UNIVERSITY OF LEICESTER

Department of Economics

A Method for Measuring Treatment Effects on the Treated without Randomization

P. A. V. B. Swamy, Federal Reserve Bond
S. G. Hall, University of Leicester
G. S. Tavlas, University of Leicester
I. Chang, The American University, Washington
H. D. Gibson, Bank of Greece
W. H. Greene, New York University
J. S. Mehta, Temple University, Philadelphia

Working Paper No. 16/02

March 2016

A Method for Measuring Treatment Effects on the Treated without Randomization*

P. A. V. B. Swamy[a], S. G. Hall[b], G. S. Tavlas[c], I. Chang[d], H. D. Gibson[e], W. H. Greene[f], and J. S. Mehta[g]

Abstract

This paper contributes to the literature on the estimation of causal effects by providing an analytical formula for individual specific treatment effects and an empirical methodology that allows us to estimate these effects. We derive the formula from a general model with minimal restrictions, unknown functional form and true unobserved variables such that it is a credible model of the underlying real world relationship. Subsequently, we manipulate the model in order to put it in an estimable form. In contrast to other empirical methodologies, which derive average treatment effects, we derive an analytical formula that provides estimates of treatment effects on each treated individual. We also provide an empirical example that illustrates our methodology.

Key words: Causality; Real-world relationship; Unique error term; Treatment effect; Non-experimental situation

JEL Classification: C13, C51

* We thank Fredj Jawadi and three referees for constructive comments.

[a] Federal Reserve Board (Retired), 6333 Brocketts Crossing, Kingstowne, VA, 22315, USA, e-mail: swamyparavastu@hotmail.com

[b] Leicester University Room Astley Clarke 116, University Road, Leicester LEI 7RH, UK and Bank of Greece, e-mail: sh222@leicester.ac.uk

[c] Member, Monetary Policy Council, Bank of Greece, 21 El. Venizelos Ave. 102 50, Athens, Greece and Visiting Professor at Leicester University, e-mail: gtavlas@bankofgreece.gr

[d] Department of Mathematics and Statistics (Retired), The American University, Washington, DC 20016, USA, e-mail: ilchang@american.edu

[e] Economic Research Department, Bank of Greece, 21 El. Venizelos Ave. 102 50, Athens, Greece, email: hgibson@bankofgreece.gr

[f] New York University, Department of Economics, 44 West Fourth Street, 7-90 New York, NY 10012 e-mail: wgreene@stern.nyu.edu

[g] Department of mathematics, Temple University, Philadelphia, PA 19122, USA, e-mail: mehta1007@comcast.net

1. Introduction

Previous studies have dealt with the issue of estimating the average treatment effect on the treated (ATET) or the treatment effects averaged over the entire population (ATE).[1] These studies have typically relied on the estimation of *average* treatment effects; random assignment to treatment aims to ensure that individuals (or units) assigned to the treatment and individuals assigned to control are identical; the average outcome among the control individuals serves as the counterfactual for the average outcome among the treated individuals. The difference between those two averages is an estimate of the central tendency of the distribution of unobservable individual-level treatment effects.[2]

Estimation of treatment effects is challenging when the treatment assignment is not completely random. In this paper, we provide a method that does not require either completely random assignment or data on pairs of individuals matched by some specific criterion -- one subjected to control and the other subjected to the treatment.[3] Our model has unique coefficients and error term and guards against incorrect functional form.[4] We provide a precise specification for the treatment effect under the condition that individuals are self-selected into treatment. In deriving this specification, we use the following definition of the treatment effect: the effect of a treatment on a treated individual minus the outcome that would have been observed

[1] For textbook expositions, see Greene (2012, pp. 888-896) and Wooldridge (2013, pp. 438-43).
[2] See Holland (1986).
[3] For a short discussion on these issues, see Greene (2012, pp. 893-895).
[4] As we explain below, the coefficients and error term of a linear-in-variables and nonlinear-in-coefficients model are unique in the sense that they are invariant under the addition and subtraction of the coefficient of an omitted regressor times any included regressor on its right-hand side. Swamy, Mehta, Tavlas and Hall (2014, 2015) showed that models with nonunique coefficients and error terms are misspecified.

had the same individual not been treated (the counterfactual).[5] Thus, in contrast to previous studies, which deal with average treatment effects, our definition is *individual* specific. There are practical difficulties in empirically implementing our definition. In what follows we describe these difficulties and provide solutions.

An intuitive explanation of the contribution of this paper is as follows. In a randomized trial it is relatively easy to calculate the effect of a given treatment. This can be done simply by estimating a standard model with a dummy variable for the treatment: since we know that the treatment is random, it can be treated as exogenous. However, in a real-world situation without randomization it is extremely unlikely that the treatment can be assumed to be exogenous. Consider the case of a new cancer treatment. Clearly, the treatment would only be given to patients who are severely ill and likely to die. The treatment is not random and, therefore, simply adding a dummy for treated individuals is likely to be highly misleading and may even lead us to conclude that the treatment causes the patients to die from the illness. The empirical literature has attempted to deal with this problem by using instrumental variables (in a variety of ways). However, the difficulties of weak or irrelevant instruments are well-known.[6] This paper offers a new approach to this problem, based on coefficients that vary. Our approach avoids both the misspecification caused by incorrect functional forms, and provides coefficients that absorb omitted regressors, measurement errors and endogeneity. These varying coefficients may then be decomposed to obtain an estimate of the true underlying treatment effect.

The remainder of the paper consists of four sections. Section 2 consists of several parts. It begins by reviewing the concepts needed to define the causal effect of

[5] This is what Greene (2012, pp. 888-889) calls "the treatment effect in a pure sense."
[6] See, for example, Swamy, Tavlas and Hall (2015).

a treatment on a treated individual. The section then develops two models that contain what we characterize as "unique coefficients and error terms" -- one model for the causal effects attributable to the treatment and the other model for the unknown values of what "response" the individuals who participated in a treatment would have had they not been treated. In this connection, we provide both a formal derivation and an intuitive account of our theoretical derivation. Finally, the section discusses the issue of identification, presents a possible method of estimation of these models, and derives the predictions of the treatment effects. Section 3 provides an empirical example to illustrate our method. Section 4 concludes.

2. Modeling the Effect of a Treatment on the Treated in Non-experimental Situations

2.1 Preparations

2.1.1 Notation: Let i index treated individuals and let i' ($\neq i$) index untreated individuals.[7] The number of treated individuals is denoted by n_1 and that of untreated individuals is denoted by n_2. Let $n_1 + n_2 = n$, the size of a sample of both treated and untreated individuals. Both n_1 and n_2 are known. We assume that the individual response to treatment is heterogeneous. The dummy variable C is defined to take the value 0 for untreated individual i' and to take the value 1 for treated individual i. For untreated individual i', $(y^*_{i'} \mid C_{i'} = 0) = y^*_{0i'}$ is the unobserved true value of the observed outcome ($y_{0i'}$) of no treatment; $y^*_{0i'}$ plus measurement error ($u^*_{0i'}$) is the observed value, $y_{0i'}$. For treated individual i, $(y^*_i \mid C_i = 1) = y^*_{1i}$, $y_{1i} = y^*_{1i} +$

[7] Most empirical studies use a treatment dummy to derive the impact of the treatment; the dummy variable takes the value 1 for the treated individuals and 0 for the untreated individuals.

u_{1i}^{*} where y_{1i}^{*} is the (unobserved) true value of the observed outcome (y_{1i}) of a treatment and u_{1i}^{*} is measurement error.

2.1.2 Potential outcome notation: Pratt and Schlaifer (1988, pp. 28 and 35) used Neyman's potential-outcome notation to state causal laws.[8] Potential outcomes can be recognized through the subscripts that are attached to counterfactual events (see Pearl 2010, p. 3). Symbolically, potential outcomes are denoted by Y_{xi}, which shows the value that outcome *Y* would take for individual *i* had the treatment *X* been at level *x*.

2.1.3 Counterfactuals: The symbol $y_{1i'}^{*}$ denotes a value of what the outcome would have been had individual i' been treated. The symbol y_{0i}^{*} denotes a value of what the outcome would have been had individual *i* not been treated. The variables $y_{1i'}^{*}$ and y_{0i}^{*} are the unobserved counterfactuals implicit in the true values $y_{0i'}^{*}$ and y_{1i}^{*}, respectively. Both the values of $y_{0i'}$ (the effects of no treatment on the untreated individuals) and y_{1i} (the effect of treatment on the treated individual) are observed but they both cannot be observed for the same individual since y_{1i} refers to a treated individual and $y_{0i'}$ refers to an untreated individual.

2.1.4 Treatment effects in a pure sense: $y_{1i'}^{*} - y_{0i'}^{*}$ and $y_{1i}^{*} - y_{0i}^{*}$.

In the treatment effect on the untreated, defined by $y_{1i'}^{*} - y_{0i'}^{*}$, $y_{0i'}^{*}$ is the unobserved true value; it differs from the observed value $y_{0i'}$ by a measurement error, and the counterfactual $y_{1i'}^{*}$ has no observations for all untreated individuals $i' = 1, \ldots, n_2$. In

[8] In doing so, Pratt and Schlaifer (1988) followed Rubin (1974, 1978).

the treatment effect on the treated, defined by $y_{1i}^{*} - y_{0i}^{*}$, y_{1i}^{*} is the unobserved true value, it differs from the observed value y_{1i} by a measurement error, and the counterfactual y_{0i}^{*} has no observations for all treated individuals $i = 1, \ldots, n_1$.

2.1.5 The purpose of the paper: In Section 2.2 below, we derive the models of y_{1i}^{*} and y_{0i}^{*} that give the predictions of their dependent variables, respectively. Following Greene (2012, p. 888), we believe that an accurate estimate of the treatment effect $y_{1i}^{*} - y_{0i}^{*}$ on the treated is more useful than an accurate estimate of the treatment effect $y_{1i'}^{*} - y_{0i'}^{*}$ on the untreated.[9] That is, it is more natural to ask, what is the treatment effect on a treated individual, rather than ask, what would have been the treatment effect on an untreated individual?[10] In the following subsections, we derive an analytical formula for $y_{1i}^{*} - y_{0i}^{*}$.

2.1.6 What is causality? Previous researchers have set-forth various definitions of causality. In this section, we show how our specification of a treatment effect relates to several of those definitions. Our aim here is not to provide a comprehensive discussion of the causality literature.

- First, Basmann (1988, p. 99) revealed that common to all of the generally accepted meanings of "causality" is the notion that causality is a property of the real world and is not an algebraic property of the mathematical representations of parts of the real world. An insight that follows from this notion is that real-world relationships do not contain specification errors. This insight suggests that statistical causation requires the need to

[9] Greene (2012, p. 888) pointed out that "The natural, ultimate objective of an analysis of a 'treatment' or intervention would be the effect of treatment on the treated."
[10] Greene (2012, p. 894) pointed out that the desired quantity is not necessarily the ATE, but ATET.

derive estimates within an environment free of specification errors. With regard to the definition of treatment effects, the notion requires that, to measure causal effects, we should take the difference between the real-world relations for the outcome of a treatment and the potential outcome of no treatment on the same individual. In what follows, we empirically implement this definition.

- Second, to show statistical causation, Skyrms (1988) proved that positive statistical relevance needs to continue to hold when all relevant pre-existing conditions are controlled-for.[11] Intuitively, the relevant pre-existing conditions can be thought of as all the factors that might affect a relationship but which cannot be captured (for example, omitted variables). For example, the typical empirical counterpart to household consumption function is derived from a utility function. We don't know how to measure the utility function, but it governs the actual structure of the consumption function. We control for such pre-existing conditions.

