On “Solutions” to the Ecological Inference Problem

by

D. A. Freedman
Statistics Department
U. C. Berkeley, CA 94720

S. P. Klein
RAND Corporation
1700 Main Street
Santa Monica, CA 90401.

M. Ostland
Statistics Department
U. C. Berkeley, CA 94720

M. Roberts
Economics Department
U. C. Berkeley, CA 94720

Abstract

In his 1997 book, King announced “A Solution to the Ecological Inference Problem”. This review discusses King’s method, and tests it on data where truth is known. In the test data, his method produces results that are far from truth, and diagnostics are unreliable. Ecological regression makes estimates that are similar to King’s, while the neighborhood model is more accurate. His announcement is premature.

Introduction

Before discussing King’s book, we explain the problem of “ecological inference”. Suppose, for instance, that in a certain precinct there are 500 registered voters of whom 100 are hispanic and 400 are non-hispanic. Suppose too that a hispanic candidate gets 90 votes in this precinct. (Such data would be available from public records.) How many of the votes for the hispanic candidate came from the hispanics? That is a typical ecological-inference problem. The secrecy of the ballot box prevents a direct solution, so indirect methods are used.

This review will compare three methods for making ecological inferences. First and easiest is the “neighborhood model”. This model makes its estimates by assuming that, within a precinct, ethnicity has no influence on voting behavior: in the example, of the 90 votes for the hispanic candidate, $90 \times 100/(100+400) = 18$ are estimated to come from the hispanic voters. The second method to consider is “ecological regression”, which requires data on many precincts (indexed by i). Let n_i^h be the number of hispanics in precinct i, and n_i^a the number of non-hispanics; let v_i be the number of votes for the hispanic candidate. (The superscript a is for “anglo”; this is only a mnemonic.) If our example precinct is indexed by $i = 1$, say, then $n_1^h = 100$, $n_1^a = 400$, and $v_1 = 90$. Ecological regression is based on the “constancy assumption”: there is a fixed propensity p for hispanics to vote for the hispanic candidate and another fixed propensity q for non-hispanics to vote for that candidate. These propensities are fixed in the sense of being constant across precincts.

On this basis, the expected number of votes for the hispanic candidate in precinct i is $pn_i^h + qn_i^a$. Then p and q can be estimated by doing some kind of regression of v on n^h and n^a.

More recently, King published "a solution to the ecological inference problem". His method will be sketched now, with a more detailed treatment below. In precinct i, the hispanics have propensity p_i to vote for the hispanic candidate, while the non-hispanics have propensity q_i: the number of votes for the hispanic candidate is then $v_i = p_i n_i^h + q_i n_i^a$. The precinct-specific propensities p_i and q_i are assumed to vary independently from precinct to precinct, being drawn at random from a fixed bivariate distribution—fixed in the sense that the same distribution is used for every precinct. (That replaces the "constancy assumption" of ecological regression.) The bivariate distribution is assumed to belong to a family of similar distributions, characterized by a few unknown parameters. These parameters are estimated by maximum likelihood, and then the precinct-level propensities p_i and q_i can be estimated too.

According to King, his "basic model is robust to aggregation bias" and "offers realistic estimates of the uncertainty of ecological estimates". Moreover, "all components of the proposed model are in large part verifiable in aggregate data" using "diagnostic tests to evaluate the appropriateness of the model to each application [pp. 19–20]". The model is validated on two main data sets, in chapters 10 and 11:

- registration by race in 275 southern counties,
- poverty status by sex in 3187 block groups in South Carolina.

In the South Carolina data, "there are high levels of aggregation bias [p. 219]", but "even in this data set, chosen for its difficulty in making ecological inferences, the inferences are accurate [p. 225]". Chapter 13 considers two additional data sets: voter turnout in successive years in Fulton county, Georgia, and literacy by race and county in the U. S. in 1910. Apparently, the model succeeds in the latter example if two thirds of the counties are eliminated (p. 243). A fifth data set, voter turnout by race in Louisiana, is considered briefly on pp. 22–33.

