Rejoinder to Discussants of "Sentencing Convicted Felons in the United States: A Bayesian Analysis Using Multilevel Covariates"

Iain Pardoe * Robert R. Weidner

Department of Decision Sciences, Charles H. Lundquist College of Business, University of Oregon, Eugene, OR 97403–1208, USA. Tel: (541) 346-3250. Fax: (541) 346-3341.

Department of Sociology Anthropology, University of Minnesota Duluth, Duluth, MN 55812, USA

We thank the discussants for their contributions. We think that they have greatly enhanced the value of our article by providing additional insights into our study and raising challenging questions about the choices we made during our analysis and alternative paths we could have taken. Our thanks also go to the executive editor and to the coordinating editor for organizing this discussion.

Space does not allow us to fully respond to all of the points raised by the discussants, so we restrict our comments to issues of disagreement or where elaboration would seem to be useful. Since many of the issues raised by the discussants overlap, we organize our rejoinder by theme rather than by discussant.

1 Bayesian modeling and choice of priors

As Professor Zaslavsky notes, hierarchical modeling and Bayesian modeling are not synonymous, and we could have approached this analysis from a frequentist perspective—maybe Professor De Leeuw would have preferred it if we had. Given our particular skills and experience, however, this would have been far from straightforward for us to accomplish, and—as remarked by Professor Browne—the Bayesian approach, at the very least, provides a relatively easy way to bolt-on a missing data imputation procedure and to produce posterior simulations that prove very useful in graphically assessing the model fit and interpreting results.

* Corresponding author.
Email addresses: ipardoe@lcbmail.uoregon.edu (Iain Pardoe), rweidner@d.umn.edu (Robert R. Weidner).
URL: http://lcb1.uoregon.edu/ipardoe (Iain Pardoe).

Our decision to follow the Bayesian path was therefore driven more by practical than philosophical concerns. Having embarked on this route, the familiar obstacle of prior sensitivity, specifically for the random effects covariance matrix, $\mathbf{\Gamma}^{-1}$, lays squarely in the way. De Leeuw would have us tackle this issue by adopting Wong and Mason's original empirical Bayes approach. This is quite possibly an excellent solution for those well-versed in the EM-algorithm, since Wong and Mason's 1985 article uses just such an approach to estimate $\mathbf{\Gamma}^{-1}$. However, it would take an adept programmer more fearless than ourselves to tackle the computations. Given the size of the dataset and the (relative) complexity of the model, a programming language such as C or Fortran would surely be needed. By contrast, the programming needed for the fully Bayesian approach in WinBUGS is reasonably straightforward, even if implementation is a little on the slow side (more on this later).

We followed standard guidance (provided in WinBUGS documentation) in using a conjugate Wishart prior for $\mathbf{\Gamma}^{-1}$, which is parameterized in WinBUGS with a matrix $\boldsymbol{R}$ that can be considered a prior guess for the mean of $\mathbf{\Gamma}^{-1}$ and degrees of freedom that can be considered an equivalent prior sample size. With little prior knowledge to go on, a prudent approach seemed to be setting the degrees of freedom as small as possible (the rank of $\boldsymbol{R}$), and picking reasonable values for the variances and covariances of the random effects (we picked ten for each variance, and zero for each covariance). We then performed an admittedly limited sensitivity analysis with two alternative choices for $\boldsymbol{R}$, and noted little change in the main results. We did however fail to quote any estimates for $\mathbf{\Gamma}^{-1}$ in the article—the estimate of the variance of the first random effect (the "intercept") was 1.5, with the remaining diagonal elements of the posterior mean of $\mathbf{\Gamma}^{-1}$ ranging between 0.0 and 0.5. The off-diagonal covariances were mostly in the -0.2 to 0.2 range.

2 Missing data

We welcome the advice provided by Browne and Zaslavsky on improving our missing data imputation methods. Having been conditioned to be wary of "double use" of data, it had not occurred to us to use the response to help impute missing covariates. And, at the time, using county-level information for individual-level imputation had seemed like one added complication too many. Inspired by the discussants, however, we investigated the following enhancements to the missing data imputation described near the end of Section 4:

- use Y as a predictor in each of the missing data models
- use CPCTAA as a predictor in one of the county-level regressions by modeling θ_6 as a random intercept in the missing data model for IBLACK (since there is high correlation between these two covariates across counties)
- similarly, use CUNEMP in the model for IACTCJS
- similarly, use CUNEMP in the model for IPPRIS

MCMC convergence was a little slower with this more complicated imputation, but there was a clear improvement in the fit of the missing data distributions. There was little change for most of the results of Table 2 (posterior means changing by up to ± 0.1), but there were some larger differences in the results for IPPRIS and its interactions (e.g. the estimated main effect increased from 1.7 to 2.3). This illustrates that there may have been scope for improving results for covariates with substantial missing data (IPPRIS had the most missing data of all the individual-level covariates).

