SOAS Department of Economics
University of London

Working Paper Series

ISSN 1753 - 5816

Please cite this paper as:

QIN, D. (2016) "Let's Take the Bias Out of Econometrics" SOAS Department of Economics Working Paper Series, No. 192, The School of Oriental and African Studies.

No. 192

Let's Take the Bias Out of Econometrics

by

Duo QIN

December, 2016

Department of Economics
School of Oriental and African Studies
London
WC1H 0XG
Phone: + 44 (0)20 7898 4730
Fax: 020 7898 4759
E-mail: economics@soas.ac.uk
http://www.soas.ac.uk/economics/

The **SOAS Department of Economics Working Paper Series** is published electronically by The School of Oriental and African Studies-University of London.

This and other papers can be downloaded without charge from:

SOAS Department of Economics Working Paper Series at
http://www.soas.ac.uk/economics/research/workingpapers/

Design and layout: O. González Dávila

Let's Take the Bias Out of Econometrics

Duo QIN[1]

Department of Economics, SOAS, University of London, UK

April 2018

Abstract

This study exposes the cognitive flaws of 'endogeneity bias'. It reviews how conceptualisation of the bias has evolved to embrace all major econometric problems, despite extensive lack of hard evidence. It reveals the crux of the bias – *a priori* rejection of causal variables as conditionally valid ones, and of the bias correction by consistent estimators – modification of those variables by non-uniquely and non-causally generated regressors. It traces the flaws to misconceived error terms in statistical modelling and estimation consistency. It highlights the need to shake off the bias to let statistical learning play an active and formal role in econometrics.

JEL classification: B23, B40, C10, C50

Keywords: simultaneity, omitted-variable bias, self-selection, consistency, causality

[1] Contacting author; email: dq1@soas.ac.uk . This is a substantial revision of the earlier paper entitled 'Time to demystify endogeneity bias'. The current version is a substantial revision of the 2016 version of the working paper. During this lengthy process, I have received invaluable support, comments and suggestions from Ruben Lee, Sophie van Hüllen, Ugur Ergun, Simon Appleton, Shi Li, Lina Song, Xuheng Zang, participants at various seminars and conferences where the paper was presented, editors of *Journal of Economic Methodology* and also anonymous referees' reports. I am grateful to them all.

'Humans are masters of cognitive dissonance' (p354)
from Homo Deus by Yuval Noah Harari

1. Introduction

The notion of endogeneity bias arguably forms the keystone of econometrics. It played a pivotal role in the formalisation of econometrics during 1940s; it acts as a fundamental attribute demarcating econometrics from statistics and other disciplines overlapping with statistics. At its most fundamental, the bias arises when the ordinary least squares (OLS) estimator is applied to an *a priori* constructed model in which a causal variable is postulated to be correlated with the error term. The bias then acts as a marker to divide endogenous variables from exogenous ones. This division is succinctly described in a popular textbook by Stock and Watson (2003: 333):

> 'Variables correlated with the error term are called **endogenous variables**, while variables uncorrelated with the error term are called **exogenous variables**. The historical source of these terms traces to models with multiple equations, in which an "exogenous" variable is determined outside the model' [bold in original].

A slightly lengthier description can be found in Wooldridge's textbook (2010: 54):

> 'You should not rely too much on the meaning of "endogenous" from other branches of economics. In traditional usage, a variable is endogenous if it is determined within the context of a model. The usage in econometrics, while related to traditional definitions, has evolved to describe any situation where an explanatory variable is correlated with the disturbance.'

To illustrate this usage, Wooldridge lists three examples – 'omitted variables', 'measurement error' and 'simultaneity' (2010: 54-5). Two other cases are listed in Kennedy's (2008: 139-40) textbook – 'autocorrelated errors' and 'sample selection'. Correction of the bias entails the device of consistent estimators. A description of such estimators thus occupies the core of econometrics textbooks.

Two points are worth noting from the above quotations. First, the concept of endogeneity bias has changed significantly from its original use in the context of applying the OLS to a simultaneous-equation model. Second, the concept is fundamental as it is used to signify virtually all the major problems which economists worry about when fitting causal postulates with data – simultaneity bias, omitted variables, measurement error, autocorrelated errors, and selection bias. Indeed, textbook econometrics advocates the use of consistent estimators as the universal solution to these perceived major problems and, in doing so, spread a phobia against endogeneity bias

widely among economists. In contrast, the causal modelling community outside econometrics has concentrated increasingly on dissecting two key conditions for the adequate closure of statistical models – the causal Markov condition and the related 'faithfulness' condition – accompanied by lively development in graphic model-assisted causal structure learning by means of computers, e.g. see Wermuth and Cox (2011), and Kalisch and Bühlmann (2014).[2] Endogeneity bias has thus played a decisive role in widening the gap in research strategies between econometrics on the one hand and statistics and other related disciplines on the other.

In order to help bridge the gap, this paper probes into the conceptualisation of endogeneity bias, as defined in textbooks, to reveal its cognitive flaws. Essentially, the probe pins down the bias to the *a priori* rejection of direct translation of causal postulates of interest into statistically conditional relations, or more precisely, the rejection of the causal variable of interest as a valid conditional variable, and the consequent bias correction to modification of the variable in question by non-uniquely and non-causally generated regressors (section 2). As such, the issue should be conceptually tackled as one of causal model specification rather than estimation, an argument which has been repeatedly raised and debated in the history (the Appendix). The estimation outlook or categorisation is, however, pivotal in maintaining the bias. Mathematical derivations of consistent estimators, shown as *the* analytical solution to correct the bias, are so impeccable that they almost completely camouflage their shaky premise – existence of the error term as autonomous as economic variables (section 3). In practice, the strength of those derivations is limited severely by the facts that the error term is the residual derivative of a model and that the causally bivariate relations based on which those derivations are elaborated are much too simplistic for the economic reality. These facts help explain not only why it is impossible to measure directly and robustly the corrections in question, but also why failures are widespread in getting consistency cross-validated empirically of those estimators. Methodologically, the estimator-centred approach cannot be scientific because it seeks to reduce and entangle different sources of key econometric problems into one symptom – the presumed correlation, and to settle on analytical solutions so long as they remove this symptom. At its core, this approach depends on *a priori* model closure. Untenability of such closure is reflected in extremely naïve translations of economic reality into causally bivariate models (Section 4). Faithful translations require the profession modify its predominantly analytical-solution based standpoint to let statistical learning play an active and systematic role, especially when it comes to decisions as whether, under what

[2] A book edited by Mayo and Spanos (2010) is a rare exception. However, a search with Google Scholar yields no citations of this book by econometricians or economists once self-citations are discounted.

circumstances and/or to what degree causal postulates are, or *indeed are not*, directly translatable into statistically conditional relations.

A paradigm shift toward *a posteriori* model closure entails releasing the profession out of the conceptual trap of endogeneity bias. To facilitate this task, the rest of this paper tries to demystify the bias by offering a common and as simple as possible explanation of the crux of the bias from different sources as well as its treatment (section 2), clarifying the cognitive deficiency in judging consistency by *a priori* analytical solutions alone (section 3), and highlighting the fundamental importance for economists to shake off the bias and engage actively in searching for causally faithful and data-consistent models (section 4).

2. Endogeneity Bias: An Anatomy

The anatomy is carried out on three key sources of endogeneity bias – simultaneity bias, omitted variable bias and self-selection bias. 'Measurement error' is discussed in relation to both simultaneity bias and self-selection bias. The anatomy aims at (a) finding a common rationale upon which these biases are believed reducible to one common symptom – the correlation in question; (b) exposing the nature of the IV treatment of the correlation, a universal remedy taught in textbooks; and (c) explaining the major findings from empirical treatments of each bias.

The anatomy is focused on causal model re-specification consequences of endogeneity bias treatments beyond the estimation box, and can be viewed as an extension of the key finding in Qin (2015), namely that the IV approach essentially achieves its effect by modifying a causal variable of interest by non-uniquely and non-causally IV-generated regressors on the ground of *a priori* rejection of that variable being a *valid* conditional variable. Mathematical demonstration is kept to a minimum and causal interpretation of various models is illustrated in causal graphs to facilitate the logical exposure here.[3]

2.1. Simultaneity Bias:

Textbook demonstration of simultaneity bias is set essentially in a bivariate model. When two variables are jointly distributed, elementary probability theory dictates the following density decomposition:

(1) $f_{x,y} = f_{y|x} f_x.$

[3] Causal graphs, also known as directed acyclic graphs, are widely used in statistics and computing, e.g. see Pearl (2009), Wermuth and Cox (2011); see also Spirtes (2005) and Elwert (2013) for their potential in econometric and social research respectively.

