CSCU Cornell Statistical Consulting Unit

Binary Logistic Regression Models and Statistical Software: What You Need to Know

Jay Barry, Haim Bar, Simon Robert, Françoise Vermeylen and Jing Yang

1 Binary Logistic Regression

Binary logistic regression is used to model the relationship between a categorical response variable and one or more explanatory variables that may be continuous or categorical. Statistical software can be used to easily obtain the estimates for such a model; however the conclusions drawn from the estimates can be completely erroneous if the researcher is not careful in checking the specification of the software that has been used. Although this is true for any type of model, logistic regression models tend to be more confusing as more coding options, taken individually or in combinations are available and they vary more between software packages. More specifically, to fit logistic regression models it is important to know: (1) the level of the dependent variable being modeled (2) the default coding method used for independent categorical variables, and (3) the reference level used for the categorical independent variables.

Regardless of how the binary response variable is being coded (e.g. 0/1 or 1/2), when estimating a logistic regression model we need to pay close attention to how the binary response is being modeled. Is the software modeling the probability of the highest response level or the lowest? Even within a specific software package the coding scheme of the response might depend on the module being used. Most statistical software packages offer several different modules to estimate a logistic regression model. As logistic regression belongs to the class of Generalized Linear Models, the main choices are often using a Generalized Linear Model module (GLM, not to be confused with General Linear Model) or a logistic regression module. In JMP the response level modeled is always the lowest level (i.e. 0 in a 0/1 coding scheme), while in R and Stata the response level modeled is the highest level. In SPSS by default the response level modeled is the highest with the LOGISTIC command, and the lowest with the GLM command. In SAS the response level modeled is by default the lowest both in PROC LOGISTIC and in PROC GENMOD which fits Generalized Linear Models. With some software there are options enabling the user to change this default such as using a “descending” option when using GLM in both SAS and SPSS. Unfortunately this default can also sometimes be reversed unintentionally for example when using ‘formats’ in SAS.

To incorporate categorical predictors into a regression model two main coding schemes are used: dummy coding and effect coding (see StatNews #72 for more details). Which coding scheme is used by default, again, not only depends on the software but also on the specific module within the software. In addition to differences in the type of coding used for categorical predictors, how the reference level is determined is also software dependent. For logistic regression Stata, SPSS and R, use dummy coding, whereas JMP uses effect coding by default. The default in SPSS is to code the last level (i.e. highest in alphanumeric order) of the categorical variable as the reference level; in Stata and R it is the first level.

In JMP by default it is the last level that will be the reference level and which will thus be coded -1 for each effect variable. In SAS the default coding will again depend on the specific procedure used. PROC LOGISTIC uses effect coding, with the last level being the reference and thus coded -1. The default coding in PROC GENMOD is dummy coding and the last level of the categorical variable is considered as the reference level. The user can change the default specifications mentioned above in most software packages. For example, when using PROC LOGISTIC with the param=glm option, SAS uses dummy coding and the last level of the categorical variable is considered as the reference level.

The example below illustrates the diversity of the results obtained across the various statistical software packages and summarizes the default coding used for both the dependent and the categorical independent variables of the most commonly used software packages.

Besides checking the manuals of the software, fitting a simple logistic regression using only one categorical independent variable and comparing the output obtained with a cross tabulation will help in understanding the coding schemes used by the software.

2 Interpreting regression coefficients in logistic regression models – example

To understand how different statistical packages report the estimates of parameters in logistic models, we use an example with "contraceptive data" [1].

The data contains a sample of 1,607 women and their use of contraceptives, as well as their age, education, and if they wanted more children. Age was coded in four categories (1 through 4) with 1: <25; 2: 25-29; 3: 30-39 and 4: 40-49. Their education was coded either L or H to represent respectively a low or high level of education. If they wanted more children was coded either yes or no.

In our logistic regression model, the dependent variable is the use of contraception (0=no use; 1=use) and the independent variables are age, education and if they want more children or not. Our goal here is to show how the results are reported when different software packages are used. To that end, we use R, JMP, SPSS, SAS and Stata. Table 2.1 contains the parameter estimates.

Table 2.1: Regression coefficients for the contraceptive data

	R or STATA	JMP or SASlog	SASglm or SPSSglm	SPSSlog
Response level: Event =	1	0	0	1
Intercept	-0.808	0.765	0.777	-0.777
Age= 1 (<25)	NA^{r}	0.622	1.189	-1.189

[1] Reference: Little, R. J. A. (1978). Generalized Linear Models for Cross-Classified Data from the WFS. World Fertility Survey Technical Bulletins, Number 5.

	R or STATA	JMP or SASlog	SASglm or SPSSglm	SPSSlog
Age= 2 (25-29)	0.389	0.232	0.800	-0.800
Age= 3 (30-39)	0.909	-0.287	0.281	-0.281
Age= 4 (40-49)	1.189	NA^j	0	NA^r
Edu=H	NA^r	-0.163	-0.325	0.325
Edu=L	-0.325	NA^j	0	NA^r
Wmo=yes	-0.833	NA^j	0	NA^r
Wmo=no	NA^r	-0.417	-0.833	0.833

Notes:

1. SPSSlog and SASlog refer to the logistic procedure of the respective software packages whereas SPSSglm and SASglm refer to the generalized linear model procedure.
2. NA^r means "dummy-coding" was used and that the estimate is not provided because this is the reference level (where the effect is assumed to be 0).

3. NA^j means that "effect-coding" is used. That is, the missing level is inferred by the estimates of the other levels. Specifically, it is minus the sum of the estimates of the other levels. For example, the effect of Age=4 (40-49) in JMP or SASlog is: $-(0.622 + 0.232 - 0.287) = -0.567$. Similarly, the effect of Edu=L is 0.163. SAS G and SPSS G report the redundant levels explicitly (as 0).

4. When "effect-coding" is used, the estimates are reported relative to the overall mean, and not relative to a reference level (as we do in "dummy coding").

Created December 2011. Last updated April 2022.