Rejoinder

Daniel F. Heitjan

I thank Dr. Jim Burridge for contributing his remarks on my review. His interesting examples give further testimony to the continuing importance and vitality of grouped data research. I can add only a few observations on some points that he raised.

DOUBLE GROUPING

Burridge's example of doubly grouped survival data amply illustrates the fact that we statisticians view inference problems through curiously distorted lenses. Why bother with the double grouping, whose effect is bound to be slight, without also troubling oneself about the Weibull assumption, whose effect could be large? Yet there is a certain logic in the position. In a sense models are the currency of scientific inquiry. Even if it is not always used appropriately, the Weibull model is a popular reference point in the analysis of survival data. In this context, concern about the effects of grouping represents a move toward equalizing, so to speak, the rates of exchange, so that model fits can be compared free of effects that are due to differences in grouping mechanisms. This is not to say that the problem of model selection is not important, for assuredly it is. However, given that people are going to apply popular models to all sorts of grouped data, the question of grouping should be given its due.

As for the example itself, I would lean toward ignoring the grouping, if the data presented by Burridge were typical. However I have recently come across another example in which the double grouping cannot be ignored. The problem arises in data gathered by the Multicenter AIDS Cohort Study, an investigation of the natural history of AIDS currently under way in the United States. A cohort of homosexual men was recruited for the study in 1984 and have submitted to biannual physical examinations and interviews since then. One of the current goals of the study is to estimate the distribution of time from HIV infection to time of AIDS onset in these men. In those who develop AIDS, the date of onset can be determined fairly well, say to the nearest month. Unfortunately many in the cohort were already HIV seropositive at the outset of the study, so that the dates of infection for these men are known only to lie somewhere between 1977 (the supposed starting date of the epidemic) and 1984. Ignoring the double grouping, particularly on the left, seems clearly wrong. A multiple imputation approach to this difficult problem has been proposed by Taylor, Schwartz and Detels (1986), and further work is under way.

GROUPING IN PREDICTORS

As Burridge points out, the problem of accommodating grouping in predictor variables is challenging. Let me review some of the salient issues. A major complication injected by the grouping is that one cannot make inferences on the regression of Y on X without assuming something about the marginal distribution of the ungrouped X. To see this, note that the likelihood in this case is computed by integrating the joint density of the ungrouped X and Y over the set of possible ungrouped data sets. This likelihood depends explicitly on the marginal distribution of X. In observational studies where X is generated by simple random sampling, we are accustomed to specifying and fitting a marginal distribution for X, even if we don't usually need to do so. When the sampling is more complex—for example if prespecified numbers of units are selected at each of several X values—then the choice of distribution is more perplexing. Moreover if the grouping is coarse it may be difficult to use standard diagnostics in determining a suitable distribution for X from the data, however they were sampled. And of course all of these potential complications are overshadowed by the practical difficulty of carrying out the necessary computations.

COMPUTATIONAL ISSUES

Everything else being equal, most would, I think, prefer to use grouped data likelihoods rather than approximate substitute likelihoods. And if what one reads and hears can be believed, many who do not now do so would prefer to employ the Bayesian paradigm in inference. The lack of numerical software attuned to these tasks has prevented all but the specialists from carrying out their preferred analyses. I wish to point out, however, that although both problems go under the general heading of "computing," in fact they have very different features, and I believe they should be considered separately.

In my experience the problem of computing grouped data likelihoods is by far the easier of the two. The major difficulty here is that one must be ready to compute probability integrals over rectangles—not just one specific rectangle but any that might occur in the sample. For univariate data this is almost routine; reliable software exists for many popular distributions, and where it does not, common quadrature rules are available. Two-dimensional numerical integrals are harder, but can often be reduced to more standard

univariate integrals. Appreciable grouping in three or more dimensions is harder still, but happily such data sets appear to be rare. With regard to optimization routines, my experience has been that in a well-conditioned problem almost anything will work, whereas in a poorly conditioned problem almost nothing will. Most models one is likely to encounter are fortunately of the well-conditioned kind, as most likelihoods are roughly quadratic (at least near the mode); consequently, finding a good optimization routine should seldom be difficult. Because not all grouped data likelihoods are as nearly quadratic as one might wish, however, it is obviously desirable to use algorithms that can detect departures from quadratic shape. This is to my mind a persuasive argument for doing the more difficult grouped likelihood calculations rather than relying on moment corrections, which cannot reveal much about the shape of L.

The problem of computing marginal posterior distributions involves integration in potentially very many dimensions and is consequently more difficult; I have little new to say here. In problems where complete data Bayesian inferences are easy to construct, multiple imputation and data augmentation should be useful in producing grouped data posterior distributions. Neither method has yet been applied to a wide range of grouped data problems, although as I have suggested they hold great promise.

THE ROLE OF BAYESIAN METHODS

Burridge and I share a sympathetic view of the Bayesian paradigm, but I sense that I am more sanguine than he about the current prospects for Bayesian data analysis. I confess that I came to appreciate Bayes's theorem for reasons that were less philosophical than practical; I simply saw for myself how useful Bayesian methods can be in handling grouped and otherwise incomplete data. I believe this is because of the central role of the predictive distribution of the complete data in the Bayesian analysis. Once one can simulate this distribution, he can construct summary inferences for any estimand simply by averaging complete data inferences. As a result, howsoever one intends to interpret summary inferences, thinking about grouped data from a Bayesian perspective can suggest effective ways of obtaining such inferences. On the other hand, frequentist approaches, with the exception of maximum likelihood, have been less successful in providing general principles for handling grouped data. The reason, I believe, is that the frequentist's goal is to construct solutions that have optimal frequency properties. In principle this means that each new problem or optimality criterion requires an entirely new (and for grouped data very complicated) analysis.

I am not trying to suggest that frequentist evaluations of grouped data methods are unnecessary or irrelevant. On the contrary, I believe it is essential to understand the sampling properties of incomplete data methods and to select those that are likely to perform well in the situation under study. However I am asserting that Bayes's theorem can be used to clarify and solve incomplete data problems (including grouping), and that methods that cannot be justified, at least loosely, from the Bayesian perspective are not likely to be very good from a frequentist perspective either (cf. Rubin, 1984). Finally, I hope I have not left the impression that we should all defer our difficult grouped data problems until the epoch when advances in hardware, software and mathematics have combined to render numerical integration routine. As hard as many grouped data problems are, potent methods for solving them, particularly from a Bayesian point of view, are already with us.

ADDITIONAL REFERENCES

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.

TAYLOR, J. M. G., SCHWARTZ, K. and DETELS, R. (1986). The time from infection with human immunodeficiency virus (HIV) to the onset of AIDS. *J. Infect. Dis.* **154** 694–697.