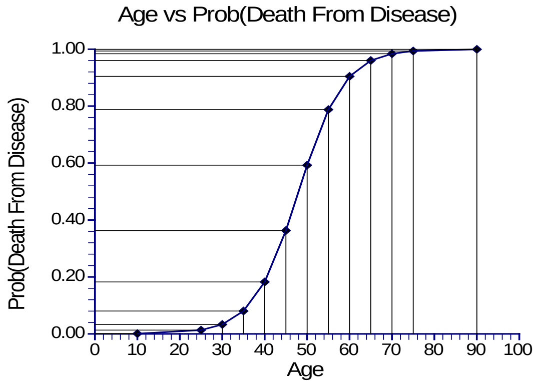
Chapter 872

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

Introduction

Logistic regression expresses the relationship between a binary response variable and one or more independent variables called *covariates*. A covariate can be discrete or continuous. This procedure deals with the specific case in which the covariate of interest is binary.

Consider a study of death from disease at various ages. This can be put in a logistic regression format as follows. Let a binary response variable Y be one if death has occurred and zero if not. Let X be the individual's age. Suppose a large group of various ages is followed for ten years and then both Y and X are recorded for each person. In order to study the pattern of death versus age, the age values are grouped into intervals and the proportions that have died in each age group are calculated. The results are displayed in the following plot.



As you would expect, as age increases, the proportion dying of disease increases. However, since the proportion dying is bounded below by zero and above by one, the relationship is approximated by an "S" shaped curve. Although a straight-line might be used to summarize the relationship between ages 40 and 60, it certainly could not be used for the young or the elderly.

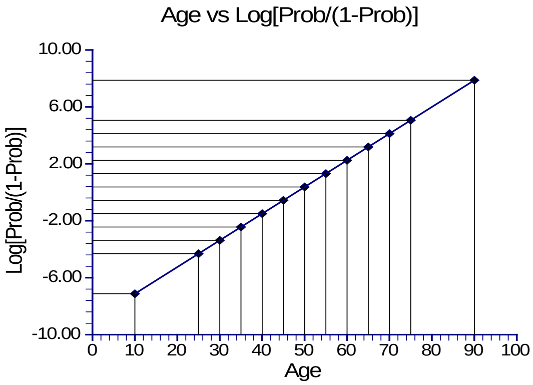
Under the logistic model, the proportion dying, P, at a given age can be calculated using the formula

$$P = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

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

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X$$

Note that the left side is the logarithm of the odds of death versus non-death and the right side is a linear equation for *X*. This is sometimes called the *logit* transformation of *P*. When the scale of the vertical axis of the plot is modified using the logit transformation, the following straight-line plot results.



In the logistic regression model, the influence of *X* on *Y* is measured by the value of the slope of *X* which we have called β_1. The hypothesis that $\beta_1 = 0$ versus the alternative that $\beta_1 = B \neq 0$ is of interest since if $\beta_1 = 0$, *X* is not related to *Y*.

Under the alternative hypothesis that $\beta_1 = B$, the logistic model becomes

$$\log\left(\frac{P_1}{1-P_1}\right) = \beta_0 + BX$$

Under the null hypothesis, this reduces to

$$\log\left(\frac{P_1}{1-P_1}\right) = \beta_0$$

To test whether the slope is zero at a given value of *X*, the difference between these two quantities is formed giving

$$\beta_0 + BX - \beta_0 = \log\left(\frac{P_1}{1-P_1}\right) - \log\left(\frac{P_0}{1-P_0}\right)$$

which reduces to

$$\begin{aligned} BX &= \log\left(\frac{P_1}{1-P_1}\right) - \log\left(\frac{P_0}{1-P_0}\right) \\ &= \log\left(\frac{P_1/(1-P_1)}{P_0/(1-P_0)}\right) \\ &= \log(OR) \end{aligned}$$

where *OR* is odds ratio of P_1 and P_0. This relationship may be solved for *OR* giving

$$OR = e^{BX}$$

This shows that the odds ratio of P_1 and P_0 is directly related to the slope of the logistic regression equation. It also shows that the value of the odds ratio depends on the value of *X*. For a given value of *X*, testing that *B* is zero is equivalent to testing *OR* is one. Since *OR* is commonly used and well understood, it is used as a measure of effect size in power analysis and sample size calculations.

Power Calculations

Suppose you want to test the null hypothesis that $\beta_1 = 0$ versus the alternative that $\beta_1 = B$. Hsieh, Block, and Larsen (1998) have presented formulae relating sample size, α, power, and *B* for two situations: when X_1 is normally distributed and when X_1 is binomially distributed.

