ecological inference

1 The ecological inference problem

For expository purposes, we discuss only an important but simple special case of ecological inference, and adopt the running example and notation from King (1997: ch. 2). The basic problem has two observed variables \(T_i\) and \(X_i\) and two unobserved quantities of interest \(\beta_i^b\) and \(\beta_i^w\) for each of \(p\) observations. Observations represent aggregate units, such as geographic areas, and each individual-level variable within these units is dichotomous.

To be more specific, in Figure 1 we observe for each electoral precinct \(i(i = 1, \ldots, p)\) the fraction of voting age people who turn out to vote \(T_i\) and who are black \(X_i\), along with the number of voting age people \(N_i\). The quantities of interest, which remain unobserved because of the secret ballot, are the proportions of blacks who vote \(\beta_i^b\) and whites who vote \(\beta_i^w\). The proportions \(\beta_i^b\) and \(\beta_i^w\) are not observed because \(T_i\) and \(X_i\) are from different data sources (electoral results and census data, respectively) and record linkage is impossible (and illegal), and so the cross-tabulation cannot be computed.

Also of interest are the district-wide fractions of blacks and whites who vote, which are respectively

\[
\begin{align*}
B^b &= \frac{\sum_{i=1}^{p} N_i X_i \beta_i^b}{\sum_{i=1}^{p} N_i}, \quad \text{and} \\
B^w &= \frac{\sum_{i=1}^{p} N_i (1 - X_i) \beta_i^w}{\sum_{i=1}^{p} N_i (1 - X_i)}.
\end{align*}
\]

These are weighted averages of the corresponding precinct-level quantities. Some methods aim to estimate only \(B^b\) and \(B^w\) without giving estimates of \(\beta_i^b\) and \(\beta_i^w\) for all \(i\).

**Race of Voting decision**

\[
\begin{array}{ccc}
\text{Race of Voting decision} & \text{Vote} & \text{No Vote} \\
\text{Black} & \beta_i^b & 1 - \beta_i^b \\
\text{White} & \beta_i^w & 1 - \beta_i^w
\end{array}
\]

**Figure 1** Notation for Precinct \(i\). Note: The goal is to estimate the quantities of interest, \(\beta_i^b\) (the fraction of blacks who vote) and \(\beta_i^w\) (the fraction of whites who vote), from the aggregate variables \(X_i\) (the fraction of voting age people who are black) and \(T_i\) (the fraction of people who vote), along with \(N_i\) (the known number of voting age people).
2 Deterministic and statistical approaches

The ecological inference literature before King (1997) was bifurcated between supporters of the method of bounds, originally proposed by Duncan and Davis (1953), and supporters of statistical approaches, proposed even before Ogburn and Goltra (1919) but first formalized by Goodman (1953; 1959). (For the historians of science among us: although these two monumental articles were written by two colleagues and friends in the same year and in the same department and university — the Department of Sociology at the University of Chicago — the principal did not discuss their work prior to completion. Even by today’s standards, nearly a half century after their publication, the articles are models of clarity and creativity.) Although Goodman and Duncan and Davis moved on to other interests following their seminal contributions, most of the ecological inference literature in the five decades since 1953 was an ongoing war between supporters of these two key approaches, and often without the usual academic decorum.

2.1 Extracting deterministic information: the method of bounds

The purpose of the method of bounds and its generalizations is to extract deterministic information, known with certainty, about the quantities of interest.

The intuition behind these quantities is simple. For example, if a precinct contained 150 African-Americans and 87 people in the precinct voted, then how many of the 150 African-American actually cast their ballot? We do not know exactly, but bounds on the answer are easy to obtain: in this case, the answer must lie between 0 and 87. Indeed, conditional only on the data being correct, \([0,87]\) is a 100 per cent confidence interval. Intervals like this are sometimes narrow enough to draw meaningful inferences, and sometimes they are too wide, but the ability to provide (non-trivial) 100 per cent confidence intervals in even some situations is quite rare in any statistical field.

