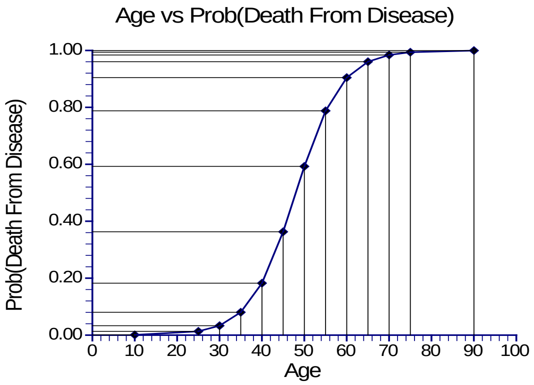
Chapter 872

Tests for the Odds Ratio in Logistic Regression with One Binary X and Other Xs (Wald Test)

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. This procedure deals with the specific case in which the covariate of interest is binary.

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.

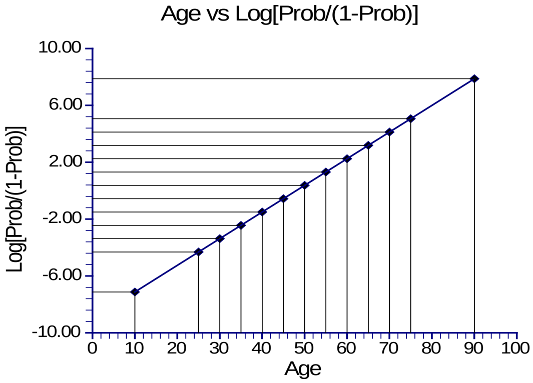
Under the logistic model, the proportion dying, *P*, at a given age can be calculated using the formula

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

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.



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

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

Under the null hypothesis, this reduces to

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

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

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

which reduces to

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

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 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

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

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

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

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.

Ryan (2013) had some reservations with this approach. We refer you to page 163 of his sample size book for more details.

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

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

Effect Size – Alternative Probability

Use P1 or Odds Ratio

This option specifies whether to specify *P1* directly or to specify it by specifying the odds ratio. Since the relationship between the odds ratio, *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 $X_1 = 1$.

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*. 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 $X_1 = 0$.

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.

Percent of N with X1 = 1

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 – Finding Power 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. They will eventually have five X's in their study. The R-squared of the remaining four variables is estimated to be 0.20.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for the Odds Ratio in Logistic Regression with One Binary X and Other Xs (Wald Test)** 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**
R-Squared of X1 with Other X's	**0.2**
Percent of N with X1=1	**50**

Annotated Output

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

Numeric Results

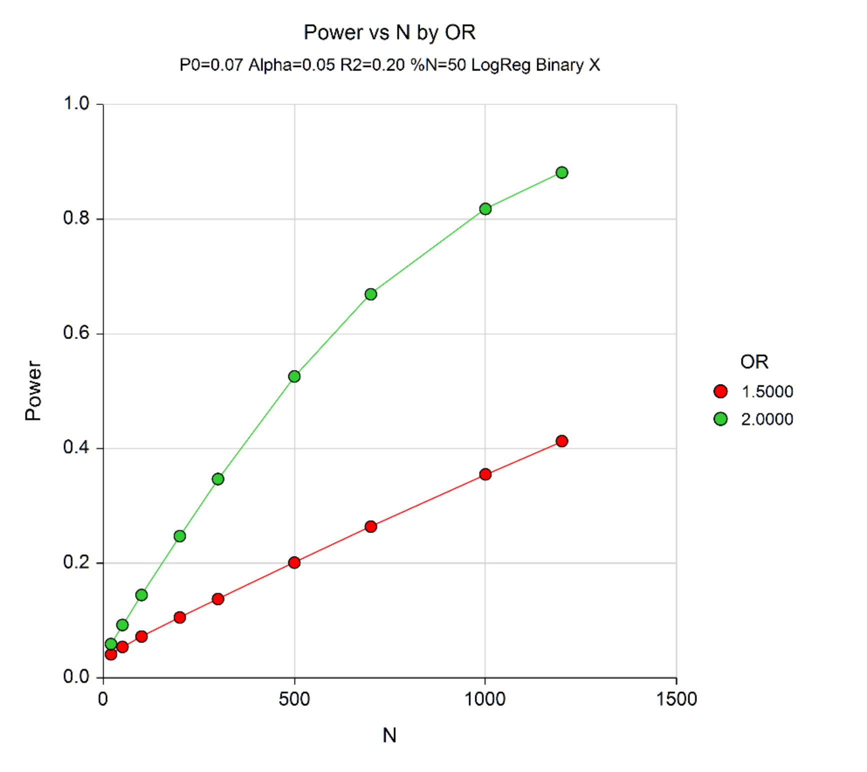
```
                   Pcnt N                         Odds          R
Power        N       X=1       P0        P1      Ratio    Squared      Alpha      Beta
0.0411      20        50   0.0700    0.1014     1.5000     0.2000       0.05    0.9589
0.0540      50        50   0.0700    0.1014     1.5000     0.2000       0.05    0.9460
0.0722     100        50   0.0700    0.1014     1.5000     0.2000       0.05    0.9278
0.1054     200        50   0.0700    0.1014     1.5000     0.2000       0.05    0.8946
0.1375     300        50   0.0700    0.1014     1.5000     0.2000       0.05    0.8625
0.2010     500        50   0.0700    0.1014     1.5000     0.2000       0.05    0.7990
0.2638     700        50   0.0700    0.1014     1.5000     0.2000       0.05    0.7362
0.3550    1000        50   0.0700    0.1014     1.5000     0.2000       0.05    0.6450
0.4129    1200        50   0.0700    0.1014     1.5000     0.2000       0.05    0.5871
0.0590      20        50   0.0700    0.1308     2.0000     0.2000       0.05    0.9410
0.0923      50        50   0.0700    0.1308     2.0000     0.2000       0.05    0.9077
0.1445     100        50   0.0700    0.1308     2.0000     0.2000       0.05    0.8555
0.2472     200        50   0.0700    0.1308     2.0000     0.2000       0.05    0.7528
0.3468     300        50   0.0700    0.1308     2.0000     0.2000       0.05    0.6532
0.5258     500        50   0.0700    0.1308     2.0000     0.2000       0.05    0.4742
0.6691     700        50   0.0700    0.1308     2.0000     0.2000       0.05    0.3309
0.8179    1000        50   0.0700    0.1308     2.0000     0.2000       0.05    0.1821
0.8814    1200        50   0.0700    0.1308     2.0000     0.2000       0.05    0.1186
```

