Whose Variance is it Anyway?

Interpreting Empirical Models with State-Level Data

Jeff Gill
University of Florida
P.O. Box 117325, 234 Anderson Hall
Gainesville, Florida 32611-7325

telephone: 352-392-0262x272, fax: 352-392-8127
e-mail: jgill@ufl.edu

State Politics and Policy Quarterly Volume 1, No.3, p.318-39.

AUTHOR INFORMATION:

Jeff Gill is an assistant professor of political science at the University of Florida. His primary research applies Bayesian statistical modeling to questions in public policy, bureaucracy, and Congress. His work has appeared in Public Administration Review, Political Research Quarterly, Journal of Public Administration Research and Theory, as well as with Georgetown University Press and Brookings Institution Press. His books include Generalized Linear Models (Sage), and What Works: A New Approach to Program and Policy Analysis (Westview, with Ken Meier).

Abstract

Researchers commonly apply inferential statistical procedures to population data from the fifty U.S. States as if estimating population parameters from sample statistics. This is incorrect because with population data there is no need to make inferences about quantities that are already known. Instead authors should simply provide evidence that their specified model provides a good fit to the data. Summary measures of variance as well as the full engine of Bayesian statistics perform this function. This research note demonstrates why the current practice of making inferences from population data with the null hypothesis significance test is wrong, provides some specific examples of problems in the literature, and gives prescriptive advice about correctly assessing and conveying empirical model results.

Introduction

A frequent and unanswered query in empirical research on state politics and policy concerns the interpretation of statistical models using population level data. This research note addresses the question: when we use population level data and construct a model, what is the appropriate interpretation of the results?

A common empirical methodology in the study of the U.S. states is to obtain population data aggregated to the state-level from government sources and then construct a parametric model (Dye 1984, Gray 1976, Hero and Tolbert 1996, Rinquist 1993). Researchers typically apply null hypothesis significance tests to such a model to assess the reliability of their findings. However, this use of significance testing is incorrect. Possession of population data precludes the need for inferring population values, which is the only motivation for performing a significance test.

Researchers should instead be concerned with how well their model "fits" the population data, meaning how much of the variance is explained relative to how much is left unexplained. With population data, the fit of the model to the data is correctly judged by the uncertainty for each coefficient relative to its magnitude. However, this vague standard lacks the invalid, but appealing, arbitrary threshold approach of a significance test.

The best means of measuring model fit for empirical analysis in state-level research with population data is to employ general measures of model fit as well as Bayesian statistical methods which are free from the flawed hypothesis testing methodology that dominates political science. The Bayesian paradigm considers parameters not as fixed points as but realizations from some estimable distribution. These distributions are recharacterized when new data are observed, thus making Bayesian analysis particularly well suited to describing the underlying iterative state-level population data generation process. Conversely, null hypothesis significance testing is a deeply flawed and widely misunderstood procedure that impairs the accumulation of knowledge (Cohen 1994, Gill 1999, Hunter 1997, Rozeboom 1960, Serlin and Lapsley 1993, Schmidt 1996).

1 The Main Point

In the classical and typical scenario where sample data are used to make inferences about population parameters, all models that can be constructed through differing specifications

are technically incorrect. However, some of these models are quite useful in understanding political phenomena of interest. Conversely, when population data are used, all models are by definition correct. Yet some models are obviously useless. This contrast exists because with population data there is no requirement to infer values that are already known, and differing population models are just alternative views of the truth. What we really care about is whether or not the researcher generated parametric specification is a good fit to the population data, as measured by unexplained variance.

Suppose we construct a simple bivariate regression model in which state-level median income is used to predict state high school graduation rate using data from 50 states for a given year. These data are produced by government agencies and are considered highly reliable descriptions of the state populations. For simplicity assume that a linear model is appropriate. This model produces a slope β, with an associated standard deviation σ. What does this mean? The parameter β is the population change in high school graduation rate for a one unit difference in median income, holding all other modeled factors constant: a fact not a supposition or inference. For any value of β, including zero, this is a meaningful and true description of reality. Now suppose we have an alternative model in which state-level median shoe size is used to predict state high school graduation rate. The slope parameter from this model is also a true description of the population, but is unlikely to be far from zero and is certainly not useful. Even if an effect is known to be true in the population, this is not a sufficient condition for the effect to have substantive utility as a description of social behavior.

A population cross-sectional model is actually a single-period observation from a stochastic process. It is still called a model because the researcher makes specification decisions with regard to explanatory variables and cases, and therefore introduces the possibility of both omitted variable bias and selection bias. In addition, population models can specify an ill-fitting parametric form, fail to consider heteroscedasticity and autocorrelation, introduced endogeneity, and ignore systematic measurement error. Unaccounted for, these unwelcome effects will fall to the error term in the specified model. Therefore population cross-sectional results are single period observations from a stochastic series of events in which the specifi-

cation determines the distribution of the error term (Rivers and Rose 1985, 185). In other words, even with population data, the researcher determines how the theoretical model is fit to the data, necessarily leaving unexplained variance of some source and magnitude.

Because these data are a cross-section from some ongoing stochastic process at the population level, statistical descriptions of a single period have no generalizability across an enclosing era. While most readers are willing to accept a description of 1973 data as resembling 1972 or 1974, there is no guarantee from a cross-sectional study that dramatic events (presidential resignations, OPEC oil embargoes, price freezes, etc.) will not occur. Thus while there may be a low probability of macro-level shocks in population data across fifty states and nearly 300 million citizens, a cross-sectional statistical model does not present any actual evidence of the continuance and stability of observed phenomena and trends.

