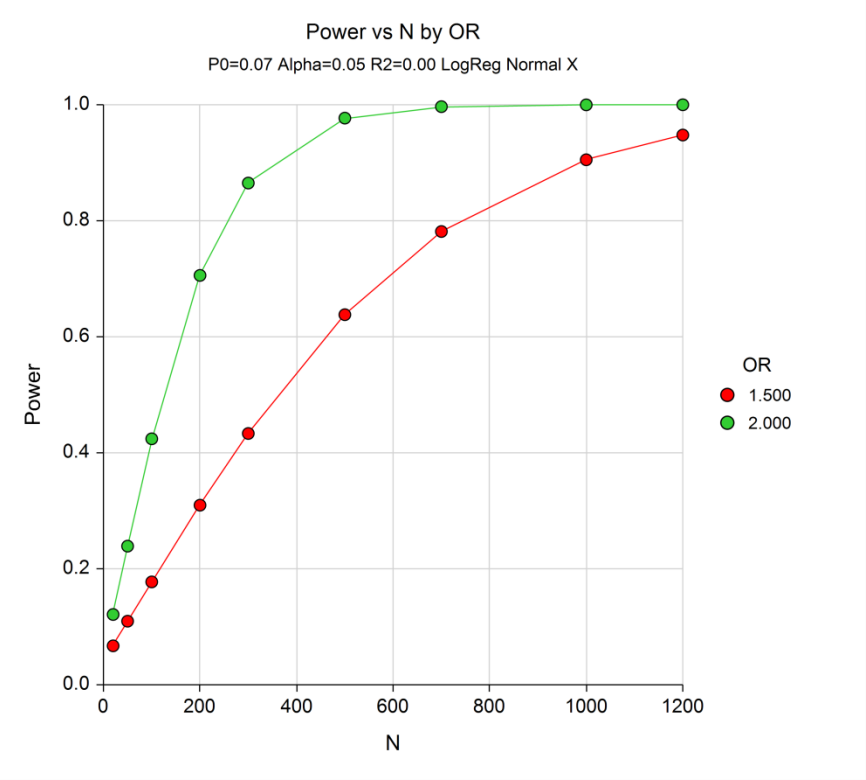
Logistic Regression (Retired)

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.
P0 is the response probability at the mean of the covariate, X.
P1 is the response probability when X is increased to one standard deviation above the mean.
Odds Ratio is the odds ratio when P1 is on top. That is, it is [P1/(1-P1)]/[P0/(1-P0)].
R-Squared is the R2 achieved when X is regressed on the other independent variables in the regression.
Alpha is the probability of rejecting a true null hypothesis.
Beta is the probability of accepting a false null hypothesis.

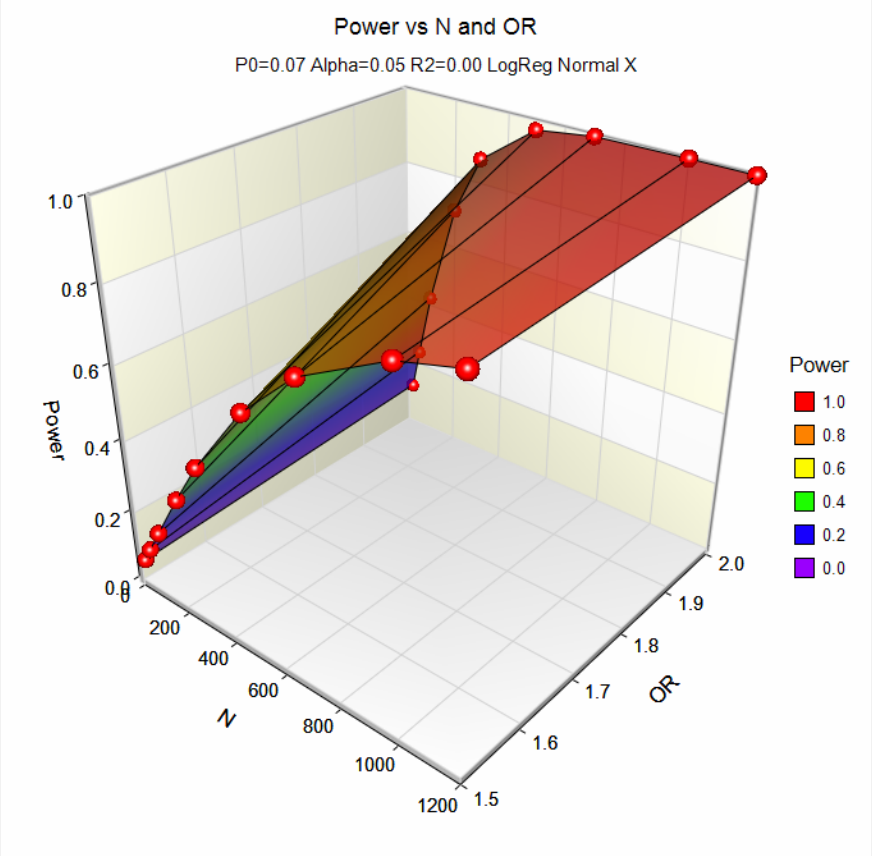
Summary Statements
A logistic regression of a binary response variable (Y) on a continuous, normally distributed, independent variable (X) with a sample size of 20 observations achieves 7% power at a 0.050 significance level to detect a change in Prob(Y=1) from the value of 0.070 at the mean of X to 0.101 when X is increased to one standard deviation above the mean.This change corresponds to an odds ratio of 1.500.

