

Micah Altman, `Micah_Altman@harvard.edu`

Jeff Gill, `jgill@polisci.ufl.edu`

Michael McDonald, `mmcdon@gmu.edu`

1 Introduction

The numerical accuracy of commonly used statistical software packages has been recently evaluated by a number of concerned authors (McCullough and Vinod, 1999; McCullough 1998; McCullough 1999a; McCullough 1999b; Altman and McDonald 2001). The central concern is that different embedded numerical methods have been shown to produce vastly different solutions from the same data and model. Furthermore, there is great variation in the quality and thoroughness of programmers of statistical software and their sensitivity to purely statistical concerns (Knuth 1997). Specific cases of incorrect analyses based on these problems have recently been documented in published research (Altman and McDonald 2003), and others are sure to exist.

Methodologists in the social sciences are increasingly sophisticated in their use of statistical software. Elaborate models are now commonly programmed into more advanced statistical packages. Such sophistication and power are not free, numerical issues remain important for ensuring high quality results, yet are often ignored. Just as worrisome, many models are becoming increasingly fragile purely due to the complexity of the specification (Achen 2001).

Among the most complex models specifications in the social sciences are those proposed as a solution to the *ecological inference problem*, inferences of individual behavior drawn from analysis of aggregate data. Until recently, the gold standard of ecological inference solutions has been simple linear a regression approach proposed by Goodman (1953). The numerical methods underpinning the regression algorithm are relatively basic, although poor implementations of the algorithm still exist in some commercial software, it is generally understood how to implement it in a numerically accurate way.¹

In contrast, recent solutions to the ecological inference problem are much more complex. King (1997), and McCue (2001) in response to King, have proposed solutions to the ecological inference problem that purport to produce more accurate (and realistic) estimates of the true individual values, but with a considerable increase in algorithmic complexity.

Although much of the attention has been positive, King’s solution has also been criticized (Tam Cho 1999; Ferree 1999; Freedman, *et al.* 1999). Recently, McCue (2001) argues that simulation

¹The Goodman procedure has other (non-computational) flaws, most notably that it may produce estimates outside the logical bounds. For example, estimating turnout rates over 100% for population sub-groups.

and the use of a constrained maximum likelihood algorithm are unnecessary. McCue’s less statistically complex model is still computationally intensive, respecting many of the same underlying assumptions as King. McCue’s method has garnered critiques of its own, although none has as yet been published. Even though McCue argues for his method almost exclusively on computational grounds, he offers no direct evidence that his computational alternative for estimating the EI model is more accurate or reliable than the method he is attempting to displace. This is unfortunate but not unusual—in social science, computational issues tend to be mentioned in passing. For example, although the software that King distributes to compute his EI model contains many numerically sophisticated features, King devotes only two out of the nearly three hundred and fifty pages in his book to computational details (see King 1997, Appendix F).² Critiques notwithstanding, since a speedy resolution to the debate over ecological inferences does not seem forthcoming, we believe it is imperative to subject the contending methods to numerical testing.

As scholars in social science work more frequently with sophisticated statistical models, attention to computational details has begun to slightly increase. In this volume, three other authors address computational issues – focussing particularly on computational efficiency: Wakefield (chapter XXX) devotes a section of his paper to various strategies for efficiently computing multi-stage Bayesian approaches to ecological inference; Grofman and Merrill devote their attention to (chapter XXX) “quick and dirty” approaches to computing ecological inference more quickly (although admittedly less statistically sophisticated), and Mattos, as part of investigating the predictive correctness of alternative models, (chapter XXXX) develops a fast alternative to MCMC methods for computing the beta-binomial hierarchical EI model.

Although we focus here on computation, we do not propose faster ways of computing EI models. Instead, for the first time in publication, to our knowledge, we provide an explicit comparison of the numerical properties of the leading proposed solutions to the EI problem. In this work we examine in detail the numerical accuracy of Goodman’s (1953), King’s (1997), and McCue’s (2001) approaches to ecological inference. We analyze the performance of these separate approaches to “solving” the ecological inference problem through data perturbation, and comparative reliability assessment. The data perturbation technique is used to evaluate the pseudo-stability of these competing techniques across identical datasets. Provided results illuminate the trade-offs among correctness, complexity and numerical sensitivity.

The paper proceeds as follows: we describe potential sources of numeric inaccuracies within King’s solution, provide ways to test for the presence of such numeric inaccuracies (not only for this context but in a broader context as well), perform tests, and describe ways to ameliorate the inaccuracies we identify. The primary purpose is to give an accurate picture of the extent to which numerical issues can affect substantive conclusions in ecological inference models.

²See however, Chapter 6 in Altman, Gill, and McDonald, (2003), in which these computational details are discussed at more length.

2 Sources of Numeric Inaccuracy in Ecological Inference

Computational problems may affect any statistical program. For ecological inference solutions there are three potential areas of concern where computation may affect estimates to the ecological inference problem: floating point error, the choice of optimization algorithms, and imperfect random number generation.

2.1 Floating Point Arithmetic

Most statistical programs, and all of the ecological inference techniques examined here, use floating point arithmetic. Numeric inaccuracies are introduced because statistical programs use a fixed number of bits to store and manipulate binary numbers and calculations. When the binary representation of a number exceeds the available number of bits, *overflow* occurs. Surprisingly, decimal numbers with fractions often do not have an exact binary representation, thus forcing software to round or truncate numbers to the level of precision. It is here that small errors occur in the translation of numbers from the world of pencil and paper to computers. Manipulating these numbers, such as adding the squares of a large and small number, may propagate these errors or introduce new ones that can produce wildly different answers from the truth. Such accumulated numerical errors in repeated, complex calculations can be deadly to statistical computing procedures (Thisted 1988).

Higham (1996) observes that researchers are often overconfident of the accuracy of their calculations because they are unaware of the consequence of floating point inaccuracy. Even a simple computation may be subject to numerical inaccuracy, and one small error propagated through a calculation can cause significantly inaccurate results. Furthermore, increasing the precision of an intermediate calculation does not monotonically increase the accuracy of the final result.

2.2 Nonlinear Optimization

Estimating a solution for a regression, such as Goodman's approach, involves a straightforward series of closed-form calculations that may be found in any intermediate level statistics book. Non-linear optimization problems, such as the maximum likelihood algorithms used in King's and McCue's methods, suffer from a further and larger complication: finding the global optimum to a likelihood function. King's programs use the `Gauss` implementation of constrained maximum likelihood solver, `cml`, to fit the truncated bivariate normal onto the unit square and estimate the parameters of its solution, while McCue proposes an unconstrained estimation process (with the Duncan and Davis bounds built within) that alternately uses an unconstrained maximum likelihood solver.

Standard techniques for optimization, such as those based on Newton-Raphson, typically involve examining the numerically calculated or analytic gradients of the likelihood function at the current

guess for the solution, and then use these to determine a direction to head “uphill.” Although we can determine conditions for the global optimum for some classes of problems, finding this point is never *guaranteed*. As one eminent set of practitioners in the field wrote: “Finding a global extremum is, in general, a very difficult problem.” (Press, *et al.*, 2002: 398).³ Most techniques for finding global optima involve some degree of guesswork, or heuristics: either the algorithm guesses at initial values for parameters and proceeds to find a local optima from there, or it perturbs a local optima in an attempt to dislodge the search from it.

However, as already noted, these programs are also prone to floating point errors discussed above. Data that is poorly conditioned with respect to the EI optimization problem or implementations with numerical inaccuracies may inadvertently create false optima. Even in the absence of numerical problems, standard techniques may become stuck on purely local optima.

2.3 Pseudo-Random Number Generation

Goodman’s regression technique uses a deterministic algorithm (least squares) for finding a solution, unlike the maximum likelihood algorithms used by other solutions (although for the linear model the MLE and LS gives the exact same answer). These other programs may make use of randomness to search for the global optimum in that starting points or candidate points are often randomly generated by the algorithm. Additionally, King’s solution uses random numbers in simulation to find solutions to some otherwise intractable aspects of his estimation process.

Random numbers provided by computer algorithms are never genuinely random. Instead, they are produced by *pseudo-random number generators* (PRNGs), deterministic processes that create a sequence that is statistically similar, in limited respects, to random draws from a uniform distribution. Pseudo-random number generators start with a single “seed” value and generate a repeating sequence with a certain fixed length, or period (p). In order for simulation or sampling results to be accurate, a PRNG should satisfy three criteria: long period, independence, and consistent in distribution. In addition, all require a truly random seed to produce independent sequences (Ripley 1987, 1988; Knuth 1997; Gentle 1998).

Random number generation is an important, but under-studied aspect of *applied* statistical computing, at least on the high-end of statistical package evaluation. Some authors have shown the deleterious effects of poorly designed random number generation procedures (Gentle 1988, Morgan 1984). Fortunately, the King solution’s use of random numbers is comparatively light, and other research suggests that the choice of random number generators affects it very little (see Altman, Gill and McDonald 2003). However, we know from a vast foundational literature that serious problems can be caused by poorly written PRNG algorithms: Atkinson (1980), Butcher (1961),

³King speculates, but does not prove, that the likelihood mode for the TBVN is globally unique (King 1997, Appendix D) We further discuss techniques for the identification of global optima in Altman, Gill and McDonald (2003).

Coveyou and MacPherson (1967), Downham (1970), Dudewicz (1976), Good (1957), Gorenstein (1967), Krawczyk (1992), Kronmal (1964), Learmonth and Lewis (1973), Marsaglia (1968), McArdle (1976), McCullough (1999), McCullough and Wilson (1999), Toothill, Robinson, and Adams (1971), and Whittlesey (1969). Much in the same way that the famously flawed, but widely used RANDU algorithm from IBM was used for quite some time although it had received quite a lot of criticism in this literature (Coveyou 1960, 1970; Fishman and Moore 1982; Hellekalek 1998).

3 Assessing the Numerical Accuracy of Statistical Inferences

3.1 Evaluating and Ensuring Accuracy

Broadly defined, a statistical estimate is a mapping between:

$$\{data, model, priors, inference\ method\} \Rightarrow \{estimates\}$$

or, symbolically

$$\{\mathbf{X}, M, \pi, Im\} \Rightarrow e$$

If, however, the estimate is too complex to calculate analytically, using only ‘pencil and paper’, we have to consider how computation may affect the results. In such a case, if the output from the computer is not necessarily equivalent to e , it can be inaccurate. Moreover, the output may be dependent upon the algorithm chosen to perform the estimation, parameters given to that algorithm, the accuracy and correctness of the implementation of that algorithm, and implementation-specific parameters. Including these factors results in a more complex mapping.

$$\{\mathbf{X}, M, \pi, Im, algorithm, algorithm\ parameters, implemtation, implementation\ parameters\} \Rightarrow output$$

By “algorithm” we intend to encompass choices made in creating output that are not part of the statistical description of the model and which are independent of a particular computer program or language: This includes the choice of mathematical approximations for elements of the model (e.g., the use of Taylor series expansion to approximate a distribution) and the method used to find estimates (e.g., non-linear optimization algorithm). Implementation is meant to capture all remaining ‘programming’ details, including bugs, and the implementation of data storage and arithmetic operations (e.g., using floating point double precision).

