B.J.Pol.S. **27,** 83–110
Printed in Great Britain
Copyright © 1997 Cambridge University Press

Review Article: Of Silicon and Political Science – Computationally Intensive Techniques of Statistical Estimation and Inference

CHRISTOPHER Z. MOONEY AND GEORGE A. KRAUSE*

If the automobile business had developed like the computer business, a Rolls Royce would cost $2.75 and run for 3 million miles on one gallon of gas.

Forrester, *High-Tech Society*[1]

The computer power available to political scientists has increased at least one thousandfold since 1970. We can now make millions of calculations in a few seconds, generate beautiful and meaningful graphics, and communicate instantly around the world, all at virtually no cost. But in what ways has this computer power helped us make more accurate statements about political phenomena? Beyond simply doing faster and more cheaply that which we had done previously, this computer revolution has made possible a qualitative leap in our analysis of political behaviour.

This article reviews three important classes of statistical techniques that increased computer power has made feasible in recent years: numerical optimization for point estimation; Monte Carlo simulation; and bootstrapping for statistical inference. These procedures have already been used by some political scientists, to a greater or lesser extent, but it is important that these techniques be understood by all political scientists who conduct or read statistical analysis as they become more widely used in the discipline. We survey each of these techniques in turn, describing what it does, the advantages it holds over the traditional approach for which it is an alternative, and some problems that can arise in its application. Our goal is not to teach the reader how to execute these techniques, but rather to provide enough information for the reader to begin a knowledgeable exploration of them in the literature cited.

* Department of Political Science, West Virginia University; and Department of Government and International Studies, University of South Carolina, respectively. This is a revised version of a paper presented at the Political Studies Association Annual Conference, University of York, April 1995. We would like to thank Neil Berch, Albert Weale and David Lublin for comments on this project, and Burak Saltoglu and David Cohen for research assistance. While working on this project, Dr Mooney was supported by the West Virginia University Senate, the University of Essex Research Promotion Fund and the British Academy, and was a visiting lecturer at the University of Essex.

[1] Tom Forrester, *High-Tech Society* (Oxford: Basil Blackwell, 1987), p. 18.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

STATISTICAL THEORY AND COMPUTATION

Statistics is a practical discipline, whose goal is to solve the data analytic problems of other scientific disciplines. These solutions, in the form of statistical models and techniques, are developed under the practical constraints of the mathematical theory and computational power existent at the time of their development. The goal is to model a process as closely as possible according to substantive theory while working within these constraints.

Ordinary Least Squares estimation (OLS) of linear regression coefficients and the parametric approach to inference regarding these coefficients is a classic example of how a less than ideal statistical technique was developed to fulfil the practical requirements of its age. In debating the best way to generate point estimates of regression coefficients at the turn of the nineteenth century, Laplace favoured the method of least absolute deviations (LAD), while Gauss advocated a least squared deviation (OLS) approach.[2] Laplace's argument was that LAD handled unusual distributions of error better than OLS. While conceding this point, Gauss argued that LAD was not mathematically convenient enough to be practical. The persuasiveness of this practicality argument is reflected in the widespread use of OLS (as opposed to the very limited use of LAD) almost 200 years later. Gauss also appreciated the need to understand the sampling distribution of a regression slope estimate in order to make inferences to its population parameters,[3] but could not do so analytically, that is using calculus, without knowing the distribution of the error term in the model. Gauss settled on the strategy of assuming that the error term was distributed normally, primarily because this made the problem mathematically tractable, and only later did he show that this assumption was plausible in many situations.

From its development in the eighteenth century, differential calculus was the primary tool used to develop point estimates of model parameters and to make inferences to them. But the limits of this analytic approach for understanding general, non-normally distributed estimators were all too apparent by the turn of the twentieth century,[4] and statisticians began considering experimental mathematics as an alternative. That is, they attempted to overcome the limits of their analytical techniques by increasing the level of computation that was brought to bear on their problems. Unfortunately, these statisticians were restricted by having only very cumbersome calculating devices with which to undertake these experiments. In the 1890s, Karl Pearson was only able to complete the breakthrough work on his family of skewed distributions after he purchased a Brunsviga calculator, an over 50 lb. ten-key affair, which he used

[2] Bradley Efron, 'Computers and the Theory of Statistics: Thinking the Unthinkable', *SIAM Review*, 21 (1979), 460–80, p. 476.

[3] R. L. Plackett, 'The Discovery of the Method of Least Squares', *Biometrika*, 59 (1972), 239–51, p. 248.

[4] Karl Pearson, 'Notes on the History of Correlation', *Biometrika*, 13 (1920), 25–45.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

until his death in 1936.[5] Great use was also made of clerks who would undertake the tedious task of evaluating algebraic formulae and recording the results with pencil and paper. The famous *Biometrika* statistical tables were developed in this way under Pearson's supervision between 1914 and 1934.[6] The image conjured up by these industrious clerks is one of Bartleby the Scrivner or of a Roman galley, rather than one of the development of the statistical tools that are the standard in political science today.

And to say that these techniques are the standard in contemporary political science is not stretching a point. OLS, for example, is by far the most commonly used single statistical technique in the discipline. A survey of the 1993 volumes of five top political science journals revealed that fully 35.6 per cent of all the articles contained therein reported OLS results.[7] This predominant use of OLS regression is probably influenced by its reasonable approximation of the theoretical models in which political scientists are interested, but it is also undoubtedly influenced by the inertia of using a method that was accepted because of its mathematical simplicity in the pre-computer era of the early twentieth century.

In the late 1930s, almost simultaneously with Karl Pearson's death, and the standardization of OLS and of Fisher's and Neyman's work on parametric inference,[8] a revolution in computing technology began with the development of the first embryonic electronic computers.[9] These machines used electricity to overcome the problems of friction and power source inherent in mechanical calculators. In 1946, the first practical, fully digital electronic computer as we would recognize it today went on line in Philadelphia – the ENIAC. The ENIAC used vacuum tubes and a digital language to calculate 5,000 additions per second – an enormous advance over clerk-driven mechanical calculation.[10]

The history of computing power since 1946 is one of exponential growth, measured by whatever standard and over whatever time period. As transistors replaced vacuum tubes, integrated circuits replaced transistors, and gallium arsenide replaced silicon, as architecture became more efficient and memory mushroomed, and as the size of these machines microscoped from 15,000 square feet to less than two, the ability to do calculations quickly has increased at a rate unparalleled in the history of any human technological development. Figure 1 shows the pattern of this growth throughout the twentieth century using a scale

[5] E. S. Pearson, 'Some Incidents in the Early History of Biometry and Statistics 1890–1894', *Biometrika*, 52 (1965), 3–18.

[6] J. B. S. Haldane, 'Karl Pearson, 1857 (1957): A Centenary Lecture Delivered at University College London', *Biometrika*, 44 (1957), 303–13.

[7] The journals surveyed were *British Journal of Political Science*, *American Political Science Review*, *Journal of Politics*, *American Journal of Political Science*, and *Political Research Quarterly*.

[8] Ronald A. Fisher, 'The Logic of Inductive Inference' (with discussion), *Journal of the Royal Statistical Society*, 98 (1935), 39–82; A. C. Aitken, 'On Least Squares and Linear Combinations of Variables', *Proceedings of the Royal Society of Edinburgh*, 55 (1935), 42–8.

[9] Paul E. Ceruzzi, *Reckoners* (Westport, Conn.: Greenwood, 1983).