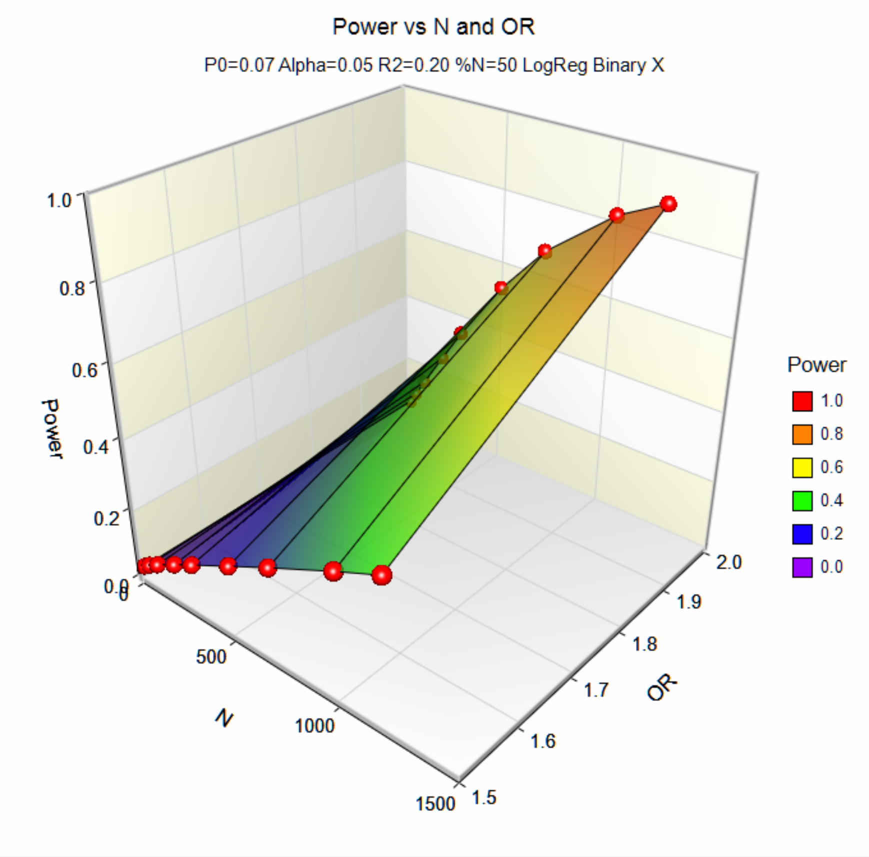
Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population.
P0 is the response probability at the mean of the covariate, X.
P1 is the response probability when X is increased to one standard deviation above the mean.
Pcnt N X=1 is the percentage of the population in which X = 1.
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 binary independent variable (X) with a sample size of 20 observations (of which 50% are in the group X=0 and 50% are in the group X=1) achieves 4% power at a 0.05 significance level to detect a change in Prob(Y=1) from the baseline value of 0.0700 to 0.1014. This change corresponds to an odds ratio of 1.5000. An adjustment was made since a multiple regression of the independent variable of interest on the other independent variables in the logistic regression obtained an R-Squared of 0.2000.

This report shows the power for each of the scenarios.

Plot Section





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

Example 2 – Finding Sample Size

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 **Tests for the Odds Ratio in Logistic Regression with One Binary X and Other Xs (Wald Test)** 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**
R-Squared of X1 with Other X's	**0.2**
Percent of N with X1=1	**50**

Output

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

Numeric Results

```
                     Pcnt N                           Odds          R
Power          N        X=1       P0        P1       Ratio    Squared      Alpha     Beta
0.9000      4158         50   0.0700    0.1014      1.5000     0.2000       0.05   0.1000
0.8996      1276         50   0.0700    0.1308      2.0000     0.2000       0.05   0.1004
```

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 1276 for an odds ratio of 2.0 and 4158 for an odds ratio of 1.5.

Example 3 – 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 prevalence of X1 is assumed to be 0.50.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for the Odds Ratio in Logistic Regression with One Binary X and Other Xs (Wald Test)** 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	**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**
Percent of N with X1=1	**50**

Output

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

Numeric Results

```
                   Pcnt N                         Odds          R
Power        N       X=1      P0        P1      Ratio    Squared     Alpha      Beta
0.9502    1282        50  0.4000    0.5000     1.5000     0.0000      0.05    0.0498
```

PASS calculates a power of 0.9502 which matches Hsieh (1998).