To be specific, we follow generally accepted meanings of causality. We follow Basmann's clarification that causal relations should be free of specification errors and Skyrms' explanation of the definition of statistical causation which stresses the need to control for pre-existing conditions. To account for Skyrms' (1988, p. 59)

[11] Skyrms distinguished among different types of causation such as deterministic, probabilistic, and statistical. He argued that the answers to questions of probabilistic causation given by different statisticians depended on their conceptions of probability. Three major concepts of probability are: rational degree of belief, limiting relative frequency, and propensity or chance. Skyrms (1988, p. 59) recognized that not all would agree with the subjectivistic gloss he put on the causal approaches of Reichenbach, Granger, Suppes, Salman, Cartwright and others. As Skyrms pointed out, "statistical causation is positive statistical relevance which does not disappear when we control for all relevant pre-existing conditions." We consider this definition of statistical causation here. Skyrms further clarified that "Within the Bayesian framework … 'controlling for all relevant pre-existing conditions' comes to much the same as identifying the appropriate partition … which together with the presence or absence of the putative cause (or value of the causal variable) determines the chance of the effect."

and Basmann's (1988, p. 99) insights on these issues, we derive real-world relations -- that is, relations free of all specification errors -- under the insight that causality is a property of the real world. Skyrms' insight leads to the conclusion that all irrelevant variables need to be eliminated from a relation. To find such a relation, we start with a general nonlinear mathematical model with unknown functional form, in which the dependent variable satisfies the normalization rule (that the coefficient on the dependent variable equals unity), and the arguments of the mathematical function include all the determinants of the dependent variable and all the relevant pre-existing conditions. We express this model as linear in variables and nonlinear in coefficients. These coefficients are the partial derivatives of the function with respect to its arguments. It can be verified that this linear-in-variables and nonlinear-in-coefficients model has the correct *functional* form. These partial derivatives keep the values of all *relevant pre-existing conditions constant*. Specifically, we: (i) follow Basmann's (1988) notion of causality because it is not restrictive (it necessitates the absence of specification errors); (ii) follow Skyrms' (1988) elucidation of statistical causation; (iii) work with the partial derivatives of some deterministic real-world (i.e., misspecification-free) relationships to control for all relevant pre-existing conditions and use the frequentist probability to measure causal effects; and (iv) work with the misspecification-free models of y_{1i}^{*} and y_{0i}^{*} to evaluate $y_{1i}^{*} - y_{0i}^{*}$.[12]

Finally, in articulating a definition of causality, we also take account of the insights provided by Zellner (1979) and Pratt and Schlaifer (1988). Zellner adopted Feigl's definition according to which causality is 'predictability according to a law or set of laws.' Pratt and Schlaifer defined a law with factors and concomitants and

[12] The list of these misspecifications is given in Section 2.2.6 below.

provided the conditions under which the laws can be observed in data.[13] In what follows, we develop both a set of laws and the necessary additional variables -- which we call coefficient drivers -- needed to empirically implement the laws.[14]

2.2 The Correctly Specified (or Misspecification-free) Models of y_{1i}^*, y_{1i}, and y_{0i}^*

2.2.1 Mathematical functions: To generate the predictions on y_{1i}^*, y_{1i}, and y_{0i}^*, we begin with their real-world relationships expressed in terms of the following *mathematical* equations.

$$y_{c\eta}^* = f_{c\eta}(x_{c\eta 1}^*, ..., x_{c\eta, L_{c\eta}}^*) \quad (1)$$

where $c \in (0,1)$, $\eta \in (i, i')$. Since equation (1) is a mathematical equation, it does not contain an error term.

Henceforth, $f_{c\eta}(x_{c\eta 1}^*, ..., x_{c\eta, L_{c\eta}}^*)$ will be written more compactly as $f_{c\eta}(.)$. The precise functional form of this function is unknown; $x_{c\eta 1}^*, ..., x_{c\eta, L_{c\eta}}^*$ are the arguments of $f_{c\eta}(.)$. These arguments are of three types: (i) observed and (ii) unobserved

[13] In his causal analyses, Pearl (2000) used the Bayesian interpretation of probability in terms of degrees of belief about events, recursive models, and in many cases finitely additive probability functions. Pearl's (2000, p. 176) Bayesian view of causality is that "[i]f something is real, then it cannot be causal because causality is a mental construct that is not well defined." This view is not consistent with Basmann's (1988) view, which is also the view that we adopt in this paper.

[14] **The principle of causal invariance** (Basmann 1988, p. 73): Causal relations and orderings are unique in the real world and they remain invariant under mere changes in the language we use to describe them. Examples of models that do not satisfy this principle are those that are built using stationarity producing transformations of observable variables (see Basmann 1988, p. 98). A related principle is that causes must precede their effects in time. Pratt and Schlaifer (1988, pp. 24-25) pointed out an interesting exception to this principle which is: "Whether or not a cause must precede its effect, engineers who design machines that really work in the real world will continue to base their designs on a law which asserts that acceleration at time *t* is proportional to force at that same time *t*." The reason why we consider real-world (misspecification-free) relationships is that they satisfy the principle of causal invariance. They do not disappear when we control for all relevant pre-existing conditions (Skyrms 1988, p. 59). We build misspecification-free models with these properties. If we do not estimate $y_{1i}^* - y_{0i}^*$ from the misspecification-free relations of y_{1i}^* and y_{0i}^*, then according to Basmann (1988), our estimate of the treatment effect $y_{1i}^* - y_{0i}^*$ will not be an estimate of the causal effect of a treatment on the treated *i*th individual.

determinants of $y^*_{c\eta}$ and (iii) all relevant pre-existing conditions; the number $L_{c\eta}$ of all these arguments is an unknown integer dependent on *c* and η, since the number of the arguments of types (ii) and (iii) is unknown. Why include type (iii) of arguments in $f_{c\eta}(.)$? The answer is provided by Skyrms' (1988, p. 59) definition that mathematical causation is positive mathematical relevance which does not disappear when we control for all relevant pre-existing conditions. To control for these conditions, we first include them directly into $f_{c\eta}(.)$ as its arguments and take the partial derivatives of $f_{c\eta}(.)$ with respect to its type (i) and type (ii) arguments that keep the values of these conditions constant. In this way, we control for all relevant pre-existing conditions.

Next, we use these partial derivatives as the coefficients of equation (2) below. There are no relevant arguments excluded from $f_{c\eta}(.)$. Therefore, there is no need to introduce an error term into $f_{c\eta}(.)$ to represent nonexistent omitted variables. Alternatively stated, all the variables constituting the econometrician's error term are treated as the arguments of $f_{c\eta}(.)$. This is done to avoid all incorrect functional forms of $f_{c\eta}(.)$. The symbols $\beta_1, \beta_2, ..., \beta_K$ may be used to denote the constant features of $f_{c\eta}(.)$. We do not treat any features of $f_{c\eta}(.)$ as constant parameters because, as Goldberger (1987) pointed out in the context of the Rotterdam school demand models, the treatment of any particular features of $f_{c\eta}(.)$ as constants may be questioned.

2.2.2 Minimally restricted relations: The only restriction that we have imposed on equation (1) is the normalization rule that the coefficient of $y^*_{c\eta}$ is equal to unity.

2.2.3 Available data for estimation of (1): We assume that $L_{1i} > K + 1 < L_{0i'}$, $K + 1 < n_1$ and n_2. Data on $x^*_{c\eta 1}$, …, $x^*_{c\eta K}$ are available. These data may contain measurement errors, i.e., $x_{c\eta 1} = x^*_{c\eta 1} + v^*_{c\eta 1}$, …, $x_{c\eta K} = x^*_{c\eta K} + v^*_{c\eta K}$, where the variables without an asterisk are observable, the variables with an asterisk are true and unobservable, and the v^*'s are measurement errors.[15] We call $x_{c\eta 1}$, …, $x_{c\eta K}$ "the included arguments of $f_{c\eta}(.)$".[16] Also available are data on $y_{0i'}$ for n_2 untreated individuals and on y_{1i} for n_1 treated individuals. For treated individuals with c = 1, y_{1i} is observed with measurement error and a non-constant proxy, denoted by $x^*_{1i,K+1}$, for the treatment variable is used as an additional included argument of $f_{1i}(.)$.[17] Let $x_{1i,K+1} = x^*_{1i,K+1} + v^*_{1i,K+1}$ where the variable without an asterisk is observable, the variable with an asterisk is true and unobservable, and the $v^*_{1i,K+1}$'s are measurement errors. No data on $x^*_{c\eta,K+2}$, …, $x^*_{c\eta,L_{c\eta}}$ are available and hence they can only be treated as omitted arguments.[18, 19]

[15] We do not treat measurement errors as random variables until we make some stochastic assumptions about them.

[16] The reason for assigning this label to them is that they are included as regressors in our regressions below.

[17] Data on $x^*_{1i,K+1}$ are not available in some experiments like medical experiments. In these cases, what all we know is whether an individual is treated or not (see Greene 2012, pp. 893-894). In these cases it is possible to obtain analytical expressions but not numerical measures for the treatment effects.

[18] The reason why we attach this label to them is that they are actually omitted from our regressions below.

[19] **Unobserved treatment variable**: In the absence of data on $x_{1i,K+1}$, the coefficient of a dummy variable is used to measure treatment effects, as in the Heckman and Schmierer's (HS) (2010) model. Greene (2012, pp. 251-254, 893) elaborated on this practice by commenting that though a treatment can be represented by a dummy variable, measurement of its effect cannot be done with multiple linear regression.

2.2.4 Correctly specified models for y_{1i} and the counterfactual y_{0i}^* for the same individual *i*:

Without misspecifying its functional form, (1) can be expressed as

$$y_{c\eta}^* = \alpha_{c\eta 0}^* + \sum_{j=1}^{K} x_{c\eta j}^* \alpha_{c\eta j}^* + x_{c\eta,K+1}^* \alpha_{c\eta,K+1}^* + \sum_{g=K+2}^{L_{c\eta}} x_{c\eta g}^* \alpha_{c\eta g}^* \quad (2)$$

where for $\ell = 1, \ldots, L_{c\eta}$, the coefficient of $x_{c\eta\ell}^*$ is equal to $\partial y_{c\eta}^* / \partial x_{c\eta\ell}^*$ unless $x_{c\eta\ell}^*$ is discrete, in which case this partial derivative is approximated by $\Delta y_{c\eta}^* / \Delta x_{c\eta\ell}^*$ with the right sign where $\Delta y_{c\eta}^*$ and $\Delta x_{c\eta\ell}^*$ are small differences in the values of $y_{c\eta}^*$ and $x_{c\eta\ell}^*$, respectively, and the intercept $\alpha_{c\eta 0}^*$ is equal to $y_{c\eta}^* - \sum_{j=1}^{L_{c\eta}} x_{c\eta j}^* \alpha_{c\eta j}^*$. This $\alpha_{c\eta 0}^*$ is the error of approximation that results from approximating $f_{c\eta}(.)$ by $\sum_{j=1}^{L_{c\eta}} x_{c\eta j}^* \alpha_{c\eta j}^*$. Equation (2) is obtained from $y_{c\eta}^* = f_{c\eta}(x_{c\eta 1}^*, ..., x_{c\eta,L_{c\eta}}^*) - \sum_{j=1}^{L_{c\eta}} x_{c\eta j}^* \alpha_{c\eta j}^* + \sum_{j=1}^{L_{c\eta}} x_{c\eta j}^* \alpha_{c\eta j}^*$ where $\alpha_{c\eta 0}^* = f_{c\eta}(x_{c\eta 1}^*, ..., x_{c\eta,L_{c\eta}}^*) - \sum_{j=1}^{L_{c\eta}} x_{c\eta j}^* \alpha_{c\eta j}^*$. From this it follows that equation (2) without $\alpha_{c\eta 0}^*$ will have the correct functional form when $\alpha_{c\eta 0}^* = 0$. For our further analysis of (1), it is convenient to express it in the form of model (2) that is linear in variables but nonlinear in coefficients. We avoid the use of any incorrect functional form of (1) by defining the coefficients of (2) as the partial derivatives of $f_{c\eta}(.)$ with respect to its arguments. The coefficient $\alpha_{c\eta,K+1}^*$ is zero for untreated individuals. It follows that the problem of estimating $f_{c\eta}(.)$ with unknown functional form is solved by changing it to that of estimating certain partial derivatives of $f_{c\eta}(.)$.