Statistical models for causal inference are commonly based on the *conditional expectation* $E_{y|x}$ of $f_{y|x}$ where f_x is marginalised out. The concept of conditional expectation is crucial in bridging causal postulates with statistical evidence via data aggregation. Here, it underpins regression models such as:

(2) $y = \beta_{yx}x + \varepsilon_y.$

Now, the decomposition in (1) is *de facto* refuted by Haavelmo (1943, 1944) (see the Appendix), although the joint distribution, $f_{x,y}$, is endorsed in his works as being fundamental in econometrics. The refutation is embodied in the rejection of model (2) in favour of a simultaneous-equation model, such as:

(3) $$\begin{aligned} y &= \beta_1 x + \epsilon_y \\ x &= \beta_2 y + \epsilon_x \end{aligned}$$

Based on (3), Haavelmo demonstrates simultaneity bias of the OLS, $\beta_{yx} \neq \beta_1$, via a proof of $cov(x\epsilon_y) \neq 0$. But such a bi-directional position on $f_{x,y}$ makes (3) mathematically impossible for statistical estimation. This impossibility is termed 'under-identification' and circumvented by identification conditions. These conditions secure ways to decompose $f_{x,y}$ *indirectly* with the help of additional exogenous variables,[4] variables which are regarded simply as *instruments* for the consistent estimation of the 'structural' parameters, such as β_1 and β_2 in (3). Consistent estimation of a single equation in a simultaneous-equation model can be generically represented by the 2-stage least squares (2SLS), e.g. for the upper equation in (3):

(4) $$\begin{aligned} x &= V'\boldsymbol{\gamma}_{xV} + u_x \\ y &= \beta_{yx_V}\hat{x}^V + \epsilon_y^V \end{aligned}$$

where *V* denotes the IV set and $\hat{x}^V$ the OLS fitted *x* from the upper equation. β_{yx_V} is known as the IV estimator for β_1 in (3).

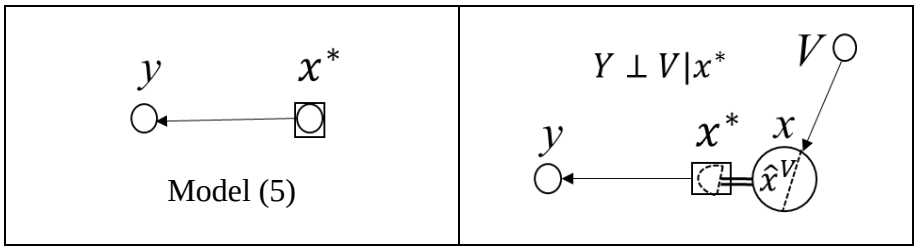
In fact, this IV estimator is not harmless with respect to (3). It acts as an implicit model modifying device to break its circular causality. In order to demonstrate this point more clearly, consider the case of an errors-in-variables model in which the explanatory variable of interest, x^*, is latent, or suffers from measurement errors:

(5) $y = \beta_{yx^*}x^* + \varepsilon_y^*; \qquad x = x^* + x''$

[4] This interpretation was implied in Wermuth's (1992) in-depth analysis of how over-parameterisation in multivariate linear structural equations results in *non-decomposable* independence hypotheses, and identification conditions help to remove the over-parameterisation so as to achieve decomposable independence.

Here, running an IV regression on the observed *x*, similar to the upper equation of (4), serves the purpose of trimming off the noisy error part, x", if feasible IVs are available. Figure 1 provides a graphic illustration of (5) and its IV treatment.

Figure 1. Errors-in-variables Model and IV Treatment



Note: The square symbol indicates a latent variable; the arrowed line indicates a probabilistically conditional relation; the disimilarity between $\hat{x}^V$ and x is shown by a semicircle (seminode) versus a circle (node); and dotted lines indicate non-uniqueness.

It is vital to note from Figure 1 that the IV treatment implies two conditions: (i) IVs should be uncorrelated to conditional expectation, $E_{y|x^*}$, and (ii) the aim of the IV regression is *not* to optimally predict *x*, as is normally expected of a regression model design, i.e. $\hat{x}^V \approx x$ must hold. The first condition is denoted by $y \perp V|x^*$ and the second by the dotted semicircle symbol in the right panel. Condition (i) is widely taught in textbooks whereas condition (ii) is apparently absent. However, this condition functions as the *de facto* backbone of the Durbin and Wu-Hausman endogeneity test, because $\hat{x}^V \approx x$ is a prerequisite for finding any possible rejections of *x* as a valid exogeneous/conditional variable. It also underpins the justification of the generalised method of moments (GMM), the generalised form of IV estimators. To see this, let us write the IV estimator of model (4) as:

$\beta_{yx_V} = (X'V(V'V)V'X)^{-1}X'V(V'V)^{-1}V'Y$, or: $\beta_{yx_V} = (X^{V\prime}X^V)^{-1}X^{V\prime}Y$, where $X^V = V'\boldsymbol{\gamma}_{xV}$.

It is clear from the above dual expression that β_{yx_V} will not differ from β_{yx} significantly without $X^V \approx X$. Hence, choice of the GMM over the OLS presumes condition (ii). This condition thus reflects the nature of the IV treatment – modifying *x* by *V* on the presumption that *x* is an invalid conditional variable whereas its IV-modelfied version, $\hat{x}^V$, is valid.

The above explains why the IV method is transferable from treating measurement errors to treating simultaneity bias. The left panel of Figure 2 depicts model (3) and the right panel the 2SLS-IV solution of (4).

Figure 2. SEM and IV Treatment via 2SLS

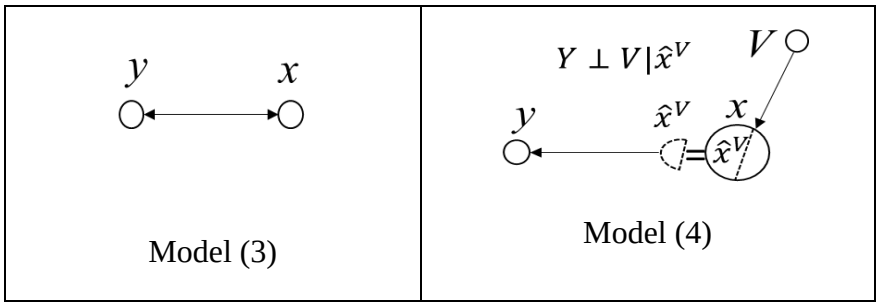


Figure 2 shows us how the bi-directional position in the left panel is broken by the addition of IVs, that is, how a simultaneous-equation model is revised into an asymmetric model through a non-unique but significant modification of *x*, because condition (i), i.e. $y \perp V|\hat{x}^V$, enables the marginalisation over $\hat{x}^V$. The modification effectively abandons the fundamental position of bi-directional simultaneity and, at the same time, utilises simultaneity to endorse the refutation of (1) in favour of the following conditional decomposition:

(6) $\quad f_{\hat{x}^V,y} = f_{y|\hat{x}^V} f_{\hat{x}^V}.$

Empirical evidence from macro modelling research, however, has not corroborated this refutation almost from the start, e.g. see Waugh (1961) and the Appendix. Although it is not difficult to find IV estimates which differ significantly from their OLS counterparts, the latter almost surely outperform the IV estimates, indicating lack of empirical consistency of those IV estimates. Nowadays, dynamics forms the core of macro-econometric models and there is plenty of evidence showing that adequately specified dynamics in Vector Auto-Regression (VAR) modelling leads to $\hat{x}^V \to x$, hence rejecting condition (ii). The evidence implies that endogenisation of *x* does not automatically nullify its conditional status with respect to another endogenous variable, i.e. the essence of Wold's 'proximity theorem'. Moreover, identification requirements have already resulted in all the estimable VARs being based a recursive structure, i.e. via the decomposition: $f_{x,y,V_{-1}} = f_{x,y|V_{-1}} f_{V_{-1}}$, e.g. see Qin (2013, Chapter 3). The postulated position of simultaneity is *de facto* abandoned. After all, statistically operational models need to start from a clearly specified 'asymmetry between cause and effect' Cox (1992: 293).

2.2. Omitted Variable Bias

This concept stems from evaluating a bivariate relation, e.g. $E_{y|x}$ from (2), on the basis of a presumed multivariate regression model. The bivariate relation could lead to biased inference when $cov(xz) \neq 0$, where z is part of the presumed multivariate model but missing in the bivariate relation. It should be emphasised that it is insufficient to equate omitted variable bias

with the mathematical discrepancy in a parameter estimates between bivariate and multivariate regressions. The discrepancy is perceived as bias only if the essential aim is to measure a partial and direct causal postulate, e.g. $x \to y$, in a multivariate model whereby other regressors are regarded as control variables. The following demonstration clarifies this point. Mathematically, there are two ways to factorise $f_{x,y,z}$ when z is considered as an extension of (1):

(7a) $f_{x,y,z} = f_{y|x,z} f_{x|z} f_z$;
(7b) $f_{x,y,z} = f_{y|x,z} f_{z|x} f_x$.