King contends that (i) his method works even if the assumptions are violated, and (ii) his diagnostics will detect the cases where assumptions are violated. With respect to claim (i), the method should of course work when its assumptions are satisfied. Furthermore, the method may work when assumptions are violated—but it may also fail, as we show by example. With respect to claim (ii), the diagnostics do not reliably identify cases where assumptions are problematic. Indeed, we give examples where the data satisfy the diagnostics but the estimates are seriously in error. In other examples, data are generated according to the model but the diagnostics indicate trouble.

We apply King's method, and three of his main diagnostics, to several data sets where truth is known:

- an exit poll in Stockton where the unit of analysis is the precinct,
- demographic data from the 1980 census in Los Angeles county where the unit of analysis is the tract, and
- registration data from the 1988 general election in Los Angeles county, aggregated to the tract level.

In these cases, as in King's examples discussed above, individual-level data are available and truth is known. We aggregate the data, deliberately losing (for the moment) information about individuals, and then use three methods to make ecological inferences:

(i) the neighborhood model,

(ii) ecological regression,
(iii) King's method.

The inferences being made, they can be compared to truth. Moreover, King's method can be compared to existing methods for ecological inference. King's method (estimation, calculation of standard errors, and diagnostic plots) is implemented in the software package EZIDOS—version 1.31 dated 8/22/97—which we downloaded in fall 1997 from his web page after publication of the book. We used this software for Tables 1 and 2 below.

The test data

The exit poll was done in Stockton during the 1988 presidential primary; the outcome measure is hispanic support for Jackson: data were collected on 1867 voters in 39 sample precincts. The data set differs slightly from the one used in Freedman et al. (1991) or Klein et al. (1991). The other data sets are based on 1409 census tracts in Los Angeles county, using demographic data from the 1980 census and registration data from the 1988 general election. Tracts that were small, or had inconsistent data, were eliminated; again, the data differ slightly from those in Freedman et al. (1991). The "high hispanic" tracts have more than 25% hispanics. The outcome measures on the demographic side are percent with high school degrees, percent with household incomes of $20,000 a year or more, percent living in owner-occupied housing units. We also consider registration in the democratic party. For demographic data, the base is citizen voting age population, and there are 314 high-hispanic tracts. For registration data, the base is registered voters, and there are 271 high-hispanic tracts.

Two artificial data sets were generated using King's model, in order to assess the quality of the diagnostics when the model is correct. In Stockton, for instance, King's software was used to fit his model to the real exit poll data, and estimated parameters were used to generate an artificial data set. In these data, King's assumptions hold by construction. The artificial data were aggregated, and run through the three estimation procedures. A similar procedure was followed for the registration data in Los Angeles (all 1409 tracts).

Empirical results

In Stockton, ecological regression gives impossible estimates: 109% of the hispanics supported Jesse Jackson for president in 1988. King's method gives estimates that are far from truth, but the SE is large too (Table 1). In the Los Angeles data, King's method gives essentially the same estimates as ecological regression. These estimates are seriously wrong, and the standard errors are much too small. For example, 55.6% of hispanics in Los Angeles are high school graduates. King's model estimates 30.1%, with an SE of 1.1%: the model is off by 23.2 SEs. The ecological regression estimate of 30.7% is virtually the same as King's, while the neighborhood model does noticeably better at 65.1%. As discussed below, the diagnostics are mildly suggestive of model failure, with indications that the high-hispanic tracts are different from others. So, we looked at tracts that are more than 25% hispanic (compare King, pp. 241ff). The diagnostic plots for the restricted data were unremarkable, but King's estimates were off by 8.1 percentage points, or 6.8 SEs. For these tracts, ecological regression does a little worse than King, while the neighborhood model is a bit better. Other lines in the table can be interpreted in the same way.

