Chapter 864

Confidence Intervals for the Odds Ratio in Logistic Regression with One Binary X

Introduction

Logistic regression expresses the relationship between a binary response variable and one or more independent variables called *covariates*. This procedure calculates sample size for the case when there is only one, binary covariate (X) in the logistic regression model and a Wald statistic is used to calculate a confidence interval for the odds ratio of Y to X. Often, Y is called the *response* variable and X is referred to as the *exposure* variable. For example, Y might refer to the presence or absence of cancer and X might indicate whether the subject smoked or not.

Sample Size Calculations

Using the *logistic model*, the probability of a binary event is

$$\Pr(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X)}$$

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

$$\log\left(\frac{\Pr(Y = 1|X)}{1 - \Pr(Y = 1|X)}\right) = \beta_0 + \beta_1 X$$

Note that the left side is the logarithm of the odds of a response event (Y = 1) versus a response non-event (Y = 0). This is sometimes called the *logit* transformation of the probability. In the logistic regression model, the magnitude of the association of X and Y is represented by the slope β_1. Since X is binary, only two cases need be considered: X = 0 and X = 1.

The logistic regression model lets us define two quantities

$$P_0 = \Pr(Y = 1|X = 0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

$$P_1 = \Pr(Y = 1|X = 1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$$

These values are combined in the odds ratio (OR) of P_1 to P_0 resulting in

$$OR_{yx} = \exp(\beta_1)$$

or, by taking the logarithm of both sides, simply

$$\log(OR_{yx}) = \log\left(\frac{\frac{P_1}{(1-P_1)}}{\frac{P_0}{(1-P_0)}}\right) = \beta_1$$

Hence the relationship between Y and X can be quantified as a single regression coefficient. It well known that the distribution of the maximum likelihood estimate of β_1 is asymptotically normal. A significance test or confidence interval for this slope is commonly formed from the Wald statistic

$$\mathrm{z} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

A (1 - α)% two-sided confidence interval for β_1 is

$$\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}}\, s_{\hat{\beta}_1}$$

By transforming this interval into the odds ratio scale by exponentiating both limits, a (1 - α)% two-sided confidence interval for OR is

$$(OR_{LL}, OR_{UL}) = \exp\left(\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}}\, s_{\hat{\beta}_1}\right)$$

Note that this interval is not symmetric about $\exp(\hat{\beta}_1)$.

Often, the goal during this part of the planning process is to find the sample size that reduces the width of the interval to a certain value $D = OR_{UL} - OR_{LL}$. A suitable D is found using a simple search of possible values of N.

Usually, the value of $s_{\hat{\beta}_1}$ is not known before the study so this quantity must be estimated. Demidenko (2007) gives a method for calculating an estimate of the variance from various quantities that can be set at the planning stage. Let p_x be the probability that X = 1 in the sample. The information matrix for this model is

$$I = \begin{bmatrix} \frac{p_x\exp(\beta_0+\beta_1)}{\left(1+\exp(\beta_0+\beta_1)\right)^2} + \frac{(1-p_x)\exp(\beta_0)}{\left(1+\exp(\beta_0)\right)^2} & \frac{p_x\exp(\beta_0+\beta_1)}{\left(1+\exp(\beta_0+\beta_1)\right)^2} \\ \frac{p_x\exp(\beta_0+\beta_1)}{\left(1+\exp(\beta_0+\beta_1)\right)^2} & \frac{p_x\exp(\beta_0+\beta_1)}{\left(1+\exp(\beta_0+\beta_1)\right)^2} \end{bmatrix}$$

The value of $\sqrt{N}s_{\hat{\beta}_1}$ is the (2,2) element of the inverse of I.

The values of β_0 and β_1 are calculated from OR_{yx} and P_0 using

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

$$\beta_1 = \log(OR_{yx}) = \log\left(\frac{\frac{P_1}{(1-P_1)}}{\frac{P_0}{(1-P_0)}}\right)$$

Thus, the confidence interval can be specified in terms of OR_{yx} and P_0. Of course, these results are only approximate. The final confidence interval depends on the actual data values.

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 *Precision (Confidence Interval Width)*, *Confidence Level*, or *Sample Size*.

