journal of peace RESEARCH

© 1998 Journal of Peace Research
vol. 35, no. 4, 1998, pp. 419–427
Sage Publications (London, Thousand Oaks, CA and New Delhi)
[0022-3433 (199805) 35:3; 419–427; 004766]

COUNTERPOINT

Quantitative Methods in Peace Research*

RON P. SMITH

Birkbeck College, London, and University of Colorado at Boulder

Quantitative methods are an important component of peace research, since many of the issues addressed are inherently quantitative – the frequency and intensity of conflict, or the determination of military expenditures, for instance. This article argues that quantitative peace research could be improved if authors put more emphasis on the substantive issues and less on the mechanical application of rule-based, statistical techniques. After some methodological discussion, seven questions are posed that quantitative researchers might ask themselves; an attempt is made to show why these questions are important. If quantitative peace researchers asked themselves these questions more often, the substantive contribution of quantitative peace research could be increased.

Introduction

The analysis of quantitative data has played an important role in peace research ever since Richardson's use of measures of military expenditures and frequency of war. Formal statistical methods are the 'language' we use to discuss such data. This is an essential language, since people have great difficulty with probabilities and tend to make elementary mistakes unless they use formal methods – which means learning the language. Unfortunately, statistics is quite a difficult language, so learners tend to focus on mastering the rules of grammar and syntax, at the expense of the *meaning* of what is being said.[1] There is a process of goal displacement from the substantive issues of peace research to the application of the rules. The literature reflects this, being excessively technique oriented, emphasizing the mechanical application of rules.[2] It seems to me that authors would benefit not from arguing about what the right rules may be, but from trying to ask themselves the right questions.

When you start writing of course it helps to have rules – do not split infinitives, do not end sentences with prepositions – but later it is more useful to have style manuals that can suggest questions to help you judge whether your grammar and syntax work to convey the meaning effectively: is 'to boldly go' the right note to start Star Trek on? Below, I suggest some questions that authors of quantitative work might ask themselves. Though in principle asking these questions should be taken for granted, in practice they bear regular repetition. I make no claim for originality; most of these are taught in introductory statistics courses, although they are often forgotten by the time one obtains a

* I am grateful to the editor of *JPR* for persuading me to write this article, to 15 editorial committee members and other *JPR* referees for responding to an earlier version and to Kristian S. Gleditsch for guiding my further reading in quantitative peace research.

[1] This metaphor I owe to Ed Leamer, who used it about econometrics.

[2] This is a problem not just for peace research but almost any discipline that uses statistics. The problem has been extensively discussed in economics and I have drawn on that discussion. Kennedy (1992: ch. 5) is a good introduction.

PhD. My prejudice against rule-based procedures is also common in statistics, where rule-based teaching is widely disparaged as 'cook book' courses. This is unfair to real cook books, which are usually much better than such courses in emphasizing the importance of the objective of the recipe, the need to adjust the recipe to the quality of the ingredients (data) and to the implements (statistical methods) available.

Most of these points are also made in King (1989).[3] I use King as a reference partly because it is good on these issues and partly because it is a statistics book written by a political scientist rather than an economist. Judging from citations, many peace researchers use econometrics texts for reference. This is unfortunate partly because peace research data are often very different from economic data and partly because most econometrics texts adopt the wrong approach: listing tricks to fix the error terms to make the assumptions hold, rather than emphasizing the need to specify the model correctly. King (1989: 251) also criticizes the assumption–violation–correction strategy taught in most econometrics texts.