Table 1. Comparison of three methods for making ecological inferences, in situations where the truth is known. King's method gives an estimate and a standard error, reported in the format "estimate ± SE", and

$$Z = (\text{King's estimate} - \text{Truth})/\text{SE}.$$

	Neighborhood Model	Ecological Regression	King's Method	Truth	Z
Stockton					
Exit Poll	46%	109%	61% ± 18%	35%	+1.4
Artificial data	39%	36%	40% ± 15%	56%	−1.1
Los Angeles					
Education	65.1%	30.7%	30.1% ± 1.1%	55.6%	−23.2
High hispanic	55.8%	38.9%	40.4% ± 1.2%	48.5%	−6.8
Income	48.5%	31.5%	32.9% ± 1.2%	48.8%	−13.2
Ownership	56.7%	51.7%	49.0% ± 1.5%	53.6%	−3.1
Party affiliation	65.0%	85.7%	90.8% ± 0.5%	73.5%	+34.6
Artificial data	67.2%	90.3%	90.3% ± 0.5%	89.5%	+1.6
High hispanic	73.4%	90.1%	90.3% ± 0.5%	81.0%	+18.6

Diagnostics

We examined plots of $E\{t|x\}$ vs x as in King (p. 206) and "bias plots" of the estimated p or q vs x as in King (p. 183). We also examined "tomography plots" as in King (p. 176); these were generally unrevealing. The diagnostics will be defined more carefully below, and some examples will be given. In brief, x is the fraction of hispanics in each area and t is the response: the $E\{t|x\}$ plot, for instance, shows the data and confidence bands derived from the model. In the Stockton exit poll data set, the $E\{t|x\}$ plot looks fine. The estimated p vs x plot has a significant slope, of about 0.6. To calibrate the diagnostics, we used artificial data generated from King's model as fitted to the exit poll. Diagnostic plots indicated no problems, but the software generated numerous error messages, for instance,

> Warning: Some bounds are very far from distribution mean. Forcing 36 simulations to their closest bound.

(Similar warning messages were generated for the real data.)

We turn to Los Angeles. In the education data, there is a slight nonlinearity in the $E\{t|x\}$ figure—the data are too high at the right. Furthermore, there is a small but significant slope in the bias plot of estimated p vs x. In the high-hispanic tracts, by contrast, the diagnostic plots are fine. For income and ownership, the diagnostics are unremarkable; there is a small but significant slope in the plot of estimated p vs x, for instance, .05 ± .02 for ownership. For party affiliation, heterogeneity is visible in the scatter plot, with a cluster of tracts that have a low proportion of hispanics but are highly democratic in registration. (These tracts are in South-Central Los Angeles, with a high concentration of black voters.) Heterogeneity is barely detectable in the tomography plot. The plot of $E\{t|x\}$ is problematic: most of the tracts are above their expected responses. An artificial data set was constructed to satisfy King's assumptions, but the $E\{t|x\}$ plot looked like the one for the real data. In the high-hispanic tracts, the diagnostic plots are unrevealing. Our overall judgments on the diagnostics for the various data sets are shown in Table 2.

Summary on diagnostics

The diagnostics are quite subjective, with no clear guidelines as to when King's model should *not* be used. Of course, some degree of subjectivity may be inescapable. In several data sets where estimates are far from truth, diagnostics are passed. On the other hand, the diagnostics indicate problems where none exist, in artificial data generated according to the assumptions of the model. Finally, when diagnostics are passed, standard errors produced by the model do not reliably indicate the magnitude of the actual errors (Tables 1 and 2).

