Policy Prediction: The Missing Tool in Experimental Econometrics and a Roadmap to Fix It

Nandan Rao

September 7, 2019

Abstract

Applied economics research is most often "applied" to policy making, yet formal frameworks for policy prediction are largely ignored in the literature of experimental and quasi-experimental econometrics. I explain the existence of this gap from a historical perspective, review the current methods that do exist, and set a research agenda to fill it using recent ideas from econometrandics and machine learning.

1 Introduction

Randomized control trials, or natural experiments that replicate them, have become the official gold standard of empirical work in economics and many related fields. More and more, policy makers are encouraged to look to RCTs to make "evidence-based" policy decisions (Manski 2013; Cartwright and Hardie 2013).

Many prominent economists have expressed a concern that RCTs, and the quasi-experimental methods that seek to replicate them (i.e. natural experiments, intrumental variables, regression discontinuity, etc.) are particularly difficult to generalize to new contexts due to their orverarching concern for internal validity, often at the expense of external validity (**Heckman1995**; Heckman 2008; Deaton 2010; Manski 2013; Deaton and Cartwright 2018).

Making evidence-based policy decisions is an act of generalizing from previous studies to decisions about the future. Charles Manski (2013), voices the concern that experimental studies have tended to "be silent" on the

question of external validity and that "from the perspective of policy choice... What matters is the informativeness of a study for policy making, which depends jointly on internal and external validity."

This silence on external validity means that, despite the huge field of research and techniques for ensuring causal identification (internal validity) via experimental or quasi-experimental methods, very little has been done to either A) create tools to prove that the same results will apply elsewhere or B) create tools to predict the results in a new location, given proven results from one or more experiments.

Banjerjee himself (A. Banerjee, Chassang, and Snowberg 2016), a large proponent of RCTs, has recently echoed the current lack of and need for more formal systems for the generalization of experimental studies, saying "it is our belief that creating a rigorous framework for external validity is an important step in completing an ecosystem for social science field experiments, and a complement to many other aspects of experimentation."

A small but growing literature has sprung up around empirically proving that results from prominent RCT or other causally identified studies do not easily extrapolate to new contexts (Pritchett and Sandefur 2015, Allcott 2015, Bisbee et al. 2017, Rosenzweig and Udry 2019). These results emphasize a need for powerful extrapolation tools if policy makers are realistically expected to make decisions based on "evidence."

Predicting the results of of a policy in a new location is, tautologically, a prediction problem. Machine learning is a field which has been very successful at formalizing prediction problems. Domain adaptation, a subfield of machine learning, formalizes the problem of moving from one (or more) domains with labelled data to a new domain where only unlaballed data is available (for a survey, see Pan and Yang 2010).

This article will take the following form: Section 2 will define external and internal validity, their role in scientific inference, review the development of the Fisherian experimental model of statistical inference, and set out the need for a formal framework of extrapolation. Section 3 will argue that a target context is necessary by reviewing the development of such an argument in mid-century econometrics and formalizing their argument in probablistic terms. Section 4 will introduce the concept of domain adaptation in machine learning. Section 5 will review current discussions and state of the art for formal frameworks within the literature of empirical economics and policy prediction. Section 6 will present recent work in machine learning that combines the ideas from structural econometrics and domain adaptation. Section 7 will present conclusions and frame a research agenda.

2 Validity and Counterfactual Identification

2.1 A Taxonomy of Validities

Critiques against the holy status of RCTs have focused on their preference for internal validity and tendency to completely ignore external validity. It is worth defining these terms and reflecting on exactly what it is that experimental methods provide us. The most agreed upon definition of these terms comes from Shadish, T. D. Cook, and D. T. Campbell 2002 (an update of their previously popular taxonomy of validities from T. Cook and D. Campbell 1979):

> **Statistical Conclusion Validity:** The validity of inferences about the correlation (covariation) between treatment and outcome.
>
> **Internal Validity:** The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.
>
> **Construct Validity:** The validity of inferences about the higher order constructs that represent sampling particulars.
>
> **External Validity:** The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

Many authors use the term external validity to refer to what Shadish, Cook, and Campbell separate into construct and external validity. As both are involved in generalization, and both are external to the particulars of the sample, I will follow that abuse of terminology and refer to all the challenges of both as "external validity."

What, then, is meant by the term "inference?"