The *accuracy* of the output actually presented to the user is thus the distance (using a well-behaved distance metric) between estimates and output⁴ : $accuracy = Distance = exp(e, output)$

⁴Since “accurate” is often used loosely in other contexts, it is important to distinguish between computational accuracy, as above, and correct inference. A perfectly accurate computer program can still lead one to incorrect results if the model being estimated is misspecified.

The choice of an appropriate distance metrics depends on the form of the estimates and the purpose for which those estimates are used. E.g. for output which is a single scalar value, we might choose “Log Relative Error” as an informative distance metric, which can be interpreted, roughly as the number of numerically ‘correct’ digits in the output:

$$\text{LRE} = -\log_{10} \left(j \frac{\text{output} - e}{e} j \right) \quad (1)$$

When $c = 0$, LRE is defined as the *log absolute error*, LAE, given by:

$$\text{LRE} = -\log_{10}(j\text{output} - e \times j) \quad (2)$$

Accuracy alone is often not enough to ensure correct inferences, because of the possibility of model mis-specification, the ubiquity of un-modelled measurement error in the data, and of rounding error in implementations. Where noise is present in the data or its storage representation, and not explicitly modelled, correct inference requires the output to be *stable*.

Stability is simply the distance of the estimate as compared to output that is based upon data with some small noise: $\text{stability} = \text{Distance} = \exp(e, \text{output}')$, where $\text{output}'|Y' \equiv Y + \Delta Y$. Note that unstable output could be due to sensitivity in the algorithm, implementation or model – regardless, if there is any error in the data from any source, inferences will be incorrect if the output is not stable. (Less formally, a stable algorithm gives, to quote Higham 1996, “almost the right answer to almost the same problem.”)

In an ideal world, we would be able to compute formal bounds on the accuracy and stability of the estimates. For some distribution functions, and many individual computations in matrix algebra, it is possible to derive analytical bounds on the accuracy of those functions, given a particular implementation and algorithm. Alternatively, methods such as interval arithmetic can be used to track the accumulated round-off errors across a set of calculations. [Higham 2002 provides an excellent introduction to the field.] Typically, however, the statistical analyses used by social scientists are too complex for bounds to be either derivable or informative. (Although see Polasek 1987 for some results on the accuracy of linear regression in the presence of rounding error, and Higham 2002, section 25.2 for some bounds on Newton’s method.)

Furthermore, in an ideal world, all statistical computation would follow the best practices for treatment of data, choice of algorithm, and programming techniques. (See Altman, Gill & McDonald 2003; Thisted 1988; for a survey of best practices.) Unfortunately, while best practices can improve accuracy and stability, they cannot guarantee it generally for estimates of nonlinear model, and best practices are often ignored, even in large commercial statistical packages.

So, what can one do to assess or ensure the accuracy and stability of one’s estimation procedure when formal bounds are not known? There are three general heuristics that can help to draw attention to potential computational problems:

1. **Test benchmark cases.** It is sometimes possible to devote extraordinary effort to compute correct estimates exactly (or to a known level of accuracy), for a particular set of test data and a particular model. One can then compare the estimates of a particular algorithm and implementation to the known results. This approach is taken in the NIST tests for accuracy of statistical software [Rogerson, et. al 2000].

This approach is useful, and we recommend it as a minimal requirement for any publicly distributed software, wherever feasible. However, it has three large drawbacks. First, it may not be feasible to create benchmark data, for which estimates can be calculated with known accuracy, that are at all realistic. Second, this method can detect some inaccuracies, but cannot prove that the program and algorithm yield accurate results outside of the data tested. One simply has to assume that inaccuracy is unlikely where the data used in the tests are sufficiently similar of the data being analyzed.

2. **Use separate information to confirm results, or necessary/sufficient conditions for results.** Generally, in any statistical analysis, the researcher should always apply substantive knowledge of the model, data, and phenomena being analyzed to check that the results are plausible. Implausible results should be held up to extensive scrutiny.

Besides this higher-level “gut-check” there may be other techniques that can be used to confirm (or disconfirm) results:

- Test that known analytic necessary conditions hold true at the solution. For example, after using a non-linear optimization algorithm to find a candidate solution, one can apply probabilistic tests that can disconfirm global optimality.
 - Examine likelihood profiles and other diagnostics [see McCullough in Altman, Gill and McDonald 2003].
 - Use sensitivity analysis. Sensitivity analyses are invaluable because they can often be applied where the benchmark tests and independent confirmation are unavailable. Furthermore, unlike benchmarks, sensitivity analysis can be performed using the actual data being analyzed, rather than being limited to restricted cases, as are the previous two methods. Sensitivity analysis cannot demonstrate that the results are correct or incorrect, nor can they be used to improve estimates of ‘correct’ values, but they can serve to draw attention to potential problems in algorithm, implementation, or model.
3. **Try alternative methods on the same problem.** One popular approach is to replicate the analysis keeping the data and model the same, but using multiple different algorithms, algorithmic parameters (such as starting values), and implementations (e.g. different PRNG’s and/or different optimization software). If results disagree, one should investigate (applying

the other techniques) until one clearly understands which set of results should be discarded. This is highly recommended where multiple implementations and algorithms are available. The effort required to create alternatives where none presently exist, however, can be prohibitively high.

A second popular, and complementary, approach is to replicate the analysis while perturbing the input data, and to observe the sensitivity of the estimates to such perturbations. (Sensitivity, or ‘pseudo-instability’ is not a measure of true computational stability, since values for the correct estimates are unknown.) This has the advantage of drawing attention to results that cannot be supported confidently given the current data, model, and algorithm/implementation, and unlike the first method, is easy to implement.

These two sensitivity tests can be combined fruitfully. A potential drawback of the second method is that pseudo-instability detected by perturbing the data could be due to problems in the algorithm/implementation, but may also be due to the interaction of model and data. For example, the results of a linear regression, running on data that is ‘almost’ multicollinear, can be highly unstable with respect to very small amounts of noise in the data, even if the regression calculations are performed exactly (without numerical inaccuracy). The instability will not be reflected in the standard errors of the estimate. These can be arbitrarily small, even in the presence of multicollinearity (Beaton, Barone and Rubin, 1977). *Combining the two methods can help to separate the portions of pseudo-instability due to model.* By running multiple implementations/algorithms on the same sets of perturbed data, if one implementation is more stable than the other, the difference in pseudo-stability is a result of implementation and algorithm, not model and data, which are kept fixed by construction.

Note that the size and form of the noise is not what serves to differentiate numerical problems from model/data problems – even simple uniform noise at the level of machine round-off can affect analyses purely because of model and data problems. It is the combination of perturbations and varying implementations that allows one to gain some insight into the sources of sensitivity. Nevertheless, regardless of the cause of sensitivity, one should be cautious if the conclusions are not pseudo-stable with respect to the amount of noise that is reasonably thought to be in the data.

The three approaches above cannot be used to prove the accuracy of a particular method, but are useful in drawing attention to potential problems. Further experimentation and analysis may be necessary to determine the specific cause of the problem. For example, if two software packages disagree on the estimates for the same model and data, the discrepancy could be a result of several factors⁵:

⁵With the exception of bugs, it is not necessarily clear-cut whether the programmer or end-user is at fault for any of these problems. Users of statistical software should pay close attention to warning messages, diagnostics, and

- **Implementation issues.** Either one or both programs have a bug, performs (some) calculations less accurately, or the results are conditioned on different implementation-level parameters (e.g. a difference in a convergence tolerance setting).
- **Algorithmic issues.** One or both programs may use an algorithm which has its preconditions violated by the particular model and data. Algorithms may afford different levels of approximation error. Or the results are conditioned on different values for algorithm-specific parameters (e.g. starting values for local optimization algorithms)
- **Data and model issues.** The problem is ill-conditioned.

3.2 Conditioning

The last of these issues is worth discussing in more detail, since there are common misconceptions regarding it. Conditioning is a often mentioned, but it's precise meaning is sometimes not well-understood by social scientists.

Following Higham 2002, (sections 1.5-6): The most general definition of conditioning is “the sensitivity of the model to perturbations of the data”. Condition numbers are used to represent the conditioning of a problem with respect to a particular set of inputs. For example, if a scalar function f is twice differentiable, a useful way to may define the relative condition number of f is $c(x) = \left| \frac{x f'(x)}{f(x)} \right|$. When defined in this way, the accuracy of the estimate is $c(x) \times \text{backwarderror}$, where ‘backward error’ is defined as the minimum $|\Delta x|$ for which our computation of y , \tilde{y} satisfies $\tilde{y} = f(x + \Delta x)$.⁶

All of this is a formalization of the notion that the accuracy of the estimate is a function of the model, data, and the computational procedure. The condition number is a particularly useful formalization because it is usually easier to derive the backward error of a computational method than the overall accuracy or stability.

Although conditioning is an important factor in the accuracy of any computation, social scientists should not assume that all computational inaccuracies problems are *simply* a matter of conditioning. In fact, a computation method with a large backward error will yield inaccurate results even where the problem itself is well-conditioned.

Moreover, the conditioning of the problem depends on these data, the model, algorithm, and the form of perturbation. There is no such thing as data that is well-conditioned with respect to every model. While it might appear tempting to use condition number estimators produced by standard

stated limitation of implementations and algorithms. Often however, software developers fail to provide adequate diagnostics, informative warning messages, or document the computational methods used and their limitations . Users should also examine data for outliers, coding errors, and other problems. However, users may have no way of knowing a-priori that a particular set of data is likely to cause computational problems, given the algorithm and implementation chosen by the programmer.

⁶These definitions and the example of $\log(x)$ below are both from Higham 2002.

statistical software (such as matlab) to calculate condition numbers for a particular dataset, the results are bound to be misleading, since the formula used by these estimators are tailored to specific types of problems in linear algebra, such as matrix inversion. These formula may be completely inappropriate to estimate the conditioning of another type of problem or computation procedure. To use a very simple example, using the condition number formula above, $x = 1$ is would be ill-conditioned for the function $\log(x)$ but is not well conditioned for the function e^x . Another example is that Higham (2002) shows that the accuracy of Newton’s method depends on the condition of the Jacobian at the solution (as well as the accuracy of the Jacobian and residual calculations), not (as one might naively assume) on the conditioning of the data matrix.⁷

3.3 Data Perturbations for Probing Numerical Stability

The accuracy and stability of an algorithm are measured both through analysis of the algorithm itself (e.g., examining every operation in the algorithm to compute bounds on the numeric error) and through an analysis of the algorithm’s overall behavior. Analysis of the algorithm is tractable when the subsequent results are well understood and the method is simple (e.g., computable using analytic techniques, or using supercomputers to compute the answers with hundreds of digits of precision).⁸

3.3.1 Perturbation Strategy

The definition of stability above suggests an exploratory test for a given problem: introduce small random perturbations to the data, on the order of the measurement error of the instruments used to collect it, and recalculate the estimate. This technique is analogous to bootstrapping, with two differences. First, in bootstrapping the sample selection is randomly perturbed but individual observations are not, whereas in our strategy the sample selection is not perturbed but the individual observations are. Second, our strategy doesn’t require any information to be discarded—as long as the perturbations are within measurement error. This technique is described in this context by Beaton, Rubin, and Barone (1976), who develop a stability index based on it.

