Learning From Data

Semiparametric Models Versus Faith-based Inference

Mark Van der Laan, Alan E. Hubbard, and Nicholas Jewell

We appreciate the thoughtful comments by Subramanian and O'Malley[1] to our paper[2] on comparing mixed models and population average models, and the opportunity this response affords us to make a stronger and more general case regarding prevalent misconceptions surrounding statistical estimation. There are several technical points made in the paper that can be debated, but we will focus on what we believe is the crux of their critique—an issue that is widely shared (either explicitly or implicitly) by analyses of a majority of researchers using statistical inference from data to support scientific hypotheses.

We start with what we hope is an accurate summary of their argument: nonparametric identifiability of a parameter of interest from the observed data, considering knowledge available on the data-generating distribution, should not be a major concern in deciding on the choice of parameter of interest within a chosen data-generating model. Instead, the scientific question should guide the types of models used to make inferences from data. Thus, the proposed model for the data-generating distribution and the resulting target parameter should not be restricted by what is actually known (and knowable) about the data-generating process. There are times when the parameters of interest are defined only within the context of a mixed model (latent variable model), and thus giving up on one's parameter of interest for the sake of semiparametric inference, they argue, is counterproductive or even illogical.

The authors' assertion that the generalized-estimating-equation (GEE) approach requires parametric assumptions for interpretable parameters suggests that they failed to understand one of our basic points. This demonstrates a need to reiterate what we meant by defining the parameter of the data-generating distribution nonparametrically, or, in general, in the context of a realistic semiparametric statistical model. Their assertion that, for instance, modeling ranks as a nonparametric solution shows that there has been an unfortunate disconnect between what is meant by a semi-parametric approach and traditional notions of nonparametric statistics (ie, estimating the distribution of ranks). Though persuasively presented, the comment serves to underscore the need for vigorous debate on how we "learn" from data and what has been a consistent failure in our discipline to distinguish, both in estimation and inference, what can be learned from data and what aspects of some models are simply not deducible from data.

We propose a new "golden rule" in statistical estimation, namely that one should be able to define what an estimation procedure estimates purely as a function of the data-generating distribution. If, in the words of Box,[3] all models are wrong, then this parameter of the observed data-generating distribution is the only quantity one knows for sure is actually being estimated. Only under further nonidentifiable assumptions can this be interpreted as a parameter of a hypothesized model. It is also known (and we have

From the Division of Biostatistics, Berkeley School of Public Health, University of California, Berkeley, CA.
Editors' note: *Related articles appear on pages 467 and 475.*
Correspondence: Alan E. Hubbard, Division of Biostatistics, University of California, Berkeley School of Public Health, Berkeley, CA 94720. E-mail: hubbard@berkeley.edu.

ISSN: 1044-3983/10/2104-0479
DOI: 10.1097/EDE.0b013e3181e13328

shown in our paper) that often this "estimate" converges to a parameter that has very little interpretability; thus, despite the desire to model complexity with complex models, the results have questionable value as scientific evidence. We also believe that if these issues were more broadly understood, then practitioners would be more careful consumers of methods that have been used widely with little regard to the many biased results that must litter the scientific literature.[4]

We wholeheartedly agree that the scientific question of interest should precede the choice of an estimation procedure that attempts to address the relevant issue. But this platitude ignores our main message—that the parameter of interest must also be connected to the data at hand. Thus, it is questionable to interpret within-neighborhood effects of a covariate change (such as crime rates) from mixed-effect models when the data do not include any variation of such attributes within neighborhoods. Such inference is based entirely on assumption and not on data (this emphasizes another important distinction between clustered data associated with longitudinal observations on individuals, and information arising from studies of neighborhoods and health: in the latter case, many health predictors of interest are necessarily cluster-constant by definition and therefore cannot vary within neighborhoods). If within-neighborhood effects are of primary interest, and there is variation of predictors within neighborhoods, it is straightforward to derive appropriate parameters to describe these effects (and relevant estimators) without recourse to latent variables. In this case, such estimators have meaningful interpretation even if some model assumptions are incorrect when those arising from a blinkered application of a mixed-effects model possess no such understanding. In more complex situations, the same issue arises but may be much less obvious to the user. We attempt below to reiterate this point in general terms.

