Author’s reply to the discussion of Ecological Inference for 2 × 2 Tables

I would first like to thank all of the discussants for their contributions; my comments will be organised into a set of general themes.

The Likelihood

Many comments concern the convolution likelihood, it is clearly an important element in the assumption-sensitive world of ecological inference, and much of the response will consider its formulation.

Dr. Best argues that “... the binomial likelihood is appropriate in many settings”, but I think this statement requires qualification in an ecological regression setting (as opposed to disease mapping where prediction is often the aim). If the covariate data are available from a survey of individuals within the area, then the binomial model only arises under specific situations.

Consider an example noted by Dr. Best in which the outcome is disease status, and the covariate is smoking. Let the number of cases and non-cases in a single area be denoted \( Y \) and \( N - Y \). Further assume that covariate samples are available from a survey of individuals within the area. Assume the binary exposure, non-smoker/smoker, with a survey of \( m \) individuals, yielding \( m_0 \) non-smokers and \( m_0 \) smokers. Letting \( p_0 \) and \( p_1 \) represent the disease probabilities for non-smokers and smokers, the likelihood of interest is given by

\[
\begin{align*}
\ell(p_0, p_1) &= \Pr(y|m_0, m, p_0, p_1, N) \\
&= \sum_{N_0=m_0}^{N-m_0} \Pr(y|N_0, N, p_0, p_1) \times \Pr(N_0|m_0, m, N) \\
&= \frac{1}{y!} \sum_{N_0=m_0}^{N-m_0} e^{-(N_0p_0 + N_1p_1)} (N_0p_0 + N_1p_1)^y \times \Pr(N_0|m_0, m, N) \quad (1)
\end{align*}
\]

where the convolution is approximated by a Poisson distribution in this rare disease context. We have

\[
\Pr(N_0|m_0, m, N) = \frac{\Pr(m_0|m, N_0, N) \times \Pr(N_0|N)}{\Pr(m_0|m, N)}, \quad (2)
\]

for \( N_0 = m_0, ..., N - m + m_0 \), with

\[
\Pr(m_0|m, N_0, N) = \binom{N_0}{m_0} \binom{N - N_0}{m - m_0} / \binom{N}{m}, \quad (3)
\]

for \( \max(0, m - N + N_0) \leq m_0 \leq \min(N_0, m) \). To evaluate the remaining terms in (2) we turn to the paper of Ericson (1969), highlighted by Dr. Forster, in which exchangeable priors for characteristics in a finite sampling context are described. Here

\[
\Pr(N_0|N) = \int_\phi \Pr(N_0|N, \phi) \pi(\phi) \, d\phi = \binom{N}{N_0} \int_\phi \phi^{N_0} (1 - \phi)^{N - N_0} \, d\phi = \frac{1}{N + 1}
\]

for \( N_0 = 0, 1, ..., N \), and where \( \phi \), the proportion of smokers in a (hypothetical) superpopulation, has been assigned a uniform prior. Evaluating the normalizing constant

\[
\Pr(m_0|m, N_0, N) = \sum_{N_0^*} \Pr(m_0|m, N_0^*, N) \Pr(N_0^*|N)
\]
gives

$$\Pr(N_0|m_0, m, N) = \binom{N_0}{m_0} \binom{N - N_0}{m - m_0} / \sum_{N_0} \binom{N_0}{m_0} \binom{N - N_0}{m - m_0}$$

(see Pearson (1928) for this exact derivation). Hence the likelihood (1) is a weighted sum of Poisson random variables with weights given by (4). The complexity of this likelihood demonstrates the benefit of the estimating functions approach of Prentice and Sheppard (1995) in which the first two moments of $Y$ only (rather than a full likelihood) are used; the above development may also provide insight into this method, with respect to the asymptotic bias in the estimating function induced by the sampling of $m$ ($< N$) covariate values. The Poisson approximation to the naive binomial likelihood is

$$Y|m_0, m, p_0, p_1, N \sim \text{Poisson} \{N[m_0 p_0 + (m - m_0)p_1]/m\}.$$  

By comparison with (1) we see that this is equivalent to assigning a weight of 1 to $N_0 = Nm_0/m$, the mean of the hypergeometric (3), and does not acknowledge the uncertainty in $N_0$, but may be a reasonable approximation if $m$ is large (and smoking is not rare). A binomial model may be appropriate in this example if smoking information came from total tobacco sales in the area, rather than from a survey of individuals.