2.2 Inference and Statistics

Following from Aristotelian tradition of logic, inference is the process of reasoning and can be broken into two parts: reasoning from particulars to generals (induction) and reasoning from generals to particulars (deduction). Deduction is often recognized by Aristotle's syllogisms, such as:

> All men are mortal
> Socrates is a man
> Therefore, Socrates is mortal

Once one has induced a general law ("all men are mortal"), one can deduce facts that might otherwise not yet be apparent ("Socrates will die"). There is, of course, a natural contradiction in this process: how can one know that all men are mortal, if one did not already know that Socrates will die? In other words, how can one ever claim that "all men are mortal" until one has seen every man die?

This problem forms the foundation of David Hume's "Problem of Induction":

> "As to past Experience, it can be allowed to give direct and certain information of those precise objects only, and that precise period of time, which fell under its cognizance: but why this experience should be extended to future times, and to other objects, which for aught we know, may be only in appearance similar, this is the main question on which I would insist" (Hume).

There are two distinct problems Hume raises, that of generalizing from one object to another (from seeing some men die to assuming all men have died) and that of generalizing from the past to the future (because all men have died, thus, all men will die). He goes on to explain that this extrapolation is only valid under strong assumptions as to the "uniformity of nature." One must assume nature is uniform in such a way as to enable extrapolation from one object to another or from the past to the future.

R.A. Fisher framed his statistical techniques in terms of "inductive inference." In the introduction to The Design of Experiments (1935), he explicitly frames his canonical book and its techniques in the terms of logicians such as Hume, arguing for the possibility of induction via statistical methods:

> "it is possible to draw valid inferences from the results of experimentation... as a statistician would say, from a sample to the population from which the sample was drawn, or, as a logician might put it, from the particular to the general." (Fisher)

Fisherian statistical inference is a tool in the process of induction that seeks to address Hume's problem of extending experience with one object to that of similar objects. In particular, the "similar objects" to which conclusions are extended are not just similar in appearance, but rather have a distinct relationship to the experienced objects: they are the population from which the experienced objects represent a sample.

Fisher's methodologies for significance testing relate to drawing conclusions about populations given a sample. The validity of the use of such techniques in a study falls squarely under the category of "statistically conclusion validity" in the taxonomy of Shadish, Cook, and Campbell.

Fisher's theory of experiments (in particular, the RCT) addresses internal validity and causality. The connection between the RCT and the identification of a causal relationship comes straight from John Stuart Mill's (1843) "method of difference" for causal discovery:

> "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon."

Fisher (1935) clearly had this in mind when he defended randomization, claiming that holding all other variables constant was not feasible, and thus, holding them to the same distribution, by making the assignment of treatment independent of those variables, was desirable (Rosenbaum 2005). It should be noted, however, that for the subjects of interest to Fisher, the difference between something being a "cause" or "a necessary part of the cause" was not especially important. What Fisher was essentially interested in was interventional prediction: what is the effect when one makes individual changes to the growing conditions of these plants.

Holland (1986) begins his landmark paper, before formulating the Rubin-Neyman causal model, by framing his goal:

> "Others are concerned with deducing the causes of a given effect. Still others are interested in understanding the details of causal mechanisms. The emphasis here will be on measuring the effects of causes because this seems to be a place where statistics, which is concerned with measurement, has contributions to make."

The success of Fisher's framework of randomization and the Rubin-Neyman causal model comes down to this razor focus in purpose: they make no claims to discover all the causes of a given effect, to discover the mechanism of the cause, or even to separate between a cause or a necessary part of a cause. They allow us to reason about counterfactuals: what would have happened, on average and in the past, had we treated

our entire population rather than a randomized part of a randomized sample drawn from that population. It operationalizes Mill's method of differences, creating a pathway to internal validity that is achievable in the real world.

The counterfactual causal model, however, provides no framework for generalizing from a specific population to a more general one or from the past to the future (Heckman 2008). The "effects of causes" is a black-box methodology for causal identification (**Heckman1995**). It does not require one to answer the question "why" and it is therefore inherently context specific: it asks, "what happened when I did X in context D?" In Fisher's line of work (agriculture), none of those shortcomings were problematic. He was able to randomly sample from the exact population (seeds of grain) to which his inferences needed to generalize.

The success of Fisher's randomization in his field of agriculture, and the subsequent success of the Rubin-Neyman causal model in epidemiology, is in no way incidental. These are fields that deal with encapsulated biological units where agreed-upon scientific theory tells us that the relationships in these systems will be invariant to a wide degree of changes that happen in the world from one year to the next or from one country to the next.

For example, it is implicitly assumed that the growth of corn will not be affected by its expectations of how it will be cut down and processed during the harvest. Corn grows according to the properties of the soil and the sun it receives. Its growth will be independent of its destination in corn flakes or arepas, conditional on those properties. Opioids will inhibit pain in humans, regardless of their political ideology, faith in their governmental institutions, or their love of Shark Tank.