[10] William Stallings, *Computer Organization and Architecture*, 3rd edn (New York: Macmillan, 1993), pp. 22–3.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

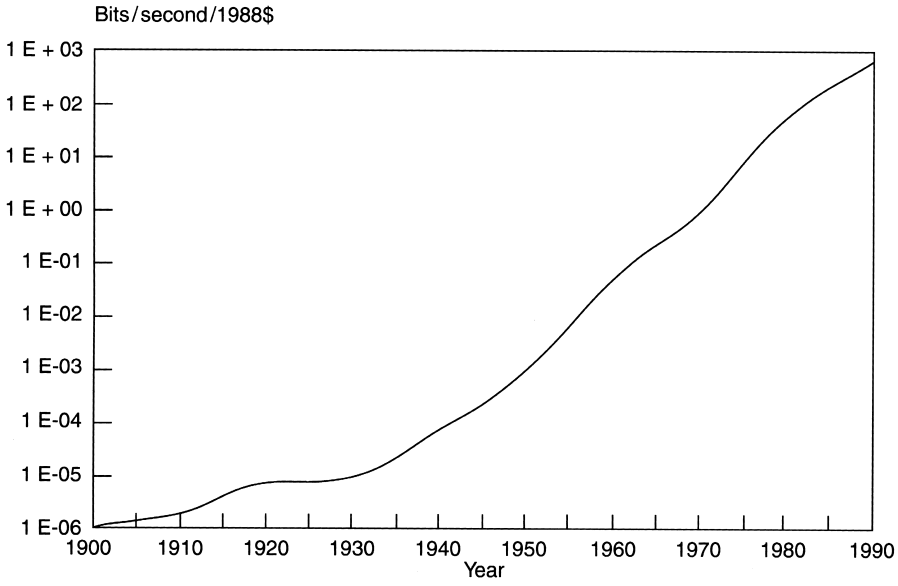


Fig. 1. Evolution of computer power in the twentieth century
Source: John D. Barrow, *Theories of Everything* (London: BCA, 1991), p. 145.

assessing speed in relation to cost. It is clear from this figure that this quantity has not only increased steadily up until today, but that this increase has been *exponential* (note the *y*-axis). For example, whereas the massive and expensive ENIAC could conduct 5,000 additions per second, by July 1995 the state of the art in personal computing, the tiny and relatively inexpensive Pentium 100 chip, could conduct over 500 million operations per second – a 100,000-fold increase in speed alone. And this trend showed no sign of stopping: Digital Equipment Corporation had even then announced the development of the Alpha AXPZ1164 chip, which is capable of over 1,000 million operations per second.[11]

With the old computational constraints lifted, statisticians have been busy developing models and techniques to allow empirical researchers more flexibility in their analyses. Non-parametric statistics, the jackknife, cross-validation, error-rate estimation in discriminant analysis, robust estimation, graphic analysis techniques and exploratory data analysis are but a few of the directions this computational revolution has inspired in the past fifty years.[12]

[11] *Washington Post*, 'New Chip Introduced by Digital: Microprocessor Called the Fastest in the World,' 8 September 1994, sec. B.
[12] Efron, 'Computers and the Theory of Statistics'.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

The techniques reviewed in this article – numerical optimization, Monte Carlo simulation and bootstrapping – have been selected for discussion for two reasons. First, they are among the most general of the new computationally intensive techniques. They represent *approaches* to estimation (numerical optimization) and inference (Monte Carlo simulation and bootstrapping), not specific estimators or models. This means that political scientists working with a variety of types of models and data may well encounter and use these techniques. The second reason we focus on these three techniques is that they are self-conscious computational spendthrifts. That is, they take obvious advantage of high-speed computation in a way that makes it quite obvious why they would not have been feasible on a large scale until quite recently. This is important in that they show the possibilities of making a clean break with past constraints, and indicate the way in which statistical estimation and inference are likely to be carried out in the future.

NUMERICAL OPTIMIZATION

The first logical step in making statements about the fit of a statistical model to a given set of data is to make point estimates of the parameters of that model. Even though most political scientists are primarily concerned with assessing the probability that a parameter has a certain theoretically interesting value given the data (for example, that a regression coefficient is equal to zero), point estimation is of vital concern for at least two reasons. First, it must be undertaken prior to any inferential testing – one cannot conduct a hypothesis test without first making a point estimate of the parameter involved. And second, the point estimates of ancillary model parameters, such as standard deviations, are used directly in conducting hypothesis tests for parameters of more theoretical interest.

Most point estimation undertaken by political scientists today proceeds analytically and requires a clearly defined mathematical solution. This is the approach used any time a standard estimation formula is applied to data. A familiar example of analytic estimation is conventional OLS point estimation. In the OLS estimation of a bivariate linear model, one uses the following well-known formulae to estimate the slope and intercept of the line that minimizes the sum of the squared residuals for a set of data:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}; \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}, \tag{1}$$

where X is the independent variable and Y is the dependent variable. These formulae yield point estimates with desirable properties under certain

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

conditions, according to the Gauss–Markov theorem.[13] Another example of this sort of estimation is in the analytic solution of a maximum likelihood problem using calculus.[14] Analytic parameter estimation is by far the most common general estimation procedure used in political science, as indeed it is in all social science disciplines.

Unfortunately, as we move away from the simplest models and estimators, analytic estimation cannot always be applied. It is often mathematically difficult or impossible to find analytic solutions for models that exhibit, among other things, non-linear relationships between the dependent and independent variables, non-normally distributed error structures, and dependent variables that are highly skewed,[15] to name but a few instances. Because of this, political scientists relying solely on analytic estimation must 'trick' a simple (and mathematically tractable) model into approximating the model that substantive theory suggests is best. For example, political scientists wishing to model non-linear policy impacts over time have often used linear models with a linear trend variable[16] or pulse/decay shifts modelled via Box–Jenkins techniques.[17] Dependent variables that come in the form of event counts (for example, the number of US states adopting a given policy, the number of prime ministerial appointments, or the number of coups in African countries) are often treated incorrectly as continuous data and subjected to OLS analysis.[18] While more theoretically appropriate models could have been written out algebraically in these cases, their parameters cannot generally be estimated analytically.

In response to these limitations of analytic estimation, numerical optimization methods of statistical estimation have been developed.[19] This class of techniques places an emphasis on computational power, as opposed to analytical solution, to arrive at an estimate of a model's parameters. This is generally implemented in an iterative manner, with the algorithm 'trying out' a large number of potential values of the parameters in a systematic fashion until the 'best' estimates are achieved, based on some criterion.[20] The criterion to be optimized varies with the model being estimated, and the algorithm being used. For example, non-linear least squares (NLLS) models estimated using numerical optimization

[13] Robert S. Pindyck and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts*, 3rd edn (New York: McGraw-Hill, 1991) pp. 46–54.

[14] Gary King, *Unifying Political Methodology* (New York: Cambridge University Press, 1989), p. 67.

[15] Jan Kmenta, *Elements of Econometrics* (New York: Macmillan, 1986), pp. 503–26.

[16] For example, Marc A. Eisner and Kenneth J. Meier, 'Presidential Control Versus Bureaucratic Power: Explaining the Reagan Revolution in Anti-Trust', *American Journal of Political Science*, 34 (1990), 269–287.

[17] For example, Richard J. Timpone, 'Mass Mobilization or Government Intervention? The Growth of Black Registration in the South', *Journal of Politics*, 57 (1995), 425–42.

[18] For example, see Thomas H. Johnson, Robert O. Slater and Pat McGowan, 'Explaining African Military Coups d'Etat, 1960–1982', *American Political Science Review*, 78 (1984), 622–40.

[19] King, *Unifying Political Methodology*, pp. 66–73.

[20] For more detailed information on numerical optimization, see Douglas M. Bates and Donald G. Watts, *Nonlinear Regression Analysis and its Applications* (New York: Wiley/Bates and Watts,

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

require the minimization of the residuals in one form or another, and the numerical optimization approach to maximum likelihood estimation (MLE) requires the maximization of the model's likelihood function.

Numerical optimization is not a new concept. It has been used for decades in disciplines with foundations in the physical sciences, such as aeronautics, robotics and vehicle dynamics.[21] These disciplines routinely work with models that are highly complex and therefore not amenable to analytic estimation. The wide use of numerical optimization in these disciplines gives it a practical advantage not enjoyed by the other techniques reviewed in this article: it is available as an easily implemented option on a range of popular software. Largely for this reason, numerical optimization is the most widely used of these three computationally intensive techniques in political science today.

The Technique

Numerical optimization can be implemented quite easily for NLLS or MLE in a variety of computer packages for a variety of statistical models by using a simple four-step approach. The first step is to decide on the proper functional form and model specification. This is determined through substantive theory and previous research experience, with perhaps the judicious use of trial and error in some instances. The second step is to specify the initial parameter values to be used as the baseline estimate, which will then be improved upon through the numerical optimization routine. For example, when estimating a non-linear regression model, a common set of starting values is the OLS estimate of each parameter in a parallel, linearized model. Usually, a packaged numerical optimization computer procedure will routinely supply starting values if it is not instructed to do otherwise. The third step is to set the auxiliary specifications used in numerical optimization, such as the convergence criterion (which specifies the condition under which the iterations will cease), the maximum number of iterations to be undertaken, and the type of iteration algorithm to be employed.

Aside from specifying the form of the model from substantive theory, the selection of the iteration algorithm is probably the most important step in the process. An algorithm is a procedure that solves a mathematical problem – in this case, numerical optimization. We devote special attention to algorithms here because their selection can affect whether a model is successfully

(F'note continued)

1988); B. Dahlquist and A. Bjorck, *Numerical Methods* (Englewood Cliffs, NJ: Prentice-Hall, 1974); John E. Dennis Jr and Robert B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (Englewood Cliffs, NJ: Prentice-Hall, 1983); R. Fletcher, *Practical Methods of Optimization*, 2nd edn (New York: Wiley, 1987); G. E. Forsythe, M. A. Malcolm and C. B. Moler, *Computer Methods for Mathematical Computation* (Englewood Cliffs, NJ: Prentice-Hall, 1977); and Phillip E. Gill, Walter Murray and Margaret H. Wright, *Practical Optimization* (New York: Academic Press, 1981).

