PART V

Quantitative-Empirical Methods

Econometric Modeling: From Measurement, Prediction, and Causal Inference to Causal-Response Estimation

Robert Franzese

One can identify four modes, or purposes, of empirical analyses of positive[1] political science and international relations: *measurement* and description, *testing* (of causal theory or ‘effects’), *prediction* or forecasting, and *estimation* (of causal models and causal responses or effects).[2] These alternative aims or ends one might have in empirical analyses will place emphasis on different methodological challenges and properties over others and so demand different methodological approaches and tools. Econometric modeling[3] is an approach and set of tools that can be useful toward all four ends, but it plays an especially crucial role in the last: causal-response estimation. *Causal responses*, as opposed to *treatment ‘effects’*, refer to how some outcomes of interest (dependent variables) respond to inputs of interest (independent variables or treatments). As such, causal responses are inclusive of the contextual conditioning and effect heterogeneity, of the temporal, spatial, and spatiotemporal dynamics, and of the causal-simultaneity feedback that *treatment ‘effects’* purposefully exclude (in order to cleanly identify tests for causal-effect *existence*), and these heterogeneities, dynamics, and feedbacks cannot be estimated without modeling of the theoretical and substantive structure. Indeed, the specification and estimation of the empirical model, far from being an unfortunate unavoidable limitation, is, from that perspective and for those aims, the very goal of the exercise. The purpose of the analysis and the aim of the theoretically structured model is for its estimates to provide a ‘useful empirical summary’ of the actual substantive processes under study.

Measurement follows (as directly as possible) on *operationalization* – the translation from theoretical concepts, *X* and *Y*, to observable empirical indicators of those concepts (see Munck et al., Chapter 19, this *Handbook*) – to assign quantitative values gauging the extent or presence of those indicators in some unit of observation. As an end-goal of empirical analysis, measurement is distinct from the other modes in its purely descriptive aims:

scholars might conduct empirical analyses to offer measures of – i.e., to describe empirically – democracy (e.g., Coppedge et al., 2008; Treier and Jackman, 2008), the ideological placement of parties (Poole, 2019), and socioeconomic cleavage-structures (Selway, 2011), to name just three. Measurement is also distinct in being a fundamental prerequisite of any of the other modes of empirical analysis: causal inference, causal estimation, or prediction/forecasting. Econometric modeling is central to some measurement analyses (for examples, see Fariss et al., Chapter 20, this *Handbook*; Leemann and Wasserfallen, Chapter 21, this *Handbook*; and Treier, Chapter 48, this *Handbook*), but the focus of this overview remains the distinction between causal-effect *testing* and causal-effect *estimation* and the essential role of econometric modeling in the latter.

Regarding testing of causal theory and 'effects', the aim of the analysis is to evaluate empirically some causal-theoretical claim, i.e., to assess *whether* some posited causal relationship or causal effect *exists* empirically. Because the analyst's central purpose is to *test* a particular theory, ideally as little as possible from beyond that theory will be brought into the empirical assessment, so as to isolate the 'empirical existence proof' of the hypothesized causal effect. An empirical implication is derived from the theoretical argument that some $X \Rightarrow Y$, and the empirical analysis aims to evaluate this argument, this causal proposition, this hypothesis; in other words, the researcher wants to verify empirically that d$x \rightarrow$dy,[4] which entails (a) demonstrating that dx associates with dy empirically and (b) substantiating that the causal arrow goes from X to Y in the expressed direction. Notice that the adverb *empirically* applies only to component (a), empirical association of x and y; it intentionally does not apply to establishing causality because *causality* is a theoretical and not an empirical concept. Thus, the validity of 'empirical tests of causal "effects"' rests on the strength of the empirical association and, separately, on the strength of the arguments establishing that the causal arrow generating that empirical association goes in the theorized direction from X to Y. That is why the gold-standard ideal for *causal inference*[5] (see Bowers and Leavitt, Chapter 41, this *Handbook*) is the randomized controlled trial (RCT).

The potential-outcomes framework (POF)[6] proposes as the estimand for causal inference, i.e., for testing the existence of a causal effect of X on Y:

$$\text{Causal Effect} = Y_{it}(X=1) - Y_{it}(X=0) \qquad (1)$$

The *fundamental problem of causal inference* arises immediately: for a single observation on unit i at time t, denoted subscript it, the treatment or causal impetus, X, either is present (X = 1) or is not (X = 0). The counterfactual cannot be observed. Empirical designs for causal inference typically then proceed to establish conditions under which the difference in the empirical sample-means of y under x = 1 vs under x = 0 can be taken as an estimate of (1).[7] In the POF, this involves designing an analysis in which the comparison treatment (X = 1) and control (X = 0) groups are identical in all ways except treatment status (x value) and, potentially, the outcome (y value). The association of x and y, d$x \rightarrow$dy, is measured or estimated, and it can be understood as indicating the causal relationship d$x \Rightarrow$dy if two alternative causal-relationship possibilities can be ruled out as having instead generated that association: (a) that $Y \Rightarrow X$ (i.e., *endogeneity*, for instance by *simultaneity* or reverse causality) and (b) that some $Z \Rightarrow Y$ and $Z \leftrightarrow X$ (*spuriousness*).[8] Now we can see why the RCT is the gold standard for causal inference. *Experimental control* (of dx) assures that movements in x could not have been caused by y; the analysts know y did not move x because the analysts themselves moved or manipulated x.[9] *Experimental randomization* in which unit-times receive dx rules out spuriousness because if the values of dx are successfully independently randomized across a very large number of observational units, then dx will be unassociated with any alternative causal-factor Z – observed or unobserved (or even unimagined!) – by definition of *independent*.

The validity of the estimated empirical associations as test statistics for causal inference rests on the strength of the argument that the causal arrow underlying those associations runs as theoretically postulated. In some rare situations, $Y \Rightarrow X$ may be ruled out *a priori*: e.g., few Y could logically possibly cause race or gender X, but experimentation – i.e., *successful* control and randomization (and large samples[10]) – usually offers the strongest possible argument. In these cases, so-called *nonparametric causal inference* – i.e., causal inference that does not rely upon a pre-specified structural model (no model beyond the additive and separable treatment effects inherent in the difference-in-means definition of *causal* '*effect*' in (1)) – may be feasible. Even here, though, the validity of the causal-effect interpretation of any empirical estimate of (1) requires that Stable Unit Treatment Value Assumption (SUTVA) hold. The SUTVA conditions can be understood as the conditions under which the empirical dx is validly as-if experimental, i.e., controlled and randomized, and they amount practically to the following:

> The probability of one unit receiving treatment, the *homogenous* magnitude of the treatment, and the *homogenous* effect of treatment are independent of each other and of any other unit(s) receiving treatment, the sizes of treatments in those others, or effects of treatments in those others.

As one of POF's founding protagonists suggests, 'The two most common ways in which SUTVA can be violated [are] when (a) there are versions of each treatment varying in effectiveness or (b) there exists interference between units' (Rubin, 1990: 282). If SUTVA is violated, for instance by treatment –'effect' heterogeneity or conditionality and/or by spillovers or contagion across units i and time periods t – and such heterogeneity and interdependence are ubiquitous in socio-politico-economics – empirical estimates of (1) by the sample-mean difference of y given $x = 0$ or $x = 1$

$$\widehat{E(y \mid x=1)} - \widehat{E(y \mid x=0)} \qquad (2)$$

will not be valid or, speaking more precisely and generally, will be *inadequate* estimates of the causal effect of X on Y, or the expected empirical response of Y, dy, to an exogenous movement in X, i.e., treatment, dx. In general, reclaiming valid causal-effect (existence) inference, and *a fortiori* any hope to claim valid causal-effect (response) *estimation*, will require econometric modeling beyond that of estimating (1) by its empirical analog (2).

In sum to this point, for purposes of establishing causality and non-spuriousness in the experimental sample, i.e., for *internal validity*, the RCT is indeed the gold-standard ideal (see Morton and Vásquez-Cortés, Chapter 51, this *Handbook*). Of course, all scholars accept that practical and ethical considerations constrain what can or should be experimentally manipulated in the purview of social science. Even beyond those feasibility limitations, however, except for purely descriptive purposes of determining what was true in the observed experimental unit-times, analysts are more concerned with inferring from that experimental sample to what would be true in new data, outside that observed sample, i.e., with *external validity*.[11] And for that purpose, three significant challenges of *representativeness* – of the experimental sample to the intended population of inference, of the experimental treatment to the causal variable of interest, and of the experimental context to that of the socio-politico-economic outcomes of interest[12] – hinder what can usefully be inferred from these high-internal-validity RCT studies to target populations of interest. As a practical, empirical, applied matter (see also note 11), it seems as though external validity dominates internal validity in mean-squared-error terms, where potentially biased observational studies in proper context generally yield smaller mean-squared errors than do unbiased experimental studies conducted in *necessarily*[13] incorrect contexts (Pritchett and Sandefur, 2015). Moreover, even an ideal RCT is generally mute on other potential causes and says little about the magnitude of the treatment effects from the study relative to other effects[14] and relative to variation in the outcome variable of interest in the actual (external) context of interest.

> Experiment[s] will have nothing whatsoever to say about other causes. What it will do, and do well, is to determine ***whether*** […treatment…] had a positive or negative effect, or none at all. (Kellstedt and Whitten, 2009: 70; emphasis added)

Regarding external-validity concerns but staying within this *causal inference* or *testing-for-causal-effects* mode of empirical analysis, there has been much advancement in extra-laboratorial field- and survey-experimental research designs (see Sinclair et al., Chapter 52, this *Handbook*) and observational-study research designs (see Bowers and Leavitt, Chapter 41, this *Handbook*; Nielsen, Chapter 42, this *Handbook*; Keele, Chapter 43, this *Handbook*; Cattaneo et al., Chapter 44, this *Handbook*) to yield pseudo-experimental conditions for this causal 'effect' as defined in the POF (equation (1)). Treatment uptake by subjects is necessarily less strongly controlled in the field than in the lab, therefore, relative to the RCT laboratory experiment, field-experimental studies essentially trade some loss of purity in control and randomization for some enhancement of representativeness, perhaps of all three sorts (see note 12). Survey experiments, somewhat analogously, buy enhanced representativeness of subjects, given a scientific survey-design appropriate to the intended population, at the cost of representativeness of the treatment – mention or emphasis in a survey question, question ordering etc., are generally quite unlike the conceptual cause of interest in the theory – and of the context – answering a survey is usually very unlike the context to which the results are intended to be inferred.