After a little discussion of methodology, the main part of my article provides some questions and a discussion of their importance. I shall follow the economics convention and distinguish methods, herein the use of particular statistical techniques, from methodology, which is the philosophical basis for those methods. I shall also use some conventions to avoid the ambiguities associated with the word 'statistics'. Statistics may be used to mean data (I will use the term 'data'); to mean functions of the observed data (I will use the terms 'estimates' or 'tests'); to mean the academic discipline concerned with methods for making inferences from data (which is how I will use the term); and finally, statistics is a branch of mathematics (which I will call mathematical statistics). The latter distinction is important, since many appear confused about what are mathematical derivations, on which we tend to agree, and what are questions about methods of inference, on which we may justifiably disagree. For instance, while we may all agree that probabilities are numbers that satisfy three axioms (an issue of mathematical statistics), we may disagree on how to estimate a particular empirical probability (an issue of statistical inference). Similarly, we may agree on the basis of mathematical statistics that a particular estimator is unbiased but disagree on whether unbiasedness is a desirable property for an estimator in a particular case.

Methodological Issues

What we are doing when we do quantitative work? At first sight this appears to be a process of induction, of seeking to infer general laws from particular observations. However, since at least Hume, it has been realized that there exists no philosophical basis for induction. Suppose that no two democracies have ever been observed to go to war: this provides no philosophical basis for concluding that we will not observe a war between democracies tomorrow. If we have a strong theoretical explanation for the observations (e.g. reasons why war between democracies is impossible), we may have more confidence that the relationship is structural or permanent, but we still cannot be sure. The structural stability, over time or space, of estimated relationships is thus a crucial statistical question.[4] In general, confidence in the structural stability of relationships is much less in the social sciences than

[3] I was made aware of King (1989) after writing the first version of this article. In my ignorance, I had not realized that a book called *Unifying Political Methodology* would in fact be about statistics.

[4] This of course forms part of the assumptions for most mathematical derivations in statistics, e.g. central limit theorems.

in the physical sciences. In some cases, the theory tells us that empirical relationships should not be stable. In economics, the Lucas critique argues that estimated behavioural parameters should reflect how people form their expectations. If something causes people to change the way they form their expectations, a change in policy regime for instance, then the estimated behavioural parameters will change, making the estimated model useless for predicting the effect of a policy change. However, the theory may suggest stability at a meta-level. The Lucas critique, for instance, implies that there should be a stable relationship between the way behavioural and expectation formation parameters move (see Alogoskoufis & Smith 1991).

Given the problems with induction, Popper proposed a more limited aim: falsification. If we cannot prove a theory true, at least we can potentially disprove it. However, this also runs into difficulties, in particular the Duhem–Quine problem: any particular prediction is a mixture of a core theoretical statement and auxiliary assumptions (for instance, about exactly how democracy and war should be measured). Thus, we cannot know whether what has been disproved is the theory or the auxiliary assumptions. Since most quantitative models involve a very large number of auxiliary assumptions (about parametric functional forms, stochastic specification, etc.), rejection is rarely robust. When we apply statistical tests, what we are almost invariably testing are characteristics of specific models, not theories.

Statisticians usually claim a much more modest methodological basis for their procedures. The usual basis is instrumental: how to use prior information and the available data to make 'better' decisions. 'Better' is evaluated in terms of the consequences of the decision to the decision-maker. To use an example to illustrate statistical procedures, you want to decide when to leave for the airport. You have prior views about how covariates (time of day, season) influence travel time and a series of observations on how long it took in the past. You make an estimate of the time required, and the outcome is that you either miss your plane or sit in the airport lounge – each of which involves costs or losses. The best estimate minimizes those expected losses. Since your loss function is probably asymmetrical, an unbiased estimate would not be sensible: missing your plane half the time is unlikely to be optimal. The consequences can be expressed either in terms of a loss-function, e.g. some cost attached to the difference between an estimate and outcome, or as in the standard Neyman–Pearson hypothesis testing framework in terms of costs attached to type I and type II errors.

The difficulty is that in most academic quantitative work neither the decisions nor the costs associated with the consequences of the decision are apparent. Thus academics tend to fall back on ritualistic conventions, such as quadratic loss functions and setting the size of the test (probability of type I error) at 5%. These conventions may have historical explanations and may provide a point of reference, but there is no reason to expect them to be appropriate for every situation. They are certainly not 'absolute scientific criteria', since they can be justified only in terms of the consequences of particular decisions.

