SOAS Department of Economics
University of London

Working Paper Series

ISSN 1753 - 5816

Please cite this paper as:

QIN, D. (2015) "Time to Demystify Endogeneity Bias" SOAS Department of Economics Working Paper Series, No. 192, The School of Oriental and African Studies.

No. 192

Time to Demystify Endogeneity Bias

by

Duo QIN

October, 2015

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.

Design and layout: O. González Dávila

Time to Demystify Endogeneity Bias

Duo QIN[1]

Department of Economics, SOAS, University of London, UK

Abstract

This study exposes the flaw in defining endogeneity bias by correlation between an explanatory variable and the error term of a regression model. Through dissecting the links which have led to entanglement of measurement errors, simultaneity bias, omitted variable bias and self-selection bias, the flaw is revealed to stem from a Utopian mismatch of reality directly with single explanatory variable models. The consequent estimation-centred route to circumvent the correlation is shown to be committing a type III error. Use of single variable based 'consistent' estimators without consistency of model with data can result in significant distortion of causal postulates of substantive interest. This strategic error is traced to a loss in translation of those causal postulates into multivariate conditional models appropriately designed through an efficient combination of substantive knowledge with data information. Endogeneity bias phobia will be uprooted once applied modelling research is centred on such designs.

JEL classification: B23, B40, C10, C50

Keywords: simultaneity, omitted variable, self-selection, multicollinearity, consistency, causal model, conditioning

[1] Email: dq1@soas.ac.uk .

The greatest truth is revealed in the simplest terms (Chinese proverb)

Endogeneity bias is arguably the most characteristic notion which differentiates econometrics from statistics and other disciplines overlapping with statistics. It played a pivotal role in the formalisation of econometrics during 1940s. In recent decades, it has posed as a major econometric deterrent to the rapid and lively developments of graphic model assisted causal structure learning and inference in statistics as well as other related disciplines, e.g. see Wermuth and Cox (2011), Kalisch and Bühlmann (2014).[2]

In fact, the two camps seem to have been moving in opposite directions when it comes to causal modelling research. While the causal modelling community outside econometrics has focused themselves increasingly intensely on dissecting two key conditions for adequate closure of statistical models – the causal Markov condition and the related faithfulness condition, econometricians have compounded almost all major problems into endogeneity bias and thus widened the scope for estimation-centred treatments of the bias. The latter move is neatly summarised in a popular textbook by Stock and Watson (2003): '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' (p. 333). A slightly lengthy explanation is given in Wooldridge's textbook: '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

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

'simultaneity' (2010, pp. 54-5). Two other cases are listed in Kennedy's textbook – 'autocorrelated errors' and 'sample selection' (2008, pp.139-40).

The central aim of this paper is to disentangle conceptual confusions in the present usage of endogeneity, which is thereafter referred to as *endogeneity bias syndrome* to better reflect its all-bias-inclusive feature. Specifically, our dissection is focused on three types of bias – simultaneity bias (SB), omitted variable bias (OVB) and self-selection bias (SSB). Their common crux in modifying conditional specifications is carefully demonstrated (Section 2). 'Measurement error' is discussed in relation to both the first and the third types. 'Autocorrelated errors' is considered as a special case of OVB in Section 3, where major methodological issues exposed in Section 2 are further discussed. The concluding section attributes the fundamental flaw of endogeneity bias syndrome to a tragic loss in the translation of substantive knowledge based causal postulates directly into appropriate conditional statistical models. Several historical causes of the loss are traced there. Costly consequences of the loss are also highlighted. The discussion implies how practically exigent it is now to make good this translation in order to dispel the endogeneity bias phobia in the present big-data era. But before proceeding to those sections, a brief historical account is provided first. Readers who are not interested in the history or already familiar with it can skip this background section.

1. From SB to SSB: A Brief History

The 'traditional usage' referred to by Wooldridge stems from Haavelmo's 1943 exposition of SB of the ordinary least squares (OLS) in the context of simultaneous-equation models (SEM). The related history has been well studies and documented, e.g. see Christ (1952), Epstein (1987; 1989), Qin (1993). What is worth repeating here is that the problem was defined, mainly through the works of the Cowles Commission (CC) group during the 1940s, as one of SEM identification and treated with the help of multiple-equation based estimators. Herman Wold was probably the only person at the time who objected to this Haavelmo-CC

diagnosis and attributed the problem to inadequately formulated causal models in terms of conditional expectations (1954; 1965; 1960). But it took several decades before Wold's causal chain model idea won a *de facto* victory through macro-econometric reforms towards dynamic modelling. The victory was best reflected from a general loss of concerns over SB as the VAR (Vector Auto-Regression) type of models became embraced by the macro modelling community.[3]

Although it took a few decades to have the shine off SB academically, empirical evidence of SB stayed thin almost from the start. Initial experiments by Haavelmo failed to yield significant OLS bias (1947), see also Girshick and Haavelmo (1947). Similar results were corroborated from 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 have been anticipated by Wold's 'proximity theorem', which shows SB becoming practically negligible in an SEM when the model is adequately specified conditionally in terms of its causal chain (Wold and Juréen, 1953, pp. 37-8).

While SB was losing grip of macro modellers, the symptom of endogeneity bias, i.e. correlation between one explanatory variable and the error term of a regression, caught new attention in microeconometric research in the context of limited dependent variable (LDV) models, pioneered mainly by Heckman (1974; 1976; 1987; 1979).[4] Prior to 1970, the OLS bias in LDV models had already been well understood and tackled by maximum likelihood based estimators such as probit and tobit. What triggered microeconometric research to reorient its direction was the practical problem that one explanatory variable of interest in a LDV model was also of the limited variable (data-truncated) type, e.g. wage rate of labour supply in the

[3] See Qin (2013) for a detailed study of the history of this reformative period.

[4] 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.

context of cross-section survey data. The problem led to an adjustment of the research zoom lens from a LDV regression as a whole to a particular explanatory variable wherein. This new lens enabled Heckman to relate the OLS bias to bias of a particular omitted variable – the inverse Mill's ratio, a derivative of the truncated error term of the LDV model. Correlation between the truncated explanatory variable of interest and the error term was thus established. Derivation of the inverse Mill's ratio led to an extension of a single-equation LDV model into a two-equation one, which closely resembled the two-stage least squares (2SLS) representation of the instrumental variable (IV) solution to SEMs (see the next section).

Mathematically, Heckman's two-equation LDV model is vital to linking OVB with endogeneity bias, since the OVB problem of the original equation is now related to the SB problem via the added binary equation. Conceptually, the power of the link derives from Heckman's interpretation of the added equation: self-selection decision by the micro agents involved. Once the truncated data feature of the explanatory variable of interest is explained by a decision-making process, its endogenous property becomes unquestionable, reinforcing the textbook labour market theory on the basis of simultaneity between wage and quantity of labour supply.

