Chapter 862

Tests for the Odds Ratio in Logistic Regression with Two Binary X's (Wald Test)

Introduction

Logistic regression expresses the relationship between a binary response variable and one or more independent variables called *covariates*. This procedure is for the case when there are two binary covariates (X and Z) in the logistic regression model and Wald tests are used to test their significance. 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.

Power 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 relationship between X and the response Y is represented by the slope β_1.

The logistic regression model defines the baseline probability

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

The significance of the slope β_1 is commonly tested with the Wald test

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

It is considered good practice to base the power analysis on the same test statistic that is used for analysis, so we base our power analysis on the above Wald test.

Demidenko (2007) gives the following formula for the power of the two-sided Wald test in this as

$$\text{Power} = \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{\beta_1\sqrt{N}}{\sqrt{V}}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\beta_1\sqrt{N}}{\sqrt{V}}\right)$$

where z is the usual quantile of the standard normal distribution and V is calculated as follows.

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

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

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

The information matrix for this model 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

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

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

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

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

The value of V 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

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

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

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

Example 1 – Sample Size for Various Odds Ratios

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 detect an odds ratio of 2.0 with 80% power at the 0.05 significance level with a two-sided Wald test. They want to look at the sensitivity of the analysis to the specification of the odds ratios, so they also want to obtain the results ORyz = 1, 1.5, 2 and ORxz = 1, 1.5, 2. The researchers determine that about 40% of the sample eat the food being studied. They also determine that 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**
Alternative Hypothesis	**Two-Sided**
Power	**0.80**
Alpha	**0.05**
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
Test Type:	Wald Test
Variables:	Y = Response, X = Exposure, Z = Confounder
Alternative Hypothesis:	H1: ORyx ≠ 1

Power	Sample Size N	Prevalence Percentages X = 1	Prevalence Percentages Z = 1	Baseline Probability Pr(Y = 1 | X = Z = 0) P0	Odds Ratio Y, X ORyx	Odds Ratio Y, Z ORyz	Odds Ratio X, Z ORxz	Alpha
0.8003	1048	40	25	0.05	2	1.0	1.0	0.05
0.8003	1056	40	25	0.05	2	1.0	1.5	0.05
0.8001	1071	40	25	0.05	2	1.0	2.0	0.05
0.8004	953	40	25	0.05	2	1.5	1.0	0.05
0.8003	959	40	25	0.05	2	1.5	1.5	0.05
0.8003	974	40	25	0.05	2	1.5	2.0	0.05
0.8001	883	40	25	0.05	2	2.0	1.0	0.05
0.8003	888	40	25	0.05	2	2.0	1.5	0.05
0.8003	902	40	25	0.05	2	2.0	2.0	0.05

Logistic Regression Equation: $\text{Log}(P/(1 - P)) = \beta0 + \beta1 \times X + \beta2 \times Z$, where $P = \Pr(Y = 1|X, Z)$ and X and Z are binary.

Power	The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N	The sample size.
P0	The response probability at X = 0, Z = 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.
Percent with Z = 1	The percent of the sample in which the confounder is 1.
ORyx	The odds ratio of Y versus X. This is the effect size. ORyx = Exp(β1).
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.
Alpha	The probability of rejecting a true null hypothesis.

Summary Statements

A two-binary-covariate logistic regression (binary response Y versus one binary X of interest and one binary confounder Z) design will be used to test whether the probability that Y equals 1 when X equals 1 (P1 = Pr(Y = 1|X = 1)) is different from the (baseline) probability that Y equals 1 when X and Z equal 0 (P0 = Pr(Y = 1|X = Z = 0)) of 0.05 (H0: ORyx = 1 versus H1: ORyx ≠ 1, given P0 = 0.05). The comparison will be made using a two-sided logistic regression Wald test of β1 (using the model Log(P / (1 - P)) = β0 + β1 × X + β2 × Z, where P = Pr(Y = 1|X, Z)), with a Type I error rate (α) of 0.05. Among subjects, 40% are assumed to have the value X = 1 (or be in the X = 1 group), and 25% are assumed to have the value Z = 1 (or be in the Z = 1 group). The odds ratio for Y and Z (ORyz) is assumed to be 1, and the odds ratio for X and Z (ORxz) is assumed to be 1. To detect a Y and X odds ratio (ORyx) of 2 with 80% power, the number of needed subjects will be 1048.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	1048	1310	262
20%	1056	1320	264
20%	1071	1339	268
20%	953	1192	239
20%	959	1199	240
20%	974	1218	244
20%	883	1104	221
20%	888	1110	222
20%	902	1128	226

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 power is computed. If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated power.
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, 1310 subjects should be enrolled to obtain a final sample size of 1048 subjects.

