StatNews #22

Logistic Regression for Repeated Measures
September 1997

Logistic regression is used to model the relationship between a categorical response variable and one or more explanatory variables that can be continuous or categorical. While many statistical software packages can fit basic logistic regression models, until recently the most frequently used packages could not fit logistic regression models for data with repeated measures on the same subject. The latest versions of both SAS and STATA include procedures that will handle this type of data appropriately.

A typical application of logistic regression arises in epidemiology. Suppose the researcher wants to study the relationship between various environmental factors and prevalence of an infection. For each subject in the study, the researcher collects data on exposure to environmental factors, and whether or not the subject suffers from the infection. She fits a logistic regression to these data, with infection (yes or no) as the response variable and level of exposure to each environmental factor as the explanatory variables, and estimates the effect of the environmental factors on the odds of contracting the infection.

Now suppose the researcher returns every year and collects new data on the same subjects. She now fits a logistic regression on the entire data set, which contains several observations for each subject. The logistic regression model assumes that the observations are independent, but since observations from the same subject are likely to be correlated, this is not usually a reasonable assumption. When the assumption of independent observations is violated, the estimated standard errors from logistic regression are incorrect, and can lead to incorrect inferences.

The method of generalized estimating equations (GEE) can be used to account for correlations among observations from the same subject. The GEE method estimates the regression parameters assuming that the observations are independent, uses the residuals from this model to estimate the correlations among observations from the same subjects, and then uses the correlation estimates to obtain new estimates of the regression parameters. This process is repeated until the change between two successive estimates is very small.

GEE can be implemented in SAS 6.12 (using the REPEATED option in PROC GENMOD) and STATA (using the XTGEE procedure). Both SAS and STATA allow the user to specify different correlation structures for the repeated observations, and to fit other generalized linear models such as Poisson, negative binomial, or multinomial logistic regression in addition to logistic regression. Contact the CSCU office if you need help implementing this procedure.

Author: Cara Olsen