These arguments are not made explicitly, but their validity is absolutely necessary to enable the application of their findings: they create laws defined in relation to the entire range of contextual changes that one might want for prediction- and decision-making and that allows the laws to be applied deductively in a wide array of useful situations.

While the validity of these arguments is taken for granted in contained biological processes, this is simply not the case in the social sciences. This is why Shadish, Cook, and Campbell lay out 19 different threats to construct and external validity of social science experiments. In the case of growing corn, the threats either simply do not apply or are implicitly neutralized through basic scientific understanding of plant growth.

In the case of economics, the outcomes are regularly based on individuals decisions to consume, work, study, invest, or move. The treatments are regularly subjected to a gauntlet of mediating factors and interacting

variables that are highly context-dependent and correlated across individuals in a single place and time. The population for which one wants to draw actionable inference is in the future and it is fundamentally different from the population the sample from drawn from. Internal validity of a study in economics is not sufficient for actionable inference to take place in a new context.

2.3 The Danger of Informal Inference

It might be argued that it is, and should be, up to the sound judgement and expert opinion of the policy maker to determine if a given counterfactual analysis should extrapolate to their context or not. Empirical studies only need to provide internal validity, according to this argument, as the extrapolation is done by experts who know their target domain.

While incorporating expert domain knowledge can only help predictions, we can differentiation between using a formal statistical framework and using an informal framework of judgement. John Stuart Mill, in his — –, reflects on these difference and the dangers of inferring without formal frameworks.

He begins by denying that the only process of inference consists of separately applying induction (particulars to generals) and then deduction (general law to particular context):

> "All inference is from particulars to particulars: General propositions are merely registers of such inferences already made, and short formulae for making more: The major premise of a syllogism, consequently, is a formula of this description: and the conclusion is not an inference drawn from the formula, but an inference drawn according to the formula: the real logical antecedent, or premise, being the particular facts from which the general proposition was collected by induction." (John Stuart Mill)

In other words, in the process of creating the general law, one has created a series of particular laws, and only once assured that all the particular laws are valid can one be assured of the more general law. He goes on to caution against the direct reasoning of particulars to particulars because it is informal and we are likely to bring our own biases into the process and make mistakes:

> "In reasoning from a course of individual observations to some new and unobserved case, which we are but imperfectly acquainted with (or we should not be inquiring into it), and in

> which, since we are inquiring into it, we probably feel a peculiar interest; there is very little to prevent us from giving way to negligence, or to any bias which may affect our wishes or our imagination, and, under that influence, accepting insufficient evidence as sufficient." (John Stuart Mill)

John Stuart Mill argues that formal procedures, such as that implied by the framework of induction and deduction, allow individuals to avoid these biases. In his world, there was no formal procedure for reasoning from particulars to particulars, but there was for reasoning from particulars to general. As such, he recommends the latter as a way to avoid biases of "wishes" and "imagination."

In an ideal world, the research community has discovered a set of general laws that are invariant to all contexts and the policy maker can apply them, via the process of deduction, to get the desired result in their circumstance. But what if that general law has not yet been discovered? Then the policy maker must look at individual studies (particulars) and attempt to infer a prediction for the result of a similar policy in their context. This is exactly the situation that John Stuart Mill has warned is fertile ground for bias and imagination.

3 The Origins of Structure

Economists during the same period as Fisher took a different approach to conceptualizing and thinking about the inference they were doing. Ragnar Frisch, theorizing about macro-dynamic analysis, set out several key ideas relating to the structure of a system and the autonomy of a structure (Frisch 1995). For Frisch, the "structure" of a system was all the characteristics of the phenomena that could be quantitatively described. In his macrodynamic systems, the structure is defined by a set of functional (simultaneous) equations. He then poses the question: what would happen to the system due to an arbitrary change in a single variable? To do so could imply a different "structure" than the one which the equations describe, requiring a different set of equations altogether to describe the new system.

> "But when we start speaking of the possibility of a structure different from what it actually is, we have introduced a fundamentally new idea. The big question will now be in what directions should we conceive of a possibility of changing the structure?... To get a real answer we must introduce some fundamentally new information. We do this by investigating what

> features of our structure are in fact the most autonomous in the sense that they could be maintained unaltered while other features of the structure were changed. . . So we are led to constructing a sort of super-structure, which helps us to pick out those particular equations in the main structure to which we can attribute a high degree of autonomy in the above sense. The higher this degree of autonomy, the more fundamental is the equation, the deeper is the insight which it gives us into the way in which the system functions, in short, the nearer it comes to being a real explanation. Such relations form the essence of 'theory'."

