An Introduction to *Quantitative Research in Political Science*

6 March 2015

Robert J. Franzese, Jr.
Professor of Political Science,
The University of Michigan, Ann Arbor

Modern academic commentators tend to characterize political-science scholarship prior to World War II as legalistic, favoring categorical enumeration of constitutional and historical details over positive-theoretical analysis, and offering anecdotal rather than systematic empirical analysis. This dismissive characterization of the prewar study of politics as almost entirely unscientific is surely unfair, but not until the behavioralism revolution swept political science in the 1950s, following on the Parsonian approach that had analogously revolutionized sociology in the interwar period, did a self-consciously positive-scientific approach to the study of politics grow ascendant (Brady et al. 2011). This new, positive, political science, as opposed to earlier exclusively normative or non-scientific approaches, offered theories about the systematic sources of regularities in political behavior, i.e., theories that posited explanations for the systematic aspects of the relationships that one could expect among socio-politico-economic variables. As *theory* in political science became this sort of positive exposition of systematic relationships between variables that, by the logical argument of the theory, one could expect to manifest regularly empirically, the primary aim, focus, and purpose of empirical analysis shifted in parallel from descriptive and factual inquiry – *what happened?* – to empirical evaluation of theory and empirical estimation of the theoretically posited relationships – *why and how, by what process, did it happen?*. Interestingly, this progression in the ends to which quantitative empirical research in political science were put can be followed in the various names King (1991) notes have been given the subfield before it took its current moniker, *political methodology*: from Rice's interwar

political statistics (Rice 1926) to Alker's post-behavioral-revolution *polimetrics* (1975).[1]

Before the beginning of its emergence as a subfield of political science in the 1970s, works in quantitative empirical-research methodology in the study of politics were scattered sporadically across the journals of political and related social sciences. Moreover, these rare early instances of political methodology were usually published only after hard-fought battles to defend each work's *bona fides* as a study of politics and the possibility of doing so with quantitative research-methods.[2] Through these battles, the subfield began to coalesce and gain a self-awareness that was beginning to show around the turn of that decade. Soon afterward, *The Society for Political Methodology* was formed for the purposes of, as is the case for most social societies, self-defense, mutual support, and advancement of the group's aims and activities, or, as one of the founding figures tells it:

> September 3, 1983 will not be the next national holiday, yet for some it is a day worthy of a celebration. Nor did it initiate ten days that shook the world, yet it marked the beginning of a quiet but profound revolution in Political Science. On that date, at the urging of a young Berkeley agitator turned methodologist named Steven Rosenstone, commenting on an APSA panel the day before, a group of young radicals gathered on the steps in the lobby of Chicago's Palmer House Hotel. (Efforts to place a plaque on the spot have been unsuccessful.) Accounts at the time put the number at twelve, but only seven have been positively identified in subsequent documents. Those seven are: Christopher H. Achen, John H. Aldrich, Larry M. Bartels, Henry E. Brady, John E. Jackson, George E. Marcus, Steven J. Rosenstone, and John E. Sullivan. Larry Bartels, the youngest possible member, now denies attending and claims he was not even in Chicago. The remaining individuals have remained unidentified and unindicted co-conspirators though there are rumors about their identity. It is also true that, much like Babe Ruth's famous Wrigley Field home run in 1932, the number of people who claim to have attended now exceeds the capacity of the lobby and some of those claims are from individuals who, at best, were in a pre-arithmetic state of development.
>
> A subsequent manifesto identifies the cell's purposes as three-fold: 1) To organize a formal meeting to create an agenda for action; 2) To take over a set of panels at the 1984 meeting of the American Political Science Association; and 3) To institutionalize themselves as afield. There were surely other even more incendiary topics covered, but because the group was very careful not to take minutes so there could be deniability these

[1] Note also Sir William Petty's political arithmetic (Petty 1672) a quarter millennium before Rice and the *politimetrics*, , *politometrics*, and *govemmetrics* of Gurr (1972), Hilton (1976), and Papayanopoulos (1973). Presumably, the *'metrics* labels from the 1970s fell from favor as political methodologists sought to avoid giving the impression that their subfield merely borrowed with rote translation from *econometrics*.

[2] Larry Bartels tells a story of a review of one of his first journal-submissions as a young graduate-student about this time informing the journal editor that his manuscript should obviously be rejected because one cannot trust any of its evidence with all its coefficients and standard errors and everything *being estimated*. [personal communications]

> are not documented. Though this manifesto hardly had the fiery rhetoric of the Port Huron Statement its intentions were as profound and its impacts more permanent. It initiated a bloodless revolution that without killing anyone or thing, save a few cross-tabs, some factor analyses, and a path analysis or two permanently altered the way empirical analysis would be done. (Jackson 2012: pp. 2-3)

The group did indeed convene a meeting for that next year. They met in Ann Arbor at the University of Michigan in the summer of 1984 for the first academic conference of the subfield of political methodology. Sixteen scholars discussed topics including psychometric measurement-models for political-science applications, ecological inference, and statistical models for binary outcomes (like voting, majoritarian elections, and wars). Three decades later, that annual Summer Conference of the Society for Political Methodology now regularly hosts 150-225 attendees and some 100-200 paper and poster presentations by faculty and graduate students, on topics ranging from simulation-estimation methods for complex structural models to nonparametric causal-inference, from Bayesian models for textual data to computational data-mining for real-time streaming data, and from methods for individual-level data (and even sub-individual-level neurological data) to time-series-cross-sectional data of aggregated units from medieval to modern times. Routinely now, many of the eventually polished products of these conference papers and posters are published in all the top journals of political science, including the Society's own top-rated journal, *Political Analysis*, in the top methodological journals across the social sciences, and in the top journals in statistics as well.

This five-volume collection selects 55 "Major Works" from the academic literature of this fast evolving and now enormous subfield called *political methodology*, focusing exclusively therein on *quantitative empirical-research methodology*.[3] These 55 selections are organized, following the editor's conceptualization of the central challenges of empirical analysis (quantitative or

[3] I.e., the collection does not include so-called formal-*theoretical* methods, some of which are *quantitative* also, nor does it seek to encompass *qualitative* empirical-methods in its purview.

otherwise) in the political and social sciences, in six parts across the five volumes:

a. Section 1 introduces the central challenges of quantitative empirical analysis in political science, and offers some overviews of one promising approach to tackling these challenges.
b. Section 2 starts this thusly-organized tour of quantitative empirical-research methodology from the starting point of all empirical analysis: measurement.
c. Section 3 reviews the main applied tools of multivariate analysis (which relates to the first central challenge: *multicausality*): least-squares linear-regression, maximum likelihood for limited and qualitative dependent-variables, and Bayesian methods of estimation and inference.
d. Section 4 covers the various ways empirical methodology has addressed heterogeneity (which relates to the second central challenge: *context conditionality*) of certain sorts – differing levels of outcomes, differing effects of variables – across the observational units.
e. Section 5 addresses empirical methods for dynamic and interdependent processes (which also relates to the second central challenge and to the third challenge: *ubiquitous endogeneity*), encompassing temporal (cross-time), spatial (cross-unit), and spatiotemporal (cross-unit-time) dependence and dynamics.
f. Section 6 surveys approaches to and methods for estimation and inference given *endogeneity*, i.e., simultaneity or mutual causality among the variables of the analysis (which is the third central challenge): systems estimation, instrumentation strategies, autoregressive/temporal-ordering based methods, discontinuity and difference-in-difference designs, matching methods, and experimentation.

The volume's six parts are organized around the editor's conception of the three central challenges for empirical research in social science. However, "the [one] fundamental problem of causal inference" (Holland 1986), empirical researchers and methodologists are now told, is that we would like to know the difference in outcome, y_i, for an observational unit, i, between getting some treatment, x_i=1, and not getting that treatment, x_i=0, but unfortunately we can logically only ever observe x_i=1 or x_i=0, never can both x_i=1 and x_i=0 happen to the same, single observed unit (at the same time). With this difference of $(y_i|x_i=1)$-$(y_i|x_i=0)$ understood to be, simply and unambiguously, ***the*** quantity of interest in empirical analysis, this Neyman-Rubin-Holland causal-model school of thought teaches that all challenges of empirical research come back to this one, unavoidable issue: the unobservability of the counterfactual. Alternatively, empirical methodologists and researchers in microeconometrics today are told that empirical analysis of micro-level units (as opposed to aggregates) "fundamentally…brings to the forefront

heterogeneity of individuals, firms, and organizations that should be properly controlled (modeled) if one wants to make valid inferences about underlying relationships" (Cameron and Trivedi 2005: p. 5). From this view, the one core problem is heterogeneity, and especially unobserved heterogeneity (Wooldridge 2002). The editor's view is that empirical methodologists and researchers in the political and social sciences are hardly so lucky as to have only just *one* fundamental or core problem. Rather, the central challenges of empirical analysis, quantitative or otherwise, in the political and social sciences number (at least[4]) three. The organization of these five volumes in six parts follows the methodological attacking of those three central challenges for quantitative empirical research in political science.