Under the linearity assumption, (7a) corresponds to a chain of regressions:

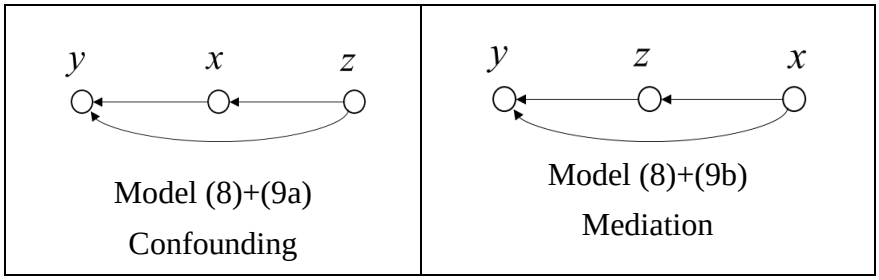
(8) $y = \beta_{yx.z} x + \beta_{yz.x} z + \nu_y$
(9a) $x = \beta_{xz} z + \nu_x$,

whereas (7b) entails (8) being followed by:

(9b) $z = \beta_{zx} x + \nu_z$.

It is only in (7a), or (8)+(9a), that the discrepancy, $\beta_{yx} - \beta_{yx.z} = \beta_{yz.x}\beta_{zx}$, is regarded problematic, a problem known as omitted variable bias from the angle of β_{yx} of the bivariate regression (2), and referred to as confounding in statistics (the left panel of Figure 3). However, it is not seen as a problem in the case of (7b) or (8)+(9b), which is known as a mediation model (see the right panel of Figure 3). The above discrepancy is no longer problematic here, because all the parameters in the dediation model are regarded as causally interpretable: $\beta_{yx.z}$ as the direct effect, the product, $\beta_{yz.x}\beta_{xz}$, as the indirect effect and β_{yx} as the total effect.

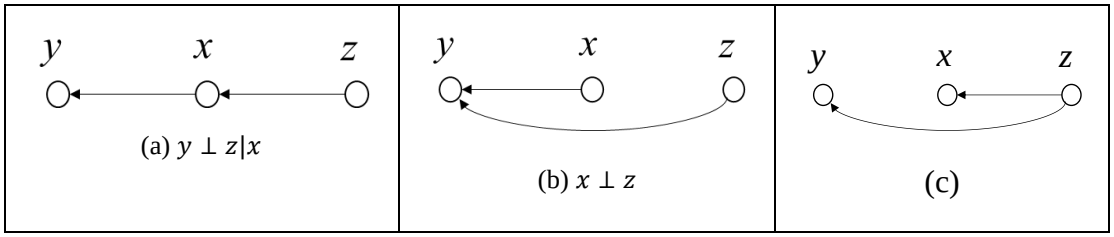
Figure 3. Two types of regression chains



In order to understand how omitted variable bias is treatable via IV estimation, we need to analyse the three special cases of (7a) illustrated in Figure 4.[5] Omission of z is permissible in case (a) because $\beta_{yz.x} = 0$, only this case is rarely feasible in economic reality. Omission of z in case (b) is also valid when $cov(xz) = 0$ or $\beta_{zx} = 0$ is verified. Case (c) is precluded as irrelevant by the partial stance of $x \to y$.[6]

[5] See Cox and Wermuth (2004) for more discussion of these cases.
[6] Notice that maintaining model (2) in case (c) leads to nonsense regression.

Figure 4. Three Special Cases of (7a) or (8)+(9a)

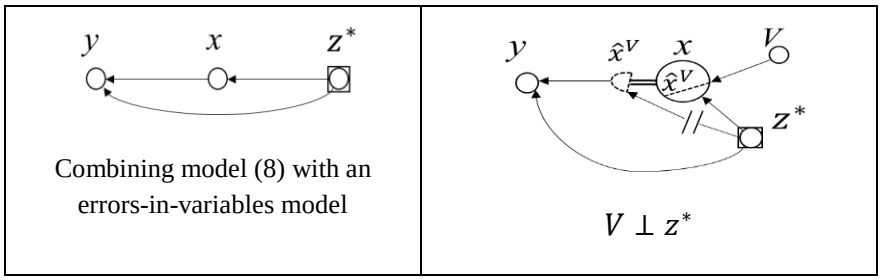


Case (b) underpins the IV treatment of omitted variable bias. Imagine the following situation: *z* is not observable from available data but is claimed to correlate with *x* by substantive knowledge. Lack of data results in *z* being included analytically into the error term of (2), $\varepsilon_y = \beta_{yz.x} z + v_y$, hence the diagnosis: $cov(x\varepsilon_y) \neq 0$. This leads to the conceptual amalgam of omitted variable bias with endogeneity bias, even though simultaneity is irrelevant here. Specifically, the IV treatment is to assume that there exist certain IVs which are known to be uncorrelated with z^*, the latent *z*, such that they help filtering out the correlation part of *x* with ε_y. The treatment amounts to a modification of the conditional decomposition (7a) to:

(10) $\quad f_{\hat{x}^V,y,z^*} = f_{y|\hat{x}^V,z^*} f_{\hat{x}^V} f_{z^*},$

similar to the simultaneity-bias case. But unlike that case, the present modification aims at restoring the feasibility of a bivariate model as far as the causal inference of $y \rightarrow x$ is concerned, i.e. $E_{y|\hat{x}^V}$ from (10), because of $V \perp z^*$ by design, even though the maintained model is still multivariate. Figure 5 illustrates this modification – from the left panel in Figure 3 into case (b) in Figure 4 through deactivating the chain effect from z^*.

Figure 5. Latent OVB and IV Treatment



The deactivating function of the IV treatment is immensely attractive to applied economists whose interest is confined to certain highly partial causal postulates and who are worried about the risk of omitted variable bias when their postulates are put directly to data. The IV treatment offers them an expedient short-cut to stay justifiably with knowingly over-simplistic models whereby the size and properties of the resulting residuals become ignorable. The assumed latent

status of *z* further fosters the belief in the IV approach as a universally effective precautious treatment against any potential risk of omitted variable bias.[7] Meanwhile, condition (i) of IV choice implies that these instruments are *non-causal* and hence must be harmless for the intended causal inference. That is why IV treatments of omitted variable bias have enjoyed far more empirical vitality than those of simultaneity bias.

Nevertheless, the nature of the IV treatment of modifying the intended causal variable is easier to be noticed here than in the simultaneity-bias case. This is reflected from repeated critiques of the credibility in interpreting $E_{y|\hat{x}^V}$ as causally equivalent to $E_{y|x,z}$, e.g. see Deaton (2010) and the Appendix. After all, case (b) is only a special case of (7a). When $x \perp z$ fails to hold, it is impossible to reduce (7a) into (b) unless we modify *x* to remove its correlation part with *z*. This modification cannot be innocuous to the causal interpretability of the model outcome with respect to the intended postulate, $y \to x$.

2.3. Self-Selection Bias

Concerns over self-selection bias arise from models using cross-section samples with incomplete observations. A classic example is the case of estimating wage elasticity in labour supply models where (reservation) wage rates for those have reported not working, i.e. zero working hours, are missing in household survey data. Let us modify a multivariate model, such as (8), for incomplete data samples:

$$y_i = \beta_{yx.z}x_i + \beta_{yz.x}z_i + \nu_{i,y} \quad \text{when} \begin{cases} y_i > 0, & x_i > 0 \\ y_i = 0, & x_i \; are \; missing \end{cases} \tag{11}$$

where i denotes sample observation, *y* is a 'limited dependent variable', and *x* is the key causal variable of interest. Obviously, replacing those missing values by zero will result in a biased estimate of $\beta_{yx.z}$ when substantive knowledge dictates that those missing observations are not zero.

This bias is conceptually amalgamated into endogeneity bias via the Heckmen two-step procedure. The procedure explains the bias as the result of a self-selection decision, namely that responses from the agents of the observed part of *x* are biased representation of the whole population because of their self-selection decision to join one of the sub-groups divided by the threshold, e.g. zero in (11). This decision is presented by a probit model of a binary variable derived from the truncation:

[7] For example, the treatment is perceived as a safeguard of the *ceteris paribus* condition, e.g. see Angrist and Pischke (2015, Introduction).

(12) $$d_i = V_i'\boldsymbol{\gamma}_{dI} + u_{i,d} \qquad d_i = 1 \text{ when } x_i > 0, \quad d_i = 0 \text{ when } x_i = 0$$

where *V* is a set of IVs, and $cov(u_d v_y) \neq 0$ is asserted in the proof of x_i in (11) suffering from omitted variable bias. To correct the bias, an inverse Mill's ratio, r_i,[8] is generated from (12) to extend (11) into:

(13) $$y_i = \beta_{yx.rz}x_i + \beta_{yz.rx}z_i + \beta_{yr.x}r_i + v_{i,y}^V \qquad when\ x_i > 0$$

A significant $\beta_{yr.x}$ is perceived as empirical verification of this self-selection bias.

