Finding New Information for Ecological Inference Models: A Comment on Jon Wakefield, "Ecological Inference in 2×2 Tables"

Gary King
Department of Government,
Harvard University*

December 4, 2003

Congratulations goes to Jon Wakefield for an unusually complete and completely insightful contribution to this fast-growing literature. Wakefield productively follows what is now standard practice by including both deterministic and statistical information in each new model and then seeking out additional sources of information.

As he clarifies, the great importance of the questions addressed by ecological inference makes any advance highly valuable, and in recognition applied researchers often rapidly adopt statistical innovations. For example, in the legislative redistricting litigation following each American decennial census, ecological inferences are required for implementing the Voting Rights Act, which often lets a political party gerrymander the electoral system in its favor. In earlier redistricting litigation, redistricters and litigants used Goodman's (1953) regression and sometimes the Duncan and Davis (1953) bounds. In the redistricting following the 2000 census, experts, legislators, courts, and litigants in most states used the methods offered in King (1997). I would be surprised if the 2010 redistricting did not make heavy use of methods inspired the ideas in Wakefield's article.

My main question today is whether the convolution model is an appropriate baseline/default or instead makes novel implicit assumptions with important empirical consequences. With data from one table, the only information about the two cell proportions comes from the Duncan-Davis bounds and the relationship in (18). This would seem to imply a flat likelihood over the tomography line, and indeed the maximum likelihood in King (1997) and King, Rosen and Tanner (1999) is a ridge over the respective tomography line.

In contrast, the maximum of the convolution likelihood is not flat; see Figure 6(a). This striking implied claim of information not in other models comes from assumptions made about people within each areal unit and then aggregating. If the assumptions are correct, then the resulting informative likelihood could be extremely valuable in applications. If they are not correct, then instead of constituting a default model that always applies, what Wakefield labels a "likelihood" may instead include parts that should really be in the prior with adjustable parameters; and so the approach may require extensive modification for each application. This too would seem useful, if the information about individuals necessary to make the required assumptions is available even when data on individuals are not. How often is this information available? And to match a reasonable range of

*David Florence Professor of Government, Center for Basic Research in the Social Sciences, Cambridge MA 02138; http://GKing.Harvard.Edu, 617-495-2027, King@Harvard.Edu.

real applications, what modified assumptions and resulting modified convolution models should be considered?

References

Duncan, O. D. and B. Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18:665–666.

Goodman, Leo. 1953. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* 18:663–666.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

King, Gary, Ori Rosen and Martin A. Tanner. 1999. "Binomial-Beta Hierarchical Models for Ecological Inference." *Sociological Methods and Research* 28(1):61–90.