For these reasons, social scientists sometimes must, and often *choose*, to work with observational data, especially for purposes of inference beyond sample, i.e., of inferring a causal relationship to exist in some target population of interest, and not only that a cause operated in some (specific, observed, and past) experimental sample. As shall be demonstrated below, the move to econometric modeling of observational data becomes especially judicious as the aim of the analysis moves beyond establishing that some causal effect exists (*causal inference*), to estimating causal responses, d*y*/d*x* (*causal estimation*). With these moves beyond internal causal inference to external causal inference and, especially, further beyond to causal estimation, empirical analyses in social-science observational studies confront not one, but at least four, fundamental challenges; namely, in socio-politico-economic reality (Franzese, 2007):

1. *Multicausality*: just about everything matters;
2. *Causal heterogeneity and context conditionality*: how everything that matters varies – how everything matters depends on just about everything else, i.e., on context;
3. *Dynamic causality*: just about everything is temporally, spatially, or spatiotemporally dynamic, not static;
4. *Omnicausality*: just about everything causes just about everything else.

Further exacerbating these four challenges is a fifth (or zero-th) challenge, which is that even with the enormous quantities of data now obtainable from internet, social-media, satellite/geospatial, and other big-data sources (Nyhuis, Chapter 22, this *Handbook*; Barberá and Steinert-Threlkeld, Chapter 23, this *Handbook*; Darmofal, Chapter 24, this *Handbook*), observational empirical analysts often find relatively little useful empirical variation with which to surmount these hurdles, even in those oceans of data. In the first instance, this is where econometric modeling becomes essential: given heterogeneity, dynamics, or simultaneity, without some structural model to reduce the parameterization of the problem, the number of quantities to estimate necessarily grows faster than the number of observations with which to estimate them.

Consider the following representation of empirical relations for one outcome of interest:

$$\begin{gathered} y_{it} = f_{it}\left(\mathbf{x}_{js}, \boldsymbol{\beta}_{js}, \varepsilon_{js}\right);\ \ \varepsilon \sim (\mathbf{0}, \Sigma_{it}); \\ i, j = 1..N,\ t, s = 1..T,\ n = NT \end{gathered} \tag{3}$$

In this perhaps most-general possible model,[15] there are k parameters, $\boldsymbol{\beta}_{it}$, linking right-hand-side variables, $\mathbf{x}_{js}$,[16] to the outcome of interest, y_{it}, plus a mean-zero stochastic component, □, characterized by an $n \times n$ variance–covariance matrix, itself possessing $\frac{1}{2}n^2 + \frac{1}{2}n$ (which is greater than n) parameters. In total, there are generally $k + \frac{1}{2}n^2 + \frac{1}{2}n$ parameters to estimate *per function, per observation*. This number of quantities to estimate grows exponentially faster than does the number of quantities observed (a $k + 1$ vector of $y, \mathbf{x}_{it}$). Thus, without some extremely strong structural-modeling assumptions, there could be no empirical analysis at all. From this perspective, we see that so-called nonparametric causal inference POF approaches applying (2) are actually highly structurally modeled: (a) empirical relations are assumed constant across all observations (within a bin if some heterogeneous effects are allowed) so there is only one function, f, to estimate; (b) random components are assumed orthogonal and homogenous across observations (or assumed zero with deterministic relationships such that the only randomness enters through the experimental manipulation); and (c) the parameters, $\boldsymbol{\beta}$, are also assumed constant across all observations, *it* (within a bin).[17] Notice that these are essentially identical to the assumptions of classical regression analysis. Indeed, the typical empirical model in POF-based studies, (2), is in many ways much more restrictive than the typical empirical model in regression-based studies: additive, separable, homogenous treatment effects (usually of a homogenous treatment: $X = 1$ or $X = 0$). The assumptions are the same because the logical necessity of some (radical) parameter reduction is the same. Therefore, the *arguments* or claims made about the research design, and not the estimation model, are the basis for POF causal-inference studies' claims to have ruled out spuriousness and simultaneity (again reflecting that causality is theoretical, not empirical).

Much of the econometric modeling deployed in the service of causal-inference studies focuses on ruling out spuriousness, i.e., some Z related to X actually causes Y: $Z \leftrightarrow X$ and $Z \Rightarrow Y$. *Matching*-based inference (see Nielsen, Chapter 42, this *Handbook*), for example, leverages the idea that, if the researcher can observe and measure all relevant $\mathbf{z}$, then comparing $y|x = 1$ to $y|x = 0$ for 'balanced' groups of data – meaning data for which the empirical *distributions* (sample means, variances, etc.) of all $\mathbf{z}$ are equal (or statistically indistinguishable) – yields a difference in means between treated and untreated observations that could not possibly be due to those $\mathbf{z}$. Note that matching, unlike the RCT, cannot control in this way for unobserved $\mathbf{z}$. Notice also that matching control for $\mathbf{z}$ is exactly like regression control for $\mathbf{z}$, except that the former is much more robust. Multiple regression controls effects of $\mathbf{z}$ that manifest in the manner modeled (e.g., linear effects only in linear regression), whereas matching controls effects of $\mathbf{z}$ in any manner they may manifest.[18] One might thus say 'Matching control is regression control on steroids'. Finally, also like regression, matching *per se* offers absolutely no address of simultaneity; like regression, claim for the matching-based estimation of (2) to be causal rests entirely on the adequacy of the controls.

Another causal-inference econometric-modeling approach that is focused on eliminating the possibility of spuriousness is the difference-in-difference (DID) design (see Keele, Chapter 43, this *Handbook*), and, relatedly, the difference-in-difference-in-difference

(DIDID or 3D) design and fixed-effects (FE) designs. The key notion underlying DID econometric modeling is that differencing the data, $y_{it} - y_{it-1}$, removes any time-constant factor in y, including any unobserved (time-constant) $\mathbf{z}$. Observed $\mathbf{z}$, including time-varying $\mathbf{z}$, may be addressed by regression or matching control; scale-variation or other functional-form issues may likewise be addressed by (structural) econometric modeling. Empirical implementation of the DID design is very simple. One needs observations on two groups, both in a pre-treatment period in which the X of interest has not been applied and a post-treatment period in which X has been applied in one but not the other group. Regression analysis with an indicator for post-treatment period, an indicator for treatment status ($x = 1$ or $x = 0$), and the interaction of those two dummies yields a coefficient on the interaction of the (treatment) difference in (time) difference, i.e., of the causal 'effect' in (2), under the maintained assumptions. Like matching and regression (after all, DID is commonly implemented by estimating a simple dummy-variable-interaction regression model), DID *per se* offers no address to the possibility of endogeneity. Units may select the treatment because of values of y or expected values of y, for instance, invalidating the causal interpretation of the DID estimate.

The (Regression) Discontinuity Design ((R)DD) is also an econometric approach to causal inference (see Cattaneo et al., Chapter 44, this *Handbook*) but one focused on addressing simultaneity bias as well as *unobserved confounds*, a.k.a. omitted-variable bias. The DD capitalizes on situations in which a treatment, $x = 1$, is triggered – a discontinuous jump in the probability of treatment suffices – as an observed continuous *index variable*, v, crosses some threshold value, v_c. For instance, a candidate wins a plurality-election office when his/her vote share crosses the plurality threshold (e.g., Caughey and Sekhon, 2011) or a party's probability of entering parliamentary government jumps discontinuously upward when its seat share crosses the plurality threshold (Hays et al., 2019). Provided (a) there are no systematic differences at the threshold in variables, $\mathbf{z}$, other than the treatment variable (which can be evaluated empirically for observed $\mathbf{z}$), and (b) no endogeneity in which observations fall near to either side of the threshold (called *sorting*, in this context), then exactly at, or at least near the threshold, it is completely random, or mostly random, whether the observation receives treatment. The causal 'effect' as defined in (1) is thus identified at the threshold as if by RCT: any observed or unobserved $\mathbf{z}$ should be equal on either side very near the threshold (and this can be verified for observed $\mathbf{z}$), and treatment is 'applied' randomly. Of course, like the RCT it aims to mimic, the DD estimate lacks external validity of its estimated treatment effect for conditions unlike those at the threshold (e.g., for not-close elections, which greatly limits the applicability of DD estimates of US Congressional incumbency advantage, e.g.). With some additional assumptions, one can estimate (2) by an RDD regressing y on a flexible polynomial in index, v, an indicator for treatment status, $x = (0,1)$, and the interaction of the polynomial terms with x. The coefficient on the x is then the RDD-identified effect.

The more full-throated econometric-modeling approach to causal inference relies upon instrumental variables (IV) (see Carter and Dunning, Chapter 40, this *Handbook*; see also selection modeling in Böhmelt and Spilker, Chapter 37, this *Handbook*) and, more full-throated still, systems estimation (see, e.g., Jackson, 2008). The causal-identification strategy of instrumentation is well known: given a causal relation, $y = f(x\beta, \square)$, about which there may be concerns that $y \Rightarrow x$ as well, find some z that (a) covaries with x but (b) not with $\square$ – alternatively, with more substantive appeal, this variable z, called an *instrument*, needs (a) to relate to x but (b) not to y, except through that relationship to x – and then estimate the relationship of y with z instead by indirect least-squares (ILS), for example:

$$\begin{aligned}
&\left.\begin{aligned}(i)\ \ & y_{it} = \beta x_{it} + \gamma z_{it}^{y} + \varepsilon_{it}^{y}\\ (ii)\ & x_{it} = \alpha y_{it} + \delta z_{it}^{x} + \varepsilon_{it}^{x}\end{aligned}\right\}\Rightarrow y_{it}\\
&\qquad = \beta(\alpha y_{it} + \delta z_{it}^{x} + \varepsilon_{it}^{x}) + \gamma z_{it}^{y} + \varepsilon_{it}^{y}\\
&\qquad \Rightarrow y_{it} = \frac{\beta\delta}{1-\beta\alpha} z_{it}^{x} + \frac{\gamma}{1-\beta\alpha} z_{it}^{y}\\
&\qquad\qquad + \frac{1}{1-\beta\alpha}\varepsilon_{it}^{y} + \frac{\beta}{1-\beta\alpha}\varepsilon_{it}^{x}
\end{aligned}\tag{4}$$

The ILS coefficient-estimates may be solved to retrieve estimates of β and the other coefficients (once the residual correlation is accounted), but, for solely causal-inference *testing* purposes, the significance of the estimated coefficient on z_{it}^{x} may be evaluated directly. Note that instrumentation strategies, even in the causal-inference modality that aims to minimize modeling assumptions, must pre-specify enough about the causal process to at least offer a causal diagram (Pearl, 1995), if not a fully specified system of equations, to establish IV-identification conditions (a) and (b).

