The Stata Journal (2004)
4, Number 4, pp. 436–441

Boolean logit and probit in Stata

Bear F. Braumoeller
Harvard University

Abstract. This paper introduces new statistical models, Boolean logit and probit, that allow researchers to model binary outcomes as the results of Boolean interactions among independent causal processes. Each process (or "causal path") is modeled as the unobserved outcome in a standard logit or probit equation, and the dependent variable is modeled as the observed product of their Boolean interaction. Up to five causal paths can be modeled, in any combination—A and B and C produce Y, A and (B or [C and D]) produce Y, etc.

Keywords: st0078, mlboolean, dichotomous dependent variable, Boolean, logit, probit, multiple causal paths, complexity, random utility

1 Overview

The impact of independent probabilistic causal processes is often thought to cumulate in a manner consistent with Boolean logic—i.e., diet *or* heredity can lead to heart failure; apathy *or* indifference can lead to nonvoting; plant survival depends on light, water, *and* proper soil; and so on (see Ragin 1987). This cumulation is nonadditive: someone with a 50% chance of having a heart attack based on diet alone and a 50% chance of having a heart attack based on heredity will not have a 100% chance of having a heart attack but rather a $1-(1-.5)\times(1-.5)=75\%$ chance, assuming that the probabilities are independent. Standard additive statistical models fail to capture the form of this interaction.

The Boolean logit and probit procedures, introduced in Braumoeller (2003) and implemented in the Stata package `mlboolean`, are designed for use in such situations. The probability calculus for a Boolean interaction is straightforward: if the combination of A and B causes Y, then $p_Y = p_A \times p_B$. If either A or B causes Y, as in the case of diet and heredity described above, then $p_Y = 1-\{(1-p_A)\times(1-p_B)\}$. The `mlboolean` procedure models the probabilities of the unobserved, or latent, dependent variables (probability of heart attack resulting from diet and probability of heart attack resulting from heredity, in the above example) as ordinary logit and probit curves, constructs a likelihood function based on the posited logic of their interaction and the observed dependent variable (heart attack), and maximizes to obtain coefficient estimates.

2 Syntax

`mlboolean` *link n* `(`*calculus*`)` `(`*depvar*`)` `(`*indepvars*$_1$`)` `...(`*indepvars*$_n$`)`
[`ystar(`*stub*`),` *mloptions*]

where *link* specifies the link function (logit or probit), n is the number of "causal paths" (i.e., latent dependent variables; at present $n \leq 5$), *calculus* contains the probability calculus—e.g., `(aorb)`, `(aand(borcord))`, etc., in which `a` corresponds to the first latent dependent variable, `b` corresponds to the second, and so forth—and $indepvars_1$–$indepvars_n$ contain the vectors of independent variables that predict the n latent dependent variables. Note that there is nothing preventing the user from including the same independent variable in more than one of the *indepvar*s; indeed, if the probabilities of the events modeled by the n latent logits or probits are not independent because they both depend on the variable in question, doing so is recommended. (Note, however, that if two lists contain exactly the same independent variables, nothing but the functional form exists to identify the model, as in [R] **heckman**, and results should be treated accordingly.)

3 Options

In the process of producing coefficient estimates, `mlboolean` also estimates the values of the latent dependent variables that are combined in the probability calculus. The names of these variables can be set with the `ystar()` option; the default is `ystar(ystar)`. In addition, `mlboolean` utilizes `ml`, to which it passes optional commands directly. See [R] **ml** for more details.

4 Postestimation commands

`mlboolfirst` calculates and graphs predicted values for a given independent variable. All other variables are held at their mean values, except for variables of type `int`, which are set to their modal values. Similarly, a curve is graphed if the independent variable is continuous, but points are graphed if it is type `int`. (To manipulate which variables are flagged as integers, use `recast`.) The syntax is

`mlboolfirst` *indepvar*

5 Variables created

For each observation, `mlboolean` estimates a predicted probability that each of the antecedent events will occur (that is, a predicted probability for each of the latent dependent variables) and then an overall predicted probability that the event measured by *depvar* will occur. These probabilities are saved in `ystar_a`, `ystar_b`, ..., and `boolpred`, respectively. The former variables can be named something other than "ystar" with the `ystar()` option.