On the applied front, evidence of SSB appeared much easier to obtain than that of SB, if judged by statistical significance of the inverse Mill's ratio. However, it gradually transpired that such evidence lacked robustness in that it depended on extensive presence of collinearity among possible control variables such that it was impossible to pin down a unique inverse Mill's ratio to verify conclusively the presence of SSB, e.g. see Puhani (2002). Moreover, the difference is frequently negligibly small between an IV treatment of SB and one of SB plus SSB on an 'endogenous' variable, such as wage rate, e.g. see Blau and Kahn (2007). These findings suggest a very weak connection between SSB and SB, but a rather strong one between SSB and multicollearity, especially that of the substitutive type. The latter deems it

unidentifiable the correspondence between an inverse Mill's ratio denominated SSB and how much agents' self-selection behaviour really matters at the aggregate level with respect to the data at hand.

The connection between SSB and SB is finally lost in model based programme evaluation methods. Econometric developments of these methods grew mainly during the post 1980 period and drew heavily from the SSB literature, e.g. see Cameron (2009, 14.5) and Wooldridge (2010, 21.1). The developments are intimately related to causal modelling since the programme under study is undoubtedly the cause and its effects contain meaningful policy implications. Obviously, outcomes of any programmes are sequential to their implementation. Simultaneity is thus out of context. But self-selection behaviour is not because it could violate randomisation, a basic sampling convention for estimation of average treatment effects (ATE). When the ATE was adopted from medical science into evaluating social programmes, randomisation failures posed as a major challenge, e.g. see Heckman (1992). 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 reasons.[5] Heckman's SEM presentation of endogenous dummy variables (1978) provided a handy framework for tackling this issue. Once the ATE was attached to an endogenous dummy variable, SSB 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 vehemently promoted by Angrist and his associates through a series studies, see Angrist (1990), Angrist and Krueger (1991), and also Angrist and Pischke (2009). Their studies have helped popularise a wide conviction of the prevalence of endogeneity bias syndrome. Meanwhile, their applications have also aroused

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

serious debates over the interpretability of those IV-generated estimates as the ATE, e.g. see Angrist *et al* (1996, with discussion). In response, those estimates have been toned down as local ATE (LATE). Noticeably, this redefinition implies a partial recognition of the causal-modifying capacity of IVs, that is, IV-modified programme dummies might no longer fully represent the programme implemented in reality. 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 highbrow style of the causality dispute by the CC group in the context of SEMs in rivalry with Wold's causal chain models over half century ago, the causal modelling issues touched by Deaton are widely and closely relevant to policy related applied researches. Applications using the IV method have reached such an extensive state that the distortionary effect of IVs on the variable of causal interest is impossible to remain unheeded. Poor credibility in IV-estimates has thus turned out as a best exposure of the non-innocuous nature of the IV route for causal modelling. Consequently, the IV route has been increasingly abandoned in development economics, a trend similar to what we have observed in macro-econometric studies decades ago.

The recurring trends raise a serious question: What has led micro-econometric researchers being side-tracked into an infertile path already experienced and abandoned by macro researchers? The present investigation identifies endogenous bias syndrome as the culprit, because it has created an almost unfathomable conceptual muddle to keep many applied economists trapped from seeing appropriate ways of conducting data-assisted causal inference.

2. SB, OVB and SSB: An Anatomy

Since correlation between the error term and a particular explanatory variable of interest is *the* identifier of endogenous bias syndrome, the anatomy aims to reveal how that correlation is actually the mirror image of negating the conditional status of the explanatory variable in

question. To cater for as wide an applied readership as possible, mathematical demonstration is kept at minimum. Graphic illustrations are also used to facilitate clarification of conceptual differences in different types of bias and their symptomatic remedies. Although drawn in accordance, as much as possible, to the basic conventions of the causal graphic modelling literature, the graphs below are not designed for the purpose of computer-based causal modelling research.

2.1. SB (Simultaneity Bias):

Let us analyse simultaneity with a bivariate case for simplicity. When two variables are jointly distributed, elementary probability theory dictates the following density decomposition:

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

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. This underpins regression models such as:

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

Now, the decomposition in (1) is *de facto* refuted in the Haavelmo-CC position to take the SEM as their maintained hypothesis, although their position endorses joint distribution, $f_{x,y}$, as fundamental in econometrics. The refutation is embodied in their rejecting (2) in favour of an SEM, such as:

(3) $$\begin{aligned} y &= \beta_1 x + \varepsilon_1 \\ x &= \beta_2 y + \varepsilon_2 \end{aligned}$$

It is based on (3) that Haavelmo demonstrates SB of the OLS, $\beta_{yx} \neq \beta_1$, via $cov(x\varepsilon_y) \neq 0$. But such a bi-directional position on $f_{x,y}$ makes (3) mathematically impossible for any regression model based empirical studies. This impossibility is termed as 'under-identification' and circumvented by identification conditions. These conditions secure ways to decompose

$f_{x,y}$ *indirectly* with the help of additional exogenous variables,[6] variables which are regarded simply as *instruments* for the consistent estimation of the 'structural' parameters of the SEM, such as β_1 and β_2 in (3). Consistent estimation of a single equation in an SEM can be generically represented by the 2SLS, for instance, the upper equation in (3):

$$(4) \qquad \begin{aligned} x &= \gamma_{xI} I + u_x \\ y &= \beta_{yx_I} \hat{x}^I + \varepsilon_y^I \end{aligned}$$

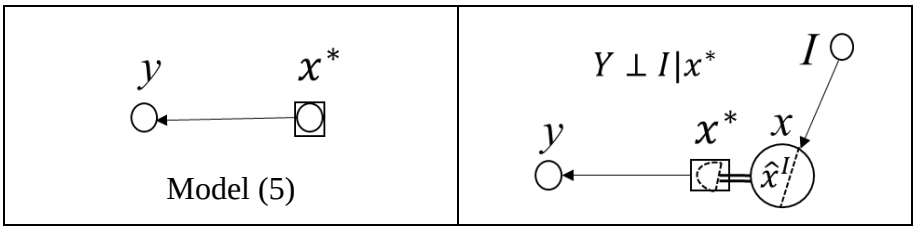
where *I* denotes the IV set and $\hat{x}^I$ the OLS fitted *x* from the upper equation. β_{yx_I} is known as the IV estimator for β_1 in (3).

In fact, this IV estimator is not innocuous for (3). It acts as an implicit model modifying device to break circular causality in SEMs, e.g. see Qin (2015). In order to better demonstrate this point, let us digress to the case of an errors-in-variables model in which the explanatory variable of interest is latent, or suffer from measurement errors, a case where the IV method was introduced into econometrics prior to its application to SEMs:

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

Here, IVs essentially serve as a means to trim off noisy errors, x'', from *x*, the observed counterpart of the latent and measurement-error free x^*. Figure 1 provides a graphic illustration of (5) and its IV treatment.

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



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

[6] 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.