Some Problems with Existing Approaches

Much empirical research on state politics and policy uses highly reliable, often government-furnished, data that describe aggregate state-level economic, social, or political population characteristics at a given point in time. Despite the commonality of the data source, scholars have employed a number different methodological approaches to questions of inference and model fit. These vary greatly in quality from essentially correct summary descriptions of population parameters to fatuous applications of inferential statistics. One notable example of the latter is the idea that standard statistical inference procedure from samples to populations can be applied to population data with the goal of estimating values from "an imaginary infinite universe of states" (Dye 1965, 593; Blaylock 1960, 302). The unearthly notion that there exists some platonic set of "potential states", from which the observed population states are drawn by some sampling scheme, runs counter to any established notion of inference. If there is an infinite number of potential states of the states from which one occurs by chance, then what is it that is desired by an estimation procedure? What value (or distribution of values for Bayesians) supplants others from this infinite set of population alternatives? Clearly the only candidate value is the observed population value, which we already have thus alleviating the need for inferring some celestial super-population parameter.

Studies Omitting Standard Errors

A seemingly obvious but pervasive mistake in reporting population cross-sectional results is the complete omission of measures of variability. The most common example is reporting correlation coefficients without the associated standard deviation (Besley and Case 1995, Bonjean and Lineberry 1970, Dawson and Robinson 1963, Dye 1961, Edwards 1990, Fiorina 1991, Gary 1973, Gray 1974, Heer 1966, Hopkins and Weber 1976). Even though this is a population correlation coefficient, the size of the studied population makes an enormous difference in judging the reliability of the observed relationship. This is because there still exist random events that deviate from the true causal phenomenon of interest and these can easily skew small-n studies: correlation coefficients are built on means making them non-resistant to outliers. In addition, without reported variances the reader has no way to assess the quality of the fit with regard to the effects of possible heteroscedasticity, autocorrelation, specification biases, measurement error, and endogeneity.

A similar mistake is the use of large proportional differences as inferential evidence of an effect (Hofferbert 1966; Jones and Miller 1984; Wiggins et al. 1992). For example, Nice (1994, 40) divides the states by the presence or absence of a public school teacher competence test and then crosstabulates by an innovation score. He suggests that large proportional differences between the two groups of states are evidence of an effect. The problem with this approach is that there is no distributional information provided and therefore no non-arbitrary thresholds.

Purely Cross-sectional Studies

Even when measures of uncertainty are reported in state studies with population data, misinterpretations exist. One mistaken view is that cross-sectional state-level data analysis is generalizable across neighboring time because it is population data (Carmines 1974, LeLoup 1978, Jennings 1979, Winters 1976). This implies that descriptive statistics have meaning outside the original period of study. The central problem here is that a linear regression coefficient that is assessed as being sufficiently far from zero, given its standard deviation, has some noticeable and reliable effect size for that period only (typically years in studies of the U.S. states). It does not mean that this effect endures as an important determinant of

outcome variable behavior in the way implied by significant coefficient estimates in a time series model. However, it is neither incorrect nor unnatural to theorize about the persistence of the observed effects in subsequent time periods. In the absence of major shocks to the modeled system, state-level policy variables are characterized by relatively slow change. There is, of course, no assurance of this type of stability provided by a statistical model.

Another flawed argument sometimes made in presenting cross-sectional population results is the assertion that correlation coefficients and their measures of uncertainty are model-neutral descriptions of the data in the same way that a mode or a range might be considered to be purely descriptive (Edwards 1996; Hedlund and Friesema 1972; Markus 1974; Marquette and Hinkley 1981). This perspective is wrong because Pearson's product moment correlation coefficient is in fact a linear model of the bivariate relationship between the two variables. The correlation coefficient, r, is deterministically related to the regression coefficient estimate, β, by: $\beta = r\frac{S_Y}{S_X}$ where S_Y is the outcome variable standard deviation and S_X is the explanatory variable standard deviation. Therefore, these are actually identical statistical models. Since authors specifying a linear regression model are typically required to substantiate the existence of a linear functional form and those providing simple correlation studies are not, then over time the literature accumulates incorrect correlational studies.

Studies Across Time

Another view of state-level studies is that we are not really interested in just a narrow cross-section of time, but instead we want to make claims about broader periods (Gray 1976; Sharkansky 1968). There are two methods for doing this: pooling data ignoring possible changes over time, and specifically modeling a time series. When pooling data it is important to determine that there have not been important changes to variables during the time period under investigation. Subject to this precaution, pooled data studies provide results that are informative over a greater period of interest than just a single point in time. Time series approaches (there are many) specifically parameterize the effects of time in order to capture cross-time seriality in variables and error terms. The result is a posterior distribution of the model coefficients of interest taking into account the order in which the data occurred.

It is not difficult to find published time series work using state-level population data that shows how much information can be missed by relying on a purely cross-sectional design. For instance, Becker et al. (1994) demonstrate how strictly cross-sectional approaches would miss an important aspect from data on taxation of cigarettes. They evaluate state-level economic factors for all fifty states and the District of Columbia to test the relative strength of short run elasticity versus long run elasticity for cigarette demand. Their model provides evidence for the theory that highly addictive products, like cigarettes, will cause short run effects to dominate long run effects since the immediate aggregate elasticity for addicts will be dramatically lower. The policy implication is that raising taxes on addictive products is likely to increase tax revenue in the short term but not necessarily in later years. This study is notable here because it very clearly illustrates the importance of time-seriality in determining the fiscal effect of a policy decision. Becker et al. also perform a cross-sectional analysis where the effect of a price increase due to new taxation is measured the year before and the year after taking effect. They find that a 10% increase in cigarette prices decreases consumption only 3% for one period, whereas the long run reduction in consumption is 7.5%. For a policy planner attempting to measure the revenue impact to the state of increasing cigarette taxes, this is an important distinction.