Summary of empirical findings

Table 2 shows for each data set which method comes closer to truth. For the artificial registration data in Los Angeles, generated to satisfy the assumptions of King's model, his method ties with ecological regression and beats the neighborhood model. Likewise, his model wins on the artificial data set generated from the Stockton exit poll. Paradoxically, his diagnostics suggest trouble in these two data sets. In all the real data sets, even those selected to pass the diagnostics, the neighborhood model prevails. The neighborhood model was introduced to demonstrate the power of assumptions in determining statistical estimates from aggregate data, not as a substantive model for group behavior (Freedman et al., 1991, pp. 682, 806; compare King, pp. 43–44). Still, the neighborhood model handily outperforms the other methods, at least in our collection of data sets.

Table 2. Which estimation procedure comes closest to truth?

Data set	Neighborhood Model	Ecological Regression	King's Method	King's Diagnostics
Stockton				
Exit Poll	x			Fails bias plot
Artificial data			x	Warning messages
Los Angeles				
Education	x			Marginal $E\{t|x\}$ plot
High hispanic	x			Passes
Income	x			Passes
Ownership	x			Passes
Party affiliation	x			Fails $E\{t|x\}$ plot
Artificial data		x	x	Fails $E\{t|x\}$ plot
High hispanic	x			Passes
Number of wins	7	1	2	

There is some possibility of error in EZIDOS. In the Los Angeles party affiliation data (1409 tracts), the mean non-hispanic propensity to register democratic is estimated by King's software as

37%, while 56% is suggested by our calculations based on his model. Such an error might explain paradoxical results obtained from the diagnostics. There is a further numerical issue: although the diagnostics that we consulted do not pick up the problem, the covariance matrix for the parameter estimates is nearly singular.

Counting success

King (p. xvii) claims that his method has been validated in a "myriad" comparisons between estimates and truth; on p. 19, the number of comparisons is said to be "over sixteen thousand". However, as far as we can see, King tests the model only on five data sets. Apparently, the figure of sixteen thousand is obtained by considering each geographical area in each data sets. For instance, "the first application [to Louisiana data on turnout by race] provides 3262 evaluations of the ecological inference model presented in [the] book—67 times as many comparisons between estimates from an aggregate model and truth as exist in the entire history of ecological inference research. [p. 22]" The Louisiana data may indeed cover 3262 precincts. However, if our arithmetic is correct, to arrive at sixteen thousand comparisons, King must count each area twice—once for each of the two groups about whom inferences are being made.

We do not believe that King's counting procedure is a good one, but let us see how it would apply to Table 1. In the education data, for instance, the neighborhood model is more accurate than King's model in 1133 out of 1409 tracts. That represents 1133 failures for King's model. Moreover, King provides 80% confidence intervals for tract-level truth. But these intervals cover the parameters only 20% of the time—another 844 failures, since $(0.80 - 0.20) \times 1409 = 844$. In the education data alone, King's approach fails two thousand times for the hispanics, never mind the non-hispanics. On this basis, Table 1 provides thousands of counterexamples to the theory. Evidently, King's way of summarizing comparisons is not a good one. What seems fair to say is only this: his model works on some data sets but not others; nor do the diagnostics indicate which are which.

A checklist

In chapter 16, King has "a concluding checklist". However, this checklist does not offer any very specific guidance in thinking about when or how to use the model. For instance, the first point advises the reader to "begin by deciding what you would do with the ecological inferences once they were made". The last point is that "it may also be desirable to use the methods described in . . . Chapter 15", but that chapter only "generalize[s] the model to tables of any size and complexity". See pp. 263, 277, and 291.

Other literature

Robinson (1950) documented the bias in ecological correlations. Goodman (1953, 1959) showed that with the constancy assumption, ecological inference was possible: otherwise, misleading results could easily be obtained. For current perspectives from the social sciences, see Achen and Shively (1995); Tam (1998) gives a number of empirical results like the ones described here. The validity of the constancy assumption for hispanics is addressed, albeit indirectly, by Massey (1981), Massey and Denton (1985), or Lieberson and Waters (1988), among others. Skerry (1995) discusses recent developments. For more background and pointers to the extensive literature, see Klein and Freedman (1993).

