1 Binary Logistic Regression

Binary Logistic regression is used to determine the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories.

1.1 Likelihood Ratio Test

The likelihood ratio test is a test of the difference between 2LL for the full model with predictors and 2LL for initial chi-square in the null model. When probability fails to reach the 5% significance level, we retain the null hypothesis that knowing the independent variables (predictors) has no increased effects (i.e. make no difference) in predicting the dependent.

1.2 Psuedo R Squared Values

1.3 The Hosmer-Lemeshow Test

The Hosmer-Lemeshow test of goodness of fit is not automatically a part of the SPSS logistic regression output. To get this output, we need to go into ‘`options` and tick the box marked Hosmer-Lemeshow test of goodness of fit. In our example, this gives us the following output:

Step	Chi-square	df	Sig.
1	142.032	6	.000

Therefore, our model is significant, suggesting it does not fit the data. However, as we have a sample size of over 13,000, even very small divergencies of the model from the data would be flagged up and cause significance. Therefore, with samples of this size it is hard to find models that are parsimonious (i.e. that use the minimum amount of independent variables to explain the dependent variable) and fit the data. Therefore, other fit indices might be more appropriate.

2 Logistic Regression

Logistic regression, also called a logit model, is used to model **dichotomous outcome** variables. In the logit model the **log odds** of the outcome is modeled as a linear combination of the predictor variables.

In logistic regression theory, the predicted dependent variable is a function of the probability that a particular subject will be in one of the categories (for example, the probability that a patient has the disease, given his or her set of scores on the predictor variables).

2.1 Introduction to the Odds Ratio

Let's begin with probability. Let's say that the probability of success is 0.8, thus p = 0.8 Then the probability of failure is

$$q = 1 - p = 0.2$$

The odds of success are defined as

$$odds(success) = p/q = 0.8/0.2 = 4$$

that is, the odds of success are 4 to 1. The odds of failure would be

$$odds(failure) = q/p = .2/.8 = .25$$

This looks a little strange but it is really saying that the odds of failure are 1 to 4. The odds of success and the odds of failure are just reciprocals of one another, i.e., 1/4 = .25 and 1/.25 = 4.

Next, we will add another variable to the equation so that we can compute an odds ratio.

Another example

Suppose that seven out of 10 males are admitted to an engineering school while three of 10 females are admitted.

- The probabilities for admitting a male are, p = 7/10 = .7 (q = 1 - .7 = .3)
- Here are the same probabilities for females, p = 3/10 = .3 (q = 1 - .3 = .7)

Now we can use the probabilities to compute the admission odds for both males and females,

- odds(male) = .7/.3 = 2.33333
- odds(female) = .3/.7 = .42857

Next, we compute the odds ratio for admission,

$$OR = 2.3333/0.42857 = 5.44$$

Thus, for a male, the odds of being admitted are 5.44 times as large than the odds for a female being admitted.

2.2 About logits

There is a direct relationship between the coefficients produced by **logit** and the odds ratios produced by the logistic procedure. First, let's define what is meant by a logit: A logit is defined as the log base e (log) of the odds,

$$logit(p) = log(odds) = log(p/q)$$

Logistic regression is in reality ordinary regression using the logit as the response variable,

$$logit(p) = a + bX$$
$$log(p/q) = a + bX$$

This means that the coefficients in logistic regression are in terms of the log odds, that is, the coefficient 1.694596 implies that a one unit change in gender results in a 1.694596 unit change in the log of the odds.

Equation [3] can be expressed in odds by getting rid of the log. This is done by taking e to the power for both sides of the equation.

$$p/q = e^{a+bX}$$

The end result of all the mathematical manipulations is that the odds ratio can be computed by raising e to the power of the logistic coefficient,

$$OR = e^b = e^1.694596 = 5.44$$

2.3 Logistic Regression: Odds Ratio

What are odds? The odds of outcome 1 versus outcome 2 are the probability (or frequency) of outcome 1 divided by the probability (or frequency) of outcome 2.

$$\hat{Y} = \frac{\text{Odds}}{1 + \text{Odds}}$$

Loglinear models essentially define a pattern of odds ratios, apply the marginals to them, and compare the resulting table with the observed table, in pretty much the same way we apply the Pearson χ^2 test for association. The big difference is the pattern we define can be much more complicated than independence.

3 The Wald Test

The Wald test is a way of testing the significance of particular explanatory variables in a statistical model.

In logistic regression we have a binary outcome variable and one or more explanatory variables. For each explanatory variable in the model there will be an associated parameter. The Wald test is one of a number of ways of testing whether the parameters associated with a group of explanatory variables are zero.

If for a particular explanatory variable, or group of explanatory variables, the Wald test is significant, then we would conclude that the parameters associated with these variables are not zero, so that the variables should be included in the model. If the Wald test is not significant then these explanatory variables can be omitted from the model.

When considering a single explanatory variable, Altman (1991) uses a t-test to check whether the parameter is significant. For a single parameter the Wald statistic is just the square of the t-statistic and so will give exactly equivalent results.

An alternative and widely used approach to testing the signicance of a number of explanatory variables is to use the likelihood ratio test. This is appropriate for a variety of types of statistical models.

Agresti (1990) argues that the likelihood ratio test is better, particularly if the sample size is small or the parameters are large.

3.1 Logistic Regression

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta_1 x)}+1} = \frac{1}{e^{-(\beta_0+\beta_1 x)}+1},$$

and

$$g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x,$$

and

$$\frac{\pi(x)}{1-\pi(x)} = e^{(\beta_0+\beta_1 x)}.$$

3.2 Wald statistic

Alternatively, when assessing the contribution of individual predictors in a given model, one may examine the significance of the Wald statistic. The Wald statistic, analogous to the t-test in linear regression, is used to assess the significance of coefficients. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution.

3.3 Pseudo-R Squared

Cox and Snell R Square and Nagelkerke R Square - These are pseudo R-squares. Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many people have tried to come up with one.

There are a wide variety of pseudo-R-square statistics (these are only two of them). Because this statistic does not mean what R-squared means in OLS regression (the proportion of variance explained by the predictors), we suggest interpreting this statistic with great caution.

3.4 Logistic Regression: Decision Rule

Our decision rule will take the following form: If the probability of the event is greater than or equal to some threshold, we shall predict that the event will take place. By default, SPSS sets this threshold to .5. While that seems reasonable, in many cases we may want to set it higher or lower than .5.

4 Multinomial Logistic Regression

Examples of multinomial logistic regression

Example 1. People's occupational choices might be influenced by their parents' occupations and their own education level. We can study the relationship of one's occupation choice with education level and father's occupation. The occupational choices will be the outcome variable which consists of categories of occupations.

Example 2. A biologist may be interested in food choices that alligators make. Adult alligators might have difference preference than young ones. The outcome variable here will be the types of food, and the predictor variables might be the length of the alligators and other environmental variables.

Example 3. Entering high school students make program choices among general program, vocational program and academic program. Their choice might be modeled using their writing score and their social economic status.

4.1 Classification Plot

The classification plot or histogram of predicted probabilities provides a visual demonstration of the correct and incorrect predictions. Also called the classplot or the plot of observed groups and predicted probabilities,it is another very useful piece of information from the SPSS output when one chooses Classification plots under the Options button in the Logistic Regression dialogue box.