[21] Roland Burlisch and Dieter Kraft, eds, *Computational Optimal Control* (Boston, Mass.: Birkhauser Verlag, 1994).

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

estimated, and a discussion of them sheds light on how numerical optimization actually works. The choice of the algorithm to be used will depend on factors such as the shape of the function to be estimated, and the size and complexity of the model.

Iteration algorithms for numerical optimization can be broken down into two general categories: grid-search and gradient methods. Grid-search techniques thoroughly and systematically scan the entire range of possible values for the parameters at certain intervals, identifying the set of values which optimize the criterion.[22] The main advantage of this approach is that it is the most likely way to ensure finding a global optimum. That is, by searching all portions of the parameter space, grid searches are less likely than gradient algorithms to overlook the best overall set of estimates. In addition, grid searches are flexible since they require neither the existence nor the approximation of first or second derivatives for the estimators.[23] The disadvantage of grid-search techniques is that they are impractical for many applications because of the extreme levels of computation required for cases where there are more than one or two parameters being estimated.[24]

Unlike grid searches, gradient methods do not search over the entire parameter space. Instead, they make incremental changes in the values of a set of parameters along a given function in accordance with the rate of change of that function in the direction which promises to optimize the criterion. Because these adjustments are made using calculus, gradient methods require the existence or approximation of first-order and second-order derivative conditions for the estimated parameters.[25] The intuitive difference between grid-searches and gradient methods can be illustrated with the analogy of a traveller trying to find the 'optimum' vacation spot in a region. A grid-search would be like the traveller going systematically from place to place, assessing the quality of each spot, and then after having searched the entire region, returning to the spot with the highest rating. A gradient 'traveller' would utilize local information as to which path he or she should take, changing direction according to what seemed

[22] Richard P. Brent, *Algorithms for Minimization Without Derivatives* (Englewood Cliffs, NJ: Prentice-Hall, 1973); Fletcher, *Practical Methods of Optimization*, pp. 33–40; Gill, Murray and Wright, *Practical Optimization*, chaps. 4 and 8.

[23] Henry E. Brady, 'Review of Gill, Murray, and Wright's *Practical Optimization*', *Political Methodologist*, 3 (1990), 5–8, pp. 7–8.

[24] Brady, 'Review of Gill, Murray, and Wright's *Practical Optimization*', p. 7; William H. Greene, *Econometric Analysis* (New York: Macmillan, 1993), p. 344.

[25] The first-order condition (often referred to as the *necessary* condition in the second-derivative test) is achieved when the first-derivative (i.e., slope) of a given function satisfies the zero-slope condition, $f'(x) = 0$, that is essential in helping to find the relative minimum or maximum for a given function. The second-order (or *sufficient*) condition involves finding the second-derivative, $f''(x)$, that is sufficient to confirm the existence of a relative minimum (or maximum). In short, establishment of first-order conditions demonstrates that a function is differentiable (i.e., continuous) but does not imply that a relative minimum or maximum exists – this latter issue can only be achieved by meeting the second-order conditions.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

the most promising path at each moment, and stopping when some predetermined optimum level of quality was reached.[26]

Gradient techniques can be divided into Newton and quasi-Newton algorithms. Newtonian gradient methods (for example, the Newton method and the Steepest Ascent) are preferable if one is working with a well-behaved function for a model of moderate size and complexity, since they often arrive at a solution more quickly than quasi-Newton algorithms. This is because these methods are partially analytical, in that they make heavy use of calculus to determine the path the search should follow. These algorithms therefore require that the first and second derivatives of the parameter estimates be calculated in some manner.

Quasi-Newton gradient methods (such as Gauss–Newton, Berndt–Hall–Hausman–Hall, Gauss–Seidel, Marquardt, and Newton–Raphson) are better able to deal with models of greater size and complexity than the Newtonian methods but they take more time to converge to a solution. These algorithms require the computation only of first derivatives while the second derivatives are approximated by the algorithm itself in a sequentially updated manner. As a result, these techniques are generally better able to arrive at a solution for models that are highly complex or contain large residuals.[27] The most flexible of this type of gradient technique are the variable-metric algorithms, such as the Broyden–Fletcher–Shanno–Goldfarb and Davidson–Fletcher–Powell procedures. However, these algorithms also require more time to converge to a solution compared to other gradient methods (i.e., Newton and more restrictive quasi-Newton approaches). But as computer speed continues to increase, the extra time it takes to arrive at estimates by more flexible gradient means (for example, variable-metric algorithms) will probably become a minor concern.

The fourth step in numerical optimization involves the actual estimation of the specified model with the selected iteration algorithm and assessing the output of this procedure. If the model successfully converges within the specified number of iterations and produces estimates that seem plausible, one can use these results for both robustness checks (for example, whether a global optimum has been obtained, testing for model restrictions) and subsequent statistical inferences. If the model does not successfully converge before the specified number of iterations, one may wish to increase the number of iterations and/or relax the convergence criterion. However, loosening the convergence criterion may be dangerous because it may lead to vastly different results.[28] In some instances, the model may 'crash' (i.e., stop iterating before the maximum number of iterations specified). When this happens one should consider re-estimating the model with a different algorithm and/or changing the starting parameter values.

[26] Brady, 'Review of Gill, Murray, and Wright's *Practical Optimization*', p. 7.

[27] Brady, 'Review of Gill, Murray, and Wright's *Practical Optimization*', p. 7; Gill, Murray and Wright, *Practical Optimization*, p. 134.

[28] Gill, Murray and Wright, *Practical Optimization*, pp. 305–12.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

Numerical Optimization and Political Science

Since the numerical optimization approach to statistical estimation allows for the modelling of complex behaviour, it can contribute to political science by serving as a practical bridge between statistical analysis and political theory. Political scientists have typically modelled complex phenomena with simple methodological tools that are not suited to capture situations where functions are, for example, inherently nonlinear.[29] The use of numerical optimization allows political scientists to move beyond these simple models to more theoretically appropriate ones.

Numerical optimization has come to be used more widely in political science as both computer speed and the methodological skills of political scientists have progressed rapidly over the past decade and a half. Perhaps the most common examples of its use come in the maximum likelihood estimation of discrete choice models, such as probit and logit,[30] and the wide range of more subtle regression-based models described succinctly by Gary King.[31] For example, it usually makes more statistical and theoretical sense to model vote choice as a discrete, two-category dependent variable using a probit model, rather than as a continuous variable, as is the case when using OLS. Nonlinear models estimated with numerical optimization are also being used more frequently. For instance, Courtney Brown uses numerical optimization to test a series of dynamic nonlinear difference equations modelling US electoral volatility.[32] Henry Chappell and William Keech use numerical optimization to estimate the set of inherently nonlinear parameters in their aggregate and disaggregate models of presidential popularity.[33] And James Alt uses numerical optimization to obtain robust results in his study of the relationship between partisan political coalitions and a nation's 'world demand'.[34] In each of these instances, conventional analytic approaches to statistical estimation were deemed inferior or intractable by the author(s). This is where numerical optimization can circumvent the limitations of analytic estimation, providing the ability to estimate the parameters of models that are closer approximations to our substantive theories.

From a practical perspective, it is now quite easy for a political scientist to use numerical optimization, particularly with certain commonly applied models, such as logit and probit. Many of the common numerical optimization

[29] Christopher H. Achen, 'The Data Analysis Revolution and S-PLUS', *Political Methodologist*, 6 (1994), 2–6.

[30] For example, Kevin B. Grier, Michael C. Munger and Brian E. Roberts, 'The Determinants of Industry Political Activity, 1978–1986', *American Political Science Review*, 88 (1994), 911–26.

[31] King, *Unifying Political Methodology*.

[32] Courtney Brown, *Ballots of Tumult: A Portrait of Volatility in American Voting* (Ann Arbor: University of Michigan Press, 1991).

[33] Henry W. Chappell Jr and William R. Keech, 'A New View of Political Accountability for Economic Performance', *American Political Science Review*, 79 (1985), 10–27.