Professor Besag and Dr. Best highlight the rationale for a binomial sampling model given by Knorr-Held and Besag (1998). A crucial step in this formulation is that the Bernoulli outcomes are independent, and in an ecological regression setting, once covariate information on individuals in the area becomes available (such as the race margin), dependence is induced. Of course, the Knorr-Held and Besag development justifies the use of the binomial distribution within each of the covariate groups; and this formulation emphasizes that there should be no clustering.

It is more difficult to develop a likelihood if the exposure information from a survey consists of continuous measures, though again the work of Ericson (1969) may be useful. To address Best, Jackson and Sylvia Richardson’s discussion of continuous covariates an environmental epidemiology setting, in which exposure information is available from a collection of pollution monitors, will be considered. Let $Z_1, ..., Z_m$ represent exposures from $m$ locations within a single area (and specifically suppose these measurements are not from study individuals), and write $\Pr(Y_1, ..., Y_N|Z_1, ..., Z_m) = \prod_{j=1}^N \Pr(Y_j|Z_1, ..., Z_m)$. Letting $z_j$ denote the exposure of individual $j$ (where we have used lower case letters to distinguish these variables from $Z_j$), we have

$$\Pr(Y_j = 1|Z_1, ..., Z_m) = E_{z_j|Z_1, ..., Z_m} \{\Pr(Y_j = 1|z_j)\}.$$  

An approach in the spirit of Prentice and Sheppard (1995) is to assume $z_j = Z_k$ with probability $1/m$ to give

$$\Pr(Y_j = 1|Z_1, ..., Z_m) = \frac{1}{m} \sum_{k=1}^m \exp(\beta_0 + \beta_1 Z_k),$$  

where we have assumed a log-linear risk exposure model. Alternatively we may assume a distribution, characterized by parameters $\phi$, for the within-area variability in exposure:

$$\Pr(Y_j = 1|\phi) = E_{z_j|\phi} \{\Pr(Y_j = 1|z_j)\}.$$  

Taking an errors-in-variables model approach we may assume a Berkson model (Berkson, 1950), $z_j = \mu + \epsilon_j$, where $\epsilon_j \sim \text{iid} N(0, \sigma^2)$, $j = 1, ..., N$. Let $\bar{z}$ denote the mean and $s^2$ the variance of $Z_1, ..., Z_m$. Then if $m$
is large we have \( \mu \approx \bar{z} \) and \( \sigma^2 \approx s^2 \) and, under normality,

\[
\Pr(Y_j = 1|\phi) = \exp(\beta_0 + \beta_1 \bar{z} + \beta_2 s^2/2),
\]  

(5)

where \( \phi = (\bar{z}, s^2) \). This is another formulation of the approach suggested by Richardson et al. (1987). It is well-known that Berkson errors do not lead to bias in regression coefficient estimates if the variance is independent of the mean (here across areas). If \( s^2 = a + b\bar{z}, \) however, then (5) takes the form \( \exp(\beta_0 + a\beta_1^2/2 + \{\beta_1 + b\beta_1^2/2\}\bar{z}) \) (Wakefield, 2003) so that, if (as usual) the variance increases with the mean \( (b > 0) \) we have over-estimation of a detrimental exposure.

This development demonstrates a point that was emphasized in the paper: to clarify assumptions it is important to build ecological models from the aggregation of individual-level models. This is why the convolution model is natural for surveyed covariates, while the binomial is not, since it cannot be derived from an individual-level model. Professor’s Draper and Waller both discuss the role of standardization, building models at the level of the individual makes clear the situations in which such standardization is valid; specifically problems of mutual standardization (Rosenbaum and Rubin, 1984) may be anticipated.