In general, before any data are seen, the unknown parameters \(\beta^B_i\) and \(\beta^w_i\) are each bounded on the unit interval. Once we observe \(T_i\) and \(X_i\), they are bounded more narrowly, as:

\[
\beta^B_i \in \left[ \max \left(0, \frac{T_i - (1-X_i)}{X_i} \right), \ min \left( \frac{T_i}{X_i}, 1 \right) \right]
\]

\[
\beta^w_i \in \left[ \max \left(0, \frac{T_i - X_i}{1 - X_i} \right), \ min \left( \frac{T_i}{1 - X_i}, 1 \right) \right].
\]

Deterministic bounds on the district-level quantities \(B^B\) and \(B^w\) are weighted averages of these precinct-level bounds.

The bounds then indicate that the parameters in each case fall within these deterministic bounds with certainty, and in practice they are almost always narrower than \([0,1]\). Whether they are narrow enough in any one application depends on the nature of the data.

2.2 Extracting statistical information: Goodman’s regression

Leo Goodman’s (1953; 1959) approach is very different from, but just as important as, Duncan and Davis’s. He looked at the same data and focused on the statistical information. His approach examines variation in the marginals \((X_i\) and \(T_i\) over the precincts to attempt to reason back to the district-wide fractions of blacks and whites who vote, \(B^B\) and \(B^w\). The outlines of this approach and the problems with it have been known at least since Ogburn and Goltra (1919). For example, if in precincts with large proportions of black citizens we observe that many people do not vote, then it may seem reasonable to infer that blacks turn out at lower rates than whites. Indeed, it often is reasonable, but not always. The problem is that it could instead be the case that the whites who happen to live in heavily black precincts are the ones who vote less frequently, yielding the opposite ecological inference to the individual-level truth.

What Goodman accomplished was to formalize the logic of the approach in a simple regression model, and to give the conditions under which estimates from such a model are unbiased. To see this, note first that the accounting identity

\[
T_i = X_i \beta^B_i + (1 - X_i) \beta^w_i
\]

holds exactly. Then he showed that a regression of \(T_i\) on \(X_i\) and \((1 - X_i)\) with no constant term could be used to estimate \(B^B\) and \(B^w\), respectively. The key assumption necessary for unbiasedness that Goodman identified is that the parameters and \(X_i\) be uncorrelated:

\[
\text{Cov}(\beta^B_i, X_i) = \text{Cov}(\beta^w_i, X_i) = 0.
\]

In the example, the assumption is that blacks vote in the same proportions in homogeneously black areas as in more integrated areas. Obviously, this is true sometimes and it is false other times. (King, 1997: ch. 3, showed that Goodman’s assumption was necessary but not sufficient. To have unbiasedness, it must also be true that the parameters and \(N_i\) are uncorrelated.)

As Goodman recognized, when this key assumption does not hold, estimates from the model will be biased. Indeed, they can be very biased, outside the deterministic bounds, and even outside the unit interval. This technique has been used extensively since the 1950s, and impossible estimates occur with considerable frequency (some estimates range to a majority of real applications; Achen and Shively, 1995).

3 Extracting both deterministic and statistical information: King’s EI approach

From 1953 until 1997, the only two approaches used widely in practice were the method of bounds and
Goodman’s regression. King’s (1997) idea was that the insights from these two conflicting literatures in fact do not conflict with each other; the sources of information are largely distinct and can be combined to improve inference overall and synergistically. The idea is to combine the information from the bounds, applied to both quantities of interest for each and every precinct, with a statistical approach for extracting information within the bounds. The amount of information in the bounds depends on the data-set, but for many data-sets it can be considerable. For example, if precincts are spread uniformly over a scatterplot of \( X_i \) by \( T_i \), the average bounds on \( \beta_i^b \) and \( \beta_i^w \) are narrowed from \([0,1]\) to less than half of that range – hence eliminating half of the ecological inference problem with certainty. This additional information also helps make the statistical portion of the model far less sensitive to assumptions than previous statistical methods which exclude the information from the bounds.