[34] James E. Alt, 'Political Parties, World Demand, and Unemployment', *American Political Science Review*, 79 (1985), 1016–40.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

TABLE 1 *Numerical Optimization Algorithms Available in Various Statistical Software Packages*

GAUSS: Berndt–Hall–Hausman–Hall, Broyden–Fletcher–Goldfarb–Shanno, Davidson–Fletcher–Powell, and Newton–Raphson
Micro-TSP: Marquardt (NLLS only)
LIMDEP: Berndt–Hall–Hausman-Hall, Davidson–Fletcher–Powell, Steepest Descent, Newton
PCGIVE: Grid Search (general procedure), Broyden–Fletcher–Goldfarb–Shanno
RATS: Gauss–Seidel, Grid Search (simplex procedure), Broyden–Fletcher–Goldfarb–Shanno
SAS: Gauss–Newton, Marquardt, and DUD
SPSS: Marquardt, Derivatives (analytical and numerical)
SYSTAT: Grid Search (simplex procedure), Quasi-Newton*
SST: Berndt–Hall–Hausman–Hall, Newton–Raphson
TSP: Gauss–Newton, Marquardt, Eicker–White, Berndt–Hall–Hausman-Hall, Davidson–Fletcher–Powell

* The quasi-Newton algorithm employed in SYSTAT is actually an alternative to Newton methods that is akin to both the Marquardt and Steepest Descent algorithms.

algorithms are contained within certain statistical procedures of most of the popular statistical software used by political scientists. Table 1 lists the algorithms that ten of these software packages can perform. In addition, a researcher may choose to obtain or write code to perform special applications, perhaps using the 'packaged' optimization routines available for GAUSS, FORTRAN, or other languages.

Potential Problems with Numerical Optimization

There are two main disadvantages associated with this general approach to statistical estimation. First, numerical optimization requires a great deal more effort on the part of the researcher than analytic forms of statistical estimation. This is partially a consequence of the fact that the researcher has more options in terms of the functional form of the model than, say, in fitting a linear model. While this is a practical 'problem', it is certainly preferable to have choice and flexibility than to be 'conveniently' restricted to a theoretically less than appropriate model. But further, a specified model may not converge for one or more of the following reasons: (1) the basic theory is fundamentally flawed; (2) the functional form has been specified incorrectly; (3) the specified model is intractable because it is too complex to arrive at a solution,[35] or (4) the data or

[35] William H. Greene, *Econometric Analysis* (New York: Macmillan, 1993).

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

the measurements being employed are poor.[36] Each of these potential sources of non-convergence must be considered and worked through, increasing the workload of the researcher. However, for the vast majority of applications a model can be estimated properly with a little bit of intuition and patience.

The second potential problem is that the results obtained from numerical optimization may seem valid when in fact they are not. This is particularly the case if the algorithm has stopped at a local optimum, rather than achieving the global optimum. That is, one set of parameter estimates may meet the convergence criterion better than any other set in the immediate 'area' in the parameter space, but there may be a different set of values at a completely different location in the parameter space that meets the criterion better. This is a problem especially for gradient methods since they do not solve for an optimum in an exhaustive fashion like grid-search techniques.[37] As a result, it is important to do robustness checks for each model when using gradient methods. One such approach is to re-estimate a given model a number of times using different starting values.

MONTE CARLO SIMULATION

Of central importance in analysing any political process is understanding the behaviour of its random component. We never hypothesize a deterministic relationship among variables because of the possibility of unmodelled and idiosyncratic influences on political behaviour. The analysis of a given sample of data will therefore yield parameter estimates that are not equal to the population parameters that generated them, and different samples will yield different estimates. The relative frequency distribution of an infinite number of these estimates from samples drawn in the same way, with the same size and from the same population, is the sampling distribution of that estimator for that process.[38] If we know this sampling distribution, we can, for example, assess the probability of an estimate from such a sample being greater than a particular size (a typical hypothesis test), or we can check if the expected value of this distribution is equal to the population parameter (to assess the bias of the estimator).

Of course, in a typical research situation we do not have the capacity to draw

[36] Greene, *Econometric Analysis*, notes that the nonlinear iteration algorithms, which are often used in numerical optimization estimation, are much more sensitive to these problems than, say, linear models estimated via OLS, because the latter can be estimated successfully (if not without bias and/or inefficiency) as long as the full rank condition is met by the regressor matrix. Therefore, it is more likely that one can successfully estimate a 'poor' linear model via an OLS model than 'poor' NLLS or MLE model via numerical optimization techniques.

[37] Brady, 'Review of Gill, Murray, and Wright's *Practical Optimization*', p. 7.

[38] Lawrence B. Mohr, 'Understanding Significance Testing', *Sage University Paper Series on Quantitative Applications in the Social Sciences*, 07–073 (Newbury Park, Calif.: Sage, 1990), pp. 18–19.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

the infinite number of samples we would need to generate such a sampling distribution. For example, how could we draw an infinite number of samples from the US electorate to assess the distribution of the mean approval rating for Bill Clinton? Therefore, we rely on statistical theory to assure us that a given statistic is unbiased or consistent, and that the sampling distribution is, for example, normally distributed with a certain variance and mean. This is the traditional parametric approach to statistical inference with which every political scientist is familiar.

The problem with parametric inference is when statistical theory fails. In order for the analytic estimation of the sampling distribution of most statistics to be tractable, a variety of more or less stringent assumptions have to be made about the population from which the data have been drawn. For example, OLS regression estimates are known to be distributed normally if the model error is distributed normally or if the sample size is 'large'. But what if neither of these conditions exists, as is sometimes the case in practical political science research? For example, Lijphart and Crepaz use a clearly non-normal dependent variable in a regression model (a typical way of inducing a non-normal error structure) to investigate the variations in corporatism across eighteen industrial countries.[39] It is not appropriate to use parametric inference here, though Lijphart and Crepaz have,[40] but it is not clear to what extent the error non-normality may affect their inferences. What difference would it make to their inferences if the true error was highly skewed or just slightly skewed? An even more troubling problem arises when we are interested in a statistic about which there is weak distributional theory. Estimators of interest to political scientists of this sort include medians and difference in medians (for voter preference and representation theories), and electoral system bias estimators developed as functions of regression coefficients.[41] The traditional parametric approach to inference offers no satisfactory way to make inferences in these cases. What is needed in both these types of statistical situation is a way to assess the properties of an estimator under a variety of conditions that does not rely on the level of statistical theory that exists for it. Monte Carlo simulation is such a technique.

Monte Carlo simulation is an experimental approach to statistical analysis that allows a researcher to evaluate the quality of an estimator and to make inferences to population parameters. This is done by generating vast amounts of data from an artificially developed population, or 'pseudo-population', which is believed to resemble a process of interest. A statistical model is then estimated

[39] Arend Lijphart and Markus M. L. Crepaz, 'Corporatism and Consensus Democracy in Eighteen Countries: Conceptual and Empirical Linkages', *British Journal of Political Science*, 21 (1991), 235–46.

[40] Lijphart and Crepaz, 'Corporatism and Consensus Democracy in Eighteen Countries: Conceptual and Empirical Linkages', p. 242.

[41] Simon Jackman, 'Measuring Electoral Bias: Australia, 1949–1993', *British Journal of Political Science*, 24 (1994), 319–57; Andrew Gelman and Gary King, 'Enhancing Democracy Through Legislative Redistricting', *American Political Science Review*, 88 (1994), 541–59.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

from these data, and the performance of this estimation is assessed, given our knowledge of the generating pseudo-population. In this way, the sampling distribution can be estimated experimentally, without any *a priori* assumptions about its functional form. Vast amounts of calculations are thereby used to bypass the assumptions required for the tractability of the parametric approach. While this sort of experimentation was used occasionally as early as the mid-1800s,[42] its extensive use was not feasible until the advent of electronic computers.

The Technique

The first step in conducting a Monte Carlo simulation is to specify the pseudo-population as completely as possible with a computer algorithm.[43] The task is to consider carefully the attributes of the population to be simulated, and then to translate these attributes into computer commands. For example, consider the simulation of the relationship between an ideology index, X, and affect towards the Conservative party in the British electorate, Y:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{2}$$

In order to set up a computer algorithm to simulate this model of the relationship, the value and structure of each of the components must be specified. The βs would be assigned constant values in the program, as they would be the pseudo-population parameters. The independent variable, X, could be set as an $(n \times 1)$ vector of fixed values (e.g., the integers $1 - n$), or as a random variable (e.g., $X \sim N[0,1]$), depending on how this variable appeared in the population, as determined from theory and previous research. The stochastic term, ε, would also be constructed based on substantive theory, and would be an $(n \times 1)$ vector of random values that might take on any of a range of distributions, for example, normal, highly skewed or uniform. Once these right-hand side terms had been defined, the dependent variable, Y, would be constructed as a function of these terms as defined by the model. Sample size is determined by the number of cases generated for the variables Y, X and ε.[44]

[42] A. Hall, 'On an Experimental Determination of π', *Messenger Mathematica*, 2 (1873), 113–14; Student, 'The Probable Error of the Mean', *Biometrika*, 6 (1908), 1.