When X_1 is binomially distributed and X_1 = 0 or 1, the sample size formula is

$$N = \frac{\left(z_{1-\alpha/2}\sqrt{\frac{\bar{P}(1-\bar{P})}{R}} + z_{1-\beta}\sqrt{P_0(1-P_0) + \frac{P_1(1-P_1)(1-R)}{R}}\right)^2}{(P_0 - P_1)^2(1-R)}$$

where P_0 is the event rate at $X_1 = 0$ and P_1 is the event rate at $X_1 = 1$, *R* is the proportion of the sample with $X_1 = 1$, and $\bar{P}$ is the overall event rate given by

$$\bar{P} = (1-R)P_0 + R(P_1).$$

Multiple Logistic Regression

The multiple logistic regression model relates the probability distribution of *Y* to two or more covariates $X_1, X_2, \cdots, X_k$ by the formula

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

where *P* is the probability that *Y* = 1 given the values of the covariates. It is a simple extension of the simple logistic regression model that was just presented. In power analysis and sample size work, attention is placed on a single covariate while the influence of the other covariates is statistically removed by placing them at their mean values.

When there are multiple covariates, the following adjustment was given by Hsieh (1998) to give the total sample size, N_m

$$N_m = \frac{N}{1-\rho^2}$$

where ρ is the multiple correlation coefficient between X_1 (the variable of interest) and the remaining covariates. Notice that the number of extra covariates does not matter in this approximation.

Ryan (2013) had some reservations with this approach. We refer you to page 163 of his sample size book for more details.

Example 1 – Finding Power for a Binary Covariate

A study is to be undertaken to study the relationship between post-traumatic stress disorder and gender. The event rate is thought to be 7% among males. The researchers want a sample size large enough to detect an odds ratio of 1.5 with 90% power at the 0.05 significance level with a two-sided test. They will eventually have five X's in their study. The R-squared of the remaining four variables is estimated to be 0.20.

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	**Power**
Alternative Hypothesis	**Two-Sided**
Alpha	**0.05**
N (Sample Size)	**20 50 100 200 300 500 700 1000 1200**
P0 (Baseline Probability that Y=1)	**0.07**
Use P1 or Odds Ratio	**Odds Ratio**
Odds Ratio (Odds1/Odds0)	**1.5 2**
R-Squared of X1 with Other X's	**0.2**
Percent of N with X1=1	**50**

Output

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

Numeric Reports

Numeric Results

Solve For: Power
Logistic Model: Log(P / (1 - P)) = B0 + B1*X1 + B2*X2 + ⋯ + Bk*Xk
Y: Binary Response
X1: Binary Independent Variable of Interest
X2, ⋯, Xk: Other Independent Variables (X's)
P: P = Pr(Y = 1)

			Probability that Y = 1			R-Squared	
	Sample	Percent				of X1 with	
	Size	of N with	Baseline	Alternative	Odds Ratio	Other X's	
Power	N	X1 = 1	P0	P1	OR	R^2	Alpha
0.0411	20	50	0.07	0.1014	1.5	0.2	0.05
0.0540	50	50	0.07	0.1014	1.5	0.2	0.05
0.0722	100	50	0.07	0.1014	1.5	0.2	0.05
0.1054	200	50	0.07	0.1014	1.5	0.2	0.05
0.1375	300	50	0.07	0.1014	1.5	0.2	0.05
0.2010	500	50	0.07	0.1014	1.5	0.2	0.05
0.2638	700	50	0.07	0.1014	1.5	0.2	0.05
0.3550	1000	50	0.07	0.1014	1.5	0.2	0.05
0.4129	1200	50	0.07	0.1014	1.5	0.2	0.05
0.0590	20	50	0.07	0.1308	2.0	0.2	0.05
0.0923	50	50	0.07	0.1308	2.0	0.2	0.05
0.1445	100	50	0.07	0.1308	2.0	0.2	0.05
0.2472	200	50	0.07	0.1308	2.0	0.2	0.05
0.3468	300	50	0.07	0.1308	2.0	0.2	0.05
0.5258	500	50	0.07	0.1308	2.0	0.2	0.05
0.6691	700	50	0.07	0.1308	2.0	0.2	0.05
0.8179	1000	50	0.07	0.1308	2.0	0.2	0.05
0.8814	1200	50	0.07	0.1308	2.0	0.2	0.05

Power The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N The size of the sample drawn from the population.
Percent of N with X1 = 1 The percentage of the population in which X1 = 1.
P0 Pr(Y = 1) when X1 = 0 and all other continuous covariates are set to their mean values.
P1 Pr(Y = 1) when X1 = 1.
OR Odds Ratio. OR = [P1 / (1 - P1)] / [P0 / (1 - P0)].
R^2 The R^2 achieved when X1 is regressed on X2, ⋯, Xk.
Alpha The probability of rejecting a true null hypothesis.