One-Sided or Two-Sided Interval

Interval Type

Specify whether the confidence interval will be two-sided, one-sided with an upper limit, or one-sided with a lower limit.

Confidence

Confidence Level (1 – Alpha)

This option specifies one or more values of the proportion of confidence intervals (constructed with this same confidence level, sample size, etc.) that would have the same width.

The range of possible values is between 0 and 1. However, the range is usually between 0.5 and 1. Common choices are 0.9, 0.95, and 0.99. You should select a value that expresses the needs of this study.

You can enter a single value such as *0.7* or a series of values such as *0.7 0.8 0.9* or *0.7 to 0.95 by 0.05*.

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.

Precision

Distance from ORyx to Limit

In a one-sided confidence interval (sometimes called a confidence bound), this is the distance between the upper or lower confidence limit of ORyx and the value of ORyx. As the sample size increases, this value decreases and thus the interval becomes more precise.

Since an odds ratio is typically between 0.2 and 10, it is reasonable that the value of this distance is also between 0.2 and 10. By definition, only positive values are possible.

You can enter a single value such as *1* or a series of values such as *0.5 1 1.5* or *0.5 to 1.5 by 0.2*.

Width of ORyx Confidence Interval

In a two-sided confidence interval, this is the difference between the upper and lower confidence limits of ORyx. As the sample size increases, this width decreases and thus the interval becomes more precise.

Since an odds ratio is typically between 0.2 and 10, it is reasonable that the value of this width is also between 0.2 and 10. By definition, only positive values are possible.

You can enter a single value such as *1* or a series of values such as *0.5 1 1.5* or *0.5 to 1.5 by 0.2*.

Baseline Probability

P0 [Pr(Y = 1 | X = 0)]

This gives the value of the baseline probability of a response, P0, when the exposure is not present.

P0 is a probability, so it must be between zero and one. It cannot be equal to P1.

Odds Ratio (Confidence Interval Term)

ORyx (Y, X Odds Ratio)

Specify one or more values of the odds ratio of Y and X, a measure of the effect size (event rate) that is to be detected by the study. This is the ratio of the odds of the outcome Y given that the exposure X = 1 to the odds of Y = 1 given X = 0. That is, odds(Y=1|X=1) / odds(Y=1|X=0). Note that odds(A) = Pr(A)/Pr(Not A)

You can enter a single value such as *1.5* or a series of values such as *1.5 2 2.5* or *0.5 to 0.9 by 0.1*.

The range of this parameter is $0 < ORyx < \infty$ (typically, $0.1 < ORyx < 10$).

Prevalence

Percent with X = 1

This is the percentage of the sample in which X = 1. It is often called the prevalence of X.

You can enter a single value or a range of values. The permissible range is 1 to 99.

Example 1 – Find Sample size

A study is to be undertaken to study the association between the occurrence of a certain type of cancer (response variable) and the presence of a certain food in the diet. The baseline cancer event rate is 7%. The researchers want a sample size large enough to create a confidence interval with a width of 0.9. They assume that the actual odds ratio with be 2.0. The confidence level is set to 0.95. They also want to look at the sensitivity of the analysis to the specification of the odds ratio, so they also want to obtain the results for odds ratios of 1.75 and 2.25. The researchers assume that between 25% and 50% of the sample eat the food being studied, so they want results for both of these values.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for the Odds Ratio in Logistic Regression with One Binary X** 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	**Sample Size**
Interval Type	**Two-Sided**
Confidence Level	**0.95**
Width of ORyx Confidence Interval	**0.90**
P0 [Pr(Y=1|X=0)]	**0.07**
Odds Ratio (Odds1/Odds0)	**1.75 2.0 2.25**
Percent with X = 1	**25 50**

Annotated Output

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

Numeric Results

Numeric Results for Two-Sided Confidence Interval of ORyx

Confidence Level	N	C.I. Width	ORyx	Lower Conf Limit of ORyx	Upper Conf Limit of ORyx	P0	Percent X=1
0.950	3525	0.8999	1.750	1.357	2.257	0.070	25.0
0.950	2979	0.8999	1.750	1.357	2.257	0.070	50.0
0.950	4294	0.9000	2.000	1.600	2.500	0.070	25.0
0.950	3727	0.9000	2.000	1.600	2.500	0.070	50.0
0.950	5136	0.9000	2.250	1.845	2.745	0.070	25.0
0.950	4561	0.9000	2.250	1.845	2.745	0.070	50.0