[43] For more detail on the execution of Monte Carlo simulation, see Christopher Z. Mooney, 'Monte Carlo Simulation', *Sage University Paper Series on Quantitative Applications in the Social Sciences*, (Newbury Park, Calif.: Sage, 1996); Byron J. T. Morgan, *Elements of Simulation* (London: Chapman & Hall, 1984); Reuven Y. Rubinstein, *Simulation and the Monte Carlo Method* (New York: Wiley, 1981); Mark E. Johnson, *Multivariate Statistical Simulation* (New York: Wiley, 1987); and J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods* (London: Chapman & Hall, 1964). The first of these focuses on the practical procedures and uses of Monte Carlo in the social sciences, and the latter four are more technical and complete in their exposition.

[44] See Mooney, 'Monte Carlo Simulation', for GAUSS code to conduct a variety of such simulations.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

When this simulation is run once, it results in pseudo-sample data for X and Y with n cases, just as if it had been drawn from a population in the real world that looked like the one specified by the algorithm. In order to investigate the properties of estimators of the coefficients, say OLS estimators, the pseudo-sample values of Y would be regressed on X, and the βs would be produced.

This procedure, from drawing the X and Y pseudo-sample to the estimation of the βs, is a single Monte Carlo *trial*. The slope and intercept estimates will be different from the value set for the βs in the computer algorithm, just as the estimates we derive from real sample data are different from the true population parameters that generated those data. This is because of the random component, ε, in the computer algorithm which simulates the random component of the substantive process being modelled. These coefficient estimates will therefore have a sampling distribution based on the generating algorithm and the estimator being used, and we are interested in the properties of these distributions. Therefore, we estimate these sampling distributions by conducting *hundreds or thousands* of these Monte Carlo trials. These multiple trials can be undertaken automatically by imbedding the pseudo-population generation and statistical estimation procedures into a looping algorithm in a computer program. The β estimates from each trial are saved, and ultimately a relative frequency distribution of these estimates is constructed for each parameter. This relative frequency distribution is the Monte Carlo estimate of the sampling distribution of that estimator for the particular statistical situation that was defined by the computer algorithm. This distribution can then be assessed as to its functional form, dispersion and central tendency, as required.

Monte Carlo Simulation and Political Science

Monte Carlo simulation can be used in a variety of ways, from assessing the theoretical properties of a new estimator to generating data for classroom use.[45] We will discuss one application of the procedure that could have wide use in political science: making inferences to population parameters when traditional parametric inference fails, either due to violations of assumptions or because of the lack of strong statistical theory for an estimator.

The assumptions of parametric inference are commonly violated in political science analysis. Some of these violations (for example, the presence of heteroscedasticity and serial correlation when using OLS) have theoretically reasonable and practical solutions (such as FGLS and a range of time-series techniques). But when not addressed, either through ignorance or technical impossibility, these violations can lead to the resulting estimators being biased, inefficient, and/or inconsistent, and with unknown functional forms, any of

[45] Mooney, 'Monte Carlo Simulation', chap. 4.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

which severely threaten the ability to make accurate probability statements about substantive hypotheses.[46]

One of the most common problems political scientists have in this respect has to do with small samples. For example, in order to be assured that sample means, OLS estimators and MLEs have normal sampling distributions under general conditions, we must rely on large sample theory.[47] But if we are analysing the fifteen countries of Latin America,[48] the fifty US states,[49] or the forty-two members of a congressional committee,[50] for example, or we have a sample that is small due to resource constraints, such as a short time series,[51] we cannot usually be sure that such a statistic is distributed normally. And this situation is common – in our survey of the 1993 volumes of five top political science journals,[52] we found that fully forty-one sets of either OLS or MLE analyses were reported with data sets of fifty or fewer cases.

As an example of the difficulties that might arise here, consider trying to assess whether a population mean is equal to zero or not based on the sample mean and its standard deviation. This is easily done if the sample is large or if the underlying variable is distributed normally. In these cases, the sample mean is distributed normally, and a standard *t*-test will provide a valid inference. But what if neither of these two conditions hold? Suppose we are assessing the mean of a non-normal variable (for example, gross national product, personal income, or Lijphart and Crepaz's corporatism index),[53] and our sample is small, how can we infer to the population mean from the sample mean?

Monte Carlo simulation can be used in this situation by allowing a researcher to try out a variety of plausible population distributions for the variable, and to assess the probability of observing the value of the original sample mean (or

[46] E.g., William D. Berry, 'Understanding Regression Assumptions', *Sage University Paper Series on Quantitative Applications in the Social Sciences*, No. 07–092 (Newbury Park, Calif.: Sage, 1993).

[47] One need not rely on large sample theory to assure normality under special circumstances. For example, the sample mean is distributed normally at any sample size if the underlying variable is distributed normally. But it is usually very difficult to establish that such special conditions hold in practice.

[48] Mark P. Jones, 'Presidential Election Laws and Multipartisanism in Latin America', *Political Research Quarterly*, 47 (1994), 41–57; Karen L. Remmer, 'The Political Impact of Economic Crisis in Latin America', *American Political Science Review*, 85 (1991), 777–800.

[49] Even J. Ringquist, 'Does Regulation Matter? Evaluating the Effects of State Air Pollution Control Programs', *Journal of Politics*, 55 (1993), 1022–45; Kim Quaille Hill and Jan E. Leighley, 'Mobilizing Institutions and Class Representation in US State Electorates', *Political Research Quarterly*, 47 (1994), 137–50.

[50] Richard Hall and Frank Wayman, 'Buying Time: Moneyed Interests and the Mobilization of Bias in Congressional Committees', *American Political Science Review*, 84 (1990), 797–820.

[51] Anne N. Costain and Steven Majstorovic, 'Congress, Social Movements and Public Opinion: Multiple Origins of Women's Rights Legislation', *Political Research Quarterly*, 47 (1994), 111–35.

[52] See fn. 7.

[53] Lijphart and Crepaz, 'Corporatism and Consensus Democracy in Eighteen Countries: Conceptual and Empirical Linkages'.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

greater) under the null hypothesis for each distribution. This would be an estimate of the probability of committing a Type I error (α) when rejecting the null hypothesis. The pseudo-population would be defined to be congruent with the null hypothesis. In our example, the pseudo-population mean would be set to be zero and the sample size would be set equal to that of the original sample. The random component of the pseudo-population (that is, the variable in this case) would be generated with some distribution based on theoretical plausibility. For example, a highly skewed distribution might be used, such as a χ^2 distribution with one degree of freedom. A variable with this distribution can be generated quite simply in most computer packages by squaring a standard normal variable.[54] Or, better still, a range of distributions that seem to cover the plausible limits of the true data characteristics could be used.

Therefore, in order to test the hypothesis that the true population mean is zero, we would simulate a variable with the same sample size as in the original sample, save its mean, draw another pseudo-sample from the pseudo-population and save its mean, and so forth, until we had generated the estimated sampling distribution for the mean of a variable with the specified distribution for that sample size. We would then compare the observed mean from the original sample, and assess the probability of its being observed *had the variable been generated under the null hypothesis in such a fashion in the population*. This probability is an estimate of the α-level of this test, given this pseudo-population. If this value is small enough (for example, less than 0.05), we could reject the null, assuming that the pseudo-population resembled the true population in all other relevant respects.

The real benefit of Monte Carlo experimentation is that we can then *replicate* this experiment with a wide range of potential distributions for the variable in question, assessing the α-level for each pseudo-population. If the α-level is uniformly low, we can reject the null with confidence that the population was accurately represented by at least one of the pseudo-populations. This sort of robustness test could be very useful in situations such as the Lijphart and Crepaz corporatism study where the statistical conditions of a test could plausibly take on a range of characteristics. Monte Carlo simulation can also be used in this way to estimate the power of a hypothesis test by estimating the level of Type II error that is committed.[55]

Potential Problems with Monte Carlo Simulation

One of the drawbacks of Monte Carlo simulation is that these experiments are

[54] Mooney, 'Monte Carlo Simulation'; Rubinstein, 'Simulation and the Monte Carlo Method', pp. 93–4; Norman L. Johnson and Samuel Kotz, *Continuous Univariate Distributions*, vols. I and II (New York: Wiley, 1970), vol. II, pp. 130–45.