To illustrate these points, we first present all the information available without making any assumptions, thus extending the bounds approach as far as possible. As a starting point, the left graph in Figure 2 provides a scatterplot of a sample data set as observed, \( X_i \) horizontally by \( T_i \) vertically. Each point in this figure corresponds to one precinct, for which we would like to estimate the two unknowns. We display the unknowns in the right graph by using the linear relationship in eq. (5). A key point is that each dot on the left graph can be expressed, without assumptions or loss of information, as what King called a ‘tomography’ line within the unit square in the right graph. It is precisely the information lost due to aggregation that causes us to have to plot an entire line (on which the true point must fall) rather than the goal of one point for each precinct on the right graph. In fact, the information lost is equivalent to having a graph of the \( \beta_i^b \) by \( \beta_i^w \) points but having the ink smear, making the points into lines and partly but not entirely obscuring the correct positions of the \((\beta_i^b, \beta_i^w)\) points. (King also showed that the ecological inference problem is mathematically equivalent to the ill-posed ‘tomography’ problem of many medical imaging procedures, such as CAT and PET scans, where one attempts to reconstruct the inside of an object by passing X-rays through it and gathering information only from the outside. Because the line sketched out by an X-ray is closely analogous to eq. (5), King labels the latter a tomography line and the corresponding graph a tomography graph.)

What does a tomography line tell us? Before we know anything, we know that the true \((\beta_i^b, \beta_i^w)\) point must lie somewhere within the unit square. After \( X_i \) and

\[
\beta_i^w = \left( \frac{T_i}{1 - X_i} \right) - \left( \frac{X_i}{1 - X_i} \right) \beta_i^b, \tag{5}
\]

which shows that \( \beta_i^w \) is a linear function of \( \beta_i^b \), where the intercept and slope are known (since they are functions of the data, \( X_i \) and \( T_i \)).

King then maps the knowns from the left graph onto the right graph by using the linear relationship in eq. (5). A key point is that each dot on the left graph can be expressed, without assumptions or loss of information, as what King called a ‘tomography’ line within the unit square in the right graph. It is precisely the information lost due to aggregation that causes us to have to plot an entire line (on which the true point must fall) rather than the goal of one point for each precinct on the right graph. In fact, the information lost is equivalent to having a graph of the \( \beta_i^b \) by \( \beta_i^w \) points but having the ink smear, making the points into lines and partly but not entirely obscuring the correct positions of the \((\beta_i^b, \beta_i^w)\) points. (King also showed that the ecological inference problem is mathematically equivalent to the ill-posed ‘tomography’ problem of many medical imaging procedures, such as CAT and PET scans, where one attempts to reconstruct the inside of an object by passing X-rays through it and gathering information only from the outside. Because the line sketched out by an X-ray is closely analogous to eq. (5), King labels the latter a tomography line and the corresponding graph a tomography graph.)

What does a tomography line tell us? Before we know anything, we know that the true \((\beta_i^b, \beta_i^w)\) point must lie somewhere within the unit square. After \( X_i \) and

![Figure 2](image1.png)  
**Figure 2** Two views of the same data. Note: The left graph is a scatterplot of the observables, \( X_i \) by \( T_i \). The right graph displays this same information as a tomography plot of the quantities of interest, \( \beta_i^b \) by \( \beta_i^w \). Each precinct \( i \) that appears as a point in the left graph is a line (rather than a point because of information lost due to aggregation) in the right graph. For example, precinct 52 appears as the dot with a little square around it in the left graph and the dark line in the right graph. Source: The data are from King (1997: Figures 5.1 and 5.5).
$T_i$ are observed for a precinct, we also know that the true point must fall on a specific line represented by eq. (5) and appearing in the tomography plot in Figure 2. In many cases narrowing the region to be searched for the true point from the entire square to the one line in the square can provide a significant amount of information. To see this, consider the point enclosed in a box in the left graph, and the corresponding dark line in the right graph. This precinct, number 52, has observed values of $X_{52} = 0.88$ and $T_{52} = 0.19$. As a result, substituting into eq. (5) gives $\beta_i^w = 1.58 - 7.33 \beta_i^b$, which when plotted appears as the dark line on the right graph. This particular line tells us that, in our search for the true $\beta_i^b, \beta_i^w$ point on the right graph, we can eliminate with certainty all area in the unit square except that on the line, which is clearly an advance over not having the data. Translated into the quantities of interest, this line tells us (by projecting the line downward to the horizontal axis) that, wherever the true point falls on the line, $\beta_i^w$ must fall in the relatively narrow bounds of $[0.07, 0.21]$. Unfortunately, in this case, $\beta_i^b$ can only be bounded (by projecting to the left) to somewhere within the entire unit interval. More generally, lines that are relatively steep, like this one, tell us a great deal about $\beta_i^w$ and little about $\beta_i^b$. Tomography lines that are relatively flat give narrow bounds on $\beta_i^w$ and wide bounds on $\beta_i^b$. Lines that cut off the bottom left (or top right) of the figure give narrow bounds on both quantities of interest.