Null Hypothesis Significance Testing with Time Series Specifications

It is not well known that time series models are more sensitive than cross-sectional models to the problems with null hypothesis significance tests as practiced in the social sciences. It has been shown (inarguably) that this hypothesis testing paradigm is deeply flawed and generally misunderstood (Bakan 1960; Cohen 1994; Hunter 1997; Rozeboom 1960; Serlin and Lapsley 1993). Many of these problems are interpretational, but the procedure also does not provide scientifically valid conclusions due to internal logical inconsistencies (Gill 1999; Meehl 1978; Schmidt 1996).

This dominant approach to hypothesis testing in political science is a synthesis of the Fisher test of significance (1925a, 1925b, 1934, 1955) and the Neyman-Pearson hypothesis test (1928a, 1928b, 1936). In the null hypothesis significance test, two hypotheses are posited: a null or restricted hypothesis (H_0) and an alternative or research hypothesis (H_1) describing

competing and exclusive notions about some substantive question. The research hypothesis is the probability model describing the author's theory about the data generation process, and is operationalized through a parameter: η.

In the most basic case, a null hypothesis asserts that $\eta = 0$, and a complementary research hypothesis asserts that $\eta \neq 0$. The test statistic (T), some function of η and the data, is produced and compared with its known distribution under the assumption that H_0 is true. Typical test statistics are sample means, chi-square statistics, and t-statistics in linear regression analysis. The test procedure assigns one of two decisions, D_0 or D_1, to all possible values in the sample space of T, and this sample space is composed of two complementary regions, S_0 and S_1, which correspond to supporting either H_0 or H_1 respectively. The p-value is equal to the area in the tail (or tails), away from the expected value, of the assumed distribution under H_0 that starts at the point designated by the placement of T on the horizontal axis and continues to infinity. If a predetermined α level has been specified, then H_0 is rejected for p-values less than α, otherwise the p-value itself is reported.

In the Neyman-Pearson test, the probability (α) that T falls in S_1 (causing decision D_1) is a predetermined null hypothesis cumulative distribution function level: the probability of getting an observed test statistic greater than or equal to the associated critical value given a specified distributional form (e.g. normal, χ^2, F, t) for the test statistic under the null hypothesis over many hypothetical iterations of the test. In Fisher's construct, the evidence against the null hypothesis is the cumulative distribution function level corresponding to the value of the test statistic under H_0: $p = \int_{S_1} P_{H_0}(T = t)dt$, and an implied support for D_1 is given if the test statistic is sufficiently atypical given the distribution under H_0. These are very different ideas. With Fisher hypothesis testing, no explicit complementary hypothesis to H_0 is identified, and the p-value that results from the model is evaluated as the strength of the evidence for the research hypothesis as a function of the data. Neyman-Pearson tests identify two complementary hypotheses: Γ_1 and Γ_2 in which rejection of one implies acceptance of the other. This rejection is based on an α level determined even before looking at the data.

The many problems with the null hypothesis significance test as practiced in the social

sciences[1] are exacerbated in time series inference because the standard test simultaneously assumes a null distribution for the parameter and a correct specification for the error term. Many authors have addressed tests for specification error in time series models of different types since failure to distinguish between residual autocorrelation and model misspecification will almost certainly lead to invalid inferences (Akaike 1969; Godfrey 1987; Knottnerus 1985; MacKinnon 1992; Shibata 1980; Thursby 1981). Specifically, p-values are calculated assuming that the null hypothesis of no effect is true and that the model has taken whatever autocorrelation structure exists in the errors over time and made it systematic by including it in the parameterization. To give a concrete example, this section derives the dependency of the null hypothesis significance test on structural assumptions for a linear regression model with autoregressive–1 disturbances. The AR(1) specification is easily the most popular time series specification in political science despite some obvious limitations (King 1989, 185).

For an outcome variable vector, $\mathbf{Y}_t$ measured at time t, the simplest AR(1) model for T periods is specified as:

$$\begin{aligned} \mathbf{Y_t} &= \mathbf{X}_t\boldsymbol{\beta} + \epsilon_t \\ \epsilon_t &= \rho\epsilon_{t-1} + u_t, \qquad |\rho| < 1 \\ u_t &\sim \text{i.i.d. } N(0, \sigma_u) \end{aligned} \tag{1}$$

where $\mathbf{X}_t$ is a matrix of explanatory variables at time t, ϵ_t and ϵ_{t-1} are residuals from period t and $t-1$ respectively, u_t is an additional zero-mean error term for the autoregressive relationship, and $\boldsymbol{\beta}$, ρ, σ_u are unknown parameters to be estimated by the model. Backward substitution through time gives $\epsilon_t = \sum_{j=1}^{T-1} \rho^j u_{t-j}$, and since $E[\epsilon_t] = \sum_{j=1}^{T-1} \rho^j E[u_{t-j}] = 0$, then $\text{var}[\epsilon_t] = E[\epsilon_t^2] = \frac{\sigma_u^2}{1-\rho^2}$, and the covariance between any two errors is: $\text{cov}[\epsilon_t, \epsilon_{t-j}] = \frac{\rho^j \sigma_u^2}{1-\rho^2}$. Assuming asymptotic normality, this setup leads to a general linear model with the following

tridiagonal $T \times T$ weight matrix:

$$\Omega = \begin{bmatrix} 1 & -\rho & & & & \\ -\rho & (1+\rho^2) & -\rho & & 0 & \\ & & \ddots & & & \\ & 0 & & -\rho & (1+\rho^2) & -\rho \\ & & & & -\rho & 1 \end{bmatrix}. \tag{2}$$

The estimated coefficient vector of interest is simply $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}$, with associated variance: $\text{var}[\hat{\boldsymbol{\beta}}] = \frac{\sigma_u^2}{1-\rho^2}(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}$. This shows that the estimate of the primary quantity of interest and its measure of uncertainty is dependent on two other model quantities, ρ, and σ_u^2, which can be readily obtained by marginal maximum likelihood (Box and Jenkins 1976, 53-66; Davidson and MacKinnon 1993, 331-341; Fomby et al. 1980, 208).