In generating positive theoretical explanations of the systematic relations among socio-politico-economic phenomena, we know or strongly suspect the following three features to manifest regularly: *multicausality*—just about everything matters; *context conditionality*—the way just about everything matters depends on just about everything else; and *ubiquitous endogeneity*—just about everything causes just about everything else. The introductory selections of Section 1 elaborate these challenges and emphasize one promising approach to successfully managing them, if not to say surmounting them, which approach is essentially to lean very heavily on the *theory* and *substance* of the scenario of study to specify empirical models for powerful leverage on this complex reality.

In Section 1: Introduction, the editor's own contribution, "Multicausality, Context-Conditionality, and Endogeneity," elaborates on the meaning and likely universal applicability in

[4] A fourth and a fifth core challenge could be mentioned as well – fourth: far too little data, or useful variation to be more precise, to surmount the first three challenges purely empirically, i.e., non-parametrically; and fifth: the truth, i.e., the DGP, is likely moving while we're trying to estimate it – although the fourth may be only *typically* and not *universally* severely manifest, and the fifth may be subsumable under the second as another form of heterogeneity in the process being analyzed.

social science of these three challenges. In essence, the three fundamental challenges imply that, with each empirical observation on however many variables, we will generally have some multiple of that many empirical quantities (e.g., relationships or averages, i.e., parameters) to estimate. With each empirical set of facts we learn, we have some times that many things more to estimate, or to learn, *unless*, that is, we impose some structure on our inquiry, like that the observations are generated according to a pre-specified model, a structural equation or set of equations, or that all observations where x=0 come from some distribution with one, constant mean, and those where x=1 come from some other distribution with one other, constant mean. The tremendously useful, in fact: *indispensable*, point of these structural pre-impositions is that they greatly reduce the parameterization, the number of parameters or things that we will have to estimate, with each observation from the multiple of the number of things observed that a truly nonparametric approach would require. The suggestion of the first article in the introduction section is that, confronted with a multi causal, context conditional, dynamic and ubiquitously interdependent world, we not only cannot avoid leaning on theoretically and substantively pre-imposed structure in our empirical models, we must have sufficient such structure to learn anything empirically, and we *should* therefore impose structure that reflects as well as possible what theory and substance tell us about the phenomena. This is the promising way forward intimated above. The next two articles in the introduction are summary or foundational in two specific ways of acting upon these principles. Granato and Scioli, in "Puzzles, Proverbs, and Omega Matrices," explain the EITM – Empirical Implications of Theoretical Models – initiative, which is a now 12 year-old program, initiated from the Political Science Program of the Social and Behavioral Sciences Directorate of the National Science Foundation (under Scioli's and Granato's leadership at the time), training young scholars in methods for integrating theoretical models (primarily formal ones) more tightly

with empirical analyses (primarily quantitative ones). Bas, Signorino, and Walker's "Statistical Backwards Induction" is a foundational work in one particular kind of tight integration between formal-theoretical and quantitative-empirical model, one in which the game-theoretic process of determining strategic behavior by solving backward, from the actors' optimal choices at the end of the game through all earlier decision back to the initial choice node, becomes a statistical model with empirical data input for unknown payoffs as parameterized functions to be estimated (by maximum likelihood). (An example of a different kind of tight integration is Franzese's "Multiple Hands on the Wheel" in Section 4, which illustrates a theoretically specified model for a particular manifestation of complex context-conditionality, estimated by nonlinear least-squares.)

Of course, before we can begin to tackle any of these of challenges empirically, we must have measures of the phenomena to be analyzed. Measurement is therefore fittingly the subject of Section 2, immediately following on the Introduction, as it is the first step in any empirical science:

> To measure is to know. — Lord Kelvin (1883). [or, as elaborated:]
>
> In…science, the first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it. I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be. — Lord Kelvin (1883).
>
> If you can think about something, you can conceptualize it; if you can conceptualize it, you can operationalize it; and if you can operationalize it, you can measure it. — J. David Singer [as quoted in Clark, Golder, and Golder 2013: pp., 143.]

In political science, many of the objects of interest for study are abstract concepts like *power* in international politics, *democracy* in comparative politics, *ideology* or *policy preferences* in domestic politics. Some of the most important achievements in political methodology have been advances in measurement. Indeed, measurement, along with specification or design, are the first-order essentials of any empirical analysis, as Lord Kelvin was very fond of noting. The most

advanced technical wizardry of advanced estimation methods, or the cleverest and most thoughtful of experimental or observational research designs, can only yield useful results insofar as the measures (and models) onto which they are applied are well taken (and specified). The selections in Section 2a represent in various ways some of the important political-methodological contributions in measurement. Achen's (1983) "Toward Theories of Data" surveys the fledgling subfield of political methodology and places measurement at the heart of its seed. Jackman's (2008) "Measurement" contribution to the *Oxford Handbook of Political Methodology* follows with a modern overview of the methods and methodology of measurement in political science 25 years later. The two virtues measures should maximize are *reliable accuracy*[5] — reliability being low variance across different measures of same item, and accuracy being small average error and so small bias of the measurement from the truth it was to measure — and *coverage*, which is measuring many instances of the item and failing to measure few. The absence or relative lack of these two virtues are, respectively, measurement error and missing data, two data problems that arise immediately along with our coverage of measurement. Groves (1991), sociologist and survey methodologist, and former Director of the U.S. Census Bureau, gives authoritative overview of "Measurement Error across Disciplines."[6] King and distinct groups of co-authors then give us two seminal contributions in political methodology, one toward each of those two ills. In "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research," King, Murray, Salomon, and Tandon (2004) show how to use "vignettes" or descriptions of particular scenarios, the same ones presented to different groups, to anchor respondents in the same places,

[5] Since both together are desirable, I merge the two properties of measures most introductions to the topic emphasize, *reliability* and *accuracy*, to make space for a third quality, *coverage*.

[6] In the first selection, Franzese had shown that, in general, arguments to reduce the number of observations by on grounds of high-quality measurement are unsustainable, so methods of improving measurement accuracy in non-tiny samples must instead be the way forward (see Table 1 and the surrounding text in that selection).

at least objectively, when they give Likert-scale or other scales or index responses to survey questions that, for instance, ask them to put themselves or other entities on left-right scales. A tightening of the link from substantive concepts in theory to empirical measures of those concepts as operationalized in variables, a tightening that in turn strengthens the connection from theoretical argument to empirical evaluation by reducing measurement error. Lastly in Section 2a: King, Honaker, Joseph, and Scheve (2001), in "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," offer the most influential address in political science to the challenge of missing data.

Section 2b turns to applications of important statistical models for measurement as developed and applied to measure political-science abstractions like *ideology* or *preferences* or *partisanship*, or *democracy*. Selection 9 provides a methodological introduction to perhaps the most-widely used innovation in measurement in political science, Poole and Rosenthal's *Nominate* score, which is a method of using U.S. Congressional roll-call votes on bills to place Representatives and the bills on a single left-right continuum or in multiple dimensions, based on how the legislators divide on each vote and how each vote divides the legislators. Although (1) the method is most useful only in legislatures where individual representatives have appreciable tendency to vote independently (as opposed to purely in party blocks), and (2) the method cannot get to "underlying" preferences but only those *revealed* by the presumably strategic votes of legislators, and (3) the method has been criticized on other grounds – just for one example: a probable leaning toward detecting few dimensions of difference among legislators and bills, usually just one dimension, which suggests the explanatory power of the likely unidimensional results should probably not be used to justify conclusions about the dimensionality of political contestation or its representation in the legislature – the importance and influence of the nominate score methodology for offering concrete, and it

seems accurate and reliable, measures of such a critical abstraction as political preferences or ideology can hardly be overstated. *Democracy* is another hugely important abstraction in political science, and so using Trier and Jackman's "Democracy as a Latent Variable" to illustrate the application of (Bayesian) item-response theory (IRT) measurement modeling, another of the most widely used methodological approaches to measurement, is very appropriate. Lastly in Section 2b, Stimson, MacKuen, and Erikson's "Dynamic Representation" gives thorough introduction to their long-term project measuring *public mood*, i.e., the general left-right leaning of the broad scope of public opinion in a jurisdiction, and *macro-partisanship*, i.e., the general left-right leaning of policy and governance in a jurisdiction, in the service of gauging the quality of democratic representation and of evaluating theories involving those abstractions. The three applications demonstrate the foundational importance of measurement methodology to the advancement of a positive scientific study of politics.

