Chapter 866

Confidence Intervals for the Odds Ratio in Logistic Regression with Two Binary X's

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 are two binary covariates (X and Z) in the logistic regression model and a Wald statistic is used to calculate a confidence interval for the odds ratio of Y and X. Often, Y is called the *response* variable, the first binary covariate, X, is referred to as the *exposure* variable and the second binary covariate, Z, is referred to as the *confounder* variable. For example, Y might refer to the presence or absence of cancer and X might indicate whether the subject smoked or not, and Z is the presence or absence of a certain gene.

Sample Size Calculations

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

$$X\ \Pr(Y = 1|X, Z) = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 Z)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 Z)}$$

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

$$\log\left(\frac{\Pr(Y = 1|X, Z)}{1 - \Pr(Y = 1|X, Z)}\right) = \beta_0 + \beta_1 X + \beta_2 Z$$

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 defines the baseline probability as

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

The odds ratio between Y and X is defined as

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

It well known that the distribution of the maximum likelihood estimate of β_1 is asymptotically normal. A 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 *ORyx* 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 *ORyx*.

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 this variance from various quantities that can be set at the planning stage. Let p_x be the probability that X = 1 in the sample. Similarly, let p_z be the probability that Z = 1 in the sample.

Define the relationship between X and Z as a logistic regression as follows

$$\Pr(X = 1|Z) = \frac{\exp(\gamma_0 + \gamma_1 Z)}{1 + \exp(\gamma_0 + \gamma_1 Z)}$$

The value of γ_0 is found from

$$\exp(\gamma_0) = \frac{Q + \sqrt{Q^2 + 4p_x(1 - p_x)\exp(\gamma_1)}}{2(1 - p_x)\exp(\gamma_1)}$$

$$Q = p_x\left(1 + \exp(\gamma_1)\right) + p_z\left(1 - \exp(\gamma_1)\right) - 1$$

The information matrix is

$$I = \begin{bmatrix} L + F + J + H & F + H & J + H \\ F + H & F + H & H \\ J + H & H & J + H \end{bmatrix}$$

where

$$L = \frac{(1 - p_z)\exp(\beta_0)}{\left(1 + \exp(\gamma_0)\right)\left(1 + \exp(\beta_0)\right)^2}$$

$$H = \frac{p_z \exp(\beta_0 + \beta_1 + \beta_2 + \gamma_0 + \gamma_1)}{\left(1 + \exp(\gamma_0 + \gamma_1)\right)\left(1 + \exp(\beta_0 + \beta_1 + \beta_2)\right)^2}$$

$$F = \frac{(1-p_z)\exp(\beta_0+\beta_1+\gamma_0)}{\left(1+\exp(\gamma_0)\right)\left(1+\exp(\beta_0+\beta_1)\right)^2}$$

$$J = \frac{p_z\exp(\beta_0+\beta_2)}{\left(1+\exp(\gamma_0+\gamma_1)\right)\left(1+\exp(\beta_0+\beta_2)\right)^2}$$

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

The values of the regression coefficients are input as P_0 and the following odds ratio as follows

$$ORyx = \exp(\beta_1)$$

$$ORyz = \exp(\beta_2)$$

$$ORxz = \exp(\gamma_1)$$

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

The value of β_0 is calculated from P_0 using

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

Thus, the confidence interval can be specified in terms of several odds ratios and P_0. Of course, these results are only approximate. The width of 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, Z = 0)]

This gives the value of the baseline probability of a response, *P*0, when neither the exposure nor confounder are present.

*P*0 is a probability, so it must be between zero and one.

Odds Ratios

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.

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 $0 < \text{ORyx} < \infty$ (typically, $0.1 < \text{ORyx} < 10$).

ORyz (Y,Z Odds Ratio)

Specify one or more values of the Odds Ratio of Y and Z, a measure of the relationship between Y and Z. This is the ratio of the odds of the outcome Y given that the exposure Z = 1 to the odds of Y = 1 given Z = 0.

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 $0 < ORyz < \infty$ (typically, $0.1 < ORyz < 10$).

ORxz (X,Z Odds Ratio)

Specify one or more values of the Odds Ratio of X and Z, a measure of the relationship between X and Z. This is the ratio of the odds of the exposure X given that the confounder Z = 1 to the odds that X = 1 given Z = 0.

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 $0 < ORxz < \infty$ (typically, $0.1 < ORxz < 10$).

Prevalences

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.

Percent with Z = 1

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

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

Example 1 – Finding 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. A second variable, the presence or absence of a certain gene, is also thought to impact the result.

The baseline cancer event rate is 5%. The researchers want a sample size large enough to create a confidence interval with a width of 0.9. They estimate that the odds ratio between Y and X (ORyx) will be 2.0 and the confidence level will be 0.95. They want to look at the sensitivity of the analysis to the specification of the other odds ratios, so they want to obtain the results ORyz = 1, 1.5, 2 and ORxz = 1, 1.5, 2. The researchers estimate that about 40% of the sample eat the food being studied and about 25% will have the gene of interest.