If the only information available to learn about the unknowns in precinct $i$ is $X_i$ and $T_i$, a tomography line like that in Figure 2 exhausts all this available information. This line immediately tells us the known bounds on each of the parameters, along with the precise relationship between the two unknowns, but it is not sufficient to narrow in on the right answer any further. Fortunately, additional information exists in the other observations in the same data set ($X_j$ and $T_j$ for all $i \neq j$) which, under the right assumptions, can be used to learn more about $\beta_i^b$ and $\beta_i^w$ in our precinct of interest.

In order to borrow statistical strength from all the precincts to learn about $\beta_i^b$ and $\beta_i^w$ in precinct $i$, some assumptions are necessary. The simplest version of King's model (that is, the one most useful for expository purposes) requires three assumptions, each of which can be relaxed in different ways.

First, the set of $(\beta_i^b, \beta_i^w)$ points must fall in a single cluster within the unit square. The cluster can fall anywhere within the square; it can be widely or narrowly dispersed or highly variable in one unknown and narrow in the other; and the two unknowns can be positively, negatively, or not at all correlated over $i$. An example that would violate this assumption would be two or more distinct clusters of $(\beta_i^b, \beta_i^w)$ points, as might result from subsets of observations with fundamentally different data generation processes (such as from markedly different regions). The specific mathematical version of this one-cluster assumption is that $\beta_i^b$ and $\beta_i^w$ follow a truncated bivariate normal density

$$TN(\beta_i^b, \beta_i^w | \tilde{\beta}, \tilde{\Sigma}) = N(\beta_i^b, \beta_i^w | \tilde{\beta}, \tilde{\Sigma}) \frac{1(\beta_i^b, \beta_i^w)}{R(\tilde{\beta}, \tilde{\Sigma})},$$

(6)

where the kernel is the untruncated bivariate normal,

$$N(\beta_i^b, \beta_i^w | \tilde{\beta}, \tilde{\Sigma}) = (2\pi)^{-1/2} |\tilde{\Sigma}|^{-1/2} \exp \left[-\frac{1}{2} (\beta_i - \tilde{\beta})' \tilde{\Sigma}^{-1} (\beta_i - \tilde{\beta}) \right],$$

(7)

and $1(\beta_i^b, \beta_i^w)$ is an indicator function that equals 1 if $\beta_i^b \in [0, 1]$ and $\beta_i^w \in [0, 1]$ and zero otherwise. The normalization factor in the denominator, $R(\tilde{\beta}, \tilde{\Sigma})$, is the volume under the untruncated normal distribution above the unit square:

$$R(\tilde{\beta}, \tilde{\Sigma}) = \int_0^1 \int_0^1 N(\beta^b, \beta^w | \tilde{\beta}, \tilde{\Sigma}) d\beta^b d\beta^w$$

(8)

When divided into the untruncated normal, this factor keeps the volume under the truncated distribution equal to 1. The parameters of the truncated density, which we summarize as

$$\psi = \{\tilde{\beta}^b, \tilde{\beta}^w, \sigma_b, \sigma_w, \rho\} = \{\tilde{\beta}, \tilde{\Sigma}\},$$

(9)

are on the scale of the untruncated normal (and so, for example, $\tilde{\beta}^b$ and $\tilde{\beta}^w$ need not be constrained to the unit interval even though $\beta_i^b$ and $\beta_i^w$ are constrained by this density).

The second assumption, which is necessary to form the likelihood function, is the absence of spatial autocorrelation: conditional on $X_i, T_i$ and $T_j$ are mean independent. Violations of this assumption in empirically reasonable (and even some unreasonable) ways do not seem to induce much if any bias.

