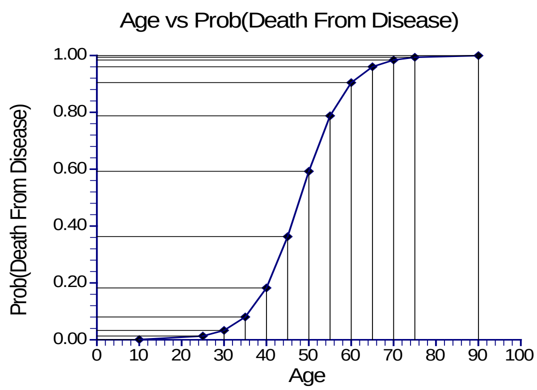
Chapter 859

Tests for the Odds Ratio in Logistic Regression with One Normal X (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, but in this procedure, the covariate is assumed to be normally distributed.

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 could be used to summarize the relationship between ages 40 and 60, it certainly could not be used for the young or the elderly.

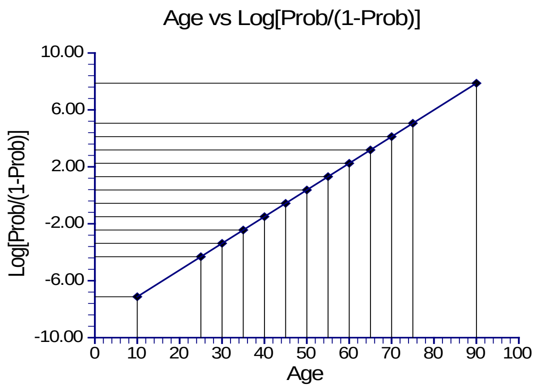
Under the logistic model, the proportion dying, *P*, at a given age can be calculated using the formula

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

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

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

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 that *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.

This procedure assumes that *X* is normally distributed. Without loss of generality, we assume that the mean of *X* is zero and the variance of *X* is one. We define X_0 to be the mean of *X* and X_1 to be the mean plus one standard deviation of *X*.

Power Calculations

We use the results of Novikov, Fund, and Freedman (2010) to compute sample size and power. This is a modification of the method of Hsieh, Block, and Larsen (1998) which is based on the Wald test. Note that their method is recommended in a simulation study by Bush (2015).

Suppose you want to test the null hypothesis that the odds ratio is one versus the alternative that it is some other positive value. Novikov *et al* (2010) presented formulae relating sample size, α, power, and odds ratio for the situation in which *X* is normally distributed and it is the only variable in the logistic regression model.

The sample size formula is

$$N = \left(\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 \frac{(\tau+\gamma)v_1}{\gamma(m_1-m_0)^2} + \frac{(\tau^2+\gamma^3)z_{1-\alpha/2}^2}{2\gamma(\tau+\gamma)^2}\right) / (1+\gamma)$$

where τ is the ratio of the variance of *X* (v_0) in the subgroup in which *Y* = 0 to the variance of *X* (v_1) in the subgroup of the population in which *Y* = 1, γ is the ratio of the size of the subgroup *Y* = 0 to the size of subgroup *Y* = 1, and m_0 and m_1 are the conditional means in the two subgroups defined by the values of Y.

Novikov *et al* (2010) implement this formula using an algorithm which they define that uses only the prevalence of *Y* (probability that *Y* = 1 in the population). They include SAS code for implementing their formula.

Errors in the SAS code of Novikov

As we implemented the Novikov *et al* formula, we found that their unnecessary use of the SAS *ceil(x)* command caused severe rounding errors in their Table IV. For example, their value for N on the first line of the table was 6552. We found that be removing the *ceil(x)* commands, our computed sample size of 6508 still gave power over 80%.

Findings of Bush (2015)

Bush (2015) conducted a simulation study of sample size estimation in logistic regression. He compared sample size formulas based on the Wald test, the likelihood ratio test, and the score test. He found that the "*power values are very similar, rarely deviating more than 2% between the three tests*." He found that the Novikov *et al* method did quite well. Hence, in this situation, it does not appear to matter upon which test the power is computed.

Example 1 – Power for a Continuous Covariate

A study is to be undertaken to study the relationship between post-traumatic stress disorder and heart rate after viewing video tapes containing violent sequences. Heart rate is assumed to be normally distributed. The event rate is thought to be 7% among soldiers. The researchers want a sample size large enough to detect an odds ratio of 1.5 or 2.0 with 90% power at the 0.05 significance level with a two-sided test. They decide to calculate the power at level sample sizes between 20 and 1200.