Two-stage least-squares (2SLS) is a convenient implementation of IV-estimation. 2SLS estimates equation (*i*) in (4), for example by regressing *x* on (any exogenous variables across the system as given by the model/graph and) *z*, or on **z** if more than one instrument is available (stage 1), and then regressing *y* on that fitted *x* (and any exogenous variables in equation (*i*)) (stage 2). The 2SLS procedure puts the fitted-*x* regressor in the same scale as the endogenous *x* so the 2SLS estimated coefficient on instrumented-*x* estimates directly, and if there are multiple instruments, the *regression* of stage 1 is the optimal procedure for projecting multidimensional information **z** to unidimensional *x* and *y*. Single-equation three-stage least-squares (3SLS), which is asymptotically equivalent to limited-information maximum-likelihood (LIML), gains efficiency relative to 2SLS by accounting the necessary non-sphericity in the stochastic component of the model seen in the compound error term in the last line of (4). For further efficiency gains, the full system of equations (*i*) and (*ii*) can be estimated jointly by multi-equation 3SLS or by (asymptotic equivalent) full-information maximum-likelihood (FIML).[19] Systems approaches also facilitate the incorporation of cross-equation substantive/theoretical knowledge into the estimation, such as. e.g., that some coefficients are equal, proportionate, or oppositely signed across equations.

Strategies for addressing simultaneity and the other challenges for applied empirical social science can perhaps be enumerated from most to least structural (excepting item 0[20]) thusly:

0 Time ('the poor man's exogeneity');
1 Full-system specification and estimation;
2 (Single-equation) instrumentation;
3 Matching;
4 Difference-in-difference;
5 Discontinuity designs;
6 Survey and field experimentation;
7 Laboratory experimentation.

For *causal-inference* testing purposes, the ordering also lists in generally increasing *credibility*, given their decreasing reliance on information beyond the theory to be tested and the empirical data. Indeed, given the tremendous advantages of controlled randomization against spuriousness and reverse causality, what could possibly argue for econometric modeling, or model-based estimation in general, over the RCT or nonparametric causal-inference strategies in general? At broadest, and in general, the answer is: external validity.

Firstly, prior to any external-validity concerns, note that for description, summarization, and measurement purposes, causality is simply irrelevant. Experimentation would be exceedingly cumbrous, and piles of nonparametric estimates – being necessarily unconnected by any formula – tend to offer only poor summary and poorer understanding.[21] The dominance of model-based approaches of diverse kinds for textual-data analysis, scaling, or classification (see Benoit, Chapter 26, this *Handbook*; Ergerod and Klemmensen, Chapter 27, this *Handbook*; Bouchat, Chapter 28, this *Handbook*), sentiment or network description (Curini and Fahey, Chapter 29, this *Handbook*; Calvo et al., Chapter 30, this *Handbook*), and latent-concept recovery (Fariss et al., Chapter 20, this *Handbook*; Leemann and Wasserfallen, Chapter 21, this *Handbook*; Treier, Chapter 48, this *Handbook*) is therefore natural and optimal. Rather than the RCT, the gold standard for measurement exercises is *usefulness* in conveying summary description or in subsequent analyses.

Likewise, for purely predictive and forecasting purposes, the gold standard is not recovery of some ideal-experimental results; it is (obviously) *out-of-sample prediction/forecast error* (see, e.g., Schrodt and Gerner, 2000; for a similar view but comparing prediction to explanation instead of causal inference, see Ward, 2016). Again, internal validity is irrelevant insofar as the aim is to predict the value of some $y_{js\neq it}$, full stop; external validity is the only relevant consideration for pure prediction (see note 11). Here, too, econometric model-based strategies dominate, but in this case it is perhaps due more to the inherent limitations of nonparametrics than those of causal inference. Nonparametric estimates, by construction, offer no connection from E($y|x = x_0$) to E($y|x = x_1$); consequently, as the possible values that a potentially large number of useful predictors, their interactions, and interdependencies may take grows, the number of nonparametric estimates needed expands at least exponentially (and possibly combinatorically).[22] The forecasting device must somehow dampen this meteoric proliferation of necessary estimands; the preferred methods, having proved most effective, i.e., performing best by the sole relevant criterion, *out-of-sample prediction/forecast error*, include sophisticated econometric modeling with Bayesian methods (see Park and Shin, Chapter 47, this *Handbook*; Gill and Heuberger, Chapter 50, this *Handbook*), particularly Bayesian model-selection and model-averaging (see Hollenbach and Montgomery, Chapter 49, this *Handbook*), Bayesian structural vector autoregression (Kilian and Lütkepohl, 2017), and machine-learning and artificial-intelligence methods (see Mikhaylov and Chatsiou, Chapter 55, this *Handbook*; Shoub and Olivella, Chapter 56, this *Handbook*).

As analysts' aims move beyond measurement, prediction, and *causal inference* or *testing* for (existence of) causal effects to *causal estimation* or *estimating causal responses*, the limitations of the experimental paradigm in the face of the four fundamental challenges for empirical research in social science become more pronounced. *Multicausality*, that there tends to be many relevant causes of effects, is least problematic, being addressed by control and, in many respects – as just discussed – ideally, by experimental control, at least for causal-inference purposes.[23] The limitations with respect to that first fundamental challenge relate to representativeness and external inference: comparing the treatment in the experimental sample and context to the intended treatments (causes) in their intended population and contexts. Effect heterogeneity raises more serious challenges. The structure of the Neyman–Holland–Rubin (NHR) causal model is of additive, constant, separable effects.[24] Effect heterogeneity or conditionality can, in principle (though see notes 22 and 24), be managed by binning observations with effects that are assumed homogenous within bins. However, to see how limiting this can be given socio-politico-economic reality, consider the simple case of the sigmoidal non-linearity implied in binary-choice and other binary-outcome contexts simply by the nature of probabilities or proportions. The essential substance of the matter dictates that for all binary outcomes, probabilities, or proportions,

$$\begin{gathered}\Pr(y=1)\equiv p(y)=f(x,\beta,\varepsilon)\\ f\text{-being sigmoidal with } 0\leq f\leq 1\end{gathered}\tag{5}$$

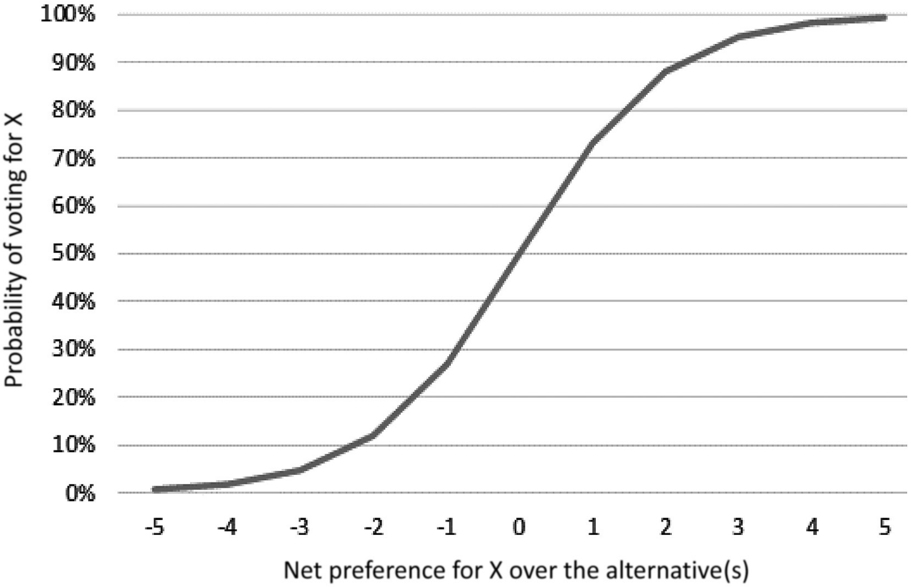


Figure 31.1 The logically necessarily sigmoidal relation p(y) = f(x)

A model of causal effects on probabilities that does not respect these first principles – the relationship between x and $p(y)$ tapers toward its 0 and 1 bounds, (because it surely would not kink at those bounds) and so is steeper in some manner in between such as in Figure 31.1 (the NHR with $\frac{dp}{dx} = c$ constant is one such non-respecting model) – is unlikely to yield very good *estimates* of those causal effects for external inference, especially for estimates traversing more curved portions of the *S*-curve and especially not beyond support. If nonlinearities like these are common, or more generally if effects in socio-politico-economic reality are typically heterogeneous and context-conditional as contended here, then the NHR causal model is a poor basis for *causal-effect estimation*, although it may remain a strong model for *causal inference* (see, e.g., Imai and Ratkovic, 2013; Egami and Imai, 2015).

The limitations of nonparametric causal inference in confronting causal heterogeneity and context conditionality are debilitating for causal-effect estimation, for which purpose econometric modeling is inescapably essential. Far from an unavoidable detraction, however, estimation of an econometric model reflecting the theory and substance of the context is the very goal of the causal-estimation exercise. In this *Handbook*, as examples, see Fukumoto, Chapter 35, for a discussion of appropriate modeling of duration or survival contexts, and see Steenbergen, Chapter 36, for effective empirical-modeling strategies for parameter heterogeneity in multilevel/hierarchical contexts. For a quick illustration of how substantively theoretically specified econometric-model estimation can yield interesting and useful empirical science beyond proofs of causal existence, consider the implications of principle-agent/multi-actor bargaining for policy outcomes (Franzese, 2003, 2010). Equilibrium policy-outcomes in principle-agent and other shared-policy-control situations are some convex combinations of the two (or more) actors' optimal policies, e.g., a linear-weighted average, such as:

$$\begin{aligned} y = & \underbrace{c_p(\mathbf{z}) \times f(\mathbf{x}_p)}_{\text{principal cntrl} \times \text{p action}} + \underbrace{[1 - c_p(\mathbf{z})] \times g(\mathbf{x}_a)}_{\text{agent control} \times \text{agent action}} \\ & \Rightarrow \frac{dy}{dx \in \mathbf{x} \equiv (\mathbf{x}_p \cup \mathbf{x}_a)} = h(c_p(\mathbf{z})) \\ & \text{and } \Rightarrow \frac{dy}{dz \in \mathbf{z}} = q(\mathbf{x}) \end{aligned} \qquad (6)$$

In words, the effect on the outcome of anything to which principal and agent (the bargainers) would respond differently depends on the degree to which each of the players is in control, which depends in a typical principal-agent model, e.g., on monitoring and enforcement conditions, represented in (6) by $c_p(\mathbf{z})$. Conversely, the effect of anything that shapes monitoring and enforcement costs and efficacy, $z \in \mathbf{z}$, depends on everything to which the players would respond (differently): $x \in \mathbf{X}$. By virtue of the shared influence of the bargainers over the outcome, the effect of any $x \in \mathbf{X}$ to which they would respond differently depends on all $\mathbf{z}$ in $c_p(\mathbf{z})$, the weight of each actor in determining the outcome, and, *vice versa*,[25] the effect of any $z \in \mathbf{z}$ that influences the actors' relative control depends on all $x \in \mathbf{X}$ to which the actors respond differently.