The instrumental or decision-theoretic framework suggests that we interpret the authors of quantitative work as having some purpose, a substantive question that they wish to answer by using the available data; some prior information (usually expressed in theory) which they embody in a mathematical model of how the data were generated. They then estimate the model, evaluate how well the model represents the data; and, depending on how good this fit is, draw some

substantive conclusions from the estimates.[5] They then hope to persuade their colleagues that they got the answer right. If this account is accepted, it raises some natural questions.

Questions for Quantitative Workers

Why Are You Doing This?

Decisions about the appropriate statistical method follow from considerations of the purpose of the exercise, the substantive issue at stake, and authors should be able to convince their readers what the issue is: why they are doing the statistics. Unfortunately, the answer is often all too obvious. Either, the authors have just discovered cointegration, correlation dimensions or chaos and want to be the first to apply it to a dispute within their discipline; or the authors feel that the paper will not look professional enough without a chi-squared test or two. While such publish-or-perish incentives encourage effort, and much can be learned from applying new techniques to old questions, the authors have to establish the relevance of the technique to the question. To do this, they will have to know what the substantive question is. Furthermore, it is not always obvious that they do.

Knowing why you are doing it is also important because models are abstractions, and different levels of abstraction are appropriate for different purposes. To use a political science example, one would use different models for the same dependent variable, Presidential Approval ratings, depending whether one wished to forecast it, influence it, or test particular hypotheses about what determined it. The relative weight that would be given to statistical fit and parsimony (the number of parameters) and the choice of information sets would differ in each case; thus, the best model would differ according to purpose. If I wanted to forecast approval over the next six months, I would not include a variable for presidential speeches in the model, because that is something I cannot forecast. If, however, I were an advisor to a president, I would include it, since choosing to make speeches is an important policy variable for a president. To use another example, many types of model have been used to estimate the economic effects of military expenditure: very large structural systems, typically estimated for policy or forecasting purposes; computable general equilibrium systems, which lack dynamics but are consistent with long-term economic theory; small structural systems of a few focus variables (e.g. military expenditure, investment and growth); small atheoretical systems, e.g. VARs or transfer functions; and single-equation structural models. All have their advantages and disadvantages, but none of them can be regarded as the 'best' type of model, since they are used for quite different purposes. However, the authors still have to convince the reader that they have chosen the correct model for their purpose.

What Are the Data?

The quality of the results depends on the quality of the data, and the sceptical reader will want to know what the sources are, how the data were constructed, how the measured variables match the theoretical categories, and what the coverage (over time or groups) of the sample is. The authors should persuade the reader that they understand the data and are able to describe their statistical characteristics. Often the best way to do this is with well chosen pictures, Tufte (1983) being the classic reference. However, it requires a lot more thought to design a good picture than to run the data through a computer package, and academic incentives will

[5] Specialists may note that I am implicitly treating the authors as being ad hoc Bayesians who use classical methods, perhaps because they cannot formulate proper priors or do the necessary integration.

play a role here. My basic empirical message in Smith (1980a) could have been made much better with a graph, but it probably would not have got published without the fancy statistical techniques.

Many articles give the impression that the authors have applied the statistical procedures without looking at the data. This is particularly a problem with large datasets. Since mistakes in standard data sources (coding errors, misplaced decimal points, etc.) are fairly common, the probability is that many statistically significant results are a product of data errors. We do not have good estimates on the frequency, since so few articles are replicated. My experience from getting graduate students to replicate articles as part of their training is that data errors are common, although they usually do not change the substantive conclusions. Again, graphical methods are a good way of discovering data errors.

What Is the Prior Probability of Your Model?

