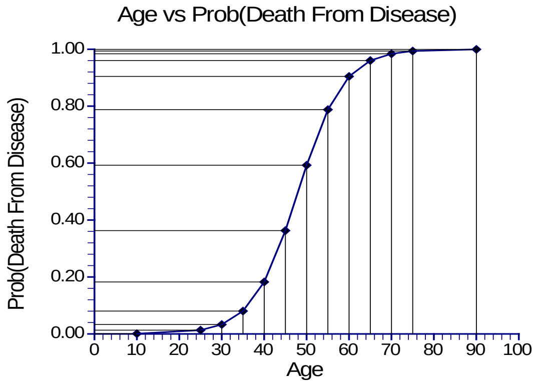
## Chapter 859

# Tests for the Odds Ratio in Logistic Regression with One Normal X (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, but in this procedure, the covariate is assumed to be normally distributed.

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 could be used to summarize the relationship between ages 40 and 60, it certainly could not be used for the young or the elderly.

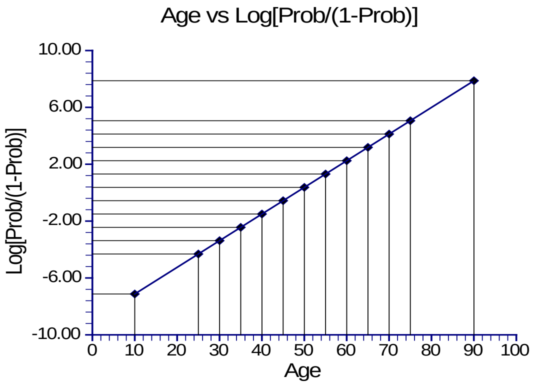
Under the logistic model, the proportion dying, *P*, at a given age can be calculated using the formula

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

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 $\beta_1$. 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 that *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.

This procedure makes the assumption that *X* is normally distributed. Without loss of generality, we assume that the mean of *X* is zero and the variance of *X* is one. We define $X_0$ to be the mean of *X* and $X_1$ to be the mean plus one standard deviation of *X*.

# Power Calculations

We use the results of Novikov, Fund, and Freedman (2010) to compute sample size and power. This is a modification of the method of Hsieh, Block, and Larsen (1998) which is based on the Wald test. Note that their method is recommended in a simulation study by Bush (2015).

Suppose you want to test the null hypothesis that the odds ratio is one versus the alternative that it is some other positive value. Novikov *et al* (2010) presented formulae relating sample size, $\alpha$, power, and odds ratio for the situation in which *X* is normally distributed and it is the only variable in the logistic regression model.

The sample size formula is

$$N = \left(\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 \frac{(\tau+\gamma)v_1}{\gamma(m_1-m_0)^2} + \frac{(\tau^2+\gamma^3)z_{1-\alpha/2}^2}{2\gamma(\tau+\gamma)^2}\right)/(1+\gamma)$$

where $\tau$ is the ratio of the variance of *X* ($v_0$) in the subgroup in which $Y = 0$ to the variance of *X* ($v_1$) in the subgroup of the population in which $Y = 1$, $\gamma$ is the ratio of the size of the subgroup $Y = 0$ to the size of subgroup $Y = 1$, and $m_0$ and $m_1$ are the conditional means in the two subgroups defined by the values of Y.

Novikov *et al* (2010) implement this formula using an algorithm which they define that uses only the prevalence of *Y* (probability that $Y = 1$ in the population). They include SAS code for implementing their formula.

## Errors in the SAS code of Novikov

As we implemented the Novikov *et al* formula, we found that their unnecessary use of the SAS *ceil(x)* command caused severe rounding errors in their Table IV. For example, their value for N on the first line of the table was 6552. We found that be removing the *ceil(x)* commands, our computed sample size of 6508 still gave power over 80%.

## Findings of Bush (2015)

Bush (2015) conducted a simulation study of sample size estimation in logistic regression. He compared sample size formulas based on the Wald test, the likelihood ratio test, and the score test. He found that the "*power values are very similar, rarely deviating more than 2% between the three tests*." He found that the Novikov *et al* method did quite well. Hence, in this situation, it does not appear to matter upon which test the power is computed.

# 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 *Sample Size* and *Power*.

## 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. 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 becomes visible when *Solve For* is set to *Power*. This option specifies the total number of observations in the sample. You may enter a single value or a list of values.

## Effect Size

### P1 (Prevalence of Y)

Enter values for P1, the prevalence of Y. That is, P1 is the overall prob(Y = 1), regardless of the value of X. It assumes that Y = 1 represents an ‘event’ and Y = 0 represents a ‘non-event’.

Note that this is the unconditional probability that Y equals 1. It is NOT conditional on the value of X..

### Odds Ratio (Odds[x+σx]/Odds[x])

Enter values for the Odds Ratio, OR. This is the ratio of the event odds at one standard deviation of X above the mean to the event odds when X is equal to its mean.