Gill, *et al.* (1981, particularly sectin 8.3.3) recommend a similar method, although it is informally described, and suggested as a pragmatic method for gauging whether a program was stable. Also, whereas Beaton, Rubin and Barone perturb only the explanatory variables in the model, Gill, *et al.* do not distinguish among the inputs to the computation.

⁷This is not to claim that the data matrix is unimportant. The data matrix will affect the conditioning of the residual calculation problem, as will the method of calculating the residual, and the function being evaluated. A standard condition number, however, does not necessarily shed any light on this type of conditioning.

⁸The ecological inference problem is not tractable for algorithmic analysis even when the true precinct level parameters are known. Since correct application of an ecological inference model does not imply that the estimates equal the true parameters, we cannot know the true estimates that an infinitely accurate model would produce.

To see how these perturbations affect the estimation process, consider two likelihood functions: a standard form based on the observed data $\ell(\theta, \mathbf{x})$, and an identical specification but with perturbed data $\ell_{\mathbf{p}}(\theta, \mathbf{x}_{\mathbf{p}})$. Here \mathbf{p} denotes an individual perturbation scheme: $\mathbf{p} = [p_1, p_2, \dots, p_n] \in \mathfrak{R}^n$ applied to the data: $\mathbf{x} = [x_1, x_2, \dots, x_n] \in \mathfrak{R}^n$. Thus we can show that comparing the two likelihood functions is analogous to comparing an un-weighted likelihood function $\ell(\theta, \mathbf{x}) = \sum_i \ell_i(\theta, \mathbf{x}_i)$ to a weighted version $\ell_{\mathbf{p}}(\theta, \mathbf{x}_{\mathbf{p}}) = \sum_i p_i \ell_i(\theta, \mathbf{x}_i)$. Or we could define the unperturbed likelihood function to be one in which there are null perturbations or weights: $\ell_{\mathbf{p}_0}(\theta, \mathbf{x}_{\mathbf{p}_0}) = \sum_i p_{0i} \ell_i(\theta, \mathbf{x}_i)$, where \mathbf{p}_0 is simply a vector of 1's. This setup gives us two maximum likelihood vectors to compare: $\hat{\theta}$ and $\hat{\theta}_{\mathbf{p}}$.

In this context, our approach is to evaluate the range of $\hat{\theta}$ produced by multiple samples of $\mathbf{x}_{\mathbf{p}}$ generated by randomly production of \mathbf{p} disturbances across different datasets, \mathbf{x} . The idea builds upon the mechanical approach of Cook (1986) who looks for maximizing and minimizing perturbances, and roughly follows a simpler test of logistic regression given by Pregibon (1981). (Lawrance [1988] applies Cook's method to develop diagnostics for linear regression.)

In addition, although this evaluation methodology does not require that the likelihood function be statistically well-behaved, it does have a natural interpretation for well-behaved maximum likelihood estimations. If the likelihood function for an MLE is well behaved (as King surmises for his model (see 1997, p.310-1), then there is a simple mapping between perturbations of data and perturbations of the model. For example, small normally-distributed noise added to the data should induce a corresponding small mean-shift in the likelihood curve (St. Laurent and Cook 1993).

Cook (1986) continues on to define the *likelihood displacement*:

$$LD_{\mathbf{p}} = -2[\ell_{\mathbf{p}}(\theta, \mathbf{x}_{\mathbf{p}}) - \ell(\theta, \mathbf{x})] \quad (3)$$

which measures the statistical influence that different perturbation schemes have on the estimation process. Not surprisingly, it can be shown that $LD_{\mathbf{p}}$ defines a large sample confidence region distributed χ_k^2 , where k is the number of specified parameters (Cox and Hinkley 1974, Cook and Weisberg 1982).

If the likelihood surface is steeply curved (in the multidimensional sense) at the MLE, then clearly we will see large values of $LD_{\mathbf{p}}$ for even small perturbation schemes. Contrary to intuition, a sharply spiked likelihood function in this way across *every* dimension is not a sign of a reliable result, it indicates a fragile finding that is heavily dependent on the exact form of the observed data. This is because the change in the likelihood function is not due to different values for the estimate (where sharp curvature is desired), it is due to changes in the data (perturbations) where sharp changes indicate serious sensitivity of the likelihood function to slightly different data: a model that is “non-resistant”.

Cook also ties this curvature definition back to the idea of statistical reliability by a geometric interpretation. Define $\mathbf{p}_a = \mathbf{p}_0 + a\mathbf{v}$ where $a \in \mathfrak{R}$, and \mathbf{v} is a unit-length vector. The interpretation

of \mathbf{p}_a is as a line that passes through \mathbf{p}_0 in the direction of the vector \mathbf{v} , where a gives the n -dimensional placement relative to \mathbf{p}_0 . The geometric curvature of $LD_{\mathbf{p}_a}$ in the direction of the vector \mathbf{v} , starting at \mathbf{p}_0 is given by:

$$C_{\mathbf{p}} = 2|(\Delta\mathbf{v})'H^{-1}(\Delta\mathbf{v})| \quad (4)$$

where Δ is the $k \times n$ matrix given by $\Delta_{ij} = \frac{\partial^2 \ell_{\mathbf{p}_0}(\hat{\theta}, \mathbf{x}_{\mathbf{p}})}{\partial \theta_i \partial p_{0j}}$ (that is, evaluated at the MLE and \mathbf{p}_0), and H is the standard Hessian matrix. Cook (p. 139) suggests calculating the maximum possible curvature by obtaining the \mathbf{v} vector that maximizes (4): $C_{\max} = \max_{\mathbf{v}} C_{\mathbf{v}}$, which gives the greatest possible change in the likelihood displacement. Other strategies include perturbing in one direction at a time as a means of understanding which dimensions (i.e. parameters) are more sensitive than others.

Recent work by Parker, Pierce and Eggert (2000) formalizes a variant of this idea, which they call ‘‘Monte Carlo Arithmetic.’’ Essentially, they replicate an analysis while introducing uniformly distributed perturbations (in the form of random rounding) to all values in all calculations. This approach is more widely applicable than formal analysis (which can be practically impossible to apply to complex problems). And even where formal analysis is possible, Parker *et al.* show that MCA can also yield tighter practical bounds. They argue this is a very successful ‘‘idiot light’’ for numerical inaccuracy.

3.3.2 Perturbations and Measurement Error

Perturbations can be considered in exactly the same way as measurement error. The effects of measurement error on statistical models is quite well known. Essentially there are two problems: zero mean measurement error and non-zero mean measurement error. The non-zero case obviously and immediately leads to biased coefficients in the opposite direction of the bias. That is, in a linear model, multiplying some non-trivial $\delta > 1$ to an every case of explanatory variable, \mathbf{X} , implies that larger increases in this variable are required to provide the same effect on the outcome variable thus reducing the magnitude of the coefficient estimate. Another way of thinking about this is that a one-unit change in \mathbf{X} now has a smaller expected change in \mathbf{Y} . This effect is also true for GLMs where there is the additional complexity of factoring in the implications of the link function.

Zero mean measurement error is the more important and more common situation. Furthermore, the effects found for zero mean measurement error also apply to non-zero mean measurement error in addition to the effects just discussed above. Here we will make our points primarily with the linear model, even though EI solutions are certainly not in this group, purely to aid in the exposition.

Suppose that the true underlying linear model meeting the standard Gauss-Markov assumptions is given by:

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}, \quad (5)$$

but \mathbf{X} and \mathbf{Y} are not directly observable where we instead get \mathbf{X}' and \mathbf{Y}' according to:

$$\mathbf{Y}' = \mathbf{Y} + \xi \quad \text{and} \quad \mathbf{X}' = \mathbf{X} + \nu \quad (6)$$

where $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ and $\nu \sim \mathcal{N}(0, \sigma_\nu^2)$. It is typically assumed that ξ and ν are independent of both each other plus \mathbf{X} and \mathbf{Y} . These are standard assumptions that can also be generalized if necessary.

Lets look first at the ramifications of substituting in the error model for \mathbf{Y} :

$$\begin{aligned} (\mathbf{Y}' - \xi) &= \beta\mathbf{X} + \epsilon \\ \mathbf{Y}' &= \beta\mathbf{X} + (\epsilon + \xi). \end{aligned} \quad (7)$$

Since ξ has zero mean and is normally distributed by assumption, then the effect of measurement error is to attenuate the overall model errors, but not to violate any assumptions. That is, measurement error in \mathbf{Y} simply falls to the regression residuals. Therefore, there is now simply a composite, zero mean, error term: $\epsilon' = (\epsilon + \xi)$. Unfortunately, the story is not quite so pleasant for measurement error in \mathbf{X} . Doing the same substitution now gives:

$$\begin{aligned} \mathbf{Y} &= \beta(\mathbf{X}' - \nu) + \epsilon \\ &= \beta\mathbf{X}' + (\epsilon - \beta\nu). \end{aligned} \quad (8)$$

While this seems benign since ν as zero mean, it does in fact lead to correlation between regressor and disturbance violating one of the Gauss-Markov assumptions. This is shown by:

$$\begin{aligned} \text{Cov}[\mathbf{X}', (\epsilon - \beta\nu)] &= \text{Cov}[(\mathbf{X} + \nu), (\epsilon - \beta\nu)] \\ &= E[(\mathbf{X} + \nu)(\epsilon - \beta\nu)] \\ &= E[\mathbf{X}\epsilon - \mathbf{X}\beta\nu + \nu\epsilon - \beta\nu^2] \\ &= -\beta\sigma_\nu^2. \end{aligned} \quad (9)$$

To directly see that this covariance leads to biased regression coefficients, let us now insert the measurement error model into the least squares calculation:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\text{Cov}[(\mathbf{X} + \nu), \mathbf{Y}]}{\text{Var}[\mathbf{X} + \nu]} = \frac{E[\mathbf{X}\mathbf{Y} + \nu\mathbf{Y}] - E[\mathbf{X} + \nu]E[\mathbf{Y}]}{E[(\mathbf{X} + \nu)^2] - (E[\mathbf{X} + \nu])^2} \\ &= \frac{E[\mathbf{X}\mathbf{Y}] - E[\mathbf{X}]E[\mathbf{Y}] + E[\nu\mathbf{Y}] - E[\nu]E[\mathbf{Y}]}{E[\mathbf{X}^2] + 2E[\mathbf{X}\nu] + \sigma_\nu^2 - (E[\mathbf{X}])^2 - 2E[\mathbf{X}]E[\nu] - (E[\nu])^2} \\ &= \frac{\text{Cov}(\mathbf{X}, \mathbf{Y}) + E[\nu\mathbf{Y}]}{\text{Var}(\mathbf{X}) + 2\text{Cov}(\mathbf{X}, \nu) + \sigma_\nu^2} \end{aligned} \quad (10)$$

where we drop out terms multiplying $E[\nu]$ since it is zero. Recall that ν is assumed to be independent of both \mathbf{X} and \mathbf{Y} . So the $2\text{Cov}(\mathbf{X}, \nu)$ term in the denominator and the $E[\nu\mathbf{Y}]$ term in the

numerator are both zero. Thus σ_ν^2 is the only effect of the measurement error on the slope estimate. Furthermore, since this is a variance and therefore non-negative, its place in the denominator means that coefficient bias will always be downward towards zero.

