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 covariate (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

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

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

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

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

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

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

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

The value of 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

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

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

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

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 *Alpha*, *Power*, *Sample Size*, or *ORyx*. Select *Sample Size* when you want to calculate the sample size needed to achieve a given power and alpha level. Select *Power* when you want to calculate the power of an experiment.

Test

Alternative Hypothesis

Specify whether the test is one-sided or two-sided. When a two-sided hypothesis is selected, the value of alpha is halved. Everything else remains the same.

Commonly, accepted procedure is to use the Two-Sided option unless you can justify using a one-sided test.

Power and Alpha

Power

This option specifies one or more values for power. Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. A type-II error occurs when you fail to reject the null hypothesis of equal probabilities of the event of interest when in fact they are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha

This option specifies one or more values for the probability of a type-I error (alpha). A type-I error occurs when you reject the null hypothesis of equal probabilities when in fact they are equal.

Values of alpha must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

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.

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, $0.1 < \text{ORyx} < 10$). Since this is the value under alternative hypothesis, $\text{ORyx} \neq 1$.

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

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

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 – 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

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for the Odds Ratio in Logistic Regression with Two Binary X's (Wald Test)** 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**
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**

Annotated Output

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

Numeric Results

Numeric Results for Two-Sided Wald Test
Alternative Hypothesis: ORyx ≠ 1

Power	N	Percent X=1	Percent Z=1	P0	ORyx	ORyz	ORxz	Alpha	Beta
0.8003	1048	40.0	25.0	0.050	2.000	1.000	1.000	0.050	0.1997
0.8003	1056	40.0	25.0	0.050	2.000	1.000	1.500	0.050	0.1997
0.8001	1071	40.0	25.0	0.050	2.000	1.000	2.000	0.050	0.1999
0.8004	953	40.0	25.0	0.050	2.000	1.500	1.000	0.050	0.1996
0.8003	959	40.0	25.0	0.050	2.000	1.500	1.500	0.050	0.1997
0.8003	974	40.0	25.0	0.050	2.000	1.500	2.000	0.050	0.1997
0.8001	883	40.0	25.0	0.050	2.000	2.000	1.000	0.050	0.1999
0.8003	888	40.0	25.0	0.050	2.000	2.000	1.500	0.050	0.1997
0.8003	902	40.0	25.0	0.050	2.000	2.000	2.000	0.050	0.1997

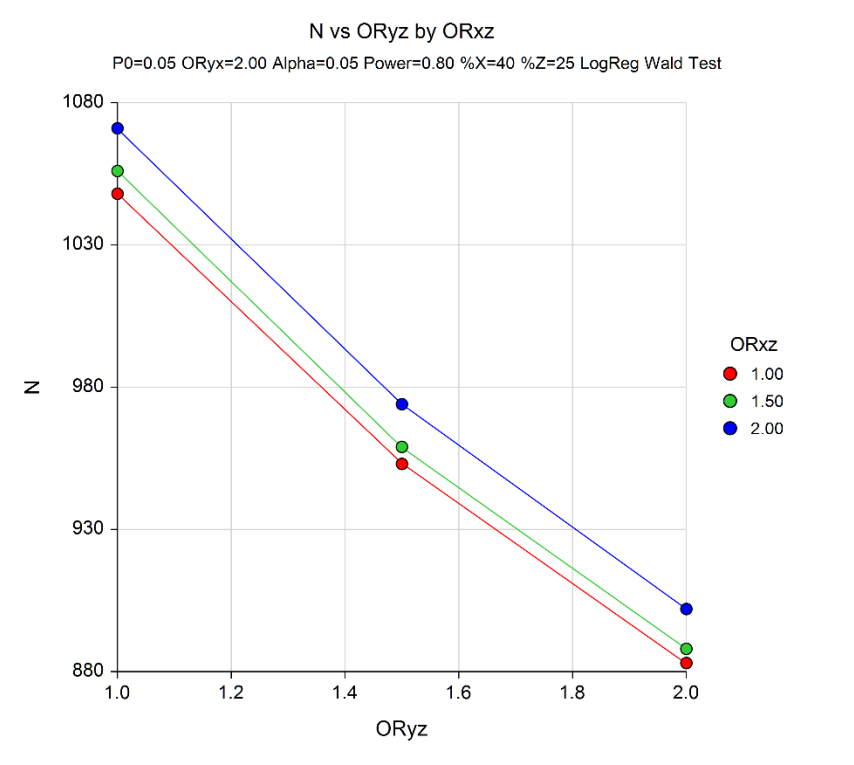
Tests for the Odds Ratio in Logistic Regression with Two Binary X's (Wald Test)