It is vital to note from Figure 1 that the IV treatment entails two key conditions. (i) IVs should be uncorrelated to conditional expectation, $E_{y|x^*}$, and (ii) the IV trimming must not aim at achieving the optimal prediction of *x*, i.e. $\hat{x}^I \approx x$. The first condition is denoted by $y \perp I|x^*$ and the second the dotted semicircle symbol in the right panel. When the IV method is applied to SEMs, these two conditions hold in spite of the fact that *x* is no longer regarded as suffering from measurement errors. 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

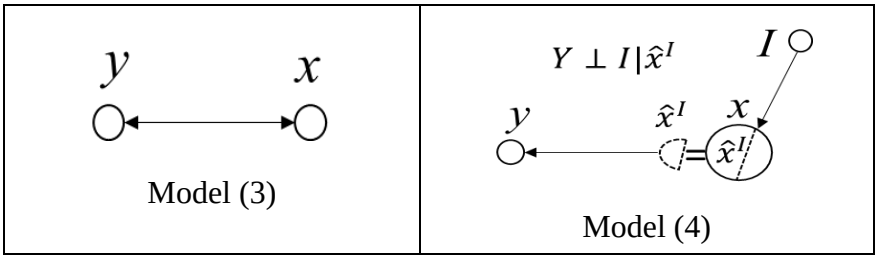


Figure 2 shows us how the bi-directional position in the left panel is broken by the introduction of IVs. In other words, how an SEM is revised into an acyclic or asymmetric model through a non-unique but significant modification of *x*. The fundamental bi-directional position of SEMs is thus falsified implicitly. The revised position suggests the following alternative to (1):

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

because condition (i), i.e. $y \perp I|\hat{x}^I$, enables the marginalisation over $\hat{x}^I$. This decomposition effectively refutes *x* as a valid conditional variable for *y*, since $\hat{x}^I \approx x$ by condition (ii).

Although unheeded in textbook econometrics, condition (ii) lies at the very foundation of the Durbin and Wu-Hausman endogeneity tests. Moreover, it explains why there are unceasing disputes over the credibility of IVs and the associated over-identification conditions. The nature of $\hat{x}^I$ being a non-optimal predictor of *x* determines $\hat{x}^I$ being non-unique, and hence there is room for arbitrary production of this 'non-causally' generated synthetic variable,

making use of high interdependence among many economic factors in a non-experimental environment. It also explains why endogeneity bias has lost its grip in macroeconometrics as the dynamic modelling approach becomes dominant. When prediction power raises with adequately specified dynamic models, such as VARs, condition (ii) fails because $\hat{x}^I \to x$, as rightly anticipated by Wold's proximity theorem. Consequently, the postulate, $E_{y|x}$, is resurrected.

A formal resurrection can be found in the re-definition of exogeneity by Richard (1980) and Engle *et al* (1983). The central theme of their work is to clarify under what conditions decomposition of (1) is valid. The conditions are related to circumstances under which f_x can be marginalised out, i.e. when a single equation model will suffice. Two conditions are identified – 'strong' exogeneity via time sequencing and 'super' exogeneity via cross-regime invariance. The latter is shown to be a valid circumstance for the marginalisation. The stance of their conditional decomposition departs methodologically from the Haavelmo-CC SEM paradigm.

The VAR project attempts to maintain the Haavelmo-CC paradigm by extending $f_{x,y}$ explicitly with a lagged information set so as to facilitate the decomposition: $f_{x,y,I_{-1}} = f_{x,y|I_{-1}} f_{I_{-1}}$. But that attempt is abortive. The thorny problem of contemporaneous conditional decomposition resurfaces via the covariance matrix of the error vector, when VARs are used for policy shock simulations. Consequent 'identification' remedies via the matrix amounts to forsaking the symmetry of joint $f_{x,y}$ for Wold's recursive model type. After all, statistically operational models have to start from a clearly specified 'asymmetry between cause and effect' Cox (1992, p293).

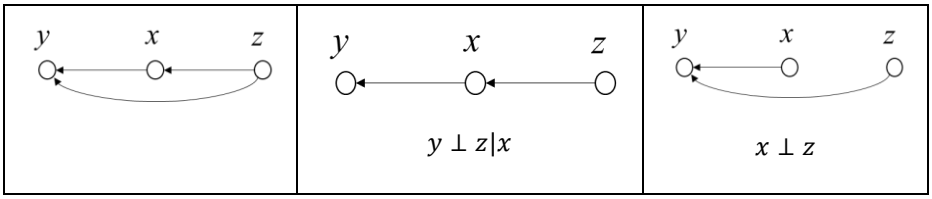
2.2 OVB (Omitted Variable Bias)

Let us now consider the possibility of (1) suffering from OVB due to omitting *z*, i.e. (1) should be extended to:

(7) $f_{x,y,z} = f_{y|x,z} f_{x|z} f_z.$

Three possible scenarios of (7) are illustrated in Figure 3, following the lucid demonstration by Cox and Wermuth (2004). In Figure 3, OVB is present only in the left panel, as *z* is correctly omitted in the middle panel whereas its omission is present in the right panel but not causing bias as far as the inference of $x \to y$ is concerned.[7]

Figure 3. Three Possible Scenarios of (7)



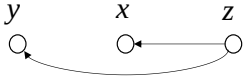
Under the linearity assumption, (7) corresponds to a chain of regressions:

(8) $y = \beta_{yx.z} x + \beta_{yz.x} z + \epsilon_y$

(9) $x = \beta_{xz} z + \epsilon_x$

with $\beta_{yx} = \beta_{yx.z} + \beta_{yz.x}\beta_{zx}$ and OVB being defined as $\beta_{yx} - \beta_{yx.z} = \beta_{yz.x}\beta_{zx}$. The mid and the right panels in Figure 3 can be expressed parametrically as (a) $\beta_{yz.x} = 0$ and (b) $\beta_{zx} = 0 = \beta_{xz}$ respectively. Case (a) is virtually non-existent in applied econometric studies for obvious reasons. Paradoxically, its omnipresence is taken for granted in all consistent estimators aiming at $cov(x\varepsilon_y) = 0$. Case (b) is relatively more realistic than (a) because it delimits the practical feasibility of partial explanations of *y*, a situation which appeals especially to analyses using non-experimental data. But the snag is that the feasibility does not apply to any explanatory variables *a priori* postulated as the substantive causes of interest. They have to satisfy $x \perp z$, a condition which looks unlikely to fulfil in general considering the extensive interdependent nature of economic variables.