Equation (2) is linear if its coefficients are constant and nonlinear otherwise. How do we ensure that $\alpha^*_{c\eta,K+1}$ is the causal effect of $x^*_{c\eta,K+1}$ on $y^*_{c\eta}$ holding the values of all arguments of $f_{c\eta}(.)$ other than $x^*_{c\eta,K+1}$ constant? This constancy condition is true because $\alpha^*_{c\eta,K+1}$ is the partial derivative of $y^*_{c\eta}$ with respect to $x^*_{c\eta,K+1}$. In the definition of this partial derivative, not only the values of all determinants of $y^*_{c\eta}$ other than $x^*_{c\eta,K+1}$ but also the values of all relevant pre-existing conditions are held constant. This is a standard way to eliminate the false relationship between $y^*_{c\eta}$ and any of its determinants (see Skyrms 1988, p.59). For example, suppose that the relation of $y^*_{c\eta}$ to $x^*_{c\eta 1}$ is false. Then the partial derivative of $y^*_{c\eta}$ with respect to $x^*_{c\eta 1}$ is zero because the values of all relevant pre-existing conditions are held constant. Also, we do not impose on (1) any restriction that makes it lose the causal invariance property of real-world relations described in Section 2. These precautions are taken to ensure that the partial derivatives used as the coefficients of (2) are the truths, meaning the properties of the real-world relationship in (1).

Treated individual *i*: We now apply the specification in (2) to the particular group of treated individuals. From (2) it follows that

$$y^*_{1i} = \alpha^*_{1i0} + \sum_{j=1}^{K} x^*_{1ij}\alpha^*_{1ij} + x^*_{1i,K+1}\alpha^*_{1i,K+1} + \sum_{g=K+2}^{L_{1i}} x^*_{1ig}\alpha^*_{1ig} \qquad (3)$$

where $y_{1i} = y^*_{1i} + u^*_{1i}$, y_{1i} is observed, y^*_{1i} is the unobserved true value, u^*_{1i} is measurement error, the treatment regressor $x^*_{1i,K+1}$ is added to the list $x^*_{1i1},...,x^*_{1i,K}$ but not to the list $x^*_{1i,K+2},...,x^*_{1i,L_{1i}}$, since equation (3) is for a treated individual. The coefficients of (3) are the truths about the real-world relationship in (1). Note that the

set of variables denoted by $\sum_{g=k+2}^{L_{1i}} x_{1ig}^{*} a_{1ig}^{*}$ are unobserved and need to be eliminated. To eliminate those unobserved variables, we regress each of these variables on all observed variables as follows.

$$x_{1ig}^{*} = \lambda_{1ig0}^{*} + \sum_{j=1}^{K+1} x_{1ij}^{*} \lambda_{1igj}^{*} \text{ (g = } K\text{+2, …, } L_{1i}) \tag{4}$$

where $\lambda_{1igj}^{*} = \partial x_{1ig}^{*} / \partial x_{1ij}^{*}$ if x_{1ij}^{*} is continuous and $= \Delta x_{1ig}^{*} / \Delta x_{1ij}^{*}$ with the right sign otherwise and $\lambda_{1ig0}^{*} = x_{1ig}^{*} - \sum_{j=1}^{K+1} x_{1ij}^{*} \lambda_{1igj}^{*}$. This definition makes equation (4) exact.

Model of y_{1i}^{*} with unique coefficients and error term: Substituting the right-hand side of equation (4) for x_{1ig}^{*} in (3) gives

$$y_{1i}^{*} = \alpha_{1i0}^{*} + \sum_{g=K+2}^{L_{1i}} \lambda_{1ig0}^{*} \alpha_{1ig}^{*} + \sum_{j=1}^{K+1} x_{1ij}^{*} (\alpha_{1ij}^{*} + \sum_{g=K+2}^{L_{1i'}} \lambda_{1igj}^{*} \alpha_{1ig}^{*}) \tag{5}$$

where $\sum_{g=K+2}^{L_{1i}} \lambda_{1ig0}^{*} \alpha_{1ig}^{*}$ and $(\alpha_{1ij}^{*} + \sum_{g=K+2}^{L_{1i}} \lambda_{1igj}^{*} \alpha_{1ig}^{*})$ are the unique error term and coefficients, respectively (see Swamy et al. 2014, p. 199). The formula $\sum_{g=K+2}^{L_{1i}} \lambda_{1igj}^{*} \alpha_{1ig}^{*}$ measures omitted-regressors bias of the coefficient of x_{1ij}^{*}. For $j = 1,\ldots, K + 1$, the α_{1ij}^{*} 's are the partial derivatives of (1) with c = 1 and $\eta = i$.

Equations (1) for $c = 1$ and $\eta = i$ and (5) are the two forms of the same real-world relation in (1).

Recall, the variables y_{1i}^{*} and x_{1ij}^{*}, $j = 1, \ldots, K + 1$, in equation (5) are the true values and are not the observed values. To express (5) in terms of the observed values, we insert measurement errors at the appropriate places in (5). Doing so gives

$$y_{1i} = \gamma_{1i0} + \sum_{j=1}^{K+1} x_{1ij}\gamma_{1ij} \tag{6}$$

where

$$\gamma_{1i0} = \alpha^*_{1i0} + \sum_{g=K+2}^{L_{1i}} \lambda^*_{1ig0}\alpha^*_{1ig} + u^*_{1i} - \sum_{x\in S_2} v^*_{1ij}(\alpha^*_{1ij} + \sum_{g=K+2}^{L_{1i}} \lambda^*_{1igj}\alpha^*_{1ig}) \tag{7}$$

$$\gamma_{1ij} = (1-\frac{v^*_{1ij}}{x_{1ij}})(\alpha^*_{1ij} + \sum_{g=K+2}^{L_{1i}} \lambda^*_{1igj}\alpha^*_{1ig}) \text{ if } x \in S_1 \tag{8}$$

$$= (\alpha^*_{1ij} + \sum_{g=K+2}^{L_{1i}} \lambda^*_{1igj}\alpha^*_{1ig}) \text{ if } x \in S_2, \tag{9}$$

S_1 is the set of all continuous regressors of equation (6) and S_2 is the set of all regressors of (6) that take the value zero with positive probability. In equations (7) and (8), $-\sum_{x\in S_2} v^*_{1ij}(\alpha^*_{1ij} + \sum_{g=K+2}^{L_{1i}} \lambda^*_{1igj}\alpha^*_{1ig})$ and $(-\frac{v^*_{1ij}}{x_{1ij}})(\alpha^*_{1ij} + \sum_{g=K+2}^{L_{1i}} \lambda^*_{1igj}\alpha^*_{1ig})$ are the measurement-error biases of γ_{1ij} if $x \in S_2$ and $x \in S_1$, respectively.[20] These measurement-error biases are not unique.

What would have been the outcome, denoted by y^*_{0i}, had the *i*th individual not been treated? We determine this outcome by setting the treatment $x^*_{1i,K+1}$ equal to zero in (3). Doing so gives

$$y^*_{0i} = y_{1i} - x^*_{1i,K+1}\alpha^*_{1i,K+1} - u^*_{1i}$$

$$= \alpha^*_{1i0} + \sum_{j=1}^{K} x^*_{1ij}\alpha^*_{1ij} + \sum_{g=K+2}^{L_{1i}} x^*_{1ig}\alpha^*_{1ig} \tag{10}$$

[20] One result that can be derived from (6)-(9) is the following: Consider two competing models of the same dependent variable with unique coefficients and error terms. Let each of these models be written in the form of (6) and let some continuous regressors be common to these two models. Of the pair of coefficients on a common regressor in the two models, the one with smaller magnitudes of omitted-regressor and measurement-error biases will be closer to the common true partial derivative component of the pair. This correct conclusion could not be drawn from the J test of two separate families of hypotheses on a misspecified model (see Greene 2012, p. 136).

where $y^*_{0i} = y_{1i} - x^*_{1i,K+1}\alpha^*_{1i,K+1} - u^*_{1i} = y_{0i} - u^*_{0i}$, y^*_{0i} is the unobserved true value, u^*_{0i} is measurement error, and $\alpha^*_{1i0} = y^*_{0i} - \sum_{\ell=1}^{L_{0i}} x^*_{1i\ell}\alpha^*_{1i\ell} + x^*_{1i,K+1}\alpha^*_{1i,K+1}$.

The treatment causal effect (TCE) on the *i*th treated individual: $y^*_{1i} - y^*_{0i} = (y_{1i} - u^*_{1i}) - (y_{0i} - u^*_{0i}) = \{(\alpha^*_{1i0} + \sum_{j=1}^{K} x^*_{1ij}\alpha^*_{1ij} + x^*_{1i,K+1}\alpha^*_{1i,K+1} + \sum_{g=K+2}^{L_{1i}} x^*_{1ig}\alpha^*_{1ig} + u^*_{1i}) - u^*_{1i}\} - \{(\alpha^*_{1i0} + \sum_{j=1}^{K} x^*_{1ij}\alpha^*_{1ij} + \sum_{g=K+2}^{L_{1i}} x^*_{1ig}\alpha^*_{1ig} + u^*_{0i}) - u^*_{0i}\} = x^*_{1i,K+1}\alpha^*_{1i,K+1}$ (11)

Thus, to derive the TCE, equations (3) and (10) enable us to derive the TCE in equation (11). However, equation (11) is an analytical equation. To *estimate* the TCE in equation (11), we need to complement equation (6) with additional equations.

2.2.5 In what sense are the coefficients and error term of (5) unique? The arguments, $x^*_{1i1},...,x^*_{1i,K+1}$, included in both (3) and (5) are called "the included regressors" and the arguments $x^*_{1i,K+2},...,x^*_{1i,L_{1i}}$ included in (3) but not in (5) are called "omitted regressors." These regressors are not unique.[21] The coefficients and error term of (5) have the correct functional forms and as a result, are unique in the sense that they are invariant under the addition and subtraction of the coefficient of any omitted regressor times any included regressor on the right-hand side of equation (3).[22] It can be shown that the error term of (5) is the unique function $\sum_{g=K+2}^{L_{1i}} \lambda^*_{1ig0}\alpha^*_{1ig}$ (with the correct functional form) of the 'sufficient sets' (λ^*_{1ig0}, $g = K + 2, \ldots, L_{1i}$) of omitted regressors, a concept due to Pratt and Schlaifer (1988, p. 34). The uniqueness

[21] A proof of this statement follows from Pratt and Schlaifer's (1988, p. 34) statement that "… some econometricians require that … [the included regressors] be independent of 'the' excluded variables themselves. We shall show … that this condition is meaningless unless the definite article is deleted and can then be satisfied only for certain 'sufficient sets' of excluded variables …"

[22] A proof of this statement is given in Swamy et al. (2014, pp.217-219).

of its coefficients and error term means that (5) possesses the causal invariance property.

