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

$$X\,\Pr(Y=1|X,Z)=\frac{\exp(\beta_0+\beta_1X+\beta_2Z)}{1+\exp(\beta_0+\beta_1X+\beta_2Z)}$$

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

$$\log\left(\frac{\Pr(Y=1|X,Z)}{1-\Pr(Y=1|X,Z)}\right)=\beta_0+\beta_1X+\beta_2Z$$

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 is 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

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

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

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

where

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

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

The values of the regression coefficients are input as *P₀*, and the following odds ratio as follows

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

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

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

The value of $\sqrt{N} s_{\hat{\beta}_1}$ 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.

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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**

Output

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

Numeric Reports

Numeric Results

Solve For: Sample Size
Interval Type: Two-Sided

Confidence Level	Sample Size N	Confidence Interval Width	Odds Ratio ORyx	Confidence Interval Limits of ORyx		Other Odds Ratios		P0	Percent with	
				Lower	Upper	ORyz	ORxz		X = 1	Z = 1
0.95	4946	0.8999	2	1.6	2.5	1.0	1.0	0.05	40	25
0.95	4984	0.8999	2	1.6	2.5	1.0	1.5	0.05	40	25
0.95	5057	0.9000	2	1.6	2.5	1.0	2.0	0.05	40	25
0.95	4497	0.8999	2	1.6	2.5	1.5	1.0	0.05	40	25
0.95	4526	0.9000	2	1.6	2.5	1.5	1.5	0.05	40	25
0.95	4596	0.9000	2	1.6	2.5	1.5	2.0	0.05	40	25
0.95	4170	0.8999	2	1.6	2.5	2.0	1.0	0.05	40	25
0.95	4191	0.9000	2	1.6	2.5	2.0	1.5	0.05	40	25
0.95	4257	0.9000	2	1.6	2.5	2.0	2.0	0.05	40	25

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 The proportion of studies with the same settings that produce a confidence interval that includes the true ORyx.
N The sample size.
C.I. Width The distance between the two boundaries of the confidence interval.
ORyx The expected sample value of the odds ratio. ORyx = exp(β1).
C.I. Limits of ORyx The lower and upper limits of the confidence interval of ORyx.
ORyz The odds ratio of Y versus Z. ORyz = exp(β2).
ORxz The odds ratio of X versus Z in a logistic regression of X on Z.
P0 The response probability at X = 0. That is, P0 = Pr(Y = 1|X = 0, Z = 0).
Percent with X = 1 The percent of the sample in which the exposure is 1 (present).
Percent with Z = 1 The percent of the sample in which the confounder is 1.

Summary Statements

A logistic regression model design with a binary response variable (Y) and two binary independent variables (X and Z) will be used to obtain a two-sided 95% confidence interval for the odds ratio of Y to X. A Wald statistic is to be used in the construction of the confidence interval. The baseline response rate of Y given X = 0 and Z = 0 is assumed to be 0.05 and the sample odds ratio between Y and X is assumed to be 2. The odds ratio of Y and Z is assumed to be 1 and the odds ratio of X and Z (between covariates) is assumed to be 1. The percent of observations with X = 1 is assumed to be 40% and the percent of observations with Z = 1 is assumed to be 25%. To produce a confidence interval with a width of no more than 0.9, 4946 subjects will be needed.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout- Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	4946	6183	1237
20%	4984	6230	1246
20%	5057	6322	1265
20%	4497	5622	1125
20%	4526	5658	1132
20%	4596	5745	1149
20%	4170	5213	1043
20%	4191	5239	1048
20%	4257	5322	1065

Dropout Rate The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.

N The evaluable sample size at which the confidence interval is computed. If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated confidence interval.

N' The total number of subjects that should be enrolled in the study in order to obtain N evaluable subjects, based on the assumed dropout rate. After solving for N, N' is calculated by inflating N using the formula N' = N / (1 - DR), with N' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)

D The expected number of dropouts. D = N' - N.

Dropout Summary Statements

Anticipating a 20% dropout rate, 6183 subjects should be enrolled to obtain a final sample size of 4946 subjects.

References

Demidenko, Eugene. 2007. 'Sample size determination for logistic regression revisited', Statistics in Medicine, Volume 26, pages 3385-3397.