How can empirical researchers effectively estimate complexly context-conditional effects like these? One strategy is to impose the substantively known structure in the empirical model. Franzese (1999) estimates an empirical model like (6) to show how the anti-inflationary effects of central bank independence (CBI) – a situation of shared monetary-policy control, agent central bank and principal government – depend on political-economic conditions that would make governments more inflationary (bigger effect) or less inflationary (smaller effect). The convex-combinatorial form of (6) implies that only one additional parameter needs be estimated to capture all of these theoretically/substantively implied interactions; namely, this parameter is the factor of proportionality by which the central bank independence measure dampens inflation from the government's to the bank's preference as CBI increases. By a further nested pair of weighted averages, Franzese (2003) extends the central bank/government domestic-actors model of 1999 to the open and institutionalized economy, wherein exchange-rate pegs effectively delegate from these two domestic actors to the peg-currency policy, and infinitesimally small capital-open economies, which effectively constrain domestic policy to the global average. Notice from the estimation model and results in Figure 31.2 that the 'theory-informed' model requires just two more parameter estimates than the linear-additive model, which completely lacks interactions, and 50 fewer than the linear-interactive model requires to generate comparable interactivity. And yet the coefficient estimates on small capital-openness, *E*, on single- and multi-currency pegs, *SP* and *MP*, and on CBI, *C*, are easily interpretable as the proportionate constraint each of those measures places on the opposite actors in its convex combination (see model (14) in Figure 31.2). The graphs illustrate two of the (many) rich substantive insights yielded about the context-conditional amplitude of partisan inflation-cycles at the top-right and about the generally declining anti-inflationary bite of CBI since about the 1970s, coinciding with the acceleration of the postwar and current great globalization.[26]

Next, consider how temporal and spatial dynamics highlight the inadequacy of the NHR model to causal-response estimation. Notice that the NHR estimand (1) and its typical empirical estimate (2) yield a scalar estimate, a single number, as the causal 'effect' of *x* on *y*. In a temporally dynamic context, in contrast, taking the simplest example to illustrate:

$$\begin{aligned} &(i)\ y_t = \rho y_{t-1} + \beta x_t + \varepsilon_t \\ &(ii) \Rightarrow \frac{dy_t}{dx_t} = \beta \\ &(iii) \Rightarrow \underbrace{dy_{LT}}_{\substack{LRSS \\ \text{response}}} = \underbrace{\beta dx}_{\text{period 0}} + \underbrace{\rho\beta dx}_{\text{period 1}} + \underbrace{\rho^2 \beta dx}_{\text{period 2}} + \underbrace{\rho^3 \beta dx}_{\text{period 3}} + \ldots \\ &\qquad = \underbrace{\sum_{s=0}^{\infty} \rho^s \beta dx}_{\text{assuming } |\rho|<1 \Rightarrow} = \frac{1}{\underbrace{1-\rho}_{\text{LR multiplier}}} \times \beta \times \underbrace{dx}_{\substack{\text{perm.} \\ \text{shock}}} \end{aligned} \qquad (7)$$

$$E(\pi)=B_0+\beta_e E\cdot\beta_{a*}\pi_a+\left(1-\beta_e E\right)\cdot\left\{\begin{array}{l}\left[\left(\beta_{gp}GP+\beta_{ey}EY+\beta_{up}UP+\beta_{bc}BC+\beta_{aw}AW+\beta_{fs}FS+\beta_{te}TE+\beta_{\pi g}\pi_a\right)\right.\\ \left.\cdot\left(1-\beta_{c1}C\right)+\beta_{c1}C\cdot\beta_{c2}\right]\\ \cdot(1-\beta_{sp}SP-\beta_{mp}MP)+\beta_{sp}SP\cdot\beta_{a*}\pi_{sp}+\beta_{mp}MP\cdot\beta_{a*}\pi_{mp}\end{array}\right\} \tag{14}$$

Table 1 Alternative models of inflation in 21 OECD democracies, 1957–1990

		Linear-interactive model (13)								
Independent variable	*Linear-additive model (12)*	$C=1$, $E=1$, $P=1$	$C=1$, $E=1$, $P=0$	$C=1$, $E=0$, $P=1$	$C=1$, $E=0$, $P=0$	$C=0$, $E=1$, $P=1$	$C=0$, $E=1$, $P=0$	$C=0$, $E=0$, $P=1$	$C=0$, $E=0$, $P=0$	*Theory-informed model (14*
Intercept	+.80	+5.93								+.53
	(6.1)	(8.40)								(.30)
Lagged inflation	+.65	+.51								+.55
(π_{t-1})	(.05)	(.06)								(.05)
Twice-lagged	−.03	−.10								−.12
inflation (π_{t-2})	(.04)	(.04)								(.04)
Government partisanship	−.14	+*.39*	*−.09*	−3.37	−1.37	−.15	−.30	+1.82	−.39	−.60
(GP $\in \mathbf{X}_g$)	(.08)	*(.80)*	(1.29)	(1.31)	(8.16)	(.47)	(.97)	(.74)	(4.68)	(.30)
Postelection year	+.59	+.75	*−2.06*	+.50	*−.88*	*−2.31*	*+6.03*	+1.87	+3.81	+2.60
(EY $\in \mathbf{X}_g$)	(.30)	(.80)	*(2.31)*	(3.07)	*(14.67)*	*(1.56)*	*(3.46)*	(1.81)	(6.88)	(1.32)
Union power	+2.19	*−16.59*	+9.51	*−3.82*	*−2.46*	*+33.95*	+2.44	*−11.88*	*−3.32*	+16.2
(UP $\in \mathbf{X}_g$)	(.74)	*(6.43)*	(17.42)	*(13.91)*	*(59.24)*	*(7.64)*	(15.92)	*(13.56)*	*(37.49)*	(4.61)
Coordination of	−1.36	*+4.38*	*+11.27*	*+6.02*	*−39.11*	*−15.61*	−11.69	*+2.20*	*+9.27*	−10.7
bargaining (BC $\in \mathbf{X}_g$)	(.41)	*(3.50)*	(5.33)	*(4.91)*	*(30.32)*	*(3.97)*	(9.79)	*(3.86)*	*(23.64)*	(2.35)
Aggregate	+.13	−.76	−2.37	+1.94	*+13.70*	−.56	−.66	−2.24	−3.43	+1.18
wealth (AW $\in \mathbf{X}_g$)	(.71)	(1.15)	(1.51)	(1.43)	*(5.37)*	(1.10)	(1.38)	(1.91)	(2.35)	(.49)
Financial-sector size	−.15	−.86	*+2.00*	*+2.11*	*−11.13*	*+.55*	−1.64	−1.00	*+4.63*	−1.09
(FS $\in \mathbf{X}_g$)	(.10)	(.36)	*(.96)*	*(.79)*	*(4.61)*	*(.36)*	(1.26)	(.71)	*(3.90)*	(.30)
Trade exposure	−.04	*+31.74*	*−50.21*	*−54.49*	*+50.81*	*−37.33*	*+104.56*	*+48.70*	*−120.5*	−8.23
(TE$\in \mathbf{X}_g$)	(.99)	*(14.33)*	*(25.31)*	*(39.85)*	*(176.99)*	*(14.87)*	*(30.40)*	*(33.74)*	*(103.79)*	(4.92)
Inflation abroad	+.39	+.24	+.89	*−.07*	*−4.01*	+.89	+.18	+.98	*+2.65*	+.64
($\pi_a \in \mathbf{X}_g$)	(.07)	(.14)	(.52)	*(.59)*	*(3.94)*	(.31)	(.78)	(.33)	*(2.58)*	(.24)
Global-financial	+.29	—								+.44
exposure (*E*)	(.75)									(.14)
Single-currency	−.33	—								+1.04
(simple) peg (*SP*)	(.49)									(.05)
Multi-currency	−.37	—								+.22
(basket) peg (*MP*)	(.38)									(.12)
Peg or global inflation	—	—								+.59
(π_{sp}, π_{mp}, π_a)										(.07)
Central bank	−1.62	—								+1.03
independence (*C*)	(.68)									(.11)
Central bank		—								−.59
target ($\bar{\pi}_c$)										(1.18)
Obs. (°Free)	660 (645)	660 (593)								660 (643)
R^2 (S.E.R.)	.72 (2.48)	.75 (2.31)								.76 (2.30)

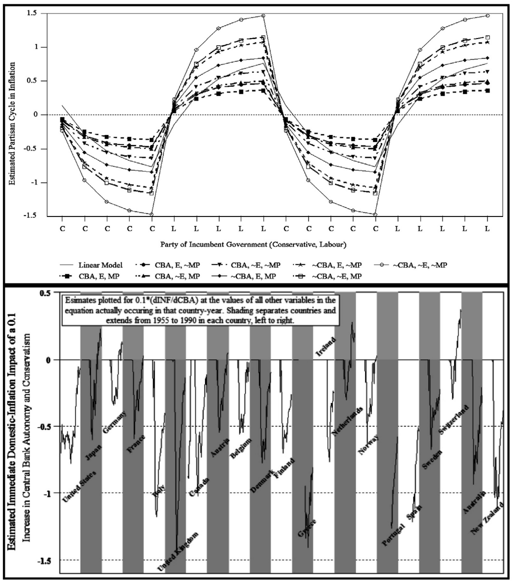


Figure 31.2 ‘Multiple Hands on the Wheel’ model of complex context-conditionality in monetary policymaking