2.2.6 What specification errors is the TCE free from? (i) We have ensured that the unknown functional forms of (3) and (4) did not become the source of specification errors. (ii) By ensuring that the coefficients and error term of (5) are unique, the specification errors resulting from non-unique coefficients and error terms are not allowed to occur. (iii) Pratt and Schlaifer (1988, p. 34) pointed out that the requirement that the included regressors be independent of the excluded regressors themselves is "meaningless". The specification error introduced by making this meaningless assumption is avoided by taking a unique function of certain 'sufficient sets' of omitted regressors as the error term of (5). (iv) The specification error of ignoring measurement errors when they are present is avoided by placing them at the *appropriate* places in (5) to obtain equation (6). The TCE in (11) is derived from equations (3) and (10) which are free of specification-errors (i)-(iv). It should be noted that when we state that (3), (6), (10) and (11) are free of specification errors, we mean that they are free of specification-errors (i)-(iv). Using (1)-(5) we have derived a real-world relationship in (6) that is free of specification-errors (i)-(iv). Thus, our approach affirms that any relationship suffering from anyone of these specification errors is definitely not a real-world relationship.

2.2.7 Specification errors and omitted-regressor biases: It would be useful to refer to a highly-influential paper by Yatchew and Griliches (YG) (1984). Those authors considered a simple binary choice model with any two regressors (X_1 *and* X_2) with nonunique coefficients and error term and omitted from it one (X_2) of its two regressors. YG showed that even if the omitted regressor is uncorrelated with the

included regressor, the coefficient on the included regressor will be inconsistent. In addition, they showed that if the disturbances in the underlying regression are heteroscedastic, then the maximum likelihood estimators that assume homoscedasticity are inconsistent and the covariance matrix is inappropriate. What is important here is that not only the omission of a regressor from the YG model, but also the omitted regressors implicit in the YG's mean-zero error term, introduce omitted variables' biases. Furthermore, the YG results are subject to the four specification errors discussed in the previous section. As noted, our approach, using equations (1)-(6), avoids these specification errors.

2.2.8 The available data are not adequate to estimate TCE: There are practical difficulties in estimating the TCE, $x^{*}_{1i,K+1}\alpha^{*}_{1i,K+1}$, because the partial derivative $\alpha^{*}_{1i,K+1}$ in (6) is corrupted by omitted-regressor and measurement-error biases. These omitted-regressor biases arise as a direct consequence of using the equations in (4) to remove x^{*}_{1ig}, $g = K + 2, \ldots, L_{1i}$, from (3) and measurement-error biases arise as a direct consequence of measurement errors in $x_{c\eta 1} = x^{*}_{c\eta 1} + \nu^{*}_{c\eta 1}, \ldots, x_{c\eta,K+1} = x^{*}_{c\eta,K+1} + \nu^{*}_{c\eta,K+1}$. Unless these biases are eliminated from $\gamma_{1i,K+1}$ we cannot obtain consistent estimate of $\alpha^{*}_{1i,K+1}$. We will show below what additional data are needed for this removal.

2.3 Variable Coefficient Regression

The model for Y_{xi} is the same as model (6) for y_{1i}. Rewrite this model as

$$Y_{xi} = y_{1i} = \boldsymbol{x}'_{1i}\,\boldsymbol{\gamma}_{1i}\,(i = 1, \ldots, n_1) \qquad (12)$$

where $\boldsymbol{x}_{1i} = (1, x_{1i1}, ..., x_{1i,K+1})'$ is the $(K+2)\times 1$ vector of regressors, $\boldsymbol{\gamma}_{1i} = (\gamma_{1i0}, \gamma_{1i1}, ..., \gamma_{1i.K+1})'$ is the $(K+2)\times 1$ vector of coefficients. We characterize this model as "the correctly specified model" Because the model is derived from the real-world

relationship in (1) without making any specification error. Similarly, equation (10) is the correctly specified model of the counterfactual (y^*_{0i}). In equation (11), it is shown that TCE = y^*_{1i} - y^*_{0i} = $x^*_{1i,K+1}\alpha^*_{1i,K+1}$. To estimate this TCE, we use (8) or (9) which shows that $\alpha^*_{1i,K+1}$ = $\gamma_{1i.K+1}$ - its omitted-regressor and measurement-error biases. We first estimate $\gamma_{1i.K+1}$ and decompose it into an estimate of $\alpha^*_{1i,K+1}$ and an estimate of $\gamma_{1i.K+1}$'s omitted-regressor and measurement-error biases.

2.3.1 Parameterization of the variable coefficient regression: Equation (12) is estimated subject to the restrictions equations (7)-(9) imposed on its coefficients. To make such estimation feasible, we assume that for j = 0, 1, …, K + 1,

$$\gamma_{1ij} = \pi_{1j0} + \sum_{h=1}^{p} z_{1ih}\pi_{1jh} + \varepsilon_{1ij} \tag{13}$$

where the z_{1ih}'s are observable and are called "the coefficient drivers", some of which may be common to different coefficients of (12), and the π's are unknown fixed parameters.[23, 24] The error term ε_{1ij} is treated as a random variable.[25]

A key justification for equation (13) is that it facilitates separate estimation of each component of the coefficients of (12), as will be obvious from equation (20) below. For each j, given the proportion of measurement error $\frac{v^*_{1ij}}{x_{1ij}}$ in (8), the

[23] We call the coefficients of (12) "the random coefficients" but not "random parameters." The reason is that there are only coefficients and no parameters in (12). We call the coefficients of (13) "the fixed parameters" to distinguish them from those of the fixed-coefficient versions of (12). We do not use the word "random parameters," since it creates confusion between (12) and its fixed coefficient versions.

[24] The definition of coefficient drivers differs from the definition of instrumental variables. The latter variables do not explain variations in the coefficients of (12) as do coefficient drivers. The coefficient on any regressor in (12) is partly dependent on the coefficient drivers in (13). In instrumental variable estimation, first the instrumental variables are used to transform both the dependent variable and the explanatory variables and then the transformed variables are used to estimate the coefficients on the regressors.

[25] The functional form of (13) is different from that of (8) or of (7) and (9). We will correct this mistake in equation (21) below.

coefficient drivers in (13) should split into 2 sets so that one set explains most of variation in the bias-free component (α^*_{1ij}) and the other set explains most of variation in omitted-regressor bias component ($\sum_{g=K+2}^{L_{1i}} \lambda^*_{1igj}\alpha^*_{1ig}$) in (8). These sets may or may not be non-overlapping. We will make use of these conditions in equation (20) below.

Note that if (1) is nonlinear, then the α^*_{1ij}'s are functionally dependent on the x^*_{1ij}'s and x^*_{1ig}'s including $x^*_{1i,K+1}$. This introduces correlations between $\boldsymbol{x}_{1\boldsymbol{i}}$ and $\boldsymbol{\gamma}_{1i}$. In the presence of these correlations we proceed as follows:

Admissibility condition: The vector $\boldsymbol{Z}_{1\boldsymbol{i}} = (Z_{1i0}, Z_{1i1}, ..., Z_{1ip})'$ in equation (13) is an admissible vector of coefficient drivers if, given $\boldsymbol{Z}_{1\boldsymbol{i}}$, the value that the coefficient vector of (12) would take in unit *i*, had $\boldsymbol{X}_{1\boldsymbol{i}} = (X_{1i1}, ..., X_{1i,K+1})'$ been $\boldsymbol{x}_{1\boldsymbol{i}} = (x_{1i1}, ..., x_{1iK+1})'$ is independent of $\boldsymbol{X}_{1\boldsymbol{i}}$ for all *i*.[26]

It is shown in (8) or (9) that the first component of the coefficient of each nonconstant regressor in (12) keeps the values of all relevant pre-existing conditions constant. Skyrms (1988, p. 59) argued that this "comes to much the same as identifying the appropriate partition, … , [(or σ-field)] which together with … (or value of the causal variable) [$x_{1i,K+1}$] determines the chance of the effect." We show below that this partition is less adequate than the partition implied by equations (13) for our purposes.

[26] A similar admissibility condition for covariates is given in Pearl (2000, p. 79). He [Pearl (2000, p. 99)] also gives an equation that forms a connection between the opaque English phrase "the value that the coefficient vector of (12) would take in unit *i*, had $\boldsymbol{X}_{1\boldsymbol{i}} = (X_{1i1}, ..., X_{1i,K+1})'$ been $\boldsymbol{x}_{1\boldsymbol{i}} = (x_{1i1}, ..., x_{1iK+1})'$" and the physical processes that transfer changes in $\boldsymbol{X}_{1\boldsymbol{i}}$ into changes in y_{1i}.

It should be understood that the coefficient drivers in (13) are not the same as the regressors in (12). The coefficient drivers explain variations in the components of the coefficients of (12), whereas the regressors of (12) in conjunction with its coefficients explain variation in the dependent variable y_{1i}. We discuss the selection of coefficient drivers below.

We use the following matrix notation: $\boldsymbol{z}_{1\boldsymbol{i}} = (1,\ z_{1i1},...,z_{1ip})'$ is $(p+1)\times 1$, $\boldsymbol{\pi}'_{1j} = (\pi_{1j0},\pi_{1j1},...,\pi_{1jp})$ is $1\times(p+1)$, Π_1 is a $(K+2)\times(p+1)$ matrix having $\boldsymbol{\pi}'_{1j}$ as its jth row, and $\varepsilon_{1i} = (\varepsilon_{1i0},...,\varepsilon_{1iK+1})'$ is the $(K+2)\times 1$ error vector, and $\gamma_{1ij} = \boldsymbol{\pi}'_{1\boldsymbol{j}}\boldsymbol{z}_{1\boldsymbol{i}} + \varepsilon_{1ij}$ is a scalar.

Substituting the right-hand side expressions of the (K+2) equations in (13) for the (K+2) coefficients in (12), respectively, gives the result which, in matrix form, can be written as[27]

$$y_{1i} = \boldsymbol{x}'_{1\boldsymbol{i}}\Pi_1\boldsymbol{z}_{1\boldsymbol{i}} + \boldsymbol{x}'_{1\boldsymbol{i}}\boldsymbol{\varepsilon}_{1\boldsymbol{i}}\ (i = 1, \ldots, n_1)^{28} \qquad (14)$$

[27] A clarification is called for here. In (12), the number of the vectors of K + 2 coefficients increases with the number of individuals in the cross-sectional sample. So many coefficients are clearly not consistently estimable. But in (14) below, the number of unknown coefficients (Π_1) is only $(K+2)\times(p+1)$. This number does not increase with n_1. So the trick that makes our estimation procedure yield a consistent estimator of Π_1 is to include the same set of coefficient drivers across all the coefficient equations in (13) and impose appropriate zero restrictions on the elements of Π_1 if different sets of coefficient drivers are needed to estimate different components of the coefficients of (12).

[28] Any variables that are highly correlated with $\boldsymbol{x}_{1\boldsymbol{i}}$ will also be correlated with both the regression part, $\boldsymbol{x}'_{1\boldsymbol{i}}\Pi_1\boldsymbol{z}_{1\boldsymbol{i}}$, and the random part, $\boldsymbol{x}'_{1\boldsymbol{i}}\boldsymbol{\varepsilon}_{1\boldsymbol{i}}$, of the dependent variable, y_{1i}, of equation (14). Furthermore, this equation is the end result of the sequence of equations (1)-(6), and (13) that is used to avoid specification errors (i)-(iv) of Section 2.2.6. These two sentences together prove that the avoidance of specification errors (i)-(iv) leads to the nonexistence of instrumental variables. There is no contradiction between this result and HS's (2010, p. 1356) instrumental variables approach because in this paper, no use is made of their threshold crossing model which assumes separability between observables Z that affect choice and an unobservable V. Their instrumental variable is a function of Z.