Some details

Let i index the units to be analyzed (precincts, tracts, and so forth). Let n_i^h be the number of hispanics in area i, and n_i^a the number of non-hispanics. These quantities are known. The total population in area i is then $n_i = n_i^h + n_i^a$. The population may be restricted to those interviewed in an exit poll, or to citizens of voting age as reported on census questionnaires, among other possibilities. Let v_i be the number of responses in area i, for instance, the number of persons who voted for a certain candidate, or the number who graduated from high school. Then $v_i = v_i^h + v_i^a$, where v_i^h is the number of hispanics with the response in question, and v_i^a is the corresponding number of non-hispanics. Although v_i is observable, its components v_i^h and v_i^a are generally unobservable. The main issue is to estimate

$$P^h = \sum_i v_i^h / \sum_i n_i^h. \tag{1}$$

Generally, the denominator of P^h is known but the numerator is not. In the Stockton exit poll, P^h is the percentage of hispanics who support Jackson; in the Los Angeles education data, P^h is the percentage of hispanics with high school degrees, for two examples. Estimating P^h from $\{v_i, n_i^h, n_i^a\}$ is an "ecological inference". In Table 1, $\{v_i^h, v_i^a\}$ are known, so the quality of the ecological estimates can be checked; likewise for the test data used by King.

Let $x_i = n_i^h/n_i$, the fraction of the population in area i that is hispanic; and let $t_i = v_i/n_i$, which is the ratio of response to population in area i. The three methods for ecological inference will be described in terms of (t_i, x_i, n_i), which are observable. The neighborhood model assumes that ethnicity has no impact within an area, so P^h can be estimated as $\sum t_i x_i n_i / \sum x_i n_i$. The ecological regression model, in its simplest form, assumes that hispanics have a propensity p to respond, constant across areas; likewise, non-hispanics have propensity q. This leads to a regression equation

$$t_i = p x_i + q(1 - x_i) + \epsilon_i, \tag{2}$$

so that p and q can be estimated by least squares. Call these estimates $\hat{p}$ and $\hat{q}$, respectively. Then P^h is estimated as $\hat{p}$. The error terms ϵ_i in (2) are not convincingly explained by the model. It is usual to assume $E\{\epsilon_i\} = 0$ and the ϵ_i are independent as i varies. Some authors assume constant variance, others assume variance inversely proportional to n_i, and so forth.

King's model is more complex. In area i, the hispanics have propensity p_i to respond and the non-hispanics have propensity q_i, so that by definition

$$t_i = p_i x_i + q_i(1 - x_i). \tag{3}$$

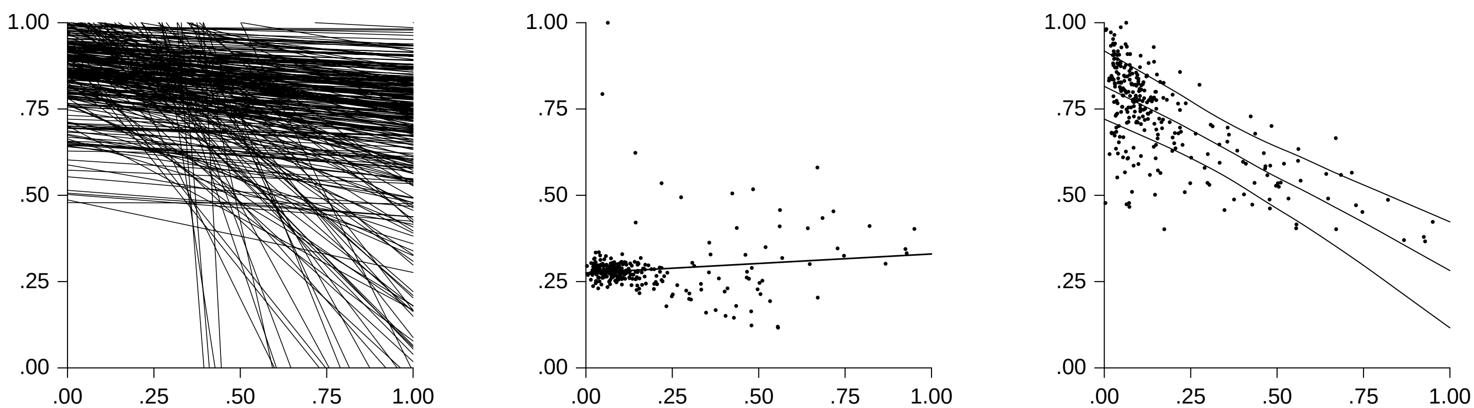
It is assumed that the pairs (p_i, q_i) are independent and identically distributed across i. The distribution is taken to be conditioned bivariate normal. More specifically, the model begins with a bivariate normal distribution covering the plane. This distribution is characterized by five parameters: two means, two standard deviations, and the correlation coefficient. The propensities (p_i, q_i) that govern behavior in area i are drawn from this distribution, but are conditioned to fall in the unit square. The five parameters are estimated by maximum likelihood. Then p_i can be estimated as

