RE: INTERPRETING MULTIPLE LOGISTIC REGRESSION COEFFICIENTS IN PROSPECTIVE OBSERVATIONAL STUDIES

Abbott and Carroll's paper (1) is motivated by the problem of estimating the effect of a reduction in triglycerides on the probability of developing coronary heart disease. If high density lipoprotein cholesterol (HDL-C) is an intermediary in the causal pathway between triglycerides and coronary heart disease, use of the triglycerides coefficient from a multiple regression model which controls for HDL-C is inappropriate when interest lies not in the effect of a change in triglycerides while HDL-C is held constant, but rather in the full effects of change in triglycerides, including those which might be mediated through changes in HDL-C. On the other hand, use of a regression model which does not include HDL-C might also be inappropriate because failure to control for HDL-C could result in confounding, since the independent effect of HDL-C on coronary heart disease may be wrongly attributed to its correlate, triglycerides. In other words, the problem is that HDL-C can act in one or both of two ways: as an "overmatcher" in Breslow and Day's first sense (2, pp. 104–105), in which case it should be excluded from the model, or as a confounder, in which case it should be included in the model.

Abbott and Carroll propose the "projected slope" as a resolution of this dilemma, but they do not explain what the projected slope measures. In fact, it is basically the univariate slope on X_1.

The problem which motivated Abbott and Carroll can also occur when the outcome variable is continuous. Applying their proposed solution in a classic multiple regression context reveals its flaw. Replace the left-hand side of their equation 6 with $E(Y_i)$, the expected value of a continuous variable observed with risk factor x_{1i} and covariate x_{2i}. In their notation, $\delta_1 = \beta_1 + \gamma_1\beta_2$, where δ_1 is the projected slope, β_1 and β_2 are the respective regression coefficients from the multiple regression of Y on X_1 and X_2, and γ_1 is the slope of the regression of X_2 on X_1. It is well known, however, that $\beta_1 + \gamma_1\beta_2 = \beta^*_1$ where β^*_1 is the slope of the univariate regression of Y on X_1 (see, for example, reference 3, pp. 394–398). Since the ε_i's, the residuals from the regression of X_2 on X_1, are uncorrelated with the X_{1i}'s, δ_1 is not affected by inclusion of the $\delta_{2\epsilon}$ term in their equation 6. So $\delta_1 = \beta^*_1$ and the projected slope simply equals the univariate coefficient and does not "consider the relationship" between X_1 and X_2. The analagous result can easily be obtained when the projected slope is used with several independent variables.

The distinction between the interpretation of coefficients from models with the same right-hand side in classic and logistic regression is that the coefficients from the former apply to changes in expected values of continuous variables, while the coefficients from the latter apply to changes in the logits of the probabilities of binary events. The estimates of projected slope and the univariate coefficients in Abbott and Carroll's tables are not identical because of the iterated reweighting of each observation in logistic regression. If the estimates of β_1, β^*_1, and γ_1 were all based on weighted regressions, with the same vector of weights in each regression, the estimates of the projected slope and the univariate coefficients would be equal, just as in classic regression. So in logistic regression, the projected slope is simply an approximation to the univariate coefficient.

The projected slope does not solve the problem raised by Abbott and Carroll. However, it may not even be an important issue in their example. Assume that a preventive measure such as a change in diet is expected to have effects on both triglycerides and HDL-C. The effect of the intervention on risk could be estimated as the sum of the estimated effects due to changes in triglycerides and HDL-C (and perhaps an interaction) using regression estimates from a model which includes both lipid variables. The issue is important, however, when, for example, the risk factor is consumption of red meat and the covariate is HDL-C. Including HDL-C in the model is inappropriate if reduction in consumption of red meat acts on coronary heart disease risk through a pathway which includes increasing HDL-C. Thus an estimate of the effect of an intervention based on a coefficient from a model including HDL-C might underestimate the impact of the intervention. On the other hand, exclusion of HDL-C from the model may confound the estimate of the effect of red meat consumption if HDL-C is an independent risk factor for coronary heart disease; i.e., if the risk of disease for individuals who consume the same amount of red meat varies with HDL-C.

References

1. Abbott RD, Carroll RJ. Interpreting multiple logistic regression coefficients in prospective observational studies. Am J Epidemiol 1984;119:830–6.
2. Breslow NE, Day NE. Statistical methods in cancer research: the analysis of case-control studies. Lyon: International Agency for Research on Cancer, 1980.
3. Snedecor GW, Cochran WG. Statistical methods. 6th ed. Ames, IA: Iowa State University Press, 1967.

Sholom Wacholder
McGill University
Department of Epidemiology and Biostatistics
Montreal PQ H3A 1A2
Canada

RE: "INTERPRETING MULTIPLE LOGISTIC REGRESSION COEFICIENTS IN PROSPECTIVE OBSERVATIONAL STUDIES"

Collinearity, or correlation among explanatory variables, is a "statistical disease" that afflicts many data sets. Abbott and Carroll (1) have recently developed a procedure which they argue reduces the problems created by collinearity. The purpose of this comment is to indicate that their remedy amounts to little more than dropping a collinear variable from the estimating equation, which is a procedure many pragmatic researchers have been practicing for years.