The dependence of quality hypothesis tests for the $\boldsymbol{\beta}$ parameter vector on the accuracy of the estimate for ρ ($\hat{\rho}$) is startlingly obvious from the foregoing example. While $\hat{\rho}$ is easily calculated from the model and the data, it is extremely sensitive to the appropriateness of the AR(1) specification. If the autocorrelation is not produced by effects between adjacent time periods, or if it is manifested in a specification different from (1), then $\hat{\rho}$ is an undersized estimate of the autocorrelation and the structure of the Ω matrix leads to an over-confident test of reliability. Furthermore, this problem exists even in the presence of highly reliable, population-level data for the 50 states.

Recommendations

This section provides two means of avoiding the problems outlined in the previous section. First, population models can be described with general goodness of fit descriptions that avoid the failed and flawed null hypothesis significance test, as well as the idea of using inferential procedures in general on population data. Second, a well developed and theoretically appropriate paradigm for analyzing population models is provided by Bayesian statistics. There is overlap between these two recommendations as Bayesian statistical output is also descriptive of population quantities (therefore a means of describing model fit), rather than being forced into a rubric appropriate for sample to population generalization.

Assessing the Fit of Population Cross-Sectional Models

In population models variance exists around the calculated coefficients, but this variance is not an indication of estimator reliability. Rather, it measures the variability of the observed effect size (β), subject to model misspecification, autocorrelation, heteroscedasticity, and measurement error, as a description of some substantive phenomenon of interest across the included cases.

Consider the two alternative bivariate linear models using population data illustrated in Figure 1. Each model has the same slope coefficient (for convenience), but they have noticeably different variance around the fit. Since population data are used here, the calculated parameters of both models are accurate descriptions of the phenomenon of interest. The fact that the first model fits the linear trend closer means that the researcher imposed specification is a better fit to the data. In other words, the error term is smaller in the first model because the factors not included in the specification are less important (influential) in the first model than in the second model.

INSERT FIGURE 1 HERE

The idea that both models in Figure 1 are correct but one is better than the other is at odds with the perspective of inferential analysis typically held by political scientists. We are trained to think of statistical models as simplified, sample-based depictions of reality that are neither correct nor unique (Blaylock 1961; Leamer 1978; Russell 1929). In inferential statistics, two models that each provide reliable results (small p-values, 95% confidence intervals bounded away from zero, and so forth) can give different substantive conclusions (Raftery 1995). We select one over the other during the process of model specification knowing that we are choosing from among incorrect, but hopefully instructive, alternatives (Box 1980; Miller 1990).

Conversely, in the case of population models, we have a specification provided by the researcher where any observed coefficient accurately describes at least some features of the population. Since all ignored effects default to the error term, along with the truly random behavior that we want in the error term, some population specifications are better than others at describing the data. Consider a simple additive linear model for the state-level

change in welfare cases (Y) using population cross-sectional data:

$$Y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + \epsilon_i \tag{3}$$

where X_{1i} denotes the i^{th} state's proportion of urban dwellers, and X_{2i} denotes the i^{th} state's proportional contribution to U.S. Gross Domestic Product (GDP). Now instead of this basic setup we perform a transformation on the GDP variable as in the following new specification:

$$Y_i = X_{1i}\beta_1 + \arctan((X_{2i})^{-\frac{1}{2}})\beta_2 + \epsilon_i. \tag{4}$$

While this specification is unlikely to produce a useful model parameter for β_2, it does give the arc-tangent, inverse, square-root population effect of state-level GDP on the change in welfare recipients where any linear GDP effect will fall to the error term (this change can also affect the β_1 term, depending on the covariance). How will we know that this is not a useful contribution? The variance of the β_2 parameter across the fifty states will be very high relative to the parameter size[2]. Since this is population data, it is not a wrong specification; it is simply a useless (and atheoretical) picture of the data.

What constitutes a "useful" population model? A useful model is one that provides support for a specific theory in which alternative model coefficients provide a poorer fit to the data as judged by the ratio of systematic effects to stochastic effects. As an illustration, consider again the two linear models in Figure 1. If we modify slightly the slope of the line in the first panel then the ratio of explained variance to unexplained variance will degrade rapidly. If we do this to the line in the second panel, then this ratio will not change very much until the line is dramatically altered. Therefore the first model constitutes a better fit than the second. This is likely because of non-modeled effects in the second model such as important but omitted explanatory effects, heteroscedasticity, and autocorrelation. Note that this determination of a useful population model is more general than just the variance ratio in a linear model; it includes the ratio of systematic effects captured to stochastic effects left-over in non-linear and non-parametric models as well.

It is also somewhat surprising that tests of significance for regression coefficients are

immaterial in population models. This is because the null hypothesis significance test, as practiced in the social sciences, is a device for making conditional probability statements about point estimates of population parameters from sample quantities. Possession of population data automatically answers the very question asked by this procedure (typically whether or not a sample provides evidence that an estimated coefficient is distinct from zero in the population). Performing a significance test on population data is actually an assessment of the fit of the parametric model rather than of point estimate reliability. However, it is a convoluted assessment because the associated probabilistic reasoning is inappropriate (Berger 1985; Cohen 1994).

Since it is illogical and inappropriate to perform significance tests with population data, we are forced to use the variance around the model parameter to assess model fit in a less formal manner. Without such a significance test there is no arbitrary acceptance threshold for the ratio of the parameter to its standard deviance. Therefore, judgments about quality of fit are simultaneously more vague and less constrained by atheoretical conventions. Authors and readers must make their own judgments about the degree of variability they are willing to tolerate. The best manner for communicating this uncertainty is to consider population parameters as having a distribution as described by the fit of the model. This decidedly Bayesian approach will be described now.

