Cornell University
Cornell Statistical Consulting Unit

StatNews #25

Logistic Regression When the Outcome is Measured With Uncertainty

October 1997

Revised May 2012

Logistic regression is often used to model the effect of explanatory variables on a binary response variable. In many situations, the binary response variable may not always be measured accurately. When this is the case, estimated coefficients and standard errors from logistic regression are too low. This newsletter discusses when and how to correct logistic regression estimates for imperfectly measured response variables.

The accuracy of a binary variable is indicated by its sensitivity and specificity. For example, one type of HIV test has a 6.7% probability of a false positive result. The specificity is 1-(probability of a false positive result), or 93.3%. The same test has a 3% probability of a false negative result. The sensitivity is 1- (probability of a false negative result), or 97%. A perfect test would have both sensitivity and specificity equal to 100%.

Response variables with imperfect sensitivity and specificity can occur in many fields. For example, a biologist may record whether or not a plant responds to a treatment, but some responses may be too small to measure, resulting in sensitivity less than 100%. In a survey, respondents may answer questions such as "have you used drugs in the past year?" incorrectly, either because they do not remember the correct answer or because they want to hide the truth. For most purposes, the researcher may be willing to assume that these discrepancies are not important, but this assumption can lead to wrong conclusions.

Logistic regression implicitly assumes that both sensitivity and specificity of the response variable are 100%. When either sensitivity or specificity is less than 100%, estimated logistic regression coefficients tend to be too close to zero. This makes it harder to detect relationships between explanatory and response variables. At the same time, the precision of the estimated coefficients is overstated, resulting in confidence intervals that tend to be too small.

A common response to this problem is to report the results from logistic regression with the caveat that the estimated coefficients are too close to zero. If the results indicate a statistically significant relationship between the response variable and a particular explanatory variable, the reader assumes that the true relationship is even stronger.

However, if the sensitivity and specificity are known, it is possible to use their values to adjust the logistic regression estimates. A SAS macro to perform this adjustment, written by Laurence S. Magder, is available on the web at http://medschool.umaryland.edu/epidemiology/software.asp
Note: the SAS macro is no longer available at that web site, or from the author's site. (April 21, 2012)

Contact the CSCU office if you need help in implementing this macro or interpreting the results.

Reference: Magder, Laurence S. and James P. Hughes, "Logistic Regression When the Outcome is Measured With Uncertainty", American Journal of Epidemiology, 146:2, pp. 195-203.

Author: Cara Olsen

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu.   Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).