Suppose that the admissibility condition on the coefficient drivers given in this section is not sufficient for the existence of the conditional expectation $E(y_{1i} \mid \boldsymbol{x}_{1i}, \boldsymbol{z}_{1i})$. Then we make

Assumption I: For all *i*, let $g(\boldsymbol{x}_{1i}, \boldsymbol{z}_{1i})$ be a Borel function of $(\boldsymbol{x}_{1i}, \boldsymbol{z}_{1i})$, $E|y_{1i}| < \infty$, and $E|y_{1i}\, g(\boldsymbol{x}_{1i}, \boldsymbol{z}_{1i})| < \infty$.

Under this assumption, the conditional expectation

$$E(\boldsymbol{y}_{1i} \mid \boldsymbol{x}_{1i}, \boldsymbol{z}_{1i}) = \boldsymbol{x}_{1i}' \Pi_1 \boldsymbol{z}_{1i} \tag{15}$$

exists (see Rao 1973, p. 97).

Assumption II: For $i = 1, \ldots, n_1$, given $\boldsymbol{z}_{1i}$, $\boldsymbol{\varepsilon}_{1i}$ is conditionally independent of $\boldsymbol{x}_{1i}$, and given $\boldsymbol{z}_{1i}$ and $\boldsymbol{x}_{1i}$, the $\boldsymbol{\varepsilon}_{1i}$'s are conditionally distributed with means zero and constant covariance matrix $\boldsymbol{E}(\boldsymbol{\varepsilon}_{1i}\boldsymbol{\varepsilon}_{1i}' | \boldsymbol{z}_{1i}, \boldsymbol{x}_{1i}) = \sigma_{1\varepsilon}^2 \Delta_{1\varepsilon}$.

Cross-sectional data for treated individuals:

$$y_{1i},\ \boldsymbol{x}_{1i}, \text{ and } \boldsymbol{z}_{1i},\ i = 1, \ldots, n_1. \tag{16}$$

The number of observations in (16) is adequate to estimate all the unknown parameters of equation (14) if $n_1 \geq (K+2)(p+1) + (K+2)(K+3)/2 + 5 - r$ where *r* is the number of restrictions on the π's. With this condition, at least 4 degrees of freedom will remain unutilized after estimating all the unknown parameters of (14).

2.3.2 Identification of model (14): Let $\otimes$ denote a Kronecker product and let vec(.) denote a column stack. Then $(K+2) \times (p+1)$ matrix Π_1 is identified if the matrix having $(\boldsymbol{z}_{1i}' \otimes \boldsymbol{x}_{1i}')$ as its *i*th row has full column rank. Even though the error vector $\boldsymbol{\varepsilon}_{1i}$

is not identifiable, the inner product $\boldsymbol{x}'_{1i}\boldsymbol{\varepsilon}_{1i}$ is identifiable, since $\boldsymbol{x}'_{1i}\boldsymbol{\varepsilon}_{1i} = y_{1i} - (\boldsymbol{z}'_{1i} \otimes \boldsymbol{x}'_{1i})\,\mathrm{vec}(\Pi_1)$. The variance-covariance matrix $\sigma^2_{1\varepsilon}\Delta_{1\varepsilon}$ is consistently estimable from feasible best linear unbiased predictors of $\boldsymbol{x}'_{1i}\boldsymbol{\varepsilon}_{1i}$. A necessary condition for the identifiability of both Π_1 and $\sigma^2_{1\varepsilon}\Delta_{1\varepsilon}$ is that the information matrix for model (14) is positive definite.

2.3.3 Identification of model (12): Because of the presence of more than one component in each coefficient of (6), we need the following identification condition: Model (12) is said to be identifiable on the basis of y_{1i}, $\boldsymbol{x}_{1i}$ and $\boldsymbol{z}_{1i}$, $i = 1, \ldots, n_1$, if the components of its coefficients are accurately estimable.

2.4 Estimation of Model (14) Under Assumptions I and II

Applying an iteratively rescaled generalized least squares (IRSGLS) method and the feasible best linear unbiased predictor to (14), we obtain the estimates of $(\Pi_1, \sigma^2_{1\varepsilon}\Delta_{1\varepsilon})$ and the predictions of $\boldsymbol{\varepsilon}_{1i}$'s.[29] Let these estimates and predictions be denoted by $(\hat{\Pi}_1, \hat{\sigma}^2_{1\varepsilon}\hat{\Delta}_{1\varepsilon})'$ and the $\hat{\boldsymbol{\varepsilon}}_{1i}$'s, respectively. Inserting these into (13) gives the estimates of the coefficients of (12). Therefore, the estimated versions of (12) and (13) can be written as

$$\hat{y}_{1i} = \hat{\gamma}_{1i0} + \sum_{j=1}^{K+1} x_{1ij}\hat{\gamma}_{1ij} \tag{17}$$

$$\hat{\gamma}_{1ij} = \hat{\pi}_{1j0} + \sum_{h=1}^{p} z_{1ih}\hat{\pi}_{1jh} + \hat{\varepsilon}_{1ij} \tag{18}$$

[29] The formulas for these estimators and predictors are given in Chang, Hallahan, and Swamy (1992) and Chang, Swamy, Hallahan and Tavlas (2000). The sampling properties of $(\hat{\Pi}_1, \hat{\sigma}^2_{1\varepsilon}\hat{\Delta}_{1\varepsilon})'$ are studied in Swamy, Tavlas, Hall and Hondroyiannis (2010).

An iteratively rescaled generalized least squares method when applied to equation (14) gives the estimates of π's and ε's in equation (18) and these estimates, in turn, give the estimates of the coefficients of (17).

2.5 Estimation of a Component of a Coefficient of (12) by Decomposition

2.5.1 Estimation of treatment effects: In this section, we estimate the TCE, $x^{*}_{1i,K+1}\alpha^{*}_{1i,K+1}$, derived in (11). If $x_{1i,K+1}$ is observed, then we use it in place of $x^{*}_{1i,K+1}$. We use (18) to estimate, $\alpha^{*}_{1i,K+1}$, which is an unobserved bias-free component of $\gamma_{1i,K+1}$, the coefficient of the treatment variable, $x_{1i,K+1}$, in (17).

Theorem: In model (6) which does not contain specification-errors (i)-(iv) discussed in Section 2.2.6, the coefficient, $\gamma_{1i,K+1}$, on the continuous treatment variable, $x_{1i,K+1}$, is equal to

$$(1- D^{*}_{1i,K+1})\, A^{*}_{1i,K+1} + (1- D^{*}_{1i,K+1})\, B^{*}_{1i,K+1} \qquad (19)$$

where $D^{*}_{1i,K+1} = (\frac{v^{*}_{1i,K+1}}{x_{1i,K+1}})$ = the proportion measurement error in the treatment variable, $A^{*}_{1i,K+1} = \alpha^{*}_{1i,K+1}$ = bias-free component, $B^{*}_{1i,K+1} = \sum_{g=K+2}^{L_{1i}} \lambda^{*}_{1ig,K+1}\alpha^{*}_{1ig}$ = omitted-regressor bias component, and $[-\hat{D}^{*}_{1i,K+1}\hat{A}^{*}_{1i,K+1} - \hat{D}^{*}_{1i,K+1}\hat{B}^{*}_{1i,K+1}]$ = measurement-error bias component of $\gamma_{1i,K+1}$. It can be seen from (19) that bias-free component and bias components of $\gamma_{1i,K+1}$ are not additively separable.

Proof: See equation (8) for the continuous $x^{*}_{1i,K+1}$ treatment variable. Q.E.D.

The choice of regressors to be included in (12) is entirely dictated by the partial derivatives of (1) we want to learn. In this paper, we want to learn only about $\alpha^{*}_{1i,K+1}$. Therefore, we reduce (12) to $y_{1i} = \gamma_{1i0} + x_{1i1}\ \gamma_{1i1}$ and consider (13) only for γ_{1i0} and γ_{1i1}.

Selection of drivers: If in the real world a causal relationship exists that determines a particular variable -- say, interest-rate spreads -- then if one of the variables -- say, *x* -- in that relationship changes, the interest-rate spread will also change. This circumstance implies that the partial derivative of the interest-rate spread with respect to *x* is nonzero. Consequently, if we had a method of obtaining consistent estimates of this partial derivative, we would be able to infer that there is a real-world relationship between the interest-rate spread and variable *x* even though we may not know the exact functional form and all the variables that comprise the relationship. Moreover, our method of obtaining consistent estimates would apply if we allow for measurement error.

To implement a parametric method for estimating consistent estimates of the partial derivative in question, two assumptions are needed. First, we assume that the stochastic coefficients of the relationship we seek to uncover are themselves determined by a set of stochastic linear equations; the set of exogenous variables in these equations are what we have above called coefficient drivers. Second, we assume that some of these drivers are correlated with the misspecification in the model -- that is the drivers "absorb" the specification errors -- and some are correlated with the variation emanating from the (true) nonlinear form. With this assumption, we can simply remove the bias from the coefficients by removing the effect of the coefficient

drivers that are correlated with the misspecification. For a valid driver, we need variables that are correlated with the misspecification.

The next step is the selection of the coefficient drivers in (13). The first point to understand is what constitutes a 'good' driver set; this issue is discussed in detail in Hall Tavlas and Swamy (2016). The basic idea presented there is that the varying coefficient γ_{1ij} will always capture the necessary variation in order to fully explain the dependent variable. This is because of the presence of the error term in the driver equation. However, to successfully decompose the coefficient into the bias free part and the biased part, the drivers must explain a large amount of the variation in the coefficient. Therefore, the first requirement for a good driver set is that it explains most of the variation in the coefficient. This result, however, can always be achieved by simply including a large number of drivers in the equation. Yet, such a prcedure would not allow a useful decomposition. Consequently, the second requirement is that the drivers must be individually relevant in explaining the movement in the coefficient -- that is, they must be statistically significant. There are several approaches that can be used for this purpose. We would suggest starting from the relevant theory in terms of selecting a large set of possible drivers by asking what variables might capture omitted variables measurement errors and non-linearities. Once a driver set is selected, there are then several options to select a suitable sub-set for actual use. This procedure amounts to using objective criteria to select relevant drivers. The procedure could include the following elements:

1. Adopt a dynamic modelling approach of general to specific, nesting down from the large set of drivers to a parsimonious, smaller set.

2. Adopt information criteria such as AIC, SBC, and pick the driver set which minimizes the criteria.
3. Hall, Tavlas, Swamy and Tsionas (2016) suggest a version of the stochastic search variable selection (SSVS) approach of George, Sun and Ni (2008), which performs well in monte carlo experiments. (See also Jochmann, Koop and Strachan (2010).)

The model can then be used in standard ways to either test an individual theory or to test between theories.

Equations (18) and (19) imply that

$$\gamma_{1i,K+1} = \hat{\pi}_{1,K+1,0} + \sum_{h=1}^{p} z_{1ih}\hat{\pi}_{1,K+1,h} + \hat{\varepsilon}_{1i,K+1} = [(1-\hat{D}^{*}_{1i,K+1})\,\hat{A}^{*}_{1i,K+1} + (1-\hat{D}^{*}_{1i,K+1})\,B^{*}_{1i,K+1}] \qquad (20)$$

This equation reconciles the discrepancies between the functional forms of the quantities on either side of its second equality sign. We have the values of all the terms on the left-hand side of the second equality sign in equation (20). From these values, it can be shown that $\hat{A}^{*}_{1i,K+1}$ and $\hat{B}^{*}_{1i,K+1}$ are equal to $(1-D^{*}_{1i,,K+1})^{-1} \times (\hat{\pi}_{1,K+1,0} + \sum_{h\in G_1} z_{1ih}\hat{\pi}_{1,K+1,h})$ and $(1-D^{*}_{1i,,K+1})^{-1}(\sum_{h\in G_2} z_{1ih}\hat{\pi}_{1,K+1,h} + \hat{\varepsilon}_{1ij})$ for some groupings G_1 and G_2 of the coefficient drivers, respectively. If these equalities hold, then we will not be committing specification errors (i)-(iv).