A causal-inference study designed to test whether x causes y will estimate β, which the second line of (7) demonstrates is only the contemporaneous response, dy_t, to a shock (treatment), dx_t, in that same period. As the lines (iii) show, if the shock persists (and nothing else occurs), the subsequent period experiences an additional $\rho\beta$ and the period after that an additional $\rho^2\beta$, and so on. If the shock persists infinitely, the long-run steady-state (LRSS) response equals the long-run or temporal steady-state multiplier of $1/(1-\rho)$ times the initial response, β (for further development and discussion, see Linn and Webb, Chapter 32, this *Handbook*; for fullest textbook treatment, Hendry, 1995). The single scalar[27] in (2) is obviously an inadequate-answer to the question, ‘what is the effect of x on y?’, in the temporally dynamic context. In fact, the question is underspecified in the dynamic context: ‘the effect of (a movement in) x, *when?*, on (movements in) y, *when?*’[28]

Temporal dynamics matter greatly for substantive conclusions about causal-effect size. Consider, e.g., the many well designed causal-inference studies on the effects of voter-registration hurdles, which typically find ‘very small effects’ on turnout (e.g., Hershey, 2009, reviews), but these are impulses, β, not *effects*, d**y**/d**x**. The considerable evidence that voting is a habit slowly acquired over repeated elections (e.g., Gerber et al., 2003) implies that voter turnout evolves dynamically, as in (7), so the response of voter turnout to registration-easing legislation is not a snapshot-in-time scalar but a vector over time, and, with ρ being large, the long-run cumulative effects, $\beta/(1-\rho)$, are many times those previously estimated ‘very small’ causal *parameters*. Another illuminating example, from Franzese (2002), shows (in Figure 31.3) dynamic estimates from an econometric model of responses of public debt in developed democracies, counterfactually (a) to the actual OECD average real interest-rate (net growth) series 1954 to 1995 and (b) to hypothetical permanent a plus-one standard-deviation shock in real interest-rates proceeding indefinitely into the future, both

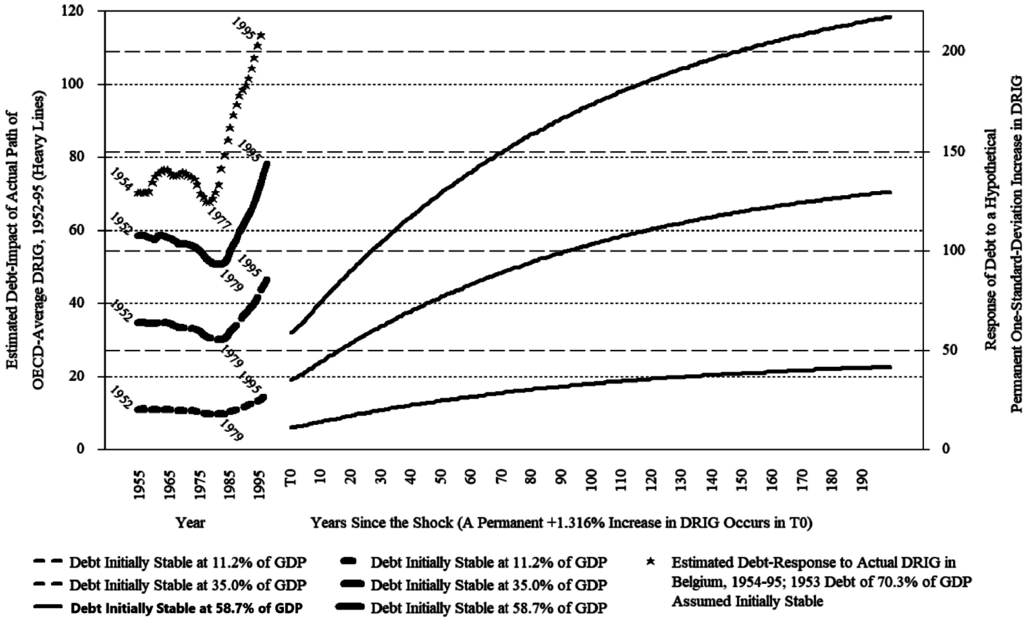


Figure 31.3 Substantive dynamic-effect estimates of real interest-rate net of growth impacts on public debt

Source: Franzese (2002).

starting from different initial debt-to-GDP ratios. Substantively, these dynamic estimates demonstrate that interest rates, i.e., monetary policy, could have enormous effects on the long-term accumulation of public debt and that, in fact, much of the post-1970s emergence of public-debt crises owed to that stagflationary era's adverse shocks to growth and unemployment, inducing deficits which were followed by tight monetary-policy that spiked interest rates on those newly accumulating debts.

The inadequacies of the NHR model and estimand are highlighted further and amplified in the time-series cross-section (TSCS) and spatially/spatiotemporally dynamic contexts (see Troeger, Chapter 33, this *Handbook*; Cook et al., Chapter 39, this *Handbook*; relatedly, for dyadic-data and network analyses, see Neumayer and Plümper, Chapter 38, this *Handbook*; Victor and Khwaja, Chapter 45, this *Handbook*; Schoeneman and Desmarais, Chapter 46, this *Handbook*). A very low-dimensional example, an n = 3-units cross-section with simultaneous first-order spatial-autoregressive interdependence (i.e., outcome contagion), suffices to demonstrate:

$$
\begin{aligned}
&(i)\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\
&= \rho \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\
&+ \beta \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \\
&(ii)\ \mathbf{y} = \rho\mathbf{W}\mathbf{y} + \beta\mathbf{x} + \varepsilon \\
&(iii) \Rightarrow \mathbf{y} = [\mathbf{I} - \rho\mathbf{W}]^{-1}[\beta\mathbf{x} + \varepsilon] \\
&(iv) \Rightarrow d\mathbf{y} = [\mathbf{I} - \rho\mathbf{W}]^{-1}\beta \times d\mathbf{x}
\end{aligned}
\tag*{$(8)^{29}$}
$$

Again, the NHR estimand (1) and its empirical estimate (2), which a well designed and conducted experiment would produce, correspond to β, which we can call the 'pre-dynamic *impulse*' from x to y, and not 'the effect of x on y', understood as how Y responds to movements in X, which is instead given by line (*iv*) of (8). Once again, the NHR estimand (1) and estimate (2) are in the wrong dimensionality, and the question is ill posed. Neither treatment nor effect are scalars: both are vectors/matrices and the effect statement is underspecified. In spatial/spatiotemporal contexts, the fully specified statement is 'the effect of movements in **x**, *where (and when)?* on movements in **y**, *where* (*and when*)*?*'. In fact, in this case, the *impulse* is not even observable: the response in unit 1 – e.g., to some dx in, say, unit 1 itself – begins with the *impulse*, $\beta \times dx$, but that *instantaneously* induces proportionate movements ρw_{1j} in units 2 and 3, which instantaneously induces proportionately smaller[30] movements $(\rho w_{1j})^2$ in units 1 and 3, and so on, reverberating through the units across space analogously to the temporal-dynamics case, but *omnidirectionally* and all *simultaneously*. Thus, the pre-dynamic impulse, β, never manifests observably at all, only its steady-state implications in (iii) and (*iv*) show empirically. Figure 31.4 maps the estimated responses in a spatiotemporal econometric model from Franzese and Hays (2006), regarding active labor-market (ALM) spending to a hypothetical 1€ (per unemployed worker) spending increase in Germany. The left panel shows the estimated response across all EU countries[31] in the time period contemporaneous to the shock (inclusive of that period's spatial feedback but exclusive of any time dynamics); the right panel shows the LRSS accumulated response, inclusive of all spatiotemporal feedback. The econometric model uncovered free-riding behavior, i.e., negative spatial interdependence; the characteristic oscillating pattern of negative autoregression is apparent in the right panel.

The case of spatial interdependence also underscores the radically limited scope for nonparametric causal inference, even with regard only to testing, in a socio-politico-economic reality characterized by omnidirectional causality (fundamental challenge number four). Notice from (8) that the effect*s* of X on Y, i.e., d**y**/d**x** in line (*iv*), impinge in general on all units, d**y**, and vary – the *vector* of effects differs – depending on (a) which units are treated, i.e., the specific allocation of treatments across units, d**x**, and (b) on **W**, the relative connectivity among units, i.e., the specific set of $\{w_{ij}\}$ connecting units according to which the contagion diffuses. Thus, proceeding nonparametrically, each possible allocation of 1s and 0s across the n units – there are 2^n such permutations – corresponds to a different treatment; the effect of each such vector of treatments depends further on **W**, which in general has $n(n-1)$ potentially unique elements (yielding $2^{n(n-1)}$ possible **W** if connectivity is binary and ∞ if continuous). Thus, there are minimally $2^n \times 2^{n(n-1)} \gg n$ treatment effects to estimate nonparametrically: obviously impossible without considerable structure (which can be productively imposed in the form of Bayesian hyperpriors in this context, see, e.g., Best et al., 2005). Because at $\rho = 0$ the allocation of treatments and contents of W are irrelevant, a sharp null hypothesis may be formed to test *whether* spatial interdependence is present, but that is the extent of the possible nonparametrically: ρ and d**y**/d**x** are inestimable without considerable structure. Indeed, notice that spatial *association*, i.e., correlation, leaving aside causality entirely, cannot even be measured until the elements of **W** are specified and thereby *proximity* defined.

