Binary Logistic Regression

(Contributed By John Pezzullo)

Background Info (just what is logistic regression, anyway?)

Ordinary regression deals with finding a function that relates a **continuous** outcome variable (dependent variable y) to one or more predictors (independent variables x_1, x_2, etc.). Simple linear regression assumes a function of the form:
$y = c_0 + c_1 * x_1 + c_2 * x_2 +...$
and finds the values of c_0, c_1, c_2, etc. (c_0 is called the "intercept" or "constant term").

Logistic regression is a variation of ordinary regression, useful when the observed outcome is **restricted to two values**, which usually represent the occurrence or non-occurrence of some outcome event, (usually coded as 1 or 0, respectively). It produces a formula that predicts the **probability of the occurrence** as a function of the independent variables.

Logistic regression fits a special s-shaped curve by taking the linear regression (above), which could produce any y-value between minus infinity and plus infinity, and transforming it with the function: $p = \text{Exp}(y) / (1 + \text{Exp}(y))$ which produces p-values between 0 (as y approaches minus infinity) and 1 (as y approaches plus infinity). This now becomes a special kind of non-linear regression, which this page performs.
Logistic regression also produces *Odds Ratios* (O.R.) associated with each predictor value. The *odds* of an event is defined as the probability of the outcome event **occurring** divided by the probability of the event **not occurring**. The odds ratio for a predictor tells the relative amount by which the odds of the outcome increase (O.R. greater than 1.0) or decrease (O.R. less than 1.0) when the value of the predictor value is increased by 1.0 units.

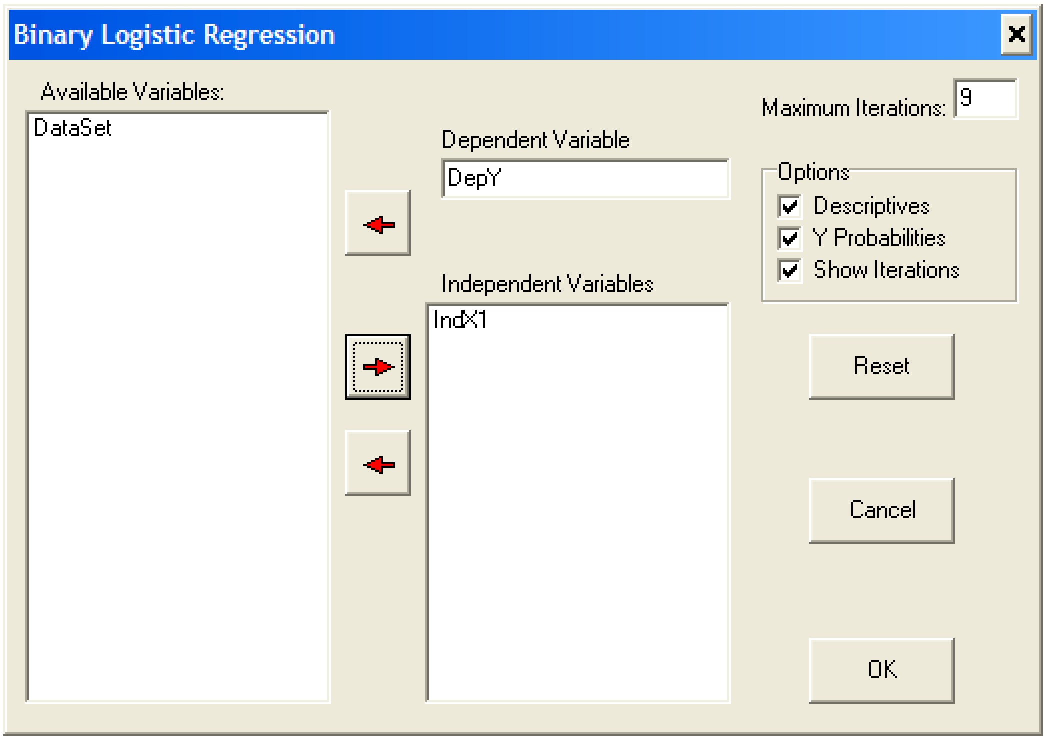
A standard iterative method is used to minimize the Log Likelihood Function (LLF), defined as the sum of the logarithms of the predicted probabilities of occurrence for those cases where the event occurred and the logarithms of the predicted probabilities of non-occurrence for those cases where the event did not occur.

Minimization is by Newton's method, with a very simple elimination algorithm to invert and solve the simultaneous equations. Central-limit estimates of parameter standard errors are obtained from the diagonal terms of the inverse matrix. Odds Ratios and their confidence limits are obtained by exponentiating the parameters and their lower and upper confidence limits (approximated by +/- 1.96 standard errors).

No special convergence-acceleration techniques are used. For improved precision, the independent variables are temporarily converted to "standard scores" (value - Mean) / StdDev. The *Null Model* is used as the starting guess for the iterations -- all parameter coefficients are zero, and the intercept is the logarithm of the ratio of the number of cases with y=1 to the number with y=0. Convergence is not guaranteed, but this page should work properly with most practical problems that arise in real-world situations.

This implementation has no predefined limits for the number of independent variables or cases. The actual limits are probably dependent on the user's available memory and other computer-specific restrictions.

When this analysis is selected from the menu, the form below is used to select the dependent and independent variables. We are using the BinLogReg2.LAZ file for an example.



Available Variables: DataSet

Dependent Variable: DepY

Independent Variables: IndX1

Maximum Iterations: 9

Options

☑ Descriptives

☑ Y Probabilities

☑ Show Iterations

The results obtained are:

```
Logistic Regression Adapted from John C. Pezzullo
Java program at http://members.aol.com/johnp71/logistic.html

Descriptive Statistics
10 cases have Y=0; 10 cases have Y=1.
Variable  Label            Average     Std.Dev.
     1           IndX1      5.5000      2.8723

Iteration History
-2 Log Likelihood =    27.7259  (Null Model)
-2 Log Likelihood =    21.0314
-2 Log Likelihood =    20.8022
-2 Log Likelihood =    20.7997
-2 Log Likelihood =    20.7997
Converged

Overall Model Fit... Chi Square =   6.9262 with df =   1 and prob. =
0.0085

Coefficients and Standard Errors...
Variable        Label     Coeff.     StdErr     p
    1           IndX1    -0.4908     0.2228     0.0276
Intercept     2.6994

Odds Ratios and 95% Confidence Intervals...
Variable            O.R.        Low   --   High
          IndX1     0.6121      0.3956     0.9473
```

```
      X         Y        Prob
     6.0000    0     0.4390
     7.0000    0     0.3238
     8.0000    0     0.2267
     9.0000    0     0.1522
    10.0000    0     0.0990
     6.0000    0     0.4390
     7.0000    0     0.3238
     8.0000    0     0.2267
     9.0000    0     0.1522
    10.0000    1     0.0990
     1.0000    0     0.9010
     2.0000    1     0.8478
     3.0000    1     0.7733
     4.0000    1     0.6762
     5.0000    1     0.5610
     1.0000    1     0.9010
     2.0000    1     0.8478
     3.0000    1     0.7733
     4.0000    1     0.6762
     5.0000    1     0.5610

Classification Table
           Predicted
        ---------------
Observed    0      1     Total
        ---------------
   0    |   9  |   1  |  10  |
   1    |   1  |   9  |  10  |
        ---------------
Total   |  10  |  10  |  20
        ---------------
```