The assertion of self-slection behavior as a cause of omitted variable bias in (11) is very appealing to economists working with incomplete data samples. It binds the bias conceptually to endogeneity bias, since equation (12) has effectively endogenised *x* into an explained variable, albeit indirectly via *d*. Simultaneity is no longer needed to justify endogeneity bias. Empirically, evidence of significant $\beta_{yr.x}$ is relatively easy to obtain, thanks to non-unique choices of *V* and ubiquitous presence of collinearity among economic variables, see the Appendix. However, it is doubtful whether significant $\beta_{yr.x}$ counts as hard evidence of self-selection bias as a particular type of omitted variable bias, because it is not only theoretically impossible to uniquely define the inverse Mill's ratio, but also practically impossible to find evidence of both $\beta_{yx.rz}$ and $\beta_{yx.rz}$ are significantly collinear with r_i in (13), especially when there are multiple control variables. The latter impossibility reveals a conceptual gap between sample selection bias and omitted variable bias. While omitted variable bias is individual regressor based, sample selection bias is model based in that it rejects the inferability of an estimated model across different categories/classifications of a hypothetical population. This difference becomes apparent if we impose $V \perp z$ in (12) to minimise the collinear effect of r_i on *z* in (13).

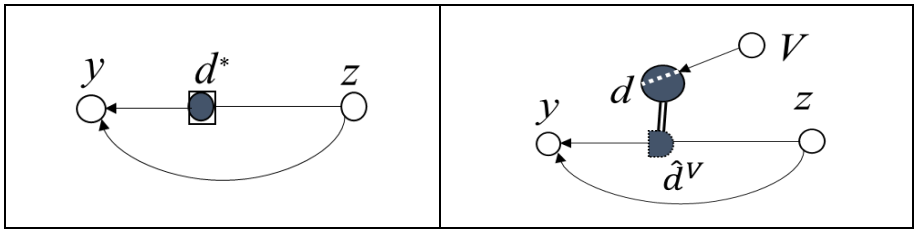
Therefore, Heckman's concept of self-selection bias is pertinent to endogeneity bias rather than to sample selection bias. This becomes transparent in the research of programme evaluation models, where self-selection behaviour forms a major threat to the randomisation condition. Such behaviour is presented as an endogenous dummy variable model:

(14) $$\begin{aligned} y_i &= \beta_{yd^*.z}d_i^* + \beta_{yz.d^*}z_i + v_{i,y} \\ d_i &= V_i'\boldsymbol{\gamma}_{d^*V} + u_{i,d^*} \end{aligned} \quad \Rightarrow \quad d_i^* = V_i'\widehat{\boldsymbol{\gamma}}_{d^*V}$$

[8] $r_i = \frac{\phi(\gamma_{x.V}V_i)}{\Phi(\gamma_{x.V}V_i)}$, where $\phi(\cdot)$ and $\Phi(\cdot)$ stand respectively for the density and cumulative density of standard normal distribution.

where d_i^* is the ideally randomised programme participation variable. The IV treatment here bears a strong resemblance to the errors-in-variables model (5), when Figure 6 is compared to Figure 1.

Figure 6. Endogenous Dummy Variable Model and IV Treatment

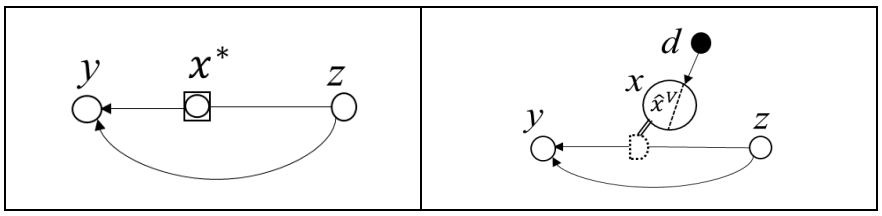


A further mutation extends (14) to a general latent variable model, where the key explanatory variable of interest is believed to suffer from measurement errors due to self-section behaviour in response to a policy change. Dummy variables representing such changes are used as IVs to circumvent the bias, e.g. see Angrist and Krueger (1991) for the case of using compulsory school laws as IVs to model the returns of education. The associated model can be written as:

$$\begin{aligned} y_i &= \beta_{yx^*.z} x_i^* + \beta_{yz.x^*} z_i + \nu_{i,y} \\ x_i &= \gamma_{x^*d} d_i + u_{i,x^*} \;\Rightarrow\; x_i^* = \hat{\gamma}_{x^*d} d_i \end{aligned} \tag{15}$$

where *x* is assumed to suffer from measurement errors. Causal diagrams of this case is given in Figure 7. Noticeably, neither (14) nor (15) targets at the missing data problem of incomplete data samples. What they target at is to produce a latent conditional variable which satisfies the ideally randomised condition, assuming that the observed causal variable is conditionally invalid. However, once the observed variable is modified, interpretation of the estimated effect by the latent variable as the observed policy effect becomes problematic. That explains why empirical results from these models have aroused serious interpretation debates, e.g. as reflected in the case of 'ATE' (average treatment effect) being modified to 'LATE' (local average treatment effect) described in the Appendix.

Figure 7. A Variation and Extension of Figure 7



The lineage of self-selection bias from endogeneity bias explains how endogeneity bias has evolved into an all-bias-inclusive concept, making it an almost professional phobia the claimed correlation between causal variables of interest and OLS-based error terms. Self-selection bias also serves us as the easiest case to recognise the nature of the IV treatment beyond the estimation box – modification of key causal variables by non-uniquely and non-causally IV-generated regressors. Clearly, few applied modellers would be willing to have their carefully postulated causal variables modified arbitrarily on non-causal grounds without empirical verification. What has misled many of them into actually doing so is the apparent logical necessity of consistent estimation after the correlation is algebraically shown to be present in their *a priori* postulated models. Hence, we need re-examine this necessity and, in particular, two key concepts involved – the error term and consistent estimation.

3. The Error Term and Consistency

Consider, first, the OLS-generated error term sustaining the claim of $cov(x\varepsilon_y) \neq 0$. It should be noted that textbook proofs of this correlation are based to a bivariate model setting. These proofs are, however, not transitive to multivariate models, because the error term has changed from ε_y to ν_y when we move from (2) to (8). Proofs of $cov(x\nu_y) \neq 0$ would require making assumptions concerning not only the possible decomposition of ν_y with respect to individual *z*s but also the relationship between those *z*s and *x* in (8). Short of empirical support for the assumptions needed, the proofs would lose credibility, if they are still mathematically tractable. This implies that proofs as such cannot be logically attempted before we can fix and uniquely define the error term. That requires us to fix model formulation. After all, the error term under concern is a by-product of fitted models, unlike variables. Since empirical models on the same topic often vary in formulation, the absence of a uniquely and unambiguously defined error term explains why it is impossible to have direct measures of the premised correlation as the conclusive evidence of the bias. Sadly, there is inadequate awareness among the profession of the lack of direct transitivity of the textbook proofs of endogeneity bias based on a bivariate-model setting to the multivariate-model setting in reality, and also the lack of model-independent nature of the error term under concern.

It should also be noted that the error term under concern is of an inferential or inductive nature, in that the bias is pertinent to the out-of-sample errors. The OLS is unbiased with respect to in-sample errors by definition. Different properties between out-of-sample errors and in-sample errors can help us pinpoint the root of the misconceived endogeneity bias. Denoting the

commonly used measure of mean square errors for these two types of errors as E_{in} and E_{out} respectively for a model, *M*, explaining *y*, we have:[9]

(16) $$E_{in} = \frac{1}{N}\sum\left(y_{in} - (\widehat{y_{in}}|M)\right)^2 = V(y_{in} - \widehat{y_{in}}|M)^2 + [E(y_{in} - \widehat{y_{in}}|M)]^2$$

where *N* is the sample size. When an unbiased estimator is used, the second term vanishes in (16) and E_{in} becomes equal to the variance of in-sample errors. However, this does not apply to the second term in E_{out}:

(17) $$E_{out} = V(y_{out} - \widehat{y_{out}}|M)^2 + E\{\left(y_{out} - (\widehat{y_{out}}|M)\right)\} = Var_M + Bias_M^2$$

even when in-sample unbiased estimators are used. That is because the in-sample errors are specified/assumed as randomly distributed and thus *known* by the estimator choice, whereas the specification/assumption does not apply to the out-of-sample errors since they are *unknown* unknowns by definition.[10] Two implications follow. First, any postulate of $cov\left(xv_{\widehat{y_{out}}}\right) \neq 0$ should not be taken as truth without post-sample validation/testing. Second, any *a priori* conviction of the presence of endogeneity bias effectively assumes the bias = $Bias_M$ in (17) for the untested model in question. This assumption amounts to taking out-of-sample errors as *known* unknowns.

The connexion of endogeneity bias with E_{out} enables us to gain a clear understanding as why the 'usage' of endogeneity bias has moved from its 'simultaneity' tradition in macro-econometrics to its current position, i.e. the situation stated in the second quotation in section 1. Severe predictive failures of the economic recession triggered by the 1973 oil price shock formed the key impetus of the dynamic modelling reforms in macro-econometrics, accompanying the rational expectations movement in macroeconomics. Those failures serve effectively as irrefutable post-sample validations that static simultaneous-equation models are severely biased for omitting dynamic features. Subsequently, the collective shift to VAR type of dynamic models and the resulting empirical evidence accrued from the shift has *de facto* verified Wold's proximity theorem that endogenization of causal variables does not necessarily invalidate their conditional

[9] Description of the two types of errors in association with model selection and assessment is given a pronouced place in statistical learning textbooks, e.g. see Abu-Mostafa *et al* (2012), James *et al* (2013), Sharlev-Shwartz and Ben-David (2014).