[7] Omitted variables are often referred to as *confounding* variables in the statistical literature. However, our discussion of OVB disregards, as irrelevant, the following case associated with a confounding variable, *z*:



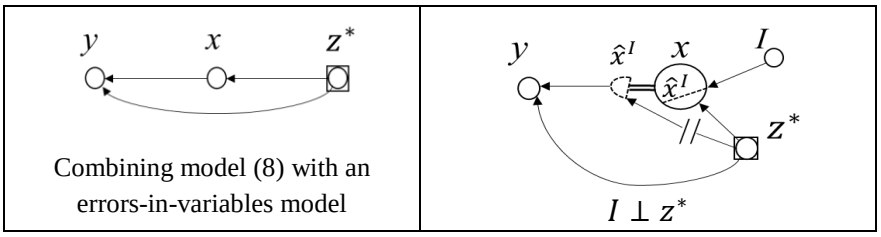
Remarkably, the IV method for endogeneity bias is used as a general way of bypassing the snag. Imagine the following situation. There lacks data information in observing *z* but it is known to be correlated with *x* in (8), on the basis of the general interdependency of economic variables. As such, (2) is an *inadequate* model as compared to (8). But the lack of data information on *z* leads to it being grouped into the error term, $\varepsilon_y = \beta_{yz.x} z + \epsilon_y$, and hence the diagnosis: $cov(x\varepsilon_y) \neq 0$. This symptomatic diagnosis entangles OVB with 'endogeneity bias' although simultaneity is a non-issue in either model. Consequently, the IV approach is taken as a generic remedy to remove the hurdle in case (b), with an additional requirement – the chosen IVs being uncorrelated with z^*, the latent *z*.

Similar to the SB case, the IV treatment is not innocuous for conditional decompositions, such as (7), which arise naturally from substantive postulates. The treatment modifies (8) by substituting *x* with $\hat{x}^I$. Since $\hat{x}^I \nsim x$, $\beta_{y\hat{x}^I.z} \neq \beta_{yx.z}$. Hence, it does not produce the parameter of interest originally specified in (8). Figure 4 illustrates this modification – from the first scenario in Figure 3 into the third scenario through deactivating the chain effect. The corresponding modification to (7) is:

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

Since z^* is latent and $\hat{x}^I \perp z^*$ is assumed to hold due to $I \perp z^*$ by design, $E_{y|\hat{x}^I}$ is thus regarded as an adequate model for measuring the causal effect: $x \rightarrow y$ on its own.[8]

Figure 4. Latent OVB and IV Treatment



[8] It should be noted that the orthogonal condition $I \perp z^*$ may results in modification of the modelled variable as well. This case is not discussed here in order to keep the demonstration simple (see more discussion on this point in the next section).

However, $E_{y|\hat{x}^I}$ on the basis of (10) leads to $\hat{x}^I \rightarrow y$ rather than $x \rightarrow y$. This explains why the IV treatment of OVB has met with unceasing scepticism over its credibility. Unfortunately, there lacks a wide awareness of the consequence of swapping $x \rightarrow y$ for $\hat{x}^I \rightarrow y$, because the two are assumed as causally equivalent in textbooks. This assumption is built on the very different status *I* is granted as compared to *x* or *z*. IVs are excluded from any joint distributions upon which regression models are formulated. In other words, the addition of *I* in (4) in the SB treatment does not augment $f_{x,y}$ in (1) to $f_{x,y,I}$, unlike what *z* does in (7) as compared to (1). Again, the IV addition in the OVB treatment does not lead to $f_{x,y,z,I}$ but $f_{\hat{x}^I,y,z^*}$ instead, as shown in (10). The exclusion reinforces the conviction that IVs are causally neutral, a substantively different property from that of exogenous variables.

The exclusion also provides us with an illuminating reason as why the IV route is so attractive empirically. It offers a shortcut in causal chain model search, probably the most daunting task in applied research. As shown in (10), the IV route allows modellers to work with $E_{y|\hat{x}^I}$ instead of $E_{y|x,\ldots}$, a magical psychosedative for them to feel immune from potential OVB risks in maintaining simple theoretical models with extremely partial explanatory power. Conceptually, the appeal of IV treatment builds vitally on the 'endogeneity bias' veneer of OVB. Since omission of factors which are potentially related to the key explanatory variables of interest is prevalent in applied econometric modelling, the argument of these variables being correlated with the error term appears much more realistic than the SB-based one. Attributing these variables as latent helps exempt the correlation from being directly testable and widen, at the same time, the general appeal of the case.

Nonetheless, there is a discernible gap between SB and OVB in terms of their behavioural connotations. A gap filler is the self-selection behaviour assumed of micro agents.

2.3 SSB (Self-Selection Bias)

The SSB issue arises from models in which the key explanatory variable of interest is data truncated. Let us modify (8) into a simple truncation model (assuming 0 as the threshold):

$$(11) \qquad \begin{array}{ll} y_i = \beta_{yx.w} x_i + \beta_{yw.x} w_i + \varepsilon_{i,y} & \text{if} \quad x_i > 0 \\ y_i = 0 & \text{if} \quad x_i = 0 \end{array}$$

where i denotes sample observation. Notice that *w* is used instead of *z* to indicate that we do not consider it as a possibly omitted variable here. It is hypothesised that the observable part of *x* is a biased representation of the population due to certain self-selection behavior of the agents concerned. This bias is treated as equivalent to OVB in Heckman's probit-OLS two-step procedure. The procedure exploits the truncation information by turning it into a binary LDV model of the presumed self-selection behavior:

$$(12) \qquad d_i = \gamma_{dI} I_i + u_{i,d} \qquad d_i = 1 \; when \; x_i > 0, \quad d_i = 0 \; when \; x_i = 0$$

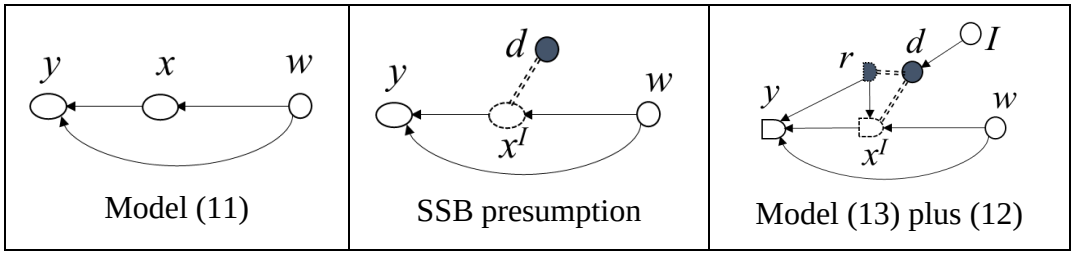
under the assumption: $cov(u_d \varepsilon_y) \neq 0$. The inverse Mill's ratio, r_i,[9] is then generated from probit regression of (12) and added to (11) to correct the presumed OVB in *x* due to SSB:

$$(13) \qquad y_i = \beta_{yx.rw} x_i + \beta_{yw.x} w_i + \beta_{yr.x} r_i + \varepsilon^I_{i,y} \qquad when \; x_i > 0$$

Let us now assume $w \perp I$ for simplicity. Although too strong as compared to $I \notin w$, the usually required exclusion restriction in practice, this assumption reveals clearest a methodologically vital difference between tobit and the Heckman procedure. While tobit corrects the LDV effect across all the explanatory variables indiscriminately, the Heckman procedure does not. The latter only targets at the parameter of a particular explanatory variable presumably suffering from self-selection bias, e.g. $\beta_{yx.w}$ alone in (11). Therefore, SSB differs from sampling bias of the general sense, in that self-selection behaviour via *d* could still be present in survey samples which are shown to be adequately representative of the population concerned. Graphic illustration of this feature is given in Figure 5.