This concept, that of "the essence of theory" being the discovery of some relationship that is autonomous and invariant to a great degree of changes we can imagine performing to a system, is taken up by Trygve Haavelmo (1944), who writes that:

> "The principal task of economic theory is to establish such relations as might be expected to possess as high a degree of autonomy as possible."

He then goes on to consider a distinction between the "invariance" of a relationship under hypothetical changes in structure versus the "persistence" of a relationship under observed changes in structure:

> ". . . if we always try to form such relations as are autonomous with respect to those changes that are in fact *most likely to occur*, and if we succeed in doing so, then, of course, there will be a very close connection between actual persistence and theoretical degree of autonomy."

This implies a connection between autonomy and the *type* of changes to which it is invariant *with respect to*. This point is made even more explicitly by Leonid Hurwicz (1966). Similar to his predecessors, his model consists of a system of equations that constrain the state of the world, given a history of states. He calls this system of equations a "behavior pattern." He states that:

> "A great deal of effort is devoted in econometrics and elsewhere to attempts at finding the behavior pattern of an observed configuration. . . But do we really need the knowledge of the behavior pattern of the configuration?. . . It will be approached here

> from the viewpoint of prediction... That is, the word 'need' in the above question will be understood as 'need' for purposes of prediction."

He then goes on to define what he calls a "structural form" as one which is identified and identical across all possible behavior changes that one *needs* to predict within. He stresses that:

> "The most important point is that the concept of structure is relative to the domain of modifications anticipated."

Thus, there is an inherent and irrevocable connection between what we consider a "law" and the degree of changes we require the law to persist across. The law is defined in relation to those changes. Folowing Hume, the performance of induction is always connected to a specific assumption about the uniformity of nature. Additionally, the exact way in which nature must be uniform, for a particular system under study, is defined by the predictions we need to make with the discovered relations in that system.

3.1 Invariant Conditionals

Robert Engle, building on the work of Frisch and Hurwicz regarding invariant/autonomous structures, creates a statistical definition of what he terms "super exogoneity." A good review of this historical evolution of autonomy can be found in Aldrich 1989.

Super exogoneity is defined by Engle as follows. Consider a model in which the "structure" (the functional form) of a relationship between outcome y and variable z is parameterized by "structural" parameters $\lambda_1, \lambda_2 \in \lambda$. If the joint density implied by the model can be factorized as follows:

$$P(y, z, \lambda) = P(y|z, \lambda_1)P(z|\lambda_2)$$

and the conditional, $P(y|z, \lambda_1)$ remains invariant to changes in the marginal $p(z)$ (presumably caused by changes in it's generating process, parameterized by λ_2), then z is super exogenous. This definition allows super exogoneity to be refuted by data:

> "It is clear that any assertion concerning super exogeneity is refutable in the data for past changes in $D(z, |X_{t-1}, \lambda_2)$ by examining the behavior of the conditional model for invariance when the parameters of the exogenous process changed...However, super exogeneity for all changes in the distribution of z, must

> remain a conjecture until refuted, both because nothing precludes agents from simply changing their behavior at a certain instant and because only a limited range of interventions will have occurred in any given sample period."

Writing a system in terms of super exogenous variables is akin to finding Frisch's "super-structure." This is the part of the system that stays invariant to a set of allowable modifications, parameterized by λ_2. This is the part of the system that must be known in order to make policy predictions, when λ_2 includes the changes in the policy, whose total causal effect (in the sense of Pearl 2000) is transmitted to the output variable y through z. The existence of such an invariant super-structure is a necessary prerequisite to successfully predict the effects of policy and therefore to successfully inform policy choice from empirical data.

This method of looking in the data for a conditional distribution that is invariant to "structural" changes that lead to differences in the marginal distribution of its "inputs" has formed the basis of new line of work in statistics and machine learning around causal discovery (Peters, Bühlmann, and Meinshausen 2015; Heinze-deml and Meinshausen 2017; Rojas-Carulla, Schölkopf, and Turner 2018). This line of research, in the terms of Engle, can be thought of as methods for model discovery by selecting covariates that are super exogenous to the output in question. They work without making any assumption as to strong or weak exogeneity (in the sense of Engle, Hendry, and Richard 1983) of the covariates, assumptions that are often provided by basic theory in economic contexts. Implications for further research based on such additional assumptions will be returned to in the sequel.

4 Domain Adaptation

Consider a domain, $\mathcal{D}$, which we define as consisting of a feature space, $\mathcal{X}$, and a marginal distribution $P(X)$ where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. A task, $\mathcal{T}$, consists of an outcome space, $\mathcal{Y}$ and a true generating mechanism $f : \mathcal{X} \to \mathcal{Y}$.