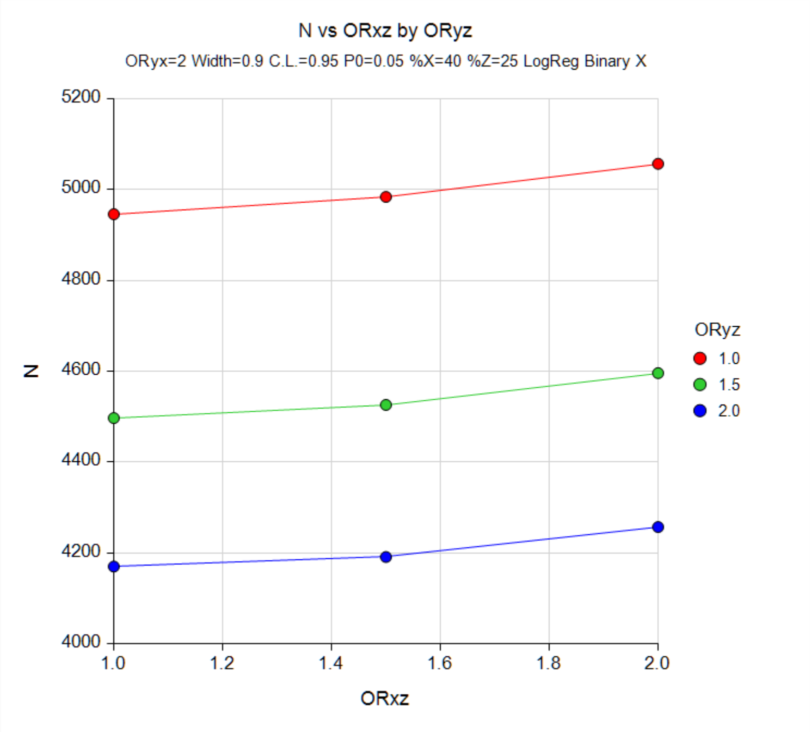
Demidenko, Eugene. 2008. 'Sample size and optimal design for logistic regression with binary interaction', Statistics in Medicine, Volume 27, pages 36-46.

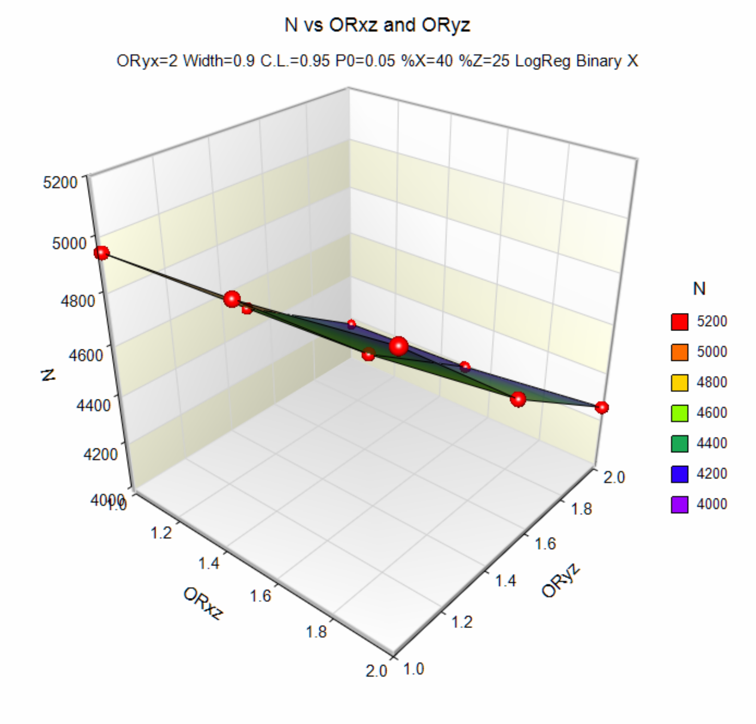
Rochon, James. 1989. 'The Application of the GSK Method to the Determination of Minimum Sample Sizes', Biometrics, Volume 45, pages 193-205.

This report shows the sample size for each of the scenarios.

Plots Section

Plots





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

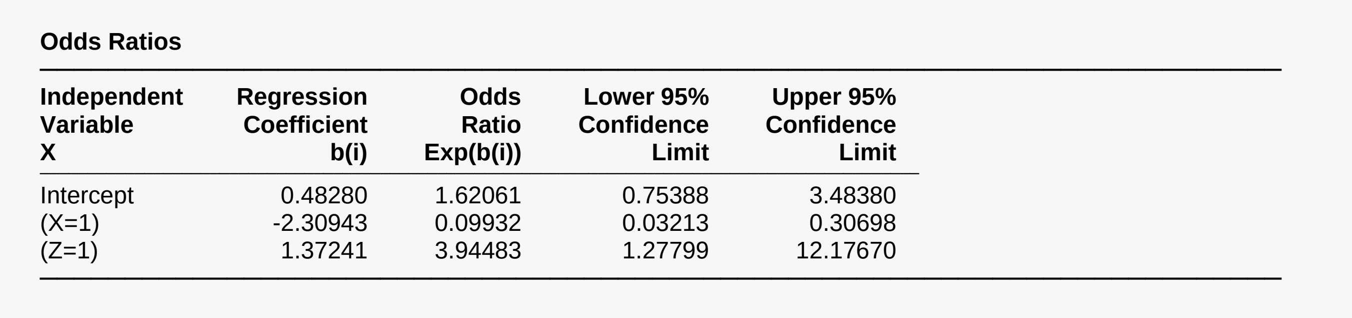
Example 2 – Validation for a Binary Covariate

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.