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

Figure 5. Truncated Regression and the Heckman Two-step Procedure



Note: Truncated variables are represented by ovals and binary variables by solid nodes; a regression using only subsample data above the threshold is indicated by truncated ovals; dotted identities represent variable transformation.

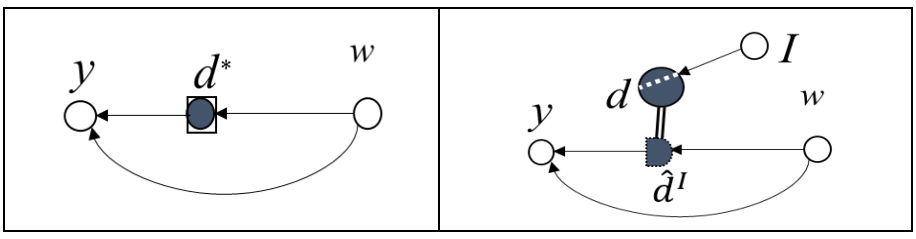
The right panel in Figure 5 shows noticeable similarity to the IV graphs in Figures 2 and 4. The similarity derives from the fact that *x* is endogenised, albeit indirectly via *d*, when (13) is appended by selection equation (12). The endogenisation is justified by presumed self-selection behaviour, diagnosed via error term correlation and circumvented by OVB correction. The substantive link between endogeneity bias and OVB is thus established, dispensing with simultaneity as a necessary theoretical justification.

The link is further turned into an obvious and direct one once *x* is replaced by *d*, e.g. see (Heckman, 1978), and the resulting model is applicable to variables not necessarily truncated:

$$\begin{aligned} y_i &= \beta_{yd.w} d_i + \beta_{yw.d} w_i + \varepsilon_{i,y} \\ d_i &= \gamma_{dI} I_i + u_{i,d} \end{aligned} \tag{14}$$

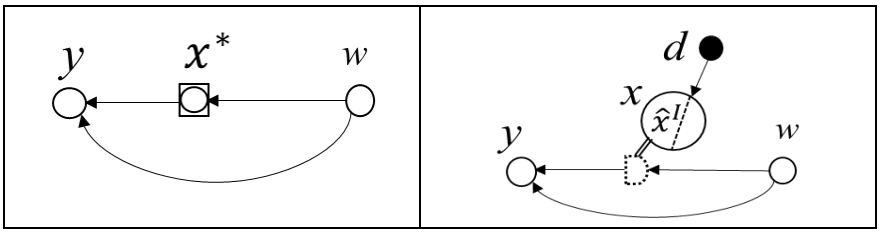
The above model forms the prototype of programme evaluation models. Since SSB is a top concern when randomisation comes under serious doubt, the IV route indicated in (14) offers an expedient remedy. The remedy effectively treats as latent the ideally randomised programme participation variable, d_i^*. The endogenized *d* thus turns (14) into an errors-in-variables model, as illustrated in Figure 6.

Figure 6. Endogenous Dummy Variable Model and IV Treatment



The link between endogeneity bias and measurement errors is further reinforced by an extension of (14) into a general latent variable model, in which SSB and/or OVB is believed to be the main problems, e.g. see Angrist and Krueger (1991). Here, social programmes are used as the ideal instruments to circumvent the bias, as illustrated in Figure 7. The extension serves as a best example of how endogeneity bias, diagnosed via $cov(x\varepsilon_y) \neq 0$, has now become an all-bias-inclusive syndrome.

Figure 7. A Variation and Extension of Figure 6



However, if we scan through the vast empirical literature, we cannot find any irrefutable evidence on how severe and prevalent endogeneity bias syndrome is. SB-based endogeneity has been virtually dismissed as practically insignificant by macro evidence so far. As for micro studies using large survey samples, it is simply impossible to seek such evidence when the size of residuals of multiple regression models takes the dominant share of data information. In other words, direct evidence of $cov(x\varepsilon_y) \neq 0$ is virtually intractable when the model fit is below 20% of data information and the control variable set is not small or unique. For one thing, there is no way to rule out that none of the control variables is not also prone to SB, SSB and/or measurement errors.

If we abandon the identifier, $cov(x\varepsilon_y) \neq 0$, and ponder over what exactly underlies the link between measurement errors and SSB, we should see that both share the *a priori* rejection of the observed *x* as a valid conditional variable. This does not apply to the OVB case, unless one stubbornly maintains as correct the model excluding those un-ignorable variables, e.g. (2) instead of (8), and justifies the exclusion by modifying the conditional variable. Only when the modification is camouflaged under consistent estimation, does it become possible to merge the

three cases. Furthermore, the OVB case shows how such covert modifications can help prolong inadequately closed empirical models from being falsified. This helps explain why endogeneity bias syndrome is placed so centrally in textbooks whose core part is devoted to consistent estimation techniques targeted at single explanatory variables.

3. Methodological Reflection

The above anatomy suggests that demystification of the endogeneity bias syndrome should entail a re-examination and clarification of two concepts – the error term and consistent estimation.

Let us look at the error term first. Error terms or model residuals have been long perceived as sundry composites of what modellers are unable and/or uninterested to explain since Frisch's time.[10] This perception is effectively refuted by textbook description of endogeneity bias syndrome. Since $cov(x\varepsilon_y) \neq 0$ is single-variable based, the contents of the error term have to be adequately 'pure', definitely not a mixture of sundry composites, to sustain its significant presence. Indeed, textbook discussions of endogeneity bias, be it associated with SB, measurement errors, OVB or SSB, are all built on simple regression models. As soon as these models are extended to multiple ones, the correlation becomes mathematically intractable. In a multiple regression, all the explanatory variables are mathematically equal. Designation of one as the causing variable of interest and the rest as control variables is purely from the

[10] Frisch classified statistical variations into three types – systematic variations, accidental variations and disturbances. The latter two formed the error term. In his view, 'accidental variations are variations due to the fact that a great number of variables have been overlooked, consciously or unconsciously, each of the variables being however of minor importance', whereas 'disturbances are variations due to the fact that one single, or a certain limited number of highly significant variables have been overlooked'; 'in economics we are actually often forced to throw so much into the bag of accidental variations that this kind of variations comes very near to take on the character of disturbances. In such cased it would perhaps be more rational to introduce a hierarchic order of types of variations, each type corresponding to the overlooking of variables of a certain order of importance.' (Frisch's 1930 Yale lecture notes, Bjerkholt and Qin, 2010, p.165). In the CC works, the error term was taken simply 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, the description has not been directly related to the error term of simple regression models. See also (Qin, 2013, Chapter 8) for a history of the error term in time-series econometrics.

substantive standpoint. The premise, $cov(x\varepsilon_y) \neq 0$, implies not only $cov(z\varepsilon_y) = 0$ for the entire set of control variables, but also the set being exhaustive. Both conditions are almost impossible to meet in practice.