6 Example

Assume, for the sake of discussion, that we want to understand why many American citizens wanted President Clinton to remain in office after the Monica Lewinsky scandal, even if he were found guilty of obstructing justice by encouraging Lewinsky to lie under oath. Impressionistic accounts at the time suggested that some Americans thought that Clinton should leave office because, regardless of the present accusation, he lacked the moral fiber to be president. Among those who found his moral values acceptable, there were three specific reasons given for not wanting him to leave office, even if he were found guilty: that he did not in fact encourage her to lie; that the issue was simply not an important one; or that, regardless of its importance, it was a private matter rather than one that should concern the public. Respondents asserting that Clinton should stay should, therefore, find his moral fiber acceptable *and* agree with at least one of the three justifications just outlined.

The latent variables in this case are (a) the probability that Clinton's moral values would be deemed acceptable, (b) the probability that Clinton would be excused because he did not encourage Lewinsky to lie, (c) the probability that Clinton would be excused because the issue was unimportant, and (d) the probability that Clinton would be excused because the matter was a private one. The logic sketched out in the previous paragraph suggests that (a) and [(b) or (c) or (d)] will produce $Y = 1$ (an evaluation that the president should stay in office).

The latent probabilities are estimated using data from CBS News Monthly Poll #1, from February 1998: responses to a question about whether Clinton should stay in office if he were to be found guilty of obstructing justice (`bcstay`) constitute the observed dependent variable, and responses to questions about whether Clinton shares the values that most Americans try to live by (`moralval`), whether the respondent thinks that Clinton encouraged Lewinsky to lie (`bcmllie`), whether the situation is of great rather than little importance to the nation (`doUcare`), and whether the issue is primarily public rather than private (`pubpriv`) predict (a)–(d), respectively. (Survey weights are contained in the variable `weight`.) Although in this case each of the latent dependent variables is predicted by a single independent variable, it is worth emphasizing that this need not be true more generally: if the survey contained more questions bearing on Clinton's moral values, for example, they could be added to the list of variables predicting (a).

`moralval` is worded in such a way that an affirmative answer should be positively related to (a); the remaining questions are worded in such a way that an affirmative answer should be negatively related to (b)–(d). The argument, therefore, suggests a probability calculus of `(aand(borcord))`, a positive coefficient for the variable predicting (a), and negative coefficients for the variables predicting (b)–(d). The results bear out these expectations.

```
. svyset [pweight=weight]
. mlboolean probit 4 (aand(borcord)) (bcstay) (moralval) (bcmllie) (doUcare)
> (pubpriv), svy
Boolean Probit Estimates
pweight:  weight                                  Number of obs    =       239
Strata:   <one>                                   Number of strata =         1
PSU:      <observations>                          Number of PSUs   =       239
                                                  Population size  =    227911
  F(   1,    238)  =     11.22
  Prob > F         =    0.0009
```

bcstay	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Path1						
moralval	.8753194	.2613543	3.35	0.001	.3604562	1.390183
_cons	.1067626	.1688836	0.63	0.528	-.225935	.4394602
Path2						
bcmllie	-6.182273	.6797094	-9.10	0.000	-7.521288	-4.843258
_cons	5.267284	.1237408	42.57	0.000	5.023517	5.511051
Path3						
doUcare	-5.724847	.8064377	-7.10	0.000	-7.313515	-4.13618
_cons	17.2037	3.114909	5.52	0.000	11.06739	23.34001
Path4						
pubpriv	-6.068586	1.571714	-3.86	0.000	-9.164833	-2.97234
_cons	12.97357	4.720715	2.75	0.006	3.673853	22.2733

```
(367 missing values generated)
Correctly predicted 182 of 239 cases, or 76.15%.
```

As anticipated, belief that Clinton shares most Americans' moral values, in conjunction with either a belief that he did not lie, a belief that the issue lacks importance, or a belief that the issue is a private rather than a public one, are related to the conclusion that Clinton should stay on even if found guilty of obstructing justice.