Report Definitions
Logistic regression equation: Log(P/(1-P)) = β0 + β1×X, where P = Pr(Y = 1|X) and X is binary.
Confidence Level is the proportion of studies with the same settings that produce a confidence interval that includes the true ORyx.
N is the sample size.
C.I. Width is the distance between the two boundaries of the confidence interval.
ORyx is the expected sample value of the odds ratio. It is the value of exp(β1).
C.I. of ORyx Lower Limit is the lower limit of the confidence interval of ORyx.
C.I. of ORyx Upper Limit is the upper limit of the confidence interval of ORyx.
P0 is the response probability at X = 0. That is, P0 = Pr(Y = 1|X = 0).
Percent X=1 is the percent of the sample in which the exposure is 1 (present).

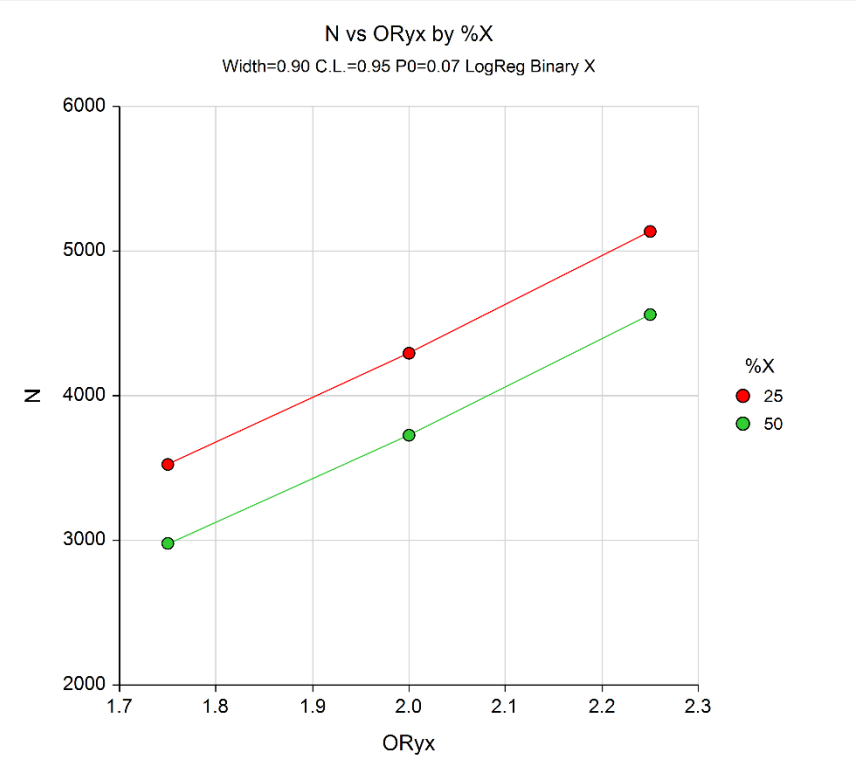
Confidence Intervals for the Odds Ratio in Logistic Regression with One Binary X

