An Introduction to Logistic Regression

Nuts and Bolts

Why use logistic regression?

There are many important research topics for which the dependent variable is "limited" (discrete not continuous). Researchers often want to analyze whether some event occurred or not, such as voting, participation in a public program, business success or failure, morbidity, mortality, a hurricane and etc.

Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable (coded 0, 1).

A data set appropriate for logistic regression might look like this:

Descriptive Statistics					
Variable	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
YES	122	.00	1.00	.6393	.4822
BAG	122	.00	7.00	1.5082	1.8464
COST	122	9.00	953.00	416.5492	285.4320
INCOME	122	5000.00	85000.00	38073.7705	18463.1274
Valid N (listwise)	122				

*This data is from a U.S. Department of the Interior survey (conducted by U.S. Bureau of the Census) which looks at a yes/no response to a question about the "willingness to pay" higher travel costs for deer hunting trips in North Carolina (a more complete description of this data can be found here).

The linear probability model

"Why shouldn't I just use ordinary least squares?" Good question.

Consider the linear probability (LP) model:

$$Y = \boldsymbol{a} + \boldsymbol{B}X + e$$

where

- Y is a dummy dependent variable, =1 if event happens, =0 if event doesn't happen,
- ***a*** is the coefficient on the constant term,
- ***B*** is the coefficient(s) on the independent variable(s),
- X is the independent variable(s), and
- e is the error term.

Use of the LP model generally gives you the correct answers in terms of the sign and significance level of the coefficients. The predicted probabilities from the model are usually where we run into trouble. There are 3 problems with using the LP model:

1. The error terms are heteroskedastic (*heteroskedasticity occurs when the variance of the dependent variable is different with different values of the independent variables*): var(e)= p(1-p), where p is the probability that EVENT=1. Since P depends on X the "classical regression assumption" that the error term does not depend on the Xs is violated.
2. e is not normally distributed because P takes on only two values, violating another "classical regression assumption"
3. The predicted probabilities can be greater than 1 or less than 0 which *can be a problem* if the predicted values are used in a subsequent analysis. Some people try to solve this problem by setting probabilities that are greater than (less than) 1 (0) to be equal to 1 (0). This amounts to an interpretation that a high probability of the Event (Nonevent) occuring is considered a sure thing.

The logistic regression model

The "logit" model solves these problems:

$$\ln[p/(1-p)] = \boldsymbol{a} + \boldsymbol{B}X + e \text{ or}$$

$$[p/(1-p)] = \exp(\boldsymbol{a} + \boldsymbol{B}X + e)$$

where:

- ln is the natural logarithm, $\log_{exp}$, where exp=2.71828…
- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"
- all other components of the model are the same.

The logistic regression model is simply a non-linear transformation of the linear regression. The "logistic" distribution is an S-shaped distribution function which is similar to the standard-normal distribution (which results in a probit regression model) but easier to work with in most applications (the probabilities are easier to calculate). The logit distribution constrains the estimated probabilities to lie between 0 and 1.

For instance, the estimated probability is:

$$p = 1/[1 + \exp(-\boldsymbol{a} - \boldsymbol{B}X)]$$

With this functional form:

- if you let $\boldsymbol{a} + \boldsymbol{B}X = 0$, then p = .50
- as $\boldsymbol{a} + \boldsymbol{B}X$ gets really big, p approaches 1
- as $\boldsymbol{a} + \boldsymbol{B}X$ gets really small, p approaches 0.

A graphical comparison of the linear probability and logistic regression models is illustrated here.

Interpreting logit coefficients

The estimated coefficients must be interpreted with care. Instead of the slope coefficients ($\boldsymbol{B}$) being the rate of change in Y (the dependent variables) as X changes (as in the LP model or OLS regression), now the slope coefficient is interpreted as the rate of change in the "log odds" as X changes. This explanation is not very intuitive. It is possible to compute the more intuitive "marginal effect" of a continuous independent variable on the probability. The marginal effect is

$$dp/d\mathbf{B} = f(\mathbf{B}X)\mathbf{B}$$

where $f(.)$ is the density function of the cumulative probability distribution function [$F(\mathbf{B}X)$, which ranges from 0 to 1]. The marginal effects depend on the values of the independent variables, so, it is often useful to evaluate the marginal effects at the means of the independent variables. (*SPSS doesn't have an option for the marginal effects. If you need to compute marginal effects you can use the LIMDEP statistical package which is available on the academic mainframe*.)

An interpretation of the logit coefficient which is usually more intuitive (especially for dummy independent variables) is the "odds ratio"-- $\exp\boldsymbol{B}$ is the effect of the independent variable on the "odds ratio" [the odds ratio is the probability of the event divided by the probability of the nonevent]. For example, if $\exp\boldsymbol{B}_3 = 2$, then a one unit change in X_3 would make the event twice as likely (.67/.33) to occur. Odds ratios equal to 1 mean that there is a 50/50 chance that the event will occur with a small change in the independent variable. Negative coefficients lead to odds ratios less than one: if $\exp\boldsymbol{B}_2 = .67$, then a one unit change in X_2 leads to the event being less likely (.40/.60) to occur. {Odds ratios less than 1 (negative coefficients) tend to be harder to interpret than odds ratios greater than one(positive coefficients).} Note that odds ratios for continuous independent variables tend to be close to one, this does NOT suggest that the coefficients are insignificant. Use the Wald statistic (see below) to test for statistical significance.

Estimation by maximum likelihood

[For those of you who just NEED to know ...] Maximum likelihood estimation (MLE) is a statistical method for estimating the coefficients of a model. MLE is usually used as an alternative to non-linear least squares for nonlinear equations.