Professor King questions the rationale for the convolution model as compared to the model of King (1997) which is flat along the tomography line for the unobserved fractions \( \tilde{p}_j = Y_j/N_j, j = 0, 1 \). We show that the predictive distribution for these fractions that arises from the convolution likelihood is flat along the tomography line, under a flat prior. Consider a finite sample approach to the derivation of the posterior \( Y_0, Y_1|N_0, N, y, \) again with the assumption of exchangeability (Ericson, 1969). Let \( S_j, j = 0, 1, \) denote the set \{\max(0, y - N_{1-j}), ..., \min(N_j, y)\} and note \( |S_0| = |S_1| = \min(N_0, N_1, Y, N - Y) + 1 \). Then

\[
\Pr(Y_0, Y_1|N_0, N, y) = \frac{\Pr(y|Y_0, Y_1, N_0, N) \Pr(Y_0, Y_1|N_0, N)}{\Pr(y|N_0, N)}
\]

and \( \Pr(y|Y_0, N_0, N) = I[y = Y_0 + Y_1] \), the indicator function for the event \( y = Y_0 + Y_1 \), so that

\[
\Pr(Y_0, Y_1|N_0, N, y) = \frac{\int_{p_0} \int_{p_1} \Pr(Y_0|p_0, N_0) \Pr(y - Y_0|p_1, N_1) \pi(p_0, p_1) \ dp_0 dp_1}{\int_{p_0} \int_{p_1} \sum_{S_0 \in S_0} \Pr(Y_0^*|p_0, N_0) \Pr(y - Y_0^*|p_1, N_1) \pi(p_0, p_1) \ dp_0 dp_1}
\]  

(6)

for \( Y_0 \in S_0, Y_1 = y - Y_0 \) and where \( p_j = \Pr(Y = 1|X = j) \) within hypothetical superpopulations, \( j = 0, 1 \). Notice that the convolution likelihood is a component of the denominator of (6). If we assume binomial distributions for \( Y_0|p_0, N_0 \) and \( y - Y_0|p_1, N_1 \) and independent uniform priors on \( p_0 \) and \( p_1 \), we obtain \( \Pr(Y_0, Y_1|N_0, N, y) = |S_0|^{-1} \), a uniform distribution, and so the likelihood for the sample fractions, \( \tilde{p}_j \), is uniform along the tomography line (and hence along the bounds, providing a link with a comment of Rice) as in the King (1997) model. This derivation also addresses one of Professor Draper’s questions; see Ericson (1969, Section 2.2) for further discussion of the, “intimate similarities” between sampling distributions and exchangeable prior distributions. Assuming a hierarchical prior \( \pi(p_0, p_1) \) in (6) leads to the model of the paper, showing that the two formulations are identical. It would interesting to explore whether the above offers any advantages, particularly in terms of computation.

The accuracy of the normal approximation to the convolution requires further investigation and I welcome the saddlepoint approximation described by Davison and Semadeni.
Combination of Individual and Ecological Data

I agree with Best, Sylvia Richardson and Jackson that small samples only may be required to supplement ecological data. I have recently been working on combining ecological data with case-control samples in an epidemiological context. This design appears promising, and is reported in detail in Haneuse and Wakefield (2004). A considerable benefit is that if the relevant variables are measured in the case-control sample, it is possible to control for within-area confounding, a major source of ecological bias. Under certain assumptions case-control data from a single area only may alleviate ecological bias (Chambers’ demonstration of this phenomenon is encouraging, see also Iversen, 1973). Supplementing ecological data with case only data alone also enforces identifiability. If case-control samples are available in each area then this also allows the possibility of investigating effect modification by area, as discussed by Best and Richardson. Without individual-level data there is no information to directly estimate these effects, and I would view such a model as part of a sensitivity analysis (as I view the hierarchical models suggested in the paper). Effect modification by area is far more difficult to detect in epidemiology because outcomes are typically far rarer than in a social science context. Individual level data also allow assessment of the assumed form of the response-covariate model. Raghunathan et al. (2003) show the benefits of combining ecological and individual data for the problem originally considered by Robinson (1950).