References

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

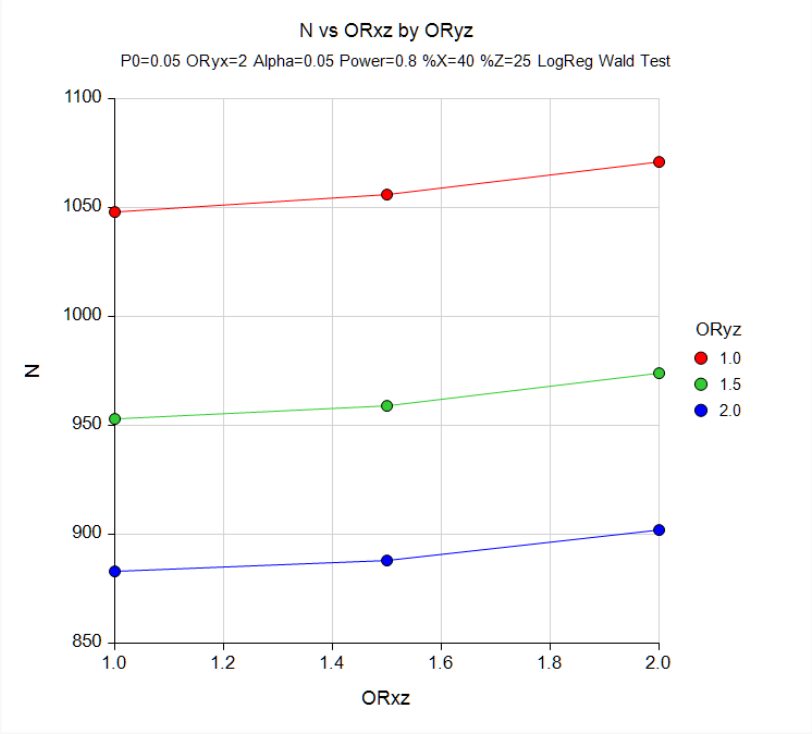
Demidenko, Eugene. 2008. 'Sample size and optimal design for logistic regression with binary interaction', Statistics in Medicine, Volume 27, pages 36-46.

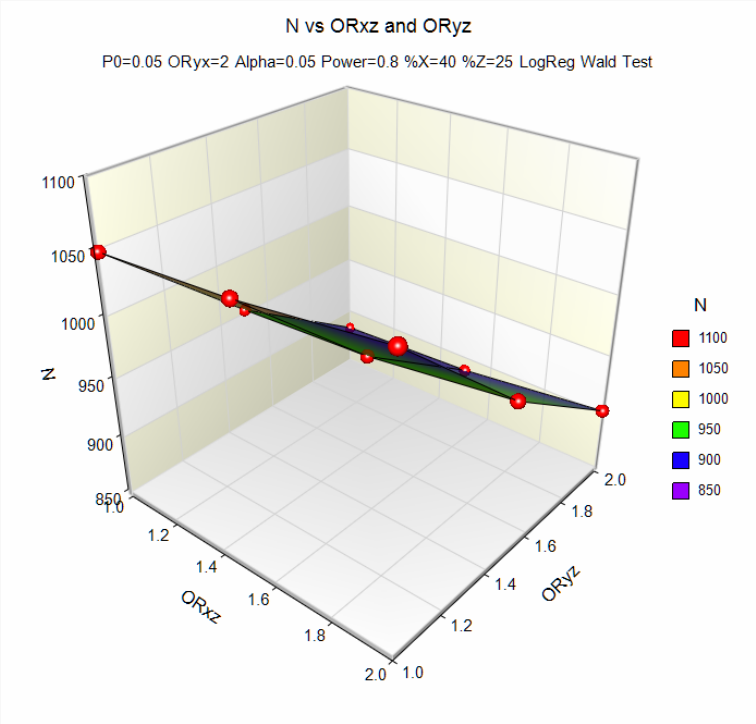
Rochon, James. 1989. 'The Application of the GSK Method to the Determination of Minimum Sample Sizes', Biometrics, Volume 45, pages 193-205.

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

Plots Section

Plots





These plots show the sample size versus the odds ratio for several scenarios.

Example 2 – Validation for a Binary Covariate using Demidenko (2007)

Demidenko (2007), page 3394, gives an example in which alpha = 0.05, power = 0.8, ORyx = 2, ORyz= 2, ORxz = 1, P0 = 0.1, percent X = 1 is 25, and percent Z = 1 is 50. These parameters give an N of 544. We calculated this amount using Demidenko's website: www.dartmouth.edu/~eugened/power-samplesize.php. We will validate this routine by running the same scenario.

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	**Sample Size**
Alternative Hypothesis	**Two-Sided**
Power	**0.8**
Alpha	**0.05**
P0 [Pr(Y = 1 | X = 0, Z = 0)]	**0.1**
ORyx (Y,X Odds Ratio)	**2**
ORyz (Y,Z Odds Ratio)	**2**
ORxz (X,Z Odds Ratio)	**1**
Percent with X = 1	**25**
Percent with Z = 1	**50**

Output

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

Numeric Results

Solve For:	Sample Size
Test Type:	Wald Test
Variables:	Y = Response, X = Exposure, Z = Confounder
Alternative Hypothesis:	H1: ORyx ≠ 1

		Prevalence Percentages			Odds Ratio			
Power	Sample Size N	X = 1	Z = 1	Baseline Probability Pr(Y = 1 | X = Z = 0) P0	Y, X ORyx	Y, Z ORyz	X, Z ORxz	Alpha
0.8005	545	25	50	0.1	2	2	1	0.05

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.

PASS calculates a sample size of 545 which is one more than Demidenko's. Note that an *N* of 544 achieves a power slightly less than the 0.8000 requested.