The final, and by far the most critical, assumption is that $X_i$ is independent of $\beta_i^b$ and $\beta_i^w$. The three assumptions together produce what has come to be known as King’s ‘basic’ EI model. (The use of EI to name this method comes from the name of his software, available at http://GKing.Harvard.edu.) King also generalizes this assumption, in what has come to be known as the ‘extended’ EI model, by allowing the truncated normal parameters to vary as functions of measured covariates, $Z_i^o$ and $Z_i^w$, giving:

$$\tilde{\beta}_i^b = [\phi_1 (\bar{\sigma}_b^2 + 0.25) + 0.5] + (Z_i^o - \bar{Z}^o) x^b$$

$$\tilde{\beta}_i^w = [\phi_2 (\bar{\sigma}_w^2 + 0.25) + 0.5] + (Z_i^w - \bar{Z}^w) x^w$$

(10)

where $x^b$ and $x^w$ are parameter vectors to be estimated along with the original model parameters and that have
as many elements as $Z_i^b$ and $Z_i^w$ have columns. This relaxes the mean independence assumptions to:

$$E(\beta_i^b|X_i, Z_i) = E(\beta_i^b|Z_i)$$
$$E(\beta_i^w|X_i, Z_i) = E(\beta_i^w|Z_i).$$

Note that this extended model also relaxes the assumptions of truncated bivariate normality, since there is now a separate density being assumed for each observation. Because the bounds, which differ in width and information content for each $i$, generally provide substantial information, even $X_i$ can be used as a covariate in $Z_i$. (The recommended default setting in EI includes $X_i$ as a covariate with a prior on its coefficient.) In contrast, under Goodman's regression, which does not include information in the bounds, including $X_i$ leads to an unidentified model (King, 1997: sec. 3.2).

These three assumptions — one cluster, no spatial autocorrelation, and mean independence between the regressor and the unknowns conditional on $X_i$ and $Z_i$ — enable one to compute a posterior (or sampling) distribution of the two unknowns in each precinct. A fundamentally important component of EI is that the quantities of interest are not the parameters of the likelihood but instead come from conditioning on $T_i$ and producing a posterior for $\beta_i^b$ and $\beta_i^w$ in each precinct. Failing to condition on $T_i$ and examining the parameters of the truncated bivariate normal only makes sense if the model holds exactly and so is much more model-dependent than King's approach. Since the most important problem in ecological inference modelling is precisely model misspecification, failing to condition on $T$ assumes away the problem without justification. This point is widely regarded as a critical step in applying the EI model (Adolph and King, with Herron and Shotts, 2003).

When bounds are narrow, EI model assumptions do not matter much. But, for precincts with wide bounds on a quantity of interest, inferences can become model dependent. This is especially the case with ecological inference problems precisely because of the loss of information due to aggregation. In fact, this loss of information can be expressed by noting that the joint distribution of $\beta_i^b$ and $\beta_i^w$ cannot be fully identified from the data without some untestable assumptions. To be precise, distributions with positive mass over any curve or combination of curves that connects the bottom left point ($\beta_i^b = 0, \beta_i^w = 0$) to the top right point ($\beta_i^b = 1, \beta_i^w = 1$) of a tomography plot cannot be rejected by the data (King, 1997: 191). Other features of the distribution are estimable. This fundamental indeterminacy is, of course, a problem because it prevents pinning down the quantities of interest with certainty, but it can also be something of an opportunity since different distributional assumptions can lead to the same estimates, especially since only those pieces of the distributions above the tomography lines are used in the final analysis.

4 Alternative approaches to ecological inference

In the continuing search for more information to bring to bear on ecological inferences, King, Rosen and Tanner (1999) extend King’s (1997) model another step. They incorporate King’s main advance of combining deterministic and statistical information but begin modelling a step earlier at the individuals who make up the counts. They also build a hierarchical Bayesian model, using easily generalizable Markov chain Monte Carlo (MCMC) technology (Tanner, 1996).

To define the model formally, let $T_i$ denote the number of voting age people who turn out to vote. At the top level of the hierarchy they assume that $T_i$ follows a binomial distribution with probability equal to $\theta_i = X_i\beta_i^b + (1 - X_i)\beta_i^w$ and count $N_i$. Note that at this level it is assumed that the expectation of $T_i$, rather than $T_i$, is equal to $X_i\beta_i^b + (1 - X_i)\beta_i^w$. In other words, King (1997) models $T_i$ as a continuous proportion, whereas King, Rosen, and Tanner (1999) recognize the inherently discrete nature of the counts of voters that go into computing this proportion. The two models are connected, of course, since $T_i/N_i$ approaches $T_i$ as $N_i$ gets large.