[55] Robert D. Duval and Leonard Groeneveld, 'Hidden Policies and Hypothesis Tests: The Implications of Type II Errors for Environmental Regulation', *American Journal of Political Science*, 31 (1987), 423–47.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

based on random processes, and as such yield only approximations of the sampling distributions of interest. This problem can be mitigated, though not eliminated, by one of two approaches. First, one may simply run ever more trials, and approach the infinity limit ever more closely. But there is roughly an inverse square-root relationship between error and the number of observations in Monte Carlo simulation, as in most statistical work,[56] so this approach has its inherent limits. The other approach to decreasing random error in Monte Carlo simulation is to use one of a set of so-called variance-reduction techniques that were developed in the 1950s, when computation was very costly.[57]

Another problem is that uncertainty about the components in the statistical model to be simulated can lead to a heavy computational burden. While some aspects of the model can be varied (notably the distribution of the random component) to understand the robustness of inferences in the face of these variations, there are severe logistical limits on the amount of this variation that can be undertaken. This is because each variation of one aspect of the pseudo-population must be matched with each variation of every other aspect in a factorial experimental design. For example, consider simulating a regression model. If the distribution of the stochastic term is varied in only five ways, and the two independent variables are each generated as random and fixed, we would need to conduct twenty full experiments (for example, 5*2*2). The less certainty there is about the structure of the original model, the more experiments will need to be undertaken, at a multiplicative rate.

A final logistical problem with the use of Monte Carlo simulation is that there currently exists little easy-to-use software for conducting the procedure. However, it appears that software engineers are beginning to develop such capabilities in popular packages, with the recent version of STATA being an example of this trend. Currently, however, more flexible and efficient packages, such as GAUSS and S+, are the preferred tools for Monte Carlo simulation.

Other problems with Monte Carlo simulation are more conceptual in nature and therefore not so easily solved by technological advances. One of the most troubling of these has to do with the final inferential decisions that are made and the possibilities of error in these. When conducting the sort of robustness test discussed above, we may be able to say that under a wide range of 'plausible' (subjectively defined, in most cases) conditions, it is highly unlikely that the null hypothesis is true. But we are not able to make any firm statements about the likelihood of making a Type I error, as we can using parametric inference when all the assumptions are met. This is because we do not know which of the pseudo-populations best represents the true population. At best, we may be able to say that α is *less than* the highest level found in the experiments. It is also important to note that the power of the test to reject the null hypothesis probably suffers from this sort of procedure. That is, we may be putting too severe a test

[56] Hammersley and Handscomb, *Monte Carlo Methods*, pp. 21–2.

[57] Rubinstein, 'Simulation and the Monte Carlo Method', pp. 121–52; Hammersley and Handscomb, *Monte Carlo Methods*, pp. 55–74.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

on the data, given an acceptable α-level. For example, we may be requiring that a regression coefficient from a true model with slightly skewed error meet the significance level as a coefficient from a pseudo-population with highly skewed error. If this is the case, we may be committing more Type II errors than necessary, given an acceptable α-level.

BOOTSTRAPPING

Bootstrapping is a non-parametric approach to statistical inference that has much in common with both traditional parametric inference and Monte Carlo simulation. Like parametric inference, it allows a researcher with a set of data to develop an α-level confidence interval around a population parameter. Like Monte Carlo simulation, it allows a researcher to relax some of the most restrictive assumptions required for parametric inference by using massive amounts of computation. The bootstrap is a recently developed approach,[58] but in combining these attractive features of parametric inference and Monte Carlo simulation, it may well prove to be the preferred inferential approach in the future.

Bootstrapping is probably most valuable to political scientists who need to make inferences with statistics that have weak statistical theory – that is, statistics which have no parametric inference alternative available. Consider factor analysis, the commonly used procedure for identifying latent variables. For example, Marcus and MacKuen[59] use factor analysis to distinguish the independence of emotional responses to presidential candidates. The specification of the number of factors to be rotated in this procedure is crucial to the satisfactory identification of all of the independent latent variables.[60] Whether or not to rotate a factor depends on whether its population eigenvalue is non-zero. Unfortunately, the functional form of an estimated eigenvalue's sampling distribution is unknown, and it is even impossible to calculate its standard error analytically. Therefore, there is no parametric way to test whether or not a population eigenvalue is zero, and therefore whether or not a factor should be rotated. The rules of thumb typically applied as an alternative to inferential testing are based on the point estimate of the eigenvalue, but these rules are largely unsupported by statistical theory and can yield misleading results.[61] The bootstrap, however, can offer a statistically sound way of making inferences with virtually any point estimator, regardless of the level of theory

[58] Bradley Efron, 'Bootstrap Methods: Another Look at the Jackknife', *Annals of Statistics*, 7 (1979), 1–26.

[59] George E. Marcus and Michael B. MacKuen, 'Anxiety, Enthusiasm, and the Vote: The Emotional Underpinnings of Learning and Involvement During Presidential Campaigns', *American Political Science Review*, 87 (1993), 672–85.

[60] R. B. Cattell, *The Scientific Use of Factor Analysis* (New York: Plenum, 1978).

[61] D. B. Clarkson, 'Estimating the Standard Errors of Rotated Factor Loadings', *Psychometrika*, 44 (1979), 297–314.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

the relative frequency distribution of which is the bootstrap estimate of the sampling distribution of that estimator.

In the multivariate case, it is not always obvious which quantity should be re-sampled. Generally, the stochastic component of the model is re-sampled, so as to simulate the random process at work in the population, just as with Monte Carlo simulation.[66] When the model involves only a single variable (for example, for a sample mean), we only need to re-sample that variable. But when the model is more complex, its random component needs to be clearly identified and re-sampled. For example, in a classic linear regression model, the appropriate quantity to re-sample is the residual, since that is the only random component of the model.[67] The procedure for doing this is only slightly more complicated than simply re-sampling the variable.[68]

The second step in making a bootstrap inference is to develop a confidence interval using this estimated sampling distribution.[69] There are several different approaches to developing bootstrap confidence intervals, reflecting the fact that bootstrapping is still in its formative stages of development. We will briefly discuss four of these approaches: the normal approximation, the percentile, the percentile-*t*, and the bias-corrected accelerated (BCa). The former two of these are interesting for the light they shed on the development and current use of the bootstrap, but the latter two techniques are the most accurate[70] and should be used in practice.

(F'note continued)
on Quantitative Applications in the Social Sciences, 07–095 (Newbury Park, Calif.: Sage, 1993); Bradley Efron and Robert J. Tibshirani, *An Introduction to the Bootstrap* (London: Chapman & Hall, 1993); and Peter Hall, *The Bootstrap and the Edgeworth Expansion* (New York: Springer-Verlag, 1992). The first two are practical introductions to the bootstrap at different levels of technical completeness, while Hall's monograph is a technical discussion of the bootstrap's theoretical properties.

[66] David A. Freedman, 'Bootstrapping Regression Models', *Annals of Statistics*, 9 (1981), 1218–28.

[67] However, if the independent variables are random, as they often are in social science data, the data needs to be re-sampled casewise (see Mooney and Duval, 'Bootstrapping: A Nonparametric Approach to Statistical Inference', pp. 15–20).

[68] Peter Hall, 'On Symmetric Bootstrap Confidence Intervals', *Journal of the Royal Statistical Society*, Series B, 50 (1988), 35–45; Freedman, 'Bootstrapping Regression Models', *Annals of Statistics*; Mooney and Duval, 'Bootstrapping: A Nonparametric Approach to Statistical Inference', pp. 15–20.

[69] Most work on bootstrap inference has been in terms of confidence intervals rather than hypothesis testing (but see Kenneth A. Bollen and Robert A. Stine, 'Bootstrapping Goodness-of-Fit Measures in Structural Equation Models', *Sociological Methods and Research*, 21 (1994), 205–29; Nicholas Fisher and Peter Hall, 'On Bootstrap Hypothesis Testing', *Australian Journal of Statistics*, 32 (1990), 177–90. Given the greater information incorporated into and simplicity in the calculation of the former, this is not likely to change in the future.

[70] 'Accurate' in this context refers to the congruity between the nominal α-level of a confidence interval and the actual α-level of the procedure (see Christopher Z. Mooney, 'Bootstrap Statistical Inference: Examples and Evaluations for Political Science', *American Journal of Political Science*, 40 (1996), 570–602.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

Normal approximation bootstrap confidence intervals are quite similar to traditional parametric confidence intervals.[71] One may sometimes be willing to assume that a statistic is distributed normally, but have no formula to calculate its standard error. For example, there is no simple formula for the standard error of the difference between two sample medians, a quantity of interest to those studying bias in congressional committees.[72] In such a case, one can use the vector of re-sampled bootstrap statistics to calculate this standard error in a straightforward manner.[73] This bootstrapped standard error can then be plugged into the traditional parametric confidence interval formula. The few instances of the use of the bootstrap in political science have been mainly of this type.[74]

The normal approximation approach has the advantage of requiring fewer re-samples than any of the other approaches (as few as 50–200 re-samples will suffice), but the need to make the fundamental assumption about the shape of the sampling distribution of the statistic is a fatal drawback. This is exactly the sort of assumption that we are trying to avoid by employing the bootstrap. If this assumption does not hold, the inferences we make will be flawed. The main conceptual problem is that this approach does not use all the information available from the bootstrap. That is, while the bootstrap estimate of the entire sampling distribution is available, the normal approximation only uses its standard deviation.