Of course the discussion so far has only considered the rather unrealistic case of a single explanatory variable. Regretfully, the effects of measurement error worsen with multiple potential explainers. Consider now a regression of \mathbf{Y} on \mathbf{X}_1 and \mathbf{X}_2 . The true regression model is expressed as in (5) is:

$$\mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \epsilon, \quad (11)$$

and assume that only \mathbf{X}_1 is not directly observable and we get instead $\mathbf{X}' = \mathbf{X} + \nu$ with the following important assumptions,

$$\begin{aligned} \nu &\sim \mathcal{N}(0, \sigma_\nu^2) & \text{Cov}(\mathbf{X}_1, \nu) &= 0 \\ \text{Cov}(\nu, \epsilon) &= 0 & \text{Cov}(\nu, \mathbf{X}_2) &= 0. \end{aligned} \quad (12)$$

Now as before, substitute in the measurement error component for the true component in (11):

$$\begin{aligned} \mathbf{Y} &= \beta_1 (\mathbf{X}_1 - \nu) + \beta_2 \mathbf{X}_2 + \epsilon \\ &= \beta_1 \mathbf{X}_1 - \beta_1 \nu + \beta_2 \mathbf{X}_2 + \epsilon \\ &= \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + (\epsilon - \beta_1 \nu). \end{aligned} \quad (13)$$

It should be intuitively obvious that problems will emerge here since β_1 affects the composite error term leading to correlation between regressor and disturbance as well as heteroscedasticity.

To show more rigorously that this simple measurement error violates the Gauss-Markov assumptions we look at the derivation of the least squares coefficient estimates. For the running example with two explanatory variables the normal equations are:

$$\begin{bmatrix} n & \sum X_{i1} & \sum X_{i2} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1} X_{i2} \\ \sum X_{i2} & \sum X_{i1} X_{i2} & \sum X_{i2}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \sum X_{i1} Y_i \\ \sum X_{i2} Y_i \\ \sum X_{i3} Y_i \end{bmatrix}. \quad (14)$$

So the inverse matrix of $\mathbf{X}'\mathbf{X}$ is:

$$\det(\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} \sum X_{i1}^2 \sum X_{i2}^2 - \sum X_{i1} X_{i2} \sum X_{i1} X_{i2} & \sum X_{i2} \sum X_{i1} X_{i2} - \sum X_{i1} \sum X_{i2}^2 & \sum X_{i1} \sum X_{i1} X_{i2} - \sum X_{i2} \sum X_{i1}^2 \\ \sum X_{i1} X_{i2} \sum X_{i2} - \sum X_{i1} \sum X_{i2}^2 & n \sum X_{i2}^2 - \sum X_{i2} \sum X_{i2} & \sum X_{i2} \sum X_{i1} - n \sum X_{i1} X_{i2} \\ \sum X_{i1} \sum X_{i1} X_{i2} - \sum X_{i1}^2 \sum X_{i2} & \sum X_{i1} \sum X_{i2} - n \sum X_{i1} X_{i2} & n \sum X_{i1}^2 - \sum X_{i1} \sum X_{i1} \end{bmatrix}$$

where:

$$\begin{aligned} \det(\mathbf{X}'\mathbf{X}) &= n \left(\sum X_{i1}^2 \sum X_{i2}^2 - \sum X_{i1} X_{i2} \sum X_{i1} X_{i2} \right) \\ &\quad + \sum X_{i1} \left(\sum X_{i1} X_{i2} \sum X_{i2} - \sum X_{i1} \sum X_{i2}^2 \right) \\ &\quad + \sum X_{i2} \left(\sum X_{i1} \sum X_{i1} X_{i2} - \sum X_{i1}^2 \sum X_{i2} \right) \end{aligned}$$

Using $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, and paying attention only to the last coefficient estimate, we replace X_{i1} with $X_{i1} + \nu$:

$$\begin{aligned} \hat{\beta}_2 = \det(\mathbf{X}'\mathbf{X})^{-1} & \left[\left(\sum (X_{i1} + \nu) \sum (X_{i1} + \nu) X_{i2} - \sum (X_{i1} + \nu)^2 \sum X_{i2} \right) \left(\sum (X_{i1} + \nu) Y_i \right) \right. \\ & + \left(\sum (X_{i1} + \nu) \sum X_{i2} - n \sum (X_{i1} + \nu) X_{i2} \right) \left(\sum X_{i2} Y_i \right) \\ & \left. + \left(n \sum (X_{i1} + \nu)^2 - \sum (X_{i1} + \nu) \sum (X_{i1} + \nu) \right) \left(\sum X_{i2}^2 \right) \right] \end{aligned} \quad (15)$$

where of course the same adjustment with ν would have to be made to every X_{i1} term in the determinant as well. The point here is to demonstrate that the measurement error in \mathbf{X}_1 can have a profound effect on the other explanatory variables, the extent of which is only now obvious when looking at the full scalar calculation of the coefficient estimate $\hat{\beta}_2$.

An easier, but perhaps less illustrative, way to show this dependency is with the handy formula:

$$\hat{\beta}_2 = \frac{\beta_{-1} - \beta_{-2}\beta_{21}}{1 - r_{21}^2} \quad (16)$$

where: β_{-1} is this same regression leaving out \mathbf{X}_1 , β_{-2} is the regression leaving out \mathbf{X}_2 , β_{21} is coefficient and r_{21}^2 is the r-square measure obtained by regressing \mathbf{X}_2 on \mathbf{X}_1 . What this shows is that the extent to which the variable with measurement error “pollutes” the coefficient estimate of interest is governed by the linear association between these explanatory variables, and the only time that measurement error in \mathbf{X}_1 will not have an effect is when there is no bivariate linear relationship between \mathbf{X}_2 and \mathbf{Y} or no bivariate linear relationship between \mathbf{X}_1 and \mathbf{X}_2 .

This section has analyzed measurement error effects in detail and demonstrated that outcome variable measurement error is benign and explanatory variable measurement error is dangerous. Perturbations are essentially researcher imposed measurement error. Having discussed the deleterious modelling problems with measurement error, we would never advise intentionally including it. Instead, the perturbations act like unintended, but completely known, measurement error as a means of testing the behavior of estimators and algorithms. That is, models that react dramatically to modest levels of measurement error are models that one should certainly be cautious about (particularly in the social sciences).

In summary, perturbation may introduce bias, but if the the problem is well-conditioned and the algorithm and implementation accurate, the bias should be small. Moreover, any bias introduced by perturbations should be the same when the same model and perturbed data are used in different implementations. So if two implementations of the same model show marked differences in pseudo-stability with respect to similar perturbation analyses, the root cause is asserted to be computational and not statistical.⁹

⁹This approach is complementary to the one proposed by Judge, Miller and Cho (chapter XXX) in this volume. Their approach uses instrumental variables, where available to reduce the effects of measurement error. Our approach provides a diagnostic of the results sensitivity to it.

3.3.3 Some Ramifications of Perturbed Models

Using the core idea of random perturbation, we can assess whether the results are reliable, whether they are consistent with respect to small perturbations in the data, and whether other implementation factors affect the estimates. This methodology complements standard diagnostic plots in two ways. First, one can use the strictly numeric results as an unambiguous check: simply evaluate whether the range of results across input perturbations still fits the original substantive conclusions about the results. Second, this methodology may sometimes reveal numerical problems that may be missed in standard diagnostic plots.

With regard to input perturbation, what is considered “small” for any particular case, is a matter of subjective judgment. There is an obvious lower limit though: perturbations of the data that are at the level below the precision of the machine should not be expected to cause meaningful changes in output. The upper limit on perturbations is less clear, but should be bounded by the accuracy of data measurement.

In political science, and many other social sciences, measurement error certainly dominates machine precision as a source of input inaccuracy. Much of the significant digits of macro data are reported as rounded, for example, to 1,000’s. Introducing perturbations to the rounding error of these data is a tractable problem to solve.¹⁰

Sometimes data are bounded, which introduces complications to perturbations. The simplest way of avoiding the bounding problem is to truncate any illegal value generated by perturbations to the constraint, but this introduces mass at the boundary points. To avoid this problem, we use resampling to draw sample perturbations from a set of truncated noise distribution, made symmetric to avoid biasing the data. A consequence of this is that observations closest to the [0,1] constraint are effectively subject to less noise. We report results using the second, more conservative, method. As a check, we replicated our results with the first method, our substantive conclusions did not change.¹¹

Choosing the number of perturbed data sets to generate is also something of an art. The extant literature does not specify a particular number of samples that is guaranteed to be sufficient for all cases. Parker (1997) and Parker, Pierce and Eggert (2000) use as many as one-hundred and as few as four samples in their monte-carlo arithmetic analysis. Parker (1997) also shows that (in all but

¹⁰The form of the perturbation is usually either uniform noise as in Beaton, *et al.*; Gill; and Parker *et al.*, (above) or normal as in St. Laurent and Cook (above). However, the proportional data used as the input to an EI analysis complicates the perturbation. Both types of perturbations can yield proportions outside of the legal [0,1] interval.

¹¹In future research, it would be interesting to model the form of the measurement error based on the substantive data-generating process. For example, in a two party race, it is possible that the primary source of error comes from miscounting individual ballots, and that each individual ballot has a small chance of being misclassified. Even if the probability of misclassification was the same for each ballot, the resulting measurement error would not necessarily be mean zero in terms of proportions: In a heavily partisan district, there would be more opportunities to misclassify votes from one party than from the other.

pathological cases) the distribution of the means of coefficients calculated under random rounding are normal, which suggests that thirty to fifty samples should be adequate. Moreover, since the perturbation technique can be replicated indefinitely, one can simply rerun the analysis, increasing the number of samples, until the variance across replications is acceptable for the substantive problem at hand.

Care must be used to distinguish between differences among implementations of a particular algorithm, and differences between algorithms used to compute the same quantity of interest. We expect that new versions of software will be made more accurate as implementations are improved, and better algorithms found. Software writers have a responsibility not only to make improvements, but also to document the range of acceptable conditions for running their software and the accuracy that may be expected from it. Furthermore, as improvements are made, facilities should be provided to replicate the results from previous versions.¹²

4 A Numerical Comparison of Ecological Inference Models

In this section, we compare the performance of three methods of ecological inference: Goodman's regression, King's solution, and McCue's new approach, with regard to both their numerical accuracy and how "correct" the estimates are – how close the answers they produce come to the truth (when the true behavior of the social system being modelled is known).

We use existing implementations of these proposed ecological inference solutions. King (1997) distributes two software versions of his model: a **Gauss** program, which King refers to as "EI," and a stand-alone DOS version, which King calls **EzI**. McCue has not released an official version of his model, but David James (directly provided, 2002) has made available to us a **Stata** version that he coded in consultation with McCue, which James has designated as "**AnEI**" (a.k.a, "**Analytical EI**").¹³ we also make use of implementations Goodman's regression routines supplied by both of these authors in their programs.¹⁴

Interest here is focused on the correctness of the three methods in estimating ecological inference parameters, and their sensitivity to perturbation. In addition, particular attention is paid to the

¹²King's EI software provides considerable built-in support for replication, diagnosis of statistical and computational problems, and support for different computational options. Such support is quite rare, and to be lauded. EI's reliance on **Gauss** as a statistical environment sometimes interferes with replication, however, as changes across **Gauss** versions cannot be controlled for in EI. Furthermore, although many computational options are provided in EI, the accuracy of these options is not always well documented.