The connection to King's tomography line can be seen in the contribution of the data from precinct $i$ to the likelihood, which is

$$(X_i\beta_i^b + (1 - X_i)\beta_i^w)^T_i (1 - X_i \beta_i^b - (1 - X_i)\beta_i^w)^{(N_i - T_i)}.$$

By taking the logarithm of this contribution to the likelihood and differentiating with respect to $\beta_i^b$ and $\beta_i^w$, King, Rosen and Tanner show that the maximum of (11) is not a unique point, but rather a line whose equation is given by the tomography line in eq. (5). Thus, the log-likelihood for precinct $i$ looks like two playing cards leaning against each other. As long as $T_i$ is fixed and bounded away from 0.5 (and $X_i$ is a fixed known value between 0 and 1), the derivative at this point is seen to increase with $N_i$, that is, the pitch of the playing cards increases with the sample size. In other words, for large $N_i$, the log-likelihood for precinct $i$ degenerates from a surface defined over the unit square into a single playing card standing perpendicular to the unit square and oriented along the corresponding tomography line.

At the second level of the hierarchical model, $\beta_i^b$ is distributed as a beta density with parameters $c_b$ and $d_b$ and $\beta_i^w$ follows an independent beta with parameters $c_w$ and $d_w$. While $\beta_i^b$ and $\beta_i^w$ are assumed a priori independent, they are a posteriori dependent. At the third and final level of the hierarchical model, the unknown parameters $c_b, d_b, c_w,$ and $d_w$ follow an exponential distribution with a large mean.

A key advantage of this model is that it generalizes immediately to arbitrarily large $R \times C$ tables. This approach was pursued by Rosen et al. (2001), who also
provided a much faster method of moment-based estimator. For an application, see King et al. (2003).


GARY KING, ORI ROSEN AND MARTIN TANNER

Bibliography


econometrics

1 What is econometrics?

Broadly speaking, econometrics aims to give empirical content to economic relations for testing economic theories, forecasting, decision making, and for ex post decision/policy evaluation. The term 'econometrics' appears to have been first used by Pawel Ciompa as early as 1910, although it is Ragnar Frisch who takes the credit for coining the term, and for establishing it as a subject in the sense in which it is known today (see Frisch, 1936, p. 95, and Bjerkholt, 1995). By emphasizing the quantitative aspects of economic relationships, econometrics calls for a 'unification' of measurement and theory in economics. Theory without measurement can have only limited relevance for the analysis of actual economic problems; while measurement without theory, being devoid of a framework necessary for the interpretation of the statistical observations, is unlikely to result in a satisfactory explanation of the way economic forces interact with each other. Neither 'theory' nor 'measurement' on its own is sufficient to further our understanding of economic phenomena.

As a unified discipline, econometrics is still relatively young and has been transforming and expanding very rapidly since an earlier version of this article was published in the first edition of *The New Palgrave: A Dictionary of Economics* in 1987 (Pesaran, 1987a). Major advances have taken place in the analysis of cross-sectional data by means of semiparametric and nonparametric techniques. Heterogeneity of economic relations across individuals, firms and industries is increasingly acknowledged, and attempts have been made to take them into account either by integrating out their effects or by modelling the sources of heterogeneity when suitable panel data exists. The counterfactual considerations that underlie policy analysis and treatment evaluation have been given a more satisfactory foundation. New time series econometric techniques have been developed and employed extensively in the areas of macroeconometrics and finance. Nonlinear econometric techniques are used increasingly in the analysis of cross-section and time-series observations. Applications of Bayesian techniques to econometric problems have been given new impetus largely thanks to advances in computer power and computational techniques. The use of Bayesian techniques has in turn provided the investigators with a unifying framework where the tasks of forecasting, decision making, model evaluation and learning can be considered as parts of the same interactive and iterative process; thus paving the way for establishing the foundation of 'real time econometrics'. See Pesaran and Timmermann (2005a).

This article attempts to provide an overview of some of these developments. But to give an idea of the extent to which econometrics has been transformed over the past