Multivariate models are becoming increasingly popular, as readers of this journal are aware. Two of the most popular are ordinary least squares or multiple regression, and the multiple logistic model. Ordinary least squares or multiple regression is most appropriately applied when the dependent variable does not have limiting values such as 0 and 1. If the dependent variable does have the limiting values of 0 and 1, the logistic model is an appropriate procedure to use. When limiting values exist, as they do when a person either does (value 1) or does not (value 0) have a disease, unlike ordinary least squares, the logistic model will not allow predicted values to go beyond the limits or provide inefficient estimates resulting from a heteroskedastic error term. Logistic regressions are generally run using iterative, nonlinear, maximum likelihood computer routines. If the dependent variable ranges in value from 0 to 1, but never assumes either of those values, then the cheaper ordinary least squares approximation to the logit maximum likelihood approach can be used in which the log-of-the-odds ratio is regressed on the explanatory variables.

A problem that frequently plagues prospective studies using a multivariate approach is collinearity. Collinearity can be of two types: exact and inexact. If in the case of two explanatory variables collinearity is exact, it means that the two variables are an exact linear combination of each other. If X_1 and X_2 are the explanatory variables, then if

$$X_1 = a + bX_2 \tag{1}$$

where "a" and "b" are constants, X_1 and X_2 are said to be exactly collinear. More typical is the inexact variety where

$$X_1 = a + bX_2 + \epsilon \tag{2}$$

and ϵ is a random error term. In equation 2, X_1 and X_2 are linearly related, but not perfectly. If the absolute values of ϵ are large, the collinearity is small. If the absolute values of ϵ are small, the inexact collinearity is severe.

When collinearity is exact, ordinary least squares estimates of parameters cannot be obtained due to a singular data matrix that requires inversion. When collinearity is high but not exact, estimated standard errors are large and, other things equal, t-statistics are small. Under these conditions, it frequently happens that the researchers obtain insignificant coefficients but significant F or chi-square tests for the overall regression. The significant F or chi-square suggests that one or both explanatory variables are important in the equation. But because of the collinearity, the ordinary least squares or logistic procedure cannot separate the influences of the two explanatory variables on the dependent variable.

Abbott and Carroll (1) suggest a procedure for dealing with this problem when logistic maximum likelihood procedures are used. The procedure is not very useful, however, since the results obtained will always be very close to the results obtained when one of the explanatory variables is dropped from the analysis.

Abbott and Carroll begin with two equations in which X_1 and X_2 are assumed to be highly collinear

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu \tag{3}$$

and

$$X_2 = \gamma_0 + \gamma_1 X_1 + \epsilon \tag{4}$$

where μ and ϵ are random and independent error terms and the subscript i is dropped for easier reading. They then substitute into equation 3 for X_2 obtaining

$$\text{logit}(p) = (\beta_0 + \gamma_0\beta_2) + (\beta_1 + \gamma_1\beta_2)X_1 + \beta_2\epsilon + \mu \tag{5}$$

The expression $(\beta_1 + \gamma_1\beta_2)$ is called the "projected slope". According to their Appendix, equation 5 (their equation (6)) can be estimated by regressing the logit(p) on X_1 and e where e are the residuals from equation 4 above, which is estimated with ordinary least squares.

Our first point is that Abbott and Carroll's procedure is not useful when ordinary least squares approximation is used. Under ordinary least squares the estimated "projected slope" will be identical to the estimated slope in a simple regression of the dependent variable on X_1. This result follows from two facts. First, the residuals in equation 4 and X_1 will have a zero covariance in sample data simply as a result of applying ordinary least squares in regressing X_2 on X_1. The second normal equation requires that $\Sigma X_1 e = 0$ which is the numerator of the sample covariance between X_1 and e (2, p. 119). Second, when explanatory variables have zero sample covariance, ordinary least squares multiple regression estimators collapse to simple (univariate) regression estimators. If X_1 and e have zero covariance and the dependent variable is regressed on X_1 alone and then in a separate equation is regressed on X_1 and e, the estimated slope on X_1 will be exactly the same in either regression (2, p. 191).

Thus, the Abbott and Carroll (1) procedure is redundant and superfluous when ordinary least squares is applied to a dependent variable with no limits or the log-of-the-odds-ratio approximation to logistic model. The estimated "projected slope" of X_1 and its standard error will be identical to those obtained in a univariate regression which ignores both X_2 and e.

Our second point is that when Abbott and Carroll's procedure is applied to an equation being estimated with an iterative maximum likelihood logistic routine, the estimated "projected slope" of X_1 and its standard