

Chapter 867

Confidence Intervals for the Interaction 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) and their interaction (XZ) in the logistic regression model and a Wald statistic is used to calculate a confidence interval for the interaction odds ratio (OR_{int}). 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

$$X \Pr(Y = 1|X, Z) = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ)}$$

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

$$\log\left(\frac{\Pr(Y = 1|X, Z)}{1 - \Pr(Y = 1|X, Z)}\right) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

In the logistic regression model, the magnitude of the association of XZ and Y is represented by the slope β_3 . Since XZ is binary, only two cases need be considered: XZ = 0 and XZ = 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 XZ is defined as

$$OR_{int} = \exp(\beta_3)$$

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

It is well known that the distribution of the maximum likelihood estimate of β_3 is asymptotically normal. A confidence interval for this slope is commonly formed from the Wald statistic

$$z = \frac{\hat{\beta}_3}{s_{\hat{\beta}_3}}$$

A $(1 - \alpha)\%$ two-sided confidence interval for β_3 is

$$\hat{\beta}_3 \pm z_{1-\frac{\alpha}{2}} s_{\hat{\beta}_3}$$

By transforming this interval into the odds ratio scale by exponentiating both limits, a $(1 - \alpha)\%$ two-sided confidence interval for OR_{int} is

$$(OR_{LL}, OR_{UL}) = \exp\left(\hat{\beta}_3 \pm z_{1-\frac{\alpha}{2}} s_{\hat{\beta}_3}\right)$$

Note that this interval is not symmetric about OR_{int} .

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 $s_{\hat{\beta}_3}$ is not known before the study so this quantity must be estimated. Demidenko (2008) 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

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

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

The information matrix for this model is

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

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

where

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

$$R = \frac{p_z \exp(\beta_0 + \beta_1 + \beta_2 + \beta_3 + \gamma_0 + \gamma_1)}{(1 + \exp(\gamma_0 + \gamma_1))(1 + \exp(\beta_0 + \beta_1 + \beta_2 + \beta_3))^2}$$

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

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

The value of $\sqrt{N}s_{\hat{\beta}_3}$ is the (4, 4) element of the inverse of I .

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

$$OR_{int} = \exp(\beta_3)$$

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

$$OR_{yz} = \exp(\beta_2)$$

$$OR_{xz} = \exp(\gamma_1)$$

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 – 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. A second variable, the presence or absence of a certain gene, is also thought to impact the result. The researchers want a sample size large enough to guarantee that the width of a confidence interval about ORint is less than 0.90.

The baseline cancer event rate is 5%. They estimate ORint will be 0.5 and they want a confidence level of 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. They estimate ORyx at 1.5. 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	0.9
P0 [Pr(Y=1 X=0, Z=0)]	0.05
ORint (X,Z Interaction Odds Ratio)	0.5
ORyx (Y,X Odds Ratio)	1.5
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

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

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 ORint	Confidence Interval Limits of ORint		Other Odds Ratios			Percent with		
				Lower	Upper	ORyx	ORyz	ORxz	P0	X = 1	Z = 1
0.95	2995	0.8999	0.5	0.223	1.123	1.5	1.0	1.0	0.05	40	25
0.95	2868	0.8999	0.5	0.223	1.123	1.5	1.0	1.5	0.05	40	25
0.95	2845	0.8999	0.5	0.223	1.123	1.5	1.0	2.0	0.05	40	25
0.95	2253	0.9000	0.5	0.223	1.123	1.5	1.5	1.0	0.05	40	25
0.95	2169	0.8999	0.5	0.223	1.123	1.5	1.5	1.5	0.05	40	25
0.95	2156	0.9000	0.5	0.223	1.123	1.5	1.5	2.0	0.05	40	25
0.95	1884	0.8998	0.5	0.223	1.123	1.5	2.0	1.0	0.05	40	25
0.95	1821	0.8998	0.5	0.223	1.122	1.5	2.0	1.5	0.05	40	25
0.95	1813	0.8999	0.5	0.223	1.123	1.5	2.0	2.0	0.05	40	25