Many traditional machine learning applications involve the assumption that the domain and the task are the same in the training data and the prediction context. Transfer learning is the generic name for all frameworks that work outside these assumptions, when either the domain, the task, or both are different.

I will consider the term domain adaptation in the sense of Ben-David (2006) and Pan and Yang (2010). Under this definition, domain adaptation is a specific form of transfer learning where the task $\mathcal{T}$ is constant and the feature space, $\mathcal{X}$ is the same across all domains. There is a target domain, $\mathcal{D}_T$ from which one has samples $\{x_1, \ldots, x_i\} \in X$ and (one or more) source domain(s), $\mathcal{D}_S$, from which one has tuples $\{(y_1, x_1), \ldots, (y_n, x_n)\} \in (Y, X)$.

The consistency of the task relates to the construct validity in the taxonomy of Shadish, T. D. Cook, and D. T. Campbell 2002. If the outcome measured in one experiment might be considered to measure a different construct than the outcome measured in another experiment, then the task can not be said to be the same and this problem formulation fails.

Allow $\mathcal{X} = \{\mathcal{W}, \mathcal{Z}\}$, where $\mathcal{W} = \{W_0, W_1\}$ a binary treatment and $\mathcal{Z} = Z_1, ..., Z_P$ a set of covariates. The assumption of the consistency of the feature space relies on the construct validity of both the treatment and the covariates. If this validity does not hold, if the covariates or the treatment across domains represent a different construct, then the formulation of domain adaptation does not hold. A formulation in which the feature space is allowed to change would be more appropriate (see ...).

If one considers a random-variable formulation of the true generating mechanism, $f(\cdot)$, then the assumption of task consistency implies the conditional distribution is consistent across tasks, thus $P(Y_S|X_S) = P(Y_T|X_T)$. This assumption is referred to as covariate-shift, as the only difference between domains $\mathcal{D}_S$ and $\mathcal{D}_T$ is that of the marginal covariate distribution, $P(X_S) \neq P(X_T)$.

Under the assumptions of covariate shift, Shimodaira (**Shimodaira2000**) shows that the optimum likelihood function for a maximum likelihood estimator of $P(Y_T|X_T)$, given sufficiently large sample size, is obtained by minimizing the weighted loss function trained on labeled data from the source domain and weighted by the ratio $w(x) = \frac{P_T(x)}{P_S(x)}$:

$$\underset{\beta}{\operatorname{argmin}} \sum_i -w(x_i) \log P_S(y_i|x_i, \beta)$$

The use of this weighting ratio, $w(x)$ (referred to as the *importance*) has led to many other importance estimation techniques to deal with covariate shift, including estimators that don't require the calculation of estimates of the marginal densities which is prohibitive in high-dimensional spaces or with limited data (**foo**).

Thus, the covariate shift problem is relatively well solved. The covariate shift assumption is equivalent to that of all the included covariates, X, being super-exogenous (as in Engle, Hendry, and Richard 1983) or to them making up the autonomous "super-structure" in the terms of Frisch.

Allow $\mathcal{X} = \{\mathcal{W}, \mathcal{Z}, \Lambda\}$, where $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ consists of all unobserved variables for which $P_S(\lambda_i) \neq P_T(\lambda_i)$.

Further, allow for a feature representation function, $g : \mathcal{Z} \to \mathbb{R}^K$

It should be clear that the covariate shift assumption for observed variables amounts to:

5 Experimental Economics Formulation as Domain Adaptation

Banerjee and Duflo (2014) discuss concerns to experimental validity and generalization. As a primary tool to address external validity of RCTs, they emphasize the need for replication studies. Indeed, they posit what can be thought of as asymptotic theory of domain adaptation for RCTs:

> "If we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any given set of covariates."

Asymptotic theory can be comforting, it's nice to know that the road we are on leads somewhere. However, the key term "any given set of covariates" must be restricted for this to be actionable in the finite lifespan of our species. The definition of those covariates comes down to defining the uniformity of nature assumption posed by Hume. If the covariate set is infinite, then we assume nothing regarding the nature of uniformity, and induction is impossible.

How, then, do we define those covariates? Which parts of nature must be uniform and which parts are allowed to change? Our formulation allows us to

This will answer the following practical question for any policy maker attempting to learn from previous studies: how many studies are enough studies? And if the answer to that question depends on where one intends to implement a new policy, then how can one relate the target context to the context of the studies?