¹³According to McCue, the James method also corrects an error in McCue's linear estimator, such that the original estimator can fall outside the interval [0,1].

¹⁴Since first releasing the program, King has provided a constant flow of updates. For the analysis reported here we use version 1.7 of EI running with **cml** version 2.0 and **Gauss** version 4.0.26. For the McCue algorithm we use **AnEI** 4.0 running in **Stata** 7.0. Elsewhere, we have noted that newer versions of the program, and even the different implementations of the EI and **EzI** program, on separate operating systems, may produce estimates outside the range of the random variance of simulation King uses in some of his calculations (Altman and McDonald 2002).

often overlooked importance of analyzing performance with different software option settings that the authors' provide, or are a component of the underlying programming language.

4.1 Goodman's Regression

Goodman (1953) proposed a simple linear regression solution to the ecological inference problem. Simplicity is often a virtue in terms of numerical accuracy. Numerous studies of the numeric accuracy that implementations of linear regression in most statistical packages are relatively accurate, compared to non-linear models. (McCullough and Vinod, 1999; McCullough 1998; McCullough 1999a; McCullough 1999b; Altman and McDonald 2001). Consequently, our expectation is that Goodman's regression model will show the same robust behavior.

Goodman's regression is implemented by both King and James programs. Goodman's regression serves as a consistency check for the estimation of the more complex models, as well as a diagnostic for the presence of aggregation bias. In King's EI implementation, Goodman's regression is estimated by executing a command within `Gauss` while in `EzI` it is estimated through a command given to the program through the user interface. In the McCue-James software implementation, Goodman's regression is only estimated if the user requests *and* if `AnEI` successfully finds an ecological inference solution. Since `AnEI` would sometimes fail to produce such an estimate, we subsequently rely only on the `Gauss` estimates of Goodman's regression for reported results.

4.2 King's EI

The EI program, and the `Gauss` statistical software, particularly the maximum likelihood algorithm that it is built upon, have numerous option settings that may affect the performance of the program. In the "Frequently Asked Questions" accompanying the EI software documentation, King explicitly states that "The method' proposed in the book is *not* what comes spinning out of EI with all of the globals set to their defaults" (King 1998: 36, original emphasis). He admonishes users to carefully check all diagnostics available to his program, as well as bring to bear common sense drawn from conventional wisdom and any other outside information concerning the program being studied.

Among the numerous option setting available, we focus on three which should, theoretically, improve the accuracy of the estimates generated by EI: the method of numerically calculating central derivatives, White's (1982) QML covariance, and the algorithm to calculate the cumulative bivariate normal distribution.

EI relies upon `Gauss`'s constrained maximum likelihood (`cm1`) library to find a solution. `cm1` uses a variety of derivative based algorithms, and offers a number of options for computing derivatives. The default method is to use "forward" differences, while the most accurate approach is for the user to supply subroutines to compute analytic derivatives (and Hessians). No analytic derivatives are known to exist for EI's likelihood function, and EI uses the default `cm1` method. However, an

intermediate option exists – “central” differences. This is more computationally expensive than the default, but is generally more accurate. (Gill, *et al.* 1981)

We also use the `cml` library’s option to calculate QML covariances. White (1981, 1982) observes that when the variance-covariance matrix can be calculated in multiple ways, and these subsequent matrices differ, then it is an indication of serious model misspecification. White’s test is related to that of Hausman (1978), but is free from some of its well-noted deficiencies (Kramer and Sonnberger 1986, Thursby 1985). The variance-covariance matrix can be calculated from either the expected or observed Fisher’s information matrix (almost sure convergence: White’s (1981) Information Matrix Equivalence theorem), so a comparison of the two variance-covariance matrices has the potential to reveal misspecification resulting from computationally introduced problems provided that the two methods *should* give the same answer in the absence of such problems (e.g., when one is computed with analytical derivatives and the other is computed with forward differences). The procedure defines a vectorized difference of the two matrices: $\nabla = \text{vec}(1/\mathbf{H}(\boldsymbol{\theta}|y) - 1/\mathbf{G}(\boldsymbol{\theta}|y))$, but the test statistic is calculated from a sub-vector, ∇^* , selected such that its asymptotic covariance, C^* , is required to be nonsingular: $W = \nabla^* C^* \nabla^*$. White gives the procedures for Wald and Lagrange tests based on the asymptotic χ^2 distribution of W with degrees of freedom equal to the rank of C^* . While this test has found its way into a number of computing packages, and is recommended in `Gauss`, some authors criticize its utility due to selection of numerically unstable estimates for the sub-vector and associated covariance (Fahrmeir and Tutz 2001, Andrews 1988).

A second source of option dependence is the method used to invert a non-positive definite Hessian. In these circumstances the program uses specialized methods to find a “close”¹⁵ Hessian that is non-singular and therefore invertible. The `_EI_vc` option controls how the EI program attempts a number of methods in sequence, and exits on the execution of the first successful method. As new versions of the program have been developed, new techniques have been devised to handle situations when the Hessian is not strictly positive definite. The sequence of methods applied, when the normal method fails, has also changed. In early versions of the EI, the first specialized method that the program will attempt is documented as a “wide step procedure” or “quadratic approximation with falloff” (King 1998: 9). In later versions of EI, the program attempts a generalized inverse Cholesky alternative proposed in a paper by Gill and King (2002) and based on the Schnabel and Eskow procedure (1990). These methods are not guaranteed to produce meaningful results in all cases; the researcher must exercise caution since there is a paucity of theoretical and empirical work in favor of any particular method. One should also note that non-invertible Hessians may signal limitations in data or in numerical methods, and that the generalized inverse method used by King is justified in terms of the former cause. If the Hessian is non-invertible because, e.g., the likelihood function or derivatives are insufficiently accurate, a more orthodox, and

¹⁵Here the definition of “close” is that the diagonal of the Hessian matrix is changed as little as possible in order to obtain a barely invertible matrix form.

well-studied approach would be to increase the accuracy of these calculations, rather than trying to correct the Hessian at a later stage.¹⁶

The cumulative bivariate normal distribution algorithm is an important factor in determining the shape of the likelihood function for the EI method. The “shape” of the likelihood function determines not only the location of the mode of posterior, but also the Hessian matrix which is the curvature around this mode as measured by the second derivative of the likelihood function at the modal value. The importance of the Hessian matrix is that it produces, by inversion, the variance-covariance matrix of the coefficient estimate. King recognizes that this process is not always straightforward and provides options for users to choose six different methods of calculating the cumulative bivariate normal distribution with the `_Ecdfbvn` option, which we refer to as `cdfbvn` (a.k.a., **C**umulative **D**ensity **F**unction, **B**i**V**ariate **N**ormal). The original default `cdfbvn` is a fast algorithm, but subject to inaccuracies for small values, while the current default represents a tradeoff between accuracy and speed. King once recommended the use of the current default (King 1998: 8), although he now provides a more accurate version.¹⁷

4.3 The McCue-James Approach

Ken McCue does not distribute a program to estimate the model he proposes in his 2001 *American Statistician* article, so instead we use an implementation of the program written for `Stata` by David James, who worked in consultation with McCue. The McCue model is similar to King’s in that it stipulates a bivariate normal specification, but McCue claims that he can avoid the Bayesian-style simulations required in EI by using Lagrange multipliers to get a generalized least squares problem instead of the more complicated posterior that King uses. The article actually leaves out some of the computational and technical details, but the James implementation seems to produce answers as advertised.

Unfortunately, `Stata` does not allow users the same degree of option control as `Gauss`. There is one option available in `Stata`’s maximum likelihood algorithm that may affect numeric accuracy, the “difficult” option, which is only invoked by user request. `Stata`’s documentation of the difficult

¹⁶In fact, in some of our replications, we discovered that while EI produced non-invertible Hessians using the replication settings, the Hessian was invertible when the analysis was run again using a more accurate version of the cumulative bivariate normal distribution function.

¹⁷The improvements to the `cdfbvn` function within EI were made following our earlier numerical accuracy investigations of the EI program. After discussing these results with King, he convinced us that much of this sensitivity could be corrected by increasing the accuracy of the cumulative distribution function for the bivariate normal distribution. We located and consulted an expert in this area, Professor Allen Genz, who supplied us with a quadruple-precision function based on an extension of Drezner and Wesolowsky (1989). After porting this function to `Gauss`, and integrating it into King’s program, we tested the areas of previous instability. These were greatly improved, although not eliminated. The more accurate function has now been incorporated into a new version of King’s programs as an option. This approach to remove numerical inaccuracies may prove fruitful for sophisticated consumers of statistical software.

option is minimal. Similar to King’s `_EI_vc` option, `Stata`’s `difficult` option appears to allow the invocation of a different method to calculate the Hessian: “difficult states that there may be regions where $-\mathbf{H}$ is not invertible and that, in those regions, `ml`’s (`Stata`’s maximum likelihood command) may not work well” (StataCorp. 1999: 385). We suspect the method is more numerically accurate (or at least numerically intense) as the documentation states that the option increases the execution time of the maximum likelihood algorithm.

In our experiments, however, the use of the `difficult` option did not improve the correctness or sensitivity of the solutions. In fact, it prevented convergence in a number of cases, without providing any noticeable change in the results of our analysis. Thus, in the tables below, we use only the default options.

4.4 Observations and Evaluations: Sensitivity and Correctness

We analyze the correctness and pseudo-stability of these three methods for solving the ecological inference problem using both simulated and real data. These simulated data is drawn from eighteen variants of the TBN distribution, each with high, low, or moderate degrees of truncation, positive, zero, or negative correlations, and either 20 or 100 observations, as described in detail by Mattos (chapter XXX).¹⁸ These real data is drawn from the seven example cases described in King’s (1997) book, as supplied in his replication archive.¹⁹ For each of the eighteen simulated and seven replication data sets, we create 50 replications with 1% normally distributed, mean zero error to X and T (using King’s 1997 notation).²⁰

In our estimation of the perturbed data sets, we use the “recommended” option settings for each program. For King’s `EI`, the recommended settings use QML covariance calculations, the most precise version of `cdfbvn`, and excluded those cases where `EI` attempted to find a non-invertible Hessian. For McCue’s `AnEI`, the default option setting proved to be the best. For Goodman’s regression, we used standard defaults.²¹

¹⁸We wish to thank Mattos for generously allowing us to use his data in this paper. Mattos also provided 9 additional datasets with 50 observations each, but we concluded from the results shown in paper that an analysis of this intermediate number of observations was unlikely to be surprising.

¹⁹These data and `Gauss` programs used to replicate the tables in King’s book are available at the ICPSR Publication Replication Archive as study #1132. In referring to this data we follow the naming conventions used in the replication archive.