This is primarily a question of theory; showing that prior information about the process that generated the data would lead to a model of that particular form.[6] To put it differently, authors need to convince the reader that their model is a plausible representation of the substantive process. A whole range of aspects of the model might need to be justified, such as: the structural stability of the parameters (why should the action–reaction coefficients in an arms race model be constant over time?), the exogeneity of the independent variables (why does the causation go from democracy to peace, not from peace to democracy?) and the particular stochastic structure (why normality?). Most statistical models were developed for quite specific purposes: regression for eugenic concerns, the theory of censored and truncated distributions to determine where to put the armour on World War II US aircraft. They can be transferred to other purposes, but then the author will need to convince the reader that the model can answer the substantive question; that the model represents what we know about the process that generates the data being analysed. Instead, authors tend to emphasize that their technique has been widely used to analyse a different type of data in another subject. The mechanical application of statistical models which do not describe the data can give nonsense results, as has been well known since at least Yule (1926).

The plausibility of the model for the process is a serious problem in peace research because the data-generation processes, how the data were constructed, are often quite complicated – measures like dyad war years. The standard procedure in this case is to construct a model of the underlying (often unobserved) process and then a model of how the observations were generated; see e.g. King (1989: 110, 116). To return to the Presidential Approval example, the series used is not a time-series but is closer to a dynamic panel, because it splices together the approval for different presidents. The computational problems this causes (you cannot use lagged dependent variables or lagged errors when presidents change) are discussed in the literature. I have not, however, seen a discussion of the fact that the basic statistical theory used to justify the properties of dynamic models employed, such as ARDL or ARIMA, does not apply. Not only is the series made up of realizations of different random variables, but it is not clear what initial observations you condition on, or how you motivate the asymptotics.[7] Given that

[6] King (1989) emphasizes this point repeatedly, e.g. pp. 41–42.

[7] The asymptotics should probably be large N fixed T, rather than large T as is usually assumed; i.e. we look at the properties of the estimators as we increase the number of presidents, keeping the incumbency period fixed, rather than letting incumbency go to infinity.

authors test for structural stability between presidents, as they often do, the fact that the statistical model does not describe the process by which the data were generated may not matter for the substantive conclusions.

Even for simple data, e.g. a continuous time-series on a continuous variable, the model needs to be justified. VARs (vector autoregressions or Very Awful Regressions, as they are sometimes known) have been widely applied in peace research; either in their own right or as vehicles to test for cointegration, if the series are I(1), or Granger causality. Maddala (1997) provides a good discussion related to work on political methodology of the problems with VARs, Granger causality, unit root tests for the order of integration and cointegration. There are some cases where a VAR is the natural statistical model, e.g. it is the reduced form of the standard bivariate Richardson arms-race model. However, even in these cases, use of a VAR tends to be justified in technical terms, or through its use in other subjects, rather than the substantive interpretation.

Confidence in the model will be increased if its parameters can be interpreted in terms of the substantive issue and the process that generated the data.[8] What parameters mean in substantive terms has always been an issue. Throughout the 19th century, there was a long debate over the validity of the arithmetic mean in terms of what it meant to say that the average family had, say, 2.5 children. In modern terminology, this would be a question about what is the parameter of interest – a point on which statistics has nothing to say. Consider the standard regression model, $y = Xb + u$, which can be written $y = Zd + u$, where $Z = XP$, $d = P^{-1}X$ and where P is any non-singular conformable matrix. The Z are linear combinations of the X. If one researcher says the effect of X_1 is b_1 and this proves my theory whereas another researcher says the effect of Z_1 is d_1 and this proves my theory, then statisticians cannot discriminate, because the theories are observationally equivalent but have different parameters of interest. An identical issue arises in psychometrics, where the dispute is over the appropriate loadings in factor analysis and in cointegration, when there is more than one cointegrating vector. One should ask whether there are observationally equivalent (reparameterized) models with quite different theoretical interpretations. King (1989; section 7.3) discusses this problem with respect to different time-series representations.

Does the Model Fit the Data?