The value entered is the assumed to be value under the alternative hypothesis, Ha. It represents that value to be detected.

Mathematically, OR is defined as

$$OR = odds(\mu x + \sigma x)/odds(\mu x)$$

$$= [P1(\mu x + \sigma x)/(1\text{-}P1(\mu x + \sigma x))] \;/\; [P1(\mu x)/(1\text{-}P1(\mu x))].$$

OR must be positive and not equal to 1, since OR = 1 under H0. Typical values range from 0.25 to 4.

# 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 ratio 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 **Tests for the Odds Ratio in Logistic Regression with One Normal X (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) | **100 to 1300 by 100** |
| P1 (Prevalence of Y) | **0.07** |
| Odds Ratio (Odds[x+σx] / Odds[x]) | **1.5 2** |

## Annotated Output

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

### Numeric Results

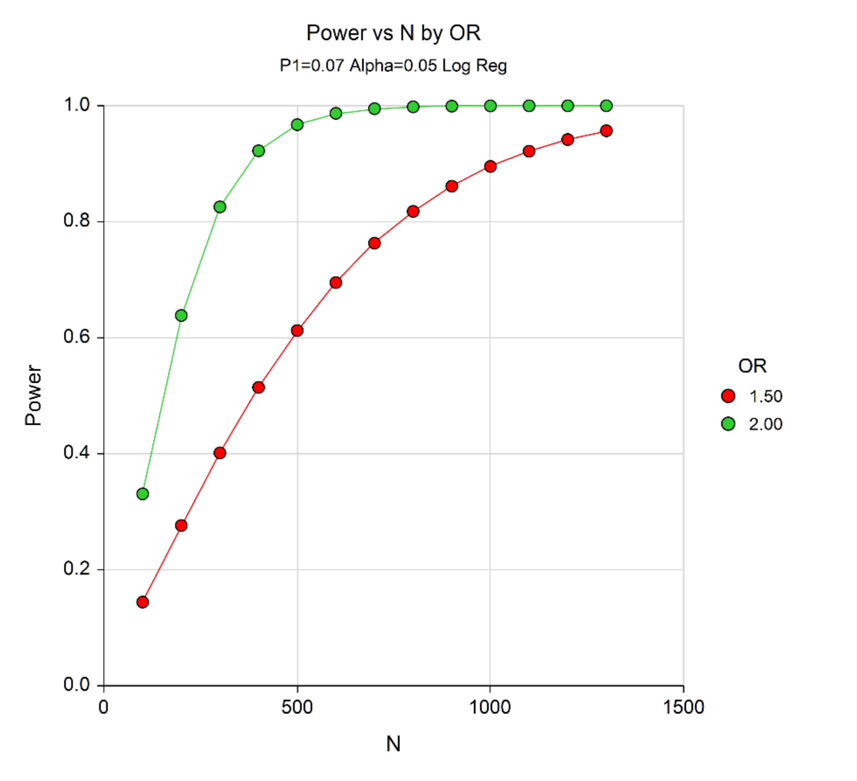
| Power | N | Events | P1 | P1(μx) | Odds Ratio | Alpha | b0 | b1 |
|---|---|---|---|---|---|---|---|---|
| 0.1443 | 100 | 7 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.2764 | 200 | 14 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.4015 | 300 | 21 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.5145 | 400 | 28 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.6125 | 500 | 35 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.6951 | 600 | 42 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.7631 | 700 | 49 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.8179 | 800 | 56 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.8613 | 900 | 63 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.8954 | 1000 | 70 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.9216 | 1100 | 77 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.9417 | 1200 | 84 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.9570 | 1300 | 91 | 0.0700 | 0.0656 | 1.5000 | 0.0500 | -2.6566 | 0.4055 |
| 0.3309 | 100 | 7 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.6384 | 200 | 14 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.8257 | 300 | 21 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.9224 | 400 | 28 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.9674 | 500 | 35 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.9869 | 600 | 42 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.9950 | 700 | 49 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.9981 | 800 | 56 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.9993 | 900 | 63 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.9998 | 1000 | 70 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 0.9999 | 1100 | 77 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 1.0000 | 1200 | 84 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |
| 1.0000 | 1300 | 91 | 0.0700 | 0.0580 | 2.0000 | 0.0500 | -2.7867 | 0.6931 |