Although we do not have the values of $\hat{D}^{*}_{1i,K+1}$ and G_1, we can only make some reasonable assumptions about them.

Assumption III: For all *i*: (i) The measurement error $v^{*}_{1i,K+1}$ forms a negligible proportion of $x_{1i,K+1}$. (ii) Alternatively, $(\frac{v^{*}_{1i,K+1}}{x_{1i,K+1}}) \times 100$ = the percentage point which the experimenter chooses using his prior information.

Assumption III does not imply that measurement error is always absent. It implies that the measurement error must be relatively small compared to the variation in the true variable..

Under Assumption III, an estimate of the TCE on the *i*th treated individual is

$$x_{1i,K+1}\,(1\text{-}\hat{D}^{*}_{1i\,„K+1})^{-1}\,(\hat{\pi}_{1,K+1,0} + \sum_{h\in G_1} z_{1ih}\hat{\pi}_{1,K+1,h}) \qquad (21)$$

It is convenient to display the estimates in (21) for all n_1 treated individuals as kernel density estimates. The standard error of the estimate in (21) can be calculated from those of the $(1\text{-}\hat{D}^{*}_{1i\,„K+1})^{-1}\,\hat{\pi}$'s involved in (21).

To recapitulate; equation (1) is the unknown true real-world relationship. In (14), the observable variables are combined in a known functional form. We go from (1) to (14) avoiding four specification errors stated in Section 2.2.6. We go from (1) to (21) making very weak assumptions: (i) The coefficient of the dependent variable of (1) is equal to -1; (ii) equation (2) with variable coefficients gives a good approximation to (1) even when the true functional form of the latter is unknown; equation (2) has the correct functional form if its approximation error $\alpha^{*}_{c\eta 0}$ is equal to zero; we try to reduce the magnitude of $\alpha^{*}_{c\eta 0}$ using (18); (iii) equation (4) not only maintains the correct relationship between omitted and the included regressors but also makes the coefficients and the error term of (5) unique; (iv) Assumptions I and II

can hold; (v) equation (13) and Assumption III lead to very accurate estimates of TCE for every sample individual if $x^{*}_{1i,K+1}$ in (11), $(\frac{v^{*}_{1i,K+1}}{x_{1i,K+1}})$ and G_1 in (21) are known.[30] These conditions are weaker than the conditions imposed by other studies on the TCEs.

It is important to understand that this methodology will give potentially different treatment effects for each individual. The reason for this is that when we split the driver set into the set correlated with the nonlinearity and the set related to misspecification, if there are any variables in the first set beyond the constant then each individual's bias free effect will be driven by different driver variables and, hence, have different values.

Throughout this paper only one individual *i* is considered. All equations in our paper refer only to individual *i*. All these equations contain only the variables for individual *i*. These are the equations used to estimate individual level treatment effects. There is no aggregation across individuals, and we have only considered individual level data. Consequently, the equations in our paper must allow the estimation of only individual level treatment effects. The only time this methodology would give rise to a common treatment effect for all individuals would be when the set of variables associated with nonlinear effects is empty except for a constant and, hence, the underlying model is linear for all individuals.

2.5.2 Some intuition

[30] We consider below the cases where these quantities are unknown.

An intuitive account of the above derivations may be helpful at this point. We began by specifying equation (1), which we called a "real world" relationship. Equation (1) contains the following attributes. First, we did not impose a specific functional form on the relationship in equation (1); the functional form is unknown. Yet, it is general enough so that it can capture any functional relationship. Second, equation (1) contains *all* the determinants of $y_{c\eta}^{*}$, that is, both the observed and unobserved determinants. Thus, there are no omitted variables in equation (1). Third, equation (1) is stated in terms of true values of the variables so that there are no measurement errors. Fourth, equation (1) includes all relevant pre-existing conditions -- that is, conditions (such as omitted variables) that may help determine the actual structure of (1), but which cannot be specified precisely.

Next, we approximated equation (1) with equation (2), which is linear in variables but nonlinear in coefficients. This relationship can capture any linear or nonlinear relationship.[31] Equation (3) is a particular case of equation (2); specifically, it applies the specification in (2) to the particular group of *treated* individuals. Note that the set of variables represented by $\sum_{g=k+2}^{L_{1i}} x_{1ig}^{*}$ are unobserved. To eliminate these variables, we regressed each unobserved variable on all the observed variables. We did this in equation (4). This equation does not contain any mis-specified functional forms and it is exact. Therefore, there is no need for an error term.

We then substituted the two determinants of each omitted variable on the right-hand-side of equation (4) into equation (3). This substitution gave equation (5).

[31] Swamy and Mehta (1975) originated the theorem stating that any nonlinear functional form can be exactly represented by a model that is linear in variables, but that has varying coefficients. The implication of this result is that, even if we do not know the correct functional form of a relationship, we can always represent this relationship as a varying-coefficient relationship and thus estimate it. Granger (2008) subsequently confirmed this theorem.

The latter equation is a real-world relationship because its coefficients and its error term are unique.

The concept of uniqueness plays an important role in this paper, and we defined it explicitly above. Intuitively, uniqueness can be thought of as follows. Any mis-specified equation has error term, the purpose of which is to capture mis-specifications. For example, every time a relevant regression is omitted from a regression, the omitted variable is put into the error term, thereby changing the composition of the error term, while, at the same time, changing the coefficients on the included variables through omitted variable bias.[32] Such a relationship, therefore, is not *unique.* Equation (5), in contrast, possesses the property of uniqueness, as we discussed above.

In equation (6), we took account of the fact that the observed dependent variable and the observed regressors are not measured accurately. Thus, to obtain equation (6), we substituted measured values for the true values. Equation (6) contains an intercept and the coefficients of the included regressors. The components of the intercept are provided in equation (7). Equation (8) presented the components of the coefficient on an included *continuous* regressor, while equation (9) gave the components of the coefficient on a regressor that takes the value of zero with positive probability.

Equation (10) is a counterfactual to equation (3). Whereas (3) referred to the effect of a treatment on individual *i*, equation (10) gives the effect of non-treatment on the same individual. Therefore, the difference between equations (3) and (10) gives the treatment effect, which, recall, is *the difference* between the effect of the treatment

[32] This bias depends on a linear specification. It also depends on a non-zero correlation between the omitted regressor and the included regressors.

on individual *i* and the effect of non-treatment on the same individual, that is, the counterfactual. Equation (11) gave the difference between equations (3) and (10). Equation (11) is causal, since equations (3) and (10) are real-world relationships. In equation (11), $x^*_{1i,k+1}$ is the true value of the unobserved treatment variable. Since this variable is unobserved, we subsequently (*e.g.*, in equation (21)) used its observed counterpart, $x_{1i,k+1}$.

In equation (11), however, we still needed to determine $a^*_{1i,k+1}$. Note that this coefficient is the bias-free component of $\gamma_{1i,k+1}$ in equations (8) or (9), and also appears in equation (6); recall, *j* goes from *1* to *K+1* in equations (6), (8) and (9) since $a^*_{1i,j}$ in these equations goes from *j=1* to *j=K+1*. We are specifically interested in *j=K+1*.

To repeat, we need to estimate $a^*_{1i,k+1}$ in equation (11) to derive the TCE. To accomplish this, we use equation (6), which has $\gamma_{1i,k+1}$ as a coefficient. In turn, the coefficient $\gamma_{1i,k+1}$ has three components, as shown in equations (8) and (9). To estimate the components, we need to estimate the coefficients of equation (6), and then decompose them into their components. For this purpose, we use equation (13), in which the γ coefficient in equation (6), or its equivalent in equation (12), are expressed as functions of coefficient drivers.

2.5.3 Does Assumption III make the treatment effect theories untestable?

Though not directly, we have already started addressing this question in footnote 20. From this footnote it follows that for the tests of hypotheses based on misspecified models the actual Type I error will be different from the stated one and the Type II error will be very large. The likelihood functions play an important role in the tests of

hypotheses. Statisticians pointed out that the likelihood functions are model based and these models can never be wholly trusted if they are misspecified. Even though the conventional models are misspecified, model (12) is not. It is free of specification-errors (i)-(iv). For this reason, statisticians' objections do not apply to the likelihood function based on model (12). However, the estimate in (21) involving certain unknown values may get distorted by our guesses of them and these distortions will affect the Type I and Type II errors of tests of hypotheses about the TCEs on treated individuals.

2.5.3 The number of components of the coefficients of (6): If all the non-constant regressors of (6) belong to S_2, then the number of components in its intercept is as large as K + 4 and the number of components in the coefficient (γ_{1ij}) of each non-constant regressor (x_{1ij}) is 2. If (7)-(9) hold, then the number of components in the intercept of (6) is (3 + the number of non-constant regressors that belong to $\boldsymbol{S}_2$, the number of components in the coefficient of each $x \in \boldsymbol{S}_1$ is 3, and the number of components in the coefficient of each $x \in \boldsymbol{S}_2$ is 2. It should be noted that the intercept of (6) contains too many components if all the non-constant regressors of (6) belong to S_2. The number of components in the intercept of (6) is larger by the number of measurement-error biases if some of the non-constant regressors of (6) belong to $\boldsymbol{S}_2$ than if all non-constant regressors of (6) belong to $\boldsymbol{S}_1$. The difficulty of estimating the components of the intercept of (6) increases with the number of its components. For this reason, the difficulty of estimating measurement-error biases is greater if they are the components of the intercept of (6) than if they are the components of the coefficients of x's $\in \boldsymbol{S}_1$.

2.5.4 Several virtues of the regressions in (12) and (13): Under Assumption I, we do not risk attributing to the TCE in (11) that should be attributed to factors that motivate both the treatment and the outcome.

It follows from equations (3) and (10) that the TCE in (11) is for treated individual *i* and is not incorrect because of the missing counterfactual (y_{0i}) for individual *i*. The reason is that for the same treated individual *i*, we could develop two correctly specified exact mathematical models, model (3) for the treatment outcome y_{1i}^{*} and model (10) for the counterfactual y_{0i}^{*} which is what would have been the outcome had individual *i* not been treated. Because the TCE is different for different treated individuals we do not average the estimate of (11) across either the entire population or the population of treated individuals. For presentation purposes, we rely on kernel density estimates of the TCE for different treated individuals. Thus, we could overcome the complication created by the fact that the treated individuals cannot also be untreated individuals.

As mentioned by Greene (2012, p. 895), other researchers dealt with this complication by considering either pairs of individuals matched by a common observation vector $\boldsymbol{x}_i$ or paired individuals with similar propensity scores, $\boldsymbol{F}(\boldsymbol{x}_i)$ = Prob($\boldsymbol{C}_i = 1|\boldsymbol{x}_i$); in either type of pair, one is untreated with $\boldsymbol{C}_i = 0$ and the other is treated with $\boldsymbol{C}_i = 1$. It can be seen from (3) and (10) that specification errors arise if $y_{1i}^{*} - y_{0i}^{*}$ is replaced by the average value of $[(y_i \,|\, C_i = 1) - (y_{i'} \,|\, C_{i'} = 0)]$ for pairs of individuals matched by some criterion.

It follows from (3) and (10) that since we do not use these misspecified pairings, our method of estimating the TCE in (11) does not need the overlap assumption: For any value of $\boldsymbol{x}$, 0 < Prob($\boldsymbol{C}_i = 1|\boldsymbol{x}$) < 1. With this assumption, we

can expect to find, for any treated individual, an identical-looking individual who is not treated (see Greene 2012, p. 889). By developing two different models for y_{1i} and y_{0i} in this paper, we have anticipated Greene (2012, p. 889) who said that a step in the model-building exercise will be to relax the assumptions that the same regression applies to both treated and untreated states and that this regression's disturbance is uncorrelated with the treatment variable. In our specification of (3) and (10) we have taken this step.