Simultaneous spatial interdependence is also illustrative as a special case of causal-systems simultaneity, with line (*i*) of (8) giving a system of equations with three endogenous variables: y_1,y_2,y_3. Thus the discussion from the spatial-interdependence

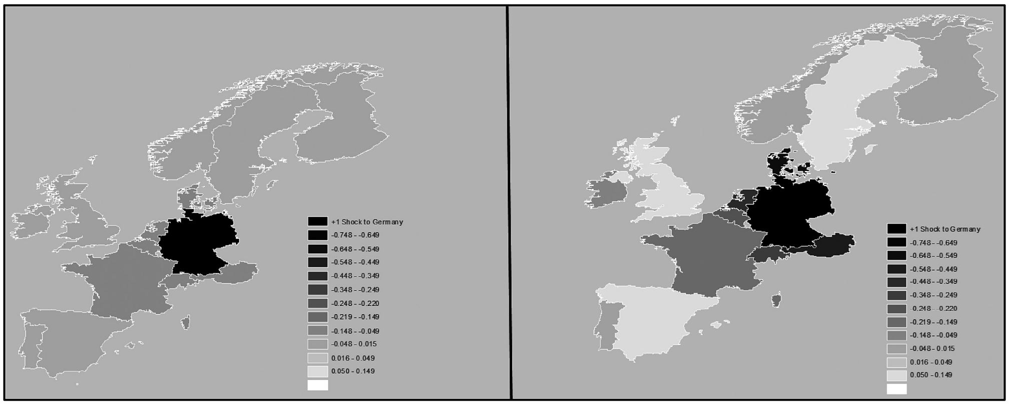


Figure 31.4 Maps depicting the initial (left panel) and LRSS (right panel) spatial ALM-spending responses to +1 shock in Germany

Source: Franzese and Hays (2006).

case applies also, *mutatis mutandis*, to causal systems of simultaneous equations more generally. Socio-politico-economic contexts with cross-unit contagion, $y_i \Leftrightarrow y_j$, or other simultaneous causality, $y_i \Leftrightarrow x_i$, imply processes like:

$$
\begin{aligned}
&(i)\ y = \alpha_0 + \alpha_1 x + \alpha_2 z_y + \varepsilon_y \\
&\quad\ x = \beta_0 + \beta_1 y + \beta_2 z_x + \varepsilon_x \\
&\Rightarrow (ii)\ y = \alpha_0 + \alpha_1\left(\beta_0 + \beta_1 y + \beta_2 z_x + \varepsilon_x\right) \\
&\qquad\quad + \alpha_2 z_y + \varepsilon_y \\
&y = \frac{\left[\alpha_0 + \alpha_1\beta_0 + \alpha_1\beta_2 z_x + \alpha_2 z_y + \varepsilon_y + \alpha_1\varepsilon_x\right]}{\left(1-\alpha_1\beta_1\right)} \\
&\Rightarrow (ii)\ \frac{dy}{dx} = \frac{\alpha_1}{1-\alpha_1\beta_1}\ ,\ \textbf{and not}\ \frac{dy}{dx} = \alpha_1
\end{aligned}
\tag{9}
$$

A well designed experiment, or a valid discontinuity or instrumentation design, will identify and estimate α_1, which is indeed sufficient to test *whether X* affects *Y*, since that effect, $\frac{dy}{dx} = \frac{\alpha_1}{1-\alpha_1\beta_1}$, is zero if $\alpha_1 = 0$, but it is clearly insufficient to *estimate* the effect, i.e., the causal response. This is because experiments work to identify the *existence* of causal effects precisely by preventing *estimation* of causal *responses* in the actual simultaneous system of interest. Specifically, causal-inference designs aim to block, *internally*, the feedback from *y* to *x* that actually occurs *externally* in the inference population. In the contexts of actual interest and intended application, if one 'moved' *x*, this would create *impulse* α_1 to *y*, but that in turn would spur β_1 further movement in *x*, which would move *y* some more, which in turn would move *x*, and so on.[32] Thus, ironically, experiments and nonparametric causal-inference designs estimate causal *parameters*, like $\frac{\partial y}{\partial x} = \alpha_1$, not causal *effects*, like $\frac{dy}{dx} = \frac{\alpha_1}{1-\alpha_1\beta_1}$.

In other words, notwithstanding that the NHR labels the estimand (1) and its empirical estimate (2) 'the causal effect', the aim of studies deploying them is actually *causal inference*, i.e., to establish *whether* a particular causal relation *exists* and not *causal-effect estimation*. The latter is to *estimate how* (not whether) outcomes of interest respond to inputs of interest, i.e., to estimate $\frac{dy}{dx}$, where those are expressly total rather than partial derivatives or differences, and the response, *dy*, and/or the treatment, *dx*, may actually be vectors or matrices of counterfactuals.

As already discussed, causal inference naturally emphasizes internal validity, whereas description and prediction instead stress external validity. Causal-response estimation, for its part, is similar to prediction in that its gold-standard ideal is an out-of-sample performance of the response estimate, emphasizing external validity; but, it is also similar to causal inference in that the external responses it aims to estimate are *causal effects*, not merely to predict $E(y_{it}|x_{js})$, but to predict how y_{it} would respond (conjunctive tense) causally to hypothetical movements in x_{js}: predictive (counterfactual) causal-response estimation. Internal causal validity is also crucial.

Given that the NHR model is inadequate, for causal-response-estimation purposes, |to meet the challenge of ubiquitous simultaneity, progress under *omnicausality* ('just about everything causes just about everything else') will rely on substantively/theoretically informed econometric modeling, as it did also in fruitfully addressing effect heterogeneity and context conditionality, and spatial, temporal, and spatiotemporal dynamics.[33] To begin, consider the general case of (linear) systems simultaneity, noticing the similarity to the spatial simultaneity in (8) and the bivariate case in (9):

$$(i)\quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_m \end{bmatrix}_i' \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1m} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mm} \end{bmatrix}$$

$$+\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_k \end{bmatrix}_i' \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{km} \end{bmatrix}$$

$$=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_m \end{bmatrix}_i$$

$$(ii)\quad \underbrace{\mathbf{y_i'}}_{1\times M}\underbrace{\Gamma}_{M\times M}+\underbrace{\mathbf{x_i'}}_{1\times K}\underbrace{\mathbf{B}}_{K\times M}$$

$$=\underbrace{\boldsymbol{\varepsilon}_i}_{1\times M}\begin{cases}\text{which normalizing diagonal } \gamma_{ii} \text{ to 1 and} \\ \text{reversing sign of } \gamma_{ij} \text{ can be written for} \\ \text{all } N \text{ obs:}\end{cases}$$

$$(iii)\quad \underbrace{\mathbf{Y}}_{N\times M}=\underbrace{\mathbf{Y}}_{N\times M}\underbrace{\Gamma^*}_{M\times M}+\underbrace{\mathbf{X}}_{N\times K}\underbrace{\mathbf{B}}_{K\times M}+\underbrace{\mathbf{E}}_{N\times M}$$

$$\Rightarrow \mathbf{Y}=(\mathbf{XB}+\mathbf{E})(\mathbf{I}-\Gamma^*)^{-1}$$

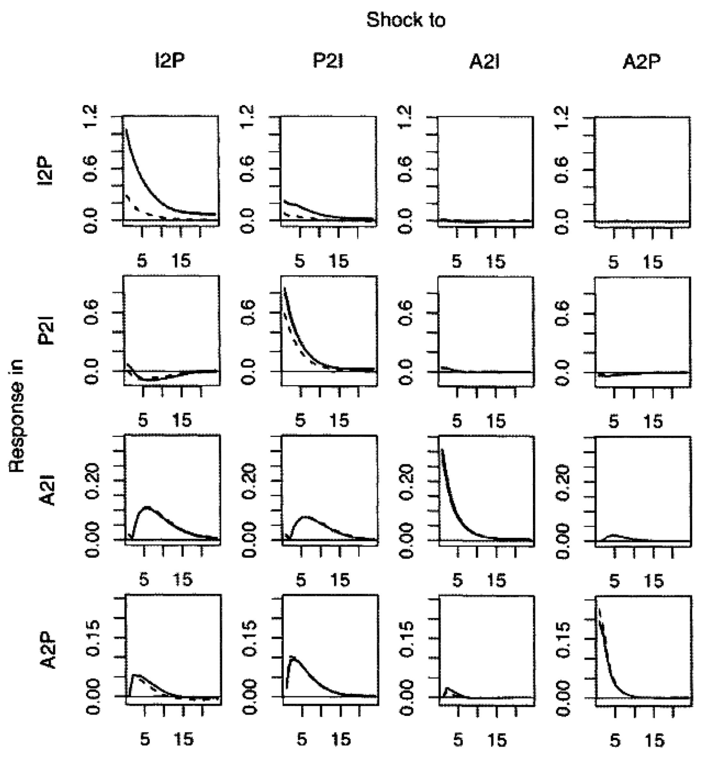
$$(iv)\quad \Rightarrow d\mathbf{Y}=(d\mathbf{X})\mathbf{B}(\mathbf{I}-\Gamma^*)^{-1}$$

(10)[34]

Thus, the translation from the causal-*parameter* estimates yielded by a well-designed experiment or single-equation causal-inference strategy like discontinuity or instrumental-variable designs to causal-*response* estimates involves a systems steady-state multiplier, $(\mathbf{I}-\boldsymbol{\Gamma}^*)^{-1}$, analogous to the temporal multiplier in (7), spatial multiplier in (8), and bivariate-system multiplier in (9). Thus, causal-effect *estimation* requires systems estimation (Jackson, 2008, is an excellent exposition[35]), or at least somehow an estimation of all the parameters of the properly modeled system relevant to some desired out-of-sample causal-response estimation. Unfortunately, political science and international relations rarely focus attention on system estimation, despite *strategic interdependence* being definitionally core to both. The emphasis on theory testing and its consequent idealization of nonparametric causal inference likely bears some blame for this, notwithstanding that systems interdependence greatly complicates even hypothesis testing and makes econometric modeling inescapably essential to causal-effect estimation.[36]

One econometric-modeling approach that does focus squarely on dynamic systems of endogenous equations is the Bayesian Structural Vector Autoregression (BSVAR) framework. In one illustrative application, Brandt et al. (2008) uncover the reciprocity and other reactions between the Israeli government and military, Palestinian groups, and US official diplomatic and foreign-policy actions (Figure 31.5).[37] The discussions above prove that these rich substantive interrelations could not be estimated in a nonparametric causal-inference approach and that, in fact, even testing for the existence of causal effects related to the alternative reciprocity, accountability, and credibility theories of these actors' strategically interdependent behavior would likely

Reciprocity Responses for the Bystander and Credibility Models



Note: I2P = Israel to Palestine; P2I = Palestine to Israel; A2I = the United States to Israel; A2P = the United States to Palestine. Bystander model results are shown with solid lines, and credibility model results are shown with dashed lines. Responses are mean estimates of positive one standard deviation shocks to each equation with 68 percent error bands computed by eigenvector decomposition method over twenty-four months.

Figure 31.5 BSVAR Estimated responses from a system of Israel↔ Palestinian, US→Israel, US→Palestinian actions

Source: Brandt et al. (2008).

fail using such a framework due to ubiquitous 'heterogeneity and spillovers' (see Rubin (1990: 282) again).