Hotz et al (2005) provide one of the only canonical models in the econometrics literature for extrapolating from a source context with experimental data and measuring predictions in a target context where such data is not available. Their technical methodology follows **Abadie2006**, using matching with replacement from the target to the source domain to create a bootstrapped dataset on which to run a regression. This can be seen as a

reinvention of importance estimation (**Shimodaira2000**; **Suigyama2008**). The implication of this choice of techniques is the assumption of covariate shift. The failure of their technique to work in certain contexts implies a violation of the covariate shift assumptions.

Gechter 2015 builds on the technique of Hotz to create bounds for prediction in a new domain, using the difference between outcomes for the control groups $P_S(Y|W_0, Z)$ and $P_T(W_0, Z)$ to create the bounds.

These methodologies are lacking two ingredients

A) they do not incorporate a methodology for discovering heterogeneous treatment effects that can be used to re-marginalize the average treatment effects given a different covariate distribution in the target population and B) they do not provide any formal framework for model selection that determines which covariates to condition on and which to marginalize out to enable a transportable prediction.

6 Current State of the Art in Dealing with External Validity

I will argue that frameworks for discovering heterogeneous treatment effects, such as causal trees (Athey and G. Imbens 2016), must form a part of any methodology that hopes to transfer treatment effects from one domain to another. Following that, I will show that ideas of conditional invariance that can be traced back to mid-century econometricians can motivate the model selection process and that such motivation exists in some recent research in the field of domain adaptation of machine learning (Rojas-Carulla, Schölkopf, and Turner 2018).

Shadish, T. D. Cook, and D. T. Campbell 2002 lay out, along with their taxonomy of validities, a taxonomy of "threats" to each type of validity. While not a formal framework, per say, by creating the taxonomy they invite researchers to address each threat in turn. If a researcher is able to argue that all the threats to external validity have been addressed and protected against, then one might consider the work of generalizing to be finished. In a sense, they provide a checkbox of things that are worth worrying about. Unfortunately, echoing again Manski's ((**Manski2008**) concern, researchers in applied economics have not systematically adopted a system of discussing threats to external validity.

GRADE is an example of this type of system applied in healthcare recommendations. It is a system, formalized to bring together expert opinion. It is not a statistical framework that starts from the original data, it starts

from PDFs, read and interpreted by humans, and lets them give grades to the evidence based on stated criteria.

It provides policy recommendation along the guidelines of patient age, setting, and intervention types.

What is "high and middle- income country" – what happens with countries on the border? this discretization is clearly problematic. But it's a start.

Tamil Nadu / Bangladesh

Hock/Imbens - They conduct a test to see whether Y|X is consistent across locations, but this does not say anything about T|X. Indeed, the Y|X they test is linear, thus by assumption all of those variables fall out of the treatment effect and the identification of those variables is not actually important...

However, if Y|X is consistent in the control, they show that it is also consistent in the treatment, which implies one can compute T|X. By using matching with replacement, they are essentially reweighting the errors by the density ratio of the covariates (show this relation!).

However if Y|X is not consistent, that shows some other latent variable that is effecting Y. This does not mean the latent variable will effect the treatment, however, it might be additive (as was the assumption with all the other variables!).

They seem to have found (through some ad-hoc method) that there is treatment effect heterogeneity among those who have previously worked and those you had not previously worked. Thus, they do the entire analysis separately for both of these groups.

Athey and Imbens (2017) similarly reflect on the recent concerns over external validity quoted previously by the likes of Manski, Deaton, and Heckman. In an article aptly titled *The State of Applied Econometrics: Causality and Policy Evaluation*, they lay out three main recent advances in addressing external validity.

The first advance is that of addressing the concerns about the Local Average Treatment Effect (LATE) measured by instrumental variables (IV) (...). The concerns relate to the generalization of the local effect caused by the variation in the instrument, to the global effect on the entire population (who were potentially not touched by the variation in the instrument in the study).

The second advance is that of addressing concerns regarding the local nature of regression discontinuity designs (...). These techniques all involve various methods to test whether the effect is only present locally at the discontinuity, or whether the effect is likely to extend to the rest of the sample.

Both of these techniques can be thought of addressing "statistical validity" in the validity taxonomy of Shadish, Cook, and Campbell, that of drawing inferences to the population from the sample. They do not address construct or external validity, they do not provide any information to the policy maker who is considering making a decision in another context.

The third advance is more salient to the major thrust of external validity: that of combining observational and experimental results.

Banjerjee and Snowberg (2016) create a formal system of "speculation" to create "falsifiable claims" of research studies as an attempt to codify the process of generalization and external validity that can be attached to any empirical study from the counterfactual paradigm. The falsifiable claims would presumably be used by other researchers to test the assumptions necessary for external validity of the original findings to hold true.