[10] Historically, the *unknown* nature of the error term has long been conceived by various leading econometricians. For example, Frisch classified statistical variations into three types – systematic variations, accidental variations and disturbances and assigned the latter two to the error term, see Bjerkholt and Qin (2010, Chapter 3). In the Cowles Commission works, the error term was described as 'the joint effect of numerous separately insignificant variables that we … presume to be independent of observable exogenous variables' Marschak (1953, p. 12). Subsequently, the error term was generally described as 'the effect of all those factors which we cannot identify for one reason or another' (Malinvaud, 1966, p. 74). However, none of these descriptions has been formally linked to the error term of bivariate regression models where endogeneity bias is defined in textbooks. See also Qin (2013, Chapter 8) for a history of the error term in time-series econometrics.

status, i.e. the induced OLS bias is negligibly small, as already mensioned in Section 2.1. In micro-econometrics, on the other hand, prediction is generally seen as irrelevant for cross-section data analyses, and the potential role of cross-validation in model inferential testing is largely disregarded.[11] In fact, 'little emphasis' is placed on 'residual analysis' of in-sample errors in the micro community (Cameron and Trivedi, 2005, p. 289), despite, paradoxically, of the fact that the widely held belief in endogeneity bias is based on the out-of-sample error term. Without explicit access to evidence on E_{out}, applied modellers are thus trapped in constant fear for endogeneity bias presumably caused by selection bias and omitted variable bias, oblivious of the fact that their fear amounts to imposing untenable assumptions on the out-of-sample error term.

The connexion of endogeneity bias with E_{out} also helps us classify the misconceived bias conviction. It amounts to a type III error – ' "errors of the third kind" (giving the right answer to the wrong question)' (Hand, 1994, p.317), see also (Kennedy, 2002, p.572). Although IV solutions are correct in removing assumed correlation between a specific regressor and the OLS-generated error term, the conviction is wrongly posed, as shown at length in the previous section. To recap, the conviction rejects without trial any untested causal variables as valid conditional variables when the causal relation under concern is formalised into a simple regression model and, at the same time, subject knowledge indicates the prospect of likely complications to that model by interdependent variables, omitted factors and inadequately representative samples. The correct question should be: Given our knowledge of likely complications to our partially and incompletely conceived hypotheses, what are empirically adequate models in which it is possible to verify a postulated causal variable as a valid conditional variable for inferential purposes?

Unfortunately, the route to the correct question is formally blocked by a powerful deterrent – the paramount importance of using consistent estimators. Now, let us turn to the notion of consistency. Judgment of consistent estimators in circumvent endogeneity bias is based on proofs of $cov(x\varepsilon_y) = 0$. But this criterion is predicated on $cov(x\varepsilon_y) \neq 0$ and hence not independent from its fallacy, as discussed above. Hence, arguments for consistent estimation serve only as a camouflage of the fallacy. In other words, the role of correct answers to a wrong question is to maintain rather than verify that question. This tautological nature of the practice of choosing consistent estimators by $cov(x\varepsilon_y) = 0$ was actually exposed by Pratt and Schlaifer three decades

[11] Cross-validation, i.e. the practice of splitting available data into training and testing subsets and utilising the testing errors as proxies for out-of-sample errors, is reviewed in a very sceptical tone by Leamer in the chapter 'Model choice and specification analysis' in the Handbook of Econometrics (1983a); it is briefly described as part of kernel estimation procedure in (Cameron and Trivedi, 2005, Chapter 9).

ago (1984; 1988).[12] Essentially, what they argue is the primacy of having an empirically adequate model before examining the possibility of any alleged correlation, as it is impossible to acquire adequate knowledge of what the error term represents without an 'exhaustive exploration' of the omitted variable problem to settle on such an adequate model.

In fact, attention on gauging estimator selection by consistency has already been dissipated and shifted towards the use of the concept for model selection purposes in the statistical profession. Conceptually, the practice of assessing estimators is defined as 'internal consistency' for its subordinate status to model design, which should be aimed at ensuring 'the consistency of the model with data' (Cox, 2006, Chapter 1 and Appendix B). Cox maintains that 'although internal consistency is desirable, to regard it overwhelmingly predominant is in principle to accept a situation of always being self-consistently wrong as preferable to some inconsistent procedure that is sometimes, or even quite often, right' (ibid, p. 199). Mathematically, quests for data-consistent models have led to rigorously defined notions of 'consistent' hypotheses, along with 'uniform convergence' and 'nonuniform learnability', and/or 'consistent' learning rules in the development of machine learning theories, e.g. see Mitchell (1997, Chapter 7), Sharlev-Shwartz and Ben-David (2014, Chapter 7). Subsequently, the need for consistent learning algorithms with respect to hypotheses under examination becomes seen as of minor importance. The asymptotic condition of consistent hypotheses helps researchers explicitly subject model design and selection to the minimisation of the bias component in E_{out} of (17). The formalisation is also concordant with active research in formalising causal inference, where various conditions have been elaborated for the empirical adequacy of causal models in terms of their conditional independence, e.g. conditions for collapsibility, ignorability, faithfulness and unconfoundedness. All these conditions are targeted at having the Markov property on the inferential error term in causal chain model designs, e.g. see Cox and Wermuth (1996; 2004), Pearl (2009), and also Shalizi (2017, Part III).[13] Imposition of the Markov condition on E_{out} is clearly more stringent than that of consistent hypotheses. Nevertheless, it reinforces conceptually the fundamentality of model-data consistency over estimator consistency.

[12] See also Swamy *et al* (2015) for a recent revisit and extension of their arguments.

[13] Methodological implications of that research have also engaged the attention of philosophers, e.g. see Glymour (2010) and Russo (2014). For recent studies on the faithfulness condition, see Spirtes (2009), Zhang and Spirtes (2011). For discussions by researchers from other social science disciplines, see Morgan (2013), and Kalisch and Bühlmann (2014).

4. Unfaithful Translation and *A Priori* Model Closure

The previous section pins the belief in endogeneity bias down to misconceptions over the error term and the over-emphasis on estimation consistency. Section 2 reveals the nature of the IV treatment of the bias – rejection of directly translating causal postulates into conditional models by modifying and substituting the causal variables of interest with IV-generated non-optimal predictors. This section delves further into the methodological roots and consequences of this mistranslation. A statement by Cox (2006: p197) summarising his life-long experience serves as the best starting point:[14]

> 'Formalization of the research question as being concerned with aspects of a specified kind of probability model is clearly of critical importance. It translates a subject-matter question into a formal statistical question and that translation must be reasonably faithful and, as far as is feasible, the consistency of the model with the data must be checked. How this translation from subject-matter problem to statistical model is done is often the most critical part of an analysis.'

In econometrics, this translation job has been formally delegated to economists, e.g. see Qin (1993). They are expected to formalise their causal postulates into mathematically consistent models, which are labelled as *structural* models. The task of econometricians is to provide statistically consistent estimators of the parameters given in those models. This task is precedent on *a priori* model closure such that the model in question is treated as *maintained* once it is relayed to the econometrics leg.[15] Under this theory-based approach, statistical learning is formally ruled out as far as the translation job is concerned.

This division of labour manifests itself clearly in the mainstream econometric teaching and practice. One and arguably the most salient feature is the core position that consistent estimation methods are placed in textbooks. Specification testing methods, in comparison, are taught mainly as means for detecting data complications, viewed from the stance of simplistically formulated models. The popularity of those tests depends frequently on whether their applications would justify and result in the use of mathematically more complicated estimators. Data-instigated modelling is prejudiced against as practising 'measurement without theory'.[16] In large cross-section sample based micro-econometric practice, in particular, low model fits in terms of tiny

[14] Insightful discussions among statisticians on the strategic importance of formulating statistical questions scientifically can also been found in Hand (1994), Senn (1998) and Breiman (2001).
[15] Methodologically, the issue on positions of model closure is closely related to the recurring debate over realism of econometrics and economics as well, e.g. see Sims (1980), Hoover (2001), Mäki (2002), Romer (2016).
[16] See the disputes on measurement without discovery versus measurement without theory between J. Koopmans and R. Vining published in *Review of Economics and Statistics* in the late 1940s, e.g. Qin (1993, Chapter 6).

R^2 statistics are seen as a normal consequence of both the very partial nature of the structural models of interest and the noisy nature of nonexperimental data. The possibility that such low fits may indicate inadequate model design in terms of significantly underfitting data is widely ignored. This attitude contrasts sharply to the prevalent view in statistical learning that *a priori* postulated models tend to suffer from underfitting bias in E_{out}, a natural consequence of theoretical abstraction.