The simplest way of using the entire bootstrap estimate of the sampling distribution to develop confidence intervals is to sort the bootstrapped point estimates by magnitude, and then select the $\alpha/2$ and $1 - (\alpha/2)$ percentile points as the α-level confidence interval endpoints.[75] This *percentile* confidence interval, as it is called, will then incorporate $1 - \alpha$ of the central values of the estimate. This is a simple and highly intuitive approach to bootstrap confidence intervals, and is widely used by statisticians and practioners.[76] Unfortunately, in order for the percentile confidence interval to be asymptotically accurate, the sampling distribution of the statistic must be symmetric and the statistic must be an unbiased estimator of the parameter.[77] This is because the bootstrap

[71] Eric W. Noreen, *Computer Intensive Methods for Hypothesis Testing* (New York: Wiley, 1989), p. 69.

[72] Tim Groseclose, 'Testing Committee Composition Hypotheses for the US Congress', *Journal of Politics*, 56 (1994), 440–58; Keith Krehbiel, 'Are Congressional Committees Composed of Outliers?' *American Political Science Review*, 84 (1990), 149–63.

[73] Mooney and Duval, 'Bootstrapping: A Nonparametric Approach to Statistical Inference', p. 35.

[74] E.g., Gary King, 'Constituency Service and Incumbency Advantage', *British Journal of Political Science*, 21 (1991), 119–28; Donald P. Green and Jonathon S. Krasno, 'Rebuttal to Jacobson's "New Evidence for Old Arguments" ', *American Journal of Political Science*, 34 (1990), 363–72; Keith J. Poole and Howard Rosenthal, 'Patterns of Congressional Voting', *American Journal of Political Science*, 35 (1991), 228–78.

[75] Bradley Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans* (Philadelphia: Society for Industrial and Applied Mathematics, 1982), chap. 10; Robert A. Stine, 'An Introduction to Bootstrap Methods', *Sociological Methods and Research*, 18 (1990), 243–91.

[76] Hall, *The Bootstrap and the Edgeworth Expansion*, p. 36.

[77] Efron and Tibshirani, *An Introduction to the Bootstrap*, chap. 14.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

estimate of the sampling distribution will *reflect* the bias of the statistic,[78] and asymmetry in the sampling distribution causes the percentile points chosen to be 'backwards' in a particularly odd way.[79]

Although making the assumption that a statistic is unbiased and has a symmetric sampling distribution is not as stringent as assuming that the sampling distribution is normal, it is still more stringent than necessary. This is because Efron has developed a method of adjusting the estimated sampling distribution for both skew and bias, the *bias-corrected accelerated (BCa)* confidence interval.[80] The BCa procedure is a bit more complicated to execute than that of the percentile or normal approximation, but it yields a confidence interval that is highly accurate under a wide range of circumstances. The basic approach is to estimate the bias and skew of the sampling distribution from its bootstrap estimate, and then apply algebraic adjustments to the percentile points chosen from the sorted bootstrap estimates for the confidence interval endpoints. A straightforward explanation of the procedure is given by Efron and Tibshirani.[81]

An alternative approach which has a level of accuracy comparable to that of the BCa, but is executed in a completely different way, is the *percentile-t* confidence interval.[82] This approach involves three basic steps. First, one standardizes the re-sampled statistics into *t*-scores. This is done by subtracting the original sample value of the statistic from its re-sample value, and then dividing by the standard error of the statistic for that re-sample. These *t*-scores are then sorted, and the $\alpha/2$ and $1 - (\alpha/2)$ values selected. The selected *t*-scores are then reconverted into the original metric of the statistic by multiplying them by the full sample standard error of the statistic and adding them to the full sample estimate. The intuitive reason for sorting out the standardized estimates instead of the unstandardized estimates is that the procedure pushes to the tails those values in which there is less confidence (that is, those with the largest standard errors).

Given these four bootstrap confidence interval options, which one should be applied in a given case? First, the normal approximation and the percentile methods should be avoided as they are asymptotically inaccurate in a wide range of cases. We discuss them only because they are often used in practice, due to their computational ease and intuitive appeal. As more researchers become

[78] Mooney and Duval, 'Bootstrapping: A Nonparametric Approach to Statistical Inference', pp. 30–5.

[79] Nathan Schenker, 'Qualms about Bootstrap Confidence Intervals', *Journal of the American Statistical Association*, 80 (1985), 360–1.

[80] Bradley Efron, 'Better Bootstrap Confidence Intervals' (with discussion), *Journal of the American Statistical Association*, 82 (1987), 171–200.

[81] Efron and Tibshirani, *An Introduction to the Bootstrap*, chap. 14.

[82] Peter Hall, 'Theoretical Comparison of Bootstrap Confidence Intervals' (with discussion) *Annals of Statistics*, 16 (1988), 927–85; Hall, *The Bootstrap and the Edgeworth Expansion*; G. Jogesh Babu and Kesar Singh, 'Inference on Means Using the Bootstrap', *Journal of Business and Economic Statistics*, 9 (1983), 223–33.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

aware of the shortcomings of these approaches, their use will certainly die out in favour of either the percentile-*t* or the BCa. The choice between these latter approaches is more difficult. Both are highly accurate asymptotically in a wide variety of circumstances. Both are a bit more complex to conduct than either the normal approximation or the percentile, but they are automatic in the sense that once one has programmed them into a computer algorithm, they are conducted entirely within the program. Efron advocates the BCa for its ability to respect transformations of the statistic being estimated,[83] and Hall advocates the percentile-*t* because it involves bootstrapping a pivotal quantity.[84] There is also evidence that these approaches perform differentially in small samples. For example, Mooney[85] found that the percentile-*t* provides a more accurate confidence interval around an OLS regression coefficient with samples of 10–40, when the stochastic term is non-normal. However, more work needs to be undertaken on the small sample behaviour of these approaches before any definitive statements as to their relative performance can be made.

The Bootstrap and Political Science

The bootstrap will be most useful to political scientists in the same circumstances described above in which Monte Carlo simulation is useful, namely when the assumptions of parametric inference are violated, and when inferences are made using an estimator with weak distributional theory. The advantages of the bootstrap in these cases are that there is not the level of indeterminacy that there is in Monte Carlo simulation. That is, it is not necessary to specify a variety of simulated models, making choices as to their plausibility if conflicting inferences result. The bootstrap allows inferences to be made directly from the actual data, while loosening the parametric assumptions that may not hold in a particular case.

The bootstrap's distinctive advantage over parametric inference is clearest when there is no parametric inference alternative. As political scientists have enhanced their methodological toolkits beyond the use of linear regression models, means and proportions, this situation has arisen with increasing frequency. When one estimates eigenvalues and principal components,[86] the switch point in a switching regression,[87] indirect effects in a structural equation

[83] Efron and Tibshirani, *An Introduction to the Bootstrap*, p. 187.
[84] Hall, *The Bootstrap and the Edgeworth Expansion*.
[85] Mooney, 'Bootstrap Statistical Inference: Examples and Evaluations for Political Science'.
[86] Ichikawa and Konishi, 'Applications of the Bootstrap Methods in Factor Analysis.'
[87] Stratford M. Douglas, 'Improving the Estimation of a Switching Regression Model: An Analysis of Problems and Improvements using the Bootstrap' (unpublished doctoral dissertation, University of North Carolina, Chapel Hill, 1987).

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

model,[88] certain parameters in LISREL models,[89] or the parameters of a non-parametric regression model,[90] for examples, there is often no way to make parametric inferences to the population parameters, because the analytics of these problems have yet to be tackled. The point estimation of a parameter is always a much easier mathematical problem to solve than assessing the estimator's sampling distribution and using it to make inferences. In some of these cases, an *ad hoc* non-parametric test has been developed to make inferences.[91] But these typically are cumbersome procedures, requiring very specific statistical conditions to exist, and are only applicable for a particular estimator. By bootstrapping any of these estimators, a researcher can make reasonable, probability-based inferences in a very general fashion. And Monte Carlo simulation becomes a less and less realistic option in these complex models, because of the increasing difficulty in simulating them accurately in all relevant respects. Bootstrapping allows the estimation of the sampling distribution of virtually any statistic, merely by re-sampling and re-estimating it multiple times.