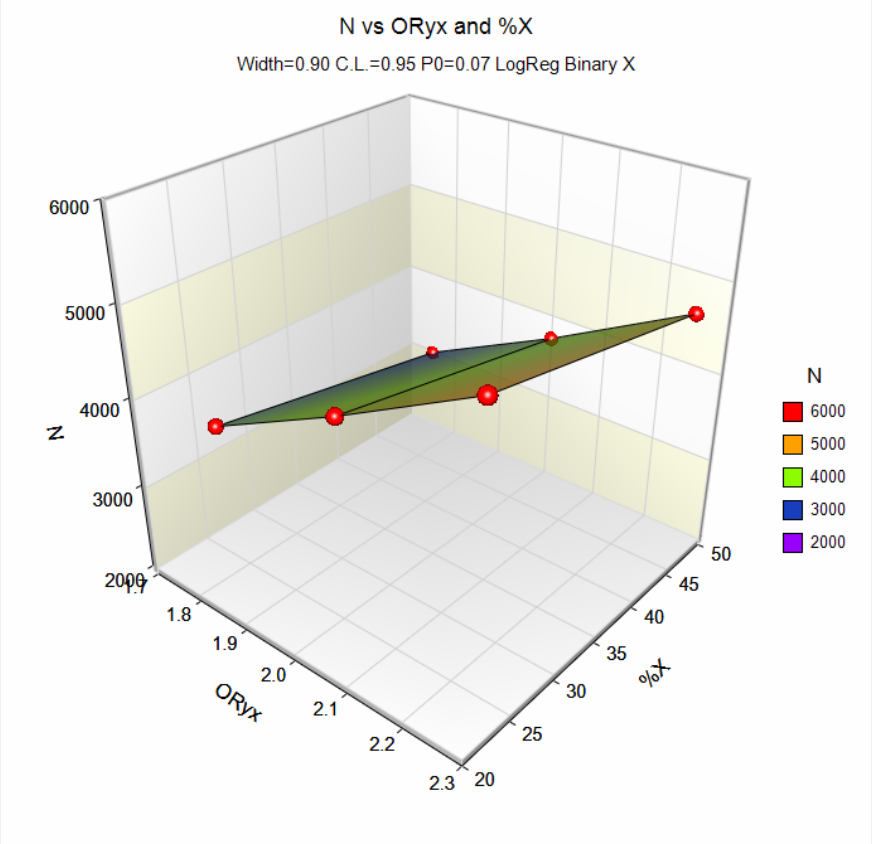
Summary Statements

A logistic regression of a binary response variable (Y) on a binary independent variable (X) with a sample size of 3525 observations (of which 25.0% are in the group X=1) at a 0.950 confidence level produces a two-sided confidence interval with a width of 0.8999. The baseline response rate is assumed to be 0.070 and the sample odds ratio is assumed to be 1.750. A Wald statistic is used to construct the confidence interval.

This report shows the power for each of the scenarios.

Plot Section





These plots show the sample size for the various values of the other parameters.

Example 2 – Validation for a Binary Covariate

We could not find a direct validation result in the literature, so we decided to create one. This is easy to do in this case because we can create a dataset, analyze it with a statistical program such as **NCSS**, and then compare these results to those obtained with the above formulas in **PASS**.

Here is a summary of the data that was used to generate this example. The numeric values are counts of the number of items in the corresponding cell.

Group	X=1	X=0	Total
Y=1	8	26	34
Y=0	31	10	41
Total	39	36	75

Here is a printout from **NCSS** showing the estimated odds ratio (0.09926) and confidence interval (0.03419 to 0.28816).

Odds Ratios

Independent Variable X	Regression Coefficient b(i)	Odds Ratio Exp(b(i))	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Intercept	0.95551	2.60000	1.25383	5.39149
(X="B")	-2.31006	0.09926	0.03419	0.28816

Note that the simple odds ratio can also be calculated directly from the above table using the definition of the odds ratio. The formula gives (8 x 10) / (31 x 26) = 80 / 806 = 0.09926 which matches the value in the printout.

Note that the value of *P0* is 26 / 36 = 0.72222222 and *Percent with X = 1* is 100 x 39 / 75 = 52%.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for the Odds Ratio in Logistic Regression with One Binary X** 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	**Precision**
Interval Type	**Two-Sided**
Confidence Level	**0.95**
Sample Size	**75**
P0 [Pr(Y=1|X=0)]	**0.72222222**
Odds Ratio (Odds1/Odds0)	**0.09925558**
Percent with X = 1	**52**

Output

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

Numeric Results

Numeric Results for Two-Sided Confidence Interval of ORyx

Confidence Level	N	C.I. Width	ORyx	Lower Conf Limit of ORyx	Upper Conf Limit of ORyx	P0	Percent X=1
0.950	75	0.2540	0.099	0.034	0.288	0.722	52.0

Using the above settings, **PASS** also calculates the confidence interval to be (0.034, 0.288) which leads to a C. I. Width of 0.254. This validates the procedure with an independent calculation.