²⁰We also repeated the experiment with an additional set of perturbations in the form of 5% uniformly distributed error with mean zero. Since the results were not substantially different, we prefer to make our points using the more modest perturbations. The number of replications was chosen based on practice in the literature (see above) and computational tractability – the computations for this study required 6 months of computer time at 1000 MIPS. Moreover, three additional replications using other forms of input perturbations (as above) did not yield substantively different results.

²¹For `AnEI` we replicated all parts of our analysis using the `Stata` “difficult” option. For `EI` we replicated the analysis using the defaults. In addition we reanalyzed the `EI` results including the corrected Hessians. Combined with variations of the perturbations (1% normal and 5% uniform, with and without re-sampling and adjustment for

We are interested in two characteristics of these perturbed data: the sensitivity and correctness of the resulting estimates. Sensitivity (or “pseudo-stability”), described above, refers to the consistency of the estimates across perturbed data sets, where less variation in the results is an indication of less sensitivity to error induced from measurement error or numerical errors found in the algorithm and implementation. We again caution readers that these data perturbations are diagnostic tests, not classical statistical tests. As mentioned above, estimates of perturbed data may be sensitive to other aspects of the model, such as a broad and flat curvature around the mle solution in the data-space (in the multidimensional sense) or data that is ill-conditioned with respect to the estimation. There is no threshold of the degree of statistical significance to attach to the results we observe; however, caution is warranted when small amounts of measurement or numerical error could substantively change the inferences drawn from a statistical model. As a rule of thumb, we recommend caution where the confidence intervals reported in the original estimates are much narrower than the simulated confidence intervals generated when the data is slightly perturbed.

Correctness in this context refers to the reproduction of true values, where estimates closer to the truth are an indication of an internally valid estimation process. For the real data, all but one of King’s examples, **NJ**, are based on aggregate data generated from individual data, so the estimates generated by these models can be compared to the “truth.” Ironically, even a numerically accurate solution is not guaranteed to produce correct answers (answers that match the truth), because the model itself could be misspecified.

Furthermore, we demonstrated above that perturbations introduces bias to the estimates. For these reasons, we do not expect a particular solution to provide a perfectly correct estimate, although we believe it is important to gain insight into how perturbations affect the correctness of the answer.

We begin with a sensitivity analysis using simulated data. Figure 2 shows one indication of the comparative sensitivity of the King, McCue and Goodman methods under different conditions. (Shorter bars are better.) Each inset chart measures the sensitivity of the algorithm and software by plotting the standard deviation of the β^b parameter across fifty perturbation of simulated data. As described above, we examine twelve variations of numbers of observations, type of correlation and degree of truncation.

Several general patterns emerge from this data. First, for all three solutions, sensitivity to perturbations decreases, *ceteris paribus*, as the number of observations in the original data increases. Second, sensitivity to perturbations tends to decrease as the degree of truncation in these data become greater. Third, sensitivity is worst , in all methods, when observations are few, truncation

observations near the boundaries) we examined 32 variations of Table 1. The table presented was based on the most conservative assumptions about noise, and used the options most favorable to each software package. None of these variations, produced results that were substantially different from those in Table 1.

Figure 1: Comparing Stability: King, McCue, Goodman

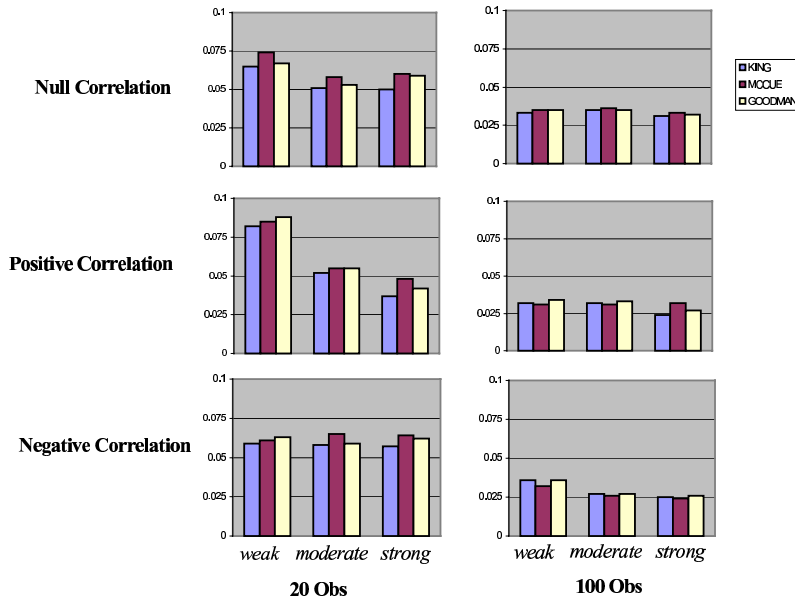


Figure 1: Relative stability (shorter is better) of of the King, McCue and Goodman methods. Based 1% perturbations of simulated data from TBVN distributions with typed of correlation, degrees of truncation, and numbers of observations. The height of the bars show the standard deviation of the BetaB parameter across fifty perturbations, at a scale of [0,0.1].

is weak, and X and T are positively correlated. It is reassuring that these trends follow one's intuition that increasing the amount of information in the data improves the performance of these methods.

Comparatively, despite EI's complexity, King's method is often less sensitive to perturbations than Goodman's regression and McCue's simpler alternative, when the number of observations is small, and the sensitivity to perturbations worst for all methods. When there are a larger number of observations, all three solutions show similar sensitivity to data perturbations, although EI tends to perform the slightly better when there is null correlation or when the degree of truncation is strong and AnEI tends to perform slightly in the presence of correlation and weak or moderate truncation. Surprisingly, despite its simple algorithm, Goodman's regression never exhibits the least sensitivity to data perturbations.

We next turn to the examples provided in King (1997). The analysis includes a run each of the 50 separate perturbed data sets using EI and AnEI, and a further estimate of Goodman's regression implemented in EI. These results are summarized for the 50 separate estimations of each case: for the log likelihood, the coefficients β^W and β^B , and their associated standard errors. For each perturbation run where an algorithm produced an estimate, we list the mean, standard deviation,

and 95% range (2.5% and 97.5% quantiles) based on the observed standard deviation across the 50 runs. For comparison purposes, we report the “truth,” available in all but the NJ example.

TABLE 1 ABOUT HERE

There are some important similarities. All three proposed solutions produced mean estimates of β^b and β^w across the perturbed data sets close to one another and close to the “truth” (not available for NJ) for **Cens1910**, **LAVOTE**, and **MATPROII**. For these data sets, the standard errors of the estimates were also close to one another. For these examples, the three proposed solutions appear to be relatively insensitive to perturbations and to yield estimates close to the truth.

For the remaining three examples, **Fulton**, **KYCK88**, and **SCSP**, Goodman’s regression produced incorrect estimates outside the $[0 : 1]$ unit square, and in addition produced more sensitive estimates than in the other four examples. Goodman’s regression estimates outside the unit square are a diagnostic test recommended by King to identify aggregation bias, which occurs when X and T are both related to an additional vector of variables Z . For these three King examples where Goodman’s regression estimates fell outside the unit square, **FULTON**, **SCSP**, and **KYCK88**, we also note sensitivity to perturbations among the EI and AnEI estimates.

Despite the aggregation bias in **FULTON**, both EI and AnEI estimate β^b close to the “truth.” However, EI has difficulty arriving at a solution, finding an invertible Hessian only 23 of 50 times, and produced solutions that are consistent, but far from the “truth” for β^w . AnEI reaches a solution for all 50 perturbed data sets, which were on average close to the “truth.”

For **KYCK88**, King notes a ridge in the tomography plot for β^b , meaning that the likelihood surface is likely to be flat for a large region around the solution for this parameter. The perturbation results mirror the lack of information in the data; estimates of β^b are sensitive to data perturbations, whereas estimates of β^w are not. Both algorithms produce estimates of β^b that are on average far from the “truth,” with EI generally closer than AnEI. Furthermore, EI finds invertible Hessians for 39 of 50 cases, and AnEI only converges 12 times. Reassuringly, although the range of the β^b estimates is quite large, the standard error of the β^b is similarly large, here and as originally estimated by King (1997).

For **SCSP**, AnEI still performs well but EI again encounters difficulties in inverting the Hessian, resorting to using the inverse Cholesky method of inverting the Hessian for all but 2 of the 50 runs.²² In this case, the large number of warnings encountered during the perturbation runs seems to signal that numerical issues have caused problems for EI – in King’s original analysis, the results

²²Generally, as discussed above, it seemed more conservative to analyze only the runs where no warnings about Hessians were issued. Since this approach sometimes leads to few final observations, such as in this case, we reanalyzed all of our results with the excluded runs included. In no case would including these additional runs substantially affect our conclusions.

were much closer to the “truth.” In contrast, **AnEI** found a solution on average closer to the “truth,” with only small variation in the estimates of the beta coefficients.

In sum, our perturbation analysis of simulated data and examples from King (1997) show that any solution—Goodman’s regression, **EI**, and **AnEI**—may exhibit sensitivity to perturbation and may fail to find an accurate estimate close to the “truth.” In the simulated data analysis, Goodman’s regression is never the least sensitive solution, though sometimes **EI** or **AnEI** are less robust. **EI** performs slightly better than **AnEI**, especially when all methods exhibit sensitivity to perturbations. The analysis of King’s examples shows that when the problem is “easy” all three algorithms are relatively insensitive to perturbations, and yield estimates close to the truth. When the problem is “hard,” when there are indications of aggregation bias or other potential estimation problems such as a flat likelihood surface, then Goodman’s regression produces results that are incorrect. The other two algorithms are better than Goodman’s but neither clearly dominates in terms of robustness to perturbation or correctness of estimates. Since the “truth” is usually not available in ecological inference applications, the analysis presented here demonstrates that caution should be applied to making inference using estimates produced by any currently proposed ecological inference solution when estimates are sensitive to data perturbations.

4.5 Conclusion: We Can Handle the Truth?

The ecological inference problem stands out in social science methodological research because it has shown to be substantively important and there are now multiple competing methods proposed to “solve” this same problem. Goodman’s (1953) proposes a simplistic regression approach, while King’s **EI** (1997) and McCue’s **AnEI** (2001) methods are much more complex.

The presence of multiple, differing methodologies provides the opportunity to study the tradeoff between complexity and accuracy. Indeed, when the problem is “easy,” there is little lost in the tradeoff as all three approaches produce consistent estimates with one another and the “truth.” However, when a difficult problem is encountered, Goodman’s regression may produce invalid results. **EI** and **AnEI**’s assumptions preclude nonsensical results, but at the same time, the complexity of the method slightly can lead to more sensitivity in the underlying statistical computations in particular cases. (Note however, that Goodman’s regression was not consistently the least sensitive method.)

Expert witnesses who report Goodman’s regression in courtroom testimony often must note the nonsensical estimates of more than 100% of a minority group voting for their candidate of choice since ordinary least squares regression is not constrained to fall within the logical bounds from the marginals. These unreasonable estimates, while statistically embarrassing, are routinely interpreted by bounding on $[0, 100]$. King’s and McCue’s methods constrain the estimates within the $[0 : 1]$ interval to avoid obvious errors that may occur with Goodman’s regression. Ironically, while McCue’s method is promoted on computational grounds, we see no evidence that it performs

consistently better.