Potential Problems with the Bootstrap

The bootstrap is at an early stage in its development, and despite the vast amount of interest that it has generated in the statistical community in the past fifteen years, it is still not as easy to use nor as well understood as are numerical optimization methods or Monte Carlo simulation. All of the potential problems with the bootstrap seem to derive from this fact. It may, therefore, be the case that these problems will be rectified in the coming years with more theoretical and applied research.

The first practical problem with the bootstrap is that, as with Monte Carlo simulation, little easy-to-use software exists to execute the procedure. RATS and STATA offer a canned re-sampling algorithm, and STATA will even report bootstrap confidence intervals. The only confidence intervals currently available from STATA, however, are the percentile and normal approximation, and as noted above these are the least accurate. A few free-standing software programs for bootstrapping exist, such as SIMSTAT,[92] but these programs are invariably

[88] Kenneth A. Bollen and Robert Stine, 'Direct and Indirect Effects: Classical and Bootstrap Estimates of Variability', in Clifford C. Clogg, ed., *Sociology Methodology 1990* (Oxford: Basil Blackwell, 1990).

[89] Walter R. Mebane Jr, Jasjeet S. Sekhom and Martin T. Wells, 'The Robustness of Normal-theory LISREL Models: Tests Using a New Optimizer, the Bootstrap, and Sampling Experiments, with Applications' (unpublished paper presented at the 12th Annual Political Methodology Summer Conference, Indiana University, Bloomington, July 1995).

[90] Achen, 'The Data Analysis Revolution and S-PLUS'.

[91] See, for example, Wayne W. Douglas, *Applied Nonparametric Statistics*, 2nd edn (Boston, Mass.: PWS-Kent, 1990), p. 238, on the difference between two medians.

[92] Normand Peladeau and Yves Lacouture, 'SIMSTAT: Bootstrap Computer Simulation and Statistical Program for IBM Personal Computers', *Behavior Research Methods, Instruments, & Computers*, 25 (1993), 410–13.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

limited in the types of models that they estimate, and they do not provide all the bootstrap confidence intervals. The best current approach is to write bootstrapping macros using a flexible language or program, such as FORTRAN, BASIC, GAUSS or even SAS. There are an increasing number of sources of examples and suggestions for help in writing this code.[93] Also, there exists an entire library of S + code for bootstrapping various statistics, available free and on-line.[94] It is likely to be only a matter of time, however, before commercial software packages are given the capacity to construct bootstrap confidence intervals generally and simply.

Another practical problem is that re-sampling complex models must be done carefully, so as to mimic correctly the random process at work. For example, data to be used to estimate a regression model need to be assessed to understand whether the cases or the residuals are the truly random component, and re-sampling then needs to be carried out accordingly.[95] Also, in working with autocorrelated data, one must either completely model out the autocorrelation, or modify the re-sampling plan to preserve the stochastic structure.[96] This means that the researcher must understand his or her model and data very well (certainly not an undesirable circumstance), and programming the procedure becomes more complicated.

Perhaps the most important implication of bootstrapping being in this early stage of theoretical development has to do with its performance in finite samples. After over a decade of debate in the statistical literature about the best approach to developing bootstrap confidence intervals,[97] there appears now to be a consensus that the BCa and the percentile-*t* are the most accurate, at least asymptotically. The problem arises in finite sample cases, where their results may differ. How does one make a judgement in such a case? Currently, there is no clear answer. Further, since the small sample properties of the bootstrap

[93] Christopher Z. Mooney, 'Constructing Bootstrap Confidence Intervals Using GAUSS', *Political Methodologist*, 6 (1994), 27–9; Christopher Z. Mooney, 'GAUSS Code for Re-sampling Cases and Residuals in Bootstrapping a Regression Model', *Political Methodologist*, 6 (1994), 26–7; Mooney and Duval, 'Bootstrapping: A Nonparametric Approach to Statistical Inference', Appendix; William G. Jacoby, 'PROCIML Statements for Creating a Bootstrap Distribution of OLS Regression Coefficients (Assuming Random Regressors)', (unpublished, Department of Government and International Studies, University of South Carolina, 1992); Noreen, *Computer Intensive Methods for Hypothesis Testing*, Appendices A, B and C.

[94] Efron and Tibshirani, *An Introduction to the Bootstrap*, pp. 399–412.

[95] Mooney and Duval, 'Bootstrapping: A Nonparametric Approach to Statistical Inference', pp. 15–20.

[96] Dimitris N. Politis and Joseph P. Romano, 'A Circular Block-Resampling Procedure for Stationary Data', in Raoul LePage and Lynne Billard, eds, *Exploring the Limits of Bootstrap* (New York: John Wiley, 1992); Hans R. Kunsch, 'The Jackknife and the Bootstrap for General Stationary Observations', *Annals of Statistics*, 17 (1989), 1217–41.

[97] See David W. Hinckley, 'Bootstrap Methods', *Journal of the Royal Statistical Society*, Series B, 50 (1988), 321–37; Hall, 'Theoretical Comparison of Bootstrap Confidence Intervals'.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

are not known generally, there is the question of how large a sample needs to be before the asymptotics of the bootstrap proofs can be trusted, just as there is with the asymptotic proofs of the consistency of MLEs, for example. Monte Carlo simulation experiments seem the best approach to tackling this problem,[98] but progress on this type of research is slow due to the many estimators and statistical situations that need to be assessed.

CONCLUSION

This review of these relatively new and general approaches to estimation and inference sheds light on what the advent of cheap and fast computation can do for our understanding of political phenomena. These and related techniques give us the ability to match our statistical models more closely with our theoretical models. We are no longer confined to using a simple linear model when a non-linear model is more theoretically appropriate. So if our theory holds that the effect of the strength of party identification on the probability of voting Conservative diminishes non-linearly as it approaches neutrality, we can estimate and conduct tests on a model congruent with that theory. We no longer need to test if the mean rather than the median ideologies of two congressional committees are different, merely because there exists no parametric way to test the latter. We no longer have to make bold distributional assumptions about our data that have no evidentiary support, but whose violation could cause error in our inferences. So if we wish to model the gross national product of industrialized nations in a regression context, we will no longer need to worry in ignorance about how that variable's high level of skew will affect our inferences. These computer-intensive techniques allow us to represent our theories more accurately in our statistical models, and to be more confident in our models' performance under a variety of conditions. In short, the computer's increased power and the techniques developed to take advantage of this power will give us greater flexibility and robustness in the quantitative analysis of political data.

It should also be understood, however, that this increase in flexibility and robustness necessarily imposes at least two burdens on political scientists. First, there is the practical burden of becoming familiar with the technically feasible models and tests now available, and assessing the greater range of statistical models in trying to determine the most appropriate approach based on the theory and data at hand. This extra work arises out of the professional obligation to test theories with the best available mechanisms. For example, a study modelling

[98] E.g., J. S. Huang, 'Efficient Computation of the Performance of Bootstrap and Jackknife Estimators of the Variance of L-Statistics', *Journal of Statistical Computing and Simulation*, 38 (1991), 45–56; Norman H. Josephy, Amir D. Aczel and Hans R. Kunsch, 'Bootstrap Estimates of the Sample Bivariate Autocorrelation and Partial Autocorrelation Distributions', *Journal of Statistical Computing and Simulation*, 46 (1993), 235–49; Mooney, 'Bootstrap Statistical Inference: Examples and Evaluations for Political Science'.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press

judicial voting decisions using OLS may have been acceptable twenty years ago, but with the development of numerical optimization approaches to estimating probit and logit models, a serious study must now employ one of these more appropriate models. A second set of burdens that arises because of this increased flexibility in modelling is the need to specify more clearly the criteria used in choosing a specific model to test a theory. When the primary feasible method of testing for interdependent relationships on quasi-experimental data was with an OLS regression model and parametric inference, criteria of model choice was less of an issue because there was not much choice. But as researchers can now estimate a variety of potentially theoretically appropriate models and conduct inferential tests on them with the new techniques described in this article, we will have to consider more carefully why one model is preferable to another.

It should be clear that these new opportunities and burdens can only be good for the discipline. The extra work they impose on us will force us to think more deeply not only about the interface between statistical models and political theories, but also about our political theories independently. As we consider, for example, modelling the non-linearity of the relationship between party identification and vote, we will of necessity think more carefully about how these concepts actually interact with one another in the political world. This will lead to richer theories of politics which are more firmly based on sound empirical evidence, leading to a better understanding of the political world.

https://doi.org/10.1017/S0007123497000045 Published online by Cambridge University Press