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

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

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

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

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.

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

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)]	**0.07**
Odds Ratio (Odds1/Odds0)	**1.75 2.0 2.25**
Percent with X = 1	**25 50**

Output

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

Numeric Reports

Numeric Results for Two-Sided Confidence Interval of ORyx

Solve For: Sample Size

Confidence Level	N	C.I. Width	ORyx	Lower Conf Limit of ORyx	Upper Conf Limit of ORyx	P0	Percent X = 1
0.95	3525	0.8999	1.75	1.357	2.257	0.07	25
0.95	2979	0.8999	1.75	1.357	2.257	0.07	50
0.95	4294	0.9000	2.00	1.600	2.500	0.07	25
0.95	3727	0.9000	2.00	1.600	2.500	0.07	50
0.95	5136	0.9000	2.25	1.845	2.745	0.07	25
0.95	4561	0.9000	2.25	1.845	2.745	0.07	50

Logistic Regression Equation: $\text{Log}(P/(1 - P)) = \beta 0 + \beta 1 \times X$, where $P = \Pr(Y = 1|X)$ and X is 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. of ORyx Lower Limit	The lower limit of the confidence interval of ORyx.
C.I. of ORyx Upper Limit	The upper limit of the confidence interval of ORyx.
P0	The response probability at X = 0. That is, P0 = Pr(Y = 1|X = 0).
Percent X = 1	The percent of the sample in which the exposure is 1 (present).

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% are in the group X=1) at a 0.95 confidence level produces a two-sided confidence interval with a width of 0.8999. The baseline response rate is assumed to be 0.07 and the sample odds ratio is assumed to be 1.75. A Wald statistic is used to construct the confidence interval.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	3525	4407	882
20%	2979	3724	745
20%	4294	5368	1074
20%	3727	4659	932
20%	5136	6420	1284
20%	4561	5702	1141

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, 4407 subjects should be enrolled to obtain a final sample size of 3525 subjects.

References

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

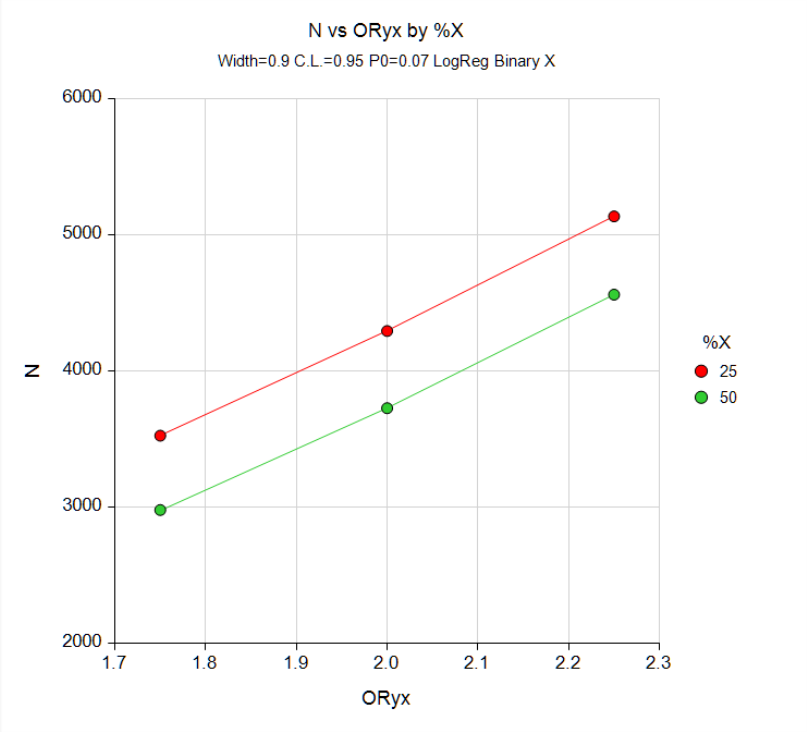
Demidenko, Eugene. 2008. 'Sample size and optimal design for logistic regression with binary interaction', Statistics in Medicine, Volume 27, pages 36-46.

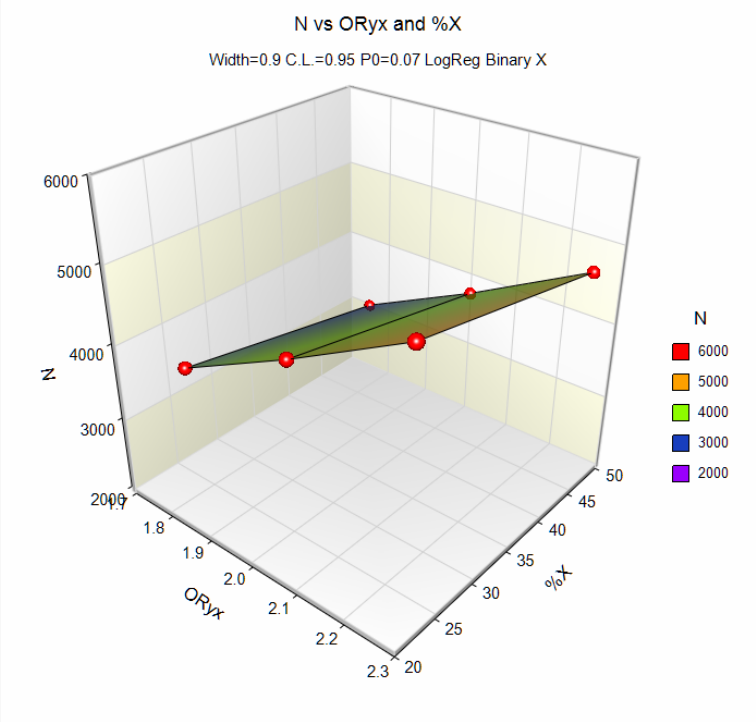
Rochon, James. 1989. 'The Application of the GSK Method to the Determination of Minimum Sample Sizes', Biometrics, Volume 45, pages 193-205.

This report shows the power 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 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	1.14272	3.13529	1.77260	5.54557
(X=1)	-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

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**
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 for Two-Sided Confidence Interval of ORyx

Solve For: Precision (C.I. Width)

Confidence Level	N	C.I. Width	ORyx	Lower Conf Limit of ORyx	Upper Conf Limit of ORyx	P0	Percent X = 1
0.95	75	0.254	0.099	0.034	0.288	0.722	52

Logistic Regression Equation: $\text{Log}(P/(1 - P)) = \beta 0 + \beta 1 \times X$, where $P = \Pr(Y = 1|X)$ and X is binary.

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.