Volume II and Section 3 of the collection turns from understanding the central challenges of social-science empirical-analysis and proposing strategies to address them, and from the starting point to political-science empirical-analysis, (operationalization and) measurement of its often abstract theoretical conceptions, to the practical implementation of empirical-analytical methods suggested by the former and made possible by the latter. The first central challenge, *multicausality*, or that "just about everything matters," is more or less precisely what the methods of multivariate analysis covered in Section 3 are designed to address: to control for other potential explanators of the outcomes of interest besides the one of the researcher's main focus and interest.[7,8] The

[7] The editor would argue against conceiving that the entire or only aim for empirical analysis is to test for existence of one causal mechanism, and so that the *only* point of multivariate analysis is to control for other possible causes, for *confounds* to use the term currently *en vogue*, this is: *omitted variables*, to use the old and still perfectly serviceable term, when testing or estimating one's preferred factor's effect.

[8] A modern focus on the possibility of unobservable, or at least unobserved, other causal factors, leads some to stress very strongly the tremendous advantage of experimental control over statistical control in observational data in that randomization of experimental treatment eliminates (in large samples, assuming perfect randomization) the possibility

selections in this part overview the main workhorse techniques in applied empirical analysis in the political and social sciences: least-squares linear-regression, maximum-likelihood estimation of limited and qualitative dependent-variable models, and Bayesian methods for estimation and inference. Achen's (1982) Sage "Little Green Book", *Interpreting and Using Regression* is the classic, and remains the best, practical introduction to linear regression. The selection from that wonderful monograph that is Section 3a here introduces regression as a tool for empirical analysis of social science theory, which exactly fits the organizational theme of this collection. King's (1998 [1989]) *Unifying Political Methodology*, which leads off Section 3b, analogously served as many subsequent generations of political methodologists' introduction to maximum-likelihood estimation and inference, in particular for limited and qualitative dependent-variable models. The selections in this collection provide the highlights of that classic introduction. The selections from Gill's (2001) *Generalized Linear Models* (GLM) lucidly introduces and explains the GLM approach to the broad class of limited and qualitative dependent-variable models that can be encompassed under the generalized-linear framework (e.g., exponential-family and so log-linear models). King, Tomz, and Wittenberg's (2000) "Making the Most of Statistical Analyses", which concludes Section 3b, effectively pressed upon the by-then almost thoroughly positive-theoretical and quantitative-empirical political science the crucial importance of working through the substantive meaning of their estimates, in terms of how outcomes of interest, y, would respond to hypothetical movements in explanators, x, which empirical methodologists and researchers will have well-learned by then also (if they were paying attention), were <u>*not*</u> simply the estimated coefficient on x, except in the simplest purely linear-additive-and-separable models. This was and

of any alternative cause, including unobserved or even as-yet unconceived ones, but the treatment. Again, the editor would argue against concluding that this great virtue, and the great peril of *unobserved confounds*, great though they are, being the only weighty considerations in choosing a method of empirical analysis.

is a critically important lesson the selection imparts, along with providing a very useful software tool for calculating these *quantities of interest* along with estimates of their standard errors.

Section 3c covers the Bayesian framework for estimation and inference, which framework is highly amenable to the approach to the three central challenges pushed in the introduction, which might be summarized as seeking empirical models to reflect, as accurately and statistically powerfully as possible, the theoretical argument or model. Designing the empirical analysis into the Bayesian posterior (priors times likelihood) in this way is logically straightforward (although not uniquely so: one could also write the empirical version of the theoretical model into the likelihood of that framework or as the expected outcome in least-squares framework). The selection "Specifying Bayesian Model's" from Gill's *Bayesian Methods: A Social and Behavioral Science Approach* introduces the framework in such terms. To elaborate: in the least-squares framework, one can build the theoretical model of what we expect the outcome, y, to be as a function of certain explanators or covariates, x, connected according to a theoretically specified $f(\mathbf{x},\boldsymbol{\beta})$ such that $E(y) = f(\mathbf{x},\boldsymbol{\beta})$ so that $y = f(\mathbf{x},\boldsymbol{\beta}) + \varepsilon$. In the baseline classical linear regression model, we assume that $f(\mathbf{x},\boldsymbol{\beta})$ is linear-additive and separable, but later on we learn that the key for least-squares estimation is not the linearity, but the separability (of the stochastic component, ε, from the systematic, $f(\mathbf{x},\boldsymbol{\beta})$). In the likelihood framework, we can easily extend such an approach further, modeling as some function of explanatory **x**'s the parameters of any distribution that could characterize an outcome of interest, whether that distribution affords separability or not. The Bayesian framework allows us to go a step further. The Bayesian posterior being simply the likelihood times the prior, we have in the prior an additional place to put substantive and theoretical information and structure, and that knowledge we bring to the empirical model can be of just about any sort. The selection from Bartels, "Pooling Disparate Observations", shows for example how

incorporate substantive and theoretical knowledge we may have about the varying extents to which different observations fit the ideal of the population to which the theory applies. Jackman's "Estimation and Inference Are Missing Data Problems: Unifying Social-Science Statistics via Bayesian Simulation" then very nicely brings us back to a unified Bayesian perspective on the measurement, specification, and estimation that form the core of any empirical analysis.

Multivariate analysis, whether by least squares regression, maximum likelihood estimation, or Bayesian methods of estimation and inference, is the first and still greatest advance in empirical methods for sciences, like the political and social, in which perfect isolation of objects of study to perfectly insulated experimental laboratory, so that a multicausal world can be reduced to a monocausal observational environment, is impossible and undesirable.[9] Since we cannot study much empirically by experimental *control* both well and with much applicability, these methods of *statistical control* in observational data must instead be the main tool.[10] Statistical control, whether by multivariate analyses of the traditional methods covered in Section 3 or by matching methods covered in Section 6e, requires that the multiple explanatory variables to be controlled statistically – or, as better fits the approach of these volumes' emphasis to say, requires that the multiple explanatory variables whose partial associations are to be estimated – be *observed*. Especially according to microeconometricians, as the quotes from Cameron and Trivedi and from Wooldridge above testify, the great concern in any cross-section or panel (i.e., time-series cross-section) dataset is cross-unit heterogeneity, especially *unobserved heterogeneity*.

Heterogeneity and heterogeneous effects are the subject of Section 4, starting with unit (and

[9] Such isolation is undesirable because, even if it were possible, the resulting ideal causal inferences would be useless for application to the environments in which we would want to use the knowledge gained since we have no reason to believe the causal relations in the non-insulated reality would mimic those from the isolated laboratory.

[10] Figure 1, equation 6, and the surrounding text in the first introductory selection from Franzese shows that, even in the simplest possible case – purely linear-additive and separable trivariate regression model properly accounting for all the variations and covariations among the three variables needed to produce the correct partial association of each *x* with *y* non-quantitatively would not seem humanly possible.

period) *fixed effects* in Section 4a of Volume II. Oddly, what econometricians call *unit effects* are, ion their essence, not actually *effects* at all as most empirical methodologists, ironically including economists, understand that latter term. *Effects*, say the effect of x on y, are differences or derivatives, like dy/dx or $\Delta y/\Delta x$, or $\partial y/\partial x$. *Unit effects*, on the other hand, are the differences in expected outcomes from one *unit* to the next, *ceteris paribus* , i.e., at the same **x** values): $E(y_i \mid \mathbf{x}) - E(y_j \mid \mathbf{x})$. Empirical researchers, especially microeconometricians, worry about the possibility of these mean-level differences across units, especially if these differences are unobserved or insufficiently modeled because of situations like those commonly depicted as in Figure 1. In the Figure, the *unit effects* are the unmodeled differences in mean-level across units, i.e., $E(y_i \mid \mathbf{x}) - E(y_j \mid \mathbf{x})$, which are also given by the y intercepts in the figure. The observations in each unit show a strong positive relationship between x and y, as separate regression lines unit by unit would reveal. Unfortunately, the mean values of x also correlate, strongly negatively in this case, with the unit effects. If unmodeled, these unit effects would induce an enormous negative bias on a regression line estimated from the pooled data.