A Bayesian Ecology Approach

In field biological and ecological research, differing ecosystems can be observed, studied, and compared, but they cannot be directly manipulated in the same way that a chemist or physicist does in the laboratory. In some respects this very much like working with state-level political science data, as there exist population-level variables, observations over time, substantial diversity in regional settings, great commonality of underlying principles, non-cooperative subjects, and measurement error. In this setting a question arises as to what is the scope of the system being evaluated. Ecologists have two primary perspectives on this question which are both directly analogous to state-level study in political science. One perspective emphasizes the completeness of the system at a particular moment and focuses on studying the population of plants and animals in a specific region or area (states

in our case). The other perspective, of equal importance in ecology, is that whatever can be observed is simply a cross-section from a time series of biological and physiological change: plants and animals evolve, atmospheres change, weather is non-constant, and tumultuous events occur. The recent introduction of Bayesian methods to ecology has synthesized these two perspectives and provided an overarching framework for analyzing such systems over time by updating descriptions of unknown parameters as new observations occur (Ludwig 1996; Reckhow 1990; Solow 1993; Taylor et al. 1996; Wolfson et al. 1996), although not without some criticism (Dennis 1996; Edwards 1996)

The central problem with analyzing ecological systems with standard population statistics is that most of these systems are uniquely occurring events in space and time, whereas frequentist statistical analysis is built upon expected occurrences over many repeated trials (Reckhow 1990). Restated, the conventional frequentist statistical approach requires an assumption that the calculated statistics have sampling distributions based on long-run behavior keeping explanatory factors constant, but population-level studies of enclosed and evolving systems (whether U.S. states or ecological areas of study) are non-replicated events over time. Therefore, the underpinnings of null hypothesis significance testing (p-values, predetermined α levels, constancy of assumptions) have no obvious theoretical interpretation (Berger 1985; Ellison 1996).

With Bayesian analysis, assertions about unknown model parameters are not expressed as point estimates with reliability assessed using the null hypothesis significance test. Instead, population parameters are assumed to be random variables themselves. Bayesian statistical information about these parameters is summarized in probability statements applied to either samples or populations. These summaries include quantiles of the posterior distribution, the probability of occupying some region of the sample space, the posterior predictive distribution, and Bayesian forms of confidence intervals such as the credible set and the highest posterior density region (Box and Tiao 1992, Casella and Berger 1990, Lee 1989, Leonard and Hsu 1999).

In the Bayesian setup, the unnormalized posterior (sampling) distribution for the un-

known coefficients is calculated by:

$$P(\boldsymbol{\eta}|D, H_0) \propto P(D|\boldsymbol{\eta}, H_0)P(\boldsymbol{\eta}|H_0) \tag{5}$$

where $\boldsymbol{\eta}$ is the unknown parameter vector, D is the data, and H_0 indicates the assumption of H_0.[3] So the desired probability statement is a product of the likelihood function, $P(D|\boldsymbol{\eta}, H_0)$, and some prior belief about the distribution of the quantity of interest, $P(\boldsymbol{\eta}|H_0)$.

The strongest argument for the inclusion of priors is that there often exists scientific evidence before a statistical model is developed and it would be foolish to ignore such previous knowledge (Tiao and Zellner 1964; Press 1989). The formal statement of a prior distribution is an overt, non-ambiguous assertion within the model specification that the reader can accept or dismiss (Box and Tiao 1973; Gelman et al. 1995). In addition, imprecise or vague knowledge often justifies a diffuse, or even a bounded uniform, prior (Jeffreys 1961; Howson and Urbach 1993), and certain probability models lead logically to particular forms of the prior for mathematical reasons (Good 1950; Press 1989).

A primary payoff for applying this Bayesian framework to population cross-sectional data is that it facilitates the explicit comparison of rival models about the system under study, even if these models are not nested: one being a restricted case of the other. Suppose Γ_1 and Γ_2 represent two competing hypotheses about the state of some unknown parameter, γ, which together form a partition of the sample space: $\Gamma = \Gamma_1 \cup \Gamma_2$. Initially, prior probabilities are assigned to each of the two outcomes: $\pi_1 = p(\gamma \in \Gamma_1)$ and $\pi_2 = p(\gamma \in \Gamma_2)$. This allows us to calculate the competing posterior distributions from the two priors and the likelihood function: $p_1 = p(\gamma \in \Gamma_1|D)$ and $p_2 = p(\gamma \in \Gamma_2|D)$, where D represents the observed data. The Bayes Factor combines the prior odds, π_1/π_2, and the posterior odds, p_1/p_2, as evidence for Γ_1 versus Γ_2 by calculating the ratio $B = (\pi_1/\pi_2)/(p_1/p_2)$ (Berger 1985; Kass and Raftery 1989; Lee 1995). Thus the Bayes Factor is the odds favoring Γ_1 versus Γ_2 given the observed data incorporating both prior and posterior information.

The Bayesian philosophy is centered on the belief that population parameters are not fixed quantities, rather they are realizations from an underlying distribution that can be described with a posterior (post-data) distribution. By treating the observed data as a prior

distribution, information about the parameters can be iteratively updated as new data are observed. For population cross-sectional data, this gradual pooling avoids the single period problem associated with frequentist analysis, and for data where there is evidence that the serial time periods are important determinants of serial autocorrelation in the errors, a Bayesian time series model can be specified (Broemeling and Shaarawy 1988; Poskitt 1986; Sims 1988; West and Harrison 1989).