The previous sections have discussed at length how proofs of endogeneity bias are predicated on *a priori* closure of over-simplistic bivariate models. Here, it is worth pointing out a hazardous consequence of this premature closure. While the pursuit of bias-free consistent parameter estimators appears utterly faithful to the *a priori* formulated structural model, solutions of the pursuit are *de facto* betrayal of this faithfulness in that the key causal variable of interest has been modified before any empirical tests are tried out on whether and under what circumstances this variable is translatable into a conditional variable, the most commonly accepted translation in statistical modelling. Sadly, the modification has evaded most economists' attention. Once their postulated parameters are granted the *structural* status, the issue that there is no unique estimator is out of the realm of theorists' concerns, and accordingly there lacks awareness that estimator choices could possibly alter the intended causal postulates. This blind spot is further camouflaged by the process of 'identification', namely the categorisation of all *ad hoc* model amendments to incompletely closed empirical models. Since identification is taught as a necessary step for estimation, any additional model amendments here is assumed harmless to the initial causal postulates. This explains why disputes over arbitrary identification conditions have never ceased among applied economists, while such disputes have hardly touched the theoretical community.[17] When the step of identification is systematically incorporated into IV estimators, its distortional effect on the causal postulate of interest, if noticed at all, is typically blamed on poor IV choices in practice. Although the IV method was shown to function as a 'generated regressor' producer by Pagan (1984) decades ago, implications of his description on causal modelling have been long overlooked.

The above discussion highlights the strategic defect in the theory-based modelling approach. But what has led econometrics being dominated by such an approach in general, and ended up in the conceptual impasse of endogeneity bias in particular? Several historical factors stand out. The need for recognition from economists of econometrics as a new subdiscipline was

[17] The latest resurgence can be found in Romer (2016). Another careful and recent examination of the conceptual links between identification, IVs, exogeneity and omitted variables from the angle of the experimental approach versus the structuralist approach can be found in Erik Biørn (2017).

crucial in the advocacy for the *structural* approach, which sanctioned the division of labour discussed above. The choice to formalise econometric methods on the basis of a simultaneous-equation model setting reinforced the belief that those methods were universally applicable to economic issues because the model setting was targeted at the very foundation of mainstream economics – the general equilibrium theory. Of those major figures who pioneered the formalisation, few had first-hand experience with data and none was the data-exploratory type. It is not surprising that they failed to comprehend Wold's arguments on the importance of clearly specified *conditional* models for causal modelling. As a result, the nexus between causal postulates and stochastically conditional models was virtually lost in the formalisation. In fact, econometricians' awareness of the link between conditional expectation and regression models remains rather limited to this day.[18] Although the concept of conditional independence is included in textbooks nowadays, it is taught merely as a statistical assumption associated with the error term.

Obviously, lack of computing technology, adequate and good quality data should also be taken into account. Noisy data, often in small samples, plus computational difficulties have led econometricians to lean heavily on mathematic formalisation of statistical techniques, a tendency encountered also by statisticians to a certain extent, e.g. see Freedman (1991), Hand (1994) and Breiman (2001). In a culture where inadequate attention is given to the strategic issue of whether those *a priori* formalised models match faithfully to the empirical tasks at hand, mathematical trackability has naturally driven those formalisation attempts to be based on highly simplistic models. Once the model type is fixed on a bivariate regression, it is almost irresistible not to perceive any data complications as a single symptom – correlation between the regressor and the OLS-generated error term.

The bivariate model base and its *a priori* closure destines 'endogeneity bias' to a fictitious existence. That existence, in turn, confines applied research in a fictitious world. The concept loses its grip in empirical studies whose findings rely heavily on forecasting accuracy, e.g. a wide range of macro-modelling research as mentioned before. It remains thriving in areas where empirical results are evaluated virtually solely by the statistical significances of estimates of one or two predestined *structural* parameters in models offering highly partial causal explanations of the data at hand. These models are usually presented to serve the practical purpose of policy

[18] In the *Econometric Theory* (ET) interview of David Hendry, he recalled how the audience at the 1977 European Econometric Society conference was bewildered by J.-F. Richard's presentation, which used conditional-expectation based sequencing to formalise the concept of exogeneity (Ericsson and Hendry, 2004). Another telling example of related communication failure can be found in the discussion of Wermuth (1992) between A.S. Goldberger and statisticians.

evaluation. Since conclusive empirical evidence is hard to come by for policies implemented in uncontrolled environments, making a good story becomes the essential goal. In the circumstances, it is widely regarded as inconsequential to assess the empirical results by the goodness of model fit, or the degrees of precision and constancy of the structural parameter estimates. On the other hand, precautionary measures against endogeneity bias are taken as necessary when the bias is perceived as a fundamental feature of economic data. Hence, use of consistent estimators enhances the persuasive power of the story by helping maintain the unfalsifiable status of the models. This enhancement is, however, likely to lose effect when cross-validation is performed on those consistent estimators. For example, none of the IV estimates demonstrates trends of convergence whereas none of the OLS counterparts trends of bias in E_{out} with increasing sample sizes in a cross-validation exercise in the re-examination of the US returns-to-education case by van Hüllen and Qin (2017); the same finding is presented by Young (2017) with a more elaborate inference design and on a much wider scale – 1533 IV estimates in 1400 2SLS regressions from 32 publications in the journals of the American Economic Association. These inferential test results indicate that direct translation of postulated causal variables into conditional variables is more faithful than modifying them by IV-generated synthetics.

From a discipline perspective, although belief in endogeneity bias has worked in favour of research topics where empirical findings are relatively hard to falsify, knowledge gain from data there is often dismally low, especially in studies working with large data samples. Meanwhile, many practical topics which beg for statistical learning have been demarcated into other disciplines such as management, marketing and business administration. A paradigm shift in this research culture is imperative if we intend to keep up with the rapid advances in computing power, artificial intelligence, data collection and accumulation, e.g. see Einav and Levin (2014), Rust (2016).

A decisive step needed for the shift is to recognise the conceptual flaw of endogeneity bias, thereby removing its traction. By probing into the roots of the bias, our discussion has laid bare the specious qualities of the bias and its shaky cognitive foundation. The success of applied models should stem from drawing together the relative advantages of both substantive knowledge and data analysis. Few can dispute the following. On the one hand, substantive knowledge is relatively good at identifying key causes, but not good at identifying the appropriate functional forms of empirical models or other minor causes which are not ignorable in estimating the effects of the key causes. On the other hand, data is the best possible source for obtaining the missing

knowledge necessary for the formulation of empirically adequate models. Econometric practice that disregards data knowledge in model design, and that camouflages deficiencies in model design by estimators which effectively modify key causal variables in non-causal ways against what has originally intended in theory, can only be called 'alchemy', not 'science' (Hendry, 1980).

Appendix: A Brief History of Endogeneity Bias

The traditional usage of the term 'endogeneity bias' referred to by Wooldridge stems from Haavelmo's 1943 exposition of simultaneity bias when the OLS was applied to a simultaneous-equation model. The related history has been well studied, e.g. see Christ (1952), Epstein (1987; 1989), Qin (1993). In the work of the Cowles Commission during the 1940s, solutions to the problem were formalised into the device of multiple-equation based consistent estimators preceded by a step of parameter identification of SEMs. Herman Wold (1954; 1956; 1960) was almost the sole voice standing against the above approach at the time and attributed the problem to inadequately formulated causal models in terms of conditional expectations. It took several decades, however, before Wold's causal modelling idea won a *de facto* victory through reforms in dynamic macro-econometric modelling. The victory was best reflected in a general reduction of concern about simultaneity bias as the VAR type of models became embraced by the macro modelling community, e.g. Sims (1980) and also Qin (2013).

Although it took a few decades for econometricians' attention on simultaneity bias to wane, empirical evidence of the bias was scant almost from the start. Initial experiments by Haavelmo failed to yield significant OLS bias (1947), see also Girshick and Haavelmo (1947). Similar results were also obtained in subsequent investigations, e.g. see Christ (1960), and led to Waugh's verdict (1961) endorsing the OLS as adequate for applied purposes. This verdict has been repeatedly verified in various applied cases since then. Amazingly, all these empirical results were anticipated by Wold's 'proximity theorem', which shows simultaneity bias to be practically negligible in a simultaneous-equation model when the model is adequately specified in terms of its causal chain - see Wold and Juréen, (1953: 37-8).

One conceptual issue which emerged as problematic during the dynamic macro-econometric modelling was the endogenous-exogenous classification, e.g. see Aldrich (1993). This has led to a formal redefinition of 'exogeneity' by Richard (1980) and Engle *et al* (1983). Their aim was to clarify under what conditions certain explanatory variables of interest were valid conditional variables in statistical models. Three levels of exogeneity were identified – 'weak'

exogeneity based essentially on causal reasoning, 'strong' exogeneity via time sequencing and 'super' exogeneity via cross-regime invariance.[19] Noticeably, the latter two are based on statistical criteria, and their evaluations are shown to rest on the parameters of interest, i.e. those representing the effects of the exogenous variables. The redefinition thus highlights the close link between the parameters of interest and causal variable specification, as well as the empirical importance of moving away from the simultaneous-equation-model based tradition to asymmetric causal model specifications. Pursuit of empirical dynamic modelling has led to the development of a more data-instigated research route, e.g. see Hendry (1995; 2009), Hendry and Doornik (2014).