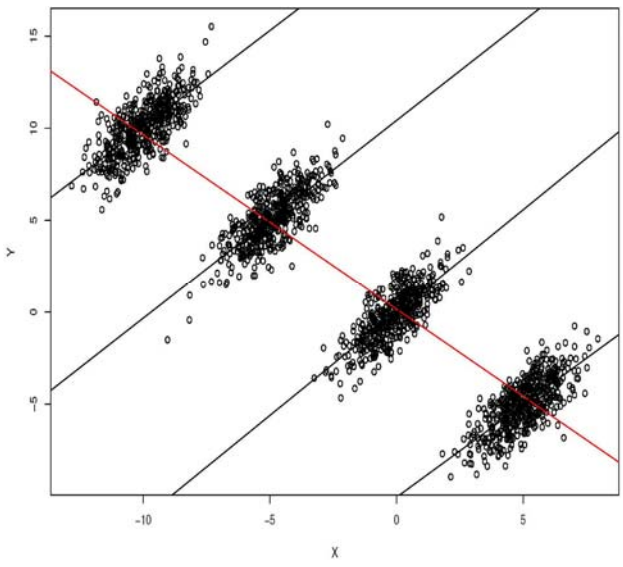

Figure 1: Sign-Flipping Bias from Unmodeled Unit Effects

This is the sort of horror story that leads the microeconometricians, and Green, Kim, and Yoon, in the first selection of Section 4a "Dirty Pool," to stress the fixed-effects guard against unobserved heterogeneity. Very simply, if one includes indicator variables for each unit, or mean-differences all the data unit-by-unit, the unit heterogeneity will be absorbed by these unit-dummy *fixed effects* (or swept away in the mean differencing), and the nightmare seen in Figure 1 cannot materialize. Period effects refers to some time-period by time-period difference, which, analogously, raises the specter of bias from unobserved period-heterogeneity if those unmodeled or unobserved mean-differences across time periods correlates with the means of the **x**'s (across units) period-by-period. The problem against which unit effects so effectively guards can be real and can be as severe as Figure 1 illustrates, but these fixed-effect cures also come at severe efficiency costs. They "sweep away" *all* of the cross-unit variation, and the remaining useful cross-time variation may often be small. This means that many, many time periods of data across the units may be necessary to find evidence even of reasonable strong true associations between x and y. In the most-extreme case, as Beck and Katz's "Throwing Out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon" emphasize, units without over-time variation are effectively eliminated from the sample, and this can make fixed-effect estimation biased in an attenuation direction if the lack of variation is simply because variation is rare or small and sample sizes are non-infinite, as may very well be the case in the substantive context of those papers: always-peaceful dyads in interstate-conflict datasets. Furthermore, as Troeger shows in "Problematic Choices: Testing for Correlated Unit Specific Effects in Panel Data," these unit dummies tend to overfit in limited samples, exaggerating any unit heterogeneity truly present and tending to find unit heterogeneity even where it is not present, and this overfitting results in poor performance of statistical tests, such as the Hausman Test of fixed-effect versus non-fixed-effect regression in panel or time-series

cross-section data, of whether the scenario depicted in Figure 1 is manifest. The upshot of the demonstration in Section 4a of a potential carcinogenic problem in unaddressed unit heterogeneity but also of the grave side effects of the fixed-effects chemotherapy for that cancer is not that empirical methodologists should freeze in indecision or give up in hopelessness, but to recognize that thoughtless insistence on always applying fixed-effects because the Figure 1 scenario is so scare reflects an insufficiently nuanced understanding of the problems and the methods for its address. One simply must tread carefully, weighing the very real and sizable cons and not just the remarkable pros, of sweeping or dummying away cross-unit and/or cross-time heterogeneity.

As Troeger also notes elsewhere (n.d.), this standard textbook heterogeneity of fixed mean differences across units (or time period) is perhaps the least interesting and certainly the simplest sort of heterogeneity that may concern us. The second of the central challenges for empirical analysis in the political and social sciences, context conditionality, speaks to a much richer and more interesting sort of heterogeneity, which is a *bona fide* heterogeneity in actual *effects*. That is, the effect of **x**'s tend to vary across contexts. The selection from Kam and Franzese *Modeling and Interpreting Interactive Hypotheses in Regression Analysis* shows how to go from theoretic statements about how the effects of x on y depend on z to the simplest empirical models reflecting such a consideration, the linear-interactive model. In this case, one very simply creates a third variable, xz, by multiplying x and z together and includes x and z and xz into the regression (or nonlinear, limited, or qualitative dependent-variable model). The result is a model in which the effect of x on y, i.e. dy/dx or $\Delta y/\Delta x$, or $\partial y/\partial x$, is no longer some single constant, given by the coefficient on x, but rather the effect of x on y is a function of z (and the coefficients on x and xz). The linear interaction is a simple yet powerful device for effectively modeling a limited amount of relatively simple interactivity, but when the substantively and theoretically implicated nature of

context conditionality is complex, such as for instance in modeling policy when multiple actors have some hand in policymaking, the linear-interactive model can grow unwieldy with the proliferation of numerous colinear (i.e., correlated) regressors. This is the same problem against which highly nonparametric approaches to empirical analysis in the political and social sciences sooner or later run afoul as well: when one eschews imposing structure one empirical estimation models, the number of things to learn (estimate) empirically grows faster than one-for-one with the number of contexts observed. Given this unavoidable fact that empirical analysts must impose structure to make any headway, the approach providing the lens through which we view the contributions to this collection suggests, as does Franzese in "Multiple Hands on the Wheel" that we impose theoretically and substantively derived structure, so that the empirical results may be theoretically and substantively informative. In the case of monetary policymaking in the open and institutionalize do context, for example, control of monetary policy is shared between governments and central banks, with the latter weighing more the greater the bank's independence (Franzese 1999), and that domestic duo controls monetary insofar as they have not delegated it to foreign banks and governments via an exchange-rate peg or effectively cannot maneuver domestically (even to peg their exchange rate) because theirs is a small and financially open economy. Insofar as those latter two conditions, explicit delegation via exchange-rate peg or de facto via small-openness, hold, the country's monetary policy is effectively given by the peg-currency's policy or some global equilibrium policy. All these considerations add to a conclusion that the effect of every factor, domestic or , that might influence monetary policy depends on the degree of central bank independence, the extent of exchange-rate-peg efficacy, and the small-openness of the economy, all at home and abroad, and through those international connections on all the domestic factors of all the foreign countries to which the current country is connected. This complex context-

conditionality would be a multicolinear nightmare to model by linear interactions, but the selection shows how to impose the structure implied by this nested set of "shared policy-controls" in an empirical model that can be estimated by nonlinear least-squares with only very few more parameters, 3 or 4, to estimate than a linear additive model. The selection thus illustrates both the effectiveness of the theoretical-structure imposition approach for the second central challenge of context conditionality, and introduces the extremely useful empirical-estimation tool of nonlinear least-squares.

From this perspective, Section 4c redresses a limitation of the prior two subsections' approaches to heterogeneity, which is that the cross-unit mean-differences or the heterogeneity in effect parameters across contexts is modeled as varying deterministically. The oddness of this is perhaps best seen in the linear-interaction model. In that case, we are modeling the effect of x on y (and of z on y symmetrically) to depend <u>*without error*</u> on z (on x, symmetrically). Odd that we do not know $E(y|x)$ with certainty, instead $y = E(y|x) + \varepsilon$, but we do know with certainty $E[\mathrm{d}(\mathrm{d}y/\mathrm{d}x)/\mathrm{d}z]$, it's just β_{xz}, without error. One way to view the multilevel or hierarchical model, the random effects and/or random coefficients model, the mixed effects model, the error components model, or shrinkage estimators – these are all basically synonyms, alternative labels for the same class of models that emphasize different aspects or uses for that single kind of model with error terms or stochastic elements added to the intercept and slope coefficients – is that these allow us to model the heterogeneity in the intercepts and coefficients across units, or across any sets of observations really, to have random error in them, i.e., to be random (across repeated samples). The selections in Section 4c, beginning with Western, "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach," which is addressed to comparative political economists in particular, introduce political scientists to this use and

approach for the Bayesian hierarchical model: theory specifies how and why *expect* coefficients to vary across contexts but there is also an error component added to these systematic components modeled as interaction terms. In "Modeling Multilevel Data Structures," Steenbergen and Jones gives thorough introduction to the empirical methodological approach from an angle that emphasizes the multilevel-data context, in which the researcher observes multiple *microlevel* observations within each of multiple *macrolevel* units and a key part of the aim is, typically, to model the variation in parameters of the *microlevel* model (often behavioral) depending on differing macrolevel contexts or differing values of conditions in the *macrolevel* contexts (often institutional). Lastly in section 4c, Park, Gelman, and Bafumi's "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls" represents a current emphasis in the use of multilevel models, from leading figures in their development, one that returns in fact to one of its original motivations in the education literature, which is to "borrow leverage" across many macrolevel units (originally: across schools, for instance; in Park et al.: across states) to estimate quantities at the microlevel wherein the number of observations is few (some or all schools have few surveyed classrooms or students; some states have few respondents in national-level polls).