Using the recommended, most numerically accurate options may help, but does not guarantee success. King's admonishments to users of his EI program should be well-heeded. Users should carefully scrutinize all available diagnostic tools available to them, including common sense. To the diagnostic tests incorporated into the EI program, we herein suggest data perturbations as an additional diagnostic test: Numeric and data problems can be revealed by failure to converge across a large proportion of slightly perturbed data sets, or by variance of the estimated coefficients that substantially exceeds the originally reported confidence intervals

These results raise the question: is numerical accuracy for ecological inference related to the statistical problems? Changes in estimated results may be more sensitive when the problem is ill-conditioned, where small changes in these data may result in relatively large changes in the shape of the likelihood function to be estimated. As such, data perturbations may provide another diagnostic test for the presence of aggregation bias or for other mismatches between the data and the statistical assumptions of the model. It is not a procedure recommended for initial data exploration, as the amount of time to conduct this suggested test increases linearly with the number of perturbed data sets – but it should be used before publication of results.

5 References

- Achen, Christopher H. 2003. "An Agenda for the New Political Methodology: Microfoundations and ART." *Annual Review of Political Science*, Forthcoming. Also available as a Political Methodology Working paper, <http://web.polmeth.ufl.edu>.
- Altman, Micah, Jeff Gill, and Michael P. McDonald. 2003. *Numerical Methods in Statistical Computing for the Social Sciences*, New York: John Wiley & Sons.
- Altman, Micah and Michael P. McDonald. 2003. "Replication with Attention to Numerical Accuracy" *Political Analysis*. Forthcoming.
- Altman, Micah and Michael P. McDonald. 2001. "Choosing Reliable Statistical Software" *PS: Political Science and Politics*. **XXXIV**, 681-7.
- Andrews, D. W. K. 1988. "Chi-Square Diagnostic Tests for Econometric Models: Theory." *Econometrica* **56**, 1419-53.
- Atkinson, A. C. 1980. "Tests of Pseudo-Random Numbers." *Applied Statistics* **29**, 164-71.
- Albert E. Beaton, Donald B. Rubin, John L. Barone 1976. "The Acceptability of Regression Solutions: Another Look at Computational Accuracy" *Journal of the American Statistical Association* **71**, 158-168.
- Albert E. Beaton, Donald B. Rubin, John L. Barone 1977. "More on Computational Accuracy in Regression: Comment" *Journal of the American Statistical Association* **72**, 600-601.

- Butcher, J. C. 1961. "A Partition Test for Pseudo-Random Numbers." *Mathematics of Computation* **15**, 198-9.
- Cook, R. Dennis. 1986. "Assessment of Local Influence." *Journal of the Royal Statistical Society* **48**, 133-69.
- Cook, R. Dennis and Sanford Weisberg. 1982. *Residuals and Influence in Regression*. New York: Chapman & Hall.
- Coveyou, R. R. 1960. "Serial Correlation in the Generation of Pseudo-Random Numbers." *Journal of the Association for Computing Machinery* **7**, 72-4.
- Coveyou, R. R. 1970. "Random Numbers Fall Mainly in the Planes (Review)." *ACM Computing Reviews*, 225.
- Coveyou, R. R. and R. D. MacPherson. 1967. "Fourier Analysis of Uniform Random Number Generators." *Journal of the ACM* **14**, 100-19.
- Cox, D. R. and D. V. Hinkley. 1974. *Theoretical Statistics*. New York: Chapman & Hall.
- Davidson, Russell and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
- Downham, D. Y. 1970. "The Runs Up and Test." *Applied Statistics* **19**, 190-92.
- Drezner, Z. and G.O. Wesolowsky. 1989. "On the Computation of the Bivariate Normal Integral." *Journal of Statistical Computation and Simulation* **35**, 101-7
- Dudewicz, E. J. 1976. "Speed and Quality of Random Numbers for Simulation." *Journal of Quality Technology* **8**, 171-8.
- Duncan, Otis Dudley and Beverly Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* **18** 665-66.
- Fahrmeir, L. and G. Tutz. 2001. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Second Edition. New York: Springer-Verlag.
- Ferree, Karen, (1999). Iterative Approaches to $R \times C$ Ecological Inference Problems: Where They Can Go Wrong. Presented at *Summer Methods Conference*, July 1999, College Station, TX.
- Fishman, G. S. and L. R. Moore. 1982. "A Statistical Evaluation of Multiplicative Congruential Random Number Generators With Modulus $2^{31} - 1$." *Journal of the American Statistical Association* **77**, 129-36.
- Freedman, D. A., M. Ostland, M. R. Roberts, and S. P. Klein. 1999. "Response to King's Comment." *Journal of the American Statistical Association* **94**, 355-7.
- Freedman, D. A., S. P. Klein, M. Ostland, and M. R. Roberts. 1998. "Review of *A Solution to the Ecological Inference Problem*, by Gary King." *Journal of the American Statistical Association* **93**, 1518-22.
- Gentle, J. E. 1999. *Random Number Generation and Monte Carlo Methods*. New York: Springer-

Verlag.

- Gill, Jeff. and Gary King. 2002. "What to do When Your Hessian is Not Invertible: Alternatives to Model Respecification in Nonlinear Estimation." *University of Florida, Department of Political Science Technical Report*, <http://web.clas.ufl.edu/jgill/papers/help.pdf>
- Gill, Phillip E., Walter Murray, and Margaret H. Wright. 1981. *Practical Optimization*. San Diego: Academic Press, Inc.
- Good, I. J. 1957. "On the Serial Test for Random Sequences." *Annals of Mathematical Statistics* **28**, 262-4.
- Goodman, Leo. 1953. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* **18**, 663-6.
- Gorenstein, S. 1967. "Testing a Random Number Generator." *Communications of the Association for Computing Machinery* **10**, 111-8.
- Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica* **46**, 1251-71.
- Hellekalek, P. 1998. "Good Random Number Generators Are (not so) Easy to Find." *Mathematics and Computers in Simulation* **46**, 485-505.
- Higham, Nicholas J. 1996. *Accuracy and Stability of Numerical Algorithms*. Philadelphia: SIAM.
- Higham, Nicholas J. 2002. *Accuracy and Stability of Numerical Algorithms. Second edition* Philadelphia: SIAM.
- King, Gary. 1999. "The Future of Ecological Inference Research: A Comment on Freedman, *et al.*" *Journal of the American Statistical Association* **94**, 352-5.
- King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.
- Knuth, Donald E. 1997. *The Art of Computer Programming*. Third Edition. Reading, MA: Addison-Wesley.
- Kramer, W. and H. Sonnberger. 1986. "Computational pitfalls of the Hausman test." *Journal of Economic Dynamics and Control* **10**, 163-5. Kramer, W. ; Sonnberger, H.
- Krawczyk, H. 1992. "How to Predict Congruential Generators." *Journal of Algorithms* **13**, 527-45.
- Kronmal, R. 1964. "The Evaluation of a Pseudorandom Normal Number Generator." *Journal of the Association for Computing Machinery* **11**, 357-63.
- Lawrance, A.J., 1988, "Regression Transformation Diagnostics Using Local Influence." *Journal of the American Statistical Association* **83**, 1067-72.
- Learmonth, G. P. and P. A. W. Lewis. 1973. "Some Widely Used and Recently Proposed Uniform Random Number Generators." *Proceedings of Computer Science and Statistics: Seventh Annual Symposium on the Interface*, W. J. Kennedy (ed.). Ames, IA: Iowa State University. pp. 163-71.
- Marsaglia, G. 1968. "Random Numbers Fall Mainly in the Planes." *Proceedings of the National*

- Academy of Sciences* **61**, 25-8.
- McArdle, J. J. 1976. "Empirical Test of Multivariate Generators." In *Proceedings of the Ninth Annual Symposium on the Interface of Computer Science and Statistics*, D. C. Hoaglin and R. Welsch (eds.). Boston: Prindle, Weber, and Schmidt. pp. 263-7.
- McCue, Kenneth. 2001. "The Statistical Foundations of the 'EI' Method." *The American Statistician* **55**, 106-11.
- McCullough, B. D. 1998. "Assessing the Reliability of Statistical Software: Part I." *The American Statistician* **53**, 149-59.
- McCullough, B. D. 1999a. "Econometric Software Reliability: Eviews, LIMDEP, SHAZAM, and TSP." *Journal of Applied Econometrics* **14**, 191-202.
- McCullough, B. D. 1999b. "Assessing the Reliability of Statistical Software: Part II." *The American Statistician* **53**, 149-59.
- McCullough, B. D. and B. Wilson. 1999. "On the Accuracy of Statistical Procedures in Microsoft Excel 97." *Computational Statistics and Data Analysis* **31**, 27-37.
- McCullough, B. D., H. Vinod, 1999. "The Numerical Reliability of Econometric Software." *Journal of Economic Literature* **37**, 633-65
- Morgan, Byron J. T. 1984. *Elements of Simulation*. New York: Chapman & Hall.
- Parker, D. Stott. 1997. "Monte Carlo Arithmetic: Exploiting Randomness in Floating-Point Arithmetic" Technical Report CSD-970002, UCLA Computer Science Dept.
<http://www.cs.ucla.edu/stott/mca/>
- Parker, D. Stott, Brad Pierce, and Paul R. Eggert. 2000. "Monte Carlo Arithmetic." *Computing in Science and Engineering* **July**, 58-68.
- Polasek, W. 1987, "Bounds on Rounding Errors in Linear Regression Models" *Statistician* 36(2): 221-7.
- Pregibon, D. 1981. "Logistic Regression Diagnostics." *Annals of Statistics* **9**, 705-24.
- Press, William H., Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery. 2002. *Numerical Recipes in C++: The Art of Scientific Computing*. Second Edition. Cambridge: Cambridge University Press.
- Revkin, Andrew C. 2002. "Data Revised on Soot in Air and Deaths." *New York Times*, June 5.
- Ripley, Brian D. 1988. "Uses and Abuses of Statistical Simulation." *Mathematical Programming* **42**, 53-68.
- Ripley, Brian D. 1987. *Stochastic Simulation*. New York: John Wiley & Sons.
- Rogers, Janet, James Filliben, Lisa Gill, William Guthrie, Eric Lagergren and Mark Vangel. (2000). StRD: Statistical Reference Datasets for Testing the Numerical Accuracy of Statistical Software. *NIST # 1396*. Washington, DC: National Institute of Standards and Technology.

- Schnabel, Robert B. and Elizabeth Eskow. 1990. "A New Modified Cholesky Factorization." *SIAM Journal of Scientific Statistical Computing* **11**, 1136-58.
- StataCorp., 1999. "*Stata Statistical Software Release 6.0*. College Station, TX: Stata Corporation.
- St. Laurent, Roy T. and R. Dennis Cook. 1993. "Leverage, Local Influence and Curvature in Nonlinear Regression." *Biometrika* **80**, 99-106.
- Tam Cho, Wendy K. 1998. "Iff the Assumption Fits...: A Comment on the King Ecological Inference Model." *Political Analysis* **7**, 143-64.
- Thisted, R. 1988. *Elements of Statistical Computing*. New York: Chapman & Hall.
- Thursby, Jerry G. 1985. "The Relationship Among the Specification Tests of Hausman, Ramsey, and Chow." *Journal of the American Statistical Association* **80**, 926-8.
- Toothill, J. P. R., W. D. Robinson, and A. G. Adams. 1971. "The Runs Up and Down Performance of Tausworthe Pseudo-Random Number Generators." *Journal of the Association for Computing Machinery* **18**, 381-99.
- White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* **50**, 1-26.
- White, Halbert. 1981. "Consequences and Detection of Misspecified Nonlinear Regression Models." *Journal of the American Statistical Association* **76**, 419-33.
- Whittlesey, J. R. B. 1969. "On the Multidimensional Uniformity of Pseudo-Random Generators." *Communications of the Association for Computing Machinery* **12**, 247.

