Comment

Melinda Drum and Peter McCullagh

We find ourselves in substantial agreement with the authors on most of the issues raised in the paper. This makes criticism difficult, so our comments are restricted to two points, namely, the role of marginal regression models in applied work and the role of the so-called robust variance estimator of empirical sandwich variance estimator.

MARGINAL REGRESSION MODELS

The purpose of this note is to comment on the role of marginal regression models in the context of longitudinal studies. The authors' arguments are presented in terms of parameter interpretation or "reproducibility," or what McCullagh and Nelder (1989) in a similar context call "upward compatibility." In the marginal model, one focuses primarily on a model for the marginal mean vector. In a fully specified model, the focus is usually on the conditional distribution of the response at time t given the observed history of that individual on previous examinations.

Most studies have multiple purposes, some specified in advance, others after the fact. We take the view that, in general, the choice of model must depend not just on the nature of the variables and the choice of design, but also on the purpose of the study. Different questions usually require different models to be fitted to the same data. Some purposes are well served by a sequence of conditional models given the individual's history: other purposes are better served by a marginal regression model. The megalomaniacal strategy of fitting a grand unified model, supposedly capable of answering any conceivable question that might be posed, is, in our view, dangerous, unnecessary and counterproductive. It violates that basic principle of applied statistics, the avoidance of unnecessary modelling.

For present purposes, it is helpful to consider purposes of medical investigations in human subjects under four headings:

(i) Scientific understanding;
(ii) Clinical prediction;
(iii) Public policy decisions;
(iv) Epidemiological purposes.

A major distinction is immediately apparent in that purposes (i) and (ii) focus on the individual, whereas (iii) and (iv) focus on the population, however defined. Purposes (i) and (ii) are therefore best served by an individual-specific model focusing on the conditional distribution of an individual's disease status given the relevant past history for that individual. Purposes (iii) and (iv) are more directly served by a marginal regression model in which population rates or averages are the primary parameters under study.

The following simplified example serves to illustrate how the purpose of a study can have a drastic effect on model choice and parameter interpretation. Consider a study set up to investigate the effect of mother's smoking habits, X_1, on infant perinatal mortality, Y. Any such study will, as a matter of course, collect information on a large number of variables, many of which have little bearing on the advertised purpose of the study. For present purposes, we consider only two other variables, namely, age of mother X_2 and duration of pregnancy Z. Some of these variables may be discrete, others continuous. For example, Z might be measured in days (essentially continuous), or it might be an indicator for premature birth, however defined. Measurement scales, however, are irrelevant to the discussion that follows. Duration of pregnancy has the status of an intermediate variable, temporally subsequent to X_1, X_2, but preceding the ultimate response Y. In the trivariate regression of Y on X_1, X_2, Z, duration of pregnancy is by far the most effective predictor of perinatal mortality. Premature babies have a much higher mortality rate than full-term babies. Smoking habit of the mother has little additional predictive power. Whether or not it has a causal interpretation suited to purpose (i), this trivariate regression model is the model of choice for clinical prediction.

From a public-policy perspective, however, the conditional mortality rate given duration of pregnancy is of little interest. It is a plausible supposition that the major effect of smoking occurs in utero, reducing the duration of pregnancy and increasing the proportion of premature births. The direct perinatal effect of smoking during the week following birth is likely to be small. Smoking kills, but only slowly. In the regression of Y on X_1, X_2, Z, the major effect of increased tobacco consumption may be masked by a consequent reduction in duration of pregnancy. It is, of course, the total effect of smoking on perinatal mortality that is chiefly of interest for public health policy purposes. For that purpose, the marginal regression model of Y on X_1, X_2, omitting the intermediate variable Z, is required. In the absence of interaction with age, the coefficient of

Melinda Drum is Assistant Professor, Department of Biostatistics, University of Illinois at Chicago, Chicago, Illinois 60612. Peter McCullagh is Professor, Department of Statistics, University of Chicago, Chicago, Illinois 60637.

X_1 measures the total effect of smoking on perinatal mortality, not simply the additional effect adjusted for duration of pregnancy. The irony here is that the appropriate model for purposes (iii) and (iv) is obtained by omitting the most effective predictor variable, a lesson seldom taught in courses on model selection. For public-policy purposes, one would ordinarily also wish to study the effect of smoking on duration of pregnancy via a further regression model of Z on X_1, X_2. In other words, two distinct but related public-policy oriented questions lead to a model for the dependence of the marginal means $E(Y)$ and $E(Z)$ on covariates X_1, X_2.