7 Methods and Formulas

Because the likelihood function for a given analysis is derived directly from the probability calculus, formulas vary depending on which probability calculus is chosen. For example, in the case of conjunctural ("and") causes,

$$
\begin{array}{ccccccc}
p_Y & = & p_A & \times & p_B & \times & \dots \\
\downarrow & & \downarrow & & \downarrow & & \\
\Pr(y_i = 1 \mid \beta, x_i) & = & y^*_{iA} & \times & y^*_{iB} & \times & \dots \\
\downarrow & & \downarrow & & \downarrow & & \\
\Pr(y_i = 1 \mid \beta, x_i) & = & \Phi(\boldsymbol{\beta_1 x_{i1}}) & \times & \Phi(\boldsymbol{\beta_2 x_{i2}}) & \times & \dots
\end{array}
$$

$$= \prod_{j=1}^{J} \left\{ \Phi(\boldsymbol{\beta_j x_{ij}}) \right\}$$

$$L(Y \mid \beta, X) = \prod_{i=1}^{N} \left[\prod_{j=1}^{J} \left\{ \Phi(\boldsymbol{\beta_j x_{ij}}) \right\} \right]^{y_i} \left[1 - \prod_{j=1}^{J} \left\{ \Phi(\boldsymbol{\beta_j x_{ij}}) \right\} \right]^{1-y_i}$$

In the case of substitutable ("or") causes, a different likelihood function is implied:

$$\begin{array}{rcccccc}
p_Y & = & 1 & - & \{(1-p_A) & \times & (1-p_B) \quad \times \quad \ldots\} \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
\Pr(y_i = 1 \mid \beta, x_i) & = & 1 & - & \{(1-y^*_{iA}) & \times & (1-y^*_{iB}) \quad \times \quad \ldots\} \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
\Pr(y_i = 1 \mid \beta, x_i) & = & 1 & - & \left[\left\{ 1 - \Phi(\boldsymbol{\beta_1 x_{i1}}) \right\} \right. & \times & \left. \left\{ 1 - \Phi(\boldsymbol{\beta_2 x_{i2}}) \right\} \quad \times \quad \ldots \right]
\end{array}$$

$$= 1 - \prod_{j=1}^{J} \left\{ 1 - \Phi(\boldsymbol{\beta_j x_{ij}}) \right\}$$

$$L(Y \mid \beta, X) = \prod_{i=1}^{N} \left[1 - \prod_{j=1}^{J} \left\{ 1 - \Phi(\boldsymbol{\beta_j x_{ij}}) \right\} \right]^{y_i} \left[\prod_{j=1}^{J} \left\{ 1 - \Phi(\boldsymbol{\beta_j x_{ij}}) \right\} \right]^{1-y_i}$$

and so forth.

8 Acknowledgments

Thanks are due to the Weatherhead Center for International Affairs at Harvard University, which provided the research funds necessary for the creation of this program; to Jacob Kline for research assistance; to Ronán Conroy for his assistance in flagging and zapping bugs; and to an anonymous reviewer for comments. A current copy of all relevant files can be found at the *Stata Journal* web site or at the author's home page, *http://www.people.fas.harvard.edu/~bfbraum*.

9 References

Braumoeller, B. F. 2003. Causal complexity and the study of politics. *Political Analysis* 11(3): 209–233.

CBS News. 1998. CBS News Monthly Poll #1, February 1998 [Computer file]. ICPSR version. New York: CBS News [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

Ragin, C. C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

About the Author

Bear F. Braumoeller is an associate professor in the Department of Government, Harvard University.