COMMENTS ON PARDOE, WEITNER, AND FRASE: SENTENCING CONVICTED FELONS IN THE UNITED STATES: A BAYESIAN ANALYSIS USING MULTILEVEL COVARIATES

JAN DE LEEUW

ABSTRACT. We discuss a paper by Pardoe et al. that uses Bayesian multilevel regression analysis to predict sentencing behavior.

This paper presents a regression analysis of the factors related to sentencing, which is obviously a very complicated and important problem. I will make some very general remarks about the paper, because I consider it to be a representative example of a general approach to statistical analysis. As an example, it looks quite good, although I am not an expert on sentencing. But neither are the authors, I think. Most important, from my point of view, is that the general approach is deeply flawed.

From the methodological point of view, the paper uses techniques which differ from logistic regression analysis popular in criminology. It replaces logistic regression by multilevel logistic regression, and, as a further elaboration, it replaces the conventional multilevel analysis of Wong and Mason [1985] by a fully Bayesian approach. The question we have to answer is simple: does this analysis improve on standard logistic regression analysis, and can we expect the results of the analysis to be more reliable and interesting. Unfortunately, there are several problems with the paper, or more precisely with this type of statistical analysis, that make it difficult to give a positive answer to these questions.

Why do Pardoe et al. (and many other social and behavioral scientists with similar data structures) argue that it makes sense to use multilevel analysis

Date: April 11, 2004.

in this context ? If we look in the multilevel literature for a general answer, we are in for some disappointment. As in various other social science statistics contexts, arguments are often replaced by references to presumably unassailable expert sources. We see this in early discussion of factor analysis, in the literature around LISREL and structural equation models, and now again in multilevel analysis. Pardoe et al. maintain, for instance, that conventional logistic regression techniques do not *correctly* (page 3) or *properly* (page 4) account for the effects of the various covariates. Hierarchical modelling is more *appropriate* (page 4) or even *required* (page 7). But, of course, no such thing is true a priori. We are dealing with an empirical question here, and we need empirical comparisons of different regression analyses results to arrive at an answer.

Multilevel analysis is clearly a form of regression analysis. It is quite customary for social and educational statisticians to wax poetic about being able for the first time to integrate the effects of individual and contextual variation in a single analysis, in the same way as they were able for the first time to model complicated causal systems with LISREL. When push comes to shove, however, multilevel analysis fits a regression model with interaction terms between individual-level and county-level variables, and with a dependence structure in which the covariance of individuals within the same county is a scalar product of the individual-level variates. The first obvious difference with the conventional approach is that potentially we have many more parameters in our regression model, because the multilevel model encourages us to look at interactions. Again, this is something which is typical of much of social science statistics [De Leeuw, 2004]. The more parameters we have, the closer one is able to approximate the "truth", and thus one tends to make the models complicated. This leads to instability, to horrendous model selection problems, and ultimately to non-cumulative science because the results of an analysis can never be replicated.

Of course one can argue, in the Pardoe et al. context, that it is sensible to assume that individuals in the same county have correlated residuals. But this

general argument does not imply anything about the form of the correlations, and it does provide the happy researchers with even more parameters to play with. Moreover there is no way in which we can actually find out if the postulated dependence structure is realistic or not, we simply do not have enough data to study that particular problem.

In the multilevel context, the interesting measure of stability is not the standard error. We have large samples, so the standard errors are bound to be small, and they are computed anyway on the assumption that the model is true. What *is* interesting is stability under model selection, and the likelihood that there exist a huge number of qualitatively different models with approximately the same fit. At the solution we only know we are sitting in a valley corresponding to a local minimum of the deviance, and with complicated models we feel that there could easily be better solutions over the hills. And we don't even mention the solutions that can be found in entirely different landscapes resulting from different deviance functions. The large number of parameters, and the substantial multicollinearity of the cross-level interactions, introduce major instabilities in model selection, even for relatively small models [Kreft and de Leeuw, 1998]. But typically, in multilevel analysis applications, only a single model is considered, and all statistical "inference" is conditional on the appropriateness of this model. This approach, I think, will often lead to throw-away science. See also Berk [2004] for a related discussion.

The second refinement in Pardoe et al. is the use of a fully Bayesian approach. It seems to me that this choice is driven mostly by the fashions of the moment, because no clear arguments are given why full Bayes is better than the more customary empirical Bayes. As in many similar analyses, the authors propose an arbitrary but convenient prior, for which there is no scientific basis, and then argue that the samples are so large that the prior is irrelevant anyway. But this argument undermines the whole reason for doing Bayesian statistics in the first place. One might as well use an equally silly improper prior and find oneself in the dreaded frequentist framework again. A minor point in this context is that the description of the model on

page 8 cannot make up its mind if it is Bayesian or frequentist. No clear notational distinction is made between random variables and fixed parameters. The η parameters are called fixed effects, but of course in a fully Bayesian analysis they are random variables as well, and we only have fixed parameters at the hyper-parameter level.

From the point of view of computational statistics, there is is not much difference between the Bayesian and the frequentist approaches. Bayesians use a general purpose mechanism to smooth their estimation problems and to shrink their parameter estimates, but there is so much freedom in their choice of smoothing parameters that they have to make many arbitrary choices. In addition, Bayesians go one step further than frequentists in replacing the unobservables in their models by random variables. Only one step, however, because the hyperparamers are still fixed constants. It is entirely unclear to me why this minor variation in modelling, combined with a rather clunky stochastic optimization method, leads to so much obnoxious triumphalism. Or, in this particular case, why anyone would prefer the Bayesian approach to a straightforward likelihood approach. The landscape is complicated enough as it is.

As I said at the start of these comments, the problem of sentencing is interesting and extremely important. Pardoe et al. present the results of their analysis in considerable detail at the end of the paper, reviewing some of the relevant literature. Predictably, some of the earlier results agree with their results, and some do not agree. Of course even if the results seem to agree, we are not really sure if they actually agree, because previous authors often define their variables differenty and use very different samples. And different regression techniques as well. It seems to me that the proper reaction to summaries of this and related research is desperation. Or the statistical version of desperation, which is known as meta-analysis. The question I asked at the beginning was if Bayesian multilevel analysis produces results on sentencing which inspire more confidence than previous results. I, for one, don't think it does.

REFERENCES

R. A. Berk. *Regression Analysis. A Constructive Critique.* Number 11 in Advanced Quantitative Techniques in the Social Sciences. Sage Publications, Thousand Oaks, CA, 2004.

J. De Leeuw. Review of J. Hox, Multilevel Analysis. *Journal of Educational Measurement*, (in press), 2004.

G.G. Kreft and J. de Leeuw. *Introduction to Multilevel Modelling*. Sage Publications, Thousand Oaks, CA, 1998.

G.Y. Wong and W.M. Mason. The Hierarchical Logistic Regression Model for Multilevel Analysis. *Journal of the American Statistical Association*, 80:513–524, 1985.

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu