Ratio Mathematica Volume 36, 2019, pp. 5-26

How to define and test explanations in populations

Peter J. Veazie*

Abstract

Solving applied social, economic, psychological, health care and public health problems can require an understanding of facts or phenomena related to populations of interest. Therefore, it can be useful to test whether an explanation of a phenomenon holds in a population. However, different definitions for the phrase "explain in a population" lead to different interpretations and methods of testing. In this paper, I present two definitions: The first is based on the number of members in the population that conform to the explanation's implications; the second is based on the total magnitude of explanation-consistent effects in the population. I show that claims based on either definition can be tested using random coefficient models, but claims based on the second definition can also be tested using the more common, and simpler, population-level regression models. Consequently, this paper provides an understanding of the type of explanatory claims these common methods can test.

Keywords: Explanation, statistical testing, population regression models, random coefficient models, mixture models

2010 AMS subject classification: 62A01; 62F03.†

* University of Rochester, Rochester NY, USA; peter_veazie@urmc.rochester.edu.
† Received on February 12th, 2019. Accepted on May 3rd, 2019. Published on June 30th, 2019. doi: 10.23755/rm.v36i1.463. ISSN: 1592-7415. eISSN: 2282-8214.

1. Introduction

Science provides explanations for facts, phenomena, and other explanations. In applied research that draws on theories from disciplines such as Economics, Psychology, Sociology, and Organizational Science, among others, this can require testing whether a proposed explanation explains a given fact, phenomenon, and other explanation in a specified population. For example, one might wish to test whether a proposed explanation based on Psychology's Regulatory Focus Theory [1, 2] explains physician risk tolerance in treatment choice (the phenomenon) among primary care physicians in the United States (the population). However, what is meant by the phrase *explains in a population*? Is it that the proposed explanation accounts for the behavior of every member of the population? This is a high bar: one member of the population for whom the explanation does not hold falsifies the claim. Is it that the proposed explanation accounts for the behavior of at least one member? This is equally extreme: only one member of a population for whom the explanation holds warrants the claim. The claim is ambiguous. Specific definitions are required if such claims are to be understood and tested.

This paper provides definitions and identifies methods for testing corresponding explanatory claims. These definitions and the identification of corresponding methods are new contributions that provide conceptual and methodological guidance for researchers who seek to test explanations in populations. The methods themselves, however, are in common use: random coefficient models and population-level regression models. Therefore, whereas a goal of this paper is to show which methods can be used to test specific explanatory claims, I do not present the implementation of the methods: there are many textbooks and articles that provide this information [e.g. 3, 4]. For simplicity of presentation, I only reference phenomena as the target of explanation rather than also facts and other explanations; however, any of these are applicable throughout.

2. Defining *explain*

Before providing the required definitions, I will clarify what I mean by *to explain* and by *an explanation*. For this paper, to explain something is to provide a way of understanding it through a conceptual structure that accounts, at least in part, for that which is being explained [5, Ch. 9]. The conceptual structure is the explanation. One might imagine there is a single explanation for any given phenomenon. However, for macro-level phenomena, such as organization and human behaviors, there may be multiple ways of understanding them. For example, a human behavioral phenomenon may have sociological explanations, psychological explanations, physiological explanations, and more. Any one of the explanations could be referred to as

an explanation, and no one of them referred to as exclusively *the explanation*. Moreover, an explanation need not be complete. There may be many causal factors or mechanisms that contribute to the phenomenon; however, an explanation might focus only on a subset.

An explanation can be intended to provide an understanding of a phenomenon as it is [6, Ch. 4], a *de re* explanation; or, it can be intended to provide an understanding that, nonetheless, contains explicitly presumed falsehoods [7, 8], a *de ficta* explanation. All terms of a *de re* explanation refer to presumed real objects, qualities, characteristics, and relationships. Designation as a *de re* explanation does not guarantee truth, nor does it imply the researcher believes it is true; indeed, if the researcher believed the explanation was in fact true, there is no need for further inquiry [9]. Moreover, it is common to expect even a well-established theory-based explanation to be incorrect in some unknown way. It is the ontological commitments (the presumption that explanatory terms intend to have real referents) of the explanation's terms that qualify it as a *de re* explanation. However, a *de ficta* explanation contains at least one identified term that is presumed to be false. These are often explanations that contain idealizations (e.g. the discrete energy levels in the Bohr model of the atom [10-12], and the rationality of the rational choice model in classic microeconomics [13, 14]) or analogies (e.g. the computer analogy or corporate analogy of information processing in cognitive science [15]). Given there need only be a single presumed false term to warrant designation as a *de ficta* explanation, the remaining terms have substantive ontological commitments. Such *de ficta* explanations are presumed to be partially true [7]. Although these definitions do not restrict explanations to those that are amenable to empirical investigation, this paper is written to provide guidance for empirical researchers. Consequently, the focus of the discussion herein is on scientific explanations that have empirical implications.

In the applied sciences, the goal of both *de re* and *de ficta* explanations is to guide interventions, actions, or policy. The pursuit and use of a *de re* explanation are based on the belief that understanding the world as it is provides assurance that consequent interventions, actions, and policies are more likely to work and generalize, and the causes for their failure are more likely to be identified. The *de ficta* explanation does not carry as great an assurance in these regards as it includes identified false claims. However, the *de ficta* explanation can be simpler, easier to develop and understand, and easier to apply. Both types of explanation are usefully employed.

Explanations are often assessed in terms of explanatory power. Explanatory power characterizes explanations in terms of explanatory virtues such as generality, coherence, accuracy, and predictive ability, among others [8, 16]. It has been qualitatively defined in terms of the scope of questions it

can address [16], and it has been the basis for formal probability-based measures [17-20]. However, for the purposes of applied science another aspect of power can be useful: effective power.

Applied researchers often focus on the ability to influence specific outcomes and therefore seek explanations to inform actions that can produce specific effects. For example, researchers may seek to reduce systolic blood pressure, decrease expected expenditures, or expand social networks rather than seek to account for variation. To achieve such goals, it can be important to assess a phenomenon's responsiveness to an explanation, its effective power. Effective power is different from accuracy and predictive power (the abilities to account for and predict phenomena and behavior). Consider an explanation of the relationship between behavior Y and explanatory factor X for two individuals w and v. Suppose the effect of the explanation on Y can be modeled as a simple linear function of X with a positive coefficient, in which variable X completely determines Y for individual w and only partially determines Y for individual v:

$$Y_w = \beta_w \cdot X_w$$

and

$$Y_v = \beta_v \cdot X_v + E_v.$$

The predictive power for w is greater than that for v; indeed, the predictive power for w is perfect, whereas it is only partial for v, due to the additional term E_v. However, if $\beta_w = \beta_v$, then variable X has the same relationship with behavior Y for both and thereby having the same effective power: a difference in X corresponds to the same difference in Y for both w and v. If $\beta_v > \beta_w$, then the explanation has greater effective power for v, even though it has greater predictive power for w. Effective power represents the responsiveness to the explanation whereas accuracy and predictive power represents the extent of Y accounted for by the explanation. As an analogy, consider a regression analysis, in the above example effective power is analogous to β and predictive power is analogous to the coefficient of determination (commonly termed R-square) or an out-of-sample prediction metric. Like Schupbach and Sprenger's [18] definition of explanatory power, effective power can be negative for a proposed explanation, if the response is counter to that implied by the explanation: for example, the case in which the β's in the preceding example were in fact negative, contrary to the explanatory implication of positive β's.

We can understand a population-level *de re* or *de ficta* explanatory claim as a reductive explanation: an explanation that applies to a population in virtue of an aggregation of the explanation's application to its members. This is kin to what Strevens terms an aggregative explanation [8]. For example, where I

may seek to explain physician risk tolerance in treatment choice among primary care physicians in the United States, the proposed explanation is regarding its members' relevant behaviors (the behaviors of individual physicians). So, regardless of the number of members in the population, which can be as few as one, our definition of the phrase *a potential explanation explains a given phenomenon in a population* represents an aggregation of an individual-level explanation across the members of the population.

As stated in the introduction, definitions that require explanation of either every member or only one member of a population are extreme. Appropriate definitions are likely somewhere in between. This paper focuses on two:

Definition 1. An explanation explains a phenomenon in a population if, and only if, it has positive effective power for most members of the population.

Definition 2. An explanation explains a phenomenon in a population if, and only if, its cumulative magnitudes of effective power among the members of the population for whom the explanation holds exceeds its cumulative magnitudes of effective power among the members of the population for whom the explanation does not hold.

These definitions are based on minimal criteria. In the first case, it would be difficult to support an explanatory claim regarding scope if the possible explanation only applied to a minority of population members. In the second case, it would be difficult to support an explanatory claim regarding cumulative power if the possible explanation was associated with less cumulative power than the counter-explanation in a population. However, this is arbitrary, and we need not take the minimal stance. We can generalize the definitions to vary with a definitional parameter q:

General Definition 1. An explanation explains a phenomenon in a population if, and only if, it has effective power for at least q percent of the members of the population.

General Definition 2. An explanation explains a phenomenon in a population if, and only if, its cumulative magnitudes of effective power among the members of the population for whom the explanation holds exceeds q times its cumulative magnitudes of effective power among the members of the population for whom the explanation does not hold.

The remaining sections focus on the minimal definitions, however the general testing method in Section 4.1 can be used to test these general definitions as well.

3. Defining Testable Implications

To test claims based on the preceding definitions, we required corresponding operational definitions in terms of testable implications:

Operational Definition 1. If an explanation explains a phenomenon in a population, then the implications of the explanation hold for most of the members of the population. And, under reasonable presumption (i.e. credible alternative explanations are accounted for), if the implications of the explanation hold for most of the members of the population, then an explanation explains a phenomenon in a population.

Operational Definition 2. If an explanation explains a phenomenon in a population, then the cumulative strength of the explanation's implications among the members of the population for whom the explanation holds exceeds the cumulative strength of the counter-implications among the members of the population for whom the explanation does not hold. And, under reasonable presumption (i.e. credible alternative explanations are accounted for), if the cumulative strength of the explanation's implications among the members of the population for whom the explanation holds exceeds the cumulative strength of the counter-implications among the members of the population for whom the explanation does not hold, then an explanation explains a phenomenon in a population.

The first conditional in each operational definition allows evidence against each consequent (the testable implications) to provide evidence against the explanatory claim. The second conditional allows evidence for each antecedent (the testable implications) to provide evidence for the explanatory claim. The first conditionals are typically derived from the explanation. The second conditionals draw more upon the weaker condition of presumption-based reasoning [21], which is grounded in current background knowledge and is thereby defeasible: future changes in scientific understanding can negate the conditional. A strong reasonable presumption for the second conditionals is achieved if there are no credible alternative explanations for the testable implications.

Regarding operational definition 1, we might say, for example, that a Regulatory-Focus-Theory-based explanation explains physician risk tolerance in treatment choice among primary care physicians in the United States if a higher promotion focus (a term in Regulatory Focus Theory [1, 22]) leads physicians to have higher risk tolerance (the explanation's implication) for more than half of the physicians, accounting for alternative explanations. Regarding operational definition 2, we might say that a Regulatory-Focus-Theory-based explanation explains physician risk tolerance in treatment choice among primary care physicians in the United States if the cumulative

magnitudes of effect of promotion focus on risk tolerance among physicians for whom a higher promotion focus leads the physician to have higher risk tolerance exceeds the cumulative magnitudes of effect of promotion focus on risk tolerance among physicians for whom a higher promotion focus leads the physician to have lower risk tolerance (or no relationship).

We can generalize the operational definitions, as we did with the original definitions, to vary with a definitional parameter *q*:

General Operational Definition 1. If an explanation explains a phenomenon in a population, then the implications of the explanation hold for *q* percent of the members of the population. And, under reasonable presumption (i.e. credible alternative explanations are accounted for), if the implications of the explanation hold for *q* percent of the members of the population, then an explanation explains a phenomenon in a population

General Operational Definition 2. If an explanation explains a phenomenon in a population, then the cumulative strength of the explanation's implications among the members of the population for whom the explanation holds exceeds *q* times the cumulative strength of the counter-implications among the members of the population for whom the explanation does not hold. And, under reasonable presumption (i.e. credible alternative explanations are accounted for), if the cumulative strength of the explanation's implications among the members of the population for whom the explanation holds exceeds *q* times the cumulative strength of the counter-implications among the members of the population for whom the explanation does not hold, then an explanation explains a phenomenon in a population.

To test claims based on the preceding definitions, we start by identifying the proposed explanation's implications. Specifically, we presume an explanation-implied relationships *g* between variables *Y* and *X* (as defined in the context of the phenomenon and explanation), with parameter θ:

$$y = g(x;\theta) \text{ , such that } \frac{\partial g(x;\theta)}{\partial x} \in \mathbb{D}_e \text{ , } \forall x \in \mathbb{R}_X . \quad (1)$$

This is to say that we have a proposed explanation *e* of a phenomenon that implies variables *X* and *Y* are related by some, perhaps unknown, function *g* such that for all values *x* in range $\mathbb{R}_X$ the derivative of *g* with respect to *x* (or the difference quotient if $\mathbb{R}_X$ is a discrete set) is in the set $\mathbb{D}_e$. Note that the implications can be more general: The $\frac{\partial g(x;\theta)}{\partial x}$ term can be a vector of derivatives across multiple *X* variables. And, the implications for any given derivative can be multi-part, having different ranges for the derivative across

different *x*-values. However, for ease of presentation this paper focuses on single-part implications.

A simple example is *g* specified as a linear relationship, $y = \alpha + \beta \cdot x$, such that the proposed explanation *e* implies $dy/dx > 0$, i.e. $\mathbb{D}_e = (0,\infty)$, for all positive values of *X*, i.e. $\mathbb{R}_X = (0,\infty)$. Applying this equation to all members of Ω, we can say that if β is positive for most members of a population Ω, then *e* explains by definition 1. If the sum of the magnitude of β's across all members of Ω for whom $\beta>0$ exceeds the sum of the magnitudes of β's across all members for whom $\beta\leq 0$, then *e* explains by definition 2.

To formalize the concept of *explain*, consider the following variable Δ defined for $w \in \Omega$ and $x \in \mathbb{R}_X$:

$$\Delta(w,x) = h\left(\frac{\partial g(x;\Theta(w))}{\partial x}\right). \tag{2}$$

The function *h* provides the relevant interpretation for *explain*. The two functions considered in this paper for *h* provide interpretations for *explain* as the scope of the explanation (definition 1 above) and as the power of the explanation (definition 2 above). These are detailed below.

We can use two functions to separate the Δ's into groups. The first picks out Δ for the explanation-implied range of values for $\partial g/\partial x$, and the second picks out Δ for the range of values outside of the explanation-implied range—the counter-explanation range:

$$\Delta^{+}(w,x) = \begin{cases} \Delta(w,x) & \text{if } \dfrac{\partial g(x;\Theta(w))}{\partial x} \in \mathbb{D}_e \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

and

$$\Delta^{-}(w,x) = \begin{cases} \Delta(w,x) & \text{if } \dfrac{\partial g(x;\Theta(w))}{\partial x} \notin \mathbb{D}_e \\ 0 & \text{Otherwise} \end{cases}. \tag{4}$$

The sum of the magnitudes of Δ^+ across population Ω at value *x* reflects the extent of the proposed explanation's implications in the population at *x* (the interpretation depending on *h*). The sum of the magnitudes of Δ^- across population Ω at value *x* reflects the extent of counter-explanation implications in the population at *x*.

For both specifications of *h* discussed below, a useful formalization of *explain* is to say that the proposed explanation explains a phenomenon in a population if the accumulated magnitudes of Δ is larger in the explanation-

implied region than in the counter-explanation region for all points in a specified set *B* of *x*-values. For arbitrary value *x* in *B*, this implies for both definitions 1 and 2 that

$$\sum_{w\in\{w:X(w)=x\}}\left(\left|\Delta^{+}(w,x)\right|\right) > \sum_{w\in\{w:X(w)=x\}}\left(\left|\Delta^{-}(w,x)\right|\right). \tag{5}$$

For the generalized definitions this is

$$\sum_{w\in\{w:X(w)=x\}}\left(\left|\Delta^{+}(w,x)\right|\right) > q^{\circ}\cdot\sum_{w\in\{w:X(w)=x\}}\left(\left|\Delta^{-}(w,x)\right|\right), \tag{6}$$

where $q^{\circ} = q/(100 - q)$ for generalized definition 1, and $q^{\circ} = q$ for generalized definition 2.

Denoting the statement *e explains p in* Ω *on set B* as $E(e, p, \Omega, B)$, the corresponding claims are $E(e, p, \Omega, B) = True$ and $E(e, p, \Omega, B) = False$. The claim that the proposed explanation holds (i.e. $E(e, p, \Omega, B) = True$) is asserted if for all points *x* in the set *B* the proposed explanation's implication exceeds that for the counter-explanation implication. The claim that the proposed explanation does not hold (i.e. $E(e, p, \Omega, B) = False$) is asserted if there exists at least one point in *B* for which the counter-explanation implication exceeds the proposed explanation's implication.

It is useful to take *B* to be one of two sets: either a singleton $\{x\}$ or the phenomenologically-relevant range $\mathbb{R}_X$. Claims $E(e, p, \Omega, \mathbb{R}_X)$ are what we may consider when testing whether a proposed explanation explains, whereas point-wise claims $E(e, p, \Omega, \{x\})$ are useful in understanding where in the range of *x*-values the claims $E(e, p, \Omega, \mathbb{R}_X)$ fail, if indeed they fail, or at which points of *X* is the underlying proposed explanation is either least or most powerful. There are occasions, however, when $E(e, p, \Omega, \mathbb{R}_X)$ is too strict: do we really want to say a proposed explanation does not explain in a population because it doesn't hold at a single point *x*? For example, suppose economic demand follows the predicted relationship with price at all prices except at $1, do we say the price-demand theory does not hold in the population because of this singular exception? Perhaps we should account for how important it is that the explanation hold at $1, or account for how many people face a price of $1 for the good being considered. We can address these concerns by taking a weighted average of *x*-specific effects across the range of *x*-values in $\mathbb{R}_X$ using a probability distribution for *X* conditional on $x \in \mathbb{R}_X$. Denoting this general explanatory claim as $E(e, p, \Omega)$, it requires the weighted sum across all *x*-values being considered and thereby can balance non-explanatory points of $\mathbb{R}_X$ with other strongly explanatory points. Its interpretation depends on

the definition of the probability for X [23]. For example, it can be helpful to consider claims regarding $E(e, p, \Omega)$ in terms of random variables defined on population Ω, with equal probabilities assigned to each member of Ω. Using Ω as its domain, the variable X provides the value x that each member is facing. The probability distribution of X therefore represents the actual normalized frequency of X in the population, and consequently $E(e, p, \Omega)$ is based on the corresponding weighted average across this distribution.

Figure 1 presents an example in which the explanation implies negative derivatives of g with respect to x, i.e. $\mathbb{D}_e = (-\infty, 0)$ for all values of x in $\mathbb{R}_X$, but for which the actual g is as shown. It is clear, regarding the point-wise explanations, that the claim $E(e, p, \Omega, \{x\}) = True$ holds true only for x less than x^*, but $E(e, p, \Omega, \{x\}) = False$ for all x greater than x^*. Consequently, due to the existing values of X for which the explanatory implications do not hold (i.e. for $x > x^*$), the overall claim is therefore $E(e, p, \Omega, \mathbb{R}_X) = False$. On the other hand, for $f(x)$ denoting the density of X based on $P(x \mid x \in \mathbb{R}_X)$, the general claim weighted by this probability is $E(e, p, \Omega) = True$ as there is little probability associated with x-values in the contra-explanatory range of derivatives.

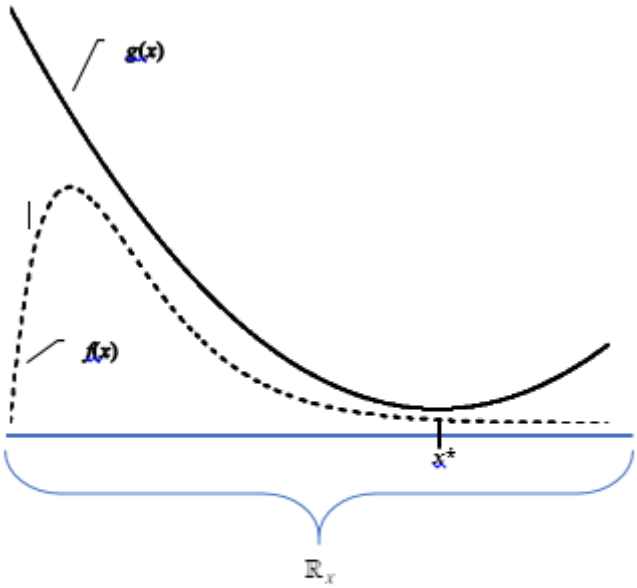

Figure 1. Example of $E(e, p, \Omega, \mathbb{R}_X) = No$ because dg/dx > 0 for some x in $\mathbb{R}_X$ (i.e. for $x > x^*$), and $E(e, p, \Omega) = Yes$ because the density f weights dg/dx heavily in the explanation-consistent region (i.e. where dg/dx < 0) and trivially in the non-explanation consistent region (i.e. where dg/dx > 0).

As mentioned above, two specifications for *h* are considered here. The first, for definition 1, specifies *h* as a constant function with value 1:

$$\Delta(w, x) = 1, \text{ for all } w \text{ and } x. \tag{7}$$

This leads to

$$\Delta^{+}(w, x) = \begin{cases} 1 \text{ if } \dfrac{\partial g(x; \Theta(w))}{\partial x} \in \mathbb{D}_e \\ 0 \quad \text{Otherwise} \end{cases} \tag{8}$$

and

$$\Delta^{-}(w, x) = \begin{cases} 1 \text{ if } \dfrac{\partial g(x; \Theta(w))}{\partial x} \notin \mathbb{D}_e \\ 0 \quad \text{Otherwise} \end{cases}. \tag{9}$$

By this definition, the sum of the absolute values of Δ^{+} is the number of people whose *X* and *Y* relationship follows the proposed explanation's prediction at specified *x*-values. The sum of absolute value of Δ^{-} is the number of people whose *X* and *Y* relationship do not follow the proposed explanation's prediction. A proposed explanation explains at *x*, by equation 5, if more people in the population follow the prediction than do not when $X = x$.

The second specification, which is used for definition 2, is to define *h* as the identity function, and therefore Δ is

$$\Delta(w, x) = \frac{\partial g(x; \Theta(w))}{\partial x}. \tag{10}$$

This leads to

$$\Delta^{+}(w, x) = \begin{cases} \dfrac{\partial g(x; \Theta(w))}{\partial x} \text{ if } \dfrac{\partial g(x; \Theta(w))}{\partial x} \in D_e \\ 0 \quad \text{Otherwise} \end{cases} \tag{11}$$

and

$$\Delta^{-}(w, x) = \begin{cases} \dfrac{\partial g(x; \Theta(w))}{\partial x} \text{ if } \dfrac{\partial g(x; \Theta(w))}{\partial x} \notin D_e \\ 0 \quad \text{Otherwise} \end{cases}. \tag{12}$$

The corresponding definition for explain compares the accumulated magnitudes of Δ between the explanation-implied region and the counter-

explanation region, which reflects the cumulative effective power of the explanation in the population.

The difference between these two corresponding specifications for *h* is that the first claim, E_1, focuses on the scope (the number or proportion of the population consistent with the explanation), whereas the second claim, E_2, focuses on the cumulative power of the explanation. It is possible for an explanation to apply to a minority of people in the population, but it does so with greater strength in the magnitude of Δ among this minority than is the magnitude of Δ for the majority, who are not in the implied region. In this case the explanation would be considered as explaining in terms of E_2, which uses the identity function for *h*, but not in terms of E_1, which uses the constant function for *h*. On the other hand, in the case where a majority has only a tiny magnitude of Δ in the implied region but a minority has a large magnitude of Δ in the non-implied region, the explanation would be considered as explaining in terms of E_1 but not in terms of E_2. This is analogous to considering the importance of whether a treatment has a larger total positive effect among those that benefit relative to the total negative affect among those who do not benefit (E_2), or whether the treatment simply positively affects a greater proportion of people regardless of how small the effect (E_1). Which definition is appropriate depends on the research goal.

These definitions are population-specific. Consequently, it is possible for a proposed explanation to explain in one population but not another. Moreover, it is possible to not explain in a population but to explain in one of its subpopulations, and vice versa. Consider a population Ω made up of two subpopulations Ω_1 and Ω_2: it is possible for $E(e, p, \Omega, \mathbb{R}_X) = False$, and yet $E(e, p, \Omega_1, \mathbb{R}_X) = True$. This is often the advantage of doing subgroup analysis, to determine if a proposed explanation holds better in one group than another. Indeed, the primary scientific aim of a study may be to identify for which population the proposed explanation holds.

4. Testing explanations

4.1 General tests using random coefficient models

How do we empirically test a hypothesis of the form $E(e, p, \Omega, \mathbb{R}_X) = True$ or $E(e, p, \Omega, \mathbb{R}_X) = False$? A general approach is conceptually straightforward, albeit empirically challenging. This approach is based on the idea that if we can estimate the distribution of Δ, we can estimate the conditions for $E(e, p, \Omega, \mathbb{R}_X) = True$ and $E(e, p, \Omega, \mathbb{R}_X) = False$. To estimate

the distribution of Δ, assuming our data generating process can support it, we can use a random coefficient model [3].

Suppose we define random variables (or random vectors) *Y*, *X*, Θ, and $\mathcal{E}$ on the population Ω, representing a population model such that

$$Y(w) = g(X(w); \Theta(w)) + \mathcal{E}(w), \text{ for } w \in \Omega. \tag{13}$$

If we have a data generating process with *N* observations, $i \in \{1, \ldots N\}$, we can consider the mixture model for the regression of *Y* on *X*:

$$E(Y_i \mid x_i) = \int E(Y_i \mid x_i, \theta_i) \cdot dF(\theta_i \mid x_i). \tag{14}$$

Substituting equation 13 for Y_i on the right-hand side of equation 14, yields

$$E(Y_i \mid x_i) = \int g(x_i, \theta_i) \cdot dF(\theta_i \mid x_i) + \int E(\mathcal{E}_i \mid x_i, \theta_i) \cdot dF(\theta_i \mid x_i), \tag{15}$$

which is the expected value of g plus the expected value of $\mathcal{E}$, each conditioned on $X = x$:

$$E(Y_i \mid x_i) = \int g(x_i, \theta_i) \cdot dF(\theta_i \mid x_i) + E(\mathcal{E}_i \mid x_i). \tag{16}$$

Under the assumption that the expected value of the error terms is 0 for all values of *X*, the regression is

$$E(Y_i \mid x_i) = \int g(x_i, \theta_i) \cdot dF(\theta_i \mid x_i). \tag{17}$$

The derivative of *g* and the estimated distribution for *F* can be used to obtain a distribution for Δ and thereby estimate the conditions for the explanation to hold. Notice, however, from equation 17 the function *g* must be the expected value of *Y* conditional on values of *X* and Θ, i.e. equation 14. Consequently, if a statistically adequate model [24] for $E(Y_i \mid x_i, \theta_i)$ can be empirically determined, an explicit a priori specification for *g* is not required, only hypotheses regarding implications (e.g. derivatives or difference quotients) are required a priori.

Estimation can be achieved using a mixture model, or random parameters model, if the study design and context allow for estimation of such a model. It is best to use a non-parametric estimator for $F(\theta \mid x)$ since results in this case are likely to be very sensitive to the distribution (we are integrating under different regions of the distribution, rather than merely estimating parameters of the distribution). For example, we may consider using Fox et al's non-parametric estimator for the distribution of random effects [25, 26].

Suppose we can assume the error term is independent of *X* and that we have a relationship such that

$$g(x_i, \theta_i) = e^{x_i \cdot \theta_i}, \tag{18}$$

which has the derivative

$$\frac{dg(x_i, \theta_i)}{dx} = \theta_i \cdot e^{x_i \cdot \theta_i}. \tag{19}$$

The expected value of *Y* conditional on *X* is

$$E(Y_i \mid x_i) = \int e^{(x_i \cdot \theta_i)} \cdot dF(\theta_i \mid x_i). \tag{20}$$

With an estimator for *F*, denoted as $\hat{F}$, we can estimate, using numeric integration, the population proportion of those whose derivative falls in the explanation-implied range for any *x*,

$$\hat{p}(x) = \int I(\theta \cdot e^{\theta \cdot x} > 0) \cdot d\hat{F}(\theta \mid x), \tag{21}$$

in which $I(\cdot)$ is an indicator function returning 1 if its argument is true, 0 otherwise. Equation 21 can be used to test *E1*.

For the general *E1*, based on the population distribution for *X* and representative sampling, we would average estimates from equation 21 for each observation in the data to obtain

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}(x_i). \tag{22}$$

In this case, because $e^{x \cdot \theta}$ is always positive, the sign of the derivative is determined by the sign of θ. Therefore, we can estimate $\hat{p}$ based solely on an indicator of $\theta > 0$:

$$\hat{p}(x) = \int I(\theta > 0) \cdot d\hat{F}(\theta \mid x). \tag{23}$$

If we can assume the distribution *F* is independent of *x*, i.e. $F(\theta \mid x) = F(\theta)$ for all *x*, then $\hat{p}$ is not a function of *x*, and $\hat{p}(x)$ is the same for all *x*; therefore

$$\hat{p} = \int I(\theta > 0) \cdot d\hat{F}(\theta) = 1 - \hat{F}(0). \tag{24}$$

In this case we can base our test on $1 - \hat{F}(0)$. Using a bootstrap distribution for $\hat{p}$ (for either equation 23 or equation 24), if a legitimate bootstrap method applies [27], we can test whether *E1* is the case using the p-value $P(p \geq \hat{p} \mid p = 0.5)$ if $\hat{p} \geq 0.5$, and p-value $P(p \leq \hat{p} \mid p = 0.5)$ if $\hat{p} \leq 0.5$ [28].

For testing *E2* at specific *x*-values we calculate

$$c(x) = \int \left[\left(I(\theta > 0) \cdot \left| \theta \cdot e^{\theta \cdot x} \right| \right) - \left(I(\theta \leq 0) \cdot \left| \theta \cdot e^{\theta \cdot x} \right| \right) \right] \cdot d\hat{F}(\theta). \qquad (25)$$

For testing the general E_2 we average $c(x)$ across the data. Again, we can use the bootstrap distribution for F to obtain p-values $P(c \geq \hat{c} \mid c = 0)$ or $P(c \leq \hat{c} \mid c = 0)$.

4.2 Testing E_2 using population-level regression models

The preceding method, which uses random coefficient models and numeric integration, is complicated—particularly for E_2, which represents definition 2. We can greatly simplify our method for testing E_2, if the explanation's implications are regarding positive vs non-positive (or negative vs non-negative) derivatives. In this case, with an additional statistical assumption, we can use population-level regression models to test the explanation. The argument is as follows: As above, we say that *e* explains phenomenon *p* at *x* if inequality 5 holds. Under the definition for E_2, in the case of $\mathbb{D}_e$ being either positive, negative, non-positive or non-negative, the absolute values can be moved outside of the summations,

$$\left| \sum_{w \in (w:X(w)=x\}} \left(\Delta^{+}(w,x) \right) \right| > \left| \sum_{w \in (w:X(w)=x\}} \left(\Delta^{-}(w,x) \right) \right|. \qquad (26)$$

Consider $\mathbb{D}_e = (0,\infty)$, i.e. the explanation implies positive derivatives. In this case, for the left-hand side of inequality 26 the summation of the Δ^+ across the population with $X = x$ is the same as the summation of the product of each Δ-value and its frequency for Δ-values greater than 0:

$$\sum_{w \in (w:X(w)=x\}} \left(\Delta^{+}(w,x) \right) = \sum_{\Delta > 0} \Delta \cdot Freq(\Delta \mid x). \qquad (27)$$

Similarly, regarding Δ^-,

$$\sum_{w \in (w:X(w)=x\}} \left(\Delta^{-}(w,x) \right) = \sum_{\Delta \leq 0} \Delta \cdot Freq(\Delta \mid x). \qquad (28)$$

Therefore, to determine E_2 we can consider whether

$$\left| \sum_{\Delta > 0} \Delta \cdot Freq(\Delta \mid x) \right| > \left| \sum_{\Delta \leq 0} \Delta \cdot Freq(\Delta \mid x) \right|. \qquad (29)$$

However, the inequality remains true if both sides are multiplied by the same positive constant. So, if we multiply by $1/N_x$, denoting the inverse of the population size with value $X = x$, then

$$\left|\sum_{\Delta>0}\Delta\cdot\frac{Freq(\Delta\mid x)}{N_x}\right|>\left|\sum_{\Delta\le 0}\Delta\cdot\frac{Freq(\Delta\mid x)}{N_x}\right|, \tag{30}$$

which is

$$\left|\sum_{\Delta>0}\Delta\cdot f(\Delta\mid x)\right|>\left|\sum_{\Delta\le 0}\Delta\cdot f(\Delta\mid x)\right| \tag{31}$$

for f denoting a probability mass function (however, the above logic and derivation also applies to Δ as a continuous variable in which f is a density, and the summation is replaced with an integral).

Multiplying the left side of inequality 31 by 1 written as

$$\frac{P(\Delta>0\mid x)}{P(\Delta>0\mid x)},$$

and multiplying the right side by 1 written as

$$\frac{P(\Delta\le 0\mid x)}{P(\Delta\le 0\mid x)},$$

yields

$$\left|\sum_{\Delta>0}\Delta\cdot f(\Delta\mid x)\cdot\frac{P(\Delta>0\mid x)}{P(\Delta>0\mid x)}\right|>\left|\sum_{\Delta\le 0}\Delta\cdot f(\Delta\mid x)\cdot\frac{P(\Delta\le 0\mid x)}{P(\Delta\le 0\mid x)}\right|. \tag{32}$$

Because on the left side of this inequality

$$\frac{f(\Delta\mid x)}{P(\Delta>0\mid x)}=f(\Delta\mid\Delta>0,x), \tag{33}$$

and on the right side of the inequality

$$\frac{f(\Delta\mid x)}{P(\Delta\le 0\mid x)}=f(\Delta\mid\Delta\le 0,x), \tag{34}$$

the inequality can be rewritten as

$$\left|\sum_{\Delta>0}\Delta\cdot f(\Delta\mid\Delta>0,x)\cdot P(\Delta>0\mid x)\right|>\left|\sum_{\Delta\le 0}\Delta\cdot f(\Delta\mid\Delta\le 0,x)\cdot P(\Delta\le 0\mid x)\right|. \tag{35}$$

Note that on the left side of inequality 35

$$\sum_{\Delta>0}\Delta\cdot f(\Delta\mid\Delta>0,x)=E(\Delta\mid\Delta>0,x), \tag{36}$$

and on the right side of the inequality

$$\sum_{\Delta \le 0} \Delta \cdot f(\Delta \mid \Delta \le 0, x) = E(\Delta \mid \Delta \le 0, x). \quad (37)$$

By substitution into equation 35, this yields

$$\left|E(\Delta \mid \Delta > 0, x) \cdot P(\Delta > 0 \mid x)\right| > \left|E(\Delta \mid \Delta \le 0, x) \cdot P(\Delta \le 0 \mid x)\right|. \quad (38)$$

Subtracting the right side of inequality 38 from both sides yields

$$\underbrace{\left|E(\Delta \mid \Delta > 0, x) \cdot P(\Delta > 0 \mid x)\right|}_{\text{Part A}} - \underbrace{\left|E(\Delta \mid \Delta \le 0, x) \cdot P(\Delta \le 0 \mid x)\right|}_{\text{Part B}} > 0. \quad (39)$$

Since Part A of inequality 39 is the absolute value of a positive number (note we are conditioning on $\Delta > 0$), the absolute value function can be dropped. Similarly, since Part B is the absolute value of a non-positive number (note we are conditioning on $\Delta \le 0$), its subtraction from *A* is just the addition of the non-positive number. The absolute value operation can be dropped as well, if we add the components rather than subtract them. This yields

$$E(\Delta \mid \Delta > 0, x) \cdot P(\Delta > 0 \mid x) + E(\Delta \mid \Delta \le 0, x) \cdot P(\Delta \le 0 \mid x) > 0. \quad (40)$$

However, the left-hand side of this inequality is the expected value of Δ conditional on *x*. Therefore, explanation E_2 implies that

$$E(\Delta \mid x) > 0 \quad \forall x \in \mathbb{R}_X. \quad (41)$$

Since $\Delta = \partial g / \partial x$ and derivatives are linear operators (and assuming we can interchange the derivative and integral operations), we have

$$E(\Delta \mid x) = E\left(\frac{\partial g(x)}{\partial x}\middle| x\right) = \frac{dE(g(x) \mid x)}{dx}, \quad (42)$$

and therefore, the implication of the explanation we seek to test is the direction of the derivative of the expected value of *g*:

$$\frac{dE(g(x) \mid x)}{dx} > 0 \quad \forall x \in \mathbb{R}_X. \quad (43)$$

Unfortunately, whereas we are likely able to empirically evaluate E($Y \mid x$) in a regression analysis, we are not likely able to directly evaluate E($g \mid x$). This is okay, if we can we use E($Y \mid x$) to evaluate E($g \mid x$). When can we do this? The requirements are identified by taking the derivative of equation 16 with respect to *x*:

$$\frac{dE(Y \mid x)}{dx} = \int \frac{\partial g(x;\theta)}{\partial x} \cdot f(\theta \mid x) \cdot d\theta + \underbrace{\int g(x;\theta) \cdot \frac{\partial f(\theta \mid x)}{\partial x} \cdot d\theta}_{\text{Part A}} + \underbrace{\frac{\partial E(\mathcal{E} \mid x)}{\partial x}}_{\text{Part B}}. \quad (44)$$

If the distribution of parameter Θ is independent of X (which, in econometrics, is often considered as there is no selection on the gains [29]), then df/dx = 0 and consequently Part A of equation 44 is zero. If the error is mean independent of X, then Part B is zero (which in econometrics, is often considered as there is no selection on the outcome [29]). Under these conditions we have

$$\frac{dE(Y \mid x)}{dx} = \int \frac{\partial g(x;\theta)}{\partial x} \cdot f(\theta) \cdot d\theta \,. \tag{45}$$

But, the right-hand side of equation 45 is the E($\Delta \mid x$), which is what we seek to evaluate for our test. Consequently, our empirical claim regarding $E(e, p, \Omega, \mathbb{R}_X) = True$ for E_2 is

$$\frac{dE(Y \mid x)}{dx} \in \mathbb{D}_e, \quad \forall x \in \mathbb{R}_X \,. \tag{46}$$

Given the independence assumptions required for parts A and B to equal 0 in equation 44, we can test our proposed explanation E_2 by evaluating the derivative of a population-level regression function (the left-hand side of equation 45). If an empirically identified statistically adequate regression function can be used, an explicit functional form for g need not be specified a priori.

5. Conclusion

Knowing how to test a proposed explanation in a population requires having a definition for what is meant by explaining in a population. In this paper I gave definitions in terms of the scope of an explanation and in terms of the power of an explanation. I provided a general method for testing proposed explanations using random parameters models, and I showed when population-level regression models can be used to test proposed explanations in terms of effective power.

Although the tests were presented in terms of the minimal definitions, the tests can be extended to generalized definitions as described above. Using the random parameters method, we can define our explanations in terms of the explanation-implied region being a multiple of that for the non-implied region. For example, the proposed explanation explains if it applies to at least 90 percent of the population (rather than at least 50 percent as used in the minimal definitions).

I focused on defining and testing proposed explanations; however, in practice the requirements for such a test to provide evidence must be kept in mind. Specifically, a proposed explanation's testable empirical implications need to be specified such that alternative potential explanations for empirical

implications are accounted for or ruled out, typically by statistical or experimental control. The extent of evidence provided by the test depends on the confidence we have that alternative explanations for empirical findings are indeed ruled out: the less confident we are, the less evidence is provided by the test. This concern is addressed by calibrating our interpretation accordingly.

This paper addressed defining and testing explanations in populations. However, it should be noted that the general definition can be the basis for addressing estimation goals as well as testing goals. Using the random coefficients method the proportion of a population that conforms to the explanation's implications or the effective power can be estimated along with corresponding bootstrapped confidence intervals.

References

[1] E.T. Higgins, Promotion and prevention: Regulatory focus as a motivational principle, in: M.P. Zanna (Ed.) Adv Exp Soc Psychol, Academic Press, New York, 1998, pp. 1-46.

[2] P.J. Veazie, S. McIntosh, B. Chapman, J.G. Dolan, Regulatory focus affects physician risk tolerance, Health Psychology Research, 2 (2014) 85-88.

[3] E. Demidenko, Mixed models : theory and applications, Wiley-Interscience, Hoboken, N.J., 2004.

[4] R.B. Darlington, A.F. Hayes, Regression analysis and linear models : concepts, applications, and implementation, Guilford Press, New York, 2017.

[5] M. Bunge, Philosophy of science: From Explanation to Justification, Rev. ed., Transaction Publishers, New Brunswick, N.J., 1998.

[6] T. Sider, Writing the book of the world, Clarenson Press ; Oxford University Press, Oxford, New York, 2011.

[7] N.C.A. da Costa, S. French, Science and partial truth : a unitary approach to models and scientific reasoning, Oxford University Press, Oxford ; New York, 2003.

[8] M. Strevens, Depth : an account of scientific explanation, Harvard University Press, Cambridge, Mass., 2008.

[9] P.J. Veazie, Understanding Scientific Inquiry, Science and Philosophy, 6 (2018) 3-14.

[10] N. Bohr, On the Constitution of Atoms and Molecules, Philos Mag, 26 (1913) 857-875.

[11] N. Bohr, On the Constitution of Atoms and Molecules, Philos Mag, 26 (1913) 476-502.

[12] N. Bohr, On the Constitution of Atoms and Molecules, Philos Mag, 26 (1913) 1-25.

[13] S. DellaVigna, Psychology and Economics: Evidence from the Field, J. Econ. Lit., 47 (2009) 315-372.

[14] M. Rabin, A perspective on psychology and economics, Eur Econ Rev, 46 (2002) 657-685.

[15] E.F. Loftus, J.W. Schooler, Information-Processing Conceptualizations of Human Cognition: Past, present, and future, in: G.D. Ruben (Ed.) Informatoin and Behavior, Transaction Books, New Brunswick, NJ, 1985, pp. 225-250.

[16] P. Ylikoski, J. Kuorikoski, Dissecting explanatory power, Philos Stud, 148 (2010) 201-219.

[17] M.P. Cohen, On Three Measures of Explanatory Power with Axiomatic Representations, Brit J Philos Sci, 67 (2016) 1077-1089.

[18] J.N. Schupbach, J. Sprenger, The Logic of Explanatory Power, Philosophy of Science, 78 (2011) 105-127.

[19] J.N. Schupbach, Comparing Probabilistic Measures of Explanatory Power, Philosophy of Science, 78 (2011) 813-829.

[20] V. Crupi, K. Tentori, A Second Look at the Logic of Explanatory Power (with Two Novel Representation Theorems), Philosophy of Science, 79 (2012) 365-385.

[21] J.B. Freeman, Acceptable premises : an epistemic approach to an informal logic problem, Cambridge University Press, Cambridge, UK ; New York, 2005.

[22] E.T. Higgins, Beyond pleasure and pain, Am. Psychol., 52 (1997) 1280-1300.

[23] P. Veazie, What makes variables random : probability for the applied researcher, CRC Press, Taylor & Francis Group, Boca Raton, 2017.

[24] A. Spanos, Revisiting Haavelmo's structural econometrics: bridging the gap between theory and data, Journal of Economic Methodology, 22 (2015) 171-196.

[25] J.T. Fox, K.I. Kim, C.Y. Yang, A simple nonparametric approach to estimating the distribution of random coefficients in structural models, Journal of Econometrics, 195 (2016) 236-254.

[26] J.T. Fox, K.I. Kim, S.P. Ryan, P. Bajari, A simple estimator for the distribution of random coefficients, Quant Econ, 2 (2011) 381-418.

[27] G. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[28] P.J. Veazie, Understanding Statistical Testing, Sage Open, 5 (2015).

[29] J.J. Heckman, E. Vytlacil, Econometric Evaluation of Social Programs, Part 1: Causal models, structural models and econometric policy evaluation, in: J. Heckman, E. Leamer (Eds.) Handbook of Econometrics, Elsevier, Amsterdam, 2007, pp. 4779-4874.