Section 5 turns from heterogeneity across observational units, of various forms, modeled in various ways, to dynamics, temporal dynamics within units and/or spatial *cum* network dynamics across units. Dynamics, it turns out, also provide transition from the second central challenge of context conditionality, because outcomes and effects of explanators vary in a sense over the courses of dynamic processes, to the third central challenge, ubiquitous endogeneity, because dynamic processes tend to leave outcomes in different observations interdependent with each other, directly and without caveat in the spatial-interdependence / spatial-dynamics case and in a

way or potentially in temporally dynamic contexts. The crucial features that change as in the move to observations in dynamic processes are, first, that observations cannot be taken as independent, i.e., each new observation is not wholly new information, as it depends on, overlaps with, comes from previous or other-units' observations. The empirical methodology, therefore, must properly account the inter-observational dependence so as not to lie, usually egregiously, about the amount of information in the data and so about the certainty of estimates. Secondly, and still at this rather technical level, failure to model the dynamics omits an important part of the causal process from the model, and so likely induces bias in the parameter estimates of the usual omitted-variable sort. The omitted factor, *the past* or *history* in the temporal dynamics context for instance, is probably among the, if not *the*, most important feature of the causal process, so its omission can have enormous bias consequences.[11] More importantly from the perspective of the present collection even than these large technical issues of obtaining quality parameter and certainty estimates in dynamic contexts, however, is that dynamic processes imply that the responses in outcomes to causal factors, x, cannot be captured by one, scalar number, like a single coefficient, β, or a difference in means, $E(y_i \mid x_i = 1) - E(y_i \mid x_i = 0)$. The response in y to some movement in x is *dynamic*. This means that the response to y to some shift in x in a temporally dynamic process, for example, will be a *stream* of how y responds over time to this hypothetical movement in x, a whole set of y's period by period after the shock to x. And the hypothetical needs to be specified more

[11] In the temporal dynamics case, this "bias" is partly a small-sample issue (i.e., samples with T, the number of time-periods, small) and partly a "bias" of different sign in each sample, such that they may cancel across repeated samples. The upshot in that case is that the properties of estimates in temporally dynamic data that fail to model that process are unbiasedness and consistency and *merely* inefficiency. It's just that the *inefficiency* in the temporal-dynamics case is generally enormous, such that the unbiasedness across many samples or in infinite samples is cold comfort since the only one estimate we have in a sample decidedly less than infinite is so variable that we have little reason to believe the estimates are anywhere close to the parameters. In the spatial dynamics case, biasedness (as well as inconsistency and inefficiency) from omitting the interdependence follows more directly and surely, without need of such further explanation, although the magnitude of these flaws tends be smaller than in the temporal dynamics case because cross-unit dependence is typically not as strong as over-time dependence within the same unit.

precisely, i.e., dynamically, also: how exactly does the hypothetical *stream* of *x*'s values occur over time, a +1 unit shift in *x* in some period t_0, reversed back to 0 the next period and 0 thereafter, or a permanent +1 shift starting in t_0 or some other stream of movements in *x*? In space even more so, we must specify the hypothetical movements in *x* across all units, as well as the connections between units, before can begin to calculate the response in *y*, which likewise is a vector, and not a scalar, of responses across all units *i*. All this adds to the important conclusion that we must *model* the dynamic processes and then interpret the estimation results in a manner appropriate to that dynamic context, i.e., *dynamically*.

Section 5a begins the treatment of dynamics with a focus on temporal dependence, first with an extraordinarily helpful overview of the main temporally dynamic models in Beck's "Comparing Dynamic Specifications: The Case of Presidential Approval." That article introduced political-science empirical researchers and methodologists to the DHSY (Davidson, Hendry, Srba, and Yeo 1978) form of the error-correction model, which gives a single-equation expression of the possibly complex differentiation of equilibrium and transitory dynamics in relationships of *x* to *y*. In "Taking Time Seriously," De Boef and Keele give an updated overview of alternative time-dynamic models, showing the close relations between several of them and their possible grouping under the general rubric of, as various specifications of, the autoregressive distributed lag model. Both these first two selections also focus appealingly on the interpretation of estimation results from dynamic models. In "Taking Time Seriously: Time-Series–Cross-Section Analysis with a Binary Dependent Variable," Beck, Katz, and Tucker offer a simple way to account time dependence in models for binary outcomes. Carter and Signorino's "Back to the Future: Modeling Time Dependence in Binary Data" gives an even simpler expedient that can properly account time dependence on binary-outcome serial observations at least as effectively. These are crucially

important contributions insofar as binary outcomes are very common in political science and that the first-order directive to empirical methodologists and researchers is, Hippocrates-like, "First, do not lie" with statistics about how much information you have. Either of these two methods makes feasibly expedient the accounting of time dependence in binary outcomes that had previously been rather intractably challenging methodologically. Directly autoregressive specification of temporal dependence would afford fuller interpretation of time-dynamic processes in binary outcomes, but these more directly modeled attacks on dynamics in binary outcomes would have to await methodological and computational advances that appear in this collection under the topic of spatiotemporal binary-outcome models. In "Time Is of the Essence: Event History Models in Political Science," Box-Steffensmeier and Jones overview event-history models in which the time and timing of events, like the end of parliamentary governments or of wars or the timing of position-taking by legislators or by candidates, is the outcome variable of interest. Event-history modeling is one of the most flexible and powerful tools in the empirical methodologist's toolbox for making temporal processes directly the object of study, and Box-Steffensmeier and Jones are the leading political methodologists in the event-history domain.

Section 5b in Volume IV moves from time series, generally of single units, to time-series cross-section or panel data in which several or many units are observed for multiple time periods.[12] Analysts typically distinguish subtypes of time-series cross-section datasets by numerous different overlapping considerations that are, unfortunately, not applied universally across literatures or authors in the same way. Some emphasize whether the units are the same from one period to the next – in which case, by one definition in usage, the data are panel or time-series cross-section

[12] Technically, all data are time-series cross-section data, since everything observed is observed in some place at some time. It simply happens that some datasets have only one unit on the cross-sectional dimension and some datasets have only one time-period on the temporal dimension.

data – or each time period is a new sample of cross-sectional units, in which case the data are "repeated cross-sections". Others differentiate between panel datasets, in which by this schema the number of units, *N*, is large (often a microlevel survey and so around or more than a thousand) and the number of time periods, *T*, is small (typically literally just a few), and time-series cross-section datasets in which the number of time periods is relatively large (at least around 20 or so) and typically the number of units is smaller (not more than a couple hundred, and usually in the 20-50 range, in most political-science applications). None of these distinctions makes any difference in terms of what sorts of theoretical and empirical models we might write, but, in practice, the sizes of the cross-sectional and temporal dimensions make a great deal of difference in what statistical procedures for estimating these models are advisable or even feasible. Summarizing in brief, longer-*T* datasets can afford richer temporally dynamic specifications, larger-*N* datasets are necessary for models that imply and estimation procedures that rely on variation across units.

Stimson's "Regression in Space and Time: A Statistical Essay" introduced empirical researchers and methodologists in political science to the complications that arise along with the enormous advantages from having data across multiple units and multiple time-periods. Ultimately, these complications are the increased strain on the maintained assumptions of constant coefficient parameters, of observations drawn from distributions all with the same constant stochastic-term variance, and of independence of those observations (conditional on the explanators, **x**). Especially cross-sectionally, we tend to suspect non-constant variance (i.e., heteroskedasticity across units[13]) and cross-unit dependence; and with respect to over-time: we already know that the existence of temporally autoregressive dependence is a certainty. Stimson follows the first mooted strategies for this complex terrain, which were to model the cross-unit

[13] Heteroskedasticity over time is also possible, and methods for modeling and/or properly accounting such cross-temporal nonconstant conditional variance (i.e., heteroskedasticity) also exist elsewhere in the literature.

heteroskedasticity and interdependence, and the temporal autocorrelation as features of the stochastic error-term, to be estimated by Feasible Generalized Least-Squares (FGLS) in the so-called *Parks* (1967) *Procedure*. In two enormously influential articles, Beck and Katz (1995, 1996) famously explained the shortcomings of Parks Procedure FGLS in samples with *T* not notably greater than N (i.e., *T* two times *N* or better). In a nutshell, the first stage of FGLS estimates the error-covariance structure, the "inverse of the square-root of which error variance-covariance" (loosely speaking) weights the data to obtain in a second stage *asymptotically* efficient coefficient and proper standard-error estimates. When *T* is not some multiple two or greater times larger than *N*, the number of parameters estimated in the error covariance specified by the Parks Procedure is high relative to the effective number of observations available for estimating them, and the asymptotic efficiency and quality standard-error estimates of FGLS do not materialize. In these seminal articles, Beck and Katz (1995, 1996) argue for using robust standard-error estimation, specifically what they call *Panel-Corrected Standard-Errors* or *PCSE*'s in samples of these dimensions instead since the asymptotic efficiency of Parks' FGLS will not materialize and the standard errors are even worse. This collection includes a later review article by Beck and Katz, "Modeling Dynamics in Time-Series–Cross-Section Political Economy Data," which reviews these issues and more in summarizing how "best practices" in time-series-cross-section data-analysis are understood to look today.