In our approach based on (3), (10), (11) and (15) there is no need for identification by functional form (e.g., relying on bivariate normality) and identification by exclusion restrictions (e.g., relying on instrumental variables). Greene (2012, p. 889) calls these identification methods fragile assumptions. Our method also does not require computing y_{1i} - $y_{0i'}$ for pairs of individuals (i, i') matched by a common $\boldsymbol{x}_i$ or, alternatively, by similar propensity score. If these are what Greene (2012, p. 889) calls "certain minimal assumptions … necessary to make any headway at all" to estimate treatment effects, then our method does not need them.

A regression analysis of treatment effects presented in Greene (2012, 890) is based on

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \delta C_i + \varepsilon_i \qquad (22)$$

where C_i = 1 if individual i is treated and = 0 if individual i is not treated, $y_i = y_{0i} + C_i(y_{1i} - y_{0i})$ and δ is the treatment effect. The individuals themselves decide whether or not they will receive the treatment.

Greene (2012, p. 890) models program participation as

$\boldsymbol{C_i^*} = \boldsymbol{w_i'\gamma} + u_i$,

$C_i = 1$ if $\boldsymbol{C_i^*} > 0$, 0 otherwise (23)

where u_i and ε_i are correlated.

Equations (22) and (23) are not free from specification errors (i)-(iv).

It is shown in the econometric literature that C_i in (23) represents simply an endogenous variable in a linear equation. The parameterization in (23) is very different from that in (13). The approach utilizing (13) has the virtue of greater generality and of avoiding specification-errors (i)-(iv). The problem of the endogeneity of the treatment variable $x_{1i,K+1}$ in (3) and (10) does not arise because they are exact mathematical equations. The conditional expectation in (15) implies that the treatment variable in equation (14) which has several error terms is exogenous. This equation does not have C_i as its explanatory variable. We make the admissibility condition for the coefficient drivers and Assumption I.

As Greene (2012, p. 892) pointed out, there are studies casting some skepticism on the normality assumption about the error terms of selection models. Fortunately, this circumstance does not apply to either the unique error term of (5), which has a nonzero mean, or the error term of (14), which is heteroscedastic.

Underlying the mathematical equations in (3) and (10) there is no assumption that the same equation applies to both treated and untreated. This is a strong assumption. Equations (3) and (10) are for the same treated individual. Equation (11) is equal to (3) minus (10). The mathematical formula for the TCE in (11) is novel. This exact analytical measure of the treatment effect on the treated makes our study analytically complete even when data on $x_{1i,K+1}$ are not available.

Selection on some unobservables created a problem for Greene's (2012, p. 891) study. This problem is nothing but the familiar problem of the missing counterfactual y_{0i}^{*} that led to Greene's (2012, p. 891) inability to estimate an off-diagonal element of an error covariance matrix. It is not encountered in this paper.

Note that the non-constant proxy $x_{1i,K+1}$ for the treatment variable in (3) is different from the binary 0/1 treatment dummy C_i used in (20). If it were not different, then the variable $x_{1i0} \equiv 1$ for the intercept in (14) would be exactly collinear with $x_{1i,K+1} = C_i \equiv 1$ because the dependent variable of (14) is y_{1i} for all *i*. To avoid this collinearity, data on the non-constant proxy $x_{1i,K+1}$ are assumed to be available. Even if data on this proxy are not available, then this non-availability is no hindrance to the derivation of the formula in (11).

3. An Example using the ECB's Securities Market Program

In response to the global financial crisis, which erupted in 2007 with the collapse of the U.S. subprime market, and then intensified in September 2008 with the failure of Lehman Brothers, and the outbreak of the euro-area's sovereign debt crisis in late-2009 and early-2010, the ECB's Governing Council adopted a number of non-standard measures to support financial conditions and credit flows to the euro-area economy over-and-above what could be achieved through reductions in key interest rates.[33] Among those measures was the Securities Market Program (SMP). The SMP was launched in May 2010 as a response to drying up of some secondary markets for government bonds. The aim of the program was to improve the functioning of the monetary-policy transmission mechanism by providing depth and liquidity in

[33] Asset purchase programs were a part of the ECB's overall response to the two crises. For detailed review of the ECB's responses, see Cour-Thimann and Winkler (2013).

segments of the sovereign-bond market that had become dysfunctional. The program can, therefore, be thought of as a treatment for the malaise that was facing the financial system at the time.

In this section we examine the effects of the SMP on spreads on euro-area sovereigns bonds for five euro-area stressed countries -- Greece, Ireland, Italy, Portugal and Spain. There are thus five individuals, in the sense defined in our theoretical discussion above. Our data are monthly and cover the period from January 2004 through January 2013.[34] Previous studies have generally used dummy variables in an attempt to capture the effects of the program, with the exception of De Pooter, Martin and Pruitt (2015), who approximate SMP purchases based on data available from Barclays as a counterparty to the ECB, and Eser and Schwaab (2013) and Ghysels et al (2014) who use actual SMP purchases. We also use the actual amounts of sovereigns purchased under the program. These data are confidential, but were made available to us for use by the ECB. In contrast to most previous papers, we use monthly (rather than daily or intraday) data. There are two reasons why we use monthly data: (1) the confidential data on actual SMP purchases that were made available to us by the ECB are monthly; and (2) the use of monthly data allows us to control for the fundamental determinants of sovereign bond spreads.

3.1 Program description

The SMP initially focused on the purchase of Greek, Irish and Portuguese government bonds; from August 2011, Spanish and Italian government bonds were also purchased. The impact of the program should thus have been felt most in

[34] We start our estimation well before the beginning of the SMP program as we believe the longer sample period is helpful in determining the other parameters of the model and hence gives us a more accurate set of parameters to remove the omitted variable bias. We choose to use monthly data as most of the fundamental driver variables are only available at a monthly frequency.

sovereign debt markets in the stressed countries, causing the prices of sovereigns in these countries to rise and, thus, spreads (compared to the German bund) to fall. A total of 240 billion euros was spent during the course of the SMP transactions. Figure 1 shows the 10 years bond spreads for our five countries -- Greece, Ireland, Italy, Portugal, and Spain. It also delineates two periods during which purchases were being made and the timing of the Draghi announcement (in July 2012) that the ECB would do whatever was necessary to preserve the euro. As shown in the figure, the SMP took place at a time of rising spreads. Therefore, any simple correlation analysis would find that the effect of the SMP was to raise, rather than lower, spreads. Thus, finding the correct treatment effect is a matter of finding the unbiased coefficient in the presence of serious omitted variables' and measurement errors. This is precisely what we claim our technique is able to do.

The basic relationship we are interested in evaluating (based on (12)) is

$$y_{it} = \gamma_{it0} + \gamma_{it1} x_{it} \tag{24}$$

where y_{it} is the spread on sovereign bonds in country i for period t and x_{it} is the SMP expenditure in country i in period t. This equation is our basic varying parameter equation (12). As discussed earlier in this paper, each of the coefficients of (24) comprises a bias free component which we want to study. This component is corrupted by an effect for measurement error and omitted variable bias. We need to uncover the unbiased coefficient which will then be our estimate of $\partial y_{it} / \partial x_{it}$, that is the partial derivative of spreads with respect to the amount of purchases under the SMP.

To estimate (24) we proceed as follows. Our data sample includes five countries: Greece, Ireland, Portugal, and Spain. Our monthly data cover the period

from 2004M1 through 2014M7. We use the following equations for the coefficients -- that is, these are the empirical counterpart to (13), the coefficient driver equations, where the coefficient drivers are chosen as a set of fundamental variables which are widely thought to determine the soverign spreads

$$\gamma_{itj} = \pi_{j0} + \pi_{j1} RP_{it} + \pi_{j2} GB_{it} + \pi_{j3} POL_{it} + \pi_{j4} DGDP_{it} + \pi_{j5} DEBT_{it} + \pi_{j6} NEWS_{it} + \varepsilon_{it} \quad (25)$$

where RP is relative prices between the country in question and Germany, GB is the government fiscal balance relative to GDP for country i, POL is an indicator of political stability for country i, DGDP is the growth rate of GDP for country i, Debt is the stock of government debt relative to GDP for country i and NEWS is a measure of news effects with respect to the fiscal deficit of country i.[35] Detailed data definitions and sources are provided in Annex 1.

Estimation of equation for γ_{1i0} and γ_{1i1} (*i.e.*, the constant and the coefficient on SMP purchases in equation (24)) yields the following results.

$$\gamma_{it0} = \underset{(1.3)}{-1.96} + \underset{(13.6)}{97.5} RP_{it} - \underset{(0.2)}{0.007} GB_{it} - \underset{(0.6)}{0.09} POL_{it} - \underset{(2.4)}{55.4} DGDP_{it} + \underset{(3.0)}{0.03} DEBT_{it} - \underset{(8.3)}{0.01} NEWS_{it} + \varepsilon_{it}$$

$$\gamma_{it1} = \underset{(0.2)}{-0.0003} + \underset{(0.4)}{0.0009} POL_{it} + \underset{(0.7)}{0.08} DGDP_{it} + \underset{(0.006)}{0.0000001} DEBT_{it} + \underset{(0.4)}{0.000001} NEWS_{it} + \varepsilon_{it}$$

where the figures given in parentheses below each coefficient estimate are standard errors. Some variables had to be excluded from the second equation in order to

[35] NEWS is calculated from updates to forecasts of the general government balance found in the EC's Spring and Autumn forecasts.

estimate the above equations, because the SMP was only applied for a few months in each country resulting in an over-parameterization of the second equation. Since a number of these coefficients are insignificant, we successively restricted the driver equations based on the significance of variables following a general to specific methodology. The following model emerged.

$$\gamma_{it0} = \underset{(1.4)}{-1.63} + \underset{(12.9)}{83.4}RP_{it} - \underset{(1.1)}{0.14}POL_{it} + \underset{(4.1)}{0.03}DEBT_{it} - \underset{(8.4)}{0.01}NEWS_{it} + \varepsilon_{it}$$

$$\gamma_{it1} = \underset{(0.23)}{-0.0002} + \varepsilon_{it}$$

As is evident, the effect of the SMP is constant and equal to - 0.0002. Thus, our estimate of the unbiased (meaning, free of omitted-regressor and measurement-error biases) effect of the SMP on spreads is $-0.0002 \times$ SMP expenditure$_{it}$, as in (11). As there are no country specific variables in the second equation above, this implies that, in this case, the effect of the SMP programme is the same for each country. If there had been any variables remaining in the second equation then our estimate of the bias free effect in each country would have been potentially different for each country.

To put the results into context, the largest amount spent on the program in any country in any single month was 47,590 million euros. Such a purchase would have lowered the spread by 9.5 percentage points (*i.e.*, - .0002 x 47,590) or 950 basis points. In other periods, where the purchases were less, the effect would have been correspondingly less. To put this into context, Eser and Schwaab (2013) give a range of possible effects for an expenditure of 1 billion euros ranging from -1 basis point to -21 basis points. At the upper end, an expenditure of 47.5 billion euros would have

reduced spreads by 997.5 basis points.[36] Ghysels et al. (2014) report a long-run effect of the SMP within a range of 0.1 to 7 basis points for an expenditure of 100 million euros; thus a purchase of 47,950 million would have led to a 3,360 basis points fall in spreads, well above our findings that have been based on a varying-coefficient methodology and the actual purchases under the SMP.