$E\{p_i|t_i\}$, the expectation being computed using estimated values for the parameters. Finally, P^h in (1) can be estimated as $\sum_i \hat{p}_i x_i n_i / \sum_i x_i n_i$. King seems to use average values generated by Monte Carlo rather than conditional means. There also seems to be a fiducial twist to his procedure, which resamples parameter values as it goes along; see chapter 8.

As a minor technical point, there may be a slip in King's value of the normalizing constant for the density of the truncated normal. One factor in this constant is the probability that a normal variate falls in an interval, given that it falls along a line. The conditional mean is incorrectly reported on pp. 109, 135, 307. In these formulas, $\omega_i \epsilon_i / \sigma_i$ should probably be $\omega_i \epsilon_i / \sigma_i^2$, as on pp. 108 and 304.

The tomography plot shows for each i the locus of points (p_i, q_i) in the unit square that satisfy equation (3). With King's method, $(\hat{p}_i, \hat{q}_i)$ falls on the line defined by (3), so that bounds are respected. (The neighborhood model also makes estimates falling on the tomography lines; ecological regression does not obey the constraints, and sometimes gives impossible estimates.) The $E\{t|x\}$ plot superimposes the data (x_i, t_i) on the graphs of three functions of x: the lower 10%-point, the mean, and the upper 10% of the distribution of $px + q(1 - x)$, with (p, q) drawn from the conditioned normal with estimated values of the parameters. The estimated p vs x plot shows $(x_i, \hat{p}_i)$ for each area i; likewise for estimated q vs x: these are called "bias plots". See Figure 1 for the Los Angeles education data. Data are shown only for every fifth tract; otherwise, the figure would be unreadable.

Figure 1. Diagnostic plots for the Los Angeles education data. Data for every fifth tract are shown. The tomography plot on the left has one line per tract, representing the possible combinations of the propensities (p_i, q_i). The hispanic propensity p_i is on the horizontal axis and q_i on the vertical. The plot seems uninformative. The middle panel plots $(x_i, \hat{p}_i)$. There is one dot per tract, with the fraction x_i of hispanics on the horizontal axis and the estimated hispanic propensity $\hat{p}_i$ on the vertical. The slope of the regression line is small but significant. The right hand panel plots (x_i, t_i). There is one dot per tract: x_i is on the horizontal axis and t_i, the fraction of persons in the tract with a high school education, is on the vertical. Also shown are 80% confidence bands derived from the model; the middle line is the estimated $E\{t|x\}$. The dots may be too high at the far right, hinting at nonlinearity.



To generate artificial data for Stockton, we fitted King's model to the exit poll data using

EZIDOS. As explained after equation (3), the key to the model is a bivariate normal distribution, with five parameters:

hispanic mean, non-hispanic mean, the two SDs, and the correlation.