The likelihood function (L) measures the probability of observing the particular set of dependent variable values (p_1, p_2, ..., p_n) that occur in the sample. It is written as the probability of the product of the dependent variables:

$$L = \text{Prob}\ (p_1 * p_2 * * * p_n)$$

The higher the likelihood function, the higher the probability of observing the ps in the sample. MLE involves finding the coeffients (***a***, ***B***) that makes the log of the likelihood function (LL < 0) as large as possible or -2 times the log of the likelihood function (-2LL) as small as possible. The maximum likelihood estimates solve the following condition:

$$\{Y - p(Y=1)\}X_i = 0, \text{ summed over all observations}$$

{*or something like that ...* }

Hypothesis testing

Testing the hypothesis that a coefficient on an independent variable is significantly different from zero is similar to OLS models. The Wald statisitic for the ***B*** coefficient is:

$$\text{Wald} = [\boldsymbol{B}/\text{s.e.}_{\boldsymbol{B}}]^2$$

which is distributed chi-square with 1 degree of freedom. The Wald is simply the square of the (asymptotic) t-statistic.

The probability of a YES response from the data above was estimated with the logistic regression procedure in SPSS (click on "statistics," "regression," and "logistic"). The SPSS results look like this:

Variables in the Equation							
Variable	**B**	**S.E.**	**Wald**	**df**	**Sig**	**R**	**Exp(B)**
	[1]	[2]	[3]		[4]	[5]	[6]
BAG	0.2639	0.1239	4.5347	1	0.0332	0.1261	1.302
INCOME	4.63E-07	1.07E-05	0.0019	1	0.9656	0	1
COST	-0.0018	0.0007	6.5254	1	0.0106	-0.1684	0.9982
Constant	0.9691	0.569	2.9005	1	0.0885		
Notes:							
[1] B is the estimated logit coefficient							
[2] S.E. is the standard error of the coefficient							
[3] Wald = $[B/S.E.]^2$							
[4] "Sig" is the significance level of the coefficient: "the coefficient on BAG is significant at the .03 (97% confidence) level."							
[5] The "Partial R" = sqrt{[(Wald-2)/(-2*LL(a)]}; see below for LL(a)							
[6] Exp(B) is the "odds ratio" of the individual coefficient.							

Evaluating the overall performance of the model

There are several statistics which can be used for comparing alternative models or evaluating the performance of a single model:

1. The model likelihood ratio (LR), or chi-square, statistic is

LR[i] = -2[LL(***a***)- LL(***a***,***B***)]

or as you are reading SPSS printout:

LR[i] = [-2 Log Likelihood (of beginning model)]

- [-2 Log Likelihood (of ending model)].

where the model LR statistic is distributed chi-square with i degrees of freedom, where i is the number of independent variables. The "unconstrained model", LL(**a**,$\mathbf{B}_i$), is the log-likelihood function evaluated with all independent variables included and the "constrained model" is the log-likelihood function evaluated with only the constant included, LL(**a**).

Use the Model Chi-Square statistic to determine if the overall model is statistically significant.

2. The "Percent Correct Predictions" statistic assumes that if the estimated p is greater than or equal to .5 then the event is expected to occur and not occur otherwise. By assigning these probabilities 0s and 1s the following table is constructed:

Classification Table for YES The Cut Value is .50				
		Predicted		% Correct
		0	1	
Observed	0	9	35	20.25%
	1	4	74	94.87%
Overall				68.03%

the bigger the % Correct Predictions, the better the model.

3. Most OLS researchers like the R^2 statistic. It is the proportion of the variance in the dependent variable which is explained by the variance in the independent variables. There is NO equivalent measure in logistic regression. However, there are several "Pseudo" R^2 statistics. One psuedo R^2 is the McFadden's-R^2 statistic (sometimes called the likelihood ratio index [LRI]):

McFadden's-R^2= 1 - [LL(***a***,***B***)/LL(***a***)]

= 1 - [-2LL(***a***,***B***)/-2LL(***a***)]

where the R^2 is a scalar measure which varies between 0 and (somewhat close to) 1 much like the R^2 in a LP model. Expect your Pseudo R^2s to be much less than what you would expect in LP

model, however. Because the LRI depends on the ratio of the beginning and ending log-likelihood functions, it is very difficult to "maximize the R^2" in logistic regression.

The Pseudo-R^2 in logistic regression is best used to compare different specifications of the same model. Don't try to compare models with different data sets with the Pseudo-R^2 [referees will yell at you ...].

Other Pseudo-R^2 statistics are printed in SPSS output but [YIKES!] I can't figure out how these are calculated (even after consulting the manual and the SPSS discussion list)!?!

Source: SPSS Output			
(-2)*Initial LL	[1]	159.526	
(-2)*Ending LL	[2]	147.495	
Goodness of Fit	[3]	123.18	
Cox & Snell-R^2		0.094	
Nagelkerke-R^2		0.129	
	Chi-Square [4]	df	Significance
Model	12.031	3	0.0073
Notes:			
[1] LL(a) = 159.526/(-2) = -79.763			
[2] LL(a,B) = 147.495/(-2) = -73.748			
[3] GF = $[Y - P(Y=1)]^2/[Y - P(Y=1)]$			
[4] Chi-Square = -2[LL(a)-LL(a,B)] = 159.526 - 147.495			
McFadden's-R^2 = 1- (147.495/159.526) = 0.075			

That's it! You are now a logistic regression expert!