While concern about simultaneity bias was dissipating among macro modellers, the possibility of correlation between one causal variable and the error term attracted new attention in micro-econometric research. This research was pioneered mainly by James Heckman (1976; 1978; 1979) in the context of models explaining either categorical variables or censored/truncated variables using incomplete cross-section data samples, a type of models known as 'limited dependent variable' models, e.g. see Maddala (1983).[20] The research followed the direction set by James Tobin nearly two decades earlier. When he tried to model durable goods consumption, Tobin encountered the situation where some observations of the dependent variable in his cross-section sample were missing effectively due to censoring. Replacing those missing observations by zeros to produce a complete sample, Tobin (1958) showed that the OLS estimator was bias and devised a maximum likelihood estimator to circumvent the bias following the principle of probit, which is now widely referred to as 'tobit'. Heckman's research during the 1970s was to extend the strategy of tobit to a situation where some observations of a causal variable of interest was missing in relation to a censored or truncated dependent variable, such as the (reserved or potential) wage rate of non-labour-force participants in labour supply models of cross-section survey data, see van der Klaauw (2014) for a more detailed review. This led him to the interpretation of this causal variable as the result of self-selection bias, i.e. a decision choice on group participation, and to the translation of the truncation-induced OLS bias into a bias caused by an omitted but correlated variable, the inverse Mill's ratio, a variable derived from a binary probit model representing the decision choice (see Section 2.3). Derivation of the inverse Mill's

[19] Outside econometrics, invariance is shown to be a strong condition for causal linear stochastic dependence by Steyer (1984; 1988) in psychometrics. Cross-sample invariance is subsequently shown by Steyer *et al* (2000) as a necessary condition to ensure causal regression models not suffering from OVB. See also Freedman (2004) on the importance of invariance in non-experimental data regression analyses.

[20] A brief historical account of this research and also the subsequent developments in programme evaluation methods is given in Qin (2015, 2.2). The following description is written to complement rather than repeat that account.

ratio led to an extension of the single-equation limited dependent model into a two-equation one, hence the term 'the Heckman two-step procedure'. The extension has also brought the truncated causal variable closer to an endogenous variable because its truncated feature is now explained by the decision equation with the two-equation model.

On the applied front, evidence of self-selection bias appeared much easier to obtain than that of simultaneity bias, if judged by the statistical significance of the inverse Mill's ratio. However, it gradually transpired that such evidence lacked robustness in that it depended on the extensive presence of collinearity among possible control variables. It thus proved impossible to pin down a unique inverse Mill's ratio to verify conclusively the presence of self-selection bias, e.g. see Puhani (2002). Moreover, it was found that there is frequently a negligibly small difference between an IV treatment of endogeneity bias in its simultaneity sense and the bias combined with self-selection bias on an 'endogenous' variable, such as wage rate, e.g. see Blau and Kahn (2007), Qin *et al* (2016). These findings suggest a very weak link between self-selection bias and simultaneity bias, but a rather strong one between self-selection bias and multicollinearity consequent to model extension to avoid omitted variable bias. The latter severely questions the inverse Mill's ratio as an effective measure of the truncation effect explained by Heckman's self-selection bias model.

The link of self-selection bias with sample truncation effects disappeared while its link with endogeneity bias strengthened in the subsequent development of programme evaluation methods. These were developed mainly during the post 1980 period and drew heavily on the self-selection bias literature, e.g. see Cameron (2009, 14.5) and Wooldridge (2010, 21.1). Obviously, outcomes of any policy-driven programmes come after their implementation. Simultaneity is thus irrelevant by construction. But self-selection behaviour is not because some participants of the programmes could be self-selected rather than randomly selected. When average treatment effect (ATE) models were adopted from medical science for evaluating social programmes, randomisation failures were regarded as a major challenge, e.g. see Heckman (1992).[21] In addition to sample selection problems concerning the comparability between the treated group and the control group, self-selection behaviour was considered un-ignorable on substantive grounds.[22] Heckman's presentation of endogenous dummy variables demonstrated an attractive route to tackling this issue along the path of simultaneous-equation models (1978). Once the ATE was attached to an

[21] A wider methodological discussion on randomisation and related model specification issues is by Leamer (1983b), from which the title of the present paper stems.

[22] Such behaviour is referred to as 'selection on unobservable' in textbooks as opposed to 'selection on observable', which covers both omitted variable bias and sampling selection concerning comparability of the two groups.

endogenous dummy variable, self-selection bias correction became associated with randomisation and the IV route was resorted to naturally, e.g. see Heckman (1996).

On the applied side, this IV route has been strongly promoted by Angrist and his associates through a series of studies, see Angrist (1990), Angrist and Krueger (1991), and Angrist and Pischke (2009). While their studies helped popularise the prevalence of endogeneity bias, their applications gave rise to serious debate over the interpretability of IV-generated estimates such as the ATE, e.g. see Angrist *et al* (1996, with discussion). As a result, their interpretation was narrowed down to *local* ATE (LATE). Noticeably, this revised interpretation implies a partial recognition of the causal-modifying capacity of IVs, and in particular that IV-modified programme dummies might no longer fully represent the programme implemented in reality. Angrist and Pischke (2015, p227) also acknowledged the possibility of the IV choice ending up with 'a failed research design'.

Similar debates have recurred in the field of development economics (see *Journal of Economic Literature*, 2010, no 2). There, the key problem of IV-assisted quasi-randomisation is criticised as a fundamental misunderstanding of exogeneity (Deaton, 2010). In contrast to the highly theoretical style of causality analysis by the Cowles Commission in rivalry with Wold's causal chain modelling arguments over half a century ago, the issues examined by Deaton are widely and closely relevant to policy related applied modelling research, , e.g. see also Deaton and Cartwright (2017). The accumulation of fragile and imprecise IV estimates, which have been produced out of concern over the presence of endogeneity bias, has reached such a state that it is no longer possible for the wide applied community to maintain faith in this approach.

References

Abu-Mostafa, Y. S., Magdon-Ismail, M. and Lin, H.-T. (2012) *Learning From Data*, AMLBook.

Aldrich, J. (1993) Cowles exogeneity and core exogeneity, *Discussion Papers in Economics and Econometrics*, University of Southampton, No 9308.

Angrist, J. (1990) Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review* **80**: 313-36.

Angrist, J., and Krueger, A. (1991) Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics* **106**: 979-1014.

Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**: 444-55.

Angrist J.D. and Pischke, J. (2009) *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist J.D. and Pischke, J. (2015) *Mastering 'Metrics: The Path from Cause to Effect*, Pinceton University Press.

Biørn, Erik (2017) Identification, instruments, omitted variables, and rudimentary models: Fallacies in the 'experimental Approach' to econometrics, Memorandum No 13/2017, Department of Economics, University of Oslo.

Bjerkholt, O. and Qin, D. (eds.) (2010) *A Dynamic Approach to Economic Theory: The Yale Lectures of Ragnar Frisch in 1930*, Routledge.

Blau, F.D. and L.M. Kahn (2007) Changes in the labor supply behaviour of married women: 1980-2000. *Journal of Labor Economics* **25**: 393-438.

Breiman, Leo (2001) Statistical modelling: The two cultures, *Statistical Science* **16**(3): 199-231.

Cameron, A.C. (2009) Microeconometrics: Current methods and some recent developments, in K. Patterson and T.C. Mills (eds.), *Palgrave handbook of econometrics*, vol. 2, Palgrave MacMillan, pp. 729-74.

Cameron, A.C. and P.K. Trivedi (2005) *Microeconometrics: Methods and Applications*, Cambridge University Press.

Christ, C.F. (1952) History of the Cowles Commission, 1932-1952, in *Economic Theory and Measurement: A Twenty Year Research Report 1932-1952*. Cowles Commission for Research in Economics, pp. 3-65.

Christ, C.F. (1960) Simultaneous equations estimation: Any verdict yet? *Econometrica* **28**: 835-45.

Cox, D.R. (1992) Causality: Some statistical aspects. *Journal of Royal Statistical Society Series A.* **155**: 291–301.

Cox, D.R. (2006) *Principles of Statistical Inference*, Cambridge University Press.

Cox, D.R. and N. Wermuth (1996) *Multivariate Dependencies: Models, Analysis and Interpretation*, Chapman & Hall.

Cox, D.R. and N. Wermuth (2004) Causality: A statistical view. *International Statistical Review* **72**: 285-305.

Deaton, A. (2010) Instruments, randomization, and learning about development. *Journal of Economic Literature* **48**: 424-55.

Deaton, A. and N. Cartwright (2017) Understanding and misunderstanding randomized controlled trials, *Social Science & Medicine*, https://doi.org/10.1016/j.socscimed.2017.12.005

Einav, L. and Levin J. (2014) Economics in the age of big data. *Science*. **346**(6210): 715-21.

Elwert, F. (2013) Graphical causal models, in S.L. Morgan ed. *Handbook of Causal Analysis for Social Research*, Springer, Chapter 13, pp. 245-73.

Engle, R.F., Hendry, D.F., and Richard, J.-F. (1983). Exogeneity. *Econometrica* **51**: 277–304.