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 Two Binary X's** 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, Z=0)]	**0.05**
ORyx (Y, X Odds Ratio)	**2**
ORyz (Y, Z Odds Ratio)	**1 1.5 2**
ORxz (X, Z Odds Ratio)	**1 1.5 2**
Percent with X = 1	**40**
Percent with Z = 1	**25**

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

Conf Level	N	C.I. Width	ORyx	Lower C.L. ORyx	Upper C.L. ORyx	ORyz	ORxz	P0	Pct X=1	Pct Z=1
0.950	4946	0.8999	2.000	1.600	2.500	1.000	1.000	0.050	40.0	25.0
0.950	4984	0.8999	2.000	1.600	2.500	1.000	1.500	0.050	40.0	25.0
0.950	5057	0.9000	2.000	1.600	2.500	1.000	2.000	0.050	40.0	25.0
0.950	4497	0.8999	2.000	1.600	2.500	1.500	1.000	0.050	40.0	25.0
0.950	4526	0.9000	2.000	1.600	2.500	1.500	1.500	0.050	40.0	25.0
0.950	4596	0.9000	2.000	1.600	2.500	1.500	2.000	0.050	40.0	25.0
0.950	4170	0.8999	2.000	1.600	2.500	2.000	1.000	0.050	40.0	25.0
0.950	4191	0.9000	2.000	1.600	2.500	2.000	1.500	0.050	40.0	25.0
0.950	4257	0.9000	2.000	1.600	2.500	2.000	2.000	0.050	40.0	25.0

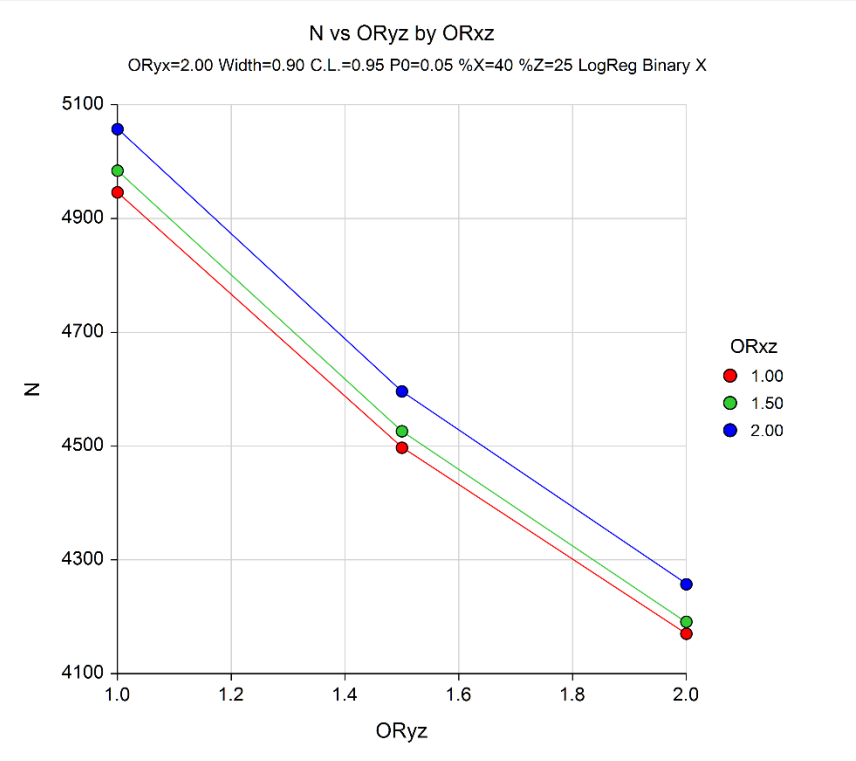
References
Demidenko, Eugene. 2007. 'Sample size determination for logistic regression revisited', Statistics in Medicine, Volume 26, pages 3385-3397.
Demidenko, Eugene. 2008. 'Sample size and optimal design for logistic regression with binary interaction', Statistics in Medicine, Volume 27, pages 36-46.
Rochon, James. 1989. 'The Application of the GSK Method to the Determination of Minimum Sample Sizes', Biometrics, Volume 45, pages 193-205.