EZIDOS estimated these parameters as 0.68, 0.37, 0.43, 0.21, and 0.45, respectively. There were 39 precincts. Following the model, we generated 39 random picks (p_i^*, q_i^*) from the estimated bivariate normal distribution, conditioning our picks to fall in the unit square. For precinct i, we computed t_i^* as $p_i^* x_i + q_i^*(1 - x_i)$, using the real x_i. Then we fed $\{t_i^*, x_i, n_i\}$ back into EZIDOS. In our notation, n_i is the total number of voters interviewed in precinct i, while x_i is the fraction of hispanics among those interviewed. Truth—the 56% in line 2 of Table 1—was computed as $\sum p_i^* x_i n_i / \sum x_i n_i$. The procedure for the registration data in Los Angeles was similar.

The extended model

The discussion so far covers the "basic model". In principle, the model can be modified so the distribution of (p_i, q_i) depends on covariates, although we found no real examples in the book. See chapter 9. The specification seems to be the following. Let u_i and w_i be covariates for area i. Then (p_i, q_i) is modeled as a random draw from the distribution of

$$\alpha_0 + \alpha_1 u_i + \delta_i, \quad \beta_0 + \beta_1 w_i + \epsilon_i. \tag{4}$$

Here $\alpha_0, \alpha_1, \beta_0, \beta_1$ are parameters, constant across areas. The disturbances (δ_i, ϵ_i) are independent across areas, with a common bivariate normal distribution, having mean 0 and a covariance matrix Σ that is constant across areas; but the distribution of (4) is conditioned for each i to lie in the unit square. Setting $\alpha_1 = \beta_1 = 0$ gives the basic model—only the notation is different.

King does not really explain when to extend the model, when to stop extending it, or how to tell if the extended model fits the data. He does advise putting a prior on α_1, β_1: cf. pp. 288–89. For the Los Angeles registration data, he recommends using variables like "education, income, and rates of home ownership . . . to solve the aggregation problem in these data [p. 171]". So, we ran the extended model with u_i and w_i equal to the percentage of persons in area i with household incomes above \$20,000 a year. The percentage of hispanics registered as democrats is 73.5%; see Table 1. The basic model gives an estimate of 90.8% $\pm$ 0.5%. The extended model gives 91.3% $\pm$ 0.5%. The change is tiny, and in the wrong direction. With education as the covariate, the extended model does very well: the estimate is 76.0% $\pm$ 1.5%. With housing as the covariate, the extended model goes back up to 91.0% $\pm$ 0.6%. In practice, of course, truth would be unknown and it would not be at all clear which model to use, if any. The diagnostics cannot help very much. In our example, all the models fail: the scatter diagram is noticeably higher than the confidence bands in the $E\{t|x\}$ plots. There is also a "non-parametric" model (pp. 191–6); no real examples are given, and we made no computations of our own.

Identifiability and other *à priori* arguments

King's basic model constrains the observables:

(5) the t_i are independent across areas.

Moreover, the expected value for t_i in area i is a linear function of x_i, namely,

$$E\{t_i|x_i\} = ax_i + b(1 - x_i), \tag{6}$$

where a is the mean of p and b is the mean of q, with (p, q) being drawn at random from the conditioned normal distribution. Finally, the variance of t_i for area i is a quadratic function of x_i:

$$\mathrm{var}(t_i|x_i) = c^2x_i^2 + d^2(1 - x_i)^2 + 2rcdx_i(1 - x_i), \tag{7}$$

where c^2 is the variance of p, d^2 is the variance of q, and r is the correlation between p and q.

One difference between King's method and the ecological regression equation (2) is the heteroscedasticity expressed in (7). Another difference—perhaps more critical—is that King's estimate for area i falls on the tomography line (3). When ecological regression makes impossible estimates, as in Stockton, this second feature has some impact. When ecological regression makes sensible-looking (if highly erroneous) estimates, as in Los Angeles, there is little difference between estimates made by ecological regression and estimates made by King's method: the heteroscedasticity does not seem to matter very much. See Table 1.