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).. **100 to 1300 by 100**
P1 (Prevalence of Y)..................................... **0.07**
Odds Ratio (Odds[x+σx] / Odds[x])................ **1.5 2**

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 X
Variables: Y = Binary Response, X = Continuous Covariate, P = Pr(Y = 1)

Power	Sample Size N	Number of Cases where Y = 1 Events	Prevalence of Y (Proportion where Y = 1) Overall P1	for X = μx P1(μx)	Odds Ratio	Alpha	Logit Model Intercept B0	Slope B1
0.1443	100	7	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.2764	200	14	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.4015	300	21	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.5145	400	28	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.6125	500	35	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.6951	600	42	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.7631	700	49	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.8179	800	56	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.8613	900	63	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.8954	1000	70	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.9216	1100	77	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.9417	1200	84	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.9570	1300	91	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.3309	100	7	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.6384	200	14	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.8257	300	21	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.9224	400	28	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.9674	500	35	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.9869	600	42	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.9950	700	49	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.9981	800	56	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.9993	900	63	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.9998	1000	70	0.07	0.0580	2.0	0.05	-2.7867	0.6931
0.9999	1100	77	0.07	0.0580	2.0	0.05	-2.7867	0.6931
1.0000	1200	84	0.07	0.0580	2.0	0.05	-2.7867	0.6931
1.0000	1300	91	0.07	0.0580	2.0	0.05	-2.7867	0.6931

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.
Events The expected number of cases in which Y = 1.
P1 The overall proportion of the population in which Y = 1.
P1(μx) The proportion of the population in which Y = 1 if X = μx (mean of X).
Odds Ratio Defined as Odds Ratio = odds(μx + σx) / odds(μx) = [P1(μx + σx) / (1 - P1(μx + σx))] / [P1(μx) / (1 - P1(μx))].
Alpha The probability of rejecting a true null hypothesis.
B0 The intercept in the logit model, log(P/(1 - P)) = B0 + B1 X.
B1 The slope in the logit model, log(P/(1 - P)) = B0 + B1 X.

Summary Statements

A logistic regression (binary response Y versus one normally distributed X) design will be used to test whether the odds ratio (odds that Y = 1 when X is one standard deviation above its mean to the odds that Y = 1 when X is equal to its mean) is different from 1. The comparison will be made using a two-sided logistic regression test of B1 (using the model Log(P / (1 - P)) = B0 + B1 X) with a Type I error rate (α) of 0.05. The prevalence of Y (probability that Y = 1) in the population is assumed to be 0.07. To detect an odds ratio [odds(μx + σx) / odds(μx)] of 1.5 with a sample size of 100 subjects, the power is 0.1443.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	100	125	25
20%	200	250	50
20%	300	375	75
20%	400	500	100
20%	500	625	125
20%	600	750	150
20%	700	875	175
20%	800	1000	200
20%	900	1125	225
20%	1000	1250	250
20%	1100	1375	275
20%	1200	1500	300
20%	1300	1625	325

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, 125 subjects should be enrolled to obtain a final sample size of 100 subjects.