The Bayesian approach outlined here solves three primary problems with current state-level empirical work: cross-sectional population studies are not generalizable across time, they cannot validly apply inferential tests, and time series specifications are sensitive to the flaws of the null hypothesis significance test. Finally, in the analysis of ecological systems and U.S. states data, the idea that there is a single true parameter of interest that remains fixed over some time period of interest is quite suspect. A model that explicitly accounts for the stochastic process generating the population data across cases and time as a distributional phenomenon is far more likely to be useful.

Conclusion

Authors have periodically lamented the paucity of serious academic attention by political scientists to the politics of the states (Brace and Jewett 1995, Herson 1957; Jewell 1982). One contributing reason for this historical disregard is attributable to the data. Political science wholeheartedly embraced survey research as the dominant quantitative methodological focus starting in the 1960s. The two primary reasons were a shift to the behavioral orientation in the discipline, and the 1950 admonition by Robinson (reinforced by others) not to make ecological inferences from aggregate data.

In this transition aggregate data, including state-level data, was subsumed in importance by individual level data from surveys. As a result some lingering questions went unaddressed. This article addresses an important but rarely discussed problem in the study of state politics: how should population models of state-level data be interpreted? The answer presented here is two-fold. First, models based on population cross-sectional data do not accommodate significance testing because the researcher already has the values that this procedure seeks. Therefore summary statistics and their measures of uncertainty are completely sufficient

measures of model quality in the possible presence of misspecification, autocorrelation, heteroscedasticity, endogeneity, and measurement error. However, since there is no supported testing mechanism in the regimented manner of the null hypothesis significance test, arguing the strength of evidence is a far more abstract task. Second, when the researcher seeks *any* generalizability beyond a specific cross-section, a model specification that incorporates time is required: either pooled or parameterized. However, time series models are particularly affected by the deeply flawed and widely misunderstood null hypothesis significance testing procedure which dominates empirical political science.

The best solutions to the problem of analyzing data collected at the level of U.S. state population are to describe the general model fit and to embrace Bayesian methodologies. Model assumptions, estimating procedures, and parametric characteristics all vary widely and have a profound effect on the quality of the fit to the observed population data. It is therefore important to use descriptive statistics to describe how closely the systematic component fits these data. In addition, the Bayesian focus on describing and updating posterior parameter distributions is exactly suited to the process of measuring and modeling population behavior across the 50 states. The Bayesian analog of hypothesis testing, the Bayes Factor, allows us to evaluate competing model specifications in a systematic search for the best description of the structure of the underlying population trends. It is hoped that this research note will focus some direction in state-level research towards this more appropriate and potentially productive methodological approach.

Endnotes

[1]A core problem with the null hypothesis significance test is that researchers pretend to select α levels a priori as in experiments based on Neyman-Pearson, but actually report p-values (or worse yet, ranges of p-values indicated by asterisks) as the strength of evidence (Gill 1999). This is because the discipline is cursed with Fisher's arbitrary thresholds (even he later recanted), despite the fact that there has *never* been a theoretical justification to support 0.01 and 0.05. Because the test is performed once on a set of unique state-level population data, the reported p-value is not a long run frequentist probability. Furthermore, since only one model specification is tested, an infinite number of alternate specifications

are not ruled out. A second problem with the null hypothesis significance test that pertains directly to research in public policy is that there is no explicitly modeled consequence of making the wrong decision (Gill and Meier 2000; Pollard and Richardson 1987). Unlike purely academic research, decisions taking place in policy analysis and implementation have direct consequences for citizens, employees, managers, and agencies in general. Yet hypothesis testing confuses inference and decision-making since it "does not allow for the costs of possible wrong actions to be taken into account in any precise way" (Barnett 1973). Decision theory is the logical adjunct to hypothesis testing that formalizes the cost of alternatives by explicitly defining the cost of making the wrong decision by specifying a loss function and associated risk for each alternative (Berger 1985; Pollard 1986).

[2]If for some reason the arc-tangent, inverse, square-root population effect of state-level GDP was important in a linear model, then we would see the opposite effect: a small variance for this model parameter.

[3]Proportionality is used here instead of equality because the $P(D, H_0)$ term in Bayes law, $P(\boldsymbol{\eta}|D, H_0) = \frac{P(\boldsymbol{\eta}, H_0)}{P(D, H_0)} P(D|\boldsymbol{\eta}, H_0)$, does not depend on $\boldsymbol{\eta}$ and can therefore be dropped. Bayes law is true whether in the context of Bayesian or frequentist statistical models, the difference being that Bayesians use this principle to update prior knowledge with new observations.

References

Akaike, H. 1969. "Fitting Autoregressive Models for Prediction." *Annals of the Institute of Statistical Mathematics* 21:243-7.

Bakan, David. 1960. "The Test of Significance in Psychological Research." *Psychological Bulletin* 66:423-37.

Barnett, Vic. 1973. *Comparative Statistical Inference.* New York: John Wiley & Sons.

Becker, Gary S., Michael Grossman, and Kevin M. Murphy. 1994. "An Empirical Analysis of Cigarette Addiction." *American Economic Review* 84:396-418.

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis.* 2nd ed. New York: Springer-Verlag.

Besley, Timothy, and Anne Case. 1995. "Incumbent Behavior: Vote-Seeking, Tax-Setting, and Yardstick Competition." *The American Economic Review* 85:25-45.

Blaylock, Hubert M. 1960. *Social Statistics.* New York: McGraw Hill.

Blaylock, Hubert M. 1961. *Causal Inferences in Nonexperimental Research.* Chapel Hill,NC : University of North Carolina Press.

Bonjean, Charles M., and Robert L. Lineberry. 1970. "The Urbanization-Party Competition Hypothesis: A Comparison of all United States Countries." *Journal of Politics* 32:305-321.

Box, George E. P. 1980. "Sampling and Bayes' Inference in Scientific Modeling and Robustness." *Journal of the Royal Statistical Society A* 143:383-430.

Box, G. E. P., and G. M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control.* San Francisco,CA : Holden-Day.