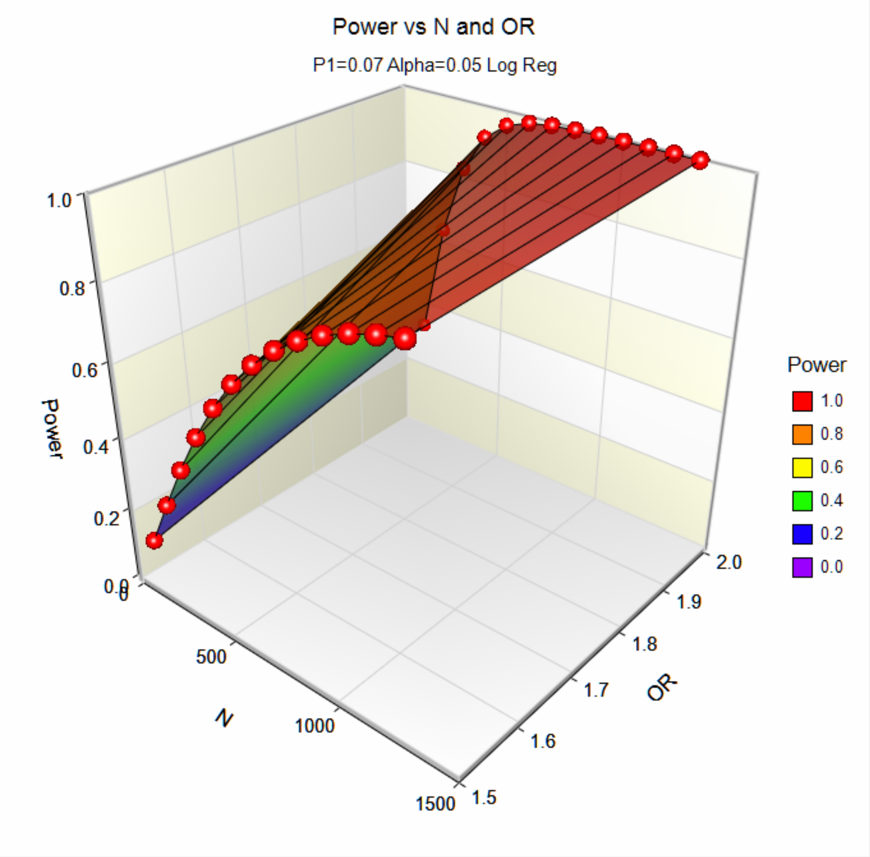
**Report Definitions**
Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population.
Events is the expected number of cases in which Y = 1.
P1 is the overall proportion of the population in which Y = 1.
P1(μx) is the proportion of the population in which Y = 1 if X = μx (mean of X).
Odds Ratio is the ratio: odds(μx + σx)/odds(μx) = [P1(μx + σx)/(1-P1(μx + σx))] / [P1(μx)/(1-P1(μx))].
Alpha is the probability of rejecting a true null hypothesis.
b0 is the intercept in the logit model log(P/(1-P)) = b0 + b1 X.
b1 is slope in the logit model log(P/(1-P)) = b0 + b1 X.

**Summary Statements**
A logistic regression of a binary response variable (Y) on a continuous, normally distributed variable (X) with a sample size of 100 observations achieves 14% power at a 0.05000 significance level to detect an odds ratio of 1.500 when the prevalence of Y in the population is 0.07000. This odds ratio is the ratio of the odds that Y = 1 when X is one standard deviation above its mean to the odds that Y = 1 when X is equal to its mean.

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 400 for an odds ratio of 2.0 and 1000 for an odds ratio of 1.5.

## Plot Section





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

# Example 2 – Sample Size for a Normal 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 **Tests for the Odds Ratio in Logistic Regression with One Normal X (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** |
| P1 (Prevalence of Y) | **0.07** |
| Odds Ratio (Odds[x+σx] / Odds[x]) | **1.5 2** |

## Output

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

### Numeric Results

```
Power        N   Events       P1     P1(μx) Odds Ratio      Alpha        b0        b1
0.9000    1016       71   0.0700     0.0656     1.5000     0.0500   -2.6566    0.4055
0.9003     370       26   0.0700     0.0580     2.0000     0.0500   -2.7867    0.6931
```

This report shows the necessary sample size for each odds ratio.

# Example 3 – Validation for a Normal Covariate

Novikov (2010) page 102, Table IV, first line, gives sample size as 6551 when alpha = 0.05 (two-sided), power = 0.8, *P1* = 0.02, and the odds ratio is log(0.25) = 1.28402542. As we discussed above, the sample size of 6551 is high because of an error in their SAS code. The sample size should be 6508, which we found using manual calculation.

## 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 Normal X (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 | **Sample Size** |
| Alternative Hypothesis | **Two-Sided** |
| Power | **0.8** |
| Alpha | **0.05** |
| P1 (Prevalence of Y) | **0.02** |
| Odds Ratio (Odds[x+σx] / Odds[x]) | **1.28402542** |

## Output

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

### Numeric Results

```
Power       N   Events      P1    P1(μx) Odds Ratio    Alpha        b0       b1
0.8000   6508      130  0.0200    0.0194     1.2840   0.0500   -3.9218   0.2500
```

This report achieves an N of 6508 which validates this procedure.