Once we accept that credible applied econometric models cannot be simple regression based, it becomes clear that all endeavours to devise and apply IV-based estimators to circumvent $cov(x\varepsilon_y) \neq 0$ are on a wrong track. They amount to committing type III errors, i.e. producing 'the right answer to the wrong question' (Kennedy, 2002, p572). The anatomy of the previous section indicates to us what the correct question should be – whether and under what circumstances a postulated explanatory variable is a valid conditional variable, upon which credible causal inference is attainable from data. It is surely a serious blunder to modify this variable by non-causal information before the question has even been considered empirically.

This wrong question is, unfortunately, well camouflaged by the notion of consistent estimation. Here, consistency is anchored on $cov(x\varepsilon_y) = 0$. The flaw of this anchor has, hopefully, been adequately exposed in the above analysis. In fact, the meaninglessness of this anchor was criticised vehemently by Pratt and Schlaifer three decade ago (1984; 1988).[11] Essentially, their criticism builds on the primacy of tackling exhaustively the omitted variable problem in non-experimental data modelling. Without such an 'exhaustive exploration' first, in their view, it is impossible to acquire adequate knowledge of what the error term represents, upon which the textbook condition for consistency is based on. This viewpoint is simply summarised by Cox into an essential step of statistical inference – checking 'the consistency of the model with data' (2006, Chapter 1 and Appendix B). Referring to consistent estimation as internal consistency, Cox points out that 'although internal consistency is desirable, to regard it overwhelmingly predominant is in principle to accept a situation of always being self-

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

consistently wrong as preferable to some inconsistent procedure that is sometimes, or even quite often, right' (ibid, p199).

Even if we confine ourselves to the narrow notion of internal consistency, there is something seriously missing in practice – empirical verification of sample-to-population convergence, the very nature of consistency as an asymptotic property. Since consistent estimators targeting at $cov(x\varepsilon_y) = 0$ are not unique, it falls on applied economists to select appropriate ones for their specific purposes. Convergence is clearly unattainable with IV estimates which differ significantly across different samples of the same population and whose precision deteriorates with increasing sample sizes. Evidence of this kind is actually not difficult to find, e.g. see Qin *et al* (2014) for the case of labour supply models, and van Hüllen and Qin (2015) for the case of programme evaluation models. Interestingly, OLS estimates often turn out to be more precise and far less divergent than their IV counterparts under the circumstance, a contrast which can only be inferred as a rejection of those IV-modified synthetics being valid conditional variables. These findings teach us that internal consistency should not be presumed. It needs to be empirically tested via cross-sample invariance.[12]

Now, if endogeneity bias syndrome is the wrong anchor for more sophisticated estimators than the OLS, what should be the right anchor? Or in what situation are specifically designed estimators called for and proven to be useful? The case of cointegration techniques stands out from a historical reflection. It is worth noting that Granger conceived of what is known as the Engle-Granger two-step procedure on the basis of a dynamic model, particularly an error-correction model in which short-run and long-run effects were separately parameterised, e.g. see Ericsson and Hendry (2004) and also Qin (2013, Chapters 4 & 7). In such a dynamic model,

[12] 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.

the *a priori* parameters of interest are already identified as the steady state solution of the long-run effect. This identification offered Granger the anchor for his derivation of a consistent estimator. On the other hand, estimators, such as the Cochrane-Orcutt procedure, which were derived for a static model to circumvent its missing dynamics observed via residual autocorrelation, have turned out to be less useful for estimating the long-run effect. In fact, differentiation of the long-run effects from the short-run effects in a dynamic model has extended the types of causal parameters beyond what a static model or a growth model can embody. Estimators anchored on either model may not remain consistent to the extension.

The above contrasting cases deserve further attention, as they may shed light on the unceasing disputes over credibility of parameters of IV-based corrections for OVB in micro-econometrics. Observed residual autocorrelation of a static model using time-series data indicates that the model suffers significantly from omitted dynamic variable bias, a special case of OVB. The Cochrane-Orcutt estimator tries to correct the bias via the autocorrelated error term. However, this back-door route turns out to be too restrictive for the task of estimating the long-run parameters of a dynamic model in general, e.g. (Hendry, 1995, Section 7.7). Notice that the Cochrane-Orcutt estimator and similar devices are essentially IV estimators and the IV-based correction is shown as implicitly making a partial difference transformation on all the variables, e.g. both x and y in (2), see (Greene, 2003, Section 12.8). A special case of the transformation is growth models, or difference-in-difference models better known in the micro context. The transformation is thus not innocuous to the causal connotation of the parameter involved because of the redefinition of the original variables. It is not surprising that this error-term anchored route becomes disused once static models are abandoned for their inconsistency with data and *a priori* parameters of interest are mapped to the long-run steady state solution of dynamic models.

The above example illustrates clearly how misleading the error-term anchored estimation route can be for rectifying inconsistency of model with data. The dynamic modelling reforms in macroeconometrics in the wake of the 1973 oil shock can thus be seen as a major reorientation of research strategy. The reorientation has deepened our understanding of the importance of empirical verification of conditional premises via, at least, three key interrelated aspects. First, choose models with smallest error terms possible and ensure their probabilistic distributions in as close vicinity of the white-noise distribution as possible; second, choose models with relatively constant parameters associated with conditional variables, especially during periods with known regime shifts (the principle of super exogeneity); and third, reparameterise models with separate long-run and short-run effects to facilitate interpretation.

The third aspect has reduced the significance of SB considerably because possible simultaneity is now contained within the long-run part of a dynamic model, a part whose equilibrating effect is found to be generally small as compared to various short-run effects. The estimation aspect of this long-run part has been discussed above. Noticeably, the first two aspects are in concord with the direction of statistical research, over the last few decades, on the causal Markov condition in connection to adequacy of causal chain model designs, e.g. see Dawid (1979), Cox and Wermuth (1996; 2004), Pearl (2009).[13] A central issue of that research is how to determine the adequacy of causal model closure against potential OVB risk. The issue is turned into defining conditions for the empirical adequacy of conditional independence. The resulting conditions for collapsibility, ignorability, faithfulness and unconfoundedness are all anchored on empirical verification of the Markov property via the error term and the invariance capacity of the model under different sampling situation. Although the empirical background of that research is made of mainly cross-section data concerning medical and psychological

[13] Methodological implications of that research have also engaged the attention of philosophers, e.g. see Glymore (2010) and Russo (2014).

trials, it is striking how much in common those model design and evaluation conditions are with the criteria used in time-series based dynamic econometric modelling, as well as the principle of general-to-specific model searching approach, see Hendry (1995; 2009).