Logistic Regression Equation: $\text{Log}(P/(1 - P)) = \beta_0 + \beta_1 \times X + \beta_2 \times Z + \beta_3 \times X \times Z$, where $P = \text{Pr}(Y = 1|X, Z)$ 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 ORint.
N	The sample size.
C.I. Width	The distance between the two boundaries of the confidence interval.
ORint	The expected sample value of the interaction odds ratio. $\text{ORint} = \exp(\beta_3)$.
C.I. Limits of ORint	The lower and upper limits of the confidence interval of ORint.
ORyx	The expected sample value of the odds ratio. $\text{ORyx} = \exp(\beta_1)$.
ORyz	The odds ratio of Y versus Z. $\text{ORyz} = \exp(\beta_2)$.
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, $P0 = \text{Pr}(Y = 1 X = 0, Z = 0)$.
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 interaction odds ratio of X to Z. 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 of the interaction is assumed to be 0.5. The odds ratio of Y and X is assumed to be 1.5, 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, 2995 subjects will be needed.

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

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout- Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	2995	3744	749
20%	2868	3585	717
20%	2845	3557	712
20%	2253	2817	564
20%	2169	2712	543
20%	2156	2695	539
20%	1884	2355	471
20%	1821	2277	456
20%	1813	2267	454

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, 3744 subjects should be enrolled to obtain a final sample size of 2995 subjects.

References

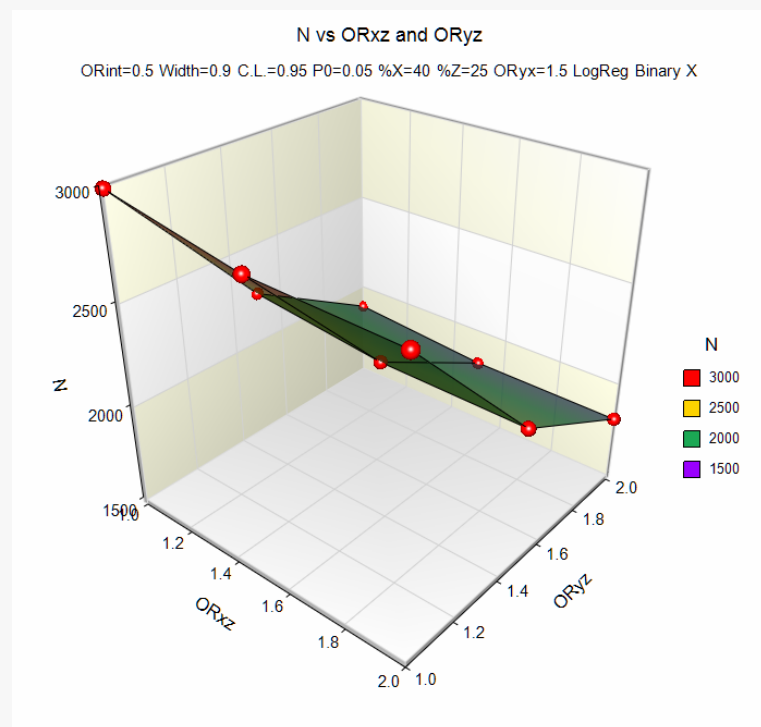
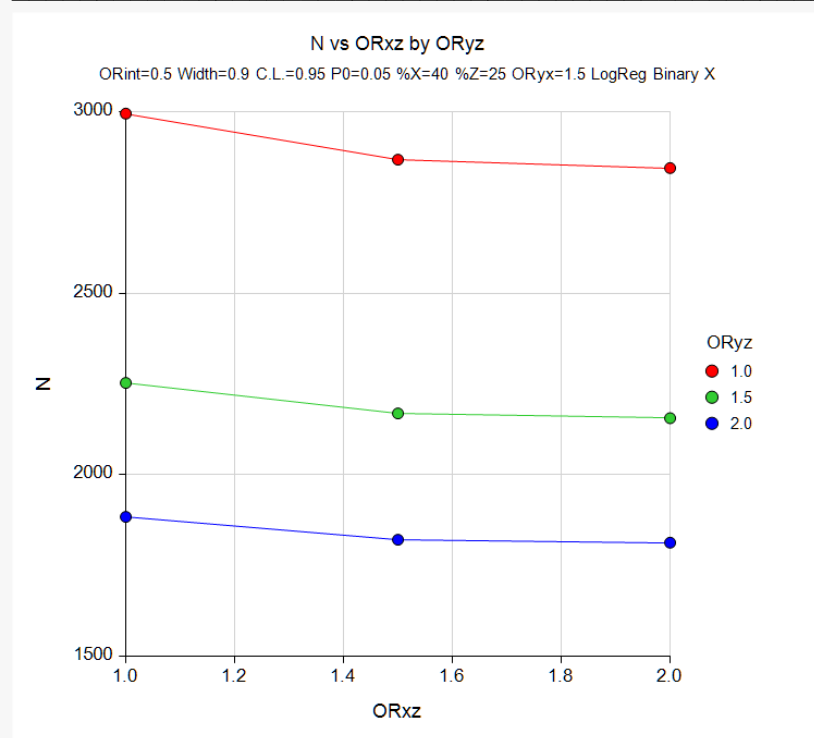
- Demidenko, Eugene. 2007. 'Sample size determination for logistic regression revisited', Statistics in Medicine, Volume 26, pages 3385-3397.
- Demidenko, Eugene. 2008. 'Sample size and optimal design for logistic regression with binary interaction', Statistics in Medicine, Volume 27, pages 36-46.
- 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.