4. Conclusions

The problem of estimating a general function with unknown functional form can be solved without introducing a single specification error by changing this problem to the problem of estimating a corresponding relationship which is linear in variables but nonlinear in coefficients. Using this solution, we showed that the causal effects of a treatment on the treated individuals can be estimated. For this estimation, we use a real-world (*i.e.*, misspecification-free) relationship between the treatment and its effect. The treatment's effect depends on the definition of causality used. In our definition, causality is treated as a property of the real world. To measure the causal effect of a treatment on the treated individuals, we take the difference between the two real-world relations, one for the effect of a treatment on a treated and another for the potential outcome of no treatment on the same individual. Obtaining pairs of treated and untreated individuals matched by similar propensity scores leads to specification errors.

[36] It is difficult to make a comparison with the results of De Pooter, Martin and Pruitt (2015) because they define their SMP variable as a percentage of outstanding debt rather than the absolute value of purchases. As far as we can discern, their result seems to yield a similar order of magnitude to our measure.

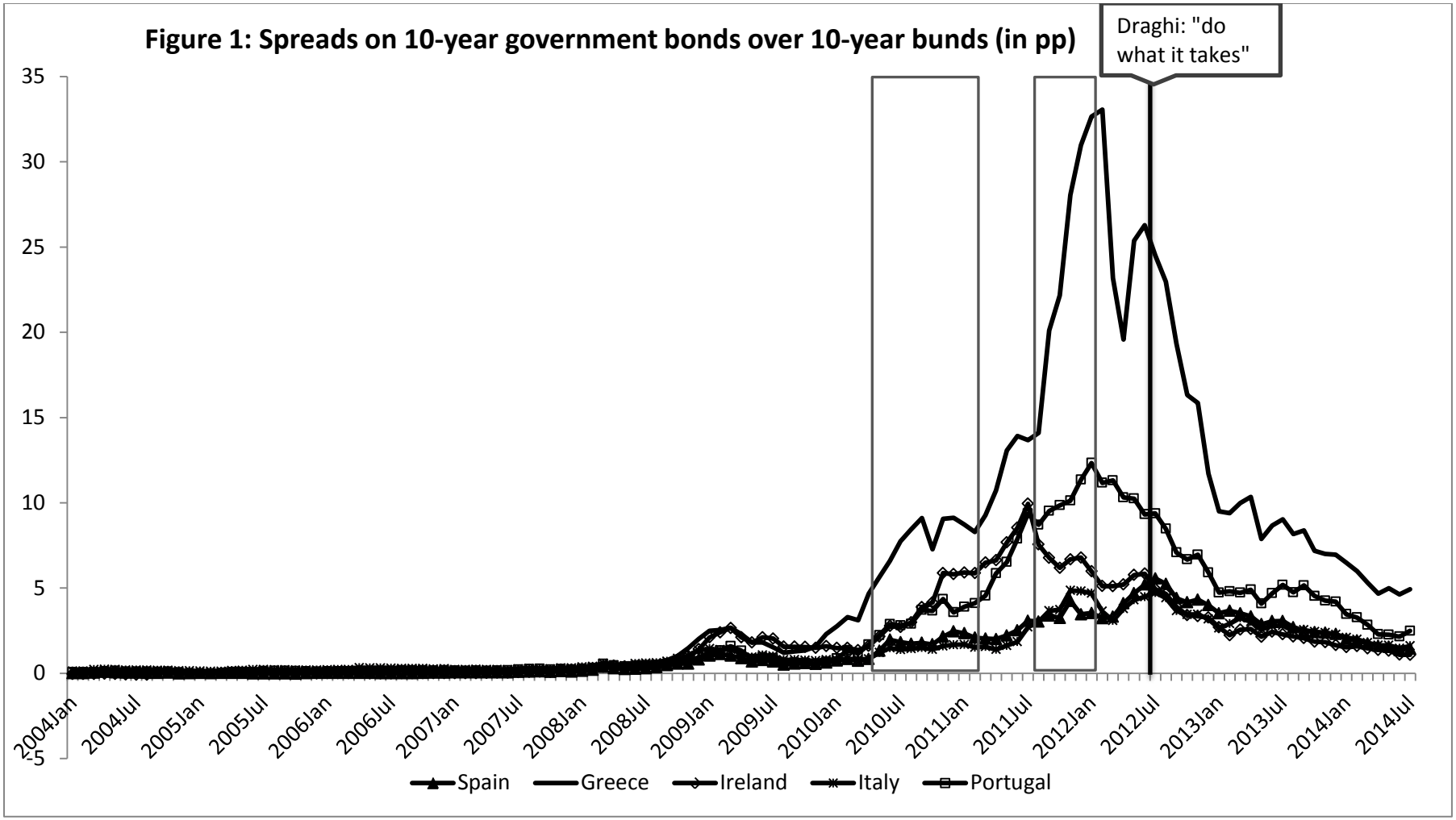
Figure 1: Spreads on 10-year government bonds over 10-year bunds (in pp)
Draghi: "do what it takes"
35
30
25
20
15
10
5
0
-5
2004Jan
2004Jul
2005Jan
2005Jul
2006Jan
2006Jul
2007Jan
2007Jul
2008Jan
2008Jul
2009Jan
2009Jul
2010Jan
2010Jul
2011Jan
2011Jul
2012Jan
2012Jul
2013Jan
2013Jul
2014Jan
2014Jul
Spain
Greece
Ireland
Italy
Portugal

References

Basmann, R. L. (1988): Causality Tests and Observationally Equivalent Representations of Econometric Models, Journal of Econometrics, Annals, 39, 69-104.

Chang, I., Hallahan, C. and Swamy, P. A. V. B. (1992): Efficient Computation of Stochastic Coefficients Models, pp. 43-53 in: Computational Economics and Econometrics. Boston: Kluwer Academic Publishers.

Chang, I., Swamy, P. A. V. B., Hallahan, C. and Tavlas, G. S. (2000): A Computational Approach to Finding Causal Economic Laws, Computational Economics, 16, 105-136.

Cour-Thimann, P., and Winkler, B., (2013): The ECB's non-standard monetary policy measures: the role of institutional factors and financial structure, Working Paper Series 1528, European Central Bank.

De Pooter M, Martin R.F. and Pruitt S. (2015): The Liquidity Effects of official Bond Market Intervention, International Finance Discussion Papers, Board of Governors of the Federal Reserve system, no. 1138.

Eser, F. and Schwaab B. (2013): Assessing asset purchases within the ECB's Securities Markets Programme, ECB Working Paper no. 1587.

Ghysels, E., Idier, J., Manganelli, S., and Vergote, O. (2014): A high frequency assessment of the ECB Securities Markets Programme, ECB Working Paper, no. 1642.

George, E., Sun, D. and Ni, S. (2008). "Bayesian stochastic search for VAR model restrictions," Journal of Econometrics, 142, 553-580.

Goldberger, A. S. (1987): Functional Form and Utility: A Review of Consumer Demand Theory. Boulder: Westview Press.

Greene, W. H. (2012): Econometric Analysis. Seventh Edition, One Lake Street, Upper Saddle River, New Jersey: Pearson/Prentice Hall.

Hall s.g., G.S. Tavlas and P.A.V.B. Swamy, 'Time Varying Coefficient Models; A Proposal for Selecting the Coefficient Driver Sets' with Macroeconomic Dynamics January, 2016. doi:10.1017/S1365100515000279

Hall, S.G, P.A.V.B. Swamy, G. Tavlas, Mike G. Tsionas (2016) Performance of the time-varying parameters model, Mimeo

Heckman, J. J. and Schmierer, D. (2010): Tests of Hypotheses Arising in the Correlated Random Coefficient Model, Economic Modelling, 27, 1355-1367. Holland, P. W. (1986): Statistics and Causal Inference, Journal of the American Statistical Association, 81, 945-960.

Jochmann, M., G. Koop, and R.W. Strachan (2010). "Bayesian forecasting using stochastic search variable selection in a VAR subject to breaks", International Journal of Forecasting 26 (2), 326-347.

Pearl, J. (2000): Causality: Models, Reasoning, and Inference, New York: Cambridge University Press.

Pearl, J. (2010): An Introduction to Causal Inference, The International Journal of Biostatistics, 6, 1-59.

Pratt, J. W. and Schlaifer, R. (1988): On the Interpretation and Observation of Laws, Journal of Econometrics, Annals, 39, 23-52.

Rao, C. R. (1973): Linear Statistical Inference and its Applications, Second edition, New York: John Wiley & Sons

Rubin, D. (1974): Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, Journal of Educational Psychology, 55, 688-701.

Rubin, D. (1978): Bayesian Inference for Causal Effects, Annals of Statistics, 6, 34-58.

Skyrms, B. (1988): Probability and Causation, Journal of Econometrics, Annals, 39, 53-68.

Swamy, P. A. V. B., Tavlas, G. S., Hall, S. G. F., and Hondroyiannis, G. (2010): Estimation of Parameters in the Presence of Model Misspecification and Measurement Error, Studies in Nonlinear Dynamics & Econometrics, 14, 1-33.

Swamy, P. A. V. B., Mehta, J. S., Tavlas, G. S. and Hall, S. G. F. (2014): Small Area Estimation with Correctly Specified Linking Models. In: Jun Ma and M. Wohar (eds.), Recent Advances in Estimating Nonlinear Models With Applications in Economics and Finance, New York: Springer.

Swamy, P. A. V. B., Mehta, J. S., Tavlas, G. S., and Hall, S. G. F. (2015): Two Applications of the Random Coefficient Procedure: Correcting for misspecifications in a small-area level model and resolving Simpson's Paradox, Economic Modelling, 45, 93-98.

Swamy, P. A. V. B., Tavlas, G. S., and Hall, S. G. (2015): On the Interpretation of Instrumental Variables in the Presence of Specification Errors, Econometrics, 3, 55-64,

Wooldridge, J. M. (2013): Policy Analysis with Pooled Cross Sections. Introductory Econometrics: A Modern Approach, Mason, OH: Thomson South-Western, 438-443.

Yatchew, A. and Griliches, Z. (1984): Specification Error in Probit Models, Review of Economics and Statistics, 66, 134-139.

Zellner, A. (1979): Causality and Econometrics, pp. 9-54 in: K. Brunner and A. H. Meltzer, eds., Three Aspects of Policy and Policymaking, Amsterdam: North-Holland Publishing Company.

Annex 1: data sources and information

Spreads (in percentage points). 10-year benchmark on each country's government 10-year bond yield minus the 10-year benchmark German government bond yield – ECB Statistical Data Warehouse – monthly average.

Covered-bond price indices. Euro area covered-bond price indices for bonds with any maturity and for those with greater than 10 years to maturity. Source: Thomson-Reuters DataStream.

Ratings. We take the ratings of each of the major credit rating agencies - Fitch, Moody's, and Standard & Poor's (S&Ps) – and construct a single series based on the agency that moved first. Ratings are mapped to a cardinal series running from 1 (AAA) to 22 (default).

Relative prices. Log difference of the monthly seasonally adjusted harmonised index of consumer prices (HICP) between each of the five countries and Germany – Thomson-Reuters DataStream.

Debt-to-GDP ratio. The ratio of the general government debt to GDP – quarterly data interpolated to monthly – Thomson-Reuters DataStream.

Political stability. We use the IFO World Economic Survey Index of Political Stability which takes values of between 0 and 10. A rise in the index implies greater stability.

Fiscal news. We construct real-time fiscal data, using the revisions to forecast general government budget balances published in the European Commission Spring and Autumn forecasts. Thus, for example, the revision to the Spring 2006 forecast is the forecast 2006 deficit/GDP ratio in the Spring compared to the forecast for 2006 made in the Autumn of 2005. This procedure allows us to generate a series of revisions (in percentage points), which, when cumulated over time, provides a real time cumulative fiscal news variable. We interpolate the series in such a way that news does not appear in the variable before it actually came out.

Economic activity. The rate of change of real GDP is interpolated to a monthly frequency – Thomson-Reuters DataStream.

SMP and CBPP. ECB.