3 Model selection

Browne and De Leeuw both mention what was to us probably the most challenging aspect of the data analysis. With so many individual-level covariates, county-level covariates, and individual-county interactions, it is hard to know where to start with variable selection. We did experiment with various ad-hoc procedures for reducing the number of interactions in the model, but were dissuaded from this approach by referees' comments on an earlier version of the article. As suggested by Browne, reversible jump MCMC methods can be used to guide model selection. However, given the somewhat ponderous nature of WinBUGS (the final model run described in the article took on the order of 24 hours to run on a reasonably fast personal computer), this was impractical for this dataset.

We ultimately decided to keep all possible individual-county interactions (although not, as noted by Browne, individual-level interactions) and let the analysis sort out important effects from unimportant ones. This seemed to us far less dangerous than fixing certain interaction effects at zero by excluding them from the model. De Leeuw's concern that there may be many qualitatively different models with approximately the same fit does not seem to be borne out by our experience. The reduced models that we did experiment with (not reported in the article) each gave very similar results to the final (reported) model.

Of course, deciding which covariates and interactions to include presupposes that you have all the relevant data from which to make your choice—Zaslavsky provides a nice discussion of this point. As with many analyses of this nature, the time spent researching and obtaining relevant data for this study is not reflected in the column-space devoted to this issue in the article. We can perhaps add a few more details here though.

When considering which contextual covariates to include in our models, we faced the dual challenge of identifying indicators for complex, sometimes multidimensional, constructs and then actually obtaining the indicator for the relevant period (i.e., 1998). For example, in the case of the concept "economic circumstances," we contemplated using either a county's poverty level or its level

of unemployment. We chose to include the latter (CUNEMP) in our analyses, as it is used in a plurality of contextual sentencing studies and prior research comparing prison use across jurisdictions. We would have considered using an alternative indicator—economic inequality as gauged by the Gini index—had we been able to obtain it at the county level for 1998. Our decision-making process in regard to the other factors was similar; the other five contextual covariates were included on the basis of their availability and the fact that they were favored in prior relevant research. Of course this is not to suggest that the resultant model specifications are ideal; as we indicate in Section 7, there are several contextual covariates that, for various reasons, we were unable to incorporate in our models.

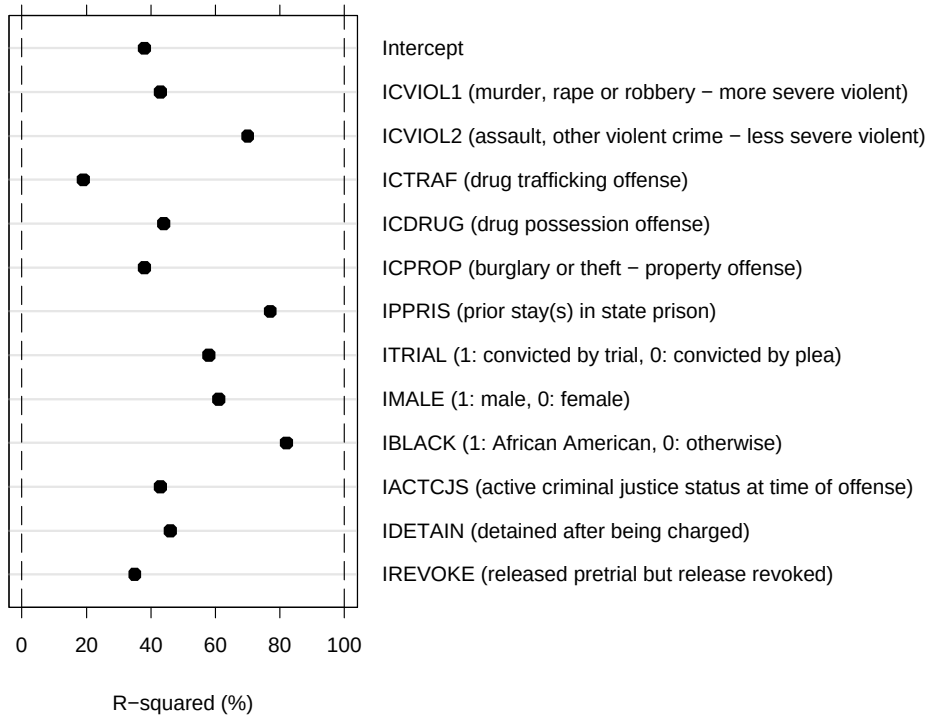
Browne extols the virtues of the DIC diagnostic for comparing models, and we wholeheartedly concur. This methodology came to our attention too late for the original analyses, but we have now been able to apply this diagnostic to our data analysis. The relevant calculations are particularly easy when using Bugs.R software (Gelman, 2004) as an interface between WinBUGS and R (R Development Core Team, 2004)—for example, the final model reported in the article has a DIC value of 52,857. For comparison, the model that excludes the five interactions with the highest coefficients of variation has a DIC value of 52,803, suggesting a slightly better fit for this reduced model. We may have been able to use the DIC diagnostic in this way to reduce the number of interactions in the model, although, as indicated above, computer-time limitations would have made this difficult to implement in anything other than an ad-hoc manner.