Epstein, R. (1987) *A History of Econometrics*. Amsterdam: North-Holland.

Epstein, R., (1989) The fall of OLS in structural estimation. *Oxford Economic Papers* **41**: 94-107.

Ericsson, N.R. and D.F. Hendry (2004) The ET Interview: Professor David F. Hendry, *Econometric Theory* **20**: 745-806.

Freedman, D.A. (1991) Statistical models and shoe leather, *Sociological Methodology* **21**: 291-313.

Freedman, D.A. (2004) On specifying graphical models for causation, and the identification problem. *Evaluation Review* **28**: 267-93.

Girshick, M.A. and T. Haavelmo (1947) Statistical analysis of the demand for food: Examples of simultaneous estimation of structural equations. *Econometrica* **15**: 79-110.

Glymour, C. (2010) Explanation and truth, in Mayo and Spanos (eds.) (2010), pp. 331-50.

Haavelmo, T. (1943) The statistical implications of a system of simultaneous equations. *Econometrica* **11**: 1–12.

Haavelmo, T. (1944) The probability approach in econometrics, *Econometrica* **12**: Supplement.

Haavelmo, T. (1947) Methods of measuring the marginal propensity to consume. *Journal of the American Statistical Association* **42**: 105-22.

Hand, David J. (1994) Deconstructing statistical questions, *Journal of Royal Statistical Society A*, **157**, Part 3, 317-56.

Heckman, J. (1976) A life-cycle model of earnings, learning, and consumption. *Journal of Political Economy* **84**: S11-S44.

Heckman, J. (1978) Dummy endogenous variables in a simultaneous equation system. *Econometrica* **46**: 931-59.

Heckman, J. (1979) Sample selection bias as a specification error. *Econometrica*, **47**: 153-61.

Heckman, J. (1992) Randomization and social program, in C. Manski and I. Garfinkel (eds.), *Evaluating Welfare and Training Programs*, Harvard University Press, pp. 201-230.

Heckman, J. (1996) Randomization as an instrumental variable. *Review of Economics and Statistics* **78**: 336-41.

Hendry, D.F. (1980) Econometrics: Alchemy or science. Economica, **47**: 387-406.

Hendry, D.F. (1995) *Dynamic econometrics*. Oxford University Press.

Hendry, D.F. (2009) The methodology of empirical econometric modelling: Applied econometrics through the looking-glass, in Patterson, K. and Mills, T. C. (eds.) *Palgrave Handbook of Econometrics*, vol. 2. Palgrave MacMillan, pp. 3-67.

Hendry, D.F. and J.A. Doornik (2014) *Empirical Model Discovery and Theory evaluation*, MIT.

Hoover, K.D. (2000) *The Methodology of Empirical Macroeconomics*, Cambridge University Press.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani (2013) *An Introduction to Statistical Learning*, Springer.

Kalisch, M. and P. Bühlmann (2014) Causal structure learning and inference: A selective review. *Quality Technology & Quantitative Management* **11**: 3-21.

Kennedy, P. (2002) Sinning in the basement: what are the rules? The ten commandments of econometrics, *Journal of Economic Survey* **16**: 569-89.

Kennedy, P. (2008) *A Guide to Econometrics* (6th edition), Wiley-Blackwell.

Leamer, E. E. (1983a) Model choice and specification analysis, in Griliches and Intriligator eds. *Handbook of Econometrics*, vol 1, North-Holland, pp. 285-328.

Leamer, E. E. (1983b) Let's take the con out of econometrics, *American Economic Review* **73**: 31-43.

Maddala, G.S. (1983) *Limited-Dependent and Qualitatively Variables in Econometrics*, Cambridge University Press.

Mäki, U. (ed.) (2002) *Fact and Fiction in Economics*, Cambridge University Press.

Malinvaud, E. (1966) *Statistical Methods in Econometrics*. North-Holland.

Marschak, J. (1953) Economic Measurements for Policy and Prediction, in Hood, W. C. and T. Koopmans (eds.), *Studies In Econometric Method*, Yale University Press, pp. 1-26.

Mayo, D.G. and A. Spanos (eds.) (2010) *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, Cambridge University Press.

Mitchell, Tom, M. (1997) *Machine Learning*, McGraw Hill.

Morgan, S.L. (ed.) (2013) *Handbook of Causal Analysis for Social Research*, Springer.

Pagan, A. (1984) Econometric issues in the analysis of regressions with generated regressors. *International Economic Review* **25**: 221-47.

Pearl, J. (2009) *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pratt, J.W. and R. Schlaifer (1984) On the nature and discovery of structure. *Journal of the American Statistical Association* **79**: 9-21.

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

Puhani, P.A. (2002) The Heckman correction for sample selection and its critique. *Journal of Economic Survey* **14**: 53–68.

Qin, D. (1993) *Formation of Econometrics: A historical perspective,* Oxford University Press.

Qin, D. (2013) *A History of Econometrics: The Reformation from the 1970s*, Oxford University Press.

Qin, D. (2015) Resurgence of the endogeneity-backed instrumental variable methods. *Economics: The Open-Access, Open-Assessment E-Journal*, 9(2015-7), 1-35.

Qin, D., van Hüllen, S., and Wang, Q.-C. (2016) How Credible Are Shrinking Wage Elasticities of Married Women Labour Supply? *Econometrics*, **4**(1).

Richard, J.-F. (1980) Models with several regimes and changes in exogeneity. *Review of Economic Studies* **47**: 1-20.

Romer, P. (2016) The trouble with macroeconomics, Working Paper, Stern School of Business, New York University.

Russo, F. (2014) What invariance is and how to test for it, *International Studies in the Philosophy of Science* **28**: 157-83.

Rust, John (2016) Mostly useless econometrics? Assessing the causal effect of econometric theory, *Foundations and Trends® in Accounting* **10**: 2-4, 125-203.

Senn, Stephen (1998) Mathematics: governess or handmaiden? *Statistician* **47** Part 2: 251-59.

Shalev-Shwartz, Shai and Shai Ben-David (2014) *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press.

Shalizi, Cosma Rohilla (2017) *Advanced Data Analysis from an Elementary Point of View* (manuscript from http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/.).

Sims, C.A. (1980) Macroeconomics and reality. *Econometrica* **48**: 1-48.

Spirtes, P. (2005) Graphical models, causal inference, and econometric models, *Journal of Economic Methodology*. **12**:1, 1-33.

Spirtes, P. (2009) Variable definition and causal inference, *Proceedings of the 13th International Congress of Logic Methodology and Philosophy of Science*, pp. 514-53.

Steyer, R. (1984) Causal linear stochastic dependencies: The formal theory, in E. Degreef, and J. van Buggenhaut (eds.) *Trends in Mathematical Psychology*. North-Holland, pp. 317-46.

Steyer, R. (1988) Conditional expectations: An introduction to the concept and its applications in empirical sciences. *Methodika* **2**(1): 53-78.

Steyer, R., S. Gabler, A.A. von Davier and C. Nachtigall (2000) Causal regression models II: Unconfoundedness and causal unbiasedness, *Methods of Psychological Research Online*, **5**, No. 3.

Stock, J.H. and M.W. Watson (2003) *Introduction to Econometrics*, Addison-Wesley.

Swamy, P.A.V.B., G.S. Tavlas and S.G. Hall (2015) On the interpretation of instrumental variables in the presence of specification errors. *Econometrics* **3**: 55-64.

Tobin, J. (1958) Estimtion of relationships of limited dependent variables. *Econometrica* **26**: 24-36.

van der Klaauw, B. (2014) From micro data to causality: Forty years of empirical labor economics, *Labour Economics* **30**: 88-97.

van Hüllen, S. and D. Qin (2017) Compulsory Schooling and the Returns to Education: A Re-examination, revised version of *SOAS Department of Economics Working Paper Series*, No. 199.

Waugh, F.V. (1961) The place of Least Squares in econometrics. *Econometrica* **29**: 386-96.

Wermuth, N. (1992) On block-recursive regression equations (with discussion). *Brazilian Journal of Probability and Statistics* **6**: 1-56.

Wermuth, N. and D.R. Cox (2011) Graphic Markov models: Overview, in J. Wright (ed.) *International Encyclopedia of Social and Behavioral Sciences* (2nd ed.), Elesevier, **10**: 341-50.

Wold, H.O.A. (1954) Causality and econometrics. *Econometrica* **22**: 162–177.

Wold, H.O.A. (1956) Causal inference from observational data: A review of ends and means. *Journal of Royal Statistical Society,* Series A, **119**: 28–61.

Wold, H.O.A. (1960) A generalization of causal chain models (Part III of a Triptych on Causal Chain Systems). *Econometrica* **28**: 443–463.

Wold, H.O.A. and Juréen, L. (1953) *Demand Analysis: A Study in Econometrics*, Wiley and Sons, New York.

Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data* (2nd edition), The MIT Press.

Young, Alwyn (2017) Consistent without inference: Instrumental variables in practical application, LSE Working Paper.

Zhang, J., and Spirtes, P. (2011) Intervention, determinism, and the causal minimality condition, *Synthese*, **182**:13, 335-47.