Table 1: Comparison of Three Ecological Inference Solution Algorithms

King Example	Parameter	King EI (v1.7) (best settings, invertible Hessians)	McCue-James AnEI default settings	Goodman's Regression	Truth	
CENSI910	(obs=1040)	LL	2.405e+03 (1.2e+01) [2.378e+03,2.432e+03] n=50	2.404e+03 (5.8e+00) [2.395e+03,2.412e+03] n=7	-	-
		β^b	6.396e-01 (1.8e-03) [6.359e-01,6.435e-01] n=50	5.742e-01 (2.2e-03) [5.708e-01,5.773e-01] n=7	6.125e-01 (1.1e-03) [6.099e-01,6.146e-01] n=50	7.34e-01
		$SE(\beta^b)$	3.438e-03 (2.8e-04) [2.927e-03,3.973e-03] n=50	1.179e-01 (1.2e-03) [1.161e-01,1.198e-01] n=7	6.542e-03 (5.0e-05) [6.500e-03,6.600e-03] n=50	-
		β^w	9.494e-01 (9.4e-04) [9.476e-01,9.513e-01] n=50	9.502e-01 (1.3e-03) [9.485e-01,9.517e-01] n=7	9.347e-01 (6.2e-04) [9.335e-01,9.360e-01] n=50	9.339e-01
		$SE(\beta^w)$	1.426e-03 (1.2e-04) [1.200e-03,1.672e-03] n=50	4.542e-02 (4.4e-04) [4.457e-02,4.577e-02] n=7	3.786e-03 (3.5e-05) [3.700e-03,3.800e-03] n=50	-
		LL	5.865e+02 (2.4e+00) [5.820e+02,5.929e+02] n=24	7.260e+02 (7.1e+00) [7.123e+02,7.417e+02] n=50	-	-
FULTON	(obs=289)	β^b	5.724e-01 (1.2e-03) [5.703e-01,5.763e-01] n=24	5.482e-01 (3.1e-03) [5.407e-01,5.553e-01] n=50	6.725e-01 (2.4e-03) [6.674e-01,6.779e-01] n=50	5.530e-01
		$SE(\beta^b)$	2.129e-03 (1.7e-04) [1.800e-03,2.500e-03] n=24	9.725e-02 (9.6e-04) [9.463e-02,9.925e-02] n=50	1.179e-02 (1.4e-04) [1.143e-02,1.200e-02] n=50	-
		β^w	3.826e-02 (1.1e-03) [3.620e-02,4.030e-02] n=24	6.366e-02 (6.9e-03) [5.127e-02,8.259e-02] n=50	-1.947e-01 (5.0e-03) [-2.061e-01,-1.841e-01] n=50	6.472e-02
		$SE(\beta^w)$	5.033e-03 (4.2e-04) [4.300e-03,6.000e-03] n=24	5.230e-02 (3.4e-03) [4.678e-02,6.572e-02] n=50	2.437e-02 (2.9e-04) [2.380e-02,2.480e-02] n=50	-
		LL	2.140e+02 (1.1e+00) [2.118e+02,2.168e+02] n=39	2.165e+02 (2.2e+00) [2.134e+02,2.195e+02] n=12	-	-
		β^b	4.152e-01 (5.1e-02) [3.520e-01,6.333e-01] n=39	3.752e-01 (8.8e-02) [1.291e-01,4.525e-01] n=12	-4.139e-01 (4.5e-02) [-5.021e-01,-2.968e-01] n=50	6.660e-01
KYCK88	(obs=118)	$SE(\beta^b)$	9.954e-02 (2.3e-02) [6.210e-02,1.828e-01] n=39	1.579e-01 (2.3e-02) [1.381e-01,2.225e-01] n=12	2.056e-01 (4.9e-03) [1.928e-01,2.147e-01] n=50	-
		β^w	7.705e-01 (4.4e-03) [7.546e-01,7.783e-01] n=39	8.209e-01 (4.2e-03) [8.163e-01,8.324e-01] n=12	8.516e-01 (2.3e-03) [8.460e-01,8.552e-01] n=50	7.533e-01
		$SE(\beta^w)$	7.079e-03 (1.6e-03) [4.500e-03,1.340e-02] n=39	1.023e-01 (2.3e-03) [9.704e-02,1.055e-01] n=12	1.173e-02 (1.9e-04) [1.123e-02,1.200e-02] n=50	-
		LL	2.140e+02 (1.1e+00) [2.118e+02,2.168e+02] n=39	2.165e+02 (2.2e+00) [2.134e+02,2.195e+02] n=12	-	-

Table 1: (Continued) Comparison of Three Ecological Inference Solution Algorithms

King Example	Parameter	King EI (v1.7) (best settings, invertible Hessians)	McCue-James AnEI default settings	Goodman's Regression	Truth	
LAVOTE	(obs=3262)	LL	6.784e+03 (8.0e+00)	6.913e+03 (1.0e+01)	–	–
		β^b	[6.768e+03,6.799e+03] n=50 6.252e-01 (5.1e-04)	[6.890e+03,6.931e+03] n=50 5.627e-01 (4.8e-04)	6.278e-01 (4.7e-04)	6.168e-01
		SE(β^b)	[6.241e-01,6.263e-01] n=50 1.120e-03 (7.8e-05)	[5.618e-01,5.637e-01] n=50 2.182e-01 (2.8e-04)	[6.267e-01,6.288e-01] n=50 3.386e-03 (3.5e-05)	–
		β^w	[1.000e-03,1.300e-03] n=50 7.067e-01 (2.6e-04)	[2.175e-01,2.187e-01] n=50 7.089e-01 (2.3e-04)	[3.300e-03,3.400e-03] n=50 7.111e-01 (2.3e-04)	6.842e-01
		SE(β^w)	[7.060e-01,7.072e-01] n=50 4.020e-04 (2.5e-05)	[7.084e-01,7.094e-01] n=50 7.096e-02 (2.5e-04)	[7.107e-01,7.116e-01] n=50 1.794e-03 (2.4e-05)	–
			[3.275e-04,5.000e-04] n=50	[7.050e-02,7.163e-02] n=50	[1.700e-03,1.800e-03] n=50	
MATPROII	(obs=268)	LL	4.072e+02 (2.6e+00)	4.159e+02 (2.5e+00)	–	–
		β^b	[3.994e+02,4.112e+02] n=41 6.143e-01 (6.8e-03)	[4.112e+02,4.209e+02] n=50 5.255e-01 (9.7e-03)	5.152e-01 (4.6e-03)	5.847e-01
		SE(β^b)	[6.006e-01,6.279e-01] n=41 3.602e-02 (2.6e-03)	[5.045e-01,5.432e-01] n=50 1.947e-01 (3.0e-03)	[5.061e-01,5.249e-01] n=50 4.547e-02 (2.5e-04)	–
		β^w	[2.835e-02,4.000e-02] n=41 8.078e-01 (2.4e-03)	[1.891e-01,2.024e-01] n=50 8.565e-01 (3.2e-03)	[4.485e-02,4.597e-02] n=50 8.673e-01 (1.6e-03)	8.242e-01
		SE(β^w)	[8.009e-01,8.139e-01] n=41 1.022e-02 (7.8e-04)	[8.508e-01,8.639e-01] n=50 1.255e-01 (1.1e-03)	[8.630e-01,8.703e-01] n=50 1.787e-02 (9.7e-05)	–
			[8.040e-03,1.149e-02] n=41	[1.231e-01,1.277e-01] n=50	[1.760e-02,1.807e-02] n=50	

Table 1: (Continued) Comparison of Three Ecological Inference Solution Algorithms

King Example	Parameter	King EI (v1.7) (best settings, invertible Hessians)	McCue-James AnEI default settings	Goodman's Regression	Truth
NJ (obs=268)	LL	1.038e+03 (3.3e+00) [1.031e+03,1.045e+03] n=50	1.097e+03 (6.6e+00) [1.081e+03,1.111e+03] n=50	-	-
	β^b	6.136e-02 (2.7e-03) [5.636e-02,6.890e-02] n=50	7.508e-02 (3.5e-03) [6.686e-02,8.368e-02] n=50	2.267e-02 (3.3e-03) [1.285e-02,3.015e-02] n=50	-
	SE(β^b)	6.602e-03 (5.4e-04) [5.528e-03,7.718e-03] n=50	2.542e-02 (8.6e-03) [1.250e-02,5.127e-02] n=50	2.676e-02 (1.4e-04) [2.643e-02,2.707e-02] n=50	-
	β^w	3.796e-01 (1.0e-03) [3.778e-01,3.817e-01] n=50	4.003e-01 (6.0e-04) [3.989e-01,4.015e-01] n=50	4.150e-01 (6.5e-04) [4.139e-01,4.164e-01] n=50	-
	SE(β^w)	1.528e-03 (1.3e-04) [1.300e-03,1.772e-03] n=50	7.929e-02 (1.1e-03) [7.705e-02,8.105e-02] n=50	4.850e-03 (5.1e-05) [4.800e-03,4.900e-03] n=50	-
	LL	5.613e+03 (1.2e+00) [5.612e+03,5.614e+03] n=2	5.317e+03 (9.5e+00) [5.295e+03,5.334e+03] n=50	-	-
SCSP (obs=3185)	β^b	5.720e-02 (7.4e-03) [5.200e-02,6.240e-02] n=2	1.454e-01 (6.6e-04) [1.438e-01,1.469e-01] n=50	-1.847e-01 (5.0e-03) [-1.942e-01,-1.722e-01] n=50	1.313e-01
	SE(β^b)	3.050e-03 (9.2e-04) [2.400e-03,3.700e-03] n=2	1.004e-01 (1.1e-03) [9.817e-02,1.029e-01] n=50	2.486e-02 (9.7e-05) [2.470e-02,2.507e-02] n=50	-
	β^w	2.429e-01 (6.9e-03) [2.380e-01,2.478e-01] n=2	1.777e-01 (5.9e-04) [1.764e-01,1.789e-01] n=50	4.838e-01 (4.6e-03) [4.725e-01,4.927e-01] n=50	1.732e-02
	SE(β^w)	2.800e-03 (8.5e-04) [2.200e-03,3.400e-03] n=2	1.693e-01 (8.7e-04) [1.672e-01,1.713e-01] n=50	2.276e-02 (8.8e-05) [2.260e-02,2.290e-02] n=50	-
	LL	5.613e+03 (1.2e+00) [5.612e+03,5.614e+03] n=2	5.317e+03 (9.5e+00) [5.295e+03,5.334e+03] n=50	-	-