Box, George E. P., and George C. Tiao. 1992. *Bayesian Inference in Statistical Analysis.* New York: John Wiley & Sons.

Brace, Paul, and Aubrey Jewett. 1995. "The State of State Politics Research." *Political Research Quarterly* 48:643-81.

Broemeling, Lyle, and Samir Shaarawy. 1988. "Time Series: A Bayesian Analysis in the Time Domain." In *Bayesian Analysis of Time Series Models and Dynamic Models*, ed. James C. Spall. New York: Marcell Dekker.

Carmines, Edward. 1974. "The Mediating Influence of State Legislatures on the Linkage Between Interparty Competition and Welfare Policies." *American Political Science Review* 68:1118-24.

Casella, George, and Roger L. Berger. 1990. *Statistical Inference.* Belmont, CA: Wadsworth.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112:115-59.

Cohen, Jacob. 1994. "The Earth is Round ($p < .05$)." *American Psychologist* 12:997-1003.

Dawson, Richard E., and James A. Robinson. 1963. "Inter-Party Competition, Economic Variables, and Welfare Policies in the American States." *Journal of Politics* 25:265-89.

Brian Dennis. 1996. "Discussion: Should Ecologists Become Bayesians?" *Ecological Applications* 6:1095-103.

Dye, Thomas R. 1965. "Malapportionment and Public Policy in the States." *Journal of Politics* 27:586-601.

Dye, Thomas R. 1961. "A Comparison of Constituency Influences in the Upper and Lower Chambers of a State Legislature." *Western Political Quarterly* 14:473-81.

Dye, Thomas R. 1984. "Party and Policy in the States" *Journal of Politics* 46:1097-116.

Edwards, Don. 1996. "Comment: THe First Data Analysis Should Be Journalistic." *Ecological Applications* 6: 1090-4.

Edwards, George. 1990. *Presidential Influence in Congress.* San Francisco,CA : W.H. Freeman and Company.

Ellison, Aaron M. 1996. "An Introduction to Bayesian Inference for Ecological Research and Environmental Decision-Making." *Ecological Applications* 64: 1036-46.

Fiorina, Morris P. 1991. "Divided Government in the States." *PS: Political Science & Politics* 24:646-50.

Fisher, Ronald A. 1925a. *Statistical Methods for Research Workers.* Edinburgh, Scotland: Oliver and Boyd.

Fisher, Ronald A. 1925b. "Theory of Statistical Estimation." *Proceedings of the Cambridge Philosophical Society* 22:700-25.

Fisher, Ronald A. 1934. *The Design of Experiments.* 1st ed. Edinburgh, Scotland: Oliver and Boyd.

Fisher, Ronald A. 1955. "Statistical Methods and Scientific Induction." *Journal of the Royal Statistical Society B* 17: 69-78.

Fomby, Thomas B., R. Carter Hill, and Stanley R. Johnson. 1980. *Advanced Econometric Methods.* New York: Springer-Verlag.

Gary, Lawrence E. 1973. "Policy Decisions in the Aid to Families with Dependent Children Program: A Comparative State Analysis." *Journal of Politics* 35: 886-923.

Gray, Virginia. 1974. "Expenditures and Innovation as Dimensions of "Progressivism": A Note on the American States" *American Journal of Political Science* 18:693-9.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis.* New York: Chapman & Hall.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52:647-74.

Gill, Jeff, and Kenneth J. Meier. 2000. "Public Administration Research and Practice: A Methodological Manifesto." *Journal of Public Administration Research and Theory* 10:157-200.

Godfrey, Leslie G. 1987. "Discrimating Between Autocorrelation and Misspecification in Regression Analysis: An Alternative Test Strategy." *Review of Economics and Statistics* 69:128-34.

Good, I. J. 1950. *Probability and the Weighing of Evidence.* New York: Hafner.

Gray, Virginia. 1976. "Models of Comparative State Politics: A Comparison of Cross-Sectional Time Series Analyses." *American Journal of Political Science* 20:235-56.

Hedlund, Ronald D., and H. Paul Friesema. 1972. "Representatives' Perception of Constituency Opinion." *Journal of Politics* 34:730-52.

Heer, David M. 1966. "Economic Development and Fertility." *Demography* 3:423-44.

Hero, Rodney, and Caroline J. Tolbert. 1996. "A Racial/Ethnic Diversity Interpretation of Politics and Policy in the States of the U.S." *American Journal of Political Science* 40:851-71.

Herson, Lawrence J. R. 1957. "The Lost World of Municipal Government." *American Political Science Review* 51:330-45.

Hofferbert, Richard I. 1966. "Ecological Development and Policy Change." *Midwest Journal of Political Science* 10:464-86.

Hopkins, Anne H. and Ronald E. Weber. 1976. "Dimensions of Public Policies in the American States." *Polity* 8:475-89.

Howson, Colin, and Peter Urbach. 1993. *Scientific Reasoning: The Bayesian Approach.* 2nd ed. Chicago: Open Court.

Hunter, John E. 1997. "Needed: A Ban on the Significance Test." *Psychological Science* January, Special Section 8:3-7.

Lee, Peter M. 1989. *Bayesian Statistics: An Introduction.* New York: John Wiley & Sons.

Leonard, Thomas, and John S.J. Hsu. 1999. *Bayesian Methods.* Cambridge, England: Cambridge University Press.

Ludwig, D. 1996. "Uncertainty and the Assessment of Extinction Probabilities." *Ecological Applications* 6:1067-76.

Jeffreys, Harold. 1961. *The Theory of Probability.* Oxford, England: Clarendon Press.

Jennings, Edward T., Jr. 1979. "Competition, Constituencies, and Wlefare Policies in American States." *American Political Science Review* 73:414-29.