CONCLUSION

Empirical analyses in political science and international relations have at least one of four different goals: description, prediction, causal inference, and causal estimation.[38] The different aims carry with them different sets of weights on desiderata like internal and external validity, robustness, flexibility, efficiency, and richness. For causal-inference theory-*testing* purposes, the aim is to establish empirically the *existence* of a causal relationship as credibly as possible. For these purposes, the (a) randomized (b) controlled trial, being the strongest-possible guard against spuriousness and reverse causality and with the greatest possible nonparametric purity against 'model dependence', represents the gold-standard ideal. However, these causal internal-validity strengths generally come at some costs in terms of representativeness of experimental sample, treatment, or context, thereby limiting the utility of the causal inferences in practical applications, which necessarily require extra-sample inference and therefore external validity, which are not strengths of *laboratory* experimentation. Accordingly, social-science experimentation moves into the field and the survey to enhance representativeness, and from there into econometric-modeling techniques like matching, difference-in-difference, discontinuity, and instrumentation designs in observational data, which tend to maximize representativeness of the sample to the population and context of intended inference. From this perspective, the remaining threats to valid causal inference are effect heterogeneity and spillovers, violations of SUTVA that would bias causal hypothesis tests. In fact, the challenges of multicausality – effect heterogeneity and context-conditionality, temporal, spatial, spatiotemporal dynamics and interdependence (spillovers), and omnicausality (spillovers) – are ubiquitous, indeed virtually definitionally central, in social-science and in socio-politico-economic reality. From a causal-inference perspective, econometric modeling is aimed to address these challenges to valid designed-based testing of positive social-science theory. Beyond causal inference, however, this chapter has noted that, in the first instance, causality is irrelevant for measurement and prediction empirical purposes. The gold-standard ideal for measurement is neither internal nor external validity, but rather *usefulness*,[39] toward providing summary descriptions and conveying information and understanding thereof. For prediction, the ideal is out-of-sample performance, i.e., prediction or forecast error. Whether the prediction input-output algorithm involves causal relationships or simply associations is irrelevant; external rather than internal validity is crucial. Econometric modeling can play central roles in both prediction and measurement, but its essential and most important role lies in causal-response estimation. As this chapter has demonstrated, valid causal-inference studies can at best provide estimates of causal *parameters*, not causal *effects*. Given the complex causal heterogeneity and context conditionality that characterizes socio-politico-economic reality, in fact, even causal-parameter estimation is impossible without considerable 'structure', ideally via theoretically/substantively informed econometric-model specification. The limitations of (so-called) nonparametric causal inference, even solely for testing theories, and the necessity and virtues of econometric modeling become even more pronounced given the temporal, spatial, spatiotemporal, and systems causal interdependence of socio-politico-economics. Careful theoretically/substantively informed specification of econometric models is the essential heart of empirical analysis for purposes of estimating causal effects, and the aim of and gold standard for empirical modeling is to provide *useful empirical simplifications* of the actual, empirical causal processes of interest.[40] From this perspective, far from an unavoidable detraction, estimation of an econometric **model** reflecting the theory and substance of the context is the very goal of the causal-estimation exercise.

Notes

[1] *Positive* here opposes *normative*, positive theory being about how the world works in actuality as opposed to how the world *ought* to work normatively or would work in some fictional *ideal*.

[2] One can also distinguish two types of empirical questions: factual questions ('what happened or will happen?') and causal questions ('why did or will something happen?'). The former have empirically extant, finite populations and deterministically true answers – 'what percentage of the citizens of certain country approve of the government's performance?' – and the latter have hypothetical populations and uncertainly estimated answers (of theoretical *because*s): 'what characteristics of citizens, governments, and performance affect citizens' approval of governments?'.

[3] To offer a definition, (the purpose and evaluative standard of) an *econometric model*, analogously to a theoretical model in Clarke and Primo (2012), is (to be) a *useful empirical simplification*.

[4] The variables x and y are empirical measures, here assumed to be wholly unproblematic, of the theoretical concepts, X and Y; dx⇒dy is the empirical implication derived from the theoretical argument, $X \Rightarrow Y$.

[5] Notice the word used in this testing context is *inference*, and not *estimation*; this is because the central aim is to infer the existence of a causal effect, i.e., to establish that $dY/dX \neq 0$, rather than to estimate it. The empirical estimand from a causal-inference design is most usually but not necessarily, a difference in means, $E(y|x = 1)$-$E(y|x = 0)$ (see Bowers and Leavitt, Chapter 41, this *Handbook*), which will only in very specific (and likely exceedingly rare in social science) conditions equate to an empirical estimate of the true causal *effect* of x on y, understood as dy/dx, i.e., how y responds to a causal impetus from dx.

[6] The POF, also called the Neyman–Rubin or Holland–Neyman–Rubin *causal model*, is indeed a model: the causal effects described in (1) are discrete, static, additive, and separable. Indeed, precisely these characteristics of the POF/(H)NR causal model simultaneously make it so powerful for *causal inference* (testing theorized causal-effect existence) and yet so limited for *causal estimation* (estimating empirical causal effects or responses). See also note 17 and the discussion throughout the rest of this chapter.

[7] Other counterfactually defined estimates have been proposed for causal inference/testing (in political methodology, e.g., see Bowers, 2013), but by far the most common practice is to define the causal quantity of interest as in (2).

[8] Double-lined arrows indicate causal relationships and single-lined arrows empirical ones, i.e., associations.

[9] Some scholars go so far as to suggest that if x cannot be manipulated, *race* for example, then it cannot be causal, but this confuses the empirically implementable with the logically possible. *Causality*, being a theoretical and not empirical concept, involves only the latter; the former is irrelevant (see, e.g., Woodward, 2016 for a fuller discussion).

[10] If empirical outcomes, y, are less than perfectly fully determined by the experimentally controlled x, such that there remains some residual component in dy, even if orthogonally random but especially if possibly systematically caused or related to alternative causes Z, then a large sample, successful randomization, and some reliance upon some form of central limit theorem are also essential to proper interpretation of test statistics from a RCT.

[11] Some scholars contend oppositely, that internal validity has lexical priority over external validity, that internal is more important and without it external has no value. Imbens (2010), e.g., suggests that instances where one could conduct the appropriate experiments and would choose observational data instead are inconceivable. To debate whether internal or external validity is more important or, especially, which is lexically prior is obviously inane: of course, one wants *both*, the aim being to infer (a) validly and (b) from the observed and already known to new contexts. If we must debate priority though, clearly the more defensible position is the reverse: external validity without internal validity (i.e., non-causal empirical associations within sample that obtain also beyond sample) is still useful, e.g., for prediction, whereas an internally validated causal relationship with no external validity has only descriptive value within the already observed and known sample and zero use in any context beyond the study, i.e., for *inference*.

12 The representativeness of an experimental sample to the intended population of inference refers to the equivalence of the subjects in the experimental sample and the external units to which the results of the study are to be inferred: college sophomores in a certain class compared to voters in actual democracies, for example. Representativeness of treatment analogously refers to the equivalence of the experimentally manipulated treatment to the concept theoretically understood as causal in the population of inferential interest, e.g., mention of party affiliation in a paragraph that trial subjects are given to read compared to the actual partisanship of actual bill sponsors in actual political contexts. Representativeness of context refers to the equivalence of the situation of the experimental subjects and treatments relative to each other and relative to the relevant socio-politico-economic reality outside the experiment compared with the situations in these regards of the intended inference population: randomized application of a campaign strategy that subjects read about or experience in a media lab in an experiment's contrived campaign compared to campaign *strategies* (by definition) strategically (which means interdependently) chosen by competing parties in actual campaigns where publics are going about their lives, not engaging in a social-science experiment.

13 The RCT obtains its strong causal-inference properties precisely by designing an unnatural context: feedback, which exists in nature, is severed by experimental control, and experimentally randomly independently assigned treatments are in nature likely non-randomly and often even strategically assigned. See also note 12.

14 Conjoint experiments offer some advances in this specific regard (see, e.g., Hainmueller et al., 2014).

15 The *i*,*j*,*s*,*t* subscripts are intended to signify that y_{it} may be a function of $\mathbf{x}$, $\boldsymbol{\beta}$, ε in any units *i* or *j* and periods *s* or *t*.

16 To be as fully general as possible, $\mathbf{x}_{js}$ may include y_{js} and/or temporal and/or spatial lags of $\mathbf{x}_{it}$ as well.

17 To elaborate these points more precisely, the POF estimand can be conceived as nonparametric estimate of some *average* treatment 'effect' (ATE), regardless of what functions, *f*, may have generated that average difference in means, and this may be adequate for purposes of testing whether this ATE is non-zero (the orthogonality of unobserved random components still seems necessary). However, to interpret this ATE as an *effect*, i.e., as an estimate of how *y* would respond to some exogenous d*x outside of the observed sample*, is to treat it as a model.

18 In fact, the similarity of matching and regression control extends further: regression controls **z** *to the degree* its effects manifest as modeled; matching controls any manifestation of effects of **z** provided, or *to the degree*, the appropriate form of **z** is included in the matching balancing.

19 Jackson (2008) offers a more complete introduction to instrumental-variable and systems estimation.

20 *Time* here refers to arguments that 'it happened yesterday, therefore it's exogenous', which is not guaranteed in socio-politico-economic applications, where human foresight can give causal weight to current expectations of futures. Moreover, exclusive reliance on temporal precedence for identification is highly susceptible to specification error.

21 *Nonparametric* here references methods that yield large numbers of discrete, unconnected values as distinct from methods explicitly intended to produce (likely graphical) descriptions that are not *a priori* structured but are smoothed descriptions (see Pagan and Ullah, 1999, for far fuller coverage of nonparametric econometrics).

22 Again, see Pagan and Ullah (1999) for a much fuller view of nonparametric analyses; here, we intend nonparametric causal inference or causal estimation specifically, which necessarily entail distinct causal 'effect' estimates for each and every context, there being allowed no functional smoothing connections between 'effects' in different conditions. A paradigm labeled *evidence-based medicine*, which carries considerable weight in the biomedical sciences, is illustrative here. The notion is that, if a well and credibly designed RCT yields reliable results that treatments of certain medicines in certain doses to patients with certain conditions, characteristics, and histories produces some estimated net-benefits, then, regardless of whether that RCT-estimated effect has some theoretical explanation, the treatment is to be applied. This is a purely predictive approach but one that attempts to retain the nonparametric foundations of the RCT. As such, the model on which it relies for external validity, i.e., the predictive basis on which to prescribe the treatment, is like that of matching: matching treatments applied to patients with matching conditions will have the same effect. No basis is provided for applying only *similar* treatments to only *similar* patients; that would require more of a model.

23 Chapter 2 of the classic text *Statistics* (Freedman et al., 2007 [1978]) extols the two great virtues of experimentation. Even in the examples

mooted there, though, some doubts of universal unmitigated virtue may be raised. For instance, when double-blind randomization is assumed vindicated because surgeons who know the health of their patients and the nature and severity of their ills yielded significantly beneficial results of an experimental surgery whereas blinded ones insignificantly so, this could, instead of suggesting pernicious bias, suggest effect heterogeneity, about which the non-blinded surgeons in the study know, as would – crucially – surgeons in actual practice. Chapter 3 then warns severely of the dangers of observational studies, lacking those two great experimental virtues. An interesting pattern develops, however: each example observational study's conclusion is overturned later by …another observational study, plus *arguments* that the latter was better designed…because *causality is ultimately a theoretical, not an empirical, matter*. Finally, the examples have also shifted from primarily clinical-medical in Chapter 2 to primarily epidemiological in Chapter 3, and epidemiology, like '[macro]economics [and most political science and international relations] is not an experimental science' (Sims, 2010).