Browne also wondered whether much was gained by using all 13 sets of random effects rather than just a random intercept. Indeed, the model with just a random intercept (but including all the individual-county interactions present in the final model in the article) does appear to provide a comparable fit, with a DIC value of 52,695. Results for this simpler model were very similar to those in Table 2. In retrospect then, this simpler approach may have been better, although it appears that no harm was done with the more complicated approach described in the article.

4 Variance components and explained variance

Browne notes how estimates of $\mathbf{\Gamma}^{-1}$ can be helpful in discerning the importance of the county in partitioning explained variance between the different levels of the model; Zaslavsky also discusses the importance of this issue. Recent work in Gelman and Pardoe (2004b) proposes a new approach to defining explained variance at each level of a multilevel model based on comparing variances in a single fitted model rather than comparing to a null model (as previous methods have done). Application of this method to this application produces explained variance (R^2) measures of 40% at the individual-level, and between 19% and 82% for each of the county-level models (equation (2) in the article)—see Figure 1.

Fig. 1. Measure of explained variation (R^2) for each of the county-level models.



So, for example, the county-level covariates explain almost 40% of the variation (in sentencing) among counties, while they explain about 60% of the systematic variation in the gender effects across counties.

Zaslavsky also goes on to note that it can be useful to describe the difference between predictions for observations at low and high values of a covariate, with other covariates fixed at their means. This concept is related to work in Gelman and Pardoe (2004a) which proposes methodology for calculating the expected change in an outcome measure associated with a unit change in one of the predictors, for models with nonlinearity, interactions, and variance components.

5 Does a hierarchical approach work here?

De Leeuw wonders whether all our endeavors were in vain—would a standard logistic regression analysis have worked as well, and have we really moved our understanding of sentencing variation in the U.S. forward, or just left it more confused than ever? Browne and Zaslavsky offer some welcome encouragement here, and the model assessment carried out in Section 5 provides further support that our analysis may be useful (the article referenced here—Pardoe 2004—also shows quite clearly that a standard logistic regression analysis of this dataset is flawed).

What then of the impact on the field of sentencing. Fortunately, one of us could be

considered an expert in this area, and rather than throwing our arms up in desperation, we believe a more measured, optimistic appraisal is warranted. To our knowledge, there is only one published study that, like ours, uses multilevel modeling to consider the effects of contextual and case-level factors on individual sentence outcomes using a sample of cases from multiple jurisdictions located in states in every region of the country. This novel research strategy can be seen as an improvement upon prior sentencing studies for two key reasons. First, it allows us to circumvent a critique that can be made of the vast majority of sentencing studies which are based on a single jurisdiction: focusing on a single jurisdiction runs the risk of arriving at results that are the product of idiosyncratic features that may not be representative of other courts from a similar jurisdiction or state. Second, this national sample allowed us to account for geographic region—a factor which has been extensively considered in studies explaining interjurisdictional differences in prison use, but up until now has been absent from contextual sentencing studies.

Thus, when viewed in relation to the cumulative body of sentencing research, these characteristics could be seen as distinct advantages which at least partially mitigate some of the concerns that De Leeuw expresses. It is undeniable (and inevitable) that our modeling technique, sample, and analytic choices vary from other studies. Yet, these would not seem to be reasons for chagrin, given that no single social scientific study can conclusively answer the questions that it poses. Most scientists would surely agree that one study does not make a body of evidence. However, we believe that this study clearly demonstrates that the type of sentence one receives and the reason one receives it partially depend on where it is meted out.

References

Gelman, A. (2004). *Bugs.R: functions for calling Bugs from R.* www.stat.columbia.edu/∼gelman/bugsR/.

Gelman, A. and I. Pardoe (2004a). Average predictive effects for models with nonlinearity, interactions, and variance components. Technical report, Department of Statistics, Columbia University.

Gelman, A. and I. Pardoe (2004b). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. Technical report, Department of Statistics, Columbia University.

Pardoe, I. (2004). Model assessment plots for multilevel logistic regression. *Computational Statistics and Data Analysis 46*, 295–307.

R Development Core Team (2004). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.