There are two aspects to this question, one relating to the assumed model and the other to possible alternative models. The first asks whether the auxiliary assumptions embodied in the model hold in the data. For instance, in regression, these assumptions include linearity, constant parameters, exogeneity, constant variance and independence. Various diagnostic tests are available to check whether the estimated model replicates these assumed properties of the data-generating process. The usual measures of fit, like R^2 and chi-squared, are completely uninformative in this respect. These tests may provide useful information, but then interpreting this information is fraught with difficulty. Thus the Hendry rule 'Test, Test and Test again' is no more absolute than the other rules.[9]

The second aspect asks how well the model performs compared to alternative models of the process. The usual classical estimates and tests are conditional on the

[8] King (1989: 102) says 'if β has no substantive interpretation for a particular model, then that model should not be estimated'.

[9] In Smith (1991), I discuss some problems with structural stability testing, using as an example an equation determining France's military expenditure.

correct specification of the model. The data might have been generated by a quite different process and you could still have obtained these very significant results. Sensitivity analysis, examining how well alternative models do, is thus vital. One way to do this is to start with a very general model which encompasses a wide range of possible alternatives and then test down to a more specific model. This approach is useful and should be more widely applied in peace research, but it also runs into difficulties in many cases. The major problems here are overparametrization of the general model (you may not be able to estimate it) and the danger that statistically significant effects may be peculiarities of the particular sample. A model developed from one sample needs to be tested on another independent sample. I have given an example of this in Smith (1989), where a model of military expenditure developed from UK data was tested on French data.

Can You Tell the Difference Between Statistical and Substantive Significance?

Academic quantitative work is often characterized by a psychological process of goal displacement, from the importance of the result to the statistical significance of the result. However, important results may be statistically insignificant, and unimportant ones statistically significant. Any difference, however small, will be statistically significant if the sample size is large enough. Some would even argue that t ratios or p values are nothing more than noisy measures of the sample size. McCloskey & Ziliak (1996) is a good introduction to these issues. Of course, being able to judge the importance of the estimates requires being able to interpret the parameters in terms of the substantive issues (see the section on 'why are you doing this?').

Can Your Results be Reproduced?

In principle, replication is essential to the scientific process; in practice, it is rather unimportant. Most published scientific results are not worth the effort of replicating them; only really important results (like cold fusion) justify the effort of replication, and these are rather few. However, if you think your results are important, you should make sure they can be replicated. Evidence on the lack of replicability of econometric estimates is given in Dewald et al. (1986); a large number of articles on verification/replication (many of them against it) can be found in *PS* (1995). In that symposium, Nagler (1995) gives excellent advice on procedures which will make it more likely that you will be able to reproduce your own results.

A range of journals, including the *JPR*, have made various efforts to make replication easier, for instance by making it a publication requirement that the data be available on a website where they can be downloaded by others. But this has not changed the incentives for replication and, thus, the incentives to avoid mistakes. When I use articles for teaching, the students who pore through them find many mistakes. Most of them are minor, as were the majority of those discovered by Dewald et al. (1986). In such cases, experienced academics can say the expression is indeed wrong, but it doesn't matter; however, the mistakes live on in the literature. Sometimes mistakes can be major. In my own work, Smith (1980b), on the demand for military expenditure, contains a major mistake, one which changed the interpretation of the results and which could be easily discovered from the information in the paper. Although the paper was widely cited nobody checked the algebra, until one of my students did. Smith (1987) corrects that particular problem.

How Much Would You Bet on the Predictions of This Model?

This is an economist's question, because the purpose of many (though not all) economics

models is to make money: if a model for predicting the exchange rate will not make money, it is no use. However, the approach is more generally applicable: substantive issues in international relations, political science and peace research can be phrased in terms of bets. What are (or were?) the odds on: the Soviet Union collapsing; two democracies going to war; the India–Pakistan confrontation escalating to a nuclear interchange; cuts in military expenditure causing a world depression; terrorists using NBC weapons in the USA? Prior information, empirical evidence and a statistical framework provide some basis for making a judgement about the odds in substantive issues. These odds, which should reflect both the statistical confidence intervals and the degree of model uncertainty, are a way to estimate an author's subjective judgement about the importance of the results.