When the number of time periods in a time-series cross-section dataset are few, Hurwicz-Nickell (1950, 1981) bias – a small-sample bias of order $1/T$, in which temporal-dependencies expressed as coefficients on autoregressive time-lagged dependent variables are underestimated in time-series models with intercepts (Hurwicz 1950) and, by exactly the same set of issues, in panel-data models with fixed effects, i.e., unit-specific intercepts (Nickell 1981) – can be appreciable.

For instance, in panel datasets, where T is often not more than 4 or 5, the downward bias in estimated autoregressive parameters, ρ, is on the of -20% to -25%. The bias in estimated long-run multipliers, which are equal to $1/\rho$, are even more worrisome. The selection "Estimating Dynamic Panel Data Models in Political Science" from Wawro explains the instrumental-variable strategies (associated in econometrics with the work of Arellano and Bond 1991 among others) that leverage the panel structure of the dataset to obtain consistent estimates of these temporal dynamics. In time-series cross-section data analysis, then, we return to the thorny issue of "To FE (or RE) or Not To FE (or RE)" that was raised in the context of cross-unit heterogeneity in Section 3. If we include unit fixed-effects, they will overfit and jeopardize thereby our estimation of substantive effects, perhaps most especially manifesting we now learn in an underestimation of temporal dependence. If we omit unit fixed-effects, we remember from the fright-inducing Figure 1, we may be vulnerable to badly biasing our estimates of substantive effects, including perhaps especially we now note over-estimating temporal dependence. And, if we try random effects instead, we should have learned, those random-effects estimates are identified off the assumption that the unobserved cross-unit heterogeneity is uncorrelated with the observed heterogeneity (i.e., included regressors), and so we simply assume there would be no bias from any unmodeled cross-unit heterogeneity. Brandon Bartels, in a selection entitled "Beyond 'Fixed versus Random Effects': A Framework for Improving Substantive and Statistical Analysis of Panel, Time-Series Cross-Sectional, and Multilevel Data," tries to offer a strategic framework for navigating between this *Scylla* and *Charybdis* of random versus fixed effects specifically in dynamic models.

Section 5c turns our attention toward cross-unit, i.e., spatial interdependence, and spatial and spatiotemporal dynamics. In "Empirical Models of Spatial Inter-Dependence," Franzese and Hays emphasize first the likely ubiquity of cross-unit interdependence, i.e., that the outcomes in some

units depend on outcomes in other units, in *social* science. They then show that failing to model adequately this spatial interdependence will bias coefficient estimates of other substantive, non-spatial, explanatory factors. Worse, failing to recognize the spatial (cross-unit) interdependence of social-science data neglects the dynamic nature of *effects* – i.e., of *responses* in outcomes, **y**, to hypothetical changes in **X** – implied by the spatial interdependence (by completely omitting these dynamics). Spatial interdependence gives rise to spatial dynamics, which means the effects of any change in some *x* in any unit *i* affects not only the outcome in *i* but also the outcomes in other units *j* to which *i* is connected, and, from there, to the units with which *j* is interdependent, including, as space is omnidirectional, feeding back into *i*, and from there back out to the original *j*'s, and so on and so forth, in feedback ripples expanding outward and bouncing back inward. Franzese and Hays then provide overview of some of the statistical methods for properly estimating empirical models with spatial interdependence (and temporal dependence), namely: spatial-autoregressive models, and then explain appropriate methods for interpreting the spatially and spatiotemporally dynamic implications of such spatial-autoregressive models.

Spatial-statistical analysis and network analysis are closely related, and overlap in some ways, although they are not identical topics. The cross-unit spatial dependencies and the resulting spatially dynamic effects that Franzese and Hays emphasize are identical with one sort of *network dependence* and *network effects* to which network analysts refer. Indeed, the spatial-weights matrix of relative connectivity between units from spatial analysis, **W**, *is* **the network** of connections (a.k.a., *edges*, *ties*) between units (a.k.a., *nodes*) from network analysis. One sort of *network effect* of which network analysts speak is exactly that which spatial-lag autoregressive models specify: the effect on the outcome or behavior of unit *i* of the outcomes or behaviors of units *j* (spatial-lag **y** model) and/or other characteristics of units *j* (spatial-lag **X** model). *Network effects*, however,

may include also a second and/or third sort of effects. Second, effects on units (nodes) of the network structure itself – e.g., networks dense with ties, or with ties in particular configurations like hub-and-spike, may induce different unit behaviors than networks sparse with ties, or in other configurations like rings or chains. And third, effects on units/nodes of their position within or relation to the broader network structure – e.g., nodes with many ties may behave differently than nodes with few, or nodes in lynchpin connecting positions between clusters of connected other nodes may behave differently than others by virtue of that position (a strategically advantageous one in many substantive contexts). Furthermore, the spatial-econometric approaches that the selection from Franzese and Hays cover emphasize *contagion* processes, generally taking the connections between units as given exogenously, and stress *Galton's problem* of distinguishing contagion processes from *correlated exposure* to exogenous environmental conditions as the source of, or mechanism underlying, diffusion or cross-unit correlation; network analytic approaches, contrarily, tend to emphasize the formation of the network, i.e., of these connections between units as the outcome of interest to be explained, and stress the challenge of distinguishing this *network-tie selection/network formation* from *contagion* as the central empirical challenge.[14] The selection from Ward, Stovel, and Sacks, "Network Analysis and Political Science," gives very helpful introductory survey of network analysis and network-analytic methods for political-science empirical researchers and methodologists. Cranmer and Desmarais' "Inferential Network Analysis with Exponential Random Graph Models" gives closer methodological coverage of network-analytic methods, with an approach that emphasizes substantive and theoretical specification and

[14] Notice, too, that each approach tends to neglect, relatively at least, the third process from its perspective. Spatial econometrics generally ignores network-selection by assuming tie formation given exogenously; network analysis, conversely, in focusing on selection versus contagion, as in the classic case of whether smoking is contagious among friends or that smokers and nonsmokers are more likely to become and remain friends, tend to relatively downplay the possibility that some exogenous other factors, like certain values of individuals' sociodemographics may make them both more likely to smoke or not to smoke and to become and stay or not to become and stay friends.

interpretation of the network models in a way that very well befits the orientation of this collection. The selection by Franzese, Hays, and Cook, lastly in this section, tackle squarely the computational and methodological challenges of estimating and interpreting empirical models of binary outcomes that directly specify the simultaneity of spatial-, temporal-, and spatiotemporal-autoregressive processes.

Simultaneous cross-unit dependence connects the topic of Section 5, dynamics, to the topic of Section 6, endogeneity. In terms of the issues raised for statistical inference, simultaneous cross-unit dependence – y_i causes y_j simultaneously with y_j causing y_i – is isomorphic with the textbook endogeneity of x_i causing y_i simultaneously with y_i causing x_i. Both can be represented as systems of causal equations. From the time scholars first coined the very old adage that *correlation does not imply causality*, empirical researchers have known that endogeneity renders estimates of *empirical* relationships potentially misleading as estimates of *causal* relationships. This sharp segregation of empirical relationships from causal relationships hints at a different *fundamental problem* of empirical causal inference: namely, causality is fundamentally a theoretical property, not an empirical one. The logical fact that causality is a theoretical property shows also how we might gain empirical leverage on the theoretical entity of causal relationships. As a simple example illustrates, each observation of a pair of x and y when x may cause y and y cause x simultaneously, each *one* paired observation of an x and y, brings with it *two* parameters to estimate, the partial relationship of x to y and the partial relationship of y to x:

$$\left.\begin{array}{l} y = \beta_{x \to y} x \\ x = \beta_{y \to x} y \end{array}\right\} \Rightarrow y = \beta_{x \to y}\beta_{y \to x} y \Rightarrow \beta_{x \to y}\beta_{y \to x} = 1 \qquad (1).$$
$$\Rightarrow \textbf{any } \beta_{x \to y} = \beta_{y \to x}^{-1} \text{ are solutions.}$$

To get empirical estimates of causal relationships, therefore, we must somehow add some *extra-empirical information*; we must bring more information to the analysis beyond the empirical

information we can find from observation, be that observation of things in the "real world" or of things in laboratories. The empirical strate*gies* for causal inference covered in Section 6 ***all*** do exactly this: they bring some extra-empirical information, some imposition as known fact, rather than unknowns to be estimated, of some aspects of the analytic variables' relationships.