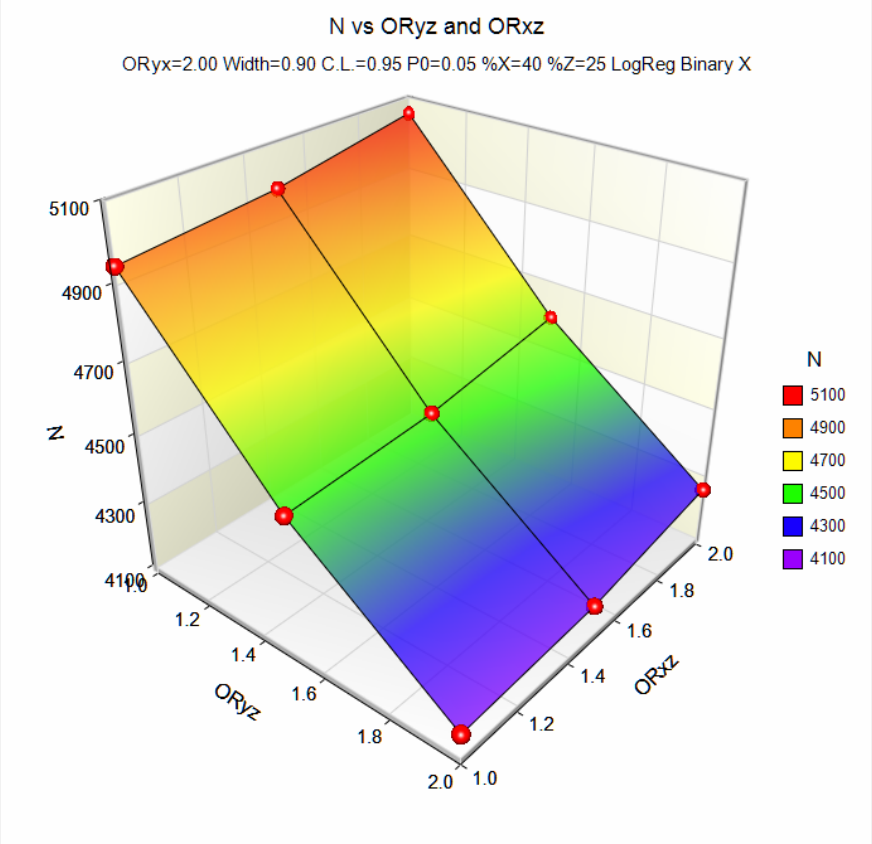
Report Definitions
Logistic regression equation: Log(P/(1-P)) = β0 + β1×X + β2×Z, where P = Pr(Y = 1|X, Z) and X and Z are binary.
Confidence Level is the proportion of studies with the same settings that produce a confidence interval that includes the true ORyx.
N is the sample size.
C.I. Width is the 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).
ORyz = Exp(β2) is the odds ratio of Y versus Z.
ORxz is the odds ratio of X versus Z in a logistic regression of X on Z.
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, Z = 0).
Percent X=1 is the percent of the sample in which the exposure is 1 (present).
Percent Z=1 is the percent of the sample in which the confounder is 1.

Summary Statements
A logistic regression of a binary response variable (Y) on two binary independent variables (X and Z) with a sample size of 4946 observations at a 0.950 confidence level produces a two-sided confidence interval for the odds ratio of Y and X with a width of 0.8999. The sample odds ratio between Y and X is assumed to be 2.000. Other settings are ORyz = 1.000, ORxz = 1.000, and P0 (prevalence of Y given X = 0 and Z = 0) = 0.050. The prevalence of X is 40.0% and the prevalence of Z is 25.0%. A Wald statistic is used to construct the confidence interval.

This report shows the sample size 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 will create one by hand. 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	Y=1	Y=0	Total
X=1, Z=1	**5**	**10**	**15**
X=1, Z=0	**3**	**21**	**24**
X=0, Z=1	**17**	**3**	**20**
X=0, Z=0	**9**	**7**	**16**
Total	**34**	**41**	**75**

Here is a printout from **NCSS** showing the estimated odds ratio (0.09932) and confidence interval (0.03213 to 0.30698). The width is 0.27485.

Odds Ratios

Independent Variable X	Regression Coefficient b(i)	Odds Ratio Exp(b(i))	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Intercept	0.48280	1.62061	0.75388	3.48380
(X=1)	-2.30943	0.09932	0.03213	0.30698
(Z=1)	1.37241	3.94483	1.27799	12.17670

Note that the value of *P0* is 9 / 16 = 0.5625. The value of *Percent with X = 1* is 100 x 39 / 75 = 52%. The value of *Percent with Z = 1* is 100 x 35 / 75 = 46.67%. Also, a logistic regression of X on Z produced an *ORxz* of 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 **Confidence Intervals for the Odds Ratio in Logistic Regression with Two Binary X's** 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 (C.I. Width)**
Interval Type	**Two-Sided**
Confidence Level	**0.95**
N (Sample Size)	**75**
P0 [Pr(Y=1|X=0, Z=0)]	**0.5625**
ORyx (Y, X Odds Ratio)	**0.09932**
ORyz (Y, Z Odds Ratio)	**3.94483**
ORxz (X, Z Odds Ratio)	**0.50**
Percent with X = 1	**52**
Percent with Z = 1	**46.6666667**

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

Conf Level	N	C.I. Width	ORyx	Lower C.L. ORyx	Upper C.L. ORyx	ORyz	ORxz	P0	Pct X=1	Pct Z=1
0.950	75	0.27567	0.0993	0.0321	0.3077	3.9448	0.5000	0.5625	52.0	46.7

Using the above settings, **PASS** calculates the confidence interval to be (0.0321, 0.3077) which leads to a C. I. Width of 0.27567. The width given in the **NCSS** run was 0.27485. These values are slightly different because the regression model presented in this chapter does not include the interaction, so it is not a saturated model. Hence, it does not reproduce the results exactly.