Summary Statements

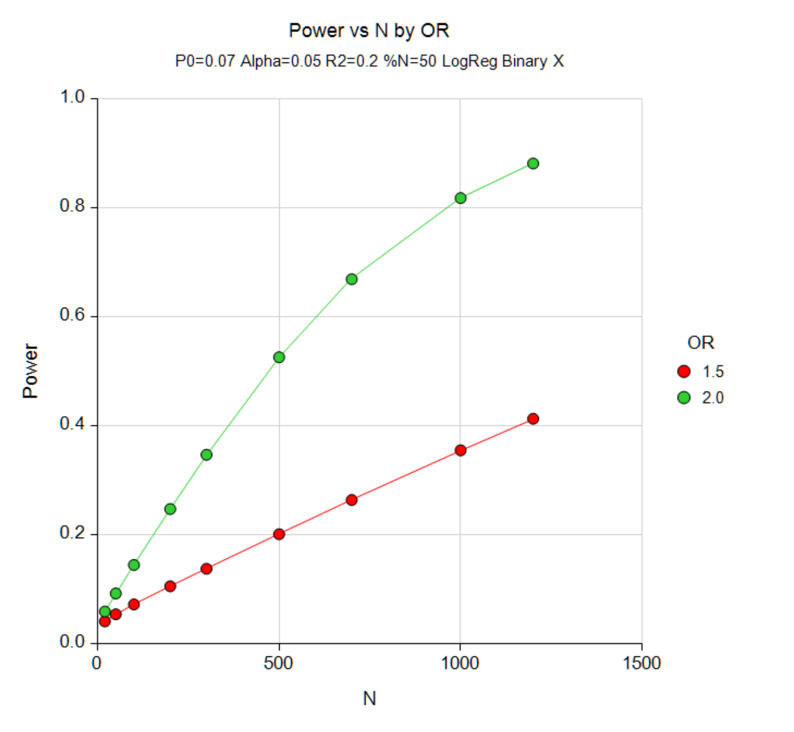
A logistic regression (binary response Y versus one binary X1 and other X's) design will be used to test whether the odds ratio (odds that Y = 1 when X1 is 1 to the odds that Y = 1 when X1 is 0) is different from 1. The comparison will be made using a two-sided logistic regression Wald test of B1 (using the model Log(P / (1 - P)) = B0 + B1*X1 + B2*X2 + ⋯ + Bk*Xk, where P = Pr(Y = 1)), with a Type I error rate (α) of 0.05. The test will use a baseline probability that Y = 1 (the probability that Y = 1 when X1 is 0 and all other X's are at their means, P0) of 0.07. Among subjects, 50% are assumed to have the value X = 1 (or be in the X = 1 group), and the remaining 50% are assumed to have the value X = 0. The R-squared of X1 with the other X's in the model is assumed to be 0.2. To detect an odds ratio (odds[X1 = 1] / odds[X1 = 0]) of 1.5 (or a P1 [probability that Y = 1 when X1 is 1] of 0.1014) with a sample size of 20, the power is 0.0411.

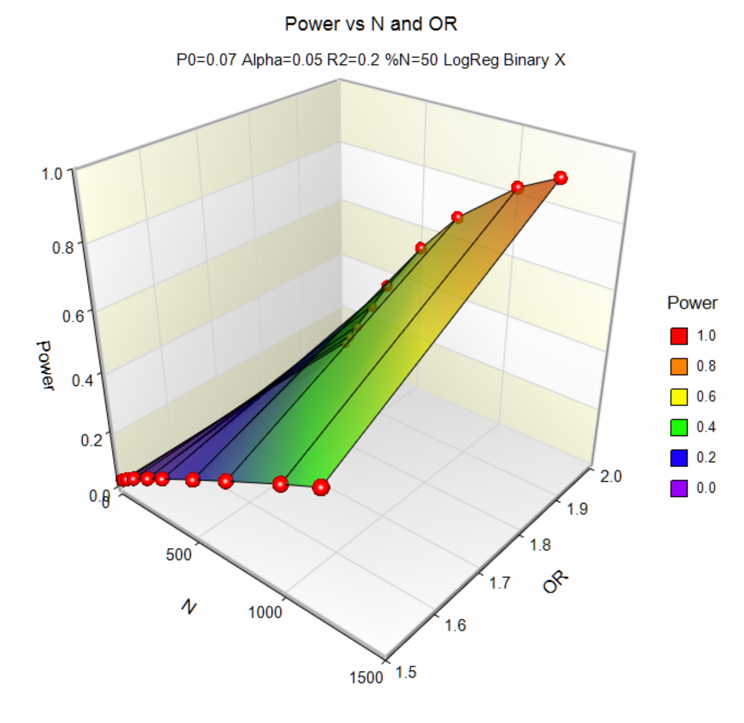
```
Dropout-Inflated Sample Size
─────────────────────────────────────────────────────────────────────────────────────────
                                    Dropout-
                                    Inflated      Expected
                                   Enrollment    Number of
                      Sample Size  Sample Size    Dropouts
Dropout Rate                    N           N'           D
───────────────────────────────────────────────────────────
20%                            20           25           5
20%                            50           63          13
20%                           100          125          25
20%                           200          250          50
20%                           300          375          75
20%                           500          625         125
20%                           700          875         175
20%                          1000         1250         250
20%                          1200         1500         300
─────────────────────────────────────────────────────────────────────────────────────────
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 (as entered by the user). 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. 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, 25 subjects should be enrolled to obtain a final sample size of 20 subjects.
─────────────────────────────────────────────────────────────────────────────────────────


References
─────────────────────────────────────────────────────────────────────────────────────────
Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. 'A Simple Method of Sample Size Calculation for Linear and
   Logistic Regression', Statistics in Medicine, Volume 17, pages 1623-1634.
─────────────────────────────────────────────────────────────────────────────────────────
```