Experimental methods, e.g., rely on knowledge that the researcher *controlled* application or not of the value of treatment x that observation i received and so the outcome of interest, y, being measured could not cause the values of x. Randomization of that controlled treatment, then suffices, in sufficient sample-sizes, if the randomization was effective, to ensure that the controlled values of x that were assigned also could not correlate with other potential causes of y, even if these other potential causes are unobserved or even as yet unconceived. This is truly an enormous pair of advantages in ensuring that, if one finds a relationship between x and y, *the relationship must contain a causal process*. The careful phrasing here is intentional: we know there's some causal process connecting x to y if we find an (well-designed and -conducted) experimental association; we of course do not know that the causal process is the one the researcher thinks it is. In social science, the researcher's experiment and treatment, for instance, are typically very small parts of the much larger games the subjects are playing. (Students in classrooms are playing "18-22 year-olds in college," for example; not the professor's ultimatum game.) This is sometimes called the problem of *external validity of the treatment*: the treatment as felt by the subject may not correspond to the treatment as conceived by the experimenter. Another well-known converse external-validity cost to the great internal-validity advantage of experimentation in establishing causal-effect existence is that the experimental subjects may not be representative of the population to which we want to apply the experimental result. (Again, students in college classrooms may not be a representative sample of voters, never mind of policymakers…) This is *external validity of*

the experimental sample concern. A less well-appreciated form of external-validity concern with experimentation might be labeled the *external validity of the context* concern, and this one gets back to the simultaneous-equations representation of multidirectional causality in equation (1) and the dynamics of Section 5. In the presence of feedback, such as the cross-unit feedback of spatial interdependence, or even just the forward dependence of temporal dependence, or the cross-variable interdependence of simultaneous causality like in equation (1), the experimental demonstration of the ***existence*** of some causal process connecting *x* to *y* is very poor answer to *what is the effect of x on y* because the experimental result does not and cannot estimate these dynamics or feedback in the response of *y* to *x*. *Cannot* because the experimental researcher controls *x*; therefore, the response in y which in the true simultaneous system would feedback to further movement in *x* is broken by the experimental control. This is another way of understanding the *external invalidity of context* inherent in experimental social science: we don't want to know the effect of randomly assigned U.N. mosquito-netting programs; we'd like to know what the effect of an actual U.N.-administered mosquito-netting program would be.

As was the case regarding the explanation of the weighty tradeoffs between controlling or omitting cross-unit heterogeneity by fixed or random effects, the point here of explaining the very real, and at least sometimes very severe, external-validity costs, and costs in terms of well-estimating causal dynamic responses with feedback, of empirical causal-inference strategies based in experimentalism – and the sometimes-argued notion that internal validity trumps external validity is nonsensical, or perhaps even exactly wrong: perfect determination of causality with zero applicability beyond the sample is at least as useless as an estimated association with zero causal content but universal applicability – the point of emphasizing these limitations is not that we should not do experimental social science, or that we should give up on causality in social science, but

rather that we understand and recognize the unavoidable necessity of bringing extra-empirical information to the analysis to obtain causal leverage and that the numerous different strategies for doing so, some more "structural" and some more "non-parametric", bring different strengths in these tradeoffs between causal-*effect inference* and causal-*response estimation*, and between internal validity and external validity (both of which are generally desirable, of course, the latter at least equally so as the former).

Section 6, therefore, reviews a wide gamut of empirical strategies for causal inference or causal estimation, starting in Section 6a with the classical instrumental-variables (IV) approach (a more-structural approach). A reader following these volumes and selections sequentially who needs an introduction to instrumental-variables estimation will want to jump ahead to read first the selection from Jackson that leads Section 6b to get excellent coverage of standard IV-methodology without embellishments or critical amendments. Bartels' "Instrumental and 'Quasi-Instrumental' Variables," which leads 6a, shows that the standard proof that IV estimation is consistent and asymptotically efficient, regardless of the strength of covariance between the instruments and the instrumented variables, as long as the covariance between the instruments and the true residual is exactly zero, i.e., if the instruments are perfectly exogenous – this selection from Bartels shows first that this result, of the consistency and asymptotic efficiency of IV regardless of the strength of the instruments, requires perfect exogeneity (and infinite sample). If the instruments are to any degree, however tiny, imperfect, if their residual correlation is not perfectly zero, then the asymptotic bias and inefficiency of the IV estimator grows with the ratio of the instruments imperfectness (correlation with ε) to its strength (correlation with the instrumented variables). In other words, in practical, applied terms, the quality of IV estimates deteriorates proportionately with the ratio of the instruments' imperfection to their strength. Bartels' piece then underscores an

advantage of Bayesian estimation in this context that one could easily apply the "exogeneity of the instruments assumption" as a matter of continuous degree in the Bayesian framework, whereas, strictly speaking, exogeneity/endogeneity is indivisibly dichotomous in classical approaches: instruments either are endogenous or are exogenous. In "Instrumental Variables Estimation in Political Science," Sovey and Green then give "a [skeptical] reader's guide" survey of IV usage in the field, and in "Model Specification in Instrumental-Variables Regression," Dunning highlights a hidden assumption in IV estimation (in all empirical causal strategies) that the causal relationship identified by the instrumentation (by the extra-empirical information imposed on the empirical analysis) is the same causal relationship that applies beyond that identification context. In IV, as the article shows, this means that "the instrumented part of endogenous regressor *x*" must have the same relationship to y as the "endogenous part of *x*" "instrumented away". If we use rainfall to instrument for economic growth in an empirical model trying to estimate the causal effect of economic growth on civil war, to use an example from the selection, we need that rainfall/drought-induced economic booms and busts have the same effect on civil conflict as does economic growth or recession of other provenance. This is an important point to keep in mind, and it applies in general to causal estimation, not just to IV. (In a perfect randomized controlled trial, for example, we must assume that treatment applied randomly has the same effect as would the nonrandomly applied treatment the response to which we actually want to know.)

A more-efficient means of estimating causal systems of endogenous equations, working from the same or similar spirit as IV estimation is *Full-Information Maximum-Likelihood (FIML)*. As the name suggests, the FIML strategy in a nutshell is to specify whole systems of endogenous equations and estimate them in some fashion simultaneously. From a perspective that would emphasize causal estimation of *dynamic* responses in outcomes of interest to *x*'s, and to each other,

inclusive of the cross-unit and cross-variable feedback that perhaps ubiquitously characterizes socio-politico-economic reality, FIML would be the gold standard and, in this sense, the structural end of a causal-estimation / causal-inference spectrum, the other end of which would be the non-parametric experimental analysis. Jackson's "Endogeneity and Structural Equation Estimation in Political Science" covers FIML from an instrumental-variables perspective. Hays and Kachi's "Interdependent Duration Models in Political Science" shows the statistical isomorphism of cross-unit and cross-variable simultaneity, and through this isomorphic lens shows how the multivariate change-in-variables theorem used in spatial analysis to yield the joint likelihood for the N (number of units) equations of a spatial cross-section can be applied as well to estimate the $N \times M$ (number of endogenous variables) simultaneous equations of a spatially (or spatiotemporally) dynamic system of endogenous equations. In single-equation instrumental-variables estimation, we impose, ideally from theoretical or substantive knowledge but extra-empirically in any case, some features from other equations to "tie down" some aspects of the equation with endogenous regressors to be estimated. In FIML, we specify the entire system (ideally, in as theoretically and substantively motivated fashion as possible) for joint-likelihood maximization. Reed's "A Unified Statistical Model of Conflict Onset and Escalation" provides another substantive example of application of this "whole-systems" approach, one that emphasizes temporal sequencing as well and so nicely transitions to a third strategy for causal estimation.

Another strategy that yields some potential for empirically estimated relationships to be give some causal interpretation is to rely on the time sequencing of cause and effect, i.e., cause before effect, being hopefully reflected in the time sequencing of measured movements in variables. Miller's "Temporal Order and Causal Inference" is the classical statement for political science empirical researchers and methodologists of what we need, at a general and abstract level, for this

statistical *post hoc ergo propter hoc* strategy to shed causal light. Vector Autoregression (VAR) (Granger 1969) takes a full-throated temporal-sequencing approach to estimating simultaneous relationships. In brief, simple terms, VAR regresses each endogenous variable on its lags and lags of all the other variables to conduct Granger-Causality tests, and to estimate the temporally dynamic responses of the endogenous variables to each other, called *impulse-response paths*, and to ascribe shares of variation in each variable's dynamic responses to shocks to each variable, called variance decompositions. Freeman, Williams, and Lin give authoritative coverage of vector autoregression (VAR and relatives like VEC, vector error-correction) for political science in "Vector Autoregression and the Study of Politics". Sattler, Brandt, and Freeman's "Democratic Accountability in Open Economies" is an application to the extremely interesting question of how globalization may be affecting democratic-government accountability for economic performance of a modern variant, Bayesian Structural VAR, in which the *structural* modifier indicates the imposition of, in addition to the temporal-ordering imposed by VAR, instrumental-variable type exclusion restrictions on which endogenous and exogenous variables enter which equation.