What is probably more striking is how mainstream micro-econometrics has resisted evolving in the same direction, considering its largely cross-section data based background. Seeking explanations to this divergence brings us to SSB. The pivotal role of SSB in consolidating endogeneity bias syndrome has already been shown in the previous section. Here, four reasons are identified to explain the divergence in connection with SSB. First, research topics of applied micro-econometric studies are mostly on measuring the effects of one or two particular causes of interest. These causes are taken as 'structural' *a priori* on substantive ground. The behavioural based SSB premise reinforces their 'structural' existence and curtails the realm of empirical model design effectively to selection of control variables, which are substantively unimportant anyway. Hence, the data-driven exploratory approach of causal structure learning pursued by statisticians appears redundant. Secondly, the majority of micro-econometric studies are *secondary* of data samples, i.e. data samples which are not specifically tailor-made for them. It is thus widely accepted that empirical micro models are extremely partial as far as their explanatory power of the modelled variables is concerned. Statistical devices which use certain desired distributions of the error terms as model selection criteria seem to be too remote to be applicable. In contrast, the IV-based remedy for SSB offers a simple and logically convincing way to bypass the key concerns in extremely partial model closure. Once consistency appears secured, failures from error-term diagnostic tests are seen as harmless. Thirdly, empirical evidence of SSB using the Heckman procedure is easy to obtain, thanks to widespread multicollinearity among economic variables from a wide variety

of data sources.[14] Here, IV applications have effectively turned multicollinearity from a pest into a prerequisite. It ensures adequate supply of IVs and widens the range of consistent estimates of the parameters of interest; these in turn justify micro modellers to stick to partial and simple models as long as the selection-on-unobservable problem is taken care of. Finally and probably most importantly, there lacks real-world demand for those parameters of interest being as precisely estimated as possible. Policy persuasion and evaluation is often the most practical purpose for empirical micro studies. While intricacy of the estimation procedure is likely to enhance persuasive power, neither predictive capacity nor precision of the estimates has counted much in the making of a good story. After all, precise magnitudes of parameters, such as wage elasticity and returns to education, seldom constitutes a prerequisite for social programme recommendations or evaluation. Once the structural status to those parameters have already been granted *a priori*, there seems little need to check how precise and invariant those parameter estimates are across different samples.

Although arbitrary choice of IVs has been recognised as the key weakness of the IV route in practice, the solidity of its SSB foundation remains unchallenged. This foundation has actually sustained microeconometrics evolving along a relatively self-contained methodology so far. The weakness appears rather minor, as it is confined to the 1st stage formulation, if viewed from the 2SLS representation, a stage which is not part of the assumed structural model anyway. But why has the substantive consequence of that stage – non-unique modifications of the explanatory variables of interest – fallen off the radar of econometric research for so long?

4. Lost in Translation

It transpires, hopefully adequately, from the last two sections that it should be a primary task of applied econometric studies to examine whether and under what circumstances

[14] The inherent linear feature of inverse Mill's ratio is clearly demonstrated by Puhani (2002). The link between multicollinearity and 'confluence', a fundamental concept for interdependency of economic variables, is discussed at length in Qin (2014).

explanatory variables of *a priori* postulated causal models are valid conditional variables, but visibility of the task has been blocked by endogeneity bias syndrome; the syndrome acts as a taboo barring translation of causal postulates directly into conditional decompositions such as (1).

This section delves mainly into two questions: What has led to this lost in translation? How costly is the lost? A statement by Cox summarising his life-long experience serves as the best starting point: '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' (2006, p197).

As shown above, what the Haavelmo-CC endeavour has done essentially is a formal translation of general-equilibrium models (GEM) into joint probability distribution based statistical models. Meanwhile, direct conditional decomposition of the joint is *de facto* rejected on the account of simultaneity. It is not until the redefinition of exogeneity that possible validity of such a conditional translation has been formally considered. Knowledge of specific dynamic circumstances, such as co-movement of long memory variables, upon which empirical validity of such conditional translations relies has accrued from further dynamic econometric research. The resulting evidence is so overwhelmingly strong that few macro economists would take static SEMs as a faithful translation of GEMs nowadays.

Unfortunately, *de facto* rejection of the direct conditional decomposition has been widely spread through textbook teaching of SB against the OLS. Emulations of the Haavelmo-CC paradigm in microeconometrics have resulted in OVB and SSB being conceptually entangled with SB to strengthen the taboo against the OLS. This preconceived prejudice is so strong that

there are numerous empirical cases where OLS results are simply dismissed when they are actually within the substantively feasible value range, much more precise and also less variant across different samples than their IV counterparts, or even when the value range of the latter falls outside prior expectations.

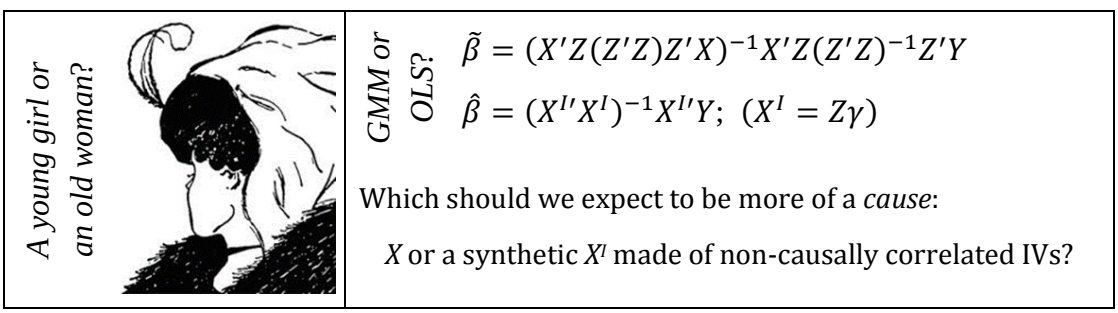
To a large extent, this prejudice manifests the striking difference between mainstream econometricians and statisticians in their attitudes towards data evidence. The former group is far more armchair-bound than the latter. This characteristic is almost hereditary. Of those major players at the CC who formalised econometrics into an academic discipline, few had first-hand experience with data and none was the data-experimental type. Their endeavour to anchor econometrics on provision of statistically optimal estimators for *a priori* postulated parameters of interest was highly restricted and Utopian – giving full faith in theorists for the supply of completely and correctly formulated models. It is not surprising that they failed to comprehend Wold's vehement arguments on the importance of conditioning via causal chain model designs. In fact, econometricians' awareness of the link between conditional expectation and regression models has been rather limited.[15] The basic correspondence between causal postulates and parametrically conditional models in statistics has been long lost in translation. Although the concept of conditional independence is included in textbooks nowadays, it is taught merely as an assumption pertinent to the statistical properties of the error term. The front door of mapping explanatory variables of interest directly to conditional variables is blocked solidly by endogeneity bias syndrome.