[24] In practice, treatments are also nominal, x = (0,1). Although claim is often made to straightforward extensions for continuous treatments, in fact the extension is generally complicated and incompatible with nonparametric causal inference, as explained in the surrounding text (see also note 22).

[25] All interactions are symmetric in this way: how $\mathbf{z}$ moderates the effect of $\mathbf{x}$ on y, $d(dy/d\mathbf{x})/d\mathbf{z}$, is identical to how $\mathbf{x}$ moderates the effect of $\mathbf{z}$ on y, $d(dy/d\mathbf{z})/d\mathbf{x}$, because interactive effects, i.e., effects on effects, are cross derivatives, and the order of differentiation in a cross derivative is irrelevant.

[26] In 'The multiple effects of multiple policymakers'; Franzese (2010) shows how one can leverage the distinct aspects of multiple policymakers – effective (common pool) vs raw numbers (veto actors) of parties, variance (common pool) vs range (veto actors) polarization of parties, and the ideological distribution of parties (bargaining compromise) – along with the different ways these different aspects of multiparty government affect policy outcomes (common pool: proportionate over/under-action; veto actor: adjustment-rate retardation; bargaining-compromise: convex combinations) to separately model, and so to separately identify and estimate, the veto-actor, common-pool, and bargaining-compromise effects of multiple policymakers.

[27] So-called 'dynamic' nonparametric-causal-effect estimates are either estimates of β from the static (2) in moving-windows of data, or estimates using static (2) of the period effects in (7) (*iii*), without the model, i.e., not estimates of the model and its parameters ρ and β separately, without which they are incapable of generating dynamic response-path or LRSS estimates.

[28] Indeed, we could expand here to note that all data, all outcomes of interest, occur in some (space and) time, and so these issues actually arise universally, ubiquitously in all applied empirical analysis, experimental or observational.

[29] The elements w_{ij} of the spatial–weights matrix, $\mathbf{W}$, give the relative connectivity from j to i, and ρ the strength of interdependence (contagion) operating in that predetermined pattern.

[30] Assuming $\rho\mathbf{W}$ is the matrix equivalent of 'less than 1', such that $|\mathbf{I}-\rho\mathbf{W}|\neq 0$ so the inverse spatial-multiplier exists. Note that the spatial multiplier derives from an infinite sum of the reverberating spatial feedback analogously to the temporal case (of forward-propagating-only 'feedback'): $(\mathbf{I}-\rho\mathbf{W})^{-1} = \mathbf{I}+\rho\mathbf{W}+\rho^2\mathbf{W}^2+\rho^3\mathbf{W}^3+\ldots+\rho^\infty\mathbf{W}^\infty$.

[31] Franzese and Hays (2006) use a modified border-contiguity $\mathbf{W}$ to define proximity in this application.

[32] These feedback reverberations are dampening provided $\alpha_1\beta_1<1$, so the system is not explosive (see also note 30).

[33] Semi- and flexible parametric designs offer a promising way forward for the (likely ubiquitous) *combination* of causal heterogeneity and causal simultaneity (see, e.g., Marra and Radice, 2011).

[34] Γ in line (*ii*) has $\mathbf{1}$ on its diagonal; Γ^* in line (*iii*) has $\mathbf{0}$ on its diagonal and reverses sign of all off-diagonal elements from Γ.

[35] Also in this context, see methods specifically designed for complex or high-dimensional systems of endogenous dynamic equations, such as structural vector-autoregression (e.g., Kilian and Lütkepohl, 2017, for textbook exposition, and Pickup, Chapter 34 in this *Handbook*).

[36] In this regard, empirical-methodological practices in physics could serve as better exemplar for social science than biomedicine (see note 23). In physics, experimental statistics often yield not only tests of causal theoretical hypotheses but also estimates of the parameters in well-specified theoretical models, and it is the empirical-estimate-calibrated model rather than the experiment's test statistics that are used for causal-response estimates and prediction.

[37] The model assumes the United States influences but is not influenced by the other two actors.

[38] Furthermore, their empirical questions can be factual – 'who voted for Hitler?' – and so pertain to

defined, finite, and extant *populations* or theoretical – 'what characteristics of voters and contexts contribute to right-wing populist support?' – and so have populations of intended inference that are hypothetical and unlimited.

[39] The analogy to Clarke and Primo's (2012) declaration of *usefulness* as the aim of theoretical modeling is intentional and perfect.

[40] Again, the analogy to Clarke and Primo's (2012) declaration that theoretical models are to be *useful simplifications* is intentional and perfect.

REFERENCES

Best, N., Richardson, S., Thomson, A. 2005. 'A Comparison of Bayesian Spatial Models for Disease Mapping', *Statistical Methods in Medical Research* 14(1):35–59.

Bowers, J., Fredrickson, M., Panagopoulos, C. 2013. 'Reasoning about Interference between Units: A General Framework.' *Political Analysis* 21(1):97–124.

Brandt, P.T., Colaresi, M., Freeman, J.R. 2008. 'The Dynamics of Reciprocity, Accountability, and Credibility', *Journal of Conflict Resolution* 52(3):343–74.

Caughey, D., Sekhon, J.S. 2011. 'Elections and the Regression Discontinuity Design: Lessons from Close US House Races, 1942–2008', *Political Analysis* 19(4):385–408.

Clarke, K., Primo, D. 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford: Oxford University Press.

Coppedge, M., Alvarez, A., Maldonado, C. 2008. 'Two Persistent Dimensions of Democracy: Contestation and Inclusiveness', *The Journal of Politics* 70(3):632–47.

Egami, N., Imai, K. 2015. *Causal Interaction in High Dimension*. Unpublished manuscript,, Princeton, NJ: Princeton University Press.

Franzese, R.J. 1999. 'Partially Independent Central Banks, Politically Responsive Governments, and Inflation', *American Journal of Political Science* 43(3):681–706.

Franzese, R.J. 2002. *Macroeconomic Policies of Developed Democracies*. Cambridge: Cambridge University Press.

Franzese, R.J. 2003. 'Multiple Hands on the Wheel: Empirically Modeling Partial Delegation and Shared Policy Control in the Open and Institutionalized Economy', *Political Analysis* 11(4):445–74.

Franzese, R.J. 2007. 'Multicausality, Context-Conditionality, and Endogeneity', in C. Boix & S.C. Stokes, eds, *The Oxford Handbook of Comparative Politics*, Oxford: Oxford University Press.

Franzese, R.J. 2010. 'The Multiple Effects of Multiple Policymakers: Veto Actors Bargaining in Common Pools', *Rivista Italiana di Scienza Politica* 40(3):341–70.

Franzese, R.J., Hays, J.C. 2006. 'Strategic Interaction among EU Governments in Active Labor Market Policy-Making: Subsidiarity & Policy Coordination under the European Employment Strategy', *European Union Politics* 7(2):167–89.

Freedman, D., Pisani, R., Purves, R. 2007 [1978]. *Statistics*, 4th ed. New York: W.W. Norton.

Gerber, A.S., Green, D.P., Shachar, R. 2003. 'Voting May Be Habit-Forming: Evidence from a Randomized Field Experiment', *American Journal of Political Science* 47(3):540–50.

Hainmueller, J., Hopkins, D.J., Yamamoto, T. 2014. 'Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices *via* Stated-Preference Experiments', *Political Analysis* 22(1):1–30.

Hays, J.C., Ornstein, J.T., Franzese, R.J. 2019. *Estimating the Interest-Premium Cost of Left Government by Regression-Discontinuity Analysis of Close Elections*, Unpublished manuscript, St. Louis, MO: Washington University.

Hendry, D.F. 1995. *Dynamic Econometrics*. Oxford University Press on Demand.

Hershey, M.R. 2009. 'What We Know About Voter-ID Laws, Registration, and Turnout', *PS: Political Science & Politics* 42(1):87–91.

Imai, K., Ratkovic, M. 2013. 'Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation', *The Annals of Applied Statistics* 7(1):443–70.

Imbens, G.W. 2010. 'Better LATE Than Nothing', *Journal of Economic Literature* 48(2):399–423.

Jackson, J.E. 2008. 'Endogeneity and Structural Equation Estimation in Political Science', in *The Oxford Handbook of Political Methodology*, Oxford: Oxford University Press.

Kellstedt, P.M., Whitten, G.D. 2018. *The Fundamentals of Political Science Research*. Cambridge: Cambridge University Press.

Kilian, L., Lütkepohl, H. 2017. *Structural Vector Autoregressive Analysis*. Cambridge: Cambridge University Press.

Marra, G., Radice, R. 2011. 'A Flexible Instrumental Variable Approach', *Statistical Modelling* 11(6):581–603.

Pagan, A., Ullah, A. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Pearl, J. 1995. 'Causal Diagrams for Empirical Research', *Biometrika* 82(4):669–88.

Poole, K. 2019. *Spatial Models of Parliamentary Voting*. Cambridge: Cambridge University Press.

Pritchett, L, Sandefur, J. 2015. 'Learning from Experiments When Context Matters', *American Economic Review* 105(5): 471–5.

Schrodt, P.A., Gerner, D.J. 2000. 'Cluster-Based Early Warning Indicators for Political Change in the Contemporary Levant', *American Political Science Review* 94(4):803–17.

Selway, J. 2011. 'The Measurement of Cross-cutting Cleavages and Other Multidimensional Cleavage Structures', *Political Analysis* 19(1):48–65.

Sims, C.A. 2010. 'But Economics Is Not an Experimental Science', *Journal of Economic Perspectives* 24(2):59–68.

Treier, S., Jackman, S. 2008. 'Democracy as a Latent Variable', *American Journal of Political Science* 52(1):201–17.

Ward, M.D. 2016. 'Can We Predict Politics? Toward What End?', *Journal of Global Security Studies* 1(1):80–91.

Woodward, J. 2016. 'Causation and Manipulability', in E.N. Zalta, ed., *Stanford Encyclopedia of Philosophy* (Winter). Available at https://plato.stanford.edu/archives/win2016/entries/causation-mani/ (Accessed on 31 January 2020).