This report shows the power for each of the scenarios.

Plots Section

Plots





These plots show the power versus the sample size for the two values of the odds ratio.

Example 2 – Finding Sample Size

Continuing with the previous study, determine the exact sample size necessary to attain a power of 90%.

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............................................................0.90
Alpha.............................................................0.05
P0 (Baseline Probability that Y=1) ................0.07
Use P1 or Odds Ratio...................................Odds Ratio
Odds Ratio (Odds1/Odds0) ...........................1.5 2
R-Squared of X1 with Other X's....................0.2
Percent of N with X1=1 .................................50
```

Output

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

```
Numeric Results

Solve For:          Sample Size
Logistic Model:     Log(P / (1 - P)) = B0 + B1*X1 + B2*X2 + ··· + Bk*Xk
Y:                  Binary Response
X1:                 Binary Independent Variable of Interest
X2, ···, Xk:        Other Independent Variables (X's)
P:                  P = Pr(Y = 1)

                                   Probability that Y = 1                   R-Squared
             Sample    Percent    ------------------------                 of X1 with
               Size   of N with    Baseline   Alternative   Odds Ratio    Other X's
Power             N     X1 = 1          P0            P1           OR           R²     Alpha

0.9000         4158         50        0.07        0.1014          1.5          0.2      0.05
0.8996         1276         50        0.07        0.1308          2.0          0.2      0.05
```

This report shows the power for each of the scenarios. The report shows that a power of 90% is achieved at a sample size of 1276 for an odds ratio of 2.0 and 4158 for an odds ratio of 1.5.

Example 3 – Validation for a Binary Covariate

Hsieh (1998) page 1626 gives the power as 95% when *N* = 1282 (equal sample sizes for both groups), alpha = 0.05 (two-sided), *P0* = 0.4, and the *P1* = 0.5. The prevalence of X1 is assumed to be 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 3** 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 .... **Power**
Alternative Hypothesis .... **Two-Sided**
Alpha.. **0.05**
N (Sample Size)... **1282**
P0 (Baseline Probability that Y=1) **0.4**
Use P1 or Odds Ratio **P1**
P1 (Alternative Probability that Y=1) **0.5**
R-Squared of X1 with Other X's **0**
Percent of N with X1=1 .... **50**

Output

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

```
Numeric Results
─────────────────────────────────────────────────────────────────────────────────────────
Solve For:        Power
Logistic Model:   Log(P / (1 - P)) = B0 + B1*X1 + B2*X2 + ··· + Bk*Xk
Y:                Binary Response
X1:               Binary Independent Variable of Interest
X2, ···, Xk:      Other Independent Variables (X's)
P:                P = Pr(Y = 1)
─────────────────────────────────────────────────────────────────────────────────────────
                                    Probability that Y = 1                R-Squared
              Sample    Percent   ───────────────────────                of X1 with
                Size  of N with     Baseline   Alternative  Odds Ratio   Other X's
Power              N     X1 = 1           P0            P1          OR          R²    Alpha
─────────────────────────────────────────────────────────────────────────────────────────
0.9502          1282         50          0.4           0.5         1.5           0     0.05
─────────────────────────────────────────────────────────────────────────────────────────
```

PASS calculates a power of 0.9502 which matches Hsieh (1998).