In principle, the constraints (5), (6), and (7) are testable. On the other hand, assumptions about unobservable area-specific propensities are—obviously—not testable. Failure of such assumptions may have radical implications for the reliability of the estimates. For instance, suppose that hispanics and non-hispanics alike have propensity π_i to respond in area i: the π_i are assumed to be independent across areas, with a mean that depends linearly on x_i as in (6) and a variance that is a quadratic function of x_i as in (7). Indeed, we can choose (p_i, q_i) from King's distribution and set $\pi_i = p_i x_i + q_i(1 - x_i)$. This "equal-propensity" model cannot on the basis of aggregate data be distinguished from King's model but leads to very different imputations. Of course, the construction applies not only to the basic model but also to the extended model, a point King seems to overlook on pp. 175–83. No doubt, the specification of the equal-propensity model may seem a bit artificial. On the other hand, King's specifications cannot be viewed as entirely natural. Among other questions, why are the propensities independent across areas? why the bivariate normal?

According to King (p. 43), the neighborhood model "can be ruled out on theoretical grounds alone, even without data, since the assumptions are not invariant to the districting plan". This argument applies with equal force to his own model. If, for example, the model holds for a set of geographical areas, it will not hold when two adjacent areas are combined—even if the two areas have exactly the same size and demographic makeup. Equation (7) must be violated, because averaging reduces variance.

Summary and conclusions

King does not really verify conditions (5), (6), and (7) in any of his examples, although he compares estimated propensities to actual values. Nor does he say at all clearly how the diagnostics would be used to decide *against* using his methods. The critical behavioral assumption in his model cannot be validated on the basis of aggregate data. Empirically, his method does no better than ecological regression or the neighborhood model, and the standard errors are far too small. The diagnostics cannot distinguish between cases where estimates are accurate, and cases where estimates are far off the mark. In short, King's method is not a solution to the ecological inference problem.

References

Achen, C. H. and Shively, W. P. (1995), *Cross-Level Inference*, University of Chicago Press.

Freedman, D. A., Klein, S. P., Sacks, J., Smyth, C. A., and Everett, C. G. (1991), "Ecological regression and voting rights," *Evaluation Review*, 15, 659–817 (with discussion).

Goodman, L. (1953), "Ecological regression and the behavior of individuals," *American Sociological Review*, 18, 663–4.

Goodman, L. (1959), "Some alternatives to ecological correlation," *American Journal of Sociology*, 64, 610–25.

King, G. (1997), *A Solution to the Ecological Inference Problem*, Princeton University Press.

Klein, S. P. and Freedman, D. A. (1993), "Ecological regression in voting rights cases," *Chance*, 6, 38–43.

Klein, S. P., Sacks, J., and Freedman, D. A. (1991), "Ecological regression *versus* the secret ballot," *Jurimetrics*, 31, 393–413.

Lieberson, S. and Waters, M. (1988), *From Many Strands: Ethnic and Racial Groups in Contemporary America*, New York: Russell Sage Foundation.

Massey, D. S. (1981), "Dimensions of the new immigration to the United States and the prospects for assimilation," *Annual Review of Sociology*, 7, 57–85.

Massey, D. S. and Denton, N. A. (1985), "Spatial assimilation as a socioeconomic outcome," *American Sociological Review*, 50, 94–105.

Robinson, W. S. (1950), "Ecological correlations and the behavior of individuals," *American Sociological Review*, 15, 351–7.

Skerry, P. (1995). *Mexican Americans: The Ambivalent Minority*, Harvard University Press.

Tam, W. K. (1998), "Iff the assumption fits," *Political Analysis* 7, in press.

10 June 1998
Technical Report No. 515

Statistics Department
UC Berkeley
CA 94720