For illustrative purposes, assume we have an outcome of interest Y_{ij}, measured repeatedly on the same unit, i, (j = 1, n), an explanatory variable of interest, A_i, and a potential confounder, W_i; the data are assumed to be independent and identically-distributed realizations of $O_i = (\vec{Y}_i, W_i, A_i)$, where $\vec{Y}_i = (Y_{1i}, Y_{2i}, \ldots, Y_{ni})$, and $O_i \sim P_O$. Thus, in principal, we can estimate the distribution P_O without any additional identifiability assumptions; in fact, this distribution is the most one can learn from the data O_i. In parametric mixed models or other latent variable models, the observed data are assumed to be a byproduct of augmented (unobserved) data α, leading to a hypothetical random variable $X_i = (O_i, \alpha_i) \sim P_{O,\alpha}$. It is assumed that the distribution of this X is an element of a parametric model, $P_\psi = M_\psi(O_i, \alpha_i)$, indexed by simple (finite-dimensional) parameters, ψ. Denote the maximum-likelihood estimator (MLE) of ψ from the observed data as $\psi(P_n)$, where P_n is the empirical distribution of O. Then the analytic problem to be addressed before estimating ψ is "what does the estimator $\psi(P_n)$ converge to as the number of units increases?" That is, find the limit h in the consistency result $\psi(P_n) \xrightarrow{n\to\infty} h(P_O)$.

If the parametric model $M_\psi(O_i, \alpha_i)$ is misspecified (a certainty, if "all models are wrong"), then whatever h is, it represents what is actually estimated by the procedure. If $h(P_O)$ for a particular estimator, $\psi(P_n)$, does not have a close connection to the scientific question of interest, then it's time to rethink the approach unless there is a compelling external validation that the supposed parametric model is correct (but we know that never happens). This is ignored in many disciplines (particularly those inclined to assume that latent variable models solve the problem of being unable to measure the variables of interest). In many cases, it can be shown that h has no useful interpretation, and in fact that should be the default expectation. In mixed models, therefore, one can discuss the results as evidence relevant to the hypothesis of interest only under the unlikely premise that the guessed model $M_\psi(O_i, \alpha_i)$ is correct or "close" enough to correct. One cannot dispute the suggestion by Subramanian and O'Malley[1] that inferences derived from semi-parametric estimation versus more parametric approaches will sometimes be similar, but this should never be assumed.

Our original paper showed that the population—average model is at least a model of O, and that the estimates produced via the GEE approach therefore have an interpretable h (a type of projection of the regression model onto the true regression form $E(Y|A,W)$). Thus, as a reply to whether one needs parametric assumptions using the GEE estimating approach, the answer is "no" if one defines the parameter of interest appropriately as an interesting approximation. On the other hand, there is no general principal that what one estimates from a latent-variable model converges to a usefully interpretable parameter.

In conclusion, a general rule for most estimators is that they should reflect an appropriate model for P_O, and this model should exploit known restrictions on the distribution of O, and possibly nontestable assumptions, based only on real knowledge and not choices of convenience. How does this work for latent-variable models? One could define the target parameter as the limit of the maximum likelihood estimator of the posed latent-variable model, thereby defining the target parameter without restrictions on P_O, but, in most scenarios, these parameters are typically meaningless, and thus more care is needed. For now, many, including Subramanian and O'Malley,[1] are encouraging leaps of faith, not only turning a blind eye to the meaning of such estimators in a world where the assumed $M_\psi(O_i, \alpha_i)$ is fiction, but also ignoring the data themselves, and thus failing to adapt a model to the new information the data provide. Statistical analysis should avoid such leaps of faith and instead regain its status as a rigorous tool that learns from data.

REFERENCES

1. Subramanian SV, O'Malley JA. Modeling neighborhood effects: the futility of comparing mixed and marginal approaches. *Epidemiology*. 2010;21:475–478.
2. Hubbard AE, Ahern J, Fleischer NL, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*. 2010;21:467–474.
3. Box GE. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN, eds. *Robustness in Statistics*. New York: Academic Press; 1979.
4. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124. doi:10.1371/journal.pmed. 0020124.