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

Plots Section

Plots



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

Example 2 – Validation for the Interaction

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 (0.79412) and confidence interval (0.08304 to 7.59380). The width is 7.51076.

Odds Ratios

Independent Variable X	Regression Coefficient b(i)	Odds Ratio Exp(b(i))	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Intercept	0.43853	1.55042	0.64509	3.72629
(X=1)	-2.19722	0.11111	0.02331	0.52968
(Z=1)	1.48329	4.40741	0.91195	21.30077
(X=1)*(Z=1)	-0.23052	0.79412	0.08304	7.59380

Note that the value of $P0$ is $9 / 16 = 0.5625$. The value of *Percent with X = 1* is $100 \times 39 / 75 = 52\%$. The value of *Percent with Z = 1* is $100 \times 35 / 75 = 46.67\%$. Also, a logistic regression of X on Z produced an OR_{xz} 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**
ORint (Interaction Odds Ratio) **0.79412**
ORyx (Y,X Odds Ratio) **0.1111**
ORyz (Y,Z Odds Ratio) **4.40741**
ORxz (X,Z Odds Ratio) **0.50**
Percent with X = 1 **52**
Percent with Z = 1 **46.666667**

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 Level	Sample Size N	Confidence Interval Width	Odds Ratio ORint	Confidence Interval Limits of ORint		Other Odds Ratios			Percent with		
				Lower	Upper	ORyx	ORyz	ORxz	P0	X = 1	Z = 1
0.95	75	7.51103	0.79412	0.08304	7.59407	0.1111	4.40741	0.5	0.5625	52	46.7

Logistic Regression Equation: $\text{Log}(P/(1 - P)) = \beta_0 + \beta_1 \times X + \beta_2 \times Z + \beta_3 \times X \times Z$, where $P = \text{Pr}(Y = 1|X, Z)$ and X and Z are binary.

Using the above settings, **PASS** calculates the confidence interval to be (0.08304, 7.59407) which leads to a C. I. Width of 7.51103. This matches the results obtained from the data to within rounding.