Within textbook econometrics, theoretical parameters are assumed as known links between causal variables, i.e. their 'structural' meaning is taken as *autonomous* to whatever

[15] In the *Econoemtric Theory* (ET) interview of David Hendry, he recalled how the audience at the 1977 European Econometric Society conference in 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 from the discussion of Wermuth (1992) between A.S. Goldberger and statisticians.

estimators from which their sample estimates are calculated. It is thus out of the realm of theorists' concerns that there lacks a unique estimator for a single 'structural' parameter. There is virtually no awareness that choice of estimators could affect the nature of their causal postulates. This oversight is further camouflaged by 'identification', i.e. 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 information brought in through identification conditions is assumed 'harmless' to the initial causal postulates (Angrist and Pischke, 2009). This explains why outbreaks of disputes over arbitrary identification conditions have never ceased among applied economists, whereas these disputes have hardly rippled the theoretical community. When identification information is systematically incorporated into IV estimators, its possibly distortional effect on the causal postulate of interest, if noticed at all, is put to the blame of poor IV choices in practice. The dual interpretation of the IV route remains overlooked, although the IV method was shown to function as 'generated regressor' producer by Pagan (1984) decades ago. From the econometric side, it is not difficult to explain the overlook either. After all, the IV method was originally adopted as an estimator to combat endogeneity bias syndrome. Its consolidation into the generalised method of moments (GMM) and the associate matrix notation further conceals its 2SLS representation, where IVs as an arbitrary modifier of the variable of causal interest is the most noticeable. A simple analogy of the issue of dual interpretability is shown in Figure 8.

Figure 8. Dual Interpretability: An Analogy



The overlook plus the predominantly *a priori* stance has thus kept mainstream econometrics trapped in endogeneity bias phobia for decades. The associate wrong question over correlation of a regressor with the error term arises from a direct mapping of bivariate regression models with the highly multivariate reality. In other words, endogeneity bias lives effectively within a prematurely closed bivariate model, as shown in section 2, whereas causal interpretation of its explanatory variable is directly linked, albeit elusively, with the most general economic phenomena – interdependence, omitted factors, measurement errors and self-selection behaviour. The bias will lose its grip and the wrong question be uprooted if we abandon such a schizophrenic position and start from a multivariate model setting, where model closure is to be empirically determined and where specific causal postulates are carefully embedded as testable null hypotheses.

Costly consequences of the translation loss are too manifest to ignore. While an impressive amount of techniques of amazing mathematical complexity have been produced along the estimation-centred route, their empirical yield is disproportionately meagre in terms of accuracy and invariance of parameter estimates, two basic properties of statistical inference.[16] Textbook preaching on endogeneity bias syndrome has misguided many applied economists into producing, accepting and justifying poor model fits and parameter estimates of low inference capacity, after toiling with various estimators on data, especially large household survey samples. The knowledge gained from their empirical research is simply not up to par with the advances of computing techniques and data availability. In the context of cross-section data modelling research, in particular, endogeneity bias phobia has effectively blocked the route of empirical verification of invariance as an asymptotic property in spite of available large sample sizes. Such verification entails recursive estimation, which is

[16] This situation actually fits what Freedman (1991) describes as an infertile way of using 'technical fixes' to rescue poorly designed models.

conditioned on a specific data ordering scheme. When the explanatory variables of interest are scale dependent, such as wage rate and income, data ordering schemes by their scales can reveal to us not only how asymptotically invariant their corresponding parameters are, but also finer gradient of possible nonlinear scale effects of easy interpretability than what quadratic forms of variables or quantile estimation could deliver, as shown in Qin and Liu (2013) and Qin *et al* (2014). Obviously, such data ordering schemes are out of consideration when endogeneity bias syndrome is taken for granted *a priori*.

On the other hand, the costs are not without academic reward. As shown in Section 2, the IV treatment of endogeneity bias syndrome serves as a powerful shortcut to facilitate applied modellers in churning out seemingly theory-consistent results from over-simplistic (inadequately closed) empirical models. It works almost like an efficient production line of apparently convincing stories. This situation is unlikely to change soon, not until there is a substantial rise of real-world demand for hard quantity based policy analyses. Stark warnings such as 'R^2 and *F* are irrelevant' (Pratt and Schlaifer, 1988, Abstract) in the absence of adequate empirical models for non-experimental data will remain being unheeded widely, until the day when exhaustive data exploration and experiments become so routinely indispensable for economists that the armchair culture is out of fashion in econometrics.

Nevertheless, a paradigm change looks inevitable in the long run. Rapid advances in computing power, software ease, data availability and the speed of knowledge exchanges across disciplines are all catalytic to the change, e.g. see Morgan (2013), Kalisch and Bühlmann (2014). Recent developments in statistics and machine learning have not only lowered the technical barrier in mapping causal postulates into statistical models, but also deepened our knowledge of basic conditions necessary for statistically adequate model closure. None of these conditions depends upon choices of complicated estimators. When it comes to the estimation step, statisticians' view is that parameters should 'have clear subject-matter interpretations' and

'statistical theory for estimation should be simple' (Cox, 2006, p13). This view is actually a fairly good summary of solid empirical studies by numerous experienced and data-experimental economists if we examine the history closely. Their choice through accruing experience with data suggests that conditional variables translated directly from substantive knowledge based causal postulates are often far more reliable than IV-generated synthetics which distort those variables through correlation-based association.

Admittedly, the above view is clearly against the teaching of textbook econometrics. Our lengthy discussion has already attributed this opposing position to the labyrinth of endogeneity bias syndrome. Here, the closing argument appeals to common sense. After all, success of applied models should stem from drawing together the relative advantages of both substantive knowledge and data information. Few can dispute the following. Substantive knowledge is relatively good at identifying key causes but not the precise functional forms through which these causes affect the modelled variable, nor other minor causes, with respect to the data at hand, which are not ignorable in estimating the effects of those key causes; on the other hand, data is the best possible source for filling the missing information. Is it not strategically wrong not to exploit data information to amend inadequately formulated empirical models, and instead, to close them prematurely by estimators which effectively modify the key causing variables non-causally, thus depriving them without trial of their conditional status originally implied in theory?

References

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.

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.

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.

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.

Dawid, A.P. (1979) Conditional independence in statistical theory (with discussion), *Journal of Royal Statistical Society B*. **41**, 1-31.

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

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.

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

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. (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.

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.

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.

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. (2014) Inextricability of confluence and autonomy in econometrics, *Oeconomia*, **4**(3), 321-41.

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. and Y.-M. Liu (2013) Modelling Scale Effect in Cross-section Data: The Case of Hedonic Price Regression. *Department of Economics Working Papers*, No. 184, School of Oriental and African Studies, London.

Qin, D., van Huellen, S., and Wang, Q.-C. (2014) What happens to wage elasticities when we strip playometrics? Revisiting married women labour supply model. *Department of Economics Working Papers*, No. 190, School of Oriental and African Studies, London.

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

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

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.

van Huellen, S. and D. Qin (2015) Compulsory schooling and the returns to education: A re-examination (memo), Department of Economics, SOAS.

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.