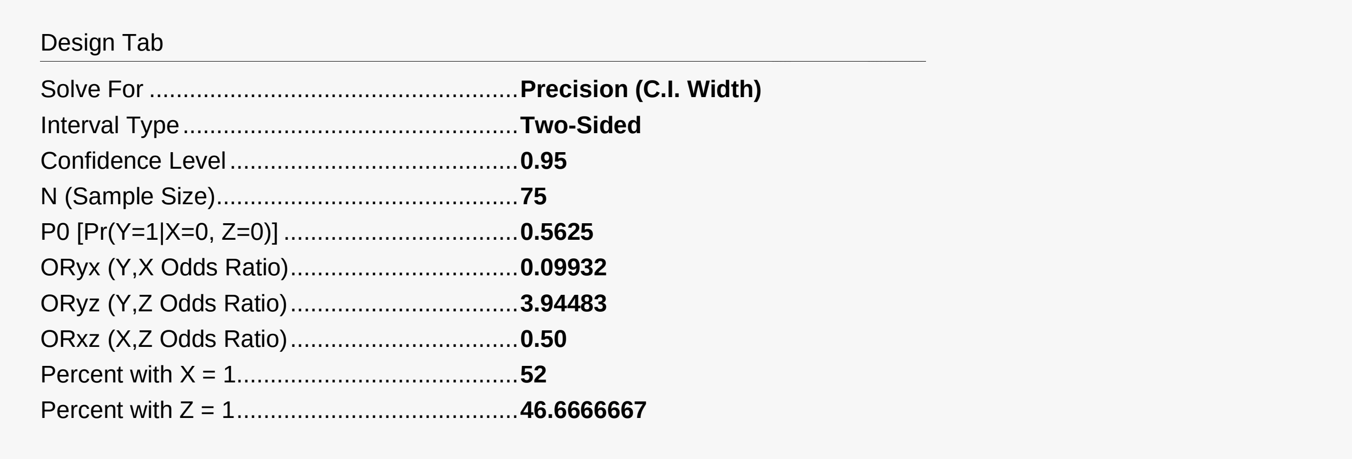


```
Odds Ratios
──────────────────────────────────────────────────────────────────────────────────
Independent        Regression        Odds      Lower 95%      Upper 95%
Variable          Coefficient       Ratio     Confidence     Confidence
X                        b(i)   Exp(b(i))          Limit          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

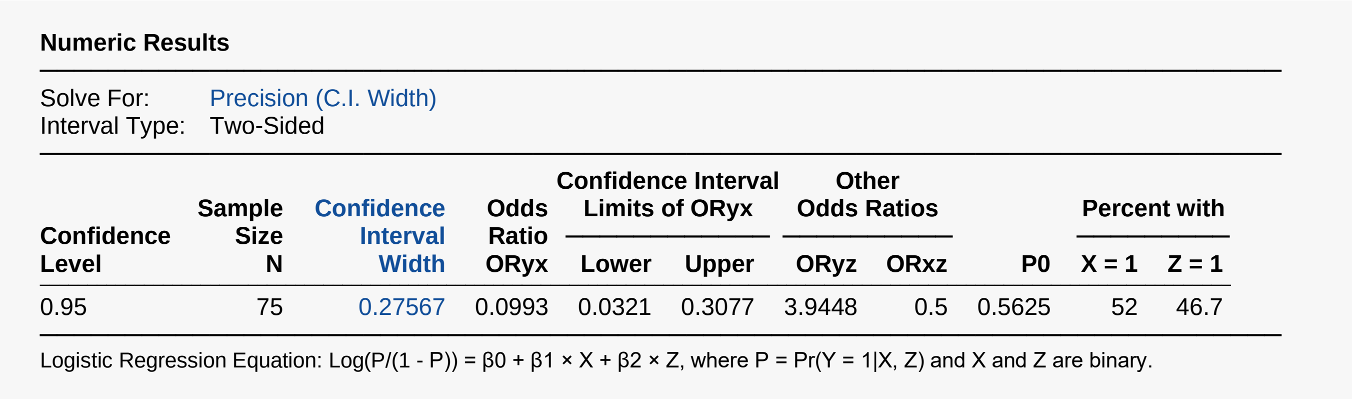
If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.



```
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

Solve For:      Precision (C.I. Width)
Interval Type:  Two-Sided

                                                 Confidence Interval      Other
              Sample   Confidence    Odds        Limits of ORyx        Odds Ratios                 Percent with
Confidence      Size     Interval    Ratio     -----------------    ---------------               --------------
Level              N        Width     ORyx      Lower     Upper      ORyz     ORxz        P0     X = 1    Z = 1
0.95              75      0.27567   0.0993     0.0321    0.3077    3.9448      0.5    0.5625        52     46.7

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.
```

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.