(TODO: something on meta-analysis methods – Meager)

Nancy Cartwright and Jeremy Hardie (2013) lay out an intuitive, qualitative method for predicting the treatment effect of a policy in your context, given evidence from other contexts.

7 Invariance in Domain Adaptation

8 Conclusions and Future Research

Nancy Carwright (**Cartwright2016**) presents an example of a policy decision by policy makers in New York City to implement a new program called Opportunity NYC. The program was modelled after a program in Mexico, Opportunidades, which had proved good results in reducing poverty there. Opportunity NYC was implemented in 2007 and shut down in 2010 due its failure to produce the desired effects.

What went wrong? Should the policy makers in New York simply have known that Mexico is clearly different and no program implemented there should be expected to work in New York? Did the writers of the Opportunidades study have an obligation to investigate and consider the threats to external validity that such a study might have been in danger of?

The purpose of this article was to argue that each of these cases is wrong:

1. Generalization, causal mechanisms, structure, support factors, etc. are all relative to a set of changes to which they must be invariant.

One cannot answer the question about external validity without information about the differences between the source and target contexts.

2. The difference between New York and Mexico must be measured over a well-defined covariate space in order to make sense of the question of whether it is "too different." A core part of applied economics research must, therefore, consist of determining, for a given intervention, what part of nature must be uniform, what makes up the super-structure, what are the support factors that will change, or similarly, what are the invariant mechanisms. New research from machine learning has shown that the discovery of invariant mechanisms, given muti-domain datasets, is feasible and allows for effective predictions in new domains. Thus, while the data requirement has gone up (more than one study is required to make a prediction in a new context without many further assumptions), the ability to give a formal answer to the question of whether it will probably work in New York or whether one has no idea is within our reach and should be pursued.

9

References

[1] By John Aldrich. "By JOHN ALDRICH* 1. Introduction only to those". In: *Oxford economic papers* 41 (1989), pp. 15–34.

[2] Hunt Allcott. "Site Selection Bias in Program Evaluation". In: (2015), pp. 1117–1165. DOI: 10.1093/qje/qjv015.Advance.

[3] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.27 (2016), pp. 7353–7360. ISSN: 10916490. DOI: 10.1073/pnas.1510489113.

[4] Susan Athey and Guido W Imbens. "The State of Applied Econometrics: Causality and Policy Evaluation". In: 31.2 (2017), pp. 3–32.

[5] Abhijit V. Banerjee and Esther Duflo. "The experimental approach to development economics". In: *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences* (2014), pp. 78–114. DOI: 10.1146/annurev.economics.050708.143235.

[6] Abhijit Banerjee, Sylvain Chassang, and Erik Snowberg. “Decision Theoretic Approaches to Experiment Desig”. In: (2016).

[7] Shai Ben-David et al. “Analysis of Representations for Domain Adaptation”. In: (2006).

[8] James Bisbee et al. “Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect”. In: *Journal of Labor Economics* 35.S1 (2017), S99–S147. ISSN: 0734-306X. DOI: 10.1086/691280.

[9] Nancy Cartwright and Jeremy Hardie. “Evidence-based policy: a practical guide to doing it better”. In: *Choice Reviews Online* 50.10 (2013), pp. 50–5831–50–5831. ISSN: 0009-4978. DOI: 10.5860/choice.50-5831.

[10] Thomas D Cook and D T Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. English. Houghton Mifflin, 1979.

[11] Angus Deaton. “Instruments, Randomization, and Learning about Development”. In: *Journal of Economic Literature* 48.2 (June 2010), pp. 424–455. ISSN: 0022-0515. DOI: 10.1257/jel.48.2.424. URL: http://pubs.aeaweb.org/doi/10.1257/jel.48.2.424.

[12] Angus Deaton and Nancy Cartwright. “Understanding and misunderstanding randomized controlled trials”. In: *Social Science & Medicine* 210.October 2017 (2018), pp. 2–21. ISSN: 0277-9536. DOI: 10.1016/j.socscimed.2017.12.005. URL: https://doi.org/10.1016/j.socscimed.2017.12.005.

[13] Robert F Engle, David F Hendry, and Jean-francois Richard. “Exogeneity”. In: *Econometrica* 51.2 (Mar. 1983), p. 277. ISSN: 00129682. DOI: 10.2307/1911990. URL: https://www.jstor.org/stable/1911990?origin=crossref.

[14] Ra Fisher. “The Design of Experiments”. In: (May 1935). ISSN: 0002-8312.