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.
Power is the probability of rejecting a false null hypothesis.
N is the sample size.
P0 is the response probability at X = 0, Z = 0. That is, P0 = Pr(Y = 1|X = 0, Z = 0).
Percent X=1 is the percent of the population in which the exposure is 1.
Percent Z=1 is the percent of the population in which the confounder is 1.
ORyx = Exp(β1) is the odds ratio of Y versus X. This is the effect size.
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.
Alpha is the probability of rejecting a true null hypothesis.
Beta is the probability of accepting a false null hypothesis.

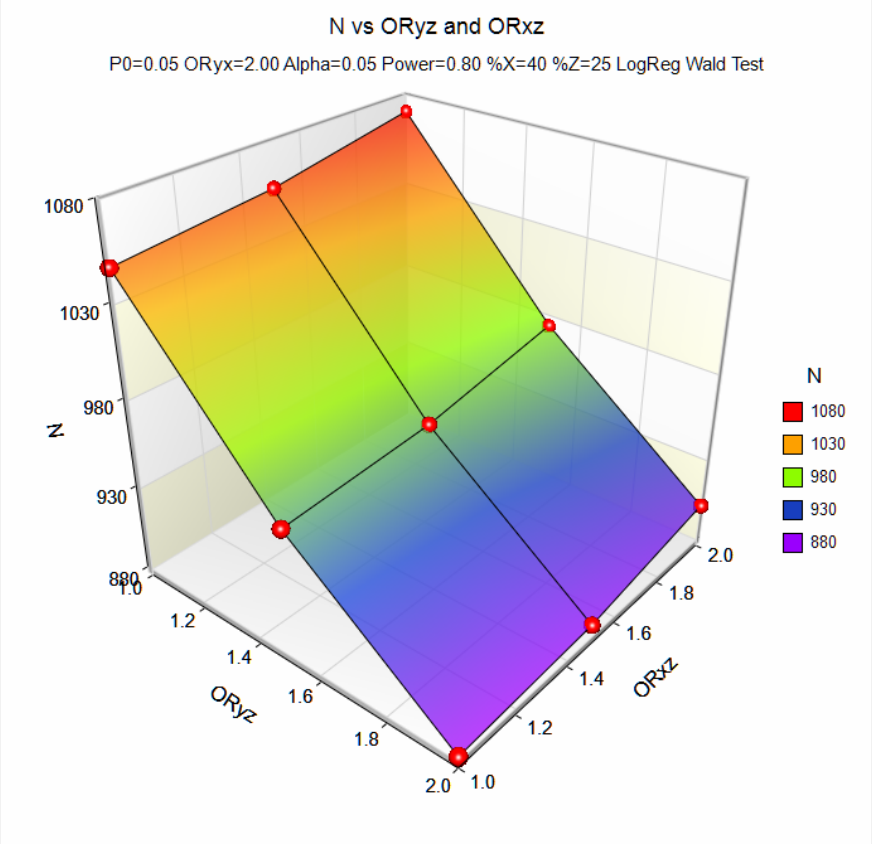
Summary Statements
A logistic regression of a binary response variable (Y) on a binary independent variable (X) and a binary confounder variable (Z) with a sample size of 1048 observations achieves 80% power at a 0.050 significance level to detect an odds ratio between Y and X of 2.000 . Other odds ratio 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%. Calculations assume that a two-sided Wald test is used.

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

Plot Section



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



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 problem.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for the Odds Ratio in Logistic Regression with Two Binary X's (Wald Test)** 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	**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

Numeric Results for Two-Sided Wald Test
Alternative Hypothesis: ORyx ≠ 1

Power	N	Percent X=1	Percent Z=1	P0	ORyx	ORyz	ORxz	Alpha	Beta
0.8005	545	25.0	50.0	0.100	2.000	2.000	1.000	0.050	0.1995

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.