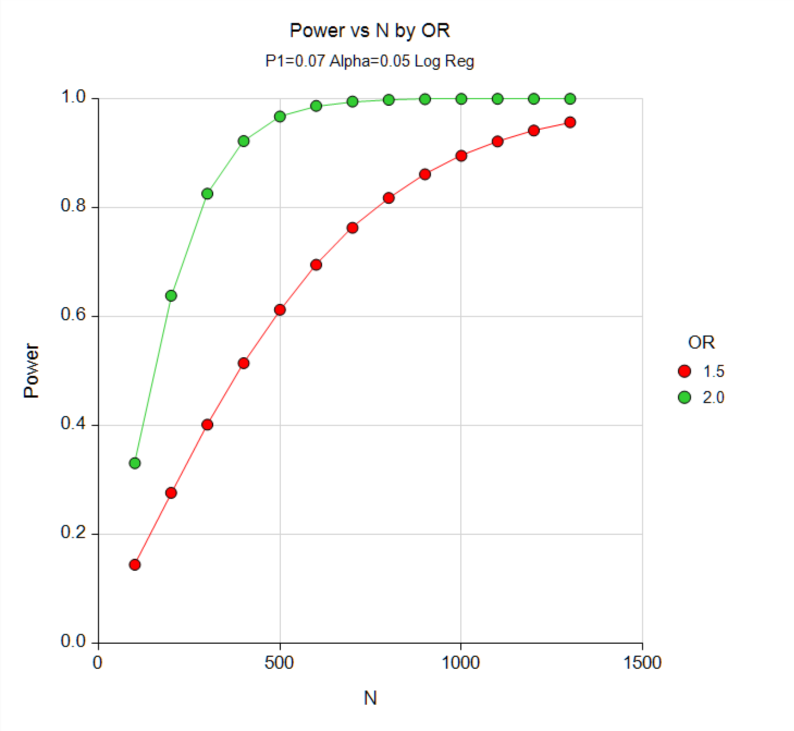
References

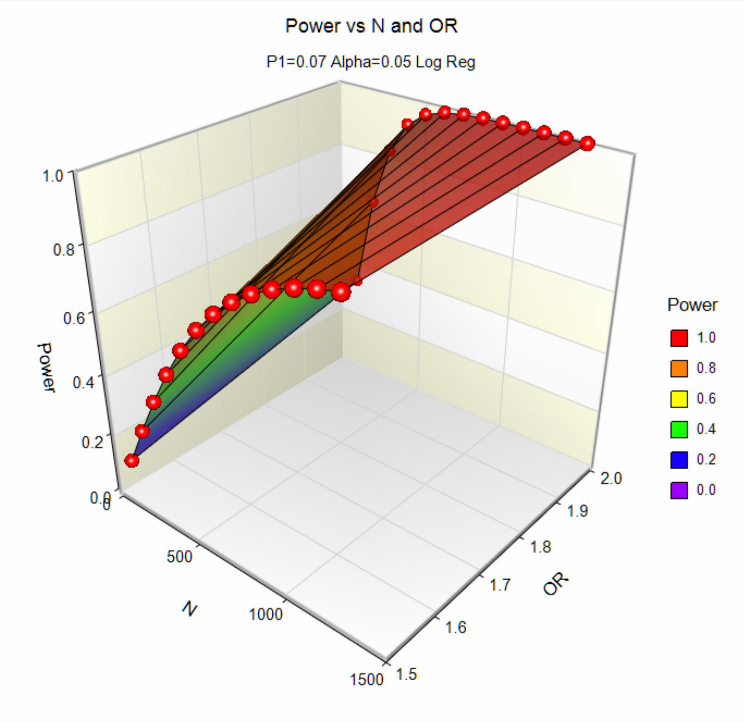
Novikov, N., Fund, N., Freedman, L.S. 2010. 'A modified approach to estimating sample size for simple logistic regression with one continuous covariate', Statistics in Medicine, Volume 29(1), pages 97-107.

This report shows the power for each of the scenarios. The report shows that a power of 90% is reached at a sample size of about 400 for an odds ratio of 2.0 and 1000 for an odds ratio of 1.5.

Plots Section

Plots





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

Example 2 – Sample Size for a Normal Covariate

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**
P1 (Prevalence of Y)..................................... **0.07**
Odds Ratio (Odds[x+σx] / Odds[x])................ **1.5 2**

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 X
Variables: Y = Binary Response, X = Continuous Covariate, P = Pr(Y = 1)

		Number of Cases where	Prevalence of Y (Proportion where Y = 1)				Logit Model	
Power	Sample Size N	Y = 1 Events	Overall P1	for X = μx P1(μx)	Odds Ratio	Alpha	Intercept B0	Slope B1
0.9000	1016	71	0.07	0.0656	1.5	0.05	-2.6566	0.4055
0.9003	370	26	0.07	0.0580	2.0	0.05	-2.7867	0.6931

This report shows the necessary sample size for each odds ratio.

Example 3 – Validation for a Normal Covariate

Novikov (2010) page 102, Table IV, first line, gives sample size as 6551 when alpha = 0.05 (two-sided), power = 0.8, *P1* = 0.02, and the odds ratio is log(0.25) = 1.28402542. As we discussed above, the sample size of 6551 is high because of an error in their SAS code. The sample size should be 6508, which we found using manual calculation.

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	**Sample Size**
Alternative Hypothesis	**Two-Sided**
Power	**0.8**
Alpha	**0.05**
P1 (Prevalence of Y)	**0.02**
Odds Ratio (Odds[x+σx] / Odds[x])	**1.28402542**

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 X
Variables: Y = Binary Response, X = Continuous Covariate, P = Pr(Y = 1)

Power	Sample Size N	Number of Cases where Y = 1 Events	Prevalence of Y (Proportion where Y = 1) Overall P1	Prevalence of Y (Proportion where Y = 1) for X = μx P1(μx)	Odds Ratio	Alpha	Logit Model Intercept B0	Logit Model Slope B1
0.8	6508	130	0.02	0.0194	1.284	0.05	-3.9218	0.25

This report achieves an N of 6508 which validates this procedure.