[15] Ragnar Frisch. *The Foundations of Econometric Analysis*. Ed. by David F. Hendry and Mary S. Morgan. Cambridge: Cambridge University Press, 1995, pp. 407–419. ISBN: 9781139170116. DOI: 10.1017/CBO9781139170116. URL: http://ebooks.cambridge.org/ref/id/CBO9781139170116.

[16] Michael Gechter. “Generalizing the Results from Social Experiments : Theory and Evidence from Mexico and India”. In: 2008 (2015), pp. 1–50.

[17] Trygve Haavelmo. “The Probability Approach in Econometrics”. In: *Econometrica* 12 (July 1944), p. iii. ISSN: 00129682. DOI: 10.2307/1906935. URL: https://www.jstor.org/stable/1906935?origin=crossref.

[18] James J Heckman. “Econometric Causality”. In: (2008).

[19] Christina Heinze-deml and Nicolai Meinshausen. “Conditional Variance Penalties and Domain Shift Robustness”. In: 2009 (2017). arXiv: arXiv:1710.11469v5.

[20] Paul W Holland. “Statistics and Causal Inference: Rejoinder”. In: *J Am Stat Assoc* 81.396 (1986), p. 968. ISSN: 0162-1459. DOI: 10.2307/2289069.

[21] V. Joseph Hotz, Guido W. Imbens, and Julie H. Mortimer. “Predicting the efficacy of future training programs using past experiences at other locations”. In: *Journal of Econometrics* 125.1-2 SPEC. ISS. (2005), pp. 241–270. ISSN: 03044076. DOI: 10.1016/j.jeconom.2004.04.009.

[22] Leonid Hurwicz. “On the Structural Form of Interdependent Systems”. In: *Studies in Logic*. Vol. 44. Board of Trustees of the Leland Stanford Junior University, 1966, pp. 232–239. DOI: 10.1016/S0049-237X(09)70590-7. URL: http://dx.doi.org/10.1016/S0049-237X(09)70590-7%20https://linkinghub.elsevier.com/retrieve/pii/S0049237X09705907.

[23] Charles F. Manski. “Public Policy in an Uncertain World”. In: *Public Policy in an Uncertain World* (2013). DOI: 10.4159/harvard.9780674067547.

[24] Sinno Jialin Pan and Qiang Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (Oct. 2010), pp. 1345–1359. ISSN: 1041-4347. DOI: 10.1109/TKDE.2009.191. arXiv: 1603.06111. URL: https://doi.org/10.1016/j.artmed.2018.03.006%7B%5C%%7D0Ahttp://arxiv.org/abs/1801.06146%7B%5C%%7D0Ahttp://arxiv.org/abs/1902.01382%7B%5C%%7D0Ahttp://arxiv.org/abs/1902.09092%7B%5C%%7D0Ahttp://arxiv.org/abs/1703.06345%7B%5C%%7D0Ahttp://dblp.org/rec/conf/aaai/ZhouPTH16%7B%5C%%7D0Awww.aaai.org%7B%5C%%7D0Ahttp://arxiv.or%20http://ieeexplore.ieee.org/document/5288526/.

[25] Judea Pearl. *Causality Second Edition*. 2000, p. 386. ISBN: 0521773628. DOI: citeulike-article-id:3888442.

[26] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. “Causal inference using invariant prediction : identification and confidence intervals”. In: (2015), pp. 1–51. arXiv: arXiv:1501.01332v3.

[27] Lant Pritchett and Justin Sandefur. "Learning from Experiments when Context Matters". In: *American Economic Review* 105.5 (May 2015), pp. 471–475. ISSN: 0002-8282. DOI: 10.1257/aer.p20151016. URL: http://eds.b.ebscohost.com.ezproxy.lib.usf.edu/eds/pdfviewer/pdfviewer?vid=5%7B%5C&%7Dsid=aef21225-20b8-458e-842b-e860243ea5f7%7B%5C%%7D40sessionmgr104%7B%5C&%7Dhid=104%20http://pubs.aeaweb.org/doi/10.1257/aer.p20151016.

[28] Mateo Rojas-Carulla, Bernhard Schölkopf, and Richard Turner. "Invariant Models for Causal Transfer Learning". In: 19 (2018), pp. 1–34.

[29] Paul R. Rosenbaum. "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies". In: *American Statistician* 59.2 (2005), pp. 147–152. ISSN: 00031305. DOI: 10.1198/000313005X42831.

[30] Mark R. Rosenzweig and Christopher Udry. "External Validity in a Stochastic World: Evidence from Low-Income Countries". In: (2019).

[31] William R Shadish, Thomas D Cook, and Donald T. Campbell. *Experimental and Designs for Generalized Causal Inference*. 2002. ISBN: 0395615569.