This report shows the power for each of the scenarios. The report shows that a power of 90% is reached at a sample size of about 300 for an odds ratio of 2.0 and 1000 for an odds ratio of 1.5.

Plot Section



Logistic Regression (Retired)



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

Example 2 – Sample Size for a Continuous Covariate

Continuing with the previous study, determine the exact sample size necessary to attain a power of 90%.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Logistic Regression (Retired)** procedure. 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	**Sample Size**
Alternative Hypothesis	**Two-Sided**
Power	**0.90**
Alpha	**0.05**
P0 (Baseline Probability that Y=1)	**0.07**
Use P1 or Odds Ratio	**Odds Ratio**
Odds Ratio (Odds1/Odds0)	**1.5 2.0**
R-Squared of X1 with Other X's	**0**
X1 (Independent Variable of Interest)	**Continuous (Normal)**

Output

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

Numeric Results

Power	N	P0	P1	Odds Ratio	R Squared	Alpha	Beta
0.89978	981	0.070	0.101	1.500	0.000	0.05000	0.10022
0.89920	335	0.070	0.131	2.000	0.000	0.05000	0.10080

This report shows the power for each of the scenarios. The report shows that a power of 90% is achieved at a sample size of 335 for an odds ratio of 2.0 and 981 for an odds ratio of 1.5.

Example 3 – Effect Size for a Continuous Covariate

Continuing the previous study, suppose the researchers can only afford a sample size of 500 individuals. They want to determine if a meaningful odds ratio can be detected with this sample size.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Logistic Regression (Retired)** procedure. 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	**P1 > P0 or Odds Ratio > 1**
Alternative Hypothesis	**Two-Sided**
Power	**0.90**
Alpha	**0.05**
N (Sample Size)	**500**
P0 (Baseline Probability that Y=1)	**0.07**
R-Squared of X1 with Other X's	**0**
X1 (Independent Variable of Interest)	**Continuous (Normal)**

Output

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

Numeric Results

Power	N	P0	P1	Odds Ratio	R Squared	Alpha	Beta
0.90000	500	0.070	0.117	1.765	0.000	0.05000	0.10000

This report shows that this experimental design can detect an odds ratio of 1.765. That is, it can detect a shift in the event rate from 0.070 to 0.117.

Example 4 – Sample Size for a Binary Covariate

A study is to be undertaken to study the relationship between post-traumatic stress disorder and gender. The event rate is thought to be 7% among males. The researchers want a sample size large enough to detect an odds ratio of 1.5 with 90% power at the 0.05 significance level with a two-sided test.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Logistic Regression (Retired)** procedure. 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**
Power	**0.90**
Alpha	**0.05**
P0 (Baseline Probability that Y=1)	**0.07**
Use P1 or Odds Ratio	**Odds Ratio**
Odds Ratio (Odds1/Odds0)	**1.5**
R-Squared of X1 with Other X's	**0**
X1 (Independent Variable of Interest)	**Binary (X=0 or 1)**
Percent of N with X1=1	**50**

Output

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

Numeric Results

Power	N	Pcnt N X=1	P0	P1	Odds Ratio	R Squared	Alpha	Beta
0.89997	3326	50.000	0.070	0.101	1.500	0.000	0.05000	0.10003

The sample size is estimated at 3326. This should be evenly divided among males and females.

Example 5 – Validation for a Continuous Covariate

Hsieh (1998) page 1628 gives the power as 95% when *N* = 317, alpha = 0.05 (two-sided), *P0* = 0.5, and the odds ratio is 1.5. The covariate is assumed to be continuous.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Logistic Regression (Retired)** procedure. You may then make the appropriate entries as listed below, or open **Example 5** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Alternative Hypothesis	**Two-Sided**
Alpha	**0.05**
N (Sample Size)	**317**
P0 (Baseline Probability that Y=1)	**0.50**
Use P1 or Odds Ratio	**Odds Ratio**
Odds Ratio (Odds1/Odds0)	**1.5**
R-Squared of X1 with Other X's	**0**
X1 (Independent Variable of Interest)	**Continuous (Normal)**

Output

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

Numeric Results

Power	N	P0	P1	Odds Ratio	R Squared	Alpha	Beta
0.95049	317	0.500	0.600	1.500	0.000	0.05000	0.04951

PASS calculates a power of 0.95049 which matches Hsieh.

Example 6 – Validation for a Binary Covariate

Hsieh (1998) page 1626 gives the power as 95% when *N* = 1282 (equal sample sizes for both groups), alpha = 0.05 (two-sided), *P0* = 0.4, and the *P1* = 0.5. The covariate is assumed to be binary.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Logistic Regression (Retired)** procedure. You may then make the appropriate entries as listed below, or open **Example 6** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Alternative Hypothesis	**Two-Sided**
Alpha	**0.05**
N (Sample Size)	**1282**
P0 (Baseline Probability that Y=1)	**0.4**
Use P1 or Odds Ratio	**P1**
P1 (Alternative Probability that Y=1)	**0.5**
R-Squared of X1 with Other X's	**0**
X1 (Independent Variable of Interest)	**Binary (X = 0 or 1)**
Percent of N with X1=1	**50**

Output

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

Numeric Results

Power	N	Pcnt N X=1	P0	P1	Odds Ratio	R Squared	Alpha	Beta
0.95021	1282	50.000	0.400	0.500	1.500	0.000	0.05000	0.04979

PASS calculates a power of 0.95021 which matches Hsieh.