November 2002: Number of Events per Variable (Rule 4.6)

Rules of the month are numbered in accordance with the numbering in the book. Thus, Rule 1.1 refers to the first rule in Chapter 1. And so on. These comments do not repeat the material in the book but highlights and amplifies it. A rule is stated as found in the book and then discussed.

"Obtain at Least Ten Subjects for Every Variable Investigated." (Rule 4.6)

"In logistic regression situations about 10 events per variable are necessary in order to get reasonably stable estimates of the regression coefficients."

Further Comments on the Rule

The rule is based on the work of Peduzzi et al. (1996) who simulated logistic regressions using estimates from a study reported in Peduzzi et al. (1985).

Additional comments can be found in the second edition of Hosmer and Lemeshow (2000)—the standard text on logistic regression, pages 346-347. They, citing the same source, explain that they prefer to consider the number of events per parameter (rather than events per variable since a variable may be represented by multiple terms as, for example, a categorical variable represented by several dummy variables each one of which has a slope estimate or parameter) that are needed to have good estimation properties. They also note that the "relevant quantity is the frequency of the least frequent outcome." This can be either the event or its complement; for example, if there are 500 subjects and 400 die we can either talk about 400 deaths or 100 survivors. Hosmer and Lemeshow therefore define $m = \min(n_1, n_0)$ where n_1 and n_0 are as illustrated in the previous sentence.

They then turn the problem around and ask, suppose there are m subjects, how many parameters can be estimated based on this rule of 10. The answer is no more than $m/10-1$. That is, if p is the number of parameters that can be estimated then, $p+1 \leq \min(n_1, n_0)/10$. In the above example m=min(400, 100)=100 and the recommended number of parameters to be studied is (100/10)-1=9. The intercept constitutes an additional one parameter, independent of the number of covariates.

This formulation makes it clear that given, say n subjects in an observational study, the maximum number of variables in any situation is (n/20)-1, since the most favorable situation is where n/2 of the subjects have the event and n/2 do not. So in this formulation it's more like a rule of 20 than a rule of 10.

Correction: The heading in the text for the rule refers to the number of "subjects" while the rule refers, more correctly, to the number

of “events.” In the next printing of the text I will change the term in the heading from “subjects” to “events.”

References

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression.* Second edition. John Wiley and Sons, New York, NY.

Peduzzi, P., Concato, J., Kemper E., Holford, T.R. and Feinstein, A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology,* **49:** 1372-1379.

Peduzzi, P., Detre, K. and Gage, A. (1985). Veterans administration cooperative study of medical versus surgical treatment for stable angina—Progress report: Section 2—Design and baseline characteristics. *Progress in Cardiac Disease,* **28:** 235-243.

Responses

This section is intended to contain reader comments and perhaps responses from me. It provides a forum for discussion and further reflection.