These first three approaches to the challenge of ubiquitous endogeneity emphasize causal estimation and perhaps prioritize external validity. Being more structural, and in particular yielding results that are estimates of (causal) *models*, they afford estimates of dynamic responses, inclusive of feedback in complex contexts. Starting from the other end, the approaches that emphasize causal inference, i.e., of internal validity in identifying existence of causal effects, the gold standard is the randomized controlled experiment. In Section 6d, McDermott surveys "Experimental Methods in Political Science" and Druckman, Green, Kuklinski, and Lupia celebrate "The Growth and Development of Experimental Research in Political Science." These reviews include laboratory, field, and survey experimentation in their purview. Laboratory experimentation in political science

presumably comes closest to being able to approximate the ideal of the randomized controlled trial, the lacks being perhaps minimal and related perhaps solely to the addition of the human element in the experimental subjects from the natural-science laboratory. Survey experimentation limits the sorts of treatments that can be applied (question wording or ordering and the like) and adds some distance from the treatment as applied to as conceived. (Wording can typically only "prime" certain considerations, like partisanship of a bill supporter for instance, to some respondents and not others, when what we may want to know is the response to actual sponsorship change not to more or different mention of it.) The outcomes measured in survey experiments are also often more distant in this way from the outcomes of interest: survey-response statement about policy preference, not any actual political action taken, like votes cast, for example. Field experimentation, finally, has much to argue for it, in that the experiment occurs closer to or exactly in the substantive context of interest. A certain amount of control and effective pureness of randomization is typically sacrificed in moving to the field from the laboratory, but perhaps not always over-much. (One cannot control what other get-out-the-vote stimuli voters may or may not have received than the experimenters' randomly sent postcards or phone calls, for example, nor whether recipients read/answer them. Nor is to what else subjects may be exposed randomized, but rather a matter of subject choice just as much as in observational data, perhaps even choice made in response to receipt or not of treatment.) These limitations, though real and non-negligible, do however come along with some steps toward the converse enormous advantages in blocking reverse-causality or *confounding* (i.e., omitted variables), even unobserved or as-yet unconceived *confounding* that perfect experimentation provides fully.

Sections 6e through 6g review methods for empirical causal-*inference* in "observational", i.e., non-experimental, studies. In "Matching as Nonparametric Preprocessing for Reducing Model

Dependence in Parametric Causal Inference," Ho, Imai, King, and Stuart emphasize a particular use to which to put matching methods, which are strategies based on control not (directly) by the partialing of multiple regression, but rather by a pre-estimation procedure of finding comparison pairs or sets of observations that "match" values or distributions of values on (observable) potential *confounds*. In multiple regression, one controls for linear relations of (observable) potential confounds, called control variables in that context; in matching, one controls any sort of relations of (observable) potential confounds since we compare observations that get our treatment to ones that do not but with both treated and controlled having effectively (as determined by the matching procedure) the same values on the confounds. The selection from Ho et al. then notes that one could therefore view matching as a way to control non-parametrically in observational studies. (Notice how, from a causal-*inference* view like this, "model dependence" is pejorative. From a causal-*estimation* view, estimation of models was essential, inherent to the aim of estimating dynamic causal responses inclusive of feedback.) In "Opiates for the Matches: Matching Methods for Causal Inference," Sekhon *critically* surveys matching methods for political science empirical researchers and methodologists from a causal-inference perspective. The fundamental weakness of matching methods as a tool for causal inference from this perspective is that, exactly as in regression analysis, matching can only control for observables. Unobserved heterogeneity or causal confounding, the popular core challenge from the microeconometric perspective emphasized in Section 4, remains unaddressed.

The last two methods of the collection, the Regression Discontinuity Design (RDD) and Difference-in-Difference (D-in-D) Methods, address most centrally this challenge of unobserved heterogeneity. Regression discontinuity uses the *as-if* randomized feature of natural experiments

(or, to the critic, the hoped-for as-if randomization[15]). If we have some observable random variable, called the index variable, above some threshold value of which index variable the treatment of interest is applied and below which critical value the treatment is not applied, then exactly at the threshold value, only the randomness of index variable determined whether a unit got the treatment or did not. So, right at that threshold on the index, it is as-if the treatment were randomly assigned. Near the threshold, mostly the random noise determined if the treatment is assigned. Therefore, one could perhaps use observations near the threshold to achieve equivalently nearly the experimental advantages of randomization, including the control for "unobserved confounds". The classic origins of the design are an exam, the score on which determines a scholarship awarded or not above or below some threshold. Comparing students who got the scholarship from those who did not just to either side of the threshold should effectively randomize assignment of scholarship and so lend strong internal validity to a causal interpretation of the difference in means across the threshold. In "A Regression Discontinuity Design Analysis of the Incumbency Advantage and Tenure in the U.S. House," Butler applies this approach to estimate incumbency advantage, which has been its most-common application in political science. Caughey and Sekhon give another application to that subject, sounding a cautionary note in "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942–2008," about the most-likely violation of the conditions under which RDD can support causal interpretation, which is "non-sorting". Essentially, observations near the threshold cannot be special, and in particular they cannot self-select or be selected to be near the threshold. Caughey and Sekhon find some warning signs in the (cynically unsurprising) correlation of the partisanship of local election officials and that of the winners of close elections.

[15] The critic's favorite joke about natural experiments may be revealing: like the Moral Majority, *natural experiments* are neither. –anonymous.

Finally, difference-in-difference designs use the fact that time-differencing observations on a unit will difference away any fixed unit characteristics. (This is the same fact deployed in the fixed-effect estimation of Section 4a.) If we observe two groups at times t_1 and t_0, one of which received the treatment of interest and another that did not in time t_1, then the difference between the units of the difference from t_1 and t_0 will be the causal effect from that beginning to end period for that treated group, under the assumption of *parallel trends*, which boils down to the two groups' outcomes trending the same way absent the treatment, which in turn means only the treatment differentiates the two groups' different experiences of t_0 versus t_1.[16] Donald and Lang's "Inference with Difference-in-Differences and Other Panel Data" covers D-in-D for this collection.

In spanning the range of political science empirical methodology, from its origins historically and substantively in, first, conceptualizing the domain of political science and the central challenges to empirical research in that domain, to modern emphases on sophisticated models for causal estimation and careful strategies for causal inference, two very different and both very important enterprises, this collection aims to provide a library for empirical political-science methodology. In doing so, it also seems to have highlighted the enormous progress of the subfield, from extremely productive and forward-moving efforts at its origins to continuing diversification of that fruitfulness and its ever-further advancement. If the next thirty years of political methodology can sustain the vibrancy of the first thirty even while advancing and diversifying, the future for empirical research in political science will be a very bright one.

[16] Or anything else that differentiates the two units' differences from t_0 to t_1 is either irrelevant to the outcome or uncorrelated with treatment (i.e., the usual conditions for an omitted variable not to bias included variables' estimates).

REFERENCES

Arellano, Manuel, and Stephen Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58(2): pp., 277-297.

Beck, Nathaniel, and Jonathan N. Katz. 1995. "What to Do (and Not to Do) with Time-Series Cross-Section Data," American Political Science Review 89(3): pp., 634-647.

Beck, Nathaniel, and Jonathan N. Katz. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models." Political Analysis 6(1): pp., 1-36.

Brady, Henry E., David Collier, and Janet M. Box-Steffensmeier. 2011. "Overview of Political Methodology: Post-Behavioral Movements and Trends," in *Oxford Handbook of Political Science*, Robert E. Goodwin, ed.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Clark, William Roberts, Matt Golder, and Sona Nadenichek Golder. 2013. *Principles of Comparative Politics*, 2nd ed. Los Angeles: Sage Publications / CQ Press.

Davidson, J. E., Hendry, D. F., Srba, F., & Yeo, S. 1978. "Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom." *The Economic Journal* 88, pp. 661-692.

Franzese, Robert J., Jr. 1999. "Partially Independent Central Banks, Politically Responsive Governments, and Inflation," American Journal of Political Science 43(3): pp. 681-706.

Granger, Clive W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37(3): pp., 424-38.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): pp. 945-960.

Hurwicz, Leonid. 1950. "Least-Squares Bias in Time Series," in *Statistical Inference in Dynamic Economic Models*, Tjalling C. Koopmans, ed. (New York: Wiley).

Jackson, John E. 2012. "On the Origins of the Society [We're Not Lost, But How Did We Get Here]," *The Political Methodologist* 19(2): pp. 2-7.

Kelvin, Lord (Sir William Thomson, Baron Kelvin of Largs). 1883. *Popular Lectures and Addresses*, "Electrical Units of Measurement" [1883-05-03].

King, Gary. 1991. "On Political Methodology," *Political Analysis*, Vol. 2: pp. 1-30.

Nickell Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49(6): pp., 1417-1426

Parks, Richard W. 1967. "Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated." *Journal of the American Statistical Association* 62(318): 500-509.

Troeger, Vera. (n.d.) Lecture Notes: "Analysis of Panel and Time-Series Cross-Section Data," Essex University. Shared by personal correspondence.

Wooldridge, Jeffrey. 2002. *Econometric Analysis of Cross-Section and Panel Data*. Cambridge, MA: MIT Press.