As Sheppard comments, true contextual effects are uncommon in an environmental epidemiological setting, though may be induced by missing variables that are related to the exposure of interest (Sheppard and Wakefield, 2004).

So far as design is concerned, I welcome the references provided by Professor Atkinson. In an environmental epidemiology setting information matrices may be evaluated in individual, ecological and combined (hybrid) settings, allowing the partition of information to be determined (see Wakefield and Shaddick, 2004, for details); Steel, Beh and Chambers (2004) also give a convenient form for the loss of information in the ecological setting of the paper, which may also be of use in a design context.

Prior Distributions

With respect to Jiang and Tanner’s comments, as mentioned in the paper, I would not universally endorse the Exp(0.01) hyperprior; in an ecological context the priors need to be as data set specific as possible. Jiang and Tanner’s suggestion for prior choice has an empirical Bayes feel and given the assumption of no contextual effects would be inaccurate if such effects are present; priors based on substantive knowledge or additional data are clearly preferable.

For simplicity, I did not consider the use of additional variables beyond race in the Louisiana data but any serious analysis would use such variables. Corder and Worbrecht discuss improvement of ecological inference via the use of second stage covariates. I think this is an important step though if the covariates are associated with the proportions \( x \), prior information is likely to be crucial. If the variability in \((p_{0i}, p_{1i})\) across areas \( i \) can be explained using additional variables (which may be inducing contextual effects), then inference is likely to be much more reliable; such an approach is perhaps in the spirit of Thomsen’s model (for a critique of which see King, 1997,p. 121–122); this model appears dependent on assumptions that are made for mathematical convenience which detracts from its appeal. With additional variables I would again advocate building aggregate models from the level of the individual. Lasserre et al. (1999) examine
the ecological model induced by two binary risk factors in a rare disease setting, and illustrate the aggregate information that is required to identify individual effects.

I appreciate Dr. Rice’s contribution linking his own very interesting work to the ecological inference context, and again illustrating the paucity of information within the margins of a $2 \times 2$ table concerning the odds ratio. Zhu and Reid (1994, Example 2) show the loss in information for the latter when the margins are conditioned upon. Forster’s suggestion of a reparameterization to clarify the amount of information on the odds ratio is deserving of further investigation.

Computation

Professor’s Chambers and Firth both highlight the computational burden of the Markov chain Monte Carlo scheme that is used in the paper for inference. While there will always be a place for simple models that are straightforward to apply, such as ecological regression (at least for exploratory purposes), it is not possible to attach reliable uncertainty measures to point estimates obtained from such approaches (including the method proposed by Chambers and Steel, the more sophisticated method of Brown and Payne, and the recent ad hoc approach of Adolph et al. 2003). Unfortunately the subtlety of information in an ecological setting suggest that for correct propagation of uncertainty an exact method of inference is required.

Thomas Richardson’s link with the counterfactuals literature, and in particular Pearl (2000), is appreciated. A cursory look at this reference reveals that, when converted to a conditional probability, $P(A|B)$, and taking $P(A) = 1 - x$ and $P(B) = q$, expression (9.10) of Pearl (2000) is identical to the bound (14), in Section 3.7 of the paper, but equation (9.9) is derived under specific assumptions which appear less related to the present context. This would be an interesting research direction to investigate, as would the related area of missing data, specifically with respect to tables for which only partial information is available due to confidentiality concerns. Much of the concern for the latter is in developing computationally efficient algorithms (e.g. Dobra et al. 2003), which may be useful in the present context.

In conclusion I would again like to thank all of the discussants for their contributions, and hope that these further stimulate research into the frustrating yet intriguing world of ecological inference.

References