Jewell, Malcom E. 1982. "The Neglected World of State Politics." *Journal of Politics* 44:638-57.

Jones, Ruth S., and Warren E. Miller. 1984. "State Polls: Promising Data Sources for Political Research." *Journal of Politics* 46:1182-92.

Kass, R. E., and A. E. Raftery. 1989. "Bayes Factors." *Journal of the American Statistical Association* 90:773-95.

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* Cambridge, England: Cambridge University Press.

Knottnerus, Paul. 1985. "A Test Strategy for Discriminating between Autocorrelation and Misspecification in Regression Analysis: A Critical Note." *The Review of Economics and Statistics* 67:175-77.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: John Wiley & Sons.

Lee, P. M. 1989. *Bayes Statistics: An Introduction.* New York: Oxford University Press.

LeLoup, Lance T. 1978. "Reassessing the Mediating Impact of Legislative Capability." *American Political Science Review* 72:616-21.

Lindsey, James K. 1997. *Applying Generalized Linear Models.* New York: Springer-Verlag.

MacKinnon, James G. 1992. "Model Specification Tests and Artificial Regressions." *Journal of Economic Literature* 30:102-46.

Markus, Gregory B. 1974. "Electoral Coalitions and Senate Roll Call Behavior: An Ecological Analysis." *American Journal of Political Science* 18:595-607.

Marquette, Jesse F., and Katherine A. Hinckley. 1981. "Competition, Control and Spurious Covariation: A Longitudinal Analysis of State Spending." *American Journal of Political Science* 25:362-75.

Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Counseling and Clinical Psychology* 46:806-34.

Miller, Alan J. 1990. *Subset Selection in Regression.* New York: Chapman & Hall.

Neyman, Jerzy, and Egon S. Pearson. 1928a. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part I" *Biometrika* 20(A):175-240.

Neyman, Jerzy, and Egon S. Pearson. 1928b. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part II" *Biometrika* 20(A):263-94.

Neyman, Jerzy, and Egon S. Pearson. 1936. "Contributions to the Theory of Testing Statistical Hypotheses." *Statistical Research Memorandum* 1:1-37.

Nice, David C. 1994. *Policy Innovation in State Government.* Ames, IA: Iowa State University Press.

Pollard, P., and J. T. E. Richardson. 1987. "On the Probability of Making Type One Errors." *Psychological Bulletin* 102:159-63.

Pollard, W. E. 1986. *Bayesian Statistics for Evaluation Research.* Newbury Park, CA: Sage.

Poskitt, D. S. 1986. "A Bayes Procedure for the Identification of Univariate Time Series Models." *Annals of Statistics* 14:502-16.

Press, S. James. 1989. *Bayesian Statistics: Principles, Models and Applications.* New York: John Wiley & Sons.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." In *Sociological Methodology*, ed. Peter V. Marsden. Cambridge, MA: Blackwells.

Reckhow, Kenneth H. 1990. "Bayesian Inference in Non-Replicated Ecological Studies." *Ecology* 71(6): 2053-9.

Rivers, Douglas, and Nancy L. Rose 1985. "Passing the President's Program: Public Opinion and Presidential Influence in Congress." *American Journal of Political Science* 29:183-96.

Rinquist, Evan J. 1993. "Does Regulation Matter?: Evaluating the Effects of State Air Pollution Control Programs." *Journal of Politics* 55:1022-45.

Robinson, W.S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15:351-7.

Rozeboom, William W. 1960. "The Fallacy of the Null Hypothesis Significance Test." *Psychological Bulletin* 57:416-28.

Russell, Bertrand. 1929. *Mysticism and Logic and Other Essays.* New York: W.W. Norton & Company.

Schmidt, Frank L. 1996. "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers." *Psychological Methods* 1:115-29.

Schmidt, Frank L., and John E. Hunter. 1977. "Development of a General Solution to the Problem of Validity Generalization." *Journal of Applied Psychology* 62:529-40.

Serlin, Ronald C., and Daniel K. Lapsley. 1993. "Rational Appraisal of Psychological Research and the Good-enough Principle." In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, G. Keren, and C. Lewis, eds. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sharkansky, Ira. 1968. *Spending in the American States.* Chicago: Rand McNally.

Sharkansky, Ira and Richard I. Hofferbert. 1969. "Dimensions of States Politics, Economics, and Public Policy." *American Political Science Review* 63:867-79.

Shibata, R. 1980. "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process." *Annals of Statistics* 8:147-64.

Sims, Christopher. 1988. "Bayesian Skepticism on Unit Root Econometrics." *Journal of Economic Dynamics and Control* 12:463-74.

Solow, Andrew R. 1993. "Inferring Extinction from Sighting Data." *Ecology* 74(3): 962-4.

Taylor, B. L., P. R. Wade, R. A. Stehn, and J. F. Cochrane. 1996. "A Bayesian Approach to Classification Criteria for Spectacled Eiders." *Ecological Applications* 6:1077-89.

Thursby, Jerry G. 1981. "A Test Strategy For Discriminating Between Autocorrelation and Misspecification in Regression Analysis." *Review of Economics and Statistics* 63:117-23.

Tiao, George C., and Arnold Zellner. 1964. "Bayes's Theorem and the Use of Prior Knowledge in Regression Analysis." *Biometrika* 51:219-30.

West, Mike, and Jeff Harrison. 1989. *Bayesian Forecasting and Dynamic Models.* New York: Springer-Verlag.

Wiggins, Charles W., Keith E. Hamm, and Charles G. Bell. 1992. "Interest Group and Party Influence Agents in the Legislative Process: A Comparative State Analysis." *Journal of Politics* 54:82-100.

Wolfson, L. J., J. B. Kadane, and M. J. Small. 1996. "Bayesian Environmental Policy Decisions: Two Case Studies." *Ecological Applications* 6:1056-66.