In order to forecast the effect of an anti-smoking publicity campaign, it is necessary to know the amount by which smoking will be reduced, the effect of this reduction on perinatal mortality and its effect on the incidence of premature births. Only rarely would one also wish to know the effect of the reduction on the survival rate of premature infants. In other words, it is ordinarily unnecessary to model the joint distribution of (Y, Z) to answer most public-policy-related questions. A marginal regression model is entirely adequate for these purposes.

ROBUSTNESS OF THE SANDWICH

Robustness is a pejorative term with positive connotations used to describe the overall statistical properties of the empirical sandwich variance estimator, $\hat{H}_1^{-1}\hat{H}_2\hat{H}_1^{-1}$, as described in subsection 2.1. Consistency, even when the assumed variance function is incorrect, is the sole property used to justify the adjective robust. Now, consistency is a puny requirement of dubious importance in samples of moderate size. To advertize as robust a method whose only demonstrated property is consistency is to invite the wrath of SASA, the Statistical Advertising Standards Authority.

In a completely randomized design with continuous observations y_{ij}, the data may be reduced to $(\bar{y}_r, s_r^2, m_r)$, the sample mean, sample variance and sample size for each treatment. For the treatment effect $\bar{y}_r - \bar{y}_t$, the conventional variance estimate is

$$(1) \qquad s^2\left(\frac{1}{m_r} + \frac{1}{m_t}\right),$$

where s^2 is the pooled estimate of variance on $m_\cdot - k$ degrees of freedom. The empirical sandwich, however, gives the "robust" estimate

$$(2) \qquad s_r^2\frac{m_r - 1}{m_r^2} + s_t^2\frac{m_t - 1}{m_t^2}$$

without assigning degrees of freedom. By failing to pool variances, all risk of contamination is avoided. But the cost of protection seems high, particularly when some of the sample sizes are not large. The prosecution might argue that by using the sandwich estimate we pay a high cost for unnecessary protection and we get a shoddy estimate. Making due allowance for courtroom hyperbole, this seems to be an accurate assessment, at least in many applications.

For binary data, it is possible that the comparison might be more favourable to the sandwich estimator because of the absence of a variance parameter to be estimated. We have calculated the sandwich matrix explicitly for a number of designs involving correlated binary data. To describe just one of these, suppose that the design is completely randomized, but that instead of one observation per subject we have a sequence of k binary observations with exchangeable correlation matrix. In this balanced design, response probabilities are estimated by treatment means, $\hat{\pi}_r = \bar{y}_{r\cdot\cdot} = \sum_i \bar{y}_{ri\cdot}/m_\cdot$. This is the average on treatment r of the subject averages. The empirical sandwich variance estimate for the contrast $\hat{\pi}_r - \hat{\pi}_s$ is again given by (2) in which

$$s_r^2 = \frac{1}{m_r - 1}\sum_i (\bar{y}_{ri\cdot} - \hat{\pi}_r)^2$$

is the treatment-specific sample variance. By contrast, the model-based variance estimator, $\hat{H}_1^{-1}$, gives

$$\left(\frac{\hat{\pi}_r(1 - \hat{\pi}_r)}{m_r} + \frac{\hat{\pi}_t(1 - \hat{\pi}_t)}{m_t}\right)(1 + (k - 1)\hat{\rho}),$$

where $\hat{\rho}$ is a pooled estimate of the common correlation. Alternatively, the dispersion parameter $\varphi = 1 + (k - 1)\rho$ may be estimated directly, without assuming correlation exchangeability, by pooling across treatments (McCullagh and Nelder, 1989). Once again, failure to pool information on variability protects against contamination at the cost of substantially increased variability in the estimate. The empirical and model-based expressions agree only if $k = 1$.

The overall conclusion that we draw from these and other examples is that the empirical sandwich estimator can be a useful tool in applied work if all sample sizes are sufficiently large. It is not a panacea, nor does it supersede the model-based estimate. Although its probability limit is unaffected by the failure of certain model assumptions, the estimate itself is not robust in the usual senses of that word. For these reasons, unless there is good reason to believe that the assumed variance function is substantially incorrect, the model-based estimator seems to be preferable in applied work, particularly where small samples are involved. Ideally, one should compute both estimates and aim to understand any differences that occur.

ACKNOWLEDGMENT

P. McCullagh would like to acknowledge that his research was supported by NSF Grant DMS-91-01333.