Requiring that models be used for prediction and then evaluating the models on their predictions, a form of replication, is becoming more common in peace research. This is a trend that should be encouraged. The predictions do not have to be true ex ante predictions: they can be conditional predictions, which can be evaluated ex post when the values of the independent variables are known. The large technical literature on forecast evaluation can provide a valuable learning experience, particularly, as economists know well, in teaching you a little humility: the role of economic forecasts is to make weather forecasters look good.

Conclusion

I believe that, by asking themselves the questions indicated here, quantitative peace researchers can make their work more meaningful in terms of the substantive questions of peace research. I also believe that quantitative work is valuable because many, though not all, questions in peace research are inherently quantitative. However, to return to my original metaphor, I do not believe that answering my questions will make the work any more comprehensible to those who do not speak the language. Answering my questions may even make it less comprehensible as simple inappropriate models are replaced by models which capture the complexity of peace research data. There will therefore still be a need for books and articles which can convey the substantive conclusions to a wider audience in a non-technical way.

It is left to the reader to judge whether peace researchers do ask themselves these questions enough, and whether my list contains the right questions. I am not claiming that all quantitative work in peace research suffers from these faults – much of it is very good. I have not cited cases of what I believe to be good, or bad, work because I felt it would be unfair to single out particular examples. I believe that these faults are common because there are clear academic incentives to emphasize technique at the expense of substance, to adopt ritualistic formal rules (such as statistical significance) because they convey an appearance of objectivity, and to pay little attention to replicability. But these are questions of academic sociology on which I, as an economist, possess no comparative advantage.

References

Alogoskoufis, George & Ron P. Smith, 1991. 'The Phillips Curve, the Persistence of Inflation and the Lucas Critique', *American Economic Review* 81(5): 1254–1275.

Dewald, William G.; Jerry G. Thursby & Richard G. Anderson, 1986. 'Replication in Empirical Economics', *American Economic Review* 76(4): 587–603.

Kennedy, Peter, 1992. *A Guide to Econometrics*, 3rd edn. Cambridge, MA: MIT Press.

King, Gary, 1989. *Unifying Political Methodol-*

ogy: The Likelihood Theory of Statistical Inference. New York: Cambridge University Press.

Maddala, G. S., 1997. 'Recent Developments in Dynamic Econometric Modeling: A Personal Viewpoint', paper presented at the Annual Meeting of the Political Methodology Group, Ohio State University, Columbus, OH.

McCloskey, Deirdre N. & Stephen T. Ziliak, 1996. 'The Standard Error of Regressions', *Journal of Economic Literature* 34(1): 97–114.

Nagler, Jonathan, 1995. 'Coding Style and Good Computing Practices', *PS: Political Science & Politics* 28(3): 488–492.

PS, 1995. 'Verification/Replication', *PS: Political Science & Politics* 28(3): 443–499.

Smith, Ron P., 1980a. 'Military Expenditure and Investment', *Journal of Comparative Economics* 4(1): 19–32.

Smith, Ron P., 1980b. 'The Demand for Military Expenditure', *Economic Journal* 90(December): 811–820.

Smith, Ron P., 1987. 'The Demand for Military Expenditure; A Correction', *Economic Journal* 97(December): 989–990.

Smith, Ron P., 1991. 'Spurious Structural Stability', *The Manchester School* 59(4): 419–423.

Tufte, Edward R., 1983. *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press.

Yule, G. Udny, 1926. 'Why Do We Sometimes Get Nonsense Correlations Between Time-series? A Study in Sampling and the Nature of Time-series', *Journal of the Royal Statistical Society* 60: 812–854.

RON SMITH, b. 1946, PhD in Economics (Cambridge, 1974); Professor of Applied Economics at Birkbeck College, University of London; Visiting Professor, Department of Economics, University of Colorado at Boulder (1997–98).